JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

On the Born-Infeld electron: Spin effects

Guy Boillat[a)]
C.I.R.A.M., Università, Via Saragozza 8, 40123 Bologna, Italy

Alberto Strumia
Dipartimento di Matematica, Università, Via Orabona 4, 70125 Bari, Italy

(Received 4 November 1997; accepted for publication 30 September 1998)

We start from a natural generalization of the Born–Infeld Lagrangian which involves two constants h,k and two parameters s,θ_0 and show, by investigating static solutions of a first order approximation with respect to a small parameter ($\epsilon = h/k$) that $s=1/2$ and that h is directly proportional to Planck's constant. It seems reasonable to interpret $s=1/2$ as the spin of the electron and the angle θ_0 as its orientation. Thus we obtain solutions that appear to reflect the influence of the states of spin on the electromagnetic field. © *1999 American Institute of Physics.* [S0022-2488(99)02901-1]

I. A GENERALIZATION OF THE BORN–INFELD LAGRANGIAN

The equations of nonlinear electrodynamics are derived from a nonlinear Lagrangian density $L(Q,R)$ which is a function of the electromagnetic invariants

$$Q=\tfrac{1}{4}F^{\alpha\beta}F_{\alpha\beta}=\tfrac{1}{2}(\mathbf{B}^2-\mathbf{E}^2),\quad R=\tfrac{1}{4}F^{\alpha\beta}F^{*}_{\alpha\beta}=\mathbf{E}\cdot\mathbf{B}, \tag{1}$$

and are given by

$$\partial_t(L_Q\mathbf{E}-L_R\mathbf{B})-\nabla\times(L_Q\mathbf{B}+L_R\mathbf{E})=0, \tag{2}$$

$$\partial_t\mathbf{B}+\nabla\times\mathbf{E}=0, \tag{3}$$

$$\nabla\cdot(L_Q\mathbf{E}-L_R\mathbf{B})=0, \tag{4}$$

$$\boldsymbol{\nabla}\cdot\mathbf{B}=0. \tag{5}$$

We observe that $\mathbf{E}$ and

$$\mathbf{H}=L_Q\mathbf{B}+L_R\mathbf{E}, \tag{6}$$

are the components of the *main field*[1] that symmetrizes the system of the evolutive equations (2) and (3).

According to the Maxwell equations the static electric field of a charged particle is given by the Coulomb's law. To solve the paradox of the infinite field at the center of the particle Born and Infeld[2] introduced new field equations derived from the nonlinear Lagrangian,

$$L(Q,R)=\sqrt{-R^2+k(2Q+k)}, \tag{7}$$

where k is a positive constant. In this way they obtained the famous static solution with spherical symmetry,

[a)]Permanent address: Department of Applied Mathematics, University of Clermont.

$$\mathbf{E}_0=\frac{\sqrt{k}}{\sqrt{1+\xi^4}}\,\frac{\mathbf{r}}{r},\quad \mathbf{B}_0=0\,,\quad \xi=\frac{r}{r_0},\tag{8}$$

which for $r=0$ yields the finite value $\sqrt{k}$ of the electric field also called the *absolute field* and for $r \gg r_0$ is equivalent to the Coulomb's law, provided that

$$\sqrt{k}=\frac{e}{r_0^2},\tag{9}$$

where e is the charge of the electron and r_0 is interpreted as the *radius of the electron*.

Calculating the total energy of the electrostatic field by integrating the Hamiltonian density over the whole space and assuming that the rest energy of the electron is of electromagnetic origin, they also find the relation

$$m_0c^2=1.2361\,\frac{e^2}{r_0},$$

which shows that r_0 is larger than the classical electron radius by the factor 1.2361,

$$r_0=3.47\times 10^{-13}\ \text{cm}.$$

It follows by (9) that

$$\sqrt{k}=3.97\times 10^{15}\ \text{e.s.u.}=1.19\times 10^{20}\ \text{V/m}.$$

''The enormous magnitude of this field, observe the authors, justifies the application of Maxwell's equations in their classical form in all cases, except those where the inner structure of the electron is concerned (field of the order $b[=\sqrt{k}]$, distance of wavelength of the order r_0).''

Now, as a consequence of nonlinearity, the wave fronts $\varphi(x^\alpha)=0$ are no longer null surfaces $(g^{\alpha\beta}\partial_\alpha\varphi\partial_\beta\varphi=0)$, but satisfy the characteristic equations,

$$\{F^{\alpha\rho}F^{\beta}{}_{\rho}-(2Q+\zeta)g^{\alpha\beta}\}\partial_\alpha\varphi\partial_\beta\varphi=0,\tag{10}$$

where ζ takes on two values $\zeta_1(Q,R)$, $\zeta_2(Q,R)$ depending on the choice of the Lagrangian.[3]

These two values coincide and turn out to be equal to the positive constant k only if L is the Born–Infeld Lagrangian (7).[3–5]

Thus the velocities given by (10) are double and as a consequence the corresponding waves are *exceptional (linearly degenerated)*. This concept introduced by Lax[6] appears to be of special importance in the applications.[1] The waves behave in a linear way and characteristic shocks traveling with the wave velocities exist. Therefore it seemed quite natural to look for Lagrangians leading to this property for each family of waves (10). The most general one has the form

$$L=F(\zeta)f(Q,R;\zeta)+RG(\zeta)+H(\zeta),\tag{11}$$

with

$$f(Q,R;\zeta)=\sqrt{-R^2+\zeta(2Q+\zeta)},$$

where $\zeta(Q,R)$ is obtained by solving the equation

$$\frac{\partial L}{\partial\zeta}=F'f+RG'+H'+(Q+\zeta)\,\frac{F}{f}=0.\tag{12}$$

Clearly the field equations obtained in this way are a direct generalization of the Born–Infeld equations, condition (12) resulting from the variation of ζ in the variational principle. Now either ζ_1 or ζ_2 may be used to yield the (same) Lagrangian. As a result the functions F,G,H are completely determined and using for instance ζ_1 for ζ they read[3]

$$F^2=\frac{1}{s(1+\gamma^2)}\left\{\frac{\gamma^2}{\zeta^2}(k^2-h^2)+2K\frac{\gamma}{\zeta}-1\right\}, \tag{13}$$

$$G=-\frac{\gamma}{\zeta}H,$$

$$H^2=\frac{1}{s(1+\gamma^2)}\{h^2-(\zeta-k)^2\}, \tag{14}$$

with

$$2K=\frac{k}{\gamma}\{s+1+\gamma^2(s-1)\},\quad s>0.$$

Setting

$$\gamma=\tan\frac{\theta_0}{2},$$

the two values ζ_1,ζ_2 are obtained by the following change of constants:

$$\zeta_1=\zeta(Q,R;k,h,s,\theta_0),\quad \zeta_2=\zeta(Q,R;k,h,-s,\theta_0+\pi), \tag{15}$$

and

$$(\zeta_1-k)^2\leqslant h^2\leqslant(\zeta_2-k)^2.$$

From (14) it appears that $\zeta=k$ when $h=0$ and (11) becomes equivalent to the Lagrangian of Born–Infeld (7).

Let

$$h=\epsilon k,$$

where ϵ is a small dimensionless parameter. Having in mind a model of a single electron it seems reasonable to think that, in this case, h is related to the Planck constant and s to the spin of the electron. As a matter of fact by (9) and the formula of the fine-structure constant,

$$\alpha=\frac{e^2}{\hbar c}\approx\frac{1}{137},$$

it follows

$$h=\epsilon\frac{c}{137r_0^4}\hbar.$$

II. FIRST ORDER APPROXIMATION

To illustrate the role played by the additional constants s,θ_0 appearing in the generalized Lagrangian (11) we look for an extension to the first order in ϵ of the classical Born–Infeld solution (8).

Since the exact solution of Eqs. (12)–(14) is not easy to find we expand the various quantities in terms of ϵ. We set

$$\zeta = k(1+\epsilon u+\cdots),$$

and retaining only the first order terms we get

$$F = 1 - \frac{s+1}{2s}\,\epsilon u,$$

$$G = \frac{\gamma}{\sqrt{s(1+\gamma^2)}}\,\epsilon\sqrt{1-u^2},$$

$$H = -\frac{k}{\sqrt{s(1+\gamma^2)}}\,\epsilon\sqrt{1-u^2},$$

$$f = f(Q,R;k) + \frac{\epsilon k(k+Q)u}{f(Q,R;k)},$$

$$L = f(Q,R;k) + \epsilon\left[\frac{k(k+Q)u}{f(Q,R;k)} - \frac{s+1}{2s}\,u f(Q,R;k) - (k-\gamma R)\sqrt{\frac{1-u^2}{s(1+\gamma^2)}}\right].$$

As we shall see later in Sec. IV, the present sign of H corresponds to the choice of a positive ϵ.

By (12) equating to zero the derivative with respect to u we obtain

$$\frac{u}{\sqrt{1-u^2}} = \frac{\sqrt{s(1+\gamma^2)}}{k-\gamma R}\left[\frac{s+1}{2s}\,f(Q,R;k) - \frac{k(k+Q)}{f(Q,R;k)}\right].$$

Or, simply when $R=0$,

$$u = \frac{[2Q-k(s-1)]\sqrt{1+\gamma^2}}{\sqrt{[k(s+1)+2Q]^2+\gamma^2[k(s-1)-2Q]^2}}.$$

To the first order in ϵ we have

$$\mathbf{E} = \mathbf{E}_0 + \epsilon\mathbf{E}_1 + \cdots\,, \quad \mathbf{B} = \epsilon\mathbf{B}_1 + \cdots\,, \tag{16}$$

$$Q = Q_0 + \epsilon Q_1 + \cdots = Q_0 - \epsilon\mathbf{E}_0\cdot\mathbf{E}_1 + \cdots\,, \quad R = \epsilon R_1 + \cdots = \epsilon\mathbf{E}_0\cdot\mathbf{B}_1 + \cdots\,,$$

$$u = \frac{(\phi-s)\sqrt{1+\gamma^2}}{\sqrt{(s+\phi)^2+\gamma^2(s-\phi)^2}}, \quad \phi = \frac{\xi^4}{1+\xi^4}, \tag{17}$$

$$L_Q = \frac{F\zeta}{f(Q,R;\zeta)} = \frac{1}{\sqrt{\phi}}\left\{1 - \frac{\epsilon}{2}\left[\left(\frac{1}{\phi}+\frac{1}{s}\right)u + 2\,\frac{Q_1}{k\phi}\right]\right\} + \cdots\,, \tag{18}$$

$$L_R = -R\,\frac{F}{f(Q,R;\zeta)} + G = -\epsilon\left[\frac{R_1}{k\sqrt{\phi}} - \frac{\gamma}{\sqrt{s(1+\gamma^2)}}\,\sqrt{1-u^2}\right] + \cdots\,. \tag{19}$$

With the substitution of (17) into (18) and (19),

$$u=\frac{\phi-s}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}, \tag{20}$$

$$L_Q=\frac{1}{\sqrt{\phi}}\left\{1-\epsilon\left[\frac{Q_1}{k\phi}+\frac{1}{2\phi s}\,\frac{\phi^2-s^2}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right]\right\}, \tag{21}$$

$$L_R=-\epsilon\left(\frac{R_1}{k\sqrt{\phi}}-\frac{\sin\theta_0\sqrt{\phi}}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right). \tag{22}$$

According to (2) and (3) the static solutions we are interested in, are derived from scalar potentials. Hence, by (16), (21), and (22),

$$\mathbf{H}_1=(L_Q\mathbf{B}+L_R\mathbf{E})_1=\frac{1}{\sqrt{\phi}}\,\mathbf{B}_1-\left(\frac{R_1}{k\sqrt{\phi}}-\frac{\sin\theta_0\sqrt{\phi}}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right)\mathbf{E}_0=\nabla\chi, \tag{23}$$

$$\mathbf{E}_1=\nabla\psi,$$

$$(L_Q\mathbf{E}-L_R\mathbf{B})_1=\frac{1}{\sqrt{\phi}}\,\mathbf{E}_1-\frac{1}{\phi^{3/2}}\left(\frac{Q_1}{k}+\frac{1}{2s}\,\frac{\phi^2-s^2}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right)\mathbf{E}_0. \tag{24}$$

Further from (4) and (5) we have

$$\nabla\cdot\left[\frac{1}{\sqrt{\phi}}\,\mathbf{E}_1-\frac{1}{\phi^{3/2}}\left(\frac{Q_1}{k}+\frac{1}{2s}\,\frac{\phi^2-s^2}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right)\mathbf{E}_0\right]=0, \tag{25}$$

$$\boldsymbol{\nabla}\cdot\mathbf{B}_1=0. \tag{26}$$

Multiplying (23) by $\mathbf{E}_0$ we find

$$R_1=\frac{1}{\sqrt{\phi}}\,\mathbf{E}_0\cdot\nabla\chi-\frac{\mathbf{E}_0^2\sin\theta_0}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}, \tag{27}$$

$$\mathbf{B}_1=\sqrt{\phi}\nabla\chi+\mathbf{E}_0\left(\frac{1}{k\sqrt{\phi}}\,\mathbf{E}_0\cdot\nabla\chi-\frac{\sin\theta_0}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right). \tag{28}$$

Substituting (24) into (25) we get:

$$\nabla^2\psi+\frac{1}{\xi^4}\,\frac{\partial^2\psi}{\partial r^2}-\frac{4}{r_0\xi^5}\,\frac{\partial\psi}{\partial r}-\frac{\sqrt{k}}{2s\sqrt{1+\xi^4}}\,\frac{d}{dr}\left(\frac{\phi^2-s^2}{\phi\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right)=0,$$

or

$$\frac{\partial}{\partial r}\left[\frac{1}{\phi\sqrt{1-\phi}}\,\frac{\partial\psi}{\partial r}-\frac{\sqrt{k}}{2s}\,\frac{\phi^2-s^2}{\phi\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}\right]+\frac{1}{r_0^2\sqrt{\phi}}\,D\psi=0, \tag{29}$$

where the Laplacian, in spherical coordinates, is given by

$$\nabla^2\psi=\frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial\psi}{\partial r}\right)+\frac{1}{r^2}\,D\psi$$

with

$$D\psi = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\varphi^2}.$$

III. ELECTRIC FIELD

The general solution to (29) is $\psi = \hat{\psi} + \psi_s$, where $\hat{\psi}$ satisfies

$$\nabla^2\hat{\psi} + \frac{1}{\xi^4}\frac{\partial^2\hat{\psi}}{\partial r^2} - \frac{4}{r_0\xi^5}\frac{\partial\overline{\psi}}{\partial r} = 0, \tag{30}$$

and $\psi_s(r)$ is a particular solution of the full equation.

From (8), (16), and (24) it follows that, at the first order, the electric field is given by

$$\mathbf{E} = \frac{\sqrt{k}}{\sqrt{1+\xi^4}}\frac{\mathbf{r}}{r} + \epsilon\left(\nabla\hat{\psi} + E_s\frac{\mathbf{r}}{r}\right), \tag{31}$$

where

$$\mathbf{E}_s = \boldsymbol{\nabla}\psi_s(r) = E_s\frac{\mathbf{r}}{r},$$

and E_s from (29) can be taken as

$$E_s(\xi;s,\theta_0) = \frac{\sqrt{k}}{2s}\frac{(\phi^2 - s^2)\sqrt{1-\phi}}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}. \tag{32}$$

On introducing the two vectors $\boldsymbol{\phi},\mathbf{s}$ of the respective moduli ϕ, s, making an angle θ_0, we can also write

$$E_s(\xi;s,\theta_0) = \frac{E_0(\xi)}{2s}\frac{\boldsymbol{\phi}+\mathbf{s}}{|\boldsymbol{\phi}+\mathbf{s}|}\cdot(\boldsymbol{\phi}-\mathbf{s}).$$

Since E_s is not a monotonic function of ϕ we require an extremum to take place on the sphere $r = r_0$, i.e., for $\xi = 1$, which implies

$$\cos\theta_0 = -\frac{1}{2}\frac{(4s^2+1)^2+(4s)^2}{4s(4s^2+1)},$$

a function of s which is always smaller than -1 except when

$$4s^2+1 = 4s \Leftrightarrow s = \tfrac{1}{2} \Leftrightarrow \theta_0 = \pi.$$

Thus the values of s, θ_0 are determined. This would not have been the case had the extremum occurred for another value of ξ (since a single equation—the vanishing of the derivative of E_s—is not sufficient, in general, to determine two quantities). Then the solution (32) becomes

$$E_s(\xi;1/2,\pi) = \sqrt{k}(\phi + \tfrac{1}{2})\sqrt{1-\phi}\ \mathrm{sgn}(\phi - \tfrac{1}{2}).$$

This electric field is a decreasing function of r inside the electron ($r < r_0$), is discontinuous across the surface $r = r_0$, decreases outside ($r > r_0$), and tends to zero at infinity like $1/r^2$ [see Fig. 1(f)]. To obtain the solution corresponding to the other orientation of the spin we take $\theta_0 = 0$ in (32) to get

$$E_s(\xi;1/2,0)=\sqrt{k}(\phi-\tfrac{1}{2})\sqrt{1-\phi}.$$

This field is continuous, vanishes on the surface $r=r_0$, has a maximum for $\phi=5/6$, and tends to zero like $1/r^2$ [see Fig. 1(a)]. Some intermediate states are illustrated in Fig. 1.

IV. ABSOLUTE FIELD

In nonlinear electrodynamics the quantities $\zeta_1^{1/2}$, $\zeta_2^{1/2}$ represent the absolute value of the electric field in a frame traveling at the ray velocity.[3]

From (15), (20),

$$\zeta_1=k+h\,\frac{\phi-s}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}=k+h\,\frac{|\boldsymbol{\phi}|-|\mathbf{s}|}{|\boldsymbol{\phi}+\mathbf{s}|},$$

$$\zeta_2=k+h\,\frac{\phi+s}{\sqrt{s^2+\phi^2+2s\phi\cos\theta_0}}=k+h\,\frac{|\boldsymbol{\phi}|+|\mathbf{s}|}{|\boldsymbol{\phi}+\mathbf{s}|},$$

and therefore with $s=1/2$,

$$\theta_0=\pi,\quad \zeta_1=\begin{cases}k-h, & r<r_0\\ k+h, & r>r_0\end{cases},\quad \zeta_2=\begin{cases}k+h(\tfrac{1}{2}+\phi)/(\tfrac{1}{2}-\phi) & r<r_0\\ k+h(\phi+\tfrac{1}{2})/(\phi-\tfrac{1}{2}) & r>r_0\end{cases},$$

$$\theta_0=0,\quad \zeta_1=k+h(\phi-\tfrac{1}{2})/(\phi+\tfrac{1}{2}),\quad \zeta_2=k+h.$$

Since both ζ's must be positive (for hyperbolicity),[3] also h is positive and this justifies the choice of the sign of H in Sec. II. Further $\zeta_1\leqslant\zeta_2$ and the inequality $\mathbf{E}^2\leqslant\zeta_1$ must hold, so that the wave and shock velocities do not exceed the speed of light.[1]

According to (10) a signal travels at light velocity when $r=r_0$ where ζ_2 is infinite.

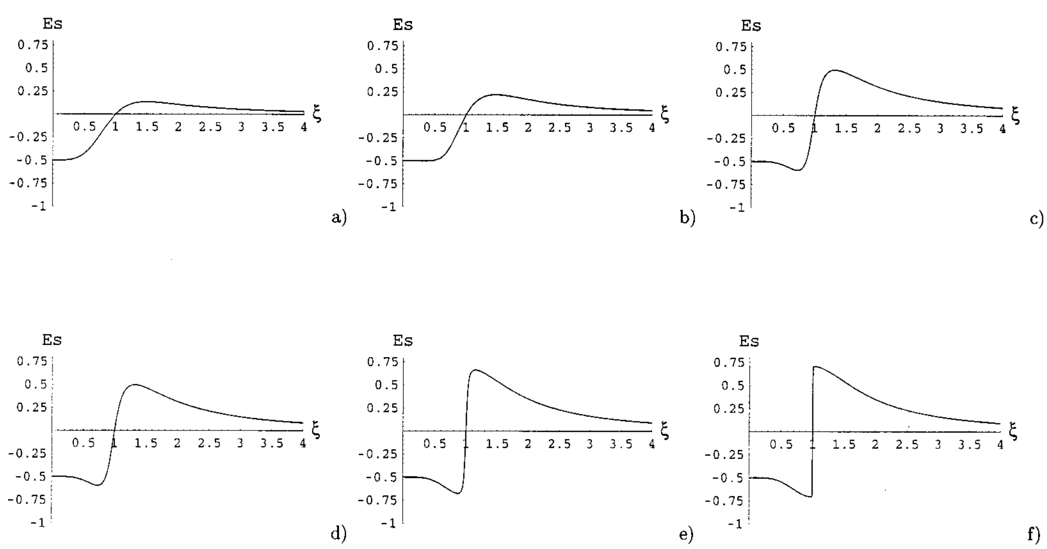

FIG. 1. Plot of E_s vs ξ, $(k=1)$: (a) $\theta_0=0$, (b) $\theta_0=\pi/5$, (c) $\theta_0=3\pi/5$, (d) $\theta_0=4.5\pi/5$, (e) $\theta_0=4.9\pi/5$, (f) $\theta_0=\pi$.

V. MAGNETIC FIELD

Turning back to the scalar field χ, substituting (28) into (26) we get

$$\frac{\partial}{\partial r}\left(\sqrt{1+\xi^4}\,\frac{\partial\chi}{\partial r}-\frac{\sqrt{k\phi}\,\sin\,\theta_0}{\sqrt{s^2+\phi^2+2s\phi\,\cos\,\theta_0}}\right)+\frac{\sqrt{\phi}}{r_0^2}\,D\chi=0.$$

The general solution is $\chi=\hat{\chi}+\chi_s$, where $\hat{\chi}$ is the general solution of the homogeneous equation and $\chi_s(r)$ is a particular solution of the full equation.

At the first order the field **H** is given by

$$\mathbf{H}=\epsilon\left(\nabla\hat{\chi}+H_s\,\frac{\mathbf{r}}{r}\right), \tag{33}$$

where

$$H_s(\xi;s,\theta_0)=\frac{\sqrt{k}}{\sqrt{1+\xi^4}}\,\frac{\sqrt{\phi}\,\sin\,\theta_0}{\sqrt{s^2+\phi^2+2s\phi\,\cos\,\theta_0}},$$

i.e.,

$$H_s=\frac{\sqrt{k}}{\xi^2}\,\frac{|(\boldsymbol{\phi}+\mathbf{s})\times\mathbf{s}|}{|\mathbf{s}||\boldsymbol{\phi}+\mathbf{s}|}.$$

This function vanishes together with its derivative for $\xi=0$ and inside the sphere of radius r_0 for

$$\xi^4=\frac{1}{\sqrt{5+4\,\cos\,\theta_0}}$$

has a positive maximum $H_s^{\max}$ defined by

$$(H_s^{\max})^2=\frac{2k}{1+2\,\cos\,\theta_0+\sqrt{5+4\,\cos\,\theta_0}}\,\sin^2\,\theta_0$$

(see Fig. 2).

When $\theta_0\rightarrow\pi$ this ratio, by L'Hospital's rule, tends to k. While for $\theta_0=0$ the H_s field is identically zero, for $\theta_0=\pi$ it is zero everywhere except on the sphere $r=r_0$, where it has a peak of height equal to $\sqrt{k}$ [see Fig. 2(f)].

To see this we write

$$s^2+\phi^2+2s\phi\,\cos\,\theta_0=(s-\phi)^2+4s\phi\,\cos^2\,\frac{\phi_0}{2}=\rho^2$$

with

$$s-\phi=\rho\,\cos\,\alpha,\quad 2\sqrt{s\phi}\,\cos\,\frac{\theta_0}{2}=\rho\,\sin\,\alpha,\quad 0\leqslant\alpha\leqslant\pi.$$

Then,

$$H_s=\frac{\sqrt{k}}{\sqrt{s(1+\xi^4)}}\,\sin\,\frac{\theta_0}{2}\,\sin\,\alpha.$$

When $\xi\to 1$, $\phi\to s=1/2$, $\theta_0\to\pi$, $H_s\to\sqrt{k}\,\sin\alpha$ for any value of α and therefore the maximum of H_s is $\sqrt{k}$. For another value of ξ then, $\rho\neq 0$, $\alpha=0$, $H_s=0$.

Now from (27) it follows

$$R_1=\frac{\sqrt{k}}{\xi^2}\,\frac{\partial\hat{\chi}}{\partial r}$$

and from (23),

$$\mathbf{B}_1=\frac{1}{\sqrt{1+\xi^4}}\left(\xi^2\nabla\hat{\chi}+\frac{1}{\xi^2}\,\frac{\partial\hat{\chi}}{\partial r}\,\frac{\mathbf{r}}{r}\right). \tag{34}$$

VI. SOLUTION TO THE HOMOGENEOUS EQUATIONS

The complete solutions for the electric (31) and magnetic field (33) include also the solutions to the homogenous equations for $\hat{\psi},\hat{\chi}$. Substituting (34) into (26) the latter equation can be written,

$$(1+\xi^4)\,\frac{\partial^2\hat{\chi}}{\partial\xi^2}+2\xi^3\,\frac{\partial\hat{\chi}}{\partial\xi}+\xi^2 D\hat{\chi}=0.$$

This means that $\hat{\chi}$ is a harmonic function with respect to the metric defined by

$$d\sigma^2=\phi d\xi^2+\xi^2(d\theta^2+\sin^2\,\theta\,d\varphi^2).$$

The former equation (31),

$$\left(1+\frac{1}{\xi^4}\right)\frac{\partial^2\hat{\psi}}{\partial\xi^2}+\frac{2}{\xi}\left(1-\frac{2}{\xi^4}\right)\frac{\partial\hat{\psi}}{\partial\xi}+\frac{1}{\xi^2}\,D\hat{\psi}=0$$

shows that $\hat{\psi}$ is a harmonic function for the conformal metric,

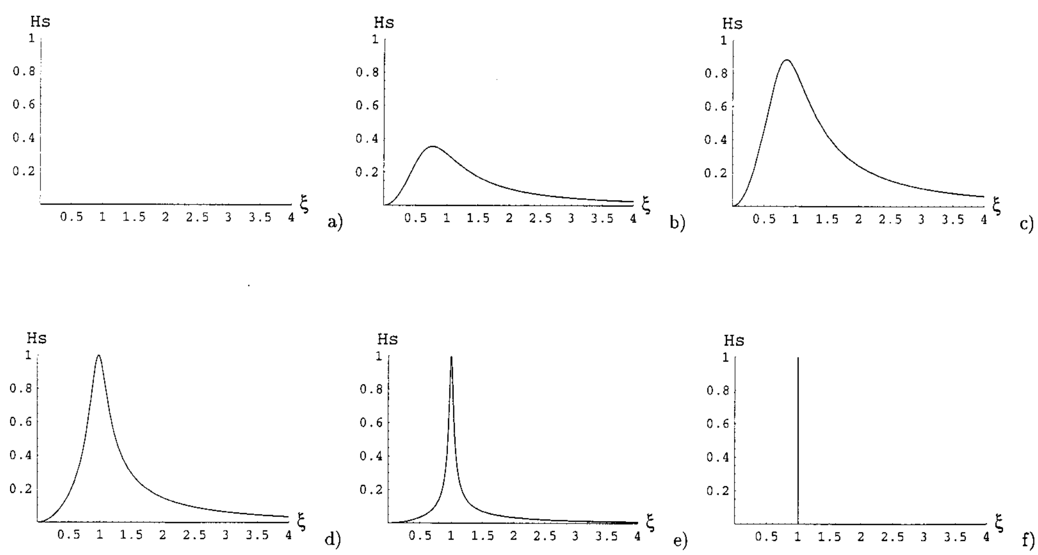

FIG. 2. Plot of H_s vs ξ, $(k=1)$: (a) $\theta_0=0$, (b) $\theta_0=\pi/5$, (c) $\theta_0=3\pi/5$, (d) $\theta_0=4.5\pi/5$, (e) $\theta_0=4.9\pi/5$, (f) $\theta_0=\pi$.

$$d\sigma^* = \frac{1}{\phi}\, d\sigma.$$

From (31), (32) we have:

$$\mathbf{E}^2 = \mathbf{E}_0^2 + 2\epsilon \mathbf{E}_0 \cdot \mathbf{E}_s + 2\epsilon \mathbf{E}_0 \cdot \nabla \hat{\psi}$$

and for $r=0$, at the center of the particle, we must have[3]

$$\mathbf{E}^2 = \zeta_1 = k - h,$$

i.e.,

$$\left(\frac{\partial \hat{\psi}}{\partial r}\right)_{r=0} = 0.$$

We point out that even if the potentials $\psi = \hat{\psi} + \psi_s$, $\chi = \hat{\chi} + \chi_s$ include the general solutions $\hat{\psi}, \hat{\chi}$ to the homogenous equations, of which nothing more at present can be said, the only terms which are completely determined, are the special solutions $\psi_s(\xi; s, \theta_0)$, $\chi_s(\xi; s, \theta_0)$ of the inhomogeneous equations, which have been analyzed in the previous sections.

VII. CONCLUSION

Recently there has been a renewal of interest for the Born–Infeld theory due partly to its link with relativistic strings, membranes[7] and gravitation theory[8] and partly to its nonlinear structure. In fact, in the last years, more researchers seem to be interested in the nonlinear nature of the physical phenomena.

In our investigation of the static field of a single electron we started from the observation that a natural generalization of the Born–Infeld Lagrangian involves, apart from the classical absolute field k (already introduced by Born and Infeld), a constant h, a dimensionless parameter s, and an angle θ_0.

By looking for static solutions at the first order of approximation with respect to the small parameter $\epsilon = h/k$ we find that the field is the sum of two contributions; one of them is a harmonic function in a suitable metric space, and not much can be said about it. The other one depends explicitly on the mentioned constants and can be studied. It results that a maximum for the electric field occurs on the surface of the electron $(r = r_0)$ only for the values $s = 1/2$, $\theta_0 = \pi$. Furthermore when $\theta_0 \neq \pi$ this solution vanishes for $r = r_0$ and this seems to reflect the influence of the states of spin on the electromagnetic field which is discontinuous when the angle θ_0 is equal to π.

So it appears reasonable, in this context (single electron), to relate s to the spin of the electron and θ_0 to its orientation (up and down), while h is proportional to the Planck's constant.

ACKNOWLEDGMENTS

We are grateful to Professor Tommaso Ruggeri for his kind hospitality at the C.I.R.A.M. We thank the referee whose observations were of great help in improving the manuscript. This work was supported by the Gruppo Nazionale per la Fisica Matematica (G.N.F.M.) of the Consiglio Nazionale delle Ricerche (C.N.R.).

[1] G. Boillat, ''Nonlinear Hyperbolic Fields and Waves,'' in *CIME Course Recent Mathematical Methods in Nonlinear Wave Propagation, Montecatini, 1994, Lecture Notes in Mathematics, edited by T. Ruggeri (Springer, Berlin, 1996), Vol. 1640, pp. 1–47.*
[2] M. Born and L. Infeld, Proc. R. Soc. London, Ser. A **144**, 425 (1934).
[3] G. Boillat, J. Math. Phys. **11**, 941 (1970).
[4] J. Plebański, *Lectures on Nonlinear Electrodynamics*, presented at the Niels Bohr Institute and NORDITA in 1968 (NORDITA, Copenhagen, 1970).
[5] I. Bialynicki-Birula, *Nonlinear Electrodynamics: Variations on a Theme by Born and Infeld* in *Quantum Theory on*

Particles and Fields, edited by B. Jancewicz and J. Lukierski (World Scientific, Singapore, 1983).
[6] P. D. Lax, Commun. Pure Appl. Math. **10**, 537 (1957).
[7] M. De Groot and P. Mansfield, Phys. Lett. B **231**, 245 (1989); A. Strumia, Nuovo Cimento B **110**, 1497 (1995); A. A. Tseytlin, Nucl. Phys. B **469**, 51 (1996).
[8] D. Palatnik, ''Born-Infeld theory of gravitation: Spherically symmetric static solutions,'' preprint, quant-ph/9701017, Vol. 4, 20 Jan 1997.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

Thermalization of quantum states

Dorje C. Brody[a)]
Department of Applied Mathematics and Theoretical Physics, Cambridge University, Silver Street, Cambridge CB3 9EW, United Kingdom

Lane P. Hughston[b)]
Department of Mathematics, King's College London, The Strand, London WC2R 2LS, United Kingdom

(Received 22 June 1998; accepted for publication 3 September 1998)

An exact stochastic model for the thermalization of quantum states is proposed. The model has various physically appealing properties. The dynamics are characterized by an underlying Schrödinger evolution, together with a nonlinear term driving the system toward an asymptotic equilibrium state and a stochastic term reflecting fluctuations. There are two free parameters, one of which can be identified with the heat bath temperature, while the other determines the characteristic time scale for thermalization. Exact expressions are derived for the evolutionary dynamics of the system energy, the system entropy, and the associated density operator. [S0022-2488(99)01901-5]

I. INTRODUCTION

The construction of a microscopic description for dynamical systems out of equilibrium is a topic of considerable interest in statistical mechanics. While many studies have been pursued in this direction,[1] the subject remains refractory owing to its nonlinear character. In particular, a realistic framework must be simple enough to be tractable, yet sufficiently sophisticated to capture some of the physical essentials. Our goal here is to propose in this spirit an elementary but compelling stochastic model that characterizes the thermalization process for a quantum system in an arbitrary given initial state.

The approach here, though carried out largely within the framework of standard quantum theory, involves a reexamination of the traditional hypotheses forming the basis of quantum statistical mechanics. In particular, we shall take the view that a quantum statistical ensemble, or ''state,'' is correctly described, not simply by a density matrix ρ^{α}_{β}, but rather by a probability distribution on the space of pure quantum mechanical states. For measurements associated with ordinary quantum mechanical observables, this description reduces to the usual one, for which the unconditional expectation of a linear observable is given by the familiar trace formula involving ρ^{α}_{β}. The advantage of the present formulation, however, is that we can give meaning to the concept of ensembles and other mixed states in nonlinear quantum mechanics.[2,3] Furthermore, the fact that the space of pure states has a symplectic structure allows us to introduce the concept of a quantum phase space Γ, and hence the use of probability distributions on Γ as ensembles. In particular, for thermal equilibrium in certain situations, we propose the distribution given in (12) below, which we call the canonical Γ ensemble. The corresponding density matrix differs in some respects from the conventional density matrix of quantum statistical mechanics, and thus the applicability of the theory may, in principle, be testable.

Our first step is to derive the Fokker–Planck equation for the time-dependent probability density function associated with a general Brownian motion with drift on the space of pure quantum states. Since Γ is compact, we can apply a theorem of Zeeman to show that there is a special class of drift terms for which, given any initial condition, the resulting asymptotic station-

[a)]Electronic mail: d.brody@damtp.cam.ac.uk
[b)]Electronic mail: lane_hughston@yahoo.com

ary solution is the canonical Γ ensemble described above. The nonequilibrium dynamics we consider here is governed by a nonlinear Schrödinger evolution coupled to a thermal noise term, described by the stochastic differential equation (14). This choice is the simplest possible that leads to thermal equilibrium for arbitrary initial states.

In what follows we use the dynamical law (14) to derive a formula for the total system energy (18), leading to a formula for the heat capacity (19). These ideas are applied, by way of illustration, to the case of a spin one-half particle in a heat bath. Then we derive an expression for the dynamics of the system entropy, given in (24), from which we are able to deduce a fundamental thermodynamic inequality (25). Finally, we study the evolution of the density matrix associated with the underlying process. We show that the time development of the density matrix involves a higher moment of the projection operator onto pure states.

II. QUANTUM HAMILTONIAN DYNAMICS

We begin by considering the quantum mechanical phase space CP^n. This is the space of pure quantum states, given by rays through the origin in the associated Hilbert space.[2] The quantum phase space can be viewed as a $2n$-dimensional real manifold Γ, endowed with the Fubini–Study metric g_{ab}. The reason that Γ plays the role of a quantum phase space is that Hamilton's equations on Γ can be lifted to the Hilbert space to give the Schrödinger dynamics. This follows from the fact that Γ has a natural symplectic structure, given by a nondegenerate antisymmetric tensor Ω_{ab}, compatible with the metric g_{ab} in the sense that $\nabla_a\Omega_{bc}=0$ and $\Omega_{ab}\Omega^{bc}=-\delta_a^c$, where $\Omega^{ab}=g^{ac}g^{bd}\Omega_{cd}$ and ∇_a denotes the standard covariant derivative, satisfying $\nabla_a g_{bc}=0$. Then the Schrödinger dynamics take the Hamiltonian form

$$dx^a=2\Omega^{ab}\ \nabla_b H\ dt, \tag{1}$$

where dx^a/dt is tangent to the quantum phase space trajectory. The Hamiltonian function $H(x)$ on Γ is given by the expectation of the energy operator at each point $x\in\Gamma$. More specifically, the expectation of the Hermitian operator H^α_β with respect to the Hilbert space vector ψ^α $(\alpha,\beta=0,1,...,n)$ is given by $H(x)=H^\alpha_\beta\bar{\psi}_\alpha\psi^\beta/\bar{\psi}_\gamma\psi^\gamma$. Since each point $x\in\Gamma$ corresponds to an equivalence class $\{\lambda\psi^\alpha,\lambda\in\mathbf{C}-0\}$ for some Hilbert space vector ψ^α, it follows that $H(x)$ is defined globally on Γ. Conversely, one can show that such functions correspond to global solutions of the equation

$$\nabla^2 H=(n+1)(\bar{H}-H), \tag{2}$$

where ∇^2 is the Laplace–Beltrami operator on Γ. Here $\bar{H}=H^\alpha_\alpha/(n+1)$, H^α_α being the trace of the Hamiltonian. Thus $\bar{H}$ is the uniform average of the energy eigenvalues, corresponding to the thermal equilibrium energy of the system at high temperature.

III. STOCHASTIC DYNAMICS

Now we generalize the Schrödinger dynamics by consideration of a diffusion process x_t taking values in Γ. For this purpose we make use of standard techniques of stochastic differential geometry.[4,5] We shall examine the case of a Brownian motion with drift on Γ, determined by the covariant stochastic differential equation,

$$dx^a=\mu^a\ dt+\kappa\sigma_i^a\ dW_t^i, \tag{3}$$

where κ is a constant, μ^a is a vector field, and the vectors σ_i^a $(i=1,2,...,2n)$ constitute an orthonormal basis in the tangent space of Γ such that $g^{ab}=\sigma_i^a\sigma_j^b\delta^{ij}$ and $\sigma_i^a\sigma_j^b g_{ab}=\delta_{ij}$. Here dx^a is the covariant Ito differential,[5] and the standard $2n$-dimensional Wiener process W_t^i satisfies $dW_t^i\ dW_t^j=\delta^{ij}\ dt$. In local coordinates $x^{\mathbf{a}}$ $(\mathbf{a}=1,2,...,2n)$, the Ito differential is given by

$$dx^a = \delta^a_{\mathbf{a}}(dx^{\mathbf{a}} + \tfrac{1}{2}\kappa^2 \Gamma^{\mathbf{a}}_{\mathbf{bc}} g^{\mathbf{bc}}\, dt). \tag{4}$$

Here $\delta^a_{\mathbf{a}}$ is a coordinate basis for the given patch of Γ, and $\delta^{\mathbf{a}}_a$ the dual basis, such that for the covariant derivative $\nabla_a \xi^b$ of a vector field ξ^a with components $\xi^{\mathbf{a}} = \delta^{\mathbf{a}}_a \xi^a$, we can write

$$\delta^b_{\mathbf{b}} \delta^{\mathbf{a}}_a (\nabla_b \xi^a) = \frac{\partial \xi^{\mathbf{a}}}{\partial x^{\mathbf{b}}} + \Gamma^{\mathbf{a}}_{\mathbf{bc}} \xi^{\mathbf{c}}. \tag{5}$$

The Ito differential (4) is constructed in such a way that the stochastic differential equation (3) is fully tensorial.

Suppose $\phi(x)$ is a smooth function on Γ, and define the process $\phi_t = \phi(x_t)$. It follows from Ito's lemma that $d\phi_t = \nabla_a \phi\, dx^a + \frac{1}{2}\nabla_a \nabla_b \phi\, dx^a\, dx^b$. Then the relation $dx^a\, dx^b = \kappa^2 g^{ab}\, dt$ implies

$$d\phi_t = (\mu^a\, \nabla_a \phi + \tfrac{1}{2}\kappa^2\, \nabla^2 \phi) dt + \kappa\, \nabla_a \phi\, \sigma^a_i\, dW^i_t, \tag{6}$$

where $\nabla^2 \phi = g^{ab} \nabla_a\, \nabla_b \phi$. Since the process x_t is Markovian, there is a normalized density function $\rho(x,t)$ for the state at time t, characterized by a partial differential equation subject to specified initial conditions. The expectation of the process ϕ_t is thus

$$E[\phi_t] = \int_\Gamma \phi(x) \rho(x,t) dV, \tag{7}$$

where dV denotes the volume element on Γ. On the other hand, it follows from Ito's lemma (6) that

$$\phi_t = \phi_0 + \int_0^t \left(\mu^a\, \nabla_a \phi + \frac{1}{2}\, \kappa^2\, \nabla^2 \phi \right) ds + \kappa \int_0^t \sigma^a_i\, \nabla_a \phi\, dW^i_s. \tag{8}$$

Here the integrands are valued at time s at the point $x_s \in \Gamma$. Since the second integral above has vanishing expectation, we obtain

$$E[\phi_t] = \phi_0 + E\left[\int_0^t \left(\mu^a\, \nabla_a \phi + \frac{1}{2}\, \kappa^2\, \nabla^2 \phi \right) ds \right]. \tag{9}$$

Hence, differentiating (7) and (9) with respect to t and equating the results, we have

$$\int_\Gamma \dot{\rho}(x,t) \phi(x) dV = \int_\Gamma \left(\mu^a\, \nabla_a \phi + \frac{\kappa^2}{2}\, \nabla^2 \phi \right) \rho(x,t) dV, \tag{10}$$

where $\dot{\rho} = \partial \rho / \partial t$. Integrating by parts, and using the fact that the resulting relation must hold for all $\phi(x)$, we find that $\rho(x,t)$ satisfies

$$\frac{\partial}{\partial t}\, \rho(x,t) = -\nabla_a(\mu^a \rho) + \frac{1}{2}\, \kappa^2\, \nabla^2 \rho. \tag{11}$$

This is the covariant Fokker–Planck equation for the density function on Γ, corresponding to the stochastic process (3). The solution of (11) thus characterizes the distribution of the diffusion (3) at time t, given an initial distribution $\rho(x,0)$. In the case of a singular initial distribution, e.g., an initially pure state, we interpret $\rho(x,t)$ as a generalized function, i.e., it has to satisfy (10) for all smooth $\phi(x)$.

The foregoing results are valid on arbitrary compact Riemannian manifolds. We are concerned, however, with the case where Γ is the quantum phase space CP^n, endowed with the Fubini–Study metric. Since Γ is compact, Eq. (11) may admit nontrivial asymptotic ($t \to \infty$)

stationary solutions. We shall examine a simple situation in which this is the case, namely, when the drift is given by a gradient flow $\mu^a = -(1/2)\kappa^2\beta\,\nabla^a H$ generated by $H(x)$, where β is a parameter. Then we can use a theorem of Zeeman[6] to show that there is a unique stationary solution for (11) of the form

$$\rho(x) = \frac{\exp(-\beta H(x))}{Z(\beta)}, \tag{12}$$

where $Z(\beta) = \int_\Gamma \exp(-\beta H(x))dV$. It follows that the probability density function $p(E)$ for the distribution of the energy function $H(x)$ is $P(E) = \Omega(E)\exp(-\beta E)/Z(\beta)$, where

$$\Omega(E) = \int_\Gamma \delta(H(x) - E)dV \tag{13}$$

is the density of states in Γ for which $E \leqslant H(x) < E + dE$, and $Z(\beta)$ is the partition function. Another way of characterizing the distribution (12) is that it maximizes the entropy $S = -\int_\Gamma \rho(x)\ln \rho(x)dV$ for a given value of system energy $U = \int_\Gamma \rho(x)H(x)dV$. The theorem of Zeeman implies in the present context that (12) is the asymptotic distribution of the process x_t under essentially arbitrary initial conditions, and that $p(E)$ is the asymptotic energy distribution.

We have so far considered a process of the form (3) with the drift $\mu^a = -(1/2)\kappa^2\beta\,\nabla^a H$. This process, as such, does not yet make reference to the Schrödinger dynamics. If we view the drift term as a nonlinear correction to the underlying quantum evolution, we can represent the complete dynamics according to the prescription

$$dx^a = \left(2\Omega^{ab}\,\nabla_b H - \frac{1}{2}\,\kappa^2\beta\,\nabla^a H\right)dt + \kappa\sigma_i^a\,dW_t^i, \tag{14}$$

which in the limit $\kappa \to 0$ reduces to the ordinary Schrödinger dynamics. Due to the antisymmetry of Ω_{ab}, inclusion of the symplectic term does not affect the resulting asymptotic state (12), since $\rho(x)$ is a function of $H(x)$. Hence, the analysis above shows that, given an arbitrary initial state (pure or general) on the quantum phase space, the dynamical law (14) necessarily takes that state into the thermal equilibrium state (12). In particular, the process (14) is ergodic on the energy surface, and asymptotically leads to a uniform distribution on each such surface. We note that (14) involves two parameters, β and κ. The stationary solution for $\rho(x,t)$ depends only on β, which we identify as the inverse of the heat bath temperature. The parameter κ, which has the units $s^{-1/2}$, and may depend on β, determines the thermalization time scale.

IV. HAMILTONIAN PROCESS

Our next step is to study the stochastic process $H_t = H(x_t)$ associated with the Hamiltonian function. From Ito's lemma (6) for the process (14) we find

$$dH_t = \tfrac{1}{2}\kappa^2(\nabla^2 H_t - \beta V_t)dt + \kappa\,\nabla_a H\,\sigma_i^a\,dW_t^i. \tag{15}$$

Here $V_t = V(x_t)$, where $V(x) = g^{ab}\,\nabla_a H\,\nabla_b H$ is the squared energy uncertainty at the point $x \in \Gamma$, conditional on the pure state to which x corresponds. Integrating (15) and taking its expectation, we have

$$E[H_t] = H_0 + \frac{1}{2}\,\kappa^2\int_0^t E[\nabla^2 H_s - \beta V_s]ds. \tag{16}$$

Let U_t denote $E[H_t]$, the unconditional energy expectation, which can be interpreted as the total system energy at time t. Then by differentiating (16) we obtain

$$\frac{\partial U_t}{\partial t} = \frac{1}{2}\kappa^2((n+1)(\overline{H}-U_t)-\beta E[V_t]), \tag{17}$$

by use of the Laplace–Beltrami equation for H_t. This relation can be integrated to yield the solution for the time development of the system energy:

$$U_t = \overline{H} + (H_0 - \overline{H})e^{-(1/2)\kappa^2(n+1)t} - \frac{1}{2}\kappa^2\beta\int_0^t\int_\Gamma e^{(1/2)\kappa^2(n+1)(s-t)}V(x)\rho(x,s)dV\,ds. \tag{18}$$

In the limit $t\to\infty$ the only contribution is given by the trace term $\overline{H}$ and the integral of $V(x)$, and the resulting energy approaches the internal energy of the system in thermal equilibrium. In particular, in the high-temperature limit $\beta\to 0$ the contribution from $V(x)$ vanishes, and we recover the uniform average of the energy eigenvalues. In the low-temperature limit $\beta\to\infty$, the gradient term in the drift of (14) dominates, and the system is forced to fall to the ground state. It is interesting to observe, as a consequence of (17), that once thermal equilibrium is reached, we have the identity

$$kT^2C = \mathrm{Var}[H] + (n+1)kT(U-\overline{H}), \tag{19}$$

for the heat capacity $C=\partial U/\partial T$, where $\beta = 1/kT$ and $\mathrm{Var}[H]=(\Delta H)^2_\rho$ is the unconditional energy variance. This follows from the conditional variance formula,

$$\mathrm{Var}[H] = E[\mathrm{Var}_x[H]] + \mathrm{Var}[E_x[H]], \tag{20}$$

where $E_x[H]$ and $\mathrm{Var}_x[H]$ are, respectively, the conditional expectation and the conditional variance of the energy, when the system is in the pure state x. This relation expresses the total energy uncertainty by the sum of terms corresponding to quantum and thermal uncertainties.

As an illustration, consider the case of a spin one-half particle in a heat bath. The state space is CP^1, which we view as a 2-sphere, the north and the south poles corresponding to the upper and lower energy eigenstates, with energies $+h$ and $-h$, where h is the magnetic moment of the particle times the external field strength. The symplectic flow gives rise to latitudinal circular orbits on the sphere, while the gradient flow is in the direction along the great circles passing through the two eigenstates, pushing the state toward the south pole. The equilibrium energy latitude is obtained by balancing the gradient flow and the Brownian fluctuations. If we let θ denote the angular coordinate for the state as measured from the north pole, then the squared energy uncertainty, given that the system is in a pure state at latitude θ, is $\mathrm{Var}_\theta[H]=h^2\sin^2\theta$. The conditional energy expectation is $E_\theta[H]=h\cos\theta$. Since $\overline{H}$ vanishes, the only contribution to the energy in (18) is from the volume integral, which gives $U=\beta^{-1}-h\coth(\beta h)$, in agreement with the result of a direct calculation of $E[E_\theta[H]]$. At infinite temperature the system energy corresponds to that of the equator, whereas for a finite temperature the system energy corresponds to that of a latitude in the southern hemisphere. At zero temperature, the system collapses to the ground state.

V. ENTROPY PRODUCTION

Next we shall consider the dynamics of the total entropy of the system, given by $S_t = -\int_\Gamma \rho(x,t)\ln\rho(x,t)dV$. Without loss of generality, we can write

$$\rho(x,t) = \frac{\exp(-\beta(H(x)+\eta(x,t)))}{\int_\Gamma \exp(-\beta(H(x)+\eta(x,t)))dV}, \tag{21}$$

for a general density function, where $\eta(x,t)$ determines the departure from thermal equilibrium. Substituting this expression into the formula for S_t, and making use of the Fokker–Planck equation (11) with drift as in (14), we find

$$\frac{\partial S_t}{\partial t} = \frac{1}{2}\,\kappa^2\beta^2 \int_\Gamma (\nabla^a\eta\,\nabla_a H + \nabla^a\eta\,\nabla_a\eta)\rho\,dV. \tag{22}$$

On the other hand, for the total energy U_t we have

$$\frac{\partial U_t}{\partial t} = \frac{1}{2}\,\kappa^2\beta \int_\Gamma (\nabla^a\eta\,\nabla_a H)\rho\,dV. \tag{23}$$

Therefore, we obtain

$$\frac{\partial S_t}{\partial t} = \beta\,\frac{\partial U_t}{\partial t} + \frac{1}{2}\,\kappa^2\beta^2 \int_\Gamma (\nabla^a\eta\,\nabla_a\eta)\rho\,dV. \tag{24}$$

This is the general formula for the entropy production associated with the process (14), which shows that

$$\frac{\partial S_t}{\partial t} \geqslant \beta\,\frac{\partial U_t}{\partial t}, \tag{25}$$

in all circumstances. In particular, if U_t increases as a consequence of heating, then the entropy of the system also necessarily increases.

VI. EVOLUTION OF THE DENSITY MATRIX

Finally, we derive the dynamics of the density matrix $\rho^\alpha_\beta(t)$ associated with the process (14). Let us revert to the use of homogeneous coordinates for the state space CP^n, and write $\Pi^\alpha_\beta(x) = \psi^\alpha(x)\overline{\psi}_\beta(x)/\psi^\gamma(x)\overline{\psi}_\gamma(x)$ for the projection operator onto the pure state $x\in\Gamma$ represented by the Hilbert space vector $\psi^\alpha(x)$. From Eqs. (7) and (10) it follows that for a general (possibly nonlinear) observable $\phi(x)$ on Γ we have

$$\frac{\partial}{\partial t}\,E[\phi(x_t)] = \int_\Gamma \left(2\Omega^{ab}\nabla_a\phi\,\nabla_b H + \frac{1}{2}\,\kappa^2\nabla^2\phi - \frac{1}{2}\,\kappa^2\beta g^{ab}\,\nabla_a\phi\,\nabla_b H \right)\rho(x,t)dV. \tag{26}$$

However, if $\phi(x)$ is given, more specifically, by the conditional expectation of an ordinary linear quantum mechanical observable ϕ^α_β, then $\phi(x) = \phi^\beta_\alpha\Pi^\alpha_\beta(x)$. Thus, for a linear observable we have $E[\phi(x_t)] = \phi^\beta_\alpha\rho^\alpha_\beta(t)$, where

$$\rho^\alpha_\beta(t) = \int_\Gamma \Pi^\alpha_\beta(x)\rho(x,t)dV \tag{27}$$

is the time-dependent density matrix associated with the state $\rho(x,t)$. To obtain the equation of motion for $\rho^\alpha_\beta(t)$ we need to evaluate the three terms in the integrand on the right of (26). These are given as follows. For the commutator term we have

$$\Omega^{ab}\,\nabla_a\phi\,\nabla_b H = \tfrac{1}{2}i(\phi^\beta_\gamma H^\gamma_\alpha - H^\beta_\gamma\phi^\gamma_\alpha); \tag{28}$$

for the term involving the Laplace–Beltrami operator we find

$$\nabla^2\phi = (\phi^\gamma_\gamma\delta^\beta_\alpha - (n+1)\phi^\beta_\alpha)\Pi^\alpha_\beta, \tag{29}$$

and for the anticommutator term we obtain

$$g^{ab}\,\nabla_a\phi\,\nabla_b H = \tfrac{1}{2}(\phi^\beta_\gamma H^\gamma_\alpha + H^\beta_\gamma\phi^\gamma_\alpha)\Pi^\alpha_\beta - (\phi^\beta_\alpha H^\delta_\gamma)\Pi^\alpha_\beta\Pi^\gamma_\delta. \tag{30}$$

Inserting these expressions into (26), and noting that the result must hold for arbitrary ϕ_α^β, we deduce that

$$\frac{\partial}{\partial t}\rho_\beta^\alpha = i[H_\gamma^\alpha\rho_\beta^\gamma - \rho_\gamma^\alpha H_\beta^\gamma] - \frac{1}{4}\kappa^2\beta(H_\gamma^\alpha\rho_\beta^\gamma + \rho_\gamma^\alpha H_\beta^\gamma) + \frac{1}{2}\kappa^2(\delta_\beta^\alpha - (n+1)\rho_\beta^\alpha + \beta H_\gamma^\delta\rho_{\beta\delta}^{\alpha\gamma}). \quad (31)$$

Here the density matrix $\rho_\beta^\alpha(t)$ is given by (27), the first moment of the projection operator $\Pi_\beta^\alpha(x)$, whereas $\rho_{\alpha\gamma}^{\beta\delta}(t)$ is the second moment of $\Pi_\beta^\alpha(x)$, given by

$$\rho_{\beta\delta}^{\alpha\gamma}(t) = \int_\Gamma \Pi_\beta^\alpha(x)\Pi_\delta^\gamma(x)\rho(x,t)dV. \quad (32)$$

Equation (31) is the dynamical law for the density operator, which takes an arbitrary initial state into an equilibrium state, with heat bath temperature *T*. The first term on the right of (31) leads to the Liouville equation of linear quantum dynamics, while the general nonequilibrium process has a richer structure. The emergence of the second moment term in (31) can also be interpreted by analogy with the renormalization group equations.

VII. DISCUSSION

The model we have described here is surprisingly tractable, and has many attractive features. Experimental support could be pursued in two stages. First, we point out that it is not difficult, in the case of the canonical Γ ensemble, to derive explicit formulas for the partition function $Z(\beta)$, the state density $\Omega(E)$, the density matrix ρ_β^α, and the second moment $\rho_{\beta\delta}^{\alpha\gamma}$. If these results turn out to give a better account of equilibrium phenomena than the conventional approach in some contexts, then the next step would be to look for effects resulting from the higher moment term in the nonequilibrium dynamical law (31).

ACKNOWLEDGMENT

D.C.B. is grateful to PPARC for financial support.

[1] L. E. Reichl, *A Modern Course in Statistical Physics* (Edward Arnold, London, 1980); R. Kubo, M. Toda, N. Hashizume, and N. Saito, *Statistical Physics I, II* (Springer-Verlag, Berlin, 1983, 1985).
[2] T. W. B. Kibble, Commun. Math. Phys. **65**, 189 (1979); G. W. Gibbons, J. Geom. Phys. **8**, 147 (1992); L. P. Hughston, in *Twistor Theory*, edited by S. Huggett (Marcel Dekker, New York, 1995); D. C. Brody and L. P. Hughston, Phys. Rev. Lett. **77**, 2581 (1996); Proc. R. Soc. London **454**, 2445 (1998); A. Ashtekar and T. A. Schilling, in *On Einstein's Path*, edited by A. Harvey (Springer-Verlag, Berlin, 1998).
[3] S. Weinberg, Phys. Rev. Lett. **62**, 485 (1989).
[4] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Process* (North-Holland, Amsterdam, 1989); M. Emery, *Stochastic Calculus on Manifolds* (Springer-Verlag, Berlin, 1989).
[5] L. P. Hughston, Proc. R. Soc. London **452**, 953 (1996).
[6] E. C. Zeeman, Nonlinearity **1**, 115 (1988).

Existence and nonexistence in Chern–Simons–Higgs theory with a constant electric charge density

Dongho Chae[a] and Jongmin Han[b]
Department of Mathematics, Seoul National University, Seoul 151-742, Korea

Oleg Yu. Imanuvilov[c]
School of Mathematics, Korea Institute for Advanced Study, Seoul 130-012, Korea

(Received 17 February 1998; accepted for publication 17 June 1998)

In this paper we are devoted to proving the existence and nonexistence of self-dual equations arising in Chern–Simons–Higgs theory with a constant electric charge density. There are three kinds of boundary conditions that admit solitonic structures. It is shown that there exist solutions in two cases of them. In the other case, we prove that there is a critical electric charge density with negative value such that above the value there exists a solution and below it we have no solution. We also study asymptotic behaviors for solutions as the electric charge density goes to zero. It is found that they converge to solutions of a topological Chern–Simons system without constant electric charge density.
[S0022-2488(99)01801-0]

I. INTRODUCTION

It is well known that the $(2+1)$ Chern–Simons–Higgs model without a Maxwell term admits vortex solutions. These Chern–Simons (CS) vortices, called anyons, are different from Nielsen–Olesen-type vortices, in that they carry electric charge as well as magnetic flux due to the CS term. A special choice for the potential $V(|\phi|)$ produces a Bogomol'nyi-type energy lower bound, which is achieved by fields satisfying a set of first-order differential equations, so-called self-dual equations.[1,2] The Maxwell–Higgs system has self-dual configurations only when the coupling constant is equal to some critical value, but in CS theory such configurations hold for all coupling constants. These equations possess topological and nontopological soliton solutions according to asymptotic behaviors of the Higgs field.[3] The mathematical existence results of topological multivortex solutions for those equations can be found in Refs. 4 and 5. Unlikely to the Maxwell–Higgs system, whether the solution is uniquely determined by the zeros of the Higgs field remains open. In the nontopological case the existence of radially symmetric solutions was established in Ref. 6, and a rigorous result of the existence of arbitrary vortex solutions is established recently in Ref. 7. On the other hand, it is still open whether the first-order equations are equivalent to the second-order equations (the Euler–Lagrange equations of the energy functional).

An interesting generalization of such self-dual models arises when the system is coupled to an external charge density or an external magnetic field. For example, the Maxwell–Higgs system with the background electric charge density is more closely related to the real superconductor rather than the system without it. In that case the self-dual equations can be reduced to those of the pure Maxwell–Higgs system.[8] In this paper we consider the self-dual Chern–Simons models coupled to an external background charge density introduced in Ref. 9. In this situation the self-dual equations are quite different from those of the system without the background electric charge density, as we shall see.

The Higgs field $\phi \in \mathbf{C}$ is coupled to a CS gauge field $A_\mu \in \mathbf{R}$. The gauge field is coupled to a

[a]Electronic mail: dhchae@math.snu.ac.kr
[b]Electronic mail: jmhan@math.snu.ac.kr
[c]Electronic mail: oleg@kias.kaist.ac.kr

constant background electric charge density ρ, so that the Lagrangian of the model can be written as

$$\mathcal{L}=\frac{\kappa}{2}\,\epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_p+|D_\mu\phi|^2-U-\rho A_0, \tag{1.1}$$

where the self-dual potential is given by

$$U=\frac{1}{4\kappa^2}\,|\phi|^2(2|\phi|^2-\sigma)^2-\frac{\rho}{2\kappa}\,(2|\phi|^2-\sigma). \tag{1.2}$$

Here $D_\mu\phi=(\partial_\mu+iA_\mu)\phi$ and $\epsilon^{\mu\nu\rho}$ is a totally skew-symmetric tensor with $\epsilon^{012}=1$. The parameters σ and ρ can be either positive or negative, but we assume that the Chern–Simons coupling constant $\kappa>0$. The system is invariant under the local gauge transformation. The static energy of the system is given by

$$E=E(\phi,A)=\int_{\mathbf{R}^2}(|D_j\phi|^2+|\phi|^2A_0^2+U)\,dx, \tag{1.3}$$

with the Gauss' constraint (the variational equation with respect to A_0):

$$\kappa F_{12}+2|\phi|^2A_0-\rho=0, \tag{1.4}$$

where $F_{12}=\partial_1A_2-\partial_2A_1$ is the magnetic field. We shall consider cases where the background charge density ρ is canceled at spatial infinity by either magnetic field F_{12} or the Higgs charge density $-2|\phi|^2A_0$. The first class of sectors are called the symmetric phase and the latter the asymmetric phase. We will pay attention to the asymmetric phase so that we have the boundary condition $F_{12}\to 0$ as $|x|\to\infty$. Using Gauss' constraint and integrating by parts, we rewrite the energy as

$$E=\int_{\mathbf{R}^2}|D_1\phi-iD_2\phi|^2+|\phi|^2\left((A_0-\frac{1}{2\kappa}\,(2|\phi|^2-\sigma))^2\right)dx+\frac{\sigma}{2}\int_{\mathbf{R}^2}F_{12}\,dx. \tag{1.5}$$

We consider the excited states of finite total magnetic flux $\Psi=\int_{\mathbf{R}^2}F_{12}\,dx$, and for those configurations we find the following self-dual bound:

$$E\geqslant\frac{\sigma}{2}\,\Psi.$$

This self-dual bound is saturated by the following self-dual equations:

$$D_1\phi-iD_2\phi=0, \tag{1.6}$$

$$\kappa F_{12}+\frac{1}{\kappa}\,|\phi|^2(2|\phi|^2-\sigma)-\rho=0. \tag{1.7}$$

Let $\phi=fe^{i\theta}/\sqrt{2}$, $f\geqslant 0$. A self-dual vortex-like configuration always consists of antivortices only,

$$\theta=-\sum_{k=1}^{N}\ \arg(x-p_k)+\eta,$$

where p_k's are the positions of the antivortices, not necessarily distinct, and η is a single-valued smooth function. We may combine the coupled first-order equations (1.6), (1.7) to obtain a single second-order elliptic equation for f;

$$-\Delta\ \ln f^2+\sum_k\ 4\pi\delta(x-p_k)+\frac{1}{\kappa^2}\, f^2(f^2-\sigma)-\frac{2\rho}{\kappa}=0. \tag{1.8}$$

There are three cases of boundary conditions that admit solitonic structure. These are obtained from the consideration of $F_{12}\to 0$;

$$(\text{BC}\ \ 1)\quad \sigma>0,\quad \rho>0,\quad f^2\to(f_*)^2=\frac{\sigma+\sqrt{\sigma^2+8\kappa\rho}}{2},$$

$$(\text{BC}\ \ 2)\quad \sigma>0,\quad -\sigma^2/8\kappa<\rho<0,\quad f^2\to(f_*)^2=\frac{gj+\sqrt{\sigma^2+8\kappa\rho}}{2},$$

$$(\text{BC}\ \ 3)\quad \sigma<0,\quad \rho>0,\quad f^2\to(f_*)^2=\frac{\sigma+\sqrt{\sigma^2+8\kappa\rho}}{2}.$$

If the boundary condition is given by (BC1) or (BC2), then the substitution

$$\ln\left(\frac{f^2}{\sigma}\right)=u,\quad \text{i.e.,}\ \frac{f^2}{\sigma}=e^u, \tag{1.9}$$

transforms (1.8) into the following form:

$$\begin{gathered}\Delta u=\frac{\sigma^2}{\kappa^2}\, e^u(e^u-1)-\frac{2\rho}{\kappa}+\sum_\kappa\ 4\pi\delta(x-p_k),\quad \text{in}\ \mathbf{R}^2,\\ u\to u_*=\ln\left(\frac{f_*^2}{\sigma}\right),\quad \text{as}\ |x|\to\infty,\end{gathered} \tag{1.10}$$

and u_* satisfies

$$\frac{\sigma^2}{\kappa^2}\, e^{u_*}(e^{u_*}-1)-\frac{2\rho}{\kappa}=0. \tag{1.11}$$

Let

$$v=u-u_*.$$

Then by (1.10) and (1.11) we have

$$\begin{gathered}\Delta v=\frac{\sigma^2}{\kappa^2}\, e^{2u_*}(e^v-1)(e^v+1-e^{-u_*})+\sum_k\ 4\pi\delta(x-p_k),\quad \text{in}\ \mathbf{R}^2,\\ v\to 0,\quad \text{as}\ |x|\to\infty.\end{gathered} \tag{1.12}$$

When the boundary condition is given by (BC3), a similar substitution is possible, and it will be induced in Sec. IV.

We state the main theorem.

Theorem 1.1: *Let $p_1,p_2,\dots,p_N$ be arbitrary points in $\mathbf{R}^2$ not necessarily distinct. Then the equations (1.6) and (1.7) with boundary conditions (BC1) or (BC3) have a solution (ϕ,A) so that the zeros of ϕ are exactly $p_1,p_2,\dots,p_N$ and the corresponding energy $E(\phi,A)$ is finite. Moreover, the energy E and the magnetic flux Ψ are quantized as*

$$E=2\sigma\pi N\quad \text{and}\quad \Psi=4\pi N. \tag{1.13}$$

If the boundary condition is given by (BC2), there is a critical electric charge density ρ_0 *depending on N so that (1.6) and (1.7) have a finite energy solution* (ϕ,A) *for all* $\rho_0<\rho<0$ *with zeros of* ϕ *being* $p_1,p_2,\ldots,p_N$ *and there holds (1.13), while we have no solution for all* ρ *satisfying* $-\sigma^2/8\kappa<\rho<\rho_0$.

When the boundary condition is given by (BC1) or (BC3), the existence of solutions are to be shown in Sec. II and IV by virtue of the super- and subsolution method. The existence and nonexistence for the case (BC2) are to be shown in Sec. III. In Sec. V we prove that solutions of (1.10) converge to solutions of topological CS vortex equations as ρ goes to 0.

For notations we denote the $L^2(\mathbf{R}^2)$-norm by $\|\cdot\|_2$. We use the notation $H^k(\mathbf{R}^2)=W^{k,2}(\mathbf{R}^2)$ for $k\in\mathbf{Z}$ as Sobolev spaces.

II. Existence in the case (BC 1)

In this section we proceed with the boundary condition (BC 1). If we define

$$u_0=-\sum_k \ln(1+|x-p_k|^{-2}),\quad \mu>0,$$

the substitution $w=v-u_0$ gives the following form:

$$\Delta w=\frac{\sigma^2}{\kappa^2}e^{2u_*}(e^{u_0+w}-1)(e^{0+w}+1-e^{-u_*})+g,\quad \text{in } \mathbf{R}^2,$$
$$w\to 0,\quad \text{as } |x|\to\infty, \tag{2.1}$$

where $u_*>0$ and

$$g=4\sum_k (1+|x-p_k|^2)^{-2}.$$

We will construct a solution to (2.1) by an iteration scheme. First of all, we recall some results in the Abelian Higgs model.

Theorem 2.1: *Consider the following equation:*

$$\Delta u=a(e^u-1)+\sum_{k=1}^N 4\pi\delta(x-p_k),\quad in\ \mathbf{R}^2,$$
$$u\to 0,\quad as\ |x|\to\infty, \tag{2.2}$$

where a is a positive constant. Let $u=u_0+w$ *and rewrite (2.2) as*

$$\Delta w=a(e^{u_0+w}-1)+g,\quad in\ \mathbf{R}^2,$$
$$w\to 0,\quad as\ |x|\to\infty. \tag{2.3}$$

Then (2.3) admits a unique solution in $C^\infty(\mathbf{R}^2)\cap H^k(\mathbf{R}^2)$ *for all* $k\geq 2$ *satisfying* $u_0+w<0$. *Moreover, for given* $0<\delta<1$ *there exists a constant* $M=M(a,\delta)$ *and* $m=m(a,\delta)$ *so that*

$$0<1-e^{u_0+w}\leq M\exp(-(1-\delta)m|x|). \tag{2.4}$$

Proof: See Ref. 10, Chap. 3.

Theorem 2.2: *There is a solution* $\bar{w}$ *to (2.1) satisfying* $\bar{w}\in C^\infty(\mathbf{R}^2)\cap H^k(\mathbf{R}^2)$ *for all* $k\geq 2$.

Proof: Let us consider the unique solutions U and W satisfying

$$\Delta U=\frac{\sigma^2}{\kappa^2}\, e^{2u_*}(1-e^{-u_*})(e^{u_0+U}-1)+g, \quad \text{in } \mathbf{R}^2,$$
$$U\to 0, \quad \text{as } |x|\to\infty, \tag{2.5}$$

and

$$\Delta W=\frac{\sigma^2}{\kappa^2}\, e^{2u_*}(2-e^{-u_*})(e^{u_0+W}-1)+g, \quad \text{in } \mathbf{R}^2,$$
$$W\to 0, \quad \text{as } |x|\to\infty. \tag{2.6}$$

It is easily checked that U and W are sub- and supersolutions to (2.1), respectively, and satisfy the inequality $U\leqslant W$.

Choosing a constant $K>2e^{2u_*}\sigma^2/\kappa^2$, we define the following iterative sequence:

$$\Delta w_{n+1}-Kw_{n+1}=\frac{\sigma^2}{\kappa^2}\, e^{2u_*}(e^{u_0+w_n}-1)(e^{u_0+w_n}+1-e^{-u_*})+g-Kw_n^-,$$
$$w_{n+1}\to 0, \quad \text{as } |x|\to\infty, \tag{2.7}$$
$$w_0=W.$$

For $n=1$,

$$\Delta w_1-Kw_1=\frac{\sigma^2}{\kappa^2}\, e^{2u_*}(e^{u_0+w_0}-1)(e^{u_0+w_0}+1-e^{-u_*})+g-Kw_0\in L^2(\mathbf{R}^2). \tag{2.8}$$

Here we used the fact that $u_0+w_0\in L^2$ and $|e^{u_0+w_0}-1|\leqslant|u_0+w_0|$, which follows from the fact that $u_0+w_0\leqslant 0$. Since $\Delta-K:H^2(\mathbf{R}^2)\to L^2(\mathbf{R}^2)$ is bijective, there exists $w_1\in H^2(\mathbf{R}^2)$ satisfying (2.8) so that w_1 goes to 0 as $|x|\to\infty$. Inductively $w_n\in H^2(\mathbf{R}^2)$ is well defined and vanishes at infinity.

Using standard iteration methods and a maximum principle, we find out that there exists a unique function $\overline{w}\in L^2(\mathbf{R}^2)$ satisfying

$$U\leqslant\overline{w}\leqslant\cdots\leqslant w_n\leqslant\cdots\leqslant w_1\leqslant w_0=W,$$

so that $w_n\to\overline{w}$ in $H^2(\mathbf{R}^2)$ and $\overline{w}\to 0$ as $|x|\to\infty$. Consequently, $\overline{w}$ is a solution to (2.1), and standard elliptic arguments show that $\overline{w}\in C^\infty(\mathbf{R}^2)\cap H^k(\mathbf{R}^2)$ for all $k\geqslant 2$. □

We can also construct an iterative sequence from below using the subsolution (2.5). Let us consider the following equation:

$$\Delta\tilde{w}_{n+1}-K\tilde{w}_{n+1}=\frac{\sigma^2}{\kappa^2}\, e^{2u_*}(e^{\tilde{w}_n+u_0}-1)(e^{\tilde{w}_n+u_0}+1-e^{-u_*})+g-K\tilde{w}_n,$$
$$\tilde{w}_{n+1}\to 0, \quad \text{as } |x|\to\infty, \tag{2.9}$$
$$\tilde{w}_0=U.$$

The constant K satisfies the inequality $K>2e^{2u_*}\sigma^2/\kappa^2$. Then as above we can find a solution to (2.1) $\underline{w}\in C^\infty(\mathbf{R}^2)\cap H^k(\mathbf{R}^2)$ for all $k\geqslant 2$, so that

$$U=\tilde{w}_0\leqslant\tilde{w}_1\leqslant\cdots\leqslant\underline{w}\leqslant W.$$

Proposition 2.3: Any solutions of (2.1) obtained by the iteration methods (2.7) are maximal in the $C^2(\mathbf{R}^2)$-solution class. In particular, the solutions $\overline{w}$ via iteration (2.7) are independent of K. Similarly $\underline{w}$ is a minimal solution and independent of K.

Proof: Let w be any given solution in $C^2(\mathbf{R}^2)$. It suffices to show that $w \leqslant w_n$ for all n, where w_n is given by (2.7). Let $u = u_* + u_0 + w$. We observe that $u \leqslant u_*$. Otherwise, $u(x') > u_*$ at the maximum point x' of u and $\Delta u(x') \leqslant 0$. But then

$$\Delta u(x') > \frac{\sigma^2}{\kappa^2} e^{u_*}(e^{u_*}-1) - \frac{2\rho}{\kappa} + \sum_k 4\pi\delta(x'-p_k) \geqslant 0,$$

a contradiction. Since $u_0 + w = u - u_* \leqslant 0$,

$$\Delta w \geqslant \frac{\sigma^2}{\kappa^2} e^{2u_*}(2-e^{-u_*})(e^{u_0+w}-1)+g,$$

and thus

$$\Delta(w-w_0) \geqslant \frac{\sigma^2}{\kappa^2} e^{2u_*}(2-e^{-u_*})(e^{u_0+w}-e^{u_0+w_0}) = \frac{\sigma^2}{\kappa^2} e^{2u_*}(2-e^{-u_*})e^{u_0+w'}(w-w_0),$$

for some w' between w and w_0. Hence $w \leqslant w_0$. Now suppose that $w \leqslant w_k$ for all $0 \leqslant k \leqslant n$. Then

$$(\Delta - K)(w-w_{n+1}) = \frac{\sigma^2}{\kappa^2} e^{2u_*+2u_0}(e^{2w}-e^{2w_n}) - \frac{\sigma^2}{\kappa^2} e^{u_*+u_0}(e^{w}-e^{w_n}) - K(w-w_n)$$

$$\geqslant K(e^{2u_0+2w'}-1)(w-w_n) \geqslant 0.$$

So $w \leqslant w_{n+1}$. □

Theorem 2.4: *Let w be any solution to (2.1). Then, given $0<\delta<1$, there are positive constants $M = M(w,\delta)$ and $m = m(w,\delta)$, so that*

$$0 < 1 - e^{u_0+w} \leqslant M \exp(-(1-\delta)m|x|). \tag{2.10}$$

Proof: By Theorem 2.1, there exist constants $M_1 = M_1(\overline{w},\delta)$, $M_2 = M_2(\underline{w},\delta)$, $m_1 = m_1(\overline{w},\delta)$, and $m_2 = m_2(\underline{w},\delta)$, so that

$$1 - e^{u_0+\overline{w}} \leqslant M_1 \exp(-(1-\delta)m_1|x|), \tag{2.11}$$

$$1 - e^{u_0+\underline{w}} \leqslant M_2 \exp(-(1-\delta)m_2|x|). \tag{2.12}$$

Since $\underline{w} \leqslant w \leqslant \overline{w}$, estimates (2.11) and (2.12) give (2.10). □

Now given any solution w to (2.1), we can construct a solution pair (ϕ,A) to (1.6) and (1.7) by standard arguments.[10] In fact, we have

$$\phi = \frac{1}{\sqrt{2}} f e^{i\theta} = \frac{\sqrt{\sigma}}{\sqrt{2}} \exp\left((u_0+w+u_*)/2 - i\sum_{k=1}^{N} \arg(x-p_k) \right), \tag{2.13}$$

$$A_1 = -2\,\mathrm{Re}(i\overline{\partial} \ln \overline{\phi}), \quad A_2 = -2\,\mathrm{Im}(i\overline{\partial} \ln \overline{\phi}), \tag{2.14}$$

where $\overline{\partial} = (\partial_1 + i\partial_2)/2$. We show that these solution pairs (ϕ,A) of the self-dual equations (1.6) and (1.7) are indeed of finite energy.

Theorem 2.5: *Let (ϕ,A) be any solution pairs of (1.6) and (1.7) with the boundary condition (BC1). Then (ϕ,A) is of finite energy, i.e., $E(\phi,A)<\infty$. Furthermore, the energy E and the magnetic flux Ψ are quantized as*

$$E=2\sigma\pi N \quad and \quad \Psi=4\pi N. \tag{2.15}$$

Proof: Since $u \leqslant u_*$ (Proposition 2.3), it follows that $f^2 \leqslant f_*^2$. For given $0<\delta<1$, Theorem 2.4 states that

$$0 \leqslant f_*^2 - f^2 \leqslant f_*^2 M \exp(-(1-\delta)m|x|), \tag{2.16}$$

for some positive constants M and m. From (1.4) and (1.7) it turns out that

$$f^2 A_0^2 = \frac{1}{4\kappa^2} f^2(f^2-\sigma)^2,$$

and hence

$$\frac{1}{2} f^2 A_0^2 + U = \frac{1}{4\kappa^2} f^2(f^2-\sigma)^2 - \frac{\rho}{2\kappa}(f^2-\sigma) = \frac{1}{4\kappa^2}(f_*^2-f^2)(f^2-\sigma)(f_*^2+f^2-\sigma).$$

This equation, combined with (2.16) and the fact that $f^2 \leqslant f_*^2$, yields

$$|\phi|^2 A_0^2 + U = \tfrac{1}{2} f^2 A_0^2 + U \in L^1(\mathbf{R}^2).$$

Next, using (1.6), we find that

$$|D_j\phi|^2 = \tfrac{1}{2}(\partial_j f\ )^2 + \tfrac{1}{2} f^2(\partial_j\theta + A_1)^2 = \tfrac{1}{2}|\nabla f|^2.$$

Since $f=\sqrt{\sigma}e^{u/2}=\sqrt{\sigma}e^{u_*/2}e^{(u_0+w)/2}$, we have

$$\partial_j f = \tfrac{1}{2}\sqrt{\sigma}e^{u_*/2}(\partial_j u_0 e^{u_0/2}\cdot e^{w/2} + \partial_j w \cdot e^{(u_0+w)/2}) \in L^2(\mathbf{R}^2).$$

Therefore $|D_1\phi|^2 + |D_2\phi|^2 = |\nabla f|^2 \in L^1(\mathbf{R}^2)$.

For the proof of (2.15) we return to the equation (2.1) and rewrite (2.1) as the following form:

$$\Delta w = F(e^{u_0+w}) + g.$$

Letting $v=u_0+w$, we have

$$\Delta \partial_j v = F'(e^v)\partial_j v,$$

for all $|x|>\sup|p_k|$. Since $F'(1)>0$, there exist constants $R>\sup|p_k|$ and $m>0$, such that $F'(e^v)\geqslant m$ for all $|x|\geqslant R$. Using a suitable comparison function, we can show that $\partial_j v$ is exponentially decreasing. Consequently, integrating (2.1), we get

$$\int_{\mathbf{R}^2} F(e^{u_0+w})dx = -4\pi N. \tag{2.17}$$

Now the equations (1.7) and (2.17) give (2.15). □

III. Existence and nonexistence in the case (BC2)

A similar background substitution of (1.12) as in Sec. II gives the final form of the equation with the boundary condition (BC2);

$$\Delta w=\frac{\sigma^2}{\kappa^2}e^{2u_*}(e^{u_0+w}-1)(e^{u_0+w}+1-e^{-u_*})+g,\quad \text{in } \mathbf{R}^2,$$
$$w\to 0,\quad \text{as } |x|\to\infty, \tag{3.1}$$

where

$$-\ln 2<u_*<0,$$

$$u_0=-\sum_k \ln(1+|x-p_k|^{-2}),$$

$$g=4\sum_k (1+|x-p_k|^2)^{-2}.$$

We rewrite (3.1) as

$$\Delta w=\lambda(e^{u_0+w}-1)^2+\lambda(1-\epsilon)(e^{u_0+w}-1)+g. \tag{3.2}$$

Here

$$\lambda=\lambda(\epsilon)=\frac{\sigma^2}{\kappa^2}e^{2u_*},\quad \epsilon=e^{-u_*}-1.$$

We note that $0<\epsilon<1$ and $0\leqslant\lambda<\lambda_0=\sigma^2/\kappa^2$.

Theorem 3.1: *For all sufficiently small $\epsilon>0$, (3.2) has a maximal solution $w\in C^\infty(\mathbf{R}^2)\cap H^k(\mathbf{R}^2)$ for all $k\geqslant 2$. There exist positive constants $R=R(\epsilon)$, $C=C(\epsilon)$, and $m=m(\epsilon)$ such that*

$$0\leqslant 1-e^{u_0+w}\leqslant Ce^{-m|x|},$$

for all $|x|\geqslant R$. Furthermore, the corresponding solution pair (ϕ,A) given by (2.13) and (2.14) is of finite energy and there holds (1.13).

Proof: Without loss of generality we may assume $0<\epsilon<1/2$ so that

$$\frac{4}{9}\lambda_0<\lambda<\lambda_0 \text{ and } \frac{\lambda}{2}(1-\epsilon)>\frac{1}{9}\lambda_0.$$

Let $R_0=2\sup_k|p_k|$ and choose a smooth function $\tilde{g}:\mathbf{R}^2\to\mathbf{R}$ so that $0\leqslant\tilde{g}\leqslant\lambda_0$, $\operatorname{supp}\tilde{g}\subset B(0,2R_0)$, and $\tilde{g}=\lambda_0$ on $B(0,R_0)$. Let us consider the following equation:

$$\Delta g=\frac{\lambda_0}{9}e^{u_0+q}(e^{u_0+q}-1)+g+\tilde{g},\quad \text{in } \mathbf{R}^2,$$

$$q\to 0,\quad \text{as } |x|\to\infty.$$

This is a slight variation of the topological CS equation and its existence result is established in Refs. 4 and 5.

For $|x|\leqslant R_0$ we have

$$\Delta q\geqslant\lambda(1-\epsilon)e^{u_0+q}(e^{u_0+q}-1)+g+\lambda_0\geqslant\lambda(e^{u_0+q}-1)^2+\lambda(1-\epsilon)(e^{u_0+q}-1)+g.$$

Since $u_0+q\to 0$ as $|x|\to\infty$, there exists $\epsilon_0=\epsilon_0(N)>0$ depending only on the number of vortices so that $e^{u_0+q}\geqslant 2\epsilon_0$ for $|x|\geqslant R_0$. Hence, if $0<\epsilon<\epsilon_0$, then

$$e^{u_0+q} \geqslant \frac{2\epsilon}{1+\epsilon},$$

which implies that for $|x| \geqslant R_0$,

$$\Delta q \geqslant \lambda(1-\epsilon)e^{u_0+q}(e^{u_0+q}-1)+g-\frac{\lambda}{2}(1-\epsilon)e^{u_0+q}(e^{u_0+q}-1)$$
$$\geqslant \lambda(e^{u_0+q}-1)^2+\lambda(1-\epsilon)(e^{u_0+q}-1)+g.$$

Consequently, q is a subsolution to $(3.2)_\epsilon$ for all $\epsilon \in (0,\epsilon_0)$. Since $-u_0$ is a supersolution, we can use monotone iteration techniques to obtain a maximal solution to (3.2).

It remains to show that the corresponding solution pair (ϕ,A) is of finite energy. Let $v=u_0+w \leqslant 0$. Then v goes to 0 as $|x| \to 0$, and since $0<\epsilon<1$, there is a constant $R>\sup_k|p_k|$ such that $-(1-\epsilon)/2 \leqslant v \leqslant 0$ and $e^v \geqslant 1/2$, if $|x| \geqslant R$.

For $|x| \geqslant R$, (3.2) can be rewritten as

$$\Delta v = \lambda(e^v-1)^2+\lambda(1-\epsilon)(e^v-1)$$
$$= \lambda e^{v'} v[e^{v'}v+(1-\epsilon)] \quad (v \leqslant v' \leqslant 0)$$
$$\leqslant \frac{\lambda(1-\epsilon)}{4} v \equiv m^2 v. \tag{3.3}$$

Comparing (3.3) with the function $Ce^{-m|x|}$, by a maximum principle we find that

$$-Ce^{-m|x|} \leqslant v \leqslant 0,$$

for $|x| \geqslant R$. Here $C=C(\epsilon)=-e^{mR}\inf\{v(x):|x|=R\}$. Hence

$$1-e^v \leqslant |v| \leqslant Ce^{-m|x|}.$$

As a consequence, we can prove as in Theorem 2.5 that $E(\phi,A)<\infty$ and the quantization (1.13) holds. □

Corollary 3.2: Let $0<\tilde{\epsilon}<1$ *be given so that there exist a solution* $\tilde{w}$ *to* $(3.2)_{\tilde{\epsilon}}$. *Then* $(3.2)_\epsilon$ *has a solution for each* $0<\epsilon<\tilde{\epsilon}$. *Consequently, there exists a critical number* $\epsilon_0=\epsilon_0(N) \in (0,1]$ *depending only on the number of vortices such that* $(3.2)_\epsilon$ *has a solution for all* $\epsilon \in (0,\epsilon_0)$ *and no solution for all* $\epsilon \in (\epsilon_0,1)$.

Proof: We see that $\tilde{w}$ is a subsolution to $(3.2)_\epsilon$. Thus we can construct a solution to $(3.2)_\epsilon$ by iteration arguments. □

Our next aim is to show that $\epsilon_0(N)$ is less than 1. Indeed, (3.2) has no solution when $\epsilon=1$. Since g is strictly positive, we may expect that (3.2) has no solution for all ϵ sufficiently close to 1. We first prove that $\epsilon_0(1)<1$, showing the nonexistence of a radially symmetric solution of an auxiliary equation by a shooting argument.

Theorem 3.3: *Suppose that* $N=1$ *and the vortex point is the origin. Then* $\epsilon_0(1)<1$.

Proof: We assume that $\epsilon \geqslant \frac{1}{2}$ and the vortex point is the origin. To clarify the dependence of the equation on ϵ, we rewrite (3.2) as

$$\Delta w_\epsilon = \lambda_\epsilon(e^{u_1+w_\epsilon}-1)^2+\lambda_\epsilon(1-\epsilon)(e^{u_1+w_e}-1)+g_1,$$
$$w_\epsilon \to 0, \quad \text{as } |x| \to \infty, \tag{3.4}$$

where

$$u_1=-\ln(1+|x|^{-2}), \quad g_1=4/(1+|x|^2)^2, \quad \lambda_\epsilon=\frac{\sigma^2}{\kappa^2}\frac{1}{(1+\epsilon)^2}.$$

Let us consider the function

$$f_\epsilon(a)=\lambda_\epsilon(e^a-1)^2+\lambda_\epsilon(1-\epsilon)(e^a-1).$$

We construct a function

$$h_\epsilon(a)=\begin{cases} m_\epsilon, & -\infty<a\leqslant -1,\\ f_\epsilon(a), & -1<a<\infty\end{cases}$$

where $m_\epsilon=f_\epsilon(-1)\geqslant m_0$ for some $m_0>0$ independent of ϵ. To prove the theorem we need the following lemma.

Lemma 3.4: There exists $\epsilon_1\in(0,1)$ *such that there is no radially symmetric solution of*

$$\Delta b_\epsilon=h_\epsilon(b_\epsilon)+4\pi\delta(x),\quad b_\epsilon\to 0,\quad as\quad |x|\to\infty, \tag{3.5}$$

for all $\epsilon\in(\epsilon_1,1)$.

Proof of lemma: The proof is given by contradiction. Let us assume that there is a solution to (3.5). We drop the subscript ϵ for simplicity. It is easily seen by a maximum principle that $b(x)<0$ for all $x\neq 0$. By definition of h, there exists $r_0=r_0(\epsilon)>0$ such that

$$\Delta b=m+4\pi\delta(x),\quad \text{for } |x|\leqslant r_0. \tag{3.6}$$

But the solution of (3.6) is given by the following form:

$$b(x)=2\ln|x|+\tfrac{1}{4}m|x|^2+H(x), \tag{3.7}$$

where $H(x)$ is harmonic in $B(0,r_0)$. Furthermore, by definition of h we also have the boundary condition

$$b(r_0)=-1,$$

and the harmonic function in (3.7) becomes $H(x)=\text{const}$. Since

$$\frac{\partial b}{\partial r}(r_0)=\frac{2}{r_0}+\frac{1}{2}mr_0,$$

we have

$$\frac{\partial b}{\partial r}(r_0)\geqslant 2\sqrt{m}\geqslant 2\sqrt{m_0}. \tag{3.8}$$

Since $f(a)>0$ for $-\infty<a<\ln\epsilon$, there is $r_1=r_1(\epsilon)>r_0$ satisfying $b(r_1)=\ln\epsilon$ such that

$$\Delta b=f(b)\geqslant 0, \tag{3.9}$$

for $r_0\leqslant|x|\leqslant r_1$. We wish to estimate $(\partial b/\partial r)(r_1)$. Let us define a function,

$$\tilde{b}(x)=r_0\frac{\partial b}{\partial r}(r_0)\ln\frac{|x|}{r_0}-1,$$

for $|x|\geqslant r_0$. We observe that

$$\tilde{b}(r_0)=b(r_0)=-1, \tag{3.10}$$

$$\frac{\partial\tilde{b}}{\partial r}(r_0)=\frac{\partial b}{\partial r}(r_0). \tag{3.11}$$

Since $\Delta\tilde{b}=0$, using (3.9) and (3.11), we have

$$\frac{\partial b}{\partial r}(r)\geqslant\frac{\partial\tilde{b}}{\partial r}(r), \tag{3.12}$$

for $r_0\leqslant r\leqslant r_1$. Let $\hat{r}=\hat{r}(\epsilon)$ be given by $\tilde{b}(\hat{r})=0$. Since $b(r)$ is negative, it follows from (3.10) and (3.12) that $\hat{r}\geqslant r_1$. Noting that $\partial\tilde{b}/\partial r$ is monotonically decreasing, we find that

$$\frac{\partial b}{\partial r}(r_1)\geqslant\frac{\partial\tilde{b}}{\partial r}(r_1)\geqslant\frac{\partial\tilde{b}}{\partial r}(\hat{r}). \tag{3.13}$$

Since

$$\ln\frac{\hat{r}}{r_0}=\frac{2}{4+mr_0^2}\leqslant\frac{1}{2},$$

we have

$$\frac{1}{\sqrt{e}}\leqslant\frac{r_0}{\hat{r}}.$$

Consequently, (3.8) and (3.13) yield

$$\frac{\partial b}{\partial r}(r_1)\geqslant\frac{2\sqrt{m_0}}{\sqrt{e}}. \tag{3.14}$$

Now for $|x|\geqslant r_1$, we rewrite (3.5) in polar coordinates as

$$\frac{\partial}{\partial r}\left(r\frac{\partial b}{\partial r}\right)=rh(b).$$

Integrating both sides and using (3.14), we have

$$\frac{\partial b}{\partial r}(r)=\frac{r_1}{r}\frac{\partial b}{\partial r}(r_1)+\frac{1}{r}\int_{r_1}^{r}sh(b(s))ds\geqslant\frac{r_1}{r}\frac{2\sqrt{m_0}}{\sqrt{e}}-(r-r_1)\sup_{s\in[r_1,r]}|h(b(s))|.$$

Since $\sup_{x\leqslant 0}|h(x)|$ is uniformly bounded with respect to ϵ, we can find a number $\alpha>0$ and a constant $C_0>0$ independent of ϵ, satisfying

$$\frac{\partial b}{\partial r}(r)\geqslant C_0,$$

for all $r\in[r_1,r_1+\alpha]$. Since $b(r_1)=\ln\epsilon\to 0$ as $\epsilon\to 1$, we conclude that $b(r_1+\alpha)>0$ for all ϵ sufficiently close to 1, which contradicts the fact that $b(r)$ is negative. □

Let us assume that there exists a solution w_ϵ of (3.4) for $\epsilon\in(\epsilon_1,1)$. If we make substitutions $y_\epsilon=b_\epsilon-u_1$ in (3.5), we get

$$\Delta y_\epsilon=h_\epsilon(u_1+y_\epsilon)+g_1,\quad y_\epsilon\to 0,\quad \text{as } |x|\to\infty. \tag{3.15}$$

By Lemma 3.4 the equation (3.15) has no radially symmetric solution for $\epsilon\in(\epsilon_1,1)$. Let us drop ϵ for brevity. We shall construct a radially symmetric solution of (3.15), using w_ϵ, to conclude a contradiction. Let us define an iterative sequence,

$$\Delta y_{n+1} - K y_{n+1} = h(u_1 + y_n) + g_1 - K y_n,$$
$$y_{n+1} \to 0, \quad \text{as } |x| \to \infty, \tag{3.16}$$

where

$$y_0 = -u_1, \quad K = \|h'\|_\infty + 1.$$

Since the right-hand side of (3.16) is radially symmetric, the unique solution of (3.16) is also radially symmetric. It is easily checked that $y_1 \leqslant y_0$. Moreover, inductively we have

$$(\Delta - K)(y_{n+1} - y_n) = h(u_1 + y_n) - h(u_1 + y_{n-1}) - K(y_n - y_{n-1})$$
$$= (h'(u_1 + \xi) - K)(y_n - y_{n-1}) \geqslant 0,$$

where $y_n \leqslant \xi \leqslant y_{n-1}$. Thus, by the maximum principle $y_{n+1} \leqslant y_n$. Now let us show that w is a subsolution of (3.15). It is easily seen that $w \leqslant y_0$. Inductively,

$$(\Delta - K)(w - y_{n+1}) = f(u_1 + w) - h(u_1 + y_n) - K(w - y_n)$$
$$\geqslant h(u_1 + w) - h(u_1 + y_n) - K(w - y_n)$$
$$= (h'(\xi + u_1) - K)(w - y_n) \geqslant 0,$$

where $w \leqslant \xi \leqslant y_n$. Therefore $y_{n+1} \geqslant w$. Consequently, we get a sequence of radially symmetric functions satisfying

$$y_0 \geqslant y_1 \geqslant \cdots \geqslant w.$$

Then it is easy to show that y_n converges to a radially symmetric solution of (3.15), a contradiction to Lemma 3.4.

We recall from Corollary 3.2 that $\epsilon_0(1)$ is the maximal number satisfying that there exists a solution to (3.4) for all $\epsilon \in (0, \epsilon_0(1))$. Since there is no solution to (3.4) for $\epsilon \in (\epsilon_1, 1)$, we conclude that $\epsilon_0(1) \leqslant \epsilon_1 < 1$. □

Corollary 3.5: Suppose that the number of vortices is $N > 1$. Then $\epsilon_0(N) < 1$.

Proof: By Theorem 3.3 it suffices to show that $\epsilon_0(N) \leqslant \epsilon_0(1)$. Let $\epsilon \in (0, \epsilon_0(N))$ and w be a solution of (3.2). Then returning to (1.12), we have a solution v and

$$\Delta v \geqslant \frac{\sigma^2}{\kappa^2} e^{2u_*}(e^v - 1)(e^v + 1 - e^{-u_*}) + 4\pi\delta(x - p_1).$$

We may assume that $p_1 = 0$. Letting $\tilde{w} = v - u_1$, we see that $\tilde{w}$ is a subsolution of (3.4). Since $-u_1$ is a supersolution to (3.4), we can construct a solution to (3.4) by standard iterative schemes we have developed. Hence $\epsilon_0(N) \leqslant \epsilon_0(1) < 1$. □

IV. EXISTENCE IN THE CASE (BC3)

In the (BC3) case, if we transform σ into $-\sigma$ so that $\sigma > 0$, then the background substitution as in Sec. I gives

$$\Delta w = \frac{\sigma^2}{\kappa^2} e^{2u_*}(e^{u_0 + w} - 1)(e^{u_0 + w} + 1 + e^{-u_*}) + g, \quad \text{in } \mathbf{R}^2,$$
$$w \to 0, \quad \text{as } |x| \to \infty, \tag{4.1}$$
$$-\infty < u_* < \infty,$$

where

$$u_0=-\sum_k \ln(1+|x-p_k|^{-2}), \quad \mu>0,$$

$$g=4\sum_k (1+|x-p_k|^2)^{-2},$$

and u_* satisfies that

$$\frac{\sigma^2}{\kappa^2} e^{u_*}(e^{u_*}+1)-\frac{2\rho}{\kappa}=0.$$

As in Sec. II, we consider the following two equations:

$$\Delta U=\frac{\sigma^2}{\kappa^2} e^{2u_*}(2+e^{-u_*})(e^{u_0+U}-1)+g, \tag{4.2}$$

$$U\to 0, \quad \text{as } |x|\to\infty,$$

and

$$\Delta W=\frac{\sigma^2}{\kappa^2} e^{2u_*}(1+e^{-u_*})(e^{u_0+W}-1)+g,$$
$$W\to 0, \quad \text{as } |x|\to\infty. \tag{4.3}$$

Then it is easily checked that U and W are sub- and supersolutions of (4.1) as in Sec. II. Now, all existence results via an iteration process are parallel to the case (BC1) so that we achieve the following theorem.

Theorem 4.1: *There exist maximal and minimal solutions of (4.1) for all parameters $-\infty<u_*<\infty$ satisfying (2.10). For any solution of (4.1) the corresponding pair (ϕ,A) given by (2.13) and (2.14) is a finite energy solution of (1.6) and (1.7). Moreover, there holds (1.13).*

V. ASYMPTOTICS FOR SOLUTIONS AS $\rho\to 0$

In this section we study asymptotics for solutions to (1.10) when the boundary conditions are given by (BC1) or (BC2). We first recall the existence results of topological CS vortex equations.

Theorem 5.1: *Consider the following equation:*

$$\Delta u=ae^u(e^u-1)+\sum_{k=1}^N 4\pi\delta(x-p_k), \quad in \ \mathbf{R}^2,$$
$$u\to 0, \quad as \ |x|\to\infty, \tag{5.1}$$

where a is a positive constant. Let $u=u_0+w$ and rewrite (5.1) as

$$\Delta w=ae^{u_0+w}(e^{u_0+w}-1)+g, \quad in \ \mathbf{R}^2,$$
$$w\to 0, \quad as \ |x|\to\infty. \tag{5.2}$$

Then (5.2) admits a maximal solution in $C^\infty(\mathbf{R}^2)\cap H^k(\mathbf{R}^2)$ for all $k\geqslant 2$ satisfying $u_0+w<0$. Moreover, for given $0<\delta<1$ there exist a constant $M=M(a,\delta)$ such that

$$0<1-e^{u_0+w}\leqslant M \exp(-(1-\delta)\sqrt{a}|x|). \tag{5.3}$$

Proof: See Ref. 4.

We established maximal solutions w to (2.1) and (3.1) in Secs. II and III, and hence maximal solutions u to (1.10) for all sufficiently small $|\rho|$. To clarify the dependence of solutions on ρ let us denote the maximal solutions w_ρ and rewrite (2.1) and (3.1) as

$$\Delta w_\rho = \frac{\sigma^2}{\kappa^2} e^{2u_*^\rho}(e^{u_0+w_\rho}-1)(e^{u_0+w_\rho}+1-e^{-u_*^\rho})+g, \quad \text{in } \mathbf{R}^2,$$
$$w_\rho \to 0, \quad \text{as } |x|\to\infty. \tag{5.4}$$

Let us choose $\rho_1>0$ such that (5.4) has a maximal solution for all $|\rho|\leqslant\rho_1$. We notice that $w_{\tilde\rho} \leqslant w_\rho$ if $\tilde\rho\leqslant\rho$. In fact, the proof of Corollary 3.2 can also be applied to the case that one or two of ρ and $\tilde\rho$ are positive or negative. In the sequel, there exist two functions w^+ and w^- so that, for $\tilde\rho<0<\rho$,

$$w_{\tilde\rho}\leqslant w^-\leqslant w^+\leqslant w_\rho,$$

and $w_{\tilde\rho}$ and w_ρ converge pointwise and in $L^2(\mathbf{R}^2)$ to w^+ and w^-, respectively.

Theorem 5.2: *The functions w^+ and w^- are solutions of the following topological CS vortex equation 8*

$$\Delta w = \frac{\sigma^2}{\kappa^2} e^{u_0+w}(e^{u_0+w}-1)+g, \quad \textit{in } \mathbf{R}^2,$$
$$w\to 0, \quad \textit{as } |x|\to\infty, \tag{5.5}$$

and w_ρ converges in $H^k(\mathbf{R}^2)$ for all $k\geqslant 1$ to w^+ as $\rho\to 0+$ and w^- as $\rho\to 0-$. Moreover, w^+ is the maximal solution appearing in Theorem 5.1 with $a=\sigma^2/k^2$.

Proof: First, we prove the theorem for the case $\rho>0$. Suppose that $0<\tilde\rho<\rho$. Then

$$\Delta(w_\rho - w_{\tilde\rho}) = \frac{\sigma^2}{\kappa^2} e^{2u_*^\rho}(e^{u_0+w_\rho}-1)(e^{u_0+w_\rho}+1-e^{-u_*^\rho})$$
$$-\frac{\sigma^2}{\kappa^2} e^{2u_*^{\tilde\rho}}(e^{u_0+w_{\tilde\rho}}-1)(e^{u_0+w_{\tilde\rho}}+1-e^{-u_*^{\tilde\rho}}). \tag{5.6}$$

Multiplying $-(w_\rho - w_{\tilde\rho})$ on both sides of (5.6) and integrating by parts, we get

$$\int_{\mathbf{R}^2}|\nabla w_\rho - \nabla w_{\tilde\rho}|^2\,dx \leqslant C\int_{\mathbf{R}^2}(|e^{u_0+w_\rho}-1|+|e^{u_0+w_{\tilde\rho}}-1|)|w_\rho - w_{\tilde\rho}|dx$$
$$\leqslant C(\|e^{u_0+w_\rho}-1\|_2+\|e^{u_0+w_{\tilde\rho}}-1\|_2)\|w_\rho - w_{\tilde\rho}\|_2$$
$$\leqslant C\|e^{u_0+w_{-\rho_1}}-1\|_2\|w_\rho - w_{\tilde\rho}\|_2 \to 0.$$

Here $C>0$ is a constant that may vary at different places. The last inequality follows from the inequality

$$0<1-e^{u_0+w_\rho}\leqslant 1-e^{u_0+w_{-\rho_1}}.$$

Consequently, $w_\rho\to w^+$ in $H^1(\mathbf{R}^2)$. A differentiation of (5.6) gives

$$\Delta(\partial_j w_\rho - \partial_j w_{\tilde\rho}) = \frac{\sigma^2}{\kappa^2} e^{2u_*^\rho}\partial_j(u_0+w_\rho)e^{u_0+w_\rho}(2e^{u_0+w_\rho}-e^{-u_*^\rho})$$
$$-\frac{\sigma^2}{\kappa^2} e^{2u_*^{\tilde\rho}}\partial_j(u_0+w_{\tilde\rho})e^{u_0+w_{\tilde\rho}}(2e^{u_0+w_{\tilde\rho}}-e^{-u_*^{\tilde\rho}}).$$

Multiplying $-(\partial_j w_\rho - \partial_j w_{\tilde\rho})$ on both sides and integrating by parts, we obtain

$$\int_{\mathbf{R}^2}|\nabla(\partial_j w_\rho - \partial_j w_{\tilde\rho})|^2\, dx \leqslant C\int_{\mathbf{R}^2}(2|\partial_j u_0 e^{u_0}|+|\partial_j w_\rho|+|\partial_j w_{\tilde\rho}|)|\partial_j w_\rho - \partial_j w_{\tilde\rho}|dx$$
$$\leqslant C(\|\partial_j u_0\; e^{u_0}\|_2+\|\partial_j w_\rho\|_2+\|\partial_j w_{\tilde\rho}\|_2)\|\partial_j w_\rho - \partial_j w_{\tilde\rho}\|_2 \to 0.$$

Thus, $w_\rho \to w^+$ in $H^2(\mathbf{R}^2)$. Successively, it is shown that $w_\rho \to w^+$ in $H^k(\mathbf{R}^2)$ for all $k \geqslant 3$, and we are led to

$$\Delta w^+ = \frac{\sigma^2}{\kappa^2} e^{u_0+w^+}(e^{u_0+w^+-1})+g, \quad \text{in } \mathbf{R}^2,$$

$$w_\rho \to 0, \quad \text{as } |x|\to\infty.$$

This implies that w^+ is a solution to (5.5). Similarly, w^- is a solution to (5.5) and $w_\rho \to w^-$ in $H^k(\mathbf{R}^2)$ for all $k \to 1$ when $\rho < 0$.

To verify that w^+ is a maximal solution to (5.5), it suffices to prove that $w \leqslant w_\rho$ for given any solution w to (5.5) because $\{w_\rho\}$ is a decreasing sequence converging to w^+ for $\rho > 0$. We recall that for $\rho > 0$, w_ρ was constructed by the iterative sequence w_n^ρ defined by (2.7). Thus, it is enough to show that $w \leqslant w_n^\rho$ for all $k \geqslant 0$.

Let W^ρ be the unique solution of (2.6) corresponding to $\rho > 0$, satisfying

$$\Delta W^\rho = \frac{\sigma^2}{\kappa^2} e^{2u_*^\rho}(2-e^{-u_*^\rho})(e^{u_0+W^\rho}-1)+g, \quad \text{in } \mathbf{R}^2,$$

$$W^\rho \to 0, \quad \text{as } |x|\to\infty.$$

Since $u_*^\rho > 0$, we have

$$\Delta w \geqslant \frac{\sigma^2}{\kappa^2} e^{2u_*^\rho}(2-e^{-u_*^\rho})(e^{u_0+w}-1)+g.$$

Hence

$$\Delta(w - W^\rho) \geqslant \frac{\sigma^2}{\kappa^2} e^{2u_*^\rho}(2-e^{-u_*^\rho})e^{u_0+w'}(w-W^\rho),$$

for some w' between w and W^ρ. The maximum principle implies that $w \leqslant W^\rho = w_0^\rho$. Suppose that $w \leqslant w_k^\rho$ for all $0 \leqslant k \leqslant n$. Then

$$(\Delta - K)(w - w_{n+1}^\rho) \geqslant \frac{\sigma^2}{\kappa^2} e^{2u_*^\rho}e^{2u_0}(e^{2w}-e^{2w_n^\rho}) - \frac{\sigma^2}{\kappa^2} e^{u_*^\rho}e^{u_0}(e^{w}-e^{w_n^\rho}) - K(w-w_n^\rho)$$
$$\geqslant K(e^{2u_0+2w'}-1)(w-w_n^\rho) \geqslant 0,$$

where $w \leqslant w' \leqslant w_n^\rho \leqslant w_0^\rho$. Now the maximum principle verifies that $w \leqslant w_{n+1}^\rho$, which completes the proof of maximality of w^+. □

ACKNOWLEDGMENTS

The first author has been supported partially by GARC-KOSEF, BSRI-MOE, KOSE F(K95070) and the SNU Research Fund, while the second author has been supported partially by GARC-KOSEF.

[1] J. Hong, Y. Kim, and P. Y. Pac, ''Multivortex solutions of the Abelian Chern–Simons–Higgs theory,'' Phys. Rev. Lett. **64**, 2230–2233 (1990).
[2] R. Jackiw and E. J. Weinberg, ''Self-dual Chen-Simons vortices,'' Phys. Rev. Lett. **64**, 2234–2237 (1990).
[3] R. Jackiw, K. Lee, and E. J. Weinberg, ''Self-dual Chern–Simons solitons,'' Phys. Rev. D **42**, 3488–3499 (1990).
[4] J. Spruck and Y. Yang, ''Topological solutions in the self-dual Chern–Simons theory,'' Ann. Inst. Henri Poincaré **12**, 75–97 (1995).
[5] R. Wang, ''The existence of Chern–Simons vortices,'' Commun. Math. Phys. **137**, 587–597 (1991).
[6] J. Spruck and Y. Yang, ''The existence of non-topological solitons in the self-dual Chern–Simons theory,'' Commun. Math. Phys. **149**, 361–376 (1992).
[7] D. Chae and O. Yu. Imanuvilov, ''The existence of non-topological multivortex solutions in the relativistic self-dual Chern–Simons theory,'' RIM-GARC Preprint Series 97–54 (1997).
[8] K. Lee, ''Vortex dynamics in the self-dual Maxwell-Higgs systems with a uniform background electric charge density,'' Phys. Rev. D **49**, 4265–4276 (1994).
[9] K. Lee and P. Yi, ''Self-dual anyons in uniform background fields,'' Phys. Rev. D **52**, 2412–2421 (1995).
[10] A. Jaffe and C. H. Taubes, *Vortices and Monopoles* (Birkhäuser, Boston, 1980).

On the spectral theory of dispersive *N*-body Hamiltonians

Mondher Damak
Université Paris 7, UMR 9994 Laboratoire de Physique, Mathématique et Géométrie, Case 7012 2, Place Jussieu, F-75251, cedex 05, Paris

(Received 20 June 1997; accepted for publication 14 September 1998)

In this work we describe a general class of dispersive *N*-body Hamiltonians for which we prove the Hunziker, van Winter, and Zislin (HWZ) theorem and a Mourre estimate outside a closed and countable set of energies called thresholds. As a consequence of the Mourre estimate we prove a strong form of the limiting absorption principle, which implies the absence of a singular continuous spectrum and gives criteria of local smoothness. © *1999 American Institute of Physics.* [S0022-2488(99)00901-9]

I. INTRODUCTION

The basic Hamiltonian describing *N* nonrelativistic interacting particles is

$$H=\sum_{i=1}^{N}\frac{-1}{2m_i}\Delta_{x_i}+\sum_{i<j}V_{ij}(x_i-x_j)=\sum_{i=1}^{N}\frac{1}{2m_i}P_i^2+\sum_{i<j}V_{ij}(x_i-x_j),$$

where $x_i\in\mathbb{R}^3$, $m_i>0$ and $P_i=-i\nabla_{x_i}$ are the position, mass, and momentum of particle number *i*. Important results concerning the spectral theory of this type of Hamiltonian have been obtained during the last years (see Ref. 1). One of the oldest (but fundamental results, the theorem of Hunziker, van Winter, and Zislin (HWZ) concerning the essential spectrum of such an operator, has been established in Refs. 2–4; also see Refs. 5–6. This theorem says that the bottom of the essential spectrum of a nonrelativistic multiparticle operator is determined by the two-cluster decompositions of the system of particles. In Ref. 7 Lewis, Siedentop, and Vugalter show that the result still holds for multiparticle relativistic Hamiltonians too, in this case the kinetic energy $(1/2m_i)P_i^2$ being replaced by $(P_i^2+m_i^2)^{1/2}$.

A fundamental step in the study of more subtle spectral properties of *N*-body Schrödinger operators *H* is the proof that they obey a Mourre estimate at all nonthreshold points. The first proof of this result is due to Mourre, Ref. 8, for $N=3$ and to Perry, Sigal, and Simon, Ref. 9, for general *N*; these proofs were simplified afterward by Froese–Herbst, Ref. 10.

Our purpose here is to prove the HWZ theorem and the Mourre estimate (outside a set of points called thresholds) for a rather general class of *N*-body dispersive systems, i.e., for Hamiltonians of the form

$$H=h(P)+\sum_{i<j}V_{ij}(x_i-x_j),$$

where $P=-i\nabla$ and *h* is a function continuous and divergent at infinity.

In particular, the relativistic and nonrelativistic Hamiltonians are included. The class of interactions that we consider is sensibly more general than usual, e.g., they can be nonlocal and quite singular. In the case of relativistic operators the Coulomb potentials are allowed.

The first proof that a dispersive *N*-body operator satisfies a Mourre estimate is due to Derezinski, Ref. 11, but his hypotheses contain implicit conditions that are difficult to check on examples. After this, Gerard, Ref. 12, gave a new proof of the estimate for what he called ‘‘regular dispersive systems.’’ Both proofs are geometric in nature, being natural extensions of the proof of

Froese and Herbst. Also, the proof of the HWZ theorem in the relativistic case due to Lewis, Siedentop, and Vugalter follows the geometric ideas introduced by Zislin in this context (see Ref. 6).

In this paper we use quite different methods, namely, the C^*-algebra techniques introduced in Refs. 13–14 that require neither partitions of unity in configuration space nor equations of the Weinberg–van Winter type. The algebraic techniques allows us to prove in a rather simple way the main results of this paper (Theorems III.1 and IV.1) for a very general class of kinetic energies. For example, in the HWZ theorem h is an arbitrary continuous function such that $h(k)\to\infty$ as $|k|\to\infty$. For the Mourre estimate the only regularity condition that we need is that h be of class C^1 and $|kh'(k)|\leqslant c(1+|h(k)|)$ (plus positivity conditions of the same nature as those of Gerard). Moreover, the interaction terms of our Hamiltonians can be nonlocal operators. The results of Lewis, Siedentop, and Vugalter cover only relativistic particles with two-body local interactions. Gerard requires h to be a smooth symbol and the interactions to be multiplication operators.

As a consequence of the Mourre estimate we get the absence of singular continuous spectrum, discreteness of eigenvalues outside the thresholds, and local (outside the thresholds) H smoothness of operators decaying like $\langle x\rangle^{-s}$ for all $s>\frac{1}{2}$.

The organization of this paper is as follows. In Sec. II we introduce the algebraic framework in which our investigations will be done. The main results of Sec. III are Theorem III.1, which is a general criterion for an operator to be affiliated to the C^*-algebra $\mathcal{I}$, and the dispersive HWZ theorem. We also show that the N-body relativistic Hamiltonians with Coulomb interactions are covered by our general results. Section IV is devoted to the proof of the Mourre estimate for a rather general class of N-body dispersive systems. Our conditions (both on the kinetic energy and on the interactions) are of the same nature as those of Gerard, but are simpler and more general. In Sec. IV.4 we present several explicit examples of dispersive Hamiltonians verifying all our conditions. Finally, the Appendix contains a proof of the fact that the algebras $\mathcal{I}(a)$ we are using are independent of the choice of the spaces X^a.

II. PRELIMINARIES

II.1. Throughout this paper $\mathcal{L}$ denotes a finite lattice, i.e., a finite partially ordered set such that for any two elements $a,b\in\mathcal{L}$ their upper and lower bounds $a\vee b, a\wedge b$ exist in $\mathcal{L}$. In the usual setting of the N-body problem, $\mathcal{L}$ is the lattice of partitions of the set $\{1,2,...,N\}$. The order relation in $\mathcal{L}$ is denoted $a\leqslant b$ and $a<b$ means strict inequality; $a_{\min}$ (resp., $a_{\max}$) are the least (resp., the largest) element of $\mathcal{L}$. We denote by $\mathcal{L}(2)$ the set of maximal elements of $\mathcal{L}\setminus\{a_{\max}\}$ (in the N-body case this is the set of two-cluster partitions).

II.2. Let X be a finite-dimensional real vector space equipped with a volume element dx, i.e., dx is a translation-invariant positive Radon measure on X. Equivalently, one may give a nonzero element e of the exterior product $\wedge^n X(n=\dim X)$, and if $(e_1,\ldots,e_n)$ is a basis in X such that $e_1\wedge\cdots\wedge e_n=e$ then one sets $\int f\,dx=\int_{\mathbb{R}^n} f(x_1e_1+\cdots+x_ne_n)dx_1\cdots dx_n$. Then $\mathcal{H}(X)$ will be the Hilbert space $L^2(X;dx)$ and we denote $\mathbb{B}(X)$ [resp., $\mathbb{K}(X)$] the C^*-algebra of bounded (resp., compact) operators in $\mathcal{H}(X)$. Note that $\mathbb{B}(X)$ and $\mathbb{K}(X)$ do not depend on the choice of the volume element.

Let X^* be the dual space of X. The Fourier transformation of a function $f:X\to\mathbb{C}$ is the function $\mathcal{F}f\equiv\hat{f}:X^*\to\mathbb{C}$ defined by $\hat{f}(k)=\int \exp[-i\langle x,k\rangle]f(x)dx$. There is a unique volume element dk on X^* such that $\mathcal{F}$ extends to an isometric operator of $L^2(X;dx)$ onto $L^2(X^*;dk)$. We shall always equip X^* with this volume element, and we set $\mathcal{H}(X^*)=L^2(X^*;dk)$.

If $f:X\to\mathbb{C}$ is a Borel function, we denote $f(Q)$ the operator of multiplication by f in the space $\mathcal{H}(X)$. If $g:X^*\to\mathbb{C}$ is a Borel function, then $g(P)$ is the operator in $\mathcal{H}(X)$ such that $\mathcal{F}g(P)\mathcal{F}^*$ is the operator of multiplication by g in $\mathcal{H}(X^*)$. Let $\mathbb{T}(X)$ be the C^*-subalgebra of $\mathbb{B}(X)$ consisting of operators of the form $f(P)$, where $f:X^*\to\mathbb{C}$ is continuous and convergent to zero at infinity.

II.3. If Y is a subspace of X provided with a volume element dy, then the Hilbert space $\mathcal{H}(Y)$, the Fourier transform $\mathcal{F}^Y$, and the C^*-algebras $\mathbb{B}(Y)$, $\mathbb{K}(Y)$, and $\mathbb{T}(Y)$ are well defined. By convention, if $Y=0\equiv\{0\}$ we put $\mathcal{H}(0)=\mathbb{K}(0)=\mathbb{T}(0)=\mathbb{C}$.

Now let Z be a subspace of X supplementary to Y, so $X=Y\oplus Z$. Then there is a unique volume element dz on Z such that $dx=dy\otimes dz$. Indeed, if e and u are elements of $\wedge^n X$ and $\wedge^k Y$ ($k=\dim Y$), which define the volume elements dx and dy, respectively, then there is a unique $v\in\wedge^m Z$ ($m=\dim Z$) such that $e=u\wedge v$; dz will be the volume element on Z associated to v. We then get a canonical isomorphism,

$$\mathcal{H}(X)\cong\mathcal{H}(Y)\otimes\mathcal{H}(Z)\cong L^2(Z;\mathcal{H}(Y)).$$

The decomposition $X=Y\oplus Z$ induces also a canonical identification $X^*=Y^*\oplus Z^*$, and if we make a Fourier transformation in the variable $z\in Z$ only we get a canonical isomorphism,

$$\mathcal{H}(X)\cong\mathcal{H}(Y)\otimes\mathcal{H}(Z^*)\cong L^2(Z^*;\mathcal{H}(Y)).$$

II.4. Let $\{X^a,X_a\}_{a\in\mathcal{L}}$ be a family of couples of vector subspace of X equipped with volume elements dx^a, dx_a such that the subspaces X^a and X_a are supplementary in X and $dx=dx^a\otimes dx_a$, $\forall a\in\mathcal{L}$. We suppose the following:
(i) if $b\leq a$ then $X^b\subset X^a$ and $X_a\subset X_b$;
(ii) $X^a+X^b=X^{a\vee b}$ and $X_a\cap X_b=X_{a\vee b}$;
(iii) $X^{a_{\max}}=X_{a_{\min}}=X$ and $X^{a_{\min}}=X_{a_{\max}}=\{0\}$.
Then for each a we have a canonical factorization,

$$\mathcal{H}(X)\cong\mathcal{H}(X^a)\otimes\mathcal{H}(X_a)\cong L^2(X_a;\mathcal{H}(X^a)).$$

If $S\in\mathbb{B}(X^a)$, $T\in\mathbb{B}(X_a)$ we shall denote by $S\otimes_a T$ the operator in $\mathcal{H}(X)$ that corresponds to $S\otimes T$ by the preceding isomorphism. If $\mathcal{C}_1\subset\mathbb{B}(X^a)$ and $\mathcal{C}_2\subset\mathbb{B}(X_a)$ are C^*-subalgebras, we denote by $\mathcal{C}_1\otimes_a\mathcal{C}_2$ the C^*-subalgebra of $\mathbb{B}(X)$ generated by the operators of the form $S\otimes_a T$ with $S\in\mathcal{C}_1$, $T\in\mathcal{C}_2$.

For each $a\in\mathcal{L}$, we introduce now the following C^*-subalgebra of $\mathbb{B}(X)$:

$$\mathcal{I}(a):=\mathbb{K}(X^a)\otimes_a\mathbb{T}(X_a),$$

and for each $a\in\mathcal{L}$ we define

$$\mathcal{I}_a:=\sum_{b\leq a,b\in\mathcal{L}}\mathcal{I}(b),\quad \mathcal{I}=\mathcal{I}_{a_{\max}}=\sum_{a\in\mathcal{L}}\mathcal{I}(a).$$

The sums that appear on the right-hand side (r.h.s.) have to be interpreted in the vector space sense. Note the following particular cases: $\mathcal{I}(a_{\min})=\mathbb{T}(X)$, $\mathcal{I}(a_{\max})=\mathbb{K}(X)$. The main property of these spaces are the following: $\mathcal{I}$ is a C^*-algebra, $\mathcal{I}(a)\mathcal{I}(b)\subset\mathcal{I}(a\vee b)$ for each $a,b\in\mathcal{L}$; the preceding sums are direct in the Banach space sense and each $\mathcal{I}_a$ is a C^*-subalgebra of $\mathcal{I}$.

These assertions can be proved without difficulty by mimicking the proof of Theorem 2.1 from Ref. 15 (or see Chap. 9 in Ref. 16). These facts can also be deduced from the more general formalism developed in Ref. 17 by observing that there are canonical isomorphisms between the algebras $\mathcal{I}(a)$ and the algebras denoted $\mathcal{C}(X_a^\sigma)$ in Sec. II of Ref. 17 (see pp. 35–36). This is interesting because it shows that *the C^*-algebras $\mathcal{I}(a)$ depend only on the choice of the subspace X_a of X; if other choices for the volume element on X_a and for the supplementary space X^a are made, then we get the same algebras $\mathcal{I}(a)$*. The freedom that one has in the choice of the subspaces X^a is quite useful in applications; cf. the examples at the end of Sec. IV. For this reason we shall give a direct and elementary proof of the independence of $\mathcal{I}(a)$ on X^a in an Appendix to this paper.

If $b\leq a$ then $X^b\subset X^a$ and we put $X_b^a=X^a\cap X_b$ so $X^a=X^b\oplus X_b^a$. We equip X_b^a with the volume element dx_b^a associated to dx^a and dx^b by the rule $dx^a=dx^b\otimes dx_b^a$, and we define

$$\mathcal{I}^a(b)=\mathbb{K}(X^b)\otimes\mathbb{T}(X^a_b),\quad \mathcal{I}^a=\sum_{b\leqslant a,b\in\mathcal{L}}\mathcal{I}^a(b).$$

The tensor product here is relative to the factorization $L^2(X^a)\cong L^2(X^b)\otimes L^2(X^a_b)$. Using the fact that $X=X^a\oplus X_a=X^b\oplus X^a_b\oplus X_a$ and $\mathbb{T}(X_b)=\mathbb{T}(X^a_b\oplus X_a)=\mathbb{T}(X^a_b)\otimes\mathbb{T}(X_a)$, it is easily shown that one has the following relation between $\mathcal{I}^a$ and $\mathcal{I}_a$:

$$\mathcal{I}_a=\mathcal{I}^a\otimes_a\mathbb{T}(X_a),$$

relative to the factorization $\mathcal{H}(X)\cong\mathcal{H}(X^a)\otimes\mathcal{H}(X_a)$.

II.5. Let $\mathcal{S}(X)$ be the space of a rapidly decreasing test function and $\mathcal{S}^*(X)$ its adjoint space, the space of temperate distributions on X. The operators $f(Q)$ and $g(P)$ have natural extensions to the space $\mathcal{S}^*(X)$ if f and g are of class C^∞ and are polynomially bounded together with all their derivatives. We denote $|\cdot|$ some Euclidean norms on X or X^* and we put $\langle x\rangle=(1+|x|^2)^{1/2}$. This allows us to define the operators $\langle Q\rangle^s$ and $\langle P\rangle^s$ for any $s\in\mathbb{R}$; they act in the space $\mathcal{S}^*(X)$. Then we define the Sobolev spaces $\mathcal{H}^s(X)=\{f\in\mathcal{S}^*(X)|\langle P\rangle^s f\in\mathcal{H}(X)\}$ with the natural topology. These spaces do not depend on the chosen norms.

We introduce now a class of weighted Sobolev spaces. Let $\theta\in C_0^\infty(X)$ be such that $\theta(x)>0$ if $2^{-1}<|x|<2$ and $\theta(x)=0$ otherwise. Choose one more function $\eta\in C_0^\infty(X)$ such that $\eta(x)>0$ if $|x|<1$. Then for any $s,\ t\in\mathbb{R}$ and $1\leqslant p\leqslant\infty$ let $\mathcal{H}^s_{t,p}$ be the space of distributions u that locally belong to $\mathcal{H}^s$, and such that

$$\|\eta(Q)u\|_{\mathcal{H}^s}+\left(\int_1^\infty\|r^t\theta(r^{-1}Q)u\|^p_{\mathcal{H}^s}\frac{dr}{r}\right)^{1/p}<\infty.$$

If $p=\infty$ the second term is interpreted as $\sup_{r\geqslant1}\|r^t\theta(r^{-1}Q)u\|_{\mathcal{H}^s}$. The left-hand side above is a norm on $\mathcal{H}^s_{t,p}$ that provides this space with a Banach space structure. We set $\mathcal{H}^s_t\equiv\mathcal{H}^s_{t,2}$; these are the usual weighted Sobolev spaces defined by the norms $\|\langle P\rangle^s\langle Q\rangle^t u\|$. If $t_1<t<t_2$, $t=(1-\lambda)t_1+\lambda t_2$ and $p,p_1,p_2\in[1,\infty]$, $s\in\mathbb{R}$, then

$$\mathcal{H}^s_{t,p}=(\mathcal{H}^s_{t_1,p_1},\mathcal{H}^s_{t_2,p_2})\lambda,p.$$

Moreover, if $1\leqslant p<\infty$ and $1/p+1/p'=1$, then $(\mathcal{H}^s_{t,p})^*=\mathcal{H}^{-s}_{-t,p'}$.

III. A DISPERSIVE HWZ THEOREM

We describe now some explicit classes of self-adjoint operators affiliated to $\mathcal{I}$. We shall keep the notations introduced above. Let us fix a continuous function $h:X^*\to\mathbb{R}$ such that $h(k)\to+\infty$ as $|k|\to\infty$. We denote by H_0 the self-adjoint operator $h(P)$ in the Hilbert space $\mathcal{H}(X)$. Since h is bounded from below H_0 is bounded from below too.

Recall the canonical isomorphism $X^*=X^{a*}\oplus X_a^*$. For each $a\in\mathcal{L}$, we define $h^a:X^{a*}\to\mathbb{R}$ by $h^a(k^a)=\inf\{h(k^a+k_a)|k_a\in X_a^*\}$. Observe that h^a is continuous and divergent at infinity. Then let $\mathcal{H}^h(X^a)$ be the form domain in $\mathcal{H}(X^a)$ of the operator $h^a(P^a)$ [we denote by P^a the momentum operator in $\mathcal{H}(X^a)$] equipped with the graph norm $\langle f,(1+|h^a(P^a)|)f\rangle^{1/2}$. After identifying $\mathcal{H}(X^a)$ and $\mathcal{H}(X^a)^*$, we get

$$\mathcal{H}^h(X^a)\subset\mathcal{H}(X^a)\subset\mathcal{H}^h(X^a)^*.$$

We shall consider Hamiltonians H of the form $H=H_0+V$ with perturbations of the form $V=\Sigma_a V^a\otimes_a 1$, where $\otimes_a$ is the tensor product determined by the factorization $\mathcal{H}(X)\cong\mathcal{H}(X^a)\otimes\mathcal{H}(X_a)$ and V^a is a sesquilinear form in $\mathcal{H}(X^a)$. Since $\mathcal{H}(X^{a_{\min}})=\mathbb{C}$, $V^{a_{\min}}$ is a real constant.

The next lemma allows us to define the total Hamiltonian for the interactions we have in mind.

Lemma III.1: Let V^a be a symmetric form on $\mathcal{H}(X^a)$ that is $h^a(P^a)$-form bounded with relative bound $\nu \geqslant 0$. Then $V_a = V^a \otimes_a 1$ is a symmetric form in $\mathcal{H}(X)$, $h(P)$-form bounded with relative bound equal to ν.

Proof: By hypothesis, V^a is a continuous sesquilinear form on the Hilbert space $\mathcal{H}^h(X^a)$ and there is a constant $c < \infty$ such that $|\langle u, V^a u\rangle| \leqslant \nu \langle u, h^a(P^a)u\rangle + c\|u\|^2$ for all $u \in \mathcal{H}^h(X^a)$. Then $V_a = V^a \otimes_a 1$ is a continuous symmetric form on $\mathcal{H}^h(X^a)\otimes \mathcal{H}(X_a)$. $\mathcal{H}^h(X)$ is the form domain of $h(P)$ and $\mathcal{H}^h(X) \subset \mathcal{H}^h(X^a)\otimes\mathcal{H}(X_a)$. We work in the representation $\mathcal{H}(X) \cong L^2(X_a; \mathcal{H}(X^a))$ and we make a Fourier transformation in the variable $x_a \in X_a$. Then we have $(V_a \hat{f})(k_a) = V^a \hat{f}(k_a)$ for $\hat{f} \in L^2(X_a^*; \mathcal{H}^h(X^a))$ and $k_a \in X_a^*$, while the operator $h(P)$ becomes the operator of multiplication by the operator valued function $k_a \mapsto h(P^a + k_a)$. Then, for each $\epsilon > 0$ there is a constant $c_\epsilon < \infty$ such that

$$
\begin{aligned}
\pm\langle f, V_a f\rangle &= \pm \int_{X_a^*} \langle \hat{f}(k_a), V^a \hat{f}(k_a)\rangle dk_a \\
&\leqslant (\nu+\epsilon)\int_{X_a^*}\langle \hat{f}(k_a), h^a(P^a)\hat{f}(k_a)\rangle dk_a + c_\epsilon \int_{X_a^*} \|\hat{f}(k_a)\|^2\, dk_a \\
&\leqslant (\nu+\epsilon)\int_{X_a^*}\langle \hat{f}(k_a), h(P^a+k_a)\hat{f}(k_a)\rangle dk_a + c_\epsilon \int_{X_a^*} \|\hat{f}(k_a)\|^2\, dk_a \\
&\leqslant (\nu+\epsilon)\langle f, h(P)f\rangle + c_\epsilon \|f\|^2, \quad \text{for all } f \in \mathcal{H}^h(X).
\end{aligned}
$$

This proves the lemma. ■

Let us recall the notion of smallness at infinity that is the weak decay assumption that we need.

Definition: Let s, $t \in \mathbb{R}$, let Y be a vector space and let L: $\mathcal{H}^s(Y) \to \mathcal{H}^t(Y)$ a linear continuous operator. We shall say that L is small at infinity if it satisfies one (and therefore all) of the following equivalent conditions:
(a) L: $\mathcal{H}^s(Y) \to \mathcal{H}^{t'}(Y)$ is compact for some $t' < t$;
(a') the preceding assertion holds for all $t' < t$;
(b) there is $\theta \in C^\infty(Y)$ with $\theta(y) = 0$ (resp., 1) if $|y| \leqslant 1$ (resp., $|y| \geqslant 2$) and there is $t' < t$ such that $\lim_{r\to\infty}\|\theta(Qr^{-1})L\|_{\mathbb{B}(\mathcal{H}^s, \mathcal{H}^{t'})} = 0$;
(b') L has the preceding property for all $\theta \in C^\infty(Y)$ with $\theta(y) = 0$ (resp., 1) near zero (resp., near ∞) and all $t' < t$.

For the proof of the equivalence see Definition 3.5 in Ref. 15. If $t = -s$ then giving a continuous linear operator $\mathcal{H}^s \to \mathcal{H}^{-s}$ is equivalent with giving a continuous sesquilinear form on $\mathcal{H}^s$. In particular, we get the notion of sesquilinear form on $\mathcal{H}^s$ small at infinity.

The algebraic techniques developed in Ref. 16 apply only to the self-adjoint operators H on $\mathcal{H}$ that are affiliated to the C^*-algebra $\mathcal{I}$. One says that H is *affiliated* to $\mathcal{I}$ if $(H-z)^{-1} \in \mathcal{I}$ for some complex number z outside the spectrum of H. If this is the case then $\varphi(H) \in \mathcal{I}$ for each continuous function on $\mathbb{R}$, which tends to zero at infinity.

Now we state and prove a criterion for an operator to be affiliated to $\mathcal{I}$. Observe that V^a may be nonlocal and that it could be a pseudodifferential operator of the same order as $h(P)$.

Let $\mu = 2s$, $s > 0$. From now on we suppose that *there are two constants c_1 and c_2 such that*

$$c_1\langle k\rangle^\mu \leqslant h(k) \leqslant c_2 \langle k\rangle^\mu, \quad \text{for all } k \in X^*.$$

Then we also have $c_1'\langle k^a\rangle^\mu \leqslant h^a(k^a) \leqslant c_2'\langle k^a\rangle^\mu$ for all $k^a \in X^{a*}$. It is clear that $\mathcal{H}^h(X^a) = \mathcal{H}^s(X^a)$.

Theorem III.1: *Assume that for each $a \in \mathcal{L}$ a continuous symmetric and small at infinity sesquilinear form V^a with domain $\mathcal{H}^s(X^a)$ is given such that $\Sigma_a \mu_a < 1$, where μ_a is the relative* *form bound of* V^a *with respect to* $h^a(P^a)$. *Identify* V^a *with an operator* $\mathcal{H}^s(X^a)\to\mathcal{H}^{-s}(X^a)$ *and assume that there is a* $t>s$ *such that* $V^a\mathcal{H}^t(X^a)\subset\mathcal{H}^{t-2s}(X^a)$. *Then* $H:=h(P)+\Sigma_{a\in\mathcal{L}}V_a$ *is a self-adjoint operator in* $\mathcal{H}(X)$ *with form domain equal to* $\mathcal{H}^s(X)$. *The same assertion holds for each* $H_a=h(P)+\Sigma_{b\leqslant a}V_b$. *The operator H is a affiliated to* $\mathcal{I}$ *and* H_a *is affiliated to* $\mathcal{I}_a$.

Proof: Our proof is similar to that of Proposition 3.6 in Ref. 15. By the closed graph theorem, V^a is continuous when considered as operator $V^a:\mathcal{H}^t(X^a)\to\mathcal{H}^{t-2s}(X^a)$. Without loss of generality we assume $s<t\leqslant 2s$. For any $\epsilon>0$ let us define $V^a_\epsilon=(1+\epsilon h^a(P^a))^{-1}V^a(1+\epsilon h^a(P^a))^{-1}$. Since $V^a:\mathcal{H}^{2s}(X^a)\to\mathcal{H}^{-s}(X^a)$ is a compact symmetric operator, we have $V^a_\epsilon\in\mathbb{K}(X^a)$ and V^a_ϵ is symmetric too. Since $(1+\epsilon h^a(P^a))^{-1}\to 1$ as $\epsilon\to 0$ strongly in each $\mathcal{H}^r(X^a)$, we see that $V^a_\epsilon\to V^a$ strongly in $B(\mathcal{H}^r(X^a),\mathcal{H}^{r-2s}(X^a))$ for all $r\in[2s-t,t]$ and in norm in $B(\mathcal{H}^r(X^a),\mathcal{H}^{r-2s-\delta}(X^a))$ for $r\in[2s-t,t]$ and $\delta>0$. Let $H_\epsilon:=h(P)+V_\epsilon$ with $V_\epsilon=\Sigma_a(V_a)_\epsilon$ and $(V_a)_\epsilon=V^a_\epsilon\otimes_a 1$. Then $(H_\epsilon)_a:=h(P)+\Sigma_{b\leqslant a}(V_b)_\epsilon$. Proposition 9.3.4 in Ref. 16 implies that H_ϵ is affiliated to $\mathcal{I}$. Hence, it is enough to show that $\lim_{\epsilon\to 0}H_\epsilon=H$ in a norm-resolvent sense. For this we use the identity ($\lambda>0$ large enough)

$$(\lambda+H_\epsilon)^{-1}-(\lambda+H)^{-1}=(\lambda+H_\epsilon)^{-1}[V-V_\epsilon](\lambda+H)^{-1}.$$

Since $(\lambda+H)^{-1}$ and $(\lambda+H_\epsilon)^{-1}$ are isomorphisms of $\mathcal{H}^{-s}(X)$ onto $\mathcal{H}^s(X)$ and $V-V_\epsilon:\mathcal{H}^s(X)\to\mathcal{H}^{-s}(X)$ the identity holds in $B(\mathcal{H}^{-s}(X),\mathcal{H}^s(X))$. We shall show two things: $(\lambda+H)^{-1}\mathcal{H}(X)\subset\mathcal{H}^r(X)$ for some $r>s$ and $\|(\lambda+H_\epsilon)^{-1}\|_{B(\mathcal{H}^{-s},\mathcal{H}^s)}\leqslant$const independently of ϵ; here λ is large enough and also independent of ϵ. Assuming these facts are true, we get

$$\|(\lambda+H_\epsilon)^{-1}-(\lambda+H)^{-1}\|_{B(\mathcal{H},\mathcal{H}^s)}\leqslant c\|V-V_\epsilon\|_{B(\mathcal{H}^r,\mathcal{H}^{-s})}\to 0,$$

because we may take $\delta=r-s>0$ above. Observe that $V^a_\epsilon\to V^a$ in norm in $B(\mathcal{H}^r(X^a),\mathcal{H}^{-s}(X^a))$, which implies $(V_a)_\epsilon\to V_a$ in norm in $B(\mathcal{H}^r(X),\mathcal{H}^{-s}(X))$.

Now we prove the assertion concerning $(\lambda+H_\epsilon)^{-1}$. Using the equality $(\lambda+H_\epsilon)^{-1}=(\lambda+h(P))^{-1/2}[1+(\lambda+h(P))^{-1/2}V^\epsilon(\lambda+h(P))^{-1/2}]^{-1}(\lambda+h(P))^{-1/2}$ it is enough to show that there are λ and $\mu<1$ independent of ϵ, such that $\|(\lambda+h(P))^{-1/2}V_\epsilon(\lambda+h(P))^{-1/2}\|\leqslant\mu<1$. But

$$(\lambda+h(P))^{-1/2}V_\epsilon(\lambda+h(P))^{-1/2}=\sum_a(\lambda+h(P))^{-1/2}(\lambda+\epsilon h^a(P^a))^{-1}V^a(\lambda+\epsilon h^a(P^a))^{-1}\otimes_a 1.(\lambda+h(P))^{-1/2};$$

hence

$$\|(\lambda+h(P))^{-1/2}V_\epsilon(\lambda+h(P))^{-1/2}\|\leqslant\sum_a\|(\lambda+h(P))^{-1/2}V_a(\lambda+h(P))^{-1/2}\|\leqslant\sum_a\left(\mu_a+\frac{c}{\lambda}\right),$$

where c/λ can be made as small as we want by choosing λ large enough. Finally, the proof of the assertion concerning $(\lambda+H_\epsilon)^{-1}$ is the same as in Ref. 15. ■

The main result of this section is the following generalization of the HWZ Theorem.

Theorem III.2: *For each* $a\in\mathcal{L}$ *define* $\tau_a=inf\ \sigma(H_a)$ *and let* $\tau=\min_{a\in\mathcal{L}(2)}\tau_a$. *If the hypotheses of Theorem III.1 hold, then* $\sigma_{ess}(H)=[\tau,\infty)$.

Proof: By Theorem III.1, H is affiliated to $\mathcal{I}$ and H_a is affiliated to $\mathcal{I}_a$. According to Proposition 8.4.2 in Ref. 16 (see also Theorem 3.2 in Ref. 15) we shall then have $\sigma_{ess}(H)=\cup_{a\in\mathcal{L}(2)}\sigma(H_a)$. So it suffices to show that for each $a\in\mathcal{L}(2),\sigma(H_a)$ is a half-line. Since H is bounded from below we may assume, without loss of generality, that H is positive; this implies $H_a\geqslant 0$. $\mathcal{I}_a$ is the tensor product of $\mathcal{I}^a$ with the Abelian C^*-algebra $\mathbb{T}(X_a)$. Working in the spectral representation of $\mathbb{T}(X_a)$ we see that there is a family $\{H^a(k_a),\ k_a\in X^*_a\}$ of Hamiltonians $H^a(k_a)$ affiliated to $\mathcal{I}^a$ such that $\mathcal{F}_a(H_a+1)^{-1}\mathcal{F}^*_a$ (where $\mathcal{F}_a=1\otimes_a\mathcal{F}^{X_a}$ is the operator of Fourier transformation in the variable $x_a\in X_a$ only) is the operator of multiplication by the operator-valued function $k_a\mapsto(H^a(k_a)+1)^{-1}\equiv F(k_a)$ (see remark 9.2.6 in Ref. 16). Moreover, this function is norm continuous, tends in norm to zero at infinity, and $0<F(k_a)\leqslant 1$. So $m(k_a)\equiv\sup F(k_a)=\|F(k_a)\|\to 0$ as $k_a\to\infty$. But $m(k_a)$ is the upper bound of the spectrum of $F(k_a)$ and depends continuously on k_a (by the norm continuity of F). Since X_a^* is a connected set we see that the range of m has to be an interval. Finally, one has $\sigma(H_a)=\cup_{k_a\in X_a^*}\sigma(H^a(k_a))$ (see Sec. 8.2.4 in Ref. 16), which clearly shows that $\sigma(H_a)$ is an interval which has to be a half-line. This finishes the proof. ■

Application: Relativistic Hamiltonians with Coulomb potentials

We consider the quantum mechanical many-body problem of electrons and fixed nuclei interacting via Coulomb forces with a relativistic form for the kinetic energy. The HWZ theorem for this model has been established in Ref. 7; here we show that this situation is covered by our general theorem, Theorem III.2.

We recall the framework from Ref. 7. The system consists of N_n identical nuclei of mass M and charge Z, and N_e electrons of mass m; let $N=N_n+N_e$. If the coordinates of the electrons and nuclei are denoted by $r_1,\ldots,r_{N_e}$ and $R_1,\ldots,R_{N_n}$, respectively, then the potential energy is

$$V=-\sum_{\nu=1}^{N_e}\sum_{\kappa=1}^{N_n}\frac{\alpha Z}{|r_\mu-R_\kappa|}+\sum_{1\leqslant\mu<\nu\leqslant N_e}\frac{\alpha}{|r_\mu-r_\nu|}+\sum_{1\leqslant\kappa<\lambda\leqslant N_n}\frac{\alpha Z^2}{|R_\kappa-R_\lambda|},$$

where α is the fine structure constant. Then the total Hamiltonian is

$$H=\sum_{\nu=1}^{N_e}\sqrt{|P_\nu|^2+m^2}+\sum_{\kappa=1}^{N_n}\sqrt{|P_\kappa|^2+M^2}+V,$$

with natural notations. It is convenient to introduce new notations: $x_i=r_i$, $m_i=m$ if $1\leqslant i\leqslant N_e$, and $x_{N_e+i}=R_i$, $m_{N_e+i}=M$ if $1\leqslant i\leqslant N_n$. We work in $L^2(X)$ with $X=\{x\in\mathbb{R}^{3N}|\Sigma_{i=1}^N m_i x_i=0\}$. Then $\mathcal{L}$ is the lattice of partitions of the set $\{1,2,...,N\}$, and for $a\in\mathcal{L}$ we take $X^a=\{x\in X|\Sigma_{k\in C}m_k x_k=0,\ \text{for each}\ C\in a\}$; $X_a=\{x\in X|x_i=x_j$ if i,j belong to the same cluster of $a\}$.

As we explained in II.4 only the choice of the spaces X_a is really important.

Theorem III.3: If $\alpha Z<2/\pi$ and $\alpha<\frac{1}{47}$, then $\sigma_{ess}(H)=[\tau,\infty)$, where $\tau=\min_{a\in\mathcal{L}(2)}\inf\sigma(H_a)$.

Remark: In corollary 2 (p. 24) of Ref. 7 two different sets of conditions are imposed. We have chosen the simplest one, the second being treated in the same way.

Proof: We define the function $h:X^*\to\mathbb{R}$ by $h(k)=\Sigma_{i=1}^N\sqrt{|k_i|^2+m_i^2}$, where $X^*=\{k\in\mathbb{R}^{3N}|\Sigma_{i=1}^N k_i=0\}$ is the dual (momentum) space of X. Then h is continuous and divergent at infinity, $\mathcal{H}^h(X^a)=\mathcal{H}^{1/2}(X^a)$ and $\mathcal{H}^h(X^a)^*=\mathcal{H}^{-1/2}(X^a)$ $(s=\frac{1}{2})$. Since, in our case H contains only two-body forces (i.e., $V^a\neq 0$ only if $a=l$ is a pair) each $V^a:\mathcal{H}^{1/2}(X^a)\to\mathcal{H}^{-1/2}(X^a)$ is small at infinity. Indeed, if $V^a\neq 0$ then $X^a\cong\mathbb{R}^3$ and each V^a is of the form $\gamma/|x|(\gamma\in\mathbb{R})$, which be may written as $V^a(x)=V_1(x)+V_2(x)$ with $V_1\in L^2(\mathbb{R}^3)$ and $V_2\in C_\infty(\mathbb{R}^3)$. We also have that $V^a\mathcal{H}^1(\mathbb{R}^3)\subset\mathcal{H}(\mathbb{R}^3)$ $(t=1>\frac{1}{2})$. If α is small the relative bound of V will be less than one, and we may apply directly Theorem III.2 to finish the proof. In order to get the stated result we argue as follows.

Let $V_-=\sup(0,-V)$ and $V_+=V+V_-$; then $V=V_+-V_-$ and $V_\pm\geqslant 0$. As it is explained in Ref. 7, it follows from the results of Lieb and Yau (Ref. 18) that there are constants $0<\mu<1$ and $\delta>0$ such that $0\leqslant V_-\leqslant\mu H_0+\delta$. Let $c>0$ and $H^c=H_0+cV_+-V_-$. Since $V_\pm$ are positive H_0-form bounded perturbations and the H_0-form bound of V_- is strictly less than 1, it follows that for each $c\geqslant 0$ (in fact, small negatives c are allowed) the operator H^c (defined as a form sum) is self-adjoint. Moreover, the resolvent of H^c depends analytically on $c\in(0,\infty)$. For small c we may apply now the theorems III.1 and III.2. But the result will remain true for all $c>0$ for analyticity reasons. Indeed, by series expansion (in c) one sees that the resolvent of H^c remains in the algebra

$\mathcal{I}$ (which is norm closed) for all c. Similarly, the resolvent of H_a^c will belong to $\mathcal{I}_a$ and the equality $\mathcal{P}_a[(H^c-z)^{-1}]=(H_a^c-z)^{-1}$, valid for small c, will hold for all $c>0$. So the HWZ Theorem holds for all $c>0$.

IV. THE MOURRE ESTIMATE

In this section we prove that for N-body dispersive systems the Mourre estimate holds outside a set of points (called thresholds), which can be described explicitly in terms of eigenvalues of some sub-Hamiltonians. The conjugate operator is the generator of dilations $D=\frac{1}{2}(\langle Q,P\rangle+\langle P,Q\rangle)$ on X. We will denote by D^a the generator of dilations on X^a ($D=D^{a_{\max}}$). We shall keep the notations introduced above.

IV.1. We recall some results concerning the conjugate operator method (see Ref. 16 for details). Let $\mathcal{H}$ be a Hilbert space and A a self-adjoint, densely defined operator in $\mathcal{H}$. If $S\in B(\mathcal{H})$, we denote $\mathcal{W}_\tau(S)=e^{-iA\tau}Se^{iA\tau}$; then $\{\mathcal{W}_\tau\}_{\tau\in\mathbb{R}}$ is a group of automorphisms of $B(\mathcal{H})$. The derivative $(d/d\tau)\mathcal{W}_\tau(S)|_{\tau=0}=i[S,A]$ exists as a sesquilinear form on $D(A)$, and it exists in the strong topology of $B(\mathcal{H})$ if and only if the sesquilinear form $[S,A]$ extends to a bounded operator in $\mathcal{H}$ (which will be denoted by the same symbol). In this case we say that S is of class $C^1(A)$ and we write $S\in C^1(A)$. If the derivative exists in the norm topology of $B(\mathcal{H})$, we say that S is of class $C_u^1(A)$. For the development of the theory, for example, in order to show that the limiting absorption principle holds, one has to require more regularity than C_u^1 (see Ref. 16 and references there). We say that $S\in B(\mathcal{H})$ is of class $\mathcal{C}^{1,1}(A)$ if $\int_0^1\|(\mathcal{W}_\tau-1)^2S\|(d\tau/\tau^2)<\infty$. If S is of class $\mathcal{C}^{1,1}(A)$ it will be automatically of class $C_u^1(A)$.

Now let H be a self-adjoint operator in $\mathcal{H}$. We shall say that H is of class $C^1(A)$, $C_u^1(A)$ or $\mathcal{C}^{1,1}(A)$ if $R(z)=(H-z)^{-1}$ has the corresponding property for some $z\in\mathbb{C}\setminus\sigma(H)$.

The proof of the Mourre estimate is simplified by the introduction of two functions $\rho_H\equiv\rho_H^A$ and $\tilde{\rho}_H\equiv\tilde{\rho}_H^A$ defined on $\mathbb{R}$ with values in $(-\infty,+\infty]$. Let H be any self-adjoint operator of class $C^1(A)$. For any $\lambda\in\mathbb{R}$ we denote the following:

$$\rho_H(\lambda)=\sup\{\alpha\in\mathbb{R}|\exists\varphi\in C_0^\infty(\mathbb{R})\text{real},\ \ \varphi(\lambda)\neq0,\ \text{ such that }\ \varphi(H)[H,iA]\varphi(H)\geqslant\alpha\varphi(H)^2\},$$

$$\tilde{\rho}_H(\lambda)=\sup\{\alpha\in\mathbb{R}|\exists\varphi\in C_0^\infty(\mathbb{R})\text{real},\ \ \varphi(\lambda)\neq0,\ \text{ and } K \text{ a compact operator, such that}$$
$$\varphi(H)[H,iA]\varphi(H)\geqslant\alpha\varphi(H)^2+K\}.$$

For a detailed description of the properties of the functions ρ and $\tilde{\rho}$, see Ref. 16.

We say that A is conjugate to H at the point $\lambda\in\mathbb{R}$ if $\tilde{\rho}_H(\lambda)>0$. Let $J\subset\mathbb{R}$ a Borel set; if $\tilde{\rho}_H(\lambda)>0$ for all $\lambda\in J$ we say that A is locally conjugate to H on J.

IV.2. For each $a\in\mathcal{L}$, we define the operator $H^a=h(P^a)+\Sigma_{b\leqslant a}V_b^a$ where $V_b^a=V^b\otimes1$, the tensor product being determined by the factorization $L^2(X^a)\cong L^2(X^b)\otimes L^2(X_b^a)$. We use the notation $\sigma_p(H)$ for the set of eigenvalues of a self-adjoint operator H. If H is a dispersive N-body Hamiltonian, one associates to it two sets: the set of thresholds of H, defined as $\tau(H):=\cup_{a<a_{\max}}\sigma_p(H^a)$, and the critical set of H, defined as $\kappa(H):=\cup_{a\in\mathcal{L}}\sigma_p(H^a)$. We define also the set $\mu^D(H)=\mathbb{R}\setminus\kappa(H)$.

The next lemma contains a technical results that will be needed in the proof of the main theorem of this section.

Lemma IV.1: Let X be a vector space, A a self-adjoint operator in a Hilbert space $\mathcal{H}$ and $\{H_x\}_{x\in X}$ a family of self-adjoint operators of class $C^1(A)$. Assume that for some (hence for all) $z\in\mathbb{C}\setminus\mathbb{R}$ the maps $x\mapsto(H_x-z)^{-1}$ and $x\mapsto[(H_x-z)^{-1},A]$ are norm continuous. Let $\lambda\in\mathbb{R}$; then
(a) The function $x\mapsto\rho_{H_x}(\lambda)$ is lower semicontinuous.
(b) Let U be a compact subset of X and $c_0>0$ a constant such that $c_0<\rho_{H_x}(\lambda)$ for all $x\in U$. For each $c_1\in(0,c_0)$ there exists $\varphi\in C_0^\infty(\mathbb{R})$ real, with $\varphi(\lambda)\neq0$, such that for all $x\in U$: $\varphi(H_x)[H_x,iA]\varphi(H_x)\geqslant c_1\varphi^2(H_x)$.

Proof: (a) By using, for example, the Helfer–Sjöstrand formula [or see (8.1.1) in Ref. 16]. One easily sees that the functions $x\mapsto\varphi(H_x)$ and $x\mapsto[\psi(H_x),iA]$ are continuous in norm for all

$\varphi,\ \psi\in C_0^\infty(\mathbb{R})$. On the other hand $\varphi(H_x)[H_x,iA]\varphi(H_x)=\varphi(H_x)[\psi(H_x),iA]\varphi(H_x)$ where ψ is any function in $C_0^\infty(\mathbb{R})$ such that $\psi(x)=x$ on supp φ, so the function $x\mapsto\varphi(H_x)[H_x,iA]\varphi(H_x)$ is also norm continuous.

Let $x_0\in X$ and $r\in\mathbb{R}$ be such that $\rho_{H_{x_0}}(\lambda)>r$. We must show that there is a neighborhood of x_0 on which $\rho_{H_x}(\lambda)>r$. There exist $a_2>a_1>r$ and $\varphi_0\in C_0^\infty(\mathbb{R})$ real with $\varphi_0(\mu)=1$ on a neighborhood of λ and such that $\varphi_0(H_{x_0})[H_{x_0},iA]\varphi_0(H_{x_0})\geqslant a_2\varphi_0^2(H_{x_0})$.

Let $\epsilon_0>0$. By a continuity argument there is a neighborhood V_0 of x_0 in X such that $\varphi_0(H_x)[H_x,iA]\varphi_0(H_x)\geqslant a_2\varphi_0^2(H_x)-\epsilon_0$ for all $x\in V_0$. Let $\psi_0\in C_0^\infty(\mathbb{R})$ be real with $\psi_0(\lambda)\neq 0$ and $\varphi_0(\mu)=1$ on supp ψ_0. Multiplying left and right the preceding inequality by $\psi_0(H_x)$ we get $\psi_0(H_x)[H_x,iA]\psi_0(H_x)\geqslant(a_2-\epsilon_0)\psi_0^2(H_x)$ for all $x\in V_0$. Finally, choose ϵ_0 such that $a_2-\epsilon_0=a_1$. Then $\psi_0(H_x)[H_x,iA]\psi_0(H_x)\geqslant a_1\psi_0^2(H_x)$; hence $\rho_{H_x}(\lambda)\geqslant a_1>r$ for all $x\in V_0$.

(b) As in the proof of (a) we show that each $x_0\in U$ has a neighborhood $U(x_0)$ in U such that there is $\varphi_0\in C_0^\infty(\mathbb{R})$ real with $\varphi_0(\mu)=1$ on a neighborhood of λ, with the property

$$x\in U(x_0)\Rightarrow c_1\varphi_0^2(H_x)\leqslant\varphi_0(H_x)[H_x,iA]\varphi_0(H_x).$$

Now the proof can be completed as follows. Since U is compact, on may choose a finite number of neighborhoods $U(x_1),\ldots,U(x_n)$ covering U. Let $\varphi_1,\ldots,\varphi_n$ be the corresponding functions constructed as above. Then for each $k\in\{1,...,n\}c_1\varphi_k^2(H_x)\leqslant\varphi_k(H_x)[H_x,iA]\varphi_k(H_x)$ for all $x\in U(x_k)$. Finally, let $\varphi\in C_0^\infty(\mathbb{R})$ real with $\varphi(\lambda)\neq 0$ and $\varphi_k(\mu)=1$ on supp φ for all k. Multiplying left and right the preceding inequality by $\varphi(H_x)$, we get $c_1\varphi^2(H_x)\leqslant\varphi(H_x)[H_x,iA]\varphi(H_x)$ for all $x\in U$. ■

We now state the main result of this section. Note that we only need that H to be of class $C_u^1(A)$. For physically interesting and nontrivial functions h satisfying all the conditions below (e.g., relativistic three-body kinetic energy) see Ref. 12.

Theorem IV.1: *Besides the hypotheses of Theorem III.1 we assume that h is of class* $C^1,|kh'(k)|\leqslant c(1+|h(k)|)$ *for all* $k\in X^*$, *and that* $k_b^a h'(k^b+k_b^a+k_a)\geqslant 0$ *for all* $a,b\in\mathcal{L}$ *with* $b<a$ *and all* $k^b\in X^{b*}, k_b^a\in X_b^{a*}, k_a\in X_a^*$. *Moreover, in the case* $a=\max\mathcal{L}$ *assume that the stronger inequality* $k_b h'(k^b+k_b)>0$ *holds if* $k_b\neq 0$. *Finally, assume that* $[D^a,V^a]$ *is a compact operator* $\mathcal{H}^{2s}(X^a)\to\mathcal{H}^{-2s}(X^a)$. *Then*
(i) $\tau(H)$ *is a closed and countable real set;*
(ii) the eigenvalues of H which do not belong to $\tau(H)$ *[i.e., the points of* $\kappa(H)\setminus\tau(H)$*] are of finite multiplicity and can accumulate only to points from* $\tau(H)$;
(iii) D is locally conjugate to H on $\mathbb{R}\setminus\tau(H)$.

Proof: At steps (a)–(e) below we shall prove that D is locally conjugate to H on $\mathbb{R}\setminus\tau(H)$. Now we explain the ''standard'' proof of (i) and (ii). The assertion (ii) is a consequence of the abstract results of Mourre (see, for example, Corollary 7.2.11 in Ref. 16). The set $\tau(H)$ is clearly countable, but the fact that it is closed is far from obvious. But note that the operators H^a on $L^2(X^a)$ have the same properties as the operator H on $L^2(X)$ [see step (c) below]. If $a\neq\max\mathcal{L}$ then the lattice $\mathcal{L}_a=\{b\in\mathcal{L}|b\leqslant a\}$ associated to H^a is strictly smaller than $\mathcal{L}$. So one can make the induction hypothesis that $\tau(H^a)=\cup_{b<a}\sigma_p(H^b)$ is closed and that the generator of dilations on $L^2(X^a)$ is locally conjugate to H^a on $\mathbb{R}\setminus\tau(H^a)$ (for $a=\min\mathcal{L}$ this is trivial). Then $\kappa(H^a)$ is closed and $\tau(H)=\cup_{a<\max\mathcal{L}}\kappa(H^a)$ is closed too.

(a) By a simple calculation one shows that $e^{-iDt}\mathcal{I}(a)e^{iDt}\subset\mathcal{I}(a)$, $\forall a\in\mathcal{L}$ and $t\in\mathbb{R}$. Moreover, one can see easily that $[(H+1)^{-1},iD]\in\mathcal{I}$, which implies that H is of class $C_u^1(D)$ (see p. 420 of Ref. 16). Hence, from Theorem 5.10 in Ref. 15 (or see Theorem 8.4.3 in Ref. 16), it follows that $\tilde{\rho}_H(\lambda)=\min_{a\in\mathcal{L}(2)}\rho_{H_a}(\lambda)$, so for the proof of (iii) it is sufficient to show that if $\lambda\notin\tau(H)$, then $\rho_{H_a}(\lambda)>0$ for all $a\in\mathcal{L}(2)$.

(b) Using the relation $D=D^a\otimes_a 1+1\otimes_a D_a$ (where D_a is the generator of dilations on X_a) we write $[H_a,iD]=[H_a,iD^a\otimes_a 1]+[H_a,i(1\otimes_a D_a)]$. If we work in the representation $\mathcal{H}(X)\cong L^2(X_a;\mathcal{H}(X^a))$ and we make a Fourier transformation in the variable $x_a\in X_a$; then for each real function $\varphi \in C_0^\infty(\mathbb{R})$ the operator $\varphi(H_a)[H_a, iD]\varphi(H_a)$ becomes the operator of multiplication by the operator-valued function:

$$k_a \mapsto \varphi(H^a(k_a))\{[H^a(k_a), iD^a] + k_a h'(P^a + k_a)\}\varphi(H^a(k_a)). \tag{1}$$

Since the family $\{H^a(k_a)\}_{k_a \in X_a^*}$ is proper ($a \neq a_{\max}$), we have $\varphi(H^a(k_a)) = 0$ if k_a is large enough and if φ has support in a fixed compact. Let us fix $\lambda \notin \tau(H)$ and a compact neighborhood K of λ. Then for each $a \in \mathcal{L}(2)$ there is a compact set U_a in X such that $\varphi(H^a(k_a)) = 0$ if $k_a \notin U_a$ and supp $\varphi \subset K$. So it is enough to show that there are $c_0 > 0$ and $\varphi \in C_0^\infty(\mathbb{R})$ with $\varphi(\lambda) \neq 0$ and supp $\varphi \subset K$ such that the r.h.s of (1) is larger than $c_0 \varphi^2(H^a(k_a))$ for all $k_a \in U_a$.

(c) Since $\lambda \notin \tau(H)$, the inequality $\rho_{H^a(0)}(\lambda) > 0$ is a straightforward consequence of the definition of $\tau(H)$. Note that $H^a(0) = H^a$. From (a) of the preceding lemma we conclude that the function $k_a \mapsto \rho_{H^a(k_a)}$ is lower semicontinuous. So there exists a compact neighborhood $\mathcal{V}_a$ of 0 in X_a^* and a constant $c_1 > 0$ such that $\rho_{H^a(k_a)}(\lambda) \geq c_1$ for each $k_a \in \mathcal{V}_a$. Now, from (b) of the preceding lemma, if we choose $c_2 \in (0, c_1)$, then there exists $\varphi \in C_0^\infty(\mathbb{R})$ with $\varphi(\lambda) \neq 0$ and supp $\varphi \subset K$, such that

$$\forall k_a \in \mathcal{V}_a, \varphi(H^a(k_a))[H^a(k_a), iD^a]\varphi(H^a(k_a)) \geq c_2 \varphi^2(H^a(k_a)). \tag{2}$$

(d) By induction over a and by using the equality $\tilde{\rho}_H(\lambda) = \min_{a \in \mathcal{L}(2)} \rho_{H_a}(\lambda)$, we easily get that $\tilde{\rho}_H(\lambda) \geq 0$. From the decomposition $\mathcal{I}_a = \mathcal{I}^a \otimes_a \mathbb{T}(X_a)$ we see that $H^a(k_a)$ has in $\mathcal{H}(X^a)$ properties similar to those of H in $\mathcal{H}(X)$. More precisely, the kinetic energy part of $H^a(k_a)$ is given by the function $k^a \mapsto h(k^a + k_a)$, and by hypothesis we have $k_b^a h'(k^b + k_b^a + k_a) \geq 0$ if $b < a$. It follows that $\tilde{\rho}_{H^a(k_a)}(\lambda) \geq 0$. Now, as a consequence of Lemma 7.2.12 in Ref. 16 we get $\forall \epsilon > 0$, $\exists \varphi \in C_0^\infty(\mathbb{R})$ with $\varphi(\lambda) \neq 0$, such that

$$\forall k_a \in U_a, \varphi(H^a(k_a))[H^a(k_a), iD^a]\varphi(H^a(k_a)) \geq -\epsilon \varphi^2(H^a(k_a)). \tag{3}$$

(e) From our hypotheses it follows that there is $c_3 > 0$ such that $k_a h'(P^a + k_a) \geq c_3 > 0$ for each $k_a \notin \mathcal{V}_a$. Note that this strict positivity condition has not been used until now.

Finally, if we take $\epsilon_0 \leq \inf(c_2, c_3)$ and $\varphi_0 \in C_0^\infty(\mathbb{R})$ with $\varphi_0(\lambda) \neq 0$, such that (3) holds for ϵ_0 and φ_0, we get

$$\varphi_0(H^a(k_a))\{[H^a(k_a), iD^a] + k_a h'(P^a + k_a)\}\varphi_0(H^a(k_a)) \geq \inf(c_2, c_3)\varphi_0^2(H^a(k_a));$$

this completes the proof of the theorem. ■

IV.3. We summarize in the next theorem some consequences concerning the spectral properties of the dispersive N-body Hamiltonian H. Here we need that H be of class $\mathcal{C}^{1,1}(D)$.

Theorem IV.2: *Besides the hypotheses of Theorem III.1 we assume that h is of class* C^2, $(|kh'(k)| + |k^2 h''(k)|) \leq c(1 + |h(k)|)$ *for all* $k \in X^*$, *and that* $k_b^a h'(k^b + k_b^a + k_a) \geq 0$ *for all* $a, b \in \mathcal{L}$ *with* $b < a$ *and all* $k^b \in X^{b*}$, $k_b^a \in X_b^{a*}$, $k_a \in X_a^*$. *Moreover, in the case* $a = \max \mathcal{L}$ *assume that the stronger inequality* $k_b h'(k^b + k_b) > 0$ *holds if* $k_b \neq 0$. *Assume also that* V^a *can be decomposed into a sum* $V^a = V_S^a + V_L^a$, *where* V_S^a: $\mathcal{H}^s(X^a) \to \mathcal{H}^{-s}(X^a)$ *(the short-range component) and* V_L^a: $\mathcal{H}^s(X^a) \to \mathcal{H}^{-s}(X^a)$ *(the long-range component) are symmetric operators satisfying the following decay conditions at infinity. Choose* $\theta \in C_0^\infty(X)$ *such that* $\theta(x) > 0$ *if* $0 < 1 < |x| < 2$ *and* $\theta(x) = 0$ *otherwise and denote* $\|\cdot\|_{\mathcal{X}}$ *is the norm in* $\mathcal{X} \equiv B(\mathcal{H}^s(X^a), \mathcal{H}^{-s}(X^a))$. *Then we assume* $[V_L^a, iD^a] \in \mathcal{X}$ *and*

$$\int_1^\infty \left\{ \left\| \theta\left(\frac{Q^a}{r}\right) V_S^a \right\|_{\mathcal{X}} + \left\| \frac{1}{r}\, \theta\left(\frac{Q^a}{r}\right)[V_L^a, iD^a] \right\|_{\mathcal{X}} \right\} dr < \infty.$$

Then $\tau(H)$ *is a closed countable set, the eigenvalues of H outside* $\tau(H)$ *are of finite multiplicity and can accumulate only at points belonging to* $\tau(H)$, *and H has no singularly continuous spectrum.*

The limits $\lim_{\mu\to\pm 0}(H-\lambda-i\mu)^{-1}$ *exist in the* weak* *topology of* $B(\mathcal{H}^{-s}_{1/2,1},\mathcal{H}^{s}_{-1/2,\infty})$ *uniformly in* λ *on each compact subset of the open real set* $\mu^D(H)$. *Finally, if* $\mathcal{K}$ *is a Hilbert space and T:* $\mathcal{H}(X)\to\mathcal{K}$ *is a linear operator that is continuous when* $\mathcal{H}(X)$ is equipped with the topology induced by $\mathcal{H}^{s}_{-1/2,\infty}$, *then T is locally H-smooth on* $\mu^D(H)$.

Proof: The first assertion will hold inductively given the conclusion of the second assertion for subsystems and the definition of thresholds. Our hypotheses imply that H is of class $\mathcal{C}^{1,1}(D)$; for this we use Proposition 7.5.7 and Theorem 7.5.8 in Ref. 16 with $\Lambda=\langle Q\rangle$. Since H is bounded from below, it has a spectral gap. Then the second and the third assertion are consequences of Theorem 7.3.1 and Theorem 7.4.2 of Ref. 16 and also of the first assertion. The last assertion follows from Theorem 7.4.1 and Proposition 7.4.4 of Ref. 16. ■

IV.4. We give now three classes of examples of dispersive systems, for which we prove that the hypotheses of Theorem IV.1 hold.

Example 1: Assume that X is an Euclidean space identified with X^* and let g be a function g: $[0,\infty)\to\mathbb{R}$, such that

$$g'(t)\geqslant c>0,$$

$$tg'(t)\leqslant \mathrm{const}(1+g(t)),$$

for all $t\geqslant 0$. Then all the hypotheses of Theorem IV.1 hold for $h(k)=g(|k|)$ if we take $X^a=X_a^\perp$ $(\forall a)$.

Example 2: Let $X=\mathbb{R}^\nu\times\mathbb{R}^\nu$ where $\nu\geqslant 1$, and let

$$H=\omega(P_1)+\omega(P_2)+V_1(x_1)+V_2(x_2)+V_{12}(x_1-x_2).$$

We assume that $\omega:\mathbb{R}^\nu\to\mathbb{R}$ is a function such that $\omega(k)=U(|k|)$ for some increasing and convex function U. In particular, the Hamiltonian of two relativistic particles with the same mass is included.

We identify $X^*=X$, so $h:X^*\to\mathbb{R}$ is given by $h(k)=\omega(k_1)+\omega(k_2)$ for $k=(k_1,k_2)$. The set of subspace $\{X_a\}$ is the following: if $a=\{(1,2)\}$ then $X_a=\{x\in X|x_1=x_2\}$; if $a=i$ $(i=1,2)$ then $X_a=\{x\in X|x_i=0\}$. We take $X_{a_{\max}}=0$ and $X_{a_{\min}}=X$.

Let us check that (1) if $a<b$ then $k_a^b h'(k^a+k_a^b+k_b)\geqslant 0$;

(2) if $b=a_{\max}$ then $k_a h'(k^a+k_a)>0$ for $k_a\neq 0$.

(i) For $a=a_{\min}$: we have $X^{a_{\min}}=0$.
(a) $kh'(k)=k_1\omega'(k_1)+k_2\omega'(k_2)=|k_1|U'(|k_1|)+|k_2|U'(|k_2|)>0$ if $k\neq 0$ because U is increasing, so (2) holds.
(b) In this case, (1) is equivalent to check that $k^b h'(k^b+k_b)\geqslant 0$ if $b=1,2,\{(1,2)\}$. For $b=i$, $(i=1,2)$ we take $X^b=\{x\in X|x_j=0, j\neq b\}$ and we check easily that $k^b h'(k^b+k_b)\geqslant 0$. For example, if $b=1$, then $k^b=(k_1,0)$, $k_b=(0,k_2)$ and $k^b h'(k^b+k_b)=k_1\omega'(k_1)=|k_1|U'(|k_1|)\geqslant 0$.

For $b=\{(1,2)\}$ we take $X^b=\{x\in X|x_1=-x_2\}$, then $k^b=(k_2,-k_2)$, $k_b=(k_1,k_1)$. Let us set $x=k_1+k_2$ and $y=k_1-k_2$; then

$$k^b h'(k^b+k_b)=k_2\omega'(k_1+k_2)-k_2\omega'(k_1-k_2)=1/2\left[(x-y)\,\frac{x}{|x|}\,U'(|x|)-(x-y)\,\frac{y}{|y|}\,U'(|y|)\right].$$

We use polar coordinates: $x=re$, $y=\rho f$, where e,f are two unitary vectors and $r,\rho\geqslant 0$, and we set $\theta=e\cdot f\in[-1,1]$. Then $k^b h'(k^b+k_b)\geqslant 0$ is equivalent to $rU'(r)+\rho U'(\rho)\geqslant\theta(rU'(\rho)+\rho U'(r))$. So it suffices to check that the last inequality holds for $\theta=1$, which is true because U' is increasing.

(ii) If $a=1,2$ then $b=a_{\max}$ and the hypotheses (1) and (2) are easy to check. For example, if $a=1$ then $k^a=(k_1,0)$, $k_a=(0,k_2)$ and $k_a h'(k^a+k_a)=k_2\omega'(k_2)=|k_2|U'(|k_2|)>0$ for $k_2\neq 0$.

(iii) For $a=\{(1,2)\}$ then $k^a=(k_1,-k_1)$, $k_a=(k_2,k_2)$, and as in (i) we have that $k_a h'(k^a+k_a)=k_2\omega'(k_1+k_2)+k_2\omega'(k_2-k_1)=k_2\omega'(k_1+k_2)-k_2\omega'(k_1-k_2)>0$ if $k_a\neq 0$.

Example 3: In this example we show the usefulness of the freedom one has in the choice of the spaces X^a supplementary to X_a. We consider two models of Hamiltonians describing particles with distinct kinetic energy:

$$H=\omega_1(P_1)+\omega_2(P_2)+V_1(x_1)+V_2(x_2)+V_{12}(x_1-x_2).$$

Then $h(k)=\omega_1(k_1)+\omega_2(k_2)$ for $k=(k_1,k_2)$. We shall keep the space X and the set of subspaces $\{X_a\}$ introduced in Example 2.

(a) The first model is simply the Hamiltonian of nonrelativistic particles. Then $\omega_i(k_i)=|k_i|^2/2m_i$ ($i=1,2$), where $m_i>0$ is the mass of particles i. By Example 2, we have to check only that: $k^a h'(k^a+k_a)\geqslant 0$ and $k_a h'(k^a+k_a)>0$ if $k_a\neq 0$ for $a=\{(1,2)\}$. Let $X^a=\{x\in X|x=(y,\alpha y),y\in\mathbb{R}^\nu\}$, with some $\alpha\in\mathbb{R}$. Then $k^a=(k_1,\alpha k_1)$, $k_a=(k_2,k_2)$ and

$$\begin{aligned}k^a h'(k^a+k_a)&=\langle k_1,\omega_1'(k_1+k_2)\rangle+\langle \alpha k_1,\omega_2'(\alpha k_1+k_2)\rangle\\&=\frac{1}{m_1}\langle k_1,(k_1+k_2)\rangle+\frac{\alpha}{m_2}\langle k_1,(\alpha k_1+k_2)\rangle.\end{aligned}$$

Let us set $t=|k_1|^2$ and $s=\langle k_1,k_2\rangle$. Then the last expression is positive for all $k_1,k_2\in\mathbb{R}^\nu$ if and only if $(m_2+\alpha^2 m_1)t+(m_2+\alpha m_1)s\geqslant 0$ for all $t\geqslant 0$ and $s\in\mathbb{R}$. But for a fixed t this is true for all $s\in\mathbb{R}$ only if $\alpha=-m_2/m_1$. Now with this α it is easy to check the second assertion.

(b) The second model appears in the theory of elasticity: we take $\omega_i(k_i)=c_i|k_i|^4$ ($i=1,2$) where $c_i>0$ is a constant. Here we suppose that $\nu=1$ and as in the last example we take $X^a=\{x\in X|x=(s,\alpha s),s\in\mathbb{R}\}$, with some $\alpha\in R$. Then $k^a=(s,\alpha s)$, $k_a=(t,t)$, and

$$\begin{aligned}k^a h'(k^a+k_a)&=s\omega_1'(s+t)+\alpha s\omega_2'(\alpha s+t)\\&=4(c_1+c_2\alpha^4)s^4+12(c_1+c_2\alpha^2)s^2t^2+4(c_1+c_2\alpha)st^3+12(c_1+c_2\alpha^3)s^3t.\end{aligned}$$

Then it is clear that (1) and (2) hold if $\alpha=-c_1/c_2$.

ACKNOWLEDGMENTS

The author takes this opportunity to express his gratitude to Vladimir Georgescu for suggesting to him the subject of this work and for many helpful discussions and to Anne Boutet de Monvel for useful advices and for the inspiring environment of the Laboratoire de Physique Mathématique et Géométrie. Finally, I am indebted to the referee for comments and criticism that allowed me to correct and clarify several points of the initial version of the paper.

APPENDIX

We shall prove here that the algebra $\mathcal{I}(a)$ is independent of the choice of the supplementary space X^a and of the various volume elements (see II.4).

Let us fix a vector subspace Y of X and let us choose a supplementary subspace Z to Y in X (so $X=Y+Z$, $Y\cap Z=\{0\}$) and some volume elements dy, dz on Y and Z. We identify $L^2(X)=L^2(Z)\otimes L^2(Y)$ and we define the C^*-algebra $\mathbb{K}(Z)\otimes\mathbb{T}(Y)$ as the norm closure in $\mathbb{B}(X)$ of the linear space $\mathcal{K}(Z,Y)$ generated by the operators of the form $S\otimes T$ with S compact in $L^2(Z)$ and $T\in\mathbb{T}(Y)$. The space $\mathcal{K}(Z,Y)$ depends on the choice of Z (see the last identity in the proof of the next lemma). Our purpose here is to show that its norm closure $\mathbb{K}(Z)\otimes\mathbb{T}(Y)$ is, however, independent of Z. For this we shall introduce a class of operators K, which clearly has the same norm closure as $\mathcal{K}(Z,Y)$, and then we show that this class of operators is independent of Z.

If $k\in C_c(Y\times Z\times Z)$ (spaces of continuous functions with compact support on $Y\times Z\times Z$) then

$$(Kf)(y+z)=\int_Y\int_Z k(y-y',z,z')f(y'+z')dy'\ dz'$$

defines an element of $\mathbb{B}(X)$. Indeed, if K_y is the operator in $L^2(Z)$ defined by the kernel $k(y,\cdot,\cdot)$ with respect to the volume element *dz*, then K_y is a Hilbert–Schmidt operator and

$$\|(Kf)(y+\cdot)\|_{L^2(Z)}\leqslant\int_Y\|K_{y-y'}f(y'+\cdot)\|_{L^2(Z)}\ dy'\leqslant\int_Y\|K_{y'}\|_{\mathbb{B}(Z)}\|f(y-y'+\cdot)\|_{L^2(Z)}\ dy'.$$

We may assume that $dx=dy\otimes dz$. Then we get

$$\|Kf\|_{L^2(X)}=\left[\int_Y\|(Kf)(y+\cdot)\|^2_{L^2(Z)}\ dy\right]^{1/2}\leqslant\int_Y\|K_{y'}\|_{\mathbb{B}(Z)}\ dy'\cdot\|f\|_{L^2(X)},$$

and the r.h.s. is a finite number.

Lemma: The class of operators of the form K is independent of the choice of Z, dy, dz.

Proof: Let Z_1,Z_2 be two supplemenary spaces to *Y* in *X* and let p_1 be the projection of *X* onto *Y* defined by Z_1. If $x\in X$ and $x=y_1+z_1=y_2+z_2$ with $y_1,\ y_2\in Y$ and $z_1\in Z_1,\ z_2\in Z_2$ then $p_1z_1=0$, $p_1y_2=y_2$, hence $y_1=y_2+p_1z_2$. Moreover, if we set $q_1=1-p_1$ then $q_1y_2=0$, so $z_1=q_1z_2$. Observe that $q_1|Z_2$ is a bijective linear map of Z_2 onto Z_1.

Now let $k_1\in C_c(Y\times Z_1\times Z_1)$ and choose some volume elements *dy*, dz_1 on *Y* and Z_1. Then define *K* by

$$(Kf)(x)=(Kf)(y_1+z_1)=\int_Y\int_{Z_1}k_1(y_1-y_1',z_1,z_1')f(y_1'+z_1')dy_1'\ dz_1'.$$

In the integral over Z_1 we make the change of variable $z_1'=q_1z_2'$. The volume element dz_1' is the image of a volume element z_2' on Z_2 through this map and we get

$$(Kf)(x)=\int_Y\int_{Z_2}k_1(y_1-y_1',z_1,q_1z_2')f(y_1'+q_1z_2')dy_1'\ dz_2'.$$

In the integral over *Y* (for a fixed z_2') we make the translation $y_1'=y_2'+p_1z_2'$ and get

$$(Kf)(x)=\int_Y\int_{Z_2}k_1(y_1-y_2'-p_1z_2',z_1,q_1z_2')f(y_1'+p_1z_2'+q_1z_2')dy_2'\ dz_2'.$$

We have $p_1z_2'+q_1z_2'=z_2'$. Now $x=y_1+z_1=y_2+z_2$ and $y_1=y_2+p_1z_2$, $z_1=q_1z_2$. We get

$$(Kf)(y_2+z_2)=\int_Y\int_{Z_2}k_1(y_2-y_2'+p_1(z_2-z_2',q_1z_2,q_1z_2')f(y_2'+z_2')dy_2'\ dz_2'$$

$$\equiv\int_Y\int_{Z_2}k_2(y_2-y_2',q_1z_2,q_1z_2')f(y_2'+z_2')dy_2'\ dz_2'.$$

The function $k_2:Y\times Z_2\times Z_2\to\mathbb{C}$ defined by $k_2(y,z_2,z_2')=k_1(y+p_1(z_2-z_2',q_1z_2,q_1z_2'))$ is continuous with compact support because $q_1|Z_2:Z_2\to Z_1$ is a bijective linear map. Now the assertion of the lemma is clear. ■

[1] J. Derezinski and Ch. Gerard, *Scattering Theory of Classical and Quantum N-Particle Systems* (Springer-Verlag, Berlin, 1997).
[2] W. Hunziker, ''On the spectra of Schrödinger multiparticle Hamiltonians,'' Helv. Phys. Acta **39**, 451–462 (1966).
[3] C. van Winter, ''Theory of finite systems of particles I.,'' Mat. Fys. Skr. Danske Vid. Selsk. **1**, 1–60 (1964).

[4] G. M. Zislin, ''Discussion of the spectrum of the Schrödinger operator for system of many particles,'' Trudy Mask. Mat. Obs. **9**, 81–120 (1960).
[5] M. Reed and B. Simon, Methods of Modern Mathematical Physics, Volume 4: Analysis of Operators (Academic, New York, 1978).
[6] H. L. Cycon, R. G. Froese, W. Kirsch, and B. Simon, Schrödinger Operators with Applications to Quantum Mechanics and Global Geometry (Springer-Verlag, Berlin, 1987).
[7] R. T. Lewis, H. Siedentop, and A. Vugalter, ''The essential spectrum of relativistic multi-particle operators,'' Ann. Inst. Henri Poincaré Phys. Theor. **67**, 1–28 (1997) (preprint number 96-68 at http://www.ma.utexas.edu/mp_arc).
[8] E. Mourre, ''Absence of singular continuous spectrum for certain self-adjoint operators,'' Commun. Math. Phys. **78**, 391–408 (1981).
[9] P. Perry, I. M. Sigal, and B. Simon, ''Spectral analysis of N-body Schrödinger operators,'' Ann. Math. **114**, 519–567 (1981).
[10] R. G. Froese and I. Herbst, ''A new proof of the Mourre estimate,'' Duke Math. J. **49**, 1075–1085 (1982).
[11] J. Derezinski, ''The Mourre estimate for dispersive N-body Schrödinger operators,'' Trans. Am. Math. Soc. **317**, 773–798 (1990).
[12] Ch. Gerard, ''The Mourre estimate for regular dispersive systems,'' Ann. Inst. Henri Poincaré Phys. Theor. **54**, 59–88 (1991).
[13] A. Boutet de Monvel and V. Georgescu, ''Graded C*-algebras in the N-body problem,'' J. Math. Phys. **32**, 3101–3110 (1991).
[14] A. Boutet de Monvel and V. Georgescu, ''Graded C*-algebras and many-body perturbation theory: II The Mourre estimate,'' in Méthodes Semi-Classiques, Colloque International, Nantes, June 1991, Astérisque, Vol. 210, Société Mathématique de France, 1992, Vol. 2, pp. 75–96.
[15] A. Boutet de Monvel V. Georgescu, and A. Soffer, ''N-Body Hamiltonians with hard core interactions,'' Rev. Math. Phys. **6**, 1–82 (1994).
[16] W. Amrein, A. Boutet de Monvel, and V. Georgescu, *C_0-Groups, Commutator Methods and Spectral Theory of N-Body Hamiltonians*, Birkhauser, Progress in Math. Series No. 135, 1996.
[17] A. Boutet de Monvel and V. Georgescu, ''Graded C*-algebras associated to symplectic spaces and spectral analysis of many-channel Hamiltonians,'' in *Dynamics of Complex and Irregular Systems* Bielefeld encounters in Math. Physics VIII, 1991 (World Scientific, Singapore, 1993), pp. 22–66.
[18] E. H. Lieb and H. T. Yau, ''The stability and instability of relativistic matter,'' Commun. Math. Phys. **118**, 177–213 (1988).

Gauge transformations in quantum mechanics and the unification of nonlinear Schrödinger equations

H.-D. Doebner[a)]
Arnold Sommerfeld Institute for Mathematical Physics and Institute for Theoretical Physics (A), Technical University of Clausthal, D-38678 Clausthal-Zellerfeld, Germany

G. A. Goldin[b)]
Departments of Mathematics and Physics, Rutgers University, New Brunswick, New Jersey 08903

P. Nattermann[c)]
Institute for Theoretical Physics (A), Technical University of Clausthal, D-38678 Clausthal-Zellerfeld, Germany

(Received 29 September 1997; accepted for publication 1 July 1998)

Beginning with ordinary quantum mechanics for spinless particles, together with the hypothesis that all experimental measurements consist of positional measurements at different times, we characterize directly a class of nonlinear quantum theories physically equivalent to linear quantum mechanics through nonlinear gauge transformations. We show that under two physically motivated assumptions, these transformations are uniquely determined: they are exactly the group of time-dependent, nonlinear gauge transformations introduced previously for a family of nonlinear Schrödinger equations. The general equation in this family, including terms considered by Kostin, by Bialynicki-Birula and Mycielski, and by Doebner and Goldin, with time-dependent coefficients, can be obtained from the linear Schrödinger equation through gauge transformation and a subsequent process we call *gauge generalization*. We thus unify, on fundamental grounds, a rather diverse set of nonlinear time evolutions in quantum mechanics. © *1999 American Institute of Physics.* [S0022-2488(98)04011-0]

I. INTRODUCTION

Recently a group $\mathscr{N}$ of nonlinear gauge transformations was introduced and shown to act as a transformation group in a family $\mathscr{F}$ of nonlinear Schrödinger equations (NLSEs).[1] The family $\mathscr{F}$ consists of equations with nonlinear terms of the type introduced by Kostin,[2] by Bialynicki-Birula and Mycielski,[3] by Guerra and Pusterla,[4] and by Doebner and Goldin,[5,6] with time-dependent coefficients.

A transformation $N_{(\gamma,\Lambda)}\in\mathscr{N}$ is labeled by two real, time-dependent parameters γ and Λ (with $\Lambda\neq 0$), and acts as a nonlinear analog of a gauge transformation in quantum mechanics. Letting the time-dependent wave function $\psi(\mathbf{x},t)$ on $\mathbf{R}^3$ be an arbitrary solution of any particular NLSE in $\mathscr{F}$, $N_{(\gamma,\Lambda)}$ is given by

$$\psi' = N_{(\gamma,\Lambda)}[\psi] = |\psi| \exp[i(\gamma \ln|\psi| + \Lambda \arg \psi)]. \tag{1}$$

Then ψ' solves a transformed equation that also belongs to $\mathscr{F}$.

The physical interpretation of this construction, developed briefly below, was elaborated on in some detail in Ref. 1. However, the underlying mathematical structure, and the physical reasons for the form of (1), remained somewhat hidden. Equation (1) was motivated in earlier work by the

a)Electronic mail: ashdd@pt.tu-clausthal.de
b)Electronic mail: gagoldin@dimacs.rutgers.edu
c)Electronic mail: aspn@pt.tu-clausthal.de

desire to linearize the equations in a special subset of $\mathscr{F}$, and to obtain stationary and nonstationary solutions.[6–11] The present paper takes a different, more fundamental approach to nonlinear gauge transformations and their consequences.

We begin with linear, nonrelativistic quantum mechanics for spinless particles in $\mathbf{R}^3$, together with the assumption, advocated for instance in Refs. 12–14, and discussed in Refs. 1, 9, 10, 15, and 16, that all experimental measurements consist fundamentally of positional measurements made at different times. Defining as usual the positional probability density $\rho(\mathbf{x},t) = \overline{\psi(\mathbf{x},t)}\psi(\mathbf{x},t)$, where ψ conventionally is a normalized solution of the linear Schrödinger equation, we are therefore interested in transformations N which leave $\rho(\mathbf{x},t)$ invariant—i.e., such that for all ψ in an appropriate domain of the unit sphere in the Hilbert space $\mathscr{H}$,

$$\overline{N[\psi](\mathbf{x},t)}N[\psi](\mathbf{x},t) = \overline{\psi(\mathbf{x},t)}\psi(\mathbf{x},t). \tag{2}$$

In addition, N should respect the prescription for writing the wave function subsequent to an ideal positional measurement. A conventional prescription for such a measurement at time t_1 consists of a projection in a region B of position space (a Borel subset of $\mathbf{R}^3$), with normalization,

$$\psi_s(\mathbf{x},t_1) = \begin{cases} \dfrac{\psi(\mathbf{x},t_1)}{(\int_B |\psi(\mathbf{x},t_1)|^2 d^3x)^{1/2}}, & \mathbf{x} \in B \\ 0, & \mathbf{x} \notin B \end{cases} \tag{3}$$

followed by time-evolution of $\psi_s(\mathbf{x},t)$ for $t > t_1$ (here the subscript ''s'' stands for ''subsequent''). As N should respect this prescription, we need for all ψ, $\mathbf{x}$ and $t \geqslant t_1$,

$$|N[\psi]_s(\mathbf{x},t)|^2 = |\psi_s(\mathbf{x},t)|^2, \tag{4}$$

and because of (2),

$$|N[\psi]_s(\mathbf{x},t)|^2 = |N[\psi_s](\mathbf{x},t)|^2. \tag{5}$$

We remark that in writing (3) we do not intend to express a commitment to a particular formalism for describing measurement. We merely note that the justification of N as a gauge transformation requires that in addition to (2) it leave invariant the outcomes of *sequences* of positional measurements at various times. Equation (3) is one prescription for predicting such outcomes in quantum mechanics.

Now if all actual measurements (outcomes of experiments) are obtained from positional measurements performed at various times, it can be argued that a system with states ψ obeying the Schrödinger equation, and one with states $N[\psi]$ obeying a transformed equation, have the same physical content. But we make two essential observations:

(a) Equations (2) and (4)–(5) do not require N to be a linear transformation—nonlinear N are also possible.
(b) Such nonlinear choices of N will transform a system governed by the usual, linear Schrödinger equation to physically equivalent systems obeying NLSEs that are, of course, linearizable (by construction).

The usual formulation and interpretation of quantum mechanics is based quite deeply on linearity and linear structures—superposition principle, on observables modeled by self-adjoint linear operators, on a linear time-evolution equation for the states, on a measurement process involving orthogonal projection onto linear subspaces for all sorts of observables, and on the description of mixed states by density matrices. Any proposal for nonlinearity in quantum mechanics requires a revised mathematical formulation and physical interpretation of all these ideas.

Here the linearizable NLSEs obtained using N can be useful. Due to their physical equivalence with linear quantum mechanics, they serve as a kind of ''laboratory'' for exploring how to generalize quantum mechanics to accommodate nonlinearities.

When N is *assumed* to be linear (and densely defined), Eq. (2) implies that it is a *unitary multiplication operator* for each t. Then N is labeled by a measurable function $\theta(\mathbf{x},t)$, and we have

$$\psi'(\mathbf{x},t)=(\mathbf{U}_\theta\psi)(\mathbf{x},t)=\exp[i\theta(\mathbf{x},t)]\psi(\mathbf{x},t). \tag{6}$$

Any such $\mathbf{U}_\theta$ commutes with the projection in (3), thus ensuring (5) and respecting the conventional prescription for wave functions subsequent to a positional measurement.

If θ is independent of $\mathbf{x}$ and t, we have just introduced a fixed phase, sometimes called a ''gauge transformation of the first kind.'' This changes neither the Schrödinger equation nor the form of position and momentum operators. A space- and time-dependent, linear $U(1)$-gauge transformation, implemented by (6), is sometimes called a ''gauge transformation of the second kind.'' Such transformations constitute an Abelian group $\mathscr{U}_{\mathrm{loc}}$ of *local* unitary operators acting on $\mathscr{H}$. The physical equivalence of the two theories, with states ψ and ψ', respectively, is guaranteed by the invariance of the outcomes of sequences of positional observations at all times.

A system with wave functions governed by the (linear) Schrödinger equation

$$i\hbar\partial_t\psi=-\frac{\hbar^2}{2m}\nabla^2\psi+V\psi \tag{7}$$

is transformed by (6) to a physically equivalent system, with wave function ψ' and Schrödinger equation

$$i\hbar\partial_t\psi'=\frac{1}{2m}\left(\frac{\hbar}{i}\nabla-\hbar\nabla\theta\right)^2\psi'+(V-\hbar\partial_t\theta)\psi'. \tag{8}$$

This observation suggests a *generic way* to construct new systems that are *not* physically equivalent to the original family given by (7) and (8). The scalar term $-\hbar\partial_t\theta$ and the vector term $\hbar\nabla\theta$ are merely special choices. If we replace $-\hbar\partial_t\theta$ by a *general* scalar field $\hat{\Phi}(\mathbf{x},t)$ and $\hbar\nabla\theta$ by a *general* vector field $\hat{\mathbf{A}}(\mathbf{x},t)$, calling them (Abelian) *gauge fields*, we obtain two well-known and well-established structures:

(a) a family $\mathscr{F}_{(\hat{\Phi},\hat{\mathbf{A}})}$ of time-evolution equations labeled by the gauge fields $\hat{\Phi}$ and $\hat{\mathbf{A}}$:

$$i\hbar\partial_t\psi=\frac{1}{2m}\left(\frac{\hbar}{i}\nabla-\hat{\mathbf{A}}\right)^2\psi+(V+\hat{\Phi})\psi; \tag{9}$$

and

(b) an action on this family by the gauge transformations $\mathbf{U}_\theta$, to establish equivalence classes—so that Schrödinger equations with $(\hat{\Phi},\hat{\mathbf{A}})$ and with $(\hat{\Phi}',\hat{\mathbf{A}}')=(\hat{\Phi}-\hbar\partial_t\theta,\hat{\mathbf{A}}+\hbar\nabla\theta)$ describe physically equivalent systems—with the family being closed under the action of gauge transformations.

This generic construction, which we here call *gauge generalization*, is physically relevant because external electromagnetic fields $(\Phi,\mathbf{A})$ interacting with a charged particle provide a realization of $(\hat{\Phi},\hat{\mathbf{A}})$ in nature: In Gaussian units, $\hat{\Phi}=e\Phi$ and $\hat{\mathbf{A}}=(e/c)\mathbf{A}$, where e is the charge of the particle. The gauge-transformed Schrödinger equations are physically equivalent to the original, but those obtained from them by gauge generalization are not. These well-known results provide a model for similar arguments involving the nonlinear transformations N.

In Sec. II, we demonstrate that two straightforward, physically motivated conditions precisely specify the group $\mathcal{N}$ of time-dependent, nonlinear gauge transformations introduced in Ref. 1. These assumptions are: (a) strict locality and (b) a separation condition. We observe that (5) is then ensured.

In Sec. III, we apply various subgroups of $\mathcal{N}$ to the linear Schrödinger equation (7). This leads to physically equivalent systems satisfying NLSEs, where the coefficients obey certain constraints. Then, in structural analogy to the way (8) motivates (9), we construct new, physically *inequivalent* systems by generalizing the parameters so as to break the constraints.

In Sec. IV, following this analogy, we consider the parameters as *gauge parameters*. We thus obtain a family of NLSEs through gauge generalization and gauge closure, labeled by the gauge parameters, on which the gauge group acts to establish physical equivalence classes. In this way, we derive naturally—as a unified class—equations containing the terms proposed by Kostin, Bialynicki-Birula and Mycielski, and Doebner and Goldin, with coefficients that are (in general) time dependent. The subfamily that includes the equations of Guerra and Pusterla turns out to be equivalent to linear quantum mechanics.

We believe this to offer a fundamentally new perspective, partially elucidating the hidden mathematical and physical structure behind certain nonlinear quantum time evolutions.

II. CONDITIONS ON NONLINEAR GAUGE TRANSFORMATIONS

A. Locality

We have from Eq. (2) that

$$N[\psi](\mathbf{x},t)=\exp[iG_{\psi}(\mathbf{x},t)]|\psi(\mathbf{x},t)|, \tag{10}$$

where G_ψ is a real-valued function of $\mathbf{x}$ and t depending on ψ. It is apparent that G_ψ must be further restricted if for instance we hope to ensure (5) for all Borel subsets B of $\mathbf{R}^3$. Suppose the value of $G_\psi(\mathbf{x},t)$ at $\mathbf{x}=\mathbf{x}_1$ depends nontrivially on values of $\psi(\mathbf{x},t)$ for $\mathbf{x}\neq\mathbf{x}_1$ and the evolution equation is local. Then we will be unable to satisfy (4) and (5) in the general case of a region B where $\mathbf{x}_1\in B$ but $\mathbf{x}\notin B$. Therefore let us assume N to be a *local* transformation, in analogy with the linear gauge transformations $\mathbf{U}_\theta$. This is taken here in the strict sense that the value of $N[\psi]$ at $(\mathbf{x},t)$ should depend only on $\mathbf{x}$, t, and the value of $\psi(\mathbf{x},t)$—not on any other space or time points, and not on derivatives of ψ. Then we must have

$$\psi'(\mathbf{x},t)=N_F[\psi](\mathbf{x},t)=\exp[iF(\psi(\mathbf{x},t),\mathbf{x},t)]|\psi(\mathbf{x},t)|, \tag{11}$$

where F is a real-valued function (defined up to integer multiples of 2π) of the three variables whose values are provided by $\psi(\mathbf{x},t)$, $\mathbf{x}$, and t. The possible dependence of F on the value of $\psi(\mathbf{x},t)$ allows nonlinearity in N_F. With $R(\mathbf{x},t)=|\psi(\mathbf{x},t)|$ and $S(\mathbf{x},t)=\arg\psi(\mathbf{x},t)$, we can consider F to be a function of the real variables R, S, $\mathbf{x}$, and t, relaxing for now the requirement that F take the same value at S and $S+2\pi n$.

Note that a weaker assumption, in which F is permitted to depend on finitely many derivatives of ψ at $(\mathbf{x},t)$, may still be compatible with (5). We make a stricter assumption here, which limits the resulting time-evolution equations to second order.

B. A separation condition

We consider now systems of n particles described by normalized states in $\mathcal{H}^{(n)}=L^2(\mathbf{R}^{3n},d^{3n}x)$. For simplicity, take each individual particle to evolve under the same time-evolution operator $T^{(1)}$. We suppose a hierarchy of time-evolutions $T^{(n)}$ of n-particle states, fulfilling the *separation condition*. For linear time evolutions this condition requires that product states $\psi^{(n)}=\psi_1\otimes\ldots\otimes\psi_n$, $\|\psi_j\|=1$, $j=1,\ldots,n$, evolve into product states:

$$T^{(n)}[\psi^{(n)}]=T^{(1)}[\psi_1]\otimes T^{(1)}[\psi_2]\otimes\cdots\otimes T^{(1)}[\psi_n]. \tag{12}$$

It ensures that in the absence of interaction terms, initially uncorrelated subsystems remain uncorrelated, and $T^{(n)}$ is extended (by linearity) from product states to all of $\mathscr{H}^{(n)}$.

It is physically plausible to assume (12) for nonlinear time evolutions $T^{(n)}$ as well.[3,17] Then nonlinear gauge transformations $N_F^{(n)}$ should respect this condition. Here, the states $\psi_j'(\mathbf{x},t) = N_F[\psi_j]$ in (11) are governed by a nonlinear time evolution $T^{(1)'}$, and our separation condition becomes for product states $\psi^{(n)'} = \psi_1' \otimes \ldots \otimes \psi_n'$, $\|\psi_j'\| = 1$, $j = 1,\ldots,n$,

$$T^{(n)'}[\psi^{(n)'}] = T^{(1)'}[\psi_1'] \otimes \cdots \otimes T^{(1)'}[\psi_n']. \tag{13}$$

We thus want a nonlinear gauge transformation $N_F^{(n)}$ acting on the unit sphere in $\mathscr{H}^{(n)}$, with $\|N_F^{(n)}[\psi]\| = 1$, so that on the product states $\psi^{(n)}$

$$N_F^{(n)}[\psi] = N_F[\psi_1] \otimes \cdots \otimes N_F[\psi_n]. \tag{14}$$

Unitary gauge transformations $\mathbf{U}^{(n)}$ in $\mathscr{H}^{(n)}$ may be written as

$$(\mathbf{U}\psi^{(n)})(\mathbf{x}_1,\ldots,\mathbf{x}_n,t) = \exp[i\,\theta_n(\mathbf{x}_1,\ldots,\mathbf{x}_n,t)]\,\psi(\mathbf{x}_1,\ldots,\mathbf{x}_n,t). \tag{15}$$

On product states, using (6), we want

$$\mathbf{U}^{(n)}\psi^{(n)} = (\mathbf{U}_\theta\psi_1) \otimes (\mathbf{U}_\theta\psi_2) \otimes \cdots \otimes (\mathbf{U}_\theta\psi_n), \tag{16}$$

so that

$$\theta_n(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,t) = \sum_{j=1}^{n} \theta(\mathbf{x}_j,t). \tag{17}$$

And of course, for this case, the operators are linear and can be extended by linearity from product states to the whole Hilbert space. But $N_F^{(n)}$ in Eq. (14) is nonlinear, so we cannot extend it uniquely to $\mathscr{H}^{(n)}$.

The apparently weak condition (14) leads nevertheless to a sharp restriction on F. To see this it is sufficient to discuss the case $n=2$. With $R_j = |\psi_j|$, $S_j = \arg\psi_j$, $j = 1,2$, Eqs. (11) and (14) imply that

$$F(R_1,S_1,\mathbf{x}_1,t) + F(R_2,S_2,\mathbf{x}_2,t) \tag{18}$$

depends only on the product $R = R_1R_2$ and the sum $S = S_1 + S_2$. Thus

$$F(R_1,S_1,\mathbf{x}_1,t) + F(R_2,S_2,\mathbf{x}_2,t) = F(R,S,\mathbf{x}_1,t) + F(1,0,\mathbf{x}_2,t), \tag{19}$$

for all $R,S,\mathbf{x}_1,\mathbf{x}_2,t$, whence $F(R_2,S_2,\mathbf{x}_2,t) - F(1,0,\mathbf{x}_2,t)$ must be independent of $\mathbf{x}_2$. Setting $F(1,0,\mathbf{x},t) = \theta(\mathbf{x},t)$ and $L(R,S,t) = F(R,S,\mathbf{x},t) - \theta(\mathbf{x},t)$, we have the functional equation

$$L(R_1,S_1,t) + L(R_2,S_2,t) = L(R_1R_2,S_1+S_2,t). \tag{20}$$

The smooth solutions of (20) are given by $L(R,S,t) = \gamma(t)\ln R + \Lambda(t)S$, where γ,Λ are real functions of t. Nondegeneracy of the transformation requires $\Lambda(t) \neq 0$. Finally, we have

$$F(R,S,\mathbf{x},t) = \gamma(t)\ln R + \Lambda(t)S + \theta(\mathbf{x},t). \tag{21}$$

The above argument is similar to the way in which generalized homogeneity of the time evolution is deduced from the separation property.[17]

Note that our separation condition is *weak* in the sense that for nonlinear $T^{(n)}$ and $N_F^{(n)}$ it is only defined on product states; for nonproduct (entangled) initial states, noninteracting subsystems may yet acquire new correlations. The nonlocal effects in some nonlinear evolution equations can be traced back to this fact.[18] A *strong* version of the separation condition, more adapted to the physical situation and valid for general states, can be formulated along the lines given in Ref. 18.

C. The result

In short, the locality and the separation condition required on n-particle product states boil down the transformations N_F for single particle states to those labeled by two real functions γ and Λ of time, with Λ nonvanishing, and a real function θ of space and time:

$$(N_{(\gamma,\Lambda,\theta)}[\psi])(\mathbf{x},t)=|\psi(\mathbf{x},t)|\exp[i(\gamma(t)\ln|\psi(\mathbf{x},t)|+\Lambda(t)\arg\,\psi(\mathbf{x},t)+\theta(\mathbf{x},t))]. \tag{22}$$

The set $N_{(\gamma,\Lambda,\theta)}$ forms a group $\mathscr{G}$, with multiplication law

$$N_{(\gamma',\Lambda',\theta')}\circ N_{(\gamma,\Lambda,\theta)}=N_{(\gamma'+\Lambda'\gamma,\Lambda'\Lambda,\theta'+\Lambda'\theta)}\,. \tag{23}$$

This can be expressed in terms of 3×3 matrices,

$$N_{(\gamma,\Lambda,\theta)}\simeq\begin{pmatrix}1&0&0\\ \theta&\Lambda&0\\ \gamma&0&\Lambda\end{pmatrix} \tag{24}$$

with entries $\Lambda=\Lambda(t)$, $\gamma=\gamma(t)$, and $\theta=\theta(\mathbf{x},t)$ taken from the corresponding function spaces. We thus have here a group $\mathscr{G}$ of nonlinear gauge transformations, strictly local and separating on n-particle product states, labeled by time-dependent parameters γ and Λ together with a function $\theta(\mathbf{x},t)$. The group is a semidirect product of the group of gauge transformations of the second kind $\mathscr{U}_{\mathrm{loc}}=\{\mathbf{U}_\theta\}$ and the group $\mathscr{N}$, mentioned in Sec. I, of ''pure nonlinear'' gauge transformations (where $\theta\equiv0$):

$$\mathscr{G}=\mathscr{N}\otimes_s\mathscr{U}_{\mathrm{loc}}\,. \tag{25}$$

$\mathscr{G}$ can be viewed as a *nonlinear generalization* of $\mathscr{U}_{\mathrm{loc}}$, i.e., as the group of ''gauge transformations of the *third kind*.'' [1]

The transformations $N_{(\gamma,\Lambda,\theta)}$ are not uniquely defined on the Hilbert space. If we restrict the range of Λ to the integers, $\Lambda(t)\in\mathbf{Z}$, then $N_{(\gamma,\Lambda,\theta)}$ is well defined. Then if Λ is a continuous function of time, Λ has to be a constant; $N_{(\gamma,\Lambda,\theta)}$ is invertible with this restriction only for $\Lambda=\pm1$. $\Lambda=-1$ corresponds to complex conjugation: $N_{(0,-1,0)}\psi=\overline{\psi}$. $N_{(\gamma,1,\theta)}$ is strongly continuous,[15] and the set of these transformations is an Abelian subgroup of $\mathscr{G}$,

$$\mathscr{G}\supset\mathscr{G}_0=\mathscr{N}_0\otimes\mathscr{U}_{\mathrm{loc}}, \tag{26}$$

where $\mathscr{N}_0:=\{N_\gamma:=N_{(\gamma,\Lambda\equiv1,\theta\equiv0)}\}$.

For noninteger Λ, $N_{(\gamma,\Lambda,\theta)}$ may be specified uniquely on certain domains in the Hilbert space, e.g., by imposing continuity of the phase of $N[\psi]$ on a domain of nonvanishing functions ψ. However, such a domain is not needed explicitly for our further considerations.

D. A generalization

Because of the difficulties with the separation condition mentioned above, a more general group structure is also of interest. This can be obtained, without assuming separation, by making a physically motivated, weaker assumption: an intertwining relation that follows from requiring compatibility with linear gauge transformations.

The group of linear gauge transformations $\mathscr{U}_{\mathrm{loc}}$ is commutative, but this need not be the case for the set $\{N_F\}\supset\mathscr{U}_{\mathrm{loc}}$. In particular, $\mathbf{U}_\theta$ might not commute with N_F. We explore the condition that N_F be consistent with the usual notion of physical equivalence under gauge transformations of the second kind. That is, the result of applying N_F to a gauge-transformed theory with wave functions $\mathbf{U}_\theta\psi$ should be expressible as a transform by $\mathbf{U}_{\theta'}$ of the theory with wave functions $N_F[\psi]$, where, in general, $\theta'(\mathbf{x},t)\neq\theta(\mathbf{x},t)$. Thus we require an *intertwining relation*

$$N_F[\mathbf{U}_\theta\psi]=\mathbf{U}_{\theta'}N_F[\psi]. \tag{27}$$

Here the function $\theta'(\mathbf{x},t)$ depends on both of the functions F and θ.

Then Eq. (27) implies the functional equation

$$\exp\, i[F(R,S+\theta;\mathbf{x},t)]=\exp\, i[\theta'(\mathbf{x},t)+F(R,S;\mathbf{x},t)], \tag{28}$$

valid for each R, S, $\mathbf{x}$, and t. It is straightforward to show that smooth solutions F of (28) take the form $F(R,S;\mathbf{x},t)=k(R,\mathbf{x},t)+\lambda(\mathbf{x},t)S$, where k and λ are real-valued functions of the indicated variables. Nondegeneracy of the transformation requires $\lambda(\mathbf{x},t)\neq 0$ for all $\mathbf{x},t$. Thus N_F is parametrized by k and λ, and given by

$$N_{(k,\lambda)}[\psi](\mathbf{x},t)=\exp\, i[k(|\psi(\mathbf{x},t)|,\mathbf{x},t)+\lambda(\mathbf{x},t)\arg\,\psi(\mathbf{x},t)]|\psi(\mathbf{x},t)|. \tag{29}$$

One easily checks that (2), (11), and (27) are fulfilled, with

$$\theta'(\mathbf{x},t)=\lambda(\mathbf{x},t)\theta(\mathbf{x},t). \tag{30}$$

The set $\{N_{(k,\lambda)};\lambda(\mathbf{x},t)\neq 0\}$ is a noncommutative, infinite dimensional group $\tilde{\mathscr{G}}$ with multiplication law

$$N_{(k,\lambda)}\circ N_{(k',\lambda')}=N_{(k+k'\lambda,\lambda\lambda')}. \tag{31}$$

$N_{(0,1)}$ acts as the identity on ψ, and $N_{(-k/\lambda,1/\lambda)}$ is the (formal) inverse of $N_{(k,\lambda)}$. The group law may be expressed as multiplication of 2×2 matrices

$$N_{(k,\lambda)}\simeq\begin{pmatrix}1 & 0\\ k & \lambda\end{pmatrix} \tag{32}$$

with entries $k(|\psi|,\mathbf{x},t)$ and $\lambda(\mathbf{x},t)$ taken from suitable function spaces. Such matrices span a linear representation Aff(1) of the one-dimensional affine group.

The nonlinear transformations $N_{(\gamma,\Lambda,\theta)}$ are special cases of $N_{(k,\lambda)}$, i.e., the separation condition restricts k and λ to the form

$$k(|\psi|,\mathbf{x},t)=\gamma(t)\ln|\psi|+\theta(\mathbf{x},t), \tag{33}$$

$$\lambda(\mathbf{x},t)=\Lambda(t); \tag{34}$$

and $\mathscr{G}$ is a subgroup of $\tilde{\mathscr{G}}$.

III. NONLINEAR QUANTUM-MECHANICAL EVOLUTION EQUATIONS FROM GAUGE GENERALIZATION

A. Linearizable NLSEs

In accordance with the discussion in Sec. I, we are now interested in the evolution equation of

$$\psi'(\mathbf{x},t)=N_{(\gamma,\Lambda,\theta)}[\psi](\mathbf{x},t), \tag{35}$$

when $\psi(\mathbf{x},t)$ is a solution of a linear Schrödinger equation

$$i\partial_t\psi=(\nu_1\Delta+\mu_0 V)\psi. \tag{36}$$

Let us regard (36) as belonging to a parametrized family $\mathscr{F}_0(\nu_1,\mu_0)$, $\nu_1\neq 0$, depending on the two real parameters ν_1,μ_0; in Eq. (7), $\nu_1=-\hbar/2m$ and $\mu_0=1/\hbar$.

Due to (27) linear gauge transformations can be treated independently, and we shall here restrict ourselves to the case $\theta\equiv 0$. Applying the group $\mathscr{N}$ to $\mathscr{F}_0$, we obtain a family $\overline{\mathscr{F}_0}$ of NLSEs,

$$i\partial_t\psi'=(\nu_1'(t)\Delta+\mu_0'(t)V+F_{\mathrm{DG}}^{(0)}[\psi']+F_{\mathrm{BM}}[\psi']+F_{\mathrm{K}}[\psi'])\psi', \tag{37}$$

where the subscripts DG, BM, and K, refer to terms of types considered respectively by Doebner and Goldin, Bialynicki-Birula and Mycielski, and Kostin; specifically:

$$F_{\mathrm{DG}}^{(0)}[\psi']=\mu_1'(t)\left(\nabla\left(\mathrm{Im}\left\{\frac{\nabla\psi'}{\psi'}\right\}\right)+\frac{i}{2}\,\frac{\Delta|\psi'|^2}{|\psi'|^2}\right)+2\kappa'(t)\,\frac{\Delta|\psi'|}{|\psi'|}, \tag{38}$$

$$F_{\mathrm{BM}}[\psi']=\alpha_1'(t)\log|\psi'|^2, \tag{39}$$

$$F_{\mathrm{K}}[\psi']=\alpha_2'(t)\arg\,\psi'. \tag{40}$$

The coefficients ν_1', μ_0', μ_1', κ', α_1', and α_2' are constrained, and depend on both ν_1,μ_0, and on $\Lambda(t),\gamma(t)$:

$$\nu_1'(t)=\frac{1}{\Lambda(t)}\nu_1,\quad \mu_0'(t)=\Lambda(t)\mu_0,\quad \mu_1'(t)=-\frac{\gamma(t)}{\Lambda(t)}\nu_1,\quad \kappa'(t)=\frac{\gamma(t)^2+\Lambda(t)^2-1}{2\Lambda(t)}\nu_1,$$
$$\alpha_1'(t)=\gamma(t)\frac{\dot\Lambda(t)}{2\Lambda(t)}-\frac{1}{2}\,\dot\gamma(t),\quad \alpha_2'(t)=-\frac{\dot\Lambda(t)}{\Lambda(t)}. \tag{41}$$

This family $\overline{\mathscr{F}_0}$ is closed under $\mathscr{N}$, i.e., it is the *gauge closure* of $\mathscr{F}_0(\nu_1,\mu_0)$ under the action of the group $\mathscr{N}$. It is, up to questions of domain mentioned above, linearizable. It depends on the independent quantities ν_1, μ_0, $\gamma(t)$ and $\Lambda(t)$. One could also write $\overline{\mathscr{F}_0}$ as $\mathscr{F}(\nu_1,\mu_0,\mu_1,\kappa,\alpha_1,\alpha_2)$ labeled by time-dependent coefficients that are constrained.

Note that if Λ and γ are independent of t, the coefficients are time independent, and $\alpha_1'=\alpha_2'=0$.

If we restrict $\mathscr{N}$ to the subgroup $\mathscr{N}_0$, then starting with $\mathscr{F}_0(\nu_1,\mu_0)$, we obtain a family $\overline{\mathscr{F}_0}^0$ closed under $\mathscr{N}_0$ and contained in $\overline{\mathscr{F}_0}$; here the indexed bar denotes the closure with respect to $\mathscr{N}_0$. The elements in $\overline{\mathscr{F}_0}^0$ are by construction linearizable NLSEs. The parameters are

$$\nu_1'=\nu_1,\quad \mu_0'=\mu_0,\quad \mu_1'(t)=\gamma(t)\nu_1,\quad \kappa'(t)=\frac{\gamma(t)^2}{2}\nu_1,$$
$$\alpha_1'(t)=-\tfrac{1}{2}\dot\gamma(t),\quad \alpha_2'(t)=0. \tag{42}$$

Now the term F_{K} disappears, everything is well-defined, and ν_1' and μ_0' are time-independent invariants. Strictly speaking, these NLSEs are *defined* using the continuity and invertibility of $N_{(\gamma,1,0)}$.

For later purposes we mention that $F_{\mathrm{DG}}^{(0)}$ decomposes into independent nonlinear real functionals R with the following properties: $R[\psi]$ is Euclidean invariant, complex homogeneous of degree zero and a rational function of $\psi,\overline{\psi}$ with derivatives not higher than second order in the numerator only. There exist five functionals of this type (see Ref. 6):

$$R_1[\psi]=\frac{\nabla\cdot\mathbf{J}}{\rho},\quad R_2[\psi]=\frac{\Delta\rho}{\rho},$$
$$R_3[\psi]=\frac{\mathbf{J}^2}{\rho^2},\quad R_4[\psi]=\frac{\mathbf{J}\cdot\nabla\rho}{\rho^2},\quad R_5[\psi]=\frac{(\nabla\rho)^2}{\rho^2}, \tag{43}$$

where $\rho=\bar{\psi}\psi$ and $\mathbf{J}=(1/2i)(\bar{\psi}\nabla\psi-(\nabla\bar{\psi})\psi)$ are the probability density and current corresponding to ψ. With this notation $F_{\mathrm{DG}}^{(0)}$ in (38) is a complex linear combination:

$$F_{\mathrm{DG}}^{(0)}[\psi]=\mu_1(t)(R_1[\psi]-R_4[\psi])+i\nu_2(t)R_2[\psi]+\kappa(t)(R_2[\psi]-\tfrac{1}{2}R_5[\psi]), \tag{44}$$

with

$$\nu_2(t)=-\tfrac{1}{2}\mu_1(t). \tag{45}$$

The term $R_3[\psi]$ will appear in Sec. III B.

B. Generalizing linearizable NLSEs; gauge parameters

The nonlinear gauge transformations $N_{(\gamma,\Lambda)}$ generate special linearizable NLSEs, i.e., nonlinear partial differential equations with constrained coefficients, physically equivalent to linear Schrödinger equations. Hence the situation is similar to the case of gauge transformations $\mathbf{U}_\theta$ in Sec. I. It is possible to construct generically through *gauge generalizations* and *gauge closures* a sequence of new families of evolution equations physically *in*equivalent to the linear Schrödinger equation. We obtain the sequence of these families in three steps:

Step 1: We break the constraints (41) in $\overline{\mathscr{F}_0}$ (gauge generalization), i.e., we take the six constrained coefficients $\nu_1',\mu_0',\mu_1',\kappa',\alpha_1',\alpha_2'$ as independent functions of time. Thus we obtain a family $\mathscr{F}_1(\nu_1,\mu_0,\mu_1,\kappa,\alpha_1,\alpha_2)$ with six independent parameters. The gauge transformations $\mathscr{N}$ are automorphisms of this family. That is, $\overline{\mathscr{F}_1}=\mathscr{F}_1$; the family is gauge closed. In the notation of Ref. 1, $\kappa=\mu_2-\frac{1}{2}\nu_1$.

Step 2: We break the constraint (45) for F_{DG}^0 in $\mathscr{F}_1$ (gauge generalization),

$$F_{\mathrm{DG}}^{(1)}[\psi]=i\nu_2(t)R_2[\psi]+\mu_1(t)(R_1[\psi]-R_4[\psi])+\kappa(t)(R_2[\psi]-\tfrac{1}{2}R_5[\psi]), \tag{46}$$

and obtain a seven-parameter family $\mathscr{F}_2(\nu_1,\nu_2,\mu_0,\mu_1,\kappa,\alpha_1,\alpha_2)$ of NLSEs (37), with $F_{\mathrm{DG}}^{(1)}$ replacing $F_{\mathrm{DG}}^{(0)}$.

The action of the group $\mathscr{N}$, however, does not leave this family invariant. The gauge closure $\overline{\mathscr{F}_2}$ of $\mathscr{F}_2$ consists of NLSEs (37) with

$$F_{\mathrm{DG}}^{(2)}[\psi]=i\nu_2'(t)R_2[\psi]+\mu_1'(t)(R_1[\psi]-R_4[\psi])+\kappa'(t)R_2[\psi]+\xi'(t)R_5[\psi] \tag{47}$$

in place of $F_{\mathrm{DG}}^{(1)}$. Now there are eight coefficients which already are unconstrained. Thus we write $\overline{\mathscr{F}_2}$ as a family $\mathscr{F}_3(\nu_1,\nu_2,\mu_0,\mu_1,\kappa,\xi,\alpha_1,\alpha_2)$ with eight time-dependent parameters, that is invariant by construction under $\mathscr{N}$, i.e., $\overline{\mathscr{F}_2}=\mathscr{F}_3=\overline{\mathscr{F}_3}$. In the notation of Ref. 1, $\xi=\mu_5+\frac{1}{4}\nu_1$. The explicit formula for these coefficients is given by Eq. (50) below, with $\mu_3=-\nu_1$ and $\xi=-\frac{1}{2}\kappa$.

Step 3: We write $\Delta\psi$ as a complex linear combination of $R_j[\psi]\psi$,

$$\Delta\psi=(iR_1[\psi]+\tfrac{1}{2}R_2[\psi]-R_3[\psi]-\tfrac{1}{4}R_5[\psi])\psi, \tag{48}$$

insert into (37) and obtain an additional term $(\mu_3+\nu_1)R_3[\psi]$ in F_{DG}, and a constraint $\mu_3(t)=-\nu_1(t)$.

We break this constraint, and obtain from $\mathscr{F}_3$ a family $\mathscr{F}_4(\nu_1,\nu_2,\mu_0,\mu_1,\kappa,\mu_3,\xi,\alpha_1,\alpha_2)$ depending on nine time-dependent parameters.

The closure $\overline{\mathscr{F}_4}$ is larger than $\mathscr{F}_4$ and contains all NLSEs (37) with

$$F_{\mathrm{DG}}[\psi]=i\nu_2'R_2[\psi]+\mu_1'R_1[\psi]+\kappa' R_2[\psi]+(\mu_3'+\nu_1')R_3[\psi]+\mu_4'R_4[\psi]+\xi' R_5[\psi], \quad (49)$$

where the time-dependent coefficients are given by

$$\nu_1'=\frac{\nu_1}{\Lambda},\quad \nu_2'=-\frac{\gamma}{2\Lambda}\nu_1+\nu_2,\quad \mu_0'=\Lambda\mu_0,\quad \mu_1'=-\frac{\gamma}{\Lambda}\nu_1+\mu_1,$$

$$\kappa'=\frac{\gamma^2+\Lambda^2-1}{2\Lambda}\nu_1-\gamma\nu_2-\frac{\gamma}{2}\mu_1+\Lambda\kappa,\quad \mu_3'=\frac{1}{\Lambda}\mu_3,\quad \mu_4'=\frac{\gamma}{\Lambda}\nu_1-\mu_1-\frac{\gamma}{\Lambda}\mu_3,$$

$$\xi'=\frac{1-\gamma^2-\Lambda^2}{4\Lambda}\nu_1+\frac{\gamma}{2}\mu_1+\frac{\gamma^2}{4\Lambda}\mu_3+\Lambda\xi, \quad (50)$$

$$\alpha_1'=\Lambda\alpha_1-\frac{\gamma}{2}\alpha_2+\gamma\frac{\dot\Lambda}{2\Lambda}-\frac{1}{2}\dot\gamma,\quad \alpha_2'=\alpha_2-\frac{\dot\Lambda}{\Lambda}.$$

These coefficients are actually independent, so that $\overline{\mathscr{F}_4}$ is a ten-parameter family. For a more symmetrical notation, we now go over to using $\mu_2=\kappa+\frac{1}{2}\nu_1$ and $\mu_5=\xi-\frac{1}{4}\nu_1$, denoting the family by $\mathscr{F}_5(\nu_1,\nu_2,\mu_0,\ldots,\mu_5,\alpha_1,\alpha_2)$:

$$i\partial_t\psi=i\sum_{j=1}^{2}\nu_jR_j[\psi]\psi+\sum_{k=1}^{5}\mu_kR_k[\psi]\psi+\mu_0V\psi+\alpha_1\log|\psi|^2\psi+\alpha_2(\arg\psi)\psi, \quad (51)$$

or in a form which exhibits the linear part separately, with Laplacian Δ,

$$i\partial_t\psi=(\nu_1\Delta+\mu_0V)\psi+i\nu_2R_2[\psi]\psi+\mu_1R_1[\psi]\psi+(\mu_2-\tfrac{1}{2}\nu_1)R_2[\psi]\psi+(\mu_3+\nu_1)R_3[\psi]\psi$$
$$+\mu_4R_4[\psi]\psi+(\mu_5+\tfrac{1}{4}\nu_1)R_5[\psi]\psi+\alpha_1\log|\psi|^2\psi+\alpha_2(\arg\psi)\psi. \quad (52)$$

$\mathscr{F}_5$ is invariant under the action of the group $\mathscr{N}$, i.e., $\overline{\mathscr{F}_5}=\mathscr{F}_5$.

Starting with the linear family $\mathscr{F}_0$, through iterated gauge generalizations and gauge closures with respect to the pure nonlinear gauge group $\mathscr{N}$, we have thus obtained a sequence

$$\mathscr{F}_0\subset\overline{\mathscr{F}_0}\subset\mathscr{F}_1=\overline{\mathscr{F}_1}\subset\mathscr{F}_2\subset\overline{\mathscr{F}_2}=\mathscr{F}_3=\overline{\mathscr{F}_3}\subset\mathscr{F}_4\subset\overline{\mathscr{F}_4}=\mathscr{F}_5 \quad (53)$$

of families of nonlinear Schrödinger equations.

The same procedure can be followed for the restricted gauge group $\mathscr{N}_0$. It turns out that there is an analogous sequence of families $\mathscr{R}_j$ of NLSEs:

$$\mathscr{F}_0\equiv\mathscr{R}_0\subset\overline{\mathscr{R}_0}^0\subset\mathscr{R}_1=\overline{\mathscr{R}_1}^0\subset\mathscr{R}_2\subset\overline{\mathscr{R}_2}^0=\mathscr{R}_3=\overline{\mathscr{R}_3}^0\subset\mathscr{R}_4\subset\overline{\mathscr{R}_4}^0=\mathscr{R}_5. \quad (54)$$

The families $\mathscr{R}_j$ are subsets of the $\mathscr{F}_j$:

$$\mathscr{R}_j=\mathscr{F}_j\restriction_{\nu_1(t)=\nu_1,\mu_0(t)=\mu_0,\mu_3(t)=-\nu_1,\alpha_2(t)=0}. \quad (55)$$

The only type of term that is not obtained in these families is the term F_{K} (which is technically not well defined). Note furthermore that here the parameters of the original linear family $\mathscr{R}_0\equiv\mathscr{F}_0$ remain invariant, $\nu_1'=\nu_1$ and $\mu_0'=\mu_0$.

IV. DISCUSSION OF THE GAUGE-GENERALIZED NLSE

A. Gauge-invariant parameters, Ehrenfest relations, and Galilei invariance

The group $\mathscr{N}$ transforms the family $\mathscr{F}_5$ into itself. In fact, $N_{(\gamma,\Lambda)}$ acts (for all t) linearly on the eight gauge parameters $\boldsymbol{\nu} = (\nu_1,\ldots,_5)$,

$$\begin{pmatrix} \nu_1' \\ \nu_2' \\ \mu_0' \\ \mu_1' \\ \mu_2' \\ \mu_3' \\ \mu_4' \\ \mu_5' \end{pmatrix} = \begin{pmatrix} \frac{1}{\Lambda} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{\gamma}{2\Lambda} & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \Lambda & 0 & 0 & 0 & 0 & 0 \\ -\frac{\gamma}{\Lambda} & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{\gamma^2}{2\Lambda} & -\gamma & 0 & -\frac{\gamma}{2} & \Lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\Lambda} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{\gamma}{\Lambda} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\gamma^2}{4\Lambda} & -\frac{\gamma}{2} & \Lambda \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \mu_0 \\ \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix}. \tag{56}$$

One can show that the orbits of $\mathscr{N}$, for a fixed time t, are two dimensional on the space $\dot{\mathbf{R}}_t^8 := \{\boldsymbol{\nu} \in \mathbf{R}_t^8 | \nu_1 \neq 0\}$, and foliate $\dot{\mathbf{R}}_t^8$ in two-dimensional leaves. Hence there exist (in general, at least locally; but here in fact globally) six functionally independent parameters $\iota_0,\ldots,\iota_5$ invariant under the action of $\mathscr{N}$,[8,9]

$$\begin{gathered} \iota_0 = \nu_1\mu_0, \quad \iota_1 = \nu_1\mu_2 - \nu_2\mu_1, \quad \iota_2 = \mu_1 - 2\nu_2, \quad \iota_3 = 1 + \mu_3/\nu_1, \\ \iota_4 = \mu_4 - \mu_1\mu_3/\nu_1, \quad \iota_5 = \nu_1(\mu_2 + 2\mu_5) - \nu_2(\mu_1 + 2\mu_4) + 2\nu_2^2\mu_3/\nu_1. \end{gathered} \tag{57}$$

On the remaining two parameters $\boldsymbol{\alpha} = (\alpha_1,\alpha_2)$, the transformation $N_{(\gamma,\Lambda)}$ acts as an affine transformation,

$$\begin{pmatrix} \alpha_1' \\ \alpha_2' \end{pmatrix} = \begin{pmatrix} \Lambda & -\frac{\gamma}{2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \frac{1}{2}\left(\gamma\frac{\dot{\Lambda}}{\Lambda} - \dot{\gamma}\right) \\ -\frac{\dot{\Lambda}}{\Lambda} \end{pmatrix}. \tag{58}$$

Thus there are two further independent parameters invariant under the action of $\mathscr{N}$ on the *control space* $\dot{\mathbf{R}}_t^{10}$ spanned by $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$,

$$\iota_6 = \nu_1\alpha_1 - \nu_2\alpha_2 + \nu_2\frac{\dot{\nu}_1}{\nu_1} - \dot{\nu}_2, \quad \iota_7 = \alpha_2 - \frac{\dot{\nu}_1}{\nu_1}, \tag{59}$$

generalizing the result in Refs. 8 and 9 for the family of NLSEs derived in Refs. 5 and 6. We call $\boldsymbol{\iota} = (\iota_0,\ldots,\iota_7)$ *gauge-invariant parameters*. They are important for interpreting $\mathscr{F}_5$ and its subfamilies; for details, we refer to Ref. 1, where gauge-invariant parameters have been discussed in a slightly different context.

TABLE I. Classification of subfamilies of $\mathscr{F}_5$ using gauge invariants.

	ι_0	ι_1	ι_2	ι_3	ι_4	ι_5	ι_6	ι_7
$\mathscr{F}_0$	×	×	0	0	0	0	0	0
$\mathscr{F}_1$	×	×	0	0	0	0	×	×
$\mathscr{F}_3$	×	×	×	0	0	×	×	×
$\mathscr{F}_5$	×	×	×	×	×	×	×	×

The subfamilies $\mathscr{F}_j$ and $\mathscr{R}_j$, that are closed under the gauge groups $\mathscr{N}$ and $\mathscr{N}_0$ respectively, can now be characterized in terms of the vanishing of gauge-invariant parameters. Such a characterization is given in Tables I and II, respectively. Note that ν_1 and μ_3 are themselves gauge-invariant parameters of the subfamilies $\mathscr{R}_j$.

Some of these families show interesting behavior. In the family $\mathscr{F}_1$ consider the time dependence of the expectation values $\langle \mathbf{x}\rangle_{\psi(t)} = \int_{\mathbf{R}^3} \mathbf{x}\rho_t(\mathbf{x})d^3x = \int_{\mathbf{R}^3} \mathbf{x}\overline{\psi(\mathbf{x},t)}\psi(\mathbf{x},t)d^3x$. Then

$$\frac{d}{dt}\langle \mathbf{x}\rangle_{\psi(t)} = -2\nu_1\langle -i\nabla\rangle_{\psi(t)}, \tag{60}$$

$$\frac{d^2}{dt^2}\langle \mathbf{x}\rangle_{\psi(t)} = -2\iota_0\langle - \ V\rangle_{\psi(t)} + \iota_7\frac{d}{dt}\langle \mathbf{x}\rangle_{\psi(t)}, \tag{61}$$

i.e., we have the analog of the first and second Ehrenfest relations for $\mathscr{F}_1$. The center of a nonstationary solution behaves like a classical system under a conservative force and a frictional force proportional to the velocity. For the linearizable subfamily $\overline{\mathscr{F}_0} \subset \mathscr{F}_1$, the frictional term disappears ($\iota_7 = 0$). This is plausible: A linear or nonlinear quantum-mechanical evolution equation and its $\mathscr{N}$ transform describe physically equivalent systems.[1]

The first Ehrenfest relation (60) holds for all members of $\mathscr{F}_5$. This shows that the physical systems described by $\mathscr{F}_5$ have something in common. For $\mathscr{F}_2,\dots,\mathscr{F}_5$ there are additional terms in the second Ehrenfest relation (61), which are connected with the quantum-mechanical diffusion current.[5,6]

The free linear SE ($V\equiv 0$) is invariant under the centrally extended Galilei group $G_e(3)$ including time translations. Consider the $G_e(3)$ invariance of $\mathscr{F}_5$ and its subfamilies. $\mathscr{F}_5$ is invariant under $T(t)$ if the gauge-invariant parameters $\boldsymbol{\iota}$ are time independent. If in addition $\iota_3 = \iota_4 = \iota_7 = 0$, the equations are invariant under $G_e(3)$. The generator of time translations is represented via a nonlinear operator H_{nl}, $i\partial_t\psi = H_{\mathrm{nl}}[\psi]$, as in Eqs. (51) and (52), while all other generators of $G_e(3)$ are as usual represented linearly. Hence, one has a nonlinear representation of $G_e(3)$ (see also Refs. 8, 19, and 20).

B. Gauge-generalized NLSE as a unification

Now we are ready to understand the connection between various proposals for nonlinear terms to be added to the linear Schrödinger equation. Such terms have often been chosen in a physically guided, but *ad hoc* way. Some proposed terms have been based directly on fundamental considerations. Our attempt is of the latter type. Its foundation, the physical equivalence of theories and

TABLE II. Classification of subfamilies of $\mathscr{R}_5$ using gauge invariants.

	ι_1	ι_2	ι_3	ι_4	ι_5	ι_6	ν_1	μ_0
$\mathscr{R}_0$	×	0	0	0	0	0	×	×
$\mathscr{R}_1$	×	0	0	0	0	×	×	×
$\mathscr{R}_3$	×	×	0	0	×	×	×	×
$\mathscr{R}_5$	×	×	×	×	×	×	×	×

the resulting group of nonlinear gauge transformations (together with gauge generalization and gauge closure), reflects some of the structure of quantum mechanics. Consequently the family $\mathscr{F}_5$ exhibits a common, fundamental basis for some of the proposed NLSEs. Let us consider some of the particular nonlinearities that have been proposed.

1. Logarithmic nonlinearity

Based on the observation that all linear evolution equations for physical quantities are known to be approximations of nonlinear evolutions (except for the Schrödinger equation), Bialynicki-Birula and Mycielski[3] added a (local) nonlinear term $F(|\psi|^2)$. They used the separation property to show that F has to be logarithmic, $F(|\psi|^2) = -b \ln|\psi|^2$. Their NLSE (the BM family) is

$$i\hbar\partial_t\psi = \left(-\frac{\hbar^2}{2m}\Delta + V - b \ln|\psi|^2 \right)\psi. \tag{62}$$

This NLSE is contained in $\mathscr{F}_5$ with

$$\nu_1 = \frac{\hbar}{2m}, \quad \mu_2 = \frac{\hbar}{4m}, \quad \mu_3 = -\frac{\hbar}{2m}, \quad \mu_5 = -\frac{\hbar}{8m}, \quad \mu_0 = \frac{1}{\hbar}, \quad \alpha_1 = -\frac{b}{\hbar}, \tag{63}$$

and the other coefficients vanishing. Note that in order to obtain this logarithmic term in our gauge generalization, we had to allow for a time-dependent group parameter $\gamma = \gamma(t)$.

2. Nonlinearity proportional to the phase

One of many examples of a heuristic implementation of dissipation in quantum mechanics is the approach by Kostin.[2] Starting with a frictional term proportional to the expectation of the momentum operator in the (second) Ehrenfest relation, Kostin motivated adding a nonlinear term proportional to the phase of ψ to the linear Schrödinger equation, i.e. (with $f \in \mathbf{R}$),

$$i\hbar\partial_t\psi = \left(-\frac{\hbar^2}{2m}\Delta + V + \frac{\hbar f}{m} \arg \psi \right)\psi. \tag{64}$$

Kostin's NLSE (the K family) is contained in $\mathscr{F}_5$ with

$$\nu_1 = \frac{\hbar}{2m}, \quad \mu_2 = \frac{\hbar}{4m}, \quad \mu_3 = -\frac{\hbar}{2m}, \quad \mu_3 = -\frac{\hbar}{8m}, \quad \mu_0 = \frac{1}{\hbar}, \quad \alpha_2 = \frac{f}{m}, \tag{65}$$

and the other coefficients vanishing. To obtain this term in our approach, we had to assume that $\Lambda = \Lambda(t)$ can be a function of time. Obviously, arg ψ is not well defined; this is reflected in the problem of gauge transformations with $\Lambda \neq \pm 1$, discussed in Sec. II C.

3. Nonlinearity from diffeomorphism group representations

The approach of Doebner and Goldin[5,6] is motivated by fundamental considerations. The generic kinematical symmetry algebra $S(\mathbf{R}^3)$ on $\mathbf{R}^3$ is a semidirect sum of the Lie algebra of real smooth functions $f \in C^\infty(\mathbf{R}^3)$, and the Lie algebra of vector fields $X \in \mathrm{Vect}(\mathbf{R}^3)$, or equivalently a local current algebra on $\mathbf{R}^3$.[21–23] Vect($\mathbf{R}^3$) is the Lie algebra of a subgroup of the group of diffeomorphisms of $\mathbf{R}^3$ (diffeomorphisms trivial at infinity). The functions $f \in C^\infty(\mathbf{R}^3)$ can be interpreted physically as classical position observables and the vector fields $X \in \mathrm{Vect}(\mathbf{R}^3)$ as classical kinematical momenta. Then a quantization map $\mathscr{Q}$ represents the kinematical algebra $S(\mathbf{R}^3)$ by self-adjoint operators in the single particle Hilbert space $\mathscr{H}^{(1)}$. Under physically motivated assumptions, all such representations $\mathscr{Q}$ can be classified up to unitary equivalence by a real parameter D with the dimensionality of a diffusion coefficient $[\text{length}^2/\text{time}]$. The presence of such a family of inequivalent representations reflects the richness of Vect($\mathbf{R}^3$). The method can be generalized to any smooth manifold.[24–26]

To obtain some information about the evolution equation of ψ, local probability conservation (for pure states) is assumed,[5] or a generalized first Ehrenfest relation is postulated.[25,26] Then the time-dependent probability density and current are related through an equation of Fokker–Planck type,

$$\partial_t \rho = -\frac{\hbar}{m}\nabla\cdot\mathbf{J} + D\Delta\rho. \tag{66}$$

This restricts the evolution equation of ψ to the form

$$i\hbar\,\partial_t\psi = \left(-\frac{\hbar^2}{2m}\Delta + V\right)\psi + i\,\frac{\hbar D}{2}\,\frac{\Delta\rho}{\rho}\,\psi + R[\psi]\psi, \tag{67}$$

where $R[\psi]$ is an arbitrary real-valued (nonlinear) operator. The form of the pure imaginary functional, $\Delta\rho/\rho$, is enforced. If $R[\psi]$ is assumed to be of a similar form, i.e., if it is (i) complex homogeneous of degree zero, (ii) a rational function with derivatives of no more than second order occurring only in the numerator, and (iii) invariant under the three-dimensional Euclidean group $E(3)$, then a five-parameter family of NLSEs (the DG family) is obtained:

$$R[\psi] = \hbar D' \sum_{j=1}^{5} c_j R_j[\psi], \tag{68}$$

with the R_j as in Eq. (43). Obviously this is a special case of $\mathscr{F}_5$, where $\alpha_1 = \alpha_2 = 0$ and all gauge parameters are time-independent:

$$\begin{gathered}\nu_1 = -\frac{\hbar}{2m},\quad \nu_2 = \frac{\hbar D}{2},\quad \mu_0 = \frac{1}{\hbar},\quad \mu_1 = \hbar D' c_1,\quad \mu_2 = \hbar D' c_2 - \frac{\hbar}{4m},\\ \mu_3 = \hbar D' c_3 + \frac{\hbar}{2m},\quad \mu_4 = \hbar D' c_4,\quad \mu_5 = \hbar D' c_5 + \frac{\hbar}{8m},\quad \alpha_1 = \alpha_2 = 0.\end{gathered} \tag{69}$$

The equation proposed by Guerra and Pusterla in connection with de Broglie's double solution theory[4] is contained in this family, with $D=0$, $c_1 = c_3 = c_4 = 0$, $c_5 = -\frac{1}{2}c_2$.

V. SUMMARY

To summarize, we have taken a small step toward a nonlinear quantum theory which could be physically relevant by discussing nonlinear evolution equations derived from fundamental considerations.

Under the assumption that all measurements are positional measurements performed at different times, we derived a group of nonlinear gauge transformations $\mathscr{G}$, including the usual linear ones. Applying these transformations to a linear Schrödinger equation, we obtained nonlinear ones, and after gauge generalization and gauge closure we reached a family $\mathscr{F}_5$ of nonlinear Schrödinger equations. Certain subfamilies of $\mathscr{F}_5$ were motivated originally by different physical ideas and different mathematical structures. Thus $\mathscr{F}_5$ is a *unification* of these NLSEs: the BM family, the K family, and the DG family. It is surprising, and also satisfying, when different structures and lines of reasoning yield the same or compatible results. This is an indication that these structures have a common origin. If there is some deeper reason for this, beyond the gauge generalization process described here, we have not yet unveiled it.

Moreover, our discussion may show how to circumvent some formal arguments against nonlinear quantum theory put forth by Gisin and others;[27–29] in connection with nonlocal effects, we refer especially to Ref. 18. We have not touched on other problems of nonlinear quantum theory, such as the concept of mixed states (see Ref. 16), or discussed the physical interpretation of a (necessarily non-self-adjoint) nonlinear Hamiltonian.

ACKNOWLEDGMENTS

We would like to thank W. Lücke for fruitful discussions on the topic of this paper. G.A.G. acknowledges hospitality from the Arnold Sommerfeld Institute for Mathematical Physics, Technical University of Clausthal, and travel support from the DAAD (Deutscher Akademischer Austauschdienst) and Rutgers University. H.D.D. acknowledges the essential help of Dr. G. A. Oswald FRCP and the Coronary Care Unit, Princess Elizabeth Hospital, Guernsey and of Professor Kreuzer and Dr. Schulze, Unit 1023 ''HIS,'' University Hospital Göttingen, where an earlier draft of this paper was written.

[1] H.-D. Doebner and G. A. Goldin, Phys. Rev. A **54**, 3764 (1996).
[2] M. D. Kostin, J. Chem. Phys. **57**, 3589 (1972).
[3] I. Bialynicki-Birula and J. Mycielski, Ann. Phys. (N.Y.) **100**, 62 (1976).
[4] F. Guerra and M. Pusterla, Lett. Nuovo Cimento **34**, 351 (1982).
[5] H.-D. Doebner and G. A. Goldin, Phys. Lett. A **162**, 397 (1992).
[6] H.-D. Doebner and G. A. Goldin, J. Phys. A **27**, 1771 (1994).
[7] P. Nattermann, in *Proceedings of the XXVI Symposium on Mathematical Physics, Torun, Poland, 7–10 December, 1993* (Nicolas Copernicus University Press, Torun, 1994), p. 47.
[8] P. Nattermann, Rep. Math. Phys. **36(2/3)**, 387 (1995).
[9] H.-D. Doebner, G. A. Goldin, and P. Nattermann, in *Quantization, Coherent States, and Complex Structures*, edited by J.-P. Antoine, S. T. Ali, W. Lisiecki, I. M. Mladenov, and A. Odzijewicz (Plenum, New York, 1995), p. 27.
[10] G. A. Goldin, in *Nonlinear, Deformed, and Irreversible Quantum Systems*, edited by H.-D. Doebner, V. K. Dobrev, and P. Nattermann (World Scientific, Singapore, 1995), p. 125.
[11] P. Nattermann and W. Scherer, in Ref. 10, p. 188.
[12] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw–Hill, New York, 1965), p. 96.
[13] E. Nelson, Phys. Rev. **150**, 1079 (1966).
[14] This perspective is consistent with the discussion given by B. Mielnik, Commun. Math. Phys. **37**, 221 (1974).
[15] W. Lücke, in Ref. 10, p. 140.
[16] P. Nattermann, *Generalized Quantum Mechanics and Nonlinear Gauge Transformations*, in Symmetry in Science, Vol. IX, edited by B. Gruber and M. Ramek (Plenum, New York, 1997), p. 269; quant-ph/9703017.
[17] G. A. Goldin and G. Svetlichny, J. Math. Phys. **35**, 3322 (1994).
[18] W. Lücke and P. Nattermann, *Nonlinear Quantum Mechanics and Locality*, in Symmetry in Science Vol. X, edited by B. Gruber and M. Ramek (New York, Plenum, 1998), p. 197; quant-ph/9707055.
[19] G. Rideau and P. Winternitz, J. Math. Phys. **34**, 558 (1993).
[20] H.-D. Doebner and G. A. Goldin, in *Particles and Fields*, edited by H.-D. Doebner, R. Raczka, and M. Pawlowski (World Scientific, Singapore, 1993), p. 115.
[21] R. Dashen and D. H. Sharp, Phys. Rev. **165**, 1867 (1968).
[22] G. A. Goldin and D. H. Sharp, in *Group Representations in Mathematics and Physics*, edited by V. Bargmann, Lecture Notes in Physics Vol. 6 (Springer, Berlin, 1970), p. 300.
[23] G. A. Goldin, J. Math. Phys. **12**, 462 (1971).
[24] B. Angermann, H.-D. Doebner, and J. Tolar, in *Nonlinear Partial Differential Operators and Quantization Procedures*, edited by S. I. Andersson and H.-D. Doebner, Lecture Notes in Mathematics Vol. **1037** (Springer, Berlin, 1983), p. 171.
[25] H.-D. Doebner and J. D. Hennig, *A Quantum Mechanical Evolution Equation for Mixed States from Symmetry and Kinematics*, in Symmetries in Science Vol. VIII, edited by B. Gruber (Plenum, New York, 1995), p. 85.
[26] H.-D. Doebner and P. Nattermann, Acta Phys. Pol. B **27**, 2327 (1996); **27**, 4003 (1996).
[27] N. Gisin, Phys. Lett. A **143**, 1 (1990).
[28] J. Polchinski, Phys. Rev. Lett. **66**, 397 (1991).
[29] N. Gisin, in Ref. 10, p. 109.

A rigorous real time Feynman path integral

Ken Loo[a)]
P.O. Box 9160, Portland, Oregon 97207

(Received 24 June 1998; accepted for publication 21 September 1998)

Using improper Riemann integrals, we will formulate a rigorous version of the real-time, time-sliced Feynman path integral for the L^2 transition probability amplitude. We will do this for nonvector potential Hamiltonians with potential which has, at most, a finite number of discontinuities and singularities. We will also provide a Nonstandard Analysis version of our formulation. © *1999 American Institute of Physics.* [S0022-2488(99)03801-3]

I. INTRODUCTION AND NOTATIONS

In this paper, we will formulate a rigorous version of the real-time, time-sliced Feynman path integral for the L^2 transition probability amplitude,

$$\left\langle \phi^*, \exp\left(\frac{-it\overline{H}}{\hbar}\right)\psi\right\rangle_{L^2} = \int_{\mathbb{R}^n} \phi(\mathbf{x})\left[\exp\left(\frac{-it\overline{H}}{\hbar}\right)\psi\right](\mathbf{x})\,d\mathbf{x}, \tag{1.1}$$

where $\phi,\psi\in L^2$, $H=(-\hbar^2/2m)\Delta+V(\mathbf{x})$ is essentially self-adjoint, $\overline{H}$ is the closure of H, and ϕ, ψ, V each carries, at most, a finite number of singularities and discontinuities. In favor of physics literature, we will formulate the Feynman path integral with improper Riemann integrals. In hope that with further research we can formulate a rigorous polygonal path integral, we will also provide a Nonstandard Analysis version of the Feynman path integral. Using Nonstandard Analysis is not essential to our formulation, and the idea of using Nonstandard Analysis on the Feynman path integral is not a new concept. For readers interested in Nonstandard Analysis, and its applications to Feynman path integrals, see Refs. 1–5 and references within. We will assume that the reader is familiar with Nonstandard Analysis.

In physics, the Feynman path integral is formulated on the propagator, and it is formally given by (see Refs. 6–8)

$$K_t(\mathbf{x},\mathbf{x}_0)=\lim_{k\to\infty} w_{n,k}\int_{\mathbb{R}^{(k-1)n}} \exp\left[\frac{i\epsilon}{\hbar}S_k(\mathbf{x},\ldots,\mathbf{x}_0)\right]d\mathbf{x}_1\cdots d\mathbf{x}_{k-1}, \tag{1.2}$$

where

$$w_{n,k}=\left(\frac{m}{2i\pi\hbar\epsilon}\right)^{nk/2}, \qquad \epsilon=\frac{t}{k},$$
$$S_k(\mathbf{x}=\mathbf{x}_k,\ldots,\mathbf{x}_0)=\sum_{j=1}^{k}\left[\frac{m}{2}\left(\frac{\mathbf{x}_j-\mathbf{x}_{j-1}}{\epsilon}\right)^2-V(\mathbf{x}_j)\right], \tag{1.3}$$

and all integrals are improper Riemann integrals.

In mathematics, there is a rigorous time-sliced Feynman path integral for the wave function (see Refs. 9 and 10)

---

a)Electronic mail: look@math.purdue.edu

$$\left[\exp\left(\frac{-it}{\hbar}\right)\psi\right](\mathbf{x})=\lim_{k\to\infty}w_{n,k}\int_{\mathbb{R}^{kn}}\exp\left[\frac{i\epsilon}{\hbar}S_k(\mathbf{x},\ldots,\mathbf{x}_0)\right]\psi(\mathbf{x}_0)d\mathbf{x}_0\cdots d\mathbf{x}_{k-1}, \tag{1.4}$$

where the integrals in (1.4) are improper Lebesgue integrals and their convergence is in the L^2 norm. Other popular rigorous versions of the Feynman path integral are the Wiener integral (see Refs. 11–14 and 10), generalization of Fresnel integrals (see Ref. 15), and Henstock integrals (see Ref. 16). For a more detailed exposition and further references, see Refs. 15 and 17.

Our main concern in this paper is to provide a rigorous version of (1.2) for the transition probability amplitude given in (1.1) by using (1.4). We will show that for any essentially self-adjoint Hamiltonian with potential that carries a finite number of singularities and discontinuities, and for any $\phi,\psi\in L^2$ which also has a finite number of singularities and discontinuities the following holds:

$$\int_{\mathbb{R}^n}\phi(\mathbf{x})\left[\exp\left(\frac{-it\overline{H}}{\hbar}\right)\psi\right](\mathbf{x})d\mathbf{x}$$
$$=\lim_{k\to\infty}w_{n,k}\int_{\mathbb{R}^{(k+1)n}}\phi(\mathbf{x}_k)\exp\left[\frac{i\epsilon}{\hbar}S_k(\mathbf{x}_k,\ldots,\mathbf{x}_0)\right]\psi(\mathbf{x}_0)d\mathbf{x}_0\cdots d\mathbf{x}_k. \tag{1.5}$$

In the last line of (1.5), the integral is an improper Riemann integral over $\mathbb{R}^{(k+1)n}$.

A trivial application of Nonstandard Analysis on the k limit in (1.5) yields (1.6),

$$\int_{\mathbb{R}^n}\phi(\mathbf{x})\left[\exp\left(\frac{-it\overline{H}}{\hbar}\right)\psi\right](\mathbf{x})d\mathbf{x}$$
$$=st\left\{w_{n,\omega}\int_{{}^*\mathbb{R}^{(\omega+1)n}}\phi(\mathbf{x}_\omega)\exp\left[\frac{i\epsilon}{\hbar}S_\omega(\mathbf{x}_\omega,\ldots,\mathbf{x}_0)\right]\psi(\mathbf{x}_0)d\mathbf{x}_0\cdots d\mathbf{x}_\omega\right\}, \tag{1.6}$$

where the integral in the last line of (1.6) is a $*$-transformed improper Riemann integral over ${}^*\mathbb{R}^{(\omega+1)n}$, and $\omega\in{}^*\mathbb{N}-\mathbb{N}$.

The main idea in the proof of (1.5) is the following. For simplicity, suppose $f(x)\in L^2(\mathbb{R})$, $g(x,y)\in L^2(\mathbb{R}\times\mathbb{R})$ are such that they are bounded and continuous. Further, suppose that both

$$h(x)=\int_{-a}^{b}g(x,y)dy,$$
$$p(x)=\lim_{a,b\to\infty}\int_{-a}^{b}g(x,y)dy \tag{1.7}$$

are in $L^2(\mathbb{R})$ as a function of x. In (1.7), we take the integral to be Lebesgue integrals and the limits are taken independent of each other in the L^2 topology. Notice that for $p(x)$, we can interpret the integral as an improper Lebesgue integral with convergence in the L^2 topology. Let us denote $\chi_{[-c,d]}$ to be the characteristic function on $[-c,d]$. Schwarz's inequality then implies

$$\left|\int_{\mathbb{R}}f(x)p(x)-\int_{-c}^{d}\int_{-a}^{b}f(x)g(x,y)dx\;dy\right|\leqslant\|f\|_2\|p-h\|_2+\|f-\chi_{[-c,d]}\,f\,\|_2\|h\|_2\to 0. \tag{1.8}$$

Thus, we can write

$$\int_{\mathbb{R}}f(x)p(x)=\lim_{a,b,c,d\to\infty}\int_{-c}^{d}\int_{-a}^{b}f(x)g(x,y)dx\;dy, \tag{1.9}$$

where the limits are all taken independent of each other. Since f and g are bounded and continuous, the Lebesgue integral over $[-a,b]\times[-c,d]$ in (1.9) can be replaced by a Riemann integral. Since the limits are taken independent of each other, we can then interpret the right-hand side of (1.9) as an improper Riemann integral. If f and g carry singularities and discontinuities, care must be taken in the region of integration so that the replacement of the Lebesgue integral with Riemann integrals can be done.

We now set some notations to deal with n-dimensional integrations, singularities and discontinuities. Let $k\in\mathbb{N}$ and $0\leqslant l\leqslant k$. We will denote the interior of the lth box by

$$A^l=(-a_1^l,b_1^l)\times\cdots\times(-a_n^l,b_n^l), \tag{1.10}$$

for positive and large a's and b's. Let $K=\{\mathbf{y}_1\cdots\mathbf{y}_p\}$ be the set of discontinuous and singular points of ϕ, ψ and V. For each $\mathbf{y}_q=(y_1^q,\ldots,y_n^q)\in K$, denote the lth box centered at $\mathbf{y}_q$ by

$$B_q^l=\left(y_1^q-\frac{1}{c_1^{q,l}},y_1^q+\frac{1}{d_1^{q,l}}\right)\times\cdots\times\left(y_n^q-\frac{1}{c_n^{q,l}},y_n^q+\frac{1}{d_n^{q,l}}\right), \tag{1.11}$$

for positive and large c's and d's. Let

$$C^l=A^l-\left\{\bigcup_{q=1}^{p}B_q^l\right\}. \tag{1.12}$$

For arbitrary large a's, b's, c's and d's, C^l is a box which encloses the set K and at each point of K, a small box centered at that point is taken out. Associated with C^l is a set of indices,

$$\{j_l\}=\{a_1^l,\ldots,a_n^l,b_1^l,\ldots,b_n^l,c_1^{1,l},\ldots,c_n^{1,l},\ldots,c_1^{p,l},\ldots,c_n^{p,l},d_1^{1,l},\ldots,d_n^{1,l},\ldots,d_1^{p,l},\ldots,d_n^{p,l}\}. \tag{1.13}$$

We will denote by $\{j_l\}\rightarrow\infty$ to mean

$$a_1^l,\ldots,a_n^l,b_1^l,\ldots,b_n^l,c_1^{1,l},\ldots,c_n^{1,l},\ldots,c_1^{p,l},\ldots,c_n^{p,l},d_1^{1,l},\ldots,d_n^{1,l},\ldots,d_1^{p,l},\ldots,d_n^{p,l}\rightarrow\infty, \tag{1.14}$$

where all indices go to infinity independent of each other. Notice that as $\{j_l\}\rightarrow\infty$, we recover $\mathbb{R}^n$ *a.e.* from C^l. We will denote by $\chi_{\{j_l\}}$ the characteristic function on C^l. Notice that for $f\in L^2(\mathbb{R}^n)$,

$$\lim_{\{j_l\}\rightarrow\infty}\chi\{j_l\}f=f,\quad a.e. \tag{1.15}$$

Last, let us write

$$D_{\{J_0^k\}}=C^0\times\cdots\times C^k. \tag{1.16}$$

Associated with $D_{\{J_0^k\}}$ is a set of indices,

$$\{J_0^k\}=\bigcup_{l=0}^{k}\{j_l\}, \tag{1.17}$$

and as before, we will use the notation $\{J_0^k\}\rightarrow\infty$ to mean

$$\{j_0\}\rightarrow\infty,\ldots,\{j_k\}\rightarrow\infty, \tag{1.18}$$

where the indices are taken to infinity independent of each other.

From here on, we will assume that $\phi,\psi\in L^2$ and V are such that they have, at most, a finite number of singularities and discontinuities and the set of those points are denoted as $K=\{\mathbf{y}_1\cdots\mathbf{y}_p\}$. Finally, we will denote by $\int_{rO}$ to be Riemann or improper Riemann integration over the region O and $\int_O$ to be Lebesgue integration over the region O.

II. FEYNMAN PATH INTEGRALS

The standard derivation of (1.4) is via the Trotter product formula (see Refs. 9, 14, and 10), which says that for any essentially self-adjoint $H=H_0+V$ with $H_0=(-\hbar^2/2m)\Delta$ and any $\psi\in L^2$,

$$\exp\left(\frac{-it\overline{H}}{\hbar}\right)\psi=\lim_{k\to\infty}\left\{\exp\left(\frac{-itV}{k\hbar}\right)\exp\left(\frac{-itH_0}{k\hbar}\right)\right\}^k\psi, \tag{2.1}$$

where the limit is taken in the L^2 norm.

Thus, we have the following.

Lemma 2.1: Suppose $H=H_0+V$ is essentially self-adjoint. Let $\psi,\phi\in L^2$; then

$$\int_{\mathbb{R}^n}\phi(\mathbf{x})\left[\exp\left(\frac{-it\overline{H}}{\hbar}\right)\psi\right](\mathbf{x})d\mathbf{x}=\lim_{k\to\infty}\int_{\mathbb{R}^n}\phi(\mathbf{x})\left[\left\{\exp\left(\frac{-itV}{k\hbar}\right)\exp\left(\frac{-itH_0}{k\hbar}\right)\right\}^k\psi\right](\mathbf{x})d\mathbf{x}, \tag{2.2}$$

where the limit in (2.2) is taken pointwise in t.

Proof: The proof is just an application of (2.1) and Schwarz's inequality. □

It is well known that (see Refs. 9 and 10) for $\nu\in L^1(\mathbb{R}^n)\cap L^2(\mathbb{R}^n)$,

$$\left[\exp\left(\frac{-i\epsilon H_0}{\hbar}\right)\nu\right](\mathbf{x}_1)=\left(\frac{m}{2i\pi\hbar\epsilon}\right)^{n/2}\int_{\mathbb{R}^n}\exp\left[\frac{im\epsilon}{2\hbar}\left(\frac{\mathbf{x}_1-\mathbf{x}_0}{\epsilon}\right)^2\right]\nu(\mathbf{x}_0)d\mathbf{x}_0, \tag{2.3}$$

and that the operator $\exp(-i\epsilon H_0/\hbar)$ is unitary. Thus, we can write

$$\begin{aligned}\left[\exp\left(\frac{-i\epsilon H_0}{\hbar}\right)\psi\right](\mathbf{x}_1)&=\left[\exp\left(\frac{-i\epsilon H_0}{\hbar}\right)[\lim_{\{j_0\}\to\infty}\chi\{j_0\}\psi]\right](\mathbf{x}_1)\\&=\lim_{\{j_0\}\to\infty}\left[\exp\left(\frac{-i\epsilon H_0}{\hbar}\right)\chi_{\{j_0\}}\psi\right](\mathbf{x}_1)\\&=\lim_{\{j_0\}\to\infty}\left(\frac{m}{2i\pi\hbar\epsilon}\right)^{n/2}\int_{C^0}\exp\left[\frac{im\epsilon}{2\hbar}\left(\frac{\mathbf{x}_1-\mathbf{x}_0}{\epsilon}\right)^2\right]\psi(\mathbf{x}_0)d\mathbf{x}_0,\end{aligned} \tag{2.4}$$

where the limits are taken in L^2. Notice that by construction of the region C^0, ψ is bounded and continuous on C^0, hence the Lebesgue integral in the last line of (2.4) can be replaced by a Riemann integral.

For notation convenience, we will denote

$$\begin{gathered}\rho(\mathbf{x}_k,\{J_0^{k-1}\})=w_{n,k}\int_{D_{\{J_0^{k-1}\}}}\exp\left[\frac{it}{(k+1)\hbar}S_k(\mathbf{x}_k,\ldots,\mathbf{x}_0)\right]\psi(\mathbf{x}_0)d\mathbf{x}_0\cdots d\mathbf{x}_{k-1},\\T=\exp\left(\frac{-itV}{k\hbar}\right)\exp\left(\frac{-itH_0}{k\hbar}\right),\quad T^k=\left\{\exp\left(\frac{-itV}{k\hbar}\right)\exp\left(\frac{-itH_0}{k\hbar}\right)\right\}^k.\end{gathered} \tag{2.5}$$

Lemma 2.2: Suppose $H=H_0+V$ is essentially self-adjoint. Let $\psi\in L^2$, then for $k\in\mathbb{N}$ the following holds:

$$\left\{\exp\left(\frac{-itV}{k\hbar}\right)\exp\left(\frac{-itH_0}{k\hbar}\right)\right\}^k\psi = \lim_{\{J_0^{k-1}\}\to\infty} w_{n,k}\int_{D_{\{J_0^{k-1}\}}}\exp\left[\frac{i\epsilon}{\hbar}S_k(\mathbf{x}_k,\ldots,\mathbf{x}_0)\right]\psi(\mathbf{x}_0)d\mathbf{x}_0\cdots d\mathbf{x}_{k-1}, \tag{2.6}$$

where the limits are taken in the L^2 norm.

Proof: We will proof (2.6) by induction. Suppose $k=2$; then (2.4) implies

$$T^2\psi = T\left\{\lim_{\{J_0^0\}\to\infty}\rho(\mathbf{x}_1,\{J_0^0\})\right\} = \exp\left(\frac{-itV}{2\hbar}\right)\exp\left(\frac{-itH_0}{2\hbar}\right)\left\{\lim_{\{j_1\}\to\infty}\chi_{\{j_1\}}\left[\lim_{\{J_0^0\}\to\infty}\rho(\mathbf{x}_1,\{J_0^0\})\right]\right\}. \tag{2.7}$$

Since multiplication by a characteristic function $\exp(-itV/2\hbar)$, and $\exp(-itH_0/2\hbar)$ are all continuous operators from L^2 to L^2, we can take the L^2 limits in (2.7) outside of the operators and we can do this in any order we wish. Hence, (2.6) is true for $k=2$. Assuming (2.6) to be true for k, then

$$T^{k+1}\psi = \exp\left(\frac{-itV}{(k+1)\hbar}\right)\exp\left(\frac{-itH_0}{(k+1)\hbar}\right)\left\{\lim_{\{j_k\}\to\infty}\chi_{\{j_k\}}\left[\lim_{\{J_0^{k-1}\}\to\infty}\rho(\mathbf{x}_k,\{J_0^{k-1}\})\right]\right\}. \tag{2.8}$$

By the same reasoning as for the case of $k=2$, we can take all the L^2 limits in (2.8) outside of the operators, and we can do this in any order we wish. Hence, (2.6) is true for all $k\in\mathbb{N}$. □

Proposition 2.3: Suppose $H=H_0+V$ is essentially self-adjoint. Let $\psi,\phi\in L^2$; Let ψ, ϕ, V be such that they have at most a finite number of discontinuities and singularities then for all $k\in\mathbb{N}$ the following is true:

$$\int_{\mathbb{R}^n}\phi(\mathbf{x})\left[\left\{\exp\left(\frac{-itV}{k\hbar}\right)\exp\left(\frac{-itH_0}{k\hbar}\right)\right\}^k\psi\right](\mathbf{x})d\mathbf{x} = w_{n,k}\int_{r\mathbb{R}^{(k+1)n}}\phi(\mathbf{x}_k)\exp\left[\frac{i\epsilon}{\hbar}S_k(\mathbf{x}_k,\ldots,\mathbf{x}_0)\right]\psi(\mathbf{x}_0)d\mathbf{x}_0\cdots d\mathbf{x}_k. \tag{2.9}$$

Proof: Lemma 2.2 implies that

$$\begin{aligned}\int_{\mathbb{R}^n}\phi(\mathbf{x})\left[\left\{\exp\left(\frac{-itV}{k\hbar}\right)\exp\left(\frac{-itH_0}{k\hbar}\right)\right\}^k\psi\right](\mathbf{x})d\mathbf{x} &= w_{n,k}\int_{\mathbb{R}^n}\left\{\lim_{\{j_k\}\to\infty}\chi_{\{j_k\}}\phi(\mathbf{x}_k)\right\}\\ &\times\left\{\lim_{\{J_0^{k-1}\}\to\infty}w_{n,k}\int_{D_{\{J_0^{k-1}\}}}\exp\left[\frac{i\epsilon}{\hbar}S_k(\mathbf{x}_k,\ldots,\mathbf{x}_0)\right]\psi(\mathbf{x}_0)d\mathbf{x}_0\cdots d\mathbf{x}_{k-1}\right\}d\mathbf{x}_k.\end{aligned} \tag{2.10}$$

We now apply the idea in (1.8) and (1.9). Since all limits in (2.10) are taken independent of each other, we can use Schwarz's inequality and take all the L^2 limits outside of the integral as pointwise limits. Thus,

$$\int_{\mathbb{R}^n} \phi(\mathbf{x}) \left[\left\{ \exp\left(\frac{-itV}{k\hbar} \right) \exp\left(\frac{-itH_0}{k\hbar} \right) \right\}^k \psi \right](\mathbf{x}) d\mathbf{x}$$

$$= w_{n,k} \lim_{\{J_0^k\} \to \infty} \int_{D_{\{J_0^k\}}} \phi(\mathbf{x}_k) \exp\left[\frac{i\epsilon}{\hbar} S_k(\mathbf{x}_k, \ldots, \mathbf{x}_0) \right] \psi(\mathbf{x}_0) d\mathbf{x}_0 \cdots d\mathbf{x}_k, \tag{2.11}$$

where the limits are pointwise in t.

By construction of $D_{\{J_0^k\}}$, the integrand $\phi(\mathbf{x}_k)\exp[(i\epsilon/\hbar)S_k(\mathbf{x}_k,\ldots,\mathbf{x}_0)]\psi(\mathbf{x}_0)$ in (2.11) is a bounded and continuous function on $D_{\{J_0^k\}}$. Hence, we can replace the Lebesgue integrals in (2.11) by Riemann integrals. Since all limits in (2.11) are taken independent of each other, we can interpret (2.11) as an improper Riemann integral. □

We are now ready to prove (1.5).

Theorem 2.4: *Suppose $H = H_0 + V$ is essentially self-adjoint. Let $\psi, \phi \in L^2$. Furthermore, suppose that ψ, ϕ and V have, at most, a finite number of singularities and discontinuities. With our previously defined notations, the following is true:*

$$\int_{\mathbb{R}^n} \phi(\mathbf{x}) \left[\exp\left(\frac{-it\overline{H}}{\hbar} \right) \psi \right](\mathbf{x}) d\mathbf{x}$$

$$= \lim_{k \to \infty} w_{n,k} \int_{r\mathbb{R}^{(k+1)n}} \phi(\mathbf{x}_k) \exp\left[\frac{i\epsilon}{\hbar} S_k(\mathbf{x}_k, \ldots, \mathbf{x}_0) \right] \psi(\mathbf{x}_0) d\mathbf{x}_0 \cdots d\mathbf{x}_k. \tag{2.12}$$

Proof: Follows from Lemma 2.1 and Proposition 2.3. □

III. NONSTANDARD FEYNMAN PATH INTEGRALS

A trivial application of Nonstandard Analysis on the K limit in (2.2) will produce (1.6). It is our hope that with further research, a rigorous nonstandard polygonal path integral can be formulated.

Theorem 3.1: *Suppose $H = H_0 + V$ is essentially self-adjoint. Let $\psi, \phi \in L^2$. Furthermore, suppose that ψ, ϕ and V has, at most, a finite number of singularities and discontinuities. With our previously defined notations, the following is true:*

$$\int_{\mathbb{R}^n} \phi(\mathbf{x}) \left[\exp\left(\frac{-it\overline{H}}{\hbar} \right) \psi \right](\mathbf{x}) d\mathbf{x}$$

$$= st \left\{ w_{n,\omega} \int_{r{}^*\mathbb{R}^{(\omega+1)n}} \phi(\mathbf{x}_\omega) \exp\left[\frac{i\epsilon}{\hbar} S_\omega(\mathbf{x}_\omega, \ldots, \mathbf{x}_0) \right] \psi(\mathbf{x}_0) d\mathbf{x}_0 \cdots d\mathbf{x}_\omega \right\}, \tag{3.1}$$

where the integral in the last line of (3.1) is a $*$-transformed improper Riemann integral over ${}^*\mathbb{R}^{(\omega+1)n}$, and $\omega \in {}^*\mathbb{N} - \mathbb{N}$.

Proof: The nonstandard equivalent of Lemma 2.1 is that, for any $\omega \in {}^*\mathbb{N} - \mathbb{N}$,

$$\int_{\mathbb{R}^n} \phi(\mathbf{x}) \left[\exp\left(\frac{-it\overline{H}}{\hbar} \right) \psi \right](\mathbf{x}) d\mathbf{x} = st \left\{ \int_{\mathbb{R}^n} \phi(\mathbf{x}) \left[\left\{ \exp\left(\frac{-itV}{\omega\hbar} \right) \exp\left(\frac{-itH_0}{\omega\hbar} \right) \right\}^\omega \psi \right](\mathbf{x}) d\mathbf{x}. \tag{3.2}$$

After $*$-transforming Proposition 2.3, Eq. (3.1) follows from (3.2). □

ACKNOWLEDGMENTS

The author would like to give thanks to his Ph.D. thesis advisor James Thurber and his Ph.D. thesis committee: Daniel Gottlieb, Leonard Lipshitz, Herman Rubin, and James Thurber. Also, special thanks to Patricia Bauman.

[1] S. Albeverio, J. Fenstat, R. Høegh-Krohn, and T. Lindstrøm, *Nonstandard Methods in Stochastic Analysis and Mathematical Physics* (Academic, New York, 1986).
[2] N. Cutland, *Nonstandard Analysis and Its Applications* (Cambridge University Press, Cambridge, 1988).
[3] A. E. Hurd and P. A. Loeb, *An Introduction to Nonstandard Real Analysis* (Academic, New York, 1985).
[4] A. Robinson, *Nonstandard Analysis* (Princeton University Press, Princeton, 1996).
[5] K. Stroyan and W. Luxemburg, *Introduction to the Theory of Infinitesimals* (Academic, New York, 1976).
[6] R. Feynman and A. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).
[7] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, and Polymer Physics* (World Scientific, Singapore, 1995).
[8] L. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981).
[9] J. Blank, P. Exner, and M. Havlīček, *Hilber Space Operators in Quantum Physics* (American Institute of Physics, New York, 1994).
[10] M. Reed and B. Simon, *Functional Analysis I and II* (Academic, New York, 1980).
[11] R. Cameron, ''A family of integrals serving to connect the Wiener and Feynman integrals,'' J. Math. Phys. **39**, 126–141 (1961).
[12] R. Cameron, ''The Ilstow and Feynman integrals,'' J. Anal. Math **10**, 287–361 (1962/1963).
[13] J. Glimm and A. Jaffe, *Quantum Physics, A Functional Integral Point of View*, 2nd ed. (Springer-Verlag, Berlin, 1987).
[14] E. Nelson, ''Feynman integrals and the Schrodinger equation,'' J. Math. Phys. **5**, 332–343 (1964).
[15] S. Albeverio and R. Høegh-Krohn, *Mathematical Theory of Feynman Path Integrals*, Lecture Notes in Mathematics (Springer-Verlag, Berlin, 1976), Vol. 523.
[16] P. Muldowney, *A General Theory of Integration in Function Spaces* (Longman Scientific & Technical and Wiley, New York, 1987).
[17] S. Albeverio *et al.*, eds., *Feynman Path Integrals*, Lecture Notes in Physics (Springer-Verlag, Berlin, 1979), Vol. 106.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

Second-order mixtures in relativistic Schrödinger theory

M. Mattes and M. Sorg
II. Institut für Theoretische Physik der Universität Stuttgart, Pfaffenwaldring 57, D 70550 Stuttgart, Germany

(Received 24 September 1997; accepted for publication 19 August 1998)

In relativistic Schrödinger theory the mixtures and pure states can be treated from a unified point of view such that a pure state merely emerges as a special case of a mixture. Here the concept of mixture is of purely local nature and therefore the mixture character (*degree of order*) can change over space and time. Although the general dynamics does not forbid the transitions from mixtures to pure states (and vice versa), the considered models do admit these transitions only in an asymptotic sense. The general concepts and results are demonstrated by considering the four-component Dirac theory for spinning matter over the Robertson–Walker universes. A detailed study is made for a specific subclass of second-order mixtures sharing many of their properties with the pure states (i.e., wave functions). © *1999 American Institute of Physics.* [S0022-2488(98)01312-7]

I. INTRODUCTION AND SURVEY

The right interpretation of quantum theory was a permanent point of concern for those people who were not satisfied with merely extracting numerical predictions from the quantum formalism. The recent controversy[1] may be especially considered as further evidence for the fact that no commonly accepted viewpoint among physicists exists about such fundamental questions as the following.

(i) Are the true laws of physics time symmetric? (If yes, how do you then explain the time-asymmetric reduction of the wave function during the measuring process?)
(ii) Is there some continuous dynamical process which links a pure state to a mixture with progression of time?

The last problem in particular, originally put forward by Schrödinger[2] in the form of his famous cat paradox, acquires more and more experimental relevance. The reason is that certain features of the wave function (e.g., its phase) increasingly become accessible to experimental observation[3] and this provides us with the certainty that physical systems can actually occur as pure states, even in the macroscopic domain. Moreover, the experimental situation also supports Schrödinger's claim that the wave function ψ must develop according to his famous equation

$$i\hbar\,\frac{\partial\psi}{\partial t}=\hat{H}\psi. \tag{I.1}$$

On the other hand, it was already pointed out by von Neumann[4] that not any quantum system can be described by some wave function (or state vector) ψ, but rather must be characterized by a more general concept: the density matrix $\hat{\rho}$. Its equation of motion was written down by von Neumann as

$$i\hbar\,\frac{d\hat{\rho}}{dt}=[\hat{H},\ \hat{\rho}], \tag{I.2}$$

so that Schrödinger's approach (I.1) was recovered as a special subcase hereof, namely when the density matrix is composed of a single wave function only:

$$\hat{\rho}\rightarrow\psi\otimes\overline{\psi}. \tag{I.3}$$

The density matrix approach to the quantum phenomena has received much credit (for a review see, e.g., Ref. 5), but nevertheless the pure states (to be described by a single wave function) must retain their physical significance. In such a situation, one could have expected that some unified theory would have been readily established so that one can predict when a pure state decays into a mixture (and vice versa). Indeed there were a few attempts to construct such a combined theory, which were especially aimed at the resolution of the well-known ''*quantum paradoxes*,'' e.g., the Einstein–Podolsky–Rosen paradox (for a review, see Ref. 6). However, it seems that the relatively moderate success of these attempts is due to the fact that some basic structural elements of the conventional quantum theory have been unduly distorted [e.g., Eq. (I.1) itself or the conservation laws]. Thus it seems reasonable to look for a completely new modification of the conventional theory so that its truly basic structure is preserved to a larger extent (albeit only in a more formal respect).

The present paper is concerned with such a completely new modification and presents an investigation of the relationships between mixtures and pure states within the framework of the recently established relativistic Schrödinger theory.[7–9] In particular, the smooth transitions from mixtures into pure states (and vice versa) are considered in detail. Here, the original Schrödinger equation (I.1) is replaced by its relativistic counterpart

$$i\hbar c\mathcal{D}_\mu\psi=\mathcal{H}_\mu\psi. \tag{I.4}$$

The (gauge plus coordinate) covariant derivative $\mathcal{D}_\mu$ is defined as usual through

$$\mathcal{D}_\mu\psi:=\partial_\mu\psi+\mathcal{A}_\mu\cdot\psi, \tag{I.5}$$

where the anti-Hermitian gauge potential $\mathcal{A}_\mu(=-\overline{\mathcal{A}}_\mu)$ generates the ''field strength'' (i.e., bundle curvature) $\mathcal{F}_{\mu\nu}$ in the standard way

$$\mathcal{F}_{\mu\nu}=\nabla_\mu\mathcal{A}_\nu-\nabla_\nu\mathcal{A}_\mu+[\mathcal{A}_\mu,\ \mathcal{A}_\nu]. \tag{I.6}$$

The Hamiltonian $\mathcal{H}_\mu$, corresponding to Schrödinger's $\hat{H}$ (I.1), is now a $\mathcal{GL}(N,\mathbb{C})$-valued one-form and acts upon the N_f-component wave function ψ as a section of the underlying complex vector bundle. However, contrary to Schrödinger's $\hat{H}$, the present relativistic Hamiltonian $\mathcal{H}_\mu$ is itself a dynamical object to be first determined from its field equations. These consist of the ''*integrability condition*''

$$\mathcal{D}_\mu\mathcal{H}_\nu-\mathcal{D}_\nu\mathcal{H}_\mu+\frac{i}{\hbar c}[\mathcal{H}_\mu,\ \mathcal{H}_\nu]=i\hbar c\mathcal{F}_{\mu\nu} \tag{I.7}$$

and of the ''*conservation equation*'' which reads in the present case of the four-component Dirac theory $(N_f=4)$

$$\gamma^\mu\cdot\mathcal{H}_\mu=Mc^2\cdot\mathbf{1}. \tag{I.8}$$

Clearly, M is the mass parameter and the objects $\gamma^\mu(\mu=0,1,2,3)$ are the Dirac matrices generating the Clifford algebra $\mathbb{C}(1,3)$ over pseudo-Riemannian space–time. Thus, any solution $\psi(x)$ of the relativistic Schrödinger's equation (I.4) also obeys the Dirac equation

$$i\gamma^\mu\mathcal{D}_\mu\psi=m\psi,\quad \left(m:=\frac{Mc}{\hbar}\right). \tag{I.9}$$

After Schrödinger's ideas have thus been incorporated into the new framework, let us now turn to von Neumann's equation (I.2). (The equation of motion for the field strength $\mathcal{F}_{\mu\nu}$ is of the Yang–Mills type not to be considered here.) The relativistic analog of $\hat{\rho}$ is the ''*intensity matrix*'' $\mathcal{I}$ and obeys the following equation of motion:

$$i\hbar c\mathcal{D}_\mu\mathcal{I}=[\mathcal{H}_\mu\cdot\mathcal{I}-\mathcal{I}\cdot\overline{\mathcal{H}}_\mu]. \tag{I.10}$$

Observe here that the Hamiltonian in general, will not be Hermitian ($\mathcal{H}_\mu \neq \overline{\mathcal{H}_\mu}$), but nevertheless the conservation laws do apply. For instance, consider the continuity equation

$$\nabla^\mu j_\mu = 0, \tag{I.11}$$

and define the current density j_μ through

$$j_\mu = \mathrm{tr}\,(\mathcal{I}\cdot\gamma_\mu), \tag{I.12}$$

resp. for the case of a pure state

$$j_\mu = \overline{\psi}\cdot\gamma_\mu\cdot\psi, \tag{I.13a}$$

$$(\mathcal{I}\rightarrow\psi\otimes\overline{\psi}). \tag{I.13b}$$

Now carry through the differentiation process (I.10), i.e.,

$$\nabla^\mu j_\mu = \mathrm{tr}\,(\gamma^\mu(\mathcal{D}_\mu\mathcal{I})), \quad (\mathcal{D}_\mu\gamma_\nu \equiv 0). \tag{I.14}$$

Furthermore, observe the relativistic von Neumann equation (I.10) as well as the conservation equation (I.8) and then find that the continuity equation (I.11) actually is valid. Thus, we simultaneously can attain the relativistic realization of both Schrödinger's and von Neumann's ideas together with the validity of the conservation laws!

Thus, Hermiticity of the Hamiltonian is not necessary and not even desirable for the present approach, contrary to the nonrelativistic Schrödinger theory (where $\hat{H}=\hat{H}^{+}$). Neither is the Hermiticity of the Hamiltonian in our theory needed for obtaining real energy eigenvalues because the nonrelativistic concept of energy eigenvalue must in any case be incorporated into a more general relativistic context. Consequently, the ''energy'' will emerge here by solving the nonlinear field equations for the Hamiltonian $\mathcal{H}_\mu$ (see below), but not by solving some linear eigenvalue problem for $\mathcal{H}_\mu$ which, in fact, would have required its Hermiticity. It is well known for the nonrelativistic approach that it is just the Hermiticity of the Hamiltonian $\hat{H}$ which forbids the possibility of transitions from mixtures into pure states (and vice versa).[10] However, since the relativistic Hamiltonian $\mathcal{H}_\mu$ is not Hermitian, it becomes possible in relativistic Schrödinger theory (RST) to study those transitions between mixtures and pure states. More precisely, one expects that a physical system may be a true mixture over one part of space–time [→general $\mathcal{I}(x)$], but may be in a pure state (I.13b) in some other part. The purely temporal transitions have been studied for a scalar Higgs doublet in a preceding paper,[9] and it has been found that a mixture cannot evolve into a pure state, except if one admits the existence of a ''Fierz potential.'' Such a potential minimizes the energy of the system for the pure states but equips the mixtures with an additional energy content. Although the system is pressed into a pure state by such a potential, the pure-state configuration could be attained only asymptotically after sufficiently large expansion of the universe.

In the present paper, we are concerned with the analogous questions for Dirac's theory of spinning matter; especially we want to know whether RST admits *spatial* changes of the mixture character.

First, in Sec. II we present a collection of the kinematical properties of the pure states (i.e., Dirac spinors) which are most concisely expressed by the well-known Fierz identities for the physical densities [see Eqs. (II.5a)–(II.5c)]. From this presentation arises a natural classification of the mixtures in terms of their ''*degree of order*'' $N'(\geqslant 1)$ so that the pure states appear as ''mixtures of order $N'=1$.'' Such a unified view of mixtures and pure states is convenient for the study of transitions between different order parameters N'! Moreover, it thus becomes obvious that for any order N' there must exist the corresponding generalization of the Fierz identities.

In Sec. III we introduce the deviation operator $\mathfrak{D}_\mathrm{F}$ (''*deviator*'') which measures the deviation of a mixture from a pure-state configuration [see Eq. (III.1)]. For the Dirac fiber dimension $N_f=4$ there arise $4\times 4=16$ deviation fields $\{\Delta_\mathrm{F},\tilde{\Delta}_F,\Delta_\mu,\tilde{\Delta}_\mu,\Delta_{\mu\nu}\}$. It turns out that there is a special subclass of *second-order mixtures* ($N'=2$) whose kinematic structure is especially similar to that of the pure states and which is characterized by vanishing tensor deviation ($\Delta_{\mu\nu}\equiv 0$). For instance, the polarization density $S_{\mu\nu}$ is the same as for the pure states, if expressed in terms of the scalar and vector densities [see Eq. (II.5c)], and, moreover, the remaining deviation fields can all be parametrized by the single scalar Δ_F. Thus for $\Delta_F = 0$ one can be sure to deal with a pure state. Because of these nice properties, such a mixture is chosen for the subsequent investigations.

In Sec. IV we present the dynamical equation for the deviator $\mathfrak{D}_F$ [see Eq. (IV.1)]. Clearly, *any* pure state ($\mathfrak{D}_F \equiv 0$) must fit into this equation of motion, otherwise it could not be the final stage of a mixture. However, this does not guarantee the *stability* of the pure states and this question cannot be settled here in full generality.

In Sec. V the fiber distributions are studied in terms of projectors. The latter objects are relevant for the relationship between mixtures and pure states because any pure state is associated to some one-dimensional projector. Thus it becomes important to elaborate the field equation for the projectors [see Eqs. (V.3a)–(V.3b)].

In Sec. VI we specify our model Hamiltonian for the subsequent investigation of the space–time behavior of the second-order mixtures [see Eq. (VI.1)]. It is true that this Hamiltonian may appear somewhat unsuitable for the present purpose because it originally was set up to describe the homogeneous and isotropic matter distributions in agreement with the cosmological principle.[7] However, it turns out that this Hamiltonian can also account for SO(3) symmetric, localized matter distributions which are then used in Sec. VII for the study of the spatial behavior of mixtures. Concerning their temporal evolution, the selected class of Hamiltonians forbids the transition into a pure state and the situation is somewhat similar to the previous case of the Klein–Gordon–Higgs fields.[9]

In the final section (Sec. VII), the difference between mixtures and pure states is elaborated with respect to some specific spatial pattern. As expected, the relativistic von Neumann equation (I.10) admits more physically meaningful configurations than the Schrödinger equation (I.4). For instance, for the considered SO(3) symmetric configuration the wave function $\psi(x)$, as a solution of (I.4), would develop a singularity at the symmetry center and therefore would forbid itself whereas the more general equation (I.10) admits mixtures with completely regular behavior.

II. FIRST-ORDER IDENTITIES

In Dirac's one-particle theory, the wave function $\psi(x)$ is some section of a complex vector bundle over pseudo-Riemannian space–time with the complex vector space $\mathbb{C}^4$ as typical fiber. Such a fiber space admits the existence of 16 Hermitian operators ($\mu=0,1,2,3$), namely $\mathbf{1}$, γ^μ, $\varepsilon = (1/4!)\varepsilon_{\mu\nu\lambda\sigma}\gamma^\mu\gamma^\nu\gamma^\lambda\gamma^\sigma$, $\tilde{\gamma}^\mu = \varepsilon\cdot\gamma^\mu$, and $\Sigma^{\mu\nu} = \frac{1}{4}[\gamma^\mu, \gamma^\nu]$. Any other Hermitian operator essentially is a linear combination of these basic elements e.g.,

$$ {}^*\Sigma^{\mu\nu} := -\varepsilon\cdot\Sigma^{\mu\nu} = \tfrac{1}{2}\varepsilon^{\mu\nu}{}_{\lambda\sigma}\Sigma^{\lambda\sigma} \tag{II.1} $$

or

$$ \{\Sigma^{\mu\nu}, \gamma^\lambda\} \equiv \Sigma^{\mu\nu}\gamma^\lambda + \gamma^\lambda\Sigma^{\mu\nu} = \varepsilon^{\mu\nu\lambda}_{\sigma}\tilde{\gamma}^\sigma. \tag{II.2} $$

These (and similar) relations may easily be verified by referring to the defining equations of the Clifford algebra $\mathbb{C}(1,3)$,

$$ \gamma_\mu\cdot\gamma_\nu + \gamma_\nu\cdot\gamma_\mu = 2g_{\mu\nu}\cdot\mathbf{1}, \tag{II.3} $$

with $g_{\mu\nu}$ denoting the pseudo-Riemannian metric tensor.

As soon as a vector section $\psi(x)$ is at hand together with these operator sections, one can immediately form the corresponding physical densities, i.e.,

$$ \rho(x) := \bar{\psi}\cdot\psi, \tag{II.4a} $$

$$ \tilde{\rho}(x) := \bar{\psi}\cdot\varepsilon\cdot\psi, \tag{II.4b} $$

$$ j_\mu(x) := \bar{\psi}\cdot\gamma_\mu\cdot\psi, \tag{II.4c} $$

$$\tilde{j}_\mu(x) := i\bar{\psi}\cdot\tilde{\gamma}_\mu\cdot\psi, \tag{II.4d}$$

$$S_{\mu\nu}(x) := \frac{i}{2}\bar{\psi}\cdot\Sigma_{\mu\nu}\cdot\psi. \tag{II.4e}$$

Since these 16 real densities have been built up by the four-component wave function ψ, there must exist certain constraints among them, the ‘‘*Fierz identities*.’’[11,12] In compact notation, these relations read

$$j^\mu j_\mu \equiv -\tilde{j}^\mu\tilde{j}_\mu \equiv \rho^2+\tilde{\rho}^2, \tag{II.5a}$$

$$j^\mu\tilde{j}_\mu \equiv 0, \tag{II.5b}$$

$$S_{\mu\nu} \equiv \frac{1}{4}\frac{\tilde{\rho}}{\rho^2+\tilde{\rho}^2}[j_\mu\tilde{j}_\nu - j_\nu\tilde{j}_\mu] - \frac{1}{4}\frac{\rho}{\rho^2+\tilde{\rho}^2}\varepsilon_{\mu\nu\lambda\sigma}j^\lambda\tilde{j}^\sigma. \tag{II.5c}$$

Further relationships between the densities are immediately deduced hereof, e.g., the *vector* identities

$$\tilde{\rho}j_\mu \equiv -4S_{\mu\nu}\tilde{j}^\nu, \tag{II.6a}$$

$$\tilde{\rho}\tilde{j}_\mu \equiv -4S_{\mu\nu}j^\nu, \tag{II.6b}$$

$$\rho\tilde{j}_\mu \equiv -4{}^*S_{\mu\nu}j^\nu, \tag{II.6c}$$

$$\rho j_\mu \equiv -4{}^*S_{\mu\nu}\tilde{j}^\nu, \tag{II.6d}$$

$$({}^*S_{\mu\nu} := \tfrac{1}{2}\varepsilon_{\mu\nu\lambda\sigma}S^{\lambda\sigma}),$$

the *scalar* identities

$${}^*S_{\mu\nu}S^{\mu\nu} \equiv -\tfrac{1}{4}\rho\tilde{\rho}, \tag{II.7a}$$

$$S^{\mu\nu}S_{\mu\nu} = \tfrac{1}{8}(\rho^2-\tilde{\rho}^2); \tag{II.7b}$$

or the *tensor* identities

$$S_{\lambda\mu}S^\lambda_\nu \equiv \tfrac{1}{16}(\rho^2 g_{\mu\nu} + \tilde{j}_\mu\tilde{j}_\nu - j_\mu j_\nu). \tag{II.8}$$

[Hint: Check the consistency of the scalar identity (II.7b) with the tensor case (II.8) by contracting the latter equation with the metric tensor $g^{\mu\nu}$, etc.]. Concerning the number of independent field degrees of freedom, we observe that the nine constraints (II.5a)–(II.5c) for the 16 densities (II.4a)–(II.4e) leave us with seven independent field variables which are represented by the four complex components of the spinor field ψ minus an irrelevant U(1) phase factor dropping out of the bispinor densities. In view of this circumstance, it is convenient to reparametrize the ten densities ρ, $\tilde{\rho}$, j_μ, $\tilde{j}_\mu$ by ten other variables in such a way that the three constraints (II.5a)+(II.5b) are automatically satisfied. Choosing here the ten new variables as the scalar density ρ itself, further some orthonormal tetrad $\{b_\mu, \hat{g}_\mu, \hat{\pi}_\mu, \hat{\lambda}_\mu\}$ with its six degrees of freedom

$$b^\mu b_\mu = -\hat{g}^\mu\hat{g}_\mu = -\hat{\pi}^\mu\hat{\pi}_\mu = -\hat{\lambda}^\mu\hat{\lambda}_\mu = +1, \tag{II.9a}$$

$$b^\mu\hat{g}_\mu = b^\mu\hat{\pi}_\mu = \hat{g}^\mu\hat{\lambda}_\mu = \cdots = 0, \tag{II.9b}$$

and three additional scalar fields z, κ, χ, the reparametrization looks as follows:

$$\tilde{\rho} = z\rho\sinh 2\kappa\sin\chi, \tag{II.10a}$$

$$j_\mu = \rho(\cosh 2\kappa b_\mu + \sinh 2\kappa[\hat{g}_\mu \cdot \cos \chi + \hat{\pi}_\mu (1-z^2)^{1/2} \cdot \sin \chi]), \tag{II.10b}$$

$$\tilde{j}_\mu = -\rho(z \sinh 2\kappa \cdot \cos \chi b_\mu + z \cosh 2\kappa \cdot \hat{g}_\mu + (1-z^2)^{1/2} \cdot \hat{\lambda}_\mu). \tag{II.10c}$$

The polarization density $S_{\mu\nu}$ (II.5c) is then easily found by use of the results (II.10a)–(II.10c) but need not be reproduced here (cf. Ref. 13).

Clearly, the Fierz identities (II.5a)–(II.5c) with all its implications do apply only because the physical densities (II.4a)–(II.4e) have been constructed by means of the wave function ψ. However, the very concept of a ''wave function'' ψ can easily be generalized to the notion of ''intensity matrix'' $\mathcal{I}$ which accounts for the simultaneous presence of several wave functions. In a purely formal respect, this generalization process is mathematically equivalent to the replacement of the state vector $|\psi\rangle$ by the intensity matrix $\hat{\rho}$ in nonrelativistic quantum mechanics.[14] For the present context one observes that the physical densities (II.4a)–(II.4e) can be generated also in the following alternative way: first form the tensor product of the wave function ψ and its Hermitian adjoint $\bar{\psi}$

$$\mathcal{I} = \psi \otimes \bar{\psi} \tag{II.11}$$

and then find the physical densities in terms of the intensity matrix $\mathcal{I}$ (II.11) as

$$\rho = \mathrm{tr}\,(\mathcal{I} \cdot \mathbf{1}), \tag{II.12a}$$

$$\tilde{\rho} = \mathrm{tr}\,(\mathcal{I} \cdot \varepsilon), \tag{II.12b}$$

$$j_\mu = \mathrm{tr}\,(\mathcal{I} \cdot \gamma_\mu), \tag{II.12c}$$

$$\tilde{j}_\mu = i\ \mathrm{tr}\,(\mathcal{I} \cdot \tilde{\gamma}_\mu), \tag{II.12d}$$

$$S_{\mu\nu} = \frac{i}{2}\ \mathrm{tr}\,(\mathcal{I} \cdot \Sigma_{\mu\nu}). \tag{II.12e}$$

Of course the Fierz identities (II.5a)–(II.5c) are still valid here, because both densities (II.4a)–(II.4e) and (II.12a)–(II.12e) are identical, but the identities can now be recast into the compact form

$$\mathcal{I}^2 = \rho \mathcal{I} \tag{II.13}$$

with the intensity matrix $\mathcal{I}$ being given by[15]

$$\mathcal{I} = \tfrac{1}{4}(\rho \cdot \mathbf{1} - \tilde{\rho} \cdot \varepsilon + j_\mu \gamma^\mu - i \tilde{j}_\mu \tilde{\gamma}^\mu + 4 i S_{\mu\nu} \Sigma^{\mu\nu}). \tag{II.14}$$

[Hint: Check the consistency of this new operator form (II.13) for the Fierz identities by action upon the wave function ψ which is an eigenvector of $\mathcal{I}$ with eigenvalue ρ, i.e., $\mathcal{I} \cdot \psi = \rho \psi$].

However, once we are equipped with an intensity matrix $\mathcal{I}$ as a Hermitian operator ($\bar{\mathcal{I}} = \mathcal{I}$) over the fiber space $\mathbb{C}^4$, we can construct the physical densities (II.12a)–(II.12e) also in those cases where $\mathcal{I}$ is built up by more than one single wave function

$$\mathcal{I} = \psi_1 \otimes \bar{\psi}_1 + \psi_2 \otimes \bar{\psi}_2 + \cdots \tag{II.15}$$

($\psi_1, \psi_2, \ldots$ being neither orthogonal nor normalized in general). However, as a consequence of this generalization, the Fierz identity (II.13), or (II.5a)–(II.5c) resp., will no longer apply as it stands. Is it possible that those constraints (II.5a)–(II.5c) survive in some generalized form corresponding to the generalized intensity matrix $\mathcal{I}$ (II.15)? Obviously, it becomes desirable to subdivide the whole set of all possible intensity matrices into different classes, being characterized by some ''degree of order'' N', so that for any $N' \leq N_f$ (presently $N=4$) there exists the corresponding ''*Fierz identity of order* N''' for the physical densities (II.12a)–(II.12e). The desired classification is obtained in a most natural way by identifying the *order parameter* N' with the number of nonzero eigenvalues of the intensity matrix $\mathcal{I}$. Thus the previous case (II.13), from which we started off, has order parameter $N'=1$ (''*pure state*''). For $N'>1$ we have a ''*mixture*.''

For $N'>1$ the first-order identities (II.5a)–(II.5c) are generalized in some way, but whenever some nonzero eigenvalue for a higher-order mixture tends to zero, the physical densities must (at least asymptotically) obey the Fierz identities of the next lower order. In this sense the discrete order parameter N' can undergo a (discontinuous) change. Clearly, such a specific process represents some kind of intrinsic transition of the physical system and must be expected to be accompanied by some singular effects. In order to see this phenomenon in some more detail, we briefly consider the transition of a second-order mixture ($N'=2$) into a pure state ($N'=1$). By the very definition of the second order, there must exist *two* nonzero eigenvalues of $\mathcal{I}$, say ζ_e and η_e:

$$\mathcal{I}\cdot\psi=\zeta_e\psi, \tag{II.16a}$$

$$\mathcal{I}\cdot\phi=\eta_e\phi. \tag{II.16b}$$

[Since $\mathcal{I}$ is (pseudo-)Hermitian, $\mathcal{I}=\overline{\mathcal{I}}$, the eigenvalues ζ_e, η_e are real.] Consequently, the *second-order* generalization of the *first-order* identities (II.13) must read

$$\mathcal{I}^3-(\zeta_e+\eta_e)\mathcal{I}^2+(\zeta_e\cdot\eta_e)\mathcal{I}\equiv 0. \tag{II.17}$$

[Hint: Convince yourself by applying this operator relation to the eigenvalue system (II.16a) and (II.16b).] Now, remember here that the sum of the eigenvalues must always equal the trace ρ (II.12a) of the intensity matrix $\mathcal{I}$, i.e.,

$$\zeta_e+\eta_e\equiv\rho. \tag{II.18}$$

However, this result recasts the second-order identity (II.17) into the following form:

$$\mathcal{I}^3-\rho\mathcal{I}^2+(\zeta_e\cdot\eta_e)\mathcal{I}\equiv 0. \tag{II.19}$$

(For expressing the product of eigenvalues in terms of the physical densities, see below.) Now it may be possible that in some region of space–time (distant past or future, say) one of the eigenvalues tends to zero and the other to the density ρ, according to Eq. (II.18). For such a situation the second-order identity (II.19) leaves us with the truncated form

$$\mathcal{I}^3-\rho\mathcal{I}^2\equiv 0, \tag{II.20}$$

which immediately leads us back to the first-order identity (II.13). Thus we see that the degree of order N' is truly indicated by the corresponding Fierz identities for the physical densities.

After the kinematics of those order transitions has been settled, we next have to inquire whether they are really admitted by the dynamics of the system. (In conventional quantum theory they are forbidden.) In parlicular, one would like to know whether the dynamics equips the mixtures or the pure states with a higher stability. For a Higgs doublet of scalar fields (i.e., $N_f=2$), the pure states have already been shown to be unstable[9] and this leads us to study a similar question for the spinor fields. However, before turning to the general equation of motion, let us first look for a convenient parametrization of the second-order mixtures in order to get some variable measuring the ''distance'' from a pure state.

III. SECOND-ORDER MIXTURES

The most general mixture evidently is encountered when the degree of order (N') coincides with the ''degree of complexity'' (N_f), i.e., when the number of nonzero eigenvalues of $\mathcal{I}$ agrees with the fiber dimension N_f and thus the intensity matrix is regular ($\det\mathcal{I}\neq 0$). If the physical system tends now to adopt all the properties of a pure state, there may exist various routes to ultimately reach such a state. For instance, the order parameter N' could be lowered successively by one unit until the pure state value $N'=1$ is reached, or N' may continue to equal N_f during the ''distance'' from the final pure state shrinks to zero. Conversely, if at the beginning the system is already in a pure state, one wants to know whether the dynamics admits the persistence of such a special state or whether the pure state has to decay into a mixture. In any case we need some measure over the configuration space for those deviations.

For that purpose, we define the *deviation operator* (''deviator''[9]) $\mathfrak{D}_{\mathrm{F}}$ through

$$\mathfrak{D}_{\mathrm{F}}=\rho\mathcal{I}-\mathcal{I}^2, \tag{III.1}$$

so that the pure states are uniquely characterized by the vanishing of $\mathfrak{D}_{\mathrm{F}}$. Like any operator over the typical fiber $\mathbb{C}^N$, $\mathfrak{D}_{\mathrm{F}}$ may be decomposed with respect to the basic operators, i.e., for $N_f=4$,

$$\mathfrak{D}_{\mathrm{F}}=\tfrac{1}{4}(\Delta_{\mathrm{F}}\cdot\mathbf{1}-\tilde{\Delta}_{\mathrm{F}}\cdot\varepsilon+\Delta_\mu\gamma^\mu-i\tilde{\Delta}_\mu\tilde{\gamma}^\mu+4i\Delta_{\mu\nu}\Sigma^{\mu\nu}). \tag{III.2}$$

The deviation densities occurring here are then easily found in terms of the physical densities as

$$\Delta_{\mathrm{F}}\equiv\mathrm{tr}\,(\mathfrak{D}_{\mathrm{F}}\cdot\mathbf{1})=\tfrac{1}{4}(3\rho^2+\tilde{\rho}^{\,2}-j^\mu j_\mu+\tilde{j}^\mu\tilde{j}_\mu-8S^{\mu\nu}S_{\mu\nu}), \tag{III.3a}$$

$$\tilde{\Delta}_{\mathrm{F}}\equiv\mathrm{tr}\,(\mathfrak{D}_{\mathrm{F}}\cdot\varepsilon)=\tfrac{1}{2}(\rho\tilde{\rho}+4{}^*S^{\mu\nu}S_{\mu\nu}), \tag{III.3b}$$

$$\Delta_\mu\equiv\mathrm{tr}\,(\mathfrak{D}_{\mathrm{F}}\cdot\gamma_\mu)=\tfrac{1}{2}(\rho j_\mu+4{}^*S_{\mu\nu}\tilde{j}^\nu), \tag{III.3c}$$

$$\tilde{\Delta}_\mu=i\;\mathrm{tr}\,(\mathfrak{D}_{\mathrm{F}}\cdot\tilde{\gamma}_\mu)=\tfrac{1}{2}(\rho\tilde{j}_\mu+4{}^*S_{\mu\nu}j^\nu), \tag{III.3d}$$

$$\Delta_{\mu\nu}=\tfrac{1}{2}(\rho S_{\mu\nu}-\tilde{\rho}{}^*S_{\mu\nu}+\tfrac{1}{4}\varepsilon_{\mu\nu\lambda\sigma}j^\lambda\tilde{j}^\sigma). \tag{III.3e}$$

Clearly, all the deviation fields Δ_{F}, $\tilde{\Delta}_{\mathrm{F}}$, Δ_μ, $\tilde{\Delta}_\mu$, $\Delta_{\mu\nu}$ must vanish whenever the system is in a pure state, i.e., when the Fierz identities for the physical densities (II.5a)–(II.5c) do apply. [Hint: Check this by use of the implications (II.6a)–(II.8)]. Observe here again that though the vanishing of all the deviation fields (III.3a)–(III.3e) is equivalent to the existence of a pure state, there are further deviation variables not being contained in the set (III.3a)–(III.3e). For instance, it is easily verified that for a pure state also the following deviations have to vanish:

$$\Delta'_\mu:=\tfrac{1}{2}(\tilde{\rho}\tilde{j}_\mu+4S_{\mu\nu}j^\nu), \tag{III.4a}$$

$$\tilde{\Delta}'_\mu:=\tfrac{1}{2}(\tilde{\rho}j_\mu+4S_{\mu\nu}\tilde{j}^\nu), \tag{III.4b}$$

$$Z_{\mu\nu}:=j_\mu j_\nu-\tilde{j}_\mu\tilde{j}_\nu-\rho^2 g_{\mu\nu}+16S_{\mu\lambda}S^\lambda_{\ \nu}. \tag{III.4c}$$

Here the situation is somewhat similar to that for the Fierz identities (II.5a)–(II.5c) which themselves are the necessary and sufficient conditions for the existence of a pure state but nevertheless admit the additional identities (II.6a)–(II.8).

Now we want to restrict ourselves to mixtures of the *second* order. Here the intensity matrix $\mathcal{I}$ obeys the algebraic condition (II.17). Eliminating from that equation $\mathcal{I}^2$ in favor of the deviator $\mathfrak{D}_{\mathrm{F}}$ yields

$$\{\mathcal{I},\mathfrak{D}_{\mathrm{F}}\}=2\zeta_e\eta_e\cdot\mathcal{I}, \tag{III.5}$$

or in coefficient form

$$(\Delta_{\mathrm{F}}-4\zeta_e\eta_e)\rho-\tilde{\rho}\tilde{\Delta}_{\mathrm{F}}+\Delta_\mu j^\mu-\tilde{\Delta}_\mu\tilde{j}^\mu-8\Delta_{\mu\nu}S^{\mu\nu}=0, \tag{III.6a}$$

$$(\Delta_{\mathrm{F}}-4\zeta_e\eta_e)\tilde{\rho}+\rho\tilde{\Delta}_{\mathrm{F}}-{}^8\Delta_{\mu\nu}{}^*S^{\mu\nu}=0, \tag{III.6b}$$

$$(\Delta_{\mathrm{F}}-4\zeta_e\eta_e)j_\mu+\rho\Delta_\mu-4{}^*S_{\mu\nu}\tilde{\Delta}^\nu-4{}^*\Delta_{\mu\nu}\tilde{j}^\nu=0, \tag{III.6c}$$

$$(\Delta_{\mathrm{F}}-4\zeta_e\eta_e)\tilde{j}_\mu+\rho\tilde{\Delta}_\mu-4{}^*S_{\mu\nu}\Delta^\nu-4{}^*\Delta_{\mu\nu}j^\nu=0, \tag{III.6d}$$

$$(\Delta_{\mathrm{F}}-4\zeta_e\eta_e)S_{\mu\nu}+\tilde{\Delta}_{\mathrm{F}}{}^*S_{\mu\nu}+\tilde{\rho}{}^*\Delta_{\mu\nu}+\rho\Delta_{\mu\nu}+\tfrac{1}{4}\varepsilon_{\mu\nu\lambda\sigma}(\tilde{j}^{\,\lambda}\Delta^\sigma-j^\lambda\tilde{\Delta}^\sigma)=0. \tag{III.6e}$$

Finally, after expressing here the product of the eigenvalues in terms of the physical and deviation densities by taking the trace in Eq. (III.5),

$$\Delta_{\mathrm{F}}-4\cdot\zeta_e\eta_e=\frac{1}{\rho}\,(\tilde{\rho}\tilde{\Delta}_{\mathrm{F}}+\Delta_\mu j^\mu-\tilde{\Delta}_\mu\tilde{j}^\mu+8\Delta_{\mu\nu}S^{\mu\nu}) \tag{III.7}$$

then leaves us with a certain set of identities (III.6a)–(III.6e) which are generally valid for all mixtures of the second order. Obviously, these ''*Fierz identities of second order*'' (III.6a)–(III.6e) are nothing else than a set of constraints between the 16 deviation variables Δ_{F}, $\tilde{\Delta}_{\mathrm{F}}$, Δ_μ, $\tilde{\Delta}_\mu$, $\Delta_{\mu\nu}$.

For our subsequent investigation of the transitions between mixtures and pure states we do not only want to restrict ourselves to the present second-order mixtures $N_f=4$ and $N'=2$, but we additionally confine the discussions to those cases where the deviation tensor $\Delta_{\mu\nu}$ (III.3e) vanishes identically: $\Delta_{\mu\nu}\equiv 0$. Here, it immediately follows from the very definition (III.3e) that the polarization density $S_{\mu\nu}$ adopts just that form for a pure state, cf. (II.5c). Clearly the idea is that if transitions from mixtures into pure states are possible, then they should be observable most easily for those situations where the mixture is already as ''close'' as possible to the pure state. Without defining a rigorous measure for this ''closeness,'' one surely may expect that the closer a mixture is to a pure state, the more identities of the first order are satisfied. In this sense, our choice $\Delta_{\mu\nu}\equiv 0$ for a second-order mixture ($N'=2$) has put us already fairly ''close'' to the pure states! The pure states can differ from these specific second-order mixtures ($\Delta_{\mu\nu}\equiv 0$) only through the invalidation of the first two identities (II.5a) and (II.5b)! However, this implies that all the remaining deviation variables can be expressed in terms of three scalar deviation parameters (a_0, $\tilde{a}_1$, and a_2, say).

A convenient choice for these parameters is the following:

$$a_0=\frac{j^\mu j_\mu}{\rho^2+\tilde{\rho}^2}-1, \tag{III.8a}$$

$$\tilde{a}_1=\frac{j^\mu\tilde{j}_\mu}{\rho^2+\tilde{\rho}^2}, \tag{III.8b}$$

$$a_2=\frac{\tilde{j}^\mu\tilde{j}_\mu}{\rho^2+\tilde{\rho}^2}+1. \tag{III.8c}$$

However, the situation becomes even further simplified because the second-order identities (III.6a)–(III.6e) impose the following constraint:

$$a_0\cdot a_2=\tilde{a}_1^{\,2}. \tag{III.9}$$

Consequently, we need only two parameters (a_0,a_2) for rewriting the remaining deviation fields (III.3a)–(III.3e):

$$\Delta_{\mathrm{F}}=\tfrac{1}{2}\rho^2(a_2-a_0), \tag{III.10a}$$

$$\tilde{\Delta}_{\mathrm{F}}=\frac{1}{2}\,\rho\tilde{\rho}(a_2-a_0)=\frac{\tilde{\rho}}{\rho}\,\Delta_{\mathrm{F}}, \tag{III.10b}$$

$$\Delta_\mu=\tfrac{1}{2}\rho(a_2\cdot j_\mu-\tilde{a}_1\cdot\tilde{j}^\mu), \tag{III.10c}$$

$$\tilde{\Delta}_\mu=\frac{1}{2}\,\rho(\tilde{a}_1\cdot j_\mu-a_0\tilde{j}_\mu)=\frac{a_0}{\tilde{a}_1}\,\Delta_\mu\,. \tag{III.10d}$$

Moreover, the product of eigenvalues (III.7) is found in terms of the new parameters as

$$\zeta_e \cdot \eta_e = \tfrac{1}{4}\rho^2(a_2 - a_0) \equiv \tfrac{1}{2}\Delta_{\mathrm{F}}, \tag{III.11}$$

or, similarly, the additional deviation fields (III.4a)–(III.4c) read now

$$\Delta'_\mu = \frac{1}{2}\,\tilde{\rho}(\tilde{a}_1 \cdot j_\mu - a_0 \tilde{j}_\mu) = \frac{\tilde{\rho}}{\rho}\,\tilde{\Delta}_\mu, \tag{III.12a}$$

$$\tilde{\Delta}'_\mu = \frac{1}{2}\,\tilde{\rho}(a_2 \cdot j_\mu - \tilde{a}_1 \cdot \tilde{j}_\mu) = \frac{\tilde{\rho}}{\rho}\,\Delta_\mu, \tag{III.12b}$$

$$Z_{\mu\nu} = a_2 \cdot j_\mu j_\nu + a_0 \tilde{j}_\mu \tilde{j}_\nu - \tilde{a}_1(j_\mu \tilde{j}_\nu + j_\nu \tilde{j}_\mu) + (a_0 - a_2)\rho^2 g_{\mu\nu}. \tag{III.12c}$$

Summarizing, we can say that there are specific mixtures of the second order ($N' = 2$) which are characterized by only two scalar deviation degrees of freedom (i.e., a_0 and a_2) and thus it may be expected that the deviation dynamics (for the transition between mixtures and pure states) will be cut down to an effectively two-dimensional problem. It is very instructive to see in some more detail to what extent the present second-order mixtures differ, if at all, from the pure states. To this end, we start with the observation that all the deviation vectors point into the same four-direction, cf. (III.10d), (III.12a), and (III.12b). Thus, let this common direction be denoted by the timelike unit vector field $b_\mu (\rightarrow b^\mu b_\mu = +1)$,

$$\Delta_\mu = a b_\mu . \tag{III.13a}$$

One directly concludes from the reparametrizations (III.10a)–(III.10d)

$$(\rho^2 + \tilde{\rho}^{\,2}) b_\mu = I \cdot j_\mu - \tilde{I} \cdot \tilde{j}_\mu, \tag{III.13b}$$

where we have put

$$I := b^\mu j_\mu, \tag{III.14a}$$

$$\tilde{I} := b^\mu \tilde{j}_\mu . \tag{III.14b}$$

Now multiply through the new result (III.13b) by b_μ and find the scalar relationship

$$\rho^2 + \tilde{\rho}^{\,2} = I^2 - \tilde{I}^{\,2}. \tag{III.15}$$

On the other hand, look at the pure-state densities (II.10a)–(II.10c) and find that the latter relation (III.15) also holds for those pure states which have $z = 1$. As a consequence, the interrelationship between the present second-order mixtures ($\Delta_{\mu\nu} \equiv 0$) and the $z = 1$ subset of pure states is threefold: both configurations

(i) share the same form of the polarization density $S_{\mu\nu}$, cf. (II.5c),
(ii) are characterized by a timelike unit vector b_μ, and
(iii) obey the same scalar relationship (III.15).

It does not seem possible to find mixtures which have more than these three *structural* elements in common with the pure states! For later purpose, the situation may even be made somewhat more transparent by reformulating the deviation fields in terms of the new elements b_μ and a of (III.13a). Here the new factor a is found to be related to the former deviation parameters (III.8a)–(III.8c) through

$$a = \frac{1}{2}\,\frac{\rho}{\tilde{I}}\,(\rho^2 + \tilde{\rho}^{\,2})\tilde{a}_1 = \frac{1}{2}\,\frac{\rho}{I}\,(\rho^2 + \tilde{\rho}^{\,2}) a_2 = \frac{1}{2}\,\frac{I \cdot \rho}{\tilde{I}^{\,2}}\,(\rho^2 + \tilde{\rho}^{\,2}) a_0, \tag{III.16}$$

and this yields the desired result, namely,

$$\Delta_F = \frac{\rho}{I} a, \tag{III.17a}$$

$$\Delta_\mu = a b_\mu = \frac{I}{\rho} \Delta_F b_\mu, \tag{III.17b}$$

$$\tilde{\Delta}_\mu = \frac{\tilde{I}}{I} \Delta_\mu = \frac{\tilde{I}}{\rho} \Delta_F b_\mu, \tag{III.17c}$$

$$\Delta'_\mu = \frac{\tilde{\rho} \cdot \tilde{I}}{\rho^2} \Delta_F b_\mu, \tag{III.17d}$$

$$\tilde{\Delta}'_\mu = \frac{\tilde{\rho} \cdot I}{\rho^2} \Delta_F b_\mu, \tag{III.17e}$$

$$Z_{\mu\nu} = \frac{2}{\rho^2} \Delta_F [(\rho^2 + \tilde{\rho}^2) b_\mu b_\nu - \rho^2 g_{\mu\nu}]. \tag{III.17f}$$

Indeed, this is a nice result because it says that all the relevant deviation fields are proportional to the scalar deviation Δ_F so that they must tend to zero if only Δ_F does ($\rightarrow$ pure state). As a consequence, we merely have to look for one single dynamical equation (for Δ_F, namely) in order to describe the transition dynamics for mixtures and pure states.

IV. DEVIATION DYNAMICS

Once the kinematical presumptions have been settled, one can now take the next step towards the central question of whether the dynamics will admit the transitions between mixtures of different order N'. Of course the integer N' cannot change continuously but a transition between different orders N' may be understood in the sense that some of the nonzero eigenvalues of the intensity matrix $\mathcal{I}$ become infinitely small (albeit not exactly zero), or conversely, an initially very small eigenvalue raises to the order of magnitude of the others. It must be expected that the answer to the questions of this kind will strongly depend upon the specific dynamics to be applied to the intensity matrix $\mathcal{I}$. The proper choice of dynamics (Sec. VI) has already been attained through some previous work[7,16] and therefore it need not be further discussed here. Instead, let us readily try to transfer the equation of motion for the intensity matrix $\mathcal{I}$ of (I.10) to the general deviator $\mathfrak{D}_F$ of (III.1) and see whether there arises a homogeneous or an inhomogeneous equation for $\mathfrak{D}_F$. Clearly, if we find a homogeneous equation, there exists a permanently pure state ($\mathfrak{D}_F \equiv 0$) as a solution of the equation of motion whereas in the inhomogeneous case the inhomogeneity will act as a force driving the initially pure state ($N' = 1$) into the configuration space of the mixtures $N' > 1$. Evidently this leads to the question of stability of the pure states with respect to their decay into a mixture.

Thus the dynamical equation for the deviator $\mathfrak{D}_F(\mathcal{I})$ in (III.1) is deduced from the original equation of motion (I.10) as

$$\mathcal{D}_\mu \mathfrak{D}_F = \frac{i}{\hbar c} [\mathfrak{D}_F \cdot \overline{\mathcal{H}}_\mu - \mathcal{H}_\mu \cdot \mathfrak{D}_F] + \{\mathcal{L}_\mu, \mathfrak{D}_F\} + \mathfrak{C}_\mu . \tag{IV.1}$$

Here, the Hamiltonian $\mathcal{H}_\mu$ has been split up into its (anti-) Hermitian constituents $\mathcal{K}_\mu$, $\mathcal{L}_\mu$ as usual,

$$\mathcal{H}_\mu = \hbar c (\mathcal{K}_\mu + i \mathcal{L}_\mu), \tag{IV.2}$$

and the inhomogeneous term $\mathfrak{C}_\mu$ on the right of (IV.1) is found in terms of the ''localization field'' $\mathcal{L}_\mu$ as

$$\mathfrak{E}_\mu=[\mathcal{I},\ [\mathcal{I},\ \mathcal{L}_\mu]]+\mathcal{I}\cdot \mathrm{tr}\ \{\mathcal{I},\mathcal{L}_\mu\}-\{\mathcal{I},\mathcal{L}_\mu\}\cdot \mathrm{tr}\ \mathcal{I}. \tag{IV.3}$$

Now, as a consequence of the emergence of the inhomogeneity $\mathfrak{E}_\mu$, it may appear that the permanency of a pure state ($\mathfrak{D}_\mathrm{F}\equiv 0$) would not be admitted by the deviation dynamics (IV.1). However, we have to remember here that the characterization of a pure state by the vanishing of the deviator $\mathfrak{D}_\mathrm{F}(\equiv 0)$ entails further identities, e.g., Eqs. (III.4a)–(III.4c). Therefore one cannot exclude that the inhomogeneous term $\mathfrak{E}_\mu$ in (IV.3) eventually is built up completely by those secondary deviation objects which, in fact, are not contained in the original set of deviation fields (III.3a)–(III.3e), but nevertheless ensure the vanishing of the inhomogeneous term $\mathfrak{E}_\mu$ for the case of a pure state.

Let us inspect this supposition in more detail for Dirac's spinor field ($N_f=4$). Here the $\mathbb{C}(1,3)$-valued localization field $\mathcal{L}_\mu$ is decomposed with respect to the basis operators, mentioned at the beginning of Sec. II, according to

$$\mathcal{L}_\mu=L_\mu\cdot \mathbf{1}+\tilde{L}_\mu\cdot\varepsilon+L_{\lambda\mu}\gamma^\lambda+i\tilde{L}_{\lambda\mu}\tilde{\gamma}^\lambda+iL_{\lambda\sigma\mu}\Sigma^{\lambda\sigma}, \tag{IV.4}$$

where the coefficients L_μ, $\tilde{L}_\mu$, etc. are ordinary tensor fields. However, since the inhomogeneity $\mathfrak{E}_\mu$ (IV.3) is linear with respect to the localization field $\mathcal{L}_\mu$,

$$\mathfrak{E}_\mu=\mathfrak{E}(\mathcal{L}_\mu)=L_\mu\cdot\mathfrak{E}(\mathbf{1})+\tilde{L}_\mu\cdot\mathfrak{E}(\varepsilon)+L_{\lambda\mu}\cdot\mathfrak{E}(\gamma^\lambda)+i\tilde{L}_{\lambda\mu}\cdot\mathfrak{E}(\tilde{\gamma}^\lambda)+iL_{\lambda\sigma\mu}\mathfrak{E}(\Sigma^{\lambda\sigma}), \tag{IV.5}$$

it is sufficient for its vanishing that the value of $\mathcal{E}$ upon any one of the basis operators be zero, i.e.,

$$\mathfrak{E}(\mathbf{1})=\mathfrak{E}(\varepsilon)=\mathfrak{E}(\gamma_\lambda)=\mathfrak{E}(\tilde{\gamma}_\lambda)=\mathfrak{E}(\Sigma_{\lambda\sigma})=0. \tag{IV.6}$$

Indeed, the case for the identity operator is trivial: $\mathfrak{E}(\mathbf{1})\equiv 0$; next for the pseudo-scalar case one finds with a little bit of Clifford algebra and by use of the intensity matrix $\mathcal{I}$ in (II.14),

$$\begin{aligned}\mathfrak{E}(\varepsilon):&=[\mathcal{I},\ [\mathcal{I},\ \varepsilon]]+\mathcal{I}\cdot \mathrm{tr}\ \{\mathcal{I},\varepsilon\}-\{\mathcal{I},\varepsilon\}\cdot \mathrm{tr}\ \mathcal{I}\\ &=\frac{1}{4}\ (j_\lambda j^\lambda-\tilde{j}^\lambda\tilde{j}_\lambda-2(\rho^2+\tilde{\rho}^2))\cdot\varepsilon+\frac{1}{2}\ \tilde{\Delta}'_\mu\cdot\gamma^\mu-\frac{i}{2}\ \Delta'_\mu\cdot\tilde{\gamma}^\mu+4i{}^*\Delta_{\mu\nu}\cdot\Sigma^{\mu\nu}.\end{aligned} \tag{IV.7}$$

However, remember here the first-order identities (II.5a) as well as the vanishing of the deviation fields $\Delta_{\mu\nu}$ [(III.3e)], Δ'_μ [(III.4a)], and $\tilde{\Delta}'_\mu$ [(III.4b)] for a pure state and thus actually find $\mathcal{E}(\varepsilon)\equiv 0$ as required by Eq. (IV.6). Quite similar arguments do apply also to the vector inhomogeneity, e.g.,

$$\begin{aligned}\mathfrak{E}(\gamma^\mu)&:=[\mathcal{I},\ [\mathcal{I},\ \gamma^\mu]]+\mathcal{I}\ \mathrm{tr}\ \{\mathcal{I},\gamma^\mu\}-\{\mathcal{I},\gamma^\mu\}\cdot \mathrm{tr}\ \mathcal{I}\\ &=-\tfrac{1}{2}\ \tilde{\Delta}'^\mu\cdot\varepsilon+\tfrac{1}{4}\ (j^\nu j_\nu-\rho^2-\tilde{\rho}^2)\cdot\gamma^\mu+\tfrac{1}{4}\ Z^\mu_\nu\\ &\quad\cdot\gamma^\nu-\frac{i}{4}\ (j_\nu\cdot\tilde{j}^\nu)\cdot\tilde{\gamma}^\mu-4i{}^*\Delta^\mu_\nu\cdot\tilde{\gamma}^\nu-i\Delta'_\nu\cdot\Sigma^{\mu\nu}-2i\tilde{\Delta}_\nu\cdot{}^*\Sigma^{\nu\mu}.\end{aligned} \tag{IV.8}$$

Observe here again that all the coefficient fields must vanish for a pure state either on account of the first-order identities (II.5a) and (II.5b) or because they directly coincide with one of the deviation fields (III.3a)–(III.3e) or (III.4a)–(III.4c). Thus we actually end up again with $\mathfrak{E}(\gamma^\mu)\equiv 0$ in agreement with the claim (IV.6). Finally, it can be shown by quite similar arguments that the two remaining objects of inhomogeneity $\mathfrak{E}(\tilde{\gamma}^\mu)$ and $\mathcal{E}(\Sigma^{\mu\nu})$ do indeed also vanish for a pure state, but for the sake of brevity the corresponding formulas are not reproduced here. What is more important lies in the fact that the inhomogeneous term $\mathfrak{E}_\mu$ in (IV.3) *always* vanishes for a pure state and therefore $\mathfrak{D}_\mathrm{F}\equiv 0$ becomes an exact solution of the deviation dynamics (IV.1).

However, even if we have to concede now the theoretical persistency of the pure states, it remains to be clarified whether the pure states are stable or unstable configurations. The qualitative computation for a Higgs doublet of scalar fields ($N_f=2$) have shown that the pure state submanifold of the configuration space for the physical densities (i.e., the ''*Fierz cone*'') is equipped with some repulsive potential of singular nature.[9] As a consequence, a pure-state configuration could never be reached exactly. Thus, the repulsive potential appears as the physical counterpart of the mathematical fact which says that an integer (N') cannot change continuously. It is just from this purely formal reason that one tends to believe the repulsive phenomena occurring also in connection with the present spinor fields.

For the subsequent study of this effect we return again to those second-order mixtures described in the preceding section. Moreover, we further simplify the situation by considering a $\Delta_{\mu\nu}=0$ mixture which admitted the introduction of some unit vector field b_μ [(III.13a)] and finally led to that specific form of the deviation fields [see Eqs. (III.17a)–(III.17f)]. As a consequence, the deviator $\mathfrak{D}_\mathrm{F}$ in (III.2) also adopts a very specific shape, namely,

$$\mathfrak{D}_\mathrm{F}=\tfrac{1}{2}\,\Delta_\mathrm{F}\cdot\mathcal{P}_\parallel, \tag{IV.9}$$

and thus becomes proportional to some two-dimensional projector $\mathcal{P}_\parallel$:

$$\mathcal{P}_\parallel=\frac{1}{2}\left(\mathbf{1}-\frac{\tilde{\rho}}{\rho}\,\varepsilon+\frac{I}{\rho}\,\beta-i\,\frac{\tilde{I}}{\rho}\,\tilde{\beta}\right), \tag{IV.10}$$

i.e.,

$$\mathcal{P}_\parallel\cdot\mathcal{P}_\parallel=\mathcal{P}_\parallel, \tag{IV.11a}$$

$$\mathrm{tr}\,\mathcal{P}_\parallel=2. \tag{IV.11b}$$

These projector properties are easily verified by observing the constraint (III.15) together with the fact that the corresponding operators $\{1,\varepsilon,\beta:=b^\mu\gamma_\mu,\tilde{\beta}:=b^\mu\tilde{\gamma}_\mu\}$ obey a very simple subalgebra of $\mathbb{C}(1,3)$, namely,

$$\beta\cdot\beta=\tilde{\beta}\cdot\tilde{\beta}=-\varepsilon\cdot\varepsilon=+\mathbf{1}, \tag{IV.12a}$$

$$\varepsilon\cdot\beta=-\beta\cdot\varepsilon=\tilde{\beta}, \tag{IV.12b}$$

$$\varepsilon\cdot\tilde{\beta}=-\tilde{\beta}\cdot\varepsilon=-\beta, \tag{IV.12c}$$

$$\tilde{\beta}\cdot\beta=-\beta\cdot\tilde{\beta}=\varepsilon. \tag{IV.12d}$$

Clearly the projector $\mathcal{P}_\parallel$ in (IV.10) is nothing else than the identity operator for that subspace in which the intensity matrix $\mathcal{I}$ is living in, i.e.,

$$\mathcal{P}_\parallel\cdot\mathcal{I}=\mathcal{I}. \tag{IV.13}$$

Here, it is a nice check to reassure that all the algebraic properties (including $\Delta_{\mu\nu}\equiv0$) of the present second-order mixture just follow from that compact characterization (IV.13). For instance, insert the projector result for the deviator $\mathfrak{D}_\mathrm{F}$ (IV.9) into the second-order identity (III.5), then observe the projective relation (IV.13) and thus readily verify that the scalar deviation Δ_F just equals twice the product of the eigenvalues $\zeta_e\eta_e$, cf. Eq. (III.11).

However, when the central dynamical equation (I.10) respects the permanency of the pure states, does it also respect the permanency of those second-order mixtures which have vanishing tensor deviation $(\Delta_{\mu\nu}=0)$? Obviously, this question must first be clarified now before we legitimately can restrict ourselves to that subclass of mixtures.

V. FIBER DISTRIBUTIONS

The emergence of the projector $\mathcal{P}_\parallel$ indicates the significance of certain subspaces of the typical fiber $\mathbb{C}^4$ over space–time. More concretely, consider all $\mathbb{C}^4$ elements $\psi(x)$ at any event x of space–time such that

$$\mathcal{P}_\parallel\cdot\psi=\psi. \tag{V.1}$$

Obviously, this set $\{\psi\}_x$ defines some two-dimensional moving subspace of $\mathbb{C}^4$ (''distribution''). Let the orthogonal projector be $\mathcal{P}_\perp$,

$$\mathcal{P}_\parallel + \mathcal{P}_\perp = \mathbf{1}, \tag{V.2a}$$

$$\mathcal{P}_\perp^2 = \mathcal{P}_\perp, \tag{V.2b}$$

$$\mathcal{P}_\perp \cdot \mathcal{P}_\parallel = 0, \tag{V.2c}$$

$$\mathrm{tr}\ \mathcal{P}_\perp = 2; \tag{V.2d}$$

we get a moving (2+2)-splitting of the typical fiber. More generally, a $(N_\parallel + N_\perp)$-splitting is generated by the orthogonal projectors $\mathcal{P}_\parallel(x)$ and $\mathcal{P}_\perp(x)$ such that $\mathrm{tr}\ \mathcal{P}_\parallel + \mathrm{tr}\ \mathcal{P}_\perp = N_\parallel + N_\perp = N_f$, where N_f is the dimension of the typical vector fiber $\mathbb{C}^{N_f}$ in which the wave functions (i.e., sections) are living.

The question with these fiber distributions is now whether they are consistent with the general dynamics, or, more concretely: let the intensity matrix $\mathcal{I}$ be a solution of the general equation of motion (I.10). How then must the field equation for $\mathcal{P}_\parallel$ look in order that those relations as, e.g., (IV.13), do hold over the whole space–time? A possible answer is the following:

$$\mathcal{D}_\mu \mathcal{P}_\parallel = \frac{i}{\hbar c}\left(\mathcal{P}_\parallel \cdot \overline{\mathcal{H}}_\mu - \mathcal{H}_\mu \cdot \mathcal{P}_\parallel\right) - 2 \overset{(\parallel)}{\mathcal{L}}_\mu, \tag{V.3a}$$

$$\mathcal{D}_\mu \mathcal{P}_\perp = \frac{i}{\hbar c}\left(\mathcal{P}_\perp \cdot \mathcal{H}_\mu - \overline{\mathcal{H}}_\mu \cdot \mathcal{P}_\perp\right) + 2 \overset{(\perp)}{\mathcal{L}}_\mu, \tag{V.3b}$$

where the projective localization fields have been defined in a nearby manner, i.e.,

$$\overset{(\parallel)}{\mathcal{L}}_\mu = \mathcal{P}_\parallel \cdot \mathcal{L}_\mu \cdot \mathcal{P}_\parallel, \tag{V.4a}$$

$$\overset{(\perp)}{\mathcal{L}}_\mu = \mathcal{P}_\perp \cdot \mathcal{L}_\mu \cdot \mathcal{P}_\perp. \tag{V.4b}$$

It is a nice exercise to make sure that all the projector properties (IV.11a), (IV.11b), (IV.13), and (V.2a)–(V.2d) are correctly respected by the proposed equations of motion (V.3a) and (V.3b). Furthermore, the necessary bundle identities[8] are also valid:

$$(\mathcal{D}_\mu \mathcal{D}_\nu - \mathcal{D}_\nu \mathcal{D}_\mu)\mathcal{P}_\parallel = [\mathcal{F}_{\mu\nu}, \mathcal{P}_\parallel], \tag{V.5a}$$

$$(\mathcal{D}_\mu \mathcal{D}_\nu - \mathcal{D}_\nu \mathcal{D}_\mu)\mathcal{P}_\perp = [\mathcal{F}_{\mu\nu}, \mathcal{P}_\perp]. \tag{V.5b}$$

Thus we see that if the Hamiltonian $\mathcal{H}_\mu$ admits some fiber splitting (V.3a) and (V.3b), such that $N_\parallel + N_\perp = N_f$, then the dimensions of the corresponding distributions must necessarily remain constant ($\rightarrow N_\perp = \text{const}$, $N_\parallel = \text{const}$). [Hint: Take the traces for the equations (V.3a) and (V.3b)]. However, since an intensity matrix $\mathcal{I}(x)$ of order degree $N'(\equiv N_\parallel)$ always annihilates the complementary fiber space

$$\mathcal{P}_\perp \cdot \mathcal{I} \equiv 0, \tag{V.6}$$

and this constraint must then be valid at any event, we conclude that the mixture's degree of order (N') can *never* be *changed*. In any case, a pure state $(N'=1)$ cannot decay into a mixture $(N' > 1)$, at least in the strict sense which confirms the result of the preceding section. Thus we conjecture that the relativistic Schrödinger theory could eventually admit the transition from a mixture $(N'>1)$ into a pure state $(N'=1)$ only in the asymptotic sense. Observe, however, that the present conclusions heavily rely upon the validity of the splitting dynamics (V.3a) and (V.3b) and, therefore, they readily have to be abandoned if such a fiber distribution does not exist.

In order to keep the situation as simple as possible, we shall restrict ourselves to the second-order mixtures with vanishing tensor deviation. Since this subclass of mixtures has its deviator $\mathfrak{D}_{\mathrm{F}}$

proportional to the projector $\mathcal{P}_\|$, remember (IV.9), we have to make sure that both field equations for the deviator $\mathfrak{D}_F$ in (IV.1) and for the projector $\mathcal{P}_\|$ in (V.3a) are really compatible with each other.

To this end, the special result for the deviator $\mathfrak{D}_F$ in (IV.9) is introduced into its general dynamical equation (IV.1). This first yields for the derivative by use of Eq. (V.3a)

$$\mathcal{D}_\mu \mathfrak{D}_F = \frac{1}{2}(\partial_\mu \Delta_F)\cdot \mathcal{P}_\| - \Delta_F \cdot \overset{(\|)}{\mathcal{L}}_\mu + \frac{i}{\hbar c}(\mathfrak{D}_F \cdot \overline{\mathcal{H}}_\mu - \mathcal{H}_\mu \cdot \mathfrak{D}_F). \tag{V.7}$$

Comparing this now to the original equation (IV.1) puts the inhomogeneous term $\mathfrak{E}_\mu$ into the following form:

$$\mathfrak{E}_\mu = \tfrac{1}{2}(\partial_\mu \Delta_F)\cdot \mathcal{P}_\| - \Delta_F \cdot \overset{(\|)}{\mathcal{L}}_\mu - \tfrac{1}{2}\Delta_F \{\mathcal{L}_\mu, \mathcal{P}_\|\}. \tag{V.8}$$

Now observe here that the trace of the inhomogeneous term $\mathfrak{E}_\mu$ just yields the derivative of the scalar Δ_F, cf. (IV.1),

$$\partial_\mu \Delta_F = 4 \operatorname{tr}(\mathfrak{D}_F \cdot \mathcal{L}_\mu) = 2\Delta_F \operatorname{tr}(\overset{(\|)}{\mathcal{L}}_\mu), \tag{V.9}$$

and thus rewrite the inhomogeneous term $\mathfrak{E}_\mu$ (V.8) in its final form as

$$\mathfrak{E}_\mu = \Delta_F(\mathcal{P}_\| \cdot \operatorname{tr} \overset{(\|)}{\mathcal{L}}_\mu - \overset{(\|)}{\mathcal{L}}_\mu - \tfrac{1}{2}\mathcal{L}_\mu \cdot \mathcal{P}_\| - \tfrac{1}{2}\mathcal{P}_\| \cdot \mathcal{L}_\mu). \tag{V.10}$$

From here it is realized once more that $\mathfrak{E}_\mu$ must vanish for a pure state ($\Delta_F = 0$) [see the arguments below (IV.6)].

However, the significance of Eq. (V.10) rather lies in the fact that it appears to establish some constraint upon the localization field $\mathcal{L}_\mu$. The reason is that the object $\mathfrak{E}_\mu$ has been already defined in terms of $\mathcal{L}_\mu$ through the equation (IV.3) and, therefore, it may be possible that both Eqs. (IV.3) and (V.10) are compatible only if $\mathcal{L}_\mu$ obeys some restrictive condition. This would readily imply the conclusion that our specific second-order mixtures are preserved only when the corresponding Hamiltonian $\mathcal{H}_\mu$ (IV.2) is of some special form. However, we want to stress here that such a constraint upon $\mathcal{H}_\mu$ does not actually exist. In order to verify this claim we merely decompose the localization field $\mathcal{L}_\mu$ into its parallel and orthogonal parts

$$\mathcal{L}_\mu = \overset{(\|)}{\mathcal{L}}_\mu + \overset{(\perp)}{\mathcal{L}}_\mu + \mathcal{P}_\| \cdot \mathcal{L}_\mu \cdot \mathcal{P}_\perp + \mathcal{P}_\perp \cdot \mathcal{L}_\mu \cdot \mathcal{P}_\|, \tag{V.11}$$

and use this decomposition for the compatibility test of both Eqs. (V.10) and (IV.3). This test leads to the following remaining requirement of purely parallel character:

$$[\mathcal{I}, [\mathcal{I}, \overset{(\|)}{\mathcal{L}}_\mu]] + \mathcal{I} \cdot \operatorname{tr}\{\mathcal{I}, \overset{(\|)}{\mathcal{L}}_\mu\} - \{\mathcal{I}, \overset{(\|)}{\mathcal{L}}_\mu\} \cdot \operatorname{tr} \mathcal{I} = -2\Delta_F(\overset{(\|)}{\mathcal{L}}_\mu - \tfrac{1}{2}\mathcal{P}_\| \cdot \operatorname{tr} \overset{(\|)}{\mathcal{L}}_\mu). \tag{V.12}$$

Though this may appear as a rather unwieldy constraint binding together the intensity matrix $\mathcal{I}$ and the parallel part $\overset{(\|)}{\mathcal{L}}_\mu$ of the localization field, it actually is an *identity* (for $N_\| = 2$) which holds for all matrices $\mathcal{I}$ of the type (II.17) with (IV.13) and for *all* localization fields $\mathcal{L}_\mu$. (For a quick check put $\mathcal{L}_\mu$ proportional to the identity operator: $\mathcal{L}_\mu \rightarrow l_\mu \cdot \mathbf{1}$.)

Consequently, the original presumption was correct, i.e., the special second-order mixtures ($\Delta_{\mu\nu} \equiv 0$) with their characteristic deviator $\mathfrak{D}_F$ in (IV.9) are preserved by *any* Hamiltonian $\mathcal{H}_\mu$! Observe that this result is not quite trivial because, in general, it is only the order (N') of the mixture which is preserved in the strict sense during the Hamiltonian space–time evolution, but the preservation of certain specific subclasses (due to the same N') cannot be expected. In any case, the present result provides us with a very convenient possibility to study the (asymptotic) transitions between mixtures and pure states within that specific subclass ($\Delta_{\mu\nu} = 0$) characterized by only one single deviation variable (Δ_F).

VI. HAMILTONIAN DYNAMICS

The preceding result provides us with the freedom to choose an arbitrary Hamiltonian $\mathcal{H}_\mu$ being subject only to the field equations (I.7) and (I.8). Clearly one will select an object $\mathcal{H}_\mu$ whose properties have already been studied in sufficient detail:[16]

$$\frac{1}{\hbar c}\cdot\mathcal{H}_\mu=\frac{m}{4}\gamma_\mu+\frac{3}{2}ib_\mu(N\cdot\mathbf{1}-\tilde{N}\cdot\varepsilon)+(4b_\mu b_\lambda-g_{\mu\lambda})(W\cdot\mathbf{1}+\tilde{W}\cdot\varepsilon)\gamma^\lambda$$
$$-ib^\lambda(N\cdot\Sigma_{\mu\lambda}+\tilde{N}\cdot{}^*\Sigma_{\mu\lambda}). \tag{VI.1}$$

This Hamiltonian is due to a Robertson–Walker (RW) universe whose Hubble flow vector b_μ is thus identified with the present deviation vector (III.13a). Furthermore, m is the rescaled mass $M(:=\hbar m/c)$ and, according to the RW symmetry, the four complex scalar fields $W,\tilde{W},N,\tilde{N}$ are homogeneous, i.e., they exclusively depend upon the cosmic time θ $(\partial_\mu\theta\equiv b_\mu)$.

Introducing the present ansatz for $\mathcal{H}_\mu$ (VI.1) into the field equations (I.7) and (I.8) must yield the field equations for those four ansatz scalars $W,\tilde{W},N,\tilde{N}$, i.e.,

$$b^\mu\partial_\mu W:=\dot{W}=-H\cdot\left(W-\frac{m}{4}\right)-(N+H)\left(3W+\frac{m}{4}\right), \tag{VI.2a}$$

$$\dot{\tilde{W}}=-H\cdot\tilde{W}-3(N+H)\cdot\tilde{W}, \tag{VI.2b}$$

$$\dot{\tilde{N}}=-H\cdot\tilde{N}, \tag{VI.2c}$$

$$\dot{N}+\dot{H}+\frac{\sigma}{\mathcal{R}^2}=N\cdot(N+H)+16W^2+16\tilde{W}^2-4mW-\tilde{N}^2. \tag{VI.2d}$$

However, besides these dynamical equations one also finds some algebraic constraints for the ansatz fields:

$$\tilde{N}\cdot\tilde{W}=0, \tag{VI.3a}$$

$$\tilde{N}\cdot\left(W-\frac{m}{4}\right)=0, \tag{VI.3b}$$

$$\tilde{N}\cdot(N+H)=0, \tag{VI.3c}$$

$$\frac{\sigma}{\mathcal{R}^2}=(N+H)^2+4\left(W-\frac{m}{4}\right)^2+4\tilde{W}^2-\tilde{N}^2. \tag{VI.3d}$$

Here, the scale parameter of the RW line-element has been denoted by $\mathcal{R}$ (''radius of the universe''), $H(\equiv\dot{\mathcal{R}}/\mathcal{R})$ is the Hubble expansion rate, and σ is the topological index ($\sigma=+1$: open universe; $\sigma=0$: flat; $\sigma=-1$: closed). Perhaps the most striking feature of the present Hamiltonian $\mathcal{H}_\mu$ in (VI.1) is the fact that it falls apart into two subsets, which are rather different from the topological point of view, according to whether the ansatz scalar $\tilde{N}$ does vanish or not, cf. Ref. 16.

However, for the present context it is more important to consider the splitting of $\mathcal{H}_\mu$ into its (anti-)Hermitian parts $\mathcal{K}_\mu$, $\mathcal{L}_\mu$, of (IV.2) because the localization field $\mathcal{L}_\mu$ explicitly governs the deviation dynamics [see Eq. (V.7)]. The kinetic field $\mathcal{K}_\mu$ as the Hermitian part of the Hamiltonian $\mathcal{H}_\mu$ of (VI.1) is found as

$$\mathcal{K}_\mu=\frac{m}{4}\gamma_\mu+\frac{3}{2}N_r b_\mu\cdot\mathbf{1}-\frac{3}{2}\tilde{N}_r b_\mu\cdot\varepsilon+(4b_\mu b_\lambda-g_{\mu\lambda})(W_r\cdot\mathbf{1}+i\tilde{W}_c\cdot\varepsilon)\gamma^\lambda$$
$$-ib^\lambda(N_c\cdot\Sigma_{\mu\lambda}+\tilde{N}_c\cdot{}^*\Sigma_{\mu\lambda}) \tag{VI.4}$$

and similarly for the anti-Hermitian part:

$$\mathcal{L}_\mu = \frac{3}{2} N_c b_\mu \cdot \mathbf{1} - \frac{3}{2} \tilde{N}_c b_\mu \cdot \varepsilon + (4 b_\mu b_\lambda - g_{\mu\lambda})(W_c \cdot \gamma^\lambda - i \tilde{W}_r \tilde{\gamma}^\lambda) + i b^\lambda (N_r \cdot \Sigma_{\mu\lambda} + \tilde{N}_r \cdot {}^*\Sigma_{\mu\lambda}). \tag{VI.5}$$

Observe here that the ansatz scalars have been split up into their real and imaginary parts according to

$$W = W_r + i W_c , \tag{VI.6a}$$

$$\tilde{W} = \tilde{W}_r + i \tilde{W}_c , \tag{VI.6b}$$

$$N = N_c - i N_r , \tag{VI.6c}$$

$$\tilde{N} = \tilde{N}_c - i \tilde{N}_r , \tag{VI.6d}$$

and that the Spin(1,3) generators $\Sigma_{\mu\nu} = \frac{1}{4}[\gamma_\mu , \gamma_\nu]$ are anti-Hermitian ($\overline{\Sigma}_{\mu\nu} = -\Sigma_{\mu\nu}$) as well as the axial velocity operator $\tilde{\gamma}_\lambda := \varepsilon \gamma_\lambda = -\bar{\tilde{\gamma}}_\lambda$.

Once the Hamiltonian $\mathcal{H}_\mu$ is known, one can now study the dynamics for the deviation scalar Δ_{F} (V.9). However, the information carried by Δ_{F} (with respect to the existence of a pure state) can be improved by relating it to the scalar density ρ. As the physical densities (II.10a)–(II.10c) are demonstrating clearly, the scalar ρ essentially determines their absolute magnitudes; and since the deviation fields (III.3a)–(III.3e) are certain combinations of these physical densities, their magnitude will also be determined by the scalar ρ. Thus if we want to have an adequate measure for the presence of a pure state, it is better to consider the *reduced* deviation field $\Delta_{\mathrm{F}}/\rho^2$. Therefore, let us combine the derivative for Δ_{F} of (V.9) with that of ρ,

$$\partial_\mu \rho = \mathrm{tr}\, (\mathcal{D}_\mu \mathcal{I}) = 2\, \mathrm{tr}\, (\mathcal{I} \cdot \mathcal{L}_\mu), \tag{VI.7}$$

to find

$$\partial_\mu \left(\frac{\Delta_{\mathrm{F}}}{\rho^2} \right) = \frac{4}{\rho^2}\, \mathrm{tr} \left[\left(\mathfrak{D}_{\mathrm{F}} - \frac{\Delta_{\mathrm{F}}}{\rho} \mathcal{I} \right) \cdot \mathcal{L}_\mu \right]. \tag{VI.8}$$

Introducing here the intensity matrix $\mathcal{I}$ of (II.14), the specific deviation fields (III.17a)–(III.17f), and the present localization field $\mathcal{L}_\mu$ (VI.5) then yield

$$\partial_\mu \left(\frac{\Delta_{\mathrm{F}}}{\rho^2} \right) = \frac{4}{\rho} \left(\frac{\Delta_{\mathrm{F}}}{\rho^2} \right) [-W_c (I b_\mu - j_\mu) + \tilde{W}_r (\tilde{I} b_\mu - \tilde{j}_\mu) + 2 N_r b^\lambda S_{\lambda\mu} + 2 \tilde{N}_r b^\lambda {}^*S_{\lambda\mu}]. \tag{VI.9}$$

This is the desired result because, if we succeed in integrating this equation, we can judge in which region of space–time there prevails a pure state ($\Delta_{\mathrm{F}}/\rho^2 \ll 1$) and where one has a mixture ($\Delta_{\mathrm{F}}/\rho^2 \sim 1$). A special property of the result (VI.9) consists in the fact that the reduced deviation does not change with cosmic time θ. Indeed, multiply through the equation (VI.9) by the Hubble flow b^μ and find

$$b^\mu \partial_\mu \left(\frac{\Delta_{\mathrm{F}}}{\rho^2} \right) = \frac{d}{d\theta} \left(\frac{\Delta_{\mathrm{F}}}{\rho^2} \right) \equiv 0. \tag{VI.10}$$

Clearly, this time independence of the reduced deviation $\Delta_{\mathrm{F}}/\rho^2$ is due to our special assumptions concerning the choice of Hamiltonian $\mathcal{H}_\mu$, and will not apply in the general case. However, on account of the time independence, one can concentrate now upon the study of the purely spatial variations of $\Delta_{\mathrm{F}}/\rho^2$.

VII. SO(3) SYMMETRIC MIXTURES

The most striking difference between mixtures and pure states lies in the fact that certain configurations of the physical densities cannot be generated by a pure state but are available only by a mixture. This circumstance becomes especially clear in the context of certain symmetries of the physical densities. For instance, in the present case of spinning (Dirac) matter, one would expect that spin effects should show up in (at least some of) the physical densities which therefore could not be isotropic. The cosmological principle, requiring strict isotropy and homogeneity, should therefore not apply to all of the physical densities carried by spinning matter.[17] However, even if one admits some lower symmetry, the mixtures develop a higher potentiality for realizing those symmetries.

As an example, return for a moment to the former relationship (III.13b) which establishes a linear dependence of three vector fields $\{b_\mu, j_\mu \tilde{j}_\mu\}$ and also implies the scalar relationship (III.15). However, though the latter scalar constraint does apply also for the pure state densities (II.10a)–(II.10c) (for $z=1$), the former vector relationship (III.13b) does not hold in general for the pure-state case (II.10a)–(II.10c)! Consequently, the immediate conclusions, to be drawn from that vector relationship (III.13b), cannot be obeyed by the pure-state configurations. We now demonstrate that among those properties, not being available for the pure states, is just the SO(3) symmetry of certain physical densities. Clearly, one furthermore needs the breaking of homogeneity down to (e.g.) the SO(3) symmetry in order to demonstrate the spatial variations of the reduced deviation Δ_F/ρ^2 as a measure for the ''distance'' between mixtures and pure states.

The point of departure for this demonstration is the observation that the crucial vector relation (III.13b) admits the introduction of a space–like vector $\tilde{c}_\mu$:

$$\tilde{c}_\mu = \frac{I}{\rho^2+\tilde{\rho}^2}\cdot \tilde{j}_\mu - \frac{\tilde{I}}{\rho^2+\tilde{\rho}^2}\cdot j_\mu . \tag{VII.1}$$

Evidently this vector field is orthogonal to the Hubble flow b_μ,

$$b^\mu \cdot \tilde{c}_\mu \equiv 0, \tag{VII.2}$$

and is found to be of length $\tilde{c}$

$$\tilde{c}^{\,2} := -\tilde{c}^\mu \tilde{c}_\mu = 1 - 2\,\frac{\Delta_F}{\rho^2}, \tag{VII.3}$$

when the properties of the second-order mixtures (Sec. III) are used. Thus the values of the reduced deviation are restricted to the following range:

$$0 \leqslant \frac{\Delta_F}{\rho^2} \leqslant \frac{1}{2}. \tag{VII.4}$$

Indeed, this result is now the crucial point for the emergence of new spatial structures which are *forbidden for the pure states* $(\Delta_F \equiv 0)$. The reason is that for $\Delta_F \to 0$ the vector $\tilde{c}_\mu$ (VII.1) becomes a unit vector which, by its very definition, can nowhere vanish. However, the existence of a zero of the vector field $\tilde{c}_\mu$ is *necessary*, e.g., for the case of a rotationally symmetric configuration, otherwise $\tilde{c}_\mu$ would not be differentiable at the symmetry center. However, this is *no problem for a mixture* because, in view of Eq. (VII.3), one merely has to demand for the center $(r=0)$

$$\lim_{r\to 0} \frac{\Delta_F}{\rho^2} = \frac{1}{2} \tag{VII.5}$$

and the vector field $\tilde{c}_\mu$ becomes zero there as desired.

In what follows we want to elaborate this point in some more detail but, in order to see the essential point, it will be sufficient to select the most simple Hamiltonian from the set (VI.1), namely the case of nonvanishing scalar $\tilde{N}$ for an open RW universe (i.e., $\sigma=+1$). For this presumption the Hamiltonian constraint (VI.3d) yields

$$\frac{1}{\mathcal{R}^2}=-\tilde{N}^2=-\tilde{N}_c^2+\tilde{N}_r^2+2i\tilde{N}_c\cdot\tilde{N}_r\,. \tag{VII.6}$$

Thus $\tilde{N}_c$ must be zero and the derivative of the scalar density ρ (VI.7), with the localization field $\mathcal{L}_\mu$ being given by (VI.5), tells us that we will then get some nontrivial spatial structure over the time slices $\theta=$const. Consequently, the choice for the localization field $\mathcal{L}_\mu$ is the following:

$$\mathcal{L}_\mu\rightarrow{}^{(\sim)}\mathcal{L}_\mu=-\tfrac{3}{2}Hb_\mu\cdot\mathbf{1}+i\tilde{N}_r\cdot{}^*\Sigma_{\mu\lambda}b^\lambda, \tag{VII.7}$$

and the corresponding kinetic field $\mathcal{K}_\mu$ (VI.4) becomes

$$\mathcal{K}_\mu\rightarrow{}^{(\sim)}\mathcal{K}_\mu=-\tfrac{3}{2}\tilde{N}_rb_\mu\cdot\varepsilon+mb_\mu\cdot\beta+iH\cdot\Sigma_{\mu\lambda}b^\lambda. \tag{VII.8}$$

Next, the RW homogeneity on the time slices must be broken down to SO(3) symmetry. This is attained most conveniently by rewriting the RW line-element in terms of ''polar coordinates'' $\{r,\vartheta,\phi\}$ (Ref. 18) as

$$ds^2=d\theta^2-\mathcal{R}^2(dr^2+\sinh^2\, r(d\vartheta^2+\sin^2\,\vartheta\;d\phi^2)) \tag{VII.9}$$

such that the origin $(r=0)$ becomes the symmetry center. The radial coordinate r may be considered as a scalar field over space–time, whose unit normal r_μ is then given by[17]

$$r_\mu=\mathcal{R}\cdot\partial_\mu r, \tag{VII.10a}$$

$$r^\mu r_\mu=-1, \tag{VII.10b}$$

$$b^\mu r_\mu=0. \tag{VII.10c}$$

Clearly the coordinate vector r_μ is an SO(3) symmetric vector field and obeys the field equation[17]

$$\nabla_\mu r_\nu=\frac{1}{\mathcal{R}\cdot\tanh\, r}\,(g_{\mu\nu}-b_\mu b_\nu+r_\mu r_\nu)-Hb_\nu r_\mu\,. \tag{VII.11}$$

Therefore, if the physical densities (II.12a)–(II.12e) are coined SO(3) symmetric by the present Hamiltonian ${}^{(\sim)}\mathcal{H}_\mu=\hbar c({}^{(\sim)}\mathcal{K}_\mu+i{}^{(\sim)}\mathcal{L}_\mu)$ via the relativistic von-Neumann equation (I.10), then the spacelike vector field $\tilde{c}_\mu$ (VII.1) must obey the SO(3) symmetry condition

$$\nabla_\nu\tilde{c}_\mu=\tilde{N}_r(g_{\mu\nu}-b_\mu b_\nu+\tilde{c}_\mu\tilde{c}_\nu)-Hb_\mu\tilde{c}_\nu \tag{VII.12a}$$

or, resp., for its normalized version $\hat{c}_\mu(:=\tilde{c}^{-1}\cdot\tilde{c}_\mu,\hat{c}^\mu\hat{c}_\mu=-1)$:

$$\nabla_\nu\hat{c}_\mu=\tilde{N}_r\cdot\frac{g_{\mu\nu}-b_\mu b_\nu+\hat{c}_\mu\hat{c}_\nu}{\tilde{c}}-Hb_\mu\hat{c}_\nu\,. \tag{VII.12b}$$

Here, the origin of the symmetry claim (VII.12a) lies in the observation that the spatial derivatives of the physical densities are governed mainly by the vector field $\tilde{c}_\mu$. However, it is then immediately obvious that the total density configuration will be SO(3) symmetric if $\tilde{c}_\mu$ obeys that symmetry condition (VII.12a) (observe that the geometric background remains RW symmetric). In order to present a few examples, consider first the reduced deviation $\Delta_{\mathrm{F}}/\rho^2$ (VI.9). Since our choice of Hamiltonians (VII.7) and (VII.8) lets only the ansatz scalar $\tilde{N}_r$ survive, the derivative (VI.9) is first simplified to

$$\partial_\mu\left(\frac{\Delta_F}{\rho^2}\right)=\frac{8}{\rho}\frac{\Delta_F}{\rho^2}\,\tilde{N}_r b^{\lambda *}S_{\lambda\mu}\,. \tag{VII.13}$$

But on the other hand, the former second-order identities (III.6a)–(III.6e) imply

$$b^{\lambda *}S_{\lambda\mu}=\tfrac{1}{4}\rho\tilde{c}_\mu\,, \tag{VII.14a}$$

$$b^{\lambda}S_{\lambda\mu}=\tfrac{1}{4}\tilde{\rho}\tilde{c}_\mu\,, \tag{VII.14b}$$

so that the polarization density $S_{\mu\nu}$ of (II.5c) becomes

$$S_{\mu\nu}=\tfrac{1}{4}\tilde{\rho}(b_\mu\tilde{c}_\nu-b_\nu\tilde{c}_\mu)-\tfrac{1}{4}\rho\varepsilon_{\mu\nu\lambda\sigma}b^\lambda\tilde{c}^\sigma. \tag{VII.15}$$

Thus the derivative of the reduced deviation (VII.13) is actually found proportional to the vector $\tilde{c}_\mu$,

$$\partial_\mu\left(\frac{\Delta_F}{\rho^2}\right)=2\,\frac{\Delta_F}{\rho^2}\,\tilde{N}_r\tilde{c}_\mu\,, \tag{VII.16}$$

as expected. Further examples refer to the case of the scalar density ρ of (VI.7), pseudo-scalar $\tilde{\rho}$ of (II.12b), or the other scalar products I of (III.14a) and $\tilde{I}$ of (III.14b):

$$\partial_\mu\rho=-3H\rho b_\mu-\tilde{N}_r\rho\tilde{c}_\mu\,, \tag{VII.17a}$$

$$\partial_\mu\tilde{\rho}=-(2m\tilde{I}+3H\tilde{\rho})b_\mu-\tilde{N}_r\tilde{\rho}\tilde{c}_\mu\,, \tag{VII.17b}$$

$$\partial_\mu I=-3(HI+\tilde{N}_r\tilde{I})b_\mu-\tilde{N}_r I\tilde{c}_\mu\,, \tag{VII.17c}$$

$$\partial_\mu\tilde{I}=-3(H\tilde{I}+\tilde{N}_r I-\tfrac{2}{3}m\tilde{\rho})b_\mu-\tilde{N}_r\tilde{I}\tilde{c}_\mu\,. \tag{VII.17d}$$

Moreover, the space parts of both current densities j_μ of (II.12c) and $\tilde{j}_\mu$ of (II.12d) also point to the direction of $\tilde{c}_\mu$:

$$j_\mu=Ib_\mu+\tilde{I}\tilde{c}_\mu\,, \tag{VII.18a}$$

$$\tilde{j}_\mu=\tilde{I}b_\mu+I\tilde{c}_\mu\,, \tag{VII.18b}$$

which is easily found by inverting both Eqs. (III.13b) for b_μ and (VII.1) for $\tilde{c}_\mu$. Thus, the whole matter configuration is safely SO(3) symmetric if only the SO(3) symmetry condition (VII.12a) upon the vector field $\tilde{c}_\mu$ is valid. But observe again here that the vector and tensor objects (VII.15) and (VII.18a) and (VII.18b) can be nonsingular at the symmetry center $r=0$ only for a mixture $(\Delta_F\neq 0)$ [see the arguments following (VII.4)].

However, the desired symmetry is readily established by explicit differentiation. It may be sufficient to demonstrate the technique for the current-density j_μ. Differentiating the defining equation (II.12c) yields first

$$\nabla_\nu j_\mu=\mathrm{tr}\,((\mathcal{D}_\nu\mathcal{I})\cdot\gamma_\mu) \tag{VII.19}$$

or, by use of the relativistic von-Neumann equation (I.10),

$$\nabla_\nu j_\mu=\frac{i}{\hbar c}\,\mathrm{tr}\,(\mathcal{I}[\overline{\mathcal{H}}_\nu\cdot\gamma_\mu-\gamma_\mu\cdot\mathcal{H}_\nu])=i\,\mathrm{tr}\,(\mathcal{I}\cdot[\mathcal{K}_\nu,\gamma_\mu])+\mathrm{tr}\,(\mathcal{I}\cdot\{\mathcal{L}_\nu,\gamma_\mu\}). \tag{VII.20}$$

Then, introducing here the adopted form for the kinetic field $\mathcal{K}_\mu$ (VII.8) and localization field $\mathcal{L}_\mu$ (VII.7) finally leads to the result

$$\nabla_\nu j_\mu = 2m\tilde{\rho} b_\nu \tilde{c}_\mu + (\tilde{N}_r \cdot \tilde{I} + H \cdot I) g_{\mu\nu} - 3(\tilde{N}_r \cdot \tilde{j}_\mu + H \cdot j_\mu) b_\nu - (\tilde{N}_r \cdot \tilde{j}_\nu + H \cdot j_\nu) b_\mu . \tag{VII.21}$$

As a consistency check one immediately verifies here the validity of the continuity equation

$$\nabla^\mu j_\mu = 0, \tag{VII.22}$$

or multiply through the equation (VII.23) with the Hubble flow b_μ, whose derivative reads

$$\nabla_\mu b_\nu = H B_{\mu\nu} \equiv H(g_{\mu\nu} - b_\mu b_\nu), \tag{VII.23}$$

and just find the equation (VII.17c) for the derivative of the scalar product $I(\equiv j^\mu b_\mu)$.

In any case, with this technique of computing the derivatives of the physical densities one can actually verify the correctness of the SO(3) symmetry condition (VII.12a). In the next step, one then wants to see the dependence of the reduced deviation Δ_F/ρ^2 upon the radial variable r (VII.9). For that purpose, the corresponding derivative (VII.16) is recast into the following form:

$$\hat{c}^\mu \partial_\mu \left(\frac{\Delta_F}{\rho^2} \right) \equiv \frac{1}{\mathcal{R}} \frac{d}{dr} \left(\frac{\Delta_F}{\rho^2} \right) = -2\tilde{N}_r \cdot \frac{\Delta_F}{\rho^2} \cdot \tilde{c}(r). \tag{VII.24}$$

On the other hand, the length $\tilde{c}(r)$ of the surface vector $\tilde{c}_\mu$ depends upon the reduced deviation as is shown in Eq. (VII.3) so that we end up with a closed equation for Δ_F/ρ^2.

$$\frac{d}{dr} \left(\frac{\Delta_F}{\rho^2} \right) = -2\mathcal{R}\tilde{N}_r \frac{\Delta_F}{\rho^2} \sqrt{1 - 2\frac{\Delta_F}{\rho^2}}. \tag{VII.25}$$

Putting here [in agreement with the Hamiltonian constraint (VII.6)]

$$\mathcal{R}\tilde{N}_r = +1, \tag{VII.26}$$

the obvious solution of (VII.25) for the reduced deviation is found as

$$\frac{\Delta_F}{\rho^2} = \frac{1}{2} (1 - \tanh^2 r) = \frac{1}{2\cosh^2 r}. \tag{VII.27}$$

Since this solution just obeys the smoothness condition (VII.5) for the radially symmetric tensor objects [e.g., $j_\mu, \tilde{j}_\mu$ (VII.18a) and (VII.18b) or $S_{\mu\nu}$ (VII.15)], it seems as if one would have found the desired SO(3) symmetric mixture configuration which in the outside region ($r \to \infty$) becomes a pure state ($\Delta_F/\rho^2 \to 0$).

However, the situation here is not quite so favorable because we applied a certain type of Hamiltonian $\mathcal{H}_\mu$ (VII.7) and (VII.8) which originally was designed for cosmological purposes (Ref. 19). For such a situation, the cosmological principle demands homogeneity and isotropy of the matter distribution, which is just the contrary of a well-localized distribution of the physical densities, as it is desired for the present context. Therefore one should not be surprised that the spatial concentration of the reduced deviation (VII.27) is actually spurious, namely in the sense that the scalar deviation Δ_F is constant and the scalar density ρ is increasing exponentially in the outside region ($r \to \infty$). These assertions can immediately be checked by explicitly solving the field equation for the scalar density ρ (III.17a) to yield

$$\rho(\theta, r) = \rho_{in} \left(\frac{\mathcal{R}_{in}}{\mathcal{R}(\theta)} \right)^3 \cosh r, \tag{VII.28}$$

which implies the spatial constancy for the scalar deviation Δ_F:

$$\Delta_F = \frac{1}{2} \rho_{in}^2 \left(\frac{\mathcal{R}_{in}}{\mathcal{R}(\theta)} \right)^6 . \tag{VII.29}$$

However, this drawback of the present model should not be taken too seriously. It is expected that the unboundedness of the physical densities will be removed by considering a closed universe ($\sigma=-1$) or by applying some Hamiltonian due to an attractive potential. Therefore, despite its deficiencies, the results for the present model may be understood as a distinct hint upon the existence of well-localized mixtures which become pure states in the outside region ($\Delta_F/\rho^2 \to 0$).

[1] S. Hawking and R. Penrose, *The Nature of Space and Time* (Princeton U.P., Princeton, NJ, 1996).
[2] E. Schrödinger, Naturwissenschaften **23**, 48–50 (1935).
[3] M. Silvermann, *More Than One Mystery. Explorations in Quantum Interference* (Springer, New York, 1995).
[4] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1955).
[5] R. Omnès, Rev. Mod. Phys. **64**, 339 (1992).
[6] F. Selleri, *Quantum Paradoxes and Physical Reality* (Kluwer, Dordrecht, 1990).
[7] M. Sorg, *Relativistic Schrödinger Equations*, preprint Stuttgart (1992).
[8] M. Sorg, Nuovo Cimento B **112**, 23 (1997).
[9] M. Mattes and M. Sorg, Int. J. Theor. Phys. **36**, 395 (1997).
[10] G. C. Ghirardi, A. Rimini, and T. Weber, Nuovo Cimento B **31**, 177 (1976); *ibid.* **36**, 97 (1976).
[11] M. Fierz, Z. Phys. **104**, 553 (1937).
[12] J. P. Crawford, J. Math. Phys. **26**, 1439 (1985).
[13] M. Mattes and M. Sorg, J. Phys. A **26**, 3013 (1993).
[14] See any textbook about advanced quantum mechanics, e.g., A. Messiah, *Quantum Mechanics* (North-Holland Amsterdam 1965), Vols. I and II.
[15] M. Mattes and M. Sorg, Nuovo Cimento B **109**, 1097 (1994).
[16] M. Sorg, J. Phys. A **30**, 5517 (1997).
[17] M. Sorg, Lett. Math. Phys. **33**, 113 (1995).
[18] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[19] U. Ochs and M. Sorg, J. Phys. A **28**, 7263 (1995).

8*D* oscillator and 5*D* Kepler problem: The case of nontrivial constraints

M. V. Pletyukhov[a)] and E. A.Tolkachev[b)]
Institute of Physics, National Academy of Sciences of Belarus, Minsk 220072 Scorina av. 70, Belarus

(Received 29 December 1997; accepted for publication 22 September 1998)

The topologically nontrivial correspondence between the $5D$ $SU(2)$-Kepler problem and the $8D$ singular oscillator is considered. It is shown that both ''isospinor'' and ''isovector'' particles on the $SU(2)$ instanton + $1/R$ background in $5D$ can be described in terms of an $8D$ singular oscillator. The energy spectrum is calculated for an arbitrary ''isospin'' case. © *1999 American Institute of Physics.* [S0022-2488(99)01301-8]

I. INTRODUCTION

Nonbijective bilinear transformations of Levi-Civita (LCT) $R^2 \to R^2$, Kustaanheimo–Stiefel (KST) $R^4 \to R^3$, Hurwitz (HT) $R^8 \to R^5$ and the transformation $R^{16} \to R^9$, their extensions and applications have been the objects of numerous studies (e.g., see Refs. 1–4 and references therein). These transformations constitute a series of mappings $R^{2p} \to R^{p+1}$ which possess the property $r_\mu r_\mu = (x_m x_m)^2$ for $r_\mu \in R^{p+1}$, $\mu = 0, \ldots, p$; $x_m \in R^{2p}$, $m = 1, \ldots, 2p$; $p = 2^q = 1,2,4,8$. They are closely connected with division algebras of real, complex numbers, quaternions and octonions. The LCT, KST and HT are inherent to the Hopf fibrations, namely real $S^1/Z_2 = S^1$, complex $S^3/S^1 = S^2$ and quaternion $S^7/S^3 = S^4$, respectively.

In particular, these transformations are exploited to establish the correspondence between oscillator and Kepler problems, both classical and quantum, in the respective dimensions.[5] Moreover, using KST it has been shown (e.g., see Refs. 6,7) that the Schrödinger equation for a $4D$ isotropic oscillator with a constraint is equivalent to that for a $3D$ particle on the background of the $U(1)$ Dirac monopole+Kepler (Coulomb) $1/R$+extra $1/R^2$. The latter is known as MIC-Kepler problem.[8] Introducing the corresponding singular term into the $4D$ equation one can remove the extra centrifugal—Zwanziger's[9]—potential in $3D$ and the equation for the dyogen atom arises.[10] Its energy spectrum and eigenfunctions have been obtained on the basis of the dynamical symmetries of the $4D$ singular isotropic oscillator.[11]

As for HT, it establishes the relation between an $8D$ oscillator and $5D$ Kepler problems. In the quantum case it is usually supposed[5,12,13] that the corresponding wavefunctions are related as $\Psi^{(8)}(x_m) = \Psi^{(5)}(r_\mu)$. Thus, the Hilbert space of the $8D$ oscillator appears to be in one-to-one correspondence with that of a $5D$ problem. A geometric quantization procedure for this case has been developed in Ref. 14.

Recently attempts to consider a topologically nontrivial version imposing a less strict constraint on $\Psi^{(8)}(x_m)$ have been made.[15,16] The results obtained allow us to describe a $5D$ ''isospin'' particle on the background of $SU(2)$ instanton+Kepler $1/R$+extra $1/R^2$ in the oscillator framework as well. To make it explicit, it is necessary to choose the certain parametrization for $SU(2)$ representations of the gauge group. The Eulerian one complicates computations and therefore seems to be inefficient for explicit usage of nontrivial constraints.[17] Those introduced in Refs. 18,19 makes the geometric nature of the problem less transparent. The parametrization proposed by the authors[15] helped to choose the fiber coordinate in a simple form and to fulfill the straight-

[a)]Electronic mail: plet@dragon.bas-net.by
[b)]Electronic mail: tea@dragon.bas-net.by

forward transition to Hilbert space of the $SU(2)$ Kepler problem in the ''isovector'' (integer ''isospins'') case.

In this paper, in Sec. II, we describe the well-known correspondence between an $8D$ harmonic oscillator and a $5D$ $SU(2)$-Kepler problem on the background of a centrifugal Zwanziger-like term ($5D$ MIC-Kepler problem) in suitable parametrization, half-integer ''isospins'' being treated as well. Then, in Sec. III, we introduce the singular term into the oscillator's Hamiltonian and show that this ''corrected'' equation is associated with the $5D$ Kepler problem on $SU(2)$ instanton background (a higher dimensional dyogen atom's counterpart). The energy spectrum of the latter problem is obtained for the first time.

II. 8*D* OSCILLATOR AND 5*D* KEPLER PROBLEM

Let us start with the Schrödinger equation for an $8D$ isotropic oscillator,

$$-\frac{1}{2}\Delta_8\Psi^{(8)}+\frac{\omega^2}{2}(u^2+v^2)\Psi^{(8)}=E\Psi^{(8)}, \tag{1}$$

where $\Psi^{(8)}=\Psi^{(8)}(u_i,v_i)$ is a scalar wave function.

It is well-known that Hurwitz transformation (HT) mapping $\dot{R}^8\equiv R^8\setminus\{0\}\to\dot{R}^5\equiv R^5\setminus\{0\}$,

$$\begin{gathered} r_0=u_iu_i-v_iv_i\equiv u^2-v^2, \\ r_a=2(-u_0v_a+u_av_0-\varepsilon_{abc}u_bv_c), \quad r_4=2u_iv_i, \end{gathered} \tag{2}$$

$(a,b,c=1,2,3;\ i=0,1,2,3)$, possesses the property

$$(u^2+v^2)^2=r_0^2+r_a^2+r_4^2\equiv r_\mu^2\equiv R^2 \quad (\mu=0,\ldots,4). \tag{3}$$

HT makes projection in the principal fiber bundle, $\dot{R}^8$, $\dot{R}^5$ and $SU(2)$ being total bundle, base, and structure group, respectively. The operators removing the function, which depends on the base coordinates r_μ only, i.e., $X_af(r_\mu)=0$, are called vertical and have the form[20]

$$X_a=-u_a\frac{\partial}{\partial u_0}+u_0\frac{\partial}{\partial u_a}-\varepsilon_{abc}u_b\frac{\partial}{\partial u_c}-v_a\frac{\partial}{\partial v_0}+v_0\frac{\partial}{\partial v_a}-\varepsilon_{abc}v_b\frac{\partial}{\partial v_c}. \tag{4}$$

The operators $\hat{X}_a\equiv(i/2)X_a$ obey standard $SU(2)$ commutation relations.

It is convenient to introduce fiber coordinates $z_a=u_a/u_0$ and pass from one set (u_0,u_a,v_0,v_a) to another (r_0,r_a,r_4,z_a). This transformation is valid locally, i.e., in the region, where $|u|\neq 0$, or $R\neq -r_0$, only. Then

$$\hat{X}_a\to T_a=\frac{i}{2}\left(z_az_b\frac{\partial}{\partial z_b}+\frac{\partial}{\partial z_a}-\varepsilon_{abc}z_b\frac{\partial}{\partial z_c}\right). \tag{5}$$

These vertical operators exactly coincide with $SU(2)$ generators acting in the group space expressed in terms of vector-parameters.[21] Note that vector parametrization of $SU(2)(SO(3))$ is related with Eulerian one via

$$\begin{gathered} z_1=\tan\frac{\theta}{2}\cos\frac{\varphi-\psi}{2}\Big/\cos\frac{\varphi+\psi}{2}, \\ z_2=\tan\frac{\theta}{2}\sin\frac{\varphi-\psi}{2}\Big/\cos\frac{\varphi+\psi}{2}, \quad z_3=\tan\frac{\varphi+\psi}{2}. \end{gathered} \tag{6}$$

In the new set of coordinates the $8D$ Laplacian reads as

$$\Delta_8=-4R(-i\partial_\mu+A_\mu^a O_{ab}(\mathbf{z})T_b(\mathbf{z}))^2-\frac{4}{R}\mathbf{T}^2, \tag{7}$$

where

$$A_\mu^a dr_\mu=\frac{1}{R(R+r_0)}(-r_4 dr_a+r_a dr_4-\varepsilon_{abc}r_b dr_c) \tag{8}$$

is an instantonic connection form;[22]

$$O_{ab}(\mathbf{z})=\delta_{ab}+\frac{2}{1+\mathbf{z}^2}(z_a z_b-\mathbf{z}^2\delta_{ab}+\varepsilon_{adb}z_d) \tag{9}$$

is the orthogonal matrix in the vector parametrization.[21]

It can be readily shown that

$$O_{ab}(\mathbf{z})T_b(\mathbf{z})=\frac{i}{2}\left(z_a z_b\frac{\partial}{\partial z_b}+\frac{\partial}{\partial z_a}+\varepsilon_{abc}z_b\frac{\partial}{\partial z_c}\right)=-T_a(-\mathbf{z})\equiv -K_a(\mathbf{z})\equiv -K_a. \tag{10}$$

The following relations are valid:

$$[K_a,T_b]=0,\quad [K_a,K_b]=i\varepsilon_{abc}K_c,\quad \mathbf{K}^2\equiv K_aK_a=\mathbf{T}^2\equiv T_aT_a. \tag{11}$$

Then the initial equation (1) can be identically rewritten as

$$\frac{1}{2}(-i\partial_\mu-A_\mu^a K_a)^2\Psi^{(8)}+\left(\frac{\mathbf{K}^2}{2R^2}-\frac{\alpha}{R}\right)\Psi^{(8)}=\varepsilon\Psi^{(8)}, \tag{12}$$

where $\varepsilon=-\omega^2/8$, $\alpha=E/4$.

To complete the transition to $\dot{R}^5$, one should impose additional constraints onto the function $\Psi^{(8)}(r_\mu,\mathbf{z})$. Instead of the topologically trivial condition—the independence of $\Psi^{(8)}$ upon the fiber coordinates[5,12,13]—we require the equivariance condition to hold:[15,16]

$$\mathbf{K}^2\Psi^{(8)}=l(l+1)\Psi^{(8)}. \tag{13}$$

Note that the trivial case corresponds to ''isospin'' $l=0$.

Consider now the $SU(2)$ Wigner representation constructed according to the classic scheme:

$$\begin{gathered}\mathbf{K}^2\psi^l_{mm'}=l(l+1)\psi^l_{mm'},\quad K_3\psi^l_{mm'}=m\psi^l_{mm'},\\ (-T_3)\psi^l_{mm'}=m'\psi^l_{mm'},\quad -l\leqslant m,m'\leqslant l.\end{gathered} \tag{14}$$

For $\psi=(\psi^l_{mm'})$, the identity

$$K_a\psi^T=\psi^T\Lambda_a \tag{15}$$

is valid, provided that Λ_a are the generators of the respective representation with ''isospin'' l.

Let us expand $\Psi^{(8)}$ in a series of $\psi^l_{mm'}$,

$$\Psi^{(8)}=\sum_{m,m'}\psi^l_{mm'}\Phi_{mm'}=Sp[\psi^T(\mathbf{z})\Phi(r_\mu)]. \tag{16}$$

Then

$$(-i\partial_\mu - A_\mu^a K_a)\Psi^{(8)} = Sp[\psi^T(\mathbf{z})(-i\partial_\mu - A_\mu^a \Lambda_a)\Phi(r_\mu)]. \tag{17}$$

Substituting (17) into (12), we obtain the equation for $\phi \equiv \Phi_{ml}$ (without loss of generality it is chosen as $m' = l$):

$$\left(\frac{\pi_\mu^2}{2} + \frac{l(l+1)}{2R^2} - \frac{\alpha}{R}\right)\phi = \varepsilon \phi, \tag{18}$$

where $\pi_\mu = -i\partial_\mu - A_\mu^a \Lambda_a$. Equation (18) is the $5D$ counterpart of the MIC-Kepler problem arising in electrodynamics with magnetic charge.[8] The energy spectrum of (18) is known[16] to be

$$\varepsilon_N = -\frac{\alpha^2}{2[(N/2)+2)^2}, \quad N = 0,1,2,\ldots. \tag{19}$$

Though there is no explicit dependence on l, N takes only even values: $N = 0,2,4,\ldots$, for $l = 0$.[12]

To finish this section, we recall the relation between the Hermitian products in $\dot{R}^8$ and $\dot{R}^5$ (see, for instance, Ref. 13). From the definition

$$\langle \Psi_1 | \Psi_2 \rangle_8 = \frac{2^7}{\pi^2} \int \Psi_1^*(u_i, v_i)\Psi_2(u_i, v_i)(|u|^2 + |v|^2) d^4u_i d^4v_i, \tag{20}$$

and the natural requirement

$$\langle \Psi_1 | \Psi_2 \rangle_8 = \langle \varphi_1 | \varphi_2 \rangle_5,$$

one obtains

$$\langle \varphi_1 | \varphi_2 \rangle_5 = \int \varphi_1^+(r_\mu)\varphi_2(r_\mu) d^5 r_\mu. \tag{21}$$

The latter is valid because

$$\frac{2^7}{\pi^2} \int_{S^3} (|u|^2 + |v|^2) d^4u_i d^4v_i = d^5 r_\mu, \tag{22}$$

where factor 2^8 is due to the bilinearity of HT and $2\pi^2$ is $SU(2)$ group volume.

III. ENERGY SPECTRUM OF $SU(2)$ KEPLER PROBLEM

In Ref. 11 it has been shown that introduction of the certain singular term into the $4D$ oscillator's equation leads to the $3D$ dyogen atom's problem. The Hamiltonian of the latter does not contain a nonphysical Zwanziger's term.

Let us consider on the analogy a singular $8D$ Hamiltonian,

$$H = -\frac{1}{2}\sum_{s=1}^{8} \frac{\partial^2}{\partial x_s^2} + \frac{\omega^2 x^2}{2} - \frac{2l(l+1)}{x^2} \quad \left(x^2 = \sum_{s=1}^{8} x_s x_s\right). \tag{23}$$

Here and below we mean $(x_1,\ldots,x_4,x_5,\ldots,x_8) = (u_0,\ldots,u_3,v_0,\ldots,v_3)$. The eigenvalues of (23) are given by[23,24]

$$E_{nj} = 2\omega(n + 1/2 + \sqrt{-l(l+1) + j(j+3) + 9/4}), \tag{24}$$

where j takes integer and half-integer values.

Under the HT Hamiltonian (23) is associated with a $5D$ counterpart of the dyogen atom:

$$\left(\frac{\pi_\mu^2}{2}-\frac{\alpha}{R}\right)\varphi=\varepsilon\varphi. \tag{25}$$

Since $\varepsilon=-\omega^2/8$ and $\alpha=E/4$, one can readily find

$$\varepsilon_{nj}=-\frac{\alpha^2}{2(n+1/2+\sqrt{-l(l+1)+j(j+3)+9/4})^2}. \tag{26}$$

But such treatment does not clarify what values j *exactly* takes.

To obtain the expression (26) in the closed form, we shall use the ideas based on the dynamical symmetry approach. Consider the operators

$$B_2^+=B_{20}^+-i\frac{2l(l+1)}{x^2},\quad B_2=B_{20}+i\frac{2l(l+1)}{x^2},$$
$$H=H_0-\frac{2l(l+1)}{x^2}, \tag{27}$$

expressed via nonsingular (harmonic) ones,

$$B_{20}^+=-i\omega a_s^+a_s^+,\quad B_{20}=i\omega a_s a_s,\quad H_0=\omega(a_s^+a_s+4). \tag{28}$$

The creation a_s^+ and annihilation a_s operators obey $[a_s,a_t^+]=\delta_{st}$ as usual.

The operators (28) satisfy the relations

$$[H,B_2^+]=2\omega B_2^+,\quad [H,B_2]=-2\omega B_2,\quad [B_2,B_2^+]=4\omega H, \tag{29}$$

as well as the operators (28) do,

$$[H_0,B_{20}^+]=2\omega B_{20}^+,\quad [H_0,B_{20}]=-2\omega B_{20},\quad [B_{20},B_{20}^+]=4\omega H_0. \tag{30}$$

Let us introduce

$$K_0=\frac{H}{2\omega},\quad K_+=\frac{B_2^+}{2\omega},\quad K_-=\frac{B_2}{2\omega}, \tag{31}$$

which satisfy $SU(1,1)$ commutation relations,

$$[K_0,K_\pm]=\pm K_\pm,\quad [K_-,K_+]=2K_0. \tag{32}$$

The representation of $SU(1,1)$ is constructed as follows:[24]

$$K_0|n,k\rangle=(n+k)|n,k\rangle,$$
$$K_+|n,k\rangle=\sqrt{(n+1)(n+2k)}|n+1,k\rangle, \tag{33}$$
$$K_-|n,k\rangle=\sqrt{n(n+2k-1)}|n-1,k\rangle,$$

where $n=0,1,\ldots,$ and $k(k-1)$ are the eigenvalues of the Casimir operator,

$$C_2=K_0^2-K_0-K_+K_-. \tag{34}$$

From the formal relation

$$\frac{E}{2\omega}=n+k, \tag{35}$$

it follows that

$$\varepsilon_{nk}=-\frac{\omega_{nk}^2}{8}=-\frac{\alpha^2}{2(n+k)^2}. \tag{36}$$

Here n takes the arbitrary integer non-negative values. One should figure out what values k takes.

The Casimir operators C_2 of $SU(1,1)$ (34) and C_{20} of $SU(1,1)$ generated by $H_0/2\omega$, $B_{20}^+/2\omega$, $B_{20}/2\omega$, are related as

$$C_2=C_{20}-l(l+1), \tag{37}$$

where

$$C_{20}=\frac{1}{4\omega^2}(\omega^2(a_s^+a_s+4)^2-2\omega^2(a_s^+a_s+4)-\omega^2(a_s^+)^2(a_t)^2)=-\frac{1}{8}(a_s^+a_t-a_t^+a_s)^2+2. \tag{38}$$

Taking into account $x_s\partial_t-x_t\partial_s=a_s^+a_t-a_t^+a_s$ it is equal,

$$C_{20}=-\frac{1}{8}(x_s\partial_t-x_t\partial_s)^2+2=\frac{1}{4}(-x^2\Delta_8+7x_s\partial_s+x_sx_t\partial_s\partial_t)+2. \tag{39}$$

The spatial symmetry of the Hamiltonian (23) is $SO(8)$. Its 28 generators are $F_{st}=i(x_s\partial_t-x_t\partial_s)$, $1\leqslant s<t\leqslant 8$. One can express C_{20} through the operator of the squared $SO(8)$ angular momentum F^2,

$$C_{20}=\frac{1}{8}(F_{st})^2+2=\frac{1}{4}\sum_{s<t}F_{st}F_{st}+2=\frac{1}{4}F^2+2. \tag{40}$$

The complete set of the wavefunctions of the $8D$ harmonic oscillator has been constructed[21] by means of the consideration of the subgroup chain $SO(8)\supset SO(4)\times SO(4)\supset U(1)\times U(1)\times U(1)\times U(1)$. The obtained wavefunctions $Y^j_{j_1m_{11}m_{12}j_2m_{21}m_{22}}$ are the simultaneous eigenfunctions of the mutually commuting operators F^2, $\mathbf{I}^2=\mathbf{I}_1^2=\mathbf{I}_2^2$, $\mathbf{N}^2=\mathbf{N}_1^2=\mathbf{N}_2^2$, I_{1_3},I_{2_3}, N_{1_3},N_{2_3} with eigenvalues

$$\begin{gathered}F^2Y^f=f(f+6)Y^f,\quad \mathbf{I}^2Y^f=j_1(j_1+1)Y^f,\\ \mathbf{N}^2Y^f=j_2(j_2+1)Y^f,\quad I_{1_3}Y^f=m_{11}Y^f,\\ I_{2_3}Y^f=m_{12}Y^f,\quad N_{1_3}Y^f=m_{21}Y^f,\quad N_{2_3}Y^f=m_{22}Y^f,\end{gathered} \tag{41}$$

where

$$0\leqslant f\leqslant N,\ N \text{ even},\ f \text{ even},$$

$$1\leqslant f\leqslant N,\ N \text{ odd},\ f \text{ odd};$$

$$0\leqslant j_1+j_2\leqslant f/2,\ f \text{ even},\ j_1+j_2 \text{ integer},$$

$$1/2 \leqslant j_1 + j_2 \leqslant f/2, \quad f \text{ odd}, \quad j_1 + j_2 \quad \text{half-integer};$$

$$-j_1 \leqslant m_{11}, m_{12} \leqslant j_1,$$

$$-j_2 \leqslant m_{21}, m_{22} \leqslant j_2,$$

and N is the oscillator's occupation number: $E_N = \omega(N+4)$, $N = 0,1,2,\ldots$.

Comparing the explicit form of the operators $\hat{\mathbf{X}}$ (4) and $\mathbf{I}_1$, $\mathbf{I}_2$, $\mathbf{N}_1$, $\mathbf{N}_2$,

$$I_{1_a} = \frac{i}{2}\left(-u_a \frac{\partial}{\partial u_0} + u_0 \frac{\partial}{\partial u_a} - \varepsilon_{abc} u_b \frac{\partial}{\partial u_c} \right),$$

$$I_{2_a} = \frac{i}{2}\left(-u_a \frac{\partial}{\partial u_0} + u_0 \frac{\partial}{\partial u_a} + \varepsilon_{abc} u_b \frac{\partial}{\partial u_c} \right),$$

$$N_{1_a} = \frac{i}{2}\left(-v_a \frac{\partial}{\partial v_0} + v_0 \frac{\partial}{\partial v_a} - \varepsilon_{abc} v_b \frac{\partial}{\partial v_c} \right),$$

$$N_{2_a} = \frac{i}{2}\left(-v_a \frac{\partial}{\partial v_0} + v_0 \frac{\partial}{\partial v_a} + \varepsilon_{abc} v_b \frac{\partial}{\partial v_c} \right),$$

we deduce that $\hat{\mathbf{X}} = \mathbf{I}_1 + \mathbf{N}_1$.

We recall that there exist the constraint (13) which in this context can be written as

$$\hat{\mathbf{X}}^2 Y^f = l(l+1) Y^f. \tag{42}$$

According to the angular momentum addition rules $l = j_1 - j_2, j_1 - j_2 + 1, \ldots, j_1 + j_2$ (we put $j_1 \geqslant j_2$ for certainty), i.e., l takes values which depend on j_1 and j_2. We shall rearrange the scheme (65) in order to make l the free index, i.e., taking the arbitrary integer and half-integer values: $l = 0,1/2,1,\ldots$.

Let us define $j = f/2$ and put $l = q - p$, where $q = j_1 + j_2$, $p = 0,1,\ldots,2j_2$. From the comment to (41) it follows that $j = q, q+1, \ldots = q + p'$, where $p' = 0,1,2,\ldots$. Fixing l we find $j = l + (p + p')$, where $p'' \equiv p + p' = 0,1,2,\ldots$. At fixed j, one has $N = 2j, 2j+1, \ldots$. It is natural to define the free quantum number $n = N/2 - j = 0,1,2,\ldots$.

From (37), (40) and (41) it follows that the relation

$$k(k-1) = j(j+3) + 2 - l(l+1), \tag{43}$$

which allows us to express k through j,

$$k = 1/2 + \sqrt{-l(l+1) + j(j+3) + 9/4}. \tag{44}$$

Substituting (44) into (36), we obtain the energy spectrum for the $5D$ Kepler problem on the background of the $SU(2)$ instanton,

$$\varepsilon_{nk} \to \varepsilon_{nj} = -\frac{\alpha^2}{2(n + 1/2 + \sqrt{-l(l+1) + j(j+3) + 9/4})^2}. \tag{45}$$

It exactly coincides with (26), but now we are convinced that

$$j = l, l+1, \ldots .$$

In the particular case $l = 0$,

$$\varepsilon_{nj} = -\frac{\alpha^2}{2(n+j+2)^2}, \quad n=0,1,2,\ldots, \quad j=0,1,2,\ldots. \tag{46}$$

Taking into account the relation between n,j and N, one can conclude that (46) is in full correspondence with (19) considered at $l=0$ as well.

In conclusion, we note that the Hurwitz transformation and related topics may be useful in some nuclear models (see the references in Ref. 20).

ACKNOWLEDGMENT

The authors would like to thank I. E. Pris' for fruitful discussions.

[1] I. Polubarinov, preprint JINR E2-84-607, Dubna, 1984.
[2] D. Lambert and M. Kibler, J. Phys. A **21**, 307 (1988).
[3] M. Kibler and P. Labastie, preprint LYCEN 8829, Lyon, 1988 (also available as hep-th\9409196).
[4] M. Hage Hassan and M. Kibler, preprint LYCEN 9110, Lyon, 991 (also available as hep-th\9409051).
[5] I. Polubarinov, preprint JINR P2-83-872, Dubna, 1983.
[6] T. Iwai and Y. Uwano, J. Math. Phys. **27**, 1523 (1986).
[7] T. Iwai and Y. Uwano, J. Phys. A **21**, 4083 (1988).
[8] A. Cisneros and H. McIntosh, J. Math. Phys. **10**, 277 (1969).
[9] D. Zwanziger, Phys. Rev. **176**, 1480 (1968).
[10] C. Gerry and J. Keifer, Phys. Rev. A **37**, 665 (1988).
[11] I. E. Pris' and E. A. Tolkachev, Sov. J. Nucl. Phys. **54**, 582 (1991).
[12] L. Davtyan *et al.*, J. Phys. A **20**, 6121 (1987).
[13] M. Trunk, Int. J. Mod. Phys. A **11**, 2329 (1996).
[14] I. Mladenov and V. Tsanov, J. Geom. Phys. **2**, 17 (1985).
[15] M. Pletyukhov and E. Tolkachev, *Proceedings of the International Workshop on Quantum Systems: New Trends and Methods (QS-96)* (World Scientific, Singapore, 1997), pp. 77–80.
[16] T. Iwai and T. Sunako, *Proceedings of the VII International Conference on Symmetry Methods in Physics*, edited by A. Sissakian and G. Pogosyan (Joint Institute for Nuclear Research, Dubna, 1996), Vol. I, pp. 273–281; T. Iwai and T. Sunako, J. Geom. Phys. **20**, 250 (1996).
[17] L. Mardoyan, A. Sissakian, and V. Ter-Antonyan, preprint JINR E2-96-24, Dubna, 1996, submitted to Phys. Lett. **A** (also available as hep-th\9601093).
[18] Kh. Karayan, L. Mardoyan, and V. Ter-Antonyan, preprint JINR E2-94-359, Dubna, 1994.
[19] Kh. Karayan, L. Mardoyan, and V. Ter-Antonyan, preprint JINR E2-94-439, Dubna, 1994.
[20] M. Kibler and P. Winternitz, J. Phys. A **21**, 1787 (1988).
[21] F. I. Fedorov, *Lorentz Group* (Nauka, Moscow, 1979).
[22] C. N. Yang, J. Math. Phys. **19**, 320 (1978).
[23] I. Malkin and V. Man'ko, *Dynamical Symmetries and Coherent States of Quantum Systems* (Nauka, Moscow, 1979).
[24] A. Perelomov, *Generalized Coherent States and Their Applications* (Springer-Verlag, Berlin, 1986).

Isolated versus nonisolated periodic orbits in variants of the two-dimensional square and circular billiards

R. W. Robinett
Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802

(Received 20 July 1998; accepted for publication 5 October 1998)

Square and circular infinite wells are among the simplest two-dimensional potentials which can completely solved in both classical and quantum mechanics. Using the methods of periodic orbit theory, we study several variants of these planar billiard systems which admit both singular isolated and continuous classes of nonisolated periodic orbits. (In this context, isolated orbits are defined as those which are *not* members of a continuous family of paths whose orbits are all of the same length.) Examples include (i) various ''folded'' versions of the standard infinite wells (i.e., potentials whose geometrical shapes or ''footprints'' can be obtained by repeated folding of the basic square and circular shapes) and (ii) a square well with an infinite-strength repulsive δ-function ''core,'' which is a special case of a Sinai billiard. In each variant case considered, new isolated orbits are introduced and their connections to the changes in the quantum mechanical energy spectrum are explored. Finally, we also speculate about the connections between the periodic orbit structure of supersymmetric partner potentials, using the two-dimensional square well and it superpartner potential as a specific example. © *1999 American Institute of Physics.* [S0022-2488(99)03101-1]

I. INTRODUCTION

Many of the physical principles and mathematical tools necessary to understand quantum mechanics are first approached through a variety of familiar one-dimensional examples, including problems involving infinite potential wells. The extension to two-dimensional (2D) systems, such as square and circular infinite wells, allows for the study of new features such as degeneracy of energy levels (accidental or otherwise) as well as providing some of the simplest systems which exhibit both classical and quantum mechanical chaos.

It is also in such two-dimensional systems that several novel methods of analysis become nontrivially applicable, with examples including the study of energy-level statistics[1] and periodic orbit (PO) theory.[2] In this report, we study how the second method, namely periodic orbit theory, can be applied to a variety of variants of the simplest square and circular well geometries and our main focus will be on the relationship between the symmetries of the potential well system (geometrical and otherwise), the structure of the allowed energy eigenvalue space, and their relationships to the presence of isolated periodic orbits, that is, orbits which are *not* part of a continuous family of paths, all of which are characterized by identical path lengths.

The infinite well potentials upon which we will first focus our attention are the familiar square and circular wells,[3] as well as several variants thereon which can be constructed by ''folding'' the basic geometrical footprints in obviously symmetrical ways. Specifically we will derive results for rectangular (2×1 aspect ratio) wells (which do not have new isolated orbits) and triangular (isosceles) wells (which do!), both of which can be obtained from the square well by folding along horizontal (vertical) or diagonal axes respectively, as well as the half- and quarter-circular wells obtained by repeated foldings of the circle along an axis of symmetry. We concentrate on these examples for several reasons:

(i) Because of the high degree of symmetry involved in all these cases, there will be simple connections between both the energy eigenvalues of the folded wells and their more standard precursors, as well as for the classes of allowed periodic orbits in these geometries. As we will see, the way in which the energy eigenvalue spectrum and the periodic orbits are changed are remarkably similar for the ''half'' circular well and the triangular folding of the square billiard.

(ii) Folded versions of the basic well shapes will have a reduced amount of degeneracy compared to their original counterparts and this will allow us to study the effects that such degeneracies have on the appearance of new features corresponding to isolated periodic orbits.

(iii) The use of square and circular well geometries ensures that, due to the infinite number of bound state energies, one can readily obtain a large number of energy levels to increase the statistical sample in any numerical analysis. In addition, these simple geometries allow one to easily calculate the energy levels using standard mathematical techniques. More complex geometries, such as stadium-shaped billiards[4] or even elliptical wells,[5,6] require more sophisticated mathematical methods (or even experimental[7] techniques) in order to evaluate the energy eigenvalues.

In addition to these folded-wells, we will also consider a 2D square well with an infinite-strength repulsive δ-function ''core'' placed at the center of the well which will be seen to also induce new isolated periodic orbits in an even more dramatic fashion. For this case, which is a limiting case of the Sinai billiard,[8] we find that an infinite number of isolated new periodic orbits are generated, with a correspondingly more dramatic change in the energy eigenvalue spectrum.

Finally, we will speculate on the connections between the periodic orbit structure of systems whose potential energy functions are related by supersymmetric quantum mechanics. Because of the high degree of correlation between the energy eigenvalue spectra of two systems whose potentials are related by supersymmetry, we expect similarly close connections between the allowed periodic orbits in the classical theory. This somewhat obvious, but nonetheless interesting question, deserves to be examined more fully and formally, but as a starting point we will consider the superpartner potential of the 2D infinite square well and we will note that the relative importance of isolated versus nonisolated orbits may also play a significant role there as well.

In the next sections, we first briefly review the main ideas of periodic orbit theory and their applications in the circular and square well (Secs. II A and II B). We then apply these ideas to two folded versions of the square well, namely the triangular well (Sec. II C) and the rectangular well (Sec. II D). We discuss in some detail the square well plus repulsive δ-function ''core'' in Sec. II E, while we present a discussion of the relationship between quantum mechanical supersymmetry and periodic orbit theory in Sec. II F, with our conclusions appearing in Sec. II G. Finally, in the Appendix, we discuss some of the technical details required to evaluate the energy eigenvalue spectrum we use in Sec. II E, using a matrix mechanics formulation of quantum theory.

II. PERIODIC ORBIT THEORY

Of the possible approaches which can be used to explore the many relationships between the classical and quantum mechanical descriptions of nature, periodic orbit (PO) theory[2] is certainly one of the most compelling. An intimate connection between the quantum mechanical density of energy eigenvalues and the set of classical closed paths has been discussed by Gutzwiller and others[9–11] and is described by the formula

$$\sum_n \delta(E-E_n) \equiv \rho(E) = \rho_0(E) + \sum_{p=1}^{\infty} \sum_{\gamma} \rho_{(\gamma,p)} \cos\left[p\left(\frac{S_\gamma(E)}{\hbar} - \phi_\gamma \right) \right]. \tag{1}$$

In this expression, the E_n are the quantized energy eigenvalues in the bound state system which therefore define the energy level density, $\rho(E)$. This can be decomposed into a smooth part, $\rho_0(E)$, which has no information on the classical orbits, and an oscillatory piece depending on the

S_γ which are the values of the classical action corresponding to all possible closed paths (characterized by the label γ) of the classical motion; the integral values of $p=1,2,\ldots$ simply correspond to repetitions or recurrences of the basic ''primitive'' periodic orbits or closed trajectories. In this context, Eq. (1) has been used to make connections with the energy level structure of important physical systems, such as metallic clusters[12] or the hydrogen atom in large magnetic fields.[13]

This connection can be simplified even more in so-called billiard systems, two- and three-dimensional free particle systems confined in infinite well type potentials. There one finds for the quantity $\rho(L)$ [essentially the Fourier transform of $\rho(E)$]

$$\rho(L) \equiv \sum_{n=1}^{\infty} e^{ik_n L} = \rho_0(L) + \sum_{p=1}^{\infty} \sum_{\gamma} \rho_{(\gamma,p)}\, \delta(L - pL_\gamma). \tag{2}$$

(We note that Balian and Bloch[11] use a slightly different weighting in their study of the three-dimensional spherical billiard system.) In this relation, L_γ is the length of an individual primitive (or ''once-around'') periodic or closed orbit and $p=1,2,3,\ldots$ once again label the recurrences. [The $\rho_{(\gamma,\rho)}$ are constants which will not be relevant to our present discussion; the $\rho_0(L)$ term arises from the smooth $\rho_0(E)$ term and is highly peaked around $L=0$.] Thus, the Fourier transform of the quantum energy spectrum, defined via $\rho(L)$ using knowledge of the quantized k_n values, will show features at values of L corresponding to the lengths of classical periodic orbits (and their repetitions). [We recall that one of our interests is in exploring the relative importance of isolated versus nonisolated periodic orbits whose path lengths will both appear as features in the right hand side of Eq. (2).] Since we necessarily use a finite number of energy eigenvalues in a numerical evaluation of $\rho(L)$, we will actually evaluate

$$\rho(L, N_{\mathrm{TOT}}) \equiv \sum_{n=1}^{N_{\mathrm{TOT}}} e^{ik_n L} \tag{3}$$

and we will find it useful to examine the scaling properties of $\rho(L,N_{\mathrm{TOT}})$, as a function of N_{TOT}, for various obvious features corresponding to both isolated and nonisolated orbits.

The analysis of simple billiard systems such as 2D square[3] and rectangular[14] wells as well as 2D circular wells[3] rely heavily on geometrical ideas and symmetry principles to visualize both the patterns of energy eigenvalues and periodic orbits. Using just such ideas, it was possible in Ref. 3 to extend the PO theory analysis of the circular well to simple variants thereof, such as the half- and quarter-well cases. (A much more complete analysis of the general case of an ''angular slice'' well, characterized by an infinite well with a circular arc boundary and walls separated by an arbitrary angle Θ was performed in Ref. 14; the half- and quarter-well cases are then special cases of this more general result corresponding to $\Theta=\pi$ and $\Theta=\pi/2$, respectively.) For both completeness sake, as well as for use in later discussions, we very briefly review these results for various versions of the circular well before proceeding to variants on the square well.

A. Circular and half-circular well

The periodic orbits or closed trajectories in a circular infinite well (of radius R) or billiard can be characterized by two integers, p (which counts the number of ''hits'' on the inner wall) and q (which counts the number of net revolutions made by the particle before it returns to the same location in phase space) and the two are related by the condition $p\Theta = q(2\pi)$, where Θ is the angle subtended by the chord length between bounces. The total path length for such a primitive or once-around closed orbit is

$$L(p,q) = 2pR \sin\left(\frac{\pi q}{p}\right) \tag{4}$$

and integral multiples of this, corresponding to multiple repetitions of the basic trajectory, are also allowed. Each set of (p,q) actually corresponds to a continuous set of closed paths, each equivalent to each other in length, and only differing in the initial angle. These continuous families of trajectories are our first example of sets of nonisolated orbits.

The corresponding quantum mechanical system has an energy spectrum given by

$$E_{(m,n_r)}=\frac{\hbar^2[a_{(m,n_r)}]^2}{2\mu R^2}=\left(\frac{\hbar^2}{2\mu}\right)k_n^2, \tag{5}$$

where we denote the particle mass by μ in order to avoid confusion with the angular momentum quantum number, m. The $a_{(m,n_r)}$ denote the n_rth zero of the $|m|$th regular cylindrical Bessel function and the orbital angular momentum is given by $L_z=m\hbar$. The angular wave functions are given by $\Theta_m(\theta)\propto\cos(m\theta)$ and $\sin(\theta)$ so that for each value of $m>0$ there is a twofold degeneracy (corresponding to the indistinguishability of clockwise versus counterclockwise motion), while for $m=0$ only one copy of the constant cos(0) solutions is present.

Using the values of the $a_{(m,n_r)}$, the Fourier transform defined by Eq. (2) is shown in Fig. 1 for the full well case where many of the path lengths described by Eq. (4) are clearly visible. The large feature at $L=0$ corresponds to the Fourier transform of the smooth $\rho_0(E)$ piece, while the dashed vertical line shows the limiting value of $(p,q)\rightarrow(\infty,1)$ where $L(\infty,1)=2\pi R$; evaluating $\rho(L)$ with a finite number of eigenvalues implies limited resolution near such locations where the path-length features are increasingly closely spaced.

For the half-circular well, defined by folding the circular well along a diameter, simple geometrical constructions guarantee that any periodic orbit in the full well will continue to exist as a closed trajectory in the half-well, due to the mirroring effect of the new infinite wall along the diameter. The only new feature to be expected is an isolated ''back-and-forth'' orbit through the center of the well, perpendicular to the diametrical infinite wall; this is our first example of an isolated periodic orbit. This singular new feature will have $L/R=2.0$ (and recurrences) in addition

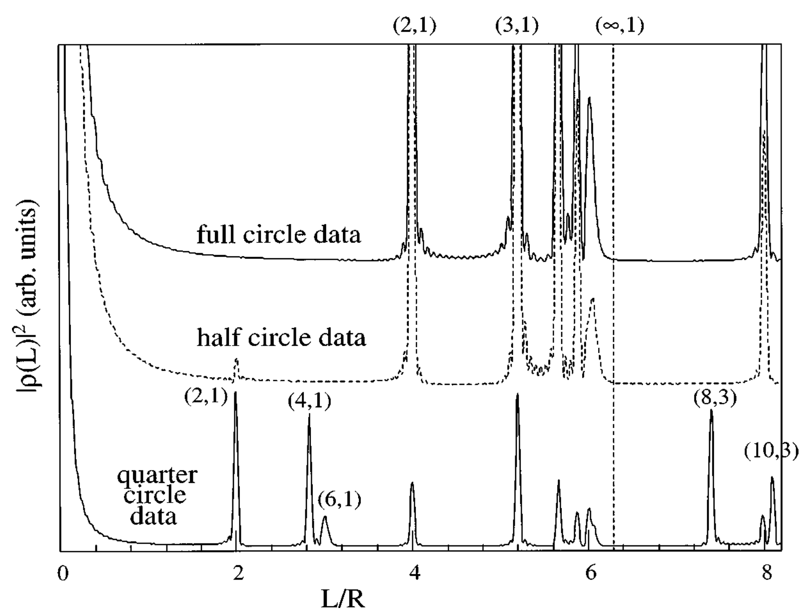

FIG. 1. Fourier transform of the energy level density $|\rho(L)|^2$ [defined by Eq. (3)] versus L/R for the circular well (top), the half-circular well (middle) and quarter-circular well (bottom) geometries. The lowest-lying 2000, 1500, and 860 energy levels have been used in each case, respectively. Note the emergence of a path length feature at $L/R=2.00$ corresponding to an isolated, ''back-and-forth'' orbit in the half-well case.

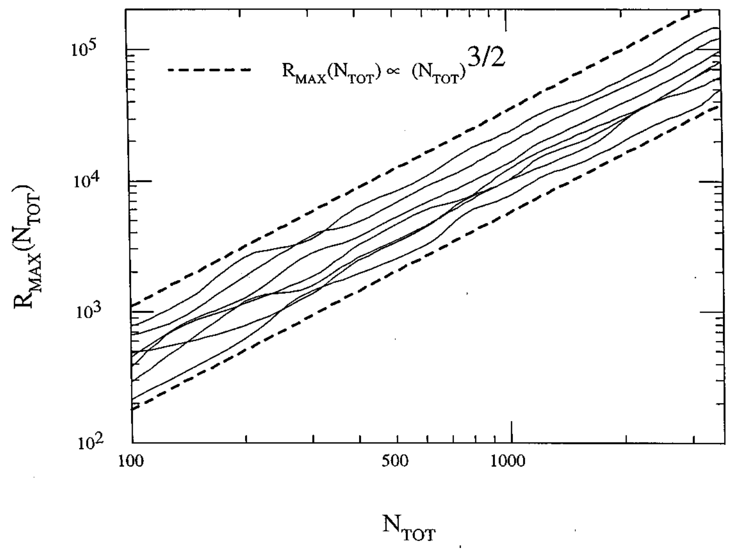

FIG. 2. Scaling behavior of $R_{MAX}(N_{TOT})$ [defined in Eq. (7)] for the first seven nonisolated orbit features as a function of N_{TOT} illustrating the power law behavior consistent with an exponent of $\alpha=3/2$.

to the continuous set of $(p,q)=(2,1)$ features with $L/R=4.0$ (and recurrences) and would be expected to show up, if at all, in the $\rho(L)$ plots as a much smaller ''bump.'' To confirm this, we require knowledge of the quantum energy spectrum in the half-well which we can easily obtain from the solutions of the full well. The $\Theta_m(\theta)\propto\sin(m\theta)$ solutions will remain solutions of the half-well as they satisfy the new boundary condition of vanishing along the diametrical wall. Thus, the energy spectrum consists of one-half of the $m>0$ spectrum for the full well, with no contributions from the nondegenerate $m=0$ eigenstates. The large amount of degeneracy present in the full well is completely lost in the half-well as there are no ''accidental'' degeneracies in this system as there are in the square well.[16]

The resulting $\rho(L)$ versus L plot for the half-well using this set of eigenvalues is also shown in Fig. 1 (middle data) and does indeed show ''spikes'' at the same locations as in the full circular well, plus the expected additional small feature at $L/R=2.0$; recurrences of this new feature at $L/R=4.0$ and 6.0 are hidden by more standard $(p,q)=(2,1)$ and (6,1) spikes.

In order to examine any differences between the features corresponding to isolated orbits (such as the $L/R=2.0$ feature) and nonisolated paths, we also explore the dependence of all the observed features in the $\rho(L)$ spectrum as the number of energy eigenvalues N_{TOT} used in the evaluation of Eq. (3) is increased. We evaluate the ''peak height'' for the first seven recognizable features corresponding to obvious nonisolated paths for increasing values of N_{TOT}, namely,

$$R_{MAX}(N_{TOT})\equiv|\rho(L(p,q),N_{TOT})|^2 \tag{6}$$

and we plot the results in Fig. 2. We note that the behavior of $R_{MAX}(N_{TOT})$ for these features is consistent with a power law scaling of the form

$$R_{MAX}(N_{TOT})\sim A(N_{TOT})^{\alpha} \tag{7}$$

with $\alpha=3/2$ so that the magnitude of these path length features certainly increases with N_{TOT}. In order to compare this dependence with that for the isolated orbits as well as the ''background,'' we evaluate $|\rho(L,N_{TOT})|^2$ for several values of N_{TOT} (namely, 100, 200, 400, 800, 1600, and 3200) for each value of L in the range $0.0\leqslant L/R\leqslant 14.0$, fit each data set to a power law of the form in Eq. (7), and plot the resulting fitted values of α versus L/R in Fig. 3. Directly below each such plot, we also show $|\rho(L)|^2$ itself using a larger ($N_{TOT}=3200$) data set than in Fig. 1 for the half-well for comparison. We immediately note several features:

(i) Even with the increased statistics used, obvious features corresponding to isolated back-and-forth orbits with $L/R=2.0$ and even 10.0 are still obvious (indicated by bold arrows). Contributions of these orbits to features at $L/R=4.0$, 8.0, and 12.0 are, as mentioned above, hidden under the larger peaks corresponding to $L(p,q)$ features with $(p,q)=(2,1)$ and $(6,1)$. It is sometimes possible, however, to separate the otherwise ''degenerate'' contributions from various types of periodic orbits in the circular well by using only appropriate subsets of energy eigenvalues, corresponding in obvious ways to the appropriate classical orbits[15] and that same observation can be used in this case to observe separately the contributions of the isolated orbits.

(ii) The effective value of α drops rapidly from its $L=0$ value of $\alpha=2$ to something consistent with unity or smaller, except for isolated features. [Recall that the definition in Eq. (3) implies that $\rho(L=0,N_{\mathrm{TOT}})\equiv N^2$.] This means that the scaling behavior of the (uninteresting) $\rho_0(L)$ term is relatively unimportant for $L/R\gtrsim 0.1$.

(iii) As noted above, peaks in the α versus L/R plot with values of $\alpha\approx 1.5$ are obvious for each resolvable, nonisolated path length feature in the $\rho(L)$ spectrum, implying that the standard $L(p,q)$ features in the $\rho(L,N_{\mathrm{TOT}})$ spectrum will grow in relative importance in the $N_{\mathrm{TOT}}\to\infty$ limit, just as expected.

(iv) Small features in the α versus L/R spectrum near $L/R=2.0$ and 10.0 indicate that the scaling properties of these features are different from the ''background'' values nearby; in

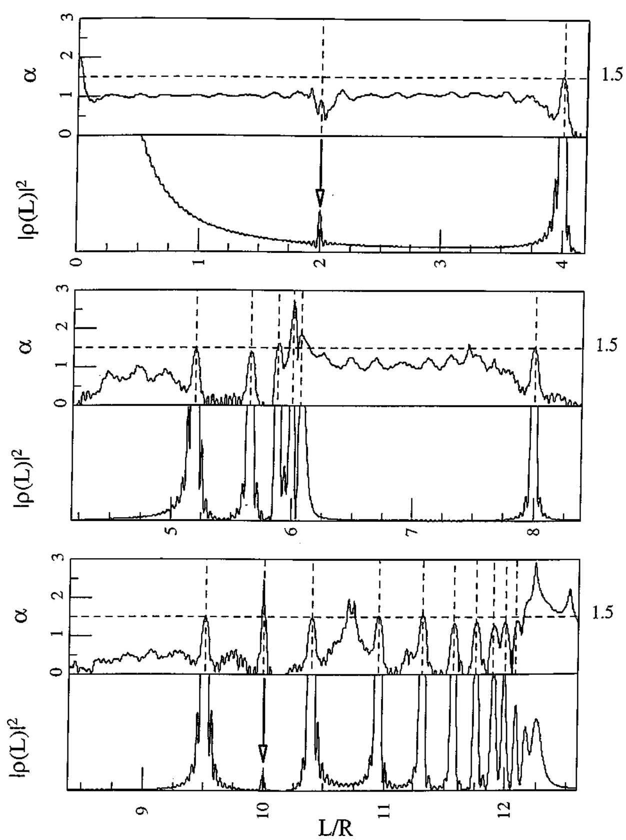

FIG. 3. Fitted values of the power law exponent α [from Eq. (7)] versus L/R for the half-circular well (top plots) with corresponding values of $|\rho(L)|^2$ directly below. A total of $N_{\mathrm{TOT}}=3200$ eigenvalues was used in the numerical evaluation of $\rho(L)$ in this case compared to $N_{\mathrm{TOT}}=1500$ in Fig. 1. Only standard nonisolated path length features given by the $L(p,q)$ in Eq. (4) (indicated by vertical dashed lines) and a new isolated back-and-forth orbit feature at $L/R=2.00,10.00$ are visible. The horizontal dashed line indicates the value of $\alpha=3/2$.

the case of the $L/R=2.0$ feature, the value of $\alpha\approx 1$ is separated from the nearby values, while the $L/R=10.0$ value of $\alpha\approx 2.5$ indicates that the feature will continue to remain significant as $N_{\mathrm{TOT}}\to\infty$.

(v) No other new features (due, for example, to diffraction effects) are obvious in the $\rho(L)$ spectrum in the region explored, except those due to isolated orbits introduced by the folding.

Finally, a second folding, resulting in a quarter-well yields qualitatively new path length features. Orbits in the half-well corresponding to values of p which are *even* can be mirrored in the quarter-well with resulting path lengths which are half as long as in the precursor well, while the features for *odd* values of q remain unchanged in length. (This geometry is not so different from the familiar optics problem of counting the number of images visible when an object is placed between two pivoted mirrors with an opening angle Θ or other similarly kaleidoscopic arrangements of mirrors.) The quantized energy eigenvalues required to evaluate $\rho(L)$ for this geometry can be obtained from those of the standard circular well by choosing one copy of the $m>0$ eigenvalues for *even* values of m as the corresponding eigenfunctions will then satisfy the boundary conditions along both the x- and y-axes. The Fourier transform $\rho(L)$ using these values is also shown in Fig. 1 and shows the expected pattern. Continued foldings into quarter-wells and beyond can be analyzed in a very general manner and the results have been extensively discussed and visualized in Ref. 14.

B. Square well

The general pattern of allowed orbits in a two-dimensional square well will consist of a ''plaid'' pattern of line segments traced by the particles trajectory. This pattern will come to increasingly ''paint'' the square area of the well increasingly uniformly (for the visualization of such orbits, see Ref. 3) unless certain conditions are satisfied which lead to closed orbits. In order for trajectories to eventually close on themselves, we require that

$$\frac{v_y}{v_x}=\frac{n_y}{n_x}, \tag{8}$$

where v_x, v_y are the magnitudes of the classical velocity components in the x,y directions and $n_x, n_y=0,1,2,\ldots$ take on integer values. This condition is then independent of the total energy of the particle, depending only on geometry as it should for a billiard system of this type. The integers (n_x, n_y) give the number of bounces experienced by the particle (during a single primitive periodic orbit) from the vertical (horizontal) walls as $2n_x$ $(2n_y)$. As with the circular well, each label corresponds to a continuous family of paths, of identical path length, differing only by initial position; once again this is a continuous set of nonisolated periodic orbits, one family for each

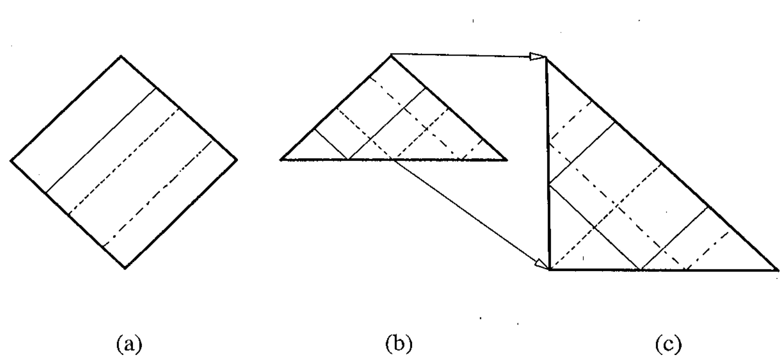

FIG. 4. Typical periodic orbits for the square well (a), the once-folded triangular well (b), and the twice-folded triangular well (c) corresponding to the integer labels $(n_x, n_y)=(1,0)/(0,1)$.

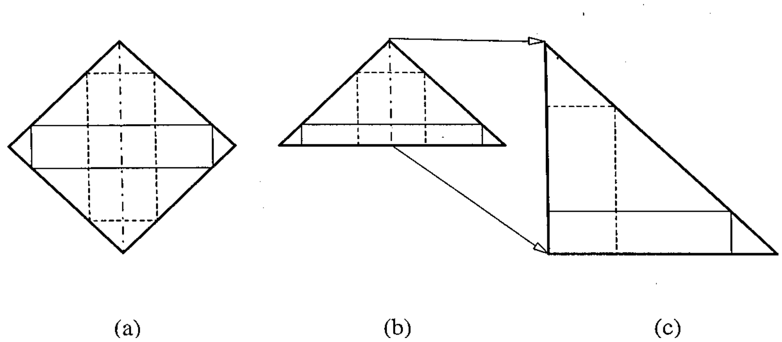

FIG. 5. Same as Fig. 4, but for $(n_x,n_y)=(1,1)$. For case (b), note the appearance of an isolated orbit of length $L/a=\sqrt{2}$ (dotted-dashed line) in addition to the continuous family of nonisolated paths with $L/a=2\sqrt{2}$ (exemplified by the solid and dashed paths.)

(n_x,n_y) label. A few of the simplest such patterns are shown in Figs. 4(a), 5(a), and 6(a). The corresponding primitive or once-around path lengths are then given by a simple geometrical construction to be

$$L(n_x,n_y;a)=2a\sqrt{n_x^2+n_y^2}. \tag{9}$$

A simple extension of this system is to a rectangular well (of dimensions $a_x\times a_y$) where the classification of closed orbits can still be described by an (n_x,n_y) pair, but where the path lengths are now given by

$$L(n_x,n_y;a_x,a_y)=2\sqrt{(n_xa_x)^2+(n_ya_y)^2}. \tag{10}$$

The quantum mechanical case is equally simple to analyze with bound state energies (with wave numbers k_n) given by

$$E_{(i,j)}=\frac{\hbar^2\pi^2}{2ma^2}(i^2+j^2)=\frac{\hbar^2k_n^2}{2m} \qquad \text{or} \qquad k_n=\frac{\pi}{a}\sqrt{i^2+j^2} \tag{11}$$

and wave functions

$$\psi_{(i,j)}(x,y)=\left[\sqrt{\frac{2}{a}}\sin\left(\frac{i\pi x}{a}\right)\right]\left[\sqrt{\frac{2}{a}}\sin\left(\frac{j\pi y}{a}\right)\right], \tag{12}$$

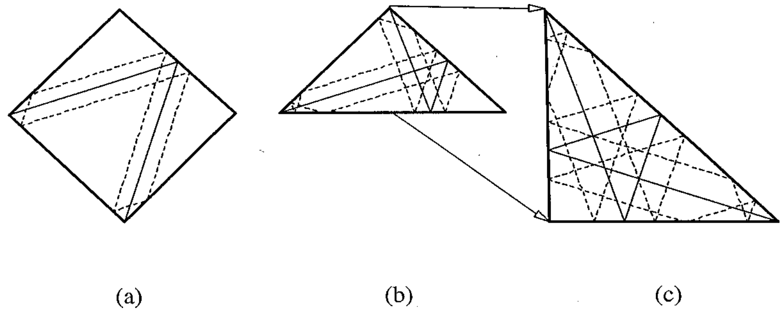

FIG. 6. Same as Fig. 4, but for $(n_x,n_y)=(2,1)/(2,1)$.

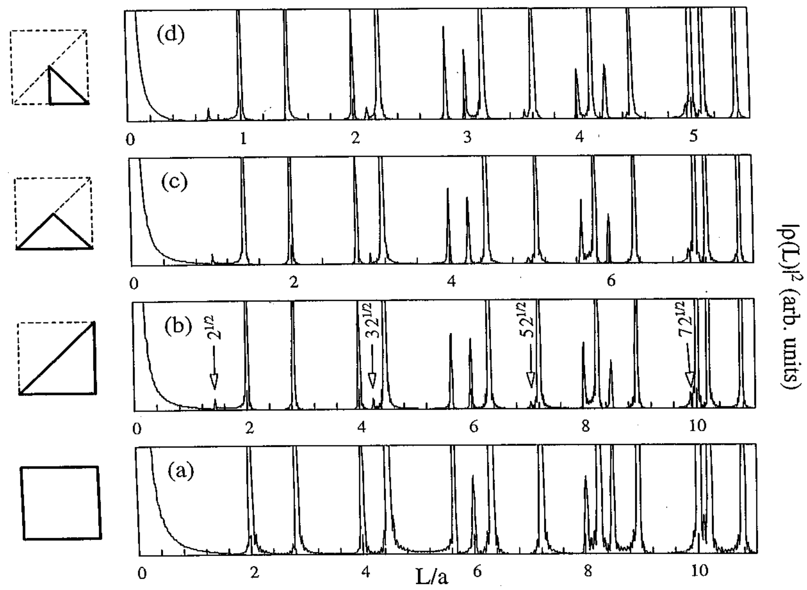

FIG. 7. Fourier transform of the energy level density, $|\rho(L)|^2$ [defined by Eq. (3)] versus L/a for the full square well (a), the once-folded triangular well (b), the twice-folded triangular well (c), and the three-times-folded triangular well (d). Note that the path length spectrum of the once-folded well is basically identical to the square well except for the appearance of new small features at $L/a=n\sqrt{2}$ where n is odd, corresponding to the isolated back-and-forth orbits shown in Fig. 5(b). Cases (c) and (d) are shown with the horizontal axis scaled down by factors of $\sqrt{2}$ and 2, respectively, to show that the same structures are visible for all three isosceles triangles, but with L scaled appropriately.

where $i,j=1,2,3,\ldots$. For future reference we note that odd (even) values of i,j correspond to wave functions which are even (odd) with respect to a parity operation about horizontal or vertical lines through the center of the square. One of the most important features of this spectrum is the large degree of degeneracy, since $E_{(i,j)}=E_{(j,i)}$, and a symmetry group analysis of this double degeneracy has recently been presented in Ref. 16.

A standard periodic orbit theory analysis applied to such a billiard system then proceeds by taking a large number of low-lying energy eigenvalues (in this case we use $N_{\mathrm{TOT}}=4000$) and evaluating Eq. (3) using the corresponding k_n in Eq. (11). The resulting plot of $|\rho(L)|^2$ versus L/a is shown (at the bottom) in Fig. 7, with the (a) label corresponding to the initial square well. All of the expected features at values of $L/a=\sqrt{n_x^2+n_y^2}$ (and recurrences) are observed in the region shown.

C. Triangular foldings of the square well

The basic square well geometry can be folded in a variety of ways which result in soluble quantum mechanical and classical systems and the ''triangular'' foldings illustrated in Figs. 4, 5 and 6 are closest in spirit to the half-circular well case considered above. If we fold the square well footprint along a diagonal, we obtain an isosceles triangle and the allowed set of closed orbits in the full well is reproduced in this triangular half-well; this is shown in parts (a) and (b) of those figures where we consider cases corresponding to the $(n_x,n_y)=(1,0)/(0,1)$ (Fig. 4), (1,1) (Fig. 5), and (1,2)/(2,1) (Fig. 6) cases, respectively. In each case shown there, an allowed orbit in the square well is supported by the triangular well with the identical overall path length. The only new feature arises from a special isolated orbit of the type shown in Fig. 5(b) (as the vertical dot-dash line) for the (1,1) case. This back-and-forth orbit is very similar to that seen in the half-circular well case and should give rise to a small feature at integral multiples of $L/a=\sqrt{2}$ and the *odd* multiples $(1,3,5,\ldots)$ of this feature will not be hidden under the recurrences of the more standard $(n_x,n_y)=(1,1)$ features at integral multiples of $L/a=2\sqrt{2}$.

Although it is seldom stressed in textbooks, the energy eigenvalues and eigenfunctions of the isosceles triangular well can be easily obtained from those of the square well. Taking for definiteness a square well defined over the intervals $x,y \in (0,a)$, we can form a triangular well by restricting the area further to satisfy $y \leqslant x$ with the $y = x$ line defining another infinite potential wall. While eigenfunctions of the type in Eq. (12) do not satisfy the necessary boundary conditions, appropriate linear combinations of pairs of degenerate solutions, defined by

$$\psi^{(b)}_{(i,j)}(x,y) \;\propto\; \psi_{(i,j)}(x,y) - \psi_{(j,i)}(x,y) \tag{13}$$

(with $i > j$ and $x \geqslant y$), do vanish on all infinite wall boundaries and so are the appropriate wave functions. The corresponding energy eigenvalues are then given by Eq. (11), but with the restriction that $i > j$. In this way, a large amount of the degeneracy which is obvious in the square well is removed. (In fact, linear combinations of just this type have been recently used in discussions of the symmetry group for the square well potential.[16]) The ground state probability density in the triangular well [corresponding to $(i,j) = (2,1]$, as well as its ''precursor'' in the full well are shown in Fig. 8 (left column) as a contour plot; the wave function in the once-folded triangular well (lower left) is more peaked (note the contours!), since in order to be properly normalized in the smaller well it must have a larger magnitude.

In this regard, the energy spectrum of the triangular folding is obtained in a very similar manner to the half-circular well considered above. The nondegenerate states ($m = 0$ in the circular case, $i = j$ in the square case) are not used and one-half of the remaining degenerate eigenvalues ($m \neq 0$ in the circular case, $i \neq j$ in the square case) are retained. Because of this, much of the original degeneracy of the system is removed. Some of the original accidental degeneracies[16] are also eliminated, such as those degenerate states labeled by $(i,j) = (1,7)/(5,5)$ or $(7,17)/(13,13)$,

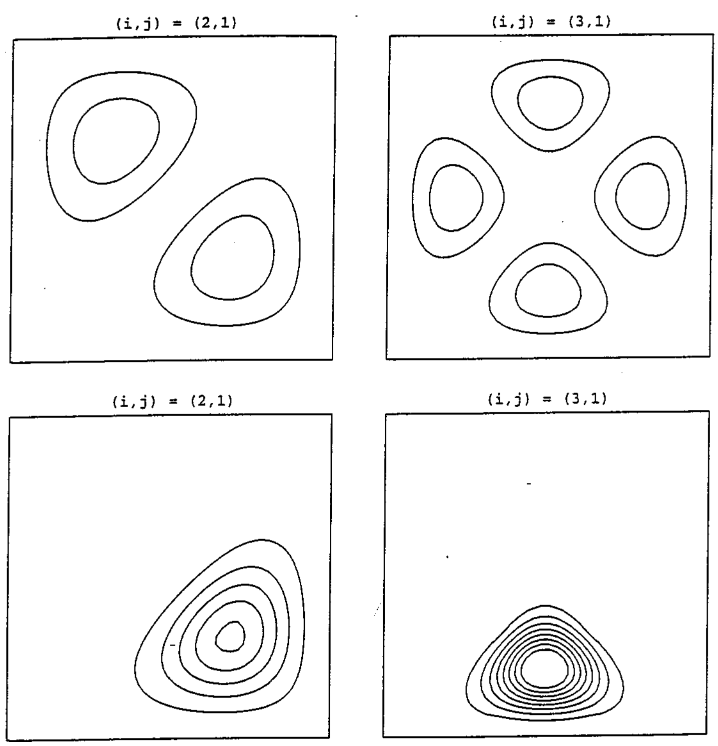

FIG. 8. Contour plots of the full square well probability densities from Eq. (13) for $(i,j) = (2,1)$ (upper left) and (3,1) (upper right). The corresponding ground state wave functions for the once-folded triangular well (lower left) and twice-folded triangular well (lower right) are also shown.

TABLE I. Energy eigenvalues, $E_n = E_{(i,j)} = (\hbar^2\pi^2/2ma^2)(i^2+j^2)$, for the triangular wells resulting from the first (b), second (c), and third (d) triangular foldings of the square well, scaled to the value of $E_0 = \hbar^2\pi^2/2ma^2$. The values of (i,j) for the second-folded triangular well are given by $(i+j, i-j)$ which explains why $E_n^{(c)} = 2E_n^{(b)}$.

First triangle		Second triangle		Third triangle	
(i,j) $i>j$	$E_n^{(b)}/E_0$	(i,j) $i>j$ and either both even or both odd	$E_n^{(c)}/E_0$	(i,j) $i>j$ and both even	$E_n^{(d)}/E_0$
(2,1)	5	(3,1)	10	(4,2)	20
(3,1)	10	(4,2)	20	(6,2)	40
(3,2)	13	(5,1)	26	(6,4)	52
(4,1)	17	(5,3)	34	(8,2)	68
(4,2)	20	(6,2)	40	(8,4)	80
(4,3)	25	(7,1)	50	(8,6)	100
(5,1)	26	(6,2)	52	(10,2)	104
(5,2)	29	(7,3)	58	(10,4)	116
(5,3)	34	(8,2)	68	(10,6)	136
(6,1)	37	(7,5)	74	(12,2)	148

because (i,i) states are no longer present. However, the majority of such accidental degeneracies are still present, for example, those of the form (4,7)/(1,8), (2,9)/(6,7), and (5,20)/(8,19)/13/16). (We note that a large fraction of such accidental degeneracies are, in fact, lifted in the square well variant considered in Sec. II E as discussed in detail in the Appendix.)

Given these simple connections between the energy eigenvalues for the full square well and its triangular folding, it is then perhaps not surprising that when we evaluate $\rho(L)$ using this energy spectrum, we obtain the result shown in Fig. 7(b) which shows all of the same path length features as the full square well, but with additional small features obvious at $L/a = \sqrt{2}, 3\sqrt{2}, 5\sqrt{2}$; the next feature at $L/a = 7\sqrt{2} \approx 9.9$ is almost hidden under the more standard 5(1,0) and 1(3,4) peaks which fall at $L/a = 10.0$.

Further foldings of the isosceles triangular well along a perpendicular bisector result in congruent isosceles triangular footprints, but with an area of half the size and edges a factor $1/\sqrt{2}$ as long. The closed paths allowed in the first-folded wells in Figs. 4(b), 5(b), and 6(b), upon a second folding, are mirrored into periodic orbits in the smaller versions of the triangular wells, but corresponding to a different set of (n_x, n_y) labels. For example, in going from (b) to (c) in Figs. 4, 5, and 6, we find that the (n_x, n_y) labels correspond to $(1,0)/(0,1) \rightarrow (1,1)$, $(1,1) \rightarrow (1,0)/(0,1)$, and $(1,2)/(2,1) \rightarrow (3,1)/(1,3)$ respectively. Thus, the entire set of allowed orbits are indeed reproduced, with path length features reduced by a factor of $1/\sqrt{2}$.

The quantum mechanical eigenvalues in the ''second-folded'' triangular well can also be obtained from the first-folded triangular well by further restrictions in the allowed values of i,j using symmetry arguments. If one restricts the second-folded well to lie below the line $y = a - x$ (i.e., the other diagonal of the original square), then a subset of the states in Eq. (13) will still be solutions in the restricted well and also satisfy the necessary boundary conditions on the new infinite wall boundary. Specifically, those $i>j$ states with (i,j) either both even *or* both odd will have a nodal line along $y = a - x$ and comprise the allowed solution space. We compile, in Table I, the lowest-lying energy eigenvalues with the corresponding (i,j) labels for the first-folded (b) and second-folded (c) triangular wells. To visualize the new ground state wave function in the second-folded well, we show in Fig. 8 (right column) the $(i,j) = (3,1)$ state, both in the full precursor well (upper right) and in the second-folded triangular well (lower right). Note that the ground state wave function is even more highly peaked than in the larger triangular well due to normalization constraints, but its shape is identical to that in the first-folded triangular case.

From Table I, we see that the pattern of allowed i,j values in the second-folded well is given by $(i+j, i-j)$ in which case the energies are related to those in the first-folded well by

$$E^{(c)}_{(i,j)}=\frac{\hbar^2\pi^2}{2ma^2}[(i+j)^2+(i-j)^2]=2\left[\frac{\hbar^2\pi^2}{2ma^2}(i^2+j^2)\right]=2E^{(b)}_{(i,j)}. \tag{14}$$

[The combination $(i+j,i-j)$ which arises in this case is easily seen, either via Fig. 8 or using analytic constructions, to be connected to the fact that the wave functions in the once-folded well can be related to those in the twice-folded well by rotations through various multiples of 45° and subsequent scaling of lengths by factors of $1/\sqrt{2}$.] So, since the energy spectrum in the second-folded well consists of values which are exactly twice those in the once-folded well, i.e., $\{E_n^{(c)}\}=\{2E_n^{(b)}\}$, we also know that the allowed wave numbers satisfy $\{k_n^{(c)}\}=\{\sqrt{2}k_n^{(b)}\}$. This simple scaling result implies that the Fourier transform defining $\rho(L)$ in case (c) will then have the identical spectral features in case (b), simply scaled down in L by the appropriate factor of $1/\sqrt{2}$. We illustrate this in Fig. 7 where the horizontal $\rho(L)$ ''spectrum'' for part (c) is identical to that for part (b) when plotted with the appropriately scaled value of L/a.

Finally, additional foldings along axes of symmetry result in ever smaller isosceles triangular wells, with appropriately smaller path lengths and increasingly large energy eigenvalues. For example, the third-folding (not shown in Figs. 4, 5, or 6) will give path length features at half the original lengths. The energy eigenvalues which lead to this are obtained from the original well by further restricting the $E^{(b)}_{(i,j)}$ to (i,j) values which are *both* even and the resulting energy spectrum (see Table I once again) consists of values which are obviously 4 times the original, with twice the resulting k_n values and path length features at half the precursor values. Examples of such path length features are shown in Fig. 7(d).

We have also performed a scaling analysis of $\rho(L,N_{\text{TOT}})$ versus N_{TOT} similar to that for the half-circle as shown in Figs. 2 and 3, and we find very similar results, specifically, the scaling coefficient for the standard, nonisolated features is completely consistent with $\alpha=3/2$ for all of the path length features shown in Fig. 7.

D. Rectangular foldings of the square well

As another simple example of a folding of the square well, let us examine both the energy eigenvalue spectrum and the periodic orbit structure of the rectangular well of dimensions $a\times a/2$ obtained by restricting ourselves to the region of the initial square well defined by y

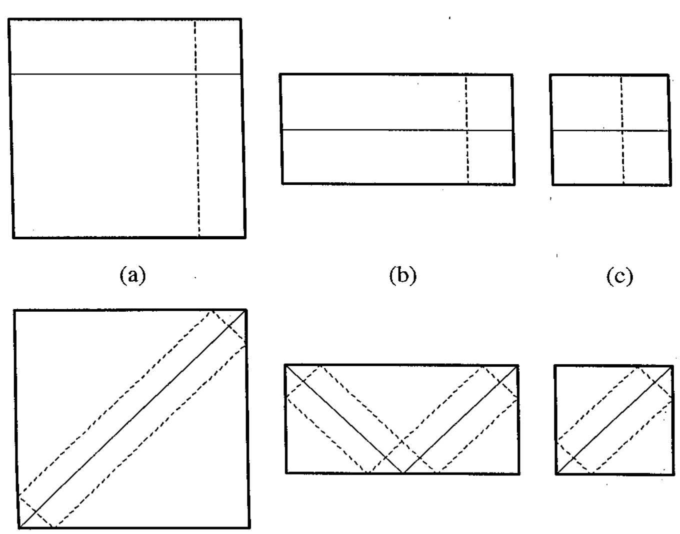

FIG. 9. Typical periodic orbits for the square well (a), the once-folded rectangular well (b), and the twice-folded square well of half the original size (c). The plots correspond to the integer labels $(n_x,n_y)=(1,0)$ (solid) and (0,1) (dashed) at the top, and to several versions of $(n_x,n_y)=(1,1)$ on the bottom.

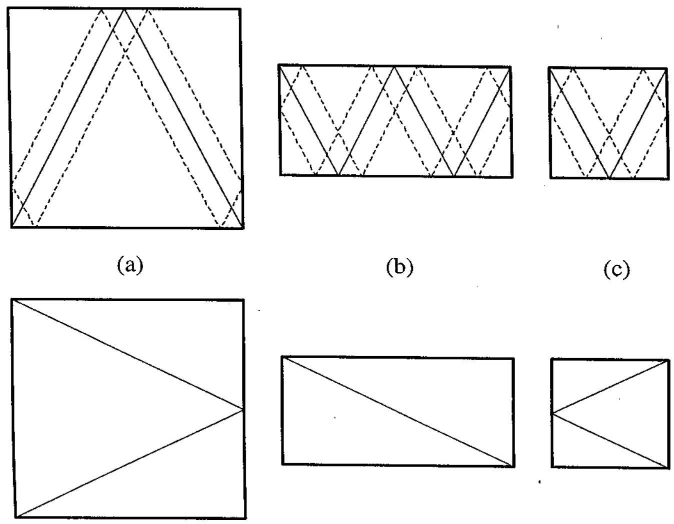

FIG. 10. Same as Fig. 9, but for $(n_x, n_y) = (1,2)$ (top) and (2,1) (bottom).

$\geqslant a/2$, i.e., the upper half. (We will find no isolated orbits in this ''folding,'' but we consider this case for completeness.) The wave functions in Eq. (12) are still solutions of the appropriate Schrödinger equation within the restricted well and those with even j values also satisfy the boundary condition that $\psi(x,y=a/2)=0$ on the new boundary. Thus, the appropriate energy eigenvalues are still given by Eq. (11), but with j restricted to be an even integer.

The corresponding path lengths, given by Eq. (10), can be written in the form

$$L_{(b)}(n_x,n_y)=2\sqrt{(n_x a)^2+n_y(a/2)^2}=\tfrac{1}{2}[2a\sqrt{(2n_x)^2+n_y^2}]=\tfrac{1}{2}L_{(a)}(2n_x,n_y;a). \tag{15}$$

The changes in path lengths for allowed orbits can also be visualized by folding over allowed trajectories in the square box as shown in Figs. 9 and 10. The corresponding spectrum of allowed path lengths, L, will then be somewhat similar to that of the square well, but with the following changes:

(i) For each path length feature in the full square well characterized by a value of (n_x, n_y), there will be a corresponding feature in the once-folded, rectangular well with half the path length, provided that n_x and n_y are not both odd.

This pattern can be confirmed by evaluating $\rho(L)$ using Eq. (2), but with the $N_{\mathrm{TOT}}=4000$ lowest eigenvalues given by Eq. (11), but where j is an even integer. The resulting plot for the rectangular well is shown in Fig. 11, directly above that for the square well, and with an appropriately scaled horizontal axis. One sees immediately that the resulting path length spectrum is as described above, similar to the square well result, scaled down by a factor of 2, but with the appropriate $(n_x, n_y) = (\mathrm{odd}, \mathrm{odd})$ peaks missing.

A subsequent second folding of the rectangular well into a square well of one-quarter the area of the original square is an obvious continuation of the pattern. The resulting possible path lengths can be obtained by folding or reflections as shown in Figs. 9(c) and 10(c) and are, of course, identical to the original well, but with path lengths given by

$$L_{(c)}(n_x,n_y)=2\sqrt{(n_x a/2)^2+(n_y a/2)^2}=\tfrac{1}{2}[2a\sqrt{n_x^2+n_y^2}]=\tfrac{1}{2}L_{(a)}(n_x,n_y), \tag{16}$$

i.e., simply scaled down by the appropriate factor of 2. The energy eigenvalues are given by Eq. (11), but in order to satisfy the boundary conditions along $y=a/2$ and $x=a/2$, we require that *both i and j* be even integers, so that $\tilde{i}=2i$, $\tilde{j}=2j$ which gives

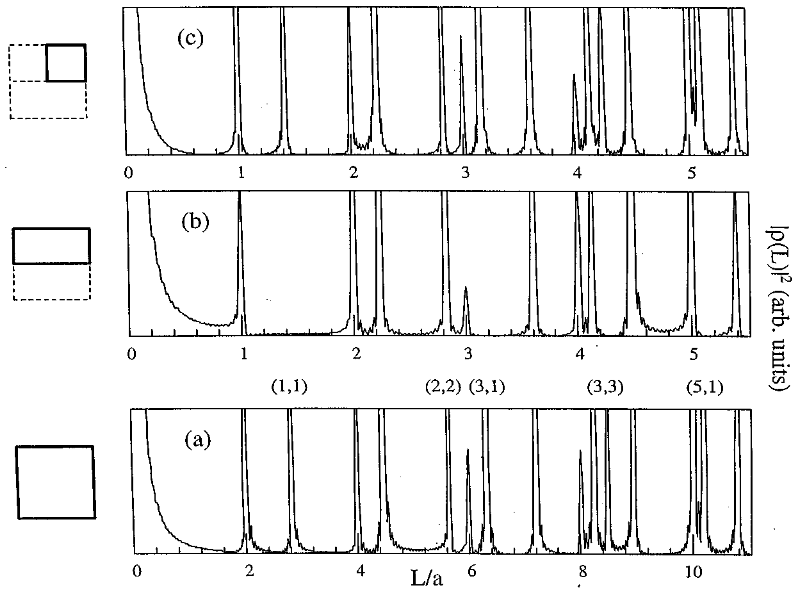

FIG. 11. Fourier transform of the energy level density, $|\rho(L)|^2$ [defined by Eq. (2)] versus L/a for the full square well (a), the once-folded rectangular well (b) and the twice-folded square well of half the original size (c). Note the different horizontal scales used in (b) and (c) compared to case (a). Labels corresponding to (n_x, n_y) both odd are shown.

$$\tilde{E}_{(\tilde{i},\tilde{j})}=\frac{\hbar^2\pi^2}{2ma^2}(\tilde{i}^2+\tilde{j}^2)=4\left[\frac{\hbar^2\pi^2}{2ma^2}(i^2+j^2)\right]=4E_{(i,j)} \tag{17}$$

so that the wave numbers required in Eq. (2) satisfy $\tilde{k}_n=2k_n$. This immediately implies that the $\rho(L)$ evaluated with this $\tilde{k}_n$ spectrum will have peaks at values of $\tilde{L}=L/2$ as expected. Additional foldings will continue to reproduce these patterns, but with 2×1 rectangles and squares of increasingly smaller dimensions. For example, after $2n$ foldings, the allowed energy eigenvalues in the smaller square well will, by symmetry considerations alone, be given by $\tilde{E}_{(\tilde{i},\tilde{j})}=(2n)^2E_{(i,j)}$ and wave numbers related by $\tilde{k}_n=(2n)k_n$ immediately implying square well path length features given by $\tilde{\rho}_{(2n)}(L)=\rho(L/2n)$.

E. Another variant of the square well

Other variants of the square well which also exhibit induced isolated periodic orbits are possible beyond those derived by foldings and the next such system we consider is the square well with an ''infinite-strength'' repulsive δ-function potential at the center. This variant is similar in spirit to the study of the annular circular well or circular disk considered in Ref. 17 with the radius of the inner annulus made vanishingly small. It can also be obtained by adding a repulsive potential of the form

$$V_{\mathrm{REP}}(x,y)=+\tilde{g}\,\delta(\mathbf{r}-\mathbf{r}_0), \tag{18}$$

where $\mathbf{r}_0$ defines the center of the square well and we then let $\tilde{g}\to\infty$.

The structure of the classical periodic orbits in this slightly modified system is clear. Any standard orbit of the type in Sec. II B will still be allowed, provided it misses the central obstruction. Exactly one isolated orbit for each value of (n_x, n_y) which *does* hit the center will also form a closed trajectory, but with half the path length. Thus, the path length spectrum exhibited in $\rho(L)$ should consist of the standard large spike features at values given by $L/a=2\sqrt{n_x^2+n_y^2}$ (and repetitions), but will also show much smaller features (similar in magnitude to the isolated path length features for special back-and-forth orbits in the half-circular and triangular wells) at values of L/a equal to half the standard ones. (We note that for a square well with a repulsive core consisting of a circular infinite barrier with a finite radius, one obtains a Sinai billiard[8] which is known to exhibit classical chaotic motion; we are therefore considering a very simple special case.)

A large part of the quantum mechanical energy spectrum for this problem can once again be trivially obtained by symmetry arguments. Energy eigenfunctions in the original square well with either *i or j* even will have an appropriate node at the middle of the well and will be unaffected by the presence of the central δ-function ''core.'' States characterized by (i,j) both *odd* can, in principle, be increased in energy, but once again symmetry arguments restrict the states which are affected. For such (odd,odd) states which have $i\neq j$, we can construct two linear combinations of the standard eigenfunctions in Eq. (12), namely the one already discussed in Eq. (13) and the orthogonal state corresponding to the sum of the two basic states. The antisymmetric linear combinations vanish at the center of the well and so are unaffected by the δ-function addition; thus, something just less than (1/2) of the remaining (odd, odd) states remain eigenstates of the new problem since the $i=j$ (odd, odd) states will be affected. There then remain a set of states characterized by (i,j) with $i\leqslant j$ and both odd which will be shifted up in energy, but in contrast to the much simpler one-dimensional case, the resulting spectrum is not easily related to the standard spectrum in the same obvious way. (The one-dimensional case[18] is reviewed in the Appendix as a model upon which we base our solution for the 2D problem of interest here.) Crudely speaking then, something like $(3/4)+(1/8)_{-}\approx 7/8$ of the energy values in the square well plus core problem are identical to the standard square well and will obviously contribute to the large array of standard path length features when used to evaluate $\rho(L)$. In order to obtain estimates for the remaining energy eigenstates which *are* affected by the central core, we will make use of the matrix mechanics approach which is discussed at some length in the Appendix. (We note that a perturbative approach is not useful in the case when $\tilde{g}\rightarrow\infty$, but we can check the results of the matrix approach for small values of $\tilde{g}$ using perturbation theory as a consistency check on our method; we find complete agreement in that limit.)

Using the lowest 800 eigenvalues (roughly 100 or 1/8 of which are those affected by the additional δ-function), we numerically evaluate $\rho(L)$ as usual and obtain the results shown in Fig. 12. The spectrum including the infinite-strength δ-function is shown in part (b) and does indeed exhibit many new features [most of which are labeled by the appropriate values of (n_x,n_y)] at half the ordinary path length values familiar from the ordinary 2D well spectrum which is shown in Fig. 12(a) for comparison. The new spectrum is also shown in part (c) on an expanded scale. In this case there is now a large array of new (smaller) features corresponding to a large class of distinct new isolated orbits.

F. Supersymmetric version of the square well

We have focused extensively in this report on the connections between systems with geometries which are related by simple symmetry operations and have found, not surprisingly, that the structure of the allowed space of periodic orbits and the quantized energy eigenvalues are highly correlated with one another. In several such cases, we have found that the energy spectra in two related systems are almost identical to one another, but with special strings of eigenvalues not appearing in one case or the other. One of the most familiar examples of a symmetry for which this pattern is known to happen quite generally is found in supersymmetric quantum mechanics.[19] In more familiar one-dimensional supersymmetry (SUSY) problems, the energy spectra of two superpartner potentials, $V_{(-)}(x)$ and $V_{(+)}(x)$, are identical except that the zero-energy ground state in $V_{(-)}(x)$ is missing for $V_{(+)}(x)$. While commonly discussed for one-dimensional problems, it is just as easy to derive the supersymmetric pair potentials corresponding to the two-dimensional square well under discussion here. Since the problem is separable, we need only consider the 1D SUSY partner potentials for the standard square well[20] and combine results.

In order to obtain a ground state solution with vanishing energy (necessary for the construc-

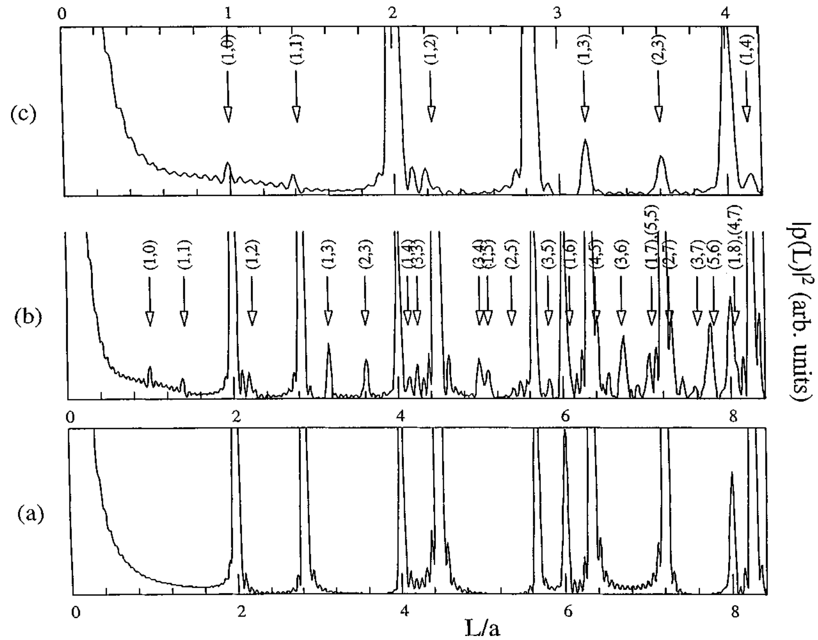

FIG. 12. Comparison of $|\rho(L)|^2$ versus L/a for the standard square well [bottom plot labeled (a)], and the square well with an infinite-strength δ-function spike at the center (middle and top plots). New features due to isolated orbits which hit the central spike are present in case (b) and (c) at half the path lengths of the more standard nonisolated orbits. Case (c) is identical to case (b), but with an expanded scale for easier comparison.

tion of the superpotential), we subtract the ground state energy from the standard 1D infinite square potential ($V_\square(x)$) thus defining

$$V_{(-)}(x)=V_\square(x)-\frac{\hbar^2\pi^2}{2ma^2} \tag{19}$$

with a similar result for the y coordinate. The resulting energy spectrum is then given by

$$E^{(-)}_{(i,j)}=\frac{\hbar^2\pi^2}{2ma^2}[i(i+2)+j(j+2)], \quad \text{where} \quad i,j=0,1,2,3,\ldots . \tag{20}$$

The supersymmetric partner potentials for either coordinate can be derived using standard results[20] and we find that

$$V_{(+)}(x)=\frac{\hbar^2\pi^2}{2ma^2}\left[2\csc^2\left(\frac{\pi x}{a}\right)-1\right] \tag{21}$$

so that $V_{(+)}(x,y)=V_{(+)}(x)+V_{(-)}(y)$.

A contour plot of this potential is shown in Fig. 13 and the limiting case of very high energies approaches the dashed (square well) boundary shown; thus, when $E^{(+)}_{(i,j)}\gg\hbar^2\pi^2/2ma^2$ we recover the familiar 2D square well geometry. If, for example, we set $(\hbar^2\pi^2/2ma^2)(i(i+2)+j(j+2))=V_{(+)}(x,y)$, the distance from the (smooth) corner of the corresponding contour of $V_{(+)}(x,y)$ to the (sharp) corner of the $V_{(-)}(x,y)$ potential is approximately

$$x\approx a\left(\frac{2}{\pi\sqrt{i^2+j^2}}\right)\to 0 \quad \text{as} \quad i,j\to\infty. \tag{22}$$

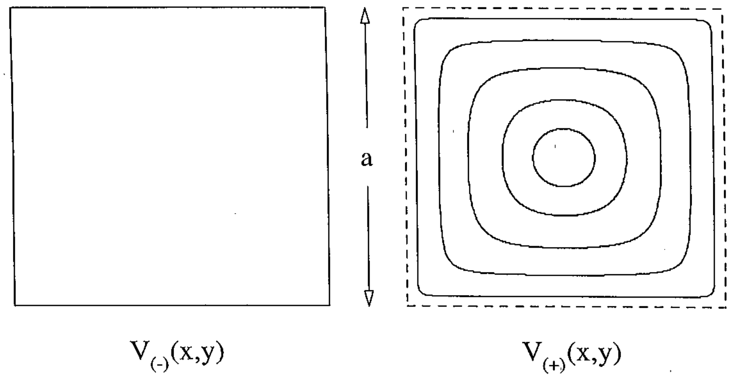

FIG. 13. Plots of the supersymmetric partner potentials defined by the two-dimensional square well problem. A contour plot of the supersymmetric partner potential, $V_+(x,y)$ [defined by Eq. (21)], is shown on the right along with the dashed (square well) boundaries which is the limiting case for $E \gg \hbar^2\pi^2/2ma^2$.

The corresponding energy eigenvalues are given by Eq. (20), but now with the restriction that $i,j=1,2,3,\ldots$ since the zero-energy ground state in each direction is excluded.

An obvious question is how the periodic orbits in the two wells are related to each other, since given Eq. (1), the energy level densities, $\rho(E)$, for the two systems are very highly correlated. Because the SUSY partner potential defined by $V_{(+)}(x,y)$ is not a purely free-particle billiard-like system, we cannot use Eq. (2) directly. However, because the system does approach the square well case in the limit of high energies, we do expect a very close correspondence between the structure of the periodic orbits in the two systems. In order to model what the likely differences are in the two model systems, we calculate $\rho(L)$ using Eq. (2), but instead of using the full energy spectrum in Eq. (20), we evaluate $\rho(L)$ using the energy (and wave number) values with the case $i,j=0$ excluded. We compare the two evaluations of $\rho(L)$ in Fig. 14 the dashed (solid) curves correspond to the $V_{(-)}$ ($V_{(+)}$) spectrum.

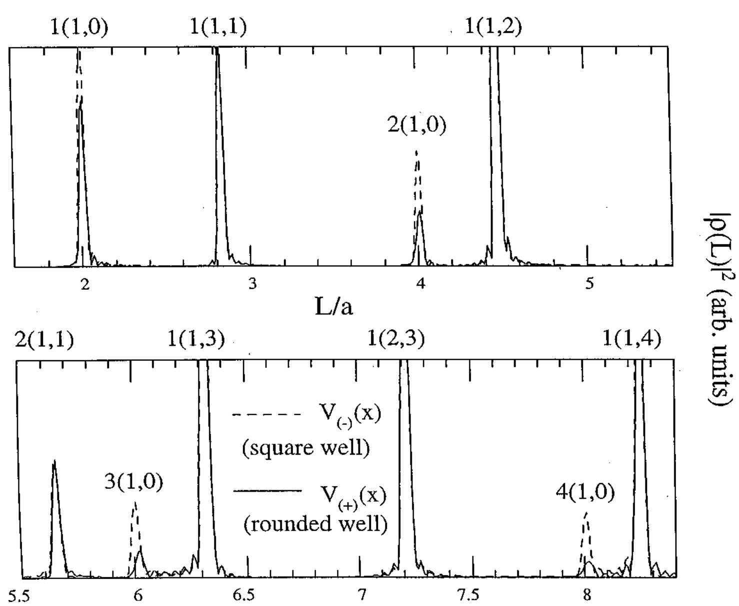

FIG. 14. Comparison of $|\rho(L)|^2$ versus L/a using the energy eigenvalues, $E^{(-)}_{(i,j)}$ with $i,j=0,1,2,\ldots$ defined by Eq. (20) (dashed curve) versus those in the supersymmetric partner potential, $E^{(+)}_{(i,j)}$ for $i,j=1,2,3,\ldots$ (solid curve). Only the path length features corresponding to the (1,0),(0,1) purely horizontal or vertical closed orbits are seemingly affected.

Perhaps not surprisingly, the only observable differences are the relative importance of the (1,0)/(0,1) features corresponding to purely back-and-forth or up-and-down motions. In the full square well, the paths characterized by these integer labels form a continuous set and so appear as ''full spikes'' in the $\rho(L)$ spectrum; in the SUSY partner case, the relative importance of these features, and especially of their recurrences, become increasingly smaller. This can likely be attributed to the fact that only purely horizontal or vertical closed paths *through* the center of the ''rounded'' $V_{(+)}(x,y)$ potential will support closed orbits and paths which are only *near* the center will increasingly become ''defocused,'' repetition after repetition. We reiterate that this is not, in any sense, a complete or rigorous analysis of this problem, but it does point out that, because of Eq. (1), there will likely be a very interesting relationship between the classes of allowed periodic orbits in superpartner potentials. In cases where the 2D potential is separable, as here, we do expect the general pattern of energy level differences, namely that the $E^{(+)}$ spectrum will be identical to the $E^{(-)}$ one, but with two one-dimensional arrays of eigenvalues missing (those corresponding here to $i,j=0$.) Examples for further study could also include the SUSY partner potentials for the circular infinite well, in which case one would likely have to construct $V_{(-)}^{(m)}(r)$ and $V_{(+)}^{(m)}(r)$ potentials (for the corresponding radial equation) separately for each value of the orbital angular momentum quantum number *m*.

One can already see, in broad strokes, how the construction of a superpartner potential, $V_{(+)}(x,y)$, from a zero-energy version of any infinite well potential, $V_{(-)}(x,y)$, using superpotentials will ensure that the resulting $V_{(+)}(x,y)$ will have the same set of infinite wall boundaries for large energies. The vanishing of the ground state solution in $V_{(-)}(x,y)=V_{(-)}(\mathbf{r})$ on the boundary (labeled *C*), schematically given by $\psi(\mathbf{r})_C=0$, ensures that the superpotential, which consists of factors such as $\psi'(\mathbf{r})/\psi(\mathbf{r})$ will give a partner potential which will diverge (due to the vanishing denominator) on the boundary of the original infinite well geometry. This connection between the two potential wells is a specific example of the shape invariance relationship for SUSY partner potentials discussed in Ref. 20 and a general analysis of the periodic orbit theory interpretation of supersymmetric quantum mechanics will, no doubt, rely heavily on this property.

G. Conclusions

We have studied two folded versions of standard two-dimensional billiard systems, namely the half-circular well and the ''triangular'' folding of the square well, and found that the changes in the energy eigenvalue spectrum and the new classical periodic orbits induced by the folding are very similar. We have discussed the scaling properties of the isolated and nonisolated orbit features as a function of N_{TOT} and have found that the values of $|\rho(L)|^2$ for standard, nonisolated path length features increase as $(N_{\mathrm{TOT}})^{3/2}$ in both cases, at least with the simple normalization used in Eq. (3); the isolated path ''spikes'' also exhibit well-separated local maxima in plots of α versus *L*. In each case, with the largest number of states used, the only new features obvious in the $\rho(L)$ spectra are those corresponding to the back-and-forth orbits and these new features corresponding to isolated orbits continue to be apparent above the background as N_{TOT} is increased.

For the case of the square well plus δ-function, we find that a much larger fraction of the energy eigenstates are affected by the introduction of the central barrier with a corresponding effect on the number of new isolated orbits. Compared to the folded cases where a one-dimensional set of energy eigenvalues is changed (discarded) and a single isolated periodic orbit (and its repetitions) are induced, in this case a significant fraction of the energy spectrum is altered and an infinite series of new path length features (as well as their repetitions) are then observed in the $\rho(L)$ spectrum.

Finally, we have discussed some possible connections between the periodic orbits in pairs of partner potentials related by supersymmetry and the corresponding energy eigenvalue spectra, focusing on the 2D square well and its superpartner as a specific example, in the context of the other cases considered here. We consider this discussion as motivation for a more thorough study of this interesting connection.

APPENDIX A:

For the discussion of the square well variant in Sec. II E, we require the energy eigenvalue spectrum for the case of a square well of side a with the addition of an infinite strength, repulsive δ-function potential of the form

$$V_{\mathrm{REP}}(x,y)=+\tilde{g}\,\delta(\mathbf{r}-\mathbf{r}_0)=+\tilde{g}\,\delta(x-a/2)\,\delta(y-a/2), \tag{A1}$$

where we let $\tilde{g}\to\infty$. To gain some guidance on how to solve this problem, we can consider the simpler one-dimensional analog of this case, namely a particle in a 1D well with an additional potential

$$\tilde{V}(x)=g\,\delta(x-a/2). \tag{A2}$$

This problem is solvable analytically (by simple matching of the appropriate boundary conditions at the central δ-function) and one can therefore take the limit where $g\to\infty$ exactly. [We note that the dimensions of $\tilde{g}$ (in 2D) and g (in 1D) are different.] In that case one finds that the 1D states labeled by odd values of the integer i, i.e., those states which are nonvanishing at the center of the well and are thus affected by the δ-function at all, are increased in energy until they become degenerate with the even i states just above them in the spectrum. This problem can also be approached by using perturbation theory, with expansions up at least third-order[21] having appeared in the literature. In our case, where we are interested in the nonperturbative limit where $g\to\infty$, this approach is not useful, but we can check our matrix results with a similar perturbative expansion in the limit of small $\tilde{g}$ as a consistency check on our method.

For the 2D case under consideration in this study, an analytic solution is not possible since the potential defined by Eq. (A1) is not separable and a different approach must be followed. A relatively straightforward matrix mechanics approach to the 1D problem[22] can be extended to the two-dimensional case under consideration and that is the approach we follow in obtaining our estimates of the energy eigenvalue spectrum used in Sec. II E. In this Appendix, we briefly describe the method used in the 1D case and then discuss how it is extended to the two-dimensional problem.

In the purely one-dimensional case, we only consider the basis states given by

$$\psi_{(i)}(x)=\sqrt{\frac{2}{a}}\sin\left(\frac{i\pi x}{a}\right) \tag{A3}$$

with i odd as these are the states which are affected by the new potential. The matrix elements required are given by

$$\mathcal{H}_{ij}(g)=\langle\psi_i|\hat{H}_0+g\,\delta(x-a/2)|\psi_j\rangle=(i^2E_0)\,\delta_{i,j}+\Delta\,\sin\left(\frac{i\pi}{2}\right)\sin\left(\frac{j\pi}{2}\right), \tag{A4}$$

where

$$E_0=\frac{\hbar^2\pi^2}{2ma^2}\quad\text{and}\quad\Delta=\Delta(g)=\frac{2g}{a} \tag{A5}$$

or in matrix form

TABLE II. Estimated energy eigenvalues for the one-dimensional square well plus repulsive δ-function potential of the form Eq. (A2) using the matrix diagonalization method discussed in the Appendix. The unperturbed energies are characterized by odd values of i and are given by $E_{(i)}=i^2E_0$, where $E_0=\hbar^2\pi^2/2ma^2$. The exact solutions as $g\rightarrow\infty$ correspond to each such odd state approaching the even state just above it in energy. We show the resulting estimate for increasing values of N_{TOT} and g as well as the ratio of the approximate result to the limiting value of $(i+1)^2E_0$.

Unperturbed energies (in units of E_0)	$N_{\mathrm{TOT}}=5$	$N_{\mathrm{TOT}}=10$	$N_{\mathrm{TOT}}=20$	$N_{\mathrm{TOT}}=40$	Exact values for $g\rightarrow\infty$
1	4.34 (1.08)	4.17 (1.04)	4.08 (1.02)	4.04 (1.01)	4.00
9	17.41 (1.09)	16.67 (1.04)	16.39 (1.02)	16.16 (1.01)	16.00
25	39.38 (1.09)	37.53 (1.04)	36.74 (1.02)	36.37 (1.01)	36.00
49	70.76 (1.10)	66.76 (1.04)	65.32 (1.02)	64.65 (1.01)	64.00

$$\mathcal{H}(g)=\begin{pmatrix} E_0+\Delta & -\Delta & +\Delta & \cdots \\ -\Delta & 9E_0+\Delta & -\Delta & \cdots \\ +\Delta & -\Delta & 25E_0+\Delta & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{A6}$$

In order to obtain increasingly accurate values for the eigenvalues of this system, we diagonalize $\mathcal{H}(g)$ for increasingly larger values of N_{TOT} where N_{TOT} is the dimensionality of the matrix. At the same time, we also use increasingly large values of g to approach the desired $g\rightarrow\infty$ limit. Specifically, we use a value of g given by

$$\frac{\Delta(g)}{E_0}=10\cdot(2N_{\mathrm{TOT}}-1)^2 \tag{A7}$$

which corresponds to off-diagonal terms, Δ, which are ten times larger than the largest diagonal energy term used in the matrix diagonalization. In this way, the lowest $N_{\mathrm{TOT}}-1$ eigenvalues approach the desired energies, while the largest eigenvalue increases without bound, going roughly as N_{TOT}^3, as we increase g via Eq. (A7); this special eigenvalue is an artifact of the diagonalization process we use. As an estimate of the convergence of this method as N_{TOT} and g are increased, we show the values of the first 4 eigenvalues (and their ratio to the expected values in the large g limit) for several values of N_{TOT} in Table II. We see that the convergence is fairly rapid and uniform.

For the two-dimensional case, we employ the same strategy, but as a basis set we use only those 2D wave functions which are affected by the presence of the central δ-function, namely those characterized by integer labels (i,j) where *both* i and j are odd. When $i=j$, we simply use the standard basis states given in Eq. (12), while for $i\neq j$, we use the linear combination which is orthogonal to the antisymmetric states given by Eq. (13), namely

$$\psi_{(i,j)}^{(+)}(x,y)\propto\psi_{(i,j)}(x,y)+\psi_{(j,i)}(x,y) \tag{A8}$$

as these states are the ones which are nonvanishing at the center of the potential well before the δ-function is added. (The normalized version of this combination is used.) We evaluate the Hamiltonian matrix just as above, sorting the states in order of increasing (unperturbed) energies as N_{TOT} increases. We show in Table III the results of such a resulting diagonalization process, up to values of $N_{\mathrm{TOT}}=80$, along with the corresponding values of energies for the $\tilde{g}=0$ case for comparison. The convergence is not quite so rapid as in the 1D case where the unperturbed energy eigenvalues are initially more separated in energy. For the numerical application required in Sec.

TABLE III. Estimated energy eigenvalues for the two-dimensional square well plus repulsive δ-function potential of the form Eq. (A1) using the matrix diagonalization technique discussed in the Appendix. The matrix diagonalization is made using basis states characterized by those states with i and j both odd; for the case of $i \neq j$, we use the symmetric combination wave function in Eq. (A8). For states with an accidental pairwise degeneracy (such as the (5,5) and (1,7) states), the numerical calculation finds the appropriate linear combination which vanishes at the origin (and is thus unaffected by the δ-function potential); we see this for the (1,7),(5,5) case where one eigenvalue given by $E=50E_0$ is always found, within the numerical accuracy of the program.

(i,j)	Unperturbed energies (in units of E_0)	$N_{\mathrm{TOT}}=10$	$N_{\mathrm{TOT}}=20$	$N_{\mathrm{TOT}}=40$	$N_{\mathrm{TOT}}=80$
(1,1)	2	3.46	3.22	3.05	2.92
(1,3)	10	13.45	12.87	12.44	12.12
(3,3)	18	20.17	19.71	19.40	19.19
(1,5)	26	29.69	29.06	28.60	28.25
(3,5)	34	40.92	39.03	37.92	37.20
(1,7)	50	50.00	50.00	50.00	50.00
(5,5)	50	55.02	54.24	53.69	53.26

II E, we make use of the lowest-lying 99 energy eigenvalues of this type in addition to as many of the ''unaffected'' energy eigenvalues as required so as to make a total of 800 energies in the numerical evaluation of Eq. (2).

As a check on the program used to numerically evaluate the eigenvalues and eigenvectors in this matrix approach (we use MATHEMATICA®), we note that when we use this set of basis functions, there still are a number of ''accidental'' pairwise degeneracies[23] present in the original spectrum; for example $(i,j)=(5,5)$ and (7,1) both give $E_{(i,j)}/E_0=50$ for the unperturbed states. We note that when we examine the resulting energy eigenvalues, the appropriate linear combination of such states is found which is *unaffected* by the δ-function perturbation and its energy

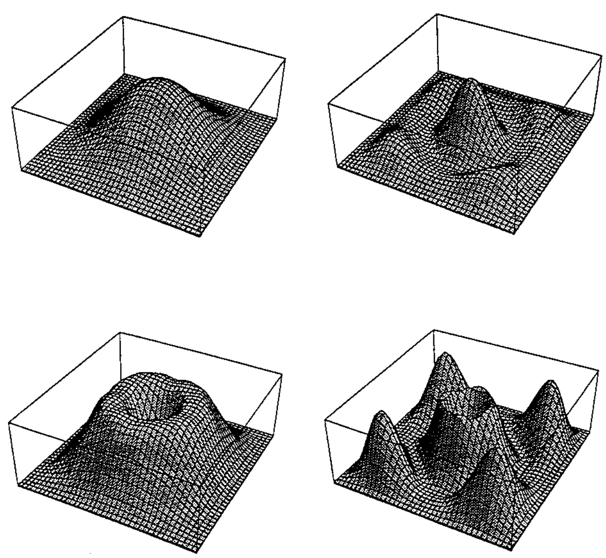

FIG. 15. Plots of $|\psi(x,y)|^2$ versus (x,y) for the two lowest energy states in the 2D well which are nonvanishing at the center of the well. The top row shows the $(i,j)=(1,1)$ state (the true ground state) on the upper left and the $(i,j)=(3,1)$ symmetric state [as defined in Eq. (A8)], for the unperturbed square well. The bottom row shows the same states now including the effect of the infinite δ-function at the origin given by Eq. (A1). In the evaluation of the perturbed eigenstates, the lowest-lying $N_{\mathrm{TOT}}=20$ states are used in the matrix diagonalization approach followed in the Appendix. We note that while each wave function is properly normalized, different scales are used in each figure to emphasize the shape of $|\psi(x,y)|^2$.

eigenvalue is, within machine errors, the same as the unperturbed value. For an example, see Table III for the case of $(1,7),(5,5)$.

As a final visualization of the effect of the infinite strength δ-function addition to the 2D well, we plot $|\psi(x,y)|^2$ versus (x,y) in Fig. 15 the $(i,j)=(1,1)$ and $(3,1)$ states, both without the δ-function (top row) and with $V_{\text{REP}}(x,y)$ added in the limit where $\tilde{g}\rightarrow\infty$ (bottom row). The presence of the cusp in the wave function at the center of the well due to $V_{\text{REP}}(x,y)$ is obvious.

[1] For one of the original references, see M. V. Berry and M. Tabor, Proc. R. Soc. London, Ser. A **356**, 375–395 (1977).
[2] For a recent and very readable review of semiclassical methods, including periodic orbit theory, see M. Brack and R. K. Bhaduri, *Semiclassical Physics* (Addison–Wesley, Reading, 1997).
[3] For an elementary review of these systems see R. W. Robinett, Am. J. Phys. **65**, 1167–1175 (1997), and many references therein.
[4] S. W. Macdonald and A. N. Kaufman, Phys. Rev. Lett. **42**, 1189 (1979); Phys. Rev. A **37**, 3067–3086 (1988).
[5] M. V. Berry, Eur. J. Phys. **2**, 91–102 (1981).
[6] R. van Zon and Th. W. Ruijgrok, Eur. J. Phys. **19**, 77–84 (1998).
[7] H.-J. Stockman and J. Stein, Phys. Rev. Lett. **64**, 2215–2218 (1990).
[8] Y. Sinai, Russ. Math. Surv. **25**, 137–189 (1970); Usp. Mat. Nauk **25**, 141–192 (1970).
[9] M. C. Gutzwiller, J. Math. Phys. **11**, 1791 (1970).
[10] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer, Berlin, 1990),
[11] R. Balian and C. Bloch, Ann. Phys. **69**, 76 (1972).
[12] T. P. Martin, Phys. Rep. **273**, 199 (1996).
[13] H. Friedrich and D. Wintgen, Phys. Rep. **183**, 37–39 (1989).
[14] R. W. Robinett, J. Math. Phys. **39**, 278–298 (1998).
[15] R. W. Robinett, ''Periodic orbit theory analysis of the circular disk or annular billiard: Nonclassical effects and the distribution of energy eigenvalues,'' Am. J. Phys. (to be published). The basic ''trick'' is to consider subsets of energy eigenstates with different values of $|m|/a_{(m,n_r)}$ which would correspond classically to varying values of $R_{\min}/R=\cos(\pi q/p)$ or ''circularity.'' For example, in the purely circular well case, for the (2,1) or (6,1) orbits one has $R_{\min}/R=0.0$ or 0.866, so that ''cuts'' on the $|m|/a$ values used in the evaluation of Eq. (3) centered around these values will enhance the features in $\rho(L)$ at $L/R=4.0,8.0,12.0$ or $6.0,12.0$, respectively.
[16] For a discussion at an elementary level, see F. Leyvraz, A. Frank, R. Lemus, and M. V. Andrès, Am. J. Phys. **65**, 1087–1094 (1997).
[17] R. W. Robinett, Surf. Rev. Lett. **5**, 519 (1998).
[18] See, for example, R. W. Robinett, *Quantum Mechanics: Classical Results, Modern Systems, and Visualized Examples* (Oxford University Press, New York, 1997), p. 191.
[19] For a review, see F. Cooper, A. Khare, and U. Sukhatme, Phys. Rep. **251**, 267–385 (1995).
[20] R. Dutt, A. Khare, and U. Sukhatme, Am. J. Phys. **56**, 163–172 (1988).
[21] I. R. Lapidus, Am. J. Phys. **55**, 172–174 (1987); see also S. T. Epstein, *ibid.* **28**, 495–496 (1959).
[22] See Ref. 18, pp. 237–239.
[23] W.-K. Lee, Am. J. Phys. **50**, 666 (1982).

Chaotic observables for a free quantum particle

Clasine van Winter
Department of Mathematics and Department of Physics and Astronomy, University of Kentucky, Lexington, Kentucky 40506

(Received 14 April 1998; accepted for publication 14 September 1998)

This paper is devoted to the time evolution of observables in the quantum mechanics of a single particle without interaction. It is assumed that wave functions belong to a certain set K^2 that is dense in L^2. The paper applies to observables represented by positive self-adjoint operators A on L^2 with the property that $A^{1/2}$ maps K^2 into L^2. Quadratic forms with form domains K^2 are used to generate a topology for operators A, defining a topological space X. The space X provides the framework to define sensitive dependence on initial conditions, topological transitivity, and existence of a dense set of periodic points, the three aspects of chaos in Devaney's definition of chaos for maps on metric spaces. It is shown that every neighborhood of every operator A in X contains operators that establish sensitive dependence on initial conditions, and similarly for the other aspects of chaos. Hence, the time evolution of operators in the Heisenberg picture is chaotic in the sense of this paper.
[S0022-2488(99)00101-2]

I. INTRODUCTION

Chaos in classical dynamical systems is characterized by three distinctive features.

Definition 1.1 (Devaney):[1] Let X be a set. A map $F:X\to X$ is said to be chaotic on X if

(1) F has sensitive dependence on initial conditions.

(2) F is topologically transitive.

(3) Periodic points are dense in X.

Definition 1.2 (Devaney):[1] Let X be a metric space. Let $F:X\to X$ be a map.

(1) F has sensitive dependence on initial conditions if there exists $\delta>0$ such that, for any $x\in X$ and any neighborhood V of x, there exists $y\in V$ and $n\geqslant 0$ such that $|F^{(n)}(x)-F^{(n)}(y)|>\delta$.

(2) F is said to be topologically transitive if for any pair of open sets $V,W\subset X$ there exists $n>0$ such that $F^{(n)}(V)\cap W\neq\emptyset$.

(3) A point $x\in X$ is a periodic point of period n if $F^{(n)}(x)=x$.

In this paper a version of chaos is introduced that applies to quantum mechanics. The points x become operators; the map F is replaced by the time evolution. The quantities corresponding to $F^{(n)}(x)$ are time-dependent operators representing observables in the Heisenberg picture. The time evolution of such observables is examined in a one-particle system without interaction.

Let f be a wave function in L^2; let H_0 be the Hamiltonian. In the Schrödinger picture the time evolution is represented by $f(t)=\exp(-iH_0t)f$. Let A be a self-adjoint operator on L^2. In the Heisenberg picture the time-dependent operator,

$$A(t):=\exp(iH_0t)A\exp(-iH_0t),$$

acts on time-independent wave functions f.

To find quantum analogs of the concepts in Definitions 1.1 and 1.2, we make some restrictive assumptions on wave functions f and observables A. It is assumed that f belongs to a set K^2 which is dense in L^2 in the L^2-norm, in addition to being a Hilbert space in its own right. If $f \in K^2$, then $\exp(-iH_0 t) f \in K^2$ if $t>0$, but not necessarily if $t<0$. Hence $t \geqslant 0$ in all of the following, corresponding to $n \geqslant 0$ in Definitions 1.1 and 1.2. If A corresponds to $x \in X$, then $A(t)$ corresponds to $F^{(n)}(x)$.

We only consider positive operators A. Hence, let A be a positive self-adjoint operator on L^2. Let $\{E(\lambda)\}$ be the corresponding family of spectral projections. Suppose we have a system in state f and we measure the observable represented by A. The probability of finding the measured value in a Borel set Δ is $\int_\Delta d(E(\lambda)f,f)$. If $A^{1/2} f \in L^2$, the expectation value of A is

$$\begin{aligned}\int_0^\infty \lambda \; d(E(\lambda)f,f) &= \int_0^\infty \lambda^{1/2} d_\lambda \int_0^\lambda \mu^{1/2} \; d(E(\mu)f,f)\\ &= \int_0^\infty \lambda^{1/2} \; d_\lambda \int_0^\infty \mu^{1/2} d_\mu (E(\lambda)f, E(\mu)f)\\ &= \int_0^\infty \lambda^{1/2} \; d(E(\lambda)f, A^{1/2}f)\\ &= (A^{1/2}f, A^{1/2}f).\end{aligned}$$

Hence we allow all positive self-adjoint operators A on L^2 with the property that $A^{1/2} f \in L^2$ for every $f \in K^2$. For the purpose of calculating expectation values with $f,g \in K^2$, the set of allowed A can be divided into equivalence classes characterized by positive self-adjoint A such that K^2 is a core for $A^{1/2}$. The set of operators A with the latter properties is denoted by Γ. All bounded positive A belong to Γ. If $A \in \Gamma$, then $A(t) \in \Gamma$ if $t>0$. For $f,g \in K^2$, we define

$$\begin{aligned}\langle f|A|g\rangle &:= (A^{1/2}f, A^{1/2}g),\\ \langle f|A(t)|g\rangle &:= (A^{1/2} \exp[-iH_0 t] f, A^{1/2} \exp[-iH_0 t] g).\end{aligned} \tag{1.1}$$

There is a one-to-one correspondence between operators $A \in \Gamma$ and operators $\tilde{A}$ in a cone $\tilde{\Gamma}$ in the space of bounded self-adjoint operators on K^2. This fact is used to define a topology on Γ. The result is a topological space X to replace the metric space X in Definition 1.2. If a net $\{Z_\tau\}(\tau \geqslant 0)$ of operators in X tends to $A \in X$ as $\tau \to \infty$, then $\langle f|Z_\tau|g\rangle$ tends to $\langle f|A|g\rangle$ for every fixed $f,g \in K^2$.

The above framework leads to the following quantum analogs of the three parts of Devaney's definition of chaos. Instead of one $\delta>0$ as in Definition 1.2, we may allow any $N>0$.

Theorem S (sensitive dependence on initial conditions): *Given $A \in X$ and $N>0$, there is a net $\{Z_{N\tau}\} \in X(\tau \geqslant 0)$ with the property that $f,g \in K^2$ and $\epsilon>0$ determine $T>0$ such that*

$$|\langle f|A|g\rangle - \langle f|Z_{N\tau}|g\rangle| < \epsilon \quad \text{if} \quad \tau > T,$$

$$|\langle f|A(t)|f\rangle - \langle f|Z_{N\tau}(t)|f\rangle| > N||f||^2, \quad \text{if} \quad t \geqslant \tau.$$

Theorem T (topological transitivity): *Given $A,B \in X$, there is a net $\{Z_\tau\} \in X(\tau \geqslant 0)$ with the property that $f,g \in K^2$ and $\epsilon>0$ determine $T>0$ such that*

$$|\langle f|A|g\rangle - \langle f|Z_\tau|g\rangle| < \epsilon, \quad \text{if} \quad \tau > T,$$

$$|\langle f|B|g\rangle - \langle f|Z_\tau(\tau)|g\rangle| < \epsilon, \quad \text{if} \quad \tau > T.$$

Theorem P (existence of a dense set of periodic points): *Given $A \in X$, there is a net $\{Z_\tau\} \in X(\tau \geqslant 0)$ with the property that $f,g \in K^2$ and $\epsilon>0$ determine $T>0$ such that*

$$|\langle f|A|g\rangle - \langle f|Z_\tau|g\rangle| < \epsilon, \text{ if } \tau > T.$$

Moreover,

$$Z_\tau(t+n\tau) = Z_\tau(t), \text{ if } t \geqslant 0, \; n = 0,1,2, \ldots$$

In each theorem, Z_τ belongs to an ϵ-neighborhood of A when $\tau > T$. Hence, suppose an experiment is meant to determine $\langle f|A(t)|f\rangle$, but actually measures $\langle f|Z_\tau(t)|f\rangle$. The error could be due to design imperfections. If ϵ is of the order of the error in reading off experimental data, one cannot distinguish between the two expectation values at time $t=0$, but at later times the difference may be substantial. In the case of Theorem S, the error may increase to any value $N||f||^2$. Theorem T says that Z_τ in an ϵ-neighborhood V of A may develop into $Z_\tau(\tau)$ in an ϵ-neighborhood W of B. This corresponds to $F^{(n)}(V) \cap W \neq \emptyset$ in Definition 1.2. According to Theorem P, cases may occur in which $Z_\tau(t)$ is periodic with period τ. Since there is an operator Z_τ with this property in any neighborhood of any $A \in X$, periodic operators Z_τ form a dense set in X.

The nets $\{Z_\tau\}$ to which our theorems refer do not depend on $f,g \in K^2$. It is a property of the topological space X that convergence of $\{Z_\tau\}$ in X implies convergence of expectation values $\langle f|Z_\tau|g\rangle$ for fixed $f,g \in K^2$. The proofs of Theorems T and P are carried out in the language of the space X. This determines the topology of these results, although it might appear that the theorems are formulated in terms of the weak operator topology on L^2.

Sensitive dependence on initial conditions calls for a distance concept, but it is not necessary that this distance generates the topology. In Theorem S the topology of X is augmented by a distance measured in terms of the diagonal matrix elements of the identity operator. Once we adopt this augmented topology, the time evolution of operators $A(t) \in X$ is chaotic in the sense of Definitions 1.1 and 1.2.

The space of wave functions K^2 is described in Sec. II. Section III is devoted to quadratic form techniques to be used later on. The topological space X is defined in Sec. IV. That operators Z_τ exist as desired is due to properties of the time evolution on K^2 derived in Sec. V. It is shown in Sec. VI that the time evolution takes $A \in X$ into $A(t) \in X$. Theorems S, T, and P are proved in Secs. VII, VIII, and IX. The method of proof is inspired by symbolic dynamics. To indicate the relation, in Sec. X a comparison is made between the time evolution on K^2 and canonical models of chaos in classical dynamical systems.

There is an extensive literature on chaos in quantum mechanics from an entirely different point of view.[2–4] Its focus is on quantum systems with chaotic classical limits. There is strong evidence that such systems have spectral properties and other characteristics that do not occur if the underlying classical system is integrable. It is not clear, however, whether these quantum signatures of chaos can be related to the definition of chaos in classical dynamical systems.

II. THE SPACE OF WAVE FUNCTIONS

This paper is best understood in the momentum representation. Wave functions $f(\mathbf{k})$ belong to $L^2(\mathbf{R}^3)$. We write $|\mathbf{k}| = k$ and introduce two spherical polar coordinates ω. Identifying $f(\mathbf{k})$ and $f(k,\omega)$, we write L^2 instead of $L^2(\mathbf{R}^3)$. Units are chosen so that the Hamiltonian H_0 acts as multiplication by k^2.

The dilation operator D is the self-adjoint operator on L^2 which acts on C_0^∞-functions $f \in L^2$ as the differential operator,

$$D = \frac{i}{2}(\mathbf{k}\cdot\nabla_k + \nabla_k\cdot\mathbf{k}).$$

In a previous paper[5] we defined $J := \exp(-\pi D/2)$. The operator J is positive and self-adjoint but not bounded. The set of all $f \in L^2$ with the property that $Jf \in L^2$ was denoted by K^2. This set is dense in L^2 in the L^2-norm. It was shown that $f(k,\omega)\in K^2$ if and only if $f(k,\omega)$ is the mean-square boundary value of a function $f(ke^{i\phi},\omega)$ which is analytic in the sector $-\pi/2<\phi<0$ for almost every $\omega\in S^2$ and has the property that the integral,

$$\int_{S^2}\int_0^\infty |ke^{i\phi}f(ke^{i\phi},\omega)|^2\, dk\, d\omega,$$

is bounded uniformly in the sector.

It is known[6,7] that all functions $f(ke^{i\phi},\omega)$ with the above properties have mean-square boundary values $f(k,\omega)$ and $f(ke^{-i\pi/2},\omega)$. Under the inner product,

$$\langle f,g\rangle=\int_{S^2}\int_0^\infty [f(k,\omega)\overline{g}(k,\omega)+f(ke^{-i\pi/2},\omega)\overline{g}(ke^{-i\pi/2},\omega)]k^2\, dk\, d\omega, \tag{2.1}$$

the set is a Hilbert space which we denote by G^2.

It was shown before[5] that boundary values of $f\in G^2$ satisfy

$$Jf(k,\omega)=e^{-3i\pi/4}f(ke^{-i\pi/2},\omega). \tag{2.2}$$

In terms of inner products $(\cdot,\cdot)$ on L^2, we may therefore rewrite Eq. (2.1) as

$$\langle f,g\rangle=(f,g)+(Jf,Jg). \tag{2.3}$$

Since G^2 is a Hilbert space and $f\in K^2$ if and only if $f(k,\omega)$ is the boundary value at $\phi=0$ of $f\in G^2$, the set K^2 is complete under the inner product (2.3), hence a Hilbert space. We denote the Hilbert space by K^2. It is easy to verify that

$$\langle f,g\rangle=([1+J^2]^{1/2}f,[1+J^2]^{1/2}g). \tag{2.4}$$

Summarizing, inner products on L^2 are denoted by $(\cdot,\cdot)$, inner products on K^2 by $\langle\cdot,\cdot\rangle$. The expressions in Eqs. (2.3) and (2.4) are linear in f, antilinear in g. Henceforth norms on L^2 are denoted by $||\cdot||$, norms on K^2 by $|||\cdot|||$. The restriction to K^2 of an operator $T:L^2\to L^2$ is $T\restriction K^2$. The domain of T is $\mathrm{Dom}(T)$. The adjoint on L^2 of an operator $T:L^2\to L^2$ is denoted by T^*, the adjoint on K^2 of an operator $S:K^2\to K^2$ is denoted by $S^\dagger$. For example, if T is a bounded operator on L^2 and $f\in K^2$ is such that $(1+J^2)f\in L^2$, it follows from the definition of an adjoint that

$$T^\dagger f=(1+J^2)^{-1}T^*(1+J^2)f. \tag{2.5}$$

Lemma 2.1: If T is a closed operator on L^2 with $\mathrm{Dom}(T)\supseteq K^2$, *then* $(1+J^2)^{-1/2}T$ *is a bounded operator on* K^2.

Proof: Suppose $f\in K^2$. By the data and Eq. (2.4), $(1+J^2)^{-1/2}Tf$ belongs to K^2. Hence $S:=(1+J^2)^{-1/2}T$ can be viewed as an operator on K^2. If we can show that S is closed, the proposition follows from the closed-graph theorem.

To verify that S is closed, consider a Cauchy sequence $\{g_n\}(n=1,2,\dots)$ in K^2 with the property that $|||Sg_n-Sg_m|||\to 0$. Since $|||g_n-g_m|||\to 0$, an element $g\in K^2$ exists such that $|||g_n-g|||\to 0$. Since $g\in K^2$, the data imply that $Tg\in L^2$, hence $Sg\in K^2$. We have to show that $|||Sg_n-Sg|||\to 0$.

Since the L^2-norm is weaker that the K^2-norm, $||g_n-g||\to 0$. Due to Eq. (2.4), $|||Sg_n-Sg_m|||\to 0$ implies $||Tg_n-Tg_m||\to 0$. Hence $h\in L^2$ exists such that $||Tg_m-h||\to 0$. Because T is closed by assumption, it follows not only that $Tg\in L^2$, but also that $Tg=h$. This means that $||Tg_n-Tg||\to 0$. With Eq. (2.4) it follows that $|||Sg_n-Sg|||\to 0$. Hence S is closed. The closed-graph theorem completes the proof. □

III. QUADRATIC FORMS

Let A be a positive self-adjoint operator on L^2 such that Dom $(A^{1/2}) \supseteq K^2$. It follows from Lemma 2.1 that $(1+J^2)^{-1/2}A^{1/2}$ is a bounded operator on K^2.

Now consider the positive quadratic form,

$$q[f,g]:=(A^{1/2}f,A^{1/2}g), \tag{3.1}$$

on L^2 with form domain $\mathrm{Dom}(q)=K^2$. If K^2 happens to be equal to $\mathrm{Dom}(A^{1/2})$, the form q is closed. Otherwise q is closable.[8] In any case,

$$q[f,g]=\langle (1+J^2)^{-1/2}A^{1/2}f,(1+J^2)^{-1/2}A^{1/2}g\rangle=\langle [(1+J^2)^{-1/2}A^{1/2}]^{\dagger}(1+J^2)^{-1/2}A^{1/2}f,g\rangle. \tag{3.2}$$

In an obvious notation, we write

$$q[f,g]=\langle \tilde{A}f,g\rangle. \tag{3.3}$$

This defines a bounded positive operator $\tilde{A}$ on K^2. In case A is bounded on L^2, it follows with Eq. (2.5) that $\tilde{A}=(1+J^2)^{-1}A$.

In Eq. (3.1) the operator A determines the form q. We now let $q[f,g]$ be a positive quadratic form on L^2 with form domain $\mathrm{Dom}(q)=K^2$. Suppose q is closed or closable. Denote the closure by $\overline{q}$ and its form domain by $\mathrm{Dom}(\overline{q})$. The form $\overline{q}$ determines a positive self-adjoint operator A on L^2 with domain $\mathrm{Dom}(A)\subset\mathrm{Dom}(\overline{q})$ and

$$\overline{q}[f,g]=(Af,g), \tag{3.4}$$

for every $f\in\mathrm{Dom}(A)$ and $g\in\mathrm{Dom}(\overline{q})$. Moreover, $\mathrm{Dom}(A^{1/2})=\mathrm{Dom}(\overline{q})$ and

$$\overline{q}[f,g]=(A^{1/2}f,A^{1/2}g), \tag{3.5}$$

for every $f,g\in\mathrm{Dom}(\overline{q})$. A subset D' of $\mathrm{Dom}(\overline{q})$ is a core for $\overline{q}$ if and only if it is a core for $A^{1/2}$.

The above propositions follow from two representation theorems for quadratic forms.[9] The assumption that q with form domain K^2 has closure $\overline{q}$ means that K^2 is a core for $\overline{q}$. Hence we may let D' be K^2, showing that K^2 is a core for $A^{1/2}$ in Eq. (3.5).

Restricting Eq. (3.5) to K^2, we find Eq. (3.1). On the other hand, in setting up Eq. (3.1) we did not assume that K^2 was a core for $A^{1/2}$. Given the operator A that was the starting point for Eq. (3.1), it may happen that $A^{1/2}\restriction K^2$ has more than one self-adjoint extension. Among all possibilities, the representation theorems select $(A^{1/2}\restriction K^2)^*$. This is the operator $A^{1/2}$ in Eq. (3.5) whose square is A in Eq.(3.4).

Since we are looking only at expectation values of the form (3.1), with $f,g\in K^2$, all positive self-adjoint operators A with the same $A^{1/2}\restriction K^2$ form an equivalence class characterized by the particular A that satisfies the representation theorems. Hence we denote by Γ the set of all positive self-adjoint operators A on L^2 with the property that K^2 is a core for $A^{1/2}$.

The set of positive closable forms on L^2 with form domain K^2 is denoted by Γ_q. Since $q\in\Gamma_q$ determines $A\in\Gamma$ and $A\in\Gamma$ determines the bounded positive operator $\tilde{A}$ on K^2 as in Eqs. (3.2) and (3.3), each $q\in\Gamma_q$ is of the form $\langle\tilde{A}f,g\rangle$ with some $\tilde{A}$. We denote the set of operators $\tilde{A}$ by $\tilde{\Gamma}$. Given $\tilde{A}\in\tilde{\Gamma}$, we can construct $q\in\Gamma_q$ which determines $A\in\Gamma$. In this sense A is a function of $\tilde{A}$. For future reference we write $A=\gamma(\tilde{A})$. The inverse relation $\tilde{A}=\gamma^{-1}(A)$ is illustrated by Eqs. (3.2) and (3.3).

In proving Theorems S, T, and P, our strategy is to construct operators $\tilde{Z}_\tau\in\tilde{\Gamma}$ giving rise to forms $q\in\Gamma_q$. Via the representation theorems, the forms determine operators $Z_\tau\in\Gamma$ and expectation values $\langle f|Z_\tau(t)|g\rangle$ as desired.

IV. THE TOPOLOGY

Let $A,B \in \Gamma$ determine the forms $q_A, q_B \in \Gamma_q$. The form $q := q_A + q_B$ with domain Dom(q) $= K^2$ is closable.[8] Hence $q \in \Gamma_q$. The closure $\overline{q}$ determines an operator[9] $A \dotplus B \in \Gamma$ known as the form sum of A and B. For $f,g \in K^2$, it follows from Eqs. (3.1) and (3.3) that

$$\begin{aligned}(A^{1/2}f, A^{1/2}g) + (B^{1/2}f, B^{1/2}g) &= \langle \tilde{A}f,g\rangle + \langle \tilde{B}f,g\rangle \\ &= \langle [\tilde{A}+\tilde{B}]f,g\rangle \\ &= ([A \dotplus B]^{1/2}f, [A \dotplus B]^{1/2}g).\end{aligned}$$

Expectation values satisfy

$$\langle f|A|g\rangle + \langle f|B|g\rangle = \langle f|A \dotplus B|g\rangle.$$

This shows that we can define a sum on Γ. If $A \in \Gamma$ and c is a positive constant, $cA \in \Gamma$. Hence multiplication by positive constants is defined on Γ. Summarizing, Γ is a cone corresponding to a cone $\tilde{\Gamma}$ in the vector space $\tilde{B}$ consisting of bounded self-adjoint operators on K^2.

The weak operator topology on $\tilde{B}$ is a suitable starting point for defining a topology on Γ. It is generated by the family of seminorms,

$$\rho := \{f,g \in K^2, \tilde{T} \in \tilde{B}, \rho_{fg}(\tilde{T}) := |\langle \tilde{T}f,g\rangle|\}.$$

A neighborhood base at 0 is formed by the sets

$$\tilde{W}_\epsilon(K_m^2, K_n^2) := \{\tilde{T} \in \tilde{B} | \rho_{fg}(\tilde{T}) < \epsilon, f \in K_m^2, g \in K_n^2\},$$

for all $\epsilon > 0$ and any finite sets $K_m^2 \in K^2, K_n^2 \in K^2$.

Let us denote the topology by w and let $\tilde{W}$ be a typical open set. The pair $(\tilde{B},w)$ is a locally convex space. It induces on $\tilde{\Gamma}$ the relative topology[10] v with open sets $\tilde{W} \cap \tilde{\Gamma}$,

$$v := \{\tilde{W} \cap \tilde{\Gamma} | \tilde{W} \in w\}.$$

We denote the topological space $(\tilde{\Gamma}, v)$ by $\tilde{X}$. A net $\{\tilde{Z}_\tau\} \in \tilde{X} (\tau \geqslant 0)$ tends to $\tilde{A} \in \tilde{X}$ as $\tau \to \infty$ if and only if $\langle \tilde{Z}_\tau f,g\rangle$ tends to $\langle \tilde{A}f,g\rangle$ for every fixed $f,g \in K^2$. By Eqs. (3.1) and (3.3),

$$\lim_{\tau \to \infty} (Z_\tau^{1/2}f, Z_\tau^{1/2}g) = (A^{1/2}f, A^{1/2}g). \tag{4.1}$$

Conversely, if Eq. (4.1) is true for every $f,g \in K^2$, then $\tilde{Z}_\tau$ tends to $\tilde{A}$ in $\tilde{X}$. Hence we want Eq. (4.1) to express what it means for a net $\{Z_\tau\}$ to converge to A. This convergence concept makes the function $\gamma: \tilde{\Gamma} \to \Gamma$ introduced at the end of Sec. III continuous with a continuous inverse. Since γ maps $\tilde{\Gamma}$ onto Γ, the only topology on Γ to make γ and γ^{-1} continuous is the quotient topology[10] defined by

$$v_\gamma := \{W \subseteq \Gamma | \gamma^{-1}(W) \in v\}.$$

The topological space (Γ, v_γ) is denoted by X. This is our space of observables A.

As a map between topological spaces, the function $\gamma: \tilde{X} \to X$ is continuous and maps open sets onto open sets. In addition γ is one-to-one and onto. Hence γ is a homeomorphism. A net $\{Z_\tau\} \in X$ tends to $A \in X$ if and only if $\{\tilde{Z}_\tau\} \in \tilde{X}$ tends to $\tilde{A} \in \tilde{X}$. By Equation (1.1), this condition is necessary and sufficient in order that expectation values $\langle f|Z_\tau|g\rangle$ tend to $\langle f|A|g\rangle$.

V. THE TIME EVOLUTION

In order that we can use operators on K^2 as intermediate steps in proving Theorems S, T, and P, we need to know how the time evolution $\exp(-iH_0t)$ acts on K^2. Since H_0 acts as multiplication by k^2, the operator $\exp(-iH_0t)$ acts on L^2 as multiplication by $\exp(-ik^2t)$. If $-\pi/2<\phi<0$, then $\exp(-ik^2e^{2i\phi}t)$ is bounded if and only if $t\geqslant 0$. Hence, if $f(ke^{i\phi},\omega)\in G^2$, then $\exp(-ik^2e^{2i\phi}t)f(ke^{i\phi},\omega)\in G^2$ if $t\geqslant 0$, but not necessarily if $t<0$. By the same token, if $f(k,\omega)\in K^2$, then $\exp(-ik^2t)f(k,\omega)\in K^2$ if $t\geqslant 0$ but not necessarily if $t<0$. Hence $t\geqslant 0$ in all of the following.

If $f\in K^2$ and $t\geqslant 0$, it follows from Eq. (2.2) that

$$J \exp\,(-ik^2t)f(k,\omega)=\exp(ik^2t)Jf(k,\omega). \tag{5.1}$$

For $t\geqslant 0$ only, we define the operator $U(t)$ on L^2 by

$$U(t)f(k,\omega):=\exp(-ik^2t)f(k,\omega). \tag{5.2}$$

This operator is unitary. Its adjoint $U^*(t)$ acts as multiplication by $\exp(ik^2t)$. Due to Eqs. (2.3) and (5.1),

$$\langle U(t)f,U(t)g\rangle=\langle f,g\rangle. \tag{5.3}$$

Hence $U(t)$ is an isometry on K^2. It follows that $U^\dagger(t)U(t)$ acts on K^2 as the identity. By Eq. (2.5), $U^\dagger(t)$ is not the same as $U^*(t)$. Since $U(t)$ is an isometry on K^2, the set $U(t)K^2$ is a subspace of K^2 which is closed in the K^2-norm. Let the subspace be denoted by $K^2(t)$ and let $Q(t)$ be the orthogonal projection of K^2 onto $K^2(t)$. By general properties of isometries,[11]

$$U(t)U^\dagger(t)=Q(t). \tag{5.4}$$

To see how $Q(t)$ acts on K^2, we first consider the space G^2 and use the fact that G^2 is isometrically isomorphic to the Hardy space H^2 of the lower half-plane.[6] For $f\in G^2$, we define

$$v+iw:=k^2e^{2i\phi},\quad F(v+iw,\omega):=(ke^{i\phi}/2)^{1/2}f(ke^{i\phi},\omega).$$

When $ke^{i\phi}$ varies in the sector $-\pi/2<\phi<0$, then $v+iw$ runs through the lower half-plane. If $f,g\in G^2$, the inner product $\langle f,g\rangle$ satisfies

$$\langle f,g\rangle=\int_{S^2}\int_{-\infty}^{\infty}F(v-i0,\omega)\overline{G}(v-i0,\omega)\;dv\;d\omega.$$

Now define the Fourier transform,

$$\hat{f}(s,\omega)=(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(ivs)F(v-i0,\omega)\;dv.$$

By the Paley–Wiener theorem,[12] $\hat{f}(s,\omega)=0$ if $s<0$, for almost every ω. The Fourier transform is a unitary map taking $f\in G^2$ into $\hat{f}(s,\omega)\in L^2(\mathbf{R}^+\times S^2)$.

If $f\in K^2$, then $U(t)f(k,\omega)\in K^2$ is the boundary value at $\phi=0$ of $\exp(-ik^2e^{2i\phi}t)f(ke^{i\phi},\omega)$. The latter function corresponds to $\exp(-ivt+wt)F(v+iw,\omega)$ with Fourier transform

$$(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(-ivt+ivs)F(v-i0,\omega)\;dv=\hat{f}(s-t,\omega). \tag{5.5}$$

For fixed $t\geqslant 0$, this Fourier transform vanishes if $s<t$.

We now return to the space K^2. Since K^2 and G^2 are isometrically isomorphic, the Fourier transform on G^2 gives rise to a unitary map taking $f(k,\omega)\in K^2$ into $\hat{f}(s,\omega)\in L^2(\mathbf{R}^+\times S^2)$. With Eq. (5.5) this point of view yields

$$\begin{aligned}\langle U(t)f,g\rangle &= \int_{S^2}\int_t^\infty \hat{f}(s-t,\omega)\overline{\hat{g}}(s,\omega)\,ds\,d\omega\\ &=\langle f,U^\dagger(t)g\rangle\\ &=\int_{S^2}\int_0^\infty \hat{f}(s,\omega)\overline{\hat{g}}(s+t,\omega)\,ds\,d\omega.\end{aligned} \tag{5.6}$$

Replacing f by $U^\dagger(t)f$ gives

$$\langle Q(t)f,g\rangle=\int_{S^2}\int_t^\infty \hat{f}(s,\omega)\overline{\hat{g}}(s,\omega)\,ds\,d\omega.$$

Since $Q(t)$ is an orthogonal projection,

$$|||Q(t)f|||^2=\int_{S^2}\int_t^\infty |\hat{f}(s,\omega)|^2\,ds\,d\omega. \tag{5.7}$$

This quantity tends to 0 as $t\to\infty$.

Since $U(t)$ takes K^2 into K^2 and $U(s)U(t)$ equals $U(s+t)$, the family $\{U(t)|0\leqslant t<\infty\}$ is a semigroup of operators on K^2. If $0\leqslant t<s$, it follows from the semigroup property and Eq. (5.4) that

$$\begin{aligned}Q(s)U(t)&=U(s)U^\dagger(s)U(t)\\ &=U(t)U(s-t)U^\dagger(s-t)U^\dagger(t)U(t)\\ &=U(t)U(s-t)U^\dagger(s-t)\\ &=U(t)Q(s-t),\end{aligned}$$

$$\begin{aligned}Q(s)Q(t)&=U(s)U^\dagger(s)U(t)U^\dagger(t)\\ &=U(s)U^\dagger(s-t)U^\dagger(t)U(t)U^\dagger(t)\\ &=U(s)U^\dagger(s-t)U^\dagger(t)=U(s)U^\dagger(s)=Q(s).\end{aligned}$$

With similar calculations for $0\leqslant s\leqslant t$, we find

$$\begin{aligned}Q(s)U(t)&=U(t),\ \text{if }0\leqslant s\leqslant t,\\ &=U(t)Q(s-t),\ \text{if }0\leqslant t<s,\end{aligned} \tag{5.8}$$

$$Q(s)Q(t)=Q(\max s,t). \tag{5.9}$$

VI. THE HEISENBERG PICTURE

An operator $A\in\Gamma$ gives rise to a form q as in Eqs. (3.1) and (3.3). Replacing $f,g\in K^2$ by $U(t)f,U(t)g\in K^2$ gives

$$q_t[f,g]:=\langle U^\dagger(t)\tilde{A}U(t)f,g\rangle=(A^{1/2}U(t)f,A^{1/2}U(t)g), \tag{6.1}$$

with form domain K^2. According to Lemma 6.1 below, q_t is closable, hence determines $A_t \in \Gamma$. On the other hand, we can apply $U^*(t)AU(t)$ to $U^*(t)\mathrm{Dom}(A)$. The operator $A(t)$ so defined is symmetric. By Lemma 6.2 it is self-adjoint and equal to A_t. Hence $A(t) \in \Gamma$ if $A \in \Gamma$. This means that the Heisenberg picture can be used without ambiguity. Given a form $q \in \Gamma_q$, we can extract $A \in \Gamma$, then construct $A(t) \in \Gamma$. Or we can let q develop into q_t, then determine $A_t \in \Gamma$. The result is the same. Similarly, $A(t)$ and A_t act in the same way as elements of the topological space X.

Lemma 6.1: Let T be a symmetric operator on L^2 with domain K^2. Let $U(t)$ be defined by Eq. (5.2). Let q and q_t be the forms defined by

$$q[f,g]:=(Tf,Tg),$$

$$q_t[f,g]:=(TU(t)f,TU(t)g),$$

each with form domain K^2.

The forms are closable. The form domains of the closures satisfy

$$\mathrm{Dom}(\overline{q})=U(t)\mathrm{Dom}(\overline{q_t}),\ \ \mathrm{Dom}(\overline{q_t})=U^*(t)\mathrm{Dom}(\overline{q}). \tag{6.2}$$

Let A and A_t be the positive self-adjoint operators defined by q and q_t, respectively. Then

$$A_t^{1/2}=U^*(t)A^{1/2}U(t), \tag{6.3}$$

where each side is self-adjoint with domain $\mathrm{Dom}(\overline{q_t})$.

Proof: Since T is symmetric, T is closable. Hence q is closable. Since $U(t)K^2 \subseteq K^2$, the operator $U^*(t)TU(t)\restriction K^2$ is symmetric, hence closable. This is sufficient for q_t to be closable.

By the representation theorems,[9] A and A_t are well-defined. The domains of $A^{1/2}$ and $A_t^{1/2}$ are equal to $\mathrm{Dom}(\overline{q})$ and $\mathrm{Dom}(\overline{q_t})$, respectively. The common domain $\mathrm{Dom}(q)=\mathrm{Dom}(q_t)=K^2$ is a core for both $A^{1/2}$ and $A_t^{1/2}$.

Choose $f \in \mathrm{Dom}(\overline{q_t})$. Since K^2 is a core for $A_t^{1/2}$, there exists a sequence $\{f_n\} \in K^2 (n=1,2,\dots)$ such that $\|f_n - f\| \to 0$ and $\|A_t^{1/2}(f_n-f_m)\| \to 0$ as $n,m \to \infty$. Since $U(t)f_n \in K^2$, the sequence $\{f_n\}$ has the property that $\|U(t)f_n - U(t)f\| \to 0$ and

$$\begin{aligned}\|A_t^{1/2}(f_n-f_m)\| &= \|TU(t)(f_n-f_m)\| \\ &= \|T[U(t)f_n - U(t)f_m]\| \\ &= \|A^{1/2}[U(t)f_n - U(t)f_m]\| \to 0.\end{aligned}$$

Since $A^{1/2}$ is closed, it follows that $U(t)f \in \mathrm{Dom}(A^{1/2})$. Hence

$$U(t)\mathrm{Dom}(\overline{q_t}) \subseteq \mathrm{Dom}(\overline{q}). \tag{6.4}$$

That K^2 is a core for $A_t^{1/2}$ also implies that

$$A_t^{1/2}=[U^*(t)TU(t)\restriction K^2]^*. \tag{6.5}$$

Likewise $A^{1/2}=T^*$. Now choose $g \in U^*(t)\mathrm{Dom}(\overline{q})$. Clearly $g=U^*(t)h$ with $h=U(t)g \in \mathrm{Dom}(\overline{q})$. If $f \in K^2$,

$$\begin{aligned}(U^*(t)TU(t)f,g) &= (TU(t)f,h) \\ &= (f,U^*(t)T^*h) \\ &= (f,U^*(t)A^{1/2}h) \\ &= (f,U^*(t)A^{1/2}U(t)g).\end{aligned} \tag{6.6}$$

Due to Eq. (6.5), this shows that $g \in \mathrm{Dom}(A_t^{1/2})$, hence

$$U^*(t)\mathrm{Dom}(\overline{q}) \subseteq \mathrm{Dom}(\overline{q_t}). \tag{6.7}$$

The relations (6.4) and (6.7) give the desired equality (6.2). Hence g in Eq. (6.6) may be any element of $\mathrm{Dom}(\overline{q_t})$. With Eq. (6.5), it follows that $A_t^{1/2}$ acts on $\mathrm{Dom}(\overline{q_t})$ in the same way as $U^*(t)A^{1/2}U(t)$, confirming Eq. (6.3). It is known from the representation theorems that $A_t^{1/2}$ with domain $\mathrm{Dom}(\overline{q_t})$ is self-adjoint.

It remains to show that $U^*(t)A^{1/2}U(t)$ with domain $U^*(t)\mathrm{Dom}(\overline{q})$ is self-adjoint. It is easy to see that the operator is symmetric. To prove that it is self-adjoint, it is sufficient to show that the operators $U^*(t)A^{1/2}U(t) \pm i$ have range L^2. Hence we let f run through $\mathrm{Dom}(\overline{q})$ and examine

$$U^*(t)(A^{1/2} \pm i)U(t)U^*(t)f = U^*(t)(A^{1/2} \pm i)f.$$

Since $A^{1/2}$ with domain $\mathrm{Dom}(\overline{q})$ is self-adjoint, $(A^{1/2} \pm i)f$ runs through L^2. This is sufficient for $U^*(t)(A^{1/2} \pm i)f$ to run through L^2, as we wanted to show. □

Lemma 6.2: Let the data be as in Lemma 6.1. The operator

$$A(t) := U^*(t)AU(t)$$

with domain $U^*(t)\mathrm{Dom}(A)$ *is self-adjoint and equal to* A_t.

Proof: The reasoning applied to $U^*(t)A^{1/2}U(t)$ at the end of the proof of Lemma 6.1, can be repeated to show that $A(t)$ with domain $U^*(t)\mathrm{Dom}(A)$ is self-adjoint. The operator on the right-hand side of Eq. (6.3) is the square root $[A(t)]^{1/2}$. Since it is equal to $A_t^{1/2}$, it follows that $A(t) = A_t$. □

Corollary 6.3: If $A \in X$ *and* $t \geqslant 0$, *then* $A(t) \in X$. □

For future reference we define

$$\tilde{A}(t) := U^\dagger(t)\tilde{A}U(t).$$

By Eq. (6.1), this is the operator in $\tilde{X}$ determined by $A(t)$ in X.

Since $\tilde{A}(t)$ is a bounded operator on K^2 whenever $t \geqslant 0$, this quantity is easy to work with. In the notation of Lemma 6.1, the domain of $A(t)$ is a subset of $U^*(t)\mathrm{Dom}(\overline{q})$. We do not know whether the intersection $\cap_{t\geqslant 0}\mathrm{Dom}[A(t)]$ is sufficiently large to be useful. On the other hand, the operator $[A(t)]^{1/2}$ can be applied to every $f \in K^2$ at all times $t \geqslant 0$. This is the reason why it is convenient to express expectation values of observables in the form (1.1).

VII. SENSITIVE DEPENDENCE ON INITIAL CONDITIONS

Proof of Theorem S: The operator $A \in X$ determines the positive operator $\tilde{A}$ on K^2 as in Eqs. (3.2) and (3.3). For $N > 0$ and $\tau \geqslant 0$, define

$$\tilde{Z}_{N\tau} := \tilde{A} + (1+\tau)^{-1} + NQ(\tau). \tag{7.1}$$

This is a positive operator on K^2. Since $\tilde{A}$ is bounded by Lemma 2.1 and $Q(\tau)$ is a projection operator, $\tilde{Z}_{N\tau}$ is bounded.

We claim that the form

$$q[f,g] := \langle \tilde{Z}_{N\tau}f, g\rangle$$

with form domain K^2 is closed for every fixed N, τ. To prove this, we choose a sequence $\{f_n\} \in K^2 (n=1,2,\ldots)$ such that $\|f_n - f_m\| \to 0$ and

$$\langle \tilde{Z}_{N\tau}(f_n - f_m), f_n - f_m \rangle \to 0, \tag{7.2}$$

as $n,m \to \infty$. It follows that $f \in L^2$ exists such that $\|f_n - f\| \to 0$. We have to show that $f \in K^2$ and $\langle \tilde{Z}_{N\tau}(f_n - f), f_n - f \rangle \to 0$.

By Eq. (7.1),

$$\langle \tilde{Z}_{N\tau}(f_n - f_m), f_n - f_m \rangle \geq (1+\tau)^{-1} |||f_n - f_m|||^2. \tag{7.3}$$

Hence $|||f_n - f_m||| \to 0$. Since K^2 is a Hilbert space, it follows that $g \in K^2$ exists such that $|||f_n - g||| \to 0$. From this it follows that $\|f_n - g\| \to 0$. Hence $f = g$ and $f \in K^2$. Moreover,

$$|\langle \tilde{Z}_{N\tau}(f_n - f), f_n - f \rangle| \leq [|||\tilde{A}||| + (1+\tau)^{-1} + N] |||f_n - f|||^2 \to 0. \tag{7.4}$$

This is sufficient for the form $q[f,g]$ to be closed.

By the representation theorems[9] and Lemmas 6.1, 6.2, the form $q[f,g]$ determines operators $Z_{N\tau} \in X$ and $Z_{N\tau}(t) \in X$. The operator $[Z_{N\tau}(t)]^{1/2}$ has domain $U^*(t)K^2$ $(t \geq 0)$.

Given any fixed $f \in K^2$, it is clear that $(1+\tau)^{-1}|||f|||$ tends to 0 as $\tau \to \infty$. By Equation (5.7) $|||Q(\tau)f||| \to 0$. Since N is fixed, it follows that $|||(\tilde{Z}_{N\tau} - \tilde{A})f||| \to 0$ as $\tau \to \infty$. Hence $\tilde{Z}_{N\tau} \to \tilde{A}$ in $\tilde{X}$ and $Z_{N\tau} \to A$ in X. Given $f,g \in K^2$ and $\epsilon > 0$, we can find T such that

$$|\langle \tilde{Z}_{N\tau} f, g \rangle - \langle \tilde{A} f, g \rangle| < \epsilon, \quad \text{if } \tau > T. \tag{7.5}$$

Via Eqs. (1.1) and (3.1), (3.3), this proves that expectation values at time $t = 0$ agree with Theorem S.

At time $t \geq \tau$ we have to examine

$$\langle \tilde{Z}_{N\tau}(t) f, f \rangle - \langle \tilde{A}(t) f, f \rangle = (1+\tau)^{-1} \langle f, f \rangle + N \langle U^\dagger(t) Q(\tau) U(t) f, f \rangle.$$

By Eqs. (5.3) and (5.8),

$$U^\dagger(t) Q(\tau) U(t) = U^\dagger(t) U(t) = I,$$

where I is the identity operator. Hence

$$|\langle f|A(t)|f\rangle - \langle f|Z_{N\tau}(t)|f\rangle| \geq N|||f|||^2 \geq N\|f\|^2,$$

as desired. □

Chaos in classical dynamical systems can often be characterized in terms of Lyapunov exponents related to the rate at which the distance between nearby orbits increases.[4,13–17] In an experiment in which we want to measure $\langle f|A(t)|f\rangle$ but actually observe $\langle f|Z_{N\tau}(t)|f\rangle$, the operator $(1+\tau)^{-1} + NQ(\tau)$ on K^2 represents the error. Let us examine the rate at which its expectation value increases. If ϵ in Eq. (7.5) is small and $\tau > T$, then τ is large. This makes $(1+\tau)^{-1}\langle f, f \rangle$ small. Hence we focus on the error

$$E_\tau(t) := N \langle U^\dagger(t) Q(\tau) U(t) f, f \rangle.$$

As long as $0 \leq t \leq \tau$, it follows from Eqs. (5.7) and (5.8) that

$$E_\tau(t) = N \langle U^\dagger(t) U(t) Q(\tau - t) f, f \rangle = N \langle Q(\tau - t) f, f \rangle = N \int_{S^2} \int_{\tau - t}^{\infty} |\hat{f}(s,\omega)|^2 \, ds \, d\omega.$$

This is a nondecreasing function of t which reaches its maximum $N|||f|||^2$ at time $t = \tau$. By Eq. (5.8), $E_\tau(t)$ is constant and equal to $N|||f|||^2$ if $t \geq \tau$.

When $0 \leq t \leq \tau$, the relative rate of increase at time t is

$$R_\tau(t)=[E_\tau(t)]^{-1}\ dE_\tau(t)/dt=|||Q(\tau-t)f|||^{-2}\int_{S^2}|\hat{f}(\tau-t,\omega)|^2\ d\omega.$$

Averaging over the time period $0\leqslant t\leqslant\tau$ yields

$$R_\tau=\tau^{-1}\int_0^\tau R_\tau(t)\ dt=\tau^{-1}\ln\,[|||f|||^2\ |||Q(\tau)f|||^{-2}].$$

Take $\tau>T$ as in Eq. (7.5). As an example, now suppose that

$$\int_{S^2}|\hat{f}(s,\omega)|^2\ d\omega=C\ \exp(-\lambda^p s^p) \tag{7.6}$$

when $s>T$, with some positive constants C,λ, and p. This gives

$$R_\tau=\tau^{-1}\ \ln\ |||f|||^2-\tau^{-1}\ \ln\int_\tau^\infty C\ \exp(-\lambda^p s^p)\ ds.$$

To get an estimate of R_τ at large τ, we evaluate $\lim_{\tau\to\infty}R_\tau$ by applying L'Hospital's rule twice. The result is

$$\lim_{\tau\to\infty}R_\tau=\lim_{\tau\to\infty}p\lambda^p\tau^{p-1}=0,\ \text{if}\ \ 0<p<1;=\lambda,\ \text{if}\ \ p=1;=\infty,\ \text{if}\ \ p>1.$$

If $0<p<1$ and the equality sign in Eq. (7.6) is replaced by $\geqslant$, it remains true that R_τ tends to 0. Similarly, if $p>1$ and the equality sign is replaced by $\leqslant 1$, it remains true that R_τ tends to ∞.

A more complicated example is

$$\int_{S^2}|\hat{f}(s,\omega)|^2\ d\omega=\exp(-\lambda s)\ \cos^2\ \mu s.$$

In this case $\lim_{\tau\to\infty}R_\tau$ does not exist. On the other hand, since

$$|||Q(\tau)f|||^2\leqslant\int_\tau^\infty\exp(-\lambda s)\ ds=\lambda^{-1}\exp(-\lambda\tau),$$

we find that

$$R_\tau\geqslant\tau^{-1}\ \ln\ |||f|||^2+\tau^{-1}\ \ln\ \lambda+\lambda.$$

Given any $\delta>0$, it follows that Θ exists such that $R_\tau>\lambda-\delta$ if $\tau>\Theta$.

VIII. TOPOLOGICAL TRANSITIVITY

Proof of Theorem T: Let $\tilde{A}$ and $\tilde{B}$ be the operators in $\tilde{X}$ determined by A and B in X. Define

$$\tilde{Z}_\tau:=[1-Q(\tau)]\tilde{A}[1-Q(\tau)]+U(\tau)\tilde{B}U^\dagger(\tau)+(1+\tau)^{-1}. \tag{8.1}$$

This is a bounded operator on K^2 satisfying

$$|||\tilde{Z}_\tau|||\leqslant|||\tilde{A}|||+|||\tilde{B}|||+(1+\tau)^{-1}. \tag{8.2}$$

The form $\langle\tilde{Z}_\tau f,g\rangle$ with form domain K^2 is closed. To show this, we choose a sequence $\{f_n\}\in K^2(n=1,2,\dots)$ such that $||f_n-f_m||\to 0$ and the analog of Eq. (7.2) is satisfied. As with Eq. (7.3), it follows that $f\in K^2$ exists such that $|||f_n-f|||\to 0$. In fact, the term $(1+\tau)^{-1}$ is included in $\tilde{Z}_\tau$ to make sure the form is closed or closable, even in case $\tilde{A}$ and $\tilde{B}$ have nonempty nullspaces. Due to Eq. (8.2), the analog of Eq. (7.4) is true. Hence the form is closed.

At time $t=0$, we examine

$$|||(\tilde{Z}_\tau-\tilde{A})f|||=|||[-Q(\tau)\tilde{A}-\tilde{A}Q(\tau)+Q(\tau)\tilde{A}Q(\tau)+U(\tau)\tilde{B}U^\dagger(\tau)+(1+\tau)^{-1}]f|||$$
$$\leqslant|||Q(\tau)\tilde{A}f|||+2|||\tilde{A}|||\ |||Q(\tau)f|||+|||\tilde{B}|||\ |||U^\dagger(\tau)f|||+(1+\tau)^{-1}|||f|||.$$

Since $\tilde{A}f$ is fixed, $|||Q(\tau)\tilde{A}f|||\to 0$ as $\tau\to\infty$. Similarly, $|||Q(\tau)f|||\to 0$. Furthermore,

$$|||U^\dagger(\tau)f|||^2=\langle U(\tau)U^\dagger(\tau)f,f\rangle=\langle Q(\tau)f,f\rangle=|||Q(\tau)f|||^2\to 0.$$

Hence $\tilde{Z}_\tau\to\tilde{A}$ in $\tilde{X}$.

At time $t=\tau$, we have

$$\tilde{Z}_\tau(\tau)=\tilde{B}+(1+\tau)^{-1}.$$

This is true because $[1-Q(\tau)]U(\tau)=0$ by Eq. (5.8) and $U^\dagger(\tau)U(\tau)=I$ by Eq. (5.3). It follows that $\tilde{Z}_\tau(\tau)\to\tilde{B}$ in $\tilde{X}$. To be specific, if $f,g\in K^2$ and T is so large that $(1+T)^{-1}|\langle f,g\rangle|<\epsilon$, then

$$|\langle\tilde{Z}_\tau(\tau)f,g\rangle-\langle\tilde{B}f,g\rangle|<\epsilon,\ \text{if}\quad \tau>T.$$

By choosing T sufficiently large, we can also arrange for Eq. (7.5) to be true (with $\tilde{Z}_\tau$ instead of $\tilde{Z}_{N\tau}$). That expectation values agree with Theorem T can now be proved as in Sec. VII. □

Recall Definitions 1.1 and 1.2. In classical dynamical systems, topological transitivity means that the space X does not contain disjoint sets V and W that are each invariant under the map F. As a corollary, if a continuous function $\psi:X\to\mathbf{R}$ is invariant under F, it has to be constant on X. In Hamiltonian systems the implication is that conserved quantities must be constant throughout the phase space.

In the present context there are operators representing conserved quantities, but there are no nontrivial invariant sets. For example, $A\in X$ may be the projection onto angular momentum l_A. In that case $\text{Dom}(A)=L^2$ and $A(t)=A$. The proof of Theorem T shows that any neighborhood of A contains operators Z_τ with the property that $Z_\tau(t)\neq Z_\tau$ for some or all $t>0$. Hence, although $A(t)$ is invariant, typical neighborhoods of $A(t)$ are not. The only exception is X interpreted as a neighborhood of $A(t)$. By Corollary 6.3, the time evolution takes X into X.

Suppose we want to measure $\langle f|A(t)|f\rangle$ in the above example, but inadvertedly measure $\langle f|Z_\tau(t)|f\rangle$ as in Eq. (8.1), with B projecting onto angular momentum $l_B\neq l_A$. The error causes measurements taken at different times to have entirely different outcomes, although the objective is to determine a conserved quantity.

IX. EXISTENCE OF A DENSE SET OF PERIODIC POINTS

Proof of Theorem P: For any $\tau>0$, define $Q(0):=I$ and

$$\Delta Q(n):=Q(n\tau)-Q([n+1]\tau)\ (n=0,1,2,\dots). \tag{9.1}$$

By Eq. (5.9),

$$\Delta Q(n)\Delta Q(m)=\delta_{mn}\Delta Q(n).$$

The operators $\Delta Q(n)$ are projections on K^2 with mutually orthogonal ranges.

Let $\tilde{A}\in\tilde{X}$ be the operator determined by $A\in X$ and define

$$\tilde{Z}_\tau := \sum_{n=0}^{\infty} \Delta Q(n) U(n\tau)[\tilde{A} + (1+\tau)^{-1}] U^\dagger(n\tau) \Delta Q(n). \tag{9.2}$$

The sum converges in the strong operator topology on K^2. To show this, we introduce the abbreviation

$$\tilde{A}_n := U(n\tau)[\tilde{A} + (1+\tau)^{-1}] U^\dagger(n\tau).$$

This gives

$$\begin{aligned} \left|\left|\left| \sum_{n=N}^{M} \Delta Q(n) \tilde{A}_n \Delta Q(n) f \right|\right|\right|^2 &= \sum_{n=N}^{M} |||\Delta Q(n) \tilde{A}_n \Delta Q(n) f|||^2 \\ &\leqslant \sum_{n=N}^{M} |||\tilde{A}_n \Delta Q(n) f|||^2 \\ &\leqslant [|||\tilde{A}||| + (1+\tau)^{-1}]^2 \sum_{n=N}^{M} |||\Delta Q(n) f|||^2 \\ &\leqslant [|||\tilde{A}||| + (1+\tau)^{-1}]^2 \, |||Q(N\tau) f|||^2. \end{aligned} \tag{9.3}$$

Since $|||Q(N\tau)f|||$ tends to 0 as $N \to \infty$, the sum for $\tilde{Z}_\tau$ converges. Moreover,

$$|||\tilde{Z}_\tau||| \leqslant |||\tilde{A}||| + (1+\tau)^{-1}. \tag{9.4}$$

By Eq. (5.9),

$$\Delta Q(n) U(n\tau) U^\dagger(n\tau) \Delta Q(n) = \Delta Q(n) Q(n\tau) \Delta Q(n) = \Delta Q(n).$$

It follows that the terms with $(1+\tau)^{-1}$ in Eq. (9.2) add up to $(1+\tau)^{-1}$. Since each term with $\tilde{A}$ is non-negative,

$$\langle \tilde{Z}_\tau f, f \rangle \geqslant (1+\tau)^{-1} \, |||f|||^2. \tag{9.5}$$

Using Eqs. (9.4) and (9.5), we can repeat the reasoning in Secs. VII and VIII to show that the form $\langle \tilde{Z}_\tau f, g \rangle$ with form domain K^2 is closed.

By Eq. (5.8),

$$\Delta Q(n) U(n\tau) = U(n\tau) \Delta Q(0).$$

It follows that

$$U^\dagger(\tau) \tilde{Z}_\tau U(\tau) = \sum_{n=0}^{\infty} U^\dagger(\tau) U(n\tau) \Delta Q(0) [\tilde{A} + (1+\tau)^{-1}] \Delta Q(0) U^\dagger(n\tau) U(\tau).$$

The term with $n=0$ actually vanishes. This is due to

$$\Delta Q(0) U^\dagger(0) U(\tau) = \Delta Q(0) U(\tau) = [1 - Q(\tau)] U(\tau) = 0,$$

where we have used Eqs. (5.8) and (9.1). In the terms with $n = 1, 2, \ldots,$ we have

$$U^\dagger(n\tau) U(\tau) = U^\dagger([n-1]\tau) U^\dagger(\tau) U(\tau) = U^\dagger([n-1]\tau).$$

Hence

$$U^{\dagger}(\tau)\tilde{Z}_{\tau}U(\tau)=\tilde{Z}_{\tau}.$$

By iteration and the semigroup property of $\{U(t)\}$,

$$\tilde{Z}_{\tau}(t+n\tau)=U^{\dagger}(t)U^{\dagger}(n\tau)\tilde{Z}_{\tau}U(n\tau)U(t)=\tilde{Z}_{\tau}(t)\ (n=0,1,2,\dots).$$

We now show that $\tilde{Z}_{\tau}\to\tilde{A}$ in $\tilde{X}$ when $\tau\to\infty$. To this end, we denote $\Sigma_{n=1}^{\infty}\ \dots$ in Eq. (9.2) by $\Delta\tilde{Z}_{\tau}$. By the reasoning in Eq. (9.3),

$$|||\Delta\tilde{Z}_{\tau}|||\leqslant|||\tilde{A}|||+(1+\tau)^{-1}.$$

By Eqs. (5.9) and (9.1),

$$\Delta Q(n)Q(\tau)=\Delta Q(n)\ (n=1,2,\dots),$$

hence

$$\Delta\tilde{Z}_{\tau}=Q(\tau)(\Delta\tilde{Z}_{\tau})Q(\tau).$$

It follows that

$$\begin{aligned}|||(\tilde{Z}_{\tau}-\tilde{A})f|||&=|||\{[1-Q(\tau)][\tilde{A}+(1+\tau)^{-1}][1-Q(\tau)]+Q(\tau)(\Delta\tilde{Z}_{\tau})Q(\tau)-\tilde{A}\}f|||\\&=|||\{-Q(\tau)\tilde{A}-\tilde{A}Q(\tau)+Q(\tau)(\tilde{A}+\Delta\tilde{Z}_{\tau})Q(\tau)+(1+\tau^{-1})[1-Q(\tau)]\}f|||\\&\leqslant|||Q(\tau)\tilde{A}f|||+[3|||\tilde{A}|||+(1+\tau)^{-1}]|||Q(\tau)f|||+(1+\tau)^{-1}\ |||f|||.\end{aligned}$$

This tends to 0 as $\tau\to\infty$.

Given $f,g\in K^2$ and $\epsilon>0$, we first find T such that

$$|\langle\tilde{Z}_{\tau}f,g\rangle-\langle\tilde{A}f,g\rangle|<\epsilon,\quad\text{if}\quad\tau>T.$$

As in Secs. VII and VIII, this guarantees that the expectation values of A and Z_{τ} are close if $\tau>T$. Next we choose τ. This results in a periodic operator $\tilde{Z}_{\tau}(t)$ with period τ. By Lemmas 6.1 and 6.2, the quadratic form,

$$q_t[f,g]:=\langle U^{\dagger}(t)\tilde{Z}_{\tau}U(t)f,g\rangle=(Z_{\tau}^{1/2}U(t)f,Z_{\tau}^{1/2}U(t)g),$$

uniquely determines $Z_{\tau}(t)$. Notice that the form is equal to $\langle\tilde{Z}_{\tau}(t)f,g\rangle$. By the periodicity of $\tilde{Z}_{\tau}(t)$, the operators $Z_{\tau}(t)$ determined at times t and $t+n\tau$ are equal. Hence $Z_{\tau}(t)$ has the same periodicity as $\tilde{Z}_{\tau}(t)$. □

Since there is an operator Z_{τ} with periodic time evolution in any ϵ-neighborhood of any $A\in X$, the set of periodic operators is dense in X.

X. CONCLUDING REMARKS

The set of operators A to which this paper applies contains all bounded positive operators on L^2. In particular, all orthogonal projections are included. It is assumed that wave functions belong to the space K^2 which is dense in L^2.

The important point about K^2 is that the time evolution $\{U(t)\ |0\leqslant t<\infty\}$ on K^2 is unitarily equivalent to a semigroup of shift operators, as shown by Eq. (5.6). This is suggestive of Bernoulli systems and other K-maps in which the dynamics can be represented by a shift on sequences of symbols.[1,13,15–18] Our construction of operators $\tilde{Z}_{\tau}$ is based on ideas from the symbolic dynamics[1,13,15–18] used in choas proofs for K-maps. A canonical model for K-flows is due to Sinai.[19] On the orthogonal complement of the constant functions, a K-flow can be represented in terms of Hilbert spaces N and $L^2(\mathbf{R})\otimes N$, and a unitary group $\{V(t)|-\infty<t<\infty\}$ such that $\hat{g}(s)\in L^2(\mathbf{R})\otimes N$ transforms according to

$$(V(t)\hat{g})(s)=\hat{g}(s-t). \tag{10.1}$$

The same canonical form occurs in the Lax–Phillips scattering theory,[20] where it is called the translation representation.

Although this paper has captured some of the chaos-related features of the translation representation, it must be emphasized that the Lax–Phillips theory was developed for the wave equation. It cannot be used in quantum mechanics. Corresponding to the Hilbert space N, we have $L^2(S^2)$, the space of square-integrable functions on the unit sphere in $\mathbf{R}^3$. It makes a difference, however, that we have a translation semigroup on $L^2(\mathbf{R}^+)$, not a group on $L^2(\mathbf{R})$.

The starting point for this paper is the group $\{\exp(-iH_0t)\}$ on L^2 with self-adjoint generator H_0. The spectrum of the generator is $[0,\infty)$. The group $\{V(t)\}$ in Eq. (10.1) has self-adjoint generator $-id/ds$ with spectrum $(-\infty,\infty)$. The semigroup $\{U(t)\}$ on K^2 has generator k^2. If we let Dom(k^2) be the set of all $f\in K^2$ with the property that $k^2f(k,\omega)\in K^2$, it follows from Eq. (2.2) that $Jk^2f=-k^2Jf$ for all $f\in\text{Dom}(k^2)$. This makes it easy to show that the operator k^2 is closed and symmetric. To identify its spectrum, we examine the resolvent $(k^2-\lambda)^{-1}$. For fixed λ, the resolvent is a bounded operator on K^2 if and only if $(k^2e^{2i\phi}-\lambda)^{-1}f(ke^{i\phi},\omega)\in G^2$ for every $f\in G^2$. Since $k^2e^{2i\phi}$ runs through the lower half-plane, the open upper half-plane is the resolvent set, and the closed lower half-plane is the spectrum of the operator k^2 on K^2. This is an example of an operator that does not have self-adjoint extensions.

Although the Hamiltonian H_0 does not belong to the cone Γ consisting of allowed operators A on L^2, we can define its counterpart $\tilde{H}_0$ on K^2 as follows. Let Dom(q) be the set of all $f\in K^2$ with the property that $kf\in L^2$. Define the quadratic form

$$q[f,g]:=(kf,kg) \tag{10.2}$$

on K^2 with form domain Dom(q). This form is densely defined and positive. By Lemma 10.1 below, the form is closed. Hence it defines a positive self-adjoint operator $\tilde{H}_0$ on K^2 with $\text{Dom}(\tilde{H}_0)\subset\text{Dom}(q)$. For $f\in\text{Dom}(\tilde{H}_0)$ and $g\in\text{Dom}(q)$,

$$q[f,g]=\langle\tilde{H}_0f,g\rangle.$$

For $f,g\in\text{Dom}(q)$ we have Eq. (10.2) and

$$q[f,g]=\langle\tilde{H}_0^{1/2}f,\tilde{H}_0^{1/2}g\rangle.$$

If $f,g\in K^2$ and both $k^2f\in L^2$ and $k^2g\in L^2$,

$$q[f,g]=(k^2f,g)=\langle(1+J^2)^{-1}k^2f,g\rangle=(f,k^2g)=\langle f,(1+J^2)^{-1}k^2g\rangle.$$

Hence $\tilde{H}_0$ is the Friedrichs extension[9] of $(1+J^2)^{-1}k^2$.

The foregoing illustrates that the generator of the semigroup $\{U(t)\}$ is very different from $\tilde{H}_0$. If $f\in\text{Dom}(q)$, then $U(t)f\in\text{Dom}(q)$. For $f,g\in\text{Dom}(q)$,

$$\langle\tilde{H}_0^{1/2}U(t)f,\tilde{H}_0^{1/2}U(t)g\rangle=(kU(t)f,kU(t)g)=(kf,kg)=\langle\tilde{H}_0^{1/2}f,\tilde{H}_0^{1/2}g\rangle.$$

Hence the energy represented by $\tilde{H}_0$ on K^2 is conserved under the time evolution represented by the semigroup $\{U(t)\}$ on K^2.

It remains to prove that the form (10.2) is closed.

Lemma 10.1: Let Dom(q) *be the set of all* $f\in K^2$ *with the property that* $kf\in L^2$*. The form (10.2) on* K^2 *with form domain* Dom(q) *is closed.*

Proof: Let $\{f_n\}(n=1,2,\ldots)$ be a sequence in Dom(q) such that $|||f_n-f|||\to 0$ and $q[f_n-f_m,f_n-f_m]\to 0$ as $n,m\to\infty$. We have to show that $f\in$ Dom(q) and $q[f_n-f,f_n-f]\to 0$.

By assumption $f\in K^2$. Moreover, $||f_n-f||\to 0$ and $||k(f_n-f_m)||\to 0$. Since k is a closed operator on L^2, it follows that $kf\in L^2$ and $||k(f_n-f)||\to 0$. The last statement means that $q[f_n-f,f_n-f]\to 0$. Since $f\in K^2$ and $kf\in L^2$, we have $f\in$ Dom(q). Hence all requirements for the form to be closed are satisfied. □

[1] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems* (Benjamin/Cummings, Menlo Park, 1986).
[2] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer-Verlag, New York, 1990).
[3] F. Haake, *Quantum Signatures of Chaos* (Springer-Verlag, Berlin, 1991).
[4] L. E. Reichl, *The Transition to Chaos in Conservative Classical Systems: Quantum Manifestations* (Springer-Verlag, New York, 1992).
[5] C. van Winter, ''An increasing entropy for a free quantum particle,'' J. Math. Phys. **39**, 3600–3618 (1998).
[6] C. van Winter, ''Fredholm equations on a Hilbert space of analytic functions,'' Trans. Am. Math. Soc. **162**, 103–139 (1971).
[7] C. van Winter, ''Complex dynamical variables for multiparticle systems with analytic interactions. I,'' J. Math. Anal. Appl. **47**, 633–670 (1974).
[8] T. Kato, *Perturbation Theory for Linear Operators* (Springer-Verlag, Berlin, 1966), Chap. VI, Sec. 1.
[9] Reference 8, Chap. VI, Sec. 2.
[10] J. L. Kelley, *General Topology* (Van Nostrand, New York, 1955).
[11] Reference 8, Chap. V, Sec. 2.
[12] K. Hoffman, *Banach Spaces of Analytic Functions* (Prentice–Hall, Englewood Cliffs, 1962).
[13] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer-Verlag, New York, 1983).
[14] A. J. Lichtenberg and M. A. Lieberman, *Regular and Chaotic Dynamics*, 2nd ed. (Springer-Verlag, New York, 1992).
[15] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, Cambridge, 1993).
[16] C. Beck and F. Schlögl, *Thermodynamics of Chaotic Systems* (Cambridge University Press, Cambridge, 1993).
[17] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems* (Cambridge University Press, Cambridge, 1995).
[18] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding* (Cambridge University Press, Cambridge, 1995).
[19] J. G. Sinai, ''Dynamical systems with countably-multiple Lebesgue spectrum. I,'' Izv. Akad. Nauk SSSR Ser. mat. **25**, 899–924 (1961) [Am. Math. Soc. Transl. Ser. 2 **39**, 83–110 (1964)].
[20] P. D. Lax and R. S. Phillips, *Scattering Theory* (Academic, New York, 1967).

Nonlocal potentials, isolated states, and Levinson's theorem

T. A. Weber
Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011

(Received 22 June 1998; accepted for publication 15 September 1998)

A compact expression for the calculation of phase shifts is derived for a potential which is the sum of local and nonlocal parts. Nonlocal potentials can support positive energy bound states, that is, states embedded in the continuous energy spectrum. These states, sometimes referred to as ''isolated'' states, are not associated with any poles of the S matrix. Some controversy exists in the literature on how such bound states are included in Levinson's theorem; it is found that the phase shift should be taken continuous at the energy of the bound state rather than taken to have a discontinuity of π. For simplicity, the analysis is restricted to the radial s wave Schrödinger equation and separable nonlocal potentials. © *1999 American Institute of Physics.* [S0022-2488(99)03601-4]

I. INTRODUCTION

As is well known, in an attempt to account for the suppression of degrees of freedom, nonlocal potentials are used in the description of many particle systems to create an effective equation for the relative motion of two particles. Such potentials are also useful in creating phase-equivalent potentials for the study of nucleon–nucleus and nucleus–nucleus scattering and cluster models of light nuclei. For these reasons it would be advantageous to have a compact expression from which the phase shifts could be found so that the effect of changes in the potential could more easily be gauged.

Nonlocal potentials can support bound states with energy embedded in the continuous spectrum. These states are sometimes called ''isolated'' since the S matrix element does not have a pole at the energy of such a state.[1] How to include such states in Levinson's theorem has been controversial.

In the rest of Sec. I, the standard analysis of nonlocal potentials[2] is reviewed. Then in Sec. II, it is shown that, for Levinson's theorem to hold, the phase shift should be taken continuous at bound states with energy embedded in the continuum. In Secs. III and IV, the potential is extended to include the addition of a local potential and in Sec. V, it is shown that the phase shift is just the negative of the phase of a compact expression which plays the same role as the Jost function does for local potentials. It is seen that Levinson's theorem also holds for this general case. And finally, in Sec. VI, an example is given.

The analysis is restricted to the radial s wave Schrödinger equation

$$-\frac{d^2}{dr^2}\,\psi(r)+V(r)\psi(r)+\int_0^\infty U(r,r')\psi(r')dr'=k^2\psi(r), \tag{1}$$

where $\hbar=2m=1$ and $E=k^2$.

The nonlocal potential is taken to be separable,

$$U(r,r')=\sum g_i v_i(r)v_i(r'), \tag{2}$$

where $v_i(r)$ are real and well behaved, and the g_i are real constants, positive or negative. These conditions ensure that the Hamiltonian is Hermitian. Only one term in the sum will be used in the following analysis; the results are easily generalized.

The physical solution of Eq. (1) is defined by the mixed boundary conditions,

$$\psi(k,0)=0, \tag{3a}$$

and, for large r,

$$\psi(k,r)\rightarrow \tfrac{1}{2}i[e^{-ikr}-S(k)e^{ikr}], \tag{3b}$$

where $S(k)$ is the scattering matrix element which can be written in terms of the phase shift,

$$S(k)=e^{2i\delta(k)}. \tag{4}$$

In these equations, k is taken to be real positive. The usual notation of a superscript ''+'' indicating the physical solution and the corresponding Green's function is not used here since there is little likelihood of confusion. In the following, the Fredholm determinant of the physical solution, written as an integral equation, is constructed. The zeros of this determinant give the energies of solutions to the homogeneous equation and these solutions, if normalizable, are bound states. In this section the local potential is taken to be zero.

Taking the nonlocal potential to have only one term, the physical solution to the Schrödinger equation is

$$\psi(k,r)=\sin kr+\int_0^\infty ds\ K(r,s)\psi(k,s), \tag{5}$$

with kernel,

$$K(r,s)=gv(s)\int_0^\infty dr'\ G(k,r,r')v(r'), \tag{6}$$

and Green's function,

$$G(k,r,r')=-\frac{1}{k}\, e^{ikr_>}\sin kr_<, \tag{7}$$

where $r_>$ indicates the larger of r and r' while $r_<$ denotes the smaller. Note that the integral for the kernel is convergent for Im $k>0$, that is, the solutions will be analytic in the upper-half of the complex k plane. If $v(r)$ is of finite range, the solutions will also be analytic functions of k in the lower half plane while if $v(r)$ is exponentially damping, the solutions will be analytic in a strip along the real k axis in the lower half plane.

The Fredholm determinant is easily found to be

$$D(k)=1-\int_0^\infty K(r,r)dr=1-g\int_0^\infty dr\ v(r)\int_0^\infty dr'\ G(k,r,r')v(r'). \tag{8}$$

This determinant can be written in the momentum representation on the half-line using the transform

$$v(r)=\frac{2}{\pi}\int_0^\infty p\ dp\ \tilde{v}(p)\sin pr, \tag{9}$$

with

$$\tilde{v}(p)=\int_0^\infty v(r)\ \frac{\sin pr}{p}\ dr. \tag{10}$$

In the resulting integrals, it is assumed that k has a positive imaginary part so that the need for distributions is avoided. The result,

$$D(k)=1+\frac{2g}{\pi}\int_0^\infty p^2\ dp\ \frac{\tilde{v}^2(p)}{p^2-k^2}, \tag{11}$$

can be easily generalized to the sum of separable terms by replacing in the integrand,

$$g\tilde{v}^2 \Rightarrow g_1\tilde{v}_1^2(p)+g_2\tilde{v}_2^2(p)+\cdots\ . \tag{12}$$

The expression for the Fredholm determinant given in Eq. (11) is analytic in the upper half complex k plane. To evaluate $D(k)$ for real momentum values, k is taken to approach the real positive axis from above. Then,

$$D(k)=1+\frac{2g}{\pi}\ \wp\int_0^\infty p^2\ dp\ \frac{\tilde{v}^2(p)}{p^2-k^2}+igk\ \tilde{v}^2(k), \tag{13}$$

where k is real positive and $\wp$ indicates the Cauchy principal part. For the conjugate physical solution, k is taken to approach the negative real axis.

The determinant $D(k)$ will be analytic in the lower half k plane within a strip along the real axis provided $v(r)$ is exponentially damping for large r. To continue $D(k)$ into the lower half plane, the path of integration is distorted around the simple pole at k as k is moved from the upper half plane to the lower. Finally the contour of integration is pinched off as k passes into the lower half plane. The result is exactly the same as given in Eq. (11) but with the addition of the residue at the pole:

$$D(k)=1+\frac{2g}{\pi}\int_0^\infty p^2\ dp\ \frac{\tilde{v}^2(p)}{p^2-k^2}+2igk\ \tilde{v}^2(k). \tag{14}$$

It is not difficult to show that $D(k)$ given by Eq. (11) or the continuation given in Eq. (14) satisfies the condition,

$$D(k)=D^*(-k^*), \tag{15}$$

where (*) indicates complex conjugate.

For a solution of the homogeneous equation at k_0 real and positive, both the real and imaginary parts of the Fredholm determinant given in Eq. (13) must be zero,

$$1+\frac{2g}{\pi}\ \wp\int_0^\infty p^2\ dp\ \frac{\tilde{v}^2(p)}{p^2-k_0^2}=0, \tag{16a}$$

and

$$\tilde{v}^2(k_0)=0. \tag{16b}$$

Such a solution may represent a bound state in the continuous spectrum provided it is normalizable. The homogeneous equation is

$$\psi(k,r)=-\frac{g}{k}\int_0^\infty ds'\ v(s)\psi(k,s)\left[e^{ikr}\int_0^r dr'\ \sin kr' v(r')+\sin kr\int_r^\infty dr'\ e^{ikr'}\ v(r')\right]. \tag{17}$$

From Eq. (16b) it is seen that this solution is zero for $r \to \infty$. If this approach to zero is fast enough, there will be a bound state which can be written as

$$\psi(k_0, r) = C \int_0^\infty d\xi \sin k_0 \xi \ v(\xi + r), \tag{18}$$

where C is a constant. It is apparent that if $v(r)$ is of finite range or decreases exponentially, the solution will be normalizable. If $v(r)$ goes as $r^{-\alpha}$ then the wave function also goes as $r^{-\alpha}$ so that, if α is greater than 1/2, the solution will be normalizable. These normalizable wave functions represent bound states with energy embedded in the continuous spectrum.

II. THE PHASE SHIFT AND LEVINSON'S THEOREM

Levinson's theorem states that

$$\delta(0) = \pi(n + q/2), \tag{19}$$

where the phase shift $\delta(k)$ is evaluated at zero momentum, n is the number of bound states, and q is either 1 or 0 depending on whether or not a ''half bound'' state exists. There has been some disagreement in the literature on the application of this theorem to the case of nonlocal potentials which support bound states with energy embedded in the continuous spectrum. Bolsterli[3] suggests that a jump of π should be added to the phase shift at the energy of the bound state whereas Newton[4] states that the phase shift should be taken continuous. The following derivation confirms Newton's analysis.

It is not difficult to show that the t matrix is proportional to the reciprocal of the Fredholm determinant and that the Jost function is proportional to the determinant[3] with a real proportionality factor.[5] Therefore, the phase shift can be taken to be the negative phase of $D(k)$, that is,

$$D(k) = |D(k)| e^{-i\delta(k)}. \tag{20}$$

This phase shift has a discontinuity of π at k_0, the value of k associated with the bound state in the continuum. This jump of π comes about because $D(k)$ goes through zero at k_0 where the real part changes sign but not the imaginary. But the S matrix element,

$$S(k) = \frac{D^*(k)}{D(k)} = e^{2i\delta(k)}, \tag{21}$$

is continuous at k_0. There are no zeros for $D(k)$ in the upper half plane except possibly on the positive imaginary k axis. Such zeros correspond to conventional bound states with negative real energies. If the continuation to the lower half plane is possible then $D(k)$, given by Eq. (14), may also have zeros there. In the rest of the analysis, it will be assumed that $D(k)$ has a zero at k_0 which may be complex. Then it follows from Eq. (15) that $D(k)$ has a zero at $-k_0^*$, symmetrically placed with respect to the imaginary axis, and that for real k,

$$\delta(k) = -\delta(-k). \tag{22}$$

Suppose that these two zeros at k_0 and $-k_0^*$ can be moved around in the complex k plane by varying the parameters of the potential. With zeros only in the lower half plane, that is, no bound states, the following integral is zero:

$$I = \int_{-\infty}^{\infty} d \ln D(k) = \int_{-\infty}^{\infty} \frac{1}{D(k)} \frac{dD(k)}{dk} dk = 0. \tag{23}$$

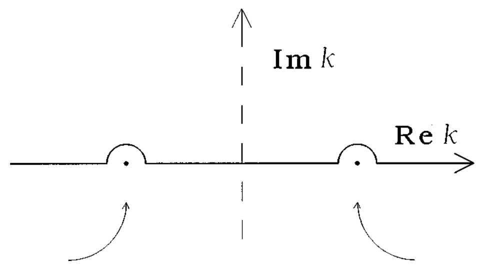

FIG. 1. Path of integration for Levinson's theorem. The zeros of $D(k)$, symmetric about the imaginary axis, approach the real axis from below.

This is readily seen by closing the contour of integration by a large semicircle in the upper half plane and letting the radius go to infinity. Doing the integral along the real axis and noting the definition of the phase shift in terms of $D(k)$, one obtains

$$\delta(0)-\delta(\infty)=0. \tag{24}$$

As usual, the phase shift for infinite wave number is taken to be zero and, since there are no zeros on the real axis, the phase shift is continuous. In the following, continuous phase shifts will be denoted by a subscript c.

Imagine how the phase shift changes when k is varied along the real axis and passes near a simple pole due to a zero of $D(k)$ in the lower half k plane. The phase shift will increase sharply but continuously through $\pi/2$. This behavior of the phase shift is usually taken to indicate a resonance at the real part of the wave number of the zero.

Now suppose that the parameters are varied such that the zeros approach the real axis from below. The path of integration is deformed around the zeros as shown in Fig. 1. It should be noted that it is not possible to have the zero pop up into the upper half k plane for Re $k>0$ and still have a Hermitian Hamiltonian. The integral is again zero as can be seen by closing the contour by a large semicircle. Assuming that the zeros are simple, one obtains

$$I=\wp\int_{-\infty}^{\infty} d\ \ln D(k)-2i\pi=2i\delta_c(0)-2\pi i, \tag{25}$$

where the $-2\pi i$ comes from the integration around the two semicircles and the principal part defines a continuous phase shift,

$$\wp\int_{-\infty}^{\infty} d\ \ln D(k)=2i\delta_c(0). \tag{26}$$

Since I given in Eq. (25) is zero, Levinson's theorem reads, for a bound state with energy embedded in the continuum,

$$\delta_c(0)=\pi, \tag{27}$$

with the phase shift defined continuously. If the zero is of order m then

$$\delta_c(0)=m\pi. \tag{28}$$

It may surprise one that the principal part defines a continuous phase shift since, for a simple pole at k_0 real and positive, the integration from 0 to k_0 will give a logarithmic singularity. But one must remember to include the integration from $-k_0$ to 0 which results in a cancellation of the logarithmic singularity and gives a continuous phase shift as k passes through k_0.

In an actual experiment, the measurement of the phase shift will be continuous since any jump of π will not affect the cross section. There appears to be nothing in the phase shift at the energy of the bound state that would signify its existence. The existence of the bound state, however, shows up in the modification of the phase shift at zero energy as described by Levinson's theorem.

III. CASE WITH AN ADDITIONAL LOCAL POTENTIAL

The basis functions to be used for the case of a local plus a nonlocal potential are the solutions of the Schrödinger equation with only the local potential. These solutions will be denoted by a subscript 0. The Fredholm determinant for the integral equation of the physical solution for the combined potentials is

$$D(k)=1-g\int dr\ dr'\ G(k,r,r')v(r)v(r'), \tag{29}$$

where the Green's function is

$$G(k,r,r')=-\frac{f_0(k,r_>)\phi_0(k,r_<)}{\mathcal{F}_0(k)}. \tag{30}$$

The Jost solution is defined by its asymptotic form for large r

$$f_0(k,r)\rightarrow e^{ikr}, \tag{31}$$

while the regular solution, $\phi_0(k,r)$, is defined such that, at the origin, the function and its first derivative are zero and one, respectively. The Jost function,

$$\mathcal{F}_0(k)=f_0(k,0), \tag{32}$$

will have zeros on the imaginary axis in the upper half complex k plane at the location of bound states of the local potential.

For reference,

$$\phi_0(k,r)=\frac{1}{2ik}\,[\mathcal{F}_0^*(k)f_0(k,r)-\mathcal{F}_0(k)f_0^*(k,r)], \tag{33}$$

for real values of k. Furthermore, the regular solution is real for real k or for purely imaginary k and, for the usual restrictions on the potential, it is regular for all k while the Jost solution is analytic in the upper half k plane. For a discussion of these properties see Newton.[6]

The nonlocal potential function $v(r)$ can be expanded in terms of the Jost solution and the regular solution of the local potential by use of the completeness relation,[7]

$$\int_C \frac{k\ dk\ \phi_0(k,r)f_0(k,r')}{\mathcal{F}_0(k)}=i\pi\delta(r-r'), \tag{34}$$

where the contour C goes along the real k axis from minus infinity, over any zeros of $\mathcal{F}_0(k)$ in the upper half plane, to plus infinity and as shown in Fig. 2. In this way the bound states of the local potential are included in the expansion.

Using the completeness relation gives

$$v(r)=\int_C \frac{p\ dp}{\mathcal{F}_0(p)}\ \phi_0(p,r)\overline{v}(p), \tag{35}$$

where

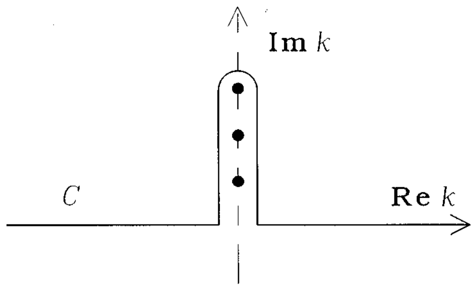

FIG. 2. Integration contour for the completeness relation. The dots on the imaginary axis represent the zeros of the Jost function.

$$\overline{v}(p)=\frac{1}{\pi i}\int_0^\infty dr\ v(r)f_0(p,r). \tag{36}$$

Substituting for $v(r)$ from Eq. (35) into $D(k)$ gives

$$D(k)=1+g\int_C \frac{p\ dp}{\mathcal{F}_0(p)}\frac{\overline{v}(p)\tilde{v}(p)}{p^2-k^2}, \tag{37}$$

where

$$\tilde{v}(p)=\int_0^\infty dr\ v(r)\phi_0(p,r). \tag{38}$$

To arrive at this result it is useful to use the relation

$$\frac{d}{dr}W[\phi_0(k_1,r),\psi_0(k_2,r)]=(k_1^2-k_2^2)\phi_0(k_1,r)\psi_0(k_2,r), \tag{39}$$

for the Wronskian, valid for any solution to the Schrödinger equation with the local potential regardless of boundary conditions.

For the particular case of no bound states of the local potential, Eq. (37) reduces to

$$D(k)=1+\frac{2g}{\pi}\int_0^\infty \frac{p^2\ dp\ \tilde{v}^2(p)}{(p^2-k^2)|\mathcal{F}_0(p)|^2}, \tag{40}$$

as given in Ref. 2.

But $D(k)$ given in Eq. (37) cannot tell the whole story since for $v(r)=0$, $D(k)=1$. This appears to say there are no bound states even if the local potential supports a bound state. Examination of the t matrix[8] suggests that for the local and nonlocal potentials combined, Eq. (37) should be modified by the overall factor $\mathcal{F}_0(k)$ on the right-hand side.

IV. GENERAL CASE WITH ALL ZEROS OF BOUND STATES CONTAINED IN $D(k)$

For the general case, one could proceed by finding the Fredholm determinant for the integral equation of the physical solution with the free particle states as the basis. But then to get an expression like Eq. (37) would require changing the basis. Rather than searching for the Fredholm determinant, it is simpler to find the Jost function for the combined potentials in the basis of the solutions of the local potentials. The integral equation for the Jost solution is

$$f(k,r)=f_0(k,r)+\int_r^\infty dr'\ G(k,r,r')\int_0^\infty ds\ U(r',s)f(k,s), \tag{41}$$

where the subscript 0 denotes solutions for the local potential. Substituting the nonlocal potential given in Eq. (2) yields

$$f(k,r)=f_0(k,r)+gI(k)\int_r^\infty dr'\ G(k,r,r')v(r'), \tag{42}$$

where

$$I(k)\equiv\int_0^\infty ds\ v(s)f_0(k,s). \tag{43}$$

The function $I(k)$ is found by multiplying Eq. (42) by $v(r)$, integrating, and then solving the resulting algebraic equation to get

$$I(k)=\frac{\int_0^\infty dr\ f_0(k,r)v(r)}{1-g\int_0^\infty dr'\ v(r')\int_{r'}^\infty dr\ v(r)G(k,r,r')}. \tag{44}$$

The Green's function that is 0 for $r'<r$ and satisfies

$$\left[-\frac{d^2}{dr^2}+V(r)-k^2\right]G(k,r,r')=-\delta(r-r') \tag{45}$$

can be constructed out of solutions of the Schrödinger equation with the local potential. One finds that

$$G(r,r',k)=-\frac{\phi_0(k,r)f_0(k,r')}{\mathcal{F}_0(k)}+\frac{\phi_0(k,r')f_0(k,r)}{\mathcal{F}_0(k)} \quad \text{for } r<r',$$

$$=0 \quad \text{for } r'<r, \tag{46}$$

which has the necessary unit step at $r=r'$ in the first derivative.

The relation given in Eq. (39) is useful for doing the integrals involving the Green's function. Again the function $v(r)$ is expanded in terms of the basis functions using the completeness relation, Eq. (34). After some straightforward manipulation it is found that

$$I(k)=\pi i\left[\overline{v}(k)\Big/\left(1+g\int_C\frac{pdp}{\mathcal{F}_0(p)}\frac{\tilde{v}(p)}{p^2-k^2}[\overline{v}(p)-\overline{v}(k)]\right)\right], \tag{47}$$

so that the Jost function, found by evaluating the Jost solution at $r=0$, is

$$\mathcal{F}(k)=\mathcal{F}_0(k)\left(1+g\int_C\frac{pdp}{\mathcal{F}_0(p)}\frac{\overline{v}(p)\tilde{v}(p)}{p^2-k^2}\right)\Big/\left(1+g\int_C\frac{pdp}{\mathcal{F}_0(k)}\frac{\tilde{v}(p)}{p^2-k^2}[\overline{v}(p)-\overline{v}(k)]\right). \tag{48}$$

It is not difficult to show that the denominator of the Jost solution on the right-hand side of Eq. (48) is the Fredholm determinant of the regular solution in the basis of the solutions for the local potential. Hence the Jost function for the general case is just the product of the Jost function for the local potential by itself times the ratio of the Fredholm determinants of the physical solution and the regular solution. For a bound state in the continuous spectrum, say at k_0, the numerator of the Jost function or equivalently $D(k)$ given in Eq. (37) must be zero. This requires that the real and imaginary parts must be separately zero, that is,

$$1+g\wp\int_C\frac{pdp}{\mathcal{F}_0(p)}\frac{\overline{v}(p)\tilde{v}(p)}{p^2-k_0^2}=0, \tag{49a}$$

and

$$\tilde{v}^2(k_0)=0. \tag{49b}$$

These conditions also imply that the denominator of the Jost function is also zero at k_0. For the Jost function, there is a jump in phase of π at k_0 for both the denominator and the numerator. Thus the Jost function is continuous at k_0 unlike the function $D(k)$.

Since the Fredholm determinant of the regular solution is real for real values of k, the denominator of the Jost function is real so that it can be dropped in calculating the S matrix element formed by the ratio $\mathcal{F}^*(k)/\mathcal{F}(k)$. This justifies using

$$D(k)=\mathcal{F}_0(k)\left[1+g\int_C \frac{p\ dp}{\mathcal{F}_0(p)}\,\frac{\overline{v(p)}\tilde{v}(p)}{p^2-k^2}\right], \tag{50}$$

for calculating the phase shift. Note that the results of Sec. II on Levinson's theorem apply here also and therefore the phase shift is taken continuous at bound states in the continuous spectrum. This compact form for $D(k)$ is the main result of this paper.

V. EXAMPLE

It is simple to arrange things so that the only effect of the nonlocal potential is to shift the energies of any bound state of the local potential. Take $v(r)$ for the nonlocal potential to be the same as one of the bound states eigenfunctions of the local potential, say at $E=k_0^2<0$. Then the contour of integration for completeness reduces to a contour around the simple pole due to the zero of $\mathcal{F}_0(k)$ at k_0. Using the linear dependence of $f_0(k,r)$ and $\phi_0(k,r)$ at a zero of $\mathcal{F}_0(k)$, one obtains

$$D(k)=\mathcal{F}_0(k)\left\{1+\frac{4ig\ k_0^2\tilde{v}^2(k_0)}{\mathcal{F}_0^*(k_0)[d\mathcal{F}_0(k)/dk]_{k_0}(k^2-k_0^2)}\right\}. \tag{51}$$

That this function is not equal to zero at k_0 can be seen by rewriting $D(k)$ as

$$D(k)=\mathcal{F}_0(k)\left[\frac{k^2+K^2-gN^2\tilde{v}^2(k_0)}{k^2+K^2}\right], \tag{52}$$

where

$$k_0=iK, \tag{53}$$

with K real and positive, and N the normalization of the regular solution of the local potential,

$$N^2\int\ dr\ \phi_0^2(k_0,r)=1. \tag{54}$$

It is apparent that the zero of $D(k)$ now appears at

$$k^2=-K^2+gN^2\tilde{v}^2(k_0), \tag{55}$$

which, for a positive g large enough, is a bound state embedded in the continuous spectrum. In a specific example studied in an earlier paper,[9] a bound state of the local potential was moved to positive energy by use of a nonlocal potential. The phase shift was found to be the negative phase of the Jost function of the local potential, that is, the term in braces in Eq. (51) was missing. The resulting cross section is the same in either case since the missing term is real. But in the earlier work, Levinson's theorem held because of the zero in the Jost function at the original energy of the bound state, as if its ghost remained. The situation described in this paper is much more satisfying.

[1] A. M. Shirokov, Yu. F. Smirnov, and S. A. Zaytsev, Rev. Mex. Fis. **40**, 74 (1994).
[2] K. Chardan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1989), pp. 112–118 and 122–124.
[3] M. Bolsterli, Phys. Rev. **182**, 1095 (1969).
[4] R. G. Newton, J. Math. Phys. **18**, 1582 (1977).
[5] B. Mulligan, L. G. Arnold, B. Bagchi, and T. O. Krause, Phys. Rev. C **13**, 2131 (1976).
[6] Roger G. Newton, *Scattering Theory of Waves and Particles* (McGraw–Hill, New York, 1996), pp. 331–341.
[7] T. A. Weber and C. L. Hammer, J. Math. Phys. **21**, 24 (1980).
[8] See Ref. 6, pp. 266–276.
[9] T. A. Weber, *Proceedings of the Tenth Workshop on High Energy Physics and Quantum Field Theory*, edited by B. B. Levchenko (Moscow State University, Moscow, 1997), pp. 419–424.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

An estimate of the ground state energy of the fractional quantum Hall effect

Jingbo Xia[a)]
Department of Mathematics, State University of New York, Buffalo, New York 14214-3093

(Received 24 June 1998; accepted for publication 28 August 1998)

Suppose that there are N electrons in a disk of radius R with a perpendicular magnetic field. We give an estimate for the ground state energy of such a system in the case $N \approx \nu R^2$. © *1999 American Institute of Physics.* [S0022-2488(99)02101-5]

I. INTRODUCTION

In the study of the fractional quantum Hall effect,[1] one considers the two-dimensional Coulomb interactions of electrons in a magnetic field. In this paper we will give an estimate for the ground state energy of such a system. Suppose that N electrons are confined to a disk of radius R. We will show that in the limit $R \to \infty$ and $N/R^2 \to \nu$, where ν is the *Landau level filling factor*, the ground state energy is on the order of $N^{3/2}$. Our result is motivated by the observation that, in the Hamiltonian for N electrons, there are $N(N-1)/2$ terms of Coulomb potentials compared with only N terms for kinetic energy. Therefore, for large N, the total energy of the system can be close to its minimum, even when the kinetic energy is not. This will be seen from the state we use to obtain the estimate $N^{3/2}$, which imposes a uniform spacing between the electrons and is considerably different from the Laughlin state.[2]

More specifically, let us consider N electrons in the plane $\mathbf{R}^2$ interacting with each other, with a randomly distributed, bounded potential, and with a constant magnetic field $\mathbf{B}$ in the perpendicular direction. In units of magnetic length $(2c\hbar/e|\mathbf{B}|)^{1/2}$ and cyclotron energy $e|\mathbf{B}|\hbar/2m_e c$, the Hamiltonian representing this system reads as

$$H_N = \sum_{j=1}^{N} \left(\frac{1}{2} D_j + Q(z_j) \right) + \lambda \sum_{1 \leq j < k \leq N} |z_j - z_k|^{-1}.$$

Here, $D_j = (-i(\partial/\partial x_j) + y_j)^2 + (-i(\partial/\partial y_j) - x_j)^2$, $z_j = (x_j, y_j) \in \mathbf{R}^2$, Q is a bounded, real-valued function, and $\lambda = (2e^3 m_e^2 c/\hbar^3 |\mathbf{B}|)^{1/2}$. Since $e^3 m_e^2 c/\hbar^3 \approx 2.43 \times 10^9$ G, the dimensionless constant λ is large under earthly conditions.

For $R > 0$, let $\Delta(N,R)$ be the polydisk $\{(z_1, \ldots, z_N) : |z_j| < R,\ j = 1,...,N\}$ in $(\mathbf{R}^2)^N$. Let $H(N,R)$ be the restriction of H_N to $\Delta(N,R)$ with the Dirichlet boundary condition. By a well-known construction[3,4] of Friedrichs, $H(N,R)$ is realized as a self-adjoint operator [which we also denote by $H(N,R)$] on a dense domain in the Hilbert space $L^2(\Delta(N,R))$. Let $\mathcal{D}_a(N,R)$ be the collection of the functions in the domain of $H(N,R)$ that are antisymmetric under the interchange of any two variables. We define

$$H_a(N,R) = H(N,R) | \mathcal{D}_a(N,R),$$

the restriction of $H(N,R)$ to $\mathcal{D}_a(N,R)$. If the electrons are confined to the disk $\{z \in \mathbf{R}^2 : |z| < R\}$, then, taking the exclusion principle into account, the more appropriate operator to consider is $H_a(N,R)$ rather than H_N.

Let $I(N,R)$ be the ground state energy of $H_a(N,R)$. That is,

[a)]jxia@acsu.buffalo.edu

$$I(N,R)=\inf\{\langle H_a(N,R)f,f\rangle : f\in\mathcal{D}_a(N,R), \|f\|=1\}.$$

A natural question concerning the operator $H_a(N,R)$ is, what is $I(N,R)$? Or, at the very least we would like to know the asymptotics of $I(N,R)$ as $R\to\infty$ and $N/R^2\to\nu$, where $\nu>0$ is the Landau level filling factor of the system under consideration. This is precisely the situation that arises in the study of the fractional quantum Hall effect (FQHE). See Ref. 1 and Sec. VII in Ref. 5.

The Dirichlet boundary condition guarantees that the operator $\Sigma_{j=1}^N D_j$ is positive. Hence, it follows that $I(N,R)\geqslant -\|Q\|_\infty N+(\lambda N(N-1)/2)\inf_{|z|<R,|w|<R}|z-w|^{-1}\geqslant -\|Q\|_\infty N+\lambda(N(N-1)/4R)$. Thus, for large R, if $N/R^2\approx\nu$, then $I(N,R)\geqslant\frac{1}{4}\lambda\sqrt{\nu}N^{3/2}-KN$ for some $0<K<\infty$, which is independent of N. That is,

$$\liminf_{\substack{R\to\infty\\ N/R^2\to\nu}} N^{-3/2}I(N,R)\geqslant 0.25\times\lambda\sqrt{\nu}.$$

Note that this lower-bound estimate for $I(N,R)$ is extremely crude; it does not even take the exclusion principle into account. What is amazing is that crude as it is, this estimate is only off by a factor of less than 4. The purpose of this paper is to report the following.

Theorem: *Suppose that* $\nu>0$. *Then*

$$\limsup_{\substack{R\to\infty\\ N/R^2\to\nu}} N^{-3/2}I(N,R)\leqslant\frac{8}{3\pi}\lambda\sqrt{\nu}=0.8488...\times\lambda\sqrt{\nu}. \tag{1}$$

A consequence of our theorem is that it sets a cap on the energy of any approximation to the ground state of the FQHE. In particular, this cap applies to the Laughlin state, which lies at the heart of the incompressible quantum fluid approach to the FQHE. Recall that the Laughlin m state for N electrons is

$$\Phi_N^m(z_1,\ldots,z_N)=\exp\left(-\sum_{j=1}^N\frac{1}{2}|z_j|^2\right)\prod_{1\leqslant j<k\leqslant N}(z_j-z_k)^m,$$

where $m>1$ is an odd integer.[1,2] This function is generally taken as an approximation to the ground state of FQHE with Landau level filling factor $1/m$. In light of our theorem, if Φ_N^m truly approximates the ground state of FQHE, then we should have

$$\limsup_{N\to\infty}\sum_{1\leqslant j<k\leqslant N}\frac{\langle|z_j-z_k|^{-1}\Phi_N^m,\Phi_N^m\rangle}{N^{3/2}\|\Phi_N^m\|^2}\leqslant\frac{8}{3\pi}\sqrt{\frac{1}{m}}. \tag{2}$$

So far, our attempts have failed to produce a proof of this. In fact, we are not even able to ascertain that the left-hand side of (2) is finite. But this seems to be an interesting problem, for it serves to test how closely Φ_N^m approximates the ground state of FQHE.

The original version of our theorem was stated as

$$\limsup_{\substack{R\to\infty\\ N/R^2\to\nu}} N^{-3/2}I(N,R)\leqslant C\lambda\sqrt{\nu}, \tag{3}$$

with a finite C that was not explicitly determined. And this was derived with a proof that involves the eigenfunctions of $D_1,\ldots,D_N$ and that is rather long and technical. Friesecke subsequently found a much simpler and shorter proof of (3). Using local gauge transformations, his proof avoids the eigenfunctions of $D_1,\ldots,D_N$ completely and circumvents several technical difficulties. More important, his use of local gauge transformations makes it possible to impose a uniform spacing between the electrons. As it turns out, with the electrons uniformly spaced, it is possible to obtain a reasonably good value for C. The proof of our theorem presented here is a refinement of

Friesecke's proof of (3). That is, by carefully tracking the various constants that appear in Friesecke's proof of (3), we are able to show that $8/3\pi$ will do for C. We thank Friesecke for the permission to use his proof of (3) here.

The proof of our theorem involves estimates that are quite technical. In order not to lose sight of the main idea, therefore, it is perhaps desirable to give a general outline before presenting the details.

Given N electrons in a disk of radius $R=O(\sqrt{N})$, we need to find a state Ψ whose total energy is $O(N^{3/2})$. If at all possible, one tries to achieve this in the form of a single Slater determinant,

$$\Psi(z_1,\ldots,z_N)=(N!)^{-1/2}\det(\varphi_i(z_j))_{i,j=1}^N.$$

The arrangement of the electrons, $\varphi_1,\ldots,\varphi_N$, are determined by the following considerations: If there is a cap on the kinetic energy of each electron, then the total kinetic energy will only be $O(N)$, which is negligible in comparison to $O(N^{3/2})$. The simplest way to obtain such a cap is to take an equally spaced lattice Γ of N points in the disk of radius R and translate a single electron state (which we call χ_β below) by the vectors in Γ. Because of the presence of the magnetic field, we also need to apply gauge transformations to the translated electron states to ensure that their kinetic energy is the same as that for χ_β, which provides the desired cap. Thus $\varphi_1,\ldots,\varphi_N$ are simply an enumeration of

$$\{U_kT_k\chi_\beta : k\in\Gamma\},$$

where T_k is the translation determined by the vector k and U_k the corresponding gauge transformation. This takes care of the kinetic part of the Hamiltonian. Since the random potential Q is assumed to be bounded, its contribution to the Hamiltonian is also $O(N)$.

To facilitate the estimate of the potential energy of Ψ, which is the dominant part, we confine the original state χ_β to a small disk so that there is no overlapping among $\varphi_1,\ldots,\varphi_N$. With such an arrangement of the electrons, when N is large, each individual electron sees a charge distribution on the disk that is approximately uniform. Accordingly, because of the $1/r$ falloff of the Coulomb potential, the potential energy attributable to each electron is on the order of

$$\int_{|z|\leqslant R}\frac{1}{|z|}\,dA(z)=O(R)=O(\sqrt{N}).$$

Hence, the total potential energy is $N\times O(\sqrt{N})=O(N^{3/2})$. By carefully working out the details of this argument, we obtain the numerical factor $8/3\pi$ in (1).

II. PROOF OF THEOREM

Throughout the proof, we let $B_r(w)$ denote the open disk $\{z\in\mathbf{R}^2 : |w-z|<r\}$ in $\mathbf{R}^2$. We start by picking an ϵ such that $0<\epsilon<\nu$. Suppose that $N\geqslant 10^6$ and R are so large that

$$N/R^2\leqslant\nu+\epsilon. \tag{4}$$

Note that $4(N^{1/6}-4)^{-1}<1$, which will be relevant later. Let $\alpha>0$ be such that

$$(N+17\sqrt{N})\alpha^2=\pi R^2. \tag{5}$$

We have $1/\alpha^2=(N/R^2)(1+17N^{-1/2})/\pi\leqslant(\nu+\epsilon)(1+17N^{-1/2})/\pi\leqslant 2\nu$.

Let η be a C^∞ function on $[0,\infty)$ such that $\eta=0$ on $[\frac{1}{3},\infty)$, $\eta=\text{const}$ on $[0,\frac{1}{4}]$, and $2\pi\int_0^\infty|\eta(r)|^2r\,dr=1$. Define $\chi(z)=\eta(|z|)$, $z\in\mathbf{R}^2$. Then $\chi(z)=0$ if $|z|\geqslant\frac{1}{3}$ and χ is a unit vector in $L^2(\mathbf{R}^2)$. Let $\beta=\min\{1,\alpha\}$ and $\chi_\beta(z)=\beta^{-1}\chi(\beta^{-1}z)$. Then $\|\chi_\beta\|=\|\chi\|=1$. Since χ_β is a radial function, we have $(y(\partial/\partial x)-x(\partial/\partial y))\chi_\beta=0$. Now $\chi_\beta(z)=0$ if $|z|\geqslant\beta/3$. Therefore $\||z|^2\chi_\beta\|\leqslant\beta^2\|\chi_\beta\|=\beta^2\leqslant 1$. Also, $(\Delta\chi_\beta)(z)=\beta^{-2}\{\beta^{-1}(\Delta\chi)(\beta^{-1}z)\}$, which implies $\|\Delta\chi_\beta\|=\beta^{-2}\|\Delta\chi\|\leqslant(1+2\nu)\|\Delta\chi\|$. Let $D=(-i(\partial/\partial x)+y)^2+(-i(\partial/\partial y)-x)^2$. Then

$$\|D\chi_\beta\|\leqslant\|\Delta\chi_\beta\|+\||z|^2\chi_\beta\|\leqslant(1+2\nu)\|\Delta\chi\|+1. \tag{6}$$

For each $k=(k_1,k_2)\in\mathbf{R}^2$, define the unitary operators

$$(T_k f)(x,y)=f(x-k_1,y-k_2) \quad \text{and} \quad (U_k f)(x,y)=e^{i(k_1y-k_2x)}f(x,y),$$

on $L^2(\mathbf{R}^2)$. If we set $D(k)=(-i(\partial/\partial x)+y-k_2)^2+(-i(\partial/\partial y)-x+k_1)^2$ for $k=(k_1,k_2)$, then it is easy to verify that $DU_k=U_kD(k)$ and $D(k)T_k=T_kD$. Thus, by (6),

$$\|DU_k\,T_k\chi_\beta\|=\|U_kD(k)T_k\chi_\beta\|=\|U_kT_k\,D\chi_\beta\|=\|D\chi_\beta\|\leqslant(1+2\nu)\|\Delta\chi\|+1. \tag{7}$$

Let

$$\Gamma=\left\{\left(\alpha i+\frac{\alpha}{2},\alpha j+\frac{\alpha}{2}\right):i,j\in\mathbf{Z},\alpha^2(i+a)^2+\alpha^2(j+b)^2<R^2 \quad \text{for all} \quad a,b\in\{0,1\}\right\}.$$

We claim that $\mathrm{Card}(\Gamma)\geqslant N$. For this purpose let N' be the number of squares of the size $\alpha\times\alpha$ with vertices of the form $(\alpha i,\alpha j)$, $i,j\in\mathbf{Z}$, which have a nonempty intersection with $\overline{B_R(0)}$, and let N'' be the number of such squares that are contained in $B_R(0)$. That is, $N''=\mathrm{Card}(\Gamma)$. If such a square intersects the circle $|z|=R$, then it is contained in the annulus $\{z\in\mathbf{R}^2:R-\sqrt{2}\alpha\leqslant|z|\leqslant R+\sqrt{2}\alpha\}$ since the diagonal of the square is $\sqrt{2}\alpha$. Hence $(N'-N'')\alpha^2\leqslant4\sqrt{2}\pi\alpha R$. That is, $N'-N''\leqslant4\sqrt{2}\pi R/\alpha\leqslant4\sqrt{2}\pi\sqrt{2N/\pi}=8\sqrt{\pi}\sqrt{N}\leqslant16\sqrt{N}$. On the other hand, $N'\alpha^2\geqslant\pi R^2=(N+17\sqrt{N})\alpha^2$. That is, $N'\geqslant N+17\sqrt{N}\geqslant N+N'-N''$. This proves that $\mathrm{Card}(\Gamma)=N''\geqslant N$.

Thus we can choose a subset $\Omega\subset\Gamma$ such that $\mathrm{Card}(\Omega)=N$. For each $k\in\Omega$, let $\psi_k=U_kT_k\chi_\beta$. Thus, each ψ_k is supported in $B_{\beta/3}(k)\subset B_{\alpha/3}(k)\subset B_R(0)$ and $\psi_k\psi_{k'}=0$ for all $k\neq k'$ in Ω. Let $\varphi_1,\ldots,\varphi_N$ be an enumeration of $\{\psi_k:k\in\Omega\}$, which is an orthonormal set. Define the Slater determinant $\Psi(z_1,\ldots,z_N)=(N!)^{-1/2}\det(\varphi_i(z_j))_{i,j=1}^N$. Since $\langle\Sigma_{j=1}^N D_j\Psi,\Psi\rangle=N\langle D_1\Psi,\Psi\rangle=N\Sigma_{\sigma\in S_N}\langle D\varphi_{\sigma(1)},\varphi_{\sigma(1)}\rangle/N!$, where S_N is the group of permutations of $\{1,...,N\}$, and since each $\varphi_{\sigma(1)}$ is some $U_kT_k\chi_\beta$, it follows from (7) that

$$\sum_{j=1}^N\langle D_j\Psi,\Psi\rangle\leqslant\{(1+2\nu)\|\Delta\chi\|+1\}N.$$

Next, we estimate the repulsive potential $Y=\Sigma_{1\leqslant i<j\leqslant N}\langle|z_i-z_j|^{-1}\Psi,\Psi\rangle$.

Clearly $Y_{ij}=\langle|z_i-z_j|^{-1}\Psi,\Psi\rangle$ is independent of the indices $i<j$. Hence $Y=N(N-1)Y_{12}/2$. Now, if σ, $\sigma'\in S_N$ are such that $\sigma(j)\neq\sigma'(j)$ for some $j>2$, then $\langle|z_1-z_2|^{-1}\varphi_{\sigma(1)}(z_1)\cdots\varphi_{\sigma(N)}(z_N),\ \varphi_{\sigma'(1)}(z_1)\cdots\varphi_{\sigma'(N)}(z_N)\rangle=0$. That is,

$$Y_{12}=\frac{1}{N!}\sum_{\sigma\in S_N}\langle|z-w|^{-1}\varphi_{\sigma(1)}(z)\varphi_{\sigma(2)}(w),\{\varphi_{\sigma(1)}(z)\varphi_{\sigma(2)}(w)-\varphi_{\sigma(2)}(z)\varphi_{\sigma(1)}(w)\}\rangle.$$

Since $\varphi_1,\ldots,\varphi_N$ are an enumeration of $\{\psi_k:k\in\Omega\}$, we can write $Y_{12}=(Z-Z')/N(N-1)$ with

$$Z=\sum_{\substack{k,k'\in\Omega\\k\neq k'}}\langle|z-w|^{-1}\psi_k(z)\psi_{k'}(w),\psi_k(z)\psi_{k'}(w)\rangle,$$

$$Z'=\sum_{\substack{k,k'\in\Omega\\k\neq k'}}\langle|z-w|^{-1}\psi_k(z)\psi_{k'}(w),\psi_{k'}(z)\psi_k(w)\rangle.$$

Now, if $k\neq k'$, then $\psi_k(z)\psi_{k'}(w)\overline{\psi_{k'}(z)\psi_k(w)}=0$ because the supports of ψ_k and $\psi_{k'}$ are disjoint. Therefore $Z'=0$. That is, $Y=N(N-1)Y_{12}/2=Z/2$. Thus, recalling that $\epsilon\in(0,\nu)$ is arbitrary, the theorem will follow once we establish

$$Z \leqslant \frac{16}{3\pi} \cdot \frac{(1+17N^{-1/2})^2}{1-4(N^{1/6}-4)^{-1}} \sqrt{\nu+\epsilon} N^{3/2} + O(N^{4/3}). \tag{8}$$

To prove (8), let $P=\{(k,k'):k,k' \in \Omega, k \neq k', |k-k'| > \alpha N^{1/6}\}$ and $L=\{(k,k'):k,k' \in \Omega, k \neq k', |k-k'| \leqslant \alpha N^{1/6}\}$. Then $Z=U+V$, where

$$U=\sum_{(k,k') \in P} \langle |z-w|^{-1} \psi_k(z)\psi_{k'}(w), \psi_k(z)\psi_{k'}(w)\rangle,$$

$$V=\sum_{(k,k') \in L} \langle |z-w|^{-1} \psi_k(z)\psi_{k'}(w), \psi_k(z)\psi_{k'}(w)\rangle.$$

For each $k \in \Omega$, if $n(k)$ is the number of k''s in Ω such that $|k-k'| \leqslant \alpha N^{1/6}$, then $n(k)\alpha^2 \leqslant \pi(2\alpha N^{1/6})^2$. That is, $n(k) \leqslant 4\pi N^{1/3}$. Thus, $\mathrm{Card}(L) \leqslant 4\pi N^{4/3}$. Now every term in V is bounded by $(3/\alpha)\cdot\beta^4\cdot(\|\chi\|_\infty/\beta)^4$. Therefore V is accounted for by the $O(N^{4/3})$ term in (8).

For each $k=(\alpha i(k)+(\alpha/2), \alpha j(k)+(\alpha/2)) \in \Omega$, where $i(k)$, $j(k) \in \mathbf{Z}$, let $\Sigma(k)$ be the $\alpha \times \alpha$ square whose vertices are $(\alpha(i(k)+a),\ \alpha(j(k)+b))$, a, $b \in \{0,1\}$. Of course, $\Sigma(k)$ contains the support of ψ_k. Suppose that $(k,k') \in P$ and (z,w), $(u,v) \in \Sigma(k)\times\Sigma(k')$. Then $|z-w| \geqslant |u-v|-4\alpha=(1-4\alpha|u-v|^{-1})|u-v| \geqslant (1-4(N^{1/6}-4)^{-1})|u-v|$. This shows that, if $(k,k') \in P$, then

$$\left\langle \frac{1}{|z-w|} \psi_k(z)\psi_{k'}(w), \psi_k(z)\psi_{k'}(w) \right\rangle \leqslant \frac{\alpha^{-4}}{1-4(N^{1/6}-4)^{-1}} \int_{\Sigma(k)} \int_{\Sigma(k')} |u-v|^{-1}\, dA(u)dA(v),$$

where dA is the area measure on $\mathbf{R}^2$. Because $\cup_{(k,k') \in P}\Sigma(k)\times\Sigma(k') \subset B_R(0)\times B_R(0)$,

$$U \leqslant \frac{\alpha^{-4}}{1-4(N^{1/6}-4)^{-1}} \int_{|z| \leqslant R} \int_{|w| \leqslant R} |z-w|^{-1}\, dA(z)dA(w) = \frac{R^3/\alpha^4}{1-4(N^{1/6}-4)^{-1}} \cdot J,$$

where

$$J=\int_{|z| \leqslant 1} \int_{|w| \leqslant 1} |z-w|^{-1}\, dA(z)dA(w).$$

It follows from (4) and (5) that $R^3/\alpha^4=(R/\alpha)^3(1/\alpha) \leqslant \pi^{-2}N^{3/2}(1+17N^{-1/2})^2\sqrt{\nu+\epsilon}$. Thus, the proof of (8) will be complete once we show that $J=16\pi/3$.

Define $f(w)=\int_{|z| \leqslant 1}|z-w|^{-1}\, dA(z)=\int_{|z+w| \leqslant 1}|z|^{-1}\, dA(z)$ for $|w| \leqslant 1$ and consider z and w also as complex variables. For $0 \leqslant r < 1$, the equation

$$\rho=-r\cos\theta+\sqrt{1-r^2\sin^2\theta}$$

describes the circle $\{z:|z+r|=1\}$ in the polar coordinates (θ,ρ). Thus

$$f(r)=\int_0^{2\pi}\int_0^{-r\cos\theta+\sqrt{1-r^2\sin^2\theta}} \frac{1}{\rho}\,\rho\, d\rho\, d\theta=4\int_0^{\pi/2}\sqrt{1-r^2\sin^2\theta}\,d\theta.$$

It is obvious that $f(w)=f(e^{it}w)$ if $t \in \mathbf{R}$. Therefore

$$J=\int_{|w| \leqslant 1} f(w)dA(w)=2\pi\int_0^1 f(r)r\, dr=8\pi\int_0^{\pi/2}\left(\int_0^1 r\sqrt{1-r^2\sin^2\theta}\,dr\right)d\theta.$$

Evaluate the double integral in the order indicated, and we obtain $J=16\pi/3$. This completes the proof of the theorem.

ACKNOWLEDGMENT

This work was supported in part by National Science Foundation, Grant No. DMS-9703515.

[1] R. Prange and S. Girvin, in *The Quantum Hall Effect*, Graduate Texts in Physics (Springer-Verlag, New York, 1990).
[2] R. Laughlin, Phys. Rev. Lett. **50**, 1395 (1983).
[3] T. Kato, *Perturbation Theory for Linear Operators* (Springer-Verlag, New York, 1976).
[4] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, II, Fourier Analysis, Self-Adjointness* (Academic, New York, 1975).
[5] J. Bellissard, A. van Elst, and H. Schulz-Baldes, J. Math. Phys. **35**, 5373 (1994).

Rigorous application of the stochastic functional method to plane-wave scattering from a random cylindrical surface

Nikolaos C. Skaropoulos and Dimitrios P. Chrissoulidis[a]
Department of Electrical and Computer Engineering, Faculty of Technology, Aristotle University of Thessaloniki, GR-54006 Thessaloniki, Greece

(Received 16 March 1998; accepted for publication 2 October 1998)

The stochastic functional method is applied to plane-wave scattering from a random cylindrical surface, whereupon the Dirichlet boundary condition is rigorously imposed. Analytical results, accurate to second and fourth order in surface roughness, are obtained for the coefficients of the Wiener–Hermite expansion of the secondary scattered wave field. The validity of approximate solutions is numerically investigated by means of the boundary condition criterion and of the energy consistency criterion. The former, which is introduced herein, states that any approximate solution should be in conformity with the boundary condition, whereas the latter pertains to the energy conservation law. The numerical investigation indicates that the rigorous application of the stochastic functional method yields more accurate results in terms of both criteria than did previous treatments of the problem under consideration. Moreover, it is suggested that applicability limits should be set through the mean boundary condition criterion instead of the energy consistency criterion; the latter may lead to underestimating deficiencies of the approximate solution under test.
[S0022-2488(99)02401-9]

I. INTRODUCTION

Several studies of wave scattering from random surfaces were lately based on the stochastic functional method,[1–6] which may be summarized as follows. The scattering surface is represented by a homogeneous, Gaussian random field; the scattered wave field is regarded as a nonlinear stochastic functional of surface roughness and it is expanded into a series of orthogonal Wiener–Hermite (WH) functionals by application of the Cameron–Martin theorem[7] and extensions thereof.[8,9] The expansion coefficients are determined through an approximation of the boundary condition that is imposed upon the random surface. By use of the orthogonality properties of the WH functionals, the approximate boundary condition is transformed into a set of linear deterministic equations that yields the expansion coefficients. Algebraic manipulations are greatly facilitated via a group-theoretic consideration of the stochastic homogeneity of surface roughness, which accounts for the symmetries of the scattered wave field.[5]

Previous applications of the stochastic functional method include studies of wave scattering from a random planar surface,[1–3] as well as from a random cylindrical[4] or spherical[5] surface; a detailed list of pertinent references can be found in Ref. 6. Thus far, the expansion coefficients have been determined through a first-order approximation of the boundary condition and, as a rule, the accuracy of the resulting solution has been evaluated through tests of energy consistency.

In the present paper, it is shown that the expansion coefficients can be more accurately determined, should the boundary condition be rigorously imposed on the random surface. The application of the stochastic functional method presented herein makes use of the complete Taylor series expansion of the boundary condition; the truncation of that series is deferred until the

[a]Electronic mail: dpchriss@vergina.eng.auth.gr

formulation of a hierarchical set of linear equations for the expansion coefficients. The resulting second- and fourth-order approximations of that set of equations include terms that would have been unnecessarily discarded otherwise and they yield more accurate expressions for the expansion coefficients.

Moreover, it is argued that energy conservation may not always be an adequate criterion for the validity of approximate solutions under test. Apart from energy consistency, conformity with the boundary condition on the random surface should be examined. To this end, a boundary condition test, which consists in numerically investigating whether the mean field satisfies the boundary condition on the random surface, is introduced herein; conformity of higher-order statistical moments of the field with the boundary condition could also be investigated.

The above are applied to the two-dimensional problem of plane-wave scattering from a random cylindrical surface, whereupon the Dirichlet boundary condition is imposed; other scattering surfaces and boundary conditions could also be treated, yet more complicated cases are deferred to future work. The numerical results indicate that the approximate solutions obtained by use of the rigorous application of the stochastic functional method are more accurate than those of a previous treatment of the problem in hand,[4] both in terms of the boundary condition criterion and of the energy consistency criterion. Furthermore, it is seen that the limits of validity should be set by means of the boundary condition criterion instead of the energy consistency criterion; the latter may lead to underestimating the inherent error of the approximate solution under test.

The paper is organized as follows: in Sec. II we deal with the analytical formulation of the stochastic functional method, second- and fourth-order approximations of the expansion coefficients are obtained in Sec. III, expressions for the characteristics of farfield scatter are determined in Sec. IV, and a numerical application is presented in Sec. V.

II. ANALYTICAL FORMULATION

A. Homogeneous random circular surface

A two-dimensional (2-D) random rough surface S, circular in the mean and centered at the origin O of a cylindrical coordinate system $(O;r\theta)$ (Fig. 1), is considered. S may be described by the equation $r_s=a+f(\theta,\omega)$, where a stands for the mean radius and $f(\theta,\omega)$ is a zero-mean random field on the circle S_2; the parameter $\omega\in\Omega$, denoting a sample point in the space Ω, will be suppressed from this point onward. Assuming that $f(\theta)$ is a homogeneous Gaussian random field, the following spectral representation can be written:[4]

$$f(\theta)=\sum_{n=-\infty}^{\infty} e^{in\theta}F_nB_n; \tag{1}$$

$\{B_n,n\in Z\}$ is a set of zero-mean, mutually orthogonal, Gaussian random variables; orthogonality suggests that $\langle\overline{B_m}B_n\rangle=\delta_{m,n}$, the overbar implying the complex conjugate; $\delta_{m,n}$ is the Kronecker delta. Since $f(\theta)$ is real valued, $B_{-n}=\overline{B_n}$ and $F_{-n}=\overline{F_n}$. As can be verified from the corresponding spectral representation of the correlation function,

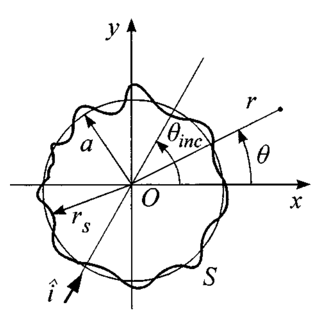

FIG. 1. Geometric configuration.

$$R(\theta_1,\theta_2)=\langle f(\theta_1)f(\theta_2)\rangle=\sum_{n=-\infty}^{\infty}|F_n|^2 e^{in(\theta_1-\theta_2)}, \tag{2}$$

the coefficients F_n of (1), actually the squared magnitude thereof, constitute the power spectrum of $f(\theta)$. Accordingly, the variance σ^2 of the roughness of the random surface S can be expressed as follows:

$$\sigma^2=\langle f^2(\theta)\rangle=\sum_{n=-\infty}^{\infty}|F_n|^2. \tag{3}$$

B. Wave field expressions

A random wave field $\psi(kr,\theta)$ outside S (i.e., for $r\geqslant r_s$) is a solution to the 2-D Helmholtz equation,

$$(\nabla^2+k^2)\psi(kr,\theta)=0, \tag{4}$$

where k stands for the wave number. By use of the shorthand notation $\rho=kr$, a boundary condition of the Dirichlet type on S is expressed as follows:

$$\psi(\rho,\theta)|_{\rho=\rho_s=kr_s}=0. \tag{5}$$

The wave field $\psi=\psi^{\text{inc}}+\psi^{\text{sca}}$ consists of two terms, which represent the incident and the scattered wave field, respectively. If excitation by a plane wave, propagated along the direction $\hat{i}=\hat{x}\cos\theta_{\text{inc}}+\hat{y}\sin\theta_{\text{inc}}$ (Fig. 1), is considered, ψ^{inc} can be written as follows:

$$\psi^{\text{inc}}(\rho,\theta)=\sum_{m=-\infty}^{\infty} i^m \psi_m^{\text{inc}}(\rho,\theta)e^{-im\theta_{\text{inc}}}, \tag{6}$$

where $\psi_m^{\text{inc}}(\rho,\theta)=J_m(\rho)e^{im\theta}$ is the mth-order cylindrical wave function and $J_m(\cdot)$ is the mth-order cylindrical Bessel function of the first kind.[10] A similar expression can be written for the scattered wave field:

$$\psi^{\text{sca}}(\rho,\theta)=\sum_{m=-\infty}^{\infty} i^m \psi_m^{\text{sca}}(\rho,\theta)e^{-im\theta_{\text{inc}}}; \tag{7}$$

(6) and (7) imply that ψ_m^{inc} gives rise to ψ_m^{sca}. Such simple coupling between excitation and scatter results from the statistical homogeneity of the random surface S; a proof, based on group-theoretic considerations, may be found in Ref. 4. The mth-order term of the scattered wave field, $\psi_m^{\text{sca}}={}_{(p)}\psi_m^{\text{sca}}+{}_{(s)}\psi_m^{\text{sca}}$, consists of two terms, which are referred to as the primary and the secondary wave. On the one hand, the primary wave ${}_{(p)}\psi_m^{\text{sca}}$ corresponds to unperturbed scatter from a smooth cylindrical surface of radius a, and it is expressed as follows:

$${}_{(p)}\psi_m^{\text{sca}}(\rho,\theta)=a_m^0 H_m^{(1)}(\rho)e^{im\theta}, \tag{8}$$

$a_m^0=-J_m(\rho_0)/H_m^{(1)}(\rho_0)$, $\rho_0=ka$, and $H_m^{(1)}(\cdot)$ is the mth-order cylindrical Hankel function of the first kind.[10] On the other hand, the secondary wave ${}_{(s)}\psi_m^{\text{sca}}$ incorporates the effect of surface roughness, and it can be regarded as a nonlinear functional of the random surface S. By use of the Cameron–Martin theorem,[7–9] ${}_{(s)}\psi_m^{\text{sca}}$ can be expanded in a series of complex, orthogonal, n-variate, Wiener–Hermite (WH) functionals $\hat{H}_n$ of the complex Gaussian random sequence B_n,[4,11]

$$ {}_{(s)}\psi_m^{\text{sca}}(\rho,\theta)=\sum_{n=0}^{\infty}\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}\cdots\sum_{j_n=-\infty}^{\infty}A_m^n(j_1,j_2,\ldots,j_n)H^{(1)}_{m+j_1+j_2+\cdots+j_n}(\rho) \times e^{i(m+j_1+j_2+\cdots+j_n)\theta}\hat{H}_n(B_{j_1},B_{j_2},\ldots,B_{j_n}), \tag{9} $$

where $\hat{H}_n$ are defined by

$$ \hat{H}_0=1,\quad \hat{H}_1(B_{j_1})=B_{j_1},\quad \hat{H}_2(B_{j_1},B_{j_2})=B_{j_1}B_{j_2}-\delta_{-j_1,j_2}, \tag{10} $$

and the recurrence relation:

$$ \hat{H}_{n+1}(B_{j_1},B_{j_2},\ldots,B_{j_{n+1}})=\hat{H}_1(B_{j_1})\hat{H}_n(B_{j_2},\ldots,B_{j_{n+1}}) -\sum_{k=2}^{n}\hat{H}_{n-1}(B_{j_2},\ldots,B_{j_{k-1}},B_{j_{k+1}},\ldots,B_{j_n})\,\delta_{-j_1,j_k}. \tag{11} $$

The cylindrical wave function $H^{(1)}_{m+j_1+j_2+\cdots+j_n}(\rho)e^{i(m+j_1+j_2+\cdots+j_n)\theta}$ satisfies the 2-D Helmholtz equation as well as the radiation condition at infinity; hence, so does ${}_{(s)}\psi_m^{\text{sca}}$, as expanded in (9). A_m^n are deterministic n-variate coefficients, which are symmetric with respect to their arguments; these coefficients are determined in Sec. III by application of the Dirichlet boundary condition on S.

III. CALCULATION OF THE EXPANSION COEFFICIENTS

A. Taylor series expansion of boundary condition

The first step toward the determination of A_m^n is taken by expanding the left-hand side of (5) into a Taylor series around the point $\rho=\rho_0$:

$$ \sum_{p=0}^{\infty}\frac{1}{p!}\,k^p f^p(\theta)\,\psi^{(p)}(\rho,\theta)\big|_{\rho=\rho_0}=0\,. \tag{12} $$

The abbreviation $\psi^{(p)}=\partial^p\psi/\partial\rho^p$ has been used in (12). By use of (6)–(9), (12) yields

$$ \sum_{p=1}^{\infty}\frac{1}{p!}\,k^p W_m{}^{(p)}(\rho_0)\sum_{n_1=-\infty}^{\infty}\sum_{n_2=-\infty}^{\infty}\cdots\sum_{n_p=-\infty}^{\infty}e^{i(n_1+n_2+\cdots+n_p)\theta}\prod_{i=1}^{p}F_{n_i}\hat{H}_1(B_{n_i}) + \sum_{n=0}^{\infty}\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}\cdots\sum_{j_n=-\infty}^{\infty}A_m^n(j_1,j_2,\ldots,j_n)e^{i(j_1+j_2+\cdots+j_n)\theta}\hat{H}_n(B_{j_1},B_{j_2},\ldots,B_{j_n}) \times\Bigg[H^{(1)}_{m+j_1+j_2+\cdots+j_n}(\rho_0)+\sum_{p=1}^{\infty}\frac{1}{p!}\,k^p(H^{(1)}_{m+j_1+j_2+\cdots+j_n})^{(p)}(\rho_0) \times\sum_{n_1=-\infty}^{\infty}\sum_{n_2=-\infty}^{\infty}\cdots\sum_{n_p=-\infty}^{\infty}e^{i(n_1+n_2+\cdots+n_p)\theta}\prod_{i=1}^{p}F_{n_i}\hat{H}_1(B_{n_i})\Bigg]=0, \tag{13} $$

where $W_m=J_m+a_m^0H_m^{(1)}$. It should be pointed out that in the present paper the complete Taylor series expansion of the boundary condition is used instead of the first-order approximation thereof, which was used in the previous treatment of this problem[4] as well as in other applications of the stochastic functional method.[1–3,5,6] The truncation of the infinite Taylor series is postponed until the formulation of a hierarchical set of equations for A_m^n. As shown in Sec. V, this approach results in more accurate expressions for the expansion coefficients.

B. Hierarchy of equations

Linear equations for A_m^n can be obtained from (13) by multiplying both sides by $\overline{\hat{H}_l(B_{k_1},B_{k_2},\ldots,B_{k_l})}$, where $l=0,1,\ldots$, and averaging over Ω. Use is then made of the following formula for the expectation value of a product of WH functionals:[11]

$$\langle \hat{H}_{n_1}(B_{j_1},\ldots,B_{j_{n_1}})\hat{H}_{n_2}(B_{j_{n_1+1}},\ldots,B_{j_{n_1+n_2}})\cdots\overline{\hat{H}_{n_m}(B_{j_{n_1+\ldots+n_{m-1}+1}},\ldots,B_{j_{n_1+\ldots+n_m}})}\rangle$$
$$=\begin{cases}\displaystyle\sum_{\substack{\text{distinct}\\ \text{exogamous}\\ \text{pairs}}}\prod_{\substack{\text{exogamous}\\ \text{pairs}}}\delta_{-j_k,j_l}, & \text{if } (n_1+n_2+\cdots+n_m) \text{ is } \begin{cases}\text{even,}\\ \text{odd.}\end{cases}\\ 0 \end{cases} \tag{14}$$

Each one of δ_{-j_k,j_l} on the right-hand side of (14) has an exogamous pair of indices, which means that j_k,j_l do not originate from the same $\hat{H}$ function; if either of j_k,j_l originates from an $\overline{\hat{H}}$ function, δ_{j_k,j_l}, instead of δ_{-j_k,j_l}, should be used. The indices $j_1,j_2,\ldots,j_{n_1+\ldots+n_m}$ are used once in any product of Kronecker deltas, and the summation is taken over all possible arrangements of all indices into distinct exogamous pairs. According to these constraints, $\langle \hat{H}_1(B_{j_1})\overline{\hat{H}_1(B_{j_2})}\hat{H}_2(B_{j_3},B_{j_4})\rangle$ is equal to $\delta_{-j_1,j_3}\delta_{-j_2,j_4}+\delta_{-j_1,j_4}\delta_{-j_2,j_3}$, whereas $\langle \hat{H}_1(B_{j_1})\hat{H}_1(B_{j_2})\hat{H}_2(B_{j_3},B_{j_4})\rangle$ equals $\delta_{j_1,j_3}\delta_{j_2,j_4}+\delta_{j_1,j_4}\delta_{j_2,j_3}$.

C. Second-order approximation

Approximate expressions for A_m^n, accurate to $O((k\sigma)^2)$, are determined next. Since, as suggested by (3), $kF_j=O(k\sigma)$, the triad of equations obtained for $l=0,\ 1,\ 2$ is the following:

$$\frac{1}{2}(k\sigma)^2W_m^{(2)}(\rho_0)+A_m^0H_m^{(1)}(\rho_0)+\sum_{j_1=-\infty}^{\infty}k\overline{F_{j_1}}A_m^1(j_1)(H_{m+j_1}^{(1)})^{(1)}(\rho_0)=0, \tag{15a}$$

$$kF_{k_1}W_m^{(1)}(\rho_0)+A_m^1(k_1)H_{m+k_1}^{(1)}(\rho_0)=0, \tag{15b}$$

$$k^2F_{k_1}F_{k_2}W_m^{(2)}(\rho_0)+kF_{k_2}A_m^1(k_1)(H_{m+k_1}^{(1)})^{(1)}(\rho_0)+kF_{k_1}A_m^1(k_2)(H_{m+k_2}^{(1)})^{(1)}(\rho_0)$$
$$+2A_m^2(k_1,k_2)H_{m+k_1+k_2}^{(1)}(\rho_0)=0. \tag{15c}$$

The second-order approximations to A_m^n are readily obtained from (15),

$$A_m^0=-(k\sigma)^2\frac{W_m^{(2)}(\rho_0)}{2H_m^{(1)}(\rho_0)}+\frac{W_m^{(1)}(\rho_0)}{H_m^{(1)}(\rho_0)}\sum_{j_1=-\infty}^{\infty}|kF_{j_1}|^2\frac{(H_{m+j_1}^{(1)})^{(1)}(\rho_0)}{H_{m+j_1}^{(1)}(\rho_0)}, \tag{16a}$$

$$A_m^1(j_1)=-kF_{j_1}\frac{W_m^{(1)}(\rho_0)}{H_{m+j_1}^{(1)}(\rho_0)}, \tag{16b}$$

$$A_m^2(j_1,j_2)=-k^2F_{j_1}F_{j_2}$$
$$\times\frac{\left(W_m^2(\rho_0)-\frac{(H_{m+j_1}^{(1)})^{(1)}(\rho_0)}{H_{m+j_1}^{(1)}(\rho_0)}W_m^{(1)}(\rho_0)-\frac{(H_{m+j_2}^{(1)})^{(1)}(\rho_0)}{H_{m+j_2}^{(1)}(\rho_0)}W_m^{(1)}(\rho_0)\right)}{2H_{m+j_1+j_2}^{(1)}(\rho_0)}, \tag{16c}$$

and, evidently, $A_m^n = 0$, if $n \geqslant 3$. The expressions of (16) are identical to those obtained through the single-scattering approximation of the previous treatment of this problem,[4] apart from the first term on the right-hand side of (16a), which does not appear in the corresponding result of Ref. 4; this is due to the first-order approximation of the boundary condition used therein. Furthermore, it should be noted that Eqs. (15) were obtained with the *a priori* assumptions $A_m^0 = O((k\sigma)^2)$ and $A_m^n = O((k\sigma)^n)$, $n \geqslant 1$, which are *a posteriori* justified by the results of (16).

D. Fourth-order approximation

A similar procedure yields fourth-order approximations to A_m^n. The tetrad of equations obtained for $l = 0, 1, 2, 3$ is given below:

$$\begin{aligned}
&\tfrac{1}{2}(k\sigma)^2 W_m^{(2)}(\rho_0) + \tfrac{1}{8}(k\sigma)^4 W_m^{(4)}(\rho_0) + A_m^0[H_m^{(1)}(\rho_0) + \tfrac{1}{2}(k\sigma)^2 (H_m^{(1)})^{(2)}(\rho_0)]\\
&\quad + \sum_{j_1=-\infty}^{\infty} k\overline{F_{j_1}} A_m^1(j_1)((H_{m+j_1}^{(1)})^{(1)}(\rho_0) + \tfrac{1}{2}(k\sigma)^2 (H_{m+j_1}^{(1)})^{(3)}(\rho_0))\\
&\quad + \sum_{j_1=-\infty}^{\infty} \sum_{j_2=-\infty}^{\infty} k^2 \overline{F_{j_1} F_{j_2}} A_m^2(j_1, j_2)(H_{m+j_1+j_2}^{(1)})^{(2)}(\rho_0) = 0,
\end{aligned} \tag{17a}$$

$$\begin{aligned}
&kF_{k_1}[W_m^{(1)}(\rho_0) + \tfrac{1}{2}(k\sigma)^2 W_m^{(3)}(\rho_0)] + kF_{k_1} A_m^0 H_m^{(1)}(\rho_0) + A_m^1(k_1)[H_{m+k_1}^{(1)}(\rho_0)\\
&\quad + \tfrac{1}{2}(k\sigma)^2 (H_{m+k_1}^{(1)})^{(2)}(\rho_0)] + \sum_{j_1=-\infty}^{\infty} k\overline{F_{j_1}}[kF_{k_1} A_m^1(j_1)(H_{m+j_1}^{(1)})^{(2)}(\rho_0)\\
&\quad + 2A_m^2(k_1, j_1)(H_{m+k_1+j_1}^{(1)})^{(1)}(\rho_0)] = 0,
\end{aligned} \tag{17b}$$

$$\begin{aligned}
&k^2 F_{k_1} F_{k_2}[W_m^{(2)}(\rho_0) + \tfrac{1}{2}(k\sigma)^2 W_m^{(4)}(\rho_0) + A_m^0 (H_m^{(1)})^{(2)}(\rho_0)] + kF_{k_2} A_m^1(k_1)[(H_{m+k_1}^{(1)})^{(1)}(\rho_0)\\
&\quad + \tfrac{1}{2}(k\sigma)^2 (H_{m+k_1}^{(1)})^{(3)}(\rho_0)] + kF_{k_1} A_m^1(k_2)[(H_{m+k_2}^{(1)})^{(1)}(\rho_0) + \tfrac{1}{2}(k\sigma)^2 (H_{m+k_2}^{(1)})^{(3)}(\rho_0)]\\
&\quad + \sum_{j_1=-\infty}^{\infty} k^3 F_{k_1} F_{k_2} \overline{F_{j_1}} A_m^1(j_1)(H_{m+j_1}^{(1)})^{(3)}(\rho_0) + 2A_m^2(k_1, k_2)[H_{m+k_1+k_2}^{(1)}(\rho_0)\\
&\quad + \tfrac{1}{2}(k\sigma)^2 (H_{m+k_1+k_2}^{(1)})^{(2)}(\rho_0)] + 2\sum_{j_1=-\infty}^{\infty} k^2 F_{k_2} \overline{F_{j_1}} A_m^2(k_1, j_1)(H_{m+k_1+j_1}^{(1)})^{(2)}(\rho_0)\\
&\quad + 2\sum_{j_1=-\infty}^{\infty} k^2 F_{k_1} \overline{F_{j_1}} A_m^2(k_2, j_1)(H_{m+k_2+j_1}^{(1)})^{(2)}(\rho_0) + 6\sum_{j_1=-\infty}^{\infty} k\overline{F_{j_1}} A_m^3(k_1, k_2, j_1)\\
&\quad \times (H_{m+k_1+k_2+j_1}^{(1)})^{(1)}(\rho_0) = 0,
\end{aligned} \tag{17c}$$

$$\begin{aligned}
&k^3 F_{k_1} F_{k_2} F_{k_3} W_m^{(3)}(\rho_0) + k^2 F_{k_2} F_{k_3} A_m^1(k_1)(H_{m+k_1}^{(1)})^{(2)}(\rho_0) + k^2 F_{k_1} F_{k_3} A_m^1(k_2)(H_{m+k_2}^{(1)})^{(2)}(\rho_0)\\
&\quad + k^2 F_{k_1} F_{k_2} A_m^1(k_3)(H_{m+k_3}^{(1)})^{(2)}(\rho_0) + 2kF_{k_1} A_m^2(k_2, k_3)(H_{m+k_2+k_3}^{(1)})^{(1)}(\rho_0) + 2kF_{k_2} A_m^2(k_1, k_3)\\
&\quad \times (H_{m+k_1+k_3}^{(1)})^{(1)}(\rho_0) + 2kF_{k_3} A_m^2(k_1, k_2)(H_{m+k_1+k_2}^{(1)})^{(1)}(\rho_0) + 6A_m^3(k_1, k_2, k_3)\\
&\quad \times H_{m+k_1+k_2+k_3}^{(1)}(\rho_0) = 0.
\end{aligned} \tag{17d}$$

Elimination of A_m^0 and A_m^3 from (17a) and (17d), respectively, yields an infinite set of linear equations for A_m^1 and A_m^2:

$$\sum_{j_1=-\infty}^{\infty}\left[C_1(j_1;m,k_1)A_m^1(j_1)+\sum_{j_2=-\infty}^{\infty}C_2(j_1,j_2;m,k_1)A_m^2(j_1,j_2)\right]=B_1(m,k_1), \tag{18a}$$

$$\sum_{j_1=-\infty}^{\infty}\left[C_3(j_1;m,k_1,k_2)A_m^1(j_1)+\sum_{j_2=-\infty}^{\infty}C_4(j_1,j_2;m,k_1,k_2)A_m^2(j_1,j_2)\right]=B_2(m,k_1,k_2). \tag{18b}$$

The following abbreviations are used in (18):

$$\begin{aligned}C_1(j_1;m,k_1)=&\delta_{j_1,k_1}[H_{m+k_1}^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_{m+k_1}^{(1)})^{(2)}(\rho_0)][H_m^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)]\\&-k^2F_{k_1}\overline{F_{j_1}}[(H_{m+j_1}^{(1)})^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_{m+j_1}^{(1)})^{(3)}(\rho_0)](H_m^{(1)})^{(1)}(\rho_0)\\&+k^2F_{k_1}\overline{F_{j_1}}[H_m^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)](H_{m+j_1}^{(1)})^{(2)}(\rho_0),\end{aligned} \tag{19a}$$

$$\begin{aligned}C_2(j_1,j_2;m,k_1)=&2\delta_{j_2,k_1}k\overline{F_{j_1}}(H_{m+j_1+j_2}^{(1)})^{(1)}(\rho_0)[H_m^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)]\\&-k^3F_{k_1}\overline{F_{j_1}F_{j_2}}(H_{m+j_1+j_2}^{(1)})^{(2)}(\rho_0)(H_m^{(1)})^{(1)}(\rho_0),\end{aligned} \tag{19b}$$

$$\begin{aligned}C_3(j_1;m,k_1,k_2)=&k(F_{k_2}\delta_{j_1,k_1}+F_{k_1}\delta_{j_1,k_2})[H_m^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)]\\&\times\left[(H_{m+j_1}^{(1)})^{(1)}(\rho_0)-(H_{m+j_1}^{(1)})^{(2)}(\rho_0)\sum_{j=-\infty}^{\infty}|kF_j|^2\frac{(H_{m+k_1+k_2+j}^{(1)})^{(1)}(\rho_0)}{H_{m+k_1+k_2+j}^{(1)}(\rho_0)}\right.\\&\left.+\frac{1}{2}(k\sigma)^2(H_{m+j_1}^{(1)})^{(3)}(\rho_0)\right]-k^3F_{k_1}F_{k_2}\overline{F_{j_1}}(H_m^{(1)})^{(2)}(\rho_0)\\&\times[(H_{m+j_1}^{(1)})^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_{m+j_1}^{(1)})^{(3)}(\rho_0)]+k^3F_{k_1}F_{k_2}\overline{F_{j_1}}[H_m^{(1)}(\rho_0)\\&+\tfrac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)]\\&\times\left[(H_{m+j_1}^{(1)})^{(3)}(\rho_0)-(H_{m+j_1}^{(1)})^{(2)}(\rho_0)\frac{(H_{m+k_1+k_2+j_1}^{(1)})^{(1)}(\rho_0)}{H_{m+k_1+k_2+j_1}^{(1)}(\rho_0)}\right],\end{aligned} \tag{19c}$$

$$\begin{aligned}&C_4(j_1,j_2;m,k_1,k_2)\\&=2\delta_{j_1,k_1}\delta_{j_2,k_2}[H_m^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)]\\&\times\left[H_{m+k_1+k_2}^{(1)}(\rho_0)-(H_{m+k_1+k_2}^{(1)})^{(1)}(\rho_0)\sum_{j=-\infty}^{\infty}|kF_j|^2\frac{(H_{m+k_1+k_2+j}^{(1)})^{(1)}(\rho_0)}{H_{m+k_1+k_2+j}^{(1)}(\rho_0)}\right.\\&\left.+\frac{1}{2}(k\sigma)^2(H_{m+k_1+k_2}^{(1)})^{(2)}(\rho_0)\right]+2k^2\overline{F_{j_1}}(F_{k_2}\delta_{j_2,k_1}+F_{k_1}\delta_{j_2,k_2})\\&\times\left[H_m^{(1)}(\rho_0)+\frac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)\right]\\&\times\left[(H_{m+j_1+j_2}^{(1)})^{(2)}(\rho_0)-(H_{m+j_1+j_2}^{(1)})^{(1)}(\rho_0)\frac{(H_{m+k_1+k_2+j_1}^{(1)})^{(1)}(\rho_0)}{H_{m+k_1+k_2+j_1}^{(1)}(\rho_0)}\right]\\&-k^4F_{k_1}F_{k_2}\overline{F_{j_1}F_{j_2}}(H_m^{(1)})^{(2)}(\rho_0)(H_{m+j_1+j_2}^{(1)})^{(2)}(\rho_0),\end{aligned} \tag{19d}$$

$$B_1(m,k_1)=-kF_{k_1}[W_m^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2W_m^{(3)}(\rho_0)][H_m^{(1)}(\rho_0)+\tfrac{1}{2}(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)]$$
$$+\tfrac{1}{2}(k\sigma)^2kF_{k_1}(H_m^{(1)})^{(1)}(\rho_0)[W_m^{(2)}(\rho_0)+\tfrac{1}{4}(k\sigma)^2W_m^{(4)}(\rho_0)], \tag{20a}$$

$$B_2(m,k_1,k_2)=-k^2F_{k_1}F_{k_2}[H_m^{(1)}(\rho_0)+\tfrac{1}{2}\,(k\sigma)^2(H_m^{(1)})^{(2)}(\rho_0)]$$
$$\times\left[W_m^{(2)}(\rho_0)-W_m^{(3)}(\rho_0)\sum_{j=-\infty}^{\infty}|kF_j|^2\,\frac{(H^{(1)}_{m+k_1+k_2+j})^{(1)}(\rho_0)}{H^{(1)}_{m+k_1+k_2+j}(\rho_0)}\right.$$
$$\left.+\frac{1}{2}\,(k\sigma)^2W_m^{(4)}(\rho_0)\right]+\tfrac{1}{2}(k\sigma)^2k^2F_{k_1}F_{k_2}(H_m^{(1)})^{(2)}(\rho_0)[W_m^{(2)}(\rho_0)$$
$$+\tfrac{1}{4}(k\sigma)^2W_m^{(4)}(\rho_0)]. \tag{20b}$$

The set of (18) may be solved by truncation of the infinite summations and numerical inversion. Let M be the appropriate truncation number; the choice of M is discussed in Sec. V. Since the coefficients $A_m^2(j_1,j_2)$ are symmetric with respect to j_1,j_2, there are $(2M+1)(M+2)$ unknowns in (18) with j_1 and j_2 varying in the range $[-M,M]$ and $[-M,j_1]$, respectively. A set of $(2M+1)(M+2)$ equations can be obtained by allowing k_1 to vary in the range $[-M,M]$ and k_2 to vary in the range $[-M,k_1]$. This set yields $A_m^1(j_1)$ and $A_m^2(j_1,j_2)$ for $j_1,j_2\in[-M,M]$.

IV. CHARACTERISTICS OF FARFIELD SCATTER

A. Scattering amplitude

The scattered wave field far from the origin of coordinates (i.e., for $\rho\gg1$), is usually expressed as a diverging cylindrical wave $\psi^{\mathrm{sca}}(\rho,\theta)=e^{i\rho}\Phi(\theta)/\sqrt{\rho}$. The direction-dependent factor,

$$\Phi(\theta)=\sqrt{\frac{2}{\pi i}}\sum_{m=-\infty}^{\infty}e^{im\theta}\left(a_m^0+\sum_{n=0}^{\infty}\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}\cdots\sum_{j_n=-\infty}^{\infty}A_m^n(j_1,j_2,\ldots,j_n)\right.$$
$$\left.\times\frac{e^{i(j_1+j_2+\cdots+j_n)\theta}}{i^{(j_1+j_2+\cdots+j_n)}}\,\hat{H}_n(B_{j_1},B_{j_2},\ldots,B_{j_n})\right) \tag{21}$$

is the scattering amplitude, which is given above for $\theta_{\mathrm{inc}}=0$; this assumption implies that incidence is along the x axis (Fig. 1), and it has no effect on generality whatsoever.

B. Coherent scattering

The coherent part of the scattering amplitude $\langle\Phi(\theta)\rangle$ can be obtained from (21) by averaging

$$\langle\Phi(\theta)\rangle=\sqrt{\frac{2}{\pi i}}\sum_{m=-\infty}^{\infty}e^{im\theta}(a_m^0+A_m^0). \tag{22}$$

The coherent differential scattering cross section is defined as $\sigma_c(\theta)=|\langle\Phi(\theta)\rangle|^2$ and, if integrated with respect to θ in the range $[0,\pi]$, it yields the coherent scattering cross section:

$$\sigma_c=4\sum_{m=-\infty}^{\infty}|a_m^0+A_m^0|^2. \tag{23}$$

C. Incoherent scattering

The incoherent differential scattering cross section is defined as $\sigma_{ic}(\theta)=\langle|\Phi(\theta)|^2\rangle-|\langle\Phi(\theta)\rangle|^2$ and, by use of (21), the following expression can be obtained:

$$\sigma_{ic}(\theta)=\frac{2}{\pi}\sum_{n=1}^{\infty} n!\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}\cdots\sum_{j_n=-\infty}^{\infty}\left|\sum_{m=-\infty}^{\infty} A_m^n(j_1,j_2,\ldots,j_n)e^{im\theta}\right|^2. \tag{24}$$

Integration of (24) with respect to θ in the range $[0,\pi]$ yields the incoherent scattering cross section:

$$\sigma_{ic}=4\sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty} n!\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}\cdots\sum_{j_n=-\infty}^{\infty}|A_m^n(j_1,j_2,\ldots,j_n)|^2. \tag{25}$$

V. NUMERICAL RESULTS

The numerical results are aimed at investigating the accuracy and applicability of the second- and fourth-order solutions obtained in Sec. III; these solutions are compared to those obtained in a previous treatment of the problem under consideration.[4] The validity of approximate solutions is tested by means of the following two criteria: first, that any approximate solution should be in conformity with the boundary condition and, second, that it should be energy consistent. All calculations make use of the following coefficients:

$$|F_n|^2=\sigma^2\left(e^{-K^2n^2/2}\bigg/\sum_{n=-\infty}^{\infty} e^{-K^2n^2/2}\right), \tag{26}$$

for the power spectrum of $f(\theta)$; K (rad) is the correlation distance of surface roughness. With regard to the infinite summations of the expressions for the characteristics of farfield scatter, three-digit convergence can be achieved in all cases with truncation number $M\leqslant 12$.

A. Boundary condition test

The boundary condition test is performed by means of the following equation, which is obtained from the Dirichlet boundary condition (5) by averaging and by use of (14):

$$\begin{aligned}\sum_{l=1}^{\infty}\frac{(2l-1)!!}{(2l)!}(k\sigma)^{2l}W_m^{(2l)}(\rho_0)=A_m^0&\left[H_m^{(1)}(\rho_0)+\sum_{l=1}^{\infty}\frac{(2l-1)!!}{(2l)!}(k\sigma)^{2l}(H_m^{(1)})^{(2l)}(\rho_0)\right]\\&+\sum_{j_1=-\infty}^{\infty}k\overline{F_{j_1}}A_m^1(j_1)\sum_{l=0}^{\infty}\frac{(2l+1)!!}{(2l+1)!}(k\sigma)^{2l}(H_{m+j_1}^{(1)})^{(2l+1)}(\rho_0)\\&+\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}k^2\overline{F_{j_1}F_{j_2}}A_m^2(j_1,j_2)\\&\times\sum_{l=1}^{\infty}\frac{(2l-1)!!}{(2l-1)!}(k\sigma)^{2l-2}(H_{m+j_1+j_2}^{(1)})^{(2l)}(\rho_0)\\&+\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}\sum_{j_3=-\infty}^{\infty}k^3\overline{F_{j_1}F_{j_2}F_{j_3}}A_m^3(j_1,j_2,j_3)\\&\times\sum_{l=1}^{\infty}\frac{(2l-1)!!}{(2l-1)!}(k\sigma)^{2l-2}(H_{m+j_1+j_2+j_3}^{(1)})^{(2l+1)}(\rho_0).\end{aligned} \tag{27}$$

The abbreviation $(2l-1)!!=1.3\cdots(2l-1)$ has been used in (27). The left-hand side of (27) represents the mean value of $\psi_m^{\text{inc}}+{}_{(p)}\psi_m^{\text{sca}}$ at $\rho=\rho_s$, whereas the right-hand side represents the mean value of $-{}_{(s)}\psi_m^{\text{sca}}$, also at $\rho=\rho_s$. The former can be calculated exactly, thus serving as a reference for approximate calculations of the latter, which involves the approximate expressions for A_m^n. Results for the modulus of both sides of (27) are shown in Fig. 2. It may be seen that curves 1 and 2, corresponding to approximate calculations of the right-hand side of (27) by use of

(16) and (17), respectively, follow closely the reference curve for the left-hand side as long as $k\sigma$ is small; the deviation of the approximate curves from the reference curve increases with $k\sigma$. The fourth-order approximation is evidently more accurate than the second-order approximation, roughly up to $k\sigma=0.1$; from that value onward the fourth-order solution deteriorates. Curves 3 and 4, corresponding to the previous treatment of this problem,[4] are quite far from the reference curve, which is probably due to the first-order approximation of the boundary condition that is used therein.

B. Energy consistency test

Energy consistency can be checked through the optical theorem,[4] which states that $\sigma_c+\sigma_{ic}=\sqrt{8\pi}\ \mathrm{Im}\{\langle\Phi(0)\rangle/\sqrt{i}\}$. The ratio of the left- to right-hand side of this equation may serve as a measure of energy consistency, any deviation from the value 1 revealing deficiencies of the approximate solution under test. The energy consistency test can be performed more effectively by discarding the dominant zeroth-order terms from both sides of the optical theorem by use of the identity $\Sigma_{m=-\infty}^{\infty}|a_m^0|^2=-\Sigma_{m=-\infty}^{\infty}\mathrm{Re}\{a_m^0\}$; thus, the following equations, which are accurate to $O((k\sigma)^2)$ and $O((k\sigma)^4)$ respectively, are obtained:

$$\sum_{m=-\infty}^{\infty}\left[|A_m^0|^2+2\ \mathrm{Re}\{\overline{a_m^0}A_m^0\}+\sum_{j_1=-\infty}^{\infty}|A_m^1(j_1)|^2\right]=-\sum_{m=-\infty}^{\infty}\mathrm{Re}\{A_m^0\}, \tag{28a}$$

$$\sum_{m=-\infty}^{\infty}\left[|A_m^0|^2+2\ \mathrm{Re}\{\overline{a_m^0}A_m^0\}+\sum_{j_1=-\infty}^{\infty}|A_m^1(j_1)|^2+4\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}|A_m^2(j_1,j_2)|^2\right]$$
$$=-\sum_{m=-\infty}^{\infty}\mathrm{Re}\{A_m^0\}. \tag{28b}$$

Numerical results for the ratio of the left- to right-hand side of (28a) or (28b) are displayed in Fig. 3. It is readily seen that curve 1 is constantly equal to 1 regardless of $k\sigma$; the second-order approximation is, therefore, energy consistent, which can also be verified analytically. The fourth-order approximation results in a nonzero energy error, which is, however, less than 1% in all

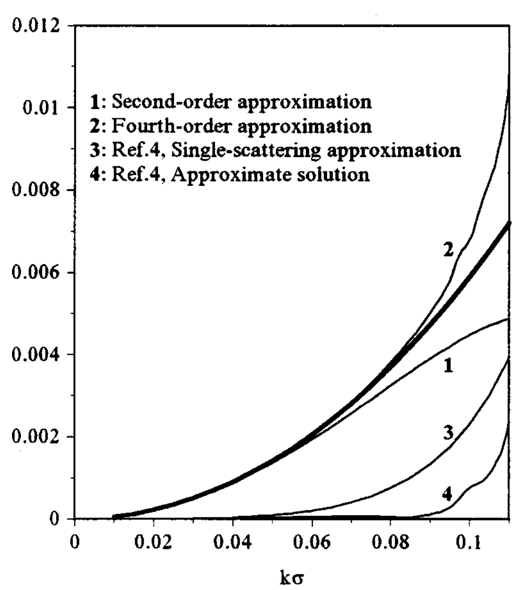

FIG. 2. Boundary condition test ($ka=1, K=0.6$ rad): the thick curve depicts the modulus of the left-hand side of (25) and serves as reference; curves 1–4 correspond to approximate calculations of the modulus of the right-hand side of (25).

cases. In general, all approximate solutions meet the energy consistency criterion more satisfactorily than they meet the boundary condition criterion. This is more evident for the approximate solutions of Ref. 4, which were seen in Fig. 2 to fail with regard to the latter criterion.

C. Applicability limit

It can be argued, after inspection of Figs. 2 and 3, that the boundary condition test is more sensitive than the energy conservation test to inherent errors of approximate solutions. It is, therefore, plausible to investigate the limit of applicability of approximate solutions through calculations of the relative mean boundary condition error, which is defined as the ratio of the left- minus the right- to left-hand side of (27). Numerical results for the relative mean boundary condition error are displayed in Fig. 4, the main conclusion being that it increases rapidly with $k\sigma$, especially for $K=0.4$ rad. The applicability limit, defined as the highest value of $k\sigma$ for the small mean boundary condition error, is well below the corresponding value that would have been estimated by use of the energy consistency criterion. This argument is most convincing when applied to the second-order approximate solution of this paper, which yields a substantial mean boundary condition error, even for small $k\sigma$, although, as mentioned above, it is energy consistent for any $k\sigma$.

D. Angular distribution of farfield scatter

Calculations of the angular distribution of farfield scatter are presented in Fig. 5. It can be observed that the effect of surface roughness is most pronounced in the backscattering direction (i.e., for $\theta=180°$), where it manifests itself as a reduction of coherent scatter; this is a well-known result from studies of light scattering from irregular particles.[12] Incoherent scatter is strongest in the backscattering direction, but it is roughly two orders of magnitude weaker than coherent scatter.

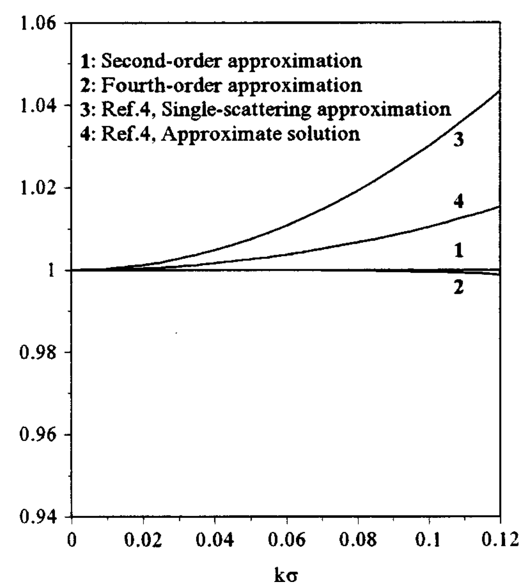

FIG. 3. Energy consistency test ($ka=1, K=0.6$ rad): plots of the ratio of the left- to right-hand side of (26), which should equal 1 for an energy consistent solution. Curve 1 corresponds to (26a); the expressions for $A_m^0, A_m^1(j_1)$ are those of (14). Curve 2 corresponds to (26b); $A_m^1(j_1), A_m^2(j_1, j_2)$ are determined from (16), whereas A_m^0 are obtained from (15a). Curves 3 and 4 correspond to (26a) and (26b), respectively, with A_m^n as expressed in Ref. 4.

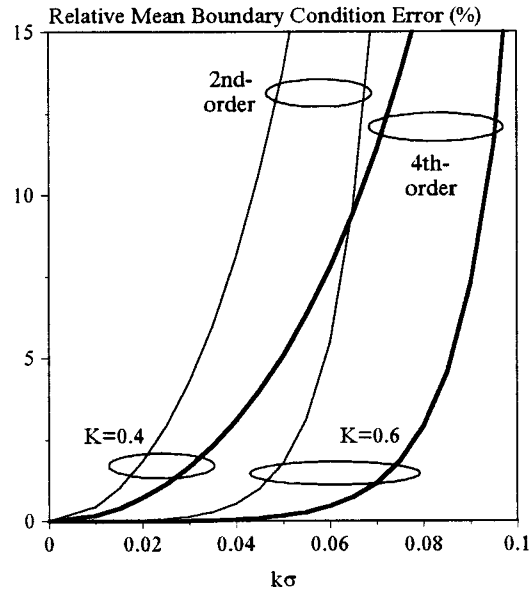

FIG. 4. Relative mean boundary condition error versus normalized variance $k\sigma$ ($ka=1$). This error is defined as the modulus of the ratio of the left- minus the right- to left-hand side of (25), and it should equal 0 for an accurate solution.

VI. CONCLUSIONS

The rigorous application of the boundary condition improves the results obtained by the stochastic functional method, both in terms of the boundary condition criterion and of the energy consistency criterion. The former places stricter restrictions to the applicability of approximate solutions and it is, therefore, more reliable than the latter. The second-order approximate solution obtained herein exhibits considerable boundary condition error and, yet, it is energy consistent,

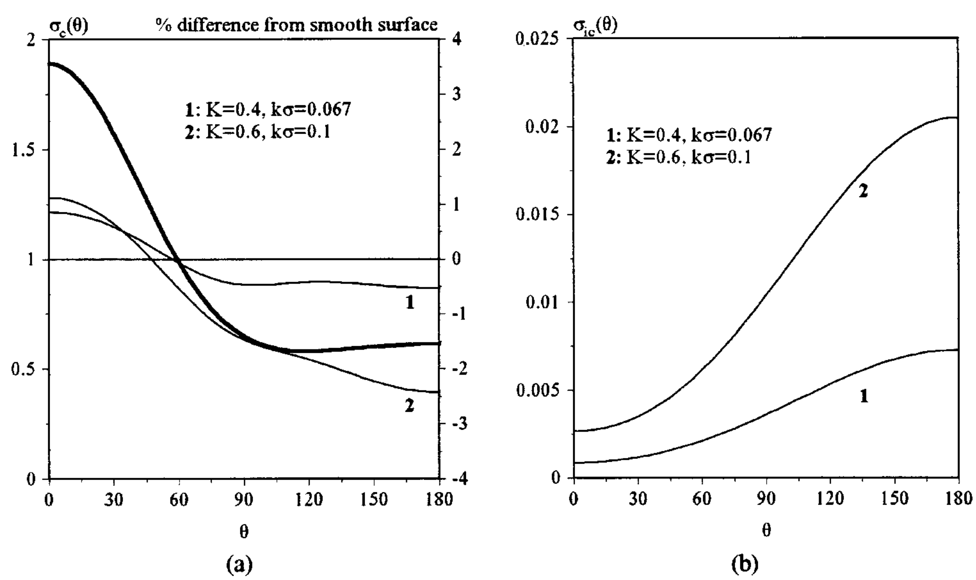

FIG. 5. Angular distribution of farfield scatter ($ka=1$). (a) Coherent differential scattering cross section versus scattering angle θ(°). The thick curve is associated with the left-hand vertical axis, and it corresponds to scattering from a smooth cylindrical surface ($k\sigma=0$); curves 1 and 2 are associated with the right-hand vertical axis and they display the percentage difference in $\sigma_c(\theta)$ of a random cylindrical surface from the corresponding smooth surface. (b) Incoherent differential scattering cross section.

which indicates the limitations of the energy consistency criterion as a means of validating approximate solutions. The fourth-order approximate solution exhibits negligible energy error and it is more accurate with respect to the boundary condition criterion than the second-order solution of this paper and the solutions obtained in Ref. 4.

[1] J. Nakayama, H. Ogura, and B. Matsumoto, Radio Sci. **15**, 1049 (1980).
[2] J. Nakayama, H. Ogura, and M. Sakata, J. Math. Phys. **22**, 471 (1981).
[3] J. Nakayama, H. Ogura, and M. Sakata, Radio Sci. **16**, 831, 847 (1981).
[4] H. Ogura and H. Nakayama, J. Math. Phys. **29**, 851 (1988).
[5] H. Ogura and N. Takahashi, J. Math. Phys. **31**, 61 (1990).
[6] H. Ogura, T. Kawanishi, N. Takahashi, and Z. L. Wang, Waves Random Media **5**, 461 (1995).
[7] R. H. Cameron and W. T. Martin, Ann. Math. **48**, 385 (1947).
[8] K. Ito, J. Math. Soc. Jpn. **13**, 157 (1951).
[9] K. Ito, Jpn. J. Math. **22**, 63 (1952).
[10] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).
[11] T. Imamura, W. Meecham, and A. Siegel, J. Math. Phys. **6**, 695 (1965).
[12] D. W. Schuerman, in *Light Scattering by Irregularly Shaped Particles* (Plenum, New York, 1980).

Lattice electromagnetic theory from a topological viewpoint[a]

F. L. Teixeira[b] and W. C. Chew
Center for Computational Electromagnetics, Electromagnetics Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801-2991

(Received 7 July 1998; accepted for publication 25 September 1998)

The language of differential forms and topological concepts are applied to study classical electromagnetic theory on a lattice. It is shown that differential forms and their discrete counterparts (cochains) provide a natural bridge between the continuum and the lattice versions of the theory, allowing for a natural factorization of the field equations into topological field equations (i.e., invariant under homeomorphisms) and metric field equations. The various potential sources of inconsistency in the discretization process are identified, distinguished, and discussed. A rationale for a consistent extension of the lattice theory to more general situations, such as to irregular lattices, is considered.
[S0022-2488(99)02301-4]

I. INTRODUCTION

For historical reasons, the prevalent approach to study initial and/or boundary value problems for Maxwell's equations is based on the vector calculus language. Relatively few references make use of differential forms[1–23] as an alternative mathematical language to describe classical electromagnetic (EM) field theory, despite its adequacy and strong geometrical content. This is in marked contrast with the current tendency of geometrization of other areas of physics.[13–23]

The classical approach for deriving a lattice electromagnetic (EM) theory utilizes the vector calculus language. Such description assumes a space (and time) infinitely divisible. The discrete theory is then obtained by a finite-difference, finite-volume, or finite-element approximation. There has been a number of consistent (in the sense that properties exhibited by the continuum theory, such as divergence-preserving conditions, reciprocity, and conservation laws, are retained) and self-contained formulations of a lattice EM theory in the past.[24–26] However, until recently, these formulations were restricted for the finite-difference case and regular lattices.[24,26]

Apart from exhibiting some interesting physical phenomena not present in its continuum counterpart (such as high-frequency cutoff and rotational symmetry breaking), the interest in the development of a discrete EM theory is driven basically by the recent surge of interest in the numerical simulation of EM fields in complex environments using differential equation solvers, made possible by advances in computer technology. The absence of consistency in more general lattices usually leads to harmful effects on the numerical simulations for hyperbolic equations (time–domain simulations), such as unconditional late-time instabilities.[11,27] In the case of elliptic equations (frequency–domain simulations), they are usually associated with the presence of spurious modes.[3]

Traditionally, the derivation of a lattice EM theory using the classical, vector calculus approach has some inherent drawbacks, which hamper its application for developing more general discrete models on irregular lattices. First, it involves an approximation whereby derivatives are replaced by finite differences (e.g., in the finite-difference method). Second, in the case of structured lattices, the discretization is dependent on the underlying coordinate system; in the case of

[a]This work is dedicated to the memory of Professor George A. Deschamps.
[b]E-mail: fteixeir@cspark.ece.uiuc.edu

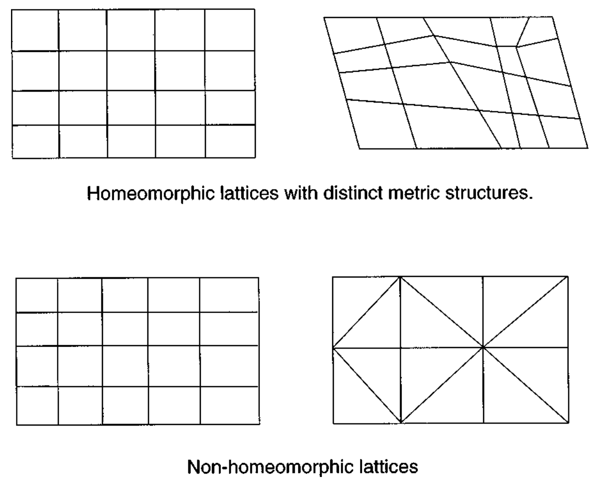

FIG. 1. Topological and metric structure of general, irregular lattices. The differential form language reveals the metric independence of Maxwell's equations. When put on a lattice, Maxwell's equations (but not Maxwell's system) become invariant for any lattice with the same topology.

unstructured lattices, the differential operators are replaced by integral operators and the evaluation of fields involves an averaging procedure (e.g., in the finite-volume method) or a projection on a functional space (e.g., in the finite-element method). More importantly, the use of vector calculus implies that the placement of the theory on the lattice depends on the metric on which the continuum field theory was first cast. This leads, in most cases, to lengthy formulas. Furthermore, the underlying geometrical concepts and the metric independence of Maxwell's equations are not explored. When written in the vector calculus language, the metric independence of Maxwell's equations is hidden because the topological structure is intertwined with their metric structure.

More recently, an alternative approach to the finite-difference discretization of Maxwell's equations for irregular quadrilateral lattices based on the support operator method (SOM)[28,29] has been described in Ref. 30. Such an approach is still based on the vector calculus language, but, compared to traditional discretization schemes, has the distinct advantage of being consistent *by construction* and, therefore, free from spurious solutions and numerical instabilities. Moreover, such an approach also explores the metric independence of Maxwell's equations, since the resulting discrete operators can be written as a composition of a topological part (formal differences) and a metric part.[28]

Here, we explore and discuss the application of differential forms and their discrete counterparts (cochains) to study lattice EM theory. One of the advantages of using differential forms is that the metric independence of Maxwell's equations is already factored out in the continuum, and, therefore, explicitly manifested.[2] This fact implies that the continuum Maxwell's equations written in the differential forms language are invariant under diffeomorphisms, while their lattice counterparts are invariant under homeomorphisms (in Fig. 1, we illustrate the concepts of the topological and metric structure of a lattice). Metric concepts are present only in the so-called Hodge star operators, which also generalize the constitutive relations of the medium. In the lattice, these operators can be thought of either defining *a priori* the local metric structure of the lattice, or being defined *a posteriori* by a given metric structure of the lattice.

In the discrete counterpart of the differential forms language, the continuum derivative operations are replaced not by finite-difference approximations, but as exact exterior derivatives on the lattice cell complex. The discrete exterior derivative corresponds not to a discrete *approximation*, but to a discrete *counterpart*. The exterior derivative is an operator that can be related to a simple evaluation of quantities on the boundary of the elements of the lattice complex, and which makes no assumptions about differentiability. These observations illustrate yet another advantage of the use of differential forms: their discrete counterparts are objects amenable to analysis using the powerful tools of algebraic topology.

Algebraic topological tools have been used to study discrete models for many years (see, e.g., Ref. 17 and references therein). The generality achieved by using algebraic topology was recently illustrated in Ref. 23, where it provided a conceptual basis to analyze the general similarities and differences among various discretization schemes, in the context of thermostatics.

The main objective of this paper is to tackle the general problem of the consistency of lattice EM theory within the framework of algebraic topology. By *general*, we mean lattices with arbitrary *metric* and *topological* structures. We distinguish three basic classes of consistency requirements. The first class (based on topological considerations only) is common to all field theories cast on a discrete form, and it is associated with the correct implementation of the boundary operator on the lattice. Discrete schemes that satisfy this first class can be classified as divergence-preserving schemes. The second class (also based on topological considerations only) is related to the topological structure of EM theory and the dual nature of ordinary and twisted cell complexes. The third class is the metric-dependent one, associated with the Hodge operators. We point out that each requirement is a separate, necessary condition for an overall consistent lattice EM theory.

The remainder of this work is organized as follows. In Sec. II, we write Maxwell's equations using the language of differential forms and discuss their factorization into topological and metric equations. In Sec. III, we review the discretization of differential forms on a lattice using algebraic topological tools. In Sec. IV, we put Maxwell's equations on the lattice using the concepts of the previous sections, stressing that it provides an exact counterpart to the continuum theory that is invariant under homeomorphisms. We also discuss the topological consistency requirements associated with the correct implementation of the boundary operator, and their connection with the usual theorems of vector calculus. In Sec. V, we discuss the concept of dual lattices and how it arises from the necessity of a proper discretization of the different geometrical objects representing the EM fields. In Sec. VI, we treat some additional algebraic properties of the resulting discrete Maxwell's equations by discussing additional topological consistency requirements associated with the dual structure of the ordinary and twisted cell complexes (important to guarantee reciprocity of the discrete Maxwell's equations). In Sec. VII, we discuss the problem of the discretization of the constitutive relations, where metric concepts are present and approximations are involved through the discretization of the Hodge operators. We do not present explicit constructions for the Hodge operators (these are highly problem specific); instead, we discuss general rationales for this, and describe basic requirements that *any* consistent version of the discrete Hodge should satisfy. Finally, in Sec. VIII, we summarize the conclusions. We use a $(3+1)$ representation with the $e^{-i\omega t}$ time convention assumed. Throughout this work, the term discretization refers to *spatial* discretization, unless indicated otherwise.

II. MAXWELL'S EQUATIONS AND DIFFERENTIAL FORMS

In the language of differential forms,[1–23] Maxwell's equations are written as

$$dE = i\omega B, \tag{1}$$

$$dH = -i\omega D + J_E, \tag{2}$$

$$dB = 0, \tag{3}$$

$$dD = \rho_E. \tag{4}$$

In the above, E and H are electric and magnetic field intensity 1-forms, D and B are electric and magnetic flux density 2-forms, J_E is the electric current density 2-form, and ρ_E is the electric charge density 3-form.

The operator d is the usual exterior derivative, which simultaneously plays the role of the curl and div operators of vector calculus. The exterior derivative is an operator applicable to any differentiable manifold, even without a metric defined on it. This is in contrast to the vector calculus operators, which depend on metric factors and have different expressions when written in different coordinate systems. The Maxwell's equations in the above form (27)–(30) are metric independent and retain the same form irrespective of the coordinate system used.[1,2]

Constitutive parameters of a given medium relate the 1-forms E,H to the 2-forms D,B and are given in terms of *Hodge operators*, $\star_e$ and $\star_h$,[6–8] as

$$D = \star_e E, \tag{5}$$

$$B = \star_h H. \tag{6}$$

These relations close the Maxwell's system. In this paper, the term Maxwell's equations will refer to Eqs. (1)–(4), while the term Maxwell's system will refer to (1)–(6). In the case of a three-dimensional manifold, the Hodge operator establishes a natural isomorphism between the space of 1-forms as E and H and the space of 2-forms as D and B. This isomorphism is usually called a Hodge duality map. The Hodge operators depend on a metric and, in the equations (1)–(6), all the information about the metric of space is contained in the constitutive relations (5) and (6). Any modification on the metric tensor preserves the form of Maxwell's equations.

The possibility of decomposing the field equations into a purely topological part and a metric one is not a special property of the EM theory. Such a decomposition is equally possible in the context of other classical field theories.[23]

III. DIFFERENTIAL FORMS ON A LATTICE

In this section, we will briefly review the correspondence between continuum and lattice equations provided by the known mapping of differential forms onto linear functions on the space of some lattice elements. For brevity, we have deliberately chosen a somewhat sloppy approach to any topological subtleties (these are discussed elsewhere, e.g., in Refs. 17–20), by focusing on the important concepts behind the terminology.

To make the right correspondence between the continuum and the lattice, the latter should be considered as a *cell complex* (or cell decomposition).[13,14,16–20,22,23] A cell complex is a partitioning of some space X into a finite number of k cells of different sizes, covering X without overlap), which form a set χ. In our case of interest, X is just a region of the three-dimensional Euclidean space. A k-cell s_i^k is an object homeomorphic to $\mathbf{R}^k$ so that, a 0-cell is a point, a 1-cell is a link (edge), etc. In general, k-cells are k-dimensional ($k=0,1,2,3$) elements of the lattice. The set of all k-cells is denoted by χ^k. The cell complex is the direct sum of such sets,

$$\chi = \bigoplus_{k=0}^{n} \chi^k. \tag{7}$$

Each k-dimensional element of χ^k, for $k=0,1,2,3$, corresponds to a point (vertex), link (edge), plaquette (face), and polyhedron (volume), respectively.

We assume a fixed orientation for each k-cell s_i^k on χ^k, which results in an *oriented cell complex*. For brevity, the term complex will refer to oriented complex in the remaining of this work. With this assumption, a *cell k-chain*, or simply k-chain, is defined as a linear combination of k-cells in χ^k through

$$S^k = \sum_i \alpha_i s_i^k \in \chi^k. \tag{8}$$

The weights α_i belong, in general, to any additive Abelian group. For our purposes, we do not need such a generality and assume that they are just integer numbers. From this definition, a 0-chain is a linear combination of points, a 1-chain is a linear combination of links, etc. A chain is always one of these types; there are no mixed chains.

The k-cells s_i^k form a basis for the space of k-chains. An arbitrary decomposition of X is not a cell complex χ. To characterize a cell complex, certain conditions must be observed. In particular, the *boundary* of any k-cell on χ should be the union of lower-dimensional cells in χ, and no overlapping cells are allowed. The boundary operator ∂ is an operator on χ, ∂: $\chi^k \mapsto \chi^{k-1}$, which carries the usual geometric interpretation and is subject to the requirement $\partial^2 \equiv \partial \circ \partial = 0$. It connects the algebra of k-chains with the algebra of $(k-1)$-chains on the lattice: if S^k is a k-chain, then ∂S^k is a $(k-1)$-chain. Moreover, the boundary operator acts linearly on the space of chains.

If Ω is a k-form and γ is a k-dimensional integration surface, then integration defines a pairing,

$$\int_\gamma \Omega, \tag{9}$$

which gives a scalar as a result. Therefore, the space of k-forms can be thought of as being *dual* to the space of k-dimensional surfaces. This motivates the definition of *cochains*: if the space of k-dimensional surfaces on the continuum is identified with the space of k-chains on the lattice; then the space of k-forms in the continuum is identified as the space of cochains, which are linear functionals on the space of chains. Cochains constitute the discrete representation for the differential forms on the lattice, or the discrete counterparts of forms. Here, to emphasize the connection between the lattice and the continuum, we will also refer to cochains as *lattice differential forms*.

The continuum pairing given by the previous equation has the following *exact counterpart* on the lattice, in terms of a k-chain S^k and a k-lattice form Θ^k,

$$\int_\gamma \Omega \rightarrow \langle S^k, \Theta^k \rangle. \tag{10}$$

Such pairing defines a *contraction* between S^k and Θ^k. From the basis for k-chains, s_i^k, we define a dual basis of lattice forms, θ_i^k, such that $\langle s_i^k, \theta_j^k \rangle = \delta_{ij}$. This basis generates the space of k-lattice forms so that the generic k-lattice form Θ^k is written as

$$\Theta^k = \sum_i \beta_i \theta_i^k. \tag{11}$$

In the above, $\beta_i \in G$, where G is some Abelian group. To make the correspondence with the continuum, G is assumed to be **R**, and the composition law assumed to be the usual algebraic law of addition. The contraction of chains and lattice forms then gives

$$\langle S^k, \Theta^k \rangle = \left\langle \sum_i \alpha_i s_i^k, \sum_j \beta_j \theta_j^k \right\rangle = \sum_i \sum_j \alpha_i \beta_j \langle s_i^k, \theta_j^k \rangle = \sum_i \alpha_i \beta_i. \tag{12}$$

From Eqs. (1)–(4), we see that the only spatial operator present in the differential forms language version of Maxwell's equations is the exterior derivative. The concept of duality makes the definition of the exterior derivative on a lattice very natural, usually called the *coboundary operator*. Here, to emphasize the connection between the lattice and the continuum, we will also refer to it as the *lattice exterior derivative* and write it with the same symbol, d, as in the continuum case. The lattice exterior derivative d is defined in terms of its adjoint, the boundary operator ∂, as

$$\langle S^k, d\Theta^{k-1} \rangle = \langle \partial S^k, \Theta^{k-1} \rangle. \tag{13}$$

This definition has some interesting properties. First, it defines d on a lattice in an *exact* manner. The coboundary does not correspond to an approximation to the continuum exterior derivative d, but instead, as a counterpart to it. Second, it is defined without any need for differentiability. Third, it automatically satisfies the generalized Stokes' theorem.

In this framework, the usual differential operators of vector calculus are replaced by a *single* operator on the lattice: the boundary operator ∂ acting on the elements of the lattice (cell complex). The div operator corresponds to the action of ∂ on a 3-cell, the curl to the action of ∂ on a 2-cell, and the grad to the action of ∂ on a 1-cell. Furthermore, in this sense, these operators are distilled from an unnecessary metric structure, becoming purely topological operations.

In the next section, these remarks will be substantiated when discussing Maxwell's equations on a lattice.

IV. MAXWELL'S EQUATIONS ON A LATTICE

In this section Maxwell's equations are put on a lattice using the previously discussed concepts. The lattice counterpart to Maxwell's equations (1)–(4) are formally the same as the continuum ones, written as

$$dE = i\omega B, \tag{14}$$

$$dH = -i\omega D + J_E, \tag{15}$$

$$dB = 0, \tag{16}$$

$$dD = \rho_E, \tag{17}$$

but with E and H properly interpreted as lattice 1-forms, B and D interpreted as lattice 2-forms, and d interpreted as the coboundary operator. Since lattice forms are operators on the space of chains, we need to contract the above equations with any 2- and 3-chains $S^2, \tilde{S}^2 S^3, \tilde{S}^3$ to get actual numbers,

$$\langle S^2, dE \rangle = i\omega \langle S^2, B \rangle, \tag{18}$$

$$\langle \tilde{S}^2, dH \rangle = -i\omega \langle \tilde{S}^2, D \rangle + \langle \tilde{S}^2, J_E \rangle, \tag{19}$$

$$\langle S^3, dB \rangle = 0, \tag{20}$$

$$\langle \tilde{S}^3, dD \rangle = \langle \tilde{S}^3, \rho_E \rangle. \tag{21}$$

Using the definition of the coboundary operator (the generalized Stokes' theorem), we get

$$\langle \partial S^2, E \rangle = i\omega \langle S^2, B \rangle, \tag{22}$$

$$\langle \partial \tilde{S}^2, H \rangle = -i\omega \langle \tilde{S}^2, D \rangle + \langle \tilde{S}^2, J_E \rangle, \tag{23}$$

$$\langle \partial S^3, B \rangle = 0, \tag{24}$$

$$\langle \partial \tilde{S}^3, D \rangle = \langle \tilde{S}^3, \rho_E \rangle. \tag{25}$$

We use an overtilde to distinguish between the chains belonging to the cell complex, χ (i.e., the cell complex over which E and B live), from the chains belonging the cell complex, $\tilde{\chi}$ (i.e., the cell complex over which D and H live). *These cell complexes are not necessarily the same* since the pair of equations for (E,B) and for (D,H) are independent of each other. Indeed, we will show that there are strong reasons to use different cell complexes for these quantities (this is discussed in the next section). To find the lattice forms at each edge or face, we apply the above procedure for each edge or face of the cell complex, or, equivalently, to each element s_i^k of the basis of k-chains,

$$\langle \partial s_i^2, E \rangle = i\omega \langle s_i^2, B \rangle, \tag{26}$$

$$\langle \partial \tilde{s}_i^2, H \rangle = -i\omega \langle \tilde{s}_i^2, D \rangle + \langle \tilde{s}_i^2, J_E \rangle, \tag{27}$$

$$\langle \partial s_i^3, B \rangle = 0, \tag{28}$$

$$\langle \partial \tilde{s}_i^3, D \rangle = \langle \tilde{s}_i^3, \rho_E \rangle. \tag{29}$$

Since ∂ is an operator from χ^k to χ^{k-1} (or $\tilde{\chi}^k$ to $\tilde{\chi}^{k-1}$), the boundaries $\partial s_i^1, \partial s_i^2, \partial s_i^3$ (or $\partial \tilde{s}_i^2, \partial \tilde{s}_i^2, \partial \tilde{s}_i^3$) can be expressed in terms of a basis of 0-, 1-, and 2-chains, respectively,

$$\partial s_i^1 = \sum_j \alpha_{ij} s_j^0, \tag{30}$$

$$\partial s_i^2 = \sum_j \beta_{ij} s_j^1, \tag{31}$$

$$\partial s_i^3 = \sum_j \gamma_{ij} s_j^2, \tag{32}$$

$$\partial \tilde{s}_i^1 = \sum_j \tilde{\alpha}_{ij} \tilde{s}_j^0, \tag{33}$$

$$\partial \tilde{s}_i^2 = \sum_j \tilde{\beta}_{ij} \tilde{s}_j^1, \tag{34}$$

$$\partial \tilde{s}_i^3 = \sum_j \tilde{\gamma}_{ij} \tilde{s}_j^2. \tag{35}$$

The matrices $\alpha_{ij}, \beta_{ij}, \gamma_{ij}$ (and $\tilde{\alpha}_{ij}, \tilde{\beta}_{ij}, \tilde{\gamma}_{ij}$) define the *incidence relations* for the operator ∂ in a given complex χ (and $\tilde{\chi}$). They are the discrete-topological counterpart to the grad, curl, and div operators, respectively. The elements of these matrices are integers having the values ± 1 (depending on their relative chosen orientation) when $s_j^{(k-1)} \in \partial s_i^k$, and zero otherwise. Furthermore, for any χ, the identity $\partial^2 = 0$ implies

$$\sum_j \beta_{ij} \alpha_{jk} = 0, \tag{36}$$

$$\sum_j \gamma_{ij} \beta_{jk} = 0, \tag{37}$$

and similar relations in $\tilde{\chi}$. Equations (36)–(37) are the topological equivalents on χ of the familiar identities curl $=0$ and div curl$=0$, respectively. The fact that these identities are preserved in the numerical discretization scheme are necessary to ensure that the theorems of the continuum are preserved, although not sufficient. Schemes satisfying (36)–(37) may be classified as divergence-preserving schemes, for obvious reasons.

Substituting (30)–(35) in (26)–(29), we have

$$\sum_j \beta_{ij} \langle s_j^1, E \rangle = i\omega \langle s_i^2, B \rangle, \tag{38}$$

$$\sum_j \tilde{\beta}_{ij} \langle \tilde{s}_j^1, H \rangle = -i\omega \langle \tilde{s}_i^2, D \rangle + \langle \tilde{s}_i^2, J_E \rangle, \tag{39}$$

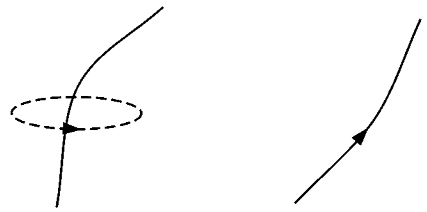

FIG. 2. The concept of two orientations applied for one-dimensional objects on a three-dimensional space. The external orientation is based on the circulation around the object and makes use of additional dimensions. The internal orientation is based on a direction along the object and does not require additional dimensions other than that of the object itself. The two orientations can be related to each other if a screw sense is defined. If the right-hand rule is adopted, the orientations chosen in this figure become equivalent. In the differential form language, two different objects may be defined according to the orientation they possess. Ordinary forms have an internal orientation. Twisted forms have an external orientation.

$$\sum_j \gamma_{ij}\langle s_j^2,B\rangle=0, \tag{40}$$

$$\sum_j \tilde{\gamma}_{ij}\langle \tilde{s}_j^2,D\rangle=\langle \tilde{s}_i^3,\rho_E\rangle. \tag{41}$$

These equations gives the *exact* lattice counterparts of Maxwell's equations in terms of the lattice variables $\langle s_i^2,B\rangle$, $\langle \tilde{s}_i^2,D\rangle$, $\langle \tilde{s}_i^1,H\rangle$, and $\langle s_i^1,E\rangle$. The lattice 2-forms *B, D* that live on 2-chains $s_i^2,\tilde{s}_i^2$ are related to the lattice 1-forms *E, D*, respectively, which live on 1-chains $\partial s_i^2,\partial\tilde{s}_i^2$.

The above equations in terms of lattice variables are the same for *any* lattice with the same topological structure. Maxwell's equations in this form are invariant under homeomorphisms. This topological equivalence leads to an equivalence relation among lattices.

We also note that the fundamental dynamic variables in the lattice theory are not the lattice forms *E,H,D,B* (i.e., the field values themselves) anymore, but their contraction with the cell complex elements. The latter quantities are the usual ones of interest associated with a finite region of space (i.e., global quantities like electric voltages and magnetic fluxes). The discretization process just described can be viewed as a process of limiting the (originally infinite) degrees of freedom in accessing these global quantities.

The problem of obtaining a *continuum* representation for the *continuum* forms *E,H,D,B* over *X* from the knowledge of $\langle s_i^1,E\rangle,\langle \tilde{s}_i^1,H\rangle,\langle \tilde{s}_i^2,D\rangle,\langle s_i^2,B\rangle$ over χ, $\tilde{\chi}$, is nevertheless important for discretizing the Hodge operators (constitutive relations) in (5) and (6) and to achieve the full discretization of the Maxwell's system (1)–(6). This is a part of the general problem of obtaining a consistent but approximate continuum representation for a differential *k*-form Ω on *X* from the knowledge of its discrete counterpart (*k*-cochain) Θ^k on χ^k. The discussion of this general problem is postponed until Sec. VII.

V. DUAL LATTICES AND TWISTED FORMS

In this section, we will discuss the concept of dual lattices for EM field simulations (such as in the Yee scheme[31] or usual finite-volume discretizations[32]) and show that its convenience arises not only for computational purposes, but also from geometrical reasons not obviated by the vector language. These reasons are connected with the concept of *orientation*.

We start by noting that there are two fundamental ways to define an orientation in three dimensions.[22,23] This is illustrated in Fig. 2 in the case of a one-dimensional object (line) in a three-dimensional space. The first way is to specify a (inner) *direction* along the line. This does not make use of additional dimensions other than the one defined by the line itself (one dimensional) and is referred to as *internal* orientation. The second way is to specify a (transversal)

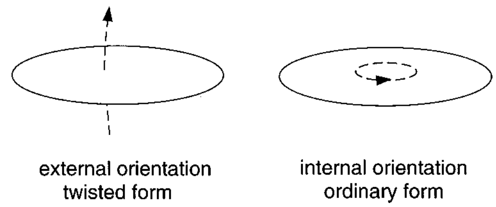

FIG. 3. The concept of two orientations now applied for two-dimensional objects on a three-dimensional space. The same observations made for Fig. 2 also apply here.

circulation along the line. In this case, additional dimensions are required and is referred to as *external* orientation. Internal and external orientations behave differently under coordinate reflection. These two kinds of orientations necessitate the definition of two different kinds of forms. Forms with internal orientation are called *ordinary* differential forms (for historical reasons only, since there is nothing about them to make them more ''ordinary'' than the twisted forms). Forms with external orientation are called *twisted* differential forms.[22] In the case of vectors, this distinction is not present because the vector calculus language comes with a predefined *screw sense* (in addition to a metric structure). For instance, if the right-hand rule is used for the objects (1-forms) of Fig. 2, then their vector counterparts automatically will have the same orientation. However, using the concept of two different kinds of orientations, no *a posteriori* right-hand rules are necessary.

Figure 3 illustrates the concept of internal and external orientation for two-dimensional objects in the three-dimensional space (this concept may also be applied for zero and three-dimensional objects[22,23]).

As expected, the boundary operator ∂ preserves orientation in the sense that the boundary of an ordinary/twisted form is another ordinary/twisted form. More interestingly, we note from Figs. 2 and 3 that the same concept (direction) that gives the internal orientation for one-dimensional objects gives the external orientation for two-dimensional objects. Similarly, the second concept (circulation), while giving internal orientation for two-dimensional objects, gives external orientation for one-dimensional ones.

If a given cell complex χ is chosen to discretize the space X so that its links (edges) have internal orientation and its faces have external orientation (consistently through the boundary operator), then there is a dual cell complex $\tilde{\chi}$ where edge orientations are given by circulations and the face orientations given by directions. This is depicted in Fig. 4. The edges (faces) of the dual cell complex are associated with faces (edges) of the primary cell complex. If the lattice ordinary forms live on the primary cell complex χ, then the lattice twisted forms should live on the dual cell complex $\tilde{\chi}$. This is one important distinction between differential forms in the continuum and the

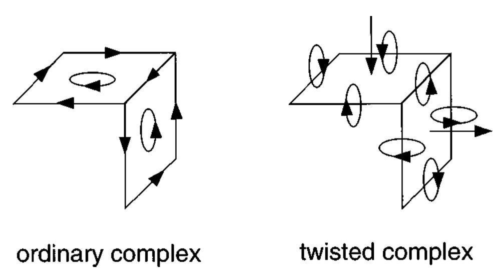

FIG. 4. Oriented cell complexes give rise to two distinct kinds of cell complexes. In the ordinary complex, each cell complex is endowed with an internal orientation and the associated lattice forms (cochains) are ordinary forms. In the twisted complex, each cell complex is endowed with an external orientation and the associated lattice forms (cochains) are twisted forms. On an EM lattice, these ordinary and twisted cell complexes are combined such that k-cells of one complex are associated to $(n-k)$-cells of the other, where n is the dimensionality of the space. This gives rise to the concept of a dual lattice (complex). The fields E and B live on the ordinary complex, while D and H live on the twisted complex (also see Fig. 5).

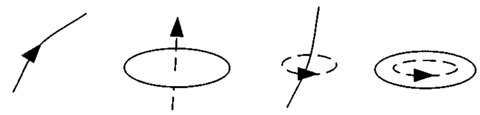

FIG. 5. Each EM field is a different geometrical object with distinct properties. This is hidden in the vector calculus language and best revealed through the use of the differential form language. In the continuum theory this distinction (usually) does not have an important role, but on a lattice, where the degrees of freedom in accessing global quantities are limited and exhibit a specific interdependence (e.g., edges as a boundary of faces) each field must be associated with a proper geometric object (1- or 2-cell) on the proper lattice (ordinary or twisted).

lattice forms. For the continuum forms, ordinary and twisted forms live on the same space X; but for the lattice forms, ordinary and twisted forms live on different spaces, χ and $\tilde{\chi}$. The role of the orientation concept in defining dual lattices was first discussed in Ref. 23.

To stress their geometrical properties, we will refer to the primary cell complex, χ, as the *ordinary cell complex* and to the dual cell complex, $\tilde{\chi}$, as the *twisted cell complex*.

The concept of distinct orientations applies directly to three-dimensional EM fields. The electric field E is associated with internally oriented lines (ordinary 1-form), the magnetic field H is associated with externally oriented lines (twisted 1-form), the electric flux D with externally oriented surfaces (twisted 2-form), and the magnetic flux B with internally oriented surfaces (ordinary 2-form). In addition, electric charge density is associated with externally oriented volumes (twisted 3-form), and the electric current density is associated with externally oriented surfaces (twisted 2-form). Figure 5 illustrates this classification for EM fields. The E and B lattice forms live on χ, while D and H (and J_E, ρ_E, which are also twisted forms) live on $\tilde{\chi}$.

The ordinary EM forms (E and B) are associated with the concept of forces (Lorentz formula), while the twisted EM forms (D and H) are associated with the concept of sources (ρ_E and J_E). Indeed, in the four-dimensional space–time notation, the forms E and B are components of a single two-form F (Faraday), while the forms D and H are components of its Hodge dual, $\star F$ (Maxwell two-form).[1]

Here, we appreciate the amount of geometric structure that is lost when representing the EM fields in the vector language. Each EM field is a distinct geometric object (Fig. 5), but in vector calculus, all are under the same umbrella as three-dimensional (contravariant) vectors. The need for two cell complexes arises not only as a computational device but also to account for the inherent geometric differences among the electromagnetic fields.

VI. ALGEBRAIC PROPERTIES OF MAXWELL'S EQUATIONS ON A LATTICE

The expansions (30)–(35) should satisfy, by construction, some conditions resulting from the properties of the boundary operator ∂ and from the dual complex construction. In this section, we shall describe and discuss these conditions. The fact that they are preserved in the lattice theory is important to preserve the continuum theorems and ensure an overall consistent theory.

The dual complex construction is such that, in the three-dimensional case, to each 2-cell of the ordinary cell complex, there corresponds a 1-cell on the twisted cell complex, and vice-versa. In this natural one-to-one pairing, the 1-cells (links), on the ordinary complex cross associate 2-cells (faces) on the dual complex, and vice-versa. A similar pairing also exist between 0-cells and 3-cells and vice-versa. For the two-dimensional case, the pairing is between ordinary 0-cells and twisted 2-cells (and vice-versa), and between ordinary and twisted 1-cells. This is true not only for hexahedral cells but for cell complexes with different cell topologies.

If the indices chosen for the basis elements of the cell complexes χ and $\tilde{\chi}$ reflects this natural pairing, i.e., if the cell s_i^k on the ordinary cell complex has the same i index of the associated cell $\tilde{s}_i^{(n-k)}$ (on an n-dimensional space) on the twisted cell complex, then it is easy to show that the coefficients for the incidence relations in (30)–(35) are related through (see Fig. 6)

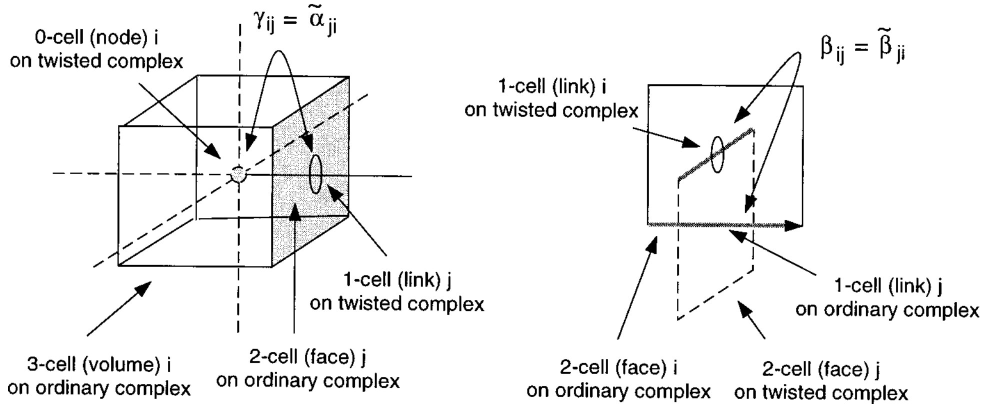

FIG. 6. Illustration of the reciprocal relationship between the boundary operator coefficients (incidence relations) for the ordinary and twisted lattices. At the left, we illustrate that a volume-to-face incidence relation on the ordinary complex is mirrored by a link-to-node incidence relation in the twisted complex. At the right, we illustrate that a face-to-link incidence relation in the ordinary complex is mirrored by another face-to-link incidence relation in the twisted complex. For visualization purposes, we depicted a regular hexahedral lattice, but this is valid for general lattices as well.

$$\alpha_{ij}=\tilde{\gamma}_{ji}, \tag{42}$$

$$\beta_{ij}=\tilde{\beta}_{ji}, \tag{43}$$

$$\gamma_{ij}=\tilde{\alpha}_{ji}. \tag{44}$$

These relations do not depend on the primary orientation chosen for χ and $\tilde{\chi}$. Similar to Eqs. (36)–(37), Eqs. (42)–(44) are another example of consistency relations derived from topological considerations only. The fact that these relations are satisfied in a particular discretization scheme is of key importance to preserve symmetry and positive definiteness of the resulting matrix systems, and, consequently, stability in time–domain updates of numerical methods. Combined with symmetry properties to be observed on the discrete representation for the Hodge operators $\star_\epsilon,\star_\mu$ (discussed in Sec. VII), they mimic, on a lattice, the reciprocal nature of the continuum Maxwell's equations. Equations (42)–(44) are also related with the consistency requirements for the boundary operator on the twisted complex $\tilde{\chi}$. This can be seen by substituting Eqs. (42)–(44) into Eqs. (36), (37),

$$\sum_j \beta_{ij}\alpha_{jk}=\sum_j \tilde{\beta}_{ji}\tilde{\gamma}_{kj}=\sum_j \tilde{\gamma}_{kj}\tilde{\beta}_{ji}=0, \tag{45}$$

$$\sum_j \gamma_{ij}\beta_{jk}=\sum_j \tilde{\alpha}_{ji}\tilde{\beta}_{kj}=\sum_j \tilde{\beta}_{kj}\tilde{\alpha}_{ji}=0, \tag{46}$$

which is equivalent to having the identity $\partial^2=0$ fulfilled on $\tilde{\chi}$.

For lattices with simple topology, the relations (42)–(44) hold true because of the staggered nature of the ordinary and twisted cells. However, for more exotic lattices, such as those encountered in subgridding or in modeling curved boundaries through locally distorted lattice elements, they do not necessarily hold true in commonly employed discretizations schemes (with naive interpolatory rules), so that the reciprocity of the continuum Maxwell's equations is lost. In these cases, the relations (42)–(44) should be enforced by construction to ensure that reciprocity is maintained. This is also discussed in Refs. 33–35, but from a completely different point of view.

VII. HODGE OPERATORS ON A LATTICE

The discretization of the Hodge operators $\star_\epsilon, \star_\mu$ (constitutive relations) is a central step for the formulation a general lattice EM theory. The discrete Hodge operators relate the lattice forms of the ordinary grid to the lattice forms of the twisted grid, and involve material properties of the particular medium. From the continuum equations, where the Hodge operators are metric-dependent objects, such relationship should also involve concepts such as lengths, angles, etc. Contrary to the topological equations treated before, the discrete Hodge operators are *approximations* to the continuum operators. The continuum Hodge operators are linear mapping of the space of k-forms into the space of $(n-k)$-forms, where n is the dimensionality of the space. In the EM case, the constitutive relations are written in terms of those operators, $D=\star_\epsilon E$, $B=\star_\mu H$, connecting the 2-forms D,B on one cell complex with the 1-forms E,H on the other cell complex. Since they are linear mappings, the discrete version of the Hodge operators in the lattice can be represented as a generic linear mapping connecting the dynamical discrete variables on χ and $\tilde{\chi}$ as follows:

$$[\star_\epsilon]:\chi\rightarrow\tilde{\chi},$$

$$\langle \tilde{s}_i^2, D\rangle = \sum_j [\star_\epsilon]_{ij}\langle s_j^1, E\rangle, \tag{47}$$

$$[\star_\mu]:\tilde{\chi}\rightarrow\chi,$$

$$\langle s_i^2, B\rangle = \sum_j [\star_\mu]_{ij}\langle \tilde{s}_j^1, H\rangle. \tag{48}$$

In the above, $[\star_\epsilon]$ and $[\star_\mu]$, are square, nonsingular, sparse matrices representing the discrete Hodge operators for a general dispersive and anisotropic linear media (in the general case of bianisotropic media, the following discussion remains essentially unchanged, except for the appearance of cross terms, $[\star_\zeta]$ and $[\star_\xi]$, in the above equations, relating quantities on the same cell complex). These approximate equations close the Maxwell's system (1)–(6) and, along with the exact discrete Maxwell's equations (38)–(41), constitute the discrete *approximation* to the Maxwell's system.

We now discuss some rationales for the construction of the Hodge operator on a lattice, but we do not claim such rationales to be unique. More importantly, we draw attention to the basic consistency requirements that any discrete Hodge should obey.

A rationale for a systematic construction of the discrete Hodge operators for a particular lattice geometry (metric) is described in Route *A* below.

Route A:

(i) An *approximate* continuum representation for the electromagnetic forms E and H is built from the knowledge of the discrete quantities $\langle s_i^1, E\rangle$ and $\langle \tilde{s}_i^1, H\rangle$ over χ and $\tilde{\chi}$, respectively (this will be discussed shortly).

(ii) The Hodge star operators $\star_\epsilon$ and $\star_\mu$ are applied to the resultant continuum representation for E and H, respectively, to yield the corresponding approximate continuum representations for D and B.

(iii) These resulting approximate representations for D and B are then paired with the elements of the cell complexes $\tilde{\chi}$ and χ, respectively, to yield $\langle \tilde{s}_i^2, D\rangle$ and $\langle s_i^2, B\rangle$. When $\langle \tilde{s}_i^2, D\rangle$ and $\langle s_i^2, B\rangle$ are written as functions of $\langle s_i^1, E\rangle$ and $\langle \tilde{s}_i^1, H\rangle$, respectively, we have determined the matrix elements $[\star_\epsilon]_{ij}$ and $[\star_\mu]_{ij}$. Note that in Eqs. (47), (48), the approximate continuum representations for the electromagnetic forms E,B,D,H do not appear. They are used only as an auxiliary tool to obtain $[\star_\epsilon]$ and $[\star_\mu]$. This is only natural, since, in a true discrete theory, E,B,D,H are not primary quantities and $[\star_\epsilon]$, $[\star_\mu]$ should be treated as being given *a priori*, in the same manner as the material constitutive tensors, $\overline{\boldsymbol{\epsilon}}$ and $\overline{\boldsymbol{\mu}}$, for the continuum theory.

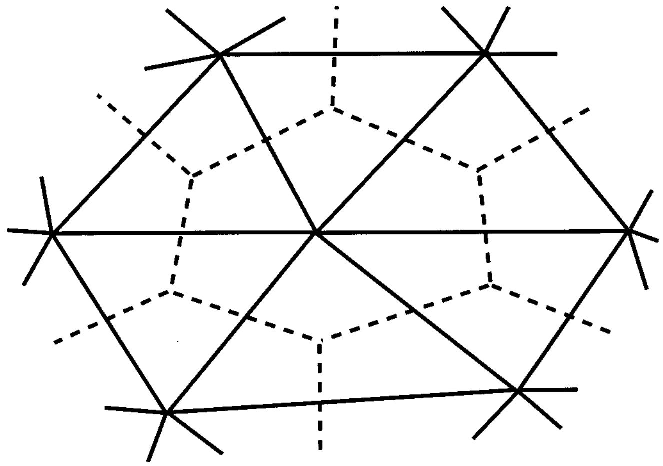

FIG. 7. A two-dimensional, irregular simplicial lattice (solid lines) and its dual lattice (dashed lines), which is not simplicial anymore (hexagonal elements).

As a result, an alternative route to (i)–(iii) can be followed by using, instead of *E,B,D,H*, their vector counterparts **E**,**B**,**D**,**H**, without changing the final formalism. This alternative route translates the metric dependency of the Hodge operators isolated in step (ii) to the metric-dependent notions of vector fields and surface integrals in modified versions of steps (i) and (iii). In this case, the modified step (ii) becomes just a tensorial product on vectors. This alternative route to determine the discrete Hodge operator may be summarized as follows.

Route B:

(i) An *approximate* continuum representation for the electromagnetic vector fields **E** and **H** is built from the knowledge of the discrete quantities $\langle s_i^1,E\rangle$ and $\langle \tilde{s}_i^1,H\rangle$ over χ and $\tilde{\chi}$, respectively.

(ii) The approximate continuum representations for **D** and **B** are found using the tensorial products $\mathbf{D}=\overline{\boldsymbol{\epsilon}}\cdot\mathbf{E}$ and $\mathbf{B}=\overline{\boldsymbol{\mu}}\cdot\mathbf{H}$.

(iii) These resulting approximate representations for the vector fields **D** and **B** are then integrated over 2-cells to give $\langle \tilde{s}_i^2,D\rangle$ and $\langle s_i^2,B\rangle$.

Variant schemes from the above are possible, where instead of first using the contractions of 1-forms, $\langle s_i^1,E\rangle$ and $\langle \tilde{s}_i^1,H\rangle$ over χ and $\tilde{\chi}$, to find the approximate continuum representations, the contractions used are located over the same cell complex, say χ. In such a case, the continuum representations are derived from $\langle s_i^1,E\rangle$ and $\langle s_i^2,B\rangle$ and the discrete operators obtained are approximations to $\star_\epsilon$ and $\star_\mu{}^{-1}$. This does not result in equivalence, however, because, in general, $[\star_\mu{}^{-1}]\neq[\star_\mu]^{-1}$. In particular, the matrix $[\star_\mu{}^{-1}]$ is sparse, but $[\star_\mu]^{-1}$ is not. The use of a same cell complex to obtain the continuum field representation from the contractions is sometimes of interest because, in general, the primary and dual cell complexes have different topological structures (not only metric structure), as illustrated in Fig. 7. As will become clear shortly, a proper continuum representation construction at step (i) depends directly on the topology of the cell complex over which the contractions are defined.

Both Routes *A* and *B* above are inevitably metric dependent, and, in principle, there is no conceptual advantage in adopting one over another. However, since Route *B* involves vector calculus concepts only, which supposedly results in a more familiar operational approach, we will base our remaining discussion mainly on it.

To find the continuum approximation for the vector fields **E** and **H**, a continuum basis of *k*-forms $\Omega_{s_i^k}$, counterpart to the basis of *k*-cochains θ_i^k (duals to s_i^k) should be constructed. Such a basis is built to obey the following properties.

(i) $\Omega_{s_i^k}=0$ outside s_i^k and its neighborhood (i.e., such forms should be compactly supported), which ensures the sparsity of $[\star_\epsilon]$, $[\star_\mu]$.

(ii) $\Omega_{s_i^k} = d\Omega_{\partial s_i^k}$, consistent with Eq. (13) and indicating a proper relationship between the $\Omega_{s_i^k}$'s for the different k's and a hierarchical construction that can be invoked.

(iii) The lattice form θ_i^k and its continuum counterpart $\Omega_{s_i^k}$ give the same results when contracted or integrated, respectively, over corresponding elements of χ and X, i.e.,

$$\int_\gamma \Omega_{s_i^k} = \langle s_j^k, \theta_i^k \rangle, \tag{49}$$

where γ is the k-dimensional region in X corresponding to the k-cell s_j^k in χ.

A basis for the space of forms obeying (i)–(iii) is introduced in Ref. 13 for a *simplicial* lattice (cell complex), and the resulting continuum forms are usually called *Whitney forms*. A simplicial lattice is one having the property that *all* its cell elements are *simplices*, i.e., cells whose boundaries are the union of a minimal number of lower-dimensional cells. Therefore, in a simplicial lattice, a 0-cell is a point (0-simplex), a 1-cell a link (1-simplex), a 2-cell a triangle (2-simplex), a 3-cell a tetrahedron (3-simplex), etc. The Whitney form associated with k-cell (simplex) s_i^k is written as[3,13]

$$\Omega_{s_i^k} = k! \sum_{j=0}^{k} (-1)^j \zeta_{i,j} d\zeta_{i,0} \wedge \cdots d\zeta_{i,j-1} \wedge d\zeta_{i,j+1} \wedge \cdots \wedge d\zeta_{i,k}, \tag{50}$$

where $\zeta_{i,j}$, $0 \leqslant j \leqslant k$, are the barycentric coordinates of the simplex s_i^k, and the wedge denotes the usual exterior product. These are piecewise linear forms. Higher-order forms are also possible. The Whitney k-forms are just linear interpolants for simplicial cochains and are uniquely determined from their integration over the k-simplices, which completely defines their degrees of freedom.

Using a Euclidean metric for the continuum three-dimensional case, the Whitney forms for the various degrees can be easily written using the vector calculus language[3] as basis functions for **E** and **H** [step (i) of Route *B* above]. For the $k=0$ case, we simply have, for each node i,

$$\Omega_{s_i^0} \overset{g_E}{\rightarrow} \tau_{s_i^0} = \zeta_{i,0}, \tag{51}$$

where $\zeta_{i,0}$ is the (single) barycentric coordinate associated with the 0-simplex s_i^0 (node), and g_E denotes the isomorphism (0-form to scalar) governed by the Euclidean metric. In this case, the Whitney form is the barycentric coordinate itself (a scalar) and the continuum approximation for a lattice 0-form is the usual node (point-based) interpolation through scalar functions $\tau_{s_i^0}$. The value of this function is equal to unity at s_i^0, and equal to zero at all other 0-cells s_j^0, $j \neq i$.

For the $k=1$ case, we have, for each edge i,

$$\Omega_{s_i^1} \overset{g_E}{\rightarrow} \boldsymbol{\tau}_{s_i^1} = \zeta_{i,0} \nabla \zeta_{i,1} - \zeta_{i,1} \nabla \zeta_{i,0}, \tag{52}$$

where $\zeta_{i,0}$ and $\zeta_{i,1}$ are the barycentric coordinates associated with the *two* vertices of the 1-simplex s_i^1 (edge). The Whitney forms (1-forms) in this case translate to a vector field, $\boldsymbol{\tau}_{s_i^1}$ (from the isomorphism between 1-forms and vectors governed by the Euclidean metric). The resultant interpolation scheme for a lattice 1-form using the above elements is the so-called edge interpolation. The line integral of this function is equal to unity along its associated edge s_i^1, and equal to zero at all other 1-cells s_j^1, $j \neq i$.

For the $k=2$ case, we have, for each face i,

$$\Omega_{s_i^2} \overset{g_E}{\rightarrow} \boldsymbol{\tau}_{s_i^2} = 2(\zeta_{i,0} \nabla \zeta_{i,1} \times \nabla \zeta_{i,2} + \zeta_{i,1} \nabla \zeta_{i,2} \times \nabla \zeta_{i,0} + \zeta_{i,2} \nabla \zeta_{i,0} \times \nabla \zeta_{i,1}), \tag{53}$$

where now $\zeta_{i,j}$, $j=0,1,2$ are the barycentric coordinates associated with the three vertices of the 2-simplex s_i^2 (triangular face). The Whitney forms (2-forms) in this case translate to a (pseudo-) vector field, $\boldsymbol{\tau}_{s_i^2}$ (from the isomorphism between 2-forms and pseudovectors governed by the Euclidean metric). The surface integral of this function is equal to unity over its associated face s_i^2, and equal to zero at all other 2-cells s_j^2, $j\neq i$.

The $k=3$ case, we have, for each volume i,

$$\Omega_{s_i^3}\overset{g_E}{\rightarrow}\tau_{s_i^3}=6[(\zeta_{i,0}\,\nabla\zeta_{i,1}\times\nabla\zeta_{i,2})\cdot\nabla\zeta_{i,3}+(\zeta_{i,1}\,\nabla\zeta_{i,2}\times\nabla\zeta_{i,3})\cdot\nabla\zeta_{i,0}+(\zeta_{i,2}\,\nabla\zeta_{i,3}\times\nabla\zeta_{i,0})\cdot\nabla\zeta_{i,1}$$
$$+(\zeta_{i,3}\,\nabla\zeta_{i,0}\times\nabla\zeta_{i,1})\cdot\nabla\zeta_{i,2}], \tag{54}$$

which results in a (pseudo-)scalar function, $\tau_{s_i^3}$ (from the isomorphism between 3-forms and pseudoscalars governed by the Euclidean metric) associated with each 3-simplex (tetrahedron). Despite the complicated appearance of Eq. (54), these are simple, step-like functions, which are constant on the associated tetrahedron and zero elsewhere. The corresponding volume integral is equal to unity over its associated volume s_i^3, and equal to zero over all other 3-cells s_j^3, $j\neq i$.

Using the Whitney forms, we may write the 1-chain approximation for the **E** fields as a sum of the $k=1$ vector basis functions running over all 1-simplices s_i^1 of the simplicial cell complex χ,

$$\mathbf{E}=\sum_i \langle s_i^1,E\rangle\boldsymbol{\tau}_{s_i^1}. \tag{55}$$

For the **H** field, the sum runs over the 1-cells of the dual complex $\tilde{\chi}$ (not simplicial anymore),

$$\mathbf{H}=\sum_i \langle \tilde{s}_i^1,H\rangle\boldsymbol{\tau}_{\tilde{s}_i^1}. \tag{56}$$

Alternatively, as discussed before, a discrete approximation may be first sought for the inverse operator $\star_\mu{}^{-1}$, so that

$$[\star_\mu{}^{-1}]:\chi\rightarrow\tilde{\chi},$$

$$\langle\tilde{s}_i^1,H\rangle=\sum_j [\star_\mu{}^{-1}]_{ij}\langle s_j^2,B\rangle, \tag{57}$$

and **B** is expanded over 2-simplices s_i^2 of the simplicial cell complex χ,

$$\mathbf{B}=\sum_i \langle s_i^2,B\rangle\boldsymbol{\tau}_{s_i^2}. \tag{58}$$

Note that Eqs. (55), (56), and (58) are approximations for the total vector fields (and not for each of their components separately).

The Whitney functions are commonly used as vector basis functions for the finite-element method to avoid the appearance of spurious solutions.[3] According to their order, k, they are sometimes referred to as node interpolants $(k=0)$, edge elements $(k=1)$, or face elements $(k=2)$. Such elements have been generalized for other types of complexes also (e.g., with hexahedral cells). However, as opposed to the simplicial case, there are no established mathematical results behind such generalizations. More importantly, some of the basic properties of Whitney forms on a simplicial lattice are not preserved in more general lattices. Among them is the divergence-free condition. It can be shown that, although for *regular* hexahedral lattices, these vector basis functions are divergence-free inside each element, in the case of general hexahedral elements, this is not true.[36] The divergence in this case is proportional to the amount of deviation from a regular lattice.

For the general cell complex case, i.e., not necessary simplicial [note that even for the simplicial case, its dual lattice is not simplicial anymore, as exemplified by Fig. 7 and Eq. (56)], and in the present absence of definitive mathematical results, the use of simpler *ad hoc* interpolatory schemes to obtain $[\star_\epsilon]$ and $[\star_\mu]$ are of interest for practical purposes. Any such scheme is highly dependent on the type of problem and geometry considered. However, *any* interpolatory scheme should meet some basic consistency requirements, described next.

For a reciprocal medium, $\overline{\boldsymbol{\mu}}=\overline{\boldsymbol{\mu}}^t$, $\overline{\boldsymbol{\epsilon}}=\overline{\boldsymbol{\epsilon}}^t$, the continuum Hodge operators $\star_\mu$, $\star_\epsilon$ on a Riemannian manifold are symmetric, nondegenerate, and positive definite operators. These properties are a simple consequence from the fact that a Riemannian metric tensor itself is a symmetric, nondegenerate, positive definite tensor.

However, for nonorthogonal lattices, usual interpolatory schemes for the discrete Hodge operators do not generally preserve the symmetry of the continuum operators. This is because the discrete versions of the Hodge are not strictly local. The lattice variables $\langle s_i^1, E\rangle$, $\langle \tilde{s}_i^1, H\rangle$, $\langle s_i^2, B\rangle$, $\langle \tilde{s}_i^2, D\rangle$ are defined over different geometric elements that span *finite* regions of space. Elements of the ordinary (twisted) cell complex that contribute to a given element of the twisted (ordinary) cell complex (and therefore define the local interpolatory stencil) are associated with metric elements defined at different points of space. Since the metric itself is a function of position, the *local* symmetry of $\star_\mu$, $\star_\epsilon$ may be lost on $[\star_\mu]$, $[\star_\epsilon]$, if not enforced by construction.[27] In addition, even if symmetry is enforced by construction on $[\star_\mu]$, $[\star_\epsilon]$, the positive definiteness condition on the discrete Hodge may be violated when highly skewed lattices are employed.[11,27]

Symmetric, positive-definite discrete Hodge operators yield real, positive eigenvalues for the matrices $[\star_\mu]$, $[\star_\epsilon]$. This means the resultant discrete Maxwell's system will not contain spurious eigenmodes with exponential growth in time. Nonsymmetric, nonpositive definite matrices would give rise to negative or complex eigenvalues. Regardless of their magnitude, negative or complex eigenvalues lead to spurious eigenmodes with *unconditional* exponential time growth that eventually contaminate the solution (late-time instabilities). This can be seen by substituting Eqs. (47), (48) into Eqs. (38)–(41) and solving, e.g., for $\langle s_i^1, E\rangle$. As a result, we get

$$\sum_m \left(\sum_{j,k,l} ([\star_\epsilon]^{-1})_{ij} \tilde{\beta}_{jk} ([\star_\mu]^{-1})_{kl} \beta_{lm} - \omega^2 \delta_{im} \right) \langle s_m^1, E\rangle = 0, \tag{59}$$

for all *i*. The eigenfunctions of the corresponding system of differential equations for $\langle s_i^1, E\rangle$ in the time–domain are given by

$$\langle s_i^1, E\rangle \rightarrow \boldsymbol{\phi}(t) = e^{\pm i \overline{\mathbf{A}}^{1/2} t} \boldsymbol{\phi}^0, \tag{60}$$

where $\boldsymbol{\phi}(t)$ is the column vector of eigenfunctions, $\boldsymbol{\phi}^0$ is the initial value of $\boldsymbol{\phi}(t)$ at $t=0$, and the elements of the matrix $\overline{\mathbf{A}}$ are given by

$$A_{im} = \sum_{j,k,l} ([\star_\epsilon]^{-1})_{ij} \tilde{\beta}_{jk} ([\star_\mu]^{-1})_{kl} \beta_{lm}. \tag{61}$$

The above functional operations on matrices are understood in the usual manner, by using similarity transformations and operating on the matrix eigenvalues. We let $\overline{\mathbf{A}}=\overline{\mathbf{W}}\cdot\overline{\boldsymbol{\lambda}}\cdot\overline{\mathbf{V}}^t$, or $\overline{\mathbf{A}}\cdot\overline{\mathbf{W}} = \overline{\mathbf{W}}\cdot\overline{\boldsymbol{\lambda}}$, where $\overline{\mathbf{W}}$ contains the right eigenvectors of $\overline{\mathbf{A}}$, while $\overline{\mathbf{V}}$ contains the left eigenvectors (for a symmetric $\overline{\mathbf{A}}$, they are the same), and $\overline{\mathbf{V}}^t\cdot\overline{\mathbf{W}}=\overline{\mathbf{I}}$. Consequently, $\overline{\mathbf{A}}\cdot\overline{\mathbf{W}}\cdot\mathbf{u}=\overline{\mathbf{W}}\cdot\overline{\boldsymbol{\lambda}}\cdot\mathbf{u}$, and $\overline{\mathbf{A}}^n\cdot\overline{\mathbf{W}}\cdot\mathbf{u}=\overline{\mathbf{W}}\cdot\overline{\boldsymbol{\lambda}}^n\cdot\mathbf{u}$, or, in general, $f(\overline{\mathbf{A}})\cdot\overline{\mathbf{W}}\cdot\mathbf{u}=\overline{\mathbf{W}}\cdot f(\overline{\boldsymbol{\lambda}})\cdot\mathbf{u}$, by using a Taylor expansion on $f(\cdot)$, where $\overline{\boldsymbol{\lambda}}$ is a diagonal matrix containing the eigenvalues of $\overline{\mathbf{A}}$. Using $\overline{\mathbf{V}}^t\cdot\overline{\mathbf{W}}=\overline{\mathbf{I}}$, we can let $\boldsymbol{\phi}^0=\overline{\mathbf{V}}^t\cdot\overline{\mathbf{W}}\cdot\boldsymbol{\phi}^0$, i.e., expand $\overline{\mathbf{f}}^0$ in terms of the eigenvectors of $\overline{\mathbf{W}}$. Equation (60) then becomes

$$\boldsymbol{\phi}(t) = \overline{\mathbf{W}}\cdot e^{\pm i \overline{\boldsymbol{\lambda}}^{1/2} t} \overline{\mathbf{V}}^t \cdot \boldsymbol{\phi}^0. \tag{62}$$

Equation (62) gives the solutions of the semidiscrete problem, i.e., without considerations about the time discretization. For any convergent time-discretization scheme (e.g., independent of the time step chosen), Eq. (62) will lead to *bounded* solutions, $\boldsymbol{\phi}(t)$, if all the eigenvalues of $\overline{\mathbf{A}}$ are real and positive (note that for a lossless, dispersionless media, $\overline{\mathbf{A}}$ is real and, therefore, any complex eigenvalues will occur in conjugate pairs). If $\tilde{\beta}_{ij}=\beta_{ji}$ [Eq. (43)], it can be easily shown that this is true if $[\star_\epsilon]^{-1}$ and $[\star_\mu]^{-1}$ (and, consequently, $[\star_\epsilon]$ and $[\star_\mu]$) are *simultaneously* nonsingular, symmetric, negative definite or positive definite. The positive definite is the case of interest to recover the continuum Hodge operators.

As observed in Sec. IV, the discretization process can be viewed as a process of limiting the degrees of freedom in accessing global dynamic quantities of interest. The original infinite degrees of freedom in the continuum theory are reduced to a finite number over the cell complex elements. In the semidiscrete dynamic equations, Eq. (59), this is reflected in the reduction of the spectral content of the solution to a finite number of poles (eigenfrequencies). Usually this may also be viewed as a low-pass filtering, which is determined by the lattice spacing size and nature of approximation; but in general terms corresponds to a rearrangement of the spectral content. The requirement for symmetric, positive definite discrete Hodge operators corresponds to assuring (in lossless media) that no spurious poles are introduced in the upper-half complex ω plane after discretization.

An additional, interesting point revealed by the differential forms language is that, since the metric is entirely defined in the Hodge operators, the simulation of Maxwell's equations on an *irregular* lattice and homogeneous medium can be mimicked by a dual theory, where we view the simulation performed on a *regular* lattice, but on an inhomogeneous, particular class of orthotropic medium with electric and magnetic constitutive tensors proportional to each other, i.e., $\overline{\boldsymbol{\epsilon}}=\epsilon\overline{\boldsymbol{\Gamma}}$, $\overline{\boldsymbol{\mu}}=\mu\overline{\boldsymbol{\Gamma}}$. In this latter case, metric factors are incorporated into the medium properties so that lattice irregularities become orthotropic inhomogeneities. By making use of such an observation, the properties that should be explicitly enforced on the final, approximate matrix representation of the discrete Hodge operators on a general irregular lattice can be established on very simple physical grounds only. This is simply because the violation of symmetry or positive definiteness would render the orthotropic media of the dual theory nonreciprocal or active, giving rise to spurious numerical artifacts.

As mentioned, symmetry and positive definiteness for $[\star_\epsilon]$ and $[\star_\mu]$ (or their procedural equivalents in the vector calculus language) are not always met by some of the commonly employed interpolations for finite-volume or finite-difference simulations. Symmetry is guaranteed only if it is explicitly enforced at each lattice point (e.g., through a perfectly symmetric numerical evaluation of the metric coefficients[27]) and the positive definiteness is usually violated when using highly skewed or curved meshes.[11,27]

It should be stressed that these (metric-dependent) conditions over $[\star_\epsilon]$ and $[\star_\mu]$ are not sufficient for the consistency of the lattice theory. They should be enforced *in addition to* the topological consistency conditions previously discussed in Secs. IV and VI.

VIII. CONCLUSIONS

In this work, we discussed the application of differential forms and topological concepts to the study of lattice EM theory.

Differential forms provide a very concise and elegant language to treat the classical EM theory on a lattice. It allows for the factorization of the field equations into a topological part and a metric part. The resultant topological equations are invariant under homeomorphisms, viz., invariant for lattices with the same topological structure. All the usual vector calculus operators are unified by a single operator, the exterior derivative, which admits a trivial and exact discretization on an arbitrary lattice through the use of its discrete adjoint, the boundary operator. This allows for a more general interpretation for the derivative on the lattice, not as a finite-difference approximation, but as an evaluation of fields at boundaries.

Consistency conditions for the lattice theory, such as divergence-free conditions and reciprocity, are discussed in a very general setting using purely topological concepts. Metric concepts need to be invoked only in connection with the Hodge operators, which also generalize the constitutive relations of the medium. General consistency requirements on the discrete Hodge operator are also discussed.

Lattice differential forms provide a richer geometrical language to discuss some aspects of the discretization procedure. The potential sources of inconsistency can be adequately identified and classified. The treatment of the EM fields E and B as ordinary forms, and D and H as twisted forms reveals a geometric reason of the dual lattice construction, common to EM discretization schemes for numerical simulations, such as the celebrated Yee scheme.[31]

ACKNOWLEDGMENTS

G. A. Deschamps was among the first to advocate the use of forms as a natural language for electromagnetic theory.[2] He passed away on June 20, 1998, while this paper was being prepared. This paper is dedicated to his memory. The authors would like to acknowledge the reviewers for suggestions, E. A. Forgy for discussions and for pointing out Ref. 16, and M. Shashkov and R. Schuhmann for providing preprints of their recent works. This work is supported by the Air Force Office of Scientific Research under MURI Grant No. F49620-96-1-0025, Office of Naval Research under Grant No. N00014-95-1-0872, the National Science Foundation under Grant No. ECS93-02145, and a CAPES Fellowship.

[1] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, New York, 1973).
[2] G. A. Deschamps, ''Electromagnetics and differential forms,'' Proc. IEEE **69**, 676–696 (1981).
[3] A. Bossavit, ''Whitney forms: a new class of finite elements for three-dimensional computations in electromagnetics,'' IEE Proc., Part A: Phys. Sci., Meas. Instrum., Manage. Educ. **135**, 493–500 (1988).
[4] P. R. Kotiuga, ''Helicity functionals and metric invariance in three dimensions,'' IEEE Trans. Magn. **25**, 2813–2815 (1989).
[5] A. Bossavit, ''Differential forms and the computation of fields and forces in electromagnetism,'' Eur. J. Mech. B/Fluids **10**, 5,474–488 (1991).
[6] K. F. Warnick, R. H. Selfridge, and D. V. Arnold, ''Electromagnetic boundary conditions and differential forms,'' IEE Proc. Microwave Ant. Prop. **142**, 326–332 (1995).
[7] K. F. Warnick, R. H. Selfridge, and D. V. Arnold, ''Teaching electromagnetic field theory using differential forms,'' IEEE Trans. Edu. **40**, 1,53–68 (1997).
[8] K. F. Warnick and D. V. Arnold, ''Green forms for anisotropic, inhomogeneous media,'' J. Electromagn. Waves Appl. **11**, 1145–1164 (1997).
[9] A. de La Bourdonnaye and S. Lala, ''Duality between finite elements and Hodge operator in three dimensions,'' Proceedings of the 13th Ann. Rev. Prog. Appl. Comp. Electromag., Monterey, California, 1997, pp. 606–612.
[10] A. Arkko, T. Tarhasaari, and L. Kettunen, ''A time domain method for high frequency problems exploring the Whitney complex,'' Proc. 14th. Ann. Rev. Prog. Appl. Comp. Electromag., Monterey, California, 1998, pp. 121–126.
[11] C. Kraus and R. Ziolkowsky, ''Topological and geometrical considerations for Maxwell's equations on unstructured meshes,'' *Proceedings of the URSI Meeting*, Montreal, Canada, 1997, p. 714.
[12] F. L. Teixeira and W. C. Chew, ''Differential forms, metrics, and the reflectionless absorption of electromagnetic waves,'' J. Electromagn. Waves Appl. **13**, 665–686 (1999).
[13] H. Whitney, Geometric Integration Theory (Princeton University Press, Princeton, 1957).
[14] I. Montvay and G. Munster, *Quantum Fields on a Lattice* (Cambridge University Press, Cambridge, 1994).
[15] R. W. R. Darling, *Differential Forms and Connections* (Cambridge University Press, Cambridge, 1994).
[16] T. J. Honan, ''The geometry of lattice field theory,'' Ph.D. dissertation, University of Maryland, 1986.
[17] A. A. Dezin, *Multidimensional Analysis and Discrete Models* (CRC Press, Boca Raton, FL, 1995).
[18] A. S. Schwarz, *Topology for Physicists* (Springer-Verlag, New York, 1994).
[19] C. Nash and S. Sen, *Topology and Geometry for Physicists* (Academic, London, 1983).
[20] I. M. Singer and J. A. Thorpe, *Lecture Notes on Elementary Topology and Geometry* (Scott, Foresman and Company, Glenview, IL, 1967).
[21] R. L. Bishop and S. I. Goldberg, *Tensor Analysis on Manifolds* (Dover, New York, 1980).
[22] W. L. Burke, *Applied Differential Geometry* (Cambridge University Press, Cambridge, 1985).
[23] C. Mattiussi, ''An analysis of finite volume, finite element, and finite difference methods using some concepts from algebraic topology,'' J. Comput. Phys. **133**, 289–309 (1997).
[24] T. Weiland, ''On the numerical solutions of Maxwell's equations and applications in the field of accelerator physics,'' Part. Accel. **15**, 245–292 (1984).
[25] T. Weiland, ''Time domain electromagnetic field computations with finite difference methods,'' Int. J. Num. Model. **9**, 295–319 (1996).
[26] W. C. Chew, ''Electromagnetic theory on a lattice,'' J. Appl. Phys. **75**, 4843–4850 (1994).

[27] R. Schuhmann and T. Weiland, ''Stability of the FDTD algorithm on nonorthogonal grids related to the spatial interpolation scheme,'' IEEE Trans. Magn. **34**, 2751–2754 (1998).
[28] J. M. Hyman and M. Shashkov, ''Natural discretization for the divergence, gradient and curl on logically rectangular grids,'' Comput. Math. Appl. **33**, 81–104 (1997).
[29] J. M. Hyman and M. Shashkov, ''Adjoint operators for the natural discretization for the divergence, gradient and curl on logically rectangular grids,'' Appl. Numer. Math. **25**, 413–442 (1997).
[30] J. M. Hyman and M. Shashkov, ''Mimetic discretizations for Maxwell's equations and equations of magnetic diffusion,'' preprint LA-UR-98-1032, Los Alamos National Laboratory, 1998.
[31] K. S. Yee, ''Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media,'' IEEE Trans. Antennas Propag. **14**, 302–307 (1966).
[32] K. S. Yee, ''Ten years of evolution of the FDTD-like conformal techniques,'' Proc. 11th. Ann. Rev. Prog. Appl. Comp. Electromag., Monterey, California, 1995, pp. 1241–1257.
[33] I. J. Craddock, C. J. Railton, and J. P. McGeehan, ''Derivation and application of a passive equivalent circuit for the finite difference time domain algorithm,'' IEEE Microwave Guid. Wave Lett. **6**, 40–42 (1996).
[34] K. M. Krishnaiah and C. J. Railton, ''Passive equivalent circuit of FDTD: an application to subgridding,'' Electron. Lett. **33**, 1277–1278 (1997).
[35] P. Thoma and T. Weiland, ''A consistent subgridding scheme for the finite difference time domain method,'' Int. J. Num. Model. **9**, 359–374 (1996).
[36] J. M. Jin, *The Finite Element Method in Electromagnetics* (Wiley, New York, 1993).

Pseudo-orthogonal groups and integrable dynamical systems in two dimensions

Juan A. Calzada[a)]
Departamento de Matemática Aplicada a la Ingeniería, Universidad de Valladolid, E-47011 Valladolid, Spain

Mariano A. del Olmo[b)]
Departamento de Física Teórica, Universidad de Valladolid, E-47011 Valladolid, Spain

Miguel A. Rodríguez[c)]
Departamento de Física Teórica, Universidad Complutense, E-28040 Madrid, Spain

(Received 23 March 1998; accepted for publication 25 September 1998)

Integrable systems in low dimensions, constructed through the symmetry reduction method, are studied using phase portrait and variable separation techniques. In particular, invariant quantities and explicit periodic solutions are determined. Widely applied models in Physics are shown to appear as particular cases of the method. © *1999 American Institute of Physics.* [S0022-2488(99)01201-3]

I. INTRODUCTION

Integrable Hamiltonian systems play a fundamental role in the study and description of physical systems, due to their many interesting properties, both from the mathematical and physical points of view. The construction of such models represents a contribution to this field, and many of them have proved to be of an extraordinary physical interest. Let us be reminded here of the Morse[1] and Pöschl–Teller[2] potentials in one dimension, or the Calogero[3–5] and Sutherland[6] potentials. These constructions have also been considered from many points of view. See, for instance, the reviews in Refs. 7 and 8.

A method used to construct these systems is the Marsden–Weinstein reduction procedure,[9] or its extensions,[10] to free Hamiltonians lying on an N-dimensional homogeneous space under a suitable Lie group. In this way (using an appropriate momentum map), one assures the integrability, or even the superintegrability[11] of the system. In the first case, there exists N constants of motion in involution, one of them the Hamiltonian. The superintegrability requires more than N constants of motion (not all of them in involution) and more than one subset of N constants in involution. There are good reasons to suspect that any integrable system may be constructed in this way, as a reduction of a free one,[10] so the problem to construct these systems and study their properties is a profitable and very interesting field. A related topic is the problem of separation of variables for the associated Hamilton–Jacobi (HJ) equations. As it is well known, the existence of quadratic invariants allows us to classify and construct these systems,[12,13] relating them in many occasions to subgroups of the invariance group.

A series of articles appeared in the last years, and has been devoted to the study of these superintegrable systems constructed using homogeneous spaces over the pseudounitary groups SU (p,q).[14,15] In particular, using the maximal Abelian subalgebras (MASA) of the corresponding algebras, one can build a family of integrable systems of arbitrary dimension, and present their invariants and the coordinate systems in which the HJ equation is separated. The explicit solutions and a unifying view of the compact Cartan subalgebra case have been presented in Ref. 16.

a)Electronic mail: juacal@wmatem.eis.uva.es
b)Electronic mail: olmo@fta.uva.es
c)Electronic mail: rodrigue@eucmos.sim.ucm.es

Our aim in this article is to work in detail with the low dimensional cases. The reason is twofold. On one side, the one-dimensional case allows an easy geometric description of the systems, through their phase portrait. The potentials we obtain are not new, but have been applied successfully in many physical models (for instance, the Pöschl–Teller and Morse potentials). They also appear in the study of quasi-exactly solvable (QES) models,[17,18] as the case of exactly solvable systems, providing examples in which, from the quantum point of view, the corresponding Schrödinger equation can be solved algebraically (a part of the spectrum for QES systems or an arbitrary number of states for the exactly solvable ones). On the other side, the two-dimensional case can be studied from the point of view of variable separation, and we can solve the HJ equation in a wide class of coordinate systems, especially in the noncompact case.[19,20] The results we present here (in the two-dimensional case) should be considered in a local context. Considerations about global behavior, which will differ from the compact to the noncompact case, will not be addressed in this work.

The article is organized as follows. In Sec. II we present a concise description of the method used to construct these Hamiltonian systems. Section III is devoted to the one-dimensional case, while the two-dimensional case is studied in Sec. IV. In each case we present the list of all the conserved quantities for these systems in terms of the generators of the corresponding algebras, and the explicit form in the chosen coordinate system. Conclusions and further outlook of this research are discussed in Sec. V.

II. INTEGRABLE HAMILTONIAN SYSTEMS AND PSEUDOUNITARY GROUPS

The results presented in this section are a summary of the contents of Refs. 14 and 15. We will include some of them to set the notations which will be used in the following sections.

We will consider the free Hamiltonian ($\mu,\nu=0,\dots,N=p+q-1$),

$$H=4cg^{\mu\bar{\nu}}p_\mu\overline{p_\nu} \tag{2.1}$$

(the bar denoting complex conjugate) defined in the Hermitian hyperbolic space (with coordinates $y^\mu\in\mathbf{C}$, satisfying $g_{\bar{\mu}\nu}\overline{y^\mu}y^\nu=1$, and conjugate momenta p_μ),

$$\mathrm{SU}(p,q)/\mathrm{SU}(p-1,q)\times\mathrm{U}(1) \tag{2.2}$$

whose geometry is described in Ref. 21. The real constant c is related to the sectional curvature of the Hermitian space. See also Ref. 16 for a detailed analysis of this space and its properties.

Using a maximal Abelian subalgebra of $su(p,q)$,[22] we carry a reduction procedure,[9] in order to obtain a reduced Hamiltonian (which is not free) in the reduced space, a homogeneous $\mathrm{SO}(p,q)$ space,

$$H=c\left(\tfrac{1}{2}g^{\mu\nu}p_{s^\mu}p_{s^\nu}+V(s)\right), \tag{2.3}$$

where $V(s)$ is a potential depending on the real coordinates s^μ. The set of complex coordinates y^μ is transformed in the reduction procedure into a set of ignorable variables x^μ (which are the parameters of the transformation associated to the MASA of $u(p,q)$ used in the reduction) and the coordinates s^μ with the constraint $g_{\mu\nu}s^\mu s^\nu=1$.

If Y_μ, $\mu=0,\dots,N$, is a basis of the considered MASA of $u(p,q)$, formed by pure imaginary matrices, the relation between old (y^μ) and new coordinates (x^μ,s^μ) is

$$y^\mu=B(x)^\mu_\nu s^\nu,\quad B(x)=\exp(x^\mu Y_\mu), \tag{2.4}$$

which assures the ignorability of the x coordinates (the vector fields corresponding to the MASA are straightened out in these coordinates). The Jacobian matrix, J, is easily obtained. If

$$A^{\mu}_{\nu}=\frac{\partial y^{\mu}}{\partial x^{\nu}}=(Y_{\nu})^{\mu}_{\rho}y^{\rho}, \tag{2.5}$$

then

$$J=\frac{\partial(y,\overline{y})}{\partial(x,s)}=\begin{pmatrix} A & B \\ \overline{A} & \overline{B} \end{pmatrix}. \tag{2.6}$$

The Hamiltonian calculated in the new coordinates is written as

$$H=c(\tfrac{1}{2}g^{\mu\nu}p_{\mu}p_{\nu}+V(s)), \quad V(s)=p_x^T(A^{\dagger}KA)^{-1}p_x, \tag{2.7}$$

where p_x are the constant momenta associated to the ignorable coordinates x and K is the matrix defined by the metric g.

Note that, to obtain these Hamiltonians, we need MASAs of $su(p,q)$ of dimension $N=p+q-1$ (corresponding to MASAs of $u(p,q)$ of dimension $p+q$). We also require that these MASAs have a representation in terms of imaginary matrices that allows to write the Hamiltonian in the form (2.7). Once we have chosen a particular MASA, we can obtain a set of invariants and also the corresponding coordinate systems in which the HJ equation separates. The MASAs of $su(p,q)$ are classified in Ref. 22, and for low ranks are completely determined. The corresponding potentials have been obtained for SU(N) in Ref. 19, for SU(2,1) in Ref. 20, for SU(2,2) in Ref. 15 and for any SU(p,q), choosing as the MASA one of the Cartan subalgebras, in Ref. 14. From now on, we will always use contravariant coordinates, but we will write the indices as subscripts in order to simplify the notation and avoid the use of unnecessary brackets.

III. ONE-DIMENSIONAL HAMILTONIANS

One-dimensional Hamiltonians are always integrable and the phase portrait gives a complete description of the allowed motions. We shall expose the main ideas in order to achieve a better understanding of the more complicate systems we will study in the next section. We have two cases: $su(2)$ and $su(1,1)$.

A. $su(2)$

We will use as a basis for $su(2)$ the operators X_1,X_2,X_3, which are given in the natural 2×2 matrix representation by

$$X_1\rightarrow\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad X_2\rightarrow\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad X_3\rightarrow\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

in the metric $K=\mathrm{diag}(1,1)$.

In the compact algebra $su(2)$ there is only one class of MASAs, corresponding to the Cartan subalgebra (CC).[22] A basis of a representative of this class of MASA is

$$\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}. \tag{3.1}$$

A basis of the corresponding MASA of $u(2)$ can be chosen as

$$Y_0=\begin{pmatrix} i & 0 \\ 0 & 0 \end{pmatrix}, \quad Y_1=\begin{pmatrix} 0 & 0 \\ 0 & i \end{pmatrix}. \tag{3.2}$$

The relation between old and new coordinates is given by the matrix $B(x)$ in (2.4). Note that a change of basis in the corresponding MASA of $u(2)$ changes only the parameters appearing in the potential,

$$y_0 = s_0 e^{ix_0}, \quad y_1 = s_1 e^{ix_1}. \tag{3.3}$$

The Hamiltonian, following the general expression (2.7), is

$$H = c\left[\tfrac{1}{2}(p_{s_0}^2 + p_{s_1}^2) + V(s)\right], \quad V(s) = \frac{m_0^2}{s_0^2} + \frac{m_1^2}{s_1^2} \tag{3.4}$$

with the constraint $s_0^2 + s_1^2 = 1$. Parameterizing the circle S^1 in spherical coordinates, $s_0 = \cos\phi$, $s_1 = \sin\phi$, we get the Hamiltonian $(c=1)$,

$$H(\phi) = \frac{1}{2}p_\phi^2 + V(\phi), \quad V(\phi) = \frac{m_0^2}{\cos^2\phi} + \frac{m_1^2}{\sin^2\phi}. \tag{3.5}$$

We have only one second order conserved quantity, the Hamiltonian, which is equal to the Casimir of the algebra, C, up to an additive constant,

$$\hat{Q}_1 = X_2^2 + X_3^2. \tag{3.6}$$

The square of the generator in the compact Cartan subalgebra, $C_1 = X_1^2$, is constant after the reduction and $C = C_1 + \hat{Q}_1$.

The specific values of the real positive constants m_0, m_1 play no essential role in the qualitative description of the orbits and trajectories of this system. The potential has singularities (in the generic case) in $\phi = 0, \pi/2, \pi, 3\pi/2$. When m_0 or m_1 are equal to zero we have only two singularities in $0,\pi$ or $\pi/2, 3\pi/2$, respectively.

The particles are confined inside a sector, and there, the motion is periodic, with an equilibrium point (a center in the phase space) corresponding to the unique minimum of the potential, in $\tan\phi = \sqrt{m_1/m_0}$. The solution can be easily computed, using Hamilton equations. The potential is bounded from below, and the energy is always positive $(E \geqslant (m_0 + m_1)^2)$. Though the use of the HJ equation is not necessary in this context of one-dimensional systems, we will write down the equation in order to compare it to the two dimensional cases. In fact, when we will make separation of variables there, we will find again this equation,

$$\frac{1}{2}\left(\frac{\partial S}{\partial \phi}\right)^2 + \frac{m_0^2}{\cos^2\phi} + \frac{m_1^2}{\sin^2\phi} = E \tag{3.7}$$

with the solution $(u = \cos^2\phi)$,

$$u = \frac{1}{2E}\left(b + \sqrt{b^2 - 4m_0^2 E}\cos 2\sqrt{2E}t\right) \tag{3.8}$$

and $b = m_0^2 - m_1^2 + E$. This solution is obviously much simpler to find if we consider the equation of orbits in the phase portrait of this system.

For instance, if $m_0 = 0$, $m_1 = 1$, the solution is

$$s_0 = \cos\phi = \sqrt{1 - \frac{1}{E}}\cos\sqrt{2E}t, \tag{3.9}$$

that is, a system with similar solutions to a harmonic oscillator, but now the frequency depends on the energy.

B. $su(1,1)$

The noncompact algebra $su(1,1)$ has three nonconjugate classes of MASAs, compact Cartan subalgebra, noncompact Cartan subalgebra and a class of nilpotent maximal Abelian subalgebras (MANS).[22] We will fix the metric to be

$$K=\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{3.10}$$

and the basis $\{X_1,X_2,X_3\}$ is given in the 2×2 matrix representation through the correspondence

$$X_1\to\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},\quad X_2\to\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\quad X_3\to\begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}.$$

1. Compact Cartan subalgebra (CC)

We choose, as a representative of this class, the same matrices as in (3.2). Hence, the old and new coordinates are related in the same way they did in the $su(2)$ case (3.3), and the Hamiltonian is now

$$H=c[\tfrac{1}{2}(p_{s_0}^2-p_{s_1}^2)+V(s)],\quad V(s)=\frac{m_0^2}{s_0^2}-\frac{m_1^2}{s_1^2}, \tag{3.11}$$

with the constraint $s_0^2-s_1^2=1$. This hyperbola can be described with a coordinate ϕ varying in the real line, $s_0=\cosh\phi$, $s_1=\sinh\phi$, and the Hamiltonian in these coordinates is $(c=-1)$,

$$H(\phi)=\tfrac{1}{2}p_\phi^2+V(\phi),\quad V(\phi)=-\frac{m_0^2}{\cosh^2\phi}+\frac{m_1^2}{\sinh^2\phi}. \tag{3.12}$$

The second order invariant (the Hamiltonian) is

$$\hat{Q}_1=X_2^2+X_3^2 \tag{3.13}$$

and the trivial constant associated to the MASA is $C_1=X_1^2$. Hence, the Casimir in terms of these two quantities is $C=C_1-\hat{Q}_1$.

The HJ equation is

$$\frac{1}{2}\left(\frac{\partial S}{\partial\phi}\right)^2-\frac{m_0^2}{\cosh^2\phi}+\frac{m_1^2}{\sinh^2\phi}=E \tag{3.14}$$

and the solution depends on the values of E and the parameters.

Considering different values of the parameters m_0,m_1 we obtain three different systems.

(a) If $m_1\neq 0$ the potential has a singularity in $\phi=0$. It is easy to check that, if $m_1\geqslant m_0$, there are no minima for the potential and all the motions are unbounded (with a turning point). The energy is always positive. The parameters m_0,m_1 do not modify qualitatively the phase portrait or the form of the solutions. The solution can be written as $(u=\cosh^2\phi)$

$$u=\frac{1}{2E}(-b+\sqrt{b^2+4m_0^2E}\cosh 2\sqrt{2E}t) \tag{3.15}$$

and $b=m_0^2-m_1^2-E$. If $m_0=0$, $m_1=1$, the solution is

$$s_0(t)=\sqrt{1+\frac{1}{E}\cosh\sqrt{2E}t}. \tag{3.16}$$

(b) If $m_0 > m_1 > 0$, the potential has two minima, symmetric respect to the origin where it has the singularity. The energy is bounded from below, $E \geqslant -(m_0 - m_1)^2$, the value $E = -(m_0 - m_1)^2$ corresponding to the equilibrium solution in the center of the phase space. The other solutions are easily calculated (b has the same value as in case a),

(i) $-(m_0 - m_1)^2 < E < 0$

$$s_0 = \frac{1}{\sqrt{2|E|}} [b + \sqrt{b^2 + 4Em_0^2} \cos 2\sqrt{2|E|}t]^{1/2}. \tag{3.17}$$

(ii) $E = 0$

$$s_0 = \left[\frac{m_0^2}{m_0^2 - m_1^2} + 2(m_0^2 - m_1^2)t^2 \right]^{1/2}. \tag{3.18}$$

When $E > 0$ we get the solution (3.15).

(c) If $m_1 = 0$ there is no singularity in the potential, which has a minimum in $\phi = 0$, with periodic motions of negative energy and unbounded motions of positive or zero energy. The multiplicative constant m_0 plays no essential role for the qualitative description of the system. The solutions can be read off from case (b) with $m_1 = 0$.

2. Noncompact Cartan subalgebra (NC)

A representative subalgebra of this class has the basis [in the metric (3.10)],

$$Y_1 = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \tag{3.19}$$

and we will add the matrix $Y_0 = iI$ to get a MASA of $u(1,1)$. The new and old coordinates are related in a slightly more complicated way,

$$y_0 = e^{ix_0}(s_0 \cosh x_1 + is_1 \sinh x_1), \quad y_1 = e^{ix_0}(-is_0 \sinh x_1 + s_1 \cosh x_1) \tag{3.20}$$

and the Hamiltonian is written in the new coordinates as

$$H = c[\tfrac{1}{2}(p_{s_0}^2 - p_{s_1}^2) + V(s)], \quad V(s) = \frac{m_0^2 - m_1^2 + 4m_0 m_1 s_0 s_1}{1 + 4s_0^2 s_1^2}, \tag{3.21}$$

and the constraint $s_0^2 - s_1^2 = 1$. Using again the ϕ coordinate as in the previous case, we get

$$H(\phi) = \frac{1}{2} p_\phi^2 + V(\phi), \quad V(\phi) = -\frac{m_0^2 - m_1^2 + 2m_0 m_1 \sinh 2\phi}{\cosh^2 2\phi}. \tag{3.22}$$

The Casimir is written as $C = \hat{Q}_1 - C_1$, where $C_1 = X_3^2$, the square of the generator of the noncompact Cartan subalgebra, and

$$\hat{Q}_1 = X_1^2 - X_2^2, \tag{3.23}$$

which is equal to the Hamiltonian.

We will also write down HJ equation for future references,

$$\frac{1}{2}\left(\frac{\partial S}{\partial \phi}\right)^2 - \frac{m_0^2 - m_1^2 + 2m_0 m_1 \sinh 2\phi}{\cosh^2 2\phi} = E. \tag{3.24}$$

If m_0 or m_1 are equal to 0, we obtain similar results to those of the case of compact Cartan MASA, an attractive potential if $m_0^2 > m_1^2$ and a repulsive one in the opposite case. If $m_0 m_1 \neq 0$, the potential is qualitatively the same for all values of m_0 and m_1.

The potential $V(\phi)$ has two extrema in $\sinh 2\phi = m_1/m_0, -m_0/m_1$. The first point corresponds to a minimum (a center in the phase portrait), and the potential takes the value $V = -m_0^2$. The second point is a maximum (a saddle point in the phase portrait), and $V = m_1^2$ there. The energy is bounded from below ($E \geqslant -m_0^2$) and the explicit solutions are ($u = \sinh 2\phi$),

(i) $-m_0^2 < E < 0$

$$u = \frac{1}{|E|}\left[m_0 m_1 + \sqrt{(E+m_0^2)(m_1^2 - E)} \cos 2\sqrt{2|E|}t\right]. \tag{3.25}$$

(ii) $E = 0$

$$u = -\frac{m_0^2 - m_1^2}{2m_0 m_1} + 4m_0 m_1 t^2. \tag{3.26}$$

(iii) $0 < E < m_1^2$

$$u = \frac{1}{E}\left[-m_0 m_1 + \sqrt{(E+m_0^2)(m_1^2 - E)} \cosh 2\sqrt{2|E|}t\right]. \tag{3.27}$$

(iv) $E = m_1^2$

$$u = -m_0 + e^{2\sqrt{2}m_1 t}. \tag{3.28}$$

(v) $E > m_1^2$

$$u = \frac{1}{E}\left[-m_0 m_1 + \sqrt{(E+m_0^2)(E - m_1^2)} \sinh 2\sqrt{2E}t\right]. \tag{3.29}$$

3. Nilpotent subalgebra (NIL)

Though the simplest representative of this class of subalgebras is obtained in the skew-diagonal metric, we will use again the diagonal one, because in this way, the kinetic term is also diagonal. We will take as a basis,

$$Y_1 = \begin{pmatrix} i & i \\ -i & -i \end{pmatrix}, \tag{3.30}$$

which is a nilpotent matrix. As in the noncompact case we will also use $Y_0 = iI$ to complete the basis of a $u(1,1)$ MASA. Old and new coordinates satisfy

$$y_0 = e^{ix_0}((1+ix_1)s_0 + ix_1 s_1), \quad y_1 = e^{ix_0}(-ix_1 s_0 + (1-ix_1)s_1). \tag{3.31}$$

The Hamiltonian is

$$H = c\left[\tfrac{1}{2}(p_{s_0}^2 - p_{s_1}^2) + V(s)\right], \quad V(s) = \frac{2m_0 m_1}{(s_0+s_1)^2} - \frac{m_1^2}{(s_0+s_1)^4}, \tag{3.32}$$

with the constraint (which is the same for all the subalgebras in the $su(1,1)$ case, as we are using the same metric), $s_0^2 - s_1^2 = 1$. The expression of the Hamiltonian in terms of the ϕ coordinate $(c = -1)$ is

TABLE I. One-dimensional potentials.

Algebra	MASA	Kinetic term	Potential
$su(2)$	Compact Cartan	p_ϕ^2	$\dfrac{m_0^2}{\cos^2\phi}+\dfrac{m_1^2}{\sin^2\phi}$
$su(1,1)$	Compact Cartan	p_ϕ^2	$-\dfrac{m_0^2}{\cosh^2\phi}+\dfrac{m_1^2}{\sinh^2\phi}$
	Noncompact Cartan	p_ϕ^2	$-\dfrac{m_0^2-m_1^2+2m_0m_1\sinh 2\phi}{\cosh^2 2\phi}$
	Nilpotent	p_ϕ^2	$m_1^2e^{-4\phi}-2m_0m_1e^{-2\phi}$

$$H(\phi)=\frac{1}{2}p_\phi^2+V(\phi),\quad V(\phi)=m_1^2e^{-4\phi}-2m_0m_1e^{-2\phi}. \tag{3.33}$$

The Hamiltonian in terms of the second order operators in the enveloping algebra is again ($\{X_i,X_j\}=X_iX_j+X_jX_i$)

$$\hat{Q}_1=2X_1^2+\{X_1,X_3\}-X_2^2 \tag{3.34}$$

and the trivial constant C_1 is equal to $(X_1+X_3)^2$, with $C=\hat{Q}_1-C_1$.

The HJ equation is

$$\frac{1}{2}\left(\frac{\partial S}{\partial\phi}\right)^2-2m_0m_1e^{-2\phi}+m_1^2e^{-4\phi}=E. \tag{3.35}$$

We will assume $m_1\neq 0$ to obtain nontrivial results. Depending on the constants m_0,m_1 we get essentially two classes of systems.

(1) If $m_0m_1\leqslant 0$ there is no extremum for the potential, and the energy is always positive. The solutions, with a unique turning point, are given by

$$e^{2\phi}=\frac{m_1}{E}[-m_0+\sqrt{E+m_0^2}\cosh\sqrt{2E}t]. \tag{3.36}$$

(2) If $m_0m_1>0$ the potential has a minimum, in $\phi=(1/2)\log(m_1/m_0)$ and the energy is bounded from below, $E>-m_0^2$. As in the first case, the values of the parameters are not essential if they satisfy the constraints. The solutions for the different values of the energy are

(i) $-m_0^2<E<0$

$$e^{2\phi}=\frac{m_1}{|E|}[m_0+\sqrt{E+m_0^2}\cos\sqrt{2E}t]. \tag{3.37}$$

(ii) $E=0$

$$e^{2\phi}=\frac{m_1}{2m_0}+4m_0m_1t^2. \tag{3.38}$$

If $E>0$ the solution is the same as (3.36).

This case completes the set of one-dimensional Hamiltonians obtained through a reduction procedure out of free Hamiltonians, invariant under SU(p,q), $p+q=2$, and defined over a homogeneous space of the corresponding group. In Table I, we present a summary of these

Hamiltonians in the one-dimensional case.

In the next section we will treat the two-dimensional Hamiltonians associated with the rank 2 algebras $su(3)$ and $su(2,1)$.

IV. THE TWO-DIMENSIONAL CASE

There are two pseudounitary algebras to be used to construct superintegrable Hamiltonians of dimension 2, $su(3)$ and $su(2,1)$. We will treat separately both cases.

A. $su(3)$

We will use as a basis for $su(3)$ the operators $\{X_1,\dots,X_8\}$ which are given in the 3×3 matrix representation by

$$X_1\to\begin{pmatrix} i & 0 & 0\\ 0 & -i & 0\\ 0 & 0 & 0\end{pmatrix},\quad X_2\to\begin{pmatrix} 0 & 0 & 0\\ 0 & i & 0\\ 0 & 0 & -i\end{pmatrix},\quad X_3\to\begin{pmatrix} 0 & 1 & 0\\ -1 & 0 & 0\\ 0 & 0 & 0\end{pmatrix},$$

$$X_4\to\begin{pmatrix} 0 & i & 0\\ i & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\quad X_5\to\begin{pmatrix} 0 & 0 & 1\\ 0 & 0 & 0\\ -1 & 0 & 0\end{pmatrix},\quad X_6\to\begin{pmatrix} 0 & 0 & i\\ 0 & 0 & 0\\ i & 0 & 0\end{pmatrix},$$

$$X_7\to\begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & 1\\ 0 & -1 & 0\end{pmatrix},\quad X_8\to\begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & i\\ 0 & i & 0\end{pmatrix},$$

when the metric is $K=\mathrm{diag}(1,1,1)$.

In the compact case there is only one MASA, the Cartan subalgebra, generated by the matrices

$$\begin{pmatrix} i & & \\ & -i & \\ & & 0\end{pmatrix},\quad \begin{pmatrix} 0 & & \\ & i & \\ & & -i\end{pmatrix}, \tag{4.1}$$

and we shall use the following basis for the corresponding MASA in $u(3)$:

$$Y_0=\begin{pmatrix} i & & \\ & 0 & \\ & & 0\end{pmatrix},\quad Y_1=\begin{pmatrix} 0 & & \\ & i & \\ & & 0\end{pmatrix},\quad Y_2=\begin{pmatrix} 0 & & \\ & 0 & \\ & & i\end{pmatrix}. \tag{4.2}$$

The coordinates s are related to the coordinates y in the same way as in $su(2)$, (3.3),

$$y_0=s_0e^{ix_0},\quad y_1=s_1e^{ix_1},\quad y_2=s_2e^{ix_2}, \tag{4.3}$$

and the Hamiltonian has also the same form of all cases using compact Cartan subalgebras (3.4),

$$H=\tfrac{1}{2}(p_0^2+p_1^2+p_2^2)+V(s),\quad V(s)=\frac{m_0^2}{s_0^2}+\frac{m_1^2}{s_1^2}+\frac{m_2^2}{s_2^2}, \tag{4.4}$$

with the constraint $s_0^2+s_1^2+s_2^2=1$.

In the one-dimensional case, there was only one invariant, which was the Hamiltonian. In this case, we can construct three invariants, only two of them in involution at the same time (one of them the Hamiltonian). The system is superintegrable in the sense of Ref. 12. These invariants are[14]

$$R_{01}=(s_0p_1-s_1p_0)^2+\left(m_0\frac{s_1}{s_0}+m_1\frac{s_0}{s_1}\right)^2,$$
$$R_{02}=(s_0p_2-s_2p_0)^2+\left(m_0\frac{s_2}{s_0}+m_2\frac{s_0}{s_2}\right)^2, \tag{4.5}$$
$$R_{12}=(s_1p_2-s_2p_1)^2+\left(m_1\frac{s_2}{s_1}+m_2\frac{s_1}{s_2}\right)^2.$$

The sum of these three invariants is the Hamiltonian up to an additive constant. In order to study the solutions of this problem we need construct a coordinate system in which the corresponding HJ equation separates into a system of ordinary differential equations. As in Ref. 16 we will use spherical coordinates,[19] defined by

$$s_0=\cos\phi_2\cos\phi_1,\quad s_1=\cos\phi_2\sin\phi_1,\quad s_2=\sin\phi_2, \tag{4.6}$$

and the Hamiltonian is written as $(c=1)$

$$H=\frac{1}{2}\left(p_{\phi_2}^2+\frac{p_{\phi_1}^2}{\cos^2\phi_2}\right)+V(\phi_1,\phi_2),$$
$$V(\phi_1,\phi_2)=\frac{1}{\cos^2\phi_2}\left(\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1}\right)+\frac{m_2^2}{\sin^2\phi_2}, \tag{4.7}$$

where the constants m_0, m_1, m_2 are chosen to be non-negative.

The second order conserved quantities (4.5) (we will follow the notation $\hat{Q}$ for these operators) can be written in terms of the basis $\{X_1, \ldots, X_8\}$,

$$\hat{Q}_1=X_3^2+X_4^2,\quad \hat{Q}_2=X_5^2+X_6^2,\quad \hat{Q}_3=X_7^2+X_8^2 \tag{4.8}$$

with commutation relations (the commutator is a third order element which plays no essential role in the method),

$$[\hat{Q}_1,\hat{Q}_2]=[\hat{Q}_2,\hat{Q}_3]=[\hat{Q}_3,\hat{Q}_1].$$

The Casimir is

$$C=4C_1+2C_2+4C_3+3\hat{Q}_1+3\hat{Q}_2+3\hat{Q}_3,$$

where

$$C_1=X_1^2,\quad C_2=\{X_1,X_2\},\quad C_3=X_2^2$$

are the second order operators in the enveloping algebra of the compact Cartan subalgebra.

The Hamiltonian is

$$H=Q_1+Q_2+Q_3+\text{constant}$$

where Q_i is the expression of $\hat{Q}_i$ in spherical coordinates,[19]

$$Q_1=\frac{1}{2}p_{\phi_1}^2+\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1},$$

$$Q_2=\tan^2\phi_2\left(\frac{1}{2}p_{\phi_1}^2\sin^2\phi_1+\frac{m_0^2}{\cos^2\phi_1}\right)+\cos^2\phi_1\left(\frac{1}{2}p_{\phi_2}^2+\frac{m_2^2}{\tan^2\phi_2}\right)$$
$$+\frac{1}{2}p_{\phi_1}p_{\phi_2}\sin 2\phi_1\tan\phi_2,$$

$$Q_3=\tan^2\phi_2\left(\frac{1}{2}p_{\phi_1}^2\cos^2\phi_1+\frac{m_1^2}{\sin^2\phi_1}\right)+\sin^2\phi_1\left(\frac{1}{2}p_{\phi_2}^2+\frac{m_2^2}{\tan^2\phi_2}\right)$$
$$-\frac{1}{2}p_{\phi_1}p_{\phi_2}\sin 2\phi_1\tan\phi_2.$$

The HJ equation is

$$\frac{1}{2}\left(\frac{\partial S}{\partial\phi_2}\right)^2+\frac{m_2^2}{\sin^2\phi_2}+\frac{1}{\cos^2\phi_2}\left(\frac{1}{2}\left(\frac{\partial S}{\partial\phi_1}\right)^2+\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1}\right)=E, \tag{4.9}$$

and separates into two ordinary differential equations using $S(\phi_1,\phi_2)=S_1(\phi_1)+S_2(\phi_2)-Et$,

$$\frac{1}{2}\left(\frac{\partial S_1}{\partial\phi_1}\right)^2+\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1}=\alpha_1, \tag{4.10}$$

$$\frac{1}{2}\left(\frac{\partial S_2}{\partial\phi_2}\right)^2+\frac{m_2^2}{\sin^2\phi_2}+\frac{\alpha_1}{\cos^2\phi_2}=\alpha_2, \tag{4.11}$$

where $\alpha_2=E$ and α_1 are the separation constants (which are positive). These equations have the same form as those in (3.7) The solutions are easily computed and can be found as particular cases in Ref. 16. The potential has singularities along the coordinate lines, $\phi_1=0,\pi/2,\pi,3\pi/2$, and $\phi_2=\pi/2,3\pi/2$ in the generic case. It has a unique minimum inside each regularity domain. An analysis of the associated dynamical system (Hamilton equations) shows that all the orbits in a neighborhood of the critical point (center) are closed and hence, the corresponding trajectories are periodic (a direct consequence of the correspondence between extrema of the potential and critical points of the phase space). Let us restrict to the domain $0<\phi_1,\phi_2<\pi/2$, where the minimum is $\tan\phi_1=\sqrt{m_1/m_0}$, $\tan\phi_2=\sqrt{m_2/(m_0+m_1)}$. The value of the potential at this point is $(m_0+m_1+m_2)^2$, hence the energy E is bounded from below ($E\geqslant(m_0+m_1+m_2)^2$). The explicit solutions are

$$\cos^2\phi_2=\frac{1}{2E}[b_2+\sqrt{b_2^2-4\alpha_1E}\cos 2\sqrt{2E}t], \tag{4.12}$$

$$\cos^2\phi_1=\frac{1}{2\alpha_1}\left[b_1+\frac{1}{\cos^2\phi_2}\left[\frac{b_1^2-4\alpha_1m_0^2}{b_2^2-4\alpha_1E}\right]^{1/2}((b_2\cos^2\phi_2-2\alpha_1)\sin 2\sqrt{2\alpha_1}\beta_1\right.$$
$$\left.+2\sqrt{\alpha_1}[(b_2-E\cos^2\phi_2)\cos^2\phi_2-\alpha_1]^{1/2}\cos 2\sqrt{2\alpha_1}\beta_1)\right], \tag{4.13}$$

where $b_1=\alpha_1+m_0^2-m_1^2$ and $b_2=E+\alpha_1-m_2^2$.

Let us remark that these results reflect essentially the case $su(2)$. In fact all systems we can construct using Cartan subalgebras can be described in a unified way as it was shown in Refs. 14 and 16.

B. $su(2,1)$

The basis we will use is formed by the set of operators $\{X_1, \ldots, X_8\}$ which are given in the 3×3 matrix representation by

$$X_1 \to \begin{pmatrix} i & 0 & 0 \\ 0 & -i & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad X_2 \to \begin{pmatrix} 0 & 0 & 0 \\ 0 & i & 0 \\ 0 & 0 & -i \end{pmatrix}, \quad X_3 \to \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$X_4 \to \begin{pmatrix} 0 & i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad X_5 \to \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad X_6 \to \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{pmatrix},$$

$$X_7 \to \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad X_8 \to \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & i \\ 0 & -i & 0 \end{pmatrix}.$$

According to general results,[22] $su(2,1)$ has four MASAs, two of them Cartan subalgebras (compact and noncompact), one orthogonally decomposable subalgebra, with one nilpotent element, and one nilpotent subalgebra. We will discuss these four cases in the following. Although some of these subalgebras have a simpler expression in some skew diagonal metrics, we will always use the diagonal one,

$$K = \begin{pmatrix} 1 & & \\ & 1 & \\ & & -1 \end{pmatrix} \tag{4.14}$$

because, in this way, the kinetic part is always diagonal. There are nine coordinate systems associated with $O(2,1)$ free Hamiltonians, spherical, hyperbolic, elliptic (I and II), complex elliptic, horospheric, elliptic parabolic, hyperbolic parabolic, and semicircular parabolic.[20,23] Not all of them will separate our systems because these are not free. However, the appropriate systems have been computed in Ref. 20 and we will use their results.

1. Compact Cartan subalgebra (CC)

The compact Cartan subalgebra has a basis formed by the same two matrices we used in $su(3)$, and the same situation happens for the corresponding MASA in $u(2,1)$ (4.2). The coordinates s are also related to the coordinates y as they did in the compact case $su(3)$ (4.3).

However, the Hamiltonian reflects the noncompact character of $su(2,1)$

$$H = c\left(\tfrac{1}{2}(p_0^2 + p_1^2 - p_2^2) + V(s)\right), \quad V(s) = \frac{m_0^2}{s_0^2} + \frac{m_1^2}{s_1^2} - \frac{m_2^2}{s_2^2}, \tag{4.15}$$

where the constraint $s_0^2 + s_1^2 - s_2^2 = 1$ must be satisfied.

This Hamiltonian separates in four coordinate systems, spherical, hyperbolic, and elliptic I and II.[20] We will use spherical coordinates to discuss the explicit solution,

$$s_0 = \cosh\phi_2 \cos\phi_1, \quad s_1 = \cosh\phi_2 \sin\phi_1, \quad s_2 = \sinh\phi_2. \tag{4.16}$$

Choosing $c = -1$, we have the Hamiltonian in these coordinates,

$$H=\frac{1}{2}\left(p_{\phi_2}^2-\frac{p_{\phi_1}^2}{\cosh^2\phi_2}\right)+V(\phi_1,\phi_2),$$
$$V(\phi_1,\phi_2)=-\frac{1}{\cosh^2\phi_2}\left(\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1}\right)+\frac{m_2^2}{\sinh^2\phi_2}. \tag{4.17}$$

Due to the form of the potential the constants m_0, m_1, m_2 can be chosen to be non-negative. The potential is regular inside the domain $0<\phi_1<\pi/2$, $0<\phi_2<\infty$. It has a saddle point, $\tan\phi_1=\sqrt{m_1/m_0}$, $\tanh\phi_2=\sqrt{m_2/(m_0+m_1)}$, if $m_0+m_1>m_2$. However, due to the special form of the kinetic term (which is not positive definite), it is easy to check that the associated dynamical system has all the orbits in a neighborhood of the critical point (which is also a center as in the compact case) closed and again, the corresponding trajectories are periodic.

The second order operators in the enveloping algebra of this MASA are

$$C_1=X_1^2,\quad C_2=\{X_1,X_2\},\quad C_3=X_2^2.$$

The quadratic constants of motion lying in the enveloping algebra of $su(2,1)$ and commuting with the elements in the compact Cartan subalgebra are

$$\hat{Q}_1=X_3^2+X_4^2,\quad \hat{Q}_2=X_5^2+X_6^2,\quad \hat{Q}_3=X_7^2+X_8^2 \tag{4.18}$$

with commutation relations

$$[\hat{Q}_1,\hat{Q}_2]=[\hat{Q}_3,\hat{Q}_2]=-[\hat{Q}_3,\hat{Q}_1].$$

The Casimir is written in terms of these second order operators as

$$C=4C_1+2C_2+4C_3+3\hat{Q}_1-3\hat{Q}_2-3\hat{Q}_3,$$

and the Hamiltonian is

$$H=-Q_1+Q_2+Q_3+\text{constant}$$

where Q_i are the conserved quantities in spherical coordinates,

$$Q_1=\frac{1}{2}p_{\phi_1}^2+\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1},$$

$$Q_2=\tanh^2\phi_2\left(\frac{1}{2}p_{\phi_1}^2\sin^2\phi_1+\frac{m_0^2}{\cos^2\phi_1}\right)+\cos^2\phi_1\left(\frac{1}{2}p_{\phi_2}^2+\frac{m_2^2}{\tanh^2\phi_2}\right)$$
$$-\frac{1}{2}p_{\phi_1}p_{\phi_2}\sin 2\phi_1\tanh\phi_2,$$

$$Q_3=\tanh^2\phi_2\left(\frac{1}{2}p_{\phi_1}^2\cos^2\phi_1+\frac{m_1^2}{\sin^2\phi_1}\right)+\sin^2\phi_1\left(\frac{1}{2}p_{\phi_2}^2+\frac{m_2^2}{\tanh^2\phi_2}\right)$$
$$+\frac{1}{2}p_{\phi_1}p_{\phi_2}\sin 2\phi_1\tanh\phi_2.$$

The HJ equations corresponding to the Hamiltonian (4.17) are

$$\frac{1}{2}\left(\frac{\partial S_1}{\partial \phi_1}\right)^2+\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1}=\alpha_1, \tag{4.19}$$

$$\frac{1}{2}\left(\frac{\partial S_2}{\partial \phi_2}\right)^2+\frac{m_2^2}{\sinh^2\phi_2}-\frac{\alpha_1}{\cosh^2\phi_2}=\alpha_2. \tag{4.20}$$

The first one is the same as we got in $su(3)$ (4.10), α_1 is always positive and $\alpha_2=E$.

The solutions depend on the values of the parameters and energy

(i) $E<0$

$$u_2=\frac{1}{2|E|}\left[-b_2+\sqrt{b_2^2+4\alpha_1 E}\cos 2\sqrt{2|E|}t\right]. \tag{4.21}$$

(ii) $E=0$

$$u_2=\frac{\alpha_1}{\alpha_1-m_2^2}+2(\alpha_1-m_2^2)t^2. \tag{4.22}$$

(iii) $E>0$

$$u_2=\frac{1}{2E}\left[b_2+\sqrt{b_2^2+4\alpha_1 E}\cosh 2\sqrt{2E}t\right], \tag{4.23}$$

where $u_2=\cosh^2\phi_2$, $b_2=E-\alpha_1+m_2^2$. The other equation can be solved as we did in the previous cases. The result is

$$u_1=\frac{1}{2\alpha_1}\left[b_1+\frac{1}{u_2}\left[\frac{b_1^2-4\alpha_1 m_0^2}{b_2^2+4\alpha_1 E}\right]^{1/2}(-(b_2u_2+2\alpha_1)\sin 2\sqrt{2\alpha_1}\beta_1\right.$$
$$\left.+2\sqrt{\alpha_1}[(Eu_2-b_2)u_2-\alpha_1]^{1/2}\cos 2\sqrt{2\alpha_1}\beta_1)\right], \tag{4.24}$$

where $u_1=\cos^2\phi_1$ and $b_1=\alpha_1+m_0^2-m_1^2$.

2. Noncompact Cartan subalgebra (NC)

There is only one noncompact Cartan subalgebra. A representative can be chosen according to the same criteria we used in $su(1,1)$ (3.19), keeping one element compact and the other (as in $su(1,1)$) noncompact,

$$\begin{pmatrix} 2i & & \\ & -i & \\ & & -i \end{pmatrix},\quad \begin{pmatrix} 0 & & \\ & 0 & i \\ & -i & 0 \end{pmatrix}, \tag{4.25}$$

and the basis for the corresponding MASA of $u(2,1)$ will be

$$Y_0=\begin{pmatrix} i & & \\ & 0 & \\ & & 0 \end{pmatrix},\quad Y_1=\begin{pmatrix} 0 & & \\ & i & \\ & & i \end{pmatrix},\quad Y_2=\begin{pmatrix} 0 & & \\ & 0 & i \\ & -i & 0 \end{pmatrix}. \tag{4.26}$$

The coordinates are as in $su(1,1)$,

$$y_0=e^{ix_0}s_0,$$

$$y_1 = e^{ix_1}(s_1 \cosh x_2 + i s_2 \sinh x_2), \tag{4.27}$$
$$y_2 = e^{ix_1}(-i s_1 \sinh x_2 + s_2 \cosh x_2).$$

The Hamiltonian is

$$H = c\left(\tfrac{1}{2}(p_0^2 + p_1^2 - p_2^2) + V(s)\right),$$
$$V(s) = \frac{m_0^2}{s_0^2} + \frac{(m_1^2 - m_2^2)(s_1^2 - s_2^2) + 4 m_1 m_2 s_1 s_2}{(s_1^2 + s_2^2)^2}, \tag{4.28}$$

where we will take $m_0 > 0$ and m_1, m_2 can take any value, and the coordinates satisfy the constraint (the same for all $su(2,1)$ MASAs, as we have chosen the same metric in all cases), $s_0^2 + s_1^2 - s_2^2 = 1$.

There are two systems of coordinates in which the associated HJ equation separates, hyperbolic and complex elliptic.[20] We will use hyperbolic coordinates, defined as

$$s_0 = \cosh \phi_2, \qquad s_1 = \sinh \phi_2 \sinh \phi_1, \qquad s_2 = \sinh \phi_2 \cosh \phi_1, \tag{4.29}$$

and the new Hamiltonian is $(c = -1)$

$$H = \frac{1}{2}\left(p_{\phi_2}^2 - \frac{p_{\phi_1}^2}{\sinh^2 \phi_2} \right) + V(\phi_1, \phi_2),$$
$$V(\phi_1, \phi_2) = -\frac{m_0^2}{\cosh^2 \phi_2} + \frac{1}{\sinh^2 \phi_2}\left(\frac{m_1^2 - m_2^2 - 2 m_1 m_2 \sinh 2\phi_1}{\cosh^2 2\phi_1} \right). \tag{4.30}$$

Note that the potential follows the same pattern as the corresponding case in $su(1,1)$. It is regular inside the domain, $-\infty < \phi_1 < \infty$, $0 < \phi_2 < \infty$, and has also a saddle point at $\sinh 2\phi_1 = -m_2/m_1$, $\tanh \phi_2 = \sqrt{|m_1/m_0|}$, when $|m_1| < |m_0|$. As in the previous case, the associated dynamical system has a center and the trajectories in a neighborhood of it are periodic.

The basis for this MASA is $\{2X_1 + X_2, X_8\}$, and the corresponding second order elements are

$$C_1 = (2X_1 + X_2)^2, \quad C_2 = \{2X_1 + X_2, X_8\}, \quad C_3 = X_8^2.$$

The second order conserved quantities, commuting with $2X_1 + X_2$ and X_8, and belonging to the enveloping algebra of $su(2,1)$ are

$$\hat{Q}_1 = X_2^2 - X_7^2,$$
$$\hat{Q}_2 = X_3^2 + X_4^2 - X_5^2 - X_6^2, \tag{4.31}$$
$$\hat{Q}_3 = \{X_3, X_5\} + \{X_4, X_6\},$$

with commuting relations,

$$[\hat{Q}_3, \hat{Q}_1] = [\hat{Q}_2, \hat{Q}_3], \;\; [\hat{Q}_1, \hat{Q}_2] = 0.$$

The Casimir is written as

$$C = C_1 - 3C_3 + 3\hat{Q}_1 + 3\hat{Q}_2,$$

and the Hamiltonian is

$$H = Q_1 + Q_2 + \text{constant}.$$

Finally, the conserved quantities are expressed in hyperbolic coordinates by

$$Q_1 = \frac{1}{2} p_{\phi_1}^2 - \frac{m_1^2 - m_2^2 - 2m_1 m_2 \sinh 2\phi_1}{\cosh^2 2\phi_1},$$

$$Q_2 = \frac{1}{2} p_{\phi_2}^2 - \frac{m_0^2}{\cosh^2 \phi_2} - \frac{1}{\tanh^2 \phi_2}\left(\frac{1}{2} p_{\phi_1}^2 - \frac{m_1^2 - m_2^2 - 2m_1 m_2 \sinh 2\phi_1}{\cosh^2 2\phi_1}\right),$$

$$Q_3 = \frac{1}{2} \sinh 2\phi_1 \left(p_{\phi_2}^2 + \frac{1}{\tanh^2 \phi_2} p_{\phi_1}^2\right) - \frac{\cosh 2\phi_1}{\tanh \phi_2} p_{\phi_1} p_{\phi_2}$$
$$+ m_0^2 \tanh^2 \phi_2 \sinh 2\phi_1 - \frac{(m_1^2 - m_2^2)\sinh 2\phi_1 + 2m_1 m_2}{\tanh^2 \phi_2 \cosh^2 2\phi_1}.$$

The Hamilton–Jacobi equation separates into two ordinary differential equations,

$$\frac{1}{2}\left(\frac{\partial S_1}{\partial \phi_1}\right)^2 - \frac{m_1^2 - m_2^2 - 2m_1 m_2 \sinh 2\phi_1}{\cosh^2 2\phi_1} = \alpha_1, \tag{4.32}$$

$$\frac{1}{2}\left(\frac{\partial S_2}{\partial \phi_2}\right)^2 - \frac{m_0^2}{\cosh^2 \phi_2} - \frac{\alpha_1}{\sinh^2 \phi_2} = E. \tag{4.33}$$

The solutions have the same form we have found before.

(i) $E<0$

$$u_2 = \frac{1}{2|E|}\left[b_2 + \sqrt{b_2^2 - 4\alpha_1 E} \cos 2\sqrt{2|E|}t\right]. \tag{4.34}$$

(ii) $E=0$

$$u_2 = -\frac{\alpha_1}{\alpha_1 + m_0^2} + 2(\alpha_1 + m_0^2)t^2. \tag{4.35}$$

(iii) $E>0$

$$u_2 = \frac{1}{2E}\left[-b_2 + \sqrt{b_2^2 - 4\alpha_1 E} \cosh 2\sqrt{2E}t\right], \tag{4.36}$$

where $u_2 = \sinh^2 \phi_2$, $b_2 = E + \alpha_1 + m_0^2$.

The solution for the other coordinate is obtained in the same way (with the change $u_1 = \sinh 2\phi_1$). The equation for this coordinate is the same as that given in formula (3.24) and its solutions can be found in formulas (3.25)–(3.29). Due to the possible different signs of the energy and the constant α_1, one should take care of the square roots appearing in all the formulas.

3. Orthogonally decomposable subalgebra (OD)

The orthogonally decomposable subalgebra (a representative of the class) has a basis formed by a compact element and a nilpotent one,

$$\begin{pmatrix} 2i & & \\ & -i & \\ & & -i \end{pmatrix}, \quad \begin{pmatrix} 0 & & \\ & i & i \\ & -i & -i \end{pmatrix}, \tag{4.37}$$

and the basis for the corresponding MASA of $u(2,1)$ is

$$Y_0 = \begin{pmatrix} i & & \\ & 0 & \\ & & 0 \end{pmatrix}, \quad Y_1 = \begin{pmatrix} 0 & & \\ & i & \\ & & i \end{pmatrix}, \quad Y_2 = \begin{pmatrix} 0 & & \\ & i & i \\ & -i & -i \end{pmatrix}. \tag{4.38}$$

The coordinates have also a similar form to those in $su(1,1)$ (3.31)

$$\begin{aligned} y_0 &= e^{ix_0} s_0, \\ y_1 &= e^{ix_1}((1+ix_2)s_1 + ix_2 s_2), \\ y_2 &= e^{ix_1}(-ix_2 s_1 + (1-ix_2)s_2). \end{aligned} \tag{4.39}$$

The Hamiltonian is

$$\begin{aligned} H &= c\left(\tfrac{1}{2}(p_0^2 + p_1^2 - p_2^2) + V(s)\right), \\ V(s) &= \frac{m_0^2}{s_0^2} - \frac{m_2^2(s_1 - s_2)}{(s_1+s_2)^3} + \frac{2m_1 m_2}{(s_1+s_2)^2} \end{aligned} \tag{4.40}$$

with $s_0^2 + s_1^2 - s_2^2 = 1$.

There are four coordinate systems associated to this subalgebra, hyperbolic, horospheric, elliptic parabolic, and hyperbolic parabolic.[20] We will use again the hyperbolic ones, defined as in (4.29).

The Hamiltonian is $(c=-1)$

$$\begin{aligned} H &= \frac{1}{2}\left(p_{\phi_2}^2 - \frac{p_{\phi_1}^2}{\sinh^2 \phi_2} \right) + V(\phi_1, \phi_2), \\ V(\phi_1, \phi_2) &= -\frac{m_0^2}{\cosh^2 \phi_2} - \frac{1}{\sinh^2 \phi_2}(m_2^2 e^{-4\phi_1} + 2m_1 m_2 e^{-2\phi_1}). \end{aligned} \tag{4.41}$$

The potential is regular inside the domain, $-\infty < \phi_1 < \infty$, $0 < \phi_2 < \infty$, and, as it happens in all the $su(2,1)$ cases, has a saddle point at $\phi_1 = (1/2)\log(|m_2/m_1|)$, $\tanh \phi_2 = \sqrt{|m_1/m_0|}$, when $|m_0| > |m_1|, m_1 m_2 < 0$. The situation is the same as in all other cases in $su(2,1)$.

The second order operators in the enveloping algebra of the MASA under consideration are given by

$$C_1 = (2X_1 + X_2)^2, \quad C_2 = \{2X_1 + X_2, X_2 + X_8\}, \quad C_3 = (X_2 + X_8)^2,$$

and the quadratic constants of motion,

$$\begin{aligned} \hat{Q}_1 &= X_3^2 + X_4^2 - X_5^2 - X_6^2, \\ \hat{Q}_2 &= (X_3 + X_5)^2 + (X_4 + X_6)^2, \end{aligned} \tag{4.42}$$

$$\hat{Q}_3 = X_7^2 + 2\{X_1, X_2 + X_8\},$$

satisfy the commutation relations

$$[\hat{Q}_1, \hat{Q}_2] = [\hat{Q}_2, \hat{Q}_3], \quad [\hat{Q}_1, \hat{Q}_3] = 0.$$

The Casimir is given in terms of these operators by

$$C = C_1 + 3C_2 - 3C_3 + 3\hat{Q}_1 - 3\hat{Q}_3$$

and the Hamiltonian is

$$H = -Q_1 + Q_3 + \text{constant}.$$

We can write the conserved quantities in hyperbolic coordinates,

$$Q_1 = \frac{1}{\tanh^2 \phi_2}\left(\frac{1}{2} p_{\phi_1}^2 + m_2^2 e^{-4\phi_1} + 2 m_1 m_2 e^{-2\phi_1}\right) - \left(\frac{1}{2} p_{\phi_2}^2 - \frac{m_0^2}{\cosh^2 \phi_2}\right),$$

$$Q_2 = e^{2\phi_1}\left(\frac{1}{2} p_{\phi_2}^2 + m_0^2 \tanh^2 \phi_2 + \frac{1}{\tanh^2 \phi_2}\left(\frac{1}{2} p_{\phi_1}^2 + m_2^2 e^{-4\phi_1}\right) - \frac{1}{\tanh \phi_2} p_{\phi_1} p_{\phi_2}\right),$$

$$Q_3 = \frac{1}{2} p_{\phi_1}^2 + m_2^2 e^{-4\phi_1} + 2 m_1 m_2 e^{-2\phi_2}.$$

The HJ equation is separated into the following equations:

$$\frac{1}{2}\left(\frac{\partial S_1}{\partial \phi_1}\right)^2 + (m_2^2 e^{-4\phi_1} + 2 m_1 m_2 e^{-2\phi_1}) = \alpha_1, \tag{4.43}$$

$$\frac{1}{2}\left(\frac{\partial S_2}{\partial \phi_2}\right)^2 - \frac{m_0^2}{\cosh^2 \phi_2} - \frac{\alpha_1}{\sinh^2 \phi_2} = E. \tag{4.44}$$

Equation (4.44) is integrated using $u_2 = \sinh^2 \phi_2$. The result is the same as in the previous case (4.33). Equation (4.43) is solved using $u_1 = e^{2\phi_1}$ (the same change we use in the nilpotent MASA of the $su(1,1)$ case), and the results are essentially the same we have found above (see 3.35).

4. Nilpotent subalgebra (NIL)

The nilpotent subalgebra has a basis formed by two nilpotent elements (one of order 2 and the other of order 3),

$$\begin{pmatrix} 0 & & \\ & i & i \\ & -i & -i \end{pmatrix}, \quad \begin{pmatrix} 0 & i & i \\ i & 0 & 0 \\ -i & 0 & 0 \end{pmatrix}, \tag{4.45}$$

and the basis for the MASA of $u(2,1)$ can be obtained adding to these two matrices the identity times the imaginary unit.

The new coordinates are defined through

$$y_0 = e^{ix_0}(s_0 + i x_2(s_1 + s_2)),$$

$$y_1 = e^{ix_0}\left(ix_2 s_0 + \left(1 - \frac{x_2^2}{2} + ix_1\right)s_1 + \left(-\frac{x_2^2}{2} + ix_1\right)s_2\right), \tag{4.46}$$

$$y_2 = e^{ix_0}\left(-ix_2 s_0 + \left(\frac{x_2^2}{2} - ix_1\right)s_1 + \left(1 + \frac{x_2^2}{2} - ix_1\right)s_2\right).$$

The Hamiltonian is

$$H = C(\tfrac{1}{2}(p_0^2 + p_1^2 - p_2^2) + V(s)),$$
$$V(s) = \frac{2m_0 m_1 + m_2^2}{(s_1+s_2)^2} - \frac{4m_1 m_2 s_0}{(s_1+s_2)^3} + \frac{m_1^2(4s_0^2-1)}{(s_1+s_2)^4}, \tag{4.47}$$

with the constraint, $s_0^2 + s_1^2 - s_2^2 = 1$.

We have now two separable coordinate systems: horospheric and semicircular parabolic.[20] We will use now the horospheric ones, defined by

$$s_0 = \phi_1 e^{\phi_2}, \quad s_1 = \cosh \phi_2 - \tfrac{1}{2}\phi_1^2 e^{\phi_2}, \quad s_2 = \sinh \phi_2 + \tfrac{1}{2}\phi_1^2 e^{\phi_2}. \tag{4.48}$$

The Hamiltonian is $(c = -1)$

$$H = \tfrac{1}{2}(p_{\phi_2}^2 - e^{-2\phi_2} p_{\phi_1}^2) + V(\phi_1, \phi_2),$$
$$V(\phi_1,\phi_2) = m_1^2 e^{-4\phi_2} - e^{-2\phi_2}(m_2^2 + 2m_0 m_1 + 4m_1\phi_1(m_1\phi_1 - m_2)). \tag{4.49}$$

The potential has no singularity in the whole plain (ϕ_1, ϕ_2). It has a saddle point at $\phi_1 = m_2/2m_1$, $\phi_2 = (1/2)\log(m_1/m_0)$, when $m_0 m_1 > 0$. The situation is the same as in all other cases in $su(2,1)$.

The second order elements in the enveloping algebra of the nilpotent subalgebra are

$$C_1 = (X_2 + X_8)^2, \quad C_2 = \{X_2 + X_8, X_4 + X_6\}, \quad C_3 = (X_4 + X_6)^2,$$

and the constants of motion

$$\hat{Q}_1 = 3(X_3 + X_5)^2 - 2\{2X_1 + X_2, X_2 + X_8\},$$
$$\hat{Q}_2 = \{2X_1 + X_2, X_4 + X_6\} + 6\{X_4, X_2 + X_8\} - 3\{X_7, X_3 + X_5\}, \tag{4.50}$$
$$\hat{Q}_3 = 4X_1^2 + 3X_2^2 - 2\{X_1, X_2\} + 6X_3^2 + 6X_4^2 - 3X_7^2 - \{4X_1 - X_2, X_8\} + 3\{X_3, X_5\} + 3\{X_4, X_6\},$$

have the following commutation relations:

$$[\hat{Q}_1, \hat{Q}_2] = [\hat{Q}_3, \hat{Q}_2], \quad [\hat{Q}_1, \hat{Q}_3] = 0.$$

The Casimir is

$$C = -3C_1 - 3C_3 - \hat{Q}_1 + \hat{Q}_3,$$

and the Hamiltonian

$$H = Q_1 - Q_3 + \text{constant}.$$

Finally, the second order constant of motion are given in horospheric coordinates by the following expressions:

$$Q_1=\tfrac{1}{2}p_{\phi_1}^2+4m_1\phi_1(m_1\phi_1-m_2),$$

$$Q_2=\tfrac{1}{2}\phi_1 p_{\phi_1}^2-\tfrac{1}{2}p_{\phi_1}p_{\phi_2}+(m_2^2+2m_0m_1)\phi_1$$
$$-m_1e^{-2\phi_2}(2m_1\phi_1-m_2)+4m_1\phi_1^2(m_1\phi_1-m_2),$$

$$Q_3=(1+e^{-2\phi_2})(\tfrac{1}{2}p_{\phi_1}^2+4m_1\phi_1(m_1\phi_1-m_2))$$
$$-(\tfrac{1}{2}p_{\phi_2}^2+m_1^2e^{-4\phi_2}-(m_2^2+2m_0m_1)e^{-2\phi_2}).$$

This is the most interesting case, in the sense that the others are easily reduced to the cases in dimension 1, while this nilpotent subalgebra does not appear in the $su(1,1)$ case. However, the solutions are still very similar to those found before. It is worth mentioning here, that all the potentials we have constructed (and any potential we could construct by using this method) are always inverse quadratic potentials in the coordinates, and the solutions have always similar forms (though they depend on the specific characteristics of these potentials and the constants involved).

The HJ equation is separated into the following equations:

$$\frac{1}{2}\left(\frac{\partial S_1}{\partial\phi_1}\right)^2+m_2^2+2m_0m_1+4m_1\phi_1(m_1\phi_1-m_2)=\alpha_1, \tag{4.51}$$

$$\frac{1}{2}\left(\frac{\partial S_2}{\partial\phi_2}\right)^2+m_1^2e^{-4\phi_2}-\alpha_1e^{-2\phi_2}=E. \tag{4.52}$$

The change $u_2=e^{2\phi_2}$ allows us to solve the second equation, and the other one is solved directly. The solutions are

(i) $E<0$

$$u_2=\frac{1}{2|E|}[\alpha_1+\sqrt{\alpha_1^2+4m_1^2E}\cos 2\sqrt{2|E|}t]. \tag{4.53}$$

(ii) $E=0$

$$u_2=\frac{m_1^2}{\alpha_1}+2\alpha_1t^2. \tag{4.54}$$

(iii) $E>0$

$$u_2=\frac{1}{2E}[-\alpha_1+\sqrt{\alpha_1^2+4m_1^2E}\cosh 2\sqrt{2E}t]. \tag{4.55}$$

The first equation gives the value of the ϕ_1 coordinate,

$$\phi_1=\frac{1}{2m_1}\left[m_2+\frac{1}{u_2}\left[\frac{\alpha_1-2m_0m_1}{\alpha_1^2+4m_1^2E}\right]^{1/2}((\alpha_1u_2-2m_1^2)\sin 2\sqrt{2}\beta_1m_1\right.$$
$$\left.+2m_1[(Eu_2-\alpha_1)u_2-m_1^2]^{1/2}\cos 2\sqrt{2}\beta_1m_1)\right]. \tag{4.56}$$

In Table II, we present a summary of these Hamiltonians in the two-dimensional case.

TABLE II. Two-dimensional potentials.

Algebra	MASA	Kinetic term	Potential
$su(3)$	CC	$\frac{1}{2}\left(p_{\phi_2}^2+\frac{p_{\phi_1}^2}{\cos^2\phi_2}\right)$	$\frac{1}{\cos^2\phi_2}\left(\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1}\right)+\frac{m_2^2}{\sin^2\phi_2}$
$su(2,1)$	CC	$\frac{1}{2}\left(p_{\phi_2}^2-\frac{p_{\phi_1}^2}{\cosh^2\phi_2}\right)$	$-\frac{1}{\cosh^2\phi_2}\left(\frac{m_0^2}{\cos^2\phi_1}+\frac{m_1^2}{\sin^2\phi_1}\right)+\frac{m_2^2}{\sinh^2\phi_2}$
	NC	$\frac{1}{2}\left(p_{\phi_2}^2-\frac{p_{\phi_1}^2}{\sinh^2\phi_2}\right)$	$-\frac{m_0^2}{\cosh^2\phi_2}+\frac{1}{\sinh^2\phi_2}\left(\frac{m_1^2-m_2^2-2m_1m_2\sinh 2\phi_1}{\cosh^2 2\phi_1}\right)$
	OD	$\frac{1}{2}\left(p_{\phi_2}^2-\frac{p_{\phi_1}^2}{\sinh^2\phi_2}\right)$	$-\frac{m_0^2}{\cosh^2\phi_2}-\frac{1}{\sinh^2\phi_2}(m_2^2e^{-4\phi_1}+2m_1m_2e^{-2\phi_1})$
	NIL	$\frac{1}{2}(p_{\phi_2}^2-e^{-2\phi_2}p_{\phi_1}^2)$	$m_1^2e^{-4\phi_2}-e^{-2\phi_2}(m_2^2+2m_0m_1+4m_1\phi_1(m_1\phi_1-m_2))$

V. CONCLUSIONS

We have presented in this work a complete analysis of a series of one- and two-dimensional integrable Hamiltonians, which in the two-dimensional case are superintegrable in the sense described in the Introduction. Though the one-dimensional case is always an integrable system, let us remark the importance and applications of the potentials described in Sec. II. Regarding the two-dimensional ones, we have provided them with a set of conserved quantities which allows us to study the HJ equations in several separable coordinate systems and compute in some interesting cases the explicit solutions.

The one-dimensional Hamiltonians obtained here have been extensively studied in the literature from many other points of view. See for instance Ref. 24 for a recent application of Morse potentials. As an example of these different approaches, all of them appear in the classification of quasi-exactly solvable Schrödinger operators,[17,18] as particular types of these systems corresponding to the so called exactly solvable systems. Following the classification in Ref. 18, the exactly solvable potentials of Cases 1 and 2 are just the ones we have obtained associated to $su(1,1)$ and its compact and noncompact Cartan subalgebras. The first one (3.12) is the celebrated Pöschl–Teller potential. That appearing in case 3 is the Morse potential (3.33), which we get using the nilpotent subalgebra of $su(1,1)$. Finally the potential (3.5), associated to the Cartan subalgebra of $su(2)$ is related to the modified harmonic oscillators appearing in Ref. 18 as cases 4 and 5. One has to take into account in this case, that QES potentials, as studied in Ref. 18, are defined in the line (or half-line), and we are working here in a sector of S^1 (see also Ref. 25 for a study of harmonic oscillators in a sector). The relation is not surprising at all if one considers that QES systems in the line are related to the complex Lie algebra $sl(2)$,[17] and the ones we get here reflect the invariance under $su(2)$ and $su(1,1)$, the real forms of $sl(2)$. In the QES setting, Schrödinger operators belong to the enveloping algebra of a Lie algebra, while in our approach, the corresponding classical Hamiltonians are second order Casimirs of the algebra, and hence, they are particular cases (exactly solvable) of the former.

Two prolongations of this study are now in progress. One of them is the use of contractions in Lie algebras to obtain other Hamiltonian systems associated to different algebras, not necessarily semisimple. In this sense, the Hamiltonians, the conserved quantities and the coordinate systems can be obtained by contraction.[26,27] The second one is to apply this approach to the quantum case, considering the Schrödinger equation with these potentials.[28] We also plan to study the links of this theory with QES systems and the possibility of considering partial integrability and partial variable separation in HJ equations.[29]

ACKNOWLEDGMENTS

The authors would like to thank J. Negro for many interesting discussions. This work has been partially supported by CICYT (Spain, Projects PB95-0719 and PB95-0401) and Junta de Castilla y León (Spain).

[1] P. M. Morse, Phys. Rev. **34**, 57 (1929).
[2] G. Pöschl and E. Teller, Z. Phys. **83**, 143 (1933).
[3] F. Calogero, J. Math. Phys. **10**, 2191 (1969); **10**, 2197 (1969).
[4] D. Kazhdan, B. Kostant, and S. Sternberg, Commun. Pure Appl. Math. **31**, 481 (1978).
[5] J. Moser, *Various Aspects of Integrable Hamiltonian Systems, Dynamical Systems, CIME Lectures*, Bressanone, Italia, 1978, Progress in Mathematics (Birkhauser, Boston, 1980), Vol. 86.
[6] B. Sutherland, Phys. Rev. A **4**, 2019 (1971).
[7] J. Hietarinta, Phys. Rep. **147**, 87 (1987).
[8] A. M. Perelomov, *Integrable Systems of Classical Mechanics and Lie Algebras* (Birkhäuser, Basel, 1991).
[9] J. Marsden and A. Weinstein, Rep. Math. Phys. **5**, 121 (1974).
[10] J. Grabowski, G. Landi, G. Marmo, and G. Vilasi, Fortschr. Phys. **42**, 393 (1996).
[11] N. W. Evans, Phys. Rev. A **41**, 5666 (1990); Phys. Lett. A **147A**, 483 (1990); J. Math. Phys. **32**, 3369 (1991).
[12] L. P. Eisenhart, Ann. Math. **35**, 284 (1934); Phys. Rev. **74**, 87 (1948).
[13] W. Miller, Jr., J. Patera, and P. Winternitz, J. Math. Phys. **22**, 251 (1981).
[14] M. A. de Olmo, M. A. Rodriquez, and P. Winternitz, J. Math. Phys. **34**, 5118 (1993).
[15] M. A. de Olmo, M. A. Rodríguez, and P. Winternitz, Fortschr. Phys. **44**, 91 (1996).
[16] J. A. Calzada, M. A. de Olmo, and M. A. Rodriquez, J. Geom. Phys. **23**, 14 (1997).
[17] A. V. Turbiner, Commun. Math. Phys. **118**, 467 (1988).
[18] A. González-López, N. Kamran, and P. J. Olver, Commun. Math. Phys. **153**, 117 (1993).
[19] C. P. Boyer, E. G. Kalnins, and P. Winternitz, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **16**, 93 (1985).
[20] C. P. Boyer, E. G. Kalnins, and P. Winternitz, J. Math. Phys. **24**, 2022 (1983).
[21] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1969).
[22] M. A. del Olmo, M. A. Rodriquez, P. Winternitz, and H. Zassenhaus, Linear Algebr. Appl. **135**, 79 (1990).
[23] M. P. Olevski, Mat. Sb. **27**, 379 (1950).
[24] A. de Sol Mesa, C. Quesne, and Yu. F. Smirnov, J. Phys. A **31**, 321 (1998).
[25] J. D. Louck, M. Moshinsky, and K. B. Wolf, J. Math. Phys. **14**, 696 (1973).
[26] A. A. Izmest'ev, G. S. Pogosyan, A. N. Sissakian, and P. Winternitz, J. Phys. A **29**, 5949 (1996).
[27] J. A. Calzada, J. Negro, M. A. del Olmo, and M. A. Rodríguez, ''Contraction of classical integrable systems'' (preprint, 1998).
[28] J. A. Calzada, J. Negro, and M. A. del Olmo, *Quantum superintegrable Hamiltonian systems*. V. *Wigner Symposium*, Vienna, 1997, edited by P. Kasperkovitz and D. Grau (World Scientific, Singapore, 1998), p. 233.
[29] P. Havas, J. Math. Phys. **16**, 2476 (1975).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

Symmetries of Hamiltonian systems with two degrees of freedom

P. A. Damianou[a] and C. Sophocleous[b]
Department of Mathematics and Statistics, University of Cyprus, P. O. Box 537, Nicosia, Cyprus

(Received 11 June 1998; accepted for publication 25 September 1998)

We classify the Lie point symmetry groups for an autonomous Hamiltonian system with two degrees of freedom. With the exception of the harmonic oscillator or a free particle where the dimension is 15, we obtain all dimensions between 1 and 7. For each system in the classification we examine integrability. © *1999 American Institute of Physics.* [S0022-2488(99)01101-9]

I. INTRODUCTION

The objective of this paper is a complete classification of the Lie point symmetry groups for a Hamiltonian system with two degrees of freedom. We should clarify that we are dealing only with point transformations. In other words, the generators are functions of the dependent and independent variables; there is no dependence on the derivatives. We study the equations in Newtonian form since a first order system always has an infinite number of symmetries. We consider the motion of a particle of unit mass in the plane (q_1,q_2) under the influence of a potential of the form $V(q_1,q_2)$. We will assume that the Hamiltonian is time independent. This is not really a restriction because a time-dependent n-dimensional system is equivalent to a time-independent $(n+1)$-dimensional system by regarding the time variable as the new coordinate. For the most part we assume that the system is two-dimensional with Hamiltonian

$$H(q_1,q_2,p_1,p_2)=\tfrac{1}{2}p_1^2+\tfrac{1}{2}p_2^2+V(q_1,q_2). \tag{1}$$

However, in Sec. VI we generalize some of the results to the n-dimensional case. The real valued function $V(q_1,q_2)$ is assumed to be smooth on some open, connected subset of $\mathbf{R}^2$.

Hamilton's equations, in Newtonian form, become

$$\ddot{q}_1=-\frac{\partial V}{\partial q_1},\quad \ddot{q}_2=-\frac{\partial V}{\partial q_2}. \tag{2}$$

We search for point symmetries of the system (2). That is, we search for the infinitesimal transformations of the form

$$\begin{aligned} t'&=t+\epsilon\ T(t,q_1,q_2)+O(\epsilon^2),\\ q_1'&=q_1+\epsilon\ Q_1(t,q_1,q_2)+O(\epsilon^2),\\ q_2'&=q_2+\epsilon\ Q_2(t,q_1,q_2)+O(\epsilon^2). \end{aligned} \tag{3}$$

The functions $V(q_1,q_2)$ such that the system (2) admit such transformations are completely classified. Therefore, in the following analysis we determine the functions V, T, Q_1 and Q_2.

Equations (2) admit Lie transformations of the form (3) if and only if

[a]Electronic mail: damianou@ucy.ac.cy
[b]Electronic mail: christod@ucy.ac.cy

$$\Gamma^{(2)}\{\ddot{q}_1+V_{q_1}\}=0, \quad \Gamma^{(2)}\{\ddot{q}_2+V_{q_2}\}=0, \tag{4}$$

where $\Gamma^{(2)}$ is the second prolongation of

$$\Gamma=T\frac{\partial}{\partial t}+Q_1\frac{\partial}{\partial q_1}+Q_2\frac{\partial}{\partial q_2}. \tag{5}$$

For the reader who is unfamiliar with the definition and properties of Lie point symmetries, there are a number of excellent books on the subject, e.g., Refs. 1–4.

Equations (4) give two identities of the form

$$E_1(t,q_1,q_2,\dot{q}_1,\dot{q}_2)=0, \quad E_2(t,q_1,q_2,\dot{q}_1,\dot{q}_2)=0, \tag{6}$$

where we have used that $\ddot{q}_1=-\partial V/\partial q_1$ and $\ddot{q}_2=-\partial V/\partial q_2$. The functions E_1 and E_2 are explicit polynomials in $\dot{q}_1$ and $\dot{q}_2$. We impose the condition that Eqs. (6) are identities in five variables t, q_1, q_2, $\dot{q}_1$ and $\dot{q}_2$ which are regarded as independent. These two identities enable the infinitesimal transformations to be derived and ultimately impose restrictions on the functional forms of V, T, Q_1 and Q_2.

After some straightforward calculations one can show, see, e.g., Ref. 5, that the generators necessarily have the following form:

$$\begin{aligned} T&=a(t)+b_1(t)q_1+b_2(t)q_2, \\ Q_1&=b_1'(t)q_1^2+b_2'(t)q_1q_2+c_{11}(t)q_1+c_{12}(t)q_2+d_1(t), \\ Q_2&=b_1'(t)q_1q_2+b_2'(t)q_2^2+c_{21}(t)q_1+c_{22}(t)q_2+d_2(t). \end{aligned} \tag{7}$$

In this paper we classify the symmetry groups of the system according to the form of the generators. Here is a preview of the various cases and the potentials that appear.

Case 1. $b_1\neq 0$, $b_2\neq 0$.

In this case the potential is of the form

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\lambda_1 q_1+\lambda_2 q_2.$$

The symmetry group has maximum dimension. It is a 15-parameter group of transformations isomorphic to $sl(4,\mathbf{R})$.

Case 2. $b_1=b_2=0$.

In other words, T is function of time only. We consider two possibilities according to $a''\neq 0$ or $a''=0$.

First subcase: $a''\neq 0$

(2a)

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{\mu}{(q_1+\kappa q_2)^2}.$$

In this case we obtain a 6-parameter group.

(2b)

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{\mu}{q_2^2}.$$

This is a special case of the previous one with $\kappa=0$.
(2c)

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{1}{q_2^2}\Phi(\xi),$$

where $\xi=q_1/q_2$. For Φ arbitrary we end up with a 3-parameter symmetry group. For special types of Φ we obtain a 4-parameter group.
Second subcase: $a''=0$
This is the case where T is a linear function of time.
(2d)

$$V=q_2^2+\lambda_1 q_1 q_2+\Phi(\xi),$$

where $\xi=q_1-\lambda q_2$. For Φ arbitrary we end up with a 4-parameter symmetry group. For Φ quadratic we obtain a 7-parameter group and setting $\lambda_1=0$, for some special forms of Φ (exponential, logarithmic, nth power) results in a 5-parameter group of symmetries.
(2e)

$$V=\lambda_1 q_2^2+\Phi(q_1).$$

The case $\lambda_1=0$ is the case of a separable potential with one variable missing. We will comment on this case separately. If $\lambda_1\neq 0$, we end up with a 4-parameter group.
(2f) The dimensions of the symmetry groups in this case are all equal to 2 except for the last three systems where the dimension is 3. Specifically, we obtain the following list of potentials.

1.

$$V=q_2^N\Phi\left(\frac{q_1}{q_2}\right).$$

2.

$$V=\lambda_1\log q_2+\Phi\left(\frac{q_1}{q_2}\right).$$

3.

$$V=e^{\mu q_1}\Phi(q_2).$$

4

$$V=e^{\mu q_1}\Phi(q_1-\lambda q_2).$$

5–10.

$$V=\lambda_1 f_1(q_1)+\lambda_2 f_2(q_2),$$

where $f_1(q_1)=q_1^n$, $\log q_1$, $e^{\mu q_1}$ and $f_2(q_2)=q_2^m$, $\log q_2$, $e^{\mu q_2}$.

11.

$$V=\Phi(q_1^2+q_2^2).$$

12.

$$V=\lambda(q_1^2+q_2^2)^n,\quad n\neq -1,0,1\,.$$

13.

$$V=\lambda \log(q_1^2+q_2^2)$$

14.

$$V=\lambda \sin^{-1}\frac{q_1^2-q_2^2}{q_1^2+q_2^2}.$$

Finally, we note that the potential $V(q_1,q_2)=q_1^k$ has a 15-parameter group of symmetries for $k=0,1$, a 7-parameter group for $k=2$, a 6-parameter group for $k=-2$ and a 5-parameter group of symmetries otherwise. These dimensions generalize for arbitrary $n\geqslant 2$ to $(n+2)^2-1$, n^2+3, n^2+2, n^2+1 respectively.

In this paper we will not consider the general case of motion in $\mathbf{R}^n$, however, for the case $n=3$, the classification of symmetry groups is in progress; see Ref. 6. In this case, we end up again with a maximal dimension of $(n+2)^2-1=24$ for the harmonic oscillator or a free particle, but the dimensions of the other groups in the classification vary from 1 to 12 (n^2+3). We did not obtain the dimension 8 and any dimension between 13 and 23. For $n=4$ the dimensions 13 and 14 are probably missing.

In Sec. II we will also consider the simplest system, with one degree of freedom, mainly to illustrate the procedure we use for the two-dimensional case. The classification for a general ordinary differential equation of second order with one dependent and one independent variable goes back to Sophus Lie.[7] He showed that the dimensions of a maximal admitted algebra take only the values 1, 2, 3 and 8. Lie actually gave a group classification of all arbitrary order O.D.E.s. In this way he identified all equations that can be reduced to lower-order equations or completely integrated by group theoretical methods.[8] The problem of classifying symmetry groups for a system of differential equations is open. This is mentioned in Ref. 9 where some known facts are presented. Some results for linear systems of second order ordinary differential equations can be found in Ref. 10.

Of course, the ultimate goal in classical mechanics is to integrate explicitly the equations of motion. Such systems are called integrable. For the definition and basic concepts of Hamiltonian Systems and Symplectic Geometry there are a number of good references, e.g., Refs. 11–15. The key result is the following theorem of Liouville which in the 2-dimensional case translates as follows: *Consider a Hamiltonian system with two degrees of freedom. If in addition to the Hamiltonian H there is a second integral of motion I, independent of H, then the system is integrable, i.e., in principle one can solve the equations by quadratures.* Even though most of the well known systems of classical mechanics are Liouville integrable, the fact is that most Hamiltonian systems are not integrable, a result demonstrated by Poincaré. It is not surprising that most of the symmetry groups that appear in our classification correspond to integrable potentials. The nonintegrable potentials appear mainly in case 2f, where the size of the symmetry group is small. The integrability of two-dimensional systems has been the subject of numerous investigations; see, for example, Refs. 16–22. Of course, a system with symmetries should be expected to be integrable, after all this is the essence of Noether's theorem; in this direction see, for example, the review.[23] On the other hand, one can give a number of examples of integrable systems whose symmetry group is trivial (i.e., one-dimensional). Some of the chaotic systems that do not appear in our list, for example, the famous Hénon–Heiles system, are known to have only $\partial/\partial t$ as a single symmetry. Of course, we should point out that all the systems which do not appear in our classification will have $\partial/\partial t$ as a single symmetry. One can construct a number of systems possessing only one symmetry. For example, one can take

$$V=q_1^2+2q_2^2+q_1^kq_2, \tag{8}$$

with $k>1$.

As was discussed in Ref. 24 there are integrable systems which possess only one symmetry. This situation is also investigated in Ref. 25. We would like to point out another example:

$$V=4q_1^2+q_2^2+\frac{r_1}{q_1}+\frac{r_2}{q_2^2}. \tag{9}$$

In this system the associated Hamilton–Jacobi equation is separable in Cartesian coordinates. We should point out that integrable systems have symmetries other than point symmetries. One may allow the infinitesimals T, Q_i to depend on t, q_i and the derivatives of q_i. Transformations of this type are commonly called Lie–Bäckland or generalized transformations. For more details see, for example, Refs. 26–28 or the books Refs. 1–4. There is also the notion of dynamical or contact symmetries (a subset of Lie–Bäckland transformations) where the generators are velocity dependent. The connection between first integrals and symmetries is more transparent when these types of symmetries are used. Furthermore, these generalizations cover a wider range of systems. The drawback, of course, is the complexity of calculating and classifying these types of symmetry groups. This is an obvious direction of future research. Another direction is to consider velocity-dependent potentials.

For the integrable systems that appear in the list we actually give the second integral (whenever it is not obvious) or a reference. Since integrability is preserved under various transformations, e.g., translations, rotations, scalings, time reflections, we construct the second invariant for a representative of that class. Generally we choose the potential to be as simple as possible in order to illustrate the symmetry group and demonstrate integrability.

II. SYSTEMS WITH ONE DEGREE OF FREEDOM

Before attacking the two-dimensional case, we classify the symmetries for a one-dimensional system, just to illustrate the techniques we use on the two-dimensional case. We consider a Hamiltonian of the form

$$H=1/2p^2+V(q). \tag{10}$$

The equation of motion of the particle is

$$\ddot{q}=-\frac{\partial V}{\partial q}. \tag{11}$$

We search for symmetries for Eq. (11) of the form

$$t'=t+\epsilon T(t,q)+O(\epsilon^2), \quad q'=q+\epsilon Q(t,q)+O(\epsilon^2). \tag{12}$$

Equation (11) admits symmetries of the form (12) if and only if

$$\Gamma^{(2)}\{\ddot{q}+V_q\}=0, \tag{13}$$

where $\Gamma^{(2)}$ is the second prolongation of

$$\Gamma=T\frac{\partial}{\partial t}+Q\frac{\partial}{\partial q}. \tag{14}$$

The definition of the second prolongation is the following: First we define the first prolongation,

$$\Gamma^{(1)}=\Gamma+[-T_q\dot{q}^2+(Q_q-T_t)\dot{q}+Q_t]\frac{\partial}{\partial \dot{q}}. \tag{15}$$

The second prolongation of Γ is an extension of $\Gamma^{(1)}$ given by

$$\Gamma^{(2)}=\Gamma^{(1)}+[(Q_q-2T_t-3T_q\dot{q})\ddot{q}-T_{qq}\dot{q}^3+(Q_{qq}-2T_{tq})\dot{q}^2+(2Q_{tq}-T_{tt})\dot{q}+Q_{tt}]\frac{\partial}{\partial\ddot{q}}. \tag{16}$$

Equation (13) becomes an identity of the form

$$E(t,q,\dot{q})=0, \tag{17}$$

using the fact that $\ddot{q}=-V_q$.

The coefficient of $\dot{q}^3$ in (17) gives $T_{qq}=0$. Similarly, the coefficient of $\dot{q}^2$ gives $Q_{qq}=2T_{tq}$. Therefore,

$$T=a(t)+b(t)q, \quad Q=b'(t)q^2+c(t)q+d(t). \tag{18}$$

Using Eqs. (18), identity (17) becomes

$$\left[3b\frac{\partial V}{\partial q}+3qb''-a''+2c'\right]\dot{q}+(q^2b'+qc+d)\frac{\partial^2 V}{\partial q^2}+(2a'-c)\frac{\partial V}{\partial q}+q^2b'''+qc''+d''=0. \tag{19}$$

The coefficient of $\dot{q}$ in (19) gives

$$3b\frac{\partial V}{\partial q}+3qb''-a''+2c'=0. \tag{20}$$

We split the analysis into two exclusive cases: Case 1. $b\neq 0$. Case 2. $b=0$.

Case 1. $b\neq 0$.

From Eq. (20) we obtain

$$V=\frac{\lambda_1}{2}q^2+\lambda_2 q+\lambda_3. \tag{21}$$

One easily calculates that the algebra of symmetries has dimension 8. It is a simple Lie algebra isomorphic to $sl(3,\mathbf{R})$.

Case 2. $b=0$.

From (20) we have

$$c=\tfrac{1}{2}a'+c_1,$$

and Eq. (19) becomes

$$[q(a'+2c_1)+2d]\frac{\partial^2 V}{\partial q^2}+(3a'-2c_1)\frac{\partial V}{\partial q}+a'''q+2d''=0. \tag{22}$$

From Eq. (22) we deduce that V satisfies an o.d.e. of the form

$$(\lambda_1 q+\lambda_2)V_{qq}+\lambda_3 V_q=\lambda_4 q+\lambda_5. \tag{23}$$

In order to solve Eq. (23) we consider the following five possibilities.

1. $\lambda_1=\lambda_2=\lambda_3=\lambda_4=\lambda_5=0$. In this case we get $a(t)=\text{constant}$, $c_1=0$ and $d(t)=0$. Therefore for V arbitrary we have $T=c_2$, $Q=0$. In other words the symmetry group is trivial (one-dimensional).
2. $\lambda_1=\lambda_3=0$, $\lambda_2\neq 0$. In this case, V is quadratic, a case already examined.

3. $\lambda_1=0$, $\lambda_2\neq 0$, $\lambda_3\neq 0$. From (23) we get $V=\lambda e^{\mu q}$, where we have ignored linear terms. We obtain a 2-parameter group of symmetries with $T=-2c_1t+c_2$ and $Q=4c_1/\mu$.
4. $\lambda_1\neq 0$, $\lambda_3=0$. From (23) we obtain $V=q\log q$. The symmetry group here is also trivial.
5. $\lambda_1\neq 0$, $\lambda_3\neq 0$. Without loss of generality we take $\lambda_2=0$ in (23) and we obtain either (i) $V=\lambda q^n$, $n\neq 0,1,2$, or (ii) $V=\lambda \log q$.

(i) We substitute $V=\lambda q^n$ into (22) to obtain $a(t)=c_2t^2+2c_3t+c_4$, $d(t)=c_1=0$ if $n=-2$ and $a(t)=[2(2-n)/n+2]c_1t+c_2$, $d(t)=0$ if $n\neq -2$. To summarize, we have for $V=\lambda/q^2$,

$$T=c_2t^2+2c_3t+c_4,\quad Q=(c_2t+c_3)q, \tag{24}$$

and for $V=\lambda q^n$, $n\neq -2,0,1,2$ the generators have the following form:

$$T=\frac{2(2-n)}{n+2}c_1t+c_2,\quad Q=\frac{4c_1}{n+2}q. \tag{25}$$

In other words, we obtain either a two-parameter or a three-parameter group of symmetries.
(ii) When $V=\lambda\log q$ we get $a(t)=2c_1t+c_2$ and $d(t)=0$. Therefore,

$$T=2c_1t+c_2,\quad Q=2c_1q. \tag{26}$$

This is a two-parameter group of symmetries.

To summarize the results: In the case of one degree of freedom we obtain a maximal dimension of 8 for the harmonic oscillator or a free particle, but the dimensions in the other groups in the classification vary from 1 to 3. We do not obtain any dimension between 4 and 7.

III. SYSTEMS WITH TWO DEGREES OF FREEDOM

We return now to the case of two-degrees of freedom. The analysis is analogous to the one used in the case of one-degree of freedom. We substitute the forms of T, Q_1, Q_2 in (7) into Eqs. (6).

The coefficient of $\dot{q}_1$ in Eq. (6) $[E_2=0]$ gives

$$2\left(\frac{\partial V}{\partial q_2}b_1+\frac{\partial^2 b_1}{\partial t^2}q_2+\frac{\partial c_{21}}{\partial t}\right)=0. \tag{27}$$

On the other hand, the coefficient of $\dot{q}_2$ in Eq. (6) $[E_1=0]$ implies

$$2\left(\frac{\partial V}{\partial q_1}b_2+\frac{\partial^2 b_2}{\partial t^2}q_1+\frac{\partial c_{12}}{\partial t}\right)=0. \tag{28}$$

Similarly, the coefficient of $\dot{q}_1$ in Eq. (6) $[E_1=0]$ gives

$$3b_1\frac{\partial V}{\partial q_1}+b_2\frac{\partial V}{\partial q_2}+3q_1\frac{\partial^2 b_1}{\partial t^2}+q_2\frac{\partial^2 b_2}{\partial t^2}-\frac{\partial^2 a}{\partial t^2}+2\frac{\partial c_{11}}{\partial t}=0, \tag{29}$$

while the coefficient of $\dot{q}_2$ in Eq. (6) $[E_2=0]$ implies

$$3b_2\frac{\partial V}{\partial q_2}+b_1\frac{\partial V}{\partial q_1}+3q_2\frac{\partial^2 b_2}{\partial t^2}+q_1\frac{\partial^2 b_1}{\partial t^2}-\frac{\partial^2 a}{\partial t^2}+2\frac{\partial c_{22}}{\partial t}=0. \tag{30}$$

If $b_1(t)\neq 0$ and $b_2(t)\neq 0$, then from Eqs. (27) and (28) we deduce that V is quadratic in q_1 and q_2. We also note that if $b_1=0$, $b_2\neq 0$ (or $b_1\neq 0$, $b_2=0$), then from Eqs. (28) and (29) V has again a quadratic form. We therefore split the analysis into two exclusive cases: **Case 1.** $b_1\neq 0$, $b_2\neq 0$. **Case 2.** $b_1=b_2=0$.

IV. CASE 1

From Eqs. (27) and (28) we deduce that

$$V=\lambda_1 q_1^2+\lambda_2 q_2^2+\lambda_3 q_1+\lambda_4 q_2+\lambda_5. \tag{31}$$

Now, substituting (31) into (6) the coefficients of $q_1\dot{q}_1$ in E_1 and $q_2\dot{q}_1$ in E_2 give, respectively,

$$3(b_1''+2\lambda_1 b_1)=0, \quad 3(b_1''+2\lambda_2 b_1)=0. \tag{32}$$

Hence, it follows that $\lambda_1=\lambda_2=0$ or $\lambda_1=\lambda_2\neq 0$.

In the case $\lambda_1=\lambda_2=0$, V is linear. We shall present the symmetries for this case, but in the remaining part of the analysis we shall ignore linear terms in the form of V. Adding a constant to Eqs. (2) has no effect on the symmetry groups.

Subcase 1a: $V=\lambda_1 q_1+\lambda_2 q_2+\lambda_3$.

Note that we have taken $\lambda_1=\lambda_2=0$ in (31) and then renamed the constants. Without presenting any calculations, we state that the system (2) with V linear has the following 15 symmetries:

$$\Gamma_1=\frac{\partial}{\partial t}, \quad \Gamma_2=\frac{\partial}{\partial q_1}, \quad \Gamma_3=\frac{\partial}{\partial q_2},$$

$$\Gamma_4=q_1\frac{\partial}{\partial t}-\left(\frac{\lambda_1^2}{4}t^3+\frac{3}{2}\lambda_1 t q_1\right)\frac{\partial}{\partial q_1}-\left(\frac{\lambda_1\lambda_2}{4}t^3+\frac{1}{2}\lambda_1 t q_2+\lambda_2 t q_1\right)\frac{\partial}{\partial q_2},$$

$$\Gamma_5=t\frac{\partial}{\partial t}+\left(\frac{1}{2}q_1-\frac{3}{4}\lambda_1 t^2\right)\frac{\partial}{\partial q_1}+\left(\frac{1}{2}q_2-\frac{3}{4}\lambda_2 t^2\right)\frac{\partial}{\partial q_2},$$

$$\Gamma_6=t^2\frac{\partial}{\partial t}+\left(q_1 t-\frac{\lambda_1}{2}t^3\frac{\partial}{\partial q_1}+\left(q_2 t-\frac{\lambda_2}{2}t^3\right)\frac{\partial}{\partial q_2},\right.$$

$$\Gamma_7=q_2\frac{\partial}{\partial t}-\left(\frac{\lambda_1\lambda_2}{4}t^3+\lambda_1 t q_2+\frac{1}{2}\lambda_2 t q_1\right)\frac{\partial}{\partial q_1}-\left(\frac{\lambda_2^2}{4}t^3+\frac{3}{2}\lambda_2 t q_2\right)\frac{\partial}{\partial q_2},$$

$$\Gamma_8=(2tq_2+\lambda_2 t^3)\frac{\partial}{\partial t}+\left(2q_1 q_2+\lambda_2 t^2 q_1-\lambda_1 t^2 q_2-\frac{1}{2}\lambda_1\lambda_2 t^4\right)\frac{\partial}{\partial q_1}+\left(2q_2^2-\frac{1}{2}\lambda_2^2 t^4\right)\frac{\partial}{\partial q_2},$$

$$\Gamma_9=(2tq_1+\lambda_1 t^3)\frac{\partial}{\partial t}+\left(2q_1^2-\frac{1}{2}\lambda_1^2 t^4\right)\frac{\partial}{\partial q_1}+\left(2q_1 q_2+\lambda_1 t^2 q_2-\lambda_2 t^2 q_1-\frac{1}{2}\lambda_1\lambda_2 t^4\right)\frac{\partial}{\partial q_2},$$

$$\Gamma_{10}=\left(q_2+\frac{1}{2}\lambda_2 t^2\right)\frac{\partial}{\partial q_1}, \quad \Gamma_{11}=\left(q_1+\frac{1}{2}\lambda_1 t^2\right)\frac{\partial}{\partial q_2}, \quad \Gamma_{12}=t\frac{\partial}{\partial q_1},$$

$$\Gamma_{13}=t\frac{\partial}{\partial q_2}, \quad \Gamma_{14}=\left(q_1+\frac{1}{2}\lambda_1 t^2\right)\frac{\partial}{\partial q_1}, \quad \Gamma_{15}=\left(q_2+\frac{1}{2}\lambda_2 t^2\right)\frac{\partial}{\partial q_2}.$$

Remark: This system is, of course, integrable. The second integral is $I=1/2\,p_1^2+\lambda_1 q_1$ *or* $I=1/2\,p_2^2+\lambda_2 q_2$. *It also has constants of motion linear in the momenta, for example,* $I=-\lambda_2 p_1+\lambda_1 p_2$.

Subcase 1b: $V=(\lambda/2)(q_1^2+q_2^2)$.

We will choose $\lambda=1$. We substitute the form of V into Eqs. (6) and equate coefficients.

In E_1, $\dot{q}_1 q_1=0$ implies

$$\frac{d^2 b_1}{dt^2}+b_1=0.$$

In E_1, $\dot{q}_1 q_2=0$ implies

$$\frac{d^2 b_2}{dt^2}+b_2=0.$$

In E_1, $\dot{q}_2=0$ implies

$$\frac{dc_{12}}{dt}=0,$$

therefore c_{12} is constant. Similarly, by examining the coefficient of $\dot{q}_1$ in E_2 we see that c_{21} is also constant.

In E_1, $\dot{q}_1=0$ and in E_2, $\dot{q}_2=0$, imply that

$$\frac{d^2 a}{dt^2}-2\frac{dc_{11}}{dt}=0,$$

and

$$\frac{d^2 a}{dt^2}-2\frac{dc_{22}}{dt}=0.$$

Similarly, using the coefficient of $q_1=0$ in E_1 and $q_2=0$ in E_2, we obtain

$$\frac{d^2 c_{11}}{dt^2}+2\frac{da}{dt}=0,$$

and

$$\frac{d^2 c_{22}}{dt^2}+2\frac{da}{dt}=0.$$

Finally, $E_1=0$ and $E_2=0$ imply that the functions $d_1(t)$ and $d_2(t)$ are solutions of the equation

$$\frac{d^2 x}{dt^2}+x=0.$$

Therefore the form of the generators in this case is the following:

$$T=k_1+k_2\cos 2t+k_3\sin 2t+(k_4\cos t+k_5\sin t)q_1+(k_6\cos t+k_7\sin t)q_2,$$

$$Q_1=(-k_4\sin t+k_5\cos t)q_1^2+(-k_6\sin t+k_7\cos t)q_1q_2 +(-k_2\sin 2t+k_3\cos 2t+c_{11})q_1+c_{12}q_2+k_8\cos t+k_9\sin t,$$

$$Q_2=(-k_4\sin t+k_5\cos t)q_1q_2+(-k_6\sin t+k_7\cos t)q_2^2+c_{21}q_1 +(-k_2\sin 2t+k_3\cos 2t+c_{22})q_2+k_{10}\cos t+k_{11}\sin t.$$

We note that the system (2) with $V=(\lambda/2)(q_1^2+q_2^2)$ admits a 15-parameter group of transformations isomorphic to $sl(4,\mathbf{R})$.

Remark: This system is the 2-dimensional isotropic oscillator. A second integral is $I_1=1/2\,p_1^2+1/2\,q_1^2$ *or* $I_2=1/2\,p_2^2+\frac{1}{2}q_2^2$. *We also have constants of motion linear in the momenta, for example,* $I_3=q_2p_1-q_1p_2$.

Remark: Cases 1a and 1b give the most general form of Hamiltonian for which the second invariant is linear in the momenta.[21]

V. CASE 2

We use the identities $E_1=0$ and $E_2=0$ in (6) to obtain

$$E_1:\ \dot{q}_2=0\ \Leftrightarrow\ c_{12}'=0,\quad E_2:\ \dot{q}_1=0\Leftrightarrow c_{21}'=0,$$
$$E_1:\dot{q}_1=0\ \Leftrightarrow\ 2c_{11}'-a''=0,\quad E_2:\dot{q}_2=0\ \Leftrightarrow\ 2c_{22}'-a''=0. \tag{33}$$

Therefore, $c_{12}=c_1$, $c_{21}=c_2$, $c_{11}=1/2\,a'+c_3$ and $c_{22}=1/2\,a'+c_4$. Here and elsewhere the c_i are constants. Using these results, Eqs. (6) take the form

$$(a'q_2+2c_2q_1+2c_4q_2+2d_2)V_{q_1q_2}+(a'q_1+2c_3q_1+2c_1q_2+2d_1)V_{q_1q_1} +(3a'-2c_3)V_{q_1}-2c_1V_{q_2}+a'''q_1+2d_1''=0, \tag{34}$$

and

$$(a'q_1+2c_3q_1+2c_1q_2+2d_1)V_{q_1q_2}+(a'q_2+2c_2q_1+2c_4q_2+2d_2)V_{q_2q_2} +(3a'-2c_4)V_{q_2}-2c_2V_{q_1}+a'''q_2+2d_2''=0. \tag{35}$$

We differentiate Eqs. (34) and (35) with respect to t to obtain, respectively,

$$(a''q_2+2d_2')V_{q_1q_2}+(a''q_1+2d_1')V_{q_1q_1}+3a''V_{q_1}+a'''{}'q_1+2d_1'''=0, \tag{36}$$

$$(a''q_1+2d_1')V_{q_1q_2}+(a''q_2+2d_2')V_{q_2q_2}+3a''V_{q_2}+a'''{}'q_2+2d_2'''=0. \tag{37}$$

We now split the analysis into two parts: $a''\neq 0$ or $a''=0$.

Nonlinear *T*: $a''\neq 0$

We divide Eqs. (36) and (37) by a'' and then differentiate with respect to t to obtain, respectively,

$$2\left(\frac{d_2'}{a''}\right)'V_{q_1q_2}+2\left(\frac{d_1'}{a''}\right)'V_{q_1q_1}+\left(\frac{a'''{}'}{a''}\right)'q_1+2\left(\frac{d_1'''}{a''}\right)'=0, \tag{38}$$

$$2\left(\frac{d_1'}{a''}\right)' V_{q_1q_2}+2\left(\frac{d_2'}{a''}\right)' V_{q_2q_2}+\left(\frac{a''''}{a''}\right)' q_2+2\left(\frac{d_2'''}{a''}\right)'=0. \tag{39}$$

From Eqs. (38) and (39) we deduce that the function $V(q_1,q_2)$ satisfies two partial differential equations of the form

$$\lambda_1 V_{q_1q_2}+\lambda_2 V_{q_1q_1}+\lambda_3 q_1+\lambda_4=0, \tag{40}$$

$$\lambda_2 V_{q_1q_2}+\lambda_1 V_{q_2q_2}+\lambda_3 q_2+\lambda_5=0. \tag{41}$$

In order to solve Eqs. (40) and (41) we consider the following cases:

(i) $\lambda_1\neq 0,\ \lambda_2\neq 0$

(ii) $\lambda_1=0,\ \lambda_2\neq 0\ \ (\text{or}\ \lambda_1\neq 0,\ \lambda_2=0)$

(iii) $\lambda_1=\lambda_2=\lambda_3=\lambda_4=0$.

In the following three subcases we determine the form of V from Eqs. (40) and (41). The corresponding generators may be obtained with the employment of Eqs. (34)–(39).

Subcase 2a:

In this case, V takes the form

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{\mu}{(q_1+\kappa q_2)^2}, \tag{42}$$

with the corresponding generators

$$\begin{gathered} T=a(t),\\ Q_1=\tfrac{1}{2}a'(t)q_1+\kappa^2 c_4 q_1-\kappa c_4 q_2+\kappa d_2(t),\\ Q_2=\tfrac{1}{2}a'(t)q_2-\kappa c_4 q_1+c_4 q_2-d_2(t), \end{gathered} \tag{43}$$

where $a''+4\lambda a=c_8$ and d_2 is a solution of $d_2''+\lambda d_2=0$ (6-parameter group).

Subcase 2b:

Setting $\lambda_1=0$ in (40) and (41) we deduce that V has the form

$$V=k_1q_1^3+k_2q_1^2+k_3q_1q_2^2+k_4q_1q_2+k_5q_1+\Phi(q_2). \tag{44}$$

Using Eqs. (36) and (37) we get $k_1=k_3=k_4=0$ and $\Phi=k_2q_2^2+k_6q_2+k_7+\mu/q_2^2$. Therefore, ignoring the linear terms, V takes the form

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{\mu}{q_2^2}. \tag{45}$$

Finally, using Eqs. (34) and (35) we obtain the forms of the group generators:

$$\begin{gathered} T=a(t),\quad Q_1=c_3q_1+\tfrac{1}{2}a'(t)q_1+d_1(t),\\ Q_2=\tfrac{1}{2}a'(t)q_2, \end{gathered} \tag{46}$$

where $a''+4\lambda a=c_8$ and $d_1''+\lambda d_1=0$ (6-parameter group).

The case $\lambda=0$ is equivalent to the system $V(q_1,q_2)=1/q_1^2$ with generators

$$\begin{aligned} T&=c_1t^2+2c_2t+c_3, \quad Q_1=(c_1t+c_2)q_1, \\ Q_2&=(c_1t+c_2+c_4)q_2+c_5t+c_6. \end{aligned} \tag{47}$$

We choose the following basis for the Lie algebra of symmetries:

$$X_1=\frac{\partial}{\partial t},$$

$$X_2=2t\frac{\partial}{\partial t}+q_1\frac{\partial}{\partial q_1}+q_2\frac{\partial}{\partial q_2},$$

$$X_3=t^2\frac{\partial}{\partial t}+q_1t\frac{\partial}{\partial q_1}+q_2t\frac{\partial}{\partial q_2},$$

$$X_4=q_2\frac{\partial}{\partial q_2}, \quad X_5=t\frac{\partial}{\partial q_2}, \quad X_6=\frac{\partial}{\partial q_2},$$

with (nonzero) bracket relations

$$\begin{aligned} &[X_1,X_2]=2X_1, \quad [X_1,X_3]=X_2, \\ &[X_1,X_5]=X_6, \quad [X_2,X_3]=2X_3, \\ &[X_2,X_5]=X_5, \quad [X_2,X_6]=-X_6, \\ &[X_3,X_6]=-X_5, \quad [X_4,X_5]=-X_5, \quad [X_4,X_6]=-X_6. \end{aligned} \tag{48}$$

This algebra is not semi-simple since the ideal generated by X_5, X_6 is Abelian. It is not solvable either because $L^{(1)}=\{X_1,X_2,X_3,X_5,X_6\}$, and $L^{(2)}=L^{(1)}$.

Remark: It is clear that Subcase 2b is a special case of Subcase 2a by setting $\kappa=0$.

Subcase 2c:

Since all the coefficients of the terms in Eqs. (38) and (39) vanish, the functions $a(t)$, $d_1(t)$ and $d_2(t)$ may be determined. From Eqs. (36) and (37) we deduce that

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{1}{q_2^2}\Phi(\xi), \tag{49}$$

where $\xi=q_1/q_2$. We now use Eqs. (34) and (35) to determine the forms of the generators.

Without presenting any more calculations we state the following results.

Φ *arbitrary:*

$$T=a(t), \quad Q_1=\tfrac{1}{2}a'(t)q_1, \quad Q_2=\tfrac{1}{2}a'(t)q_2, \tag{50}$$

where $a''+4\lambda a=c_8$ (3-parameter group).

For $\lambda=0$ the Lie algebra has a basis given by

$$X_1=\frac{\partial}{\partial t}, \quad X_2=2t\frac{\partial}{\partial t}+q_1\frac{\partial}{\partial q_1}+q_2\frac{\partial}{\partial q_2},$$

$$X_3=t^2\frac{\partial}{\partial t}+tq_1\frac{\partial}{\partial q_1}+tq_2\frac{\partial}{\partial q_2},$$

with bracket relations of a simple Lie algebra:

$$[X_1,X_2]=2X_1,\quad [X_1,X_3]=X_2,\quad [X_2,X_3]=2X_3.$$

$\Phi=\mu/\xi^2=\mu(q_2^2/q_1^2)$:

It follows from (49) that

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{\mu}{q_1^2}. \tag{51}$$

This potential already appeared in 2b.

$\Phi=\mu=$constant:

It follows from (49) that

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{\mu}{q_2^2}. \tag{52}$$

This system is not different from the previous one.

$\Phi=\mu/(\xi^2+1)]\,e^{c\tan^{-1}\xi}$:

It follows from (49) that

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{\mu}{q_1^2+q_2^2}e^{c\tan^{-1}(q_1/q_2)}. \tag{53}$$

The generators take the form

$$T=a(t),\quad Q_1=\tfrac{1}{4}cc_1q_1+c_1q_2+\tfrac{1}{2}a'(t)q_1, \tag{54}$$

$$Q_2=\tfrac{1}{4}cc_1q_2-c_1q_1+\tfrac{1}{2}a'(t)q_2,$$

where $a''+4\lambda a=c_8$ (4-parameter group).

The Lie algebra for this system is a direct sum of an $sl(2,\mathbf{R})$ and a one-dimensional Lie algebra. It has a basis consisting of the vectors

$$X_1=\frac{\partial}{\partial t},$$

$$X_2=2t\frac{\partial}{\partial t}+q_1\frac{\partial}{\partial q_1}+q_2\frac{\partial}{\partial q_2},$$

$$X_3=t^2\frac{\partial}{\partial t}+tq_1\frac{\partial}{\partial q_1}+tq_2\frac{\partial}{\partial q_2},$$

$$X_4=\left(\frac{1}{4}cq_1+q_2\right)\frac{\partial}{\partial q_1}+\left(\frac{1}{4}cq_2-q_1\right)\frac{\partial}{\partial q_2},$$

with bracket relations

$$[X_1,X_2]=2X_1,\quad [X_1,X_3]=X_2,$$

$$[X_2,X_3]=2X_3,\quad [X_i,X_4]=0,\quad i=1,2,3.$$

This example generalizes an n dimension to a Lie algebra which is a direct sum $sl(2,\mathbf{R})\oplus so(n,\mathbf{R})$.

Remark: In polar coordinates this system is

$$V=\frac{\lambda}{2}(r^2)+\frac{\mu}{r^2}e^{c\theta}.$$

It is integrable, by taking $B(\theta)=\mu e^{c\theta}$ *and* $I=\frac{1}{2}l^2+B(\theta)$, *where* $l=q_1p_2-p_1q_2$.

Taking $\Phi(\xi)=r_1/\xi^2+r_2$ *we end up with the system*

$$V=\frac{\lambda}{2}(q_1^2+q_2^2)+\frac{r_1}{q_1^2}+\frac{r_2}{q_2^2}.$$

The associated Hamilton–Jacobi equation is separable in Cartesian and polar coordinates. This system is an example of a system with closed trajectories under the influence of a Noncentral field.[29] *For* $\lambda=1$, *the generators take the form*

$$T=c_1+c_2\cos 2t+c_3\sin 2t,$$

$$Q_1=(-c_2\sin 2t+c_3\cos 2t)q_1,$$

$$Q_2=(-c_2\sin 2t+c_3\cos 2t)q_2.$$

They form a 3-dimensional Lie algebra,

$$X_1=\frac{\partial}{\partial t},$$

$$X_2=\cos 2t\frac{\partial}{\partial t}-q_1\sin 2t\frac{\partial}{\partial q_1}-q_2\sin 2t\frac{\partial}{\partial q_2},$$

$$X_3=\sin 2t\frac{\partial}{\partial t}+q_1\cos 2t\frac{\partial}{\partial q_1}+q_2\cos 2t\frac{\partial}{\partial q_2},$$

with bracket relations

$$[X_1,X_2]=-2X_3,\quad [X_1,X_3]=2X_2,\quad [X_2,X_3]=2X_1. \tag{55}$$

In other words, it is a simple Lie algebra of type A_1 *isomorphic to* $so(3,\mathbf{R})$. *For* $\lambda=0$ *we obtain a Lie algebra isomorphic to* $sl(2,\mathbf{R})$. *In Ref.* 30 *the most general form of a differential equation invariant under the action of the generators of* $sl(2,\mathbf{R})$ *is determined.*

We note that the similar system,

$$V=\frac{\lambda}{2}(4q_1^2+q_2^2)+\frac{r_1}{q_1}+\frac{r_2}{q_2^2}, \tag{56}$$

has $\partial/\partial t$ *as the only symmetry.*

In general, the system $V=(\lambda/2)(q_1^2+q_2^2)+(1/q_2^2)\Phi(q_1/q_2)$ *is integrable: Changing to polar coordinates* $q_1=r\cos\theta$, $q_2=r\sin\theta$ *we find that*

$$V=\frac{\lambda}{2}r^2+\frac{\Phi(\cot\theta)}{r^2\sin^2\theta}=\frac{\lambda}{2}r^2+\frac{B(\theta)}{r^2}.$$

Letting $l=q_1p_2-p_1q_2$, *the second integral is* $I=1/2\,l^2+B(\theta)$.

The system $V=(\lambda/2)(q_1^2+q_2^2)+\mu/(q_1+\kappa q_2)^2$ *is integrable. It is a special case of (49) with* $\Phi(\xi)=\mu/(\xi+\kappa)^2$.

T linear: $a''=0$

In this case $a(t)$ is a linear function of time.

From Eqs. (36) and (37) we deduce that the function $V(q_1,q_2)$ satisfies two partial differential equations of the form

$$\lambda_1 V_{q_1 q_1}+\lambda_2 V_{q_1 q_2}+\lambda_3=0, \tag{57}$$

$$\lambda_1 V_{q_1 q_2}+\lambda_2 V_{q_2 q_2}+\lambda_4=0. \tag{58}$$

In order to solve these equations, we consider the following cases:

(i) $\lambda_1\neq 0,\ \lambda_2\neq 0,$

(ii) $\lambda_1=0,\ \lambda_2\neq 0\ \ (\text{or}\ \ \lambda_1\neq 0,\ \lambda_2=0),$

(iii) $\lambda_1=\lambda_2=\lambda_3=\lambda_4=0.$

Without giving any more details, using Eqs. (34) and (35), we are led to the following results:

Subcase 2d: $V=\lambda_1 q_2^2+\lambda_2 q_1 q_2+\Phi(\xi)$,

where $\xi=q_1-\lambda q_2$. We obtain various forms of the generators depending on the form of the function Φ.

$\Phi=\lambda_3\xi^2$:

That is, V is quadratic of the form

$$V=\lambda_1 q_2^2+\lambda_2 q_1 q_2+\lambda_3 q_1^2. \tag{59}$$

We have followed the common practice of renaming the constants. The corresponding generators are

$$\begin{gathered} T=c_6,\quad Q_1=c_1 q_2+c_3 q_1+d_1(t),\\ Q_2=c_1 q_1+\left(2\left(\frac{\lambda_1}{\lambda_2}-\frac{\lambda_3}{\lambda_2}\right)c_1+c_3\right)q_2+d_2(t), \end{gathered} \tag{60}$$

where $d_1(t)$ and $d_2(t)$ satisfy the o.d.e.'s $d_1''+2\lambda_3 d_1(t)+\lambda_2 d_2(t)=0$, and $d_2''+2\lambda_1 d_2(t)+\lambda_2 d_1(t)=0$ (7-parameter group).

If $\lambda_2=0$ then $\lambda_1\neq\lambda_3$, $Q_1=c_3 q_1+d_1(t)$, $Q_2=c_4 q_2+d_2(t)$ and $d_1(t)$, $d_2(t)$ satisfy the same o.d.e.'s with $\lambda_2=0$.

We describe explicitly the Lie algebra for the potential $V(q_1,q_2)=-\frac{1}{2}q_1^2+\frac{1}{2}q_2^2$. The Lie algebra is 7-dimensional with generators:

$$X_1=\frac{\partial}{\partial t},\quad X_2=q_1\frac{\partial}{\partial q_1},\quad X_3=e^t\frac{\partial}{\partial q_1},$$

$$X_4=e^{-t}\frac{\partial}{\partial q_1},\quad X_5=q_2\frac{\partial}{\partial q_2},\quad X_6=\cos t\,\frac{\partial}{\partial q_2},\quad X_7=\sin t\,\frac{\partial}{\partial q_2},$$

with (nonzero) bracket relations

$$[X_1,X_3]=X_3, \quad [X_1,X_4]=-X_4, \quad [X_1,X_6]=-X_7,$$

$$[X_1,X_7]=X_6, \quad [X_2,X_3]=-X_3,$$

$$[X_2,X_4]=-X_4, \quad [X_5,X_6]=-X_6, \quad [X_5,X_7]=-X_7.$$

This Lie algebra L is solvable with $L^{(1)}=[L,L]=\{X_3,X_4,X_6,X_7\}$ and $L^{(2)}=\{0\}$.

Remark: The system with the Hamiltonian,

$$H=\tfrac{1}{2}p_1^2+\tfrac{1}{2}p_2^2+\lambda_1 q_1^2+\lambda_2 q_1 q_2+\lambda_3 q_2^2, \tag{61}$$

is integrable. We can actually rotate the Hamiltonian to a separable one, obtain the second integral and then rotate back to obtain the invariant in the original coordinates. So, we set

$$q_1=\cos\theta\ \ Q_x+\sin\theta\ \ Q_y,$$

$$q_2=-\sin\theta\ \ Q_x+\cos\theta\ \ Q_y,$$

$$p_1=\cos\theta\ \ p_x+\sin\theta\ \ p_y,$$

$$p_2=-\sin\theta\ \ p_x+\cos\theta\ \ p_y.$$

The Hamiltonian H will be transformed to a new Hamiltonian which is a function of Q_x, Q_y, p_x and p_y. The coefficient of Q_xQ_y in the rotated Hamiltonian is

$$(\lambda_1-\lambda_3)\sin 2\theta+\lambda_2\cos 2\theta.$$

If $\lambda_1=\lambda_3$, we choose $\theta=\pi/4$. If $\lambda_1\neq\lambda_3$, then we choose θ to satisfy

$$\tan 2\theta=\frac{\lambda_2}{\lambda_3-\lambda_1}.$$

Therefore, in the new coordinates the Hamiltonian is separable of the form

$$\tfrac{1}{2}p_x^2+\tfrac{1}{2}p_y^2+\mu_1 Q_x^2+\mu_2 Q_y^2.$$

We may choose the second integral to be $p_x^2+\mu_1 Q_x^2$. The second integral for the original system is

$$I=(\cos\theta\, p_1-\sin\theta\, p_2)^2+\mu_1(\cos\theta\, q_1-\sin\theta\, q_2)^2.$$

Φ arbitrary:

In this case, V has the form

$$V=\lambda_2\left(q_1q_2+\frac{1-\lambda^2}{2\lambda}q_2^2\right)+\Phi(q_1-\lambda q_2). \tag{62}$$

The corresponding generators are

$$T=c_6,$$

$$Q_1=\lambda c_2 q_1+c_2 q_2+\lambda d_2(t), \tag{63}$$

$$Q_2=c_2 q_1+\frac{1}{\lambda}c_2 q_2+d_2(t),$$

where $d_2(t)$ satisfies the o.d.e. $d_2''+(\lambda_2/\lambda)d_2(t)=0$ (4-parameter group).

Remark: Assume $\lambda=1$. *The system with Hamiltonian*

$$H=\tfrac{1}{2}p_1^2+\tfrac{1}{2}p_2^2+\lambda_2 q_1 q_2+\Phi(q_1-q_2), \tag{64}$$

is integrable. We can actually transform the Hamiltonian to a separable one, obtain the second integral and then rotate back to obtain the invariant in the original coordinates. So, we set

$$q_1=\frac{1}{\sqrt{2}}\ Q_x+\frac{1}{\sqrt{2}}\ Q_y,\quad q_2=-\frac{1}{\sqrt{2}}\ Q_x+\frac{1}{\sqrt{2}}\ Q_y,$$

$$p_1=\frac{1}{\sqrt{2}}\ p_x+\frac{1}{\sqrt{2}}\ p_y,\quad p_2=-\frac{1}{\sqrt{2}}\ p_x+\frac{1}{\sqrt{2}}\ p_y.$$

The Hamiltonian H will be transformed to a new Hamiltonian which is a function of Q_x, Q_y, p_x *and* p_y. *Therefore, in the new coordinates the Hamiltonian is separable of the form*

$$\frac{1}{2}p_x^2+\frac{1}{2}p_y^2+\Phi(\sqrt{2}Q_x)+\frac{\lambda_2}{2}(Q_y^2-Q_x^2).$$

We may choose the second integral to be $\frac{1}{2}p_x^2+\Phi(\sqrt{2}Q_x)-(\lambda_2/2)Q_x^2$. *The second integral for the original system is*

$$I=\frac{1}{4}(p_1-p_2)^2+\Phi(q_1-q_2)-\frac{\lambda_2}{4}(q_1-q_2)^2.$$

$\Phi=\lambda_3\xi^n$, $n\neq 0,1,2$:

In this case, V has the form

$$V=\lambda_2\left(q_1q_2+\frac{1-\lambda^2}{2\lambda}q_2^2\right)+\lambda_3(q_1-\lambda q_2)^n, \tag{65}$$

with $n\neq 0,1,2$.

The generators are

$$T=2c_5t+c_6,$$

$$Q_1=c_2(\lambda q_1+q_2)-\frac{4}{n-2}c_5q_1+\lambda d_2(t), \tag{66}$$

$$Q_2=c_2\left(q_1+\frac{1}{\lambda}q_2\right)-\frac{4}{n-2}c_5q_2+d_2(t),$$

where $d_2(t)$ satisfies the o.d.e. $d_2''+\lambda_2/\lambda\, d_2(t)=0$.

If $\lambda_2=0$ then $d_2(t)=c_3+c_4t$ and we end-up with a 5-parameter group. Note that for $n=-2$ we are in subcase 2a with a 6-parameter group.

If $\lambda_2\neq 0$, then we set $c_5=0$ and we end up with a 4-parameter group (the same as Φ arbitrary).

For example, if $V(q_1,q_2)=q_1q_2+(q_1-q_2)^3$ then the Lie algebra is generated by

$$X_1=\frac{\partial}{\partial t},\quad X_2=(q_1+q_2)\left(\frac{\partial}{\partial q_1}+\frac{\partial}{\partial q_2}\right),$$

$$X_3=\cos t\left(\frac{\partial}{\partial q_1}+\frac{\partial}{\partial q_2}\right),\quad X_4=\sin t\left(\frac{\partial}{\partial q_1}+\frac{\partial}{\partial q_2}\right),$$

with (nonzero) bracket relations

$$[X_1,X_3]=-X_4,\quad [X_1,X_4]=X_3,$$

$$[X_2,X_3]=-2X_3,\quad [X_2,X_4]=-2X_4.$$

This Lie algebra L is solvable with $L^{(1)}=[L,L]=\{X_3,X_4\}$ and $L^{(2)}=\{0\}$.

On the other hand, the potential $V(q_1,q_2)=(q_1-q_2)^3$ gives a five-dimensional Lie algebra. This Lie algebra is isomorphic with the symmetry Lie algebra for the potential $V(q_1,q_2)=q_1^3$ which we examine later.

$\Phi=\lambda_3 e^{\mu\xi}$:

In this case, V has the form

$$V=\lambda_2\left(q_1q_2+\frac{1-\lambda^2}{2\lambda}q_2^2\right)+\lambda_3 e^{\mu(q_1-\lambda q_2)}. \tag{67}$$

The generators are

$$T=2c_5t+c_6,$$

$$Q_1=c_2(\lambda q_1+q_2)-\frac{4}{\mu}c_5+\lambda d_2(t), \tag{68}$$

$$Q_2=c_2\left(q_1+\frac{1}{\lambda}q_2\right)+d_2(t),$$

where $d_2(t)$ satisfies the o.d.e. $d_2''+(\lambda_2/\lambda)d_2=0$.

If $\lambda_2=0$ then $d_2(t)=c_3+c_4t$ and we end up with a 5-parameter group.

If $\lambda_2\neq 0$, then we set $c_5=0$ and we end up with a 4-parameter group (the same as Φ arbitrary).

The case $\lambda=\mu=1$ and $\lambda_2=0$ is the Toda Lattice, a well-known integrable system.[31,32] We will calculate the Lie algebra of symmetries for the potential of the Toda lattice $V(q_1,q_2)=e^{q_1-q_2}$. We obtain a five-dimensional Lie algebra with generators

$$X_1=\frac{\partial}{\partial t},\quad X_2=2t\frac{\partial}{\partial t}-4\frac{\partial}{\partial q_1},$$

$$X_3=(q_1+q_2)\left(\frac{\partial}{\partial q_1}+\frac{\partial}{\partial q_2}\right),\quad X_4=\left(\frac{\partial}{\partial q_1}+\frac{\partial}{\partial q_2}\right),$$

$$X_5=t\left(\frac{\partial}{\partial q_1}+\frac{\partial}{\partial q_2}\right),$$

with (nonzero) bracket relations

$$[X_1,X_2]=2X_1,\quad [X_1,X_5]=X_4,$$

$$[X_2,X_3]=-4X_4,\quad [X_2,X_5]=2X_5,$$

$$[X_3,X_4]=-2X_4,\quad [X_3,X_5]=-2X_5.$$

This Lie algebra L is solvable with $L^{(1)}=[L,L]=\{X_1,X_4,X_5\}$, $L^{(2)}=\{X_4\}$ and $L^{(3)}=\{0\}$.

For the case $\lambda_2\neq 0$ we obtain a Lie algebra which is identical with the one in Φ arbitrary.

$\Phi=\lambda_3\log\xi$:

Setting $\lambda_2=0$, V takes the form

$$V=\lambda_3\log(q_1-\lambda q_2), \tag{69}$$

$$T=2c_5t+c_6,$$

$$Q_1=\lambda c_2q_1+c_2q_2+2c_5q_1+\lambda d_2(t), \tag{70}$$

$$Q_2=c_2q_1+\frac{1}{\lambda}c_2q_2+2c_5q_2+d_2(t),$$

where $d_2(t)=c_3t+c_4$ (5-parameter group).

If $\lambda_2\neq 0$ the result again is the same as in Φ arbitrary.

Subcase 2e: $V=\lambda_1q_2^2+\Phi(q_1)$.

We obtain various forms of the generators depending on the form of the function Φ.

Φ arbitrary:

The generators take the form

$$T=c_6,\quad Q_1=0,\quad Q_2=c_4q_2+d_2(t), \tag{71}$$

where $d_2(t)$ satisfies the o.d.e. $d_2''+2\lambda_1d_2(t)=0$ (4-parameter group).

$\Phi=\lambda_2q_1^n$, $n\neq-2,0,1,2$:

If $\lambda_1=0$, then

$$T=2c_5t+c_6,\quad Q_1=\frac{4}{2-n}c_5q_1,\quad Q_2=c_4q_2+c_1t+c_2 \tag{72}$$

(5-parameter group).

If $\lambda_1\neq 0$, we set $c_5=0$. We end up with a 4-parameter group. It is the same as in Φ arbitrary.

We will calculate explicitly the Lie algebra for the potential $V(q_1,q_2)=q_1^3$. For a basis we choose the following five vector fields:

$$X_1=\frac{\partial}{\partial t},\quad X_2=t\frac{\partial}{\partial t}-2q_1\frac{\partial}{\partial q_1},$$

$$X_3=q_2\frac{\partial}{\partial q_2},\quad X_4=t\frac{\partial}{\partial q_2},\quad X_5=\frac{\partial}{\partial q_2},$$

with (nonzero) bracket relations

$$[X_1,X_2]=X_1,\quad [X_1,X_4]=X_5,$$

$$[X_2,X_4]=X_4,\quad [X_3,X_4]=-X_4,\quad [X_3,X_5]=-X_5.$$

This Lie algebra L is solvable with $L^{(1)}=[L,L]=\{X_1,X_4,X_5\}$, $L^{(2)}=\{X_5\}$ and $L^{(3)}=\{0\}$.

On the other hand, for the potential $V(q_1,q_2)=\frac{1}{2}q_2^2+q_1^3$ we obtain a 4-parameter group with basis

$$X_1=\frac{\partial}{\partial t},\quad X_2=q_2\frac{\partial}{\partial q_2},\quad X_3=\cos t\,\frac{\partial}{\partial q_2},\quad X_4=\sin t\,\frac{\partial}{\partial q_2}.$$

This Lie algebra L is also solvable with $L^{(1)}=[L,L]=\{X_3,X_4\}$ and $L^{(2)}=\{0\}$. It is isomorphic with the algebra of symmetries of potential (65) which we already examined.

$\Phi=\lambda_2 e^{\mu q_1}$:

If $\lambda_1=0$, then

$$T=2c_5t+c_6,\quad Q_1=\frac{-4}{\mu}c_5,\quad Q_2=c_4q_2+d_2(t),\tag{73}$$

where $d_2(t)$ satisfies the o.d.e. $d_2''+2\lambda_1 d_2=0$ (5-parameter group).

If $\lambda_1\neq 0$, we set $c_5=0$. It is the same case as in Φ arbitrary.

$\Phi=\lambda_2\log q_1$:

We set $\lambda_1=0$. Then $V=\lambda_2\log q_1$ and the generators take the form

$$T=2c_5t+c_6,\quad Q_1=2c_5q_1,\quad Q_2=c_4q_2+c_7t+c_8,\tag{74}$$

(5-parameter group).

Remark: The potentials that appear in this case are clearly integrable, being separable potentials. At this point we have completed the analysis of a separable potential with one variable missing. The potential $1/q_1^2$ was considered in subcase 2b. The potentials q_1^n for $n=0,1$ are covered by Case 1. The potential q_1^2 was considered in subcase 2d. The potential $f(q_1)$ falls under subcase 2e.

Subcase 2f:

Equations (36) and (37) are satisfied $[d_1(t)=\text{constant}$, $d_2(t)=\text{constant}]$. From Eqs. (34) and (35) we obtain the following results.

1.

$$V=q_2^N\Phi\left(\frac{q_1}{q_2}\right),\quad T=\tfrac{1}{2}c_3(2-N)t+c_6,\quad Q_1=c_3q_1,\quad Q_2=c_3q_2,\tag{75}$$

a 2-parameter group of transformations. The Lie algebra in this case is the two-dimensional non-Abelian Lie algebra with bracket $[X_1,X_2]=\tfrac{1}{2}(2-N)X_1$ if $N\neq 2$ and an Abelian 2-dimensional Lie algebra if $N=2$. We should mention that for certain choices of Φ we may obtain a larger symmetry group, e.g., for $\Phi(x)=x^N$, but generically the Lie algebra is 2-dimensional. Some values of N will also give different results. For example, $N=-2$ falls under subcase 2c.

Remark: In general, this system is not integrable, however, there are some integrable examples. We mention the Holt potentials,[19,20,33]

$$V(q_1,q_2)=q_2^{-2/3}(cq_2^2+q_1^2),\tag{76}$$

where $c=\frac{3}{4}$, $c=\frac{9}{2}$ and $c=12$.

Also the Fokas–Lagerström potential,[34]

$$V(q_1,q_2)=\frac{1}{(q_1^2-q_2^2)^{-2/3}}.\tag{77}$$

Case 2f includes Hénon–Heiles type potentials of the form $cq_2^3+q_1^2q_2$. They are integrable for the following values of c: $c=\frac{1}{3}$, $c=2$ and $c=\frac{16}{3}$.[35–37]

Finally, we mention the potential

$$V(q_1,q_2)=\frac{q_1}{q_2}.\tag{78}$$

It was shown by Hietarinta[38] *that the second integral for this potential is a transcendental function. It can be expressed as a combination of* W_+ *and* W_-, *the standard Whittaker functions, i.e., the solutions of the equation*

$$y'' + (\tfrac{1}{4}x^2 - a)y = 0. \tag{79}$$

2.

$$V = \lambda_1 \log q_2 + \Phi\left(\frac{q_1}{q_2}\right), \quad T = c_3 t + c_6,$$
$$Q_1 = c_3 q_1, \quad Q_2 = c_3 q_2, \tag{80}$$

a 2-parameter group of transformations. The Lie algebra in this case is the two-dimensional non-Abelian Lie algebra with bracket $[X_1, X_2] = X_1$.

3.

$$V = e^{\mu q_1}\Phi(q_2), \quad T = \tfrac{1}{2}c_5 t + c_6,$$
$$Q_1 = -\frac{c_5}{\mu}, \quad Q_2 = 0, \tag{81}$$

a 2-parameter group of transformations.

Remark: Taking $\Phi(q_2) = e^{-\mu q_2}$ *we obtain again the Toda lattice. However, we already have seen that this system has a 5-parameter group of transformations. Therefore, for the specific potential we do not obtain a maximal admitted algebra. Generically, the symmetry group is 2-dimensional. For example, taking* $V(q_1, q_2) = e^{q_1} q_2^3$ *gives a two-dimensional non-Abelian algebra with basis* $X_1 = \partial/\partial t$ *and* $X_2 = t(\partial/\partial t) - 2(\partial/\partial q_1)$.

4.

$$V = e^{\mu q_1}\Phi(q_1 - \lambda q_2), \quad T = -\tfrac{1}{2}\lambda\mu c_8 t + c_6$$
$$Q_1 = \lambda c_8, \quad Q_2 = c_8, \tag{82}$$

a 2-parameter group of transformations. The Lie algebra is again the two-dimensional non-Abelian Lie algebra with bracket $[X_1, X_2] = -\tfrac{1}{2}\lambda\mu X_1$.

Remark: We should mention that because of symmetry we do not list potentials of the form $V(q_1, q_2) = e^{\mu q_2}\Phi(q_2 - \lambda q_1)$. *We can also replace* $q_1 - \lambda q_2$ *with* $\alpha q_1 + \beta q_2$. *Taking* $\mu = 1$, $\alpha = 1$ *and* $\beta = -2$ *we obtain the potential* $V(q_1, q_2) = e^{q_1 - q_2} + e^{q_2}$. *This is a generalized Toda lattice associated with a Lie algebra of type* B_2, *first considered by Bogoyavlensky in Ref. 39.*

5.

$$V = \lambda_1 q_1^n + \lambda_2 q_2^m, \quad T = \tfrac{1}{2}c_5 t + c_6,$$
$$Q_1 = \frac{c_5}{2-n} q_1, \quad Q_2 = \frac{c_5}{2-m} q_2, \tag{83}$$

a 2-parameter group of transformations. Here, $n \neq 0,1,2$ and $m \neq 0,1,2$ and m, n not both equal to

-2. The Lie algebra in this case is the two-dimensional non-Abelian Lie algebra with bracket $[X_1,X_2]=\frac{1}{2}X_1$. The symmetry Lie algebra for the potentials 6–10 satisfies precisely the same bracket relation.

6.

$$V=\lambda_1 q_1^n+\lambda_2 \log q_2, \quad n\neq 0,1,2, \quad T=\tfrac{1}{2}c_5 t+c_6, \quad Q_1=\frac{c_5}{2-n}q_1, \quad Q_2=\frac{c_5}{2}q_2. \tag{84}$$

7.

$$V=\lambda_1 q_1^n+\lambda_2 e^{\mu q_2}, \quad n\neq 0,1,2, \quad T=\tfrac{1}{2}c_5 t+c_6, \quad Q_1=\frac{c_5}{2-n}q_1, \quad Q_2=-\frac{c_5}{\mu}. \tag{85}$$

8.

$$V=\lambda_1 \log q_1+\lambda_2 \log q_2, \quad T=\tfrac{1}{2}c_5 t+c_6, \quad Q_1=\frac{c_5}{2}q_1, \quad Q_2=\frac{c_5}{2}q_2. \tag{86}$$

9.

$$V=\lambda_1 \log q_1+\lambda_2 e^{\mu q_2}, \quad T=\tfrac{1}{2}c_5 t+c_6, \quad Q_1=\frac{c_5}{2}q_1, \quad Q_2=-\frac{c_5}{\mu}. \tag{87}$$

10.

$$V=\lambda_1 e^{\mu_1 q_1}+\lambda_2 e^{\mu_2 q_2}, \quad T=\tfrac{1}{2}c_5 t+c_6, \quad Q_1=-\frac{c_5}{\mu_1}, \quad Q_2=-\frac{c_5}{\mu_2}. \tag{88}$$

11.

$$V=\Phi(q_1^2+q_2^2), \quad T=c_6, \quad Q_1=c_1 q_2, \quad Q_2=-c_1 q_1, \tag{89}$$

a 2-parameter group of transformations. This is a unit mass in 2-dimensional space moving in a central field, i.e., a potential which is a function of r only. The function $q_1p_2-q_2p_1$ is a second integral. Note that in this case the Lie algebra is Abelian.

12.

$$V=\lambda(q_1^2+q_2^2)^n, \quad n\neq -1,0,1,$$

$$T=2c_5 t+c_6, \quad Q_1=c_1 q_2-\frac{2}{n-1}c_5 q_1, \quad Q_2=-c_1 q_1-\frac{2}{n-1}c_5 q_2, \tag{90}$$

a 3-parameter group of transformations. The Lie algebra is 3-dimensional with only nonzero bracket $[X_1,X_2]=2X_1$. The case $n=-\frac{1}{2}$ is a Kepler problem. For $n=-1$ the Lie algebra is 4-dimensional; it falls under subcase 2c. See (53) with $\lambda=c=0$ and $\mu=1$.

Remark: This potential is a special case of system 1 with $N=2n$. Taking $n=2$, we have a system of the form $aq_1^4+bq_1^2q_2^2+cq_2^4$. In general (for a, b, c non-zero) this system has a 2-dimensional group of symmetries unless $b=2a=2c$. Generically the potential $V(q_1,q_2)=aq_1^4+bq_1^2q_2^2+cq_2^4$ is not integrable, but for certain values of the parameters it becomes integrable. That is the case when $b=6a=6c$ or $a=16c$, $b=12c$ or $b=6a$, $c=8a$.[19,21,40]

13.

$$V=\lambda \log(q_1^2+q_2^2),$$

$$T=2c_5 t+c_6, \quad Q_1=c_1 q_2+2c_5 q_1, \quad Q_2=-c_1 q_1+2c_5 q_2, \tag{91}$$

a 3-parameter group of transformations. The Lie algebra, which is the same as in the previous case, may not seem interesting, but in n dimensions it is a direct sum of a 2-dimensional Lie algebra with $so(n, \mathbf{R})$.

14.

$$V=\lambda \sin^{-1}\frac{q_1^2-q_2^2}{q_1^2+q_2^2},$$
$$T=2c_5t+c_6, \quad Q_1=c_1q_2+2c_5q_1, \quad Q_2=-c_1q_1+2c_5q_2, \tag{92}$$

a 3-parameter group of transformations. This potential can be written in the form $\lambda_1+\lambda_2\theta$, in polar coordinates. In other words, it is a linear function of θ.

VI. GENERALIZATIONS

Case 1 generalizes in n dimensions. Consider a Hamiltonian with n degrees of freedom,

$$H=\frac{1}{2}\sum_{i=1}^{n} p_i^2+V(q_1,q_2,\ldots,q_n),$$

and the associated Lagrange–Newton equations,

$$\ddot{q}_i+V_{q_i}=0, \qquad i=1,2,\ldots,n \ . \tag{93}$$

As in the case of two degrees of freedom we seek point symmetries of Eqs. (93). We consider the equations

$$\Gamma^{(2)}\{\ddot{q}_i+V_{q_i}\}=0, \quad i=1,2,\ldots,n, \tag{94}$$

where $\Gamma^{(2)}$ is the second prolongation of

$$\Gamma=T\frac{\partial}{\partial t}+\sum_{i=1}^{n} Q_i\frac{\partial}{\partial q_i}. \tag{95}$$

Equations (94) give n identities of the form

$$E_i(t,q_1,q_2,\ldots,q_n,\dot{q}_1,\dot{q}_2,\ldots,\dot{q}_n)=0, \qquad i=1,2,\ldots,n, \tag{96}$$

where, we have used that $\dot{q}_i=-\partial V/\partial q_i$. The functions E_i are explicit polynomials in $\dot{q}_1,\dot{q}_2,\ldots,\dot{q}_n$. We impose the condition that Eqs. (96) are identities in the variables t, q_i, $\dot{q}_i$ which are regarded as independent.

Again, the functions T and Q_i must be of the form

$$T=a(t)+\sum_{i=1}^{n} b_i(t)q_i,$$
$$Q_i=\sum_{k=1}^{n} b_k'(t)q_iq_k+\sum_{k=1}^{n} c_{ik}(t)q_k+d_i(t), \quad i=1,2,\ldots,n \ . \tag{97}$$

We substitute (97) into (96). By considering the coefficient of $\dot{q}_k$ in E_j we obtain the following n^2 equations:

For $j\neq k$,

$$b_k''(t)q_j+c_{jk}'(t)+\frac{\partial V}{\partial q_j}b_k=0, \tag{98}$$

and for $j=k$,

$$2c_{jj}'(t)-a''(t)+3b_j''(t)q_j+3\frac{\partial V}{\partial q_j}b_j+\sum_{i\neq j}\left(b_i''(t)q_i+\frac{\partial V}{\partial q_i}b_i(t)\right)=0. \tag{99}$$

It follows from Eqs. (98) that V is quadratic of the form

$$V=\sum_{i=1}^{n}\lambda_i q_i^2+\sum_{i=1}^{n}\mu_i q_i, \tag{100}$$

unless $b_i(t)=0$ for $i=1,2,\ldots,n$.

Substituting (100) into (99), we obtain

$$b_i''+2\lambda_i b_i=0, \quad i=2,3,\ldots,n$$

and

$$b_i''+2\lambda_1 b_i=0, \quad i=2,3,\ldots,n.$$

Therefore for nonzero $b_i(t)$, we necessarily have

$$\lambda_1=\lambda_2=\cdots=\lambda_n.$$

Hence, V is of the form

$$V=\frac{\lambda}{2}\sum_{i=1}^{n}q_i^2,$$

where the linear terms are ignored.

One can easily deduce the form of the generators $a(t)$ is a solution of a second order equation of the form $a''+4\lambda a=c$. (3 parameters); $b_i(t)$ is a solution of $b_i''+\lambda b_i=0$. ($2n$ parameters); $d_i(t)$ is a solution of $d_i''+\lambda d_i=0$. ($2n$ parameters); $c_{ij}(t)$ are constant for $i\neq j$ and $c_{kk}=ct+c_k-2\lambda\int a(t)dt$ (n^2 parameters).

Therefore, the dimension of the symmetry algebra is $3+2n+2n+n^2=(n+2)^2-1$.

The case of the harmonic oscillator has been studied in Ref. 41 where it is shown that the symmetry group for a time-dependent harmonic oscillator is $SL(n+2, \mathbf{R})$.

When $\lambda=0$, the potential energy is zero and we have a free particle moving in $\mathbf{R}^n$. In this case the generators take the following simple form:

$$\begin{gathered} a(t)=c_1+c_2t+ct^2, \quad b_i(t)=\alpha_i+\beta_i t, \\ d_i(t)=\gamma_i+\delta_i t, \quad c_{ii}=\epsilon_i+ct, \\ c_{ij}=\kappa_{ij}, \quad i\neq j. \end{gathered} \tag{101}$$

The dimension is again $(n+2)^2-1$.

This dimension is in agreement with the results in Ref. 42, where upper bounds for the dimension of symmetry groups are obtained.

In case 2, $b_i(t)=0$ for $i=1,2,\ldots,n$ and Eqs. (99) and (100) imply that

$$c_{jk}(t)=c_{jk}, \quad \text{for } j\neq k,$$

$$c_{jj}(t)=\tfrac{1}{2}a'(t)+c_{jj},$$

where c_{jk} are constants. Equations (96) now become

$$\frac{1}{2}a'''(t)q_i+d_i''(t)+\sum_{k=1}^{n}\frac{\partial^2 V}{\partial q_i \partial q_k}\gamma_k+\frac{3}{2}a'(t)\frac{\partial V}{\partial q_i}-\sum_{k=1}^{n}c_{ik}\frac{\partial V}{\partial q_k}=0, \tag{102}$$

for $i=1,2,\ldots,n$, where

$$\gamma_k=\frac{1}{2}q_k a'(t)+\sum_{j=1}^{n}c_{kj}q_j+d_k(t). \tag{103}$$

Using Eqs. (102) we can prove the following.

(1) The potential q_1^2 has an n^2+3 parameter group of symmetries.
(2) The potential $1/q_1^2$ has an n^2+2 parameter group of symmetries.
(3) The potential $q_1^k, k\notin\{-2,0,1,2\}$ has an n^2+1 parameter group of symmetries.
(4) The potential $f(q_1)$ where f is arbitrary but not exponential, logarithmic or a power has an n^2 parameter group of symmetries.

We give the proof for the potential

$$V(q_1,q_2,\ldots,q_n)=\tfrac{1}{2}q_1^2.$$

The proof for the other three cases is similar. Since the variables $q_2,q_3,\ldots,q_n$ are missing, Eqs. (102) for $k=2,3,\ldots,n$ become

$$\tfrac{1}{2}a'''q_k+d_k''-c_{k1}q_1=0.$$

Therefore, $a'''=0$, $d_k(t)=\xi_k+\eta_k t$, and $c_{k1}=0$ for $k=2,3,\ldots,n$. On the other hand, the first equation in (102) gives

$$d_1''+\gamma_1+\left(\frac{3}{2}a'-c_{11}\right)q_1=0,$$

where

$$\gamma_1=\frac{1}{2}q_1 a'+\sum_{j=1}^{n}c_{1j}q_j+d_1.$$

Therefore, $a'(t)=0$, $d_1(t)=c_2\cos t+c_3\sin t$, and $c_{1j}=0$ for $j=2,3,\ldots,n$.

We obtain the following form for the generators:

$$\begin{aligned}T&=c_1,\\ Q_1&=c_{11}q_1+c_2\cos t+c_3\sin t,\\ Q_k&=\sum_{j=2}^{n}c_{kj}q_j+\xi_k+\eta_k t,\quad k=2,\ldots,n.\end{aligned} \tag{104}$$

Therefore, the dimension of the algebra of symmetries is $4+(n-1)^2+2(n-1)=n^2+3$.

ACKNOWLEDGMENTS

We would like to thank Peter Leach for sending us preprints of his work and also for some valuable remarks on the transcript. We would also like to thank Rui Fernandes, Giuseppe Gaeta, Peter Olver and the Referee for some helpful comments.

[1] G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations*, Applied Mathematical Science (Springer-Verlag, New York (1974), Vol. 13.
[2] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations*, Applied Mathematical Science (Springer-Verlag, New York, 1989), Vol. 81.
[3] P. J. Olver, *Applications of Lie groups to Differential Equations*, GTM (Springer, Verlag, New York, 1986), Vol. 107.
[4] L. V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).
[5] P. G. L. Leach, J. Math. Phys. **22**, 679 (1981).
[6] P. A. Damianou and C. Sophocleous, ''Symmetries of Hamiltonian systems with three degrees of freedom,'' Technical Report TR/8/1998, Department of Mathematics and Statistics, University of Cyprus.
[7] S. Lie, Arch. Math. **VIII**, 187 (1883).
[8] N. H. Ibragimov, *CRC Handbook of Lie Group Analysis of Differential Equations*, (CRC Press, Boca Raton, FL, 1996), Vols. 1–3.
[9] G. Gaeta, ''Nonlinear symmetries and nonlinear equations,'' *Mathematics and its Applications* (Kluwer, Dordrecht, 1994), Vol. 299.
[10] V. M. Gorringe and P. G. L. Leach, Quaest. Math. **11**, 95 (1988).
[11] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed. (Benjamin Cummings, Reading, MA, 1978).
[12] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, (Springer-Verlag, New York, 1978).
[13] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison-Wesley, Reading, MA, 1980).
[14] P. Libermann and C. M. Marle, *Symplectic Geometry and Analytical Mechanics* (Reidel, Dordrecht, MA, 1989).
[15] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed. (Dover, New York, 1944).
[16] B. Dorizzi, B. Grammaticos, and A. Ramani, J. Math. Phys. **24**, 2282 (1983).
[17] B. Dorizzi, B. Grammaticos, and A. Ramani, J. Math. Phys. **25**, 481 (1984).
[18] B. Grammaticos, B. Dorizzi, and A. Ramani, J. Math. Phys. **24**, 2289 (1983).
[19] B. Grammaticos, B. Dorizzi, and A. Ramani, J. Math. Phys. **25**, 3470 (1984).
[20] J. Hietarinta, Phys. Lett. A **96**, 273 (1983).
[21] J. Hietarinta, Phys. Rep. **147**, 87 (1987).
[22] A. M. Perelomov, *Integrable Systems of Classical Mechanics and Lie Algebras*, Vol. 1 (Birkhauser-Verlag, Basel, 1990).
[23] W. Sarlet and F. Cantrijn, SIAM (Soc. Ind. Appl. Math.) Rev. **23**, 467 (1981).
[24] P. G. L. Leach and K. S. Govinder, Lie Groups Appl. **1**, 146 (1994).
[25] F. Gonzalez-Gaston and A. Gonzalez-Lopes, Phys. Lett. A **129**, 153 (1988).
[26] A. A. Makarov, J. A. Smorodinsky, K. Valiev, and P. Winternitz, Nuovo Cimento A **L11a**, 1061 (1967).
[27] A. S. Fokas, ''Invariants, Lie–Bäcklund operators and Bäcklund transformations,'' Ph.D. thesis, California Institute of Technology, 1979.
[28] P. G. L. Leach, J. Aust. Math. Soc. B, Appl. Math. **23**, 173 (1981).
[29] I. Frish, J. A. Smorodinsky, M. Uhlirz, and P. Winternitz, Sov. J. Nucl. Phys. **4**, 444 (1967).
[30] K. S. Govinder and P. G. L. Leach, Quaest. Math. **16**, 405 (1993).
[31] H. Flaschka, Phys. Rev. **9**, 1924 (1974).
[32] M. Toda, *Theory of Nonlinear Lattices* (Springer-Verlag, New York, 1981).
[33] C. R. Holt, J. Math. Phys. **23**, 1037 (1982).
[34] A. Fokas and P. Lagerström, J. Math. Anal. Appl. **74**, 325 (1980).
[35] M. Henon and C. Heiles, Astron. J. **69**, 73 (1964).
[36] B. Grammaticos, B. Dorizzi, and R. Padjen, Phys. Lett. A **89**, 111 (1982).
[37] L. S. Hall, Physica D **8**, 90 (1983).
[38] J. Hietarinta, Phys. Rev. Lett. **52**, 1057 (1984).
[39] O. I. Bogoyanlensky, Commun. Math. Phys. **51**, 201 (1976).
[40] A. Ramani, B. Dorizzi, and B. Grammaticos, Phys. Rev. Lett. **49**, 1539 (1982).
[41] G. E. Prince and C. J. Eliezer, J. Phys. A **13**, 815 (1980).
[42] F. Gonzalez-Gaston and A. Gonzalez-Lopez, J. Math. Phys. **24**, 2006 (1983).

Superintegrability of the Calogero–Moser system: Constants of motion, master symmetries, and time-dependent symmetries

Manuel F. Rañada[a)]
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, 50009 Zaragoza, Spain

(Received 4 June 1998; accepted for publication 9 July 1998)

The classical n-dimensional Calogero–Moser system is a maximally superintegrable system endowed with a rich variety of symmetries and constants of motion. In the first part of the article some properties related with the existence of several families of constants of motion are analyzed. In the second part, the master symmetries and the time-dependent symmetries of this system are studied. © *1999 American Institute of Physics.* [S0022-2488(99)01401-2]

I. INTRODUCTION

A superintegrable system is a system that is integrable (in the sense of Liouville–Arnold) and that, in addition to this, possesses more globally defined constants of motion than degrees of freedom. If the number N of independent constants takes the value $N=2n-1$ (n the number of degrees of freedom) then the system is called maximally superintegrable (see Refs. 1–19). There are three well known cases of this very particular class of systems, namely, the Kepler problem, the isotropic harmonic oscillator, and the nonisotropic oscillator with rational frequencies. The Kepler possesses not only the energy and the angular momentum but also the Runge–Lenz vector; five of these integrals are functionally independent. The harmonic oscillator is a system trivially integrable since it is a direct sum of one degree of freedom systems. If the oscillator is isotropic then it has the angular momentum as an additional integral of motion. If the oscillator is nonisotropic then the angular momentum is not preserved but in the very particular case in which the quotients of the frequencies are rational the system has other additional nonlinear integrals. In all these three cases it is known that all the orbits became closed for the case of bounded motions. This high degree of regularity (the existence of periodic motions) is a consequence of the superintegrable character.

The system of Calogero–Moser[20–26] is a system of n particles in which the interaction force between every two particles is given by the inverse of the square of their relative distance. The Lagrangian of this system is given by

$$L_{CM}=\frac{1}{2}\sum_{j=1}^{n} v_j^2-\sum_{i<j} V_{ij}, \quad V_{ij}=\frac{c_0^2}{q_{ij}^2},$$

where $q_{ij}=q_i-q_j$, $i,j=1,2,\ldots,n$, and c_0 is an arbitrary constant (the masses of the particles are set equal to unity). This system has been extensively studied from many different viewpoints (see Refs. 27–29 for general reviews on the Calogero–Moser system and Refs. 30–36 for some articles published in these last years). Wojciechowsky proved in Ref. 4 that the Calogero–Moser system is not only integrable but even maximally superintegrable, that is, there exist $N=2n-1$ independent integrals of motion. More recently Kutnetsov[33] and Gonera[35] have also proved that the corresponding quantum system is superintegrable as well.

[a)]Electronic mail: mfran@posta.unizar.es

In this article we will study some properties closely related with the superintegrability of the Calogero–Moser system. The article is organized as follows: In Sec. II we discuss the existence and properties of the additional families of constants of motion. In Sec. III, we discuss the existence and the geometric properties of master symmetries, and in Sec. IV we study the existence of time-dependent constants of motion. Finally, in Sec. V we make some final comments.

II. SUPERINTEGRABILITY OF THE CALOGERO–MOSER SYSTEM

Moser proved[26] that the Calogero system can be presented as a Lax equation,

$$\frac{dA}{dt}=[A,B], \quad A=A_1+\mathrm{i}c_0A_2, \quad B=\mathrm{i}c_0(B_1-B_2),$$

where A_1 and B_1 denote the diagonal matrices,

$$A_1=\mathrm{diagonal}[v_1,v_2,\ldots,v_n], \quad B_1=\mathrm{diagonal}\left[\sum_{j\neq 1} x_{1j}^2,\sum_{j\neq 2} x_{2j}^2,\ldots,\sum_{j\neq n} x_{nj}^2\right]$$

and A_2 and B_2 take the form

$$A_2=[(1-\delta_{ij})x_{ij}], \quad B_2=[(1-\delta_{ij})x_{ij}^2],$$

where, for ease of notation, we use x_{ij} for denoting $1/q_{ij}$. The important point is that, because of the Lax equation, the traces of the powers of the matrix A are constants of motion,

$$I_k=\left(\frac{1}{k}\right)\mathrm{tr}\,A^k, \quad \frac{d}{dt}\,I_k=0, \quad k=1,2,\ldots,n.$$

Moreover, these n functions I_k, $k=1,2,\ldots,n$, are globally defined, independent, and in involution. They take the form

$$I_k=\left(\frac{1}{k}\right)(v_1^k+v_2^k+\cdots+v_n^k)+\text{terms of lower order in the velocities}.$$

Wojciechowsky proved in Ref. 4 the existence of an additional family of integrals K_j, $j=2,3,\ldots,n$. One of the main differences between this new family and the old one is that the first or leading term of the n functions I_k is q-independent (this term is dominant for large values of the q_i); whereas the first term of the $n-1$ functions K_j is linear in the coordinates q.

It is known (see, e.g., Refs. 18, 19) that the potential $V(x,y)=k/x^2$ (k is an arbitrary constant) is superintegrable. Two fundamental constants of motion, I_1 and I_2, are

$$I_1=v_y, \quad I_2=\left(\frac{1}{2}\right)(v_x^2+v_y^2)+\left(\frac{k}{x^2}\right),$$

and, concerning the third constant, one can choose any one of the two following quadratic functions:

$$K_2=(xv_y-yv_x)v_x-2k\left(\frac{y}{x^2}\right),$$

$$J_2=(xv_y-yv_x)^2+2k\left(\frac{y}{x}\right)^2.$$

This potential can be transformed into the $n=2$ Calogero–Moser potential by a rotation. Then the function K_2 becomes the $n=2$ particular case of the integral obtained by Wojciechowsky. Next we will prove that J_2, which is quadratic in the angular momentum, can also be considered as a very particular case of a family of integrals for the general n-particle Calogero–Moser system.

In the following we will make use of the matrix Q defined by

$$Q=\text{diagonal}[q_1,q_2,\dots,q_n].$$

Notice that the time-evolution of Q can be written as follows:

$$\dot{Q}=[Q,B]+A.$$

Let us denote by $T_r^{(q)}$ the following functions:

$$T_r^{(q)}=\text{tr}(Q^qA^r).$$

Then we have the following proposition.

Proposition 1: The traces $T_r^{(2)}$ and $T_r^{(1)}$ satisfy the following two properties:

$$(a)\quad \frac{d}{dt}T_r^{(2)}=2T_{r+1}^{(1)},$$

$$(b)\quad \frac{d}{dt}T_r^{(1)}=T_{r+1}^{(0)}.$$

Proof: The time derivative of $T_r^{(2)}$ can be written as

$$\frac{d}{dt}T_r^{(2)}=\frac{d}{dt}[\text{tr}(Q^2A^r)]=\text{tr}\left[\frac{d}{dt}(Q^2A^r)\right]=\text{tr}(M_1+M_2),$$

where the two matrices, M_1 and M_2, are given by

$$\begin{aligned}M_1=\dot{Q}QA^r+Q\dot{Q}A^r&=([Q,B]+A)QA^r+Q([Q,B]+A)A^r\\&=(Q^2BA^r-BQ^2A^r)+(AQA^r+QA^{r+1})\end{aligned}$$

and

$$M_2=\sum_{p=1}^{r}Q^2A^{p-1}\dot{A}A^{r-p}=\sum_{p=1}^{r}Q^2A^{p-1}[A,B]A^{r-p}=Q^2A^rB-Q^2BA^r.$$

Hence

$$\begin{aligned}\text{tr}(M_1+M_2)&=\text{tr}(Q^2A^rB-BQ^2A^r)+\text{tr}(AQA^r+QA^{r+1})\\&=\text{tr}([Q^2A^r,B])+2\,\text{tr}(QA^{r+1})=2T_{r+1}^{(1)}.\end{aligned}$$

Property (b) is proved in a similar way:

$$\frac{d}{dt}T_r^{(1)}=\frac{d}{dt}[\text{tr}(QA^r)]=\text{tr}\left[\frac{d}{dt}(QA^r)\right]=\text{tr}(N_1+N_2),$$

where the two matrices, N_1 and N_2, take the form

$$N_1=\dot{Q}A^r=(QB-BQ)A^r+A^{r+1}$$

and

$$N_2 = \sum_{p=0}^{r} QA^{p-1}\dot{A}A^{r-p} = \sum_{p=0}^{r} QA^{p-1}[A,B]A^{r-p} = QA^rB - QBA^r.$$

Hence

$$\mathrm{tr}(N_1+N_2) = \mathrm{tr}(QA^rB - BQA^r) + \mathrm{tr}\ A^{r+1} = \mathrm{tr}([QA^r,B]) + \mathrm{tr}\ A^{r+1} = T^{(0)}_{r+1}.$$

Notice that this proposition means that the traces $T^{(1)}_r$ are nonconstant functions generating integral of motion by a time derivation. Similarly, the $T^{(2)}_r$ generate also constants of motion but now making use of two successive time derivations.

Making use of the above proposition, it is easy to prove that the following two families of functions, K_{rs} and J_{rs}, defined by

$$K_{rs} = rI_rT^{(1)}_{s-1} - sI_sT^{(1)}_{r-1},$$

$$J_{rs} = (r+1)I_{r+1}T^{(2)}_{s-1} + (s+1)I_{s+1}T^{(2)}_{r-1} - 2T^{(1)}_rT^{(1)}_s,$$

are constants of motion. The first family of integrals, K_{rs}, $r,s=1,2,\ldots,n$, has been studied in Ref. 4.

Of course, the existence of these two families means an excessive number of integrals since the maximum number of (time-independent) functionally independent integrals is $N=2n-1$. Notice that $K_{sr}=-K_{rs}$ and $J_{sr}=J_{rs}$; so the total number of elements in every one of these two-parametric families is $(1/2)n(n-1)$ and $(1/2)n(n+1)$, respectively. We will focus our attention in the following more reduced one-parameter families $K_r \equiv K_{1r}$ and $J_{2r} \equiv (1/2)J_{rr}$. That is

$$K_r = I_1T^{(1)}_{r-1} - rI_rT^{(1)}_0,$$

$$J_{2r} = (r+1)I_{r+1}T^{(2)}_{r-1} - (T^{(1)}_r)^2.$$

The functions J_{2r}, $r=1,2\ldots,n$, generalize the above-mentioned function J_2 of the $n=2$ particle system.

Next we present the form of the integrals K_r and J_{2r} for the particular case $n=3$. We denote by L_{kj} the components of the angular momentum, that is, $L_{kj}=q_kv_j-q_jv_k$.

(a) The functions K_r have the following expressions:

$$\begin{aligned} K_2 &= [\mathrm{tr}(QA)]I_1 - [\mathrm{tr}(Q)](2I_2) \\ &= L_{21}(v_2-v_1) + L_{32}(v_3-v_2) + L_{13}(v_1-v_3) \\ &\quad + \text{other terms of lower order}, \end{aligned}$$

$$\begin{aligned} K_3 &= [\mathrm{tr}(QA^2)]I_1 - [\mathrm{tr}(Q)](3I_3) \\ &= L_{21}(v_2^2-v_1^2) + L_{32}(v_3^2-v_2^2) + L_{13}(v_1^2-v_3^2) \\ &\quad + \text{other terms of lower order}, \end{aligned}$$

$$\begin{aligned} K_4 &= [\mathrm{tr}(QA^3)]I_1 - [\mathrm{tr}(Q)](4I_4) \\ &= L_{21}(v_2^3-v_1^3) + L_{32}(v_3^3-v_2^3) + L_{13}(v_1^3-v_3^3) + \cdots. \end{aligned}$$

(b) The functions J_{2r} have the following expressions:

$$J_2 = [\mathrm{tr}(Q^2)](2I_2) - [\mathrm{tr}(QA)]^2 = L_{21}^2 + L_{32}^2 + L_{13}^2 + \text{other terms of lower order},$$

$$J_4=[\mathrm{tr}(Q^2A)](3I_3)-[\mathrm{tr}(QA^2)]^2$$
$$=L_{21}^2(v_2v_1)+L_{32}^2(v_3v_2)+L_{13}^2(v_1v_3)$$
$$+\text{other terms of lower order,}$$

$$J_6=[\mathrm{tr}(Q^2A^2)](4I_4)-[\mathrm{tr}(QA^3)]^2=L_{21}^2(v_2v_1)^2+L_{32}^2(v_3v_2)^2+L_{13}^2(v_1v_3)^2+\cdots .$$

Therefore, the leading term of the the family K_r is linear in the angular momentum, and the leading term of the family J_{2r} is quadratic.

III. MASTER SYMMETRIES

From this point we will make use of the Hamiltonian formalism.

The Hamiltonian phase space M is the $2n$-dimensional cotangent bundle $M=T^*Q$ of the n-dimensional configuration space Q. Cotangent bundles are manifolds endowed, in a natural or canonical way, with a symplectic structure ω_0 that in coordinates, $\{(q_i,p_i);\ i=1,2,\ldots,n\}$, is given by[37–40]

$$\omega_0=dq_i\wedge dp_i$$

(summation on the index i is understood). This symplectic structure defines a one-to-one relationship between the set $\mathfrak{X}(M)$ of vector fields on M and the set $\wedge^1(M)$ of one-forms on M as follows: To every vector field X we associate a one-form α_X given by the contraction of X with ω_0; that is, $X\mapsto\alpha_X=i(X)\omega_0$. Conversely, to every one-form α we can associate the vector field X_α uniquely determined as a solution of the equation $i(X_\alpha)\omega_0=\alpha$.

Given a differentiable function $F=F(q,p)$, the vector field X_F defined as the solution of the equation

$$i(X_F)\omega_0=dF,$$

is called the Hamiltonian vector field of the function F. At this point we make following three observations.

(1) The Hamiltonian vector field of a given function is uniqueley defined (without ambiguities). This uniqueness is a consequence of the symplectic character of the two-form ω_0.

(2) Suppose that we are given a Hamiltonian $H=H(q,p)$. Then the dynamics is given by the Hamiltonian vector field Γ_H of the Hamiltonian function. That is, $i(\Gamma_H)\omega_0=dH$.

(3) If I is a constant of motion for H then, according to the Noether approach to the dynamics, we can consider that I arises from a symmetry. In this case the symmetry is geometrically represented by the Hamiltonian vector field X_I of the function I.

Let us now return to the Calogero–Moser system. For ease of notation, we will frequently denote $T_r^{(1)}$ just by T_r, and we write rI_r instead of $T_r^{(0)}$. We will not change the notation for the constants of motion and we will continue writing I_r, K_r, J_{2r}, $r=1,2,\ldots,n$, for the corresponding functions but now considered as functions of the momenta. We will denote by Γ_r the Hamiltonian vector field of I_r except for the case Γ_2 corresponding to $I_2=H$ that will be denoted just by Γ.

Proposition 2: Let $X_r^{(1)}$ be the Hamiltonian vector field of the function $T_r^{(1)}$ with respect the canonical symplectic structure ω_0, and let $\widetilde{X}_r^{(1)}$ be the vector field defined by $\widetilde{X}_r^{(1)}=[X_r^{(1)},\Gamma]$. Then the vector field $\widetilde{X}_r^{(1)}$ is a symmetry of the dynamical vector field Γ.

Proof: Let us denote by Γ_r^K the Hamiltonian vector field of the function K_r,

$$i(\Gamma_r^K)\omega_0=dK_r.$$

The function K_r is given by

$$K_r = I_1 T_{r-1} - r I_r T_0 ;$$

therefore

$$dK_r = I_1 dT_{r-1} - (rI_r) dT_0 + T_{r-1} dI_1 - (rT_0) dI_r .$$

Thus the vector field Γ_r^K takes the form

$$\Gamma_r^K = I_1 X_{r-1}^{(1)} - (rI_r) X_0^{(1)} + T_{r-1}\Gamma_1 - (rT_0)\Gamma_r .$$

Hence, the Lie bracket of Γ_r^K with the dynamical vector field Γ can be written as follows:

$$[\Gamma_r^K,\Gamma] = I_1[X_{r-1}^{(1)},\Gamma] - (rI_r)[X_0^{(1)},\Gamma] + T_{r-1}[\Gamma_1,\Gamma] - (rI_0)[\Gamma_r,\Gamma]$$
$$-\Gamma(I_1)X_{r-1}^{(1)} + r\Gamma(I_r)X_0^{(1)} - \Gamma(T_{r-1})\Gamma_1 + r\Gamma(T_0)\Gamma_r .$$

The functions I_r, K_r, $r=1,2,\dots,n$, are constants of motion, and the associated vector fields Γ_r, Γ_r^K, are symmetries of Γ,

$$\Gamma(I_r)=0, \quad [\Gamma_r,\Gamma]=0, \quad [\Gamma_r^K,\Gamma]=0, \quad r=1,2,\dots,n,$$

Thus, if we denote by $\tilde{X}_r^{(1)}$ the vector fields,

$$\tilde{X}_r^{(1)} = [X_r^{(1)},\Gamma], \quad r=1,2,\dots,n,$$

we arrive at

$$I_1\tilde{X}_{r-1}^{(1)} = (rI_r)\tilde{X}_0^{(1)} + \Gamma(T_{r-1})\Gamma_1 - r\Gamma(T_0)\Gamma_r = (rI_r)[\tilde{X}_0^{(1)} + \Gamma_1] - rI_1\Gamma_r .$$

Notice that Γ_1 and $X_0^{(1)}$ that are the Hamiltonian vector fields of the functions $I_1 = \mathrm{tr}(A) = p_1 + p_2 + \cdots + p_n$, and $T_0^{(1)} = \mathrm{tr}(Q) = q_1 + q_2 + \cdots + q_n$, take the following form in coordinates:

$$\Gamma_1 = \frac{\partial}{\partial q_1} + \frac{\partial}{\partial q_2} + \cdots + \frac{\partial}{\partial q_n}, \qquad X_0^{(1)} = -\frac{\partial}{\partial p_1} - \frac{\partial}{\partial p_2} - \cdots - \frac{\partial}{\partial p_n}.$$

Making use of these two expressions we obtain

$$[X_0^{(1)},\Gamma] = -\Gamma_1 .$$

So, we thus arrive at the following two equations:

$$\tilde{X}_{r-1}^{(1)} = -r\Gamma_r , \quad [\tilde{X}_{r-1}^{(1)},\Gamma]=0,$$

and the proposition is proved.

In differential geometric terms, a symmetry of the dynamics is a vector field $X \in \mathfrak{X}(M)$ such that $[X,\Gamma_H]=0$ (this geometric approach is also valid in the Lagrangian case but then M is given by the tangent bundle $M=TQ$). A vector field X that satisfies the following two properties:

$$[X,\Gamma_H] \neq 0, \quad [[X,\Gamma_H],\Gamma_H]=0,$$

is called a ''master symmetry'' or a ''generator of symmetries'' of degree $m=1$ for Γ_H (see Refs. 41–43). If X is such that

$$[X,\Gamma_H] \neq 0, \;\; [[X,\Gamma_H],\Gamma_H] \neq 0, \quad \text{and } [[[X,\Gamma_H],\Gamma_H],\Gamma_H]=0,$$

then it is called a ''master symmetry'' or a ''generator of symmetries'' of degree $m=2$.

Next we illustrate this situation with a simple example. The Hamiltonian H and the vector field Γ_H of the $n=1$ free particle are given by

$$H=\left(\frac{1}{2}\right)p^2, \quad \Gamma_H=p\,\frac{\partial}{\partial q}.$$

Then the vector field $X=\partial/\partial p$ satisfies

$$\left[\frac{\partial}{\partial p},p\,\frac{\partial}{\partial q}\right]=\frac{\partial}{\partial q}, \quad \text{and} \left[\frac{\partial}{\partial q},p\,\frac{\partial}{\partial q}\right]=0.$$

So $X=\partial/\partial p$ is a ''master symmetry'' of degree $m=1$ for the free particle.

Consequently, and according to proposition 2, the vector fields $X_r^{(1)}$, $r=1,2\ldots,n$, are ''master symmetries'' of degree $m=1$ for the Calogero-Moser system. Now we consider the case of the functions $T_r^{(2)}$.

Proposition 3: The Hamiltonian vector field $X_r^{(2)}$ of the function $T_r^{(2)}$ is a ''generator of symmetries'' of degree $m=2$.

Proof: Let us denote by Γ_{2r}^J the Hamiltonian vector field of the function J_{2r},

$$i(\Gamma_{2r}^J)\omega_0=dJ_{2r}.$$

The function J_{2r} is given by

$$J_{2r}=(r+1)\,I_{r+1}T_{r-1}^{(2)}-(T_r)^2;$$

therefore

$$dJ_{2r}=(r+1)I_{r+1}dT_{r-1}^{(2)}+(r+1)T_{r-1}^{(2)}dI_{r+1}-2T_r dT_r.$$

Hence the vector field Γ_{2r}^J takes the form

$$\Gamma_{2r}^J=(r+1)I_{r+1}X_{r-1}^{(2)}+(r+1)T_{r-1}^{(2)}\Gamma_{r+1}-2T_rX_r^{(1)}.$$

The Lie bracket of Γ_{2r}^J with the dynamical vector field Γ is given by

$$\begin{aligned}[\Gamma_{2r}^J,\Gamma]&=(r+1)I_{r+1}[X_{r-1}^{(2)},\Gamma]+(r+1)T_{r-1}^{(2)}[\Gamma_{r+1},\Gamma]-2T_r[X_r^{(1)},\Gamma]\\&\quad-(r+1)\Gamma(I_{r+1})X_{r-1}^{(2)}-(r+1)\Gamma(T_{r-1}^{(2)})\Gamma_{r+1}+2\Gamma(T_r)X_r^{(1)}.\end{aligned}$$

We recall first that the functions I_r, $r=1,2,\ldots,n$ are constants of motion, and that the associated vector fields Γ_r are symmetries of Γ, and second that the Lie derivatives of $T_r^{(1)}$ and $T_r^{(2)}$ with respect to Γ are given by

$$\Gamma(T_r^{(1)})=T_{r+1}^{(0)}=(r+1)I_{r+1}, \qquad \Gamma(T_r^{(2)})=2T_{r+1}^{(1)}.$$

Thus, if we denote by $\widetilde{X}_r^{(1)}$ and $\widetilde{X}_r^{(2)}$ the following vector fields:

$$\widetilde{X}_r^{(1)}=[X_r^{(1)},\Gamma], \quad \widetilde{X}_r^{(2)}=[X_r^{(2)},\Gamma], \quad r=1,2,\ldots,n,$$

the Lie bracket of Γ_{2r}^J with Γ becomes

$$\begin{aligned}[\Gamma_{2r}^J,\Gamma]&=(r+1)I_{r+1}\widetilde{X}_{r-1}^{(2)}-2T_r\widetilde{X}_r^{(1)}-2(r+1)T_r^{(1)}\Gamma_{r+1}+2(r+1)I_{r+1}X_r^{(1)}\\&=(r+1)I_{r+1}[\widetilde{X}_{r-1}^{(2)}+2X_r^{(1)}].\end{aligned}$$

The vector field Γ_{2r}^J is a symmetry of the dynamics; so it satisfies, $[\Gamma_{2r}^J,\Gamma]=0$. Therefore, we have arrived at the following three properties:

$$[X_r^{(2)},\Gamma]=\tilde{X}_r^{(2)}\neq 0,\quad r=1,2,\ldots,n,$$

$$[\tilde{X}_r^{(2)},\Gamma]=-2\tilde{X}_{r+1}^{(1)}=2(r+2)\Gamma_{r+2},$$

$$[[\tilde{X}_r^{(2)},\Gamma],\Gamma]=0,$$

which state that $X_r^{(2)}$ is a ''generator of symmetries'' for Γ of degree $m=2$.

We close this section with the following observations.

It is known that if R and S are constants of motion for a Hamiltonian system then so is the Poisson bracket $\{R,S\}$. For the Calogero–Moser system the situation is more general since the functions M_{rs} defined by $M_{rs}=\{T_r,I_s\}$ are also constants of motion,

$$\frac{d}{dt}M_{rs}=\left\{\frac{dT_r}{dt},I_s\right\}+\left\{T_r,\frac{dI_s}{dt}\right\}=0.$$

Nevertheless they do not represent new functions. We have directly verified that

$$M_{rs}=(r+s-1)I_{r+s-1},\quad r,s=1,2,3.$$

In geometric terms this property is as follows. The vector field Z_{rs} defined by $Z_{rs}=[X_r,\tilde{X}_s]$ is a symmetry of the dynamics

$$[Z_{rs},\Gamma]=[[\Gamma,\tilde{X}_s],X_r]+[\tilde{X}_r,\tilde{X}_s]=0.$$

Then it can be proven that

$$i(Z_{rs})\omega_0=(s+1)d\{T_r,I_{s+1}\}.$$

Consequently, the constant M_{rs+1} is the Hamiltonian function of the vector field Z_{rs}.

An important point is that if H is an arbitrary Hamiltonian, and F a function such that its Hamiltonian vector field X_F is a master symmetry for Γ_H, then $I_F=\Gamma_H(F)$ is a constant of motion for H. Conversely, if I is a constant of motion for H, and if the two equations,

$$i(X_I)\omega_0=dI,\quad [X,\Gamma_H]=X_I,$$

admit as a solution a well defined complete vector field X, then X is a master symmetry for Γ_H.

Therefore, the existence of master symmetries can be considered as a property closely related with the theory of constants of motion. Nevertheless, in most of cases, obtaining master symmetries implies a very difficult calculus. The important point concerning the Calogero–Moser system is that, as this system is very peculiar, we can obtain, in a simple and direct way, the Hamiltonian functions, $T_r^{(1)}$ and $T_r^{(2)}$, of the master symmetries of the system.

IV. TIME-DEPENDENT SYMMETRIES

The time-dependent Hamiltonian formalism has as phase space the manifold M defined as $M=T^*Q\times\mathbb{R}$, and as a fundamental geometric structure the (nonsymplectic) two-form Ω_H defined as

$$\Omega_H=\omega_0+dH\wedge dt.$$

The dynamics is now given by the unique time-dependent vector field Γ_H defined as solution of the following two equations:

$$i(\Gamma_H)\Omega_H=0, \quad i(\Gamma_H)dt=1.$$

Notice that, in fact, the dynamical equation is the first one; the second equation must be considered as a restriction that leads to the uniqueness of the solution. In coordinates Γ_H takes the form

$$\Gamma_H=\left(\frac{\partial H}{\partial p_i}\right)\frac{\partial}{\partial q_i}-\left(\frac{\partial H}{\partial q_i}\right)\frac{\partial}{\partial p_i}+\frac{\partial}{\partial t}.$$

We now apply the results obtained in Sec. III to the theory of time-dependent symmetries of the Calogero–Moser system. We first recall that a time-independent Hamiltonian $H=H(q,p)$ can also be endowed with time-dependent constants of motion $I=I(q,p,t)$.

Let H be a time-independent Hamiltonian, and X a (time-independent) master symmetry for Γ_H. Then the time-dependent vector field Y_X defined by[42–43]

$$Y_X=X+t[X,\Gamma_H]+(\tfrac{1}{2})t^2[[X,\Gamma_H],\Gamma_H]$$

is a time-dependent symmetry of H. If the degree of X is $m=1$ then Y_X will determine a time-dependent constant of motion I^t linear in the time t; if $m=2$ then this constant will be quadratic in t.

As an example, let us consider once more the case of the $n=1$ free particle. Then we have

$$X=\frac{\partial}{\partial p}, \quad Y_X=t\frac{\partial}{\partial q}+\frac{\partial}{\partial p}$$

and

$$\left[t\frac{\partial}{\partial q}+\frac{\partial}{\partial p},p\frac{\partial}{\partial q}+\frac{\partial}{\partial t}\right]=0.$$

Notice that in this case,

$$i(Y_X)\Omega_H=-d(q-tp),$$

so, the time-dependent constant of motion $I=I(q,p,t)$ arising from the vector field Y_X is $I=q-tp$.

We denote by $Y_r^{(1)}$ the time-dependent symmetry of the Calogero–Moser system determined by $X_r^{(1)}$,

$$Y_r^{(1)}=X_r^{(1)}+t[X_r^{(1)},\Gamma].$$

In order to obtain the associated time-dependent constant of motion, we must calculate the contraction of the vector field $Y_r^{(1)}$ with the two-form Ω_H,

$$i(Y_r^{(1)})\Omega_H=i(X_r^{(1)})\omega_0+[i(X_r^{(1)})dH]dt+ti(\tilde{X}_r^{(1)})\omega_0+t[i(\tilde{X}_r^{(1)})dH]dt.$$

We have

$$i(X_r^{(1)})\omega_0=dT_r^{(1)},$$

$$i(X_r^{(1)})dH=X_r^{(1)}(dH)=-\Gamma(T_r^{(1)})=-(r+1)I_{r+1},$$

$$i(\tilde{X}_r^{(1)})\omega_0=-(r+1)i(\Gamma_{r+1})\omega_0=-(r+1)dI_{r+1},$$

$$i(\tilde{X}_r^{(1)})dH=\tilde{X}_r^{(1)}(H)=-(r+1)\Gamma_{r+1}(H)=0.$$

Thus, we arrive at

$$i(Y_r^{(1)})\Omega_H = d[T_r^{(1)} - (r+1)tI_{r+1}].$$

Consequently the time-dependent functions I_r^t, $r=1,2,\ldots,n$, defined by

$$I_r^t = T_{r-1}^{(1)} - rtI_r,$$

are time-dependent constants of motion for the Calogero–Moser system.

Next we consider the time-dependent symmetry of the Calogero–Moser system determined by $X_r^{(2)}$,

$$Y_r^{(2)} = X_r^{(2)} + t\tilde{X}_r^{(2)} + (\tfrac{1}{2})t^2[\tilde{X}_r^{(2)},\Gamma].$$

In this second case the contraction of the vector field $Y_r^{(2)}$ with the two-form Ω_H is given by

$$i(Y_r^{(2)})\Omega_H = i(X_r^{(2)})\omega_0 + [i(X_r^{(2)})dH]dt + ti(\tilde{X}_r^{(2)})\omega_0 + t[i(\tilde{X}_r^{(2)})dH]dt$$
$$+ (\tfrac{1}{2})t^2 i([\tilde{X}_r^{(2)},\Gamma])\omega_0 + (\tfrac{1}{2})t^2[i([\tilde{X}_r^{(2)},\Gamma])dH]dt.$$

Now we have

$$i(X_r^{(2)})\omega_0 = dT_r^{(2)},$$

$$i(X_r^{(2)})dH = X_r^{(2)}(H) = -\Gamma(T_r^{(2)}) = -2T_{r+1}^{(1)},$$

$$i(\tilde{X}_r^{(2)})\omega_0 = -2i(X_{r+1}^{(1)})\omega_0 = -2dT_{r+1}^{(1)},$$

$$i(\tilde{X}_r^{(2)})dH = -2i(X_{r+1}^{(1)})(dH) = 2\Gamma(T_{r+1}^{(1)}) = 2(r+2)I_{r+2},$$

$$i([\tilde{X}_r^{(2)},\Gamma])\omega_0 = 2(r+2)i(\Gamma_{r+2})\omega_0 = 2(r+2)dI_{r+2},$$

$$i([\tilde{X}_r^{(2)},\Gamma])dH = 2(r+2)\Gamma_{r+2}(H) = 0.$$

Thus, we arrive at

$$i(Y_r^{(2)})\Omega_H = d[T_r^{(2)} - 2tT_{r+1}^{(1)} + (r+2)t^2 I_{r+2}].$$

Consequently the time-dependent functions J_r^t, $r=1,2,\ldots,n$, defined by

$$J_r^t = T_{r-1}^{(2)} - 2tT_r^{(1)} + (r+1)t^2 I_{r+1},$$

are time-dependent constants of motion for the Calogero–Moser system.

Now, with the knowledge of these time-dependent integrals, we can consider a new approach to the two time-independent families of constants obtained in Sec. II.

It can be proven that, after some calculus, the integrals K_{rs} and J_{rs} can be rewritten as follows:

$$K_{rs} = (rI_r)I_s^t - (sI_s)I_r^t,$$

$$J_{rs} = (r+1)I_{r+1}J_s^t + (s+1)I_{s+1}J_r^t - 2I_{r+1}^t I_{s+1}^t.$$

So, in a sense, the existence of K_{rs} and J_{rs} can be considered as a consequence of the existence of I_r^t and J_r^t. Everyone of the K_{rs} (with the two indices) appears as arising from an algebraic pairing of I_r^t with I_s^t. Similarly, every one of the J_{rs} arises from a pairing of J_r^t and I_{r+1}^t with J_s^t and I_{s+1}^t. Notice that, in both constructions, the previous existence of the fundamental family I_r is necessary.

We close this section with the $n=3$ particular case as an example. In this simple case the time-dependent functions I_1^t, I_2^t, and J_1^t, have the following expressions:

$$I_1^t=(q_1+q_2+q_3)-t(p_1+p_2+p_3),$$

$$I_2^t=(q_1p_1+q_2p_2+q_3p_3)-t[p_1^2+p_2^2+p_3^2+2(V_{12}+V_{23}+V_{13})],$$

$$J_1^t=(q_1^2+q_2^2+q_3^2)-2t(q_1p_1+q_2p_2+q_3p_3)-t^2[p_1^2+p_2^2+p_3^2+2(V_{12}+V_{23}+V_{13})].$$

V. FINAL COMMENTS

The Calogero–Moser system is a maximally superintegrable system that is endowed with a great amount of different symmetries. The associated constants can be grouped in families. The first and fundamental family is the one studied by Moser in Ref. 26 which is directly related with the Lax equation. There exist, in addition to this fundamental one, other different families. In this article we have analyzed the properties of four families. Two of them are time-independent,

$$K_{rs}=rI_rT_{s-1}^{(1)}-sI_sT_{r-1}^{(1)},$$

$$J_{rs}=(r+1)I_{r+1}T_{s-1}^{(2)}+(s+1)I_{s+1}T_{r-1}^{(2)}-2T_r^{(1)}T_s^{(1)},$$

and the other two are time-dependent,

$$I_r^t=T_{r-1}^{(1)}-rtI_r,$$

$$J_r^t=T_{r-1}^{(2)}-2tT_r^{(1)}+(r+1)t^2I_{r+1}.$$

Finally, let us comment that the existence of some integrable Calogero-related systems, e.g., Calogero interacting particles in an external field have been proven. It is known that for the particular case $n=2$ the constant J_2 is compatible with a harmonic oscillator (see, e.g., Refs. 2, 18, 19). We think that a prolongation of this study could be the analysis of similar properties for the n-dimensional Calogero–Moser system wih harmonic oscillators.

Note added in proof. We have found the additional Ref. 44 that also studies the existence of master symmetries and the Calogero–Moser system.

ACKNOWLEDGMENTS

I thank L. J. Boya, J. F. Cariñena, and M. Santander, for comments on the theory of superintegrable systems and hidden symmetries. This article is part of a project supported by the grants PB-96-0717 and AEN-97-1680 (DGICYT, Madrid).

[1] M. Henon, ''Numerical exploration of Hamiltonian systems,'' in *Chaotic Behaviour of Deterministic Systems*, edited by G. Iooss, R. H. Helleman, and R. Stora (North-Holland, Amsterdam, 1983).
[2] T. I. Fris, V. Mandrosov, Y. A. Smorodinsky, M. Uhlir, and P. Winternitz, Phys. Lett. **16**, 354–356 (1965).
[3] A. A. Makarov, Y. A. Smorodinsky, K. Valiev, and P. Winternitz, Nuovo Cimento A **52**, 1061–1084 (1967).
[4] S. Wojciechowski, Phys. Lett. A **95**, 279–281 (1983).
[5] L. Chetouani, L. Guechi, and T. F. Hammann, Phys. Lett. A **125**, 277–281 (1987).
[6] M. V. Carpio-Bernido and C. C. Bernido, Phys. Lett. A **134**, 395–399 (1987).
[7] J. Hietarinta, Phys. Rep. **147**, 87–154 (1987).
[8] M. Kibler and P. Winternitz, Phys. Lett. A **147**, 338–342 (1990).
[9] N. W. Evans, Phys. Rev. A **41**, 5666–5676 (1990).
[10] N. W. Evans, Phys. Lett. A **147**, 483–486 (1990).
[11] N. W. Evans, J. Math. Phys. **32**, 3369–3375 (1991).
[12] M. A. del Olmo, M. A. Rodríguez, and P. Winternitz, J. Math. Phys. **34**, 5118–5139 (1993).

[13] M. A. del Olmo, M. A. Rodríguez, and P. Winternitz, Fortschr. Phys. **44**, 199–233 (1996).
[14] J. A. Calzada, M. A. del Olmo, and M. A. Rodríguez, J. Geom. Phys. **23**, 14–30 (1997).
[15] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, Fortschr. Phys. **43**, 453–521 (1995).
[16] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, Fortschr. Phys. **43**, 523–563 (1995).
[17] A. L. Salas-Brito, H. N. Núñez-Yépez, and R. P. Martínez-Romero, Int. J. Mod. Phys. A **12**, 271–276 (1997).
[18] M. F. Rañada, J. Math. Phys. **38**, 4165–4178 (1997).
[19] M. F. Rañada and M. Santander, ''Superintegrable systems on the two-dimensional sphere S^2,'' preprint Univ. Zaragoza, 1998.
[20] F. Calogero, J. Math. Phys. **10**, 2191–2196 (1969).
[21] F. Calogero, J. Math. Phys. **10**, 2197–2200 (1969).
[22] C. Marchioro, J. Math. Phys. **11**, 2193–2196 (1970).
[23] B. Sutherland, J. Math. Phys. **12**, 246–250 (1971).
[24] F. Calogero, J. Math. Phys. **12**, 419–436 (1971).
[25] F. Calogero and C. Marchioro, J. Math. Phys. **15**, 1425–1430 (1974).
[26] J. Moser, Adv. Math. **16**, 197–220 (1975).
[27] M. A. Olshanetsky and A. M. Perelomov, Phys. Rep. **71**, 313–400 (1981).
[28] F. Calogero, ''Integrable dynamical systems,'' in *Nonlinear Phenomena*, Lecture Notes in Physics, edited by K. B. Wolf (Springer, 1983), Vol. 189.
[29] A. M. Perelomov, *Integrable Systems of Classical Mechanics and Lie Algebras* (Birkhauser, Basel, 1990).
[30] J. F. van Diejen, J. Math. Phys. **35**, 2983–3004 (1994).
[31] J. F. van Diejen, J. Math. Phys. **36**, 1299–1323 (1994).
[32] V. Z. Enolskii and J. C. Eilbeck, J. Phys. A **28**, 1069–1088 (1995).
[33] V. B. Kuznetsov, Phys. Lett. A **218**, 212–222 (1996).
[34] O. Babelon and M. Talon, Phys. Lett. A **236**, 462–468 (1997).
[35] C. Gonera, Phys. Lett. A **237**, 365–368 (1998).
[36] M. F. Rañada, Phys. Lett. A **238**, 29–34 (1998).
[37] R. Abraham and J. Marsden, *Foundations of Mechanics* (Benjamin, Reading, MA, 1978).
[38] G. Marmo, E. J. Saletan, A. Simoni, and B. Vitale, *Dynamical Systems: A Differential Geometric Approach* (Wiley, Chichester, 1985).
[39] M. de León and P. R. Rodrigues, *Methods of Differential Geometry in Analytical Mechanics* (North-Holland, Amsterdam, 1989).
[40] G. Morandi, C. Ferrario, G. Lo Vecchio, G. Marmo, and C. Rubano, Phys. Rep. **188**, 147–284 (1990).
[41] B. Fuchssteiner, Prog. Theor. Phys. **70**, 1508–1522 (1983).
[42] P. A. Damianou, J. Phys. A **26**, 3791–3796 (1993).
[43] R. L. Fernandes, J. Phys. A **26**, 3797–3803 (1993).
[44] W. Devel and M. Falck, Prog. Theor. Phys. **75**, 1328–1341 (1986).

The study of dromion interactions of $(2+1)$-dimensional integrable systems

Hang-Yu Ruan
Institute of Modern Physics, Zhejiang University, Hangzhou, 310027, People's Republic of China and Institute of Modern Physics, Normal College of Ningbo University, Ningbo, 315211, People's Republic of China[a)]

Yi-Xin Chen
Institute of Modern Physics, Zhejiang University, Hangzhov, 310027, People's Republic of China

(Received 18 May 1998; accepted for publication 9 September 1998)

Starting from a two-line soliton solution of an integrable $(2+1)$-dimensional system in bilinear form, one can find a dromion solution that is localized in all directions for a suitable potential. The interaction between two dromions is studied in detail through the method of figure analysis for a $(2+1)$-dimensional modified Korteweg–deVries (KdV) system, sine–Gordon system and Sawada–Kotera system. Except for a phase shift, there are no changes in the shape and velocity of the dromions after interactions for these models. The interactions of dromions for these models are not only elastic (there is no exchange of energy) but also irrotational (there is no exchange of angular momentum). © *1999 American Institute of Physics.* [S0022-2488(98)03112-0]

I. INTRODUCTION

Recently, since the pioneering work of Boiti *et al.*,[1] the study of the exponentially localized soltion solutions, called dromions, in $(2+1)$-dimensions has been attracting much attention from physicists and mathematicians. Usually, dromion solutions are driven by two or more nonparallel straight-line ghost solitons. For instance, for the Davey–Stewartson (DS)[2] and the Nizhnik–Novikov–Veselov (NVV)[3] equations, their dromion solutions are driven by two perpendicular line ghost solitons.[1,4] For the Kadomtsev–Petviashivili (KP) equation, the dromion solutions are driven by nonperpendicular line ghost solitons.[5] For one type of nonlinear models, say, the DS, NNV, and asymmetrical NNV(ANNV)[6] equations, the dromion solutions exist for the physical fields. However, for other types of equations like the KP and the breaking soliton equations, the dromion solutions exist only for some suitable potentials of the field.[5,7] More recently, even more generalized dromion solutions which are driven by curved and straight-line solitons for some types of $(2+1)$-dimensional nonlinear modes are found.[8]

In this paper, we are interested in the interaction of dromions for $(2+1)$-dimensional integrable systems. It is well known that the interactions of $(1+1)$-dimensional solitons are elastic. There is no exchange of energy (no change of shape and velocity) among interacting solitons. We hope to know whether the interactions among dromions for $(2+1)$-dimensional integrable systems are elastic or not. Different from $(1+1)$-dimensional cases, the interactions between two $(2+1)$-dimensional objects with finite size, say, dromions, may lead to rotations. So we hope to know also whether there are exchanges of angular momentum among interacting dromions. In other words, whether the interactions among dromions are rotational or irrotational.

The paper is organized as follows. In Sec. II, the multi-dromion solutions are given for three $(2+1)$-dimensional integrable systems, modified KdV (MKdV), Sawada–Kotera (SK), and sine–Gordon (SG) systems. The plots of interaction of two dromions for two systems are shown in Sec. III. Section IV includes a summary and discussion.

[a)]Mailing address.

II. MULTI-DROMION SOLUTION OF A THREE (2+1)-DIMENSIONAL INTEGRABLE SYSTEM

A. Multi-dromion solutions of the MKdV system

The bilinear form of the $(2+1)$-dimensional MKdV system can be written as

$$A(D_X)(f\cdot f+g\cdot g)\equiv A(D_x,D_y,D_t)(f\cdot f+g\cdot g)=0, \tag{1}$$

$$B(D_X)f\cdot g\equiv B(D_x,D_y,D_t)f\cdot g=0, \tag{2}$$

where A and B are even and odd functions of their variables $X=(x,y,t)$ and $D_X=(D_x,D_y,D_t)$, respectively. The D-operators are defined by[9,10]

$$D_x^nD_y^mD_t^pf\cdot g\equiv(\partial_x-\partial_{x'})^n(\partial_y-\partial_{y'})^m(\partial_t-\partial_{t'})^p(f(X)\cdot g(X'))\big|_{X'=X}. \tag{3}$$

It can be proved that a single dromion solution of the equation system (1) and (2) exists if a physical field is defined suitably.

$$u=L(\partial_X)K(\partial_X)(\tan^{-1}(g/f))\equiv(a_1\partial_x+b_1\partial_y)(a_1\partial_x+b_2\partial_y)(\tan^{-1}(g/f)), \tag{4}$$

where a_1, b_1, a_2 and b_2, should be selected such that the linear operators $L(\partial_X)$ and $K(\partial_X)$ annihilate two-line solitons,

$$f=1+a_{12}\exp(\eta_1+\eta_2),\quad g=\exp(\eta_1)+\exp(\eta_2), \tag{5}$$

$$\eta_i=p_ix+q_iy+w_it+\text{const}\equiv P_i\cdot X+\text{const}, \tag{6}$$

with

$$P_i=(p_i,q_i,w_i)\quad(i=1,2),\quad B(P_i)=0,\quad a_{12}=-\frac{A(P_1-P_2)}{A(P_1+P_2)}. \tag{7}$$

That is to say, in the space–time (x,y,t), the dromion solutions are driven by ghost line solitons which are nonparallel to each other. Two-line solitons are annihilated by two-linear operators $L(\partial_X)$ and $K(\partial_X)$ while a dromion which is located at the cross point of the two-line solitons is survived. Taking a space transformation

$$p_1x+q_1y=px_1,\quad p_2x+q_2y=qy_1,\quad \Delta\equiv p_1q_2-p_2q_1\neq0, \tag{8}$$

and fixing the constants a_i and b_i, in Eq. (4) as $a_1=-q_1q/\Delta$, $b_1=p_1q/\Delta$, $a_2=q_2p/\Delta$, and $b_2=-p_2p/\Delta$, we can rewrite Eqs. (4)–(6) as

$$u=(a_1\partial_x+b_1\partial_y)(a_2\partial_x+b_2\partial_y)(\tan^{-1}(g/f))\equiv\partial_{x_1}\partial_{y_1}(\tan^{-1}(g/f)), \tag{9}$$

$$f=1+a_{12}\exp(\eta_1+\eta_2),\quad g=\exp(\eta_1)+\exp(\eta_2),\quad \eta_1=px_1+\text{const},\quad \eta_2=qy_1+\text{const}. \tag{10}$$

Now, let us discuss in detail the dromion structures for the following $(2+1)$-dimensional integrable MKdV equation:[11]

$$A(D_X)(f\cdot f+g\cdot g)=D_x^2(f\cdot f+g\cdot g)=0, \tag{11}$$

$$B(D_X)f\cdot g=(D_x^2D_t+D_x^5+D_y)f\cdot g=0. \tag{12}$$

By means of the general method developed by Hirota, the N-line soliton solution of the equation system (11) and (12) can be written as

$$f(x,y,t)=\sum_{n=0}^{N/2}\sum_{{}_NC_{2n}}a(i_1,i_2,\ldots,i_{2n})\exp(\eta_{i1}+\eta_{i2}+\cdots+\eta_{i2n}), \tag{13}$$

$$g(x,y,t)=\sum_{m=0}^{[(N-1)/2]}\sum_{{}_NC_{2m+1}} a(j_1,j_2,\ldots,j_{2m+1})\exp(\eta_{j1}+\eta_{j2}+\cdots+\eta_{j2m+1}), \tag{14}$$

$$a(i_1,i_2,\ldots,i_n)=\begin{cases}\prod_{k,l}^{(n)} a(i_k,i_l) & \text{for } n\geqslant 2,\\ 1, \quad n=0,1,\end{cases} \tag{15}$$

$$a(i_k,i_l)=-\frac{A(P_{ik}-P_{il})}{A(P_{ik}+P_{il})}=-\frac{(p_{ik}-p_{il})^2}{(p_{ik}+p_{il})^2}, \tag{16}$$

$$\eta_i=p_ix+q_iy+w_it+\eta_{i0}, \tag{17}$$

$$B(P_{ik})=p_i^2w_i^2+p_i^5+q_i=0. \tag{18}$$

Here $|N/2|$ denotes the maximum integer which does not exceed $N/2$ and n_{i0} is an arbitrary but finite real constant related to the phase of the ith soliton. ${}_NC_n$ indicates summation over all possible combination of n elements taken from N, and $\Pi_{i,1}^{(n)}$ indicates the product of all possible combinations of the n elements. It can be known from Eqs. (8) and (9) that the multi-dromion solutions for the potential u given by (9) are allowed only for a special form such that two linear operators $a_i\partial_x+b_i\partial_y(i=1,2)$ with fixed a_i and b_i annihilate all the line solitons, since a_i and b_i are dependent of q_i and p_i. In other words the only allowed line solitons must be perpendicular to the axes in the new space coordinates x_1 and y_1. So the multi-dromion solutions exist only for the following potential form.

$$u=\partial_{x1}\partial_{y1}(\tan^{-1}(g(x_1,y_1,t)/f(x_1,y_1,t))), \tag{19}$$

where the forms of $g(x_1,y_1,t)$ and $f(x_1,y_1,t)$ are the same as those of (13) and (14), but η_i should be taken as

$$\eta_i=p_ix+q_iy+w_it+\eta_{i0}=p_i'x_1+w_it+\eta_{i0} \quad \text{or} \quad \eta_i=p_ix+q_iy+w_it+\eta_{i0}=q_i'y_1+w_it+\eta_{i0}. \tag{20}$$

As an example, we write down the explicit forms of f and g for $N=3$:

$$f(x,y,t)=1+a(1,2)\exp(\eta_1+\eta_2)+a(1,3)\exp(\eta_1+\eta_3)+a(2,3)\exp(\eta_2+\eta_3), \tag{21}$$

$$g(x,y,t)=\exp(\eta_1)+\exp(\eta_2)+\exp(\eta_3)+a(1,2)a(1,3)a(2,3)\exp(\eta_1+\eta_2+\eta_3), \tag{22}$$

$$\eta_i=p_i'x+w_it+\eta_{i0} \quad \text{or} \quad \eta_i=q_iy+w_it+\eta_{i0}, \tag{23}$$

$$p_i^2w_i+p_i^5+q_i=0, \tag{24}$$

$$a(i,j)=-\frac{(p_i-p_j)^2}{(p_i+p_j)^2}. \tag{25}$$

B. Multi-dromion solutions of a $(2+1)$-dimensional sine-Gordon system

The bilinear form of a $(2+1)$-dimensional sine-Gordon system is given by

$$A(D_X)(f\cdot f-g\cdot g)\equiv A(D_x,D_y,D_t)(f\cdot f-g\cdot g)=0, \tag{26}$$

$$B(D_X)f\cdot g\equiv B(D_x,D_y,D_t)f\cdot g=0, \tag{27}$$

where A and B are even functions of their variables. For simplicity, we shall discuss the dromion structure of the sine–Gordon system for a special selection of the operators A and B:

$$A(D_X)(f\cdot f-g\cdot g)\equiv D_xD_t(f\cdot f-g\cdot g)=0, \tag{28}$$

$$B(D_X)f\cdot g=(D_x^3D_t+D_yD_t+a)f\cdot g=0. \tag{29}$$

Performing the similar procedure leading to Eqs. (13)–(15), we find that the N-line soliton solution of the equation system (28) and (29) possesses the same form as Eqs. (13)–(15) but with

$$a(i_k,i_l)=\frac{A(p_{ik}-p_{il})}{A(p_{ik}+p_{il})}=\frac{(p_{ik}-p_{il})(w_{ik}-w_{il})}{(p_{ik}+p_{il})(w_{ik}+w_{il})}, \tag{30}$$

$$\eta_i=p_ix+q_iy+w_it=\eta_{i0}, \tag{31}$$

$$B(P_{ik})=p_i^3w_i+q_iw_i+a=0. \tag{32}$$

From the dispersion relation (32), we see that both p_i and q_i can be selected as zero. So in order to obtain the dromion solution of the sine–Gordon system (28) and (29), we can simply take $\eta_i=p_ix+w_it+\eta_{i0}$, or $\eta_i=q_iy+w_it+\eta_{i0}$. In this case, the linear operators to annihilate all the line solitons can be taken as $L(\partial_X)=\partial_x$ and $K(\partial_X)=\partial_y$. Then the physical field with dromion solutions reads

$$u=\partial_x\partial_y(\tan^{-1}(g(x,y,t)/f(x,y,t))). \tag{33}$$

C. Multi-dromion solutions of a $(2+1)$-dimensional Sawada–Kotera equation

The n-line soliton solution of the following $(2+1)$-dimensional SK equation in bilinear form,

$$A(D_X)=(D_x^6+5D_yD_x^3-5D_y^2+D_xD_t)F\cdot F=0, \tag{34}$$

can be written as

$$F(x,y,t)=1+\sum_{i=1}^{n}\exp(\eta_i)+\sum_{i<j}^{n}a_{ij}\exp(\eta_i+\eta_j)+\sum_{i<j<k}a_{ij}a_{jk}a_{ik}\exp(\eta_i+\eta_j+\eta_k)+\cdots+\left(\prod_{i<j}a_{ij}\right)\exp\sum_{k=1}^{n}\eta_k, \tag{35}$$

where

$$a_{ij}=-\frac{A(P_i-P_j)}{A(P_i+P_j)}=-\frac{(p_i-p_j)^6+5(q_i-q_j)(p_i-p_j)^3-5(q_i-q_j)^2+(p_i-p_j)(w_i-w_j)}{(p_i+p_j)^6+5(q_i+q_j)(p_i+p_j)^3-5(q_i+q_j)^2+(p_i+p_j)(w_i+w_j)}, \tag{36}$$

$$\eta_i=p_ix+q_iy+w_it+\eta_{i0}, \tag{37}$$

$$p_i^6+5q_ip_i^3-5q_i^2+p_iw_i=0. \tag{38}$$

It may be proved that multi-dromion solutions of Eq. (33) exist if the physical field is defined as

$$u=\partial_{x1}\partial_{y1}\ln F(x_1,y_1,t), \tag{39}$$

where x_i and y_1 are new space coordinates. In the new space–time (x_1,y_1,t), the only allowed line solitons must be perpendicular to the axes and η_i in function $F(x_1,y_1,t)$ should be taken as

$$\eta_i=p_ix+q_iy+w_it+\eta_{i0}=p_i'x_1+w_it+\eta_{i0}\quad\text{or}\quad\eta_i=p_ix+q_iy+w_it+\eta_{i0}=q_i'y_1+w_it+\eta_{i0}. \tag{40}$$

Substituting (34) with (35), (37), and (39) into (38), we obtain a single $(2+1)$-dimensional dromion solution which is localized in all directions for $N=2$, two dromion solution for $N=3$,

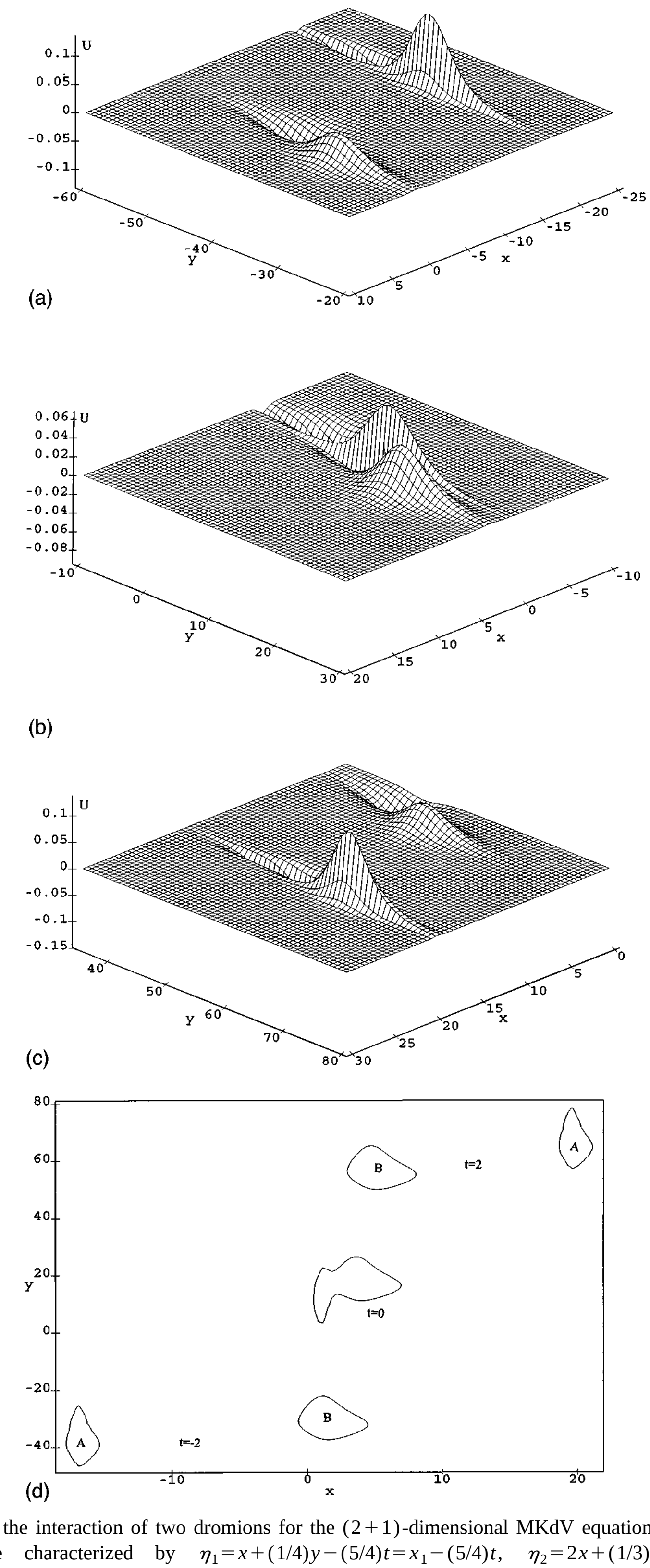

FIG. 1. The plots of the interaction of two dromions for the $(2+1)$-dimensional MKdV equation formed by three-line solitons which are characterized by $\eta_1 = x + (1/4)y - (5/4)t = x_1 - (5/4)t$, $\eta_2 = 2x + (1/3)y - (97/12)t = (1/3)y_1 - (97/12)t$, and $\eta_3 = 3x + (1/2)y - (487/18)t = 3x_1 - (487/18)t$. The times of the figures read: (a) $t = -2$, (b) $t = 0$, (c) $t = 2$, (d) is a cross-section plot $(u = 0.005)$ in correspondence with (a)–(c).

three or four dromion solutions for $N = 4$, and so on. The number of the dromions of an n-line soliton solution is determined by how many intersect points exist in the lines. As an example we write down the explicit forms of F for $N = 3$:

$$F(x_1,y_1,t)=1+\exp(\eta_1)+\exp(\eta_2)+\exp(\eta_3)+a_{12}\exp(\eta_1+\eta_2)+a_{13}\exp(\eta_1+\eta_3)$$
$$+a_{23}\exp(\eta_2+\eta_3)+a_{12}a_{13}a_{23}\exp(\eta_1+\eta_2+\eta_3), \tag{41}$$

$$\eta_i=p_ix+q_iy+w_it+\eta_{i0}=p_i'x_1+w_it+\eta_{i0} \quad \text{or} \quad \eta_i=p_ix+q_iy+w_it+\eta_{i0}=q_i'y_1+\eta_{i0}, \tag{42}$$

$$p_i^6+5q_ip_i^3-5q_i^2+p_iw_i=0, \tag{43}$$

$$a_{ij}=-\frac{A(p_i-p_j)}{A(p_i+p_j)}. \tag{44}$$

III. DROMION INTERACTIONS

It is known that in $(1+1)$-dimensions, there is no exchange of physical quantities like energy and momentum of the solitons after collision. Except for the phase shifts, the velocities and shapes all remained unchanged.

We hope to know whether the similar property is valid or not for the interactions among dromions. Especially, we hope to known whether the dromions rotate or not. In other words, is there exchange of angular momentum when the $(2+1)$-dimensional ''objects,'' dromions, collide?

It is difficult to study analytically the interaction of the dromions because of the complexity of the multi-dromion solutions. It is more straightforward to study the interaction of the dromions graphically.

Figure 1 is the interaction plot of two dromions (which are formed by three ghost line solitons) for the MKdV equation, where three ghost line solitons are characterized by

$$\eta_1=x+\frac{1}{4}y-\frac{5}{4}t=x_1-\frac{5}{4}t,$$
$$\eta_2=2x+\frac{1}{3}y-\frac{97}{12}t=\frac{1}{3}y_1-\frac{97}{12}t, \tag{45}$$
$$\eta_3=3x+\frac{1}{2}y-\frac{487}{18}t=\frac{1}{2}y_1-\frac{487}{18}t,$$

respectively. In Figs. 1(a)–(c), the time t is taken as -2, 0, and 2, respectively. Figure 1(d) is a cross-section plot of the two dromions before and after interaction in correspondence with Figs. 1(a)–(c), while $u=\text{const.}=0.005$. From Figs. 1(a)–(d), we see that the shapes of dromions are not changed. Especially from Fig. 1(d), one can see clearly that the shapes and the directions (if we locate a vector on every dromion) of two dromions are totally the same, which means there is also no exchange of the angular momentum between the dromions when they are interacting. The only change found from Fig. 1(d) is the phase shifts of the interacting dromions.

Figure 2 is the interaction plot of two dromions for sine–Gordon equation, where

$$\eta_1=x-t, \quad \eta_2=\frac{1}{2}y-2t, \quad \eta_3=\frac{3}{2}x-\frac{8}{27}t, \tag{46}$$

and Fig. 2(d) is a cross-section plot of the two dromions in correspondence with Figs. 2(a)–(c) at time $t=-15$, $t=0$, and $t=15$ for $u=\text{const}=-0.001$. As in Fig. 1, from the Figs. 2(a)–(d), we can see all interacting phenomenon between the interaction of the dromions for the sine–Gordon system (28) and (29). There are no exchanges of the energy, momentum, and angular momentum except the phase shifts.

For the same way, the interacting property between dromions for the SK equation is studied and the same conclusions are obtained. Except for the phase shifts, there is no change at all for the dromions after collision; especially, there is no rotation.

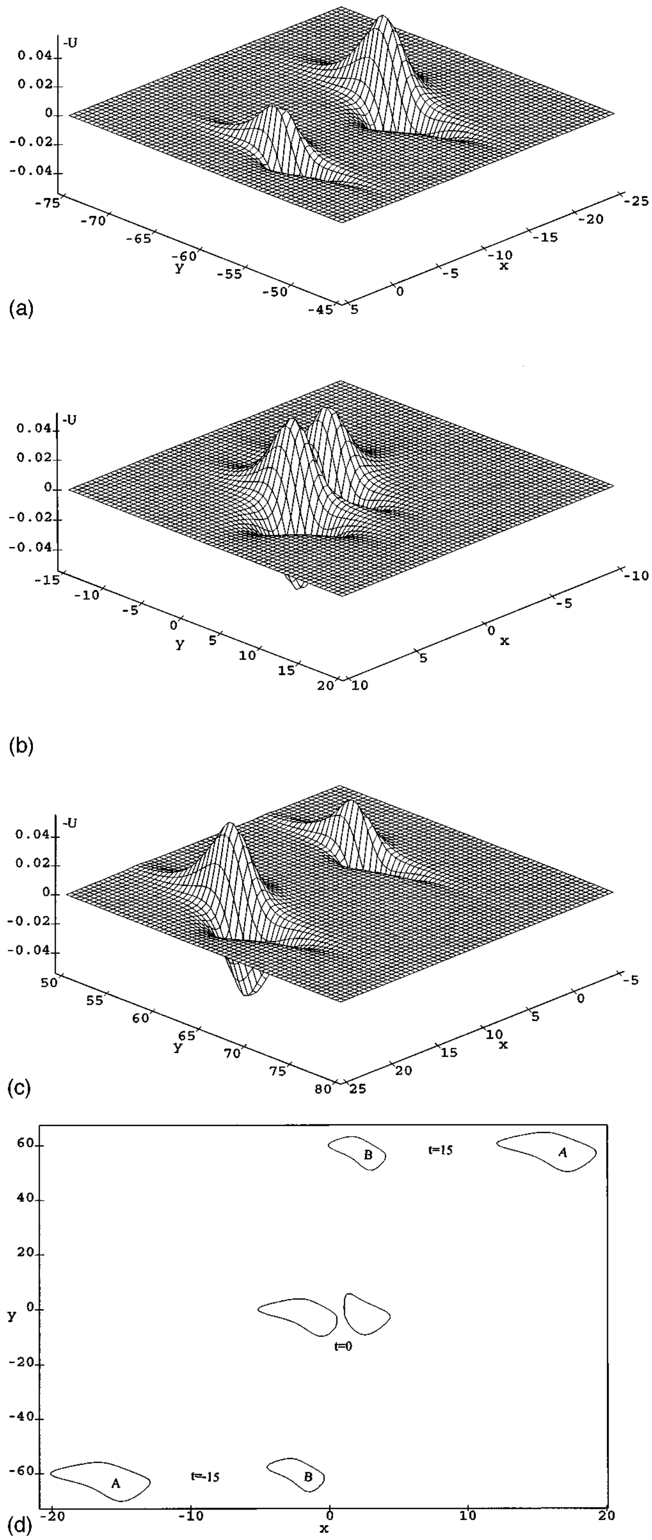

FIG. 2. The plots of the interaction of two dromions for a $(2+1)$-dimensional sine—Gordon equation. The related three-line solitons are determined by $\eta_1 = x - t$, $\eta_2 = (1/2)y - 2t$, and $\eta_3 = (3/2)x - (8/27)t$. The times of the figures read: (a) $t=-15$, (b) $t=0$, (c) $t=15$; (d) is a cross-section plot ($u=-0.001$) in correspondence with (a)–(c).

IV. SUMMARY AND DISCUSSIONS

In summary, we have constructed multidromion solutions of the $(2+1)$-dimensional MKdV-type, $(2+1)$-dimensional sine–Gordon-type, and the $(2+1)$-dimensional SK-type equations for some suitable potentials. The multidromions are constructed by multi-line solitons, say, a single dromion is constructed by two-line solitons, two-dromion solutions are constructed by three line solitons. All the line solitons should be parallel to the new axises $\{x_1, y_1\}$.

For $(1+1)$-dimensional integrable models, like the KdV equation, the interaction among solitons are completely elastic. There is no energy and momentum exchange among solitons when they are interacting. The only effect of the soliton interaction is the phase shifts. In $(2+1)$-dimensional cases, the similar conclusions for three types of models, MKdV, sine–Gordon, and SK models, are obtained. Figures 1(c) and 2(c) show us that the ''shapes'' and velocities of two dromions in the two systems are totally the same before and after interactions. The only effect observed from Figs. 1(d) and 2(d) of the interactions is the phase shifts. That is to say there is no energy and momentum exchange when the dromions finishing the interaction. Different from the $(1+1)$-dimensional case, there may be exchange of another physical quantity, angular momentum, for the collision of $(2+1)$-dimensional objects. However, for the three types of $(2+1)$-dimensional integrable models studied in this paper, we find that no rotations occur when the dromions are interacting. In other words, there is no angular momentum exchange among the interaction dromions of the three types of $(2+1)$-dimensional models.

We believe that the conclusions obtained in this paper are true for all $(2+1)$-dimensional integrable models and $(3+1)$-dimensional models (if there exist), though we obtained the results only from three types of $(2+1)$-dimensional models. We hope this conviction can be checked in a future study.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China and the Nature Science foundation of Zhejiang Province. We would like to thank Professor S-y Lou for his helpful discussions.

[1] M. Boiti, J. J. P. Leon, L. M. Martina, and F. Pempinelli, Phys. Lett. A **132**, 432 (1988).
[2] A. Davey and K. Stewartson, Proc. R. Soc. London, Ser. A **360**, 592 (1978).
[3] L. P. Nizhnik, Sov. Phys. Dokl. **25**, 707 (1980); A. P. Veselovand and S. P. Novikov, Sov. Math. Dokl. **30**, 588, 705 (1984); S. P. Novikov and A. P. Veselov, Physica D **18**, 267 (1986).
[4] R. Radha and M. Lakshmanan, J. Math. Phys. **35**, 4746 (1994).
[5] J. Hietarintr, Phys. Lett. A **149**, 133 (1990).
[6] M. Boiti, J. J. P. Leon, M. Manna, and F. Penpinelli, Inverse Probl. **2**, 271 (1986); **3**, 25 (1987).
[7] R. Radha and M. Lakshmanan, Phys. Lett. A **197**, 7 (1995).
[8] S-y. Lou, J. Phys. A **28**, 7227 (1995).
[9] R. Hirota, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Spinger, Berlin, 1980).
[10] R. Hirota, Phys. Rev. Lett. **27**, 1192 (1971).
[11] K. M. Tamizhmani, A. Ramani, and B. Grammaticos, J. Math. Phys. **32**, 2635 (1991).

Spin and wave function as attributes of ideal fluid

Yuri A. Rylov[a)]
Institute for Problems in Mechanics, Russian Academy of Sciences, 101, bild.1 Vernadskii Avenue, Moscow, 117526, Russia

(Received 17 April 1998; accepted for publication 18 September 1998)

An ideal fluid whose internal energy depends on density, density gradient, and entropy is considered. Dynamic equations are integrated, and a description in terms of hydrodynamic (Clebsch) potentials occurs. All essential information on the fluid flow (including initial and boundary conditions) appears to be carried by the dynamic equations for hydrodynamic potentials. Information on initial values of the fluid flow is carried by arbitrary integration functions. Initial and boundary conditions for potentials contain only nonessential information concerning the fluid particle labeling. It is shown that the description in terms of n-component complex wave function is a kind of such description in terms of hydrodynamic potentials. Spin determined by the irreducible number n_m of the wave function components appears to be an attribute of the fluid flow. Classification of fluid flows by the spin appears to be connected with invariant subspaces of the relabeling group. [S0022-2488(99)01001-4]

I. INTRODUCTION

An ideal (nondissipative) fluid with the internal energy E of a very general form is considered. The internal energy E is supposed to depend on the fluid density ρ, density gradient $\nabla\rho$, and entropy per unit mass S. The stress tensor for such fluid has the form

$$P^{\alpha\beta}=\delta_{\alpha\beta}\left[\rho^2\,\frac{\partial E}{\partial\rho}+\frac{\partial(\rho E)}{\partial\rho_\gamma}\,\rho_\gamma\right]-\rho\partial_\alpha\,\frac{\partial(\rho E)}{\partial\rho_\beta},\quad \alpha,\beta=1,2,3$$

$$\rho_\alpha\equiv\partial_\alpha\rho,\quad \partial_i\equiv\frac{\partial}{\partial x^i},\quad i=0,1,2,3. \tag{1}$$

If $E=E(\rho,S)$ does not depend on $\nabla\rho$, the stress tensor has the form

$$P^{\alpha\beta}=p\,\delta^\alpha_\beta,$$

where $p=\rho^2\partial E/\partial\rho$ is the pressure. Conventionally the dependence of the internal energy on $\nabla\rho$ is not considered. There are two motives for the consideration of such an unusual fluid.

First, the proper dependence of E on $\nabla\rho$ prevents sound waves from tilting. Indeed, let the internal energy have the form

$$E=E_0(\rho,S)+a(\nabla\rho/\rho)^2, \tag{2}$$

where a is a small positive quantity. For usual laminar flows, where $\nabla\rho/\rho$ is small, the last term of (2) does not give a significant contribution to the stress tensor (1), and it is of no importance whether or not there is the last term in (2). In the case of the wave tilting the last term in (2) comes to be principal. At the front of the tilted sound wave E tends to ∞, and the tilting of the wave may be stopped.

[a)]Electronic mail: rylov@ipmnet.ru

Second, fluid models with an internal energy of a very general form are used for the description of statistical ensembles of stochastic particles. By definition a statistical ensemble $\mathcal{E}[\mathcal{S}_{st}]$ of stochastic particles $\mathcal{S}_{st}$ is a set of many independent identical stochastic particles $\mathcal{S}_{st}$. Usually the term ''statistical ensemble'' is associated with some tool for the calculation of average values of physical quantities. But this tool is effective, provided the statistical ensemble is a set of deterministic (nonstochastic) particles. In reality the principal property of the statistical ensemble is formulated as follows. *The statistical ensemble of many stochastic (or deterministic) particles is a deterministic dynamic system*. This statement seems rather unexpected, because for a statistical ensemble of deterministic systems this property looks like a trivial one. In statistical physics, statistical ensembles of deterministic systems are mainly considered (the only exception is the statistical ensemble of Brownian particles), and the statistical ensemble's property of being a deterministic dynamic system needs some explanation.[1,2]

The result of an experiment with a single stochastic particle $\mathcal{S}_{st}$ is irreproducible. But distributions of results of similar experiments with many independent stochastic particles are reproducible. Projecting many independent identical stochastic particles $\mathcal{S}_{st}$ to the same space–time region, one obtains a cloud $\mathcal{E}[N,\mathcal{S}_{st}]$ of N independent identical particles $\mathcal{S}_{st}$ moving randomly. With the number N of particles tending to ∞, this cloud $\mathcal{E}[\infty,\mathcal{S}_{st}]$ may be considered as a continuous medium, or as a fluid. This fluid is a deterministic dynamic system, because experiments with the fluid $\mathcal{E}[\infty,\mathcal{S}_{st}]$ are reproducible. Besides, any reproducible experiments with the stochastic particle can be described in terms of the fluid $\mathcal{E}[\infty,\mathcal{S}_{st}]$ without reference to any probabilistic construction (i.e., without reference to the statistical ensemble's property of being a tool for the calculation of average values). The probabilistic constructions are effective only if the statistical ensemble $\mathcal{E}[\infty,\mathcal{S}_d]$ consists of deterministic particles $\mathcal{S}_d$, whose properties can be determined independent of $\mathcal{E}[\infty,\mathcal{S}_d]$. In the case of $\mathcal{E}[\infty,\mathcal{S}_{st}]$ these probabilistic constructions (probability density, or probability amplitude) are needed only for the interpretation of the fluid in terms of a single stochastic particle (see Ref. 3 for details).

For instance, let us consider a single electron $\mathcal{S}_{st}$, flying from an electron gun, passing through a narrow slit in a diaphragm and hitting a screen at a point x_1. Another electron $\mathcal{S}_{st}$, prepared in the same way, hits the screen at another point x_2 which does not coincide with x_1. In other words, an experiment with a single electron is irreproducible in general. It means that a single electron is a stochastic particle. Let us consider a series of N ($N\rightarrow\infty$) experiments with identically prepared independent electrons. The distribution of N impact points over the screen is reproducible, i.e., it is approximately the same for other series of N experiments. It means that a set $\mathcal{E}[N,\mathcal{S}_{st}]$ of N ($N\rightarrow\infty$) independent identical electrons $\mathcal{S}_{st}$ is a deterministic dynamic system, although a single electron $\mathcal{S}_{st}$ is a stochastic system. If $\mathcal{E}[\infty,\mathcal{S}_{st}]$ can be considered to be a fluid, then by solving dynamic equations for this fluid and calculating the flux of the fluid $\mathcal{E}[\infty,\mathcal{S}_{st}]$ through the screen, one can calculate the diffraction picture (the distribution of the impact points over the screen). For such calculation one needs only characteristics of the dynamic system $\mathcal{E}[\infty,\mathcal{S}_{st}]$ (dynamic equations, expressions for the particle flux and the energy–momentum tensor). Any quantum axiomatics and corresponding probabilistic constructions (wave function, linear operators, commutation relations, etc.) are not needed. It means that quantum effects can be explained and calculated as purely dynamical effects.[3]

On the other hand, quantum particles are described conventionally in terms of wave functions. The wave function is considered to be a fundamental object which cannot be defined via other more fundamental objects. As a result, like any fundamental object, the wave function and its properties are defined by a system of axioms (quantum axiomatics, or quantum principles). Some connection of the wave function with the irrotational flow of some quantum (Madelung) fluid[4–10] has been known for a long time. (The connection of the wave function with the irrotational flow was discovered comparatively recently.[11]) But always the wave function is considered to be a fundamental object, whereas the quantum fluid is considered as a derivative object. In this paper the fluid is considered as a fundamental object, connected directly with the statistical description of stochastic particles, whereas the wave function is considered to be a derivative construction whose properties can be expressed via the properties of the fluid.

In general, the wave function as a property of the fluid satisfies the quantum principles (linearity of dynamic equations, etc.) only in some special cases. For instance, in the case when the internal energy (2) depends only on $\mathbf{v}_{\text{dif}}=-\hbar(2m)^{-1}\nabla\rho/\rho$ and has the form

$$E=E(\rho,\nabla\rho)=\frac{\mathbf{v}_{\text{dif}}^2}{2},\quad \mathbf{v}_{\text{dif}}=-\frac{\hbar\nabla\rho}{2m\rho}, \tag{3}$$

where m is the particle mass and $\hbar$ is the Planck constant. To avoid misunderstanding and to distinguish between the wave function as a fundamental object, satisfying the quantum axiomatics, and the wave function as a property of a fluid, we shall use two different terms—''wave function'' in the first case and ''ψ function'' in the second one.

It is very important that the quantum phenomena be connected directly with the fluid model, i.e., such connection does not contain any reference to the quantum principles. There is hope that quantum superfluids like liquid helium may be described as an ideal fluid with the internal energy depending on $\nabla\rho$.

In the present paper some mathematical properties of conservative dynamic systems are investigated. Such a system $\mathcal{S}$ is a continuous set of particles, interacting via some self-consistent potential force field V. The dynamic system $\mathcal{S}$ is described by the action of the form

$$\mathcal{A}_L[\mathbf{x}]=\int\left\{\frac{m}{2}\left(\frac{d\mathbf{x}}{dt}\right)^2-V\right\}\rho_0(\boldsymbol{\xi})dt\,d\boldsymbol{\xi}, \tag{4}$$

where $\mathbf{x}=\{x^\alpha(t,\boldsymbol{\xi})\}$, $\alpha=1,2,3$ are functions of time t and of particle labels $\boldsymbol{\xi}=\{\xi_1,\xi_2,\xi_3\}$. $\rho_0(\boldsymbol{\xi})$ is some non-negative weight function, and $m=\text{const}$ is the mass of the fluid particle. V is a self-consistent potential depending on $\mathbf{x}$ and the derivatives of $\mathbf{x}$ with respect to $\boldsymbol{\xi}$. This function is supposed to have such a form that the potential V is a given function of variables ρ, $\nabla\rho$, and S. Here

$$\rho\equiv m\rho_0(\boldsymbol{\xi})\,\frac{\partial(\xi_1,\xi_2,\xi_3)}{\partial(x^1,x^2,x^3)}, \tag{5}$$

and $S=S_0(\boldsymbol{\xi})$ is some fixed function of variables $\boldsymbol{\xi}$. In this case the dynamic system $\mathcal{S}$ may be considered to be some ideal fluid. It will be shown that in the Euler description, where $x=\{t,\mathbf{x}\}$ are independent variables, and $\boldsymbol{\xi},\rho,\mathbf{v}\equiv d\mathbf{x}/dt$, S are dependent variables, the action (4) generates dynamic equations of the form

$$\frac{\partial\rho}{\partial t}+\nabla(\rho\mathbf{v})=0, \tag{6}$$

$$\frac{\partial v^\alpha}{\partial t}+(\mathbf{v}\nabla)v^\alpha=-\frac{1}{\rho}\,\partial_\beta P^{\alpha\beta},\quad \alpha=1,2,3, \tag{7}$$

$$\frac{\partial S}{\partial t}+(\mathbf{v}\nabla)S=0, \tag{8}$$

where $x^0=t$ is the time, $\mathbf{x}=\{x^1,x^2,x^3\}$ is the position vector, ρ and $\mathbf{v}=\{v^1,v^2,v^3\}$ are the fluid mass density and the fluid velocity, respectively, considered to be functions of $x=\{t,\mathbf{x}\}$. $P^{\alpha\beta}$ is a stress tensor, defined by (1), and $E(\rho,\nabla\rho,S)=V(\rho,\nabla\rho,S)/m$ is the internal energy of a unit mass. E depends on the density ρ, on the density gradient $\nabla\rho$, and on the entropy S per unit mass.

In the case of a usual fluid, when V does not depend on $\nabla\rho$, the stress tensor $P^{\alpha\beta}$ is isotropic, and Eq. (7) turns to the Euler equations for the ideal fluid

$$\frac{\partial \mathbf{v}}{\partial t}+(\mathbf{v}\boldsymbol{\nabla})\mathbf{v}=-\frac{1}{\rho}\boldsymbol{\nabla}p, \quad p=\rho^2\frac{\partial E}{\partial \rho}, \tag{9}$$

where p is the pressure, and $E=E(\rho,S)$ is the internal energy of a unit mass considered as a function of ρ and S. Thus, if V depends only on variables $\rho,\boldsymbol{\nabla}\rho,S$ the dynamic system $\mathcal{S}$, described by the action (4) will be referred to as nondissipative (ideal) fluid.

The system of hydrodynamic equations (6)–(8), as well as the system (6), (9), and (8) is a closed system of differential equations which has a unique solution inside some space–time region Ω, provided dependent dynamic variables ρ and $\mathbf{v}=\{v^1,v^2,v^3\}$, S are given as functions of three arguments on the space–time boundary Γ of the region Ω. Nevertheless, being closed, system (6)–(8) is incomplete, because it describes only momentum-energetic characteristics of the fluid. The action (4) generates additional dynamic equations,

$$\frac{\partial \boldsymbol{\xi}}{\partial t}+(\mathbf{v}\boldsymbol{\nabla})\boldsymbol{\xi}=0, \tag{10}$$

known as Lin constraints.[12] These equations describe the motion of fluid particles along their trajectories.

If Eq. (10) is solved and $\boldsymbol{\xi}$ is determined as a function of $(t,\mathbf{x})$, the finite relations

$$\boldsymbol{\xi}(t,\mathbf{x})=\boldsymbol{\xi}_{\rm in}={\rm const}$$

describe implicitly a fluid particle trajectory and a motion along it.

The system of eight equations (6)–(8), (10) forms a complete system of dynamic equations describing a fluid, whereas the system of five equations (6)–(8) forms a curtailed system of dynamic equations. The last system is closed, but to be a complete system, it must be supplemented by the kinematic equations

$$\frac{d\mathbf{x}}{dt}=\mathbf{v}(t,\mathbf{x}), \quad \mathbf{x}=\mathbf{x}(t,\boldsymbol{\xi}), \tag{11}$$

or by the Lin constraints (10) which are equivalent to (11).

The fact that the complete system (6)–(8), (10) of dynamic equations admits a closed subsystem (6)–(8) is connected with the invariance of the system (6)–(8), (10) with respect to the group of relabeling transformations (relabeling group),

$$\xi_\alpha\to\tilde{\xi}_\alpha=\tilde{\xi}_\alpha(\boldsymbol{\xi}), \quad D=\det\|\partial\tilde{\xi}_\alpha/\partial\xi_\beta\|\neq 0, \quad \alpha,\beta=1,2,3, \tag{12}$$

$$\varphi=\xi_0\to\tilde{\xi}_0=\tilde{\varphi}=\tilde{\xi}_0(\xi_0)+a_0(\boldsymbol{\xi}), \quad \partial\tilde{\xi}_0/\partial\xi_0>0, \tag{13}$$

where $\xi=\{\xi_0,\boldsymbol{\xi}\}$ are curvilinear Lagrangian coordinates in the space–time and $\tilde{\xi}=\{\tilde{\xi}_0,\tilde{\boldsymbol{\xi}}\}$ is another system of curvilinear Lagrangian coordinates. $\tilde{\boldsymbol{\xi}}$ and a_0 are arbitrary functions of $\boldsymbol{\xi}$. $\tilde{\xi}_0$ is an arbitrary function of ξ_0. ξ_0 is a temporal Lagrangian coordinate, and $\boldsymbol{\xi}$ are spatial ones.

The relabeling group properties have been used in hydrodynamics comparatively recently.[13–20] The action (4) is invariant with respect to the relabeling group (12), (13), provided the weight function $\rho_0(\boldsymbol{\xi})$ transforms as a scalar density,

$$\rho_0(\boldsymbol{\xi})\to\tilde{\rho}_0(\tilde{\boldsymbol{\xi}})=D^{-1}\rho_0(\boldsymbol{\xi}), \quad D=\frac{\partial(\tilde{\boldsymbol{\xi}})}{\partial(\boldsymbol{\xi})}\equiv\frac{\partial(\tilde{\xi}_1,\tilde{\xi}_2,\tilde{\xi}_3)}{\partial(\xi_1,\xi_2,\xi_3)}. \tag{14}$$

The group of relabeling transformations appears to be a symmetry group of the dynamic system (fluid). Any special particle labeling is unessential from a physical viewpoint. It is a reason why several equations (6)–(8) of the complete system form a closed system describing conservation laws. This symmetry group also permits one to integrate the complete system (6)–(8), (10) in the form (see the proof below Sec. III)

$$S(t,\mathbf{x})=S_0(\boldsymbol{\xi}), \tag{15}$$

$$\rho(t,\mathbf{x})=\rho_0(\boldsymbol{\xi})\,\frac{\partial(\xi_1,\xi_2,\xi_3)}{\partial(x^1,x^2,x^3)}\equiv\rho_0(\boldsymbol{\xi})\,\frac{\partial(\boldsymbol{\xi})}{\partial(\mathbf{x})}, \tag{16}$$

$$\mathbf{v}(t,\mathbf{x})=\boldsymbol{\pi}(\varphi,\boldsymbol{\xi},\eta,S)\equiv\boldsymbol{\nabla}\varphi+g^{\alpha}(\boldsymbol{\xi})\boldsymbol{\nabla}\xi_\alpha-\eta\boldsymbol{\nabla}S, \tag{17}$$

where $S_0(\boldsymbol{\xi})$, $\rho_0(\boldsymbol{\xi})$, $\mathbf{g}(\boldsymbol{\xi})=\{g^{\alpha}(\boldsymbol{\xi})\}$, $\alpha=1,2,3$ are arbitrary integration functions of argument $\boldsymbol{\xi}$, and φ, η are new dependent variables, satisfying the dynamic equations

$$\frac{\partial\varphi}{\partial t}+\boldsymbol{\pi}(\varphi,\boldsymbol{\xi},\eta,S)\boldsymbol{\nabla}\varphi-\frac{1}{2}\,[\boldsymbol{\pi}(\varphi,\boldsymbol{\xi},\eta,S)]^2+\frac{\partial(\rho E)}{\partial\rho}-\partial_\alpha\,\frac{\partial(\rho E)}{\partial\rho_\alpha}=0, \tag{18}$$

$$\frac{\partial\eta}{\partial t}+\boldsymbol{\pi}(\varphi,\boldsymbol{\xi},\eta,S)\boldsymbol{\nabla}\eta=-\frac{\partial E}{\partial S}. \tag{19}$$

If five dependent variables $\varphi,\boldsymbol{\xi},\eta$ satisfy the system of equations (10), (18), (19), the five dynamic variables $S,\rho,\mathbf{v}$ (15)–(17) satisfy dynamic equations (6)–(8). Indefinite functions $S_0(\boldsymbol{\xi})$, $\rho_0(\boldsymbol{\xi})$, $\mathbf{g}(\boldsymbol{\xi})$ can be determined from initial and boundary conditions in such a way that the initial and boundary conditions for the variables $\varphi,\boldsymbol{\xi},\eta$ are universal in the sense that they do not depend on the fluid flow.

The integration of the complete system (6)–(8), (10) and some corollaries of this integration correlate with the Hamilton properties of the ideal fluid.[14,19,20,23,24,27] It is connected with the fact that the curtailed system (6)–(8) is not a Hamiltonian system in itself, whereas the complete system (6)–(8), (10) is a Hamiltonian one. Constructing Hamiltonian mechanics of the ideal fluid, one uses (implicitly or explicitly) the Lin constraints (or part of them). It is this expansion of the curtailed system (but not Hamiltonian properties) that is important for integration and derivation of other useful results. To show this, the Hamiltonian technique and Hamiltonian properties of the ideal fluid will not be used at all.

According to (16) and (17) the physical quantities $\rho,\mathbf{v}$ are obtained as a result of differentiation of the variables $\varphi,\boldsymbol{\xi},S$, and the variables $\varphi,\boldsymbol{\xi},\eta$ can be regarded as hydrodynamic potentials. These potentials appear in the Hamilton fluid dynamics[23] as dependent variables. They are associated with Clebsch,[21,22] who introduced these quantities for the incompressible fluid. Such quantities as $g^{\alpha}(\boldsymbol{\xi})$ also appear in the Hamilton fluid mechanics,[23] but they appear as dependent variables (Lagrange invariants) satisfying dynamic equations of the type (10). They also are regarded as hydrodynamic potentials. Note that in the Hamilton fluid mechanics[23] the quantities g^{α} are considered simply as dependent variables, but not as indefinite functions of $\boldsymbol{\xi}$ arising as a result of integration, although corresponding dynamic equations for g^{α} can be integrated easily.

Integration of the dynamic equations admits a description of any ideal fluid in terms of hydrodynamic potentials $\xi=\{\xi_0,\boldsymbol{\xi}\}$. The hydrodynamic potentials ξ are Lagrangian coordinates considered as functions of independent Eulerian coordinates $x=\{t,\mathbf{x}\}$. Spatial Lagrangian coordinates $\boldsymbol{\xi}=\{\xi_\alpha\}$, $\alpha=1,2,3$ label fluid particles, whereas the temporal Lagrangian coordinate $\xi_0=\xi_0(t,\mathbf{x})$ means some generalized time for the fluid particle placed at the space–time point $x=\{t,\mathbf{x}\}$.

The description of any ideal fluid in terms of hydrodynamic potentials ξ can transform into a description in terms of a complex n-component hydrodynamic potential $\psi=\{\psi_\alpha\}$, $\alpha=1,2,...,n$ which associates with the wave function, used in the quantum mechanics, whereas the irreducible (minimally possible) number n_m of ψ-function components associates with the spin of the flow (not of the particle).

In the presented paper it is shown that the wave function is a way of describing any ideal fluid. The spin is a natural property of any flow of the ideal fluid. The appearance of these enigmatic quantities at the description of quantum particles may be explained merely as a result of a quantum particle description in terms of an ideal fluid (statistical ensemble). Note that the curtailed system (6)–(8) has the same order as the integrated system (10), (18), (19), but takes into account neither initial conditions, nor kinematic equations (11). The fact that the ideal fluid considered as a dynamic system admits both the curtailed system (6)–(8) and the integrated system (10), (18), (19) is connected closely with the group of the relabeling transformation (12).

Section II is devoted to a presentation of the space–time symmetric Jacobian technique which is needed for the integration of hydrodynamic equations. Use of Jacobians in hydrodynamics has had a long history, dating back to the time of Clebsch.[21,22] It was the use of Jacobians that allowed us to introduce the Clebsch potentials and integrate hydrodynamic equations. The Jacobian technique has been used in Refs. 14, 20, 23, 24, 25, and many other papers. It seems that the progress in the integration of hydrodynamic equations is connected mainly with the developed Jacobian technique.

Further it will be proved (Sec. III) that the complete system of hydrodynamic equations (6)–(8), (10) can be integrated in the form (10), (15)–(19) that leads to a special form of a description in terms of hydrodynamic potentials (DTHP). In Sec. IV the initial and boundary conditions are used for a determination of function **g**. In Sec. V a special type of complex hydrodynamic potentials is considered and the fluid flows are classified on the irreducible number of wave function components which appears to be an invariant of the relabeling group.

II. JACOBIAN TECHNIQUE

Let us consider such a space–time symmetric mathematical object as the Jacobian

$$J \equiv \frac{\partial(\xi_0,\xi_1,\xi_2,\xi_3)}{\partial(x^0,x^1,x^2,x^3)} \equiv \det\|\xi_{i,k}\|, \quad \xi_{i,k} \equiv \partial_k \xi_i \equiv \frac{\partial \xi_i}{\partial x^k}, \quad i,k=0,1,2,3. \tag{20}$$

Here $\xi=\{\xi_0,\boldsymbol{\xi}\}=\{\xi_0,\xi_1,\xi_2,\xi_3\}$ are four scalars considered as functions $\xi=\xi(x)$ of $x=\{x^0,\mathbf{x}\}$. The functions $\{\xi_0,\xi_1,\xi_2,\xi_3\}$ are supposed to be independent in the sense that $J\neq 0$. It is useful to consider the Jacobian J as a four-linear function of variables $\xi_{i,k}\equiv\partial_k\xi_i$, $i,k=0,1,2,3$. Then one can introduce derivatives of J with respect to $\xi_{i,k}$. The derivative $\partial J/\partial \xi_{i,k}$ appears as a result of the substitution of ξ_i by x^k in relation (20),

$$\frac{\partial J}{\partial \xi_{i,k}} \equiv \frac{\partial(\xi_0,\dots,\xi_{i-1},x^k,\xi_{i+1},\dots,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}, \quad i,k=0,1,2,3. \tag{21}$$

For instance,

$$\frac{\partial J}{\partial \xi_{0,i}} \equiv \frac{\partial(x^i,\xi_1,\xi_2,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}, \quad i=0,1,2,3. \tag{22}$$

This rule is also valid for higher derivatives of J,

$$\begin{aligned}\frac{\partial^2 J}{\partial \xi_{i,k}\partial \xi_{s,l}} &\equiv \frac{\partial(\xi_0,\dots,\xi_{i-1},x^k,\xi_{i+1},\dots,\xi_{s-1},x^l,\xi_{s+1},\dots,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}\\ &\equiv \frac{\partial(x^k,x^l)}{\partial(\xi_i,\xi_s)}\,\frac{\partial(\xi_0,\xi_1,\xi_2,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}\\ &\equiv J\left(\frac{\partial x^k}{\partial \xi_i}\frac{\partial x^l}{\partial \xi_s}-\frac{\partial x^k}{\partial \xi_s}\frac{\partial x^l}{\partial \xi_i}\right), \quad i,k,l,s=0,1,2,3.\end{aligned} \tag{23}$$

It follows from (20) and (21) that

$$\frac{\partial x^k}{\partial \xi_i} \equiv \frac{\partial(\xi_0,\ldots,\xi_{i-1},x^k,\xi_{i+1},\ldots,\xi_3)}{\partial(\xi_0,\xi_1,\xi_2,\xi_3)}$$
$$\equiv \frac{\partial(\xi_0,\ldots,\xi_{i-1},x^k,\xi_{i+1},\ldots,\xi_3)}{\partial(x^0,x^1,x^2,x^3)} \frac{\partial(x^0,x^1,x^2,x^3)}{\partial(\xi_0,\xi_1,\xi_2,\xi_3)}$$
$$\equiv \frac{1}{J}\frac{\partial J}{\partial \xi_{i,k}}, \quad i,k=0,1,2,3 \tag{24}$$

and (23) may be written in the form

$$\frac{\partial^2 J}{\partial \xi_{i,k}\partial \xi_{s,l}} \equiv \frac{1}{J}\left(\frac{\partial J}{\partial \xi_{i,k}}\frac{\partial J}{\partial \xi_{s,l}} - \frac{\partial J}{\partial \xi_{i,l}}\frac{\partial J}{\partial \xi_{s,k}}\right), \quad i,k,l,s=0,1,2,3. \tag{25}$$

The derivative $\partial J/\partial \xi_{i,k}$ is a cofactor to the element $\xi_{i,k}$ of the determinant (20). Then one has the following identities:

$$\xi_{l,k}\frac{\partial J}{\partial \xi_{s,k}} \equiv \delta_l^s J, \quad \xi_{k,l}\frac{\partial J}{\partial \xi_{k,s}} \equiv \delta_l^s J, \quad l,s=0,1,2,3, \tag{26}$$

$$\partial_k \frac{\partial J}{\partial \xi_{i,k}} \equiv \frac{\partial^2 J}{\partial \xi_{i,k}\partial \xi_{s,l}}\partial_k\partial_l \xi_s \equiv 0, \quad i=0,1,2,3. \tag{27}$$

Here and subsequently a summation on two repeated indices is produced (0–3) for Latin indices and (1–3) for the Greek ones. The identity (27) can be considered as a corollary of the identity (25) and a symmetry of $\partial_k\partial_l\xi_s$ with respect to the permutation of indices k,l. The convolution of (25) with ∂_k, or ∂_l also vanishes.

Relations (20)–(25) are written for four independent variables x, but they are valid in an evident way for an arbitrary number $n+1$ of variables $x=\{x^0,x^1,\ldots,x^n\}$ and $\boldsymbol{\xi}=\{\xi_0,\boldsymbol{\xi}\}$, $\boldsymbol{\xi}=\{\xi_1,\xi_2,\ldots,\xi_n\}$.

Application of the Jacobian J to hydrodynamics is founded on the property that the fluid flux

$$j^i = m\frac{\partial J}{\partial \xi_{0,i}}, \quad j=\{j^i\}=\{\rho,\rho\mathbf{v}\}, \quad i=0,1,2,3 \tag{28}$$

constructed on the basis of the variables $\boldsymbol{\xi}=\{\xi_1,\xi_2,\xi_3\}$ satisfies Lin constraints (10) and the continuity equation

$$\partial_i j^i = 0 \tag{29}$$

identically for any choice of variables $\boldsymbol{\xi}$, as it follows from the identity (27) for $i=0$. The continuity equation (29) is used without approximations in all hydrodynamic models, and the change of variables $\{\rho,\rho\mathbf{v}\}\leftrightarrow\boldsymbol{\xi}$ described by (28) is very important.

In particular, in the case of two-dimensional established flow of incompressible fluid the variables $\boldsymbol{\xi}$ reduce to one variable $\xi_1=\psi$, known as the stream function. In this case there are only two essential dependent variables $x^0=x$, $x^1=y$, and the relations (28), (29) reduce to the relations

$$\rho^{-1}j_x = v_x = \frac{\partial\psi}{\partial y}, \quad \rho^{-1}j_y = v_y = -\frac{\partial\psi}{\partial x}, \quad \frac{\partial v_x}{\partial x}+\frac{\partial v_y}{\partial y}=0. \tag{30}$$

Defining the stream line as a line tangent to the flux j,

$$\frac{dx}{j_x} = \frac{dy}{j_y}, \tag{31}$$

one obtains that the stream function is constant along the stream line, because according to the first two equations of (30), $\psi=\psi(x,y)$ is an integral of Eq. (31).

In the general case, when the space dimensionality is n and $x=\{x^0,x^1,\dots,x^n\}$, $\boldsymbol{\xi}=\{\xi_0,\boldsymbol{\xi}\}$, $\boldsymbol{\xi}=\{\xi_1,\xi_2,\dots,\xi_n\}$, the quantities $\boldsymbol{\xi}=\{\xi_\alpha\}$, $\alpha=1,2,...,n$ are constant along the line $\mathcal{L}$ tangent to the flux vector $j=\{j^i\}$, $i=0,1,...,n$,

$$\mathcal{L}: \quad \frac{dx^i}{d\tau}=j^i(x), \quad i=0,1,...,n, \tag{32}$$

where τ is a parameter along the line $\mathcal{L}$ which is described parametrically by the equation $x=x(\tau)$. This statement is formulated mathematically in the form

$$\frac{d\xi_\alpha}{d\tau}=j^i\partial_i\xi_\alpha=m\,\frac{\partial J}{\partial \xi_{0,i}}\,\partial_i\xi_\alpha=0, \quad \alpha=1,2,...,n.$$

The last equality follows from the first identity (26) taken for $s=0$, $l=1,2,...,n$

Interpretation of the line (32) tangent to the flux is different for different cases. If $x=\{x^0,x^1,\dots,x^n\}$ contains only spatial coordinates, the line (32) is a line in the usual space. It is regarded as a streamline, and $\boldsymbol{\xi}$ can be interpreted as quantities which are constant along the streamline (i.e., as a generalized stream function). If x^0 is the time coordinate, Eq. (32) describes a line in the space–time. This line (known as a world line of a fluid particle) determines a motion of the fluid particle. Variables $\boldsymbol{\xi}=\{\xi_1,\xi_2,\dots,\xi_n\}$ which are constant along the world line are different, generally, for different particles. If ξ_α, $\alpha=1,2,...,n$ are independent, they may be used for the fluid particle labeling.

Thus, although interpretation of the relation (28) considered as a replacement of dependent variables j by $\boldsymbol{\xi}$ may be different, from a mathematical viewpoint this transformation means a replacement of the continuity equation by some equations for the labeling (or generalized stream function) $\boldsymbol{\xi}$. The difference of the interpretation is of no importance in this context.

Note that the expressions

$$j^i=m\rho_0(\boldsymbol{\xi})\,\frac{\partial J}{\partial \xi_{0,i}}\equiv m\rho_0(\boldsymbol{\xi})\,\frac{\partial(x^i,\xi_1,\xi_2,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}, \quad i=0,1,2,3, \tag{33}$$

can be also considered as four-flux satisfying the continuity equation (29). Here m is a constant and $\rho_0(\boldsymbol{\xi})$ is an arbitrary function of $\boldsymbol{\xi}$. It follows from the identity

$$m\rho_0(\boldsymbol{\xi})\,\frac{\partial(x^i,\xi_1,\xi_2,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}\equiv m\,\frac{\partial(x^i,\tilde{\xi}_1,\xi_2,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}, \quad \tilde{\xi}_1=\int_0^{\xi_1}\rho_0(\xi_1',\xi_2,\xi_3)d\xi_1'.$$

As an example of the application of the Jacobian technique, let us show that (5) satisfies (6) by virtue of (10). Let us multiply (10) by (5) and introduce new variables $\mathbf{j}=\rho\mathbf{v}=\{j^1,j^2,j^3\}$. One obtains three equations

$$m\rho_0(\boldsymbol{\xi})\,\frac{\partial J}{\partial \xi_{0,0}}\,\xi_{\beta,0}+j^\alpha\xi_{\beta,\alpha}=0, \quad \beta=1,2,3. \tag{34}$$

Considering (34) as a system of three linear equations for j^α, $\alpha=1,2,3$ and resolving it with respect to j^α, one obtains

$$j^\alpha=m\rho_0(\boldsymbol{\xi})\,\frac{\partial J}{\partial \xi_{0,\alpha}}, \quad \alpha=1,2,3. \tag{35}$$

It is easy to verify this, substituting (35) into (34) and using (26). One obtains that $j=\{j^0,\mathbf{j}\}=\{\rho,\rho\mathbf{v}\}$ is described by the relations (33) which satisfy the continuity equation (29) identically. Thus, (6) is satisfied by (16) by virtue of (10).

III. VARIATIONAL PRINCIPLE

In general, the equivalency of the system (10), (18), (19) and the system (6)–(8), (10) can be verified by a direct substitution of the variables $\rho,S,\mathbf{v}$, defined by relations (15)–(17) into Eqs. (6)–(8). Using Eqs. (10), (18), and (19), one obtains identities after subsequent calculations. But such computations do not display a connection between the integration and the invariance with respect to the relabeling group (12). Besides, the meaning of the new variables φ,η is not clear. We shall use for our investigations a variational principle. Note that for a long time a derivation of a variational principle for hydrodynamic equations (6)–(8) existed as a self-dependent problem.[12,14,16,23,24,26,27] The existence of this problem was connected to a lack of understanding that the system of hydrodynamic equations (6)–(8) is a curtailed system, and the full system of dynamic equations (6)–(8), (10) includes equations (10) describing a motion of the fluid particles in the given velocity field. The variational principle can generate only the complete system of dynamic variables (but not its closed subsystem). Without understanding this, one tried to form the Lagrangian for the system (6)–(8) as a sum of some quantities taken with Lagrange multipliers. The left-hand side of dynamic equations (6)–(8) and some other constraints were taken as such quantities.

Now this problem has been solved (see the review by Salmon[23]) on the basis of the Eulerian version of the variational principle for the Lagrangian description (4), where equations (10) appear automatically and cannot be ignored. In our version of the variational principle we follow Ref. 23 with some modifications which underline a curtailed character of hydrodynamic equations (6)–(8), because an understanding of the curtailed character of the system (6)–(8) removes the problem of derivation of the variational principle for the hydrodynamic equations (6)–(8).

We consider the ideal fluid as a conservative dynamic system whose dynamic equations can be derived from the variational principle. This dynamic system is a continuous set of many identical particles moving in some self-consistent potential force field. The action functional has the form (4). Variation of the action with respect to $\mathbf{x}$ generates six first-order dynamic equations for six dependent variables $\mathbf{x},\mathbf{v}=d\mathbf{x}/dt$, considered as functions of t and of independent curvilinear Lagrangian coordinates $\boldsymbol{\xi}$. It is a Lagrangian representation of hydrodynamic equations.

We prefer to work with Eulerian representation, when Lagrangian coordinates (particle labeling) $\xi=\{\xi_0,\boldsymbol{\xi}\}$, $\boldsymbol{\xi}=\{\xi_1,\xi_2,\xi_3\}$ are considered as dependent variables, and Eulerian coordinates $x=\{x^0,\mathbf{x}\}=\{t,\mathbf{x}\}$, $\mathbf{x}=\{x^1,x^2,x^3\}$ are considered as independent variables. Here ξ_0 is a temporal Lagrangian coordinate which evolves along the particle trajectory in an arbitrary way. Now the ξ_0 is a fictitious variable, but after the integration of equations the ξ_0 stops being fictitious and turns to the variable φ, appearing in the integrated system (10), (18), (19).

Further, space–time symmetric designations will mainly be used, which simplifies considerably all computations. In the Eulerian description the action functional (4) is to be represented as an integral over independent variables $x=\{x^0,\mathbf{x}\}=\{t,\mathbf{x}\}$. One uses the Jacobian technique for such a transformation of the action (4),

Let us note that according to (22) the derivative $d\mathbf{x}/dt$ can be written in the form

$$v^\alpha=\frac{dx^\alpha}{dt}\equiv\frac{\partial J}{\partial\xi_{0,\alpha}}\left(\frac{\partial J}{\partial\xi_{0,0}}\right)^{-1},\quad \alpha=1,2,3.$$

Then components of the four flux $j=\{j^0,\mathbf{j}\}\equiv\{\rho,\rho\mathbf{v}\}$ can be written in the form (33), provided the designation (5)

$$j^0=\rho=m\rho_0(\boldsymbol{\xi})\,\frac{\partial J}{\partial\xi_{0,0}}\equiv m\rho_0(\boldsymbol{\xi})\,\frac{\partial(x^0,\xi_1,\xi_2,\xi_3)}{\partial(x^0,x^1,x^2,x^3)}\tag{36}$$

is used.

At such a form of the mass density ρ the four flux $j=\{j^i\}$, $i=0,1,2,3$ satisfies identically the continuity equation (29) which takes place by virtue of identities (26) and (27). Besides, by virtue of identities (26) and (27) the Lin constraints (10) are fulfilled identically,

$$j^i \partial_i \xi_\alpha = 0, \quad \alpha = 1,2,3. \tag{37}$$

The components j^i are invariant with respect to the relabeling group (12), provided the function $\rho^0(\boldsymbol{\xi})$ transforms according to (14).

One has

$$\rho_0(\boldsymbol{\xi}) dt\ d\boldsymbol{\xi} = \rho_0(\boldsymbol{\xi}) \frac{\partial J}{\partial \xi_{0,0}} dt\ d\mathbf{x} = \frac{\rho}{m} dt\ d\mathbf{x},$$

$$\frac{m}{2}\left(\frac{dx^\alpha}{dt}\right)^2 = \frac{m}{2}\left(\frac{\partial J}{\partial \xi_{0,\alpha}}\right)^2 \left(\frac{\partial J}{\partial \xi_{0,0}}\right)^{-2} = \frac{m}{2}\left(\frac{j^\alpha}{\rho}\right)^2,$$

and the variational problem with the action functional (4) is written as a variational problem with the action functional

$$\mathcal{A}_E[\boldsymbol{\xi}] = \int \left(\frac{\mathbf{j}^2}{2\rho} - \rho E\right) dt\ d\mathbf{x}, \quad E = \frac{V}{m}, \tag{38}$$

where $\rho = j^0$ and $\mathbf{j} = \{j^1, j^2, j^3\}$ are fixed functions of $\xi = \{\xi_0, \boldsymbol{\xi}\}$ and of $\xi_{\alpha,i} \equiv \partial_i \xi_\alpha$, $\alpha = 1,2,3$, $i = 0,1,2,3$, defined by relations (33). E is the internal energy of the fluid which is supposed to be a fixed function of $\rho, \boldsymbol{\nabla}\rho, S_0(\boldsymbol{\xi})$,

$$E = E(\rho, \boldsymbol{\nabla}\rho, S_0(\boldsymbol{\xi})), \tag{39}$$

where ρ is defined by (36) and $S_0(\boldsymbol{\xi})$ is some fixed function of $\boldsymbol{\xi}$, describing the initial distribution of the entropy over the fluid.

The action (38) is invariant with respect to the subgroup $\mathcal{G}_{S_0}$ of the relabeling group (12). The subgroup $\mathcal{G}_{S_0}$ is determined in such a way that any surface $S_0(\boldsymbol{\xi}) = \text{const}$ is invariant with respect to $\mathcal{G}_{S_0}$. In general, the subgroup $\mathcal{G}_{S_0}$ is determined by two arbitrary functions of $\boldsymbol{\xi}$.

The action (38) generates the six order system of dynamic equations, consisting of three second-order equations for three dependent variables $\boldsymbol{\xi}$. Invariance of the action (38) with respect to the subgroup $\mathcal{G}_{S_0}$ admits one to integrate the system of dynamic equations. The order of the system reduces, and two arbitrary integration functions appear. The order of the system reduces to five (but not to four), because the fictitious dependent variable ξ_0 stops being fictitious as a result of the integration.

Unfortunately, the subgroup $\mathcal{G}_{S_0}$ depends on the form of the function $S_0(\boldsymbol{\xi})$ and cannot be obtained in a general form. In the special case, when $S_0(\boldsymbol{\xi})$ does not depend on $\boldsymbol{\xi}$, the subgroup $\mathcal{G}_{S_0}$ coincides with the whole relabeling group $\mathcal{G}$, and the order of the integrated system reduces to four.

In the general case it is convenient to introduce a new dependent variable,

$$S = S_0(\boldsymbol{\xi}).$$

According to (37) the variable S satisfies the dynamic equation (8),

$$j^i \partial_i S = 0. \tag{40}$$

By virtue of designations (28) and identities (26) and (27), Eqs. (40) and (37) are fulfilled, identically. Hence, they can be added to the action functional (38) as side constraints without a change of the variational problem. Adding (40) to the Lagrangian of the action (38) by means of a Lagrange multiplier η, one obtains

$$\mathcal{A}_E[\boldsymbol{\xi},\eta,S]=\int\left\{\frac{\mathbf{j}^2}{2\rho}-\rho E+\eta j^k\partial_k S\right\}dt\ d\mathbf{x}, \tag{41}$$

where the quantities $j=\{\rho,\mathbf{j}\}$ are determined by (33), and $E=E(\rho,\boldsymbol{\nabla}\rho,S)$. The action (41) is invariant with respect to the relabeling group $\mathcal{G}$ which is determined by three arbitrary functions of $\boldsymbol{\xi}$.

To obtain the dynamic equations, it is convenient to introduce new dependent variables j^i, defined by (33). Let us introduce the new variables j^i by means of designations (33) taken with the Lagrange multipliers p_i, $i=0,1,2,3$. Then the action (41) takes the form

$$\mathcal{A}_E[\rho,\mathbf{j},\boldsymbol{\xi},p,\eta,S]=\int\left\{\frac{\mathbf{j}^2}{2\rho}-\rho E-p_k\left[j^k-m\rho_0(\boldsymbol{\xi})\ \frac{\partial J}{\partial\xi_{0,k}}\right]+\eta j^k\partial_k S\right\}dt\ d\mathbf{x}. \tag{42}$$

It is useful to keep in mind that four designations (33), introducing variables ρ, $\mathbf{j}=\rho\mathbf{v}$ via variables $\boldsymbol{\xi}$, are equivalent to three Lin constraints (10) together with the designation (36), as was shown at the end of Sec. II. The addition of relations (33) to the action (41) as side constraints is equivalent to the addition of relations (10) and (36) considered as side constraints.

For obtaining dynamic equations, the variables $\rho,\mathbf{j},\boldsymbol{\xi},p,\eta,S$ are to be varied. Let us eliminate the variables p_i from the action (42). Dynamic equations arising as a result of a variation with respect to ξ_α have the form

$$\frac{\delta\mathcal{A}_E}{\delta\xi_\alpha}\equiv\hat{\mathcal{L}}_\alpha p=-m\partial_k\left[\rho_0(\boldsymbol{\xi})\ \frac{\partial^2 J}{\partial\xi_{0,i}\partial\xi_{\alpha,k}}\ p_i\right]+m\ \frac{\partial\rho_0(\boldsymbol{\xi})}{\partial\xi_\alpha}\ \frac{\partial J}{\partial\xi_{0,k}}\ p_k=0,\quad \alpha=1,2,3, \tag{43}$$

where $\hat{\mathcal{L}}_\alpha$ are linear operators acting on variables $p=\{p_i\}$, $i=0,1,2,3$. These equations can be integrated in the form

$$p_i=bg^0(\xi_0)\partial_i\xi_0+bg^\alpha(\boldsymbol{\xi})\partial_i\xi_\alpha,\quad i=0,1,2,3, \tag{44}$$

where b is an arbitrary scale constant, ξ_0 is some new variable (temporal Lagrangian coordinate), $g^\alpha(\boldsymbol{\xi})$, $\alpha=1,2,3$ are arbitrary functions of the labels $\boldsymbol{\xi}$, $g^0(\xi_0)$ is an arbitrary function of ξ_0. The relations (44) satisfy equations (43) identically. Indeed, substituting (44) into (43) and using identities (25) and (26), one obtains

$$-m\partial_k\left\{\rho_0(\boldsymbol{\xi})\left[\frac{\partial J}{\partial\xi_{\alpha,k}}\ g^0(\xi_0)-\frac{\partial J}{\partial\xi_{0,k}}\ g^\alpha(\boldsymbol{\xi})\right]\right\}+m\ \frac{\partial\rho_0(\boldsymbol{\xi})}{\partial\xi_\alpha}\ Jg^0(\xi_0)=0,\quad \alpha=1,2,3. \tag{45}$$

Differentiating braces and using identities (27) and (26), one concludes that (45) is an identity.

Setting for simplicity

$$\partial_k\varphi=g^0(\xi_0)\partial_k\xi_0,\quad k=0,1,2,3$$

one obtains

$$p_k=b\partial_k\varphi+bg^\alpha(\boldsymbol{\xi})\partial_k\xi_\alpha,\quad k=0,1,2,3. \tag{46}$$

Substituting (46) in (42), one can eliminate variables p_i, $i=0,1,2,3$ from the functional (42). The term $g^\alpha(\boldsymbol{\xi})\partial_k\xi_\alpha\partial J/\partial\xi_{0,k}$ vanishes, the term $\partial_k\varphi\partial J/\partial\xi_{0,k}$ gives no contribution to the dynamic equations. The action functional takes the form

$$\mathcal{A}_g[\rho,\mathbf{j},\varphi,\boldsymbol{\xi},\eta,S]=\int\left\{\frac{\mathbf{j}^2}{2\rho}-\rho E-j^k[b\partial_k\varphi+bg^\alpha(\boldsymbol{\xi})\partial_k\xi_\alpha-\eta\partial_k S]\right\}dt\ d\mathbf{x},\tag{47}$$

where $g^\alpha(\boldsymbol{\xi})$ are considered as fixed functions of $\boldsymbol{\xi}$ which are determined from initial conditions. The action (47) is a functional of indefinite fixed functions $\mathbf{g}(\mathbf{x})$. Varying the action (47) with respect to $\varphi,\boldsymbol{\xi},\eta,S,\mathbf{j},\rho$, one obtains dynamic equations;

$$\delta\varphi:\quad \partial_k j^k=0,\tag{48}$$

$$\delta\xi_\alpha:\quad \Omega^{\alpha\beta}j^k\partial_k\xi_\beta=0,\quad \alpha=1,2,3,\tag{49}$$

$$\delta\eta:\quad j^k\partial_k S=0,\tag{50}$$

$$\delta S:\quad j^k\partial_k\eta=-\rho\frac{\partial E}{\partial S},\tag{51}$$

$$\delta\mathbf{j}:\quad \mathbf{v}\equiv\mathbf{j}/\rho=b\boldsymbol{\nabla}\varphi+bg^\alpha(\boldsymbol{\xi})\boldsymbol{\nabla}\xi_\alpha-\eta\boldsymbol{\nabla}S,\tag{52}$$

$$\delta\rho:\quad -\frac{\mathbf{j}^2}{2\rho^2}-\frac{\partial(\rho E)}{\partial\rho}+\partial_\alpha\frac{\partial(\rho E)}{\partial\rho_\alpha}-b\partial_0\varphi-bg^\alpha(\boldsymbol{\xi})\partial_0\xi_\alpha+\eta\partial_0 S=0.\tag{53}$$

Here $\Omega^{\alpha\beta}$ is defined by

$$\Omega^{\alpha\beta}=b\left(\frac{\partial g^\alpha(\boldsymbol{\xi})}{\partial\xi_\beta}-\frac{\partial g^\beta(\boldsymbol{\xi})}{\partial\xi_\alpha}\right),\quad \alpha,\beta=1,2,3.\tag{54}$$

Deriving relations (49) and (51), the continuity equation (48) was used. It is easy to see that (49) is equivalent to (10), provided

$$\det\|\Omega^{\alpha\beta}\|\neq 0.\tag{55}$$

Then Eqs. (50) and (48) can be integrated in the form of (15) and (16) respectively. Equations (51) and (52) are equivalent to (19) and (17). Finally, eliminating $\partial_0\xi_\alpha$ and $\partial_0 S$ from (53) by means of (49) and (50), one obtains Eq. (18) and, hence, the system of dynamic equations (10), (18), and (19), where designations (15)–(17) are used.

The curtailed system (6)–(8) can be obtained from Eqs. (48) to (53) as follows. Equations (48) and (50) coincide with (6) and (8). For deriving (7) let us note that the vorticity $\boldsymbol{\omega}\equiv\boldsymbol{\nabla}\times\mathbf{v}$ and $\mathbf{v}\times\boldsymbol{\omega}$ are obtained from (52) in the form

$$\boldsymbol{\omega}=\boldsymbol{\nabla}\times\mathbf{v}=\tfrac{1}{2}\Omega^{\alpha\beta}\boldsymbol{\nabla}\xi_\beta\times\boldsymbol{\nabla}\xi_\alpha-\boldsymbol{\nabla}\eta\times\boldsymbol{\nabla}S,\tag{56}$$

$$\mathbf{v}\times\boldsymbol{\omega}=\Omega^{\alpha\beta}\boldsymbol{\nabla}\xi_\beta(\mathbf{v}\boldsymbol{\nabla})\xi_\alpha+\boldsymbol{\nabla}S(\mathbf{v}\boldsymbol{\nabla})\eta-\boldsymbol{\nabla}\eta(\mathbf{v}\boldsymbol{\nabla})S.\tag{57}$$

Let us form a difference between the time derivative of (52) and the gradient of (53). Eliminating $\Omega^{\alpha\beta}\partial_0\xi_\alpha$, $\partial_0 S$, and $\partial_0\eta$ by means of Eqs. (49), (50), (51), one obtains

$$\begin{aligned}\partial_0\mathbf{v}+\boldsymbol{\nabla}\frac{\mathbf{v}^2}{2}+\frac{\partial^2(\rho E)}{\partial\rho^2}\boldsymbol{\nabla}\rho+\frac{\partial^2(\rho E)}{\partial\rho\partial S}\boldsymbol{\nabla}S+\boldsymbol{\nabla}\rho_\beta\frac{\partial^2(\rho E)}{\partial\rho_\beta\partial\rho}-\boldsymbol{\nabla}\partial_\beta\frac{\partial^2(\rho E)}{\partial\rho_\beta}-\frac{\partial E}{\partial S}\boldsymbol{\nabla}S\\ -\Omega^{\alpha\beta}\boldsymbol{\nabla}\xi_\beta(\mathbf{v}\boldsymbol{\nabla})\xi_\alpha+\boldsymbol{\nabla}\eta(\mathbf{v}\nabla)S-\boldsymbol{\nabla}S(\mathbf{v}\boldsymbol{\nabla})\eta=0.\end{aligned}\tag{58}$$

Using (56) and (57) expression (58) reduces to

$$\partial_0\mathbf{v}+\boldsymbol{\nabla}\frac{\mathbf{v}^2}{2}-\mathbf{v}\times(\boldsymbol{\nabla}\times\mathbf{v})+\frac{1}{\rho}\boldsymbol{\nabla}\left(\rho^2\frac{\partial E}{\partial\rho}\right)-\frac{1}{\rho}\partial_\beta\left[\rho\boldsymbol{\nabla}\frac{\partial^2(\rho E)}{\partial\rho_\beta}\right]=0.\tag{59}$$

By virtue of the identity

$$\mathbf{v}\times(\nabla\times\mathbf{v})\equiv\nabla\,\frac{\mathbf{v}^2}{2}-(\mathbf{v}\nabla)\mathbf{v}$$

the last equation is equivalent to (7). The form of the stress tensor (1) appears as a result of transformations of the relation (59) to the form (7). The stress tensor (1) is determined to within the tensor with a vanishing divergence.

Thus, differentiating Eqs. (52) and (53) and eliminating the variables $\varphi,\boldsymbol{\xi},\eta$, one obtains the curtailed system (6)–(8), whereas the system (10), (18), and (19) follows directly from the system (48) to (53) (i.e., without differentiating). It means that the system (10), (18), and (19) is an integrated system, whereas the curtailed system (6)–(8) is not, although formally they have the same order.

The action of the form (47), or close to this form was obtained by some authors,[23,27] but the quantities g^α, $\alpha=1,2,3$ are always considered as additional dependent variables (but not as indefinite functions of $\boldsymbol{\xi}$ which can be expressed via initial conditions). The action was not considered as a functional of fixed indefinite functions $g^\alpha(\boldsymbol{\xi})$.

The variable η was introduced to make the action invariant with respect to the transformations of the whole relabeling group (12). To understand what the η means from a mathematical viewpoint, let us return to the action (38), where the internal energy E has the form (39). Adding new variables j by means of designations (33), one obtains instead of (42)

$$\mathcal{A}_E[\rho,\mathbf{j},\boldsymbol{\xi},p]=\int\left\{\frac{\mathbf{j}^2}{2\rho}-\rho E-p_k\left[j^k-m\rho_0(\boldsymbol{\xi})\,\frac{\partial J}{\partial\xi_{0,k}}\right]\right\}dt\,d\mathbf{x}, \tag{60}$$

where E has the form (39).

Variation of (60) with respect to ξ_α leads to the equation

$$\hat{\mathcal{L}}_\alpha p=\rho\,\frac{\partial E(\rho,S_0(\boldsymbol{\xi}))}{\partial S_0}\,\frac{\partial S_0}{\partial\xi_\alpha},\quad \alpha=1,2,3, \tag{61}$$

where linear operators $\hat{\mathcal{L}}_\alpha$ are defined by (43). Equations (61) are linear nonuniform equations for the variables p. A solution of (61) is a sum of the general solution (46) of the uniform equations (43) and of a particular solution of the nonuniform equations (61). This particular solution depends on the form of the function S_0 and cannot be found in a general form. Adding to Eq. (41) an extraterm $-\eta j^k\partial_k S$ with η satisfying (51), a reduction of nonuniform equations (61) to uniform equations (43) appears to be possible. Thus, the extravariable η is responsible for the particular solution of (61).

From the viewpoint of the action (60) the dependence of the internal energy E on the entropy simply means a dependence of E on the labels $\boldsymbol{\xi}$ via a function $S(\boldsymbol{\xi})$. If such a dependence cannot be expressed through one function (for instance, $E=E[\rho,S_1(\boldsymbol{\xi}),S_2(\boldsymbol{\xi})]$) the ideal fluid is described by two entropies S_1 and S_2 and by two temperatures $T_1=\partial E/\partial S_1$, $T_2=\partial E/\partial S_2$. Such a situation may appear for a conducting fluid in a strong magnetic field, where there are two temperatures—longitudinal and transversal.

Thus five equations (10), (18), and (19) with S, ρ, and $\mathbf{v}$, defined, respectively, by (15), (16), and (17), constitute the fifth-order system for five dependent variables $\xi=\{\xi_0,\boldsymbol{\xi}\},\eta$. Equations (6), (8), (10), (18), (19) constitute the seventh-order system for seven variables ρ,ξ,φ,η,S.

IV. INITIAL AND BOUNDARY CONDITIONS

Boundary conditions describing vessel walls can be taken into account by means of a proper choice of the internal energy $E(x,\rho,\nabla\rho,S)$ which can include the energy of the fluid in an external potential U,

$$E = E_0(\rho, \nabla\rho, S) + U(t, \mathbf{x}),$$

where U is some given external potential. For instance, let the fluid move inside a volume $\mathcal{V}$. Then

$$U(\mathbf{x}) = \begin{cases} 0 & \text{inside } \mathcal{V} \\ \infty & \text{outside } \mathcal{V}. \end{cases}$$

Such a choice of energy E provides that the fluid does not escape the volume $\mathcal{V}$.

In this section let us set for simplicity the scale constant $b=1$, and consider the case when E does not depend on $\nabla\rho$, and the fluid flow is considered in the space–time region Ω defined by inequalities

$$\Omega: \quad t \geqslant 0, \quad x^3 \geqslant 0.$$

The region Ω has two boundaries: $\mathcal{I}$ defined by the relations $t=0$, $x^3 \geqslant 0$, and $\mathcal{B}$ defined by the relations $x^3=0$, $t \geqslant 0$. The initial conditions for the system of equations (6)–(8), and (10) have the form

$$\rho(0,\mathbf{x}) = \rho_{\text{in}}(\mathbf{x}), \quad v^{\alpha}(0,\mathbf{x}) = v^{\alpha}_{\text{in}}(\mathbf{x}), \quad \alpha = 1,2,3, \tag{62}$$

$$S(0,\mathbf{x}) = S_{\text{in}}(\mathbf{x}), \quad \xi_{\alpha}(0,\mathbf{x}) = \xi^{\alpha}_{\text{in}}(\mathbf{x}), \quad \alpha = 1,2,3 \tag{63}$$

at $\mathbf{x} \in \mathcal{I}$ ($t=0$, $x^3 \geqslant 0$). Here ρ_{in}, $\mathbf{v}_{\text{in}}$, S_{in}, and $\boldsymbol{\xi}_{\text{in}}$ are given functions of argument $\mathbf{x}$. The boundary conditions on the boundary $\mathcal{B}$ of Ω have the form:

$$\rho(x)|_{x^3=0} = \rho_b(t,\mathbf{y}), \quad S(x)|_{x^3=0} = S_b(t,\mathbf{y}), \quad \{t,\mathbf{y}\} \in \mathcal{B}, \tag{64}$$

$$v^{\alpha}(x)|_{x^3=0} = v^{\alpha}_b(t,\mathbf{y}), \quad \alpha = 1,2,3, \quad \{t,\mathbf{y}\} \in \mathcal{B}, \tag{65}$$

$$\xi_{\alpha}(x)|_{x^3=0} = \xi^{\alpha}_b(t,\mathbf{y}), \quad \alpha = 1,2,3, \quad \{t,y\} \in \mathcal{B}, \tag{66}$$

where

$$\mathbf{y} \equiv \{x^1, x^2\}. \tag{67}$$

Here ρ_b, S_b, $\mathbf{v}_b$, and $\boldsymbol{\xi}_b$ are given functions of the argument $\{t,\mathbf{y}\}$.

Let us show that indefinite functions $\mathbf{g}, S_0, \rho_0$ can be expressed via initial and boundary conditions (62)–(66). The initial conditions for the system (48)–(53) have the form

$$\xi_{\alpha}(0,\mathbf{x}) = \xi^{\alpha}_{\text{in}}(\mathbf{x}), \quad \alpha = 1,2,3, \tag{68}$$

$$\rho(0,\mathbf{x}) = \rho_{\text{in}}(\mathbf{x}), \quad S(0,\mathbf{x}) = S_0[\boldsymbol{\xi}_{\text{in}}(\mathbf{x})], \tag{69}$$

$$\varphi(0,\mathbf{x}) = \varphi_{\text{in}}(\mathbf{x}), \quad \eta(0,\mathbf{x}) = \eta_{\text{in}}(\mathbf{x}), \tag{70}$$

(68)–(70) take place at $\mathbf{x} \in \mathcal{I}$. The functions $\varphi_{\text{in}}(\mathbf{x})$, $\eta_{\text{in}}(\mathbf{x})$ as well $g^{\alpha}(\boldsymbol{\xi})$ are to be determined from the relations

$$\partial_{\alpha}\varphi_{\text{in}}(\mathbf{x}) + g^{\beta}[\boldsymbol{\xi}_{\text{in}}(\mathbf{x})]\partial_{\alpha}\xi^{\beta}_{\text{in}}(\mathbf{x}) - \eta_{\text{in}}(\mathbf{x}) \frac{\partial S_0[\boldsymbol{\xi}_{\text{in}}(\mathbf{x})]}{\partial \xi^{\beta}_{\text{in}}} \partial_{\alpha}\xi^{\beta}_{\text{in}}(\mathbf{x}) = v^{\alpha}_{\text{in}}(\mathbf{x}), \quad \alpha = 1,2,3, \quad \mathbf{x} \in \mathcal{I}. \tag{71}$$

It is clear that five functions $\mathbf{g}, \varphi_{\text{in}}, \eta_{\text{in}}$ cannot be determined unambiguously from three relations (71).

There are at least two different approaches for the determination of the functions $\boldsymbol{\xi}_{\text{in}}(\mathbf{x})$ and $\mathbf{g}(\boldsymbol{\xi}.)$

(1) One fixes the functions $\xi_{\rm in}^{\alpha}(\mathbf{x})$ in some conventional way, sets

$$\varphi_{\rm in}(\mathbf{x})=0, \quad \eta_{\rm in}(\mathbf{x})=0, \quad \mathbf{x}\in\mathcal{I}, \tag{72}$$

and determines functions $\mathbf{g}$ from the three relations (71).

(2) Functions $\mathbf{g}$ are fixed in some conventional way, and the remaining functions are determined from relations (71).

The first way. Let condition (68) be given in the form

$$\xi_{\alpha}(0,\mathbf{x})=\xi_{\rm in}^{\alpha}(\mathbf{x})=x^{\alpha}, \quad \alpha=1,2,3, \quad \mathbf{x}\in\mathcal{I}. \tag{73}$$

In other words, at $t=0$ the labels $\boldsymbol{\xi}$ coincide with the Eulerian coordinates. By virtue of (72), the relations (71) take the form

$$g^{\beta}[\boldsymbol{\xi}_{\rm in}(\mathbf{x})]=v_{\rm in}^{\beta}(\mathbf{x}), \quad \alpha=1,2,3, \quad \mathbf{x}\in\mathcal{I}, \tag{74}$$

which are resolved in the form

$$g^{\alpha}(\boldsymbol{\xi})=v_{\rm in}^{\alpha}(\boldsymbol{\xi}), \quad \alpha=1,2,3, \quad \xi_3>0. \tag{75}$$

Thus, the functions $\mathbf{g}$ are expressed through initial conditions (62).

The boundary conditions for the system of equations (48)–(53) have the form

$$\xi_{\alpha}(x)|_{x^3=0}=\xi_b^{\alpha}(t,\mathbf{y}), \quad \alpha=1,2,3, \quad \{t,\mathbf{y}\}\in\mathcal{B}, \tag{76}$$

$$S(x)|_{x^3=0}=S_0[\boldsymbol{\xi}_b(t,\mathbf{y})]=S_b(t,\mathbf{y}), \quad \{t,\mathbf{y}\}\in\mathcal{B}, \tag{77}$$

$$\rho(x)|_{x^3=0}=\rho_b(t,\mathbf{y}), \quad \mathbf{v}(x)|_{x^3=0}=\mathbf{v}_b(t,\mathbf{y}), \quad \{t,\mathbf{y}\}\in\mathcal{B}, \tag{78}$$

$$\varphi(x)|_{x^3=0}=\eta(x)|_{x^3=0}=0, \quad \{t,\mathbf{y}\}\in\mathcal{B}. \tag{79}$$

Let us set

$$\xi_b^{\alpha}(t,\mathbf{y})=x^{\alpha}, \quad \alpha=1,2, \quad \xi_b^{3}(t,\mathbf{y})=-ct, \quad (t,\mathbf{y})\in\mathcal{B}, \tag{80}$$

where c is a constant.

Writing relations (10) and (53) for $\xi_3<0$ on the boundary $\mathcal{B}$ and using (79) and (80), one obtains constraints for the functions $\mathbf{g}(\boldsymbol{\xi})$,

$$g^{\beta}[\boldsymbol{\xi}_b(t,\mathbf{y})]\partial_{\alpha}\xi_b^{\beta}(t,\mathbf{y})=v_b^{\alpha}(t,\mathbf{y}), \quad \alpha=1,2, \quad \{t,\mathbf{y}\}\in\mathcal{B}, \tag{81}$$

$$g^{\beta}[\boldsymbol{\xi}_b(t,\mathbf{y})]\partial_0\xi_b^{\beta}(t,\mathbf{y})=-K_b(t,\mathbf{y}), \quad \{t,\mathbf{y}\}\in\mathcal{B}, \tag{82}$$

where

$$K_b(t,\mathbf{y})\equiv\frac{\mathbf{v}_b^2(t,\mathbf{y})}{2}+\frac{\partial\{\rho_b(t,\mathbf{y})E[\rho_b(t,\mathbf{y}),S_b(t,\mathbf{y})]\}}{\partial\rho_b(t,\mathbf{y})}, \quad \{t,\mathbf{y}\}\in\mathcal{B}, \tag{83}$$

Substituting relations (80) into (81) and (82), one obtains three equations for the determination of functions $\mathbf{g}(\boldsymbol{\xi})$. Resolving this system of equations with respect to $\mathbf{g}$, one obtains

$$\begin{aligned} g^{\alpha}(\boldsymbol{\xi})&=v_b^{\alpha}(-\xi_3/c,\xi_1,\xi_2), \quad \alpha=1,2, \quad \xi_3<0,\\ g^{3}(\boldsymbol{\xi})&=c^{-1}K_b(-\xi_3/c,\xi_1,\xi_2), \quad \xi_3<0. \end{aligned} \tag{84}$$

Thus, $\mathbf{g}(\boldsymbol{\xi})$ is determined by (75) for $\xi_3>0$ and by (84) for $\xi_3<0$. In other words, the boundary conditions and the initial conditions determine the vector field $\mathbf{g}(\boldsymbol{\xi})$ in different regions of the argument $\boldsymbol{\xi}$. The field $\mathbf{g}(\boldsymbol{\xi})$ can describe both initial and boundary conditions. For any fluid flow the system (10), (18), and (19) of dynamic equations for variables φ, η, and $\boldsymbol{\xi}$ is to be solved under universal initial conditions (72) and (73) and under universal boundary conditions (79) and (80). All essential information on the fluid flow is found in the dynamic equations (10), (18), and (19), where the quantities $S,\rho,\mathbf{v}$ are determined by (15)–(17).

The second way. Let us choose the functions $\mathbf{g}$ in a simple form. Let for instance,

$$g^1(\boldsymbol{\xi})=\xi_2, \quad g^2(\boldsymbol{\xi})=0, \quad g^3(\boldsymbol{\xi})=0.$$

Let us set

$$\chi=\varphi, \quad \lambda=\xi_2, \quad \mu=\xi_1.$$

Then expression (17) takes the form

$$\mathbf{u}(\chi,\lambda,\mu,\eta,S)\equiv\nabla\chi+\lambda\nabla\mu-\eta\nabla S=\mathbf{v}, \tag{85}$$

where χ, λ, and μ, are Clebsch potentials.[21,22] Now six equations (6), (8), (49)–(53), and (55) [(49) for $\alpha=3$ is of no importance] for six dependent variables ρ, χ, λ, μ, η, and S do not contain indefinite functions and have an unambiguous form,

$$\partial_0\rho+\nabla(\rho\mathbf{u})=0, \quad \partial_0\lambda+(\mathbf{u}\nabla)\lambda=0, \quad \partial_0\mu+(\mathbf{u}\nabla)\mu=0, \quad \partial_0 S+(\mathbf{u}\nabla)S=0, \tag{86}$$

$$\partial_0\eta+(\mathbf{u}\nabla)\eta=-\frac{\partial E}{\partial S}, \quad \partial_0\chi+\lambda\partial_0\mu-\eta\partial_0 S+\frac{1}{2}\mathbf{u}^2+\frac{\partial(\rho E)}{\partial\rho}=0,$$

where $\mathbf{u}$ is defined by (85).

The initial conditions for variables ρ, χ, λ, μ, η, and S are determined by the relations

$$\rho(0,\mathbf{x})=\rho_{\text{in}}(0,\mathbf{x}), \quad S(0,\mathbf{x})=S_{\text{in}}(0,\mathbf{x}), \tag{87}$$

$$\nabla\chi_{\text{in}}+\lambda_{\text{in}}\nabla\mu_{\text{in}}-\eta_{\text{in}}\nabla S_{\text{in}}=-\mathbf{v}_{\text{in}}. \tag{88}$$

The three equations (87) and (88) do not determine the initial conditions

$$\chi(0,\mathbf{x})=\chi_{\text{in}}(\mathbf{x}), \quad \lambda(0,\mathbf{x})=\lambda_{\text{in}}(\mathbf{x}), \tag{89}$$

$$\mu(0,\mathbf{x})=\mu_{\text{in}}(\mathbf{x}), \quad \eta(0,\mathbf{x})=\eta_{\text{in}}(\mathbf{x}), \tag{90}$$

unambiguously.

If the fluid is described in terms of Clebsch potentials, the dynamic equations contain neither arbitrary functions, nor information about the initial conditions. It should be interpreted in the sense that the description (85)–(86) in terms of the Clebsch potentials is a result of a change of variables in dynamic equations (6)–(8), whereas the description (48)–(53) is a result of the integration of the dynamic equations (6)–(8) and (10). In other words, the description (85)–(86) in terms of Clebsch potentials relates to the description (48)–(53) in the same way, as a particular solution of a system of differential equations relates to a general solution of the same system. Let us note that there are many other ways for a determination of indefinite functions $\mathbf{g}(\boldsymbol{\xi})$.

V. WAVE FUNCTION AND SPIN

Equations (6), (8), (10), (18), and (19) can be derived from the action functional

$$\mathcal{A}[\rho,\varphi,\boldsymbol{\xi},\eta,S]=\int \rho\left[-\pi_0(\varphi,\boldsymbol{\xi},\eta,S)-\frac{1}{2}\boldsymbol{\pi}^2(\varphi,\boldsymbol{\xi},\eta,S)-E(x,\rho,\boldsymbol{\nabla}\rho,S)\right]d^4x, \tag{91}$$

where $\boldsymbol{\pi}=\{\pi_1,\pi_2,\pi_3\}$, and $\pi_k=p_k-\eta\partial_kS$, $k=0,1,2,3$ are determined by relations (46),

$$\pi_k(\varphi,\boldsymbol{\xi},\eta,S)\equiv b[\partial_k\varphi+g^\alpha(\boldsymbol{\xi})\partial_k\xi_\alpha]-\eta\partial_kS, \quad k=0,1,2,3. \tag{92}$$

The action (91) results from the action (47) after elimination of the variable $\mathbf{j}$ from the relations (47) and (52). The functions $\mathbf{g}=\{g^\beta(\boldsymbol{\xi})\}$, $\beta=1,2,3$ are considered as fixed functions of their arguments. Equations (6), (8), (10), (18), and (19) can be obtained as a result of variation with respect to φ, η, $\boldsymbol{\xi}$, S, and ρ, respectively. Equation (10) is obtained, provided the field $\mathbf{g}$ is nonpotential. If the field $\mathbf{g}$ is potential $g^\alpha(\boldsymbol{\xi})\doteq\partial\Phi/\partial\xi_\alpha$, it can be included in the variable φ by means of the substitution

$$\varphi+\Phi\rightarrow\varphi.$$

In this case the action (91) does not depend on $\boldsymbol{\xi}$, and (10) may be omitted.

Let us introduce an n-component complex function $\psi=\{\psi_\alpha\}$, $\alpha=1,2,\ldots,n$, defining it by the relations

$$\psi_\alpha=\sqrt{\rho}e^{i\varphi}u_\alpha(\boldsymbol{\xi}), \quad \psi_\alpha^*=\sqrt{\rho}e^{-i\varphi}u_\alpha^*(\boldsymbol{\xi}), \quad \alpha=1,2,\ldots,n,$$

$$\psi^*\psi\equiv\sum_{\alpha=1}^n\psi_\alpha^*\psi_\alpha,$$

where an asterisk (*) means the complex conjugate, $u_\alpha(\boldsymbol{\xi})$, $\alpha=1,2,\ldots,n$ are only functions of variables $\boldsymbol{\xi}$, and satisfy the relations

$$-\frac{i}{2}\sum_{\alpha=1}^n\left(u_\alpha^*\frac{\partial u_\alpha}{\partial\xi_\beta}-\frac{\partial u_\alpha^*}{\partial\xi_\beta}u_\alpha\right)=g^\beta(\boldsymbol{\xi}), \quad \beta=1,2,3, \quad \sum_{\alpha=1}^n u_\alpha^*u_\alpha=1. \tag{93}$$

n is such a natural number that equations (93) admit a solution. In general n may depend on the form of the arbitrary integration functions $\mathbf{g}=\{g^\beta(\boldsymbol{\xi})\}$, $\beta=1,2,3$.

It is easy to verify that

$$\rho\pi_k(\varphi,\boldsymbol{\xi},\eta,S)=-\frac{ib}{2}(\psi^*\partial_k\psi-\partial_k\psi^*\cdot\psi)-\eta\partial_kS\psi^*\psi, \quad k=0,1,2,3, \tag{94}$$

$$\rho=\psi^*\psi, \quad \mathbf{j}=-\frac{ib}{2}(\psi^*\boldsymbol{\nabla}\psi-\boldsymbol{\nabla}\psi^*\cdot\psi)-\eta\boldsymbol{\nabla}S\psi^*\psi. \tag{95}$$

The variational problem with the action (91) appears to be equivalent to the variational problem with the action functional

$$\mathcal{A}[\psi,\psi^*,\eta,S]=\int\left\{\frac{ib}{2}(\psi^*\partial_0\psi-\partial_0\psi^*\cdot\psi)+\eta\partial_0S\psi^*\psi-\frac{1}{2\psi^*\psi}\left[\frac{ib}{2}(\psi^*\boldsymbol{\nabla}\psi-\boldsymbol{\nabla}\psi^*\cdot\psi)\right.\right.$$
$$\left.\left.+\eta\boldsymbol{\nabla}S\psi^*\psi\right]^2-E[x,\psi^*\psi,\boldsymbol{\nabla}(\psi^*\psi),S]\psi^*\psi\right\}d^4x. \tag{96}$$

Note that the function ψ considered as a function of independent variables $\{t,\mathbf{x}\}$ is very indefinite in the sense that the same fluid flow may be described by different ψ functions. There are two reasons for such an indefiniteness. First, the functions $u_\alpha(\boldsymbol{\xi})$ are not determined uniquely by differential equations (93). Second, their arguments $\boldsymbol{\xi}$ as functions of x are determined only to within the transformation (12). Description of a fluid in terms of the function ψ is more indefinite than the description in terms of the hydrodynamic potentials $\boldsymbol{\xi}$. Information about initial and boundary conditions contained in the functions $\mathbf{g}(\boldsymbol{\xi})$ is lost at the description in terms of the ψ function. The ψ function can be obtained from the Clebsch variables by means of a proper change of variables.[11]

Let the function ψ have n components. Regrouping components of the function ψ in the action (96), one obtains the action in the form

$$\mathcal{A}_E[\psi,\psi^*,\eta,S]=\int\left\{\frac{1}{2}\,[\psi^*(ib\partial_0+A_0)\psi+(-ib\partial_0\psi^*+A_0\psi^*)\psi]\right.$$
$$-\frac{1}{2}\,(ib\boldsymbol{\nabla}\psi^*-\mathbf{A}\psi^*)(-ib\boldsymbol{\nabla}\psi-\mathbf{A}\psi)$$
$$\left.+\frac{b^2}{4}\sum_{\alpha,\beta=1}^{n}Q^*_{\alpha\beta,\gamma}Q_{\alpha\beta,\gamma\rho}+\frac{b^2}{8\rho}\,(\boldsymbol{\nabla}\rho)^2-\rho E\right\}d^4x,\quad \rho\equiv\psi^*\psi, \tag{97}$$

where

$$A=\{A_0,\mathbf{A}\},\quad A_0\equiv\eta\partial_0 S,\quad \mathbf{A}\equiv\eta\boldsymbol{\nabla}S,$$
$$Q_{\alpha\beta,\gamma}=\frac{1}{\psi^*\psi}\begin{vmatrix}\psi_\alpha & \psi_\beta\\ \partial_\gamma\psi_\alpha & \partial_\gamma\psi_\beta\end{vmatrix},\quad \alpha,\beta=1,2,...,n,\quad \gamma=1,2,3. \tag{98}$$

Corresponding dynamic equations have the form

$$\frac{\delta\mathcal{A}}{\delta\psi^*_\alpha}=(ib\partial_0+A_0)\psi_\alpha-\frac{1}{2}\,(ib\boldsymbol{\nabla}+\mathbf{A})^2\psi_\alpha-\frac{b^2}{4}\sum_{\mu,\nu=1}^{n}Q^*_{\mu\nu,\gamma}Q_{\mu\nu,\gamma}\psi_\alpha+\frac{b^2}{2}\sum_{\nu=1}^{n}Q_{\alpha\nu,\gamma}\partial_\gamma\psi^*_\nu$$
$$+\frac{b^2}{2}\sum_{\nu=1}^{n}\partial_\gamma(Q_{\alpha\nu,\gamma}\psi^*_\nu)+\frac{\partial}{\partial\rho}\left[\frac{b^2}{8\rho}\,(\boldsymbol{\nabla}\rho)^2-\rho E\right]\psi_\alpha$$
$$-\partial_\gamma\left\{\frac{\partial}{\partial\rho_\gamma}\left[\frac{b^2}{8\rho}\,(\boldsymbol{\nabla}\rho)^2-\rho E\right]\right\}\psi_\alpha=0,\quad \alpha=1,2,...,n, \tag{99}$$

$$\frac{\delta\mathcal{A}}{\delta S}=\partial_i(j^i\eta)-\frac{\partial(\rho E)}{\partial S}=0, \tag{100}$$

$$\frac{\delta\mathcal{A}}{\delta\eta}=-\partial_i(j^iS)=0, \tag{101}$$

where $j=\{\rho,\mathbf{j}\}=\{j^k\}$, $k=0,1,2,3$ is defined by (95).

In the case of the irrotational flow, when $g^\alpha(\boldsymbol{\xi})=\partial\Phi(\boldsymbol{\xi})/\partial\xi_\alpha$, equations (93) have a solution for $n=1$, and the function ψ may have one component. Then all $Q_{\alpha\beta,\gamma}\equiv 0$, as it follows from Eq. (98).

Let us consider an irrotational flow of a fluid with the internal energy per unit mass defined by relation (3), where m is the mass of a stochastic particle associated with the fluid. The internal energy does not depend on the entropy, and according to (3) and (100) the variable η is a function of only labels $\boldsymbol{\xi}$. Then the expression $\eta\partial_k S$ has the form $f^\alpha(\boldsymbol{\xi})\partial_k\xi_\alpha$. It may be included in the term $g^{\alpha}(\boldsymbol{\xi})\xi_{\alpha,k}$. It means that without a loss of generality one may set $\eta\equiv 0$, $S\equiv 0$. Then for an irrotational flow, when the ψ function is one component, Eq. (99) takes the form

$$\frac{\delta\mathcal{A}}{\delta\psi^*}=ib\partial_0\psi-\frac{b^2}{2}\boldsymbol{\nabla}^2\psi+\left(b^2-\frac{\hbar^2}{m^2}\right)\left\{\frac{\partial}{\partial\rho}\frac{(\boldsymbol{\nabla}\rho)^2}{8\rho}-\partial_{\gamma}\left[\frac{\partial}{\partial\rho_{\gamma}}\frac{(\boldsymbol{\nabla}\rho)^2}{8\rho}\right]\right\}\psi=0. \tag{102}$$

Choosing arbitrary constant b in the form $b=-\hbar/m$, one obtains instead of Eq. (102) the well-known Schrödinger equation

$$i\hbar\partial_0\psi+\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2\psi=0,$$

where the complex variable ψ is known as the wave function. The Schrödinger equation describes an irrotational flow of the Madelung fluid.[4]

On this basis it is possible in general to identify the function ψ with the wave function and consider the wave function as a way of description of any ideal fluid. If the fluid flow is rotational, the dynamic equation in terms of the ψ function is nonlinear, even in the case (3) and at $b=-\hbar/m$. In this case the ψ function is not one component, and the quantities $Q_{\alpha\beta,\gamma}$ do not vanish generally.

In general, the dynamic equation (99) for the ψ function is nonlinear and rather complicated. But for a special form (3) of the internal energy and for a special form of the arbitrary phase constant b the dynamic equation in terms of the ψ function becomes linear and simple.

It is worth noting that the internal energy per unit mass (3) associates with the mean diffusion velocity $\mathbf{v}_{\text{dif}}=-D\boldsymbol{\nabla}\rho/\rho$ describing the mean motion of random wandering of stochastic particles (D is the diffusion coefficient). The diffusion velocity is characteristic for any stochastic particles (both Brownian and quantum). The Brownian fluid is dissipative, and the evolution of the fluid state ρ is described directly by $\mathbf{v}_{\text{dif}}$ by means of the continuity equation

$$\partial_0\rho+\boldsymbol{\nabla}(\rho\mathbf{v}_{\text{dif}})=0.$$

For the ideal Madelung fluid the diffusion velocity influences the fluid flow via the internal fluid energy per unit mass determined by means of relation (3). Besides, the diffusion coefficients D are different for Brownian particles and for quantum ones, because the origin of the stochasticity is different in the two cases.

The number n of the ψ-function components in the actions (96) and (97) is arbitrary. A formal variation of the action with respect to ψ_{α} and ψ^*_{α}, $\alpha=1,2,...,n$ leads to $2n$ real dynamic equations, but not all of them are independent. There are such combinations of variations $\delta\psi_{\alpha},\delta\psi^*_{\alpha}$, $\alpha=1,2,...,n$ which do not change expressions (94) and (95). Such combinations of variations $\delta\psi_{\alpha},\delta\psi^*_{\alpha}$, $\alpha=1,2,...,n$ do not change the action (96), and corresponding combinations of dynamic equations $\delta\mathcal{A}/\delta\psi_{\alpha}=0$, $\delta\mathcal{A}/\delta\psi^*_{\alpha}=0$ are identities that associate with a correlation between dynamic equations. Thus, increasing the number n, one increases the number of dynamic equations, but the number of independent dynamic equations remains the same.

In such a situation it is important to determine the minimal number n_m of the ψ-function components, sufficient for a description of the given vector field $g^{\beta}(\boldsymbol{\xi})$ in the space V_{ξ} of the labels $\boldsymbol{\xi}$.

Note that under the relabeling transformations (12) the quantity $\mathbf{g}(\boldsymbol{\xi})$ transforms as a vector

$$g^{\beta}(\boldsymbol{\xi})\to\tilde{g}^{\beta}(\tilde{\boldsymbol{\xi}})=\frac{\partial\xi_{\alpha}}{\partial\tilde{\xi}_{\beta}}g^{\alpha}(\boldsymbol{\xi}),\quad \beta=1,2,3.$$

It is necessary for the quantities (94), (95) and the action (91) to be invariant with respect to the transformation (12)

Let $\mathcal{G}$ be a set of all vector fields $g^{\beta}(\boldsymbol{\xi})$ in V_{ξ}, and $\mathcal{G}_n$ be a set of such vector fields $g^{\beta}(\boldsymbol{\xi})$ in V_{ξ} which can be presented in the form

$$g^{\beta}(\boldsymbol{\xi})=\sum_{k=1}^{n} \eta_k^2(\boldsymbol{\xi})\partial\zeta_k(\boldsymbol{\xi})/\partial\xi_{\beta}, \quad \beta=1,2,3, \quad \eta_1\equiv 1, \tag{103}$$

where n is a fixed natural number and the functions η_k, ζ_k, $k=1,2,...,n$ are scalars in V_{ξ}. Under the relabeling transformation (12) the functions (103) transform as follows:

$$\eta_k(\boldsymbol{\xi})\to\tilde{\eta}_k(\tilde{\boldsymbol{\xi}})=\eta_k(\boldsymbol{\xi}), \quad \zeta_k(\boldsymbol{\xi})\to\tilde{\zeta}_k(\tilde{\boldsymbol{\xi}})=\zeta_k(\boldsymbol{\xi}), \quad k=1,2,...,n,$$

$$g^{\beta}(\boldsymbol{\xi})\to\tilde{g}^{\beta}(\tilde{\boldsymbol{\xi}})=\frac{\partial\xi_{\alpha}}{\partial\tilde{\xi}_{\beta}}\,g^{\alpha}(\boldsymbol{\xi})=\frac{\partial\xi_{\alpha}}{\partial\tilde{\xi}_{\beta}}\sum_{k=1}^{n}\eta_k^2(\boldsymbol{\xi})\,\frac{\partial\zeta_k(\boldsymbol{\xi})}{\partial\xi_{\alpha}}=\sum_{k=1}^{n}\tilde{\eta}_k^2(\tilde{\boldsymbol{\xi}})\,\frac{\partial\tilde{\zeta}_k(\tilde{\boldsymbol{\xi}})}{\partial\tilde{\xi}_{\alpha}}.$$

In other words, a vector field $g^{\beta}(\boldsymbol{\xi})$ of the form (103) transforms into the vector field $\tilde{g}^{\beta}(\tilde{\boldsymbol{\xi}})$ of the same form (103), and the set $\mathcal{G}_n$ is invariant with respect to the group (12) of the relabeling transformations.

It is easy to see that

$$\mathcal{G}_{n-1}\subseteq\mathcal{G}_n, \quad \mathcal{G}_0=\varnothing, \quad n=1,2,...$$

because the nth term of the sum (103) can be combined with the first one, if ζ_n is a function of η_n. Let

$$\mathcal{S}_n=\mathcal{G}_n\backslash\mathcal{G}_{n-1}, \quad n=1,2,... .$$

Then

$$\mathcal{G}=\bigcup_{s=1}^{s=n_m}\mathcal{S}_s, \quad \mathcal{S}_l=\varnothing, \quad l=n_m+1,n_m+2,...,$$

where n_m is the number of nonempty invariant subsets of the set $\mathcal{G}$. Each subset $\mathcal{S}_k$ contains only such vector fields $g^{\beta}(\boldsymbol{\xi})$ which associate with the k-component ψ-function $\psi=\{\psi_{\alpha}\}$, $\alpha=1,2,...,k$, having the components

$$\psi_1=\left\{\left(1-\sum_{\alpha=2}^{k}\eta_{\alpha}^2\right)\rho\right\}^{1/2}\exp[i(\varphi+\zeta_1)],$$

$$\psi_{\alpha}=\eta_{\alpha}\sqrt{\rho}\,\exp[i(\varphi+\zeta_{\alpha}+\zeta_1)], \quad \alpha=2,3,...,k.$$

In particular, the set $\mathcal{S}_1$ associates with an irrotational flow, described by a one-component ψ function determined by one scalar ζ_1; and the set $\mathcal{S}_2$ associates with a rotational flow described by a two-component ψ function, determined by three scalar functions ζ_1, η_2, and ζ_2 (Clebsch variables).

In conventional quantum mechanics the number n of ψ-function components is connected with the spin s of the particle, described by the ψ, by means of the relation $s=(n-1)/2$. The spin is considered as an internal property of a quantum particle. Particles with different spins are considered as different physical objects, described by different dynamic equations.

In a like manner the irrotational flow, when $g^{\beta}(\boldsymbol{\xi})$ is described by one function ζ_1, associates with the kinematic spin (k-spin) $s=0$, whereas the rotational flow, when $g^{\beta}(\boldsymbol{\xi})$ is described by three scalar functions ζ_1, η_2, and ζ_2, associates with the kinematic spin $s=1/2$. The term ‘‘kinematic spin’’ (instead of ‘‘spin’’ simply) is used now for the following reasons.

First, the kinematic spin (k-spin) is determined by the form of the vector field $g^{\beta}(\boldsymbol{\xi})$, which arises essentially as a result of an integration of equations (43). The vector field $g^{\beta}(\boldsymbol{\xi})$ is a kinematic structure, because it does not depend on the form of the internal energy. At the same time the k-spin is not an internal property of a particle in itself, because the action (96) describes, at least, two different k-spins ($s=0$ and $s=1/2$) simultaneously, and the k-spin looks rather like an integration constant, than a property of single fluid particles.

Second, there is a distinction between the transformation properties of the spin and those of the k-spin under the space rotation group. Components of the n-component ψ function are scalars under the space rotation group for any value of n (and for any value of the k-spin s). In the conventional quantum mechanics the wave function transforms according to a representation of the rotation group. In particular, a one-component wave function $\psi(s=0)$ is a scalar, whereas the two-component ψ function $\psi(s=1/2)$ transforms as a spinor under the rotation group.

Taking into account the indefiniteness of the ψ function, it is possible to change transformation properties of the ψ function accompanying any spatial rotation by a proper transformation of the group (12). The additional transformations (12) can be chosen in such a way that the two-component ψ function becomes a spinor under spatial rotations. Then the formal distinction between the ''k-spin'' and the ''spin'' vanishes, and one can identify them.

For instance, let the two-component ψ function be written in the form

$$\psi=\sqrt{\rho}\,\exp[i(\varphi+\boldsymbol{\sigma}\boldsymbol{\xi})]\chi,$$

where ρ, φ, and $\boldsymbol{\xi}$ are scalar functions of x, $\boldsymbol{\sigma}=\{\sigma_{\alpha}\}$, $\alpha=1,2,3$ are the Pauli matrices

$$\sigma_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix},\quad \sigma_2=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix},\quad \sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix},$$

and χ is a two-component constant column $(\chi^{*}\chi=1)$.

Let any infinitesimal spatial rotation

$$x^0\to\tilde{x}^0=x^0,\quad \mathbf{x}\to\tilde{\mathbf{x}}=\mathbf{x}+\boldsymbol{\omega}\times\mathbf{x}+O(\boldsymbol{\omega}^2),\quad |\omega|\ll 1 \tag{104}$$

be accompanied by an infinitesimal transformation

$$\boldsymbol{\xi}\to\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}-\boldsymbol{\omega}/2. \tag{105}$$

Then the ψ function transforms as a spinor with respect to the combined transformation (104) and (105),

$$\psi(x)\to\tilde{\psi}(\tilde{x})=\sqrt{\tilde{\rho}(\tilde{x})}\,\exp[i(\tilde{\varphi}(\tilde{x})+\boldsymbol{\sigma}\tilde{\boldsymbol{\xi}}(\tilde{x}))]\chi=\exp(-i\boldsymbol{\omega}\boldsymbol{\sigma}/2)\psi(x)+O(\boldsymbol{\omega}^2),$$

and as a scalar with respect to the space rotation (104) alone.

If the dynamic system is described in terms of a two-component ψ function, the labels $\boldsymbol{\xi}$ are not mentioned at all, and the ψ function can be considered equally ready as a scalar and as a spinor.

For the two-component ψ function the following identity takes place:

$$(\boldsymbol{\nabla}\rho)^2-(\psi^{*}\boldsymbol{\nabla}\psi-\boldsymbol{\nabla}\psi^{*}\psi)^2\equiv 4\rho\boldsymbol{\nabla}\psi^{*}\boldsymbol{\nabla}\psi-\rho^2\mathbf{s}^2, \tag{106}$$

$$\rho\equiv\psi^{*}\psi,\quad \mathbf{s}\equiv\psi^{*}\boldsymbol{\sigma}\psi/(2\rho),\quad \boldsymbol{\sigma}=\{\sigma_{\alpha}\},\quad \alpha=1,2,3, \tag{107}$$

where σ_{α} are Pauli matrices. By virtue of the identity (106) the action (96) reduces to the form

$$\mathcal{A}[\psi,\psi^*,\eta,S]=\int\Bigg\{\frac{1}{2}[\psi^*(ib\partial_0+A_0)\psi+(-ib\partial_0+A_0)\psi^*\psi]-\frac{1}{2}[ib\nabla\psi^*-\mathbf{A}\psi^*][-ib\nabla\psi-\mathbf{A}\psi]+\frac{b^2}{2}(\nabla s_\alpha)(\nabla s_\alpha)\rho+\frac{b^2}{8\rho}(\nabla\rho)^2-\rho E\Bigg\}d^4x, \tag{108}$$

$$\rho\equiv\psi^*\psi,\quad A_k\equiv\eta\partial_k S,\quad k=0,1,2,3$$

where s_α are defined by Eq. (107). The quantity $\mathbf{s}=\{s_\alpha\}$, $\alpha=1,2,3$ associates with the mean spin (especially, if $b=-\hbar/m$), because it is constructed on the base of the Pauli matrices. As one can see, in Eq. (108) s_α convolutes only with s_α, but not with ∇_α. As a result the action (108) is invariant with respect to space–time rotations and relabeling transformations (12) separately.

It is interesting to note in this connection that the action $\mathcal{A}_P[\psi,\psi^*]$ for the dynamical system $\mathcal{S}_P[\psi]$, described by the Pauli equation, implies the convolution between $\mathbf{s}$ and ∇. The action $\mathcal{A}_P[\psi,\psi^*]$ is invariant only with respect to the combined transformation (104) and (105), i.e., if the ψ is considered as a spinor, but it is not invariant with respect to the transformation (104), when ψ is considered as a scalar. In other words, the action $\mathcal{A}_P[\psi,\psi^*]$ is not invariant with respect to the rotation group (104), if $\mathcal{S}_P[\psi]$ is considered as a fluidlike dynamic system. The same action $\mathcal{A}_P[\psi,\psi^*]$ can be made invariant with respect to the rotation group (104) alone, provided ψ is considered as a fundamental object (not as an attribute of a dynamic system). In the last case the ψ is declared as a spinor, but the mathematical object, described by the action $\mathcal{A}_P[\psi,\psi^*]$, stops being a dynamic system. It may be regarded, for instance, as a ‘‘*dynamic system restricted by quantum axiomatics*,’’ but it is not a dynamic system in the conventional sense of this term, because a possibility of change of dynamic variables is restricted [any rotation (104) is accompanied by a proper relabeling (105)]. Of course, one may insist on the fundamental character of ψ and state that ψ is a spinor, but then $\mathcal{S}_P[\psi]$ stops being a dynamic system, and this fact may have far-reaching consequences (see the details in Ref. 28). There is a similar problem with the relativistic invariance of the dynamic system $\mathcal{S}_D[\psi]$ described by the Dirac equation.[29]

Thus, the *k*-spin is an integral property of a fluid flow, connected with kinematic properties of a dynamic system. Locally any vector field $\mathbf{g}(\boldsymbol{\xi})$ in the three-dimensional V_ξ can be written in the form

$$\mathbf{g}(\boldsymbol{\xi})=\nabla\zeta_1+\eta_2\nabla\zeta_2 \tag{109}$$

(expressed via Clebsch potentials), and one should expect that $s=1/2$ is a maximal *k*-spin of any flow in the three-dimensional space. But possible singular points of the representation (109) may lead to the circumstance that the spin of the total flow appears to be higher than $s=1/2$. It is connected with the so-called helicity of a vector field. Examples and a discussion of such a velocity field can be found in Refs. 16 and 30.

VI. CONCLUDING REMARKS

Taking into account dynamic equations for the labels $\boldsymbol{\xi}$ (Lagrangian coordinates considered as dynamic variables), one succeeds in integrating the system of hydrodynamic equations for an ideal fluid. This integration leads to the appearance of three arbitrary functions g^α, $\alpha=1,2,3$ of labels $\boldsymbol{\xi}$. The functions g^α form a vector $\mathbf{g}$ in the space V_ξ of labels $\boldsymbol{\xi}$. The vector $\mathbf{g}$ can be expressed via initial and boundary values of the fluid velocity. Dynamic equations appear to carry all essential information about the fluid motion. This form of the fluid description may appear to be important in such problems, where the character of the fluid motion depends essentially on the character of the initial and boundary conditions, and one needs to investigate dynamic equations together with boundary and initial conditions. For instance, such a necessity arises when investigating phenomena connected with a transition to an irregular motion of a fluid (turbulence).

The appearance of the field $\mathbf{g}$ activates the relabeling group. Invariant subsets of this group can be used for a classification of the fluid flows. The field $\mathbf{g}$ allows one to introduce such concepts as ψ function and k-spin which are new for the fluid dynamics. In some special cases these new constructive concepts can be identified with the wave function and the spin. Concepts of the wave function and of the spin are fundamental concepts in the sense that they cannot be defined via other more fundamental concepts. In the quantum mechanics the concepts of the wave function and of the spin are defined by their properties, i.e., by a system of axioms (quantum axiomatics).

On one hand there are derivative constructions of ψ function and k-spin, connected with stochastic electron via the construction of the statistical ensemble (ideal fluid). On the other hand, there are fundamental concepts of wave function and spin, connected with the stochastic electron via a system of axioms (quantum axiomatics). Sometimes the ψ function coincides with the wave function, but not always. Then such a question arises. Which of the two conception is valid? ψ function, or wave function?

A like problem arose in the theory of thermal phenomena. On one hand there was an axiomatic thermodynamics with its fundamental concepts of thermodynamic potentials. On the other hand there was the statistical physics, where the thermodynamic potentials were constructive quantities derived from the conception of the heat as a chaotic motion of molecules. Then the constructive theory (statistical physics) appeared to be more successful, than the axiomatic one (thermodynamics). Now the question is yet open, although there is a series of arguments in favor of the constructive approach which seems to be more reasonable and less enigmatic.

The considered general approach to the fluid dynamics is interesting from the point of view that sometimes it permits one to use advantages of the quantum technique in the dynamics of usual fluids, as well as the general technique of the fluid dynamics in application to quantum mechanics.

[1] Yu. A. Rylov, Phys. Essays **4**, 300 (1991).
[2] Yu. A. Rylov, Int. J. Theor. Phys. **19**, 123 (1980).
[3] Yu. A. Rylov, Found. Phys. **28**, 245 (1998).
[4] E. Madelung, Z. Phys. **40**, 322 (1926).
[5] L. de Broglie, Comptes Rendus **183**, 447 (1926).
[6] D. Bohm, Phys. Rev. **85**, 166 (1952); **85**, 180 (1952).
[7] F. Halbwachs, *Théorie Relativiste des Fluides á Spin* (Gauthier-Villar, Paris, 1960), and references therein.
[8] F. J. Belinfante, *A Survey of Hidden-Variables Theories* (Pergamon, Oxford, 1973), and references therein.
[9] D. Bohm and B. J. Hiley, Phys. Rep. **172**, 93 (1989), and references therein.
[10] P. Holland, *The Quantum Theory of Motion* (Cambridge University Press, Cambridge, 1993), and references therein.
[11] Yu. A. Rylov, J. Math. Phys. **30**, 2516 (1989).
[12] C. C. Lin, *Proceedings of the International School of Physics ''Enrico Fermi,''* Course XXI, Liquid Helium (Academic, New York, 1963), pp. 93–146.
[13] C. Eckart, Phys. Rev. **54**, 920 (1938).
[14] C. Eckart, Phys. Fluids **3**, 421 (1960).
[15] M. G. Calkin, Can. J. Phys. **41**, 2241 (1963).
[16] F. P. Bretherton, J. Fluid Mech. **44**, 19 (1970).
[17] J. L. Friedman and B. F. Schutz, Astrophys. J. **221**, 937 (1978).
[18] R. Salmon, AIP Conf. Proc. **88**, 127 (1982).
[19] V. E. Zakharov and E. A. Kuznetsov Sov. Sci. Rev. **91**, 1310 (1986).
[20] V. E. Zakharov and E. A. Kuznetsov, Usp. Fiz. Nauk **167**, 1137 (1997) (in Russian).
[21] A. Clebsch, J. Reine Angew. Math. **54**, 293 (1857).
[22] A. Clebsch, J. Reine Angew. Math. **56**, 1 (1859).
[23] R. Salmon, Annu. Rev. Fluid Mech. **20**, 225 (1988).
[24] J. W. Herivel, Proc. Cambridge Philos. Soc. **51**, 344 (1955).
[25] V. L. Berdichevski, *Variational Principles of the Continuum Medium Mechanics* (Nauka, Moscow, 1983) (in Russian).
[26] B. Davydov, Dok. Akad. Nauk USSR, **69**, 165 (1949) (in Russian).
[27] R. L. Seliger and F. R. S. Whitham, Proc. R. Soc. London, Ser. A **305**, 1 (1967).
[28] Yu. A. Rylov, Found. Phys. **25**, 1055 (1995).
[29] Yu. A. Rylov, Adv. Appl. Clif. Alg. **5**, 1 (1995).
[30] H. K. Moffatt, J. Fluid Mech. **35**, 117 (1969).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

The Stäckel systems and algebraic curves

A. V. Tsiganov[a)]
Department of Mathematical and Computational Physics, Institute of Physics, St. Petersburg University, 198 904, St. Petersburg, Russia

(Received 15 October 1997; accepted for publication 11 May 1998)

We show how the Abel–Jacobi map provides all the principal properties of an ample family of integrable mechanical systems associated to hyperelliptic curves. We prove that derivative of the Abel–Jacobi map is just the Stäckel matrix, which determines *n*-orthogonal curvilinear coordinate systems in a flat space. The Lax pairs, *r*-matrix algebras and explicit form of the flat coordinates are constructed. An application of the Weierstrass reduction theory allows us to construct several flat coordinate systems on a common hyperelliptic curve and to connect among themselves different integrable systems on a single phase space. © *1999 American Institute of Physics.* [S0022-2488(98)02009-X]

I. INTRODUCTION

In the classical mechanics the arrow from the initial physical variables to the action-angle variables is provided by the separation of variables and then by the Arnold construction of the action-angle representation.[1] The motion in the opposite direction ought to allow us to construct various mechanical integrable systems. However, in the action-angle representation all the mechanical systems with fixed number of degrees of freedom are indistinguishable. To describe some particular integrable system one should present an explicit construction of the initial physical variables as functions on the action-angle variables. This mapping contains all the information about a given integrable system. By using a variety of these mappings the different integrable models may be connected together via the common action-angle variables. For instance, mechanical systems may be related to nonlinear equations and to gauge field theory.

As an example, investigation of the finite-gap solutions of the nonlinear problems leads to the introduction[2–4] of analytic symplectic form Ω_g on the Jacobian fibrations and to the definition of the action-angle variables on the complex space of Liouville variables. In Ref. 5 it is shown that possible obstructions to the existence of global systems of action-angle variables on symplectic vector bundles are a nontrivial first Chern class and the presence of monodromy at singularities. Introduction of the action-angle representation enables one to consider mechanical integrable systems as systems associated with these variables on a torus bundle with base $\mathscr{M}$, moduli space of complex polynomials

$$F(\lambda)=\prod_{j=1}^{2g+1}(\lambda-\lambda_j), \tag{I.1}$$

and with a fiber $J(\mathscr{C})$, the g-dimensional complex Jacobian of auxiliary curve $\mathscr{C}$ defined by the Abel–Jacobi map $\mathscr{U}$.[1,6] The fact that action-angle variables could be used for quantization of classical systems leads to the introduction of semiclassical geometric phases. This approach results, for instance, in quantum conditions on the moduli of n-dimensional Jacobi varieties.[7]

By using this Abel–Jacobi map $\mathscr{U}$ and the Jacobi problem of inversion, the so-called root variables $\{p_j,q_j\}_{j=1}^n$ (Refs. 8 and 9) on an associated Riemann surface $\mathscr{C}$ may be constructed instead of the action-angle variables. In these root variables on the level of integrals of motion the action is represented as a sum of items depending on one coordinate only, i.e., these variables are separated variables. The corresponding Riemann surface $\mathscr{C}$ depends on parameters (moduli), parameterizing the moduli space $\mathscr{M}$ of $\mathscr{C}$.[4,10] In terms of mechanical integrable systems, the curve

[a)]Electronic mail: tsiganov@mph.niif.spbu.ru

$\mathscr{C}$ is interpreted as a time-independent spectral curve, integrals of motion are some specific coordinates on the moduli space $\mathscr{M}$, and the Jacobian $J(\mathscr{C})$ is a common level of the involutive integrals of the system.[1]

In what follows, we have to describe appropriate mechanical systems together with their phase space in initial physical coordinates $\{p_j, x_j\}_{j=1}^n$. In particular, separated coordinate systems ought to be orthogonal curvilinear coordinate systems on the flat Riemannian manifold.[11,12] In this case, these separated coordinate systems are associated to some solutions to the Lamé equation.[13–15] Recently, the solutions to this equation have been obtained in an explicit form with the help of the ''dressing procedure,''[13] the Baker-Akhiezer function[14,16] and the Lie algebraic construction[15] within framework of the inverse problem method.

The main objective of this paper is to illustrate how fixed mapping from the action-angle variables[1,6,14] to separated variables completely defines all the principal properties of mechanical systems. We shall consider the uniform Stäckel models associated to the Abel–Jacobi map $\mathscr{U}$ on the hyperelliptic curve $\mathscr{C}$ and the well-known elliptic, parabolic and spherical curvilinear coordinate systems on $\mathbb{R}^n$. Also, we discuss relations of these mechanical systems with other integrable models associated to the same algebraic curve.

II. THE STÄCKEL SYSTEMS

One of the oldest problems of the Hamiltonian mechanics is to find the quadratures for the integrable Hamiltonian systems. The simplest models integrable in quadratures are the Liouville systems and the Stäckel systems[17] (the Liouville systems are a particular case of the Stäckel systems).

Before proceeding farther it is useful to recall the classical work of Stäckel.[17] The system associated with the name of Stäckel[17] is a holonomic system on the phase space $\mathbb{R}^{2n}$. Their Hamiltonian is

$$H=\sum_{j=1}^{n} g_j(q_1,\dots,q_n)(p_j^2+U_j). \tag{II.1}$$

Here $\{p_j, q_j\}_{j=1}^n$ are canonical variables in $\mathbb{R}^{2n}$ with the standard symplectic structure and with the following Poisson brackets

$$\Omega_n=\sum_{j=1}^{n} dp_j\wedge dq_j, \quad \{p_j, q_k\}=\delta_{jk}. \tag{II.2}$$

Let us recall the Stäckel theorem.

Theorem 1: For a Hamiltonian system with Hamiltonian H of the form (II.1) the following assertions are equivalent:

(i) The associated Hamilton–Jacobi equation is separable.

(ii) There exists a nondegenerate $n\times n$ Stäckel matrix $\mathbf{S}$, whose elements $\mathfrak{s}_{kj}$ depend only on q_j:

$$\det \mathbf{S}\neq 0, \quad \frac{\partial \mathfrak{s}_{kj}}{\partial q_m}=0, \quad \text{for } j\neq m, \text{ and such that } \sum_{j=1}^{n} \mathfrak{s}_{kj}(q_j) g_j(q_1,\dots,q_n)=\delta_{k1}. \tag{II.3}$$

(iii) There exist n functionally independent integrals of motion which are quadratic in momenta.

Let $C=[c_{ik}]$ denote the inverse matrix to S such that $c_{j1}=g_j$. Then the Stäckel theorem[11,17] asserts that there are n first integrals of motion, namely

$$I_k=\sum_{j=1}^{n} c_{jk}(p_j^2+U_j), \quad I_1=H. \tag{II.4}$$

The common level surface of these integrals,

$$M_\alpha=\{z\in\mathbb{R}^{2n}: I_k(z)=\alpha_k, \quad k=1,\dots,n\},$$

is diffeomorphic to the n-dimensional real torus and one immediately obtains

$$p_j^2=\left(\frac{\partial S}{\partial q_j}\right)^2=\sum_{k=1}^n \alpha_k \mathfrak{s}_{kj}(q_j)-U_j(q_j), \tag{II.5}$$

where $S(q_1,\dots,q_n)$ is an action function.[1] If this real torus is part of a complex algebraic torus, then the corresponding mechanical system is called an algebraic completely integrable system.[18]

The Stäckel theorem allows one to reduce the solution of the equations of motion to a problem in algebraic geometry. We can regard each expression (II.5) as being defined on the Riemann surface

$$\mathscr{C}_j: \quad y_j^2=F_j(\lambda), \quad F_j(\lambda)=\sum_{k=1}^n \alpha_k \mathfrak{s}_{kj}(\lambda)-U_j(\lambda), \tag{II.6}$$

which depends on the values α_k of integrals of motion. All the pairs of variables (p_j,q_j) lie on these Riemann surfaces (II.6). Considered together, they determine an n-dimensional Lagrangian submanifold in $\mathbb{R}^{2n}$:

$$\mathscr{C}^{(n)}: \quad \mathscr{C}_1(p_1,q_1)\times\mathscr{C}_2(p_2,q_2)\times\cdots\times\mathscr{C}_n(p_n,q_n). \tag{II.7}$$

The associated Hamilton–Jacobi equation,

$$\frac{\partial S}{\partial t}+H\left(t,\frac{\partial S}{\partial q_1},\dots,\frac{\partial S}{\partial q_n},q_1,\dots,q_n\right)=0\Rightarrow g^{jj}\frac{\partial S}{\partial q_j}\frac{\partial S}{\partial q_j}=E, \tag{II.8}$$

on the local manifold $\mathscr{V}_n$ with diagonal metric $g^{jj}=g_j(q_1,\dots,q_n)$ analytic in the local coordinates $\{q_j\}$ has the following additive solution:

$$S(q_1,\dots,q_n)=\sum_{j=1}^n s_j(q_j), \quad s_j(q_j)=\int \sqrt{F_j(q_j)}\,dq_j, \tag{II.9}$$

with the functions $F_j(\lambda)$ defined in (II.6). Coordinates $q_j(t,\alpha_1,\dots,\alpha_n)$ are determined from the equations

$$\begin{gathered}\sum_{j=1}^n \int_{\gamma_0(p_0,q_0)}^{\gamma_j(p_j,q_j)} \frac{\mathfrak{s}_{1j}(\lambda)d\lambda}{\sqrt{\sum_{k=1}^n\alpha_k\mathfrak{s}_{1j}(\lambda)-U_j(\lambda)}}=\beta_1=t,\\ \sum_{j=1}^n \int_{\gamma_0(p_0,q_0)}^{\gamma_j(p_j,q_j)} \frac{\mathfrak{s}_{kj}(\lambda)d\lambda}{\sqrt{\sum_{k=1}^n\alpha_k\mathfrak{s}_{kj}(\lambda)-U_j(\lambda)}}=\beta_k, \quad k=2,\dots,n,\end{gathered} \tag{II.10}$$

where points $\gamma_j(p_j,q_j)$ and $\gamma_0(p_0,q_0)$ be on the curve $\mathscr{C}_j$ of (II.6). Notice that bounded motion in this case will not be periodic in general, but only conditionally periodic.[11,17] If λ_j^0 and λ_j are the turning points determined by the conditions that functions $F_j(\lambda)$ [(II.6) and (II.9)] vanish, the periods of the motion w_{jk} are equal to

$$w_{kj}=\int_{\lambda_j^0}^{\lambda_j}\frac{\mathfrak{s}_{kj}(\lambda)d\lambda}{\sqrt{F_j(\lambda)}}. \tag{II.11}$$

Thus, Stäckel[17] showed that the orthogonal coordinates $\{q_j\}_{j=1}^n$ permit separation in the Hamilton–Jacobi equation (II.8) if the metric

$$ds^2=\sum_{j=1}^n g_{jj}(q_1,\dots,q_n)(dq^j)^2, \quad g_{jj}(q_1,\dots,q_n)\equiv g_j(q_1,\dots,q_n), \tag{II.12}$$

is in the Stäckel form

$$g_{jj}(q_1,\ldots,q_n)=\mathsf{H}_j^2(q_1,\ldots,q_n)=\frac{\det \mathbf{S}}{\mathbf{S}^{j1}}, \qquad \text{(II.13)}$$

where S^{j1} means the cofactor of $\mathfrak{s}_{j1}$ in matrix S [(II.3)]. Here g_{jj} is a diagonal metric and H_j are called the Lamé coefficients. For a modern approach to construction of the Lamé coefficients, see Refs. 13–15.

Henceforth, we shall restrict our attention to the uniform Stäckel systems, where all the potentials $U_j(q_j)=U(q_j)$ and curves $\mathscr{C}_j$ [(II.6)] are equal. Variables $\{s_k,w_k\}$ [(II.9) and (II.11)] on a single curve $\mathscr{C}$ are the action-angle variables for the uniform Stäckel systems. To construct the metric $g_j(q_1,\ldots,q_n)$ and the potentials $U(q_j)$ in an explicit form we shall identify periods w_k [(II.11)] with periods of the Abel differentials on a common hyperelliptic curve $\mathscr{C}$ [(II.5)] along the elements of a homology basis.[1,6] In this case definition of the separated variables $\{q_j\}$ [(II.10)] leads to the Jacobi inversion problem. In the next section we prove that the Stäckel matrix **S** [(II.3)] is completely defined by the derivative of the Abel–Jacobi map $\mathscr{U}$ on $\mathscr{C}$ at generic point (the so-called Brill–Noether matrix).

III. UNIFORM STÄCKEL SYSTEMS AND ALGEBRAIC CURVES

To begin with, let us briefly recall some necessary facts about the action-angle variables on the Jacobian $J(\mathscr{C})$.[1,6,4] The main ingredient of this construction is a universal configuration space, which is the moduli space[10] of all algebraic curves with fixed jets of local coordinates at a fixed number of punctures. This concept is closely related to the notion of the Baker–Akhiezer function on admissible curves[14] and to the theory of algebraic completely integrable systems.[18]

Let us consider a genus g Riemann surface $\mathscr{C}$ with N ordered punctures P_j and with two special Abelian integrals y and λ with poles of order at most $l=(l_j)_{j=1}^N$ and $m=(m_j)_{j=1}^N$ at the punctures. The universal configuration space $\mathscr{M}_g(l,m)$ can then be defined as a moduli space of $\mathscr{C}$ under certain constraints on the set of algebraic geometrical data.[14,4] In this case the space $\mathscr{M}_g(l,m)$ is a complex manifold with only orbifold singularities. To introduce the local coordinates on $\mathscr{M}_g(l,m)$ we cut apart the Riemann surface $\mathscr{C}$ along a homology basis A_i, B_j, $j=1,\ldots,g$, with canonical intersection matrix

$$A_i\circ A_j=B_i\circ B_j=0,\quad A_i\circ B_j=\delta_{ij}. \qquad \text{(III.1)}$$

By selecting cuts from P_1 to other P_j for each $2\leqslant j\geqslant N$ one gets a well-defined branch of the Abelian integrals y and λ. Among the complete set of local coordinates on $\mathscr{M}_g(l,m)$, the following moduli are distinguished:

$$s_j=\oint_{A_j} y\,d\lambda,\quad j=1,\ldots,g. \qquad \text{(III.2)}$$

The universal configuration space $\mathscr{M}_g(l,m)$ is a base space for a hierarchy of fibrations $\mathscr{C}^{(k)}(l,m)$ of particular interest to us. These are the fibrations whose fiber above each point of $\mathscr{M}_g(l,m)$ is the kth symmetric power $S^k(\mathscr{C})$ of $\mathscr{C}$. This fiber $\mathscr{C}^{(k)}(l,m)$ is the set of all effective divisors $D=\gamma_1+\cdots+\gamma_k$ (the γ_j's may not be mutually distinct) of deg k of $\mathscr{C}$, i.e., $\mathscr{C}^{(k)}(l,m)$ can be identified with the set of all unordered k-tuples $\{\gamma_1,\ldots,\gamma_k\}$, where the γ_j's are arbitrary elements of $\mathscr{C}$.

Let $\mathscr{D}$ be the open set in $\mathscr{M}_g(l,m)$, where the zero divisors of dy and $d\lambda$ do not intersect. Fixing all the local coordinates on $\mathscr{M}_g(l,m)$ except s_j [(III.2)], one can determine a smooth g-dimensional foliation of $\mathscr{D}$, independent of the choice we made to define the coordinates themselves.[4] Hereafter, by abuse of notation, one leaf of this foliation is denoted just by $\mathscr{M}$, and $\mathscr{C}^{(k)}$ means the above fibrations restricted to $\mathscr{M}$.

Let $dS=y\,d\lambda$ be a meromorphic one-form on $\mathscr{C}$ with the special Abelian integrals y and λ, which have fixed expansions near the punctures P_j.[4] It means that we have imposed certain constraints on the algebraic geometrical data (according to Ref. 14 we used admissible data). These constraints ensure the existence of a global system of action-angle variables and the presence of the corresponding symplectic form.[5] The fact that we impose some constraints provides us with additional properties of dS. Namely, generating one-form dS possesses the property

$$\frac{\partial dS}{\partial s_j}=\frac{\partial y d\lambda}{\partial s_j}=dw_j\,,\quad j=1,...,g, \tag{III.3}$$

where s_j are action coordinates of (III.2) on the moduli space $\mathscr{M}$ and differentials of the angle variables dw_j form some basis of holomorphic differentials (normally, even if the differential is holomorphic, its moduli derivative is not). Moreover, form dS gives rise to differentials spanning a whole space $\mathscr{H}_1(\mathscr{C})$ of holomorphic differentials. Hence, for any generic divisor $D=\gamma_1+\cdots\gamma_g$ on $\mathscr{C}$ the standard two-form on $\mathscr{C}^{(g)}$,

$$\Omega_g=\delta\left(\sum_{j=1}^{g} y(\gamma_j)d\lambda(\gamma_j)\right)=\sum_{j=1}^{g}\delta y(\gamma_j)\wedge d\lambda(\gamma_j)=\sum_{j=1}^{g} ds_j\wedge dw_j \tag{III.4}$$

is a desired holomorphic symplectic form Ω_g on $\mathscr{C}^{(g)}$. The set of variables $\{s_j,w_j\}_{j=1}^{g}$ is the complete set of action-angle variables on $J(\mathscr{C})$. These action-angle variables $\{s_j,w_j\}$ have been obtained by generalizing the definition of actions introduced for integrable systems on tori in the form of periods of holomorphic differentials dw_j along the elements of a homology basis in Refs. 1 and 6.

Now we turn to the uniform Stäckel systems. The corresponding algebraic curve (II.6) is a hyperelliptic curve given by an equation of the form

$$\mathscr{C}:\quad y^2=\prod_{i=1}^{2g+1}(\lambda-\lambda_i), \tag{III.5}$$

and puncture P is the point at infinity $\lambda=\infty$. Recall that the moduli λ_j of $\mathscr{C}$ are integrals of motion (II.6). Solution to the inverse Jacobi problem and associated Abel–Jacobi map on $\mathscr{C}$ relate a set of the action-angle variables and the separated variables.

Variables of separation $q_j(t)$ give the solution to the inverse Jacobi problem (II.10). The associated Abel–Jacobi map $\mathscr{U}$: Div $(\mathscr{C})\to J(\mathscr{C})$ is restricted to Lagrangian submanifold $\mathscr{C}^{(k)}$:

$$\mathscr{U}:\quad \mathscr{C}^{(k)}\to J(\mathscr{C}). \tag{III.6}$$

Note that whenever we discuss the Abel–Jacobi map, we shall tacitly assume that a base point γ_0 [(II.10)] on $\mathscr{C}$ has already been fixed in an appropriate position.

Suppose that point $D=\gamma_1+\cdots+\gamma_k$, $k\leqslant g$ belongs to $\mathscr{C}^{(k)}$. The differential of the Abel–Jacobi map (III.6) at the point D is a linear mapping from the tangent space $T_D(\mathscr{C}^{(g)})$ of $\mathscr{C}^{(g)}$ at the point D into the tangent space $T_{\mathscr{U}(D)}(J(\mathscr{C}))$ of $J(\mathscr{C})$ at the point $\mathscr{U}(D)$:

$$\mathscr{U}_D^*:\quad T_D(C^{(k)})\to T_{\mathscr{U}(D)}(J(\mathscr{C})).$$

Now suppose that D is a generic divisor, and z_j is a local coordinate on $\mathscr{C}$ near the point γ_j. Then $(z_1,\ldots,z_k)$ yields a local coordinate system near the point D in $\mathscr{C}^{(k)}$. Let dw_k $(k=1,...,g)$ be a basis for a space $\mathscr{H}_1(\mathscr{C})$ of holomorphic differentials on $\mathscr{C}$, and near γ_j

$$dw_k=\phi_{kj}(z_j)dz_j\,,$$

where $\phi_{kj}(z_j)$ is holomorphic. It follows that the Abel–Jacobi map $\mathscr{U}$ can be expressed near D as

$$\mathscr{U}(z_1,\ldots,z_k)=\left(\sum_{j=1}^{k}\int_{\gamma_0}^{z_j}\phi_{1j}(z_j)dz_j\,,\ldots,\sum_{j=1}^{k}\int_{\gamma_0}^{z_j}\phi_{gj}(z_j)dz_j\right).$$

Hence

$$\mathscr{U}_D^*=\begin{pmatrix}\phi_{11}(\gamma_1) & \cdots & \phi_{g1}(\gamma_1)\\ \vdots & \ddots & \vdots\\ \phi_{1k}(\gamma_k) & \cdots & \phi_{gk}(\gamma_k)\end{pmatrix} \tag{III.7}$$

is the so-called Brill–Noether matrix.

Theorem 2: Transpose the Brill–Noether matrix $\mathscr{U}_D^*$ on the genus $g \geqslant n$ hyperelliptic curve $\mathscr{C}$, which is the derivative of the Abel–Jacobi map $\mathscr{U}$ at generic divisor D ($\deg D = n$) is equal to the Stäckel matrix $\mathbf{S}$ for the uniform Stäckel system on $\mathbb{C}^{2n}$ with metric

$$g_{jj}(q_1,\ldots,q_n) = \frac{\det \mathbf{S}}{\mathbf{S}^{j1}}.$$

At generic point D, $\deg D = g$ matrix $\mathbf{S} = U_D^{*t}$ is a regular matrix satisfying the Stäckel theorem.

At $g > n$ we have to consider restriction of the Abel–Jacobi map (III.6) onto $\mathscr{C}^{(n)}$. In this case the symplectic form Ω_n on $\mathscr{C}^{(n)}$ is an appropriate projection of Ω_g [(III.4)] and $\mathscr{C}^{(n)}$ is a Lagrangian submanifold in the phase space $\mathbb{C}^{2n}$. The separated variables $\{p_j, q_j\}_{j=1}^n$ are constructed from the first $2n$ action-angle variables (III.9) only, and the action differential $dS = \Sigma_{j=1}^n p_j dq_j$ gives rise to an n-dimensional chart of the whole space $\mathscr{H}_1(\mathscr{C})$. The corresponding $n \times n$ Stäckel matrix is the left upper $n \times n$ block of the general matrix $\mathbf{S} = U_D^{*t}$ and, therefore, unless otherwise indicated, we assume $n = g$.

As an example, let us consider some basis for $\mathscr{H}_1(\mathscr{C})$, for instance,

$$dw_j = \frac{\lambda^{j-1}}{y(\lambda)}\, d\lambda, \quad j = 1,\ldots,g. \tag{III.8}$$

By choosing this basis we fix a basis of action-angle variables (III.2)–(III.4). To solve the Jacobi inversion problem (II.10) one obtains variables of separation

$$p_j = y(\gamma_j), \quad q_j = \lambda(\gamma_j), \quad j = 1,\ldots,g, \tag{III.9}$$

for a generic point $D = \gamma_1 + \cdots + \gamma_g$ on $\mathscr{C}$, which coincides with the divisor of simple poles of the corresponding Baker–Akhiezer function.[14] In the real case (when p_j and q_j are real), the separated variables q_j from (III.9) (so-called root variables) vary along cycles A_j [(III.1)] over basic cuts on $\mathscr{C}$ and, therefore, our problem is defined on a g-dimensional real torus. The holomorphic symplectic form Ω_g on $\mathscr{C}^{(g)}$ coincides with standard ones of (II.2) and a fiber $\mathscr{C}^{(g)}$ is a complex Lagrangian submanifold of the phase space $\mathbb{C}^{2g}$ [(II.7)]:

$$\mathscr{C}^{(n)} \equiv S^n(\mathscr{C}): \quad (\mathscr{C}(\lambda) \times \mathscr{C}(\mu) \times \cdots \times \mathscr{C}(\nu))/\sigma_n, \quad n \leqslant g, \tag{III.10}$$

where σ_n is the permutation group on n letters.

Recall that the derivative $\mathscr{U}_D^*$ bears a great resemblance to the usual Gauss mapping. The map $\mathscr{U}_D^*$ induces a canonical mapping from $\mathscr{C}$ into the $(g-1)$-dimensional projective space $\mathscr{C} \to \mathbb{P}^{g-1}$. On the other hand, the canonical mapping is defined by the derivative of the Abel–Jacobi map. For a hyperelliptic curve $\mathscr{C}$ of genus $g \geqslant 2$, the canonical map $\mathscr{C} \to \mathbb{P}^{g-1}$ is the composition of the double covering map $\mathscr{C} \to \mathbb{P}^1$, sending (y,λ) to λ, with the Veronese map $\mathbb{P}^1 \to \mathbb{P}^{g-1}$ given by a basis for the polynomial ring of degree $g-1$. With respect to the basis of $\mathscr{H}_1(\mathscr{C})$ in (III.8), the canonical map of $\mathscr{C}$ has an extremely simple expression:

$$(y,\lambda) \to \lambda \to [\lambda^{g-1}, \lambda^{g-2}, \ldots, \lambda, 1].$$

By using this map we introduce the $g \times g$ matrix

$$\mathbf{S}(\lambda,\mu,\ldots,\nu) = \begin{pmatrix} \lambda^{g-1} & \mu^{g-1} & \cdots & \nu^{g-1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda & \mu & \cdots & \nu \\ 1 & 1 & \cdots & 1 \end{pmatrix}, \tag{III.11}$$

determined on the Lagrangian submanifold (III.10). For a generic point $D = \gamma_1 + \cdots + \gamma_g$ in (III.9), the Stäckel matrix is equal to

$$\mathbf{S}(q_1, q_2, \ldots, q_g) = \mathbf{S}(\lambda, \mu, \ldots, \nu)|_{\lambda = q_1, \mu = q_2, \ldots, \nu = q_g}, \quad \mathfrak{s}_{kj}(q_j) = \lambda^{g-k}|_{\lambda = q_j}. \tag{III.12}$$

Recall that the diagonal metric g_{jj} is completely determined by the corresponding Stäckel matrix (II.13). Nevertheless, we introduce another equivalent definition of the metric. Substituting the Stäckel matrix (III.12) in the algebraic equation (II.3) one obtain,

$$\sum_{j=1}^{g} \mathfrak{s}_{kj}(q_j) g_{jj}(q_1,q_2,\ldots,q_g) = \delta_{k1} = \sum_{j=1}^{g} \operatorname{Res}|_{\lambda=q_j} \frac{\lambda^{k-1}}{e(\lambda)} = \frac{1}{2\pi i} \oint_C \frac{\lambda^{k-1}}{e(\lambda)}, \qquad \text{(III.13)}$$

where, by definition,

$$g_{jj}(q_1,q_2,\ldots,q_n) = \operatorname{Res}|_{\lambda=q_j} \frac{1}{e(\lambda)}.$$

Here we introduced function $e(\lambda)$, which has zeroes at the points q_j giving the solution of the inverse Jacobi problem.

In general, function $e(\lambda)$ with g zeroes, which are solutions of inverse Jacobi problem, is expressed in the Riemann theta-function

$$e(\lambda) = \theta(\mathcal{U}(\gamma_1,\ldots,\gamma_g) - \beta - K), \quad \mathcal{U}(\gamma_1,\ldots,\gamma_g) = \mathcal{U}(\gamma_1) + \cdots + \mathcal{U}(\gamma_g). \qquad \text{(III.14)}$$

Here K is a vector of the Riemann constants and $\beta = (\beta_1,\ldots,\beta_g) \in \mathbb{C}^g$ is a fixed vector.[3] The principal properties of the function $e(\lambda)$ in (III.14) are considered in Ref. 3.

Proposition 1: Function $e(\lambda)$ on $\mathcal{C}$ with g zeroes (p_j, q_j) giving solution to the Jacobi inversion problem is completely defined by the metric $g_{jj}(q_1,q_2,\ldots,q_n)$ [(III.13)] for a uniform Stäckel system.

We prove this proposition in the polynomial ring only. In this case

$$e(\lambda) = \prod_{k=1}^{g} (\lambda - q_k), \qquad \text{(III.15)}$$

and

$$g_{jj}(q_1,q_2,\ldots,q_g) = \operatorname{Res}|_{\lambda=q_j} e^{-1}(\lambda) = \frac{1}{\Pi_{j\neq k}^{g}(q_j - q_k)}. \qquad \text{(III.16)}$$

To prove (III.13) for this metric, it suffices to consider the following integral:

$$\frac{1}{2\pi i} \oint_C \frac{\lambda^k}{e(\lambda)} = \sum_{j=1}^{g} \operatorname{Res}|_{\lambda=q_j} \frac{\lambda^k}{e(\lambda)} = -\operatorname{Res}|_{\lambda=\infty} \frac{\lambda^k}{e(\lambda)} = \delta_{k,g-1}, \qquad \text{(III.17)}$$

where C encloses all q_j.

The function $e(\lambda)$ is defined on the universal configuration space, i.e., it is independent on the moduli λ_j of $\mathcal{C}$ (integrals of motion) and on a choice of the basis of holomorphic differentials in $\mathcal{H}_1(\mathcal{C})$. For instance, in the polynomial ring let us consider a set of the equivalent Stäckel matrices with the following entries (III.7):

$$\mathfrak{s}_{kj}(\lambda)|_{q_j} = \phi_{kj}(\lambda)|_{q_j}, \quad \phi_{kj}(\lambda) = \lambda^{g-k} + a_1^{(j)} \lambda^{g-k-1} + \cdots + a_{g-k}^{(j)}, \qquad \text{(III.18)}$$

where polynomials ϕ_{kj} form various bases for the polynomial ring of degree $g-1$. Substituting (III.18) in (III.13) and (III.17) one obtains at once universal solution $e(\lambda)$ of (III.15). Below we shall see that the Hamiltonian H [(II.1)] with the diagonal metric g_{jj} [(III.16)] is closely related to the distinguished puncture P at infinity $\lambda = \infty$ on the hyperelliptic curve $\mathcal{C}$ [(III.5)]. The different Stäckel matrices (III.18) correspond to the distinct sets of the integrals of motion in the involution for a single Hamiltonian H. The completeness and functional independence of these integrals directly follows from the completeness and independence of the basis elements (III.18) for a polynomial ring.

Finally, we look at other fibrations $\mathcal{C}^{(n)}$ at $n \neq g$. At $g > n$, to construct the metric $g_{jj}(q_1,\ldots,q_n)$ on $\mathcal{C}^{(n)}$, we expand the initial curve $\mathcal{C}$ [(III.5)] by

$$y^2=\prod_{i=1}^{2g+1}(\lambda-\lambda_i)=U_{2g+1}(\lambda)+\prod_{i=1}^{2n+1}(\lambda-\tilde{\lambda}_i),\quad n\leqslant g. \tag{III.19}$$

Here $U_{2g+1}(\lambda)$ is an at most $2g+1$ order polynomial, which is regarded as a potential in (II.6). The $n\times n$ Stäckel matrix and the corresponding function $e(\lambda)$ may be associated to the auxiliary genus n curve

$$\tilde{\mathscr{C}}:\quad \tilde{y}^2=\prod_{i=1}^{2n+1}(\lambda-\tilde{\lambda}_i). \tag{III.20}$$

The function $e(\lambda)$ is independent on the moduli of $\mathscr{C}$ [(III.19)] and, therefore, uniform potential U_{2g+1} in (III.19) has an arbitrary form and decomposition (III.19) determines the highest power of the polynomial $U(\lambda)$ only.

At $n>g$ the above holomorphic symplectic form Ω_g on the leaves $\mathscr{M}$ is degenerate. However, a nondegenerate form on $\mathscr{C}^{(n)}$ may be obtained by restricting $\mathscr{C}^{(n)}$ to the larger leaves $\tilde{\mathscr{M}}$ of the foliation.[4] The leaves $\tilde{\mathscr{M}}$ correspond to the level sets of all the local coordinates except to holomorphic s_j [(III.2)] and to some additional $(n-g)$ coordinates associated to meromorphic differentials $d\tilde{w}_j$ in (III.3) and (III.4). In fact, to construct the action-angle variables, we have to add several meromorphic differentials to holomorphic angle variables. Thus, at $n>g$ the symplectic two-form Ω_n on $\mathscr{C}^{(n)}$ is meromorphic.[4]

As an example, at $n=g+1$, we can add one local coordinate in the neighborhood of puncture P at infinity.[4] This additional coordinate occurs in the Stäckel matrix and in the metric in the following way:

$$\begin{gathered}\mathbf{S}^{(g+1)}(\lambda,\mu,\ldots,\nu)=q_0\mathbf{S}^{(g)}(\lambda,\mu,\ldots,\nu),\\ e(\lambda)=q_0\prod_{j=1}^{n}(\lambda-q_j),\quad g_{00}=\mathrm{Res}|_{\lambda=\infty}\frac{\lambda^{g-1}}{g(\lambda)}.\end{gathered} \tag{III.21}$$

At $n>g$ these systems with meromorphic form Ω_n possess several reductions of the additional meromorphic coordinates, for instance $q_0=\text{const}$ in (III.21).[19]

The above formulas are well adjusted for generalization. If the curve $\mathscr{C}$ in (III.5) is substituted by

$$\mathscr{C}:\quad y^2=F(\lambda)=\frac{P_l(\lambda)}{Q_m(\lambda)}=\frac{\Pi_{j=1}^{2g+1}(\lambda-\lambda_j)}{\Pi_{k=1}^{m}(\lambda-\delta_k)},\quad m\leqslant 2g+1, \tag{III.22}$$

where $\{\delta_k\}$ is a set of m arbitrary constant, one obtains

$$\mathfrak{s}_{kj}(\lambda)|_{\lambda=q_j}=\left.\frac{\phi_{kj}(\lambda)}{Q_m(\lambda)}\right|_{q_j},\quad e(\lambda)=\frac{\Pi_{j=1}^{g}(\lambda-q_j)}{Q_m(\lambda)}. \tag{III.23}$$

Note that the algebraic equation (III.13) is covariant with respect to the transformations

$$\mathbf{S}\to R^{-1}(\lambda)\mathbf{S},\quad e(\lambda)\to R^{-1}(\lambda)e(\lambda),$$

which leads to the general form of the metric

$$g_{jj}(q_1,q_2,\ldots,q_n)=\mathrm{Res}|_{\lambda=q_j}\left(\frac{Q_m(\lambda)R(\lambda)}{\Pi_{j=1}^{g}(\lambda-q_j)}\right) \tag{III.24}$$

associated to the curve $\mathscr{C}$. We shall use this freedom to consider the standard curvilinear coordinate systems.[11,12,20]

So, the hyperelliptic genus g curve $\mathscr{C}$ may be associated to a family of the uniform Stäckel systems on the phase space $\mathbb{C}^{2n}$ by using the Abel–Jacobi map $\mathscr{U}$, its differential $\mathscr{U}_D^*$, and their restrictions on $\mathscr{C}^{(n)}$. Diagonal metric $g_{jj}(q_1,q_2,\dots,q_n)$ [(II.13)] in the Hamiltonian (II.1) is completely defined by number of degrees of freedom *n*, and potential $U(\lambda)$ is at most $2g+1$ order arbitrary polynomial.

On the other hand, one fixed metric $g_{jj}(q_1,\dots,q_n)$ may be associated to an infinite set of the hyperelliptic curves $\mathscr{C}$. The corresponding Hamiltonian systems differ from each other by the power and by the form of polynomial potentials $U(\lambda)$ in (III.19). Among these systems we must distinguish systems on $\mathscr{C}^{(n)}$ at $n>g$ [(III.10)] for which the number of degrees of freedom *n* is more than the genus g of the associated curve $\mathscr{C}$. In this case the corresponding symplectic two-form on $\mathscr{C}$ is meromorphic.[4] In the next section, we shall identify these systems with the degenerate or superintegrable systems.[21] Recall that for the degenerate system the number of independent integrals of motion is more than the number of degrees of freedom.

IV. THE LAX REPRESENTATIONS

Let us recall that the key idea, which has started the modern age in the study of classical integrable systems, is to bring them into the Lax form. All the properties of the uniform Stäckel systems may be recovered from the properties of the Abel map. Nevertheless, now we want to obtain the Lax representations for all the uniform Stäckel systems associated to the hyperelliptic curve $\mathscr{C}(y,\lambda)$ in (III.5). We consider construction of the Lax representation as a necessary intermediate step to study quantum counterparts of the Stäckel systems.

In the simplest case the Lax matrices $L(\lambda)$ or $L(y)$ are defined as the matrix-valued functions on bare spectral curves F_λ, $\lambda\in F_\lambda$ [(I.1)] or F_y, $y\in F_y$, while the full spectral curve $\mathscr{C}(y,\lambda)$ is given by the Lax eigenvalue equations

$$\mathscr{C}:\quad \det(L(\lambda)-y)=0,\quad \det(L(y)-\lambda)=0. \tag{IV.1}$$

As a result, $\mathscr{C}$ arises as a ramified covering over the bare spectral curve F_λ or F_y.[22]

Until now a delicate question was how to construct the Lax matrices $L(\lambda)$ or $L(y)$ for a given integrable system. The one integrable system may be associated to the different curves and one curve $\mathscr{C}$ may be associated to the different mechanical integrable system on a common phase space. As an example, the *n*-particles Toda lattice can be equivalently formulated in terms of two different Lax representations[23] associated to the single hyperelliptic curve $\mathscr{C}$.

Here we consider the equation for a general algebro-geometric symplectic structure associated to the spectral curve $\mathscr{C}$ of the given Lax representation *L*,

$$\Omega_n=-\sum_\alpha \operatorname{Res}_{P\alpha}\frac{\langle \delta\psi^+\wedge\delta L\,\psi\rangle}{\langle\psi^+\psi\rangle}, \tag{IV.2}$$

proposed in Ref. 4. Here Ω_n is the restriction of the algebro-geometrical symplectic form (III.4) on $\mathscr{C}$ generated by two differentials dy and $d\lambda$ having poles at punctures P_α. Functions ψ and ψ^+ are the Baker–Akhiezer function on $\mathscr{C}$ and its dual function. If we fix some two-form Ω_n and the Baker–Akhiezer functions ψ, ψ^+ on a given curve $\mathscr{C}$, then one can attempt to recover the associated Lax matrix *L*.

For some particular Stäckel systems the 2×2 Lax matrices[20,24] and the corresponding vector Baker–Akhiezer function $\vec{\psi}$ associated to natural vector fields on the Jacobian of any hyperelliptic curve are known. On the other hand, we know the general scalar Baker–Akhiezer function ψ on *C* defined by its analytical properties on $\mathscr{C}$, which corresponds to geodesic systems with diagonal metric.[14]

Note here we have the vector Baker–Akhiezer function $\vec{\psi}$, which is the eigenfunction of the matrix *L* associated to the curve $\mathscr{C}$, and the scalar Baker–Akhiezer ψ, which is completely defined by analytical properties on the same curve $\mathscr{C}$.

For the uniform Stäckel systems let us identify the preassigned symplectic structure Ω [(II.2)] with the symplectic structure (III.4) defined on a hyperelliptic algebraic curve $\mathscr{C}$. Next we try to recover Lame matrix for a geodesic motion under the following additional assumptions:

(1) $L(\lambda)$ is a generic 2×2 matrix associated to a spectral hyperelliptic curve $\mathscr{C}$ of genus $g=[(n-1)/2]$.

(2) The associated vector Baker–Akhiezer function $\vec{\psi}$ has a constant normalization $\vec{\alpha}$:[25]

$$\langle\vec{\alpha},\vec{\psi}\rangle=\alpha_1\psi_1+\alpha_2\psi_2=1,\quad \vec{\alpha}=(0,1).$$

(3) The first component of $\vec{\psi}$ in (IV.2) is proportional to the unique Baker–Akhiezer function ψ on $\mathscr{C}$ with fixed analytical properties.[14]

At the first assumption n is a number of integrals of motion, which are moduli of C ($n=2g+1$) and, therefore, form Ω_n in (IV.2) is a restriction of meromorphic symplectic form Ω_g in (III.4) to the minimal n-dimensional leaf $\widetilde{\mathscr{M}}$ (Ref. 4) for integrable systems on $\mathbb{C}^{2n}$. The second assumption allows us to reduce the vector Baker–Akhiezer function to a scalar one. In this case, the solution of (IV.2) is completely defined by the function ψ on $\mathscr{C}$ only. At first we present this particular solution in terms of the function $e(\lambda)$ associated to the Abel map $\mathscr{U}$. We introduce the function $e(\lambda)$ and its time derivative

$$e(\lambda)=\frac{\Pi_{j=1}^{n}(\lambda-q_j)}{\Pi_{k=1}^{m}(\lambda-\delta_k)},\quad m\leqslant n,\quad e_x(\lambda)=\{H,e(\lambda)\},\tag{IV.3}$$

where $\{\delta_k\}$ is a set of m arbitrary constant and H is a Hamiltonian of the geodesic motion

$$H=\sum_{j=1}^{n} g_{jj}(q_1,\ldots,q_n)p_j^2,\quad g_{jj}(q_1,\ldots,q_n)=\mathrm{Res}|_{\lambda=q_j}\frac{1}{e(\lambda)}.\tag{IV.4}$$

Thus, in the Lax equation for a geodesic motion

$$L_x(\lambda)=\{H,L\}=[L,A],$$

matrices L and A are given by

$$L(\lambda)=\begin{pmatrix}-e_x/2 & e\\ -e_{xx}/2 & e_x/2\end{pmatrix}(\lambda)\equiv\begin{pmatrix}h & e\\ f & -h\end{pmatrix}(\lambda),$$
$$A(\lambda)=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}.\tag{IV.5}$$

The Hamiltonian H in (IV.4) is equal to a highest residue at the distinguished Weierstrass point on $\mathscr{C}$ at infinity $\lambda=\infty$,

$$H=-\mathrm{Res}|_{\lambda=\infty}\lambda^{n-m}\det L(\lambda),\tag{IV.6}$$

where the full spectral curve $\mathscr{C}$ is equal to

$$\mathscr{C}:\quad y^2=F(\lambda)\equiv\det L(\lambda),$$
$$F(\lambda)=-h^2(\lambda)-e(\lambda)f(\lambda)=\frac{e\cdot e_{xx}}{2}-\frac{e_x^2}{4}.\tag{IV.7}$$

By definition the zeroes of $e(\lambda)$ are separation variables and conjugated variables p_j are given by

$$p_j=h(\lambda)|_{\lambda=q_j},\quad h(\lambda)=-e_x/2=e(\lambda)\sum_{j=1}^{n}\frac{g_{jj}(q_1,\ldots,q_n)p_j}{\lambda-q_j}.\tag{IV.8}$$

In accordance with Ref. 25 pairs of separation variables (q_j,p_j) lie on the spectral curve $\mathscr{C}$:

$$y^2(\gamma_j)=p_j^2=h^2(\lambda)|_{\lambda=q_j}=-F(\lambda=q_j)=-F(\lambda)|_{\gamma_j}.$$

As usual, the rational function $F(\lambda)$ admits some different representations:

$$F(\lambda)=\frac{\Sigma_{j=1}^{n}I_j\lambda^{n-j}}{\Pi_{k=1}^{m}(\lambda-\delta_k)}=\sum_{k=1}^{m}\frac{J_k}{(\lambda-\delta_k)}+\sum_{k=m+1}^{n}J_k\lambda^{n-k-1}.\tag{IV.9}$$

Here $\{I_j\}_{j=1}^n$ and $\{J_k\}_{k=1}^n$ are two sets of independent integrals of motion in the involution. The first set of integrals $\{I_j\}_{j=1}^n$ in (IV.9) corresponds to the Stäckel matrix in (III.11). The set of equivalent Stäckel matrices (III.18) relate to another decomposition of a numerator of $F(\lambda)$ in (IV.9). The second set of integrals $\{J_k\}_{k=1}^n$ in (IV.9) is associated to an expansion near punctures $\{\delta_k,\infty\}$ on $\mathscr{C}$.

The spectral curve $\mathscr{C}$ in (IV.1) is a time-independent curve and, therefore,

$$\{H,F(\lambda)\}=0\Rightarrow\partial_x^3 e(\lambda)=e_{xxx}=0. \tag{IV.10}$$

Thus, in fact,[26] we consider the polynomial solutions $e(\lambda)=\Pi(\lambda-q_j(t))$ to the equation (IV.10) and describe the Hamiltonian dynamics of their zeroes $q_j(t)$ (recall that ∂_x means derivative by time).

Substituting function $e(\lambda)$ from (IV.3) and Hamiltonian H from (IV.4) into (IV.10) one obtains the equations in the metric $g_{jj}=\mathsf{H}_j^2$ from (II.12). If we introduce so-called rotation coefficients,

$$\beta_{ij}=\frac{\partial_i\mathsf{H}_j}{\mathsf{H}_i},\quad i\neq j, \tag{IV.11}$$

these equations may be reduced to the following equations:[13]

$$\begin{gathered}\partial_k\beta_{ij}=\beta_{ik}\beta_{kj},\quad i\neq j\neq k,\\ \partial_i\beta_{ij}+\partial_j\beta_{ji}+\sum_{m\neq i,j}\beta_{mi}\beta_{mj}=0,\quad i\neq j,\end{gathered} \tag{IV.12}$$

where the notation $i\neq j\neq k$ means that indices i,j,k are distinct.

Of course, these equations may be obtained without any Lax representation by using definition (II.13) of the metric, properties of the Abel-Jacobi map, and preassigned asymptotic behavior of $e(\lambda)$ at the distinguished point $\lambda=\infty$.

The equations (IV.12) are equivalent to the vanishing conditions of all *a priori* nontrivial components of the curvature tensor.[13–15] Therefore, using (IV.12) we conclude that the local Riemannian submanifold $(\mathscr{V}_n,g|_{\mathscr{V}})$ in (II.8) of the Riemannian manifold $(\mathbb{C}^n,g)$ is a flat manifold whose metric is diagonal with respect to the coordinates $\{q_j\}$. Imposing some additional restrictions on the space of solutions to (IV.2),[19] one could get the Bourlet type equations[15] related to another Riemannian manifold of constant curvature.

To construct more general solutions to (IV.2) associated to hyperelliptic curve $\mathscr{C}$ of higher genus we begin with the calculation of the Poisson bracket relations for the initial Lax matrix $L(\lambda)$. It allows us to identify the space of solutions to Eq. (IV.2) with the loop algebra $\mathscr{L}$(sl(2)) in fundamental representation[23] and then to use the representation theory of the underlying algebra sl(2).[27]

Theorem 3: The Poisson bracket relations for the matrix $L(\lambda)$ [(IV.5)] are closed into the following r-matrix algebra at $m\leqslant n$ only:

$$\{\overset{1}{L}(\lambda),\overset{2}{L}(\mu)\}=[r_{12}(\lambda,\mu),\overset{1}{L}(\lambda)]-[r_{21}(\lambda,\mu),\overset{2}{L}(\mu)]. \tag{IV.13}$$

Here the standard notations are introduced:

$$\begin{gathered}\overset{1}{L}(\lambda)=L(\lambda)\otimes I,\quad \overset{2}{L}(\mu)=I\otimes L(\mu),\\ r_{12}(\lambda,\mu)=\frac{\Pi}{\lambda-\mu},\quad r_{21}(\lambda,\mu)=\Pi r_{12}(\mu,\lambda)\Pi,\end{gathered} \tag{IV.14}$$

and Π is the permutation operator of auxiliary spaces.[23]

The Poisson bracket relations for the Lax matrix $L(\lambda)$ in (IV.5) are preassigned by the initial symplectic structure (III.4). It is necessary to calculate two brackets only:

$$\{e(\lambda),e(\mu)\}=0 \tag{IV.15}$$

and

$$\begin{aligned}\{h(\lambda),e(\mu)\}&=\left\{e(\lambda)\sum_{j=1}^{n}\frac{g_{jj}(q_1,\ldots,q_n)p_j}{\lambda-q_j},\frac{\Pi_{j=1}^{n}(\lambda-q_j)}{\Pi_{k=1}^{m}(\lambda-\delta_k)}\right\}\\&=-e(\lambda)e(\mu)\sum_{j=1}^{n}\frac{g_{jj}}{(\lambda-q_j)(\mu-q_j)}\\&=\frac{e(\lambda)e(\mu)}{\lambda-\mu}\sum_{j=1}^{n}\left(\frac{g_{jj}}{\lambda-q_j}-\frac{g_{jj}}{\mu-q_j}\right)=\frac{1}{\lambda-\mu}[e(\mu)-e(\lambda)],\end{aligned} \tag{IV.16}$$

where we used a standard decomposition of rational function

$$e^{-1}(\lambda)=\sum_{j=1}^{n}\frac{g_{jj}}{\lambda-q_j},\quad g_{jj}=\mathrm{Res}|_{\lambda=q_j}e^{-1}(\lambda).$$

Another Poisson bracket may be directly derived from these brackets and by definition of the entries of the Lax matrix $L(\lambda)$ in (IV.5) via derivative of the single function $e(\lambda)$

$$\begin{aligned}&\{h(\lambda),h(\mu)\}=0,\\&\{f(\lambda),e(\mu)\}=\partial_x\{h(\lambda),e(\mu)\}=\frac{2}{\lambda-\mu}[h(\lambda)-h(\mu)],\\&\{f(\lambda),h(\mu)\}=-\frac{1}{2}\partial_x^2\{h(\lambda),e(\mu)\}=\frac{1}{\lambda-\mu}[f(\lambda)-f(\mu)],\\&\{f(\lambda),f(\mu)\}=-\tfrac{1}{2}\partial_x^3\{h(\lambda),e(\mu)\}=0.\end{aligned} \tag{IV.17}$$

To derive the first bracket we have to combine second and first derivatives of the brackets (IV.15) and (IV.16), respectively. At the last bracket one substitutes the equation of motion (IV.10).

If, contrary to our geometric conventions, the order of polynomial $Q_m(\lambda)$ is more than an order of polynomial P_l in (III.22), i.e., if $m>n$ in the metric (IV.5), then the rational function $e(\lambda)$ admits another representation,

$$e^{-1}(\lambda)=\sum_{j=1}^{n}\frac{g_{jj}}{\lambda-q_j}+\xi(\lambda,q_1,\ldots,q_n),$$

where remainder $\xi(\lambda)$ is a certain polynomial. Substituting this function $e(\lambda)$ into (IV.15)–(IV.17) one obtains

$$\frac{\partial\xi(\lambda,q_1,\ldots,q_n)}{\partial\lambda}=0.$$

This constraint to remainder $\xi(\lambda,q_1,\ldots,q_n)$ directly follows from the symmetry of the last Poisson bracket in (IV.17).

The r-matrix algebra (IV.13) is the so-called linear case of the r-matrix algebras corresponding to integrable systems, which are modelled on coadjoint orbits of the Lie algebra sl(2). The r-matrix in (IV.14) is a standard rational r-matrix on $\mathscr{L}$(sl(2)).[28] The general form of the function $e(\lambda)$ in (III.14) leads to the elliptic and trigonometric r-matrices.[28,29]

Thus, for a geodesic motion (IV.4), the Lax representation (IV.5) with arbitrary poles $\{\delta_k\}_{k=1}^m$ in (IV.3) may be regarded as a generic point at the loop algebra $\mathscr{L}$(sl(2)) in fundamental representation after an appropriate completion.[28] To construct the Lax representation for a potential motion with the fixed metric $g_{jj}(q_1,\dots,q_n)$ in (IV.3), we can use the outer automorphism of the space of infinite-dimensional representations of sl(2) proposed in Ref. 27.

Applying this automorphism of the underlying algebra sl(2) directly to the Lax representation $L(\lambda)$ in (IV.5) on $\mathscr{L}$(sl(2)) we obtain a family of the new Lax pairs:

$$L'(\lambda)=L(\lambda)-\sigma_-\cdot[\phi(\lambda)^{-1}(\lambda)]_N,\quad \sigma_-=\begin{pmatrix}0&0\\1&0\end{pmatrix},$$
$$A'(\lambda)=A-\sigma_-\cdot[\phi(\lambda)e^{-2}(\lambda)]_N=\begin{pmatrix}0&1\\u_N(\lambda)&0\end{pmatrix}. \tag{IV.18}$$

Here $\phi(\lambda)$ is a function on spectral parameter and $[z]_N$ means restriction of z onto the ad_R^*-invariant Poisson subspace of the initial r-bracket.[27,29,30] For the rational r-matrix (IV.14) we can use the linear combinations of the following Taylor projections

$$[z]_N=\left[\sum_{k=-\infty}^{+\infty} z_k\lambda^k\right]_N\equiv\sum_{k=0}^{N} z_k\lambda^k, \tag{IV.19}$$

or the Laurent projections.[27,29]

The mappings (IV.18) from the representation of the loop algebra $\mathscr{L}$(sl(2)) to representations of the universal enveloping algebra $U(\mathscr{L})$ play the role of a dressing procedure, allowing us to construct the Lax matrices $L_N'(\lambda)$ for an infinite set of new integrable systems starting from the single known Lax matrix $L(\lambda)$ associated to one integrable model. This mapping preserves the metric $g_j(q_1,\dots,q_n)$ in the Hamiltonian (II.1), but changes the potential $U(q_j)$ and associated curve $\mathscr{C}$.

The new Lax matrix $L'(\lambda)$ in (IV.18) obeys the linear r-bracket (IV.13), where constant r_{ij}-matrices substituted by r_{ij}'-matrices depend on dynamical variables:[27,29]

$$r_{12}(\lambda,\mu)\to r_{12}'=r_{12}-\frac{([\phi(\lambda)e^{-2}(\lambda)]_N-[\phi(\mu)e^{-2}(\mu)]_N)}{(\lambda-\mu)}\cdot\sigma_-\otimes\sigma_-\,. \tag{IV.20}$$

We have to distinguish systems on $\mathscr{C}^{(n)}$ at $n>g$ [(III.10)] for which the number of degrees of freedom n is more than genus g of the associated curve $\mathscr{C}$. According to Ref. 4 the corresponding symplectic form is meromorphic. In this case the action differential $dS=y\,d\lambda$ gives rise to a whole space $\mathscr{H}_1(\mathscr{C})$ and, in addition, several meromorphic differentials on $\mathscr{C}$. We can identify these systems with the degenerate or superintegrable systems.[21]

Theorem 4: The complete set of noncommutative integrals of motion for the degenerate uniform Stäckel systems with meromorphic symplectic form Ω_g is determined by the generalized spectral surface

$$\mathscr{C}(y,\lambda,\mu):\quad \det(yI+\Pi L'(\lambda)\otimes L'(\mu))=0.$$

Here we used the outer product of the 2×2 Lax matrices $L'(\lambda)$ with $L'(\mu)$ and Π means 4×4 permutation matrix in $\mathbb{C}^2\times\mathbb{C}^2$. The equation of motion for the matrix $L(\lambda,\mu)=\Pi L'(\lambda)\otimes L'(\mu)$ is equal to

$$\frac{d}{dt}L(\lambda,\mu)=L(\lambda,\mu)A(\lambda,\mu)-\Pi A(\lambda,\mu)\Pi^{-1}L(\lambda,\mu),$$
$$A(\lambda,\mu)=A(\lambda)\otimes I+I\otimes A(\mu), \tag{IV.21}$$

where the matrix $A(\lambda)$ is a second Lax matrix and I is a unit matrix.

It is easy to derive from (IV.18) that $n>g$ iff $n\geqslant N$, where N is a highest power in the Taylor projection (IV.19). In this case the corresponding r-matrix (IV.20) preserves the simple pole at the puncture P at $\lambda=\infty$ and the associated second Lax matrix A' remains a constant in spectral sense $\partial A(\lambda)/\partial\lambda=0$ under the mapping (IV.18).

Thus, for the degenerate systems, $A(\lambda,\mu)=\Pi A(\lambda,\mu)\Pi^{-1}$, and Eq. (IV.21) takes the standard Lax form and it proves the theorem.

As usual, spectral curve $\mathscr{C}$ [(IV.1)] of $L'(\lambda)$ is a generating function of the involutive family of integrals of motion. Substituting functions $\phi(\lambda)=\lambda^n Q_m^{-1}(\lambda)U_N(\lambda)$ into $L'(\lambda)$ [(IV.18)] one obtains their spectral curve in the form

$$\mathscr{C}:\quad y^2=F'(\lambda)=\det L'(\lambda)=U_N(\lambda)+\frac{\sum_{j=1}^{n}I'_j\lambda^{n-j}}{\Pi_{k=1}^{m}(\lambda-\delta_k)},$$

where $\{I'_j\}$ are integrals of motion. It is a time-independent curve and, therefore,

$$\frac{dF'(\lambda)}{dt}=0\Rightarrow\left[\frac{1}{4}\,\partial_x^3+u_N(\lambda)\partial_x+\frac{1}{2}\,u_{N,x}(\lambda)\right]\cdot e(\lambda)=0. \tag{IV.22}$$

Of course, this equation may be obtained directly in the framework of symplectic geometry.[31] Let us briefly explain an origin of this equation in the theory of nonlinear equations that allows us to relate the scalar Baker–Akhiezer function ψ and function $e(\lambda)$.

The same algebro-geometrical symplectic form Ω_g [(III.4)] on hyperelliptic curve $\mathscr{C}$ [(III.5)] leads directly to a Hamiltonian structure for soliton equations.[2,4] As an example, we consider the Kortewege–de Vries (KdV) equation associated to hyperelliptic curve (III.5) with one puncture $P(N=1)$ at infinity $\lambda=\infty$ and at $l=1$, $m=2$.[4] Let us select one leaf of foliation corresponding to $d\lambda$ with all zero periods

$$\oint_C d\lambda=0$$

for an arbitrary cycle C. In this case, the Abelian integral $\lambda(P)$ is a single-valued function, with only a pole of second order at $P(m=2)$. For finite-gap solutions of the KdV equations, moduli s_j from (III.2) are canonically conjugated with respect to the Gardner–Faddeev–Zakharov symplectic structure to angle variables w_j (see Ref. 32 and references within). Thus, the uniform Stäckel systems have a common set of the action-angle variables with solutions of the KdV equations.

Starting with this set of variables we consider general algebro-geometric equation (IV.2) for nonlinear systems. Solution of the equation (IV.2) in a ring of second order differential operators with the standard Baker–Akhiezer function ψ on $\mathscr{C}$ [(III.5)] is well known.[2–4,9] The associated Shrödinger operator has the form

$$\mathscr{L}(\lambda)=-\frac{\partial^2}{\partial x^2}+u(x,t,\lambda), \tag{IV.23}$$

where λ is a parameter. In some simple cases, such as the KdV equation, this parameter λ appears as an eigenvalue and one ultimately equates the potential u with a solution of the nonlinear equation itself. Let us look for a solution $\mathscr{A}(\lambda)$ of the Lax system in the ring of differential operators

$$\mathscr{L}(\lambda)\psi=0,$$
$$\left(\frac{\partial\mathscr{L}(\lambda)}{\partial t}+[\mathscr{L},\mathscr{A}]\right)\psi=0 \tag{IV.24}$$

of the form

$$\mathscr{A}(\lambda)=e(\lambda)\,\frac{\partial}{\partial x}-\frac{1}{2}\,\frac{\partial e(\lambda)}{\partial x}. \tag{IV.25}$$

Substituting the given form of $\mathscr{A}$ into the Lax system, one obtains

$$\frac{\partial u}{\partial t}=-2\left[\frac{1}{4}\,\partial_x^3+u(\lambda)\partial_x+\frac{1}{2}\,u_x(\lambda)\right]\cdot e(\lambda). \tag{IV.26}$$

Equation (IV.26) is called the generating equation. For different choices of the form of $e(\lambda)$ and $u(\lambda)$, this procedure leads to different hierarchies of integrable equations, as an example to the KdV, nonlinear Shrödinger, and sine–Gordon hierarchies[8,9] or to the Dym hierarchy.[33] If we consider the solutions of the equation (IV.26) in the form of polynomial (III.15), then the roots q_j of $e(\lambda)$ define the root variables and, as a result, finite-gap solutions of the problem of the geodesic (see Refs. 8, 9, and 33 and references within).

Substitution of the special form of second operator $\mathscr{A}(\lambda)$ [(IV.25)] into the Lax system (IV.24) allows us to eliminate the Baker–Akhiezer function ψ and to construct a 2×2 Lax matrix in $e(\lambda)$. In fact, we replace the Baker–Akhiezer function ψ on $\mathscr{C}$ to the mutually disjoint function $e(\lambda)$ on $\mathscr{C}$, which has a transparent mechanical interpretation (III.15). Recall that the function $e(\lambda)$ is defined as a function with zeroes, which gives solution to the Jacobi inversion problem[3] on the hyperelliptic curve $\mathscr{C}$.

V. THE FLAT COORDINATES

According to Ref. 14, at $n=g$ the orthogonal curvilinear coordinates $\{p_j,q_j\}_{j=1}^{g}$ form a generic divisor of the simple poles of the Baker–Akhiezer function ψ, which is defined by their analytical properties on $\mathscr{C}$. The evaluation of ψ at a set of punctures on $\mathscr{C}$ determines the flat coordinates $\{p_j,x_j\}_{j=1}^{g}$ for the diagonal metric (II.12). It turns out that, up to constant factors, the Lamé coefficients H_j are equal to the leading terms of the expansion of the same function ψ at the punctures on $\mathscr{C}$.[14]

Next we reach the same conclusions by using the function $e(\lambda)$ and the corresponding Lax representation $L(\lambda)$ on $\mathscr{C}$. As usual, we reduce the study of algebraic geometrical data to the analysis of the associated geodesic motion. The crucial observation is that the equations of motion in coordinates $\{p_j,x_j\}_{j=1}^{n}$ on the Riemannian manifolds of constant curvature have a Newtonian form and the corresponding Hamiltonian has a natural form

$$\ddot{x}_j=\xi_j(x_1,\ldots,x_n),\quad H=\sum a_{ij}p_ip_j+V(x_1,\ldots,x_n),\quad a_{ij}\in\mathbb{C}, \tag{V.1}$$

where $\xi_j(x_1,\ldots,x_n)$ and potential $V(x_1,\ldots,x_n)$ are functions on coordinates only. Let us introduce new function $\mathscr{B}(\lambda)$,

$$\mathscr{B}^2(\lambda)=e(\lambda)=\mathsf{H}^{-2}(\lambda), \tag{V.2}$$

which is "inverse" to the Lamé coefficients H_j in (II.13). One immediately obtains

$$F(\lambda)=\mathscr{B}^3\mathscr{B}_{xx},\quad F'(\lambda)=\mathscr{B}^3\mathscr{B}_{xx}+\mathscr{B}^4\left[\frac{\phi(\lambda)}{\mathscr{B}^4}\right]_N. \tag{V.3}$$

These equations have the form of Newton's equations for the function $\mathscr{B}$:

$$\mathscr{B}_{xx}=F(\lambda)\mathscr{B}^{-3},\qquad \mathscr{B}_{xx}=F'(\lambda)\mathscr{B}^{-3}-\mathscr{B}\left[\frac{\phi(\lambda)}{\mathscr{B}^4}\right]_N, \tag{V.4}$$

To expand function $\mathscr{B}(\lambda)$ at the Laurent set,

$$\mathscr{B}=\sum_{j=0}^{N}x_{N-j}\lambda^j,$$

it is easy to prove that coefficients x_j obey the Newtonian equation of motion (V.4) [see (IV.10) and references within Ref. 26]. Here we reinterpret the coefficients of the bare curves $F(\lambda)$ and $F'(\lambda)$ in (V.4) not as functions on the phase space, but rather as integration constants. In variables x_j, mapping (IV.18) affects only the potential (x dependent) part of the integrals of motion I_k. The kinetic (momentum dependent) part of I_k remains unchanged. So, the dressing mapping (IV.18) allows us to get over from a free motion on $\mathbb{C}^{2n}$ to a potential motion on $\mathbb{C}^{2n}$.

As an example, from (V.4) we get some well known orthogonal curvilinear coordinates on $\mathbb{R}^n$ (see Refs. 11, 12, and 19):

elliptic coordinates $m=n$ in (III.23):

$$e(\lambda,q_1,\dots,q_n)=\frac{\Pi_{j=1}^n(\lambda-q_j)}{\Pi_{k=1}^n(\lambda-\delta_k)}=1+\sum_{k=1}^n \frac{x_k^2}{\lambda-\delta_k}=\mathscr{B}^2(\lambda,x_1,\dots,x_n),$$

$$\delta_1<x_1<\delta_2\cdots<\delta_n<x_n\,,$$

parabolic coordinates $m=n-1$ in (III.23):

$$e(\lambda,q_1,\dots,q_n)=\frac{\Pi_{j=1}^n(\lambda-q_j)}{\Pi_{k=1}^{n-1}(\lambda-\delta_k)}=\lambda-x_n+\sum_{k=1}^{n-1} \frac{x_k^2}{\lambda-\delta_k}=\mathscr{B}^2(\lambda,x_1,\dots,x_n),$$

$$x_1<\delta_1<x_2\cdots<\delta_{n-1}<x_n\,,$$

spherical coordinates $m=n+1$ in (III.21):

$$e(\lambda,q_0,\dots,q_n)=\frac{q_0\Pi_{j=1}^n(\lambda-q_j)}{\Pi_{k=1}^{n+1}(\lambda-\delta_k)}=\sum_{k=1}^{n+1} \frac{x_k^2}{\lambda-\delta_k}=\mathscr{B}^2(\lambda,x_1,\dots,x_{n+1}).$$

Curvilinear coordinates $\{q_j\}$ are zeroes of function $e(\lambda)$ and flat coordinates $\{x_j\}$ are residues of $e(\lambda)=\mathscr{B}^2(\lambda)$ at the punctures, in accordance with the Baker–Akhiezer function approach.[14,25]

All the separable orthogonal curvilinear coordinate systems in $\mathbb{R}^n$ may be obtained from these coordinate systems.[12,19,20] According to Ref. 34, all the possible separables in these coordinates potentials, which are polynomials or rational functions of the Cartesian coordinates x_j, belong to the set of uniform Stäckel systems. Thus, we can claim that every such mechanical system is embedded into a proposed scheme.

A. Quasi-point canonical transformations

In conclusion, we discuss another parametrizations of the function $e(\lambda)$. Of course, function $e(\lambda)$ admits various representations in different variables and we can use this freedom as an example to solve equations of motion.[26] The parametrization considered above describes the point canonical transformation only. Here we discuss an application of the Weierstrass reduction theory to construct other Cartesian coordinates on $\mathscr{C}$.

It is obvious that the Lax representation,

$$\dot{L}(\lambda)=[L(\lambda),A(\lambda)],$$

is covariant with respect to the transformation of the first Lax matrix,

$$L(\lambda)\to\phi(\lambda,\lambda_1,\dots,\lambda_k)L(\lambda),$$

with an arbitrary function $\phi(\lambda,\lambda_1,\dots,\lambda_k)$ on time-independent moduli $\{\lambda_j\}$ of $\mathscr{C}$ and on spectral parameter λ. However, this transformation drastically changes the Poisson bracket relations (IV.14) and parametrization of $L(\lambda)$ in the flat coordinates $\{p_j,x_j\}$. Hence, in addition to the flat coordinates $\{p_j,x_j\}_{j=1}^n$ considered above, the same function $e(\lambda,q_1,\dots,q_n)$ may be associated to another set of flat coordinates. Now we show that to introduce these new variables $\{p_j,x_j\}$ we can use various coverings of the initial curve $\mathscr{C}$, for example, the covering listed in Ref. 35.

Let us assume that the initial torus $J(\mathscr{C})=T^{2g}$ may be decomposed in a direct product of several tori:

$$T^{2g}=T^{2g_1}\times\cdots\times T^{2g_k},\quad \sum_{j=1}^k g_j=g. \tag{V.5}$$

The corresponding Riemann matrix has a block form $B=B_1\times B_2\ldots\times B_k$, where B_j are the $g_j\times g_j$ Riemann matrices and the corresponding Baker–Akhiezer function on $\mathscr{C}$ is factorized. In this case we can consider curve $\mathscr{C}$ as a K-sheeted covering of tori T^{2g_j}. Such covers are known to exist for any $K>1$ and for arbitrary tori.[35]

First of all, we can introduce the separated variables $\{q_j\}$ associated to a whole torus T^{2g}. For dynamics on $J(\mathscr{C})=T^{2g}$, the corresponding Lax representations $L(\lambda)$ in (IV.5) are 2×2 matrices.

Second, we can introduce another set of separated variables $\{\tilde{q}_j\}$ associated to each torus T^{2g_j} in (V.5). For dynamics splitting on several tori T^{2g_j}, the Lax representations have a block form

$$L(\lambda)=\begin{pmatrix} L_1 & & \\ & \ddots & \\ & & L_k \end{pmatrix}(\lambda), \tag{V.6}$$

where $L_j(\lambda)$ are the 2×2 matrices defined by functions $e_j(\lambda)$ on each torus T^{2g_j}.[29] Two sets of variables $\{q_j\}$ and $\{\tilde{q}_j\}$ are related by canonical transformation induced by the covering, which allows us to obtain a 2×2 Lax matrix instead of matrix (V.6). This means that we have two isomorphic integrable systems with different Lax representations and the corresponding canonical transformation is a quasi-point transformation.[36]

To illustrate this construction we take as an example several systems at $n=2$. Starting with a hyperelliptic curve $\mathscr{C}$ of genus $g=n=2$ we define variables (p_1,q_1) and (p_2,q_2) on the Lagrangian submanifold $\mathscr{C}^{(2)}$ from (III.9). The Jacobi inversion problem is the problem of finding these variables from the equations (II.10) with the Stäckel matrix S given by (III.12). This problem is solved by using the Kleinian $\wp$-functions, which are second logarithmic derivatives of the Kleinian σ-function:

$$\wp_{ij}=-\frac{\partial \ln \sigma(\beta_1,\beta_2)}{\partial\beta_i\partial\beta_j}, \quad \wp_{22}=q_1+q_2, \quad \wp_{12}=-q_1q_2$$

(for details, see Refs. 35 and 37). The function $e(\lambda)$ [(III.15)] on $\mathscr{C}$ with zeroes at the points q_1,q_2 is equal to

$$e(\lambda)=\lambda^2-\wp_{22}\lambda-\wp_{12}=(\lambda-q_1)(\lambda-q_2)=\lambda^2+2\lambda x_1+(2x_2+x_1^2) \tag{V.7}$$

or

$$e(\lambda)=\frac{(\lambda-q_1)(\lambda-q_2)}{(\lambda-\delta_1)(\lambda-\delta_2)}=1+\frac{x_1'^2}{\lambda-\delta_1}+\frac{x_2'^2}{\lambda-\delta_2}. \tag{V.8}$$

Here we used the freedom (III.23), and Cartesian coordinates $\{x_j\}$ or $\{x_j'\}$ are derived from the ‘‘inverse’’ Lamé function $\mathscr{B}(\lambda)$ in (V.2). Applying the outer additive automorphism of sl(2), we can construct the Lax matrices $L'(\lambda)$ for an infinite set of integrable mechanical systems with the following Hamiltonians:

$$\begin{aligned} H&=p_1p_2+V_N(x_1,x_2),\\ H&=p_1'^2+p_2'^2+V_N'(x_1',x_2'). \end{aligned} \tag{V.9}$$

Among them, we distinguish the Hénon–Heiles systems at $N=3$ and the systems with quartic potential at $N=4$. For these systems the genus of associated curve $\mathscr{C}$ is equal to the number of degrees of freedom $g=n=2$.

The function $e(\lambda)$ in (III.13) is independent on the moduli of $\mathscr{C}$ and, therefore, the above construction of the integrable systems (V.9) readily gets over on the reducible curve $\mathscr{C}$. To construct this reducible curve, let us take two tori $T_{1,2}^2$,

$$w_\pm^2=\xi(1-\xi)(1-k_\pm^2\xi), \tag{V.10}$$

with a Jacobi moduli

$$k_{\pm}^2=-\frac{(\sqrt{\alpha}\mp\sqrt{\beta})^2}{(1-\alpha)(1-\beta)}.$$

Making the rational order two $(K=2)$ change of variables

$$w_{\pm}=-\sqrt{(1-\alpha)(1-\beta)}\,\frac{\lambda\mp\sqrt{\alpha\beta}}{(\lambda-\alpha)^2(\lambda-\beta)^2}\,y,\quad \xi=\frac{(1-\alpha)(1-\beta)}{(\lambda-\alpha)(\lambda-\beta)}\,\lambda, \tag{V.11}$$

one obtains the hyperelliptic curve

$$\mathscr{C}:\quad y^2=\lambda(\lambda-1)(\lambda-\alpha)(\lambda-\beta)(\lambda-\alpha\beta), \tag{V.12}$$

which gives a two-sheeted covering of two tori $T_{1,2}^2$ [(V.10)]. It is a well-known example of the reduction of hyperelliptic integrals to elliptic ones by using the rational change of variables proposed by Legendre and generalized by Jacobi.[35]

The complex torus T^2 is isomorphic to the curve of genus $g=1$ given by the equation $w^2=f(\xi)$. In the above, we have presented the covering for the two odd curves (V.10) at $\deg(f)=2g+1=3$. All computations concerning the even curves at $\deg(f)=2g+2=4$ give similar covering,[35] so we do not present these formulas. The odd and even curves at $g=1$ are associated to the integrable cases of the Hénon–Hailes system and the system with quartic potential, respectively.

Next we can introduce two pairs of variables $(\tilde p_1,\tilde q_1)$ and $(\tilde p_2,\tilde q_2)$ on the tori $T_{1,2}^2$. Functions $e_{1,2}(\lambda)$ on $T_{1,2}^2$ are equal to

$$e_1(\lambda)=\lambda-\tilde q_1,\quad e_2(\lambda)=\lambda-\tilde q_2. \tag{V.13}$$

Variables $\{\tilde p_j,\tilde q_j\}$ are separated Cartesian coordinates for the integrable systems on $T_1^2\times T_2^2$ with the Hamiltonians

$$H_{3,4}=\tilde p_1^2+\tilde p_2^2+V_{3,4}(\tilde q_1)+V_{3,4}(\tilde q_2), \tag{V.14}$$

which is a sum of two one-dimensional Hamiltonians on $T_{1,2}^2$. The corresponding 4×4 Lax representation has a block form (V.6), whose blocks are determined by the functions $e_{1,2}(\lambda)$ of (V.13).

The covering (V.11) induces canonical transformation of variables $\{\tilde p_j,\tilde q_j\}$ to $\{p_j,q_j\}$.[37] These pairs of variables lie on the different curves $T_{1,2}^2$ and $\mathscr{C}$, respectively. The common moduli α and β of these curves are integrals of motion. On the orbit $\mathscr{O}$ ($\alpha=$const, $\beta=$const) this canonical transformation (V.11) becomes a point transformation. It is the so-called quasi-point transformation.[36]

By using the change of variables induced by covering (V.11) one can construct the 2×2 Lax matrix for the evolution (V.14) splitting on two tori. In variables $\{\tilde q_j\}$ the matrix $L(\lambda)$ is determined by the function

$$e(\lambda)=\frac{(\lambda-\alpha)(\lambda-\beta)}{(1-\alpha)(1-\beta)}\,\tilde e(\lambda),\quad \tilde e(\lambda)=(\lambda-\tilde q_1)(\lambda-\tilde q_2). \tag{V.15}$$

In fact, we add two additional zeroes α and β into the function $e(\lambda)$ of (III.13) on the reducible curve $\mathscr{C}$ in (V.12) and, therefore, change the parametrization of the Lax matrices in flat coordinates $\{p_j,x_j\}$.

In general, to introduce new flat coordinates, we can take any tori $T_{1,2}^{2g_j}$ of arbitrary genus $g_{1,2}>1$ and consider the two-dimensional evolution (V.14) splitting on these curves with an arbitrary one-dimensional potential $V_{2g_j+1}(q_j)$. The standard change of variables,

$$\tilde q_j=\frac{\tilde x_1\pm\tilde x_2}{2}\Rightarrow\tilde e(\lambda)=\lambda^2-\tilde x_1+\frac{\tilde x_1^2-\tilde x_2^2}{4}, \tag{V.16}$$

preserves the natural form of the Hamiltonians (V.14) for arbitrary potentials $V_{2g_j+1}(q_j)$. The equations of motion remain the Newtonian equations in these variables $\{\tilde p_j,\widetilde{x_j}\}$.

In the example (V.10) considered above both independent hyperelliptic integrals are reduced to elliptic ones by using a common substitution $\xi\to\lambda$ [(V.11)]. It relates to the existence of the second order automorphism of a hyperelliptic curve (V.12):[35]

$$\tau:\quad (\lambda,y)\to\left(\frac{\alpha\beta}{\lambda},\ \frac{y}{\lambda^3}\ \sqrt{\alpha^3\beta^3}\right). \tag{V.17}$$

It allows us to introduce another parametrization of the function $\tilde{e}(\lambda)$ in Cartesian coordinates, which preserves the natural form of the Hamiltonian. Namely, in addition to (V.16), we can use the following canonical transformation of variables $\{\tilde{p}_j,\tilde{q}_j\}$ to the Cartesian coordinates $\{\hat{p}_j,\hat{x}_j\}$

$$\tilde{e}(\lambda)=\lambda^2-\frac{Q_++Q_-}{\hat{x}_1}+\frac{(Q_+-Q_-)^2}{4\hat{x}_1}. \tag{V.18}$$

Here functions $Q_\pm(\hat{p}_j,\hat{x}_j)$ are the classical counterparts of the supercharges in two-dimensional supersymmetric quantum mechanics (SUSY)[36] with the following properties:

$$\{H,Q_\pm\}=\pm f(\hat{p}_j,\hat{x}_j)Q_\pm,\quad \{H,Q_+Q_-\}=0.$$

At $g=2$ [$N=3$ or $N=4$ in (V.14)] these functions $Q_\pm$ and f on variables $\{\tilde{p}_j,\tilde{q}_j\}$ or $\{\hat{p}_j,\hat{x}_j\}$ are listed in Ref. 36. Moduli α and β in (V.17) are integrals of motion; therefore, automorphism τ induces a second quasi-point transformation associated to torus $T_1^2\times T_2^2$ [(V.10)].

Two quasi-point transformations (V.11) and (V.18) for the physical variables $\{x_j'\}$, $\{\tilde{x}_j\}$ and $\{\hat{x}_j\}$ bind together all the integrable cases of the Hénon–Hailes system at $N=3$ and three integrable cases of the system with quartic potential at $N=4$. Of course, these systems have a common set of action-angle variables. Moreover, the same variables are associated to the Kowalewski top,[3] which is a supersymmetric quantum model as well.

Thus, several supersymmetric models are related to evolution splitting on the tori, when the number of degrees of freedom is equal to the genus $g=n$ of the associated covering curve $\mathscr{C}=T_1\times T_2$. It would be interesting to get a geometrical interpretation of these supersymmetric objects arising from finite-dimensional SUSY quantum mechanics.

VI. CONCLUSION

It is known that curves $y+y^{-1}=F(\lambda)$ together with the one-forms

$$dS^{(4)}=\lambda\ \frac{dy}{y},\quad dS^{(5)}=\log\lambda\ \frac{dy}{y},$$

are implied by integrable models of the Toda chain family (standard and relativistic models). The corresponding Lax representations are defined on the Poisson–Lie groups with quadratic r-matrix algebra. The corresponding mapping from the action-angle variables to separated variables has been proposed in Ref. 32.

On the other hand, we can consider the umbilic solutions of the KdV equation.[9,33] These systems are defined on a generalized Jacobi variety of the symmetric product of n logarithmic Riemannian surfaces in place of the Liouville tori. Nevertheless, it is possible to introduce variables that linearize the corresponding Hamiltonian flows. These systems may be interpreted as counterparts of the discrete-time Stäckel systems.

Both these sets of models are associated to the change of parametrization of the hyperelliptic curve from ''plane'' parametrization to ''annulus'' ones ($\lambda\to\log\lambda$). Another possible generalization relates to the interpretation of the parameter λ as a coordinate on elliptic curve.

For all these integrable models it would be interesting to estimate the possibility of application of the usual Stäckel approach. In this way we should consider mapping between action-angle variables and separated variables, and should study the differential of this map. In the presented paper there are the Jacobi inversion problem and the differential of the Abel-Jacobi map.

ACKNOWLEDGMENTS

This work was partially supported by RFBR Grant No. 98-0101075 and a GRACENAS grant.

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed. (Springer, New York, 1989).
[2] S. P. Novikov and A. P. Veselov, Sov. Math. Doklady **26**, 357 (1982); Proc. Steklov Inst. Math. **165**, 53 (1985).
[3] B. A. Dubrovin, Russ. Math. Surveys **36**, 11 (1981).
[4] I. M. Krichever and D. H. Phong, J. Diff. Geom. **45**, 349 (1997).
[5] V. N. Nekhoroshev, Trans. Moscow Math. Soc. **26**, 181 (1972); H. J. Duistermaat, Commun. Pure Appl. Math. **32**, 687 (1980); H. Knörer, J. Reine Angew. Math. **355**, 67 (1985).
[6] J. P. Francoise, Bull. Am. Math. Soc. **17**, 301 (1987).
[7] M. V. Berry, J. Phys. A **18**, 15 (1985).
[8] S. J. Alber, Commun. Pure Appl. Math. **34**, 259 (1981).
[9] M. S. Alber and J. E. Marsden, Commun. Math. Phys. **149**, 217 (1992).
[10] Yu. I. Manin, ''Frobenius manifolds, quantum cohomology and moduli spaces,'' Preprint Max-Plank-Institute für Mathematik (1996).
[11] L. P. Eisenhart, Ann. Math. **35**, 284 (1934).
[12] N. Ya. Vilenkin, Trans. Math. Mon. **22** (1968).
[13] V. E. Zakharov, ''Description of the *n*-orthogonal curvilinear coordinate systems and hamiltonian integrable systems of the hydrodynamic type. Part I. Integration of the Lame' equation,'' preprint (1996).
[14] I. M. Krichever, ''Algebraic-geometrical *n*-orthogonal curvilinear coordinate systems and solutions to the associativity equations,'' hep-th/9611158 (1996).
[15] A. V. Razumov and M. V. Saveliev, ''Riemannian manifolds with diagonal metric. The Lame' and Bourlet systems,'' solv-int/96612004 (1996).
[16] I. M. Krichever, Commun. Pure Appl. Math. **47**, 1 (1994).
[17] P. Stäckel, Comp. Rend. **116**, 485 (1893); **121**, 489 (1895).
[18] M. Adler and P. van Moerbeke, *Algebraic Completely Integrable Systems: A Systematic Approach*, Perspect. Math. (Academic, Boston, 1989).
[19] E. G. Kalnins, V. B. Kuznetsov, and W. Miller, Jr., J. Math. Phys. **35**, 1710 (1994).
[20] E. G. Kalnins, *Separation of Variables for Riemann Spaces of Constant Curvature*, Pitman Monographs and Surveys in Pure and Applied Mathematics (Longman, New York, 1986).
[21] A. V. Tsiganov, J. Phys. A **31**, 2075 (1998).
[22] I. M. Krichever, Funct. Anal. Appl. **14**, 45 (1980).
[23] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer, Berlin, 1987).
[24] D. Mumford, Tata Lectures on Theta (Birkhaüser, Boston, 1984).
[25] E. K. Sklyanin, Prog. Theor. Phys. Suppl. **118**, 35 (1995).
[26] F. Calogero and J. P. Francoise, J. Phys. A **30**, 211 (1997).
[27] A. V. Tsiganov, J. Math. Phys. **39**, 650 (1998).
[28] A. G. Reyman and M. A. Semenov-Tian Shansky, in *Dynamical Systems VII*, edited by V. I. Arnold and S. P. Novikov, Vol. EMS 16 (Springer, Berlin, 1993).
[29] P. P. Kulish, S. Rauch-Wojciechowski, and A. V. Tsiganov, J. Math. Phys. **37**, 3365 (1996).
[30] J. C. Eilbeck, V. Z. Enolskii, V. B. Kuznetsov, and A. V. Tsiganov, J. Phys. A **27**, 567 (1994).
[31] N. Ercolani, Proc. Commun. Appl. Math. **49**, 87 (1989).
[32] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevskii, *Soliton Theory: Inverse Scattering Method* (Nauka, Moscow, 1980).
[33] M. S. Alber, R. Camassa, D. D. Holm, and J. E. Marsden, Lett. Math. Phys. **32**, 137 (1995).
[34] E. G. Kalnins, S. Benenti, and W. Miller, Jr., J. Math. Phys. **38**, 2345 (1997).
[35] A. Krazer, *Lehrbuch der Thetafunktionen* (Teübner, Leipzig, 1903).
[36] S. Rauch-Wojciechowski and A. V. Tsiganov, J. Phys. A **29**, 7769 (1996).
[37] V. Z. Enolskii and M. Salerno, J. Phys. A **17**, L425 (1996).

Lax pairs for integrable lattice systems

R. S. Ward[a)]
Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, United Kingdom

(Received 19 August 1998; accepted for publication 4 September 1998)

This paper studies the structure of Lax pairs associated with integrable lattice systems (where space is a one-dimensional lattice, and time is continuous). It describes a procedure for generating examples of such systems, and emphasizes the features that are needed to obtain equations which are local on the spatial lattice. © *1999 American Institute of Physics.* [S0022-2488(99)02001-0]

I. INTRODUCTION

There has long been interest in integrable differential-difference equations (integrable lattice systems), especially since the discovery of the Toda lattice.[1] Such systems can have direct applications, for example, in condensed-matter physics; and also occur as spatially discrete versions of integrable partial differential equations.[2] Associated with each integrable lattice is a Lax pair, involving a matrix L that ''steps'' along the lattice, together with another matrix V that generates the time evolution. Our purpose in this paper is to investigate the structure of this Lax pair, and how it affects the nature of the associated integrable systems.

Throughout the paper, we work with functions $\varphi_n(t)$ which depend on time t, and on an integer variable n. Such a function will be written simply as φ, its dependence on t and n being understood; then φ_+ denotes $\varphi_{n+1}(t)$, and φ_- denotes $\varphi_{n-1}(t)$. The symbol Δ denotes the forward-difference operator, i.e., $\Delta\varphi=\varphi_+-\varphi$. For example, the Toda lattice equation (in first-order form) is

$$\frac{d}{dt}\varphi=\psi,\quad \frac{d}{dt}\psi=\exp\Delta\varphi-\exp\Delta\varphi_- . \tag{1}$$

We shall take the integer n to be unrestricted (i.e., the lattice is infinite). Our primary interest is in systems that are local, in the sense that the time derivative of a variable at site n equals some expression involving the variables at sites $n-1$, n and $n+1$: i.e., nearest neighbors only.

The point of view adopted here is that a lattice equation is integrable if it can be written as the consistency condition for a linear system (Lax pair) of a suitable type. This involves two $q\times q$ matrices L and V, the entries of which are functions of a ''spectral parameter'' λ, as well as of t and n. In what follows, we shall, for the sake of simplicity, restrict to the case $q=2$ (i.e., 2×2 matrices). The linear system is

$$\Psi_+=L\Psi,\qquad \frac{d}{dt}\Psi=V\Psi, \tag{2}$$

where Ψ is a column 2-vector (depending on λ, t, and n). The consistency condition for (2) is

$$\frac{d}{dt}L=V_+L-LV. \tag{3}$$

[a)]Electronic mail: Richard.Ward@durham.ac.uk

The crucial feature of (3) is that it specifies the evolution only of L, and not of V; so in order to get a meaningful equation, V has to be determined in terms of L. In the next section, we shall see how this happens. The subsequent sections illustrate how this structure can be used to generate integrable lattice systems. We shall see how known examples fit into this framework; and as a new example, construct a system that couples lattice versions of the Heisenberg ferromagnet and the derivative nonlinear Schrödinger equation.

II. HOW $L(\lambda)$ DETERMINES $V(\lambda)$

In order to analyze the structure of $L(\lambda)$ and $V(\lambda)$, one needs to impose some requirements on the way that they depend on λ. Let us assume that L, L^{-1}, and V are rational functions of λ, with poles at constant values of λ (that is, the location of the poles does not depend on t or n). This is not the only possibility: for example, there is the well-known case of the lattice Landau–Lifshitz equation,[3] which involves elliptic functions of λ. But we shall restrict to the rational case.

By making a Möbius transformation on λ, we may ensure that $V(\lambda)$ is finite at $\lambda=\infty$, i.e., that its poles occur at finite values of λ. Furthermore, since (3) is homogeneous in L, we have the freedom to multiply L by a scalar function of λ (not depending on t or n). We can use this freedom to ensure that L is a (matrix) polynomial in λ that is nonzero at each of the poles of V. Let p denote the degree of L as a polynomial in λ.

Equation (3) determines the evolution of each matrix coefficient in the polynomial $L(\lambda) = A\lambda^p + B\lambda^{p-1} + \cdots + D$; so at this stage it is a set of coupled equations for $q^2(p+1)$ functions (with $q=2$ in what follows). As was emphasized above, the matrix V has to be determined in terms of L, since (3) does not specify its evolution: let us now examine how this happens.

Assume for the time being that the poles of V are all simple. So V has the form

$$V(\lambda) = \sum_{\alpha=1}^{N} V^{(\alpha)}(\lambda-\lambda_\alpha)^{-1} + V^{(0)}, \tag{4}$$

where $V^{(0)}, V^{(1)}, \ldots, V^{(N)}$ are matrices, independent of λ. The general idea is that $V^{(0)}$ is determined by a choice of gauge, whereas each $V^{(\alpha)}$ for $1 \leqslant \alpha \leqslant N$ is determined by the residue of (3) at the pole $\lambda=\lambda_\alpha$. Note that Eqs. (2) and (3) are invariant under the gauge transformations,

$$\Psi \mapsto \Lambda\Psi, \qquad L \mapsto \Lambda_+ L \Lambda^{-1}, \qquad V \mapsto \Lambda V \Lambda^{-1} + \left(\frac{d}{dt}\Lambda\right)\Lambda^{-1}, \tag{5}$$

where Λ is a nonsingular 2×2 matrix depending on t and n (but not on λ). A choice of gauge involves the following steps.

(i) Choose a form for $L(\lambda)$ (an algebraic condition on the entries in the matrix L), such that a necessary condition for this form to be preserved under gauge transformations (5) is that $\Lambda_+ = \Lambda$.

(ii) Then choose any $V^{(0)}$ that is consistent with (i) and the evolution equation (3).

As an example to illustrate how this works, choose the coefficient of λ^p in $L(\lambda)$ (i.e., the leading term) to be the identity matrix. Then the remaining gauge freedom is (5), with Λ independent of n, as required. And the leading term in (3) gives $V_+^{(0)} = V^{(0)}$, so any choice $V^{(0)} = V^{(0)}(t)$ then fixes the gauge completely; to obtain an autonomous system of equations, one chooses $V^{(0)}$ to be a constant matrix. This is the most straightforward choice of gauge; for other gauges, $V^{(0)}$ will depend on the functions appearing in $L(\lambda)$, and this dependence is, in general, nonlocal, as we shall see below.

Consider, next, the pole at $\lambda=\lambda_\alpha$. Clearly the residue of the right-hand side of (3) at this pole must vanish, i.e.,

$$V_+^{(\alpha)} L^{(\alpha)} - L^{(\alpha)} V^{(\alpha)} = 0, \tag{6}$$

where $L^{(\alpha)}=L(\lambda_\alpha)$. The idea is that this constraint determines $V^{(\alpha)}$. Since $L^{(\alpha)}$ is a nonzero 2×2 matrix, there are two gauge-invariant possibilities: the rank of $L^{(\alpha)}$ could be either 2 or 1. In the next section we deal with the rank 2 case; thereafter, we concentrate on the rank 1 case.

III. RANK 2 CASE: NONLOCAL SYSTEMS

If $L^{(\alpha)}$ is invertible, then (6) is a difference equation,

$$V_+^{(\alpha)}=L^{(\alpha)}V^{(\alpha)}L^{(\alpha)-1},$$

which determines $V_n^{(\alpha)}$ in terms of (say) $L_{n-1}^{(\alpha)}$, $L_{n-2}^{(\alpha)},\ldots,$ and $V_{-\infty}^{(\alpha)}$. In other words, $V^{(\alpha)}$ is a nonlocal function of the entries in $L^{(\alpha)}$.

To obtain a simple example that illustrates this case, take $p=N=1$. Without loss of generality, we may set $\lambda_1=0$. Write $L=A\lambda+B$, where B is invertible; and choose a gauge by specifying

$$A=\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and $V^{(0)}=0$. So $V(\lambda)=V^{(1)}/\lambda$, where $V^{(1)}$ is the solution of the difference equation,

$$V_+^{(1)}=BV^{(1)}B^{-1}. \tag{7}$$

In general, the 2×2 matrix B contains four functions; let us effect a reduction to one function $y_n(t)$ by taking B to have the form

$$B=\begin{pmatrix} e^y & 0 \\ 0 & e^{-y} \end{pmatrix}. \tag{8}$$

In order for the reduction to be consistent, we need $V_{-\infty}^{(1)}$ to have the form

$$V_{-\infty}^{(1)}=\begin{pmatrix} 0 & b \\ c & 0 \end{pmatrix}$$

(plus a multiple of the identity, which has no eventual effect). The resulting system of equations for y is

$$\frac{d}{dt}\,y=b\,\exp(Y)-c\,\exp(-Y), \tag{9}$$

where $Y_n(t)$ is given by the nonlocal expression

$$Y_n=y_n+2\sum_{k=-\infty}^{n-1} y_k .$$

This equation can be transformed to a form that looks local, by writing $y_n=\phi_n-\phi_{n+1}$. In terms of $\phi_n(t)$, Eq. (9) becomes

$$\frac{d}{dt}\,(\phi_+-\phi)=c\,\exp(\phi_++\phi)-b\,\exp(-\phi_+-\phi). \tag{10}$$

If (for example) $b=0$, then (10) is a differential-difference version of the Liouville equation,

$$\phi_{tx}=\exp(2\phi). \tag{11}$$

To get (11) from (10), we interpret c as the lattice spacing, put $x=nc$, and take the continuum limit $c\to 0$. Similarly, putting $b=c$ in (10) gives a differential-difference version of the sine-Gordon equation.[4] But for these lattice Liouville and sine-Gordon systems, x and t (or n and t) do not represent space and time; rather they are characteristic (null) coordinates. In particular, one cannot specify arbitrary initial data at $t=0$. In these lattice equations, one of the characteristic coordinates has become discrete, while the other (namely t) remains continuous.

This example can be generalized in several directions, as follows. If one does not make the reduction (8), then one obtains an equation for the matrix B. Choosing the slightly different gauge $A=\mathbf{I}$ (the identity matrix) and $V^{(0)}=0$, leads to

$$\frac{d}{dt}B-\Delta V^{(1)}=0, \tag{12}$$

together with (7). Now (7) and (12) ensure that there exists a matrix $R_n(t)$ with

$$B=1+R_+R^{-1}, \quad V^{(1)}=-\left(\frac{d}{dt}R\right)R^{-1}.$$

Then (12) becomes

$$\frac{d}{dt}(R_+R^{-1})+\Delta\left[\left(\frac{d}{dt}R\right)R^{-1}\right]=0. \tag{13}$$

This is a differential-difference version of the principal chiral equation,

$$(R_xR^{-1})_t+(R_tR^{-1})_x=0, \tag{14}$$

in which, as before, one of the characteristic coordinates has become discrete.

This chiral equation generalizes, of course, to larger matrices $(q>2)$. Similarly, the Liouville and sine-Gordon examples generalize to differential-difference versions of other Toda field equations.

Finally, it might be noted that there are difference-difference versions of the principal chiral[5] and Toda field[6] equations in which both characteristic coordinates (here x and t) become discrete. Another, very general, example of this type is the Hirota bilinear difference equation.[7]

So systems with $\det L^{(\alpha)}\neq 0$ may be thought of as time-evolution equations that are nonlocal on the spatial lattice, or as equations where a characteristic coordinate (neither space nor time) has become discrete. To get local evolution equations, it is necessary for each $L^{(\alpha)}$ to have rank 1. From now on, we shall concentrate on this case.

IV. RANK 1 CASE: LOCAL SYSTEMS

Given that $L^{(\alpha)}$ has rank 1 the constraint (6) may be solved as follows. Write $K^{(\alpha)}=L_-^{(\alpha)}L^{(\alpha)}$. Assuming that $\operatorname{tr} K^{(\alpha)}$ is nonzero, the general solution of (6) is

$$V^{(\alpha)}=\frac{1}{\operatorname{tr} K^{(\alpha)}}[f^{(\alpha)}K^{(\alpha)}+g^{(\alpha)} \operatorname{adj} K^{(\alpha)}], \tag{15}$$

where $\operatorname{adj} K^{(\alpha)}$ denotes the adjoint matrix of $K^{(\alpha)}$, and where $f^{(\alpha)}$ and $g^{(\alpha)}$ are scalar functions, with $f_+^{(\alpha)}=f^{(\alpha)}$. So the constraint (6) does not determine $V^{(1)},\ldots,V^{(N)}$ uniquely: in particular, one has the arbitrary functions $g^{(\alpha)}$. However, there is a further constraint, namely that the condition $\det L^{(\alpha)}=0$ has to be preserved by the evolution (3). This gives equations on the $g^{(\alpha)}$, which are precisely that they are constant on the lattice, just as the $f^{(\alpha)}$ are: $g_+^{(\alpha)}=g^{(\alpha)}$. Then (15) can be rewritten as

$$V^{(\alpha)}=\frac{c^{(\alpha)}}{\operatorname{tr} K^{(\alpha)}}\, K^{(\alpha)}+d^{(\alpha)}\mathbf{I},$$

where $c^{(\alpha)}$ and $d^{(\alpha)}$ are functions of t only. It is clear that the $d^{(\alpha)}$ term will not contribute in the evolution equations, and so only the $c^{(\alpha)}$ remain; we may as well set $d^{(\alpha)}=0$, and take

$$V^{(\alpha)}=\frac{c^{(\alpha)}}{\operatorname{tr} K^{(\alpha)}}\, K^{(\alpha)}. \tag{16}$$

At this stage, the $c^{(\alpha)}$ could still be functions of time t; for simplicity, let us take them to be constants. One point to note about (16) is that $V^{(\alpha)}$ is local: $V_n^{(\alpha)}$ is expressed in terms of L_n and L_{n-1}.

So to obtain local evolution equations with 2×2 matrix Lax pairs, one first specifies the integer p [the degree of $L(\lambda)$]; the integer N appearing in (4) equals $2p$, since the λ_α all have to be roots of $\det L(\lambda)$. The matrices $V^{(\alpha)}$ for $1\leqslant\alpha\leqslant 2p$ are given by (16), and involve the $2p$ constants $c^{(\alpha)}$. Finally, there is the choice of gauge, which determines $V^{(0)}$. In general, $V^{(0)}$ turns out to be nonlocal, and special gauge choices are needed to ensure that it is local.

One can relate all this to the r-matrix description (see, for example, Ref. 3; and also Ref. 8, which addresses the construction of an r-matrix from a given Lax pair). Suppose one has an $L(\lambda)$, a Poisson bracket and an r-matrix such that the Fundamental Poisson Bracket Relations are satisfied. Suppose also that there exist $\lambda_1,\ldots,\lambda_N$ such that $\det L(\lambda_\alpha)=0$ for each α. Let $\tau(\lambda)$ be the trace of the monodromy matrix $\ldots L_2(\lambda)L_1(\lambda)L_0(\lambda)L_{-1}(\lambda)\ldots$ (which propagates from $n=-\infty$ to $n=+\infty$). Then

$$H=\sum_{\alpha=1}^{N} c^{(\alpha)} \log\, \tau(\lambda_\alpha)$$

is a local Hamiltonian, and the corresponding Hamiltonian equations are just (3); the constants $c^{(\alpha)}$ are the same as those appearing in (16). The problem from this point of view is to choose $L(\lambda)$, in a suitable gauge, such that a compatible r-matrix structure exists.

More generally, one wants $\tau(\lambda)$ to be conserved in time, for all λ—this then gives infinitely many conserved quantities. If one has a Lax pair (2) and boundary conditions which imply that $V_{+\infty}=V_{-\infty}$, then $\tau(\lambda)$ is indeed conserved. When $V(\lambda)$ depends locally on the fields, then then it is easy to ensure that this condition is met; if, on the other hand, V is nonlocal, then the conservation of $\tau(\lambda)$ is not guaranteed. This is one reason why locality is desirable, in the present context.

If $p=1$ [in other words, $L(\lambda)$ is linear in λ], then $\det L(\lambda)$ is a quadratic polynomial in λ, the roots of which are λ_1 and λ_2. For the time being, let us assume that these roots are distinct; and by translating λ set $\lambda_1=1$ and $\lambda_2=-1$. It follows that $L(\lambda)$ has the form

$$L(\lambda)=\tfrac{1}{2}(\lambda+1)L^{(1)}-\tfrac{1}{2}(\lambda-1)L^{(2)}, \tag{17}$$

where $L^{(1)}$ and $L^{(2)}$ are 2×2 matrices each having a zero determinant. So the entries in $L^{(\alpha)}$ involve six independent functions of t and n [in effect, the requirement that λ_1 and λ_2 be constant has reduced the number of functions in $L(\lambda)$ from eight to six]. The two $L^{(\alpha)}$ satisfy evolution equations obtained by expanding (3): these are

$$\frac{d}{dt}\, L^{(\alpha)}=V_+^{(0)}L^{(\alpha)}-L^{(\alpha)}V^{(0)}-\tfrac{1}{2}\Xi, \tag{18}$$

where $\Xi=V_+^{(1)}L^{(2)}-L^{(2)}V^{(1)}-V_+^{(2)}L^{(1)}+L^{(1)}V^{(2)}$. The gauge choice reduces the number of functions by four (since Λ contains four entries), and we end up with a system involving two functions. A number of examples of this type are described in the following section.

V. SOME $p=1$ EXAMPLES

In this section we exhibit some systems of the type described in the previous section. Such examples are simple to generate; but before doing so, we should ask when two lattice equations are to be regarded as being ''the same.'' More specifically, is there an appropriate equivalence relation on the set of all such systems? Certainly such an equivalence would include gauge transformations in which Λ was constant; and strictly local redefinitions of the functions appearing in L (i.e., the new functions at lattice site n depend on the old functions at site n only). However, it is customary to allow more general transformations than just these. A well-known case is that of the Toda lattice (1). If one replaces φ by

$$r=\Delta\varphi_{-}=\varphi-\varphi_{-}, \tag{19}$$

then (1) becomes

$$\frac{d^2}{dt^2}r=e^{r_+}-2e^{r}+e^{r_-}; \tag{20}$$

and this is regarded as simply another form of the Toda lattice equation.

But any equivalence relation that admitted (1) and (20) to the same class, would also have to allow the transformations $\varphi\mapsto\Delta^k\varphi$ for all integers k (negative as well as positive). If one allows such highly nonlocal transformations, then ones ends up with rather few equivalence classes; in fact, one might as well transform to action-angle variables, and say that the ''only'' integrable lattice is linear. Clearly this is inappropriate.

The point of this argument is to conclude that there is no useful equivalence relation on integrable systems of the type that we are considering [unless we insist that (1) and (20) are to be regarded as distinct]. This means that the task of listing such systems in a systematic way is not really well defined. The best that one can do is to exhibit examples, and indicate how they are related to one another.

Example (i). Choose a gauge such that $L^{(1)}-L^{(2)}=2\mathbf{I}$. This is the gauge that was mentioned as an example in Sec. II. As was remarked there, the gauge is then fixed completely by specifying some $V^{(0)}(t)$. The simplest choice is to set $V^{(0)}=0$. Note that $L^{(1)}$ and $L^{(2)}$ must have the form

$$L^{(1)}=\mathbf{I}+M,\quad L^{(2)}=-\mathbf{I}+M,$$

where M is trace-free and det $M=-1$. So we may write $M=\mathbf{f}\cdot\sigma$, where σ_1, σ_2, and σ_3 are the Pauli matrices, and $\mathbf{f}=\mathbf{f}_n(t)$ is a unit 3-vector. The dot denotes the usual three-dimensional Euclidean scalar product (and $\wedge$ below will denote the vector product). The evolution equation for $\mathbf{f}$, derived from (18), is then

$$\frac{d}{dt}\mathbf{f}=\Delta[(1+\mathbf{f}_-\cdot\mathbf{f})^{-1}(\mu\mathbf{f}_-+\mu\mathbf{f}+\nu\mathbf{f}_-\wedge\mathbf{f})], \tag{21}$$

where $\mu=\frac{1}{2}(c^{(1)}-c^{(2)})$ and $\nu=\frac{1}{2}i(c^{(1)}+c^{(2)})$ are constants. Equation (21), then, is an integrable equation for the unit-vector function $\mathbf{f}=\mathbf{f}_n(t)$. If the parameter ν is nonzero, it can be set to unity by scaling t; so the system effectively depends on the single parameter μ. The case $\mu=0$ is the ''Lattice Heisenberg Model,'' [3,9] so called because it has the equation of the Heisenberg ferromagnet as a continuum limit. Indeed, if we set $\nu=2/h^2$, $\mu=\hat{\mu}/h$, and let $h\to0$, then (21) becomes

$$\mathbf{f}_t=\hat{\mu}\mathbf{f}_x+\mathbf{f}\wedge\mathbf{f}_{xx}; \tag{22}$$

the Heisenberg model corresponds to $\hat{\mu}=0$. A slightly different choice of gauge, namely one in which $V^{(0)}$ is a nonzero constant matrix, yields a lattice nonlinear Schrödinger equation (different from the one in the next example). This is the lattice counterpart of the well-known gauge equivalence of the nonlinear Schrödinger and Heisenberg systems.

Example (ii). Choose a gauge such that

$$L^{(1)}=\begin{pmatrix}1 & 0\\ u & 0\end{pmatrix},\quad L^{(2)}=-\begin{pmatrix}0 & v\\ 0 & 1\end{pmatrix}, \tag{23}$$

where u and v are functions of t and n. The remaining gauge freedom has Λ being a diagonal matrix function of t only. Substituting (23) into the evolution equation (18) determines $V^{(0)}$ as

$$V^{(0)}=\frac{1}{2}\begin{pmatrix}-c^{(1)}vu_- & c^{(1)}v+c^{(2)}v_-\\ c^{(2)}u+c^{(1)}u_- & -c^{(2)}uv_-\end{pmatrix},$$

plus a diagonal matrix function of t, which by the residual gauge freedom may be set to zero. In addition, (18) gives the equations for u and v, namely

$$\begin{aligned}\frac{d}{dt}u&=\frac{1}{2}\,(c^{(2)}\Delta u+c^{(1)}\Delta u_-+c^{(1)}vuu_--c^{(2)}vuu_+),\\ \frac{d}{dt}v&=\frac{1}{2}\,(c^{(1)}\Delta v+c^{(2)}\Delta v_-+c^{(2)}uvv_--c^{(1)}uvv_+).\end{aligned} \tag{24}$$

This is exactly the Ablowitz–Ladik system;[2] their L operator is slightly different from the one presented here (it is, in effect, quadratic rather than linear in λ); but it is easily seen to be equivalent. In particular, if we choose $c^{(1)}=2i=-c^{(2)}$, and impose the (consistent) reduction $v=\pm u^*$, then (24) reduces to a lattice nonlinear Schrödinger equation,

$$i\,\frac{d}{dt}u=u_+-2u+u_-\mp uu^*(u_++u_-).$$

Example (iii). Here we choose a gauge such that $L^{(1)}$ is constant. Without loss of generality, we may take $L^{(1)}$ to be

$$L^{(1)}=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}. \tag{25}$$

The most general form for the matrix $L^{(2)}$ is

$$L^{(2)}=\begin{pmatrix}uv & uw\\ v & w\end{pmatrix}, \tag{26}$$

where u, v, and w are functions of n and t. The evolution equation (18) for $L^{(1)}$ implies that $V^{(0)}$ must have the form

$$V^{(0)}=-\frac{c^{(1)}}{2}\begin{pmatrix}0 & uw\\ v_- & 0\end{pmatrix}-\frac{c^{(2)}}{2(vu_-+w)}\begin{pmatrix}vu_- & wu_-\\ v & 0\end{pmatrix}+\begin{pmatrix}A & 0\\ 0 & D\end{pmatrix}, \tag{27}$$

where $\Delta A=0$ (i.e., A depends only on t). The residual gauge freedom, i.e., that which preserves (25), is (5) with

$$\Lambda=\begin{pmatrix}f(t) & 0\\ 0 & g(n,t)\end{pmatrix};$$

this has to be used to determine A and D, and to eliminate one of the three functions u, v, or w. In fact, the role of $f(t)$ is simply to fix A: let us choose $A=\frac{1}{2}c^{(1)}$. The remaining freedom now is

$$u \mapsto g_+^{-1} u, \qquad v \mapsto g_+ v, \qquad w \mapsto g^{-1} g_+ w. \tag{28}$$

Equation (18) for $L^{(2)}$ gives equations for u, v, and w, one form of which is

$$\begin{aligned}
&\frac{d}{d} \log v = D_+ + \frac{1}{2} c^{(1)}(wv_- / v - uv) + \frac{1}{2} c^{(2)} \frac{v_+ + vw_+}{v(uv_+ + w_+)}, \\
&\frac{d}{dt} \log w = \Delta\left(D - \frac{1}{2} c^{(2)} \frac{vu_-}{vu_- + w} \right), \\
&\frac{d}{dt}(uv) = \Delta\left(-\frac{1}{2} c^{(1)} uwv_- + \frac{1}{2} c^{(2)} \frac{vu_-}{vu_- + w} \right), \\
&\frac{d}{dt} \log(uw) = -D - \frac{1}{2} c^{(1)}(u_+ w_+ / u - uv) - \frac{1}{2} c^{(2)} \frac{uw + u_-}{u(vu_- + w)}.
\end{aligned} \tag{29}$$

(Any three of these equations implies the fourth.) We see from (28) that in order to remove the remaining gauge freedom, i.e., fix g (at least up to a function of t), there are three possibilities. Namely, we can specify either w, or uw, or v as a function of the gauge-invariant combination uv. This, in turn, will determine D, and hence $V^{(0)}$, on eliminating the relevant variable from (29). For example, specifying uw as a function of uv will give a local formula for D. Similarly, v can be specified as any function of uv. But if we impose $w = F(uv)$, then we need F to be an exponential in order to get a local expression for D. This illustrates the way in which some choices of gauge lead to a nonlocal expression for $V^{(0)}$.

As an example, let us take the gauge $w = \text{const}$. Choose $w = -1$, and write $u = -e^x$, $uv = y$. Then (29) reduces to the system

$$\begin{aligned}
&\frac{d}{dt} y = \frac{1}{2} c^{(1)} \Delta(y_- e^{\Delta x_-}) + \frac{1}{2} c^{(2)} \Delta\left(\frac{y}{y - \exp \Delta x_-} \right), \\
&\frac{d}{dt} x = \frac{1}{2} c^{(1)}(y + e^{\Delta x}) - \frac{c^{(2)}}{2(y - \exp \Delta x_-)}.
\end{aligned} \tag{30}$$

This is a version of the relativistic Toda lattice.[10–14]

Example (iv). The gauge choice,

$$L^{(1)} = \begin{pmatrix} -1 & e^x \\ 0 & 0 \end{pmatrix}, \quad L^{(2)} = \begin{pmatrix} e^y & 0 \\ -ke^{y-x} & 0 \end{pmatrix},$$

where k is a constant, gives

$$V^{(0)} = \frac{c^{(1)}}{2} \begin{pmatrix} 1 - ke^{y_- + \Delta x_-} & -e^x \\ -ke^{y_- - x_-} & 0 \end{pmatrix} + \frac{c^{(2)}}{2} \begin{pmatrix} 0 & e^{x-y} \\ ke^{-x_-} & 1 - ke^{-y + \Delta x_-} \end{pmatrix},$$

and again leads to the relativistic Toda system,[10,14,15] this time in the form

$$\begin{aligned}
&\frac{d}{dt} y = -\frac{1}{2} c^{(1)} k \Delta(e^{y_- + \Delta x_-}) - \frac{1}{2} c^{(2)} k \Delta(e^{-y + \Delta x_-}), \\
&\frac{d}{dt} x = -\frac{1}{2} c^{(1)} e^y (1 + ke^{\Delta x}) + \frac{1}{2} c^{(2)} e^{-y}(1 + ke^{\Delta x_-}).
\end{aligned} \tag{31}$$

The limit $k\rightarrow 0$ gives the Toda system (1); it is worth examining this in more detail. In order to get (1), one may replace the variables x and y by φ and ψ, where

$$e^{x}=-\sqrt{k}e^{\varphi},\quad e^{y}=-1+\sqrt{k}\psi,$$

and set $c^{(1)}=c^{(2)}=-1/\sqrt{k}$. Then the $k\rightarrow 0$ limit of (31) is indeed (1). But since $c^{(\alpha)}\rightarrow\infty$ in this limit, we need to reinterpret the associated Lax pair. The way to get a well-behaved limit is to replace λ by $2\lambda/\sqrt{k}$. The roots of det $L(\lambda)$ now occur at $\lambda=\pm\frac{1}{2}\sqrt{k}$, and so in the $k\rightarrow 0$ limit they coincide. In fact, when $k=0$ we have

$$L(\lambda)=-\begin{pmatrix} 1+\psi\lambda & e^{\varphi}\lambda \\ -e^{-\varphi}\lambda & 0 \end{pmatrix},$$

and det $L(\lambda)$ has a double zero at $\lambda=0$ (cf. Ref. 3). The corresponding expression for $V(\lambda)$ is obtained by taking the $k\rightarrow 0$ limit after first subtracting a constant multiple of the identity matrix: this yields

$$V(\lambda)=\begin{pmatrix} 0 & -e^{\varphi} \\ e^{-\varphi_{-}} & \lambda^{-1} \end{pmatrix}.$$

VI. SOME $p=2$ EXAMPLES

In the $p=2$ case, $L(\lambda)$ has the form $L(\lambda)=A\lambda^{2}+B\lambda+C$, where A, B, and C are 2×2 matrices; so to begin with, one has 12 functions of n and t. The requirement that the zeros λ_{α} of the quartic polynomial det $L(\lambda)$ be constant imposes four equations on these functions, and choice of gauge imposes a further four, so one is left with four independent functions. In other words, the generic system in this $p=2$, $q=2$ case is a system of coupled evolution equations for four lattice variables.

Reductions of such systems, so that fewer functions are involved, are, of course, possible. One example that has been known for some time is a lattice version of the sine-Gordon equation in which space is discrete and time continuous (by contrast with the version mentioned in Sec. III). Here the L operator has the form[16,3]

$$L=\begin{pmatrix} \lambda f(\varphi)e^{i\eta} & \frac{1}{4}h(e^{-i\varphi/2}-\lambda^{2}e^{i\varphi/2}) \\ \frac{1}{4}h(\lambda^{2}e^{-i\varphi/2}-e^{i\varphi/2}) & \lambda f(\varphi)e^{-i\eta} \end{pmatrix},$$

where φ and η are functions of n and t, h is a constant corresponding to the lattice spacing, and $f(\varphi)=(1+\frac{1}{8}h^{2}\cos\varphi)^{1/2}$. The resulting integrable lattice has sine-Gordon in ''laboratory coordinates'' as a continuum limit: if we replace n by $x=nh$ and let $h\rightarrow 0$, then φ satisfies $\varphi_{tt}-\varphi_{xx}+\sin\varphi=0$.

In order to obtain another example, let us choose a different gauge, namely $A=\mathbf{I}$, $V^{(0)}=0$. On B and C we impose the four constraints tr $B=0$, det $C=l$ const, $\mathrm{tr}(BC)=0$, and tr $C+\det B=-2k$ constant. It then follows that det $L(\lambda)=\lambda^{4}-2k\lambda^{2}+l$ has constant zeros. The matrices B and C now involve four independent functions, and their evolution is given by

$$\frac{d}{dt}B=\Delta Q,\qquad \frac{d}{dt}C=R_{+}C-CR, \tag{32}$$

where $Q=\Sigma_{\alpha}V^{(\alpha)}$, $R=-\Sigma_{\alpha}\lambda_{\alpha}^{-1}V^{(\alpha)}$, and the $V^{(\alpha)}$ are constructed as in (16). There are four parameters, namely the $c^{(\alpha)}$.

We can get an idea of what this system represents by looking at a continuum limit. To keep things simple, we assign particular values to the parameters, and this leads to the following continuum integrable system.

Let $\mathbf{B}$ and $\mathbf{C}$ be 3-vectors, functions of x and t, satisfying the constraints $\mathbf{B}\cdot\mathbf{C}=0$ and $\mathbf{C}\cdot\mathbf{C}=1$. Their time evolution is given by

$$\begin{aligned} \mathbf{B}_t &= (\mathbf{C}\wedge\mathbf{B}_x)_x - \{(\mathbf{C}\wedge\mathbf{C}_x\cdot\mathbf{B})\mathbf{C}\}_x + \tfrac{1}{2}\{(\mathbf{B}\cdot\mathbf{B})\mathbf{B}\}_x, \\ \mathbf{C}_t &= (\mathbf{C}\wedge\mathbf{C}_x)_x + (\mathbf{C}\wedge\mathbf{C}_x\cdot\mathbf{B})\mathbf{B}\wedge\mathbf{C} + \tfrac{1}{2}(\mathbf{B}\cdot\mathbf{B})\mathbf{C}_x. \end{aligned} \tag{33}$$

This is integrable: it has a Lax pair of the form $\Psi_x = U\Psi$, $\Psi_t = V\Psi$, where $U = -\frac{1}{2}i(\mathbf{B}\lambda^{-1} + \mathbf{C}\lambda^{-2})\cdot\sigma$ corresponds to $L(\lambda)$, and $V = \Sigma_{k=1}^{4} V_k \lambda^{-k}$ is a limiting version of the $V(\lambda)$ of the lattice system.

The equations (33) have two obvious reductions: if $\mathbf{B}=0$, then we are left with the Heisenberg ferromagnet equation for $\mathbf{C}$; while if $\mathbf{C}$ is a constant unit vector, then $\mathbf{B}$ satisfies the derivative nonlinear Schrödinger equation. So (33) may be viewed as a coupled Heisenberg–DNLS system; and (32) is a spatially discrete version of this coupled system.

VII. CONCLUDING REMARKS

It is clear that the examples given above provide only a very small sample of integrable lattice systems. One may envisage a classification involving the three integers q (the size of the matrices L and V), p [the degree of $L(\lambda)$], and r [the maximum order of the poles of $V(\lambda)$]. But in view of the remarks at the beginning of Sec. V, a complete classification would require a way of dealing with the problem of equivalence.

We conclude with a remark on higher values of r. One can take a given $L(\lambda)$ (thereby fixing q and p), and allow $r>1$: in other words, higher-order poles in $V(\lambda)$. This leads to hierarchies of lattice systems, of which the $r=1$ cases are the first members. So hierarcies also fit naturally into this framework.

[1] M. Toda, *Theory of Nonlinear Lattices* (Springer-Verlag, Berlin, 1988).
[2] M. J. Ablowitz and J. F. Ladik, J. Math. Phys. **17**, 1011 (1976).
[3] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer-Verlag, Berlin, 1987).
[4] S. J. Orfanidis, Phys. Rev. D **18**, 3822 (1978).
[5] M. Jimbo and T. Miwa, Publ. RIMS, Kyoto Univ. **19**, 943 (1983).
[6] R. S. Ward, Phys. Lett. A **199**, 45 (1995).
[7] R. Hirota, J. Phys. Soc. Jpn. **50**, 3785 (1981).
[8] H. W. Braden, J. Phys. A **30**, L485 (1997).
[9] N. Papanicolaou, J. Phys. A **20**, 3637 (1987).
[10] S. N. M. Ruijsenaars, Commun. Math. Phys. **133**, 217 (1990).
[11] I. Merola, O. Ragnisco and G. Z. Tu, Inverse Probl. **10**, 1315 (1994).
[12] B. Deconinck, Phys. Lett. A **223**, 45 (1996).
[13] Yu. B. Suris, J. Phys. A **30**, 1745 (1997).
[14] Yu. B. Suris, ''A collection of integrable systems of the Toda type in continuous and discrete time, with 2×2 Lax representations,'' preprint solv-int 9703004.
[15] Yu. B. Suris, Phys. Lett. A **145**, 113 (1990).
[16] A. G. Izergin and V. E. Korepin, Lett. Math. Phys. **5**, 199 (1981).

Solutions of Penrose's equation

E. N. Glass[a)]
Physics Department, University of Michigan, Ann Arbor, Michigan 48109

Jonathan Kress
School of Mathematics and Statistics, University of Sydney, NSW 2006, Sydney, Australia

(Received 21 July 1998; accepted for publication 22 September 1998)

The computational use of Killing potentials which satisfy Penrose's equation is discussed. Penrose's equation is presented as a conformal Killing–Yano equation and the class of possible solutions is analyzed. It is shown that solutions exist in space–times of Petrov type *O*, *D*, or *N*. In the particular case of the Kerr background, it is shown that there can be no Killing potential for the axial Killing vector. © *1999 American Institute of Physics.* [S0022-2488(99)03001-7]

I. INTRODUCTION

In a space–time which admits a Killing vector k^a it is straightforward to find its Killing potential. Killing potentials are real bivectors Q^{ab} whose divergence returns the Killing vector $(1/3)\nabla_b Q^{ab}=k^a$. Killing potentials attain physical importance when they are used in the Penrose–Goldberg (PG)[1] superpotential for computing conserved quantities such as mass and angular momentum. The PG superpotential is

$$U^{ab}_{\mathrm{PG}}=\sqrt{-g}\,\frac{1}{2}\,G^{ab}{}_{cd}Q^{cd}, \tag{1}$$

where $G^{ab}{}_{cd}=-{}^*R^{*ab}{}_{cd}$, the negative right and left dual of the Riemann tensor. When the Ricci tensor is zero then $G^{ab}{}_{cd}=C^{ab}{}_{cd}$, the Weyl tensor. If Q^{ab} satisfies Penrose's equation (4) then

$$\nabla_b U^{ab}_{\mathrm{PG}}=\sqrt{-g}G^{ab}k_b \tag{2}$$

for Einstein tensor G^{ab}. The current density

$$J^a=\sqrt{-g}G^{ab}k_b \tag{3}$$

is conserved independently of the left-hand side of Eq. (2). It is the PG superpotential that allows the Noether quantities to be computed by integrating over closed two-surfaces, which is Penrose's quasilocal construction.[2] If one views the Killing vector itself as a conserved current then its integral over a three-surface is identically equal to 1/3 the integral of its Killing potential over the bounding two-surface and no new information can be obtained.

The tensor version of Penrose's equation[3] is

$$P^{abc}:=\nabla^{(a}Q^{b)c}-\nabla^{(a}Q^{c)b}+g^{a[b}Q^{c]e}{}_{;e}=0. \tag{4}$$

With $j^a:=(1/3)\nabla_b Q^{*ab}$, and $k^a:=(1/3)\nabla_b Q^{ab}$, an equivalent equation[4] to $P^{abc}=0$ is

$$\nabla_c Q^{ab}=-2\,\delta_c^{[a}k^{b]}+2(\delta_c^{[a}j^{b]})^*. \tag{5}$$

If Q^{ab} is a solution of the Penrose equation then $k_{(b;c)}=-(1/2)Q_{a(b}R^a{}_{c)}$ with a similar relation connecting j^a and Q^{*ab}. For Ricci-flat space–times j^a and k^a are Killing vectors.

[a)]Permanent address: Physics Department, University of Windsor, Ontario, Canada.

For a particular space–time the number of independent Killing vectors is between zero and ten. Penrose[3] gave the complete solution to Eq. (4) in Minkowski space for ten real independent Q^{ab}.

This work discusses the existence of Killing potentials which satisfy Penrose's equation or equivalently the conformal Killing–Yano (CKY) equation for two-form $\mathcal{Q}$. The fact that such tensors only exist in space–times of Petrov type *D*, *N*, or *O* is discussed in Sec. III B and Appendices *C* and *D*.

In the Kerr background, it has previously been shown that there is no Killing potential for the axial Killing vector.[5] We show, in Sec. III C, how this can be anticipated from properties of the curvature and the fact that the axial Killing vector must vanish along the axis of symmetry.

We use both the abstract index notation familiar to relativists and some coordinate free notation for which we provide Appendix A as a reference. We use boldface characters for index free tensor notation, excepting differential forms which appear in calligraphic type. Appendix B describes some aspects of the Petrov classification in a way convenient for our purposes.

II. PREVIOUS RESULTS

An exact solution of the Penrose equation for Kerr's vacuum solution is given in Eq. (8). This solution was first used in the context of the PG superpotential construction in Ref. 6. The Kerr solution has two Killing vectors (KVs), stationary $k_{(t)}$ and axial $k_{(\varphi)}$, and the metric is

$$g^{\mathrm{Kerr}} = l\otimes n + n\otimes l - m\otimes\overline{m} - \overline{m}\otimes m, \tag{6}$$

where $\{l,n,m,\overline{m}\}$ is the Newman–Penrose principal null coframe, given in Boyer–Lindquist coordinates by

$$\begin{aligned} l &= dt - (\Sigma/\Delta)dr - a\sin^2\theta\, d\varphi, \\ n &= \frac{\Delta}{2\Sigma}[dt + (\Sigma/\Delta)dr - a\sin^2\theta\, d\varphi], \\ m &= \frac{1}{\sqrt{2}\overline{R}}[ia\sin\theta dt - \Sigma d\theta - i(r^2+a^2)\sin\theta\, d\varphi], \end{aligned} \tag{7}$$

where $R = r - ia\cos\theta$, $\Sigma = R\overline{R}$, and $\Delta = r^2 + a^2 - 2m_0 r$. The Killing potential for $k_{(t)}$ is the bivector $Q^{ab}_{(t)}$ obtained by raising the components of the two-form

$$\mathcal{Q}_{(t)} = -(R\mathcal{M} + \overline{R}\overline{\mathcal{M}}), \tag{8}$$

where $\mathcal{M} := l\wedge n - m\wedge\overline{m}$ is an anti-self-dual two-form, that is $*\mathcal{M} = -i\mathcal{M}$. We mention that $Q^{ab}_{(t)}$ is a global solution since the quasilocal PG mass, resulting from integration of the PG superpotential over two-surfaces of constant t and r, is *independent of choice of two-surface*

$$\oint_{S^2} U^{ab}_{\mathrm{PG}}\, dS_{ab} = -8\pi\, m_0 \tag{9}$$

for any r beyond the outer event horizon.

The next interesting result involves the axial Kerr symmetry. Goldberg[1] found asymptotic solutions of the Penrose equation for the Bondi–Sachs metric which includes the Kerr solution as a special case. But Glass[5] showed that the axial Killing potential could not be a solution of the Penrose equation at finite r.

The bivector $Q^{ab}_{(t)}$ generally has six independent components and so enough information to describe two Killing vectors. Since the Kerr solution has two KVs, can the dual of $Q^{ab}_{(t)}$ yield $k_{(\varphi)}$? Direct differentiation shows

$$\nabla_b Q^{*ab}_{(t)} = 0, \tag{10}$$

and so $Q^{ab}_{(t)}$ can only yield $k_{(t)}$. In fact $Q^{*ab}_{(t)}$ satisfies the Killing–Yano (KY) equation, which for an antisymmetric tensor A_{ab} can be written as

$$A_{a(b;c)} = 0. \tag{11}$$

This generalizes Killing's equation to antisymmetric tensors and can be further generalized to antisymmetric tensors of arbitrary valence. Modern usage reserves the name KY tensor for antisymmetric tensors. For the Kerr solution a symmetric tensor K_{ab} is constructed from the dual Killing potential by

$$K_{ab} = Q^{*e}_{(t)a} Q^{*(t)}_{eb} = 2\Sigma l_{(a} n_{b)} - r^2 g_{ab}. \tag{12}$$

This ''hidden'' symmetry of the Kerr solution was discovered by Carter[7] and later shown to be the ''square'' of a two-index Killing spinor,[8] or equivalently, the ''square'' of a Killing–Yano tensor. Though K_{ab} satisfies Eq. (11) it is symmetric and generally referred to as a Killing tensor.

Collinson[9] found that all vacuum metrics of Petrov type D, with the exception of Kinnersley's type $IIIB$, possess a KY tensor. He gave an explicit expression for both the KY tensor and its associated Killing tensor.

III. EXISTENCE OF SOLUTIONS

A. Conformal Killing–Yano tensors

Many of the arguments in this work depend on the conformal covariance of Penrose's equation. Penrose and Rindler[10] established the conformal covariance of its spinor form $\nabla_{A'}{}^{(A}\sigma^{BC)} = 0$ for a symmetric spinor σ^{BC}. The tensor version was previously discovered by Tachibana as the conformally covariant generalization of the KY equation.[11] In this paper it was written in the form

$$Q_{a(b;c)} = (1/3)[g_{bc} Q_{a;e}^{\;\;\;e} - g_{a(b} Q_{c)}{}^{e}{}_{;e}]. \tag{13}$$

In that same work Tachibana showed that in a Ricci-flat space, for Q_{ab} a CKY bivector satisfying Eq. (13), $(1/3)\nabla^b Q_{ab}$ is a Killing vector.

From Eq. (13) we can obtain an expression for $Q_{ab;c}$ by writing out the symmetrization brackets explicitly:

$$Q_{ab;c} = -Q_{ac;b} + \tfrac{2}{3} g_{bc} Q_{a;e}^{\;\;\;e} - \tfrac{1}{3} g_{ab} Q_{c;e}^{\;\;\;e} - \tfrac{1}{3} g_{ac} Q_{b;e}^{\;\;\;e}.$$

Now, since $Q_{ab;c}$ is antisymmetric in the first two indices, we have

$$\begin{aligned} 3Q_{ab;c} &= Q_{ab;c} + Q_{ab;c} - Q_{ba;c} \\ &= Q_{ab;c} - Q_{ac;b} + \tfrac{2}{3} g_{bc} Q_{a;e}^{\;\;\;e} - \tfrac{1}{3} g_{ab} Q_{c;e}^{\;\;\;e} - \tfrac{1}{3} g_{ac} Q_{b;e}^{\;\;\;e} \\ &\quad + Q_{bc;a} - \tfrac{2}{3} g_{ac} Q_{b;e}^{\;\;\;e} + \tfrac{1}{3} g_{ba} Q_{c;e}^{\;\;\;e} + \tfrac{1}{3} g_{bc} Q_{a;e}^{\;\;\;e} \end{aligned}$$

and so from (13) we can deduce that

$$3Q_{ab;c} = 3Q_{[ab;c]} - 2 g_{c[a} Q_{b]}{}^{e}{}_{;e}. \tag{14}$$

It is easily verified that given Eq. (14) we recover Eq. (13) and hence Eq. (14) is an alternative form of the CKY equation. Furthermore Penrose's Eq. (4) can easily be rewritten as Tachibana's Eq. (13) and so is another form of the CKY equation.

Since Q is an antisymmetric tensor, it is natural to discuss its properties in the language of differential forms. Equation (14) is manifestly antisymmetric in the first two indices, and so it is straightforward to verify that it is the abstract index equivalent of the CKY two-form equation of Benn *et al.*,[12]

$$3\nabla_z \mathcal{Q} = Z \lrcorner d\mathcal{Q} - Z^\flat \wedge \delta\mathcal{Q}, \quad \forall Z. \tag{15}$$

In this form, since $*$ commutes with ∇_Z, it is readily verified using the identities given in Appendix A, that whenever $\mathcal{Q}$ is a CKY two-form so is $*\mathcal{Q}$. Thus any solution to the CKY equation can be decomposed into self-dual and anti-self-dual CKY two-forms.

B. Existence of CKY two-forms

On a flat background the CKY equation has many solutions, while, as will be explained, in a more general space–time the curvature imposes tight consistency conditions and there can be at most two independent solutions, one self-dual and one anti-self-dual with respect to the Hodge star. This result appears to be closely tied to the four-dimensional nature of space–time and the properties of these solutions are almost universally discussed in their spinor form, where the utility of the two-component spinor formalism simplifies the calculations. A detailed discussion of this can be found in spinor form in Ref. 12 or in terms of differential forms in Ref. 13.

Since any CKY two-form can be decomposed into self-dual and anti-self-dual parts that are themselves CKY two-forms, in discussing their existence, it is sufficient to consider only two-forms of definite Hodge-duality.

In order to understand how the curvature of the underlying space–time restricts the solutions to Eq. (15) two steps are required. First, it can be shown directly from the CKY two-form equation that the real eigenvectors of (anti-) self-dual CKY two-forms are shear-free and hence principal null directions of the conformal tensor. Second, by differentiating Eq. (15) an integrability condition can be obtained that restricts the Petrov type by showing these eigenvectors to be *repeated* principal null directions.

In the case of non-null self-dual two-forms, Dietz and Rüdiger[14] used spinor methods to obtain both of these results for a scaling covariant generalization of Eq. (15). It was later shown, again using spinor methods, that similar results can be obtained for the null case.[12]

An outline of these results in the notation of differential forms is given in Appendices C and D. It is shown that apart from conformally flat space–times, non-null (anti-) self-dual CKY two-forms can only exist in space–times of Petrov type *D*, while null (anti-) self-dual CKY two-forms require a background space–time of Petrov type *N*.

C. The divergence of a CKY two-form

In order to apply the PG superpotential method using a given CKY two-form $\mathcal{Q}$, its divergence (coderivative) $\delta\mathcal{Q}$ must be dual to a Killing vector. Tachibana showed that this was always the case in a Ricci flat background[11] (the result also holds for the slightly more general case of an Einstein space–time).

In the Kerr background, there are two independent Killing vectors and two independent CKY two-forms (one of each Hodge-duality). However the divergence of either of these CKY two-forms is proportional to the timelike Killing vector, leaving the axial KV without a Killing potential. This allows a divergence free linear combination of the self-dual and anti-self-dual CKY two-forms to be found. The Hodge-dual of this two-form is known as a Killing–Yano two-form and satisfies the Killing–Yano equation (11), which can be written in a similar fashion to Eq. (15) as

$$3\nabla_X \mathcal{Q} = X \lrcorner d\mathcal{Q}. \tag{16}$$

However, this leaves open the question of why it is that the timelike rather than the axial KV possesses a Killing potential? To answer this question, we note that the axial Killing vector must vanish along the symmetry axis and we show that a Killing vector obtained as the divergence of a CKY two-form must be nowhere vanishing.

First consider a non-null anti-self-dual CKY two-form $\mathcal{Q}^-$. From Eq. (15) we can write $d(\mathcal{Q}^{-2})$ in terms of $\mathcal{Q}^-$ and $\delta\mathcal{Q}^-$:

$$d(\mathcal{Q}^{-2}) = \tfrac{4}{3}(\delta\mathcal{Q}^-)^\sharp \lrcorner \mathcal{Q}^-,$$

which after contracting with $\mathcal{Q}^-$ leads to

$$\delta\mathcal{Q}^- = -\tfrac{3}{2}(d(\mathcal{Q}^{-2}))^\sharp \lrcorner \mathcal{Q}^-.$$

Hence $\delta\mathcal{Q}^-$ vanishes if and only if $d(\mathcal{Q}^{-2})$ vanishes.

In a vacuum type D background we can deduce that $\mathcal{Q}^{-2}$ is a constant multiple of $\Psi_2{}^{-2/3}$ from the fact that $\mathcal{Q}^-$ is an eigen-two-form of $\mathbf{C}$ and both $(\mathcal{Q}^{-2})^{-3/2}\mathcal{Q}^-$ and $\mathbf{C}\mathcal{Q}^-$ are Maxwell fields. Hence if $\mathcal{Q}^-$ vanishes, then so does Ψ_2 and the background becomes conformally flat.

Further, it can be deduced from the Bianchi identities that for a type D vacuum space–time, the gradient of Ψ_2 vanishes if and only if the Ψ_2 itself vanishes. [In the Newman–Penrose (NP) formalism, using a principal null tetrad, the vacuum type D condition implies that the only nonzero curvature component is Ψ_2 and $\kappa=\sigma=\nu=\lambda=0$. Then, imposing $\nabla_{X_a}\Psi_2=0$, the Bianchi identities lead to either $\rho=\mu=\tau=\pi=0$ or $\Psi_2=0$. If we assume the former, then the NP equations for the derivatives of the spin coefficients immediately force the conclusion that Ψ_2 vanishes.] We therefore conclude that $\mathcal{Q}^{-2}$ is nowhere constant and hence $\delta\mathcal{Q}^-$ is nowhere vanishing and Kerr's axial Killing vector *cannot* have a Killing potential.

IV. SUMMARY

We have shown here that Penrose's equation for Killing potentials is equivalent to the conformal Killing–Yano equation for two-forms. With no appeal to Ricci-flatness existence of solutions was proven for space–times of Petrov type D, N or O. It was further shown, for type D vacuum backgrounds possessing a Killing–Yano two-form, that Killing vectors with zeros cannot have Killing potentials.

ACKNOWLEDGMENTS

E.N. Glass was partially supported by an NSERC of Canada grant and thanks the Physics Department of the University of Michigan and Professor Jean Krisch for their hospitality. Jonathan Kress began this work at the University of Newcastle.

APPENDIX A: DIFFERENTIAL FORMS

We denote a basis for vector fields by $\{X_a\}$. The natural dual of this we denote by $\{e^a\}$, a basis for covector or one-form fields. A coordinate basis is $X_a=\partial/\partial x^a$ and $e^a=dx^a$. The metric gives a natural bijection between vector and one-form fields, which we denote by $^\sharp$ and $^\flat$; $X^\flat$ is the one-form metric dual to the vector X and $\alpha^\sharp$ is the vector field metric dual to the one-form α.

The one-forms, along with the wedge product $\wedge$, generate the algebra of differential forms. The wedge product is antisymmetric and so the differential forms of degree p can be thought of as the subset of covariant tensors of valence p that are antisymmetric in their arguments. If α and β are one-forms with components $\alpha_a=\alpha(X_a)$ and $\beta_a=\beta(X_a)$, then

$$\alpha\wedge\beta=\alpha_{[a}\beta_{a]}e^a\otimes e^b=\alpha_a\beta_b e^a\wedge e^b. \tag{A1}$$

A vector can be contracted with the p-form $\mathcal{P}$ to give a $(p-1)$-form $X\lrcorner\mathcal{P}$ so that

$$(X\lrcorner\mathcal{P})(X_{a_1},X_{a_2},\ldots,X_{a_{p-1}})=p\mathcal{P}(X,X_{a_1},X_{a_2},\ldots,X_{a_{p-1}}),$$

and so the components of a p-form can be expressed using the hook as

$$\mathcal{P}_{ab\cdots c}=\mathcal{P}(X_a,X_b,\ldots,X_c)=\frac{1}{p!}X_c\lrcorner\cdots\lrcorner X_b\lrcorner X_a\lrcorner\mathcal{P}.$$

We can define an inner product between any pair of two-forms:

$$\mathcal{P}\cdot\mathcal{Q}=\tfrac{1}{2}X_a\lrcorner X_b\lrcorner\mathcal{P}X^a\lrcorner X^b\lrcorner\mathcal{Q}=2\mathcal{P}_{ab}\mathcal{Q}^{ab}.$$

For $\mathcal{P}\cdot\mathcal{P}$ we write $\mathcal{P}^2$.

The metric defines a natural map from p-forms to $(n-p)$-forms called the Hodge star. In four dimensions, this maps two-forms to two-forms, and is defined so that

$$\mathcal{P}\wedge*\mathcal{Q}=(\mathcal{P}\cdot\mathcal{Q})*1,$$

where $*1$ is the volume four-form. For a Lorentzian metric, this map squares to -1 and so has eigenvalues, $\pm i$. Elements of the eigenspace corresponding to $(-i)+i$ are called (anti-) self-dual two-forms. Any two-form can be decomposed into self-dual and anti-self-dual parts

$$\mathcal{P}=\mathcal{P}^+ + \mathcal{P}^- \quad \text{where} \quad *\mathcal{P}^{\pm}=\pm i\mathcal{P}.$$

The Hodge star relates the hook and wedge operations by

$$X\lrcorner*\mathcal{P}=*(\mathcal{P}\wedge X^\flat). \tag{A2}$$

The two-form commutator is given by

$$[\mathcal{P},\mathcal{Q}]=-2X_a\lrcorner\mathcal{P}\wedge X^a\lrcorner\mathcal{Q} \tag{A3}$$

for two-forms $\mathcal{P}$ and $\mathcal{Q}$. The Lie algebra of two-forms under commutation is the Lie algebra of the Lorentz group.

It is often useful to work with a null coframe (basis for one-forms) $\{l,n,m,\overline{m}\}$ dual to a Newman–Penrose tetrad, that is, one for which all inner products vanish except

$$l\cdot n=-m\cdot\overline{m}=1. \tag{A4}$$

From this we can construct a basis for the anti-self-dual two-forms:

$$\mathcal{U}=-n\wedge\overline{m}, \quad \mathcal{M}=l\wedge n-m\wedge\overline{m}, \quad \mathcal{V}=l\wedge m \tag{A5}$$

with the property that all inner products vanish except

$$\mathcal{U}\cdot\mathcal{V}=1, \quad \mathcal{M}\cdot\mathcal{M}=-2. \tag{A6}$$

In this basis, the two-form commutator can be calculated from

$$[\mathcal{M},\mathcal{U}]=-4\mathcal{U}, \quad [\mathcal{M},\mathcal{V}]=4\mathcal{V}, \quad [\mathcal{U},\mathcal{V}]=-\mathcal{M}. \tag{A7}$$

The null basis elements $\mathcal{U}$ and $\mathcal{V}$ for each have one two-dimensional eigenspace, with corresponding zero eigenvalue, spanned by $\{n^\sharp,\overline{m}^\sharp\}$ and $\{l^\sharp,m^\sharp\}$, respectively. These are also the eigenspaces of $\mathcal{M}$ for which they have eigenvalues $+1$ and -1. Note that choosing $\mathcal{M}$ determines $\mathcal{U}$ and $\mathcal{V}$ up to their relative scaling or interchange.

We denote the torsion-free metric compatible covariant derivative of a two-form $\mathcal{Q}$ with respect to a vector field Z by $\nabla_Z\mathcal{Q}$. In terms of this, the exterior derivative d and coderivative $\delta=*d*$ can be expressed:

$$d\equiv e^a\wedge\nabla_{X_a}, \quad \delta\equiv-X^a\lrcorner\nabla_{X_a}.$$

APPENDIX B: THE PETROV CLASSIFICATION

In a vacuum background, the Riemann curvature tensor **R** is equal to the Weyl conformal curvature tensor **C**. The symmetries of these tensors allow them to be written as the sum of terms made of symmetric tensor products of two-forms (i.e., terms like $\mathcal{P}\otimes\mathcal{Q}+\mathcal{P}\otimes\mathcal{Q}$). So, both can be considered as self-adjoint maps on two-forms; if C_{abcd} are components of **C** and $\mathcal{P}_{ab}$ the components of a two-form, then the definition

$$(\mathbf{C}\mathcal{P})_{ab}=\tfrac{1}{2}C_{abcd}\mathcal{P}^{cd}$$

gives the components of the two-form $\mathbf{C}\mathcal{P}$. As a map on two-forms, the conformal tensor preserves the eigenspaces of $*$ and so may be decomposed into a part made from self-dual two-forms alone and a part made from anti-self-dual two-forms. That is, we can write

$$\mathbf{C}=\mathbf{C}^{(+)}+\mathbf{C}^{(-)},$$

where $\mathbf{C}^{(\pm)}\mathcal{Q}^{\mp}=0$. Note that since the conformal tensor is real, $\mathbf{C}^{(-)}$ is the complex conjugate of $\mathbf{C}^{(+)}$, and so it is sufficient to classify only one of these.

The action of $\mathbf{C}^{(-)}$ on the Newman–Penrose two-form basis described in Appendix A is the same as the action of **C** on this basis and can be written as

$$\mathbf{C}^{(-)}\begin{bmatrix}\mathcal{U}\\ \mathcal{M}\\ \mathcal{V}\end{bmatrix}=\begin{bmatrix}-\Psi_2 & \Psi_3 & -\Psi_4\\ -2\Psi_1 & 2\Psi_2 & -2\Psi_3\\ -\Psi_0 & \Psi_1 & -\Psi_2\end{bmatrix}\begin{bmatrix}\mathcal{U}\\ \mathcal{M}\\ \mathcal{V}\end{bmatrix}.$$

Note that the matrix of this transformation is trace-free and the mapping is self-adjoint (that is, $\mathcal{Q}\cdot\mathbf{C}\mathcal{P}=\mathbf{C}\mathcal{Q}\cdot\mathcal{P}$).

The Petrov classification is a classification of this mapping. The space–time is known as algebraically general when there are three distinct eigenvalues, and algebraically special otherwise. Two special cases of interest here are that of type *D* and *N*, for which a basis can be chosen so that the matrix above takes the forms,

$$\begin{bmatrix}-\Psi_2 & 0 & 0\\ 0 & 2\Psi_2 & 0\\ 0 & 0 & -\Psi_2\end{bmatrix},\quad\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ -\Psi_0 & 0 & 0\end{bmatrix},$$

respectively.

The real null direction of a null anti-self-dual two-form $\mathcal{Q}$ is said to be a *principal null direction* (PND) of the conformal tensor if $\mathcal{Q}\cdot\mathbf{C}\mathcal{Q}=0$. We will call such a $\mathcal{Q}$, a *principal null* (PN) two-form. There can be at most four independent PNDs and their number and ''multiplicities'' provide another description of the Petrov types.[3] The multiplicities can be determined in the present formulation by the following (with $\mathcal{P}$ an anti-self-dual two-form):

Multiplicity	Equivalent conditions		
1	$\mathcal{Q}\cdot\mathbf{C}\mathcal{Q}=0$		$\Psi_4=0$
2	$[\mathcal{Q},\mathbf{C}\mathcal{Q}]=0$	$\mathbf{C}\mathcal{Q}\propto\mathcal{Q}$	$\Psi_3=\Psi_4=0$
3	$\mathcal{Q}\cdot\mathbf{C}\mathcal{P}=0\ \forall\mathcal{P}$	$\mathbf{C}\mathcal{Q}=0$	$\Psi_2=\Psi_3=\Psi_4=0$
4	$[\mathcal{Q},\mathbf{C}\mathcal{P}]=0\ \forall\mathcal{P}$	$\mathbf{C}\mathcal{P}\propto\mathcal{Q}\ \forall\mathcal{P}$	$\Psi_1=\Psi_2=\Psi_3=\Psi_4=0$

APPENDIX C: CKY TWO-FORMS AND SHEAR-FREE CONGRUENCES

Defining the shear of a null geodesic vector field requires the choice of a ''screen space,'' and so is not an intrinsic property of the vector field. However, if the shear vanishes for one choice of screen space, then it does for all and hence the notion of a shear-free null vector field is well defined. For definitions and discussion of optical scalars see Ref. 3.

Robinson[15] showed that the real null eigenvector l of a (anti) self-dual null two-form ϕ is geodesic and shear-free if and only if ϕ is proportional to a source-free Maxwell field, that is $d\phi=0$. Note that the eigenspace of such a two-form is two-dimensional, isotropic and integrable. So we can use this fact or the Frobenius integrability condition, that $d\phi=\alpha\wedge\phi$ for some α, for the vanishing of the shear of l. It is convenient here to use these results interchangeably as our criterion for a shear-free null geodesic.

Note that a shear-free null geodesic is a PND of the conformal tensor.

1. Null CKY two-forms

Now, suppose that $\mathcal{Q}$ is a null anti-self-dual CKY two-form. Since the right-hand side of CKY two-form Eq. (15) is simply the anti-self-dual part of $-2Z^\flat\wedge\delta\mathcal{Q}$, we have that

$$0=\mathcal{Q}\cdot 3\nabla_Z\mathcal{Q}=-2(Z^\flat\wedge\delta\mathcal{Q})\cdot\mathcal{Q}=2Z\lrcorner(\delta\mathcal{Q})^\sharp\lrcorner\mathcal{Q}.$$

Hence we can find an α such that $\delta\mathcal{Q}=\alpha^\sharp\lrcorner\mathcal{Q}$ or equivalently $d\mathcal{Q}=-\alpha\wedge\mathcal{Q}$. So the real null eigenvector of $\mathcal{Q}$ is shear-free.

2. Non-null CKY two-forms

We wish to show that the eigenspaces of a non-null CKY two-form $\mathcal{Q}$ are integrable and hence contain a shear-free null geodesic vector field. That is, we want to show that if X and Y are elements of the same eigenspace of $\mathcal{Q}$ with eigenvalue λ ($X\lrcorner\mathcal{Q}=\lambda X^\flat$ and $Y\lrcorner\mathcal{Q}=\lambda Y^\flat$), then so is $[X,Y]$. Since $[X,Y]=\nabla_x Y-\nabla_Y X$, we will show that $\nabla_X Y\lrcorner\mathcal{Q}=\lambda\nabla_X Y^\flat$. Note that this eigenspace is isotropic, that is $g(X,Y)=0$.

Since the map $\alpha\mapsto\alpha^\sharp\lrcorner\mathcal{Q}$ is of maximal rank for non-null $\mathcal{Q}$, it can always be inverted and a one-form α found such that $\delta\mathcal{Q}=-\alpha^\sharp\lrcorner\mathcal{Q}$ and $d\mathcal{Q}=\alpha\wedge\mathcal{Q}$. Using these expressions for $\delta\mathcal{Q}$ and $d\mathcal{Q}$, and the CKY two-form Eq. (15), we have

$$\nabla_X Y\lrcorner\mathcal{Q}=\nabla_X(Y\lrcorner\mathcal{Q})-Y\lrcorner\nabla_X\mathcal{Q}=\lambda\nabla_X Y^\flat+X\lambda Y^\flat-\tfrac{1}{3}\lambda\,\alpha(X)Y^\flat.$$

Rearranging and writing the vector equation dual to this shows that

$$(\nabla_X Y\lrcorner\mathcal{Q})^\sharp-\lambda\nabla_X Y=(X\lambda-\tfrac{1}{3}\lambda\,\alpha(X))Y. \tag{C1}$$

Note that the right-hand side is a multiple of Y and hence an eigenvector of $\mathcal{Q}$ with eigenvalue λ. However, upon contracting the left-hand side with $\mathcal{Q}$, we find that it is an element of the other eigenspace, having eigenvalue $-\lambda$. Hence we must conclude that

$$\nabla_X Y\lrcorner\mathcal{Q}-\lambda\nabla_X Y^\flat=0, \tag{C2}$$

and we have the required result.

Since each eigenspace of $\mathcal{Q}$ is integrable they each give rise to a null self-dual two-form proportional to a Maxwell field, and hence the real eigenvectors of $\mathcal{Q}$ are shear-free.

APPENDIX D: INTEGRABILITY OF CKY TWO-FORMS

Apart from conformally flat space–times, CKY two-forms can only exist in space–times of Petrov type D or N. To understand this it is sufficient to consider only CKY tensors of definite Hodge-duality, for which we give an integrability condition. For an anti-self-dual CKY two-form $\mathcal{Q}$,

$$[\mathcal{Q},C\mathcal{P}]=\tfrac{1}{2}[\mathcal{P},C\mathcal{Q}], \quad \forall \text{ two-forms } \mathcal{P}. \tag{D1}$$

If we let $\mathcal{P}=\mathcal{Q}$, it follows that

$$[\mathbf{C}\mathcal{Q},\mathcal{Q}]=0.$$

Then, from the commutator algebra of anti-self-dual two-forms Eq. (A7), it can be deduced that $C\mathcal{Q}$ must be proportional to $\mathcal{Q}$, i.e.,

$$\mathbf{C}\mathcal{Q}=\mu\mathcal{Q}, \tag{D2}$$

where μ is a scalar. From this, we can deduce the Petrov type as described in Appendix B.

1. Null CKY two-forms

When $\mathcal{Q}$ is null this implies that the real null eigenvector of $\mathcal{Q}$ is a repeated principal null direction. However, if we write out Eq. (D1) in an anti-self-dual two-form basis chosen so that $\mathcal{U}=\mathcal{Q}$ and $\mathcal{V}\propto\mathcal{P}$, we find that $\mu=-\Psi_2=0$. Not only does this immediately tell us that $C\mathcal{Q}=0$, but upon substitution into Eq. (D1) we have that $[\mathcal{Q},C\mathcal{P}]=0$ for all anti-self-dual two-forms $\mathcal{P}$. Hence the real null direction defined by $\mathcal{Q}$ is a fourfold PND and the space–time is of Petrov type *N*.

2. Non-null CKY two-forms

When $\mathcal{Q}$ is non-null, we concluded in Appendix C that the real null eigenvectors of $\mathcal{Q}$ are shear-free. If we align our anti-self-dual two-form basis so that $\mathcal{M}\propto\mathcal{Q}$ then $\mathcal{U}$ and $\mathcal{V}$ have shear-free eigenvectors and hence are PN two-forms. From this we conclude that $\Psi_0=\Psi_4=0$. The integrability condition Eq. (D2) immediately requires that Ψ_1 and Ψ_3 vanish and hence the space–time is of Petrov type *D*.

This reasoning made no use of Ricci-flatness wherein the Goldberg-Sachs theorem[16] would imply the same result.

[1] J. N. Goldberg, Phys. Rev. D **41**, 410 (1990).
[2] R. Penrose, Proc. R. Soc. London, Ser. A **381**, 53 (1982).
[3] R. Penrose and W. Rindler, *Spinors and Space-Time* (Cambridge Univerity Press, Cambridge, 1986), Vol. 2, Eq. (6.4.6).
[4] K. P. Tod, Proc. R. Soc. London, Ser. A **388**, 457 (1983).
[5] E. N. Glass, J. Math. Phys. **37**, 421 (1996).
[6] E. N. Glass and M. G. Naber, J. Math. Phys. **35**, 4178 (1994).
[7] B. Carter, Commun. Math. Phys. **10**, 280 (1968).
[8] M. Walker and R. Penrose, Commun. Math. Phys. **18**, 265 (1970).
[9] C. D. Collinson, Int. J. Theor. Phys. **15**, 311 (1976).
[10] See Ref. 3, Sec. 6.5.
[11] S. Tachibana, Tohoku Math. J. **21**, 56 (1969).
[12] I. M. Benn, P. Charlton, and J. Kress, J. Math. Phys. **38**, 4504 (1997).
[13] J. Kress, Ph.D. thesis, Dept. of Mathematics, Newcastle University, NSW, Australia, 1998.
[14] W. Dietz and R. Rüdiger, Gen. Relativ. Gravit. **12**, 545 (1980).
[15] I. Robinson, J. Math. Phys. **2**, 290 (1961).
[16] J. N. Goldberg and R. K. Sachs, Acta Phys. Pol. **22**, 13 (1962).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

Black hole interacting with matter as a simple dynamical system

Petr Hájíček
Institute for Theoretical Physics, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland

(Received 21 July 1998; accepted for publication 28 August 1998)

Recently, a variational principle has been derived from Einstein–Hilbert and a matter Lagrangian for the spherically symmetric system of a dust shell and a black hole. The so-called physical region of the phase space, which contains all physically meaningful states of the system defined by the variational principle, is specified; it has a complicated boundary. The principle is then transformed to new variables that remove some problems of the original formalism: the whole phase space is covered (in particular, the variables are regular at all horizons), the constraint has a polynomial form, and the constraint equation is uniquely solvable for two of the three conserved momenta. The solutions for the momenta are written down explicitly. The symmetry group of the system is studied. The equations of motion are derived from the transformed principle and are shown to be equivalent to the previous ones. Some lower-dimensional systems are constructed by exclusion of cyclic variables, and some of their properties are found. © *1999 American Institute of Physics.* [S0022-2488(99)03701-9]

I. INTRODUCTION

The spherically symmetric gravitating shell has been used as a simplified quantum model at many occasions. One can mention the quantum effects of domain walls in the early Universe,[1] quantum aspects of gravitational collapse,[2] quantum theory of black holes[3] or Hawking evaporation of black holes.[4]

The equations of motion for the shells has been derived by Dautcourt[5] and transformed to a geometric (i.e., ''gauge invariant'') form by Israel.[6] However, for a quantum theory, one needs a variational principle rather than dynamical equations. Such a variational principle has as yet been just guessed from the dynamical equations, Refs. 7 and 8, or from some intermediate variational principle, Refs. 4 and 9. Quite a number of different Hamiltonians have resulted and the corresponding quantum theories have not been unitarily equivalent. There are two sources of the ambiguity.

The first source of ambiguity is the invariance of the general relativity with respect to any coordinate transformations on one hand, and the property of Hamiltonians to generate dynamics with respect to a particular time coordinate on the other. For example, the choice of the proper time along the shell as such a coordinate leads to the $\cosh p$ Hamiltonian, Refs. 7 and 10, the Schwarzschild time coordinate inside the shell to the square-root Hamiltonian, Ref. 2, and the Schwarzschild time coordinate outside the shell to a merely implicitly determined Hamiltonian of Refs. 4 and 9. An attempt to work out a gauge invariant theory for at least some class of the guessed Hamiltonians is Ref. 8; based on a technical assumption that the super-Hamiltonian is quadratic in the momenta, it gives a unique action principle.

The second source of ambiguity is the fact that equations of motion do not determine the corresponding variational principle uniquely, in general. This ambiguity can be removed by a direct derivation of the variational principle for the shell from the Einstein–Hilbert and shell matter Lagrangian. Three such derivations exist: Refs. 11, 12, and 13; they all lead to the same variational principle for the same system. In Refs. 12 and 13, the reduction of the second order formalism to the spherically symmetric case is done first, followed by a reduction to dynamical variables of the shell alone and by the transformation to the second order formalism. In Ref. 11, a first and second order formalism for a general shell and gravitational field is derived (no symmetry); this second order formalism is reduced by the spherical symmetry and to shell variables in Ref. 14.

We are then left with the following dilemma. The super-Hamiltonian of Ref. 8 is simple and amenable to the existing gauge invariant quantization methods (like, e.g., Refs. 15 and 16). The super-Hamiltonian of Refs. 12, 14 and 13 follows from the Einstein–Hilbert and shell-matter Lagrangian, but it is extremely complicated: it contains nested square roots and hyperbolic functions, and for a good measure it is formulated in coordinates that diverge on horizons. It would be difficult to quantize.

A natural question then arises: what is the relation between the guessed and simple variational principle of Ref. 8 and the derived but complicated one of Refs. 12, 14 and 13? If the simple principle were equivalent to the complicated one, we could use the simple principle and forget the complicated one. However, the answer turns out to be rather surprising (see Ref. 17): the two descriptions are only locally, but not globally, equivalent. The local equivalence explains why the equations of motion are the same. For the study of the problem, a geometric approach of Ref. 18 has been applied. All gauge invariant properties of a constrained system are encoded in two objects: the constraint manifold Γ and the presymplectic form Ω on Γ. The equivalence of two such systems, (Γ_1,Ω_1) and (Γ_2,Ω_2), is then a well-defined mathematical concept independent of a choice of extended phase space, a choice of constraint functions or a choice of gauge.

In this situation, it seems natural to look for a transformation of the variational principle of Refs. 12, 14 and 13 to a better set of coordinates, and this will be the main topic of the present paper. The method will again be based on the gauge invariant description (Γ,Ω) of reparametrization invariant systems. We shall perform the transformation in three steps; to motivate the steps, we take into account the symmetry of the system, the structure of the so-called *Cartan form* Θ, which is defined by $d\Theta := \Omega$, and the topology of the space (Γ,Ω). The topological problem involved here is the existence of transversal surfaces in Γ; such surfaces are nowhere tangential to the dynamical trajectories. Functions whose levels are transversal are not only helpful, if one looks for nice coordinates, because they simplify Θ. Quite generally, they have to do with the existence of Hamiltonian and with the possibility to give the quantum dynamics the form of the Schrödinger equation (cf. Ref. 18). We shall, therefore, also look for transversal surfaces in a more systematic way. Finally, we shall find that the structure of the phase space of the spherically symmetric shell is not simple. Only a proper subset, which we call a *physical region*, contains physical states of the shell. Its nature can be roughly specified as follows. All points inside it correspond to the system consisting of a shell of positive mass and radius, interacting with a black hole. The boundary of the region is complicated; it contains shells with zero radius, shells with zero rest mass (which cannot, however, be regarded as shells made of light-like matter), as well as self-gravitating isolated shells with positive radius and mass—these are the points at the boundary that are physically meaningful. The points outside the region that describe shells with positive radius and mass correspond to systems consisting of the shell and a negative mass source. A conclusion seems to be that the system will be difficult to quantize even if the new variables have made it algebraically simple. One problem will be to keep the spectra of observables contained within the physical region. Another problem will be to construct a unitary dynamics, because the classical dynamics breaks down at the boundary of the physical region.

The plan of the paper is as follows. In Sec. II, we collect the relevant results of the previous papers, mainly Refs. 14 and 13. The variational principle that we describe allows also for an internal degree of freedom of the shell; it depends, therefore, on two additional variables in comparison with Refs. 12 and 14: the proper time along the shell trajectory and the rest mass of the shell. We determine the constraint manifold Γ and the Cartan form Θ that follow from the variational principle. We study the physical region.

In Sec. III, we introduce three new functions on the phase space that are regular at the horizons and that will replace the momentum P of the shell and the two Schwarzschild times $T^\pm$ that diverge there. We describe their physical and geometrical meaning, also using some results of previous papers. In Sec. IV, we specify a new set of coordinates on Γ, and show that the transformation from the old to the new coordinates is differentiable and invertible. We also find the ranges of these coordinates in the physical region. In Sec. V we collect and extend the results about symmetries of the system that were obtained in Ref. 17 and express the symmetry transformations in the new coordinates; the form of the transformation simplifies. The new times are cyclic coordinates, as the old have been. In Sec. VI, we transform the Cartan form into new coordinates. The square roots and hyperbolic functions disappear, and an equivalent form becomes regular at the horizons so that it can be smoothly matched across them. However, the Cartan form is still complicated and this motivates further transformations. The final result thereof can be described as follows. The extended phase space is $(\mathbf{R}^8,\bar{\Omega})$. The natural coordinates of $\mathbf{R}^8$ form a Darboux chart for $\bar{\Omega}$. The equation of the constraint surface Γ in $\mathbf{R}^8$ is polynomial in all momenta and the presymplectic form $d\Theta$ is the pull-back of $\bar{\Omega}$ to Γ.

In Sec. VII, we study the polynomial constraint. We show that it defines two disjoint submanifolds in $\mathbf{R}^8$. One, denoted by Γ_6, is six-dimensional and lies well outside the physical region. The other, denoted by Γ_7, is seven-dimensional and intersects the physical region. In the intersection of Γ_7 with the physical region, we find two different foliations by transversal surfaces. They are not globally transversal, however, because none of them is intersected by all dynamical trajectories. The problem is that each dynamical trajectory starts or finishes at the $(R=0)$-subset of the boundary of the physical region. The two different Hamiltonians corresponding to the two foliations are written down explicitly; they contain second and third roots.

In Sec. VIII, we repeat the study for some important lower-dimensional cases; they can be obtained by exclusion of cyclic coordinates from the original system. In particular, a (dynamical) black hole and the shell without its internal degree of freedom, the shell in the field of a fixed black hole external field or the isolated self-gravitating shell (with the flat spacetime inside) are considered.

II. DESCRIPTION OF THE SYSTEM

In this section, we shall collect some results of Refs. 8, 14 and 13 so that the paper becomes relatively self-consistent. We shall also describe and study the physical region.

A spherically symmetric thin-shell spacetime solution of Einstein equations can be constructed as follows. Consider two Schwarzschild spacetimes $\mathcal{M}_1$ and $\mathcal{M}_2$ with Schwarzschild masses E_1 and E_2. Let Σ_1 be a time-like hypersurface in $\mathcal{M}_1$ and Σ_2 be one in $\mathcal{M}_2$. Let Σ_1 divide $\mathcal{M}_1$ into two subspacetimes, $\mathcal{M}_{1+}$ and $\mathcal{M}_{1-}$, and similarly Σ_2 divide $\mathcal{M}_2$ into $\mathcal{M}_{2+}$ and $\mathcal{M}_{2-}$. As everything is spherically symmetric, all spacetimes are effectively two-dimensional; we chose fixed time and space orientation in the two-dimensional Schwarzschild spacetimes so that future and past as well as right and left are unambiguous; then let $\mathcal{M}_{2+}$ and $\mathcal{M}_{1+}$ be right with respect to $\mathcal{M}_{2-}$ and $\mathcal{M}_{1-}$. Let Σ_1 and Σ_2 be isometric; then the spacetime $\mathcal{M}_{1-}$ can be pasted together with the spacetime $\mathcal{M}_{2+}$ along the boundaries Σ_1 and Σ_2. The result is a shell spacetime $\mathcal{M}_s$. Given a shell spacetime, we shall leave out the indices 1 and 2, having right (left) energy E_+ (E_-), shell trajectory Σ, and the right (left) subspacetime $\mathcal{M}_+$ ($\mathcal{M}_-$). Thus, $\mathcal{M}_+=\mathcal{M}_2\cap\mathcal{M}_s$ and $\mathcal{M}_-=\mathcal{M}_1\cap\mathcal{M}_s$. In Refs. 14, 17 and 8, the Schwarzschild masses E_ϵ, $\epsilon=\pm 1$ were assumed to be *positive*.

Each subspacetime $\mathcal{M}_\epsilon$ has the metric

$$ds_\epsilon^2=-F_\epsilon(R)(dT^\epsilon)^2+F_\epsilon^{-1}(R)dR^2, \tag{1}$$

where

$$F_\epsilon(R):=1-\frac{2E_\epsilon}{R},$$

and it is split up by its horizons into four quadrants: we denote by Q_I that which is adjacent to the right infinity, Q_{II} to the left infinity, Q_{III} to the future singularity and Q_{IV} to the past singularity. Suitable notation (introduced in Ref. 14) enables us to write all formulas in a form valid in any quadrant: we define four sign functions a_ϵ and b_ϵ distinguishing the quadrants of $\mathcal{M}_\epsilon$: $a_\epsilon := \mathrm{sgn}\, F_\epsilon$, and b_ϵ equals to $+1$ (-1) in the past (future) of the event horizon in the spacetime $(\mathcal{M}_\epsilon, g_\epsilon)$. We also include the dust internal degree of freedom, M, and the proper time T along the shell trajectory using the results of Ref. 13. Then the total action for the dust and gravity can be rewritten in the form

$$S_\Sigma[\mathsf{T},\mathsf{M};T^\pm,E_\pm;R,P;\nu] = \int dt(P\dot{R} - E_+\dot{T}^+ + E_-\dot{T}^- + \mathsf{M}\dot{\mathsf{T}} - \nu\mathcal{C}), \tag{2}$$

where ν is a Lagrange multiplier,

$$\mathcal{C} = \mathsf{M} - R\sqrt{F_+ + F_- - 2a_-b_+b_-\sqrt{|F_+F_-|}\mathrm{sh}_{a_+a_-}(P/R)}, \tag{3}$$

is the super-Hamiltonian, $\mathrm{sh}_a x := (e^x + ae^{-x})/2$ for any $a = \pm 1$ and $x \in (-\infty, \infty)$. The momentum P conjugate to the radial coordinate R can be defined as follows (for more details, see Ref. 14). Let (n_ϵ, m_ϵ) be the orthonormal dyad at a shell point in the subspacetime $(\mathcal{M}_\epsilon, g_\epsilon)$, the vector n_ϵ being tangential to the shell and future oriented, m_ϵ being right oriented; we call it *shell dyad*. Further, let the so-called *Schwarzschild dyad* $(n_{S\epsilon}, m_{S\epsilon})$ be defined at each point of the shell in the spacetime $(\mathcal{M}_\epsilon, g_\epsilon)$ by the requirement that $n_{S\epsilon}$ be time-like future oriented, $m_{S\epsilon}$ be right oriented and that one of the vectors be tangential to the $(R = \text{const})$-curves. Clearly, in each quadrant, we have a different formula for the components of the Schwarzschild dyad; with respect to the Schwarzschild coordinates:

$$\text{for } F_\epsilon > 0, \quad n_{S\epsilon} = (b_\epsilon/\sqrt{F_c}, 0), \quad m_{S\epsilon} = (0, b_\epsilon\sqrt{F_\epsilon});$$

$$\text{for } F_\epsilon < 0, \quad n_{S\epsilon} = (0, b_\epsilon\sqrt{|F_\epsilon|}), \quad m_{S\epsilon} = (-b_\epsilon/\sqrt{|F_\epsilon|}, 0).$$

Then, P_ϵ/R is the hyperbolic angle between the Schwarzschild and shell dyads:

$$n_\epsilon = n_{S\epsilon} \cosh P_\epsilon/R + m_{S\epsilon} \sinh P_\epsilon/R,$$

$$m_\epsilon = n_{S\epsilon} \sinh P_\epsilon/R + m_{S\epsilon} \cosh P_\epsilon/R.$$

Finally, $P := [P]$, where we use the common short-hand for the jump $[A] := A_+ - A_-$ of a quantity A across the shell.

The extended phase space of the system is eight-dimensional, split up into 16 disjoint sectors, each sector being a pair of quadrants, one chosen from the left and one from the right subspacetimes. Each of these sectors can be covered with the coordinates P, R, E_+, T^+, E_-, T^-, M and T. The constraint surface that we call Γ is defined by the constraint equation $\mathcal{C} = 0$ in the extended phase space. It does not intersects all 16 sectors, but it is split up into at most (if $E_- > \mathsf{M}$) nonempty intersections of the sectors with it. (Observe that we admit unphysical sectors, i.e., those in which the constraint equation $\mathcal{C} = 0$ has no solutions. This simplifies the boundary of the extended phase space a great deal and does not lead to any problem: it is the very purpose of extended phase spaces to enclose also nonphysical poits, if this leads to any simplification.) These can be covered by the system of seven coordinates P, R, E_+, T^+, E_-, T^- and T. The following coordinates are assumed to have nontrivial ranges:

$$R > 0, \quad E_+ > 0, \quad E_- > 0, \quad \mathsf{M} > 0. \tag{4}$$

The pull-back of the Liouville form of the action (2) to the constraint surface Γ in these coordinates is simply

$$\Theta = P\ dR - E_+\ dT^+ + E_-\ dT^- + \mathsf{M}\ d\mathsf{T}, \tag{5}$$

where M is now a function of the seven coordinates defined by

$$\mathsf{M} = R\sqrt{F_+ + F_- - 2a_-b_+b_-\sqrt{|F_+F_-|}sh_{a_+a_-}(P/R)}. \tag{6}$$

The form (5) is called *Cartan form* of the system; it contains all information about the dynamics and about the Poisson brackets (see, e.g., Refs. 19 or 17).

In the present paper we start from these formulas and quantities. The first new observation that we make is that all these formulas remain valid also for the values $E_+ = 0$ or $E_- = 0$, if the notation and conventions are adapted a little. These cases are interesting because they describe an isolated self-gravitating shell (with the flat spacetime inside). Indeed, the metric of the flat spacetime is covered by formula (1). However, the flat spacetime does not contain any horizon and is not split into quadrants. Still, a piece of the flat spacetime can be used in four distinct ways in the construction of the shell spacetime: either it lies to the left ($E_- = 0$) or to the right ($E_+ = 0$) of the shell, and, in each case, the R coordinate can either increase to the right or to the left (R is always space-like in the flat spacetime). We can, therefore, formally define the *E-spacetime* for $E > 0$ as before to be the Kruskal spacetime of Schwarzschild mass E with four quadrants, and for $E = 0$ as consisting of two topologically separated quadrants Q_I and Q_{II}, each isometric to the flat spacetime, but with opposite time and space orientations: in Q_I, the radial (''Schwarzschild'') coordinate R increases to the right and the ''Schwarzschild'' time T to the future; in Q_{II}, R increases to the left and T to the past. It is amusing to observe that the validity of the formulas (2) and (3) can then be extended to vanishing E's, if we just use the values $a_\epsilon = 1$ everywhere, $b_\epsilon = +1$ in Q_I and $b_\epsilon = -1$ in Q_{II}. This works because the components of the Schwarzschild dyad that are crucial for the derivation of Eq. (3) in Ref. 14 retain their form in the quadrants Q_I and Q_{II} of the E-spacetime for $E = 0$, and there are no quadrants Q_{III} and Q_{IV}. Similarly, the sign relation (14) of Ref. 8, which is basic for the form of the dynamical equations, remains valid with our orientation and quadrant conventions. Observe that the presence of two topologically separated subspacetimes is not observable and does not imply any physical assertion; similarly, the orientation of coordinates is a mere coordinate convention without any physical meaning.

A very important but rather embarrassing observation is now that a formally impeccable shell spacetime can be pasted together from Schwarzschild spacetimes of any mass, even a negative one. The metric is still given by Eq. (1) and the spacetime has a similar global structure as the flat one: there are no horizons and there are two possible orientations of how it can be built in. Again, all equations remain valid, if we accept the same sign, quadrant and orientation conventions as for the flat spacetime and define the (two-quadrant) E-spacetime for $E < 0$ analogously to what we have done for $E = 0$. The corresponding shell dynamics is then again given by the action (2).

Let us quickly discuss the dynamical equations following from the action (2) under these conventions. There are three conservation laws,

$$\dot{E}_+ = \dot{E}_- = \dot{\mathsf{M}} = 0, \tag{7}$$

resulting from varying the action with respect to the cyclic coordinates T^+, T^- and T. The variations with respect of the conserved momenta and P with the subsequent simplification by the constraint (6) yield

$$\dot{\mathsf{T}} = \nu, \tag{8}$$

$$\dot{T}^\epsilon = -\epsilon\nu\,\frac{R}{\mathsf{M}}\left(1 - a_-b_+b_-\,\frac{\sqrt{|F_+F_-|}}{F_\epsilon}\,\mathrm{sh}_{a_+a_-}\frac{P}{R}\right), \tag{9}$$

$$\dot{R} = \nu a_-b_+b_-\,\frac{R}{\mathsf{M}}\,\sqrt{|F_+F_-|}\mathrm{sh}_{-a_+a_-}\frac{P}{R}. \tag{10}$$

The equation determining $\dot{P}$ can be obtained either by varying the action with respect to R, or by differentiating the constraint equation $\mathcal{C}=0$ with respect to t followed by substitution for all other t-derivatives from Eqs. (7)–(10). Let us rewrite the constraint equation in the form

$$-\frac{\mathsf{M}}{2R}+\frac{R}{2\mathsf{M}}(F_+ + F_-)=a_-b_+b_-\frac{R}{\mathsf{M}}\sqrt{|F_+F_-|}\,\mathrm{sh}_{a_+a_-}\frac{P}{R}. \tag{11}$$

Equations (7)–(11) represent a complete system of dynamical equations for the shell.

We derive some important consequences from the dynamical equations. Equations (8)–(11) imply that

$$-F_\epsilon\left(\frac{dT^\epsilon}{d\mathsf{T}}\right)^2+\frac{1}{F_\epsilon}\left(\frac{dR}{d\mathsf{T}}\right)^2=-1,$$

confirming that T is the proper time along shell trajectories. Equations (8), (9) and (11) deliver the *time equation*:

$$F_\epsilon\frac{dT^\epsilon}{d\mathsf{T}}=-\epsilon\frac{\mathsf{M}}{2R}+\frac{E_+-E_-}{\mathsf{M}}. \tag{12}$$

From Eqs. (10) and (11), the *radial equation* follows:

$$\frac{dR}{d\mathsf{T}}=-\omega\sqrt{-V}, \tag{13}$$

where

$$V(R)=-\frac{\mathsf{M}^2}{4R^2}-\frac{E_++E_-}{R}-\frac{(E_+-E_-)^2}{\mathsf{M}^2}+1, \tag{14}$$

and $\omega=\pm 1$ is a suitable sign [see Eq. (2) of Ref. 17]. The square of the radial equation can be rewritten in one of the two forms

$$F_\epsilon+\left(\frac{dR}{d\mathsf{T}}\right)^2=J_\epsilon^2, \tag{15}$$

according as $\epsilon=\pm 1$, where

$$J_\epsilon:=\frac{\mathsf{M}}{2R}-\epsilon\frac{E_+-E_-}{\mathsf{M}} \tag{16}$$

(cf. Ref. 8). It follows immediately from this definition of J_ϵ that

$$J_++J_-=\frac{\mathsf{M}}{R},$$

and this can be transformed with the help of Eq. (15) to

$$\mathrm{sgn}\,J_+\sqrt{F_++\left(\frac{dR}{d\mathsf{T}}\right)^2}+\mathrm{sgn}\,J_-\sqrt{F_-+\left(\frac{dR}{d\mathsf{T}}\right)^2}=\frac{\mathsf{M}}{R}.$$

However, comparing Eqs. (12) and (16), we observe that

$$\text{sgn } J_\epsilon = -\epsilon \text{ sgn}\left(F_\epsilon \frac{dT^\epsilon}{d\mathsf{T}} \right).$$

The sign of the expression $F_\epsilon(dT^\epsilon/d\mathsf{T})$ deserves a name; let us call it τ_ϵ (cf. Ref. 8). Then, we obtain, finally,

$$-\tau_+ \sqrt{F_+ + \left(\frac{dR}{d\mathsf{T}} \right)^2} + \tau_- \sqrt{F_- + \left(\frac{dR}{d\mathsf{T}} \right)^2} = \frac{\mathsf{M}}{R}. \tag{17}$$

This is the so-called *Israel equation* for a dust shell [cf. Eqs. (14), (21) and (22) of Ref. 8].

Equations (12)–(17) are valid for all values of Schwarzschild masses $E_+ \in (-\infty, +\infty)$ and $E_- \in (-\infty, +\infty)$. Some features of the unphysical (negative E_ϵ) cases are described by the following theorem.

Theorem 1: *Let E_ϵ be nonpositive and not larger than $E_{-\epsilon}$ for some value of ϵ. Then the timelike (connected) center $(R=0)$ curve lies in the subspacetime $\mathcal{M}_\epsilon$ on the ϵ-side of the shell.* To prove the theorem, we observe that $E_\epsilon \leq E_{-\epsilon}$ implies

$$\sqrt{F_\epsilon + (dR/d\mathsf{T})^2} \geq \sqrt{F_{-\epsilon} + (dR/d\mathsf{T})^2}.$$

Thus, the left-hand side of Eq. (17) can only be positive, if $\tau_\epsilon = -\epsilon$. The assumption $E_\epsilon \leq 0$ of the theorem implies that $\mathcal{M}_\epsilon$ is the two-quadrant E-spacetime with $\tau_\epsilon > 0$ in the quadrant Q_I and $\tau_\epsilon < 0$ in the quadrant Q_{II}; this follows from the definition of τ_ϵ and from the positivity of F_ϵ (in fact, for $E_\epsilon \leq 0$, we have $F_\epsilon \geq 1$). Hence, to the right $(\mathcal{M}_+)$ of the shell, there is a part of Q_{II} with the center to the right, and to the left of the shell $(\mathcal{M}_-)$, there must be Q_I with the center to the left. Q.E.D.

An interpretation of the theorem is that the shell spacetime containing one or two negative-mass subspacetimes is unphysical: there must be at least one negative-mass source somewhere outside the shell. The theorem also says that a shell spacetime constructed from one flat and one positive-mass subspacetimes must have a regular center inside the flat subspacetime.

To summarize: The action (2) generates a regular dynamics in the enlarged phase space just defined by

$$\mathsf{M} > 0, \quad R > 0. \tag{18}$$

For $E_+ > 0$ and $E_- > 0$, it describes a shell interacting with a black hole of mass E_- (or E_+, depending on from where we observe the shell). For $E_+ > 0$ and $E_- = 0$ (or $E_+ = 0$ and $E_- > 0$), it describes a self-gravitating shell with flat spacetime inside. If any of E_+ and E_- are negative, it describes the shell interacting with a negative mass source. Only the points with

$$E_+ \geq 0, \quad E_- \geq 0, \tag{19}$$

are physically sensible; the subset specified by the inequalities (18) and (19) is the *physical region* of the phase space. There seems to be nothing about the variational principle (2) that would help us to enforce the validity of the Eqs. (19) in any natural, automatic way.

III. RADIAL AND KRUSKAL MOMENTA

From the definition of the functions $T^\pm$ and P as given in the previous section, it follows that they are singular at the horizons: T^ϵ at $R = 2E_\epsilon$ and P at $R = 2E_-$ as well as at $R = 2E_+$. In this section, we introduce three functions that are well-defined on the whole constraint surface to replace $T^\pm$ and P. Another replacement of $T^\pm$ and P with similar properties has been tried in Ref. 14. There, the functions were constructed from the Kruskal coordinates; however, the Kruskal coordinates make no sense for flat spacetime and so the important case $E_- = 0$ could not be incorporated. In this subsection, we remove this problem.

Let us define the new functions q, $[T_1]$ and $\overline{T_1}$ by

$$q:=a_-b_+b_-R\sqrt{|F_+F_-|}\mathrm{sh}_{-a_+a_-}\frac{P}{R}, \tag{20}$$

$$\overline{T_1}:=\frac{T^++T^-}{2}+E_+\ \ln\left|\frac{u_+}{v_+}\right|+E_-\ \ln\left|\frac{u_-}{v_-}\right|, \tag{21}$$

$$[T_1]:=T^+-T^-+2E_+\ \ln\left|\frac{u_+}{v_+}\right|-2E_-\ \ln\left|\frac{u_-}{v_-}\right|, \tag{22}$$

and the auxiliary functions u_ϵ and v_ϵ by

$$u_\epsilon:=-\frac{q}{\mathsf{M}}-\frac{[E]}{\mathsf{M}}+\epsilon\frac{\mathsf{M}}{2R},\quad v_\epsilon:=-\frac{q}{\mathsf{M}}+\frac{[E]}{\mathsf{M}}-\epsilon\frac{\mathsf{M}}{2R}, \tag{23}$$

in Eq. (23), the functions q and M are given by Eqs. (20), (6) and $[E]:=E_+-E_-$.

The meaning of the momentum q can be seen if it is expressed by means of velocities along dynamical trajectories. Comparing Eqs. (8), (10) and (20), we find that

$$q=\mathsf{M}\frac{dR}{d\mathsf{T}}; \tag{24}$$

hence, q is a kind of radial momentum, and it is regular everywhere along dynamical trajectories.

The meaning of the functions $\overline{T_1}$ and $[T_1]$ can be inferred from that of T_1^ϵ, which are defined by

$$T_1^\epsilon:=\overline{T_1}+\frac{\epsilon}{2}[T_1].$$

Equations (21) and (22) lead then to

$$T_1^\epsilon=T^\epsilon+2E_\epsilon\ \ln\left|\frac{u_\epsilon}{v_\epsilon}\right|,$$

for each $\epsilon=\pm1$. Let us limit ourselves to the physical region and distinguish two cases. $\underline{E_\epsilon=0}$. Then, $T_1^\epsilon=T^\epsilon$ in both quadrants of the E-spacetime, so $b_\epsilon T_1^\epsilon$ is the time coordinate of the inertial system of the Minkowski spacetime $\mathcal{M}_\epsilon$ in which the center of mass of the shell is in rest. $\underline{E_\epsilon>0}$. Then, it is possible to introduce Kruskal coordinates U_ϵ and V_ϵ in the Kruskal spacetime $\mathcal{M}_\epsilon$ by

$$R=2E_\epsilon\kappa(-U_\epsilon V_\epsilon),$$
$$T^\epsilon=2E_\epsilon\ \ln\left|\frac{V_\epsilon}{U_\epsilon}\right|, \tag{25}$$

and by the requirement that both functions U_ϵ and V_ϵ increase to the future, where κ is the well-known monotonous function defined on the interval $(-1,\infty)$ by its inverse,

$$\kappa^{-1}(x)=(x-1)e^x.$$

At each point of the shell trajectory $U_\epsilon=U_\epsilon(t)$ and $V_\epsilon=V_\epsilon(t)$, the so-called *Kruskal momentum*, P_K^ϵ, was defined in Ref. 14 by

$$P_K^\epsilon := \frac{R}{2} \ln \frac{dV_\epsilon}{dU_\epsilon}.$$

In Ref. 17, four further functions u_ϵ and v_ϵ were defined by

$$u_\epsilon := \frac{U_\epsilon e^{P_K^\epsilon/R}}{\sqrt{\kappa_\epsilon e^{\kappa_\epsilon}}}, \quad v_\epsilon := \frac{V_\epsilon e^{-P_K^\epsilon/R}}{\sqrt{\kappa_\epsilon e^{\kappa_\epsilon}}}, \tag{26}$$

where $\kappa_\epsilon := \kappa(-U_\epsilon V_\epsilon)$.

The following relation between the pairs u_ϵ, v_ϵ and R, E_ϵ was shown to hold in Ref. 17 [Eqs. (32) and (33)]:

$$u_\epsilon = -\frac{[E]}{\mathsf{M}} + \epsilon \frac{\mathsf{M}}{2R} + \omega\sqrt{-V}, \quad v_\epsilon = \frac{[E]}{\mathsf{M}} - \epsilon \frac{\mathsf{M}}{2R} + \omega\sqrt{-V}.$$

A comparison with Eqs. (13) and (14) shows that our functions u_ϵ and v_ϵ defined by Eqs. (23) coincide with the functions u_ϵ and v_ϵ of Ref. 17, if the dynamical equations are satisfied. Equation (26) then shows that u_ϵ and v_ϵ vanish at horizons and that their logarithms diverge there. It then also follows for the functions T_1^ϵ along dynamical trajectories that

$$T_1^\epsilon = \frac{4E_\epsilon P_K^\epsilon}{R}. \tag{27}$$

We can see that the functions T_1^ϵ are directly determined by the geometry of the dynamical trajectory in the spacetime $\mathcal{M}_\epsilon$ and are regular everywhere along the trajectory.

IV. A REGULAR COORDINATE SYSTEM

In this section, we describe a coordinate system that is regular at the horizons, and at the same time remains meaningful for the special values $E_\epsilon = 0$ of Schwarzschild masses.

As such coordinates on Γ, we suggest the seven functions q, R, M, T, $[E]$, $\overline{T_1}$, and $[T_1]$. Let us show that the Jacobian of the transformation from these coordinates to P, R, E_+, T^+, E_-, T^- and T is nonzero. To calculate the determinant, we observe that the four columns R, T^+, T^- and T of the Jacobian,

$$\frac{\partial(q,\ldots,[T_1])}{\partial(P,\ldots,\mathsf{T})},$$

contain each, at most, two nonzero elements; if we expand the determinant along these columns, we obtain

$$\frac{\partial(q,\ldots,[T_1])}{\partial(P,\ldots,\mathsf{T})} = -\frac{\partial q}{\partial P}\left(\frac{\partial \mathsf{M}}{\partial E_+} + \frac{\partial \mathsf{M}}{\partial E_-}\right) + \frac{\partial \mathsf{M}}{\partial P}\left(\frac{\partial q}{\partial E_+} + \frac{\partial q}{\partial E_-}\right). \tag{28}$$

For the derivatives involved here, Eqs. (6) and (20) yield immediately

$$\frac{\partial q}{\partial P} = a_- b_+ b_- \sqrt{|F_+ F_-|}\,\mathrm{sh}_{a_+ a_-} \frac{P}{R}, \quad \frac{\partial \mathsf{M}}{\partial P} = -a_- b_+ b_- \sqrt{|F_+ F_-|}\,\frac{R}{\mathsf{M}}\,\mathrm{sh}_{-a_+ a_-} \frac{P}{R},$$

$$\frac{\partial \mathsf{M}}{\partial E_+} + \frac{\partial \mathsf{M}}{\partial E_-} = \frac{R}{\mathsf{M}}\left(-2 + a_+ b_+ b_- \frac{F_+ + F_-}{\sqrt{|F_+ F_-|}}\,\mathrm{sh}_{a_+ a_-} \frac{P}{R}\right),$$

$$\frac{\partial q}{\partial E_+}+\frac{\partial q}{\partial E_-}=-a_+b_+b_-\frac{F_++F_-}{\sqrt{|F_+F_-|}}\,\mathrm{sh}_{-a_+a_-}\frac{P}{R}.$$

Substituting this in Eq. (28) and using again Eq. (6), we arrive at the simple result

$$\frac{\partial(q,..,[T_1])}{\partial(P,..,\mathsf{T})}=-\frac{\mathsf{M}}{R},$$

which holds in each sector of Γ.

It follows that the transformation is regular within each sector and can be inverted. In fact, it is not difficult to find the functions defining the inverse transformation. First, we show that

$$\overline{E}=\frac{R}{2}\left(1+\frac{q^2}{\mathsf{M}^2}-\frac{[E]^2}{\mathsf{M}^2}-\frac{\mathsf{M}^2}{4R^2}\right), \tag{29}$$

in each sector. To this aim, we substitute for q and M from Eqs. (6) and (20) into Eq. (29) and use the identities

$$F_++F_-=2\left(1-\frac{2\overline{E}}{R}\right),\quad F_+F_-=\left(1-\frac{2\overline{E}}{R}\right)^2-\frac{[E]^2}{R^2}. \tag{30}$$

In this way, we obtain

$$\frac{2\overline{E}}{R}-1=\frac{R^2}{\mathsf{M}^2}\left(-2\left(1-\frac{2\overline{E}}{R}\right)^2+2a_-b_+b_-\left(1-\frac{2\overline{E}}{R}\right)\sqrt{|F_+F_-|}\mathrm{sh}_{a_+a_-}\frac{P}{R}\right).$$

Now, the equality follows immediately from Eq. (6).

The next nontrivial part of the inverse transformation is the relation

$$P=\frac{R}{2}\ln\left|\frac{u_+v_-}{u_-v_+}\right|, \tag{31}$$

which again holds in each sector. To show Eq. (31), we use Eqs. (23) to obtain

$$\frac{u_+v_-}{u_-v_+}=\frac{(q-\mathsf{M}^2/2R)^2-[E]^2}{(q+\mathsf{M}^2/2R)^2-[E]^2}.$$

The substitution for q and M gives

$$q+\eta\frac{\mathsf{M}^2}{2R}=\eta(R-2\overline{E})-\eta a_{-\eta}b_+b_-R\sqrt{|F_+F_-|}e^{\eta P/R},$$

where $\eta=\pm 1$. After some rearrangement and applying Eqs. (30), we have

$$\left(q+\eta\frac{\mathsf{M}^2}{2R}\right)^2-[E]^2=-a_{-\eta}b_+b_-R^2\sqrt{|F_+F_-|}e^{\eta P/R}\left(F_++F_--2a_-b_+b_+\mathrm{sh}_{a_+a_-}\frac{P}{R}\right).$$

Then, Eq. (6) leads immediately to

$$\left(q+\eta\frac{\mathsf{M}^2}{2R}\right)^2-[E]^2=-a_{\eta}b_+b_-\mathsf{M}^2\sqrt{|F_+F_-|}e^{\eta P/R},$$

and this implies Eq. (31).

The rest of the inverse transformation is easy: the functions R and T are the same in both sets of variables,

$$E_\epsilon = \overline{E} + \frac{\epsilon}{2}[E] \tag{32}$$

and

$$T^\epsilon = \overline{T_1} + \frac{\epsilon}{2}[T_1] - 2E_\epsilon \ln\left|\frac{u_\epsilon}{v_\epsilon}\right|; \tag{33}$$

in the last equation, one has to substitute Eq. (32) for E_ϵ and Eqs. (23) for u_ϵ and v_ϵ.

The ranges of the variables q, R, M, T, $[E]$, $\overline{T_1}$ and $[T_1]$ in the physical region are implied by the conditions (18), (19) and Eq. (29). Let us work them out. Equation (29) together with $\overline{E}\geqslant 0$ implies that

$$[E]^2 \leqslant \mathsf{M}^2 + q^2 - \frac{\mathsf{M}^4}{4R^2}; \tag{34}$$

hence,

$$\mathsf{M}^2 + q^2 \geqslant \frac{\mathsf{M}^4}{4R^2}. \tag{35}$$

The two inequalities $E_+\geqslant 0$ and $E_-\geqslant 0$ are equivalent to $[E]^2 \leqslant 4\overline{E}^2$. Equation (29) implies that this inequality is, in turn, equivalent to

$$\left([E]^2 - \left(\mathsf{M}^2 + q^2 + \frac{\mathsf{M}^4}{4R^2}\right)\right)^2 \geqslant \left(\frac{\mathsf{M}^2}{R}\sqrt{\mathsf{M}^2+q^2}\right)^2. \tag{36}$$

From Eq. (34), it follows that

$$[E]^2 - \left(\mathsf{M}^2 + q^2 + \frac{\mathsf{M}^4}{4R^2}\right) \leqslant -\frac{\mathsf{M}^4}{2R^2} < 0.$$

Equation (36) is, therefore, equivalent to

$$[E]^2 \leqslant \left(\sqrt{\mathsf{M}^2+q^2} - \frac{\mathsf{M}^2}{2R}\right)^2. \tag{37}$$

Equation (35) shows that the inequality (37) is not weaker than (34). Hence, all information is contained in the following inequality:

$$|[E]| \leqslant \sqrt{\mathsf{M}^2+q^2} - \frac{\mathsf{M}^2}{2R}. \tag{38}$$

Inequality (38) together with $\mathsf{M}>0$ and $R>0$ define the ranges of the variables q, R, M, T, $[E]$, $\overline{T_1}$ and $[T_1]$ in the physical region.

V. SYMMETRY OF THE SYSTEM

In this section, we collect the results of Ref. 17 on the symmetry of the shell action and describe the action of the symmetry transformations on the new variables.

In Ref. 17, we have found that there is a continuous symmetry group generated by an arbitrary function of E_+ and E_-. A simple generalization of the argument given in Ref. 17 implies that the following finite transformation of the variables E_ϵ, u_ϵ, v_ϵ and $\tilde{P}_\epsilon$, which were used there, is a symmetry of the system

$$\tilde{P}_\epsilon \mapsto \tilde{P}_\epsilon - \epsilon \frac{\partial \Lambda}{4 \partial E_\epsilon}, \quad \epsilon = \pm 1,$$

the other variables being invariant, where $\Lambda(E_+, E_-)$ is an arbitrary smooth function of two variables (the factor 1/4 is introduced for convenience). In addition to this continuous infinitely dimensional group, there were two reflections, which we denote here by σ_1 and σ_2. σ_1 is a time reflection defined by

$$E_\epsilon \mapsto E_\epsilon, \quad u_\epsilon + u_\epsilon \mapsto -(u_\epsilon + v_\epsilon), \quad u_\epsilon - u_\epsilon \mapsto u_\epsilon - v_\epsilon, \quad \tilde{P}_\epsilon \mapsto -\tilde{P}_\epsilon,$$

and σ_2 is a left–right reflection defined by

$$E_+ \leftrightarrow E_-, \quad u_+ \leftrightarrow v_-, \quad u_- \leftrightarrow v_+, \quad \tilde{P}_+ \leftrightarrow -\tilde{P}_-.$$

These results can be easily expressed in our variables q, R, $[E]$, $\overline{T_1}$, $\overline{E}$ and $[T_1]$ and extended to M and T. Consider the function q; Eqs. (23) yield

$$q = -\mathsf{M}(u_\epsilon + v_\epsilon), \quad \epsilon = \pm 1.$$

It is clear that M is invariant with respect to the whole group; hence, q is invariant with respect to the Λ-transformation for any $\Lambda(E_+, E_-)$, it changes sign by the time reflection σ_1 and is invariant with respect to σ_2. The function R was written in terms of u_c and v_ϵ in Ref. 17 in the form

$$R = \frac{2E_\epsilon}{1 + u_\epsilon v_\epsilon}, \quad \epsilon = \pm 1.$$

Realizing that $4u_\epsilon v_\epsilon = (u_\epsilon + v_\epsilon)^2 - (u_\epsilon - v_\epsilon)^2$, we can see that R is an invariant of all above transformations. The function $[E]$ transforms nontrivially only by σ_2,

$$[E] \mapsto -[E],$$

and $\overline{E}$ is an invariant like R. A comparison of our Eq. (27) with Eq. (43) of Ref. 17 reveals that $T_1^\epsilon = 4\tilde{P}_\epsilon$, so that its Λ-transformation is

$$T_1^\epsilon \mapsto T_1^\epsilon - \epsilon \frac{\partial \Lambda}{\partial E_\epsilon}.$$

Using this formula, one finds easily that

$$\overline{T_1} \mapsto \overline{T_1} - \frac{\partial \Lambda}{\partial [E]}, \quad [T_1] \mapsto [T_1] - \frac{\partial \Lambda}{\partial \overline{E}}. \tag{39}$$

The time reflection σ_1 changes the signs of both times:

$$\overline{T_1} \mapsto -\overline{T_1}, \quad [T_1] \mapsto -[T_1],$$

but the left–right reflection σ_2 acts as follows:

$$\overline{T_1} \mapsto -\overline{T_1}, \quad [T_1] \mapsto [T_1].$$

From the physical meaning of M and T, it follows that they are invariant with respect to the continuous transformation for any $\Lambda(E_+,E_-)$ and of the left–right reflection σ_2, whereas the time reflection σ_1 must give

$$\mathsf{M}\mapsto\mathsf{M},\quad \mathsf{T}\mapsto-\mathsf{T}.$$

However, as T is a new cyclic coordinate, the new system has larger continuous symmetry: we can extend Λ to an arbitrary function of the three variables E_+, E_- and M, and define the action on T as follows:

$$\mathsf{T}\mapsto\mathsf{T}+\frac{\partial\Lambda}{\partial\mathsf{M}}. \tag{40}$$

Let us postpone the proof that this extended Λ-transformation is a symmetry to Sec. VI. It generates three independent constant shifts of the cyclic variables $\overline{T_1}$, $[T_1]$ and T that are different in different shell spacetimes.

We can see from these results that our new variables transform particularly simply. This was, in fact, the idea that helped to find the variables in the first place.

VI. TRANSFORMATIONS OF THE CARTAN FORM

Let us transform the Cartan form Θ in each sector of Γ to the variables q, R, M, T, $[E]$, $\overline{T_1}$ and $[T_1]$ and check that it can then be extended smoothly accross the boundaries of the sectors.

If we substitute for P, T^+, T^- and M, we obtain

$$\Theta=-\overline{E}d[T_1]-[E]d\overline{T_1}+\mathsf{M}\ d\mathsf{T}+2E_+^2d\left(\ln\left|\frac{u_+}{v_+}\right|\right)-2E_-^2d\left(\ln\left|\frac{u_-}{v_-}\right|\right)$$
$$+\ln\left|\frac{u_+v_-}{u_-v_+}\right|d\left(\frac{R^2}{4}\right)+\ln\left|\frac{u_+}{v_+}\right|d(E_+^2)-\ln\left|\frac{u_-}{v_-}\right|d(E_-^2).$$

The logarithms can be rearanged as follows:

$$\Theta=-\overline{E}d[T_1]-[E]d\overline{T_1}+\mathsf{M}\ d\mathsf{T}+d\left(\left(\frac{R^2}{4}+E_+^2\right)\ln\left|\frac{u_+}{v_+}\right|-\left(\frac{R^2}{4}+E_-^2\right)\ln\left|\frac{u_-}{v_-}\right|\right)$$
$$-\left(\frac{R^2}{4}-E_+^2\right)d\left(\ln\left|\frac{u_+}{v_+}\right|\right)+\left(\frac{R^2}{4}-E_-^2\right)d\left(\ln\left|\frac{u_-}{v_-}\right|\right).$$

The total differential of a function (that diverges badly at the horizons) can be left out. We obtain

$$\Theta=-\overline{E}d[T_1]-[E]d\overline{T_1}+\mathsf{M}\ d\mathsf{T}-\left[\left(\frac{R^2}{4}-E^2\right)d\left(\ln\left|\frac{u}{v}\right|\right)\right], \tag{41}$$

where we must substitute for E_+, E_-, $\overline{E}$, u_ϵ and v_ϵ from Eqs. (32), (29) and (23). This form of Θ has been obtained in Ref. 17 (without the matter term); there, however, u_ϵ and v_ϵ were regarded as independent variables. Observe that Eqs. (23), (29) and (32) imply

$$\frac{2E_\epsilon}{R}-1=u_\epsilon v_\epsilon, \tag{42}$$

so that u_ϵ and v_ϵ are not independent functions for $E_\epsilon=0$.

Now, we can show that the last term in Θ is regular: we just rewrite it using Eq. (42):

$$\left[\left(\frac{R^2}{4}-E^2\right)d\left(\ln\left|\frac{u}{v}\right|\right)\right]=\left[\frac{R^2}{4}\left(1+\frac{2E}{R}\right)(v\ du-u\ dv)\right]. \tag{43}$$

Thus, our new variables cover all of the constraint hypersurface Γ as promised.

Let us express Θ explicitly by means of the variables q, R, M and $[E]$. Employing Eqs. (23), we have

$$\overline{v\ du-u\ dv}=2\left(-\frac{[E]}{\mathsf{M}^2}\,dq+\frac{q}{\mathsf{M}^2}\,d[E]\right),$$

$$[v\ du-u\ dv]=2\left(\frac{1}{R}\,dq+\frac{q}{R^2}\,dR-\frac{2q}{R\mathsf{M}}\,d\mathsf{M}\right).$$

Then, applying the well-known identity $[AB]=\overline{A}[B]+\overline{B}[A]$ that holds for jumps and averages of any two functions A and B, and after using Eq. (29), we arrive at

$$\left[\left(\frac{R^2}{4}-E^2\right)d\left(\ln\left|\frac{u}{v}\right|\right)\right]=\frac{Rq[E]}{\mathsf{M}^2}\,d[E]-\left(\frac{2Rq}{\mathsf{M}}+\frac{Rq^3}{\mathsf{M}^3}-\frac{Rq[E]^2}{\mathsf{M}^3}-\frac{q\mathsf{M}}{4R}\right)d\mathsf{M}$$
$$+\left(R+\frac{Rq^2}{2\mathsf{M}^2}-\frac{3R[E]^2}{2\mathsf{M}^2}-\frac{\mathsf{M}^2}{8R}\right)dq+\left(q+\frac{q^3}{2\mathsf{M}^2}-\frac{q[E]^2}{2\mathsf{M}^2}-\frac{q\mathsf{M}^2}{8R^2}\right)dR. \tag{44}$$

This, together with Eq. (41), gives Θ as a function of new variables. It is a complicated one, so we have to transform the coordinates further in order to get a simple expression.

The following question will give some direction to this search for simplicity: is the variable $[T_1]$ suitable for the role of time? More concretely, the time levels in constraint surface should be transversal (this is shown in Ref. 10). A surface is transversal, if dynamical trajectories intersect it transversally and only once. We can see if the surface $[T_1]=\text{const}$ has this property as follows. The direction of motion on the constraint surface coincides with the direction of degeneration of the presymplectic form $d\Theta$; hence, the pull-back $d\Theta_f$ of $d\Theta$ to any transversal surface $f=\text{const}$ must be nondegenerate. If the surface is $2n$-dimensional, then the $2n$-form $d\Theta_f\wedge\cdots\wedge d\Theta_f$ must be everywhere nonzero. In our case, $n=3$. Let us calculate $d\Theta_{[T_1]}\wedge d\Theta_{[T_1]}\wedge d\Theta_{[T_1]}$. Equations (41) and (44) imply that

$$d\Theta_{[T_1]}=-d[E]\wedge d\overline{T}_1+d\mathsf{M}\wedge d\mathsf{T}+A_1dR\wedge dq+A_2dR\wedge d[E]$$
$$+A_3dR\wedge d\mathsf{M}+A_4dq\wedge d[E]+A_5dq\wedge d\mathsf{M}+A_6d[E]\wedge d\mathsf{M},$$

where A_i, $i=1,\dots,6$ are some functions of the variables R, q, $[E]$ and M. It is clear, therefore, that

$$d\Theta_{[T_1]}\wedge d\Theta_{[T_1]}\wedge d\Theta_{[T_1]}=-6A_1d[E]\wedge d\overline{T}_1\wedge d\mathsf{M}\wedge d\mathsf{T}\wedge dR\wedge dq,$$

and the points where the six-form vanishes coincide with those where A_1 vanishes. A_1 can be calculated from Eq. (44) with the result

$$A_1=\frac{q^2}{\mathsf{M}^2}+\frac{[E]^2}{\mathsf{M}^2}-\frac{\mathsf{M}^2}{4R^2}. \tag{45}$$

Thus, A_1 vanishes for

$$[E]^2=\frac{\mathsf{M}^4}{4R^2}-q^2.$$

Do these points lie in the physical region that is given by inequality (38)? Let us substitute the value of $[E]^2$ into Eq. (38) and rewrite the result in the form

$$\frac{\mathsf{M}^2}{R^2} \leqslant 4\left(1+\frac{q^2}{\mathsf{M}^2}\right)-\frac{3\mathsf{M}^2}{\mathsf{M}^2+q^2}.$$

This inequality implies inequality (35); hence, all zero points of A_1 lie in the physical region. We conclude that $[T_1]=\text{const}$ is *no* transversal surface.

Let us try to shift $[T_1]$ by a function dependent on the variables R, q, $[E]$ and M:

$$[T_1]=T_2+X(R,q,[E],\mathsf{M}).$$

Our motivation for choosing such a form is that T_2 is then transformed by the Λ-symmetry (39) in the same way as $[T_1]$ is, and that it is also a cyclic coordinate. Substitution for $[T_1]$ into $d\Theta$ changes the coefficient A_1 at $dR\wedge dq$ in $d\Theta_{T_2}$ by

$$A_1 \to A_1+\frac{\partial\overline{E}}{\partial q}\frac{\partial X}{\partial R}-\frac{\partial\overline{E}}{\partial R}\frac{\partial X}{\partial q}.$$

Comparing A_1, Eq. (45), with $\partial\overline{E}/\partial R$ calculated from Eq. (29),

$$\frac{\partial\overline{E}}{\partial R}=\frac{1}{2}+\frac{q^2}{2\mathsf{M}^2}-\frac{[E]^2}{2\mathsf{M}^2}+\frac{\mathsf{M}^2}{8R^2},$$

we observe that

$$A_1+2\frac{\partial\overline{E}}{\partial R}=2+2\frac{q^2}{\mathsf{M}^2},$$

which is always positive; thus, one possible shift is

$$[T_1]=T_2-2q. \tag{46}$$

If we perform this transformation in Θ, we find out easily that another shift,

$$\overline{T}_1=T_3+2\frac{Rq[E]}{\mathsf{M}^2}, \tag{47}$$

miraculously cancels some pesky cross terms in $d\Theta$. Indeed, shifting the ''times'' in the Cartan form, we have

$$-\overline{E}d[T_1]-[E]d\overline{T}_1+\mathsf{M}\,d\mathsf{T}=-\overline{E}\,dT_2-[E]dT_3+\mathsf{M}\,d\mathsf{T}-\left(R+\frac{Rq^2}{\mathsf{M}^2}-\frac{3R[E]^2}{\mathsf{M}^2}-\frac{\mathsf{M}^2}{4R}\right)dq$$
$$+\frac{2q[E]^2}{\mathsf{M}^2}\,dR+\frac{2qR[E]}{\mathsf{M}^2}\,d[E]-\frac{4qR[E]^2}{\mathsf{M}^3}\,d\mathsf{M}.$$

Using Eqs. (41) and (44), we obtain

$$\Theta=-\overline{E}\,dT_2-[E]dT_3+\mathsf{M}\,d\mathsf{T}-d\left(\frac{3qR[E]^2}{2\mathsf{M}^2}+\frac{q\mathsf{M}^2}{8R}\right)$$
$$+\frac{Rq^2}{2\mathsf{M}^2}\,dq-\left(q+\frac{q^3}{2\mathsf{M}^2}\right)dR+\frac{2R}{\mathsf{M}}\left(q+\frac{q^3}{2\mathsf{M}^2}\right)d\mathsf{M}.$$

The last three terms do not contain other variables than q, R and M and so the terms with $d[E]\wedge dR$, $d[E]\wedge dq$ and $d[E]\wedge d\mathsf{M}$ disappear from $d\Theta$. We have not found any simple physical or geometrical interpretation of the new times T_2 and T_3; observe that the shifts (46) and (47) ''mix the sides.''

The fact that the last three terms in Θ contain only three variables means that they can be simplified to just one term. For example,

$$\frac{Rq^2}{2\mathsf{M}^2}\,dq-\left(q+\frac{q^3}{2\mathsf{M}^2}\right)dR+\frac{2R}{\mathsf{M}}\left(q+\frac{q^3}{2\mathsf{M}^2}\right)d\mathsf{M}=d\left(\frac{Rq^3}{6\mathsf{M}^2}\right)-\left(q\mathsf{M}^2+\frac{2q^3}{3}\right)d\left(\frac{R}{\mathsf{M}^2}\right).$$

This suggests the second step of our transformation:

$$p:=q\mathsf{M}^2+\frac{2q^3}{3}, \tag{48}$$

$$x:=\frac{R}{\mathsf{M}^2}. \tag{49}$$

We arrive so at the final shape of the Cartan form,

$$\Theta=-\overline{E}\;dT_2-[E]dT_3+\mathsf{M}\;d\mathsf{T}-p\;dx, \tag{50}$$

where $\overline{E}$ is given by Eq. (29) in which q and R are expressed in terms of p and x; such expressions can be obtained by solving Eqs. (48) and (49) for q and R.

Consider Eq. (48). If we keep M fixed, p is an increasing function of q in the whole interval $(-\infty,\infty)$, for

$$\frac{\partial p}{\partial q}=\mathsf{M}^2+2q^2>0.$$

Hence, the function maps the range $(-\infty,\infty)$ of q onto the range $(-\infty,\infty)$ of p in a bijective way, and it possesses a unique inverse. One can explicitly write down this inverse using second and third roots; this will make $\overline{E}$ a complicated function of p, x, M and $[E]$ [we shall write down this function later; cf. Eq. (63)].

Equation (50) implies that

$$d\Theta_{T_2}\wedge d\Theta_{T_2}\wedge d\Theta_{T_2}=6d[E]\wedge dT_3\wedge d\mathsf{M}\wedge d\mathsf{T}\wedge dp\wedge dx,$$

$$d\Theta_{T_3}\wedge d\Theta_{T_3}\wedge d\Theta_{T_3}=6\,\frac{\partial\overline{E}}{\partial[E]}\;d[E]\wedge dT_2\wedge d\mathsf{M}\wedge d\mathsf{T}\wedge dp\wedge dx,$$

$$d\Theta_{T}\wedge d\Theta_{T}\wedge d\Theta_{T}=-6\,\frac{\partial\overline{E}}{\partial\mathsf{M}}\;d[E]\wedge dT_3\wedge d\mathsf{M}\wedge dT_2\wedge dp\wedge dx,$$

and Eq. (29) gives

$$\frac{\partial\overline{E}}{\partial[E]}=-\frac{R}{\mathsf{M}^2}\,[E],$$

$$\frac{\partial\overline{E}}{\partial\mathsf{M}}=-\frac{\mathsf{M}}{4R}.$$

Thus, the surface $T_3=\text{const}$ is not transversal at $[E]=0$, but $T_2=\text{const}$ and $\mathsf{T}=\text{const}$ are transversal everywhere.

The Cartan form (50) is not yet very simple, because $\overline{E}$ is a complicated function of p, x, $[E]$ and M. What we can still do is to extend the phase space to an eight-dimensional manifold with the coordinates $\overline{E}$, $[T_2]$, $[E]$, $\overline{T}_2$, M, T, p and x, and to express the constraint between these variables as a polynomial. Indeed, applying Eq. (49), we can rewrite Eq. (29) as follows:

$$q^2=\frac{1}{4x^2}+\frac{2\overline{E}}{x}+[E]^2-\mathsf{M}^2;$$

squaring Eq. (48) and substituting the above expression for q^2 into the result, we obtain

$$p^2=\frac{1}{9}\left(\frac{1}{4x^2}+\frac{2\overline{E}}{x}+[E]^2-\mathsf{M}^2\right)\left(\frac{1}{2x^2}+\frac{4\overline{E}}{x}+2[E]^2+\mathsf{M}^2\right)^2. \tag{51}$$

This is a constraint that is polynomial in the momenta p, $\overline{E}$, $[E]$ and M. The inequalities defining the ranges of the variables $\overline{E}$, $[T_2]$, $[E]$, $\overline{T}_2$, M, T, p and x in the physical region are also simplified:

$$[E]^2\leqslant 4\overline{E}^2,\quad \mathsf{M}>0,\quad x>0. \tag{52}$$

Let us, finally, find the transformation of the variables p, x, M, T, $\overline{E}$, T_2, $[E]$ and T_3 by the symmetry group. The results are as follows. We have for the Λ-transformation,

$$p\mapsto p,\quad x\mapsto x,\quad \mathsf{M}\mapsto\mathsf{M},\quad \mathsf{T}\mapsto\mathsf{T},\quad \overline{E}\mapsto\overline{E},\quad [E]\mapsto[E],$$

$$T_2\mapsto T_2-\frac{\partial\Lambda}{\partial\overline{E}},\quad T_3\mapsto T_3-\frac{\partial\Lambda}{\partial[E]},\quad \mathsf{T}\mapsto\mathsf{T}+\frac{\partial\Lambda}{\partial\mathsf{M}};$$

for the time reflection σ_1,

$$p\mapsto -p,\quad x\mapsto x,\quad \mathsf{M}\mapsto\mathsf{M},\quad \mathsf{T}\mapsto-\mathsf{T},\quad \overline{E}\mapsto\overline{E},\quad [E]\mapsto[E],\quad T_2\mapsto -T_2,\quad T_3\mapsto -T_3;$$

and for the left–right reflection σ_2,

$$p\mapsto p,\quad x\mapsto x,\quad \mathsf{M}\mapsto\mathsf{M},\quad \mathsf{T}\mapsto\mathsf{T},\quad \overline{E}\mapsto\overline{E},\quad [E]\mapsto-[E],\quad T_2\mapsto T_2,\quad T_3\mapsto -T_3.$$

Thus, the simple form of the symmetry transformations is preserved by the shifts (46) and (47) as well as by transformations (48) and (49). Equation (50) implies that

$$\Theta\mapsto\Theta+d\left(\frac{\partial\Lambda}{\partial\overline{E}}\,\overline{E}+\frac{\partial\Lambda}{\partial[E]}\,[E]+\frac{\partial\Lambda}{\partial\mathsf{M}}\,\mathsf{M}-\Lambda\right),$$

by a Λ-transformation, so our extended Λ-transformation is a symmetry,

$$\Theta\mapsto-\Theta,$$

by the time reflection σ_1, and

$$\Theta\mapsto\Theta,$$

by the left–right reflection σ_2. (The sign change of Θ by σ_1 is necessary, because time reflections are anti-symplectic transformations.) The constraint equation (51) is clearly invariant of the group. Thus, the action of the symmetry group is well-defined in the whole phase space and it preserves the physical region.

VII. THE CONSTRAINT SURFACE

In this section, we are going to study the properties of the constraint (51). We are interested in the topology of the constraint surface; we shall investigate the salient question of whether or not any spurious solutions have been unintentionally included during the process of transforming the constraint to the polynomial form; and we shall look for solutions of the constraint (51) with respect to some momenta.

Our new variational principle reads as follows:

$$S=\int dt(-\overline{E}\dot{T}_2-[E]\dot{T}_3+\mathsf{M}\dot{\mathsf{T}}-p\dot{x}-\nu_1\mathcal{C}_1), \tag{53}$$

where

$$\mathcal{C}_1=9p^2-\left(\frac{1}{4x^2}+\frac{2\overline{E}}{x}+[E]^2-\mathsf{M}^2\right)\left(\frac{1}{2x^2}+\frac{4\overline{E}}{x}+2[E]^2+\mathsf{M}^2\right)^2. \tag{54}$$

The constraint surface is trivial in the direction of the three times T_2, T_3 and T, so we can limit ourselves to its projection to the five-dimensional space spanned by $\overline{E}$, $[E]$, M, p and x. Let us observe that $\mathcal{C}_1$ depends on $\overline{E}$, $[E]$ and x through a function that we shall call B,

$$B:=\frac{1}{4x^2}+\frac{2\overline{E}}{x}+[E]^2, \tag{55}$$

and that the gradient of B is nonzero. It is, therefore, advantageous, to write $\mathcal{C}_1$ as a composed function,

$$\mathcal{C}_1=9p^2-(B-\mathsf{M}^2)(2B+\mathsf{M}^2)^2.$$

Let us first look for the singular points of the constraint surface (that is, where the gradient of $\mathcal{C}_1$ vanishes). We have

$$\frac{\partial\mathcal{C}_1}{\partial p}=18p, \tag{56}$$

$$\frac{\partial\mathcal{C}_1}{\partial\mathsf{M}}=6\mathsf{M}^3(2B+\mathsf{M}^2), \tag{57}$$

$$\frac{\partial\mathcal{C}_1}{\partial B}=-3(2B+\mathsf{M}^2)(2B-\mathsf{M}^2). \tag{58}$$

As all three derivatives must vanish at a singular point, all such points are determined by the equations

$$p=0,\quad 2B+\mathsf{M}^2=0. \tag{59}$$

All points satisfying Eqs. (59) are solutions of Eq. (51); however, the physical ranges (52) of the variables $\overline{E}$, $[E]$ and x do not allow B to be negative. We conclude that the constraint surface is regular (smooth) in the physical region (52).

The topology of the constraint surface can be found in the shortest way, if we consider $\mathcal{C}_1$ as a function $\mathcal{C}_1(B)$ of B keeping all other variables constant. Studying Eqs. (54) and (58) we can easily see that $\mathcal{C}_1(B)$ decreases in the interval $(-\infty,-\mathsf{M}^2/2)$ from ∞ to $9p^2$, it increases in the interval $(-\mathsf{M}^2/2),\mathsf{M}^2/2)$, from $9p^2$ to $9p^2+2\mathsf{M}^6$ and it decreases again in the interval $(\mathsf{M}^2/2,\infty)$ from $9p^2+2\mathsf{M}^6$ to $-\infty$. It follows that there is only one solution for $p\neq 0$ [because $\mathcal{C}_1(B)=9p^2$ for $B=\mathsf{M}^2$]. As $\mathcal{C}_1(\mathsf{M}^2)=9p^2>0$, the solution must satisfy the inequality $B>\mathsf{M}^2$. This solution depends smoothly on the values of all other variables. For $p=0$, there are, however, two solutions: one with $B=\mathsf{M}^2$, which is a continuous extension of the previous $p\neq 0$ case, and a new one with $B=-\mathsf{M}^2/2$, which appears when the local minimum at $-\mathsf{M}^2/2$ touches the B-axis. This shows that the constraint surface consists of two components: one is a seven-dimensional smooth manifold, let us denote it by Γ_7, that lies in the region $B\geqslant\mathsf{M}^2$, and one is a six-dimensional smooth manifold, let us denote it by Γ_6, that is defined by Eqs. (59).

Next, let us study the solvability of the constraint equation (51) with respect to the conserved momenta. The unique solution of the constraint (51) with respect to B at Γ_7 is given by

$$B=f(p,\mathsf{M}), \tag{60}$$

where

$$f(p,\mathsf{M}):=\tfrac{1}{2}\sqrt[3]{9p^2+\mathsf{M}^6+3p\sqrt{9p^2+2\mathsf{M}^6}}+\tfrac{1}{2}\sqrt[3]{9p^2+\mathsf{M}^6-3p\sqrt{9p^2+2\mathsf{M}^6}}. \tag{61}$$

The equation (51) has, in fact, three independent solutions, but only one of them is real. For $p=0$, B reaches its minimum M^2 on Γ_7; thus, at Γ_7,

$$B\geqslant\mathsf{M}^2. \tag{62}$$

Given B, one can solve uniquely for $\overline{E}$ and there is also a unique solution for $[E]^2$. The solution for $\overline{E}$ is

$$\overline{E}=-\frac{1}{8x}-\frac{x[E]^2}{2}+\frac{x}{2}f(p,\mathsf{M}). \tag{63}$$

Similarly, we can find the solution of the constraint (51) with respect to M. Consider $\mathcal{C}_1$ as a function $\mathcal{C}_1(\mathsf{M})$ of M, keeping all other variables fixed. $\mathcal{C}_1(\mathsf{M})$ is symmetric with respect to the reflection on the $\mathcal{C}_1$-axis. We have to distinguish two cases: $\underline{B<0,\ B=0:}$ $\mathcal{C}_1(\mathsf{M})$ decreases in the interval $(-\infty,2B)$ from ∞ to $9p^2$, it increases in the interval $(2B,0)$ from $9p^2$ to the local maximum $9p^2-4B^3$ at $\mathsf{M}=0$, it decreases again in the interval $(0,-2B)$ from $9p^2-4B^3$ to $9p^2$ and, finally, it increases in $(-2B,\infty)$ from $9p^2$ to ∞. A solution exists only for $p=0$, and there are two solutions; then $\mathsf{M}=\pm\sqrt{-2B}$. This is at the surface Γ_6. $\underline{B>0:}$ then, there is only one minimum, at $\mathsf{M}=0$, with the value $9p^2-4B^3$; there are no other extrema. Consider Eq. (51) as an equation for M^{-2}. There are three real solutions for M^{-2}, but all of them are negative unless

$$4B^3-9p^2\geqslant 0, \tag{64}$$

and then there is only one non-negative solution. A reasonable solution is, therefore, unique; it depends continuously of the other variables and it lies at Γ_7. We can write it in the form

$$\mathsf{M}=\pm\sqrt[4]{\frac{4B^3-9p^2}{4B}}\left(\cos\frac{1}{3}\arctan\frac{3p}{\sqrt{4B^3-9p^2}}\right)^{-1/2}.$$

This solution can, of course, be also expressed by means of square and third roots, but only if one employs complex numbers. We can see that squaring of M introduced nonphysical solutions with negative M; we have to choose the positive branch. Observe that all points at the surface Γ_7 must satisfy the inequalities (62) and (64).

The equations of motion that follow from the action (53) can be written as follows:

$$\dot{\overline{E}}=[\dot{E}]=\dot{\mathsf{M}}=0, \tag{65}$$

$$\dot{T}_2=-\nu_1\frac{\partial\mathcal{C}_1}{\partial B}\frac{\partial B}{\partial\overline{E}}, \tag{66}$$

$$\dot{T}_3=-\nu_1\frac{\partial\mathcal{C}_1}{\partial B}\frac{\partial B}{\partial[E]}, \tag{67}$$

$$\dot{\mathsf{T}}=\nu_1\frac{\partial\mathcal{C}_1}{\partial\mathsf{M}}, \tag{68}$$

$$\dot{p}=\nu_1\frac{\partial\mathcal{C}_1}{\partial B}\frac{\partial B}{\partial x}, \tag{69}$$

$$\dot{x}=-\nu_1\frac{\partial\mathcal{C}_1}{\partial p}. \tag{70}$$

It is easy to derive the radial equation from them. On Γ_6, we have only a static solution:

$$\dot{x}=\dot{T}=\dot{p}=\dot{T}_2=\dot{T}_3=0.$$

On Γ_7, $\dot{\mathsf{T}}>0$ for $\nu_1>0$ and Eqs. (57), (68) and (70) yield

$$\frac{dx}{d\mathsf{T}}=-\frac{3p}{\mathsf{M}^3(2B+\mathsf{M}^2)}.$$

Using the constraint (51) to exclude p, we obtain the radial equation on Γ_7:

$$\frac{dx}{d\mathsf{T}}=\pm\frac{1}{\mathsf{M}^2}\sqrt{\frac{B}{\mathsf{M}^2}-1}. \tag{71}$$

Let us compare it with Eq. (13). Eqs. (65) and (49) imply that

$$\frac{dx}{d\mathsf{T}}=\frac{1}{\mathsf{M}^2}\frac{dR}{d\mathsf{T}},$$

whereas Eqs. (55), (49) and (14) give

$$V=1-\frac{B}{\mathsf{M}^2}.$$

Hence, the radial equations (71) and (13) are equivalent, and the dynamics of the new action on Γ_7 coincides with that of the old action. Γ_6 consists of the unintentionally added unphysical solutions, at least on the phase space defined by Eqs. (18), but nothing is added in the physical region.

Finally, we shall study the question of the monotonicity of the time functions T_2, T_3 and T along the dynamical trajectories at Γ_7. Equations (57) and (58) show that the right-hand sides of Eqs. (66) and (68) cannot vanish, and we have for $\nu_1>0$,

$$\dot{T}_2>0, \quad \dot{\mathsf{T}}>0.$$

Hence, T_2 and T are good times and, as we have seen above, the constraint can be solved uniquely for the corresponding momenta, $\overline{E}$ and M. As for T_3, Eq. (67) implies that $\dot{T}_3>0$ for $[E]>0$, $\dot{T}_3=0$ for $[E]=0$ and $\dot{T}_3<0$ for $[E]<0$. The solvability with respect to $[E]$ is also only partial: there is a unique positive and a unique negative solution. Thus T_3 is a good time only in some special cases, as the next section will show.

VIII. SOME INTERESTING SPECIAL CASES

In this section, we shall remove some degrees of freedom and describe the resulting simpler models in terms of the new variables.

A. Dust degrees of freedom removed

Here, we remove the variables M and T and return to the system considered in Ref. 14. It has two degrees of freedom: the black hole mass E_- and the position x of the shell. To this aim, we first choose a particular value of M and demote it so to a mere parameter. In this way, a submanifold Γ_{7M} of Γ_7 emerges. Second, we take a quotient of Γ_{7M} by the T-curves (see Ref. 17 for more details of this *exclusion of a cyclic variable*). The system that results has the action

$$S_M=\int dt(-\overline{E}\dot{T}_2-[E]\dot{T}_3-p\dot{x}-\nu_1\mathcal{C}_1),$$

where $\mathcal{C}_1$ is given by Eq. (54) as before, only M is a parameter now. The equations of motion comprise Eq. (65) without $\dot{\mathsf{M}}=0$, as well as Eqs. (66), (67), (69) and (70).

The properties of this system are analogous to that of the original one. In particular, the constraint is regular everywhere on Γ_{7M} because the derivative (56) and (58) cannot both simultaneously vanish, T_2 is a good time variable and the constraint equation is uniquely solvable for $\overline{E}$.

B. Black hole degree of freedom removed

Here, we demote the variable E_- to a parameter and remove the corresponding cyclic variable $T_2/2-T_3$; the remaining two degrees of freedom are the internal energy M and the position x of the shell. The system obtained in this way will describe the dynamics of the dust shell in the field of a fixed black hole if $E_->0$ or in the Minkowski space if $E_-=0$. One returns so to the system considered in Ref. 13; the procedure, in different coordinates, has been performed in Ref. 17.

To begin with, we transform variables as follows:

$$\overline{E}=\frac{E_++E_-}{2},\quad T_2^+=T_2/2+T_3,$$

$$[E]=E_+-E_-,\quad T_2^-=T_2/2-T_3.$$

Then, we choose a particular value for E_-; this defines the black hole mass, and formally, a six-dimensional submanifold Γ_{7E} of Γ_7. Finally, we take the quotient Γ_{7E}/T_- of Γ_{7E} by the T_--curves. The result is the action

$$S_E=\int dt(-E_+\dot{T}_2^++\mathsf{M}\dot{\mathsf{T}}-p\dot{x}-\nu_1\mathcal{C}_2),$$

where

$$\mathcal{C}_2=9p^2-(B_2-\mathsf{M}^2)(2B_2+\mathsf{M}^2)^2,$$

and

$$B_2 = \frac{1}{4x^2} + \frac{E_+ + E_-}{x} + (E_+ - E_-)^2.$$

The variable T is a good time and $C_2 = 0$ is, of course, solvable for M exactly as in the general case. The variable T_2^+ is, for large values of the parameter E_-, not a good time, because it is a combination $T_2/2 + T_3$ of T_2 that increases along all dynamical trajectories and T_3 that increases along those with $E_+ > E_-$ and decreases along those with $E_+ < E_-$. One can show that $\dot{T}_2^+$ is positive for $E_+ - E_- > 0$ and negative for $E_+ - E_- < -\mathsf{M}$ and that there is no surface for $E_- > \mathsf{M}$ that would be transversal to both the dynamical trajectories and the orbits of the continuous subgroup. [Still, one can transform T_2^+ so that the resulting time levels are transversal to the dynamical trajectories, but a mere shift is not sufficient; one must screw the level to a helix form with the helix axis in the E_+-direction; this may be done by a singular ''shift'' proportional to $(E_+ - E_-)^{-1}$. We shall not go into detail.] The derivative of $\mathcal{C}_2$ with respect to E_+ changes sign somewhere if E_- is sufficiently large and so there are more than one, or no solution of the constraint with respect to E_+.

There is one prominent exception: suppose that $E_- = 0$. Then, $E_+ - E_-$ is positive if E_+ is. Hence, in the physical region (where $E_+ > 0$), T_2^+ is a good time and the constraint has a unique solution for E_+ there, namely,

$$E_+ = -\frac{1}{2x} + \sqrt{f(p, \mathsf{M})},$$

where f is defined by Eq. (61).

One can also remove all four variables T, M, E_- and T_2^- so that only one degree of freedom, the position x of the shell, remains, but it is not difficult to work out the properties of the system using the results obtained in this section, and we left it as an easy exercise to the reader.

The special case $E_- = 0$ is also interesting, because it is a part of the boundary of the physical region given by inequalities (52) on Γ_7. The variational principle (53) does not break down at the boundary, that is, it defines a regular dynamics there. From this, it must clearly follow that the action defines a regular dynamics even at those points of Γ_7 that do not satisfy the conditions (52). We have studied the meaning of this dynamics and of these points in Sec. II.

ACKNOWLEDGMENTS

Important discussions with Karel V. Kuchař are acknowledged. This work was supported in part by the Swiss Nationalfonds, and by the Tomalla Foundation, Zurich.

[1] E. Farhi, A. H. Guth, and J. Guven, Nucl. Phys. B **339**, 417 (1990).
[2] P. Hájíček, B. S. Kay, and K. V. Kuchař, Phys. Rev. D **46**, 5439 (1992).
[3] T. Dray and G. 't Hooft, Commun. Math. Phys. **99**, 613 (1985).
[4] P. Kraus and F. Wilczek, Nucl. Phys. B **433**, 403 (1995).
[5] R. Dautcourt, Math. Nachr. **27**, 277 (1964).
[6] W. Israel, Nuovo Cimento B **44**, 1 (1966); **48**, 463 (1967).
[7] V. A. Berezin, Phys. Lett. B **241**, 194 (1990).
[8] P. Hájíček and J. Bičák, Phys. Rev. D **56**, 4706 (1997).
[9] J. L. Friedman, J. Louko, and S. N. Winters-Hilt, Phys. Rev. D **56**, 7674 (1997).
[10] P. Hájíček, Commun. Math. Phys. **150**, 545 (1993).
[11] P. Hájíček and J. Kijowski, Phys. Rev. D **57**, 914 (1998).
[12] A. Ansoldi, A. Aurilia, R. Balbinot, and E. Spalluci, Class. Quantum Grav. **14**, 2727 (1997).
[13] K. V. Kuchař, ''Time for space: canonical dynamics of gravitating shells,'' in preparation.
[14] P. Hájíček, Phys. Rev. D **57**, 936 (1998).
[15] P. Hájíček and K. V. Kuchař, Phys. Rev. D **41**, 1091 (1990).
[16] P. Hájíček and K. V. Kuchař, J. Math. Phys. **31**, 1723 (1990).
[17] P. Hájíček, Phys. Rev. D **58**, 084005 (1998).
[18] P. Hájíček, Nucl. Phys. B (Proc. Suppl.) **57**, 115 (1997).
[19] J.-M. Souriau, Structure des Systèmes Dynamiques (Dunod, Paris, 1970).

Asymptotic behavior in polarized T^2-symmetric vacuum space–times

James Isenberg[a)]
Department of Mathematics and Institute for Theoretical Science, University of Oregon, Eugene, Oregon 97403

Satyanad Kichenassamy[b)]
Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstrasse 22-26, D-04103 Leipzig, Germany

(Received 23 March 1998; accepted for publication 20 July 1998)

We use the Fuchsian algorithm to study the behavior near the singularity of a class of solutions of Einstein's vacuum equations. These solutions admit two commuting spacelike Killing fields like the Gowdy space–times, but their twist does not vanish. The space–times are also polarized in the sense that one of the ''gravitational degrees of freedom'' is turned off. Examining an analytic family of solutions with the maximum number of arbitrary functions, we find that they are all asymptotically velocity-term dominated as one approaches the singularity. © *1999 American Institute of Physics.* [S0022-2488(99)00601-5]

I. INTRODUCTION

There is increasing evidence that cosmological solutions exhibit rather special dynamical behavior in the neighborhood of their singularities. The evidence is still essentially limited to families of solutions with at least one Killing field. However, it is quite striking that although the Hawking–Penrose singularity theorems[1] require nothing more than geodesic incompleteness in generic cosmological solutions, every study to date indicates that the solutions under investigation are either ''asymptotically velocity-term dominated'' (AVD) or show ''Mixmaster'' behavior (see Refs. 2–8).

In a space with AVD behavior, the metric tensor $g_{ab}(\mathbf{x},t)$ evolves in such a way that an observer with fixed $\mathbf{x}_0$ moving toward the singularity sees the dynamics of $g_{ab}(\mathbf{x}_0,t)$ asymptotically approach that of a Kasner space–time, with there being generally a different Kasner limit for each different $\mathbf{x}_0$ (see Refs. 9, 3, 5, and 7, and references therein). Mixmaster behavior is similar, except that this observer sees $g_{ab}(\mathbf{x}_0,t)$ move through an infinite sequence of Kasner epochs, with regular intermittent bounces from one epoch to another (see Ref. 10). Again, different observers generally see different sequences (see, for instance, Refs. 3 and 8). While neither AVD nor Mixmaster behavior as described above is trivial, the Einstein equations, even with the simplification of an assumed symmetry, are sufficiently complicated that the prevalence of these special behaviors is quite remarkable.

The earliest verifications of AVD behavior in a family of inhomogeneous solutions, the polarized Gowdy space–times, took the form of a theorem.[5,11] The techniques developed in proving that result have not, however, been readily extended to more general families. Instead, most of the recent evidence for AVD and Mixmaster behavior in cosmological space–times has been based on numerical work: Berger and Moncrief[12] provide strong numerical evidence for AVD behavior in general ($\mathbf{T}^3$) Gowdy space–times, but find that the Kasner exponents should satisfy some inequalities in generic solutions (the solutions should be ''low-velocity''); they also

[a)]Electronic mail: jim@newton.uoregon.edu
[b)]Electronic mail: satyanad.kichenassamy@univ-reims.fr Permanent address: Laboratoire de Mathématiques, Université de Reims, Moulin de la Housse, B. P. 1039, F-51687 Reims Cedex 2, France.

have evidence in polarized $\mathbf{U}(1)$-symmetric space–times.[13] One should note that it is not always easy to be sure, in numerical computations, that the constraint equations do hold, except in the Gowdy class. Note also that Weaver, Berger, and Isenberg[8] provide similar evidence that locally $\mathbf{T}^2$-symmetric space–times with certain magnetic fields have Mixmaster behavior.

This numerical evidence motivates the search for a theoretical explanation for the prevalence of these behaviors and numerical observations such as the distinction between high- and low-velocity solutions, and, if possible, a means to predict which behavior occurs. The recent work of Kichenassamy and Rendall[7] introduces a new tool for obtaining such information. They use the Fuchsian algorithm to prove that there is a family of general (non-polarized) Gowdy space–times parametrized by the maximum number of free functions, namely four, which all exhibit AVD ''low-velocity'' behavior. If the derivative of one of these functions vanishes, ''high-velocity'' behavior is allowed. This family of solutions includes all of the previously known solutions in this class. The results also shed new light on other features of the numerical computations.

It is very likely that one can show that these new Gowdy space–times are stable under smooth perturbation of Cauchy data, by adapting the techniques described in Refs. 14 and 15. The general strategy consists in showing, using the Nash–Moser implicit function theorem, that the free functions which determine the solutions given by the Fuchsian algorithm can be used to parametrize solutions much in the same way as one uses Cauchy data on a hypersurface to label regular solutions. In a sense, one therefore generates systematically an ''asymptotic phase space'' for families of solutions, as was called for in Ref. 5.

In this work, we show that the Fuchsian algorithm is an effective tool for proving that AVD behavior occurs in a wider class of space–times: those which possess, like the Gowdy space–times, a $\mathbf{T}^2$ isometry group with spacelike generators, but in which, unlike the Gowdy case, the Killing vectors have a nonvanishing twist. The main new difficulty is that this nonvanishing twist prevents the constraint equations from decoupling from the evolution equations, resulting in a considerably more complicated partial differential equation (PDE) system than what obtains in the case of Gowdy space–times.[16–18] This difficulty is overcome by abandoning the separation of constraint and evolution equations. It is found that, combining some of the constraints with some of the ''evolution'' equations, one can form a system which is sufficient to determine the metric. One then proves directly that the remaining constraints hold everywhere if they hold asymptotically at the singularity. This latter condition can be expressed explicitly in terms of the data which determine the asymptotics at the singularity, or ''singularity data'' for short.

The Fuchsian algorithm has been extensively studied and takes a variety of forms (see Refs. 15, 19, 20, and 7). In Sec. II, we briefly review the form of the algorithm we use here, and a few relevant results we will need. Next, we describe in Sec. III the $\mathbf{T}^2$-symmetric space–times, noting some of their properties and defining the polarized subfamily. Then, in Sec. IV, we propose an AVD ansatz for the metric coefficients and show that the ''regular part'' of the field is indeed negligible in comparison with the leading terms. Finally, we discuss in Sec. V our conclusions and plans for future work.

II. THE FUCHSIAN ALGORITHM

The Fuchsian algorithm was initially developed to understand the behavior of solutions of differential equations in the neighborhood of a possible singularity of unknown location. The rationale was that if singularities are to form, it would be desirable to figure out by what mechanism they form: Which components of the solution become singular? Do singularities occur only in higher derivatives? Is the locus of the singularity arbitrary? How does it vary with Cauchy data given on a surface where the solution is smooth?

Existing results prior to Fuchsian techniques gave some information on the time of the first singularity, but did not shed light on the mechanism of singularity formation, except for special classes of singularities, such as shock waves in low dimensions, or caustic formation.

The questions asked above would be answered if it were possible to establish an expansion of the solution to relatively high order. To achieve this, one needs to establish a formal solution, and to prove that this formal expansion does characterize the solution. In practice, one is not primarily interested in the convergence or divergence of a series representation. Rather, one would like to know whether the parameters entering in a formal series representation do *determine uniquely* the solution, or whether there are infinitely many solutions differing from each other by, say, exponentially small corrections.

The Fuchsian method tackles this problem by seeking a reduction of the given system of PDEs to a Fuchsian system; that is, one which has a regular singular point with respect to one of the variables, which we call t. Using a change of coordinates if necessary, one may assume that the locus of the singularity is $t=0$. It is also possible to set things up so that one always deals with *first-order* Fuchsian systems, by the introduction of new variables. This is a familiar procedure for the Cauchy problem: for instance, if u solves the wave equation in Minkowski space, it is easy to check that the quantities $(u,\partial_a u)$ satisfy a first-order system.

Let us consider a PDE system which we write symbolically as

$$F[u]=0.$$

The exact form of the nonlinearity is not important for what follows. Generally, u can have any number of components.

Schematically, the Fuchsian algorithm has three parts:

Step 1. Identify the *leading part* of the desired expansion for u. This can be done in many cases by seeking a leading balance; that is, a leading term $a(t)$ such that, upon substitution into the equation, the most singular terms cancel each other.

Step 2. Introduce a *renormalized unknown*. This means that one writes

$$u=a(t)+t^s v(t), \tag{1}$$

where v is the new unknown. It is generally useful to compute a to relatively high order if possible, so that any arbitrary functions in the expansion are already included in a. If a is a solution up to order n, one may usually take $s=n+\varepsilon$, where ε is small.

Step 3. Obtain and solve a *Fuchsian system* for v. Indeed, one finds under rather general circumstances that the function v solves an equation of a very particular form, namely

$$(t\partial_t+A)v=t^\varepsilon f(t,x^\rho,v,\nabla_\rho v), \tag{2}$$

where ρ stands for spatial indices in this formula. The matrix A depends at most on the spatial variables, and f is, say, bounded. By taking t^ε to be a new time variable, one may always assume that ε is equal to one. Observe how spatial derivatives are effectively switched off from the equation when t goes to zero: *the Fuchsian algorithm provides a systematic procedure to guarantee AVD behavior.*

There is a variety of existence results for Fuchsian systems.[19,21,7] For our purposes, it suffices to note the following (see, Refs. 7 and 19).

Theorem 1: *There is a unique local solution which is continuous in time and analytic in space, and vanishes as t goes to zero, provided that (a) f is continuous in t and analytic in its other arguments and satisfies an estimate of the form*

$$|f(t,x^\rho,v,\nabla_\rho v)-f(t,x^\rho,w,\nabla_\rho w)|\leqslant C[|v-w|+|\nabla_\rho v-\nabla_\rho w|]$$

for some constant C provided v and w are bounded, and (b) the matrix $\sigma^A[=\exp(A\ln\sigma)]$ *is uniformly bounded for* $0<\sigma<1$.

Condition (b) is usually most conveniently checked by simply computing the matrix exponential.

We emphasize that we are not allowed to prescribe arbitrarily the initial value of v. The free data (usually called ''singularity data'') which label the solution u are already built into the choice of a in (1), and are subsequently incorporated into the function f in (2). A straightforward extension of the theorem can be made if we assume only that $v(0)$ belongs to the null space of A. By considering the equation satisfied by $v-v(0)$, one can reduce the problem to an equation to which the theorem applies. In such a case, $v(0)$ must be added to the list of singularity data.

General strategies for carrying out the algorithm can be found in Refs. 15, 20, 19, and 7, with applications to several examples. Let us simply describe here what these steps entail for the first PDE to which these ideas were applied successfully:

$$\eta^{ab}\partial_{ab}u=e^{u} \tag{3}$$

in Minkowski space, where u is a scalar field (there are similar results for power nonlinearities as well). Let $t=\psi(x)$ be the locus of the (yet unknown) singularity, and let x stand for the spatial variables. For one space dimension, Eq. (3) has a closed-form solution (''Liouville field theory''); however, we allow here the number of space dimensions to be arbitrary. Let $T=t-\psi(x)$. If we choose the leading part of u so that $\exp(u)\sim\varphi(x)T^{s}$ where s and φ are unknown, one readily finds that, to eliminate the most singular term in the expansion of (3), we need to choose $s=-2$ and $\varphi=2(1-|\nabla_x\psi|^2)$ which must therefore be positive. Hence the leading part of u takes the form

$$u\approx\ln\frac{2}{T^2}+\ln(1-|\nabla_x\psi|^2).$$

This completes the first step.

It is useful to write out the rest of the leading part $a(t)$ of u up to order two in T for two reasons: (a) this reveals that the solution contains logarithmic terms, which disappear, in fact, only if the scalar curvature of the singularity manifold vanishes identically; and (b) this shows that the coefficient of T^2 in the expansion is arbitrary. We therefore compute, by direct substitution,

$$u\approx\ln\frac{2}{T^2}+u_0(x)+u_1(x)T+u_{1,1}(x)T^2\ln T+u_2(x)T^2+\ldots,$$

where u_0, u_1, and $u_{1,1}$ are entirely determined by ψ; in particular, $u_0=\ln(1-|\nabla_x\psi|^2)$. However, u_2 remains arbitrary. One then sets

$$u=\ln\frac{2}{T^2}+u_0(x)+u_1(x)T+u_{1,1}(x)T^2\ln T+vT^2, \tag{4}$$

so that the arbitrary function u_2 appears as an 'initial value' for the renormalized unknown v. This completes the second step.

The singularity data in this case are ψ and $u_2=v(0)$. Once they are known, the formal solution is completely determined.

For the third step, we now substitute expression (4) for u into (3), and find that v solves an equation which can be thought of as a nonlinear perturbation of the Euler–Poisson–Darboux equation. One then checks that v, Tv_T, and $T\nabla_\rho v$ solve a Fuchsian system. This has the following consequences:

(a) There is a formal solution to all orders, in powers of T and $T\ln T$; for $T<0$, one replaces $T\ln T$ by $T\ln|T|$. The series are convergent if ψ and u_2 are analytic; otherwise, they are valid as far as the differentiability of the free functions allows. As already mentioned, the logarithmic terms cannot be dispensed with, except for very special solutions. The example of Gowdy space–times shows that one should even allow terms such as $T^{k(x)}$ for the application to Einstein's equations.

(b) The solution is uniquely determined by the ''singularity data'' $\{\psi,u_2\}$. In fact, the correspondence between these data and the Cauchy data on a hypersurface where the solution is regular can be inverted, using the Nash–Moser version of the inverse function theorem (see Ref. 14). From a practical viewpoint, this means that *(i)* the smoother the Cauchy data, the smoother the singularity surface; and *(ii)* there is a recipe for computing how the singularity data vary when the Cauchy data vary. The question in the opposite direction is simpler: the series representation gives explicitly the solution in terms of the singularity data.

All this makes the series representations generated by Fuchsian techniques as reliable as an exact solution in the vicinity of the singularity—a place where at present no other representation is available.

It should be stressed that the Fuchsian method applies without symmetry or integrability restrictions. For this reason, it enables one to study directly the stability of solutions furnished by generation techniques, even under fully ''inhomogeneous'' or ''asymmetric'' perturbations, and the information it provides yields concrete analytical insight into the properties of solutions.

III. POLARIZED T^2-SYMMETRIC SPACE–TIMES

While the Gowdy $\mathbf{T}^3$ space–times[22] have been extensively studied over the years,[5,16,6,12,7] and are relatively well understood, the more general $\mathbf{T}^2$-symmetric space–times have only recently begun to be considered.[16–18] The technical condition which distinguishes the Gowdy subfamily is the requirement that the Killing fields X and Y which generate the isometry group have vanishing twist constants $\kappa_x := \varepsilon^{abcd} X_a Y_b \nabla_c X_d$ and $\kappa_y := \varepsilon^{abcd} X_a Y_b \nabla_c Y_d$, where ε^{abcd} is the Levi-Civita tensor. The essential difference in practice is that, if one chooses the constant orbit area time-foliation (''Gowdy time''),[22] the constraint equations decouple from the evolution equations in the Gowdy case, and can therefore more or less be ignored in the analysis. If, however, either κ_x or κ_y is nonzero, then no such decoupling occurs.

The general form of the metric and the field equations for the $\mathbf{T}^2$-symmetric space–times is presented in Ref. 17, along with a proof that the Gowdy time always exists globally for these space–times. To write the metric, we assume that all metric components depend on two coordinates, the Gowdy time t and spatial coordinate $\theta \in S^1$ with $\partial/\partial x$ and $\partial/\partial y$ generating the $\mathbf{T}^2$ isometry. By choosing X and Y to be suitable linear combinations of the generators, we may always assume without loss of generality that $\kappa_x = 0$. We then drop the subscript from κ_y. We now focus our attention on the subclass of polarized space–times, which have $A \equiv 0$ in the notation of Ref. 17. The metric takes the form

$$ds^2 = e^{2(\nu-u)}(-\alpha dt^2 + d\theta^2) + \lambda e^{2u}(dx + G_1 d\theta + M_1 dt)^2 + \lambda e^{-2u} t^2 (dy + G_2 d\theta + M_2 dt)^2, \tag{5}$$

where λ is a positive constant and the functions u, ν, α, G_1, M_1, G_2, and M_2 depend on t and θ. The vacuum field equations take the form, writing u_t for $\partial u/\partial t$, etc., $D = t\partial_t$, $D^2 = t^2\partial_t^2 + t\partial_t$, and $m = \lambda \kappa^2$,

$$D^2 u - t^2 \alpha u_{\theta\theta} = \frac{1}{2\alpha} D\alpha D u + \frac{t^2}{2} \alpha_\theta u_\theta, \tag{6a}$$

$$D\alpha = -\frac{\alpha^2}{t^2} m e^{2\nu}, \tag{6b}$$

$$D\nu = (Du)^2 + t^2 \alpha u_\theta^2 + \frac{\alpha}{4t^2} m e^{2\nu}, \tag{6c}$$

$$\partial_\theta \nu = 2 u_\theta D u - \frac{\alpha_\theta}{2\alpha}, \tag{6d}$$

$$G_{1,t} = M_{1,\theta}, \tag{6e}$$

$$G_{2,t}=M_{2,\theta}+\frac{\kappa\alpha^{1/2}}{t^3}e^{2\nu}, \tag{6f}$$

$$\kappa_t=0, \tag{6g}$$

$$\kappa_\theta=0. \tag{6h}$$

Note that the Gowdy case is recovered if $\kappa=0$, $\alpha=1$, and $G_1=G_2=M_1=M_2=0$. Since $G_{1,t}=M_{1,\theta}$, $G_1 d\theta+M_1 dt$ is locally an exact differential $d\varphi$. Replacing x by $x+\varphi$, we may assume *locally* that $G_1=M_1=0$. Similarly, one can set $M_2=0$ by redefining y. Since these reductions are only local and may be incompatible with global requirements, we do not consider them further, even though they do make the geometric ''degrees of freedom'' more clear.

Equations (6a) constitute an initial-value problem for the polarized space–times, in which the equations (6a)–(6d) decouple from the rest. They form an independent system for $\{u,\alpha,\nu\}$. Once these three functions are known, the other equations can be solved easily.

We note that Eqs. (6b)–(6d) in particular—three of the four equations which constitute the heart of the Cauchy problem for these space–times—actually derive from the constraint equations of Einstein's theory. Unlike the Gowdy case, the wave equation (6a) does not decouple from the constraints, since it contains the function α. We therefore take (6a)–(6d) as our basic equations, treating (6a)–(6c) as evolution equations, and (6d) as the only effective constraint.

The local well-posedness of the initial-value problem away from the singularity at $t=0$ is not quite straightforward, for we must prove that Eq. (6d) propagates. This is not an immediate consequence of standard results because we are not using any of the standard setups for the initial-value problem. It nevertheless does hold, and this can be ascertained in two ways.

One approach is as follows (this is basically the argument used by Refs. 17 and 16): if we choose $\{u,u_t,\alpha,\nu,\ldots\}$ at some initial time $t_0>0$ so that they satisfy the constraint (6d), then we can view these as an initial data set for the Einstein equations without any symmetry and construct a local solution in the standard way. One then uses the results of Ref. 16 to introduce coordinates in this region so that the metric takes the form (5).

We can also give a direct argument, which will be useful later. We first deal with the analytic case, which is all we need for the results of Sec. IV. In view of its independent interest, we show in the appendix how to deal with the nonanalytic Cauchy problem as well.

Away from $t=0$, the PDE system (6a)–(6c) is of Cauchy–Kowalewska type. More precisely, we can reduce it to the following first-order system for $(z_0,z_1,z_2,\alpha,\nu):=(u,u_t,u_\theta,\alpha,\nu)$:

$$\partial_t z_0=z_1,$$

$$\partial_t z_1=\alpha\partial_\theta z_2-\frac{z_1}{t}-\frac{m}{2t^3}z_1e^{2\nu}+\tfrac{1}{2}z_2\alpha_\theta,$$

$$\partial_t z_2=\partial_\theta z_1,$$

$$\partial_t\alpha=-\frac{\alpha^2}{t^3}me^{2\nu},$$

$$\partial_t\nu=tz_1^2+t\alpha z_2^2+\frac{\alpha}{4t^3}me^{2\nu}.$$

In particular, ignoring the constraint (6d), we obtain a unique solution of the remaining equations by prescribing the data $\{u,u_t,\alpha,\nu\}$ for $t=t_0$. Now let us set

$$N:=\nu_\theta-2u_\theta Du+\frac{\alpha_\theta}{2\alpha}. \tag{7}$$

Calculating

$$0 = D\nu_\theta - \partial_\theta D\nu = DN + D\left(2u_\theta Du - \frac{\alpha_\theta}{2\alpha}\right) - \partial_\theta D\nu,$$

we find, using (6a)–(6c),

$$DN - \frac{1}{2\alpha} ND\alpha = 0. \tag{8}$$

This is a linear *ordinary* differential equation (ODE) for N (there are no θ-derivatives). Hence if we choose data $\{u, u_t, \alpha, \nu\}$ for $t = t_0$ so that $N(t_0) = 0$, the uniqueness theorem for ODEs guarantees that N is identically zero for all time.

We therefore have proved the well posedness of the initial-value problem. The results of Ref. 17 ensure that the solution remains bounded for $t > \rho$, where $\rho \geqslant 0$ is independent of θ. It is expected that $\rho > 0$ in special cases only, such as exact Kasner space–times.[23] We are interested in asymptotics near $t = 0$. Note that Fuchsian techniques may be useful for analyzing singularities for t near $\rho > 0$; however, if these solutions are nongeneric in some reasonable sense, they should not contain the full number of free parameters, and they may be nonpolarized as well. It does appear that there are consistent asymptotics of the form $u \approx u_0$, $\nu \approx 1/2 \ln(t-\rho) + \nu_0(\theta)$, and $\alpha \approx \alpha_0(\theta) \times (t-\rho)^{-2}$.

As far as the number of free functions in the metric is concerned, one might expect that there will only be two, since one of the gravitational degrees of freedom has been turned off. Indeed, while the initial data for (6a)–(6c) consist of four functions $\{u, u_t, \alpha, \nu\}$, they are constrained by one relation, *viz.* (6d), and, if we set aside the choice of the initial value for the lapse function α, we obtain two arbitrary functions in the solution.

Similarly, we will obtain a family of *singular* solutions of (6a)–(6c) depending on four arbitrary functions occurring in its singular expansion, and will show that if these ''singularity data'' are constrained by one relation, the constraint (6d) holds for all time as well.

IV. APPLICATION OF THE FUCHSIAN ALGORITHM

We are interested in generating solutions to (6) which have controlled asymptotics near $t = 0$ and which are parametrized by as many arbitrary singularity data as possible. We achieve this by following the program outlined in Sec. II.

Step 1. Leading-order asymptotics. Since we expect Kasner-like behavior at the singularity, and since u and ν appear in the metric exponentially, we choose logarithmic leading terms for u and ν:

$$u \approx k(\theta)\ln t + u_0(\theta) + \cdots; \tag{9a}$$

$$\nu \approx (1 + \sigma(\theta))\ln t + \nu_0(\theta) + \cdots; \tag{9b}$$

$$\alpha \approx \alpha_0(\theta) + \cdots. \tag{9c}$$

For Eq. (6b) to hold at leading order, it is sufficient that $\sigma > 0$. For (6c) to hold at leading order, one needs $D\nu$ and $(Du)^2$ to balance each other, which requires that

$$k^2 = 1 + \sigma, \tag{10}$$

which we assume from now on. The function α_0 should be taken to be positive, to ensure the metric has the correct signature.

Note that there are four free functions, namely (k,u_0,α_0,ν_0), in these leading term expansions, just as there were four Cauchy data in the discussion of Sec. III. These four free functions are the singularity data for this system. They are 2π-periodic; furthermore, α_0 and $\sigma=k^2-1$ are positive.

These asymptotics may be compared with those of the solutions obtained in the Gowdy case in Ref. 7. If k_G denotes the parameter called k in Ref. 7, the correspondence is $\pm k_G=2k-1$. This means that the solutions we obtain here, with $k^2>1$, are similar to the ''high-velocity'' Gowdy solutions, for which $k_G>1$. The asymptotics (9a)–(9c) are not compatible with Eqs. (6) if $0<k<1$, unless $m=0$, which is the Gowdy case. Indeed, (6b) implies that α is of the order $t^{2\sigma}$, which is singular if $\sigma=k^2-1$ is negative. This makes the term $D\alpha Du/(2\alpha)$ in (6a) more singular than all the other terms in this equation, so that (6a) cannot hold. There are two ways to circumvent this: (1) take $k=0$, so that Du vanishes to leading order, giving a consistent balance, at the expense of losing the freedom to vary k; and (2) add terms to the field equations which would compensate the most singular term in (6a)—which is possible by going over to the nonpolarized field equations. These possibilities will be addressed when we deal with nonpolarized space–times, in a forthcoming paper.

Step 2. Renormalized unknown. We now introduce new unknowns which will provide an exact form for the remainders indicated with '$\cdots$' in (9a)–(9c). Because of the $e^{2\nu}$ term, we see that it is not possible to assume that the remainder terms are of order t. We do expect them to be of order t^ε if ε is small compared to the minimum of σ. We therefore define the renormalized unknowns (v,μ,β) by

$$u(\theta,t)=k(\theta)\ln t+u_0(\theta)+t^\varepsilon v(\theta,t); \tag{11a}$$

$$\nu(\theta,t)=k^2(\theta)\ln t+\nu_0(\theta)+t^\varepsilon\mu(\theta,t); \tag{11b}$$

$$\alpha(\theta,t)=\alpha_0+t^\varepsilon\beta(\theta,t). \tag{11c}$$

Step 3. Fuchsian system. We shall now show that the renormalized field variables solve a Fuchsian problem. Consequently, once the functions (k,u_0,α_0,ν_0) have been specified, and ε has been chosen small enough, the unknowns v, μ, and β are uniquely determined via Theorem 1.

To achieve this, let us first, since we are looking for a first-order system, introduce first-order derivatives of v as new unknowns. This suggests letting

$$\vec{v}=(v_1,v_2,v_3,v_4,v_5):=(v,Dv,t^\varepsilon v_\theta,\beta,\mu).$$

Let us also introduce the abbreviation $E=m\exp(2\nu_0+2t^\varepsilon\mu)$. It is helpful to remove the t-derivatives of α in the right-hand side of (6a) by using

$$\frac{D\alpha}{\alpha}=-\alpha t^{2\sigma(\theta)}E, \tag{12}$$

which follows from (6b) and (11b). We then find the following evolution equations for $\vec{v}$:

$$Dv_1=v_2; \tag{13a}$$

$$\begin{aligned}Dv_2+2\varepsilon v_2+\varepsilon^2 v_1=&\,t^{2-\varepsilon}(\alpha_0+t^\varepsilon\beta)(k_{\theta\theta}\ln t+u_{0,\theta\theta}+v_{3,\theta})\\&-\tfrac{1}{2}E\alpha t^{2\sigma-\varepsilon}(k+t^\varepsilon(v_2+\varepsilon v_1))\\&+\tfrac{1}{2}t^{2-\varepsilon}(\alpha_0+t^\varepsilon\beta)(k_\theta\ln t+u_{0,\theta}+v_3);\end{aligned} \tag{13b}$$

$$Dv_3=t^\varepsilon\partial_\theta(\varepsilon v_1+v_2); \tag{13c}$$

$$(D+\varepsilon)v_4=-t^{2\sigma-\varepsilon}(\alpha_0+t^\varepsilon\beta)^2E; \tag{13d}$$

$$(D+\varepsilon)v_5=2k(v_2+\varepsilon v_1)+t^{\varepsilon}(v_2+\varepsilon v_1)^2+\tfrac{1}{4}Et^{2\sigma-\varepsilon}(\alpha_0+t^{\varepsilon}\beta)$$
$$+\alpha t^{2-\varepsilon}(k_\theta \ln t+u_{0,\theta}+v_3)^2. \tag{13e}$$

This system has the general form

$$(D+A)\vec{v}=t^{\varepsilon}\vec{f}(t,x,\vec{v},\partial_\theta\vec{v}),$$

where

$$A=\begin{pmatrix} 0 & -1 & 0 & 0 & 0 \\ \varepsilon^2 & 2\varepsilon & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \varepsilon & 0 \\ -2k\varepsilon & -2k & 0 & 0 & \varepsilon \end{pmatrix},$$

and $\vec{f}$ is a five-component object containing all the terms in the system that are not already included in the right-hand side.

By taking ε small (less than the smaller of 1 and any possible value of σ), we can ensure that $\vec{f}$ is continuous in t and analytic in all the remaining variables. Since the eigenvalues of A are ε and 0, of multiplicities four and one, respectively, we conclude that the boundedness condition of Theorem 1 holds. Explicitly, we have $P^{-1}AP=A_0$, hence $\sigma^A=P\sigma^{A_0}P^{-1}$, where

$$A_0=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 & 1 \\ 0 & 0 & \varepsilon & 0 & 0 \\ 0 & 0 & 0 & \varepsilon & 2k \\ 0 & 0 & 0 & 0 & \varepsilon \end{pmatrix} \quad \text{and} \quad P=\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & -\varepsilon & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

so that

$$\sigma^{A_0}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^{\varepsilon} & 0 & 0 & \sigma^{\varepsilon}\ln\sigma \\ 0 & 0 & \sigma^{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 & \sigma^{\varepsilon} & 2k\sigma^{\varepsilon}\ln\sigma \\ 0 & 0 & 0 & 0 & \sigma^{\varepsilon} \end{pmatrix}.$$

We conclude from Theorem 1 that there is a unique solution of the Fuchsian system (13) which vanishes as t tends to zero, and which is analytic in θ and continuous in time. We note in particular that if we construct u, ν, and α from (11a)–(11c) with $v=v_1$, $\mu=v_5$, and $\beta=v_4$, then (u,ν,α) is a solution of equations (6a)–(6c). To verify this, we note that equations (13a)–(13c) imply that

$$D(v_3-t^{\varepsilon}v_{1,\theta})=0,$$

so that any solution which tends to zero with t has also the property that $v_2=tv_{1,t}$ and $v_3=t^{\varepsilon}v_{1,\theta}$.

We now wish to show that, by imposing a constraint on the singularity data (k,u_0,α_0,ν_0), we can guarantee that the solution (u,ν,α) of (6a)–(6c) obtained by solving the Fuchsian system (13) will satisfy the constraint (6d) as well, in order to obtain genuine solutions of Einstein's vacuum equations. We achieve this using (8), which in turn has been derived using only (6a)–(6c).

First of all, since $\vec{f}$ is bounded, we know that $(D+A)\vec{v}$ is actually $O(t^{\varepsilon})$, which implies in particular that α and $D\alpha$ are of order 1 and t^{ε}, respectively. In particular, $D\alpha/\alpha = t\alpha_t/\alpha = O(t^{\varepsilon})$. This means, using (8), that

$$\frac{\partial_t N}{N} = \frac{\alpha_t}{2\alpha} = O(t^{\varepsilon-1}),$$

which is integrable up to $t=0$. [One could also have estimated $D\alpha/\alpha$ directly from (12).] Letting $z(t,\theta)$ be the integral of this function from 0 to t, we find that

$$N(t,\theta) \propto \exp\ z(t,\theta).$$

Thus, if we can choose the data so that $N \to 0$ as $t \to 0$ for fixed θ, we will know that N is, in fact, identically zero, and therefore that the constraint is satisfied. Now

$$N = \nu_\theta - 2u_\theta Du + \frac{\alpha_\theta}{2\alpha}$$

$$= \nu_{0,\theta} - 2ku_{0,\theta} + \frac{\alpha_{0\theta}}{2\alpha_0} + o(1),$$

where $o(1)$ is some expression which tends to zero with t. We conclude that the constraint is satisfied if and only if the singularity data satisfy

$$\nu_{0,\theta} - 2ku_{0,\theta} + \frac{\alpha_{0\theta}}{2\alpha_0} = 0. \tag{14}$$

Note also that all the considerations in this paper are, in fact, local in θ, and therefore allow in principle for other spatial topologies.

To summarize, we have proved the following result:

Theorem 2: *For any choice of the singularity data $k(\theta)$, $u_0(\theta)$, $\nu_0(\theta)$, and $\alpha_0(\theta)$, subject to condition (14), the* $\mathbf{T}^2$*-symmetric vacuum Einstein equations have a solution of the form (11) where β, v, and ν are bounded near, $t=0$. It is unique once the twist constant κ has been fixed, except for the freedom in the functions G_1, G_2, M_1, and M_2. Each of these solutions generates space–times with AVD asymptotics.*

V. CONCLUDING REMARKS

We have therefore obtained a family of singular $\mathbf{T}^2$-symmetric space–times with precise asymptotics at the singularity, which is of AVD type, and which depends on the maximum number of singularity data, that is, as many singularity data as there are Cauchy data for solutions away from the singularity. Fuchsian techniques therefore apply even if the constraints do not decouple from the ''evolution'' equations as in the Gowdy case.

We may also note the following.

First, it is likely that, as in the case of scalar fields, these singular solutions are stable in a Sobolev topology, by application of the Nash–Moser theorem, in which case these solutions form an open set in the space of all solutions. This means that this type of AVD behavior is *stable* in this class, and is therefore not a special feature of some closed-form solution.

Second, the polarized U(1)-symmetric solutions are believed to be AVD as well,[13] and work is underway to address this class by Fuchsian methods.

Third, it appears that the general (nonpolarized) $\mathbf{T}^2$-symmetric space–times may show Mixmaster behavior.[24] Numerical and analytical work to explore this possibility is being carried out.

ACKNOWLEDGMENTS

We thank the Max-Planck-Institut für Gravitationsphysik in Potsdam, where this work was initiated, for its hospitality, and we thank A. D. Rendall for helpful discussions. Partial support for this research has come from NSF Grant No. PHY-9308117 at Oregon.

APPENDIX: NONANALYTIC INITIAL-VALUE PROBLEM

In this appendix, we consider the nonanalytic version of the initial-value problem for $\mathbf{T}^2$-symmetric space–times. The strategy is as follows: We first promote α_θ to a new field variable $\zeta := \alpha_\theta$, and produce an evolution equation for ζ by differentiating (6b) with respect to θ. We then use (6b) to eliminate $D\alpha$ from (6a), and equation (6d) to express $\partial_\theta \nu$ in terms of the other field variables. This gives us a symmetric-hyperbolic system (A1) for $(z_0, z_1, z_2, \alpha, \zeta, \nu)$. Standard theorems then ensure that (A1) admits a unique solution, defined in a small time interval, for nonanalytic, but sufficiently smooth, initial data. We then show that the constraints $\zeta = \alpha_\theta$ and $N=0$ do propagate, by a variant of the argument used for the propagation of the constraint $N=0$. This will establish that we do obtain solutions to (6a)–(6d) with nonanalytic initial data.

We proceed with the details of this argument. The symmetric–hyperbolic system is

$$\partial_t z_0 = z_1, \tag{A1a}$$

$$\partial_t z_1 = \alpha \partial_\theta z_2 - \frac{z_1}{t} - \frac{m}{2t^3} z_1 e^{2\nu} + \frac{1}{2} z_2 \zeta, \tag{A1b}$$

$$\alpha \partial_t z_2 = \alpha \partial_\theta z_1, \tag{A1c}$$

$$\partial_t \alpha = -\frac{\alpha^2}{t^3} m e^{2\nu}, \tag{A1d}$$

$$\partial_t \nu = t z_1^2 + t \alpha z_2^2 + \frac{\alpha}{4t^3} m e^{2\nu}, \tag{A1e}$$

$$\partial_t \zeta = -\frac{2m\alpha}{t^3} e^{2\nu} \left[\zeta + \alpha \left(2t z_1 z_2 - \frac{\zeta}{2\alpha} \right) \right]. \tag{A1f}$$

One verifies by inspection that this system is symmetric–hyperbolic, so that if we prescribe sufficiently smooth initial data $\{u, u_t, \alpha, \zeta, \nu\}$ for $t = t_0$, we obtain a unique solution. The first and third equations ensure respectively that $z_1 = \partial_t z_0$ and $\partial_t (z_2 - \partial_\theta z_0) = 0$; we may thus set $z_0 = u$, $z_1 = u_t$ and $z_2 = u_\theta$. Equations (6a)–(6c) therefore hold, with α_θ replaced by ζ in (6a).

Now, let us set

$$R := \zeta - \alpha_\theta \quad \text{and} \quad N' := \nu_\theta - 2u_\theta D u + \frac{\zeta}{2\alpha}. \tag{A2}$$

We proceed to derive a first-order system of ODEs for R and N'. For the rest of this section, we write N for N', for convenience.

First of all, using Eqs. (A1d) and (A1f),

$$
\begin{aligned}
DR &= D(\zeta - \alpha_\theta) \\
&= -\frac{2m\alpha}{t^2} e^{2\nu}\left[\zeta + \alpha\left(2u_\theta Du - \frac{\zeta}{2\alpha}\right)\right] - \partial_\theta\left(-\frac{\alpha^2}{t^2} m e^{2\nu}\right) \\
&= -\frac{2m\alpha}{t^2} e^{2\nu}[R - \alpha N] \\
&= 2\frac{D\alpha}{\alpha}[R - \alpha N]. \qquad \text{(A3)}
\end{aligned}
$$

Using the expression for N from (A2), taking the relation $\zeta = \alpha_\theta + R$ into account, we have

$$DN = (D\nu)_\theta - 2DuDu_\theta - 2u_\theta D^2 u + D\left(\frac{\alpha_\theta + R}{2\alpha}\right),$$

or

$$DN - D\left(\frac{R}{2\alpha}\right) = (D\nu)_\theta - 2DuDu_\theta - 2u_\theta D^2 u + D\left(\frac{\alpha_\theta}{2\alpha}\right).$$

Then, from (A1a), (A1b), and (A1d) and the definition of R, we find

$$
\begin{aligned}
DN - D\left(\frac{R}{2\alpha}\right) &= \partial_\theta\left(-\frac{D\alpha}{4\alpha}\right) - \frac{D\alpha}{\alpha} u_\theta Du - t^2 u_\theta^2 R + D\left(\frac{\alpha_\theta}{2\alpha}\right) \\
&= -t^2 u_\theta^2 R + \partial_\theta\left(\frac{D\alpha}{4\alpha}\right) - \frac{D\alpha}{2\alpha}(2u_\theta Du).
\end{aligned}
$$

Since, from (6b), one has

$$\left(\frac{D\alpha}{4\alpha}\right)_\theta = \frac{D\alpha}{2\alpha}\left(\nu_\theta + \frac{\alpha_\theta}{2\alpha}\right),$$

it follows that

$$DN - D\left(\frac{R}{2\alpha}\right) + t^2 u_\theta^2 R = \frac{D\alpha}{2\alpha}\left(N - \frac{R}{2\alpha}\right). \qquad \text{(A4)}$$

Thus, combining (A3) and (A4), we have

$$
\begin{aligned}
DN &= N\frac{D\alpha}{2\alpha} + R\left(D\left(\frac{1}{2\alpha}\right) - t^2 u_\theta^2\right) - \frac{RD\alpha}{4\alpha^2} + \frac{D\alpha}{\alpha^2}[R - \alpha N] \\
&= R\left[\frac{D\alpha}{\alpha^2}\left(-\frac{1}{2} - \frac{1}{4} + 1\right) - t^2 u_\theta^2\right] - N\frac{D\alpha}{2\alpha} \\
&= R\left[\frac{D\alpha}{4\alpha^2} - t^2 u_\theta^2\right] - N\frac{D\alpha}{2\alpha}. \qquad \text{(A5)}
\end{aligned}
$$

Equations (A3) and (A5) constitute a linear, homogeneous system of ODEs for R and N. Therefore, if the initial data are such that these quantities are zero for $t = t_0$, they remain so for all time, QED.

[1] S. W. Hawking and G. F. R. Ellis, *The Large-Scale Structure of Spacetime* (Cambridge U. P., Cambridge, England, 1973).
[2] J. Wainwright and L. Hsu, ''A dynamical systems approach to Bianchi cosmologies: orthogonal models of class A,'' Class. Quantum Grav. **6**, 1409–1431 (1989).
[3] V. A. Belinskii, I. M. Khalatnikov, and E. M. Lifshitz, ''A general solution of the Einstein equations with a time singularity,'' Adv. Phys. **13**, 639–667 (1982).
[4] A. D. Rendall, ''Global dynamics of the mixmaster model,'' Class. Quantum Grav. **14**, 2341–2356 (1997).
[5] J. Isenberg and V. Moncrief, ''Asymptotic behavior of the gravitational field and the nature of singularities in Gowdy spacetimes,'' Ann. Phys. (N.Y.) **199**, 84–122 (1990).
[6] P. T. Chruściel, ''On uniqueness in the large of solutions of Einstein's equations ('Strong Cosmic Censorship'),'' Proceedings of the Centre for Mathematical Analysis, Australian National University (Canberra) **27**, ANU (1991).
[7] S. Kichenassamy and A. D. Rendall, ''Analytic description of singularities in Gowdy spacetimes,'' Class. Quantum Grav. **15**, 1339–1355 (1998).
[8] M. Weaver, J. Isenberg, and B. K. Berger, ''Mixmaster behavior in inhomogeneous cosmological spacetimes,'' Phys. Rev. Lett. **80**, 2984–2987 (1998).
[9] D. Eardley, E. Liang, and R. Sachs, ''Velocity-dominated singularities in irrotational dust cosmologies,'' J. Math. Phys. **13**, 99–106 (1972).
[10] D. Hobill, A. Burd, and A. Cohen, *Deterministic Chaos in General Relativity* (Plenum, New York, 1994).
[11] P. T. Chruściel, J. Isenberg, and V. Moncrief, ''Strong cosmic censorship in polarized Gowdy spacetimes,'' Class. Quantum Grav. **7**, 1671–1680 (1990).
[12] B. K. Berger and V. Moncrief, ''Numerical investigation of cosmological singularities,'' Phys. Rev. D **48**, 4676–4687 (1994).
[13] B. K. Berger and V. Moncrief, ''Numerical evidence for velocity-dominated singularities in polarized U(1)-symmetric cosmologies,'' unpublished.
[14] S. Kichenassamy, ''The blow-up problem for exponential nonlinearities,'' Comm. PDE **21**(1&2), 125–162 (1996).
[15] S. Kichenassamy, ''WTC expansions and non-integrable equations,'' to appear in Stud. Appl. Math. (1999).
[16] P. T. Chruściel, ''On space-times with $U(1)\times U(1)$ symmetric compact Cauchy horizons,'' Ann. Phys. (N.Y.) **202**, 100–150 (1990).
[17] B. K. Berger, P. T. Chruściel, J. Isenberg, and V. Moncrief, ''Global foliations of vacuum spacetimes with $\mathbf{T}^2$ isometry,'' Ann. Phys. (N.Y.) **260**, 117–148 (1997).
[18] A. D. Rendall, ''Existence of constant mean curvature foliations in space-times with two-dimensional local symmetry,'' Commun. Math. Phys. **189**, 145–164 (1997).
[19] S. Kichenassamy, *Nonlinear Wave Equations* (Dekker, New York, 1996).
[20] S. Kichenassamy and G. K. Srinivasan, ''The structure of WTC expansions and applications,'' J. Phys. A **28**, 1977–2004 (1995).
[21] S. Kichenassamy, ''Fuchsian equations in Sobolev spaces and blow-up,'' J. Diff. Eqns. **125**, 299–327 (1996).
[22] R. H. Gowdy, ''Vacuum spacetimes and compact invariant hypersurfaces: topologies and boundary conditions,'' Ann. Phys. (N.Y.) **83**, 203–241 (1974).
[23] A Kasner space–time has a three-dimensional isometry group $\mathbf{T}^3$. If we wish to study a Kasner solution as a $\mathbf{T}^2$-symmetric space–time, we must choose a two-dimensional subgroup of $\mathbf{T}^3$. For certain choices of this subgroup, the spatial volume of the $\mathbf{T}^3$ orbits may go to zero before the spatial area of the chosen subgroup orbit does. Hence, evolving backward in time, the space–time stops before the Gowdy time t, which is proportional to the area of the $\mathbf{T}^2$-orbits, hits zero.
[24] J. Isenberg and M. Weaver, unpublished.

The kinematical role of automorphisms in the orthonormal frame approach to Bianchi cosmology

R. T. Jantzen[a)]
Department of Mathematical Sciences, Villanova University, Villanova, PA 19085, and International Center for Relativistic Astrophysics, University of Rome, I-00185 Roma, Italy

C. Uggla
Department of Engineering Sciences, Physics and Mathematics, University of Karlstad, S-65188 Karlstad, Sweden

(Received 14 August 1998; accepted for publication 15 September 1998)

The automorphism group and frame commutator relations in the orthonormal frame approach to Bianchi cosmology are used to construct an explicit coordinate representation of the orthonormal frame itself (and hence of the spacetime metric) which depends algebraically on the connection coefficients. This is not possible in general inhomogeneous models where differential equations must instead be solved. The shift vector field required for this procedure is intimately related to the true Smarr–York minimal strain and minimal distortion shifts. © *1999 American Institute of Physics.* [S0022-2488(99)01701-6]

I. INTRODUCTION

When studying spatially homogeneous (SH) Bianchi cosmology, two complementary approaches have been taken, one using orthonormal frames in which the metric components are fixed and the dynamics resides in the commutation functions (see references in Ref. 1) and the other using computational frames[2] in which the commutation functions are fixed and the dynamics resides in the metric components (see references in Ref. 3). The relationship between these two has been clouded by the fact that one usually uses a synchronous frame in the computational frame approach to Bianchi cosmology. By instead choosing a computational frame based on a suitable shift vector field intimately related to the true Smarr–York minimal strain and minimal distortion shifts,[4,5] one can construct an explicit coordinate representation of all the orthonormal frame vectors (and therefore a coordinate representation of the spacetime metric) using an algebraic procedure involving the commutator functions and commutator relations of the orthonormal frame approach.

For general inhomogeneous models, such an algebraic procedure is not possible and one is instead forced to solve differential equations resulting from the commutator relations in order to obtain a coordinate representation of the orthonormal frame. The closest one might come to the present SH construction in more general inhomogeneous cases would be to use a computational frame with a minimal strain or minimal distortion shift vector field. It is remarkable that the usual orthonormal frame approach to Bianchi cosmology is so closely connected to these general ideas about fixing the coordinate gauge freedom in evolving a spacetime from initial data.

Bianchi cosmology has long served as a testing ground for exploring features of general relativity both in generalizing aspects of these highly symmetric models to the broader context of more general inhomogeneous spacetimes and in specializing results from the general theory to explore them in SH models which facilitate computations. While one may not need the metric explicitly to answer questions about Bianchi models alone, an explicit representation of the metric is essential for answering many interesting questions about general inhomogeneous spacetimes.

[a)]Electronic mail: jantzen@ucis.vill.edu

TABLE I. Canonical structure constants for each Bianchi symmetry type. Note the special case Bianchi type III=VI_{-1} with $\hat{a}=1$.

	Class A						Class B			
Type	IX	VIII	VII_0	VI_0	II	I	VII_h	VI_h	IV	V
$\hat{n}_1$	1	1	1	1	1	0	1	1	1	0
$\hat{n}_2$	1	1	1	-1	0	0	1	-1	0	0
$\hat{n}_3$	1	-1	0	0	0	0	0	0	0	0
$\hat{a}\geqslant 0$	0	0	0	0	0	0	$\hat{a}$	$\hat{a}$	1	1

Thus relating results of Bianchi cosmology to a wider setting requires the construction of the metric.

The outline of this article is as follows. In Sec. II the symmetry group properties of this class of spacetimes are summarized. In Sec. III the essentials of the orthonormal frame approach are reviewed. In Sec. IV the general framework is described for constructing the metric in a computational frame starting from the commutation functions of an orthonormal frame and then a discussion of why it works is given. In Sec. V the metric is explicitly constructed for each Bianchi type. The last section ends with concluding remarks.

II. SYMMETRY PROPERTIES

In the present discussion we consider only those symmetry types for which the full spacetime symmetry group admits a simply transitive 3-dimensional subgroup G acting on the spatially homogeneous (SH) hypersurfaces, i.e., a Bianchi group action. Such spacetimes admit a class of spatial frames $\{\hat{\mathbf{e}}_i\}$ $(i=1,2,3)$ tangent to each hypersurface which are not only invariant under the action of the group but which have structure or commutator functions $\hat{C}^k_{ij}$ which are constants throughout the spacetime. These invariant vector fields thus themselves generate a transformation group which turns out to be isomorphic to the original Bianchi group. The constant structure functions are defined by

$$[\hat{\mathbf{e}}_i,\hat{\mathbf{e}}_j]=\hat{C}^k_{ij}\hat{\mathbf{e}}_k, \tag{2.1}$$

and may be represented in the form (see e.g., Ref. 6)

$$\hat{C}^k_{ij}=2\hat{a}_{[i}\delta^k_{j]}+\epsilon_{ijl}\hat{n}^{lk}, \tag{2.2}$$

where the Jacobi identities $\hat{C}^l{}_{m[i}\hat{C}^m{}_{jk]}=0$ require

$$0=\hat{n}^{ij}\hat{a}_j. \tag{2.3}$$

A useful parameter h may be defined by

$$\hat{a}_i\hat{a}_j=\frac{1}{2}h\epsilon_{ikl}\epsilon_{jmn}\hat{n}^{km}\hat{n}^{ln}. \tag{2.4}$$

The Bianchi symmetry types may be divided into 2 symmetry classes, A and B, depending on whether $\hat{a}_i$ is zero or not.

It is often convenient to choose a gauge in which the structure constants $\hat{n}_{ij}$ are diagonal, and the covector $\hat{a}_i$ is aligned with one of the basis directions, here chosen to be the third one,

$$\hat{n}_{ij}=\mathrm{diag}(\hat{n}_1,\hat{n}_2,\hat{n}_3), \quad \hat{a}_i=(0,0,\hat{a}). \tag{2.5}$$

This ''diagonal-alignment'' gauge will be assumed here. Canonical choices of the structure constants for each Bianchi type are given in this form in Table I.[6] The Bianchi symmetry group action

TABLE II. Dimensions of the adjoint and automorphism groups for each Bianchi symmetry type.

Type	VIII, IX	IV, VI, VII	III	V	II	I
dim(Ad)	3	3	2	3	2	0
dim(Aut)	3	4	4	6	6	9

is assumed to be a global action by a simply connected group. (The discussion of local symmetry actions which are not global is considerably more complicated.)

The automorphism matrix group of the Lie algebra with structure constants $\hat{C}^i{}_{jk}$ is the subgroup of linear transformations of its basis $\{\hat{\mathbf{e}}_i\}$ which leaves those constants invariant,

$$B^k{}_l \hat{C}^l{}_{mn} B^{-1m}{}_i B^{-1n}{}_j = \hat{C}^k{}_{ij} . \tag{2.6}$$

The Lie algebra of this matrix group consists of the matrix derivations of the original Lie algebra,

$$F^k{}_l \hat{C}^l{}_{ij} = \hat{C}^k{}_{lj} F^l{}_i + \hat{C}^k{}_{il} F^l{}_j . \tag{2.7}$$

The matrix adjoint group is the subgroup of inner automorphisms generated by the matrix Lie algebra whose basis consists of a linearly independent subset of the adjoint matrices $\hat{k}_i$ defined by $[\hat{k}_i]^j{}_k = \hat{C}^j{}_{ik}$ representing the inner derivations (Lie bracketing by elements of the original Lie algebra),

$$\pounds_{\hat{\mathbf{e}}_i} \hat{\mathbf{e}}_j = [\hat{k}_i]^l{}_j \hat{\mathbf{e}}_l . \tag{2.8}$$

These matrices satisfy the derivation property (2.7) due to the Jacobi identities. Automorphisms which are not inner are called outer automorphisms. The dimensions of the adjoint and automorphism groups are given in Table II. Their differences represent the number of independent outer automorphism generators which exist in any basis of the full matrix automorphism Lie algebra which includes a basis of the matrix adjoint Lie algebra. The automorphism structure summarized in Table II is important for the algebraic procedure for constructing the metric from the commutator functions.

III. THE ORTHONORMAL FRAME APPROACH

Let $\{\mathbf{e}_a\}$ be a SH orthonormal frame $(a=0,1,2,3)$ with dual frame $\{\boldsymbol{\omega}^a\}$ (satisfying $\langle \boldsymbol{\omega}^a, \mathbf{e}_b \rangle = \delta^a{}_b$) so that the metric takes the form

$$\mathbf{g} = \eta_{ab} \boldsymbol{\omega}^a \boldsymbol{\omega}^b , \tag{3.1}$$

where $(\eta_{ab}) = \mathrm{diag}(-1,1,1,1)$. Choose $\mathbf{e}_0 = \mathbf{n} = n^a \mathbf{e}_a$ to be the unit normal vector field of the SH hypersurfaces. The remaining frame vector fields are then tangent to the SH hypersurfaces and so are related to any set of invariant spatial frame vectors $\hat{\mathbf{e}}_i$ by a linear transformation which is constant on any given such hypersurface.

The full set of commutator functions $\gamma^a{}_{bc}$ are defined by

$$[\mathbf{e}_a , \mathbf{e}_b] = \gamma^c{}_{ab} \mathbf{e}_c , \quad \mathbf{d}\boldsymbol{\omega}^a = -\tfrac{1}{2} \gamma^a{}_{bc} \boldsymbol{\omega}^b \wedge \boldsymbol{\omega}^c . \tag{3.2}$$

As a consequence of the symmetry and hypersurface-forming condition, the normal $\mathbf{n}$ has zero acceleration and rotation. Thus, making a $3+1$ decomposition of these functions leads to[1]

$$\gamma^\alpha{}_{0\beta} = -\theta^\alpha_\beta + \epsilon^\alpha{}_{\beta\gamma} \Omega^\gamma , \qquad \gamma^\gamma{}_{\alpha\beta} = 2 a_{[\alpha} \delta^\gamma{}_{\beta]} + \epsilon_{\alpha\beta\delta} n^{\delta\gamma} , \tag{3.3}$$

where $\epsilon^{\alpha\beta\gamma}$ is the permutation symbol satisfying $\epsilon^{123}=1$ and $\alpha=1,2,3$. The quantity Ω^{α} can be interpreted as the local angular velocity of a spatial frame $\{\mathbf{e}_\alpha\}$ with respect to a second spatial frame $\{\overline{\mathbf{e}}_\alpha\}$ which is Fermi-propagated along $\mathbf{e}_0=\mathbf{n}$. The quantity θ_{ab} is the expansion tensor, which is often represented in terms of the trace-free shear tensor σ_{ab}, the expansion scalar Θ, and the spatial metric $h_{ab}=g_{ab}+n_a n_b$ as $\theta_{ab}=\sigma_{ab}+\frac{1}{3}\Theta h_{ab}$. The purely spatial components $\gamma^{\alpha}{}_{\beta\gamma}$ have been decomposed in the same way as $\hat{C}^{k}{}_{ij}$ and are assumed to have the same diagonal-alignment gauge form as the canonical structure constants for each Bianchi type, with the correspondingly defined structure functions (n_1,n_2,n_3,a) having the same signs (when nonzero) as the corresponding structure constants $(\hat{n}_1,\hat{n}_2,\hat{n}_3,\hat{a})$. The symbol n_α will be used only to designate these structure functions in the remainder of the paper, and not the covariant spatial components of the normal vector field, which are no longer needed.

The $3+1$ decomposition of the Jacobi identities,

$$[[\mathbf{e}_a,\mathbf{e}_b],\mathbf{e}_c]+[[\mathbf{e}_b,\mathbf{e}_c],\mathbf{e}_a]+[[\mathbf{e}_c,\mathbf{e}_a],\mathbf{e}_b]=0 \Leftrightarrow \mathbf{e}_{[a}(\gamma^{d}{}_{bc]})+\gamma^{e}{}_{[ab}\gamma^{d}{}_{c]e}=0, \tag{3.4}$$

leads to

$$\mathbf{e}_0(a^\alpha)=-a_\beta\theta^{\alpha\beta}-\epsilon^{\alpha\beta\gamma}a_\beta\Omega_\gamma, \tag{3.5}$$

$$\mathbf{e}_0(n^{\alpha\beta})=-\Theta n^{\alpha\beta}+2\theta^{(\alpha}{}_{\gamma}n^{\beta)\gamma}-2\epsilon^{\gamma\delta(\alpha}n^{\beta)}{}_{\gamma}\Omega_\delta, \tag{3.6}$$

$$0=a_\beta n^{\alpha\beta}. \tag{3.7}$$

The diagonal-alignment conditions imposed on the structure functions cause certain components of the first two derivative equations to have an identically zero left hand side, leading to certain relationships among components of Ω^{α} and $\theta_{\alpha\beta}$ when the right hand side is not also identically zero. Two relationships follow from the first two components of Eq. (3.5) when $a\neq 0$ and three from the off-diagonal components of (3.6), all of which must be identically zero. These restrictions are

$$a(\theta_{13}-\Omega_2)=0=a(\theta_{23}+\Omega_1), \tag{3.8}$$

and

$$(n_1+n_2)\theta_{12}+(n_1-n_2)\Omega_3=0, \tag{3.9}$$

and its two cyclic permutations.

IV. CONSTRUCTION OF THE METRIC: GENERAL FRAMEWORK

Once the commutation functions have been obtained as explicit functions of some time function t (for example, by solving the constraint and evolution equations for some metric theory with a particular choice of the lapse function N), it is possible to construct the spacetime metric explicitly in terms of local coordinates $\{t,x^i\}$ $(i=1,2,3)$ adapted to the SH hypersurfaces. This is accomplished without solving any additional differential equations, using the particular structure that is associated with the SH symmetry.

Here the orthonormal spatial frame $\{\mathbf{e}_\alpha\}$ will be expressed first in terms of an invariant spatial frame $\{\hat{\mathbf{e}}_i\}$ with canonical structure constant values for its structure functions and then in terms of local coordinates. The full orthonormal frame is related to the computational frame $\{\boldsymbol{\partial}_t,\hat{\mathbf{e}}_i\}$ by

$$\mathbf{e}_0=N(t)^{-1}(\boldsymbol{\partial}_t-\hat{N}^i(t,\mathbf{x})\hat{\mathbf{e}}_i), \quad \mathbf{e}_\alpha=\hat{e}(t)_\alpha{}^i\hat{\mathbf{e}}_i, \tag{4.1}$$

where N is the lapse function and $\vec{\mathbf{N}}=\hat{N}^i\hat{\mathbf{e}}_i$ is the shift vector field. The corresponding dual 1-forms are given by

$$\boldsymbol{\omega}^0 = N(t)\mathbf{d}t, \quad \boldsymbol{\omega}^\alpha = \hat{e}^\alpha{}_i[\hat{\boldsymbol{\omega}}^i + \hat{N}(t,\mathbf{x})^i\mathbf{d}t], \tag{4.2}$$

where $\hat{D} = (\hat{e}^\alpha{}_i)$ is the inverse matrix of $(\hat{e}_\alpha{}^j)$ (i.e., $\hat{e}^\alpha{}_i \hat{e}_\alpha{}^j = \delta^j{}_i$).

This leads to the following spacetime metric:

$$\mathbf{g} = -N(t)^2\ \mathbf{d}t^2 + \hat{g}_{ij}(t)[\hat{\boldsymbol{\omega}}^i + \hat{N}^i(t,\mathbf{x})\mathbf{d}t][\hat{\boldsymbol{\omega}}^j + \hat{N}^j(t,\mathbf{x})\mathbf{d}t], \tag{4.3}$$

where the SH components of the spatial metric tensor in the spatial frame $\{\hat{\mathbf{e}}_i\}$ are given by

$$\hat{g}_{ij} = \delta_{\alpha\beta}\hat{e}^\alpha{}_i \hat{e}^\beta{}_j\,, \tag{4.4}$$

or simply $(\hat{g}_{ij}) = \hat{D}^T\hat{D}$ in matrix notation.

The computational frame $\{\boldsymbol{\partial}_t, \hat{\mathbf{e}}_i\}$ is characterized by the Lie dragging condition $\pounds_{\boldsymbol{\partial}_t}\hat{\mathbf{e}}_i = 0$ which implies the time-independent local coordinate expression $\hat{\mathbf{e}}_i = \hat{e}_i{}^j(x^k)\boldsymbol{\partial}_j$ for the invariant spatial frame, whose constant commutator functions $\hat{C}^k{}_{ij}$ are assumed to have their canonical values given in Table I. Explicit coordinate expressions for $\hat{e}_i{}^j(x^k)$ follow from the representation of the left invariant vector fields in canonical coordinates of the second kind in Refs. 7 and 8. These spatial coordinates will be assumed throughout this article. Similarly, explicit coordinate expressions for a basis of the homogeneity Killing vector fields follow from the representation of the right invariant vector fields in these coordinates.[7,8]

The lapse function is SH, but the associated shift vector field is not necessarily SH. In fact exploiting the action of the outer automorphisms requires an inhomogeneous shift. However, the shift Lie derivative of the spatial frame $\hat{\mathbf{e}}_i$ and its dual must be SH,

$$\pounds_{\vec{\mathbf{N}}}\hat{\mathbf{e}}_i = -\hat{A}(t)^j{}_i\hat{\mathbf{e}}_j\,, \quad \pounds_{\vec{\mathbf{N}}}\hat{\boldsymbol{\omega}}^j = \hat{A}(t)^j{}_i\hat{\boldsymbol{\omega}}^i, \tag{4.5}$$

in order that the Lie derivative of the induced spatial metric be SH for any component matrix $(\hat{g}_{ij})$,

$$\pounds_{\vec{\mathbf{N}}}\hat{g}_{ij} = 2\hat{g}_{k(i}\hat{A}^k{}_{j)}\,. \tag{4.6}$$

This in turn guarantees that the extrinsic curvature (sign-reversed expansion tensor) be SH under the same condition,

$$\hat{K}_{ij} = \frac{1}{2}N^{-1}[-\dot{\hat{g}}_{ij} + \pounds_{\vec{\mathbf{N}}}\hat{g}_{ij}] = -\hat{e}_i{}^\alpha e_j{}^\beta\theta_{\alpha\beta}\,. \tag{4.7}$$

Thus, for a fixed value of t, this restricts the shift to the finite-dimensional space of derivations of the Lie algebra of invariant spatial vector fields.[7] This derivation Lie algebra (containing the homogeneity Killing vector fields which correspond to the trivial zero derivation) generates an action of the automorphism-translation group of the Bianchi homogeneity group G on each SH hypersurface which induces the action of the matrix automorphism group on the invariant spatial vector fields under Lie dragging. Given a basis for the matrix derivation Lie algebra $\{\kappa_P\}$ (so that P is an index taking values from 1 to the dimension of the automorphism group given in Table II), one can construct a corresponding basis $\{\boldsymbol{\xi}_P\}$ for the Lie algebra of derivation vector fields modulo Killing vector fields by the relation

$$\pounds_{\boldsymbol{\xi}_P}\hat{\mathbf{e}}_i = [\kappa_P]^j{}_i\hat{\mathbf{e}}_j\,. \tag{4.8}$$

One can then express the shift and its corresponding derivation matrix as time-dependent linear combinations of these respective bases with SH coefficients,

$$-\hat{A}(t) = M^P(t)\kappa_P \rightarrow \vec{\mathbf{N}} = M^P(t)\boldsymbol{\xi}_P\,. \tag{4.9}$$

This relationship may then be used to determine the shift vector field from the matrix $\hat{A}$. The basis vector fields $\boldsymbol{\xi}_P$ can be taken from a subset of invariant spatial frame vector fields $\hat{\mathbf{e}}_i$ which generate the inner automorphisms, plus some independent outer automorphism generators. Coordinate expressions for the additional independent outer automorphism vector field generators for each symmetry type, most of which may be found in Appendix 3 of Ref. 9, will be given below, while expressions for the invariant spatial vector fields themselves have already been discussed. All of these follow from known results for 3-dimensional Lie groups.

In the diagonal-alignment gauge the action of the matrix automorphism group on the spatial metric induced by the action of Lie dragging by the automorphism-translations,

$$\hat{g}_{ij} \rightarrow B^m{}_i B^n{}_j \hat{g}_{mn}, \tag{4.10}$$

has orbits which may be parametrized by a submanifold of the space of diagonal metric matrices (not unique when diagonal automorphism matrices exist). It is exactly this fact which allows one to assume a shift for which the spatial metric component matrix is confined to such a diagonal submanifold, i.e., making the matrix $\hat{D}$ diagonal. This shift generates a time-dependent matrix automorphism which transforms the orthogonal (zero shift) gauge metric matrix into the diagonal matrix $(\hat{g}_{ij})$.

One may evaluate the relationship between the structure functions of the original orthonormal frame and the computational frame by inserting the expressions (4.1) into Eqs. (3.2) and (3.3) leading to

$$\gamma^\alpha{}_{0\beta} = N^{-1}[-\dot{\hat{e}}^\alpha{}_i \hat{e}_\beta{}^i + \hat{e}^\alpha{}_i \hat{A}^i{}_j \hat{e}_\beta{}^i] = -\theta^\alpha_\beta + \epsilon^\alpha{}_{\beta\gamma}\Omega^\gamma, \tag{4.11}$$

$$\gamma^\alpha{}_{\beta\gamma} = \hat{e}^\alpha{}_i \hat{C}^i{}_{jk} \hat{e}_\beta{}^j \hat{e}_\gamma{}^k. \tag{4.12}$$

The first of these in matrix notation takes the form

$$(\gamma^\alpha{}_{0\beta}) = N^{-1}[-\dot{\hat{D}}\hat{D}^{-1} + \hat{D}\hat{A}\hat{D}^{-1}] = (-\theta^\alpha{}_\beta + \epsilon^\alpha{}_{\beta gg}\Omega^\gamma). \tag{4.13}$$

The shift may be chosen so that the matrix $\hat{D}$ is diagonal and positive-definite,

$$\hat{D} = (\hat{e}^\alpha{}_i) = \mathrm{diag}(e^{\beta^1}, e^{\beta^2}, e^{\beta^3}), \tag{4.14}$$

as will always be assumed here, with the number of independent components equal to three minus the number of independent diagonal automorphisms. This matrix represents the time-dependent rescaling of the orthogonal spatial frame $\{\hat{\mathbf{e}}_i\}$ which normalizes it to the orthonormal spatial frame $\{\mathbf{e}_\alpha\}$ and transforms the nonzero structure constants of the first frame into the corresponding nonzero time-dependent structure functions of the second frame by Eq. (4.12),

$$n_1 = e^{\beta^1 - \beta^2 - \beta^3}\hat{n}_1, \quad n_2 = e^{\beta^2 - \beta^3 - \beta^1}\hat{n}_2, \quad n_3 = e^{\beta^3 - \beta^1 - \beta^2}\hat{n}_3, \tag{4.15}$$

$$a = e^{-\beta^3}\hat{a}. \tag{4.16}$$

When $\hat{D}$ is diagonal, the (index-lowered) symmetric part of Eq. (4.13) is equivalent to the orthonormal components of the mixed form of Eq. (4.7), evaluating the extrinsic curvature or sign-reversed expansion tensor,

$$N^{-1}[-\dot{\hat{D}}\hat{D}^{-1} + \tfrac{1}{2}(\hat{D}\hat{A}\hat{D}^{-1}) + \tfrac{1}{2}(\hat{D}\hat{A}\hat{D}^{-1})^T] = \hat{D}(\hat{K}^i{}_j)\hat{D}^{-1}, \tag{4.17}$$

while the antisymmetric part of Eq. (4.13) relates the shift Lie derivative term involving the matrix $\hat{A}$ to the local angular velocity Ω^{α} of the spatial orthonormal frame [or to the off-diagonal components of the expansion tensor, due to Jacobi identities of the form (3.8) and (3.9)]. The particular way in which this diagonal matrix $\hat{D}$ is fixed when some freedom remains allows one to specialize the shift vector field either to a true minimal strain or minimal distortion shift. These shifts minimize the contribution of the diagonal time derivative term,

$$-\dot{\hat{D}}\hat{D}^{-1} = -(\ln \hat{D})^{\cdot} = \text{diag}(\dot{\beta}^1, \dot{\beta}^2, \dot{\beta}^3), \tag{4.18}$$

to the formula (4.17) for the extrinsic curvature tensor in two different ways.

A true minimal strain shift is one for which the diagonal time-derivative term $-\dot{\hat{D}}\hat{D}^{-1}$ in Eq. (4.17) is orthogonal to the remaining two $\hat{A}$ terms in that expression under the trace inner product for second rank tensors, while the true minimal distortion shift is the one for which the trace-free part of $-\dot{\hat{D}}\hat{D}^{-1}$ is instead orthogonal to the trace-free part of the remaining two $\hat{A}$ terms in that expression.[5] The assumption that $\hat{D}$ is diagonal is consistent with the generic off-diagonal part of the second term in Eq. (4.13) for all Bianchi types in the diagonal-alignment gauge, leaving only the diagonal orthogonality conditions to be analyzed. By representing the logarithm of $\hat{D}$ as an arbitrary linear combination of a set of diagonal matrices which are each orthogonal to the matrix generators of the matrix automorphism group, one obtains a true minimal strain shift vector field. By representing it instead so that the trace-free parts are orthogonal, one obtains a true minimal distortion shift vector field. Since only the true minimal strain and distortion shifts are relevant to the SH case, the modifier ''true'' will be implicitly understood below.

A. The construction procedure

The procedure for constructing a coordinate representation of the metric consists of the following steps.

(1) Represent the diagonal matrix $\hat{D}$ in terms of a minimal set of variables which parametrize the quotient space of the diagonal metric matrices under the action of the diagonal automorphisms. When 1 or 2 independent diagonal automorphism generators exist, there is no natural choice for these variables, and a parameter ζ describes the most useful variations, allowing one to specialize to a minimal strain or minimal distortion shift if desired. This is done by choosing to parametrize $\ln \hat{D}$ or its trace-free part, respectively, so that it is always orthogonal to the diagonal matrix automorphism generators.

(2) Express the minimal diagonal variables in terms of the spatial commutation functions using Eqs. (4.15) and (4.16).

(3) Construct the diagonal (κ_D) and off-diagonal (κ_O) matrix automorphism generators, so that the matrix $-\hat{A}$ can be expressed as a linear combination of them, $-\hat{A} = M^D \kappa_D + M^O \kappa_O$.

(4) Use the diagonal components of Eq. (4.13) to express the time derivatives of the minimal $\hat{D}$ variables as functions of the diagonal components of $\theta_{\alpha\beta}$ and then use these results in the solution of the same equation for the automorphism coefficients (M^D, M^O) to express the latter entirely in terms of the commutation functions. Note that some of these may be expressed in several equivalent ways due to the Jacobi identities (3.8) and (3.9).

(5) Give coordinate expressions for the basis $\{\boldsymbol{\xi}_P\}$ of the automorphism vector fields (modulo Killing vector fields). $\{\hat{\mathbf{e}}_i\}$ provide a basis of the inner automorphism generators corresponding to the adjoint matrices $\{\hat{k}_i\}$ so one only needs coordinate expressions for the remaining independent outer automorphism vector fields which may be easily found from their matrices using the condition (4.8).

(6) Re-express the automorphism matrices as a linear combination of a linearly independent subset of the adjoint matrices $\hat{k}_i$ and the remaining outer automorphisms, so that one can then re-express the shift vector field $\vec{\mathbf{N}} = M^D \boldsymbol{\xi}_D + M^O \boldsymbol{\xi}_O$ as the same linear combination of the

corresponding invariant vectors $\hat{\mathbf{e}}_i$ and the remaining inhomogeneous outer automorphism generators.

The final result then gives a coordinate representation of the orthonormal frame vectors and of the metric whose time dependence is completely determined by the commutation functions through the diagonal matrix $\hat{D}=(\hat{e}^{\alpha}{}_i)$ and the shift coefficients M^D. The spatial homogeneity determines the remaining spatial coordinate dependence.

B. Why it works

Before carrying out these steps explicitly for the various symmetry types, it is worth explaining why the symmetry allows this procedure to work. As discussed in Ref. 7, the ''minigauge group'' of symmetry compatible diffeomorphisms for Bianchi cosmology is the subgroup of the spacetime diffeomorphism group which maps into itself both the space of SH spatial vector fields and also leaves invariant the normal vector field $\mathbf{n}$ to the SH hypersurfaces. Its corresponding Lie algebra consists of vector fields of the form $\mathbf{X}=X^{\perp}\mathbf{n}+\vec{\mathbf{X}}$, where $X^{\perp}$ is SH and the spatial vector field $\vec{\mathbf{X}}$ belongs to the ''automorphism-translation'' Lie algebra on each SH hypersurface.

Since the Bianchi symmetry group G acts simply transitively on its orbits, each orbit is diffeomorphic to G with its action on the orbit corresponding to its action on itself by left translation, and one may map the semidirect product group of automorphisms and (left or right) translations $\mathrm{Aut}(G)\otimes_s L(G)=\mathrm{Aut}(G)\otimes_s R(G)$ onto each orbit, the generators of which define this ''automorphism-translation'' Lie algebra. It is characterized by the condition that the Lie derivatives of the SH spatial vector fields by its elements are themselves SH. The SH spatial vector fields correspond to the left invariant vector fields on G, while the spacetime Killing vector fields generating the action of G correspond to the right invariant vector fields. A minigauge group diffeomorphism of the spacetime is then a 1-parameter (i.e., time-dependent) family of such automorphism-translations acting on the family of SH orbits.

The subgroup of diffeomorphisms generated by the SH spatial vector fields, when acting on the space of SH spatial vector fields by Lie dragging, induces the action of the linear inner automorphism or adjoint group on that space, while the full symmetry compatible diffeomorphism subgroup of automorphism-translations induces the action of the whole linear automorphism group. When expressed in terms of a given invariant spatial frame, these groups are represented by their corresponding matrix groups, which are entirely determined by the values of the structure constants for that frame.

Associated with every computational frame $\{\boldsymbol{\partial}_t,\hat{\mathbf{e}}_i\}$ is an equivalence class of comoving coordinate systems $\{t,x^i\}$ which establish an identification of the spacetime manifold with the product manifold $R\times G$. The usual synchronous gauge frame has the time lines aligned with the unit normal vector field $\mathbf{n}$, and any other symmetry compatible computational frame with the same structure constants is related to it by the action of a time-dependent automorphism matrix induced by the action of the related shift vector field in Lie dragging the original invariant spatial frame. Conversely time-dependent changes of an invariant spatial frame by a time-dependent automorphism matrix are equivalent to the choice of a new time direction for the computational frame. Thus one can reconstruct the associated shift for a new computational frame from a knowledge of the time-dependent automorphism which relates the spatial frames, modulo spacetime Killing vector fields which commute with the spatial frame vectors and induce no change in them nor in the spatial metric or extrinsic curvature, but only change the direction of the time lines of the associated comoving coordinate system since $\boldsymbol{\partial}_t=N\mathbf{n}+\vec{\mathbf{N}}$.

Given a choice of SH spatial frame $\{\hat{\mathbf{e}}_i\}$, the action of the symmetry compatible diffeomorphism group of automorphism-translations on this frame by Lie dragging induces the action (4.10) of the matrix automorphism group on the space of SH inner product matrices. For the diagonal-alignment gauge choice of the structure constants of such a frame, the orbits of this action can be parametrized by a submanifold of the diagonal inner product matrices (corresponding to orthogonal spatial frames). Thus, starting from a frame with arbitrary inner products in synchronous gauge, one can always produce from it an orthogonal frame with minimal freedom in those diagonal inner product components by using the automorphism matrix freedom to choose an appropriate new time direction via the corresponding generating shift vector field. Conversely, given any orthogonal invariant spatial frame with constant structure functions, one can always pick a vector field to complete it to a spacetime computational frame. This is equivalent to picking the shift which generates the time-dependent automorphism matrix which transforms the synchronous gauge inner product matrix to the one of the orthogonal computational frame. Solving the key equation (4.13) for a symmetry compatible shift vector field under the assumption that $\hat{D}$ is a diagonal matrix parametrizing the orbits of the matrix automorphism group determines this desired shift.

V. CONSTRUCTION OF THE METRIC: SPECIFIC BIANCHI TYPES

The metric is simultaneously constructed for each subset of Bianchi types listed in Table II according to common adjoint and automorphism dimensions, except for Bianchi type III$=$VI$_{-1}$, which is included with the type VI and VII discussion, and type IV which is treated separately. This is an artifact of the choices of automorphism matrix parametrizations made for the convenience of calculation.

A. Bianchi types VIII and IX

In this case there are no diagonal or outer automorphisms, $\hat{a}=0=a$, and $n_1n_2n_3\neq 0 \neq \hat{n}_1\hat{n}_2\hat{n}_3$, and the equations (4.15) uniquely determine the β^α, yielding

$$e^{-2\beta^\alpha}=(n_\beta n_\gamma)/(\hat{n}_\beta\hat{n}_\gamma), \tag{5.1}$$

where (α,β,γ) is a cyclic permutation of (1,2,3). Given the canonical choice of $(\hat{n}_\alpha)$ for each symmetry type, the β^α variables are then expressed in terms of the structure functions n_α.

The relationship (4.13) may then be used to solve for the matrix automorphism generator $\hat{A}^i{}_j$. For these symmetry types, the matrix automorphism generators are off-diagonal, and consist entirely of inner automorphism generators belonging to the adjoint Lie algebra of the original Lie algebra with structure constants $\hat{C}^i{}_{jk}$, a basis for which consists of the three off-diagonal adjoint matrices,

$$[\kappa_i]^j{}_k=[\hat{k}_i]^j{}_k=\hat{C}^j{}_{ik}=\epsilon_{ikl}\hat{n}^{jl}, \tag{5.2}$$

so that

$$-\hat{A}^j{}_k=M^i[\kappa_i]^j{}_k=M^i[\hat{k}_i]^j{}_k\,. \tag{5.3}$$

Since there are no diagonal automorphisms, there is no need to examine the diagonal components of the key equation (4.13). Its off-diagonal components immediately determine the off-diagonal matrix $\hat{A}$ in terms of Ω^α and the off-diagonal components of $\theta^\alpha{}_\beta$, the latter of which are related by the off-diagonal components of the Jacobi identities (3.9). The results are

$$N^{-1}M^3=(-\theta_{12}+\Omega_3)/(n_1e^{\beta^3})=(\theta_{12}+\Omega_3)/(n_2e^{\beta^3}), \tag{5.4}$$

and its two cyclic permutations.

The shift vector field is then SH and given by

$$\vec{\mathbf{N}}=M^i\boldsymbol{\xi}_i=M^i\hat{\mathbf{e}}_i\,, \tag{5.5}$$

which is both a minimal strain and a minimal distortion shift.

B. Bianchi types VI and VII

For this category of symmetry types, both class A and B including type III$=$VI$_{-1}$, the matrix automorphism group has one independent diagonal automorphism generator so that one relationship may be imposed on the β^α, which may be parametrized as follows:

$$\ln \hat{D}=\text{diag}(\beta^1,\beta^2,\beta^3)=\beta^- \ \text{diag}(\sqrt{3},-\sqrt{3},0)+\beta^\times \ \text{diag}(\zeta,\zeta,1)$$
$$=\text{diag}(\zeta\beta^\times+\sqrt{3}\beta^-,\zeta\beta^\times-\sqrt{3}\beta^-,\beta^\times). \tag{5.6}$$

The arbitrary parameter ζ, which can be chosen to have any convenient value, reflects a freedom in the shift vector field. The β variables are determined in terms of the structure functions by the two independent components of the equation (4.15),

$$e^{-2\beta^\times}=(n_1n_2)/(\hat{n}_1\hat{n}_2), \quad e^{4\sqrt{3}\beta^-}=(n_1\hat{n}_2)/(\hat{n}_1n_2). \tag{5.7}$$

Note that in the class B case $a^2=hn^1n^2$ and $\hat{a}^2=h\hat{n}_1\hat{n}_2$ leading to an equivalent expression $e^{-2\beta^\times}=(a/\hat{a})^2$.

A basis of the matrix automorphism Lie algebra consists of the following three off-diagonal matrices whose nonzero entries are given by

$$[\kappa_i]^j{}_k=\epsilon_{ikl}\hat{n}^{jl}, \tag{5.8}$$

for each of the three cyclic permutations of (i,j,k) and the fourth diagonal automorphism generator,

$$\kappa_4=\text{diag}(1,1,0). \tag{5.9}$$

One can then express the matrix $\hat{A}$ as a linear combination of these matrices $\hat{A}=M^i\kappa_i+M^4\kappa_4$. (One could have chosen instead the basis $\{k_1,k_2,k_3-a\kappa_4,\kappa_4\}$ more closely related to the adjoint matrices but then k_1 and k_2 are linearly dependent for type III.)

The key equation (4.13) then becomes

$$N^{-1}[-\dot{\beta}^- \ \text{diag}(\sqrt{3},-\sqrt{3},0)-\dot{\beta}^\times \ \text{diag}(\zeta,\zeta,1)+\hat{D}\hat{A}\hat{D}^{-1}]=(-\theta^\alpha{}_\beta+\epsilon^\alpha{}_{\beta\gamma}\Omega^\gamma), \tag{5.10}$$

where the choice $\zeta=0$ corresponds to a minimal strain shift and the choice $\zeta=1$ corresponds to a minimal distortion shift.

The third diagonal component of this equation yields

$$N^{-1}\dot{\beta}^\times=\theta_{33}. \tag{5.11}$$

Using this in the sum of the first two diagonal components of the same equation leads to

$$N^{-1}M^4=-\tfrac{1}{2}(\theta_{11}+\theta_{22})+\zeta\theta_{33}. \tag{5.12}$$

The off-diagonal components of Eq. (5.10) may be used to determine the coefficients of the three off-diagonal automorphism generators in the same way as in the previous case of types VIII and IX,

$$N^{-1}M^1=(-\theta_{23}+\Omega_1)/(n_2e^{\beta^1}), \quad N^{-1}M^2=(\theta_{13}+\Omega_2)/(n_1e^{\beta^2}), \tag{5.13}$$

$$N^{-1}M^3=(-\theta_{12}+\Omega_3)/(n_1e^{\beta^3})=(\theta_{12}+\Omega_3)/(n_2e^{\beta^3}).$$

The shift vector field itself is then determined once a vector field generator corresponding to the diagonal automorphism is identified. Let $\boldsymbol{\xi}_4$ be the unique time-independent (inhomogeneous) spatial vector field which satisfies

$$\pounds_{\xi_4}\hat{\mathbf{e}}_i=[\mathrm{diag}(1,1,0)]^j{}_i\hat{\mathbf{e}}_j\,. \tag{5.14}$$

It has the expression $\boldsymbol{\xi}_4=-x^1\partial_1-x^2\partial_2$ in coordinates that correspond to canonical coordinates of the second kind.[9]

Except for Bianchi type III (VI_h with $h=-1$) the first three automorphism matrices may be expressed in terms of the adjoint matrices $\hat{k}_i$ and the matrix κ_4 in the following way:

$$\kappa_1=(1+h)^{-1}[\hat{k}_1+(\hat{a}/\hat{n}_1)\hat{k}_2],\quad \kappa_2=(1+h)^{-1}[\hat{k}_2-(\hat{a}/\hat{n}_2)\hat{k}_1],\quad \kappa_3=\hat{k}_3-\hat{a}\kappa_4, \tag{5.15}$$

allowing one to expand $\hat{A}$ in terms of these latter four matrices instead. The desired shift vector field is then the same linear combination of the corresponding vector fields $\hat{\mathbf{e}}_i,\boldsymbol{\xi}_4$,

$$\vec{\mathbf{N}}=(1+h)^{-1}[M^1-(\hat{a}/\hat{n}_2)M^2]\hat{\mathbf{e}}_1+(1+h)^{-1}[M^2+(\hat{a}/\hat{n}_1)M^1]\hat{\mathbf{e}}_2+M^3\hat{\mathbf{e}}_3+(M^4-\hat{a}M^3)\boldsymbol{\xi}_4. \tag{5.16}$$

In the type III case, there is one less independent adjoint matrix and one more independent outer automorphism matrix whose corresponding vector field generator must be evaluated. Introduce the combination

$$\kappa_5=\kappa_1+\kappa_2=\begin{pmatrix}0&0&1\\0&0&1\\0&0&0\end{pmatrix},\quad \boldsymbol{\xi}_5=-x^3(\partial_1+\partial_2)=-x^3(\hat{\mathbf{e}}_1+\hat{\mathbf{e}}_2). \tag{5.17}$$

Then the shift vector field can be chosen to be instead

$$\vec{\mathbf{N}}=\tfrac{1}{2}(M^1+M^2)\hat{\mathbf{e}}_1+\tfrac{1}{2}(M^1-M^2)\boldsymbol{\xi}_5+M^3\hat{\mathbf{e}}_3+(M^4-\hat{a}M^3)\boldsymbol{\xi}_4. \tag{5.18}$$

C. Bianchi type IV

Bianchi type IV is very similar to the previous case, but with slightly different algebra. One can use the same parametrization of $\hat{D}$ in terms of β^- and $\beta^\times$ but their relation to the structure functions following from Eqs. (4.15) and (4.16) is now

$$e^{-\beta^\times}=a/\hat{a},\quad e^{2\sqrt{3}\beta^-}=(n_1\hat{a})/(\hat{n}_1a). \tag{5.19}$$

A basis of the matrix automorphism Lie algebra consists of the following three off-diagonal matrices,

$$\kappa_1=\hat{k}_1,\quad \kappa_2=\hat{k}_2,\quad \kappa_3=\hat{k}_3-\hat{a}\ \mathrm{diag}(1,1,0), \tag{5.20}$$

and the fourth diagonal automorphism generator,

$$\kappa_4 = \mathrm{diag}(1,1,0). \tag{5.21}$$

One can then express the matrix $\hat{A}$ as a linear combination of these matrices $\hat{A} = M^i \kappa_i + M^4 \kappa_4$.

The key equation (5.10) and its immediate consequence (5.11) remain the same, leading to the same expression (5.12) for M^4 and to the same values of ζ for the minimal strain and minimal distortion shifts. The off-diagonal components of this equation lead to the following expressions for the remaining automorphism coefficients:

$$N^{-1}M^1 = -[(\theta_{13} + \Omega_2) - (n_1/a)(-\theta_{23} + \Omega_1)]/(ae^{\beta^1}),$$

$$N^{-1}M^2 = (\theta_{23} - \Omega_1)/(ae^{\beta^2}), \quad N^{-1}M^3 = (\theta_{12} - \Omega_3)/(n_1 e^{\beta^3}). \tag{5.22}$$

Let $\boldsymbol{\xi}_4$ be the same as in the previous case. The first three automorphism matrices may be expressed in terms of the adjoint matrices $\hat{k}_i$ and κ_4,

$$\kappa_1 = -[\hat{k}_2 + (\hat{n}_1/\hat{a})\hat{k}_1]/\hat{a}, \quad \kappa_2 = (-\hat{n}_1/\hat{a})\hat{k}_1, \kappa_3 = k_3 - \hat{a}\kappa_4, \tag{5.23}$$

allowing one to expand $-\hat{A}$ in terms of these latter four matrices instead. Then the desired shift vector field is

$$\vec{\mathbf{N}} = -[(\hat{n}_1 M^1/\hat{a}^2) + (\hat{n}_1 M^2/\hat{a})]\hat{\mathbf{e}}_1 - (M^1/\hat{a})\hat{\mathbf{e}}_2 + M^3 \hat{\mathbf{e}}_3 + (M^4 - \hat{a}M^3)\boldsymbol{\xi}_4, \tag{5.24}$$

with the same remarks about the minimal strain and distortion conditions holding as in the previous case.

D. Bianchi type V

The parametrization of $\hat{D}$ for types VI, VII, IV with $\beta^- = 0$ is appropriate here, with the one independent β variable now given by

$$e^{-\beta^\times} = a/\hat{a}, \tag{5.25}$$

as in the type IV case. In addition to the three adjoint matrices $\kappa_i = \hat{k}_i$, one must introduce three outer automorphism matrices:

$$\kappa_4 = \mathrm{diag}(1,-1,0), \quad \kappa_5 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \kappa_6 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{5.26}$$

so that one has $-\hat{A} = M^i \hat{k}_i + M^4 \kappa_4 + M^5 \kappa_5 + M^6 \kappa_6$. Note that $\hat{k}_3 = \hat{a}$ diag(1,1,0) is diagonal and $\hat{a} = 1$ by assumption. Since the diagonal matrix κ_4 is trace-free and orthogonal to ln $\hat{D}$, and $\hat{k}_3$ is the same diagonal automorphism matrix generator as in the previous two cases, the same respective values of ζ yield the minimal strain and minimal distortion shifts.

As before the third diagonal component of the key equation (4.13) yields the same result $N^{-1}\dot{\beta}^\times = \theta_{33}$, while its diagonal components together yield

$$N^{-1}M^3=[-\tfrac{1}{2}(\theta_{11}+\theta_{22})+\zeta\theta_{33}],\quad N^{-1}M^4=\tfrac{1}{2}(-\theta_{11}+\theta_{22}). \tag{5.27}$$

The off-diagonal components then give the remaining automorphism coefficients:

$$\begin{aligned} N^{-1}M^1&=(\theta_{13}+\Omega_2)/(ae^{\beta^1}),\quad N^{-1}M^2=(\theta_{23}-\Omega_1)/(ae^{\beta^2}),\\ N^{-1}M^5&=(-\theta_{12}+\Omega_3)e^{\beta^2-\beta^1},\quad N^{-1}M^6=-(\theta_{21}+\Omega_3)e^{\beta^1-\beta^2}. \end{aligned} \tag{5.28}$$

Let

$$\boldsymbol{\xi}_4=-x^1\partial_1+x^2\partial_2,\quad \boldsymbol{\xi}_5=-x^2\partial_1,\quad \boldsymbol{\xi}_6=-x^1\partial_2, \tag{5.29}$$

be the explicit expressions for the outer automorphism vector fields corresponding to $\kappa_4,\kappa_5,\kappa_6$.[9] Then the desired shift vector field is

$$\vec{\mathbf{N}}=M^i\hat{\mathbf{e}}_i+M^4\boldsymbol{\xi}_4+M^5\boldsymbol{\xi}_5+M^6\boldsymbol{\xi}_6, \tag{5.30}$$

with the same minimal distortion and strain features as in the previous Class B models.

E. Bianchi type II

For this symmetry type, the matrix automorphism group has two additional diagonal automorphism generators leading to two relationships among the β^α, which may be parametrized as follows:

$$\ln \hat{D}=\mathrm{diag}(\beta^1,\beta^2,\beta^3)=\beta^\dagger\ \mathrm{diag}((4\zeta-1)/3,\ (2\zeta+1)/3,\ (2\zeta+1)/3). \tag{5.31}$$

The arbitrary parameter ζ, which can be chosen to have any convenient value, reflects a freedom in the shift vector field. The variable $\beta^\dagger$ is determined in terms of one nonvanishing structure function by the one independent component of the equation (4.15),

$$e^{-\beta^\dagger}=n_1/\hat{n}_1. \tag{5.32}$$

A basis of the matrix automorphism Lie algebra consists of the following off-diagonal matrices:

$$\kappa_2=\hat{k}_2,\quad \kappa_3=\hat{k}_3,\qquad \kappa_1=\begin{pmatrix}0&0&0\\0&0&1\\0&0&0\end{pmatrix},\quad \kappa_4=\begin{pmatrix}0&0&0\\0&0&0\\0&1&0\end{pmatrix}, \tag{5.33}$$

plus the additional two diagonal matrices,

$$\kappa_5=\mathrm{diag}(2,1,1),\quad \kappa_6=\mathrm{diag}(0,1,-1). \tag{5.34}$$

One can then express the matrix $\hat{A}$ as a linear combination of these matrices $-\hat{A}=M^1\kappa_1+M^2\hat{k}_2+M^3\hat{k}_3+M^4\kappa_4+M^5\kappa_5+M^6\kappa_6$.

The key equation (4.13) then becomes

$$N^{-1}[-\dot{\beta}^{\dagger}\ \mathrm{diag}((4\zeta-1)/3,\ (2\zeta+1)/3,\ (2\zeta+1)/3)+\hat{D}\hat{A}\hat{D}^{-1}]=(-\theta^{\alpha}{}_{\beta}+\epsilon^{\alpha}{}_{\beta\gamma}\Omega^{\gamma}), \tag{5.35}$$

where the choice $\zeta=0$ corresponds to a minimal strain shift and the choice $\zeta=1$ corresponds to a minimal distortion shift as in the previous cases. The first diagonal component of this equation minus the sum of the last two diagonal components yields

$$N^{-1}\dot{\beta}^{\dagger}=-\theta_{11}+\theta_{22}+\theta_{33}. \tag{5.36}$$

Using this in the average of the last two diagonal components of the same equation leads to

$$N^{-1}M^{5}=-\zeta\theta_{11}+(\zeta-1)(\theta_{22}+\theta_{33}). \tag{5.37}$$

Taking half the difference of the last two diagonal components of that equation then yields the remaining diagonal automorphism coefficient,

$$N^{-1}M^{6}=-\theta_{22}+\theta_{33}. \tag{5.38}$$

The off-diagonal components yield the remaining coefficients,

$$\begin{aligned} N^{-1}M^{1}&=(-\theta_{23}+\Omega_{1})e^{\beta^{3}-\beta^{2}}, \quad N^{-1}M^{2}=-(\theta_{13}+\Omega_{2})/(n_{1}e^{\beta^{2}}),\\ N^{-1}M^{3}&=(\theta_{12}-\Omega_{3})/(n_{1}e^{\beta^{3}}), \quad N^{-1}M^{4}=-(\theta_{23}+\Omega_{1})e^{\beta^{2}-\beta^{3}}. \end{aligned} \tag{5.39}$$

The generating vector fields ξ_1,ξ_4,ξ_5,ξ_6, corresponding to outer automorphism matrices expressed in canonical coordinates of the second kind are[9]

$$\begin{aligned} \boldsymbol{\xi}_1&=-x^{3}\partial_{2}+\tfrac{1}{2}(x^{3})^{2}\partial_{1}, \quad \boldsymbol{\xi}_5=-2x^{1}\partial_{1}-x^{2}\partial_{2}-x^{3}\partial_{3},\\ \boldsymbol{\xi}_4&=-x^{2}\partial_{3}+\tfrac{1}{2}(x^{2})^{2}\partial_{1}, \quad \boldsymbol{\xi}_6=-x^{2}\partial_{2}+x^{3}\partial_{3}. \end{aligned} \tag{5.40}$$

Finally the desired shift vector field is

$$\vec{\mathbf{N}}=M^{1}\xi_{1}+M^{2}\hat{\mathbf{e}}_{2}+M^{3}\hat{\mathbf{e}}_{3}+M^{4}\xi_{4}+M^{5}\xi_{5}+M^{6}\xi_{6}. \tag{5.41}$$

F. Bianchi type I

For this symmetry type, there exist no nontrivial inner matrix automorphisms and the matrix automorphisms are just the entire general linear group, so the problem is somewhat simpler. Let

$$[E^{i}{}_{j}]^{k}{}_{l}=\delta^{i}{}_{l}\delta^{k}{}_{j} \tag{5.42}$$

be the component definition of the natural basis of 3×3 matrices. In Cartesian coordinates adapted to the translational symmetry, a basis of the corresponding vector field Lie algebra is then[9]

$$\boldsymbol{\xi}^{i}{}_{j}=-[E^{i}{}_{j}]^{k}{}_{l}x^{l}\partial_{k}. \tag{5.43}$$

One may set $\hat{D}$ equal to the identity, so that (4.13) reduces to

$$N^{-1}\hat{A}^{i}{}_{j}=\delta^{i}{}_{\beta}\delta^{\alpha}{}_{j}(-\theta^{\alpha}{}_{\beta}+\epsilon^{\alpha}{}_{\beta\gamma}\Omega^{\gamma}), \tag{5.44}$$

and the corresponding (inhomogeneous) shift vector field is

$$\vec{\mathbf{N}} = -\hat{A}^i{}_j \boldsymbol{\xi}^j{}_i\,, \tag{5.45}$$

where the entries of the matrix $\hat{A}$ themselves play the role of the automorphism coefficients M^D, M^O of the previous cases.

This is clearly both a minimal strain shift and a minimal distortion shift. However, the minimal distortion shift for all the other Bianchi types still has a time-dependent isotropic part of $\hat{D}$ present. To restore this possibility, one can relax the condition on $\hat{D}$ to allow it to have a purely isotropic time-dependent part ($\beta^1 = \beta^2 = \beta^3$) while retaining the minimal distortion condition on the shift, though not the minimal strain condition. Then the parametrization,

$$\ln \hat{D} = \beta^0 \ \mathrm{diag}(1,1,1), \tag{5.46}$$

leads to an arbitrary contribution to the pure trace part of $\hat{A}$,

$$\mathrm{Tr}\ \hat{A} = -\Theta + 3N^{-1}\dot{\beta}^0. \tag{5.47}$$

VI. CONCLUDING REMARKS

The essential ingredient of the above procedure for constructing a coordinate representation of the orthonormal frame is the hypersurface transitivity of the symmetry group. It can therefore be carried over to the temporally homogeneous case and to some extent to their corresponding hypersurface self-similar cases. While the latter are not hypersurface-homogeneous, they are conformally hypersurface-homogeneous with a particular conformal factor.

The construction procedure for SH models is valid for any choice of lapse function. However, it suggests a preferred class of lapse functions, namely those which depend only on the commutator functions through an algebraic relationship. A lapse of this type is used to produce a dimensionless time variable in the dynamical systems formulation[1] of the orthonormal frame equations.

Geometrical objects on spacetime can be assigned dimensions under constant spacetime conformal transformations. In a computational frame the lapse has dimension 1 and the shift dimension 0 provided that a dimensionless time variable is used, while the spatial metric has dimension 2 and hence $\hat{D}$ dimension 1, except for Bianchi type I where $\hat{D}$ is the identity and the spatial coordinates instead have dimension 1. By re-introducing an isotropic degree of freedom in $\hat{D}$ to carry the dimension as described above for this case, the spatial coordinates become dimensionless. In an orthonormal frame the commutation functions have dimension -1.

In the dynamical systems analysis of the orthonormal frame equations, dimensionless variables are obtained by dividing the commutation functions by a function of them which has the same dimension. Usually one chooses the expansion scalar Θ, which leads to the so-called expansion normalized variables.[1] To produce a dimensionless time variable, the lapse must be a function constructed from the commutator functions with dimension 1. In the expansion normalized approach, the associated lapse is $N \propto \Theta^{-1}$.

The minimal distortion shift choice of the present approach is naturally adapted to any such dimensionless formulation since it allows the factorization of the metric into an overall conformal factor carrying the dimension and the remaining part which is therefore automatically expressed in terms of dimensionless variables. This can be seen from the fact that by choosing the minimal distortion shift for all Bianchi types provides a purely isotropic β variable parametrizing part of $\hat{D}$ that carries the dimension, making the remaining variables, if any, dimensionless. Allowing the extra isotropic degree of freedom for Bianchi type I, one can re-express the metric for all Bianchi types in the form in which the square of the lapse is an overall conformal factor times a new unphysical metric. In the minimal distortion gauge, if the lapse is chosen to be a function of the commutation functions with dimension 1, then the new unphysical metric is then expressed entirely in terms of dimensionless combinations of the commutator functions.

In the special case of a Bianchi spacetime with an additional homothetic Killing vector not tangent to the SH hypersurfaces, all dimensionless quantities take constant values[10,11] under these conditions, and the spacetime metric in the minimal distortion gauge is explicitly stationary except for an overall conformal factor which is exponential in the dimensionless time variable. Thus one automatically obtains the standard form for a transitively self-similar spacetime metric adapted to the homothetic Killing vector field (compare Ref. 9).

[1] J. Wainwright and G. F. R. Ellis, *Dynamical Systems in Cosmology* (Cambridge University Press, Cambridge, 1997).
[2] J. W. York, Jr., *Sources of Gravitational Radiation*, edited by L. Smarr (Cambridge University Press, Cambridge, 1979).
[3] K. Rosquist and R. T. Jantzen, Phys. Rev. D **166**, 90 (1988).
[4] L. Smarr and J. W. York, Jr., Phys. Rev. D **17**, 2529 (1978).
[5] R. T. Jantzen, Ann. Inst. Henri Poincaré **33**, 121 (1980).
[6] G. F. R. Ellis and M. A. H. MacCallum, Commun. Math. Phys. **12**, 108 (1969).
[7] R. T. Jantzen, Commun. Math. Phys. **64**, 211 (1979).
[8] R. T. Jantzen, *Proc. Int. Sch. Phys. ''E Fermi'' Course LXXXVI on ''Gamov Cosmology,''* edited by R. Ruffini and F. Melchiorri (North-Holland, Amsterdam, 1987); *Cosmology of the Early Universe*, edited by L. Z. Fang and R. Ruffini (World Scientific, Singapore, 1984).
[9] R. T. Jantzen and K. Rosquist, Class. Quantum Grav. **3**, 281 (1986).
[10] R. T. Jantzen and K. Rosquist, Class. Quantum Grav. **2**, L129 (1985).
[11] L. Hsu and J. Wainwright, Class. Quantum Grav. **3**, 1105 (1986).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

Inverse Laplace transform and perturbation theory

T. Biswas and Satish D. Joglekar[a)]
Department of Physics, Indian Institute of Technology, Kanpur, Kanpur 208016, India

(Received 29 May 1998; accepted for publication 9 July 1998)

We prove an extension of the result on the inverse Laplace transform. This extension will help toward making the applications of Borel techniques to perturbation theory in Quantum Field Theories be placed on a more rigorous foundation. © *1999 American Institute of Physics.* [S0022-2488(98)02812-6]

I. INTRODUCTION

The known elementary particle processes are very well described by the standard model,[1] which is a Quantum Field Theory (QFT) in four dimensions. No QFTs that are of direct physical interest are exactly soluble in four dimensions. Hence a recourse has to be made to the perturbation theory in the coupling constant(s) in order to evaluate testable consequences of the QFTs. However, as it turns out,[2] even for the simple $\lambda\phi^4_{[4]}$ theory, the series in powers of coupling λ is only an asymptotic series.[3] While a suitable truncation of an asymptotic series can offer a good approximation to a physical quantity $P(\lambda)$ being calculated for a given coupling λ, nonetheless, as the asymptotic series is a strictly divergent one, it cannot be regarded as a rigorous mathematical expression for $P(\lambda)$ and may miss out on some features of $P(\lambda)$. Moreover, an attempt at the evaluation of $P(\lambda)$ in a power series in λ inherently assumes the analytic nature of P at $\lambda=0$, and this will miss out contributions to $P(\lambda)$ which may be singular at $\lambda=0$. Such contributions are known to exist, say, in QCD.[4]

In view of the divergent nature of the perturbation series for $P(\lambda)$, attempts have been made to relate this perturbation series to a series with improved convergence properties, and then from this new series extract information about $P(\lambda)$ that was originally lost in its asymptotic expansion.[5] The various techniques of Borel/Laplace transforms have been utilized in these works.[6–10] In this process, results have had to be used whose mathematical validity has not been known for the class of functions to which they have been applied. Our purpose in this paper is to try to fill this gap by extending the existing results in this regard further to the set of functions that may be of interest in physical theories but not covered in the conditions that are necessary for existing results on the Borel/Laplace transform to hold.

We now summarize the plan of the paper. In Sec. II, we summarize the various methods and uses involving Borel/Laplace transforms that have been employed in QFT literature.

We summarize the known results on these in the mathematical literature.[11]

In Sec. III, we state the extended version of the theorem and prove it. In Sec. IV we make comments on the usefulness of the results and suggest further extensions of the proof.

II. PREVIOUS RESULTS

A. Uses of Borel sum, Borel transforms

First we shall review, in a nonrigorous fashion, the use of a Borel sum/transform in the literature. As mentioned in the Introduction, the perturbation expansion in powers of the coupling λ for a physical quantity $P(\lambda)$ is generally an asymptotic expansion,

$$P(\lambda)=\sum_{n=0}^{\infty} a_n\lambda^n. \tag{1}$$

[a)]Electronic mail: sdj@iitk.ernet.in

In addition, there could be nonperturbative contributions to $P(\lambda)$, singular at $\lambda=0$, say of the form $e^{A/\lambda}$, which are missed in the usual Feynman diagrammatic perturbation expansion. The series in such cases, has zero radius of convergence; yet it can give a good approximation to $P(\lambda)$ for small enough λ if one keeps a correct number of terms in (1) to get the best approximation. Now it may happen that the additional information about $P(\lambda)$ may be hidden in (1) itself but as (1) is not an expansion for $P(\lambda)$ it cannot directly be deduced from the asymptotic series. Hence, one constructs a companion series to (1); viz., the ''Borel sum,''

$$B_1(z) \equiv \sum_{n=0}^{\infty} a_n \frac{z^n}{n!}. \tag{2}$$

The series in (2) may have a finite radius of convergence (depending upon the behavior of a_n's), even though (1) had radius of convergence zero.

[Of course, if (1) has a finite radius of convergence, then it can be shown that (2) represents an entire function.[2]] Now (2) could then be, in principle, summable in the disk $|z|<R$ and could possibly be analytically continued to the entire complex plane. The function $P(\lambda)$ can be recovered formally from $B_1(z)$ via

$$\lambda P(\lambda) = \int_0^{\infty} e^{-z/\lambda} B_1(z) dz. \tag{3}$$

The reconstruction of $P(\lambda)$ from $B_1(z)$ is possible only if $B_1(z)$ has no singularities on the positive real axis. In such cases, the theory is said to be Borel summable. On the other hand, the presence of a singularity of $B_1(z)$ on the positive real axis makes (3) ill defined and any procedure of giving meaning to the right-hand side of (3) by deformation of the contour makes the process ambiguous, as the latter could be chosen in multiple ways. In such cases the theory is said to be Borel nonsummable. QCD is an example of a Borel nonsummable theory; and this is due to the presence of singularities of $B_1(z)$ called Renormalons on the positive real axis. In these latter cases, the perturbation series alone is unable to unambiguously obtain $P(\lambda)$ via (3) uniquely.[5]

Equation (3) allows one to construct, in the case of Borel summable theories, the exact analytic function $P(\lambda)$ of which (1) was only a perturbative asymptotic expansion.

Alternate forms of Borel/Laplace transforms have also been used. Here are some important ones.

First we define[7]

$$B'(s) \equiv \int_0^{\infty} e^{s/f} Z(f)\, d\left(\frac{1}{f}\right), \tag{4}$$

$$B_2(s_r) \equiv \frac{1}{2\pi} \text{disc } B'(s); \tag{5}$$

disc=discontinuity across the real axis at $s=s_r$. Then in some cases we can construct back $Z(f)$,

$$Z(f) = \int_0^{\infty} B_2(s) e^{-s/f}\, ds. \tag{6}$$

The one we shall be dealing with extensively in this work is defined as follows:
Let $P(\lambda)$ be the function we are interested in.

Define

$$F(z) = P\left(\frac{1}{z}\right), \tag{7}$$

assuming this is possible.

Now define the inverse Laplace transform,[6]

$$B_3(s)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}e^{sz}F(z)dz; \tag{8}$$

c is, at present, arbitrary. Then we can reconstruct $F(z)$ by the Laplace transform

$$F(z)=\int_0^\infty e^{-tz}B_3(t)dt. \tag{9}$$

We note that if $P(\lambda)=\Sigma_{n=1}^{\infty}c_n\lambda^n$, and the series $F(z)$ defined by (7) is substituted in (8) one would obtain (if summations and integrations could commute)

$$B_3(t)=\sum_{n=1}^{\infty} c_{n+1}\frac{t^n}{n!}, \tag{10}$$

which is very similar to $B_1(t)$ [c_n being replaced by c_{n+1}, and, in fact, by a slight modification of the definition of the Borel sum, as in Ref. 8, it may be made to coincide with $B_3(t)$].

We now make comments on the rigor of the above result. (1) is valid generally only as an asymptotic series, enabling one to estimate P to a good accuracy. $B_1(z)$ of (2) is defined only within a finite radius of convergence $|z|<R$ outside; it has to be analytically continued, and this analytic continuation is used in (3) for $|z|>R$. For small λ, and Borel summable theories, one can argue[5] that (3) yields the correct asymptotic series (1), as the region $|z|<\lambda<R$ is what contributes to the right-hand side of (3) where the uniformly convergent expansion (2) for $B_1(z)$ holds and term by term integration becomes possible. We then *assume* that $P(\lambda)$ of (3) is indeed the physical quantity we are looking for in any λ.

The expression for $F(z)$ of (7) is only valid as an asymptotic expansion and not as a uniformly convergent series. Hence on the right-hand side of (8) the interchange of summation and integration that would lead to (10) is not really justified.

In this work, we shall deal mainly with the reconstruction via $B_3(t)$ of (8). Our aim is not so much to apply it yet to the asymptotic series such as (7) but to extend and rigorously establish the sufficient conditions under which such a reconstruction is valid for a function $F(z)$. We will discuss a few applications of our extension in Sec. IV. However, we will not make the ultimate contact of establishing a rigorous procedure for reconstruction starting from asymptotic series $P(\lambda)$. We aim only at the extension of the known results in this connection, which we now recapitulate.

B. Known results on inverse Laplace transform

The Complex Laplace Inverse Theorem.[11]

Theorem: Let $F(z)$ be analytic in the whole of the Complex plane, except that it has a finite number of poles. Suppose that $F(z)$ is analytic on half-plane $\{z|\mathrm{Re}\, z>\sigma\}$ and that $|F(z)| \leqslant M/|z|^\beta$ for all $|z|\geqslant R$ for constants M, $\beta>0$. Then for $t\geqslant 0$, let

$$\tilde{F}(t)=\sum \{\text{residue of } e^{zt}F(z) \text{ at each of its pole}\}. \tag{11}$$

Then

$$\mathcal{L}_z\tilde{F}(t)=F(z), \quad \forall\ \mathrm{Re}\ z>\sigma. \tag{12}$$

A corollary to this theorem is

$$\tilde{F}(t)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+\infty}e^{zt}F(z)dz, \quad \forall\ \mathrm{Re}\ z>\sigma. \tag{13}$$

We have tried to extend this theorem to cover a larger class of functions. Following is the modified version.

III. EXTENSION OF THE RESULT ON LAPLACE TRANSFORM

In this section, we shall present the extension of the result on the Laplace transform.

A. The theorem

Let $F(z)$ be analytic on the half-plane $\{z|\text{Re } z>\sigma\}$. Let $c\equiv\max(\sigma,0)$. We then define, for $t\geqslant 0$,

$$\tilde{F}(t)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+\infty}e^{zt}F(z)dz \tag{14}$$

(whenever integral exists). Then

$$\mathcal{L}_z\tilde{F}(t)\equiv\int_0^{\infty}\tilde{F}(t)e^{-zt}\,dt=F(z), \tag{15}$$

for all Re $z>c$, if $F(z)$ satisfies the following (weaker) conditions. These require that $F(z)$ can be written as a finite sum of functions:

$$F(z)=\sum_{i=1}^{N}F_i(z), \tag{16}$$

with each $F_i(z)$ satisfying the following conditions.

(a) With some R_i, $|F_i(z)|$ can be bounded, for $|z|>R_i$, as

$$|F_i(z)|\leqslant\frac{M_i e^{-b\text{ Re }z}}{|z|^{\beta_i}}, \tag{17}$$

with $M_i,b_i,\beta_i>0$.

(b) $F_i(z)$ has only a finite number of singularities w_{ij} $(j=1,2\ldots m_i)$.

(c) $F_i(z)$ has only a finite number B of separated branch cuts, which can be chosen to be of horizontal type, extending parallel to the negative real axis.

The proof, being lengthy, is being broken up. We first prove several lemmas. In the proofs, we shall constantly refer to various contours that we shall first define.

The contours K_j $(j=1...,B)$ are around the branch point K_j. Of these k_i represents the branch point with maximum real part. $C_{21},C_{22},\ldots,C_2(B+1)$ represents a broken up vertical contour, which we shall denote collectively by C_2. Then C represents the closed contour consisting of the segments C_0,C_1,C_2,C_3 and K_j $(j=1,...,B)$. x_1,y_1,y_2,x_1' are all positive. By assumption x_1, y_1, and y_2 have been chosen to be large enough and K_j narrow enough so that C always encloses all isolated singularities of $F_k(z)$ for a given k, except those lying on the branch cut itself. The contour C_2' is always to the right of C_0 (i.e., $x_1'>c$). C_0,C_1',C_2',C_3' together constitute the contour C'. We assume that K_j's have been chosen symmetrically around the branch cut at a distance ϵ_j such that there are no singularities in the strip of width $2\epsilon_j$, except those lying on the branch cut itself.

B. The lemmas

In all the three lemmas below we shall always restrict ourselves to a function $g(z)$ satisfying the condition that there exist $R>0$ such that $|g(z)|$ can be bounded as

$$|g(z)|\leqslant\frac{Me^{-b\text{ Re }z}}{|z|^{\beta}}\quad |z|>R \tag{18}$$

with some positive real M, b, β. We shall call this condition A.

Lemma 1: Let $g(z)$ satisfy condition A. We define

$$I(\tilde{C})=\int_{\tilde{C}}g(z)e^{zt}\,dz. \tag{19}$$

Then (i) for $t<b$,

$$\lim_{\tilde{C}\to\infty} I(\tilde{C})=0, \quad \tilde{C}=C_1',C_2',C_3'; \tag{20}$$

(ii) for $t>b$,

$$\lim_{\tilde{C}\to\infty} I(\tilde{C})=0, \quad \tilde{C}=C_1,C_2,C_3; \tag{21}$$

and (iii) for $t=b$ if $\beta>1$,

$$\lim_{\tilde{C}\to\infty} I(\tilde{C})=0, \quad \tilde{C}=C_1,C_2,C_3,C_1',C_2',C_3'. \tag{22}$$

In (ii) above, the limit $\tilde{C}\to\infty$ implies, whenever applicable, that y_1 and/or $y_2\to\infty$ after x_1' or $x_1\to\infty$.

Proof: As the contour $\tilde{C}$ is tending to infinity, we can assume, without loss of generality, that all points on $\tilde{C}$ are at a distance $\geqslant R$ from the origin. Then, $y_1,y_2,x_1'-c,x_1+c\geqslant R$. (i) We note that

$$I(C_1')=\int_c^{x_1'} g(z)e^{xt+iyt}\,dx, \tag{23}$$

so that in view of the fact that $y_1>R$ and the condition A,

$$|I(C_1')|\leqslant\int_c^{x_1'}\frac{Me^{-bx}e^{xt}}{(x^2+y_1^2)^{\beta/2}}\,dx\leqslant\frac{M}{y_1^\beta}\int_c^{x_1'}e^{(t-b)x}\,dx \tag{24}$$

$$=\frac{M}{y_1^\beta}\frac{1}{t-b}\left[e^{-(b-t)x_1'}-e^{-(b-t)c}\right], \quad t\neq b$$

$$\to 0, \quad \text{as } y_1\to\infty, \tag{25}$$

for any x_1 and for $t<b$ and $I(C_1')\to 0$ for $x_1'\to\infty$, $y_1\to\infty$ in any order.

(iia) The argument also goes through with $C_1'\to C_3'$. It also works with $C_1'\to C_1$ or C_3 (with $x_1'\to -x_1$), provided $t>b$.

(ib)

$$I(C_2')=i\int_{-y_2}^{y_1} g(z)e^{x_1't+iyt}\,dy, \tag{26}$$

so that with $x_1'>R$ and the condition A on $g(z)$,

$$|I(C_2')|\leqslant\int_{-y_2}^{y_1}\frac{Me^{-bx_1'}e^{x_1't}}{(x_1'^2+y_1^2)^{\beta/2}}\,dy \tag{27}$$

$$\leqslant Me^{-(b-t)x_1'}\int_{-y_2}^{y_1}\frac{dy}{(R^2+y^2)^{\beta/2}}\to 0, \quad \text{as } x_1'\to\infty, \tag{28}$$

for $t<b$, assuming y_1,y_2 to be finite.

(iib) The argument above also goes through with $C_2'\to C_2$ and $x_1'\to -x_1$ if $t>b$.

(iii) For $t=b$ and $\beta=1+4\alpha>1$, we have from (24),

$$|I(C_1')|\leqslant\int_c^{x_1'}\frac{M}{(x^2+y_1^2)^{1/2+2\alpha}}\,dx$$

$$\leqslant M\int_c^{x_1'}\frac{dx}{(x^2+y_1^2)^{1/2+\alpha}(y_1^2)^{\alpha}}$$

$$\leqslant\frac{M}{y_1^{2\alpha}}\int_c^{x_1'}\frac{dx}{x^{1+2\alpha}}$$

$$=\frac{M}{y_1^{2\alpha}}\frac{(x_1')^{-2\alpha}-(c)^{-2\alpha}}{(-2\alpha)}\to 0,\quad \text{as } y_1\to\infty,$$

for any fixed x_1' (and also as $x_1'\to\infty$, $y_1\to\infty$ in any order). The same argument goes through if $C_1'\to C_3'$. It also goes through for $C_1'\to C_1$, with minor modifications.

Further, for $t=b$ and $\beta=1+4\alpha>1$, we have, from (27),

$$|I(C_2')|\leqslant\int_{-y_2}^{y_1}\frac{M}{(x_1'^2+y^2)^{1/2+2\alpha}}\,dy$$

$$\leqslant\int_{-y_2}^{y_1}\frac{M}{(x_1^{2\prime}+y^2)^{1/2}x_1'^{2\alpha}}\,dy$$

$$=\frac{M}{x_1'^{2\alpha}}\tilde{F}(y_1,y_2)\to 0,\quad \text{as } x_1'\to\infty,$$

for any finite y_1,y_2 [since $\tilde{F}(y_1,y_2)$ is then finite].

The same argument goes through with $C_2'\to C_2$ and $x_1'\to -x_1$.

Lemma 2: Let $g(z)$ satisfy condition A and be continuous on C_0. Then for Re $z>c$.

$$\lim_{r\to\infty} e^{-rz}\int_C\frac{dz'g(z')e^{rz'}}{z'-z}\equiv\lim_{r\to\infty} I(r)=0,$$

where C is chosen as in Fig. 1 with $x_1,y_1,y_2>R$.

Proof: On $C_1,C_3,C_{21},C_{22},\ldots,C_{2(B+1)}$; $|z'|>R$. So that condition A yields that on these,

$$|g(z')|\leqslant\frac{Me^{-b\text{ Re }z'}}{|z'|^{\beta}}\leqslant\frac{Me^{bx_1}}{R^{\beta}}\equiv g_1.\tag{29}$$

Also, $g(z)$ is continuous on the finite segment C_0 and hence bounded. Hence

$$|g(z)|\leqslant g_2,\quad \text{on } C_0.\tag{30}$$

Further, $g(z)$ is analytic on $K_j's$, $j=1,2,\ldots,B$, which are finite in length. Hence it is bounded on the set of disjoint contours $K_1+K_2+\cdots+K_B$. Let

$$|g(z)|\leqslant g_3,\quad \text{for } z\epsilon K_j,\quad \forall j.\tag{31}$$

Let $g_0\equiv\max(g_1,g_2,g_3)$. Then

$$|g(z)|\leqslant g_0,\quad \text{on } C.\tag{32}$$

Also,

$$|e^{rz'}|=e^{rx'}\leqslant e^{rc}\quad \text{on } C,\tag{33}$$

$$\frac{1}{|z-z'|}=\frac{1}{[(x'-x)^2+(y-y')^2]^{1/2}}\leqslant\frac{1}{|x-x'|}\leqslant\frac{1}{(x-c)}\quad \text{on } C.\tag{34}$$

Further, the length of the contour C [$\epsilon_{\max}\equiv\max(\epsilon_1,\ldots,\epsilon_B)$],

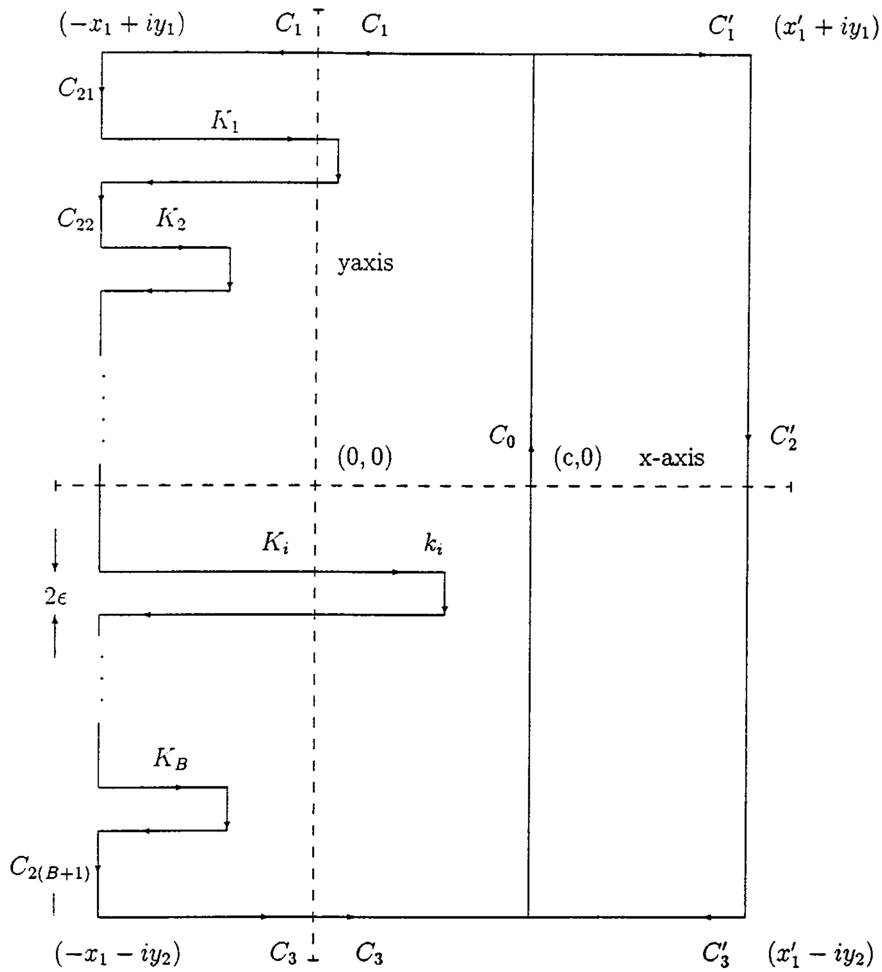

FIG. 1. The contours used in the proof of the theorem.

$$L \leqslant 2(y+y_1)+2(c+x_1)+B[2(x_1+k_i)+4\epsilon_{\max}] \equiv L_0. \tag{35}$$

Thus,

$$\left| e^{-rz}\int_C \frac{dz' g(z') e^{rz'}}{|z'-z|}\right| \leqslant L_0 g_0 \frac{1}{(x-c)} e^{rc} e^{-rx} \to 0, \quad \text{as } r\to\infty, \quad \text{for } \mathrm{Re}\, z=x>c. \tag{36}$$

Lemma 3: Let $g(z)$ satisfy the condition A. Then, for $\mathrm{Re}\, z>c$, we have

$$I=\lim_{K_j\to\infty}\int_{K_j} g(z')dz' \int_b^\infty e^{(z-z')}\, dt = \int_b^\infty dt\, e^{-zt}\int_{K_j\to\infty} e^{z't} g(z')dz' \tag{37}$$

[where K_j is the contour around the branch cut consisting of (i) a straight line parallel to the x axis: $y=y_+\equiv \mathrm{Im}\, k_j+\epsilon$; $-x_1\leqslant x\leqslant x_0\equiv \mathrm{Re}\, k_j+\epsilon$; (ii) a straight line oriented antiparallel to the y axis: $x=x_0$, $\mathrm{Im}\, k_j-\epsilon\equiv y_-\leqslant y\leqslant y_+$; (iii) A straight line antiparallel to the x axis: $y=y_-$, $x_0>x>-x_1$.]

Proof: Introducing the notation

$$G(x',y',t)\equiv e^{(z'-z)t} g(z') \equiv G_R(x',y',t)+iG_I(x',y',t), \tag{38}$$

we could express I as

$$I=I_+-I_2-I_-, \tag{39}$$

with

$$I_{\pm} \equiv \int_{-\infty}^{x_0} dx' \int_b^{\infty} dt\ G_R(x',y_{\pm},t) + i \int_{-\infty}^{x_0} dx' \int_b^{\infty} dt\ G_I(x',y_{\pm},t), \tag{40}$$

$$I_2 \equiv i \int_{y_-}^{y_+} dy' \int_b^{\infty} dt\ G_R(x_0,y',t) - \int_{y_-}^{y_+} dy' \int_b^{\infty} dt\ G_I(x_0,y',t). \tag{41}$$

Equation (37) is established if the order of integration in each of the six real integrals occurring in (40)–(41) can be interchanged. According to the Tonelli–Hobson Theorem,[12] this is possible in each case if the real integral under consideration or one obtained by interchange of the order of integrations is shown to exist. Thus, (37) would be established if, in particular, the six real integrals present in (40) and (41) are (themselves) shown to exist. However, these six integrals exist iff I_+, I_-, I_2 themselves exist. We thus need to show that

$$I_{\pm} = \int_{-\infty}^{x_0} dx' \int_b^{\infty} dt\ e^{(z'-z)t} g(z') \big|_{y'=y_{\pm}} \tag{42}$$

and

$$I_2 = \int_{y_-}^{y_+} dy' \int_b^{\infty} dt\ e^{(z'-z)t} g(z') \big|_{x'=x_0} \tag{43}$$

exist. Performing the t integrals in both cases (and noting that $\mathrm{Re}\, z > c > \mathrm{Re}\, z'$), we find $[\tilde{g}(z') \equiv -g(z') e^{(z'-z)b}]$,

$$I_{\pm} = \int_{-\infty}^{x_0} dx'\ \frac{\tilde{g}(z')}{z'-z} \bigg|_{y'=y_{\pm}}, \tag{44}$$

$$I_2 = \int_{y_-}^{y_+} dy'\ \frac{\tilde{g}(z')}{z'-z} \bigg|_{x'=x_0}. \tag{45}$$

Now, as for I_2, $g(z')/(z'-z)$ is analytic on the finite interval $y_- \leqslant y' \leqslant y_+$, $x' = x_0$ and thus bounded. The range of y' is also bounded. Hence

$$|I_2| = (y_+ - y_-)\{[\tilde{g}(z')/z'-z]_{x'=x_0}\}_{\max} < \infty. \tag{46}$$

As for $I_{\pm}$,

$$I_{\pm} = \int_{-R}^{x_0} dx'\ \frac{g(z')}{z'-z} \bigg|_{y'=y_{\pm}} + \int_{-\infty}^{-R} dx'\ \frac{g(z')}{z'-z}. \tag{47}$$

The first integral on the right-hand of (46) is finite and will approach a finite limit as $R \to \infty$ if the second residual integral (I_1) goes to zero as $R \to \infty$. Now

$$|I_1| \leqslant \int_{-\infty}^{-R} dx'\ \frac{|g(z')|}{|z'-z|}. \tag{48}$$

We now note that on the contour I_1, $|z'| > R$ so that condition A then implies

$$|\tilde{g}(z')| = |e^{b(z'-z)}||g(z')| \leqslant e^{b(x'-x)}\ \frac{M e^{-bx'}}{|z'|^{\beta}} = e^{-bx}\ \frac{M}{|x'|^{\beta}}. \tag{49}$$

Further, noting that $x > 0$ while $x' < -R$,

$$\frac{1}{|z'-z|}<\frac{1}{|x'-x|}<\frac{1}{|x'|}, \tag{50}$$

so that

$$I_1 \leqslant \int_{-\infty}^{-R} dx' \, \frac{Me^{-bx}}{|x'|^{1+\beta}} = \left[\int_R^{\infty} \frac{dw}{w^{1+\beta}} \right] Me^{-bx} = \frac{Me^{-bx}}{\beta R^{\beta}} \to 0, \quad \text{as} \quad R \to \infty, \tag{51}$$

Hence, etc.

C. Proof of the theorem

As $F(z)$ is expressible as the finite sum (16): if we could prove the result for one $F_i(z)$, the result for $F(z)$ follows.

The proof proceeds as follows: We first evaluate $\tilde{F}_i(t)$ of (14) in three steps.

(1) We show that $\tilde{F}_i(t)$ vanishes for $t<b_i$.

(2) We then cast $\tilde{F}_i(t)$ for $t>b_i$ in a form amenable to the evaluation of $\mathcal{L}_z \tilde{F}_i(t)$ via (15).

(3) $\mathcal{L}_z \tilde{F}_i(t)$ is then evaluated as a limit of the integral

$$\lim_{r\to\infty} \int_0^r \tilde{F}_i(t) e^{-zt} \, dt. \tag{52}$$

This we then show indeed is $F_i(z)$, $\forall$ Re $z>c$.

In the following proof, we drop the suffix i on b_i, β_i, and M_i for convenience.

(1) Let $C' \equiv C_0 + C_1' + C_2' - C_3'$. Consider

$$\tilde{F}(t,C') \equiv \oint_{C'} F(z) e^{zt} \, dz = \int_{C_0} F(z) e^{zt} \, dz + \sum_{i=1}^{3} \int_{C_i'} F(z) e^{zt} \, dz. \tag{53}$$

By Lemma 1, for $t<b$,

$$\lim_{C_i' \to \infty} \int_{C_i'} F(z) e^{zt} \, dt \equiv 0. \tag{54}$$

Also, on account of analyticity of $F(z)e^{zt}$ on and within C', $\tilde{F}(t,C') = 0 = \lim_{C' \to \infty} \tilde{F}(t,C')$. Hence, taking the limit $C' \to \infty$ in (53) and noting (54), we obtain, for $t<b$,

$$\tilde{F}(t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{zt} F(z) dz = 0. \tag{55}$$

(2) Now consider, for $t>b$,

$$\tilde{F}(t,C) \equiv \oint_C e^{zt} F(z) dz, \tag{56}$$

where the closed contour C has been defined in Sec. III A. Since C, in particular, continues to contain the same singularities of $F(z)$ as $C \to \infty$*, we have

$$\tilde{F}(t,C) = \lim_{C\to\infty} \tilde{F}(t,C) = 2\pi i \tilde{F}(t) + \sum_{j=1}^{3} \lim_{Cj\to\infty} \int e^{zt} F(z) dz + \lim_{K_j \to \infty} \sum_{j=1}^{B} \int_{K_j} e^{zt} F(z) dz. \tag{57}$$

(The singularities lying on the branch cut (if at all) are always outside the contour C.)

Lemma 1 ensures that the second term on the right-hand side vanishes. This implies that

$$\tilde{F}(t)=\frac{1}{2\pi i}\left[\tilde{F}(t,C)-\lim_{K_j\to\infty}\sum_{j=1}^{B}\int_{K_j}e^{zt}F(z)dz\right]. \tag{58}$$

(3) Now we reexpress

$$\mathcal{L}_z\tilde{F}(t)\equiv\int_0^\infty e^{-zt}\tilde{F}(t)dt\equiv\lim_{r\to\infty}\int_0^r e^{-zt}\tilde{F}(t)dt=\lim_{r\to\infty}\int_b^r e^{-zt}\tilde{F}(t)dt\equiv\lim_{r\to\infty}F(z,r), \tag{59}$$

in view of (55). Now (56) above leads to

$$\begin{aligned}F(z,r)&=\int_b^r e^{-zt}\tilde{F}(t)dt\\&=\frac{1}{2\pi i}\int_b^r e^{-zt}\,dt\left[\tilde{F}(t,C)-\lim_{K_j\to\infty}\sum_{j=1}^{B}\int_{K_j}e^{z't}F(z')dz'\right]\\&\equiv\frac{1}{2\pi i}[J-K].\end{aligned} \tag{60}$$

Now consider the first term on the right-hand side:

$$J\equiv\int_b^r e^{-zt}\,dt\tilde{F}(t,C)=\int_b^r e^{-zt}\,dt\oint_C e^{z't}F(z')dz'. \tag{61}$$

Now the integrals over finite ranges commute. Hence

$$\begin{aligned}J=\oint_C dz'\,F(z')\int_b^r dt\,e^{(z'-z)t}&=\oint_C dz'\,F(z')\,\frac{1}{z'-z}\,[e^{r(z'-z)}-e^{b(z'-z)}]\\&=e^{-zr}\oint_C\frac{F(z')e^{rz'}\,dz'}{z'-z}-\oint_C\frac{F(z')e^{b(z'-z)}\,dz'}{z'-z}\\&\equiv J_1-J_2.\end{aligned} \tag{62}$$

Now consider

$$J_2\equiv\oint_C\frac{F(z')e^{b(z'-z)}\,dz'}{z'-z}\equiv\oint_C H(z')e^{tz'}dz'|_{t=b}. \tag{63}$$

Now, on C_1, C_2, and C_3, in view of the fact that $|z'|>R$, we have

$$|H(z')|=\left|\frac{F(z')}{z'-z}\right|\leqslant\frac{Me^{-b\ \mathrm{Re}\ z'}}{|z'|^\beta|z'-z|}\leqslant\frac{Me^{-b\ \mathrm{Re}\ z'}}{|z'|^{\beta+1}\left|1-\frac{z}{z'}\right|}. \tag{64}$$

For a given z, we now choose C such that on C_1, C_2, and C_3 $|1-z/z'|>1/2$, which is possible without loss of generality, as we are ultimately going to let $C\to\infty$. Then

$$|H(z')|\leqslant 2\,\frac{Me^{-b\ \mathrm{Re}\ z'}}{|z'|^{\beta'}}, \tag{65}$$

where $\beta'=1+\beta>1$. Thus on account of Lemma 1 part (iii),

$$\lim_{C\to\infty}\oint_{C_1+C_2+C_3}H(z')e^{tz'}\,dz'|_{t=b}=0. \tag{66}$$

Thus, in view of the fact that C always encloses the same singularities of $F(z')$,

$$J_2 = \oint_C \frac{F(z')e^{b(z'-z)}\,dz'}{z'-z}$$

$$= \lim_{C\to\infty} \oint_C \frac{F(z')e^{b(z'-z)}\,dz'}{z'-z}$$

$$= \lim_{C_0\to\infty} \int_{C_0} \frac{F(z')e^{b(z'-z)}\,dz'}{z'-z} + \sum_{j=1}^{B} \lim_{K_j\to\infty} \int_{K_j} \frac{F(z')e^{b(z'-z)}\,dz'}{z'-z}. \tag{67}$$

Applying Lemma 1 part (iii) once again, we can replace $C_0 \to C'$ in the first term. Thus

$$J_2 = \lim_{C'\to\infty} \int_{C'} \frac{F(z')e^{b(z'-z)}\,dz'}{z'-z} + \sum_{j=1}^{B} \lim_{K_j\to\infty} \int_{K_j} \frac{F(z')e^{b(z'-z)}\,dz'}{z'-z}. \tag{68}$$

The integrand in the first term on the right-hand side is analytic everywhere on and within (the clockwise contour) C', except for the simple pole at $z'=z$. Hence

$$J_2 = -2\pi i F(z) + \sum_{j=1}^{B} \lim_{K_j\to\infty} \int_C \frac{F(z')e^{b(z'-z)}\,dz'}{z'-z}. \tag{69}$$

Thus, using (60), (62), and (69), we obtain

$$F(z,r) = \frac{1}{2\pi i}[J_1 - J_2 - K] = \frac{1}{2\pi i}\left[J_1 + 2\pi i F(z) - \sum_{j=1}^{B} \lim_{K_j\to\infty} \int_{K_j} \frac{F(z')e^{b(z'-z)}dz'}{z'-z} - K\right]. \tag{70}$$

We now take the limit as $r\to\infty$ in (70) and note that Lemma 2 ensures that $\lim_{r\to\infty} J_1 = 0$ [refer to (62) for a definition]. On the other hand Lemma 3 ensures that

$$\lim_{r\to\infty} K = \sum_{j=1}^{B} \int_b^\infty e^{-zt}\,dt \int_{K_j\to\infty} F(z')e^{z't}\,dz' = \sum_{j=1}^{B} \int_{K_j\to\infty} dz'\; F(z') \int_b^\infty e^{(z'-z)t}\,dt$$

$$= \sum_{j=1}^{B} \int_{K_j\to\infty} dz'\; F(z')\,\frac{(-1)e^{b(z'-z)}}{z'-z}, \tag{71}$$

in view of Re $z > c \geqslant$ Re z' always. Thus, the last two terms on the right-hand side of (70) vanish as $r\to\infty$. Thus, as $r\to\infty$, (70) leads to

$$\mathcal{L}_z \tilde{F}(t) = \lim_{r\to\infty} F(z,r) = F(z), \quad \forall \text{ Re } z > c. \tag{72}$$

(72) is proved for a component F_i of F as in (16). The result (15) then follows from the additive nature of both sides of (15). Hence, etc.

IV. DISCUSSION AND AN APPLICATION

A. Scope of results

Having proved the extension of the Laplace Inverse Theorem, the next natural step is to investigate its utility. For this purpose let us consider a general physical quantity $P(\lambda)$. It seems reasonable to assume that there exists a continuation of $P(\lambda)$ to an entire complex plane of λ,

TABLE I. Description of functions to which the previously existing theorem and our extended version apply.

$\mathcal{F}_1$	$\mathcal{F}_2$
1. Only a finite number of poles in F are allowed, which implies only a finite number of poles in P are permissible.	Only a finite number of isolated singularities are allowed, which implies that a finite number of isolated singularities in P are permissible (these need not be only poles).
2. $\|F\|<1/\|z\|^{\beta}$, $\forall \|z\|>R$ $\Rightarrow F\to 0$ as $z\to\infty$ $\Rightarrow P\to 0$ as $\lambda\to 0$, i.e., $P(0)=0$ and hence functions singular at the origin are not covered.	$\|F\|<e^{-b\,\mathrm{Re}\,z}/\|z\|^{\beta}$, $\forall \|z\|>R$ $\Rightarrow F\to e^{-b\,\mathrm{Re}\,z}/\|z\|^{\beta}$, as $z\to\infty$ (i.e., it is not necessary that $F\to 0$ as $z\to\infty$ in all directions $\Rightarrow P\sim\lambda^{\beta}e^{-b/\lambda}$, $\lambda\to 0$, i.e., some types of essential singularities at $\lambda=0$ are allowed. (In fact, in Field Theory this is the most common type of singularity encountered at $\lambda=0$.)
3. No branch cuts are allowed.	A finite number of branch cuts are allowed.

which is analytic everywhere except possibly at some singularities and branch cuts. Thus, this class of functions (which we will refer to as $\mathcal{F}$) contains the most general physically acceptable functions.

We define

$$F(z)\equiv P\left(\frac{1}{z}\right). \tag{73}$$

We then observe the following.

(1) If there is a singularity at $z=a$ in P, then it is manifested as a singularity in F at $z=1/a$.

(2) If there is a branch cut in P, we get corresponding branch cuts in F. For example, (a) if $P\sim 1/(z-a)^{1/2}$ near $z=a$, $F\sim z^{1/2}/a^{1/2}(1/a-z)^{1/2}$, i.e., branch cuts at $z=0$ and $z=1/a$; and (b) if $P\sim\ln(a+z)$, $F\sim\ln a-\ln z+\ln(z+1/a)$, i.e., branch cuts at $z=-1/a$ and $z=0$.

(3) The behavior of P at $z=0$ is related to the behavior of F as $z\to\infty$.

We are now ready to describe the subclass of $\mathcal{F}$ for which the previously existing theorem and our extended version guarantees the existence of the Laplace Inverse. We will refer to them as $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively (see Table I).

Thus we find that our extended version has definitely enlarged the subclass in $\mathcal{F}$ whose Laplace inverse exists. However, for an application to the Field Theory it is desirable to enlarge the class further, and we discuss this in some detail in Sec. IV C and also suggest how it can be achieved.

B. An application

Let us now look at a physical example where the relevent function belongs to $\mathcal{F}_2-\mathcal{F}_1$. Consider a one-dimensional (1-D) spin chain defined by the action

$$S=\sum_{i=1}^{N}(1-\hat{\phi}_i\cdot\hat{\phi}_{i+1}), \tag{74}$$

where $\hat{\phi}_i$ are unit 3-D vectors. Choosing a free boundary condition, one finds the partition function given by

$$Z_N(f)=\int\prod_i d\hat{\phi}_i\, e^{-S[\phi]/f}=\left(\frac{f}{2}\right)^N(1-e^{-2/f})^N. \tag{75}$$

For illustration purposes we take $N=1$. Then

$$Z(f)=\frac{f}{2}(1-e^{-2/f})=P_1(f)+P_2(f), \tag{76}$$

where $P_1(f)=(f/2)e^{-0/f}=f/2$ and $P_2(f)=(f/2)e^{-2/f}$. Then $F_1(z)=1/2z$ and $F_2(z)=(1/2z)e^{-2z}$. Now, though the previous theorem does not guarantee the existence of the Laplace Inverse, the extended version does; and indeed as we will see soon, the inverse exists,

$$\tilde{F}_1(t)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\frac{e^{tz}}{2z}\,dz,\quad c>0. \tag{77}$$

We can close the contour from the right for $t<0$, since the integral over the right half circle vanishes, and hence the integral is zero (no pole is contained). However, for $t>0$, we have to close the contour from the left and then since the residue at $z=0$ is $-\frac{1}{2}$, we get $\tilde{F}_1(t)=\frac{1}{2}$. Hence, finally we have

$$\tilde{F}_1(t)=\tfrac{1}{2}\theta(t). \tag{78}$$

Similarly, one can compute $\tilde{F}_2(t)$, which comes out to be

$$\tilde{F}_2(t)=-\tfrac{1}{2}\theta(t-2), \tag{79}$$

$$\tilde{F}(t)=\tfrac{1}{2}[\theta(t)-\theta(t-2)]=\tfrac{1}{2}\theta(t)\theta(2-t)=B_3(t). \tag{80}$$

One can verify that by taking the Laplace Transform of B_3 we get back F and hence P,

$$\int_0^\infty e^{-zt}\,\frac{1}{2}\,\theta(t)\theta(2-t)dt=\frac{1}{2}\int_0^2 e^{-zt}\,dt=\frac{1}{2z}(1-e^{-2z}), \tag{81}$$

and we get back $Z(f)=(f/2)(1-e^{-2/f})$.

In Field Theory we also find instances where $P(\lambda)$ have branch cuts. In Ref. 8, for example, it is mentioned that in expression (1), if $a_{k+1}=(1/z_i)^k k^\gamma a_k(\gamma>0)$, then B_1 has a singularity proportional to $(z-z_i)^{-\gamma-1}$; so if γ is a positive integer we have a pole; and for noninteger γ a branch point in the z plane at $z=z_i$. It is easy to check that such a Borel transform indicates branch cuts in the original function $P(\lambda)$.

C. Possible generalization

Let us first try to find the types of functions that belong to $\mathcal{F}-\mathcal{F}_2$.

First, most functions that behave badly at $\lambda=0$, say having poles at 0, does not belong to $\mathcal{F}_2$. Here, however, we present an argument that suggests that such functions can be handled by switching to generalized functions and slightly modifying the definition of the Laplace transform that does not affect our earlier discussion. We define the Laplace transform of $\tilde{F}(t)$ (which can now be a generalized function) as

$$\mathcal{L}_z[\tilde{F}(t)]\equiv\lim_{\epsilon\to 0}\int_{-\epsilon}^{\infty}\tilde{F}(t)e^{-zt}\,dt. \tag{82}$$

Say now $P(\lambda)=a/\lambda^n$, $n\geqslant 0$, is an integer. Then $F(z)=az^n$, and we have

$$\tilde{F}(t)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}F(z)e^{tz}\,dz=\frac{1}{2\pi}\int_{-\infty}^{\infty}(c+iy)^n e^{ct}e^{iyt}\,dy. \tag{83}$$

Now we also have $2\pi\delta(t)=\int_{-\infty}^{\infty}e^{iyt}\,dy$ or

$$2\pi[\delta(t)e^{ct}]_n=\int_{-\infty}^{\infty}(c+iy)^n e^{ct}e^{iyt}\,dy \tag{84}$$

(the suffix n denotes differentiation with respect to t). Thus

$$\tilde{F}(t)=a[\delta(t)e^{ct}]_n=a\,\delta^n(t). \tag{85}$$

Now

$$\mathcal{L}_z[\tilde{F}(t)]=\lim_{\epsilon\to 0}\int_{-\epsilon}^{\infty}\tilde{F}(t)e^{-zt}\,dt=\lim_{\epsilon\to 0}a\int_{-\epsilon}^{\infty}\delta^n(t)e^{-zt}\,dt=a(-z)^n(-1)^n=az^n. \tag{86}$$

Thus, we see that it works and, in fact, it will now work for any polynomial in λ also.

[1] See, for example, T-P. Cheng and L. F. Li, *Gauge Theory of Elementary Particle Physics* (Clarendon, Oxford, 1984).
[2] See, e.g., R. Rajaraman, *Solitons and Instantons* (North-Holland, New York, 1982).
[3] For a discussion of asymptotic series, see, e.g., E. T. Whittaker and C. N. Watson, *A Course of Modern Analysis*, 4th ed. (University Press, Cambridge, 1952).
[4] See, e.g., C. G. Callan, Jr., R. Dashen, and D. Gross, Phys. Lett. B **63**, 334 (1976).
[5] See, for example, the discussion in S. Weinberg, *The Quantum Theory of Fields* (Cambridge University Press, Cambridge, 1996), Vol. 2, and references therein.
[6] A. H. Mueller, Nucl. Phys. B **250**, 327 (1985).
[7] A. Duncan and S. Pernice, Phys. Rev. D **51**, 1956 (1995).
[8] C. N. Lovett-Turner and C. J. Maxwell, Nucl. Phys. B **432**, 147 (1994).
[9] J. Fischer, Int. J. Mod. Phys. A **12**, 3625 (1997).
[10] A. H. Mueller, Phys. Lett. B **308**, 355 (1993).
[11] See any standard Complex analysis textbook for, e.g., E. Marsden, *Basic Complex Analysis* (Freeman & Co., San Francisco, 1973).
[12] The Tonelli–Hobson Theorem applies for two-dimensional Lebesgue measurable functions and it is well known that piecewise continuous functions (since f is analytic and hence continuous the functions we are dealing with trivially satisfy this condition) are measurable. For a reference see T. Apostol, *Mathematical Analysis* (Addison-Wesley, New York, 1974).

The Eikonal equation in flat space: Null surfaces and their singularities. I

Simonetta Frittelli
Physics Department, Duquesne University, Pittsburgh, Pennsylvania 15282 and Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Ezra T. Newman
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Gilberto Silva-Ortigoza
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260 and Facultad de Ciencias Físico Matemáticas de la Universidad Autónoma de Puebla, Apartado Postal 1152, 72001 Puebla, Pue., México

(Received 8 June 1998; accepted for publication 3 August 1998)

The level surfaces of solutions to the eikonal equation define null or characteristic surfaces. In this paper we study, in Minkowski space, properties of these surfaces. In particular, we are interested both in the singularities of these ''surfaces'' (which can, in general, self-intersect and be only piecewise smooth) and in the decomposition of the null surfaces into a one-parameter family of two-dimensional wavefronts which can also have self-intersections and singularities. We first review a beautiful method for constructing the general solution to the flat-space eikonal equation; it allows for solutions either from arbitrary Cauchy data or for time-independent (stationary) solutions of the form $S=t-S_0(x,y,z)$. We then apply this method to obtain global, asymptotically spherical, null surfaces that are associated with shearing (''bad'') two-dimensional cuts of null infinity; the surfaces are defined from the normal rays to the cut. This is followed by a study of the caustics and singularities of these surfaces and those of their associated wavefronts. We then treat the same set of issues from an alternative point of view, namely from Arnold's theory of generating families. This treatment allows one to deal (parametrically) with the regions of self-intersection and nonsmoothness of the null surfaces, regions which are difficult to treat otherwise. Finally, we generalize the analysis of the singularities to the case of families of characteristic surfaces. © *1999 American Institute of Physics.* [S0022-2488(99)00801-4]

I. INTRODUCTION

The study of the propagation of electromagnetic wavefronts, in the high-frequency or geometric optics limit, is almost ubiquitous in physics; it is a basic staple of elementary physics courses, it arises in the practical area of optical equipment, in applied subjects, too numerable to mention in detail, involving materials with variable refractive index and in atmospheric and astrophysical studies. They have been a prime illustration of V. I. Arnold's theory of Lagrangian and Legendre maps.[1–4] In a different guise, similar problems arise in catastrophe theory. In addition to the various applications to more standard physics problems, they also play a most fundamental role in general relativity, e.g., the continuous propagation of the two-dimensional wavefronts, i.e., the one-parameter family of evolving wavefronts, form null (or characteristic) three-surfaces that are determined by the dynamics of the curved space–time in which the wavefronts propagate.[5,6] In this context it forms the basis for the theory of gravitational lensing.[7] The converse statement is also true, namely that sets of null surfaces define, up to a conformal factor, the space–time geometry itself.[8,9] In arbitrary space–times, the high-frequency limit is completely governed by the eikonal equation,

$$g^{ab}\partial_a S_b S = 0, \tag{1}$$

where $x^a = (x^i, t)$ are local space–time coordinates, $g^{ab}(x^c)$ is a given space–time metric, and $S = S(x^a)$. The level surfaces of S, i.e., $S(x^a) = \text{const}$, define the characteristic or null three-surfaces (or what Arnold calls the ''big wave fronts''[2]) and the $S(x^i, t) = \text{const}$ for const t define the two-dimensional ''small'' wavefront in the three-dimensional space of $t = \text{const}$. The vector $p^a = g^{ab}\partial_b S$ is tangent to the null geodesics that rule the characteristic surface. Though we are basically interested in Eq. (1) for arbitrary space–times, here we will confine ourselves to a study of its solutions only in flat (Minkowski) space. (A future paper, in preparation, will generalize the present material to curved space–times.) Equation (1) then becomes

$$\eta^{ab}\partial_a S \partial_b S = (\partial_t S)^2 - (\partial_x S)^2 - (\partial_y S)^2 - (\partial_z S)^2 = 0. \tag{2}$$

The level surfaces of the solutions to Eq. (1) or (2), the null surfaces, can be viewed as being generated by the evolution of two-dimensional wavefronts. Specifically, a wavefront evolves by following light rays that are normal to it, generating the null three-surface. A smooth wavefront in three-space, in general, progresses into a singular one, either to the past or the future, i.e., a generic null surface in space–time has singularities. The singular wavefronts are two-surfaces that are continuous with existing first derivatives, but where (piecewise) the second derivatives are singular, being either undefined or infinite. The structure of the singularities are generically cusp ridges and swallowtails. There are unstable exceptions.

A textbook example[2,3] of flat space singular wavefronts (and associated big wavefront) are from imploding triaxial ellipsoids, where an initially ellipsoidal wavefront is evolved inwardly, self-intersects for some finite period of time, and eventually expands out to infinity, becoming spherical in the limit.

The singularities of wavefronts are also interpreted as the location of focusing regions, where the intensity of light becomes very high. At the focusing regions, neighboring null geodesics meet, and the cross-sectional area of the bundle of light rays collapses to zero, which leads to the increase in intensity. Spherical wavefronts focus at a single point (which are unstable under small perturbations of the front) whereas generic wavefronts trace spatial curves of focusing points (cusp ridges and swallowtails).

In Sec. II we will review a beautiful method for giving the general solution to the flat-space eikonal; it allows for solutions either from arbitrary Cauchy data or for stationary solutions that arise from the ansatz, $S = t - S_0(x, y, z)$.

In Sec. III, we will apply the method of Sec. II to obtain global asymptotically spherical null surfaces that are associated with shearing (''bad'') cuts of null infinity.[10,11] They will be defined from the normal rays to a ''bad'' cut. This construction can be thought of as beginning with a deformed, initial, two-sphere in a finite region of space–time. Then, construct the future outward directed null normals to the two surface which generates a null surface and finally ''slide'' the initial two-surface along the null geodesics that generate the null surface, to future null infinity. This limit is the ''bad'' cut of null infinity.

In Sec. IV, we will study the caustics and singularities of these characteristic surfaces and their associated wavefronts.

In Sec. V, we treat the same problems of the singularities of these surfaces but now from an alternate point of view, namely from Arnold's theory of generating families.[4] This treatment allows one to handle (parametrically) the regions of self-intersection and nonsmoothness of the null surfaces.

In Sec. VI, we discuss a generalization of the ideas presented to this point. Though this generalization is primarily intended for use in nonflat Lorentzian space–times, nevertheless we believe that it is quite useful to see it in the simpler case of flat space–time; it allows for the clarification of certain points that would be difficult in more general situations. Specifically we will consider solutions of the eikonal equation that depend on two parameters—that are different from the two-parameter family of plane waves. We will see the slightly surprising result that the singularity structure of the individual characteristic surfaces can be studied via the parameter behavior of nearby solutions. More precisely, if the two-parameter set of solutions is given by $Z(x^a,\mu,\bar{\mu})$, the singularities of the level surfaces of Z for fixed values of $(\mu,\bar{\mu})$ can be studied and expressed in terms of $(\mu,\bar{\mu})$ derivatives of Z. These results become important in asymptotically flat space–times where the $Z(x^a,\mu,\bar{\mu})$ can be chosen to represent the family of past null cones from all the points of future null infinity.

II. SOLUTIONS OF THE EIKONAL EQUATION

We review a powerful method for solving the flat-space eikonal equation with arbitrary given Cauchy data. We begin with a solution S^* of the eikonal equation that depends on three arbitrary parameters, i.e.,

$$S^* = S^*(x^i,t,\alpha_i) = x^i\alpha_i - t\sqrt{\Sigma(\alpha_i)^2} \tag{3}$$

called a complete integral. A ''general integral'' (which involves an arbitrary function) can be constructed from the complete integral in the following manner: we first add to it an arbitrary function of the three α_i, i.e., we consider

$$S^{**} = S^*(x^i,t,\alpha_i) - F(\alpha_i), \tag{4}$$

with the weak condition that (aside from lower-dimensional regions)

$$\left|\frac{\partial^2 S^{**}}{\partial\alpha_i\partial\alpha_j}\right| \neq 0. \tag{5}$$

We next demand that $\partial S^{**}/\partial\alpha_i = \partial S^*/\partial\alpha_i - \partial F/\partial\alpha_i = 0$, which implies that there are three functions of the form $\alpha_i = A_i(x^i,t)$. (In general these solutions are not unique and they must be expressed on different sheets. See Sec. IV for a complete discussion of this issue.) Finally, via $\alpha_i = A_i(x^i,t)$, the α_i are eliminated in the S^{**}, yielding (perhaps multivalued)

$$S(x^i,t) = S^*(x^i,t,A_i(x^i,t)) - F(A_i(x^i,t)). \tag{6}$$

The level surfaces of this S might self-intersect and be only piecewise differentiable.

It is not difficult to show that the S so constructed satisfies the eikonal equation.[12] This follows immediately from the fact that

$$\partial_a S = \partial_a S^* + (\partial_{A_i}S^* - \partial_{A_i}F)\partial_a A_i = \partial_a S^*.$$

This solution now depends on an arbitrary function of three variables, namely the F. The task is now to determine the $F(\alpha_i)$ in terms of (appropriate) Cauchy data, $S_C(x^i)$. This is accomplished as follows; define $\alpha_i = \partial S_C/\partial x^i$ and algebraically invert it in the form of the three equations $x^i = X^i(\alpha_i)$. At $t = t_0$ we have that

$$S(x^i,t_0) = S^*(x^i,t_0,A_i(x^i,t_0)) - F(A_i(x^i,t_0)). \tag{7}$$

Replacing all the A_i by α_i and all the x_i by $X^i(\alpha_i)$, we have that

$$F(\alpha_i) = S^*(X^i,t_0,\alpha_i) - S_C(X^i), \tag{8}$$

i.e., the free $F(\alpha_i)$ is now expressed in terms of the free Cauchy data, $S_C(x^i)$.

This allows us to find (in principle, modulo algebraic inversions or parametrizations) solutions of the flat-space eikonal equation with arbitrary Cauchy data.

There exists a special class of solutions that are not studied or found easily via the Cauchy problem, namely the ''stationary'' solutions which have the form

$$S=t-S_0(x^i).$$

To generate solutions of this form we modify the complete integral, Eq. (3), making it a function of only two free parameters by imposing the condition that $\Sigma(\alpha_i)^2=\text{const}$ [for convenience chosen as $\Sigma(\alpha_i)^2=1/\sqrt{2}$].

We then write the modified complete integral as

$$S^*=x^a l_a(\zeta,\overline{\zeta}), \tag{9}$$

where

$$l_a(\zeta,\overline{\zeta})=\frac{1}{\sqrt{2}(1+\zeta\overline{\zeta})}[(1+\zeta\overline{\zeta}),-(\zeta+\overline{\zeta}),-i(\overline{\zeta}-\zeta),(1-\zeta\overline{\zeta})]. \tag{10}$$

The complex number ζ, which plays the role of two of the independent parameters among the three α_i, can be thought of as the complex stereographic coordinate on the sphere; the $l_a(\zeta,\overline{\zeta})$ is a Minkowski null vector $\eta^{ab}l_a l_b=0$ that spans the entire lightcone as ζ ranges over the sphere. Equation (9) represents a spheres worth of different families of plane waves parametrized by the direction ζ.

If we now take

$$S^{**}=x^a l_a(\zeta,\overline{\zeta})+\alpha(\zeta,\overline{\zeta}) \tag{11}$$

and construct $\eth S^{**}=\bar{\eth} S^{**}=0$, i.e.,

$$\omega\equiv x^a m_a(\zeta,\overline{\zeta})+\eth\alpha(\zeta,\overline{\zeta})=0, \tag{12}$$

$$\overline{\omega}\equiv x^a \overline{m}_a(\zeta,\overline{\zeta})+\eth\alpha(\zeta,\overline{\zeta})=0, \tag{13}$$

where

$$m_a(\zeta,\overline{\zeta})=\eth l_a(\zeta,\overline{\zeta})\equiv(1+\zeta\overline{\zeta})\,\frac{\partial l_a(\zeta,\overline{\zeta})}{\partial\zeta}, \tag{14}$$

$$\overline{m}_a(\zeta,\overline{\zeta})=\bar{\eth} l_a,(\zeta,\overline{\zeta})\equiv(1+\zeta\overline{\zeta})\,\frac{\partial l_a(\zeta,\overline{\zeta})}{\partial\overline{\zeta}}. \tag{15}$$

For generic $\alpha(\zeta,\overline{\zeta})$, Eq. (12) can be solved for

$$\zeta=\Upsilon(x,y,z), \tag{16}$$

where again these solutions need not be unique and must be expressed on different sheets. (See Sec. IV for a full treatment of this problem.) Note that Eq. (12) does not depend on the time t and hence Υ is a function only of the spatial coordinates. When the $\Upsilon(x,y,z)$ is substituted into Eq. (11), i.e.,

$$S(t,x,y,z)=\frac{t}{\sqrt{2}}-S_0(x,y,z)=\frac{t}{\sqrt{2}}+x^i l_i(\Upsilon,\overline{\Upsilon})+\alpha(\Upsilon,\overline{\Upsilon}), \tag{17}$$

we have a solution of the eikonal equation depending on an arbitrary function of two variables, $\alpha(\Upsilon,\overline{\Upsilon})$. The level surfaces of S could in general self-intersect and be only piecewise differentiable.

The procedure of beginning with a complete solution and obtaining the general solution via the two (or three) variable arbitrary function *is* geometrically equivalent to the construction of an envelope from the family of plane waves as the two (or three) constants in the complete solutions are varied.[13]

Since in this work we will only be interested in individual null surfaces and their properties, we can and will confine ourselves to the level surfaces of the solutions of the form given in Eq. (17).

III. NULL SURFACES GENERATED BY NORMALS TO TWO-SURFACES

We want to give a slightly different geometric interpretation to the method of the previous section for generating the stationary solutions of the eikonal equation. Given any (spatial) two-surface (for example, consider any two-dimensional slice of the past light cone of an arbitrary space–time point), the normal rays to the surface (either the outgoing or incoming ones) generate a null surface. In this section we will consider a particular case of this construction where this *past light cone* is taken to be the future null infinity, $\mathfrak{I}^+$, of Penrose.[10]

The future null boundary, $\mathfrak{I}^+$ (the endpoints of future directed null geodesics), of any asymptotically simple space–time is a null surface with topology $R\times S^2$. A choice of Bondi coordinates $(u,\zeta,\bar{\zeta})$ can be assigned to $\mathfrak{I}^+$, where $\zeta=\cot(\theta/2)e^{i\phi}$ for the S^2 sector. The intersection of the future lightcone $\mathfrak{C}_x$ of a point x^a with $\mathfrak{I}^+$ is a two-surface, locally imbedded in $R\times S^2$; it can generically be described locally by $u=Z(x^a,\zeta,\bar{\zeta})$. The two-surface is referred to as a light cone cut,[14] whereas the function $Z(x^a,\zeta,\bar{\zeta})$ is referred to as a light cone cut function, and is a two-point real function on the space–time and the boundary, $\mathfrak{I}^+$.

Remark 1: Though for this work it is irrelevant, we mention that the light cone cut function $Z(x^a,\zeta,\bar{\zeta})$ is one of two fundamental variables in a reformulation of general relativity via null surfaces.[8,9] It encodes all the conformal information of the space–time.

In the remainder of this work we will confine ourselves to flat space–time where (modulo Poincare transformations) a natural choice of Bondi coordinates $(u_n,\zeta,\bar{\zeta})$ exists; the $u_n=\text{const}$ is constructed from the intersection of the future light cone, $\mathfrak{C}_{(t,o,o,o)}$, of the spatial origin, at time $t=u_n$, with $\mathfrak{I}^+$; the $(\zeta,\bar{\zeta})$ are just the null directions, at the origin, carried along by the null generators of the lightcone. Using Cartesian coordinates x^a for the space–time and these natural Bondi coordinates,[15] the light cone cuts can be described as

$$u_n=x^a l_a(\zeta,\bar{\zeta}), \tag{18}$$

where l_a represents the covariant version of a null vector l^a with Cartesian components given as

$$\begin{aligned} l^a(\zeta,\bar{\zeta}) &= \frac{1}{\sqrt{2}(1+\zeta\bar{\zeta})}\,((1+\zeta\bar{\zeta}),(\zeta+\bar{\zeta}),i(\bar{\zeta}-\zeta),(\zeta\bar{\zeta}-1)) \\ &= \frac{1}{\sqrt{2}}\,(1,\sin\,\theta\,\cos\,\phi,\sin\,\theta\,\sin\,\phi,\cos\,\theta). \end{aligned} \tag{19}$$

Adding the radial coordinate r, this natural choice of Bondi coordinates is identical to the standard null polar coordinates $(u_n,r,\zeta,\bar{\zeta})$ given by

$$x^a=ut^a+rl^a(\zeta,\bar{\zeta}),\quad t^a=\sqrt{2}(1,0,0,0). \tag{20}$$

Note that Eqs. (18) and (19) are identical with Eqs. (9) and (10) though their meanings are different; Eq. (18) has the dual meaning of being, for fixed value of the x^a, the light cone cut of $\mathfrak{I}^+$ and also, for fixed values of $(u_n,\zeta,\bar{\zeta})$, it describes the plane wave (null surface) intersecting the time axis at $t=u_n$ in the direction of $(\zeta,\bar{\zeta})$.

By Eq. (18), the light cone cuts of any points $x^a=(t,0,0,0)$ along the time axis take, as we mentioned earlier, a constant value on $\mathfrak{I}^+$, namely, they are the constant-u_n slices. The natural Bondi cuts are light cone cuts as well. By following inwardly the null geodesics that leave the natural Bondi cuts orthogonally, we find no focusing other than at the apex (on the time axis) of the light cone.

By a slight modification of the above we can find other null surfaces leaving $\mathfrak{I}^+$ that have much more complicated focusing properties than that of a light cone. If we consider the one-parameter family of cuts of $\mathfrak{I}^+$ given, say by,

$$u_n=-\alpha(\zeta,\bar{\zeta})+u, \tag{21}$$

where $\alpha(\zeta,\bar{\zeta})$ is a given but arbitrary regular function on S^2, and u is a parameter on R, we can ask for the null surfaces generated by the null normals to the family of cuts.

• Note that Eq. (21) can be rewritten as $u=u_n+\alpha(\zeta,\bar{\zeta})$ and reinterpreted as a [Bondi–Metzner–Sachs (BMS)] supertranslation[11] between the coordinates u_n and u on $\mathfrak{I}^+$.

We now construct the null surface formed by the normal rays to the cuts, Eq. (21), determined by $u=$const; replacing the u_n in Eq. (21) by the null planes Eq. (18), $[u_n=x^a l_a(\zeta,\bar{\zeta})]$, we have

$$u=x^a l_a(\zeta,\bar{\zeta})+\alpha(\zeta,\bar{\zeta}), \tag{22}$$

which is identical to Eq. (11). The envelope formed from all the null planes that are normal to the family of cuts are found by setting to zero the ζ and $\bar{\zeta}$ derivatives of Eq. (22) and eliminating the $(\zeta,\bar{\zeta})$ from (22), a procedure *identical* to that followed in the previous section to obtain Eq. (17), i.e., we now have the one-parameter family of null surfaces

$$S^{**}\equiv u=\frac{t}{\sqrt{2}}+x^i l_i(\Upsilon,\bar{\Upsilon})+\alpha(\Upsilon,\bar{\Upsilon}) \tag{23}$$

with $\zeta=\Upsilon(x,y,z)$, a solution of Eqs. (12) and (13). The procedure of setting to zero the ω and $\bar{\omega}$ in

$$\omega\equiv x^a m_a(\zeta,\bar{\zeta})+\bar{\eth}\alpha(\zeta,\bar{\zeta})=0, \tag{24}$$

$$\bar{\omega}\equiv x^a \bar{m}_a(\zeta,\bar{\zeta})+\bar{\eth}\alpha(\zeta,\bar{\zeta})=0 \tag{25}$$

selects the null ray at each point of $\mathfrak{I}^+$ that is orthogonal to the cut given by Eq. (21).

• Note that α can be chosen to contain only spherical harmonics of order $l\geq 2$ since any $l=0,1$ components of α could be absorbed by $x^a l_a$ with no modification other than displacing the origin of the coordinates x^a, since l_a is precisely the collection of spherical harmonics of order 0 and 1.

We can give a parametric description of the family of null surfaces, Eq. (23), by the following procedure: we consider the four functions

$$\zeta=\Upsilon(x,y,z)\Leftrightarrow x^a m_a(\zeta,\bar{\zeta})+\eth\alpha(\zeta,\bar{\zeta})=0, \tag{26}$$

$$\bar{\zeta}=\bar{\Upsilon}(x,y,z)\Leftrightarrow x^a \bar{m}_a(\zeta,\bar{\zeta})+\bar{\eth}\alpha(\zeta,\bar{\zeta})=0, \tag{27}$$

$$u=u(t,x,y,z)=\{x^a l_a(\zeta,\bar{\zeta})+\alpha(\zeta,\bar{\zeta})\}|_{\Upsilon,\bar{\Upsilon}}, \tag{28}$$

$$r=r(x,y,z)=\{x^a(n_a-l_a)(\zeta,\bar{\zeta})+\eth\bar{\eth}\alpha(\zeta,\bar{\zeta})\}|_{\Upsilon,\bar{\Upsilon}}, \tag{29}$$

and consider them as a coordinate transformation between the $\{u,\zeta,\overline{\zeta},r\}$ and the $\{x^a\}$. We have used $m^a \equiv \eth l^a$, $\overline{m^a} \equiv \bar{\eth} l^a$, and $n^a \equiv \eth\bar{\eth} l^a + l^a$. From the fact that $(l^a,m^a,\overline{m^a},n^a)$ form a null tetrad for every fixed value of $(\zeta,\overline{\zeta})$, this coordinate transformation can be readily inverted into the form

$$x^a = (u-\alpha)(n^a+l^a) + (r-\eth\bar{\eth}\alpha)l^a + (\eth\alpha)\overline{m^a} + (\bar{\eth}\alpha)m^a. \tag{30}$$

This relationship can alternatively be looked on as the parametric version of the one-parameter family of null surfaces, Eq. (23) [where, for fixed u, the $(r,\zeta,\overline{\zeta})$ parametrize the surface], or as the coordinate transformation between the $\{x^a\}$ and the null-geodesic coordinates, $(u,\zeta,\overline{\zeta},r)$; u labels the null surfaces, the pair $(\zeta,\overline{\zeta})$ labels null geodesics (via their intersection with $\mathfrak{I}^+$), and r is an affine parameter along the null geodesics. That this is true can be easily seen from the parametric form, Eq. (30), by simply constructing

$$\frac{dx^a}{dr} = l^a(\zeta,\overline{\zeta}) \tag{31}$$

and observing that l^a is a null tangent vector with affine normalization.

The transformation between the x^a and the $(u,r,\zeta,\overline{\zeta})$ breaks down when the Jacobian of the transformation, Eq. (30), vanishes, i.e., when

$$D \equiv \frac{\partial(t,x,y,z)}{\partial(u,r,\zeta,\overline{\zeta})} = r^2 - \eth^2\alpha\bar{\eth}^2\alpha = 0. \tag{32}$$

Geometrically, this is where the null surface develops singularities. In the projection to the three-space (x,y,z) it is a two-surface; the ''caustic surface.'' To see this explicitly, we return to Eq. (30) where we have (for fixed u) that $(x,y,z) = x^i = X^i(r,\zeta,\overline{\zeta})$, i.e., are known functions of $(r,\zeta,\overline{\zeta})$. If the r in X^i is replaced by the r from Eq. (32) we have the parametric form of the caustic,

$$x^i = X^i(r(\zeta,\overline{\zeta}),\zeta,\overline{\zeta}) = \hat{X}^i(\zeta,\overline{\zeta}). \tag{33}$$

We will return to this in the next section.

It is interesting to note that the coordinates $(u,\zeta,\overline{\zeta},r)$ represent a type of null coordinate system that we could call *asymptotic null-polar coordinates* which are the flat space case of an *interior Bondi coordinate system*,[16,11] i.e., the extension into the interior of the space–time of the Bondi coordinates $(u,\zeta,\overline{\zeta})$ on $\mathfrak{I}^+$. They differ from the standard null polar coordinates by the fact that the null geodesics that rule these surfaces possess nonvanishing shear while for the standard ones the shear vanishes.

The complex shear is defined as $\sigma = M^aM^b\nabla_aL_b$, where L_b is tangent to the null geodesics and M^a is complex null, orthogonal to L_b, and such that $M^a\overline{M}_a = -1$. In our case, because of Eqs. (30) and (31) and the fact that $(l^a,m^a,\overline{m^a},n^a)$ forms a null tetrad, we have that $L_b = l_b$ and $M^a = m^a$. Furthermore, the gradient of l_b is $\nabla_a l_b = m_a\zeta_{,b} + \overline{m}_a\overline{\zeta}_{,b}$ and thus $\sigma = -m_a^b\overline{\zeta}_{,b}$. To obtain the derivative of $\overline{\zeta}$ along m^b we take the gradients of Eqs. (24) and (25) which yields

$$m_b(\zeta,\overline{\zeta}) + \{x^a(n_a - l_a) + \eth\bar{\eth}\alpha\}\overline{\zeta}_{,b} + \eth^2\alpha\zeta_{,b} = 0, \tag{34}$$

$$\overline{m}_b(\zeta,\overline{\zeta}) + \{x^a(n_a - l_a) + \eth\bar{\eth}\alpha\}\zeta_{,b} + \bar{\eth}^2\alpha\overline{\zeta}_{,b} = 0. \tag{35}$$

Using Eq. (29) and contracting Eqs. (34) and (35) by m^b, we obtain

$$rm^b\overline{\zeta}_{,b} + \bar{\eth}^2\alpha m^b\zeta_{,b} = 0, \tag{36}$$

$$-1+rm^b\zeta_{,b}+\bar{\eth}^2\alpha m^b\overline{\zeta_{,b}}=0. \tag{37}$$

By eliminating the $m^b\zeta_{,b}$ from these equations, we find

$$(r^2-\eth^2\alpha\bar{\eth}^2\alpha)m^b\overline{\zeta_{,b}}+\eth^2\alpha=0 \tag{38}$$

or

$$\sigma=\frac{\sigma^0}{r^2-\sigma^0\overline{\sigma^0}}, \tag{39}$$

where $\sigma^0=\eth^2\alpha$. This is also a confirmation of the Sachs theorem on the transformation of the asymptotic shear, σ^0, under a BMS transformation. Equation (39) represents a special (nontwisting) case of a more general result valid for generic null congruences in flat space.[17]

We point out that the flat-space line element, using Eq. (30), can easily be expressed in terms of these shearing Bondi coordinates as[18]

$$ds^2=\eta_{ab}dx^a dx^b=2du\left(du+dr-\bar{\eth}\eth^2\alpha\,\frac{d\zeta}{P}-\eth\bar{\eth}^2\alpha\,\frac{d\overline{\zeta}}{P}\right)-\frac{2r}{P^2}\,(\eth^2\alpha d\zeta^2+\bar{\eth}^2\alpha d\overline{\zeta}^2)$$
$$-2(r^2+\eth^2\alpha\bar{\eth}^2\alpha)\,\frac{d\zeta d\overline{\zeta}}{P^2}, \tag{40}$$

where $P=1+\zeta\overline{\zeta}$. The determinant of g_{ab} is given by $|g|=(1/P^4)\{r^2-\eth^2\alpha\bar{\eth}^2\alpha\}^2=D^2/P^4$ whose vanishing agrees with the vanishing of the Jacobian, Eq. (32).

Note that the asymptotic $r=\text{const}\Rightarrow\infty$ sections, at $u=\text{const}$, become metric spheres.

• It is perhaps interesting to speculate on the use of Eq. (40) as the Minkowski space lowest-order term, in perturbation calculations, for solutions of the Einstein equations.

We complete this section by showing how a null surface can be constructed explicitly from the normals to an arbitrary spacelike two surface, $\mathfrak{S}$, in a manner virtually identical to those constructed from a cut or slice of null infinity.

We begin from Eq. (22)

$$u=0=x^a l_a(\zeta,\overline{\zeta})+\alpha(\zeta,\overline{\zeta})$$

with Eqs. (24) and (25). The issue is, given the surface $\mathfrak{S}$, how is one to choose $\alpha(\zeta,\overline{\zeta})$?

First $\mathfrak{S}$ is defined parametrically, $x^a=x_0^a(\zeta,\overline{\zeta})$, where the parameters are chosen as follows: consider a timelike world line at the spacial origin and the family of light cones centered on the line. The null geodesics ruling these cones are labeled by their directions $(\zeta,\overline{\zeta})$ on the sphere and coincide with labeling of the generators of null infinity. The points on $\mathfrak{S}$ are now (locally) parametrized by the labels of the null geodesics passing thru those points. With this parametrization the function $\alpha(\zeta,\overline{\zeta})$ is defined by

$$\alpha(\zeta,\overline{\zeta})=-x_0^a(\zeta,\overline{\zeta})l_a(\zeta,\overline{\zeta})$$

so that Eq. (23) becomes

$$u=0=(x^a-x_0^a(\zeta,\overline{\zeta}))l_a(\zeta,\overline{\zeta})$$

and Eqs. (24) and (25) become

$$(x^a-x_0^a(\zeta,\overline{\zeta}))m_a(\zeta,\overline{\zeta})-l_a(\zeta,\overline{\zeta})\eth x_0^a=0 \quad \text{and} \quad \text{c.c.} \tag{41}$$

We see that the null surface so defined goes thru $\mathfrak{S}$, i.e., thru $x^a = x_0^a(\zeta,\bar{\zeta})$. To see that it is also normal to $\mathfrak{S}$, we notice that at $\mathfrak{S}$ the first two terms of Eq. (41) cancel out and we are left at $\mathfrak{S}$ with

$$l_a(\zeta,\bar{\zeta})\eth x_0^a = 0.$$

Thus as was claimed, the tangent vectors to $\mathfrak{S}$, namely $\eth x_0^a(\zeta,\bar{\zeta})$, are normal to the null tangent vectors to the null surface, l_a.

We see that the earlier construction of null surfaces from cuts of null infinity actually includes those constructed from finite surfaces.

IV. WAVEFRONT EVOLUTION AND SINGULARITIES

In the previous section, we mentioned that the null coordinate system broke down and the associated shearing null surfaces developed caustics at the points where $r^2 = \eth^2\alpha\bar{\eth}^2\alpha$.

Here we focus our attention on the two-dimensional wavefronts associated with these null surfaces. We show that the wavefronts develop singularities, and we locate the singularities via the standard method of singularity theory, and via our light cone cut approach. The evolution of these singularities as the wave fronts evolve become the caustics.

A wavefront is, by definition, the intersection of our null surface, $u = u_0$, with a constant-time t_0 surface. This represents an instant in the progression of a wave. In our case, this requires fixing the time coordinate $x^0 = t_0$ in Eq. (30) and solving for

$$r = \sqrt{2}t_0 - 2u_0 + 2\alpha + \eth\bar{\eth}\alpha. \tag{42}$$

The remaining coordinates (x,y,z), using Eq. (42) to eliminate r, trace a two-surface (a ''small'' wavefront) in the Euclidean three-space, parametrized by $(\zeta,\bar{\zeta})$ [or (θ,ϕ) under the transformation $\zeta = e^{i\phi}\cot\theta/2$]:

$$x = \frac{1}{\sqrt{2}}\left[(\sqrt{2}t_0 - 2u_0 + 2\alpha)\,\frac{(\zeta+\bar{\zeta})}{(1+\zeta\bar{\zeta})} + \bar{\eth}\alpha\,\frac{(1-\bar{\zeta}^2)}{(1+\zeta\bar{\zeta})} + \eth\alpha\,\frac{(1-\zeta^2)}{(1+\zeta\bar{\zeta})}\right], \tag{43}$$

$$y = \frac{i}{\sqrt{2}}\left[(\sqrt{2}t_0 - 2u_0 + 2\alpha)\,\frac{(\bar{\zeta}-\zeta)}{(1+\zeta\bar{\zeta})} - \bar{\eth}\alpha\,\frac{(1+\bar{\zeta}^2)}{(1+\zeta\bar{\zeta})} + \eth\alpha\,\frac{(1+\zeta^2)}{(1+\zeta\bar{\zeta})}\right], \tag{44}$$

$$z = \frac{1}{\sqrt{2}}\left[(\sqrt{2}t_0 - 2u_0 + 2\alpha)\,\frac{(\zeta\bar{\zeta}-1)}{(1+\zeta\bar{\zeta})} + \bar{\eth}\alpha\,\frac{2\bar{\zeta}}{(1+\zeta\bar{\zeta})} + \eth\alpha\,\frac{2\zeta}{(1+\zeta\bar{\zeta})}\right]. \tag{45}$$

The map $(\zeta,\bar{\zeta}) \to (x,y,z)$ is singular at points where the Jacobian matrix

$$\begin{pmatrix} \eth x & \bar{\eth}x \\ \eth y & \bar{\eth}y \\ \eth z & \bar{\eth}z \end{pmatrix} \tag{46}$$

drops rank, from 2 to 1 or 0. The drop in rank takes place if the three two-determinants vanish simultaneously:

$$\eth x\bar{\eth}y - \eth y\bar{\eth}x = 0, \tag{47}$$

$$\eth y\bar{\eth}z - \eth z\bar{\eth}y = 0, \tag{48}$$

$$\eth z \bar{\eth} x - \eth x \bar{\eth} z = 0. \tag{49}$$

Since

$$\eth x^a|_{u,t} = \eth^2 \alpha \overline{m}^a + r m^a, \tag{50}$$

$$\bar{\eth} x^a|_{u,t} = \bar{\eth}^2 \alpha m^a + r \overline{m}^a, \tag{51}$$

the explicit expressions of the two-determinants are as follows:

$$\eth x \bar{\eth} y - \eth y \bar{\eth} x = (r^2 - \eth^2 \alpha \bar{\eth}^2 \alpha)(m^x \overline{m}^y - \overline{m}^x m^y), \tag{52}$$

$$\eth y \bar{\eth} z - \eth z \bar{\eth} y = (r^2 - \eth^2 \alpha \bar{\eth}^2 \alpha)(m^y \overline{m}^z - \overline{m}^y m^z), \tag{53}$$

$$\eth z \bar{\eth} x - \eth x \bar{\eth} z = (r^2 - \eth^2 \alpha \bar{\eth}^2 \alpha)(m^z \overline{m}^x - \overline{m}^z m^x). \tag{54}$$

Thus, all three determinants vanish at points where

$$D \equiv r^2 - \eth^2 \alpha \bar{\eth}^2 \alpha = 0, \tag{55}$$

which, with r from Eq. (42), determines a curve, the wavefront singularities. Note that the evolution of the wavefront singularities (obtained by varying t_0) yields the caustic surface.

Since α is a regular function on the sphere, so is $\eth^2 \alpha \bar{\eth}^2 \alpha$; therefore, Eq. (55) [with r given by Eq. (42)] admits solutions ζ only within a finite interval of time t. Thus the wavefronts are singular only during a closed interval of time. On the other hand, at very long times the wavefronts become spherical, which follows from the line element Eq. (40).

The wavefront singularities (curves) are places where neighboring null geodesics meet. We have a null surface $x^a(r,\zeta,\overline{\zeta})$ foliated by null geodesics. At every fixed value of r, there are two connecting vectors $\eth x^a|_{u,r}$ and $\bar{\eth} x^a|_{u,r}$. The null geodesics in this congruence meet wherever the area orthogonal to the congruence, spanned by the connecting vectors, vanishes. The connecting vectors are, explicitly,

$$\eth x^a|_{u,r} = -\bar{\eth}\eth^2 \alpha l^a + \eth^2 \alpha \overline{m}^a + r m^a, \tag{56}$$

$$\bar{\eth} x^a|_{u,r} = -\eth\bar{\eth}^2 \alpha l^a + r \overline{m}^a + \bar{\eth}^2 \alpha m^a. \tag{57}$$

The area spanned by the connecting vectors (calculated from their skew product) is simply $D \equiv r^2 - \eth^2 \alpha \bar{\eth}^2 \alpha$. The vanishing of this area takes place at exactly the points given by Eq. (55).

We close this section with two examples.

Example 1: $\alpha = Y_{20} = 3\cos^2\theta - 1$. Due to axial symmetry, the wavefronts and their singularities for this choice of α can be completely worked out analytically, which gives insights into more general cases. The wavefronts at a given time t are given by

$$x = \frac{1}{\sqrt{2}} \sin\theta \cos\phi(\sqrt{2}t - 2u - 2 - 6\cos^2\theta), \tag{58}$$

$$y = \frac{1}{\sqrt{2}} \sin\theta \sin\phi(\sqrt{2}t - 2u - 2 - 6\cos^2\theta), \tag{59}$$

$$z = \frac{1}{\sqrt{2}} \cos\theta(\sqrt{2}t - 2u + 10 - 6\cos^2\theta). \tag{60}$$

These are axially symmetric. For a closed interval of time, all the wavefronts are singular. For early and late times, however, the wavefronts are smooth.

The singular points are located by eliminating r from Eqs. (55) and (42), yielding in this case, the two solutions or ''sheets''

$$(\sqrt{2}t-2u+10-18\cos^2\theta)(\sqrt{2}t-2u-2-6\cos^2\theta)=0. \tag{61}$$

There is a solution θ only at times $\sqrt{2}(u-5)\leqslant t\leqslant\sqrt{2}(4+u)$. This is the interval where every wavefront is singular. A smooth wavefront and its corresponding profile at a late time are shown in Fig. 1. A wavefront at a time when both the cusp ridge singularities and the z-axis singularities are occurring, and its corresponding profile, are shown in Fig. 2. In Fig. 3, we have a later wavefront and profile with only the cusp ridge singularity. In Fig. 4, we display the evolution of the singularities forming the caustic two-surface.

The high symmetry of this case is responsible for the lack of resemblance of the singular points on the z axis with standard cusps. At these points, null geodesics labeled by different, but neighboring, values of ϕ meet. This is clear from the fact that $\partial x/\partial\phi=\partial x/\partial\phi=\partial z/\partial\phi=0$ at these points, therefore the vector that connects geodesics with different values of ϕ vanishes.

In order to make a comparison, Fig. 5 shows a wavefront in the evolution of an imploding ellipsoid of revolution which is very similar to that of example 1. In this case, an ellipsoid of revolution sends an incoming wavefront, which develops singularities during a certain interval of time. The standard cuspoidal ridges are clearly visible as rings at both ends of the figure. However, the crossover points in between are also singular, of the same type of singularity as that one developed in our example. Assuming a speed of light of 1, the formulas for the imploding wavefront in this case are

$$x=a\sin\theta\cos\phi\left(1-\frac{t}{\sqrt{a^2\sin^2\theta+(a^2/c)^2\cos^2\theta}}\right), \tag{62}$$

$$y=a\sin\theta\sin\phi\left(1-\frac{t}{\sqrt{a^2\sin^2\theta+(a^2/c)^2\cos^2\theta}}\right), \tag{63}$$

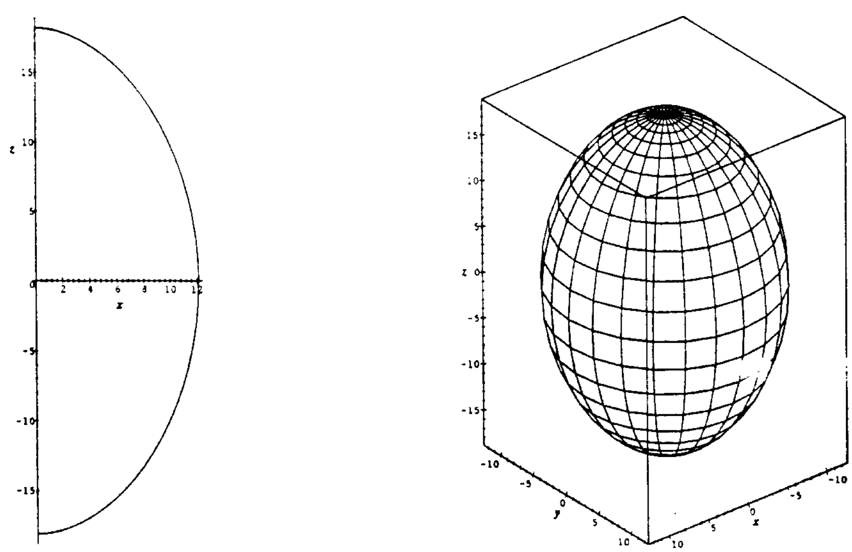

FIG. 1. The wavefront and profile ($t=10$, $u=0$) of the shearing Bondi congruence with $\alpha=Y_{20}(\theta,\phi)$.

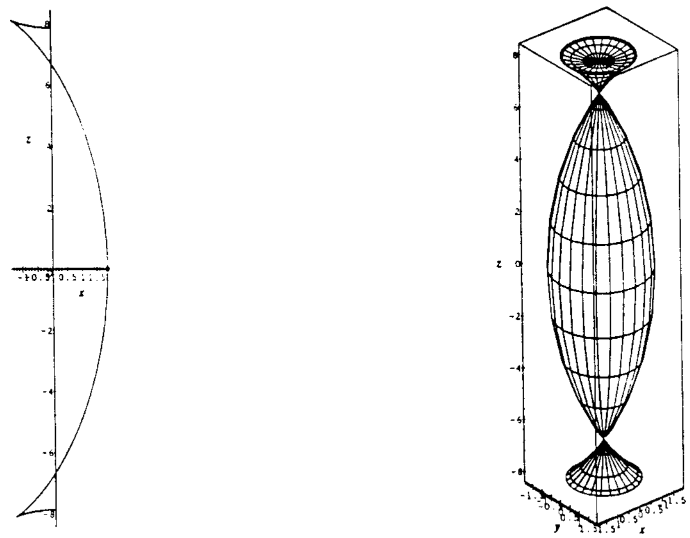

FIG. 2. The wavefront and profile ($t=2.73$, $u=0$) of the shearing Bondi congruence with $\alpha=Y_{20}(\theta,\phi)$. Notice the focusing at the crossover point.

$$z=c\,\cos\,\theta\left(1-\frac{t}{\sqrt{(c^2/a)^2\,\sin^2\,\theta+c^2\,\cos^2\,\theta}}\right). \tag{64}$$

Example 2: $\alpha=\mathrm{Real}\,(Y_{21})=(\zeta+\bar{\zeta})(\zeta\bar{\zeta}-1)/(1+\zeta\bar{\zeta})^2$. In this case, there is no advantage in writing the wavefronts explicitly. However, they can be plotted with ease, displaying the typical singularities of three-dimensional wavefronts, namely swallowtails and cusp ridges. Cusp ridges are clearly visible in Fig. 6, which represents a wavefront at $u=0$ and $t=1.5$. Swallowtails are exemplified in Fig. 7, which represents another wavefront in the same $u=0$ family, but at a later time of $t=2.35$. Locally all swallowtails have the form of Fig. 8.

Both Figs. 6 and 7 compare remarkably well with a typical wavefront in the evolution of a triaxial ellipsoid, in which case an ellipsoid $(x/a)^2+(y/b)^2+(z/c)^2=1$ emits a wavefront of light

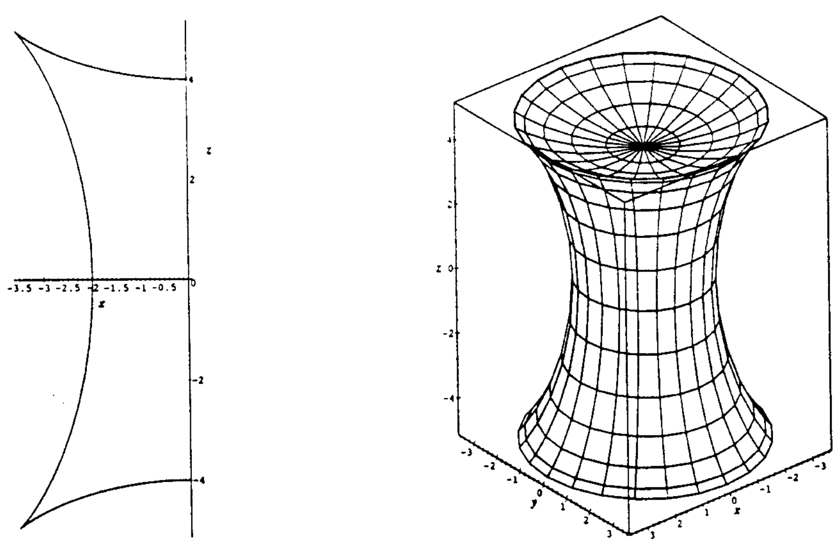

FIG. 3. The wavefront and profile ($t=0$, $u=0$) of the shearing Bondi congruence with $\alpha=Y_{20}(\theta,\phi)$. The only singularities are the standard cups, showing as rings.

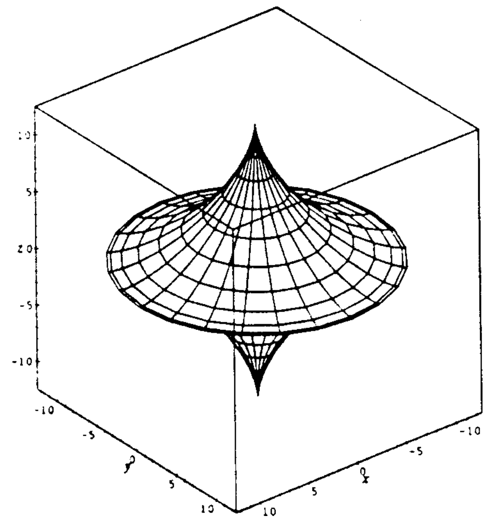

FIG. 4. The caustic surface ($u=0$) of the shearing Bondi congruence with $\alpha = Y_{20}(\theta,\phi)$. The caustic is the time evolution of the cusp ridges, projected in three-space. There is another sheet of the caustic surface, representing the time evolution of the crossover singularities, which is a line along the z axis, not visible in this picture.

inwardly, which develops singularities for a period of time. A typical singular imploding wavefront is shown in Fig. 9. The formulas for the imploding triaxial-ellipsoidal wavefront are the following:

$$x = a \sin\theta \cos\phi \left(1 - \frac{t}{a^2 \sqrt{(\sin\theta\cos\phi/a^2)^2 + (\sin\theta\sin\phi/b^2)^2 + (\cos\theta/c^2)^2}} \right),$$

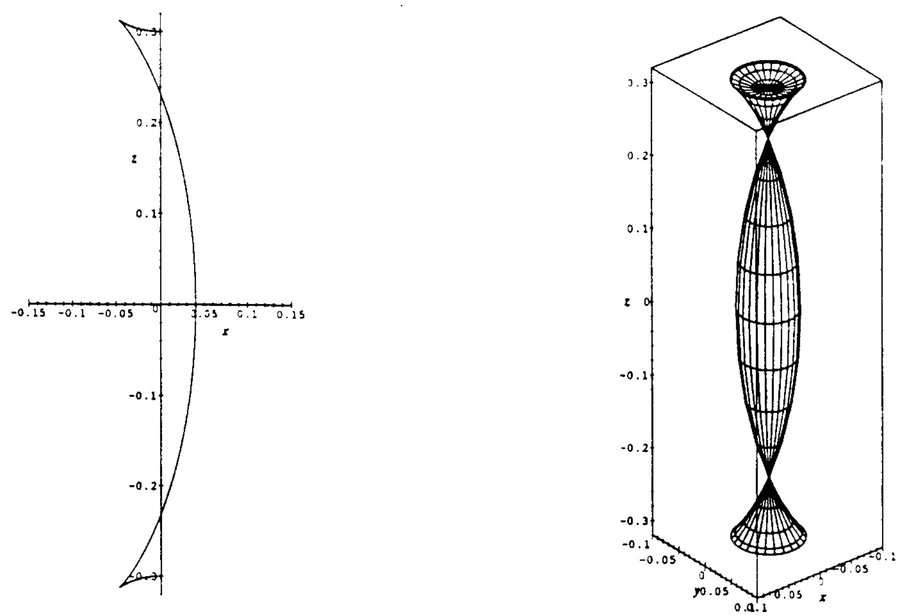

FIG. 5. A wavefront, and its profile, in the evolution of an imploding ellipsoid of revolution. The semiaxes of the initial ellipsoid $(x/a)^2+(y/b)^2+(z/c)^2=1$ have been chosen as $a=b=1$ and $c=1.2$, and the instant of time is $t=0.9$ assuming a speed of light of 1. Notice the cuspoidal ridges and the focusing at the crossover points.

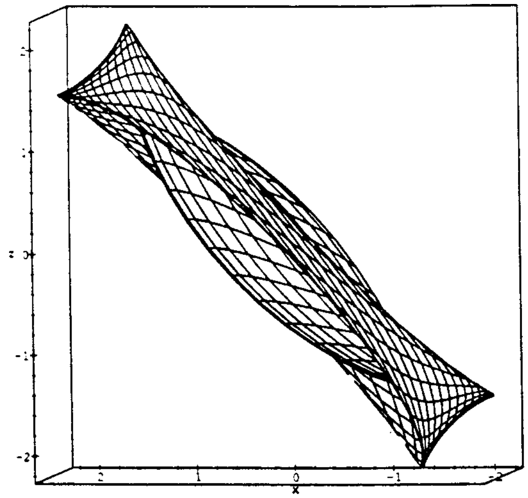

FIG. 6. A wavefront ($t=1.5$, $u=0$) of the shearing Bondi congruence with $\alpha=\Re(Y_{21}(\theta,\phi))$. Notice the cuspoidal ridges.

$$y=b\,\sin\,\theta\,\sin\,\phi\left(1-\frac{t}{b^2\sqrt{(\sin\,\theta\,\cos\,\phi/a^2)^2+(\sin\,\theta\,\sin\,\phi/b^2)^2+(\cos\,\theta/c^2)^2}}\right),\qquad(65)$$

$$z=c\,\cos\,\theta\left(1-\frac{t}{c^2\sqrt{(\sin\,\theta\,\cos\,\phi/a^2)^2+(\sin\,\theta\,\sin\,\phi/b^2)^2+(\cos\,\theta/c^2)^2}}\right).$$

V. GENERATING FAMILIES

In this section we will study the subject of the caustics of the null surfaces and the wavefront singularities via an alternative method, namely from the use of generating families for the construction of Lagrangian and Legendre submanifolds (developed by V. I. Arnold and his colleagues[1–4]) associated with cotangent and contact bundles over space–time. The value of this treatment is that it allows one to deal (via a parametric representation) with the regions of self-intersection and nondifferentiability of the null surfaces.

We first give a brief review of a special case of this theory that is adapted to the problem of null surfaces in four-dimensional space–time. Consider a four-dimensional Lorentzian manifold (with local coordinates, $\{t,x^i\}$) foliated by the constant t surfaces. Now consider the x^i as the coordinates of a configuration space M and p_i as the conjugate momentum so that we have the six-dimensional cotangent bundle T^*M, with local coordinates (x^i,p_i). We now describe the construction of a three-dimensional submanifold of T^*M (a Lagrangian submanifold—a maximal submanifold such that the symplectic form, restricted to it, vanishes) that plays a fundamental role in the discussion of the singularities of wavefronts and their associated caustics. We begin with a general description without any particular choice of dynamics, later restricting ourselves to null geodesic motion.

Choose a scalar function (determined later from the dynamics), referred to as a generating family, of the six variables

$$f=F(t,x^i,\zeta,\bar{\zeta});\qquad(66)$$

the x^i are spatial points, the $(\zeta,\bar{\zeta})$ are parametric labels (for convenience we are using a complex representation) for points on a given spatial two-surface, $\mathfrak{s}$, i.e., we have a two-point function, while t is the time for the (dynamic) particle to go from a point on $\mathfrak{s}$ to the point x^i. For a constant value of f, we consider Eq. (66) as defining t implicitly as a function of $(x^i,\zeta,\bar{\zeta})$, i.e.,

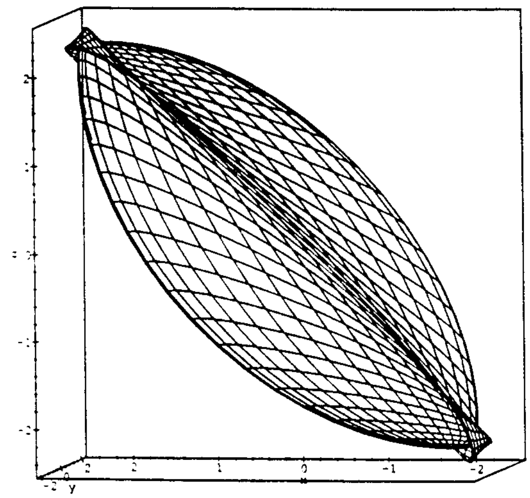

FIG. 7. A wavefront ($t=2.35$, $u=0$) of the shearing Bondi congruence with $\alpha=\Re(Y_{21}(\theta,\phi))$. Notice the swallowtails and the cuspoidal ridges.

$$t=T(f;x^i,\zeta,\overline{\zeta})$$

or simply

$$t=T(x^i,\zeta,\overline{\zeta}). \tag{67}$$

Note that T might be a multivalued function of its arguments, in which case it must be considered separately on the different sheets.

We now ask for the relationship between the $(x^i,\zeta,\overline{\zeta})$ when T is an extremal under variations of the $(\zeta,\overline{\zeta})$; i.e., we require that

$$\partial T/\partial\zeta=\partial T/\partial\overline{\zeta}=0, \tag{68}$$

which in turn forces

$$\partial F/\partial\zeta=\partial F/\partial\overline{\zeta}=0. \tag{69}$$

Finally, a rank condition is imposed on the choice of F; the following 2×5 matrix must have rank 2:

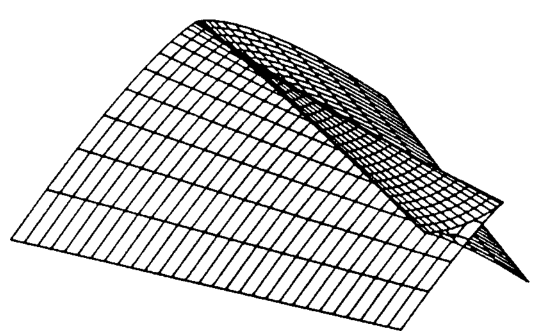

FIG. 8. The general form of a swallowtail singularity.

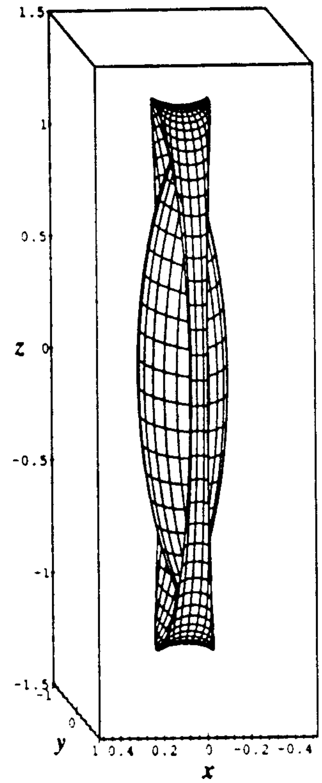

FIG. 9. An imploding triaxial-ellipsoidal wavefront, corresponding to $t=0.8$ and semiaxes $a=1$, $b=1.5$, and $c=2$ and assuming a speed of light of 1. Notice the cuspoidal ridges. Earlier instants, such as $t=0.65$, display swallowtails as well, not visible here.

$$\begin{bmatrix} \frac{\partial^2 F}{\partial \zeta^2} & \frac{\partial^2 F}{\partial \zeta \partial \overline{\zeta}} & \frac{\partial^2 F}{\partial \zeta \partial x^i} \\ \frac{\partial^2 F}{\partial \zeta \partial \overline{\zeta}} & \frac{\partial^2 F}{\partial \overline{\zeta}^2} & \frac{\partial^2 F}{\partial \overline{\zeta} \partial x^i} \end{bmatrix}. \tag{70}$$

The meaning of this condition is that the two equations [(69) or (68)] can be solved (locally) in at least one of a variety of possible ways for two of the five variables $(x^i,\zeta,\overline{\zeta})$; often it is necessary to solve them in different ways in different regions. We then have three different possible cases:

(1) $\zeta=\Upsilon(x^i)$, $\overline{\zeta}=\overline{\Upsilon}(x^i)$; the simplest of the three cases. It allows F to be treated as a function of just x^i. In the other cases F must be treated parametrically.
(2) $\zeta=\Psi(x^A,\overline{\zeta})$, $x^J=X^J(x^A,\overline{\zeta})$, where x^A are any two of the three x^i and x^J is the third one, or the conjugate version, $\overline{\zeta}=\overline{\Psi}(x^A,\zeta)$, $x^J=X^J(x^A,\zeta)$.
(3) $x^A=X^A(x^J,\zeta,\overline{\zeta})$, where again x^A are any two of the three x^i and x^J is the third one.

Case 1 can occur when the determinant

$$\hat{D}=\begin{vmatrix} \frac{\partial^2 F}{\partial \zeta^2} & \frac{\partial^2 F}{\partial \zeta \partial \overline{\zeta}} \\ \frac{\partial^2 F}{\partial \zeta \partial \overline{\zeta}} & \frac{\partial^2 F}{\partial \overline{\zeta}^2} \end{vmatrix} \neq 0; \tag{71}$$

$\hat{D}=0$ when there is a critical point of the Lagrangian map defined shortly.

The Lagrangian submanifold obtained from F is defined in the following way:

First we have $p_i=\partial F/\partial x^i$. [Note that this involves only the explicit x^i dependence in F since any implicit dependence, via the $(\zeta,\bar{\zeta})$, does not enter into the definition of p_i because of Eq. (69).] Now depending on which case, 1, 2, or 3, is relevant, we eliminate two of the five variables $(x^i,\zeta,\bar{\zeta})$, in the $p_i=(\partial F/\partial x^i)(x^i,\zeta,\bar{\zeta})$. This leaves the result that the six coordinates of T^*M can be expressed in terms of three parameters, thus defining a three-dimensional submanifold in each of the cases;

(i) In case 1, $p_i=P_i(x^i)$, $x^i=x^i$; the three parameters $\{x^i\}=\chi^\alpha$.
(ii) In case 2, $p_i=P_i(x^A,\bar{\zeta})$, $x^A=x^A$, $x^J=X^J(x^A,\bar{\zeta})$; the three parameters $\{x^A,\bar{\zeta}\}=\chi^\alpha$.
(iii) In case 3, $p_i=P_i(x^J,\zeta,\bar{\zeta})$, $x^A=X^A(x^J,\zeta,\bar{\zeta})$, $x^J=x^J$; the three parameters $\{x^J,\zeta,\bar{\zeta}\}=\chi^\alpha$.

To simplify the discussion, we have referred to the three parameters, in each of the cases, simply as χ^α. In each case we thus have

$$x^i=X^i(\chi^\alpha), \quad p_i=P_i(\chi^\alpha). \tag{72}$$

Of immediate relevance to us is the projection (Lagrange map) of the Lagrange submanifold into the configuration space which becomes in each of the cases, $x^i=X^i(\chi^\alpha)$ or

(i) in case 1, $x^i=x^i$, trivial diffeomorphism,
(ii) in case 2, $x^A=x^A$, $x^J=X^J(x^A,\bar{\zeta})$,
(iii) in case 3, $x^A=X^A(x^J,\zeta,\bar{\zeta})$, $x^J=x^J$.

The caustics of this problem are the regions in the configuration space where the mappings 1, 2, or 3 have rank 2 or 1; i.e, where the Jacobian of the mapping vanishes. They occur when the determinant $\hat{D}=0$. The inverse image to the caustics in the parameter space are referred to as the critical points of the Lagrange map. It is clear that in case 1, the Jacobian is one and rank reduction can only occur in cases 2 and 3.

This treatment of caustics can be extended into the full four-space by eliminating, in the expression for $t=T(f;x^i,\zeta,\bar{\zeta})$, or in the implicit version, $f=F(t,x^i,\zeta,\bar{\zeta})$, two of the five variables $(x^i,\zeta,\bar{\zeta})$ via the cases 1, 2, 3. This results in t now are a function of the three parameters, χ^α. Though we will not need the full theory here, this construction leads to a seven-dimensional manifold, (t,x^i,p_i) {an example of a contact manifold}, and a three-dimensional submanifold of the contact manifold (a Legendre submanifold) defined by

$$t=T(\chi^\alpha), \quad x^i=X^i(\chi^\alpha), \quad p_i=P_i(\chi^\alpha), \tag{73}$$

as well as the Legendre mapping from the Legendre submanifold, to space–time, (t,x^i),

$$\{t,x^i,p_i\}(\chi^\alpha)\Rightarrow t=T(\chi^\alpha), \quad x^i=X^i(\chi^\alpha), \tag{74}$$

a three-surface in space–time—the ''big wavefronts'' in Arnold's language—in our case a null surface. The singularities of the map (74), where the rank drops below 3, are the null surface singularities. These singularities, at fixed t, are the (''small'') wavefront singularities of the previous section.

We now return to the question of the determination of the function $f=F(t,x^i,\zeta,\bar{\zeta})$ of Eq. (66) for use in the study of null surfaces. Our choice will be, from Eq. (11),

$$F(t,x^i,\zeta,\bar{\zeta})=S^{**}(t,x^i,\zeta,\bar{\zeta})\equiv x^a l_a(\zeta,\bar{\zeta})+\alpha(\zeta,\bar{\zeta}). \tag{75}$$

There are three independent reasons for this choice;

(1) It was the method of generating an arbitrary null surface from the complete solution, $x^a l_a(\zeta,\bar{\zeta})$; see Sec. II.
(2) It was the method for the construction of a null surface such that the generators were orthogonal to a given two-surface; see Sec. III.
(3) It arises from a variant of Fermat's principle of stationary time: Consider a timelike worldline, $\mathfrak{L}$ (in a Lorentzian space–time), of the form, in local coordinates, ($x^i=$const, t varies) and a two-dimensional spacelike surface, $\mathfrak{s}(\zeta,\bar{\zeta})$. Assume, locally, that from every point of $\mathfrak{s}(\zeta,\bar{\zeta})$ there is a null geodesic that reaches $\mathfrak{L}$ at a time $t=T(x^i,\zeta,\bar{\zeta})$. Then t is extremized by those curves that are normal to $\mathfrak{s}(\zeta,\bar{\zeta})$. This result (which will be described in detail elsewhere) follows from Schrödinger's derivation[19] of the gravitational frequency shift.

Note that the rank condition on the matrix, Eq. (70), is satisfied by direct calculation.

From the discussion of generating families, we see that the treatment of the null surfaces that we proposed, in Secs. II and III, namely to solve for the $\zeta=\Upsilon(x,y,z)$, was really only valid for case 1, but we actually used a version of case 3 where the additional parameter r was introduced in order not to single out any particular Cartesian coordinate. Case 1 broke down precisely on the caustic given by

$$r^2=\eth^2\alpha\bar{\eth}^2\alpha, \tag{76}$$

which is where cases 2 and 3 must be applied.

What follows is a straightforward application of case 3 of these ideas—on one patch. For completeness we repeat some of the earlier steps.

Starting with

$$u=S^{**}(x^a,\zeta,\bar{\zeta})=x^a l_a(\zeta,\bar{\zeta})+\alpha(\zeta,\bar{\zeta}), \tag{77}$$

then

$$\begin{aligned}
\eth S^{**}&=\frac{1}{\sqrt{2}(1+\zeta\bar{\zeta})}\,[\bar{\zeta}^2W-\bar{W}-2z\bar{\zeta}]+\eth\alpha,\\
\bar{\eth}S^{**}&=\frac{1}{\sqrt{2}(1+\zeta\bar{\zeta})}\,[\zeta^2\bar{W}-W-2z\zeta]+\bar{\eth}\alpha,
\end{aligned} \tag{78}$$

with $W=(x+iy)$. From $\eth S^{**}=0$ and $\bar{\eth}S^{**}=0$, we obtain [from case 3, where x^A are (x,y) or $(W,\bar{W})$] that

$$\begin{aligned}
W&=\frac{\sqrt{2}(\bar{\eth}\alpha+\zeta^2\eth\alpha)-2z\zeta}{1-\zeta\bar{\zeta}},\\
\bar{W}&=\frac{\sqrt{2}(\eth\alpha+\bar{\zeta}^2\bar{\eth}\alpha)-2z\bar{\zeta}}{1-\zeta\bar{\zeta}}.
\end{aligned} \tag{79}$$

From Eq. (77) we have that

$$p_x=\frac{\partial S^{**}}{\partial x}=-\frac{\zeta+\bar{\zeta}}{\sqrt{2}(1+\zeta\bar{\zeta})},$$

$$p_y = \frac{\partial S^{**}}{\partial y} = \frac{i(\zeta - \bar{\zeta})}{\sqrt{2}(1+\zeta\bar{\zeta})}, \tag{80}$$

$$p_z = \frac{\partial S^{**}}{\partial z} = \frac{(1-\zeta\bar{\zeta})}{\sqrt{2}(1+\zeta\bar{\zeta})}.$$

Taking $p = \sqrt{2}(p_x + i p_y)$, we obtain

$$p = -\frac{2\zeta}{1+\zeta\bar{\zeta}}. \tag{81}$$

The Lagrange submanifold, parametrized by $(z, \zeta, \bar{\zeta})$, is given by

$$z = z, \tag{82}$$

$$W = \frac{\sqrt{2}(\bar{\eth}\alpha + \zeta^2 \eth\alpha) - 2z\zeta}{1-\zeta\bar{\zeta}}, \tag{83}$$

$$\bar{W} = \frac{\sqrt{2}(\eth\alpha + \bar{\zeta}^2 \bar{\eth}\alpha) - 2z\bar{\zeta}}{1-\zeta\bar{\zeta}}, \tag{84}$$

$$p = -\frac{2\zeta}{1+\zeta\bar{\zeta}}, \tag{85}$$

$$\bar{p} = -\frac{2\bar{\zeta}}{1+\zeta\bar{\zeta}}, \tag{86}$$

$$p_z = \frac{(1-\zeta\bar{\zeta})}{\sqrt{2}(1+\zeta\bar{\zeta})}. \tag{87}$$

The projection to the configuration space is given by

$$z = z,$$

$$W = \frac{\sqrt{2}(\bar{\eth}\alpha + \zeta^2 \eth\alpha) - 2z\zeta}{1-\zeta\bar{\zeta}}, \tag{88}$$

$$\bar{W} = \frac{\sqrt{2}(\eth\alpha + \bar{\zeta}^2 \bar{\eth}\alpha) - 2z\bar{\zeta}}{1-\zeta\bar{\zeta}}.$$

The critical points of the Lagrange map are obtained from the condition that the Jacobian

$$J = \frac{\partial(z, W, \bar{W})}{\partial(z, \zeta, \bar{\zeta})}. \tag{89}$$

vanishes. The vanishing of J is equivalent to $D=r^2-\eth^2\alpha\bar{\eth}^2\alpha=0$.

To construct the Legendrian submanifold [in the seven-dimensional contact space, (t,x^i,p_i)] we take the generating family $u=S^{**}(x^a,\zeta,\bar{\zeta})$ where u is constant and solve for t expressing the contact coordinate t in terms of the three parameters $(z,\zeta,\bar{\zeta})$ by

$$t=\frac{\zeta\bar{W}+\bar{\zeta}W-z(1-\zeta\bar{\zeta})}{(1+\zeta\bar{\zeta})}+\sqrt{2}[u-\alpha(\zeta,\bar{\zeta})] \tag{90}$$

with

$$W=\frac{\sqrt{2}(\bar{\eth}\alpha+\zeta^2\eth\alpha)-2z\zeta}{1-\zeta\bar{\zeta}},\qquad \bar{W}=\frac{\sqrt{2}(\eth\alpha+\bar{\zeta}^2\bar{\eth}\alpha)-2z\bar{\zeta}}{1-\zeta\bar{\zeta}}. \tag{91}$$

The full Legendre submanifold is then given by Eqs. (82)–(87) and (90).

Note that this entire construction, using Case 3, was valid where $1-\zeta\bar{\zeta}\neq 0$ (or equivalently where $p_z\neq 0$). To include the region where $p_z=0$, a different choice of parametrization would be necessary, e.g., $(x,\zeta,\bar{\zeta})$, which is valid in the region where $p_x\neq 0$ or $(y,\zeta,\bar{\zeta})$, valid where $p_y\neq 0$.

Using the example (2), from Sec. IV, given by

$$S^{**}(x^a,\zeta,\bar{\zeta})=x^a l_a(\zeta,\bar{\zeta})+\alpha(\zeta,\bar{\zeta}),\quad \alpha=-\frac{(1-\zeta\bar{\zeta})(\zeta+\bar{\zeta})}{(1+\zeta\bar{\zeta})^2} \tag{92}$$

with

$$\eth S^{**}=0\Leftrightarrow\{x(-1+\bar{\zeta}^2)+iy(1+\bar{\zeta}^2)-2z\bar{\zeta}](1+\zeta\bar{\zeta})=\sqrt{2}[1+\bar{\zeta}^3\zeta-3\bar{\zeta}(\zeta+\bar{\zeta})] \tag{93}$$

and

$$\bar{\eth}S^{**}=0\Leftrightarrow\{x(-1+\zeta^2)-iy(1+\zeta^2)-2z\zeta](1+\zeta\bar{\zeta})=\sqrt{2}[1+\zeta^3\bar{\zeta}-3\zeta(\zeta+\bar{\zeta})], \tag{94}$$

one could try to solve for different pairs from the set $(x,y,z,\zeta,\bar{\zeta})$. When $D\neq 0$ one could always solve for $(\zeta,\bar{\zeta})$, though, in general, there would be more than one solution; i.e., for fixed (x,y,z) there would in general be more than one ray going thru that space-point, either at the same or at different times. Alternately one could try to solve in different regions for (x,y), (y,z), (z,x), etc. Solving for (x,y) or W we have that the Lagrange map [from the $(z,\zeta,\bar{\zeta})$ parameter space] becomes

$$W=\frac{\sqrt{2}(2\zeta^2+4\zeta\bar{\zeta}-1-\zeta^2\bar{\zeta}^2)-2z\zeta(1+\zeta\bar{\zeta})}{1-\zeta^2\bar{\zeta}^2}, \tag{95}$$

$$\bar{W}=\frac{\sqrt{2}(2\bar{\zeta}^2+4\zeta\bar{\zeta}-1-\zeta^2\bar{\zeta}^2)-2z\bar{\zeta}(1+\zeta\bar{\zeta})}{1-\zeta^2\bar{\zeta}^2}, \tag{96}$$

which in turn becomes the Legendre map when the contact coordinate t is added in:

$$t=\sqrt{2}\left\{u+\frac{1}{(1-\zeta\bar{\zeta})(1+\zeta\bar{\zeta})^2}\left[4\zeta\bar{\zeta}(\zeta+\bar{\zeta})-\frac{z}{\sqrt{2}}(1+3\zeta\bar{\zeta}+3\zeta^2\bar{\zeta}^2+\zeta^3\bar{\zeta}^3)\right]\right\}. \tag{97}$$

Though none of the details of this analysis is particularly enlightening, it nevertheless shows how in principle one constructs the Lagrange submanifold and the Lagrange map even in the presence of the caustics.

VI. FAMILIES OF FOLIATIONS

In this section we will generalize, in the following sense, the ideas of Sec. II. Recently there has been a reformulation of general relativity, referred to as the null surface formulation (NSF) where the basic idea has been to use a family (a sphere's worth) of null foliations of space–time, so that there are a spheres worth of null surfaces passing through each point of space–time. These surfaces are described as the level surfaces of the function

$$u=Z(x^a,\zeta,\bar{\zeta});$$

x^a are the local space–time coordinates and $(\zeta,\bar{\zeta})$ are the complex stereographic coordinate on the sphere which labels the family of foliations. The function Z, for every fixed value of $(\zeta,\bar{\zeta})$, satisfies the eikonal equation,

$$g^{ab}\partial_a Z\partial_b Z=0. \tag{98}$$

Knowing these families of foliations one can construct the (conformal) metric in terms of Z. The idea was then to express the Einstein equations in terms of these surfaces, i.e., in terms of Z and a conformal factor. Though this was successfully accomplished, a technical difficulty in fully understanding the equations arose due to the fact that the null surfaces developed singularities (caustics) and self-intersections. It was clear that the development of caustics was a generic feature of the equations but it was not at all clear how to see and study their existence directly in terms of the function $Z(x^a,\zeta,\bar{\zeta})$ and its derivatives. In this section, we will study, in flat space, the construction of such families and show explicitly how to calculate the structure of the null surface singularities (the caustics and wavefront singularities) directly in terms of the Z function.

Locally (up to first derivatives) there is no direct curvature involvement in the eikonal equation, so that the form of the caustics in terms of Z should apply equally in curved space as in Minkowski space. The results obtained here for Minkowski space will thus likely apply to the curved space situation.

Starting with the two-parameter family of plane waves used earlier, $Z_0(x^a,\zeta,\bar{\zeta})=x^a l_a(\zeta,\bar{\zeta})$, we will first construct a general two-parameter family of solutions to the flat-space eikonal equation, $Z(x^a,\mu,\bar{\mu})$, with $(\mu,\bar{\mu})$ parametrizing the sphere; we then study the singularities and caustics of this new family.

We begin by generalizing Eq. (11), namely

$$S^{**}=x^a l_a(\zeta,\bar{\zeta})+\alpha(\zeta,\bar{\zeta}),$$

by writing α as a function on $S^2\times S^2$; i.e., as $\alpha=\alpha(\zeta,\bar{\zeta},\mu,\bar{\mu})$, and then repeating the earlier procedure of setting to zero, the ð and $\bar{ð}$ derivatives with respect to the $(\zeta,\bar{\zeta})$. Considering

$$u=Z^{**}(x,\zeta,\bar{\zeta},\mu,\bar{\mu})=x^a l_a(\zeta,\bar{\zeta})+\alpha(\zeta,\bar{\zeta},\mu,\bar{\mu}) \tag{99}$$

and

$$ð_\zeta Z^{**}=\bar{ð}_\zeta Z^{**}=0, \tag{100}$$

and then solving them (when possible) for $(\zeta,\bar{\zeta})$ obtaining

$$\zeta=\Upsilon(x,y,z,\mu,\bar{\mu}), \quad \bar{\zeta}=\bar{\Upsilon}(x,y,z,\mu,\bar{\mu}) \tag{101}$$

so that when substituted into Eq. (99), we obtain the new family which depends on the choice of $\alpha(\zeta,\bar{\zeta},\mu,\bar{\mu})$:

$$Z(x^a,\mu,\bar{\mu})=x^a l_a(\Upsilon,\bar{\Upsilon})+\alpha(\Upsilon,\bar{\Upsilon},\mu,\bar{\mu}). \tag{102}$$

[Alternatively we could use the different cases of Sec. V, when one can not solve for $(\zeta,\bar{\zeta})$.] It is obvious from the previous discussion that Eq. (102) satisfies the eikonal equation for each fixed value of $(\mu,\bar{\mu})$. All we have done so far is create a new sphere's worth of null foliations (wavefront families) of Minkowski space, different from the plane wave case of $S=x^a l_a(\zeta,\bar{\zeta})$. As in the earlier sections we could have analyzed the null surfaces for each value of $(\mu,\bar{\mu})$ separately but now in this generalization the null surfaces are smoothly connected to each other through the variable $(\mu,\bar{\mu})$ and it becomes of interest to see the development of the caustics via the variation of the $(\mu,\bar{\mu})$, or through the $(\mu,\bar{\mu})$ derivatives.

Remark 2: We will use, respectively, the notation $(\eth_\mu,\bar{\eth}_\mu)$ *for the eth and ethbar derivatives with respect to the variables* $(\mu,\bar{\mu})$ *and* $(\eth_\zeta,\bar{\eth}_\zeta)$ *for the variables* $(\zeta,\bar{\zeta})$.

We begin by defining several derivatives of Z:

$$\omega=\eth_\mu Z, \quad \bar{\omega}=\bar{\eth}_\mu Z \tag{103}$$

$$R=\bar{\eth}_\mu \eth_\mu Z, \tag{104}$$

$$\Lambda=\eth_\mu^2 Z, \quad \bar{\Lambda}=\bar{\eth}_\mu^2 Z. \tag{105}$$

A level surface of Z, with fixed $(\mu,\bar{\mu})$, is ruled by null geodesics, whose tangent vectors are given by $l_a(\Upsilon,\bar{\Upsilon})$. A pencil of rays defined from a pair of geodesic deviation vectors (from a given geodesic) has an area A that can be given[20] up to a proportionality by

$$A=K\frac{\Omega^{-2}}{\sqrt{(1-\Lambda_{,1}\bar{\Lambda}_{,1})}}, \tag{106}$$

where K is a constant determined by the initial area and

$$\Omega^2=l^a R_{,a}$$

and

$$\Lambda_{,1}=\Omega^{-2} l^a \Lambda_{,a}.$$

The derivation of Eq. (106) is lengthy and will not be given here. It will, however, be shown in this case to be proportional to the area.

We now want to see the behavior of ω, Λ, and R, as well as the area A, in the neighborhood of a caustic.

By direct calculation we have, from Eqs. (102) and (100), that

$$\omega=\eth_\mu \alpha, \quad \bar{\omega}=\bar{\eth}_\mu \alpha, \tag{107}$$

and hence is singularity free.

After a rather lengthy calculation, using Eqs. (99) and (100), we obtain

$$R=\eth_\mu\bar\eth_\mu\alpha+\frac{1}{D}\{(\eth_\zeta\eth_\mu\alpha)[(\eth_\zeta\bar\eth_\mu\alpha)(\bar\eth_\zeta^2u)-(\bar\eth_\zeta\bar\eth_\mu\alpha)(\eth_\zeta\bar\eth_\zeta u)]+(\bar\eth_\zeta\eth_\mu\alpha)[(\bar\eth_\zeta\bar\eth_\mu\alpha)(\eth_\zeta^2u) -(\eth_\zeta\bar\eth_\mu\alpha)(\eth_\zeta\bar\eth_\zeta u)]\}, \tag{108}$$

where

$$D=(\eth_\zeta\bar\eth_\zeta u)^2-\eth_\zeta^2u\bar\eth_\zeta^2u. \tag{109}$$

In a similar way we obtain

$$\Lambda=\eth_\mu^2\alpha+\frac{1}{D}\{(\eth_\zeta\eth_\mu\alpha)[(\eth_\zeta\eth_\mu\alpha)(\bar\eth_\zeta^2u)-(\bar\eth_\zeta\eth_\mu\alpha)(\eth_\zeta\bar\eth_\zeta u)] +(\bar\eth_\zeta\eth_\mu\alpha)[(\bar\eth_\zeta\eth_\mu\alpha)(\eth_\zeta^2u)-(\eth_\zeta\eth_\mu\alpha)(\eth_\zeta\bar\eth_\zeta u)]\}. \tag{110}$$

First we see that for fixed values of $(\mu,\overline{\mu})$, D from Eq. (109) is the same as in Eq. (32), namely $D=r^2-\eth_\zeta^2\alpha\bar\eth_\zeta^2\alpha$ and hence vanishes at the caustic. We can now see that ω is regular at the caustic while both R and Λ have singularities of the form D^{-1} at the caustic.

In order to find the area A we first need $R_{,a}$ and $\Lambda_{,a}$. After a lengthy calculation we obtain

$$\Omega^2\equiv l^aR_{,a}=\frac{1}{D^2}\{[(\eth_\zeta^2u)(\bar\eth_\zeta^2u)+(\eth_\zeta\bar\eth_\zeta u)^2][(\eth_\zeta\eth_\mu\alpha)(\bar\eth_\zeta\bar\eth_\mu\alpha)+(\bar\eth_\zeta\eth_\mu\alpha)(\eth_\zeta\bar\eth_\mu\alpha)] -2[(\eth_\zeta\eth_\mu\alpha)(\eth_\zeta\bar\eth_\mu\alpha)(\bar\eth_\zeta^2u)+(\bar\eth_\zeta\bar\eth_\mu\alpha)(\bar\eth_\zeta\eth_\mu\alpha)(\eth_\zeta^2u)](\eth_\zeta\bar\eth_\zeta u)\} \tag{111}$$

and

$$\Omega^2\Lambda_{,1}\equiv l^a\Lambda_{,a}=\frac{2}{D^2}\{[(\eth_\zeta^2u)(\bar\eth_\zeta^2u)+(\eth_\zeta\bar\eth_\zeta u)^2](\eth_\zeta\eth_\mu\alpha)(\bar\eth_\zeta\eth_\mu\alpha) -[(\eth_\zeta\eth_\mu\alpha)^2(\bar\eth_\zeta^2u)+(\bar\eth_\zeta\eth_\mu\alpha)^2(\eth_\zeta^2u)](\eth_\zeta\bar\eth_\zeta u)\}. \tag{112}$$

Though it is not immediately obvious, from Eqs. (111) and (112) one can show that $(1-\Lambda_{,1}\overline{\Lambda}_{,1})$ is proportional to D^2 or that $|\Lambda_{,1}|\Rightarrow 1+O(D)$ at the caustic. From these results we have that

$$A=K\frac{\Omega^{-2}}{\sqrt{(1-\Lambda_{,1}\overline{\Lambda}_{,1})}}=\frac{\pm KD}{\hat K} \tag{113}$$

with

$$\hat K=|\eth_\zeta\eth_\mu\alpha|^2-|\bar\eth_\zeta\eth_\mu\alpha|^2.$$

From Eqs. (100) and (101), we have that $\hat K=\hat K(\Upsilon,\overline{\Upsilon})$ from which it can be shown that

$$l^a\hat K_{,a}=0,$$

i.e., $\hat K$ is constant along the geodesic flow. If we chose $K=\pm\hat K$ we have that $A=D$, in agreement with the area obtained in Sec. III from geodesic deviation.

Several important observations can now be made:

(1) In our particular case of flat space, we have seen that the quantities R and Λ diverge as D^{-1} at the caustic. It appears virtually certain that this is a general result and remains true in a general curved space.
(2) As was to be expected, the area of a pencil of null geodesics vanishes at the caustic. This is clearly true in general and the result here is a confirmation that Eq. (106) really is the area formula.
(3) The quantity Ω (which plays a central role in the NSF version of GR) diverges as D^{-1} at the caustic.
(4) Though Λ diverges at the caustic, the absolute value of its weighted derivative

$$|\Lambda_{,1}|=|\Omega^{-2}\Lambda_{,a}l^a|$$

approaches one as $1-O(D)$. From this one sees that $\Lambda_{,a}l^a$ diverges as D^{-2}.

VII. DISCUSSION

Our main interest in the study of wavefronts and their associated null surfaces lies in our desire to understand and describe their singularity structure in curved Lorentzian manifolds and in particular to find the most appropriate variables and representations for their analysis. Though locally the classification of generic singularities and caustics is complete and is the same in both flat and curved spaces,[5] however, in general spaces, curvature effects are large and must eventually be taken into account for global questions. (For example, the structure of the light cone in a curved space is very different from that of a light cone in flat space.) The present work is intended to begin this study with the description of singular, global, asymptotically spherical, null surfaces in flat spaces. A follow-up second paper will be devoted to *the same* issues as here but in asymptotically flat space–times. We will see that beginning with a two-parameter family of solutions of the eikonal equation—analogous to the plane wave solutions of flat space–time—it will be possible to construct any other null surface and then analyze its singularity structure. In particular, it is possible to construct, in terms of the two-parameter family, the light cone of any space–time point. These insights are important for applications of the null surface formulation of GR.[8,9]

ACKNOWLEDGMENTS

We would like to thank Carlos Kozameh for many very useful arguments and even for some peaceful conversations. Also we want to point out that there has been a fair amount of unwitting and uncoordinated overlap between the work of the present authors (in Secs. II and III) and the group of Bishop, Gomez, Lehner, Szilagyi, and Winicour.[18] Since the motivation, notation, and final results of the work of the two groups are quite different, it has been decided to keep the work separate and publish independently. In addition, we mention that several of the issues discussed here were explored earlier in a beautiful paper by K. P. Tod.[21] We also thank the referee, Juergen Ehlers, for valuable improvements and corrections. SF and EN thank the NSF for support under Grant No. PHY 92-05109. GSO acknowledges the financial support from the Sistema Nacional de Investigadores and from the Consejo Nacional de Ciencia y Tecnología (CONACyT) México.

[1] V. I. Arnold, S. M. Gusein-Zade, and A. N. Varchenko, *Singularities of Differentiable Maps, Vol. I* (Birkhauser, Boston, 1985).
[2] V. I. Arnold, *Catastrophe Theory* (Springer-Verlag, Berlin, Heidelberg, 1986).
[3] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, Berlin, 1980).
[4] V. I. Arnold and S. P. Novikov, eds., *Dynamical Systems, Vol. IV* (Springer-Verlag, Berlin, 1990).
[5] H. Friedrich and J. Stewart, Proc. R. Soc. London, Ser. A **385**, 345 (1983).
[6] W. Hasse, M. Kriele, and V. Perlick, Class. Quantum Grav. **13**, 161 (1996).
[7] P. Schneider, J. Ehlers, and E. E. Falco, *Gravitational Lenses* (Springer-Verlag, New York, 1992).
[8] S. Frittelli, C. Kozameh, and E. Newman, J. Math. Phys. **36**, 4984 (1995).
[9] S. Frittelli, C. Kozameh, and E. Newman, J. Math. Phys. **36**, 6397 (1995).
[10] R. Penrose and W. Rindler, *Spinors and Space-Time* (Cambridge U.P., Cambridge, England, 1994).

[11] E. T. Newman and K. P. Tod, ''Asymptotically Flat Space-Times,'' in *General Relativity and Gravitation, Vol. 2*, edited by A. Held (Plenum, New York, 1980), and references cited therein.
[12] L. Landau and E. Lifschitz, *Classical Mechanics* (Pergamon, Headington Hill Hall, Oxford, 1960).
[13] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience New York, 1962), Vol. II.
[14] C. Kozameh and E. Newman, J. Math. Phys. **24**, 2481 (1983).
[15] S. L. Kent and E. T. Newman, J. Math. Phys. **24**, 949 (1983).
[16] S. Frittelli, C. Kozameh, and E. Newman, Phys. Rev. D **56**, 4729 (1997).
[17] R. W. Lind and E. T. Newman, J. Math. Phys. **15**, 1103 (1974).
[18] A recent preprint by Bishop, Gomez, Lehner, Szilagyi, and Winicour uses this same form of the metric for the study of numerical solutions of the Einstein equations.
[19] I. Schrodinger, *Expanding Universes* (Cambridge U.P., Cambridge, England, 1956).
[20] C. Kozameh and E. Newman, ''A Non-Local Approach to the Vacuum Maxwell, Yang-Mills and Einstein equations,'' in *Topological Properties and Global Structure of Space-Time*, edited by P. Bergmann and V. de Sabbata (Plenum, New York, 1986).
[21] K. P. Tod, in *Recent Advances in General Relativity*, edited by A. Janis and J. Porter (Birkhauser, Boston, Basel, 1992).

S_4 symmetry of $6j$ symbols and Frobenius–Schur indicators in rigid monoidal C^* categories

J. Fuchs[a)]
Max-Planck-Institut für Mathematik, Gottfried-Claren-Str. 26, D-53225 Bonn, Germany

A. Ch. Ganchev
Institute for Nuclear Research and Nuclear Energy, 72 Tsarigradsko Chaussee, BG-1784 Sofia, Bulgaria

K. Szlachányi and P. Vecsernyés
Central Research Institute for Physics, P.O. Box 49, H-1525 Budapest 114, Hungary

(Received 2 April 1998; accepted for publication 11 August 1998)

We show that a left-rigid monoidal C^* category with irreducible monoidal unit is also a sovereign and spherical category. Defining a Frobenius–Schur-type indicator we obtain selection rules for the fusion coefficients of irreducible objects. As a main result we prove S_4 invariance of $6j$ symbols in such a category. © *1999 American Institute of Physics.* [S0022-2488(99)00701-X]

I. INTRODUCTION

The quantities that are known as $6j$ *symbols* or *F coefficients* appear in various disguises in mathematics and physics, for instance, as recoupling coefficients in the theory of groups[1–3] and quantum groups,[4] as (partially gauge-fixed) fusing matrices[5–7] in conformal field theory, as the Boltzmann weights for triangulations of three-manifolds[8–11] giving rise to topological lattice field theories,[11–13] as expansion coefficients of exchange algebra relations in the algebraic theory of superselection sectors,[14] as the components of the three-cocycle in Ocneanu's non-Abelian cohomology,[15] as structure constants in the theory of three-algebras,[16] as specific endomorphisms of a von Neumann factor M in the theory of subfactors,[17] and as components of the coassociator φ of a rational Hopf algebra.[18–20] More generally, the F coefficients can be described as the projections of the associativity constraint φ of a semisimple, rigid, monoidal category onto irreducible objects, hence they map a pair of basic intertwiner spaces to another pair. The F coefficients depend on the choice of basis (gauge choices) in the basic intertwiner spaces (spaces of three-point functions).

An important property of the F coefficients is their ''tetrahedral symmetry'' or, more precisely, the fact that the gauge freedom that is present in their definition can be fixed in such a way that they are invariant under a set of transformations which form the permutation group S_4. This symmetry property is, for example, needed in three-dimensional lattice theories in order for the partition function that is defined in terms of the F coefficients to be independent of the triangulation so that the theory is topological (see, e.g., Refs. 21 and 9). Other places where this symmetry plays an important role are the derivation of identities for ''higher-order fusion coefficients''[22] and the construction of solutions to the ''big pentagon equation.''[23] Also, in pursuing a project for finding (possibly with the aid of computers) explicit solutions of the Moore–Seiberg polynomial equations for ''small'' fusion rings, the assumption of tetrahedral symmetry allows for the substantial increase in the number of accessible fusion rings. (In conformal field theory the explicit computation of fusing matrices is, e.g., required for the calculation of the operator product coefficients.)

Let us remark that in conformal field theory it seems to be common lore that the F coefficients

[a)]Electronic mail: jfuchs@x4u2.desy.de

possess S_4 invariance (see, e.g., Refs. 7 and 24); but, as a matter of fact, no complete and detailed proof has ever been published. In the theory of subfactors it was realized that the S_4 symmetry follows from Frobenius reciprocity of intertwiners between bimodules,[17] and in algebraic field theory the same aspects of the symmetry are implicit in the results of Ref. 14 (see, in particular, the appendix of that paper). The Frobenius maps between basic intertwiners generate the group S_3, and the coherent choice of bases in the orbits of this S_3 group lead to S_4-symmetric F coefficients. In Ref. 9 it was shown that in order to have S_4 symmetry one does not need a braiding—it is sufficient that the category be spherical; however, in that paper the possibility of nontrivial Frobenius–Schur (FS) indicators[25–27] was not considered.

In this paper we present a proof that in a left-rigid monoidal C^* category the F coefficients possess S_4 symmetry. We would like to emphasize two fine points. The first is that, in general, due to the possibility of having nontrivial FS indicators, the Frobenius maps provide only a $\mathbb{Z}_2$-projective representation of S_3; but nevertheless the corresponding signs cancel in the transformation of the F coefficients so that the F's are truly S_4 invariant. The second point is the following. The Frobenius transformations of order 3 map the basic intertwiners space $(p\times p,\hat{p})$ into itself; hence its eigenvalues are third roots of unity (here $\hat{p}$ is the conjugate of the irreducible object p). When considering F's that involve such intertwiner spaces, then in order to verify S_4 invariance we must calculate the different transforms of F in different bases of the space $(p\times p,\hat{p})$. If, instead, we use only a single basis in this space, then the transforms of F will possibly carry factors of third roots of unity (this is illustrated in an example in the Appendix).

Let us also briefly mention a possible application of our results to the quest of explicitly solving the polynomial equations for the braiding and fusing matrices, which constitutes, e.g., a part of the problem of classifying all rational conformal field theories. Namely, one expects that when the category is modular, then the trace of the braiding matrices $R^{pp,q}$ can be expressed completely in terms of the modular data, i.e., of the fusion rules, the modular S matrix and balancing phases (in the case of doubles of finite groups this was proved in Ref. 27). Assuming that this expectation is indeed correct, it follows that the Frobenius–Schur indicators as well as the multiplicities of the above-mentioned third roots of unity do not constitute independent data, but are already determined uniquely by the modular data. In particular, one can immediately write down all the F coefficients for which at least one line is colored by the unit object as well as the R coefficients. In fact, there are further special F coefficients that can easily be solved for. Finally, our results concerning the S_4 symmetry allow us to reduce the number of unknowns in the polynomial equations drastically, namely for generic F coefficients by a factor of 24. [Further simplifications come from the action of simple currents on the F-coefficients, which should be derivable in a way similar to the S_4-action. Moreover, one would expect the pentagon equation itself, which can be regarded as the boundary of a four-simplex, to possess S_5-symmetry, just like the tetrahedra (F-coefficients)—the boundaries of three-simplices—possess S_4-symmetry.]

II. RIGID MONOIDAL C^* CATEGORIES WITH IRREDUCIBLE MONOIDAL UNIT

Our starting point is a left-rigid monoidal C^* category $(\mathcal{C};(\varepsilon,\times,\{\lambda,\rho,\varphi\});(\hat{\ },\{e,c\});*)$ with the restriction that the monoidal unit ε is irreducible. Here $\times$ is the monoidal product and $\lambda_a,\rho_a,\varphi_{a,b,c}$ with $a,b,c\in\mathrm{Obj}\,\mathcal{C}$ are natural isomorphisms,

$$\begin{aligned}&\lambda_a:\ a\rightarrow a\times\varepsilon,\quad \rho_a:\ a\rightarrow\varepsilon\times a,\\&\varphi_{a,b,c}:\ a\times(b\times c)\rightarrow(a\times b)\times c,\end{aligned}\tag{2.1}$$

that satisfy the triangle identity

$$\varphi_{a,\varepsilon,c}=(\lambda_a\times 1_c)(1_a\times\rho_c^{-1}),\tag{2.2}$$

and the pentagon identity

$$\varphi_{a\times b,c,d}\varphi_{a,b,c\times d}=(\varphi_{a,b,c}\times 1_d)\varphi_{a,b\times c,d}(1_a\times\varphi_{b,c,d}).\tag{2.3}$$

The C^* property requires the arrows (the intertwiners)

$$(a,b)\equiv \mathrm{Hom}(a,b):=\{T:\ \ a\to b\} \tag{2.4}$$

between two objects *a,b* to form a complex Banach space. The $*$ is an involutive monoidal contravariant functor acting as identity on the objects and antilinearly on the intertwiner spaces. The norm of the intertwiners satisfies the C^* property, $\|T^*T\|=\|T\|^2$, and the natural equivalences $\{\lambda,\rho,\varphi\}$ are isometries. Irreducibility of ε implies that $(\varepsilon,\varepsilon)=\mathbb{C}\cdot 1_\varepsilon$.

The left conjugation L: $\mathcal{C}\to\mathcal{C}$ is an antimonoidal contravariant (linear) functor, which is the identity on the monoidal unit. The functor L is built up from a left rigidity structure $(\hat{\ },\{e,c\})$, where $\hat{\ }$ is the left conjugation on objects: $\hat{a}\equiv L(a)$. The left evaluation and coevaluation maps $e_a:\ \hat{a}\times a\to\varepsilon$ and $c_a:\ \varepsilon\to a\times\hat{a}$, $a\in\mathrm{Obj}\ \mathcal{C}$, satisfy the left rigidity equations

$$\lambda_a^*(1_a\times e_a)\varphi^*_{a,\hat{a},a}(c_a\times 1_a)\rho_a=1_a,\quad \rho^*_{\hat{a}}(e_a\times 1_{\hat{a}})\varphi_{\hat{a},a,\hat{a}}(1_{\hat{a}}\times c_a)\lambda_{\hat{a}}=1_{\hat{a}}, \tag{2.5}$$

and lead to the definition

$$(a,b)\ni T\mapsto L(T):=\rho^*_{\hat{a}}(e_b\times 1_{\hat{a}})((1_{\hat{b}}\times T)\times 1_{\hat{a}})\varphi_{\hat{b},a,\hat{a}}(1_{\hat{b}}\times c_a)\lambda_{\hat{b}}\in(\hat{b},\hat{a}) \tag{2.6}$$

of the conjugated intertwiners $L(T)$. Note that two left conjugations L_1 and L_2 that arise from left rigidity structures are always related by canonical natural equivalences μ: $L_1\to L_2$ given by

$$\begin{aligned}\mu_a&=\rho^*_{a^2}(e^1_a\times 1_{a^2})\varphi_{a^1,a,a^2}(1_{a^1}\times c^2_a)\lambda_{a^1}\in(a^1,a^2),\\ \mu_a^{-1}&=\rho^*_{a^1}(e^2_a\times 1_{a^1})\varphi_{a^2,a,a^1}(1_{a^2}\times c^1_a)\lambda_{a^2}\in(a^2,a^1),\end{aligned}\quad \text{for every}\ \ a\in\mathrm{Obj}\ \mathcal{C}, \tag{2.7}$$

where $a^i\equiv L_i(a)$ for $i=1,2$.

The C^* property allows us to define a right rigidity structure within $\mathcal{C}$:

$$(R(a),e^R_a,c^R_a):=(L(a),c^{L*}_a,e^{L*}_a)\equiv(\hat{a},c^*_a,e^*_a)\quad \text{for}\ \ a\in\mathrm{Obj}\ \mathcal{C}, \tag{2.8}$$

leading to a right conjugate functor R: $\mathcal{C}\to\mathcal{C}$ similarly to (2.6):

$$\begin{aligned}(a,b)\ni T\mapsto R(T)&:=\lambda^*_{R(a)}(1_{R(a)}\times e^R_b)(1_{R(a)}\times(T\times 1_{R(b)}))\varphi^*_{R(a),a,R(b)}(c^R_a\times 1_{R(b)})\rho_{R(b)}\\ &\equiv\lambda^*_{\hat{a}}(1_{\hat{a}}\times c^*_b)(1_{\hat{a}}\times(T\times 1_{\hat{b}}))\varphi^*_{\hat{a},a,\hat{b}}(e^*_a\times 1_{\hat{b}})\rho_{\hat{b}}\in(\hat{b},\hat{a}).\end{aligned} \tag{2.9}$$

Since the right and left conjugated objects are identical, $R(a)=L(a)\equiv\hat{a}$ for all $a\in\mathrm{Obj}\ \mathcal{C}$, the canonical natural isomorphisms

$$\{\kappa_a:\ \ R(L(a))\to a\},\quad \{\tilde{\kappa}_a:\ \ L(R(a))\to a\} \tag{2.10}$$

are given by

$$\begin{aligned}\kappa_a&=\lambda^*_a(1_a\times e^R_{L(a)})\varphi^*_{a,L(a),R(L(a))}(c^L_a\times 1_{R(L(a))})\rho_{R(L(a))}\equiv\lambda^*_a(1_a\times c^*_{\hat{a}})\varphi^*_{a,\hat{a},\hat{\hat{a}}}(c_a\times 1_{\hat{\hat{a}}})\rho_{\hat{\hat{a}}},\\ \tilde{\kappa}_a&=\rho^*_a(e^L_{R(a)}\times 1_a)\varphi_{L(R(a)),R(a),a}(1_{L(R(a))}\times c^R_a)\lambda_{L(R(a))}\equiv\rho^*_a(e_{\hat{a}}\times 1_a)\varphi_{\hat{\hat{a}},\hat{a},a}(1_{\hat{\hat{a}}}\times e^*_a)\lambda_{\hat{\hat{a}}},\end{aligned} \tag{2.11}$$

which satisfy

$$\begin{aligned}\kappa_a^{-1}&=\lambda^*_{\hat{\hat{a}}}(1_{\hat{\hat{a}}}\times e_a)\varphi^*_{\hat{\hat{a}},\hat{a},a}(e^*_{\hat{a}}\times 1_a)\rho_a=\tilde{\kappa}^*_a,\\ \tilde{\kappa}_a^{-1}&=\rho^*_{\hat{\hat{a}}}(c^*_a\times 1_{\hat{\hat{a}}})\varphi_{a,\hat{a},\hat{\hat{a}}}(1_a\times c_{\hat{a}})\lambda_a=\kappa^*_a.\end{aligned} \tag{2.12}$$

One can define (positive) left and right inverses for $a\in\mathrm{Obj}\ \mathcal{C}$:[28]

$$\Phi_a^L:\ (a\times b,a\times c)\rightarrow(b,c),\quad \Phi_a^R:\ (b\times a,c\times a)\rightarrow(b,c),\quad \text{for } b,c\in\text{Obj }\mathcal{C}, \tag{2.13}$$

using evaluation and coevaluation maps. In the special case $b=c=\varepsilon$ these maps,

$$\begin{aligned}(a,a)\ni T\mapsto \Phi_a^L(T)&:=e_a(1_{\hat a}\times T)e_a^*\in(\varepsilon,\varepsilon)=\mathbb{C}\cdot 1_\varepsilon,\\ (a,a)\ni T\mapsto \Phi_a^R(T)&:=c_a^*(T\times 1_{\hat a})c_a\in(\varepsilon,\varepsilon)=\mathbb{C}\cdot 1_\varepsilon,\end{aligned} \tag{2.14}$$

are faithful positive linear functionals. Since they are bounded from below,[28]

$$\begin{aligned}(a,a)\ni T^*T&\leqslant\rho_a^*(c_a^*c_a\times 1_a)\rho_a\cdot\lambda_a^*(1_a\times\Phi_a^L(T^*T))\lambda_a,\\ (a,a)\ni T^*T&\leqslant\lambda_a^*(1_a\times e_ae_a^*)\lambda_a\cdot\rho_a^*(\Phi_a^R(T^*T)\times 1_a)\rho_a,\end{aligned}\quad \text{for } T\in(a,b), \tag{2.15}$$

it follows that End $a\equiv(a,a)$ is finite dimensional for every $a\in$ Obj $\mathcal{C}$; therefore $\mathcal{C}$ is semisimple. We denote the subset of irreducible objects by $\mathcal{I}\subset$ Obj $\mathcal{C}$.

Semisimplicity allows us to construct the so-called *standard* rigidity intertwiners,[28] leading to a left conjugation functor L_s among the equivalent ones [cf. (2.7)] that obeys $L_s=R_s$ by (2.8) and (2.9). This conjugation is achieved by the following procedure.

In the case of an irreducible object $p\in\mathcal{I}$, one uses the scalar freedom $e_p\mapsto z_pe_p$, $c_p\mapsto z_p^{-1}c_p$ with $z_p\in\mathbb{C}\setminus\{0\}$ to set

$$e_pe_p^*=d_p1_\varepsilon=c_p^*c_p\quad\text{for every } p\in\mathcal{I}, \tag{2.16}$$

where d_a is the *quantum* or *statistics dimension* of an arbitrary object $a\in\mathcal{C}$, defined to be

$$d_a:=\|e_a\|\|c_a\|. \tag{2.17}$$

Then for an arbitrary object a one chooses two orthogonal and complete sets of partial isometries

$$\begin{aligned}V_a^{p\alpha}&:\ p\rightarrow a,\\ W_a^{p\alpha}&:\ \hat p\rightarrow\hat a,\end{aligned}\quad \text{for } p\in\mathcal{I},\ \ \alpha=1,...,m_a^p, \tag{2.18}$$

satisfying

$$V_a^{p\alpha*}V_a^{p'\alpha'}=\delta_{pp'}\delta_{\alpha\alpha'}1_p,\quad W_a^{p\alpha*}W_a^{p'\alpha'}=\delta_{pp'}\delta_{\alpha\alpha'}1_{\hat p}, \tag{2.19}$$

and

$$\sum_{p,\alpha}V_a^{p\alpha}V_a^{p\alpha*}=1_a,\quad \sum_{p,\alpha}W_a^{p\alpha}W_a^{p\alpha*}=1_{\hat a}, \tag{2.20}$$

to define

$$e_a:=\sum_{p,\alpha}e_p(W_a^{p\alpha*}\times V_a^{p\alpha*})\in(\hat a\times a,\varepsilon),\quad c_a:=\sum_{p,\alpha}(V_a^{p\alpha}\times W_a^{p\alpha})c_p\in(\varepsilon,a\times\hat a). \tag{2.21}$$

Then $(\hat{\ },\{e_a,c_a\})$ becomes a left rigidity structure, i.e., it satisfies (2.5) due to naturality of $\{\lambda,\rho,\varphi\}$ and (2.19) and (2.20). The rigidity intertwiners (e_a,c_a) of an object a satisfying (2.16) and (2.21) are called *standard*.

Now the corresponding right rigidity structure [cf. (2.8) and (2.9)] $(\hat{\ },\{c_a^*,e_a^*\})$ leads to a right conjugation functor R that is identical to the left conjugation L. As a matter of fact, by an explicit calculation one obtains

$$(b,a) \ni T \mapsto R(T) = L(T) = \sum_{p,\alpha,\beta} W_b^{p\beta} t_p^{\alpha\beta} 1_{\hat{p}} W_a^{p\alpha *} \in (\hat{a},\hat{b}), \tag{2.22}$$

with

$$t_p^{\alpha\beta} 1_p := V_a^{p\alpha *} T V_b^{p\beta}, \quad t_p^{\alpha\beta} \in \mathbb{C}. \tag{2.23}$$

Therefore, in the case of standard conjugation, we use the notation $\hat{T} \equiv R(T) = L(T)$ for the conjugated intertwiners.

One can also prove[28] that if (e_a, c_a) and (e'_a, c'_a) are both standard, then

$$e'_a = e_a(1_{\hat{a}} \times U_a), \quad c'_a = (U_a^* \times 1_{\hat{a}}) c_a, \tag{2.24}$$

where $U_a \in (a,a)$ is unitary. Moreover,

$$\begin{aligned} e_{(a,b)} &:= e_b(1_{\hat{b}} \times \rho_b)(1_{\hat{b}} \times (e_a \times 1_b))(1_{\hat{b}} \times \varphi_{\hat{a},a,b}) \varphi^*_{\hat{b},\hat{a},a\times b} \in ((\hat{b} \times \hat{a}) \times (a \times b), \varepsilon), \\ c_{(a,b)} &:= \varphi^*_{a\times b,\hat{b},\hat{a}}(\varphi_{a,b,\hat{b}} \times 1_{\hat{a}})((1_a \times c_b) \times 1_{\hat{a}})(\lambda_a \times 1_{\hat{a}}) c_a \in (\varepsilon, (a \times b) \times (\hat{b} \times \hat{a})) \end{aligned} \tag{2.25}$$

are standard if (e_a, c_a) and (e_b, c_b) are standard. Thus in the case of standard conjugation, $L = R$, the natural equivalence $\{\alpha_{a,b}: \widehat{a \times b} \to \hat{b} \times \hat{a}\}$ that expresses the antimonoidality of the conjugation functor is given by

$$\begin{aligned} \alpha_{a,b} &= \rho^*_{\hat{b}\times\hat{a}}(e_{a\times b} \times 1_{\hat{b}\times\hat{a}}) \varphi_{\widehat{a\times b},a\times b,\hat{b}\times\hat{a}}(1_{\widehat{a\times b}} \times c_{(a,b)}) \lambda_{a\times b} \in (\widehat{a\times b}, \hat{b}\times\hat{a}), \\ \alpha^{-1}_{a,b} &= \rho^*_{\widehat{a\times b}}(e_{(a,b)} \times 1_{\widehat{a\times b}}) \varphi_{\hat{b}\times\hat{a},a\times b,\widehat{a\times b}}(1_{\hat{b}\times\hat{a}} \times c_{a\times b}) \lambda_{\hat{b}\times\hat{a}} \in (\hat{b}\times\hat{a}, \widehat{a\times b}). \end{aligned} \tag{2.26}$$

III. SOVEREIGNTY, TRACES, AND SPHERICITY

We note that in the case of standard rigidity intertwiners the choice $\psi_a := 1_{\hat{a}}$ for every $a \in \mathrm{Obj}\,\mathcal{C}$ leads to a natural equivalence ψ: $R \to L$ and makes $\mathcal{C}$ a *sovereign* category.[29] Indeed, $\psi_\varepsilon = 1_\varepsilon$ and monoidality of ψ is clear, one has only to show the commutativity of the sovereignty diagram, which reads in this case as $\kappa_a = \tilde{\kappa}_a$. Since κ and $\tilde{\kappa}$ are natural equivalences it is enough to show this equality for $p \in \mathcal{I}$. But then one has

$$\kappa_p^* \kappa_p = x_p 1_{\hat{\hat{p}}} \quad \text{and} \quad \tilde{\kappa}_p^* \tilde{\kappa}_p = \tilde{x}_p 1_{\hat{\hat{p}}} \quad \text{for all} \quad p \in \mathcal{I}, \tag{3.1}$$

with $x_p, \tilde{x}_p \in \mathbb{R}_+$ and $x_p = \tilde{x}_p^{-1}$ due to (2.12). Using (2.16) and rigidity one obtains

$$\begin{aligned} x_p d_{\hat{p}} 1_\varepsilon &= e_{\hat{p}}(x_p 1_{\hat{\hat{p}}} \times 1_{\hat{p}}) e^*_{\hat{p}} = e_{\hat{p}}(\kappa_p^* \kappa_p \times 1_{\hat{p}}) e^*_{\hat{p}} = c_p^* c_p = d_p 1_\varepsilon, \\ \tilde{x}_p d_{\hat{p}} 1_\varepsilon &= c^*_{\hat{p}}(1_{\hat{p}} \times \tilde{x}_p 1_{\hat{\hat{p}}}) c_{\hat{p}} = c^*_{\hat{p}}(1_{\hat{p}} \times \tilde{\kappa}_p^* \tilde{\kappa}_p) c_{\hat{p}} = e_p e_p^* = d_p 1_\varepsilon, \end{aligned} \tag{3.2}$$

which imply $x_p = \tilde{x}_p$, hence $x_p = \tilde{x}_p = 1$ and $d_{\hat{p}} = d_p$. Then (3.1) and (2.12) lead to the equalities $\tilde{\kappa}_p = \kappa_p$, $p \in \mathcal{I}$, of isometries. Owing to naturality, the maps $\tilde{\kappa}_a \equiv \kappa_a : \hat{\hat{a}} \to a$, $a \in \mathrm{Obj}\,\mathcal{C}$, are isometries.

Having reached a standard conjugation, $L = R$, and sovereignty, let us consider the conjugates of standard rigidity intertwiners. Using (2.26) one obtains

$$\alpha_{\hat{\alpha},a} \hat{e}_a = c_{\hat{a}} \in (\varepsilon, \hat{a} \times \hat{\hat{a}}), \quad \hat{c}_a \alpha^{-1}_{a,\hat{a}} = e_{\hat{a}} \in (\hat{\hat{a}} \times \hat{a}, \varepsilon), \tag{3.3}$$

and the rigidity intertwiners of conjugated objects are related to the original ones as

$$e_{\hat{a}}=c_a^*(\kappa_a\times 1_{\hat{a}}),\quad c_{\hat{a}}=(1_{\hat{a}}\times\kappa_a^{-1})e_a^*,\tag{3.4}$$

as is seen by using (2.11) and (2.12). From (3.4) it follows that further specification of the rigidity intertwiners, e.g., identification of the (e_a,c_a) and $(c_{\hat{a}}^*,e_{\hat{a}}^*)$ rigidity pairs, may be achieved only in the case of involutive conjugation, which will be discussed in Sec. V.

In a sovereign monoidal category one can define left and right traces $\mathrm{tr}_a^{L/R}:(a,a)\to(\varepsilon,\varepsilon)$ for each $a\in\mathrm{Obj}\,\mathcal{C}$:

$$\begin{matrix}\mathrm{tr}_a^L(T):=e_a^L(\psi_a\times T)c_a^R,\\ \mathrm{tr}_a^R(T):=e_a^R(T\times\psi_a)c_a^L,\end{matrix}\quad\text{for}\quad\psi_a:\ R(a)\to L(a),\tag{3.5}$$

obeying the property

$$\mathrm{tr}_a^{R/L}(TS)=\mathrm{tr}_b^{R/L}(ST)\quad\text{for all}\quad T\in(b,a),\ S\in(a,b).\tag{3.6}$$

In the case of standard conjugation in a C^* category, together with the previously chosen sovereignty maps, $\{\psi_a=1_{\hat{a}}\}$, these left (right) traces become identical to left (right) inverses (2.14), hence they are positive traces. Moreover, due to standardness they are equal:

$$\mathrm{tr}_a^L(T)=\mathrm{tr}_a^R(T)=:\mathrm{tr}_a(T)\quad\text{for all}\quad T\in(a,a);\tag{3.7}$$

hence $\mathcal{C}$ is also a *spherical* category.[9]

IV. FROBENIUS–SCHUR INDICATORS

For any finite group G one defines the Frobenius–Schur element σ of the group algebra $\mathbb{C}G$ as in Ref. 25, i.e. (up to normalization),

$$\sigma:=\frac{1}{|G|}\sum_{g\in G}g^2.\tag{4.1}$$

The Frobenius–Schur element is a central self-adjoint element of $\mathbb{C}G$; its central decomposition reads

$$\sigma=\sum_{r\in\mathcal{I}}\frac{\nu_r}{d_r}e_r\quad\text{with}\quad\nu_r=\begin{cases}0 & \text{for } r\neq\hat{r},\\ \pm1 & \text{for } r=\hat{r},\end{cases}\tag{4.2}$$

where e_r, $r\in\mathcal{I}$, are the minimal central projectors and $d_r=\mathrm{tr}_r(e_r)$ is the (integral) dimension of the corresponding simple ideal of the group algebra $\mathbb{C}G$. The three-valent indicator ν_r that is defined by (4.2) is the *Frobenius–Schur indicator*, which is zero on non-self-conjugate simple ideals while on self-conjugate simple ideals the $+(-)$ sign indicates (pseudo)reality.

There is an extension of σ for C^*-Hopf[30] or weak C^*-Hopf algebras,[23] which is obtained using the Haar integral $h\in H$ of the (weak) C^*-Hopf algebra H:

$$\sigma:=\frac{1}{\epsilon(\mathbf{1})}h^{(1)}h^{(2)};\tag{4.3}$$

here we use Sweedler notation for the coproduct, i.e., $h^{(1)}\otimes h^{(2)}\equiv\Delta(h)$, and $\mathbf{1}$ and ϵ are the unit and the counit of H, respectively. Since weak C^*-Hopf algebras contain C^*-Hopf algebras as special cases, we prove the property (4.2) for σ as defined in (4.3) only for the former case.

There is a unique positive $g\in H$ (Ref. 31) that implements the square of the antipode, i.e., $gxg^{-1}=S^2(x)$ for all $x\in H$, and satisfies the normalization condition $\mathrm{tr}\,D_r(g)=\mathrm{tr}\,D_r(g^{-1})=:\tau_r\in\mathbb{R}_+$. Then $S(g)=g^{-1}$ holds, and if the unit representation is irreducible, then $\tau_r=d_r/\epsilon(\mathbf{1})$. Using the properties[32]

$$h^2=h^*=S(h)=h,\quad h^{(1)}\otimes xh^{(2)}y=S(x)h^{(1)}S^{-1}(y)\otimes h^{(2)}\quad \text{for all } x,y\in H \tag{4.4}$$

and

$$h^{(2)}\otimes h^{(1)}=h^{(1)}\otimes gh^{(2)}g,\quad S(h^{(1)})\otimes h^{(2)}=\sum_{r\in\mathcal{I}}\frac{1}{\tau_r}\sum_{i,j}e_r^{ij}g^{-1/2}\otimes g^{-1/2}e_r^{ji} \tag{4.5}$$

of the Haar integral $h\in H$, one obtains

$$\sigma^*=\frac{1}{\epsilon(\mathbf{1})}h^{(2)}h^{(1)}=\frac{1}{\epsilon(\mathbf{1})}h^{(1)}gh^{(2)}g=\frac{1}{\epsilon(\mathbf{1})}h^{(1)}S^{-1}(g)gh^{(2)}=\sigma \tag{4.6}$$

and

$$\begin{aligned}\sigma\cdot y&=\frac{1}{\epsilon(\mathbf{1})}h^{(1')}h^{(1)}h^{(2')}h^{(2)}\cdot y\\ &=\frac{1}{\epsilon(\mathbf{1})}S(h^{(1)})h^{(1')}h^{(2')}h^{(2)}\cdot y\\ &=\frac{1}{\epsilon(\mathbf{1})}S(h^{(1)}S^{-1}(y))h^{(1')}h^{(2')}h^{(2)}=y\cdot\sigma\end{aligned} \tag{4.7}$$

for all $y\in H$, i.e., σ is a central self-adjoint element of H. From the second identity in (4.5) it follows that σ vanishes on non-self-conjugate ideals, hence $\nu_r=0$ for $r\neq\hat{r}$. For a self-conjugate ideal, $e_r=e_{\hat{r}}$, one obtains

$$\epsilon(\mathbf{1})\cdot e_r\sigma=\frac{1}{\tau_r}\sum_{i,j}S^{-1}(e_r^{ij}g^{-1/2})g^{-1/2}e_r^{ji}=\frac{1}{\tau_r}\sum_{i,j}g^{1/2}S^{-1}(e_r^{ij})g^{-1/2}e_r^{ji}=\frac{1}{\tau_r}\sum_{i,j}S_0(e_r^{ij})e_r^{ji}, \tag{4.8}$$

where $S_0:=\mathrm{Ad}_{g^{1/2}}\circ S^{-1}$ is the involutive antipode: $S_0^2=id_H$, $S_0\circ *=*\circ S_0$. Since $e_r=e_{\hat{r}}$ for matrix units, it follows that $S_0(e_r^{ij})=v_re_r^{ji}v_r^*$ with $v_r\in e_rH$ unitary. Moreover,

$$e_r^{ij}=S_0^2(e_r^{ij})=S_0(v_r^*)v_re_r^{ij}v_r^*S_0(v_r) \tag{4.9}$$

implies that $S_0(v_r^*)v_r$ is a central unitary element in e_rH, hence $S_0(v_r^*)=\pm v_r^*=:\nu_rv_r^*$ for $r=\hat{r}$ due to the involutivity of S_0. But then

$$\begin{aligned}\sigma&=\sum_{\substack{r\in\mathcal{I}\\ r=\hat{r}}}\frac{1}{\epsilon(\mathbf{1})\tau_r}\sum_{i,j}S_0(e_r^{ij})e_r^{ji}\\ &=\sum_{\substack{r\in\mathcal{I}\\ r=\hat{r}}}\frac{1}{d_r}\sum_{i,j}v_re_r^{ji}v_r^*e_r^{ji}\\ &=\sum_{\substack{r\in\mathcal{I}\\ r=\hat{r}}}\frac{1}{d_r}v_rv_r^{*T}=\sum_{\substack{\tau\in\mathcal{I}\\ r=\hat{r}}}\frac{1}{d_r}v_rv_r^*S_0(v_r^*)v_r=\sum_{r\in\mathcal{I}}\frac{\nu_r}{d_r}e_r,\end{aligned} \tag{4.10}$$

proving (4.2).

There exists a purely categorical definition of the Frobenius–Schur indicator. Let $\mathcal{C}$ be a category as in Sec. II and p an irreducible object. If p is self-conjugate, then there exists an invertible intertwiner $J_p:p\rightarrow\hat{p}$. Let us define the self-intertwiner

$$\nu_p:=J_p^{-1}\hat{J}_p\kappa_p^{-1}\in(p,p), \tag{4.11}$$

which is independent of the choice of J_p due to irreducibility of p and linearity of the conjugation. Hence ν_p is an isometry, because an isometric J_p can be chosen. In order to prove that even $\nu_p = \pm 1_p$ holds, one first computes that

$$\Phi_p^R(\nu_p) := c_p^*(\nu_p \times 1_{\hat{p}}) c_p = c_p^*(J_p^{-1} \times J_p) e_p^*, \tag{4.12}$$

using (2.6) and (2.12). However, the rigidity intertwiners

$$(e_p', c_p') := (c_p^*(J_p^{-1} \times J_p), (J_p^{-1} \times J_p) e_p^*) \in ((\hat{p} \times p, \epsilon), (\epsilon, p \times \hat{p})) \tag{4.13}$$

are standard; hence, owing to (2.24) we have

$$e_p' = e_p u_p \quad \text{and} \quad c_p' = u_p^* c_p \quad \text{for some } u_p \in \mathbb{C} \text{ with } |u_p| = 1. \tag{4.14}$$

Therefore,

$$u_p d_p = u_p e_p e_p^* = e_p' e_p^* = \Phi_p^R(\nu_p) = c_p^* c_p' = u_p^* c_p^* c_p = u_p^* d_p, \tag{4.15}$$

which proves that $u_p = \pm 1$ and $\nu_p = \pm 1_p$.

Defining ν_p as the zero intertwiner for non-self-conjugate irreducible objects p, one can obtain a natural map ν between the identity functors.

Since the representation category of a (pure weak) C^*-Hopf algebra is a rigid monoidal C^* category with irreducible unit object, the consistency of the Hopf algebraic and the categorical definitions of the FS indicator requires that $\nu_p = d_p D_p(\sigma)$ for $p \in \mathcal{I}$, where $D_p : H \to \text{End } V_p$ is an irreducible representation of H. This is indeed the case: One can use the natural equivalences λ and ρ, the standard rigidity intertwiners that were given in Ref. 23, and the involutive conjugation on the objects $D \mapsto \hat{D} := D^T \circ S_0$ to arrive at $\kappa_D = 1_D$. Then for an irreducible self-conjugate object D_p the choice $J_p^{\hat{\alpha}\alpha} := D_p^{\hat{\alpha}\alpha}(S_0(v_p^*))$ for the matrix elements of the map $J_p : V_p \to \hat{V}_p$ leads to the desired result.

V. INVOLUTIVE CONJUGATION

The existence of the natural isomorphism $\{\kappa_a : \hat{\hat{a}} \to a\}$ suggests that a convenient choice $\hat{\hat{a}} = a$ can be achieved. This requires, however, a further assumption on the category $\mathcal{C}$; namely, the cardinalities of the sets of objects in any two equivalence classes that are related by conjugation must be equal. This quite harmless assumption will always hold after adjoining, if necessary, new objects to $\mathcal{C}$, and this procedure does not change the category within equivalence.

Therefore, from now on we assume that the object map $a \mapsto \hat{a}$ of our conjugation functor is involutive, $\hat{\hat{a}} = a$ for all $a \in \text{Obj } \mathcal{C}$, and that $\hat{a} = a$ whenever $\hat{a} \simeq a$. It follows that every choice of standard rigidity intertwiners leads to a conjugation functor $L \equiv R$ that is involutive on the arrows as well. As a matter of fact, for each $T \in (a,b)$ we have $T^{LL} = T^{LR} = \kappa_b^{-1} T \kappa_a = T$.

Due to the involutivity of the conjugation, for irreducible objects p the natural isomorphism $\kappa_p : \hat{\hat{p}} = p \to p$ becomes $\kappa_p = \chi_p 1_p$ with $\chi_p \sim \mathbb{C}$ and $|\chi_p| = 1$. The question arises whether the scalar χ_p that appears here has something to do with the Frobenius–Schur indicator or it can be ''gauged'' away. The relations (3.4) now read as

$$\begin{aligned} e_{\hat{p}} &= c_p^*(\kappa_p \times 1_{\hat{p}}) = \chi_p c_p^*, \\ c_{\hat{p}} &= (1_{\hat{p}} \times \kappa_p^{-1}) e_p^* = \chi_p^{-1} e_p^* \equiv \overline{\chi_p} e_p^*. \end{aligned} \tag{5.1}$$

Inserting $p = \hat{q}$ in the first equation, we obtain $\chi_{\hat{q}} e_{\hat{q}}^* = c_{\hat{q}} = \overline{\chi_q} e_q^*$, implying that $\chi_{\hat{q}} = \overline{\chi_q} \equiv \chi_q^{-1}$. For self-conjugate irreducibles $q = \hat{q}$ we obtain $\chi_q = \pm 1$. In this case by choosing the map $J_q : q \to \hat{q}$ in the definition (4.12) of the FS indicator to be 1_q, we obtain

$$\nu_q = J^{-1} \hat{J} \kappa_q^{-1} = 1_q 1_q (\chi_q^{-1} 1_q) = \chi_q^{-1} 1_q \quad \text{for all } q \in \mathcal{I} \text{ with } q = \hat{q}. \tag{5.2}$$

If $p \neq \hat{p}$ for $p \in \mathcal{I}$, then we can employ the freedom (2.24) so as to change the rigidity intertwiners of one of the irreducible objects of the pair $\{p,\hat{p}\}$, let us say of $\hat{p}$:

$$e'_{\hat{p}} := \chi_p^{-1} e_{\hat{p}}, \quad c'_{\hat{p}} := \chi_p c_{\hat{p}}, \quad e'_p := e_p, \quad c'_p := c_p. \tag{5.3}$$

Then $e'_{\hat{p}} = c'^*_p$ and $c'_{\hat{p}} = e'^*_p$, i.e., the coefficients χ that are present in (5.1) are gauged away for $p \in \mathcal{I}$, $p \neq \hat{p}$.

To summarize: after enlarging $\mathcal{C}$ appropriately, there exists a *standard conjugation functor* that is involutive both on the objects and on the arrows and has the property that for $p \in \mathcal{I}$

$$\begin{aligned} e_{\hat{p}} &= \chi_p c_p^*, \\ c_{\hat{p}} &= \chi_p^{-1} e_p^*, \end{aligned} \quad \text{with} \quad \chi_p = \begin{cases} 1 & \text{for } p \neq \hat{p}, \\ \nu_p & \text{for } p = \hat{p}. \end{cases} \tag{5.4}$$

VI. FROBENIUS MAPS ON BASIC INTERTWINER SPACES

Let us call the space $(p, q \times r)$, for $p,q,r \in \mathcal{I}$, a *basic intertwiner* space. It is a Hilbert space with scalar product determined by

$$(t_1, t_2) 1_p := t_1^* t_2 \quad \text{for} \quad t_1, t_2 \in (p, q \times r). \tag{6.1}$$

We define two antilinear maps

$$x: (p, q \times r) \to (r, \hat{q} \times p), \quad y: (p, q \times r) \to (q, p \times \hat{r}) \tag{6.2}$$

by

$$x(t) := (1_{\hat{q}} \times t^*) \varphi^*_{\hat{q},q,r} (e_q^* \times 1_r) \rho_r \sqrt{\frac{d_r}{d_p}}, \quad y(t) := (t^* \times 1_{\hat{r}}) \varphi^*_{q,r,\hat{r}} (1_q \times c_r) \lambda_r \sqrt{\frac{d_q}{d_p}} \tag{6.3}$$

and call them the Frobenius maps. These maps are antilinear isometries, i.e., for the scalar product the formula

$$\big(x(t_1), x(t_2)\big) = (t_2, t_1) = \big(y(t_1), y(t_2)\big) \tag{6.4}$$

holds due to the trace property (3.6). They generate a $\mathbb{Z}_2$-graded antilinear and projective action of S_3 on the basic intertwiner spaces. A $\mathbb{Z}_2$-graded antilinear action of a $\mathbb{Z}_2$-graded group means linear (antilinear) action for even (odd) elements of the group. A permutation group is naturally $\mathbb{Z}_2$ graded by distinguishing even and odd permutations. Moreover, the Frobenius maps generate only a projective action since using (5.4) one proves that

$$x^2 = \chi_q \cdot id_{(p,q \times r)}, \quad y^2 = \chi_r \cdot id_{(p,q \times r)}, \tag{6.5}$$

as well as

$$xyx = yxy: (p, q \times r) \to (\hat{p}, \hat{r} \times \hat{q}), \tag{6.6}$$

which constitutes just a projective version of the defining S_3 relations for the transpositions: $\sigma_{12} \leftrightarrow x$, $\sigma_{23} \leftrightarrow y$. In a generic case the Frobenius maps lead to an orbit of six different intertwiner spaces:

$$\begin{aligned} &\mathrm{Im}(id) = (p, q \times r), \quad \mathrm{Im}(xy) = (\hat{r}, \hat{p} \times q), \\ &\mathrm{Im}(x) = (r, \hat{q} \times p), \quad \mathrm{Im}(yx) = (\hat{q}, r \times \hat{p}), \\ &\mathrm{Im}(y) = (q, p \times \hat{r}), \quad \mathrm{Im}(yxy) = (\hat{p}, \hat{r} \times \hat{q}). \end{aligned} \tag{6.7}$$

It is easy to give the action of the Frobenius maps on canonical basic intertwiners $\rho_p \in (p,\varepsilon \times p)$ and $\lambda_p \in (p,p\times\varepsilon)$, namely,

$$x(\rho_p)=\rho_p\,,\quad y(\rho_p)=\frac{1}{\sqrt{d_p}}\,c_p\,,\quad xy(\rho_p)=\lambda_{\hat p}\,, \tag{6.8}$$

and

$$x(\lambda_p)=\frac{1}{\sqrt{d_p}}\,e_p^*\,,\quad y(\lambda_p)=\lambda_p\,,\quad yx(\lambda_p)=\rho_{\hat p}\,. \tag{6.9}$$

Therefore, they are not generic but length three orbits of S_3 with a possible projective $\mathbb{Z}_2$ extension by χ_p.

There are other cases, too, where the orbits are not generic. Then we have a representation of the corresponding nontrivial stabilizer subgroup of S_3 on a basic intertwiner space. In the list below we present all the possible irreducible representations of the stabilizer subgroups in the nongeneric cases. In two cases we obtain certain restrictions on fusion coefficients.

(1) $q=\hat q,\ r=p\neq q$: orbit$=\{(p,p\times q),(q,\hat p\times p),(\hat p,q\times\hat p)\}$. Therefore there is a $\mathbb{Z}_2$-graded antilinear (projective if $\chi_q=-1$) representation of $\mathbb{Z}_2$ on $(p,p\times q)$ generated by *y*. But an antilinear unitary is always of order 2, therefore this intertwiner space is the null space if $\chi_q=-1$, i.e., the corresponding fusion coefficient is $N_{pq}^{p}=0$. If $\chi_q=1$, we have a $\mathbb{Z}_2$-graded antilinear representation of $\mathbb{Z}_2$, and such irreducible representations are unique up to unitary equivalence.

(2) $q=r=\hat p\neq p$: orbit$=\{(p,\hat p\times\hat p),(\hat p,p\times p)\}$. Therefore there is a linear representation of $\mathbb{Z}_3$ on $(p,\hat p\times\hat p)$, generated by *xy*. One can use an orthonormal basis in $(p,\hat p\times\hat p)$ in which the basis elements carry one of the three possible (one-dimensional) irreducible representations of $\mathbb{Z}_3$. (A simple example is given in the Appendix.)

(3) $q=r=p=\hat p$: orbit$=\{(p,p\times p)\}$. Therefore there is a $\mathbb{Z}_2$-graded antilinear (projective if $\chi_p=-1$) representation of S_3 on $(p,p\times p)$. Similarly to the first case we have $N_{pp}^{p}=0$ if $\chi_p=-1$. If $\chi_p=1$, we get a $\mathbb{Z}_2$-graded antilinear proper representation of S_3. Such irreducible representations are one-dimensional and can be labelled by the three possible irreducible representations of $\mathbb{Z}_3$. One can use a basis in $(p,p\times p)$ where basis elements carry these irreducible representations.

VII. S_4-TRANSFORMED SIMPLICIAL MAPS OF A TETRAHEDRON

Let us consider the following Hilbert space containing certain fourfold tensor products of basic intertwiner spaces as orthogonal subspaces:

$$\begin{aligned}\mathcal{H}\equiv\bigoplus_{\mathbf{s}}\mathcal{H}_{\mathbf{s}}:=&\bigoplus_{p,q,r,t,u,v\in\mathcal{I}}(p\times q,u)\otimes(u\times r,t)\otimes(v,q\times r)\otimes(t,p\times v)\\ \equiv&\bigoplus_{p,q,r,t,u,v\in\mathcal{I}}(u,p\times q)^*\otimes(t,u\times r)^*\otimes(v,q\times r)\otimes(t,p\times v),\end{aligned} \tag{7.1}$$

where we put $(u,p\times q)^*:=\{t^*|t\in(u,p\times q)\}$. Vectors in subspaces $\mathcal{H}_{\mathbf{s}}$ of $\mathcal{H}$ that correspond to a fourfold tensor product of basic intertwiner spaces can be labeled by the two-dimensional simplicial complex of the boundary of a tetrahedron Δ^3. More precisely, there are simplicial maps

$$\nu=(\nu^0,\nu^1,\nu^2):\ \ \partial\Delta^3\rightarrow\mathcal{H}_{\mathbf{s}} \tag{7.2}$$

as follows. ν^0 maps all vertices into one point, thus the vertices of the tetrahedron remain unlabeled. (Nontrivial ν^0 is needed only in two-categories, which is out of the scope of the present paper.) ν^1 associates to each oriented edge of Δ^3 an irreducible object from the set of representants $\mathcal{I}$. Thus the edges become labeled with $\mathcal{I}$:

$$\begin{aligned} \nu^1(12)&:=p, \quad \nu^1(14):=t,\\ \nu^1(23)&:=q, \quad \nu^1(13):=u,\\ \nu^1(34)&:=r, \quad \nu^1(24):=v. \end{aligned} \tag{7.3}$$

Edges with opposite orientation carry the conjugate objects, e.g., $\nu^1(21)=\hat{p}$.

To every element of a basic intertwiner space one can associate an oriented face by a right-hand rule:

$(u,p\times q)\ni\alpha=$ 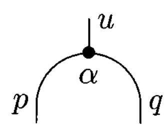

$\longleftrightarrow$ 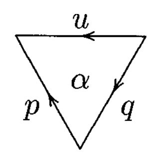

$=\nu^2(123)_+$

$(u,p\times q)^*\ni\alpha^*=$ 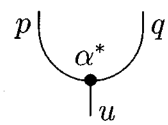

$\longleftrightarrow$ 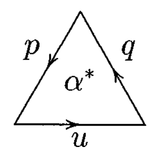

$=\nu^2(123)_-$

(In the vertex pictures, all arrows point downwards; correspondingly we can safely suppress them.)

The two possible orientations tell us whether we are in a basic intertwiner space $(u,p\times q)$ or in its adjoint $(u,p\times q)^*\equiv(p\times q,u)$. Now the images of the faces of the tetrahedron are defined to be elements of basic intertwiner spaces:

$$\begin{aligned} \nu^2(123)&:=\alpha^u_{pq}\in(u,p\times q)\equiv(\nu^1(13),\nu^1(12)\times\nu^1(23)),\\ \nu^2(134)&:=\beta^t_{ur}\in(t,u\times r)\equiv(\nu^1(14),\nu^1(13)\times\nu^1(34)),\\ \nu^2(234)&:=\gamma^v_{qr}\in(v,q\times r)\equiv(\nu^1(24),\nu^1(23)\times\nu^1(34)),\\ \nu^2(124)&:=\delta^t_{pv}\in(t,p\times v)\equiv(\nu^1(14),\nu^1(12)\times\nu^1(24)). \end{aligned} \tag{7.4}$$

Permutation of the index set $\{1,2,3,4\}$ of the points of the tetrahedron defines, using the Frobenius maps x,y, an S_4-action on the maps $\nu:\partial\Delta^3\rightarrow\mathcal{H}_s$ as follows:

$$(\nu_\tau)^0:=\nu^0\circ\tau, \quad (\nu_\tau)^1(ij):=\nu^1(\tau(i),\tau(j)) \quad \text{for all} \quad \tau\in S_4 \quad \text{and} \quad i,j\in\{1,2,3,4\}. \tag{7.5}$$

The definition of the transforms of ν^2 uses the Frobenius maps. Let τ_{12}, τ_{23}, and τ_{34} denote the generating transpositions of S_4. Then

$$\begin{aligned} (\nu_{\tau_{12}})^2(123)&:=x(\nu^2(123)), \quad (\nu_{\tau_{12}})^2(234):=\nu^2(134),\\ (\nu_{\tau_{12}})^2(134)&:=\nu^2(234), \quad (\nu_{\tau_{12}})^2(124):=x(\nu^2(124)); \end{aligned} \tag{7.6}$$

in other words, introducing the notation

$$\alpha^* \otimes \beta^* \otimes \gamma \otimes \delta = \nu^2(123)^* \otimes \nu^2(134)^* \otimes \nu^2(234) \otimes \nu^2(124) \in \mathcal{H}_s\,, \tag{7.7}$$

one has the induced transformations

$$\begin{aligned} \tau_{12}:&\;\; \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta \to x(\alpha)^* \otimes \gamma^* \otimes \beta \otimes x(\delta), \\ \tau_{23}:&\;\; \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta \to y(\alpha)^* \otimes \delta^* \otimes x(\gamma) \otimes \beta, \\ \tau_{34}:&\;\; \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta \to \delta^* \otimes y(\beta)^* \otimes y(\gamma) \otimes \alpha \end{aligned} \tag{7.8}$$

on the ν^2 images. The images of the original and the transformed maps ν, $\nu_{\tau_{12}}$, $\nu_{\tau_{23}}$, and $\nu_{\tau_{34}}$ can be pictured as

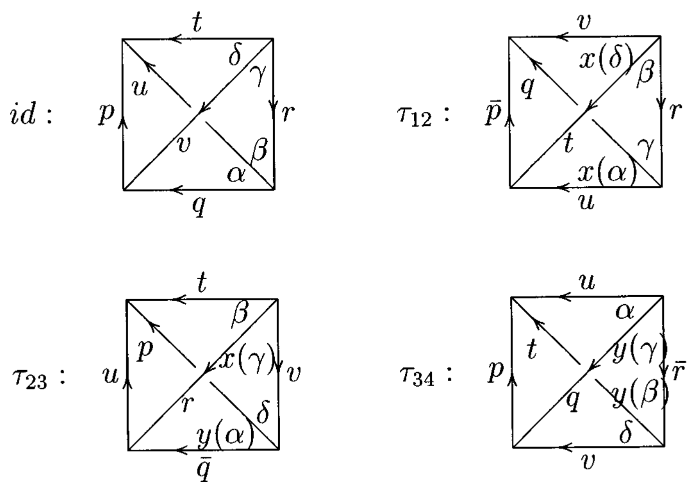

Extending the maps in (7.8) antilinearly, one obtains an action of S_4 on the Hilbert space $\mathcal{H}$ in (7.1). The important property of this S_4 action is the following:

Lemma: The above-defined action of the transpositions τ_{12}, τ_{23}, *and* τ_{34} *induces a* $\mathbb{Z}_2$*-graded antilinear representation of* S_4 *on* $\mathcal{H}$.

Proof: One has to check only the defining relations of S_4:

$$\tau_{12}^2 = \tau_{23}^2 = \tau_{34}^2 = id, \quad \tau_{12}\tau_{23}\tau_{12} = \tau_{23}\tau_{12}\tau_{23}, \quad \tau_{23}\tau_{34}\tau_{23} = \tau_{34}\tau_{23}\tau_{34}. \tag{7.9}$$

Indeed one obtains

$$\begin{aligned} \tau_{12}^2(\alpha^* \otimes \beta^* \otimes \gamma \otimes \delta) &= x^2(\alpha)^* \otimes \beta^* \otimes \gamma \otimes x^2(\delta) = \chi_p^2 \cdot \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta = \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta, \\ \tau_{23}^2(\alpha^* \otimes \beta^* \otimes \gamma \otimes \delta) &= y^2(\alpha)^* \otimes \beta^* \otimes x^2(\gamma) \otimes \delta = \chi_q^2 \cdot \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta = \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta, \\ \tau_{34}^2(\alpha^* \otimes \beta^* \otimes \gamma \otimes \delta) &= \alpha^* \otimes y^2(\beta)^* \otimes y^2(\gamma) \otimes \delta = \chi_r^2 \cdot \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta = \alpha^* \otimes \beta^* \otimes \gamma \otimes \delta, \end{aligned} \tag{7.10}$$

while

$$\begin{aligned}\tau_{12}\tau_{23}\tau_{12}(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta) &= xyx(\alpha)^*\otimes x(\beta)^*\otimes x(\delta)\otimes x(\gamma)\\ &= yxy(\alpha)^*\otimes x(\beta)^*\otimes x(\delta)\otimes x(\gamma)\\ &= \tau_{23}\tau_{12}\tau_{23}(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta),\\ \tau_{23}\tau_{34}\tau_{23}(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta) &= y(\beta)^*\otimes y(\alpha)^*\otimes xyx(\gamma)\otimes y(\delta)\\ &= y(\beta)^*\otimes y(\alpha)^*\otimes yxy(\gamma)\otimes y(\delta)\\ &= \tau_{34}\tau_{23}\tau_{34}(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta).\end{aligned} \tag{7.11}$$

■

Notice that although the Frobenius maps lead to a ($\mathbb{Z}_2$-graded antilinear) *projective* representation of S_3 on each basic intertwiner space, due to the fact that $\partial\Delta^3$ is a closed orientable surface without boundary, the action of S_4 on $\mathcal{H}$ is a *proper* (i.e., nonprojective) representation. As a matter of fact, according to (7.10), the Frobenius transformations x^2 and y^2 that lead to the signs χ are always coming in pairs.

VIII. AN S_4-INVARIANT LINEAR FUNCTIONAL ON $\mathcal{H}$

Let us define a linear functional Φ: $\mathcal{H}\to\mathbb{C}$ on the Hilbert space $\mathcal{H}$ (7.1) by

$$\Phi(\alpha_{pq}^{u*}\otimes\beta_{ur}^{t*}\otimes\gamma_{qr}^{v}\otimes\delta_{pv}^{t}) := \sqrt{\frac{d_p d_q d_r}{d_t}}\cdot e_t(1_{\hat t}\times\beta_{ur}^{t*})(1_{\hat t}\times(\alpha_{pq}^{u*}))$$
$$(1_{\hat t}\times\varphi_{p,q,r})(1_{\hat t}\times(1_p\times\gamma_{qr}^{v}))(1_{\hat t}\times\delta_{pv}^{t})e_t^*. \tag{8.1}$$

Pictorially, the value is

$$\Phi(\alpha_{pq}^{u*}\otimes\beta_{ur}^{t*}\otimes\gamma_{qr}^{v}\otimes\delta_{pv}^{t}) = \sqrt{\frac{d_p d_q d_r}{d_t}}\cdot \tag{8.2}$$

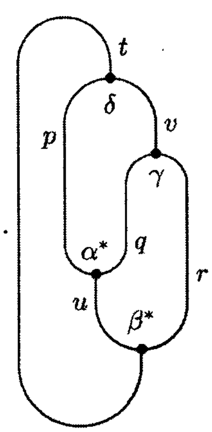

(In order not to overburden the picture, we do not indicate the maps λ, ρ, and φ; due to coherence and naturality they can be put back unambiguously.)

The action of $\tau\in S_4$ on Φ is the one induced by the action on $\mathcal{H}$,

$$(\tau\Phi)(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta) := \overline{\Phi(\tau^{-1}(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta))}^{\deg\tau}, \tag{8.3}$$

where the notation ''overline to the power of deg τ'' means complex conjugation for odd S_4 elements and the identity operation for even elements. This transformation property is required by the linearity of Φ and the antilinearity of S_4 on $\mathcal{H}$ to be compatible:

$$(\tau\Phi)(\lambda X)=\overline{\Phi(\tau^{-1}(\lambda X))}^{\deg\tau}=\overline{\overline{\lambda}^{\deg\tau}\Phi(\tau^{-1}(X))}^{\deg\tau}=\lambda\cdot\overline{\Phi(\tau^{-1}(X))}^{\deg\tau} \tag{8.4}$$

for $X\in\mathcal{H}$. Hence the functional Φ is S_4 invariant if

$$(\tau\Phi)(X):=\overline{\Phi(\tau^{-1}(X))}^{\deg\tau}=\Phi(X) \tag{8.5}$$

for all $\tau\in S_4$, or equivalently

$$\Phi(\tau^{-1}(X))=\overline{\Phi(X)}^{\deg\tau} \tag{8.6}$$

for all $\tau\in S_4$.

Proposition: The functional Φ is constant on S_4 orbits.

Proof: It is sufficient to show this property for the generators of S_4. Below there is a diagrammatic proof where we used the definitions of the Frobenius maps, the trace property, rigidity, the property End $\varepsilon=\mathbb{C}\cdot 1_\varepsilon$, and sphericity in the last series of pictures.

Invariance with respect to τ_{12} is shown by the following chain of equalities:

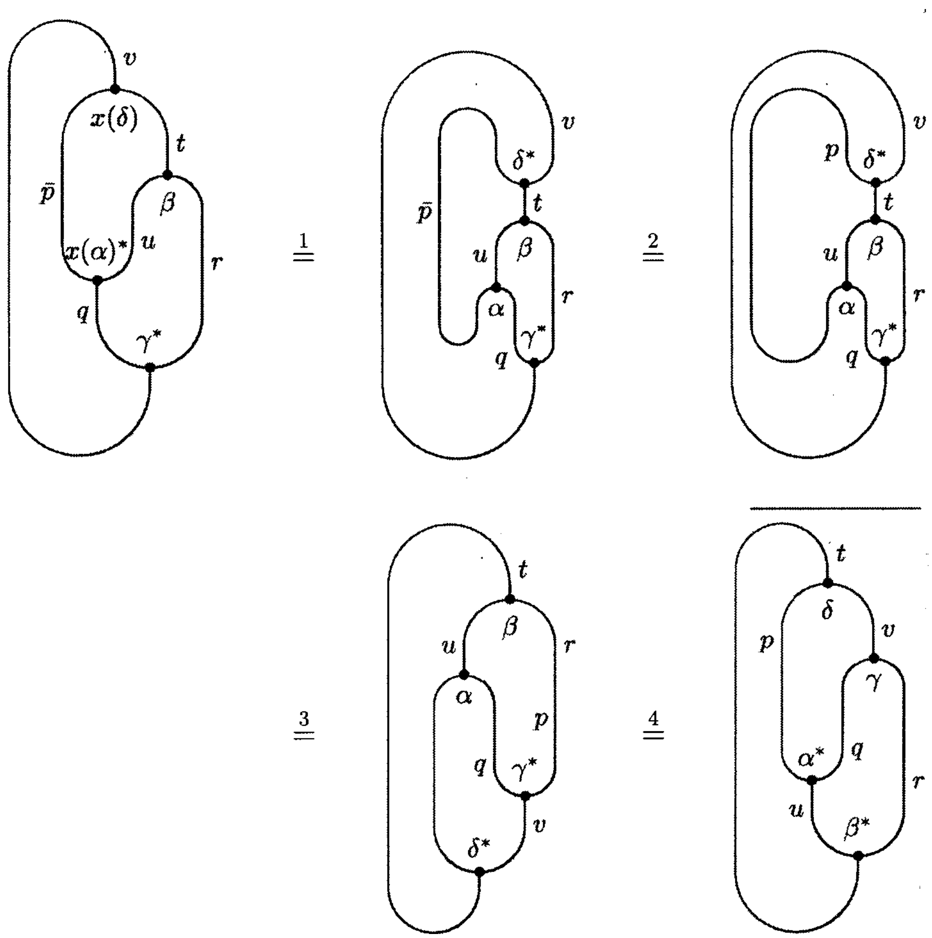

Here we have used the definition of the Frobenius map x in the first equality, the monoidality of the functor of taking duals in the second, used the trace property in the third, and the definition of ''*'' in the fourth equality.

Proof of the invariance with respect to τ_{23}:

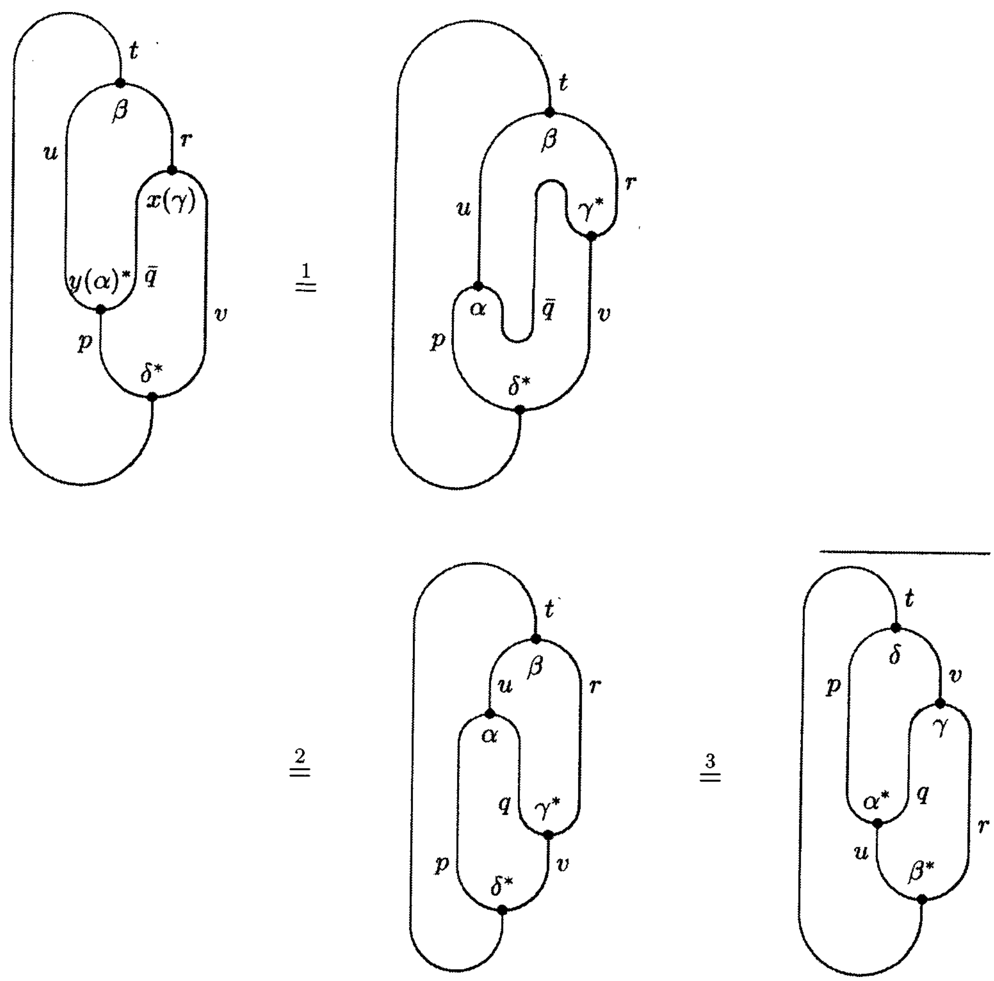

This chain of equalities is obtained as follows. In the first equality we used the definition of the Frobenius maps x and y; in the second equality the rigidity identity is used; and in the third equality the definition of ''*'' is implemented.

Proof of the invariance with respect to τ_{34}:

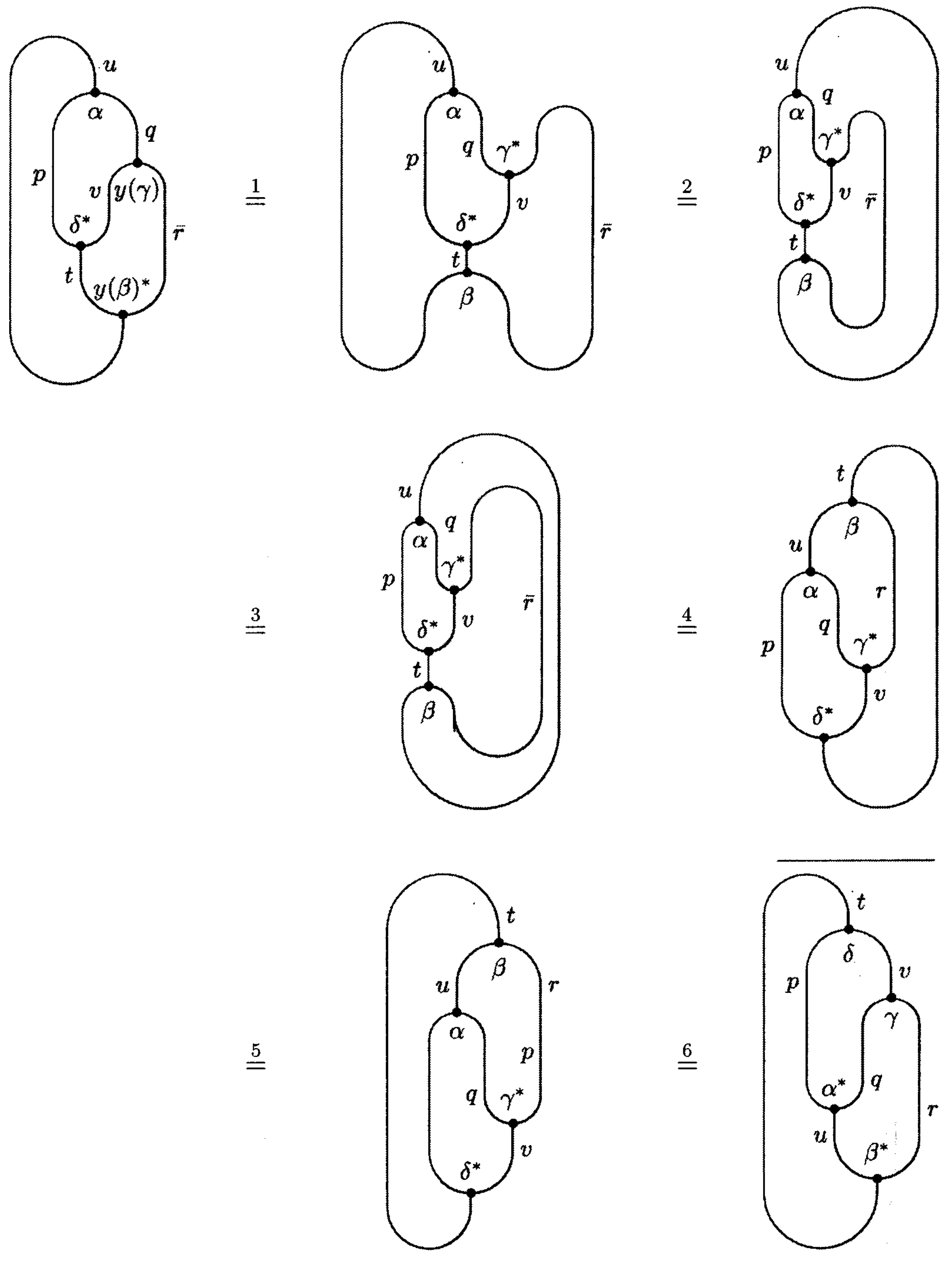

In this chain of equalities we have used the definition of the Frobenius map y in the first equality; the spherical property in the second; the monoidality of the functor of taking duals in the third; used the trace property in the fourth; used the spherical property in the fifth; and used the definition of "*" in the sixth. ■

By definition the normalized F-coefficients are

$$\check{F}^{(pqr)_t}_{u,v}(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta) := \Phi(\alpha^{u*}_{pq}\otimes\beta^{t*}_{ur}\otimes\gamma^{v}_{qr}\otimes\delta^{t}_{pv}), \tag{8.7}$$

where α, β, γ, and δ are elements of orthonormal bases of the corresponding basic intertwiner spaces. Their relation to the $6j$-symbols $\{F^{(pqr)_t}_{\alpha u\beta,\gamma v\delta}\}$ is given by

$$F^{(pqr)_t}_{\alpha u\beta,\gamma v\delta}=\frac{1}{\sqrt{d_pd_qd_rd_t}}\check{F}^{(pqr)_t}_{u,v}(\alpha^*\otimes\beta^*\otimes\gamma\otimes\delta). \tag{8.8}$$

Because of the isometry property of φ they are unitary matrices in the multi-labels $(\alpha u\beta,\gamma v\delta)$. The proposition leads to the possibility of computing $\check{F}$ in a single point of an S_4 orbit and determine the $6j$ symbols on all the other elements of that orbit.

ACKNOWLEDGMENTS

A. Ganchev acknowledges the support of the Alexander von Humboldt Foundation and the hospitality of FB Physik, University of Kaiserlautern, and of II. Institut für Theoretische Physik, University of Hamburg. The work of A. Ganchev, K. Szlachányi, and P. Vecsernyés was supported in part by a joint project of the Bulgarian and Hungarian Academies of Science and by the Hungarian Scientific Research Fund (OTKA T 020 895).

APPENDIX: S_4 SYMMETRY FOR C^* CATEGORIES BASED ON $\mathbb{Z}_3$

As a simple but nevertheless nontrivial illustration of the results of the main text, let us consider the following three (degenerate) rational Hopf algebras[18,19] that can be obtained as deformations of the Hopf algebra $\mathbb{C}\mathbb{Z}_3$, i.e., of the group algebra of the cyclic group $\mathbb{Z}_3$. The structural data can be summarized as follows:

$$H=\mathbb{C}e_0\oplus\mathbb{C}e_1\oplus\mathbb{C}e_2,\quad e_p^*=e_p^2=e_p\quad\text{for } i=0,1,2, \tag{A1}$$

$$\Delta(e_p)=\sum_{\substack{q,r=0\\ q+r=p \bmod 3}}^{2} e_q\otimes e_r,\quad S(e_p)=e_{-p \bmod 3}, \tag{A2}$$

$$\lambda=\rho=\mathbf{1}\equiv e_0+e_1+e_2=l=r\in H, \tag{A3}$$

$$\varphi=\sum_{p,q,r=0}^{2}\omega_{pqr}\cdot e_p\otimes e_q\otimes e_r\in H\otimes H\otimes H. \tag{A4}$$

Here $\omega_{111}=\omega_{222}=\omega_{112}=\omega_{221}=\omega_{211}=\omega_{122}=:\omega$ is a third root of unity, $\omega^3=1$, which parametrizes the three different rational Hopf algebras, while in all three cases one has $\omega_{pqr}=1$ for all other combinations of indices.

The representation category **Rep** H is a rigid monoidal C^*-category. The irreducible representations D_p, $p=0,1,2$, are one-dimensional and obey $D_p(e_q)=\delta_{p,q}$. The basic intertwiner spaces are one-dimensional at most, and we can choose $(p,q\times r)=\mathbb{C}\cdot 1^p_{qr}$ for the nontrivial ones, where 1^p_{qr} maps the tensor product of the chosen unit vectors into the chosen unit vector of the corresponding one-dimensional representation spaces, i.e., $1^p_{qr}(v_q\otimes v_r)=v_p$. The natural isometries connected to monoidality and the standard rigidity intertwiners are given by

$$\lambda_p=1^p_{p0},\quad \rho_p=1^p_{0p},\quad \varphi_{p,q,r}=(D_p\otimes D_q\otimes D_r)(\varphi), \tag{A5}$$

$$e_p=1^{0*}_{\hat{p}p},\quad c_p=1^0_{p\hat{p}}. \tag{A6}$$

The values of χ (5.4) are all trivial: $\chi_p=1$ for $p=0,1,2$. Using the definitions (6.3) of the Frobenius maps, one obtains

$$x(1^p_{qr})=\overline{\omega}_{\hat{q}qr}1^r_{\hat{q}p},\quad y(1^p_{qr})=\omega_{qr\hat{r}}1^q_{p\hat{r}}, \tag{A7}$$

with ω_{pqr} as in (A4). In the case $(p,q\times r)=(1,2\times 2)$, this leads to

$$xy(1^1_{22}) = \omega \cdot 1^1_{22}, \tag{A8}$$

which serve as examples of a degenerate orbit of type 2 in Sec. VI with the possible third roots of unity.

The S_4 symmetry is valid in a nontrivial way in the following sense. First note that there are five S_4 orbits of normalized F-coefficients $\check{F}^{(pqr)_t}_{u,v}(\alpha^{u*}_{pq}\otimes\beta^{t*}_{ur}\otimes\gamma^{v}_{qr}\otimes\delta^{t}_{pv})$ that have the elements (only their pqr edges indicated):

$$\{000\}, \quad \{001,200,120,012\}, \quad \{002,100,210,021\},$$
$$\{010,020,102,201,121,212\}, \tag{A9}$$
$$\{101,011,110,202,022,220,111,222,112,122,221,211\},$$

respectively, together with their complex conjugated quantities. One can easily compute the first normalized F coefficient of the fifth orbit:

$$\check{F}^{(101)_2}_{1,1}(1^{1*}_{10}\otimes 1^{2*}_{11}\otimes 1^{1}_{01}\otimes 1^{2}_{11}) = 1; \tag{A10}$$

then due to the S_4 invariance one deduces that the other coefficients in the same orbit have the value 1, too. Computation of the τ_{12}-transformed quantity

$$\begin{aligned}1 = \tau_{12}(\check{F}^{(101)_2}_{1,1}(1^{1*}_{10}\otimes 1^{2*}_{11}\otimes 1^{1}_{01}\otimes 1^{2}_{11})) &= \overline{\check{F}^{(211)_1}_{0,2}(x(1^{1}_{10})^*\otimes 1^{1*}_{01}\otimes 1^{2}_{11}\otimes x(1^{2}_{11}))}\\ = \overline{\check{F}^{(211)_1}_{0,2}(\omega_{210}\overline{\omega}_{211}\cdot 1^{0*}_{21}\otimes 1^{1*}_{01}\otimes 1^{2}_{11}\otimes 1^{1}_{22})} &= \omega\cdot\overline{\check{F}^{(211)_1}_{0,2}(1^{0*}_{21}\otimes 1^{1*}_{01}\otimes 1^{2}_{11}\otimes 1^{1}_{22})} = \omega\cdot\overline{\omega}\end{aligned} \tag{A11}$$

then shows that one may not have S_4 invariance for a fixed set of basic intertwiners in general, so it is important to take the action of the Frobenius maps into account. In this example even the following stronger statement holds: in the case of $\omega\neq 1$ there is *no* such choice for the set of orthonormal basic intertwiners, $\{b^{q\times r}_{qr} := \omega_{qr}\cdot 1^{q\times r}_{qr} \mid q,\ r=0,1,2\}$ with arbitrary but fixed values of $\omega_{qr}\in\mathbb{C}$, $|\omega_{qr}|=1$, that leads to a constant value of normalized F coefficients within the chosen set of intertwiners. Indeed, the relation

$$\begin{aligned}\check{F}^{(101)_2}_{1,1}(b^{1*}_{10}\otimes b^{2*}_{11}\otimes b^{1}_{01}\otimes b^{2}_{11}) &= \tau_{13}(\check{F}^{(101)_2}_{1,1}(b^{1*}_{10}\otimes b^{2*}_{11}\otimes b^{1}_{01}\otimes b^{2}_{11}))\\ &= \overline{\check{F}^{(101)_2}_{1,1}(y(b^{1}_{10})^*\otimes b^{2*}_{11}\otimes x(b^{1}_{01})\otimes b^{2}_{11})}\end{aligned} \tag{A12}$$

implies that the common value should be ± 1, but then the product of the $\tau_{12}\tau_{34}$- and $\tau_{23}\tau_{34}\tau_{12}\tau_{23}$-transformed F-coefficients leads to the contradiction

$$\begin{aligned}1 \neq \check{F}^{(222)_0}_{1,1}&(x(b^{2}_{11})^*\otimes y(b^{1}_{01})^*\otimes y(b^{2}_{11})\otimes x(b^{1}_{10}))\\ &\times\check{F}^{(111)_0}_{2,2}(xy(b^{2}_{11})^*\otimes xy(b^{1}_{01})^*\otimes xy(b^{2}_{11})\otimes yx(b^{1}_{10})) = \omega^2.\end{aligned} \tag{A13}$$

We also note that for a nontrivial third root of unity, $\omega\neq 1$, there are no solutions of the hexagon equations, that is **Rep***H* cannot be made into a braided category in those cases.

[1] U. Fano and G. Racah, *Irreducible Tensorial Sets* (Academic, New York, 1959).
[2] L. C. Biedenharn, J. M. Blatt, and M. E. Rose, ''Some properties of the Racah and associated coefficients,'' Rev. Mod. Phys. **24**, 249 (1952).
[3] V. K. Agrawala and J. G. Belinfante, ''Graphical formulation of recoupling theory of any compact group,'' Ann. Phys. **49**, 130 (1968).
[4] A. N. Kirillov and N. Yu. Reshetikhin, ''Representations of the algebra $U_q(sl(2))$, q-orthogonal polynomials and

invariants of links,'' in *Infinite-dimensional Lie Algebras and Groups* [Adv. Series in Math. Phys. 7], edited by V. G. Kac (World Scientific, Singapore, 1989), p. 285.
[5] G. Moore and N. Seiberg, ''Classical and quantum conformal field theory,'' Commun. Math. Phys. **123**, 177 (1989).
[6] G. Felder, J. Fröhlich, and G. Keller, ''On the structure of unitary conformal field theory II: Representation theoretic approach,'' Commun. Math. Phys. **130**, 1 (1990).
[7] M. Li and M. Yu, ''Braiding matrices, modular transformations and topological field theories in $2+1$ dimensions,'' Commun. Math. Phys. **127**, 195 (1990).
[8] V. G. Turaev and O. Viro, ''State sum invariants of 3-manifolds and quantum $6j$-symbols,'' Topology **31**, 865 (1992).
[9] J. W. Barrett and B. W. Westbury, ''Invariants of piecewise-linear 3-manifolds,'' Trans. Am. Math. Soc. **348**, 3997 (1996).
[10] V. G. Turaev, *Quantum Invariants of Knots and 3-manifolds* (de Gruyter, New York, 1994).
[11] M. Karowski, W. Müller, and R. Schrader, ''State sum invariants of compact 3-manifolds with boundary and $6j$-symbols,'' J. Phys. A **25**, 4847 (1992).
[12] B. Durhuus, H. P. Jakobsen, and R. Nest, ''A construction of topological quantum field theories from generalized $6J$-symbols,'' Nucl. Phys. B (Proc. Suppl.) **25A**, 109 (1992).
[13] S.-w. Chung, M. Fukuma, and A. Shapere, ''Structure of topological lattice field theories in three dimensions,'' Int. J. Mod. Phys. A **9**, 1305 (1994).
[14] K. Fredenhagen, K.-H. Rehren, and B. Schroer, ''Superselection sectors with braid group statistics and exchange algebras, II: Covariance properties,'' Rev. Math. Phys. (special issue) 111 (1992).
[15] A. Ocneanu, ''Quantum cohomology, quantum groupoids, and subfactors,'' talk presented at the First Caribic School of Mathematics and Theoretical Physics, Guadeloupe, 1993 (unpublished).
[16] R. J. Lawrence, ''Algebras and triangle relations,'' J. Pure Appl. Alg. **100**, 43 (1995).
[17] D. E. Evans and Y. Kawahigashi, ''From subfactors to 3-dimensional topological quantum field theories and back—a detailed account of Ocneanu's theory,'' Int. J. Math. **6**, 537 (1995).
[18] P. Vecsernyés, ''On the quantum symmetry of the chiral Ising model,'' Nucl. Phys. B **415**, 557 (1994).
[19] J. Fuchs, A. Ch. Ganchev, and P. Vecsernyés, ''Rational Hopf algebras: Polynomial equations, gauge fixing, and low-dimensional examples,'' Int. J. Mod. Phys. A **10**, 3431 (1995).
[20] J. Fuchs, A. Ch. Ganchev, and P. Vecsernyés, ''On the quantum symmetry of rational field theories,'' in *Quantum Groups. Formalism and Applications*, edited by J. Lukierski, Z. Popowicz, and J. Sobczyk (PWN Polish Scientific Publishers, Warsaw 1994), p. 473.
[21] D. N. Yetter, ''State-sum invariants of 3-manifolds associated to artinian semisimple tortile categories,'' Topology Appl. **58**, 47 (1994).
[22] P. Bantay and P. Vecsernyés, ''Mapping class group representations and generalized Verlinde formula,'' preprint hep-th/9506186.
[23] G. Böhm and K. Szlachányi, ''A coassociative C^*-quantum group with non-integral dimensions,'' Lett. Math. Phys. **38**, 437 (1996).
[24] P. Degiovanni, ''Moore and Seiberg's equations and $3D$ topological field theory,'' Commun. Math. Phys. **145**, 459 (1992).
[25] I. M. Isaacs, *Character Theory of Finite Groups* (Academic, New York, 1976).
[26] A. A. Kirillov, *Elements of the Theory of Representations* (Springer Verlag, Berlin, 1975).
[27] P. Bantay, ''The Frobenius-Schur indicator in conformal field theory,'' Phys. Lett. B **394**, 87 (1997).
[28] R. Longo and J. E. Roberts, ''A theory of dimension,'' K-Theory **11**, 103 (1997).
[29] P. Freyd and D. N. Yetter, ''Coherence theorems via knot theory,'' J. Pure Appl. Alg. **78**, 49 (1992).
[30] M. E. Sweedler, *Hopf Algebras* (Benjamin/Cummings, Menlo Park, 1969).
[31] K. Szlachányi, ''Weak Hopf algebras,'' in *Proceedings of the Conference on Operator Algebras and Quantum Field Theory* (Rome 1996), edited by S. Doplicher, R. Longo, J. E. Roberts, and L. Zsidó (International, 1997).
[32] G. Böhm, F. Nill, and K. Szlachányi, ''Weak Hopf algebras,'' preprint No. Sfb 288, Berlin 1998.

Hecke algebraic properties of dynamical *R*-matrices. Application to related quantum matrix algebras

L. K. Hadjiivanov[a)]
International Centre for Theoretical Physics (ICTP), I-34014 Trieste, Italy

A. P. Isaev[b)]
Dipartimento di Fisica, Università di Pisa, I-56100 Pisa, Italy

O. V. Ogievetsky[c)]
Centre de Physique Théorique, Luminy, F-13288 Marseille, France

P. N. Pyatov[d)]
Bogoliubov Laboratory of Theoretical Physics, JINR, Dubna, 141980 Moscow Region, Russia

I. T. Todorov[e)]
Erwin Schrödinger Institute for Mathematical Physics (ESI), A-1090 Wien, Austria

(Received 26 February 1998; accepted for publication 5 June 1998)

The quantum dynamical Yang–Baxter (or Gervais–Neveu–Felder) equation defines an *R*-matrix $\hat{R}(p)$, where p stands for a set of mutually commuting variables. A family of $SL(n)$-type solutions of this equation provides a new realization of the Hecke algebra. We define quantum antisymmetrizers, introduce the notion of quantum determinant and compute the inverse quantum matrix for matrix algebras of the type $\hat{R}(p)a_1a_2=a_1a_2\hat{R}$. It is pointed out that such a quantum matrix algebra arises in the operator realization of the chiral zero modes of the WZNW model. © *1999 American Institute of Physics.* [S0022-2488(99)02801-7]

I. INTRODUCTION

Let $\{v^{(i)}, i=1,\dots,n\}$ be a "barycentric basis" in a Cartan subalgebra $\mathfrak{h}\subset \mathrm{sl}(n)$. Viewed as operators in the n-dimensional complex space $V=\mathbb{C}^n$, $v^{(i)}$ can be realized as real traceless diagonal $n\times n$ matrices:

$$(v^{(i)})^j_j=\delta_{ij}-\frac{1}{n} \quad \Rightarrow \quad \sum_{i=1}^{n} v^{(i)}=0\ . \tag{1.1}$$

Let further $\{p_i\}_{i=1}^n$ span the dual Lie algebra $\mathfrak{h}^*$. Introduce the traceless diagonal matrix

$$p=p_iv^{(i)}\left(\equiv\sum_{i=1}^{n} p_iv^{(i)}\right), \quad [p_i,p_j]=0, \quad \sum_{i=1}^{n} p_i=0. \tag{1.2}$$

[a)]On leave of absence from Division of Theoretical Physics, Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Tsarigradsko Chaussee 72, BG-1784 Sofia, Bulgaria; electronic mail address: lhadji@inrne.bas.bg

[b)]On leave of absence from Bogoliubov Laboratory of Theoretical Physics, JINR, Dubna, 141 980 Moscow Region, Russia; electronic mail address: isaevap@ thsun1.jinr.ru

[c)]On leave of absence from P. N. Lebedev Physical Institute, Theoretical Department, 117924 Moscow, Leninsky Prospect 53, Russia; electronic mail address: oleg@cpt.univ-mrs.fr

[d)]Electronic mail address: pyatov@thsun1.jinr.ru

[e)]On leave of absence from Division of Theoretical Physics, Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Tsarigradsko Chaussee 72, BG-1784 Sofia, Bulgaria; electronic mail address: todorov@ inrne.bas.bg

We define a Hecke-type quantum dynamical R-matrix $\hat{R}(p)$ as a map from $\mathfrak{h}^*$ to End $(V\otimes V)$ satisfying the *twisted braid relation*

$$\hat{R}_{12}(p)\hat{R}_{23}(p-v_1)\hat{R}_{12}(p)=\hat{R}_{23}(p-v_1)\hat{R}_{12}(p)\hat{R}_{23}(p-v_1) \tag{1.3}$$

and the Hecke condition

$$\hat{R}(p)^2=\mathbb{I}+(q-\overline{q})\hat{R}(p)\,, \quad \overline{q}:=q^{-1}\,. \tag{1.4}$$

(Although the notation is tailored to the special case in which the parameter q takes values on the unit circle, we shall not use this property in the main body of the paper.) The subscripts in (1.3) refer to the, by now standard, tensor product notation of Faddeev *et al.* (see, e.g., Ref. 1); in particular, $\hat{R}_{23}(p-v_1)\in \mathrm{End}(V^{\otimes 3})$ has matrix elements

$$(\hat{R}_{23}(p-v_1))^{i_1 i_2 i_3}_{j_1 j_2 j_3} = \delta^{i_1}_{j_1}\hat{R}(p-v^{(i_1)})^{i_2 i_3}_{j_2 j_3}\,. \tag{1.5}$$

The twisted braid relation (1.3) is equivalent to the *quantum dynamical*[2,3] (or *deformed*[4]) *Yang–Baxter equation* (QDYBE) for the matrix $R(p)$ related to the braid operator $\hat{R}(p)$ by $\hat{R}(p)=PR(p)$, where P stands for the permutation operator $Px_1y_2=y_1x_2, x,y\in V, P^2=1$. Abusing terminology we shall also refer to Eq. (1.3) by the above abbreviation. The term ''dynamical R-matrix'' for $\hat{R}(p)$ is suggested by the fact that in the physical applications its arguments play the role of (commuting) dynamical variables and that $\hat{R}(p)$ satisfies a finite difference rather than a purely algebraic equation.

The important concept of a *quantum matrix algebra* $\mathcal{A}=\mathcal{A}(\hat{R}(p),\hat{R})$ (Sec. V) can be introduced as a (complex) associative algebra with 1 generated by rational functions of $q^{p_i}, i=1,\dots,n$, and the (noncommuting) entries of an $n\times n$ matrix $a=(a^i_\alpha)$ satisfying the quadratic exchange relations

$$\hat{R}(p)\,a_1\,a_2 = a_1\,a_2\,\hat{R}, \tag{1.6}$$

where $\hat{R}\equiv\hat{R}^{\alpha_1\alpha_2}_{\beta_1\beta_2}$ in the right-hand side is a constant (i.e., p-independent) solution of (1.3), (1.4), and all entries in a matrix row $a^i=\{a^i_\alpha\}^n_{\alpha=1}$ are acting equivalently as shift operators for p:

$$p\,a^i=a^i(p+v^{(i)})\text{or}\;\; p_{jk}a^i=a^i(p_{jk}+\delta^i_j-\delta^i_k)\;\;\text{for}\;\; p_{jk}=p_j-p_k\,. \tag{1.7}$$

Remark 1.1: It has been pointed out[4,5] that, in the su(2) case, the matrix a generates the q-Clebsh–Gordan coefficients while $\hat{R}(p)$ plays the role of a ''quantum $6j$-symbol.''[6–8]

Remark 1.2: Eq. (1.6) is related to the one with indices 1 and 2 interchanged,

$$\hat{R}(p)\,a_2\,a_1 = a_2\,a_1\,\hat{R}, \tag{1.8}$$

by the substitution $\hat{R}\rightarrow P\hat{R}P$, $\hat{R}(p)\rightarrow P\hat{R}(p)\,P$. It is, on the other hand, formally obtained from

$$\hat{R}\,\overline{a}_2\,\overline{a}_1 = \overline{a}_2\,\overline{a}_1\,\hat{R}(p) \tag{1.9}$$

by the substitution $\overline{a}=a^{-1}$; the same substitution relates (1.8) to

$$\hat{R}\,\overline{a}_1\,\overline{a}_2 = \overline{a}_1\,\overline{a}_2\,\hat{R}(p). \tag{1.10}$$

Since

$$\hat{R}(p)_{21}=P\,\hat{R}(p)P \tag{1.11}$$

satisfies conditions of the same type as $\hat{R}(p)$, we can start with either of these relations.

The QDYBE, introduced by Gervais and Neveu[2] for the exchange algebra associated with the Liouville equation and applied to the zero mode algebra[5] of the Wess–Zumino–Novikov–Witten (WZNW) model[9–11] is attracting ever more attention. Its classical counterpart, introduced in Ref. 3 (see also Ref. 12) has been displayed in Ref. 5 for the sl(2) case and in Ref. 13 for an arbitrary simple Lie algebra. The quantum $\hat{R}(p)$ is central to a continuing study[4] of q-deformed cotangent bundles on group manifolds and quantum model spaces. A quasi-Hopf-algebraic point of view is taken in Refs. 14,15, where $\hat{R}(p)$ is obtained by a Drinfeld twist of the constant R-matrix. Felder[16] explores the more general case of (classical and) QDYBE depending on a spectral parameter and finds elliptic solutions of this equation. These solutions are applied in Ref. 17 to quantize Calogero–Moser and Ruijsenaars–Schneider models. A class of $\mathrm{SL}(n|m)$-type solutions of the QDYBE (and related trigonometric solutions of the equation with spectral parameter) are described in Ref. 18. A more mathematically minded approach to the subject in terms of ''$\mathfrak{h}$-algebroids'' is being developed in Ref. 19.

The present work was motivated in part by earlier study[20–22] of the canonical quantization of the WZNW model (following Refs. 6,10,5,11,23). It was noticed, in particular, that the exchange relations[21] for the chiral zero modes a^i_α that diagonalize the $U_q(\mathrm{sl}(2))$ monodromy matrix can be written in the form (1.8). As a result, the operator realization of the chiral group valued field was understood as a quantization of the (deformed) classical Poisson bracket relations of Ref. 13 thus opening the way to its generalization for $\mathrm{SU}(n)$. Here we show that a special solution of the QDYBE (1.3) yields a new matrix representation of the Hecke algebra. We concentrate on a general study of the Hecke algebra properties of this solution and the ensuing properties of the quantum matrices satisfying (1.6) relegating applications to the WZNW model to a subsequent publication[24] which is highlighted in Sec. VI. A central result is the computation of the quantum determinant of a and the (based on it) evaluation of the inverse quantum matrix.

The paper is organized as follows. We review and extend in Sec. II results of Gurevich[25] on quantum (anti)symmetrizers and illustrate them in Sec. III on the known example of a constant $\hat{R}$. We proceed in Sec. IV to a study of a family of $\mathrm{SL}(n)$-type dynamical R-matrices and describe two types of symmetry transformations for this family: the twist transformation (a version of Drinfeld's twist for dynamical R-matrices) and the canonical shifts. In Sec. V we show that these dynamical matrices provide a new realization of the Hecke algebra. This allows us to define ''dynamical'' (p-dependent) analogs of quantum antisymmetrizers, including the Levi-Cività $\mathcal{E}$ tensor. In Sec. VI we study the quantum algebra $\mathcal{A}$, define the quantum determinant $\det(a)$, and compute the inverse matrix a^{-1}. We demonstrate that $\mathcal{A}$ provides a realization of a reflection equation algebra which is interpreted as a quantum monodromy algebra in the WZNW theory. An Appendix is devoted to deriving some useful identities for the parameters determining the solution of the QDYBE found in Ref. 18 and to computing the normalization of the dynamical Levi-Cività tensor.

II. HECKE ALGEBRAS AND *q*-ANTISYMMETRIZERS

In this section we collect some basic notions on Hecke algebras and describe the q-antisymmetrizers technique which is to be applied later on. We follow closely the approach of Gurevich.[25]

In the present context by a Hecke algebra $\mathcal{H}_k(q)$ we understand a $\mathbb{C}$-algebra with generators $1, g_1, g_2, \ldots, g_{k-1}$, a nonzero parameter $q \in \mathbb{C}$, and defining relations

$$g_i\, g_{i+1}\, g_i = g_{i+1}\, g_i\, g_{i+1} \quad \text{for } 1 \leqslant i \leqslant k-2, \tag{2.1}$$

$$g_i^2 = 1 + (q - \overline{q})\, g_i \quad \text{for } 1 \leqslant i \leqslant k-1, \tag{2.2}$$

$$g_i\, g_j = g_j\, g_i \quad \text{if } |i-j| \geqslant 2, \tag{2.3}$$

where $\overline{q} := q^{-1}$.

We shall consider the set of idempotents $A^{(j)} \in \mathcal{H}_k(q), j=1,\ldots,k$, associated with single column Young diagrams containing j nodes — the so-called q-antisymmetrizers. Their inductive definition is given by

$$A^{(1)}=1, \quad A^{(j)}=\frac{1}{[j]}A^{(j-1)}(q^{j-1}-[j-1]g_{j-1})A^{(j-1)}. \tag{2.4}$$

Here $[j]=(q^j-\overline{q}^j)/(q-\overline{q})$ and we assume $[j]\neq 0$, for $j=2,\ldots,k$. Note, that $A^{(k)}$ is a central idempotent in the algebra $\mathcal{H}_k(q)$.

Equivalently, one may write

$$A^{(j)}=\frac{1}{[j]}A^{(2,j-1)}(q^{j-1}-[j-1]g_1)A^{(2,j-1)}, \tag{2.5}$$

where we have adopted the notation $A^{(i,j)}$, $1\leqslant i\leqslant j$, for the central idempotent of the subalgebra $\mathcal{H}_{i,j}(q)\subset\mathcal{H}_j(q)$ generated by the subset $1,g_i,\ldots,g_{j-1}$. In particular, $A^{(1,j)}=A^{(j)}$, $A^{(j,j)}=1$.

Remark 2.1: All the subalgebras $\mathcal{H}_{i,r+i}(q)\subset\mathcal{H}_k(q)$, $i=1,\ldots,k-r$, are isomorphic by definition. Moreover, they are related by inner $\mathcal{H}_k(q)$-automorphisms. For example, the automorphism $\phi_i:\mathcal{H}_{i,r+i}(q)\to\mathcal{H}_{i+1,r+i+1}(q)$ is given by

$$\phi_i(t)=g_ig_{i+1}\ldots g_{r+i}\,t\,(g_ig_{i+1}\ldots g_{r+i})^{-1}, \quad \forall t\in\mathcal{H}_{i,r+i}(q). \tag{2.6}$$

The term q-antisymmetrizer for the elements $A^{(j)}$ is justified by the following properties:

$$(g_i+\overline{q})A^{(j)}=A^{(j)}(g_i+\overline{q})=0 \quad \text{for} \quad 1\leqslant i\leqslant j-1, \tag{2.7}$$

$$A^{(j)}A^{(i,l)}=A^{(i,l)}A^{(j)}=A^{(j)} \quad \text{for} \quad 1\leqslant i\leqslant l\leqslant j. \tag{2.8}$$

Remark 2.2: Replacing q by $(-q^{-1})$ in (2.4) leads to another sequence of projectors, called symmetrizers. Abstractly, inside the Hecke algebra, it is a matter of convention — which projectors one calls symmetrizers, and which — antisymmetrizers. We use the common convention. However, on the level of representations, when one can calculate the ranks of projectors and see which sequence of projectors terminates, the distinction between symmetrizers and antisymmetrizers becomes meaningful.

Consider a representation $\rho_{W,k}:\mathcal{H}_k(q)\to\mathrm{Aut}(W)$ of the algebra $\mathcal{H}_k(q)$ in a vector space W.

Definition 2.1: We shall say that $\rho_{W,k}$ is a representation of height n in one of the following two cases:

(a) $n<k$ and the conditions

$$\rho_{W,k}(A^{(n+1)})=0, \tag{2.9}$$

$$\mathrm{rank}\,\rho_{W,k}(A^{(n)})=1, \tag{2.10}$$

are fulfilled, or

(b) $n=k$ *and* $\mathrm{rank}\,\rho_{W,n}(A^{(n)})=1$.

Remark 2.3: The notion of height of a Hecke algebra representation was introduced in Ref. 25 for the special case of the representations generated by constant R-matrices. There it was named *the rank of the R-matrix*. We have changed the name here in order to avoid a possible confusion with the standard notion of rank of a matrix. Note that the use of the term ''height'' is suggested by the fact that imposing condition (2.9) for the representation $\rho_{W,k}$ results in vanishing of any central (and primitive) idempotent related to a Young diagram (standard tableaux) containing more than n boxes in one of its columns.

Remark 2.4: In view of Remark 2.1, the whole sets of q-antisymmetrizers $\{\rho_{W,k}(A^{(i,n+i)})\}_{i=1,\ldots,k-n}$ and $\{\rho_{W,k}(A^{(j,n+j-1)})\}_{j=1,\ldots,k-n+1}$ satisfy conditions (2.9) and (2.10), respectively.

Remark 2.5: Instead of using (2.9) one can impose the condition

$$A^{(n+1)}=0 \tag{2.11}$$

at the algebraic level. This is the way how generalized Temperley–Lieb–Martin algebras are defined (cf. Ref. 26). Below we present several useful equivalent forms of this condition.

Lemma 2.1: The condition (2.11) is equivalent to any of the following relations:

$$A^{(n)}g_n\cdots g_2g_1=(-1)^{n-1}q\,[n]A^{(n)}A^{(2,n+1)}, \tag{2.12a}$$

$$g_1g_2\cdots g_nA^{(n)}=(-1)^{n-1}q\,[n]A^{(2,n+1)}A^{(n)}, \tag{2.12b}$$

$$g_n\cdots g_2g_1A^{(2,n+1)}=(-1)^{n-1}q\,[n]A^{(n)}A^{(2,n+1)}, \tag{2.12c}$$

$$A^{(2,n+1)}g_1g_2\cdots g_n=(-1)^{n-1}q\,[n]A^{(2,n+1)}A^{(n)}, \tag{2.12d}$$

$$A^{(n)}A^{(2,n+1)}A^{(n)}=[n]^{-2}A^{(n)}, \tag{2.12e}$$

$$A^{(2,n+1)}A^{(n)}A^{(2,n+1)}=[n]^{-2}A^{(2,n+1)}. \tag{2.12f}$$

Proof: Applying repeatedly (2.4) for the q-antisymmetrizers that appear as last factors in the resulting products and using (2.7), (2.8) we find

$$\begin{aligned}A^{(n+1)}&=\frac{1}{[n+1]}A^{(n)}(q^n-[n]g_n)A^{(n)}\\&=\frac{1}{[n+1]}\{q^nA^{(n)}-A^{(n)}g_n(q^{n-1}-[n-1]g_{n-1})A^{(n-1)}\}\\&=\frac{1}{[n+1]}\{A^{(n)}(q^n-q^{n-1}g_n)+A^{(n)}g_ng_{n-1}(q^{n-2}-[n-2]g_{n-2})A^{(n-2)}\}\cdots\\&=\frac{1}{[n+1]}A^{(n)}(q^n-q^{n-1}g_n+\cdots+(-1)^ng_ng_{n-1}\cdots g_1).\end{aligned} \tag{2.13}$$

Next, we apply $A^{(2,n+1)}$ to the both sides of Eq. (2.13). Using again (2.7) and (2.8) we obtain

$$A^{(n+1)}=\frac{1}{[n+1]}A^{(n)}\{q[n]A^{(2,n+1)}+(-1)^ng_ng_{n-1}\cdots g_1A^{(2,n+1)}\}.$$

Taking into account the relation $(g_ng_{n-1}\cdots g_1)A^{(2,n+1)}=A^{(n)}(g_ng_{n-1}\cdots g_1)$ which is a consequence of (2.1) we end up with

$$A^{(n+1)}=\frac{1}{[n+1]}\{q[n]A^{(n)}A^{(2,n+1)}+(-1)^nA^{(n)}g_ng_{n-1}\cdots g_1\}. \tag{2.14}$$

This proves the equivalence of Eqs. (2.11) and (2.12a). A similar argument using iteratively a substitution of the first q-antisymmetrizer in the right-hand side of (2.4) implies the equivalence of (2.11) and (2.12b). Condition (2.11) is transformed to the forms (2.12c) and (2.12d) in the same manner starting from Eq. (2.5).

To show equivalence of (2.11) to (2.12e) and to (2.12f) one should employ Eqs. (2.4) and (2.5), respectively. We shall treat the case of (2.12e) here.

Consider the difference

$$[n]^2 A^{(n)} A^{(2,n+1)} A^{(n)} - A^{(n)} = A^{(n)}([n]^2 A^{(2,n+1)} - 1) A^{(n)}$$
$$= A^{(n)}\{[n] A^{(2,n)}(q^{n-1} - [n-1] g_n) A^{(2,n)} - 1\} A^{(n)}$$
$$= [n-1] A^{(n)}(q^n - g_n) A^{(n)} = [n-1][n+1] A^{(n+1)},$$

where we have again used the definition (2.4) and the relations (2.8). Comparing the first and the last lines of the calculation we deduce the equivalence of conditions (2.11) and (2.12e). ■

Equations (2.12a)–(2.12f) display properties of the rank 1 idempotents $\rho_{W,k}(A^{(n)})$ that are hidden in (2.11). In fact, they are the basic technical tools which one needs to effectively deal with the height n Hecke algebra representations.

In the rest of the paper we make use of a special type of representations of the algebras $\mathcal{H}_k(q)$ for which the representation space is given by kth tensor power of an [n-dimensional, in the case of SL(n)] vector space V: $W = V^{\otimes k}$. These representations are generated by constant or dynamical R-matrices of Hecke-type. The representations we are dealing with have the specific feature that their height, when defined (i.e., for $k \geqslant n$), coincides with the dimension of the space V.[27]

Below we first illustrate the general notions introduced above on the well known case of constant SL(n)-type R-matrices relegating the study of dynamical R-matrices to Secs. IV and V.

III. REPRESENTATIONS GENERATED BY A CONSTANT *R*-MATRIX OF THE SL(*n*) TYPE

The R-matrix corresponding to the Drinfeld–Jimbo deformation of SL(n) (Refs. 28,29) is an operator acting in a tensor square of an n-dimensional vector space V and given by

$$\hat{R}^{\alpha_1\alpha_2}_{\beta_1\beta_2} = q^{\delta_{\alpha_1\alpha_2}} \delta^{\alpha_1}_{\beta_2}\delta^{\alpha_2}_{\beta_1} + (q - \overline{q})\, \theta_{\alpha_2\alpha_1}\, \delta^{\alpha_1}_{\beta_1}\delta^{\alpha_2}_{\beta_2}, \tag{3.1}$$

(no summation in the right-hand side) where the indices α, β take values from 1 to n, and $\theta_{\alpha\beta} = \{1 \text{ if } \alpha > \beta, \ 0 \text{ if } \alpha \leqslant \beta\}$.

This R-matrix is a particular representative of a family of constant Hecke R-matrices, i.e., it satisfies the braid relation and the Hecke condition

$$\hat{R}_{12}\hat{R}_{23}\hat{R}_{12} = \hat{R}_{23}\hat{R}_{12}\hat{R}_{23}, \tag{3.2}$$

$$\hat{R}^2 = \mathbb{I} + (q - \overline{q})\, \hat{R}. \tag{3.3}$$

Equations (3.2), (3.3) imply that the matrices $\hat{R}_{12}, \hat{R}_{23}$ generate a representation of $\mathcal{H}_3(q)$ in $V^{\otimes 3}$. For an arbitrary k the representation $\rho_{\hat{R},k} : \mathcal{H}_k(q) \to \mathrm{Aut}(V^{\otimes k})$ generated by a constant Hecke R-matrix is defined by

$$\rho_{\hat{R},k}(g_i) = \hat{R}_{i\, i+1}. \tag{3.4}$$

For representations $\rho_{\hat{R},k}$ generated by the R-matrix (3.1) we have

$$\text{height}\, \rho_{\hat{R},k} = n \quad \text{if } k \geqslant n.$$

The rank 1 q-antisymmetrizers $\rho_{\hat{R},k}(A^{(i,n+i-1)})$ are most conveniently described in terms of q-analogs of (co- and contravariant) Levi-Cività tensors which are solutions of the equations

$$\hat{R}^{\alpha_i\alpha_{i+1}}_{\beta_i\beta_{i+1}} \varepsilon^{\alpha_1\cdots\beta_i\beta_{i+1}\cdots\alpha_n} = -\overline{q}\, \varepsilon^{\alpha_1\cdots\alpha_i\alpha_{i+1}\cdots\alpha_n},$$

$$\varepsilon_{\alpha_1\cdots\beta_i\beta_{i+1}\cdots\alpha_n} \hat{R}^{\beta_i\beta_{i+1}}_{\alpha_i\alpha_{i+1}} = -\overline{q}\, \varepsilon_{\alpha_1\cdots\alpha_i\alpha_{i+1}\cdots\alpha_n}, \quad i = 1,2,\ldots,n-1.$$

It is straightforward to prove that these equations have unique (up to normalization) solutions. The rank1 condition (2.10) follows as a corollary.

In the special case of representations $\rho_{\hat{R},k}$ generated by the R-matrix (3.1) the only nonvanishing components of the ε-tensors have pairwise different indices $\alpha_1,\alpha_2,\ldots,\alpha_n$, and can be chosen as

$$\varepsilon^{\alpha_1\alpha_2\cdots\alpha_n}=\overline{q}^{n(n-1)/2}(-q)^{l(\sigma)},\quad \varepsilon_{\alpha_1\alpha_2\cdots\alpha_n}=(-q)^{l(\sigma)}. \tag{3.5}$$

Here $l(\sigma)$ is the length of the permutation $\sigma=\binom{1,\ \ 2,\ \ \ldots,\ \ n}{\alpha_1,\ \alpha_2,\ \ldots,\ \alpha}$.

The rank 1 q-antisymmetrizers are expressed in terms of the ε-tensors as

$$\rho_{\hat{R},k}(A^{(i,n+i-1)})=\frac{1}{[n]!}\,\varepsilon^{|i\cdots n+i-1\rangle}\varepsilon_{\langle i\cdots n+i-1|},\quad i=1,\ldots,k-n+1. \tag{3.6}$$

Here (by analogy with the matrix notation) we substitute the vector space indices of ε-tensors by their labels: $\alpha_i\to i$. The ''bra'' and ''ket'' notation of ε-tensor indices is used in order to distinguish labels of matrix spaces from those of vector spaces. One should have in mind the following symbolic decomposition for the matrix space label: $i=|i\rangle\otimes\langle i|$. For example, the equation $A_i\,u^{|i\rangle}(\equiv A^{|i\rangle}{}_{\langle i|}\,u^{|i\rangle}=v^{|i\rangle}$ is to be understood as $\Sigma_{\beta_i}A^{\alpha_i}_{\beta_i}u^{\beta_i}=v^{\alpha_i}$.

Finally, we shall adapt for $\rho_{\hat{R},k}$ those formulas (2.12a)–(2.12f) which will be used in Sec. VI. Written in terms of the ε-tensors the relations (2.12b), (2.12c), (2.12e) and (2.12f) assume the form

$$\rho_{\hat{R},n+1}(g_1\cdots g_n)\varepsilon^{|1\cdots n\rangle}\equiv\hat{R}_{12}\cdots\hat{R}_{n\,n+1}\varepsilon^{|1\cdots n\rangle}=q\,\varepsilon^{|2\cdots n+1\rangle}\,N^{|1\rangle}{}_{\langle n+1|}, \tag{3.7}$$

$$\rho_{\hat{R},n+1}(g_n\cdots g_1)\varepsilon^{|2\cdots n+1\rangle}\equiv\hat{R}_{n\,n+1}\cdots\hat{R}_{12}\varepsilon^{|2\cdots n+1\rangle}=q\,\varepsilon^{|1\cdots n\rangle}\,K^{|n+1\rangle}{}_{\langle 1|}, \tag{3.8}$$

$$K\,N=N\,K=\mathbb{I}. \tag{3.9}$$

Here the matrices N and K are defined as

$$N^{|1\rangle}{}_{\langle n+1|}=\frac{(-1)^{n-1}}{[n-1]!}\,\varepsilon_{\langle 2\cdots n+1|}\varepsilon^{|1\cdots n\rangle}, \tag{3.10}$$

$$K^{|n+1\rangle}_{\langle 1|}=\frac{(-1)^{n-1}}{[n-1]!}\,\varepsilon_{\langle 1\cdots n|}\varepsilon^{|2\cdots n+1\rangle}, \tag{3.11}$$

and for the ε-tensors given by Eq. (3.5) we have just

$$N=K=\mathbb{I}. \tag{3.12}$$

IV. SL(n)-TYPE DYNAMICAL R-MATRICES

We now turn to the dynamical R-matrix defined in the Introduction. In order to present the QDYBE in a form suitable for our purposes we shall introduce a set of commutative variables

$$X^i,\quad i=1,\ldots,n,\quad [X^i,X^j]=0\,,\quad \prod_{i=1}^{n}X^i=1, \tag{4.1a}$$

which play the role of elementary shift operators for p_i:

$$p\,X^i=X^i\,(\,p+v^{(i)}\,)\,. \tag{4.1b}$$

The elements X^i and q^{p_i} provide a realization of (the Weyl's form of) the canonical commutation relations. Note that in concrete applications they can be naturally identified with (a subset of) dynamical variables of a model (see, e.g., Ref. 5).

Let us arrange the auxiliary variables X^i into a unimodular diagonal matrix,

$$X=\mathrm{diag}\{X^1,\ldots,X^n\},\quad \det(X)=1. \tag{4.2}$$

The *Hecke-type dynamical R-matrix* is characterized by the following set of relations:

$$\hat{R}_{12}(p)(X_1\hat{R}_{23}(p)X_1^{-1})\hat{R}_{12}(p)=(X_1\hat{R}_{23}(p)X_1^{-1})\hat{R}_{12}(p)(X_1\hat{R}_{23}(p)X_1^{-1}), \tag{4.3}$$

$$\hat{R}(p)^2=\mathbb{I}+(q-\overline{q})\hat{R}(p), \tag{4.4}$$

$$\hat{R}_{12}(p)X_1X_2=X_1X_2\hat{R}_{12}(p). \tag{4.5}$$

Here the first and second relations are the dynamical Yang–Baxter equation and the Hecke condition, respectively. A condition of type (4.5), although not always imposed on dynamical *R*-matrices, is also necessary in our treatment. As we shall see below, it ensures that conditions (2.3) for the Hecke algebra representations generated by $\hat{R}(p)$ are satisfied.

Following Ref. 18 we shall consider dynamical *R*-matrices of the form

$$\hat{R}^{i_1i_2}_{j_1j_2}(p)=a_{i_1i_2}(p)\,\delta^{i_1}_{j_2}\delta^{i_2}_{j_1}+b_{i_1i_2}(p)\,\delta^{i_1}_{j_1}\delta^{i_2}_{j_2},\quad i_{1,2},j_{1,2}=1,\ldots,n \tag{4.6}$$

(there is no summation over repeated upper and lower indices in the right-hand side); in order to have a unique decomposition in terms of the unit and the permutation matrices in the tensor square of spaces we impose the condition $b_{ii}(p)=0$. This special class of *p*-dependent Hecke *R*-matrices will be called *dynamical R-matrices of an* SL*(n)-type*.[30]

The unknown functions $a_{ij}(p)$, $b_{ij}(p)$ in the Ansatz (4.6) are to be fixed by the conditions (4.3)–(4.5). The Hecke condition (4.4) gives

$$b_{ii}=0,\ \ b_{ij}+b_{ji}=q-\overline{q},\quad \text{for}\ \ i\neq j, \tag{4.7}$$

$$a_{ij}\,a_{ji}-b_{ij}\,b_{ji}=1,\quad \text{for}\ \ i\neq j, \tag{4.8}$$

$$a_{ii}^2-(q-\overline{q})\,a_{ii}=1. \tag{4.9}$$

The last equation has two solutions: $a_{ii}=\pm\, q^{\pm 1}$ for each *i*. Below we consider only the case $a_{ii}=q,\ \forall i$ (the other cases correspond, in particular, to quantum supergroups and have been considered in Ref. 18). Finally, the dynamical Yang–Baxter equation (4.3) and Eq. (4.5) impose the constraints

$$a_{ij}(p_1,\ldots,p_n)=a_{ij}(p_{ij}),\quad b_{ij}(p_1,\ldots,p_n)=b_{ij}(p_{ij}),$$

$$b_{ij}b_{jk}b_{ki}+b_{ik}b_{kj}b_{ji}=0, \tag{4.10}$$

$$b_{ij}(p_{ij}+1)=\frac{b_{ij}(p_{ij})\,q}{\overline{q}+b_{ij}(p_{ij})}, \tag{4.11}$$

where $p_{ij}:=p_i-p_j$. For $a_{ii}=q$ the general solution of (4.7)–(4.11) can be written as[18]

$$a_{ij}(p)=\alpha_{ij}(p_{ij})\,\xi_{ij}(p_{ij}),\quad b_{ij}(p)=q-\xi_{ij}(p_{ij}), \tag{4.12}$$

where $\xi_{ij}(p)$are expressed as the following ratios:

$$\xi_{ij}(p)=\frac{f(p_{ij}-1,\beta_{ij})}{f(p_{ij},\beta_{ij})}, \quad f(p,\beta)=\overline{q^p}+[p]\beta. \tag{4.13}$$

Here $\beta_{ij}(p_{ij})=\beta_{ij}(p_{ij}+1)$. We shall consider β_{ij} as constant parameters since their functional dependence does not change any of the results below. The function $f(p,\beta)$ satisfies the finite difference equation

$$f(p+1,\beta)+f(p-1,\beta)=[2]f(p,\beta), \tag{4.14}$$

with the initial conditions

$$f(0,\beta)=1, \quad f(1,\beta)=\overline{q}+\beta. \tag{4.15}$$

Equations (4.14) can be deduced from (4.11).

Not all of the remaining in (4.12), (4.13) parameters $\alpha_{ij}(p_{ij})$ and β_{ij} are independent. The relations between them are given by

$$\alpha_{ii}=1\,, \quad \alpha_{ij}(p_{ij})\alpha_{ji}(p_{ji})=1, \tag{4.16}$$

$$\beta_{ii}=0\,, \quad \beta_{ij}+\beta_{ji}=q-\overline{q} \qquad \text{for} \qquad i\neq j, \tag{4.17a}$$

$$\beta_{ij}\beta_{jk}\beta_{ki}+\beta_{ik}\beta_{kj}\beta_{ji}=0. \tag{4.17b}$$

An easy way to solve Eqs. (4.17a), (4.17b) is to make the substitution

$$\beta_{ij}=\frac{q-\overline{q}}{1-\pi_{ij}} \quad \Leftrightarrow \quad \pi_{ij}=\frac{\beta_{ij}-q+\overline{q}}{\beta_{ij}}, \qquad \text{for } i\neq j. \tag{4.18a}$$

We stress that the parameters π_{ii} are not fixed here and can be chosen as arbitrary constants. In terms of the new variables π_{ij} Eqs. (4.17a), (4.17b) take the simple form

$$\pi_{ij}\,\pi_{ji}=1, \quad \pi_{ij}\,\pi_{jk}\,\pi_{ki}=1, \tag{4.18b}$$

and are solved by $\pi_{ij}=\Pi_{k=i}^{j-1}\pi_{k\,k+1}=\pi_{ji}^{-1}$, for $i<j$. Hence,

$$\beta_{ij}=\frac{(q-\overline{q})\Pi_{k=i}^{j-1}\beta_k}{\Pi_{k=i}^{j-1}\beta_k-\Pi_{k=i}^{j-1}(\beta_k-q+\overline{q})}, \quad \text{for } i<j, \tag{4.19}$$

where the remaining $(n-1)$ parameters $\beta_i\equiv\beta_{i\,i+1}$ are independent.

We shall describe two types of transformations on the set of SL(n)-type dynamical R-matrices.

(1) a simple version of *twisting* for dynamical R-matrices:

$$\hat{R}(p)_{21}\rightarrow{}^F\hat{R}_{21}=\hat{F}\,\hat{R}(p)\,\hat{F}^{-1},$$

where $\hat{F}(p)=F_{12}(p)P_{12}$ and

$$F_{12}\equiv F_{j_1j_2}^{i_1i_2}(p)=\delta_{j_1}^{i_1}\,\delta_{j_2}^{i_2}\,\psi_{i_1i_2}(p), \tag{4.20}$$

(an analog of Drinfeld's twist, see Ref. 31). An explanation on how this twist works is given in the two Lemmas below.

Lemma 4.1: Let F(p) be an operator acting in V⊗V. If

$$\hat{F}_{12}^{-1} X_1^{-1} \hat{F}_{23} = \hat{A}_{123} X_3^{-1} \hat{A}_{123} \tag{4.21}$$

for some operator A(p) acting in V⊗V⊗V and

$$\hat{R}(p)_{12} \hat{A}_{123} = \hat{A}_{123} \hat{R}(p)_{23} \tag{4.22}$$

then the matrix ${}^F\hat{R}_{21} = \hat{F}\hat{R}(p)\hat{F}^{-1}$ *satisfies the QDYBE (4.3).*

Proof: Substitute ${}^F\hat{R}_{21}$ in the QDYBE taking into account that the QDYBE has two equivalent forms:

$$\hat{R}(p)_{23} X_3^{-1} \hat{R}(p)_{12} X_3 \hat{R}(p)_{23} = X_3^{-1} \hat{R}(p)_{12} X_3 \hat{R}(p)_{23} X_3^{-1} \hat{R}(p)_{12} X_3 ,$$

$$\hat{R}(p)_{21} X_1^{-1} \hat{R}(p)_{32} X_1 \hat{R}(p)_{21} = X_1^{-1} \hat{R}(p)_{32} X_1 \hat{R}(p)_{21} X_1^{-1} \hat{R}(p)_{32} X_1 .$$

The first equation results from repeated application of (4.5) to (4.3). The second equation is obtained from the first one by simply permuting the subscripts 1 and 3. ■

Lemma 4.2: Let F_{12} *be the diagonal matrix (4.20) where* $\psi_{ij}\psi_{ji}=1$ *and* $\psi_{ii}=1$. *Assume that* ψ_{ij} *depends on* p_{ij} *only. Then, Eqs. (4.21), (4.22) are satisfied by* $\hat{A} = AP_{23}P_{12}$, *where the matrix A is diagonal,* $A^{ijk}_{abc} = a_{ijk}\delta^i_a \delta^j_b \delta^k_c$, *the elements* a_{ijk} *being given by*

$$a_{ijk} = \begin{cases} \psi_{ik}\psi_{jk} & \text{if } i \neq j, \\ [\psi_{ik}(p_{ik}+1)]^2 & \text{if } i = j. \end{cases} \tag{4.23}$$

Proof: The operator A_{123} is symmetric in the first two indices, $A_{123} = A_{213}$ which implies that it commutes with any *R*-matrix of the form (4.6). Therefore, (4.22) is satisfied.

Equation (4.21) can be checked directly. ■

These Lemmas demonstrate that the operator (4.20) indeed generates a twist leading to the changes $\alpha_{ij} \to \alpha_{ij}\psi_{ji}^2$, $\beta_{ij} \to \beta_{ij}$ of the parameters in (4.12), (4.13).

(2) *Canonical transformations* of the dynamical parameters $p_i \to p_i + c_i$, where c_i, $i = 1, \ldots, n$ are arbitrary constants satisfying the condition $\Sigma_{i=1}^n c_i = 0$.

We conclude the section by a brief discussion of the structure of the family of SL(n)-type dynamical *R*-matrices (4.6), (4.12), (4.13). There are two essentially different domains for the parameters β_i of this family.

(a) $\beta_i \neq 0$ and $\beta_i \neq q - \overline{q}$, for all *i*.

In this case the whole family (4.6), (4.12), (4.13) can be generated starting from any particular representative with the use of the two types of transformations described above.

Indeed, the parameters α_{ij} can be excluded with the help of a twist. Then, performing a canonical transformation of the form

$$q^{2p_{ij}} \to q^{2p_{ij}} \pi_{ij} = q^{2p_{ij}} \prod_{k=i}^{j-1} \frac{\beta_k - q + \overline{q}}{\beta_k} \quad \text{for } i<j,$$

for instance, one excludes the parameters β_{ij} from the ansatz (4.6), (4.12), (4.13) and passes to a dynamical *R*-matrix with

$$\xi_{ij}(p) = \frac{[p_{ij}-1]}{[p_{ij}]} \tag{4.24}$$

[cf. (4.13)–(4.17b)]. This *R*-matrix is the limiting case $\beta_i \to \infty$ of our family, and it is this type of dynamical *R*-matrix which is discussed in Ref. 5.

(b) Either all $\beta_i = q - \overline{q}$, or all $\beta_i = 0$.

We shall consider the first case $\beta_i = q - \overline{q}$ for which

$$\beta_{ij} = \begin{cases} q - \overline{q} & \text{for} \quad i < j \\ 0 & \text{for} \quad i \geqslant j \end{cases} \tag{4.25}$$

and put $\alpha_{ij}(p_{ij}) = \mathrm{const}_{ij}$. In this case the R-matrix (4.6), (4.12), (4.13) becomes independent on the dynamical variables p_i and is reduced to the constant R-matrix describing the multiparametric[32] deformations of $\mathrm{GL}(n)$ which are all twist-equivalent.

With the particular choice

$$\alpha_{ij} = \begin{cases} q & \text{for} \quad i < j \\ 1 & \text{for} \quad i = j \\ \overline{q} & \text{for} \quad i > j \end{cases} \tag{4.26}$$

one reproduces the standard $\mathrm{SL}(n)$-type R-matrix (3.1).

Remark 4.1: In the intermediate cases (where only a part of the parameters β_i are equal to 0 or $q - \overline{q}$) the corresponding dynamical R-matrix $\hat{R}(p)$ contains the (dynamical and constant) R-matrices described in (a) and (b) as submatrices.

V. REPRESENTATIONS GENERATED BY THE SL(n)-TYPE DYNAMICAL R-MATRICES

Now we are in a position to introduce the Hecke algebra representations associated with Hecke-type dynamical R-matrices.

Proposition 5.1: Let $\hat{R}(p)$ *be a dynamical R-matrix of the Hecke type. The matrices*

$$\rho_{\hat{R}(p),k}(g_i) = (X_1 X_2 \cdots X_{i-1}) \hat{R}_{i,i+1}(p) (X_1 X_2 \cdots X_{i-1})^{-1}, \quad i = 1, \ldots, k-1, \tag{5.1a}$$

generate a Hecke algebra representation, $\rho_{\hat{R}(p),k} : \mathcal{H}_k(q) \rightarrow \mathrm{Aut}(V^{\otimes k})$.

Proof: Obviously, Eq. (4.3) implies that the matrices $\rho_{\hat{R}(p),k}(g_i)$ and $\rho_{\hat{R}(p),k}(g_{i+1})$ satisfy the braid relations (2.1). Then, the conditions (4.5) ensure that the matrices (5.1a) satisfy (2.3) and, therefore, (5.1a) represent the generators of the braid algebra $\mathcal{B}_k$. Finally, the Hecke conditions (2.2) for the generators (5.1a) follow from the Hecke property (4.4) of the dynamical R-matrix (2.2). ∎

Remark 5.1: In contrast with the case of constant Hecke R-matrix (3.4) the representations generated by a dynamical Hecke R-matrix are nonlocal; in other words, the matrices $\rho_{\hat{R}(p),k}(g_i)$ act nontrivially as diagonal matrices on V_j with $j < i$ (and not merely on $V_i \otimes V_{i+1}$). Only the representation of the first generator with $i = 1$ has the usual ''locality'' property.

Remark 5.2: One can construct representations equivalent to $\rho_{\hat{R}(p),k}$ in which some other generator is ''localized'' instead g_1. For instance, the representation $\overline{\rho}_{\hat{R}(p),k}$ which localizes $\overline{\rho}_{\hat{R}(p),k}(g_{k-1})$ is given by

$$\overline{\rho}_{\hat{R}(p),k}(t) = (X_1 X_2 \cdots X_k)^{-1} \rho_{\hat{R}(p),k}(t) (X_1 X_2 \cdots X_k), \quad \forall t \in \mathcal{H}_k(q) \tag{5.1b}$$

so that

$$\overline{\rho}_{\hat{R}(p),k}(g_i) = (X_{i+2} \cdots X_k)^{-1} \hat{R}_{i,i+1}(p) (X_{i+2} \cdots X_k). \tag{5.1c}$$

Note that in addition to the nonlocal property the representation matrices of $\overline{\rho}_{\hat{R}(p),k}$ depend explicitly on k.

From now on we shall restrict ourselves to discussing those representations $\rho_{\hat{R}(p),k}$ which are generated by the $\mathrm{SL}(n)$-type dynamical R-matrices (4.6), (4.12), (4.13). For $k \geqslant n$ all these representations are of height *n*. The rank *1q*-antisymmetrizers are conveniently expressed in terms of dynamical $\mathcal{E}$-tensors $\mathcal{E}^{|1\cdots n\rangle}(p)$ and $\mathcal{E}_{\langle 1\ldots n|}(p)$, which are the unique (up to normalization) solutions of the equations

$$\begin{aligned}\rho_{\hat{R}(p),k}(g_i)\mathcal{E}^{|1\cdots n\rangle}(p) &= -\overline{q}\mathcal{E}^{|1\cdots n\rangle}(p),\\ \mathcal{E}_{\langle 1\ldots n|}(p)\rho_{\hat{R}(p),k}(g_i) &= \overline{q}\mathcal{E}_{\langle 1\ldots n|}(p),\end{aligned}\quad 1\leqslant i\leqslant n-1. \tag{5.2}$$

The only nonvanishing components of these $\mathcal{E}$-tensors have pairwise different indices $i_1, i_2, \ldots, i_n$ and look like

$$\mathcal{E}^{i_1 i_2\cdots i_n}(p) = (-1)^{l(\sigma)}\prod_{(j,i)\in J(\sigma)}\alpha_{ji}(p_{ji})\prod_{1\leqslant a<b\leqslant n}\xi_{i_a i_b}(p_{i_a i_b}), \tag{5.3}$$

$$\mathcal{E}_{i_1 i_2\cdots i_n}(p) = (-1)^{l(\sigma)}\prod_{(j,i)\in J(\sigma)}\alpha_{ij}(p_{ij}). \tag{5.4}$$

Here $l(\sigma)$ is the length of the permutation $\sigma = \binom{1,\ 2,\ \ldots,n}{i_1,i_2,\ldots,i_n}$, and

$$J(\sigma) = \{(i_a, i_b) : a<b, i_a>i_b\}.$$

The dynamical $\mathcal{E}$-tensors (5.3), (5.4) are normalized so that they would coincide with the constant ε-tensors (3.5) in the case (4.25), (4.26).

Now the expressions for rank 1q-antisymmetrizers in the representations $\rho_{\hat{R}(p),k}$ are given by

$$\rho_{\hat{R}(p),k}(A^{(i,n+i-1)}) = \frac{1}{[n]!}(X_1\cdots X_{i-1})\mathcal{E}^{|i\cdots n+i-1\rangle}(p)\mathcal{E}_{\langle i\cdots n+i-1|}(p)(X_1\cdots X_{i-1})^{-1}. \tag{5.5}$$

The numerical coefficient in this formula is calculated with the use of the relation

$$\mathcal{E}_{\langle 1\cdots n|}(p)\mathcal{E}^{|1\cdots n\rangle}(p) = [n]!, \tag{5.6}$$

which is proved in the Appendix.

We conclude the discussion on dynamical *R*-matrices by writing down formulas (2.12a), (2.12d), and (2.12e), (2.12f) for the representation $\rho_{\hat{R}(p),n+1}$:

$$\mathcal{E}_{\langle 1\cdots n|}(p)\,\rho_{\hat{R}(p),n+1}(g_n g_{n-1}\cdots g_1) = q\,K^{|n+1\rangle}{}_{\langle 1|}(p)\,X_1\mathcal{E}_{\langle 2\cdots n+1|}(p)X_1^{-1}, \tag{5.7}$$

$$X_1\mathcal{E}_{\langle 2\cdots n+1|}(p)X_1^{-1}\rho_{\hat{R}(p),n+1}(g_1 g_2\cdots g_n) = q\,N^{|1\rangle}{}_{\langle n+1|}(p)\,\mathcal{E}_{\langle 1\,\ldots\,n|}(p), \tag{5.8}$$

$$K(p)\,N(p) = N(p)\,K(p) = \mathbb{I}. \tag{5.9}$$

Here the matrices $N(p)$, $K(p)$ are defined as

$$N^{|1\rangle}{}_{\langle n+1|}(p) = \frac{(-1)^{n-1}}{[n-1]!}X_1\mathcal{E}_{\langle 2\cdots n+1|}(p)X_1^{-1}\mathcal{E}^{|1\cdots n\rangle}(p), \tag{5.10}$$

$$K^{|n+1\rangle}{}_{\langle 1|}(p) = \frac{(-1)^{n-1}}{[n-1]!}\mathcal{E}_{\langle 1\cdots n|}(p)X_1\mathcal{E}^{|2\cdots n+1\rangle}(p)X_1^{-1}. \tag{5.11}$$

For the SL(n)-type dynamical *R*-matrices the matrices $N(p)$, $K(p)$ are diagonal. Inserting formulas (5.3), (5.4) for the dynamical ε-tensors into (5.10), (5.11) and using (5.6), one ends up with the following expressions for their diagonal components:

$$N_i^{\,i}(p) = (K_i^{\,i}(p))^{-1} = \prod_{j \neq i} \alpha_{ij}(p_{ij} - \theta_{ji})\, \xi_{ij}(p_{ij}), \tag{5.12}$$

where $\theta_{ji} = \{1 \text{ if } j > i, 0 \text{ if } j < i\}$.

VI. QUANTUM MATRIX ALGEBRA $\mathcal{A}(\hat{R}(P),\hat{R})$: QUANTUM DETERMINANT AND INVERSION FORMULA

We shall apply the above technique to the quantum matrix algebra $\mathcal{A}$ which is defined as follows (cf. the Introduction).

Definition 6.1: Let $\mathbb{F}$ *be the field of the complex meromorphic functions of the (commuting) variables* $p_j, j = 1,\ldots,n$*. Let* $\hat{R}(p)$ *be a dynamical R-matrix of an* $\mathrm{SL}(n)$*-type and* $\hat{R}$ *be a constant* $\mathrm{SL}(m)$*-type R-matrix. Assume that both* $\hat{R}(p)$ *and* $\hat{R}$ *satisfy the Hecke condition (1.4) with the same value of q. Then* $\mathcal{A} = \mathcal{A}(\hat{R}(p),\hat{R})$ *is a complex algebra with* 1 *that is generated by* $\mathbb{F}$ *and the mn elements* a_α^i $(i = 1,\ldots,n$ *and* $\alpha = 1,\ldots,m)$*, satisfying the relations*

$$\hat{R}(p)_{12}\, a_1\, a_2 = a_1\, a_2\, \hat{R}_{12}, \tag{6.1}$$

$$a f(p) = X f(p) X^{-1} a, \quad \forall f(p) \in \mathbb{F}, \tag{6.2}$$

where X is a unimodular diagonal matrix (4.2) whose diagonal elements X^i *satisfy (4.1a), (4.1b).*

Remark 6.1: The definition above is given for arbitrary *m* and *n*. However in the sequel we shall discuss the case $m = n$ only.

Remark 6.2: For the applications envisaged here, the field $\mathbb{F}$ of meromorphic functions of p_j can be replaced by its subfield of rational functions of q^{p_j} (as stated in the Introduction). Then we should just require

$$X^k q^{p_{ij}} (X^k)^{-1} = q^{p_{ij} + \delta_{jk} - \delta_{ik}}$$

instead of (4.1b). Note that for q a root of unity, $q^{2h} = 1$ [cf. Eq. (7.8) below] p_{ij} are only determined up to an additive integer multiple of $2h$.

Remark 6.3: More general matrix algebras are of interest in which the *R*-matrices on both sides of the quadratic relations (6.1) are allowed to depend on possibly different sets of commuting variables p and p'

$$\hat{R}(p)_{12} Q_1 Q_2 = Q_1 Q_2 \hat{R}'(p')_{12}, \tag{6.3}$$

while the shift properties assume the form

$$p_{kl}\, Q_j^i = Q_j^i\, (p_{kl} + \delta_k^i - \delta_l^i), \quad p'_{kl}\, Q_j^i = Q_j^i\, (p'_{kl} + \delta_{jk} - \delta_{jl}). \tag{6.4}$$

Such $\mathcal{A}(\hat{R}(p),\hat{R}'(p'))$ can be treated in much the same way or reduced to the study of two matrix algebras of the above type setting $Q_j^i = a_\alpha^i \overline{a}_j^\alpha$, where a and $\overline{a}$ satisfy exchange relations of the type (1.6) and (1.10), respectively (see Ref. 24). Note that dynamical quantum groups (introduced in Ref. 16) are defined by relations similar to (6.3) and (6.4) but with the dynamical *R*-matrices (and momenta p,p') related to each other by some equivalence transformation $\hat{R}'(p') = X^{-1}\hat{R}(p)X$. Another desirable modification of the matrix algebra (6.1) corresponds to the case when $\hat{R}$ is an SO_q or Sp_q constant *R*-matrix. In this case $\hat{R}$ and $\hat{R}(p)$ satisfy a third order (Birman–Wenzl) condition instead of the Hecke property (2.2) and the QDYBE (4.3) have to be modified correspondingly.

Remark 6.4: The algebra $\mathcal{A}$ differs from the one considered in Ref. 33 where the counterpart of a matrix $b = X^{-1}a$ [denoted by $u_\alpha^i(\overline{\omega})$ in Eq. (11) of Ref. 33] which commutes with p is used for changing the basis of chiral vertex operators. It is assumed in Ref. 33 that the elements of b

only depend on p and hence commute among themselves while in our case this is not so. Indeed, the reflection equation subalgebra $\mathcal{M}(\hat{R})$, defined in Proposition 6.5 below, is noncommutative although its elements commute with the p's. The difference is essential: as a result, Cremmer and Gervais do not recover the standard (constant) $\mathrm{SL}_q(n) R$-matrix for $n>2$ but introduce instead new solutions of the Yang–Baxter equation. One of the authors (I.T.) would like to thank J.-L. Gervais and E. Cremmer for an enlightening discussion on this point.

The term *''matrix algebra''*[34] for the algebra $\mathcal{A} \equiv \mathcal{A}(\hat{R}(p), \hat{R})$ is justified by the fact that we shall be able to (define and) compute the determinant of a — as a function of p — and to find the inverse of a. In the case of 2×2 matrices the determinant of a was constructed in Ref. 4 (see also Ref. 36) for the special choice $\beta_i \to \infty, \alpha_{ij}=1$ of the parameters. We shall present the definition of the determinant in a general setting.

Definition 6.2: Let $a=||a^i_\alpha||$ *be the matrix of generators of the algebra* $\mathcal{A}(\hat{R}(p), \hat{R})$. *The determinant of the matrix a is given by*

$$\det(a) = \frac{1}{[n]!} \mathcal{E}_{\langle 1 \cdots n|}(p) a_1 a_2 \cdots a_n \varepsilon^{|1 \cdots n\rangle}. \tag{6.5}$$

The meaning of this definition is made clear by the following three Propositions. The first and the third of them are the quantum analogs of the basic determinant properties. The second one allows to perform an $\mathrm{SL}(n)$-reduction in the algebra $\mathcal{A}(\hat{R}(p), \hat{R})$.

Proposition 6.1: The product $(a_1 a_2 \cdots a_n)$ *intertwines between constant and dynamical* ε*-tensors:*

$$\mathcal{E}_{\langle 1 \cdots n|}(p) a_1 a_2 \cdots a_n = \det(a)\, \varepsilon_{\langle 1 \cdots n|}, \tag{6.6}$$

$$a_1 a_2 \cdots a_n \epsilon^{|1 \dots n\rangle} = \mathcal{E}^{|1 \cdots n\rangle}(p) \det(a). \tag{6.7}$$

Proof: First, observe that due to the relations (6.1), (6.2) the product of k matrices $(a_1 a_2 \cdots a_k)$ intertwines between the representations $\rho_{\hat{R},k}$ and $\rho_{\hat{R}(p),k}$ of the algebra $\mathcal{H}_k(q)$. Indeed,

$$\begin{aligned}(a_1 \cdots a_k) \rho_{\hat{R},k}(g_i) &= (a_1 \cdots a_k) \hat{R}_{i(i+1)} \\ &= a_1 \cdots a_{i-1} (\hat{R}_{i(i+1)}(p) a_i a_{i+1}) a_{i+2} \cdots a_k \\ &= (X_1 \cdots X_{i-1}) \hat{R}_{i(i+1)}(p) (X_1 \cdots X_{i-1})^{-1} (a_1 \cdots a_k) = \rho_{\hat{R}(p),k}(g_i)(a_1 \cdots a_k).\end{aligned} \tag{6.8}$$

In particular, one has

$$(a_1 \cdots a_n) \rho_{\hat{R},n}(A^{(n)}) = \rho_{\hat{R}(p),n}(A^{(n)})(a_1 \cdots a_n).$$

Multiplying both sides by $\rho_{\hat{R},n}(A^{(n)})$ from the right or by $\rho_{\hat{R}(p),n}(A^{(n)})$ from the left and using projector property of the q-antisymmetrizer one comes to the equations

$$\rho_{\hat{R}(p),n}(A^{(n)})(a_1 \cdots a_n) = \rho_{\hat{R}(p),n}(A^{(n)})(a_1 \cdots a_n) \rho_{\hat{R},n}(A^{(n)}), \tag{6.9}$$

$$(a_1 \cdots a_n) \rho_{\hat{R},n}(A^{(n)}) = \rho_{\hat{R}(p),n}(A^{(n)})(a_1 \cdots a_n) \rho_{\hat{R},n}(A^{(n)}). \tag{6.10}$$

Finally, expressing (3.6), (5.5) for constant and dynamical q-antisymmetrizers in terms of the ε-tensors, one transforms (6.9), (6.10) to the form (6.6), (6.7). ■

Proposition 6.2: The element $\det(a)$ *of the algebra* $\mathcal{A}(\hat{R}(p), \hat{R})$ *commutes with the generators* p_i *and its commutation with the generators* a^i_α *is described by*

$$\det(a)\,a = K(p)\,a\,\det(a), \tag{6.11}$$

where the diagonal matrix $K(p)$ is given in (5.11), (5.12).

Proof: Consider the permutation of $\det(a)$ with an arbitrary function $h(p)$:

$$\begin{aligned}\det(a)h(p) &= \mathcal{E}_{\langle 1\cdots n|}(p)a_1\cdots a_n\,h(p)\varepsilon^{|1\cdots n\rangle}/[n]!\\ &= \mathcal{E}_{\langle 1\cdots n|}(p)(X_n\cdots X_1)h(p)\\ &\quad\times(X_n\cdots X_1)^{-1}a_1\cdots a_n\,\varepsilon^{|1\cdots n\rangle}/[n]!.\end{aligned} \tag{6.12}$$

Since the only nonvanishing components of the tensor $\mathcal{E}_{\langle 1\cdots n|}(p)$ are those with pairwise different indices and due to the diagonal structure of the matrix X one has

$$\mathcal{E}_{\langle 1\cdots n|}(p)X_n\cdots X_1 = \mathcal{E}_{\langle 1\cdots n|}(p)\det(X) = \mathcal{E}_{\langle 1\cdots n|}(p),$$

where in the last equality the unimodularity of X [see (4.2)] is taken into account. Now we can complete the transformation of (6.12):

$$\begin{aligned}\det(a)h(p) &= \cdots\\ &= h(p)\mathcal{E}_{\langle 1\cdots n|}(p)(X_n\cdots X_1)^{-1}a_1\cdots a_n\,\varepsilon^{|1\cdots n\rangle}/[n]! = h(p)\det(a).\end{aligned}$$

This proves commutivity of $\det(a)$ and p_i.

Consider now permutation of $\det(a)$ with the matrix a. It is technically convenient to take a living in the matrix space with label $(n+1)$:

$$\begin{aligned}\det(a)\,a_{n+1} &= \mathcal{E}_{\langle 1\cdots n|}(p)\{a_1\cdots a_n\,a_{n+1}\}\varepsilon^{|1\cdots n\rangle}/[n]!\\ &= \{\mathcal{E}_{\langle 1\cdots n|}(p)\,\rho_{\hat{R}(p),n+1}(g_n\cdots g_1)\}a_1\cdots a_{n+1}\{\rho^{-1}_{\hat{R},n+1}(g_n\cdots g_1)\,\varepsilon^{|1\cdots n\rangle}\}/[n]!\\ &= K^{|n+1\rangle}{}_{\langle 1|}(p)\{X_1\mathcal{E}_{\langle 2\cdots n+1|}(p)X_1^{-1}\,a_1\}\{a_2\cdots a_{n+1}\,\varepsilon^{|2\cdots n+1\rangle}\}K^{-1|1\rangle}{}_{\langle n+1|}/[n]!\\ &= K^{|n+1\rangle}{}_{\langle 1|}(p)\,a_1\,\{\mathcal{E}_{\langle 2\cdots n+1|}(p)\mathcal{E}^{|2\cdots n+1\rangle}(p)\}\det(a)\,K^{-1|1\rangle}{}_{\langle n+1|}/[n]!\\ &= (K(p)aK^{-1})_{n+1}\det(a).\end{aligned}$$

The following formulas are used in the course of the calculation: (6.5) and (6.8) in the first line, (5.7) and (3.8) in the second line, (6.2) and (6.7) in the third line, and (5.6) in passing to the last line. For clarity we put into braces those expressions which are to be transformed in the next step.

Finally, substituting $\mathbb{I}$ for the matrix K [see (3.12)] we obtain (6.11). ■

Corollary 6.1: The element

$$\Delta = \det(a)\prod_{i<j}\frac{\varphi_{ij}(p_{ij})}{f(p_{ij})}, \tag{6.13}$$

where $f(p_{ij}) = \overline{q}^{p_{ij}} + [p_{ij}]\beta_{ij}$ and the functions φ_{ij} are defined by the relations

$$\alpha_{ij}(p_{ij}) = \frac{\varphi_{ij}(p_{ij}+1)}{\varphi_{ij}(p_{ij})} \tag{6.14}$$

belongs to the center of the algebra $\mathcal{A}(\hat{R}(p),\hat{R})$. The $\mathrm{SL}(n)$-reduction in the algebra $\mathcal{A}(\hat{R}(p),\hat{R})$ can be performed by imposing the condition $\Delta = 1$.

Proof: We shall search for the central element in $\mathcal{A}(\hat{R}(p),\hat{R})$in the form $\Delta = U(p)\det(a)$, where $U(p)$ is some function of p_i which is to be fixed. As follows from Proposition 6.2, the element Δ commutes with p_i and its commutivity with the generators a^i_α imposes the following conditions on the function U

$$X^i U(p)(X^i)^{-1} = U(p) K_i^i(p), \quad i = 1, \cdots, n. \tag{6.15}$$

Now using (4.1b), (4.13) and (5.12) it is straightforward to check that with the choice (6.13), (6.14) one satisfies conditions (6.15). ■

Proposition 6.3: Let the algebra $\mathcal{A}(\hat{R}(p), \hat{R})$ *be completed with the inverse determinant of a:* $(\det\ a)^{-1}\det(a) = \det(a)(\det a)^{-1} = 1$. *Then the left and right inverse of a is given by*

$$(a^{-1})^{|1\rangle}{}_{\langle n+1|} = \frac{(-1)^{n-1}}{[n-1]!} (\det a)^{-1}\, \mathcal{E}_{\langle 2 \cdots n+1|}(p)\, a_2 \cdots a_n\, \varepsilon^{|1 \cdots n\rangle}. \tag{6.16}$$

Proof: We first check that the expression (6.16) is a left inverse of a:

$$\begin{aligned}(a^{-1})^{|1\rangle}{}_{\langle n+1|}\, a_{n+1} &= \frac{(-1)^{n-1}}{[n-1]!} (\det a)^{-1} \{\mathcal{E}_{\langle 2 \cdots n+1|}(p) a_2 \cdots a_n\, a_{n+1}\} \varepsilon^{|1 \cdots n\rangle} \\ &= \frac{(-1)^{n-1}}{[n-1]!} (\det a)^{-1} \det(a)\, \varepsilon_{\langle 2 \cdots n+1|} \varepsilon^{|1 \cdots n\rangle} = N^{|1\rangle}{}_{\langle n+1|} = \mathbb{I}^{|1\rangle}{}_{\langle n+1|}.\end{aligned}$$

Here we have used successively Eqs. (6.6), (3.10) and (3.12).

Checking that (6.16) is also a right inverse is slightly more complicated:

$$\begin{aligned} a_1\, (a^{-1})^{|1\rangle}{}_{\langle n+1|} &= \frac{(-1)^{n-1}}{[n-1]!} \{a_1\, (\det a)^{-1}\} \mathcal{E}_{\langle 2 \cdots n+1|}(p)\, a_2 \cdots a_n\, \varepsilon^{|1 \cdots n\rangle} \\ &= \frac{(-1)^{n-1}}{[n-1]!} (\det a)^{-1} K_1(p) \{a_1\, \mathcal{E}_{\langle 2 \cdots n+1|}(p)\} a_2 \cdots a_n\, \varepsilon^{|1 \cdots n\rangle} \\ &= \frac{(-1)^{n-1}}{[n-1]!} (\det a)^{-1} K_1(p)\, X_1\, \mathcal{E}_{\langle 2 \cdots n+1|}(p) X_1^{-1} \{a_1 \cdots a_n\, \varepsilon^{|1 \cdots n\rangle} \\ &= (\det a)^{-1} K_1(p)\, N^{|1\rangle}{}_{\langle n+1|}(p)\, \det(a) = \mathbb{I}^{|1\rangle}{}_{\langle n+1|},\end{aligned}$$

where we have applied successively Eqs. (6.11), (6.2), (6.7), (5.10) and (5.9). ■

The existence of inverse matrix a^{-1} is needed in many applications of the algebra $\mathcal{A}(\hat{R}(p), \hat{R})$. As an example of such application we shall construct a realization of a reflection equation algebra $\mathcal{M}(\hat{R})$ (for definition of this algebra see, e.g., Ref. 37, and references therein) in terms of the generators of $\mathcal{A}(\hat{R}(p), \hat{R})$. We have to use here the following general property of $\mathrm{SL}(n)$-type dynamical R-matrices (which has been noticed in Ref. 5 for the $\mathrm{SL}(2)$ case, see also Refs. 4,14).

Proposition 6.4: The dynamical matrix $\hat{R}(p)$ *(4.6), (4.12), (4.13) satisfies the equation*

$$D_1\, \hat{R}(p)\, D_2^{-1} = \hat{R}(p)^{-1}\, \sigma_{12}, \tag{6.17}$$

where the diagonal matrices D and σ

$$D_j^i \equiv q^{d_i} \delta_j^i, \quad (\sigma_{12})_{j_1 j_2}^{i_1 i_2} = \delta_{j_1}^{i_1} \delta_{j_2}^{i_2}\, \sigma_{i_1 i_2} \tag{6.18}$$

are fixed by (6.17) as

$$q^{d_i - d_j} = q^{-2 p_{ij}}\, \pi_{ij} \quad (i \neq j), \tag{6.19}$$

$$\sigma_{ij} = q^{2\delta_{ij}} \tag{6.20}$$

so that d_i *are functions of* p.

Proof: First of all we note that from the Hecke condition (4.4) [and (4.7)] one can deduce

$$\hat{R}(p)^{-1}=(a_{i_1 i_2}(p)-(q-\overline{q})\,\delta_{i_1 i_2})\,\delta^{i_1}_{j_2}\delta^{i_2}_{j_1}-b_{i_2 i_1}(p)\,\delta^{i_1}_{j_1}\delta^{i_2}_{j_2}. \tag{6.21}$$

Substitution of (4.6), (6.18) and (6.21) into (6.17) gives the following equations for the parameters σ_{ij} and d_i:

$$a_{ij}=(a_{ij}-(q-\overline{q})\,\delta_{ij})\,\sigma_{ji}, \tag{6.22}$$

$$q^{d_i-d_j}\,b_{ij}=-b_{ji}\,\sigma_{ij}. \tag{6.23}$$

Equation (6.22) leads to (6.20) while (6.23) is equivalent [in view of (A7)] to (6.19). ■

Now we construct the matrix $M^{\alpha}{}_{\beta}$ which is diagonalized with the help of the matrix a^i_α and the spectrum of which is defined by the matrix D (6.18), (6.19)

$$M=a^{-1}\,D\,a. \tag{6.24}$$

It is clear that $[D_1,D_2]=0$ and therefore the spectrum of the matrix M gives a commutative set of elements.

Proposition 6.5: The elements of the matrix M (6.24) *satisfy a reflection equation of the form*

$$M_2\,\hat{R}^{-1}\,M_2\,\hat{R}^{-1}=\hat{R}^{-1}\,M_2\,\hat{R}^{-1}\,M_2, \tag{6.25}$$

and thus provide a realization of a reflection equation subalgebra $\mathcal{M}(\hat{R})$ *in* $\mathcal{A}(\hat{R}(p),\hat{R})$. *The matrix elements of M satisfy the following exchange relations with the generators of* $\mathcal{A}(\hat{R}(p),\hat{R})$:

$$[D_2,M_1]=0, \quad M_1\,a_2=q^{2/n}\,a_2\,\hat{R}^{-1}\,M_2\,\hat{R}^{-1}. \tag{6.26}$$

Proof: Using (6.2), one can bring the commutation relations of the matrix D with the elements a^i_α to the form

$$a_1\,D_2=q^{-2/n}\,\sigma_{12}\,D_2\,a_1, \tag{6.27}$$

where the diagonal matrix σ_{12} is given by (6.18), (6.20). Equations (6.24) and (6.27) imply $[D_2,M_1]=0$. Then one proves (6.25) and the second relation in (6.26) by using (6.24), (6.1), (6.27) and (6.17). ■

VII. APPLICATION TO THE SU(n) WZNW MODEL

As an application of quantum matrix algebras we briefly describe here a typical problem of the two dimensional conformal field theory in which such matrices arise (see Ref. 24 for more details).

Let G be a connected compact Lie group and $g=g(t,x)$ be a map from the cylinder $\mathbb{R}\times\mathbb{S}^1$ into G which satisfies the Wess–Zumino–Novikov–Witten (WZNW) equations of motion. The general periodic solution $g(t,x)=g(t,x+2\pi)$ of these equations factorizes into a product of group valued chiral fields

$$g^A_B(t,x)=u^A_\alpha(x-t)\,\overline{u}^\alpha_B(x+t) \quad (\text{classically, } g,u,\overline{u}\in G), \tag{7.1}$$

each of which satisfies a twisted periodicity condition; in particular,

$$u(x+2\pi)=u(x)M \quad (M\in G), \tag{7.2}$$

where M is the monodromy.

Furthermore, the quantum chiral fields obey quadratic exchange relations[5,6,10,11,20,21,23,38,39]

$$u(y)_2 u(x)_1 = u(x)_1 u(y)_2 R(x-y) \Leftrightarrow P u(y)_2 u(x)_1 = u(x)_2 u(y)_1 \hat{R}(x-y) . \tag{7.3}$$

Here the matrix $R(x)$ is a solution of the quantum Yang–Baxter equation whose x-dependence is given by a step function, while $\hat{R}(x)$ is the associated braid operator:

$$\hat{R}(x) = \hat{R}\ \theta(x) + \hat{R}^{-1}\ \theta(-x) , \quad \hat{R}(x) = P\ R(x) = \hat{R}^{\varepsilon(x)} \tag{7.4}$$

$[\varepsilon(x) = \theta(x) - \theta(-x)]$.

Since $\hat{R}$ enters Eq. (7.3) in pair with P it should be normalized to have determinant $\det \hat{R} = \det P$. For $G = \mathrm{SU}(n)$ this implies the relation

$$\hat{R}_{i i+1} = \overline{q}^{1/n}\, \rho(g_i) \quad (\text{for } g_i^2 = \mathbb{I} + (q - \overline{q}) g_i) \Rightarrow \det \hat{R} = \det P = (-1)^{\binom{n}{2}} \tag{7.5}$$

so that we have to renormalize $\hat{R}$ of (3.1) by multiplying it by $\overline{q}^{1/n}$. (The resulting $\hat{R}$ has eigenvalues $q^{1-1/n}$ and $-\overline{q}^{1+1/n}$ of multiplicities $\binom{n+1}{2}$ and $\binom{n}{2}$, respectively; thus the product of all n^2 eigenvalues of $\hat{R}$ is indeed $(-1)^{\binom{n}{2}}$.)

We expand, following Refs. 21 and 4, $u(x)$ into a basis of zero modes that diagonalizes the monodromy matrix M at least for ''physical weights'' (satisfying $p_{1n} < h$):

$$u_\alpha^A(x) = a^i_\alpha\, u_i^A(x,p) , \quad a\, M = D\, a , \quad D^i_j = q^{d_i} \delta^i_j . \tag{7.6}$$

Here $d_i = -2p_i - 1/n + 1$, $p = \{p_i\}$ are central elements of the reflection equation algebra $\mathcal{M}(\hat{R})$; in the quantum field theoretic representation at hand they form a commuting set of operators such that Eq. (1.7) takes place. The eigenvalues of the differences $p_{i i+1} (= p_i - p_{i+1})$ are natural numbers that can be identified with the extended weights, $\lambda_i + 1$ labeling the (finite dimensional) irreducible representations of $\mathrm{SU}(n)$. The labels of the $\binom{n}{j}$ dimensional fundamental representation are given by $\lambda_i^{(j)} = \delta_i^j$, $1 \leqslant i,\ j \leqslant n-1$. Under these assumptions Eq. (7.3) implies exchange relations of the type

$$\hat{R}(p)\ a_1\, a_2 = a_1\, a_2\, \hat{R}_{21} , \tag{7.7}$$

where $\hat{R}(p)$ obeys the QDYBE (1.3). Hence, the results displayed in Secs. IV and V can be applied with slight modifications. (Since $\hat{R}(p)$ and $\hat{R}_{21}$ enter (7.7) homogeneously, the factor $\overline{q}^{1/n\rangle}$ of (7.5) cancels in the two sides.) Thus we can also apply the results of Sec. VI to the (chiral zero mode) quantum matrix algebra $\mathcal{A}$ of the $\mathrm{SU}(n)$ WZNW model. It should be noted that in this case q is a root of -1 associated with the level k $\widehat{su}(n)$ Kac–Moody algebra:

$$q = e^{i(\pi/h)} , \quad [2] = q + \overline{q} = 2 \cos \frac{\pi}{h} , \quad h = n + k (\geqslant n+1) . \tag{7.8}$$

The eigenvalues q^{d_i} of the diagonal matrix D can be expressed as differences of conformal dimensions. Indeed, according to Ref. 21, the chiral vertex operators $u_j(x,p)$ satisfy

$$u_j(x + 2\pi, p) = u_j(x,p)\ e^{2\pi i (\Delta_h(p) - \Delta_h(p + v^{(j)}))} , \tag{7.9}$$

where the matrices $v^{(j)}$ and p are defined by (1.1) and (1.2). Here the conformal dimensions are expressed in terms of the $\mathrm{SU}(n)$ Casimir operator,

$$2\, h\, \Delta_h(p) = C_2(p) = \frac{1}{n} \sum_{1 \leqslant i < k \leqslant n} p_{ik}^2 - \frac{n(n^2-1)}{12} , \tag{7.10}$$

so that

$$d_j = C_2(p) - C_2(p+v^{(j)}) = -2(p|v^{(j)}) - |v^{(j)}|^2 = (1/n) - 1 - 2p_j . \tag{7.11}$$

Inserting this in (6.19), we deduce that $\pi_{ij}=1$ so that we arrive at the special solution (4.24) for $\hat{R}(p)$, allowing us to present (7.7) in the form

$$\alpha_{ij}(p)[p_{ij}-1]\, a^j_\alpha a^i_\beta = [p_{ij}]\, a^i_\beta a^j_\alpha - q^{\epsilon_{\beta\alpha} p_{ij}} a^i_\alpha a^j_\beta \tag{7.12}$$

(here $\epsilon_{\alpha\beta}$ is equal 1 for $\alpha>\beta$, 0 for $\alpha=\beta$ and -1 for $\alpha<\beta$). According to the analysis of Sec. IV we can reduce (7.12) to the case $\alpha_{ij}(p)=1$ by a suitable twist.

An important consequence of (7.8) and (7.12) is the existence of an ideal $\mathcal{I}_h$ of $\mathcal{A}$ generated by n^2 elements $(a^i_\alpha)^h$ such that the factor algebra $\mathcal{A}/\mathcal{I}_h$ is finite dimensional.[24] This allows to define a finite dimensional ''Fock space representation'' of $\mathcal{A}$ with a unique vacuum vector $|\mathrm{vac}\rangle$ corresponding to the trivial su(n) weight $\lambda_i=0$ ($p_{i\,i+1}=1$, $i=1,\ldots,n-1$) such that

$$a^i_\alpha |\mathrm{vac}\rangle = 0 \quad \text{for } i>1 , \quad \mathcal{I}_h |\mathrm{vac}\rangle = 0 . \tag{7.13}$$

ACKNOWLEDGMENTS

This work started while three of us, L.K.H., O.V.O. and I.T.T. were visiting the Bogoliubov Laboratory of Theoretical Physics of the JINR, Dubna and was completed while A.P.I., L.K.H. and I.T.T. were visiting the Department of Physics of the University of Pisa, ICTP (Trieste) and the Erwin Schrödinger Institute for Mathematical Physics, respectively. We thank all these Institutions for hospitality and support. This work was also supported in part by the RFBR (Grant No. 97-01-01041), INTAS (Grant No. 93-127-ext), by the exchange program between INFN and JINR (Dubna) and by the Bulgarian National Foundation for Scientific Research under Contract No. F-404.

APPENDIX: NORMALIZATION OF DYNAMICAL LEVI-CIVITÀ TENSORS

The definitions (5.3), (5.4) lead to the expression

$$\mathcal{E}_{i_1\cdots i_n}(p)\, \mathcal{E}^{i_1\cdots i_n}(p) = \prod_{1\leqslant a<b\leqslant n} \xi_{i_a i_b} \tag{A1}$$

(there are no summations over the indices i_k), and the normalization condition (5.6) for the dynamical $\mathcal{E}$-tensors follows from

Proposition A: Let $\xi_{ij}=d-b_{ij}$*, where d is a constant (comparing with (4.12), one gets* $d=q$*) and the elements* b_{ij} *satisfy (4.7), (4.10). Then the following identity holds:*

$$I_k \equiv \sum_{S_k} \prod_{1\leqslant a<b\leqslant k} \xi_{i_a i_b} = [k]_d! , \tag{A2}$$

where $k\leqslant n$, $[k]_d = (d^k-(d-\lambda)^k)/\lambda$ $(\lambda = q-\overline{q})$ and S_k denotes all permutations of indices $(i_1,\ldots,i_k)$ $(i_a\neq i_b$ for $a\neq b)$. Note that $[k]_d=[k]$ for $d=q$ [as it is needed in (5.6)].

Proof: We shall proceed by induction. For $k=2$ we have

$$I_2 = \xi_{i_1 i_2} + \xi_{i_2 i_1} = 2\,d - \lambda = [2]_d .$$

Let (A2) be correct for some $k(1<k<n)$, then for $k+1$ we derive

$$I_{k+1}=\sum_{S_{k+1}}\left[\left(\prod_{l=1}^{k}\xi_{i_l i_{k+1}}\right)\prod_{1\leqslant a<b\leqslant k}\xi_{i_a i_b}\right]$$
$$=\sum_{r=1}^{k+1}\left[\left(\prod_{\substack{l=1\\ l\neq r}}^{k+1}\xi_{i_l i_r}\right)\sum_{S_k}\prod_{\substack{a\neq r\neq b\\ 1\leqslant a<b\leqslant k+1}}\xi_{i_a i_b}\right]$$
$$=I_k\sum_{r=1}^{k+1}\prod_{\substack{l=1\\ l\neq r}}^{k+1}\xi_{i_l i_r}.$$

Therefore we should prove the identity

$$[k+1]_d=\sum_{r=1}^{k+1}\prod_{\substack{l=1\\ l\neq r}}^{k+1}\xi_{i_l i_r}=\sum_{r=1}^{k+1}\prod_{\substack{l=1\\ l\neq r}}^{k+1}(d-b_{i_l i_r}). \tag{A3}$$

This identity follows from the relation

$$\sum_{r=1}^{m}\prod_{\substack{l=1\\ l\neq r}}^{m}b_{i_l i_r}=\lambda^{m-1}(m\leqslant k+1), \tag{A4}$$

which can be obtained by induction. Indeed, from Eqs. (4.7), (4.10) we have for $m=2,3$,

$$b_{i_1 i_2}+b_{i_2 i_1}=\lambda,\quad b_{i_2 i_1}b_{i_3 i_1}+b_{i_1 i_2}b_{i_3 i_2}+b_{i_1 i_3}b_{i_2 i_3}=\lambda^2.$$

Then we deduce

$$\prod_{l=2}^{m}b_{i_l i_1}=b_{i_2 i_1}\left(\lambda^{m-2}-\sum_{r=3}^{m}(\lambda-b_{i_r i_1})\prod_{\substack{l=3\\ l\neq r}}b_{i_l i_r}\right)$$
$$=b_{i_2 i_1}\sum_{r=3}^{m}b_{i_r i_1}\prod_{\substack{l=3\\ l\neq r}}b_{i_l i_r}$$
$$=\sum_{r=3}^{m}(\lambda^2-b_{i_1 i_2}b_{i_r i_2}-b_{i_1 i_r}b_{i_2 i_r})\prod_{\substack{l=3\\ l\neq r}}b_{i_l i_r}$$
$$=\lambda^{m-1}-\sum_{r=2}^{m}\prod_{\substack{l=1\\ l\neq r}}b_{i_l i_r},$$

which proves (A4). Expanding the right hand side of (A3) in power series of d and taking into account (A4) we verify the relations (A3) and, thus, complete the proof. ■

One can reformulate the statement of Proposition A in more concise form (only in terms of elements ξ_{ij}).

Proposition B: Let ξ_{ij} *satisfy*

$$\xi_{ij}+\xi_{ji}=[2]=\xi_{ij}\,\xi_{jk}\,\xi_{ki}+\xi_{ik}\,\xi_{kj}\,\xi_{ji}(i\neq j\neq k\neq i).$$

We rewrite these conditions as

$$\sum_{r=1}^{k} \prod_{\substack{l=1 \\ l\neq r}}^{k} \xi_{i_l i_r} = [k] \quad \text{for } k=2,3. \tag{A5}$$

Then, Eq. (A5) is also valid for $4 \leqslant k \leqslant n$, *and the following identity holds:*

$$I_k \equiv \sum_{S_k} \prod_{1 \leqslant a < b \leqslant k} \xi_{i_a i_b} = [k]!, \tag{A6}$$

where S_k *denotes all permutations of the indices* $(i_1, \cdots, i_k)$ *and* $i_a \neq i_b$ *for* $a \neq b$.

Proof: The proof is similar to that of Proposition A. ■

Remark: There are many other interesting relations among the elements b_{ij} (4.7), (4.10) (as well as among ξ_{ij}). For example, one can easily deduce the identity

$$b_{i_1 i_2} b_{i_2 i_3} \cdots b_{i_{k-1} i_k} b_{i_k i_1} = (-1)^k b_{i_1 i_k} b_{i_k i_{k-1}} \cdots b_{i_3 i_2} b_{i_2 i_1},$$

which generalizes (4.10) and follows from the relation

$$-\frac{b_{ji}(p)}{b_{ij}(p)} q^{2p_{ij}} = \pi_{ij}. \tag{A7}$$

Note that we consider π_{ij} as constants which are independent of p_i.

[1] L. D. Faddeev, N. Yu. Reshetikhin, and L. A. Takhtajan, ''Quantized Lie groups and Lie algebras,'' Algebra i Analiz **1**, 178 (1989) [English translation, Leningrad Math. J. **1**, 193 (1990)].
[2] J.-L. Gervais and A. Neveu, Nucl. Phys. B **238**, 125 (1984).
[3] E. Sklyanin, ''Dynamical *r*-matrices for elliptic Calogero–Moser model,'' Algebra i Analiz **6**, 227 (1994) [English translation, St. Petersburg Math. J. **6**, 397 (1995)] hep-th/9308060.
[4] A. G. Bytsko and L. D. Faddeev, J. Math. Phys. **37**, 6324 (1996), q-alg/9508022.
[5] A. Yu. Alekseev and L. D. Faddeev, Commun. Math. Phys. **141**, 413 (1991).
[6] L. D. Faddeev, Commun. Math. Phys. **132**, 131 (1990).
[7] A. N. Kirillov and N. Yu. Reshetikhin, ''Representations of the algebra $U_q(sl(2))$, *q*-orthogonal polynomials and invariants of links,'' in *Infinite Dimensional Lie Algebras and Groups, Proceedings of the 1988 CIRM Conference*, edited by V. Kac (World Scientific, Singapore, 1989), pp. 285–339.
[8] O. Babelon, D. Bernard, and E. Billey, Phys. Lett. B **375**, 89 (1996), q-alg/9511019; O. Babelon, Commun. Math. Phys. **139**, 619 (1991).
[9] E. Witten, Commun. Math. Phys. **92**, 455 (1984).
[10] A. Alekseev and S. Shatashvili, Commun. Math. Phys. **133**, 353 (1990).
[11] L. D. Faddeev, Gargèse lectures 1991, in *New Symmetry Principles in Quantum Field Theory*, edited by J. Fröhlich *et al.* (Plenum, New York, 1992), pp. 159–175.
[12] J.-M. Maillet, Phys. Lett. **162B**, 137 (1985).
[13] A. Yu. Alekseev and I. T. Todorov, Nucl. Phys. B **421**, 413 (1994).
[14] A. G. Bytsko, ''Fusion of *q*-tensor operators: Quasi-Hopf-algebraic point of view,'' Zapiski Seminarov POMI **245**, 107 (1997), q-alg/9609007.
[15] A. G. Bytsko and V. Schomerus, ''Vertex operators: From a toy model to lattice algebras,'' Commun. Math. Phys. **191,** 87 (1998)
[16] G. Felder, ''Elliptic quantum groups,'' in *Proceedings of the ICMP*, Paris, 1994, edited by D. Iagolnitzer (International, Boston, 1995), pp. 211–218, hep-th/9412207; ''Conformal field theory and integrable systems associated with elliptic curves,'' in *Proceedings of the ICM, Birkhäuser, Zürich*, 1994, hep-th/9407154.
[17] J. Avan, O. Babelon, and E. Billey, Commun. Math. Phys. **178**, 281 (1996); G. Arutyunov, L. Chehov, and S. Frolov, ''*R*-matrix quantization of the elliptic Ruijsenaars–Schneider model,'' Steklov Mathematical Institute preprint SM1-96-59 (1996), q-alg/9612032; G. Felder and A. Varchenko, ''Elliptic quantum groups and Ruijsenaars models,'' q-alg/9704005.
[18] A. P. Isaev, J. Phys. A **29**, 6903 (1996).
[19] P. Etingof and A. Varchenko, ''Solutions of the quantum dynamical Yang–Baxter equation and dynamical quantum groups,'' q-alg/9708015.
[20] P. Furlan, L. K. Hadjiivanov, and I. T. Todorov, ''Canonical approach to the quantum WZNW model,'' ICTP Trieste and ESI Vienna preprint IC/95/74, ESI 234, 1995.
[21] P. Furlan, L. K. Hadjiivanov, and I. T. Todorov, Nucl. Phys. B **474**, 497 (1996), hep-th/9602101; P. Furlan, L. K. Hadjiivanov, and I. T. Todorov, Int. J. Mod. Phys. A **12**, 23 (1997), hep-th/9610202.
[22] M. Dubois-Violette and I. T. Todorov, Lett. Math. Phys. **42**, 183 (1997), hep-th/9704069.
[23] K. Gawędzki, Commun. Math. Phys. **139**, 201 (1991); F. Falceto and K. Gawędzki, J. Geom. Phys. **11**, 251 (1993).

[24] P. Furlan, L. K. Hadjiivanov, A. P. Isaev, P. N. Pyatov, and I. T. Todorov (''Quantum matrix algebra for the SL(n) WZNW model,'' in preparation).
[25] D. I. Gurevich, ''Algebraic aspects of the quantum Yang–Baxter equation,'' Algebra i Analiz **2**, 119 (1990) [English translation, Leningrad Math. J. **2**, 801 (1991)].
[26] P. Martin, *Potts Models and Related Problems in Statistical Mechanics* (World Scientific, Singapore, 1991).
[27] In general, this need not be the case. Examples with height $\rho \neq \dim V$ were constructed in Ref. 25.
[28] V. G. Drinfeld, ''Quantum groups,'' in *Proceedings of the International Congress of Mathematicians,* Berkeley, 1986 (Academic, New York, 1986), Vol. 1, pp. 798–820.
[29] M. Jimbo, Lett. Math. Phys. **11**, 247 (1986).
[30] In Ref. 19 these R-matrices were called GL(n)-type R-matrices. We call them SL(n)-type R-matrices instead since we impose the additional condition $\Sigma_{i=1}^{n} p_i = 0$ [see (1.2)] on the variables p_i.
[31] V. G. Drinfeld, ''Quasi Hopf algebras,'' Algebra i Analiz **1**, 141 (1989) [English translation, Leningrad Math. J. **1**, 1419 (1990)]; N. Yu. Reshetikhin, Lett. Math. Phys. **20**, 331 (1990).
[32] Yu. I. Manin ''Quantum groups and noncommutative geometry,'' Université de Montréal preprint CRM-1561, 1989; Commun. Math. Phys. **122**, 163 (1989).
[33] E. Cremmer and J.-L. Gervais, Commun. Math. Phys. **134**, 619 (1990); A. Bilal and J.-L. Gervais, Nucl. Phys. B **318**, 579 (1989).
[34] More conventional quantum matrix algebras have been introduced in Refs. 1,32 and their matrix nature were further investigated in Refs. 25,35.
[35] P. N. Pyatov and P. A. Saponov, J. Phys. A **28**, 4415 (1995); D. I. Gurevich, P. N. Pyatov, and P. A. Saponov, Lett. Math. Phys. **41**, 255 (1997); A. P. Isaev, O. V. Ogievetsky, P. N. Pyatov, and P. A. Saponov, ''Characteristic polynomials and Newton identities for quantum matrices,'' preprint CPT-97/P3471, 1997.
[36] G. Felder and A. Varchenko, Commun. Math. Phys. **181**, 741 (1996).
[37] P. P. Kulish and R. Sasaki, Prog. Theor. Phys. **89**, 741 (1993).
[38] O. Babelon, Phys. Lett. B **215**, 523 (1988); B. Blok, *ibid.* **233**, 359 (1989).
[39] M. Chu, P. Goddard, I. Halliday, D. Olive, and A. Schwimmer, Phys. Lett. B **266**, 71 (1991); M. Chu and P. Goddard, Nucl. Phys. B **445**, 145 (1995).

Lie subalgebras of real and complex orthogonal groups in dimension four

Federico G. Lastaria[a)]
Dipartimento di Matematica, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

(Received 29 April 1998; accepted for publication 3 August 1998)

We give an explicit description of all Lie subalgebras of the Lie algebras $\mathfrak{so}_4\mathbb{R}$ and $\mathfrak{so}_4\mathbb{C}$ of the orthogonal groups $O(4,\mathbb{R})$ and $O(4,\mathbb{C})$. [S0022-2488(99)00501-0]

I. INTRODUCTION

In view of applications to mathematical physics as well as to differential geometry, we find it worthy to give a description, as explicit as possible, of the Lie subalgebras of the orthogonal Lie algebras $\mathfrak{so}_4\mathbb{R}$ and $\mathfrak{so}_4\mathbb{C}$.

Our approach is elementary, in the sense that we do not rely on Dynkin diagrams or general structure theorems: only some linear algebra and elementary facts about Borel subalgebras are needed.

II. LIE SUBALGEBRAS OF $\mathfrak{so}_4\mathbb{R}$

Let $\mathbb{H}$ be the skew field of quaternions and let $S^3 \subset \mathbb{H}$ be the Lie group of unit quaternions, isomorphic to the three-dimensional sphere. For each $(p,q) \in S^3 \times S^3$, let $\varphi_{p,q}$ be the $\mathbb{R}$-linear map defined by

$$\varphi_{p,q}: \quad \mathbb{H} \to \mathbb{H},$$

$$x \mapsto pxq^{-1},$$

for every $x \in \mathbb{H}$. Since $\|\varphi_{p,q}x\| = \|p\|\|x\|\|q^{-1}\| = \|x\|$, for all $p,q \in S^3$ the map $\phi_{p,q}$ belongs to $O(4,\mathbb{R})$, where we identify $\mathbb{R}^4$ with $\mathbb{H}$. Let φ be the map

$$\varphi: \quad S^3 \times S^3 \to O(4,\mathbb{R}),$$

$$(p,q) \mapsto \phi_{p,q}.$$

Since φ is continuous and $S^3 \times S^3$ is connected, the image $\varphi(S^3 \times S^3)$ of φ is in fact contained in the connected component $SO(4,\mathbb{R})$. Then we have a commutative diagram of Lie group homomorphisms:

$$\begin{array}{ccc} S^3 & \overset{\Delta}{\hookrightarrow} & S^3 \times S^3 \\ \Big\downarrow \psi & & \Big\downarrow \varphi \\ SO(3,\mathbb{R}) & \hookrightarrow & SO(4,\mathbb{R}) \end{array}$$

where Δ is the diagonal map of the Lie group S^3 and the inclusion of $SO(3,\mathbb{R})$ in $SO(4,\mathbb{R})$ is induced by the identification of $\mathbb{R}^3$ with the subspace of pure quaternions. Then the Lie group S^3,

[a)]Electronic mail: fedlas@mate.polimi.it readily seen; is the universal covering of $SO(3,\mathbb{R})$ and the Lie group $S^3\times S^3$ is the universal covering of $SO(4,\mathbb{R})$ (see, e.g., Ref. 1). As a consequence, *we have a Lie algebra isomorphism*:

$$\mathfrak{so}_4\mathbb{R}\simeq\mathfrak{so}_3\mathbb{R}\oplus\mathfrak{so}_3\mathbb{R}. \tag{II.1}$$

Henceforth, denote by $\mathfrak{so}_3E$ and $\mathfrak{so}_3F$ the first and the second summand of the direct sum (II.1):

$$\mathfrak{so}_4\mathbb{R}\simeq\mathfrak{so}_3E\oplus\mathfrak{so}_3F. \tag{II.2}$$

Given any subalgebra $\mathfrak{g}\subset\mathfrak{so}_4\mathbb{R}\simeq\mathfrak{so}_3E\oplus\mathfrak{so}_3F$, choose a linear subspace $\mathfrak{g}'$ (in general, not uniquely determined) in such a way that

$$\mathfrak{g}=(\mathfrak{g}\cap\mathfrak{so}_3E)\oplus(\mathfrak{g}\cap\mathfrak{so}_3F)\oplus\mathfrak{g}',$$

as a direct sum of vector spaces. Let

$$pr_E:\mathfrak{so}_3E\oplus\mathfrak{so}_3F\to\mathfrak{so}_3E,\quad pr_F:\mathfrak{so}_3E\oplus\mathfrak{so}_3F\to\mathfrak{so}_3F$$

be the projections, and denote by

$$\bar{\mathfrak{g}}_E=pr_E(\mathfrak{g}),\quad \bar{\mathfrak{g}}_F=pr_F(\mathfrak{g})$$

the projections of the Lie algebra $\mathfrak{g}$ on $\mathfrak{sl}E$ and $\mathfrak{sl}F$, respectively. It is easy to prove that $\mathfrak{g}\cap\mathfrak{so}_3E$ and $\mathfrak{g}\cap\mathfrak{so}_3F$ are ideal of the Lie algebras $\bar{\mathfrak{g}}_E$ and $\bar{\mathfrak{g}}_F$, respectively.

From the linear isomorphisms,

$$\frac{\mathfrak{g}}{\mathfrak{g}\cap\mathfrak{so}_3F}\simeq\bar{\mathfrak{g}}_E;\quad \frac{\mathfrak{g}}{\mathfrak{g}\cap\mathfrak{so}_3F}\simeq(\mathfrak{g}\cap\mathfrak{so}_3E)\oplus\mathfrak{g}',$$

we see that

$$\bar{\mathfrak{g}}_E\simeq(\mathfrak{g}\cap\mathfrak{so}_3E)\oplus\mathfrak{g}',$$

as vector spaces. At the same way, we see, of course, that

$$\bar{\mathfrak{g}}_F\simeq(\mathfrak{g}\cap\mathfrak{so}_3F)\oplus\mathfrak{g}'.$$

Since $\dim\mathfrak{g}'\leqslant 3$, *a priori* the following cases may occur:

(a) $\dim\mathfrak{g}'=0$.

Then $\mathfrak{g}=(\mathfrak{g}\cap\mathfrak{so}_3E)\oplus(\mathfrak{g}\cap\mathfrak{so}_3F)$ is direct sum of Lie subalgebras of $\mathfrak{so}_3E$ and $\mathfrak{so}_3F$;

(b) $\dim\mathfrak{g}'=1$.

Since there are no two-dimensional Lie subalgebras in $\mathfrak{so}_3\mathbb{R}$ (because $\mathfrak{so}_3\mathbb{R}$ is isomorphic to the Lie algebra $(\mathbb{R}^3,\wedge)$, where $\wedge$ is the ordinary vector product), we have $\mathfrak{g}\cap\mathfrak{so}_3E=\mathfrak{g}\cap\mathfrak{so}_3F=0$ and therefore $\mathfrak{g}=\mathfrak{g}'$ is one-dimensional;

(c) $\dim\mathfrak{g}'=2$.

This would imply that $\dim\bar{\mathfrak{g}}_E=2$ or 3. Either cases are impossible, because on the one hand there are no two-dimensional subalgebras of $\mathfrak{so}_3E$ and, on the other, if $\dim\bar{\mathfrak{g}}_E$ were equal to 3, then $\mathfrak{so}_3E=\bar{\mathfrak{g}}_E$ would contain a one-dimensional ideal $\mathfrak{g}_E$, which is impossible;

(d) $\dim\mathfrak{g}'=3$.

TABLE I. *Lie subalgebras of* $\mathfrak{so}_3\mathbb{R}\simeq\mathfrak{so}_3\mathbb{R}\oplus\mathfrak{so}_3\mathbb{R}$. $\mathfrak{so}_3E$ and $\mathfrak{so}_3F$ are, respectively, the first and the second summand of $\mathfrak{so}_4\mathbb{R}\simeq\mathfrak{so}_3\mathbb{R}\oplus\mathfrak{so}_3\mathbb{R}$. (e_1,e_2,e_3) is a basis of $\mathfrak{so}_3E$ for which $[e_1,e_2]=e_3$, $[e_2,e_3]=e_1$, $[e_3,e_1]=e_2$. (e_1',e_2',e_3') is a basis of $\mathfrak{so}_3F$ for which $[e_1',e_2']=e_3'$, $[e_2',e_3']=e_1'$, $[e_3',e_1']=e_2'$.

dim 1	dim 2	dim 3	dim 4	dim 5
$\mathbb{R}e_1$	$\mathbb{R}e_1+\mathbb{R}e_1'$	$\mathfrak{so}_3E$	$\mathbb{R}e_1\oplus\mathfrak{so}_3F$	
$\mathbb{R}e_1'$		$\mathfrak{so}_3F$	$\mathfrak{so}_3E\oplus\mathbb{R}e_1'$	
$\mathbb{R}(e_1+e_1')$		$\{(A,A),\ A\in\mathfrak{so}_3\mathbb{R}\}$		

If this case occurs, then $\mathfrak{g}\cap\mathfrak{so}_3E=\mathfrak{g}\cap\mathfrak{so}_3F=0$. Therefore $\mathfrak{g}\simeq\bar{\mathfrak{g}}_E\simeq\bar{\mathfrak{g}}_F\simeq\mathfrak{so}_3\mathbb{R}$ is the ''diagonal'' Lie subalgebra,

$$\mathfrak{g}=\{(A,A)\in\mathfrak{so}_3\mathbb{R}\oplus\mathfrak{so}_3\mathbb{R},\quad A\in\mathfrak{so}_3\mathbb{R}\}.$$

Results are summarized in Table I. In particular, we find the following results, already stated in Ref. 2.

(1) *The Lie algebra* $\mathfrak{so}_4\mathbb{R}$ *has no five-dimensional Lie subalgebras*.

(2) *Every subalgebra of* $\mathfrak{so}_4\mathbb{R}$ *of dimension four is a direct sum of two summands: a one-dimensional subalgebra contained either in the ideal* $\mathfrak{so}_3E$ *or in* $\mathfrak{so}_3F$, *and the ideal* $\mathfrak{so}_3F$ *or* $\mathfrak{so}_3E$, *respectively*.

III. LIE SUBALGEBRAS OF $\mathfrak{so}_4\mathbb{C}$

It is well known that $\mathfrak{so}_4\mathbb{C}\simeq\mathfrak{sl}_2\mathbb{C}\oplus\mathfrak{sl}_2\mathbb{C}$ (see, e.g., Ref. 3). We therefore begin by investigating the Lie subalgebras of $\mathfrak{sl}_2\mathbb{C}$.

Each nonzero matrix $A\in\mathfrak{sl}_2\mathbb{C}$ is $GL_2\mathbb{C}$ conjugate—and hence $SL_2\mathbb{C}$ conjugate—to one of the following matrices in Jordan canonical form:

$$\begin{vmatrix}\lambda & 0\\ 0 & -\lambda\end{vmatrix}\quad\text{or}\quad\begin{vmatrix}0 & 1\\ 0 & 0\end{vmatrix}.$$

Therefore the only one-dimensional Lie subalgebra of $\mathfrak{sl}_2\mathbb{C}$ are, up to conjugation, $\mathbb{C}H$ and $\mathbb{C}E_+$, where

$$H=\begin{vmatrix}1 & 0\\ 0 & -1\end{vmatrix}\quad\text{and}\quad E_+=\begin{vmatrix}0 & 1\\ 0 & 0\end{vmatrix}.$$

Since two-dimensional Lie algebras are solvable and it is well known that Borel subalgebras (maximal solvable Lie subalgebras) are conjugate (see Ref. 3), the only two-dimensional Lie subalgebra of $\mathfrak{sl}_2\mathbb{C}$ is

$$\left\{\begin{vmatrix}a & b\\ 0 & -a\end{vmatrix} q \quad a,b\in\mathbb{C}\right\}=\mathbb{C}H\ltimes\mathbb{C}E_+,$$

where we use the following:

Notation. $\mathfrak{g}=\mathfrak{a}\ltimes\mathfrak{h}$ means that the Lie algebra $\mathfrak{g}$ is a *semidirect sum* of the Lie subalgebra $\mathfrak{a}$ and of the *ideal* $\mathfrak{h}$.

We now go back to the problem of describing Lie subalgebras of $\mathfrak{so}_4\mathbb{C}$. Let $\mathfrak{g}$ be a Lie subalgebra of $\mathfrak{so}_4\mathbb{C}\simeq\mathfrak{sl}_2\mathbb{C}\oplus\mathfrak{sl}_2\mathbb{C}=\mathfrak{sl}E\oplus\mathfrak{sl}F$. Here we simply call $\mathfrak{sl}E$ the first summand and $\mathfrak{sl}F$ the second, with $E\simeq F\simeq\mathbb{C}^2$. Fix a linear *subspace* $\mathfrak{g}'\subseteq\mathfrak{sl}E\oplus\mathfrak{sl}F$ such that

$$\mathfrak{g}=(\mathfrak{g}\cap\mathfrak{sl}E)\oplus(\mathfrak{g}\cap\mathfrak{sl}F)\oplus\mathfrak{g}',$$

as a direct sum of *vector spaces*. (Of course, such a $\mathfrak{g}'$ is not uniquely determined.)

Set

$$\mathfrak{g}_E = \mathfrak{g} \cap \mathfrak{sl}E, \quad \mathfrak{g}_F = \mathfrak{g} \cap \mathfrak{sl}F;$$

let

$$pr_E : \mathfrak{sl}E \oplus \mathfrak{sl}F \to \mathfrak{sl}E, \quad pr_F : \mathfrak{sl}E \oplus \mathfrak{sl}F \to \mathfrak{sl}F$$

be the natural projections, and denote by

$$\bar{\mathfrak{g}}_E = pr_E(\mathfrak{g}), \quad \bar{\mathfrak{g}}_F = pr_E(\mathfrak{g})$$

the projections of $\mathfrak{g}$ on $\mathfrak{sl}E$ and $\mathfrak{sl}F$.

Remark: $\mathfrak{g}_E$ is an *ideal* of $\bar{\mathfrak{g}}_E$ (and similarly $\mathfrak{g}_F$ is an ideal of $\bar{\mathfrak{g}}_F$).

Proof: Let $x \in \mathfrak{g}_E$, $y = pr_E(y+z) \in \bar{\mathfrak{g}}_E$, with $y+z \in \mathfrak{g}$, $z \in \mathfrak{sl}F$; then $[x,y] = [x,y+z] \in \mathfrak{g}$, because both x and $y+z$ are in $\mathfrak{g}$. But $[x,y] \in \mathfrak{sl}E$, because $x,y \in \mathfrak{sl}E$; therefore $[x,y] \in \mathfrak{g} \cap \mathfrak{sl}E = \mathfrak{g}_E$.

Remark: dim $\mathfrak{g}' \leqslant 3$ (by the Grassmann formula).

Let us examine the different cases: dim $\mathfrak{g}' = 0,1,2,3$.

Case 0: dim $\mathfrak{g}' = 0$.

Then the Lie subalgebra $\mathfrak{g}$ is a direct sum, as a Lie algebra, of the two Lie subalgebras $\mathfrak{g} \cap \mathfrak{sl}E \subset \mathfrak{sl}E$ and $\mathfrak{g} \cap \mathfrak{sl}F \subset \mathfrak{sl}F$:

$$\mathfrak{g} = (\mathfrak{g} \cap \mathfrak{sl}E) \oplus (\mathfrak{g} \cap \mathfrak{sl}F).$$

Case 1: dim $\mathfrak{g}' = 1$.

From the vector space isomorphisms

$$\frac{\mathfrak{g}}{\mathfrak{g} \cap \mathfrak{sl}F} \simeq (\mathfrak{g} \cap \mathfrak{sl}E) \oplus \mathfrak{g}' \quad \text{and} \quad \frac{\mathfrak{g}}{\mathfrak{g} \cap \mathfrak{sl}F} \simeq \bar{\mathfrak{g}}_E,$$

we have $\bar{\mathfrak{g}}_E \simeq (\mathfrak{g} \cap \mathfrak{sl}E) \oplus \mathfrak{g}'$. Therefore

$$\dim \bar{\mathfrak{g}}_E + 1 \quad \text{and} \quad \dim \bar{\mathfrak{g}}_F = \dim \mathfrak{g}_F + 1.$$

Hence dim $\bar{\mathfrak{g}}_E < 3$ and dim $\bar{\mathfrak{g}}_F < 3$, because the simple Lie algebras $\mathfrak{sl}E$ and $\mathfrak{sl}F$ can not contain two-dimensional ideals. The following cases are possible: if dim $\bar{\mathfrak{g}}_E = 2$, that is $\bar{\mathfrak{g}}_E = \mathbb{C}H \ltimes \mathbb{C}E_+$, then $\mathfrak{g}_E = \mathbb{C}E_+$, because $\mathbb{C}H$ is not an ideal in $\mathbb{C}H \ltimes \mathbb{C}E_+$; if dim $\bar{\mathfrak{g}}_E = 1$ (i.e., $\bar{\mathfrak{g}}_E = \mathbb{C}E_+$), then $\mathfrak{g}_E = 0$. Analogous results hold for $\bar{\mathfrak{g}}_F$, $\mathfrak{g}_F$.

Case 2: dim $\mathfrak{g}' = 2$.

By the same argument as in Case 1, we have

$$\dim \bar{\mathfrak{g}}_E = \dim \mathfrak{g}_E + 2 \quad \text{and} \quad \dim \bar{\mathfrak{g}}_F = \dim \mathfrak{g}_F + 2;$$

since $\mathfrak{sl}_2\mathbb{C}$ is simple, dim $\bar{\mathfrak{g}}_E = $ dim $\bar{\mathfrak{g}}_F = 2$. Thus $\bar{\mathfrak{g}}_E = \mathrm{st}\ E$, $\bar{\mathfrak{g}}_F = \mathrm{st}\ F$, and therefore

$$\mathfrak{g} = \left\{ (A,A) \in \mathfrak{sl}_2\mathbb{C} \oplus \mathfrak{sl}_2\mathbb{C}, \quad A = \begin{vmatrix} a & b \\ 0 & -a \end{vmatrix} \quad a,b \in \mathbb{C} \right\} = \mathbb{C}(H+H') \ltimes \mathbb{C}(E_+ + E'_+).$$

Case 3: dim $\mathfrak{g}' = 3$.

This implies $\bar{\mathfrak{g}}_E = \mathfrak{sl}E$ and $\bar{\mathfrak{g}}_F = \mathfrak{sl}F$. Therefore $\mathfrak{g}$ is the "diagonal" Lie subalgebra,

$$\mathfrak{g} = \{ (A,A) \in \mathfrak{sl}_2\mathbb{C} \oplus \mathfrak{sl}_2\mathbb{C}, \quad A \in \mathfrak{sl}_2\mathbb{C} \}.$$

See Table II for a summary of results.

TABLE II. *Lie subalgebras of* $\mathfrak{so}_4\mathbb{C}\simeq\mathfrak{sl}_2\mathbb{C}\oplus\mathfrak{sl}_2\mathbb{C}$.

dim 1	dim 2	dim 3	dim 4	dim 5
$\mathbb{C}H$	st E	$\mathfrak{sl}E$	$\mathbb{C}H+\mathfrak{sl}F$	st $E+\mathfrak{sl}F$
$\mathbb{C}H'$	st F	$\mathfrak{sl}F$	$\mathfrak{sl}E+\mathbb{C}H'$	$\mathfrak{sl}E+$st F
$\mathbb{C}E_+$	$\mathbb{C}(H+H')\ltimes\mathbb{C}(E_++E'_+)$	$\{(A,A),A\in\mathfrak{sl}_2\mathbb{C}\}$	$\mathbb{C}E_+\oplus\mathfrak{sl}F$	
$\mathbb{C}E'_+$	$\mathbb{C}H\oplus\mathbb{C}H'$	st $E\oplus\mathbb{C}H'$	$\mathfrak{sl}E\oplus\mathbb{C}E'_+$	
$\mathbb{C}(H+H')$	$\mathbb{C}E_+\oplus\mathbb{C}E'_+$	$\mathbb{C}H\oplus$st F	st $E\oplus$st F	
$\mathbb{C}(E_++E'_+)$	$\mathbb{C}H\oplus\mathbb{C}E'_+$	$\mathbb{C}E_+\oplus$st F		
$\mathbb{C}(E_++H')$	$\mathbb{C}E_+\rtimes\mathbb{C}(H+H')$	st $E\oplus\mathbb{C}E'_+$		
	$\mathbb{C}E'_+\rtimes\mathbb{C}(H+H')$	$(\mathbb{C}E_++\mathbb{C}E'_+)\rtimes\mathbb{C}(H+H')$		
	$\mathbb{C}E_+\rtimes\mathbb{C}(H+E'_+)$			
	$\mathbb{C}E'_+\rtimes\mathbb{C}(E_++H')$			

$\mathfrak{sl}E$ and $\mathfrak{sl}F$ are, respectively, the first and the second summand of $\mathfrak{sl}2\mathbb{C}\oplus\mathfrak{sl}_2\mathbb{C}$.

$H=\begin{vmatrix}1 & 0\\ 1 & -1\end{vmatrix}$, $E_+=\begin{vmatrix}0 & 1\\ 0 & 0\end{vmatrix}\in\mathfrak{sl}E$; $H'=\begin{vmatrix}1 & 0\\ 0 & -1\end{vmatrix}$, $E'_+=\begin{vmatrix}0 & 1\\ 0 & 0\end{vmatrix}\in\mathfrak{sl}F$.

st $E=\left\{\begin{vmatrix}a & b\\ 0 & -a\end{vmatrix}a,b\in\mathbb{C}\right\}=\mathbb{C}H\oplus\mathbb{C}E_+$ (Borel subalgebra of $\mathfrak{sl}E$).

st $F=\left\{\begin{vmatrix}a & b\\ 0 & -a\end{vmatrix}a,b\in\mathbb{C}\right\}=\mathbb{C}H'\oplus\mathbb{C}E_+$ (Borel subalgebra of $\mathfrak{sl}F$).

$\mathfrak{g}=\mathfrak{g}_1\rtimes\mathfrak{g}_2$ is *the semidirect sum* of the Lie subalgebras $\mathfrak{g}_1$ and $\mathfrak{g}_2$; $\mathfrak{g}_1$ is an *ideal*.

ACKNOWLEDGMENT

The author acknowledges his debt to D. Alekseevsky for discussions and mathematical help.

[1] J. Tits, *Liesche Gruppen und Algebren* (Springer-Verlag, Berlin, 1983).
[2] D. Meyer, ''Représentations linéaires en dimension inférieure ou égale a 4,'' Exposé II in *Géométrie Riemannienne en Dimension 4*, Séminaire Arthur Besse 1978/79, Cedic Fernand Nathan, Paris, 1981.
[3] W. Fulton and J. Harris, *Representation Theory* (Springer-Verlag, New York, 1987).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

String structure over the Brownian bridge

Rémi Léandre
Département de Mathématiques, Institut Elie Cartan, Faculté des Sciences, Université de Nancy I, 54000. Vandoeuvre-les-Nancy, France

(Received 22 September 1997; accepted for publication 18 September 1998)

We give the definition of a line bundle over the Brownian bridge by using its section. This allows us to define a Hilbert space of sections of a line bundle over the Brownian bridge associated to the transgression of a representative of an element of $H^3(M;Z)$. We consider the case of a string structure over the Brownian bridge: this allows us to define a Hilbert space of spinor fields over the Brownian bridge, when the first Pontryaguin class of the spin bundle over the manifold is equal to 0. [S0022-2488(99)00401-6]

I. INTRODUCTION

Let M be a compact-oriented even-dimensional manifold. A spin structure over M is a lifting of the frame bundle by the spinor group ''spin'' which is a $Z/2Z$ extension of $\mathrm{SO}(n)$ where n is the dimension of M. We choose a $Z/2Z$ cover of $\mathrm{SO}(n)$ because the spin representation in finite dimension is a projective representation of $\mathrm{SO}(n)$ with fiber $Z/2Z$.[1] There is a topological obstruction to constructing a spin structure over M,[2] which is the second Stiefel–Whitney class of M. It cannot be seen generally at the level of differential forms and of the de Rham theory over M.

Our goal is to go at the infinite-dimensional case, in order to construct spin structures over some infinite-dimensional objects. Namely in finite dimension, the Dirac operator over a manifold is related to the construction of a spin structure. In infinite dimension, the infinite-dimensional Dirac operator will be related to the construction of infinite-dimensional spin structures. Let us consider the free loop space $L_{\mathrm{fin}}(M)$ of M: it is the space of finite energy maps from the circle S^1 into M. In other words, it is the set of loops $s \to \gamma_s$ such that $\int_0^1 \| d/ds\, \gamma_s \|^2 ds < \infty$. Infinite-dimensional operators over $L_{\mathrm{fin}} M$ give topological invariants of the manifold,[3,4] if we consider their index. These are the Witten genus, the elliptic genus, and the elliptic genus of level N, which play an important role in the theory of cobordism. In order to work with these operators, we need a measure. The latter is useful for the computation of the adjoint of the operator. The index theorem is a statement at the level of Sobolev spaces. It is true at least in finite dimension. Our choice of measure is to take this Brownian bridge measure. The loops under study are only continuous.[5] A beginning of construction of these operators is given in Refs. 6–9. The case studied in Ref. 6 is related to the Witten genus of the manifold and to the Dirac operator over the free loop space.

Unfortunately, Ref. 6 gives only an approach to the construction of the Dirac operator over the free loop space, because the fiber is not what it should be. Namely, the construction of the suitable fiber is related to the theory of representations of a loop group.[1] Namely, the projective spin representation of a loop group is related to a central extension of this loop group by the circle and not by $Z/2Z$. It uses differential form unlike the finite-dimensional construction of the universal cover of $\mathrm{SO}(n)$.[10] If we consider the case of the smooth free loop space, the structure group is the loop group of $\mathrm{SO}(n)$. A spin structure is related to a lifting of the principal bundle in a loop group by a central extension of this loop group.[11,4] The obstruction is related to the first Pontryaguin class of the original bundle,[3,4,12–14] and, more precisely, to its transgression; but, of course, these considerations work only for the smooth loop space. On the other hand, the BHK measure is used in order to compute the adjoint of this type of operator and the Brownian bridge is only continuous. Let us recall the BHK measure. It is the measure over the free loop space $[p_1(x,x)dx \otimes dP_{1,x}]/[\int p_1(x,x)dx]$ where $p_t(x,y)$ is the heat kernel over the Riemannian manifold M and

$dP_{1,x}$ is the Brownian bridge measure over the based loop space of continuous loops starting from x. Gross[15] had shown that the tangent space of a continuous loop remains a Hilbert space in infinite-dimensional analysis, although the support of the measure is a Banach manifold. In this paper, we will follow the choice of Ref. 16 or 17 for the tangent space of a continuous loop.

Reference 13 has given a geometrical construction of such lifting, called a string structure, in the case of the based loop space of finite energy loops $L_{\mathrm{fin}}M$. Our goal is to construct such lifting for the Brownian bridge. The purpose of this work is to try to unify the considerations of Ref. 13 over the smooth loop space with the stochastic considerations of Refs. 17 and 18 over the continuous loop space. In order to construct a string structure, Ref. 13 introduces a commutative diagram. The holonomy over a loop plays an important role in this commutative diagram. Namely, let us consider a G principal bundle P over M. We consider the based loop space LP of P of loop starting from (x,e) in P: it is an LG principal bundle. Reference 13 constructs a U_1 bundle over LP denoted $\tilde{L}P$ which is an $\tilde{L}G$ bundle for a given central extension of LG. In order to give an explicit construction, they suppose that LP and LG are simply connected. Under this hypothesis, we can construct this bundle for a given representative of a integral cohomology class of LP of degree two. They avoid in this way all torsion phenomena: a circle bundle over a simply connected manifold is indeed determined by its curvature. So the construction of this lift is strongly related to a theory of forms and a theory of construction of a bundle by starting from a given cohomology class. The treatment is simplified, because the loop space of P is supposed simply connected.

In Ref. 17 or 18 the continuous loop space is introduced. A suitable Hilbert space which is the tangent space of a continuous loop is introduced in order to get the integration by parts formulas.[5,18,19] Reference 17 developed a theory of forms over the loop space without differential operations. In Refs. 20 and 21 a theory of stochastic cohomology of the Brownian loop is developed, where the stochastic Chen forms play a key role.

The purpose of this paper is to go one step further, by starting from elementary geometrical considerations: that is, to construct bundles by starting from a stochastic two-form. A preliminary work is Ref. 9. A closed Chen form of degree 1 is integrated by patching together its local integrals, when the path space is supposed simply connected. The key tool is the positivity theorem of Ref. 22 or 23. (The reader can find in Ref. 24 a jump version of this theorem.) Of course, we cannot integrate explicitly any closed one-form. Namely, we meet the problem that the deformation of a loop which is useful in order to integrate a form over a path in the continuous loop space does not keep the measure. Generally, the law of the deformed stochastic loop is a foreign law with respect to the Brownian bridge law. For this reason, we restrict our attention to the case where the form is a stochastic Chen form. This form is defined for a larger class of processes.

In the first part of this paper, we start from a closed Chen form of degree 2 with integral values over $L_{\mathrm{fin}}M$. An example is the transgression $\tau(\omega)$ of a representative ω of an element of $H^3(M,Z)$. A surface in the loop space is constituted by a volume in the manifold. The main property is that the integral of the transgression over a surface in the smooth loop space is equal to the integral of the underlying form over the underlying volume in M.

Let us recall[14,1] how we construct a line bundle over a simply connected manifold (of finite dimension or of infinite dimension) by starting from its curvature $\tilde{\omega}$. It is the set of couples of the shape $(l_\cdot(x),\alpha)$ [$l_\cdot(x)$ is a path starting from a point of reference and arriving at x in the manifold and α is an element of the complex line]. We say that the couple $(l_\cdot(x),\alpha)$ and $(l'_\cdot(x'),\alpha')$ are equivalent if the endpoints of the paths x and x' are equal and if

$$\alpha'=\alpha\ \exp\left[-2i\pi\int_S\tilde{\omega}\right]. \tag{1.1}$$

S is any surface bounded by the loop constituted of $l_\cdot(x)$ and $l'_\cdot(x')$ circled in the opposite sense. Such surface exists because we suppose that the manifold is simply connected. In the stochastic context, $\int_S\tau(\omega)$ leads to the study of some stochastic integrals. We have the choice of some distinguished surfaces in order to be able to integrate this transgression over a given surface in the stochastic loop space. Instead of constructing the bundle globally as in (1.1), we construct it by using a system of transition functions. However, we cannot construct the stochastic line bundle associated to this Chen form over the stochastic-based loop space. Namely, the transition functions are only almost surely defined. We request that the transition functions belong to all the Sobolev spaces, in order to give some rigidity to them. Let us recall namely that all bundles over a finite-dimensional manifold are measurably trivial. However, we construct the Hilbert space of sections of this formal line bundle, and a connection which operates over the space of sections. We hope that there are some applications to the geometric quantization of the loop space.[25] In order to do this, we suppose that the based loop space of loops with finite energy $L_{\mathrm{fin}}(M)$ is simply connected. We can determine explicitly the system of the measurable transition functions. A connection allows us to define the space of C^1 Sobolev sections of this formal line bundle. The bundle obtained is an extension in a certain sense of the underlying bundle over the finite energy loop space. Namely, if we take the polygonal approximation of a continuous loop, the transition functions are surely defined. They are, in fact, the transition functions which defined the underlying line bundle over the smooth loop space. Moreover, they tend almost surely to the transition functions which define the formal line bundle over the continuous loop space, when the length of the subdivision goes to the infinity.

We summarize the results of the second part of this work in this theorem-definition. (We refer to Ref. 20 for the theory of forms over the loop space.)

Definition I.I: Let $L_x(M)$ be the based loop space of the Riemannian manifold M endowed with the Brownian bridge measure. Let us suppose that M is two connected, and let us consider a three-form ω which is closed Z-valued. We can construct the two-form over $L_x(M)\int_0^1\omega(d\gamma_s,.,.)=\tilde{\omega}$ where γ denotes the typical continuous loop in M. There exists a system of open subsets O_i for the uniform topology which is a cover of $L_x(M)$ and a set $\rho_{i,j}$ of Brownian functionals with values in the unit circle of C, defined over $O_i\cap O_j$, which are almost surely defined and which belong to all the Sobolev spaces. Over $O_i\cap O_j$, they are such that $\rho_{i,j}\rho_{j,i}=1$ almost surely and, over a triple intersection $O_i\cap O_j\cap O_k$, they satisfy almost surely to the relation $\rho_{i,j}\rho_{j,k}\rho_{k,i}=1$. Moreover, there exists a set of one-forms over O_i denoted by A_i with values in the Lie algebra of the unit circle of C, such that almost surely over $O_i\cap O_j$ we have $A_i=A_j+\rho_{i,j}^{-1}d\rho_{i,j}$ and such that $dA_i=\sqrt{-1}2\pi\tilde{\omega}$. A section of the formal line bundle associated to the set of $\rho_{i,j}$ is a collection of random variables α_i from O_i into C such that $\alpha_j=\alpha_i\rho_{i,j}$ over $O_i\cap O_j$. The Hilbert space of sections is the set of sections ψ such that $E[|\psi|^2]<\infty$ where $|\alpha_i|=|\psi|$ is intrinsically defined.

After these elementary geometrical considerations, we consider the more complicated case of the construction of a string structure over the Brownian bridge. In the second part, we consider a G principal bundle P over M. We suppose that G is simple, simply connected. Moreover, we suppose that the based loop group of loops of finite energy $L_{\mathrm{fin}}G$ is simply connected. There exists a central extension $\tilde{L}_{\mathrm{fin}}G$ of $L_{\mathrm{fin}}G$. Let us suppose that there is a unitary representation Spin of $\tilde{L}_{\mathrm{fin}}G$, which is a Hilbert space (The reader can find references in this direction in Ref. 1 or 26). We construct over *LM*, the set of the Brownian bridge, an $L_{\mathrm{fin}}G$ bundle which extends the natural $L_{\mathrm{fin}}G$ bundle which exists over *LM*. We use the parallel transport over a loop. This bundle Q is only formal, in the sense that the transition functions are only measurable, because they belong only to all the Sobolev spaces. This allows us to construct, by using the associated transition functions, a formal linear bundle whose fiber is Spin. But we can define rigorously the space of sections of this formal linear bundle, which belong to some L^p spaces. The Dirac operator over the based loop space should act over the Hilbert space of a section of this formal bundle. We have the same topological obstructions which are involved with the first Pontryaguin class P of M as in the smooth case. This part is a mixture of the geometrical construction of Ref. 11 where the holonomy along a stochastic-based loop has a key role and of the stochastic considerations of the first part.

We can summarize the results of the third part of this work by the following theorem-definition:

Definition I.2: Let $Q\rightarrow M$ be a principal bundle over a compact Riemannian manifold with compact structural group, which is supposed simple and simply connected. We suppose that M is two connected. Let $L_{\text{fin}}G$ be the based loop group of G of loops with finite energy. Let L_eQ be the space of loops in Q of the shape $\tau_s^G g_s$ where g_s is a path in G of finite energy and τ_s^G the parallel transport over a random loop in M for the Brownian bridge measure, if the principal bundle Q is endowed by a connection. Associated to L_eQ, there exists a system of subsets V_i of L_xM such that $\cup V_i = L_xM$ almost surely. Over $V_i \cap V_j$, there exists a natural measurable application $g_{i,j}$ with values in $L_{\text{fin}}G$ such that $g_{i,j}g_{j,i} = e_.$ almost surely over $V_i \cap V_j$ and such that almost surely over $V_i \cap V_j \cap V_k$, $g_{i,j}g_{j,k}g_{k,i} = e_.$. We suppose that the Pontryagin class of the bundle is equal to 0. There exists a refinement of V_i, still denoted by V_i, and a measurable application $\rho_{i,j}$ from $V_i \cap V_j$ into $\tilde{L}_{\text{fin}}G$, the basical central extension of $L_{\text{fin}}G$, which satisfies the following conditions:

(i) Over $V_i \cap V_j$, $\rho_{i,j}\rho_{j,i} = \tilde{e}$ almost surely.
(ii) Over $V_i \cap V_j \cap V_k$, $\rho_{i,j}\rho_{j,k}\rho_{k,i} = \tilde{e}$ almost surely.
(iii) $\pi \rho_{i,j} = g_{i,j}$ almost surely, where π is the projection from $\tilde{L}_{\text{fin}}G$ into $L_{\text{fin}}G$.

Let us suppose, given a unitary representation of $\tilde{L}_{\text{fin}}G$, the space of measurable sections of the associated bundle to the system of transition functionals $\rho_{i,j}$ is given by a set of measurable functionals in the representation space called ψ_i defined over V_i, satisfying $\psi_j = \rho_{j,i}\psi_i$ almost surely over $V_i \cap V_j$. The Hilbert space of the sections is the space of sections such that $E[\|\psi\|^2] < \infty$, where the random variable $\|\psi\| = \|\psi_i\|$ over V_i is intrinsically defined.

This work is an element of a set of three papers about this construction.[27,28] We thank M. Katz for helpful discussions.

II. STOCHASTIC LINE BUNDLE

Let M be a compact Riemannian manifold. Let $L_xM = LM$ be the space of continuous loops in M starting from x and coming back in time 1 in x. Let $dP_{1,x}$ be the law of the Brownian bridge starting from x and coming back in time 1 in x: it is a probability measure over LM. Let us recall how $dP_{1,x}$ is defined. Let us introduce the heat kernel $p_t(x,y)$ associated to the Laplace–Beltrami operator over M. Let $F(\gamma_{t_1},\ldots,\gamma_{t_r})$, $0 < t_1 < \cdots < t_r < 1$, be a cylindrical functional over the based loop space. We have

$$E_{1,x}[F(\gamma_{t_1},\ldots,\gamma_{t_r})] = \frac{1}{p_1(x,x)} \int \int F(x_1,\ldots,x_r) p_{t_1}(x,x_1) p_{t_2-t_1}(x_1,x_2) \cdots p_{1-t_r}(x_r,x) dx_1 \ldots dx_r; \tag{2.1}$$

$s \to \gamma_s$ is a semi-martingale with respect to the canonical filtration associated to the canonical process $\gamma_. \to \gamma_s$.

Let us recall for that the definition of a semi-martingale. Let us introduce a filtration of σ-algebras indexed by R^+ such that $F_s \subseteq F_t s > t$ and a process X_s with values in R^d such that

(i) X_s is F_s adapted.
(ii) $s \to X_s$ is continuous almost surely for a given probability P.
(iii) X_s is in L^2.
(iv) $E[X_t | F_s] = X_s$ almost surely for $s < t$.

We say that such a random process $X_.$ is a L^2 martingale adapted to the filtration F_s. A semi-martingale is, roughly speaking, the sum of a martingale and an adapted finite variational process.

The main result is the following: if $X_.^i$, $i = 1,\ldots,d$ is a semi-martingale in R^d, $X_.^i X_.^j$ is still a semi-martingale. When we consider a family of martingales, there is an adapted finite variational process associated to each product, called the bracket of $X_.^i X_.^j$ and denoted by $\langle X^i, X^j \rangle_s$.

Let us consider a martingale X_s and Y_s a continuous ADAPTED process which is supposed to be bounded. The Itô integral $\int_0^t \langle Y_s, \delta X_s\rangle$ is the limit of the sum $\Sigma_i \langle Y_{t_i}, X_{t_{i+1}} - X_{t_i}\rangle$ for a given partition when the length of the subdivision tends to 0. It is still an L^2 martingale if the process $X_.$ is a L^2 martingale. The main ingredient in the sequel is the Itô formula.

Let $X_.$ be a semi-martingale over R^d and let f be a function from R^d into R with bounded derivatives of all orders. $f(X_.)$ is still a semi-martingale and we have

$$f(X_t) = f(X_0) + \int_0^t \langle df(X_s), \delta X_s\rangle + \frac{1}{2} \sum \int_0^t \frac{\partial^2}{\partial x_i \partial x_j} f(X_s) d\langle X^i, X^j\rangle_s . \tag{2.2}$$

We will not specify the natural different criteria of integrability which appear in these definitions.

We will say that a process $X_.$ over a manifold is a semi-martingale if, for all smooth functions with bounded derivatives with values in R, the process $f(X_.)$ is a semi-martingale with values in R. In order to specify the meaning of the sentence ''the Brownian bridge'' is a semi-martingale, we have to define a probability space. It is the space LM of continuous functions $s \to \gamma_s$ from the circle into the manifold such that $\gamma_0 = x$. The formula (2.1) gives over LM a probability measure: it is not immediately clear and is the purpose of a theorem. The σ-algebra F_s is the smallest σ-algebra such that the evaluation maps $e_u: \gamma_. \to \gamma_u$ are measurable for all $u \leqslant s$. We will not specify the technical difficulties which appear here.

Let γ_t be a loop. Let τ_t be the parallel transport from γ_0 to γ_t. In order to define it, we consider the polygonal approximation γ_t^n of a loop γ, and the parallel transport τ_t^n along the polygonal approximation of this loop: $\tau_.^n \to \tau_.$ almost surely for the topology of the uniform convergence.

In order to be more precise, we have to recall quickly the theory of stochastic differential equations. Let $X_.$ be a semi-martingale in R^d. We consider $A(y)$ a linear application from R^d into R^m which depends smoothly on y in R^m. We consider the Itô stochastic differential equation:

$$dY_t = A(Y_t)\delta Y_t . \tag{2.3}$$

This means that almost surely

$$Y_t = Y_0 + \int_0^t A(Y_s)\delta Y_s . \tag{2.4}$$

It has a unique solution. It can be given by the Picard iteration method or the Peano approximation, which converges to the solution.

There is another type of stochastic differential equation, which is simpler under a geometrical point of view. Let $Y_.$ and $X_.$ be two R-valued semi-martingales. We put

$$\int_0^t Y_s dX_s = \int_0^t Y_s \delta X_s + \frac{1}{2} \int_0^t d\langle X, Y\rangle_s . \tag{2.5}$$

It is called the Stratonovich integral. The Itô formula becomes the classical formula in the Stratonovich calculus:

$$f(X_t) = f(X_0) + \int_0^t f'(X_s) dX_s . \tag{2.6}$$

This leads to the following principle, called Malliavin's transfer principle. A formula which is true over a manifold in the deterministic context, which has a meaning in the stochastic context, is still true in the stochastic context. This leads to the theory of Stratonovich stochastic differential equations:

$$dY_t = A(Y_t) dX_t , \tag{2.7}$$

instead of (1.3) or equivalently $Y_t = Y_0 + \int_0^t A(Y_s) dX_s$. This equation can be converted into an Itô stochastic differential equation.

An important case of stochastic process is the flat Brownian motion.

We consider a set $B^i_\cdot$, $i=1,...,m$, R-valued martingales in L^2 with bracket $\langle B^i, B^j \rangle_s = \delta_{i,j} s$ starting from $(x_i) = x$ in R^m. The law of such a process is uniquely determined and is called the law of the flat Brownian motion starting from x. Let us introduce the flat Laplacian Δ over R^m:

$$\Delta = \frac{1}{2} \sum_i^m \frac{\partial^2}{\partial x_i^2}. \tag{2.8}$$

Let $\exp[t\Delta]$ be the semi-group over R^m associated to the flat Laplacian. We have a stochastic representation of it. Namely, if f is a smooth function with bounded derivatives, we get

$$\exp[t\Delta] f(x) = E[f(B_t)], \tag{2.9}$$

where $B_\cdot$ is a flat Brownian motion starting from x.

Let X_i be some vector fields over the compact manifold M. Let Δ_c be the operator $\frac{1}{2}\Sigma_1^m X_i^2 + X_0$. Let us introduce the Stratonovich differential equation,

$$dX_t = \sum_1^m X_i(X_t) dB_t^i + X_0(X_t) dt, \tag{2.10}$$

starting from x. Let us introduce the heat semi-group associated to Δ_c denoted by $\exp[-t\Delta_c]$. We have the following stochastic representation:

$$\exp[t\Delta_c] f(x) = E[f(X_t)]. \tag{2.11}$$

The equation (2.10) has a meaning in local charts, and we can patch together the local solutions, because the Itô formula in the Stratonovich context is the usual one. The equation (2.10) gives Schwartz's construction of the Brownian motion over a manifold: let us recall namely that the Laplace–Beltrami operator over a Riemannian compact manifold can be written as $\Delta = \frac{1}{2}\Sigma X_i^2 + X_0$. The semi-group associated to Δ (called the heat semi-group associated to the Riemannian manifold) has a stochastic representation in terms of the semi-martingale solution of (2.10). The law of the Brownian motion starting from x gives a probability measure over the space of continuous functions $\gamma_\cdot$ starting from x into M. Let us denote by dP_1^x this probability measure. It is characterized by

$$E_{P_1^x}[f(\gamma_{t_1}, \ldots, \gamma_{t_r})] = \int \int \cdots \int f(x_1, \ldots, x_r) p_{t_1}(x, x_1) p_{t_2 - t_1}(x_1, x_2) \ldots p_{t_r - t_{r-1}}(x_{r-1}, x_r) dx_1 \ldots dx_r. \tag{2.12}$$

Compare with (2.1). This gives another approach to the Brownian bridge measure. If we condition the Brownian motion to come back to the departure, the law of γ_t is given by the density $[p_t(x,y) p_{1-t}(y,x) / p_1(x,x)] dy$. We deduce that the law of γ_t satisfies an inhomogeneous heat equation. Equation (2.10) becomes

$$dX_t = \sum X_i(X_t)(dB_t^i + \langle X_i(X_t), \ \text{gradlog } p_{1-t}(X_t, x) \rangle dt) + X_0(t) dt. \tag{2.13}$$

Therefore, the equation of the bridge is an inhomogeneous Stratonovich differential equation. It is a nontrivial result that $X_t \to x$ when $t \to 1$ and that $X_\cdot$ is a semi-martingale. The law of X_t starting from x is the measure over the based loop space given by (2.1).

Let $Q \to M$ be a principal bundle with compact structural group. We can trivialize it locally. We endow it with a connection ∇. Let Γ be the Christoffel symbol in this trivialization. Let $X_.$ be a semi-martingale over the manifold. The following Stratonovich differential equation has a meaning

$$d\tau_t = -\Gamma_{\tau_t}(X_t)dX_t \tag{2.14}$$

if we operate in the local chart of the trivialization. Since the Itô formula in the Stratonovich calculus is the traditional one, this equation is intrinsically defined and we can patch together the local solution of it, when we change of charts. We call τ_t the parallel transport over the semi-martingale X_t. This works too if we look at a linear bundle over M. This allows us to give the notion of a semi-martingale with values in Q over the semi-martingale $X_.$: $Z_.$ is a semi-martingale over $X_.$ if and only if $\tau_t^{-1}Z_t$ is a semi-martingale over X_0.

The interest of the Stratonovich calculus is twofold: first of all, as we have seen, the Itô formula in the Stratonovich calculus is the classical one. Second, if we consider the polygonal approximation of B_t^i, we get a classical solution of (2.10) or (2.13), and we get a finite energy process $X_.^n$ which tends almost surely to the solution solution of the Stratonovoch differential equation for the uniform topology. This second property can be generalized. If in (2.14) we consider the polygonal approximation of X_t, which is defined if the length of the dyadic subdivision tends to infinity, we get the classical equation of the parallel transport over a finite energy curve. Let us denote it by $\tau_.^n$ When $n \to \infty$, $\tau_.^n \to \tau_.$ almost surely. We can consider other types of finite energy approximation of the leading semi-martingale $X_.$ (See Ref. 29, for instance).

This allows us to give the notion of the Stratonovich integral of a one-form along the trajectories of a semi-martingale X_t. Let ω be a one-form over the manifold and let X_t be a semi-martingale in the manifold M. The Stratonovich integral $\int_0^1 \langle \omega(X_s), dX_s \rangle$ is the limit of the classical integrals $\int_0^1 \langle \omega(X_s^n), dX_s^n \rangle$ for the polygonal approximation $X_.^n$ of the semi-martingale $X_.$.

The tangent space of a loop $\gamma_.$ is the space of sections X_t over γ_t of TM such that

$$X_t = \tau_t H_t, \quad H_0 = H_1 = 0, \tag{2.15}$$

where $s \to H_s$ is a process in $T_{\gamma_0}(M)$. We suppose, moreover, that H_s has finite energy and we take as the Hilbertian norm

$$\|X\|^2 = \int_0^1 |H_s'|^2 ds. \tag{2.16}$$

Let us recall the definition of a Chen form. Let $\Omega_.(M)$ be the space of forms over the manifold of degree not equal to 0. Let $\tilde{\omega}_n \in \Omega_.^{\otimes}$:

$$\tilde{\omega}_n = \omega_1 \otimes \cdots \otimes \omega_n. \tag{2.17}$$

We associate a form over the based loop space by the formula

$$\Sigma \tilde{\omega}_n = \int_{0<s_1<\cdots<s_n<1} \omega_1(d\gamma_{s_1}, .) \wedge \cdots \wedge \omega_n(d\gamma_{s_n}, .). \tag{2.18}$$

We will not repeat general considerations over the Chen forms. We will give two examples in order to show how they act. First of all, let us consider an r-form. $\Sigma\omega_1$ is an $r-1$ form over the loop space given by the formula

$$\Sigma\omega_1(X^1, \ldots, X^{r-1}) = \int \omega_1(d\gamma_s, X_s^1, \ldots, X_s^{r-1}). \tag{2.19}$$

This formula does not cause any problem if we work over the smooth loop space, because X_s is a vector over γ_s and because the differential $d\gamma_s$ is defined. If we consider continuous loops, we get $X_s=\tau_s H_s$ and $H_s=\int_0^s H'_u\,du$. We apply the Fubini theorem in (2.19) and we have

$$\Sigma\omega_1(X^1,\ldots,X^{r-1})=\int\cdots\int k(s_1,\ldots,s_{r-1})H'^{1}_{s_1}\cdots H'^{r-1}_{s_{r-1}}\,ds_1\cdots ds_{r-1}, \tag{2.20}$$

where $k(s_1,\ldots,s_{r_1})$ is a stochastic kernel given by Stratonovitch differential integral.

Let us consider the case of two-forms of degree 2, ω_1 and ω_2. Let us consider $\tilde{\omega}_2=\omega_1\otimes\omega_2$. For smooth loops,

$$\Sigma\tilde{\omega}_2(X^1,X^2)=\int_{0<s_1<s_2<1}\omega_1(d\gamma_{s_1},X^1_{s_1})\omega_2(d\gamma_{s_2},X^2_{s_2})-\int_{0<s_1<s_2<1}\omega_1(d\gamma_{s_1},X^2_{s_1})\omega_2(d\gamma_{s_2},X^1_{s_2}). \tag{2.21}$$

For continuous loops, $X^1_s=\tau_s H^1_s$, $X^2_s=\tau_s H^2_s$ where H^1 and H^2 have finite energy. We apply the Fubini theorem in the stochastic integral (2.21), and we deduce a formula analogous to (2.20) for $\Sigma\tilde{\omega}_2$.

The exterior derivative of the Chen form $\Sigma\tilde{\omega}_n$ corresponds to the Chen form associated to the Hochschild coboundary of $\tilde{\omega}_n$.[30] Let us recall that the Hochschild coboundary is given by $(b_1+b_2)\tilde{\omega}_n$. Without specifying the signs which appear in the following expressions (see Refs. 30 or 20), we have

$$b_1(\tilde{\omega}_n)=\sum \text{ sign } \omega_1\otimes\cdots\otimes d\omega_i\otimes\cdots\otimes\omega_n \tag{2.22}$$

and

$$b_2(\tilde{\omega}_n)=\sum \text{ sign } \omega_1\otimes\cdots\otimes\omega_i\wedge\omega_{i+1}\otimes\cdots\otimes\omega_n\,. \tag{2.23}$$

For the smooth loop space, the cohomology of Chen forms is equal to the cohomology of the based loop space.[31] Equivalently, the Hochschild cohomology of forms of the manifold is equal to the cohomology of the smooth loop space. A version of this theorem for the continuous loop space can be seen in Refs. 20 and 21. Especially for the continuous loop space of an homogeneous manifold, the Brownian cohomology is equal to the Hochschild cohomology of the forms of the manifold. This remains true for the finite energy loop space. In particular, the part without torsion of the Z cohomology of $L_{\text{fin}}M$ for loops of finite energy is given by Chen forms. Let us take a representative ω of a cohomology class of degree 3 with integral values. The transgression

$$\tau(\omega)=\int_0^1\omega(d\gamma_s,.,.) \tag{2.24}$$

is a Chen form over $L_{\text{fin}}M$ which is closed and with values in Z. Its integral over closed surfaces in the smooth loop space is equal to an integer. Namely, the integral over this surface of $\tau(\omega)$ is equal to the integral over the underlying volume in M of ω. If the surface in the loop space has no boundary, the underlying volume in the manifold has no boundary. This comes from the fact that the circle has no boundary.[25]

Let us recall how we produce an S^1 bundle from a Z closed form of degree 2 $\tilde{\omega}$ over a simply connected manifold V.[14] This construction works, of course, for the smooth based loop space.

We fix a base point $x_0\in V$. We consider the set of triple $(x,l_{(x_0,x)},\alpha)$ where $l_{(x_0,x)}$ is a path going from x_0 to x, and α is an element of S^1. We identify two triples $(x,l_{(x_0,x)},\alpha)$ and $(x',l'_{(x_0,x')},\alpha')$ if $x=x'$ and if $\alpha'=\exp[-2i\pi\int_B\tilde{\omega}]\alpha$. B is any surface bounded by the loop going from x_0 to x by l and going back from x to x_0 by l' circled in the opposite sense. Since ω is integral, the value of $\int_B \omega$ depends on the surface by an integer. Moreover, we can find such a surface because V is simply connected. The integral over the surface depends on the surface by an integer, because $\tilde{\omega}$ is Z valued.

We construct the connection from the parallel transport. Let $l(x,x')$ be a path in V. The parallel transport from x to x' of $(x, l_{(x,x_0)}, \alpha)$ is given by the following formula: we choose the path which goes from x_0 to x', first by going from x_0 to x by $l_{(x_0,x)}$ and, second, by going from x to x' by $l_{(x,x')}$, and we choose the same α. If we choose a loop which goes from x to $x l_{(x,x)}$, the holonomy over this loop is given by $\exp[-2\pi i \int_B \omega]$ where B is a surface bounded by the loop $l_{(x,x)}$.

This procedure for constructing an S^1 bundle is global and cannot be extended to the construction of line bundles over the Brownian bridge. However, the reader can find a solution of this problem in Ref. 32, and a definition of a stochastic curve in the stochastic loop space. We will overcome this problem in another way: $\tau(\omega)$ becomes a stochastic Chen form, and the integral over B of $\tau(\omega)$ leads to the study of some stochastic integrals. So we have to choose a distinguished surface which depends in a ''nice'' way on $\gamma_.$ in order to perform these stochastic integrals. The same remark remains true for a general Z-valued closed Chen form. We present a procedure which can be adapted to the Brownian bridge.

Let x_0 be a base point in V. Let O_i be a countable system of charts of V: there is a distinguished point x_i in each O_i such that the two following conditions are satisfied:

(i) If $x \in O_i$, there is a distinguished curve $l(x_i,x)$ going from x_i to x which depends smoothly on x.

(ii) There is a distinguished curve $l_{(x_0,x_i)}$ going from x_0 to x_i.

Therefore, there is by composition a distinguished curve going from x_0 to $x \in O_i l_i(x_0,x)$ which depends smoothly on x. This will give a system of charts which will allow us to construct our S^1 bundle by using a system of transition functions. Let $\alpha \in C$ over O_i. We identify by using $\rho_{j,i}\alpha$ with $\exp[-2\pi i \int_B \tilde{\omega}]\alpha$ over O_j. B is any surface bounded by the loop going from x_0 to x by $l_i(x_0,x)$ and coming back to x_0 by $l_j(x_0,x)$.

The system of $\rho_{i,j}$ defines a system of transition functions. Namely over $O_i \cap O_j$, we have

$$\rho_{i,j}\rho_{j,i} = Id, \tag{2.25}$$

and over $O_i \cap O_j \cap O_k$, we have

$$\rho_{i,j}\rho_{j,k}\rho_{k,i} = Id \tag{2.26}$$

because we integrate in this case $\tilde{\omega}$ over the sum of three surfaces:

(i) The first one is the surface bounded by the loop constituted by the path going from x_0 to x by l_i (orientation $+$) and coming back from x to x_0 by l_k (orientation $-$).

(ii) The second one is a surface bounded by the loop constituted by the path going from x_0 to x by l_k (orientation $+$) and coming back to x_0 by l_j (orientation $-$).

(iii) The third one is a surface bounded by the loop constituted of the path going from x_0 to x by l_j (orientation $+$) and coming back to x_0 by l_i (orientation $-$).

The sum of these three surfaces has no boundary, and the integral of $\tilde{\omega}$ over the sum of these three surfaces is an integer.

In the global construction given above, the parallel transport is given by the distinguished path $l_i(x_0,x)$ going from x_0 to x and after running from x to x' by $l(x,x')$ with $x' \in O_j$, with the same $\alpha \in C$. But if x' belongs to O_j, we have chosen a distinguished path $l_j(x_0,x')$ joining x_0 to x'. So the parallel transport is given by $\exp[-2i\pi \int_B \tilde{\omega}]$ where B is a surface bounded by $l_i(x_0,x)$,

$l(x,x')$, and $l_j(x_0,x')$ circled in the opposite sense: this means that $(l_i(x_0,x),\alpha)$ is transformed into $(l_j(x_0,x'),\exp[-2i\pi\int_B\tilde{\omega}]\alpha)$. We obtain that the connection form over O_i is given modulo the normalizing constant $2\pi i$ by

$$A_i(x)=\int_0^1\tilde{\omega}\left(l'_{i,t}(x_0,x),\ \frac{\partial}{\partial x}\ l_{i,t}(x_0,x)\right), \tag{2.27}$$

where $l'_{i,t}(x_0,x)$ is the speed of $l_i(x_0,x)$ and $\partial/\partial x$ the directional derivative in the chart O_i of $l_{i,t}(x_0,x)$. Here A_i is the restriction by the map $x\rightarrow l_i(x_0,x)$ of a Chen form. Therefore $dA_i=\tilde{\omega}$.[30]

Let us consider the Brownian bridge endowed with a closed Chen form $\tilde{\omega}$ of degree 2. We suppose that this Chen form restricted to the finite energy loop space is Z valued. Equivalently, we consider a form whose restriction to smooth loop is Z valued: namely, any surface without boundary can be approximated by a surface without boundary over the smooth loop space. In order to see this, we imbed the manifold into R^d, we regularize our loops by convolution, and, if the radius of the convolution is small, we can project the regularized loops into the manifold. By this procedure we get a surface without boundary over the smooth loop space. The surface which is obtained is an approximation of the original surface over the finite energy loop space. There is no boundary in all the surfaces which are considered. The integral of the Chen form over the approximated surface is an integer; by continuity, the integral over the surface in the finite energy loop space is still an integer, because the surface over the smooth loop space is an approximation of the original surface not only for the uniform distance, but also for the energy distance, which can be obtained by imbedding the manifold into a flat space.

We will postpone the problem of the locality of the following constructions later. We will now produce a system of charts O_i of LM: let γ_i be a countable system of finite energy paths such that the balls of radius δ for the uniform norm determines an open cover of LM. We can suppose that δ is small. The loop γ_i constitutes a distinguished point in $O_i=B(\gamma_i,\delta)$. (It is in fact the center of this ball.) Here $l(\gamma_i,\gamma)$ is constructed as follows: since δ is small, $\gamma_{i,s}$ and γ_s are joined by a unique geodesic. $l_t(\gamma_i,\gamma)$ is the loop $s\rightarrow\exp_{\gamma_{i,s}}[t(\gamma_s-\gamma_{i,s})]$ where $\gamma_s-\gamma_{i,s}$ is the vector of the unique geodesic joining $\gamma_{i,s}$ to γ_s. This allows us to define over O_i a distinguished path joining $\gamma_.$ to the center γ_i. We take a distinguished path joining γ_{base} to γ_i. We get a distinguished path going from γ_{base} to $\gamma_.$, $l_i(\gamma_{\text{base}},\gamma_.)$: we glue together the two previous pieces of paths.

The second step is to specify a distinguished surface bounded by $l_i(\gamma_{\text{base}},\gamma_.)$ and $l_j(\gamma_{\text{base}},\gamma_.)$, when $\gamma\in O_i\cap O_j$.

Since $O_i\cap O_j$ is not empty, and since δ is small, there is a path $u\rightarrow\exp_{\gamma_{i,.}}[u(\gamma_{j,.}-\gamma_{i,.})]$ joining γ_i to γ_j. Because $L_{\text{fin}}M$ is supposed simply connected, we can fill in by a surface the deterministic triangle constituted by the path joining γ_{based} to γ_i, the path joining γ_i to γ_j, and the path joining γ_j to γ_{based}. We can, moreover, fill in the small stochastic triangle constituted by $l_t(\gamma_i,\gamma)$, $l_t(\gamma_j,\gamma_.)$, and the path $\exp_{\gamma_{i,.}}u(\gamma_{j,.}-\gamma_{i,.})$ by the surface defined as follows: we join $l_t(\gamma_i,\gamma_.)$ to $l_t(\gamma_j,\gamma)$ by the curve $l_{t,u,i,j}(\gamma_.):s\rightarrow\exp_{l_t(\gamma_i,\gamma)_s}u(l_t(\gamma_j,\gamma)_s-l_t(\gamma_i,\gamma)_s)$. We produce in this way a small random surface bounded by the triangle whose the vertices are $\gamma_{i,.}$., $\gamma_.$, and $\gamma_{j,.}$. We glue together these two surfaces, and we get a surface bounded by $l_i(\gamma_{\text{base}},\gamma_.)$ and $l_j(\gamma_{\text{base}},\gamma_.)$. Let $l_{u,v}(\gamma)$ be this surface called $B_{i,j}(\gamma_.)$. We have to specify a little bit what we mean by a surface. It is a sum of smooth maps from the oriented simplex in R^2 into the continuous loop space. (See Ref. 32 for a complete theory of such stochastic surfaces.) In our case, we consider the sum of a small stochastic simplex and a deterministic simplex: (u,v) describes a square.

This surface $B_{i,j}(\gamma)$ is useful in so far as we can integrate the Chen form $\tilde{\omega}$ over it by means of the theory of stochastic integrals. Namely, $l_{u,v}(\gamma)_s$ is a semi-martingale, and $(\partial/\partial u)l_{u,v}(\gamma_.)$ is a semi-martingale over the semi-martingale $l_{u,v}(\gamma_.)$. $(\partial/\partial v)l_{u,v}(\gamma_.)$ too, is, a semi-martingale over $l_{u,v}(\gamma_.)$. We recall what is a semi-martingale over a semi-martingale $X_.$ with values in a manifold: the stochastic parallel transport τ_s along a curve with respect to the Levi-Civita connection is almost surely defined. The process Y_s is a semi-martingale if $\tau_s^{-1}Y_s$ is a semi-martingale with value in the tangent space of the starting point of $X_.$, which is supposed deterministic. We can specify what is the sense of the second term, because $(\partial/\partial u)l_{u,v}(\gamma)(s)$ is a vector over $l_{u,v}(\gamma)(s)$. We parallel transport $(\partial/\partial u)l_{u,v}(\gamma)(s)$ over the path $l_{u,v}(\gamma)(s)$ circled in the opposite sense by using the time in the reversed sense. We get a semi-martingale over the tangent space of x. Therefore the integral of $\tilde{\omega}$ over $B_{i,j}(\gamma_.)$ appears as a nonanticipative Stratonovitch integral, which is smooth in $\gamma_.$. In order to be more precise, it is $\tilde{\omega}((\partial/\partial u)l_{u,v}(\gamma_.),(\partial/\partial v)l_{u,v}(\gamma_.))$, which is a stochastic Stratonovitch integral, when $l_{u,v}$ belongs to the small stochastic triangle constituting one of the two parts of $B_{i,j}(\gamma_.)$. Let us specify a little bit what it means:

$$\tilde{\omega}\left(\frac{\partial}{\partial u}\, l_{u,v}(\gamma_.),\, \frac{\partial}{\partial v}\, l_{u,v}(\gamma_.)\right)=\int_0^1 \omega\left(d_s l_{u,v}(\gamma_.),\, \frac{\partial}{\partial u}\, l_{u,v}(\gamma_.)_s\,,\, \frac{\partial}{\partial v}\, l_{u,v}(\gamma_.)_s\right). \tag{2.28}$$

Moreover, it belongs to all the Sobolev spaces. Therefore, we can consider the system of transition functions $\rho_{i,j}(\gamma_.)=\exp[-2\pi\sqrt{-1}\int_{B_{i,j}(\gamma_.)}\tilde{\omega}]$. The transition functions belong to all the Sobolev spaces.[18] Let us recall briefly what these Sobolev spaces are: let $X_s=\tau_s H_s$, H_s deterministic, be a vector field over γ. We get the integration by parts formula:

$$E[\langle dF,X\rangle]=E[F \text{ div } X] \tag{2.29}$$

true for any cylindrical functional F. We put

$$\langle dF,X\rangle=\int_0^1 \langle k_s\,,d/dsH_s\rangle ds. \tag{2.30}$$

This allows us to define the notion of first-order Sobolev space, because this notion of derivative is intrinsically defined by using (2.29):

$$\|F\|_{W_{1,p}}=E[|F|^p]^{1/p}+E\left[\left(\int_0^1 |k_s|^2 ds\right)^{p/2}\right]^{1/p}. \tag{2.31}$$

We can iterate the notion of gradient. Namely, we introduce a connection $\nabla X_.=\tau_. DH_.$, where D is the traditional H-derivative. Let us suppose that $d_\nabla^r F$ is defined inductively. We define $d_\nabla^{r+1}F$ by

$$d_\nabla^{r+1}F(X_1,\ldots,X_{r+1})=\langle d(d_\nabla^r F(X_1,..,X_r),X_{r+1}\rangle-\sum\, d_\nabla^r F(X_1,..,\nabla_{X_{r+1}}X_i\,,..,X_r). \tag{2.32}$$

Here d_∇^r is a Hilbert–Schmidt cotensor: it is given by kernels. Let us denote it by $k^r(s_1,\ldots,s_r)$. We put

$$\|F\|_{W_{r,p}}=E[|F|p]^{1/p}+\sum_{j\leqslant r}\, E\left[\left(\int\int\, |k^j(s_1,\ldots,s_j)|^2 ds_1\cdots ds_j\right)^{1/2}\right]^{1/p}. \tag{2.33}$$

The main theorem (see Ref. 18, theorems 3.6 and 3.7 for the free loop space) is that stochastic integrals belong to all the Sobolev spaces as well as the solution of stochastic differential equations.

By using the positivity theorem of Ref. 22 or 23, we deduce that the $\rho_{i,j}$ are transition functions, in the sense that they satisfy (2.25) and (2.26) almost surely. Moreover, the $\rho_{i,j}$ are smooth functions, in the sense that they belong to all the Sobolev spaces.

The previous discussion is not completely true, because these considerations work only locally. As a first step, we regularize the indicator function of O_i. Let g be a function from $[0,\delta]$ into $[1,\infty]$, which is infinite if z is larger than $\delta-\alpha/2$, which is smooth over $[0,\delta-\alpha/2[$, and which behaves as $1/(z-\delta+\alpha/2)^{+n}$ when $z \to (\delta-\alpha/2)_-$. Moreover, $z \leqslant \delta-\alpha$ is equivalent to $g(z)=1$. Let F be an auxiliary function from $[1,\infty]$ into $[0,1]$, which is equal to 1 only in 1 and with compact support. Let G_i be the functional:

$$G_i(\gamma_.)=F\left(\int_0^1 g(d(\gamma_s,\gamma_{i,s}))ds\right). \tag{2.34}$$

The functional $G_i(\gamma_.)$ is equal to 1 if and only if $d(\gamma,\gamma_i) \leqslant \delta-\alpha$ and if $G_i(\gamma_.)>0$, $d(\gamma,\gamma_i)<\delta$.

Moreover, G_i is smooth (it belongs to all the Sobolev spaces), if n is big enough. Namely, by the exponential inequality, we have

$$P\left\{\text{Sup}\,\frac{1}{(d(\gamma_s,\gamma_{i,s})-\delta+\alpha/2)^+}>1/\epsilon;\int\frac{1}{(d(\gamma_s,\gamma_{i,s})-\delta+\alpha/2)}+<C\right\}<C(p)\epsilon^p. \tag{2.35}$$

Moreover, we can find a sequence α_i, g_i, F_i such that the sequence of the corresponding $G_i(\gamma)$ tends simply to the indicator function of O_i. The reader can see Ref. 8, p. 106, and Ref. 6, Eqs. (2.12)–(2.14), for analogous considerations.

We imbed the manifold in R^d. We can extend $(x,y) \to \exp_x[u(y-x)]$ if x and y are close to a smooth map from $R^d \times R^d$ into R^d which depends smoothly on u. This allows us to extend to the whole loop space $L(M)$ the distinguished path which was defined only over O_i, and to extend to $L(M)$ the distinguished stochastic surface $B_{i,j}(\gamma_.)$. The price to pay is that $l_{i,t}(\gamma_.)$ is a process which belongs to R^d and not to M. We can extend the forms constituting the Chen form over R^d. This gives an extension of the Chen form to the semi-martingale $l_{i,t}(\gamma_.)$ and to the semi-martingale $l_{u,v}$ over the extended surface of $B_{i,j}(\gamma_.)$. This works when $\gamma_.$ describes the whole loop space and not only O_i or $O_i \cap O_j$. This allows us to produce stochastic integrals, whose restriction over $O_i \cap O_j$ is equal to the stochastic integrals which appear in the definition of $\rho_{i,j}$. These stochastic integrals are smooth over LM.

This allows us to give the following definition:

Definition II.1: A measurable section ϕ of the line bundle associated to the Chen form $\tilde{\omega}$ is a collection of random variables α_i over O_i with value in C submitted to the rule

$$\alpha_j=\alpha_i\rho_{i,j} \tag{2.36}$$

almost surely over $O_i \cap O_j$.

Over O_i, we put the metric

$$\|\alpha(\gamma_.)\|^2=\|\alpha\|^2 \tag{2.37}$$

Since the transition functions are of modulus one, this metric is intrinsically defined.

We can give the definition:

Definition II.2: The L^p space of sections of the line bundle associated to $\tilde{\omega}$ is the space of measurable sections ϕ endowed with the norm:

$$\|\phi\|_{L^p}=\|\|\phi\|\|_{L^p}. \tag{2.38}$$

Unfortunately the random measurable sections of the line bundle do not give a lot of information about the structure of the line bundle. For this purpose, we need to study C^1 sections. We will introduce a connection.

Over O_i, the connection form is given by

$$A_i(\gamma)=\int_0^1 \tilde{\omega}_{l_{i,t}(\gamma_{\text{base}},\gamma_.)}\left(l'_{i,t}(\gamma_{\text{base}},\gamma_.),\,\frac{\partial}{\partial\gamma_.}\,l_{i,t}(\gamma_{\text{base}},\gamma_.)\right)dt. \tag{2.39}$$

We have to specify what the meaning of this expression is. $l'_{i,t}(\gamma_{\mathrm{base}},\gamma_.)$ is a semi-martingale in the tangent bundle of M over $l_{i,t}(\gamma_{\mathrm{base}},\gamma_.)$. The quantity $(\partial/\partial\gamma_.)l_{i,t}(\gamma_{\mathrm{base}},\gamma_.)$ too, is a semi-martingale over $l_{i,t}(\gamma_{\mathrm{base}},\gamma_.)$ if we describe what it is. Let us recall namely that in infinite-dimensional analysis, the tangent space is smaller than the natural tangent space of the functional space which is studied.[15] In $(\partial/\partial\gamma)l_{i,t}(\gamma_{\mathrm{base}},\gamma_.)$, we have to take the derivative of $l_{i,t}(\gamma(\mathrm{base}),\gamma)(s)$ over the tangent vector $\tau_t(\gamma_.)H_t$ ($H_0=H_1=0$, H_t of finite energy). This leads to the study of some stochastic integrals whose the structure is very similar to stochastic Chen forms.[17] Namely,

$$\frac{\partial}{\partial\gamma_.}\exp_{\gamma_{i,s}}[t(\gamma_s-\gamma_{i,s})]\tau_sH_s=\tau_s(l_{i,t}(\gamma_{\mathrm{base}},\gamma_.)A_sH_s\,, \tag{2.40}$$

where A_s is a nonanticipative process of matrices. Moreover the Ito–Stratonovitch formula of Ref. 33 shows that

$$d\ \exp_{\gamma_{i,s}}[t(gg_s-\gamma_{i,s})]=\frac{\partial}{\partial\gamma_s}\exp_{\gamma_{i,s}}[t(\gamma_s-\gamma_{i,s})]d\gamma_s+\text{finite energy}. \tag{2.41}$$

These finite energy terms are nonanticipative processes. This shows that the stochastic integral part of $A_i(\gamma)$ is

$$\tilde{A}_i(\gamma)(H)=\int_0^1\langle B_s\tau_s^{-1}d\gamma_s\,,H_s\rangle, \tag{2.42}$$

where B_s is a nonanticipative semi-martingale. Of course these manipulations are true only over O_i, but they can be extended smoothly over LM as it was done for $\rho_{i,j}$. We extend $l_{i,t}$ to the whole $L(M)$, and we extend $\tilde{\omega}$ to the new set of paths by extending the forms over M which define $\tilde{\omega}$.

For a finite energy loop, the connection forms A_i are compatible with the transition functions $\rho_{i,j}$. We deduce from the positivity theorem of Refs. 22 and 23 that this remains true in the stochastic case. This theorem tells namely that a formula which is given in terms of stochastic integrals and of the parallel transport of the Brownian bridge is true almost surely if the underlying formula which is given for the smooth loop space is true surely. Let γ^n be the polygonal approximation of $\gamma_.$. Here $\rho_{i,j}(\gamma^n)$ is surely defined and checks surely (2.25) and (2.26). Moreover, by the rules of the approximation of Stratonovitch integrals by traditional integrals, $\rho_{i,j}(\gamma^n_.)$ tends almost surely to $\rho_{i,j}(\gamma_.)$, when the length of the subdivision tends to the infinity.[33] This says that the formal bundle over $L(M)$ extends the bundle over the finite energy loop space which is deduced from $\tilde{\omega}$.

For finite energy paths, the connection preserves the metric, because the parallel transport preserves the metric. Let $\alpha(\gamma_.)$ be a scalar which depends only on a finite number of coordinates. It is a cylindrical functional. Let $G_i(\gamma_.)\alpha=\phi_i(\gamma_.)$ be the corresponding section of the line bundle which is equal to 0 outside O_i. Let us consider another section of the same structure. Since the connection preserves the metric

$$\nabla_X\langle\phi,\phi'\rangle=\langle\nabla_X\phi,\phi'\rangle+\langle\phi,\nabla_X\phi'\rangle \tag{2.43}$$

for any vector field X.

Let us suppose that the vector field X has a divergence (see Refs. 16, 5, and 19). This means that

$$E[\langle dF,X\rangle]=E[F\ \mathrm{div}\ X] \tag{2.44}$$

for all cylindrical functionals. We conclude that

$$\nabla_X^* = -\nabla_X + \operatorname{div} X. \tag{2.45}$$

If H_t is deterministic, X_t has a divergence (see Refs. 16 and 5), which is in all the L^p.

For any sum ϕ of sections ϕ_i of the previous structure, $\nabla_. \phi$ is a tensor in the tangent space in $\gamma_.$ with value in the fiber. We take its Hilbert–Schmidt norm $\|\nabla\phi\|_\gamma$. The $W_{p,1}(\tilde{\omega})$ norm of this special type of section is

$$\|\phi\|_{W_{p,1}(\tilde{\omega})} = \|\|\phi\|\|_{L^p} + \|\|\nabla_.\phi\|\|_{L^p}. \tag{2.46}$$

Since we have integration by parts formulas, the operation of covariant derivative is closable and allows us to obtain the following.

Definition II.3: The Sobolev space $W_{p,1}(\tilde{\omega})$ of a section of the formal line bundle associated to $\tilde{\omega}$ is the completion of finite combination $\Sigma\phi_i$ of local sections over O_i for the norm (2.45).

The space of C^1 sections is the intersection of the Sobolev spaces $W_{p,1}(\tilde{\omega})$. Let us recall that there is the Sobolev space $W_{p,1}$ of functionals which belong to L^p such that dF belongs in L^p. We have, obviously, the following.

Proposition II.4: The space of C^1 sections $\cap W_{p,1}(\tilde{\omega})$ is a module over $\cap W_{p,1}$.

Moreover, if F belongs to all the $W_{p,1}$, if ϕ is a C^1 section, and if X is a vector field,

$$\nabla_X(F\phi) = \langle dF, X\rangle\phi + F\nabla_X\phi. \tag{2.47}$$

Remark: Over O_i, the connection form has an exterior derivative. Namely, it is given by nonanticipative stochastic integrals and the paths $l_{i,t}(\gamma_.)$ depend smoothly on $\gamma_.$. We can apply indeed to A_i the procedure of Ref. 20 in order to define a stochastic exterior derivative, although the Lie bracket of two vector fields is not a vector field. Over the finite energy loop space,

$$dA_i = \tilde{\omega} \tag{2.48}$$

by construction. As a matter of fact, A_i is a Chen form over a set of paths over $L(M)$. The connection form A_i is a Chen form over the set of distinguished paths over $L(M)$ constructed from the form $\tilde{\omega}$ over $L(M)$. Therefore,

$$dA_i(\gamma) = \tilde{\omega}(l_{i,1}l(\gamma_{\text{base}},\gamma_.))\left(\frac{\partial}{\partial\gamma}\, l_{i,1}(\gamma_{\text{base}},\gamma_.),\ \frac{\partial}{\partial\gamma}\, l_{i,1}(\gamma_{\text{base}},\gamma_.)\right) = \tilde{\omega} \tag{2.49}$$

by the rules over the exterior derivatives of a Chen form, which are still true in infinite dimension. It remains true over the continuous loop space almost surely, by the positivity theorems of Refs. 22 and 23. The curvature of the formal line bundle is $\tilde{\omega}$. This notion of curvature has, however, only a formal meaning, because $[X,Y]$ is not a vector field.

Remark: If we suppose that the loops γ_i are smooth, which is possible, we can apply the theory of Ref. 20 to the form $A_i(\gamma)$. Namely, we imbed M into R^d and we extend ω into ω_{ext}. It is possible to extend $l_{i,t}(\gamma_.)_s$ into a function $f_t(s,\gamma_s)$ where f is a smooth application from R^d into R^d, smooth in s and piecewise smooth in t.

Here $A_i(\gamma)$ is the restriction to $B(\gamma_i;\delta)$ of the global form $\tilde{A}_i(\gamma)$ given by the following formula:

$$\tilde{A}_i(\gamma) = \int_0^1\int_0^1 \omega_{\text{ext}}(f_t(s,\gamma_s))\left(d_s f_t(s,\gamma_s),\ \frac{\partial}{\partial t}\, f_t(s,\gamma_s),\ \frac{\partial}{\partial\gamma_.}\, f_t(s,\gamma_s)\right). \tag{2.50}$$

Let us recall what is a form which belongs to all the Nualart–Pardoux spaces over the loop space (see Ref. 20). Let σ be a form of degree n over the Brownian bridge:

$$\sigma(\tau_. H^1_.,\ldots,\tau_. H^n) = \int_0^1\cdots\int_0^1 k(s_1,\ldots,s_n)H'^1_{s_1}\cdots H'^n_{s_n}\, ds_1\cdots ds_n, \tag{2.51}$$

where $k(s_1,\ldots,s_n)$ is a random variable in $T^{\otimes n}_{\gamma_0}$ (We do not specify the conditions of antisymmetry which naturally appear.) Since we have a connection over the Brownian bridge [see a little bit before (2.32) the definition of this connection], we can define the iterate covariant derivative of the form σ for this connection; let us denote them by $\nabla^r\sigma$. It is given by kernels $k(s_1,\ldots,s_n;t_1,\ldots,t_r)$. The Nualart–Pardoux Sobolev norms $N.P_{p,k}(LM)$ of the n-form σ are given by the two smallest constants $C(p;k)$ and $C'(p;k)$ such that

$$\|k(s_1,\ldots,s_n;t_1,\ldots,t_k)-k(s'_1,\ldots,s'_n;t'_1,\ldots,t'_k)\|_{L^p}\leqslant C(p;k)\left(\sum \sqrt{|s_i-s'_i|}+\sum \sqrt{|t_j-t'_j|}\right) \tag{2.52}$$

if $s_1,\ldots,s_n,t_1,\ldots,t_r$ and $s'_1,\ldots,s'_n,t'_1,\ldots t'_r$ belong to the same connected component of $[0,1]^{n+k}$. where we have removed the diagonal and such that

$$\|k(s_1,\ldots,s_n;t_1,\ldots t_k)\|_{L^p}\leqslant C'(p;k) \tag{2.53}$$

for all $s_1,\ldots,s_n,t_1,\ldots t_k$. By definition, the space of forms which belong to all the Nualart–Pardoux Sobolev spaces $N.P_{k,p}(LM)$ is the space of forms which are smooth in the Nualart–Pardoux sense, denoted by $N.P_{\infty-}(LM)$, endowed with its natural topology. The exterior derivative is continuous from $N.P_{\infty-}(LM)$ into $N.P_{\infty-}(LM)$ (see Ref. 20).

Since $s\to f_t(s,.)$ is smooth, $\tilde{A}_i(\gamma)$ is a form of degree 1 which is smooth in the Nualart–Pardoux sense. We can therefore define its exterior derivative, and define by restriction the exterior derivative of $A_i(\gamma)$.

Remark: We can specify now what is a finite-dimensional bundle over $L(M)$. Let G be a compact Lie group and let R be a representation of it over C^d or over R^d. Let O_i be a countable family of open set which constitutes a cover of $L(M)$ and let $\rho_{i,j}(\gamma)$ be a set of transition functions with values in G. In other words, we suppose that over $O_i\cap O_j$

$$\rho_{i,j}\rho_{j,i}=1 \tag{2.54}$$

almost surely, and that over $O_i\cap O_j\cap O_k$

$$\rho_{i,j}\rho_{j,k}\rho_{k,i}=1 \tag{2.55}$$

almost surely. We can define the set of measurable sections of the associated to the representation R as in the Definition I.1. But we can go one step further. We imbed G into $\mathrm{SO}(N)$. For all smooth functionals F with support in $O_i\cap O_j$, we suppose that $\gamma\to F\rho_{i,j}$ belongs to all the Sobolev spaces. We would like to say that the transition functions are in some sense continuous. For that, we define the capacity of an open set:

$$Cap_{r,p}(O)=\mathrm{Inf}\{\|F\|_{W_{r,p}},F\geqslant 1_O \text{ almost surely}\}. \tag{2.56}$$

The capacity of a Borelian subset A of $L(M)$ is the infimum of the capacities of the open subset which contain A. A slim set is a set, all of whose capacities equal 0.[34] For instance, a point is a slim set, unlike the finite-dimensional case. A functional which belongs to all the Sobolev spaces should have a redefinition which is continuous outside a slim set (see Ref. 34, Chap. IV.2). This shows us that the transition functionals $\rho_{i,j}$ should have a redefinition continuous outside a slim set and (2.54) and (2.55) should be checked outside a slim set for this redefinition. In others words, we should be able to define the topological space of the principal bundle over $L(M)$ outside a set of capacity 0. We should be able to define the linear bundle associated to the representation R outside a slim set. Under this point of view, the open set O_i should be defined modulo a slim set and they should constitute a cover modulo a slim set. We do not give more details about this point of view, because we are ultimately interested in infinite-dimensional operators, which operate about a space of sections.

Remark: The system of sections of the line bundle depends apparently on the systems of curve $\gamma_{i,.}$ chosen. Let us consider another countable system of finite energy loops γ_i' such that the balls of radius δ' for δ' small enough and of center γ_i' constitute an open cover of $L(M)$ for the uniform distance. The set of $B(\gamma_j', \delta') \cap B(\gamma_i, \delta)$ constitutes still an open cover of $L(M)$. Let us call this open cover by O_i. There exists two distinguished curves going from γ_{base} to γ: the first one, $l_i(\gamma_{\text{base}}, \gamma)$, is constructed as before and is passing by γ_i. The second one, $l'(\gamma_{\text{base}}, \gamma)$, is passing by γ_i'. There exists two distinguished surfaces $B_{i,j}(\gamma)$ and $B_{i,j}'(\gamma)$ constructed as before over $O_i \cap O_j$, by choosing the system of γ_i or the system of γ_i'. Let us denote

$$\rho_{i,j}(\gamma) = \exp\left[-2\pi\sqrt{-1}\int_{B_{i,j}(\gamma)} \tilde{\omega}\right] \tag{2.57}$$

and

$$\rho_{i,j}'(\gamma) = \exp\left[-2\pi\sqrt{-1}\int_{B_{i,j}'(\gamma)} \tilde{\omega}\right]. \tag{2.58}$$

We choose a small stochastic surface as it was done before whose boundary is the small circle constructed from $l_i(\gamma_{\text{base}}, \gamma)$ and $l_i'(\gamma_{\text{base}}, \gamma)$ circled in the opposite sense. Let us call it $S_i(\gamma)$. We put

$$\chi_i(\gamma) = \exp\left[-2\pi\sqrt{-1}\int_{S_i(\gamma)} \tilde{\omega}\right]. \tag{2.59}$$

We get, since $\tilde{\omega}$ is Z valued, almost surely

$$\rho_{i,j}(\gamma) = \chi_i(\gamma)\chi_j^{-1}(\gamma)\rho_{i,j}'(\gamma). \tag{2.60}$$

Let us define the section of the line bundle by its local section α_i for $\rho_{i,j}$ and α_i' for $\rho_{i,j}'$. Let us consider the transformation:

$$\alpha_i' \to \alpha_i' \chi_i^{-1} = \alpha_i. \tag{2.61}$$

It defines an isomorphism of the measurable sections of the line bundle and of the space of L^p sections of the line bundle, since the complex number χ_i is of modulus 1. Namely the relation $\alpha_i' = \alpha_j' \rho_{j,i}'$ is equivalent to the relation $\alpha_i = \alpha_j \rho_{j,i}$ over $O_i \cap O_j$, because we have the relation (2.61).

III. STOCHASTIC STRING STRUCTURE

Let us introduce a compact, simple, simply connected Lie group G. Since G is simple, all the invariant bilinear forms on Lie G are proportional. We can find a Killing form such that

$$\omega(X,Y,Z) = \frac{1}{8\pi^2}\langle X,[Y,Z]\rangle \tag{3.1}$$

defines a G-invariant form with integral values,[1,14] since G is simple.

In all this part, we will identify a form invariant under the group action over the Lie algebra with the corresponding form over the Lie group.

Let $P_{\text{fin}}G$ be the space of the finite energy path over G starting from e. Let π be the projection from $P_{\text{fin}}G$ to G which to a path g_s associates its end point g_1. Let $L_{\text{fin}}G$ be the based loop group of loops of finite energy in G starting from e. Here π determines a $L_{\text{fin}}G$ fibration. If g_1 is in a small neighborhood $O_i(G)$ of G, we get a trivialization of this principal bundle. We introduce the local slice of $O_i(G)$ into $P_{\text{fin}}G$: $g_1 \to g_s^i(g_1)$ where the path $g_s^i(g)$ depends smoothly of g in $O_i(G)$ and joins e to g. The transition functions $\rho_{i,j}$ are equal to $(g_s^j)(g)^{-1}g_s^i(g)$ and satisfy the rule of a system of transition functions. Since G is compact, there exists ∇^G an $L_{\mathrm{fin}}G$ connection over $P_{\mathrm{fin}}G$.[35] Moreover, $L_{\mathrm{fin}}G$ is a Hilbertian Lie group. Its Lie algebra is the space of finite energy loops L_{fin} Lie G going from 0 and arriving at 0 at time 1 in the Lie algebra of G. We introduce over it the energy Hilbert norm.

In the local chart given over $O_i(G)$ by the slice $g \to g^i(g)$, the connection form is a one-form A_i with values in L_{fin} Lie G. This one-form is compatible with the transition function in $O_i(G) \cap O_j(G)$.[35]

Let $X_s \in L_{\mathrm{fin}}$ Lie G. We choose a two cocycle

$$c(X,Y) = \frac{1}{8\pi^2}\left(\int_0^1 \langle X_s, dX_s\rangle - \langle Y_s, dX_s\rangle\right). \tag{3.2}$$

This gives an invariant closed form of degree 2 over $L_{\mathrm{fin}}G$. It has integral values. Namely c is equal to the transgression of ω modulo an exact form.[1] The transgression of ω is the Chen form over $L_{\mathrm{fin}}G$:

$$\int_0^1 \omega(dg_{\mathrm{fin},s}, X_s, Y_s), \tag{3.3}$$

which is closed over $L_{\mathrm{fin}}G$. Namely $X_0 = Y_0 = 0$ and $d\omega = 0$. This transgression is integral because the circle has no boundary. The reader can see the first part for analogous considerations. The integrability condition holds classically for the smooth loop space[25] and not for the finite energy loop space: we can approximate any surface without boundary in the finite energy loop space by a surface without boundary in the smooth loop space, as it was done in the first chapter. Therefore the integrability condition holds for the finite energy loop space.

The formula (3.2) gives a form over P_{fin} Lie G, the Lie algebra of $P_{\mathrm{fin}}G$. The difference is that we do not have $X_0 = X_1 = 0$. The main property of this form is the following;[11] when we consider its version over $P_{\mathrm{fin}}G$:

$$dc = \pi^*(\omega). \tag{3.4}$$

Let us remark that in the previous considerations, we have always worked with finite energy paths.

Let LM be the space of continuous based loop endowed with the Brownian bridge measure. Let us suppose that there is principal bundle P over M with structure group G. We introduce a connection over P. Let τ_s^G be the parallel transport starting from e in G over $x = \gamma_0$. (γ_s, τ_s^G) is the horizontal lift of γ_s starting from e and is almost surely defined. Q is the set of paths g_s over γ_s such that $g_s = \tau_s^G l_s$ where l_s is of finite energy and such that $g_0 = g_1 = e$. We can give a description of Q in a more algebraic way.[13]

We have a map f from LM into G which to a stochastic loop γ associates $(\tau_1^G)^{-1}$. Q is the pullback of $P_{\mathrm{fin}}G$ by f. We have the following commutative diagram:

$$\begin{array}{ccc} Q & \rightarrow & P_{\mathrm{fin}}G \\ \downarrow & & \downarrow \\ LM & \rightarrow & G \end{array} \tag{3.5}$$

The first vertical map is the projection map p. The second vertical map is the projection map π. The lower horizontal map is f and the upper horizontal map is f^*. f is an application from $L(M)$ into G which belongs to all the Sobolev spaces, because the parallel transport is a solution of a stochastic differential equation.[18]

A system of trivializations of Q is given by the pullback of a system of trivializations of $P_{\mathrm{fin}}G$. If $(\tau_1^G)^{-1} \in O_i(G)$, Q is trivialized by the slice $\gamma \to g^i((\tau_1^G)^{-1})$, and the fiber is isomorphic to $L_{\mathrm{fin}}G$. The transition functions are $g^j((\tau_1^G)^{-1})^{-1} g^i((\tau_1^G)^{-1})$ and belong to all the Sobolev spaces from LM into the Hilbertian Lie group $L_{\mathrm{fin}}G$. They are smooth in the sense that they are restrictions of smooth functions in $L_{\mathrm{fin}}R^d$, after imbedding G into R^d, and after extending $g^j_{\cdot}(g)$ over R^d. This is possible because R^d is contractible.

Over the principal bundle $Q \to LM$ with projection p, we choose the pullback connection ∇^G. If $(\tau_1^G)^{-1} \in O_i(G)$, the pullback connection form is the one-form:

$$A_i(d(\tau_1^G)^{-1}) = A_i^{f^*}. \tag{3.6}$$

We can take the derivative of τ_1^G, because τ_1^G belongs to all the Sobolev spaces.

Let us introduce the Pontryagin class of the G bundle P. Let F^G be the curvature of the G bundle P for the connection ∇ determining the horizontal lift τ_s^G over the Brownian loop. We have

$$\mathrm{Pont}(F^G) = \frac{1}{8\pi^2} \langle F^G \wedge F^G \rangle. \tag{3.7}$$

It is a closed form over M. The transgression of $\mathrm{Pont}(F^G)$ is the closed form of degree 3 over $LM \int_0^1 \mathrm{Pont}(F^G)(d\gamma_s,.)$.

It is a stochastic Chen form.[17]

We have[13] the following.

Lemma III.1: $f^{*-1}\omega$ is cohomologuous to the transgression of the Pontryagin class.

Proof: Let us consider over LM the principal bundle P_s which is the pullback by the evaluation map $\gamma_\cdot \to \gamma_s$ of the principal bundle P over M. Let ∇_s be the pullback connection of the connection ∇ by this evaluation map. Its curvature F_s^G is given by

$$F_s^G(X,Y) = F_{\gamma_s}^G(X_s, Y_s). \tag{3.8}$$

Moreover, P_s is a trivial bundle: a trivialization is given by $\tau_s^G g$ where g belongs to the fiber of P over $x = \gamma_0$. In this system of trivializations, the connection form is given by the heuristic notation:

$$(\tau_s^G)^{-1} \nabla_\cdot \tau_s^G = \hat{A}_s. \tag{3.9}$$

In order to give a rigorous meaning to this notation, let us introduce ω_1^G, the canonical connection form associated to the connection ∇ over the G bundle P. U_s is the derivative of the couple (γ_s, τ_s^G) in P over the vector field $X_\cdot$. We transform $\gamma_\cdot$ in $\exp_{\gamma_\cdot}(lX)$, we compute the associated lift $\tau_s^G(l)$, and we take its derivative in $l=0$. We get a vector U_s over (γ_s, τ_s^G) in P and we apply ω_1^G to U_s.

Since the curvature is intrinsically defined, and since the bundle P_s is trivial, we have the important relation:

$$d\hat{A}_s + [\hat{A}_s, \hat{A}_s] = (\tau_s^G)^{-1} F_{\gamma_s}^G(\tau_s^G), \tag{3.10}$$

where we take the exterior derivative of $\hat{A}_s$ over LM. This exterior derivative has not only a formal sense. Namely, we can use the matrix notation for the group. The computations of Refs. 36 and 16 show that

$$\hat{A}_s(X) = \int_0^s (\tau_t^G)^{-1} F_{\gamma_t}^G(d\gamma_t, X_t) \tau_t^G, \tag{3.11}$$

which takes its values in the Lie algebra of G. Moreover, $X_s = \tau_s H_s$ where τ_s is the parallel transport for the Levi-Civita connection for the tangent space. We have the obvious relations

$H_0 = H_1 = 0$ and H_s is of finite energy.[16,17] $X_.$ is a vector over $\gamma_.$. Let us recall that it is the parallel transport τ which has a key role in the definition of Sobolev calculus over loop spaces[5,18] and not τ_s^G.

Let us suppose that H and H^1 are deterministic. Since the Levi-Civita connection is without torsion, we have[16,36]

$$[X,Y]_t = (\nabla_Y \tau_t) H_t - (\nabla_X \tau_t) H_t^1, \tag{3.12}$$

if $Y_t = \tau_t H_t^1$. Moreover, we have

$$\nabla_Y \tau_t = \tau_t^{-1} \int_0^t \tau_s^{-1} R(d\gamma_s, Y_s) \tau_s. \tag{3.13}$$

Moreover, we have

$$d\hat{A}_s(X,Y) = X\hat{A}_s(Y) - Y\hat{A}_s(X) - \hat{A}_s[X,Y]. \tag{3.14}$$

From (3.12)–(3.14), we deduce that $d\hat{A}_s$ exists and is a stochastic nonanticipative integral in the way described in Ref. 17.

Let us compute now $\hat{A}_1$. We consider the based loop space, such that $X_0 = X_1 = 0$. Therefore the counterterm which appears when we consider $\omega_1^G(U_1)$ which depends linearly from x_1 and from the Christoffel symbols of the connection ∇ disappears.[18] This shows us that

$$\hat{A}_1 = (\tau_1^G)^{-1} d\tau_1^G. \tag{3.15}$$

From the formula (2.12), we deduce the analogous formula then (5.7) in Ref. 13:

$$d_s \hat{A}_s(X) = (\tau_s^G)^{-1} F_{\gamma_s}(d\gamma_s, X_s) \tau_s^G. \tag{3.16}$$

Moreover, we have

$$F^G \wedge F^G(d\gamma_s, X_s, Y_s, Z_s) = F^G(d\gamma_s, X_s) F^G(Y_s, Z_s) + \text{antisymmetry}. \tag{3.17}$$

We have

$$(F^G \wedge F^G)(d\gamma_s, .) = F^G(d\gamma_s, .) \wedge F^G + F^G \wedge F^G(d\gamma_s, .). \tag{3.18}$$

Since the Killing form is symmetric, we have

$$\langle (F^G \wedge F^G)(d\gamma_s, .) \rangle = 2 \langle F^G(d\gamma_s, .) \wedge F^G \rangle. \tag{3.19}$$

We use the fact that the Killing form is invariant under the group action. We find, after using (3.10) and (3.11),

$$8\pi^2 \int_0^1 \text{Pont}(F^G)(d\gamma_t, .) = 2\int_0^1 \langle d_t \hat{A}_t \wedge d\hat{A}_t \rangle + 2\int_0^1 \langle d_t \hat{A}_t \wedge [\hat{A}_t, \hat{A}_t] \rangle = B + C. \tag{3.20}$$

In $\int_0^1 \langle d_t d\hat{A}_t \wedge \hat{A}_t(X,Y,Z) \rangle$, there are three expressions which occur of the type $\int_0^1 \langle d_t d\hat{A}_t(X,Y,Z) \hat{A}_t(Z) \rangle$. When we sum these expressions after performing an integration by parts, we find, since $d_t d = d_t d$,

$$\int_0^1 \langle d d_t \hat{A}_t \wedge \hat{A}_t \rangle (X,Y,Z) = -\left(\int_0^1 d_t A_t \wedge dA_t \right)(X,Y,Z). \tag{3.21}$$

We have used the symmetry of the Killing form. It remains to consider the integrated term. However, we have

$$\langle \hat{A}_1(Y)\hat{A}_1(X)-\hat{A}_1(X)\hat{A}_1(Y),\hat{A}_1(Z)\rangle=-\langle \hat{A}_1(X),\hat{A}_1(Y)\hat{A}_1(Z)-\hat{A}_1(Z)\hat{A}_1(Y)\rangle$$
$$=\langle \hat{A}_1(X),\hat{A}_1(Z)\hat{A}_1(Y)-\hat{A}_1(Y)\hat{A}_1(Z)\rangle. \tag{3.22}$$

because the form $\langle . \rangle$ is invariant under the G action. This shows that, by using (3.10),

$$\int_0^1 \langle d_t d\hat{A}_t \wedge \hat{A}_t\rangle = -8\pi^2 f^*\omega - \int_0^1 \langle d_t\hat{A}_t \wedge d\hat{A}_t\rangle. \tag{3.23}$$

This shows us that

$$dc(\hat{A},\hat{A}) = -8\pi^2 f^*\omega - 2\int_0^1 \langle d_t\hat{A}\wedge d\hat{A}_t\rangle, \tag{3.24}$$

or, equivalently, we have

$$2\int_0^1 d_t\hat{A}_t\wedge d\hat{A}_t = -8\pi^2 f^*\omega - dc(\hat{A},\hat{A}). \tag{3.25}$$

Let us study now

$$\int_0^1 \langle d_t\hat{A}_t(X),\hat{A}_t(Y)\hat{A}_t(Z)-\hat{A}_t(Z)-\hat{A}_t(Z)\hat{A}_t(Y)\rangle. \tag{3.26}$$

We integrate by parts. We find

$$\langle \hat{A}_1(X),\hat{A}_1(Y)\hat{A}_1(Z)-\hat{A}_1(Z)\hat{A}_1(Y)\rangle - \int_0^1 \langle \hat{A}_t(X),d_t\hat{A}_t(Y)\hat{A}_t(Z)-\hat{A}_t(Z)d_t\hat{A}_t(Y)\rangle$$
$$-\int_0^1 \langle \hat{A}_t(X),\hat{A}_t(Y)d_t\hat{A}_t(Z)-d_t\hat{A}_t(Z)\hat{A}_t(Y)\rangle$$
$$=\langle \hat{A}_1(X),\hat{A}_1(Y)\hat{A}_1(Z)-\hat{A}_1(Z)\hat{A}_1(Y)\rangle + \int_0^1 \langle d_t\hat{A}_t(Y),\hat{A}_t(X)\hat{A}_t(Z)-\hat{A}_t(Z)\hat{A}_t(X)\rangle$$
$$-\int_0^1 \langle d_t\hat{A}_t(Z),\hat{A}_t(X)\hat{A}_t(Y)-\hat{A}_t(Y)\hat{A}_t(X)\rangle. \tag{3.27}$$

Because the form $\langle . \rangle$ is invariant under the G action, the two last integrals cancel with the two others which appear in $\int_0^1 \langle d_t\hat{A}_t,[\hat{A}_t,\hat{A}_t]\rangle(X,Y,Z)$. These two other terms are in fact $-\int_0^1 \langle d_t\hat{A}_t(Y),\hat{A}_t(X)\hat{A}_t(Z)-\hat{A}_t(Z)\hat{A}_t(X)\rangle$ and $\int_0^1 \langle d_t\hat{A}_t(Z),\hat{A}_t(X)\hat{A}_t(Y)-\hat{A}_t(Y)\hat{A}_t(X)\rangle$. Namely in the first case, we deduce (Y,X,Z) from (X,Y,Z) by one permutation and (Z,X,Y) from (X,Y,Z) by two permutations.

In order to summarize,

$$2\int_0^1 \langle d_t\hat{A}_t,[\hat{A}_t,\hat{A}_t]\rangle = 8\pi^2 2\, f^*\omega. \tag{3.28}$$

We deduce by (3.20) that

$$f^*(\omega)=\int_0^1 \text{Pont } F^G(d\gamma_t,.)+dc(\hat{A},\hat{A}). \tag{3.29}$$

This proves the lemma. ◇

Let us recall how we construct a central extension of $L_{\text{fin}}G$ from the cocycle c.[37,14]

We will make in this paper the following hypothesis:

Hypothesis H.1: $L_{\text{fin}}G$ is simply connected.

We consider the set of path starting from the unit path in $L_{\text{fin}}G$. Let l_t be such a path. We consider a product of $(l_.,\alpha)$ where $\alpha\in S^1$. We endow the set of (l,α) with the group structure:

$$(l_.,\alpha)(l',\beta)=(ll',c(l,l')\alpha\beta), \tag{3.30}$$

where $c(l,l')=\exp[2\pi i\int_B c]$. The integral of c is taken over any surface bounded by the triangle in $L_{\text{fin}}G$ with vertices $(1,1l'_1,l_1l'_1)$ and edges $(l'_t,l_tl'_1)$ and $l_tl'_t$, the last one being circled in the opposite direction. Under the hypothesis H.1, we can find such surface. We identify $(l_.,\alpha)$ and $(l'_.,\alpha')$ modulo a normal subgroup. This one is constituted of (l,α) where l is a loop in $L_{\text{fin}}G$ and $\alpha=\exp[-2\pi i\int_B c]$ where B is a surface bounded by l. c being integral valued, the integral over the surface depends only on its boundary modulo an integer. We get by this procedure a central extension $\tilde{L}_{\text{fin}}G$ of $L_{\text{fin}}G$.

If we consider the diagram (3.5) for the finite energy loop, we get a bundle Q_{fin} over $L_{\text{fin}}M$, the based loop space of loops of finite energy in M. We make the following hypothesis:

Hypothesis H.2: Q_{fin} is simply connected.

This will allow us to construct a formal S^1 bundle over Q by starting from a given closed form of degree 2 with integral values over Q_{fin}.

The last hypothesis we will give is the following:

Hypothesis H.3: Pont F^G gives a trivial cohomology class.

This will allow us to construct a closed form with integral value over Q_{fin} (see Ref. 13). Namely, if Pont F^G gives a trivial cohomology class, we have

$$\text{Pont } F^G=d\mu. \tag{3.31}$$

Over the finite energy based loop space $L_{\text{fin}}G$,

$$f^*\omega=d\int_0^1 \mu(d\gamma_s,.)+dc(\hat{A},\hat{A})=d\nu. \tag{3.32}$$

Namely, we have

$$d\int_0^1 \mu(d\gamma_s,.)=\int_0^1 d\mu(d\gamma_s,.), \tag{3.33}$$

since we work over the based loop space (We refer to Ref. 30 for the rules of derivation of Chen forms). Equation (3.32) remains true for the stochastic loop space.

We follow Ref. 13. Let us put

$$F_{Q_{\text{fin}}}=f^*c-p^*\nu. \tag{3.34}$$

Over Q_{fin}, $dF_{Q_{\text{fin}}}=0$, but $F_{Q_{\text{fin}}}$ has no integral values. The Z homology of $L_{\text{fin}}M$ is finitely generated. Since $L_{\text{fin}}G$ is simply connected, we can find a lifting of the finite numbers of cycle which generate the Z homology of $L_{\text{fin}}M$. Moreover, the Hochschild cohomology is equal to the cohomology of the based loop space.[31] We can perturb ν in (3.34) by a finite combination of Chen forms of degree 2 such that $F_{Q_{\text{fin}}}$ in (3.34) has integral values over Q_{fin}. The number of string structures is related to the number of deformations by Chen forms of $F_{Q_{\text{fin}}}$. A perturbation by a

Chen form instead of a general form as in Ref. 13 is useful in so far as we can integrate Chen forms over random surfaces as it was done in the first part. We will still call F_Q the perturbated Z-valued form.

Let us recall what Ref. 13 does in order to construct the lift $\tilde{Q}_{\text{fin}}$ of Q_{fin}: they consider the space of path PQ_{fin} starting from a based point in Q_{fin}. Over $PQ_{\text{fin}} \times S^1$ [a typical element is denoted by $(q_{.}, \alpha)$], there is an action of $PL_{\text{fin}}G \times S^1$, where $PL_{\text{fin}}G$ is the set of path in $L_{\text{fin}}G$ starting from the unit path. A typical element of this last one is denoted by $(l_{.}, \alpha)$. The action is defined as follows:

$$(q_{.}, \alpha_{.})(l_{.}, \alpha') = (ql, c(q,l)\alpha\alpha'). \tag{3.35}$$

The expression $c(q,l)$ is defined as follows:

$$c(q,l) = \exp\left[2i\pi \int F_Q\right]. \tag{3.36}$$

We omit writing the distinction between F_Q and $F_{Q_{\text{fin}}}$. The integral is taken over any surface B filling in the triangle with vertices $(q_0, q_o l_1, q_1 l_1)$ and edges $(q_0 l_t, q_t l_1)$ and $(q_t l_t)$, the edge $q_t l_t$ being circled in the opposite sense. We identify (q,α) and (q', α') if q and q' have the same point and if

$$\alpha' = c(q,q')\alpha, \tag{3.37}$$

where $c(q,q') = \exp[2\pi i \int F_Q]$. (see Refs. 14 and 37 for analogous considerations). The integral is taken over a surface B whose boundary is the loop constituted from q' circled in the positive direction and q circled in the negative direction. We use the fact that F_Q is invariant under the $L_{\text{fin}}G$ action. We construct an S^1 bundle $\tilde{Q}_{\text{fin}}$ over Q_{fin} which is an $\tilde{L}_{\text{fin}}G$ principal bundle. The bundle $\tilde{Q}_{\text{fin}}$ over Q_{fin} is an S^1 bundle with a connection whose curvature is F_Q modulo the normalizing term $2\pi i$. There is a tower of bundles: $\tilde{Q}_{\text{fin}}$ which is an S^1 bundle over Q_{fin}. Q_{fin} which is an $L_{\text{fin}}G$ bundle over $L_{\text{fin}}(M)$ and $\tilde{Q}_{\text{fin}}$ which is an $\tilde{L}_{\text{fin}}G$ bundle over $L_{\text{fin}}(M)$.

The purpose of the remaining part of this paper is to construct a formal $\tilde{L}_{\text{fin}}G$ bundle over the stochastic loop space $L(M)$ which extends in some sense the bundle $\tilde{Q}_{\text{fin}}$ over $L_{\text{fin}}(M)$. For that, we have to construct over Q the analogous parts of the distinguished curves and of the distinguished surfaces of the first part. It is the purpose of a connection to lift curves over a bundle. By introducing a suitable connection over Q, we will produce a system of transition functions with values in $\tilde{L}_{\text{fin}}G$, which satisfy almost surely (2.25) and (2.26). But these transition functions cannot be defined over a system of open sets for the uniform distance over $L(M)$, because they depend on τ_1^G which is only almost surely defined.

Let γ_i be a countable set of finite energy loops such that the union of the open balls of radius δ for the uniform norm $B(\gamma_i, \delta)$ is equal to LM. If $\gamma_{.}$ belongs to $B(\gamma_i, \delta)$, there is a curve $l_t(\gamma_i, \gamma_{.})$ joining γ_i to $\gamma_{.}$. The parallel transport $\tau_1^G(l_t(\gamma_i, \gamma_{.}))$ depends almost surely on t on this curve. Therefore, starting from an element $q(\gamma_{.})$ over $\gamma_{.}$ in Q, we use the parallel transport for the connection ∇^G and we deduce a curve $q_t(\gamma)$ over $l_t(\gamma_i, \gamma_{.})$.

Let us introduce another condition over $\gamma_{.}$. It is $(\tau_1^G)^{-1} \in O_j(G)$, a small neighborhood of G. If this condition is satisfied, we have a slice $g^j(\gamma_{.})$ of Q which depends smoothly on $\gamma_{.}$. Of course this dependance is only local. However, it is the restriction of a smooth application which takes its values in the smooth manifold $P_{\text{fin}}R^{d'}$, by imbedding M into R^d and G into $R^{d'}$. As in the first part, we can regularize the indicator function of $B(\gamma_i, \delta) \cap \{(\tau_1^G)^{-1} \in O_j(G)\}$. The indicator function of the first event is regularized as in the first part. The indicator function of the second event can be easily regularized, because it depends only on (τ_1^G). In particular, we can find a sequence of smooth functions $G_{i,j}$ with support in the event $B(\gamma_i, \delta) \cap \{(\tau_1^G)^{-1} \in O_j(G)\}$ which tends simply to the indicator function of this event and with support in this event.

Let g_k be a family of elements of $L_{\mathrm{fin}}G$ such that the set of open balls $B(g_k,\delta)$ for the uniform distance constitutes a cover of $L_{\mathrm{fin}}G$. We have a cover of Q by the set of $g^j(\gamma_.)g$ for $\gamma \in B(\gamma_i,\delta)\cap\{(\tau_1^G)^{-1}\in O_j(G)\}$ and $g\in B(g_k,\delta)$.

In order to simplify the exposure, we will call only by one parameter i this family of subsets Q_i of Q.

We will define a set of transition functions $\rho_{i,j}$ which will define a formal circle bundle $\tilde{Q}$ over Q.

This means that almost surely in $\gamma\in B(\gamma_i,\delta)\cap\{(\tau_1^G)^{-1}\in O_i\}$, for all $g\in B(g_i,\delta)$, we have almost surely the following properties:

$$\rho_{i,j}\rho_{j,i}=1 \tag{3.38}$$

over $Q_i\cap Q_j$ and over $Q_i\cap Q_j\cap Q_k$

$$\rho_{i,j}\rho_{j,k}\rho_{k,i}=1. \tag{3.39}$$

The topological space $\tilde{Q}$ is not defined, because these transition functions are only almost surely defined over the basis LM. But this system will allow us to define another system of transition functions $\rho_{i,j}$ over $O_i\cap O_j$ for a suitable cover O_i of LM. It is a cover of $L(M)$ modulo a set of measure 0 and it satisfies almost surely

$$\rho_{i,j}\in\tilde{L}_{\mathrm{fin}}G, \tag{3.40}$$

$$\rho_{i,j}\rho_{j,i}=1 \tag{3.41}$$

over $O_i\cap O_j$ and

$$\rho_{i,j}\rho_{j,k}\rho_{k,i}=1 \tag{3.42}$$

over $O_i\cap O_j\cap O_k$.

Let us suppose that this last point is satisfied. Let us suppose that there is a unitary representation Spin of $\tilde{L}_{\mathrm{fin}}G$, which is a Hilbert space (see Ref. 26 and 1). The transition functions allow to define the space of measurable sections of the bundle Spin.

Definition III.2: Let ψ_i be a system of random variables over O_i with value in Spin. This system defines a section of the formal bundle Spin over LM if over $O_i\cap O_j$:

$$\psi_j=\rho_{j,i}\psi_i. \tag{3.43}$$

If ψ_i is an element of Spin over O_i, we will define its norm $\|\psi_i\|^2$, which is compatible with the transition functions. Namely the representation of $\tilde{L}_{\mathrm{fin}}G$ over Spin is supposed unitary.

Definition III.3: The space of L^p sections of Spin is the space of measurable sections ψ such that

$$\|\|\psi\|\|_{L^p}<\infty. \tag{3.44}$$

Let $(\gamma_.,g^i(\gamma_.)g)$ be in Q_i. There is a distinguished path joining g to g_j, the center of the ball for the uniform distance in $L_{\mathrm{fin}}G$ which constitutes the vertical part of Q_i. After this, there is a distinguished path joining $\gamma_.$ to γ_i, the center of the ball in the basis LM which constitutes the basical part of Q_i. We parallel transport by $\nabla^G g^i(\gamma_.)g_i$ to $\tilde{g}^i(\gamma_.)g_i$, which is over γ_i. The element $\tilde{g}^i(\gamma_.)g_i$ depends smoothly on $\gamma_.$. The main problem is that $\tilde{g}^i(\gamma_.)g_i$ is not $g^i(\gamma_i)g_i$. In order to be able to choose a nice distinguished curve joining these two elements over γ_i, we have to impose another condition over Q_i. The element $\tilde{g}^i(\gamma_.)g_i$ in the fiber of γ_i remains in a small tubular neighborhood of $P_{\mathrm{fin}}G$. This can be done if we assume, for instance, that the process of the holonomy τ_1^G over the distinguished path joining γ to γ_i remains in a small tubular neighborhood of $P_{\mathrm{fin}}G$. We use the fact that Q is simply connected. There is a distinguished path $l_{i,Q}(q)$ joining

q in Q_i to q_{base} in Q, a reference element of Q. When we have arrived at $g^i(\gamma_i)g_i$, we go by a deterministic path to q_{base}, a reference element in Q. This path depends smoothly on q. In the fiber direction, it depends smoothly on q in the traditional sense. In the basis direction, it depends smoothly on q in the generalized sense of Ref. 5 or 18.

Let us choose a distinguished surface bounded by the loop constituted of $l_{i,Q}(q)$ and of $l_{j,Q}(q)$ if $q \in Q_i \cap Q_j$. We should be able to integrate F_Q over this distinguished surface by using the theory of stochastic integrals.

Let us recall some results of the first part. There is a distinguished path joining $p(l_{i,Q,t}(q))$ to $p(l_{j,Q,t}(q))$ for t smaller than the time which is used in order to join $\tilde{g}^i(\gamma_\cdot)(g_i)$ to $g^i(\gamma_i)(g_i)$. We parallel transport $l_{i,Q,t}(q)$ over this distinguished path, and we find at the end over $p(l_{j,Q,t}(q))$ a curve $\tilde{l}_{j,Q,t}(q)$. Moreover, there is a deterministic path joining $(\gamma_i, g^i(\gamma_i)(g_i))$ to $(\gamma_j, g^j(\gamma_j) \times (g_j))$. Namely, Q_{fin} is supposed simply connected. Therefore, we get a surface whose boundary differs of the $l_{j,Q,t}(q)$ and the path joining $(\gamma_i, g^i(\gamma_i)g_i)$ and $(\gamma_j, g^j(\gamma_j)g_j)$ by a loop in $L_{\text{fin}}G$. We take any vertical surface $l_{u,v}(\gamma_\cdot)$ in $L_{\text{fin}}(G)$ bounded by this loop. We get a surface $B_{i,j}(q)$ in Q bounded by the loop constituted of $l_{i,Q,t}(q)$ and of $l_{j,Q,t}(q)$. Namely we can fill in the triangle constituted of the path joining q_{base} to γ_i, $g^i(\gamma_i)g_i$, the path joining $(\gamma_i, g^i(\gamma_i)(g_i))$ to $(\gamma_j, g^j(\gamma_j)g_j)$ by a surface, because Q_{fin} is simply connected.

In the integral of F_Q over $B_{i,j}(q)$, there are two parts:

(i) The part involved with integral in the basis. Since $c(\hat{A},\hat{A})$ is a stochastic integral, and since we have perturbed ν by a Chen form, this contribution can be treated as in the first part.

(ii) The part which is involved with the fiber integration, which is a traditional integration, does not involve any stochastic integration. In particular, we can take $l_{u,v}(\gamma_\cdot)$ arbitrary: since F_Q is integral, it does not change the value of the transition function.

Since we can take $l_{u,v}(\gamma)$ arbitrary, this shows us that $(\gamma,g) \to \rho_{i,j}$ is smooth in $(\gamma_\cdot, g)$. The transition function $\rho_{i,j}$ has a similar shape as in the first part. As a matter of fact, we are working over Q instead of over $L(M)$. The derivative in g is in the normal sense for the Hilbert manifold $L_{\text{fin}}G$, and the derivative in γ is in the general sense of Ref. 18. In particular, for almost all $\gamma_\cdot$, $g \to \rho_{i,j}(\gamma,g)$ is smooth. We localize the smoothness assumption by using a regularized version of the indicator function of Q_i. It is a regularization in the Sobolev sense in $\gamma_\cdot$, and a regularization in the normal sense in $g \in L_{\text{fin}}(G)$. In order to do this operation, we have chosen the slice $\gamma \to g^i(\gamma_\cdot)$. The regularization of a tubular neighborhood of $P_{\text{fin}}G$ is performed as in the first part. It is not clear that we get a smooth function when we compose with $\tilde{g}^i(\gamma_\cdot)$, but it has a derivative in the generalized sense which is locally integrable.

In particular, if $(\gamma, g^i(\gamma_\cdot)) \in Q_i$, and if we consider a polygonal approximation $\gamma^n_\cdot$ of $\gamma_\cdot$, $(\gamma^n, g^i(\gamma^n)g) \in Q_i$ for n big enough. We get a transition function $\rho_{i,j}(\gamma^n, g)$ which tends to $\rho_{i,j}(\gamma_\cdot, g)$ (see Ref. 31) almost surely.

But over finite energy path, we have (3.38) and (3.39) surely, if we do the same procedure to construct the transition functions. In particular, it is true for $\rho_{i,j}(\gamma^n)$. We deduce that (3.38) and (3.39) are almost surely true when we consider a fixed g. But $\rho_{i,j}$ depends in a smooth way of g in the fiber. The topological space $L_{\text{fin}}G$ has a countable dense set of elements. We deduce that (3.38) and (3.39) are almost surely true in the following sense: it is almost surely in $\gamma_\cdot$ over $p(Q_i) \cap p(Q_j)$, for all $q \in p^{-1}(\gamma)$ in $Q_i \cap Q_j$, that we have (3.38). There is the same statement for (3.39).

Let us now state (3.40)–(3.42). We are working over $\tilde{Q}_{\text{fin}}$. Since there is an action of $\tilde{L}_{\text{fin}}G$ which preserves the projection $\tilde{p}$, it is enough to find local slices of $\tilde{Q}_{\text{fin}}$ over $L_{\text{fin}}M$.

It is possible to find such slices for the continuous loop space.

The chart $O_i \subset LM$ is defined by the three following conditions:

(i) $\gamma_\cdot$ is closed of a given energy curve γ_i.

(ii) $(\tau_1^G)^{-1}(\gamma)$ is in a small neighborhood of G.

(iii) The parallel transport for ∇^G of $g^i(\gamma_\cdot)$ over the distinguished curve given in the first part which joins $\gamma_\cdot$ to γ_i is in a small tubular neighborhood of $L_{\text{fin}}G$ for the uniform distance.

Over O_i, we can find a formal slice $\tilde{q}_i(\gamma_\cdot)$ of the formal line bundle $\tilde{Q}$. Namely the parallel transport of $g^i(\gamma)$ over the distinguished curve joining γ_i to γ remains in a small neighborhood of $L_{\text{fin}}G$. There is, since Q_{fin} is simply connected, a distinguished curve joining $q_{\text{base}} \in Q$ to $(\gamma, g^i(\gamma))$. We will denote it by $q_{i,\cdot}(\gamma)$ and we put $\tilde{q}_i(\gamma) = (q_i, (\gamma), 1)$.

Let us work over $O_i \cap O_j$. Let us show that almost surely

$$(q_i, .(\gamma_\cdot), 1) = (q_j, .(\gamma_\cdot), 1)\rho_{j,i}(\gamma_\cdot). \tag{3.45}$$

The quantity $\rho_{j,i} \in \tilde{L}_{\text{fin}}G$ has an action given by (3.35).

The quantity $q_{i,1}(\gamma)$ is the path $g_t^i((\tau_1^G))$ over $\gamma_\cdot$ and $q_{j,1}(\gamma)$ is the path $g_t^j((\tau_1^G)^{-1})$ over $\gamma_\cdot$. Moreover, τ_1^G over O_i remains in a small neighborhood of G and τ_1^G over O_j remains in a small ball of G. There is, therefore, a distinguished path in $L_{\text{fin}}G$, smooth in τ_1^G joining $g_t^i((\tau_1^G)^{-1})$ to $g_t^i((\tau_1^G)^{-1})$ by right multiplication. Moreover, this path remains in a small tubular neighborhood of $L_{\text{fin}}G$. Let us call it $l_{j,i,1}(\gamma_\cdot)$. Since it remains in a small neighborhood of $L_{\text{fin}}G$, we can find a path $l_{j,i,t}(\gamma_\cdot)$ joining the unit path in $L_{\text{fin}}G$ to $l_{j,i,1}(\gamma_\cdot)$, and which is smooth in γ. Let us compute by the rule (3.35) and (3.36) the quantity

$$(q_{j,\cdot}(\gamma), 1)(l_{j,i,\cdot}(\gamma), 1) = (q_{j,\cdot}(\gamma) l_{j,i,\cdot}(\gamma), \alpha_{j,i}). \tag{3.46}$$

We have to find a distinguished suitable surface which depends on a nice way of $\gamma_\cdot$. Its vertices are given by $(q_{j,0}(\gamma), q_{j,0}(\gamma) l_{j,i,1}(\gamma), q_{j,1}(\gamma) l_{j,i,1}(\gamma))$. Its edges are given by $(q_{j,0}(\gamma) l_{j,i,s}(\gamma), q_{j,s}(\gamma) l_{j,i,1}(\gamma), q_{j,s}(\gamma) l_{j,i,s}(\gamma))$, the last one being circled in the opposite sense. We do as before in order to produce this surface by filling in first the loop over the basis LM as it was done in the first part, and afterwards by using the connection ∇^G and using the third condition in the definition of O_i. We use the parallel transport over the natural path joining $pq_{j,0}(\gamma_\cdot) l_{j,i,s}(\gamma_\cdot)$ to $pq_{j,s}(\gamma_\cdot) l_{j,i,s}(\gamma_\cdot)$. An extra loop in $L_{\text{fin}}G$ appears. We find an arbitrary surface in $L_{\text{fin}}G$ whose boundary is this extra loop. Two parts in the integral of F_Q over the global surface are obtained. The basic part leads to a nonanticipative stochastic integral. The fiber part is a traditional integral and is smooth. Namely, the surface filling in the extra loop which appears in $L_{\text{fin}}G$ can be chosen arbitrarily. We deduce that $\alpha_{j,i}$ is smooth. More precisely, $G_i G_j \alpha_{j,i}$ belongs locally to all the Sobolev spaces, where G_i is a regularization of the indicator function of the same type of the previous regularizations.

We conclude that

$$(q_{j,\cdot}, 1)(l_{j,i,\cdot}, \beta_{j,i}) \equiv (q_{i,\cdot}, 1), \tag{3.47}$$

where the $\beta_{j,i}$ are smooth in the previous sense. Namely, we have to identify $(q_{i,\cdot}(\gamma_\cdot), 1)$ to $(q_{j,\cdot} l_{j,i}, (\gamma_\cdot), \beta_{i,j})$. We have to find a nice surface which joins $q_{i,\cdot}(\gamma_\cdot)$ to $q_{j,\cdot}(\gamma_\cdot) l_{j,i,\cdot}(\gamma_\cdot)$ where we can integrate F_Q by using the theory of stochastic integrals. As in the first part, there is a distinguished surface bounded by the loop $p(q_{i,\cdot}(\gamma)$ to $p(q_{j,\cdot}(\gamma_\cdot) l_{j,i,\cdot}(\gamma_\cdot))$. We parallel transport $q_{i,t}(\gamma_\cdot)$ over the path joining $p(q_{i,t}(\gamma_\cdot))$ to $p(q_{j,t}(\gamma_\cdot))$. We find $\tilde{q}_{j,t}(\gamma_\cdot)$. But $\tilde{q}_{j,1}(\gamma_\cdot)$ is different from $q_j(\gamma_\cdot)$, but it differs by a loop in $L_{\text{fin}}G$. An arbitrary surface in $L_{\text{fin}}G$ is bounded by this loop, because the third condition is checked in the definition of O_i.

We get a global random surface over Q. They are two parts in the integral of F_Q over this surface. The basic part leads to a nonanticipative stochastic integral. The fiber part is a traditional integral, which is smooth. Namely, the surface bounded by the extra loop which arises in $L_{\text{fin}}G$ can be chosen arbitrarily.

Moreover, if $\gamma \in O_i$, $\gamma^n \in O_i$ for a polygonal approximation of γ. $\rho_{i,j}(\gamma^n)$ tends almost surely to $\rho_{i,j}(\gamma)$. Since over $\tilde{Q}_{\text{fin}}$ the $\rho_{i,j}$ are transition functions, we have (3.41) and (3.42) surely for $\rho_{i,j}(\gamma^n)$; (3.41) and (3.42) are true almost surely for $\rho_{i,j}(\gamma_\cdot)$.

Remark: If we look at the formalism of the Definition I.2, we have immediately $g_{j,i} = (g^j_{\cdot}((\tau_1 G)^{-1}))^{-1} g^i_{\cdot}((\tau_1^G)^{-1})$, and $\rho_{j,i} = (l_{j,i}, \beta_{j,i})$ checks clearly the condition of Definition I.2.

[1] A. Pressley and G. Segal, *Loop Groups* (Oxford U.P., Oxford, 1986).
[2] P. Gilkey, *Invariance theory, the heat equation and the Atiyah-Singer theorem*. Math. Lect. Series. 11. (Publish or Perish, unpublished).
[3] G. Segal, ''Elliptic cohomology,'' Séminaire Bourbaki. Exposé 695. Astérisque 161–162., 187–201 (1988).
[4] E. Witten, ''The index of the Dirac operator in loop space,'' in *Elliptic Curves and Modular Forms in Algebraic Topology*, Landweber Lecture Notes in Mathematics, Vol. 1326 (Springer, Berlin, 1988), pp. 161–181.
[5] R. Léandre, ''Integration by parts formulas and rotationally invariant Sobolev Calculus on the free loop space,'' J. Geom. Phys. **11**, 517–528 (1993).
[6] J. D. S. Jones and R. Léandre, ''A stochastic approach to the Dirac operator over the free loop space,'' in *Loop Spaces and Groups of Diffeomorphisms*, Proceedings of the Steklov Institute 217 (1997), pp. 253–282.
[7] R. Léandre and S. S. Roan, ''A stochastic approach to the Euler-Poincaré number of the loop space of a developable orbifold,'' J. Geom. Phys. **16**, 71–98 (1995).
[8] R. Léandre, ''Brownian motion over a Kahler manifold and elliptic genera of level *N*,'' in *Stochastic Analysis and Applications in Physics*, edited by R. Sénéor and L. Streit Nato ASI Ser. 449 (1994), pp. 193–217.
[9] R. Léandre, ''Stochastic Wess-Zumino-Witten model over a symplectic manifold,'' J. Geom. Phys. **21**, 307–336 (1997).
[10] R. Bott and L. W. Tu, *Differential Forms in Algebraic Topology* (Springer, Berlin, 1986).
[11] T. Killingback, ''World-sheet anomalies and loop geometry,'' Nucl. Phys. B **288**, 578–588 (1987).
[12] D. MacLaughlin, ''Orientation and string structures on loop spaces,'' Pac. J. Math. **155**, 143–156 (1992).
[13] A. L. Carey and M. K. Murray, ''String structure and the path fibration of a group,'' C.M.P. **141**, 441–452 (1991).
[14] R. Coquereaux and K. Pilch, ''String structure on loop bundles,'' C.M.P. **120**, 353–378 (1989).
[15] L. Gross, ''Potential theory on Hilbert spaces,'' J.F.A. **1**, 123–181 (1967).
[16] J. M. Bismut, ''Large deviations and the Malliavin Calculus,'' Progress in Math. **45**. (Birkhauser, 1984).
[17] J. Jones and R. Léandre, ''L^p Chen forms over loop spaces,'' in *Stochastic Analysis*, edited by M. Barlow and N. Bingham (Cambridge U.P., Cambridge, 1991), pp. 104–162.
[18] R. Léandre, ''Invariant Sobolev Calculus on the free loop space,'' Acta Applicandae Mathematicae. **46**, 267–350 (1997).
[19] R. Léandre and J. Norris, ''Integration by parts and Cameron-Martin formulas for the free path space of a compact Riemannian manifold,'' Séminaire de Probabilités XXXI, Lecture Notes in Mathematics 1655, edited by J. Azema, M. Emery, and M. Yor (1997), pp. 16–24.
[20] R. Léandre, ''Cohomologie de Bismut-Nualart-Pardoux et cohomologie de Hochschild entiere,'' Séminaire de Probabiliés XXX in honour of P. A. Meyer et J. Neveu, Lecture Notes in Mathematics 1626. edited by J. Azéma, M. Emery, and M. Yor (1996), pp. 68–100.
[21] R. Léandre, ''Brownian cohomology of an homogeneous manifold,'' in *Proceedings of the Taniguchi Conference*, edited by K. D. Elworthy, S. Kusuoka, and I. Shigekawa (World Scientific, Singapore, 1997), pp. 305–348.
[22] S. Aida, S. Kusuoka, and D. W. Stroock, ''On the support of Wiener functionals, in *Asymptotics Problems in Probability Theory: Wiener Functionals and Asymptotics*, edited by D. Elworthy and N. Ikeda, Longman Scientific 284 (1993), pp. 3–35.
[23] G. Ben-Arous and R. Léandre, ''Décroissance exponentielle du noyau de la chaleur sur la diagonale II,'' P.T.R.F. **90**, 377–402 (1991).
[24] R. Léandre, ''Strange behaviour of the heat kernel on the diagonal,'' in *Stochastic Processes, Physic and Geometry*, edited by S. Albeverio (World Scientific, Singapore, 1990), pp. 516–528.
[25] J. L. Brylinski, ''Loop spaces, characteristic classes and geometric quantization,'' Prog. Math. **107**, (1992).
[26] D. Borthwick, ''The Pfaffian line bundle,'' C.M.P. **149**, 463–493 ().
[27] R. Léandre, ''Hilbert space of spinors fields over the free loop space,'' Rev. Math. Phys. **9(2)**, 243–277 (1997).
[28] R. Léandre, ''Stochastic gauge transform of the string bundle,'' J. Geom. Phys. **26**, 1–25 (1998).
[29] M. Emery, ''On two transfer principles in stochastic differential geometry,'' Séminaire de Probabilités XXIV, Lecture Notes in Mathematics, Vol. 1426, edited by J. Azéma, P. A. Meyer, and M. Yor (1990), pp. 407–422.
[30] E. Getzler, J. D. S. Jones, and S. Petrack, ''Differential forms on loop spaces and the cyclic bar complex,'' Topology **30**, 339–371 (1991).
[31] J. F. Adams, ''On the cobar construction,'' Proc. Natl. Acad. Sci. USA **42**, 346–373 (1956).
[32] R. Léandre, ''Singular integral homology of the stochastic loop space,'' Infinite Dimensional Analysis, Quantum Probability and related topics **1(1)**, 17–31 (1998).
[33] J. M. Bismut, *Mécanique Aléatoire*, Lecture Notes in Mathematics 866. (Springer, Berlin, 1981).
[34] P. Malliavin, *Stochastic Analysis* (Springer, Berlin, 1997).
[35] S. Kobayashi and S. Nomizu, *Foundations of Differential Geometry. I.* (Interscience, New York, 1969).
[36] J. Norris, ''Covariant stochastic calculus and applications to heat kernels,'' P.T.R.F. **94**, 525–541 (1993).
[37] M. K. Murray, ''Another construction of the central extension of the loop group,'' C.M.P. **116**, 73–80 (1988).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 1 JANUARY 1999

A canonical realization of the BMS algebra

G. Longhi[a)] and M. Materasso
Department of Physics, University of Firenze, 50125 Firenze, Italy

(Received 8 December 1997; accepted for publication 18 August 1998)

A canonical realization of the BMS (Bondi–Metzner–Sachs) algebra is given on the phase space of the classical real Klein–Gordon field. By assuming the finiteness of the generators of the Poincaré group, it is shown that a countable set of conserved quantities exists (supertranslations); this set transforms under a particular Lorentz representation, which is uniquely determined by the requirement of having an invariant four-dimensional subspace, which corresponds to the Poincaré translations. This Lorentz representation is infinite-dimensional, nonunitary, reducible and indecomposable. Its representation space is studied in some detail. It determines the structure constants of the infinite-dimensional canonical algebra of the Poincaré generators together with the infinite set of the new conserved quantities. It is shown that this algebra is isomorphic with that of the BMS group. © *1999 American Institute of Physics.* [S0022-2488(99)00301-1]

I. INTRODUCTION

Our purpose in this paper is to give a canonical realization of the algebra of the Bondi–Metzner–Sachs (BMS) group[1] on the phase space of the real classical Klein–Gordon field.

The BMS group arises in general relativity as the asymptotic symmetry at null infinity of a space–time describing an isolated (radiating) gravitational source. Such space–time is expected to become flat at null infinity,[2] and to exhibit the Poincaré group as the group of asymptotic symmetry.

While an asymptotic symmetry exists, its group is not the Poincaré group. It is a larger group containing the Poincaré group as a subgroup, and is a semidirect product of the homogeneous Lorentz group with an infinite-dimensional Abelian group. Among the infinitely many generators of this Abelian group there are the generators of translations; the remaining are the so-called supertranslations.

The idea of asymptotic flatness was formalized by Penrose by a process of conformal compactification;[3] by means of this process a boundary (scri) was added to space–time, consisting of end points of null geodesics. In this way the notion of flatness acquired an intrinsic meaning.

The asymptotic symmetry at null infinity has been studied by several authors. Apart from the classic papers[1] we also quote some review articles.[4] Here we only mention some points discussed in the literature on this argument.

One problem with the BMS group is due to the presence of infinitely many copies of the homogeneous Lorentz group, obtained from one another by conjugation with a supertranslation. This happens even in the case of the Poincaré group, but the spin Casimir of the Poincaré group is not invariant under such a conjugation.[5] This causes difficulties in the definition of a unique Poincaré group, and, by consequence, in the definition of the angular momentum of the system[6] (see, however, Ref. 7).

The unitary and irreducible representations of the BMS group have been studied in Ref. 8 to give a possible interpretation of this group as a fundamental group of elementary particles, a line of research no longer pursued.

There is a situation analogous to that of null infinity in the case of spatial infinity, where a similar (bigger) asymptotic symmetry arises.[9]

[a)]Electronic mail: longhi@fi.infn.it

More recently the classical approach to null infinity was criticized,[10] and it was shown that a more general asymptotic expansion for the metric should be used (polyhomogeneous expansion). Nevertheless, the group of asymptotic symmetry at null infinity is still the BMS group.

In view of the role played by the BMS group in general relativity it seems of interest to show that its algebra can be canonically realized in terms of conserved quantities of a real classical Klein–Gordon field. While this field has been chosen for its simplicity we expect that a similar analysis could be done for any other classical field. In particular, we will give explicitly the underlying infinite-dimensional representation of the Lorentz algebra, which determines the structure constants of the BMS algebra. Indeed we will show that, besides the total energy-momentum P^μ of the Klein–Gordon field, it is possible to define a countable set of other conserved quantities, with vanishing Poisson brackets with P^μ and among themselves. This new set of conserved quantities is given in terms of a countable set of functions on momentum space, which are the basis for an infinite-dimensional representation of the Lorentz group, which is nonunitary, reducible and indecomposable.

The structure of this representation function space is the same as that of the translations and supertranslations of the BMS group. We show that the Poisson algebra of the conserved quantities built from these functions and the generators of the Lorentz group is the same as that of the BMS group. So we call these conserved quantities the generators of supertranslations.

The action of the Lorentz generators on the supertranslations defines a set of structure constants of the BMS algebra, *which are the matrix elements* of the infinite-dimensional representation of the Lorentz group. The action of the supertranslations on the field is nonlocal. Nevertheless it defines, as in the case of P^μ, a canonical transformation on the field.

In Sec. II we recall some definitions about the Klein–Gordon field and define the Laplace–Beltrami operator on the mass hyperboloid. The eigenfunctions of this operator are studied in Sec. III and, in Sec. IV, the infinite representation of the Lorentz algebra is studied in some detail. In Sec. V the Poisson algebra of the generators of supertranslations is given and it is shown that it is isomorphic with the algebra of the BMS group.

The definitions and notations for the Klein–Gordon field are given in Appendix A. In Appendix B we discuss the eigenfunction problem for the Laplace–Beltrami operator on the mass hyperboloid. Moreover, we study in some detail the representation of the Lorentz group, which is the basis for the definition of the supertranslation generators.

II. THE KLEIN–GORDON FIELD AND THE LAPLACE–BELTRAMI OPERATOR

The notations and definitions for the real Klein–Gordon field $\Phi(\mathbf{x},t)$ are given in Appendix A. For each $t\in R$ the field $\Phi(\mathbf{x},t)$ will be supposed to belong, together with its spatial and temporal derivatives, to $L^2(R^3)$, that is

$$(\Phi(\cdot,t),\dot{\Phi}(\cdot,t))\in H^1(R^3)\times L^2(R^3), \tag{2.1}$$

where $\dot{\Phi}=\partial\Phi/\partial t$ and where $H^1(R^3)$ is the $W_2^1(R^3)$ Sobolev space.[11] This implies the existence of the total energy–momentum,

$$P^\mu=\int d^3x\left[\dot{\Phi}(\mathbf{x},t)\partial^\mu\Phi(\mathbf{x},t)-\frac{1}{2}\eta^{\mu 0}(\partial_\nu\Phi(\mathbf{x},t)\partial^\nu\Phi(\mathbf{x},t)-m^2\Phi^2(\mathbf{x},t))\right], \tag{2.2}$$

or, in terms of the Fourier coefficients of the field Φ,

$$P^\mu=\int \widetilde{d}k\, k^\mu\overline{a}(\mathbf{k})a(\mathbf{k}), \tag{2.3}$$

where the measure $\widetilde{d}k$ is defined in Appendix A and $k^0=\omega(\mathbf{k})$. The energy–momentum P^μ is time-like and future oriented.

Apart from the existence of a well defined total momentum P^{μ}, we require the existence of the generators of the Lorentz group. This implies a more stringent condition on Φ, or on $a(\mathbf{k})$. We require the additional condition: $\nabla a(\mathbf{k}) \in L^2(R^3)$. From Eq. (2.3), with $\mu=0$, we have that $a(\mathbf{k}) \in L^2(R^3)$; with this additional condition we have that $a(\mathbf{k})$ is continuous and that has a vanishing limit when $|\mathbf{k}| \to \infty$.[12]

The scalar field $\Phi(x)$, where $x \equiv (\mathbf{x},t)$, transforms under a Poincaré transformation $U(\Lambda,\alpha)$ as

$$(U(\Lambda,\alpha)\Phi)(x)=\Phi(\Lambda^{-1}(x-\alpha)). \tag{2.4}$$

This induces the transformation on $a(\mathbf{k})$,

$$(U(\Lambda,\alpha)a)(\mathbf{k})=a(\Lambda^{-1}\mathbf{k})e^{i(k\cdot\alpha)}. \tag{2.5}$$

In this last equation the notation $\Lambda^{-1}\mathbf{k}$ has the meaning $\Lambda^{i}{}_{\cdot\nu}k^{\nu}, i=1,2,3$, where $k^0 = \sqrt{m^2+\mathbf{k}^2}$.

The canonical action of the Poincaré generators on $a(\mathbf{k})$ is given by (see Ref. 11)

$$\{P^{\mu},a(\mathbf{k})\}=ik^{\mu}a(\mathbf{k}), \tag{2.6}$$

$$\{M'^{\mu\nu},a(\mathbf{k})\}=D_{\mu\nu}a(\mathbf{k}), \quad \{M'^{\mu\nu},\overline{\mathbf{a}}(\mathbf{k})\}=D_{\mu\nu}\overline{\mathbf{a}}(\mathbf{k}), \tag{2.7}$$

where $M'^{\mu\nu}$ is defined in Eq. (A21) and

$$D_{\mu\nu}=(\eta^{i}_{\mu}k_{\nu}-\eta^{i}_{\nu}k_{\mu})\frac{\partial}{\partial k^{i}}. \tag{2.8}$$

The differential operators $D_{\mu\nu}$ satisfy the algebra

$$[D_{\mu\nu},D_{\rho\lambda}]=\eta_{\mu\rho}D_{\nu\lambda}+\eta_{\nu\lambda}D_{\mu\rho}-\eta_{\mu\lambda}D_{\nu\rho}-\eta_{\nu\rho}D_{\mu\lambda}. \tag{2.9}$$

We can now work with the field $a(\mathbf{k})$ and its complex conjugate $\overline{a}(\mathbf{k})$, which are defined on the mass hyperboloid. This is a Riemannian manifold which will be called H^1_3, following the notations of Ref. 13, where such manifolds are studied.

Indeed, if we define H^1_3 as the inclusion f of the submanifold $q^2-m^2=0$ with $q^0>0$ in the Minkowski space M^4 (with coordinates q^{μ}), with f defined by

$$f:\{k^i\}\to\{q^{\mu}\}, \tag{2.10}$$

with

$$q^0=\sqrt{m^2+\mathbf{k}^2} \quad \text{and} \quad q^i=k^i \quad (i=1,2,3), \tag{2.11}$$

we get for the induced metric $\hat{\eta}$,

$$\hat{\eta}=f^*\eta, \tag{2.12}$$

the following expression:

$$\hat{\eta}=\hat{\eta}_{ij}\,dk^i\,dk^j, \tag{2.13}$$

with

$$\hat{\eta}_{ij}=\frac{1}{\omega^2(k)}k_ik_j-\delta_{ij} \quad (i,j=1,2,3). \tag{2.14}$$

The inverse of the matrix $\{\hat{\eta}_{ij}\}$ is

$$\hat{\eta}^{ij}=-\left(\delta^{ij}+\frac{k^i k^j}{m^2}\right); \tag{2.15}$$

its determinant is

$$|\hat{\eta}|=\det\{\hat{\eta}_{ij}\}=-\frac{m^2}{\omega^2(\mathbf{k})}; \tag{2.16}$$

and its eigenvalues are $-1,-1,-1+\mathbf{k}^2/\omega^2(\mathbf{k})$.

So the metric $\hat{\eta}$ is proper Riemannian. The manifold H_3^1 is a space-like surface since its normal is

$$n^\mu\equiv\left(\frac{\omega(\mathbf{k})}{m},\frac{\mathbf{k}}{m}\right)\equiv\frac{k^\mu}{m},\quad n^2=1, \tag{2.17}$$

and has constant negative scalar curvature, as can be shown by explicit calculation.

The only invariant second order differential operator on H_3^1 is the Laplace–Beltrami operator Δ (no invariant differential operator of first order exists):

$$\Delta=-\frac{1}{\sqrt{|\hat{\eta}|}}\frac{\partial}{\partial k^i}\hat{\eta}^{ij}\sqrt{|\hat{\eta}|}\frac{\partial}{\partial k^j}, \tag{2.18}$$

where $|\hat{\eta}|=\det\{\hat{\eta}_{ij}\}$.

Explicitly, this operator is

$$\Delta=\frac{\omega(\mathbf{k})}{m}\frac{\partial}{\partial k^i}\left(\delta^{ij}+\frac{k^i k^j}{m^2}\right)\frac{m}{\omega(\mathbf{k})}\frac{\partial}{\partial k^j}, \tag{2.19}$$

or

$$\Delta=\left[\boldsymbol{\nabla}^2+\frac{2}{m^2}\mathbf{k}\cdot\nabla+\frac{1}{m^2}(\mathbf{k}\cdot\nabla)^2\right]. \tag{2.20}$$

This operator is invariant under a Lorentz transformation Λ,

$$k^i\rightarrow k'^i=\Lambda^i_j k^j+\Lambda^i_0\sqrt{m^2+\mathbf{k}^2}. \tag{2.21}$$

The operator Δ is formally self-adjoint with respect to the invariant measure $\widetilde{d}k$. It is an elliptic operator and has the property

$$\Delta k^\mu=\frac{3}{m^2}k^\mu. \tag{2.22}$$

Let us define

$$D=-m^2\Delta+3, \tag{2.23}$$

so that

$$Dk^\mu=0. \tag{2.24}$$

When explicitly written (see the following Section) the equation

$$Df(\mathbf{k})=0, \tag{2.25}$$

is an equation which can be separated in spherical coordinates. The radial equation has three regular singularities in the complex plane of $z=-\mathbf{k}^2/m^2$, which are in $-1,0,\infty$, so it is a hypergeometric equation. The four functions k^μ are a subset of the solutions of Eq. (2.25), with the characteristic exponents $l=0,1$ in the neighborhood of the point 0 $[k^0=\omega(\mathbf{k})$ has the exponent $l=0$ and $\mathbf{k}$ has the exponent $l=1]$. In the neighborhood of the point at infinity the characteristic exponents are $+1$ and -3. So we may expect that an infinite set of solutions could have the same asymptotic behavior like k^μ, when $|\mathbf{k}|\to\infty$; we will see in the next Section that this is indeed the case and that there is an infinite set of solutions of Eq. (2.25) which gives a set of well defined integrals, when we put them in place of k^μ in Eq. (2.3).

III. THE EIGENFUNCTIONS OF THE LAPLACE–BELTRAMI OPERATOR

The Laplace–Beltrami operator Δ of Eqs. (2.18) and (2.20) was studied in a series of papers by Raczka, Limic, and Niederle,[13] where it is called $\Delta(H_3^1)$ [see Eq. (5.10) in the first reference of Ref. 13]. There it is shown that Δ has no discrete spectrum, but only a continuous one, and the basis of its generalized eigenfunctions is determined.

Δ can be identified with one of the Casimir operators of the Lorentz group, whose Lie algebra is defined in terms of the differential operators

$$l_{\mu\nu}=iD_{\mu\nu}, \tag{3.1}$$

with $D_{\mu\nu}$ as in Eq. (2.8).

To make contact with the notations used by Naimark in his book[14] we put

$$L_i=l_i,\quad K_i=-l_{0i}, \tag{3.2}$$

$$l_i=-\tfrac{1}{2}\epsilon_{ijk}l_{jk}\quad (i,j,k=1,2,3) \tag{3.2}$$

(in Naimark's notations:

$$H_3=L_3;\quad H_\pm=L_\pm=L_1\pm iL_2; \tag{3.3}$$

$$F_3=K_3;\quad F_\pm=K_\pm=K_1\pm iK_2). \tag{3.3}$$

These operators satisfy the Lorentz algebra

$$[L_i,L_j]=i\epsilon_{ijk}L_k,\quad [K_i,L_j]=i\epsilon_{ijk}K_k,\quad [K_i,K_j]=-i\epsilon_{ijk}L_k. \tag{3.4}$$

Two invariant operators can be defined using $\mathbf{L}$ and $\mathbf{K}$,

$$\Xi_1=\mathbf{L}\cdot\mathbf{K},\quad \Xi_2=|\mathbf{K}|^2-|\mathbf{L}|^2, \tag{3.5}$$

and these are the two Casimir operators of the Lorentz group.

The operators $\mathbf{L}$ and $\mathbf{K}$ can be written as

$$\mathbf{L}=-i\mathbf{k}\wedge\boldsymbol{\nabla},\quad \mathbf{K}=i\omega(\mathbf{k})\boldsymbol{\nabla}, \tag{3.6}$$

so

$$\Xi_1=-[\mathbf{k}\wedge\boldsymbol{\nabla}\omega(\mathbf{k})]\cdot\boldsymbol{\nabla},\quad \Xi_2=m^2\nabla^2+2(\mathbf{k}\cdot\boldsymbol{\nabla})+(\mathbf{k}\cdot\boldsymbol{\nabla})^2. \tag{3.7}$$

$\omega(\mathbf{k})$ has a gradient which is parallel to $\mathbf{k}$, so $\Xi_1=0$. We are left with one Casimir only,

$$\Xi_2 = m^2\Delta. \tag{3.8}$$

In terms of $l_{\mu\nu}$ we have

$$\Delta = \tfrac{1}{2} l_{\mu\nu} l^{\mu\nu}, \quad \epsilon_{\mu\nu\rho\lambda} l^{\mu\nu} l^{\rho\lambda} = 0, \tag{3.9}$$

and of course we have

$$[D, D_{\mu\nu}] = 0, \tag{3.10}$$

where the operator D is defined in Eq. (2.23).

If we denote λ the eigenvalue of $-m^2\Delta$, the relation with Naimark's notations[14] is given by

$$\lambda = -(k_o^2 + c^2 - 1), \quad k_o c = 0, \tag{3.11}$$

where k_o and c are the eigenvalues of two operators Δ and Δ' defined in the quoted reference,

$$\Delta|_{\text{eigenvalue}} = -2(k_o^2 + c^2 - 1), \qquad \Delta'|_{\text{eigenvalue}} = -4ik_o c. \tag{3.12}$$

If we choose

$$k_o = 0, \quad c = i\Lambda, \tag{3.13}$$

we get

$$\lambda = 1 + \Lambda^2 \in [1, +\infty), \tag{3.14}$$

corresponding to the representations of the Lorentz group of the principal series, with $k_o = 0$, which are unitary; see Ref. 14.

The zero modes of the operator D of Eq. (2.23) correspond to

$$\lambda = -3. \tag{3.15}$$

This value of λ corresponds to a nonunitary representation of the Lorentz group, which is reducible but not completely reducible (indecomposable) as it will be shown later. In Appendix B we give the details of the determination of this representation. Let us show here the expression of the boost K_3 and its action on the representation $\lambda = -3$. From Eqs. (3.1), (3.2) and (2.8) we get

$$l_{ij} = -\epsilon_{ijk} L_k = i(x^i \partial_j - x^j \partial_i), \quad l_{0j} = -K_j = -i\sqrt{1 + \mathbf{x}^2}\,\partial_j, \tag{3.16}$$

where

$$\mathbf{x} = \frac{\mathbf{k}}{m}. \tag{3.17}$$

From these expressions we get

$$K_3 = i\sqrt{1 + r^2}\left(\cos\theta \frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta}\right), \tag{3.18}$$

in the spherical coordinates (r, θ, ϕ) of $\mathbf{x}$.

D becomes

$$D=-m^2\Delta+3=-\Delta|_x+3, \tag{3.19}$$

where

$$\Delta|_x=\left[(1+r^2)\frac{\partial^2}{\partial r^2}+\left(\frac{2}{r}+3r\right)\frac{\partial}{\partial r}-\frac{J^2}{r^2}\right], \tag{3.20}$$

and $J^2=\mathbf{L}^2$ as usual.

The eigenfunctions of $-\Delta|_x$ correspond to $\lambda\in[1,+\infty)$, or $\Lambda\in[0,+\infty)$ as a limit from the upper half-plane of the complex plane of Λ. They are given in Ref. 13 [also see Eqs. (B2) and (B3)]

$$u_{\lambda,l,m}(r,\theta,\phi)=N_{\lambda l}v^{(o)}_{1,\lambda,l,m}=N_{\lambda l}r^lF\left(\frac{l+1+i\Lambda}{2},\frac{l+1-i\Lambda}{2};l+\frac{3}{2};-r^2\right)Y_{l,m}(\theta,\phi), \tag{3.21}$$

where F is the hypergeometric function, with the three parameters of $F(\alpha,\beta;\gamma;-r^2)$ satisfying the relation

$$\alpha+\beta+\tfrac{1}{2}=\gamma, \tag{3.22}$$

and where $l=0,1,2,\ldots,$ and $|m|\leqslant l$.

The normalization factor $N_{\lambda l}$ is

$$N_{\lambda l}=\frac{2\pi}{m\sqrt{\Lambda}}\left|\frac{\Gamma((l+2-i\Lambda)/2)\Gamma((l+1-i\Lambda)/2)}{\Gamma(i\Lambda)\Gamma(l+(3/2))}\right|, \tag{3.23}$$

for $\Lambda\in[0,\infty)$.

With this normalization $u_{\lambda,l,m}$ is an orthonormal set with respect to the scalar product of $L_2(\tilde{d}k)$,

$$\int \tilde{d}k\,\overline{u}_{\lambda lm}(r,\theta,\phi)u_{\lambda',l',m'}(r,\theta,\phi)=\delta_{ll'}\delta_{mm'}\delta(\Lambda-\Lambda'). \tag{3.24}$$

This basis is complete in $L_2(\tilde{d}k)$, that is

$$\sum_{l\geqslant 0}\sum_{|m|\leqslant l}\int_1^\infty d\lambda\, u_{\lambda,l,m}(r,\theta,\phi)\overline{u}_{\lambda,l,m}(r',\theta',\phi')=\Omega(k)\delta^3(k-k'). \tag{3.25}$$

The action of the Lorentz generators on the basis $\{v^{(o)}_{1,\lambda,l,m}\}$ is given in Appendix B, subsection 2.

For $\lambda\in[1,+\infty)$, as shown in Ref. 13, the representation of the Lorentz group on these bases is unitary. This interval for λ corresponds to the continuous spectrum of Δ.

We are interested in the case $\lambda=-3$, which corresponds to a pure imaginary Λ [see Eq. (3.14)], with $\mathrm{Im}\,\Lambda>0$. In this case the normalization factor $N_{\lambda,l}$ becomes meaningless, and the functions $\{v^{(o)}_{1,\lambda,l,m}\}$ are no more normalizable.

Following the discussion of Appendix B the only set of solutions of Eq. (2.25) which we will take into account in the following is the set $\{v^{(o)}_{1,\lambda,l,m}\}$, with $\lambda=-3$.

The action of K_3 on $v^{(o)}_{1,\lambda,l,m}$ is

$$K_3v^{(o)}_{1,\lambda,l,m}=-i\frac{(l+1+i\Lambda)(l+1-i\Lambda)}{2l+3}C_{l+1,m}v^{(o)}_{1,\lambda,l+1,m}++i(2l+1)C_{l,m}v^{(o)}_{1,\lambda,l-1,m}, \tag{3.26}$$

as given in Eq. (B32).

The action of $K_{\pm}$ is given in Eq. (B33), where the dependence on l of the coefficients in the r.h.s. is determined by the action of K_3 in Eq. (3.26) through Eq. (4.8).

From this equation we see that, for complex values of Λ in the upper half-plane, the first term in the r.h.s. of Eq. (3.26) vanishes for

$$\Lambda = i(l+1). \tag{3.27}$$

We may conclude that the *only representation* with an invariant subspace of dimension 4 is the representation with

$$\Lambda = 2i, \quad \text{or} \quad \lambda = -3. \tag{3.28}$$

In the next section we will study in more detail this representation.

IV. THE REPRESENTATION $\lambda = -3$

In this section we will study in some detail the representation $\lambda = -3$. For $\Lambda = 2i$ let us define

$$w_{l,m}(r,\theta,\phi) = v^{(o)}_{-3,l,m}(r,\theta,\phi). \tag{4.1}$$

From Eq. (3.26) we have the action of K_3 on $w_{l,m}$:

$$K_3 w_{l,m} = -i\frac{(l-1)(l+3)}{(2l+3)} C_{l+1,m} w_{l+1,m} + +i(2l+1) C_{l,m} w_{l-1,m}, \tag{4.2}$$

where the coefficients $C_{l,m}$ are defined in Eq. (B30), and where the values of $|m|$ in the right hand side are always limited by l in the first term and by $l-1$ in the second term.

Explicitly $w_{l,m}(\mathbf{x})$ is given by

$$w_{l,m}(r,\theta,\phi) = r^l F\left(\frac{l-1}{2}, \frac{l+3}{2}; l+\frac{3}{2}; -r^2\right) Y_{l,m}(\theta,\phi), \tag{4.3}$$

where r, θ, and ϕ are the spherical coordinates of $\mathbf{x} = \mathbf{k}/m$.

Observe that for $l=0,1$,

$$w_{0,0} = \sqrt{1+r^2} Y_{0,0} = \frac{\omega(\mathbf{k})}{m} Y_{0,0}, \quad w_{1,m} = r Y_{1,m} = \frac{|\mathbf{k}|}{m} Y_{1,m}, \tag{4.4}$$

which are the components of the 4-vector k^μ/m in the spherical basis, with the Minkowski diagonal metric given by

$$\tilde{\eta} = \frac{4\pi m^2}{3}(3,-1,-1,-1), \tag{4.5}$$

where $(l,m) = (0,0),(1,-1),(1,0),(1,-1)$ and

$$\sum_{l=0,1} \sum_{|m|\leq l} \overline{w}_{l,m} \tilde{\eta}_{lm,lm} w_{l,m} \equiv \overline{w} \cdot \tilde{\eta} \cdot w = m^2. \tag{4.6}$$

It is important to observe that all the functions $w_{l,m}$ have the same asymptotic behavior as $|\mathbf{k}| \to \infty$. In other words we have an infinite sequence of zero modes of the operator D, all with the same asymptotic behavior as k^μ. This fact is explicitly shown by Eq. (B14).

The subspace spanned by the components $l=0,1$ of $w_{l,m}$ is *invariant* under the action of the generators of the Lorentz group. Indeed, from Eq. (4.2), we have

$$K_3 w_{0,0} = \frac{i}{\sqrt{3}} w_{1,0}, \quad K_3 w_{1,m} = i\sqrt{3}\,\delta_{m,0} w_{0,0}. \tag{4.7}$$

Moreover, since

$$K_{\pm} = \pm [K_3, L_{\pm}], \tag{4.8}$$

and since $L_{\pm}$ does not modify the value of l, even the action of $K_{\pm}$ leaves the subspace $l=0,1$ invariant.

It is the factor $l-1$ in the r.h.s. of Eq. (4.2) which is responsible for this fact. There are no other values of Λ, in the upper half of its complex plane, which could provide an invariant subspace of the same dimension. If we choose $\Lambda = in$, as in Eq. (3.27), with an n integer >0, we have an invariant subspace with a dimension which is determined by the maximum value of l, which is equal to $n-1$, as we can see from Eq. (3.26). In conclusion, the value $\lambda = -3$, or $\Lambda = 2i$, is determined by the requirement of the existence of an invariant subspace of dimension 4.

The representation whose basis is $\{w_{l,m}\}$ is reducible, but not decomposable (not completely reducible). That it is reducible is clear, since the matrix of K_3, determined by Eq. (3.26), is block triangular. Indeed, if we define

$$K_3 w_{l,m} = w_{l',m} (K_3)_{l',l}, \tag{4.9}$$

we have

$$(K_3)_{2,1} = 0, \tag{4.10}$$

but

$$(K_3)_{1,2} \neq 0. \tag{4.11}$$

Now, a representation like this is decomposable if there exists a similarity transformation defined by a matrix S of the form

$$\begin{pmatrix} 1 & Y \\ 0 & 1 \end{pmatrix}, \tag{4.12}$$

transforming all the generators of the Lorentz group $l_{\mu\nu}$ in a block diagonal form; see Ref. 15.

The matrix Y has four rows and infinitely many columns and, of course, it must be nonzero. It should similarly transform in a block diagonal form also the matrix of $\mathbf{L}^2$, which is already in block diagonal form. This is clearly impossible. So the representation is indecomposable.

V. THE SUPERTRANSLATIONS

The presence of a whole series of zero modes of the operator D, with the same asymptotic behavior as $|\mathbf{k}| \to \infty$, allows us to define the following set of integrals:

$$P_{l,m} = \int \tilde{d}k\, w_{l,m}(\mathbf{k}) \overline{a(\mathbf{k})} a(\mathbf{k}), \tag{5.1}$$

where

$$\overline{P}_{l,m} = (-1)^m P_{l,-m}, \tag{5.2}$$

and where the functions $w_{l,m}$ are given in Eq. (4.3).

As shown in Appendix B [see Eqs. (B14) and (B22)] all these integrals are well defined.

The canonical action of the generators of the Lorentz group $M^{\mu\nu}$ on the $P_{l,m}$ can be obtained from Eqs. (2.6) and (2.7); since from Eq. (2.8) we have

$$D^{ij}=\left(k^i\frac{\partial}{\partial k^j}-k^j\frac{\partial}{\partial k^i}\right)=i\epsilon_{ijk}L_k,\quad D^{0j}=\omega(k)\frac{\partial}{\partial k^j}=-iK_j,\tag{5.3}$$

with the definitions (3.1) and (3.2), we get

$$\{M^{i,j},P_{l,m}\}=-i\epsilon_{ijk}P_{l',m'}(L_k)_{l',m';l,m},\tag{5.4}$$

$$\{M^{0,j},P_{l,m}\}=iP_{l',m'}(K_j)_{l',m';l,m}.\tag{5.5}$$

In these equations the matrices $\|L_k\|$ and $\|K_j\|$ are those given in Eqs. (B25), (B37) and (B38), and correspond to the representation $\lambda=-3$ of the Lorentz algebra.

Moreover, we have

$$\{P_{l,m},P_{l',m'}\}=0.\tag{5.6}$$

For $l=0,1$ the $P_{l,m}$ are the spherical components of P^{μ}.

Equations (5.4), (5.5) and (5.6) show that this algebra is infinite-dimensional, that the Abelian subalgebra of the $\{P_{l,m}\}$ is a normal subalgebra, and that the factor algebra is isomorphic to the Lorentz algebra.

In order to show that this algebra is the same as that of the BMS group, we simply redefine the $P_{l,m}$ with

$$\hat{P}_{l,m}=\nu_l P_{l,m},\tag{5.7}$$

with ν_l satisfying the recurrence relation

$$\nu_{l+1}=\frac{2l+3}{l+3}\nu_l,\tag{5.8}$$

or

$$\nu_l=\frac{(2l+1)!!}{(l+2)!}2\nu_0.\tag{5.9}$$

By defining

$$R_z=iK_3,\quad R_{\pm}=iK_{\pm},\quad L'_z=iL_z,\tag{5.10}$$

we get exactly the algebra given by Sachs[5] [see in this reference Eqs. (IV.19) and (IV.20)].

As shown by Sachs,[5] the four-dimensional subgroup of translations ($P_{l,m}$ with $l=0,1$) is unique. On the other hand the homogeneous Lorentz group is not similarly unique. This is due to the fact that copies of the Lorentz group can be obtained by a conjugation with an arbitrary supertranslation, whose role is here played by the quantities $P_{l,m}, l\geqslant 2$.

Indeed, let us define the following transformation of the Lorentz generators:

$$M'^{i,j}\to M'^{i,j}+\alpha\{P_{l,m},M'^{i,j}\}=M'^{i,j}+i\alpha\epsilon_{ijk}P_{l',m'}(L_k)_{l',m';l,m},\tag{5.11}$$

$$M'^{0,j}\to M'^{0,j}+\alpha\{P_{l,m},M'^{0,j}\}=M'^{0,j}-i\alpha P_{l',m'}(L_j)_{l',m';l,m},\tag{5.12}$$

where α is an arbitrary real parameter, and Eqs. (5.4), (5.5) were used.

For a fixed infinitesimal transformation (5.11), (5.12), getting the same r.h.s. terms:

$$e^{\alpha P_{l,m}}*M'^{i,j}=M'^{i,j}+i\alpha\epsilon_{ijk}P_{l',m'}(L_k)_{l',m';l,m},\tag{5.13}$$

$$e^{\alpha P_{l,m}}*M'^{0,j}=M'^{0,j}-i\alpha P_{l',m'}(L_j)_{l',m';l,m}.\tag{5.14}$$

In these equations the $*$ operation is defined by

$$e^{A}*B=\sum_{n\geqslant 0}\frac{1}{n!}D_A^n B,\quad D_A=\{A,\cdot\}. \tag{5.15}$$

This transformation corresponds to a conjugation of the Lorentz algebra with an arbitrary fixed supertranslation. It can be verified that the transformed algebra is again the Lorentz algebra.

As a consequence, the Casimir operator of the Poincaré group given by the square of the Pauli–Lubanski four-vector, which is invariant under the transformation (5.13) and (5.14) when $l=0,1$, will in general, change.[5]

The canonical action of a fixed supertranslation on the field is determined by

$$\{P_{l,m},a(\mathbf{k})\}=iw_{l,m}(\mathbf{k})a(\mathbf{k}),\quad \{P_{l,m},\overline{a}(\mathbf{k})\}=-iw_{l,m}(\mathbf{k})\overline{a}(\mathbf{k}), \tag{5.16}$$

which, for $l=0,1$, reduces to (2.6), written in spherical coordinates.

Since $P_{l,m}$ is not real ($\overline{P}_{l,m}=(-1)^m P_{l,-m}$), it induces two different canonical transformations on the field Φ, determined by its real and its imaginary part,

$$P_{l,m}=R_{l,m}+iI_{l,m}, \tag{5.17}$$

where

$$\{R_{l,m},a(\mathbf{k})\}=i\,\mathrm{Re}(w_{l,m}(\mathbf{k}))a(\mathbf{k}),\quad \{R_{l,m},\overline{a}(\mathbf{k})\}=-i\,\mathrm{Re}(w_{l,m}(\mathbf{k}))\overline{a}(\mathbf{k}), \tag{5.18}$$

and

$$\{I_{l,m},a(\mathbf{k})\}=i\,\mathrm{Im}(w_{l,m}(\mathbf{k}))a(\mathbf{k}),\quad \{I_{l,m},\overline{a}(\mathbf{k})\}=-i\,\mathrm{Im}(w_{l,m}(\mathbf{k}))\overline{a}(\mathbf{k}). \tag{5.19}$$

Again these can be exponentiated for a fixed supertranslation, for instance,

$$e^{\alpha R_{l,m}}a(\mathbf{k})=e^{i\alpha\,\mathrm{Re}\,w_{l,m}(\mathbf{k})}a(\mathbf{k}),\quad e^{\alpha R_{l,m}}\overline{a}(\mathbf{k})=e^{-i\alpha\,\mathrm{Re}\,w_{l,m}(\mathbf{k})}\overline{a}(\mathbf{k}), \tag{5.20}$$

and the analogous expressions for $I_{l,m}$.

Equation (5.20) defines a canonical transformation of the field, since the canonical Poisson brackets (A16) are invariant under such a transformation.

This transformation is nonlocal in the field $\Phi(\mathbf{x},t)$. It is a particular case of a linear transformation,

$$(\Phi,\Pi)\rightarrow(\Phi',\Pi'), \tag{5.21}$$

realized as an integral transformation of the convolution type. Indeed, $a(\mathbf{k})$ belongs to $L^2(R^3)$, and so is a tempered distribution, and the exponentials $\exp(i\,\mathrm{Re}\,w_{l,m})$ and $\exp(i\,\mathrm{Im}\,w_{l,m})$ are tempered distributions too. Moreover, the convolution product between their Fourier transforms is well defined. Indeed, the functions $w_{l,m}(\mathbf{k})$, being the solutions of a homogeneous elliptic equation, are infinitely differentiable.[16] They and all their derivatives are polynomially bounded, due to the bound (B14). So, the functions $w_{l,m}(\mathbf{k})$ and their exponential $\exp(i\,\mathrm{Re}\,w_{l,m})$, $\exp(i\,\mathrm{Im}\,w_{l,m})$, are multipliers in $S'(R^3)$, the space of tempered distributions on R^3.[17] This implies that their Fourier transforms are convolutes,[18] that is, their convolution exists with any tempered distribution. In conclusion, we may write the relation among (Φ,Π) and (Φ',Π') as

$$\begin{aligned}\Phi'(\mathbf{x},t)&=\int d^3x'[f(\mathbf{x}-\mathbf{x}')\Phi(\mathbf{x}',t)+g(\mathbf{x}-\mathbf{x}')\Pi(\mathbf{x}',t)],\\ \Pi'(\mathbf{x},t)&=\int d^3x'[h(\mathbf{x}-\mathbf{x}')\Phi(\mathbf{x}',t)+k(\mathbf{x}-\mathbf{x}')\Pi(\mathbf{x}',t)],\end{aligned} \tag{5.22}$$

where the distributions f, g and h are defined by

$$f(\mathbf{x})=\int \tilde{d}k\,\omega(\mathbf{k})(e^{i\alpha\,\mathrm{Re}\,w_{l,m}(\mathbf{k})+i\mathbf{k}\cdot\mathbf{x}}+\mathrm{c.c.}),\quad g(\mathbf{x})=i\int \tilde{d}k(e^{i\alpha\,\mathrm{Re}\,w_{l,m}(\mathbf{k})+i\mathbf{k}\cdot\mathbf{x}}-\mathrm{c.c.}),$$
$$h(\mathbf{x})=-i\int \tilde{d}k\,\omega^2(\mathbf{k})(e^{i\alpha\,\mathrm{Re}\,w_{l,m}(\mathbf{k})+i\mathbf{k}\cdot\mathbf{x}}-\mathrm{c.c.}),\quad k(\mathbf{x})=f(\mathbf{x}),\tag{5.23}$$

where c.c. means the complex conjugated. It is easily seen that the condition

$$\int d^3y[f(\mathbf{x}-\mathbf{y})k(\mathbf{x}'-\mathbf{y})-g(\mathbf{x}-\mathbf{y})h(\mathbf{x}'-\mathbf{y})]=\delta^3(\mathbf{x}-\mathbf{x}'),\tag{5.24}$$

which must be satisfied if the transformation (5.22) has to be canonical, holds.

The transformation induced by $I_{l,m}$ is the same with the replacement $\mathrm{Re}\,w_{l,m}\rightarrow\mathrm{Im}\,w_{l,m}$.

The transformation (5.22) changes the initial configuration of the field determined by the functions $a(\mathbf{k})$, to a new one, determined by the functions $a'(\mathbf{k})=\exp(i\alpha\,\mathrm{Re}(w_{l,m}(\mathbf{k})))a(\mathbf{k})$. As we have seen, this transformation can change the spin content of the field, when $l\geqslant 2$.

As a conclusion we collect Eqs. (5.4), (5.5) and (5.6), which give the Poisson algebra of the BMS group,

$$\{M'^{\mu\nu},P_{l,m}\}=P_{l',m'}(M'^{\mu\nu})_{l',m';l,m},\quad \{P_{l,m},P_{l',m'}\}=0,\tag{5.25}$$

where, as given by Eqs. (5.4) and (5.5),

$$(M'^{ij})_{l',m';l,m}=-i\epsilon_{ijk}(L_k)_{l',m';l,m},\quad (M'^{0j})_{l',m';l,m}=i(K_j)_{l',m';l,m},\tag{5.26}$$

with the matrix elements of L_k and of K_j given in Eqs. (B25), (B37) and (B38).

VI. CONCLUSIONS

Following our initial purpose, we have found a Poisson algebra isomorphic to the BMS algebra, realized on the phase space of a real classical Klein–Gordon field.

The structure constants of this algebra are given by the matrix elements of an infinite-dimensional representation of the Lorentz group, which is nonunitary, reducible and indecomposable. We have given an explicit basis of functions for this representation, which are not normalizable in the sense of an L_2 space with Lorentz invariant measure, but for which we give an asymptotic bound.

The requirement of the existence of a four-dimensional invariant subspace selects this representation almost uniquely among all the possible representations of the Lorentz group.

Thus, we have shown that it is possible to realize the BMS algebra outside the general relativity context, in which it was originally discovered. This fact suggests a more general role of the BMS group.

As it is well known, the BMS group is the semidirect product of the Lorentz group with the finite set of translations and the infinite set of supertranslations. In the general relativity case, on the basis of physical arguments, the vanishing of the supertranslations is required.[5,4] In the case of the same algebra realized in terms of the Klein–Gordon field this requirement should be a strong restriction on the possible field configurations, which could not have a clear justification.

As we have observed in the Introduction, we may expect that a similar analysis could be worked out for other classical massive fields. Something similar should also happen for a zero mass field, by performing a harmonic analysis on the $k_\mu k^\mu=0$ cone, where the Laplace–Beltrami operator changes significantly.[19]

The existence of the supertranslations in the context of the Klein–Gordon field theory is a byproduct of a larger study on the search of a canonical set of collective and relative variables for a classical relativistic field, playing a role analogous to the center of mass and relative variables of a nonrelativistic system of particles. This study will be the argument of a forthcoming paper.

ACKNOWLEDGMENTS

The authors wish to thank L. Lusanna for many suggestions and enlightening discussions on the subject of this paper, for his useful criticism and for reading the manuscript. The authors are also indebted to K. Marathe for many suggestions and for reading the manuscript.

APPENDIX A: NOTATIONS FOR THE CLASSICAL REAL KLEIN–GORDON FIELD

We list in this appendix the various definitions concerning the real Klein–Gordon field. We will put $c=\hbar=1$, and the metric signature is $(+;-,-,-)$.

The Lagrangian and the Lagrangian density are

$$L=\int dt\,\mathcal{L},\quad \mathcal{L}=\frac{1}{2}(\partial_\mu\Phi\,\partial^\mu\Phi-m^2\Phi^2), \tag{A1}$$

where $\mu=0,1,2,3$, and

$$\overline{\Phi}=\Phi. \tag{A2}$$

The conjugate momentum and the equation of motion are

$$\Pi(x)=\dot{\Phi}(x)\equiv\partial_o\Phi(x),\quad (\Box+m^2)\Phi(x)=0, \tag{A3}$$

and the generators of the Poincaré group are

$$P^\mu=\int d^3x\, j^{0\mu}(x),$$

$$M^{\mu\nu}=\int d^3x(x^\mu j^{0\nu}-x^\nu j^{0\mu}). \tag{A4}$$

The canonical Poisson brackets are (see Ref. 11)

$$\{\Phi(\mathbf{x},x^0),\Pi(\mathbf{x}',x^0)\}=\delta^3(\mathbf{x}-\mathbf{x}'), \tag{A5}$$

with the other Poisson brackets vanishing.

The Poincaré algebra is

$$\{P^\mu,P^\nu\}=0, \tag{A6}$$

$$\{M^{\mu\nu},P^\rho\}=P^\mu\eta^{\nu\rho}-P^\nu\eta^{\mu\rho}, \tag{A7}$$

$$\{M^{\mu\nu},M^{\rho\lambda}\}=M^{\mu\lambda}\eta^{\nu\rho}+\eta^{\mu\lambda}M^{\nu\rho}-M^{\mu\rho}\eta^{\nu\lambda}-\eta^{\mu\rho}M^{\nu\lambda}. \tag{A8}$$

The Fourier expansion of the field is

$$\Phi(\mathbf{x},t)=\int \tilde{d}k[a(\mathbf{k})e^{-i(k\cdot x)}+\overline{a(\mathbf{k})}e^{i(k\cdot x)}], \tag{A9}$$

where c.c. means the complex conjugate, and $(\cdot)$ is the usual Lorentz invariant scalar product between four-vectors, and (we use the notations of Ref. 20)

$$\tilde{d}k = \frac{d^3k}{\Omega(k)}, \quad \Omega(\mathbf{k}) = (2\pi)^3 2\omega(\mathbf{k}),$$

$$\mathbf{k} \equiv \{k^i\} \quad (i=1,2,3), \quad k_i = -k^i, \tag{A10}$$

$$\omega(\mathbf{k}) = k_o = \sqrt{\mathbf{k}^2 + m^2}.$$

If we denote with $\hat{\Phi}(\mathbf{k},t)$ the Fourier transform of the field, with respect to the measure d^3k, then

$$\Phi(\mathbf{x},t) = \int d^3k\, \hat{\Phi}(\mathbf{k},t) e^{i\mathbf{k}\cdot\mathbf{x}}, \tag{A11}$$

with

$$\overline{\hat{\Phi}}(\mathbf{k},t) = \hat{\Phi}(-\mathbf{k},t), \tag{A12}$$

and

$$a(\mathbf{k}) = \frac{1}{2}\Omega(\mathbf{k}) e^{i\omega(\mathbf{k})t}\left[\hat{\Phi}(\mathbf{k},t) + \frac{i}{\omega(\mathbf{k})}\dot{\hat{\Phi}}(\mathbf{k},t)\right]. \tag{A13}$$

From this we have the bound

$$|a(\mathbf{k})|^2 \leqslant (2\pi)^6 [\mathbf{k}^2 |\hat{\Phi}(\mathbf{k},t)|^2 + |\dot{\hat{\Phi}}(\mathbf{k},t)|^2]. \tag{A14}$$

Since we have assumed that $\Phi(\mathbf{x},\cdot)$, $\dot{\Phi}(\mathbf{x},\cdot)$ and $\boldsymbol{\nabla}\Phi(\mathbf{x},\cdot)$ are functions in $L^2(R^3)$, from the Parseval identity we get that $a(\mathbf{k}) \in L^2(R^3)$.

This implies the existence of P^μ, but not of $M^{\mu\nu}$. This last will be assured if we assume $\boldsymbol{\nabla} a(\mathbf{k}) \in L^2(R^3)$. So $a(\mathbf{k})$ will go to zero as $|\mathbf{k}| \to \infty$.

The conjugate momentum is

$$\Pi(x) = -i \int \tilde{d}k\, \omega(\mathbf{k}) [a(\mathbf{k}) e^{-i(k\cdot x)} - \overline{a}(\mathbf{k}) e^{i(k\cdot x)}]. \tag{A15}$$

The Poisson brackets for the Fourier coefficients are

$$\{a(\mathbf{k}), \overline{a}(\mathbf{k}')\} = -i\Omega(\mathbf{k}) \delta^3(\mathbf{k} - \mathbf{k}'), \tag{A16}$$

with the other Poisson brackets vanishing.

The Fourier coefficients in terms of the field are given by

$$a(\mathbf{k}) = \int d^3x\, e^{i(k\cdot x)} [\omega(\mathbf{k})\Phi(x) + i\Pi(x)]. \tag{A17}$$

In terms of the Fourier coefficients the Poincaré generators are the following:

$$P^\mu = \int \tilde{d}k\, k^\mu \overline{a}(\mathbf{k}) a(\mathbf{k}), \tag{A18}$$

$$M^{ij} = -i \int \tilde{d}k\, \overline{a}(\mathbf{k}) \left(k^i \frac{\partial}{\partial k^j} - k^j \frac{\partial}{\partial k^i} \right) a(\mathbf{k}) = \overline{M}^{ij}, \tag{A19}$$

$$M_{0j}=tP_j-i\int \tilde{d}k\,\overline{a(\mathbf{k})}\,\omega(\mathbf{k})\,\frac{\partial}{\partial k^j}a(\mathbf{k})=-M^{jo}=\overline{M}^{0j}. \tag{A20}$$

We define

$$M'^{0j}=-i\int \tilde{d}k\,\overline{a(\mathbf{k})}\,\omega(\mathbf{k})\,\frac{\partial}{\partial k^j}a(\mathbf{k}),\quad M'^{ij}=M^{ij}, \tag{A21}$$

where M'^{ij} and M'^{0j} have the same Poisson algebra as M^{ij} and M^{0j}, and will be used as the Lorentz generators.

APPENDIX B: THE EIGENFUNCTIONS OF THE LAPLACE–BELTRAMI OPERATOR AND THE ACTION OF THE LORENTZ GENERATORS

1. The eigenfunctions of the Laplace–Beltrami operator

The eigenfunctions of the continuous spectrum of the Laplace–Beltrami operator, corresponding to $\lambda\in[1,\infty)$, have been studied in Ref. 13. However, we are also interested in the non-normalizable solutions of the equation,

$$(-\Delta|_x-\lambda)v_{\lambda,l,m}=0, \tag{B1}$$

for other values of λ.

A fundamental system of solutions, in the neighborhood of the origin, that is for $r\simeq 0$, of Eq. (B1) in spherical coordinates, is

$$v^{(o)}_{1,\lambda,l,m}(\mathbf{r})=u^{(o)}_{1,\lambda,l}(r)Y_{l,m}(\theta,\phi),\quad v^{(o)}_{2,\lambda,l,m}(\mathbf{r})=u^{(o)}_{2,\lambda,l}(r)Y_{l,m}(\theta,\phi), \tag{B2}$$

where $r=|\mathbf{k}|/m$, $Y_{l,m}$ are the spherical harmonics as defined in Ref. 21, and

$$\begin{aligned} u^{(o)}_{1,\lambda,l}(r)&=r^lF\left(\frac{l+1+i\Lambda}{2},\frac{l+1-i\Lambda}{2};l+\frac{3}{2};-r^2\right),\\ u^{(o)}_{2,\lambda,l}(r)&=r^{-l-1}F\left(-\frac{l+i\Lambda}{2},-\frac{l-i\Lambda}{2};\frac{1}{2}-l;-r^2\right),\end{aligned} \tag{B3}$$

$$(u^{(o)}_{2,\lambda,l}(r)=u^{(o)}_{1,\lambda,-l-1}(r)), \tag{B4}$$

where F is the hypergeometric function.

The relation between λ and Λ is given by $\lambda=1+\Lambda^2$, with $\operatorname{Im}\Lambda\geqslant 0$, as in Sec. III.

A fundamental system in the neighborhood of the point at infinite $r\simeq\infty$ is

$$v^{(\infty)}_{1,\lambda,l,m}(\mathbf{r})=u^{(\infty)}_{1,\lambda,l}(r)Y_{l,m}(\theta,\phi),\quad v^{(\infty)}_{2,\lambda,l,m}(\mathbf{r})=u^{(\infty)}_{2,\lambda,l}(r)Y_{l,m}(\theta,\phi), \tag{B5}$$

where

$$\begin{aligned} u^{(\infty)}_{1,\lambda,l}(r)&=r^{-1-i\Lambda}F\left(\frac{l+1+i\Lambda}{2},-\frac{l-i\Lambda}{2};1+i\Lambda;-\frac{1}{r^2}\right),\\ u^{(\infty)}_{2,\lambda,l}(r)&=r^{-1+i\Lambda}F\left(\frac{l+1-i\Lambda}{2},-\frac{l+i\Lambda}{2};1-i\Lambda;-\frac{1}{r^2}\right).\end{aligned} \tag{B6}$$

For $r \geqslant 0$ no other singular points are met. For the argument $z=-1$, the hypergeometric series is absolutely convergent, since the coefficients of $F(\alpha,\beta;\gamma;z)$ satisfy the condition [see Eq. (3.22)]

$$\alpha+\beta-\gamma=-\tfrac{1}{2}. \tag{B7}$$

If $i\Lambda$ is a positive integer, the solution $u^{(\infty)}_{1,\lambda,l}(r)$ requires a modification of the expression given in Eq. (B6). However, since we are only interested in the solution $u^{(o)}_{1,\lambda,l}(r)$ for $\mathrm{Im}\,\Lambda \geqslant 0$, here we do not give the necessary modification.

We are interested in the normalization properties of these solutions in the neighborhood of the points 0 and ∞, with respect to the invariant measure,

$$\tilde{d}r=\frac{d^3r}{\sqrt{1+r^2}}. \tag{B8}$$

The solution $v^{(o)}_{1,\lambda,l,m}$ is regular and normalizable in the neighborhood of the origin. The solution $v^{(o)}_{2,\lambda,l,m}$ on the other hand is normalizable in the neighborhood of origin for $l=0,1$ only. We will discard this solution, even in the case $l=0,1$, since under the action of the Lorentz generators it will be transformed into a solution with a different value of l, thus becoming non-normalizable in the neighborhood of origin. See also the discussion at the end of subsection 2.

The solution $u^{(o)}_{1,\lambda,l}$ can be analytically continued to $r\simeq\infty$, and, for a generic value of Λ, has an asymptotic expansion which is a linear combination of the two powers of r [see Eq. 2.10(2) of Ref. 22],

$$r^{-1-i\Lambda}, \quad r^{-1+i\Lambda}, \tag{B9}$$

and, for a real Λ, is normalizable in the sense of the continuous spectrum.

For real values of $\Lambda\in[0,\infty)$, $v^{(o)}_{1,\lambda,l,m}$ is an eigenfunction of the operator Δ, see Ref. 13, belonging to the continuous spectrum.

In the case $\lambda=-3$, that is $\Lambda=2i$, we must use another asymptotic expansion. In general, for $\Lambda=ni$, with an n integer, we must use the expansion given in Eq. 2.10(7) of Ref. 22; for $\Lambda=2i$ and for $r\to\infty$ we get

$$u^{(o)}_{1,-3,l}\simeq r+O(r^{-1}). \tag{B10}$$

So, this solution is no more normalizable at ∞, but only at 0.

The other solution at $r\simeq\infty$ has the behavior

$$u^{(o)}_{2,-3,l}\simeq r+O(r^{-1}), \quad \text{for } l=0,1; \tag{B11}$$

$$u^{(o)}_{2,-3,l}\simeq r+O(r^{-3}), \quad \text{for } l\geqslant 2, \tag{B12}$$

and is singular at the origin for $l>0$. So it is not normalizable at $r\simeq\infty$.

As for the second fundamental system given in Eqs. (B5) and (B6) they are linear combinations of the first set, since they can be obtained by analytic continuation using well known relations. The solution $u^{(\infty)}_{1,\lambda,l}$ has, for $r\simeq\infty$, the asymptotic behavior

$$r^{-1+\mathrm{Im}\,\Lambda-i\,\mathrm{Re}\,\Lambda}, \tag{B13}$$

and, for $\mathrm{Im}\,\Lambda\geqslant 0$, is not normalizable.

The second solution $u^{(\infty)}_{2,\lambda,l,m}$ is instead normalizable for $r\simeq\infty$.

The solutions $w_{l,m}=v^{(o)}_{1,-3,l,m}$ satisfy the following inequality:

$$|w_{l,m}(\mathbf{k})|\leqslant\sqrt{\frac{2l+1}{4\pi}}\sqrt{1+r^2}M_l\,, \tag{B14}$$

where

$$M_l=\frac{\Gamma(l+(3/2))\Gamma(2)}{\Gamma((l/2)+2)\Gamma(l/2)},\quad \text{if } l\geqslant 2,$$
$$M_l=1,\quad \text{if } l=0,1. \tag{B15}$$

Indeed, since

$$|Y_{l,m}(\theta,\phi)|\leqslant\sqrt{\frac{2l+1}{4\pi}}, \tag{B16}$$

and

$$u_l(r)=r^lF\left(\frac{l-1}{2},\frac{l+3}{2};l+\frac{3}{2};-r^2\right)=r^l(1+r^2)^{(1-l)/2}F\left(\frac{l-1}{2},\frac{l}{2};l+\frac{3}{2};\frac{r^2}{1+r^2}\right), \tag{B17}$$

using Eq. 2.9(3) of Ref. 22, and the bounds

$$0\leqslant\frac{r^2}{1+r^2}<1, \tag{B18}$$

$$|F(\alpha,\beta;\gamma;x)|\leqslant F(\alpha,\beta;\gamma;1)=\frac{\Gamma(\gamma-\alpha-\beta)\Gamma(\gamma)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}, \tag{B19}$$

where the last holds if $\alpha,\beta,\gamma>0$ and $\gamma-\alpha-\beta>0$, we have from Eq. (B17),

$$|u_l(r)|\leqslant r^l(1+r^2)^{(1-l)/2)}\frac{\Gamma(2)\Gamma(l+(3/2))}{\Gamma((l/2)+2)\Gamma((l+3)/2)}. \tag{B20}$$

Now

$$r^l(1+r^2)^{(1-l)/2}\leqslant\sqrt{1+r^2}, \tag{B21}$$

and collecting the results we get the inequality (B14) for $l\geqslant 2$.

For $l=0,1$ the last hypergeometric function in Eq. (B17) is equal to 1, and we get the inequality (B14) for $l=0,1$.

The inequality (B14) for $w_{l,m}$ implies the existence of the integrals $P_{l,m}$ defined in Eq. (5.1). Indeed,

$$|P_{l,m}|\leqslant\int\tilde{d}k|w_{l,m}(\mathbf{k})|\overline{a}(\mathbf{k})a(\mathbf{k})\leqslant\sqrt{\frac{2l+1}{4\pi}}M_l\frac{P_0}{m}. \tag{B22}$$

So, if P_0 exists, all the integrals $P_{l,m}$ will similarly exist.

2. The action of the Lorentz generators

Let us determine the action of the generators of the Lorentz group on the solutions $v^{(o)}$. The explicit expression of the generators is

$$L_3=-i\frac{\partial}{\partial\phi},\quad L_\pm=L_1\pm iL_2=e^{\pm i\phi}\left(\pm\frac{\partial}{\partial\theta}+i\cot\theta\frac{\partial}{\partial\phi}\right),$$

$$K_3=i\sqrt{1+r^2}\left(\cos\theta\frac{\partial}{\partial r}-\frac{\sin\theta}{r}\frac{\partial}{\partial\theta}\right), \tag{B23}$$

$$K_\pm=i\sqrt{1+r^2}e^{\pm i\phi}\left(\sin\theta\frac{\partial}{\partial r}+\frac{\cos\theta}{r}\frac{\partial}{\partial\theta}\pm\frac{i}{r\sin\theta}\frac{\partial}{\partial\phi}\right).$$

They satisfy the algebra

$$\begin{aligned}
&[L_+,L_-]=2L_3,\quad [L_3,L_\pm]=\pm L_\pm,\\
&[K_+,K_-]=-2L_3,\quad [K_3,K_\pm]=\mp L_\pm,\\
&[K_3,L_3]=0,\quad [K_3,L_\pm]=\pm K_\pm,\\
&[L_3,K_\pm]=\pm K_\pm,\quad [K_\pm,L_\pm]=0.
\end{aligned} \tag{B24}$$

The action of L_3 and $L_\pm$ is the usual one,

$$\begin{aligned}
&L_3 v_{..,\lambda,l,m}=m v_{..,\lambda,l,m},\\
&L_\pm v_{..,\lambda,l,m}=\sqrt{l(l+1)-m(m\pm 1)}\;v_{..,\lambda,l,m\pm 1}.
\end{aligned} \tag{B25}$$

For the action of K_3 and $K_\pm$ we will use the following formulas:

$$\frac{d}{dr}F(a_l,b_l;c_l;-r^2)=-\frac{r}{2\sqrt{1+r^2}}\frac{(2a_l)(2b_l)}{c_l}F(a_{l+1},b_{l+1};c_{l+1};-r^2), \tag{B26}$$

$$\begin{aligned}
\frac{d}{dr}[r^lF(a_l,b_l;c_l;-r^2)]=(2l+1)\frac{r^{l-1}}{\sqrt{1+r^2}}F(a_{l-1},b_{l-1};c_{l-1};-r^2)\\
-(l+1)r^{l-1}F(a_l,b_l;c_l;-r^2),
\end{aligned} \tag{B27}$$

where $F(a_l,b_l;c_l;-r^2)$ is the hypergeometric function and where, see Eq. (3.22),

$$a_l+b_l+\tfrac{1}{2}=c_l. \tag{B28}$$

Equation (B26) can be obtained by using Eqs. 2.11(10) and 2.8(20) of Ref. 22, and Eq. (B27) using Eqs. 2.8(27) and 2.11(10) of the same reference.

Moreover, we will use the properties of the spherical harmonics,[21]

$$\cos\theta\;Y_{l,m}(\theta,\phi)=C_{l+1,m}Y_{l+1,m}(\theta,\phi)+C_{l,m}Y_{l-1,m}(\theta,\phi), \tag{B29}$$

where

$$C_{l,m}=\sqrt{\frac{(l-m)(l+m)}{(2l-1)(2l+1)}}, \tag{B30}$$

and

$$\sin\theta\frac{\partial}{\partial\theta}Y_{l,m}(\theta,\phi)=l\cos\theta\;\; Y_{l,m}(\theta,\phi)-\sqrt{\frac{(2l+1)(l^2-m^2)}{(2l-1)}}Y_{l-1,m}(\theta,\phi). \tag{B31}$$

With these relations we get

$$K_3 v^{(o)}_{1,\lambda,l,m}=-i\frac{(l+1+i\Lambda)(l+1-i\Lambda)}{(2l+3)}C_{l+1,m}v^{(o)}_{1,\lambda,l+1,m}+i(2l+1)C_{l,m}v^{(o)}_{1,\lambda,l-1,m}, \tag{B32}$$

and

$$K_\pm v^{(o)}_{1,\lambda,l,m}=\pm i\Bigg[\frac{(l+1+i\Lambda)(l+1-i\Lambda)}{(2l+3)}\sqrt{\frac{(l+1\pm m)(l+2\pm m)}{(2l+1)(2l+3)}}v^{(o)}_{1,\lambda,l+1,m\pm1}$$
$$+(2l+1)\sqrt{\frac{(l\mp m)(l\mp m-1)}{(2l-1)(2l+1)}}v^{(o)}_{1,\lambda,l-1,m\pm1}\Bigg]. \tag{B33}$$

For the solution $v_{2,\lambda,l,m}$ we get similarly

$$K_3 v^{(o)}_{2,\lambda,l,m}=+i\frac{(l+i\Lambda)(l-i\Lambda)}{(2l-1)}C_{l,m}v^{(o)}_{2,\lambda,l-1,m}-i(2l+1)C_{l+1,m}v^{(o)}_{2,\lambda,l+1,m}, \tag{B34}$$

and

$$K_\pm v^{(o)}_{2,\lambda,l,m}=\pm i\Bigg[\frac{(l+i\Lambda)(l-i\Lambda)}{(2l-1)}\sqrt{\frac{(l\mp m)(l-1\mp m)}{(2l-1)(2l+1)}}v^{(o)}_{2,\lambda,l-1,m\pm1}$$
$$+(2l+1)\sqrt{\frac{(l+1\pm m)(l+2\pm m)}{(2l+1)(2l+3)}}v^{(o)}_{2,\lambda,l+1,m\pm1}\Bigg]. \tag{B35}$$

If we put $\Lambda=2i$ we get for

$$w_{l,m}=v^{(o)}_{1,-3,l,m},\qquad w'_{l,m}=v^{(o)}_{2,-3,l,m}, \tag{B36}$$

$$K_3 w_{l,m}=-i\frac{(l-1)(l+3)}{(2l+3)}C_{l+1,m}w_{l+1,m}+i(2l+1)C_{l,m}w_{l-1,m}, \tag{B37}$$

$$K_\pm w_{l,m}=\pm i\Bigg[\frac{(l-1)(l+3)}{(2l+3)}\sqrt{\frac{(l+1\pm m)(l+2\pm m)}{(2l+1)(2l+3)}}w_{l+1,m\pm1}$$
$$+(2l+1)\sqrt{\frac{(l\mp m)(l\mp m-1)}{(2l-1)(2l+1)}}w_{l-1,m\pm1}\Bigg]. \tag{B38}$$

$$K_3 w'_{l,m}=+i\frac{(l-2)(l+2)}{(2l-1)}C_{l,m}w'_{l-1,m}-i(2l+1)C_{l+1,m}w'_{l+1,m}, \tag{B39}$$

$$K_\pm w'_{l,m}=\pm i\Bigg[\frac{(l-2)(l+2)}{(2l-1)}\sqrt{\frac{(l\mp m)(l-1\mp m)}{(2l-1)(2l+1)}}w'_{l-1,m\pm1}$$
$$+(2l+1)\sqrt{\frac{(l+1\pm m)(l+2\pm m)}{(2l+1)(2l+3)}}w'_{l+1,m\pm1}\Bigg]. \tag{B40}$$

In the previous formulas, the terms on the right hand side with the function $w_{l,m}$ with l negative or with $|m|>l$ are to be considered zero.

Referring to Eq. (B37) and (B38), observe that the representation subspace corresponding to the values of $l=0,1$ is an invariant subspace. This is due to the factor $(l-1)$.

Similarly, looking at Eq. (B39) and (B40), we see that the subspace corresponding to the values $l\geqslant 2$ is invariant.

Let us define the matrices corresponding to the representations $w_{l,m}$ and $w'_{l,m}$,

$$Mw_{l,m}=w_{l',m'}(M)_{l',m';l,m}, \quad Mw'_{l,m}=w'_{l',m'}(M')_{l',m';l,m}, \tag{B41}$$

where M is any one of the Lorentz generators.

If we define the new bases,

$$\hat{w}_{l,m}=N_l w_{l,m}, \quad \text{and} \quad \hat{w}'_{l,m}=N'_l w'_{l,m}, \tag{B42}$$

with

$$N_l=N'_l=\frac{1}{\sqrt{2l+1}}, \tag{B43}$$

we get for the new matrices

$$(\hat{M})_{l',m';l,m}=\frac{N_l}{N_{l'}}(M)_{l',m';l,m}, \tag{B44}$$

$$(\hat{M}')_{l',m';l,m}=\frac{N_l}{N_{l'}}(M')_{l',m';l,m}, \tag{B45}$$

the relation

$$(\hat{M}')^{\dagger}=\hat{M}. \tag{B46}$$

That is, the representation $\hat{w}'_{l,m}$ is the adjoint of the representation $\hat{w}_{l,m}$.[14]

[1] H. Bondi, M. G. J. Van der Burg, and A. W. K. Metzner, Proc. R. Soc. London, Ser. A **269**, 21 (1962); R. K. Sachs, *ibid.* **270**, 103 (1962); R. Sachs, Phys. Rev. **128**, 2851 (1962); E. T. Newman and T. W. J. Unti, J. Math. Phys. **3**, 891 (1062).
[2] For an exact definition of asymptotic flatness see, for instance, R. M. Wald, *General Relativity* (The University of Chicago Press, Chicago and London, 1984), Co. 11; J. Stewart, ''Advanced general relativity,'' *Cambridge Monographs on Mathematical Physics* (Cambridge University Press, Cambridge, 1991), Chap. 3.
[3] R. Penrose, Phys. Rev. Lett. **10**, 66 (1963); Proc. R. Soc. London, Ser. A **284**, 159 (1965); also see R. Penrose in Ref. 4.
[4] R. Geroch, ''Asymptotic structure of space–time,'' in *Asymptotic Structure of Space–time,* edited by E. Esposito and L. Witten (Plenum, New York, 1977); E. T. Newmann and K. P. Tod, ''Asymptotically flat space–-time,'' in *General Relativity and Gravitation,* edited by H. Held (Plenum, New York, 1980), Vol. 2; R. K. Sachs, ''Gravitational radiation,'' and R. Penrose, ''Conformal treatment of infinity,'' in *Relativity Groups and Topology,* edited by C. De Witt and B. De Witt (Gordon and Breach, New York, 1964); J. Winicour, ''Angular momentum in general relativity,'' in *General Relativity and Gravitation,* edited by A. Held (Plenum, New York, 1980), Vol. 2; H. Friedrich, ''Asymptotic structure of space–time,'' in *Recent Advances in General Relativity,* edited by A. I. Janis and J. R. Porter (Birkhauser, Basel, 1992), p. 146.
[5] See the third reference in Ref. 1. See also B. Schmidt, M. Walker, and P. Sommers, Gen. Relativ. Gravit. **6**, 489 (1975).
[6] See J. Winicour in Ref. 4.
[7] O. M. Moreschi, Class. Quantum Grav. **3**, 503 (1986); J. Math. Phys. **31**, 1208 (1990); J. Jezierski, ''Bondi mass in classical field theory,'' gr-qc/9703083, 27 March 97.
[8] V. Cantoni, J. Math. Phys. **7**, 1361 (1966); P. J. McCarthy, *ibid.* **13**, 1837 (1972); Phys. Rev. Lett. **29**, 817 (1972); Proc. R. Soc. London, Ser. A **330**, 517 (1972); A **333**, 317 (1973); A **335**, 301 (1973).
[9] R. Arnowitt, S. Deser, and C. W. Misner, Phys. Rev. **117**, 1595 (1960); **118**, 1100 (1960); **121**, 1566 (1961); **122**, 997 (1961); P. G. Bergmann, *ibid.* **124**, 274 (1961); R. Geroch, J. Math. Phys. **13**, 956 (1972); A. Ashtekar and R. U. Hansen, *ibid.* **19**, 1542 (1978); A. Ashtekar, ''Asymptotic structure of the gravitational field at spatial infinity,'' in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 2.
[10] L. Andersson and P. T. Chruściel, Phys. Rev. Lett. **70**, 2829 (1993); P. T. Chruściel, M. A. H. MacCallum, and D. B.

Singleton, gr-qc/9305021, 25 May 93; see, however, P. G. Bergmann, Phys. Rev. **124**, 274 (1961), where logarithmic terms were already considered.
[11] With this hypotheses the Klein–Gordon system is Hamiltonian; see P. Chernov and J. E. Mardsen, ''Properties of infinite dimensional Hamiltonian systems,'' *Lecture Notes in Mathematics* No. 425, edited by A. Dold and B. Eckmann (Springer-Verlag, Berlin, 1974). See also R. Schmid, ''Infinite dimensional Hamiltonian systems,'' *Monographs and Textbooks in Physical Science,* Lecture Notes 3, 1987, by Bibliopolis, edizioni di filosofia e scienze, Napoli.
[12] See for instance R. D. Richtmyer, *Principles of Advanced Mathematical Physics* (Springer-Verlag, 1978), Vol. I, Chap. 5.
[13] R. Raczka, N. Limic, and J. Niederle, J. Math. Phys. **7**, 1861 (1966); **7**, 2026 (1966); **8**, 1079 (1967). See also: A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (Polish Scientific Publishing, Warszawa, 1977), Chap. 15.
[14] M. A. Naimark, *Les Représentations Linéaires du Groupe de Lorentz* (Dunod, Paris, 1962).
[15] M. Naimark and A. Stern, *Théorie des Représentations Des Groupes (MIR, Moscow, 1979).*
[16] W. Rudin, *Functional Analysis* (McGraw-Hill, New York, 1991), Chap. 8.
[17] I. M. Gel'fand and G. E. Shilov, *Generalized Functions* (Academic, New York, 1968), Vol. II, Chap. II.
[18] The same reference as in Ref. 17, Chap. III.
[19] See the second paper in Ref. 13.
[20] C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980), Chap. 3.
[21] A. Messiah, *Quantum Mechanics* (North-Holland, New York, 1961), Vol. I, Appendix B.
[22] A. Erdélyi *et al.*, *Higher Transcendental Functions* (Krieger, 1953, reprint edition 1985), Vol. I.

Involutive automorphisms and Iwasawa decomposition of some hyperbolic Kac–Moody algebras

K. C. Pati, D. Parashar, and R. S. Kaushal[a]
Department of Physics and Astrophysics, University of Delhi, Delhi-110 007, India

(Received 13 July 1998; accepted for publication 11 August 1998)

The involutive automorphisms of hyperbolic Kac–Moody algebras are computed from Satake diagrams of these algebras which are then used to furnish a general procedure of their Iwasawa decomposition. In particular $\hat{A}_1^{(1)}$ and $\hat{A}_2^{(1)}$ are considered as representative examples to illustrate the efficacy of the underlying treatment. © *1999 American Institute of Physics.* [S0022-2488(99)01501-7]

I. INTRODUCTION

It is now widely accepted that the study of infinite dimensional Lie algebras[1,2] and more particularly the hyperbolic version of Kac–Moody algebras[3,4] has played an outstanding role in the realms of string theory (with special reference to the over extended exceptional algebra E_{10}), duality properties of supersymmetric gauge theories,[5] and two-dimensional field theories.[6] This lends sufficient credence to the desirability of studying the classification theory and representations of hyperbolic Kac–Moody algebras. While the Dynkin diagrams[7] and root spaces[8–12] of such algebras have already been exhaustively investigated and classified, a complete analysis of their representations has not yet been achieved to that extent. We shall, therefore, make a modest attempt to address this problem in the present communication.

Here we carry out the Iwasawa decomposition[9] of the hyperbolic Kac–Moody algebras. The involutive automorphisms required for such a decomposition are determined from the Satake diagrams of these algebras. In addition, these diagrams are capable of providing the real forms of Lie algebras[13–15] and their supersymmetric extensions.[16] In a recent work, this technique has been exploited to obtain the involutive automorphisms of affine Kac–Moody algebras[17–19] and superalgebras.[20] Here we shall expound that the same technique can be employed to generate involutive automorphisms of hyperbolic Kac–Moody algebras once their Satake diagrams have been constructed. The essential ingredients entering the method comprise the root system and the associated Dynkin diagrams which eventually lead to the construction of Satake diagrams within the framework of a standard prescription outlined earlier.

Before embarking on a systematic analysis, it is prudent to emphasize that the generalized Kac–Moody algebras (GKMA) differ from the finite simple Lie algebras (FSLA) or the affine Kac–Moody algebras (AKMA) in terms of their Cartan matrices. While the determinant of the Cartan matrix is positive for FSLA and zero for AKMA, it can be anything for the GKMA. The hyperbolic Kac–Moody algebras are those types of GKMA which fall back to the case of AKMA or FSLA on deletion of any one simple root from their Dynkin diagrams. Fortunately, the Dynkin diagrams[7] of all such algebras are known. It is instructive to distinguish between two types of hyperbolic Kac–Moody algebras, namely, (i) the strictly hyperbolic one which yields only FSLA Dynkin diagrams on deletion of any one vertex, with maximum allowed rank being 4 and (ii) the hyperbolic one whose Dynkin diagram reverts to those of FSLA or AKMA on deletion of any vertex, with the maximum allowed rank being 10 in this case. However, note must be taken of the fact that there is no hyperbolic Kac–Moody algebra of rank larger than 10. Furthermore, though there are infinitely many hyperbolic Kac–Moody algebras of rank 2, the number of these algebras of rank from 3 to 10 is necessarily finite. However, for Iwasawa decomposition[19] we restrict ourselves to only those types of hyperbolic Kac–Moody algebras which are extensions of the

[a]UGC Research Scientist.

affine algebras by a basic representation, since the root system of these types of algebras is less complicated than their other counterparts.

The outline for the construction of Satake diagrams of the hyperbolic Kac–Moody algebras are briefly delineated in Sec. II. All possible Satake diagrams of $\hat{A}_1^{(1)}$ and $\hat{A}_2^{(1)}$ along with their root automorphisms are computed in detail and listed in Tables I and II, respectively. Section III is devoted to a quick resume of the procedure for obtaining the Iwasawa decomposition within the present context. The explicit construction of this decomposition is then carried out for the hyperbolic Kac–Moody algebras $\hat{A}_1^{(1)}$ and $\hat{A}_2^{(1)}$. Section IV contains a few concluding remarks.

II. HYPERBOLIC KAC–MOODY ALGEBRAS AND SATAKE DIAGRAMS

We commence this section by taking a look at the hyperbolic Kac–Moody algebras $\hat{G}$ which are extensions of affine algebras G by a basic representation and the construction of the Satake diagrams. Let (a_{ij}), $i,j=0,1,...,r$, be the Cartan matrix of G and e_{α_i}, $e_{-\alpha_i}$, h_{α_i} be the generators associated with the simple roots α_i. By convention, the root α_0 is the extended root of the affine algebra G and the remaining roots α_j, $(j=1,...,r)$ are those of the associated finite simple Lie algebra. We can extend G by a derivation d which commutes with Cartan generators h_{α_i} and $[d,e_{\alpha_i}]=-\delta_{i,0}e_{\alpha_i}$, $[d,e_{-\alpha_i}]=\delta_{i,0}e_{-\alpha_i}$. The Cartan matrix $\hat{a}$ of the extension of G is defined by $\hat{a}_{ij}=a_{ij}$ for $i,j=0,1,...,r$, $\hat{a}_{i,-1}=\hat{a}_{-1,i}=-\delta_{i,0}$. The algebra $\hat{G}$ is thus of rank $(r+1)$. The root α_{-1} is usually referred to as the overextended root. The algebra $\hat{G}$ can be defined as the algebra generated by the elements $\{e_{\alpha_i}, e_{-\alpha_i}, h_{\alpha_i}, i=-1,0,1,...,r\}$ with relations

$$\begin{gathered}[h_{\alpha_i},h_{\alpha_j}]=0, \quad [e_{\alpha_i},e_{-\alpha_j}]=\delta_{ij}h_{\alpha_i}, \quad [e_{\pm\alpha_j},h_i]=a_{ij}e_{\pm\alpha_j}, \\ (ad\ e_{\alpha_i})^{1-\alpha_{ij}}e_{\alpha_j}=(ad\ e_{-\alpha_i})^{1-a_{ij}}e_{-\alpha_i}=0, \quad i\neq j.\end{gathered} \tag{2.1}$$

We can now associate to each simple root system a Dynkin diagram, the details of which can be found in Ref. 7. The construction of Satake diagrams for hyperbolic Kac–Moody algebra is achieved with the help of the following prescription.

Let R be a root system of hyperbolic Kac–Moody algebra. For $\alpha\in R$, let $\overline{\alpha}=\alpha-\sigma(\alpha)$, where σ is the involutive automorphisms of root α. Let us introduce $R_-=\{\overline{\alpha}|\overline{\alpha}\neq 0,\alpha\in R\}$ and $R_0=\{\alpha\in R|\overline{\alpha}=0\}$. If B_- (resp. B) denotes the basis of R_- (resp. R) and B_0 a basis of R_0 then $B=B\cap R_0$. If $B_-=B\setminus B_0=\{\alpha_i\}$ and $B_0=\{\beta_l\}$, then it can be shown that

$$-\sigma(\alpha_i)=\alpha_{\pi(i)}+\sum_l \eta_{il}\beta_l, \tag{2.2}$$

where π is the involutive permutation of $\{-1,0,1,...,r\}$ and η_{il} are non-negative integers. We should note that $\sigma(\beta_l)=\beta_l$ and $\alpha+\sigma(\alpha)\notin R\forall\alpha\in R$. We can now associate with B its Satake diagrams. In the Dynkin diagram of B denote the roots α_i by the usual white circles ○ and roots β_l by the black circles ●. If $\pi(i)=j$, then it is indicated by

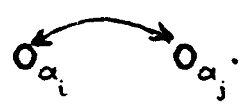

Satake diagrams determine the involution of R uniquely. As an illustration, we restrict ourselves to the consideration of two examples $\hat{A}_1^{(1)}$ and $\hat{A}_2^{(1)}$.

A. Satake diagrams of $\hat{A}_1^{(1)}$

The Cartan matrix of $\hat{A}_1^{(1)}$ is

$$\begin{bmatrix} 2 & -2 & 0 \\ -2 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}.$$

The three simple roots of $\hat{A}_1^{(1)}$ are α_1, α_0, and α_{-1}. The possible Satake diagrams of $\hat{A}_1^{(1)}$ along with their root automorphisms are represented in Table I.

B. Satake diagrams of $\hat{A}_2^{(1)}$

The Cartan matrix of $\hat{A}_2^{(1)}$ is

$$\begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

and the four simple roots of $\hat{A}_2^{(1)}$ are α_1, α_2, α_0, and α_{-1}, where $\alpha_0 = \delta - (\alpha_1 + \alpha_2)$. The possible Satake diagrams of $\hat{A}_2^{(1)}$ along with their root automorphisms can be read from Table II.

III. IWASAWA DECOMPOSITION OF HYPERBOLIC KAC–MOODY ALGEBRAS

The notion of direct determination of Iwasawa decomposition of Lie algebras is now extended to the case of hyperbolic Kac–Moody algebras. Let $\mathbf{G}$ be a real hyperbolic Kac–Moody algebra generated from its compact real form $\hat{\mathbf{G}}_k$ by an involutive automorphism defined with respect to the Cartan subalgebra $\mathbf{h}$ of $\hat{\mathbf{G}}_c$, where $\hat{\mathbf{G}}_c$ is the complexification of $\hat{\mathbf{G}}$. The following commutation relations are satisfied by the elements of $\hat{\mathbf{G}}_c$:

$$[e_\alpha, h] = \alpha(h) e_\alpha, \quad h \in \mathbf{h}, \quad \alpha \in \Delta,$$

$$[e_\alpha, e_\beta] = \begin{cases} N_{\alpha\beta} e_{\alpha+\beta} & \text{if } \alpha+\beta \text{ is a root} \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

$$[e_\alpha, e_{-\alpha}] = h_\alpha, h_\alpha \in \mathbf{h}.$$

Here Δ denotes the set of roots of $\hat{\mathbf{G}}_c$ w.r.t. $\mathbf{h}$ and the Killing form is defined by $B(e_\alpha, e_{-\alpha}) = -1$. Here $\alpha(h) = B(h, h_\alpha)$. The compact real form $\hat{\mathbf{G}}_k$ may be taken to consist of ih_α, $\alpha = \alpha_{-1}, \alpha_0, \alpha_1, \ldots, \alpha_r$ and $(e_\alpha + e_{-\alpha})$, $i(e_\alpha - e_{-\alpha}) \forall \alpha \in \Delta$. Let $\mathbf{k}$ be the maximal compact subalgebra of $\hat{\mathbf{G}}$ defined in such a way that $a \in \mathbf{k}$ if $a \in \hat{\mathbf{G}}$ and $\sigma a = a$. Let $\mathbf{p}$ be the subspace such that $a \in \mathbf{p}$ if $a \in \hat{\mathbf{G}}_c$ and $\sigma a = -a$. Thus, $\mathbf{k}$ and $\mathbf{p}$ are given by

$$\mathbf{k} = \{ih_\alpha, \alpha = \alpha_{-1}, \alpha_0, \alpha_1, \ldots, \alpha_r \text{ and } (e_\alpha + e_{-\alpha}), i(e_\alpha - e_{-\alpha}) \forall \alpha | \exp \alpha(h) = +1\}, \tag{3.2}$$

TABLE I. Satake diagrams and involutive automorphisms of $\hat{A}_1^{(1)}$.

Satake diagrams of $\hat{A}_1^{(1)}$	Involutive automorphisms
(*i*) 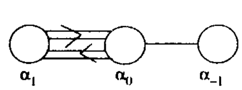	$-\sigma(\alpha_1) = \alpha_1$ $-\sigma(\alpha_0) = \alpha_0$ $-\sigma(\alpha_{-1}) = \alpha_{-1}$
(*ii*) 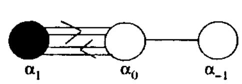	$\sigma(\alpha_1) = \alpha_1$ $-\sigma(\alpha_0) = \alpha_0 + 2\alpha_1$ $-\sigma(\alpha_{-1}) = \alpha_{-1}$

TABLE II. Satake diagrams and involutive automorphisms of $\hat{A}_2^{(1)}$.

Satake diagrams of $\tilde{A}_{1(1)}$	Involutive automorphism
(i) 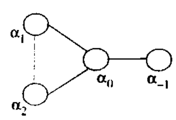	$-\sigma(\alpha_1)=\alpha_1$ $-\sigma(\alpha_2)=\alpha_2$ $-\sigma(\alpha_0)=\alpha_0$ $-\sigma(\alpha_{-1})=\alpha_{-1}$
(ii) 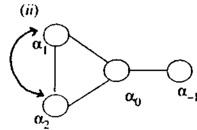	$-\sigma(\alpha_1)=\alpha_2$ $-\sigma(\alpha_2)=\alpha_1$ $-\sigma(\alpha_0)=\alpha_0$ $-\sigma(\alpha_{-1})=\alpha_{-1}$
(iii) 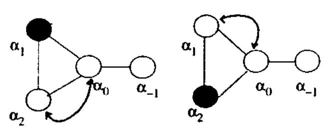	$\sigma(\alpha_1)=\alpha_i$ $-\sigma(\alpha_1)=\alpha_0+\alpha_2$ $-\sigma(\alpha_2)=\alpha_0+\alpha_1$ $\sigma(\alpha_2)=\alpha_2$ $-\sigma(\alpha_0)=\alpha_1+\alpha_2$ $-\sigma(\alpha_0)=\alpha_1+\alpha_2$ $-\sigma(\alpha_{-1})=\alpha_{-1}$ $-\sigma(\alpha_{-1})=\alpha_{-1}$
(iv) 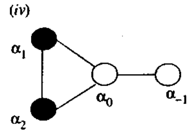	$\sigma(\alpha_1)=\alpha_1$ $\sigma(\alpha_2)=\alpha_2$ $-\sigma(\alpha_0)=\alpha_0+2\alpha_1+2\alpha_2$ $-\sigma(\alpha_0)=\alpha_{-1}$

$$\mathbf{p}=\{i(e_\alpha+e_{-\alpha}),(e_\alpha-e_{-\alpha})\forall\alpha|\exp\ \alpha(h)=-1\}. \tag{3.3}$$

Let $\mathbf{a}$ be the maximal Abelian subalgebra of $\mathbf{p}$ with dimension $|m_i|$ and $\mathbf{m}$ be the centralizer of $\mathbf{a}$ in $\mathbf{k}$. Thus, $\mathbf{a}$ may be taken to have a basis consisting of the elements of the form $i(e_\alpha+e_{-\alpha})$. Let R_A denote the set of positive roots α that appear in this way in $\mathbf{a}$. Similarly, $\mathbf{m}$ may be taken to have a basis consisting of the elements of the form $(e_\alpha+e_{-\alpha})$, with a set of positive roots α appearing this way in $\mathbf{m}$, being denoted by R_M, together possibly with some elements of $\mathbf{h}\cap\hat{\mathbf{G}}$. If $h''\in\mathbf{h}\cap\hat{\mathbf{G}}$ is a member of $\mathbf{m}$, then $\alpha(h'')=0\forall\alpha\in R_A\cup R_M$. Complexification of $\mathbf{a}\oplus\mathbf{m}$ together with the derivation d' gives a Cartan subalgebra $\mathbf{h}'$ of $\hat{\mathbf{G}}_c$ with basis $H'_{-1},H'_0,H'_1,\ldots,d'$. Now there exists an inner automorphism V: $\mathbf{h}'\rightarrow\mathbf{h}$, i.e.,

$$H_j=VH'_j,\ \ \text{where}\ \ V=\prod_\alpha V_\alpha,\ \ \alpha\in\Delta. \tag{3.4}$$

Let Δ^+ be the set of positive roots, then

$$h_\alpha=\sum_{j=-1}^{r} b_j(\alpha)H_j. \tag{3.5}$$

Thus, $\alpha\in\Delta^+$ iff $b_j(\alpha)>0$, where j is the least index such that $b_j(\alpha)\neq 0$. The positive roots can be again divided into the following classes:

$$\text{(i)}\ \ \Delta_+^+=\{\alpha|\alpha\in\Delta^+,\ \ \alpha(h)\neq\alpha(V\sigma V^{-1}h)\forall h\in\mathbf{h}\}, \tag{3.6}$$

$$\text{(ii)}\quad \Delta_{-}^{+}=\{\alpha|\alpha\in\Delta^{+},\ \ \alpha(h)=\alpha(V\sigma V^{-1}h)\forall h\in\mathbf{h}\}. \tag{3.7}$$

Let the subalgebra $\tilde{\mathbf{n}}$ be spanned by the elements $V^{-1}e_{\alpha}$ for $\alpha\in\Delta_{+}^{+}$ and $\mathbf{n}=\tilde{\mathbf{n}}\cap\hat{\mathbf{G}}_{c}$, where $\tilde{\mathbf{n}}$ and $\mathbf{n}$ are the nilpotent subalgebras of $\hat{\mathbf{G}}_{c}$ and $\hat{\mathbf{G}}$, respectively. Thus the Iwasawa decomposition of $\hat{\mathbf{G}}$ is given by

$$\hat{\mathbf{G}}=\mathbf{k}\oplus\mathbf{a}\oplus\mathbf{n}. \tag{3.8}$$

A. Iwasawa decomposition of $\hat{A}_{1}^{(1)}$

Let us consider the involutive automorphisms of $\hat{A}_{1}^{(1)}$ determined by any one of the Satake diagrams, say (ii), of Table I. The simple root automorphisms are given by

$$\alpha(\alpha_{1})=\alpha_{1},\quad -\sigma(\alpha_{0})=\alpha_{0}+2\alpha_{1},\quad -\sigma(\alpha_{-1})=\alpha_{-1}. \tag{3.9}$$

The positive roots of $\hat{A}_{1}^{(1)}$ are given by

$$\begin{aligned}\Delta=\{&n_{1}\alpha_{-1}+n_{2}\alpha_{1}+(n_{2}-1)\alpha_{0},\ \ \text{where}\ \ n_{1}=0,1,...,(n_{2}-1);\ \ n_{1}\alpha_{-1}+(n_{2}-1)\alpha_{1}\\ &+n_{2}\alpha_{0},n_{1}\alpha_{-1}+n_{2}(\alpha_{1}+\alpha_{0}),\ \ \text{where}\ \ n_{1}=0,1,...,n_{2}\ \ \text{and}\ \ n_{2}\in Z\}.\end{aligned} \tag{3.10}$$

We can apply the simple root automorphisms to find out the automorphisms of other roots and we see that the positive roots can be separated into categories, i.e.,

$$\exp\ \alpha(h)=+1\quad\text{for}\quad \alpha=n_{1}\alpha_{-1}+n_{2}\alpha_{1}+(n_{2}-1)\alpha_{0}, \tag{3.11}$$

$$\exp\ \alpha(h)=-1\quad\text{for}\quad \alpha=\begin{cases}n_{1}\alpha_{-1}+(n_{2}-1)\alpha_{1}+n_{2}\alpha_{0}\\ n_{1}\alpha_{-1}+n_{2}(\alpha_{1}+\alpha_{0})\end{cases}. \tag{3.12}$$

Thus, for $\hat{A}_{1}^{(1)}$, $\mathbf{k}$ and $\mathbf{p}$ are given by

$$\mathbf{k}=\{ih_{\alpha}\ \ \text{for}\ \ \alpha=\alpha_{-1},\alpha_{0},\alpha_{1}\ \ \text{and}\ \ (e_{\alpha}+e_{-\alpha}),\ \ i(e_{\alpha}-e_{-\alpha})\ \ \text{for}\ \ \alpha\ \ \text{given by Eq. (3.11)}\}, \tag{3.13}$$

$$\mathbf{p}=\{i(e_{\alpha}+e_{-\alpha}),(e_{\alpha}-e_{-\alpha})\ \ \text{for}\ \ \alpha\ \ \text{given by Eq. (3.12)}\}. \tag{3.14}$$

We now select a maximal Abelian subalgebra $\mathbf{a}$ which is one-dimensional and may be chosen to have a basis element

$$H_{-1}'=i(e_{\alpha_{0}}-e_{-\alpha_{0}}). \tag{3.15}$$

In fact, we may also choose $i(e_{\alpha_{-1}}-e_{-\alpha_{-1}})$ as a basis element of $\mathbf{a}$. But we shall see that later on it will pose a problem while calculating the elements of nilpotent subalgebra $\tilde{\mathbf{n}}$ because of the fact that the α_{-1} string containing $n_{1}\alpha_{-1}+(n_{2}-1)\alpha_{1}+n_{2}\alpha_{0}$ or $n_{1}\alpha_{-1}+n_{2}(\alpha_{1}+\alpha_{0})$ or $n_{1}\alpha_{-1}+n_{2}\alpha_{1}+(n_{2}-1)\alpha_{0}$ can have a large number of roots up to infinity. So, now we have $R_{A}=\{\alpha_{0}\}$ and R_{M} is empty. $\mathbf{m}$ is two-dimensional and its basis elements are given by

$$-iH_{0}'=ih_{\alpha_{1}+\alpha_{0}},\quad -iH_{1}'=ih_{2\alpha_{-1}+2\alpha_{0}+\alpha_{1}}. \tag{3.16}$$

Note that H_{-1}',H_{0}',H_{1}' together with scaling element d' are the elements of the Cartan subalgebra $\mathbf{h}'$. The inner automorphism of $\hat{A}_{1}^{(1)}$ is given by the expression

$$V_{\alpha_{0}}=\exp[ad\{ia_{\alpha_{0}}(e_{\alpha_{0}}-e_{-\alpha_{0}})\}], \tag{3.17}$$

where

$$a_{\alpha_0}=\frac{\pi}{[8(\alpha_0,\alpha_0)]^{1/2}}. \tag{3.18}$$

Applying this to the Cartan subalgebra $\mathbf{h}'$ of $\hat{A}_1^{(1)}$ we obtain

$$H_{-1}=-2^{-1/2}h_{\alpha_0},\quad H_0=-(h_{\alpha_1}+h_{\alpha_0}),\qquad H_1=-(2h_{\alpha_{-1}}+2h_{\alpha_0}+h_{\alpha_1}). \tag{3.19}$$

With respect to this Cartan subalgebra, the set of positive roots is given by

$$\begin{aligned}\Delta^+=\{&n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0\forall n_1,n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0 \text{ for } n_1>2,\\ &-(n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0) \text{ for } n_1\leqslant 2,\\ &n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0) \text{ for } n_1\neq 0,\\ &-(n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0)) \text{ for } n_1=0\}.\end{aligned} \tag{3.20}$$

Now Δ_+^+ and Δ_-^+ can be written down as

$$\Delta_-^+=\{n_2+(\alpha_1+\alpha_0),2\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0\}. \tag{3.21}$$

$$\begin{aligned}\Delta_+^+=\{&n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0,n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0 \text{ for } n_1>2,\\ &-(n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0) \text{ for } n_1<2,\\ &n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0) \quad \text{ for } n_1\neq 0\}.\end{aligned} \tag{3.22}$$

The elements of $\tilde{\mathbf{n}}$ are determined by the structures $V_{\alpha_0}^{-1}e_\alpha$, where $\alpha\in\Delta_+^+$. These are given by

$$\begin{aligned}V_{\alpha_0}^{-1}e_{n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0)}&=-i2^{-1/2}Ae_{n_1\alpha_{-1}+n_2\alpha_1+(n_2+1)\alpha_0}+i2^{-1/2}Be_{n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)a_0},\\ V_{\alpha_0}^{-1}e_{n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0}&=\tfrac{1}{2}e_{n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0}+i2^{-1/2}Be_{n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0)}\\ &\quad+\tfrac{1}{2}ABe_{n_1\alpha_{-1}+n_2\alpha_1+(n_2+1)\alpha_0},\\ V_{\alpha_0}^{-1}e_{n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0}&=\tfrac{1}{2}e_{n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0}-i2^{-1/2}Ce_{n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0)}\\ &\quad+\tfrac{1}{2}CDe_{n_1\alpha_{-1}+(n_2-1)\alpha_1+(n_2-2)\alpha_0},\\ V_{\alpha_0}^{-1}e_{-(n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0)}&=\tfrac{1}{2}e_{-(n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0)}+i2^{-1/2}Ee_{-(n_1\alpha_{-1}+(n_2-1)(\alpha_1+a_0))}\\ &\quad+\tfrac{1}{2}EFe_{-(n_1\alpha_{-1}+(n_2-1)\alpha_1+(n_2-2)\alpha_0)},\end{aligned} \tag{3.23}$$

where

$$A=\operatorname{sgn} N_{\alpha_0,n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0)},$$

$$B=\operatorname{sgn} N_{-\alpha_0,n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0)},$$

$$C=\operatorname{sgn} N_{\alpha_0,n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0)},$$

$$D=\operatorname{sgn} N_{-\alpha_0,n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0)},$$

$$E = \text{sgn } N_{-\alpha_0', -(n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0))},$$
$$F = \text{sgn } N_{\alpha_0', -(n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0))}. \tag{3.24}$$

Now $\mathbf{n}$ can be calculated from $\tilde{\mathbf{n}}$ and its elements are given by

$$-i2^{-1/2}A[e_{n_1\alpha_{-1}+n_2\alpha_1+(n_2+1)\alpha_0}+e_{-(n_1\alpha_{-1}+n_2\alpha_1+(n_2+1)\alpha_0)}]$$
$$+i2^{-1/2}B[e_{n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0}+e_{-(n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0)}],$$

$$\tfrac{1}{2}[e_{n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0}+e_{-(n_1\alpha_{-1}+n_2\alpha_1+(n_2-1)\alpha_0)}]$$
$$+i2^{-1/2}B[e_{n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0)}+e_{-(n_1\alpha_{-1}+n_2(\alpha_1+\alpha_0))}]$$
$$+\tfrac{1}{2}AB[e_{n_1\alpha_{-1}+n_2\alpha_1+(n_2+1)\alpha_0}+e_{-(n_1\alpha_{-1}+n_2\alpha_1+(n_2+1)\alpha_0)}],$$

$$\tfrac{1}{2}[e_{n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0}+e_{-(n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0)}]$$
$$-i2^{-1/2}C[e_{n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0)}+e_{-(n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0))}]$$
$$+\tfrac{1}{2}CD[e_{n_1\alpha_{-1}+(n_2-1)\alpha_1+(n_2-2)\alpha_0}+e_{-(n_1\alpha_{-1}+(n_2-1)\alpha_1+(n_2-2)\alpha_0)}],$$

$$\frac{i}{2}[e_{n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0}-e_{-(n_1\alpha_{-1}+(n_2-1)\alpha_1+n_2\alpha_0)}]$$
$$-2^{-1/2}E[e_{n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0)}-e_{-(n_1\alpha_{-1}+(n_2-1)(\alpha_1+\alpha_0))}]$$
$$+\tfrac{1}{2}EF[e_{n_1\alpha_{-1}+(n_2-1)\alpha_1+(n_2-2)\alpha_0}-e_{-(n_1\alpha_{-1}+(n_2-1)\alpha_1+(n_2-2)\alpha_0)}]. \tag{3.25}$$

The required Iwasawa decomposition is then written as

$$\hat{A}_1^{(1)} = \mathbf{k} \oplus \mathbf{a} \oplus \mathbf{n}, \tag{3.26}$$

where $\mathbf{k}$, $\mathbf{a}$, and $\mathbf{n}$ are given by Eqs. (3.13), (3.15), and (3.25), respectively.

B. Iwasawa decomposition of $\hat{A}_2^{(1)}$

As a second example, we consider the involutive automorphism of $\hat{A}_2^{(1)}$ determined by the Satake diagrams, say (iv), of Table II. The basic root automorphisms are given by

$$\sigma(\alpha_1)=\alpha_1, \qquad \sigma(\alpha_2)=\alpha_2,$$
$$-\sigma(\alpha_0)=\alpha_0+2\alpha_1+2\alpha_2, \qquad -\sigma(\alpha_{-1})=\alpha_{-1}. \tag{3.27}$$

The above equations can be written as

$$\sigma(\alpha_1)=\alpha_1, \qquad \sigma(\alpha_2)=\alpha_2,$$
$$-\sigma(\delta-\alpha_1-\alpha_2)=\delta+\alpha_1+\alpha_2, \qquad -\sigma(\alpha_{-1})=\alpha_{-1}. \tag{3.28}$$

The positive roots of $\hat{A}_2^{(1)}$ are given by

$$\Delta = \{n_1\alpha_{-1}\pm a_1+n_2\delta,\ n_1\alpha_{-1}\pm\alpha_2+n_2\delta,\ n_1\alpha_{-1}\pm(\alpha_1+\alpha_2)+n_2\delta,$$
$$n_1\alpha_{-1}+n_2\delta, \text{ where } n_1=0,1,\ldots,n_2 \text{ and } n_2\in Z^+\}. \tag{3.29}$$

We can apply the simple root automorphisms to find out the automorphisms of other roots and it is easy to see that

$$\exp\ \alpha(h)=+1 \quad \text{for} \quad \alpha=\alpha_1,\alpha_2,\ n_1\alpha_{-1}+\alpha_1+n_2\delta,\ n_1\alpha_{-1}+\alpha_2+n_2\delta,$$
$$n_1\alpha_{-1}+(\alpha_1+\alpha_2)+n_2\delta, \tag{3.30}$$

$$\exp\ \alpha(h)=-1 \quad \text{for} \quad \alpha=\alpha_{-1},\alpha_0,\ n_1\alpha_{-1}-\alpha_1+n_2\delta,$$
$$n_1\alpha_{-1}-\alpha_2+n_2\delta,\ n_1\alpha_{-1}-(\alpha_1+\alpha_2)+n_2\delta, n_2\delta+n_1\alpha_{-1}. \tag{3.31}$$

Thus, for $\hat{A}_2^{(1)}$, **k** and **p** are given by

$$\mathbf{k}=\{ih_\alpha \ \text{for} \ a=\alpha_{-1},\alpha_0,\alpha_1,\alpha_2 \ \text{and} \ (e_\alpha+e_{-\alpha}),i(e_\alpha-e_{-\alpha}) \ \text{for} \ \alpha \ \text{given by Eq. (3.30)}\}, \tag{3.32}$$

$$\mathbf{p}=\{i(e_\alpha+e_{-\alpha}),(e_\alpha-e_{-\alpha}) \ \text{for} \ \alpha \ \text{given by Eq. (3.31)}\}. \tag{3.33}$$

Again in an analogous manner, we select a maximal Abelian subalgebra **a** and we see that **a** is one-dimensional. Its basis element is given by

$$H'_{-1}=i(e_\alpha+e_{-\alpha}) \quad \text{for} \quad \alpha=-\alpha_2+m\delta. \tag{3.34}$$

So, we have $R_A=\{-\alpha_2+m\delta\}$ and R_M is empty, **m** is three-dimensional and its basis elements are given by

$$\begin{gathered} -iH'_0=i[h_{\alpha_2+(m+n)\delta}+2h_{\alpha_1+(m+n)\delta}], \\ -iH'_1=ih_{(m+n)\delta}, \qquad -iH'_2=i(h_{\alpha_{-1}}+mh_{\alpha_2}). \end{gathered} \tag{3.35}$$

The elements H'_{-1},H'_0,H'_1,H'_2 together with scaling element d' form the Cartan subalgebra $\mathbf{h}'$. Defining the inner automorphism of $\hat{A}_2^{(1)}$ as

$$V=V_{-\alpha_2+n\delta}=\exp[ad\{\alpha_{-\alpha_2+n\delta}(e_{-\alpha_2+n\delta}-e_{\alpha_2+n\delta})\}], \tag{3.36}$$

where

$$a_{-\alpha_2+n\delta}=\frac{\pi}{t^n[8(\alpha_2,\alpha_2)]^{1/2}}, \tag{3.37}$$

and applying to the Cartan subalgebra $\mathbf{h}'$, we obtain

$$\begin{aligned} H_{-1}&=2^{1/2}h_{-\alpha_2+(m+n)\delta}, \\ H_0&=-(h_{\alpha_2+(m+n)\delta}+2h_{\alpha_1+(m+n)\delta}), \\ H_1&=-h_{(m+n)\delta}, \\ H_2&=-h_{\alpha_{-1}}-\frac{m}{2}\,h_{\alpha_2}. \end{aligned} \tag{3.38}$$

The set of positive roots w.r.t. this Cartan subalgebra is given by

$$\Delta^{+}=\{(n_1\alpha_{-1}+\alpha_1+(m+n)\delta),(n_1\alpha_{-1}-\alpha_1+(m+n)\delta) \quad \text{for} \quad mn_1\neq 0,$$
$$-(n_1\alpha_{-1}-\alpha_1+(m+n)\delta) \quad \text{for} \quad mn_1=0,$$
$$(n_1\alpha_{-1}+\alpha_2+(m+n)\delta) \quad \text{for} \quad mn_1\geqslant 2,$$
$$-(n_1\alpha_{-1}+\alpha_2+(m+n)\delta) \quad \text{for} \quad mn_1=0,1,$$
$$(n_1\alpha_{-1}-\alpha_2+(m+n)\delta),(n_1\alpha_{-1}+\alpha_1+\alpha_2+(m+n)\delta) \quad \text{for} \quad mn_1\geqslant 2,$$
$$-(n_1\alpha_{-1}+\alpha_1+\alpha_2+(m+n)\delta) \quad \text{for} \quad mn_1=0,1,$$
$$(n_1\alpha_{-1}-\alpha_1-\alpha_2+(m+n)\delta),(n_1\alpha_{-1}+(m+n)\delta) \quad \text{for} \quad mn_1\neq 0,$$
$$-(n_1\alpha_{-1}+(m+n)\delta) \quad \text{for} \quad mn_1=0\}. \tag{3.39}$$

The sets Δ_{+}^{+} and Δ_{-}^{+} can similarly be written as

$$\Delta_{-}^{+}=\{(n_1\alpha_{-1}+\alpha_1+(m+n)\delta) \quad \text{for } mn_1=1,$$
$$(n_1\alpha_{-1}-\alpha_2+(m+n)\delta) \quad \text{for } mn_1=2,$$
$$(n_1\alpha_{-1}-\alpha_1-\alpha_2+(m+n)\delta) \quad \text{for } mn_1=2,$$
$$-(n_1\alpha_{-1}+(m+n)\delta) \quad \text{for } mn_1=0\}, \tag{3.40}$$

and

$$\Delta_{+}^{+}=\Delta^{+}/\Delta_{-}^{+}. \tag{3.41}$$

As before, for $\hat{A}_2^{(1)}$ the elements of $\tilde{\mathbf{n}}$ are given by $V^{-1}e_{\alpha}$, where $\alpha\in\Delta_{+}^{+}$. These structures can be calculated explicitly by applying the properties of inner automorphisms. For example, the element $V^{-1}e_{n_1\alpha_{-1}+\alpha_1+(m+n)\delta}$ is given by

$$V^{-1}e_{n_1\alpha_{-1}+\alpha_1+(m+n)\delta}=\frac{1}{1+t^{2n}}\,e_{n_1\alpha_{-1}+\alpha_1+(m+n)\delta}$$
$$+\frac{1}{1+t^{2n}}\,[\operatorname{sgn} N_{-\alpha_2,n_1\alpha_{-1}+\alpha_1+\alpha_2}]e_{n_1\alpha_{-1}+\alpha_1+\alpha_2+(m+2n)\delta}. \tag{3.42}$$

The elements of $\mathbf{n}$ can be known by considering the elements $\tilde{\mathbf{n}}\cap\tilde{\mathbf{G}}$ as before. The required Iwasawa decomposition is then written as

$$\hat{A}_2^{(1)}=\mathbf{k}\oplus\mathbf{a}\oplus\mathbf{n}. \tag{3.43}$$

IV. CONCLUDING REMARKS

The present work has concerned itself with the application of the powerful techniques of Satake diagrams to the case of hyperbolic Kac–Moody algebras with special emphasis on the Iwasawa decomposition of these algebras. The involutive automorphisms for the corresponding Satake diagrams of these algebras are determined in a rather unambiguous manner. It is pertinent to stress that the Iwasawa decomposition readily leads to Langlands decomposition from which it is straightforward to obtain parabolic subalgebras necessary for constructing induced representations. Since the Satake diagrams have been previously exploited to classify the real forms of Lie algebras,[13–15] Lie superalgebras,[16] and Z graded semisimple algebras,[21] it would indeed be possible to extend the method to classify the real forms of all hyperbolic Kac–Moody algebras. We have, however, not endeavored to discuss this aspect in this communication but expect to delve into it in the future.

TABLE III. Satake diagrams and involutive automorphisms of $G(3)$.

Satake diagrams of $G(3)$	Involutive automorphisms
(*i*) 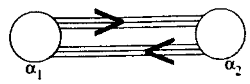	$-\sigma(\alpha_1)=\alpha_1$ $-\sigma(\alpha_2)=\alpha_2$
(*ii*) 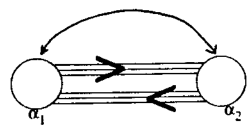	$-\sigma(\alpha_1)=\alpha_2$ $-\sigma(\alpha_2)=\alpha_1$
(*iii*) 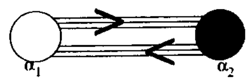	$-\sigma(\alpha_1)=\alpha_1+3\alpha_2$ $\sigma(\alpha_2)=\alpha_2$
(*iv*) 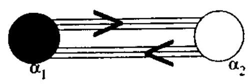	$\sigma(\alpha_1)=\alpha_1$ $-\sigma(\alpha_2)=\alpha_2+3\alpha_1$

The two examples $\hat{A}_1^{(1)}$ and $\hat{A}_2^{(1)}$ of the hyperbolic Kac–Moody algebras are typical illustrations of the working scheme and the general framework of the Satake diagrams. In principle, however, one can construct Satake diagrams for all hyperbolic Kac–Moody algebras which would indeed be a stupendous effort. We take the example of $G(3)$, which is not an extension of the affine Kac–Moody algebras and compute the Satake diagrams and the corresponding root automorphisms displayed in Table III so as to have a glimpse into the pragmatic way in which the technique has been appropriately tailored for these algebras.

[1] V. G. Kac, *Infinite Dimensional Lie Algebras*, 2nd ed. (Cambridge University Press, Cambridge, 1985).
[2] P. Goddard and D. Olive, Int. J. Mod. Phys. A **1**, 303 (1986).
[3] R. E. Borcherds, J. Algebra **115**, 501 (1988).
[4] J. Lepowsky and R. V. Moody, Math. Ann. **245**, 63 (1979).
[5] J. Harvey and G. Moore, Nucl. Phys. B **463**, 315 (1996).
[6] D. Bernard and A. Leclair, Commun. Math. Phys. **142**, 99 (1991).
[7] C. Saclioglu, J. Phys. A **22**, 3753 (1989).
[8] M. Bauer and D. Bernard, Lett. Math. Phys. **42**, 153 (1997).
[9] A. Feingold and I. Frenkel, Math. Ann. **263**, 87 (1983).
[10] A. Feingold, Proc. Am. Math. Soc. **80**, 379 (1980).
[11] R. V. Moody, Adv. Math. **33**, 144 (1979).
[12] S. J. Kang, Trans. Am. Math. Soc. **329**, 463 (1993); J. Algebra **160**, 492 (1993).
[13] O. Loos, *Symmetric Spaces* (Benjamin, New York, 1969), Vol. II.
[14] I. G. McDonald, in Proceedings of the SRC/LMS Research Symposium on Representations of Lie Groups, edited by M. F. Atiyah (Oxford University Press, Oxford, 1977), p. 91.
[15] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces* (Academic, New York, 1979).
[16] K. C. Pati and D. Parashar, J. Phys. A **31**, 767 (1998).
[17] K. C. Pati and D. Parashar, J. Math. Phys. **39**, 5015 (1998).
[18] F. Levestin, J. Algebra **114**, 489 (1989).
[19] J. F. Cornwell, J. Math. Phys. **16**, 1992 (1975); **20**, 547 (1979); J. Phys. A **25**, 2311 (1992); **25**, 2335 (1992); **25**, 3955 (1992).
[20] K. C. Pati and D. Parashar, ''Involutive automophisms and Iwasawa decomposition of affine Kac–Moody superalgebras'' (Delhi University preprint).
[21] D. Z. Djokovic, J. Algebra **76**, 367 (1982).

Propagation estimates for dispersive wave equations: Application to the stratified wave equation

David W. Pravica
Department of Mathematics, East Carolina University, Greenville, North Carolina 27858

(Received 14 March 1997; accepted for publication 19 September 1998)

The plane-stratified wave equation $(\partial_t^2+H)\psi=0$ with $H=-c(y)^2\nabla_z^2$ is studied, where $z=x\oplus y$, $x\in\mathbf{R}^k$, $y\in\mathbf{R}^1$ and $|c(y)-c_\infty|\to 0$ as $|y|\to\infty$. Solutions to such an equation are solved for the propagation of waves through a layered medium and can include waves which propagate in the x-directions only (i.e., trapped modes). This leads to a consideration of the pseudo-differential wave equation $(\partial_t^2+\omega(-\Delta_x))\psi=0$ such that the dispersion relation $\omega(\xi^2)$ is analytic and satisfies $c_1\leqslant\omega'(\xi^2)\leqslant c_2$ for $c_*>0$. Uniform propagation estimates like $\int_{|x|\leqslant|t|^\alpha}\mathcal{E}(U_t\mathcal{P}_\pm\boldsymbol{\phi}_0)d^kx\leqslant C_{\alpha,\beta}(1+|t|)^{-\beta}\int\mathcal{E}(\boldsymbol{\phi}_0)d^kx$ are obtained where U_t is the evolution group, $\mathcal{P}_\pm$ are projection operators onto the Hilbert space of initial conditions $\boldsymbol{\phi}\in\boldsymbol{\mathcal{H}}$ and $\mathcal{E}(\cdot)$ is the local energy density. In special cases scattering of trapped modes off a local perturbation satisfies the causality estimate $\|\mathcal{P}^j_{+\rho\Lambda}S\mathcal{P}^k_{-\rho\Lambda}\|\leqslant C_\nu\rho^{-\nu}$ for each $\nu<1/2$. Here $\mathcal{P}^j_{+\rho\Lambda}$ $(\mathcal{P}^k_{-\rho\Lambda})$ are remote outgoing/detector (incoming/transmitter) projections for the jth (kth) trapped mode. Also $\Lambda\Subset\mathbf{R}^+$ is compact, so the projections localize onto formally-incoming (eventually-outgoing) states. © *1999 American Institute of Physics.*
[S0022-2488(99)02601-8]

I. INTRODUCTION

The inhomogeneous-media wave equation[1] (in symmetric form[2,3]) in $\mathbf{R}^n$ is,

$$-\partial_t^2\psi=K^p\psi,\quad K^p\equiv-c_p(z)\nabla_z^2c_p(z)+V_p(z),\qquad(1)$$

$$\psi(z,t=0)=\psi_1(z),\quad \partial_t\psi(z,t=0)=\psi_2(z),$$

where the dependent variables are bounded above and below, i.e.,

$$0<\epsilon_p\leqslant c_p(z)\leqslant c_p^M,\quad 0\leqslant V_p(z)\leqslant V_p^M.\qquad(2)$$

Since K^p is a positive self-adjoint operator, we can rewrite (1) as a Schrödinger equation,

$$i\partial_t\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}=i\begin{pmatrix}0 & \sqrt{K^p}\\ -\sqrt{K^p} & 0\end{pmatrix}\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\quad \begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}_{t=0}=\begin{pmatrix}\sqrt{K^p}\psi_1\\ \psi_2\end{pmatrix},\qquad(3)$$

where the solution to the corresponding wave equation is $\psi=(1/\sqrt{K^p})\phi_1$. This form allows us to work on a vector-Hilbert space $\boldsymbol{\mathcal{H}}=\mathcal{L}^2(\mathbf{R}^n,d^nz)\oplus\mathcal{L}^2(\mathbf{R}^n,d^nz)$. By Stone's Theorem, Eq. (3) defines a unitary evolution group U_t^p on $\boldsymbol{\mathcal{H}}$ and this solves the wave equation (1) under conditions (2).

Now suppose that the space variables decompose as $z=x\oplus y\in\mathbf{R}^k\oplus\mathbf{R}^1$ and that there are bounded functions of the y-variable $c(y)\geqslant\epsilon>0, V(y)\geqslant 0$ so that,

$$\max\{|c(y)-c_p(z)|,\ |V(y)-V_p(z)|\}\to 0,\ \text{as}\ |z|\to\infty.$$

Then Eq. (1) is a perturbation of the plane-stratified wave equation,

$$-\partial_t^2\psi = K\psi, \quad K \equiv -c(y)\nabla_z^2 c(y) + V(y). \tag{4}$$

By writing Eq. (4) in the form of (3) we can define the *free* evolution group U_t which is also defined on $\boldsymbol{\mathcal{H}}$ as a family of unitary operators. Taking the Fourier transform of Eq. (4) in the x-variable gives the reduced wave equation,

$$-\partial_t^2\hat{\psi}_\xi = K_y^\xi\hat{\psi}_\xi, \quad K_y^\xi \equiv -c(y)\partial_y^2 c(y) + \xi^2 c^2(y) + V(y), \tag{5}$$

which operates on functions in $\mathcal{L}^2(\mathbf{R}^k, d^k\xi)\otimes\mathcal{L}^2(\mathbf{R}^1, dy)$. The corresponding wave equation decomposes into a *free part* associated with $\sigma_{\mathrm{cont}}(K_y^\xi)$, the continuous spectrum of K_y^ξ, and a *trapped part* associated with $\sigma_{pp}(K_y^\xi)$, the pure point spectrum of K_y^ξ. This latter set is defined by the eigenvalue equation,

$$K_y^\xi\hat{\psi}_\xi = \omega(\xi^2)\hat{\psi}_\xi, \quad \int_{\mathbf{R}^1}|\hat{\psi}_\xi(y)|^2 dy = 1. \tag{6}$$

Here $\hat{\psi}_\xi(y)\in\mathcal{L}^2(\mathbf{R}^1, dy)$ are normalized eigenmodes of the reduced stratified system. From (6) a family of eigenvalues are obtained $\{\omega_j(\xi^2)\}$ along with associated *thresholds* $\{\xi_j\}$ and eigenfunctions $\{\hat{\psi}_\xi^j\}$ so that $\omega_j(\xi^2)$ and $\hat{\psi}_\xi^j$ are locally analytic for $|\xi|>\xi_j$. We can now define the jth-*mode subspace* as,

$$\mathcal{H}_j \equiv \mathrm{closure}\left\{ \psi(x,y) = \int_{|\xi|>\xi_j} e^{ix\cdot\xi} f(\xi)\hat{\psi}_\xi^j(y)\, d^k\xi \,\middle|\, f\in\mathcal{L}^2(\mathbf{R}^k)\right\}.$$

On initial states $\psi_1, \psi_2 \in \mathcal{H}_j$ the stratified wave equation (4) becomes,

$$-\partial_t^2\psi^j = \omega_j(-\Delta_x)\psi^j, \quad \psi^j(x,y,0) = \psi_1(x,y), \quad \partial_t\psi^j(x,y,0) = \psi_2(x,y), \tag{7}$$

which is a pseudo-differential wave equation on $\mathbf{R}^k$. Solutions to (7) $\psi^j(x,y,t)$ can be expressed using Fourier transforms. We will call them *trapped* or *channeled modes* since they propagate in the x-directions only.[4]

To study (1) using the asymptotics of (4) we construct the scattering operator,

$$S \equiv \Omega_-^*\Omega_+, \quad \Omega_\pm \equiv s-\lim_{t\to\pm\infty} U_t^p U_{-t}. \tag{8}$$

Next, we introduce a class of projection operators $\mathcal{P}_\Lambda$ for $\Lambda \Subset \mathbf{R}^+$, so that $\forall\,\boldsymbol{\phi}\in\boldsymbol{\mathcal{H}}$,

$$\|\chi_+(|\rho|^\alpha - |x|)\mathcal{P}_{\rho\Lambda}\boldsymbol{\phi}\| \leqslant C_\Lambda|\rho|^{-\beta}\|\boldsymbol{\phi}\|, \tag{9}$$

for $0<\alpha\leqslant 1$ and some $\beta>0$. Here $\chi_+(s)=1$ for $s>0$ and 0 otherwise. Property (9) implies that $\mathcal{P}_{\rho\Lambda}$ localizes states outside a ball of radius $|\rho|^\alpha$, leaving a small polynomial tail inside the ball. For each modal subspace $\mathcal{H}_j$ it is possible to define orthogonal projections $\mathcal{P}_\Lambda^j$ satisfying (9). By taking the perspective that $\mathcal{P}_\Lambda$ projects onto outgoing scattering states (and $\mathcal{P}_{-\Lambda}$ onto incoming states) we can ask whether perturbations cause a leakage of incoming into outgoing modes during the scattering process. In particular, *causality estimates* of the form,

$$\|\mathcal{P}_{+\rho\Lambda}^j S\mathcal{P}_{-\rho\Lambda}^k\| \leqslant C_\nu/\rho^\nu, \tag{10}$$

are obtained, for each $\nu\in(0,1/2)$. We say that ''asymptotically, incoming channeled modes do not effect outgoing states.'' Here $\Lambda\Subset\mathbf{R}^+\equiv(0,\infty)$ is a compact set, so for $\rho>0$, $\mathcal{P}_{+\rho\Lambda}^k$ projects onto outgoing states which are formally-incoming and similarly $\mathcal{P}_{-\rho\Lambda}^j$ projects onto incoming states which are eventually-outgoing. Also, due to the difficulty of this problem, we only consider strictly positive thresholds $\xi_j>0$ in a media with a single layer. Extensions are possible to all thresholds $\xi_j\geqslant 0$ and to many cases of multi-layered media.

Inequality (10) is a type of cluster property in time, which was introduced by Taylor.[5] Our analysis uses $x\cdot\hat{\nabla}_x$ the *formal-time operator* originally used by Lavine[6] and expanded upon by Mourre.[7] For further history on this approach see the work of Perry.[8] Though the phase space techniques which we use are standard,[9] there has been little work on the behavior near thresholds. For stratified media problems, restrictions on energy-decay are due mainly to thresholds and not due to dispersion. Hence our main goal is to obtain uniform-energy-propagation estimates and then to use these estimates to obtain properties like (10). We refer to Ref. 10 for a discussion of other properties of S which are obtainable through phase space analysis.

In this paper we have chosen to study the wave equation because working with a vector-Hilbert space allows us to define a *global* time operator. In the special case that the threshold vanishes $\xi_j=0$, that there is no dispersion $\omega_j''=0$, that the dimension is odd and the perturbation is compactly supported, Lax and Phillips have shown that $\mathcal{P}_{+\rho}S\mathcal{P}_{-\rho}=0$, i.e., the system is *completely causal*.[11] Our analysis applies to noncompact perturbations in any dimension so we can only show that the scattering operator is *weakly* causal. For further discussion about the effects of dispersion on causality, we suggest consulting the work of Nussenzveig.[12]

The paper is organized as follows. In the next section we review the theory of wave equations as needed for this work. Section III then begins with an explicit definition of the projection operators $\mathcal{P}_\Lambda$ and a presentation of important properties. The section ends with a series of Propositions intended to justify the terminology incoming/outgoing. Finally in Sec. IV we consider the scattering operator for the stratified wave equation and obtain causality estimates of the form (10). Some calculations are left for the Appendix.

II. PRELIMINARIES

We begin with a discussion of the wave equation which also applies to the dispersive case. Let us rewrite (7) as a system,

$$i\partial_t\begin{pmatrix}\psi\\ \partial_t\psi\end{pmatrix}=i\begin{pmatrix}0 & 1\\ -\omega_j(-\Delta_x) & 0\end{pmatrix}\begin{pmatrix}\psi\\ \partial_t\psi\end{pmatrix},\qquad \begin{pmatrix}\psi\\ \partial_t\psi\end{pmatrix}_{(t=0)}=\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}. \tag{11}$$

For each channel define the *spectral thresholds* $\tau_j\geqslant 0$, where $\tau_j^2=\omega_j(\xi_j^2)$. Then,

$$\text{Domain}(\omega_j)=[\xi_j^2,\infty),\qquad \text{Range}(\omega_j)=[\tau_j^2,\infty).$$

Energy of the solution to (11) is, for $h=\omega^2$

$$E(\psi_0,\psi_1;t)\equiv\int_{\mathbf{R}^k}(|h(|\nabla_x|)\psi|^2+|\partial_t\psi|^2)d^kx=\int_{\mathbf{R}^k}\mathcal{E}(\psi)d^kx, \tag{12}$$

where the *local-energy density* at time t is defined,

$$\mathcal{E}(\psi)=\left|\left(\frac{1}{-\Delta_x}\omega(-\Delta_x)\right)^{1/2}\nabla_x\psi\right|^2+|\partial_t\psi|^2. \tag{13}$$

This agrees with the usual notion of local-energy density when $\omega(\cdot)$ is a homogeneous function of degree one. The total energy E is a conserved quantity and functions with finite energy define the Hilbert space on which we will work.

In this paper we suppose that $c(y)$ and $V(y)$ satisfy one of the following conditions:

(C1) (a) $V,c\in\mathcal{C}^\infty$ and $\exists\,\gamma_b>0$ so that $\|e^{\gamma_b|y|}V(y)\|_\infty+\|e^{\gamma_b|y|}(c(y)-c_\infty)\|_\infty<\infty$;

(b) $V(y)$ and $c(y)$ are finitely-piecewise constant, with $V(y)$ and $c(y)-c_\infty$ compactly supported.

Then we have the following important result due to the work of Klaus and Simon.[13]

Theorem 2.1: Suppose that (C1) (a) or (b) hold. Then $\exists\{\omega_j(\xi^2)\}$ a sequence of nondegenerate negative eigenvalues of (6) which are locally analytic for $|\xi|>\xi_j$ and continuous for $|\xi|\geqslant\xi_j$. Furthermore, $\exists c_*,\epsilon_*,l_*$ strictly positive constants, so that,

(i) $c_1\leqslant\omega'(\xi^2)\leqslant c_2$ for all $|\xi|\geqslant\xi_j$;
(ii) $|\xi^{2(l-1)}\omega_j^{(l)}(\xi^2)|\leqslant c_3$ for all $\xi^2>\xi_j^2+\epsilon_\infty$ and $l\leqslant l_\infty$;
(iii) If $\tau_j>0$ and $\xi_j>0$, then $\lim_{\xi\to\xi_{\tau_j}^+}|(\xi^2-\xi_j^2)^{\epsilon_j}\omega^{(l)}(\xi^2)|\leqslant c_4$ for all $l\leqslant_j$;
otherwise, if $\tau_j=\xi_j=0$ then $\lim_{\xi\to 0^+}|\xi^{2(l-1)}\omega^{(l)}(\xi^2)|\leqslant c_5$ for all $l\leqslant l_j$.

Now define $h(\xi)\equiv(\omega(\xi^2))^{1/2}$ where we abuse notation by writing $\xi=|\xi|$ and by ignoring the j-index, i.e., $\omega\equiv\omega_j, h\equiv h_j, \xi_\tau\equiv\xi_j$, etc. Then, to generalize our results, suppose that $h(\xi)$ has the following properties, which are weaker than actually required,

(P1) Domain$(h)=[\xi_\tau,\infty)$, Range$(h)=[\tau,\infty)$, $\tau=h(\xi_\tau)$;
(P2) $\exists c_*,\epsilon_\infty,l_\infty>0$, so that
 (a) $c_1\leqslant h'(\xi)\leqslant c_2$ for all $\xi\geqslant\xi_\tau$;
 (b) $h\in\mathcal{C}^{l_\infty}(\xi_\tau+\epsilon_\infty,\infty)$ and $|\xi^{(l-1)}h^{(l)}(\xi)|\leqslant c_3$ for all $\xi>\xi_\tau+\epsilon_\infty$ and $l\leqslant l_\infty$;
(P3) $\exists\epsilon_\tau,l_\tau>0$, so that
 if $\tau>0$ and $\xi_\tau>0$ then $\lim_{\xi\to\xi_\tau^+}|(\xi-\xi_\tau)^{\epsilon_\tau}h^{(l)}(\xi)|\leqslant c_4$ for all $l\leqslant l_\tau$;
 otherwise, if $\tau=\xi_\tau=0$ then $\lim_{\xi\to 0^+}|\xi^{l-1}h^{(l)}(\xi)|\leqslant c_4$ for all $l\leqslant l_\tau$.

Now denote $\phi_1=h\psi$, $\phi_2=\partial_t\psi$. Then the wave equation (11) has the alternate form,

$$i\partial_t\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}=i\begin{pmatrix}0 & h(|\nabla_x|)\\ -h(|\nabla_x|) & 0\end{pmatrix}\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\qquad \begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}_{(t=0)}=\begin{pmatrix}h(|\nabla_x|)\psi_1\\ \psi_2\end{pmatrix}. \tag{14}$$

Evolution is unitary on the Hilbert space,

$$\boldsymbol{\mathcal{H}}_\tau\equiv\mathcal{H}_\tau\oplus\mathcal{H}_\tau,$$

$$\mathcal{H}_\tau\equiv E^{-\Delta_x}_{[\tau^2,\infty)}\mathcal{L}^2(\mathbf{R}^k,d^kx)=\text{closure}\left\{\psi(x)=\int_{|\xi|>\xi_\tau}e^{ix\cdot\xi}f(\xi)d^k\xi\,\middle|\,f\in\mathcal{L}^2(\mathbf{R}^k)\right\},$$

where E^K_Λ is the spectral measure of a self-adjoint operator K on the set Λ. The norm of $\boldsymbol{\mathcal{H}}_\tau$ is the sum of $\mathcal{L}^2$-norms which is equivalent to the energy in (12). Furthermore, the system in (14) can be diagonalized by defining,

$$\phi_\pm=(1/2)^{1/2}(\phi_1\mp i\phi_2)=(1/2)^{1/2}(h(|\nabla_x|)\psi\mp i\partial_t\psi). \tag{15}$$

Then it is clear that,

$$E(\psi_0,\psi_1)=\|\phi_1\|^2+\|\phi_2\|^2=\|\phi_+\|^2+\|\phi_-\|^2.$$

To solve the dispersive wave equation (14) we introduce the Fourier transform on $\mathbf{R}^k$,

$$\mathcal{F}_x\psi(\xi\mu)=\Omega_k\int_{\mathbf{R}^k}e^{i(x\cdot\mu)\xi}\psi(x)d^kx,$$

for $\mu\in S^{k-1}$ and $\Omega_k\equiv(2\pi)^{-k/2}$. Making the extension,

$$h(-\xi)=-h(\xi),\qquad \mathcal{D}_\tau\equiv\text{Domain}(h)=(-\infty,-\xi_\tau]\cup[\xi_\tau,\infty),$$

allows (14) to be rewritten as,

$$i\partial_t f(\xi,\mu;t)=-h(\xi)f(\xi,\mu;t),\quad\text{on}\quad\mathcal{L}^2(\mathcal{D}_\tau\otimes S^{k-1},|\xi|^{k-1}d\xi d\mu),$$

where, in terms of solutions to (14) and the notation of (15), we define,

$$f(\xi,\mu;t)\equiv(\mathcal{F}_x\phi_\pm)(\xi,\mu;t), \quad \text{for } \xi\in\mathcal{D}_\tau \quad \text{and } \pm\xi\geqslant 0.$$

In this way the wave equation becomes first order. A study of this equation in the case that $h(\xi)$ is a homogeneous function of degree 1 with $\xi_\tau=0$ was carried out in Ref. 11, Sec. VI. We also refer to Ref. 14 for a detailed discussion of *domains of definition*.

In this paper, instead of using $\mathcal{F}_x$, we work with the transformation,

$$\mathcal{F}^h\psi(q,\mu)=\Omega_k H_k(q)\int_{\mathbf{R}^k} e^{i(x\cdot\mu)h^{-1}(q)}\psi(x)d^kx,$$

where the dispersion weight factor is defined as,

$$H_k(q)=\frac{|h^{-1}(q)|^{(k-1)/2}}{\sqrt{h'(h^{-1}(q))}},$$

for $q\in\sigma_\tau$. Here $\sigma_\tau\equiv(-\infty,-\tau]\cup[\tau,\infty)$ and in this case Eq. (14) becomes,

$$i\partial_t g(q,\mu;t)=-qg(q,\mu;t), \quad \text{on} \quad \mathcal{L}^2(\sigma_\tau\otimes S^{k-1},dq\,d\mu),$$

where $g(q,\mu;t)\equiv(\mathcal{F}^h\phi_{\mathrm{sgn}(q)})(q,\mu;t)$. Finally, applying the inverse 1-d Fourier transform,

$$(\mathcal{F}_1^{-1}g)(s)\equiv\Omega_1\int_{\sigma_\tau} e^{-isq}g(q)dq, \quad s\in\mathbf{R},$$

gives the one-dimensional wave equation,

$$\partial_t k(s,\mu,t)=-\partial_s k(s,\mu,t), \quad \text{on} \quad \mathcal{L}^2(\mathbf{R},ds;\mathcal{N}), \quad \mathcal{N}\equiv\mathcal{L}^2(S^{k-1},d\mu). \tag{16}$$

The s-variable is conjugate to $-i\partial_s$, (i.e., $i[-i\partial_s,s]=1$) and the solution of (16) is $k(s,\mu,t)=k_0(s-t,\mu)$, which is just translation to the right in the s-variable for some initial condition function $k_0(\cdot,\cdot)$. The solution to (11) is $\psi(x,t)=[h(|\nabla_x|)]^{-1}\phi(x,t)$, which can be an equivalence class of functions if $\tau=0$ (see discussions in Refs. 1 and Ref. 15). Here,

$$\phi(x,t)=\Omega_{k+1}\int_{\sigma_\tau\otimes S^{k-1}} e^{-ix\cdot\mu h^{-1}(q)}H_k(q)\left(\int_{-\infty}^{\infty} e^{isq}k_0(s-t,\mu)ds\right)dq\,d\mu,$$

and the Lax–Phillips transformation of the initial data are,

$$\begin{aligned} k_0(s,\mu)&\equiv\mathcal{LP}(\psi_0,\psi_1)(s,\mu)\\ &=\partial_s\mathcal{R}(\psi_0)(s,\mu)-\mathcal{R}(\psi_1)(s,\mu)\\ &=\Omega_{k+1}\int_{\sigma_\tau} e^{-is\overline{q}}H_k(\overline{q})\left(\int_{\mathbf{R}^k} e^{i\overline{x}\cdot\mu h^{-1}(\overline{q})}\mathrm{sgn}(\overline{q})[\overline{q}\psi_0(\overline{x})-i\psi_1(\overline{x})]d^k\overline{x}\right)d\overline{q}. \end{aligned}$$

$\mathcal{R}(\cdot)$ is called the Radon transformation and it is used to define the projection operators $\mathcal{P}_\pm^{LP}$ on $\boldsymbol{\mathcal{H}}$ so that,

$$\mathcal{LP}(\mathcal{P}_\pm^{LP}(\psi_0,\psi_1))(s,\mu)=\chi_\pm(s)\mathcal{LP}(\psi_0,\psi_1)(s,\mu).$$

Here $\chi_\pm(s)=1$ for $s\gtrless 0$ respectively and 0 otherwise. The spaces $\boldsymbol{\mathcal{D}}_\pm^{LP}\equiv\mathcal{P}_\pm^{LP}\boldsymbol{\mathcal{H}}$ are called *outgoing/incoming* in the sense of Lax–Phillips. Let U_t be the unitary group which solves (14) on $\boldsymbol{\mathcal{H}}_\tau$. Then $\boldsymbol{\mathcal{D}}_\pm^{LP}$ have the well known properties,

$$\text{(i)}\quad U_t\boldsymbol{\mathcal{D}}_{\pm}^{LP}\subset\boldsymbol{\mathcal{D}}_{\pm}^{LP},\quad t\gtrless 0;\quad \text{(ii)}\quad \bigcap_{t\in\mathbf{R}} U_t\boldsymbol{\mathcal{D}}_{\pm}^{LP}=\{\mathbf{0}\};\quad \text{(iii)}\quad \overline{\bigcap_{t\in\mathbf{R}} U_t\boldsymbol{\mathcal{D}}_{\pm}^{LP}}=\boldsymbol{\mathcal{H}},\tag{17}$$

which follow from the Weyl relations, which hold due to the translation representation of (16). The difficulty in studying the wave equation in this form is that if $\tau\neq 0$ then the variable s does not act like simple differentiation under Fourier transform. In fact,

$$\mathcal{F}_1(sk_0(\cdot,\mu))(q)=\Omega_1\int_{-\infty}^{\infty}e^{isq}sk_0(s,\mu)ds=(i\partial_q+\delta(q-\tau)-\delta(q+\tau))\mathcal{F}_1(sk_0(\cdot,\mu))(q),$$

so $\mathcal{P}_{\pm}^{LP}$ are t-dependent. However if $\tau=0$ then $\mathcal{F}_1: s\to i\partial_q$ as operators. Thus Lax–Phillips theory only applies to the $\tau=0$ case, i.e., systems with uniform multiplicity.[11,15]

III. ENERGY REPRESENTATION

In this section we introduce projection operators $\mathcal{P}_{\pm}$ for the dispersive wave equation without using the Radon transform. This leads to a decomposition of the initial condition space into $\boldsymbol{\mathcal{D}}_{\pm}=\mathcal{P}_{\pm}\boldsymbol{\mathcal{H}}_{\tau}$. These outgoing and incoming spaces agree with the Lax–Phillips definition only if $\tau=0$ and k is odd. Propagation estimates are obtained for elements in $\boldsymbol{\mathcal{D}}_{+}$. Corresponding results hold for $\boldsymbol{\mathcal{D}}_{-}$.

Define $\phi_{*0}=\phi_*(t=0)\in\mathcal{H}_{\tau}$. We also use the notations, for $q\in\mathbf{R}$,[1]

$$\hat{q}\equiv\operatorname{sgn}(q),\quad \langle q\rangle\equiv(1+q^2)^{1/2},$$

$$\hat{v}_q=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\ i\hat{q}\end{pmatrix},\quad \phi_{\hat{q}}=\frac{1}{\sqrt{2}}(\phi_1-i\hat{q}\phi_2)=\hat{v}_q^T\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},$$

[see (15)]. Then returning to (14), the solution of the wave equation is given by,

$$\begin{aligned}\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}(t)&=U_t\begin{pmatrix}\phi_{10}\\ \phi_{20}\end{pmatrix}\\ &\equiv\begin{pmatrix}\cos ht & \sin ht\\ -\sin ht & \cos ht\end{pmatrix}\begin{pmatrix}\phi_{10}\\ \phi_{20}\end{pmatrix}\\ &=\frac{1}{\sqrt{2}}\begin{pmatrix}e^{iht}\phi_{+0}+e^{-iht}\phi_{-0}\\ ie^{iht}\phi_{+0}-ie^{-iht}\phi_{-0}\end{pmatrix}\\ &=\frac{\Omega_k^2}{\sqrt{2}}\int_{\sigma_\tau\otimes\mathcal{S}^{k-1}}e^{i(qt-x\cdot\mu h^{-1}(q))}H_k^2(q)\left(\int_{\mathbf{R}^k}e^{i\overline{x}\cdot\mu h^{-1}(q)}\begin{pmatrix}\phi_{\hat{q}0}\\ i\hat{q}\phi_{\hat{q}0}\end{pmatrix}d^k\overline{x}\right)dq\,d\mu.\end{aligned}\tag{18}$$

Let $\chi_\Lambda(s)$ be the characteristic function of the set $\Lambda\subset\mathbf{R}$ and denote $\chi_{\pm}(s)\equiv\chi_{\mathbf{R}^+}(\pm s)$. Then with $\gamma\equiv\hat{q}^{(k+1)/2}$ for k odd and $\tau=0$, and $\gamma=1$ otherwise, define $\mathcal{P}_\Lambda$ so that,

$$\begin{aligned}\mathcal{P}_\Lambda\begin{pmatrix}\phi_{10}\\ \phi_{20}\end{pmatrix}&=\Omega_k\int_{\sigma_\tau\otimes\mathcal{S}^{k-1}}e^{-ix\cdot\mu h^{-1}(q)}H_k(q)\mathcal{F}^h\left[\mathcal{P}_\Lambda\begin{pmatrix}\phi_{10}\\ \phi_{20}\end{pmatrix}\right]dq\,d\mu\\ &\equiv\Omega_k\int_{\mathbf{R}\otimes\mathcal{S}^{k-1}}e^{-ix\cdot\mu h^{-1}(q_p)}H_k(q_p)\hat{v}_p(\gamma\chi_\Lambda(-i\partial_p)\gamma)\hat{v}_p^T\begin{pmatrix}\mathcal{F}^h\phi_{10}\\ \mathcal{F}^h\phi_{20}\end{pmatrix}dp\,d\mu,\end{aligned}\tag{19}$$

where we use the change of variables $p=q-\tau\hat{q}\in(-\infty,\infty)$ and define the quantity $q_p\equiv p+\tau\hat{p}\in\sigma_\tau$. Here $\mathcal{P}_\Lambda\equiv\chi_\Lambda(\mathcal{A}^h)$ where $\mathcal{A}^h$ is a self-adjoint extension (see the work of Nagy[16]) of the following symmetric operator:

$$\mathcal{A}^h\equiv i\begin{pmatrix}0 & A_h\\ -A_h & 0\end{pmatrix},\quad \text{on}\quad \boldsymbol{\mathcal{D}}(\mathcal{A}^h)=\{\boldsymbol{\phi}\in\mathcal{D}^2(A_h)\subset\boldsymbol{\mathcal{H}}_\tau|\mathcal{F}^h\boldsymbol{\phi}(\pm\tau)=\mathbf{0}\},$$

$$A_h\equiv\frac{1}{4i}\frac{1}{\sqrt{h'(|\nabla_x|)}}(\hat{\nabla}_x\cdot x+x\cdot\hat{\nabla}_x)\frac{1}{\sqrt{h'(|\nabla_x|)}}.$$

For $\Lambda=(a,b)\subset\mathbf{R}^+$, $\rho>0$, we use the notations,

$$\mathcal{P}_{+\rho\Lambda}\equiv\mathcal{P}_{(\rho a,\rho b)},\quad \mathcal{P}_{-\rho\Lambda}\equiv\mathcal{P}_{(-\rho b,-\rho a)},\quad \mathcal{P}_{\pm\rho}\equiv\mathcal{P}_{\pm[\rho,\infty)},\quad \mathcal{P}_\pm=\mathcal{P}_{\pm 0}\equiv\mathcal{P}_{\pm\mathbf{R}^+}.$$

To demonstrate that $\boldsymbol{\mathcal{D}}_\pm\equiv\mathcal{P}_\pm\boldsymbol{\mathcal{H}}_\tau$ are orthogonal spaces, we prove in the Appendix,

Lemma 3.1: The operators $\mathcal{P}_\pm$ are complementary-orthogonal projections. In particular $\mathcal{P}_+\mathcal{P}_-=0$, $\mathcal{P}_++\mathcal{P}_-=1$, $\mathcal{P}_\pm^*=\mathcal{P}_\pm$.

We also consider the following in the Appendix:

Claim 3.2: Let U_t be the unitary group defined in (18) on $\boldsymbol{\mathcal{H}}_\tau$. Then for any $\tau\geqslant 0$ properties (ii) and (iii) of Lax–Phillips hold. Furthermore,

(i)′ (a) $\mathcal{P}_\mp U_t\overset{s}{\rightarrow}0$ and $U_t\mathcal{P}_\pm\overset{w}{\rightarrow}0$ as $t\rightarrow\pm\infty$;

(b) $(1-\mathcal{P}_{+t})U_t\mathcal{P}_+=0$ for $t=(n\pi/\tau)$ if $\tau>0, n\in\mathbf{N}$ and $\forall t>0$ if $\tau=0$.

Remark: Property (i) in (17) can be restated as $\mathcal{P}_\mp U_t\mathcal{P}_\pm=0, \forall t\gtrless 0$. Lax and Phillips also discuss the operator $Z(t)\equiv(1-\mathcal{P}_{+\rho})U_t^p(1-\mathcal{P}_{-\rho})$, where U_t^p is the evolution group of a compactly perturbed system. If $\|Z(t)\|\rightarrow 0$ as $t\rightarrow\infty$ then the perturbed system will have uniform energy-decay properties. For the problem studied here, property (i)′ only says that $Z(t)$ vanishes strongly, even if there is no perturbation.

The claim is established from estimates of the form,

$$\|\chi_+(c_m'|t\pm\rho|^\alpha-|x|)U_t\mathcal{P}_{\pm\rho\Lambda}\boldsymbol{\phi}_0\|^2\leqslant\frac{C(\alpha,\beta,\Lambda)}{(1+|t\pm\rho|)^\beta}\|\boldsymbol{\phi}_0\|^2,\quad t\gtrless 0,\quad \rho>0. \tag{20}$$

If $\exists\,\alpha\geqslant 0,\beta>0,\Lambda\subseteq\mathbf{R}^+$ so that (20) holds $\forall\,\boldsymbol{\phi}_0\in\boldsymbol{\mathcal{H}}_\tau$ then it is called a *uniform-energy-propagation estimate*. Note that uniformity refers to the absence of momentum cutoff functions on either $\boldsymbol{\phi}_0$ or U_t.

Discussion: Equation (20) is not just a propagation estimate. For $t=0$ and Λ compact, it shows that $\mathcal{P}_{+\rho\Lambda}\boldsymbol{\mathcal{H}}_\tau$ are outgoing states, which were once incoming. Similarly $\mathcal{P}_{-\rho\Lambda}\boldsymbol{\mathcal{H}}_\tau$ consists of incoming states which are eventually outgoing. These propagation estimates allow us to replace strict notions of incoming/outgoing with an energy-density estimate. If $\tau=0$ and the dimension is odd, then localization is nearly exponential, but if $\tau>0$ one can only localize the energy in space in terms of inverse powers of ρ.

We now present our main results for system (14) under conditions (P1)-(P3), concentrating on the $\Lambda=\mathbf{R}^+$ case and considering only outgoing projections.

Proposition 3.3: Suppose that k is odd, $\tau=0$, $N\equiv\min\{l_\infty,l_\tau\}-1>(k+1)/2$ and $\alpha<1$. Then for $\beta\equiv 2N-\alpha k-1$ and $t,\rho>0$ the estimate,

$$\int\chi_+(c_m'|t+\rho|^\alpha-|x|)\mathcal{E}(U_t\mathcal{P}_{+\rho}\boldsymbol{\phi}_0)d^kx\leqslant C_{\alpha,\beta}(1+|t+\rho|)^{-\beta}E(\boldsymbol{\phi}_0), \tag{21}$$

holds for any $\boldsymbol{\phi}_0 \in \boldsymbol{\mathcal{H}}_\tau$. If $\alpha=1$ then for each $c_m' < c_m$ there is a $C>0$ so that (21) holds. Furthermore, if $h(\xi)=c_m\xi$ (i.e., homogeneous of degree one) then we can take $C=0$ for any $\alpha \leqslant 1$, $c_m' < c_m$.

Remark: The last statement is Huygen's (energy) principle which holds for odd $k \geqslant 1$.

Proof: We extend the methods of Theorem 2.6.1 in Ref. 8 by introducing a scaling argument. For any $B \gg 1$ define the symmetric functions,

$$\mathcal{G}(q)=1-G(q/B), \quad \text{where} \quad G(q)=\begin{cases} 0, & |q|\leqslant 1 \\ 1, & |q|\geqslant 2 \end{cases}, \quad qG'(q)\geqslant 0, \quad G\in\mathcal{C}^\infty(\mathbf{R}).$$

Then the left-hand side of (21), can be written, for $\rho=0$, $c_m'=1$, $\alpha<1$ and each $t>0$,

$$\int_{|x|\leqslant t^\alpha} \mathcal{E}(U_t\mathcal{P}_+\boldsymbol{\phi}_0)d^kx \leqslant 2\|G(|\nabla_x|/B)\mathcal{P}_+\boldsymbol{\phi}_0\|^2 + 2t^{\alpha k}\mathrm{Vol}(\mathcal{S}^{k-1}) \sup_{|x|\leqslant t^\alpha} \mathcal{E}(U_t\mathcal{G}(|\nabla_x|)\mathcal{P}_+\boldsymbol{\phi}_0).$$

The first term on the right-hand side vanishes as $B\to\infty$. As for the last term, first note that $-i\partial_q$ is a self-adjoint operator on $\mathcal{L}^2(\mathbf{R}^1,dq)$. Thus the Weyl relations hold,

$$e^{iqt}e^{s\partial_q}=e^{-ist}e^{s\partial_q}e^{iqt} \Leftrightarrow e^{iqt}\chi_\Lambda(-i\partial_q)=\chi_{\Lambda+t}(-i\partial_q)e^{iqt}. \tag{22}$$

Now, applying expression (19), then (18) and definition (13), gives,

$$\begin{aligned}
&\mathcal{E}(U_t\mathcal{G}(|\nabla_x|)\mathcal{P}_+\boldsymbol{\phi}_0) \\
&\equiv \mathcal{E}\left(U_t\Omega_k \int_{\mathbf{R}\otimes\mathcal{S}^{k-1}} e^{-ix\cdot\mu h^{-1}(q_p)}\mathcal{G}(h^{-1}(q_p))H_k(q_p)\hat{v}_p(\gamma\chi_+(-i\partial_p)\gamma)\hat{v}_p^T\mathcal{F}^h\boldsymbol{\phi}_0 dp\, d\mu \right) \\
&\equiv \frac{\Omega_k^2}{\sqrt{2}}\left| \begin{pmatrix} \hat{\nabla}_x & 0 \\ 0 & 1 \end{pmatrix} \int_{\sigma_\tau\otimes\mathcal{S}^{k-1}} e^{i(qt-x\cdot\mu h^{-1}(q))}\mathcal{G}(h^{-1}(q))H_k(q)\hat{v}_q \right. \\
&\qquad \left. \times(\gamma\chi_+(-i\partial_q)\gamma)\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}_0 dq\, d\mu \right|^2 \\
&\equiv \frac{\Omega_k^2}{2}\left| \int dq\, d\mu e^{i(qt-x\cdot\mu h^{-1})}\mathcal{G}\circ h^{-1}H_k\begin{pmatrix} -i\hat{q} \\ i\hat{q} \end{pmatrix}(\gamma\chi_+\gamma)\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}_0 \right|^2.
\end{aligned} \tag{23}$$

Consider the linear case where $q=h(\xi)=c_m\xi$ and $\tau=0$. We then use self-adjointness of $i\partial_q$ on $\mathcal{L}^2(\mathbf{R},dq)$, relation (22) and the Schwartz inequality to bound (23) by,

$$\Omega_k^2\int dq\, d\mu |\chi_{(-\infty,-t)}(-i\partial_q)e^{-ix\cdot\mu q/c_m}\mathcal{G}(q/c_m)\gamma\hat{q}|q|^{(k-1)/2}|^2 \times \int dq\, d\mu|\gamma e^{iqt}\mathcal{F}^h\boldsymbol{\phi}_0|^2.$$

Next, by the Parseval equality we obtain an upper bound of,

$$\mathrm{Vol}(S^{k-1})\sup_\mu \int_{-\infty}^{-t}\left| \int_{-2B}^{2B} e^{-iq(s+x\cdot\mu/c_m)}\mathcal{G}(q/c_m)q^{(k-1)/2}dq\right|^2 ds \times 2E(\boldsymbol{\phi}_0).$$

Hence $\exists C_*>0$ so that applying integration by parts M-times gives a larger upper bound,

$$\begin{aligned}
&C_kE(\boldsymbol{\phi}_0)\sup_\mu\int_{-\infty}^{-t}\left|\int_{-2B}^{2B}\left(\frac{i}{s+x\cdot\mu/c_m}\right)^M[\partial_q^M e^{-iq(s+x\cdot\mu/c_m)}]\mathcal{G}(q/c_m)q^{(k-1)/2}dq\right|^2 ds \\
&\leqslant C_{k,M}E(\boldsymbol{\phi}_0)\int_{-\infty}^{-t}(|s|-|x|/c_m)^{-2M}ds\left(\int_{-2B}^{2B}|\partial_q^M\mathcal{G}(q/c_m)q^{(k-1)/2}|dq\right)^2,
\end{aligned} \tag{24}$$

from which we can easily conclude that,

$$\sup_{|x|\leqslant t^{\alpha}} \mathcal{E}(U_t\mathcal{G}(|\nabla_x|)\mathcal{P}_+\boldsymbol{\phi}_0)\leqslant C_{k,M,\alpha}t^{-2M+1}(2B)^{-2M+k+1}E(\boldsymbol{\phi}_0). \tag{25}$$

Hence as $B\to\infty$ (for M sufficiently large) the right-hand side of (25) vanishes for each fixed t. Replacing t with $t+\rho$ at appropriate points verifies the last statement of Proposition 3.3.

Remark: All remaining computations required in this section are modifications of the steps in (23)–(25). Repetitions will be kept to a minimum.

Next, in case that $h(\xi)=q$ is nonlinear and $\tau=0$, let us define the phase function,

$$\varphi_0(q)=sq+x\cdot\mu h^{-1}(q)=sh(\xi)+x\cdot\mu\xi,$$

and the operators,

$$L\equiv\frac{i}{\varphi_0'(q)}\partial_q,\quad L^*\equiv-i\frac{\varphi_0''(q)}{(\varphi_0'(q))^2}+\frac{i}{\varphi_0'(q)}\partial_q,$$

or, using the change of variables identities $1=h'(\xi)\partial_q h^{-1}(q)$ and $\partial_q=(h')^{-1}\partial_\xi$,

$$L^*\equiv i\frac{x\cdot\mu h''/h'}{(sh'+x\cdot\mu)^2}+\frac{i}{sh'+x\cdot\mu}\partial_\xi.$$

For $s\leqslant -t$ and $|x|\leqslant(c_m'/c_m)t$, (24) gives,

$$\left|\int_{-2B}^{2B}e^{-i\varphi_0(q)}(L^*)^N\mathcal{G}(h^{-1}(q))\gamma\hat{q}H_k(q)dq\right|\leqslant\frac{C_N}{(|s|-c_m't/c_m)^N}\int_{-2B}^{2B}\langle q\rangle^{-N+(k-1)/2}dq, \tag{26}$$

where C_N is finite due to the smoothness conditions (P2) and (P3). The right-hand side of (26) has a finite limit as $B\to\infty$ if $N-(k-1)/2>1$. Integrating for $s\in(-\infty,-t)$ completes the proof for $\rho=0$. The $\rho>0$ cases are similarly obtained. □

For the remaining results of this section, consider $\Lambda=(a,b)$ with $0\leqslant a<b$ and define $\delta=2$ if $b<\infty$ or $\delta=1$ if $b=\infty$.

Proposition 3.4: Suppose that $k\geqslant1$ (even or odd), $\tau>0$ and $\min\{l_\infty,l_\tau\}\geqslant2$. For each $\alpha<\delta/k$ with $\alpha\leqslant1$ and $\beta\equiv\delta-\alpha k>0$, $\exists C>0$ so that (20) holds for all $\boldsymbol{\phi}_0\in\boldsymbol{\mathcal{H}}$.

Proof: We change the previous proof by defining, for $A>\tau$ and $B\geqslant A$,

$$\mathcal{G}(q)=\mathcal{G}_A(q)+\mathcal{G}_\infty(q),$$

$$\mathcal{G}_A(q)=G(q_\tau/A_\tau)(1-G(q/c_\tau)),\quad \mathcal{G}_\infty(q)=(1-G(q/B))G(q/c_\tau), \tag{27}$$

$$q_\tau\equiv q-\tau,\quad A_\tau\equiv A-\tau,\quad c_\tau\equiv1+\tau.$$

Using the $\mathcal{G}(q)$ partition on the left-hand side of (20) gives, for $t>0$ and $h\equiv h(|\nabla_x|)$,

$$\int_{|x|\leqslant t^\alpha}|U_t\mathcal{P}_\Lambda\boldsymbol{\phi}_0|^2d^kx\leqslant2\|(1-G((h-\tau)/A_\tau))\mathcal{P}_\Lambda\boldsymbol{\phi}_0\|^2+2\|G(h/B)\mathcal{P}_\Lambda\boldsymbol{\phi}_0\|^2$$
$$+Ct^{\alpha k}\sup_{|x|\leqslant t^\alpha}|U_t\mathcal{G}_A(h)\mathcal{P}_\Lambda\boldsymbol{\phi}_0|^2+C't^{\alpha k}\sup_{|x|\leqslant t^\alpha}|U_t\mathcal{G}_\infty(h)\mathcal{P}_\Lambda\boldsymbol{\phi}_0|^2. \tag{28}$$

The $\mathcal{G}_\infty(h)$ and $G(h/B)$ terms can be handled as in the previous Proposition. However, since $\tau>0$, we cannot use (22). Define the time-dependent phase function and operator,

$$\varphi_t(q) \equiv (s-t)q + x\cdot\mu h^{-1}(q) = (s-t)h(\xi) + x\cdot\mu\xi, \quad L_t \equiv \frac{i}{\varphi_t'(q)}\partial_q .$$

For the $\mathcal{G}_A$ term in (28) use integration by parts only once, i.e., $M=1$ in (24). Then,

$$\chi_\Lambda(s)\int_A^{2c_\tau} |L_t^* \mathcal{G}_A(q) H_k(q)| dq \leq \frac{C_{k,\tau}\chi_\Lambda(s)}{c_m|s-t|-|x|} + \frac{C'_{k,\tau}\chi_\Lambda(s)\|(\xi^2-\xi_\tau^2)h''\|_\infty}{(c_m|s-t|-|x|)^2}. \tag{29}$$

Thus, integrating in $s\in\mathbf{R}$, gives the following estimate, analogous to (25):

$$\sup_{|x|\leq t^\alpha} |U_t \mathcal{G}_A(|\nabla_x|)\mathcal{P}_{+\Lambda}\boldsymbol{\phi}_0|^2 \leq C_{k,\tau,\Lambda}\frac{1}{(c_m t - t^\alpha)^\delta}\|\boldsymbol{\phi}_0\|^2.$$

Returning to (28) it is clear that for $\alpha k<\delta$ one obtains an estimate independent of A, except for the first term on the right-hand side. This term vanishes as $A\to\tau^+$. The same computations hold if we replace t^α with $|t+\rho|^\alpha$ and Λ with $\rho\Lambda$. The result is complete. □

Proposition 3.5: Suppose that $k\geq 1$, $\tau=0$ and $\min\{l_\infty, l_\tau\}\geq 2+k/2$. Then, for any $\alpha\leq 1$ with $\beta>0$, where $\beta(\alpha)\equiv(1-\alpha)k+(\delta-1)-(1+(-1)^k)/2$, $\exists C>0$ so that (20) holds.

Proof: The difficulty in these cases is that even when $l_\tau=\infty$ integration by parts can be used only $M\equiv\lceil k/2\rceil$ (least integer greater than $k/2$) times, due to the singularity of $q^{(k-1)/2}$. To prove this, return to (29) with $\tau=0$ and $q=h(\xi)=c_m\xi$. Then given $N\leq\lceil k/2\rceil$, $\exists C_N$ so that,

$$\left|\int_A^{2c_\tau}(L^*)^N\mathcal{G}_A(q)q^{(k-1)/2}dq\right|^2 \leq \frac{C_N}{(c_m|s-t|-|x|)^{2N}}.$$

Now we may proceed as in the proof of Proposition 3.4. The case of nonlinear $h(\xi)$ is handled similarly. We omit the details. □

Remark: The difference between (20) and (21) is unimportant for our study of the scattering matrix since the energy norms are equivalent if $\tau>0$. It will be convenient in the next section to use the estimates in (20).

IV. SCATTERING THEORY FOR TRAPPED WAVES

A. Plane stratified media

We now apply the results of Sec. III to the study of time-asymptotic solutions for the plane-stratified wave equation. Recall that $z=x\oplus y\in\mathbf{R}^k\times\mathbf{R}^1$.

A typical trapped mode for system (4) has initial conditions of the form,

$$\phi_*(x,y) = \Omega_k\int_{\mathbf{R}^k} e^{-ix\cdot\xi} f(\xi)\hat{\psi}_\xi^j(y)d^k\xi, \tag{30}$$

for appropriate $f\in\mathcal{L}^2(\mathbf{R}^k)$ and with $\hat{\psi}_\xi^j$ defined in (6). Using the $\mathcal{F}^h$-transform on (6) gives the new eigenvalue equation,

$$\begin{gathered} H_y^q\tilde{\psi}_j(q,y) = q^2\tilde{\psi}_j(q,y), \quad \tilde{\psi}_j(q,y)\equiv\mathcal{F}^h\mathcal{F}_x^{-1}\hat{\psi}_\xi^j, \\ H_y^q \equiv \mathcal{F}^h K(\mathcal{F}^h)^{-1} = \mathcal{F}^h(\mathcal{F}_x)^{-1}K_y^\xi\mathcal{F}_x(\mathcal{F}^h)^{-1}. \end{gathered} \tag{31}$$

In this case the initial conditions will take the form,

$$\phi_*(x,y) = \Omega_k\int_{\sigma_\tau\otimes S^{k-1}} e^{-ix\cdot\mu h^{-1}(q)} g(q,\mu)\tilde{\psi}_j(q,y)H_k(q)dq\, d\mu, \tag{32}$$

for appropriate $g\in\mathcal{L}^2(\sigma_\tau\times\mathcal{S}^{k-1})$. Now, for any Borel set $\Sigma\subseteq[\tau_j^2,\infty)$, we define $P^j[\Sigma]$ to be the projection onto the ω_j eigenspace of Σ. Explicitly, if $E^K_{\{\lambda\}}$ is the spectral projection of K onto the eigenvalue λ, we have,

$$P^j[\Sigma]\phi(x,y)\equiv\int_{|\xi|^2\in\omega_j^{-1}(\Sigma)}\left(E^{K_y^\xi}_{\{\omega_j(\xi^2)\}}\phi\right)(x,y)d^k\xi$$

$$=\Omega_k\int_{|\xi|^2\in\omega_j^{-1}(\Sigma)}e^{-ix\cdot\xi}\hat{\psi}^j_\xi(y)\left(\int_{\mathbf{R}^1}\hat{\psi}^j_\xi(\overline{y})\mathcal{F}_x\phi(\xi,\overline{y})d\overline{y}\right)d^k\xi.$$

We use the notations $P^j\equiv P^j[\tau_j,\infty)$ and $\mathbf{P}^t\equiv\oplus_j P^j$. Then the *trapped space* is defined,

$$\boldsymbol{\mathcal{H}}^t\equiv\mathbf{P}^t\mathcal{H}\oplus\mathbf{P}^t\mathcal{H}\equiv\mathbf{P}^t\boldsymbol{\mathcal{H}},\tag{33}$$

where $\boldsymbol{\mathcal{H}}\equiv\mathcal{H}\oplus\mathcal{H}$ and $\mathcal{H}\equiv\mathcal{L}^2(\mathbf{R}^{k+1},d^kxdy)$. The *space of free states* is,

$$\boldsymbol{\mathcal{H}}^f\equiv\mathbf{P}^f\mathcal{H}\oplus\mathbf{P}^f\mathcal{H}\equiv\mathbf{P}^f\boldsymbol{\mathcal{H}},\tag{34}$$

where $\mathbf{P}^f=(1-\mathbf{P}^t)$. The *full Hilbert space of initial conditions* for the *stratified wave equation* is thus $\boldsymbol{\mathcal{H}}=\boldsymbol{\mathcal{H}}^f\oplus\boldsymbol{\mathcal{H}}^t$, which is the standard decomposition.[2,4]

The theory of Sec. III is now used to further decompose $\boldsymbol{\mathcal{H}}^t$ by introducing,

$$\mathcal{P}^j_\Lambda\boldsymbol{\phi}(x,y)\equiv\Omega_k\int_{\sigma_\tau\otimes S^{k-1}}e^{-ix\cdot\mu h_j^{-1}(q)}\tilde{\psi}_j(q,y)H_k(q)\hat{v}_q\chi_\Lambda(-i\partial_q)$$

$$\times\left(\int_{\mathbf{R}^1}\tilde{\psi}_j(q,\overline{y})\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}(q,\overline{y})d\overline{y}\right)dq\,d\mu.\tag{35}$$

Lemma 4.1: For each $j\in\mathbf{N}$, $\mathcal{P}^j_\pm$ are complementary-orthogonal projections on $\boldsymbol{\mathcal{H}}_j$.

The proof is similar to that of Lemma 3.1 so the details are not repeated.

Now to proceed, detailed properties of $\tilde{\psi}(q,y)$ are needed. Thus we consider the special case of a stratified operator with single layer, which is an example of (C1)(b) in Sec. II;

(C2) Assume that $V(y)=0$. Furthermore, suppose $\exists c_*,R_*$ with $0<c_m<c_\infty$ so that $c(y)=c_m$ for $R_1<y<R_2$ and $c(y)=c_\infty$ otherwise.

If $R_1=-R_2=-R$ for $R>0$ and j is even, the normalized eigenstates of (31) are,

$$\tilde{\psi}_j(q,y)=N_j(q)\begin{cases}(1/c_\infty)\cos(g_jR)\exp[-f_j(|y|-R)], & |y|>R\\ (1/c_m)\cos(g_jy), & |y|<R\end{cases},\tag{36}$$

where the q-dependent quantities are,

$$f_j\equiv\sqrt{\omega_j^{-1}(q^2)-q^2/c_\infty^2},\quad g_j\equiv\sqrt{q^2/c_m^2-\omega_j^{-1}(q^2)},$$
$$N_j(q)\equiv\left[(f_jc_\infty^2)^{-1}\cos^2(g_jR)+2(g_jc_m^2)^{-1}\int_0^{g_jR}\cos^2u\,du\right]^{-1/2},\tag{37}$$

and the thresholds are $\tau_j=j\pi/R\sqrt{c_m^{-2}-c_\infty^{-2}}$. The dispersion relations satisfy,

$$c_mf_j=c_\infty g_j\tan(g_jR),\quad f_j^2+g_j^2=q^2(c_m^{-2}-c_\infty^{-2}),$$

which ensure the existence of analytic neighborhoods for $\omega_j^{-1}(q^2)$ around τ_j and ∞ so that,

$$f_j=\sum_{k=1}^\infty a_k(q^2-\tau_j^2)^k,\quad g_j=(j+1)\pi/2R+\sum_{k=1}^\infty b_kq^{-k},$$

respectively. Similar expressions hold for the odd j modes. Using these explicit expressions we now obtain propagation estimates for outgoing trapped states with $\Lambda=(a,b)$, where $0<a<b$ and with $\delta=2$ for $b<\infty$, or with $\delta=1$ for $b=\infty$ (as defined in Sec. III). The corresponding result holds for incoming states.

Proposition 4.2: Let (C2) hold and suppose that $\tau_j>0$. Then for each $\alpha<\delta/k$ with $\alpha<1$ and $\beta\equiv\delta-\alpha k>0$, $\exists C$ so that $\forall\,\boldsymbol{\psi}_0\in\boldsymbol{\mathcal{H}}$,

$$\|\chi_+(|t+\rho|^\alpha-|z|)U_t\mathcal{P}^j_{+\rho\Lambda}\boldsymbol{\psi}_0\|^2\leqslant C(1+|t+\rho|)^{-\beta}\|\boldsymbol{\psi}_0\|^2,\quad t>0. \tag{38}$$

If $\tau_j=0$ then for each $\alpha<1$, $\beta\equiv\delta-\alpha$, $\exists C>0$ so that (38) holds.

Proof: We sketch the essential details. Set $\rho=1$. Then, proceeding as in (27) and (28), we can bound the left-hand side of (38) with,

$$\begin{aligned}\int_{y=-t^\alpha}^{t^\alpha}\int_{|x|\leqslant|t|^\alpha}|U_t\mathcal{P}^j_{+\Lambda}\boldsymbol{\psi}_0|^2d^kx\,dy&\leqslant 2\int_{\mathbf{R}^{k+1}}|(1-\mathcal{G}(h(|\nabla_x|)))\mathcal{P}^j_{+\Lambda}\boldsymbol{\psi}_0|^2d^kx\,dy\\&\quad+2\int_{y=-t^\alpha}^{t^\alpha}\int_{|x|\leqslant|t|^\alpha}|U_t\mathcal{G}(h(|\nabla_x|))\mathcal{P}^j_{+\Lambda}\boldsymbol{\psi}_0|^2d^kx\,dy.\end{aligned} \tag{39}$$

The first term on the right-hand side vanishes as $A\to\tau_j^+$, $B\to\infty$. For the second term we decompose the y-integration into oscillatory and exponential parts as defined in (36). In the oscillatory region $|y|<R$ integration by parts gives,

$$\begin{aligned}\int_{-R}^{R}\sup_{\mu}\sup_{|x|\leqslant t^\alpha}\int_{-b}^{-a}\Bigg\{&\left|\int_A^{2c_\tau}e^{i\psi_t}(\psi_t''(\psi_t')^{-2}+(\psi_t')^{-1}\partial_q)\mathcal{G}_AH_kN(q)dq\right|^2\\&+\left|\int_{c_\tau}^{2B}e^{i\psi_t}(\psi_t''(\psi_t')^{-2}+(\psi_t')^{-1}\partial_q)^{\lceil k/2\rceil}\mathcal{G}_BH_kN(q)dq\right|^2\Bigg\}ds\,dy,\end{aligned} \tag{40}$$

where the phase function is defined, suppressing the j index,

$$\varphi_t\equiv tq-sq-x\cdot\mu h^{-1}(q)\pm g(q)y,\quad \varphi_t'=t-s-x\cdot\mu(h^{-1})'\pm g'y. \tag{41}$$

By analyticity of f, g, and N, the integral in (40) is bounded by $Ct^{-\delta}$.

In the exponential region $R<|y|<t^\alpha$ we modify the previous expression by integrating $|y|$ from $R\to t^\alpha$ and replacing $N(q)$ with $N(q)\exp(-f(|y|-R))$. In the phase function (41) set $y=R$. We may now proceed as above except for terms of the form,

$$\int_R^{t^\alpha}\left|\int_{c_\tau}^{2B}e^{i\varphi_t}\mathcal{G}^{(k_1)}H_k^{(k_2)}N^{(k_3)}\partial_q^{k_4}e^{-f(y-R)}dq\right|^2dy\leqslant\sum_{l=1}^{k_4}\left(\int_0^{f(t^\alpha-R)}u^{2l}e^{-2u}du\right)C_l,$$

which are clearly bounded. Here $\Sigma k_i=\lceil k/2\rceil$. This leads to a $t^{-\delta}$ bound. When $\tau_j=0$ then the H_k function improves the bounds for $k\geqslant 2$ near A. We omit the details. □

B. Perturbed stratified media

As an application to scattering theory, consider the perturbed system (1), where the dependent variables satisfy, in addition to (2),

(C3) $V_p, c_p\in\mathcal{C}^4$ and $\exists\,\gamma_p>1/2$ so that for $\langle z\rangle\equiv(1+|z|^2)^{1/2}$,

$$\|\langle z\rangle^{2\gamma_p}(V_p(z)-V(y))\|_\infty+\|\langle z\rangle^{2\gamma_p}(c_p(z)-c(y))\|_\infty<\infty.$$

Let $F \equiv F_{(u,v)} \in C^{\infty}(\mathbf{R}^1)$ be a uniformly bounded positive function with support on (u,v), $0 < u < v < \infty$. The following is a consequence of the work in Ref. 2. See also Ref. 15 for further discussion of operator-smoothness and Ref. 3 for similar energy estimates.

Lemma 4.3: Suppose the systems in (4) and (1) satisfy (C1) and (C3), respectively. Then the following energy-decay estimate holds:

$$\int_{-\infty}^{\infty} \|\langle z\rangle^{-\gamma\#} U_t^{\#} F(K^{\#}) \boldsymbol{\psi}_0\|_{\mathcal{H}}^2 dt \leq C \|\boldsymbol{\psi}_0\|_{\mathcal{H}}^2, \tag{42}$$

for $\gamma_{\#} > 1/2$ and $\gamma_{\#} = \gamma_0$ or γ_p, $U_t^{\#} = U_t$ or U_t^p and $K^{\#} = K$ or K^p, respectively.

The scattering operator S in (8) is defined in terms of the wave operators $\Omega_{\pm}$ which, by Cook's method, have the alternative representations,

$$\Omega_{\pm} = 1 + i \int_0^{\pm\infty} U_t^p \mathcal{V} U_{-t} dt. \tag{43}$$

Here $\mathcal{V}$ is a matrix operator with localness properties (shown in the Appendix),

Lemma 4.4: Under conditions (C1) and (C3), with $F \equiv F_{(u,v)}$, the operators,

$$\mathcal{V} F(K) \langle z\rangle^{\mu}, \quad \langle z\rangle^{\mu} F(K^p) \mathcal{V},$$

$$\langle z\rangle^{2\gamma} F(K^p) \mathcal{V} F(K) \langle z\rangle^{2(\gamma_p - \gamma)}, \quad \langle z\rangle^{2\gamma} (F(K^p) - F(K)) \langle z\rangle^{2(\gamma - \gamma_p)},$$

are bounded for each $0 < u < v < \infty$ and $\mu \leq \gamma_p$ and $0 \leq \gamma \leq \gamma_p$. If $c_p(z) = c(y)$, i.e., $K^p - K = V_p - V$, then the same holds for $v = \infty$.

Using the $\mathcal{H}$-decomposition into free and trapped states, we obtain, with obvious interpretation,

$$S = \mathbf{P}^f S \mathbf{P}^f + \mathbf{P}^f S \mathbf{P}^t + \mathbf{P}^t S \mathbf{P}^f + \mathbf{P}^t S \mathbf{P}^t.$$

The paper is completed with the following:

Main Theorem 4.5: The scattering operator S exists, given conditions (C1) and (C3). Furthermore, for nondegenerate trapped modes with positive thresholds $\tau_j, \tau_k > 0$ and any finite $u > \max\{\tau_j, \tau_k\}$, $v > u$, $\exists\, \nu, C_\nu > 0$ so that,

$$\|P_{+\rho}^j S F_{(u,v)}(K) P_{-\rho}^k\| \leq C_\nu / \rho^\nu. \tag{44}$$

If (C2) and $c_p(z) = c(y)$ hold, i.e., $K^p - K = V_p$, then the uniform causality estimate in (10) holds. Furthermore, if condition (C3) holds for all $\gamma_p > 1/2$, then for each $\nu < 1/2$ a finite constant C_ν exists so that (10) holds.

Proof: The existence of S follows directly from completeness of the wave operators $\Omega_{\pm}$. The latter holds by observing that for any $\boldsymbol{\phi} \in F(K^p)\mathcal{H}$, $\boldsymbol{\psi} \in F(K)\mathcal{H}$, with $F \equiv F_{(u,v)}$,

$$\left| \left\langle \boldsymbol{\phi}, \int_0^{\infty} U_t^p \mathcal{V} U_{-t} dt\, \boldsymbol{\psi} \right\rangle \right| \leq C_F \|\boldsymbol{\phi}\| \|\boldsymbol{\psi}\|,$$

using Lemmas 4.3 and 4.4. Thus $\Omega_{\pm}$ in (43) are densely defined on $\mathcal{H}$. By a similar argument the adjoints are densely defined. Since the wave operators are all bounded linear operators, a standard $\epsilon/3$ argument verifies completeness for $u = 0$, $v = \infty$.

Inequality (44) requires propagation estimates like those in (38). Due to the momentum cutoff function $F_{(u,v)}$ we can proceed as in Proposition 4.2 and use local analyticity of eigenstates[17] to show that for each $M > 0$ and $c_m' < c_m$, $\exists C_M$ so that,

$$\|\chi_+(c_m'|t+\rho| - |z|) U_t F(K) \mathcal{P}_{+\rho}^j \boldsymbol{\psi}_0\|^2 \leq C_M (1 + |t+\rho|)^{-M} \|\boldsymbol{\psi}_0\|^2, \quad t > 0. \tag{45}$$

Choose any $\phi,\psi\in\mathcal{H}$ and let $G^3=F(K)$. Also define $G_0\equiv G(K)$ and $G_p\equiv G(K^p)$. Then (44) is obtained by considering,

$$\langle\boldsymbol{\phi},\mathcal{P}^j_{+\rho}G_0\Omega^*_-G_p\Omega_+G_0\mathcal{P}^k_{-\rho}\boldsymbol{\psi}\rangle=\langle\boldsymbol{\phi},\mathcal{P}^j_{+\rho}G_0G_pG_0\mathcal{P}^k_{-\rho}\boldsymbol{\psi}\rangle+C_1\|\boldsymbol{\phi}\|\int_0^\infty\|\langle z\rangle^{-\gamma_p}U_{-t}G_0\mathcal{P}^k_{-\rho}\boldsymbol{\psi}\|dt$$

$$+C_2\int_0^\infty\|\langle z\rangle^{-\gamma_p}U^p_tG_0\mathcal{P}^j_{+\rho}\boldsymbol{\phi}\|dt\|\boldsymbol{\psi}\|.$$

The last two terms are bounded by inverse powers of ρ due to (45). For the first term on the right-hand side, write $G_p=(G_p-G_0)+G_0$, and note that Lemma 4.4 and Eq. (45) combine to handle the difference of operators. If $j\neq k$ then the G_0 term vanishes. Else, we need to establish the following causality estimate (in the Appendix),

Lemma 4.6: Let $\mathcal{P}^j_*$ be the projections defined in (35) on $\mathcal{H}$. Then for any $F\in\mathcal{C}_0^\infty(\mathbf{R}^+)$ and $\forall M>0,\exists C_*>0$, so that,

$$\|\mathcal{P}^j_\pm F(K)\mathcal{P}^j_{\mp\rho}\|\leq C_{M,F}/\rho^M.$$

Finally we show (10), assuming that (C3) holds for all $\gamma_p>1/(2\alpha)$. Remaining results are obtained similarly. For $\tau_j\leq\tau_k$ consider the decomposition of $P^j_{+\rho\Lambda}SP^k_{-\rho\Lambda}$,

$$P^j_{+\rho\Lambda}\chi_k\Omega^*_+[(1-P^k_-)\Omega_-P^k_{-\rho\Lambda}]+[P^j_{+\rho\Lambda}\chi_k\Omega^*_+P^k_-]\Omega_-P^k_{-\rho\Lambda}, \tag{46}$$

for $\chi_k\equiv\chi_{[\tau_k,\infty)}(K)$. Write the $[\cdots]$ factor in the first term as,

$$(1-P^k_-)\Omega_-P^k_{-\rho\Lambda}=(1-P^k_-)P^k_{-\rho\Lambda}+i\int_0^{-\infty}(1-P^k_-)U^p_t\mathcal{V}U_{-t}P^k_{-\rho\Lambda}\,dt.$$

The first term on the right-hand side is clearly 0. As for the second term, we use Lemmas 4.3 and 4.4 and consider, for any $\boldsymbol{\psi}\in\mathcal{H}$,

$$\int_0^{-\infty}\|\langle z\rangle^{-\gamma_p}U_{-t}P^k_{-\rho\Lambda}\boldsymbol{\psi}\|^2dt\leq\|\boldsymbol{\psi}\|^2\int_0^{-\infty}\|\langle z\rangle^{-\gamma_p}\chi_+(|z|-|t-\rho|^\alpha)\|^2_\infty dt$$

$$+\int_0^{-\infty}\|\chi_+(|t-\rho|^\alpha-|z|)U_{-t}P^k_{-\rho\Lambda}\boldsymbol{\psi}\|^2dt$$

$$\leq\|\boldsymbol{\psi}\|^2(C_1/\rho^{(2\alpha\gamma_p)-1}+C_2/\rho^{\delta-\alpha k-1}).$$

Since $\Lambda\Subset\mathbf{R}^+$ is compact, $\delta=2$. Hence for α arbitrarily small and γ_p large we obtain a nearly $\rho^{-1/2}$ bound for the norm of the first term in (46).

Finally, if $\tau_j<\tau_k$ then the term $P^j_{+\rho\Lambda}\chi_k\Omega^*_+P^k_-$ in (46) has a stronger bound than the one above, even for $\Lambda=(a,\infty)$, $a>0$. This completes the Theorem. □

V. CONCLUSION

In this paper we introduced incoming/outgoing projection operators and obtained uniform decay estimates for solutions of the wave equation with initial conditions in the corresponding subspaces. These estimates allowed us to study the scattering operator for the perturbed-stratified wave equation. The concept of strict causality for the system is replaced by a causality estimate.

ACKNOWLEDGMENT

This is dedicated to the memory of Sanjaya Baratham.

APPENDIX

Proof of Lemma 3.1: Due to the details involved, we only consider the special case $h(\xi)=\xi$ and $\gamma=1$. However we allow $\tau\geqslant 0$ in order to demonstrate how one can *bridge the energy gap*. Let $\phi_*,\psi_*\in\mathcal{H}_\tau$ be arbitrary.

Idempotentcy: We show $\mathcal{P}_+^2=\mathcal{P}_+$ using definition (3.2). That $\mathcal{P}_-^2=\mathcal{P}_-$ follows similarly. Compute,

$$\begin{aligned}\mathcal{P}_+^2\boldsymbol{\phi}&=\mathcal{P}_+(\mathcal{F}^h)^{-1}\hat{v}_q(\chi_+(-i\partial_q))\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}\\&=(\mathcal{F}^h)^{-1}\hat{v}_r\chi_+(1/2)(\mathcal{F}^h(\mathcal{F}^h)^{-1}-i\hat{r}\mathcal{F}^h(\mathcal{F}^h)^{-1}i\hat{q})\chi_+\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}\\&=(\mathcal{F}^h)^{-1}\hat{v}_r\chi_+^2\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}=\mathcal{P}_+\boldsymbol{\phi},\end{aligned}$$

where we use the fact that $\mathcal{F}^h(\mathcal{F}^h)^{-1}=\delta(r-q)$ and $\chi_+^2=\chi_+$.

Self-adjointness: Since $\mathcal{P}_\pm$ are bounded operators, we need only show the symmetric property. Using Fubini's theorem and the δ-function properties we calculate,

$$\langle\boldsymbol{\psi},\mathcal{P}_+\boldsymbol{\phi}\rangle=\Omega_k\int_{\sigma_\tau\otimes S^{k-1}}(\mathcal{F}^h\boldsymbol{\psi})^*\hat{v}_q(\gamma\chi_+(-i\partial_q)\gamma)\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}\,dq\,d\mu.$$

Then, since $-i\partial_q$ is a self-adjoint operator on $\mathcal{L}^2(\sigma_\tau\otimes\mathcal{S}^{k-1},dq\,d\mu)$, so is $\chi_+(-i\partial_q)$.

To complete the Lemma note that $\chi_+(-s)=\chi_-(s)$, $\chi_+\chi_-=0$, $\chi_++\chi_-=1$, etc. □

Proof of Claim 3.2: From (3.3) it is clear that for each $R>0$, as $t\to\infty$, $\|\chi_+(R-|x|)U_t\mathcal{P}_+\|\to 0$ and this, along with its corresponding adjoint expression, verifies (i)′(a). For (i)′ (b), choose any $\boldsymbol{\phi}\in\boldsymbol{\mathcal{H}}$ and observe that using (3.5) gives,

$$\begin{aligned}(1-\mathcal{P}_{+t})U_t\mathcal{P}_+\boldsymbol{\phi}&=\int dq\,d\mu\, e^{-ix\cdot\mu h^{-1}(q)}H_k(q)\hat{v}_q(1-\chi_{+t})e^{iqt}\chi_+(\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi})\\&=\int dq\,d\mu\, e^{i((q+\tau\hat{q})t-x\cdot\mu h^{-1}(q))}H_k(q)\hat{v}_q[\chi_-e^{i\tau\hat{q}t}\chi_+](\hat{v}_q^T\mathcal{F}^h\boldsymbol{\phi}).\end{aligned}$$

The factor in $[\cdots]$ vanishes identically if $\tau t=n\pi$ for $t\geqslant 0$, $n\in\mathbf{N}$.

Property (ii) follows immediately from (i)′(b). For (iii) we approximate any $\boldsymbol{\phi}\in\boldsymbol{\mathcal{H}}$ by a state in $\cup U_t\mathcal{D}_+$. In particular, consider,

$$\|\boldsymbol{\phi}-U_t\mathcal{P}_+U_{-t}\boldsymbol{\phi}\|=\|(1-\mathcal{P}_+)U_{-t}\boldsymbol{\phi}\|=\|\mathcal{P}_-U_{-t}\boldsymbol{\phi}\|.$$

The right-hand side vanishes as $t\to-\infty$ by (i)′(a). □

Proof of Lemma 4.4: The perturbed matrix operator is,

$$\mathcal{V}\equiv\begin{pmatrix}0&\sqrt{K}-\sqrt{K^p}\\\sqrt{K^p}-\sqrt{K}&0\end{pmatrix}.$$

To study this quantity, we use Kato's square root formula,

$$\sqrt{K}-\sqrt{K^p}=\int_0^\infty\sqrt{\omega}\frac{1}{K^p+\omega}(K-K^p)\frac{1}{K+\omega}\,d\omega.\tag{A1}$$

Similarly we can reduce the study of $F(K)-F(K^p)$ to that of $K-K^p$ (see the discussion in Sec. VI of Ref. 2). Here we only consider the integrand on the right-hand side of (A1). A simple computation gives,

$$K-K^p=(1-c_p/c)K-K^p(1-c/c_p)-Vc_p/c+V_pc/c_p.$$

Now, for $L=K$ or K^p denote $F\equiv F_{(u,v)}(L)$ where $0\leqslant F\leqslant 1$. In the case $c_p\neq c$ we have the obvious bounds,

$$\left\|\frac{K}{K+\omega}F(K)\right\|_\infty=\left\|\frac{K^p}{K^p+\omega}F(K^p)\right\|_\infty\leqslant\frac{v}{u+\omega}.$$

If $c_p=c$ (but $V_p\neq V$) then these bounds will not be required, hence $v=\infty$ is allowed.

To verify the first two bounds in Lemma 4.4, we seek to prove,

$$\left\|\langle z\rangle^{-\mu}F(L)\frac{L^\delta}{L+\omega}\langle z\rangle^\mu\right\|\leqslant\frac{C_\mu}{\omega+1},\quad \delta=0,1,$$

where $\delta=1$ requires $v<\infty$. Here $\mu\geqslant 0$ but we consider only the $\mu=1$ case. The result is clear for $\mu=0$ and the following techniques apply for any integer μ. Interpolation obtains this bound for all $\mu\in\mathbf{R}$. The remaining bounds in Lemma 4.4 are obtained similarly.

Define the smooth cutoff functions $\{G_i\}$ so that $F=G_0=G_1^2=G_2^4=$etc. Then using a partition of $\mathbf{R}^n$, homogeneous of degree 0 in $|z|$ near ∞, it is sufficient to consider, for any $z_j\in\mathbf{R}$, $1\leqslant j\leqslant k+1$, $n\in\mathbf{N}$ with $G_*=G_*(L)$ and $i=\sqrt{-1}$,

$$\begin{aligned}G_{n-1}(L+\omega)^{-1}(z_j+i)-(z_j+i)G_{n-1}(L+\omega)^{-1}\\=G_n(L+\omega)^{-1}[G_n,z_j]-G_n(L+\omega)^{-1}[L,z_j](L+\omega)^{-1}G_n\\+[G_n,z_j](L+\omega)^{-1}G_n.\end{aligned}\tag{A2}$$

For the middle term on the right-hand side of (A2), we evaluate for $1\leqslant l\leqslant k$,

$$[K,x_l]=-2c^2(y)\partial_{x_l},\quad [K^p,x_l]=-2c^p(z)\partial_{x_l}c^p(z),$$

$$[K,y]=-2c\partial_y c,\quad [K^p,y]=-2c^p\partial_y c^p.$$

Thus for any $\phi\in\mathcal{H}$, using the Schwartz inequality,

$$\|[L,z_j](L+\omega)^{-1}G_i\phi\|^2\leqslant 4(c_p^M)^2\|(L+\omega)^{-1}G_i\|\|\phi\|^2\leqslant C(\omega+1)^{-1}\|\phi\|^2.$$

We are left to consider the remaining terms in (A2). Define $R\equiv(L+1)^{-1}$ and $J_n=(L+1)^2G_n$ if $v<\infty$ or $J_n=(L+1)^2(1-G_n)$ if $v=\infty$. Then,

$$[G_n,z_j]=RJ_n[R,z_j]+R[J_n,z_j]R+[R,z_j]J_nR.$$

To bound the middle term on the right-hand side it is standard to use the Fourier transform $\hat{J}_n$ of J_n and write,

$$R[J_n,z_j]R=\int_0^\infty\hat{J}_n(s)\int_0^s e^{iL(s-r)}R[L,z_j]Re^{-iLr}dr\,ds.$$

This is clearly a bounded operator. Iteration of these techniques verifies the result. □

Proof of Lemma 4.6: For brevity we only consider the $h(q)=c_mq,\tau=0$ case. Compute,

$$\begin{aligned}\|\mathcal{P}_+^j F(K)\mathcal{P}_{-\rho}^j \boldsymbol{\phi}\|^2 &= \int_{\mathbf{R}^k} (\mathcal{P}_+^j F \mathcal{P}_{-\rho}^j \boldsymbol{\phi})^* (\mathcal{P}_+^j F \mathcal{P}_{-\rho}^j \boldsymbol{\phi}) d^k x \\ &= \int_{\mathbf{R}^{k+1}} \left[\int_{\mathbf{R}\times\mathcal{S}^{k-1}} e^{i(x\cdot\omega/c_m)r} \tilde{\psi}_j(r,y) |r|^{(k-1)/2} \hat{v}_r^* \right. \\ &\quad \times (\gamma\chi_-(-i\partial_r) F(r^2)\chi_\rho(-i\partial_r)\gamma^*) \\ &\quad \times \left. \left(\int_{\mathbf{R}^1} \tilde{\psi}_j(r,\overline{y})(\hat{v}_r^T)^* \overline{(\mathcal{F}^h \boldsymbol{\phi})^* dy} \right) dr\, d\omega \right] (\text{c.c.}) d^k x\, dy \\ &= \int_{\mathbf{R}\times\mathcal{S}^{k-1}} |r|^{k-1} [\hat{v}_r^* (\gamma\chi_- F \chi_\rho \gamma^*)(\hat{v}_r^T)^* (\mathcal{F}^h \boldsymbol{\phi})^*] (\text{c.c.}) dr\, d\omega \\ &\leqslant 2\|\chi_- F \chi_\rho\|_{op}^2 \|\boldsymbol{\phi}\|^2,\end{aligned}$$

where c.c. means complex conjugate of the previously bracketed factor. Now, since $F \in C_0^\infty(\mathbf{R}^+)$, we can write $F(r^2) = \int_{-\infty}^{\infty} e^{-irs}\hat{G}(s)ds$ for some $\hat{G} \in \mathcal{S}(\mathbf{R})$. Then, using (3.5),

$$\begin{aligned}\|\chi_- F \chi_\rho \psi\|^2 &= \int_{\mathbf{R}\times\mathcal{S}^{k-1}} \left| \int_{-\infty}^{-\rho} e^{-irs}\hat{G}(s)\chi_{(\rho,-s)}(-i\partial_r)\psi(r,\omega)ds \right|^2 |r|^{k-1} dr\, d\omega \\ &\leqslant \int |r|^{k-1} dr d\omega \left[\int_{-\infty}^{-\rho} \langle s\rangle^{M+1} |\hat{G}(s)|^2 ds \int_{-\infty}^{-\rho} \langle t\rangle^{-(M+1)} |\chi_{(\rho,-t)}\psi(r,\omega)|^2 dt \right] \\ &\leqslant C \int_{-\infty}^{-\rho} \langle t\rangle^{-(M+1)} \left(\int |r|^{k-1} dr\, d\omega |\chi_{(\rho,-t)}\psi(r,\omega)|^2 \right) dt \leqslant C' \rho^{-M} \|\psi\|^2.\end{aligned}$$

These details easily extend to the case of nonlinear $h(q)$. □

[1] M. Reed and B. Simon, ''The scattering theory of classical waves from inhomogeneous media,'' Math. Z. **155**, 163–180 (1977).
[2] S. DeBièvre and D. W. Pravica, ''Spectral analysis for optical fibers and stratified fluids. I. The limiting absorption principle,'' J. Funct. Anal. **98**, 404–436 (1991).
[3] D. W. Pravica, ''Trapped rays in cylindrically-symmetric media and poles of the analytically continued resolvent,'' J. Math. Anal. Appl. **209**, 341–364 (1997).
[4] J. C. Guillot and C. H. Wilcox, ''Spectral analysis of the Epstein operator,'' Proc. R. Soc. Edinburgh, Sect. A. Math 85–98 (1978).
[5] J. R. Taylor, ''Time-like cluster properties in nonrelativistic scattering,'' J. Math. Phys. **8**, 2131–2137 (1967).
[6] R. Lavine, ''Commutators and scattering theory. I. Repulsive interactions,'' Commun. Math. Phys. **20**, 301–323 (1971).
[7] E. Mourre, ''Operateurs conjugues et proprietes de propagations,'' Commun. Math. Phys. **91**, 279–300 (1983).
[8] P. Perry, *Scattering Theory by the Enss Method* (Harwood Academic, Great Britain, 1983).
[9] D. Robert, ''Time delay for long-rang perturbations of the Laplace operator,'' in *Operator Calculus and Spectral Theory*, edited by M. Demuth *et al.* Vol. 57, pp. 245–254 (Lambrecht, Germany, 1991).
[10] T. Christiansen, ''Scattering theory for manifolds with asymptotically cylindrical ends,'' J. Funct. Anal. **131**, 499–530 (1995).
[11] P. Lax and R. Phillips, *Scattering Theory* (Academic, New York, 1967).
[12] H. M. Nussenzveig, *Causality and Dispersion Relations* (Academic, New York, 1972).
[13] M. Klaus and B. Simon, ''Coupling constant thresholds in nonrelativistic quantum mechanics,'' Ann. Phys. **130**, 251–281 (1980).
[14] D. Ludwig, ''The Radon transform on Euclidean space,'' Commun. Pure Appl. Math. **19**, 49–81 (1966).
[15] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. III. Scattering Theory* (Academic, New York, 1979).
[16] B. Sz. Nagy, ''Extensions of linear transformations in a Hilbert space which extend beyond this space,'' supplement to F. Riez and B. Sz Nagy, *Functional Analysis* (Frederick Ungar, New York, 1960).
[17] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. IV. Analysis of Operators* (Academic, London, 1979).

Poisson–Lie structures on infinite-dimensional jet groups and quantum groups related to them

Ognyan Stoyanov[a)]
Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903

(Received 22 December 1997; accepted for publication 17 September 1998)

We study the problem of classifying all Poisson–Lie structures on the group G_∞ of formal diffeomorphisms of the real line $\mathbb{R}^1$ which leave the origin fixed, as well as the extended group of diffeomorphisms $G_{0\infty}\supset G_\infty$ whose action on $\mathbb{R}^1$ does not necessarily fix the origin. A complete local classification of all Poisson–Lie structures on the groups G_∞ and $G_{0\infty}$ is given. This includes a classification of all Lie-bialgebra structures on the Lie algebra $\mathcal{G}_\infty$ of G_∞, which we prove to be all of the coboundary type, and a classification of all Lie-bialgebra strucutures on the Lie algebra $\mathcal{G}_{0\infty}$ (the Witt algebra) of $G_{0\infty}$ which also turned out to be all of the coboundary type. A large class of Poisson structures on the space V_λ of λ-densities on the real line is found such that V_λ becomes a homogeneous Poisson space under the action of the Poisson–Lie group G_∞. We construct a series of quantum semigroups whose quasiclassical limits are finite-dimensional Poisson–Lie quotient groups of G_∞ and $G_{0\infty}$.
[S0022-2488(99)00201-7]

I. INTRODUCTION

Quantum groups have been introduced in Refs. 1,2 as deformations of universal enveloping algebras of Lie groups and of the algebra of functions on Poisson–Lie groups.[3] The latter are Lie groups equipped with Poisson structures compatible with the group structure (from where the term Poisson–Lie group originates). In this approach to constructing quantum groups the first step is to analyze the existence of Poisson–Lie structures on the corresponding Lie group. The question of classifying all Poisson–Lie structures on a given Lie group (provided any exist) has been posed originally by Drinfel'd and Belavin.[4–6] In the same paper they give a complete solution for the case of finite-dimensional complex (semi)simple Lie groups. The problem, in general, is very difficult. It has been solved for some other groups in low dimensions. Let us give a list of groups for which the solution of the classification problem is known to us at present: (a) Finite dimensional complex (semi)simple Lie groups;[4] (b) the groups $GL(2,\mathbb{R})$, $SL(2,\mathbb{R})$, $GL(1|1)$;[7,8] (c) the 3-dimensional Heisenberg group,[9] and some unipotent subgroups of $GL(n,\mathbb{R})$ in various dimensions;[10] (d) the group of affine transformations of the line $Aff(1)$; (e) the group of motions of $\mathbb{R}^1\times\mathbb{R}^1$; (f) the Lorentz group considered as a realification of $SL(2,\mathbb{C})$.[11]

Note that all the groups mentioned above are *finite-dimensional*, and the only infinite *series* for which the classification has been completed are *complex* not *real* groups.

With this paper we initiate a program of study of Poisson–Lie structures for the important case of the group of (formal) diffeomorphisms $FDiff(\mathbb{R}^n)$. One of our principal results is the complete solution of the classification problem for the case $n=1$. Namely, in the work presented here we study the problem of local classification (up to a local change of coordinates) of all the Poisson–Lie structures on the group $G_\infty=FDiff(\mathbb{R}^1)$ of formal diffeomorphisms of the real line $\mathbb{R}^1$ which leave the origin fixed, as well as the full group of diffeomorphisms $G_{0\infty}=FDiff_0(\mathbb{R}^1)\supset G_\infty$ (considered as a formal group) whose action on $\mathbb{R}^1$ does not necessarily fix the origin.

[a)]Electronic mail: stoyanov@math.rutgers.edu, Tel: (732) 445-6587, Fax: (732) 445-5530.

The existence of Poisson–Lie structures on G_∞ and $G_{0\infty}$ is far from being obvious. For instance, since G_∞ is a projective limit of groups of finite jets (cf. Sec. III), if we consider the group of 3-jets leaving the origin fixed, then there exists a Poisson–Lie structure on this group,[12] which can not be extended to G_∞. Even though the above groups are infinite-dimensional, surprisingly, the classification problem has a complete solution. There exist countably many isomorphism classes of Poisson–Lie structures on G_∞ and $G_{0\infty}$, labeled by integers $d\in\mathbb{N}\rightarrow\mathbb{Z}_+$ and $d\in\mathbb{N}\rightarrow\mathbb{Z}_+\cup\{-1\}$, respectively. The isomorphism classes under the actions of $G_{0\infty}$ and G_∞ are explicitly described.

Let $[u\mapsto x(u)=x_0+x_1u+x_2u^2+\cdots]\in G_{0\infty}$ be a formal diffeomorphism. The coordinate ring $\mathbb{R}[[x]]$ of $G_{0\infty}$ is countably generated. Let $x=(x_0,x_1,\ldots)$ be a formal point of $G_{0\infty}$, where x_1 is assumed to be invertible. For any $f,g\in\mathbb{R}[[x]]$ a Poisson–Lie bracket, $\{,\}:\mathbb{R}[[x]]\otimes\mathbb{R}[[x]]\rightarrow\mathbb{R}[[x]]$, is defined as

$$\{f,g\}(x):=\omega_{ij}(x)\frac{\partial f}{\partial x_i}\frac{\partial g}{\partial x_j}, \tag{1}$$

where summation over repeated indices is understood. Here $\omega_{ij}=\{x_i,x_j\}\in\mathbb{R}[[x]]$, $i,j\in\mathbb{Z}_+$ are the components of a (formal) tensor that satisfy the infinite system of differential equations (the Jacobi identities):

$$\omega_{ij}\frac{\partial\omega_{kl}}{\partial x_i}+\omega_{ik}\frac{\partial\omega_{lj}}{\partial x_i}+\omega_{il}\frac{\partial\omega_{jk}}{\partial x_i}=0, \tag{2}$$

and the infinite system of functional equations [ω is required to be a (group) 1-cocycle)],

$$\omega_{ij}(z)=\omega_{kl}(x)\frac{\partial z_i}{\partial x_k}\frac{\partial z_j}{\partial x_l}+\omega_{kl}(y)\frac{\partial z_i}{\partial y_k}\frac{\partial z_j}{\partial y_l}, \tag{3}$$

where $z_i=F_i(x,y)$, $i=0,1,2,\ldots$, describe the formal group law[24] induced by the substitution of formal power series $z(u)=x(y(u))$. We introduce $\Omega(u,v;x):=\sum_{i,j=1}^{\infty}\omega_{ij}(x)u^iv^j$, a generating series for the brackets ω_{ij}. The multiplicativity (3) of the Poisson brackets is equivalent to $\Omega(u,v;x)$ satisfying the following functional equation (Lemma IV.1):

$$\Omega(u,v;xy)=\Omega(y(u),y(v);x)+\Omega(u,v;y)x'(y(u))x'(y(v)). \tag{4}$$

Here x' denotes the derivative of x with respect to its argument. The general solution of this equation is (Theorem IV.1)

$$\Omega(u,v;x)=\varphi(u,v)x'(u)x'(v)-\varphi(x(u),x(v)), \tag{5}$$

where $\varphi(u,v)=-\varphi(v,u)\in\mathbb{R}[[u,v]]$ is a formal series in u,v, and satisfies the following functional partial differential equation (Lemma IV.2):

$$\begin{aligned}\varphi(u,v)[\partial_u\varphi(w,u)+\partial_v\varphi(w,v)]+\varphi(v,w)[\partial_v\varphi(u,v)+\partial_w\varphi(u,w)]\\+\varphi(w,u)[\partial_w\varphi(v,w)+\partial_u\varphi(v,u)]=0.\end{aligned} \tag{6}$$

This equation is equivalent to the Jacobi identities (2) and is a functional realization of the Classical Yang–Baxter Equation (Sec. VI). We show that G_∞ acts on the space of solutions of (6). Further, we explicitly describe the moduli space of solutions and show that it is isomorphic to $\mathbb{Z}_+\cup\{-1\}$. By choosing the representative $\tilde{\varphi}(u,v)=uv(v^d-u^d)$, $d\in\mathbb{Z}_+\cup\{-1\}$, of each isomorphism class of solutions we obtain explicit formulas for the Poisson–Lie brackets. For each $d\in\mathbb{N}$ this gives rise to the Poisson–Lie tensor,

$$\omega_{ij}(x)=(i-d)jx_jx_{i-d}-i(j-d)x_ix_{j-d}+x_i\sum_{\Sigma_{k=1}^{d+1}s_k=j}x_{s_1}\ldots x_{s_{d+1}}-x_j\sum_{\Sigma_{k=1}^{d+1}s_k=i}x_{s_1}\ldots x_{s_{d+1}},\tag{7}$$

where $i,j\in\mathbb{Z}_+$. When $d=-1$ we have the tensor

$$\omega_{ij}(x)=i(j+1)x_ix_{j+1}-(i+1)jx_{i+1}x_j-x_i\delta_j^0+x_j\delta_i^0,\quad i,j\in\mathbb{Z}_+.\tag{8}$$

The class of the trivial Poisson–Lie bracket corresponds to $d=0$.

As we can see from the above formulas, the Poisson structures on $G_{0\infty}$ and G_∞ are given by polynomials of arbitrary degree, extending the existing list of known linear (Kirillov–Kostant brackets on the dual of a Lie algebra), quadratic (coming from solutions of CYBE for finite-dimensional groups),[13] and cubic (Toda lattice)[14] Poisson brackets. They give a series of nontrivial examples of infinite-dimensional Poisson manifolds.

The classification problem can be studied also on the level of Lie algebras. The Lie algebras of the groups $G_{0\infty}$ and G_∞ are the Witt algebra,

$$\mathcal{G}_{0\infty}=\mathrm{span}_k\{e_i|[e_i,e_j]=(i-j)e_{i+j},i,j\geqslant -1\},\quad \text{where } k=\mathbb{R}\text{ or }\mathbb{C},$$

and its principal subalgebra $\mathcal{G}_\infty=\mathrm{span}_k\{e_i|[e_i,e_j]=(i-j)e_{i+j},i\geqslant 0\}$. We prove that there is a one-to-one correspondence between the Poisson–Lie structures on G_∞ and the Lie-bialgebra structures on $\mathcal{G}_\infty$ (Sec. VI). The latter are shown to be *all* of the coboundary type (Theorem V.2). *All* Lie-bialgebra structures on $\mathcal{G}_{0\infty}$ are also of the coboundary type (Theorem V.1), they are all classified, and there is a one-to-one correspondence between them and an explicitly listed family of the Poisson–Lie structures on $G_{0\infty}$. Thus, a complete classification of all Lie-bialgebra structures on the Witt algebra $\mathcal{G}_{0\infty}$ and its principal subalgebra $\mathcal{G}_\infty$ is obtained. The result is that every Lie-bialgebra structure on $\mathcal{G}_\infty$ and $\mathcal{G}_{0\infty}$ is induced by a Lie-bialgebra structure on a 2-dimensional subalgebra generated by $\{e_0,e_d\}$ where $d\in\mathbb{N}$ and $d\in\mathbb{N}\cup\{-1\}$, respectively.

The study of Poisson homogeneous spaces under the action of Poisson–Lie groups is a subject of its own right. In this paper we construct a class of Poisson homogeneous spaces for the group G_∞. With a fixed Poisson–Lie structure the group G_∞ acts naturally (by substitution) on the modules $V_\lambda=\mathbb{R}[[u]](du)^\lambda$ of λ-densities on the line. Each Poisson–Lie structure on the group G_∞ induces a family of Poisson structures on V_λ such that V_λ becomes a homogeneous Poisson G_∞-space. Namely, the following theorem holds.

Theorem I.1: *For every $\lambda\in\mathbb{R}$ and $y(u)(du)^\lambda\in V_\lambda$ the following family of Poisson structures on V_λ:*

$$\{y(u),y(v)\}=\phi(u,v)y'(u)y'(v)+\lambda\partial_u\phi(u,v)y(u)y'(v)$$
$$+\lambda\partial_v\phi(u,v)y'(u)y(v)+\lambda^2\frac{\partial^2}{\partial u\,\partial v}\phi(u,v)y(u)y(v),\tag{9}$$

makes the action $G_\infty\times V_\lambda\to V_\lambda$ Poisson. Here $\phi(u,v)=-\phi(v,u)$, and $\phi(u,v)$ satisfies the functional partial differential equation (6).

Finally, we address the quantization problem for the Poisson–Lie brackets. We construct a series of finitely generated noncommutative noncocommutative bialgebras (quantum semigroups) whose quasi-classical limits are finite-dimensional Poisson–Lie quotient groups of G_∞ and $G_{0\infty}$. These are constructed by a formal deformation of the product of the coordinate ring of G_∞. The Poisson–Lie structures on these finite-dimensional groups are restrictions of the Poisson-Lie structures on G_∞ and $G_{0\infty}$, that is, restrictions of the families of brackets (7) to the finite-dimensional quotient groups $G_n=G_\infty \bmod u^{n+1}$ of n-jets. It is easy to show that the restriction is a well defined Poisson map. All such finite-dimensional Poisson–Lie groups for $d\leqslant 5$ and $n\leqslant 7$ have been quantized. Below we give an example of such quantization for the case $d=1$ and $n=4$. The corresponding Poisson brackets on G_4 are

$$\{x_1,x_2\}=x_1^3-x_1^2,$$

$$\{x_1,x_3\}=2x_2(x_1^2-x_1),$$

$$\{x_2,x_3\}=(3x_1-x_1^2)x_3+x_2^2(2x_1-4),$$

$$\{x_1,x_4\}=x_3(2x_1^2-3x_1)+x_2^2x_1,$$

$$\{x_2,x_4\}=x_4(4x_1-x_1^2)+x_3x_2(2x_1-6),$$

$$\{x_3,x_4\}=x_4x_2(8-2x_1)+x_3x_2^2+x_3^2(2x_1-9).$$

Their quantization is described by the following theorem.

Theorem I.2: *Let $X=\{x_1,x_2,x_3,x_4\}$ be a set of four generators and let $\langle X\rangle$ be the associative semigroup with an identity generated by X. Consider an ideal $\mathcal{I}_h$ generated by the following set of relations in $\mathbb{C}[[h]]\langle X\rangle$:*

$$x_1x_2=x_2x_1+h(x_1^3-x_1^2),$$

$$x_1x_3=x_3x_1+h(2x_2x_1^2-2x_2x_1)+h^2(2x_1^4-3x_1^3+x_1^2),$$

$$x_2x_3=x_3x_2+h(3x_3x_1-x_3x_1^2+2x_2^2x_1-4x_2^2)+h^2(3x_2x_1^2-3x_2x_1)+h^3(2-2C_3)(x_1^5-x_1^2),$$

$$\begin{aligned}x_1x_4=&x_4x_1+h(-3x_3x_1+2x_3x_1^2+x_2^2x_1)\\&+h^2x_2(3x_1-8x_1^2+5x_1^3)+h^3[(5x_1^5-12x_1^4+7x_1^3)+C_3(x_1^5-x_1^2)],\end{aligned}$$

$$\begin{aligned}x_2x_4=&x_4x_2+h(4x_4x_1-x_4x_1^2+2x_3x_2x_1-6x_3x_2+x_2^3)+h^2(3x_2^2x_1^2-10x_2^2x_1\\&+12x_2^2+12x_3x_1^2-2x_3x_1^3-15x_3x_1)+h^3x_2[(9+2C_3)x_1-17x_1^2+6x_1^3+(5-5C_3)x_1^4]\\&+h^4[(22-22C_3)x_1^2+(-4+4C_3)x_1^3+(-18+18C_3)x_1^5+C_4(x_1^6-x_1^2)],\end{aligned}$$

$$\begin{aligned}x_3x_4=&x_4x_3+h(8x_4x_2-2x_4x_2x_1+x_3x_2^2-9x_3^2+2x_3^2x_1)\\&+h^2[x_3x_2(-x_1^2+16x_1-24)+x_4(-7x_1^2+16x_1)]\\&+h^3[(10-10C_3)x_2^2x_1^3+(-9+8C_3)x_2^2+(-5+5C_3)x_3x_1^4+16x_3x_1^2-(6+9C_3)x_3x_1]\\&+h^4[(8-9C_3-2C_4)x_2x_1+(-8C_3+9)x_2x_1^2\\&+(10-10C_3)x_2x_1^4+(2C_4-18+18C_3)x_2x_1^5]\\&+h^5[(C_4+2C_3-2)x_1^2+(4-4C_3-C_4)x_1^6+(2C_3-2)x_1^5+C_5(x_1^7-x_1^2)],\end{aligned}$$

where $C_3,C_4,C_5\in\mathbb{C}$ are arbitrary parameters. Then the semigroup quotient algebra $\mathbb{C}[[h]]\langle X\rangle/\mathcal{I}_h$ is a bialgebra (quantum semigroup), denoted $G^1_{4|h}$, with a comultiplication (induced by the group multiplication) defined by

$$\Delta x_1 = x_1 \otimes x_1,$$

$$\Delta x_2 = x_1 \otimes x_2 + x_2 \otimes x_1^2,$$

$$\Delta x_3 = x_1 \otimes x_3 + x_2 \otimes x_1 x_2 + x_2 \otimes x_2 x_1 + x_3 \otimes x_1^3,$$

$$\Delta x_4 = x_1 \otimes x_4 + x_2 \otimes x_1 x_3 + x_2 \otimes x_2^2 + x_2 \otimes x_3 x_1 + x_3 \otimes x_1^2 x_2 + x_3 \otimes x_1 x_2 x_1 + x_3 \otimes x_2 x_1^2 + x_4 \otimes x_1^4.$$

Moreover, the algebra $\mathbb{C}[[h]]\langle X\rangle/\mathcal{I}_h$ has a Poincaré–Birkhoff–Witt basis.

This is a 3-parameter family of quantum semigroups and *the quantization is exact*. That is, it is *not a quantization modulo* h^6. Simply, no terms of higher order in h arise. This is due to the fact that the algebras under consideration are graded. On the generators the degree is defined by $\deg(x_i) = i-1$ and h has degree $\deg(h) = d$ ($d=1$ in the above case). Quantizations of other brackets are described in Sec. IX. The combinatorial nature of the coefficients of the noncommutative polynomials in the above relations remains a mystery.

The paper is organized as follows. In Sec. II we recall the basic concepts related to the Poisson–Lie theory and formulate the fundamental theorem of Drinfel'd relating Poisson–Lie groups and Lie-bialgebras. In Sec. III we introduce the infinite-dimensional group G_∞ and a smooth structure on it. In Sec. IV we find a class of Poisson–Lie structures on G_∞. In Sec. V we find all bialgebra structures on the Lie algebras $\mathcal{G}_{0\infty}$ and $\mathcal{G}_\infty$ and show that they are all coboundaries. In Sec. VI we show that there is a one-to-one correspondence between the Lie-bialgebra structures on $\mathcal{G}_\infty$ and the Poisson–Lie structures found on G_∞, and prove that the latter give a complete list of all Poisson–Lie structures on G_∞. In Sec. VII we classify all Poisson–Lie structures on $G_{0\infty}$. Section VIII is devoted to elements of representation theory for the Poisson–Lie group G_∞ on the homogeneous spaces V_λ. In Sec. IX we describe a series of finitely generated quantum semigroups. These we believe to be precursors of quantizations of G_∞ and $G_{0\infty}$ which are presently unknown.

II. POISSON–LIE THEORY

In this section we review the basic objects to be studied: Poisson manifolds, Poisson–Lie groups, Lie bialgebras, and some basic results about them.

Let $\mathcal{M}$ be a finite-dimensional smooth manifold. A Poisson structure (bracket) on $\mathcal{M}$ is a bilinear map $\{,\}: C^\infty(\mathcal{M}) \times C^\infty(\mathcal{M}) \to C^\infty(\mathcal{M})$, which makes $C^\infty(\mathcal{M})$ a Lie algebra, and is a derivation with respect to each argument. That is, there exists a section $\omega \in \wedge^2 T_{\mathcal{M}}$, where $T_{\mathcal{M}}$ is the tangent bundle of $\mathcal{M}$, such that for any $f,g,h \in C^\infty(\mathcal{M})$ we have $(f,g) \mapsto \{f,g\} = \langle \omega, df \wedge dg\rangle$, and

(i) $\{\{f,g\},h\} + \{\{g,h\},f\} + \{\{h,f\},g\} = 0$ (Jacobi identity);

(ii) $\{f,gh\} = \{f,g\}h + \{f,h\}g$ (derivation property);

(iii) $\{f,g\} = -\{g,f\}$ (antisymmetry),

where $\langle,\rangle$ denotes the natural pairing between $\wedge^2 T_{\mathcal{M}}$ and $\wedge^2 T^*_{\mathcal{M}}$, where $T^*_{\mathcal{M}}$ is the tangent bundle of $\mathcal{M}$. The second property, (ii), amounts to compatibility between the Lie algebra structure defined by $\{,\}$ and the multiplication in $C^\infty(\mathcal{M})$. In local coordinates,

$$\{f,g\}(x) = \omega_{ij}(x) \frac{\partial f}{\partial x_i} \frac{\partial g}{\partial x_j},$$

where $\omega_x = \omega_{ij}(x)(\partial/\partial x_i) \wedge \partial/\partial x_j \in \wedge^2 T_x$ is a bi-vector field at the point $x \in \mathcal{M}$, and $\{\partial/\partial x_i\}$ is a basis of the tangent space T_x at $x \in \mathcal{M}$ in the local coordinates (x_i).

Here and throughout this text a summation is understood over repeated nonfixed indices unless stated otherwise. Note also that our convention about the position of indices of tensors is the opposite to the standard one. Namely, all contravariant(covariant) tensors have lower(upper) indices. We found this notation more convenient when working with power series, and hope that it will not create confusion.

The Jacobi identity (i) is equivalent to the following system of equations for the components $\omega_{ij}(x) = -\omega_{ji}(x)$:

$$\omega_{ij}\frac{\partial \omega_{kl}}{\partial x_i} + \omega_{ik}\frac{\partial \omega_{lj}}{\partial x_i} + \omega_{il}\frac{\partial \omega_{jk}}{\partial x_i} = 0. \tag{10}$$

Definition II.1 (Poisson Manifold):[15] *A Poisson manifold is a smooth manifold with a Poisson structure.*

A smooth map $F:\mathcal{M}_1 \to \mathcal{M}_2$, of two Poisson manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$, is said to be *Poisson* if $F^*(\{g,h\}_{\mathcal{M}_2}) = \{F^*(g),F^*(h)\}_{\mathcal{M}_1}$, for all $g,h \in C^\infty(\mathcal{M}_2)$, where $(F^*(g))(x) := g(F(x))$, for any $x \in \mathcal{M}_1$, and $\{,\}_{\mathcal{M}_1}$, $\{,\}_{\mathcal{M}_2}$ are the Poisson brackets on $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Thus, the above condition is equivalent to $\{g,h\}_{\mathcal{M}_2} \circ F = \{g\circ F, h\circ F\}_{\mathcal{M}_1}$.

If $\mathcal{M}_1$ and $\mathcal{M}_2$ are two Poisson manifolds with Poisson structures $\omega_1 \in \wedge^2 T_{\mathcal{M}_1}$ and $\omega_2 \in \wedge^2 T_{\mathcal{M}_2}$, respectively, we define the direct product Poisson structure on $\mathcal{M}_1 \times \mathcal{M}_2$ as

$$\omega_1 \times \omega_2 := \omega_1 \times 1 + 1 \times \omega_2, \tag{11}$$

which is a map $:\wedge^2 T_{\mathcal{M}_1} \oplus \wedge^2 T_{\mathcal{M}_2} \hookrightarrow \wedge^2 T_{\mathcal{M}_1 \times \mathcal{M}_2}$. Here the space $C^\infty(\mathcal{M}_1 \times \mathcal{M}_2)$ is identified with the space $C^\infty(\mathcal{M}_1) \otimes C^\infty(\mathcal{M}_2)$ [the reason being that a Poisson structure on $C^\infty(\mathcal{M}_1 \times \mathcal{M}_2)$ is uniquely defined by the one on $C^\infty(\mathcal{M}_1) \otimes C^\infty(\mathcal{M}_2)$] under appropriate completion of the tensor product. In more detail, for any function $f \in C^\infty(\mathcal{M}_1 \times \mathcal{M}_2)$, and for each $x \in \mathcal{M}_1$ and $y \in \mathcal{M}_2$ let us define the functions f^x on $\mathcal{M}_2$ and f^y on $\mathcal{M}_1$ as follows:

$$f^x(y) = f(x,y) \quad \text{and} \quad f^y(x) = f(x,y).$$

Then (11) means

$$\{f_1,f_2\}_{\mathcal{M}_1 \times \mathcal{M}_2}(x,y) = \{f_1^x, f_2^x\}_{\mathcal{M}_2}(y) + \{f_1^y, f_2^y\}_{\mathcal{M}_1}(x),$$

for any two functions $f_1, f_2 \in C^\infty(\mathcal{M}_1 \times \mathcal{M}_2)$.

Definition II.2 (Poisson–Lie group):[16] *Let G be a Lie group. Let ω be a Poisson structure on G. The pair (G,ω) is said to be a Poisson–Lie group if the multiplication map $m:G\times G \to G$ is Poisson, where the manifold $G\times G$ is equipped with the direct product Poisson structure $\omega \times \omega$.*

Let $L_x:G\to G$ and $R_x:G\to G$ be the left and right actions of G on itself defined by $y \mapsto xy$ and $y \mapsto yx$, respectively, where $x,y \in G$. Then for any two functions $f_1,f_2 \in C^\infty(G)$ the compatibility between the product Poisson structure on $G\times G$ introduced by (11) and the Poisson structure on G can be written as

$$\begin{aligned}
\{f_1,f_2\}_G(xy) &= m^*(\{f_1,f_2\}_G)(x,y)\\
&= \{m^*f_1, m^*f_2\}_{G\times G}(x,y)\\
&= \{(m^*f_1)^x,(m^*f_2)^x\}_G(y) + \{(m^*f_1)^y,(m^*f_2)^y\}_G(x)\\
&= \{f_1\circ L_x, f_2\circ L_x\}_G(y) + \{f_1\circ R_y, f_2\circ R_y\}_G(x).
\end{aligned}$$

If $(L_x)_{*y}$ and $(R_y)_{*x}$ are the tangent maps to L_x and R_y evaluated at the points y and x, respectively, we deduce

$$\omega_{xy} = (L_x)_{*y}\omega_y + (R_y)_{*x}\omega_x. \tag{12}$$

In local coordinates,

$$\omega_{ij}(z)=\omega_{kl}(x)\frac{\partial z_i}{\partial x_k}\frac{\partial z_j}{\partial x_l}+\omega_{kl}(y)\frac{\partial z_i}{\partial y_k}\frac{\partial z_j}{\partial y_l}, \tag{13}$$

where $z=xy$.

If $e\in G$ is the identity of G, then (12) yields $2\omega_e=\omega_e$. Therefore $\omega_e=0$. This implies that ω is not a symplectic structure since the rank of ω at the identity of G is zero, and we are dealing with more general Poisson manifolds.

Locally, (10) and (13) are the defining equations of a Poisson–Lie group.

Remark II.1: In the definitions above all manifolds were finite-dimensional ($\mathcal{M}$, *respectively, the group* G). *To extend these to the infinite-dimensional case, one needs two objects:* $T_{\mathcal{M}}$ *and* $C^\infty(\mathcal{M})$. *Since we shall study infinite-dimensional groups in this text, the infinite-dimensional aspects will be addressed at the moment they are introduced.*

We now proceed with the definition of a Lie-bialgebra and formulate a theorem (again due to Drinfel'd) relating the concept of a Lie-bialgebra to the concept of a Poisson–Lie group.

Definition II.3: A Lie-bialgebra $\mathcal{G}$ *is a Lie algebra* $\mathcal{G}$ *equipped with a coalgebra map* $\alpha:\mathcal{G}\to\mathcal{G}\wedge\mathcal{G}$ *such that* α *is a 1-cocycle of* $\mathcal{G}$ *with values in the* $\mathcal{G}$*-module* $\wedge^2\mathcal{G}$*, where* $\mathcal{G}$ *acts on* $\wedge^2\mathcal{G}$ *by means of the adjoint representation, and* α *satisfies the co-Jacobi identity. Thus,* $(\mathcal{G},\alpha)$ *is a Lie bialgebra iff*

$$(i)\quad \tau\circ\alpha=-\alpha,$$

$$(ii)\quad \alpha([X,Y])=\mathrm{ad}_X\alpha(Y)-\mathrm{ad}_Y\alpha(X),\qquad X,Y\in\mathcal{G},$$

$$(iii)\quad [1\otimes 1\otimes 1+(\tau\otimes 1)(1\otimes\tau)+(1\otimes\tau)(\tau\otimes 1)](1\otimes\alpha)\circ\alpha=0,$$

where τ *is the transposition map* $\tau:\mathcal{G}\otimes\mathcal{G}\to\mathcal{G}\otimes\mathcal{G}$ *defined by* $\tau(a\otimes b)=b\otimes a$*, for any* $a,b\in\mathcal{G}$.

This definition encompasses the case when $\mathcal{G}$ is infinite-dimensional. Condition (ii) means that α is a 1-cocycle in the Chevalley–Eilenberg cohomology of Lie algebras. Therefore we will refer to (ii) as the 1-cocycle condition in the sequel. In the case when $\alpha=\delta\alpha^0$ is a 1-coboundary, $\alpha(X)=ad_X r$, where $r\in\mathcal{G}\wedge\mathcal{G}$ is a 0-cochain, which is referred to as the classical r-matrix.[17]

Let $\{e_i\}$ be a basis of $\mathcal{G}$ and let us write α in this basis as $\alpha(e_n)=\alpha_{ij}^n e_i\wedge e_j$. Let C_n^{ij} be the structure constants of $\mathcal{G}$ defining the Lie structure on $\mathcal{G}$ by $[e_i,e_j]=C_n^{ij}e_n$. Property (i) in the definition of α implies that $\alpha_{ij}^n=-\alpha_{ji}^n$. Then the equation (iii) written in terms of α_{ij}^n becomes

$$\alpha_{ij}^n\alpha^{\ j}_{sp}+\alpha_{pj}^n\alpha^{\ j}_{is}+\alpha_{sj}^n\alpha_{pi}^j=0. \tag{14}$$

Similarly, equation (ii) expressed in terms of α_{ij}^n and the structure constants C_n^{ij} of $\mathcal{G}$ becomes

$$C_n^{ij}\alpha_{kl}^n=\alpha_{ml}^jC_k^{im}+\alpha_{km}^jC_l^{im}-\alpha_{ml}^iC_k^{jm}-\alpha_{km}^iC_l^{jm}. \tag{15}$$

Thus, these two systems of equations, (14) and (15), plus the Jacobi identity for the structure constants of the Lie algebra $\mathcal{G}$,

$$C_m^{ij}C_n^{mk}+C_m^{jk}C_n^{mi}+C_m^{ki}C_n^{mj}=0, \tag{16}$$

define a Lie-bialgebra structure on $\mathcal{G}$. We have the following result.

Theorem II.1:[16] *The category of connected, simply connected finite dimensional Poisson–Lie groups is equivalent to the category of finite-dimensional Lie-bialgebras.*

For a proof see, for example, Refs. 18,19. We complete this section by proving a property of (finite-dimensional) Poisson–Lie groups, which is usually assumed to be part of the definition.

Theorem II.2: *Let* G *be a Poisson–Lie group. Then the map* $\varphi:G\to G$ *defined by* $\varphi(x)=x^{-1}$ *is an anti-Poisson map.*

Proof: We prove the statement in a neighborhood of the identity element of G. Let $z_i=z_i(x,y)$, for $i=1,\dots,n$, be the coordinate functions of $z=xy$. After solving $z_i=z_i(x,y)$ with respect to the coordinates of y we have $y_i = y_i(x,z)$. We differentiate the identities $y_i \equiv y_i(x,z(x,y))$, for $i=1,\dots,n$, with respect to y_k for each $k=1,\dots,n$ to obtain

$$\delta_i^k = \left.\frac{\partial y_i}{\partial z_l}\right|_{(x,z)} \left.\frac{\partial z_l}{\partial y_k}\right|_{(x,y)}. \tag{17}$$

Let $\varphi: G \to G$ be the map defined by $\varphi(x) = x^{-1}$, which is given in coordinates by the functions $\varphi_i = \varphi_i(x)$. Then we have $0 = z_i(x,\varphi(x))$, for $i=1,\dots,n$. We differentiate this identity with respect to x_k to obtain

$$0 = \left.\frac{\partial z_l}{\partial x_k}\right|_{(x,\varphi(x))} + \left.\frac{\partial z_i}{\partial y_l}\right|_{(x,\varphi(x))} \frac{\partial \varphi_l}{\partial x_k}. \tag{18}$$

After multiplying both sides of (13) by $\partial y_m/\partial z_i|_{(x,z)}\ \partial y_n/\partial z_j|_{(x,z)}$ and summing over i,j we get

$$\left.\frac{\partial y_m}{\partial z_i}\right|_{(x,z)} \left.\frac{\partial y_n}{\partial z_j}\right|_{(x,z)} \omega_{ij}(z) = \left.\frac{\partial y_m}{\partial z_i}\right|_{(x,z)} \left.\frac{\partial y_n}{\partial z_j}\right|_{(x,z)} \omega_{kl}(x) \frac{\partial z_i}{\partial x_k} \frac{\partial z_j}{\partial x_l} + \omega_{mn}(y), \tag{19}$$

where we have used (17).

We now set $z = e = x\varphi(x)$ in (19), and obtain

$$0 = \left.\frac{\partial y_m}{\partial z_i}\right|_{(x,e)} \left.\frac{\partial y_n}{\partial z_j}\right|_{(x,e)} \omega_{kl}(x) \left.\frac{\partial z_i}{\partial x_k}\right|_{(x,\varphi(x))} \left.\frac{\partial z_j}{\partial x_l}\right|_{(x,\varphi(x))} + \omega_{mn}(\varphi(x)).$$

After using (18) the above equality becomes equivalent to

$$0 = \left.\frac{\partial y_m}{\partial z_i}\right|_{(x,e)} \left.\frac{\partial y_n}{\partial z_j}\right|_{(x,e)} \omega_{kl}(x) \left.\frac{\partial z_i}{\partial y_p}\right|_{(x,\varphi(x))} \frac{\partial \varphi_p}{\partial x_k} \left.\frac{\partial z_j}{\partial y_s}\right|_{(x,\varphi(x))} \frac{\partial \varphi_s}{\partial x_l} + \omega_{mn}(\varphi(x)).$$

Now using again (17) as

$$\delta_i^k = \left.\frac{\partial y_i}{\partial z_l}\right|_{(x,e)} \left.\frac{\partial z_l}{\partial y_k}\right|_{(x,\varphi(x))},$$

we finally conclude that

$$\omega_{mn}(\varphi(x)) = -\omega_{kl}(x) \frac{\partial \varphi_m}{\partial x_k} \frac{\partial \varphi_n}{\partial x_l}. \qquad \square$$

In the following sections we will adapt the above given definitions for the case where G will stand for the infinite-dimensional groups $FDiff(\mathbb{R}^1)$ and $FDiff_0(\mathbb{R}^1)$. We will show that theorems analogous to Theorem II.1 and Theorem II.2 above also hold in this case.

III. THE GROUP OF INFINITE-JETS G_∞ AND ITS LIE ALGEBRA

Let $G_\infty = \{x = (x_1, x_2, \dots) \in \mathbb{R}^\infty | x \neq 0\} \subset \mathbb{R}^\infty$ be a subset of the set of infinite sequences of real numbers. (We may adopt a purely formal point of view and take sequences of letters (indeterminates) which we interpret as generators of an algebra of ''functions'' on the group. This is done in Sec. VI where the group of diffeomorphisms of the line is treated as a formal group. For the group of diffeomorphisms fixing a point, discussed in this section, both points of view are possible, and they lead to the same results, since our treatment is mostly algebraic in nature.) For each $x \in G_\infty$ consider the formal power series, $x(u) = \sum_{i=1}^\infty x_i u^i$, in the variable u. This defines a bijective map from G_∞ into the group of ∞-jets, maps from $\mathbb{R}^1 \to \mathbb{R}^1$ at $0 \in \mathbb{R}^1$, as follows. Define a multiplication $m_\infty : G_\infty \times G_\infty \to G_\infty$ on G_∞ induced by the substitution of formal power series. For any $x,y \in G_\infty$ define $xy \in G_\infty$ by $(xy)(u) := x(y(u))$. The induced multiplication makes G_∞ a group with an identity $e=(1,0,0,\ldots)$. That is, $e:\mathbb{R}^1 \to \mathbb{R}^1$ is the identity map, $e(u)=u$. The associativity of multiplication is implied by the associativity of substitution of power series. In coordinates $z(u)=(xy)(u)=\Sigma_{i=1}^{\infty} z_i u^i$ gives

$$z_k = \sum_{i=1}^{k} x_i \sum_{(\Sigma_{\alpha=1}^{i} j_\alpha)=k} y_{j_1} \ldots y_{j_i}. \tag{20}$$

The first several formulas are given below:

$$z_1 = x_1 y_1,$$

$$z_2 = x_1 y_2 + x_2 y_1^2,$$

$$z_3 = x_1 y_3 + x_2 2 y_1 y_2 + x_3 y_1^3,$$

$$z_4 = x_1 y_4 + x_2 (y_2^2 + 2 y_1 y_3) + x_3 3 y_1^2 y_2 + x_4 y_1^4,$$

$$\vdots$$

$$z_n = x_1 y_n + x_n y_1^n + y_{n-1} 2 y_1 x_2 + x_{n-1}(n-1) y_1^{n-2} y_2 + O(<n-1), \quad n>3,$$

$$\vdots$$

The group so obtained is the group of formal diffeomorphisms (∞-jets) of the line, leaving the origin fixed. It can be viewed as a projective limit of a family of finite-dimensional Lie groups with a smooth structure introduced as follows. Consider the family of Lie groups and maps $(G_n, \pi_{n+1,n})_{n\in\mathbb{N}}$, where $G_n = \mathbb{R}^n \backslash M_n$ and $M_n = \{(x_1,\ldots,x_n) \in \mathbb{R}^n | x_1 = 0\}$. The multiplication $m_n : G_n \times G_n \to G_n$ is induced by substitution $(\mathcal{X}_n \mathcal{Y}_n)(u) = \mathcal{X}_n(\mathcal{Y}_n(u)) \bmod u^{n+1}$, where $\mathcal{X}_n$ and $\mathcal{Y}_n$ are polynomials in u of degree n. That is, the group G_n is an open subset of $\mathbb{R}^n$, that carries the structure of a finite-dimensional C^∞ manifold modeled on $\mathbb{R}^n$. The maps $\pi_{n+1,n} : G_{n+1} \to G_n$ defined by $\pi_{n+1,n}(x_1,\ldots,x_{n+1}) = (x_1,\ldots,x_n)$ are homomorphisms, i.e., $\pi_{n+1,n} \circ m_{n+1} = m_n \circ (\pi_{n+1,n} \times \pi_{n+1,n})$.[20] This follows from the definition of $\pi_{n+1,n}$ and (20). The family $(G_n, \pi_{n+1,n})_{n\in\mathbb{N}}$ has a projective limit $(G_\infty, (\pi_{\infty,n})_{n\in\mathbb{N}})$, where $G_\infty = \{x \in \mathbb{R}^\infty | x_1 \neq 0\}$ is an open subset of $\mathbb{R}^\infty$. The maps $\pi_{\infty,n} : G_\infty \to G_n$ are defined by $\pi_{\infty,n}(x_1,\ldots,x_n,x_{n+1},\ldots) = (x_1,\ldots,x_n)$. Obviously, these maps satisfy $\pi_{\infty,n} = \pi_{n+1,n} \circ \pi_{\infty,n+1}$, and are homomorphisms, $\pi_{\infty,n} \circ m_\infty = m_n \circ (\pi_{\infty,n} \times \pi_{\infty,n})$, where $m_\infty : G_\infty \times G_\infty \to G_\infty$ is defined by (20).

Let us consider now the family of spaces and maps $(C^\infty(G_n), \pi^*_{n,n+1})_{n\in\mathbb{N}}$, where the maps $\pi^*_{n,n+1} : C^\infty(G_n) \to C^\infty(G_{n+1})$ are defined by $(\pi^*_{n,n+1}(f))(x_1,\ldots,x_n,x_{n+1}) = f(x_1,\ldots,x_n)$, for any $f \in C^\infty(G_n)$. Then the above family has an inductive limit $(C^\infty(G_\infty), \pi^*_{n,\infty})$, where $\pi^*_{n,\infty} : C^\infty(G_n) \to C^\infty(G_\infty)$ is defined by $((\pi^*_{n,\infty})(f))(x) = f(x_1,\ldots,x_n)$ for any $x \in G_\infty$ and $f \in C^\infty(G_n)$. Thus, by definition, the space $C^\infty(G_\infty)$ of smooth functions on G_∞ is the space of smooth functions (of *finite* number of variables) on $\mathbb{R}^\infty$, restricted to $\mathbb{R}^\infty \backslash M$, where $M = \{x \in \mathbb{R}^\infty | x_1 = 0\}$.

One can define the Lie algebra of G_∞ in different ways. Probably the most efficient one is as the Lie algebra of derivations (smooth vector fields) of the algebra $C^\infty(G_\infty)$. These are of the form

$$X = \sum_{i=1}^{\infty} v_i \frac{\partial}{\partial x_i}, \quad v_i \in C^\infty(G_\infty). \tag{21}$$

Note that if $f\in C^\infty(G_\infty)$, then $X(f)=\Sigma_{i=1}^n v_i(\partial f/\partial x_i)$ is a finite sum, for some $n\in\mathbb{N}$, since f depends only upon a finite number of variables. We also have $X(f)\in C^\infty(G_\infty)$. Every automorphism $\varphi: G_\infty\to G_\infty$ acts on the space of derivations by $(\varphi_* X)=(\varphi_*^{-1})X\varphi^*$, and on $C^\infty(G_\infty)$ it acts by $(\varphi^* f)(x)=f(\varphi(x))$. Since the functions $f\in C^\infty(G_\infty)$ are functions of a finite number of variables it is enough to describe the map φ_* on vector fields restricted to $C^\infty(G_n)$ for each $n\in\mathbb{N}$.

Lemma III.1: The set $\{X_n\}_{n\geqslant 1}$ of left-invariant vector fields on G_∞ is given by

$$X_n=\sum_{i=1}^{\infty} ix_i\frac{\partial}{\partial x_{i+n-1}}. \tag{22}$$

Proof: From (20) the map $y\mapsto xy$ is given by

$$z_n=(xy)_n=x_1y_n+y_1^n x_n+\sum_{i=2}^{n-1} x_i\sum_{(\Sigma_{\alpha=1}^i j_\alpha)=n} y_{j_1}\ldots y_{j_i},\quad \text{for each}\quad n\geqslant 1. \tag{23}$$

The matrix of the tangent to the map defined by (23) is $\partial z_n/\partial y_m|_{y=e}$. The only terms in

$$\sum_{i=2}^{n-1} x_i\sum_{(\Sigma_{\alpha=1}^i j_\alpha)=n} y_{j_1}\ldots y_{j_i},\quad \text{for each}\quad n\geqslant 1,$$

that will contribute to the tangent map are the ones for which the product $y_{j_1}\cdots y_{j_i}$ has exactly $(i-1)$ multiples equal to y_1 and the one remaining equal to y_{j_α} for some α, $2\leqslant\alpha\leqslant i$. There are exactly $\binom{i}{i-1}=i$ terms of this form. Therefore we rewrite (23) as

$$z_n=\sum_{i=1}^{n} ix_i y_1^{i-1}y_{n-i+1}+\ \ldots,$$

where the dots indicate terms that do not contribute to $\partial z_n/\partial y_m|_{y=e}$. Hence,

$$\left.\frac{\partial z_n}{\partial y_m}\right|_{y=e}=\sum_{i=1}^{n} ix_i\delta_{n-i+1}^m=(n-m+1)x_{n-m+1}.$$

If $\{\partial/\partial y_i\}$ is a basis of vector fields at the identity, then

$$\left(\varphi_*\frac{\partial}{\partial y_m}\right)_x=\sum_{i=1}^{n}\left.\frac{\partial z_i}{\partial y_m}\right|_{y=e}\frac{\partial}{\partial x_i}=\sum_{i=m}^{n}(i-m+1)x_{i-m+1}\frac{\partial}{\partial x_i}=\sum_{i=1}^{n-m+1} ix_i\frac{\partial}{\partial x_{i+m-1}}.$$

Therefore for each $n\in\mathbb{N}$, the set of vector fields $\{X_k\}_{k=1}^n$, where $X_k=\Sigma_{i=1}^{n-k+1}ix_i(\partial/\partial x_{i+k-1})$, for $1\leqslant k\leqslant n$, forms a basis of left-invariant vector fields on G_n. The set $\{X_n\}_{n\geqslant 1}$, where $X_n=\Sigma_{i=1}^{\infty}ix_i(\partial/\partial x_{i+n-1})$, forms a basis of left-invariant vector fields on G_∞. □

Lemma III.2: Every smooth vector field on G_∞ is generated by the set $\{X_n\}_{n\geqslant 1}$ of left-invariant vector fields (22) on G_∞.

Proof: Let $Y=\Sigma_{i=1}^{\infty}v_i(\partial/\partial x_i)$ be a smooth vector field on G_∞. We define inductively the following sequence of smooth vector fields. Let

$$Y_1=Y-\frac{v_1}{x_1}X_1=Y-\psi_1X_1,\quad \text{where}\quad \psi_1:=\frac{v_1}{x_1},$$

$$Y_2=Y_1-\psi_2X_2,\quad \text{where}\quad \psi_2:=\frac{1}{x_1}(v_2-2x_2\psi_1),$$

$$\vdots \tag{24}$$

$$Y_n = Y_{n-1} - \psi_n X_n, \quad \text{where} \quad \psi_n := \frac{1}{x_1}\left(v_n - \sum_{i=2}^{n} i x_i \psi_{n-i+1} \right),$$

$$\vdots$$

Summing up the first n equalities in (24) we get $Y = \Sigma_{i=1}^{n} \psi_i X_i + Y_n$. By construction Y_n is such that $Y_n|_{C^\infty(G_n)} = 0$, for any $n \in \mathbb{N}$. Hence, $Y = \Sigma_{i=1}^{\infty} \psi_i X_i$. □

We now show that $\{X_n\}_{n \geqslant 1}$ forms a Lie subalgebra of the Lie algebra of vector fields on G_∞ with a Lie bracket given by

$$[X_n, X_m] = (n-m) X_{n+m-1}. \tag{25}$$

For that we compute the commutator of two left-invariant vector fields $X_n = \Sigma_{i=1}^{\infty} i x_i (\partial/\partial x_{i+n-1})$ and $X_m = \Sigma_{j=1}^{\infty} j x_j (\partial/\partial x_{j+m-1})$:

$$\begin{aligned} [X_n, X_m] &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} i x_i j \delta^j_{i+n-1} \frac{\partial}{\partial x_{j+m-1}} - \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} j x_j i \delta^i_{j+n-1} \frac{\partial}{\partial x_{i+n-1}} \\ &= \sum_{i=1}^{\infty} i x_i (i+n-1) \frac{\partial}{\partial x_{i+n+m-2}} - \sum_{j=1}^{\infty} j x_j (j+m-1) \frac{\partial}{\partial x_{j+n+m-2}} \\ &= (n-m) \sum_{i=1}^{\infty} i x_i \frac{\partial}{\partial x_{i+n+m-2}} = (n-m) X_{n+m-1}. \end{aligned} \tag{26}$$

To make correspondence with the more familiar notation we shift the indices by 1, and introduce $e_n := X_{n+1}$. Then $[e_n, e_m] = (n-m) e_{n+m}$, for all $n,m \geqslant 0$. The algebra so obtained is the maximal subalgebra of the Witt algebra.

Let us assume now that G_n are equipped with Poisson–Lie structures $\{,\}_n : C^\infty(G_n) \times C^\infty(G_n) \to C^\infty(G_n)$. It is natural to require that the projection maps $\pi_{n+1,n} : G_{n+1} \to G_n$ are Poisson, i.e., $\pi^*_{n+1,n}(\{f,g\}_n) = \{\pi^*_{n+1,n}(f), \pi^*_{n+1,n}(g)\}_{n+1}$, where, as above, $(\pi^*_{n+1,n}(f))(x) := f(\pi_{n+1,n}(x))$ for every $f \in C^\infty(G_n)$ and $x \in G_{n+1}$.

For a smooth bivector field ω on G_∞ [that is, a rank 2 skew-symmetric tensor such that $\omega_{ij} \in C^\infty(G_\infty)$ and $\omega_x = \omega_{ij}(x) (\partial/\partial x_i) \wedge (\partial/\partial x_j)$] let us assume that a map $\{,\} : C^\infty(G_\infty) \times C^\infty(G_\infty) \to C^\infty(G_\infty)$ gives a Poisson–Lie structure on G_∞ defined by $\{f,g\}(x) = \omega_{ij}(x)(\partial f/\partial x_i)(\partial g/\partial x_j)$ for any $f,g \in C^\infty(G_\infty)$. Then it is also natural to require that the maps $\pi_{\infty,n}$ are Poisson. That is, we want the condition $\pi^*_{n,\infty}(\{f,g\}_n) = \{\pi^*_{n,\infty}(f), \pi^*_{n,\infty}(g)\}$ to be satisfied, for any $f,g \in C^\infty(G_n)$. One defines $(G_\infty, \{,\}, (\pi_{\infty,n})_{n \in \mathbb{N}})$ to be the projective limit of the family of Poisson–Lie groups and maps $(G_n, \{,\}_n, \pi_{n+1,n})_{n \in \mathbb{N}}$. Later it will be clear that for the Poisson–Lie groups studied in this text these conditions are automatically satisfied.

Are there any Poisson–Lie structures on G_∞? If such structures exist, can they be classified? Also, since for any finite n there is a one-to-one correspondence between the Poisson–Lie structures on G_n and the Lie-bialgebra structures on the Lie algebra $\mathcal{G}_n$ of G_n, one is led to inquire if there are any Lie-bialgebra structures on the Lie algebra $\mathcal{G}_\infty$ of G_∞. The same questions exist for the group $G_{0\infty}$ of formal diffeomorphisms of the line without fixed points and its Lie algebra $\mathcal{G}_{0\infty}$, which is the Witt algebra. It turns out that all these questions can be fully answered. Let us turn our attention to the group G_∞ first.

IV. POISSON–LIE STRUCTURES ON G_∞

In this section we study Poisson–Lie structures on the group G_∞. It turns out that there exists a large class of such structures, which can be described explicitly. In the next two sections we prove that, in fact, this class exhausts all Poisson–Lie structures on G_∞.

We recall the definition of a Poisson–Lie structure on the group G_∞. It is defined as a skew-symmetric map $\{,\}:C^\infty(G_\infty)\times C^\infty(G_\infty)\to C^\infty(G_\infty)$ which is multiplicative, is a derivation in both arguments, and satisfies the Jacobi identity. The derivation property implies that there is a bi-vector field $\omega\in\wedge^2 TG_\infty$ given locally by $\omega_x=\omega_{ij}(x)(\partial/\partial x_i)\wedge\partial/\partial x_j$, where $\omega_{ij}\in C^\infty(G_\infty)$ are smooth functions on G_∞. Then for every $f,g\in C^\infty(G_\infty)$ we have $\{f,g\}(x)=\omega_{ij}(x)(\partial f/\partial x_i)\times(\partial g/\partial x_j)$. Note that on the right-hand side (r.h.s.) we have in effect a finite sum since the functions f and g depend only upon a finite number of arguments. In particular, $\{\mathcal{X}_i,\mathcal{X}_j\}=\omega_{ij}$, where $\mathcal{X}_i$, $i\in\mathbb{N}$, are the coordinate functions of $x\in G_\infty$, i.e., $\mathcal{X}_i(x)=x_i$. Similarly, the 1-cocycle equation (13) for ω_{ij} is given by

$$\omega_{ij}(xy)=\omega_{kl}(x)\frac{\partial z_i}{\partial x_k}\frac{\partial z_j}{\partial x_l}+\omega_{kl}(y)\frac{\partial z_i}{\partial y_k}\frac{\partial z_j}{\partial y_l}. \tag{27}$$

Here again the sums on the right-hand-side are finite, since for every $n\in\mathbb{N}$ we have $z_n=z_n(x_1,\ldots,x_n;y_1,\ldots,y_n)$. The same is true for the sums in the Jacobi identity (10) in terms of the functions ω_{ij}.

Let us introduce the formal series $\mathcal{X}(u):=\Sigma_{i=1}^{\infty}\mathcal{X}_i u^i$. Then $x(u)=\mathcal{X}(u)(x)=\Sigma_{i=1}^{\infty}x_i u^i$. Define the formal series $\Omega(u,v;\mathcal{X}):=\Sigma_{i,j=1}^{\infty}\omega_{ij}u^i v^j$. Thus $\Omega(u,v;\mathcal{X})$ is a generating series for the brackets ω_{ij}. Evaluating at $x\in G_\infty$ we have $\Omega(u,v;x)=\Sigma_{i,j=1}^{\infty}\omega_{ij}(x)u^i v^j$.

Lemma IV.1: In terms of Ω the cocycle condition (27) assumes the form

$$\Omega(u,v;z)=\Omega(y(u),y(v);x)+\Omega(u,v;y)x'(y(u))x'(y(v)),\quad z(u)=x(y(u)). \tag{28}$$

Proof: Recall that $z(u)=x(y(u))=\Sigma_{i=1}^{\infty}x_i[y(u)]^i=\Sigma_{i=1}^{\infty}z_i u^i$, where $z_i=(xy)_i$. From the last formula we obtain that

$$\frac{\partial z(u)}{\partial x_k}=[y(u)]^k\quad\left(=\sum_{i=1}^{\infty}\frac{\partial z_i}{\partial x_k}u^i\right),$$

and

$$\frac{\partial z(u)}{\partial y_k}=\sum_{i=1}^{\infty}ix_i u^k[y(u)]^{i-1}=u^k\sum_{i=1}^{\infty}ix_i[y(u)]^{i-1}=u^k x'(y(u))\quad\left(=\sum_{i=1}^{\infty}\frac{\partial z_i}{\partial y_k}u^i\right).$$

Here $x'(u)$ denotes the derivative of $x(u)$ with respect to its argument u. If we multiply both sides of Eq. (27) by $u^i v^j$ and sum over i and j we obtain

$$\begin{aligned}\sum_{i,j=1}^{\infty}\omega_{ij}(z)u^i v^j&=\sum_{k,l=1}^{\infty}\omega_{kl}(x)\sum_{i=1}^{\infty}\frac{\partial z_i}{\partial x_k}u^i\sum_{j=1}^{\infty}\frac{\partial z_j}{\partial x_l}v^j+\sum_{k,l=1}^{\infty}\omega_{kl}(y)\sum_{i=1}^{\infty}\frac{\partial z_i}{\partial y_k}u^i\sum_{j=1}^{\infty}\frac{\partial z_j}{\partial y_l}v^j\\&=\sum_{k,l=1}^{\infty}\omega_{kl}(x)[y(u)]^k[y(v)]^l+x'(y(u))x'(y(v))\sum_{k,l=1}^{\infty}\omega_{kl}(y)u^k v^l.\end{aligned}$$

Now, using the definition of Ω we finally obtain that

$$\Omega(u,v;z)=\Omega(y(u),y(v);x)+\Omega(u,v;y)x'(y(u))x'(y(v)).$$

Notice also that both sides of the above equation are divisible by uv. □

Equation (28) has a large class of solutions $\Omega(u,v;x)$. Namely, we have the following theorem.

Theorem IV.1: *For any* $\varphi=\varphi(u,v)$ *with the properties*
(i) $\varphi(u,v)$ *is divisible by* u *and* v;
(ii) $\varphi(u,v)=-\varphi(v,u)$,
we have the following solution of (28)*:*

$$\Omega(u,v;x)=\varphi(u,v)x'(u)x'(v)-\varphi(x(u),x(v)). \tag{29}$$

Proof: Indeed, in terms of (29), the left-hand side of Eq. (28) reads as

$$\begin{aligned}\Omega(u,v;z)&=\varphi(u,v)z'(u)z'(v)-\varphi(z(u),z(v))\\&=\varphi(u,v)x'(y(u))y'(u)x'(y(v))y'(v)-\varphi(z(u),z(v)).\end{aligned}$$

The right-hand side of (28) gives

$$\begin{aligned}&\Omega(y(u),y(v);x)+\Omega(u,v;y)x'(y(u))x'(y(v))\\&\quad=\varphi(y(u),y(v))x'(y(u))x'(y(v))-\varphi(z(u),z(v))\\&\qquad+\varphi(u,v)x'(y(u))y'(u)x'(y(v))y'(v)-\varphi(y(u),y(v))x'(y(u))x'(y(v)).\end{aligned}$$

Comparing both sides we obtain an identity.

The condition (ii) is equivalent to $\Omega(u,v;\mathcal{X})=-\Omega(v,u;\mathcal{X})$ which on the other hand is equivalent to the skew-symmetry of the ω_{ij}'s.

The condition (i) is needed since, as noticed above, $\Omega(u,v;x)$ is divisible by uv. This requires that the r.h.s. of (29) is divisible by uv. From the definition of $x(u)$ it is clear that $x'(u)x'(v)$ is not divisible by uv. It begins with a term $x_1^2+2x_1x_2(u+v)+\cdots$. Suppose that $\varphi(u,v)$ is not divisible by uv. Then $\varphi(x(u),x(v))$ is also not divisible by uv, and so is the difference $\varphi(u,v)z'(u)z'(v)-\varphi(z(u),z(v))$, as an easy analysis shows. □

Next, we would like to find out for which classes of φ's the Jacobi identity is satisfied. This will be an important step in the solution of the problem of classifying all possible Lie–Poisson structures on G_∞. For this we use the following technical tool.

Let $\mathcal{U}=\{u_1,u_2,\ldots\}$ be a countably infinite set of indeterminates. Consider the ring of formal power series $C^\infty(G_\infty)[[\mathcal{U}]]$ in $\mathcal{U}$ over the algebra $C^\infty(G_\infty)$ defined as the inductive limit of the rings $\{C^\infty(G_\infty)[[u_1,\ldots,u_n]]\}_{n\in\mathbb{N}}$. Then the map $\{,\}:C^\infty(G_\infty)\times C^\infty(G_\infty)\to C^\infty(G_\infty)$ induces a map $\{,\}:C^\infty(G_\infty)[[\mathcal{U}]]\times C^\infty(G_\infty)[[\mathcal{U}]]\to C^\infty(G_\infty)[[\mathcal{U}]]$. In particular, we have

$$\{\mathcal{X}(u),\mathcal{X}(v)\}=\sum_{i,j=1}^{\infty}\{\mathcal{X}_i,\mathcal{X}_j\}u^iv^j=\Omega(u,v;\mathcal{X}),$$

where $u=u_i$ and $v=u_j$ for some $u_i,u_j\in\mathcal{U}$. Then the Jacobi identities (10), in terms of generating series, can be put together in a single equation,

$$\{\mathcal{X}(w),\{\mathcal{X}(u),\mathcal{X}(v)\}\}+\{\mathcal{X}(u),\{\mathcal{X}(v),\mathcal{X}(w)\}\}+\{\mathcal{X}(v),\{\mathcal{X}(w),\mathcal{X}(u)\}\}=0, \tag{30}$$

for any $u,v,w\in\mathcal{U}$. On the other hand, we have

$$\{\mathcal{X}(w),\{\mathcal{X}(u),\mathcal{X}(v)\}\}=\{\mathcal{X}(w),\varphi(u,v)\mathcal{X}'(u)\mathcal{X}'(v)-\varphi(\mathcal{X}(u),\mathcal{X}(v)\}$$
$$=\varphi(u,v)[\{\mathcal{X}(w),\mathcal{X}'(u)\}\mathcal{X}'(v)+\{\mathcal{X}(w),\mathcal{X}'(v)\}\mathcal{X}'(u)]$$
$$-\partial_1\varphi(\mathcal{X}(u),\mathcal{X}(v))\{\mathcal{X}(w),\mathcal{X}(u)\}$$
$$-\partial_2\varphi(\mathcal{X}(u),\mathcal{X}(v))\{\mathcal{X}(w),\mathcal{X}(v)\},$$

where ∂_1 denotes the derivative with respect to the first argument and ∂_2 is the derivative with respect to the second argument. Also

$$\{\mathcal{X}(w),\mathcal{X}'(u)\}=\partial_u\{\mathcal{X}(w),\mathcal{X}(u)\}$$
$$=\partial_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}'(u)+\varphi(w,u)\mathcal{X}'(w)\mathcal{X}''(u)+\partial_2\varphi(\mathcal{X}(w),\mathcal{X}(u))\mathcal{X}'(u),$$

and we have similar formulas coming from the remaining two terms in (30) with w,u,v cyclicly permuted.

Lemma IV.2: The solution (29) satisfies the Jacobi identity (30) iff $\varphi(u,v)$ satisfies the following functional partial differential equation:

$$\varphi(u,v)[\partial_u\varphi(w,u)+\partial_v\varphi(w,v)]+\varphi(v,w)[\partial_v\varphi(u,v)+\partial_w\varphi(u,w)]$$
$$+\varphi(w,u)[\partial_w\varphi(v,w)+\partial_u\varphi(v,u)]=0. \tag{31}$$

Proof: After substituting (29) into (30), using the formulas derived above, and collecting terms we obtain

$$(\varphi(u,v)[\partial_u\varphi(w,v)+\partial_v\varphi(w,v)]+\text{cyclic}(u,v,w))\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}'(v)$$
$$+(\varphi(\mathcal{X}(v),\mathcal{X}(u))[\partial_2\varphi(\mathcal{X}(w),\mathcal{X}(v))+\partial_2\varphi(\mathcal{X}(w),\mathcal{X}(u))]+\text{cyclic}(u,v,w))=0. \tag{32}$$

Let us define $\Phi(w,u,v)$ by

$$\Phi(w,u,v):=\varphi(u,v)[\partial_u\varphi(w,u)+\partial_v\varphi(w,v)]$$
$$+\varphi(v,w)[\partial_v\varphi(u,v)+\partial_w\varphi(u,w)]+\varphi(w,u)[\partial_w\varphi(v,w)+\partial_u\varphi(v,u)].$$

It is easily verified that $\Phi(w,u,v)$ is antisymmetric with respect to each pair of its arguments. For example $\Phi(w,u,v)=-\Phi(u,w,v)$. After evaluation at x (32) becomes

$$\Phi(x(w),x(u),x(v))=x'(w)x'(u)x'(v)\Phi(w,u,v).$$

This equation is satisfied for every $x(u)$. In particular, it is true for $x(u)=\lambda u$, where $\lambda\neq 0$. In this case the above equation is equivalent to $\Phi(\lambda w,\lambda u,\lambda v)=\lambda^3\Phi(w,u,v)$. In other words Φ is homogeneous of degree 3. But the only homogeneous function $\Phi(w,u,v)$ of degree 3 which is also antisymmetric with respect to each pair of its arguments is $\Phi=0$. Therefore the statement of the Lemma follows. Note that the last equation shows that the group of formal diffeomorphism acts on the space of solutions of (31), since it is invariant under this action. □

Theorem IV.2: *The map $x\mapsto x^{-1}$ is an anti-Poisson map.*

Proof: Let $\bar{\mathcal{X}}(u)$ denote the inverse of $\mathcal{X}(u)$. Then we have the identities

$$\bar{\mathcal{X}}(\mathcal{X}(u))=u,\quad \text{and}\quad \mathcal{X}(\bar{\mathcal{X}}(u))=u,$$

as well as (following from them)

$$\bar{\mathcal{X}}'(\mathcal{X}(u))\mathcal{X}'(u)=1,\quad \text{and}\quad \mathcal{X}'(\bar{\mathcal{X}}(u))\bar{\mathcal{X}}'(u)=1.$$

On the other hand we have

$$0=\{u,\mathcal{X}(v)\}=\{\overline{\mathcal{X}}(\mathcal{X}(u)),\mathcal{X}(v)\}=\{\overline{\mathcal{X}}(w),\mathcal{X}(v)\}|_{w=\mathcal{X}(u)}+\overline{\mathcal{X}}'(w)|_{w=\mathcal{X}(u)}\{\mathcal{X}(u),\mathcal{X}(v)\}.$$

Therefore,

$$\{\mathcal{X}(v),\overline{\mathcal{X}}(w)\}|_{w=\mathcal{X}(u)}=\overline{\mathcal{X}}'(w)|_{w=\mathcal{X}(u)}\{\mathcal{X}(u),\mathcal{X}(v)\}. \tag{33}$$

Also, we have the following chain of identities:

$$\begin{aligned}0=\{v,\overline{\mathcal{X}}(w)\}|_{w=\mathcal{X}(u)}&=\{\overline{\mathcal{X}}(\mathcal{X}(v)),\overline{\mathcal{X}}(w)\}|_{w=\mathcal{X}(u)}\\&=\{\overline{\mathcal{X}}(s),\overline{\mathcal{X}}(w)\}|_{s=\mathcal{X}(v),w=\mathcal{X}(u)}\\&\quad+\overline{\mathcal{X}}'(s)|_{s=\mathcal{X}(v)}\{\mathcal{X}(v),\overline{\mathcal{X}}(w)\}|_{w=\mathcal{X}(u)}.\end{aligned}$$

Using (29) and (33), the last identity can be rewritten as

$$\begin{aligned}0&=\varphi(\mathcal{X}(v),\mathcal{X}(u))\overline{\mathcal{X}}'(\mathcal{X}(v))\overline{\mathcal{X}}'(\mathcal{X}(u))-\varphi(v,u)\\&\quad+\overline{\mathcal{X}}'(\mathcal{X}(v))\overline{\mathcal{X}}'(\mathcal{X}(u))[\varphi(u,v)\mathcal{X}'(u)\mathcal{X}'(v)-\varphi(\mathcal{X}(u),\mathcal{X}(v))]\\&=\{\overline{\mathcal{X}}(s),\overline{\mathcal{X}}(w)\}|_{s=\mathcal{X}(v),w=\mathcal{X}(u)}+\varphi(u,v)-\overline{\mathcal{X}}'(w)\overline{\mathcal{X}}'(s)\varphi(w,s).\end{aligned}$$

Thus,

$$\{\overline{\mathcal{X}}(w),\overline{\mathcal{X}}(s)\}=-[\overline{\mathcal{X}}'(w)\overline{\mathcal{X}}'(s)\varphi(w,s)-\varphi(\overline{\mathcal{X}}(w),\overline{\mathcal{X}}(s))],$$

and the proof is finished. □

In order to classify *all* Poisson–Lie structures on G_∞ one has, in particular, to classify all solutions of the functional partial differential equation (31). The main result of this section is formulated below.

Theorem IV.3: *The moduli space of solutions of (31) is parametrized by* $\mathbb{N}$, *that is, the space of isomorphism classes of solutions of (31) under the action of the formal group of diffeomorphisms of the line is isomorphic to the set of natural numbers.*

Proof: Let $\Phi(u,v,w)$ be the left-hand side of (31) as defined above. For $w=v+\epsilon$ we expand $\Phi(u,v,v+\epsilon)$ near the diagonal,

$$\Phi(u,v,v+\epsilon)=\left.\frac{\partial\Phi}{\partial w}\right|_{w=v}\epsilon+\frac{1}{2}\left.\frac{\partial^2\Phi}{\partial w^2}\right|_{w=v}\epsilon^2+\cdots .$$

Thus $\Phi(u,v,w)=0$ is equivalent to the infinite system $\partial\Phi/\partial w|_{w=v}=0$, $\partial^2\Phi/\partial w^2|_{w=v}=0, \ldots$. However, using the fact that the equation $\Phi(u,v,w)=0$ is gauge invariant under the action of G_∞ to find all of its solutions, it is enough to find all gauge invariant solutions of $\partial\Phi/\partial w|_{w=v}=0$. Indeed, let S be the space of solutions of $\Phi(u,v,w)=0$ and S' be the space of solutions of $\partial\Phi/\partial w|_{w=v}=0$. Clearly, $S\subset S'$. Thus, restricting S' to the subspace S'_{inv} of gauge invariant solutions we can recover the space S of solutions of the original equation. As we shall see below $S'_{inv}\subset S$ and since we also have $S\subset S'_{inv}$ we shall obtain that $S=S'_{inv}$. We now set out to find S'_{inv}. We have

$$\left.\frac{\partial\Phi}{\partial w}\right|_{w=v}=\varphi(u,v)\partial_v\partial_u\varphi(v,u)+2f(v)\partial_v\varphi(u,v)+\partial_v\varphi(v,u)[f(v)+\partial_u\varphi(v,u)]+\varphi(v,u)g(v)=0,$$

where $f(v) := \partial_2 \varphi(v,v)$, $g(v) = \partial_2^2 \varphi(v,v)$ and we used the fact that $\partial_1 \partial_2 \varphi(v,v) = 0$. Here ∂_1 and ∂_2 denote the derivatives with respect of the first and second arguments. This equation is equivalent to

$$f(v)\partial_v \varphi(u,v) - g(v)\varphi(u,v) = \varphi(u,v)\partial_u \partial_v \varphi(u,v) - \partial_v \varphi(u,v)\partial_u \varphi(u,v). \tag{34}$$

We note that the right-hand side is symmetric with respect to u and v. This implies that

$$f(v)\partial_v \varphi(u,v) - g(v)\varphi(u,v) = -f(u)\partial_u \varphi(u,v) + g(u)\varphi(u,v). \tag{35}$$

Since $\varphi(u,v) = -\varphi(v,u)$, this also implies that $\varphi(u,v) = (u-v)^n \psi(u,v)$, where $\psi(u,v) = \psi(v,u)$, and some odd $n \in \mathbb{N}$. From

$$\partial_v \varphi(u,v) = (u-v)^{n-1}[-n\partial_v \psi(u,v) - (u-v)\partial_v \psi(u,v)],$$

and the definition of $f(v) = \partial_v \varphi(u,v)|_{u=v}$ we obtain that

$$f(v) = \begin{cases} -\psi(v,v), & \text{if } n=1, \\ 0, & \text{if } n>1. \end{cases} \tag{36}$$

Further, from

$$\partial_v^2 \varphi(u,v) = (u-v)^{n-2}[-n(n-1)\psi(u,v) - 2(u-v)\partial_v \psi(u,v) + (u-v)^2 \partial_v^2 \psi(u,v)],$$

and the definition of $g(v) = \partial_v^2 \varphi(u,v)|_{u=v}$ follows that $g=0$, if $n>2$. If $n=2$ we have $g(v) = -2\psi(v,v)$ but then Eq. (35) together with (36) imply that $g(u) = -g(v)$, which implies $g = 0$. If $n=1$ we have $g(v) = -2\partial_2 \psi(v,v) = -\partial_1 \psi(v,v) - \partial_2 \psi(v,v) = f'(v)$. Thus,

$$g(v) = \begin{cases} f'(v), & \text{if } n=1, \\ 0, & \text{if } n>1. \end{cases} \tag{37}$$

Let us assume that $n>1$. Then Eq. (34) implies that

$$\varphi(u,v)\partial_u \partial_v \varphi(u,v) = \partial_v \varphi(u,v)\partial_u \varphi(u,v),$$

which after integration leads to $\varphi(u,v) = \exp(\int C_1(v)dv + C_2(u))$, where $C_1(v) \in k[[v]]$ and $C_2(u) \in k[[u]]$ are formal power series. This is in clear contradiction with $\varphi(u,v) = -\varphi(v,u)$. Therefore $n=1$ and $\varphi(u,v)$ has a zero of order 1 on the diagonal. Thus, from (35) we have

$$-[f'(v)\varphi - f(v)\partial_v \varphi] = [f'(u)\varphi - f(u)\partial_u \varphi],$$

or equivalently,

$$-\partial_v \left(\frac{f(v)}{\varphi} \right) = \partial_u \left(\frac{f(u)}{\varphi} \right).$$

This implies that there is a formal Laurent series $\theta(u,v)$ such that

$$\frac{f(u)}{\varphi(u,v)} = \partial_v \theta(u,v),$$

$$\frac{f(v)}{\varphi(u,v)} = -\partial_u \theta(u,v),$$

from where we obtain

$$f(v)\frac{\partial\theta}{\partial v}+f(u)\frac{\partial\theta}{\partial u}=0. \tag{38}$$

The general solution of (38) is given by $\theta(u,v)=\zeta(F(u)-F(v))$, where $F(u)=\int du/f(u)$ is a formal Laurent series and $\zeta(t)$ is a formal Laurent series in one variable. Thus,

$$\varphi(u,v)=-\frac{1}{F'(u)F'(v)\zeta'(F(u)-F(v))}=-\frac{\eta(F(u)-F(v))}{F'(u)F'(v)}, \tag{39}$$

where $\eta(u)=1/\zeta'(u)$. Let $u\mapsto x(u)$ be a formal diffeomorphism such that $F(x(u))=u$. Then the general solution (39) of $\partial_w\Phi(u,v,w)|_{w=v}=0$ transforms to $\varphi(u,v)=-\eta(u-v)$. To find the space S'_{inv} of invariant solutions of $\partial_w\Phi(u,v,w)|_{w=v}=0$ we substitute back into (34) which leads to

$$-\eta(u-v)\eta''(u-v)+\eta'(0)\eta'(u-v)-\eta'(u-v)^2=0, \tag{40}$$

where we used the fact that $\eta(-u)=-\eta(u)$, which, in particular, implies that $\eta(0)=0=\eta''(0)$. The only nontrivial solution of (40) is $\eta(u)=\eta'(0)u$. Thus, the space S'_{inv} is described by all

$$\varphi(u,v)=\frac{F(v)-F(u)}{F'(u)F'(v)}. \tag{41}$$

By direct substitution one checks that this solution satisfies (31) and thus we have also $S'_{inv}\subset S$. Therefore all solutions of (31) are described by (41). Let

$$F(u)=\sum_{n=-d}^{\infty}C_nu^n=C_{-d}\frac{1}{u^d}+C_{-d+1}\frac{1}{u^{d-1}}+\cdots,$$

where $d>0$ or $d<0$. (By a formal diffeomorphism we can always remove the constant term.) From $F'(u)=1/f(u)$ and that $\varphi(u,v)$ is divisible by uv it follows that $d>0$ for the group G_∞. Let $x(u)$ be a formal diffeomorphism, such that $F(u)=x(u)^{-d}$. Then the solution (41) transforms to

$$\tilde{\varphi}(u,v)=\frac{1}{d^2}u^{d+1}v^{d+1}(v^{-d}-u^{-d})=\frac{1}{d^2}uv(v^d-u^d). \tag{42}$$

For each $d\in\mathbb{Z}_+$ this gives a representative of an isomorphism class of solutions of (31). For each $d\in\mathbb{N}$ formula (29) gives rise to the Poisson–Lie tensor,

$$\omega_{ij}(x)=(i-d)jx_jx_{i-d}-i(j-d)x_ix_{j-d}+x_i\sum_{\Sigma_{k=1}^{d+1}s_k=j}x_{s_1}\ldots x_{s_{d+1}}-x_j\sum_{\Sigma_{k=1}^{d+1}s_k=i}x_{s_1}\ldots x_{s_{d+1}}, \tag{43}$$

where $i,j\in\mathbb{Z}_+$, describing a class of Poisson–Lie structures on G_∞. □

We shall give a more detailed description of the Poisson–Lie structures on G_∞ in the general position. For this, we use the following equivalent formulation of the last theorem.[19]

Theorem IV.4: *For each $d\in\mathbb{N}$, and any formal power series $f_d(u),g_d(u)$ such that $f'_d(u)g_d(u)-f_d(u)g'_d(u)=-d\lambda_{1,d+1}f_d(u)$, where $\lambda_{1,d+1}\neq 0$ is an arbitrary parameter, and f_d has a zero of order $d+1$ at $u=0$, there is a solution of (31) given by the series*

$$\varphi_d(u,v)=\frac{1}{\lambda_{1,d+1}}[f_d(u)g_d(v)-f_d(v)g_d(u)].$$

The set of all solutions of (31) is described in this way.

In more detail this means the following. Let us write $\varphi(u,v)=\Sigma\lambda_{mn}u^m v^n\in k[[u,v]]$. Let

$$M=\{(\mu_1,\mu_2,\ldots)|\mu_i\in k,i\in\mathbb{N}\}$$

be the set of all sequences with elements in the field k.

Then *all* Poisson–Lie structures on G_∞ are described as follows. For each $d\in\mathbb{N}$ and

$$M_d=\{(\underbrace{0,0,\ldots,0}_{d},\mu_{d+1},\mu_{d+2},\ldots)|\mu_{d+1}\neq 0\}\subset M,$$

there is a Poisson–Lie structure on G_∞ given by

$$\{x_i,x_j\}=\sum_{p=1}^{i}\sum_{q=1}^{j}px_pqx_q\lambda_{i-p+1,j-q+1}-\sum_{p=1}^{i}\sum_{q=1}^{j}\lambda_{pq}\sum_{\Sigma_{k=1}^{p}r_k=i}x_{r_1}\ldots x_{r_p}\sum_{\Sigma_{l=1}^{q}s_l=j}x_{s_1}\ldots x_{s_q}, \tag{44}$$

where

$$\lambda_{mn}=\frac{1}{\mu_{d+1}}[\mu_m\lambda_{d+1,n}-\mu_n\lambda_{d+1,m}],\forall_{m,n\geqslant 1}.$$

Clearly $\lambda_{mn}=-\lambda_{nm}$ and we implicitly assume the convention that $\lambda_{mn}=0$ whenever $m<1$ or $n<1$. Here, $\lambda_{d+1,n}$ are given by rational functions $\lambda_{d+1,n}=f_n(\mu_{d+1},\ldots,\mu_{d+n})$ for $n\geqslant 1$, which are computed by the following recursive formula:

$$\lambda_{d+1,n}=-\frac{1}{(d-n+1)\mu_{d+1}}\left[d\mu_{d+1}\mu_{n+d}-\sum_{s=1}^{n-1}(n+d-2s+1)\mu_{n+d-s+1}\lambda_{s,d+1}\right], \tag{45}$$

where $\lambda_{1,d+1}=\mu_{d+1}$. The sequences from M_d are ''almost arbitrary'' in the sense that there is *precisely one* relation between the numbers μ_n:

$$d\mu_{d+1}\mu_{2d+1}+2(d-1)\mu_{2d}\lambda_{1,d+1}+2(d-2)\mu_{2d-1}\lambda_{2,d+1}+\cdots+2\mu_{d+2}\lambda_{d,d+1}=0, \tag{46}$$

which are otherwise subject to no other restrictions. Here

$$\lambda_{1,d+1}=\mu_{d+1},$$

$$\lambda_{2,d+1}=f_2(\mu_{d+1},\mu_{d+2})=\frac{1}{1-d}\mu_{d+2},$$

$$\lambda_{3,d+1}=f_3(\mu_{d+1},\mu_{d+2},\mu_{d+3})=\frac{2}{2-d}\mu_{d+3}+\frac{d}{(d-1)(d-2)}\frac{(\mu_{d+2})^2}{\mu_{d+1}},$$

$$\vdots$$

$$\lambda_{d-1,d+1}=f_{d-1}(\mu_{d+1},\mu_{d+2},\ldots,\mu_{2d-1}),$$

$$\lambda_{d,d+1}=f_d(\mu_{d+1},\mu_{d+2},\ldots,\mu_{2d}),$$

where the functions $f_2, f_3, \ldots, f_d$ are rational in μ_{d+1} and polynomial in $\mu_{d+2}, \ldots, \mu_{2d}$. Thus, the above relation involves only $d+1$ elements of the sequence M_d. Namely, the elements $\mu_{d+1}, \mu_{d+2}, \ldots, \mu_{2d+1}$.

For example, to each sequence,

$$M_d = \{(\underbrace{0,0,\ldots,0}_{d},1,0,0,\ldots), \quad d \in \mathbb{N},$$

there correspond the Poisson–Lie brackets (43). We compute below a set of solutions obtained by an appropriate diffeomorphism from (42). The elements of this set are labeled by the sequences

$$M_d = \{(\underbrace{0,0,\ldots,0}_{d},1,\lambda,\lambda^2,\lambda^3,\ldots), \quad d \in \mathbb{N}, \quad \text{and } \lambda \in k. \tag{47}$$

Lemma IV.3: For each $d \in \mathbb{N}\backslash\{1\}$, *if* $\lambda_{1n} = (\lambda_{1,d+2})^{n-d-1}/(\lambda_{1,d+1})^{n-d-2}$, *for every* $n \geqslant d+1$, *it follows that* $\lambda_{n,d+1} = 0$, *for every* $n \geqslant d+1$, *and* $\lambda_{n,d+1} = -[1/(d-1)][(\lambda_{1,d+2})^{n-1}/(\lambda_{1,d+1})^{n-2}]$, *for all n such that* $2 \leqslant n \leqslant d$.

Proof: From (45) we have

$$\lambda_{n,d+1} = \frac{1}{(d-n+1)\lambda_{1,d+1}}\left[d\frac{(\lambda_{1,d+1})^{n-1}}{(\lambda_{1,d+1})^{n-3}} - \sum_{s=1}^{n-1}(n+d-2s+1)\frac{(\lambda_{1,d+2})^{n-s}}{(\lambda_{1,d+1})^{n-s-1}}\lambda_{s,d+1}\right]. \tag{48}$$

For $n=2$ we obtain

$$\lambda_{2,d+1} = \frac{1}{(d-n+1)\lambda_{1,d+1}}[d\lambda_{1,d+1}\lambda_{1,\,d+2} - (d+1)\lambda_{1,d+2}\lambda_{1,d+1}] = -\frac{1}{d-1}\lambda_{1,d+2}.$$

Let as assume now that $\lambda_{k,d+1} = -[1/(d-1)][(\lambda_{1,d+2})^{k-1}/(\lambda_{1,d+1})^{k-2}]$ for all k, such that $2 \leqslant k \leqslant n-1 < d$. We will now prove that this relation is true for $k=n \leqslant d$. Indeed, we have

$$\begin{aligned}\lambda_{n,d+1} &= \frac{1}{(d-n+1)\lambda_{1,d+1}}\left[d\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-3}} + \frac{1}{d-1}\sum_{s=1}^{n-1}(n+d-2s+1)\frac{(\lambda_{1,d+2})^{n-s}}{(\lambda_{1,d+1})^{n-s-1}}\lambda_{s,d+1}\right] \\ &= \frac{1}{(d-n+1)}\left[-(n-1) + \frac{1}{d-1}(n-2)(n+d+1) - \frac{2}{d-1}\left(\frac{n(n-1)}{2} - 1\right)\right]\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-2}} \\ &= -\frac{1}{d-1}\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-2}}.\end{aligned}$$

This proves the second part of the lemma. If $n=d+1$, then the l.h.s. of (48) is zero. We check for the consistency of whether the r.h.s. is zero. Indeed, the expression in the brackets reads as

$$-d + \frac{1}{d-1}\sum_{s=2}^{d}2(d-s+1) = \frac{1}{d-1}\left[-d^2 + d + 2(d+1)(d-1) - 2\left(\frac{d(d+1)}{2} - 1\right)\right] = 0.$$

Now, we check whether the statement is true for $\lambda_{d+2,d+1}$. From (48) one has

$$\begin{aligned}\lambda_{d+2,d+1} &= -\left[-(d+1) + \frac{1}{d-1}\sum_{s=2}^{d}(2d+3-2s)\right]\frac{(\lambda_{1,d+2})^{d+1}}{(\lambda_{1,d+1})^{d}} \\ &= -\frac{1}{d-1}\left[-d^2 + 1 + (d-1)(2d+3) - 2\left(\frac{d(d+1)}{2} - 1\right)\right]\frac{(\lambda_{1,d+2})^{d+1}}{(\lambda_{1,d+1})^{d}} = 0.\end{aligned}$$

We use again an inductive argument. We assume that $\lambda_{k,d+1}=0$ for all k such that $d+1\leqslant k\leqslant n-1$, and we aim to show that this implies $\lambda_{n,d+1}=0$. With these assumptions from (48) we have

$$\lambda_{n,d+1}=\frac{1}{(d-n+)}\left[d\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-2}}-\sum_{s=1}^{d}(n+d-2s+1)\frac{(\lambda_{1,d+2})^{n-s}}{(\lambda_{1,d+1})^{n-2}}\lambda_{s,d+1}\right]$$
$$=\frac{1}{(d-n+1)}\left[-(n-1)+\frac{1}{d-1}\sum_{s=2}^{d}(n+d-2s+1)\right]\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-2}}$$
$$=\frac{1}{(d-n+1)(d-1)}\left[-(n-1)(d-1)+(d-1)(n+d+1)\right.$$
$$\left.-2\left(\frac{d(d+1)}{2}-1\right)\right]\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-2}}=0.$$

This concludes the proof of the lemma. □

Theorem IV.5: *For every $d\geqslant 2$ and $\lambda:=\lambda_{1,d+2}/\lambda_{1,d+1}$ the formal power series,*

$$\varphi_{d,\lambda}(u,v)=\frac{1}{(d-1)(1-\lambda u)(1-\lambda v)}\{(d-1)uv(v^{d}-u^{d})+\lambda\, du^{2}\, v^{2}(u^{d-1}-v^{d-1})\},\qquad(49)$$

is a solution of (31). It is a one-parameter extension of the solution $\varphi(u,v)=uv(v^{d}-u^{d})$, for every $d\geqslant 2$, which we obtain back from (49) by setting $\lambda=0$.

Proof: With the assumptions of Lemma IV.3 we have

$$f_{d}(u)=\sum_{n=d+1}^{\infty}\lambda_{1n}u^{n}=\sum_{n=d+1}^{\infty}\frac{(\lambda_{1,d+2})^{n-d-1}}{(\lambda_{1,d+1})^{n-d-2}}u^{n}=\lambda_{1,d+1}u^{d+1}\sum_{n=0}^{\infty}\left(\frac{\lambda_{1,d+2}}{\lambda_{1,d+1}}u\right)^{n}=\frac{\lambda_{1,d+1}}{1-\lambda u}u^{d+1},$$

and

$$g_{d}(u,v)=-\sum_{n=1}^{d}\lambda_{n,d+1}u^{n}=-\lambda_{1,d+1}u+\frac{1}{d-1}\sum_{n=2}^{d}\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-2}}u^{n}$$
$$=-\lambda_{1,d+1}u+\frac{1}{d-1}\lambda_{1,d+2}u^{2}\left[\frac{1-(\lambda u)^{d-1}}{1-\lambda u}\right]$$
$$=\frac{\lambda_{1,d+1}}{(d-1)(1-\lambda u)}[d\lambda u^{2}-u(d-1+(\lambda u)^{d})].$$

By Theorem IV.4 we obtain

$$\varphi_{d,\lambda}(u,v)=\frac{1}{\lambda_{1,d+1}}[f_{d}(u)g_{d}(v)-f_{d}(v)g_{d}(u)]$$
$$=\frac{1}{(d-1)(1-\lambda u)(1-\lambda v)}\{v^{d+1}[-\lambda du^{2}+u(d-1+(\lambda u)^{d})]$$
$$-u^{d+1}[-\lambda dv^{2}+v(d-1+(\lambda v)^{d})]\}$$
$$=\frac{1}{(d-1)(1-\lambda u)(1-\lambda v)}\{(d-1)uv(v^{d}-u^{d})+\lambda du^{2}v^{2}(u^{d-1}-v^{d-1})\}$$
$$=\frac{F(v)-F(u)}{F'(u)F'(v)},$$

where $F(u)=-(1/d)u^{-d}+[\lambda/(d-1)]u^{-d+1}$. □

Remark IV.1: One can obtain the solution $\varphi(u,v)=uv(v-u)$, *which gives the Poisson–Lie structure for* $d=1$ *in Theorem IV.3, from (49) in the following way. Rewriting (49) as*

$$\varphi_{d,\lambda}(u,v)=\frac{uv(v^d-u^d)}{(1-\lambda u)(1-\lambda v)}+\frac{\lambda\ du^2\ v^2}{(1-\lambda u)(1-\lambda v)}\frac{u^{d-1}-v^{d-1}}{d-1},$$

we pass to the limit $d\to 1$ *and then set* $\lambda=0$.

Summarizing, we showed the existence of countably infinite classes of Poisson–Lie structures on the group G_∞. In the following two sections we shall show that this family of classes exhausts all such possible structures on G_∞.

V. BIALGEBRA STRUCTURES ON THE WITT ALGEBRA $\mathcal{G}_{0\infty}$ AND ITS PRINCIPAL SUBALGEBRA $\mathcal{G}_\infty$

In this section we use completely elementary methods to compute the Shevalley–Eilenberg cohomology groups $H^1(\mathcal{G}_{0\infty},\mathcal{G}_{0\infty}\hat{\wedge}\mathcal{G}_{0\infty})$ and $H^1(\mathcal{G}_\infty,\mathcal{G}_\infty\hat{\wedge}\mathcal{G}_\infty)$ with coefficients in the completion of the tensor product under the natural grading induced from the gradings on $\mathcal{G}_{0\infty}$ and $\mathcal{G}_\infty$. We show that both groups are trivial. Thus, we find all 1-cocycles on the Lie algebra $\mathcal{G}_{0\infty}$ (the Witt algebra), and all 1-cocycles on the Lie algebra $\mathcal{G}_\infty$ of the group G_∞. All of them are coboundaries, and they are all explicitly enumerated. The algebras $\mathcal{G}_{0\infty}$ and $\mathcal{G}_\infty$ are taken over $\mathbb{R}$ (or $\mathbb{C}$).

Let $\{e_i=u^{i+1}(d/du)\}_{i\geqslant -1}$ be the canonical basis of the Lie algebra $\mathcal{G}_{0\infty}$ (the Witt algebra). We recall that the Lie algebra structure on $\mathcal{G}_{0\infty}$ is defined by

$$[e_n,e_m]=(n-m)e_{n+m},\quad n,m\geqslant -1.$$

To find the 1-cocycles we turn to the 1-cocycle equation,

$$\alpha([e_n,e_m])=e_n.\alpha(e_m)-e_m.\alpha(e_n),$$

which we rewrite as

$$(n-m)\alpha(e_{n+m})=e_n.\alpha(e_m)-e_m.\alpha(e_n),\quad n,m\geqslant -1. \tag{50}$$

Let $\alpha(e_n)=\Sigma_{i,j=-1}^{\infty}\alpha_{ij}^n e_i\wedge e_j\in\mathcal{G}_{0\infty}\hat{\wedge}\mathcal{G}_{0\infty}=\oplus_k\oplus_{i+j=k}\mathcal{G}_{0i}\wedge\mathcal{G}_{0j}$, where $\mathcal{G}_{0i}$ is the 1-dimensional subspace of $\mathcal{G}_{0\infty}$ generated by e_i. Then (50) is equivalent to the following infinite system of linear equations:

$$(n-m)\alpha_{ij}^{n+m}=(2n-i)\alpha_{i-n,j}^m+(2n-j)\alpha_{i,j-n}^m-(2m-i)\alpha_{i-m,j}^n-(2m-j)\alpha_{i,j-m}^n, \tag{51}$$

where $n,m,i,j\geqslant -1$. Therefore, to find all 1-cocycles, one has to describe all solutions of this system.

Theorem V.1: *All 1-cocycles on the Lie algebra* $\mathcal{G}_{0\infty}$ *are coboundaries.*

Proof: We observe that (51) consists of two independent subsystems. Namely, one for α_{ij}^n's with $n\neq i+j$ and the other for α_{ij}^n's with $n=i+j$.

Set $m=0$ in (51). [Note that (51) is invariant with respect to $m\to n$, $n\to m$.] Then from (51) we deduce that

$$(n-i-j)\alpha_{ij}^n=(2n-i)\alpha_{i-n,j}^0+(2n-j)\alpha_{i,j-n}^0,\quad \text{for every}\quad n,i,j\geqslant -1. \tag{52}$$

(a) In case $n\neq i+j$ we obtain from (52) immediately the solution of the first subsystem,

$$\alpha_{ij}^n=\frac{(2n-i)}{(n-i-j)}\alpha_{i-n,j}^0+\frac{(2n-j)}{(n-i-j)}\alpha_{i,j-n}^0.$$

This means that the coalgebra structure constants α_{ij}^{n}, where $n\neq i+j$, are given in terms of α_{ij}^{0}, $i,j\in\mathbb{Z}_{+}$, which are arbitrary.

(b) The case $n=i+j$ requires a more thorough analysis. Set $j=n+m-i$ in (51). Then we have

$$(n-m)\alpha_{i,n+m-i}^{n+m}=(2n-i)\alpha_{i-n,m+n-i}^{m}+(n-m+i)\alpha_{i,m-i}^{m} \\ -(2m-i)\alpha_{i-m,n+m-i}^{n}-(m-n+i)\alpha_{i,n-i}^{n}, \tag{53}$$

where $n,m,i\geqslant -1$. It turns out that it is enough to investigate (53) for a few values of m and i in order to obtain a complete set of reccurence relations sufficient to find the general solution for $\alpha_{i,n-i}^{n}$. First we set $m=0$ in (53). This implies that $0=(2n-i)\alpha_{i-n,n-i}^{0}+(n+i)\alpha_{i,-i}^{0}$, for every $i,n\geqslant -1$. In particular, if $n=i$ we obtain $0=i\alpha_{-i,i}^{0}$, for every $i\geqslant -1$, since $\alpha_{0,0}^{0}=0$ by the antisymmetry of $\alpha_{ij}^{n}(=-\alpha_{ji}^{n})$. Therefore $\alpha_{-i,i}^{0}=0$, for every $i\geqslant -1$. Set $m=-1=i$ in (53) to obtain

$$(n+1)\alpha_{-1,n}^{n-1}=(2n+1)\alpha_{-1-n,n}^{-1}+n\alpha_{-1,0}^{-1}+\alpha_{0,n}^{n}+(n+2)\alpha_{-1,n+1}^{n}. \tag{54}$$

Then we have the following lemma.

Lemma V.1: One has $\alpha_{-1,n+1}^{n}=[1/(n+2)([n(n+1)/2]\alpha_{0,-1}^{-1}-\Sigma_{i=1}^{n}\alpha_{0,i}^{i})$, *for every* $n\geqslant 1$.

Proof: For $n=-1,0$, (54) is identically satisfied. For $n=1$ we have from (54): $2\alpha_{-1,1}^{0}=\alpha_{-1,0}^{-1}+\alpha_{0,1}^{1}+3\alpha_{-1,2}^{1}$, where we used that $\alpha_{-2,1}^{-1}=0$. From this it follows that $\alpha_{-1,2}^{1}=\frac{1}{3}[\alpha_{0,-1}^{-1}-\alpha_{0,1}^{1}]$, Assume that $\alpha_{-1,k+1}^{k}=[1/(k+2)]([k(k+1)/2]\alpha_{0,-1}^{-1}-\Sigma_{i=1}^{k}\alpha_{0,i}^{i})$, for all k, such that $1\leqslant k\leqslant n-1$. Then from (54), using the induction hypothesis we have

$$\begin{aligned}\alpha_{-1,n+1}^{n}&=\frac{1}{(n+2)}[n\alpha_{0,-1}^{-1}-\alpha_{0,n}^{n}+(n+1)\alpha_{-1,n}^{n-1}]\\ &=\frac{1}{(n+2)}\left[n\alpha_{0,-1}^{-1}-\alpha_{0,n}^{n}+\frac{n(n-1)}{2}\alpha_{0,-1}^{-1}-\sum_{i=1}^{n-1}\alpha_{0,i}^{i}\right]\\ &=\frac{1}{(n+2)}\left[\frac{n(n+1)}{2}\alpha_{0,-1}^{-1}-\sum_{i=1}^{n}\alpha_{0,i}^{i}\right].\end{aligned}$$

□

Lemma V.2: One has $\alpha_{0,2}^{2}=\alpha_{0,1}^{1}-\alpha_{0,-1}^{-1}$, *and* $\alpha_{1,n-1}^{n}=[(n+1)/2][-\alpha_{0,-1}^{-1}+\alpha_{0,n-1}^{n-1}-\alpha_{0,n}^{n}]$, *for every* $n\geqslant 3$.

Proof: From (53), after setting $m=-1,i=0$, we obtain

$$(n+1)\alpha_{0,n-1}^{n-1}=2n\alpha_{-n,n-1}^{-1}+(n+1)\alpha_{0,-1}^{-1}+2\alpha_{1,n-1}^{n}+(n+1)\alpha_{0,n}^{n}. \tag{55}$$

For $n=-1,0,1$, (55) is identically satisfied. Next, for $n=2$ we obtain from (55) that

$$\alpha_{0,2}^{2}=\alpha_{0,1}^{1}-\alpha_{0,-1}^{-1}, \tag{56}$$

since $\alpha_{-2,1}^{-1}=0$. Finally, since $\alpha_{-n,n-1}^{-1}=0$ for $n\geqslant 2$, we obtain from (55) the second assertion of the Lemma. □

Lemma V.3: The formula $\alpha_{2,n-2}^{n}=[n(n+1)/6][\alpha_{0,n-2}^{n-2}-2\alpha_{0,n-1}^{n-1}+\alpha_{0,n}^{n}]$ *is valid for every* $n\geqslant 5$.

Proof: In (53) set $m=-1$, $i=1$. Then we have

$$(n+1)\alpha_{1,n-2}^{n-1}=(2n-1)\alpha_{1-n,n-2}^{-1}+3\alpha_{2,n-2}^{n}+n\alpha_{1,n-1}^{n}.$$

For $n=-1,0$ this formula yields trivial identitites. For $n=1$ we obtain that $3\alpha^{1}_{-1,2}=\alpha^{-1}_{0,-1}-\alpha^{1}_{0,1}(=-\alpha^{2}_{0,2})$. The case $n=2$ reduces to formula (56), and $n=3$ yields the identity $0=3\alpha^{3}_{2,1}+3\alpha^{3}_{1,2}$. For $n=4$ we obtain the relation

$$5\alpha^{3}_{1,2}=4\alpha^{4}_{1,3}. \tag{57}$$

If $n\geqslant 5$, then $\alpha^{-1}_{1-n,n-2}=0$, and using Lemma V.2 we obtain that

$$\begin{aligned}\alpha^{n}_{2,n-2}&=\frac{1}{3}[(n+1)\alpha^{n-1}_{1,n-2}-n\alpha^{n}_{1,n-1}]\\&=\frac{1}{3}\left[\frac{n(n+1)}{2}(-\alpha^{-1}_{0,-1}+\alpha^{n-2}_{0,n-2}-\alpha^{n-1}_{0,n-1})-\frac{n(n+1)}{2}(-\alpha^{-1}_{0,-1}+\alpha^{n-1}_{0,n-1}-\alpha^{n}_{0,n})\right]\\&=\frac{n(n+1)}{6}[\alpha^{n-2}_{0,n-2}-2\alpha^{n-1}_{0,n-1}+\alpha^{n}_{0,n}].\end{aligned}$$

□

The result obtained in Lemma V.3 is suggestive in finding a general formula for $\alpha^{n}_{i,n-i}$ (that is, α^{n}_{ij} with $n=i+j$) in terms of $\alpha^{n}_{0,n},\ldots,\alpha^{n-i}_{0,n-i}$, for every $i\geqslant 2$ and $n\geqslant i+3$.

Lemma V.4: For every $i\geqslant 2$ and $n\geqslant i+3$ the following formula is valid:

$$\alpha^{n}_{i,n-i}=\frac{1}{i+1}\binom{n+1}{i}\sum_{k=0}^{i}\binom{i}{k}(-1)^{i-k}\alpha^{n-k}_{0,n-k}. \tag{58}$$

Proof: We use an induction in i. For $i=2$ the statetment is true by Lemma V.3. Assume that (58) is true for all $\alpha^{n}_{j,n-j}$, where $2\leqslant j\leqslant i$. We now proceed in proving that it is true for $j=i+1$ and all $n\geqslant i+4$. Setting $m=-1$ in (53) and using that $i\geqslant 2$ and $n\geqslant i+3$ we obtain

$$(n+1)\alpha^{n-1}_{i,n-i-1}-(n+1-i)\alpha^{n}_{i,n-i}=(i+2)\alpha^{n}_{i+1,n-i-1}. \tag{59}$$

Then we use the induction hypothesis in the left-hand side of (59) and deduce that

$$\begin{aligned}\alpha^{n}_{i+1,n-i-1}&=\frac{1}{i+2}\left[\frac{(n+1)}{i+1}\binom{n}{i}\sum_{k=0}^{i}\binom{i}{k}(-1)^{i-k}\alpha^{n-k-1}_{0,n-k-1}\right.\\&\qquad\left.-\frac{(n+1-i)}{i+1}\binom{n+1}{i}\sum_{k=0}^{i}\binom{i}{k}(-1)^{i-k}\alpha^{n-k}_{0,n-k}\right]\\&=\frac{1}{i+2}\binom{n+1}{i+1}\left[\sum_{k=0}^{i}\binom{i}{k}(-1)^{i-k}\alpha^{n-k-1}_{0,n-k-1}-\sum_{k=0}^{i}\binom{i}{k}(-1)^{i-k}\alpha^{n-k}_{0,n-k}\right]\\&=\frac{1}{i+2}\binom{n+1}{i+1}\left[\sum_{k=1}^{i+1}\binom{i}{k-1}(-1)^{i-k+1}\alpha^{n-k}_{0,n-k}+\sum_{k=0}^{i}\binom{i}{k}(-1)^{i-k+1}\alpha^{n-k}_{0,n-k}\right]\\&=\frac{1}{i+2}\binom{n+1}{i+1}\left[\sum_{k=1}^{i}\left[\binom{i}{k-1}+\binom{i}{k}\right](-1)^{i-k+1}\alpha^{n-k}_{0,n-k}+\alpha^{n-i-1}_{0,n-i-1}+(-1)^{i+1}\alpha^{n}_{0,n}\right]\\&=\frac{1}{i+2}\binom{n+1}{i+1}\sum_{k=0}^{i+1}\binom{i+1}{k}(-1)^{i+1-k}\alpha^{n-k}_{0,n-k},\end{aligned}$$

which concludes the proof. It is also an immediate corollary that the relation (57) is identically satisfied. □

Let $m=1,i=-1$ in (53). Then we have

$$(n-1)\alpha^{n+1}_{-1,n+2}=(2n+1)\alpha^{1}_{-1-n,n+2}+(n-2)\alpha^{1}_{-1,2}+n\alpha^{n}_{-1,n+1}. \tag{60}$$

For $n=-1$ (60) yields $\alpha^{1}_{-1,2}=\frac{1}{3}(\alpha^{-1}_{0,-1}-\alpha^{1}_{0,1})$. For $n=0,1$, it is identically satisfied. For $n=2$, we obtain the relation

$$\alpha^{3}_{-1,4}=2\alpha^{2}_{-1,3}. \tag{61}$$

Continuing further we have

$$2\alpha^{4}_{-1,5}=\alpha^{1}_{-1,2}+3\alpha^{3}_{-1,4} \quad (n=3),$$

$$3\alpha^{5}_{-1,6}=2\alpha^{1}_{-1,2}+4\alpha^{4}_{-1,5} \quad (n=4),$$

$$\vdots$$

which, as we shall see later, are all identically satisfied.

Lemma V.5: The following formula is valid for every $n\geqslant 2$*:*

$$\alpha^{n}_{0,n}=a_n(\alpha^{1}_{0,1}-\alpha^{-1}_{0,-1}),$$

where $a_2=1$, $a_3=3$, *and the coefficients* a_n, $(n\geqslant 4)$ *are computed by the recursive formula*

$$a_{n+1}=\frac{2n}{(n-1)(n+2)}+\frac{2(n+1)}{(n+2)}a_n-\frac{(n+1)(n-2)}{(n-1)(n+2)}a_{n-1},$$

which is valid for all $n\geqslant 3$.

Proof: For $n=2$ the statement is true according to formula (56). In order to compute a_3 we use the relation (61) and Lemma V.2. From Lemma V.1 we have

$$\alpha^{2}_{-1,3}=\tfrac{1}{2}[2\alpha^{-1}_{0,-1}-\alpha^{1}_{0,1}], \quad \text{and} \quad \alpha^{3}_{-1,4}=\tfrac{1}{5}[7\alpha^{-1}_{0,-1}-2\alpha^{1}_{0,1}-\alpha^{3}_{0,3}].$$

Then from (61) it follows that $\alpha^{3}_{0,3}=3(\alpha^{1}_{0,1}-\alpha^{-1}_{0,-1})$. After setting $m=1,i=1$ in (53) one obtains

$$(n-1)\alpha^{n+1}_{1,n}=-n\alpha^{1}_{0,1}-\alpha^{n}_{0,n}+(n-2)\alpha^{n}_{1,n-1}, \quad \text{for } n\geqslant 3, \tag{62}$$

since $\alpha^{1}_{1-n,n}=0$, for $n\geqslant 3$. Next, we use Lemma V.2 to write formulas for

$$\alpha^{n}_{1,n-1}=\frac{(n+1)}{2}[-\alpha^{-1}_{0,-1}+\alpha^{n+1}_{0,n+1}-\alpha^{n}_{0,n}] \quad \text{and} \quad \alpha^{n+1}_{1,n}=\frac{(n+2)}{2}[-\alpha^{-1}_{0,-1}+\alpha^{n}_{0,n}-\alpha^{n+1}_{0,n+1}].$$

Substituting the above formulas into (62) we obtain, after some algebra, that

$$\alpha^{n+1}_{0,n+1}=\frac{2n}{(n-1)(n+2)}(\alpha^{1}_{0,1}-\alpha^{-1}_{0,-1})+\frac{2(n+1)}{(n+2)}\alpha^{n}_{0,n}-\frac{(n+1)(n-2)}{(n-1)(n+2)}\alpha^{n-1}_{0,n-1}, \quad \text{for all } n\geqslant 3. \tag{63}$$

Now, we make an induction hypothesis. Namely, we assume that for all k, $2\leqslant k\leqslant n$, it is true that $\alpha^{k}_{0,k}=a_k(\alpha^{1}_{0,1}-\alpha^{-1}_{0,-1})$, for some a_k. Then from (63) it follows that $\alpha^{n+1}_{0,n+1}=a_{n+1}(\alpha^{1}_{0,1}-\alpha^{-1}_{0,-1})$, where

$$a_{n+1}=\frac{2n}{(n-1)(n+2)}+\frac{2(n+1)}{(n+2)}a_n-\frac{(n+1)(n-2)}{(n-1)(n+2)}a_{n-1}.$$

The first several coefficients are: $a_2=1$, $a_3=3$, $a_4=5$, $a_5=\frac{64}{9}$, $a_6=\frac{28}{3}$, $a_7=\frac{451}{45}$, ... This concludes the proof. □

Corollary V.1: From Lemma V.4 *and Lemma* V.5 *it follows that*

$$\alpha^n_{i,n-i} = \frac{1}{i+1}\binom{n+1}{i}\sum_{k=0}^{i}\binom{i}{k}(-1)^{i-k}a_{n-k}(\alpha^1_{0,1}-\alpha^{-1}_{0,-1}),$$

for every $i \geqslant 2$ *and* $n \geqslant i+3$.

Lemma V.6: One has the following relation: $\alpha^{-1}_{0,-1}=\alpha^1_{0,1}$.

Proof: In (53) set $m=0, i=-1$. Then one has

$$(n-2)\alpha^{n+2}_{-1,n+3}=(2n+1)\alpha^2_{-1-n,n+3}+(n-3)\alpha^2_{-1,3}+(n-1)\alpha^n_{-1,n+1}. \tag{64}$$

For $n=-1$ we obtain from (64) that $-3\alpha^1_{-1,2}=-\alpha^2_{0,2}-4\alpha^2_{-1,3}-2\alpha^{-1}_{-1,0}$, which, after using Lemma V.1, turns into an identity. For $n=0$ we have an identity. For $n=1$ we obtain $-\alpha^3_{-1,4} = -2\alpha^2_{-1,3}$, which leads to an identity after using Lemma V.1 and Lemma V.5. The case $n=2$ leads to an identity. For $n=3$ we obtain the nontrivial relation

$$\alpha^5_{-1,6}=2\alpha^3_{-1,4}. \tag{65}$$

We use now Lemma V.1 and Lemma V.5 to reduce both sides of the relation (65) which leads to

$$\frac{1}{7}\left[15\alpha^{-1}_{0,-1}-\sum_{i=1}^{5}\alpha^i_{0,i}\right]=\frac{2}{5}\left[6\alpha^{-1}_{0,-1}-\sum_{i=1}^{3}\alpha^i_{0,i}\right],$$

$$\frac{1}{7}\left[15\alpha^{-1}_{0,-1}-\alpha^1_{0,1}-\left(1+3+5+\frac{64}{9}\right)(\alpha^1_{0,1}-\alpha^{-1}_{0,-1})\right]=\frac{2}{5}[6\alpha^{-1}_{0,-1}-\alpha^1_{0,1}-(1+3)(\alpha^1_{0,1}-\alpha^{-1}_{0,-1})],$$

$$\frac{40}{9}\alpha^{-1}_{0,-1}-\frac{22}{9}\alpha^1_{0,1}=4\alpha^{-1}_{0,-1}-2\alpha^1_{0,1},$$

$$\alpha^{-1}_{0,-1}=\alpha^1_{0,1}.$$

This concludes the proof. □

Immediately several corollaries follow.

Corollary V.2: One has $\alpha^n_{0,n}=0$, *for all* $n \geqslant 2$.

Corollary V.3: One obtains that $\alpha^n_{-1,n+1}=[1/(n+2)]([n(n+1)/2]\alpha^{-1}_{0,-1}-\alpha^1_{0,1})=[(n-1)/2]\alpha^1_{0,1}$, *for all* $n \geqslant -1$.

Corollary V.4: One has $\alpha^n_{i,n-i}=0$, *for all* $i \geqslant 2$ *and* $n \geqslant i+3$.

Corollary V.5: One has $\alpha^n_{1,n-1}=-[(n+1)/2]\alpha^1_{0,1}$, *for all* $n \geqslant 3$.

In the above formulas $\alpha^1_{0,1}$ is arbitrary. As a result we are now able to write a formula for the general solution of (50). Namely,

$$\begin{aligned}
\alpha(e_n) &= \sum_{i,j=-1}^{\infty}\alpha^n_{ij}e_i\wedge e_j \\
&= \sum_{i,j=-1, i+j\neq n}^{\infty}\alpha^n_{ij}e_i\wedge e_j+\sum_{i,j=-1,i+j=n}^{\infty}\alpha^n_{ij}e_i\wedge e_j \\
&= \sum_{i,j=-1,i+j\neq n}^{\infty}\left[\frac{(2n-i)}{(n-i-j)}\alpha^0_{i-n,j}+\frac{(2n-j)}{(n-i-j)}\alpha^0_{i,j-n}\right]e_i\wedge e_j+\sum_{i=-1}^{\infty}\alpha^n_{i,n-i}e_i\wedge e_{n-i} \\
&= e_n\cdot\left(-\sum_{i,j=-1,i+j\neq 0}^{\infty}\frac{1}{(i+j)}\alpha^0_{ij}e_i\wedge e_j\right)+\alpha^n_{-1,n+1}e_{-1}\wedge e_{n+1} \\
&\quad +\alpha^n_{0,n}e_0\wedge e_n+\alpha^n_{1,n-1}e_1\wedge e_{n-1}
\end{aligned}$$

$$=e_n.\left(-\sum_{i,j=-1,i+j\neq 0}^{\infty}\frac{1}{(i+j)}\alpha_{ij}^0 e_i\wedge e_j+\frac{1}{2}\alpha_{0,1}^1(1+\delta_{n,\pm 1})e_{-1}\wedge e_1\right),\quad \text{for all } n\geqslant -1.$$

This shows that the general solution of (50) is a coboundary. The proof of Theorem V.1 is completed. □

With a different technique it has been obtained in Ref. 21 that the first cohomology of $\mathcal{G}_{0\infty}$ with coefficients in the ordinary tensor product is trivial. This result follows from Theorem V.1 of which it is a special case.

We now proceed with describing all bialgebra structures on the Lie algebra $\mathcal{G}_\infty$ of the group G_∞. The problem now is to describe all solutions of

$$(n-m)\alpha(e_{n+m})=e_n.\alpha(e_m)-e_m.\alpha(e_n),\qquad \text{for all } n,m\geqslant 0. \tag{66}$$

Equation (66) is now equivalent to the following infinite system of equations for the coalgebra structure constants α_{ij}^n:

$$(n-m)\alpha_{ij}^{n+m}=(2n-i)\alpha_{i-n,j}^m+(2n-j)\alpha_{i,j-n}^m-(2m-i)\alpha_{i-m,j}^n-(2m-j)\alpha_{i,j-m}^n, \tag{67}$$

where $n,m,i,j\geqslant 0$.

Theorem V.2: *All 1-cocycles on the Lie algebra $\mathcal{G}_\infty$ are coboundaries.*

Proof: The begining of the argument is similar to the one in the proof of Theorem V.1. Namely, the system of equations (67) is split into two completely independent systems. One for the structure constants α_{ij}^n with $n\neq i+j$, and one for the structure constants α_{ij}^n with $n=i+j$. In the first case, after setting $m=0$ in (67) we obtain

$$(n-i-j)\alpha_{ij}^n=(2n-i)\alpha_{i-n,j}^0+(2n-j)\alpha_{i,j-n}^0,\quad \text{for every}\quad n,i,j\geqslant 0.$$

Since $n\neq i+j$ we obtain that

$$\alpha_{ij}^n=\frac{(2n-i)}{(n-i-j)}\alpha_{i-n,j}^0+\frac{(2n-j)}{(n-i-j)}\alpha_{i,j-n}^0,$$

i.e., the coalgebra structure constants α_{ij}^n, where $n\neq i+j$, are expressed in terms of α_{ij}^0, $i,j\geqslant 0$, which are arbitrary.

We now turn to the second case, which requires a more detailed analysis. The system of equations for the coalgebra structure constants α_{ij}^n with $n=i+j$ is obtained from (67) by setting $j=n+m-i$:

$$(n-m)\alpha_{i,n+m-i}^{n+m}=(2n-i)\alpha_{i-n,m+n-i}^m+(n-m+i)\alpha_{i,m-i}^m$$
$$-(2m-i)\alpha_{i-m,n+m-i}^n-(m-n+i)\alpha_{i,n-i}^n, \tag{68}$$

where $n,m,i\geqslant 0$.

Again we split the rest of the proof into lemmas.

Lemma V.7: One has $\alpha_{0,n}^n=n\alpha_{0,1}^1$, for all $n\geqslant 0$.

Proof: After setting $m=1,i=0$ in (68) one obtains $\alpha_{0,n+1}^{n+1}=\alpha_{0,1}^1+\alpha_{0,n}^n$, since $n\alpha_{-n,n+1}^1=0$ for all $n\geqslant 0$. A simple inductive argument now leads to $\alpha_{0,n}^n=n\alpha_{0,1}^1$, for all $n\geqslant 0$. □

Lemma V.8: One has $\alpha_{1,n}^{n+1}=-(n+2)\alpha_{0,1}^1$ for all $n\geqslant 2$.

Proof: Set $m=1,i=1$ in (68). Then one has

$$(n-1)\alpha_{1,n}^{n+1}=(2n-1)\alpha_{1-n,n}^1-2n\alpha_{0,1}^1+(n-2)\alpha_{1,n-1}^n. \tag{69}$$

We investigate (69) for small values of n. For $n=0,1$ one obtains identities. For $n=2$ we obtain the first nontrivial relation. Namely, $\alpha_{1,2}^3=-4\alpha_{0,1}^1$. Clearly, since $\alpha_{1-n,n}^1=0$ for all $n\geqslant 2$, (69) is equivalent to

$$(n-1)\alpha_{1,n}^{n+1}=-2n\alpha_{0,1}^{1}+(n-2)\alpha_{1,n-1}^{n},$$

for all $n\geqslant 2$. Again, an induction argument leads to $\alpha_{1,n}^{n+1}=-(n+2)\alpha_{0,1}^{1}$, for all $n\geqslant 2$. □

Lemma V.9: One has $\alpha_{i,n}^{n+i}=-(n+2i)\alpha_{0,1}^{1}+[3(1-i^2)/(n-i)]\alpha_{0,1}^{1}$ *for all* $n\geqslant 1$, *and all* $i\geqslant 1$, *and* $n\neq i$.

Proof: Set $m=i$ in (68) to obtain

$$(n-i)\alpha_{i,n}^{n+i}=(2n-i)\alpha_{i-n,n}^{i}+n\alpha_{i,0}^{i}-i\alpha_{0,n}^{n}+(n-2i)\alpha_{i,n-i}^{n}. \tag{70}$$

After using Lemma V.7, (70) reduces to

$$(n-i)\alpha_{i,n}^{n+i}=-2ni\alpha_{0,1}^{1}+(n-2i)\alpha_{i,n-i}^{n}. \tag{71}$$

If we set $n=i$ in (71) it reduces further to $2i\alpha_{0,i}^{i}=2i^2\alpha_{0,1}^{1}$, from where for $i\neq 0$ it is equivalent to the assertion of Lemma V.7. Let $n\neq i$. Then the general solution of (71) is given by

$$\alpha_{i,n}^{n+i}=-(n+2i)\alpha_{0,1}^{1}+f(n,i), \tag{72}$$

where $f(n,i)$ satisfies the equation $(n-i)f(n,i)=(n-2i)f(n-i,i)$. If we define $\alpha(n,i):=(n-i)f(n,i)$ then this equation is equivalent to $\alpha(n,i)=\alpha(n-i,i)$. From here it follows that $\alpha(n,i)=\alpha(i)$, that is, α is independent of n. Therefore the general solution (72) is given by

$$\alpha_{i,n}^{n+i}=-(n+2i)\alpha_{0,1}^{1}+\frac{\alpha(i)}{(n-i)}. \tag{73}$$

Our next goal is to determine the parameters $\alpha(i)$, for all $i\geqslant 2$. From the conclusion of Lemma V.7 we know that $\alpha(1)=0$. Set $n=1$ in (73). This yields $\alpha_{i,1}^{i+1}=-(1+2i)\alpha_{0,1}^{1}+\alpha(i)/(1-i)$. Using the result of Lemma V.7, this leads to the equations $-(1+2i)\alpha_{0,1}^{1}+\alpha(i)/(1-i)=(i+2)\alpha_{0,1}^{1}$, for all $i\geqslant 2$, from where, solving for $\alpha(i)$, we obtain $\alpha(i)=3(1-i^2)\alpha_{0,1}^{1}$, for all $i\geqslant 2$. Therefore, $\alpha_{i,n}^{n+i}=-(n+2i)\alpha_{0,1}^{1}+[3(1-i^2)/(n-i)]\alpha_{0,1}^{1}$, for all $n\geqslant 1$, and all $i\geqslant 1$, and $n\neq i$. □

Quite similarly as in the case of the Witt algebra the following observation helps to complete the argument.

Lemma V.10: One has $\alpha_{0,1}^{1}=0$.

Proof: Set $m=2,i=1$ in (68) to obtain

$$(n-2)\alpha_{1,n+1}^{n+2}=(2n-1)\alpha_{1-n,n+1}^{2}+(n-3)\alpha_{1,n-1}^{n}. \tag{74}$$

We investigate (74) for small values of n. For $n=0$ it yields an identity. For $n=1$ it gives $-\alpha_{1,2}^{3}=\alpha_{0,2}^{2}-2\alpha_{1,0}^{1}=4\alpha_{0,1}^{1}$, a fact we learned from Lemma V.9. For $n=2$ it yields a trivial identity. For $n=3$ it gives $\alpha_{1,4}^{5}=0$. But from Lemma V.8 we have $\alpha_{1,4}^{5}=-6\alpha_{0,1}^{1}$, a clear contradiction. Therefore we conclude that $\alpha_{0,1}^{1}=0$, which is what we set out to show. □

The following two corollaries are direct consequences of the last result and the preceding lemmas.

Corollary V.6: One has $\alpha_{0,n}^{n}=0$, *for all* $n\geqslant 1$.

Corollary V.7: One has $\alpha_{i,n}^{n+i}=0$ *for all* $n,i\geqslant 1$.

Thus, all coalgebra structure constants $\alpha_{ij}^{n}=0$, whenever $n=i+j$. Therefore the general solution of (66) is described by

$$\alpha(e_n)=\sum_{i,j=0}^{\infty}\alpha_{ij}^n e_i\wedge e_j=\sum_{i,j=0,i+j\neq n}^{\infty}\alpha_{ij}^n e_i\wedge e_j$$

$$=\sum_{i,j=0,i+j\neq n}^{\infty}\left[\frac{(2n-i)}{(n-i-j)}\alpha_{i-n,j}^0+\frac{(2n-j)}{(n-i-j)}\alpha_{i,j-n}^0\right]e_i\wedge e_j$$

$$=e_n.\left(-\sum_{i,j=0,i+j\neq 0}^{\infty}\frac{1}{(i+j)}\alpha_{ij}^0 e_i\wedge e_j\right).$$

This shows that for $\mathcal{G}_\infty$ all 1-cocycles are coboundaries. □

VI. THE GROUP G_∞ AND THE R-MATRIX

In this section we describe the correspondence between the solution

$$\Omega(u,v;x)=\varphi(u,v)x'(u)x'(v)-\varphi(x(u),x(v)) \tag{75}$$

of the cocycle equation and the classical r-matrix on $\mathcal{G}_\infty$.[17] We prove that there is a one-to-one correspondence between the Poisson–Lie structures on G_∞ and the r-matrices on $\mathcal{G}_\infty$. This establishes the classification of Poisson–Lie (Lie-bialgebra) structures for the group G_∞ and its Lie algebra $\mathcal{G}_\infty$. In the next section we extend this result to the group $G_{0\infty}$ and its Lie algebra $\mathcal{G}_{0\infty}$. Let $e_n=u^{n+1}(d/du)$, $n=0,1,2,\ldots$, be the canonical basis of $\mathcal{G}_\infty$. We write the classical r-matrix (taking values in the completed tensor product) as

$$\lambda=\sum_{m,n=0}^{\infty}\lambda_{m+1,n+1}u^{m+1}\frac{d}{du}\wedge v^{n+1}\frac{d}{dv}=\varphi(u,v)\frac{d}{du}\wedge\frac{d}{dv}\in\bar{\wedge}^2\mathcal{G}_\infty,$$

where $\varphi(u,v)=\sum_{m,n=1}^{\infty}\lambda_{mn}u^m v^n$, and $\lambda_{mn}=-\lambda_{nm}$. If α is a 1-cocycle, then

$$(\delta\alpha)(e_l,e_m)=e_l.\alpha(e_m)-e_m\cdot\alpha(e_l)-\alpha([e_l,e_m])=0,$$

where δ is the coboundary operator in the Chevalley–Eilenberg cohomology of Lie algebras.

In Sec. V we proved that all 1-cocycles α are coboundaries. Thus, $\alpha=\delta\lambda$. Let us define $\langle\lambda,\lambda\rangle:=\Phi(u,v,w)(d/du)\wedge d/dv\wedge d/dw\in\bar{\wedge}^3\mathcal{G}$ as

$$\langle\lambda,\lambda\rangle:=[\lambda^{12},\lambda^{13}]+[\lambda^{12},\lambda^{23}]+[\lambda^{13},\lambda^{23}],$$

where

$$[\lambda^{12},\lambda^{13}]:=[\varphi(u,v)\partial_u\varphi(u,w)-\varphi(u,w)\partial_u\varphi(u,v)]\frac{d}{du}\wedge\frac{d}{dv}\wedge\frac{d}{dw},$$

$$[\lambda^{12},\lambda^{23}]:=[\varphi(u,v)\partial_v\varphi(v,w)-\varphi(v,w)\partial_v\varphi(u,v)]\frac{d}{du}\wedge\frac{d}{dv}\wedge\frac{d}{dw},$$

$$[\lambda^{13},\lambda^{23}]:=[\varphi(u,w)\partial_w\varphi(v,w)-\varphi(v,w)\partial_w\varphi(u,w)]\frac{d}{du}\wedge\frac{d}{dv}\wedge\frac{d}{dw}.$$

Therefore, we have

$$\langle\lambda,\lambda\rangle=\{\varphi(u,v)[\partial_u\varphi(u,w)+\partial_v\varphi(v,w)]+\text{cyclic}(u,v,w)\}\frac{d}{du}\wedge\frac{d}{dv}\wedge\frac{d}{dw}.$$

Then α satisfies the co-Jacobi identity, that is, α defines a Lie-bialgebra structure on $\mathcal{G}$, if and only if $\langle\lambda,\lambda\rangle$ is $\mathcal{G}_\infty$-invariant with respect to the adjoint action of $\mathcal{G}_\infty$ on itself.[1] Let $f(u)(d/du)\in\mathcal{G}_\infty$ be an arbitrary element, where $f(u)\in\mathbb{C}[[u]]$ is a formal power series. Then ad$\mathcal{G}_\infty$-invariance implies that $f(u)\frac{d}{du}.\langle\lambda,\lambda\rangle=0$. This is equivalent to the following equation:

$$f(u)\frac{\partial\Phi}{\partial u}+f(v)\frac{\partial\Phi}{\partial v}+f(w)\frac{\partial\Phi}{\partial w}=(f'(u)+f'(v)+f'(w))\Phi(u,v,w)\,. \tag{76}$$

Lemma VI.1: The subspace of ad$\mathcal{G}_\infty$-invariants in $\widetilde{\bigwedge}^3\mathcal{G}_\infty$ is empty.

Proof: Let $e_0=u(d/du)$. Then from (76) it follows that $e_0.\langle\lambda,\lambda\rangle=0$ is equivalent to

$$u\frac{\partial\Phi}{\partial u}+v\frac{\partial\Phi}{\partial v}+w\frac{\partial\Phi}{\partial w}=3\Phi(u,v,w). \tag{77}$$

Thus, $\Phi(u,v,w)$ is homogeneous of degree 3. But $\Phi(u,v,w)$ is antisymmetric with respect of all of its arguments. The only $\Phi(u,v,w)$ with both of the above properties is $\Phi(u,v,w)=0$. We conclude that all Lie-bialgebra structures on $\mathcal{G}_\infty$ are given by solutions of the classical Yang–Baxter equation. □

Remark VI.1: It is immediately clear that the above argument can be extended to the case of the Witt algebra $\mathcal{G}_{0\infty}$ with basis $u^{n+1}(d/du)$, where $n=-1,0,2,\ldots$, and the double-sided Witt algebra with basis $u^{n+1}(d/du)$, $n\in\mathbb{Z}$. Therefore, all bialgebra structures for these algebras are also obtained as solutions of the Classical Yang–Baxter equation.

Now, we turn our attention to the Poisson brackets on the group, written in local coordinates, by writing $\Omega(u,v;x)$ in components. In order to do this we shall need the formula

$$x(u)^n=\sum_{s_1=1}^{\infty}x_{s_1}u^{s_1}\ldots\sum_{s_n=1}^{\infty}x_{s_n}u^{s_n}=\sum_{s_1=1}^{\infty}\ldots\sum_{s_n=1}^{\infty}x_{s_1}\ldots x_{s_n}u^{s_1+\cdots+s_n}$$
$$=\sum_{i=n}^{\infty}\left(\sum_{(\Sigma_{k=1}^{n}s_k)=i}x_{s_1}\ldots x_{s_n}\right)u^i.$$

Then we have

$$\Omega(u,v;x)=\varphi(u,v)x'(u)x'(v)-\varphi(x(u),x(v))$$
$$=\sum_{p,q=1}^{\infty}\lambda_{pq}u^pv^q\sum_{i=1}^{\infty}ix_iu^{i-1}\sum_{j=1}^{\infty}jx_jv^{j-1}$$
$$-\sum_{p,q=1}^{\infty}\lambda_{pq}\sum_{i=p}^{\infty}\left(\sum_{(\Sigma_{k=1}^{p}r_k)=i}x_{r_1}\ldots x_{r_p}\right)u^i\sum_{j=q}^{\infty}\left(\sum_{(\Sigma_{l=1}^{q}s_l)=j}x_{s_1}\ldots x_{s_q}\right)v^j$$
$$=\sum_{p,q=1}^{\infty}\sum_{i=p}^{\infty}\sum_{j=q}^{\infty}\Bigg(\lambda_{i-p+1,j-q+1}px_pqx_q$$
$$-\lambda_{pq}\sum_{(\Sigma_{k=1}^{p}r_k)=i}x_{r_1}\ldots x_{r_p}\sum_{(\Sigma_{l=1}^{q}s_l)=j}x_{s_1}\ldots x_{s_q}\Bigg)u^iv^j$$
$$=\sum_{i,j=1}^{\infty}\Bigg[\sum_{p=1}^{i}\sum_{q=1}^{j}\Bigg(\lambda_{i-p+1,j-q+1}px_pqx_q$$

$$-\lambda_{pq}\sum_{(\Sigma_{k=1}^{p} r_k)=i} x_{r_1}\ldots x_{r_p}\sum_{(\Sigma_{l=1}^{q} s_l)=j} x_{s_1}\ldots x_{s_q}\Bigg)\Bigg] u^i v^j.$$

Therefore for $\{x_i, x_j\}=\omega_{ij}(x)$ we obtain

$$\omega_{ij}(x)=\sum_{p=1}^{i}\sum_{q=1}^{j} p x_p q x_q \lambda_{i-p+1,j-q+1} - \sum_{p=1}^{i}\sum_{q=1}^{j}\lambda_{pq}\left(\sum_{(\Sigma_{k=1}^{p} r_k)=i} x_{r_1}\ldots x_{r_p}\sum_{(\Sigma_{l=1}^{q} s_l)=j} x_{s_1}\ldots x_{s_q}\right). \tag{78}$$

Before we continue further, let us deduce the following useful formulas:

$$\frac{\partial}{\partial x_n}\sum_{(\Sigma_{k=1}^{p} r_k)=i} x_{r_1}\ldots x_{r_p}=\sum_{(\Sigma_{k=1}^{p} r_k)=i}\left(\sum_{l=1}^{p} x_{r_1}\ldots\delta_{r_l}^{n}\ldots x_{r_p}\right) =\sum_{l=1}^{p}\left(\sum_{(\Sigma_{k=1}^{p} r_k)=i} x_{r_1}\ldots\delta_{r_l}^{n}\ldots x_{r_p}\right) =p\sum_{(\Sigma_{k=1}^{p-1} r_k)=i-n} x_{r_1}\ldots x_{r_{p-1}},$$

as well as

$$\sum_{(\Sigma_{k=1}^{p-1} r_k)=i-n} x_{r_1}\ldots x_{r_{p-1}}\big|_e=\sum_{(\Sigma_{k=1}^{p-1} r_k)=i-n}\delta_{r_1}^{1}\ldots\delta_{r_{p-1}}^{1}=\delta_{i-n}^{p-1},$$

$$\left(\frac{\partial}{\partial x_n}\sum_{(\Sigma_{k=1}^{p} r_k)=i} x_{r_1}\ldots x_{r_p}\right)\Bigg|_e=p\,\delta_{i-n}^{p-1}. \tag{79}$$

Differentiating (78) with respect to x_n we obtain

$$\frac{\partial\omega_{ij}}{\partial x_n}\bigg|_x=\sum_{p=1}^{i}\sum_{q=1}^{j}(p\,\delta_p^n q x_q\lambda_{i-p+1,j-q+1}+p x_p\delta_q^n q\lambda_{i-p+1,j-q+1}) -\sum_{p=1}^{i}\sum_{q=1}^{j}\left[p\sum_{(\Sigma_{k=1}^{p-1} r_k)=i-n} x_{r_1}\ldots x_{r_{p-1}}\sum_{(\Sigma_{l=1}^{q} s_l)=j} x_{s_1}\ldots x_{s_q}\right]\lambda_{pq} -\sum_{p=1}^{i}\sum_{q=1}^{j}\left[q\sum_{(\Sigma_{k=1}^{p} r_k)=i} x_{r_1}\ldots x_{r_p}\sum_{(\Sigma_{l=1}^{q-1} s_k)=j-n} x_{s_1}\ldots x_{s_{q-1}}\right]\lambda_{pq}.$$

From the above formula we have (keeping in mind that $x_p|_e=\delta_p^1$)

$$\beta_{ij}^{n}:=\frac{\partial\omega_{ij}}{\partial x_n}\bigg|_e=\sum_{p=1}^{i} p\,\delta_p^n\lambda_{i-p+1,j}+\sum_{q=1}^{j} q\,\delta_q^n\lambda_{i,j-q+1}-\sum_{p=1}^{i}\sum_{q=1}^{j}\lambda_{pq}[p\,\delta_{i-n+1}^{p}\delta_j^q+q\,\delta_i^p\delta_{j-n+1}^{q}] =(2n-i-1)\lambda_{i-n+1,j}+(2n-j-1)\lambda_{i,j-n+1}. \tag{80}$$

Using this correspondence we can now prove the following theorem.

Theorem VI.1: *There is a one-to-one correspondence between the coboundary Lie-bialgebra structures on $\mathcal{G}_\infty$ given by λ and the Poisson–Lie structures of the type (75) on G_∞. Since all Lie bialgebra structures on $\mathcal{G}_\infty$ are given by λ (cf. Theorem V.1), Theorem IV.3 gives a classification of all solutions of the classical Yang–Baxter equation for $\mathcal{G}_\infty$.*

Proof: Recall that ω_{mn} satisfy the infinite system of functional equations (13):

$$\omega_{mn}(z)=\omega_{kl}(x)\frac{\partial z_m}{\partial x_k}\frac{\partial z_n}{\partial x_l}+\omega_{kl}(y)\frac{\partial z_m}{\partial y_k}\frac{\partial z_n}{\partial y_l}, \quad \text{where } x,y\in G_\infty, \tag{81}$$

and $z_n=z_n(x,y)$ is given by

$$z_n=\sum_{i=1}^{n} x_i \sum_{(\Sigma_{\alpha=1}^{i} j_\alpha)=n} y_{j_1}\dots y_{j_i}. \tag{82}$$

From (82) it follows that

$$\left.\frac{\partial z_i}{\partial y_k}\right|_{y=e}=(i-k+1)x_{i-k+1} \quad \text{and} \quad \left.\frac{\partial z_i}{\partial x_k}\right|_{y=e}=\delta_i^k.$$

Let us fix $n\in\mathbb{N}$ and consider a subsystem of the system of equations (81) for all ω_{ij} with $1\leqslant i<j\leqslant n$. After differentiating (81) with respect to y_j, for each j such that $1\leqslant j\leqslant n$, and setting $y=e$ we deduce that ω_{mn} satisfy the following inhomogeneous system of linear partial differential equations:

$$\sum_{i=j}^{n}(i+1-j)x_{i+1-j}\frac{\partial\omega_{mn}}{\partial x_i}=\omega_{m+1-j,n}(x)(m+1-j)+\omega_{m,n+1-j}(x)(n+1-j)$$
$$+\sum_{k=1}^{m}\sum_{l=1}^{n}\beta_{kl}^{j}(m+1-k)(n+1-l)x_{m+1-k}x_{n+1-l}, \tag{83}$$

for $1\leqslant j\leqslant n$, and where $\beta_{kl}^{j}=\partial\omega_{kl}/\partial y_j|_{y=e}$.

The method of proof is as follows. Let β_{kl}^{j} be given by (80). For each $n\in\mathbb{N}$ the general solution of (83) is a linear combination of the general solution of the homogeneous system of equations,

$$\sum_{i=j}^{n}(i+1-j)x_{i+1-j}\frac{\partial\omega_{mn}}{\partial x_i}=\omega_{m+1-j,n}(x)(m+1-j)+\omega_{m,n+1-j}(x)(n+1-j), \tag{84}$$

and a particular solution of the inhomogeneous system (83). We now show that for each $n\in\mathbb{N}$ and $1\leqslant m<n$ the system (83) has a unique solution by demonstrating that the only solution of the homogeneous system is the zero solution. Therefore, since every solution of the system of functional equations (81) is a solution of the system of partial differential equations (83) it follows that the class of solutions of (81) found in Sec. IV exhausts all its possible solutions. We shall prove this by induction applied in several steps. Recall that

$$\omega_{mn}(e)=0, \quad \text{for every } n,m\in\mathbb{N}, \tag{85}$$

and that $\omega_{mn}=0$ whenever $n<1$ or $m<1$.

(i) For $n=1$ there is nothing to prove. Let $n=2$. Then from (84) we obtain

$$x_1\frac{\partial\omega_{12}}{\partial x_1}=3\omega_{12}. \tag{86}$$

The general solution of this equation is $\omega_{12}(x)=Cx_1^3$, where C is an arbitrary constant. From (85) it follows that $C=0$. Therefore $\omega_{12}(x)=0$ is the only solution of (86). Let $n=3$. Then from (84) we obtain

$$x_1\frac{\partial\omega_{13}}{\partial x_1}+2x_2\frac{\partial\omega_{13}}{\partial x_2}=4\omega_{13},\qquad x_1\frac{\partial\omega_{13}}{\partial x_2}=0. \tag{87}$$

From (87) it follows that $\omega_{13}(x)=\omega_{13}(x_1)$. We deduce that $\omega_{13}(x)=Cx_1^5$, and from (85) it follows that $C=0$, and thus $\omega_{13}(x)=0$. Let us assume now that $\omega_{1k}(x)=0$ for $2\leqslant k\leqslant n-1$. From (84) we have

$$\sum_{i=j}^{n}(i+1-j)x_{i+1-j}\frac{\partial\omega_{1n}}{\partial x_i}=\omega_{2-j,n}(x)(2-j)+\omega_{1,n+1-j}(x)(n+1-j),\quad \text{for}\ \ 1\leqslant j\leqslant n,$$

which implies

$$\sum_{i=1}^{n}ix_i\frac{\partial\omega_{1n}}{\partial x_i}=(n+1)\omega_{1n}(x), \tag{88}$$

$$\sum_{i=j}^{n}(i+1-j)x_{i+1-j}\frac{\partial\omega_{1n}}{\partial x_i}=0,\quad \text{for}\ \ 2\leqslant j\leqslant n, \tag{89}$$

after using the induction hypothesis: $\omega_{1k}(x)=0$ for $2\leqslant k\leqslant n-1$. Then from (88) follows that $\omega_{1n}(x_1)$ satisfies

$$x_1\frac{\partial\omega_{1n}}{\partial x_1}=(n+1)\omega_{1n}. \tag{90}$$

From (90) we have that $\omega_{1n}(x)=Cx_1^{n+1}$ for an arbitrary constant C. Applying again (85) we conclude that $\omega_{1n}(x)=0$. Therefore $\omega_{1n}(x)=0$ for every $n\in\mathbb{N}$.

(ii) Let $m=2$ and $n=3$. Then from (84) we have the following homogeneous system of partial differential equations for ω_{23}:

$$x_1\frac{\partial\omega_{23}}{\partial x_1}+2x_2\frac{\partial\omega_{23}}{\partial x_2}+3x_3\frac{\partial\omega_{23}}{\partial x_3}=5\omega_{23},$$

$$x_1\frac{\partial\omega_{23}}{\partial x_2}+2x_2\frac{\partial\omega_{23}}{\partial x_3}=\omega_{13}=0,$$

$$x_1\frac{\partial\omega_{23}}{\partial x_3}=-\omega_{12}=0.$$

Arguing in a similar manner as above we obtain that $\omega_{23}(x)=0$. Let us assume that $\omega_{2k}(x)=0$ for all k such that $3\leqslant k\leqslant n-1$. We now prove that $\omega_{2n}(x)=0$. From (84) we have

$$\sum_{i=j}^{n}(i+1-j)x_{i+1-j}\frac{\partial\omega_{2n}}{\partial x_i}=\omega_{3-j,n}(x)(3-j)+\omega_{2,n+1-j}(x)(n+1-j),\quad \text{for}\ \ 1\leqslant j\leqslant n. \tag{91}$$

After using the induction hypothesis and the already proved fact that $\omega_{1n}=0$, for every $n\in\mathbb{N}$, (91) yields

$$\sum_{i=1}^{n} i x_i \frac{\partial \omega_{2n}}{\partial x_i} = (n+2)\,\omega_{2n}(x), \tag{92}$$

$$\sum_{i=j}^{n} (i+1-j) x_{i+1-j} \frac{\partial \omega_{2n}}{\partial x_i} = 0, \quad \text{for } 2 \leqslant j \leqslant n. \tag{93}$$

Therefore from (93) and (92) it follows that $\omega_{2n}(x) = \omega_{2n}(x_1) = C x_1^{n+2}$, and imposing (85) again we obtain that $\omega_{2n}(x_1) = 0$. Thus $\omega_{2n}(x) = 0$ for every $n \in \mathbb{N}$.

(iii) Assume now that $\omega_{sn} = 0$ for all s such that $1 \leqslant s \leqslant m-1$, for some $m \geqslant 2$ and all $n > s$. We will prove that $\omega_{mn} = 0$ for all $n \geqslant m$. Let $n = m+1$. From (84) we have

$$\sum_{i=j}^{m+1} (i+1-j) x_{i+1-j} \frac{\partial \omega_{m,m+1}}{\partial x_i} = \omega_{m+1-j,n}(x)(m+1-j) + \omega_{m,m+2-j}(x)(m+2-j),$$
$$\text{for } 1 \leqslant j \leqslant m+1. \tag{94}$$

We apply now the induction hypothesis and deduce from (94) the following system of equations:

$$\sum_{i=1}^{m+1} i x_i \frac{\partial \omega_{m,m+1}}{\partial x_i} = (2m+1)\,\omega_{m,m+1}(x), \tag{95}$$

$$\sum_{i=j}^{m+1} (i+1-j) x_{i+1-j} \frac{\partial \omega_{m,m+1}}{\partial x_i} = 0, \quad \text{for } 2 \leqslant j \leqslant m+1. \tag{96}$$

From (96) it follows that $\omega_{m,m+1}(x) = \omega_{m,m+1}(x_1)$, and from (95) we deduce that $\omega_{m,m+1}(x)$ must satisfy

$$x_1 \frac{\partial \omega_{m,m+1}}{\partial x_1} = (2m+1)\,\omega_{m,m+1}.$$

The solution of the above equation is $\omega_{m,m+1}(x) = C x_1^{2m+1}$, where C is an arbitrary constant. Then from $\omega_{m,m+1}(e) = C = 0$ we obtain that $\omega_{m,m+1}(x) = 0$. Finally, we assume that $\omega_{mk} = 0$ for all k such that $m+1 \leqslant k \leqslant n-1$, and we prove it for $k = n$. Indeed, from (84), after applying the induction hypothesis, we obtain

$$\sum_{i=1}^{n} i x_i \frac{\partial \omega_{mn}}{\partial x_i} = (m+n)\,\omega_{mn}(x), \tag{97}$$

$$\sum_{i=j}^{n} (i+1-j) x_{i+1-j} \frac{\partial \omega_{mn}}{\partial x_i} = 0, \quad \text{for } 2 \leqslant j \leqslant n. \tag{98}$$

Again, from (98) it follows that $\omega_{mn}(x) = \omega_{mn}(x_1)$, and that $\omega_{mn}(x)$ must satisfy

$$x_1 \frac{\partial \omega_{mn}}{\partial x_1} = (m+n)\,\omega_{mn}.$$

From here we conclude that $\omega_{mn}(x) = C x_1^{m+n}$ for an arbitrary constant C. But the requirement $\omega_{mn}(e) = 0$ fixes the value of this constant to be $C = 0$. Therefore $\omega_{mn}(x) = 0$.

Thus, we showed that for every $m, n \in \mathbb{N}$ the only solution of (84) is the zero solution. Therefore the system of partial differential equations (83) has a unique solution. The existence of the solution follows from the existence of the solution of the system of functional equations (81) of which (83) is a consequence. Thus, the structure constants β_{kl}^{j} of the Lie-bialgebra $\mathcal{G}_\infty$, as given by (80), determine uniquely all Poisson–Lie structures on the group G_∞. The proof of Theorem VI.1 is completed. □

We conclude this section by writing an explicit formula for the family of Lie-bialgebra structures arising from the family of Poisson–Lie structures obtained in Theorem IV.3, as well as the more general one-parameter family, obtained from it by an action of a formal diffeomorphism group element, for $d \geqslant 1$. An elegant way to do this is by deriving a formula for the Lie-bialgebra structures on $\mathcal{G}_\infty$ in terms of generating series and solutions of

$$\varphi(u,v)[\partial_u\varphi(w,u)+\partial_v\varphi(w,v)]+\text{cyclic}(u,v,w)=0.$$

Let us define $\mathcal{A}_n(u,v)$ as

$$\mathcal{A}_n(u,v) := \frac{\partial}{\partial x_n}\Omega(u,v;\mathcal{X})\bigg|_e = \sum_{i,j=1}^{\infty} \alpha_{ij}^n u^i v^j.$$

Then we have the following lemma.

Lemma IV.2: The generating series $\mathcal{A}_n(u,v)$ is given by

$$\mathcal{A}_n(u,v)=n\varphi(u,v)(u^{n-1}+v^{n-1})-[u^n\partial_u\varphi(u,v)+v^n\partial_v\varphi(u,v)].$$

Proof: We use formula (75) and the following facts. If $x(u)=\Sigma_{i=1}^{\infty}x_iu^i$ then $\mathcal{X}'(u)|_x=\Sigma_{i=1}^{\infty}ix_iu^{i-1}\Rightarrow\mathcal{X}'(u)|_e=1$, and also $(\partial/\partial x_n)\mathcal{X}'(u)|_x=\Sigma_{i=1}^{\infty}i\,\delta_i^n u^{i-1}=nu^{n-1}$. From this it follows that

$$\frac{\partial}{\partial x_n}[\mathcal{X}'(u)\mathcal{X}'(v)]|_x=nu^{n-1}x'(v)+nv^{n-1}x'(u)\Rightarrow\frac{\partial}{\partial x_n}[\mathcal{X}'(u)\mathcal{X}'(v)]|_e=n(u^{n-1}+v^{n-1}).$$

Finally we have

$$\begin{aligned}\frac{\partial}{\partial x_n}\varphi(\mathcal{X}(u),\mathcal{X}(v))\bigg|_e &= \frac{\partial\mathcal{X}(u)}{\partial x_n}\partial_1\varphi(\mathcal{X}(u),\mathcal{X}(v))\bigg|_e+\frac{\partial\mathcal{X}(v)}{\partial x_n}\partial_2\varphi(\mathcal{X}(u),\mathcal{X}(v))\bigg|_e\\ &= u^n\partial_u\varphi(u,v)+v^n\partial_v\varphi(u,v).\end{aligned}$$

In the above equality we used that $\partial\mathcal{X}(u)/\partial x_n|_e=u^n$. □

Proposition VI.1: (Michaelis,[22] Taft[23]): For each $d\in\mathbb{N}$ the family of Poisson–Lie structures (75) given by $\varphi_d(u,v)=uv(v^d-u^d)$ gives rise to the following family of Lie-bialgebra structures on $\mathcal{G}_\infty$:

$$\alpha(e_n)=2ne_d\wedge e_n-2(n-d)e_0\wedge e_{d+n}\quad(n\geqslant 0),$$

where $\{e_n=u^{n+1}(d/du)\}_{n\in\mathbb{Z}_+}$ is a canonical basis for $\mathcal{G}_\infty$.

Proof: The generating series $\mathcal{A}_{n,d}$ in this case is

$$\begin{aligned}\mathcal{A}_{n,d}(u,v)&=n[uv^{d+1}-vu^{d+1}](u^{n-1}+v^{n-1})-\{u^n[v^{d+1}-(d+1)vu^d]+v^n[(d+1)uv^d-u^{d+1}]\}\\ &=(n-1)u^nv^{d+1}-(n-1)v^nu^{d+1}+(n-d-1)uv^{n+d}-(n-d-1)vu^{n+d}\\ &=\{(n-1)[\delta_i^n\delta_j^{d+1}-\delta_j^n\delta_i^{d+1}]+(n-d-1)[\delta_i^1\delta_j^{d+n}-\delta_j^1\delta_i^{d+n}]\}u^iv^j\\ &=\sum_{i,j=1}^{\infty}\alpha_{ij|d}^n u^iv^j,\end{aligned}$$

where

$$\alpha^n_{ij|d}=\{(n-1)[\delta^n_i\delta^{d+1}_j-\delta^n_j\delta^{d+1}_i]+(n-d-1)[\delta^1_i\delta^{d+n}_j-\delta^1_j\delta^{d+n}_i]\}.$$

Therefore

$$\begin{aligned}\alpha_d(e_n)=\alpha^n_{ij|d}e_i\wedge e_j&=(n-1)[e_n\wedge e_{d+1}-e_{d+1}\wedge e_n]+(n-d-1)[e_1\wedge e_{d+n}-e_{d+n}\wedge e_1]\\&=2(n-1)e_n\wedge e_{d+1}+2(n-d-1)e_1\wedge e_{n+d},\end{aligned}$$

and after shifting indices by 1 we obtain

$$\alpha_d(e_n)=-2ne_d\wedge e_n+2(n-d)e_0\wedge e_{n+d},\quad \text{for every}\quad n\in\mathbb{Z}_+.$$

This concludes the proof. □

Proposition VI.2: For every $d\geqslant 2$ the family of Poisson–Lie structures described by (49) gives rise to the following family of Lie-bialgebra structures on $\mathcal{G}_\infty$:

$$\begin{aligned}\alpha_{d,\lambda}(e_n)=&\,2\sum_{i=d+n}^{\infty}(2n-i)\lambda^{i-(n+d)}e_0\wedge e_i-2n\sum_{i=d}^{\infty}\lambda^{i-d}e_i\wedge e_n\\&+\frac{2}{d-1}\sum_{i=d+n}^{\infty}\sum_{j=1}^{d-1}(2n-i)\lambda^{i+j-(n+d)}e_i\wedge e_j\\&+\frac{2}{d-1}\sum_{i=d}^{\infty}\sum_{j=n+1}^{d+n-1}(2n-j)\lambda^{i+j-(n+d)}e_i\wedge e_j,\end{aligned}\tag{99}$$

for every $n\in\mathbb{Z}_+$.

Proof: Let $\alpha_{d,\lambda}(e_n)=\alpha^n_{ij|d}e_i\wedge e_j$, where $\{e_n\}_{n\geqslant 1}$ is a basis of $\mathcal{G}_\infty$, and

$$\alpha^n_{ij|d}=(2n-i-1)\lambda_{i-n+1,j}+(2n-j-1)\lambda_{i,j-n+1},\quad \text{for every}\ \ n,i,j\in\mathbb{N}.$$

With the assumptions of Lemma IV.3 and Theorem VI.5 we have

$$\lambda_{ij}=\frac{1}{\lambda_{1,d+1}}[\lambda_{1i}\lambda_{d+1,j}-\lambda_{1j}\lambda_{d+1,i}],$$

where

$$\begin{aligned}&\lambda_{n,d+1}=0,\quad \text{for every}\ \ n\geqslant d+1,\\&\lambda_{n,d+1}=-\frac{1}{d-1}\frac{(\lambda_{1,d+2})^{n-1}}{(\lambda_{1,d+1})^{n-2}}=-\frac{1}{d-1}\lambda_{1,d+1}\lambda^{n-1},\ \ \text{for}\ \ 2\leqslant n\leqslant d,\\&\lambda_{1n}=\frac{(\lambda_{1,d+2})^{n-d-1}}{(\lambda_{1,d+1})^{n-d-2}}=\lambda_{1,d+1}\lambda^{n-d-1},\quad \text{for every}\ \ n\geqslant d+1,\end{aligned}\tag{100}$$

and where we have introduced $\lambda:=\lambda_{1,d+2}/\lambda_{1,d+1}$. Then

$$\begin{aligned}\alpha_{d,\lambda}(e_n)&=[(2n-i-1)\lambda_{i-n+1,j}+(2n-j-1)\lambda_{i,j-n+1}]e_i\wedge e_j\\&=\frac{1}{\lambda_{1,d+1}}\{(2n-i-1)[\lambda_{1,i-n+1}\lambda_{d+1,j}-\lambda_{1j}\lambda_{d+1,i-n+1}]\\&\quad+(2n-j-1)[\lambda_{1i}\lambda_{d+1,j-n+1}-\lambda_{1,j-n+1}\lambda_{d+1,i}]\}e_i\wedge e_j\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda_{1,d+1}} \left\{ \sum_{i=d+n}^{\infty} \sum_{j=1}^{d} (2n-i-1)\lambda_{1,i-n+1}\lambda_{d+1,j} e_i \wedge e_j \right. \\
&\quad - \sum_{i=n}^{d+n-1} \sum_{j=d+1}^{\infty} (2n-i-1)\lambda_{1j}\lambda_{d+1,i-n+1} e_i \wedge e_j \\
&\quad + \sum_{i=d+1}^{\infty} \sum_{j=n}^{d+n-1} (2n-j-1)\lambda_{1i}\lambda_{d+1,j-n+1} e_i \wedge e_j \\
&\quad \left. - \sum_{i=1}^{d} \sum_{j=d+n}^{\infty} (2n-j-1)\lambda_{1,j-n+1}\lambda_{d+1,i} e_i \wedge e_j \right\} \\
&= \frac{1}{\lambda_{1,d+1}} \left\{ \sum_{i=d+n}^{\infty} \sum_{j=2}^{d} (2n-i-1)\lambda_{1,i-n+1}\lambda_{d+1,j} e_i \wedge e_j \right. \\
&\quad + \lambda_{d+1,1} \sum_{i=d+n}^{\infty} (2n-i-1)\lambda_{1,i-n+1} e_i \wedge e_1 \\
&\quad - \sum_{i=n+1}^{d+n-1} \sum_{j=d+1}^{\infty} (2n-i-1)\lambda_{1j}\lambda_{d+1,i-n+1} e_i \wedge e_j \\
&\quad - \lambda_{d+1,1} \sum_{j=d+1}^{\infty} (n-1)\lambda_{1j} e_n \wedge e_j + \sum_{i=d+1}^{\infty} \sum_{j=n+1}^{d+n-1} (2n-j-1)\lambda_{1i}\lambda_{d+1,j-n+1} e_i \wedge e_j \\
&\quad + \lambda_{d+1,1} \sum_{i=d+1}^{\infty} (n-1)\lambda_{1i} e_i \wedge e_n - \sum_{i=2}^{d} \sum_{j=d+n}^{\infty} (2n-j-1)\lambda_{1,j-n+1}\lambda_{d+1,i} e_i \wedge e_j \\
&\quad \left. - \lambda_{d+1,1} \sum_{j=d+n}^{\infty} (2n-j-1)\lambda_{1,j-n+1} e_1 \wedge e_j \right\} \\
&= \frac{1}{\lambda_{1,d+1}} \left\{ 2 \sum_{i=d+n}^{\infty} \sum_{j=2}^{d} (2n-i-1)\lambda_{1,i-n+1}\lambda_{d+1,j} e_i \wedge e_j \right. \\
&\quad + 2 \sum_{i=d+1}^{\infty} \sum_{j=n+1}^{d+n-1} (2n-j-1)\lambda_{1i}\lambda_{d+1,j-n+1} e_i \wedge e_j \\
&\quad \left. + 2\lambda_{1,d+1} \sum_{i=d+n}^{\infty} (2n-i-1)\lambda_{1,i-n+1} e_1 \wedge e_i - 2\lambda_{1,d+1} \sum_{i=d+1}^{\infty} (n-1)\lambda_{1i} e_i \wedge e_n \right\} \\
&= \lambda_{1,d+1} \left\{ \frac{2}{d-1} \sum_{i=d+n}^{\infty} \sum_{j=2}^{d} (2n-i-1)\lambda^{i+j-(n+d+1)} e_i \wedge e_j \right. \\
&\quad + \frac{2}{d-1} \sum_{i=d+1}^{\infty} \sum_{j=n+1}^{d+n-1} (2n-j-1)\lambda^{i+j-(n+d+1)} e_i \wedge e_j \\
&\quad \left. + 2 \sum_{i=d+n}^{\infty} (2n-i-1)\lambda^{i-(n+d)} e_1 \wedge e_i - 2 \sum_{i=d+1}^{\infty} (n-1)\lambda^{i-(d+1)} e_i \wedge e_n \right\},
\end{aligned}
$$

where we used formulas (100) to obtain the last equality. Hence, after normalizing by the quotient $\lambda_{1,d+1} \neq 0$ and shifting indices by 1 we obtain (99). □

Remark VI.2: One can show directly that $\alpha_{d,\lambda}$ satisfies the co-Jacobi identity. The r.h.s. of (99) is an element of the completed tensor product $\mathcal{G}_\infty \hat{\otimes} \mathcal{G}_\infty$.[24]

VII. THE GROUP $G_{0\infty}$ AND POISSON–LIE STRUCTURES ON IT

In this section we study the group $G_{0\infty}$ of which G_∞ is a subgroup. We classify all Poisson–Lie structures on $G_{0\infty}$.

Let $X=\{x_i\}_{i\in\mathbb{Z}_+}$ be a countable set of indeterminates. Let $k[[X]]$ be the ring of formal power series over X without a constant term with the standard multiplication. Here k is a commutative field assumed to be of characteristic zero. Let $Y=\{y_i\}_{i\in\mathbb{Z}_+}$ be a second set of indeterminates, and $k[[Y]]$ be the corresponding ring of formal power series over Y. Consider the formal group $G_{0\infty}$ defined by a formal group law $F=(F_i)_{i\in\mathbb{Z}_+}$[25,24] in a countably infinite number of variables, where $F_i\in k[[X,Y]]$ for every $i\in\mathbb{Z}_+$, induced by a substitution of formal power series in one variable. Let $x(u)=\Sigma_{i=0}^{\infty}x_iu^i\in k[[X]][[u]]$ and $y(u)=\Sigma_{i=0}^{\infty}y_iu^i\in k[[Y]][[u]]$ be elements in the rings of formal power series with a constant term in the variable u over the rings $k[[X]]$ and $k[[Y]]$, respectively. The multiplication of formal power series in the variable u is defined again as the substitution:

$$\begin{aligned}(xy)(u)=x(y(u))=\sum_{i=0}^{\infty} x_i(y(u))^i&=\sum_{i=0}^{\infty} x_i\left[y_0^i+\sum_{j=1}^{\infty}\left(\sum_{(\Sigma_{\alpha=1}^{i}s_\alpha)=j} y_{s_1}\dots y_{s_i}\right)u^j\right]\\ &=\sum_{i=0}^{\infty} x_iy_0^i+\sum_{i=1}^{\infty} x_i\sum_{j=1}^{\infty}\left(\sum_{(\Sigma_{\alpha=1}^{i}s_\alpha)=j} y_{s_1}\dots y_{s_i}\right)u^j\\ &=\sum_{i=0}^{\infty} x_iy_0^i+\sum_{j=1}^{\infty}\left(\sum_{i=1}^{\infty} x_i\sum_{(\Sigma_{\alpha=1}^{i}s_\alpha)=j} y_{s_1}\dots y_{s_i}\right)u^j. \quad (101)\end{aligned}$$

Therefore from (101) we obtain

$$F_0(X,Y)=\sum_{i=0}^{\infty} x_iy_0^i, \quad (102)$$

$$F_j(X,Y)=\sum_{i=1}^{\infty} x_i\sum_{(\Sigma_{\alpha=1}^{i}s_\alpha)=j} y_{s_1}\dots y_{s_i},\quad \text{for every } j\geq 1. \quad (103)$$

This is a model of the group of diffeomorphisms of $\mathbb{R}^1$ not necessarily leaving the point $u=0$ fixed. The identity here is $e=(0,1,0,0,\dots)$. Formulas (102) and (103) have the following interpretation. The ring $k[[X]]$ is naturally graded. Namely, let us introduce a degree $|\,|:X\rightarrow\mathbb{Z}_+$ defined on the generators by $|x_i|:=i$. We extend it to monomials as $|x_{i_1}\dots x_{i_n}|=i_1+\dots+i_n$. The grading on $k[[X]]$ and $k[[Y]]$ induces a grading on $k[[X,Y]]$ in an obvious way. Then $F_i(X,Y)=\Sigma_n f_n(X,Y)$, where each $f_n(X,Y)$ is a finite linear combination of monomials of degree n. Clearly G_∞, if viewed as a formal group, is a subgroup of $G_{0\infty}$. We define a Poisson structure $\omega_x=\omega_{ij}(x)\partial/\partial x_i\wedge\partial/\partial x_j$ on the group $G_{0\infty}$ as a bi-derivation $\omega_x:k[[X]]\otimes k[[X]]\rightarrow k[[X]]$, where $\omega_{ij}(x)=-\omega_{ji}(x)\in k[[X]]$, satisfying the Jacobi identity. The methods developed in analyzing the Poisson–Lie structures on G_∞ apply without major changes to the case of $G_{0\infty}$, but with two important differences. Namely, Theorem IV.1 still holds with $\Omega(u,v;\mathcal{X})$ defined as $\Omega(u,v;\mathcal{X}):=\Sigma_{i,j=0}^{\infty}\omega_{ij}u^iv^j$, but in the solution of the cocycle equation,

$$\Omega(u,v;x)=\varphi(u,v)x'(u)x'(v)-\varphi(x(u),x(v)), \quad (104)$$

$\varphi(u,v)$ does not have to be divisible by uv. Thus, this condition is dropped. This change affects the analysis of the space of solutions of the equation

$$\varphi(u,v)[\partial_u\varphi(w,u)+\partial_v\varphi(w,v)]+\text{cyclic}(u,v,w)=0. \tag{105}$$

Namely, the general solution of (105) is given by

$$\varphi(u,v)=\frac{F(v)-F(u)}{F'(u)F'(v)}, \tag{106}$$

where

$$F(u)=\sum_{n=-d}^{\infty} C_n u^n = C_{-d}\frac{1}{u^d}+C_{-d+1}\frac{1}{u^{d-1}}+\cdots,$$

is a formal Laurent series and $d>0$ or $d<0$. Since on the other hand $F'(u)=1/f(u)$, where $f(u)\in\mathbb{R}[[u]]$ is a formal power series, the only negative value of d can be $d=-1$.

We found that the solution (104) corresponds to a cocycle $\mathcal{G}_{0\infty}\to\mathcal{G}_{0\infty}\hat{\wedge}\mathcal{G}_{0\infty}$ that is a coboundary. We showed that all cocycles on $\mathcal{G}_{0\infty}$ are coboundaries. This allows us to complete their classification.

Theorem VII.1: *The moduli space of solutions of (105) is parametrized by $\mathbb{N}\cup\{-1\}$, that is, the space of isomorphism classes of solutions of (105) under the action of the formal group of diffeomorphisms of the line is isomorphic to the countable set $\mathbb{N}\cup\{-1\}$.*

For $d=-1$ formula (42) yields $\varphi(u,v)=u-v$. This solution yields in coordinates the Poisson–Lie tensor,

$$\omega_{ij}(x)=i(j+1)x_i x_{j+1}-(i+1)jx_{i+1}x_j-x_i\delta_j^0+x_j\delta_i^0,\quad i,j\in\mathbb{Z}_+. \tag{107}$$

Proposition VII.1: The functions
(i) $\varphi(u,v)=u-v$, and
(ii) $\varphi(u,v)=e^{\lambda u}-e^{\lambda v}$, where λ is an arbitrary parameter,
are the only solutions of (105) of the form $\varphi(u,v)=a(u)-a(v)$.

The proof is straightforward. The Poisson–Lie tensor that corresponds to (ii) is

$$\omega_{ij}(x)=(j+1)x_{j+1}\sum_{p=0}^{i+1}\frac{px_p}{(i-p+1)!}-(i+1)x_{i+1}\sum_{q=0}^{j+1}\frac{qx_q}{(j-q+1)!}$$
$$-\delta_j^0\sum_{p=0}^{i}\frac{1}{p!}\sum_{r_1+\cdots+r_p=i}x_{r_1}\ldots x_{r_p}+\delta_i^0\sum_{q=0}^{j}\frac{1}{q!}\sum_{r_1+\cdots+r_q=j}x_{r_1}\ldots x_{r_q}.$$

We conclude this section with a useful reformulation of Theorem VII.1.[19]

Theorem VII.2: *All solutions of (105) fall into the following two classes:*
(a) The first class is given by Theorem IV.4.
(b) The second class is given by

$$\varphi(u,v)=\frac{1}{\lambda_{01}}[f(u)g(v)-f(v)g(u)], \tag{108}$$

where $\lambda_{01}\neq 0$, for any formal power series $f(u)$ and $g(u)$ satisfying the relation

$$f'(u)g(u)-f(u)g'(u)=\lambda_{01}g(u)-2\lambda_{02}f(u). \tag{109}$$

Here, λ_{01} *and* λ_{02} *are arbitrary parameters with* λ_{01} *being subject to the above restriction:* $\lambda_{01} \neq 0$.

VIII. ELEMENTS OF REPRESENTATION THEORY

Let G be a Poisson–Lie group and $\overline{\omega}$ be a Poisson–Lie structure on G. Let V be a space on which G acts, that is, there is a map $G\times V\rightarrow V$. Such a space is called a G-space. Assume that V is equipped with a Poisson structure ω. Recall the following definition.[26]

Definition VIII.1: The action of G *on* V *is called Poisson if the map* $G\times V\rightarrow V$ *is Poisson. Here* $G\times V$ *is equipped with the product Poisson structure.*

In this section we study the following problem. Suppose that we are given the Poisson–Lie group G_∞. Consider the space $V_\lambda=\{x(u)(du)^\lambda | x(u)=\Sigma_{i=0}^\infty x_i u^i\}$, $\lambda\in\mathbb{R}$. The space V_λ is the space of λ-densities (Jacobians) over the real line. The group G_∞ acts naturally on V_λ. Let $y\in G_\infty$ and $x(u)(du)^\lambda\in V_\lambda$. Then the action of G_∞ on V_λ is defined by

$$x(u)(du)^\lambda \mapsto x(y(u))(y'(u))^\lambda (du)^\lambda,$$

where $y(u)=\Sigma_{i=1}^\infty y_i u^i$, and

$$\begin{aligned}(y'(u))^\lambda &= \left(\sum_{i=1}^\infty i y_i u^{i-1}\right)^\lambda = \left(y_1+\sum_{i=2}^\infty i y_i u^{i-1}\right)^\lambda \\ &= y_1^\lambda\left(1+\sum_{i=2}^\infty i\frac{y_i}{y_1}u^{i-1}\right)^\lambda \\ &= y_1^\lambda\left[1+\frac{\lambda}{1!}\sum_{i=2}^\infty i\frac{y_i}{y_1}u^{i-1}+\frac{\lambda(\lambda-1)}{2!}\left(\sum_{i=2}^\infty i\frac{y_i}{y_1}u^{i-1}\right)^2+\ldots\right].\end{aligned}$$

Are there Poisson structures on the space V_λ such that the above action of G_∞ on V_λ is a Poisson action? In other words, is there a Poisson strucuture ω on V_λ such that the map $G_\infty\times V_\lambda\rightarrow V_\lambda$ is a Poisson map? Here again $G_\infty\times V_\lambda$ is equipped with the product Poisson structure.

Let $y(u)=\Sigma_{i=1}^\infty y_i u^i\in G_\infty$, and $x(u)(du)^\lambda\in V_\lambda$. Let us define $z_\lambda(u)\coloneqq x(y(u))[y'(u)]^\lambda=\Sigma_{i=0}^\infty z_i u^i$, where $z_i=z_i(x,y;\lambda)$ are the coordinates of z_λ. If we also introduce the notation $J(u)\coloneqq y'(u)=\Sigma_{i=1}^\infty i y_i u^{i-1}$, we have

$$\begin{aligned}x(u)(du)^\lambda \mapsto x(y(u))(y'(u))^\lambda(du)^\lambda &= x(y(u))(J(u))^\lambda(du)^\lambda \\ &= z_\lambda(u)(du)^\lambda=\sum_{i=0}^\infty x_i(y(u))^i\left(\sum_{i=1}^\infty i y_i u^{i-1}\right)^\lambda (du)^\lambda.\end{aligned}$$

Defining $z(u)\coloneqq x(y(u))$ and using the definition $z_\lambda(u)=x(y(u))[J(u)]^\lambda$ we deduce that $z_\lambda'(u)=z'(u)(J(u))^\lambda+z(u)[(J(u))^\lambda]'$, where $'$ stands for the derivative with respect to u.

An argument analogous to the argument given in Sec. I implies that the map $G_\infty\times V_\lambda\rightarrow V_\lambda$ is Poisson if and only if

$$\omega_{ij}(z)=\omega_{kl}(x)\frac{\partial z_i}{\partial x_k}\frac{\partial z_j}{\partial x_l}+\overline{\omega}_{kl}(y)\frac{\partial z_i}{\partial y_k}\frac{\partial z_j}{\partial y_l}. \tag{110}$$

Here $\omega_{ij}(x)=\{x_i,x_j\}$ and $\overline{\omega}_{ij}(y)=\{y_i,y_j\}$, where $\{x_i\}_{i\in\mathbb{Z}_+}$ and $\{y_i\}_{i\in\mathbb{Z}_+}$ are the coordinates on V_λ and G_∞, respectively. Also, let us introduce in a manner similar to the one used in Sec. IV a generating series for the Poisson strucutures on V_λ as $\Omega(u,v;\mathcal{X})\coloneqq\Sigma_{i,j=0}^\infty \omega_{ij}u^i v^j$.

Lemma VIII.1: The multiplicativity condition (110) is equivalent to the following functional equation:

$$\Omega(u,v;z_\lambda)=\Omega(y(u),y(v);x)(J(u))^\lambda(J(v))^\lambda+\overline{\Omega}(u,v;y)z'(u)(J(u))^{\lambda-1}z'(v)(J(v))^{\lambda-1}$$
$$+\lambda\ \partial_u\overline{\Omega}(u,v;y)z(u)(J(u))^{\lambda-1}z'(v)(J(v))^{\lambda-1}$$
$$+\lambda\ \partial_v\overline{\Omega}(u,v;y)z'(u)(J(u))^{\lambda-1}z(v)(J(v))^{\lambda-1}$$
$$+\lambda^2\ \partial^2_{u,v}\overline{\Omega}(u,v;y)z(u)(J(u))^{\lambda-1}z(v)(J(v))^{\lambda-1}. \tag{111}$$

Here $\overline{\Omega}$ stands for the generating series of the Poisson–Lie strucutures on G_∞.

Proof: Multiplying both sides of Eq. (110) by u^iv^j, summing over i and j, and using the definition of Ω we obtain

$$\Omega(u,v;z_\lambda)=\omega_{kl}(x)\frac{\partial z_\lambda}{\partial x_k}\frac{\partial z_\lambda}{\partial x_l}+\overline{\omega}_{kl}(y)\frac{\partial z_\lambda}{\partial y_k}\frac{\partial z_\lambda}{\partial y_l}. \tag{112}$$

On the other hand, the following formulas are valid:

$$\frac{\partial z_\lambda}{\partial x_k}=\left(\sum_{j=1}^{\infty} y_ju^j\right)^k\left(\sum_{i=1}^{\infty} iy_iu^{i-1}\right)^\lambda=[y(u)]^k[J(u)]^\lambda,$$

$$\frac{\partial z_\lambda}{\partial y_k}=\left[\sum_{i=0}^{\infty} ix_i\left(\sum_{j=1}^{\infty} y_ju^j\right)^{i-1}u^k\right]\left(\sum_{i=1}^{\infty} iy_iu^{i-1}\right)^\lambda$$
$$+\left[\sum_{i=0}^{\infty} x_i\left(\sum_{j=1}^{\infty} y_ju^j\right)^{i}\right]\lambda\left(\sum_{i=1}^{\infty} iy_iu^{i-1}\right)^{\lambda-1}ku^{k-1}$$
$$=x'(y(u))(J(u))^\lambda u^k+\lambda x(y(u))[J(u)]^{\lambda-1}ku^{k-1}$$
$$=z'(u)[J(u)]^{\lambda-1}u^k+\lambda z_\lambda(u)[J(u)]^{-1}ku^{k-1}.$$

Therefore Eq. (112) takes the form

$$\Omega(u,v;z_\lambda)=\Omega(y(u),y(v);x)(J(u))^\lambda(J(v))^\lambda+\overline{\omega}_{kl}(y)[z'(u)(J(u))^{\lambda-1}u^k$$
$$+\lambda z(u)(J(u))^{\lambda-1}ku^{k-1}]$$
$$\times[z'(v)(J(v))^{\lambda-1}v^l+\lambda z(v)(J(v))^{\lambda-1}lv^{l-1}]$$

$$=\Omega(y(u),y(v);x)(J(u))^\lambda(J(v))^\lambda$$
$$+\overline{\omega}_{kl}(y)z'(u)(J(u))^{\lambda-1}u^kz'(v)(J(v))^{\lambda-1}v^l$$
$$+\overline{\omega}_{kl}(y)\lambda z(u)(J(u))^{\lambda-1}ku^{k-1}z'(v)(J(v))^{\lambda-1}v^l$$
$$+\overline{\omega}_{kl}(y)z'(u)(J(u))^{\lambda-1}u^k\lambda z(v)(J(v))^{\lambda-1}lv^{l-1}$$
$$+\lambda^2\overline{\omega}_{kl}(y)z(u)(J(u))^{\lambda-1}ku^{k-1}z(v)(J(v))^{\lambda-1}lv^{l-1}$$

$$=\Omega(y(u),y(v);x)(J(u))^\lambda(J(v))^\lambda$$
$$+\overline{\Omega}(u,v;y)z'(u)(J(u))^{\lambda-1}z'(v)(J(v))^{\lambda-1}$$
$$+\lambda\partial_u\overline{\Omega}(u,v;y)z(u)(J(u))^{\lambda-1}z'(v)(J(v))^{\lambda-1}$$
$$+\lambda\partial_v\overline{\Omega}(u,v;y)z'(u)(J(u))^{\lambda-1}z(v)(J(v))^{\lambda-1}$$

$$+\lambda^2\partial^2_{u,v}\overline{\Omega}(u,v;y)z(u)(J(u))^{\lambda-1}z(v)(J(v))^{\lambda-1}.$$

This concludes the proof of the Lemma. □

In the above formulas $\overline{\Omega}$ is given by (cf. Sec. IV)

$$\overline{\Omega}(u,v;y)=\varphi(u,v)y'(u)y'(v)-\varphi(y(u),y(v)), \tag{113}$$

where $\varphi(u,v)$ satisfies the equation

$$\varphi(u,v)[\partial_u\varphi(w,u)+\partial_v\varphi(w,v)]+\text{cyclic}(u,v,w)=0. \tag{114}$$

In other words, $\varphi(u,v)$ is given by $\varphi(u,v)=f(u)g(v)-f(v)g(u)$ (cf. Sec. IV), where the functions f and g satisfy the relation

$$f'(u)g(u)-f(u)g'(u)=\alpha f'(u)+\beta g'(u)(\Rightarrow f''(u)g(u)-f(u)g''(u)=\alpha f''(u)+\beta g''(u)). \tag{115}$$

Here, α and β are arbitrary constants.

Theorem VIII.1: *If $\varphi(u,v)$ satisfies Eq. (114) and $\overline{\Omega}$ is defined by (113) one has the following solution of (111):*

$$\Omega(u,v;x)=\varphi(u,v)x'(u)x'(v)+\lambda\ \partial_u\varphi(u,v)x(u)x'(v)+\lambda\ \partial_v\varphi(u,v)x'(u)x(v)$$
$$+\lambda^2\ \partial^2_{u,v}\varphi(u,v)x(u)x(v). \tag{116}$$

Proof: Using formula (113) in the r.h.s. of Eq. (111) we obtain

$$\begin{aligned}\text{r.h.s.}=&\Omega(y(u),y(v);x)(J(u))^{\lambda}(J(v))^{\lambda}+\varphi(u,v)z'(u)z'(v)(J(u))^{\lambda}(J(v))^{\lambda}\\&-\varphi(y(u),y(v))z'(u)z'(v)(J(u))^{\lambda-1}(J(v))^{\lambda-1}+\lambda\ \partial_u\varphi(u,v)z_\lambda(u)z'(v)(J(v))^{\lambda}\\&+\varphi(u,v)z(u)z'(v)[(J(u))^{\lambda}]'(J(v))^{\lambda}-\lambda\ \partial_1\varphi(y(u),y(v))z_\lambda(u)z'(v)(J(v))^{\lambda-1}\\&+\lambda\ \partial_v\varphi(u,v)z_\lambda(v)z'(u)(J(u))^{\lambda}+\varphi(u,v)z'(u)z(v)[(J(v))^{\lambda}]'(J(u))^{\lambda}\\&-\lambda\ \partial_2\varphi(y(u),y(v))z_\lambda(v)z'(u)(J(u))^{\lambda-1}+\lambda^2\ \partial^2_{u,v}\varphi(u,v)z_\lambda(u)z_\lambda(v)\\&+\lambda\ \partial_u\varphi(u,v)z_\lambda(u)z(v)[(J(v))^{\lambda}]'+\lambda\ \partial_v\varphi(u,v)z_\lambda(v)z(u)[(J(u))^{\lambda}]'\\&+\varphi(u,v)z(u)z(v)[(J(u))^{\lambda}]'[(J(v))^{\lambda}]'-\lambda^2\ \partial^2_{1,2}\varphi(y(u),y(v))z_\lambda(u)z_\lambda(v).\end{aligned}$$

Above we used the formulas

$$\partial_u\overline{\Omega}(u,v;y)=\partial_u\varphi(u,v)J(u)J(v)+\varphi(u,v)J'(u)J(v)-\partial_1\varphi(y(u),y(v))J(u),$$

$$\partial_v\overline{\Omega}(u,v;y)=\partial_v\varphi(u,v)J(u)J(v)+\varphi(u,v)J(u)J'(v)-\partial_2\varphi(y(u),y(v))J(v).$$

For the l.h.s. of Eq. (111) we have

$$\begin{aligned}\text{l.h.s.}=&\varphi(u,v)z'_\lambda(u)z'_\lambda(v)+\lambda\ \partial_u\varphi(u,v)z_\lambda(u)z'_\lambda(v)+\lambda\ \partial_v\varphi(u,v)z'_\lambda(u)z_\lambda(v)\\&+\lambda^2\ \partial^2_{u,v}\varphi(u,v)z_\lambda(u)z_\lambda(v)\\=&\varphi(u,v)\{z'(u)(J(u))^{\lambda}+z(u)[(J(u))^{\lambda}]'\}\{z'(v)(J(v))^{\lambda}+z(v)[(J(v))^{\lambda}]'\}\\&+\text{remaining terms}\\=&\varphi(u,v)z'(u)z'(v)(J(u))^{\lambda}(J(v))^{\lambda}+\varphi(u,v)z'(u)z(v)(J(u))^{\lambda}[(J(v))^{\lambda}]'\end{aligned}$$

$$+\varphi(u,v)z(u)z'(v)(J(v))^{\lambda}[(J(u))^{\lambda}]'+\varphi(u,v)z(u)z(v)[(J(u))^{\lambda}]'[(J(v))^{\lambda}]'$$
$$+\lambda\ \partial_u\varphi(u,v)z_{\lambda}(u)z'(v)(J(v))^{\lambda}+\lambda\ \partial_u\varphi(u,v)z_{\lambda}(u)z(v)[(J(v))^{\lambda}]'$$
$$+\lambda\ \partial_v\varphi(u,v)z_{\lambda}(v)z'(u)(J(u))^{\lambda}+\lambda\ \partial_v\varphi(u,v)z_{\lambda}(v)z(u)[(J(u))^{\lambda}]'$$
$$+\lambda^2\ \partial^2_{u,v}\varphi(u,v)z_{\lambda}(u)z_{\lambda}(v).$$

On the other hand, we also have

$$\begin{aligned}\Omega(y(u),y(v);x)(J(u))^{\lambda}(J(v))^{\lambda}=&\varphi(y(u),y(v))z'(u)z'(v)(J(u))^{\lambda-1}(J(v))^{\lambda-1}\\&+\lambda\ \partial_1\varphi(y(u),y(v))z_{\lambda}(u)z'(v)(J(v))^{\lambda-1}\\&+\lambda\ \partial_2\varphi(y(u),y(v))z_{\lambda}(v)z'(u)(J(u))^{\lambda-1}\\&+\lambda^2\ \partial^2_{1,2}\varphi(y(u),y(v))z_{\lambda}(u)z_{\lambda}(v).\end{aligned}$$

After comparing the terms on the l.h.s. and the r.h.s. of Eq. (111) we obtain an identity. This concludes the proof of the Theorem. □

Remark VIII.1: Notice that for $\lambda\neq 0$ *we can not have an inhomogeneous term of the form* $\overline{\varphi}(x(u),x(v))$ *in (116). Had it been the case it would impose on* $\overline{\varphi}(u,v)$ *the condition of being a homogeneous function of degree 1 in both arguments. In order to satisfy the equation (111)* $\overline{\varphi}(u,v)$ *must have the property* $\overline{\varphi}(z_{\lambda}(u),z_{\lambda}(v))=\overline{\varphi}(z(u)(J(u))^{\lambda},z(v)(J(v))^{\lambda})=\overline{\varphi}(z(u),z(v))\times((J(u))^{\lambda}(J(v))^{\lambda})$. *But since* $\overline{\varphi}(u,v)$ *must be also antisymmetric it follows that the only function with these properties is* $\overline{\varphi}=0$. *On the contrary, for* $\lambda=0$ *we have a solution of (111) of the form*

$$\Omega(u,v;x)=\varphi(u,v)x'(u)x'(v)-\overline{\varphi}(x(u),x(v)),$$

where the function $\overline{\varphi}$ *has the property* $\overline{\varphi}(u,v)=-\overline{\varphi}(v,u)$. *The Jacobi identity for* Ω *then implies that* $\overline{\varphi}$ *must satisfy (114).*

Next, we come to the following fact.

Theorem VIII.2: *If* $\varphi(u,v)$ *satisfies (114) then the solution (116) satisfies the Jacobi identity, thus yielding a class of Poisson structures on* V_{λ} *for which the action of* G_{∞} *on* V_{λ} *is a Poisson action.*

Proof: We shall use again the bracket $\{\mathcal{X}(u),\mathcal{X}(v)\}:=\Sigma_{i,j=0}^{\infty}\{\mathcal{X}_i,\mathcal{X}_j\}u^iv^j$ introduced in Sec. IV. Then we have

$$\begin{aligned}\{\mathcal{X}(u),\mathcal{X}(v)\}=&\varphi(u,v)\mathcal{X}'(u)\mathcal{X}'(v)+\lambda\ \partial_u\varphi(u,v)\mathcal{X}(u)\mathcal{X}'(v)\\&+\lambda\ \partial_v\varphi(u,v)\mathcal{X}'(u)\mathcal{X}(v)+\lambda^2\ \partial^2_{u,v}\varphi(u,v)\mathcal{X}(u)\mathcal{X}(v),\end{aligned}$$

as well as

$$\begin{aligned}\partial_u\{\mathcal{X}(u),\mathcal{X}(v)\}=&\partial_v\varphi(u,v)\mathcal{X}''(u)\mathcal{X}'(v)+\varphi(u,v)\mathcal{X}'(u)\mathcal{X}''(v)+\lambda^2\ \partial^2_{u,v}\varphi(u,v)\mathcal{X}(u)\mathcal{X}'(v)\\&+\lambda\ \partial_u\varphi(u,v)\mathcal{X}(u)\mathcal{X}''(v)+\lambda\ \partial^2_v\varphi(u,v)\mathcal{X}'(u)\mathcal{X}(v)+\lambda\ \partial_v\varphi(u,v)\mathcal{X}'(u)\mathcal{X}'(v)\\&+\lambda^2\ \partial^3_{u,v^2}\varphi(u,v)\mathcal{X}(u)\mathcal{X}(v)+\lambda^2\ \partial^2_{u,v}\varphi(u,v)\mathcal{X}(u)\mathcal{X}'(v).\end{aligned}$$

The Jacobi identity $\{\mathcal{X}(w),\{\mathcal{X}(u),\mathcal{X}(v)\}\}+\text{cyclic}(u,v,w)=0$ is equivalent to the following equation: $\mathcal{A}+\mathcal{B}+\mathcal{C}+\mathcal{D}=0$. Here,

$$\mathcal{A}=\varphi(u,v)\{\mathcal{X}(w),\mathcal{X}'(u)\mathcal{X}'(v)\}+\text{cyclic}(u,v,w)$$
$$=\varphi(u,v)[\partial_u\{\mathcal{X}(w),\mathcal{X}(u)\}\mathcal{X}'(v)+\partial_v\{\mathcal{X}(w),\mathcal{X}(v)\}\mathcal{X}'(u)]+\text{cyclic}(u,v,w),$$

$$\mathcal{B}=\lambda\ \partial_u\varphi(u,v)\{\mathcal{X}(w),\mathcal{X}(u)\mathcal{X}'(v)\}+\text{cyclic}(u,v,w)$$
$$=\lambda\ \partial_u\varphi(u,v)[\{\mathcal{X}(w),\mathcal{X}(u)\}\mathcal{X}'(v)+\partial_v\{\mathcal{X}(w),\mathcal{X}(v)\}\mathcal{X}(u)]+\text{cyclic}(u,v,w),$$

$$\mathcal{C}=\lambda\ \partial_v\varphi(u,v)\{\mathcal{X}(w),\mathcal{X}'(u)\mathcal{X}(v)\}+\text{cyclic}(u,v,w)$$
$$=\lambda\ \partial_v\varphi(u,v)[\partial_u\{\mathcal{X}(w),\mathcal{X}(u)\}\mathcal{X}(v)+\{\mathcal{X}(w),\mathcal{X}(v)\}\mathcal{X}'(u)]+\text{cyclic}(u,v,w),$$

$$\mathcal{D}=\lambda^2\ \partial^2_{u,v}\varphi(u,v)\{\mathcal{X}(w),\mathcal{X}(u)\mathcal{X}(v)\}+\text{cyclic}(u,v,w)$$
$$=\lambda^2\ \partial^2_{u,v}\varphi(u,v)[\{\mathcal{X}(w),\mathcal{X}(u)\}\mathcal{X}(v)+\{\mathcal{X}(w),\mathcal{X}(v)\}\mathcal{X}(u)]+\text{cyclic}(u,v,w).$$

For the expressions in the square brackets for each term we therefore obtain

$$\begin{aligned}\mathcal{A}'=&\partial_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}'(v)+\varphi(w,u)\mathcal{X}'(w)\mathcal{X}''(u)\mathcal{X}'(v)\\
&+\lambda\ \partial^2_{w,u}\varphi(w,u)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)+\lambda\ \partial_w\varphi(w,u)\mathcal{X}(w)\mathcal{X}''(u)\mathcal{X}'(v)\\
&+\lambda\ \partial^2_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}(u)\mathcal{X}'(v)+\lambda\ \partial_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}'(v)\\
&+\lambda^2\ \partial^3_{w,u^2}\varphi(w,u)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}'(v)+\lambda^2\ \partial^2_{w,u}\varphi(w,u)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)\\
&+\partial_v\varphi(w,v)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}'(v)+\varphi(w,v)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}''(v)\\
&+\lambda\ \partial^2_{w,v}\varphi(w,v)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)+\lambda\ \partial_w\varphi(w,v)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}''(v)\\
&+\lambda\ \partial^2_v\varphi(w,v)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}(v)+\lambda\ \partial_v\varphi(w,v)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}'(v)\\
&+\lambda^2\ \partial^3_{w,v^2}\varphi(w,v)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}(v)+\lambda^2\ \partial^2_{w,v}\varphi(w,v)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v),\end{aligned}$$

$$\begin{aligned}\mathcal{B}'=&\varphi(w,u)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}'(v)+\lambda\ \partial_w\varphi(w,u)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)\\
&+\lambda\ \partial_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}(u)\mathcal{X}'(v)+\lambda^2\ \partial^2_{w,u}\varphi(w,u)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}'(v)\\
&+\partial_v\varphi(w,v)\mathcal{X}'(w)\mathcal{X}(u)\mathcal{X}'(v)+\varphi(w,v)\mathcal{X}'(w)\mathcal{X}(u)\mathcal{X}''(v)\\
&+\lambda\ \partial^2_{w,v}\varphi(w,v)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}'(v)+\lambda\ \partial_w\varphi(w,v)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}''(v)\\
&+\lambda\ \partial^2_v\varphi(w,v)\mathcal{X}'(w)\mathcal{X}(u)\mathcal{X}(v)+\lambda\ \partial_v\varphi(w,v)\mathcal{X}'(w)\mathcal{X}(u)\mathcal{X}'(v)\\
&+\lambda^2\ \partial^3_{w,v^2}\varphi(w,v)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v)+\lambda^2\ \partial^2_{w,v}\varphi(w,v)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}'(v),\end{aligned}$$

$$\begin{aligned}\mathcal{C}'=&\partial_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}(v)+\varphi(w,u)\mathcal{X}'(w)\mathcal{X}''(u)\mathcal{X}(v)+\lambda\ \partial^2_{w,u}\varphi(w,u)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}(v)\\
&+\lambda\ \partial_w\varphi(w,u)\mathcal{X}(w)\mathcal{X}''(u)\mathcal{X}(v)+\lambda\ \partial^2_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}(u)\mathcal{X}(v)\\
&+\lambda\ \partial_u\varphi(w,u)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}(v)+\lambda^2\ \partial^3_{w,u^2}\varphi(w,u)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v)\\
&+\lambda^2\ \partial^2_{w,u}\varphi(w,u)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}(v)+\varphi(w,v)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}'(v)\\
&+\lambda\ \partial_w\varphi(w,v)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)+\lambda\ \partial^2_{w,v}\varphi(w,v)\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}(v)\\
&+\lambda\ \partial_v\varphi(w,v)\mathcal{X}'(w)\mathcal{X}'(u)\mathcal{X}(v),\end{aligned}$$

$$\mathcal{D}'=\varphi(w,u)\mathcal{X}''(w)\mathcal{X}''(u)\mathcal{X}(v)+\lambda\ \partial_w\varphi(w,u)\mathcal{X}(w)\mathcal{X}''(u)\mathcal{X}(v)+\lambda\ \partial_u\varphi(w,u)\mathcal{X}''(w)\mathcal{X}(u)\mathcal{X}(v)$$
$$+\lambda^2\ \partial^2_{w,u}\varphi(w,u)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v)+\varphi(w,v)\mathcal{X}''(w)\mathcal{X}(u)\mathcal{X}''(v)$$
$$+\lambda\ \partial_w\varphi(w,v)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}''(v)+\lambda\ \partial_v\varphi(w,v)\mathcal{X}''(w)\mathcal{X}(u)\mathcal{X}(v)$$
$$+\lambda^2\ \partial^2_{w,v}\varphi(w,u)\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v).$$

In what follows, we shall study the four terms in the Jacobi identity written as

$$\mathcal{A}=\varphi(u,v)\mathcal{A}'+\text{cyclic}(u,v,w),\quad \mathcal{B}=\lambda\partial_u\varphi(u,v)\mathcal{B}'+\text{cyclic}(u,v,w),$$

$$\mathcal{C}=\lambda\partial_v\varphi(u,v)\mathcal{C}'+\text{cyclic}(u,v,w),\quad \mathcal{D}=\lambda^2\partial^2_{u,v}\varphi(u,v)\mathcal{D}'+\text{cyclic}(u,v,w).$$

We split the analysis of the Jacobi identity into seven steps (A)–(G).

(A) Terms proportional to $\mathcal{X}''\mathcal{X}''\mathcal{X}''$. From them we obtain (after cyclicly permuting the arguments of some of them):

$$\{\varphi(u,v)[\partial_u\varphi(w,u)+\partial_v\varphi(w,v)]+\text{cyclic}(u,v,w)\}\mathcal{X}''(w)\mathcal{X}''(u)\mathcal{X}''(v).$$

But from (114) it follows that the above term is zero.

(B) Terms proportional to $\mathcal{X}\mathcal{X}''\mathcal{X}''$ cancel each other out after cyclic permutation.

(C) Terms proportional to $\mathcal{X}\mathcal{X}''\mathcal{X}$ cancel each other out after cyclic permutation.

(D) Terms proportional to $\mathcal{X}\mathcal{X}\mathcal{X}$ give

$$\lambda^3[\partial_u\varphi(u,v)\partial^3_{w,v^2}\varphi(w,v)+\partial_v\varphi(u,v)\partial^3_{w,u^2}\varphi(w,u)+\text{cyclic}(u,v,w)]\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v)$$
$$+\lambda^4[\partial^2_{u,v}\varphi(u,v)(\partial^2_{w,u}\varphi(w,u)+\partial^2_{w,v}\varphi(w,v))+\text{cyclic}(u,v,w)]\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v).$$

Since $\varphi(u,v)$ is a solution of (114) the results obtained in Sec. IV showed that $\varphi(u,v)=f(u)g(v)-f(v)g(u)$, where f and g satisfy (115). Therefore the λ^3-term becomes

$$\lambda^3\{[f'(u)g(v)-f(v)g'(u)][f'(w)g''(v)-f''(v)g'(w)]+[f(u)g'(v)-f'(v)g(u)]$$
$$\times[f'(w)g''(u)-f''(u)g'(w)]+\text{cyclic}(u,v,w)\}\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v),$$

and the λ^4-term assumes the form

$$\lambda^4\{[f'(u)g'(v)-f'(v)g'(u)][f'(w)g'(u)-f'(u)g'(w)$$
$$+f'(w)g'(v)-f'(v)g'(w)]+\text{cyclic}(u,v,w)\}\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}(v).$$

Using the identities $f'(u)g(u)-f(u)g'(u)=\alpha f'(u)+\beta g'(u)(\Rightarrow f''(u)g(u)-f(u)g''(u)=\alpha f''(u)+\beta g''(u))$ one easily shows that both terms proportional to λ^3 and λ^4, respectively, are identically zero.

(E) Terms proportional to $\mathcal{X}\mathcal{X}\mathcal{X}''$ give

$$\lambda^2[\varphi(u,v)\partial^3_{w,u^2}\varphi(w,u)+\varphi(v,w)\partial^3_{u,w^2}\varphi(u,w)+\partial_u\varphi(u,v)\partial^2_{w,v}\varphi(w,v)+\partial_w\varphi(w,u)\partial^2_u\varphi(v,u)$$
$$+\partial_u\varphi(w,u)\partial^2_w\varphi(v,w)+\partial_w\varphi(v,w)\partial^2_{u,v}\varphi(u,v)]\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}''(v)+\text{cyclic}(u,v,w).$$

The expression in the square brackets becomes identically zero after using (115) in a similar way as in (D). Also we have a term proportional to λ^3 which reads as

$$\lambda^3[\partial_u\varphi(u,v)\partial^2_{w,u}\varphi(w,u)+\partial_u\varphi(u,v)\partial^2_{w,v}\varphi(w,v)+\partial_w\varphi(v,w)\partial^2_{u,v}\varphi(u,v)$$
$$+\partial_w\varphi(v,w)\partial^2_{u,w}\varphi(u,w)+\partial_w\varphi(v,w)\partial^2_{w,u}\varphi(w,u)+\partial_u\varphi(u,v)\partial^2_{v,w}\varphi(v,w)$$
$$+\partial_w\varphi(w,v)\partial^2_{u,v}\varphi(u,v)+\partial_u\varphi(v,u)\partial^2_{w,u}\varphi(w,u)]\mathcal{X}(w)\mathcal{X}(u)\mathcal{X}'(v)+\text{cyclic}(u,v,w),$$

which is identically zero.

(F) Terms proportional to $\mathcal{X}\mathcal{X}'\mathcal{X}'$ give

$$\lambda[\varphi(u,v)\partial^2_{w,u}\varphi(w,u)+\varphi(w,u)\partial^2_{w}\varphi(v,w)+\varphi(u,v)\partial^2_{w,v}\varphi(w,v)+\varphi(v,w)\partial^2_{w}\varphi(u,w)$$
$$+\partial_w\varphi(w,u)\partial_u\varphi(v,u)+\partial_w\varphi(v,w)\partial_v\varphi(u,v)]\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)+\text{cyclic}(u,v,w).$$

The expression in the square brackets can be shown to be identically zero after using (115) in a similar way as in (D) and (E).

The other two terms of the same form are

$$\lambda^2[\varphi(u,v)\partial^2_{w,u}\varphi(w,u)+\varphi(u,v)\partial^2_{w,v}\varphi(w,v)$$
$$+\varphi(v,u)\partial^2_{w,u}\varphi(w,u)+\varphi(u,v)\partial^2_{w,v}\varphi(v,w)]\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)+\text{cyclic}(u,v,w),$$

which is identically zero, and

$$\lambda^2[\partial_w\varphi(w,u)\partial_u\varphi(u,v)+\partial_w\varphi(w,u)\partial_w\varphi(v,w)+\partial_w\varphi(w,u)\partial_u\varphi(v,u)+\partial_w\varphi(v,w)\partial_v\varphi(u,v)$$
$$+\partial_v\varphi(u,v)\partial_w\varphi(w,v)+\partial_w\varphi(v,w)\partial_w\varphi(u,w)]\mathcal{X}(w)\mathcal{X}'(u)\mathcal{X}'(v)+\text{cyclic}(u,v,w),$$

which is again identically zero.

(G) Terms proportional to $\mathcal{X}'\mathcal{X}''\mathcal{X}'$ cancel each other.

Thus all terms have been covered and the proof of Theorem VIII.2 is completed. □

As a result of Theorem VIII.1 and Theorem VIII.2 for each Poisson–Lie structure on G_∞ defined by a function φ satisfying the equation (114) there exists a Poisson structure on V_λ for which the action of G_∞ is Poisson. Thus we obtain a series of representations $\mathcal{V}_{\varphi,\lambda}$ of the Poisson–Lie group G_∞ in the Poisson spaces V_λ.

IX. QUANTIZATION

This section is devoted to the quantization of some of the Poisson–Lie structures on the group G_∞ when restricted to the finite-dimensional quotient-groups $G_n=G_\infty \bmod u^{n+1}$. We shall construct explicitly families of finitely generated quantum (semi)groups. Their quasi-classical limits are the finite-dimensional Poisson–Lie groups endowed with Poisson–Lie structures, which are restrictions of the Poisson–Lie structures on the group G_∞, and belong to the countable family obtained in Theorem IV.3. We shall consider quotient-groups G_n, for $n\geqslant 4$ (cf. Sec. III). In this aproach to quantization we shall start from the quasi-classical limits, that is, the corresponding Poisson–Lie groups, and reconstruct from this data their quantum counterparts. This quantization procedure amounts to deformation quantization of the Poisson algebra $\mathcal{A}$ of C^∞ functions on the corresponding finite-dimensional Poisson–Lie groups to a noncommutative noncocommutative bialgebra $\mathcal{A}_h$ such that $\mathcal{A}_h/h\mathcal{A}_h\cong\mathcal{A}$. The second postulate of quantization requires that the deformation be ‘‘flat,’’ that is, the dimension of $\mathcal{A}_h$ as a $k[[h]]$-module, for a field k of characteristic zero, be the same as the dimension of $\mathcal{A}$ as a k-module. For the general philosophy underlying the method we refer the reader to Refs. 1 and 27.

(i) Let $X=\{x_i\}_{i\in\mathbb{N}}$ be a set of indeterminates. Let us introduce a grading on the algebra $k[X]$, where k denotes the ground field (assumed to be of characteristic zero), by assigning a degree (denoted $||$) to each of the generators x_i of $k[X]$ by the following definition: $|x_i|=i-1$, for every $i\in\mathbb{N}$. We extend it on monomials by $|AB|=|A|+|B|$, for every two monomials A,B.

As we have mentioned above, in this section we shall be concerned with the quantization problem for the countable family of Poisson–Lie structures on the group G_∞, found in Sec. IV, the formulas for which we now recall:

$$\{x_i,x_j\}=(i-d)jx_jx_{i-d}-i(j-d)x_ix_{j-d}+x_i\sum_{(\Sigma_{k=1}^{d+1}s_k)=j}x_{s_1}\ldots x_{s_{d+1}}-x_j\sum_{(\Sigma_{k=1}^{d+1}s_k)=i}x_{s_1}\ldots x_{s_{d+1}}, \tag{117}$$

for all $d\in\mathbb{N}$. It is clear from the right-hand side of (117) that for each $d\in\mathbb{N}$ the degree of the bracket $\{x_i,x_j\}$ is given by

$$|\{x_i,x_j\}|=|x_i|+|x_j|-d=i+j-d-2.$$

(ii) Let $X=\{x_i\}_{i\in\mathbb{N}}$ be a set of indeterminates and let $\langle X\rangle$ be a free associative semigroup with identity on X. Let $k[[h]]\langle X\rangle$ be the semigroup algebra of $\langle X\rangle$ over the ring of formal power series $k[[h]]$ in the parameter h. Here k is assumed to be a field of characteristic zero. Consider the set of relations

$$\mathcal{R}_h=\{x_ix_j-x_jx_i=f_{ij|h}(x)|i,j\in\mathbb{N}\},$$

where $f_{ij|h}(x)$ are polynomials in x_i with coefficients in $k[[h]]$, such that $f_{ij|0}=0$. Let $\mathcal{I}_h$ be the ideal generated by $\mathcal{R}_h$. Define $\mathcal{A}_h:=k[[h]]\langle X\rangle/\mathcal{I}_h$, and define a grading on $k\langle X\rangle$ as was done in (i).

The following is the first postulate of quantization. As explained earlier we require that $\mathcal{A}_h/h\mathcal{A}_h\cong k\langle X\rangle$, that is,

$$[x_i,x_j]=h\{x_i,x_j\}+O(h^2).$$

Here $[x_i,x_j]=x_ix_j-x_jx_i$, and the product is the product in $\mathcal{A}_h$. In other words, we would like to recover the Poisson–Lie bracket on G_∞ (or the quotient groups $G_n=G_\infty \bmod u^{n+1}$) in the quasi-classical limit $h\to 0$. This also means that for $h\to 0$ we should have

$$f_{ij|h}(x)=h\{x_i,x_j\}+O(h^2).$$

After computing the degree of the right-hand side of the above equality $|h\{x_i,x_j\}|=|h|+|\{x_i,x_j\}|=|h|+i+j-d-2$, we deduce that for each $d\in\mathbb{N}$ the parameter h must have degree $|h|=d$, since $|[x_i,x_j]|=i+j-2$.

(iii) Consider the semigroup algebra $k[[h]]\langle X\rangle$ of $\langle X\rangle$ over the ring $k[[h]]$. For each $d\in\mathbb{N}$ consider the set of relations

$$\mathcal{R}_h^d=\{x_ix_j=x_jx_i+f_{ij|h}^d(x)|i<j,\quad \text{for } i,j\in\mathbb{N}\}, \tag{118}$$

where $f_{ij|h}^d(x)\in k[[h]]\langle X\rangle$ are linear combinations of monomials $h^nx_{i_1}^{n_1}\ldots x_{i_k}^{n_k}$ such that $i_1>\ldots>i_k$ and $nd+\Sigma_{s=1}^k n_s(i_s-1)=i+j-d-2$. We shall say that these monomials are in canonical form. Thus $f_{ij|h}^d(x)$ are linear combinations of monomials in canonical form. Recall the following definition.

Definition IX.1: The semigroup algebra $k[[h]]\langle X\rangle$ has a Poincaré–Birkhoff–Witt (PBW) property if every monomial $A\in\langle X\rangle$ can be reduced to a unique expression as a linear combination of monomials in canonical form using the set of relations (118) independently of the choice of a reduction procedure.

We shall use a version of the result known as the Diamond Lemma,[28,29] applicable here, which allows us to prove the PBW property, by only proving it for the monomials with so called ‘‘overlap’’ ambiguities.[28]

Definition IX.2: For each $d \in \mathbb{N}$ *a quantum semigroup* $G^d_{\infty|h}$ *is defined as follows. As a ''quantum space''* $G^d_{\infty|h}$ *is defined by its quotient semigroup algebra* $\mathcal{A}^d_h := k[[h]]\langle X\rangle/\mathcal{I}^d_h$*, where* $\mathcal{I}^d_h \subset k[[h]]\langle X\rangle$ *is the ideal generated by the set of relations (118), and we require that in the quasi-classical limit* $[x_i, x_j] = h\{x_i, x_j\} + O(h^2)$*, one obtains the Poisson algebra of functions on the Poisson-Lie group* G^d_∞ *defined by (117). The multiplication of formal power series in one variable under the operation of substitution* $(xy)(u) = x(y(u))$*, where* $x(u) = \Sigma_{i=1}^\infty x_i u^i$*, and* $y(u) = \Sigma_{i=1}^\infty y_i u^i$*, induces a comultiplication map* $\Delta: \mathcal{A}^d_h \to \mathcal{A}^d_h \otimes \mathcal{A}^d_h$*, which is defined on the generators by*

$$\Delta(x_n) = \sum_{i=1}^n x_i \otimes \sum_{\Sigma_{\alpha=1}^i j_\alpha = n} x_{j_1} \ldots x_{j_i}, \quad n \in \mathbb{N}, \tag{119}$$

and is required to be an algebra homomorphism. Also, one defines a counit map $c: \mathcal{A}^d_h \to k[[h]]$ *by*

$$c(x_i) = \delta_i^1, \quad i \in \mathbb{N}. \tag{120}$$

All tensor products are over $k[[h]]$*. This endows* $\mathcal{A}^d_h$ *with a structure of a bialgebra and the quantum semigroup* $G^d_{\infty|h}$ *is defined to be the bialgebra* $\mathcal{A}^d_h$*, if, in addition,* $k[[h]]\langle X\rangle$ *has the PBW basis described above.*

Does such an object exist? We do not know yet. However, if we consider the Poisson–Lie quotient groups $G^d_n = G^d_\infty \bmod u^{n+1}$, for $n \leqslant 7$, then there exist quantum objects that satisfy the above definition for small values of d, $d \leqslant 5$, and whose quasi-classical limits are the Poisson–Lie groups G^d_n. The definition of the finitely generated quantum semigroups $G^d_{n|h}$ is the same as the definition above with $\mathcal{I}^d_h$ being an ideal generated by a finite set of relations $\mathcal{R}^d_h$. Their defining semigroup algebras turn out to have other interesting properties. We shall describe in more detail the construction of the quantum semigroup $G^2_{5|h}$, while omitting parts of the construction that consist of lengthy and tedious calculations, and we shall state the results for the quantum semigroups $G^1_{4|h}$ and $G^3_{5|h}$ without entering into the details of the calculations. The construction of the last two mimics exactly the construction of $G^2_{5|h}$.

Let $G^2_5 = G^2_\infty \bmod u^{n+1}$, for $n \geqslant 5$, be the finite-dimensional (dim = 5, $d=2$) Poisson–Lie group with a Poisson–Lie structure defined by

$$\{x_i, x_j\} = (i-2) j x_j x_{i-2} - i(j-2) x_i x_{j-2} + x_i \sum_{s_1+s_2+s_3=j} x_{s_1} x_{s_2} x_{s_3} - x_j \sum_{s_1+s_2+s_3=i} x_{s_1} x_{s_2} x_{s_3}.$$

The above formulas are obtained from (117) with $d=2$, and we adopt the convention that $x_i = 0$ whenever $i < 1$. In more detail, we have

$$\{x_1, x_2\} = 0,$$

$$\{x_1, x_3\} = -x_1^2 + x_1^4,$$

$$\{x_2, x_3\} = x_2(x_1^3 - 2x_1),$$

$$\{x_1, x_4\} = x_2(3x_1^3 - 2x_1),$$

$$\{x_2, x_4\} = x_2^2(3x_1^3 - 4),$$

$$\{x_3, x_4\} = x_4(4x_1 - x_1^3) + x_3 x_2(3x_1^2 - 6),$$

$$\{x_1, x_5\} = x_3(3x_1^3 - 3x_1) + 3x_2^2 x_1^2,$$

$$\{x_2,x_5\}=3x_2^3x_1+x_3x_2(3x_1^2-6),$$

$$\{x_3,x_5\}=x_5(5x_1-x_1^3)+x_3^2(3x_1^2-9)+3x_3x_2^2x_1,$$

$$\{x_4,x_5\}=x_5x_2(10-3x_1^2)+x_4x_3(3x_1^2-12)+3x_4x_2^2x_1.$$

Theorem IX.1: *Let $X=\{x_i\}_{1\leqslant i\leqslant 5}$ be a set of indeterminates and let $\langle X\rangle$ be the associative semigroup with identity generated by X. Consider an ideal $\mathcal{I}_h^2$ generated by the set of relations $\mathcal{R}_h^2$ in $k[[h]]\langle X\rangle$:*

$$x_1x_2=x_2x_1,$$

$$x_1x_3=x_3x_1+h(-x_1^2+x_1^4),$$

$$x_2x_3=x_3x_2+hx_2(x_1^3-2x_1),$$

$$x_1x_4=x_4x_1+hx_2(3x_1^3-2x_1),$$

$$x_2x_4=x_4x_2+hx_2^2(3x_1^3-4),$$

$$x_3x_4=x_4x_3+h[x_4(4x_1-x_1^3)+x_3x_2(3x_1^2-6)]+2h^2x_2x_1,$$

$$x_1x_5=x_5x_1+h[x_3(3x_1^3-3x_1)+3x_2^2x_1^2]+h^2\left(-6x_1^4+\frac{9}{2}x_1^6+\frac{3}{2}x_1^2\right),$$

$$x_2x_5=x_5x_2+h[3x_2^3x_1+x_3x_2(3x_1^2-6)]+h^2x_2\left(6x_1-9x_1^3+\frac{9}{2}x_1^5\right),$$

$$x_3x_5=x_5x_3+h[x_5(5x_1-x_1^3)+x_3^2(3x_1^2-9)+3x_3x_2^2x_1]+h^2x_3\left(-\frac{15}{2}x_1+6x_1^3+\frac{3}{2}x_1^5\right)$$
$$+h^3C(x_1^8-x_1^2),$$

$$x_4x_5=x_5x_4+h[x_5x_2(10-3x_1^2)+x_4x_3(3x_1^2-12)+3x_4x_2^2x_1]$$
$$+h^2\left[x_4\left(-24x_1+9x_1^3+\frac{3}{2}x_1^5\right)+6x_3x_2\right]+h^3x_2[-(6+2C)x_1+3Cx_1^7],$$

where $C\in k$ is an arbitrary parameter. Then the semigroup quotient algebra $k[[h]]\langle\rangle/\mathcal{I}_h^2$ defines a quantum semigroup $G_{5|h}^2$ in the sense of Definition IX.2 with a comultiplication defined by (119). Namely,

$$\Delta x_1=x_1\otimes x_1,$$

$$\Delta x_2=x_1\otimes x_2+x_2\otimes x_1^2,$$

$$\Delta x_3=x_1\otimes x_3+x_2\otimes x_1x_2+x_2\otimes x_2x_1+x_3\otimes x_1^3, \tag{121}$$

$$\Delta x_4=x_1\otimes x_4+x_2\otimes x_1x_3+x_2\otimes x_2^2+x_2\otimes x_3x_1+x_3\otimes x_1^2x_2+x_3\otimes x_1x_2x_1+x_3\otimes x_2x_1^2+x_4\otimes x_1^4,$$

$$\Delta x_5 = x_1 \otimes x_5 + x_2 \otimes x_1 x_4 + x_2 \otimes x_2 x_3 + x_2 \otimes x_3 x_2 + x_2 x_4 x_1 + x_3 x_1^2 x_3 + x_3 \otimes x_1 x_2^2$$
$$+ x_3 \otimes x_1 x_3 x_1 + x_3 \otimes x_2 x_1 x_2 + x_3 \otimes x_2^2 x_1 + x_3 \otimes x_3 x_1^2 + x_4 \otimes x_1^3 x_2$$
$$+ x_4 \otimes x_1^2 x_2 x_1 + x_4 \otimes x_1 x_2 x_1^2 + x_4 \otimes x_2 x_1^3 + x_5 \otimes x_1^4.$$

Moreover, the semigroup algebra $k[[h]]\langle X \rangle$ *has the Poincaré–Birkhoff–Witt property.*

Proof: The proof is constructive. We look for a set of relations $\mathcal{R}_h^2$ in $k[[h]]\langle X \rangle$ in the following form:

$$x_1 x_2 = x_2 x_1,$$

$$x_1 x_3 = x_3 x_1 + h(-x_1^2 + x_1^4),$$

$$x_2 x_3 = x_3 x_2 + h x_2 (x_1^3 - 2x_1),$$

$$x_1 x_4 = x_4 x_1 + h x_2 (3x_1^3 - 2x_1),$$

$$x_2 x_4 = x_4 x_2 + h x_2^2 (3x_1^3 - 4) + h^2 f_1(x_1),$$

$$x_3 x_4 = x_4 x_3 + h[x_4(4x_1 - x_1^3) + x_3 x_2 (3x_1^2 - 6)] + h^2 x_2 f_2(x_1),$$

$$x_1 x_5 = x_5 x_1 + h[x_3(3x_1^3 - 3x_1) + 3x_2^2 x_1^2] + h^2 f_3(x_1),$$

$$x_2 x_5 = x_5 x_2 + h[x_3 x_2 (3x_1^2 - 6) + 3x_2^3 x_1] + h^2 x_2 f_4(x_1),$$

$$x_3 x_5 = x_5 x_3 + h[x_5(5x_1 - x_1^3) + x_3^2(3x_1^2 - 9) + 3x_3 x_2^2 x_1] + h^2[x_3 f_5(x_1) + x_2^2 f_6(x_1)] + h^3 f_7(x_1),$$

$$x_4 x_5 = x_5 x_4 + h[x_5 x_2 (10 - 3x_1^2) + x_4 x_3 (3x_1^2 - 12) + 3x_4 x_2^2 x_1]$$
$$+ h^2[x_4 f_8(x_1) + x_3 x_2 f_9(x_1) + x_2^3 f_{10}(x_1)] + h^3 x_2 f_{11}(x_1), \tag{122}$$

where $\{f_i(x)\}_{1 \leqslant i \leqslant 11}$ is a set of arbitrary polynomials. Since the degree of h is $|h| = 2$ and the degree of x_1 is $|x_1| = 0$ this is the most general form of the set of relations that one can have, such that their quasi-classical limit gives the Poisson–Lie structure on G_5^2.

We now require $k[[h]]\langle X \rangle / \mathcal{I}_h^2$ to be a bialgebra, that is, that Δ be an algebra homomorphism. This leads to some restrictions on the polynomials $f_i(x)$. Using the formulas for comultiplication (121) one can see that the first four relations are compatible with the coalgebra structure. One obtains functional equations for $f_i(x)$ from the remaining six relations after a reduction to a canonical form. We shall analyze first the equations that arise from terms of order h^2.

(a) The compatibility with comultiplication induces that

$$\Delta(x_2 x_4) = \Delta(x_4 x_2) + h(-4\Delta x_2^2 + 3\Delta(x_2^2 x_1^2)) + h^2 f_1(\Delta x_1).$$

After reducing both sides of this relation to a canonical form, using the comultiplication formulas (121), we obtain the following linear functional equation for $f_1(x)$: $-f_1(x_1 \otimes x_1) + x_1^2 \otimes f_1(x_1) + x_1^6 \otimes f_1(x_1) = 0$. From now on since all equations for the unknowns f_i will depend only on the variable x_1 we shall use the notation $x := x_1$. Therefore we have

$$-f_1(x \otimes x) + x^2 \otimes f_1(x) + x^6 \otimes f_1(x) = 0.$$

The most general solution of the above equation is

$$f_1(x) = C_1(x^6 - x^2),$$

where $C_1 \in k$ is an arbitrary constant. There are no terms of higher order in h that arise in the analysis of this relation. We move on to the next.

(b) Again after reducing to a canonical form both sides of

$$\Delta(x_3x_4)=\Delta(x_4x_3)+h(\Delta(4x_4x_1)-\Delta(x_4x_1^3)-\Delta(6x_3x_2)+\Delta(3x_3x_2x_1^2))+h^2\Delta(x_2f_2(x_1)),$$

we obtain two equations. One of them arises from a term proportional to $x_2\otimes 1$, that is, we have a term of the form

$$(x_2\otimes 1)[-f_2(x_1\otimes x_1)+(2-2C_1)x_1\otimes x_1+f_2(x_1)\otimes x_1^5+(-2+2C_1)x_1\otimes x_1^5],$$

which does not cancel. It leads to the equation

$$-f_2(x\otimes x)+(2-2C_1)x\otimes x+f_2(x)\otimes x^5+(-2+2C_1)x\otimes x^5=0. \tag{123}$$

The term proportional to $1\otimes x_2$ leads to

$$-f_2(x\otimes x)+x\otimes f_2(x)-2C_1x\otimes x^5+2C_1x^5\otimes x^5=0. \tag{124}$$

Solving (123) and (124) together we obtain

$$f_2(x)=(2-2C_1)x+2C_1x^5.$$

There are no terms of higher order in h arising from this relation.

At this stage of the calculation we check whether the PBW property is satisfied in the subalgebra of $k[[h]]\langle X\rangle$ generated by the set $\{x_1,x_2,x_3,x_4\}$ and subject to the first six relations of (122). By direct calculation, using the Diamond Lemma, one shows that the monomial $x_2x_3x_4$ can be reduced to a unique canonical form if and only if $C_1=0$. The other possible monomials of three variables have a unique canonical form. Therefore we obtain that

$$f_1(x)=0 \quad \text{and} \quad f_2(x)=2x.$$

(c) From the next relation one has

$$\Delta(x_1x_5)=\Delta(x_5x_1)+h(-3\Delta(x_3x_1)+3\Delta(x_2^2x_1^2)+3\Delta(x_3x_1^3))+h^2f_3(\Delta x_1);$$

after a reduction to a canonical form one is lead to the equation

$$-f_3(x\otimes x)+x^2\otimes f_3(x)+6x^2\otimes x^4+x^6\otimes f_3(x)-6x^4\otimes x^4-6x^2\otimes x^6+6x^4\otimes x^6=0.$$

The most general solution of the above equation is given by

$$f_3(x)=6x^2-6x^4+C_2(x^6-x^2),$$

where $C_2\in k$ is an arbitrary constant. There are no terms of order h^3 or higher that arise after the reduction to a canonical form of thisabove relation.

(d) The analysis of the next relation,

$$\Delta(x_2x_5)=\Delta(x_5x_2)+h(-6\Delta(x_3x_2)+3\Delta(x_2^3x_1)+3\Delta(x_3x_2x_1^2))+h^2\Delta(x_2f_4(x_1)),$$

leads to two equations. The first one is

$$-f_4(x\otimes x)+(15-2C_2)x\otimes x+f_4(x)\otimes x^5-9x^3\otimes x^3+(-15+2C_2)x\otimes x^5+9x^3\otimes x^5=0, \tag{125}$$

and comes from a term proportional to $h^2x_2\otimes 1$. The second equation comes from a term proportional to $h^2(1\otimes x_2)$ and reads as

$$-f_4(x\otimes x)+x\otimes f_4(x)+9x\otimes x^3-9x^3\otimes x^3-C_2x\otimes x^5+C_2x^5\otimes x^5=0. \tag{126}$$

Solving together (125) and (126) one obtains

$$f_4(x)=(15-2C_2)x-9x^3+C_2x^5. \tag{127}$$

There are no terms of higher order in h that do not cancel after the reduction to a canonical form.

(e) We move on to the next relation which gives

$$\Delta(x_3x_5)=\Delta(x_5x_3)+h\Delta(5x_5x_1-9x_3^2+3x_3x_2^2x_1+3x_3^2x_1^2-x_5x_1^3)$$
$$+h^2\Delta[x_3f_5(x_1)+x_2^2f_6(x_1)]+h^3\Delta f_7(x_1).$$

Terms of order h^2 give rise to five functional equations which we now describe.

(i) A term proportional to $x_3\otimes 1$ gives rise to

$$-f_5(x\otimes x)+(6-3C_2)x\otimes x+f_5(x)\otimes x^5+6x^3\otimes x^3-6x^3\otimes x^5+(-6+3C_2)x\otimes x^5=0, \tag{128}$$

and a term prortional to $1\otimes x_3$ gives rise to the equation

$$-f_5(x\otimes x)+x\otimes f_5(x)-6x\otimes x^3+6x^3\otimes x^3+(3-C_2)x\otimes x^5+(-3+C_2)x^5\otimes x^5=0. \tag{129}$$

Solving (128) and (129) together we obtain for f_5,

$$f_5(x)=(6-3C_2)x+6x^3+(-3+C_2)x^5.$$

(ii) Terms proportional to $x_2^2\otimes 1$ and $1\otimes x_2^2$ give rise to another two functional equations:

$$-f_6(x\otimes x)+x^4\otimes f_6(x)=0, \tag{130}$$

and

$$-f_6(x\otimes x)+1\otimes f_6(x)=0, \tag{131}$$

respectively. The only solution that satisfies both (130) and (131) is

$$f_6(x)=0. \tag{132}$$

(iii) The last term from the terms of order h^2 is a term proportional to $x_2\otimes x_2$ which gives rise to the following equation for f_5 and f_6:

$$-2f_5(x\otimes x)+(12-6C_2)x\otimes x-2(x\otimes x)f_6(x\otimes x)+12x^3\otimes x^3+(-6+2C_2)x^5\otimes x^5=0. \tag{133}$$

After substituting the solutions $f_5(x)$ and (132) into (133) we obtain that it is satisfied identically. There is one term of order h^3 that arises which gives rise to

$$x^2\otimes f_7(x)+f_7(x)\otimes x^8-f_7(x\otimes x)=0. \tag{134}$$

The most general solution of (134) is given by

$$f_7(x)=C_3(x^8-x^2),$$

where $C_3\in k$ is an arbitrary constant. No terms of higher order in h arise.

(f) The last relation to be analyzed is

$$\Delta(x_4x_5)=\Delta(x_5x_4)+h\Delta(10x_5x_2-12x_4x_3+3x_4x_2^2x_1+3x_4x_3x_1^2-3x_5x_2x_1^2)$$
$$+h^2\Delta[x_4f_8(x_1)+x_3x_2f_9(x_1)+x_2^3f_{10}(x_1)]+h^3\Delta(x_2f_{11}(x_1)).$$

After reducing to a canonical form both sides of the above relation we obtain ten terms of order h^2 that do not cancel and two terms of order h^3. We analyze first the terms of order h^2.

(i) Two terms proportional to $x_4\otimes 1$ and $1\otimes x_4$ give rise to the following two equations:

$$-f_8(x\otimes x)+(-6-4C_2)x\otimes x+f_8(x)\otimes x^5+9x^3\otimes x^3+(6+4C_2)x\otimes x^5-9x^3\otimes x^5=0, \tag{135}$$

$$-f_8(x\otimes x)+x\otimes f_8(x)-9x\otimes x^3+9x^3\otimes x^3+(3-C_2)x\otimes x^5+(-3+C_2)x^5\otimes x^5=0. \tag{136}$$

The most general solution of (135) and (136) is

$$f_8(x)=(-6-4C_2)x+9x^3+(-3+C_2)x^5. \tag{137}$$

(ii) Terms proportional to $x_3x_2\otimes 1$ and $1\otimes x_3x_2$ give rise to the equations

$$6(1\otimes 1)-f_9(x\otimes x)-6(1\otimes x^4)+f_9(x)\otimes x^4=0, \tag{138}$$

$$-f_9(x\otimes x)+1\otimes f_9(x)=0. \tag{139}$$

The solution of the system (138) and (139) is

$$f_9(x)=6. \tag{140}$$

(iii) A term proportional to $x_3\otimes x_2$ gives rise to

$$-3f_8(x\otimes x)+(-12-12C_2)x\otimes x-(x\otimes x)f_9(x\otimes x)+27x^3\otimes x^3+(-9+3C_2)x^5\otimes x^5=0. \tag{141}$$

After substituting (137) and (140) into (141) it yields an identity. Similarly the term proportional to $x_2\otimes x_3$ leads to an identity.

(iv) Two terms proportional to $x_2^3\otimes 1$ and $1\otimes x_2^3$ lead to the equations

$$-f_{10}(x\otimes x)+f_{10}(x)\otimes x^3=0, \tag{142}$$

$$1\otimes f_{10}(x)-(x\otimes 1)f_{10}(x\otimes x)=0. \tag{143}$$

The only solution of (142) and (143) solved together is

$$f_{10}(x)=0. \tag{144}$$

(v) The terms proportional to $x_2\otimes x_2^2$ and $x_2^2\otimes x_2$ give rise to

$$-f_8(x\otimes x)+(6-4C_2)x\otimes x-2(x\otimes x)f_9(x\otimes x)$$
$$-3(x^2\otimes x^2)f_{10}(x\otimes x)+9x^3\otimes x^3+(-3+C_2)x^5\otimes x^5=0, \tag{145}$$

and

$$12(1\otimes 1)-2f_9(x\otimes x)-3(x\otimes x)f_{10}(x\otimes x)=0, \tag{146}$$

which are identically satisfied. This becomes obvious after substituting (137), (140), and (144) into (145) and (146). The two terms of order h^3 are proportional to $x_2\otimes 1$ and $1\otimes x_2$ and give rise to

$$-f_{11}(x\otimes x)+(3-2C_2-2C_3)x\otimes x+f_{11}(x)\otimes x+(-3+2C_2+2C_3)x\otimes x^7=0, \tag{147}$$

$$-f_{11}(x\otimes x)+x\otimes f_{11}(x)-3C_3x\otimes x^7+3C_3x^7\otimes x^7=0, \tag{148}$$

respectively. The most general solution of (147) and (148) is

$$f_{11}(x)=(3-2C_2-2C_3)x+3C_3x^7.$$

We need one last step in order to complete the construction. We would like to find whether the so obtained set of relations define an algebra with the PBW property. After lengthy and tedious calculation one shows that the requirement that the monomials $x_1x_3x_5$, $x_1x_4x_5$, $x_2x_4x_5$, and $x_3x_4x_5$ can be reduced to a unique canonical form imposes the following single equation on the arbitrary constant C_2:

$$-9+2C_2=0.$$

The monomial $x_1x_2x_5$ is reducible to a canonical form without imposing any conditions. Thus $C_2=\frac{9}{2}$. If we introduce $C:=C_3$ we obtain the statement of the Theorem. This concludes the proof. □

Remark IX.1: Notice that our construction yields a one-parameter family of quantum semigroups $G_{5}^{2}\big|_{h}^{C}$ *parametrized by* C. *Theorem I.2 describes a family of quantum semigroups parametrized by even more parameters.*

Let $G_4^1=G_\infty^1 \bmod u^{n+1}$, for $n\geqslant 4$, be the finite-dimensional (dim = 4, $d=1$) Poisson–Lie group with a Poisson–Lie structure defined by

$$\{x_i,x_j\}=(i-1)jx_jx_{i-1}-i(j-1)x_ix_{j-1}+x_i\sum_{s_1+s_2=j}x_{s_1}x_{s_2}-x_j\sum_{s_1+s_2=i}x_{s_1}x_{s_2}.$$

The explicit formulas for the Poisson brackets were given in the Introduction and their quantization was described by Theorem I.2.

Remark IX.2: The proof of Theorem I.2 goes along the same lines as the proof of Theorem IX.1, that is, it is constructive. In the course of the construction five arbitrary constants C_1,C_2,C_3,C_4,C_5 *appear in solving the corresponding functional equations. The requirement for the existence of a PBW basis fixes two of them. Namely,* $C_1=1$ *and* $C_2=2-2C_3$. *Thus we obtain a 3-parameter family of quantum semigroups* $G_{4|h}^{1|C_3,C_4,C_5}$.

Finally we describe a third quantum semigroup arising after the quantization of the Poisson algebra of functions on the finite-dimensional (dim = 5, $d=3$) Poisson–Lie group G_5^3. The Poisson–Lie structure on G_5^3 is given by

$$\{x_i,x_j\}=(i-3)jx_jx_{i-3}-i(j-3)x_ix_{j-3}+x_i\sum_{s_1+s_2+s_3+s_4=j}x_{s_1}x_{s_2}x_{s_3}x_{s_4}$$
$$-x_j\sum_{s_1+s_2+s_3+s_4=i}x_{s_1}x_{s_2}x_{s_3}x_{s_4}.$$

The above formulas are obtained again from (117) with $d=3$. Writing them explicitly we have

$$\{x_1,x_2\}=0,$$

$$\{x_1,x_3\}=0,$$

$$\{x_2,x_3\}=0,$$

$$\{x_1,x_4\}=x_1^5-x_1^2,$$

$$\{x_2,x_4\}=x_2(x_1^4-2x_1),$$

$$\{x_3,x_4\}=x_3(x_1^4-3x_1),$$

$$\{x_1,x_5\}=x_2(4x_1^4-2x_1),$$

$$\{x_2,x_5\}=x_2^2(4x_1^4-4),$$

$$\{x_3,x_5\}=x_3x_2(4x_1^3-6),$$

$$\{x_4,x_5\}=x_4x_2(4x_1^3-8)+x_5(5x_1-x_1^4).$$

Then we have our last theorem.

Theorem IX.2: *Let $X=\{x_i\}_{1\leqslant i\leqslant 5}$ be a set and let $\langle X\rangle$ be the associative semigroup with identity generated by X. Consider an ideal $\mathcal{I}_h^3$ generated by the set of relations $\mathcal{R}_h^3$ in $k[[h]]\langle X\rangle$:*

$$x_1x_2=x_2x_1,$$

$$x_1x_3=x_3x_1,$$

$$x_2x_3=x_3x_2,$$

$$x_1x_4=x_4x_1+h(x_1^5-x_1^2),$$

$$x_2x_4=x_4x_2+hx_2(x_1^4-2x_1),$$

$$x_3x_4=x_4x_3+hx_3(x_1^4-3x_1),$$

$$x_1x_5=x_5x_1+hx_2(4x_1^4-2x_1),$$

$$x_2x_5=x_5x_2+hx_2^2(4x_1^4-4),$$

$$x_3x_5=x_5x_3+hx_3x_2(4x_1^3-6),$$

$$x_4x_5=x_5x_4+h[x_4x_2(4x_1^3-8)+x_5(5x_1-x_1^4)]+h^23x_2x_1.$$

Then the semigroup quotient algebra $k[[h]]\langle X\rangle/\mathcal{I}_h^3$ defines a quantum semigroup $G_{5|h}^3$ in the sense of Definition IX.2 with a comultiplication defined by (121), and the semigroup algebra $k[[h]]\langle X\rangle$ has the Poincaré–Birkhoff–Witt property.

Remark IX.3: The proof is again constructive. Note that no arbitrary parameters arise in dimension 5 for $d=3$. Arbitrary parameters arise in higher dimensions though. We have been able to construct all quantum semigroups $G_{n|h}^d$ for $n\leqslant 7$ and $d\leqslant 5$. However, we refrain from describing more examples here.

ACKNOWLEDGMENTS

The author would like to thank B. Kupershmidt for encouragement and help, and P. Etingof for helpful suggestions.

[1] V. G. Drinfel'd, ''Quantum groups,'' in *Proceedings of the International Congress of Mathematicians, Berkeley*, 1987, Vol. 1, pp. 798–819.
[2] M. Jimbo, ''A *q*-difference analogue of $\mathcal{U}(\mathcal{G})$ and the Yang–Baxter equation,'' Lett. Math. Phys. **10**, 63–69 (1985).
[3] L. Faddeev, N. Reshetikhin, and L. Takhtajan, ''Quantization of Lie groups and Lie algebras,'' Leningrad Math. J. **1**, 1 (1989).
[4] A. Belavin and V. G. Drinfel'd, ''Triangle equations and simple Lie algebras,'' Math. Phys. Rev. **4**, 93–165 (1984).

[5] A. Belavin and V. G. Drinfel'd, ''Solution of the classical Yang–Baxter equation for simple Lie algebras,'' Funct. Anal. Appl. **16**, 159–180 (1982).
[6] A. Belavin and V. G. Drinfel'd, ''Classical Yang–Baxter equation for simple Lie algebras,'' Funct. Anal. Appl. **17**, 220–221 (1982).
[7] B. Kupershmidt, ''The quantum group $gl_h(2)$,'' J. Phys. A **25**, L1239–L1244 (1992).
[8] B. Kupershmidt, ''All quantum group structures on the supergroup $gl(1|1)$,'' J. Phys. A **26**, L251–L256 (1993).
[9] B. Kupershmidt, ''Quantum subgroups of the quantum group $gl(2)$,'' Phys. Lett. A **191**, 113–116 (1994).
[10] B. Kupershmidt and O. Stoyanov, ''Classical mechanics on unipotent groups,'' in preparation.
[11] S. Zakrzewski, ''Poisson structures on the Lorentz group,'' Lett. Math. Phys. **32**, 11–23 (1994).
[12] B. Kupershmidt and O. Stoyanov, ''Classification of all Poisson–Lie structures on an infinite-dimensional jet group,'' Lett. Math. Phys. **37**, 1–9 (1996).
[13] S. K. Sklyanin, ''Some algebraic structures connected with the Yang–Baxter equation,'' Funct. Anal. Appl. **16**, 263–270 (1982).
[14] B. Kupershmidt, *Discrete Lax Equations And Differential-Difference Calculus* (Asterisque, Paris, 1985).
[15] A. Lichnerowicz, ''Les variétés de Poisson et leurs algébres de Lie associées,'' J. Diff. Geom. **12**, 253–300 (1977).
[16] V. G. Drinfel'd, ''Hamiltonian structures on Lie groups, Lie bialgebras and the geometric meaning of the classical Yang–Baxter equations,'' Sov. Math. Dokl. **27**, 68 (1983).
[17] M. A. Semenov-Tian-Shansky, ''What is a classical r-matrix?,'' Funct. Anal. Appl. **17**, 259–272 (1983).
[18] J.-H. Lu and A. Weinstein, ''Poisson–Lie groups, dressing transformations and Bruhat decompositions,'' J. Diff. Geom. **31**, 501–526 (1990).
[19] O. S. Stoyanov, *Poisson–Lie structures on infinite-dimensional jet groups and their quantization*, Ph.D. thesis, VPI&SU, 1993, q-alg/9506008.
[20] D. J. Saunders, *The Geometry of Jet Bundles*, London Mathematical Society Lecture Note Series (Cambridge University Press, Cambridge, 1989), Vol. 142.
[21] A. S. Dzhumadil'daev, ''Quasi-Lie bialgebra structures of sl_2, Witt and Virasoro algebras,'' in *Israel Mathematics Conference Proceedings*, 1993, Vol. 7, pp. 13–24.
[22] W. Michaelis, ''A class of infinite-dimensional Lie bialgebras containing the Virasoro algebra,'' Adv. Math. **107**, 365–392 (1994).
[23] E. J. Taft, ''Witt and Virasoro algebras as Lie bialgebras,'' J. Pure Appl. Alg. **87**, 301–312 (1993).
[24] J. Dieudonné, *Introduction to the Theory of Formal Groups* (Marcel Dekker, New York, 1973).
[25] J.-P. Serre, *Lie Groups and Lie Algebras*, Lecture Notes in Mathematics (Springer-Verlag, Berlin, 1992), Vol. 1500.
[26] M. A. Semenov-Tian-Shansky, ''Dressing transformations and Poisson group actions,'' Publ. RIMS **21**, 1237–1260 (1985).
[27] V. G. Drinfel'd, *On Quadratic Commutation Relations in the Quasiclassical Case* (Naukova Dumka, Kiev, 1986), pp. 25–33 (in Russian).
[28] G. M. Bergman, ''The diamond Lemma for ring theory,'' Adv. Math. **29**, 178–218 (1978).
[29] Yu. Manin, *Topics in Non-Commutative Geometry* (Princeton University Press, Princeton, 1991).

Functional versus canonical quantization of a nonlocal massive vector-gauge theory

R. Amorim[a] and J. Barcelos-Neto[b]
Instituto de Física, Universidade Federal do Rio de Janeiro, RJ 21945-970, Caixa Postal 68528, Brazil

(Received 5 May 1998; accepted for publication 26 August 1998)

It has been shown in literature that a possible mechanism of mass generation for gauge fields is through a topological coupling of vector and tensor fields. After integrating over the tensor degrees of freedom, one arrives at an effective massive theory that, although gauge invariant, is nonlocal. Here we quantize this nonlocal resulting theory both by path integral and canonical procedures. This system can be considered as equivalent to one with an infinite number of time derivatives and consequently an infinite number of momenta. This means that the use of the canonical formalism deserves some care. We show the consistency of the formalism we use in the canonical procedure by showing that the obtained propagators are the same as those of the (Lagrangian) path integral approach. The problem of nonlocality appears in the obtainment of the spectrum of the theory. This fact becomes very transparent when we list the infinite number of commutators involving the fields and their velocities.
[S0022-2488(99)03501-X]

I. INTRODUCTION

It is widely accepted that the forces of nature are described by gauge theories. These theories are characterized by the gauge symmetries which are related to massless fields. However, sometimes it is necessary that these fields become massful, as occurs, for instance, in the case of the Salam–Weinberg theory. Nowadays, it has also been widely accepted that spontaneous symmetry breaking together with the Higgs mechanism is the most probable explanation for the origin of the acquisition of mass by gauge fields. However, if this is actually true, the Higgs bosons must exist in nature. The point is that there is no precise theoretical prediction on the mass scale where these fields could be found and experiments until now have shown no evidence about them.

In this way, alternative mechanisms of mass generation for gauge fields that do not spoil what is well established and that do not contain Higgs bosons are welcome. This might be the case of vector–tensor gauge theories,[1] where vector and tensor fields are coupled in a topological way by means of a kind of Chern–Simons term. The general idea of this mechanism resides in the following: Tensor gauge fields[2] are antisymmetric quantities and consequently in $D=4$ they exhibit six degrees of freedom. By virtue of the massless condition, the number of degrees of freedom goes down to four. Since the gauge parameter is a vector quantity, this number would be zero if all of its components were independent. This is nonetheless the case because the system is reducible (which means that the gauge transformations are not all independent) and we mention that the final number of physical degrees of freedom is one. It is precisely this degree of freedom that can be absorbed by the vector gauge field in the vector–tensor gauge theory in order to acquire mass.[1,3] This peculiar structure of constraints involving tensor gauge theories implies that quantization as well as its non-Abelian formulation deserve some care and a reasonable amount of work has been done on these subjects.[4–7]

[a]Electronic mail: amorim@if.ufrj.br
[b]Electronic mail: barcelos@if.ufrj.br

Usually, the treatment of the vector–tensor gauge theory is carried out with both vector and tensor fields placed together and just at the end the integration over the tensor field is done in order to obtain the effective result for the vector theory. This procedure usually hides an important aspect of this effective theory, that is, its nonlocality. We mention that it is equivalent to a theory with an infinite number of higher derivative terms and, consequently, an infinite number of momenta also.

It is not our purpose here to advocate if a vector–tensor gauge theory with topological coupling is more suitable to explain the mass generation than the usual spontaneous symmetry breaking together with the Higgs mechanism. Our intention in the present paper is to study the quantization of a massive vector gauge field directly by means of the nonlocal effective Lagrangian.[8] We do it both by path integral, where we use a Lagrangian formulation, as well as by the canonical approach. This is the subject of Secs. II and III, respectively. We would like also to add that the non-Abelian formulation for the vector tensor gauge theory is not a simple task. This is so because the non-Abelian version loses the reducibility condition unless we consider that the Maxwell stress tensor is zero.[6] Another possibility is to introduce a kind of Stuckelberg field, that disappears in the Abelian limit, in order to keep the same number of degrees of freedom in both sectors (Abelian and non-Abelian) of the theory.[7] We shall consider only the Abelian case in this paper. We left Sec. IV for some concluding remarks and include four appendices to present details of some calculations.

II. BRIEF REVIEW OF THE VECTOR–TENSOR GAUGE THEORY AND THE PATH INTEGRAL QUANTIZATION OF THE EFFECTIVE VECTOR THEORY

The Abelian theory for vector and tensor fields coupled in a topological way is described by the Lagrangian density:[1]

$$\mathcal{L}=-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}+\frac{1}{12}H_{\mu\nu\rho}H^{\mu\nu\rho}+\frac{m}{2}\epsilon^{\mu\nu\rho\lambda}B_{\mu\nu}F_{\rho\lambda}, \tag{2.1}$$

where $F_{\mu\nu}$ and $H_{\mu\nu\rho}$ are totally antisymmetric tensors written in terms of the potentials A_μ and $B_{\mu\nu}$ (also antisymmetric) through the stress tensors

$$F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu, \tag{2.2}$$

$$H_{\mu\nu\rho}=\partial_\mu B_{\nu\rho}+\partial_\rho B_{\mu\nu}+\partial_\nu B_{\rho\mu}. \tag{2.3}$$

In expression (2.1), $\epsilon^{\mu\nu\rho\lambda}$ is the totally antisymmetric symbol and m is a mass parameter. It is easy to see, by using the (coupled) Euler–Lagrange equations for A^μ and $B^{\mu\nu}$, as well as the Jacobi identity, that $F_{\mu\nu}$ satisfy a massive Klein–Gordon equation, with a mass parameter m.[1]

We observe that the Lagrangian (2.1) is invariant under the gauge transformations

$$\delta A^\mu=\partial^\mu\Lambda, \tag{2.4}$$

$$\delta B^{\mu\nu}=\partial^\mu\Lambda^\nu-\partial^\nu\Lambda^\mu, \tag{2.5}$$

where Λ and Λ^μ are (before fixing the gauge) generic functions of space–time. This is a reducible theory, which means that not all the gauge transformations above are independent. In fact, if we choose the gauge parameter Λ^μ as the gradient of some scalar Ω we have that $B^{\mu\nu}$ does not change under the gauge transformation (2.5).

Functionally integrating over the antisymmetric tensor field $B_{\mu\nu}$ we get, after a convenient gauge fixing procedure, the effective action[3]

$$S_0[A_\mu]=-\frac{1}{4}\int d^4x\, F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)F^{\mu\nu}. \tag{2.6}$$

The action (2.6), although nonlocal, is gauge invariant. It is important to emphasize that this theory is renormalizable, a characteristic that is lost when a mass term is directly put by hand as in the Proca theory.

Let us calculate the (covariant) propagator for the field A_μ. We opt to use a Lagrangian formulation in order to avoid the problem of the infinite number of momenta. Let us use the Batalin–Vilkovisky (BV) formalism.[5,9] The nonminimum BV action can be written as

$$S = S_0 + \int dx (A^*_\mu \partial^\mu c + b\overline{c^*}), \tag{2.7}$$

where the A^*_μ and the pair $(c^*, \overline{c^*})$ are, respectively, the antifields of the gauge field A^μ and of the ghosts $(c, \bar{c})$. The auxiliary field b was introduced in order to fix the gauge in a covariant way. This can be done, for instance, with the aid of the gauge-fixing fermion functional

$$\Psi = \int dx \; \bar{c}(-\alpha b + \partial^\mu A_\mu), \tag{2.8}$$

with α being a parameter. The vacuum functional is defined by[5,9]

$$Z_\Psi = \int [dA_\mu][d\bar{c}][dc][db][dA^*_\mu][d\overline{c^*}][dc^*] \delta\left[\phi^* - \frac{\delta\Psi}{\delta\phi}\right] \exp\{iS\}. \tag{2.9}$$

The action S is given by (2.7) and ϕ is generically referring to gauge and ghost fields. After functionally integrating over the antifields as well as over the auxiliary field b, we arrive at

$$\bar{S}[J] = \int d^4x \left[-\frac{1}{4} F_{\mu\nu}\left(1 + \frac{m^2}{\Box}\right) F^{\mu\nu} - \partial_\mu \bar{c} \partial^\mu c + \frac{1}{2\alpha} (\partial^\mu A_\mu)^2 + J_\mu A^\mu \right], \tag{2.10}$$

where we have introduced an external source J^μ in order to calculate the propagator. This can be directly obtained by a straightforward calculation. The result, written in momentum space, is

$$K^{\mu\nu} = -\frac{1}{k^2 + m^2} \left[\eta^{\mu\nu} + \left(\frac{\alpha - 1}{k^2} + \frac{m^2}{k^4} \right) k^\mu k^\nu \right]. \tag{2.11}$$

We notice that there is actually a mass pole at $k_0^2 = \vec{k}^2 + m^2$.

In Sec. III we are going to see how this and other features appear in terms of a canonical quantization procedure.

III. CANONICAL QUANTIZATION

Let us consider the Lagrangian for vector fields of Eq. (2.10),

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu}\left(1 + \frac{m^2}{\Box}\right) F^{\mu\nu} - \frac{1}{2\alpha} (\partial^\mu A_\mu)^2. \tag{3.1}$$

To implement the process of canonical quantization for such system, it is necessary to obtain the canonical momenta. So, we have to isolate the time derivatives from the nonlocal operator $\Box^{-1}$. We then conveniently write

$$\frac{1}{\Box} = -(\nabla^2 - \partial_t^2)^{-1} = -\frac{1}{\nabla^2} - \frac{\partial_t^2}{\nabla^4} - \frac{\partial_t^4}{\nabla^6} - \cdots . \tag{3.2}$$

As one observes, the system described by (3.1) is effectively a system with an infinite number of time derivatives and consequently it contains an infinite number of momenta.[10,11] A practical way of obtaining the momentum expressions is to consider the variation of the action by fixing the fields and their velocities at just one of the extreme times, say, $\delta A_\mu(\vec{x},t_0)=0=\delta\dot{A}_\mu(\vec{x},t_0)=\delta\ddot{A}_\mu(\vec{x},t_0)=\cdots$. After some algebraic calculation, we get (please, see Appendix A)

$$\begin{aligned}\delta\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\mathcal{L}=&\int_{t_0}^{t}d\tau\int d^3\vec{x}\left[\left(1+\frac{m^2}{\Box}\right)\partial^\mu F_{\mu\nu}+\frac{1}{\alpha}\,\partial_\mu\partial_\nu A^\mu\right]\delta A^\nu-\int d^3\vec{x}\left[\left(1+\frac{m^2}{\Box}\right)F_{0\nu}\right.\\&\left.+\frac{m^2}{2\Box}\,\frac{\partial^i}{\nabla^2}\,\dot{F}_{i\nu}+\frac{1}{\alpha}\,\partial_\mu A^\mu\eta_{0\nu}\right]\delta A^\nu+\frac{m^2}{2}\int d^3\vec{x}\,\frac{1}{\Box}\,\frac{\partial^\mu}{\nabla^2}\,F_{\mu\nu}\delta\dot{A}^\nu\\&-\frac{m^2}{2}\int d^3\vec{x}\,\frac{1}{\Box}\left(\frac{1}{\nabla^2}\,F_{0\nu}+\frac{\partial^i}{\nabla^4}\,\dot{F}_{i\nu}\right)\delta\ddot{A}^\nu+\frac{m^2}{2}\int d^3\vec{x}\,\frac{1}{\Box}\,\frac{\partial^\mu}{\nabla^4}\,F_{\mu\nu}\delta\dddot{A}^\nu+\cdots\;.\end{aligned}\tag{3.3}$$

The coefficient of δA^ν in the first term is the equation of motion, namely,

$$\left(1+\frac{m^2}{\Box}\right)\partial^\mu F_{\mu\nu}+\frac{1}{\alpha}\,\partial_\mu\partial_\nu A^\mu=0.\tag{3.4}$$

In the remaining terms, the coefficients of δA^ν, $\delta\dot{A}^\nu$, $\delta\ddot{A}^\nu$, etc., are the canonical momentum conjugate to A^ν, $\dot{A}^\nu$, $\ddot{A}^\nu$, etc. Denoting these momenta by π_ν, $\pi_\nu^{(1)}$, $\pi_\nu^{(2)}$, etc., we have

$$\begin{aligned}\pi_\nu&=-\left(1+\frac{m^2}{\Box}\right)F_{0\nu}-\frac{m^2}{2\Box}\,\frac{\partial^i}{\nabla^2}\,\dot{F}_{i\nu}-\frac{1}{\alpha}\,\partial_\mu A^\mu\eta_{0\nu},\\ \pi_\nu^{(1)}&=\frac{m^2}{2\Box}\,\frac{\partial^\mu}{\nabla^2}\,F_{\mu\nu},\\ \pi_\nu^{(2)}&=-\frac{m^2}{2\Box}\,\frac{1}{\nabla^2}\left(F_{0\nu}+\frac{\partial^i}{\nabla^2}\,\dot{F}_{i\nu}\right),\\ \pi_\nu^{(3)}&=\frac{m^2}{2\Box}\,\frac{\partial^\mu}{\nabla^4}\,F_{\mu\nu},\\ \pi_\nu^{(4)}&=-\frac{m^2}{2\Box}\,\frac{1}{\nabla^4}\left(F_{0\nu}+\frac{\partial^i}{\nabla^2}\,\dot{F}_{i\nu}\right),\\ &\;\vdots\,.\end{aligned}\tag{3.5}$$

Systems with higher derivatives have fields and their velocities as independent coordinates. For example, a system with two derivatives has its fields (denoting them generically by ϕ) and their velocities $\dot{\phi}$ as independent coordinates. If there are no constraints in the theory, the Poisson brackets (PB) are the bridge to the quantum commutator. Thus, we must have $[\phi,\dot{\phi}]=0$. The commutators that might not be zero are those (in this example with two derivatives) involving ϕ and $\dot{\phi}$ with the higher derivatives $\ddot{\phi}$ and $\dddot{\phi}$.[11]

The problem that comes out in the system we are studying is that there is an infinite number of time derivatives and it is not clear *a priori* which commutators are not zero. In order to try to figure them out we take the Lagrangian (3.1), expanded in t derivatives, until a certain limit order n and at the end we let n go to infinity. Let us then consider the expansion (3.2) until ∂_t^2, which is the first nontrivial order,

$$\mathcal{L}_3=-\frac{1}{4}\,F_{\mu\nu}F^{\mu\nu}+\frac{m^2}{4}\,F_{\mu\nu}\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{1}{\nabla^2}\,F^{\mu\nu}-\frac{1}{2\alpha}\,(\partial_\mu A^\mu)^2.\tag{3.6}$$

The equation of motion and the momenta are given by

$$\partial^\mu F_{\mu\nu} - m^2\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{\partial^\mu}{\nabla^2}F_{\mu\nu} + \frac{1}{\alpha}\,\partial_\mu\partial_\nu A^\mu = 0, \tag{3.7}$$

$$\pi_\nu = -F_{0\nu} + m^2\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{1}{\nabla^2}F_{0\nu} + \frac{m^2}{2}\,\frac{\partial^i\partial_t}{\nabla^4}F_{i\nu} - \frac{1}{\alpha}\,\eta_{0\nu}\partial_\mu A^\mu, \tag{3.8}$$

$$\pi_\nu^{(1)} = -\frac{m^2}{2}\,\frac{\partial^\mu}{\nabla^4}F_{\mu\nu}, \tag{3.9}$$

$$\pi_\nu^{(2)} = \frac{m^2}{2}\,\frac{1}{\nabla^4}F_{0\nu}. \tag{3.10}$$

In this order, the phase-space coordinate is given by $(A_\mu, \pi^\nu)\oplus(\dot{A}_\mu, \pi^{(1)\nu})\oplus(\ddot{A}_\mu, \pi^{(2)\nu})$. We thus observe that relations (3.9) and (3.10) are constraints, as well as the zero component of π_ν. The fundamental nonvanishing PB are

$$\begin{aligned}
&\{A_\mu(\vec{x},t), \pi^\nu(\vec{y},t)\} = \delta_\mu^\nu\,\delta(\vec{x}-\vec{y}),\\
&\{\dot{A}_\mu(\vec{x},t), \pi^{(1)\nu}(\vec{y},t)\} = \delta_\mu^\nu\,\delta(\vec{x}-\vec{y}),\\
&\{\ddot{A}_\mu(\vec{x},t), \pi^{(2)\nu}(\vec{y},t)\} = \delta_\mu^\nu\,\delta(\vec{x}-\vec{y}).
\end{aligned} \tag{3.11}$$

In order to calculate the PB matrix of the constraints, it is convenient to develop them separating all the velocities. The result is

$$T_0 = \pi_0 + \left(\frac{1}{\alpha} + \frac{m^2}{2\nabla^2}\right)\dot{A}_0 + \frac{m^2}{2\nabla^4}\,\partial^i\ddot{A}_i + \frac{1}{\alpha}\,\partial_i A^i, \tag{3.12}$$

$$T_\nu^{(1)} = \pi_\nu^{(1)} - \frac{m^2}{2\nabla^4}\,[\nabla^2 A_\nu + \delta_\nu^0\partial_i\dot{A}^i - \delta_\nu^i(\ddot{A}_i - \partial_i\dot{A}_0 - \partial_i\partial_j A^j)], \tag{3.13}$$

$$T_\nu^{(2)} = \pi_\nu^{(2)} + \delta_\nu^i\,\frac{m^2}{2\nabla^4}\,(\partial_i A_0 - \dot{A}_i). \tag{3.14}$$

We observe that the last constraint for $\nu=0$ becomes $T_0^{(2)} = \pi_0^{(2)}$. We also observe that the other constraints do not contain $\ddot{A}_0$ and consequently the PB matrix for the constraints above will be singular (in fact, $\ddot{A}_0$ does not play any role in the theory). Thus, instead of the constraint (3.14), we take

$$T_i^{(2)} = \pi_i^{(2)} + \frac{m^2}{2\nabla^4}\,(\partial_i A_0 - \dot{A}_i). \tag{3.15}$$

The PB matrix of the constraints reads

$$S=\begin{pmatrix} 0 & m^2\delta_0^\nu\left(\frac{1}{\alpha}+\frac{1}{\nabla^2}\right) & m^2\,\frac{\partial^j}{\nabla^4} \\ -m^2\delta_\mu^0\left(\frac{1}{\alpha}+\frac{1}{\nabla^2}\right) & -m^2(\delta_\mu^0\eta^{k\nu}+\delta_\mu^k\eta^{0\nu})\,\frac{\partial_k}{\nabla^4} & m^2\delta_\mu^j\,\frac{1}{\nabla^4} \\ m^2\,\frac{\partial_1}{\nabla^4} & -m^2\delta_i^\nu\,\frac{1}{\nabla^4} & 0 \end{pmatrix}\delta(\vec{x}-\vec{y}). \tag{3.16}$$

Since this matrix involves space and time indices separately, the calculation of its inverse requires some care. The details of the calculation are presented in Appendix B. The result reads

$$S^{-1}=\begin{pmatrix} 0 & -\alpha\delta_0^\nu & \alpha\partial^j \\ \alpha\delta_\mu^0 & -\alpha(\delta_\mu^0\delta_k^\nu\partial^k+\delta_\mu^k\delta_0^\nu\partial_k) & -\delta_\mu^k\left(\delta_k^j\,\frac{\nabla^4}{m^2}-\alpha\partial_k\partial^j\right) \\ \alpha\partial_i & \delta_k^\nu\left(\delta_i^k\,\frac{\nabla^4}{m^2}-\alpha\partial^k\partial_i\right) & 0 \end{pmatrix}\delta(\vec{x}-\vec{y}). \tag{3.17}$$

With this inverse, we directly obtain the following Dirac brackets (DB):[12]

$$\begin{aligned} \{\dot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)\}_D&=\alpha\delta_\mu^0\delta_0^\nu\delta(\vec{x}-\vec{y}),\\ \{\ddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)\}_D&=\alpha\delta_\mu^i\delta_0^\nu\partial_i\delta(\vec{x}-\vec{y}). \end{aligned} \tag{3.18}$$

The bracket $\{\dddot{A}_\mu,A^\nu\}$ is obtained from $\{\pi_\mu,A^\nu\}$. The result is

$$\{\dddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)\}=-\delta_\mu^\nu\,\frac{\nabla^4}{m^2}\,\delta(\vec{x}-\vec{y}). \tag{3.19}$$

The commutators follow directly from expressions (3.18) and (3.19), i.e.,

$$\begin{aligned} [\dot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]&=i\alpha\delta_\mu^0\delta_0^\nu\delta(\vec{x}-\vec{y}),\\ [\ddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]&=i\alpha\delta_\mu^i\delta_0^\nu\partial_i\delta(\vec{x}-\vec{y}),\\ [\dddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]&=-i\delta_\mu^\nu\,\frac{\nabla^4}{m^2}\,\delta(\vec{x}-\vec{y}). \end{aligned} \tag{3.20}$$

It might be opportune and instructive to call our attention to the following fact. We notice that from the first commutator above we get $[\dot{A}_i(\vec{x},t),A^j(\vec{y},t)]=0$. Since there is no dependence of this result with the mass parameter m, it may appear that there is a conflict with the limit case when $m\to 0$, where the Maxwell theory is obtained. We know that this commutator is not zero in the Maxwell theory. What happens is that in the limit of $m\to 0$, the structure of constraints is not the same as in the massful case, and consequently, the results we obtain in one sector cannot be kept in the other. We find it important to explain this point with detail in order to reinforce the formalism we are using. We use Appendix C to do this.

Let us now consider the propagator calculation. One can directly show by using the path integral formalism that the propagator corresponds to the inverse of the operator that appears in the equation of motion. Considering (3.7), we have

$$\left\{\eta_{\mu\nu}\left[1-\left(1+\frac{\partial_t^2}{\nabla^2}\right)\,\frac{m^2}{\nabla^2}\right]\Box+\left[\frac{1}{\alpha}-1+\left(1+\frac{\partial_t^2}{\nabla^2}\right)\,\frac{m^2}{\nabla^2}\right]\partial_\mu\partial_\nu\right\}A^\nu=0. \tag{3.21}$$

This means that the propagator must satisfy the equation

$$\left\{ \eta_{\mu\nu}\left[1-\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{m^2}{\nabla^2}\right]\Box+\left[\frac{1}{\alpha}-1+\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{m^2}{\nabla^2}\right]\partial_\mu\partial_\nu\right\} T(A^\nu(x)A^\rho(x'))=i\,\delta^\rho_\mu\,\delta(\vec{x}-\vec{x}\,'). \tag{3.22}$$

If what we have done until now is consistent, that is to say, if the quantization is embodied in the commutators (3.20), the expression (3.22) ought to be verified. In fact, after a hard algebraic calculation, we show that this actually occurs (see Appendix D for some details).

Let us now consider the Lagrangian with the next term of the expansion of $\Box^{-1}$,

$$\mathcal{L}_4=-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}+\frac{m^2}{4}F_{\mu\nu}\left(1+\frac{\partial_t^2}{\nabla^2}+\frac{\partial_t^4}{\nabla^4}\right)\frac{1}{\nabla^2}F^{\mu\nu}-\frac{1}{2\alpha}(\partial_\mu A^\mu)^2. \tag{3.23}$$

Proceeding as before we obtain the equation of motion

$$\partial^\mu F_{\mu\nu}-m^2\left(1+\frac{\partial_t^2}{\nabla^2}+\frac{\partial_t^4}{\nabla^4}\right)\frac{\partial^\mu}{\nabla^2}F_{\mu\nu}+\frac{1}{\alpha}\partial_\mu\partial_\nu A^\mu=0 \tag{3.24}$$

and the momentum expressions

$$\pi_\nu=-F_{0\nu}+m^2\left(1+\frac{\partial_t^2}{\nabla^2}+\frac{\partial_t^4}{\nabla^4}\right)\frac{1}{\nabla^2}F_{0\nu}+\frac{m^2}{2}\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{\partial^i\partial_t}{\nabla^4}F_{i\nu}-\frac{1}{\alpha}\eta_{0\nu}\partial_\mu A^\mu,$$

$$\pi^{(1)}_\nu=-\frac{m^2}{2}\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{\partial^\mu}{\nabla^4}F_{\mu\nu},\qquad \pi^{(2)}_\nu=\frac{m^2}{2}\left(1+\frac{\partial_t^2}{\nabla^2}\right)\frac{1}{\nabla^4}F_{0\nu}+\frac{m^2}{2}\frac{\partial_t\partial^i}{\nabla^6}F_{i\nu}, \tag{3.25}$$

$$\pi^{(3)}_\nu=-\frac{m^2}{2\nabla^6}\partial^\mu F_{\mu\nu},\qquad \pi^{(4)}_\nu=\frac{m^2}{2\nabla^6}F_{0\nu}.$$

The set of independent constraints is now given by

$$T_0=\pi_0+\frac{1}{\alpha}\partial_i A^i+\left(\frac{m^2}{2\nabla^2}+\frac{1}{\alpha}\right)\dot{A}_0+\frac{m^2}{2}\frac{\partial^i}{\nabla^4}\ddot{A}_i+\frac{m^2}{2\nabla^4}\dddot{A}_0+\frac{m^2}{2}\frac{\partial^i}{\nabla^6}\ddddot{A}_i,$$

$$T^{(1)}_\mu=\pi^{(1)}_\mu-\frac{m^2}{2\nabla^2}\left(A_\mu+\delta^i_\mu\frac{\partial_i\partial_j}{\nabla^2}A^j\right)-\frac{m^2}{2}\delta^0_\mu\frac{\partial^i}{\nabla^4}\dot{A}_i-\frac{m^2}{2}\delta^i_\mu\frac{\partial_i}{\nabla^4}\dot{A}_0-\frac{m^2}{2}\delta^0_\mu\frac{1}{\nabla^4}\ddot{A}_0$$
$$-\frac{m^2}{2}\delta^i_\mu\frac{\partial_i\partial_j}{\nabla^6}\ddot{A}^j-\frac{m^2}{2}\delta^0_\mu\frac{\partial^i}{\nabla^6}\dddot{A}_i-\frac{m^2}{2}\delta^i_\mu\frac{\partial_i}{\nabla^6}\dddot{A}_0+\frac{m^2}{2}\delta^i_\mu\frac{1}{\nabla^6}\ddddot{A}_i,$$

$$T^{(2)}_i=\pi^{(2)}_i+\frac{m^2}{2}\frac{\partial_i}{\nabla^4}A_0+\frac{m^2}{2}\frac{\partial_i\partial_j}{\nabla^6}\dot{A}^j+\frac{m^2}{2}\frac{\partial_i}{\nabla^4}\ddot{A}_0-\frac{m^2}{2\nabla^6}\dddot{A}_i, \tag{3.26}$$

$$T^{(3)}_i=\pi^{(3)}_i-\frac{m^2}{2\nabla^4}A_i-\frac{m^2}{2}\frac{\partial_i\partial_j}{\nabla^6}A^j-\frac{m^2}{2}\frac{\partial_i}{\nabla^6}\dot{A}_0+\frac{m^2}{2\nabla^6}\ddot{A}_i,$$

$$T^{(4)}_i=\pi^{(4)}_i+\frac{m^2}{2}\frac{\partial_i}{\nabla^6}A_0-\frac{m^2}{2\nabla^6}\dot{A}^i,$$

where, as before, velocities were conveniently separated. In the last constraint, we have not considered the index $\mu=0$ because there is no other term involving $\ddot{\ddot{A}}_0$. We have not also considered the zero components of $T_\mu^{(2)}$ and $T_\mu^{(3)}$ because these components do not constitute independent constraints.

With these constraints, we calculate the DB involving fields and their velocities, in the same way we have done in the previous approximation. The quantization of the present approximation is expressed by the following commutators:

$$\begin{gathered}
[\dot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=i\alpha\delta_\mu^0\delta_0^\nu\delta(\vec{x}-\vec{y}),\\
[\ddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=i\alpha\delta_\mu^k\delta_0^\nu\partial_k\delta(\vec{x}-\vec{y}),\\
[\dddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=0,\quad [\ddot{\ddot{A}}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=0,\\
[\overset{(v)}{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=-i\delta_\mu^\nu\frac{\nabla^6}{m^2}\,\delta(\vec{x}-\vec{y}),
\end{gathered}\tag{3.27}$$

where $\overset{(v)}{A}_\mu$ stands for five time derivatives over A_μ. Using the commutators above, one can also show that the propagator satisfies a similar relation like (3.23) with the operator that appears in the equation of motion (3.25). This shows that the commutators above are also consistent relations.

Now it is not difficult to infer the commutator relations when all the terms of the operator $\Box^{-1}$ are taken into account. These are given by

$$\begin{gathered}
[\dot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=i\alpha\delta_\mu^0\delta_0^\nu\delta(\vec{x}-\vec{y}),\\
[\ddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=i\alpha\delta_\mu^k\delta_0^\nu\partial_k\delta(\vec{x}-\vec{y}),\\
[\dddot{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=0,\\
[\ddot{\ddot{A}}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=0,\\
[\overset{(v)}{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)]=0,\\
\vdots\\
\lim_{n\to\infty}([\overset{(2n-1)}{A}_\mu(\vec{x},t),A^\nu(\vec{y},t)])=-i\delta_\mu^\nu\frac{\nabla^{2n}}{m^2}\,\delta(\vec{x}-\vec{y}).
\end{gathered}\tag{3.28}$$

We see in this way that the canonical structure, despite its nonlocality, is perfectly consistent with the functional procedure, generating propagators for the vectorial theory which display the presence of a massive field. To develop the theory furthermore, trying to construct the Fock space by introducing creation and annihilation operators for the vectorial fields, this seems to be nontrivial. This is so because there is no way, if m does not vanish, to avoid a canonical dependence between the vectorial field and its derivative of order $2n-1$, in the limit when n goes to infinity. We may say that the set of equations (3.28) shows us where the nonlocability problem appears in the process of quantization of these theories.

IV. CONCLUSION

In this work we have considered the quantization of a nonlocal massive vector gauge invariant field theory, which can be effectively obtained from a vector–tensor theory with topological coupling. We have quantized this nonlocal system first by using the BV Lagrangian functional formalism, where the propagator could be obtained without major problems. After that, we have considered its canonical quantization, where the nonlocalibility becomes a more difficult problem to be circumvented. The nonlocalibility manifests itself through the canonical independence, at commutator level, between the gauge field and its derivative of order n, in the limit when n goes to infinity. We have shown, however, that a systematic use of the canonical quantization procedure order by order permitted us to generate the same Greens functions as those obtained from the functional formalism. On the other hand, the Fock space structure seems difficult to be displayed, due to the odd canonical structure generated by the system. As it would be expected, when the mass parameter goes to zero, the tensor and vector sector of the theory decouple, and in the effective vector theory new constraints arise as a consequence of this limit. We have also shown in the Appendix C that a careful analysis of these constraints leads to a canonical structure that is identical to the usual massless gauge theory.

We could argue about the functional Hamiltonian quantization, due to Batalin, Fradkin, and Vilkovisky (BFV),[13] of this nonlocal system. The use of this formalism here appears to be a nontrivial task. Even with the procedure of how to circumvent the infinity number of momenta, we still have an additional problem because velocities have to be considered as independent canonical coordinates in higher derivative systems. Consequently, it is necessary to distinguish in the Hamiltonian path integral formalism what is a time derivative of a coordinate and what is an independent coordinate itself.[14] This problem is presently under study and possible results shall be reported elsewhere.[15]

ACKNOWLEDGMENTS

This work is supported in part by Conselho Nacional de Desenvolvimento Cientifico e Tecnológico—CNPq, Financiadora de Estudos e Projetos—FINEP, and Fundação Universitária José Bonifácio—FUJB (Brazilian Research Agencies).

APPENDIX A: OBTAINMENT OF EQ. (3.3)

Considering the Lagrangian (3.1), we have for a general variation of the corresponding action,

$$\delta\int_{t_0}^{t} d\tau \int d^3\vec{x}\ \mathcal{L} = -\int_{t_0}^{t} d\tau \int d^3\vec{x}\bigg[\frac{1}{2}\,\partial_\mu \delta A_\nu\left(1+\frac{m^2}{\Box}\right)F^{\mu\nu} + \frac{1}{2}\,F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)\partial^\mu \delta A^\nu + \frac{1}{\alpha}\,(\partial_\mu A^\mu)\partial_\nu \delta A^\nu\bigg]. \tag{A1}$$

Let us consider each term of the above expression in a separate way. The development of the first term leads to

$$\begin{aligned} -\frac{1}{2}\int_{t_0}^{t} d\tau \int d^3\vec{x}\ \partial_\mu \delta A_\nu\left(1+\frac{m^2}{\Box}\right)F^{\mu\nu} &= -\frac{1}{2}\int_{t_0}^{t} d\tau \int d^3\vec{x}\left\{\partial_\mu\left[\delta A_\nu\left(1+\frac{m^2}{\Box}\right)F^{\mu\nu}\right] - \delta A_\nu\left(1+\frac{m^2}{\Box}\right)\partial_\mu F^{\mu\nu}\right\} \\ &= -\frac{1}{2}\int d^3\vec{x}\ \delta A_\nu\left(1+\frac{m^2}{\Box}\right)F^{0\nu} + \frac{1}{2}\int_{t_0}^{t} d\tau \int d^3\vec{x}\ \delta A_\nu\left(1+\frac{m^2}{\Box}\right)\partial_\mu F^{\mu\nu}. \end{aligned} \tag{A2}$$

For the second term, we have

$$
\begin{aligned}
-\frac{1}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\; F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)\partial^\mu\delta A^\nu
&=-\frac{1}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\left\{\partial^\mu\left[F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)\delta A^\nu\right]-\partial^\mu F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)\delta A^\nu\right\}\\
&=-\frac{1}{2}\int d^3\vec{x}\; F_{0\nu}\left(1+\frac{m^2}{\Box}\right)\delta A^\nu+\frac{1}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\partial^\mu F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)\delta A_\nu\\
&=-\frac{1}{2}\int d^3\vec{x}\; F_{0\nu}\delta A^\nu+\frac{1}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\partial^\mu F_{\mu\nu}\delta A^\nu+\frac{m^2}{2}\int d^3\vec{x}\; F_{0\nu}\left(\frac{1}{\nabla^2}+\frac{\partial_t^2}{\nabla^4}+\cdots\right)\delta A^\nu\\
&\quad-\frac{m^2}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\,\partial^\mu F_{\mu\nu}\left(\frac{1}{\nabla^2}+\frac{\partial_t^2}{\nabla^4}+\cdots\right)\delta A^\nu, \qquad \text{(A3)}
\end{aligned}
$$

where we have used the expansion (3.2). We observe that in the last term of Eq. (A3), there is an integration over time and an infinite number of time derivatives acting over δA^ν. It is necessary to use some care to deal with these terms. Let us consider some of them separately,

$$
\begin{aligned}
-\frac{m^2}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\partial^\mu F_{\mu\nu}\frac{\partial_t^2}{\nabla^4}\,\delta A^\nu
&=-\frac{m^2}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\left\{\partial_t\left[\frac{\partial^\mu}{\nabla^4}F_{\mu\nu}\partial_t\partial A^\nu\right]-\frac{\partial_t\partial^\mu}{\nabla^4}F_{\mu\nu}\partial_t\delta A^\nu\right\}\\
&=-\frac{m^2}{2}\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^4}F_{\mu\nu}\delta\dot{A}^\nu+\frac{m^2}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\left\{\partial_t\left[\frac{\partial_t\partial^\mu}{\nabla^4}F_{\mu\nu}\delta A^\nu\right]-\frac{\partial_t^2\partial^\mu}{\nabla^4}F_{\mu\nu}\delta A^\nu\right\}\\
&=-\frac{m^2}{2}\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^4}F_{\mu\nu}\delta\dot{A}^\nu+\frac{m^2}{2}\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^4}\dot{F}_{\mu\nu}\delta A^\nu-\frac{m^2}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^4}\ddot{F}_{\mu\nu}\delta A^\nu.
\end{aligned}
\qquad \text{(A4)}
$$

In a similar way, we would have for the next term

$$
\begin{aligned}
-\frac{m^2}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\partial^\mu F_{\mu\nu}\frac{\partial_t^4}{\nabla^6}\,\delta A^\nu
&=-\frac{m^2}{2}\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^6}F_{\mu\nu}\delta\dddot{A}^\nu+\frac{m^2}{2}\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^6}\dot{F}_{\mu\nu}\delta\ddot{A}^\nu-\frac{m^2}{2}\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^6}\ddot{F}_{\mu\nu}\delta\dot{A}^\nu\\
&\quad+\frac{m^2}{2}\int d^3\vec{x}\;\frac{\partial^\mu}{\nabla^6}\dddot{F}_{\mu\nu}\delta A^\nu-\frac{m^2}{2}\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\frac{\partial_t^4}{\nabla^6}\,\partial^\mu F_{\mu\nu}\delta A^\nu, \qquad \text{(A5)}
\end{aligned}
$$

and so on. Introducing these results into the initial expression (A3), we obtain

$$
\begin{aligned}
&-\frac{1}{2}\int_{t_0}^{t} d\tau \int d^3\vec{x}\, F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)\partial^\mu \delta A^\nu \\
&\quad=\frac{1}{2}\int_{t_0}^{t} d\tau \int d^3\vec{x}\left(1+\frac{m^2}{\Box}\right)\partial^\mu F_{\mu\nu}\delta A^\nu \\
&\qquad+\frac{1}{2}\int d^3\vec{x}\left(-F_{0\nu}+m^2\frac{1}{\nabla^2}F_{0\nu}+m^2\frac{\partial^\mu}{\nabla^4}\dot{F}_{\mu\nu}+\cdots\right)\delta A^\nu \\
&\qquad-\frac{m^2}{2}\int d^3\vec{x}\left(\frac{\partial^\mu}{\nabla^4}F_{\mu\nu}+\frac{\partial^\mu}{\nabla^6}\ddot{F}_{\mu\nu}+\frac{\partial^\mu}{\nabla^8}\ddot{\ddot{F}}_{\mu\nu}+\cdots\right)\delta\dot{A}^\nu \\
&\qquad+\frac{m^2}{2}\int d^3\vec{x}\left(\frac{1}{\nabla^4}F_{0\nu}+\frac{\partial^\mu}{\nabla^6}\dot{F}_{\mu\nu}+\frac{\partial^\mu}{\nabla^8}\dddot{F}_{\mu\nu}+\cdots\right)\delta\ddot{A}^\nu \\
&\qquad-\frac{m^2}{2}\int d^3\vec{x}\left(\frac{\partial^\mu}{\nabla^6}F_{\mu\nu}+\frac{\partial^\mu}{\nabla^8}\ddot{F}_{\mu\nu}+\frac{\partial^\mu}{\nabla^{10}}\ddot{\ddot{F}}_{\mu\nu}+\cdots\right)\delta\dddot{A}^\nu+\cdots. \qquad \text{(A6)}
\end{aligned}
$$

We notice in the expression above that some terms can be put together to reobtain the nonlocal operator $\Box^{-1}$. For example,

$$
\begin{aligned}
\frac{1}{\nabla^2}F_{0\nu}+\frac{\partial^\mu}{\nabla^4}\dot{F}_{\mu\nu}+\frac{\partial^\mu}{\nabla^6}\dddot{F}_{\mu\nu}+\cdots&=\left(\frac{1}{\nabla^2}+\frac{\partial_t^2}{\nabla^4}+\cdots\right)F_{0\nu}+\left(\frac{1}{\nabla^4}+\frac{\partial_t^2}{\nabla^6}+\cdots\right)\partial^i\dot{F}_{i\nu} \\
&=-\frac{1}{\Box}F_{0\nu}-\frac{1}{\nabla^2}\frac{\partial^i}{\Box}\dot{F}_{i\nu}, \qquad \text{(A7)}
\end{aligned}
$$

$$
\frac{\partial^\mu}{\nabla^4}F_{\mu\nu}+\frac{\partial^\mu}{\nabla^6}\ddot{F}_{\mu\nu}+\frac{\partial^\mu}{\nabla^8}\ddot{\ddot{F}}_{\mu\nu}+\cdots=-\frac{1}{\nabla^2}\frac{\partial^\mu}{\Box}F_{\mu\nu} \qquad \text{(A8)}
$$

and so on. Using these results into (A6), we obtain the final form of the second term of (A1),

$$
\begin{aligned}
&-\frac{1}{2}\int_{t_0}^{t} d\tau \int d^3\vec{x}F_{\mu\nu}\left(1+\frac{m^2}{\Box}\right)\partial^\mu \delta A^\nu \\
&\quad=\frac{1}{2}\int_{t_0}^{t} d\tau \int d^3\vec{x}\left(1+\frac{m^2}{\Box}\right)\partial^\mu F_{\mu\nu}\delta A^\nu \\
&\qquad-\frac{1}{2}\int d^3\vec{x}\left[\left(1+\frac{m^2}{\Box}\right)F_{0\nu}+\frac{m^2}{\nabla^2}\frac{\partial^i}{\Box}F_{i\nu}\right]\delta A^\nu+\frac{m^2}{2}\int d^3\vec{x}\,\frac{1}{\nabla^2}\frac{\partial^\mu}{\Box}F_{\mu\nu}\delta\dot{A}^\nu \\
&\qquad-\frac{m^2}{2}\int d^3\vec{x}\left(\frac{1}{\nabla^2}\frac{1}{\Box}F_{0\nu}+\frac{1}{\nabla^4}\frac{\partial^i}{\Box}\dot{F}_{i\nu}\right)\delta\ddot{A}^\nu \\
&\qquad+\frac{m^2}{2}\int d^3\vec{x}\,\frac{1}{\nabla^4}\frac{\partial^\mu}{\Box}F_{\mu\nu}\delta\dddot{A}^\nu+\cdots. \qquad \text{(A9)}
\end{aligned}
$$

We finally consider the last term of expression (A1),

$$-\frac{1}{\alpha}\int_{t_0}^{t}d\tau\int d^3\vec{x}\;\partial_\mu A^\mu\partial_\nu\delta A^\nu=-\frac{1}{\alpha}\int_{t_0}^{t}d\tau\int d^3\vec{x}[\partial_\nu(\partial_\mu A^\mu\delta A^\nu)-\partial_\nu\partial_\mu A^\mu A^\nu]$$
$$=-\frac{1}{\alpha}\int d^3\vec{x}\;\partial_\mu A^\mu\delta A^0+\frac{1}{\alpha}\int_{t_0}^{t}d\tau\int d^3\vec{x}\partial_\nu\partial_\mu A^\mu\delta A^\nu. \tag{A10}$$

Introducing the results given by expressions (A2), (A9) and (A10) into the initial expression (A1), Eq. (3.3) is obtained.

APPENDIX B: CALCULATION OF THE INVERSE OF THE MATRIX (3.16)

First we notice that matrix (3.16) has the following block structure:

$$S=\begin{pmatrix}(1\times1) & (1\times4) & (1\times3)\\ (4\times1) & (4\times4) & (4\times3)\\ (3\times1) & (3\times4) & (3\times3)\end{pmatrix}. \tag{B1}$$

Of course, since the inverse S^{-1} has the same block structure, we consider it is generically given by

$$S^{-1}=\begin{pmatrix}A & B^\rho & C^k\\ D_\nu & E^\rho_\nu & F^k_\nu\\ G_j & H^\rho_j & I^k_j\end{pmatrix}. \tag{B2}$$

We then must have

$$\int d^3\vec{y}\;S(\vec{x},\vec{y})S^{-1}(\vec{y}-\vec{z})=\begin{pmatrix}1 & 0 & 0\\ 0 & \delta^\rho_\mu & 0\\ 0 & 0 & \delta^k_j\end{pmatrix}\delta(\vec{x}-\vec{z}). \tag{B3}$$

The combination of Eqs. (3.16), (B2), and (B3) gives us the following set of equations (after integrating over the intermediary variable $\vec{y}$ and summing on mudding indices):

$$\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)D_0+\frac{m^2}{\nabla^4}\;\partial^jG_j=\delta(\vec{x}-\vec{z}),$$

$$\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)E^\rho_0+\frac{m^2}{\nabla^4}\;\partial^jH^\rho_j=0, \tag{B4}$$

$$\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)F^k_0+\frac{m^2}{\nabla^4}\;\partial^jI^k_j=0,$$

$$\delta^0_\mu\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)A+\delta^0_\mu\,\frac{m^2}{\nabla^4}\,\partial_jD^j+\delta^j_\mu\,\frac{m^2}{\nabla^4}\,\partial_jD^0-\delta^j_\mu\,\frac{m^2}{\nabla^4}\,G_j=0,$$

$$\delta^0_\mu\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)B^\rho+\delta^0_\mu\,\frac{m^2}{\nabla^4}\,\partial_jE^\rho_j+\delta^j_\mu\,\frac{m^2}{\nabla^4}\,\partial_jE^\rho_0-\delta^j_\mu\,\frac{m^2}{\nabla^4}\,H^\rho_j=\delta^\rho_\mu\delta(\vec{x}-\overline{y}), \tag{B5}$$

$$\delta^0_\mu\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)C^k+\delta^0_\mu\,\frac{m^2}{\nabla^4}\,\partial_jF^k_j+\delta^j_\mu\,\frac{m^2}{\nabla^4}\,\partial_jF^k_0-\delta^j_\mu\,\frac{m^2}{\nabla^4}\,I^k_j=0,$$

$$\partial_iA-D_i=0,$$

$$\partial_iB^\rho-E^\rho_i=0, \tag{B6}$$

$$\partial_iC^k-F^k_i=\frac{\nabla^4}{m^2}\,\delta^k_i\,\delta(\vec{x}-\vec{z}),$$

$$\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)A+\frac{m^2}{\nabla^4}\,\partial^jD_j=0,$$

$$\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)B^0+\frac{m^2}{\nabla^4}\,\partial^jE^0_j=-\,\delta(\vec{x}-\vec{z}),$$

$$\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)B^k+\frac{m^2}{\nabla^4}\,\partial^jE^k_j=0,$$

$$\left(\frac{m^2}{\nabla^2}+\frac{1}{\alpha}\right)C^k+\frac{m^2}{\nabla^4}\,\partial^jF^k_j=0,$$

$$\partial_iE^0_0-H^0_i=0, \tag{B7}$$

$$\partial_iE^k_0-H^k_i=-\,\frac{\nabla^4}{m^2}\,\delta^k_i\,\delta(\vec{x}-\vec{z}),$$

$$\partial_iD_0-G_i=0,$$

$$\partial_iF^k_0-I^k_i=0.$$

The inverse S^{-1} is obtained by solving these equations. This is just a matter of algebraic work and the solution is

$$A=0,$$

$$B^0=-\,\alpha\,\delta(\vec{x}-\vec{z}),$$

$$B^k=0,$$

$$C^k=\alpha\,\partial^k\,\delta(\vec{x}-\vec{z}),$$

$$D_0=\alpha\,\delta(\vec{x}-\vec{z}),$$

$$D_i=0,$$

$$
\begin{aligned}
&E_0^0=0,\\
&E_i^0=-\alpha\partial_i\delta(\vec{x}-\vec{z}),\\
&E_0^k=-\alpha\partial^k\delta(\vec{x}-\vec{z}),\\
&E_i^k=0,\\
&F_0^k=0,\\
&F_i^k=\left(\alpha\partial_i\partial^k-\delta_i^k\,\frac{\nabla^4}{m^2}\right)\delta(\vec{x}-\vec{z}),\\
&G_j=\alpha\partial_j\delta(\vec{x}-\vec{z}),\\
&H_j^0=0,\\
&H_j^k=-\left(\alpha\partial_j\partial^k-\delta_j^k\,\frac{\nabla^4}{m^2}\right)\delta(\vec{x}-\vec{z}),\\
&I_j^k=0.
\end{aligned}
\tag{B8}
$$

These are the elements of the matrix S^{-1} given by (3.17).

APPENDIX C: LIMIT $m\to 0$ OF EXPRESSION (3.5)

When we take the limit $m\to 0$ in expressions (3.5), we get the following expressions for the momenta:

$$
\begin{aligned}
&\pi_\nu=-F_{0\nu}-\frac{1}{\alpha}\,\eta_{0\nu}\partial_\mu A^\mu,\\
&\pi_\nu^{(1)}=0.
\end{aligned}
\tag{C1}
$$

The remaining momenta (which are all zero) do not make sense to be considered because in the limit $m\to 0$ the system does not have infinite derivatives anymore. We observe that relations (C1) are constraints. So, the commutators cannot come from the PB of A_μ and $\dot{A}^\nu$, that is actually zero, but from the Dirac one. Let us calculate the DB of A_μ and $\dot{A}_\nu$. First, we need the PB matrix of the constraints. We denote these constraints by

$$
\begin{aligned}
&T_{1\mu}=\pi_\mu+F_{0\mu}+\frac{1}{\alpha}\,\eta_{0\mu}\partial_\nu A^\nu,\\
&T_{2\mu}=\pi_\mu^{(1)}.
\end{aligned}
\tag{C2}
$$

Thus,

$$
(S^{\mu\nu})=\begin{pmatrix}\{T_1^\mu,T_1^\nu\} & \{T_1^\mu,T_2^\nu\}\\ \{T_2^\mu,T_1^\nu\} & \{T_2^\mu,T_2^\nu\}\end{pmatrix}=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\left(\eta^{\mu\nu}+\frac{1-\alpha}{\alpha}\,\eta^{0\mu}\eta^{0\nu}\right)\delta(\vec{x}-\vec{y}). \tag{C3}
$$

To calculate the DB we need the inverse of the matrix above. This is given by

$$(S_{\mu\nu})^{-1}=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}(\eta_{\mu\nu}+(\alpha-1)\eta_{0\mu}\eta_{0\nu})\,\delta(\vec{x}-\vec{y}). \tag{C4}$$

Now, the DB can be directly calculated. The result is

$$\{A_\mu(\vec{x},t),\dot{A}^\nu(\vec{y},t)\}=-(\delta_\mu^\nu+(\alpha-1)\eta_{0\mu}\delta_0^\nu)\,\delta(\vec{x}-\vec{y}). \tag{C5}$$

The commutator between A_i and $\dot{A}^j$ can be directly obtained from the DB above and it is actually nonzero when $m\to 0$, which makes consistent the procedure we are developing.

APPENDIX D: VERIFICATION OF THE IDENTITY (3.22)

Considering that

$$T(A^\nu(x)A^\rho(x'))=\theta(t-t')A^\nu(x)A^\rho(x')+\theta(t'-t)A^\rho(x')A^\nu(x) \tag{D1}$$

and using the commutators given by expression (3.20) we can obtain the following relations:

$$\begin{gathered}
\frac{\partial^2}{\partial t^2}\,T(A^\nu(x)A^\rho(x'))=i\alpha\delta_0^\nu\eta^{0\rho}\delta(x-x')+T(\ddot{A}^\nu(x)A^\rho(x')),\\
\nabla^2T(A^\nu(x)A^\rho(x'))=T(\nabla^2A^\nu(x)A^\rho(x')),\\
\Box T(A^\nu(x)A^\rho(x'))=i\alpha\delta_0^\nu\eta^{0\rho}\delta(x-x')+T(\Box A^\nu(x)A^\rho(x')),\\
\frac{1}{\nabla^2}\,\Box T(A^\nu(x)A^\rho(x'))=i\alpha\delta_0^\nu\eta^{0\rho}\,\frac{1}{\nabla^2}\,\delta(x-x')+T\left(\frac{1}{\nabla^2}\,\Box A^\nu(x)A^\rho(x')\right),\\
\frac{\partial_t^2}{\nabla^2}\,\Box T(A^\nu(x)A^\rho(x'))=i\left[\alpha\delta_0^\nu\eta^{0\rho}\,\frac{1}{\nabla^2}\,\Box+\alpha\eta^{\nu i}\delta_0^\rho\,\frac{\partial_t\partial_i}{\nabla^2}-\eta^{\nu\rho}\,\frac{\nabla^2}{m^2}\right]\delta(x-x')\\
+T\left(\frac{\partial_t^2}{\nabla^2}\,\Box A^\nu(x)A^\rho(x')\right),\\
\partial_\mu\partial_\nu T(A^\nu(x)A^\rho(x'))=i\alpha\delta_\mu^0\delta_0^\rho\delta(x-x')+T(\partial_\mu\partial_\nu A^\nu(x)A^\rho(x')),\\
\frac{\partial_\mu\partial_\nu\partial_t^2}{\nabla^4}\,T(A^\nu(x)A^\rho(x'))=i\left[\alpha\delta_\mu^0\delta_0^\rho\,\frac{\partial_t^2}{\nabla^4}+\alpha\delta_\mu^i\delta_0^\rho\,\frac{\partial_i\partial_t}{\nabla^4}-\delta_\mu^0\delta_0^\rho\,\frac{1}{m^2}-\alpha\delta_\mu^0\delta_0^\rho\,\frac{1}{\nabla^2}\right]\delta(x-x')\\
+T\left(\frac{\partial_\mu\partial_\nu\partial_t^2}{\nabla^4}\,A^\nu(x)A^\rho(x')\right).
\end{gathered} \tag{D2}$$

Using the relations above in the left-hand side of Eq. (3.22) we can show that the identity is actually satisfied.

[1] E. Cremmer and J. Scherk, Nucl. Phys. B **72**, 117 (1974); see also, T. J. Allen, M. J. Bowick, and A. Lahiri, Mod. Phys. Lett. A **6**, 559 (1991).
[2] M. Kalb and P. Ramond, Phys. Rev. D **9**, 2273 (1974).
[3] R. Amorim and J. Barcelos-Neto, Mod. Phys. Lett. A **10**, 917 (1995).
[4] R. K. Kaul, Phys. Rev. D **18**, 1127 (1978); C. R. Hagen, *ibid.* **19**, 2367 (1979); V. O. Rivelles and L. Sandoval, Rev. Bras. Fis. **21**, 274 (1991); A. Lahiri, Mod. Phys. Lett. A **8**, 2403 (1993); J. Barcelos-Neto and M. B. D. Silva, Int. J. Mod. Phys. A **10**, 3759 (1995); Mod. Phys. Lett. A **11**, 515 (1996).
[5] See also M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, 1992); J. Gomis, J. París, and S. Samuel, Phys. Rep. **259**, 1 (1995), and references therein.
[6] D. Z. Freedam and P. K. Townsend, Nucl. Phys. B **177**, 282 (1981).

[7] J. Barcelos-Neto, A. Cabo, and M. B. D. Silva, Z. Phys. C **72**, 34 (1996); A. Lahiri, Phys. Rev. D **55**, 5045 (1997); D. S. Hwang and C.-Y. Lee, J. Math. Phys. **38**, 30 (1997).
[8] We mention that quantization of nonlocal theories has been carried out by other authors, but not in the same context that will be presented here. See, for example, E. C. Marino and R. L. P. G. Amaral, J. Phys. A **25**, 5183 (1992); D. G. Barci, L. E. Oxman, and M. Rocca, Int. J. Mod. Phys. A **11**, 2111 (1996).
[9] I. A. Batalin and G. A. Vilkovisky, Phys. Lett. B **102**, 27 (1981); Phys. Rev. D **28**, 2567 (1983).
[10] For a detailed discussion on systems with higher derivatives, see D. Musik, *Degenerate Systems in Generalized Mechanics* (preprint) (Publ. Inst. Math., Beogrado, 1992), and references therein.
[11] J. Barcelos-Neto and N. R. F. Braga, Mod. Phys. Lett. A **4**, 2195 (1989).
[12] P. A. M. Dirac, Can. J. Math. **2**, 129 (1950); *Lectures on Quantum Mechanics* (Yeshiva University Press, New York, 1964).
[13] G. A. Vilkovisky, Phys. Lett. **55B**, 224 (1975); I. A. Batalin and G. A. Vilkovisky, *ibid.* **69B**, 309 (1977); E. S. Fradkin and T. E. Fradkina, *ibid.* **72B**, 343 (1978).
[14] J. Barcelos-Neto and C. P. Natividade, Z. Phys. C **51**, 313 (1991).
[15] R. Amorim and J. Barcelos-Neto (unpublished).

Modular groups of quantum fields in thermal states

H. J. Borchers
Institut für Theoretische Physik, Universität Göttingen, Bunsenstrasse 9, D 37073 Göttingen, Germany

J. Yngvason[a)]
Institut für Theoretische Physik, Universität Wien, Boltzmanngasse 5, A 1090 Wien, Austria

(Received 10 August 1998; accepted for publication 28 August 1998)

For a quantum field in a thermal equilibrium state we discuss the group generated by time translations and the modular action associated with an algebra invariant under half-sided translations. The modular flows corresponding to the algebras of the forward light cone and a space-like wedge admit a simple geometric description in two-dimensional models that factorize in light-cone coordinates. At large distances from the domain boundary compared to the inverse temperature, the flow pattern is essentially the same as time translations, whereas the zero temperature results are approximately reproduced close to the edge of the wedge and the apex of the cone. For each domain there is also a one-parameter group with a positive generator, for which the thermal state is a ground state. Formally, this may be regarded as a certain converse of the Unruh effect. © *1999 American Institute of Physics.* [S0022-2488(99)01002-6]

I. INTRODUCTION

Algebraic quantum field theory in the sense of Araki, Haag, and Kastler[1] is concerned with von Neumann algebras $\mathcal{M}(\mathcal{O})$ of observables localized in space time domains $\mathcal{O}$, together with states ω on these algebras satisfying some physical selection criterion. Due to the Reeh–Schlieder property of quantum field theory,[2] one may associate with certain regions $\mathcal{O}$ and states ω the Tomita–Takesaki modular objects $\Delta_{\mathcal{O},\omega}$ and $J_{\mathcal{O},\omega}$.[3,4] The positive operator $\Delta_{\mathcal{O},\omega}$ generates a one-parameter group ad $\Delta^{it}_{\mathcal{O},\omega}$ of automorphisms of $\mathcal{M}(\mathcal{O})$, and the conjugation ad $J_{\mathcal{O},\omega}$ defined by the antiunitary $J_{\mathcal{O},\omega}$ maps $\mathcal{M}(\mathcal{O})$ onto its commutant on the GNS Hilbert space corresponding to ω.

Important structural properties of the theory are encoded in the modular objects (see, e.g., Refs. 5, 6) but an explicit description of $\Delta^{it}_{\mathcal{O},\omega}$ and $J_{\mathcal{O},\omega}$ has so far only been obtained in the following cases with ω a vacuum state.

(a) $\mathcal{O}$ is a space-like wedge and the local algebras are generated by Wightman fields that transform covariantly with a finite-dimensional representation of the Lorentz group.[7,8]
(b) $\mathcal{O}$ is a forward light cone and $\mathcal{M}(\mathcal{O})$ is generated by a massless, noninteracting field.[9]
(c) $\mathcal{O}$ is a double cone and $\mathcal{M}(\mathcal{O})$ is generated by conformally covariant fields.[10]
(d) $\mathcal{O}$ is a space-like wedge and the local algebras are generated by generalized free fields of a certain type that break Lorentz covariance.[11]

In case (a), the modular group is the group of Lorentz boosts that leave the wedge invariant, and the conjugation is the PCT operator (combined with a rotation). Cases (b) and (c) can be reduced to case (a) by a conformal mapping onto the wedge of the forward light cone and the double cone, respectively. In (b) the modular group is the dilation group, and in (c) it consists of the conformal transformations that leave the double cone invariant. In the examples considered in

[a)]Electronic mail: yngvason@Thor.thp.univie.ac.at

(d), the action of the modular group is, in general, nonlocal, i.e., an algebra $\mathcal{M}(\mathcal{O}_1)$ with $\mathcal{O}_1$ a bounded subset of the wedge need not be mapped into an $\mathcal{M}(\mathcal{O}_2)$ with $\mathcal{O}_2$ bounded.

A key to a general understanding of possible geometric interpretations of modular groups is the interplay between the modular action and certain subgroups of the space–time translations. In Ref. 12 it was shown that the modular group of a space-like wedge in a vacuum state acts on the translation group like the Lorentz boosts that leave the wedge invariant. Subsequently, Wiesbrock[13] introduced the concept of a half-sided modular inclusion and proved a certain converse of the results of Ref. 12, namely, that the two-dimensional translation group can be recovered from the modular groups of the wedge and some of its translates.

In this paper we want to investigate the modular groups when ω is a thermodynamic equilibrium state (KMS state) rather than a vacuum state. In Sec. II we discuss the generalizations of the results of Ref. 12 to KMS states. We investigate the commutation relations between the time translations and the modular group in a KMS state for any domain that is mapped into itself under half-sided time translations. Using the results of Ref. 6 we prove that the time translations and modular action together give rise to a representation of the abstract Lie group generated by one-dimensional dilations and translations. The important observation that half-sided modular actions always lead to a representation of this group was first made by Wiesbrock.[13,14] We express all its one-parameter subgroups in terms of the translations and the modular group. Of particular interest is a subgroup with a positive generator. This group acts on the global observable algebra for positive values of the group parameter.

The group relations alone do not determine the modular action and the group with a positive generator, but they put definite restrictions on the possible disclocalization of observables by the group actions. More precisely, if $\mathcal{N}$ denotes the observable algebra of a domain invariant under half-sided translations and $\mathcal{N}(t)$ its time translate by t, then the modular group $\Delta_{\mathcal{N}}^{iu}$ of $\mathcal{N}$ transforms $\mathcal{N}(t)$ into $\mathcal{N}(\varphi(u,t))$ with a certain function $\varphi(u,t)$. Likewise, the group with a positive generator transforms $\mathcal{N}(t)$ into $\mathcal{N}(\psi(\tau,t))$, where ψ is another function of t and the group parameter τ. The precise statements are given in Theorem 2.1. We also discuss the action of $\Delta_{\mathcal{N}}^{iu}$ on individual observables in $\mathcal{N}(t)$ for t large and show that, in a sense made precise in Theorems 2.2 and 2.3, this action approximates a time translation by $-\beta u$ as $t/\beta \to \infty$.

In Sec. III we consider two-dimensional (2-D) models that factorize in the light-cone coordinates. Applying the results of the previous section to the algebras on each of the light rays, one obtains a geometric description of the actions of the groups associated with the forward light cone and a space-like wedge. In the case of the forward light cone, the algebra of a translated light cone is mapped into another such algebra. An analogous statement holds for the wedge. The flow patterns are illustrated in Figs. 1 and 2. Close to the apex of the light cone and the edge of the wedge the action of the modular flow is essentially the same as for the zero temperature case, i.e., dilations for the forward light cone and Lorentz boosts for the wedge. On the other hand, at large distances from the domain boundary compared to the inverse temperature the modular flow approaches the dynamical flow, i.e., the time translations.

The one-parameter unitary group with a positive generator associated with the forward light cone, which in the limiting case of zero temperature reduces to time translations, approximates the dynamical flow close to the apex of the light cone. It corresponds everywhere to a decelerated movement toward the origin in the space variable (Fig. 3). Formally at least, this may be regarded as a reverse Unruh effect:[15–17] In the latter the vacuum appears as a KMS state with respect to a dynamics that accelerates points toward light-like infinity; here a KMS state appears as a vacuum state with respect to a dynamics that moves points from light-like infinity toward the origin of space.

For the wedge there is also a unitary group with a positive generator, which has the KMS state as a ground state. This group operates on the observables for a restricted parameter range. It approximates the time translations close to the space axis and light-like translations far away from the space axis. The action of this group is illustrated in Fig. 4. This action may also be interpreted as a kind of reverse Unruh effect, because here the acceleration is away from the wedge, whereas in the usual Unruh effect the acceleration points in the direction of the wedge.

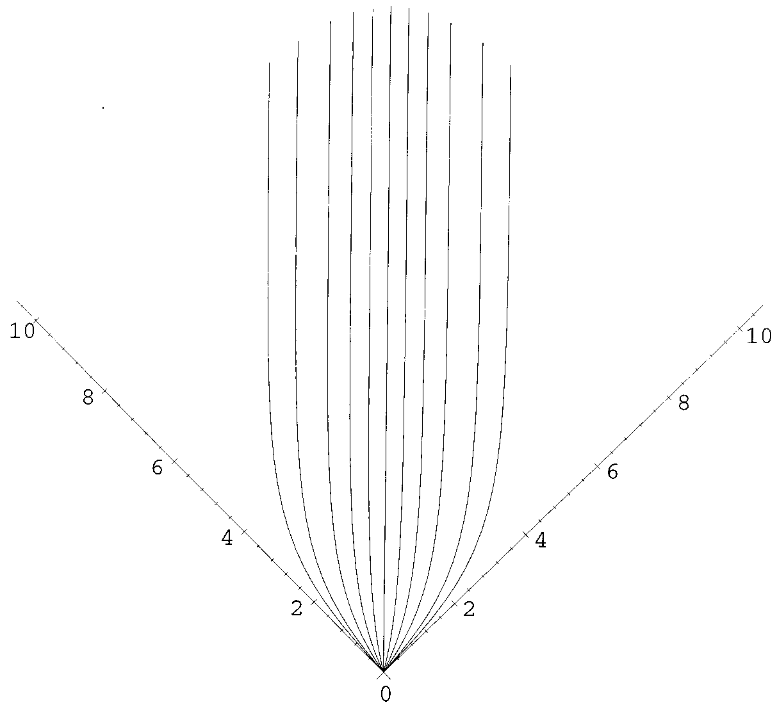

FIG. 1. The modular flow in the forward light cone. The unit is inverse temperature, β.

In Sec. IV we compute explicitly the modular groups and the groups with a positive generator for a quasifree KMS state on the Weyl algebra of a generalized free field in 2-D space time that factorizes in light-cone coordinates. For a field of minimal scaling dimension one obtains a strengthening of the general results of the previous section on the group actions: A local algebra $\mathcal{M}(\mathcal{O})$ with $\mathcal{O}$ a double cone is transformed into an algebra of the same kind. For fields of higher scaling dimension, however, double cone localization may get lost under the group action, and only a localization in a translated light cone or wedge remains.

II. THE GROUP GENERATED BY TRANSLATIONS AND THE MODULAR ACTION

Let $(\mathcal{A},\alpha_t)$ be a C^*-dynamical system and $\mathcal{B}$ a subalgebra, such that

$$\alpha_t\mathcal{B}\subset\mathcal{B}, \quad \text{for } t\geqslant 0. \tag{2.1}$$

Suppose, furthermore, that the algebra $\cup_{t\in\mathbf{R}}\alpha_t\mathcal{B}$ is norm dense in $\mathcal{A}$. Let ω be a KMS state[18] for the dynamical system $(\mathcal{A},\alpha_t)$ at inverse temperature β and denote by π the corresponding GNS representation of $\mathcal{A}$ with cyclic vector Ω, and by $T(t)$ the unitary implementation of α_t on the GNS Hilbert space $\mathcal{H}$. Put $\mathcal{M}=\pi(\mathcal{A})''$ and $\mathcal{N}=\pi(\mathcal{B})''$.

Because of the analyticity properties of the time translations in a KMS state, the vector Ω is separating for $\mathcal{M}$ and hence also for $\mathcal{N}$. Moreover, Ω is cyclic for $\mathcal{M}$ (by definition), and since $\cup_{t\in\mathbf{R}}\alpha_t\mathcal{B}$ is dense in $\mathcal{A}$ it follows by a Reeh–Schlieder-type argument that Ω is also cyclic for $\mathcal{N}$.

Let $\Delta_{\mathcal{M}}$ and $J_{\mathcal{M}}$ be the modular objects corresponding to Ω and $\mathcal{M}$. We have

$$\Delta_{\mathcal{M}}^{\mathrm{is}}=T(-\beta s), \tag{2.2}$$

where the sign is a consequence of different conventions in physics and mathematics: For $A\in\mathcal{M}$ the expression $T(t)A\Omega$ has an analytic continuation into the strip $S(0,\beta/2)$, where

$$S(a,b):=\{z\in\mathbb{C}:a<\mathrm{Im}\, z<b\}, \tag{2.3}$$

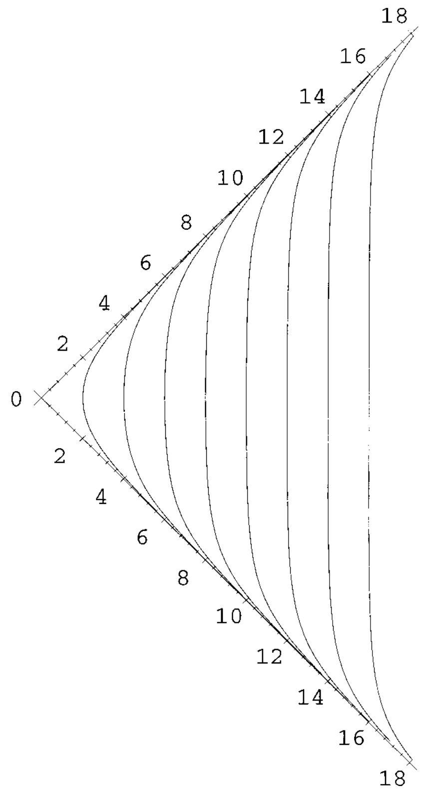

FIG. 2. The modular flow in a space-like wedge.

while $\Delta_{\mathcal{M}}^{is}A\Omega$ has an analytic continuation into $S(-1/2,0)$, by the sign convention in modular theory. Since $J_{\mathcal{M}}\Delta_{\mathcal{M}}^{1/2}A\Omega = A^*\Omega$, it follows from (2.2) that

$$T(t+i\beta)A\Omega = J_{\mathcal{M}}T(t)A^*\Omega. \tag{2.4}$$

By assumption (2.1), we have

$$T(t)\mathcal{N}T(-t)\subset\mathcal{N}, \quad \text{for } t\geqslant 0, \tag{2.5}$$

i.e., we are in the situation of a half-sided translation in the sense of Ref. 12. Because of (2.2) we are also in the situation of a half-sided modular inclusion in the sense of Wiesbrock,[13] i.e.,

$$\Delta_{\mathcal{M}}^{is}\mathcal{N}\Delta_{\mathcal{M}}^{-is}\subset\mathcal{N}, \quad \text{for } s\leqslant 0. \tag{2.6}$$

If $T(t)$ had a positive generator, then (2.5) would imply the well-known relations[12] between $T(t)$ and the modular group $\Delta_{\mathcal{N}}^{iu}$. In a KMS state, however, the spectrum of the Hamiltonian is the whole real axis and the analysis of Ref. 12 has to be generalized. The main results of this generalization are stated in Eqs. (2.20), (2.29), and (2.31) below.

We start with a heuristic discussion of the consequences of (2.6), similar to that in Ref. 13. This discussion disregards questions of domains of unbounded operators, but it leads quickly to the commutation relations between $\Delta_{\mathcal{M}}^{is}$ and $\Delta_{\mathcal{N}}^{iu}$ stated in Ref. 14. A rigorous proof of these relations follows from the results of Ref. 6 and will be given after the discussion.

Since $\mathcal{N}\subset\mathcal{M}$, it follows by standard arguments that $\Delta_{\mathcal{N}}\geqslant\Delta_{\mathcal{M}}$, and this, domain questions aside, implies that

$$G := \log\Delta_{\mathcal{N}} - \log\Delta_{\mathcal{M}} \tag{2.7}$$

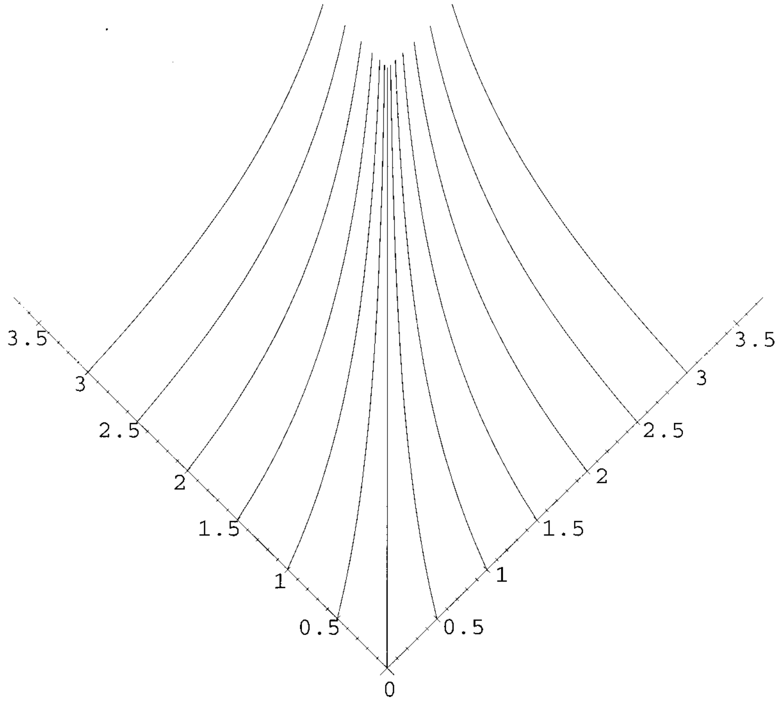

FIG. 3. The flow of $\Gamma_{V^+}(\tau)$ within the forward light cone. The whole pattern is invariant under translations in the x^0 direction.

is a non-negative operator, because log is an operator monotone function. Equations (2.1), (2.2), (2.6), and the Trotter product formula now lead to

$$e^{i\tau G}\mathcal{N}e^{-i\tau G}\subset\mathcal{N}, \quad \text{for} \ \ \tau\geqslant 0. \tag{2.8}$$

Putting $U(\tau):=\exp(i\tau G)$, Eq. (2.8) and $G\geqslant 0$ imply[12]

$$\Delta_{\mathcal{N}}^{iu}U(\tau)\Delta_{\mathcal{N}}^{-iu}=U(e^{-2\pi u}\tau), \tag{2.9}$$

for all τ, $u\in\mathbf{R}$. Hence, we obtain a unitary representation of the two-parameter Lie group $\mathcal{G}$ with elements $(\tau,u)\in\mathbf{R}^2$ and the composition law

$$(\tau,u)\circ(\tau',u')=(\tau+e^{-2\pi u}\tau',u+u'). \tag{2.10}$$

The representation $U(\tau,u)$ corresponding to (2.9) is

$$U(\tau,u):=e^{i\tau G}\Delta_{\mathcal{N}}^{iu}. \tag{2.11}$$

The group $\mathcal{G}$ defined by (2.10) is the semidirect product of $\mathbf{R}$ with itself and is the unique two-dimensional non-Abelian Lie group (''$ax+b$ group''). Some of its properties are discussed in Ref. 19.

For a discussion of the one-parameter subgroups and the Lie algebra of $\mathcal{G}$, it is convenient to realize the group in terms of 2×2 matrices:

$$(\tau,u)\leftrightarrow\begin{pmatrix}1 & \tau\\ 0 & 1\end{pmatrix}\cdot\begin{pmatrix}e^{-2\pi u} & 0\\ 0 & 1\end{pmatrix}=\begin{pmatrix}e^{-2\pi u} & \tau\\ 0 & 1\end{pmatrix}. \tag{2.12}$$

It is straightforward to determine the one-parameter subgroups, $r\mapsto g(r)$ of $\mathcal{G}$. These have the form

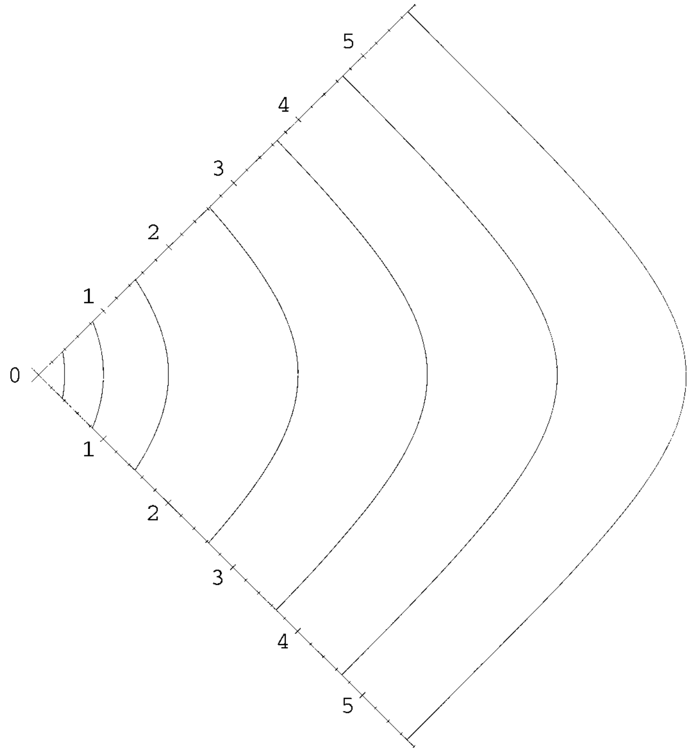

FIG. 4. The flow of $\Gamma_W(\tau)$ in a space-like wedge. The whole pattern is invariant under translations in the x^1 direction.

$$g_{a,b}(r)=\begin{pmatrix} e^{ar} & \frac{b}{a}(e^{ar}-1) \\ 0 & 1 \end{pmatrix}, \tag{2.13}$$

with $a,b\in\mathbf{R}$. In the half-plane $(\tau,e^{-2\pi u})\in\mathbf{R}\times\mathbf{R}_+$, these correspond to straight lines through (0,1). The infinitesimal generator of $g_{a,b}(r)$ is

$$\hat{g}_{a,b}=\frac{d}{dr}\, g_{a,b}(r)\bigg|_{r=0}=\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}. \tag{2.14}$$

The group $\Delta_{\mathcal{N}}^{iu}$ corresponds to $a=-2\pi$, $b=0$; the group $\exp(i\tau G)$ to $a=0$, $b=1$. Since the generator of $\Delta_{\mathcal{M}}^{is}$ is $\log\Delta_{\mathcal{N}}-G$, this one-parameter group corresponds to $a=-2\pi$, $b=-1$. Denoting for short the one-parameter subgroups of $\mathcal{G}$ in these three cases by $g_{\mathcal{N}}(u)$, $g_{\text{pos}}(\tau)$, and $g_{\mathcal{M}}(s)$, respectively, we have

$$g_{\mathcal{N}}(u)=\begin{pmatrix} e^{-2\pi u} & 0 \\ 0 & 1 \end{pmatrix},\quad g_{\text{pos}}(\tau)=\begin{pmatrix} 1 & \tau \\ 0 & 1 \end{pmatrix},\quad g_{\mathcal{M}}(s)=\begin{pmatrix} e^{-2\pi s} & \frac{1}{2\pi}(e^{-2\pi s}-1) \\ 0 & 1 \end{pmatrix}. \tag{2.15}$$

One verifies the relation

$$g_{\mathcal{N}}(u)\cdot g_{\mathcal{M}}(s)=g_{\mathcal{M}}(F(u,s))\cdot g_{\mathcal{N}}(-F(u,s)+s+u), \tag{2.16}$$

with

$$F(u,s) = -\frac{1}{2\pi}\log\{1+e^{-2\pi u}(e^{-2\pi s}-1)\}, \tag{2.17}$$

provided

$$1+e^{-2\pi u}(e^{-2\pi s}-1)>0, \tag{2.18}$$

which is always fulfilled for $s\leqslant 0$. The relation corresponding to (2.16) for the modular groups $\Delta_{\mathcal{N}}^{\mathrm{i}u}$ and $\Delta_{\mathcal{M}}^{\mathrm{i}s}$ is

$$\Delta_{\mathcal{N}}^{\mathrm{i}u}\cdot\Delta_{\mathcal{M}}^{\mathrm{i}s} = \Delta_{\mathcal{M}}^{\mathrm{i}F(u,s)}\cdot\Delta_{\mathcal{N}}^{\mathrm{i}(-F(u,s)+s+u)}. \tag{2.19}$$

This relation also appears in Ref. 14; our heuristic discussion has brought its group theoretical origin into focus.

In terms of the original translation group $T(t)$ we can, because of (2.2), write (2.19), as

$$\Delta_{\mathcal{N}}^{\mathrm{i}u}\cdot T(t)\cdot\Delta_{\mathcal{N}}^{-\mathrm{i}u} = T\left(\frac{\beta}{2\pi}\log\{1+e^{-2\pi u}(e^{2\pi t/\beta}-1)\}\right)\cdot\Delta_{\mathcal{N}}^{\mathrm{i}(1/2\pi)\log\{1+e^{-2\pi u}(e^{2\pi t/\beta}-1)\}-\mathrm{i}(t/\beta)}. \tag{2.20}$$

In the limit $\beta\to\infty$ we recover the Bisognano–Wichmann result,

$$\Delta_{\mathcal{N}}^{\mathrm{i}u}\cdot T(t)\cdot\Delta_{\mathcal{N}}^{-\mathrm{i}u} = T(e^{-2\pi u}t). \tag{2.21}$$

The one-parameter groups $g_{a,b}(r)$ can be expressed in terms of $g_{\mathcal{M}}(s)$ and $g_{\mathcal{N}}(u)$:

$$g_{a,b}(r) = g_{\mathcal{M}}(s(r))\cdot g_{\mathcal{N}}(u(r)) = g_{\mathcal{N}}(u(-r))\cdot g_{\mathcal{M}}(s(-r)), \tag{2.22}$$

with

$$s(r) = -\frac{1}{2\pi}\log\left\{1+\frac{2\pi b}{a}(e^{ar}-1)\right\},$$

$$u(r) = -s(r)-\frac{a}{2\pi}r. \tag{2.23}$$

Specializing to $a=0$, $b=1$ we obtain, for $r>-1/(2\pi)$,

$$g_{\mathrm{pos}}(\tau) = g_{\mathcal{M}}(-(2\pi)^{-1}\log(1+2\pi\tau))\cdot g_{\mathcal{N}}((2\pi)^{-1}\log(1+2\pi\tau)), \tag{2.24}$$

and hence

$$U(\tau) = \Delta_{\mathcal{M}}^{-(\mathrm{i}/2\pi)\log(1+2\pi r)}\cdot\Delta_{\mathcal{N}}^{(\mathrm{i}/2\pi)\log(1+2\pi r)} = \Delta_{\mathcal{N}}^{-(\mathrm{i}/2\pi)\log(1-2\pi\tau)}\cdot\Delta_{\mathcal{M}}^{(\mathrm{i}/2\pi)\log(1-2\pi\tau)}. \tag{2.25}$$

The first representation can be used for $\tau>-1/(2\pi)$, the second one for $\tau<1/(2\pi)$.

The group $g_{\mathcal{N}}(u)$ operates on $g_{\mathrm{pos}}(\tau)$ according to

$$g_{\mathcal{N}}(u)g_{\mathrm{pos}}(\tau)g_{\mathcal{N}}(-u) = g_{\mathrm{pos}}(\exp(-2\pi u)\tau), \tag{2.26}$$

which is just the abstract form of the basic relation (2.9). This is a special case of the general relation

$$g_{a,b}(r)g_{\mathrm{pos}}(\tau)g_{a,b}(-r) = g_{\mathrm{pos}}(\exp(ar)\tau). \tag{2.27}$$

For $a=-2\pi$, $b=-1$, i.e., $g_{\mathcal{M}}$, the corresponding relation for the unitary groups on Hilbert space is

$$\Delta_{\mathcal{M}}^{is}U(\tau)\Delta_{\mathcal{M}}^{-is}=U(\exp(-2\pi s)\tau), \tag{2.28}$$

which also follows directly from (2.25) and (2.19). We note in passing that (2.28) may be interpreted as an ''Anosov relation'' that leads to exponential clustering of matrix elements of the time translations $T(t)=\Delta_{\mathcal{M}}^{-it/\beta}$ in states of the form $A\Omega$ with A in a dense subalgebra of $\mathcal{M}$.[20]

Defining $\Gamma(\tau):=U(\tau/\beta)$ we have, by (2.25),

$$\Gamma(\tau)=T\left(\frac{\beta}{2\pi}\log\{1+(2\pi\tau/\beta)\}\right)\cdot\Delta_{\mathcal{N}}^{(i/2\pi)\log\{1+(2\pi\tau/\beta)\}}$$
$$=\Delta_{\mathcal{N}}^{-(i/2\pi)\log\{1-(2\pi\tau/\beta)\}}\cdot T\left(-\frac{\beta}{2\pi}\log\{1-(2\pi\tau/\beta)\}\right), \tag{2.29}$$

where the first equality is valid for $\tau>-\beta/(2\pi)$ and the second for $\tau<\beta/(2\pi)$. Evidently, $\Gamma(\tau)\rightarrow T(\tau)$ for $\beta\rightarrow\infty$, and

$$G/\beta=H+\frac{1}{\beta}\log\Delta_{\mathcal{N}}, \tag{2.30}$$

tends in this limit to the Hamiltonian H, which in the vacuum representation is $\geqslant 0$.

The relation (2.28) means that

$$T(t)\Gamma(\tau)T(-t)=\Gamma(\exp(2\pi t/\beta)\tau). \tag{2.31}$$

By (2.8) and our assumption that $\cup_t\alpha_t\mathcal{B}$ is norm dense in $\mathcal{A}$ (and hence $\cup$ ad $T(t)\mathcal{N}$ weakly dense in $\mathcal{M}$), we may thus conclude that

$$\text{ad }\Gamma(\tau)\mathcal{M}\subset\mathcal{M},\quad \text{for all}\quad \tau\geqslant 0. \tag{2.32}$$

A rigorous proof of the relations (2.19) and (2.25) [and hence of (2.20), (2.29), and (2.31)] can be obtained by applying Theorems A and B in Ref. 6 to the operator-valued functions,

$$V(v)=\Delta_{\mathcal{M}}^{-iv}\Delta_{\mathcal{N}}^{iv}, \tag{2.33}$$

and $W(w)=V(v(w))$, where

$$v(w)=\frac{1}{2\pi}\log(1+e^{2\pi w}). \tag{2.34}$$

The function $V(v)$ has a bounded analytic continuation into the strip $S(0,1/2)$ with continuous boundary values, and satisfies the relation

$$V\left(v+\frac{i}{2}\right)=J_{\mathcal{M}}V(v)J_{\mathcal{N}}, \tag{2.35}$$

for $v\in\mathbf{R}$. Moreover, ad $V(v)$ maps $\mathcal{N}$ into $\mathcal{N}$ for $v\geqslant 0$ and the commutant $\mathcal{N}'$ into $\mathcal{N}'$ for $v\leqslant 0$. By (2.35) it follows that ad $V(v+i/2)$ maps $\mathcal{N}'$ into $\mathcal{N}'$ for all v.

In order to apply Theorem B in Ref. 6 we have to map the strip $S(0,1/2)$ biholomorphically onto itself in such a way that $\mathbf{R}$ is mapped onto $\mathbf{R}_+$ and $\mathbf{R}+i/2$ onto $(\mathbf{R}+i/2)\cup\mathbf{R}_-$. The map (2.34) accomplishes this. It has a singularity at $w=i/2$, but as remarked in Ref. 6 such a singularity is harmless.

[The reason is as follows: Theorem B in Ref. 6 is based on the edge-of-the-wedge theorem, applied to matrix elements of the operator-valued function,

$$(u,w)\mapsto\Delta_{\mathcal{N}}^{iu}W(w)\Delta_{\mathcal{N}}^{-iu}. \tag{2.36}$$

These matrix elements have bounded analytic continuations, which are continuous at the boundary of their domain with the possible exception of points with $w = \mathrm{i}/2$. By the dominated convergence theorem and the boundedness of (2.36) this piecewise continuity is sufficient to ensure coincidence of boundary values in the sense of distributions. The edge-of-the-wedge theorem then implies analyticity in the coincidence region, so continuity in the points with $w = \mathrm{i}/2$ holds *a fortiori*.]

Theorem B in Ref. 6 leads to the general relations

$$\Delta_{\mathcal{N}}^{\mathrm{i}u} W(w) \Delta_{\mathcal{N}}^{-\mathrm{i}u} = W(w-u) \tag{2.37}$$

and

$$J_{\mathcal{N}} W(w) J_{\mathcal{N}} = W\left(w + \frac{\mathrm{i}}{2}\right). \tag{2.38}$$

Eq. (2.37) is precisely (2.19) in case (2.18) holds, but note that (2.37) is true for all $u, w \in \mathbf{R}$. As noted by Wiesbrock (Refs. 13, 14) these relations imply that $\Delta_{\mathcal{N}}^{\mathrm{i}u}$ and $\Delta_{\mathcal{M}}^{\mathrm{i}s}$ generate a unitary representation of the Lie group $\mathcal{G}$. The infinitesimal generators $\log \Delta_{\mathcal{N}}$ and $\log \Delta_{\mathcal{M}}$, together with their real linear combinations, are thus essentially self adjoint on a common core. The representation $U(\tau)$ of the one-parameter subgroup $g_{\mathrm{pos}}(\tau)$ fulfills, together with $\Delta_{\mathcal{N}}^{\mathrm{i}u}$, the relation (2.9) (because of the corresponding relation in $\mathcal{G}$), and this implies, by Ref. 21, that $U(\tau) = \exp(\mathrm{i}\tau G)$ with $G \geqslant 0$. Hence, the starting point of the heuristic discussion is rigorously justified.

The following theorem summarizes the main conclusions of the preceding discussion and states, in addition, the most important consequence of the relations (2.20), (2.29), and (2.31) for the present investigation, namely the action of the group $\mathcal{G}$ on translates of $\mathcal{N}$.

Theorem 2.1: *Let $(\mathcal{A}, \alpha_t)$ be a C^*-dynamical system and $\mathcal{B}$ a subalgebra such that $\alpha_t \mathcal{B} \subset \mathcal{B}$ for $t \geqslant 0$ and $\cup_t \alpha_t \mathcal{B}$ is norm dense in $\mathcal{A}$. In the GNS representation defined by a KMS state on $\mathcal{A}$ at inverse temperature β let $\mathcal{M}$ and $\mathcal{N}$ denote the weak closures of $\pi(\mathcal{A})$ and $\pi(\mathcal{B})$, respectively, and $T(t) = \exp(\mathrm{i}tH)$ the unitary group implementing α_t. Denote* ad $T(t)\mathcal{N} = \mathcal{N}(t)$. *Then*

(i) The translations $T(t)$ and the modular group $\Delta_{\mathcal{N}}^{\mathrm{i}u}$, defined by $\mathcal{N}$ and the KMS state vector, fulfill the relation (2.20). We have

$$\mathrm{ad}\ \Delta_{\mathcal{N}}^{\mathrm{i}u} \mathcal{N}(t) = \mathcal{N}(\varphi(u,t)), \tag{2.39}$$

with

$$\varphi(u,t) = \frac{\beta}{2\pi} \log\{1 + e^{-2\pi u}(e^{2\pi t/\beta} - 1)\}, \tag{2.40}$$

for all u, t, satisfying

$$1 + e^{-2\pi u}(e^{2\pi t/\beta} - 1) > 0. \tag{2.41}$$

In particular,

$$\mathrm{ad}\ \Delta_{\mathcal{N}}^{\mathrm{i}u} \mathcal{M} \subset \mathcal{M}, \tag{2.42}$$

for all $u \geqslant 0$, and

$$\mathcal{N} = \bigcap_{u \geqslant 0} \mathrm{ad}\ \Delta_{\mathcal{N}}^{\mathrm{i}u} \mathcal{M}. \tag{2.43}$$

(ii) The operator $G = \beta H + \log \Delta_{\mathcal{N}}$ is non-negative and essentially self-adjoint on a common core of H and $\log \Delta_{\mathcal{N}}$. The one parameter group $\Gamma(\tau) = \exp(i\tau G/\beta)$ is given by (2.29) and the groups $\Gamma(\tau)$ and $T(t)$ satisfy (2.31). We have

$$\text{ad } \Gamma(\tau)\mathcal{N}(t)=\mathcal{N}(\psi(u,t)), \tag{2.44}$$

with

$$\psi(\tau,t)=t+\frac{\beta}{2\pi}\log\left\{1+\frac{2\pi\tau}{\beta}e^{-2\pi t/\beta}\right\}, \tag{2.45}$$

for all τ, t *satisfying*

$$1+\frac{2\pi\tau}{\beta}e^{-2\pi t/\beta}<0. \tag{2.46}$$

In particular,

$$\text{ad } \Gamma(\tau)\mathcal{M}\subset\mathcal{M} \quad \text{and} \quad \text{ad } \Gamma(\tau)\mathcal{N}\subset\mathcal{N}, \tag{2.47}$$

for $\tau\geqslant 0$, *and*

$$\mathcal{N}=\text{ad } \Gamma(\beta/2\pi)\mathcal{M}. \tag{2.48}$$

Proof: As already noted, the key relations (2.20), (2.29), and (2.31), and the self-adjointness and positivity of G are a rigorous consequence of the Theorems in Refs. 6, 21, 13, and 14. Equations (2.39) and (2.44) follow directly from (2.20) and (2.29) and the fact that $\varphi(-u,\varphi(u,t))=t$, $\psi(-\tau,\psi(\tau,t))=t$ for (u,t) and (τ,t), satisfying (2.41) and (2.46), respectively. Equations (2.42), (2.43), (2.47), and (2.48) are simple consequences of (2.39) and (2.44), since $\cup_t$ ad $T(t)\mathcal{N}$ is dense in $\mathcal{M}$.

As a last topic in this section we discuss the relation between the translation group $T(t)$ and the modular group $\Delta_{\mathcal{N}}^{iu}$. Since $T(-\beta u)=\Delta_{\mathcal{M}}^{iu}$, one may expect that the actions of $T(-\beta u)$ and $\Delta_{\mathcal{N}}^{iu}$ approximately coincide on elements that have been translated far into $\mathcal{N}$, so that ''boundary effects'' are negligible. That this intuition is indeed solidly founded is the content of the next two theorems. The first concerns certain matrix elements of the unitary groups, and gives an estimate for the rate of the convergence. The second is about strong convergence of Hilbert space vectors and operators, but the error estimates are less explicit.

Theorem 2.2: *If* $A\in\mathcal{N}(t)$ *and* $B\in\mathcal{N}'$, *the following estimate holds for* $t>0$ *and all* u*:*

$$|(B\Omega,\Delta_{\mathcal{N}}^{iu}A\Omega)-(B\Omega,T(-\beta u)A\Omega)|\leqslant 2M\,\min\left\{\frac{|\exp(2\pi u)-1|}{\exp(2\pi t/\beta)-1},1\right\}, \tag{2.49}$$

with

$$M=\max\{\|A\Omega\|\|B\Omega\|,\|A^*\Omega\|\|B^*\Omega\|\}. \tag{2.50}$$

Proof: Consider the two functions

$$F^+(u)=(B\Omega,\Delta_{\mathcal{N}}^{-iu}T(-\beta u)A\Omega) \quad \text{and} \quad F^-(u)=(A^*\Omega,T(\beta u)\Delta_{\mathcal{N}}^{iu}B^*\Omega). \tag{2.51}$$

Theorem A in Ref. 6 implies that $\Delta_{\mathcal{N}}^{-iu}T(-\beta u)$ has a bounded analytic continuation into the strip $S(-\frac{1}{2},0)$. It follows that F^+ has an analytic continuation into $S(-\frac{1}{2},0)$, and F^- into $S(0,\frac{1}{2})$. Moreover, by continuity of the unitary groups, $F^\pm$ is continuous on the real axis.

Denoting $j_{\mathcal{M}}=\text{ad } J_{\mathcal{M}}$, $j_{\mathcal{N}}=\text{ad } J_{\mathcal{N}}$ we obtain

$$F^+\left(u-\frac{i}{2}\right)=(B\Omega,J_{\mathcal{N}}\,\Delta_{\mathcal{N}}^{-iu}T(-\beta u)J_{\mathcal{M}}A\Omega)=(\Delta_{\mathcal{N}}^{-iu}T(-\beta u)j_{\mathcal{M}}(A)\Omega,j_{\mathcal{N}}(B)\Omega),$$

$$F^{-}\left(u+\frac{\mathrm{i}}{2}\right)=(A^{*}\Omega,J_{\mathcal{M}}T(\beta u)\Delta_{\mathcal{N}}^{\mathrm{i}u}J_{\mathcal{N}}\,B^{*}\Omega)=(T(\beta u)\Delta_{\mathcal{N}}^{\mathrm{i}u}j_{\mathcal{N}}(B^{*})\Omega,j_{\mathcal{M}}(A^{*})\Omega).$$

In particular, $F^{\pm}$ is continuous at $u\pm \mathrm{i}/2$, $u\in\mathbf{R}$, and $j_{\mathcal{N}}(B^{*})\in\mathcal{N}$ and $j_{\mathcal{M}}(A^{*})\in\mathcal{M}'$ implies

$$F^{+}\left(u-\frac{\mathrm{i}}{2}\right)=F^{-}\left(u+\frac{\mathrm{i}}{2}\right). \tag{2.52}$$

Moreover, since $T(-s)AT(s)\in\mathcal{N}$ for $s<t$, we have

$$F^{+}(u)=F^{-}(u),\quad \text{for } u<t/\beta. \tag{2.53}$$

Hence, F^{+} and F^{-} have a common analytic continuation to a periodic function, F, with the period i and cuts $[t/\beta,\infty)+\mathrm{i}n$, $n\in\mathbf{Z}$. This function is majorized by

$$M=\max\{\|A\Omega\|\|B\Omega\|,\|A^{*}\Omega\|\|B^{*}\Omega\|\}. \tag{2.54}$$

The function $F(z)-F(0)$ vanishes at $z=\mathrm{i}n$, $n\in\mathbf{Z}$, and is bounded by $2M$. Therefore,

$$G(z)=\frac{F(z)-F(0)}{\exp(2\pi z)-1}$$

is analytic and bounded in the same domain as F. Along the cuts we have $|G(z)|\leq 2M(\exp(2\pi t/\beta)-1)^{-1}$. By the maximum modulus principle this estimate holds everywhere, and thus

$$|(B\Omega,\Delta_{\mathcal{N}}^{-\mathrm{i}u}T(-\beta u)A\Omega)-(B\Omega,A\Omega)|\leq 2M\,\frac{|\exp(2\pi u)-1|}{\exp(2\pi t/\beta)-1}. \tag{2.55}$$

This estimate blows up for $t\to 0$, but the left-hand side is trivially bounded by $2M$ for all real u and t. Replacing $B\in\mathcal{N}'$ by ad $\Delta_{\mathcal{N}}^{-\mathrm{i}u}B\in\mathcal{N}'$ does not change M, so (2.55) gives the desired estimate (2.49).

Theorem 2.3: *(i) For every $A\in\mathcal{M}$ and Hilbert space vector Ψ,*

$$\lim_{t\to\infty}\|\Delta_{\mathcal{N}}^{\mathrm{i}u}A(t)\Psi-T(-\beta u)A(t)\Psi\|=0, \tag{2.56}$$

with $A(t)=\mathrm{ad}\;T(t)A$. The convergence is uniform on half-sided u intervals $I=(-\infty,u_0]$, $u_0<\infty$.

(ii) For every A in a dense subalgebra of $\mathcal{M}$,

$$\lim_{t\to\infty}\|\mathrm{ad}\;\Delta_{\mathcal{N}}^{\mathrm{i}u}A(t)-\mathrm{ad}\;T(-\beta u)A(t)\|=0, \tag{2.57}$$

with uniform convergence on half-sided u intervals.

Proof: From (2.48), (2.2), and (2.31), it follows that

$$\Delta_{\mathcal{N}}^{\mathrm{i}u}=\Gamma(\beta/2\pi)T(-\beta u)\Gamma(-\beta/2\pi)=T(-\beta u)\Gamma((\exp(2\pi u)-1)\beta/2\pi). \tag{2.58}$$

Hence, using (2.31) again,

$$\mathrm{ad}\;\Delta_{\mathcal{N}}^{\mathrm{i}u}A(t)-\mathrm{ad}\;T(-\beta u)A(t)=\mathrm{ad}\;T(t-\beta u)[\mathrm{ad}\;\Gamma(\exp(-2\pi t/\beta)(u))A-A], \tag{2.59}$$

with $h(u)=(\exp(2\pi u)-1)\beta/2\pi$. Now (i) follows from the strong convergence of $\Gamma(\tau)$ to 1 as $\tau\to 0$, because $\sup_{u\in I}|h(u)|<\infty$ for $I=(-\infty,u_0]$.

For general $A\in\mathcal{M}$, $\|\mathrm{ad}\,\Gamma(\tau)A-A\|$ need not converge to zero as $\tau\to 0$. However, on elements of the form $A_g=\int g(\tau)\mathrm{ad}\,\Gamma(\tau)A\,d\tau$ with g continuous of compact support, this convergence holds. Moreover, if g is continuously differentiable, then (2.59) implies

$$\|\mathrm{ad}\,\Delta_{\mathcal{N}}^{\mathrm{i}u}A_g(t)-\mathrm{ad}\,T(-\beta u)A_g(t)\|\leqslant\|A\|\cdot\|dg/d\tau\|_1\sup_{u\in I}|h(u)|\cdot e^{-2\pi t/\beta}. \tag{2.60}$$

By (2.44) such regularized elements are dense in $\mathcal{M}$ if the support of g is sufficiently small, and $A_g\to A$ weakly if g tends to a delta function.

III. TWO-DIMENSIONAL MODELS

The general results of the preceding setting were formulated for a C^*-dynamical system $(\mathcal{A},\alpha_t)$ and a subalgebra $\mathcal{B}$, invariant under half-sided shifts by α_t. We shall now be more specific and consider a quasilocal algebra $\mathcal{A}$ generated by a local net $\mathcal{O}\mapsto\mathcal{A}(\mathcal{O})$ of C^*-algebras and $\mathcal{B}=\mathcal{A}(\mathcal{O}_0)$ with $\mathcal{O}_0$ a domain invariant under half-sided translations in the t direction. In the representation π generated by a KMS state ω, we denote $\pi(\mathcal{A}(\mathcal{O}))''$ by $\mathcal{M}(\mathcal{O})$ and $\pi(\mathcal{A})''$ by $\mathcal{M}$, as before.

Equations (2.39) and (2.44) describe the action of the modular and Γ groups associated with $\mathcal{N}=\mathcal{M}(\mathcal{O}_0)$ on the translated algebras $\mathcal{M}(\mathcal{O}_0+t\mathbf{e})$, where $\mathbf{e}$ is the unit vector in the t direction. We now want to investigate how the groups associated with $\mathcal{M}(\mathcal{O}_0)$ act on the algebras of more general domains than $\mathcal{O}_0+t\mathbf{e}$, in particular, $\mathcal{O}_0+\mathbf{x}$, with $\mathbf{x}$ an arbitrary vector in space–time. While a general answer to this question appears difficult, the previous results lead directly to a description of the action in the case of two-dimensional theories that factorize in the light-cone variables.

We start by considering local nets in two-dimensional space time depending only on one light-cone variable, i.e., nets on a light ray. With x^0 the time and x^1 the space coordinate of $\mathbf{x}\in\mathbf{R}^2$, the light-cone variables are $x^R=x^0+x^1$ and $x^L=x^0-x^1$. We consider either one of them and denote it simply by x. Note that translations in time $t=x^0$ are equivalent to translations in x. A local algebra corresponding to an x interval $I\subset\mathbf{R}$ is denoted by $\mathcal{M}(I)$. Local commutativity means that $\mathcal{M}(I_1)$ and $\mathcal{M}(I_2)$ commute if $I_1\cap I_2=\emptyset$.

We denote the modular group for the algebra $\mathcal{M}(\mathbf{R}_+)$ by $\Delta_+^{\mathrm{i}u}$ and the corresponding group with the positive generator $G_+/\beta=H+(1/\beta)\log\Delta_+$ by $\Gamma_+(\tau)$. We shall also consider the algebra of the negative half-axis, $\mathcal{M}(\mathbf{R}_-)$, with modular group $\Delta_-^{\mathrm{i}u}$ and the positive operator $G_-/\beta=H+(1/\beta)\log\Delta_-$, which generates the group $\Gamma_-(\tau)=\exp(\mathrm{i}\tau G_-/\beta)$. Note that $\mathrm{ad}\,\Gamma_-(\tau)$ maps $\mathcal{M}(\mathbf{R}_-)$ into itself for $\tau\leqslant 0$.

By Eq. (2.39), we have

$$\mathrm{ad}\,\Delta_+^{\mathrm{i}u}\mathcal{M}([x,\infty[)=\mathcal{M}([\varphi_+(u,x),\infty[), \tag{3.1}$$

with

$$\varphi_+(u,x)=\frac{\beta}{2\pi}\log\{1+e^{-2\pi u}(e^{2\pi x/\beta}-1)\}, \tag{3.2}$$

for all x, $u\in\mathbf{R}$, such that

$$1+\exp(-2\pi u)[\exp(2\pi x/\beta)-1]>0. \tag{3.3}$$

Note that (3.2) is just the function (2.40). We denote it here by φ_+ because there is an analogous result for $\mathcal{M}(\mathbf{R}_-)$:

$$\mathrm{ad}\,\Delta_-^{\mathrm{i}u}\mathcal{M}(]-\infty,x])=\mathcal{M}(]-\infty,\varphi_-(u,x)]), \tag{3.4}$$

with

$$\varphi_-(u,x) = -\varphi_+(-u,-x), \tag{3.5}$$

for

$$1+\exp(2\pi u)[\exp(-2\pi x/\beta)-1]>0. \tag{3.6}$$

Likewise, from Eq. (2.44),

$$\mathrm{ad}\ \Gamma_+(\tau)\mathcal{M}([x,\infty[) = \mathcal{M}([\psi_+(\tau,x),\infty[), \tag{3.7}$$

with

$$\psi_+(\tau,x) = x + \frac{\beta}{2\pi}\log\left\{1+\frac{2\pi\tau}{\beta}e^{-2\pi x/\beta}\right\}, \tag{3.8}$$

for

$$1+(2\pi\tau/\beta)\exp(-2\pi x/\beta)>0, \tag{3.9}$$

and

$$\mathrm{ad}\ \Gamma_-(\tau)\mathcal{M}(]-\infty,x]) = \mathcal{M}(]-\infty,\psi_-(u,x)]), \tag{3.10}$$

with

$$\psi_-(\tau,x) = -\psi_+(-\tau,-x), \tag{3.11}$$

for

$$1-(2\pi\tau/\beta)\exp(2\pi x/\beta)>0. \tag{3.12}$$

We now turn to models in two space–time dimensions that can be written as a tensor product of one-dimensional models in the light-cone variables, $x^R = x^0 + x^1$ and $x^L = x^0 - x^1$. For a domain $I_L \times I_R \subset \mathbf{R}^2$ with I_L and I_R intervals on the x_L and x_R axis, respectively, the local algebra is thus

$$\mathcal{M}(I_L\times I_R) = \mathcal{M}(I_L)\otimes\mathcal{M}(I_R). \tag{3.13}$$

Here $\otimes$ is the von Neumann tensor product, and $I \mapsto \mathcal{M}(I)$ is a local net of von Neumann algebras over $\mathbf{R}$. (For simplicity of notation we take identical nets on both axes.) In particular, we are interested in the algebras of the forward light cone,

$$\mathcal{M}(V^+) = \mathcal{M}(\mathbf{R}_+)\otimes\mathcal{M}(\mathbf{R}_+), \tag{3.14}$$

and the right wedge

$$\mathcal{M}(W) = \mathcal{M}(\mathbf{R}_-)\otimes\mathcal{M}(\mathbf{R}_+). \tag{3.15}$$

The modular groups for these algebras and a factorizing KMS state $\omega\otimes\omega$, where ω is a KMS state for the algebra on a light ray, are

$$\Delta_{V^+}^{\mathrm{i}u} = \Delta_+^{\mathrm{i}u}\otimes\Delta_+^{\mathrm{i}u} \tag{3.16}$$

and

$$\Delta_W^{\mathrm{i}u} = \Delta_-^{\mathrm{i}u}\otimes\Delta_+^{\mathrm{i}u}. \tag{3.17}$$

If $\mathbf{x}\in\mathbf{R}^2$, we denote the translated light cone $V^+ +\mathbf{x}$ by $V_{\mathbf{x}}^+$ and the translated wedge $W+\mathbf{x}$ by $W_{\mathbf{x}}$. From Eqs. (3.1) and (3.5), we obtain

Theorem 3.1:

$$\text{ad}\ \Delta_{V^+}^{\mathrm{i}u}\mathcal{M}(V_{\mathbf{x}}^+)=\mathcal{M}(V^+_{\varphi_{V^+(u,\mathbf{x})}}),\tag{3.18}$$

with

$$\varphi_{V^+}(u,\mathbf{x})=(\varphi_+(u,x^L),\varphi_+(u,x^R)),\tag{3.19}$$

for $u\in\mathbf{R}$ *and* $\mathbf{x}\in\mathbf{R}^2$ *such that (3.3) holds for* $x=x^L$ *and* $x=x^R$*. If* $u\geqslant 0$*, then* ad $\Delta_{V^+}^{\mathrm{i}u}\mathcal{M}(V_{\mathbf{x}}^+)\subset\mathcal{M}$ *for all* $\mathbf{x}\in\mathbf{R}^2$*, and if* $\mathbf{x}\in V^+$*, then* ad $\Delta_{V^+}^{\mathrm{i}u}\mathcal{M}(V_{\mathbf{x}}^+)\subset\mathcal{M}(V^+)$ *for all u.*

Likewise,

$$\text{ad}\ \Delta_W^{\mathrm{i}u}\mathcal{M}(W_{\mathbf{x}})=\mathcal{M}(W_{\varphi_W(u,\mathbf{x})}),\tag{3.20}$$

with

$$\varphi_W(u,\mathbf{x})=(\varphi_-(u,x^L),\varphi_+(u,x^R)),\tag{3.21}$$

for $u\in\mathbf{R}$ *and* $\mathbf{x}\in\mathbf{R}^2$ *such that (3.3) holds for* $x=x^R$ *and (3.6) for* $x=x^L$*. If* $\mathbf{x}\in W$*, then* ad $\Delta_W^{\mathrm{i}u}\mathcal{M}(W_x)\subset\mathcal{M}(W)$ *for all u.*

The flow lines of φ_{V^+} and φ_W within the respective domains are shown in Figs. 1–2.

It is evident from the figures that the character of the modular flow depends of the distance from the boundary of the domain considered (forward light cone or wedge). The natural unit of length is here the reciprocal temperature, β. Consider first the modular group of the forward light cone V^+. In terms of the original space time coordinates $x^0=(x^R+x^L)/2$ and $x^1=(x^R-x^L)/2$ the map (3.19) takes (x_0,x^1) to (x'^0,x'^1) with

$$x'^0=x^0-\beta u+R_{V^+}^0(x,u),\quad x'^1=x^1+R_{V^+}^1(x,u),\tag{3.22}$$

where

$$R_{V^+}^0(x,u)=(\beta/4\pi)\log\{(1+e^{-2\pi(x^R-\beta u)/\beta}-e^{-2\pi x^R/\beta})(1+e^{-2\pi(x^L-\beta u)/\beta}-e^{-2\pi x^L/\beta})\}\tag{3.23}$$

and

$$R_{V^+}^1(x,u)=(\beta/4\pi)\log\left\{\frac{1+e^{-2\pi(x^R-\beta u)/\beta}-e^{-2\pi x^R/\beta}}{1+e^{-2\pi(x^L-\beta u)/\beta}-e^{-2\pi x^L/\beta}}\right\}.\tag{3.24}$$

Far from the domain boundary, i.e., for x^R, x^L, $x^R-\beta u$, and $x^L-\beta u$ large compared to β, the terms $R_{V^+}^0$ and $R_{V^+}^1$ are exponentially small, and $\psi_{V^+}(\cdot,u)$ is essentially the same as translation in time by $-\beta u$ in accord with Theorems 2.2 and 2.3. On the other hand, close to the apex of the light cone (compared to β), the action is essentially the same as for $\beta=\infty$, i.e., dilation by the factor $\exp(-2\pi u)$. The deviation from a dilation is of the order $(|x|/\beta)^2$.

For the wedge W the formulas corresponding to (3.22)–(3.24) are

$$x'^0=x^0-\beta u+R_W^0(x,u),\quad x'^1=x^1+R_W^1(x,u),\tag{3.25}$$

with

$$R_W^0(x,u)=(\beta/4\pi)\log\left\{\frac{1+e^{-2\pi(x^R-\beta u)}-e^{-2\pi x^R/\beta}}{1+e^{2\pi(x^L-\beta u)/\beta}-e^{2\pi x^L/\beta}}\right\}\tag{3.26}$$

and

$$R_W^1(x,u)=(\beta/4\pi)\log\{(1+e^{-2\pi(x^R-\beta u)}-e^{-2\pi x^R/\beta})(1+e^{2\pi(x^L-\beta u)/\beta}-e^{2\pi x^L/\beta})\}. \quad (3.27)$$

Note that the wedge is characterized by $x^R \geqslant 0$ and $x^L \leqslant 0$. Again, the modular action coincides essentially with time translations far from the domain boundary. Near the edge of the wedge, the coordinate x^R is scaled by $\exp(-2\pi u)$ and x^L is scaled by $\exp(2\pi u)$, up to terms of order $(|x|/\beta)^2$. This corresponds to a Lorentz boost, i.e., the modular action at temperature zero.

From $\Gamma_\pm(\tau)$ we can form the one-parameter unitary groups,

$$\tau \mapsto \Gamma_\pm(\tau) \otimes \Gamma_\pm(\tau), \quad (3.28)$$

on the tensor product Hilbert space. These groups have the positive generators $H + (1/\beta)\log \Delta_{\pm,\pm}$, where $\Delta_{\pm,\pm} = \Delta_\pm \otimes 1 + 1 \otimes \Delta_\pm$ is the modular operator of $\mathcal{M}(\mathbf{R}_\pm \times \mathbf{R}_\pm)$. They correspond, respectively, to the forward and backward light cone ($++$ and $--$) and the left and the right wedge ($+-$ and $-+$). All four groups converge to the time translations as $\beta \to \infty$.

The group associated with the forward light cone is

$$\Gamma_{V^+}(\tau) = \Gamma_+(\tau) \otimes \Gamma_+(\tau). \quad (3.29)$$

By (2.44) we have the following theorem.

Theorem 3.2: *If* $\mathbf{x} \in \mathbf{R}^2$ *and*

$$\tau > -\beta(2\pi)^{-1} \min\{e^{2\pi x^L/\beta}, e^{2\pi x^R/\beta}\}, \quad (3.30)$$

then

$$\text{ad } \Gamma_{V^+}(\tau)\mathcal{M}(V_{\mathbf{x}}^+) = \mathcal{M}(V_{\psi_{V^+}(\tau,\mathbf{x})}^+), \quad (3.31)$$

with

$$\psi_{V^+}(\tau,\mathbf{x}) = (\psi_+(\tau,x^L), \psi_+(\tau,x^R)). \quad (3.32)$$

If

$$\tau > -\beta(2\pi)^{-1}(\min\{e^{2\pi x^L/\beta}, e^{2\pi x^R/\beta}\} - 1), \quad (3.33)$$

then

$$\text{ad } \Gamma_{V^+}(\tau)\mathcal{M}(V_{\mathbf{x}}^+) \subset \mathcal{M}(V^+). \quad (3.34)$$

The group associated with the right wedge,

$$\Gamma_W(\tau) = \Gamma_-(\tau) \otimes \Gamma_+(\tau), \quad (3.35)$$

does not induce half-sided translations on the wedge algebra, but it nevertheless acts geometrically for a restricted parameter range. In fact, by Eqs. (3.7) and (3.10) we have the following theorem.

Theorem 3.3: *If* $\mathbf{x} \in \mathbf{R}^2$ *and*

$$-\beta(2\pi)^{-1} e^{2\pi x^R/\beta} < \tau < \beta(2\pi)^{-1} e^{-2\pi x^L/\beta}, \quad (3.36)$$

then

$$\text{ad } \Gamma_W(\tau)\mathcal{M}(W_{\mathbf{x}}) = \mathcal{M}(W_{\psi_W(\tau,\mathbf{x})}), \quad (3.37)$$

with

$$\psi_W(\tau,\mathbf{x})=(\psi_-(\tau,x^L),\psi_+(\tau,x^R)). \tag{3.38}$$

If

$$-\beta(2\pi)^{-1}(e^{2\pi x^R/\beta}-1)<\tau<\beta(2\pi)^{-1}(e^{-2\pi x^L/\beta}-1), \tag{3.39}$$

then

$$\mathrm{ad}\;\Gamma_W(\tau)\mathcal{M}(W_{\mathbf{x}})\subset\mathcal{M}(W). \tag{3.40}$$

The flows of ψ_{V+} and ψ_W are shown in Figs. 3 and 4. The groups $\Gamma_{V+}(\tau)$ and $\Gamma_W(\tau)$, approximate the time translations close to the tip of the light cone and the edge of the wedge, respectively. Indeed, ψ_{V+} maps (x^0,x^1) to (x'^0,x'^1) with

$$x'^0=x^0+\tau[\exp(-2\pi x^R)+\exp(-2\pi x^L)]/2+O(\tau^2/\beta) \tag{3.41}$$

and

$$x'^1=x^1+\tau[\exp(-2\pi x^R)-\exp(-2\pi x^L)]/2+O(\tau^2/\beta). \tag{3.42}$$

For x^R and x^L both close to zero, this is close to $x'^0=x^0+\tau$, $x'^1=x^1$. More interesting, however, is the behavior of $\Gamma_{V+}(\tau)$ far from the apex of the cone. From Fig. 3 one sees clearly that the flow corresponds to a decelerated motion toward the origin of space. More quantitatively, the velocity $v=dx'^1/dx'^0$ is

$$v=-\tanh(2\pi x'^1/\beta), \tag{3.43}$$

and this differential equation has the general solution

$$x'^0(x'^1)=-(\beta/2\pi)\log(\sinh(2\pi x'^1/\beta))+C, \tag{3.44}$$

where C is an arbitrary constant. The path through the origin, $x'^1=0$, corresponds formally to $C=-\infty$. The flow pattern is invariant under a shift in the time direction, in accord with (2.31).

As already mentioned in the Introduction, this flow brings points that start out with the velocity of light at infinity gradually to rest. Formally we have the reverse of an Unruh effect for the generator of the flow of the observables is positive, with the KMS state vector as a ground state. Measured in terms of the parameter τ, it takes $\beta/2\pi$ τ units for points to reach the forward light cone from infinity. The τ parameter along the path through the origin is related to the real time t by

$$t=(\beta/2\pi)\log(1+2\pi\tau/\beta),\quad \text{i.e.,}\quad \tau=(\beta/2\pi)(\exp(2\pi t/\beta)-1). \tag{3.45}$$

The τ unit is calibrated in such a way that the t and τ scales coincide precisely where the path hits the apex of the light cone. According to Eq. (2.31), a different calibration corresponds simply to a shift of the cone in the time direction. It is clear from (3.45) that the τ parameter is ''slower'' than t, in the sense that $d\tau/dt<1$, for a point on the path outside the light cone ($\tau<0$), and ''faster'' than t, i.e., $d\tau/dt>1$, inside the light cone ($\tau>0$).

For ψ_W the equations corresponding to (3.41) and (3.42) are

$$x'^0=x^0+\tau[\exp(-2\pi x^R)+\exp(2\pi x^L)]/2+O(\tau^2/\beta) \tag{3.46}$$

and

$$x'^1=x^1+\tau[\exp(-2\pi x^R)-\exp(2\pi x^L)]/2+O(\tau^2/\beta), \tag{3.47}$$

and the velocity is

$$v = -\tanh(2\pi x'^0/\beta). \tag{3.48}$$

Thus, the velocity is small close to the space axis, but approaches ± 1 far away from the space axis. The explicit solution of (3.48) is

$$x'^1(x'^0) = -(\beta/2\pi)\log(\cosh(2\pi x'^0/\beta)) + C. \tag{3.49}$$

This flow is invariant under a translation in the x^1 direction. Here the situation is different from the light cone, since not all paths pass through the wedge, and those who do stay in the wedge only for a finite τ interval; cf. Eq. (3.39). The group even moves localized observables out of the global observable algebra in finite ''τ time,'' cf. Eq. (3.36). The direction of acceleration is in the opposite wedge here, whereas in the usual Unruh effect it points in the direction of the wedge. In this sense, here we also have a kind of reverse of the situation in the Unruh effect.

For the path passing through the origin, the τ parameter is related to $t = x^0$ by

$$t = \frac{\beta}{4\pi}\log\frac{1+(2\pi\tau/\beta)}{1-(2\pi\tau/\beta)}, \quad \text{i.e.} \quad \tau = \frac{\beta}{2\pi}\tanh\frac{2\pi t}{\beta}. \tag{3.50}$$

The relation to the proper time t_p along the path is

$$\tau = \frac{\beta}{2\pi}\sin\frac{2\pi t_p}{\beta}, \tag{3.51}$$

i.e., up to a slight deformation τ is essentially the proper time. We have

$$d\tau/dt_p = \cos\frac{2\pi t_p}{\beta} = (1-(2\pi\tau/\beta)^2)^{1/2}, \tag{3.52}$$

so ''τ time'' is everywhere slower than t_p except at the origin (calibration point), where both scales coincide with t. A change of scale corresponds to a translation of the wedge along the x^1 axis because of (2.31).

Above, we have described the actions of the modular- and Γ groups in terms of the space time coordinates (x^L, x^R) and also in terms of (x^0, x^1). The simplest description is obtained in yet another coordinate system, which is related to the others by a nonlinear transformation. For $x \in \mathbf{R}$ define

$$\xi_\pm = \pm(\beta/2\pi)(\exp(\pm 2\pi x/\beta) - 1). \tag{3.53}$$

The range of ξ_+ is $]-\beta/2\pi, \infty[$ and the range of ξ_- is $]-\infty, \beta/2\pi[$. With $x = x^L$ and $x = x^R$, respectively, we thus obtain the four coordinates, ξ_+^L, ξ_-^L, ξ_+^R, and ξ_-^R. In the case of the forward light cone we pick (ξ_+^L, ξ_+^R) and in the case of the right wedge (ξ_-^L, ξ_+^R) as a curvilinear coordinate system on Minkowski space. In these coordinates the transformations (3.19), (3.21) for the groups associated with the forward light cone V^+ become

$$(\xi_+^L, \xi_+^R) \mapsto e^{-2\pi u}(\xi_+^L, \xi_+^R), \quad (\xi_+^L, \xi_+^R) \mapsto (\xi_+^L, \xi_+^R) + \tau(1,1), \tag{3.54}$$

and the corresponding transformations (3.32) and (3.38) for the right wedge W are

$$(\xi_-^L, \xi_+^R) \mapsto (e^{2\pi u}\xi_-^L, e^{-2\pi u}\xi_+^R), \quad (\xi_-^L, \xi_+^R) \mapsto (\xi_-^L, \xi_+^R) + \tau(1,1). \tag{3.55}$$

Analogous formulas hold for the backward cone and the left wedge. Hence, in the ξ coordinates the transformations have exactly the same form for all β, including the vacuum case, $\beta = \infty$.

The four coordinate systems $(\xi_\pm^L, \xi_\pm^R)$ can be put together by defining

$$(\tilde{\xi}^L, \tilde{\xi}^R) = (\beta/2\pi)(\epsilon(x^L)(\exp(\epsilon(x^L)2\pi x^L/\beta) - 1), \epsilon(x^R)(\exp(\epsilon(x^R)2\pi x^R/\beta) - 1), \tag{3.56}$$

with $\epsilon(x)=1$ for $x\geqslant 0$ and -1 for $x<0$. This transformation is once continuously differentiable, but second derivatives have a discontinuity on the light cone. The lines corresponding to the flow of the Γ groups of the four domains (the forward and backward cones and the two wedges) pass continuously through the boundaries between the domains, although the groups themselves do *not* merge to a single one-parameter unitary group on the Hilbert space.

The transformation (3.56) is of the form $(x^L,x^R)\mapsto(f(x^L),f(x^R))$ with f an order preserving bijective map $\mathbf{R}\to\mathbf{R}$. Hence, it is a causal transformation on two-dimensional space–time, i.e., it takes light cones into light cones. Such nonlinear causal maps on Minkowski space exist only in two space–time dimensions.

Finally, we remark that all results of this section hold for general 2-D theories, provided the state satisfies a KMS condition with respect to both light-cone coordinates, x^L and x^R. For factorizing states, this holds automatically as a consequence of the KMS condition with respect to the time direction. A general proof of a KMS condition with respect to light-like translations seems out of reach, however, even if one involves the relativistic KMS condition.[22]

IV. EXPLICIT REALIZATIONS OF MODULAR GROUPS

In this section we compute explicitly the modular and Γ groups for generalized free fields on a light ray and the corresponding tensor product models on $\mathbf{R}^2$. In these examples it is possible to discuss the action of the groups on the algebras of double cones and not only of translated forward cones and wedges.

The Weyl algebra of a generalized free Bose field on a light ray is generated by elements $W(f)$, with f a real-valued Schwartz test function on $\mathbf{R}$, satisfying the following relations:

$$W(f)^*=W(-f) \tag{4.1}$$

and

$$W(f)W(g)=e^{-K(f,g)/2}W(f+g), \tag{4.2}$$

with

$$K(f,g)=\int_{-\infty}^{\infty} pQ(p^2)\tilde{f}(-p),\tilde{g}(p)dp, \tag{4.3}$$

where $Q(p^2)$ is a non-negative polynomial that characterizes the field (see Ref. 11). Here $\tilde{f}(p)=(1/2\pi)\int\exp(-ipx)f(x)dx$ is the Fourier transform of f. The kernel of $\mathcal{K}$ of K, defined by $K(f,g)=\int\mathcal{K}(y-x)f(x)g(y)dx\,dy$, is

$$\mathcal{K}(y-x)=M(-id/dy)\,\delta(y-x), \tag{4.4}$$

with $M(p)=pQ(p^2)$, so $W(f)$ and $W(g)$ commute if f and g have disjoint supports.

Translations in time are equivalent to translations along the light ray and are represented by automorphisms of the Weyl algebra,

$$\alpha_t(W(f))=W(f(\cdot-t)). \tag{4.5}$$

A quasifree KMS state ω at inverse temperature β is defined on the Weyl algebra by

$$\omega(W(f))=\exp(-\omega_2(f,f)), \tag{4.6}$$

where ω_2 is given by a positive definite kernel $\mathcal{W}_2(y-x)$ (two-point function) that is analytic in the strip $S(0,\beta)$, and satisfies

$$\mathcal{W}_2(\xi)-\mathcal{W}_2(-\xi)=\mathcal{K}(\xi), \tag{4.7}$$

for real ξ, together with the KMS condition

$$\mathcal{W}_2(\xi+\mathrm{i}\beta)=\mathcal{W}_2(-\xi). \tag{4.8}$$

It is straightforward to show that these conditions fix $\mathcal{W}_2$ (up to normalization); its Fourier transform is

$$\tilde{\mathcal{W}}_2(p)=\frac{pQ(p^2)}{1-e^{-\beta p}}. \tag{4.9}$$

The Fourier transform of the meromorphic function $(1-\exp(-\beta p))^{-1}$ is seen to be $\lim_{\varepsilon\to 0_+}(2\pi i\beta)^{-1}(\exp(\beta^{-1}2\pi\xi+\mathrm{i}\varepsilon)-1)^{-1}$ by contour integration. The Fourier transform of (4.9) for general Q follows by differentiation. In particular, we have for $Q\equiv 1$, i.e., a field of scaling dimension 1,

$$\mathcal{W}_2(\xi)=\lim_{\varepsilon\to 0_+}\frac{1}{\beta^2}\frac{1}{\left(\sinh\dfrac{\pi(\xi+i\varepsilon)}{\beta}\right)^2}, \tag{4.10}$$

and for Q a polynomial of degree n,

$$\mathcal{W}_2(\xi)=\lim_{\varepsilon\to 0_+}\frac{P\left(\cosh\dfrac{\pi\xi}{\beta},\sinh\dfrac{\pi\xi}{\beta}\right)}{\left(\sinh\dfrac{\pi(\xi+\mathrm{i}\varepsilon)}{\beta}\right)^{2n+2}}, \tag{4.11}$$

where P is a polynomial in two variables. We shall restrict ourselves to the case that $Q(p^2)=p^{2n}$, i.e., a field of a definite scaling dimension $(n+1)$, in order not to mix the effects coming from the nonzero temperature with those due to inhomogeneous polynomials Q (see Ref. 11 for the latter).

Denoting the Weyl operators corresponding to $Q(p^2)=p^{2n}$ by $W^{(n)}(f)$, it is clear from (4.3) that we may identify

$$W^{(n)}(f)=W^{(0)}(\mathrm{i}^n f^{(n)}), \tag{4.12}$$

where $f^{(n)}$ is the nth derivative of f, and from (4.9) we see also that a KMS state for any n is the same as the KMS state for $n=0$ restricted to the operators $W^{(0)}(\mathrm{i}^n f^{(n)})$. This will allow us to reduce everything to the simplest case, $n=0$.

Let π be the GNS representation defined by the KMS state (4.6) on the Weyl algebra of the $W^{(0)}(f)$'s. If $I\subset\mathbf{R}$ is an interval, bounded or unbounded, we define $\mathcal{M}^{(n)}(I)$ to be the von Neumann algebra generated by $\pi(W^{(n)}(f))$ with supp $f\subset I$. Because of the identification discussed above, these algebras are for all n realized on the same Hilbert space.

By exactly the same arguments as in Ref. 11, Sec. III, one proves the following.

4.1 Lemma: *If I is an unbounded interval, then $\mathcal{M}^{(n)}(I)\equiv\mathcal{M}(I)$ is independent of n. If I is bounded with a nonempty interior, then $\mathcal{M}^{(m)}(I)$ is a proper subalgebra of $\mathcal{M}^{(n)}(I)$ for $m>n$.*

This lemma implies, in particular, that the modular operator Δ_+ corresponding to the half-line $\mathbf{R}_+$ is the same for all n.

The main result about the modular action is the following

Theorem 4.2: *Let ω be the quasifree KMS state (4.6) and π the corresponding representation of the Weyl algebra for $n=0$. The modular group of $\mathcal{M}(\mathbf{R}_+)$ defined by ω has the form*

$$\Delta_+^{\mathrm{i}u}\pi(W^{(0)}(f))\Delta_+^{-\mathrm{i}u}=\pi(W^{(0)}(\delta_u^{(0)}f)), \tag{4.13}$$

with

$$\delta_u^{(0)}f(x)=f\left(\frac{\beta}{2\pi}\log\{1+e^{2\pi u}(e^{2\pi x/\beta}-1)\}\right), \tag{4.14}$$

for supp $f\subset\mathbf{R}_+$.

Remark 1: It is understood that if supp $f\subset\mathbf{R}_+$, then also $\delta_u^{(0)}f(x)=0$ for all $x<0$.

Remark 2: The cyclic vector Ω corresponding to ω has the Reeh–Schlieder property, in particular, it is cyclic for $\mathcal{M}(\mathbf{R}_+)$. Hence, (4.13) with supp $f\subset\mathbf{R}_+$, together with $\Delta_+^{it}\Omega=\Omega$, already fixes Δ_+^{it} as a unitary group on the GNS Hilbert space. But $\Delta_+^{iu}\pi(W^{(0)}(f))\Delta_+^{-iu}$ is, of course, a well-defined operator on the Hilbert space for all f of compact support, and, in fact, if $u\geqslant 0$, then (4.13) and (4.14) hold for functions with support outside of $\mathbf{R}_+$ with the understanding that (4.14) is zero when the argument of the logarithm is $\leqslant 0$. This is a simple consequence of (3.1)–(3.3). If $u<0$, however, the transformed operator only belongs to the observable algebra if condition (3.3) holds on the support of f.

Proof of Theorem 4.2: The formula (4.14) is motivated by Eq. (2.20). In order to show that it is the correct formula for the modular action, we have to check the following properties of $\delta_u^{(0)}$: (i) $\delta_u^{(0)}$ maps the space of test functions with support in $\mathbf{R}_+$ into itself; (ii) the group property, i.e., $\delta_u^{(0)}\circ\delta_{u'}^{0}=\delta_{u+u'}^{(0)}$; (iii) the unitary of $\delta_u^{(0)}$ in the scalar product defined by the two point function (4.10); (iv) the KMS condition: For real test functions f and g with support in $\mathbf{R}_+$ the function $u\mapsto\omega_2(f,\delta_u^{(0)}g)$ has an analytic continuation into the strip $S(-1,0)$ and

$$\omega_2(f,\delta_{u-i}^{(0)}g)=\omega_2(\delta_u^{(0)}g,f).$$

Property (i) is obvious from the definition, and (ii) and (iii) are straightforward calculations. We now check the KMS condition. Put

$$L(u,x):=\frac{\beta}{2\pi}\log\{1+e^{2\pi u}(e^{2\pi x/\beta}-1)\}. \tag{4.15}$$

Since $L(u,L(-u,y))=y$ (group property), we have

$$\omega_2(f,\delta_u^{(0)}g)=\int\int\mathcal{W}_2(L(-u,y)-x)\,\frac{\partial L(-u,y)}{\partial y}\,f(x)g(y)dx\,dy. \tag{4.16}$$

Using the addition formula for hyperbolic functions, we compute, for the two-point function, (4.10):

$$\begin{aligned}
&\mathcal{W}_2(L(-u,y)-x)\partial L(-u,y)/\partial y\\
&\quad=\frac{1}{4\beta^2}[\sinh(\pi L(-u,y)/\beta)\cosh(\pi x/\beta)-\cosh(\pi L(-u,y)/\beta)\\
&\qquad\times\sinh(\pi x/\beta)+i\varepsilon]^{-2}\,\frac{\partial L(-u,y)}{\partial y}\\
&\quad=\frac{1}{16\beta^2}[\{(1+e^{-2\pi u}(e^{2\pi y/\beta}-1))^{1/2}+(1+e^{-2\pi u}(e^{2\pi y/\beta}-1))^{-1/2}\}\cosh(\pi x/\beta)\\
&\qquad-\{(1+e^{-2\pi u}(e^{2\pi y/\beta}-1))^{1/2}+(1+e^{-2\pi u}(e^{2\pi y/\beta}-1))^{-1/2}\}\\
&\qquad\times\sinh(\pi x/\beta)+i\varepsilon]^{-2}\,\frac{e^{-2\pi u}e^{2\pi y/\beta}}{e^{-2\pi u}(e^{2\pi y/\beta}-1)+1}\\
&\quad=\frac{1}{16\beta^2}[\{e^{-2\pi u}(e^{2\pi y/\beta}-1)\}\cosh(\pi x/\beta)-\{2+e^{-2\pi u}(e^{2\pi y/\beta}-1)\}\\
&\qquad\times\sinh(\pi x/\beta)+i\varepsilon]^{-2}e^{-2\pi u}e^{2\pi y/\beta}\\
&\quad=\frac{e^{2\pi y/\beta}}{16\beta^2}[e^{-\pi u}(e^{2\pi y/\beta}-1)[\cosh(\pi x/\beta)-\sinh(\pi x/\beta)]-e^{\pi u}2\sinh(\pi x/\beta)+i\varepsilon]^{-2}.
\end{aligned}$$

For $x,y>0$, $e^{-\pi u}$ comes with a positive factor and $e^{\pi u}$ with a negative one. For $\varepsilon>0$, the total expression is therefore analytic in u in the strip $S(-1,0)$, and this analyticity is preserved in the limit $\varepsilon\rightarrow 0_+$ after smearing in x and y, with test functions with support in $\mathbf{R}_+$. The boundary value at $u-i$, $u\in\mathbf{R}$ is

$$\frac{e^{2\pi y/\beta}}{16\beta^2}\,[e^{\pi u}2\sinh(\pi x/\beta)-e^{-\pi u}(e^{2\pi y/\beta}-1)[\cosh(\pi x/\beta)-\sinh(\pi x/\beta)]+\mathrm{i}\varepsilon]^{-2}.$$

This is precisely $\mathcal{W}_2(x-L(-u,y))\partial L(-u,y)/\partial y$ (by the same computation). Hence, the KMS condition is verified.

The representation of the group $\Gamma_+(\tau)=\exp(\mathrm{i}\tau G_+/\beta)$ with the positive generator $G_+/\beta=H+(1/\beta)\log\Delta_+$ now follows immediately from Eqs. (2.25) and (4.13)–(4.14):

Theorem 4.3: *For* $\tau\geqslant 0$ *and all* f

$$\Gamma_+(\tau)\pi(W^{(0)}(f))\Gamma_+(-\tau)=\pi(W^{(0)}(\gamma_\tau^{(0)}f)) \tag{4.17}$$

with

$$\gamma_\tau^{(0)}f(x)=f\left(x+\frac{\beta}{2\pi}\log\left\{1-\frac{2\pi\tau}{\beta}e^{-2\pi x/\beta}\right\}\right). \tag{4.18}$$

Remark: It is understood that $\gamma_\tau^{(0)}f(x)=0$ if the argument of the logarithm is $\leqslant 0$, i.e., if $x\leqslant-\beta/(2\pi)\log(2\pi\tau/\beta)$. Note that if $\tau\geqslant\beta/(2\pi)$, then supp $\gamma_\tau^{(0)}f\subset\mathbf{R}_+$ for any f of compact support.

By (4.12) we obtain as a corollary of Theorems 4.2 and 4.3, the following.

Theorem 4.4: *For* $n>0$ *the action of* ad $\Delta_+^{\mathrm{i}u}$ *and* ad $\Gamma_+(\tau)$ *on* $W^{(n)}(f)$ *with* supp $f\subset\mathbf{R}_+$ *is*

$$\Delta_+^{\mathrm{i}u}\pi(W^{(n)}(f))\Delta_+^{-\mathrm{i}u}=\pi(W^{(n)}(\delta_u^{(n)}f)), \tag{4.19}$$

with

$$\delta_u^{(n)}f(x)=\int_0^x dx_1\int_0^{x_1}\cdots\int_0^{x_{n-1}}dx_n\;\delta_u^{(0)}f^{(n)}(x_n), \tag{4.20}$$

and, for $\tau\geqslant 0$,

$$\Gamma_+(\tau)\pi(W^{(n)}(f))\Gamma_+(-\tau)=\pi(W^{(n)}(\gamma_\tau^{(n)}f)), \tag{4.21}$$

with

$$\gamma_\tau^{(n)}f(x)=\int_0^x dx_1\int_0^{x_1}\cdots\int_0^{x_{n-1}}dx_n\;\gamma_\tau^{(0)}f^{(n)}(x_n). \tag{4.22}$$

Remark. It should be noted that $\delta_u^{(n)}f$ is, in general, no longer a test function if $n>0$, for it may behave like x^{n-1} for $x\rightarrow\infty$. However, it belongs to the Hilbert space defined by the two-point function, and hence the Weyl operators are well defined. The same applies to $\gamma_\tau^{(n)}f$.

Next, we investigate the localization properties of the modular groups. We recall from Lemma 4.1 that for an unbounded interval $[x,\infty[$ the algebras $\mathcal{M}^{(n)}([x,\infty[)\equiv\mathcal{M}([x,\infty[)$ are independent of n. Hence, the general result (3.1) applies. For the algebras corresponding to bounded intervals, we have the following theorem.

Theorem 4.5: *For* $-\infty<x<y<\infty$ *and* u *and* τ *restricted according to (3.3), (3.5) (3.9), (3.12),*

$$\mathrm{ad}\,\Delta_+^{\mathrm{i}u}\mathcal{M}^{(0)}([x,y])=\mathcal{M}^{(0)}([\varphi_+(u,x),\varphi_+(u,y)]), \tag{4.23}$$

and

$$\mathrm{ad}\,\Gamma_+(\tau)\mathcal{M}^{(0)}([x,y])=\mathcal{M}^{(0)}([\psi_+(\tau,x),\psi_+(\tau,y)]). \tag{4.24}$$

For $n>0$ a local algebra $\mathcal{M}^{(n)}([x,y])$ is not mapped into an $\mathcal{M}^{(n)}(I)$ with bounded I.

Proof: For fixed u and τ the maps $x\mapsto\varphi_+(u,x)$ and $x\mapsto\psi_+(\tau,x)$ are one to one for x satisfying (3.3) and (3.9), respectively, and the inverse maps correspond to $u\to-u$ and $\tau\to-\tau$. From (3.14) it is clear that f has its support in $[x,y]$, iff $\delta_u^{(0)}f$ has its support in $[\varphi_+(u,x),\varphi_+(u,y)]$ iff $\gamma_\tau^{(0)}f$ has its support in $[\psi_+(\tau,x),\psi_+(\tau,y)]$. Hence (4.23) and (4.24) follows directly from Theorems 4.2 and 4.3.

To show the dislocalization for $n>0$ we note first that neither $\delta_u^{(0)}f^{(n)}$ nor $\gamma_\tau^{(0)}f^{(n)}$ is a derivative of a function with compact support (except for $f\equiv0$). This is easily seen by considering the Fourier transforms of these functions, divided by p; the $1/p$ singularity is not compensated by the derivatives because of the nonlinear variable transformations, and analyticity is lost. Consider now a bounded interval I and a function g such that $g^{(n+1)}$ vanishes on I. Then $W^{(0)}(g)$ belongs to the commutant of $\mathcal{M}^{(n)}(I)$. If $W^{(n)}(\delta_u^{(n)}f)=W^{(0)}(\delta_u^{(0)}f^{(n)})$ would belong to $\mathcal{M}^{(n)}(I)$, then it would commute with $W^{(0)}(g)$, which means that

$$\int\delta_u^{(0)}f^{(n)}(x)g'(x)dx=0.$$

This must, in particular, hold for all g with $g'\equiv1$ on I because such g fulfill $g^{(n+1)}=0$ on I for $n>0$. Hence

$$\int_I\delta_u^{(0)}f^{(n)}(x)dx=0. \tag{4.25}$$

By isotony this should also hold for all larger intervals, and hence $\delta_u^{(0)}f^{(n)}$ would be a derivative of a function of compact support. As remarked above, this is not the case, and we have a contradiction to the assumption that $W^{(n)}(\delta_u^{(n)}f)$ belongs to $\mathcal{M}^{(n)}(I)$ with I bounded. By the same argument $W^{(n)}(\gamma_\tau^{(n)}f)$ does not belong to $\mathcal{M}^{(n)}(I)$.

Remark 1: In terms of the field operators $\Phi^{(n)}(x)$, defined by

$$\pi(W^{(n)}(f))=\exp\left(\mathrm{i}\int\Phi^{(n)}(x)f(x)dx\right), \tag{4.26}$$

Eqs. (3.13) and (3.17) say that

$$\Delta_+^{\mathrm{i}u}\Phi^{(0)}(x)\Delta_+^{-\mathrm{i}u}=\Phi^{(0)}(\varphi_+(u,x))\,\frac{\partial\varphi_+(u,x)}{\partial x} \tag{4.27}$$

and

$$\Gamma_+(\tau)\Phi^{(0)}(x)\Gamma_+(-\tau)=\Phi^{(0)}(\psi_+(\tau,x))\,\frac{\partial\psi_+(\tau,x)}{\partial x}. \tag{4.28}$$

In particular, we have

$$\Delta_+^{\mathrm{i}u}\Phi^{(0)}(0)\Delta_+^{-\mathrm{i}u}=e^{-2\pi u}\Phi^{(0)}(0) \tag{4.29}$$

and

$$\Gamma_+(\tau)\Phi^{(0)}(0)\Gamma_+(-\tau)=(1+(2\pi\tau/\beta))^{-1}\Phi^{(0)}((\beta/2\pi)\log(1+(2\pi\tau/\beta))). \tag{4.30}$$

(Although the field is only an operator-valued distribution, these equations have a rigorous meaning in terms of quadratic forms.) Conversely, (3.33) and (3.34), together with Eqs. (2.20) and (2.26), imply (3.31) and (3.32). For $n>0$, however, ad Δ_+^{iu} is a nonlocal transformation of the field operators by Theorem 4.4. For instance, we have

$$\Delta_+^{iu}\Phi^{(1)}(0)\Delta_+^{-iu}=e^{-2\pi u}\Phi^{(1)}(0)-(2\pi/\beta)e^{-4\pi u}\int_0^\infty \Phi^{(1)}(x)dx. \tag{4.31}$$

This shows clearly that there is more to the transformation law for the fields than Eqs. (2.20) and (2.29) alone.

If $\mathcal{M}(\mathbf{R}_+)$ is replaced by $\mathcal{M}(\mathbf{R}_-)$, the previous results apply with appropriate changes of signs; cf. (3.5).

Forming tensor product algebras as in (3.13), we obtain generalized free fields on two-dimensional space–time and KMS states that factorize in the light-cone variables. In the case of the field with lowest scaling dimension, i.e., $n=0$, the double cone algebras $\mathcal{M}^{(0)}(I_L\times I_R)$, with I_L, I_R bounded intervals, are again mapped into algebras of double cones. The flow lines of Figs. 1–4 describe in this case not only the movement of the apex of a forward light cone or the edge of a wedge, but also the movement of the double cones.

For fields of higher scaling dimension, i.e., $n>0$, however, double cone algebras are after the transformation no longer localized in double cones within the net $\mathcal{M}^{(n)}$. They are still localized in double cones within the net $\mathcal{M}^{(0)}(\cdot)$, because $\mathcal{M}^{(n)}(\cdot)$ is a subnet of $\mathcal{M}^{(0)}(\cdot)$.

V. CONCLUSIONS

In a KMS state at inverse temperature β the time translations coincide (up to a sign and scaling by β) with the modular group of the global observable algebra. From this fact, and the general theory of half-sided modular inclusions, algebraic relations between time translations and the modular groups for certain domains of space–time can be derived. The action of the modular groups on observables localized inside these domains far from the boundary is approximately given by the time translations. In two-dimensional models and states that satisfy a KMS condition with respect to light-like translations (in particular, models that factorize in the light-cone coordinates), a geometric interpretation can be given of the action of the modular groups of the forward light cone and a space-like wedge on observable algebras localized in translated domains of the same type. This action can be studied in detail in simple free field models. Besides the modular groups, the theory also leads to one-parameter groups with positive generators, for which the KMS state is a ground state. The actions of these groups for the forward cone and the wedge can also be described geometrically and interpreted, at least formally, as a kind of a reverse Unruh effect.

ACKNOWLEDGMENTS

We thank D. Buchholz, H. Narnhofer, P. Michor, and W. Thirring for helpful comments. Hospitality of the the Erwin Schrödinger Institute, Vienna (H.J.B.) and the University Science Institute, Reykjavik (J.Y.) is also gratefully acknowledged.

[1] R. Haag, *Local Quantum Physics*, 2nd ed. (Springer-Verlag, Berlin, 1996).
[2] H. Reeh and S. Schlieder, ''Eine Bemerkung zur Unitäräquivalenz von Lorentzinvarianten Feldern,'' Nuovo Cimento **22**, 1051 (1961).
[3] M. Takesaki, *Tomita's Theory of Modular Hilbert Algebras and its Applications*, Lecture Notes in Mathematics (Springer-Verlag Berlin, 1970), Vol. 128.
[4] R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras II* (Academic, New York, 1986).
[5] D. Buchholz, C. D'Antoni, and R. Longo, ''Nuclear maps and modular structures II: Applications to quantum field theory,'' Commun. Math. Phys. **129**, 115–138 (1990).
[6] H.-J. Borchers, ''On the use of modular groups in quantum field theory,'' Ann. Inst. Henri Poincaré Phys. Theor. **64**, 331–382 (1996).

[7] J. Bisognano and E. H. Wichmann, ''On the duality condition for a Hermitian scalar field,'' J. Math. Phys. **16**, 985–1007 (1975).
[8] J. Bisognano and E. H. Wichmann, ''On the duality condition for quantum fields,'' J. Math. Phys. **17**, 303–321 (1976).
[9] D. Buchholz, ''On the structure of local quantum fields with non-trivial interactions,'' in *Proceedings of the International Conference on Operator Algebras, Ideals and their Applications in Theoretical Physics*, Leipzig 1977, Teubner-Texte zur Mathematik, 1978, pp. 146–153.
[10] P. D. Hislop and R. Longo, ''Modular structure of the local algebra associated with a free massless scalar field theory,'' Commun. Math. Phys. **84**, 71–85 (1982).
[11] J. Yngvason, ''A note on essential duality,'' Lett. Math. Phys. **31**, 127–141 (1994).
[12] H.-J. Borchers, ''The CPT-theorem in two-dimensional theories of local observables,'' Commun. Math. Phys. **143**, 315–332 (1992).
[13] H.-W. Wiesbrock, ''Half-sided modular inclusions of von Neumann algebras,'' Commun. Math. Phys. **157**, 83 (1993).
[14] H.-W. Wiesbrock, ''Half-sided modular inclusions of von Neumann algebras,'' Commun. Math. Phys. **184**, 683–685 (1997).
[15] W. G. Unruh, ''Notes on black-hole evaporation,'' Phys. Rev. D **14**, 870–892 (1976).
[16] G. L. Sewell, ''Relativity of temperature and the Hawking effect,'' Phys. Lett. A **79**, 23–24 (1980).
[17] G. L. Sewell, ''Quantum fields on manifolds: PCT and gravitationally induced thermal states,'' Ann. Phys. **141**, 201–224 (1982).
[18] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I* (Springer-Verlag, New York, 1979).
[19] H.-J. Borchers, ''Half-sided translations and the type of von Neumann algebras,'' Lett. Math. Phys. (to appear).
[20] G. G. Emch, H. Narnhofer, W. Thirring, and G. L. Sewell, ''Anosov actions on noncommutative algebras,'' J. Math. Phys. **35**, 5582–5598 (1994).
[21] H.-W. Wiesbrock, ''A comment on a recent work of Borchers,'' Lett. Math. Phys. **25**, 157–159 (1992).
[22] J. Bros and D. Buchholz, ''Towards a relativistic KMS-condition,'' Nucl. Phys. B **429**, 2911–318 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Is there a stable hydrogen atom in higher dimensions?

Frank Burgbacher[a)]
Fakultät für Physik, Universität Konstanz, Fach M674, D-78457 Konstanz, Germany

Claus Lämmerzahl[b)]
Fakultät für Physik, Universität Konstanz, Fach M674, D-78457 Konstanz, Germany and Laboratoire de Physique des Lasers, Institut Galilée, Université Paris 13, F—93430 Villetaneuse, France

Alfredo Macias[c)]
Departamento de Física, Universidad Autónoma Metropolitana—Iztapalapa, Apartado Postal 55-534, C.P. 09340, México, D.F., México

(Received 23 July 1998; accepted for publication 19 October 1998)

The Schrödinger equation in higher dimensions is considered. It consists of the kinetic energy part given by the corresponding Laplace operator, and a term describing the interaction with the electrostatic field of a point charge. From Rutherford-type scattering experiments one can conclude that the potential of a point charge is $\sim 1/r$ irrespective of the dimension of the space where the experiment is carried through. Also the structure of the kinetic energy is unchanged in higher dimensions so that one is lead to the result that there exist stable atoms in higher spatial dimensions $d \geqslant 4$. The solutions and energy eigenvalues to this Schrödinger equation in higher dimensions are presented. As a consequence, the dimensionality of space can be read off from the spectral scheme of atoms: The three-dimensionality of space is a consequence of the existence of the Lyman series. Another consequence is that the Maxwell equations in higher dimensions must be modified in order to have the $1/r$-potential as solution for a point charge.
 [S0022-2488(99)00502-2]

I. INTRODUCTION

The idea of extra space–time dimensions continues to pervade current attempts to unify the fundamental forces, but in ways somewhat different from that originally envisaged. A modern perspective on the role of internal dimensions in physics comes mainly from the superstring theory, which is the most promising candidate for a unified field theory. The appearance of extra space–time dimensions at high energy scales is a generic feature of string theory. Typically these extra dimensions remain compactified at the Planck scale, but it is possible for new dimensions to have an effect below the Planck scale. In particular, large-radius compactification schemes have recently been discussed in a number of theoretical and phenomenological contexts.[1,2] Similarly, the effects of extra dimensions below the Planck scale have played a role in understanding the strong-coupling behavior of string theory.[3] Even the old pioneer Kaluza–Klein theory is embedded in a super-string theory; their states persist as a subset of the full string spectrum. However, string theory comes to rescue and ensures correct high-energy behavior.[4] Then, we can regard this theory as an effective ''medium'' energy model coming from finite string field theories. Therefore, the study of different higher-dimensional models is of importance for the understanding of more general theories.

One of the most interesting questions addressed to the higher-dimensional approaches concerns the stability of atoms in higher spatial dimensions, i.e., $d > 3$. These investigations started

[a)]Electronic mail: frank.burgbacher@uni-konstanz.de
[b)]Electronic mail: claus.laemmerzahl@uni-konstanz.de
[c)]Electronic mail: amac@xanum.uam.mx

with the well-known paper of Ehrenfest[5] and has inspired many additional interesting investigations. For reviews see Refs. 6 and 7, and for a recent paper on this problem see Ref. 8 where the dimensionality of space–time has been related to physical phenomena which are accessible to experiment.

According to the analysis of Ehrenfest, see also Ref. 9, there are statements in all papers that in higher dimensions it is not possible to have stable atoms. It is one of our purposes in this paper to show that it is indeed possible to have *stable atoms in higher dimensions*. The main point is that first the kinetic energy in the Schrödinger has the usual form described by the d-dimensional Laplacian and that the electrostatic interaction in the Schrödinger equation has the same form irrespective of the spatial dimension. This of course leads to modified Maxwell equations in higher dimensions. While the main characteristics of these new Maxwell equations in higher dimensions remains the same as compared with the Maxwell equations in three dimensions (the solutions have the same structure and the force between charges is the same as in three dimensions), these modified Maxwell equations do not lead to a Gaussian law for charges. This may sound strange but the results of scattering experiments, the stability of atoms in higher dimensions, and the structure of the force between charges is certainly of more basic physical content.

A second point in our paper is that the spectra of atoms are influenced by the spatial dimension. That means, as we shall show, that we *can decide from a spectroscopic experiment the dimension of our configuration space*. To be more concrete, the ratio of the frequencies of two distinguished spectral lines leads to a number from which we uniquely infer the three-dimensionality of our space. If this ratio gives a different number we would be led to four or another number of spatial dimensions.

The most important starting point of our investigation is the structure of the Schrödinger equation in higher dimensions. One way which fixes the kinetic part of the Schrödinger equation is the quantization scheme arising from the Hamilton–Jacobi equation of a point mass which also in higher dimensions has the usual form $E=p^2/2m+V$, where V is some potential energy. In addition, also from a constructive axiomatic scheme (see, for example, Refs. 8 or 10) one gets a Dirac equation in higher dimensions which nonrelativistic limit[11] necessarily possesses a kinetic term which is proportional to the Laplace operator. Therefore, any modification of this term would need a modification of the quantization scheme as well as a violation of fundamental properties (like unique evolution, superposition principle, finite propagation speed, etc., see, for example, Ref. 10) of single particle quantum systems. Since these modifications obviously changes physics drastically we do not change the structure of the usual kinetic term.

As far as the potential energy term is concerned we use results from scattering experiments to fix its form. Indeed, since the results of Rutherford-type scattering experiments are independent of the spatial dimension, we can unambiguously conclude from the experimental data, that in any dimension d the potential must be of the form $\sim 1/r$. This is of course consistent with the analysis of Ref. 5 that atoms with the usual kinetic energy coupled to a modified potential of the form $\sim/r^{d-2}$ are not stable (the exponent $d-2$ is due to the requirement that Gauss' law should be still valid in higher dimensions). Since our result for the electrostatic potential is not compatible with a Gaussian law for electrostatics, we conclude that we have to modify the structure of Maxwell's equations in higher dimensions.

Consequently, we take as general ansatz for the Hamilton operator for the hydrogen atom in higher dimensions,

$$H=\frac{p^2}{2m}+V(r), \tag{1}$$

where $V(r)$ is the spherically symmetric potential given by

$$V(r)=\frac{\alpha}{r}. \tag{2}$$

In comparison to other work on the problem of physics in higher dimensions, we do not consider the usual physical laws like the Maxwell equations (see, for example, Refs. 12 and 13) or the Schrödinger equation or Newton's field equations (see Ref. 5), or the Einstein equations (see Refs. 14–16) to be valid in higher dimensions and discuss physical implications of the solutions. Instead, we start with general *physical* properties of the class of phenomena under consideration and then try to get information of the structure of the physical laws. In general, these equations in higher dimensions are very different from the equations in three dimensions describing the same effects. An interesting approach,[17] which is in the line of our reasoning, is based on the causal structure of space–time events. It deduces the four-dimensionality of space–time from a set of axioms which do not use the notion of a differentiable manifold or of the dimensionality. Another approach having some similarities to our reasoning is given in Ref. 18 where it is shown that for a gravitational theory based on a quadratic Lagrangian the usual Newtonian limit and Huygens' principle is valid only if this theory if formulated in six space–time dimensions. In Ref. 8, a very general approach to a generalized Dirac equation in arbitrary dimensions has been used and the dimensions of space–time has been inferred from the propagation of helicity states and from the validity of Huygen's principle. In this work we do not consider the fractal dimension; see, for example, Ref. 19.

In earlier work[5–7] it has been shown that there are no stable hydrogen atoms in higher dimensions. Essential for that was the assumption that also in higher dimensions Maxwell's equations were assumed to be valid leading to a potential of a point charge of the form $\sim 1/r^{d-2}$ where d is the spatial dimension. In our approach we do not assume the usual Maxwell equations to be valid. We only use the results of scattering experiments to get information about the potential of a point charge. We use this potential in Sec. III in order to solve the hydrogen atom and then show that even in higher dimensions there are stable atoms. However, from the comparison of the calculated spectrum with the observational data we are able to determine in Sec. IV the dimensionality of our space. In Sec. V we present the full set of modified Maxwell equations in order to show that even our potential violating Gauss' law is part of a consistent set of equations governing electrodynamical phenomena in a higher dimension. Though being nonlocal in general, they are still Lorentz-covariant.

II. THE POTENTIAL OF A POINT CHARGE IN HIGHER DIMENSIONS

The electrostatic potential of the atomic nucleus which we assume to be pointlike, can be determined by means of scattering processes. Indeed, using the scattering of α-particles at gold atoms, Rutherford was able to deduce that the electrostatic potential within an atom is the Coulomb potential. We will show that this procedure and this result is true independent of the underlying spatial dimensions. This can be seen already from the fact that the classical trajectory of a point charge in a $1/r$ potential does not depend on the spatial dimension so that the relation between the deflection angle and the potential also remains the same.

Starting with (1,2), conventional quantum mechanics gives the asymptotics of scattered waves according to

$$u^{+}(\vec{r}) = \frac{1}{(2\pi)^{d/2}} \left(e^{i\vec{k}\vec{r}} + f_{\vec{k}}(\vec{e})\, \frac{e^{ikr}}{r^{(d-1)/2}} \right), \tag{3}$$

with $\vec{e} = \vec{r}/r$. This can be shown by calculating the Green's function in the energy representation

$$G(\vec{r},\vec{r}\,') = \frac{1}{(2\pi\hbar)^d} \lim_{\epsilon\to 0^+} \int \frac{e^{(i/\hbar)(\vec{r}-\vec{r}\,')\cdot\vec{p}}}{E_0 + i\epsilon - p^2/2m}\, d^d p$$

which results in a position dependence of the form $\sim |\vec{r}-\vec{r}\,'|^{-(d-2)}$ with factors depending on the dimension d and an integration over a spherical Bessel function. The scattering amplitude is then given by

$$f_{\vec{k}}(\vec{e}) \sim \int e^{-ik\vec{e}\vec{r}\,'} V(\vec{r}\,') u^{+}(\vec{r}\,') d^d x', \tag{4}$$

where $\vec{e}=\vec{r}/r$. In the Born approximation we have

$$f_{\vec{k}}(\vec{e}) \sim \int e^{i(\vec{k}-\vec{q})\vec{r}} V(\vec{r}) d^d x. \tag{5}$$

In a scattering experiment the measured quantity is the differential cross section $\sigma(\vec{e},\vec{k}_0)$ which is related to the scattering amplitude by

$$\sigma(\vec{e},\vec{k}_0) = |f_{\vec{k}_0}(\vec{e})|^2. \tag{6}$$

This is a relation which is independent of the dimension of the underlying space. In the Born approximation there is a one-to-one correspondence between the differential cross section and the potential $V(\vec{r})$. Therefore, by analyzing the standard Rutherford-type experiments we can uniquely conclude that a point charge, or the nucleus of a hydrogen atom, possess a potential of the form $\sim r^{-1}$, independent of the spatial dimension.

III. THE HYDROGEN ATOM IN HIGHER DIMENSIONS

We start with the Hamilton operator (1) in the external spherically symmetric potential (2) which gives, in position representation,

$$[\Delta - \phi + \epsilon]\psi = \psi, \tag{7}$$

where we introduced the abbreviations

$$\phi(r) = \frac{2m}{\hbar^2} V(r), \tag{8}$$

$$\epsilon = \frac{2m}{\hbar^2} E. \tag{9}$$

The following calculations are analogous to that in three dimensions. Also, in a higher dimension we can separate the Laplace operator into a radial and an angular part:

$$\Delta = \hat{R} - \frac{1}{r^2} \hat{L}, \tag{10}$$

where we introduced

$$\begin{aligned} \hat{R} &= \frac{\partial^2}{\partial r^2} + \frac{d-1}{r} \frac{\partial}{\partial r}, \\ \hat{L} &= \hat{L}(\Theta_2, \ldots, \Theta_n). \end{aligned} \tag{11}$$

With the corresponding ansatz,

$$\psi = R(r) Y(\Theta_2, \ldots, \Theta_d), \tag{12}$$

we get from the Schrödinger equation in d dimensions,

$$\frac{r^2}{R}\hat{R}R+r^2(\epsilon-\phi)=\frac{1}{Y}\hat{L}Y=l(l+d-2). \tag{13}$$

$Y(\Theta_2,\ldots,\Theta_d)$ represents the spherical harmonics in d dimensions. They are eigenstates of the angular momentum operator $\hat{L}$ with the eigenvalues $l(l+d-2)$. Thus we get for the radial part of the wave function,

$$\left[\frac{\partial^2}{\partial r^2}+\frac{d-1}{r}\frac{\partial}{\partial r}+\epsilon-\phi-\frac{l(l+d-2)}{r^2}\right]R=0. \tag{14}$$

We introduce

$$r'=\frac{r}{r_0},\quad -\epsilon=\frac{1}{r_0^2},\quad \alpha'=\frac{2m\alpha}{\hbar^2}\frac{r_0}{2}, \tag{15}$$

and assume for the potential the form

$$\phi(r')r_0^2=-\frac{\alpha'}{r'}. \tag{16}$$

We also introduce a new variable $f(r')$ through

$$R=e^{-(1/2)r'}r'^{\gamma}f(r'), \tag{17}$$

and get an equation for the function f:

$$\begin{aligned}0=r\frac{d^2f(r')}{dr'^2}+[2\gamma+d-1-r]\frac{df(r')}{dr'}\\+\left[\frac{\gamma(\gamma+d-2)-l(l+d-2)}{r'}+\frac{\alpha'}{r'}-\frac{2\gamma+d-1}{2}\right]f(r').\end{aligned} \tag{18}$$

This equation is valid for arbitrary d. In order to solve this equation we specify the value of γ by the requirement that the term $\sim 1/r'$ should vanish:

$$\gamma(\gamma+d-2)-l(l+d-2)=0. \tag{19}$$

This gives the two possibilities

$$\gamma_+=l, \tag{20}$$

$$\gamma_-=-(l+d-2), \tag{21}$$

and from (18),

$$zf''-[\vartheta-z]f'-\beta f=0, \tag{22}$$

with

$$\vartheta:=\pm(2l+d-2)+1,\quad \beta:=\frac{\pm(2l+d-2)+1}{2}-\alpha'. \tag{23}$$

Equation (22) is the confluent hypergeometric differential equation with the solution[20]

$$f(\beta,\vartheta,z)=\sum_{\nu=0}^{\infty}\frac{(\beta+\nu)!\,\vartheta!}{\beta!(\vartheta+\nu)!}\frac{z^{\nu}}{\nu!}, \tag{24}$$

which is appropriate for our problem.

It is clear that, in order to get no infinite terms, ϑ is not allowed to be a negative integer: $\vartheta\neq-1,-2,\ldots$. Therefore we cannot use the solution (21). In addition, if the sum does not terminate, then the solution diverges for large r faster than $\exp(\frac{1}{2}r')$ which leads to non-normalizable solutions. The condition for a termination of the sum is $\beta\in\mathbb{Z}^{-}$, or

$$\beta=l+\frac{d-1}{2}-\alpha'=-k, \quad k\in\mathbb{N}. \tag{25}$$

Here α' is connected with the energy eigenvalues (1,8,15). Therefore we get for the energy eigenvalues E,

$$E=\frac{2m\alpha^2}{\hbar^2}\frac{1}{\alpha'^2}=-\frac{2m\alpha^2}{\hbar^2}\frac{1}{(l+[(d-1)/2]+k)^2}=-Ry\frac{1}{n^2}=:E_n, \tag{26}$$

where the principal quantum number n is given by the series

$$n=\frac{d-1}{2},\ \frac{d-1}{2}+1,\ \frac{d-1}{2}+2,\ \frac{d-1}{2}+3,\ldots . \tag{27}$$

We also introduced the Rydberg constant Ry which, in general, may depend through α on the dimension d. In the case $d=3$ we recover the usual expressions. Note that, in general, the principal quantum number n must not be an integer.

Consequently, we have shown that for a potential of the form $\sim 1/r$ even in higher dimensions there is a lowest energy level, that is, there are stable atoms.

IV. THE INFLUENCE OF THE DIMENSION ON THE SPECTRUM

We discuss now the spectrum of stable hydrogen atoms in higher dimensions. It is clear that the spectrum depends on the dimension d. An interesting question is whether this dependence is accessible to observations. In an experiment only the difference of two energy eigenvalues,

$$\Delta E_{n',n}=E_{n'}-E_n, \quad n'>n, \tag{28}$$

can be measured. For a fixed n one gets an atomic series which now depends on the dimension d. In three dimensions $d=3$ one gets for $n=1$ the Lyman series, for $n=2$ the Balmer series, for $n=3$ the Paschen series, etc. In 4 dimension, for example, according to (27) it is not possible to have $n=1$, so that in this case there is no Lyman series. In $d=6$ dimensions there is also no Balmer series.

However, since the Rydberg constant Ry may depend on the dimension d in an unknown way, we are not able to draw any conclusions about the dimensionality of space from testing the atomic spectral series. Therefore we are forced to restrict ourselves to the ratio of two energy differences which is also independent of any unit conventions. In our case it is enough to take the ratio of the difference between the three lowest energy levels of one series characterized by n:

$$D(n)=\frac{\Delta E_{n+2,n}}{\Delta E_{n+1,n}}=\frac{4(1+n)^3}{(2+n)^2(1+2n)}. \tag{29}$$

Because this function $D(n)$ is one-to-one, the value of $D(n)$ uniquely characterizes the corresponding series. For the first few values we get $D(0)=1$, $D(\frac{1}{2})=\frac{27}{25}=1.08$, $D(1)=\frac{32}{27}=1.18519$, $D(\frac{3}{2})=\frac{125}{98}=1.275$, $D(2)=\frac{27}{20}=1.35$, $D(\frac{5}{2})=\frac{343}{243}=1.41152$, etc.

Therefore we have the following experimental method at hand in order to determine the dimensionality of our space: We consider that series which belongs to the lowest energy state. From this series we take the two highest frequency spectral lines and calculate the ratio. This gives the value $D(n_{\min})$. From this value we can calculate the corresponding $n_{\min}$ and, using (27), the dimension $d=2n_{\min}+1$ of our space. Here we used that in each dimension the lowest series contains only transitions with $l=k=0$.

We know from spectroscopy of the hydrogen atom that the two spectral lines coming from transitions to the lowest energy level have (see, for example, Ref. 21) 1215.67 Å and 1025.73 Å, so that $D(n_{\min})=1216/1026=1.18518$. A comparison with the values of $D(n)$ shows that this implies $n_{\min}=1$, and from Eq. (27) that $d=3$. Therefore we have *proven by a spectroscopic experiment that our space is three-dimensional*. In other words, because we know the spectrum of the hydrogen atom we are able to determine the dimensionality of space.

We want to stress once more that it is not the stability of the atom which one may use as argument in favor of three spatial dimensions. In our approach the stability of atoms is secured in any dimension. It is only the structure of the spectral series which leads us to the conclusion that space is three-dimensional.

V. MAXWELL EQUATIONS IN HIGHER DIMENSIONS

We have seen that the electric potential of a point charge in the Schrödinger equation in higher dimensions must be of the form $U\sim 1/r$ independent of the dimension d. Since the usual Laplacian has the same form in any dimension, the above potential cannot be a solution of the Poisson equation in $d>3$ dimensions. However, we show that it is indeed possible to present a consistent set of equations governing the electromagnetic phenomena in higher dimensions which violates no fundamental principle of electrodynamics and, in addition, possesses the above electrostatic solutions for a point charge. Of course, the structure of the Maxwell equations will be not the same as in three dimensions.

In order to determine the structure of the stationary Maxwell equation for the electric field, we use results of Riesz distributions, see, for example, Refs. 22, 23. In doing so we first define the distribution

$$G_\lambda := \frac{e^{-i\pi\lambda}\Gamma[(d/2)-\lambda]r^{2\lambda-d}}{4^\lambda \pi^{d/2}\Gamma(\lambda)}, \tag{30}$$

where r as usual is the distance $r^2=\Sigma_{i=1}^d x_i^2$. The properties of these distributions G_λ are given by the composition law

$$G_\mu * G_\lambda = G_{\lambda+\mu}, \tag{31}$$

and an explicit representation in the case of negative integers,

$$G_k = \Delta^{-k}\delta, \quad k=0,-1,-2,..., \tag{32}$$

where Δ is again the Laplace operator in an arbitrary dimension, δ the usual Dirac delta distribution, and the star $*$ the convolution operation.

We formally introduce operators $\overline{\Delta}^\lambda$ by

$$\overline{\Delta}^\lambda := G_{-\lambda}, \tag{33}$$

so that the following composition law holds:

$$\overline{\Delta}^{\mu} * G_{\lambda} = G_{\lambda-\mu}. \tag{34}$$

An important special case is given by $\mu=\lambda$:

$$\overline{\Delta}^{\mu} * G_{\mu} = \delta. \tag{35}$$

This means that G_{μ} is a Green's function corresponding to the operator $\overline{\Delta}^{\mu} *$.

Now we come back to our problem of finding the field equations which are required to possess the solution $\sim r^{-1}$ in any dimension d. That means that we require in any dimension $G_{\mu} \sim 1/r$ which implies $\mu=(d-1)/2$. Consequently,

$$\overline{\Delta}^{(d-1)/2} * G_{(d-1)/2} = \delta, \tag{36}$$

or

$$\overline{\Delta}^{(d-1)/2} * \frac{1}{r} = (4\pi)^{(d-1)/2} e^{[i\pi(d-1)/2]} \Gamma\left(\frac{d-1}{2}\right) \delta. \tag{37}$$

This means that the equation for the electric potential, or the generalized Poisson equation, reads as

$$(\overline{\Delta}^{(d-1)/2} * \phi)(x) = (4\pi)^{(d-1)/2} e^{[i\pi(d-1)/2]} \Gamma\left(\frac{d-1}{2}\right) \rho(x), \tag{38}$$

where $\rho(x)$ is the charge density in d dimensions. The operator $\overline{\Delta}^{(d-1)/2}$ replaces the Laplacian in three dimensions. In general, this operator is no differential operator.

We briefly discuss this new form of the Poisson equation in electrostatics.

(1) It is possible to reformulate the field equation for the potential ϕ in terms of the electric field strength $\mathbf{E}=-\nabla\phi$. For doing so we use (31) and (32):

$$\overline{\Delta}^{(d-1)/2} * \phi = (\overline{\Delta} * \overline{\Delta}^{(d-3)/2}) * \phi = \Delta\delta * (\overline{\Delta}^{(d-3)/2} * \phi) = \delta * (\overline{\Delta}^{(d-3)/2} * \delta\phi) = (\overline{\Delta}^{(d-3)/2} * \nabla \cdot \mathbf{E}), \tag{39}$$

so that we get as field equation for the electric field strength,

$$((\overline{\Delta}^{(d-3)/2} * \nabla) \cdot \mathbf{E})(x) = (4\pi)^{(d-1)/2} e^{[i\pi(d-1)/2]} \Gamma\left(\frac{d-1}{2}\right) \rho(x). \tag{40}$$

(2) The force between two charges still has the same form as in 3 dimensions, namely $f \sim q_1 q_2 / r^2$.

(3) For all charge densities $\rho(x)$ the solution for the potential looks as usual, i.e., $1/r * \rho$.

(4) In odd dimensions $d=1,3,5,...$, the above equation reduces via (31) and (32) to a differential equation:

$$\overline{\Delta}^{(d-1)/2} * \frac{1}{r} = \delta * \Delta^{(d-1)/2} \frac{1}{r} = \Delta^{(d-1)/2} \frac{1}{r} = (4\pi)^{(d-1)/2} e^{[i\pi(d-1)/2]} \Gamma\left(\frac{d-1}{2}\right) \delta. \tag{41}$$

For a three-dimensional space, $d=3$, we get the usual Laplace equation $\Delta\phi(x) = -4\pi\rho(x)$ and in a five-dimensional space we get $\Delta^2 \phi(x) = (4\pi)^2 \rho(x)$.

(5) In even dimensions, the operator $\overline{\Delta}^{(d-1)/2}$ is no differential operator but instead a pseudo-differential operator. Therefore the corresponding field equations are pseudo-differential operator equations. These operators are nonlocal. (Indeed, differential operators are the only local operators acting linearly and surjective on C^{∞}; see Ref. 24. For a physical discussion, see, for example, Ref. 25.)

(6) An essential difference to the usual properties of the electric field in 3 dimensions is that now the Gauss' law is no longer valid. This is easy to see by integrating the fundamental solution r^{-1} in an arbitrary dimension over the surface of a sphere with radius R:

$$\int \mathbf{E} \cdot d\mathbf{A} = \int_R \nabla \frac{1}{r} \cdot \hat{\mathbf{r}} R^{d-1} d\Omega = (4\pi)^{(d-1)/2} e^{[i\pi(d-1)/2]} \Gamma\left(\frac{d-1}{2}\right) R^{d-3}, \tag{42}$$

where $\hat{\mathbf{r}}$ is the unit vector in radial direction and $d\Omega$ is the surface element in d dimensions. The result depends on the radius of the sphere so that indeed Gauss' law does not hold. It is only in three spatial dimensions that the quantity $\mathbf{E}$ which plays the role of a force on a charged particle, is also that quantity which integral over the area enclosing a volume gives the total charge which acts as source of $\mathbf{E}$ (the field strength $\mathbf{E}$ is defined by means of the force acting on a charged particle; whether this quantity obeys a law like Gauss' law is a deduced property which holds in three dimensions).

However, Gauss' law is valid for a quantity deduced from $\mathbf{E}$, namely for $\mathfrak{E}=\overline{\Delta}^{(d-3)/2}\mathbf{E}$:

$$\nabla\cdot\mathfrak{E}=(4\pi)^{(d-1)/2}e^{i\pi[(d-1)/2]}\Gamma\left(\frac{d-1}{2}\right)\rho(x)\Leftrightarrow\oint\mathfrak{E}\cdot d\mathbf{A}=(4\pi)^{(d-1)/2}e^{i\pi[(d-1)/2]}Q, \tag{43}$$

where $Q=\int_V\rho(x)d^dx$ is the charge contained in the volume V.

It is also straightforward to give the full set of Maxwell's equations such that their static limit give the Poisson equation discussed above: Since the covariant generalization of the Laplace operator Δ is given by the d'Alambert operator $\Box$, the covariant generalization of Poisson's equation is $\bar{\Box}^{(d-1)/2}*\phi=(4\pi)^{(d-1)/2}e^{i\pi[(d-1)/2]}\Gamma((d-1)/2)\rho$. We complete the quantities to covariant 4-vectors, namely the 4-potential A^a and the 4-current j^a. Then we have, using the same methods as above,

$$\begin{aligned} j^a&=\bar{\Box}^{(d-1)/2}*A^a=(\bar{\Box}*\bar{\Box}^{(d-3)/2})*A^a=\Box\,\delta*(\bar{\Box}^{(d-3)/2}*A^a)=\delta*(\bar{\Box}^{(d-3)/2}*\Box A^a)\\ &=\bar{\Box}^{(d-3)/2}*\partial_bF^{ba}=:\overline{\partial_b}F^{ba}, \end{aligned} \tag{44}$$

where we defined a generalized partial derivative $\overline{\partial_b}:=\bar{\Box}^{(d-3)/2}*\partial_b$ and, as usual, the Maxwell field strength $F_{ab}=\partial_aA_b-\partial_bA_a$. We also used the Lorentz condition $\partial_aA^a=0$. By construction, these generalization of Maxwell's equations is covariant. Also current conservation is fulfilled. For even spatial dimensions these equations are nonlocal.

To sum up: despite the fact that the mathematical structure of the equation determining the electric potential from a given charge density changes dramatically when compared with the three-dimensional case, the physical content does not change. The solution has the same form and the force between charges is the same as in three dimensions. Only Gauss' law loses its meaning. However, we think that the specific expression for the force between charged particles and the stability of atoms are of more basic physical importance than the validity of Gauss' law.

VI. SUMMARY AND DISCUSSION

To sum up, we have shown the following.

(1) From Rutherford-type experiments we can conclude that the potential of the point charge in any spatial dimension must be $\sim 1/r$.
(2) This potential leads to stable atoms in higher dimensions.
(3) The dimensionality enters the atomic spectra thus making it possible to infer uniquely from atomic spectra the three dimensionality of space.
(4) That the Maxwell equations have to be modified in higher dimensions in order to allow solutions of the form $1/r$, leading to nonlocal equations in even spatial dimensions.

In the case that one uses the usual Maxwell equations in higher dimensions the hydrogen atom is proven to be not stable. This has been related to the fact that orbits of classical bodies in a potential derived from the usual Poisson equation in higher dimensions are not stable, as well: small perturbations of the circular orbit leads the body to fall into the central body or to leave the system. Consequently, if one wants to enlarge the above reasoning to the case of Newtonian mechanics, one has to require stable orbits, which gives the $1/r$ potential for gravity also in higher dimensions. This forces one to modify the Poisson equation for the Newtonian potential in the same way as the Poisson equation for the electrostatic potential in Sec. V. That means, in higher dimension d the field equation for the Newtonian potential $U(x)$ must be of the form $(\overline{\Delta}^{(d-1)/2} * U)(x) = (4\pi)^{(d-1)/2} e^{i\pi(d-1)/2} \Gamma((d-1)/2)\rho(x)$, where $\rho(x)$ is the mass density. As a consequence, also Einstein's equations should be modified in higher dimensions.

In conclusion, we want to say that our or similar considerations do not rule out the possibility of unifying physics in higher dimensions; we just restrict, from observations, the direct physical applicability of dynamical equations to three spatial dimensions.

ACKNOWLEDGMENTS

We thank J. Audretsch, Ch. J. Bordé, F. W. Hehl, and H. Kastrup for discussions. Moreover, C.L. thanks the German Academic Exchange Service (DAAD) and the CNRS (France) for financial support and Ch. J. Bordé and Ch. Chardonnet for their hospitality at the Laboratoire de Physique des Lasers in Villetaneuse, France. This research was supported by Conacyt, Grant No. 3544-E9311, and by the joint German-Mexican project KFA-Conacyt E130-2924.

[1] E. Caceres, V. S. Kaplunovsky, and I. M. Mandelberg, ''Large volume string compactification, revisited,'' Nucl. Phys. B **493**, 73 (1997).
[2] T. Banks and M. Dine, ''Couplings and scales in strongly coupled heterotic string theory,'' Nucl. Phys. B **479**, 173 (1996); ''The cosmology of string theoretic axions,'' *ibid.* **505**, 445 (1997).
[3] P. Hořava and E. Witten, ''Heterotic and Type I string dynamics from eleven dimensions,'' Nucl. Phys. B **460**, 506 (1996).
[4] M. J. Duff, ''Kaluza–Klein theory in perspective,'' talk delivered at the Oskar Klein Centenary Nobel Symposium, Stockholm, 19–21 September 1994 (hep-th/9410046).
[5] P. Ehrenfest, Proc. Amst. Acad. **20**, 200 (1917); P. Ehrenfest, ''Welche Rolle spielt die Dimensionalität des Raumes in den Grundgesetzen der Physik?,'' Ann. Phys. (Leipzig) **61**, 440 (1920).
[6] I. M. Freeman, ''Why is space three-dimensional?,'' based on W. Büchel, ''Warum hat der Raum drei Dimensionen?,'' Am. J. Phys. **37**, 1222 (1969).
[7] J. D. Barrow, ''Dimensionality,'' Philos. Trans. R. Soc. London, Ser. A **310**, 337 (1983).
[8] C. Lämmerzahl and A. Macias, ''On the dimensionality of space–time,'' J. Math. Phys. **34**, 4540 (1993); Addendum, *ibid.* **36**, 2131 (1995).
[9] L. Gurevich and V. Mostepanenko, ''On the existence of atoms in *n*-dimensional space,'' Phys. Lett. A **35**, 201 (1971).
[10] J. Audretsch and C. Lämmerzahl, ''A new constructive axiomatic scheme for the geometry of space–time,'' in *Semantical Aspects of Space–Time Geometry*, edited by U. Majer and H.-J. Schmidt (BI Verlag, Mannheim, 1993).
[11] C. Lämmerzahl and U. Bleyer, ''A quantum test theory for basic principles of special and general relativity with applications to atom beam interferometry and Hughes–Drever type experiments,'' preprint, University of Konstanz, 1998.
[12] L. M. Sokolowski, F. Occhionero, M. Litterio, and L. Amendola, ''Electromagnetism and the dimensionality of space-time,'' in *Relativistic Astrophysics and Cosmology*, edited by S. Gottlöber, J. P. Mücket, and V. Müller (World Scientific, Singapore, 1992).
[13] S. Sonego and V. Faraoni, ''Huygen's principle and characteristic propagation property for waves in curved space–times,'' J. Math. Phys. **33**, 625 (1992).
[14] F. R. Tangherlini, ''Schwarzschild field in *n* dimensions and the dimensionality of space problem,'' Nuovo Cimento **27**, 636 (1963).
[15] B. Müller and A. Schäfer, ''Improved bounds on the dimension of space–time,'' Phys. Rev. Lett. **56**, 1215 (1986).
[16] R. C. Myers and M. J. Perry, ''Black holes in higher dimensional space–times,'' Ann. Phys. (N.Y.) **172**, 304 (1986).
[17] H.-J. Seifert, ''Zur Dimensionszahl der Raumzeit,'' in *Festschrift Fachbereich Maschinenbau*, 10 Jahre Hochschule der Bundeswehr Hamburg, 1983.
[18] C. G. Bollini, J. J. Giambiagi, and O. Obregon, ''Are some physical theories related with a specific number of dimensions?,'' Mod. Phys. Lett. A **7**, 593 (1992).
[19] A. Zeilinger and K. Svozil, ''Measuring the dimension of space–time,'' Phys. Rev. Lett. **54**, 2553 (1985).
[20] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1983).
[21] G. W. F. Drake, ''Atomic Spectroscopy,'' in *Atomic, Molecular & Optical Physics Handbook* (American Institute of Physics, Woodbury, NY, 1996).
[22] I. M. Gel'fand, *Generalized Functions* (1966).
[23] C. G. Bollini and J. J. Giambiagi, ''Arbitrary powers of D'Alembertians and the Huygens' principle,'' J. Math. Phys. **34**, 610 (1993).
[24] J. Peetre, ''Rectification à l'article 'Une caractérisation abstraite des opérateurs différentiels','' Math. Scand. **8**, 116 (1960); J. Dieudonné, *Élements d'Analyse* (Gauthier-Villars, Paris, 1971).
[25] C. Lämmerzahl, ''The pseudo-differential operator square root of the Klein–Gordon equation,'' J. Math. Phys. **34**, 3918 (1993).

Implementation of an iterative map in the construction of (quasi)periodic instantons: Chaotic aspects and discontinuous rotation numbers

A. Chakrabarti[a)]
Centre de Physique Théorique, Ecole Polytechnique 91128 Palaiseau, France

(Received 4 March 1998; accepted for publication 15 October 1998)

An iterative map of the unit disk in the complex plane is used to explore certain aspects of self-dual, four-dimensional gauge fields (quasi)periodic in the Euclidean time. These fields are characterized by two topological numbers and contain standard instantons and monopoles as different limits. The iterations do not correspond directly to a discretized time evolution of the gauge fields. They are implemented in an indirect fashion. First, (t,r,θ,ϕ) being the standard coordinates, the (r,t) half-plane is mapped on the unit disk in an appropriate way. This provides an (r,t) parametrization of Z_0, the starting point of the iterations and makes the iterates increasingly complex functions of r and t. These are then incorporated as building blocks in the generating function of the fields. We explain in what sense and to what extent some remarkable features of our map (indicated in the title) are thus carried over into the *continuous* time development of the fields. Special features for quasiperiodicity are studied. Spinor solutions and propagators are discussed from the point of view of the mapping. Several possible generalizations are indicated. Some broader topics are also discussed.
[S0022-2488(99)03802-5]

I. INTRODUCTION

An iterative map of the unit disk centered at the origin of the complex plane is studied in the Appendix. The motivation is that it can be implemented fruitfully in the study of a hierarchy of four-dimensional, self-dual, (quasi)periodic gauge fields. Fields with spherical symmetry in R^3 are mostly used to illustrate our approach. More general possibilities are indicated at the end (Sec. VI). Such fields have been studied previously in a series of papers[1–8] which contains references to other sources. Here we reformulate them from the point of view of the iterative map. This brings remarkable new aspects to light.

Our gauge fields are (quasi)periodic in Euclidean time. The basic ingredient (one may say the generating function) for the spherically symmetric fields is a holomorphic function $g(r+it)$ satisfying certain constraints (Sec. II) in the (r,t) half-plane $(r\geqslant 0)$. Here (t,r,θ,ϕ) are the standard coordinates. The function g can have several factors (or nonfactorized terms) each with its own period in t. When all the periods are mutually commensurable the lowest common multiple of the component periods is the overall one. This is the periodic case. When at least one of the component periods is incommensurable with some others, one has, by definition, quasiperiodicity.

As is well known, fields periodic in Euclidean time provide one possible approach to field theory at finite temperature. Our explicit constructions are such that even an infinitesimal change in one single parameter (determining one of the component periods) can make a periodic solution quasiperiodic and vice versa. So they are considered in a parallel fashion.

The map studied in the Appendix is

[a)]Electronic mail: chakra@cpth.polytechnique.fr; also at: Laboratoire Propre du CNRS UPR A.0014.

$$Z_{p+1}=\frac{a_p+Z_p}{\bar{a}_p+Z_p^{-1}}, \quad |Z_0|\leq 1, \quad 0<|a_p|<1. \tag{1}$$

Note the inverse (Z_p^{-1}) in the denominator. Possible choices of a_p are discussed in the Appendix. *For implementing the map in the construction of the gauge fields the crucial step is a suitable* (r,t) *parametrization of* Z_0. The two choices considered are

$$Z_0=e^{-k(r+it)} \quad (k>0), \tag{2}$$

and

$$Z_0=\frac{\Sigma_{l=1}^{n}\lambda_l^2(e^{k_l(r+it-ic_l)}-1)^{-1}}{\Sigma_{l=1}^{n}\lambda_l^2(1-e^{-k_l(r+it-ic_l)})^{-1}} \tag{3}$$

with real parameters (λ,k,c) and

$$k_n>k_{(n-1)}>\cdots>k_2>k_1>0.$$

The uses of (2) will be amply illustrated (Secs. II, III, and the Appendix). The choice (3) is particularly suited to the construction of spinor solutions (Sec. IV). It can be shown to satisfy all the necessary constraints.

With Z_0 thus chosen one can set (Sec. II) the generating function of the gauge fields to be

$$g=\left(\prod_{j=1}^{n} Z_{p_j}^{(j)}\right), \tag{4}$$

where different sets of parameters are associated with each factor.

The iterations do not correspond to a discretized time evolution of the fields. At each step one has a different action (Sec. II), a different system. But the (r,t) parametrization of $Z_0^{(j)}$ will provide the key to the usefulness of the iterations. As p_j increases, g becomes a more and more complicated function of (r,t). But the fact that such complications are introduced in a very specific fashion stepwise makes some remarkable properties readily accessible. Some crucial properties of each block $Z_{p_j}^{(j)}$ as a whole are delivered directly. To give this statement explicit content let us look at the two most striking features of our map. [Up to Sec. IV, (2) will usually be referred to directly for simplicity. This is not an essential restriction.]

(1) For $r=0$ (the circumference of the unit disk) Z_p for any p, is a phase, the angle denoted by ψ_p, and (1) becomes a circle map. This map (see the Appendix) satisfies all the criteria for being chaotic (Ref. 9, Def.8.5, p. 50). These are as follows:

(a) a sensitive dependence on initial data, encoded by a positive characteristic index;

(b) a dense set of periodic points;

(c) transitivity.

Moreover this phase and its derivatives provide the coefficients of a series expansion in r of Z or g (the Appendix).

Suppose now, that as a consequence of (a), namely the positive index, at a certain level l of iterations the difference

$$\psi_l(\psi_0+\delta\psi_0)-\psi_l(\psi_0)$$

is appreciably large even for a small $\delta\psi_0$ (considering ψ_l as a function of its initial datum ψ_0). Due to our parametrization this means

$$\psi_l(t+\delta t)-\psi_l(t)$$

is large even for small δt.

Through (4), the result above, for each factor j, is injected into the time evolution of the fields. *This leads to sensitive time dependence.* (See, however, the remarks in Sec. VII.)

The overall phase of g (for any r) does not contribute directly to the action density, expressible in terms of $g\bar{g}$ and its derivatives (Sec. II). But the amplitude at each order of iteration involves the phases of the lower ones (the Appendix). Moreover the phase *is* directly involved in the power series expansion in r.

The phase can be studied directly and analogously for any value of r (inside the disk). The emphasis on the phase at $r=0$ is due to two factors.

(a) The relative simplicity for $r=0$ permits a transparent derivation of the crucial properties.
(b) For spherical symmetry, in our ansatz, the time dependence is damped exponentially with increasing r (going toward the center of the disk). Hence a small sphere around the origin in R^3 is the most suitable domain for studying the time evolution. The leading term of the action density for small r is given explicitly (Sec. II) to display the role of the iterated phases.

(2) We do not ignore, however, other remarkable features associated with certain values of r away from the spatial origin (the choice of the parameters a and k determines how far away or how close). These are the domains on which the $Z^{(j)}$'s can vanish and hence also g. They lead to discontinuous rotation numbers associated to the phases (the Appendix). So far as the zeros of g can be located their cumulative effects lead to staircaselike patterns. An example is given in Sec. II.

The rotation numbers are *asymptotic* quantities $[n\rightarrow\infty$ in (A58)]. Hence, discontinuities can arise even if we are dealing with well-behaved, integrable action densities.

In Ref. 10 it is emphasized (p. 20), in the context of standard circle maps, that *discontinuous* rotation numbers can arise for *smooth* maps. Here they arise in the context of "annular" maps (the Appendix) when the annulus can become a disk. It is also closely related to the central property of our map that after each iteration "on the average" the phase turns twice as fast (the Appendix).

In Sec. II the construction of the periodic fields is reformulated, with respect to our previous papers, in order to display prominently the role of the iterative map. The self-dual solutions are characterized by *two* topological integers. One is "instantonlike" (P_T) and the other is "monopolelike" (q). Their remarkable combined role in index theorems with $R^3\times S^1$ as the base manifold for periodic fields[2,4,6,7] are recapitulated in the context of explicit construction of spinor solutions (Sec. IV).

Some special features of quasiperiodic fields are studied in Sec. III. For periodic backgrounds a finite number of normalizable, zero-mass spinor solutions are obtained by imposing on them (anti)periodic boundary conditions in t (Sec. IV). For quasiperiodic backgrounds the number cannot thus be limited, unless rational approximations of the component periods are considered. Nevertheless, the spinor solutions are constructed in a way that works for both classes. The aim (not realized here) is to study the effect on the spinors of *several* mutually incommensurable periods in the background field.

For some particularly simple systems the possible effects of quasiperiodic kicks have been studied by several authors with different conclusions. (See Ref. 11, and sources cited therein.)

In a more general context (quasiperiodic state with k frequencies and a quasiperiodic attractor for the value $\mu=\mu_0$ of a continuous bifurcation parameter) the situation has been summed up as follows (Ref. 12, p. 631):

"For $k\geqslant 3$, strange attractors and positive characteristic exponent may be present for μ arbitrarily close to μ_0."

We provide (though only for zero mass and Euclidean signature) exact, explicit solutions for spinors in four dimensions in a gauge field background that can bring into play an arbitrary number of mutually incommensurable periods. This can provide an interesting starting point for further investigations.

Our iterations can also be implemented in propagators (Sec. V). This can lead to a systematic semiclassical development for our classes of background.

Several directions are indicated (Sec. VI) for possible generalizations of our study. The possibilities mentioned are—breaking spherical symmetry, magnetic charge $q \geqslant 1$, gauge group SU(N) with $N > 2$ and the use of hyperbolic coordinates.

After presenting our formalism in full, certain general questions are taken up in the concluding remarks (Sec. VII).

II. A CLASS OF PERIODIC SELF-DUAL GAUGE FIELDS

The class we consider, to start with, has spherical symmetry in R^3 and periodicity in Euclidean time. The gauge group is SU(2). Let (t, r, θ, ϕ) be the standard time and radial coordinates with

$$ds^2 = dt^2 + dr^2 + r^2(d\theta^2 + (\sin\theta)^2 d\phi^2).$$

Let $(\sigma_r, \sigma_\theta, \sigma_\phi)$ denote the projections of the Pauli matrices, respectively, on the directions indicated. The gauge potentials are given by

$$A_r \pm iA_t = \pm i(\partial_r \pm i\partial_t)\zeta \frac{\sigma_r}{2}, \tag{5}$$

$$A_\theta \pm i(\sin\theta)^{-1}A_\phi = \pm i(e^\zeta - 1)\left(\frac{\sigma_\theta \pm i\sigma_\phi}{2}\right), \tag{6}$$

where

$$e^\zeta = \frac{r}{(1 - g\bar{g})}((\partial_r^2 + \partial_t^2)(g\bar{g}))^{1/2}. \tag{7}$$

Here g is a holomorphic function $g(r+it)$ in the (r,t) half-plane $(r \geqslant 0)$, postulated to satisfy the following properties:

(1) $g\bar{g} = 1 + O(r)$ for $r \to 0$,
(2) g has no poles for $r \geqslant 0$,
(3) g falls exponentially as $r \to \infty$ (giving a constant logarithmic derivative),
(4) g is periodic in t.

We start with strict periodicity. Quasiperiodicity will be defined and studied in Sec. III. Our previous studies of periodic instantons (Refs. 1–8) will be reformulated below to display at each stage the role of the iterative map (the Appendix).

A. First iteration: From monopoles to periodic instantons

The simplest choice satisfying all the constraints is evidently

$$g = e^{-k(r+it)} \quad (k > 0). \tag{8}$$

But now $g\bar{g}$ has no time dependence and from (7)

$$e^\zeta = \frac{kr}{\sinh kr}.$$

One obtains the famous self-dual BPS monopole with the magnetic topological winding number

$$q = 1.$$

Here we have, of course, the Euclidean version, A_t replacing the Higgs field.

In (8) g is precisely Z_0, the initial point of the iterative map studied in the Appendix. Apply one iteration. Then

$$g=\frac{a_0+e^{-k(r+it)}}{\overline{a_0}+e^{k(r+it)}}. \tag{9}$$

There is a spectacular change. One now has an authentic periodic solution characterized by *two* topological integers.

(1) The magnetic number remains unchanged since as $r\to\infty$ there is no essential change in the configuration. One has still the (monopolelike) number

$$q=1.$$

(2) A second (instantonlike) topological integer P_T is now given by the total action S_T over one period $(T=2\pi k^{-1})$ divided by $8\pi^2$. One defines

$$8\pi^2 P_T=4\pi\int_0^T dt\int_0^\infty dr(\partial_r^2+\partial_t^2)\omega, \tag{10}$$

where

$$\omega=\frac{1}{2}e^{2\zeta}+\ln\left(\frac{1-g\overline{g}}{r}\right),$$

$$(\partial_r^2+\partial_t^2)\omega=\frac{1}{2}(\partial_r^2+\partial_t^2)(e^{2\zeta}-2\zeta)=\frac{1}{2}(\partial_r^2+\partial_t^2)e^{2\zeta}+\frac{1}{r^2}(1-e^{2\zeta}). \tag{11}$$

For (9) one obtains

$$P_T=2.$$

The computation of the action will be given below in a form particularly suited to our purpose. But first let us introduce a more general form of g.

In (8) set, assuming k to be sufficiently large,

$$k=\sum_{j=1}^{n} k_j \quad (k_j>0).$$

At this stage the k_j's are supposed to be mutually commensurate. Thus, in evident notations,

$$g=\prod_{j=1}^{n} g_j^{(0)}=\prod_{j=1}^{n} e^{-k_j(r+it)}.$$

Now apply one iteration to each factor giving

$$g=\prod_{j=1}^{n} g_j^{(1)}=\prod_{j=1}^{n}\left(\frac{a_0^{(j)}+e^{-k_j(r+it)}}{\overline{a_0^{(j)}}+e^{k_j(r+it)}}\right). \tag{12}$$

In the notation of the Appendix,

$$g_j^{(1)}=Z_1^{(j)}.$$

Each factor now has its own period,

$$T_j = 2\pi k_j^{-1}. \tag{13}$$

Since the k_j's are mutually commensurate one can set

$$k_j = \hat{k}\frac{P_j}{Q_j}, \tag{14}$$

where P_j, Q_j are integers without common factor. Thus there is an overall period

$$T = \frac{2\pi}{\hat{k}}\left(\prod_j Q_j\right). \tag{15}$$

We now compute the total action over T. Using (10) and Stokes' theorem with the boundary indicated by the limits, one obtains

$$S_T = 4\pi\int_\delta^R dr[\partial_t\omega]_0^T + 4\pi\int_0^T dt[\partial_r\omega]_\delta^R \quad (\delta\to 0, R\to\infty). \tag{16}$$

The first integral vanishes due to periodicity. In the second one, nonzero contributions come only from the limit $r\to 0$. These can be evaluated by using standard integrals.[3,4] *But the very first step* $(p=0)$ *of our study (the Appendix) of the circle map and the small-r expansion furnishes the result directly.*

From (12) and (A38) as $r\to 0$,

$$g\bar{g} = 1 + 2\left(\sum_j \frac{d\psi_1^{(j)}}{dt}\right)r + 2\left(\sum_j \frac{d\psi_1^{(j)}}{dt}\right)^2 r^2 + O(r^3), \tag{17}$$

where

$$e^{-i\psi_1^{(j)}} = \left(\frac{a_0^{(j)} + e^{-ik_jt}}{\bar{a}_0^{(j)} + e^{ik_jt}}\right). \tag{18}$$

From (A19) as

$$k_jt = \psi_0^{(j)} \to \psi_0^{(j)} + 2\pi, \qquad \psi_1^{(j)} \to \psi_1^{(j)} + 4\pi. \tag{19}$$

Inserting ω of (10) in (16) and using (17), the only nonzero contributions are seen to come from

$$\lim_{r\to 0}\left(\partial_r \ln\left(\frac{1-g\bar{g}}{r}\right)\right) = \sum_j \frac{d\psi_1^{(j)}}{dt}. \tag{20}$$

Hence, in terms of the previous definitions, one obtains quite simply

$$S_T = 4\pi\left(4\pi\sum_{j=1}^n \frac{T}{T_j}\right) = 8\pi T\left(\sum_{j=1}^n k_j\right) = 8\pi^2 2\left(\prod_j Q_j\right)\left(\sum_{j=1}^n \frac{P_j}{Q_j}\right). \tag{21}$$

Hence,

$$P_T = 2\left(\sum_{j=1}^n \frac{T}{T_j}\right). \tag{22}$$

Since each (T/T_j) is an integer, P_T is an even integer. Odd indices can be obtained through a simple modification described below.

B. Comments on the limiting values $a_0^{(j)}=0,\ \pm 1$

In general, the parameters a will be assumed to satisfy $0<|a|<1$. The limits indicated are associated with monopoles and standard aperiodic instantons. The following points are to be noted.

(1) For

$$a_0^{(j)}=0, \quad \psi_1^{(j)}=2k_j t. \tag{23}$$

This limit, consistent with (19), can be taken smoothly. For all $a_0^{(j)}=0$ one has a static monopole. But having determined beforehand each T_j and T, if this limit is taken *a posteriori* one obtains (formally) (22) in the simplest fashion.

(2) For

$$a_0^{(j)}=1, \quad \psi_1^{(j)}=k_j t. \tag{24}$$

As compared to (23) there is a rescaling by $\frac{1}{2}$. [The choice -1 in (24) implies no essential change.] If say, only

$$a_0^{(1)}=1$$

one obtains, instead of (22),

$$P_T=\frac{T}{T_1}+2\left(\sum_{j=2}^{n}\frac{T}{T_j}\right). \tag{25}$$

Now P_T can be odd. Consider, for example,

$$g=\left(\frac{\frac{1}{2}(1+\epsilon)+e^{-k(r+it)}}{\frac{1}{2}(1+\epsilon)+e^{k(r+it)}}\right)\prod_{j=1}^{n}\left(\frac{a_j+e^{-k(r+it)}}{\overline{a_j}+e^{k(r+it)}}\right). \tag{26}$$

For

$$\epsilon=-1, \quad P_T=2n+2; \quad \epsilon=1, \quad P_T=2n+1. \tag{27}$$

(3) For each j, taking the combined limit

$$k_j\to 0, \quad a_0^{(j)}=-1+k_j b_j \quad (b_j+\overline{b_j}>0), \tag{28}$$

one obtains from (12)

$$g=\prod_{j=1}^{n}\left(\frac{b_j-(r+it)}{\overline{b_j}+(r+it)}\right). \tag{29}$$

This is Witten's multi-instanton solution[13] with centers on the time axis and total action on R^4,

$$S=8\pi^2(n-1). \tag{30}$$

The magnetic charge is lost in this limit.

To sum up, g as given by (12), combines instantonlike and monopolelike aspects. It is a more general construction containing them as limits. As an expression of this double role such periodic self-dual configurations are characterized by the presence of *two* topological integers: one instantonlike (P^T) and one monopolelike (q). The combined role of these two in the index theorems will be discussed later (Sec. VI). So far, though, P_T can range through the entire spectrum of integers, q is restricted to unity. This restriction can be removed (Sec. VI).

Finally, we just mention that concerning the rotation numbers defined in the Appendix, the result (A59) implies for (12), which has n factors, a staircaselike pattern with n steps. It was pointed in Sec. I that jumps in the rotation numbers can arise though the action density remains smooth. We illustrate this explicitly for the simplest nontrivial case. Let

$$g=\left(\frac{a+e^{-k(r+it)}}{\overline{a}+e^{k(r+it)}}\right). \tag{31}$$

Here the crucial value of r corresponding to (A59) is

$$e^{-kr}=a. \tag{32}$$

But now from (7),

$$e^{2\zeta}=(kr)^2((\sinh kr)^{-2}-(1-a^2)(\cosh kr+a\cos t)^{-2}). \tag{33}$$

Inserting (33) in (10) one obtains a smooth action density.

The results (A39) and (A40) can be used to study the time dependence of the action density in a small sphere around $r=0$. We defer the discussion to be able to include higher iterations and quasiperiodicity.

C. Higher iterations

For higher iterations one stays within the class of (quasi)periodic instantons with unit magnetic charge. But various interesting properties arise or are accentuated as the configuration becomes more complex. Some crucial features are presented below.

A generalization of the ansatz (12) is *automatically* implemented. One more iteration on (9) gives [see (A62)]

$$g=Z_2=Z_0^{-2}\left(\frac{a_0+Z_0}{\overline{a}_0+Z_0^{-1}}\right)\left(\frac{\mu_++Z_0}{\overline{\mu}_++Z_0^{-1}}\right)\left(\frac{\mu_-+Z_0}{\overline{\mu}_-+Z_0^{-1}}\right), \tag{34}$$

where

$$\mu_\pm=\tfrac{1}{2}((a_0+\overline{a}_0a_1)\pm((a_0+\overline{a}_0a_1)^2-4a_1)^{1/2}).$$

The factor Z_0^{-2} (i.e., $e^{2k(r+it)}$) is an *increasing* exponential in r, though Z_2 as a whole falls as e^{-kr}. This aspect is generalized by setting [compare (12)]

$$g=e^{l(r+it)}\prod_{j=1}^{n}\left(\frac{a_0^{(j)}+e^{-k_j(r+it)}}{\overline{a}_0^{(j)}+e^{k_j(r+it)}}\right) \tag{35}$$

with

$$l<\left(\sum_{j=1}^{n}k_j\right).$$

The total action, instead of (21), is now (assuming l is so chosen that the overall period is still T)

$$S_T=4\pi T\left(2\sum_{j=1}^{n}k_j-l\right). \tag{36}$$

Thus one has a lower action with the same number of parameters a. In (34) the factor Z_0^{-2} lowers S_T (and P_T) to give

$$(8\pi^2)^{-1}S_T = P_T = 2(3-1) = 4. \tag{37}$$

D. Effect on rotation numbers

The rotation numbers are defined and discussed in the Appendix [(A58)–(A65)]. In the course of time evolution of the gauge field generated by (34), g can vanish for

$$e^{-kr} = |a_0|, |\mu_+|, |\mu_-|, \tag{38}$$

though $|\mu_+|$ and $|\mu_-|$ may coincide to $|a_1|^{1/2}$. Thus though S_T is doubled, the number of zeros are here (generically) tripled.

Let $r_1 < r_2 < r_3$ be the three distinct roots of (38) in r and let Ω_2 denote the rotation number after the second iteration. Then (A59) is generalized as follows:

$$\frac{\Omega_2}{\Omega_0} = 4, \frac{7}{2}, 3, \frac{5}{2}, 2, \frac{3}{2}, 1, \tag{39}$$

respectively, for

$$r < r_1, \quad r = r_1, \quad r_1 < r < r_2, \quad r = r_2, \quad r_2 < r < r_3, \quad r = r_3, \quad r > r_3.$$

Cumulative effects of higher iterations will increase the number of steps. One can factorize Z_{p+2} in terms of Z_p analogously to (34). But rather than repeating such steps our aim is to show how the results of the Appendix can lead directly to remarkable properties. An example, in a different direction, follows.

E. Action density near the spatial origin

Let us explore the role of the iterations in the action density in a small sphere about the origin where the intricate interplay of (quasi)periodic pulsations are least affected by exponential damping with increasing r. We generalize (12) by setting

$$g = \prod_{j=1}^{n} Z_{p_j}^{(j)}, \tag{40}$$

where $Z_{p_j}^{(j)}$ is $e^{-k_j(r+it)}$ iterated p_j times. For each $p_j = 1$ one obtains (12). The doubling after each iteration of the contribution of each factor to the total action has to be taken into account. A simple particular case is given by (37). The general result is evident on noting how (A19) generalizes (19) and hence the derivation of (21). For (40) one obtains, instead of (21),

$$S_T = 8\pi T\left(\sum_{j=1}^{n} 2^{(p_j-1)}k_j\right). \tag{41}$$

As the action increases, the ''weight'' of that particular field configuration in path integrals diminishes. Hence the interest of an extra factor, as in (35), bringing down the action as far as possible for a given period, as in (36). Having obtained the total action let us now take a close look at a small sphere around $r=0$. Using the results obtained in (A38), (A39), and (A40) one obtains after simplifications for (40),

$$e^{2\zeta} = \frac{r^2}{(1-g\bar{g})^2}((\partial_r^2 + \partial_t^2)(g\bar{g})) = 1 - \frac{2}{3}\Gamma_t r^2 + O(r^3), \tag{42}$$

where

$$\Gamma_t = \tfrac{1}{2}C^2 + (\ddot{C}/C) - \tfrac{3}{2}(\dot{C}/C)^2, \tag{43}$$

$$C = \sum_{j=1}^{n} \dot{\psi}_{p_j}^{(j)} \equiv \dot{\Psi}. \tag{44}$$

Here the dots denote time derivatives and $\psi_{p_j}^{(j)}$ is the phase of the jth factor (for $r=0$) in (40). This leads, for ω given by (10), to

$$(\partial_r^2 + \partial_t^2)\omega = \tfrac{4}{3}\Gamma_t r^2 + O(r^3). \tag{45}$$

The total action over R^3 and a period T is given by (41). Let $S(r)$ denote the action, at any instant t, integrated over a small sphere of radius r about the origin. Then

$$S(r) = 4\pi \int_0^r dr(\partial_r^2 + \partial_t^2)\omega = \left(\frac{4}{3}\Gamma_t^2\right)V_r + \cdots, \tag{46}$$

where (the dots indicating higher powers of r)

$$V_r = \tfrac{4}{3}\pi r^3.$$

The leading term, logically, is proportional to the volume of the sphere. [This is rendered possible by the zero coefficient of the term linear in r in (42).]

In (43) Γ_t has an interesting structure. There is a ''kinetic'' term,

$$\tfrac{1}{2}C^2 = \tfrac{1}{2}\dot{\Psi}^2$$

and the *Schwarzian derivative* of Ψ. For the simplest case [one factor with $p=1$ in (40)] using (A17), with

$$a = |a_0|, \quad \chi = \psi_0 - \alpha_0, \quad \psi_0 = kt,$$

$$C = \dot{\Psi} = 2k\frac{1 + a\cos\chi}{1 + a^2 + 2a\cos\chi} \equiv 2k\frac{X}{Y}.$$

Hence

$$\frac{\ddot{C}}{C} - \frac{3}{2}\left(\frac{\dot{C}}{C}\right)^2 = k^2 a(1-a^2)\left(\cos\chi + a(\sin\chi)^2\left(\frac{3}{2}X^{-1} + Y^{-1}\right)\right)(XY)^{-1}. \tag{47}$$

This changes sign at points determined by the choice of a_0. But for all a_0, it is positive (negative) for $\chi = 0(\pi)$, respectively. For the general case (40) the very complex structure implied by (43) determines the time dependence.

III. QUASIPERIODICITY

So far we have been considering strict periodicity, having just mentioned that the component periods T_j need not necessarily be commensurate. We now take a closer look at this possibility.

As an example let $(k_2, k_3, \ldots, k_n)$ in (12) be all mutually commensurate as in (14) but not with k_1. [This relatively simple case will suffice to exhibit some remarkable features. It is possible to consider several incommensurate periods. One can also start with Z_0 given by (3) rather than by (2).]

Let $\hat{T}$ be the common overall period for $(T_2, T_3, \ldots, T_n)$, but *not* for T_1. One can take successive rational approximations of $(\hat{T}/T_1)$,

$$\left(\frac{\hat{T}}{T_1}\right)_{(\text{appr})} = \frac{N_1}{N_2}, \tag{48}$$

N_1, N_2 being integers without common factors. Then

$$T = N_2 \hat{T}$$

is the overall period at that approximation. As the approximation is improved, N_1, N_2, T all increase without limit. (Since there is no exact period, improving the approximation one tends to cover the entire time axis.) One can keep in mind the famous example of successive rational approximations of the ''golden mean''

$$G = \tfrac{1}{2}(\sqrt{5} - 1)$$

using the Fibonacci sequence. But our considerations will not be limited to any such particular case. For the lth approximation, in evident notations, the action over the period $T^{(l)}$ is

$$S_l = 16\pi^2 \left(\frac{T^{(l)}}{T_1} + \sum_{j=2}^{n} \frac{T^{(l)}}{T_j} \right). \tag{49}$$

As $l \to \infty$ so do $T^{(l)}$ and S_l. One can now consider the limiting form of the action per unit time or the ''normalized'' action

$$S_N = \left(\frac{S_l}{T^{(l)}} \right)_{l \to \infty}. \tag{50}$$

This gives a *continuous index*, a positive real number, not an integer. The possible mathematical significance of S_N has been discussed elsewhere.[4,7] The magnetic index q does *not* vary in the successive stages described above. As one approaches the asymptotic S_2 in R^3 the exponential damping of time dependences give the same static configuration at each stage, namely that of the monopole. Hence our *quasiperiodic instantons are characterized, not by two topological integers, but by one positive real number* (S_N) *and one integer* (q). For the class of solutions under consideration $q = 1$. More general possibilities are indicated in Sec. VI.

So far we have been looking at quasiperiodic solutions from the point of view of successive approximations. Let us now look at some *exact* consequences of the postulated incommensurability.

In the case considered before, consider the Poincaré sections of the factor g_1 (corresponding to the period T_1) for

$$t = t_0 + nT \quad (n = 0,1,2,\ldots).$$

One obtains

$$g_1(n) = e^{-i2\alpha_1} \frac{|a_1| + e^{-k_1(r + i(t_0 + nT)) + i\alpha_1}}{|a_1| + e^{k_1(r + i(t_0 + nT)) - i\alpha_1}}. \tag{51}$$

We simplify notations by setting

$$a = |a_1| = a_1 e^{i\alpha_1}, \quad k_1 = k, \quad k\omega = kt_0 - \alpha_1,$$

$$X^{(n)}=e^{i2\alpha_1}g_1(n)=\frac{a+F_n}{a+F_n^{-1}},$$

where

$$F_n=e^{-k(r+i\omega+inT)}=e^{-k(r+i\omega)-i2\pi n\delta},\quad \delta\neq\frac{N_1}{N_2}\quad (N_1,N_2\ \text{integers}).$$

The irrationality of δ *is the basic quasiperiodicity postulate.* Now

$$X^{(n+p)}=\frac{a+F_{n+p}}{a+F_{n+p}^{-1}}=\frac{a+F_ne^{-i2\pi p\delta}}{a+F_n^{-1}e^{i2\pi p\delta}}. \tag{52}$$

We show that if the parameters $(r,t_0,k,\alpha,n,p,\delta)$ are fine-tuned, X may come back *just once* to an initial value but no more (*not even twice*). From (52),

$$F_n^2+e^{i2\pi p\delta}a(1-X^{(n+p)})F_n-e^{i4\pi p\delta}X^{(n+p)}=0. \tag{53}$$

Eliminating F_n from the equations for three *distinct* integer values of p say,

$$p=p_1,p_2,p_3,$$

one of which may be zero, one has as a necessary constraint the vanishing determinant

$$\det\begin{vmatrix}1 & a(1-X^{n+p_1})e^{i2\pi p_1\delta} & -X^{n+p_1}e^{i4\pi p_1\delta}\\ 1 & a(1-X^{n+p_2})e^{i2\pi p_2\delta} & -X^{n+p_2}e^{i4\pi p_2\delta}\\ 1 & a(1-X^{n+p_3})e^{i2\pi p_3\delta} & -X^{n+p_3}e^{i4\pi p_3\delta}\end{vmatrix}=0. \tag{54}$$

For, say,

$$X^{(n+p_1)}=X^{(n+p_2)}=X^{(n+p_3)}=X \tag{55}$$

this reduces to

$$X(1-X)e^{i(p_1+p_2+p_3)2\pi\delta}((1-e^{i(p_1-p_2)2\pi\delta})(1-e^{i(p_2-p_3)2\pi\delta})(1-e^{i(p_3-p_1)2\pi\delta}))=0. \tag{56}$$

The coefficient of $X(1-X)$ cannot vanish due to our quasiperiodicity postulate. Neither can $X^{(n+p)}$ be zero or unity for three distinct values of p. That would imply, for example for $X=1$,

$$F_n^2=e^{i4\pi p_j\delta}$$

for three different values of p_j and so on. Hence the constraint (55) cannot be satisfied. So $X^{(n)}$ cannot come back twice *exactly* to a previous value. There is, however, no restriction concerning repeated very close approaches.

Let us now look at the conditions necessary for a *single* return, namely,

$$X^{(n+p)}=X^{(n)}.$$

From (52) this is seen to imply

$$(F_ne^{-i\pi p\delta})^2+2\lambda(F_ne^{-i\pi p\delta})+1=0, \tag{57}$$

where

$$\lambda=a^{-1}\cos\pi p\delta.$$

Hence,

$$F_n = e^{i\pi p\delta}(-\lambda \pm \sqrt{\lambda^2 - 1}). \tag{58}$$

Case (1): $(|\lambda| \leqslant 1)$

Let $\lambda = \cos\eta$ (η real). Then

$$F_n = e^{-k(r+i\omega)-i2\pi n\delta} = -e^{i(\pi p\delta \pm \eta)}. \tag{59}$$

This implies $r=0$ and

$$e^{-i(kt_0 - \alpha_1 + \pi(2n+p)\delta \pm \eta)} = -1. \tag{60}$$

(The $\pm$ sign implies one or the other. They do not hold simultaneously.)

Case (2): $(|\lambda| > 1)$

Let $\lambda = \coth\zeta$ (ζ real). Then, retaining the solution leading to a real value of r,

$$F_n = -(\tanh(\zeta/2))e^{i\pi p\delta} \tag{61}$$

and

$$e^{-kr} = |\tanh(\zeta/2)|, \quad e^{-i(kt_0 - \alpha_1 + \pi(2n+p)\delta)} = \pm 1. \tag{62}$$

It should be noted that $X^{(n+p)} = X^{(n)}$ does not imply a periodic situation. In the solutions above n is not arbitrary for a given p. One can indeed show that given the foregoing situation

$$X^{(n+2p)} \neq X^{(n+p)} = X^{(n)}$$

consistently with the impossibility of (55). In this context the ambiguity of sign of the inverse map (expressing F_n in terms of X_n) plays an interesting role. But we will not present such details here.

Let us note one more point. *The action density and the rotation numbers now involve mutually incommensurable numbers.* It would be interesting to study in detail the time evolution of the coefficient (43)

$$\Gamma_t = \tfrac{1}{2}C^2 + (\ddot{C}/C) - \tfrac{3}{2}(\dot{C}/C)^2, \tag{63}$$

$$C = \sum_{j=1}^{n} \dot{\psi}_{p_j}^{(j)} \equiv \dot{\Psi} \tag{64}$$

with three or more incommensurable periods (namely, k_j's at the origin of the ψ_j's) varying the parameters a. Our purpose in leaving the period T and Ω_0 in (A58) is now evident. In general one will have, to start with, incommensurable $\Omega_0^{(j)}$'s corresponding to different factors of g. For the simple case considered above (only k_1 incommensurable with the rest and with F_n corresponding to a section of $Z_0^{(1)}$) one has

$$\Omega_0^{(1)} = 2\pi\delta,$$

δ being irrational. The cumulative effects of several irrational numbers have to be taken into account in a more general case.

The different roles of rational and irrational rotation numbers are well known.[9,10] Here the rotations of the phases are considered in the context of annular maps (the Appendix). The ''average'' doubling of the phase (sufficiently near the origin) for each iteration is also specific to our case. Though we have demonstrated the existence of striking discontinuities, the role of the rotation number remains to be explored.

IV. INDEX THEOREMS AND SPINOR SOLUTIONS

Again we start with the periodic case, deferring a discussion of quasiperiodicity. For periodic gauge fields the base manifold is $R^3 \times S^1$ rather than R^4 or S^4. The index theorems have to take into account the boundary effects induced by S^1. This has been discussed elsewhere.[4,6,7] Here we just mention that the number of zero modes of spinors in periodic backgrounds characterized by two topological integers (P_T, q). $= P_T - q$ for periodic spinors of isospin $\frac{1}{2}$ (see Refs. 2 and 4), $= 4P_T - 2q$ for periodic spinors of isospin 1 (see Ref. 7), $= P_T$ for antiperiodic spinors of isospin $\frac{1}{2}$ (see Ref. 6).

(The background is, of course, still periodic for antiperiodic spinors.) We construct below the spinor solutions in a more general and systematic fashion than in the previous papers. This is facilitated by introducing a gauge transformation leading from the Witten-type gauge introduced in (5) and (6) to a (quasi)periodic generalization of 't Hooft or Jackiw–Nohl–Rebbi (JNR) type solutions.[2,4,6,7] Two historical references for the standard (aperiodic) solutions are Refs. 13 and 14.

A. Gauge transformation

We start with A_μ given by (5) and (6) and gauge transform by

$$G(r,t) = e^{-id(r,t)\sigma_r/2},$$

where

$$e^{id} = \left(\frac{1-g}{1-\bar{g}}\right)\left(\frac{(\partial_r - i\partial_t)g}{(\partial_r + i\partial_t)\bar{g}}\right)^{1/2}. \tag{65}$$

One obtains

$$A'_\mu = G^{-1}A_\mu G - iG^{-1}\partial_\mu G = \sigma_{\mu\nu}\partial_\nu \ln \Sigma,$$

where $\sigma_{10} = \sigma_{23} = \sigma_1$ (cyclic), and

$$\Sigma = \frac{1}{2r}\left(\frac{1+g}{1-g} + \frac{1+\bar{g}}{1-\bar{g}}\right) = \frac{1}{r}\,\frac{(1-g\bar{g})}{(1-g)(1-\bar{g})} \tag{66}$$

satisfying, since $(\partial_r + i\partial_t)g = 0 = (\partial_r - i\partial_t)\bar{g}$,

$$\Box\Sigma = \left(\partial_t^2 + \partial_r^2 + \frac{2}{r}\partial_r\right)\Sigma = 0. \tag{67}$$

Thus we have obtained the famous ansatz leading to 't Hooft or JNR solutions. But here it is being implemented in the context of magnetically charged (quasi)periodic solutions rather than that of standard aperiodic instantons.

The poles of Σ play an essential role in the construction of spinor solutions. *The passage from g to Σ maintains contact with the iterative map implemented for g*. But this raises the problem of displaying in an additive form the poles of Σ. The roots of $g=1$ are not, in general, explicitly available. Generally there are $2n$ roots for (12), all for $r=0$. The particular case, with $(0<a<1)$ and

$$g = \left(\frac{a+e^{-(r+it)}}{a+e^{(r+it)}}\right)^n$$

is fully treated in Ref. 7. We display some typical features of the passage from g to Σ through simple examples, motivating the generalization to follow.

(1) For $g = e^{-k(rit)}$,

$$\frac{1+g}{1-g} = \coth\frac{1}{2}k(r+it),$$
$$\Sigma = \frac{1}{2r}\left(\coth\frac{1}{2}k(r+it) + \coth\frac{1}{2}k(r-it)\right) = \sum_{l=-\infty}^{\infty}\frac{1}{k^2}\frac{1}{r^2+(t-k^{-1}2\pi l)^2}. \tag{68}$$

This is the periodic form of the BPS monopole, gauge equivalent with the still better-known static form. The action over one period $(T=2\pi/k)$ can be considered, quite formally, to be $8\pi^2$ (or $P_T=1$). [See the comments following (8) and (23).]

(2) For, with $(0<a<1)$,

$$g = \left(\frac{a+e^{-k(r+it)}}{a+e^{k(r+it)}}\right), \quad P_T=2,$$

and

$$\frac{1+g}{1-g} = \frac{1+a}{2}\coth\frac{1}{2}k(r+it) + \frac{1-a}{2}\coth\frac{1}{2}k(r+it-i\pi k^{-1}). \tag{69}$$

(3) For

$$g = e^{-k(r+it)}\left(\frac{a+e^{-k(r+it)}}{a+e^{k(r+it)}}\right),$$

$$P_T=3 \quad [\text{the simplest case of (26) and (27)}]$$

and

$$\frac{1+g}{1-g} = \lambda_1^2\coth\frac{1}{2}k(r+it) + \lambda_2^2\coth\frac{1}{2}k(r+it-ick^{-1}) + \lambda_3^2\coth\frac{1}{2}k(r+it+ick^{-1}), \tag{70}$$

where

$$\cos c = -\frac{1+a}{2}, \quad \lambda_1^2 = \frac{1+a}{3+a}, \quad \lambda_2^2 = \lambda_3^2 = \frac{1}{3+a}.$$

For our normalization a general consequence is

$$\sum_i \lambda_i^2 = 1.$$

The number of coth terms in $(1+g)/(1-g)$ gives P_T if each one has the same period, say T and *distinct singularities*. A term with period (T/m) contributes m units to P_T.

(4) Once the roots of $g=1$ are obtained the residues at the poles give the λ's. But this is not strictly necessary for constructing the spinor solutions.

Consider a case with two different periods $(2\pi k^{-1}, \pi k^{-1})$, namely $(0<a<1)$ and

$$g = \left(\frac{a+e^{-k(r+it)}}{a+e^{k(r+it)}}\right)\left(\frac{a+e^{-2k(r+it)}}{a+e^{2k(r+it)}}\right).$$

Now

$$P_T = 2 + 2.2 = 6, \quad q = 1. \tag{71}$$

The zeros of g (and hence the jumps in rotation numbers) correspond to

$$e^{-kr} = a, \sqrt{a}.$$

It is sufficient to note that $g=1$ for

$$r=0, \quad e^{it} = \pm 1, \left(-\frac{1}{2} \pm i\frac{\sqrt{3}}{2}\right), \left(\frac{1-a}{2} \pm i\sqrt{1-\left(\frac{1-a}{2}\right)^2}\right).$$

One can then construct $(P_T - q) = 5$ periodic and $P_T = 6$ antiperiodic spinor solutions. The explicit solutions will follow.

More generally, using real parameters a_p at each iteration one obtains some simplifications. Thus, for example, iterating g as a whole (with $0 < a_p < 1$).

$$g_{p+1} = \frac{a_p + g_p}{a_p + g_p^{-1}}, \tag{72}$$

and

$$\frac{1+g_{p+1}}{1-g_{p+1}} = -\left(\frac{1+a_p}{2}\left(\frac{1+g_p}{1-g_p}\right) + \frac{1-a_p}{2}\left(\frac{1-g_p}{1+g_p}\right)\right). \tag{73}$$

Hence $g_{p+1} = 1$ for $g_p = \pm 1$, consistently with the doubling of the periodic action. For real a_0, a_1, and

$$g_0 = e^{-k(r+it)},$$

$$g_2 = 1 \quad \text{for } r=0; \quad \sin\frac{kt}{2} = 0, \quad \cos\frac{kt}{2} = 0, \quad \pm\sqrt{\frac{1-a_0}{2}}.$$

B. Generalization of Z_0 and g

Having noted the relation between the structures of g and Σ one can now invert the procedure. Start with

$$\Sigma = \frac{1}{2r}\sum_{l=1}^{n} \lambda_l^2 \left(\coth\frac{1}{2}k_l(r+it-ic_l) + \coth\frac{1}{2}k_l(r-it+ic_l)\right) \tag{74}$$

$$= \frac{1}{2r}\left(\frac{1+g}{1-g} + \frac{1+\bar{g}}{1-\bar{g}}\right), \tag{75}$$

where convenient conventions are

$$\sum_{l=1}^{n} \lambda_l^2 = 1, \quad k_n > k_{n-1} > \cdots > k_2 > k_1 > 0.$$

Now

$$g = \frac{\Sigma_{l=1}^{n} \lambda_l^2 (e^{k_l(r+it-ic_l)} - 1)^{-1}}{\Sigma_{l=1}^{n} \lambda_l^2 (1 - e^{-k_l(r+it-ic_l)})^{-1}}. \tag{76}$$

Thus we have motivated the parametrization (3) of Z_0 (here g). One can verify that (76) satisfies all the constraints listed below (7). In particular, g is a phase for $r=0$ and drops as $e^{-k_1 r}$ as r becomes large. *The major interest is that now, even in presence of an arbitrary number of different periods (different k's) g is directly adapted to the construction of spinor solutions to follow.* The Σ of (74) can easily be generalized to break spherical symmetry. But we defer such considerations to Sec. VI.

To give a better idea of the relation of (76) to the previous factorized form we consider again some simple examples. Different subclasses of the previous form will be found already in the simplest examples of (76).

We set $n=2$; $c_1=c_2=0$ and we use below often $\lambda_1^2+\lambda_2^2=1$.

(1) For $k_1=k_2=k$,

$$g=e^{-k(r+it)}.$$

More generally, for all k's and all c's equal one has the monopole solution.

(2) For $k_1=k,\ k_2=2k$,

$$g=\frac{\lambda_1^2+e^{-k(r+it)}}{\lambda_1^2+e^{k(r+it)}}.$$

(3) For $k_1=k,\ k_2=3k$,

$$g=e^{k(r+it)}\left(\frac{a+e^{-k(r+it)}}{\bar{a}+e^{k(r+it)}}\right)\left(\frac{\bar{a}+e^{-k(r+it)}}{a+e^{k(r+it)}}\right),$$

where $a=\frac{1}{2}(\lambda_1^2+i\lambda_1\sqrt{4-\lambda_1^2})\quad (a\bar{a}=\lambda_1^2<1)$.

(4) For $k_1=2k,\ k_2=3k$,

$$g=\left(\frac{a_1+e^{-k(r+it)}}{\bar{a}_1+e^{k(r+it)}}\right)\left(\frac{a_2+e^{-k(r+it)}}{\bar{a}_2+e^{k(r+it)}}\right),$$

where, respectively, for

$$1>\lambda_1^2>\tfrac{1}{4},\quad a_1=\bar{a}_2=\tfrac{1}{2}(1+i\sqrt{4\lambda_1^2-1});$$

$$\tfrac{1}{4}>\lambda_1^2>0,\quad a_1=\bar{a}_1=\tfrac{1}{2}(1+\sqrt{1-4\lambda_1^2});$$

$$a_2=\bar{a}_2=\tfrac{1}{2}(1-\sqrt{1-4\lambda_1^2});$$

$$\lambda_1^2=\tfrac{1}{4},\quad a_1=a_2=\tfrac{1}{2}.$$

For special values of the k's and the c's in (76) some poles can coincide, diminishing P_T accordingly. The preceding examples, apart from leading to simple factorized expressions for g also illustrate this point. One degeneracy (the common pole for $r=0$, $\sin kt=0$) diminishes P_T by 1 in each case. This is consistently incorporated in the factorized forms of g. Thus for $k_1=2k$, $k_2=3k$ one obtains $P_T=4$ instead of 5. We will assume, in general, that the choice of parameters in Σ imply distinct poles.

Spinor solutions for zero mass were constructed[2] directly in Witten-type gauge. But since the counting of the number of zero modes (and hence comparison with the index theorems) is more transparent in the 't Hooft- or JNR-type gauge, we start by generalizing (74) as follows. Let

$$A_\mu=\sigma_{\mu\nu}\partial_\nu\ln\Sigma,$$

where, with $\epsilon=\pm1$,

$$\Sigma=\frac{1}{2}(1+\epsilon)+\frac{1}{2r}\sum_{j=1}^{n}\lambda_j^2\left(\coth\frac{1}{2}k_j(r+it-ic_j)+\coth\frac{1}{2}k_j(r-it+ic_j)\right)$$
$$=\frac{1}{2}(1+\epsilon)+\sum_{j=1}^{n}\sum_{l=-\infty}^{\infty}\frac{\lambda_j^2}{k_j^2}\frac{1}{r^2+(t-k_j^{-1}2\pi l)^2}. \tag{77}$$

The role of ϵ is interesting concerning the index theorems. For $\epsilon=1$, one has a 't Hooft-like form. It can be shown that on the asymptotic sphere in R^3 one now has a dipole like configuration rather than a monopole. (See the discussion in Ref. 6 and the sources cited there.) Hence for

$$\epsilon=1, \quad q=0,$$

while for

$$\epsilon=-1, \quad q=1.$$

However, for *both* values $(\epsilon=\pm 1)$, the periodic action is the same. For an overall period T and integers n_j such that

$$T_j=2\pi k_j^{-1}, \quad T=2\pi k^{-1}, \quad k_j=n_jk, \quad S_T=8\pi^2\sum_j n_j. \tag{78}$$

This can be seen as a limiting case of the familiar ('t Hooft and JNR) solutions as follows. Start by retaining in (77) the poles covering an interval *NT*. Define, in evident notations,

$$S_T=(N^{-1}S_{NT})_{N\to\infty}.$$

The value of S_{NT} is given by the very well-known results for the 't Hooft and JNR solutions leading to

$$S_T=\frac{8\pi^2}{N}\left(N\sum_j n_j-(1-\epsilon)/2\right)_{N\to\infty}=8\pi^2\sum_{j=1}^{n}n_j.$$

Thus for the periodic case, in contrast with the aperiodic one, the change from $\epsilon=-1$ *to* $\epsilon=1$ *does not increase* P_T *by* 1 *but diminishes* q *by* 1. For $n=1$, one has quite a special case. Then for $\epsilon=-1$, one has a static monopole gauge transformed to a periodic form, whereas for $\epsilon=1$ the periodicity is authentic but $q=0$.

C. Spinor solutions

Having thoroughly prepared the ground, we at last introduce the spinors. Here we consider[2,6] only isospin

$$I=\tfrac{1}{2}$$

(see Ref. 7 for $I=1$). We consider only zero mass spinors.

For our conventions only upper (positive) helicity spinors have normalizable solution. Separating this helicity (Ψ_U) the Dirac equation reduces to

$$\overline{\alpha_\mu}(i\partial_\mu-A_\mu)\Psi_U=0,$$

where

$$\overline{\alpha_\mu}=(\tau_0,i\vec{\tau}), \quad \alpha_\mu=(\tau_0,-i\vec{\tau}).$$

(We denote space–time Pauli matrices by τ_μ and those in the isospace by σ_μ.) The isospin components are separated as

$$\Psi_U = \begin{vmatrix} \Psi_U^{(+)} \\ \Psi_U^{(-)} \end{vmatrix}, \quad I_3 \Psi_U^{(\pm)} = \pm \frac{1}{2} \Psi_U^{(\pm)}.$$

Set

$$\Psi_U^{(\pm)} = \Sigma^{1/2} \begin{vmatrix} a_\pm \\ b_\pm \end{vmatrix}$$

and define

$$u = \tfrac{1}{2}(x_3 + i x_0), \quad \overline{u} = \tfrac{1}{2}(x_3 - i x_0),$$

$$v = \tfrac{1}{2}(x_1 + i x_2), \quad \overline{v} = \tfrac{1}{2}(x_1 - i x_2).$$

Formally, the solutions are given (see Ref. 6 and the sources cited therein) by

$$a_+ = \partial_v (\Sigma^{-1} H), \quad b_+ = -\partial_u (\Sigma^{-1} H),$$

$$a_- = -\partial_{\overline{u}} (\Sigma^{-1} H), \quad b_- = -\partial_{\overline{v}} (\Sigma^{-1} H),$$

where

$$\Box H = (\partial_u \partial_{\overline{u}} + \partial_v \partial_{\overline{v}}) H = 0.$$

Normalizable solutions are obtained by matching the poles of H with the zeros of Σ^{-1}. From (77) it is evident that there are an infinite number of such possible choices of H. A finite number is obtained by imposing suitable boundary conditions relating $\Psi_U(t)$ and $\Psi_U(t+T)$, $T = 2\pi K^{-1}$ being the period.

For *periodic* spinors satisfying

$$\Psi_U(t) = \Psi_U(t+T)$$

set

$$H_{m_j} = \frac{1}{2r} \left(\coth \frac{1}{2} K(r + i(t - c_j) - i 2\pi k_j^{-1} m_j) + \coth \frac{1}{2} K(r - i(t - c_j) + i 2\pi k_j^{-1} m_j) \right) \tag{79}$$

with $T_j = 2\pi k_j^{-1}$ and

$$m_j = 1,2,\ldots,n_j, \quad k_j = n_j K \quad (j = 1,2,\ldots,n).$$

These provide all the normalizable zero modes. The total number is

$$\left(\sum_{j=1}^{n} n_j \right). \tag{80}$$

But they are not all necessarily independent since

$$\frac{1}{n_j} \sum_{m_j=1}^{n_j} \coth \frac{1}{2} K(r \pm i(t - c_j - 2\pi k_j^{-1} m_j)) = \coth \frac{1}{2} k_j (r \pm i(t - c_j)). \tag{81}$$

(This is probably most easily derived via the logarithmic derivative of the well-known product formula for $\sinh nx$.) Hence

$$\sum_j n_j^{-1}\lambda_j^2\left(\sum_{m_j} H_{m_j}\right)=\Sigma-\frac{1}{2}(1+\epsilon). \tag{82}$$

Thus for $\epsilon=-1$, in an evident notation,

$$\sum_j n_j^{-1}\lambda_j^2\left(\sum_{m_j} \Psi_U^{(m_j)}\right)=0. \tag{83}$$

There is no such constraint for $\epsilon=1$. Moreover, for *both* cases $(\epsilon=\pm 1)$, as explained before

$$P_T=\left(\sum_{j=1}^{n} n_j\right).$$

Thus in both cases, consistent with the index theorems stated at the beginning of this section, the number of zero modes $=P_T-q$.

Here q is limited to zero and one. But, on the other hand, we have done much more than counting the number of possible solutions. They have been obtained explicitly. One can now study, to take only one example, the time evolution of such spinor densities.

For *antiperiodic* spinors (in periodic backgrounds)

$$\Psi_U(t)=-\Psi_U(t+T)$$

and the correct choice[6] for H turns out to be

$$H_{m_j}=\frac{1}{2r}\left(\operatorname{cosech}\frac{1}{2}K(r+i(t-c_j)-i2\pi k_j^{-1}m_j)+\operatorname{cosech}\frac{1}{2}K(r-i(t-c_j)+i2\pi k_j^{-1}m_j)\right). \tag{84}$$

The linear constraint (83) is now absent even for $q=1$ $(\epsilon=-1)$. *There is no "magnetic defect,"* no subtraction of q, and one obtains (consistently with the result stated at the beginning of this section but now via explicit constructions)

$$\text{number of zero modes}=P_T.$$

Static configurations can be considered as a limiting case of periodic ones (infinite period) but not of antiperiodic ones. The Dirac modes in a monopole background that introduce the q subtraction as a boundary effect are not relevant for antiperiodic spinors.

D. Spinors in quasiperiodic backgrounds

Consider now the case where in (77) the k's are *not* all mutually commensurable. Generalizing the approximation (48) to several component irrational ratios one may construct (anti)periodic spinor solutions at each level of rational approximation. As this approximation is improved P_T (approx) and hence the number of spinor modes will diverge. There being no exact period for the gauge field one cannot limit the number of spinor modes by imposing (anti)periodic boundary conditions as before. *But our previous solutions provide a subset of "typical" ones which permit a comparative study of time* evolution of spinors in periodic and quasiperiodic backgrounds respectively. Consider, for example, the subset (periodic and antiperiodic, respectively, for a periodic background) corresponding to

$$H_j^{(+)}=\frac{1}{2r}\left(\coth\frac{1}{2}k_j(r+it-ic_j)+\coth\frac{1}{2}k_j(r-it+ic_j)\right), \tag{85}$$

$$H_j^{(-)}=\frac{1}{2r}\left(\operatorname{cosech}\frac{1}{2}k_j(r+it-ic_j)+\operatorname{cosech}\frac{1}{2}k_j(r-it+ic_j)\right). \tag{86}$$

Suppose we start with mutually commensurable k's and then vary some of them (infinitesimally or more) to introduce incommensurability and quasiperiodicity. $H_j^{(\pm)}$ will continue to give solutions, normalizable over any rationally approximated period or over unit time [analogously to (50)]. One can express the spinor densities

$$(\Psi_U^{(\pm)})^\dagger\Psi_U^{(\pm)}, \quad (\Psi_U^{(\pm)})^\dagger\vec{\tau}\Psi_U^{(\pm)}$$

in terms of $a_\pm$, $b_\pm$, and Σ. One can implement the constraints due to the harmonic property of Σ and the H's and (for our present case) those due to spherical symmetry. Then it might be rewarding to follow the time evolution of the densities for different values (small or some other crucial ones) of r, particularly when there are three or more incommensurable k's. Such a study is, however, entirely beyond the scope of this paper.

Let us finally note that starting with (76) as g_0 and iterating with real parameters as in (72), namely,

$$g_{p+1}=\frac{a_p+g_p}{a_p+g_p^{-1}},$$

the roots of $g_0=1$ will form a subset of those of $g_p=1$. Hence defining

$$\Sigma_p=\frac{1}{2r}\left(\frac{1+g_p}{1-g_p}+\frac{1+\overline{g_p}}{1-\overline{g_p}}\right) \tag{87}$$

and still retaining the explicit expressions (85), (86) for $H_j^{(\pm)}$ obtained for Σ_0, one obtains an interesting subset of spinor modes for the background corresponding to Σ_p, for both cases (periodic and quasiperiodic). At the level of g_0 the space–time dependence can already be quite complex. With iterations this will become much more so. But we will still have explicit solutions whose evolutions can be studied.

Note that our spinor solutions (except periodic ones for $q=0$) fall off exponentially for large r. For antiperiodic ones this is more evident. But though the leading term (for $q=1$) in (H_j/Σ) is constant for periodic spinors, the presence of derivatives introduce again exponential damping. For $q=0$ periodic spinor densities fall off as r^{-4}. We have already studied the action density of the gauge fields near the origin (Sec. II). One can now study our ''exponentially confined'' fermion densities near the origin in such backgrounds.

V. PROPAGATORS

For the gauge field backgrounds considered in the preceding sections the propagators for massless, isospin $\frac{1}{2}$ fields were given in Ref. 3. For periodic backgrounds the (anti)periodic propagators were presented in explicitly summed up, closed forms. *Our iterative map can be implemented in them* through the functions $g(\overline{g})$ corresponding to the points x and y of the propagator $\Delta(x,y)$. We recapitulate briefly the results of Ref. 3, where other sources are cited.

Define

$$t=x_0, t'=y_0, \quad r=\sqrt{\vec{x}^2}, \quad r'=\sqrt{\vec{y}^2},$$

$$\sigma_r=\frac{\vec{\sigma}\cdot\vec{x}}{r}, \quad \sigma_{r'}=\frac{\vec{\sigma}\cdot\vec{y}}{r'},$$

$$G=\frac{1+g(r+it)}{1-g(r+it)}, \quad G'=\frac{1+g(r'+it')}{1-g(r'+it')},$$

$$\Sigma=\frac{1}{2r}(G+\overline{G}),\quad \Sigma'=\frac{1}{2r'}(G'+\overline{G'}),$$

and

$$-i2F(r,t;r',t')=\frac{(1-\sigma_r)(1+\sigma_{r'})}{(t-t')+i(r-r')}(G-G')-\frac{(1+\sigma_r)(1-\sigma_{r'})}{(t-t')-i(r-r')}(\overline{G}-\overline{G'})$$
$$+\frac{(1-\sigma_r)(1-\sigma_{r'})}{(t-t')+i(r+r')}(G+\overline{G'})-\frac{(1+\sigma_r)(1+\sigma_{r'})}{(t-t')-i(r+r')}(\overline{G}+G'). \tag{88}$$

Let unprimed fields (A_μ) correspond to x and primed ones (A'_μ) to y. Let

$$\tilde{\Delta}(x,y)=\Sigma^{-1/2}\frac{F}{4\pi^2(x-y)^2}\Sigma'^{-1/2} \tag{89}$$

and

$$D_\mu=(\partial_\mu+iA_\mu)(\partial_\mu+iA_\mu).$$

Then

$$-D^2\tilde{\Delta}(x,y)=\delta^4(x-y). \tag{90}$$

Thus $\tilde{\Delta}$ gives the *aperiodic* propagator for massless scalar fields. For $G(G')$ periodic in $t(t')$ with a period T, say, the *periodic* and the *antiperiodic* propagators are, respectively, defined to be

$$\Delta_+(x,y)=\sum_{n=-\infty}^{\infty}\tilde{\Delta}(x_0+nT,\vec{x};y_0,\vec{y}), \tag{91}$$

$$\Delta_-(x,y)=\sum_{n=-\infty}^{\infty}(-1)^n\tilde{\Delta}(x_0+nT,\vec{x};y_0,\vec{y}). \tag{92}$$

Define

$$V_1=(t-t')+i|\vec{x}-\vec{y}|,\quad V_2=(t-t')-i|\vec{x}-\vec{y}|,$$
$$V_3(\epsilon,\epsilon')=(t-t')+i(\epsilon r+\epsilon' r'),\quad (\epsilon,\epsilon'=\pm 1), \tag{93}$$

and

$$S_+(\epsilon,\epsilon')=\frac{\pi}{T}\frac{\left((V_2-V_3)\cot\left(\frac{\pi}{T}V_1\right)+\text{cyclic}\right)}{(V_1-V_2)(V_2-V_3)(V_3-V_1)},$$
$$S_-(\epsilon,\epsilon')=\frac{\pi}{T}\frac{\left((V_2-V_3)\operatorname{cosec}\left(\frac{\pi}{T}V_1\right)+\text{cyclic}\right)}{(V_1-V_2)(V_2-V_3)(V_3-V_1)}. \tag{94}$$

[The indices (ϵ,ϵ') are implicit in V_3 on the right-hand sides.] Then one obtains[3]

$$\Delta_{\pm}(x,y)=\frac{i(\Sigma\Sigma')^{-1/2}}{8\pi^2}\begin{pmatrix}(G-G')S_{\pm}(1,-1)(1-\sigma_r)(1+\sigma_{r'})\\ -(\bar{G}-\bar{G}')S_{\pm}(-1,1)(1+\sigma_r)(1-\sigma_{r'})\\ +(G+\bar{G}')S_{\pm}(1,1)(1-\sigma_r)(1-\sigma_{r'})\\ -(\bar{G}+G')S_{\pm}(-1,-1)(1+\sigma_r)(1+\sigma_{r'})\end{pmatrix}. \tag{95}$$

The propagator for spinors is obtained now through a standard prescription[15] as

$$(\gamma\cdot D(x)\Delta(x,y)(1+\gamma_5)/2+\Delta(x,y)\gamma\cdot\overleftarrow{D}(y)(1-\gamma_5)/2),$$

where Δ can be $\tilde{\Delta},\Delta_+$, or Δ_-.

All that has been assumed is that g (at x,y or at any other point) satisfies the properties listed under (7). Thus g can involve an arbitrary number of iterations. When x and y are both close to the origin, G and G' will exhibit strongly and simultaneously the consequences of the chaotic aspects studied in the Appendix. (See the remarks in Sec. VII). For a quasiperiodic background one can consider the aperiodic $\tilde{\Delta}$ or $\Delta_{\pm}$ for some adequate rational approximation.

VI. GENERALIZATIONS

Only brief indications will be given below concerning some possible generalizations.

A. Beyond spherical symmetry for $q=0,1$

One can generalize (66) and (77) as follows. Let

$$A_\mu=\sigma_{\mu\nu}\partial_\nu\ln\Sigma,$$

where, with $\epsilon=\pm1$,

$$\Sigma=\frac{1}{2}(1+\epsilon)+\sum_{m=1}^{M}\frac{1}{2r_m}(G_m+\bar{G}_m) \tag{96}$$

and $G_m(r_m+it)$ is a holomorphic function with

$$r_m=|\vec{x}-\vec{x}_m|.$$

Choosing

$$G_m=\frac{1+g_m}{1-g_m}, \tag{97}$$

where g_m is now given by (40) with the origin translated to $\vec{x}_m$, iterations can be introduced independently for each center. One can have a ''gas'' (dilute or dense) of (quasi)periodic instantons. In Ref. 7 spinors were studied in such a background without iterations. One can also start by generalizing (76) including shifts of origin, starting for the mth term with

$$g_m^{(0)}=\frac{\Sigma_{l=1}^{n_m}\lambda_{l,m}^2(e^{k_{l,m}(r_m+it-ic_{l,m})}-1)^{-1}}{\Sigma_{l=1}^{n_m}\lambda_{l,m}^2(1-e^{-k_{l,m}(r_m+it-ic_{l,m})})^{-1}}. \tag{98}$$

Then one may apply iterations independently for each m. As explained in Sec. IV a subset of spinor solutions can readily be obtained.

B. $q>1$; spherical symmetry necessarily broken

The ansatz (77) [or (96)] is not suitable for generalizing beyond $q>1$. This is one of our reasons for starting with (7). The general formulation of linear pairs[16] can thus be applied specifically[1,4] to (quasi)periodic fields for constructing higher q solutions. Details can be found in those papers. Here we just indicate where iterations can be implemented.

For $q=2$, one starts with *two* functions

$$g_1=\Pi_{j=1}^{n}\left(\frac{a_j+e^{-k_j(R+it)}}{\overline{a}_j+e^{k_j(R+it)}}\right),$$
$$g_2=\Pi_{j=1}^{n}\left(\frac{b_j+e^{-k_j(\overline{R}+it)}}{\overline{b}_j+e^{k_j(\overline{R}+it)}}\right), \tag{99}$$

where, with real c,

$$R=(r^2-c^2-i2cr\cos\theta)^{1/2},$$
$$\overline{R}=(r^2-c^2+i2cr\cos\theta)^{1/2},$$

implying an *imaginary* translation (ic) parallel to the z axis. For the solutions to be regular the parameters (c,a_j,b_j) have to satisfy constraints.[1,4] More generally one starts with q functions g for charge q. Iterations can be implemented for these functions. But one must then verify carefully the regularity constraints afterwards. That, presumably, would be difficult.

C. Gauge group SU(N)

Self-dual, (quasi)periodic solutions for arbitrary N were presented in Ref. 5. Again we only indicate, in the simplest case for $N>2$, namely, SU(3), how the g functions (which can be iterated) appear in the class of solutions obtained. The details of the generalized ansatz[5] will not be reproduced here.

For SU(3) we just note that, instead of one function e^{ζ}, as in (5), (6), and (7), one needs two,

$$e^{-b_1}=d_1P^{-1}(r^2(\partial_r^2+\partial_t^2)(g\overline{g}))^{-1}((p_3-p_2)(g\overline{g})^{p_1}+\text{cyclic}),$$
$$e^{-b_2}=d_2P^{-1}(r^2(\partial_r^2+\partial_t^2)(g\overline{g}))^{-1}((p_3-p_2)(g\overline{g})^{2-p_1}+\text{cyclic}), \tag{100}$$

where d_1,d_2 are constants (given in Ref. 5),

$$P=(p_2-p_1)(p_3-p_1)(p_3-p_2)$$

and (p_1,p_2,p_3) are rational numbers satisfying

$$p_1<p_2<p_3,\quad p_1+p_2+p_3=3.$$

To avoid certain problems concerning branch points (explained in Ref. 5) we set

$$g=\left(\prod_{j=1}^{n}\left(\frac{a_j+e^{-k_j(r+it)}}{\overline{a}_j+e^{k_j(r+it)}}\right)\right)^{Q}$$

such that Qp_i $(i=1,2,3)$ are integers.

Iterations can be introduced separately for the factors of g or for g as a whole. One can also use g of (76) as a starting point.

For higher values of N a set of $N-1$ independent parameters p enters into the solutions.[5] For $N>2$ a single magnetic winding number is no longer sufficient for characterizing the asymptotic configurations in R^3. This is one reason for an increasing number of parameters.

D. Use of hyperbolic coordinates

The uses of the coordinate transformation

$$(r+it)=\tanh \tfrac{1}{2}(\rho+i\tau)$$

in the construction of instantons or the so-called hyperbolic monopoles were studied in a series of papers. (Apart from Ref. 17 they are all summarized in Ref. 18—a review article.) The metric is

$$\begin{aligned} ds^2 &= dt^2+dr^2+r^2(d\theta^2+(\sin\theta)^2 d\phi^2) \\ &= (\cosh\rho+\cos\tau)^{-2}(d\tau^2+d\rho^2+(\sinh\rho)^2(d\theta^2+(\sin\theta)^2 d\phi^2)). \end{aligned} \tag{101}$$

In Refs. 17 and 18 τ-static solutions [depending on (r,t) through ρ] were considered. In Refs. 2 and 4 we indicated the passage from the t-periodic to the τ-periodic solutions. One can replace in (6) the subscripts (r,t) by (ρ,τ), respectively, and set

$$e^{\zeta}=\frac{\sinh\rho}{(1-g\bar{g})}((\partial_\rho^2+\partial_\tau^2)(g\bar{g}))^{1/2} \tag{102}$$

and, for example,

$$g=\prod_{j=1}^{n}\left(\frac{a_j+e^{-k_j(\rho+i\tau)}}{\bar{a}_j+e^{k_j(\rho+i\tau)}}\right). \tag{103}$$

The action is evaluated in Ref. 2. To avoid irregularities at

$$\rho=0, \quad \tau=\pm\pi$$

the k's in (103) have to be integers. [The role of the conformal factor in (101) is crucial concerning this point.[18]] This is a restriction. But the formalism is more general in the following sense. A simple rescaling

$$\rho=\lambda r', \quad \tau=\lambda t', \quad A_\tau=\lambda^{-1}A_{\tau'}, \quad A_\rho=\lambda^{-1}A_{\tau'}$$

gives back the previous formalism, in the limit $\lambda\to 0$, with

$$ds'^2=4\lambda^{-2}\,ds^2=dt'^2+dr'^2+r'^2(d\theta^2+(\sin\theta)^2 d\phi^2).$$

The restriction on the k's can be lifted after the scaling limit is taken.

But let us consider the situation without any such rescaling. As it stands, g in (103) has already a quite special type of (r,t) dependence via (ρ,τ). After several iterations one can have a very complex (r,t) dependence, say for the action density. But the situation can still be studied relatively simply using (ρ,τ).

The coordinate transformation introduced maps the (r,t) half-plane on the strip

$$0\leqslant\rho<\infty, \quad -\pi\leqslant\tau\leqslant\pi.$$

More generally one can consider our solutions in the context of three noncompact and one compact dimension, all with the same signature. If T is the period associated with the last one then the condition concerning integer values, say n_j of k_j, is to be generalized to $k_j=(2\pi n_j)/T$ assuring single-valued solutions. One can also consider the possibility of embedding our solutions into such subspaces of higher dimensional spaces.

VII. REMARKS

Chaos in gauge theories is a popular topic. Many authors have studied various aspects of this domain. A convenient reference is,[19] a book devoted to this field with a long bibliography. Among more recent papers one may note Ref. 20, again citing many sources. Compared to most of the above-mentioned studies, ours is more modest in one respect but more ambitious in another.

We have shown (the Appendix) that our iterative map, though simple, is a chaotic one. But its implementation up to any given order does not automatically render our field configurations fully chaotic. The precise way in which the implemented iterations influence the time evolution of the gauge field has been pointed out in Sec. I. When the configuration is strictly periodic it comes back, by definition, to its initial state after each period T. But the sensitive time dependence discussed in Sec. I implies that any quantity (such as the action density at a given point or within a small volume) can fluctuate, within a single period, more and more with higher number of iterations and in a more complex fashion. It can wander far and waywardly before coming back. When the configuration is quasiperiodic it does not return repeatedly to an initial state (Sec. III), though arbitrarily close approaches are possible. How should one precisely characterize such fields with increasingly sensitive time dependence generated by iterations? We have not adequately explored the implications, the consequences of the two most striking features of our solutions, namely sensitive time dependence and jumps in rotation numbers. We have exhibited their existence. A more thorough exploration can probably indicate a satisfactory characterization. At this stage, after a ''large'' number of iterations, ''at the edge of chaos'' might be a convenient description (though precaution is necessary due to broad, fashionable uses of this terminology). Sufficiently accurate numerical studies can help in understanding. But that is beyond the competence of the present author.

Having noted the limitations, one may now note the positive qualities of our approach. In order to be able to apply techniques developed for one-dimensional dynamical systems authors frequently consider gauge fields depending, effectively, on one coordinate only. Thus, for example, one studies time evolutions of fields homogeneous in space (Refs. 21, 20, and a number of papers cited in Ref. 19). In Ref. 22 the static problem is reduced to the one-dimensional Duffing equation and then time dependence is introduced as a perturbation. It is known that, considered in full generality, Yang–Mills fields are nonintegrable (Ref. 19 and sources cited therein). The fully chaotic aspects are then related to this nonintegrability.

Our approach, one may say, is antithetic to the preceding one. The intriguing features we exhibit via the mapping arise in fully integrable (explicitly solved) self-dual configurations. We start with solutions, found in our previous papers, which have a whole range of remarkable properties, quite apart from those revealed in the present study. They combine topological aspects of standard instantons and monopoles and are characterized by two topological numbers (both integers for periodic solutions). Instead of considering solitonic and chaotic aspects to be entirely incompatible (and leaving it at that) we have tried to explore, using our mapping, how close and with what possible limitations, they can be brought together. This has revealed probably hitherto unsuspected possibilities.

Quark confinement has been studied using approaches as different as that of a dual superconductor and that of random fields (see Chap. 11 of Ref. 19). Our spinors (except for one subclass) provide explicit solutions damped exponentially away from the origin. This is not confinement but is worth noting. It should also be noted that this damping *increases* with the temperature (with some typical frequency K, inverse of the period), while confinement is usually supposed to break down at a sufficiently high temperature. Spinors in backgrounds with several incommensurable frequencies should be further explored for a better understanding. Rather than being fully, dully chaotic they might provide a terrain favorable to genesis of rewardingly complex patterns and structures.

Periodic instantons have been considered from the beginning[23–25] to be of particular interest in the study of gauge fields at finite temperatures. Our more general solutions (as noted in Sec. I) show very clearly that strict periodicity involves extremely fine tuning of a set of parameters (the k's). Irrationals being dense on the real line, infinitesimal shifts in one or more k's can make a periodic solution quasiperiodic and vice versa. Since such solutions exist it would be quite artificial to consider periodic solutions exclusively. Numerically, as is well known, it is a delicate task to distinguish between commensurable and incommensurable cases. But since their mathematical properties are strikingly different, can one understand better the significance of the role of quasiperiodicity in the context of finite temperature? In a realistic situation the temperature cannot be absolutely steady. A suitably chosen interval covering small fluctuations of a roughly steady temperature would cover a continuum of frequencies. Slowly varying temperatures would again need different considerations. Though such aspects might be worth considering, we have no simple adequate answer concerning the role of quasiperiodicity in the context of finite temperatures. We have studied quasiperiodicity for the light they shed on topological aspects and the possibilities formally engendered by the presence of several incommensurable periods in the background.

We have obtained the propagators for (quasi)periodic backgrounds in particularly convenient forms, where one can implement an arbitrary number of our iterations. One can next try to compute the fluctuation determinants. Then one can start to carry over the consequences of the chaotic features of our mapping into quantum domains through semiclassical developments. It would be interesting to see in what fashion and to what extent such features can thus seep through.

Our spinor solutions are limited to zero mass and to the gauge group SU(2). But they, along with the propagator for spinor fields (Sec. V), can provide a starting point, exploiting the full range of our self-dual solutions, for the study of quarks in a finite temperature gluon background.

In the context of ADHMN[26–29] and Nahm's formalism[30] for calorons (periodic instantons) recently solutions have been obtained with nontrivial holonomy and Polyakov loop.[31–34] The explicit solution has zero total magnetic charge. A limiting static form with magnetic charge has also been obtained.[35] We have exhibited the role of nonzero magnetic charge in periodic instantons concerning index theorems and linear modes with implied consequences of the ''magnetic defect'' for the moduli space.

ACKNOWLEDGMENT

I thank Pierre Collet for discussions concerning iterative maps.

APPENDIX: ITERATIVE MAP

An iterative map of the unit disk is presented and some of its properties are studied. In Sec. I we indicate how this mapping is implemented in the construction of (quasi)periodic gauge fields and with what consequences. See also the remarks in Sec. VII.

1. The map

Let Z_p be a point in the unit disk, centered at the origin, in the complex plane. Consider the map, with $0<|a_p|<1$,

$$Z_{p+1}=\frac{a_p+Z_p}{\bar{a}_p+Z_p^{-1}}=Z_p\frac{a_p+Z_p}{1+\bar{a}_pZ_p}. \tag{A1}$$

For $0\leqslant|Z_p|\leqslant 1$ one obtains $0\leqslant|Z_{p+1}|\leqslant 1$.

The inverse power in the denominator (Z_p^{-1}) leads to properties quit different from those of the standard Möbius-type maps. The most evident difference is that here zero is a fixed point. But there are other profound differences. To take just one example, the Schwarzian derivative (identically zero for Möbius maps) makes an interesting contribution in the action density near the spatial origin (Sec. II). One may set

$$a_p=a_{p-1}=\cdots=a_0=a.$$

A more general possibility is

$$a_p = f(a_{p-1}) \tag{A2}$$

with f so chosen as to guarantee $0<|a_p|<1$ for $0<|a_0|<1$. An interesting example of f will be given below. But the explicit form of f will usually be unspecified, leaving room for eventual different convenient choices.

The crucial role of a suitable (r,t) parametrization is emphasized in Sec. I. For the choice (2), ψ_0 being the phase of Z_0,

$$Z_0 = |Z_0| e^{-i\psi_0} = e^{-k(r+it)} \quad (k>0, r\geqslant 0, -\infty<t<\infty). \tag{A3}$$

For the gauge fields we are finally interested in the space–time dependence. *But in the Appendix (r,t) should be considered as convenient notations defining*

$$\ln|Z_0| = -kr, \quad \psi_0 = kt. \tag{A4}$$

If (3) is chosen for Z_0, some different notation, say $(\widehat{kr}, \widehat{kt})$, would be more appropriate. We will, however, continue to use in the Appendix (r,t) as in (A4), hoping that no confusion can arise. Moreover some of our comments will refer specifically to (2) or (A3).

Define

$$Z_p = |Z_p| e^{-i\psi_p}, \quad a_p = |a_p| e^{-i\alpha_p}, \quad \chi_p = \psi_p - \alpha_p .$$

Then from (A1),

$$|Z_{p+1}| e^{-i(\psi_{p+1} - 2\alpha_p)} = |Z_p| e^{-i\chi_p} \frac{|a_p| + |Z_p| e^{-i\chi_p}}{1 + |a_p||Z_p| e^{-i\chi_p}}. \tag{A5}$$

2. Moduli

One has

$$|Z_{p+1}|^2 = |Z_p|^2 \frac{|a_p|^2 + |Z_p|^2 + 2|a_p||Z_p|\cos\chi_p}{1 + |a_p|^2|Z_p|^2 + 2|a_p||Z_p|\cos\chi_p} \equiv |Z_p|^2 |\tilde{Z}_p|^2, \tag{A6}$$

where

$$(1 - |\tilde{Z}_p|^2) = \frac{(1-|a_p|^2)(1-|Z_p|^2)}{1 + |a_p|^2|Z_p|^2 + 2|a_p||Z_p|\cos\chi_p} \geqslant 0. \tag{A7}$$

Thus for

$$|Z_p| = 0, 1, <1, \quad |Z_{p+1}| = 0, 1, <|Z_p|,$$

respectively.

Hence not only is zero a fixed point, but the circumference of the disk ($|Z_0|=1$ or $r=0$) is stable as a whole leading to circle maps to be studied in detail soon. For the inverse map, both the roots of the quadratic

$$Z_P^2 + (a_p - \bar{a}_p Z_{p+1}) Z_p = Z_{p+1} \tag{A8}$$

must correspond to $0<|Z_p|<1$ for $0<|Z_{p+1}|<1$. (For $|Z_p|\geqslant 1, |\tilde{Z}_p|\geqslant 1$, and hence $|Z_{p+1}|\geqslant 1$.) These roots coincide for

$$Z_{p+1} = -\left(\frac{a_p}{\bar{a}_p}\right)\left(\frac{1-\sqrt{1-|a_p|^2}}{1+\sqrt{1-|a_p|^2}}\right) \tag{A9}$$

to

$$Z_p=-\left(\frac{a_p}{1+\sqrt{1-|a_p|^2}}\right). \tag{A10}$$

Thus starting with $|Z_0|<1$ (or $r>0$), $|Z|$ moves away under iterations toward the attractive fixed point ($Z=0$). Generically two values of Z_p (which may coincide) are mapped on a Z_{p+1} with a lesser modulus. But instead of a stepwise migration toward the fixed point zero, the latter might be reached abruptly due to the following feature.

3. Zeros

One has $Z_{p+1}=0$ for

$$Z_p=0$$

and for

$$Z_p=-a_p.$$

The second possibility is worth further study. As will be seen later, the more general condition $|Z_p|=|a_p|$ provides crucial domains of discontinuities of rotation numbers associated to the phases. Moreover

$$Z_p=Z_{p-1}\frac{a_{p-1}+Z_{p-1}}{1+\bar{a}_{p-1}Z_{p-1}}=-a_p$$

furnishes an interesting context for exploring the consequences of different choices for f in $a_p=f(a_{p-1})$. One has

$$\left(\frac{Z_{p-1}}{\sqrt{a_p}}\right)^2+2\lambda_p\left(\frac{Z_{p-1}}{\sqrt{a_p}}\right)+1=0, \tag{A11}$$

where

$$\lambda_p=(2\sqrt{a_p})^{-1}(a_{p-1}+a_p\bar{a}_{p-1}).$$

Hence

$$Z_{p-1}=\sqrt{a_p}(-\lambda_p\pm\sqrt{\lambda_p^2-1}).$$

The two roots coincide to $Z_{p-1}=\mp\sqrt{a_p}$ for $\lambda_p=\pm 1$ or

$$\sqrt{a_p}=\pm(\bar{a}_{p-1})^{-1}(1-\sqrt{1-|a_{p-1}|^2}). \tag{A12}$$

Denoting, with real μ_p,

$$a_p=(\tanh\mu_p)e^{-i\alpha_p}, \tag{A13}$$

(A12) gives

$$\tanh\mu_p=(\tanh\tfrac{1}{2}\mu_{p-1})^2,\quad \alpha_p=2\alpha_{p-1}.$$

This is an example of the choice of f assuring special properties (here double zeros). For comparison note that choosing, for all p,

$$a_p=\overline{a}_p=a \quad (0<a<1)$$
$$\lambda_p=\tfrac{1}{2}\sqrt{a}(1+a)=\cos\zeta \tag{A14}$$

say, ζ being real. Now (A11) gives

$$Z_{p-1}=-\sqrt{a}e^{\mp i\zeta}. \tag{A15}$$

4. Circle map on the circumference (chaotic aspects)

For $|Z_0|=1$, (i.e., $r=0$), for all p,

$$|Z_p|=1.$$

For the phases ψ_p, using the notations of (A5), one has the circle map

$$e^{-i\psi_{p+1}}=\frac{a_p+e^{-i\psi_p}}{\overline{a}_p+e^{i\psi_p}}=e^{-i2\alpha_p}\frac{|a_p|+e^{-i\chi_p}}{|a_p|+e^{i\chi_p}}. \tag{A16}$$

Hence

$$\frac{d\psi_{p+1}}{d\psi_p}=2\left(\frac{1+|a_p|\cos\chi_p}{1+|a_p|^2+2|a_p|\cos\chi_p}\right)=1+\frac{(1-|a_p|^2)}{(1-|a_p|)^2+4|a_p|(\cos(\chi_p/2))^2}>1. \tag{A17}$$

Thus ψ_{p+1} is *monotonic* (increasing) in ψ_p and

$$e^{-i(\psi_{p+1}-2\alpha_p)}=1 \quad \text{for } \chi_p=0,\pi,2\pi$$
$$=-1 \quad \text{for } \cos\chi_p=-|a_p|, \quad \sin\chi_p=\pm\sqrt{1-|a_p|^2}. \tag{A18}$$

Hence as

$$\psi_p\to\psi_p+2\pi, \quad \psi_{p+1}\to\psi_{p+1}+4\pi. \tag{A19}$$

[This result also follows from an approach analogous to the one leading to (A52) for $|Z_0|>|a_0|$, since here $|Z_p|=1>|a_p|$. But the foregoing instructive one follows the rotations in more detail.] The result (A19) is fundamental in computing the actions of the gauge fields after iterations (Sec. II). But here we concentrate on another aspect.

It is well-known (Ref. 9, p. 50 and also p. 18) that the apparently very simple circle map

$$\theta_{p+1}=2\theta_p, \quad \frac{d\theta_{p+1}}{d\theta_p}=2 \tag{A20}$$

is *chaotic*. It satisfies all the requisite conditions, the most important one being a strong sensitivity to initial conditions (here in the form of expansiveness with index $\ln 2$). Our example (A16) will be seen to satisfy the same conditions, but in a more subtle fashion. In fact, our case contains (A20) as a particularly simple limit ($a_p=0$). ''On average'' ψ_{p+1} turns twice as fast as ψ_p. But the rate is less than twice in one domain and just sufficiently more than twice in the complementary one to compensate.

For

$$\cos\chi_p=\pm 1, \quad v_p\equiv\frac{d\psi_{p+1}}{d\psi_p}=\frac{2}{1\pm|a_p|} \tag{A21}$$

and for

$$\cos\chi_p = -|a_p|, \quad v_p = 2.$$

The complementary domains are

$$1 \geqslant \cos\chi_p > -|a_p|, \quad (v_p < 2),$$

$$-1 \leqslant \cos\chi_p < -|a_p| \quad (v_p > 2).$$

As $|a_p| \to 1$ the first domain increases, but so does v_p in the other to compensate (becoming very high near $\cos\chi_p = -1$).

After p iterations (with $\psi_0 = kt$)

$$\frac{d\psi_p}{d\psi_0} = \frac{1}{k}\frac{d\psi_p}{dt} = \prod_{j=0}^{p-1} v_j = \prod_{j=0}^{p-1} \left(\frac{2(1+|a_j|\cos\chi_j)}{1+|a_j|^2+2|a_j|\cos\chi_j} \right) \tag{A22}$$

and

$$\frac{d^2\psi_p}{d\psi_0^2} = \frac{1}{k^2}\frac{d^2\psi_p}{dt^2} = \frac{d\psi_p}{d\psi_0}\left(\sum_{j=0}^{p-1} V_j\left(\frac{d\psi_j}{d\psi_0} \right) \right), \tag{A23}$$

where

$$V_j = \frac{(1-|a_j|^2)|a_j|\sin\chi_j}{(1+|a_j|\cos\chi_j)(1+|a_j|^2+2|a_j|\cos\chi_j)}.$$

In (A22) each factor $v_j > 1$. But due to the factor $\sin\chi_j$ in V_j the terms in (A23) can change sign.

The sensitive dependence on initial data should be evident from the preceding analysis. But let us formulate it more precisely in terms of a *characteristic index*. For (A20) the index is evidently $\ln 2$.[9] This is recovered in our case in the limit of each $a_j = 0$. For (A16) it depends on the sequence $(a_0, a_1, \ldots)$. *But it has a positive lower bound.* In (A22) replacing each v_j by its upper and lower bound, respectively,

$$\prod_{j=0}^{N-1} \left(\frac{2}{1-|a_j|} \right) \geqslant \frac{d\psi_N}{d\psi_0} \geqslant \prod_{j=0}^{N-1} \left(\frac{2}{1+|a_j|} \right) > 1. \tag{A24}$$

Setting for simplicity all $|a_j| = |a|$, the characteristic index λ satisfies

$$\ln\left(\frac{2}{1-|a|} \right) \geqslant \lambda \geqslant \ln\left(\frac{2}{1+|a|} \right) > 0. \tag{A25}$$

More generally $(2/1 \pm |a|)$ should be considered as the geometric means of the corresponding products in (A24).

For a map to be chaotic it must have a dense set of periodic points.[9] For (A20) the periodic points are given by[9]

$$\theta_n = 2^n\theta = \theta + 2k\pi \quad \text{or} \quad \theta = \frac{2k\pi}{2^n-1}, \tag{A26}$$

where k is an integer and

$$0 \leqslant k \leqslant 2^n - 1.$$

This is the situation for all a's zero in (A16). But for (A16) one cannot obtain as simply a general formula. One can, however, proceed stepwise to show how the periodic points remain dense but are shifted as the parameters a increase from zero. For simplicity consider all a_p's equal and real. Then

$$e^{-i\psi_{p+1}}=\frac{a+e^{-i\psi_p}}{a+e^{i\psi_p}}$$

gives

$$\psi_{p+1}=2\psi_p-2(a\sin\psi_p-\tfrac{1}{2}a^2\sin 2\psi_p+\cdots). \tag{A27}$$

Thus

$$\psi_n=2^n\psi_0-a\left(\sum_{l=1}^{n}2^l\sin 2^{n-l}\psi_0\right)+O(a^2). \tag{A28}$$

Hence up to $O(a)$ the periodic points are given by

$$\psi=\frac{2k\pi}{2^n-1}+aS_1, \tag{A29}$$

where

$$S_1=\left(\frac{2^n}{2^n-1}\right)\left(\sum_{l=1}^{n}2^{-n+l}\sin 2^{n-l}\left(\frac{2k\pi}{2^n-1}\right)\right)<\left(\frac{2^n}{2^n-1}\right)\left(\sum_{l=1}^{n}2^{-n+l}\right)=2.$$

Thus the dense set (A26) is shifted as shown above. One may now iterate to higher powers of a and find an analogous situation. We cannot produce a general solution for (A16) in a closed form, but the smooth continuity with (A26) is clear enough.

A third criterion for chaoticity[9] is topological transitivity. This is satisfied by (A16) as obviously as by (A20). Our preceding analysis of rotations makes it evident that their effects cannot remain confined in one particular segment of a circle.

Thus all three criteria for being chaotic are satisfied by our map.

5. Series expansion near the circumference (small *r*)

For the gauge field configurations principally considered in this paper the time dependence is exponentially damped as r increases. So the (quasi)periodic time evolution is best studied for small r. For our mapping this corresponds to a domain near the circumference at a distance $(1-e^{-kr})$.

Let ψ_p *continue* to denote the value of the phase for $r=0$ and let

$$Z_p=e^{-i\psi_p}(1+C_p^{(1)}r+C_p^{(2)}r^2+\cdots), \tag{A30}$$

where the C's can be *complex* since the r dependence of the total phase is included in them. The expansion (A30) is general, but particularly useful for small r due to evident reasons. (In this subsection only, ψ_p denotes not the total phase but a part. The notation $\psi_p^{(0)}$ would have been more consistent. But this simplification, leaving room for other indices to come, should not cause confusion.)

Supressing the index p temporarily and using the holomorphy condition

$$(\partial_r+i\partial_t)Z=0 \tag{A31}$$

one obtains

$$C^{(1)}+2C^{(2)}r+3C^{(3)}r^2+\cdots=-\frac{d\psi}{dt}(1+C^{(1)}r+C^{(2)}r^2+\cdots)-i\left(\frac{dC^{(1)}}{dt}r+\frac{dC^{(2)}}{dt}r^2+\cdots\right). \tag{A32}$$

Thus

$$C^{(1)}=-\frac{d\psi}{dt}$$

and, for $l>1$,

$$lC^{(l)}=C^{(1)}C^{(l-1)}-i\frac{dC^{(l-1)}}{dt}. \tag{A33}$$

The general solution is

$$C^{(l)}=\frac{e^{i\psi}}{l!}\left(-i\frac{d}{dt}\right)^l(e^{-i\psi}). \tag{A34}$$

This expansion displays precisely how the chaotic properties of the circle map for the phase are carried over through ψ and its derivatives in the coefficients.

The result (A34) is compact and elegant. But separate explicit expressions for the total phase and the amplitude of Z are useful for the gauge fields. We give the first few terms of the r expansion for both. The coefficients $(\psi^{(l)},B_l)$ will now all be real. One obtains, keeping terms up to $O(r^4)$,

$$Z=e^{-i\psi}(1+C_1r+C_2r^2+C_3r^3+C_4r^4)=e^{-i(\psi^{(0)}+\psi^{(2)}r^2+\psi^{(4)}r^4)}(1+B_1r+B_2r^2+B_3r^3+B_4r^4). \tag{A35}$$

The coefficients $\psi^{(1)}$ and $\psi^{(3)}$ turn out to be zero and one obtains, in terms of the C's given before,

$$\begin{aligned}\psi^{(2)}&=\frac{i}{2}(C^{(2)}-\overline{C}^{(2)}),\\ \psi^{(4)}&=\frac{i}{2}\left((C^{(4)}-\overline{C}^{(4)})-\frac{1}{2}((C^{(2)})^2-(\overline{C}^{(2)})^2)\right),\end{aligned} \tag{A36}$$

and

$$\begin{aligned}B_1&=C^{(1)},\\ B_2&=\tfrac{1}{2}(C^{(2)}+\overline{C}^{(2)}),\\ B_3&=\tfrac{1}{2}(C^{(3)}+\overline{C}^{(3)}),\\ B_4&=\tfrac{1}{2}(C^{(4)}+\overline{C}^{(4)})-\tfrac{1}{8}((C^{(2)})^2+(\overline{C}^{(2)})^2).\end{aligned} \tag{A37}$$

One may also note that

$$Z\overline{Z}=1+D_1r+D_2r^2+D_3r^3+D_4r^4+O(r^5), \tag{A38}$$

where

$$D_1=2C^{(1)},$$

$$D_2=2(C^{(1)})^2,$$
$$D_3=\frac{4}{3}(C^{(1)})^3-\frac{1}{3}\frac{d^2}{dt^2}C^{(1)}, \tag{A39}$$
$$D_4=\frac{2}{3}\left((C^{(1)})^4-C^{(1)}\frac{d^2}{dt^2}C^{(1)}\right).$$

Higher order terms can be evaluated stepwise. These results hold, of course, for any p. If the product of several factors $(Z_p Z'_{p'}\cdots)$ is considered, one has the same expansion with

$$C^{(1)}=-\frac{d}{dt}(\psi_p+\psi'_{p'}+\cdots). \tag{A40}$$

Consistency with this constraint is a useful check on the numerical coefficients obtained above. The notation indicates that for each factor the sequence of the parameters a, the periods involved, and also the order of iterations can be different. The results (A39) yield the leading term in the action density near the spatial origin (Sec. II).

6. Annular maps and rotation numbers

We have studied some interesting properties of our map on the circumference of the unit disk and nearby $(r=0,\ r\ll 1)$. One can continue an analogous study away from the edge. But we now concentrate on a different class of remarkable features associated to specific values of r as it increases.

For $|Z|=1$ the iterations affect only the phase giving a circle map. For $|Z|<1$ the amplitude also changes. It diminishes and becomes a function of the phases of the previous steps. The domain of variation of Z (as a function of these phases) becomes an annulus, which can, crucially, become a disk. To emphasize this aspect we use the term ''annular map.'' The rotation numbers to be defined will be associated to the phases. We are fundamentally interested in variations of functions of the phase of Z_0 as the latter moves on a circle of radius e^{-kr}. This provides, most directly through (2), the link with the time evolution of gauge fields.

From (A5) and (A6), with the notations defined there $(\chi_p=\psi_p-\alpha_p,\ldots)$

$$|Z_{p+1}|^2=|Z_p|^2\frac{|a_p|^2+|Z_p|^2+2|a_p||Z_p|\cos\chi_p}{1+|a_p|^2|Z_p|^2+2|a_p||Z_p|\cos\chi_p} \tag{A41}$$

and

$$\psi_{p+1}=\psi_p+\frac{i}{2}(\ln f_1+\ln f_2)+2\alpha_p, \tag{A42}$$

where

$$f_1=\left(\frac{|a_p|+|Z_p|e^{-i\chi_p}}{|a_p|+|Z_p|e^{i\chi_p}}\right),\quad f_2=\left(\frac{1+|a_p||Z_p|e^{i\chi_p}}{1+|a_p||Z_p|e^{-i\chi_p}}\right).$$

In particular, for $p=0$ with $Z_0=e^{-k(r+it)}$,

$$|Z_1|^2=e^{-2kr}\frac{|a_0|^2+e^{-2kr}+2|a_0|e^{-kr}\cos\chi_0}{1+|a_0|^2e^{-2kr}+2|a_0|e^{-kr}\cos\chi_0}. \tag{A43}$$

For fixed r, this can be shown to be a monotonic increasing function of $\cos\chi_0$. One obtains

$$|Z_1|_{\max}=|Z_0|\left(\frac{|a_0|+|Z_0|}{1+|a_0||Z_0|}\right), \tag{A44}$$

$$|Z_1|_{\min}=|Z_0|\left(\frac{|a_0|-|Z_0|}{1-|a_0||Z_0|}\right). \tag{A45}$$

The width of the annulus is

$$W_1=|Z_1|_{\max}-|Z_1|_{\min}.$$

For $|Z_0|>|a_0|$,

$$W_1=2|a_0||Z_0|\left(\frac{1-|Z_0|^2}{1-(|a_0||Z_0|)^2}\right). \tag{A46}$$

For $|Z_0|<|a_0|$,

$$W_1=2|Z_0|^2\left(\frac{1-|a_0|^2}{1-(|a_0||Z_0|)^2}\right). \tag{A47}$$

For $|Z_0|=|a_0|$, *the annulus becomes a disk* with

$$|Z_1|_{\min}=0$$

and

$$W_1=\frac{2|a_0|^2}{1+|a_0|^2}. \tag{A48}$$

As $|Z_1|$ touches zero the phase of Z_1 becomes undefined with crucial consequences. We will study them now.

Our aim is to compare the rates of rotation of ψ_0 and ψ_1 for the following domains of $|Z_0| = e^{-kr}$:

$$|Z_0|>|a_0|, \quad |Z_0|=|a_0|,$$

and

$$|Z_0|<|a_0|.$$

For $|Z_p|=1(>|a_p|)$ we have already seen [(A16)–(A21)] that on the average ψ_{p+1} turns twice as fast as ψ_p. In particular, ψ_1 turns twice as fast as ψ_0. This result will be seen to hold more generally for $|Z_0|>|a_0|$. But discontinuities appear as $|Z_0|$ comes down to $|a_0|$ and crosses over. We will demonstrate this now.

Let us examine (A42) for $p=0$ and in particular the term $\ln f_1$. For $|Z_0|>|a_0|$ the real part of $(|a_0|+|Z_0|e^{-i\chi_0})$ can vanish and change sign. *This feature is absent in* f_2 *since always* $|a_p||Z_p|<1$. So the term $\ln f_1$ has to be treated carefully to keep track of additive contributions as ψ_0 rotates (t increases). It is convenient to proceed as follows. For

$$|Z_0|>|a_0|,$$
$$f_1=e^{-i2\chi_0}\left(\frac{1+\dfrac{|a_0|}{|Z_0|}e^{i\chi_0}}{1+\dfrac{|a_0|}{|Z_0|}e^{-i\chi_0}}\right). \tag{A49}$$

For

$$|Z_0| = |a_0|, \quad f_1 = e^{-i\chi_0}. \tag{A50}$$

For

$$|Z_0| < |a_0|,$$
$$f_1 = \left(\frac{1 + \frac{|Z_0|}{|a_0|} e^{-i\chi_0}}{1 + \frac{|Z_0|}{|a_0|} e^{i\chi_0}} \right). \tag{A51}$$

Hence for the three domains, respectively,

$$\psi_1 = 2\psi_0 + \Lambda_{(+)}, \quad \psi_1 = \tfrac{3}{2}\psi_0 + \Lambda_{(0)}, \quad \psi_1 = \psi_0 + \Lambda_{(-)}, \tag{A52}$$

where one can now safely consider that

$$\Lambda_\delta(\psi_0 + 2\pi) = \Lambda_\delta(\psi_0) \quad (\delta = +,0,-). \tag{A53}$$

Hence for $\psi_0 \to \psi_0 + 2n\pi$ there are no cumulative, additive contributions from Λ_δ giving a supplementary term proportional to n (as does the first term proportional to ψ_0). The discontinuity involved for ψ_1 at $|Z_0| = |a_0|$ is now explicit.

In view of the crucial role of this result we present an alternative approach, closely following the argument for the circle map [(A18)–(A21)] but generalizing it for all $|Z_0| > |a_0|$. (This also generalizes the arguments for chaoticity on the circumference for the interior of the disk.)

It is sufficient to consider one single term (f_1) as follows. Define

$$e^{-i\beta} = \frac{|a_0| + |Z_0| e^{-i\chi_0}}{|a_0| + |Z_0| e^{i\chi_0}} \tag{A54}$$

when

$$\frac{d\beta}{d\psi_0} = \frac{2|Z_0|(|Z_0| + |a_0|\cos\chi_0)}{(|a_0|^2 + |Z_0|^2 + 2|a_0||Z_0|\cos\chi_0)}.$$

Hence for $|Z_0| > |a_0|$ one has strict monotonicity with

$$\frac{d\beta}{d\psi_0} > 0 \tag{A55}$$

but *not* for $|Z_0| < |a_0|$. For the exceptional value $|Z_0| = |a_0|$ one has simply

$$e^{-i\beta} = e^{-i\chi_0}. \tag{A56}$$

For $|Z_0| > |a_0|$ along with strict monotonicity one has

$$e^{-i\beta} = -1 \quad \text{for} \ \cos\chi_0 = -\frac{|a_0|}{|Z_0|}, \sin\chi_0 = \pm\sqrt{1 - \frac{|a_0|^2}{|Z_0|^2}}, \tag{A57}$$

$$e^{-i\beta}=1 \quad \text{for } \cos\chi_0=0,\pi,2\pi.$$

Hence, as for the circle map, ''on the average'' β turns twice as fast as ψ_0. This corresponds to the factor 2 in the first equation of (A52), namely,

$$\psi_1=2\psi_0+\Lambda_{(+)}.$$

Now suppose we consider the Poincaré sections for

$$t_n=t+nT \quad (n=0,1,2,...)$$

with some suitably chosen period *T*. The ''rotation number'' for $\psi_0(=kt)$ is defined in terms of

$$\psi_0^{(n)}=kt_n$$

as

$$\Omega_0=(n^{-1}(\psi_0^{(n)}-\psi_0))_{n\to\infty}=kT. \tag{A58}$$

(We reserve the term ''winding number'' for the magnetic charge q of the gauge fields.) Here, of course, it would have been natural to set

$$T=2\pi k^{-1}$$

or, rescaling, obtain an integer (say, unit) value for Ω_0. But we keep T unspecified here since for the gauge fields there will be the simultaneous presence of different periodic building blocks each with its own period, which may even be mutually incommensurable. (See Sec. III.) Now, similarly defining the rotation number Ω_1 for ψ_1, one obtains from (A52) and (A53)

$$\Omega_1=2\Omega_0 \quad \text{for } e^{-kr}>|a_0|, \quad \Omega_1=\tfrac{3}{2}\Omega_0 \quad \text{for } e^{-kr}=|a_0|, \quad \Omega_1=\Omega_0 \quad \text{for } e^{-kr}<|a_0|. \tag{A59}$$

Hence there is a step discontinuity at $r=-k^{-1}\ln|a_0|$. Let us further examine how this implies *sensitive dependence on small differences in the value of a parameter*. Let $\psi_1^{(\pm)}$ and $\hat{\psi}_1$ denote the values of ψ_1, respectively, for

$$|Z_0|=|a_0|(1\pm\epsilon), \quad |a_0|.$$

Here $\epsilon\ll 1$. From (A42) with $p=0$, up to $O(\epsilon)$, one obtains

$$\psi_1^{(\pm)}=\hat{\psi}_1\pm\epsilon\left(\frac{1}{2}\tan\frac{\chi_0}{2}-\frac{|a_0|^2\sin\chi_0}{1+|a_0|^4+2|a_0|^2\sin\chi_0}\right). \tag{A60}$$

But such a development is valid if not only $\epsilon\ll 1$ but *also* $\epsilon\tan(\chi_0/2)\ll 1$. Even if one starts with a suitably small value of the latter, the constraint will be violated as χ_0 approaches the values $(2N+1)\pi$. Near such values one may proceed as follows. The contribution to $-i2(\psi_1^{(+)}-\psi_1^{(-)})$ from the term $\ln f_1$ in (A42) is

$$\ln\left(\left(\frac{1+e^{i\chi_0}+\epsilon}{1+e^{-i\chi_0}+\epsilon}\right)\left(\frac{1+e^{-i\chi_0}-\epsilon}{1+e^{i\chi_0}-\epsilon}\right)\right)=\ln\left(\frac{\cos(\chi_0/2)-i(\epsilon/2)\sin(\chi_0/2)}{\cos(\chi_0/2)+i(\epsilon/2)\sin(\chi_0/2)}\right). \tag{A61}$$

For $\chi_0=(2N+1)\pi$ this is no longer proportional to ϵ but becomes $\ln(-1)$. The additive contribution at each turn gives a difference $(\psi_1^{(+)}-\psi_1^{(-)})$ consistent with (A59).

So far we have analyzed the effect of the passage $Z_0 \to Z_1$. For the next step, write

$$Z_2 = \left(\frac{a_1 + Z_1}{\overline{a_1} + Z_1^{-1}} \right) = Z_0^{-2} \left(\frac{a_0 + Z_0}{\overline{a_0} + Z_0^{-1}} \right) \left(\frac{\mu_+ + Z_0}{\overline{\mu_+} + Z_0^{-1}} \right) \left(\frac{\mu_- + Z_0}{\overline{\mu_-} + Z_0^{-1}} \right), \tag{A62}$$

where [compare (A11) and the discussion that follows]

$$\mu_\pm = \sqrt{a_1} (\lambda_0 \pm \sqrt{\lambda_0^2 - 1}) \tag{A63}$$

with

$$\lambda_0 = \frac{a_0 + \overline{a_0} a_1}{2 \sqrt{a_1}}.$$

One sees that for studying Ω_2 (the rotation number associated to the phase ψ_2 of Z_2) one should extend the previous considerations to *three* spherical shells

$$|Z_0| = |a_0|, \quad |\mu_+|, \quad |\mu_-|, \tag{A64}$$

where

$$|\mu_+||\mu_-| = |a_1|.$$

One may get special features corresponding to a possible double zero when $|\mu_+|$ and $|\mu_-|$ coincide to $\sqrt{|a_1|}$. The number of possibilities continue to increase with each iteration. Generically the process is systematic, but roots can coincide leading to special features. The cumulative effects of the jumps give more and more elaborate staircaselike patterns (Sec. II).

Considering $(|Z_p|, \psi_p)$ as independent variables one may carry out a similar analysis for the step $Z_p \to Z_{p+1}$ (for $p > 0$). One then obtains formally, like (A52) with the Λ's satisfying the criterion (A53),

$$\begin{aligned} \psi_{p+1} &= 2\psi_p + \Lambda_{(+)}^{(p)}, \\ \psi_{p+1} &= \tfrac{3}{2}\psi_p + \Lambda_{(0)}^{(p)}, \\ \psi_{p+1} &= \psi_p + \Lambda_{(-)}^{(p)}, \end{aligned} \tag{A65}$$

respectively, for

$$|Z_p| > |a_p|, \quad |Z_p| = |a_p|, \quad |Z_p| < |a_p|.$$

But it should be clearly noted that for $p \geqslant 1$ all $|Z_p|$'s are time dependent even for the choice (2) for Z_0. For the choice (3) even $|Z_0|$ is time dependent. The formal steps of the iterations and their consequences in the context of the mapping do not depend on the parametrization of Z_0 (provided it satisfies $Z_0 < 1$). But when the time evolution is studied in the context of the gauge fields much depends evidently on the initial choice.

[1] A. Chakrabarti, ''Periodic generalizations of static, self-dual SU(2) gauge fields,'' Phys. Rev. D **35**, 696 (1987).
[2] A. Chakrabarti, ''Spinors in periodic self-dual gauge field backgrounds,'' J. Math. Phys. **29**, 481 (1988).
[3] A. Chakrabarti, ''Propagators in periodic gauge field backgrounds,'' Phys. Lett. B **210**, 181 (1988).
[4] A. Chakrabarti, ''Quasiperiodic instantons,'' Phys. Rev. D **38**, 3219 (1988).
[5] A. Chakrabarti and C. Roiesnel, ''Periodic and quasiperiodic SU(N) instantons,'' Phys. Rev. D **39**, 2381 (1989).
[6] A. Chakrabarti, ''Antiperiodic spinor zero modes for periodic SU(2) instanton backgrounds,'' Phys. Lett. B **227**, 137 (1989).
[7] A. Chakrabarti, ''Linear fluctuations of periodic and quasiperiodic instantons,'' Phys. Rev. D **40**, 2684 (1989).

[8] A. Chakrabarti, ''Quasiperiodic gauge fields: Poincaré sections, annular maps and irrational winding numbers (3 pages, unpublished talk presented at WIGSYM, Oxford, 1993).
[9] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems* (Benjamin, New York, 1986).
[10] P. Bak, T. Bohr, and M. H. Jensen, *Circle Maps, Mode-Locking and Chaos*, Directions in Chaos Vol. 2, edited by H. Bai-lin (World Scientific, Singapore, 1988).
[11] T. Geisel, ''Quasiperiodicity versus mixing instability in a kicked quantum system,'' Phys. Rev. A **41**, 2989 (1990).
[12] J. P. Eckmann and D. Ruelle, ''Ergodic theory of chaos and strange attractors,'' Rev. Mod. Phys. **57**, 617 (1985).
[13] E. Witten, ''Some exact multipseudoparticle solutions of classical Yang–Mills theory,'' Phys. Rev. Lett. **38**, 121 (1977).
[14] R. Jackiw, C. Nohl, and C. Rebbi, ''Conformal properties of pseudoparticle configurations,'' Phys. Rev. D **15**, 1642 (1977).
[15] L. S. Brown and D. B. Creamer, ''Vacuum polarization about instantons,'' Phys. Rev. D **18**, 3695 (1978).
[16] C. Roiesnel and A. Chakrabarti, ''Geometry of linear pairs for self-dual gauge fields,'' J. Math. Phys. **31**, 1643 (1990).
[17] A. Chakrabarti, ''Construction of hyperbolic monopoles,'' J. Math. Phys. **27**, 340 (1986).
[18] A. Chakrabarti, ''Classical solutions of Yang–Mills fields (selected topics),'' Fortschr. Phys. **35**, 1 (1987).
[19] T. S. Biró, S. G. Matinyan, and B. Müller, *Chaos and Gauge Field Theory* (World Scientific, Singapore, 1994).
[20] C. Mukku, M. S. Sriram, J. Segar, B. A. Bambah, and S. Lakshmibala, ''Order-chaos transition in field theories with topological terms: A dynamical system approach, hep-th/9610071, Oct. 1996.
[21] S.-J. Chang, ''Classical Yang–Mills solutions and iterative maps,'' Phys. Rev. D **29**, 259 (1984).
[22] S. G. Matinyan, E. B. Prokhorenko, and G. K. Savvidy, ''Non-integrability of time dependent spherically symmetric Yang–Mills equations,'' Nucl. Phys. B **298**, 414 (1988).
[23] B. K. Harrington and H. K. Shepard, ''Periodic Euclidean solutions and the finite temperature Yang–Mills gas,'' Phys. Rev. D **17**, 2122 (1978).
[24] B. K. Harrington and H. K. Shepard, ''Thermodynamics of Yang–Mills gas,'' Phys. Rev. D **18**, 2990 (1978).
[25] D. J. Gross, R. D. Pisarski, and L. G. Yaffe, ''QCD and instantons at finite temperature,'' Rev. Mod. Phys. **53**, 43 (1981).
[26] M. F. Atiyah, N. J. Hitchin, V. G. Drinfeld, and Yu. I. Manin, ''Construction of instantons,'' Phys. Lett. **65A**, 185 (1978).
[27] M. F. Atiyah, *Geometry of Yang–Mills Fields, Fermi Lectures* (Scuola Normale Sup., Pisa, 1979).
[28] E. Corrigan and P. Goddard, ''Construction of instanton and monopole solutions and reciprocity,'' Ann. Phys. (N.Y.) **154**, 253 (1984).
[29] W. Nahm, *The Construction of all Self-Dual Multimonopoles by the ADHM Method*, Monopoles in Quantum Field Theory, edited by N. Craigie (World Scientific, Singapore, 1982), p. 87.
[30] W. Nahm, ''Self-dual monopoles and calorons,'' edited by G. Denardo [Lect. Notes Phys. **201**, 189 (1984)].
[31] K. Lee, ''Instantons and magnetic monopoles on $R^3 \times S^1$ with arbitrary gauge groups,'' hep-th/9802012.
[32] K. Lee and C. Lu, ''SU(2) calorons and magnetic monopoles,'' hep-th/9802108.
[33] T. C. Kraan and P. van Baal, ''Exact *T*-duality between calorons and Taub-NUT spaces,'' Phys. Lett. B **284**, 268 (1998); hep-th/9802049.
[34] T. C. Kraan and P. van Baal, ''Periodic instantons with nontrivial holonomy,'' hep-th/9805168.
[35] K. Lee and P. Yi, ''Dyons in $N=4$ supersymmetric theories and three-pronged strings,'' hep-th/9804174 v2.

The osp(1,2)-covariant Lagrangian quantization of irreducible massive gauge theories

B. Geyer[a)]
Universität Leipzig, Naturwissenschaftlich-Theoretisches Zentrum, Leipzig 04109, Germany

P. M. Lavrov
Tomsk State Pedagogical University, Tomsk 634041, Russia

D. Mülsch
Wissenschaftszentrum Leipzig e.V., Leipzig 04103, Germany

(Received 2 February 1998; accepted for publication 23 October 1998)

The osp(1,2)-covariant Lagrangian quantization of general gauge theories is formulated which applies also to massive fields. The formalism generalizes the Sp(2)-covariant Batalin–Lavrov–Tyŭtin (BLT) approach and guarantees symplectic invariance of the quantized action. The dependence of the generating functional of Green's functions on the choice of gauge in the massive case disappears in the limit $m \rightarrow 0$. Ward identities related to osp(1,2) symmetry are derived. Massive gauge theories with closed algebra are studied as an example. © *1999 American Institute of Physics.* [S0022-2488(99)01802-2]

I. INTRODUCTION

Recently, a general method for quantizing gauge theories in the Lagrangian formalism has been proposed[1–3] which is based on extended Becchi–Rouet–Stora–Tyutin (BRST) symmetry, i.e., simultaneous invariance under both BRST and antiBRST transformations. It is characterized by a quantum action functional $S = S(\phi^A, \phi^*_{Aa}, \bar{\phi}_A)$ depending on, besides the dynamical fields $\phi^A = (A^i, B^{\alpha_0}, C^{\alpha_0 a_0})$, also on related antifields (or external sources) $\phi^*_{Aa}, \bar{\phi}_A$, where A^i, B^{α_0} and $C^{\alpha_0 a_0}$ are the gauge, the auxiliary and the (anti)ghost fields, respectively, and both a and a_0 indicate members of Sp(2) doublets. To guarantee their (anti)BRST symmetry the action S (and the related gauge fixed extended action $S_{\rm ext}$) is required to satisfy two master equations which are generated by two nilpotent, anticommuting differential operators, $\bar{\Delta}^a$. The method applies to irreducible as well as reducible, complete gauge theories with either open or closed gauge algebra. (The condition of irreducibility requires the generators of the gauge transformations to be linearly independent at the stationary point of the classical action, and the condition of completeness requires the degeneracy of the Hessian of the classical action $S_{\rm cl}(A)$ to be solely due to its gauge invariance.[4,5])

Although this formalism is seemingly manifest Sp(2)-covariant among the solutions of the master equations, there are both Sp(2)-symmetric and Sp(2)-nonsymmetric ones. The symmetric solutions may be singled out by the explicit requirement of invariance under Sp(2) transformations by additional master equations whose generating differential operators $\bar{\Delta}_\alpha$ ($\alpha = 0, +, -$) are related to the generators of the symplectic group Sp(2). The algebra of these operators may be chosen to obey the orthosymplectic (super)algebra osp(1,2). (Actually, its even part is the algebra sl(2) generating the special linear transformations, but due to their isomorphism to the algebra sp(2) we will speak about symplectic transformations.) Moreover, if also massive fields should be considered to circumvent possible infrared singularities occurring in the process of subtracting ultraviolet divergences, without breaking the extended BRST symmetry, then this algebra appears

[a)]Electronic mail: geyer@rz.uni-leipzig.de necessarily. Let us also mention that the osp(1,2) superalgebra occurs in many problems where $N=1$ superconformal symmetry is involved; e.g., in the minimal $N=1$ superconformal models this symmetry appears in the light-cone approach to two-dimensional supergravity.[6] It is also of interest in this respect that topological osp(1,2)/osp(1,2) coset theories can be used to describe the noncritical Ramond–Neveu–Schwarz superstrings.[7]

The goal of the present paper will be to generalize the BLT quantization procedure to another one being osp(1,2)-covariant. For the sake of simplicity we restrict ourselves to the case of irreducible (or zero-stage) complete gauge theories and thereby we follow very closely the exposition of Ref. 1. (The extension to L-stage reducible theories will be dealt with in a succeeding paper.[19]) We also used the condensed notation introduced by DeWitt[8] and conventions adopted in Ref. 1; if not otherwise specified, derivatives with respect to the antifields are the (usual) left ones and that with respect to the fields are right ones. Left derivatives with respect to the fields are labeled by the subscript L, for example, $\delta_L/\delta\phi^A$ denotes the left derivative with respect to the fields ϕ^A.

The paper is organized as follows. In Sec. II we shortly review the standard Sp(2)-covariant approach and point out how it will be generalized to the osp(1,2)-covariant quantization procedure. As a consequence of the enlarged algebra a canonical definition of the ghost number (Faddeev–Popov charge) is obtained. Furthermore, to be able to express this algebra through operator identities and to get nontrivial solutions of the generating equations it is necessary to enlarge the set of antifields. In Sec. III the explicit construction of generating differential operators fulfilling the osp(1,2) algebra is outlined, starting with the approximation of the action S_m at lowest order in $\hbar$ which is assumed to be linear with respect to the antifields. In Sec. IV the gauge dependence of the generating functional of Green's functions is studied and corresponding Ward identities are derived. It is shown that the mass terms destroy gauge independence of the S-matrix. In Sec. V we consider massive theories with closed gauge algebra, thereby extending the solution given in Ref. 1.

II. GENERAL STRUCTURE OF osp(1,2)-COVARIANT QUANTIZATION OF IRREDUCIBLE GAUGE THEORIES

Let us consider a set of gauge (as well as matter) fields A^i with Grassmann parity $\epsilon(A^i)=\epsilon_i$ for which the classical action $S_{\mathrm{cl}}(A)$ is assumed to be invariant under the gauge transformations

$$\delta A^i = R^i_{\alpha_0}\xi^{\alpha_0}, \quad S_{\mathrm{cl},i}R^i_{\alpha_0}=0, \tag{1}$$

where ξ^{α_0} are the parameters of these transformations and $R^i_{\alpha_0}(A)$ are the gauge generators having Grassmann parity $\epsilon(\xi^{\alpha_0})=\epsilon_{\alpha_0}$ and $\epsilon(R^i_{\alpha_0})=\epsilon_i+\epsilon_{\alpha_0}$, respectively; by definition $X_{,j}=\delta X/\delta A^j$. (In the following an additional label 0 is put on the indices α_0 and a_0 to prepare the notation for later generalizations to reducible gauge theories.) We assume the set of generators $R^i_{\alpha_0}$ to be linearly independent and complete. The (open) algebra of generators has the general form[1]

$$R^i_{\alpha_0,j}R^j_{\beta_0}-(-1)^{\epsilon_{\alpha_0}\epsilon_{\beta_0}}R^i_{\beta_0,j}R^j_{\alpha_0}=-R^i_{\gamma_0}F^{\gamma_0}_{\alpha_0\beta_0}-M^{ij}_{\alpha_0\beta_0}S_{\mathrm{cl},j}, \tag{2}$$

where $F^{\gamma_0}_{\alpha_0\beta_0}(A)$, in general, are field-dependent structure functions and $M^{ij}_{\alpha_0\beta_0}(A)$ obeys the graded antisymmetry conditions

$$M^{ij}_{\alpha_0\beta_0}=-(-1)^{\epsilon_i\epsilon_j}M^{ji}_{\alpha_0\beta_0}=-(-1)^{\epsilon_{\alpha_0}\epsilon_{\beta_0}}M^{ij}_{\beta_0\alpha_0}. \tag{3}$$

Theories whose generators satisfy Eqs. (2) and (3) are called general gauge theories. In the case $M^{ij}_{\alpha_0\beta_0}=0$ the algebra is closed.

The total configuration space of fields ϕ^A and their Grassmann parities are

$$\phi^A=(A^i,B^{\alpha_0},C^{\alpha_0 a_0}),\quad \delta(\phi^A)\equiv\epsilon_A=(\epsilon_i,\epsilon_{\alpha_0},\epsilon_{\alpha_0}+1);$$

here, the auxiliary fields B^{α_0} are Sp(2)-scalar whereas the (anti)ghosts $C^{\alpha_0 a_0}$ transform as a Sp(2)-doublet. Moreover, for each field ϕ^A one introduces two sets of antifields, a Sp(2)-doublet and a Sp(2)-singlet (with respect to a):

$$\phi^*_{Aa}=(A^*_{ia},B^*_{\alpha_0 a},C^*_{\alpha_0 a a_0}),\quad \epsilon(\phi^*_{Aa})=\epsilon_A+1,\quad \overline{\phi}_A=(\overline{A}_i,\overline{B}_{\alpha_0},\overline{C}_{\alpha_0 a_0}),\quad \epsilon(\overline{\phi}_A)=\epsilon_A .$$

Raising and lowering of Sp(2)-indices is obtained by the invariant tensor of the group,

$$\epsilon^{ab}=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix},\quad \epsilon^{ac}\epsilon_{cb}=\delta^a_b .$$

Let us point to the fact that in the Sp(2)-approach the internal Sp(2) index a_0 of the third component is a dummy one, i.e., it is not affected by operations being introduced by the present approach.

Let us now shortly review the Sp(2)-covariant quantization scheme. The basic object is the bosonic action $S=S(\phi^A,\phi^*_{Aa},\overline{\phi}_A)$ satisfying the extended quantum master equations (i.e., the generating equations with respect to the extended BRST symmetry)

$$\overline{\Delta}^a\exp\{(i/\hbar)S\}=0,\quad \overline{\Delta}^a=\Delta^a+(i/\hbar)V^a, \tag{4}$$

where the odd (second-order) differential operators $\overline{\Delta}^a$ possess the important properties of nilpotency and (relative) anticommutativity:

$$\{\overline{\Delta}^a,\overline{\Delta}^b\}=0. \tag{5}$$

The solution of (4) is sought as a power series in Planck's constant $\hbar$:

$$S=\sum_{n=0}^{\infty}\hbar^n S_{(n)},$$

with the boundary condition $S|_{\phi^*_a=\overline{\phi}=\hbar=0}=S_{\rm cl}(A)$.

To remove the degeneracy of the action S a gauge has to be introduced with the following properties: First, it should lift the degeneracy in ϕ^A and, second, it should retain Eqs. (4), thereby providing the extended BRST symmetry also for the gauge fixed action denoted by $S_{\rm ext}=S_{\rm ext}(\phi^A,\phi^*_{Aa},\overline{\phi}_A)$. Introducing the gauge-fixing Sp(2)-invariant bosonic functional $F=F(\phi^A)$ the action $S_{\rm ext}$ is defined by

$$\exp\{(i/\hbar)S_{\rm ext}\}=\hat{U}(F)\exp\{(i/\hbar)S\},$$

where the operator $\hat{U}(F)$ has the general form

$$\hat{U}(F)=\exp\{(\hbar/i)\hat{T}(F)\},\quad \hat{T}(F)=\tfrac{1}{2}\epsilon_{ab}\{\overline{\Delta}^b,[\overline{\Delta}^a,F]\}.$$

Here, $\hat{T}(F)$ has been chosen such that, by virtue of (5), it commutes with $\overline{\Delta}^a$,

$$[\overline{\Delta}^a,\hat{T}(F)]=0;$$

hence $S_{\rm ext}$ obeys also Eqs. (4):

$$\overline{\Delta}^a\exp\{(i/\hbar)S_{\rm ext}\}=0.$$

Let us now briefly state the essential modifications of the Sp(2)-formalism to obtain the osp(1,2)-covariant quantization of an irreducible complete theory of massive fields whose action S_m depends on the mass m as a further parameter. First, in addition to the m-extended generalized quantum master equations

$$\overline{\Delta}_m^a \exp\{(i/\hbar)S_m\}=0, \quad \overline{\Delta}_m^a=\Delta^a+(i/\hbar)V_m^a, \tag{6}$$

which ensure (anti)BRST invariance, the action S_m is required to satisfy the generating equation of Sp(2)-invariance, too:

$$\overline{\Delta}_\alpha \exp\{(i/\hbar)S_m\}=0, \quad \overline{\Delta}_\alpha=\Delta_\alpha+(i/\hbar)V_\alpha, \tag{7}$$

where $\overline{\Delta}_m^a$ and $\overline{\Delta}_\alpha$ are odd and even (second-order) differential operators, respectively (for explicit expressions see Sec. III below).

As long as $m\neq 0$ the operators $\overline{\Delta}_m^a$ are neither nilpotent nor do they anticommute among themselves; instead, together with the operators $\overline{\Delta}_\alpha$ they generate a superalgebra isomorphic to osp(1,2) (see Appendix A):

$$[\overline{\Delta}_\alpha,\overline{\Delta}_\beta]=(i/\hbar)\epsilon_{\alpha\beta}{}^{\gamma}\overline{\Delta}_\gamma, \tag{8}$$

$$[\overline{\Delta}_\alpha,\overline{\Delta}_m^a]=(i/\hbar)\overline{\Delta}_m^b(\sigma_\alpha)_b{}^a, \tag{9}$$

$$\{\overline{\Delta}_m^a,\overline{\Delta}_m^b\}=(i/\hbar)m^2(\sigma^\alpha)^{ab}\overline{\Delta}_a, \tag{10}$$

where the Sp(2)-indices are raised or lowered corresponding to

$$(\sigma_\alpha)^{ab}=\epsilon^{ac}(\sigma_\alpha)_c{}^b=(\sigma_\alpha)^a{}_c\epsilon^{cb}=\epsilon^{ac}(\sigma_\alpha)_{cd}\epsilon^{ab}, \quad (\sigma_\alpha)_a{}^b=-(\sigma_\alpha)^b{}_a\,.$$

Here, the matrices σ_α $(\alpha=0,+,-)$ generate the group of special linear transformations:

$$\sigma_\alpha\sigma_\beta=g_{\alpha\beta}+\frac{1}{2}\epsilon_{\alpha\beta\gamma}\sigma^\gamma, \quad \sigma^\alpha=g^{\alpha\beta}\sigma_\beta, \quad \mathrm{Tr}(\sigma_\alpha\sigma_\beta)=2g_{\alpha\beta}, \quad g^{\alpha\beta}=\begin{pmatrix}1&0&0\\0&0&2\\0&2&0\end{pmatrix}, \quad g^{\alpha\gamma}g_{\gamma\beta}=\delta^\alpha_\beta,$$

$\epsilon_{\alpha\beta\gamma}$ being the antisymmetric tensor, $\epsilon_{0+-}=1$ [all relations containing the matrices σ_α hold for any representation written below Eqs. (8)–(10)]. Notice that sl$(2,R)$, the even part of osp(1,2), is isomorphic to sp$(2,R)$. From (10) it is obvious that, if and only if the action S_m is Sp(2)-invariant, it can be (anti)BRST-invariant as well. For the generators σ_α we may choose the representation

$$\sigma_0=\tau_3, \quad \sigma_\pm=-\tau_\pm=-\tfrac{1}{2}(\tau_1\pm i\tau_2), \tag{11}$$

where τ_α $(\alpha=1,2,3)$ are the Pauli matrices.

In order that the (anti)commutation relations (8)–(10) are consistent, the structure constants $\epsilon_{\alpha\beta}{}^\gamma$, $(\sigma_\alpha)_b{}^a$ and $(\sigma^\alpha)^{ab}$ on the right-hand side of these relations must and, in fact, will obey certain conditions which follow from the (graded) Jacobi identities. The requirement that the set of these identities, namely

$$[\overline{\Delta}_\alpha,[\overline{\Delta}_\beta,\overline{\Delta}_\gamma]]+[\overline{\Delta}_\beta,[\overline{\Delta}_\gamma,\overline{\Delta}_\alpha]]+[\overline{\Delta}_\gamma,[\overline{\Delta}_\alpha,\overline{\Delta}_\beta]]=0,$$

$$[\overline{\Delta}_\alpha,[\overline{\Delta}_\beta,\overline{\Delta}_m^a]]-[\overline{\Delta}_\beta,[\overline{\Delta}_\alpha,\overline{\Delta}_m^a]]-[[\overline{\Delta}_\alpha,\overline{\Delta}_\beta],\overline{\Delta}_m^a]=0,$$

$$\{\overline{\Delta}_m^a,[\overline{\Delta}_\alpha,\overline{\Delta}_m^b]\}+\{\overline{\Delta}_m^b,[\overline{\Delta}_\alpha,\overline{\Delta}_m^a]\}+[\{\overline{\Delta}_m^a,\overline{\Delta}_m^b\},\overline{\Delta}_\alpha]=0,$$

$$[\overline{\Delta}_m^a,\{\overline{\Delta}_m^b,\overline{\Delta}_m^c\}]+[\overline{\Delta}_m^b,\{\overline{\Delta}_m^c,\overline{\Delta}_m^a\}]+[\overline{\Delta}_m^c,\{\overline{\Delta}_m^a,\overline{\Delta}_m^b\}]=0,$$

be fulfilled is equivalent to the demand that in the adjoint representation of $\overline{\Delta}_\alpha$ and $\overline{\Delta}_m^a$,

$$\overline{\Delta}_\alpha=\begin{pmatrix} R_\alpha & 0 \\ 0 & S_\alpha \end{pmatrix}, \quad \overline{\Delta}_m^a=m\begin{pmatrix} 0 & U^a \\ T^a & 0 \end{pmatrix},$$

the matrices R_α and S_α should form a representation of the (even) subalgebra sl(2) of osp(1,2). The elements of these matrices are

$$(R_\alpha)_\beta{}^\gamma=\epsilon_{\beta\alpha}{}^\gamma, \quad (S_\alpha)_a{}^b=(\sigma_\alpha)_a{}^b, \quad (T^a)_b{}^\alpha=i(\sigma^\alpha)_b{}^a, \quad (U^a)_\alpha{}^b=-i(\sigma_\alpha)^{ab}.$$

Therefore, we have the following conditions: (I) the matrices R_α form the adjoint representation of sl(2) and (II) the matrices S_α form the fundamental representation of this algebra; (III) the matrix elements $(T^a)_b{}^\alpha$ and $(U^a)_\alpha{}^b$ [which do not form a representation of sl(2) because of the different indices] are numerical invariants under sl(2),

$$(T^a)_b{}^\gamma\epsilon_{\gamma\alpha}{}^\beta=(\sigma_\alpha)_b{}^c(T^a)_c{}^\beta-(T^c)_b{}^\beta(\sigma_\alpha)_c{}^a, \quad \epsilon_{\alpha\beta}{}^\gamma(U^a)_\gamma{}^b=(\sigma_\alpha)^a{}_c(U^c)_\beta{}^b-(U^a)_\beta{}^c(\sigma_\alpha)_c{}^b;$$

and (IV) there holds the cyclic identity,

$$(\sigma^\alpha)^{ab}(\sigma_\alpha)^c{}_d+\text{cyclic perm}(a,b,c)=0,$$

which restricts the possible representations $\overline{\Delta}_m^a$ of the algebra sl(2) spanned by $\overline{\Delta}_\alpha$ (in general, not every ordinary Lie algebra can be extended to a superalgebra).

Let us notice that by invoking osp(1,2)-symmetry the notion of ghost number will be a natural property of the superalgebra (8)–(10). Indeed, from (8) and (9) we observe the relations

$$(\hbar/i)[\overline{\Delta}_0,\overline{\Delta}_\pm]=\pm 2\overline{\Delta}_\pm, \quad (\hbar/i)[\overline{\Delta}_0,\overline{\Delta}_m^a]=\overline{\Delta}_m^b(\sigma_0)_b{}^a,$$

showing that $(\hbar/i)\overline{\Delta}_0=\Delta_{gh}$ is the Faddeev–Popov operator whose eigenvalues $gh(X)$ define the ghost numbers of the corresponding quantities X of the theory in question, i.e., $([\Delta_{gh},X]-gh(X)X)\exp\{(i/\hbar)S_m\}=0$.

However, insisting on osp(1,2)-symmetry this approach brings in a fundamentally new aspect. Namely, in order to express the superalgebra (8)–(10) by operator identities and to get nontrivial solutions of the generating equations (6) and (7) one is forced to enlarge the set of antifields. More precisely, because the (anti)ghost fields $C^{\alpha_0 a_0}$ as well as their related antifields transform under Sp(2) in a nontrivial way one has to introduce additional sources,

$$\eta_A=(D_i,E_{\alpha_0},F_{\alpha_0 a_0}), \quad \epsilon(\eta_A)=\epsilon_A.$$

Let us remark that D_i always could be set equal to zero (here also E_{α_0} has been introduced only for the sake of completeness and formal analogy to other fields—it could be chosen equal to zero, too). For irreducible theories $F_{\alpha_0 a_0}$ is necessary in order to close the extended BRST algebra. [For L-stage reducible theories both sources $F_{\alpha_s|a_0\cdots a_s}$ and $E_{\alpha_s|a_0\cdots a_s}$ $(s=0,...,L)$ are necessary in order to close the extended BRST algebra with respect to the more general space of auxiliary and (anti)ghost fields $B^*_{\alpha_s a|a_1\cdots a_s}$, $C^*_{\alpha_s a|a_0\cdots a_s}$, $\overline{B}_{\alpha_s|a_1\cdots a_s}$, $\overline{C}_{\alpha_s|a_0\cdots a_s}$.]

In order to set up the gauge fixing, the new generalized gauge fixed quantum action $S_{m,\text{ext}}=S_{m,\text{ext}}(\phi^A,\phi^*_{Aa},\overline{\phi}_A,\eta_A)$ will be introduced according to

$$\exp\{(i/\hbar)S_{m,\text{ext}}\}=\hat{U}_m(F)\exp\{(i/\hbar)S_m\}, \tag{12}$$

where the operator $\hat{U}_m(F)$ has to be chosen as

$$\hat{U}_m(F)=\exp\{(\hbar/i)\hat{T}_m(F)\}, \quad \overline{T}_m(F)=\tfrac{1}{2}\epsilon_{ab}\{\overline{\Delta}_m^b,[\overline{\Delta}_m^a,F]\}+(i/\hbar)^2m^2F, \tag{13}$$

$F=F(\phi^A)$ being the gauge fixing functional. Then, by virtue of (9) and (10), one establishes the relations

$$[\overline{\Delta}_m^a,\overline{T}_m(F)]=\tfrac{1}{2}(i/\hbar)m^2(\sigma_\alpha)^a{}_b[\overline{\Delta}_m^b,[\overline{\Delta}^\alpha,F]] \tag{14}$$

and

$$[\overline{\Delta}_\alpha,\hat{T}_m(F)]=\tfrac{1}{2}\epsilon_{ab}\{\overline{\Delta}_m^b,[\overline{\Delta}_m^a,[\overline{\Delta}_\alpha,F]]\}+(i/\hbar)^2m^2[\overline{\Delta}_\alpha,F]. \tag{15}$$

Restricting F to be a Sp(2) scalar by imposing

$$[\overline{\Delta}_\alpha,F]S_m=0, \tag{16}$$

it can be verified by using the explicit expressions of $\overline{\Delta}_m^a$ and $\overline{\Delta}_\alpha$ that the commutators $[\overline{\Delta}_m^a,\hat{U}_m(F)]\exp\{(i/\hbar)S_m\}=0$ and $[\overline{\Delta}_\alpha,\hat{U}_m(F)]\exp\{(i/\hbar)S_m\}=0$ vanish on the subspace of admissible actions S_m (the proof of this statement will be postponed to Sec. III). Hence, the gauge fixed action $S_{m,\mathrm{ext}}$ satisfies Eqs. (6) and (7) as well,

$$\overline{\Delta}_m^a\exp\{(i/\hbar)S_{m,\mathrm{ext}}\}=0, \quad \overline{\Delta}_\alpha\exp\{(i/\hbar)S_{m,\mathrm{ext}}\}=0. \tag{17}$$

Here, some remarks are in order. The operator $\overline{\Delta}_m^a$ is a nonlinear one; to be able to express the superalgebra (8)–(10) through operator identities the operator $\overline{\Delta}_\alpha$ must be a nonlinear operator, too. Therefore, it is not possible to impose the strong operator equation $[\overline{\Delta}_\alpha,F]=0$ except for the particular case $F=F(A^i)$ where $[\overline{\Delta}_\alpha,F]$ vanishes identically, rather it must be replaced by the weaker condition $[\overline{\Delta}_\alpha,F]S_m=0$. As a consequence, in order to prove that $S_{m,\mathrm{ext}}$ possesses the same symmetry properties as S_m one needs the explicit realization of the operators $\overline{\Delta}_m^a$ and $\overline{\Delta}_\alpha$.

In this way the Sp(2)-covariant approach is generalized to another one based on the superalgebra osp(1,2). Moreover, in this approach one can introduce a mass m (which is necessary at least intermediately in the process of BPHZL renormalization) without breaking the extended BRST symmetry.

III. EXPLICIT CONSTRUCTION OF THE OPERATORS OF osp(1,2)-ALGEBRA

After having stated the general structure of the osp(1,2) quantization procedure being an obvious extension of Sp(2) quantization we have to find an operational realization of the general operators just introduced. Thereby we proceed in such away that all formulas hold also for reducible gauge theories, except for the definition of the operator $(P_+)^{Ba}_{Ab}$ introduced below which has to be generalized. The explicit expressions for the operators $\overline{\Delta}_m^a$ and $\overline{\Delta}_\alpha$ in the generating equations (6) and (7) will be determined in two steps: First we construct a functional $\overline{S}_m$ (at lowest order of $\hbar$) which is linear with respect to the antifields and is invariant under both (anti)BRST- and Sp(2) transformations; later on we generalize to the case of nonlinear dependence. The corresponding (linear) symmetry operators being denoted by $\mathbf{s}_m^a$ and $\mathbf{d}_\alpha$ are required to fulfill the osp(1,2)-superalgebra:

$$[\mathbf{d}_\alpha,\mathbf{d}_\beta]=\epsilon_{\alpha\beta}{}^\gamma\mathbf{d}_\gamma, \quad [\mathbf{d}_\alpha,\mathbf{s}_m^a]=\mathbf{s}_m^b(\sigma_\alpha)_b{}^a, \quad \{\mathbf{s}_m^a,\mathbf{s}_m^b\}=-m^2(\sigma^\alpha)^{ab}\mathbf{d}_\alpha.$$

Let us make for $\overline{S}_m$ the following ansatz:

$$\overline{S}_m=S_{\mathrm{cl}}+W_X, \quad W_X=(\tfrac{1}{2}\epsilon_{ab}\mathbf{s}_m^b\mathbf{s}_m^a+m^2)X, \tag{18}$$

where X is assumed to be the following Sp(2)-scalar (in fact the only one we are able to build up linear in the antifields),

$$X=\overline{\phi}_A\phi^A, \quad \mathbf{d}_\alpha X=0.$$

Then, by virtue of

$$\mathbf{s}_m^c(\tfrac{1}{2}\epsilon_{ab}\mathbf{s}_m^b\mathbf{s}_m^a+m^2)=\tfrac{1}{2}m^2(\sigma^\alpha)^c{}_d s_m^d\mathbf{d}_\alpha, \quad [\mathbf{d}_\alpha,\tfrac{1}{2}\epsilon_{ab}\mathbf{s}_m^b\mathbf{s}_m^a+m^2]=0,$$

it follows that $\overline{S_m}$ is both (anti)BRST- and Sp(2)-invariant, $\mathbf{s}_m^a\overline{S_m}=0$ and $\mathbf{d}_\alpha\overline{S_m}=0$. Thereby, it is taken into account that due to gauge invariance (1) of $S_{\rm cl}(A)$ it holds $\mathbf{s}_m^a S_{\rm cl}(A)=0$ with $\mathbf{s}_m^a A^i=R^i_{\alpha_0}C^{\alpha_0 a}$ (here, the action of $\mathbf{s}_m^a$—and also of $\mathbf{d}_\alpha$—on the auxiliary and (anti)ghost fields is irrelevant and already left open).

The strategy to define the operators $\overline{\Delta}_m^a=\Delta^a+(i/\hbar)V_m^a$, $\overline{\Delta}_\alpha=\Delta_\alpha+(i/\hbar)V_\alpha$ is governed by a specific realization of the (anti)BRST- and Sp(2)-transformations of the antifields (extending the standard definitions of Ref. 1). Therefore, let us decompose $\mathbf{s}_m^a$ and $\mathbf{d}_\alpha$ into a component acting on the fields and another one acting on the antifields as follows:

$$\mathbf{s}_m^a=(\mathbf{s}_m^a\phi^A)\frac{\delta_L}{\delta\phi^A}+V_m^a, \quad \mathbf{d}_\alpha=(\mathbf{d}_\alpha\phi^A)\frac{\delta_L}{\delta\phi^A}+V_\alpha.$$

The action of V_m^a and V_α on $\overline{\phi}_A$, ϕ^*_{Aa} and η_A are uniquely defined by

$$V_m^a\overline{\phi}_A=\epsilon^{ab}\phi^*_{Ab}, \quad V_\alpha\overline{\phi}_A=\overline{\phi}_B(\sigma_\alpha)^B_A,$$

$$V_m^a\phi^*_{Ab}=m^2(P_+)^{Ba}_{Ab}\overline{\phi}_B-\delta^a_b\eta_A, \quad V_\alpha\phi^*_{Aa}=\phi^*_{Ab}(\sigma_\alpha)^b{}_a+\phi^*_{Ba}(\sigma_\alpha)^B{}_A, \tag{19}$$

$$V_m^a\eta_A=-m^2\epsilon^{ab}(P_-)^{Bc}_{Ab}\phi^*_{Bc}, \quad V_\alpha\eta_A=\eta_B(\sigma_\alpha)^B{}_A,$$

where the following abbreviations are used:

$$(P_-)^{Ba}_{Ab}\equiv(P_+)^{Ba}_{Ab}-(P_+)^B_A\delta^a_b+\delta^B_A\delta^a_b, \quad (P_+)^B_A\equiv\delta^b_a(P_+)^{Ba}_{Ab}, \quad (\sigma_\alpha)^{BA}\equiv(\sigma_\alpha)^b{}_a(P_+)^{Ba}_{Ab}.$$

Here, we introduced the matrix

$$(P_+)^{Ba}_{Ab}\equiv\begin{cases}\delta^i_j\delta^a_b & \text{for} \quad A=i,B=j,\\ \delta^{\beta_0}_{\alpha_0}\delta^a_b & \text{for} \quad A=\alpha_0,B=\beta_0,\\ 2\,\delta^{\beta_0}_{\alpha_0}S^{b_0a}_{a_0b} & \text{for} \quad A=\alpha_0a_0,B=\beta_0b_0,\\ 0 & \text{otherwise,}\end{cases}$$

where the symmetrizer $S^{b_0a}_{a_0b}$ is defined as

$$S^{b_0a}_{a_0b}\equiv\tfrac{1}{2}\frac{\partial}{\partial X^{a_0}}\frac{\partial}{\partial X^b}X^aX^{b_0}=\tfrac{1}{2}(\delta^a_b\delta^{a_0}_{b_0}+\delta^a_{a_0}\delta^{b_0}_b),$$

X^a being independent bosonic variables.

The matrices $(P_-)^B_A\equiv\delta^a_b(P_-)^{Ba}_{Ab}$ and $(\sigma_\alpha)^B{}_A$ act nontrivially on the components of the (anti) fields having an internal (dummy) Sp(2) index. For example,

$$(P_-)^B_A\overline{\phi}_B=(0,0,-\overline{C}_{\alpha_0a_0}), \quad \overline{\phi}_B(\sigma_\alpha)^B{}_A=(0,0,\overline{C}_{\alpha_0c}(\sigma_\alpha)^c{}_{a_0}).$$

Therefore, V_α acts only on the (anti)ghost part of the antifields, and V_m^a is partly of that kind. Of course, we could have used also a componentwise notation, however, then the equations would be less easy to survey [a componentwise notation of the transformations (19) is given in Appendix E].

In order to prove that the transformations (19) obey the osp(1,2)-superalgebra

$$[V_\alpha, V_\beta] = \epsilon_{\alpha\beta}{}^\gamma V_\gamma, \quad [V_\alpha, V_m^a] = V_m^b (\sigma_\alpha)_b{}^a, \quad \{V_m^a, V_m^b\} = -m^2 (\sigma^\alpha)^{ab} V_\alpha, \tag{20}$$

one needs the following two equalities:

$$\epsilon^{ad}(P_+)^{Bb}_{Ad} + \epsilon^{bd}(P_+)^{Ba}_{Ad} = -(\sigma^\alpha)^{ab}(\sigma_\alpha)^B{}_A,$$

$$\epsilon^{ad}(P_+)^{Bb}_{Ac} + \epsilon^{bd}(P_+)^{Ba}_{Ac} - (\sigma^\alpha)^{ab}(\sigma_\alpha)^e{}_c (P_-)^{Bd}_{Ae} = -(\sigma^\alpha)^{ab}((\sigma_\alpha)^d{}_c \delta^B_A + \delta^d_c (\sigma_\alpha)^B{}_A),$$

or, equivalently,

$$\epsilon^{ad}\delta^b_c + \epsilon^{bd}\delta^a_c = -(\sigma^\alpha)^{ab}(\sigma_\alpha)^d{}_c, \tag{21}$$

$$(\sigma^\alpha)^{ab}((\sigma_\alpha)^{b_0}{}_c \delta^d_{a_0} + \delta^{b_0}_c (\sigma_\alpha)^d{}_{a_0}) = (\sigma^\alpha)^{ab}((\sigma_\alpha)^d{}_c \delta^{b_0}_{a_0} + \delta^d_c (\sigma_\alpha)^{b_0}{}_{a_0}), \tag{22}$$

and the relation $(P_-)^{Ab}_{Cd}(P_+)^{Cd}_{Ba} = 0$.

Then from (18) one gets for $\overline{S}_m$ the expression

$$\overline{S}_m = S_{\rm cl} + (\eta_A - \tfrac{1}{2} m^2 (P_+)^B_A \bar{\phi}_B)\phi^A - (\mathbf{s}^a_m \phi^A)\phi^*_{Aa} + \bar{\phi}_A(\tfrac{1}{2}\epsilon_{ab}\mathbf{s}^b_m \mathbf{s}^a_m + m^2)\phi^A. \tag{23}$$

Now, the symmetry properties of $\overline{S}_m$ with respect to (anti)BRST- and Sp(2)-transformations may be expressed by the following set of equations:

$$\tfrac{1}{2}(\overline{S}_m, \overline{S}_m)^a + V_m^a \overline{S}_m = 0, \quad \tfrac{1}{2}\{\overline{S}_m, \overline{S}_m\}_\alpha + V_\alpha \overline{S}_m = 0; \tag{24}$$

here the extended antibrackets $(F,G)^a$, introduced for the first time in Ref. 1, define an odd-graded and $\{F,G\}_\alpha$ a new even-graded algebraic structure on the space of fields and antifields (see Appendix B):

$$(F,G)^a = \frac{\delta F}{\delta \phi^A}\frac{\delta G}{\delta \phi^*_{Aa}} - (-1)^{(\epsilon(F)+1)(\epsilon(G)+1)}\frac{\delta G}{\delta \phi^A}\frac{\delta F}{\delta \phi^*_{Aa}} \tag{25}$$

and

$$\{F,G\}_\alpha = (\sigma_\alpha)_B{}^A\left(\frac{\delta F}{\delta \phi^A}\frac{\delta G}{\delta \eta_B} + (-1)^{\epsilon(F)\epsilon(G)}\frac{\delta G}{\delta \phi^A}\frac{\delta F}{\delta \eta_B}\right), \tag{26}$$

and, in accordance with (19), the first-order differential operators V_m^a and V_α are given by

$$V_m^a = \epsilon^{ab}\phi^*_{Ab}\frac{\delta}{\delta \bar{\phi}_A} - \eta_A \frac{\delta}{\delta \phi^*_{Aa}} + m^2 (P_+)^{Ba}_{Ab}\bar{\phi}_B \frac{\delta}{\delta \phi^*_{Ab}} - m^2 \epsilon^{ab}(P_-)^{Bc}_{Ab}\phi^*_{Bc}\frac{\delta}{\delta \eta_A} \tag{27}$$

and

$$V_\alpha = \bar{\phi}_B (\sigma_\alpha)^B{}_A \frac{\delta}{\delta \bar{\phi}_A} + (\phi^*_{Ab}(\sigma_\alpha)^b{}_a + \phi^*_{Ba}(\sigma_\alpha)^B{}_A)\frac{\delta}{\delta \phi^*_{Aa}} + \eta_B (\sigma_\alpha)^B{}_A \frac{\delta}{\delta \eta_A}. \tag{28}$$

We also introduce the second-order differential operators Δ^a and Δ_α whose structure is extracted from (25) and (26):

$$\Delta^a=(-1)^{\epsilon_A}\frac{\delta_L}{\delta\phi^A}\frac{\delta}{\delta\phi^*_{Aa}},\quad \Delta_\alpha=(-1)^{\epsilon_A}(\sigma_\alpha)_B{}^A\frac{\delta_L}{\delta\phi^A}\frac{\delta}{\delta\eta_B}. \tag{29}$$

Let us now consider the general case where the action S_m is assumed to appear in the form of a series expansion in powers of $\hbar$ which may depend also nonlinear on the antifields. This action will be required to satisfy the following set of quantum master equations (see Appendix C):

$$\tfrac{1}{2}(S_m,S_m)^a+V^a_mS_m=i\hbar\Delta^aS_m,\quad \tfrac{1}{2}\{S_m,S_m\}_\alpha+V_\alpha S_m=i\hbar\Delta_\alpha S_m, \tag{30}$$

or, equivalently,

$$\overline{\Delta}{}^a_m\exp\{(i/\hbar)S_m\}=0,\quad \overline{\Delta}_\alpha\exp\{(i/\hbar)S_m\}=0. \tag{31}$$

Furthermore, by an explicit calculation it can be verified that $\overline{\Delta}{}^a_m=\Delta^a+(i/\hbar)V^a_m$ and $\overline{\Delta}_\alpha=\Delta_\alpha+(i/\hbar)V_\alpha$ obey the osp(1,2)-superalgebra (8)–(10).

In order to lift the degeneracy of S_m we follow the general gauge-fixing procedure suggested by (12) and (13). Let us introduce the gauge fixed action

$$\exp\{(i/\hbar)S_{m,\mathrm{ext}}\}=\hat{U}_m(F)\exp\{(i/\hbar)S_m\}, \tag{32}$$

where the operator $\hat{U}_m(F)=\exp\{(i/\hbar)\hat{T}_m(F)\}$ is defined according to the formula (13). If the gauge-fixing functional is assumed to depend only on the fields, $F=F(\phi^A)$, then one gets

$$\hat{U}_m(F)=\exp\left\{\frac{\delta F}{\delta\phi^A}\left(\frac{\delta}{\delta\overline{\phi}_A}-\frac{1}{2}m^2(P_-)^A_B\frac{\delta}{\delta\eta_B}\right)-(\hbar/i)\frac{1}{2}\epsilon_{ab}\frac{\delta}{\delta\phi^*_{Aa}}\frac{\delta^2F}{\delta\phi^A\delta\phi^B}\frac{\delta}{\delta\phi^*_{Bb}}+(i/\hbar)m^2F\right\}. \tag{33}$$

Let us prove that $S_{m,\mathrm{ext}}$ obeys the generating equations (30) and (31) as well. Clearly, since $\overline{\Delta}{}^a_m$, $\overline{\Delta}_\alpha$ and $\hat{U}_m(F)$ do not commute with each other, this proof will be more involved than in the Sp(2)-approach. This is due to the fact that, looking back at (14) and (15), neither

$$[\overline{\Delta}{}^a_m,\hat{T}_m(F)]=\tfrac{1}{2}(i/\hbar)m^2(\sigma^\alpha)^a{}_b[\overline{\Delta}{}^b_m,[\overline{\Delta}_\alpha,F]]$$

nor

$$[\overline{\Delta}_\alpha,\hat{T}_m(F)]=\tfrac{1}{2}\epsilon_{ab}\{\overline{\Delta}{}^b_m,[\overline{\Delta}{}^a_m,[\overline{\Delta}_\alpha,F]]\}+(i/\hbar)^2m^2[\overline{\Delta}_\alpha,F]$$

does vanish, since due to the nonlinearity of $\overline{\Delta}_\alpha$ one cannot require the strong condition $[\overline{\Delta}_\alpha,F]=0$. However, a direct verification shows that $\hat{T}_m(F)$ commutes with any term on the right-hand side of both previous relations, i.e., it holds

$$[\hat{T}_m(F),[\overline{\Delta}{}^a_m,\hat{T}_m(F)]]=0,\quad [\hat{T}_m(F),[\overline{\Delta}_\alpha,\hat{T}_m(F)]]=0.$$

Then, by the help of these relations one obtains

$$[\overline{\Delta}{}^a_m,\hat{U}_m(F)]=(\hbar/i)\hat{U}_m(F)[\overline{\Delta}{}^a_m,\hat{T}_m(F)],\quad [\overline{\Delta}_\alpha,\hat{U}_m(F)]=(\hbar/i)\hat{U}_m(F)[\overline{\Delta}_\alpha,\hat{T}_m(F)].$$

Let us require [see (16)]

$$[\overline{\Delta}_\alpha,F]S_m\equiv(\sigma_\alpha)_B{}^A\frac{\delta F}{\delta\phi^A}\frac{\delta S_m}{\delta\eta_B}=0,\quad (\sigma_\alpha)_B{}^A\frac{\delta F}{\delta\phi^A}\phi^B=0. \tag{34}$$

Then, taking into account that S_m solves the generating equations (31), it is easily seen that $[\overline{\Delta}_m^a, \hat{U}_m(F)]$ and $[\overline{\Delta}_\alpha, \hat{U}_m(F)]$ vanish after acting on $\exp\{(i/\hbar)S_m\}$:

$$[\overline{\Delta}_m^a, \hat{U}_m(F)]\exp\{(i/\hbar)S_m\}=0, \quad [\overline{\Delta}_\alpha, \hat{U}_m(F)]\exp\{(i/\hbar)S_m\}=0.$$

Summarizing, we have the results

$$\overline{\Delta}_m^a \exp\{(i/\hbar)S_{m,\mathrm{ext}}\}=0, \quad \overline{\Delta}_\alpha \exp\{(i/\hbar)S_{m,\mathrm{ext}}\}=0, \tag{35}$$

i.e., the gauge-fixed action $S_{m,\mathrm{ext}}$ satisfies the same generating equations (30) and (31) as S_m, which is indeed what we intended to prove.

Finally, let us return to the requirement (34) and discuss how it can be fulfilled. If we differentiate Eqs. (30) with respect to η_A, we get

$$Q_m^a\left(\frac{\delta S_m}{\delta \eta_A}-\phi^A\right)=0, \quad Q_\alpha\left(\frac{\delta S_m}{\delta \eta_A}-\phi^A\right)=0, \tag{36}$$

where the operators Q_m^a and Q_α are defined by the formulas

$$Q_m^a X \equiv (S_m, X)^a - i\hbar \overline{\Delta}_m^a X, \quad Q_\alpha X \equiv \{S_m, X\}_\alpha - i\hbar \overline{\Delta}_\alpha X, \tag{37}$$

with arbitrary functional X. These operators possess the properties

$$[Q_\alpha, Q_\beta]=\epsilon_{\alpha\beta}{}^\gamma Q_\gamma, \quad [Q_\alpha, Q_m^a]=Q_m^b(\sigma_\alpha)_b{}^a, \quad \{Q_m^a, Q_m^b\}=-m^2(\sigma^\alpha)^{ab}Q_\alpha. \tag{38}$$

The equations (36) suggest that the restriction on S_m to be linear in η_A seems to be the natural solution of the requirement (34), namely

$$\frac{\delta S_m}{\delta \eta_A}=\phi^A. \tag{39}$$

Then, the second equation in (30) simplifies into

$$(\sigma_\alpha)_B{}^A \frac{\delta S_m}{\delta \phi^A}\phi^B + V_\alpha S_m = 0,$$

since σ_α are traceless. Although the restriction (39) seems to be a particular but not the general case it should be preferred to all other solutions since, due to its linearity, it can be realized for any (renormalized) action S_m (see also Appendix D).

IV. WARD IDENTITIES, GENERATING FUNCTIONALS AND GAUGE (IN)DEPENDENCE

Next, we turn to the question of gauge independence of physical quantities, especially of the *S*-matrix. In discussing this question it is convenient to study first the symmetry properties of the vacuum functional,

$$Z_m(0)=\int d\phi^A \exp\{(i/\hbar)S_{m,\mathrm{eff}}\}, \tag{40}$$

where $S_{m,\mathrm{eff}}(\phi^A)=S_{m,\mathrm{ext}}(\phi^A, \phi^*_{Aa}, \overline{\phi}_A, \eta_A)|_{\phi^*_a=\overline{\phi}=\eta=0}$. It can be represented in the form

$$Z_m(0)=\int d\phi^A d\eta_A d\phi^*_{Aa} d\pi^{Aa} d\overline{\phi}_A d\lambda^A \delta(\eta_A)\exp\{(i/\hbar)(S_{m,\mathrm{ext}}-W_X)\} \tag{41}$$

with

$$W_X = (\eta_A - \tfrac{1}{2} m^2 (P_+)^B_A \overline{\phi}_B)\phi^A - \phi^*_{Aa}\pi^{Aa} - \overline{\phi}_A(\lambda^A - \tfrac{1}{2} m^2 (P_-)^A_B \phi^B), \tag{42}$$

where we have extended the space of variables by introducing the auxiliary fields π^{Aa} and λ^A. It is straightforward to check that W_X can be cast into the form (42). Indeed, performing in (41) first of all the integrations over λ^A and π^{Aa}, which yields the delta-functions $\delta(\overline{\phi}_A)$ and $\delta(\phi^*_{Aa})$, and, after that, carrying out the integrations over $\overline{\phi}_A$, ϕ^*_{Aa} and η_A one gets the expression (40).

Since η_A for $m \neq 0$ transforms nontrivially under osp(1,2) we express $\delta(\eta_A)$ by

$$\delta(\eta_A) = \int d\zeta^A \exp\{(i/\hbar)\eta_A \zeta^A\}$$

and change within (41) the integration variables ϕ^A and λ^A according to $\phi^A \to \phi^A + \zeta^A$ and $\lambda^A \to \lambda^A + (1/2)m^2((P_-)^A_B - (P_+)^A_B)\zeta^B$. Then, for $Z_m(0)$ this yields

$$Z_m(0) = \int d\phi^A d\eta_A d\zeta^A d\phi^*_{Aa} d\pi^{Aa} d\overline{\phi}_A d\lambda^A \exp\{(i/\hbar)(S^\zeta_{m,\mathrm{ext}} - W_X)\}, \tag{43}$$

where $S^\zeta_{m,\mathrm{ext}}$ is obtained from $S_{m,\mathrm{ext}}$ by performing the replacement $\phi^A \to \phi^A + \zeta^A$. At this stage we remark that $(\phi^A, \pi^{Aa}, \lambda^A)$ and $(\overline{\phi}_A, \phi^*_{Aa}, \eta_A)$ constitute the components of the superfield and superantifield, respectively, of the superfield quantization scheme;[9] here we changed the notation $J_A \to \eta_A$ relative to Ref. 9. Of course, the formalism introduced here may be written in that form also.

The term W_X may be cast into the osp(1,2)-invariant form

$$W_X = \tfrac{1}{2}\epsilon_{ab}(V^b_m(V^a_m X - X U^a_m) + (V^a_m X - X U^a_m)U^b_m) + m^2 X,$$

where

$$X = \overline{\phi}_A \phi^A, \quad V_\alpha X + X U_\alpha = 0,$$

with V^a_m and V_α, whose action on $\overline{\phi}_A$, ϕ^*_{Aa}, η_A are already defined in (19), satisfying the osp(1,2)-superalgebra (20), and the action of U^a_m and U_α on ϕ^A, π^{Aa}, λ^A and ζ^A being defined according to

$$\begin{aligned}
&\phi^A U^a_m = \pi^{Aa}, \quad \phi^A U_\alpha = \phi^B (\sigma_\alpha)_B{}^A, \\
&\pi^{Ab} U^a_m = \epsilon^{ab}\lambda^A + m^2 \epsilon^{ac}(P_+)^{Ab}_{Bc}\phi^B, \quad \pi^{Aa} U_\alpha = \pi^{Ab}(\sigma_\alpha)_b{}^a + \pi^{Ba}(\sigma_\alpha)_B{}^A, \\
&\lambda^A U^a_m = m^2 (P_-)^{Aa}_{Bb}\pi^{Bb}, \quad \lambda^A U_\alpha = \lambda^B (\sigma_\alpha)_B{}^A, \\
&\zeta^A U^a_m = 0, \quad \zeta^A U_\alpha = \zeta^B (\sigma_\alpha)_B{}^A.
\end{aligned} \tag{44}$$

Here, U^a_m and U_α are defined as right derivatives, in contrast to V^a_m and V_α, which was defined as left ones. In order to prove that the transformations (44) obey the osp(1,2)-superalgebra

$$[U_\alpha, U_\beta] = -\epsilon_{\alpha\beta}{}^\gamma U_\gamma, \quad [U_\alpha, U^a_m] = -U^b_m (\sigma_\alpha)_b{}^a, \quad \{U^a_m, U^b_m\} = m^2 (\sigma^\alpha)^{ab} U_\alpha, \tag{45}$$

one needs the following two equalities:

$$\epsilon^{ad}(P_+)^{Bb}_{Ad} + \epsilon^{bd}(P_+)^{Ba}_{Ad} = (\sigma^\alpha)^{ab}(\sigma_\alpha)_A{}^B,$$

$$\epsilon^{ad}(P_-)^{Bb}_{Ac} + \epsilon^{bd}(P_-)^{Ba}_{Ac} - (\sigma^\alpha)^{ab}(\sigma_\alpha)^e{}_c (P_+)^{Bd}_{Ae} = (\sigma^\alpha)^{ab}((\sigma_\alpha)_c{}^d \delta^B_A + \delta^d_c (\sigma_\alpha)_A{}^B),$$

and the relation $(P_-)^{Ab}_{Cd}(P_+)^{Cd}_{Ba}=0$. It is easily seen that both equalities are equivalent to (21) and (22), too.

Inserting into (43) the relations (32) and (33) and integrating by parts yields

$$Z_m(0)=\int d\phi^A d\eta_A d\zeta^A d\phi^*_{Aa} d\pi^{Aa} d\overline{\phi}_A d\lambda^A \exp\{(i/\hbar)(S^\zeta_m+W^\zeta_F-W_X)\} \tag{46}$$

with the following expression for W_F:

$$W_F=-\frac{\delta F}{\delta\phi^A}\left(\lambda^A+\frac{1}{2}m^2(P_+)^A_B\phi^B\right)-\frac{1}{2}\epsilon_{ab}\pi^{Aa}\frac{\delta^2 F}{\delta\phi^A\delta\phi^B}\pi^{Bb}+m^2F, \tag{47}$$

which may be recast into the osp(1,2)-invariant form

$$W_F=F(\tfrac{1}{2}\epsilon_{ab}U^b_mU^a_m+m^2),\quad FU_\alpha=0.$$

(Again, S^ζ_m and W^ζ_F are obtained from S_m and W_F, respectively, by carrying out the replacement $\phi^A\rightarrow\phi^A+\zeta^A$.)

Let us now introduce the combined (first-order) differential operators

$$L^a_mX\equiv V^a_mX-(-1)^{\epsilon(X)}XU^a_m,\quad L_\alpha X\equiv V_\alpha X+XU_\alpha,$$

X being an arbitrary functional, where U^a_m and U_α, in accordance with (44), are defined by

$$U^a_m=\frac{\delta}{\delta\phi^A}\pi^{Aa}+\frac{\delta}{\delta\pi^{Ab}}\epsilon^{ab}\lambda^A+\frac{\delta}{\delta\pi^{Ab}}m^2\epsilon^{ac}(P_+)^{Ab}_{Bc}\phi^B+\frac{\delta}{\delta\lambda^A}m^2(P_-)^{Aa}_{Bb}\pi^{Bb}$$

and

$$U_\alpha=\frac{\delta}{\delta\phi^A}\phi^B(\sigma_\alpha)_B{}^A+\frac{\delta}{\delta\lambda^A}\lambda^B(\sigma_\alpha)_B{}^A+\frac{\delta}{\delta\pi^{Aa}}(\pi^{Ab}(\sigma_\alpha)_b{}^a+\pi^{Ba}(\sigma_\alpha)_B{}^A)+\frac{\delta}{\delta\zeta^A}\zeta^B(\sigma_\alpha)_B{}^A;$$

here, the derivatives with respect to ϕ^A, π^{Ab}, λ^A and ζ^A are right ones [V^a_m and V_α are already defined in (27) and (28)]. Then, by virtue of (20) and (45), it follows that L^a_m and L_α satisfy the osp(1,2)-superalgebra

$$[L_\alpha,L_\beta]=\epsilon_{\alpha\beta}{}^\gamma L_\gamma,\quad [L_\alpha,L^a_m]=L^b_m(\sigma_\alpha)_b{}^a,\quad \{L^a_m,L^b_m\}=-m^2(\sigma^\alpha)^{ab}L_\alpha.$$

Hence, it holds

$$L^c_m(\tfrac{1}{2}\epsilon_{ab}L^b_mL^a_m+m^2)=\tfrac{1}{2}m^2(\sigma^\alpha)^c{}_dL^d_mL_\alpha,\quad [L_\alpha,\tfrac{1}{2}\epsilon_{ab}L^b_mL^a_m+m^2]=0. \tag{48}$$

We assert now that (46) is invariant under the following (global) transformations [thereby, one has to make use of the first equation in (30) and (48), respectively]:

$$\begin{gathered}\delta_m\phi^A=\phi^AU^a_m\mu_a,\quad \delta_m\zeta^A=0,\quad \delta_m\overline{\phi}_A=\mu_aV^a_m\overline{\phi}_A,\\ \delta_m\pi^{Ab}=\pi^{Ab}U^a_m\mu_a,\quad \delta_m\phi^*_{Ab}=\mu_aV^a_m\phi^*_{Ab}+\mu_a(S^\zeta_m,\phi^*_{Ab})^a,\\ \delta_m\lambda^A=\lambda^AU^a_m\mu_a,\quad \delta_m\eta_A=\mu_aV^a_m\eta_A,\end{gathered} \tag{49}$$

where μ_a, $\epsilon(\mu_a)=1$, is a Sp(2)-doublet of constant anticommuting parameters. The transformations (49) realize the m-extended BRST symmetry in the space of variables ϕ^A, $\overline{\phi}_A$, ϕ^*_{Aa}, η_A, π^{Aa}, λ^A and ζ^A.

Moreover, it is straightforward to check that (46) is also invariant under the following transformations [where one has to make use of the second equation in (30) and (48), respectively]:

$$\delta\phi^A = \phi^A U_\alpha \theta^\alpha, \quad \delta\zeta^A = \zeta^A U_\alpha \theta^\alpha, \quad \delta\overline{\phi}_A = \theta^\alpha V_\alpha \overline{\phi}_A,$$
$$\delta\pi^{Ab} = \pi^{Ab} U_\alpha \theta^\alpha, \quad \delta\phi^*_{Ab} = \theta^\alpha V_\alpha \phi^*_{Ab}, \tag{50}$$
$$\delta\lambda^A = \lambda^A U_\alpha \theta^\alpha, \quad \delta\eta_A = \theta^\alpha V_\alpha \eta_A + \theta^\alpha \{S^\zeta_m, \eta_A\}_\alpha,$$

where θ^α, $\epsilon(\theta^\alpha)=0$, are constant commuting parameters. The transformations (50) realize the Sp (2)-symmetry in the space of variables ϕ^A, $\overline{\phi}_A$, ϕ^*_{Aa}, η_A, π^{Aa}, λ^A and ζ^A.

In principle, for a general gauge functional F, μ_a may be assumed to depend on all these variables ϕ^A, $\overline{\phi}_A$, ϕ^*_{Aa}, η_A, π^{Aa}, λ^A and ζ^A. As long as F depends only on the fields it is sufficient for μ_a to depend on ϕ^A and π^{Aa} only. Then the symmetry of the vacuum functional $Z_m(0)$ with respect to the transformations (49) permits the study of the question whether the mass-dependent terms of the action violate the independence of the S-matrix on the choice of the gauge.

Indeed, let us change the gauge-fixing functional $F(\phi)\rightarrow F(\phi)+\delta F(\phi)$. Then the gauge-fixing term W_F changes according to

$$W_F \rightarrow W_{F+\delta F} = W_F + W_{\delta F}, \quad W_{\delta F} = \delta F(\phi)\left(\tfrac{1}{2}\epsilon_{ab} U^b_m U^a_m + m^2\right). \tag{51}$$

Now, performing in (46) the transformations (49), we choose

$$\mu_a = \mu_a(\phi,\pi) \equiv -(i/\hbar)\,\tfrac{1}{2}\epsilon_{ab}\delta F(\phi) U^b_m.$$

This induces the factor $\exp(\mu_a U^a_m)$ in the integration measure. Combining its exponent with W_F leads to

$$W_F \rightarrow W_F + (\hbar/i)\mu_a U^a_m = W_F - \tfrac{1}{2}\epsilon_{ab}\delta F(\phi) U^b_m U^a_m = W_F - W_{\delta F} + m^2 \delta F(\phi).$$

By comparison with (51) we see that the mass term $m^2 F$ in W_F violates the independence of the vacuum functional $Z_m(0)$ on the choice of the gauge. This result, together with the equivalence theorem,[10] is sufficient to prove that the same is true also for the S-matrix.

One may try to compensate this undesired term $m^2\delta F(\phi)$ by means of an additional change of variables. But this change should not destroy the form of the action arrived at in the previous stage. However, an additional change of variables leads to a Berezinian which is equal to one because σ_α are traceless. Therefore, the unwanted term could never be compensated.

Finally, we shall derive the Ward identities for the extended BRST-and the Sp(2)-symmetries. To begin with, let us introduce the generating functional of the Green's functions:

$$Z_m(J_A;\phi^*_{Aa},\overline{\phi}_A,\eta_A) = \int d\phi^A \exp\{(i/\hbar)(S_{m,\mathrm{ext}}(\phi^A,\phi^*_{Aa},\overline{\phi}_A,\eta_A) + J_A\phi^A)\}. \tag{52}$$

If we multiply Eqs. (35) from the left by $\exp\{(i/\hbar)J_A\phi^A\}$ and integrate over ϕ^A we get

$$\int d\phi^A \exp\{(i/\hbar)J_A\phi^A\}\overline{\Delta}^a_m \exp\{(i/\hbar)S_{m,\mathrm{ext}}(\phi^A,\phi^*_{Aa},\overline{\phi}_A,\eta_A)\} = 0,$$

$$\int d\phi^A \exp\{(i/\hbar)J_A\phi^A\}\overline{\Delta}_\alpha \exp\{(i/\hbar)S_{m,\mathrm{ext}}(\phi^A,\phi^*_{Aa},\overline{\phi}_A,\eta_A)\} = 0.$$

Now, integrating by parts and assuming the integrated expressions to vanish, we can rewrite the resulting equalities with the help of the definition (52) as

$$\left(J_A \frac{\delta}{\delta \phi^*_{Aa}} - V^a_m \right) Z_m = 0, \quad \left((\sigma_\alpha)_B{}^A J_A \frac{\delta}{\delta \eta_B} - V_\alpha \right) Z_m = 0,$$

which are the Ward identities for the generating functional of Green's functions.

Introducing as usual the generating functional of the vertex functions,

$$\Gamma_m(\phi^A; \phi^*_{Aa}, \bar{\phi}_A, \eta_A) = (\hbar/i) \ln Z_m(J_A; \phi^*_{Aa}, \bar{\phi}_A, \eta_A) - J_A \phi^A,$$

$$\phi^A = (\hbar/i) \frac{\delta \ln Z_m(J_A; \phi^*_{Aa}, \bar{\phi}_A, \eta_A)}{\delta J_A},$$

we obtain

$$\tfrac{1}{2}(\Gamma_m, \Gamma_m)^a + V^a_m \Gamma_m = 0, \quad \tfrac{1}{2}\{\Gamma_m, \Gamma_m\}_\alpha + V_\alpha \Gamma_m = 0. \tag{53}$$

For Yang–Mills theories the first identities in (53) are the Slavnov–Taylor identities of the extended BRST symmetries. Furthermore, choosing for σ_α the representation (11), the second identities in (53) express for $\alpha = 0$ the ghost number conservation and, in Yang–Mills theories, for $\alpha = (+,-)$ the Delduc–Sorella identities of the Sp(2)-symmetry.[11]

V. MASSIVE THEORIES WITH A CLOSED GAUGE ALGEBRA

To illustrate the formalism of the osp(1,2)-quantization developed here, we consider irreducible massive gauge theories with a closed algebra. Such theories are characterized by the fact, first, that, because of $M^{ij}_{\alpha_0 \beta_0} = 0$, the algebra of generators, Eq. (2), reduce to

$$R^i_{\alpha_0, j} R^j_{\beta_0} - R^i_{\beta_0, j} R^j_{\alpha_0} = -R^i_{\gamma_0} F^{\gamma_0}_{\alpha_0 \beta_0}; \tag{54}$$

for the sake of simplicity we assume throughout this and the succeeding section that the A^i are bosonic fields, and, second, that $R^i_{\alpha_0}$ has no zero modes (contrary to the case of reducible theories), i.e., any equation of the form $R^i_{\alpha_0} X^{\alpha_0} = 0$ has only the trivial solution $X^{\alpha_0} = 0$. Then, in the case of field-dependent structure functions, the Jacobi identity looks like

$$F^{\delta_0}_{\eta_0 \alpha_0} F^{\eta_0}_{\beta_0 \gamma_0} - R^i_{\alpha_0} F^{\delta_0}_{\beta_0 \gamma_0, i} + \text{cyclic perm}(\alpha_0, \beta_0, \gamma_0) = 0. \tag{55}$$

We shall restrict ourselves to consider only solutions of Eqs. (24) being linear in the antifields ϕ^*_{Aa}, $\bar{\phi}_A$ and η_A. Such solutions can be cast into the form [see Eq. (18)]

$$\bar{S}_m = S_{\text{cl}} + (\tfrac{1}{2} \epsilon_{ab} \mathbf{s}^b_m \mathbf{s}^a_m + m^2) X,$$

with $X = \bar{A}_i A^i + \bar{B}_{\alpha_0} B^{\alpha_0} + \bar{C}_{\alpha_0 c} C^{\alpha_0 c}$. A realization of the (anti)BRST- and the Sp(2)-transformations of the antifields already has been given (see Appendix E). Thus, we are left with the exercise to determine the corresponding transformations for the fields A^i, B^{α_0} and $C^{\alpha_0 c}$. Let us briefly look at the derivation of these transformations. Imposing the osp(1,2)-superalgebra

$$[\mathbf{d}_\alpha, \mathbf{d}_\beta] = \epsilon_{\alpha\beta}{}^\gamma \mathbf{d}_\gamma, \quad [\mathbf{d}_\alpha, \mathbf{s}^a_m] = \mathbf{s}^b_m (\sigma_\alpha)_b{}^a, \quad \{\mathbf{s}^a_m, \mathbf{s}^b_m\} = -m^2 (\sigma^\alpha)^{ab} \mathbf{d}_\alpha, \tag{56}$$

on the gauge fields A^i, due to $\mathbf{d}_\alpha A^i = 0$, this yields $\{\mathbf{s}^a_m, \mathbf{s}^b_m\} A^i = 0$. Then, with

$$\mathbf{s}^a_m A^i = R^i_{\alpha_0} C^{\alpha_0 a}, \tag{57}$$

by virtue of (54), we find

$$R^i_{\alpha_0}(\mathbf{s}^a_m C^{\alpha_0 b} + \mathbf{s}^b_m C^{\alpha_0 a} + F^{\alpha_0}_{\beta_0\gamma_0} C^{\beta_0 a} C^{\gamma_0 b}) = 0.$$

Because the $R^i_{\alpha_0}$ are irreducible the general solution of this equation is given by

$$\mathbf{s}^a_m C^{\alpha_0 b} = \epsilon^{ab} B^{\alpha_0} - \tfrac{1}{2} F^{\alpha_0}_{\beta_0\gamma_0} C^{\beta_0 a} C^{\gamma_0 b}. \tag{58}$$

Imposing the superalgebra (56) on the (anti)ghost fields $C^{\alpha_0 c}$ and taking into account $\mathbf{d}_\alpha C^{\alpha_0 b} = C^{\alpha_0 c}(\sigma_\alpha)_c{}^b$ it gives $\{\mathbf{s}^a_m, \mathbf{s}^b_m\} C^{\alpha_0 c} = -m^2 (\sigma^\alpha)^{ab} C^{\alpha_0 d}(\sigma_\alpha)_d{}^c$. The right-hand side of this restriction can be rewritten by means of the relations $(\sigma_\alpha)_d{}^c = \epsilon_{de}\epsilon^{fc}(\sigma_\alpha)^e{}_f$ and $(\sigma^\alpha)^{ab}(\sigma_\alpha)^e{}_f = -(\epsilon^{ae}\delta^b_f + \epsilon^{be}\delta^a_f)$ as $\{\mathbf{s}^a_m, \mathbf{s}^b_m\} C^{\alpha_0 c} = -m^2(\epsilon^{ac} C^{\alpha_0 b} + \epsilon^{bc} C^{\alpha_0 a})$. Then, with (58), by virtue of (55), we obtain

$$\begin{aligned}\{\epsilon^{bc}(\mathbf{s}^a_m B^{\alpha_0} + m^2 C^{\alpha_0 a} - \tfrac{1}{2} F^{\alpha_0}_{\beta_0\gamma_0} B^{\beta_0} C^{\gamma_0 a})\\ + \tfrac{1}{4}(F^{\alpha_0}_{\eta_0\beta_0} F^{\eta_0}_{\gamma_0\delta_0} + F^{\alpha_0}_{\eta_0\delta_0} F^{\eta_0}_{\beta_0\gamma_0} - 2R^i_{\beta_0} F^{\alpha_0}_{\gamma_0\delta_0,i}) C^{\beta_0 a} C^{\gamma_0 b} C^{\delta_0 c}\} + \mathrm{sym}(a \leftrightarrow b) = 0,\end{aligned}$$

where $\mathrm{sym}(a \leftrightarrow b)$ means symmetrization with respect to the indices a and b.

The general solution of this equation reads

$$\mathbf{s}^a_m B^{\alpha_0} = -m^2 C^{\alpha_0 a} + \tfrac{1}{2} F^{\alpha_0}_{\beta_0\gamma_0} B^{\beta_0} C^{\gamma_0 a} + \tfrac{1}{12}\epsilon_{cd}(F^{\alpha_0}_{\eta_0\beta_0} F^{\eta_0}_{\gamma_0\delta_0} + 2R^i_{\beta_0} F^{\alpha_0}_{\gamma_0\delta_0,i}) C^{\gamma_0 a} C^{\delta_0 c} C^{\beta_0 d}. \tag{59}$$

For the particular case $m=0$ the transformations (57)–(59) were already obtained earlier in Ref. 12.

Note that, due to (56), the only nonzero variation of A^i is of the form

$$\tfrac{1}{2}\epsilon_{ab}\mathbf{s}^b_m \mathbf{s}^a_m A^i = R^i_{\alpha_0} B^{\alpha_0} + \tfrac{1}{2}\epsilon_{ab} R^i_{\alpha_0,j} R^j_{\beta_0} C^{\beta_0 b} C^{\alpha_0 a}; \tag{60}$$

for the corresponding variations of $C^{\alpha_0 c}$ and B^{α_0} one gets

$$\tfrac{1}{2}\epsilon_{ab}\mathbf{s}^b_m \mathbf{s}^a_m C^{\alpha_0 c} = \tfrac{1}{2} m^2 C^{\alpha_0 c} - F^{\alpha_0}_{\beta_0\gamma_0} B^{\beta_0} C^{\gamma_0 c} - \tfrac{1}{6}\epsilon_{ab}(F^{\alpha_0}_{\eta_0\beta_0} F^{\eta_0}_{\gamma_0\delta_0} + 2R^i_{\beta_0} F^{\alpha_0}_{\gamma_0\delta_0,i}) C^{\gamma_0 c} C^{\delta_0 a} C^{\beta_0 b}, \tag{61}$$

$$\tfrac{1}{2}\epsilon_{ab}\mathbf{s}^b_m \mathbf{s}^a_m B^{\alpha_0} = -m^2 B^{\alpha_0}. \tag{62}$$

The relations (55)–(62) specify the transformations of the osp(1,2)-symmetry for gauge theories with a closed algebra. Substituting these expressions into

$$\begin{aligned}\overline{S}_m = S_{\mathrm{cl}} + A^*_{ia}(\mathbf{s}^a_m A^i) + B^*_{\alpha_0 a}(\mathbf{s}^a_m B^{\alpha_0}) - C^*_{\alpha_0 ac}(\mathbf{s}^a_m C^{\alpha_0 c}) + (F_{\alpha_0 c} - \tfrac{1}{2} m^2 \overline{C}_{\alpha_0 c}) C^{\alpha_0 c}\\ + \overline{A}_i(\tfrac{1}{2}\epsilon_{ab}\mathbf{s}^b_m \mathbf{s}^a_m A^i) + \overline{B}_{\alpha_0}(\tfrac{1}{2}\epsilon_{ab}\mathbf{s}^b_m \mathbf{s}^a_m B^{\alpha_0}) + \overline{C}_{\alpha_0 c}(\tfrac{1}{2}\epsilon_{ab}\mathbf{s}^b_m \mathbf{s}^a_m C^{\alpha_0 c}),\end{aligned} \tag{63}$$

$F_{\alpha_0 c}$ being the only nonvanishing component of η_A, a direct verification shows that the resulting action S_m satisfies Eqs. (24) identically.

Finally, let us determine the gauge-fixed action $S_{m,\mathrm{eff}}$ in (40) for the class of minimal gauges F depending only on A^i and $C^{\alpha_0 c}$. Inserting into (46) for S_m the action (63) and performing the integration over antifields and auxiliary fields we are led to the following expression for $S_{m,\mathrm{eff}}$ (at the lowest order of $\hbar$) in the vacuum functional:

$$Z_m(0) = \int dA^i dB^{\alpha_0} dC^{\alpha_0 c} \exp\{(i/\hbar) S_{m,\mathrm{eff}}\}, \quad S_{m,\mathrm{eff}} = S_{\mathrm{cl}} + W_F,$$

where W_F is given by

$$W_F = m^2 F + \frac{1}{2}\epsilon_{ab}\left(\frac{\delta F}{\delta A^i}\mathbf{s}_m^b \mathbf{s}_m^a A^i + \frac{1}{2}\epsilon_{ab}\frac{\delta F}{\delta C^{\alpha_0 c}}\mathbf{s}_m^b \mathbf{s}_m^a C^{\alpha_0 c}\right) - \frac{1}{2}\epsilon_{ab}\left(\mathbf{s}_m^a A^i \frac{\delta^2 F}{\delta A^i \delta A^j}\mathbf{s}_m^b A^j + \mathbf{s}_m^a C^{\alpha_0 c}\frac{\delta^2 F}{\delta C^{\alpha_0 c}\delta C^{\beta_0 d}}\mathbf{s}_m^b C^{\beta_0 d}\right).$$

Again, this gauge-fixing term W_F can be rewritten as

$$W_F = (\tfrac{1}{2}\epsilon_{ab}\mathbf{s}_m^b \mathbf{s}_m^a + m^2)F,$$

showing that the action $S_{m,\mathrm{eff}}$ is, in fact, osp(1,2)-invariant and that the method of gauge fixing proposed in Sec. III will actually lift the degeneracy of the classical gauge-invariant action. As has been shown recently,[13] in the particular case of Yang–Mills theories $S_{m,\mathrm{eff}}$ coincides with the gauge-fixed action in the massive Curci–Ferrari model[14] in the Delbourgo–Jarvis gauge.[15] The classical action S_{YM} is invariant under the non-Abelian gauge transformations

$$\delta A_\mu^\alpha = D_\mu^{\alpha\beta}\theta^\beta(x), \quad D_\mu^{\alpha\beta} \equiv \delta^{\alpha\beta}\partial_\mu - F^{\alpha\beta\gamma}A_\mu^\gamma,$$

where $F^{\alpha\beta\gamma}$ are the totally antisymmetric structure constants. As before, the osp(1,2)-invariance of the gauge-fixed action

$$S_{m,\mathrm{eff}} = S_{\mathrm{YM}} + (\tfrac{1}{2}\epsilon_{ab}\mathbf{s}_m^b \mathbf{s}_m^a + m^2)F, \quad F = \tfrac{1}{2}(A_\mu^\alpha A^{\mu\alpha} + \xi\epsilon_{cd}C^{\alpha c}C^{\alpha d}),$$

where ξ is the gauge parameter, is assured by construction. Using

$$\mathbf{s}_m^a A_\mu^\alpha = D_\mu^{\alpha\beta}C^{\beta a},$$

$$\mathbf{s}_m^a C^{\alpha b} = \epsilon^{ab}B^\alpha - \tfrac{1}{2}F^{\alpha\beta\gamma}C^{\beta a}C^{\gamma b},$$

$$\mathbf{s}_m^a B^\alpha = -m^2 C^{\alpha a} + \tfrac{1}{2}F^{\alpha\beta\gamma}B^\beta C^{\gamma a} + \tfrac{1}{12}\epsilon_{cd}F^{\alpha\eta\beta}F^{\eta\gamma\delta}C^{\gamma a}C^{\delta c}C^{\beta d},$$

for the gauge-fixing terms one gets

$$\tfrac{1}{4}\epsilon_{ab}\mathbf{s}_m^b \mathbf{s}_m^a (A_\mu^\alpha A^{\mu\alpha}) = A_\mu^\alpha \partial^\mu B^\alpha + \tfrac{1}{2}\epsilon_{ab}(\partial^\mu C^{\alpha b})D_\mu^{\alpha\beta}C^{\beta a},$$

$$\tfrac{1}{4}\epsilon_{ab}\epsilon_{cd}\mathbf{s}_m^b \mathbf{s}_m^a (C^{\alpha c}C^{\alpha d}) = \tfrac{1}{2}m^2\epsilon_{cd}C^{\alpha c}C^{\alpha d} + B^\alpha B^\alpha - \tfrac{1}{24}\epsilon_{ab}\epsilon_{cd}F^{\eta\alpha\beta}F^{\eta\gamma\delta}C^{\alpha a}C^{\beta c}C^{\gamma b}C^{\delta d}.$$

The elimination of B^α can be performed by Gaussian integration; it provides the gauge-fixing term $1/2\xi^{-1}(\partial^\mu A_\mu^\alpha)^2$ and, in addition, among other interactions quartic (anti)ghost terms. This shows that the degeneracy of the classical action is indeed removed.

Concluding, we shall emphasize that, up to now, we have ignored the important question whether the action (63) is the general solution of Eqs. (24), i.e., being stable against perturbations. Unfortunately, this is not the case, because the fields A^i, B^{α_0}, $C^{\alpha_0 c}$ and the antifields $\bar{A}_i$, $\bar{B}_{\alpha_0}$, $\bar{C}_{\alpha_0 c}$ have the same quantum numbers and hence mix under renormalization. Therefore, we are confronted with the complicated problem how the action (63) must be changed in order to ensure the required stability. To attack this problem one is forced to introduce within (63) the antifields in a nonlinear manner (see Ref. 13); however, then the corresponding altered action cannot be expressed in such simple form as in (63). The solution of that problem has been given for a particular case in Ref. 13; a general proof will be given in another paper.[16]

VI. CONCLUDING REMARKS

We have proved the possibility of a consistent generalization of the Sp(2) quantization scheme based on the orthosymplectic superalgebra osp(1,2). Introducing mass terms into the theory, which do not break the extended BRST symmetry, the quantum master equations of the Sp(2)-symmetry must be satisfied in order to fulfill the corresponding equations of the extended BRST symmetry. To ensure Sp(2)-invariance of the theory also in the massless case, then, besides of the requirement of extended BRST symmetry the action S must be subjected to the requirement of Sp(2)-symmetry explicitly. An open problem is the general proof, analogous to Refs. 1 and 3, of existence theorems, i.e., absence of anomalies of the theory.

Finally, let us give some remarks concerning the renormalization of the theory. In Ref. 3 this problem was discussed rather formally; thereby the standard hypothesis on the existence of a regularization respecting the Ward identities for the extended BRST symmetries had been accepted.

On the other hand, there is a well-established renormalization scheme, incidentally the most general one known until now, namely the Bogoliubov–Parasiŭk–Hepp–Zimmermann–Lowenstein (BPHZL) scheme,[17] which does not need any regularization. Moreover, the renormalized quantum action principles, which sum up all properties of the renormalized perturbation series, were first established using the BPHZL renormalization scheme (they were rederived in different other schemes confirming their character as general theorems in renormalization theory, independent of the renormalization scheme). These theorems are extremely powerful and suffice for discussing the most useful identities among renormalized Green's functions, such as Ward identities or their generalizations describing gauge invariance (e.g., the existence proofs for generating equations established in Refs. 1 and 2 could be proven also by means of the action principles without any reference to a given renormalization and regularization scheme). Furthermore, within the BPHZL scheme it is possible to perform renormalization when we are concerned with massless theories. If massless fields are involved, ultraviolet subtractions would generate drastic spurious infrared divergences. Lowenstein and Zimmermann introduced a more involved subtraction procedure which is free of spurious infrared divergences. In order to avoid these divergences one has to introduce a mass $m^2(s-1)^2$ with s being an additional infrared subtraction parameter. Then, the disease that ultraviolet subtractions lead to nonintegrable infrared singularities is cured by performing ultraviolet subtractions and extra infrared subtractions with respect to $(s-1)$. It had been proven rigorously that $s=1$ defines the renormalized massless theory; i.e., the introduction of the mass $m^2(1-s)^2$ does not lead to a contradictory massive theory. Of course, mass terms softly violate the gauge dependence of the S-matrix so that, after performing renormalization, one may take the massless limit $s=1$.

Of course, this renormalization scheme could be applied also in the Sp(2)-approach, but then the introduction of mass terms would violate the extended BRST symmetries. On the other hand, using the osp(1,2)-approach, it is possible to introduce mass terms (at the stage of gauge fixing) in such a way that the Ward identities of the extended BRST symmetries will not be violated, provided the Ward identity of the Sp(2)-symmetry is fulfilled to all orders of perturbation theory! Besides, we can use the quantum action principles to get rigorous statements without any reference to a given renormalization and regularization scheme.

Obviously, the osp(1,2)-covariant frame of quantization can be formulated within the superfield approach[9] without much effort. On the other hand, the generalization of the formalism to the case of L-stage reducible theories analogous to Ref. 2 will be more involved.

ACKNOWLEDGMENTS

The work of PML is partially supported by the Russian Foundation for Basic Research (RFBR), Project No. 96-02-16017, as well as the grant INTAS, 96-0308, and the grant of the Ministry of General and Professional Education of Russian Federation in the field of Basic Natural Sciences. He also thanks Graduate College ''Quantum Field Theory'' at Leipzig University for hospitality and fruitful cooperation. The authors are very indebted to the referee for helpful and clarifying remarks.

APPENDIX A: SUPERALGEBRA osp(1,2)

The (anti)commutation relations of the superalgebra osp(1,2) in the Cartan–Weyl basis read (see Ref. 18)

$$[L_0,L_\pm]=\pm L_\pm\,,\quad [L_+,L_-]=2L_0\,, \tag{A1}$$

$$[L_0,R_\pm]=\pm\tfrac{1}{2}R_\pm\,,\quad [L_\pm,R_\mp]=-R_\pm\,,\quad [L_\pm,R_\pm]=0, \tag{A2}$$

$$\{R_\pm,R_\pm\}=\pm\tfrac{1}{2}L_\pm\,,\quad \{R_+,R_-\}=\tfrac{1}{2}L_0\,, \tag{A3}$$

and for the fundamental representation these generators are given by

$$L_0=\begin{pmatrix}\frac{1}{2} & 0 & 0\\ 0 & -\frac{1}{2} & 0\\ 0 & 0 & 0\end{pmatrix},\quad L_+=\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\quad L_-=\begin{pmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 0 & 0 & 0\end{pmatrix},$$

$$R_+=\begin{pmatrix}0 & 0 & \frac{1}{2}\\ 0 & 0 & 0\\ 0 & \frac{1}{2} & 0\end{pmatrix},\quad R_-=\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & -\frac{1}{2}\\ \frac{1}{2} & 0 & 0\end{pmatrix}.$$

The superalgebra (8)–(10) is obtained by the following identifications:

$$(\hbar/i)\overline{\Delta}_0=2L_0\,,\quad (\hbar/i)\overline{\Delta}_\pm=L_\pm\,,\quad (\hbar/i)\overline{\Delta}_m^1=2mR_+\,,\quad (\hbar/i)\overline{\Delta}_m^2=2mR_-\,;$$

so the relations (A1)–(A3) may be written instead as follows ($\alpha=0,+,-$):

$$[\overline{\Delta}_\alpha,\overline{\Delta}_\beta]=(i/\hbar)\epsilon_{\alpha\beta}{}^{\gamma}\overline{\Delta}_\gamma\,, \tag{A4}$$

$$[\overline{\Delta}_\alpha,\overline{\Delta}_m^a]=(i/\hbar)\overline{\Delta}_m^b(\sigma_\alpha)_b{}^a, \tag{A5}$$

$$\{\overline{\Delta}_m^a,\overline{\Delta}_m^b\}=(i/\hbar)m^2(\sigma_\alpha)^{ab}\overline{\Delta}^\alpha, \tag{A6}$$

σ_α generating the group of special linear transformations with the sl(2) algebra:

$$\sigma_\alpha\sigma_\beta=g_{\alpha\beta}+\tfrac{1}{2}\epsilon_{\alpha\beta\beta}\sigma^\gamma,\quad \sigma^\alpha=g^{\alpha\beta}\sigma_\beta\,,\quad \mathrm{Tr}(\sigma_\alpha\sigma_{gb})=2g_{\alpha\beta}\,,$$

$$g^{\alpha\beta}=\begin{pmatrix}1 & 0 & 0\\ 0 & 0 & 2\\ 0 & 2 & 0\end{pmatrix},\quad g^{\alpha\gamma}g_{\gamma\beta}=\delta^\alpha_\beta\,,$$

$\epsilon_{\alpha\beta\gamma}$ being the antisymmetric tensor, $\epsilon_{0+-}=1$. Then, from (A5), the following realization of σ_α is obtained,

$$(\sigma_+)_a{}^b=\begin{pmatrix}0 & -1\\ 0 & 0\end{pmatrix},\quad (\sigma_-)_a{}^b=\begin{pmatrix}0 & 0\\ -1 & 0\end{pmatrix},\quad (\sigma_0)_a{}^b=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix},$$

and by raising the first index according to $(\sigma_\alpha)^{ab}=\epsilon^{ac}(\sigma_\alpha)_c{}^b$ we get

$$(\sigma_+)^{ab}=\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix},\quad (\sigma_-)^{ab}=\begin{pmatrix}-1 & 0\\ 0 & 0\end{pmatrix},\quad (\sigma_0)^{ab}=\begin{pmatrix}0 & -1\\ -1 & 0\end{pmatrix}.$$

The quadratic Casimir operator is given by $C_2 = 1/2\epsilon_{ab}\overline{\Delta}_m^b\overline{\Delta}_m^a + m^2\overline{\Delta}^\alpha\overline{\Delta}_\alpha$.

APPENDIX B: PROPERTIES OF THE BRACKETS

From the definitions (25), (26) and (29) it follows

$$\epsilon(\{F,G\}_\alpha) = \epsilon(F) + \epsilon(G), \quad \{F,G\}_\alpha = \{G,F\}_\alpha(-1)^{\epsilon(F)\epsilon(G)},$$

$$\epsilon((F,G)^a) = \epsilon(F) + \epsilon(G) + 1, \quad (F,G)^a = -(G,F)^a(-1)^{(\epsilon(F)+1)(\epsilon(G)+1)},$$

i.e., $\{F,G\}_\alpha((F,G)^a)$ defines an even (odd) graded bracket, and

$$\Delta_\alpha(FG) = (\Delta_\alpha F)G + F(\Delta_\alpha G) + \{F,G\}_\alpha, \tag{B1}$$

$$\Delta^a(FG) = (\Delta^a F)G + F(\Delta^a G)(-1)^{\epsilon(F)} + (F,G)^a(-1)^{\epsilon(F)}, \tag{B2}$$

where the last two relations may be understood as the definitions of the new brackets $\{F,G\}_\alpha$ and the extended antibrackets $(F,G)^a$ introduced in Ref. 1, respectively. Furthermore, it holds

$$\{F,GH\}_\alpha = \{F,G\}_\alpha H + G\{F,H\}_\alpha(-1)^{\epsilon(F)\epsilon(G)},$$

$$(F,GH)^a = (F,G)^a H + G(F,H)^a(-1)^{(\epsilon(F)+1)\epsilon(G)},$$

i.e., both brackets act on the algebra of functions under multiplications.

Let us now briefly state the properties of the new brackets $\{F,G\}_\alpha$ and $(F,G)^a$. Applying the following identities [their validity is verified by means of (29)]

$$[\Delta_\alpha,\Delta_\beta] = 0, \quad \{\Delta^a,\Delta^b\} = 0, \quad [\Delta_\alpha,\Delta^a] = 0 \tag{B3}$$

of the operators Δ^a and Δ_α to a product of two functions FG and making use of (B1) and (B2), one gets

$$\Delta_{[\alpha}\{F,G\}_{\beta]} = \{\Delta_{[\alpha}F,G\}_{\beta]} + \{F,\Delta_{[\alpha}G\}_{\beta]},$$

$$\Delta^{\{a}(F,G)^{b\}} = (\Delta^{\{a}F,G)^{b\}} + (F,\Delta^{\{a}G)^{b\}}(-1)^{\epsilon(F)+1},$$

$$\Delta_\alpha(F,G)^a - \Delta^a\{F,G\}_\alpha(-1)^{\epsilon(F)} = (\Delta_\alpha F,G)^a + (F,\Delta_\alpha G)^a - \{\Delta^a F,G\}_\alpha(-1)^{\epsilon(F)} - \{F,\Delta^a G\}_\alpha,$$

where the square (curly) bracket indicates antisymmetrization (symmetrization) in the indices α and β (a and b), respectively. Next, applying the relations (B3) to a product of three functions FGH by means of simple but cumbersome calculations, one arrives at the following Jacobi identities satisfied by the two brackets:

$$\{\{F,G\}_{[\alpha},H\}_{\beta]}(-1)^{\epsilon(F)\epsilon(H)} + \text{cyclic}(F,G,H) \equiv 0, \tag{B4}$$

$$((F,G)^{\{a},H)^{b\}}(-1)^{(\epsilon(F)+1)(\epsilon(H)+1)} + \text{cyclic}(F,G,H) \equiv 0, \tag{B5}$$

$$(\{(F,G)^a,H\}_\alpha - (\{F,G\}_\alpha,H)^a(-1)^{\epsilon(G)})(-1)^{\epsilon(F)(\epsilon(H)+1)} + \text{cyclic}(F,G,H) \equiv 0. \tag{B6}$$

Furthermore, one needs the operators V_α and V_m^a, Eqs. (27) and (28), to obey the osp(1,2)-superalgebra:

$$[V_\alpha,V_\beta] = \epsilon_{\alpha\beta}{}^\gamma V_\gamma, \quad \{V_m^a,V_m^b\} = -m^2(\sigma^\alpha)^{ab}V_\alpha, \quad [V_\alpha,V_m^a] = V_m^b(\sigma_\alpha)_b{}^a. \tag{B7}$$

Applying the identities [their validity is established by means of Eqs. (27)–(29)]

$$[\Delta_\alpha,V_\beta] + [V_\alpha,\Delta_\beta] = \epsilon_{\alpha\beta}{}^\gamma\Delta_\gamma,$$

$$\{\Delta^a,V_m^b\}+\{V_m^a,\Delta^b\}=-m^2(\sigma^\alpha)^{ab}\Delta_\alpha, \tag{B8}$$
$$[\Delta_\alpha,V_m^a]+[V_\alpha,\Delta^a]=\Delta^b(\sigma_\alpha)_b{}^a$$

to a product of two functions FG, one can establish the following relations which define the action of the operators V_α and V_m^a upon the brackets,

$$V_{[\alpha}\{F,G\}_{\beta]}=\epsilon_{\alpha\beta}{}^\gamma\{F,G\}_\gamma+\{V_{[\alpha}F,G\}_{\beta]}+\{F,V_{[\alpha}G\}_{\beta]},$$
$$V_m^{\{a}(F,G)^{b\}}=-m^2(\sigma^\alpha)^{ab}\{F,G\}_\alpha+(V_m^{\{a}F,G)^{b\}}+(F,V_m^{\{a}G)^{b\}}(-1)^{\epsilon(F)+1},$$
$$V_\alpha(F,G)^a-V_m^a\{F,G\}_\alpha(-1)^{\epsilon(F)}=(F,G)^b(\sigma_\alpha)_b{}^a+(V_\alpha F,G)^a+(F,V_\alpha G)^a$$
$$-\{V_m^aF,G\}_\alpha(-1)^{\epsilon(F)}-\{F,V_m^aG\}_\alpha,$$

where, as before, square (curly) bracket means antisymmetrization (symmetrization) in the indices α and β (a and b), respectively.

Note that the properties (B7) of V_m^a and V_α are inherited by the the operators $\overline{\Delta}_m^a=\Delta^a+(i/\hbar)V_m^a$ and $\overline{\Delta}_\alpha=\Delta_\alpha+(i/\hbar)V_\alpha$ [see Eqs. (8)–(10)], due to the properties (B3) and (B8).

Finally, let us specify $F=G=H\equiv S$ to be any bosonic functional S, $\epsilon(S)=0$. Then, the Jacobi identities simplify into

$$\{\{S,S\}_{[\alpha},S\}_{\beta]}\equiv 0,\quad ((S,S)^{\{a},S)^{b\}}\equiv 0,\quad \{(S,S)^a,S\}_\alpha-(\{S,S\}_\alpha,S)^a\equiv 0. \tag{B9}$$

For the action of the operators Δ_α, Δ^a and V_α, V_m^a upon the brackets $\{S,S\}_\alpha$ and $(S,S)^a$ one becomes

$$\tfrac{1}{2}\Delta_{[\alpha}\{S,S\}_{\beta]}=\{\Delta_{[\alpha}S,S\}_{\beta]},$$
$$\tfrac{1}{2}\Delta^{\{a}(S,S)^{b\}}=(\Delta^{\{a}S,S)^{b\}}, \tag{B10}$$
$$\tfrac{1}{2}(\Delta_\alpha(S,S)^a-\Delta^a\{S,S\}_\alpha)=(\Delta_\alpha S,S)^a-\{\Delta^aS,S\}_\alpha$$

and

$$\tfrac{1}{2}V_{[\alpha}\{S,S\}_{\beta]}=\{V_{[\alpha}S,S\}_{\beta]}+\tfrac{1}{2}\epsilon_{\alpha\beta}{}^\gamma\{S,S\}_\gamma,$$
$$\tfrac{1}{2}V_m^{\{a}(S,S)^{b\}}=(V_m^{\{a}S,S)^{b\}}-\tfrac{1}{2}m^2(\sigma^\alpha)^{ab}\{S,S\}_\alpha, \tag{B11}$$
$$\tfrac{1}{2}(V_\alpha(S,S)^a-V_m^a\{S,S\}_\alpha)=(V_\alpha S,S)^a-\{V_m^aS,S\}_\alpha+\tfrac{1}{2}(S,S)^b(\sigma_\alpha)_b{}^a.$$

APPENDIX C: COMPATIBILITY PROOF FOR GENERATING EQUATIONS

A solution $\overline{S}_m$ of the generating equations (24) at the lowest-order approximation has been constructed in (23):

$$\tfrac{1}{2}(\overline{S}_m,\overline{S}_m)^a+V_m^a\overline{S}_m=0,\quad \tfrac{1}{2}\{\overline{S}_m,\overline{S}_m\}_\alpha+V_\alpha\overline{S}_m=0. \tag{C1}$$

Let us prove that the generalization (30) of these equations at any order of $\hbar$, namely,

$$\tfrac{1}{2}(S_m,S_m)^a+V_m^aS_m+(\hbar/i)\Delta^aS_m=0,\quad \tfrac{1}{2}\{S_m,S_m\}_\alpha+V_\alpha S_m+(\hbar/i)\Delta_\alpha S_m=0, \tag{C2}$$

for the full quantum action

$$S_m = \overline{S_m} + \sum_{n=1}^{\infty} \hbar^n S_{m(n)},$$

is compatible with the algebraic properties of the brackets $(S_m, S_m)^a$ and $\{S_m, S_m\}_\alpha$ (see Appendix B).

Suppose that functionals $S_{m(k)}$, $k \leqslant n$, have been found obeying Eqs. (C2) up to and including the nth order in $\hbar$. Then for the $(n+1)$th order the following equations should be satisfied:

$$\overline{Q}^a_m S_{m(n+1)} = Y^a_{m(n+1)}, \quad \overline{Q}_\alpha S_{m(n+1)} = Y_{\alpha(n+1)}, \tag{C3}$$

where $\overline{Q}^a_m$ and $\overline{Q}_\alpha$ are the lowest-order approximations of the operators (37),

$$\overline{Q}^a_m X \equiv (\overline{S_m}, X)^a + V^a_m X, \quad \overline{Q}_\alpha X \equiv \{\overline{S_m}, X\}_\alpha + V_\alpha X, \tag{C4}$$

X being an arbitrary functional. By virtue of Eqs. (C1) the operators $\overline{Q}^a_m$ and $\overline{Q}_\alpha$ obey the osp(1,2)-superalgebra:

$$[\overline{Q}_\alpha, \overline{Q}_\beta] = \epsilon_{\alpha\beta}{}^\gamma \overline{Q}_\gamma, \quad [\overline{Q}_\alpha, \overline{Q}^a_m] = \overline{Q}^b_m (\sigma_\alpha)_b{}^a, \quad \{\overline{Q}^a_m, \overline{Q}^b_m\} = -m^2 (\sigma^\alpha)^{ab} \overline{Q}_\alpha. \tag{C5}$$

The functionals $Y^a_{m(n+1)}$ and $Y_{\alpha(n+1)}$ in Eqs. (C3) have the following form:

$$Y^a_{m(n+1)} = i\Delta^a S_{m(n)} - \frac{1}{2}\sum_{k=1}^{n} (S_{m(n+1-k)}, S_{m(k)})^a,$$

$$Y_{\alpha(n+1)} = i\Delta_\alpha S_{m(n)} - \frac{1}{2}\sum_{k=1}^{n} \{S_{m(n+1-k)}, S_{m(k)}\}_\alpha.$$

In order to ensure that Eqs. (C3) are compatible with the osp(1,2)-superalgebra (85), it is necessary that the following relations hold:

$$\begin{aligned} \overline{Q}_\alpha Y_{\beta(n+1)} - \overline{Q}_\beta Y_{\alpha(n+1)} &= \epsilon_{\alpha\beta}{}^\gamma Y_{\gamma(n+1)}, \\ \overline{Q}^a_m Y^b_{m(n+1)} + \overline{Q}^b_m Y^a_{m(n+1)} &= -m^2 (\sigma^\alpha)^{ab} Y_{\alpha(n+1)}, \\ \overline{Q}_\alpha Y^a_{m(n+1)} - \overline{Q}^a_m Y_{\alpha(n+1)} &= Y^b_{m(n+1)} (\sigma_\alpha)_b{}^a. \end{aligned} \tag{C6}$$

Let us show that the relations (C6) indeed are satisfied. To begin with, we consider the Jacobi identity $\{S_m, \{S_m, S_m\}_{[\alpha}\}_{\beta]} \equiv 0$, Eqs. (B9). By virtue of the algebraic properties (B10) and (B11) of $\{S_m, S_m\}_\alpha$, the left-hand side of this identity can be rewritten as

$$\begin{aligned} \{S_m, \tfrac{1}{2}\{S_m, S_m\}_{[\alpha}\}_{\beta]} &\equiv (S_m, \tfrac{1}{2}\{S_m, S_m\}_{[\alpha} + V_{[\alpha} S_m + (\hbar/i)\Delta_{[\alpha} S_m\}_{\beta]} - \{S_m, V_{[\alpha} S_m + (\hbar/i)\Delta_{[\alpha} S_m\}_{\beta]} \\ &= \{S_m, \tfrac{1}{2}\{S_m, S_m\}_{[\alpha} + V_{[\alpha} S_m + (\hbar/i)\Delta_{[\alpha} S_m\}_{\beta]} \\ &\quad - V_{[\alpha}(\tfrac{1}{2}\{S_m, S_m\}_{\beta]} + V_{\beta]} S_m + (\hbar/i)\Delta_{\beta]} S_m) - (\hbar/i)\Delta_{[\alpha}(\tfrac{1}{2}\{S_m, S_m\}_{\beta]} \\ &\quad + V_{\beta]} S_m + (\hbar/i)\Delta_{\beta]} S_m) + \epsilon_{\alpha\beta}{}^\gamma (\tfrac{1}{2}\{S_m, S_m\}_\gamma + V_\gamma S_m + (\hbar/i)\Delta_\gamma S_m). \end{aligned} \tag{C7}$$

Considering the generating equation (C2) in the $(n+1)$th order in $\hbar$,

$$\tfrac{1}{2}\{S_m, S_m\}_\alpha + V_\alpha S_m + (\hbar/i)\Delta_\alpha S_m = \hbar^{n+1}(\overline{Q}_\alpha S_{m(n+1)} - Y_{\alpha(n+1)}) + O(\hbar^{n+2}),$$

then for the equality (C7) in the $(n+1)$th order one gets

$$\overline{Q}_{[\beta}(\overline{Q}_{\alpha]}S_{m(n+1)}-Y_{\alpha](n+1)})+\epsilon_{\alpha\beta}{}^{\gamma}(\overline{Q}_{\gamma}S_{m(n+1)}-Y_{\gamma(n+1)})=0, \tag{C8}$$

where the definition (C4) of the operator $\overline{Q}_\alpha$ has been used. Taking into account the properties (C5) of $\overline{Q}_\alpha$ we conclude from (C8) that the first relation (C6) is fulfilled. In the same way, using the Jacobi identities $(S_m,(S_m,S_m)^{\{a})^{b\}}\equiv 0$ and $\{S_m,(S_m,S_m)^a\}_\alpha-(S_m,\{S_m,S_m\}_\alpha)^a\equiv 0$, Eqs. (B9), it can be verified that also the remaining relations (C6) are fulfilled, thus establishing the compatibility of Eqs. (C3). Obviously, by virtue of (C5), the equations (C6) are solved by

$$Y^a_{m(n+1)}=\overline{Q}^a_m X_{m(n+1)}, \quad Y_{\alpha(n+1)}=\overline{Q}_\alpha X_{m(n+1)}, \tag{C9}$$

with arbitrary functional $X_{m(n+1)}$. Here, we do not check that any solution of (C6) has the form (C9), which means that the theory is anomaly free; in order to check that, one can adopt the methods of Refs. 1 and 3. Choosing $S_{m(n+1)}=X_{m(n+1)}$ the generating equations (C2) are satisfied up to and including the $(n+1)$th order in $\hbar$ and we can proceed by induction.

APPENDIX D: NATURAL TRANSFORMATIONS OF SOLUTIONS OF THE GENERATING EQUATIONS

In this Appendix we study transformations that allow one to consider the characteristic arbitrariness of a solution of the generating equations. These transformations convert any local solution S_m of Eqs. (30) into another local solution $\tilde{S}_m$ of the same equations. For that reason let us introduce an interpolating functional $S_m(\zeta)$, $0\leqslant\zeta\leqslant 1$, where $S_m(1)\equiv\tilde{S}_m$ is the solution we are looking for and where $S_m(0)\equiv S_m$. This functional is required to satisfy the generating equations

$$\tfrac{1}{2}(S_m(\zeta),S_m(\zeta))^a-i\hbar\overline{\Delta}^a_m S_m(\zeta)=0, \quad \tfrac{1}{2}\{S_m(\zeta),S_m(\zeta)\}_\alpha-i\hbar\overline{\Delta}_\alpha S_m(\zeta)=0. \tag{D1}$$

Differentiating Eqs. (90) with respect to ζ for $\partial S_m(\zeta)/\partial\zeta$ one gets the conditions

$$Q^a_m(\zeta)\frac{\partial}{\partial\zeta}S_m(\zeta)=0, \quad Q_\alpha(\zeta)\frac{\partial}{\partial\zeta}S_m(\zeta)=0, \tag{D2}$$

where the operators $Q^a_m(\zeta)$ and $Q_\alpha(\zeta)$, analogous to (37), are defined by

$$Q^a_m(\zeta)X\equiv(S_m(\zeta),X)^a-i\hbar\overline{\Delta}_m X, \quad Q_\alpha(\zeta)X\equiv\{S_m(\zeta),X\}_\alpha-i\hbar\overline{\Delta}_\alpha X,$$

X being an arbitrary functional. Hence, analogous to (38), we have

$$[Q_\alpha(\zeta),Q_\beta(\zeta)]=\epsilon_{\alpha\beta}{}^{\gamma}Q_\gamma(\zeta),$$

$$[Q_\alpha(\zeta),Q^a_m(\zeta)]=Q^b_m(\zeta)(\sigma_\alpha)_b{}^a,$$

$$\{Q^a_m(\zeta),Q^b_m(\zeta)\}=-m^2(\sigma^\alpha)^{ab}Q_\alpha(\zeta).$$

Let us make for $\partial S_m(\zeta)/\partial\zeta$ the following ansatz:

$$\frac{\partial}{\partial\zeta}S_m(\zeta)=\hat{W}_m(\zeta)Y, \quad \hat{W}_m(\zeta)\equiv\frac{1}{2}\epsilon_{ab}Q^b_m(\zeta)Q^a_m(\zeta)+m^2, \tag{D3}$$

with $Y=Y(\phi^A,\overline{\phi}_A,\phi^*_{Aa},\eta_A)$ an arbitrary local Sp(2)-scalar. Then, by virtue of

$$Q^a_m(\zeta)\hat{W}_m(\zeta)=\tfrac{1}{2}m^2(\sigma_\alpha)^a{}_b Q^b_m(\zeta)Q^\alpha(\zeta), \quad Q_\alpha(\zeta)\hat{W}_m(\zeta)=\hat{W}_m(\zeta)Q_\alpha(\zeta),$$

the (consistency) conditions (D2) are fulfilled provided it holds that

$$Q_\alpha(\zeta)Y=0, \quad (\sigma_\alpha)_B{}^A \frac{\delta Y}{\delta\phi^A}\phi^B+V_\alpha Y=0. \tag{D4}$$

The equations

$$Q_m^a(\zeta)\left(\frac{\delta S_m(\zeta)}{\delta\eta_A}-\phi^A\right)=0, \quad Q_\alpha\left(\frac{\delta S_m(\zeta)}{\delta\eta_A}-\phi^A\right)=0,$$

which are obtained from (D1) after differentiating with respect to η_A, and

$$Q_\alpha(\zeta)Y=(\sigma_\alpha)_B{}^A\left(\frac{\delta S_m(\zeta)}{\delta\phi^A}\frac{\delta Y}{\delta\eta_B}+\frac{\delta Y}{\delta\phi^A}\left(\frac{\delta S_m(\zeta)}{\delta\eta_B}-\phi^B\right)-i\hbar\Delta_\alpha Y\right)=0$$

suggest that the natural solution of the requirement (D3) is [see (39)]

$$\frac{\delta S_m(\zeta)}{\delta\eta_A}=\phi^A, \quad \frac{\delta Y}{\delta\eta_A}=0, \tag{D5}$$

i.e., that $S_m(\zeta)$ is linear in η_A and that $Y=Y(\phi^A,\overline{\phi}_A,\phi^*_{Aa})$ must be independent of η_A. Then, as a consequence of the restrictions (D5), the second equation in (D1) becomes

$$(\sigma_\alpha)_B{}^A\frac{\delta S_m(\zeta)}{\delta\phi^A}\phi^B+V_\alpha S_m(\zeta)=0. \tag{D6}$$

The integration of Eq. (D3) leads to

$$\exp\{(i/\hbar)S_m(\zeta)\}=\hat{U}_m(\zeta Y)\exp\{(i/\hbar)S_m\},$$

$$\hat{U}_m(\zeta Y)=\exp\{(\hbar/i)\hat{T}_m(\zeta Y)\}, \quad \hat{T}_m(\zeta Y)=\tfrac{1}{2}\epsilon_{ab}\{\overline{\Delta}_m^b,[\overline{\Delta}_m^a,\zeta Y]\}+(i/\hbar)^2m^2\zeta Y.$$

From this transformation, by virtue of

$$\left[\overline{\Delta}_m^a,\frac{\delta}{\delta\eta_A}-\phi^A\right]=0, \quad \left[\overline{\Delta}_\alpha,\frac{\delta}{\delta\eta_A}-\phi^A\right]=0, \quad \left[Y,\frac{\delta}{\delta\eta_A}-\phi^A\right]=0,$$

it follows that imposing the conditions (D5) for $\zeta=0$ is sufficient to ensure their validity also for $0<\zeta\leqslant 1$, so that $S_m(\zeta)$ indeed solves the first equation in (D1) and Eq. (D6). Moreover, it is seen that the gauge (12) itself is introduced through the use of a special transformation of this kind, namely by choosing $Y=F(\phi^A)$ and $\zeta=1$.

For the iterative solution of Eq. (D3) in the form of a series expansion in powers of ζ one gets

$$S_m(\zeta)=\sum_{n=0}^{\infty}\zeta^n S_m^{(n)}, \quad S_m^{(0)}\equiv S_m,$$

$$(n+1)S_m^{(n+1)}=\frac{1}{2}\epsilon_{ab}\left\{\sum_{k=0}^{n}(S_m^{(k)},(S_m^{(n-k)},Y)^a)^b+(\hbar/i)(S_m^{(n)},\overline{\Delta}_m^a Y)^b+(\hbar/i)\overline{\Delta}_m^b(S_m^{(n)},Y)^a\right.$$
$$\left.+(\hbar/i)^2\delta_{n,0}\overline{\Delta}_m^b\overline{\Delta}_m^a Y\right\}+\delta_{n,0}m^2Y, \quad n\geqslant 0,$$

which shows that if both S_m and Y are local functions, then $S_m(\zeta)$ is a local function as well.

APPENDIX E: COMPONENTWISE NOTATION OF THE TRANSFORMATIONS (19)

In componentwise notation the extended BRST- and Sp(2)-transformations (19) of the antifields read as follows (D_i and E_{α_0} have been put equal to zero):

$$
\begin{aligned}
&V_m^a \bar{A}_i = \epsilon^{ab} A_{ib}^*, \qquad V_\alpha \bar{A}_i = 0,\\
&V_m^a A_{ib}^* = m^2 \delta_b^a \bar{A}_i, \quad V_\alpha A_{ib}^* = A_{id}^* (\sigma_\alpha)^d{}_b,\\
&V_m^a \bar{B}_{\alpha_0} = \epsilon^{ab} B_{\alpha_0 b}^*, \quad V_\alpha \bar{B}_{\alpha_0} = 0,\\
&V_m^a B_{\alpha_0 b}^* = m^2 \delta_b^a \bar{B}_{\alpha_0}, \quad V_\alpha B_{\alpha_0 b}^* = B_{\alpha_0 d}^* (\sigma_\alpha)^d{}_b,\\
&V_m^a \bar{C}_{\alpha_0 c} = \epsilon^{ab} C_{\alpha_0 bc}^*, \quad V_\alpha \bar{C}_{\alpha_0 c} = \bar{C}_{\alpha_0 d} (\sigma_\alpha)^d{}_c,\\
&V_m^a F_{\alpha_0 c} = m^2 \epsilon^{ab} (C_{\alpha_0 bc}^* - C_{\alpha_0 cb}^*), \quad V_{\alpha_0 c} = F_{\alpha_0 d} (\sigma_\alpha)^d{}_c,\\
&V_m^a C_{\alpha_0 bc}^* = m^2 (\delta_b^a \bar{C}_{\alpha_0 c} + \delta_c^a \bar{C}_{\alpha_0 b}) - \delta_b^a F_{\alpha_0 c}, \quad V_\alpha C_{\alpha_0 bc}^* = C_{\alpha_0 dc}^* (\sigma_\alpha)^d{}_b + C_{\alpha_0 bd}^* (\sigma_\alpha)^d{}_c,
\end{aligned}
\tag{D7}
$$

where the additional antifield $F_{\alpha_0 c}$ has to be introduced in order to obey the osp(1,2)-superalgebra (C7) (see Appendix B). Let us stress that expressing this algebra through operator identities is a stronger restriction than satisfying this algebra by the help of (anti)BRST transformations (D7), which could be also realized without introducing $F_{\alpha_0 c}$, namely by choosing $C_{\alpha_0 ab}^* = C_{\alpha_0 ba}^*$.

The componentwise notation of the operators V_m^a, V_α and Δ^a, Δ_α is given by [see (27)–(29)]

$$
\begin{aligned}
V_m^a =\; & \epsilon^{ab} A_{ib}^* \frac{\delta}{\delta \bar{A}_i} + m^2 \bar{A}_i \frac{\delta}{\delta A_{ia}^*} + \epsilon^{ab} B_{\alpha_0 b}^* \frac{\delta}{\delta \bar{B}^{\alpha_0}} + m^2 \bar{B}^{\alpha_0} \frac{\delta}{\delta B_{\alpha_0 a}^*} + \epsilon^{ab} C_{\alpha_0 bc}^* \frac{\delta}{\delta \bar{C}_{\alpha_0 c}}\\
& + m^2 \bar{C}_{\alpha_0 c} \left(\frac{\delta}{\delta C_{\alpha_0 ac}^*} + \frac{\delta}{\delta C_{\alpha_0 ca}^*} \right) - F_{\alpha_0 c} \frac{\delta}{\delta C_{\alpha_0 ac}^*} + m^2 \epsilon^{ab} (C_{\alpha_0 bc}^* - C_{\alpha_0 cb}^*) \frac{\delta}{\delta F_{\alpha_0 c}},
\end{aligned}
$$

$$
\begin{aligned}
V_\alpha =\; & A_{id}^* (\sigma_\alpha)^d{}_b \frac{\delta}{\delta A_{ib}^*} + B_{\alpha_0 d}^* (\sigma_\alpha)^d{}_b \frac{\delta}{\delta B_{\alpha_0 b}^*} + \bar{C}_{\alpha_0 d} (\sigma_\alpha)^d{}_c \frac{\delta}{\delta \bar{C}_{\alpha_0 c}}\\
& + (C_{\alpha_0 dc}^* (\sigma_\alpha)^d{}_b + C_{\alpha_0 bd}^* (\sigma_\alpha)^d{}_c) \frac{\delta}{\delta C_{\alpha_0 bc}^*} + F_{\alpha_0 d} (\sigma_\alpha)^d{}_c \frac{\delta}{\delta F_{\alpha_0 c}},
\end{aligned}
$$

and

$$
\Delta^a = (-1)^{\epsilon_i} \frac{\delta_L}{\delta A^i} \frac{\delta}{\delta A_{ia}^*} + (-1)^{\epsilon_{\alpha_0}} \frac{\delta_L}{\delta B^{\alpha_0}} \frac{\delta}{\delta B_{\alpha_0 a}^*} + (-1)^{\epsilon_{\alpha_0}+1} \frac{\delta_L}{\delta C^{\alpha_0 c}} \frac{\delta}{\delta C_{\alpha_0 ac}^*},
$$

$$
\Delta_\alpha = (-1)^{\epsilon_{\alpha_0}+1} (\sigma_\alpha)_d{}^c \frac{\delta_L}{\delta C^{\alpha_0 c}} \frac{\delta}{\delta F_{\alpha_0 d}}.
$$

Then, it can be verified that V_m^a, V_α and Δ^a, Δ_α fulfill the (anti)commutation relations (B3), (B7) and (B5) (see Appendix B).

[1] I. A. Batalin, P. M. Lavrov, and I. V. Tyutin, J. Math. Phys. **31**, 1487 (1990).
[2] I. A. Batalin, P. M. Lavrov, and I. V. Tyutin, J. Math. Phys. **32**, 532 (1991).
[3] I. A. Batalin, P. M. Lavrov, and I. V. Tyutin, J. Math. Phys. **32**, 2513 (1991).
[4] I. A. Batalin and G. A. Vilkovisky, Phys. Lett. **102B**, 27 (1981).
[5] I. A. Batalin and G. A. Vilkovisky, Phys. Rev. D **28**, 2567 (1983).
[6] M. Bershadsky and H. Ooguri, Phys. Lett. B **229**, 374 (1989); A. M. Polyakov and A. B. Zamolodchikov, Mod. Phys. Lett. A **3**, 1213 (1988).
[7] I. P. Ennes, J. M. Isidro, and A. V. Ramallo, Int. J. Mod. Phys. A **11**, 2379 (1996).
[8] B. S. De Witt, Dynamical Theory of Groups and Fields (Gordan and Breach, New York, 1965).
[9] P. M. Lavrov, Phys. Lett. B **366**, 160 (1996).
[10] R. E. Kallosh and I. V. Tyutin, Yad. Fiz. **17**, 190 (1973).
[11] F. Delduc and S. P. Sorella, Phys. Lett. B **231**, 408 (1989).
[12] N. Nakanishi and I. Ojima, Z. Phys. C **6**, 155 (1980); F. R. Ore and P. van Nieuwenhuizen, Nucl. Phys. B **204**, 317 (1982); L. Alvarez-Gaume and L. Baulieu, Nucl. Phys. B **212**, 255 (1983).
[13] B. Geyer and D. Mülsch, Nucl. Phys. B (Proc. Suppl.) **56**, 253 (1997).
[14] G. Curci and R. Ferrari, Nuovo Cimento A **32**, 151 (1976), **35**, 1 (1976); Phys. Lett. **63B**, 91 (1976); I. Ojima, Z. Phys. C **13**, 173 (1982).
[15] R. Delbourgo and P. D. Jarvis, J. Math. Phys. **15**, 611 (1982).
[16] B. Geyer and D. Mülsch, in preparation.
[17] J. H. Lowenstein and W. Zimmermann, Nucl. Phys. B **86**, 77 (1975); J. H. Lowenstein, Commun. Math. Phys. **47**, 53 (1976); for a review see O. Piguet and A. Rouet, Phys. Rep. **76**, 1 (1981).
[18] A. Pais and V. Rittenberg, J. Math. Phys. **16**, 2062 (1975); M. Scheunert, W. Nahm, and V. Rittenberg, *ibid.* **18**, 146 (1976); **18**, 155 (1976); F. A. Berezin and V. N. Tolstoy, Commun. Math. Phys. **78**, 409 (1981); see also L. Frappat, P. Sorba, and A. Sciarrino, *Dictionary on Lie Superalgebras*, ENSLAPP-AL-600/96 (hep-th/9607161).
[19] B. Geyer, P. M. Lavrov, and D. Mülsch, *osp(1,2)-covariant quantization of reducible massive gauge theories*, hep-th/9806117.

Constructive inversion of energy trajectories in quantum mechanics

Richard L. Hall
Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Boulevard West, Montréal, Québec H3G 1M8, Canada

(Received 22 May 1998; accepted for publication 30 September 1998)

We suppose that the ground-state eigenvalue $E=F(v)$ of the Schrödinger Hamiltonian $H=-\Delta+vf(x)$ in one dimension is known for all values of the coupling $v>0$. The potential shape $f(x)$ is assumed to be symmetric, bounded below, and monotone increasing for $x>0$. A fast algorithm is devised which allows the potential shape $f(x)$ to be reconstructed from the energy trajectory $F(v)$. Three examples are discussed in detail: a shifted power-potential, the exponential potential, and the sech-squared potential are each reconstructed from their known exact energy trajectories. © *1999 American Institute of Physics.*
[S0022-2488(99)01202-5]

I. INTRODUCTION

This paper is concerned with what may be called ''geometric spectral inversion.'' We suppose that a discrete eigenvalue $E=F(v)$ of the Schrödinger Hamiltonian

$$H=-\Delta+vf(x) \tag{1.1}$$

is known for all sufficiently large values of the coupling parameter $v>0$ and we try to use this data to reconstruct the potential shape f. The usual ''forward'' problem would be: given the potential (shape) $f(x)$, find the energy trajectory $F(v)$; the problem we now consider is the inverse of this $F\rightarrow f$.

This problem must at once be distinguished from the ''inverse problem in the coupling constant'' discussed, for example, by Chadan and Sabatier.[1] In this latter problem, the discrete part of the ''input data'' is a set $\{v_i\}$ of values of the coupling constant that all yield the identical energy eigenvalue E. The index i might typically represent the number of nodes in the corresponding eigenfunction. In contrast, for the problem discussed in the present paper, i is kept fixed and the input data is the graph $(F(v),v)$, where the coupling parameter has any value $v>v_c$, and v_c is the critical value of v for the support of a discrete eigenvalue with i nodes. We shall mainly discuss the bottom of the spectrum $i=0$ in this paper. However, on the basis of results we have obtained for the inversion (IWKB) of the Wentzel–Kramers–Brillouin (WKB) approximation,[2] there is good reason to believe that constructive inversion may also be possible starting from any discrete eigenvalue trajectory $F_i(v)$, $i>0$. In fact, perhaps not surprisingly, IWKB yields better results starting from higher trajectories; moreover, they become asymptotically exact as the eigenvalue index is increased without limit.

By making suitable assumptions concerning the class of potential shapes, theoretical progress has already been made with this inversion problem.[3–5] The most important assumptions that we retain throughout the present paper are that $f(x)$ is symmetric, monotone increasing for $x>0$, and bounded below; consequently the minimum value is $f(0)$. We assume that our spectral data, the energy trajectory $F(v)$, derives from a potential shape $f(x)$ with these features. We have discussed[3] how two potential shapes f_1 and f_2 can cross over and still preserve spectral ordering $F_1<F_2$. It is known[4] that lowest point $f(0)$ of f is given by the limit

$$f(0)=\lim_{v\to\infty}\frac{F(v)}{v}. \tag{1.2}$$

We have proved[4] that a potential shape f has a finite flat portion $(f'(x)=0)$ in its graph starting at $x=0$ if and only if the mean kinetic energy is bounded. That is to say, $s=F(v)-vF'(v)<K$, for some positive number K. More specifically, the size b of this patch can be estimated from F by means of the inequality

$$s\leqslant K\Rightarrow f(x)=f(0),\quad |x|\leqslant b,\quad \text{and}\quad b=\frac{\pi}{2}K^{-1/2}. \tag{1.3}$$

The monotonicity of the potential, which allows us to prove results like this, also yields the

Concentration Lemma:[4]

$$q(v)=\int_{-a}^{a}\psi^2(x,v)dx>\frac{f(a)-F'(v)}{f(a)-f(0)}\to 1,\quad v\to\infty, \tag{1.4}$$

where $\psi(x,v)$ is the normalized eigenfunction satisfying $H\psi=F(v)\psi$. More importantly, perhaps, if $F(v)$ derives from a symmetric monotone potential shape f which is bounded below, then f is *uniquely* determined.[5] The significance of this result can be appreciated more clearly upon consideration of an example. Suppose the bottom of the spectrum of H is given by $F(v)=\sqrt{v}$, what is f? It is well known, of course, that $f(x)=x^2\to F_0(v)=\sqrt{v}$; but are there any others? Are scaling arguments reversible? A possible source of disquiet for anyone who ponders such questions is the uncountable number of (unsymmetric) perturbations[6] of the harmonic oscillator all of which have the identical spectrum to that of the unperturbed oscillator $f(x)=x^2$.

If, in addition to symmetry and monotonicity, we also assume that a potential shape $f_1(x)$ vanishes at infinity and that $f_1(x)$ has area, then a given trajectory function $F_1(v)$ corresponding to $f_1(x)$ can be ''scaled''[5] to a standard form in which the new function $F(v)=\alpha F_1(\beta v)$ corresponds to a potential shape $f(x)$ with area -2 and minimum value $f(0)=-1$. Thus square-well potentials, which of course are completely determined by depth and area, are immediately invertible; moreover it is known that, amongst all standard potentials, the square-well is ''extremal'' for it has the lowest possible energy trajectory. In Ref. 5 an approximate variational inversion method is developed; it is also demonstrated constructively that all separable potentials are invertible. However, these results and additional constraints are not used in the present paper. When a potential has area $2A$, we first assumed, during our early attempts at numerical inversion, that it would be very useful to determine A from $F(v)$ and then appropriately constrain the inversion process. However, the area constraint did not turn out to be useful. Thus the numerical method we have established for constructing $f(x)$ from $F(v)$ does not depend on the use of this constraint, and is therefore not limited to the reconstruction of potentials which vanish at infinity and have an area.

Much of numerical analysis assumes that errors arising from arithmetic computations or from the computation of elementary functions is negligibly small. The errors usually studied in depth are those that arise from the discrete representation of continuous objects such as functions, or from operations on them, such as derivatives or integrals. In this paper we shall take this separation of numerical problems to a higher level. We shall assume that we have a numerical method for solving the eigenvalue problem in the *forward* direction $f(x)\to F(v)$ that is reliable and may be considered for our purposes to be essentially error free. Our main emphasis will be on the design of an effective algorithm for the inverse problem *assuming* that the forward problem is numerically soluble. The forward problem is essential to our methods because we shall need to know not only the given exact energy trajectory $F(v)$ but also, at each stage of the reconstruction, what eigenvalue a partly reconstructed potential generates. This line of thought immediately indicates that we shall also need a way of temporarily extrapolating a partly reconstructed potential to all x.

Our constructive inversion algorithm hinges on the assumed symmetry and monotonicity of $f(x)$. This allows us to start the reconstruction of $f(x)$ at $x=0$, and sequentially increase *x*. In Sec. II it is shown how numerical estimates can be made for the shape of the potential near $x=0$, that is for $x<b$, where *b* is a parameter of the algorithm. In Sec. III we explore the implications of the potential's monotonicity for the ''tail'' of the wave function. In Sec. IV we establish a numerical representation for the form of the unknown potential for $x>b$ and construct our inversion algorithm. In Sec. V the algorithm is applied to three test problems.

II. THE RECONSTRUCTION OF $f(x)$ NEAR $x=0$

Since the energy trajectory $F(v)$ which we are given is assumed to arise from a symmetric monotone potential, and since the spectrum generated by the potential is invariant under shifts along the *x*-axis, we may assume without loss of generality that the minimum value of the potential occurs at $x=0$. We now investigate the behavior of $F(v)$, either analytically or numerically, for large values of v. The purpose is to establish a value for the starting point $x=b>0$ of our inversion algorithm and the shape of the potential in the interval $x\in[0,b]$. First of all, the minimum value $f(0)$ of the potential is provided by the limit (1.2). Now, if the mean kinetic energy $s=(\psi,-\Delta\psi)=F(v)-vF'(v)$ is found to be bounded above by a positive number *K*, then we know[4] that the potential shape $f(x)$ satisfies $f(x)=f(0)$, $x\in[0,b]$, where *b* is given by (1.3). In this case we have a value for *b* and also the shape $f(x)$ inside the interval $[0,b]$.

If the mean potential energy *s* is (or appears numerically to be) unbounded, then we adopt another strategy: we model $f(x)$ as a shifted power potential near $x=0$. Since we never know $f(x)$ *exactly*, we shall need another symbol for the approximation we are currently using for $f(x)$. We choose this to be $g(x)$ and we suppose that the bottom of the spectrum of $-\Delta+vg(x)$ is given by $G(v)$. The goal is to adjust $g(x)$ until $G(v)$ is close to the given $F(v)$. Thus we write

$$f(x)\approx g(x)=f(0)+Ax^q,\quad x\in[0,b]. \tag{2.1}$$

Therefore we have three positive parameters to determine, *b*, *A*, and *q*. We first suppose that $g(x)$ has the form (2.1) for *all* $x\geqslant 0$. We now choose a ''large'' value v_1 of v. This is related to the later choice of *b* by a bootstrap argument; the idea is that we choose v_1 so large that the turning point determined by

$$\psi_{xx}(x,v_1)/\psi(x,v_1)=v_1f(x)-F(v)=0 \tag{2.2}$$

is equal to *b*. The concentration lemma guarantees that this is possible. By scaling arguments we have

$$G(v)=f(0)v+E(q)(vA)^{2/(2+q)}, \tag{2.3}$$

where $E(q)$ is the bottom of the spectrum of the pure-power Hamiltonian $-\Delta+|x|^q$. We now ''fit'' $G(v)$ to $F(v)$ by the equations $G(v_1)=F(v_1)$ and $G(2v_1)=F(2v_1)$ which yield the estimate for *q* given by

$$\eta=\frac{2}{2+q}=\frac{\log(F(2v_1)-2v_1f(0))-\log(F(v_1)-v_1f(0))}{\log(2)}. \tag{2.4}$$

Thus *A* is given by

$$A=((F(v_1)-v_1f(0))/E(q))^{1/\eta}/v_1. \tag{2.5}$$

We choose *b* to be equal to the turning point corresponding to the model potential $g(x)$ with the smaller value of v, that is to say so that $f(0)+Ab^q=F(v_1)/v_1$, or

$$b=\left(\frac{F(v_1)-v_1f(0)}{Av_1}\right)^{1/q}. \tag{2.6}$$

Thus we have determined the three parameters which define the potential model $g(x)$ for $x\in[-b,b]$.

III. THE TAIL OF THE WAVE FUNCTION

Let us suppose that the ground-state wave function is $\psi(x,v)$. Thus the turning point $\psi_{xx}(x,v)=0$ occurs for a given v when

$$x=x_t(v)=f^{-1}(R(v)),\quad R(v)=\left(\frac{F(v)}{v}\right). \tag{3.1}$$

The concentration lemma (1.4) quantifies the tendency of the wave function to become, as the coupling v is increased, progressively more concentrated on the patch $[-c,c]$, where $x=c$ is the point (perhaps zero) where $f(x)$ first starts to increase. This allows us to think in terms of the wave function having a ''tail.'' We think of a symmetric potential as having been determined from $x=0$ up to the current point x. The question we now ask is: what value of v should we use to determine how $f(x)$, or, more particularly, our *approximation* $g(x)$ for $f(x)$, continues beyond the current point. We have found that a good choice is to choose v so that the turning point $x_t(v)=x/2$, or some other similar fixed fraction $\sigma<1$ of the current x value. The algorithm seems to be insensitive to this choice. Since $g(x)$ has been constructed up to the current point, and $F(v)$ is known, the value of v required follows by inverting (3.1). It has been proven[4] that $R(v)$ is monotone and therefore invertible. Hence we have the following general recipe for v:

$$v=R^{-1}(g(\sigma x)),\quad \sigma=\tfrac{1}{2}. \tag{3.2}$$

Since we can only determine Schrödinger eigenvalues of $H=-\Delta+vg(x)$ if the potential is defined for all x, we must have a policy about temporarily extending $g(x)$. We have tried many possibilities and found the simplest and most effective method is to extend $g(x)$ in a straight line, with the slope to be determined.

In Fig. 1 we illustrate the ideas just discussed for the case of the sech-squared potential. The inset graph shows the sech-squared potential perturbed from $x=x_a$ by five straight line extensions; meanwhile the main graph shows the corresponding set of five wave functions which agree for $0\leqslant x\leqslant x_a$ and then continue with different ''tails'' dictated by the corresponding potential extensions. The value of the coupling v is the value that makes the turning point of the wave function occur at $x=x_a/2$. This figure illustrates the sort of graphical study that has lead to the algorithm described in this paper.

IV. THE INVERSION ALGORITHM

We must first define the ''current'' approximation $g(x)$ for the potential $f(x)$ sought. For values of x less than b, $g(x)$ is defined either as the horizontal line $f(x)=f(0)$ or as the shifted power potential (2.1). For values of x greater than b, the x-axis is divided into steps of length h. Thus the ''current'' value of x would be of the form $x=x_k=b+kh$, where k is a positive integer. The idea is that $g(x_k)$ is determined sequentially and $g(x)$ is interpolated linearly between the x_k points. We suppose that $\{g(x_k)\}$ have already been determined up to k and we need to find $y=g(x_{k+1})$. For $x\geqslant x_k$ we let

$$g(x)=g(x_k)+(y-g(x_k))\frac{x-x_k}{h}. \tag{4.1}$$

If, from a study of $F(v)$, the underlying potential $f(x)$ has been shown[5] to be bounded above, it is convenient to rescale $F(v)$ so that it corresponds to a potential shape $f(x)$ which vanishes at

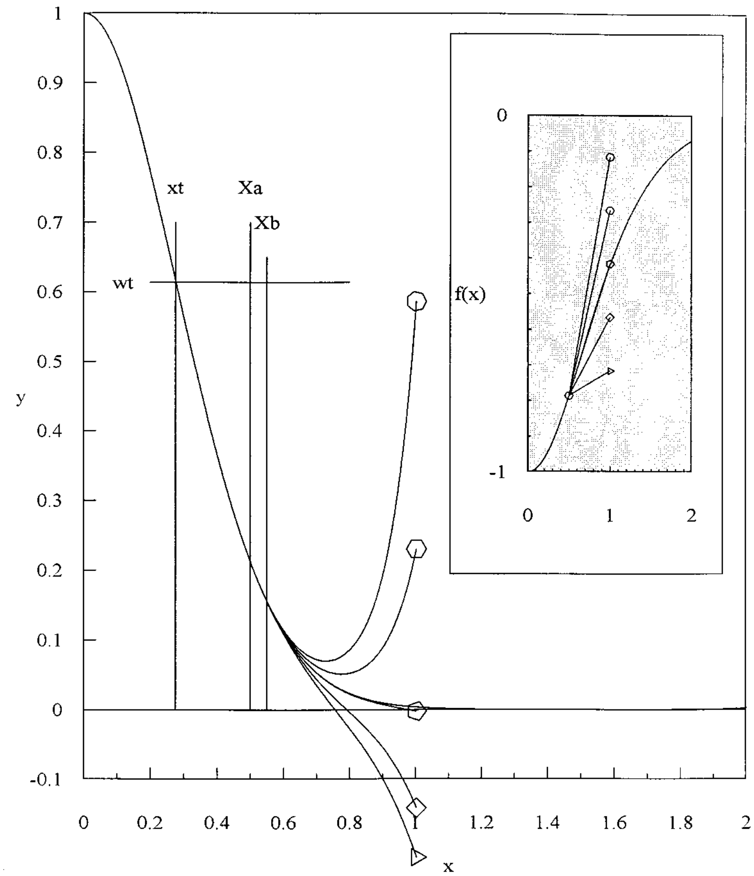

FIG. 1. The potential $f(x) = -\text{sech}^2(x)$ is perturbed from $x = x_a$ by straight-line segments. Each segment leads to a perturbation in the tail of the corresponding wave function. The coupling v is chosen so that $x_a = x_t/2$, where x_t is the turning point of the wave function.

infinity. In this case it is slightly more efficient to modify (4.1) so that for large x the straight-line extrapolation of $g(x)$ is ''cut'' to zero instead of becoming positive. In either case we now have for the current point x_k an approximate potential $g(x)$ parametrized by the ''next'' value $y = g(x_{k+1})$. The task of the inversion algorithm is simply to choose this value of y.

Let us suppose that, for given values of k and y, the bottom of the spectrum of $H = -\Delta + vg(x)$ is given by $G(v,k,y)$, then the inversion algorithm may be stated in the following succinct form in which $\sigma < 1$ is a fixed parameter. Find y such that

$$vg(\sigma x_k) = F(v) = G(v,k,y); \quad \text{then } g(x_{k+1}) = y. \tag{4.2}$$

The value of v is first chosen so that the turning point of the wave function generated by g occurs at σx_k; after this, the value of y is chosen so that G ''fits'' F for this value of v. The value of the parameter σ chosen for the examples discussed in Sec. V below is $\sigma = \frac{1}{2}$. The idea behind this choice can best be understood from a study of Fig. 1. The value of the coupling v must be such that the current value of x for which y is sought is in the ''tail'' of the corresponding wave function; that is to say, the turning point σx should be before x, but not too far away. Fortunately the inversion algorithm seems to be insensitive to the choice of σ.

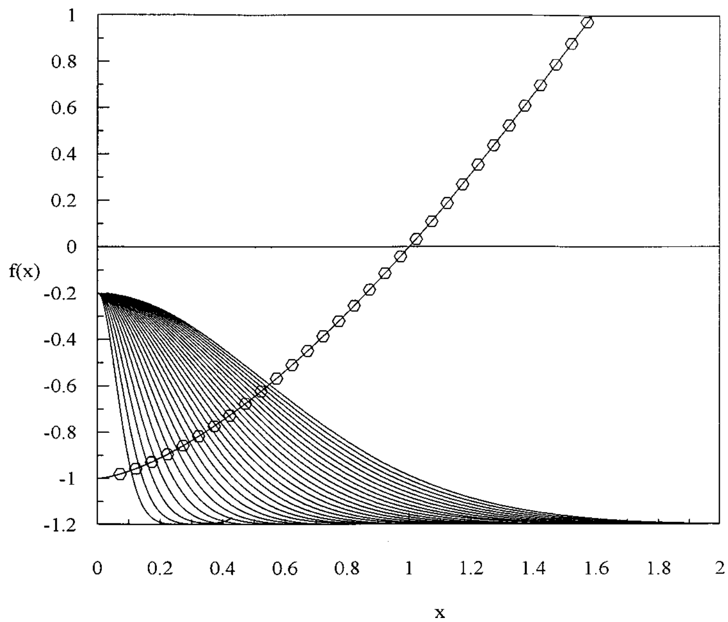

FIG. 2. Constructive inversion of the energy trajectory $F(v)$ for the shifted power potential $f(x)=-1+|x|^{3/2}$. For $x \leqslant b=0.072$, the algorithm correctly generates the model $f(x)$; for larger values of x, in steps of size $h=0.05$, the hexagons indicate the reconstructed values for the potential $f(x)$, shown exactly as a smooth curve. The unnormalized wave functions are also shown.

V. THREE EXAMPLES

The first example we consider is the unbounded potential whose shape $f(x)$ and corresponding exact energy trajectory $F(v)$ are given by the $\{f,F\}$ pair

$$f(x)=-1+|x|^{3/2} \leftrightarrow F(v)=-v+E(3/2)v^{4/7}, \tag{5.1}$$

where $E(3/2)$ is the bottom of the spectrum of $H=-\Delta+|x|^{3/2}$ and has the approximate value $E(3/2)\approx 1.001\,184$. Applying the inversion algorithm to $F(v)$ we obtain the reconstructed potential shown in Fig. 2. We first set $v_1=10^4$ and find that the initial shape is determined (as described in Sec. II) to be $-1+x^{1.5}$ for $x<b=0.072$. For larger values of x the step size is chosen to be $h=0.05$ and 40 iterations are performed by the inversion algorithm. The results are plotted as hexagons on top of the exact potential shape shown as a smooth curve. This entire computation takes less than 20 s with a program written in C++ running on a 200 MHz Pentium Pro.

The following two examples are bounded potentials both having a large-x limit zero, lowest point $f(0)=-1$, and area -2. The exponential potential[7,8] has the $\{f,F\}$ pair

$$f(x)=-e^{-|x|} \leftrightarrow J'_{2|E|^{1/2}}(2v^{1/2})=0 \equiv E=F(v), \tag{5.2}$$

where $J'_\nu(x)$ is the derivative of the Bessel function of the first kind of order ν. For the sech-squared potential[8] we have

$$f(x)=-\operatorname{sech}^2(x) \leftrightarrow F(v)=-[(v+\tfrac{1}{4})^{1/2}-\tfrac{1}{2}]^2. \tag{5.3}$$

In Fig. 3 the two energy trajectories are plotted. Since the two *potentials* have the lowest value -1 and area -2 it follows[5] that the corresponding trajectories both have the form $F(v)\approx -v^2$ for small v and they both satisfy the large-v limit $\lim_{v\to\infty}(F(v)/v)=-1$. Thus the differences between the potential shapes is somehow encoded in the fine differences between these two similar energy curves for intermediate values of v; It is the task of our inversion theory to decode this

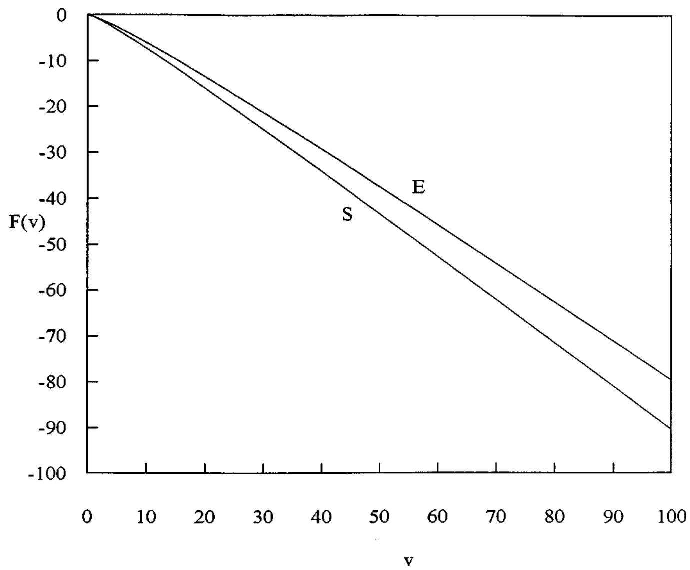

FIG. 3. The ground-state energy trajectories $F(v)$ for the exponential potential (E) and the sech-squared potential (S). For small v, $F(v) \approx -v^2$; for large v, $\lim_{v\to\infty}(F(v)/v) = -1$. The shapes of the underlying potentials are buried in the details of $F(v)$ for intermediate values of v.

information and reveal the underlying potential shape. If we apply the inversion algorithm to these two problems we obtain the results shown in Figs. 4 and 5. The parameters used are exactly the same as for the first problem described above. The time taken to perform the inversions is again less than 20 s if we discount, in the case of the exponential potential, the extra time taken to compute $F(v)$ itself.

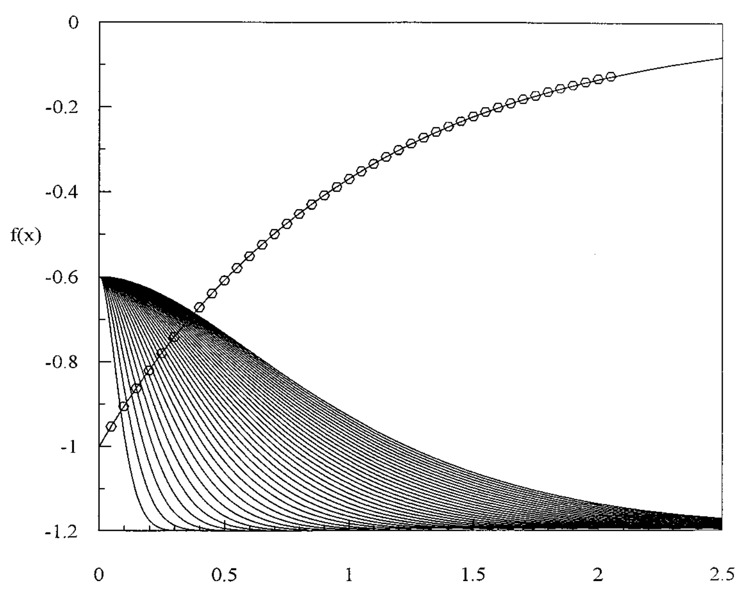

FIG. 4. Constructive inversion of the energy trajectory $F(v)$ for the exponential potential $f(x) = -\exp(x)$. For $x \leq b = 0.048$, the algorithm correctly generates the model $f(x) = -1 + |x|$; for larger values of x, in steps of size $h = 0.05$, the hexagons indicate the reconstructed values for the potential $f(x)$, shown exactly as a smooth curve. The unnormalized wave functions are also shown.

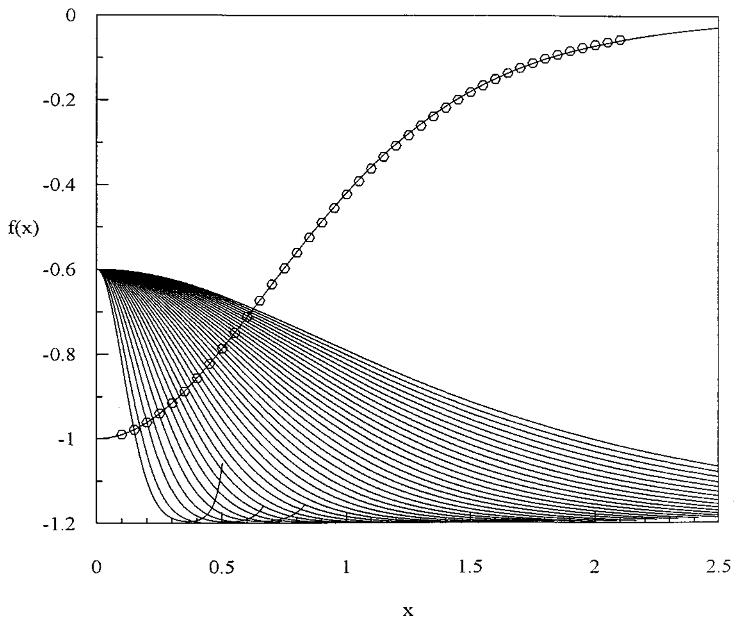

FIG. 5. Constructive inversion of the energy trajectory $F(v)$ for the sech-squared potential $f(x)=-\mathrm{sech}^2(x)$. For $x\leq b=0.1$, the algorithm correctly generates the model $f(x)=-1+x^2$; for larger values of x, in steps of size $h=0.05$, the hexagons indicate the reconstructed values for the potential $f(x)$, shown exactly as a smooth curve. The unnormalized wave functions are also shown.

VI. CONCLUSION

Once we suspect (or know) that an energy trajectory $F(v)$ is derived from a potential shape $f(x)$, it is certainly possible in principle to model the potential discretely as $g(x)$ and then find g approximately by a least-squares fit of $G(v)$ to $F(v)$. Such a ''brute force'' method would not be easy or fast, even for problems in one dimension. In terms of the reconstructions presented in this paper, one would have to consider minimizing a function of the form $\Sigma_{i=1}^{40}|G(v_i;\mathbf{Y})-F(v_i)|^2$, where the vector $\mathbf{Y}$ represents the 40 values of $g(x_k)$ to be determined. We have found that such a function of $\mathbf{Y}$ has very erratic behavior unless the starting point can be chosen quite close to the critical point.

The purpose of the approach discussed in this paper is however not so much to do with efficiency as with understanding. The method we have found is intimately linked to the basic properties of the problem; the implications of monotonicity, the relation between the position of the turning point of the wave function and the value of v, and the tail behavior. The effectiveness of the resulting algorithm stems from its systematic use of all this information. If a potential shape $f(x)$ is symmetric but *not* monotonic (on the half-axis), then for large values of the coupling v the problem will necessarily split into regimes that become more and more isolated as v increases. The situation could become arbitrarily complicated, perhaps involving resonances, and we have no idea at present whether reconstruction $F\rightarrow f$ would in principle be possible in the general case.

If the potential were unimodal and monotonic away from the minimum point, we do not at present know what might be the spectral inheritance of the additional property of the symmetry of $f(x)$. Is there nonuniqueness in this case? Could a symmetric potential be constructed that would have the same energy trajectory $F(v)$ as that of a given nonsymmetrical unimodal potential shape $f(x)$? Many interesting questions such as this which are simple to pose nevertheless appear at present to be very difficult to answer.

In our earlier papers on this topic we discussed some suggestions for applications of this form of spectral inversion. The situations that are most strongly suggestive are those such as the screened-Coulomb potentials used in atomic physics where the coupling varies with the atomic number. In such a case $F_n(v)$ or, more accurately, pair *differences* between such functions, would only be known at certain isolated points. Now that an effective form of constructive inversion is available, it will be possible to consider this more physically important type of application. Another approach which has not yet been applied to geometric spectral inversion is via control theory. Rabitz *et al.*[9,10] have successfully used ideas from control theory to reconstruct molecular potentials from sets of data that are directly measurable. This is the ultimate goal of the present work on geometric spectral inversion.

ACKNOWLEDGMENT

Partial financial support of this work under Grant No. GP3438 from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

[1] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1989). The ''inverse problem in the coupling constant'' is discussed on p. 406.
[2] R. L. Hall, Phys. Rev. A **51**, 1787 (1995).
[3] R. L. Hall, J. Phys. A **25**, 4459 (1992).
[4] R. L. Hall, Phys. Rev. A **50**, 2876 (1995).
[5] R. L. Hall, J. Phys. A **28**, 1771 (1995).
[6] O. L. De Lange and R. E. Raab, *Operator Methods in Quantum Mechanics* (Oxford University Press, Oxford, 1991). Perturbed harmonic oscillators with identical spectra to that generated by $f(x)=x^2$ are given on p. 71.
[7] H. S. W. Massey and C. B. O. Mohr, Proc. R. Soc. London **148**, 206 (1934).
[8] S. Flügge, *Practical Quantum Mechanics* (Springer, New York, 1974). The exponential potential is discussed on p. 196 and the sech-squared potential on p. 94.
[9] P. Gross, H. Singh, and H. Rabitz, Phys. Rev. A **47**, 4593 (1993).
[10] Z. Lu and H. Rabitz, Phys. Rev. A **52**, 1961 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Superintegrability in three-dimensional Euclidean space

E. G. Kalnins and G. C. Williams
Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand

W. Miller, Jr.
School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455

G. S. Pogosyan
Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, Dubna, Moscow Region 141980, Russia

(Received 19 May 1998; accepted for publication 1 September 1998)

Potentials for which the corresponding Schrödinger equation is maximally superintegrable in three-dimensional Euclidean space are studied. The quadratic algebra which is associated with each of these potentials is constructed and the bound state wave functions are computed in the separable coordinates. © *1999 American Institute of Physics.* [S0022-2488(99)02602-X]

I. INTRODUCTION

The present paper continues our study of the systems with *hidden symmetry* or so-called *superintegrable* systems in spaces with constant curvature.

The best known systems of this kind in three-dimensional Euclidean space are the harmonic oscillator and Kepler–Coulomb problems, which have many special properties distinct from other spherically symmetric potentials. These include the phenomena of separation of variables for the Hamilton–Jacobi and Schrödinger equations in more than one orthogonal coordinate system and the existence of integrals of motion in addition to the total angular momentum $\mathbf{L}^2$. In particular for the isotropic oscillator there is the Demkov tensor $D_{ik}=p_ip_k+\omega^2 x_i x_k$,[1] and, in the case of the Kepler–Coulomb problem, the Pauli–Runge–Lenz vector $\mathbf{A}=1/2([\mathbf{L}\times\mathbf{p}]-[\mathbf{p}\times\mathbf{L}])-\mathbf{r}/|r|$. Both these systems possess five functionally independent integrals of motion.[2,3] The first systematic search for all potentials for which the Schrödinger equation admits separation of variables in two or more coordinate systems was begun by Smorodinsky and Winternitz with co-workers in Refs. 4–6 and continued by Evans in Refs. 3 and 7. They found all such systems in two- and three-dimensional flat space and introduced the notion of *superintegrability*. In general, a physical system in N dimensions is called *minimally* superintegrable if it has $2N-2$ integrals of motion, and *maximally* superintegrable if it has $2N-1$ integral of motions. There are five known maximally (and some minimally) superintegrable potentials listed in Refs. 3, 8, and 10 and investigated from different points of view in the last decade.[8–13] Note also that superintegrable potentials in spaces of constant curvature were introduced in Refs. 14–16.

In previous articles[17–19] we have looked at potentials in two-dimensional Euclidean space and the two-dimensional sphere and hyperboloid, for which the Schrödinger equation is maximally superintegrable. In this article we extend this study to the case of three-dimensional Euclidean space. As previously seen in the case of two dimensions, some of these potentials (see Table I) admit bound state or finite solutions and it is these to which we draw attention in this article.

The basic equation that we investigate is of course the Schrödinger equation ($\hbar=m=1$)

$$\mathcal{H}\Psi=-\frac{1}{2}\Delta\Psi+V(x,y,z)\Psi=-\frac{1}{2}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\Psi+V(x,y,z)\Psi=E\Psi. \tag{1}$$

The idea is to find solutions of this equation via a separation of variables ansatz

TABLE I. The three-dimensional maximally superintegrable potentials.

Potential $V(x,y,z)$	Separating coordinates
$V_1=\frac{\omega^2}{2}(x^2+y^2+z^2)+\left(\frac{k_1^2-\frac{1}{4}}{x^2}+\frac{k_2^2-\frac{1}{4}}{y^2}+\frac{k_3^2-\frac{1}{4}}{z^2}\right)$	Cartesian Spherical Cylindrical polar Cylindrical elliptic Sphero-conical Oblate spheroidal Prolate spheroidal Ellipsoidal
$V_2=\frac{\omega^2}{2}(x^2+y^2+4z^2)+\frac{1}{2}\left(\frac{k_1^2-\frac{1}{4}}{x^2}+\frac{k_2^2-\frac{1}{4}}{y^2}\right)$	Cartesian Cylindrical polar Cylindrical parabolic Cylindrical elliptic Parabolic
$V_3=-\frac{\alpha}{\sqrt{x^2+y^2+z^2}}+\frac{1}{2}\left(\frac{k_1^2-\frac{1}{4}}{x^2}+\frac{k_2^2-\frac{1}{4}}{y^2}\right)$	Spheroidal-conical Spherical Parabolic Prolate spheroidal II

$$\Psi=\prod_{j=1}^{3}\psi_j(u_j)$$

for some suitable orthogonal coordinates u_j (see Table II).

In Secs. II–IV we consider three maximally superintegrable potentials (see Table I) and use the Niven-type (or Bethe[20]) ansatz for constructing the solution of the Schrödinger equation in coordinates such as spheroidal, sphero-conical, and ellipsoidal (see Table II). In addition we discuss the extension to the quadratic algebras that were in evidence in the case of two dimensions and see what their implications may be.

Section V is devoted to the calculation of interbasis expansion coefficients for the V_3 potential between spherical and parabolic bases.

II. GENERALIZED ISOTROPIC OSCILLATOR

The first potential (see Table I) on our list of three is

$$V_1(x,y,z)=\frac{\omega^2}{2}(x^2+y^2+z^2)+\frac{1}{2}\left[\frac{(k_1^2-\frac{1}{4})}{x^2}+\frac{(k_2^2-\frac{1}{4})}{y^2}+\frac{(k_3^2-\frac{1}{4})}{z^2}\right], \tag{2}$$

where the constant $k_i\geqslant\frac{1}{2}$. For $k_i=\frac{1}{2}$ we have the ordinary isotropic oscillator potential. The corresponding Schrödinger equation admits solutions via a separation of variables in eight coordinate systems: Cartesian, spherical, sphero-conical, cylindrical polar, cylindrical elliptic, prolate and oblate spheroidal, and ellipsoidal. We summarize the bound state solutions in each case.

Before considering various coordinate systems we note that a basis for the symmetries of Schrödinger's equation with the potential (2) consists of the six operators:

TABLE II. Systems of coordinate in three-dimensional Euclidean space.

Coordinate system	Coordinates
I. Cartesian $x,y,z\in\mathbf{R}$	x,y,z
II. Cylindrical polar $\rho>0,\ \varphi\in[0,2\pi)$	$x=\rho\cos\varphi,\ y=\rho\sin\varphi,\ z$
III. Cylindrical elliptic $z\in\mathbf{R},\ e_1<\mu_1<e_2<\mu_2$	$x^2=\frac{(\mu_1-e_1)(\mu_2-e_1)}{(e_2-e_1)},\quad y^2=\frac{(\mu_1-e_2)(\mu_2-e_2)}{(e_1-e_2)},z$
IV. Cylindrical parabolic $\xi,x\in\mathbf{R},\ \eta\geqslant 0$	$x,y=\xi\eta,\ z=\frac{1}{2}(\xi^2-\eta^2)$
V. Spherical $r>0,\ \theta\in[0,\pi],\ \varphi\in[0,2\pi)$	$x=r\cos\theta\cos\varphi,\ y=r\sin\theta\sin\varphi,\ z=r\cos\theta$
VI. Prolate spheroidal $e_1<u_1<e_2<u_2,\ \varphi\in[0,2\pi)$	$x^2=\frac{(u_1-e_2)(u_2-e_2)}{(e_1-e_2)}\cos^2\varphi,$ $y^2=\frac{(u_1-e_2)(u_2-e_2)}{(e_1-e_2)}\sin^2\varphi,$ $z^2=\frac{(u_1-e_1)(u_2-e_1)}{(e_2-e_1)}$
VII. Oblate spheroidal $e_1<u_1<e_2<u_2,\ \varphi\in[0,2\pi)$	$x^2=\frac{(u_1-e_1)(u_2-e_1)}{(e_2-e_1)}\cos^2\varphi,$ $y^2=\frac{(u_1-e_1)(u_2-e_1)}{(e_2-e_1)}\sin^2\varphi,$ $z^2=\frac{(u_1-e_2)(u_2-e_2)}{(e_1-e_2)}$
VIII. Sphero-conical $r\geqslant 0,\ e_1<\rho_1<e_2<\rho_2<e_3$	$x^2=r^2\frac{(\rho_1-e_1)(\rho_2-e_1)}{(e_1-e_2)(e_1-e_3)},$ $y^2=r^2\frac{(\rho_1-e_2)(\rho_2-e_2)}{(e_2-e_1)(e_2-e_3)},$ $z^2=r^2\frac{(\rho_1-e_3)(\rho_2-e_3)}{(e_3-e_2)(e_3-e_1)}$
IX. Parabolic $\xi,\eta\geqslant 0,\ \varphi\in[0,2\pi)$	$x=\xi\eta\cos\varphi,\ y=\xi\eta\sin\varphi,\ z=\frac{1}{2}(\xi^2-\eta^2)$
X. Ellipsoidal $a_1<u_1<a_2<u_2<a_3<u_3$	$x^2=\frac{(u_1-a_1)(u_2-a_1)(u_3-a_1)}{(a_3-a_1)(a_2-a_1)},$ $y^2=\frac{(u_1-a_2)(u_2-a_2)(u_3-a_2)}{(a_1-a_2)(a_3-a_2)},$ $z^2=\frac{(u_1-a_3)(u_2-a_3)(u_3-a_3)}{(a_1-a_3)(a_2-a_3)}$
XI. Paraboloidal $0<\eta_1<a_2<\eta_2<a_3<\eta_3$	$x^2=\frac{(\eta_1-a_3)(\eta_2-a_3)(\eta_3-a_3)}{(a_3-a_2)},$ $y^2=\frac{(\eta_1-a_2)(\eta_2-a_2)(\eta_3-a_2)}{(a_2-a_3)},$ $z^2=\frac{1}{2}(\eta_1+\eta_2+\eta_3-a_2-a_3)$

$$M_i=-D_{ii}-\frac{k_i^2-\frac{1}{4}}{x_i^2},\quad -\mathcal{H}=M_1+M_2+M_3, \tag{3}$$

$$J_{ij}=L_{ij}^2-\left(k_i^2-\frac{1}{4}\right)\frac{x_j^2}{x_i^2}-\left(k_j^2-\frac{1}{4}\right)\frac{x_i^2}{x_j^2}-\frac{1}{2},\quad i,j=1,2,3, \tag{4}$$

where $L_{ij}=x_i\partial_{x_j}-x_j\partial_{x_i}$, $D_{ii}=-\partial_{x_i}^2+\omega^2x_i^2$ is a diagonal components of the Demkov tensor[1] and we have the notation $x_1=x,\ x_2=y,\ x_3=z$.

The commutators of the operators (3) and (4) can be closed to form a quadratic algebra as follows:

$$[M_i,M_j]=0,\quad [M_i,J_{jk}]=0,\quad [M_i,J_{ij}]=Q_{ij}=Q_{[ij]},\quad [J_{ij},J_{ik}]=R_{[ijk]}=R,$$

where Q_{ij} is totally antisymmetric and the totally antisymmetric quantity $R_{[ijk]}$ is denoted by R. Further commutators are calculated to be

$$[M_i,Q_{jk}]=0,\quad [M_i,Q_{ij}]=4\{M_i,M_j\}+16J_{ij},\quad [M_i,R]=4\{M_k,J_{ij}\}-4\{M_j,J_{ik}\},$$

$$[J_{ij},Q_{ij}]=4\{M_i,J_{ij}\}-4\{M_j,J_{ij}\}-8(k_j^2-1)M_i+8(k_i^2-1)M_j,$$

$$[J_{ij},Q_{ik}]=4\{M_i,J_{jk}\}-4\{M_j,J_{ik}\},$$

$$[J_{ij},R]=4\{J_{ij},J_{jk}\}-4\{J_{ij},J_{ik}\}-8(k_i^2-1)J_{jk}+8(k_j^2-1)J_{ik},$$

where $\{A,B\}=AB+BA$. The expression for the commutators of the Q and R are

$$[Q_{ij},Q_{ik}]=4\{M_i,Q_{jk}\},\quad [Q_{ij},R]=-4\{J_{ij},Q_{ik}\}-4\{J_{ij},Q_{jk}\}.$$

All the commutators of the operators M_i, J_{mn}, Q_{pq}, and R can be expressed in terms of quadratic symmetric products of themselves. The algebra, therefore, is closed quadratically. There are relations between the symmetric products of the generators of this algebra. The exhaustive list of these is as follows:

$$\begin{aligned}Q_{ij}^2={}&\tfrac{8}{3}\{J_{ij},M_i,M_j\}+\tfrac{64}{3}\{M_i,M_j\}+16\omega^2J_{ij}^2-16(1-k_j^2)M_i^2\\&-16(1-k_i^2)M_j^2-\tfrac{128}{3}\omega^2J_{ij}-64\omega^2(1-k_i^2)(1-k_j^2),\end{aligned}$$

$$\begin{aligned}\{Q_{ij},Q_{ik}\}={}&\tfrac{8}{3}\{J_{ij},M_i,M_k\}+\tfrac{8}{3}\{J_{ik},M_i,M_j\}-\tfrac{8}{3}\{J_{jk},M_i,M_i\}\\&+32\omega^2(1-k_i^2)\{J_{ij},J_{ik}\}-32(1-k_i^2)M_jM_k-64\omega^2(1-k_i^2)J_{jk},\end{aligned}$$

$$\begin{aligned}\{Q_{ij},R\}={}&\tfrac{8}{3}\{J_{ij},J_{ij},M_k\}-\tfrac{8}{3}\{J_{ij},J_{ik},M_j\}-\tfrac{8}{3}\{J_{ij},J_{jk},M_i\}-\tfrac{64}{3}\{J_{ij},M_k\}-\tfrac{64}{3}\{J_{ik},M_j\}\\&-\tfrac{64}{3}\{J_{jk},M_i\}+16(1-k_i^2)\{J_{jk},M_j\}+16(1-k_j^2)\{J_{ik},M_i\}-64(1-k_i^2)(1-k_j^2)M_k,\end{aligned}$$

$$\begin{aligned}R^2={}&-\tfrac{4}{3}\{J_{ij},J_{ik},J_{jk}\}+\tfrac{64}{3}\{J_{ij},J_{ik}\}+\tfrac{64}{3}\{J_{ij},J_{jk}\}+\tfrac{64}{3}\{J_{ik},J_{jk}\}-16(1-k_k^2)J_{ij}^2\\&-16(1-k_j^2)J_{ik}^2-16(1-k_i^2)J_{jk}^2+\tfrac{128}{3}(1-k_k^2)J_{ij}\\&+\tfrac{128}{3}(1-k_j^2)J_{ik}+\tfrac{128}{3}(1-k_i^2)J_{jk}+64(1-k_i^2)(1-k_j^2)(1-k_k^2),\end{aligned}$$

where $\{A,B,C\}=ABC+CAB+BCA$. Note that only five operators from (3) and (4) are functionally independent,[7] and for all the coordinate systems that provide separable solutions for the Schrödinger equation the operators characterizing the separation are always combinations of the M_i and J_{ij}.

In the limiting case $k_i=\frac{1}{2}$, we obtain a quadratic algebra, too. In this case

$$Q_{ij}=2(L_{ij}D_{ij}+D_{ij}L_{ij}),\quad R=\{L_{ik},\{L_{ij},L_{kj}\}\},$$

and instead of operators $\{M_i,J_{ij},Q_{ij},R\}$ we can consider as a basis for the symmetries the Demkov tensor D_{ij}, and the components of orbital momentum, L_{ij}. In this regard we arrive at the Lie algebra corresponding to the symmetries of the isotropic oscillator.[1]

Of all the coordinate systems for which separation is possible, in the case of this potential there are only five which are not essentially a Euclidean two-space coordinate system supplemented by an additional Cartesian coordinate z. Such coordinate systems we do not consider further here and the corresponding solutions of the Schrödinger equation and invariant algebra are given in our previous paper[17] (see also Refs. 3 and 8). For the remaining systems we now work out bound state solutions and their corresponding symmetry characterization.

A. Oblate spheroidal basis

Let us consider what we call oblate spheroidal coordinates (see Table II). If we write these coordinates in the form

$$x=x'\cos\varphi,\quad y=x'\sin\varphi,\quad z=y',\tag{5}$$

and put $\Psi=(x')^{-1/2}\Phi$, the Schrödinger equation (1) with potential (2) assumes the form

$$\frac{\partial^2\Phi}{\partial x'^2}+\frac{\partial^2\Phi}{\partial y'^2}+\left[2E-\omega^2(x'^2+y'^2)+\frac{1}{x'^2}\left(\frac{\partial^2}{\partial\varphi^2}-\frac{k_1^2-\frac{1}{4}}{\cos^2\varphi}-\frac{k_2^2-\frac{1}{4}}{\sin^2\varphi}\right)+\frac{1}{4x'^2}-\frac{k_2^3-\frac{1}{4}}{y'^2}\right]\Phi=0.$$

If we now write

$$\Phi=\Lambda(x',y')Y(\varphi),$$

the φ dependence can be extracted by requiring that

$$\left[\frac{\partial^2}{\partial\varphi^2}-\frac{k_1^2-\frac{1}{4}}{\cos^2\varphi}-\frac{k_2^2-\frac{1}{4}}{\sin^2\varphi}\right]Y(\varphi)=-M^2Y(\varphi).\tag{6}$$

The orthonormal solution of Eq. (6) for $\varphi\in[0,\pi/2]$ has the following form:

$$Y_m^{(k_1,k_2)}(\varphi)=\sqrt{\frac{2(2m+k_1+k_2+1)m!\Gamma(m+k_1+k_2+1)}{\Gamma(m+k_1+1)\Gamma(m+k_2+1)}}$$
$$\times(\cos\varphi)^{k_1+1/2}(\sin\varphi)^{k_2+1/2}P_m^{(k_1,k_2)}(\cos2\varphi),\tag{7}$$

where $P_n^{(\alpha,\beta)}(z)$ is a Jacobi polynomial and the separation constant quantizes as

$$M=2m+k_1+k_2+1,\quad m=0,1,2,\ldots.\tag{8}$$

The remaining equation for the function $\Lambda(x',y')$ is

$$\left\{\frac{\partial^2}{\partial x'^2}+\frac{\partial^2}{\partial y'^2}+\left[2E-\omega^2(x'^2+y'^2)-\frac{k_3^2-\frac{1}{4}}{y'^2}-\frac{M^2-\frac{1}{4}}{x'^2}\right]\right\}\Lambda(x',y')=0.$$

This is exactly the equation we have already found (see Ref. 17) in the case of two-dimensional Euclidean space in elliptic coordinates. In terms of the original Cartesian coordinates the bound state solutions have the form

$$\Lambda_{nm}^{k_3}(x,y,z)=e^{-(\omega/2)(x^2+y^2+z^2)}(x^2+y^2)^{m+(k_1+k_2+1)/2}z^{k_3+1/2}\prod_{i=1}^{n}\left(\frac{x^2+y^2}{\theta_i-e_1}+\frac{z^2}{\theta_i-e_2}-1\right), \quad (9)$$

where the θ_i satisfy of the system of n nonlinear equations

$$\frac{M+1}{\theta_i-e_1}+\frac{k_3+1}{\theta_i-e_2}+\sum_{j\neq i}^{n}\frac{2}{\theta_i-\theta_j}-\omega=0.$$

We note that this prescription does correctly give a separable solution by noting the identity

$$\frac{x^2+y^2}{\theta-e_1}+\frac{z^2}{\theta-e_2}-1=-\frac{(u_1-\theta)(u_2-\theta)}{(\theta-e_1)(\theta-e_2)}.$$

The energy E is quantized according to

$$E=\omega(2n+M+k_3+3)=\omega(2N+k_1+k_2+k_3+3), \quad (10)$$

where $N=n+m$ is the principal quantum number.

Consider the Schrödinger equation in the spheroidal separable coordinates (u_1,u_2,φ). After the substitution $\Psi=\psi_1(u_1)\psi_2(u_2)Y(\varphi)$, the separation equations are

$$\frac{d^2\psi(u)}{du^2}+\frac{1}{2}\left(\frac{2}{u-e_1}+\frac{1}{u-e_2}\right)\frac{d\psi(u)}{du}+\frac{1}{4}\left\{\frac{2Eu-\omega^2(u-e_1)(u-e_2)+\lambda}{(u-e_1)(u-e_2)}+\frac{(e_2-e_1)M^2}{(u-e_1)^2(u-e_2)}\right.$$
$$\left.+\frac{(e_1-e_2)(k_3^2-\frac{1}{4})}{(u-e_1)(u-e_2)^2}\right\}\psi(u)=0, \quad (11)$$

where $u=u_1,u_2$ and λ is the oblate spheroidal separation constant. The operator whose eigenvalue is λ is

$$\mathcal{L}_1=\frac{u_2(u_1-e_1)(u_1-e_2)}{u_1-u_2}\left\{\frac{\partial^2}{\partial u_1^2}+\frac{1}{2}\left[\frac{2}{u_1-e_1}+\frac{1}{u_1-e_2}\right]\frac{\partial}{\partial u_1}\right\}-\frac{u_1(u_2-e_1)(u_2-e_2)}{u_1-u_2}$$
$$\times\left\{\frac{\partial^2}{\partial u_2^2}+\frac{1}{2}\left[\frac{2}{u_2-e_1}+\frac{1}{u_2-e_2}\right]\frac{\partial}{\partial u_2}\right\}+\frac{1}{4}\left[\omega^2(e_2e_2-u_1u_2)+\frac{M^2(e_1-e_2)}{(u_1-e_1)(u_2-e_1)}\right.$$
$$\left.\times(u_1+u_2-e_1)+\frac{(k_3^2-\frac{1}{4})(e_2-e_1)}{(u_1-e_2)(u_2-e_2)}(u_1+u_2-e_2)\right]$$
$$=J_{13}+J_{23}+J_{12}+(e_1-e_2)M_3-e_2\mathcal{H}-(k_1^2+k_2^2+k_3^2)+\tfrac{3}{4} \quad (12)$$

with eigenvalues

$$\lambda=-4e_2\sum_i^n\frac{M+1}{\theta_i-e_1}-4e_1\sum_i^n\frac{k_3+1}{\theta_i-e_2}-2[(e_1+e_2)+(e_1k_3+e_2M)]\omega-(k_1^2+k_2^2+k_3^2)+\frac{3}{4}, \quad (13)$$

and the second operator which characterizes the separation of variables in these coordinates is

$$\mathcal{L}_2\Psi=(J_{12}-k_1^2-k_2^2+1)\Psi=-M^2\Psi. \quad (14)$$

To close this paragraph let us note that in the limit $(e_2-e_1)\to 0$ and $(e_2-e_1)\to\infty$ the oblate spheroidal coordinates are changed into spherical and cylindrical polar coordinates, respectively.[8] Correspondingly, the oblate spheroidal bases transform to spherical and cylindrical polar ones.

B. Prolate spheroidal basis

For prolate coordinates the description is almost exactly the same. All that is essentially involved is the interchange of e_1 and e_2.

C. Ellipsoidal basis

For ellipsoidal coordinates (see Table II) we proceed as follows. We consider the Schrödinger equation in Cartesian coordinates,

$$\left(\frac{\partial^2\Psi}{\partial x^2}+\frac{\partial^2\Psi}{\partial y^2}+\frac{\partial^2\Psi}{\partial z^2}\right)+\left[2E-\omega^2(x^2+y^2+z^2)-\frac{(k_1^2-\frac{1}{4})}{x^2}-\frac{(k_2^2-\frac{1}{4})}{y^2}+\frac{(k_3^2-\frac{1}{4})}{z^2}\right]\Psi=0.$$

If we now write

$$\Psi(x,y,z)=e^{-\omega(x^2+y^2+z^2)}x^{2k_1+1}y^{2k_2+1}z^{2k_3+1}\Phi(x,y,z),$$

the equation for Φ becomes

$$\left[\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}+\frac{2k_1+1}{x}\frac{\partial}{\partial x}+\frac{2k_2+1}{y}\frac{\partial}{\partial y}+\frac{2k_3+1}{z}\partial_z-2\omega\left(x\frac{\partial}{\partial x}+y\frac{\partial}{\partial y}+z\frac{\partial}{\partial z}\right)\right.$$
$$\left.-2\omega(k_1+k_2+k_3+3)\right]\Psi=-2E\Psi.$$

To obtain the appropriate finite solutions we can make use of the identity

$$\frac{x^2}{\theta-e_1}+\frac{y^2}{\theta-e_2}+\frac{z^2}{\theta-e_3}-1=-\frac{(u_1-\theta)(u_2-\theta)(u_3-\theta)}{(\theta-e_1)(\theta-e_2)(\theta-e_3)}$$

and write

$$\Phi(x,y,z)=\prod_{j=1}^{N}\left(\frac{x^2}{\theta_j-e_1}+\frac{y^2}{\theta_j-e_2}+\frac{z^2}{\theta_j-e_3}-1\right), \tag{15}$$

where

$$\frac{k_1+1}{\theta_i-e_1}+\frac{k_2+1}{\theta_i-e_2}+\frac{k_3+1}{\theta_i-e_3}+\sum_{j\neq i}^{N}\frac{2}{\theta_i-\theta_j}-\omega=0$$

and the energy level E is given by Eq. (10).

Writing the Schrödinger equation in terms of the ellipsoidal coordinates u_i and using the identities

$$E\equiv\sum_{i=1}^{3}\frac{E_i^2}{\Pi_{i\neq j}(u_i-u_j)},\quad (x^2+y^2+z^2)\equiv\sum_{i=1}^{3}\frac{P(u_i)}{\Pi_{i\neq j}(u_i-u_j)},$$

$$\left[\frac{(k_1^2-\frac{1}{4})}{x^2}+\frac{(k_2^2-\frac{1}{4})}{y^2}+\frac{(k_3^2-\frac{1}{4})}{z^2}\right]\equiv\sum_{i=1}^{3}\frac{A(u_i)}{\Pi_{i\neq j}(u_i-u_j)},$$

where $P(u_i)=(u_i-a_1)(u_i-a_2)(u_i-a_3)$ and $[a_{ik}\equiv(a_i-a_k)]$ and

$$A(u)=a_{31}a_{21}\frac{(k_1^2-\frac{1}{4})}{u-a_1}+a_{12}a_{32}\frac{(k_2^2-\frac{1}{4})}{u-a_2}+a_{13}a_{23}\frac{(k_1^2-\frac{1}{4})}{u-a_3},$$

we arrive at the following equation:

$$\sum_{i=1}^{3}\frac{1}{\Pi_{i\neq j}(u_i-u_j)}\left\{4\sqrt{P(u_i)}\frac{\partial}{\partial u_i}\sqrt{P(u_i)}\frac{\partial}{\partial u_i}+2Eu_i^2-\omega^2P(u_i)-A(u_i)\right\}\Psi=0,$$

which, after the substitution $\Psi=\psi_1(u_1)\psi_2(u_2)\psi_3(u_3)$ and the introduction of the ellipsoidal constants λ_1 and λ_2, is divided into three identical differential equations

$$\sqrt{P(u)}\frac{d}{du}\sqrt{P(u)}\frac{d\psi}{du}+\frac{1}{4}\left[2Eu^2-\omega^2P(u)+\lambda_1u-\lambda_2-\frac{(k_1^2-\frac{1}{4})}{(u-a_1)}(a_3-a_1)(a_2-a_1)\right.$$
$$\left.-\frac{(k_2^2-\frac{1}{4})}{(u-a_2)}(a_1-a_2)(a_3-a_2)-\frac{(k_3^2-\frac{1}{4})}{(u-a_3)}(a_1-a_3)(a_2-a_3)\right]\psi=0,$$

where $u=u_1,u_2,u_3$. The operators that specify the eigenvalues λ_1 and λ_2 are

$$\Lambda_1=J_{12}+J_{23}+J_{23}+(a_2+a_3)M_1+(a_2+a_1)M_3+(a_1+a_3)M_2-(k_1^2+k_2^2+k_3^2)+\tfrac{3}{4}$$

and

$$\Lambda_2=a_3J_{12}+a_2J_{13}+a_1J_{23}+a_2a_3M_1+a_2a_1M_3+a_1a_3M_2$$
$$-k_1^2(a_3+a_2-a_1)-k_2^2(a_1+a_3-a_2)-k_3^2(a_1+a_2-a_3)+\tfrac{3}{4}(a_1+a_2+a_3),$$

respectively. In terms of the zeros θ_j the eigenvalues of these operators are

$$\lambda_1=-2[k_2(a_1+a_3)+k_1(a_2+a_3)+k_3(a_1+a_2)-4(a_1+a_2+a_3)]\omega-2(k_1+k_2+k_3)$$
$$\times(k_1+k_2+k_3+1)-\tfrac{9}{2}+4\left[\sum_{i=1}^{N}a_2\frac{(k_2+1)}{(\theta_i-a_2)}+\sum_{i=1}^{N}a_1\frac{(k_1+1)}{(\theta_i-a_1)}+\sum_{i=1}^{N}a_3\frac{(k_3+1)}{(\theta_i-a_3)}\right] \quad (16)$$

and

$$\lambda_2=-\tfrac{1}{2}(a_1+a_2+a_3)-2\omega[a_2a_3(k_1+1)+a_2a_1(k_3+1)+a_1a_3(k_2+1)]-a_1(k_2+k_3+1)^2$$
$$-a_2(k_1+k_3+1)^2-a_3(k_2+k_1+1)^2$$
$$-4\left[\sum_{i=1}^{N}a_3a_1\frac{(k_2+1)}{(\theta_i-a_2)}+\sum_{i=1}^{N}a_3a_2\frac{(k_1+1)}{(\theta_i-a_1)}+\sum_{i=1}^{N}a_2a_1\frac{(k_3+1)}{(\theta_i-a_3)}\right]. \quad (17)$$

D. Spherical and sphero-conical bases

For spherical-type coordinates there are two possibilities.

If we choose coordinates in Euclidean space accordingly,

$$x=rs_1,\quad y=rs_2,\quad z=rs_3, \quad (18)$$

where $s_1^2+s_2^2+s_3^2=1$, and put the wave function in the form

$$\Psi=R(r)S(\rho_1,\rho_2), \quad (19)$$

where ρ_1,ρ_2 are the spherical or sphero-conical coordinates, after separation of variables, we arrive at two equations,

$$\frac{d^2R}{dr^2}+\frac{2}{r}\frac{dR}{dr}+\left[2E-\omega^2r^2-\frac{J(J+1)}{r^2}\right]R=0, \tag{20}$$

$$\left\{L_{12}+L_{23}+L_{13}+\left[J(J+1)-\frac{(k_1^2-\frac{1}{4})}{s_1^2}-\frac{(k_2^2-\frac{1}{4})}{s_2^2}-\frac{(k_3^2-\frac{1}{4})}{s_3^2}\right]\right\}S=0, \tag{21}$$

where J is the spherical separation constant.

(1) In the spherical coordinates (see Table II) the wave function $S(\rho_1,\rho_2)$ has the separable form

$$S(\vartheta,\varphi)=Z(\vartheta)Y_m^{(k_1,k_2)}(\varphi),$$

where $Y_m^{(k_1,k_2)}(\varphi)$ is given by formula (7). This leads to the equation for Z:

$$\frac{1}{\sin\vartheta}\frac{d}{d\vartheta}\sin\vartheta\frac{dZ}{d\vartheta}+\left[J(J+1)-\frac{M^2}{\sin^2\vartheta}-\frac{k_3^2-\frac{1}{4}}{\cos^2\vartheta}\right]Z=0,\quad M=2m+k_1+k_2+1.$$

The solution of the above equation is (see Ref. 8)

$$Z(\theta)=\sqrt{\frac{2[2(m+l+1)+k_1+k_2+k_3]l!\,\Gamma(l+2m+k_1+k_2+k_3+2)}{\Gamma(l+k_3+1)\Gamma(l+2m+2+k_1+k_2)}}$$
$$\times(\cos\theta)^{1/2+k_3}(\sin\theta)^MP_l^{(M,k_3)}(\cos2\theta),\quad l\in\mathbf{N}, \tag{22}$$

and for a spherical separation constant we get

$$J=2l+M+k_3+\tfrac{1}{2}=2l+2m+k_1+k_2+k_3+\tfrac{3}{2}. \tag{23}$$

(2) If we choose the sphero-conical coordinates on the sphere (see Table II), the solution of the equation (21) has the form

$$S(\rho_1,\rho_2)=\prod_{\ell=1}^{3}s_\ell^{k_\ell+1/2}\prod_{j=1}^{n}\left(\frac{s_1^2}{\theta_j-e_1}+\frac{s_2^2}{\theta_j-e_2}+\frac{s_3^2}{\theta_j-e_3}\right), \tag{24}$$

and the spherical separation constant is quantized according to Eq. (23) where $n=l+m$. This achieves a separation of variables solution because of the identity

$$\frac{s_1^2}{\theta_j-e_1}+\frac{s_2^2}{\theta_j-e_2}+\frac{s_3^2}{\theta_j-e_3}=\frac{(\rho_1-\theta_j)(\rho_2-\theta_j)}{(\theta_j-e_1)(\theta_j-e_2)(\theta_j-e_3)}$$

and the Niven equations

$$\frac{k_1+1}{\theta_i-e_1}+\frac{k_2+1}{\theta_i-e_2}+\frac{k_3+1}{\theta_i-e_3}+\sum_{j\neq i}^{q}\frac{2}{\theta_i-\theta_j}=0.$$

The functions $S(\rho_1,\rho_2)$ have the separable form

$$S(\rho_1,\rho_2)=B_1(\rho_1)B_2(\rho_2), \tag{25}$$

and the separation equations are $[P(\rho)=(\rho-e_1)(\rho-e_2)(\rho-e_3)]$

$$\sqrt{P(\rho)}\frac{d}{d\rho}\sqrt{P(\rho)}\frac{dB}{d\rho}+\frac{1}{4}\left[\lambda-J(J+1)-\frac{(k_1^2-\frac{1}{4})}{(\rho-e_1)}(e_1-e_2)(e_1-e_3)-\frac{(k_2^2-\frac{1}{4})}{(\rho-e_2)}(e_2-e_1)(e_2-e_3)\right.$$
$$\left.-\frac{(k_3^2-\frac{1}{4})}{(\rho-e_3)}(e_3-e_2)(e_3-e_1)\right]B=0, \tag{26}$$

where $B=B_1,B_2$ according as $\rho=\rho_1,\rho_2$, respectively. The sphero-conical wave functions satisfy the eigenfunction equations

$$(J_{12}+J_{13}+J_{23})S=[(k_1^2+k_2^2+k_3^2)-(2q+2+k_1+k_2+k_3)^2-\tfrac{1}{2}]S \tag{27}$$

$$(e_1J_{23}+e_2J_{13}+e_3J_{12})S$$
$$=[k_1^2(e_2+e_3-e_1)+k_2^2(e_1+e_3-e_2)+k_3^2(e_1+e_2-e_3)-\tfrac{3}{4}(e_1+e_2+e_3)-\lambda]S, \tag{28}$$

where

$$\lambda=2[k_1(e_2+e_3)+k_2(e_1+e_3)+k_3(e_2+e_1)+e_3k_1k_2+e_2k_1k_3+e_1k_3k_2]+\tfrac{3}{2}(e_1+e_2+e_3)$$
$$-4\left[e_2e_3\sum_{i=1}^{n}\frac{k_1+1}{\theta_i-e_1}+e_1e_3\sum_{i=1}^{n}\frac{k_2+1}{\theta_i-e_2}+e_2e_1\sum_{i=1}^{n}\frac{k_3+1}{\theta_i-e_3}\right]. \tag{29}$$

Let us now go to the radial equation (20). This equation is very reminiscent of the radial equation for the three-dimensional harmonic oscillator except that the orbital quantum number l is replaced by $2l+2m+k_1+k_2+k_3+\frac{3}{2}$. The orthonormal solution of the radial equation (20), in terms of Laguerre polynomials $L_n^{\alpha}(x)$, is

$$R_{n_rJ}(r)=\sqrt{\frac{2\omega^{3/2}n_r!}{\Gamma(n_r+2l+2m+k_1+k_2+k_3+3)}}(\sqrt{\omega}r)^J\exp\left(-\frac{\omega}{2}r^2\right)L_{n_r}^{J+1/2}(\omega r^2), \tag{30}$$

and the energy spectrum is given by formula (10) where the $n_r=0,1,2,...$ is the radial quantum number and the principal quantum number now is $N=(n_r+n)=(n_r+l+m)$.

III. GENERALIZED ANISOTROPIC OSCILLATOR

The second potential (see Table I) is

$$V_2(x,y,z)=\frac{\omega^2}{2}(x^2+y^2+4z^2)+\frac{1}{2}\left[\frac{k_1^2-\frac{1}{4}}{x^2}+\frac{k_2^2-\frac{1}{4}}{y^2}\right]. \tag{31}$$

The corresponding Schrödinger equation has separable solutions in five coordinate systems: Cartesian coordinates, cylindrical polar coordinates, cylindrical elliptic coordinates, cylindrical parabolic coordinates, and parabolic coordinates. It is the last of these that gives interesting new solutions. The first four coordinate systems are of cylindrical type and can be deduced from what we already know for Euclidean two-dimensional space (see Refs. 8 and 17). Before considering the bound state solutions in the case of the parabolic coordinate system we consider the quadratic algebra of second-order symmetry operators which are associated with this potential. A basis for these operators is

$$M_1=\partial_x^2-\omega^2x^2+\frac{k_1^2-\frac{1}{4}}{x^2},\quad M_2=\partial_y^2-\omega^2y^2-\frac{k_2^2-\frac{1}{4}}{y^2},\quad P=\partial_z^2-4\omega^2z^2, \tag{32}$$

$$L=L_{12}^2-\left(k_1^2-\frac{1}{4}\right)\frac{y^2}{x^2}-\left(k_2^2-\frac{1}{4}\right)\frac{x^2}{y^2}-\frac{1}{2}, \tag{33}$$

$$S_1=-\frac{1}{2}(p_xL_{13}+L_{13}p_x)+z\left(\omega^2x^2-\frac{k_1^2-\frac{1}{4}}{x^2}\right),$$

$$S_2=-\frac{1}{2}(p_yL_{23}+L_{23}p_y)+z\left(\omega^2y^2-\frac{k_2^2-\frac{1}{4}}{x^2}\right), \tag{34}$$

where $p_{x,y}=\partial_{x,y}$.

The relations that define the quadratic algebra are obtained by exhaustive commutation. The nonzero commutators of the above basis are

$$[M_1,L]=[L,M_2]=Q,\quad [L,S_1]=[S_2,L]=B,\quad [M_i,S_i]=A_i,\quad [P,S_i]=-A_i.$$

Further nonzero commutators with Q are

$$[M_i,Q]=[Q,M_2]=4\{M_1,M_2\}+16\omega^2L,\quad [S_1,Q]=[Q,S_2]=4\{M_1,M_2\},$$

$$[L,Q]=4\{M_1,L\}-4\{M_2,L\}+16(1-k_1^2)M_1-16(1-k_2^2)M_2;$$

nonzero commutators with A_i are

$$[M_i,A_i]=16\omega^2S_i,\quad [L,A_1]=[A_2,L]=4\{M_1,S_2\}-4\{M_2,S_1\},\quad [P,A_i]=-16\omega^2S_i,$$

$$[S_i,A_i]=\{M_i,M_i\}-2\{M_i,P\}+8\omega^2(1-k_i^2),\quad [S_i,A_j]=\{M_i,M_j\}+4\omega^2L;$$

and nonzero commutators with B are

$$[M_1,B]=-4\{M_2,S_1\},\quad [M_2,B]=-4\{M_1,B\},\quad [P,B]=4\{M_2,S_1\}-4\{M_1,S_2\},$$

$$[L,B]=-4\{L,S_1\}+4\{L,S_2\}-16(1-k_2^2)S_1+16(1-k_1^2)S_2,$$

$$[S_1,B]=\{L,M_1\}-2\{L,P\}-4\{S_1,S_2\}-4(1-k_1^2)M_2,$$

$$[B,S_2]=\{L,M_2\}-2\{L,P\}-4\{S_1,S_2\}-4(1-k_2^2)M_1.$$

The remaining nonzero commutators are

$$[Q,A_i]=-4\{M_i,A_j\},\quad [Q,B]=-4\{L,A_1\}-4\{L,A_2\},\quad [A_1,A_2]=4\omega^2Q,$$

$$[A_1,B]=\{M_1,Q\}-4\{S_1,A_2\},\quad [B,A_2]=\{M_2,Q\}-4\{S_2,A_1\}.$$

There are also various relations among the generators of our quadratic algebra:

$$\{M_1,B\}=\{L,A_1\}-\{S_1,Q\}-4(1-k_1^2)A_2,$$

$$\{M_2,B\}=-\{L,A_2\}-\{S_2,Q\}+4(1-k_2^2)A_1,$$

$$\{P,Q\}=2\{S_1,A_2\}-2\{S_2,A_1\},\quad \{M_1,A_2\}-\{M_2,A_1\}-4\omega^2B=0,$$

$$Q^2=\tfrac{8}{3}\{L,M_1,M_2\}+8\omega^2\{L,L\}-16(1-k_1^2)M_1^2-16(1-k_2^2)M_2^2$$
$$+\tfrac{64}{3}\{M_1,M_2\}-\tfrac{128}{3}\omega^2L-128\omega^2(1-k_1^2)(1-k_2^2),$$

$$\{Q,A_1\}=\tfrac{8}{3}\{M_1,M_2,S_1\}-\tfrac{8}{3}\{M_1,M_1,S_2\}+16\omega^2\{L,S_1\}-64(1-k_1^2)S_2,$$

$$\{Q,A_2\}=-\tfrac{8}{3}\{M_1,M_2,S_2\}+\tfrac{8}{3}\{M_2,M_2,S_1\}-16\omega^2\{L,S_2\}+64(1-k_2^2)S_1,$$

$$\{Q,B\}=-\tfrac{8}{3}\{M_2,L,S_1\}-\tfrac{8}{3}\{M_1,L,S_2\}+16(1-k_1^2)\{M_2,S_2\}$$
$$+16(1-k_2^2)\{M_1,S_1\}-\tfrac{64}{3}\{M_1,S_2\}-\tfrac{64}{3}\{M_2,S_1\},$$

$$A_1^2=\tfrac{2}{3}\{M_1,M_1,P\}+8\omega^2\{S_1,S_1\}+16\omega^2(1-k_1^2)P-32\omega^2M_1,$$

$$\{A_1,A_2\}=\tfrac{4}{3}\{M_1,M_2,P\}+16\omega^2\{S_1,S_2\}+8\omega^2\{L,P\},$$

$$\{A_1,B\}=\tfrac{8}{3}\{M_1,S_1,S_2\}-\tfrac{8}{3}\{M_2,S_1,S_1\}+\tfrac{4}{3}\{M_1,L,P\}+\tfrac{32}{3}\{M_1,M_2\}-8(1-k_1^2)\{M_1,P\}-\tfrac{64}{3}\omega^2L,$$

$$\{A_2,B\}=-\tfrac{8}{3}\{M_2,S_2,S_1\}+\tfrac{8}{3}\{M_1,S_2,S_2\}-\tfrac{4}{3}\{M_2,L,P\}$$
$$-\tfrac{32}{3}\{M_1,M_2\}+8(1-k_2^2)\{M_2,P\}+\tfrac{64}{3}\omega^2L,$$

$$B^2=\tfrac{8}{3}\{L,S_1,S_2\}+\tfrac{2}{3}\{L,L,P\}+\tfrac{64}{3}\{S_1,S_2\}-16(1-k_1^2)S_2^2-16(1-k_2^2)S_1^2$$
$$+\tfrac{16}{3}\{L,M_1\}-\tfrac{16}{3}\{L,P\}+\tfrac{32}{3}(1-k_2^2)M_1+\tfrac{32}{3}(1-k_1^2)M_2-16(1-k_1^2)(1-k_2^2)P.$$

This completes the nonzero relations for the quadratic algebra and the associated relations among the generators. For the last coordinate system in our list we develop the bound state solutions.

Parabolic basis

The Schrödinger equation in Cartesian coordinates with this potential has the form

$$\left(\frac{\partial^2\Psi}{\partial x^2}+\frac{\partial^2\Psi}{\partial y^2}+\frac{\partial^2\Psi}{\partial z^2}\right)+\left[2E-\omega^2(x^2+y^2+4z^2)-\frac{(k_1^2-\frac{1}{4})}{x^2}+\frac{(k_2^2-\frac{1}{4})}{y^2}\right]\Psi=0.$$

If we choose the coordinates (x',y',φ) according formula (6) and the wave function Ψ in the form

$$\Psi(x',y',\varphi)=(x')^{-1/2}\Lambda(x',y')Y_m^{(k_1,k_2)}(\varphi),\quad m=0,1,...,$$

where $Y_m^{(k_1,k_2)}(\varphi)$ is given by (7), then the equation for the function $\Lambda(x',y')$ is

$$\left[\frac{\partial^2}{\partial x'^2}+\frac{\partial^2}{\partial y'^2}-\omega^2(x'^2+4y'^2)-\frac{(M^2-\frac{1}{4})}{x'^2}+2E\right]\Lambda(x',y')=0.$$

This just the problem whose solution has been found (see Ref. 1) in the case of two-dimensional Euclidean space. If now we write

$$\Lambda(x,y,z)=e^{-(\omega/2)(x^2+y^2)-\omega z^2}(x^2+y^2)^{(k_2/2+1/4)}P(x,y),$$

where

$$P(x,y)=\prod_{j=1}^{n}\left(\frac{x^2+y^2}{\theta_j^2}+2z-\theta_j^2\right),$$

then the λ_j satisfy

$$\frac{4(\lambda+1)}{\theta_\ell^2}+\sum_{i\neq l}^{n}\frac{4}{\theta_\ell^2-\theta_i^2}-2\omega\theta_\ell^2=0$$

and energy E quantizes according to

$$E=\omega(2n+M+2)=\omega(2N+k_1+k_2+3), \tag{35}$$

where the principal quantum number $N=n+m$. This method of solution is based on the identity

$$\frac{x^2+y^2}{\theta^2}+2z-\theta^2=\frac{(\xi^2-\theta^2)(\eta^2+\theta^2)}{\theta^2}.$$

In fact, the separation equations in ξ and η for solution of the Schrödinger equation

$$\Psi(\xi,\eta,\varphi)=X_1(\xi)X_2(\eta)Y_m^{(k_1,k_2)}(\varphi)$$

have the form

$$\left[\frac{d^2}{d\rho^2}+\frac{1}{\rho}\frac{d}{d\rho}+\left(2E\rho^2-\omega^2\rho^6-\frac{M^2}{\rho^2}+\epsilon\beta\right)\right]X(\rho)=0, \tag{36}$$

where $\epsilon=1$ if $\rho=\xi$ and -1 if $\rho=\eta$, and β is the parabolic separation constant. By eliminating the energy E from Eqs. (36) we produce the operator, the eigenvalues of which is β:

$$\mathcal{L}=\frac{1}{\xi^2+\eta^2}\left(\frac{\xi^2}{\eta}\frac{\partial}{\partial\eta}\eta\frac{\partial}{\partial\eta}-\frac{\eta^2}{\xi}\frac{\partial}{\partial\xi}\xi\frac{\partial}{\partial\xi}\right)+\omega^2\xi^2\eta^2(\xi^2-\eta^2)-\frac{\xi^2-\eta^2}{\xi^2\eta^2}\left(\frac{\partial^2}{\partial\varphi^2}-\frac{k_1^2-\frac{1}{4}}{\cos^2\varphi}-\frac{k_2^2-\frac{1}{4}}{\sin^2\varphi}\right). \tag{37}$$

In Cartesian coordinates the operator $\mathcal{L}$ can be written as

$$\mathcal{L}=z\left[\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\omega^2(x^2+y^2)-\frac{k_1^2-\frac{1}{4}}{x^2}-\frac{k_2^2-\frac{1}{4}}{y^2}\right]-\frac{\partial}{\partial z}\left(x\frac{\partial}{\partial x}+y\frac{\partial}{\partial y}-1\right),$$

and thus the parabolic basis satisfies two eigenvalue equations

$$L\Psi=(k_1^2+k_2^2-M^2-1)\Psi,\quad \mathcal{L}\Psi=2(S_1+S_2)\Psi=\beta\Psi,$$

where operators L, $S_{1,2}$ are given by formulas (33) and (34) and the eigenvalue β is

$$\beta=-2(M-1)\prod_{j=1}^{n}\theta_j^2\left(\sum_{k=1}^{n}\theta_k^{-2}\right). \tag{38}$$

IV. GENERALIZED KEPLER–COULOMB SYSTEM

The third potential we consider is

$$V_3(x,y,z)=-\frac{\alpha}{\sqrt{x^2+y^2+z^2}}+\frac{1}{2}\left[\frac{k_1^2-\frac{1}{4}}{x^2}+\frac{k_2^2-\frac{1}{4}}{y^2}\right]. \tag{39}$$

The corresponding Schrödinger equation has the form

$$\left(\frac{\partial^2\Psi}{\partial x^2}+\frac{\partial^2\Psi}{\partial y^2}+\frac{\partial^2\Psi}{\partial z^2}\right)+\left[2E+\frac{2\alpha}{\sqrt{x^2+y^2+z^2}}-\left(\frac{k_1^2-\frac{1}{4}}{x^2}+\frac{k_2^2-\frac{1}{4}}{y^2}\right)\right]\Psi=0.$$

This equation admits separable solutions in the four coordinate systems: spherical, spheroconical, prolate spheroidal, and parabolic.

The second-order symmetries of the corresponding Schrödinger equation are

$$J_{23}=L_{23}^2-\left(k_2^2-\frac{1}{4}\right)\frac{z^2}{y^2}-\frac{1}{2},\quad J_{13}=L_{13}^2-\left(k_1^2-\frac{1}{4}\right)\frac{z^2}{x^2}-\frac{1}{2},$$

$$J_{12}=L_{12}^2-\left(k_2^2-\frac{1}{4}\right)\frac{x^2}{y^2}-\left(k_1^2-\frac{1}{4}\right)\frac{y^2}{x^2}-\frac{1}{2},\tag{40}$$

$$L=-\frac{1}{2}[\{p_x,L_{13}\}+\{p_y,L_{23}\}]+\frac{\alpha z}{\sqrt{x^2+y^2+z^2}}-z\left(\frac{k_1^2-\frac{1}{4}}{x^2}+\frac{k_2^2-\frac{1}{4}}{y^2}\right).\tag{41}$$

These symmetry operators do not appear to close under repeated commutation. One obvious subalgebra that is quadratically closed is that generated by the elements J_{12}, J_{13}, and J_{23}. The closure relations can be readily deduced from the algebra given for the first potential with the proviso that $k_3=\frac{1}{2}$.

A. Spherical and sphero-conical bases

If we use polar coordinates according to (18) and write the wave function Ψ in the separable form $\Psi=R(r)S(\rho_1,\rho_2)$, then the separable equations are

$$\frac{d^2R}{dr}+\frac{2}{r}\frac{dR}{dr}+\left[2E+\frac{2\alpha}{r}-\frac{J(J+1)}{r^2}\right]R=0,\tag{42}$$

$$\left\{L_{12}+L_{23}+L_{13}+\left[J(J+1)-\frac{(k_1^2-\frac{1}{4})}{s_1^2}-\frac{(k_2^2-\frac{1}{4})}{s_2^2}\right]\right\}S=0.\tag{43}$$

(1) In the spherical coordinates, choosing the wave function $S(\rho_1,\rho_2)$ according to

$$S(\vartheta,\varphi)=Z(\vartheta)Y_m^{(k_1,k_2)}(\varphi),\quad m=0,1,2,...,$$

where $Y_m^{(k_1,k_2)}(\varphi)$ is given by formula (7), we go to the equation for Z:

$$\frac{1}{\sin\vartheta}\frac{d}{d\vartheta}\sin\vartheta\frac{dZ}{d\vartheta}+\left[J(J+1)-\frac{M^2}{\sin^2\vartheta}\right]Z=0,\quad M=2m+k_1+k_2+1.$$

The orthonormal solution of the above equation for $\vartheta\in[0,\pi]$ is

$$Z(\vartheta)=\frac{2^M}{\sqrt{\pi}}\sqrt{\frac{(2l+2M+1)l!}{2\Gamma(l+2M+1)}}\Gamma\left(M+\frac{1}{2}\right)(\sin\vartheta)^M C_l^{M+1/2}(\cos\vartheta),\tag{44}$$

where $l\in\mathbf{N}$ and $C_n^\lambda(x)$ is a Gegenbauer polynomial.[21] The spherical separation constant is given by

$$J=l+M=l+2m+k_1+k_2+1.\tag{45}$$

(2) The solution of the Schrödinger equation (43) in sphero-conical coordinates follows from what we have done before in Sec. II D, part (2). If we write $S(\rho_1,\rho_2)$ as

$$S(\rho_1,\rho_2)=s_3^\epsilon\prod_{\ell=1}^{2}s_\ell^{k_\ell+1/2}\prod_{j=1}^{n}\left(\frac{s_1^2}{\theta_j-e_1}+\frac{s_2^2}{\theta_j-e_2}+\frac{s_3^2}{\theta_j-e_3}\right),$$

where $\epsilon=0,1$ then the zeros satisfy the Niven equations

$$\frac{k_1+1}{\theta_i-e_1}+\frac{k_2+1}{\theta_i-e_2}+\frac{\epsilon+\frac{1}{2}}{\theta_i-e_3}+\sum_{j\neq i}^{n}\frac{2}{\theta_i-\theta_j}=0.$$

The functions $S(\rho_1,\rho_2)$ satisfy the eigenfunction equations

$$(J_{12}+J_{13}+J_{23})S=[(k_1^2+k_2^2)-(2q+\tfrac{3}{2}+k_1+k_2+\epsilon)^2-\tfrac{7}{4}]S, \tag{46}$$

$$(e_1J_{23}+e_2J_{13}+e_3J_{12})S=[(e_2-e_1)(k_1^2-k_2^2)+e_3(k_1^2+k_2^2-1)-\tfrac{1}{2}(e_1+e_2)-\lambda]S, \tag{47}$$

where the sphero-conical separation constant λ is

$$\lambda=-2[k_1(e_2+e_3)+k_2(e_1+e_3)+(\epsilon-\tfrac{1}{2})(e_2+e_1)+e_3k_1k_2+(e_2k_1+e_1k_2)(\epsilon-\tfrac{1}{2})]$$
$$-\frac{3}{2}(e_1+e_2+e_3)-4\left[e_2e_3\sum_{i=1}^{n}\frac{k_1+1}{\theta_i-e_1}+e_1e_3\sum_{i=1}^{n}\frac{k_2+1}{\theta_i-e_2}+e_2e_1\sum_{i=1}^{n}\frac{\epsilon+\frac{1}{2}}{\theta_i-e_3}\right]. \tag{48}$$

Finally, let us consider the radial equation (42). The introduction of (45) into (42) leads to

$$\frac{d^2R}{dr}+\frac{2}{r}\frac{dR}{dr}+\left[2E+\frac{2\alpha}{r}-\frac{(l+2m+k_1+k_2+1)(l+2m+k_1+k_2+2)}{r^2}\right]R=0,$$

which is the radial equation for the Coulomb problem, except the orbital quantum number l is replaced here by $(l+2m+k_1+k_2+1)$. The bound state solution of Eq. (43) is

$$R_{NJ}(r)=\frac{2(\alpha)^{3/2}}{N^2}\sqrt{\frac{\Gamma(N+J+1)}{(N-J-1)!}}\left(\frac{2\alpha r}{N}\right)^J\frac{e^{-\alpha r/N}}{\Gamma(2J+1)}\,{}_1F_1\left(-N+J+1;2J+2;\frac{2\alpha r}{N}\right)$$

and the energy spectrum given by

$$E=-\frac{\alpha^2}{N^2},\quad N=n_r+J+1=2m+n_r+l+k_1+k_2+2,\quad n_r=0,1,2,\dots.$$

B. Parabolic and prolate spheroidal bases

The remaining solutions for which separation of variables is possible can be best observed by writing the Schrödinger equation in parabolic coordinates. If we do this and choose solutions of the form

$$\Psi=S(\xi,\eta)(\xi\eta)^{-1/2}Y_m^{(k_1,k_2)}(\varphi),\quad m=0,1,2,\dots, \tag{49}$$

where $Y_m^{(k_1,k_2)}(\varphi)$ is given by formula (7), we find that the Schrödinger equation has the reduced form

$$\frac{\partial^2S}{\partial\xi^2}+\frac{\partial^2S}{\partial\eta^2}+\left[2E(\xi^2+\eta^2)+\left(M^2-\frac{1}{4}\right)\left(\frac{1}{\xi^2}+\frac{1}{\eta^2}\right)+4\alpha\right]S=0.$$

This is clearly recognizable as solvable via separation of variables in parabolic coordinates ξ and η. The separable solution for the wave function $S(\xi,\eta)$ is

$$S(\xi,\eta)=\frac{\sqrt{2}(\alpha)^{3/2}}{N^2}(\xi\eta)^{-1/2}f_{n_1}^M(\xi)f_{n_2}^M(\eta),\quad n_{1,2}\in\mathbf{N},\quad M=2m+k_1+k_2+1, \tag{50}$$

where

$$f_n^M(x)=\sqrt{\frac{\Gamma(n_1+M+1)}{n_1!}\frac{e^{-(\alpha/2N)x^2}}{\Gamma(M+1)}}\left(\frac{\alpha x^2}{N}\right)^{M/2}{}_1F_1\left(-n_1;M+1;\frac{\alpha x^2}{N}\right) \tag{51}$$

and $N=n_1+n_2+M+1=n_1+n_2+2m+k_1+k_2+2$.

It is also interesting to observe that we could contemplate an E-dependent algebra of second-order symmetries acting on the functions $H(\xi,\eta)$. Indeed, a basis for such symmetries is

$$P_1=\partial_\xi^2+\left(M^2-\frac{1}{4}\right)\frac{1}{\xi^2}+2E\xi^2,\quad P_2=\partial_n^2+\left(M^2-\frac{1}{4}\right)\frac{1}{\eta^2}+2E\eta^2,$$

$$M=(\xi\partial_\eta-\eta\partial_\xi)^2-\left(M^2-\frac{1}{4}\right)\left(\frac{\xi^2}{\eta^2}+\frac{\eta^2}{\xi^2}\right)-\frac{1}{2}.$$

The corresponding closure relations can be deduced from those given for the first potential.

Apart from the symbols this has the same form as was dealt with in two dimensions. If we now regard ξ and η as Cartesian coordinates, separation is also possible in polar and elliptical coordinates. The case of polar coordinates has essentially been done above. The case of elliptic coordinates can be done by the standard prescription. This is achieved by looking for solutions of the form

$$S(\xi,\eta)=e^{-\sqrt{-2E(x^2+y^2+z^2)}}(x^2+y^2)^{(1/2)(M+1/2)}\prod_{j=1}^{s}\left(\frac{\sqrt{x^2+y^2+z^2}+z}{\theta_m-e_1}+\frac{\sqrt{x^2+y^2+z^2}-z}{\theta_m-e_1}-1\right),$$

where we have written the solutions in the coordinate representation. (Recall that $\xi^2=\sqrt{x^2+y^2+z^2}+z$ and $\eta^2=\sqrt{x^2+y^2+z^2}-z$.) With

$$\xi=\sqrt{\frac{(u_1-e_1)(u_2-e_1)}{(e_2-e_1)}},\quad \eta=\sqrt{\frac{(u_1-e_2)(u_2-e_2)}{(e_1-e_2)}},$$

where $e_1<u_1<e_2<u_2$, the choice of Cartesian coordinates that is appropriate in this case is

$$x=\frac{1}{e_2-e_1}\sqrt{\left[\left(\frac{e_2-e_1}{2}\right)^2-\left(u_1-\frac{e_2+e_1}{2}\right)^2\right]\left[\left(u_2-\frac{e_2+e_1}{2}\right)^2-\left(\frac{e_2-e_1}{2}\right)^2\right]}\cos\varphi,$$

$$y=\frac{1}{e_2-e_1}\sqrt{\left[\left(\frac{e_2-e_1}{2}\right)^2-\left(u_1-\frac{e_2+e_1}{2}\right)^2\right]\left[\left(u_2-\frac{e_2+e_1}{2}\right)^2-\left(\frac{e_2-e_1}{2}\right)^2\right]}\sin\varphi,$$

$$z=\frac{1}{e_2-e_1}\left[\left(u_1-\frac{e_2+e_1}{2}\right)\left(u_2-\frac{e_2+e_1}{2}\right)+\left(\frac{e_2-e_1}{2}\right)^2\right].$$

This corresponds to the choice of prolate spheroidal coordinates of type II.[22,8]

V. INTERBASIS EXPANSION

According to the principles of quantum mechanics the solutions of the same Schrödinger equation in the different separable coordinate systems for a given value of energy E are connected by unitary transformations or interbasis expansions. For example, we examine here the direct calculation of the interbasis expansion between the spherical and parabolic wave functions for potential V_3. We have

$$\Psi_{n_1,n_2,m}(\xi,\eta,\varphi)=\sum_{l=0}^{n_1+n_2} W^l_{n_1n_2m}(k_1,k_2)\Psi_{n_rlm}(r,\vartheta,\varphi). \tag{52}$$

where $n_r+l=n_1+n_2$. For calculation of the coefficients of interbasis expansion in (52) we may use the ''asymptotic method,''[18,22] which is the following. Writing the parabolic wave function on the left-hand side of (52) in spherical coordinates (r,ϑ,φ) accordingly,

$$\xi^2=r(1+\cos\vartheta),\quad \eta^2=r(1-\cos\vartheta),$$

eliminating the function $Y_m^{(k_1,k_2)}(\varphi)$ on both sides of (52), and using the formula

$${}_1F_1(-n;\alpha;x)\sim\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}(-x)^n$$

for x arbitrary large, we see that the expansion (52) yields an equation which depends only on the variable ϑ. Then, by using the orthogonality relations for the functions $Z_{lm}(\vartheta)$ in the quantum number l, we arrive at the following expression for interbasis expansions coefficients:

$$\begin{aligned}W^l_{n_1n_2m}(k_1,k_2)&=(-1)^l\frac{\Gamma(M+1/2)}{2^{n_1+n_2+1}\sqrt{\pi}}\\&\times\sqrt{\frac{(2l+2M+1)\Gamma(n_1+n_2+l+2M+2)(n_1+n_2-l)!l!}{\Gamma(l+2M+1)\Gamma(n_1+M+1)\Gamma(n_1+M+1)(n_1)!(n_2)!}}\\&\times\int_0^\pi(1+\cos\vartheta)^{n_1+M}(1-\cos\vartheta)^{n_2+M}C_l^{M+1/2}(\vartheta)\sin\vartheta\,d\vartheta,\\M&=2m+k_1+k_2+1.\end{aligned}\tag{53}$$

By using the Rodrigues formula for the Gegenbauer polynomials[21]

$$C_n^\lambda(x)=\frac{(-1)^l}{l!}\frac{\sqrt{\pi}\Gamma(l+2\lambda)}{2^{l+2\lambda-1}\Gamma(\lambda)\Gamma(l+\lambda+1/2)}(1-x^2)^{-\lambda+1/2}\frac{d^l}{dx^l}(1-x^2)^{l+\lambda-1/2}$$

and comparing (53) with the integral representation for the Clebsch–Gordan coefficients of the Lie group SU(2) (Ref. 23),

$$\begin{aligned}C^{c\gamma}_{a\alpha;b\beta}=\delta_{\alpha+\beta,\gamma}&\sqrt{\frac{(2c+1)(J+1)!(J-2c)!(c+\gamma)!}{(J-2a)!(J-2b)!(a-\alpha)!(a+\alpha)!(b-\beta)!(b+\beta)!(c-\gamma)!}}\frac{(-1)^{a-c+\beta}}{2^{J+1}}\\&\times\int_{-1}^1(1-x)^{a-\alpha}(1+x)^{b-\beta}\frac{d^{c-\gamma}}{dx^{c-\gamma}}[(1-x)^{J-2a}(1+x)^{J-2b}]dx\end{aligned}$$

with $J=a+b+c$, we obtain

$$\begin{gathered}W^l_{n_1n_2m}(k_1,k_2)=(-1)^{n_2}C^{c,\alpha+\beta}_{a\alpha;a\beta},\\a=\frac{n_1+n_2+2m+k_1+k_2+1}{2},\quad c=l+2m+k_1+k_2+1,\\\alpha=\frac{n_1-n_2+2m+k_1+k_2+1}{2},\quad \beta=\frac{n_2-n_1+2m+k_1+k_2+1}{2}.\end{gathered}\tag{54}$$

Since the parameters in (54) in general are not integers or half-integers, the coefficients of interbasis expansion (51) may be considered as analytic continuation, for real values of their arguments, of the SU(2) Clebsch–Gordan coefficients. Note also that the inverse expansion of (52) follows from the orthonormality of SU(2) Clebsch–Gordan coefficients.

ACKNOWLEDGMENT

The authors thank P. Winternitz for very interesting discussions.

[1] Yu. N. Demkov, ''Symmetry Group of the Isotropic Oscillator,'' Sov. Phys. JETP **26**, 757 (1954); **36**, 63–66 (1959).
[2] E. G. Kalnins, W. Miller, Jr., and P. Winternitz, ''The Group O(4), Separation of Variables and the Hydrogen Atom,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **30**, 630–664 (1976).
[3] N. W. Evans, ''Superintegrability in Classical Mechanics,'' Phys. Rev. A **41**, 5666–5676 (1990).
[4] J. Friš, V. Mandrosov, Ya. A. Smorodinsky, M. Uhlir, and P. Winternitz, ''On Higher Symmetries in Quantum Mechanics,'' Phys. Lett. **16**, 354–356 (1965).
[5] J. Friš, Ya. A. Smorodinskii, M. Uhlir, and P. Winternitz, ''Symmetry Groups in Classical and Quantum Mechanics,'' Sov. J. Nucl. Phys. **4**, 444–450 (1967).
[6] A. A. Makarov, J. A. Smorodinsky, Kh. Valiev, and P. Winternitz, ''A Systematic Search for Nonrelativistic Systems with Dynamical Symmetries,'' Nuovo Cimento A **52**, 1061–1084 (1967).
[7] N. W. Evans, ''Super-Integrability of the Winternitz System,'' Phys. Lett. A **147**, 483–486 (1990).
[8] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, ''Path Integral Discussion for Smorodinsky-Winternitz Potentials: I. Two- and Three-Dimensional Euclidean Space,'' Fortschr. Phys. **43**, 453–521 (1995).
[9] N. W. Evans, ''Group Theory of the Smorodinsky-Winternitz System,'' J. Math. Phys. **32**, 3369–3375 (1991).
[10] P. Letourneau and L. Vinet, ''Superintegrable Systems: Polynomial Algebras and Quasi-Exactly Solvable Hamiltonians,'' Ann. Phys. **243**, 144–168 (1995).
[11] O. F. Gal'bert, Ya. I. Granovskii, and A. S. Zhedanov, ''Dynamical Symmetry of Anisotropic Singular Oscillator,'' Phys. Lett. A **153**, 177–180 (1991); ''Quadratic Algebra as a 'Hidden' Symmetry of the Hartmann Potential,'' J. Phys. A **24**, 3887–3894 (1991).
[12] M. Victoria Carpio-Bernido, ''Path integral quantization of certain noncentral systems with dynamical symmetries,'' J. Math. Phys. **32**, 1799–1807 (1991).
[13] M. Kibler, L. G. Mardoyan, and G. S. Pogosyan, ''On a Generalized Oscillator System: Interbasis Expansions,'' Int. J. Quantum Chem. **63**, 133–147 (1997); ''On a Generalized Kepler-Coulomb System: Interbasis Expansions,'' *ibid.* **52**, 1301–1316 (1994).
[14] C. P. Boyer, E. G. Kalnins, and P. Winternitz, ''Completely integrable relativistic Hamiltonian systems and separation of variables in Hermitean hyperbolic spaces,'' J. Math. Phys. **24**, 2022 (1983).
[15] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, ''Path Integral Discussion for Smorodinsky-Winternitz Potentials: II. The Two- and Three-Dimensional Sphere,'' Fortschr. Phys. **43**, 523–563 (1995); ''Path Integral Discussion for Smorodinsky-Winternitz Potentials: III. The Two-Dimensional Hyperboloid,'' Phys. Part. Nuclei **27**, 593–674 (1996).
[16] M. A. del Olmo, M. A. Rodriguez, and P. Winternitz, ''The conformal group SU(2,2) and integrable systems on a Lorentzian hyperboloid,'' Fortschr. Phys. **44**, 199–233 (1996).
[17] E. G. Kalnins, W. Miller, Jr., and G. S. Pogosyan, ''Superintegrability and associated polynomial solutions. Euclidean space and the sphere in two dimensions,'' J. Math. Phys. **37**, 6439–6467 (1996).
[18] E. G. Kalnins, W. Miller, Jr., and G. S. Pogosyan, ''Superintegrability on the two-dimensional hyperboloid,'' J. Math. Phys. **38**, 5416–5433 (1997).
[19] E. G. Kalnins, W. Miller, Jr., and G. S. Pogosyan, ''Superintegrability on two dimensional complex Euclidean space,'' Research Report No. 53, Series II (1997).
[20] M. Gaudin, *La fonction d'onde de Bethe* (Masson, Paris, 1983).
[21] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. II.
[22] L. G. Mardoyan, G. S. Pogosyan, A. N. Sissakian, and V. M. Ter-Antonyan, ''Spheroidal Analysis of Hydrogen Atom,'' J. Phys. A **16**, 711–718 (1983).
[23] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).

Bethe ansatz solution of a closed spin 1 *XXZ* Heisenberg chain with quantum algebra symmetry

Jon Links[a)]
Department of Mathematics, University of Queensland, Queensland, 4072, Australia

Angela Foerster[b)] and Michael Karowski[c)]
Institut für Theoretische Physik, Freie Universtät Berlin, Arnimmallee 14, Berlin, Germany

(Received 19 May 1998; accepted for publication 8 June 1998)

A quantum algebra invariant integrable closed spin 1 chain is introduced and analyzed in detail. The Bethe ansatz equations as well as the energy eigenvalues of the model are obtained. The highest weight property of the Bethe vectors with respect to $U_q(sl(2))$ is proved. © *1999 American Institute of Physics.* [S0022-2488(98)01512-6]

I. INTRODUCTION

The quantum inverse scattering method (QISM) has proved to be a powerful procedure in the analysis of one-dimensional integrable quantum chains or two-dimensional lattice models of statistical mechanics (e.g., see Ref. 1). Central to this formalism is the Yang–Baxter equation whose solutions are sufficient to guarantee integrability of the associated model. The advent of quantum algebras[2,3] provided a systematic treatment for obtaining solutions of the Yang–Baxter equation. However, the most common approach to the QISM, which was to impose periodic boundary conditions, was quickly realized to be incompatible with the quantum algebra symmetry of the model. Several authors were able to overcome this problem by working with a model on an open chain.[4–8] This practice made the Bethe ansatz solutions of such models more difficult and in some instances only postulated solutions are available.[9]

More recently, it has been demonstrated that it is in fact possible to construct closed chain models with preservation of quantum algebra symmetry.[10–13] Significantly, the $U_q(sl(2))$ invariant closed spin $\frac{1}{2}$ XXZ model was shown to be connected with a lattice quantization of the Liouville model.[14] The algebraic Bethe ansatz solutions of such models showed that the closed chain quantum algebra invariant case was not fraught with the same difficulties that were faced in the instances of open chains. The existence of such symmetry makes available results such as the highest weight property of the Bethe states for the fundamental representation of $U_q(sl(n))$, and furthermore a characterization of ''good'' and ''bad'' states in terms of q-dimensions when q takes values of roots of unity.[15] Initially, just quantum algebra invariant closed chains of the Hecke algebra type were analyzed. It was subsequently shown[16] that a more general formulation existed.

Here we wish to expand on the knowledge of closed chain quantum algebra invariant models by undertaking a detailed study of the $U_q(sl(2))$ invariant spin 1 model. Integrable spin 1 models built from a $U_q(sl(2))$ invariant R-matrix have already been the subject of some analysis.[17,18] Our study of the closed chain $U_q(sl(2))$ invariant spin 1 model exposes new mathematical aspects not present in the previously studied models;[10–12] viz. the model is not of Hecke algebra type and it is an example of a higher spin system where the most natural approach to the Bethe ansatz solution is to use a transfer matrix defined on an auxiliary space different from the local quantum space. We then find the eigenvalues of the transfer matrix whose auxiliary space is isomorphic to the local quantum space following the method of Babujian and Tsvelick.[19] The need to use two transfer matrices defined on different auxiliary spaces means working with more than one solution

[a)]Electronic mail: jrl@maths.uq.oz.au
[b)]Electronic mail: angela@if.ufrgs.br
[c)]Electronic mail: karowski@physik.fu-berlin.de

of the Yang–Baxter equation. Throughout we will follow the notation of[19,20] in distinguishing the spaces on which the various operators act. Specifically, we use the symbol σ to denote action on the spin $\frac{1}{2}$ space and s for action on the spin 1 space.

The paper is organized as follows. In Sec. II we define some basic quantities, e.g., R matrices, monodromy, and transfer matrices. A quantum algebra invariant closed spin-1 chain is introduced and its relation with one of the transfer matrices is presented. In Sec. III the system is analyzed through a combination of the techniques developed to handle with quantum algebra invariant closed chains[10] and higher-spin chains[18] and the Bethe ansatz equations as well as the energy eigenvalues of the model are obtained. In Sec. IV we show that the Bethe vectors are highest weight vectors with respect to $U_q(sl(2))$. We also argue that use of the $U_q(sl(2))$ generators allows use to generate a complete set of states for the model. A summary of our main results is presented in Sec. V.

II. THE MODEL

We begin by recalling the R-matrix for the spin-$\frac{1}{2}$ chain

$$ {}_{\sigma\sigma}R^{\beta_1\beta_2}_{\alpha_1\alpha_2}(x) = $$

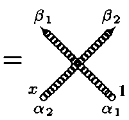

$$ = \frac{1}{{}_{\sigma\sigma}a}\left(\begin{array}{cc|cc} {}_{\sigma\sigma}a & 0 & 0 & 0 \\ 0 & {}_{\sigma\sigma}b & {}_{\sigma\sigma}c_- & 0 \\ \hline 0 & {}_{\sigma\sigma}c_+ & {}_{\sigma\sigma}b & 0 \\ 0 & 0 & 0 & {}_{\sigma\sigma}a \end{array}\right), \tag{1} $$

with

$$ {}_{\sigma\sigma}a = xq - \frac{1}{xq}, \quad {}_{\sigma\sigma}b = x - \frac{1}{x}, \quad {}_{\sigma\sigma}c_+ = x\left(q - \frac{1}{q}\right), \quad {}_{\sigma\sigma}c_- = \frac{1}{x}\left(q - \frac{1}{q}\right), \tag{2} $$

which acts in the tensor product of two two-dimensional auxiliary spaces $\mathbf{C}^2 \otimes \mathbf{C}^2$. Above α_1, α_2 (β_1 and β_2) are column (row) indices running from 1 to 2.

For the spin-1 chain the R-matrix is given by[21,22]

$$ {}_{ss}R^{j_1 j_2}_{i_1 i_2}(x) = $$

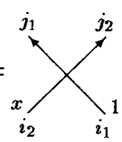

$$ = \frac{1}{{}_{ss}g}\left(\begin{array}{ccc|ccc|ccc} {}_{ss}g & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & {}_{ss}a & 0 & {}_{ss}c_- & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & {}_{ss}b & 0 & {}_{ss}d_- & 0 & {}_{ss}e_- & 0 & 0 \\ \hline 0 & {}_{ss}c_+ & 0 & {}_{ss}a & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & {}_{ss}d_+ & 0 & {}_{ss}f & 0 & {}_{ss}d_- & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & {}_{ss}a & 0 & {}_{ss}c_- & 0 \\ \hline 0 & 0 & {}_{ss}e_+ & 0 & {}_{ss}d_+ & 0 & {}_{ss}b & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & {}_{ss}c_+ & 0 & {}_{ss}a & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & {}_{ss}g \end{array}\right), \tag{3} $$

where

$$
\begin{aligned}
&{}_{ss}g=xq^2-\frac{1}{xq^2},\quad {}_{ss}a=x-\frac{1}{x},\quad {}_{ss}b=\left(x-\frac{1}{x}\right)\left(\frac{x^2-q^2}{x^2q^2-1}\right),\\
&{}_{ss}c_-=\frac{1}{x}\,{}_{ss}c,\quad {}_{ss}c_+=x\,{}_{ss}c,\quad {}_{ss}c,=\left(q^2-\frac{1}{q^2}\right),\quad {}_{ss}f={}_{ss}a+{}_{ss}e,\\
&{}_{ss}d_-=\frac{1}{x}\,{}_{ss}d,\quad {}_{ss}d_+=x\,{}_{ss}d,\quad {}_{ss}d=\left(\frac{xq}{x^2q^2-1}\right)\left(x-\frac{1}{x}\right)\left(q^2-\frac{1}{q^2}\right),\\
&{}_{ss}e_-=\frac{1}{x^2}\,{}_{ss}e,\quad {}_{ss}e+=x^2\,{}_{ss}e,\quad {}_{ss}e=\left(\frac{xq}{x^2q^2-1}\right)\left(q-\frac{1}{q}\right)\left(q^2-\frac{1}{q^2}\right),
\end{aligned}
\tag{4}
$$

and it acts in $\mathbf{C}^3\otimes\mathbf{C}^3$, with $\mathbf{C}^3$ a three-dimensional auxiliary space.

For later convenience we also introduce an R-matrix which acts on $\mathbf{C}^2\otimes\mathbf{C}^3$ (Ref. 22)

$${}_{\sigma s}R^{\beta j}_{\alpha i}(x)=$$
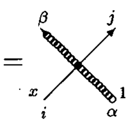

$$
=\frac{1}{{}_{\sigma s}a}\left(\begin{array}{ccc|ccc}
{}_{\sigma s}a & 0 & 0 & 0 & 0 & 0\\
0 & {}_{\sigma s}b & 0 & {}_{\sigma s}d_- & 0 & 0\\
0 & 0 & {}_{\sigma s}c & 0 & {}_{\sigma s}d_- & 0\\
\hline
0 & {}_{\sigma s}d_+ & 0 & {}_{\sigma s}c & 0 & 0\\
0 & 0 & {}_{\sigma s}d_+ & 0 & {}_{\sigma s}b & 0\\
0 & 0 & 0 & 0 & 0 & {}_{\sigma s}a
\end{array}\right),
\tag{5}
$$

where

$$
\begin{aligned}
&{}_{\sigma s}a=xq^{3/2}-\frac{1}{xq^{3/2}},\quad {}_{\sigma s}b=xq^{1/2}-\frac{1}{xq^{1/2}},\quad {}_{\sigma s}c=\frac{x}{q^{1/2}}-\frac{q^{1/2}}{x},\quad {}_{\sigma s}d_-=\frac{1}{x}\,{}_{\sigma s}d,\\
&{}_{\sigma s}d_+=x\,{}_{\sigma s}d,\quad {}_{\sigma s}d=\sqrt{\left(q-\frac{1}{q}\right)\left(q^2-\frac{1}{q^2}\right)}.
\end{aligned}
\tag{6}
$$

These R-matrices satisfy the following properties:

(i) Yang–Baxter equations

$$R^{\alpha''\beta''}_{\alpha'\beta'}(x/y)R^{\alpha'\gamma''}_{\alpha\gamma'}(x)R^{\beta'\gamma'}_{\beta\gamma}(y)=R^{\beta''\gamma''}_{\beta'\gamma'}(y)R^{\alpha''\gamma'}_{\alpha'\gamma}(x)R^{\alpha'\beta'}_{\alpha\beta}(x/y). \tag{7}$$

(ii) Generalized Cherednik reflection property[23]

$$R^{\alpha\beta}_{\alpha'\beta'}(x)(R^{-1})^{\alpha'\beta'}_{\gamma\delta}(y^{-1})=R^{\alpha\beta}_{\alpha'\beta'}(y)(R^{-1})^{\alpha'\beta'}_{\gamma\delta}(x^{-1}). \tag{8}$$

(iii) Crossing unitarity[24]

$$(R^{t_1})^{\alpha\beta}_{\alpha'\beta'}(x\eta)K^{\alpha'}_{\alpha''}((R^{-1})^{t_1})^{\alpha''\beta'}_{\gamma'\delta}(x)(K^{-1})^{\gamma'}_{\gamma}=\delta^{\alpha}_{\gamma}\delta^{\beta}_{\delta}, \tag{9}$$

where t_1 denotes matrix transposition in the first space, η is a crossing parameter and $K=K^t$ is the crossing matrix. Explicit forms for K are given below. We remark that Eq. (8) is the natural generalization of Cherednik's reflection property to the case where the R-matrix acts on two nonisomorphic spaces.

Let us now introduce the ''doubled'' monodromy matrix ${}_{ss}\mathcal{U}$

$${}_{ss}\mathcal{U}^{l\{j\}}_{k\{i\}}(x)=$$
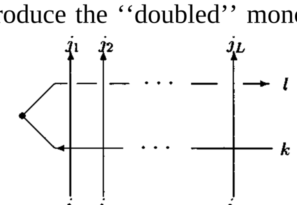

$$={}_{ss}R_{-\,l'\,j_1'}^{\;\,l_2\,j_1}\;{}_{ss}R_{-\,l_2\,j_2'}^{\;\,l_3\,j_2}\cdots{}_{ss}R_{-\,l_L\,j_L'}^{\;\,l\,j_L}\;{}_{ss}R^{l'\,j_1'}_{k_2\,i_1}(1/x)\;{}_{ss}R^{k_2\,j_2'}_{k_3\,i_2}(1/x)\ldots{}_{ss}R^{k_L\,j_L'}_{k\,i_L}(1/x),$$

(10)

which acts in the tensor product of a three-dimensional auxiliary space and a quantum space $\mathbf{C}^3 \otimes \mathbf{C}^{3L}$ and can be regarded as a 3×3 matrix of matrices acting in the quantum space

$$
{}_{ss}\mathcal{U}^l_k(x) = \begin{pmatrix} {}_{ss}\mathcal{U}^1_1 & {}_{ss}\mathcal{U}^1_2 & {}_{ss}\mathcal{U}^1_3 \\ {}_{ss}\mathcal{U}^2_1 & {}_{ss}\mathcal{U}^2_2 & {}_{ss}\mathcal{U}^2_3 \\ {}_{ss}\mathcal{U}^3_1 & {}_{ss}\mathcal{U}^3_2 & {}_{ss}\mathcal{U}^3_3 \end{pmatrix}. \tag{11}
$$

Above the constant ${}_{ss}R$-matrix is defined as

$$
{}_{ss}R_{-} = -\lim_{x\to 0}\ {}_{ss}R^{-1}(x) =
$$

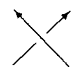

, (12)

For later convenience we also define the auxiliary doubled monodromy matrix

$$
{}_{\sigma s}\mathcal{U}^{\beta\{j\}}_{\alpha\{i\}}(x) =
$$

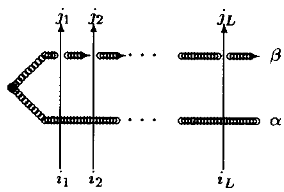

$$
= {}_{\sigma s}R_{-}{}^{\beta_2 j_1}_{\alpha' j_1'}\ {}_{\sigma s}R_{-}{}^{\beta_3 j_2}_{\beta_2 j_2'} \cdots {}_{\sigma s}R_{-}{}^{\beta j_L}_{\beta_L j_L'}\ {}_{\sigma s}R^{\alpha' j_1'}_{\alpha_2 i_1}(1/x)\ {}_{\sigma s}R^{\alpha_2 j_2'}_{\alpha_3 i_2}(1/x) \ldots {}_{\sigma s}R^{\alpha_L j_L'}_{\alpha\, i_L}(1/x)\ , \tag{13}
$$

where ${}_{\sigma s}R_{-}$ corresponds to the leading term in the limit of the matrix ${}_{\sigma s}R^{-1}(x)$ for $x\to 0$, analogously to ${}_{ss}R_{-}$ [see Eq. (12)]. It acts on $\mathbf{C}^2\otimes\mathbf{C}^{3L}$ and can be represented as a 2×2 matrix in the auxiliary space whose entries are matrices acting in the quantum space

$$
{}_{\sigma s}\mathcal{U}^{\beta}_{\alpha}(x) = \begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{14}
$$

Using Eqs. (7) and (8) we can prove that the following Yang–Baxter equations are fulfilled:

$$
\begin{aligned}
&{}_{\sigma\sigma}R^{\alpha\beta}_{\alpha'\beta'}(y/x)\ {}_{\sigma s}\mathcal{U}^{\beta'}_{\gamma'}(x)\ {}_{\sigma\sigma}R_{-}{}^{\alpha'\gamma'}_{\delta'\alpha}\ {}_{\sigma s}\mathcal{U}^{\delta'}_{\delta}(y) \\
&\quad = {}_{\sigma s}\mathcal{U}^{\alpha}_{\alpha'}(y)\ {}_{\sigma\sigma}R^{+}{}^{\alpha'\beta}_{\delta'\beta'}\ {}_{\sigma s}\mathcal{U}^{\beta'}_{\gamma'}(x)\ {}_{\sigma\sigma}(R^{-1})^{\delta'\gamma'}_{\delta\gamma}(x/y),
\end{aligned} \tag{15}
$$

$$
{}_{\sigma s}R^{\alpha i}_{\alpha' i'}(y/x)\ {}_{ss}\mathcal{U}^{i'}_{j'}(x)\ {}_{\sigma s}R_{-}{}^{\alpha' j'}_{\beta' j}\ {}_{\sigma s}\mathcal{U}^{\beta'}_{\beta}(y) = {}_{\sigma s}\mathcal{U}^{\alpha}_{\alpha'}(y)\ {}_{\sigma s}R^{+}{}^{\alpha' i}_{\beta' i'}\ {}_{ss}\mathcal{U}^{i'}_{j'}(x)\ {}_{\sigma s}(R^{-1})^{\beta' j'}_{\beta j}(x/y), \tag{16}
$$

$$
{}_{ss}R^{\alpha i}_{\alpha' i'}(y/x)\ {}_{ss}\mathcal{U}^{i'}_{j'}(x)\ {}_{ss}R_{-}{}^{\alpha' j'}_{\beta' j}\ {}_{ss}\mathcal{U}^{\beta'}_{\beta}(y) = {}_{ss}\mathcal{U}^{\alpha}_{\alpha'}(y)\ {}_{ss}R^{+}{}^{\alpha' i}_{\beta' i'}\ {}_{ss}\mathcal{U}^{i'}_{j'}(x)\ {}_{ss}(R^{-1})^{\beta' j'}_{\beta j}(x/y). \tag{17}
$$

Above we have defined (for R-matrices acting on any two spaces)

$$
R^{+} = \lim_{x\to\infty} R(x).
$$

For later use we also define

$$
R^{-} = \lim_{x\to 0} R(x).
$$

Equation (16) is depicted graphically below. Similar graphical representations apply for Eqs. (15) and (17) but will not be presented.

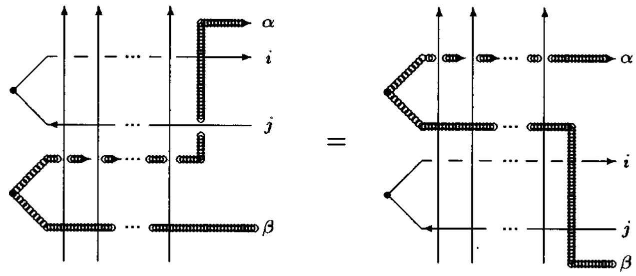

Finally, the spin-1 transfer matrix is constructed by taking the spin-1 Markov trace of the monodromy matrix (10) in the auxiliary space

$${}_{ss}\mathcal{T}^{\{j\}}_{\{i\}}(x)=\sum_{\alpha}\ {}_{s}K^{\alpha}_{\alpha}\ \ {}_{ss}\mathcal{U}^{\alpha\{j\}}_{\alpha\{i\}}\ \ = \tag{18}$$

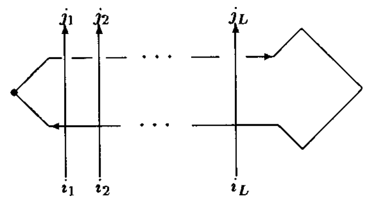

where

$${}_{s}K=\begin{pmatrix} q^{2} & & \\ & 1 & \\ & & q^{-2}\end{pmatrix}. \tag{19}$$

By using Eqs. (9) and (17) it can be shown that this transfer matrix forms a commuting family, i.e., it commutes for different spectral parameters. A quantum algebra invariant spin-1 *XXZ* Hamiltonian with closed boundary conditions will be obtained later from it. However, in order to diagonalize it, the usual algebraic Bethe ansatz scheme which applies to monodromy matrices whose auxiliary space is the fundamental representation cannot be adopted. As in other higher-spin chains[18,20,25–27] this problem can be solved by introducing an auxiliary spin-$\frac{1}{2}$ transfer matrix ${}_{\sigma s}\mathcal{T}$ which commutes with the spin-1 transfer matrix ${}_{ss}\mathcal{T}$. This spin-$\frac{1}{2}$ auxiliary transfer matrix is constructed using the auxiliary ${}_{\sigma s}R(x)$ (5) and doubled monodromy ${}_{\sigma s}\mathcal{U}(x)$ (13) matrices and is given by

$${}_{\sigma s}\mathcal{T}^{\{j\}}_{\{i\}}(x)=\sum_{\alpha}\ {}_{\sigma}K^{\alpha}_{\alpha}\ \ {}_{\sigma s}\mathcal{U}^{\alpha\{j\}}_{\alpha\{i\}}\ \ = \tag{20}$$

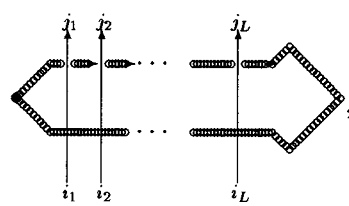

with

$${}_{\sigma}K=\begin{pmatrix} q & \\ & q^{-1}\end{pmatrix}. \tag{21}$$

Using Eqs. (9) and (16) we can also show that the above transfer matrices commute

$$[{}_{\sigma s}\mathcal{T}(x),{}_{ss}\mathcal{T}(y)]=0. \tag{22}$$

Therefore, we can simultaneously diagonalize ${}_{\sigma s}\mathcal{T}$ and ${}_{ss}\mathcal{T}$ which will be presented in the next section.

A quantum algebra invariant closed spin-1 Hamiltonian can be defined through

$$\mathcal{H}={}_{ss}\mathcal{T}'(x)\;{}_{ss}\mathcal{T}^{-1}(x)|_{x=1}, \tag{23}$$

where the prime indicates differentiation with respect to the variable x. This yields (see Ref. 16 for details about this general construction)

$$\mathcal{H}=\sum_{n=1}^{L-1} h_n+h_0, \tag{24}$$

where

$$\begin{aligned} h_n \propto \mathbf{J}_n.\mathbf{J}_{n+1}-(\mathbf{J}_n.\mathbf{J}_{n+1})^2+\frac{(q-q^{-1})^2}{2}[J_n^zJ_{n+1}^z+(J_n^z)^2+(J_{n+1}^z)^2-(J_n^zJ_{n+1}^z)^2] \\ -(q^{1/2}-q^{-1/2})^2[(J_n^xJ_{n+1}^x+J_n^yJ_{n+1}^y)J_n^zJ_{n+1}^z+J_n^zJ_{n+1}^z(J_n^xJ_{n+1}^x+J_n^yJ_{n+1}^y)], \end{aligned} \tag{25}$$

and $\mathbf{J}_n$ are spin-1 generators of $sl(2)$. The boundary term h_0 is given by

$$h_0=\underbrace{{}_{ss}\hat{R}_1^-\;{}_{ss}\hat{R}_2^-\cdots{}_{ss}\hat{R}_{L-1}^-}_{G}\,h_{L-1}\,\underbrace{{}_{ss}\hat{R}_{L-1}^+\cdots{}_{ss}\hat{R}_2^+\;{}_{ss}\hat{R}_1^+}_{G^{-1}}, \tag{26}$$

with

$${}_{ss}\hat{R}_{n\{\beta\}}^{\pm\{\gamma\}}=\mathbf{1}_{\beta_1}^{\gamma_1}\otimes\mathbf{1}_{\beta_2}^{\gamma_2}\otimes\cdots\otimes{}_{ss}R^{\pm\,\gamma_n\gamma_{n+1}}_{\;\;\beta_{n+1}\beta_n}\otimes\cdots\mathbf{1}_{\beta_L}^{\gamma_L},\quad n=1,2,...,L-1. \tag{27}$$

In Eq. (24) L is the number of lattice sites. The operators H, h_n and $\hat{R}_n^{\pm}$ $(n=1,2,...,L-1)$ act on the ''quantum space'' $\mathbf{C}^{3L}$ (for simplicity, we omit the quantum space indices and write them only whenever necessary). The model is periodic in the sense that the operator G^{-1} maps h_n into h_{n-1}

$$G^{-1}h_nG=h_{n-1},\quad n=2,...L-1, \tag{28}$$

and h_1 into h_0

$$G^{-1}h_1G=Gh_{L-1}G^{-1}. \tag{29}$$

The quantum algebra invariance of such a construction is discussed in Ref. 16.

III. BETHE ANSATZ METHOD

In this section we solve the eigenvalue problem of the transfer matrix

$${}_{ss}\mathcal{T}\Psi=(q^2\;{}_{ss}\mathcal{U}_1^1+{}_{ss}\mathcal{U}_2^2+q^{-2}\;{}_{ss}\mathcal{U}_3^3)\Psi={}_{ss}\Lambda\Psi, \tag{30}$$

[and consequently that of the Hamiltonian (24)] through a combination of the techniques developed to handle with quantum group invariant closed chains[10] and higher-spin chains.[18] First, from the fact that Eq. (22) is satisfied, ${}_{ss}\mathcal{T}$ and ${}_{\sigma s}\mathcal{T}$ have a common set of eigenvectors, which can be determined by applying the algebraic Bethe ansatz method to ${}_{\sigma s}\mathcal{T}$(20). Following Babujian,[20] the vector Ψ can be written as

$$\Psi=B(x_1)B(x_2)\cdots B(x_M)\Phi, \tag{31}$$

where Φ is the reference state defined by the equation

$$C\Phi=0,$$

whose solution is $\Phi=\otimes_{i=1}^{L}|1\rangle_i$. It is an eigenstate of A and D

$$A(x)\Phi=q^{3L/2}\Phi,$$

$$D(x)\Phi = q^{-L/2}\,\frac{{}_{\sigma s}c(1/x)^L}{{}_{\sigma s}a(1/x)^L}\,\Phi. \tag{32}$$

Next we apply $A(x)$ and $D(x)$ to Ψ (31), push them through all the $B's$ using the following commutation rules derived from the Yang–Baxter relation (15):

$$A(x)B(y) = \frac{1}{q}\,\frac{{}_{\sigma\sigma}a(x/y)}{{}_{\sigma\sigma}b(x/y)}\,B(y)A(x) - \frac{1}{q}\,\frac{{}_{\sigma\sigma}c_-(x/y)}{{}_{\sigma\sigma}b(x/y)}\,B(x)A(y) - \frac{q-1/q}{q}\,B(x)D(y),$$

$$D(x)B(y) = q\,\frac{{}_{\sigma\sigma}a(y/x)}{{}_{\sigma\sigma}b(y/x)}\,B(y)D(x) - q\,\frac{{}_{\sigma\sigma}c_-(y/x)}{{}_{\sigma\sigma}b(y/x)}\,B(x)D(y), \tag{33}$$

and apply them to the reference state Φ using Eq. (32). From the first terms of the r.h.s. of Eq. (32) we get the ''wanted'' contributions, while the other terms originate the ''unwanted'' terms, since they can never give a vector proportional to Ψ

$$A(x)\Psi = q^{3L/2-M}\prod_{i=1}^{M}\,\frac{{}_{\sigma\sigma}a(x/x_i)}{{}_{\sigma\sigma}b(x/x_i)}\,\Psi + \text{u.t.},$$

$$D(x)\Psi = q^{-L/2+M}\,\frac{{}_{\sigma s}c(1/x)^L}{{}_{\sigma s}a(1/x)^L}\prod_{i=1}^{M}\,\frac{{}_{\sigma\sigma}a(x_i/x)}{{}_{\sigma\sigma}b(x_i/x)}\,\Psi + \text{u.t.} \tag{34}$$

The cancellation of all unwanted terms ensure that Ψ, as given by (31) is an eigenstate of the transfer matrix ${}_{\sigma s}\mathcal{T}(x)$ (20) and this indeed happens if the Bethe ansatz equations (BAE) hold

$$q^{2(1+L-M)}\left(\frac{{}_{\sigma s}a(1/x_k)}{{}_{\sigma s}c(1/x_k)}\right)^L\prod_{i=1}^{M}\,\frac{{}_{\sigma\sigma}a(x_k/x_i)}{{}_{\sigma\sigma}b(x_k/x_i)}\,\frac{{}_{\sigma\sigma}b(x_i/x_k)}{{}_{\sigma\sigma}a(x_i/x_k)} = -1, \quad k=1,\dots,M. \tag{35}$$

Note that these equations are much simpler than those obtained for the quantum group invariant spin-1 chain with open boundary conditions (see Ref. 18). Also in the limit $q\to 1$ we recover the BAE for the usual periodic case.[20]

Let us now find the eigenvalues of ${}_{ss}\mathcal{T}(x)$ by acting with this transfer matrix on Ψ (31), according to (30). For this purpose we need the commutation relations between ${}_{ss}\mathcal{U}_1^1(x)$, ${}_{ss}\mathcal{U}_2^2(x)$, ${}_{ss}\mathcal{U}_3^3(x)$, and ${}_{ss}\mathcal{U}_2^1(y)$ and their action on the reference state Φ. Rewriting the Yang–Baxter equation (16) we can find the following relations:

$${}_{\sigma s}\mathcal{U}^{\alpha}_{\alpha'}(y)\ {}_{\sigma s}R^{+\alpha' i}_{\ \ \beta i'}\ {}_{ss}\mathcal{U}^{i'}_{j}(x) = {}_{\sigma s}R^{\alpha i}_{\alpha' i'}(y/x)\ {}_{ss}\mathcal{U}^{i'}_{j'}(x)\ {}_{\sigma s}R_{-\ \beta' j''}^{\ \alpha' j'}\ {}_{\sigma s}\mathcal{U}^{\beta'}_{\beta''}(y)\ {}_{\sigma s}R^{\beta'' j''}_{\beta j}(x/y), \tag{36}$$

which yield the commutation rules

$${}_{ss}\mathcal{U}_1^1(x)B(y) = \frac{1}{q^2}\,\frac{{}_{\sigma s}a(x/y)}{{}_{\sigma s}c(x/y)}\,B(y)\ {}_{ss}\mathcal{U}_1^1(x) - \frac{1}{q}\,\frac{{}_{\sigma s}d_-(x/y)}{{}_{\sigma s}c(x/y)}\ {}_{ss}\mathcal{U}_2^1(x)A(y)$$
$$- \frac{1}{q}\sqrt{\left(1-\frac{1}{q^2}\right)\left(q^2-\frac{1}{q^2}\right)}\left({}_{ss}\mathcal{U}_2^1(x)D(y) + \frac{{}_{\sigma s}d_-(x/y)}{{}_{\sigma s}c(x/y)}\ {}_{ss}\mathcal{U}_3^1(x)C(y)\right),$$

$${}_{ss}\mathcal{U}_2^2(x)B(y) = \frac{{}_{\sigma s}a(y/x)\ {}_{\sigma s}a(x/y)}{{}_{\sigma s}b(y/x)\ {}_{\sigma s}b(x/y)}\,B(y)\ {}_{ss}\mathcal{U}_2^2(x) - \frac{{}_{\sigma s}d_-(y/x)}{{}_{\sigma s}b(y/x)}\ {}_{ss}\mathcal{U}_2^1(x)D(y)$$
$$- \frac{{}_{\sigma s}d_-(x/y)}{{}_{\sigma s}b(x/y)}\left(\frac{1}{q}\ {}_{ss}\mathcal{U}_3^2(x)A(y) + q\,\frac{{}_{\sigma s}d_-(y/x)}{{}_{\sigma s}b(y/x)}\ {}_{ss}\mathcal{U}_3^1(x)C(y)\right)$$
$$- \sqrt{\left(1-\frac{1}{q^2}\right)\left(q^2-\frac{1}{q^2}\right)}\ {}_{ss}\mathcal{U}_3^2(x)D(y),$$

$$ {}_{ss}\mathcal{U}_3^3(x)B(y)=q^2\,\frac{{}_{\sigma s}a(y/x)}{{}_{\sigma s}c(y/x)}\,B(y)\,{}_{ss}\mathcal{U}_3^3(x)-q^2\,\frac{{}_{\sigma s}d_-(y/x)}{{}_{\sigma s}c(y/x)}\,{}_{ss}\mathcal{U}_3^2(x)D(y). \tag{37} $$

We also observe that

$$ {}_{ss}\mathcal{U}_1^1(x)\Phi=\Phi,\quad {}_{ss}\mathcal{U}_2^2(x)\Phi=q^{-2L}\,{}_{ss}a(1/x)^L\Phi,\quad {}_{ss}\mathcal{U}_3^3(x)\Phi=q^{-4L}\,{}_{ss}b(1/x)^L\Phi. \tag{38} $$

Then applying the transfer matrix ${}_{ss}\mathcal{T}$ on the vector Ψ (31) and using Eqs. (36) and (38) we get the eigenvalues of ${}_{ss}\mathcal{T}$

$$ {}_{ss}\Lambda(x)=q^{2-2M}\prod_{i=1}^{M}\frac{{}_{\sigma s}a(x/x_i)}{{}_{\sigma s}c(x/x_i)}+q^{-2L}\,{}_{ss}a(1/x)^L\prod_{i=1}^{M}\frac{{}_{\sigma s}a(x/x_i)\,{}_{\sigma s}a(x_i/x)}{{}_{\sigma s}b(x/x_i)\,{}_{\sigma s}b(x_i/x)} $$
$$ +q^{2(M-2L-1)}\,{}_{ss}b(1/x)^L\prod_{i=1}^{M}\frac{{}_{\sigma s}a(x_i/x)}{{}_{\sigma s}c(x_i/x)}. \tag{39} $$

We have obtained (39) by taking into account only the first terms on the r.h.s. of Eq. (36). All other terms generate unwanted contributions and the condition of their equality to zero yields the BAE (35). A simpler way to recover the BAE from (39) is by demanding that the eigenvalue ${}_{ss}\Lambda(x)$ (39) has no poles at $x=q^{\pm 1/2}x_i$, since ${}_{ss}\mathcal{T}$ is an analytic function in x. Finally, we obtain the eigenvalues of the Hamiltonian (24) from (23) and (39)

$$ E=\sum_{i=1}^{M}\frac{-2(q^2-q^{-2})}{(x_i^{-1}q^{-1/2}-x_iq^{1/2})(x_i^{-1}q^{3/2}-x_iq^{-3/2})}. \tag{40} $$

In the rational limit $q\to 1$, this expression reduces to that obtained by Babujian[20] for the usual periodic case (with appropriate rescaling).

IV. HIGHEST WEIGHT PROPERTY

In this section we show that the Bethe vectors are highest weight vectors with respect to $U_q(sl(2))$. We begin by writing

$$ {}_{\sigma s}R^+=\begin{pmatrix} q^{1/2h} & 0\\ q^{-1/2}(q-q^{-1})e & q^{-1/2h}\end{pmatrix}, $$
$$ {}_{\sigma s}R_-=\begin{pmatrix} q^{-1/2h} & -q^{1/2}(q-q^{-1})f\\ 0 & q^{1/2h}\end{pmatrix}, \tag{41} $$

where h, e, f are the $sl(2)$ generators in the spin-1 representation. Next, defining the constant auxiliary monodromy matrix as

$$ {}_{\sigma s}\mathcal{U}_\alpha^{-\beta}=\lim_{x\to 0}\,{}_{\sigma s}\mathcal{U}_\alpha^\beta(x)=({}_{\sigma s}R_-)^{\beta j}_{\alpha' j'}({}_{\sigma s}R^+)^{\alpha' j'}_{\alpha i}, \tag{42} $$

we have from (41)

$$ C^-=q^{-1/2}(q-q^{-1})q^{-1/2h}e. \tag{43} $$

The Bethe vectors (31) are highest weight vectors if

$$ C^-\Psi=0. \tag{44} $$

This can be proven by observing that from the Yang–Baxter algebra (15) we can obtain the following relation:

$$ C^-B(x)=B(x)C^-+(1-q^{-2})(A(x)D^--D^-D(x)), \tag{45} $$

which, using the fact that $C^{-}\Phi=0$, allows us to write

$$C^{-}\Psi=\sum_{i=1}^{M} Y_i W_i \Phi \tag{46}$$

where

$$W_i=B(x_1)B(x_2)\cdots B(x_{i-1})B(x_{i+1})\cdots B(x_M). \tag{47}$$

The Y_i can be computed using Eqs. (32) and (33) which yield

$$Y_i=(1-q^{-2})q^{3/2L}\ {}_{\sigma s}a(1/x_i)^L\prod_{j\neq i}^{M} q^{-1}\ \frac{{}_{\sigma\sigma}a(x_i/x_j)}{{}_{\sigma\sigma}b(x_i/x_j)}-(1-q^{-2})q^{-L/2}\ {}_{\sigma s}c(1/x_i)^L\prod_{j\neq i}^{M} q\ \frac{{}_{\sigma\sigma}a(x_j/x_i)}{{}_{\sigma\sigma}b(x_j/x_i)}. \tag{48}$$

Because of the BAE (35), each of the coefficients Y_i vanishes which implies

$$C^{-}\Psi=0.$$

It immediately follows that each of the Bethe states are highest weight states. By using the $U_q(sl(2))$ lowering operator f we obtain additional states which will also be eigenstates of the transfer matrix because of the quantum symmetry of the model.

For generic values of the deformation parameter q it is well known that the dimensions and weight spectrum of the finite dimensional irreducible representations of $U_q(sl(2))$ are in 1-1 correspondence with those of $sl(2)$. Since it is known[26,28] in the $q=1$ case that the Bethe states combined with the $sl(2)$ symmetry give a complete set of states for the model, it should be possible to prove, using methods developed in Ref. 29, that this is also true for the model described above.

V. CONCLUSIONS

We have solved a quantum algebra invariant integrable closed spin-1 chain by an algebraic Bethe ansatz approach. Particularly eigenstates of the model were constructed and their energy eigenvalues evaluated. A proof of the highest weight property of the Bethe vectors with respect to $U_q(sl(2))$ was also presented. A natural extension of this work would be to generalize the results of the spin-1 chain to corresponding chains of arbitrary spin s.

ACKNOWLEDGMENTS

J.L. is supported by an Australian Research Council Postdoctoral Fellowship. A.F. would like to thank H. M. Babujian for useful discussions and the Institute für Theoretische Physik—FUB for its kind hospitality. She also thanks DAAD—Deutscher Akademischer Austauschdienst and FAPERGS—Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul for financial support.

[1] V. E. Korepin, A. G. Izergin, and N. M. Bogoliubov, *Quantum Inverse Scattering Method, Correlation Functions and Algebraic Bethe Ansatz* (Cambridge University Press, Cambridge, 1993).
[2] M. Jimbo, Lett. Math. Phys. **10**, 63 (1985).
[3] V. G. Drinfeld, Proc. Int. Cong. Math. 798 (Berkeley, 1986).
[4] C. Destri and H. J. De Vega, Nucl. Phys. B **361**, 361 (1992); **374**, 692 (1992).
[5] A. González-Ruiz and H. J. De Vega, Nucl. Phys. B **417**, 553 (1994).
[6] A. Foerster and M. Karowski, Nucl. Phys. B **408**, 512 (1993).
[7] A. González-Ruiz, Nucl. Phys. B **424**, 468 (1994).
[8] R. Yue, H. Fan, and B. Hou, Nucl. Phys. B **462**, 167 (1996).
[9] S. Artz, L. Mezincescu, and R. I. Nepomechie, Int. J. Mod. Phys. A **10**, 1937 (1995); J. Phys. A **28**, 5131 (1995).
[10] M. Karowski and A. Zapletal, Nucl. Phys. B **419**, 567 (1994).
[11] H. Grosse, S. Pallua, P. Prester, and E. Raschhofer, J. Phys. A **27**, 4761 (1994).
[12] A. Foerster, J. Phys. A **29**, 7625 (1996).
[13] A. Lima-Santos and R. C. T. Ghiotto, J. Phys. A **31**, 505 (1998).
[14] L. D. Faddeev and O. Tirkkonen, Nucl. Phys. B **453**, 647 (1995).
[15] M. Karowski and A. Zapletal, J. Phys. A **27**, 7419 (1994).

[16] J. Links and A. Foerster, J. Phys. A **30**, 2483 (1996).
[17] M. T. Batchelor, L. Mezincescu, R. I. Nepomechie, and V. Rittenberg, J. Phys. A **23**, L141 (1990).
[18] L. Mezincescu, R. I. Nepomechie, and V. Rittenberg, Phys. Lett. A **147**, 70 (1990).
[19] H. M. Babujian and A. M. Tsvelick, Nucl. Phys. B **265**, 24 (1986).
[20] H. M. Babujian, Nucl. Phys. B **215**, 317 (1983).
[21] A. B. Zamolodchikov and V. A. Fateev, Sov. J. Nucl. Phys. **32**, 298 (1980).
[22] P. P. Kulish and N. Yu Reshetikhin, J. Sov. Math. **23**, 2435 (1983).
[23] I. Cherednik, Theor. Mat. Phys. **61**, 977 (1985).
[24] J. R. Links and M. D. Gould, Int. J. Mod. Phys. B **10**, 3461 (1996).
[25] P. P. Kulish, N. Yu Reshetikhin, and E. K. Sklyanin, Lett. Math. Phys. **5**, 393 (1981).
[26] L. Takhtajan, Phys. Lett. A **87**, 479 (1982).
[27] A. N. Kirillov and N. Yu Reshetikhin, J. Phys. A **20**, 1565 (1987).
[28] A. N. Kirillov, J. Sov. Math. **30**, 2298 (1985).
[29] A. N. Kirillov and N. A. Liskova, J. Phys. A **30**, 1209 (1997); preprint hep-th/9607012.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Deconstructing supersymmetry

N. S. Manton[a)]
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge CB3 9EW, England

(Received 15 June 1998; accepted for publication 9 November 1998)

Two supersymmetric classical mechanical systems are discussed. Concrete realizations are obtained by supposing that the dynamical variables take values in the Grassmann algebra B_2 with two generators. The equations of motion are explicitly solved, and the action of the supergroup on the space of solutions is elucidated. The Lie algebra of the supergroup is the even part of the tensor product $B_2 \otimes \mathcal{G}$, where $\mathcal{G}$ is the super Lie algebra of supersymmetries and time translations. For each system, the solutions with zero energy need to be constructed separately. For these Bogomolny-type solutions, the orbit of the supergroup is smaller than in the generic case. © *1999 American Institute of Physics.* [S0022-2488(99)03302-2]

I. INTRODUCTION

Supersymmetry is one of the most powerful ideas in theoretical physics, combining bosonic and fermionic fields into a unified framework. Most supersymmetric theories are defined by a Lagrangian, from which the classical field equations are derived. However the meaning of the fermionic fields in such equations is not always clear, because they need to be anticommuting. Moreover, there are usually sources for the bosonic fields which are bilinear in the fermionic fields, and such sources are not ordinary functions. So an interpretation of the bosonic fields as ordinary functions fails in supersymmetric theories. It is mathematically consistent to treat bosonic and fermionic fields as ordinary functions, as in classical QED or QCD, but such a treatment cannot be supersymmetric.

It might be thought that only the quantized versions of supersymmetric theories really make sense. This is not so. The formalism for making sense of classical supersymmetric theories is readily available, but perhaps not sufficiently appreciated by theoretical physicists. As emphasized by De Witt,[1] and also by Freund,[2] fields in a supersymmetric field theory must take their values in a Grassmann algebra B. B is the direct sum of an even part B_e and an odd part B_o. The bosonic fields are valued in B_e, and the fermionic fields in B_o. It is necessary to decide which algebra B to work with. B can have a finite number, n, of generators, or an infinite number, and the content of the classical theory will depend on the choice. With n generators, a scalar bosonic field is represented by 2^{n-1} ordinary functions, and by an infinite number if B is infinitely generated. This is rather daunting, and it is not clear what all these functions might mean physically. However, we shall choose $n=2$ in what follows, and the resulting equations are quite manageable. (The choice $n=1$ leads to trivial equations.) Choosing n to be as small as possible means that the theory, while including fermions, is as close as possible to the underlying classical, purely bosonic, theory. It is sometimes argued that choosing n to be finite leads to ambiguities or contradictions. The argument is connected with Green's functions and quantization. In a classical context there appear to be no problems—at least, we encounter no problems in the models discussed here, except one small matter which is dealt with in Sec. II.

Our aim is to present concrete examples of classical supersymmetric theories, and to solve the equations of motion. Rather than considering a genuine field theory, we shall simplify matters by considering supersymmetric mechanical models, whose dynamical variables depend only on time. Mechanical models, with bosonic and fermionic dynamical variables taking values in a Grassmann

[a)]Electronic mail: N.S.Manton@damtp.cam.ac.uk

algebra, were investigated by Casalbuoni[3] and by Berezin and Marinov,[4] although not solved except in very simple cases. Supersymmetry constrains the structure of such models. The dynamical variables, taken together, define a point in a supermanifold, which traces out a curve as time varies. The mathematically rigorous treatment of the geometry of dynamical systems in supermanifolds is tricky, and there is more than one point of view. For discussions of functions and their derivatives on a supermanifold, and of the topology and geometry of supermanifolds, the reader may turn to a number of papers, e.g. Refs. 5 and 6, as well as Ref. 1. We will not use supermanifold theory in any significant way here.

We analyze two supersymmetric mechanical models below. We present the Lagrangian and equations of motion, their symmetries, and the associated conserved quantities, and proceed to find the explicit form of the general solution of the equations of motion. We believe that this has not been done before. The possibility of constructing general solutions of the nonlinear coupled ordinary differential equations (ODEs) shows the power of the supersymmetry of these models. From the super Lie algebra of supersymmetries and time translations we construct, in a standard way, a genuine Lie algebra of infinitesimal symmetries, which generates a genuine Lie group of symmetries of the dynamics—the supergroup. Technically, if n is finite, the supergroup is the nth skeleton of the full supergroup. The solutions depend on a number of constants of integration, and we comment on the extent to which, when $n=2$, the supergroup relates solutions with different values of these constants.

For each of these models, the solutions with zero energy need to be constructed independently. Here, one of the bosonic equations of motion reduces to a first-order Bogomolny-type equation.[7] The solution space is still acted on by the supergroup, but the orbit is of lower dimension than in the generic case. This feature of Bogomolny equations is not unfamiliar, but the complete solution of the equations of motion, including the fermionic variables, is perhaps novel.

Section II discusses the $N=2$ supersymmetric mechanics of a particle moving in one dimension, subject to a potential. The model is a variant of the one whose quantized version was analyzed by Witten.[8] Section III is concerned with the zero energy, Bogomolny case. Section IV discusses the $N=1$ supersymmetric mechanics of a particle moving in one dimension. Again the model is a variant of the standard one, as the Lagrangian depends on a constant odd parameter. We conclude in Sec. V with some comments on the analysis, and on potential generalizations of this work.

II. $N=2$ SUPERSYMMETRIC MECHANICS

Consider the following $N=2$ supersymmetric Lagrangian (Refs. 8, 1—Sec. 5.7):

$$L=\tfrac{1}{2}\dot{x}^2+\tfrac{1}{2}U(x)^2+\tfrac{1}{2}\dot{\psi}_1\psi_1-\tfrac{1}{2}\dot{\psi}_2\psi_2+U'(x)\psi_1\psi_2, \tag{2.1}$$

where the dynamical variables x, ψ_1, and ψ_2 take values in an arbitrary Grassmann algebra B. This describes the supersymmetric mechanics of a particle moving in one dimension in a potential $-U^2$. $x(t)$ is bosonic (i.e., commuting) and $\psi_1(t)$ and $\psi_2(t)$ are fermionic (i.e., anticommuting) variables. Thus x is valued in B_e, whereas ψ_1 and ψ_2 are valued in B_o. Any function of x, e.g., $U(x)$, commutes with x. Such functions are defined as polynomials or power series with real coefficients. If $U(x)=x^p$, with p a positive integer, then $U'(x)=px^{p-1}$, with the obvious extension to polynomials and power series. An overdot denotes the derivative with respect to time t. $\dot{x}$ commutes with x, and similarly, $\dot{\psi}_1$ and $\dot{\psi}_2$ anticommute with both ψ_1 and ψ_2; hence the dynamics is classical, rather than quantized. Note that the terms $\dot{\psi}_1\psi_1$ and $\dot{\psi}_2\psi_2$ are not total time derivatives.

The Lagrangian L may be obtained by dimensional reduction of the $1+1$-dimensional $N=1$ supersymmetric field theory with Lagrangian density

$$\mathcal{L}=\frac{1}{2}\partial_+\Phi\partial_-\Phi-\frac{1}{2}U(\Phi)^2+\frac{i}{2}\psi_1\partial_-\psi_1+\frac{i}{2}\psi_2\partial_+\psi_2+i\frac{dU}{d\Phi}\psi_1\psi_2, \tag{2.2}$$

where ∂_+ and ∂_- are the standard light cone derivatives. By assuming that all fields are independent of the spatial coordinate, then absorbing certain factors of $\sqrt{i}$, etc., in the fields and potential, and finally writing Φ as x, we recover the expression (2.1). The density (2.2) is real in a certain sense,[1] but for our purposes the manifestly real expression (2.1) is a more convenient Lagrangian to discuss.

To obtain the equations of motion we calculate the formal variation ΔL due to variations Δx, $\Delta\psi_1$, and $\Delta\psi_2$. We combine $\Delta\dot{x}$, $\Delta\dot{\psi}_1$, and $\Delta\dot{\psi}_2$ into total time derivative terms, which are ignored, then move Δx, $\Delta\psi_1$, and $\Delta\psi_2$ to the left in each term. The result is

$$\Delta L=\Delta x(-\ddot{x}+UU'+U''\psi_1\psi_2)+\Delta\psi_1(-\dot{\psi}_1+U'\psi_2)+\Delta\psi_2(\dot{\psi}_2-U'\psi_1), \tag{2.3}$$

so the equations of motion are

$$\ddot{x}=UU'+U''\psi_1\psi_2, \tag{2.4a}$$

$$\dot{\psi}_1=U'\psi_2, \tag{2.4b}$$

$$\dot{\psi}_2=U'\psi_1. \tag{2.4c}$$

There is a possible ambiguity here. ΔL should be zero for all Δx in B_e and all $\Delta\psi_1$, $\Delta\psi_2$ in B_o. If B has an infinity of generators, this implies Eqs. (2.4). However, if B has n generators, then the coefficients of $\Delta\psi_1$ and $\Delta\psi_2$ need not be zero, but could be an element of highest degree in B, with arbitrary time dependence. This possibility can be excluded by careful definition of odd derivatives.[9] More trivially, below we shall assume that n is even $(n=2)$, and then this possibility is excluded because these coefficients are constrained to be odd.

There are two supersymmetry operators $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$. Together with d/dt they are a basis for a super Lie algebra over the real numbers with nontrivial relations

$$\mathcal{Q}^2=\frac{d}{dt}, \quad \tilde{\mathcal{Q}}^2=-\frac{d}{dt}, \quad \mathcal{Q}\tilde{\mathcal{Q}}+\tilde{\mathcal{Q}}\mathcal{Q}=0. \tag{2.5}$$

Formally, the algebra has a representation on the dynamical variables

$$\mathcal{Q}x=\psi_1, \quad \mathcal{Q}\psi_1=\dot{x}, \quad \mathcal{Q}\psi_2=U, \tag{2.6a}$$

$$\tilde{\mathcal{Q}}x=\psi_2, \quad \tilde{\mathcal{Q}}\psi_2=-\dot{x}, \quad \tilde{\mathcal{Q}}\psi_1=-U, \tag{2.6b}$$

with d/dt acting in the obvious way. This ''on-shell'' representation requires that the equations $\dot{\psi}_1=U'\psi_2$ and $\dot{\psi}_2=U'\psi_1$ are satisfied, so that, e.g., $\mathcal{Q}^2\psi_2=\mathcal{Q}U=U'(\mathcal{Q}x)=U'\psi_1=\dot{\psi}_2$. $\mathcal{Q}$, $\tilde{\mathcal{Q}}$, and d/dt are all symmetries of the Lagrangian, provided $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ are treated as antiderivations [an extra minus sign in the Leibniz rule when $\mathcal{Q}$ or $\tilde{\mathcal{Q}}$ goes past a fermionic variable, e.g., $\mathcal{Q}(\psi_1\psi_2)=(\mathcal{Q}\psi_1)\psi_2-\psi_1\mathcal{Q}\psi_2=\dot{x}\psi_2-\psi_1 U$]. Now although the actions of $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ given by (2.6) make formal sense, they cannot be regarded as variations of the dynamical variables. A bosonic variable cannot be varied by a fermionic one. Moreover, the vague requirement that the coefficients of $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ should be anticommuting, common in the literature, is not sufficiently precise. However, genuine variations are obtained by taking the coefficients of $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ to be arbitrary infinitesimal elements of B_o, and the coefficient of d/dt to be an infinitesimal element of B_e.

The super Lie algebra over the real numbers becomes an ordinary Lie algebra if the coefficients lie in B, as above. This construction is, in fact, well known.[2,10] If one takes a super Lie algebra $\mathcal{G}$ and tensors it with a Grassmann algebra of coefficients B, and keeps just the even part

$$(B\otimes\mathcal{G})_e=(B_e\otimes\mathcal{G}_e)\oplus(B_o\otimes\mathcal{G}_o), \tag{2.7}$$

then this is an ordinary Lie algebra over the reals, if the bracket is defined by

$$[b_1 g_1, b_2 g_2] = b_1 b_2 [g_1, g_2](-1)^{|b_2||g_1|}, \tag{2.8}$$

where $|\ |$ denotes the grade (0 for even elements, 1 for odd), and $[g_1,\ g_2]$ is the bracket in the super Lie algebra. The Lie group generated by this Lie algebra is the supergroup, and is the concrete symmetry group of our mechanical system.

Thus, the variations of the dynamical quantities generated by the supersymmetry operator $\mathcal{Q}$ are

$$\delta x = \eta \psi_1, \quad \delta\psi_1 = \eta \dot{x}, \quad \delta\psi_2 = \eta U, \tag{2.9}$$

where η is an arbitrary infinitesimal constant in B_o. It is easily shown that the variation of the Lagrangian L is a total time-derivative

$$\delta L = \eta \frac{d}{dt}\left(\frac{1}{2}\dot{x}\psi_1 + \frac{1}{2}U\psi_2\right) \tag{2.10}$$

using $\dot{U} = U'\dot{x}$. The usual Noether method gives the conserved supersymmetry charge

$$Q = \dot{x}\psi_1 - U\psi_2. \tag{2.11}$$

The conservation of Q is easily verified using the equations of motion:

$$\dot{Q} = \ddot{x}\psi_1 + \dot{x}\dot{\psi}_1 - U'\dot{x}\psi_2 - U\dot{\psi}_2 = UU'\psi_1 + U''\psi_1\psi_2\psi_1 + \dot{x}U'\psi_2 - U'\dot{x}\psi_2 - UU'\psi_1 = 0 \tag{2.12}$$

since $\psi_1\psi_2\psi_1 = -\psi_1\psi_1\psi_2 = 0$. The variations generated by the second supersymmetry operator $\tilde{\mathcal{Q}}$ are

$$\tilde{\delta} x = \eta\psi_2, \quad \tilde{\delta}\psi_2 = -\eta\dot{x}, \quad \tilde{\delta}\psi_1 = -\eta U, \tag{2.13}$$

and lead to the conserved supersymmetry charge

$$\tilde{Q} = \dot{x}\psi_2 - U\psi_1. \tag{2.14}$$

The supersymmetries relate different solutions of the equations of motion. To see this, consider the linearized variations of Eqs. (2.4),

$$(\Delta \ddot{x}) = (UU')'\Delta x + U'''\Delta x\psi_1\psi_2 + U''\Delta\psi_1\psi_2 + U''\psi_1\Delta\psi_2, \tag{2.15a}$$

$$(\Delta\dot{\psi}_1) = U''\Delta x\psi_2 + U'\Delta\psi_2, \tag{2.15b}$$

$$(\Delta\dot{\psi}_2) = U''\Delta x\psi_1 + U'\Delta\psi_1, \tag{2.15c}$$

and assume that x, ψ_1, and ψ_2 satisfy (2.4). The linear equations (2.15) are satisfied by setting $\Delta = \delta$ or $\Delta = \tilde{\delta}$, and using the variations defined in (2.9) and (2.13). Later, we shall see more concretely, and not just in the linearized approximation, how supersymmetry relates different solutions.

Since the Lagrangian (2.1) does not depend explicitly on time, we expect a conserved energy, associated with time translation symmetry. This symmetry is defined by the infinitesimal variations

$$\Delta x = \eta\dot{x}, \quad \Delta\psi_1 = \eta\dot{\psi}_1, \quad \Delta\psi_2 = \eta\dot{\psi}_2, \tag{2.16}$$

where η is now an arbitrary infinitesimal element of B_e. The energy is

$$H=\tfrac{1}{2}\dot{x}^2-\tfrac{1}{2}U^2-U'\psi_1\psi_2, \tag{2.17}$$

and its conservation is easily checked using the equations of motion.

In simple cases, it is possible to solve the equations of motion for any choice of the Grassmann algebra B, and without explicit reference to B. An example is when $U(x)=\omega x$, with ω a real constant, so $U'=\omega$ and $U''=0$, and the equations simplify. This is the analog of the supersymmetric harmonic oscillator, discussed and solved by De Witt,[1] but in our case the potential is inverted. The solution can be expressed in terms of initial data as

$$x(t)=x(0)\cosh\omega t+\frac{1}{\omega}\dot{x}(0)\sinh\omega t, \tag{2.18a}$$

$$\psi_1(t)=\psi_1(0)\cosh\omega t+\psi_2(0)\sinh\omega t, \tag{2.18b}$$

$$\psi_2(t)=\psi_2(0)\cosh\omega t+\psi_1(0)\sinh\omega t. \tag{2.18c}$$

We have not been able to solve the equations of motion (2.4) for general $U(x)$ and an arbitrary choice of the Grassmann algebra B, and in particular for B having an infinity of generators. It would be interesting to do so. If B has just one generator, then the equations are rather trivial, because $\psi_1\psi_2$ vanishes. So the simplest nontrivial case, which we discuss from here on, is where B is generated by just two elements α,β satisfying

$$\alpha^2=0,\quad \beta^2=0,\quad \alpha\beta+\beta\alpha=0. \tag{2.19}$$

A basis for the algebra is $\{1,\alpha,\beta,\alpha\beta\}$, and it follows from (2.19) that $(\alpha\beta)^2=0$. There is a matrix realization of these relations, although we will not use it. Let $\{\gamma^\mu:1\leqslant\mu\leqslant 4\}$ denote Dirac matrices in four Euclidean dimensions, and set $\alpha=\gamma^1+i\gamma^2$, $\beta=\gamma^3+i\gamma^4$.

Let us write the dynamical variables in component form as

$$x(t)=x_0(t)+x_1(t)\alpha\beta, \tag{2.20a}$$

$$\psi_1(t)=a_1(t)\alpha+b_1(t)\beta, \tag{2.20b}$$

$$\psi_2(t)=a_2(t)\alpha+b_2(t)\beta, \tag{2.20c}$$

where x_0,x_1,a_1,b_1,a_2,b_2 are ordinary functions of time. The ''body,'' $x_0(t)$, can be regarded as classical, but its partner $x_1(t)$ is a less familiar quantity.

Any positive power of x now has the truncated expansion

$$x^n=x_0^n+nx_0^{n-1}x_1\alpha\beta, \tag{2.21}$$

which extends to an arbitrary function of x as

$$U(x)=U(x_0)+U'(x_0)x_1\alpha\beta, \tag{2.22}$$

where $U'(x_0)$ denotes the usual derivative of $U(x_0)$ with respect to x_0. Henceforth, if the argument of U and its derivatives is not shown, it is x_0, with x_0 itself a function of t. The Lagrangian is the even function $L=L_0+L_1\alpha\beta$, where

$$L_0=\tfrac{1}{2}\dot{x}_0^2+\tfrac{1}{2}U^2, \tag{2.23a}$$

$$L_1=\dot{x}_0\dot{x}_1+UU'x_1+\dot{a}_1b_1-\dot{a}_2b_2+U'(a_1b_2-a_2b_1). \tag{2.23b}$$

Substituting (2.20) into (2.4), we obtain the equations of motion for the components

$$\ddot{x}_0 = UU', \tag{2.24a}$$

$$\ddot{x}_1 = (UU')'x_1 + U''(a_1 b_2 - a_2 b_1), \tag{2.24b}$$

$$\dot{a}_1 = U' a_2, \tag{2.24c}$$

$$\dot{a}_2 = U' a_1, \tag{2.24d}$$

$$\dot{b}_1 = U' b_2, \tag{2.24e}$$

$$\dot{b}_2 = U' b_1. \tag{2.24f}$$

These equations can also be derived as the variational equations of L_0 and L_1. In fact, surprisingly, they can all be derived from L_1 alone, as the equation of motion for x_0, obtained from L_0, is the same as the equation obtained from L_1 by varying x_1.

There are a host of symmetries and conservation laws associated with the component form of the system. With B generated by α and β, the supergroup is six dimensional, its Lie algebra having basis elements

$$\mathcal{Q}_\alpha = \alpha \mathcal{Q}, \quad \mathcal{Q}_\beta = \beta \mathcal{Q}, \quad \tilde{\mathcal{Q}}_\alpha = \alpha \tilde{\mathcal{Q}}, \quad \tilde{\mathcal{Q}}_\beta = \beta \tilde{\mathcal{Q}}, \quad \frac{d}{dt}, \quad \frac{\tilde{d}}{dt} = \alpha\beta \frac{d}{dt}, \tag{2.25}$$

where d/dt is the usual time derivative and $\widetilde{d/dt}$ we call the mini-time-derivative. Almost all these generators commute, except that

$$[\mathcal{Q}_\alpha, \mathcal{Q}_\beta] = -2\frac{\tilde{d}}{dt}, \quad [\tilde{\mathcal{Q}}_\alpha, \tilde{\mathcal{Q}}_\beta] = 2\frac{\tilde{d}}{dt}. \tag{2.26}$$

Note that the signs in (2.26) are consistent with (2.8).

Following (2.9), we define two variations δ_α and δ_β, generated by $\mathcal{Q}_\alpha$ and $\mathcal{Q}_\beta$, by

$$\delta_\alpha x = \epsilon \alpha \psi_1, \quad \delta_\alpha \psi_1 = \epsilon \alpha \dot{x}, \quad \delta_\alpha \psi_2 = \epsilon \alpha U(x), \tag{2.27}$$

where ϵ is infinitesimal and real, and δ_β similarly by replacing α by β. In components, the first of these variations becomes

$$\delta_\alpha (x_0 + x_1 \alpha\beta) = \epsilon b_1 \alpha\beta \tag{2.28}$$

so $\delta_\alpha x_0 = 0$ and $\delta_\alpha x_1 = \epsilon b_1$. Similarly, by expanding out, we find the complete set of component variations

$$\delta_\alpha x_1 = \epsilon b_1, \quad \delta_\alpha a_1 = \epsilon \dot{x}_0, \quad \delta_\alpha a_2 = \epsilon U, \tag{2.29a}$$

$$\delta_\beta x_1 = -\epsilon a_1, \quad \delta_\beta b_1 = \epsilon \dot{x}_0, \quad \delta_\beta b_2 = \epsilon U, \tag{2.29b}$$

with all other variations, e.g., $\delta_\beta a_1$, vanishing. The generators $\tilde{Q}_\alpha$ and $\tilde{Q}_\beta$ lead similarly to the two independent sets of variations

$$\tilde{\delta}_\alpha x_1 = \epsilon b_2, \quad \tilde{\delta}_\alpha a_1 = -\epsilon U, \quad \tilde{\delta}_\alpha a_2 = -\epsilon \dot{x}_0, \tag{2.30a}$$

$$\tilde{\delta}_\beta x_1 = -\epsilon a_2, \quad \tilde{\delta}_\beta b_1 = -\epsilon U, \quad \tilde{\delta}_\beta b_2 = -\epsilon \dot{x}_0. \tag{2.30b}$$

x_0, and hence L_0, is unchanged by all these variations.

It is easy to verify that all four sets of variations δ_α, δ_β, $\tilde{\delta}_\alpha$, $\tilde{\delta}_\beta$ are Noether symmetries of the Lagrangian L_1, giving total time derivatives. For example,

$$\delta_\alpha L_1 = \epsilon(\dot{x}_0\dot{b}_1 + UU'b_1 + \ddot{x}_0 b_1 - U'\dot{x}_0 b_2 + U'(\dot{x}_0 b_2 - Ub_1)) = \epsilon\frac{d}{dt}(\dot{x}_0 b_1). \tag{2.31}$$

In the usual way, we obtain the conserved Noether charges

$$Q_\alpha = \dot{x}_0 b_1 - Ub_2, \tag{2.32a}$$

$$Q_\beta = -\dot{x}_0 a_1 + Ua_2, \tag{2.32b}$$

$$\tilde{Q}_\alpha = \dot{x}_0 b_2 - Ub_1, \tag{2.32c}$$

$$\tilde{Q}_\beta = -\dot{x}_0 a_2 + Ua_1, \tag{2.32d}$$

and may verify their conservation using the equations of motion (2.24). Of course, these charges are just the components of the supersymmetry charges we found earlier, although with labels switched, namely

$$Q = -Q_\beta\alpha + Q_\alpha\beta, \tag{2.33a}$$

$$\tilde{Q} = -\tilde{Q}_\beta\alpha + \tilde{Q}_\alpha\beta. \tag{2.33b}$$

Both L_0 and L_1 are invariant under ordinary time translations, leading to the conservation of two energies

$$H_0 = \tfrac{1}{2}\dot{x}_0^2 - \tfrac{1}{2}U^2, \tag{2.34a}$$

$$H_1 = \dot{x}_0\dot{x}_1 - UU'x_1 - U'(a_1 b_2 - a_2 b_1). \tag{2.34b}$$

The conserved energy we found earlier is $H = H_0 + H_1\alpha\beta$.

There is a further symmetry, the mini-time-translation symmetry. The variations generated by $\widetilde{d/dt}$ are

$$\Delta x = \epsilon\alpha\beta\dot{x}, \quad \Delta\psi_1 = \epsilon\alpha\beta\dot{\psi}_1, \quad \Delta\psi_2 = \epsilon\alpha\beta\dot{\psi}_2, \tag{2.35}$$

with ϵ real. Expanding out in components, we find a single nonzero variation

$$\Delta x_1 = \epsilon\dot{x}_0. \tag{2.36}$$

The associated variation of L_1 is the total time derivative

$$\Delta L_1 = \epsilon(\dot{x}_0\ddot{x}_0 + UU'\dot{x}_0) = \epsilon\frac{d}{dt}\left(\frac{1}{2}\dot{x}_0^2 + \frac{1}{2}U^2\right), \tag{2.37}$$

and the conserved quantity is

$$\tfrac{1}{2}\dot{x}_0^2 - \tfrac{1}{2}U^2, \tag{2.38}$$

which is H_0. So we see that the equations of motion and both conserved energies, and all four components of the supersymmetry charges, can be derived from L_1.

This exhausts the infinitesimal action of the supergroup, but there are yet more symmetries which mix the functions a_1, a_2, b_1, b_2. The combined variations

$$\Delta a_1 = \epsilon b_1, \quad \Delta a_2 = \epsilon b_2 \tag{2.39}$$

leave L_1 invariant, as do the variations

$$\Delta b_1 = \epsilon a_1, \quad \Delta b_2 = \epsilon a_2. \tag{2.40}$$

Finally, L_1 is invariant under

$$\Delta a_1 = \epsilon a_2, \quad \Delta a_2 = \epsilon a_1, \quad \Delta b_1 = \epsilon b_2, \quad \Delta b_2 = \epsilon b_1. \tag{2.41}$$

These symmetries imply that

$$R_a = b_1^2 - b_2^2, \tag{2.42a}$$

$$R_b = a_1^2 - a_2^2, \tag{2.42b}$$

$$R = a_1 b_2 - a_2 b_1 \tag{2.42c}$$

are all conserved. The conservation of R can also be understood from the symmetry of the original Lagrangian L under the infinitesimal variations

$$\Delta \psi_1 = \epsilon \psi_2, \quad \Delta \psi_2 = \epsilon \psi_1 \tag{2.43}$$

with ϵ real, which implies the conservation of $\psi_1 \psi_2$.

We turn now to the solution of the coupled equations (2.24). We start with the equation for x_0. This is the classical equation of the model without fermionic variables. It has the first integral

$$\dot{x}_0^2 - U^2 = 2E, \tag{2.44}$$

where $H_0 = E$ is the conserved energy, hence

$$\dot{x}_0 = (2E + U^2)^{1/2}. \tag{2.45}$$

The solution in integral form is

$$\int_{X_0}^{x_0} \frac{dx_0'}{(2E + U(x_0')^2)^{1/2}} = t, \tag{2.46}$$

where $x_0 = X_0$ at $t = 0$.

Given $x_0(t)$, and hence $U'(x_0(t))$, we can solve the linear equations for a_1, a_2, b_1, b_2. Of course, one solution is that these four functions all vanish. The supersymmetry variations (2.29) and (2.30) suggest the solution

$$a_1 = \lambda \dot{x}_0 + \mu U, \tag{2.47a}$$

$$a_2 = \lambda U + \mu \dot{x}_0, \tag{2.47b}$$

$$b_1 = \sigma \dot{x}_0 + \tau U, \tag{2.47c}$$

$$b_2 = \sigma U + \tau \dot{x}_0, \tag{2.47d}$$

where $\lambda, \mu, \sigma, \tau$ are arbitrary real constants, and U denotes $U(x_0(t))$. These functions do satisfy the equations of motion, e.g.,

$$\dot{a}_1 = \lambda \ddot{x}_0 + \mu U' \dot{x}_0 = \lambda U U' + \mu U' \dot{x}_0 = U' a_2, \tag{2.48}$$

and the presence of four constants implies that (2.47) is the general solution. The value of the conserved supersymmetry charge component Q_α is

$$Q_\alpha=\dot{x}_0 b_1-Ub_2=\sigma(\dot{x}_0^2-U^2)=2E\sigma, \tag{2.49}$$

and similarly $Q_\beta=-2E\lambda$, $\tilde{Q}_\alpha=2E\tau$, and $\tilde{Q}_\beta=-2E\mu$. The R charges take the values $R_a=2E(\sigma^2-\tau^2)$, $R_b=2E(\lambda^2-\mu^2)$, and $R=2E(\lambda\tau-\mu\sigma)$. There is a problem, however, if $E=0$, for then

$$\dot{x}_0=\pm U \tag{2.50}$$

and the expressions (2.47) depend on only two arbitrary constants. Equation (2.50) is the Bogomolny equation for this system. We postpone discussion of the general solution in this case to Sec. III.

The remaining equation for x_1 is the inhomogeneous linear equation

$$\ddot{x}_1=(UU')'x_1+2E(\lambda\tau-\mu\sigma)U'', \tag{2.51}$$

where we have substituted the conserved value of $R=a_1b_2-a_2b_1$. The supersymmetry transformations and the mini-time-translation suggest that solutions can be constructed from U and $\dot{x}_0$. It may be verified, using (2.24a) and (2.44), that a particular integral of (2.51) is

$$x_1=(\lambda\tau-\mu\sigma)U. \tag{2.52}$$

A solution of the homogeneous equation $\ddot{x}_1=(UU')'x_1$ is $x_1=\dot{x}_0=(2E+U^2)^{1/2}$, since $d^3x_0/dt^3=d(UU')/dt=(UU')'\dot{x}_0$. A second solution must satisfy $\dot{x}_1\dot{x}_0-x_1\ddot{x}_0=C$ for some constant (Wronskian) C. Write $x_1=f(t)\dot{x}_0$. Then f must satisfy $\dot{f}=C/\dot{x}_0^2$, so

$$\frac{df}{dx_0}=\frac{C}{\dot{x}_0^3}=\frac{C}{(2E+U^2)^{3/2}}, \tag{2.53}$$

and hence the second solution is

$$x_1=C(2E+U^2)^{1/2}\int_{X_0}^{x_0(t)}\frac{dx_0'}{(2E+U(x_0')^2)^{3/2}}. \tag{2.54}$$

The complete solution of (2.51) is therefore

$$x_1=(\lambda\tau-\mu\sigma)U+C_1(2E+U^2)^{1/2}+C_2(2E+U^2)^{1/2}\int_{X_0}^{x_0(t)}\frac{dx_0'}{(2E+U(x_0')^2)^{3/2}}. \tag{2.55}$$

The value of the energy constant H_1 is C_2.

We have therefore found the general solution of the equations of motion (2.24), in terms of eight constants of integration $X_0,E,\lambda,\mu,\sigma,\tau,C_1,C_2$. Although the potential $-U(x_0)^2$ is negative, the motion can be a bounded nonlinear oscillation if the minimum of the potential is finite and occurs at finite x_0, and if $2E$ is negative but greater than this minimum. The motion is unbounded if E is positive.

From the infinitesimal action of the supergroup on the dynamical variables, one can find its action on the constants of integration of the general solution. For example, a mini-time-translation changes C_1, and has no other effect. It is clear that the supergroup has six-dimensional orbits in the space of solutions. Only E and C_2 are invariant.

III. ZERO ENERGY SOLUTIONS

When the energy $E=0$, the method described above does not give the general solution of the equations of motion (2.24). For this value of E,

$$\dot{x}_0^2 - U^2 = 0, \tag{3.1}$$

so x_0 satisfies the first-order Bogomolny equation

$$\dot{x}_0 = \pm U. \tag{3.2}$$

For either choice of sign, $\dot{x}_0$ and U are no longer independent functions of time, so the expressions (2.47) depend effectively on only two arbitrary constants, and are no longer the general solution.

For simplicity, let us choose the upper sign in (3.2). The lower sign choice is essentially the same, and corresponds to a time reversal. Then the solution of (3.2) is

$$\int_{X_0}^{x_0} \frac{dx_0'}{U(x_0')} = t. \tag{3.3}$$

To find the general solution of the equations for a_1, a_2, b_1, b_2, it helps to consider the limit $E \to 0$ of the solution given earlier. Note that for small nonzero E,

$$\dot{x}_0 = (2E + U^2)^{1/2} = U + \frac{E}{U} + O(E^2). \tag{3.4}$$

A suitable linear combination of $\dot{x}_0$ and U is proportional to $1/U$ in the limit $E \to 0$. We therefore try

$$a_1 = \frac{\lambda}{U} + \mu U. \tag{3.5}$$

Then

$$\dot{a}_1 = -\frac{\lambda}{U^2} U' \dot{x}_0 + \mu U' \dot{x}_0 = U' \left(-\frac{\lambda}{U} + \mu U \right) \tag{3.6}$$

if $\dot{x}_0 = U$. Thus $a_2 = -\lambda/U + \mu U$ gives a solution of (2.24c), and it is easily checked that (2.24d) is also satisfied. Similarly we can solve Eqs. (2.24e) and (2.24f). So the general solution of Eqs. (2.24c)–(2.24f) is

$$\begin{aligned} a_1 &= \frac{\lambda}{U} + \mu U, \quad a_2 = -\frac{\lambda}{U} + \mu U, \\ b_1 &= \frac{\sigma}{U} + \tau U, \quad b_2 = -\frac{\sigma}{U} + \tau U, \end{aligned} \tag{3.7}$$

where $\lambda, \mu, \sigma, \tau$ are arbitrary constants.

The constants of the motion take the following values:

$$\begin{aligned} Q_\alpha = -\tilde{Q}_\alpha = 2\sigma, \quad -Q_\beta = \tilde{Q}_\beta = 2\lambda, \\ R_a = 2\sigma\tau, \quad R_b = 2\lambda\mu, \quad R = 2(\lambda\tau - \mu\sigma). \end{aligned} \tag{3.8}$$

These values are generally nonzero because of the careful way the limit $E \to 0$ was taken, even though previously these quantities were proportional to E.

The remaining equation for x_1 also needs special treatment. This equation is

$$\ddot{x}_1 = (UU')' x_1 + RU'', \tag{3.9}$$

where R is the constant given in (3.8). The previous solution had a particular integral proportional to U, and one homogeneous solution proportional to $\dot{x}_0$. When $E=0$, and $\dot{x}_0 = U$, one homogeneous solution is still U. But a new particular integral is required. Again the limiting procedure suggests that this should be proportional to $1/U$, and this is correct. Finding a second homogeneous solution is as before, but with $E=0$. The result is that the general solution of (3.9) is

$$x_1 = -\frac{R}{2U} + C_1 U + C_2 U \int_{X_0}^{x_0(t)} \frac{dx_0'}{U(x_0')^3}, \tag{3.10}$$

where C_1 and C_2 are arbitrary constants. $H_1 = C_2$, as before.

Note that in the zero energy, Bogomolny case, the orbits of the supergroup on the space of solutions are four dimensional, rather than six dimensional. Only the coefficients of U in (3.7) and (3.10) can be varied by the group action. This is consistent with the observation that the supersymmetry generator $\mathcal{Q} + \tilde{\mathcal{Q}}$ produces no variation at all when $\dot{x}_0 = U$ and a_1, a_2, b_1, b_2, x_1 all vanish.

A further observation is the following. Suppose the fermionic variables ψ_1 and ψ_2 take values obtained by acting with the supergroup, starting from zero. In other words $\lambda = \sigma = 0$, and the terms proportional to $1/U$ are absent. Then all the constants (3.8) are zero, and in particular there is no inhomogeneous term in the equation for x_1. So x_1 can be zero, although it need not be. In this kind of solution the fermions have no backreaction at all on the bosonic variables x_0 and x_1. It is perhaps rather generally true for Bogomolny-type solutions in supersymmetric theories that the fermions may have no backreaction on the bosons.

IV. *N*=1 SUPERSYMMETRIC MECHANICS

Another example of a solvable supersymmetric mechanical model is that of a particle moving in one dimension with $N=1$ supersymmetry (sometimes referred to as $N=\frac{1}{2}$ supersymmetry).[3,11] The supersymmetry algebra is simply

$$\mathcal{Q}^2 = \frac{d}{dt}. \tag{4.1}$$

The dynamical variables are a bosonic variable $x(t)$ and a single fermionic variable $\psi(t)$, taking values in B_e and B_o, respectively. The Lagrangian is

$$L = \tfrac{1}{2}\dot{x}^2 + \tfrac{1}{2}\dot{\psi}\psi + \alpha U(x)\psi. \tag{4.2}$$

α is an odd constant, an element of B_o. It is necessary for α to be odd, and L even, otherwise the equations of motion are inconsistent. This model is a variant of the usual nontrivial $N=1$ supersymmetric mechanical models. Normally, such a model has two or more fermionic variables.[12] Here, one of these is replaced by the odd constant α.

Taking the variation of L, ignoring total time derivatives, and shifting the variations to the left, gives

$$\Delta L = -\Delta x(\ddot{x} - \alpha U'\psi) - \Delta\psi(\dot{\psi} + \alpha U). \tag{4.3}$$

ΔL is required to vanish for any Δx in B_e and $\Delta\psi$ in B_o, so the equations of motion are

$$\ddot{x} = \alpha U'\psi, \tag{4.4a}$$

$$\dot{\psi} = -\alpha U. \tag{4.4b}$$

As before, there is no ambiguity, provided B has an even number or an infinite number of generators. Note that Eqs. (4.4) are consistent, because both sides of (4.4a) are in B_e, and both sides of (4.4b) in B_o.

The formal action of the supersymmetry operator $\mathcal{Q}$ on the dynamical variables is

$$\mathcal{Q}x = \psi, \quad \mathcal{Q}\psi = \dot{x}, \tag{4.5}$$

so (4.1) is satisfied. Genuine variations of the dynamical variables are

$$\delta x = \eta\psi, \quad \delta\psi = \eta\dot{x}, \tag{4.6}$$

where η is an arbitrary infinitesimal odd constant. The corresponding variation of L is

$$\delta L = \eta(\tfrac{1}{2}\dot{x}\dot{\psi} + \tfrac{1}{2}\ddot{x}\psi - \alpha U\dot{x}). \tag{4.7}$$

Let us introduce $V(x)$, satisfying $V' = U$. Then we can write δL as a total time derivative

$$\delta L = \eta \frac{d}{dt}\left(\frac{1}{2}\dot{x}\psi - \alpha V\right). \tag{4.8}$$

Hence L is supersymmetric, and the conserved supersymmetry charge is

$$Q = \dot{x}\psi + \alpha V. \tag{4.9}$$

Using standard arguments, we also obtain the energy

$$H = \tfrac{1}{2}\dot{x}^2 - \alpha U\psi. \tag{4.10}$$

Its conservation follows from the equations of motion, together with $\alpha^2 = 0$.

We may again obtain a concrete realization of this model by supposing that the Grassmann algebra B is B_2, with just two generators. Without loss of generality we may suppose that α is one of these generators, and that the other is β. The algebra is then identical to that in Sec. II. Note that if B had only one generator, then $\alpha\psi$ would be zero, and the model would become trivial.

We write the component expansion of the dynamical variables as

$$x(t) = x_0(t) + x_1(t)\alpha\beta, \tag{4.11a}$$

$$\psi(t) = a(t)\alpha + b(t)\beta, \tag{4.11b}$$

where x_0, x_1, a, b are ordinary functions. The Lagrangian has the expansion $L = L_0 + L_1\alpha\beta$, where

$$L_0 = \tfrac{1}{2}\dot{x}_0^2, \tag{4.12a}$$

$$L_1 = \dot{x}_0\dot{x}_1 + \tfrac{1}{2}\dot{a}b - \tfrac{1}{2}\dot{b}a + U(x_0)b. \tag{4.12b}$$

The equations of motion become

$$\ddot{x}_0 = 0, \tag{4.13a}$$

$$\ddot{x}_1 = U'(x_0)b, \tag{4.13b}$$

$$\dot{a} = -U(x_0), \tag{4.13c}$$

$$\dot{b}=0. \tag{4.13d}$$

These can be obtained as the components of Eqs. (4.4). They are also the variational equations obtained from L_1 and L_0, and, as before, L_0 is redundant.

Equations (4.13) imply the conservation of

$$Q_\alpha=\dot{x}_0 a+V(x_0), \tag{4.14a}$$

$$Q_\beta=\dot{x}_0 b, \tag{4.14b}$$

$$H_0=\tfrac{1}{2}\dot{x}_0^2, \tag{4.14c}$$

$$H_1=\dot{x}_0\dot{x}_1-U(x_0)b, \tag{4.14d}$$

and these are the components of Q and H.

It is straightforward to solve Eqs. (4.13), starting with

$$x_0=\lambda t+\mu, \quad b=\nu, \tag{4.15}$$

where λ,μ,ν are arbitrary constants. The energy H_0 is $\frac{1}{2}\lambda^2$. We now regard Q_α as a constant of integration, obtaining

$$a=\frac{Q_\alpha}{\lambda}-\frac{1}{\lambda}V(\lambda t+\mu) \tag{4.16}$$

as the solution of (4.13c). Finally, treating H_1 similarly, we have

$$\dot{x}_1=\frac{H_1}{\lambda}+\frac{\nu}{\lambda}U(\lambda t+\mu) \tag{4.17}$$

so

$$x_1=\frac{H_1}{\lambda}t+X_1+\frac{\nu}{\lambda^2}V(\lambda t+\mu), \tag{4.18}$$

where X_1 is a constant. The general solution of the model involves six arbitrary constants λ, μ, ν, Q_α, H_1, X_1.

Let us now clarify how the supergroup acts in this model. The Lie algebra of the supergroup is obtained from the supersymmetry and time translation operators by taking coefficients in B_o and B_e, respectively. It therefore has basis

$$\mathcal{Q}_\alpha=\alpha\mathcal{Q}, \quad \mathcal{Q}_\beta=\beta\mathcal{Q}, \quad \frac{d}{dt}, \quad \frac{\tilde{d}}{dt}=\alpha\beta\frac{d}{dt}. \tag{4.19}$$

The only nontrivial bracket is

$$[\mathcal{Q}_\alpha,\mathcal{Q}_\beta]=-2\frac{\tilde{d}}{dt}. \tag{4.20}$$

Acting with $\mathcal{Q}_\alpha$ and $\mathcal{Q}_\beta$, we obtain from (4.5) two independent supersymmetry variations

$$\delta_\alpha x=\epsilon\alpha\psi, \quad \delta_\alpha\psi=\epsilon\alpha\dot{x}, \tag{4.21a}$$

$$\delta_\beta x=\epsilon\beta\psi, \quad \delta_\beta\psi=\epsilon\beta\dot{x}, \tag{4.21b}$$

where ϵ is now infinitesimal and real. Writing x and ψ in terms of components, we find

$$\delta_\alpha x_1 = \epsilon b, \quad \delta_\alpha a = \epsilon \dot{x}_0, \tag{4.22a}$$

$$\delta_\beta x_1 = -\epsilon a, \quad \delta_\beta b = \epsilon \dot{x}_0, \tag{4.22b}$$

with all other variations vanishing. These are symmetries of L_1, and trivially of L_0. In addition there is symmetry under an infinitesimal time translation of all the dynamical quantities. Finally, there is symmetry under the mini-time-translation, with generator $\widetilde{d/dt}$. As before, an infinitesimal mini-time-translation is

$$\Delta x_1 = \epsilon \dot{x}_0 \tag{4.23}$$

with ϵ real.

By considering how the supergroup generators act on the solutions of this model, we see that with the supergroup we may independently vary μ, ν, Q_α, X_1, but not the constants defining the energy λ and H_1. The orbits of the supergroup are therefore four dimensional, in general.

The solution as we have presented it does not make sense if $\lambda = 0$. This is the zero energy, Bogomolny case. If $H_0 = 0$ then $\dot{x}_0 = 0$, so x_0 takes a constant value μ, hence U and U' take constant values $U(\mu)$ and $U'(\mu)$. The general solution is then easily found to be

$$x_0 = \mu, \quad b = \nu, \tag{4.24a}$$

$$a = -U(\mu)(t - t_0), \tag{4.24b}$$

$$x_1 = \tfrac{1}{2} U'(\mu) \nu t^2 + rt + X_1, \tag{4.24c}$$

where μ, ν, t_0, r, X_1 are constants of integration. The second energy constant is $H_1 = -U(\mu)\nu$. Supersymmetry transformations and time translations change the constants r, X_1, and t_0. However, unlike in the $H_0 \neq 0$ case, Eq. (4.22b) implies that the value of b cannot be changed, and the orbits of the supergroup are three dimensional rather than four dimensional.

V. CONCLUSIONS

We have presented two supersymmetric classical mechanical models. By supposing that the dynamical variables take values in the Grassmann algebra B_2 with two generators, we have deconstructed the models into component form and obtained equations of motion which can be explicitly solved. These equations are the variational equations of a Lagrangian L_1 of nonstandard form, and in each case, the ''body'' variable x_0 obeys a classical equation unaffected by the fermionic variables. The supergroup, which is a genuine Lie group, generated from the super Lie algebra of supersymmetries and time translations by tensoring with B_2, acts on the space of solutions.

One could ask how the solutions would look if the dynamical variables were reconstructed from their components, so as to be B_2 valued, or further combined into supermanifold dynamical variables. At first sight there is only a slight gain in elegance, but this needs more careful study. It is also of interest to know whether the equations remain solvable if B is a larger algebra.

The model discussed in Sec. IV involved an odd constant α. Possibly, Grassmann-valued constants are of use in other supersymmetric models. For example, it might be possible in certain ''brane'' models to have a nonreal cosmological constant.

One of the motivations for this work was to better understand the solitons that occur in many supersymmetric field theories. These are solutions of the classical field equations, with the fermionic fields set equal to zero. They usually also satisfy first-order Bogomolny equations. It would be much more satisfactory if they could be regarded as special cases of solutions where the fermionic fields are nonzero. Our mechanical models suggest that the ''body'' fields of the soliton will be unaffected by the fermionic fields. But the general solutions will involve nonzero fermionic fields coupled to the soliton, and in addition there can be nonzero bosonic fields with values in the even, nonreal part of the Grassmann algebra.

The connection between the classical models discussed here and their quantized versions is also worth exploring. The Heisenberg equations of the quantized theory may be formally the same as the equations that we have solved, but $x, \dot{x}$ and $\psi, \dot{\psi}$ need to obey canonical commutation and anticommutation relations, respectively. It would be interesting to know whether the general classical solution describes a suitable limit of a quantum state.

ACKNOWLEDGMENTS

I am grateful to Alan Macfarlane, Jonathan Evans, and Bernd Schroers for discussions during the course of this work. I thank Roberto Casalbuoni for comments on an earlier version of this paper.

[1] B. De Witt, *Supermanifolds*, 2nd ed. (Cambridge University Press, Cambridge, 1992).
[2] P. G. O. Freund, *Introduction to Supersymmetry* (Cambridge University Press, Cambridge, 1986).
[3] R. Casalbuoni, Nuovo Cimento A **33**, 389 (1976).
[4] F. A. Berezin and M. S. Marinov, Ann. Phys. (N.Y.) **104**, 336 (1977).
[5] A. Rogers, J. Math. Phys. **21**, 1352 (1980).
[6] J. M. Rabin and L. Crane, Commun. Math. Phys. **102**, 123 (1985).
[7] E. B. Bogomolny, Sov. J. Nucl. Phys. **24**, 449 (1976).
[8] E. Witten, Nucl. Phys. B **188**, 513 (1981).
[9] A. Rogers, Commun. Math. Phys. **105**, 375 (1986).
[10] A. Rogers, J. Math. Phys. **22**, 939 (1981).
[11] M. Sakimoto, Phys. Lett. **151B**, 115 (1985).
[12] A. J. Macfarlane, Nucl. Phys. B **438**, 455 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Stochastic quantum geometry within a (1+1) manifold: A basic construction

Steven D. Miller[a]
University of Strathclyde, Glasgow, Scotland, United Kingdom

(Received 13 April 1998; accepted for publication 3 September 1998)

A bounded, globally hyperbolic, space–time manifold $\mathscr{M}^2$ with topology $I(t) \otimes \mathbb{R}$ is presented as a simple classical, deterministic, geometry, with ADM (1+1) slicing via a lapse function and shift vectors. The Cauchy development is then equivalent to the evolution of a spatial Riemann 1-metric $\beta(x,0) \rightarrow \beta(x,t)$ for a bounded spatial interval $I(t)=[0,l]$ such that $x-x'<=l$ for all $x,x' \in (t)$ for all $t \in \mathbb{R}$. The 1-metric is considered a parametrized, stochastically fluctuating variable $\{\beta(x,t)\} \in Riemm(I)$ for some critical (microscopic) correlation scale l, where $Riemm(I)$ is the space of all 1-metrics on I. If the metric fluctuations are constrained by a scale-dependent probability kernel or density distribution on [0,1], then a Fokker–Planck equation can be developed for the Cauchy evolution of the kernel. The stationary (Cauchy invariant) equilibrium limit solution is obtained. The equilibrium limit correlations $\langle \beta(x,t)\beta(x',t)\rangle$ at second order derived from the stochastic model, can be directly identified with the general, well-known form of the metric two-point (equal time) correlations obtained from linearized general relativity treated as a quantum field theory. The metric diffusion coefficients of the stochastic model are then correctly identified. The uncertainty relation $l\delta\beta>=l_*$ for nonzero metric fluctuations $\delta\beta$, emerges from the solution and is a necessary condition for the kernel to be constrained on [0,1]. The 1-metric fluctuations are exponentially damped or amplified as the spatial interval $I=[0,l]$ is expanded or contracted with respect to the Planck length l_*.
[S0022-2488(99)01601-1]

I. INTRODUCTION

The essential idea behind stochastic quantization schemes is to reformulate quantum mechanics or quantum field theory as the thermal-equilibrium limits of a hypothetical stochastic process.[1–3] One might consider this as a third approach to quantization, in addition to the standard canonical and path integral methods. The purpose of this stochastic interpretation is to find the unknown physical origins of quantum fluctuations by replacing or reformulating quantum theories with theories based on classical stochastic dynamics. The *n*-point equal time correlation or Green's function of quantum field theory are identified with the thermal equilibrium correlations of the stochastic process. The stochastic interpretation has its roots in the formal similarity between the linear Schrödinger equation and the linear diffusion equation for Brownian particles. The Schrödinger equation is essentially a diffusion equation in imaginary time with $\mathscr{D}=\hbar/2m$ playing the role of a diffusion coefficient that vanishes for classical (large) m. Both Bohm and Nelson[4,5] formulated stochastic interpretations of quantum mechanics in real time. Bohm attempted to replace quantum mechanics with a classical stochastic dynamics via hidden variables. The theories of Bohm and Nelson remain interesting from the point of view of assigning a stochastic interpretation to quantum fluctuations. A major development in stochastic quantization is due to Parisi and Wu,[1–3] which gives quantum mechanics as a thermal equilibrium limit of a hypothetical stochastic

[a]Correspondence address: 142 Cameron Drive, Kilmarnock KA3 7PL, Scotland, United Kingdom. Electronic mail: 101551.1243@Compuserve.com

process. Unlike Nelson's approach in real time, the Parisi–Wu scheme is formulated with respect to a fictitious time variable rather than ordinary time. Both theories are based on a Markovian process of the Weiner type with Gaussian white noise. Here, a d-dimensional quantum system is equivalent to a $(d+1)$-dimensional classical system with random, stochastic fluctuations. The new time variable is treated as a computational convenience with no direct physical content. The usual formulations of quantum mechanics and quantum field theory have so far utilized imaginary time or Euclidean manifolds via Wick rotations in order to have better defined path integrals. The theories of Bohm and Nelson are formulated in ordinary real time and recently some authors have also formulated viable stochastic quantum mechanical approaches in terms of ordinary Minkowski space coordinates, keeping real time within the formalism.[6–8] These theories demonstrate that the use of imaginary time or Euclidean manifolds is not inevitable. Despite success in dealing with quantum fields and gauge theories and the emergence of results not obtainable via conventional methods, stochastic interpretations have not fully been applied to the problem of the quantization of gravity and geometry. It is possible that the stochastic method may also have the potential to extend the scope of quantum gravity and geometry.

The current approaches to quantum gravity and quantum geometry are (i) canonical quantum gravity and the developments involving loop space and Ashtekar variables;[9–12] (ii) the Euclidean path integral approaches;[13] and (iii) superstring theory and M(atrix) theory.[14–16] Despite progress, unresolved mathematical, technical, and conceptual difficulties remain. It is extremely difficult to construct any self-consistent interpretation of quantum gravity. A review of the current status and difficulties can be found in Ref. 17. Despite these ongoing problems, key physical concepts can be identified, which lie at the heart of quantized gravity and geometry. These are metric fluctuations and the existence of a lower length scale truncation of $l_* = (G\hbar/c^3)^{1/2} = 10^{-33}$ cm, the Planck length. Beyond this scale, the deterministic, geometrical structure of space–time is lost. This is a model-independent feature of quantum gravity and many different arguments give rise to this same conclusion.[18] In classical general relativity, the d-dimensional space–time manifold $\mathscr{M}^d$ is pseudo-Riemannian, generally globally hyperbolic,[10] and is assumed C^2 differentiable with a casual, Lorentzian metric structure $g_{uv}(x) = g_{uv}(x,t)$. The manifold is smooth and continuous down to zero length scales, with the existence of infinitesimal differential limits on space and time such that $\delta x^u \to 0$. However, as pointed out in Ref. 19, all known physical systems possess inherent noise at some critical length scale with the manifestation of fluctuation-correlation behavior. If a vacuum space–time manifold $\mathscr{M}^4$ with metric $g_{uv}(x)$, describing a static gravitational field is itself considered as a fundamental physical system, the solutions $g_{uv}(x)$ of the vacuum Einstein equations $(R_{uv} = 0)$ should exhibit ''geometric noise'' or a strong fluctuation-correlation behavior at a critical length scale or correlation length, of the order of the Planck scale.

Wheeler[20,21] originally introduced the concept of geometrodynamics or space–time quantum ''foam'' and demonstrated that geometry itself must fluctuate near the Planck length, as exhibited by quantum mechanical fluctuations in the metric tensor itself of order $\langle g \rangle \sim (l_*/l)$, where l is the spatial resolution. Clearly for $l > l_*$, the fluctuations $\langle g \rangle$ rapidly vanish and can be ignored since $\langle g \rangle \sim 10^{-20}$, even on the nuclear scale. This suggests that as $l \to l_*$, gravity exhibits a strongly coupled fluctuating phase (the foam) on which there are no correlations on a large scale. The length l_* seems to play a role in space–time geometry similar to the role played by c in relativistic mechanics.[18] One can get closer and closer to c but never reach it; there are no velocities beyond c. Similarly, for any geometric variation $\delta x = |\mathbf{x} - \mathbf{x}'|$, one can get arbitrarily close to l_*, but the uncertainty relation $\delta x >= l_*$ strictly holds; there are no distances below l_*.

I propose that stochastic analysis may be a viable mathematical framework within which to interpret and formulate descriptions of basic features of metric fluctuations, discretization, and foam structure in geometry and gravity. Rumpf[22] presented a basic stochastic treatment for the gravitational field using real time. Most recently, Moffat[19] has considered treating the gravitational field stochastically. For strong metric fluctuations, it is shown, via stochastic Raychaudhuri and Langevin equations, how caustic singularities in space–time structure can potentially be avoided for congruences of converging geodesics and collapsing stars. This illustrates the potential of the interpretation. In this paper, the emphasis is on the conceptual issues, and future potential rather than intricate technical subtleties, focusing on well-established key physical consequences of quantum geometry (metric fluctuations and Planck length truncation). It is felt, however, that the points addressed here are a necessary prerequisite toward mature stochastic models of quantum geometry.

II. VACUUM QUANTUM FLUCTUATIONS IN SCALAR FIELD THEORY AND LINEARIZED EINSTEIN GRAVITY

We briefly consider scalar quantum field theory on a flat manifold $(\mathscr{M}^4, \eta_{ab})$ and estimate the typical magnitude of a quantum fluctuation when a massless spin-0 scalar field $\varphi(x)$ is probed at a physical length scale l, on a local inertial frame. The result can then be extended to linearized Einstein gravity for massless spin-2 fields. These basic and heuristic aspects of quantum fields will be adequate in the form presented and useful later when we come to consider the equal time correlations. Scalar field theory is paradigmatic and well known, so for our purposes we will label only the key expressions. The Hamiltonian is $H=\frac{1}{2}\int d^3x[(d\varphi(\mathbf{x},t)/dt)^2+(\boldsymbol{\nabla}\varphi)^2]$. The free field equation is $\Box\varphi(x)=0$. Quantizing, the functional Schrödinger equation is[23] $i\hbar\,\partial_t\Psi(\varphi(\mathbf{x}),t)) =H\Psi(\varphi(\mathbf{x}),t))$. The quantum field theoretic state $\Psi[\varphi(\mathbf{x}),t]=\Psi[\varphi(\mathbf{x},t)]=\Psi[\varphi(x)]$ is a function of real time $t\in\mathbb{R}$, but a functional of the field $\varphi(\mathbf{x})$. By comparison and analogy with the ordinary quantum mechanical simple harmonic oscillator, the ground state or vacuum functional for the quantum field theory is immediately of the form

$$\Psi_0[\varphi(\mathbf{x},t]\sim\exp\left[-\frac{1}{2\hbar}\int d^3x\;\varphi(\mathbf{x})(-\nabla^2)^{1/2}\varphi(\mathbf{x})\right]\exp[-iE_0/\hbar]. \tag{2.1}$$

Here $(-\nabla^2)^{1/2}$ is the formal square root of $(-\nabla^2)$, and is interpreted as a pseudodifferential operator. If we take a Fourier transform $\mathcal{F}[\nabla^2]=-k^2$ then $\mathcal{F}[(-\nabla^2)^{1/2}]=\|\mathbf{k}\|$. The ground state energy $E=\frac{1}{2}\hbar\,\mathrm{tr}(-\nabla^2)^{1/2}$ is formally infinite. However, this is of no concern: the key feature is that the ground state wave functional is peaked at $\boldsymbol{\nabla}\varphi=0$.

Suppose now the scalar field $\varphi(\mathbf{x})$ is probed at a scale l. We define the correlation function or two-point function as $\langle\varphi(\mathbf{y})\varphi(\mathbf{x})\rangle=\langle\Psi_0|\varphi(\mathbf{y})\varphi(\mathbf{x})|\Psi_0\rangle$. We also define an object $[\Delta\varphi]_l$ as

$$[\Delta\varphi]_l=[\langle\varphi(\mathbf{x}')\varphi(\mathbf{x})\rangle]^{1/2}{}_{l=|\mathrm{x}-\mathrm{x}'|}. \tag{2.2}$$

This is the quantum fluctuation probed at distance $l=|\mathbf{x}-\mathbf{x}'|$. The correlator is $\langle\varphi(x')\varphi(x)\rangle -\langle\varphi(x')\rangle\langle\varphi(x)\rangle$, but at the lower-energy length scales, one can take the correlations to be weak so that $\langle\varphi(x')\varphi(x)\rangle-\langle\varphi(x)\rangle\langle\varphi(x')\rangle\sim\epsilon$ for infinitesimal (but nonzero) ϵ. Then (2.2) is justified. The best way to determine $[\Delta\varphi]_l$ is to move to momentum space $\varphi(\mathbf{k})$ so that

$$\varphi(\mathbf{x})=\int\frac{d^3k}{2\pi^3}\,\varphi(\mathbf{k})\exp(i\mathbf{k}\cdot\mathbf{x}). \tag{2.3}$$

Then the vacuum wave functional is $\Psi_0[\varphi(\mathbf{k})]$ is

$$\Psi_0[\varphi(\mathbf{k})]\sim\exp\left[-\frac{1}{2\hbar}\int\frac{d^3x}{2\pi^2}\,k\varphi(\mathbf{k})^2\right], \tag{2.4}$$

and the momentum two-point function is $\langle\varphi(\mathbf{k})\varphi(\mathbf{k})\rangle=[\hbar(2\pi)^3/k_1]\delta^3(\mathbf{k}_1-\mathbf{k}_2)$. The position two-point correlator then follows from the inverse Fourier transform:

$$\langle\varphi(\mathbf{x}')\varphi(\mathbf{x})\rangle=\int\frac{d^3k}{2\pi^3}\frac{\hbar}{k}\exp[i\mathbf{k}(\mathbf{x}-\mathbf{x}')]=\frac{2\hbar}{(2\pi)^3|\mathbf{x}-\mathbf{x}'|^2}. \tag{2.5}$$

Hence $[\Delta\varphi]_l=(2\hbar)^{1/2}/2\pi l\sim\sqrt{\hbar}/l$, where $l=|\mathbf{x}-\mathbf{x}'|$. Quantum fluctuations therefore only become significant on small physical scales l. The characteristic size of the quantum fluctuations $\Delta\varphi$ about the average field φ depend on the physical length scale l being probed. Vacuum fluctuations in the gravitational field can be estimated in a similar manner and we can make use of Eq. (2.5). The linearized Einstein–Hilbert action S_{LEH} for gravity coupled to matter follows from expanding the Ricci scalar to second order[24] with $g_{ab}(x)=\eta_{ab}(x)+\gamma_{ab}(x)$:

$$S_{\mathrm{LEH}}=-\frac{l}{32\pi l_*^2}\int d^4x\ \gamma^{ab}(x)\Box\gamma_{ab}(x)+\int d^4x\ \gamma^{ab}(x)\left[\Theta_{ab}-\frac{1}{2}\Theta\,\eta_{ab}\right]+O(\gamma^2). \quad (2.6)$$

The field equations are (gauge fixed) linear wave equations for a massless spin-2 field γ:

$$\Box\gamma_{ab}(x)=-\frac{16\pi l_*^2}{\hbar}\left[\Theta_{ab}-\frac{1}{2}\Theta\,\eta_{ab}\right]+O(\gamma^2) \quad (2.7)$$

or $\Box\gamma_{ab}(x)=0$ in vacuum. These equations have a simple retarded potential and plane wave solutions of the tensor form $\gamma_{ab}(x)=H_{ab}\exp(i\Sigma_\mu k_\mu x^\mu)$ for $\mu=1$–3, describing massless spin-2 gravitons with two polarization states. Since the linearized Einstein action is quadratic in γ, quantization proceeds in analogy with the scalar field. This leads to two-point correlation functions $\langle\gamma(x)\gamma(x')\rangle$, $\langle\gamma_{ab}(x)\gamma(x')\rangle$, and $\langle\gamma_{ab}(x)\gamma_{cd}(x')\rangle$. If we set $\Theta_{uv}=0$, we can calculate the trace–trace correlator $\langle\gamma(x)\gamma(x')\rangle$ using (2.6) from the scalar field case so that $\langle\gamma(x)\gamma(x')\rangle=(16\pi l_*^2/\hbar)\langle\varphi(x)\varphi(x')\rangle$. The scalar correlation contains the essential quantitative and qualitative features. Then, at equal times,

$$\langle\gamma(\mathbf{x}',t)\gamma(\mathbf{x},t)\rangle=\frac{16\pi l_*^2}{\hbar}\frac{2\hbar}{(2\pi)^2|\mathbf{x}-\mathbf{x}'|^2}=\left(\frac{8}{\pi}\right)\frac{l_*^2}{|\mathbf{x}-\mathbf{x}'|^2}. \quad (2.8)$$

When we probe geometry at a distance l, the vacuum quantum metric fluctuations $[\Delta\gamma]l$ are

$$[\Delta\gamma]_l=[\langle\gamma(\mathrm{x},t)\gamma(\mathrm{x}',t)\rangle]^{1/2}{}_{l=|\mathrm{x}-\mathrm{x}'|}\sim\langle\gamma(\mathrm{x},t)\rangle\sim(l_*/l), \quad (2.9)$$

or $[\Delta\gamma]l=(8/\pi)^{1/2}(l_*/l)$. These become significant only as we approach the Planck length $l\to l_*$ and rapidly damp out as l increases, even at the nuclear scale, where $l_*/l\sim10^{-20}$.

III. THE ADM (3+1) AND (1+1) SLICING OF SPACE–TIME

Let $(\mathscr{M}^d,g_{ab})$ be a generic, globally hyperbolic d-dimensional space–time; then $(\mathscr{M}^d,g_{uv})$ is stably casual and permits a foliation of $(d-1)$-dimensional space-like hypersurfaces $\{\Sigma\}$. A global time function $t\equiv\mathbb{R}$ can be chosen such that each $(d-1)$-geometry of constant $t\in\mathbb{R}$ is a space-like Cauchy surface Σ with parametrization $\Sigma(t)$. The (fixed) topology of the space–time is then $\mathscr{M}^d\sim\mathbb{R}\times\Sigma$. A point $p\in\mathscr{M}^4$ can be locally represented as $x=(\mathbf{x},t)$ where $\mathbf{x}\in\Sigma(t)$ and $\mathbf{x}=(x_1,x_2,\ldots,x_{d-1})$ and $t\in\mathbb{R}$. A crucial feature of global hyperbolicity is that it permits a well-defined initial value (Cauchy) problem for the Green's function or two-point propagators $\{G(x,x')\}$ or $\langle\varphi(x)\varphi(x')\rangle$ in classical and quantum field theory, such as in (2.6). Solutions of the massless wave equations for scalars, photons, and gravitons ($\Box\varphi=0$, $\Box A^a=0$, or $\Box\gamma_{ab}=0$) must vanish outside the forward light cone.[25]

Given $d=4$, for example, then explicitly, the ADM (3+1) decomposition of the 4-geometry $(\mathscr{M}^4,g_{uv}(\mathbf{x},t))$ consists of the following geometric data[26]

(i) The induced fundamental form or Riemmann 3-metric $h_{ij}(\mathbf{x},t)$ on each $\Sigma(t)$, for $i,j=1,2,3$ and $\mathbf{x}\in\Sigma(t)$ and $t\in\mathbb{R}$. This is the intrinsic geometry of the 3-space and h_{ij} is positive definite.

(ii) The manner in which each Σ is embedded in $(\mathscr{M}^4,g_{uv})$. This is ascertained once we are able to compute the spatial part of the covariant derivative of the normal n to $\Sigma(t)$. If $\boldsymbol{\nabla}$ is the four-dimensional connection of $(\mathscr{M}^4,g_{uv})$, one can define the extrinsic curvature tensor $K_{ij}\equiv-\boldsymbol{\nabla}_j n_i$, so that symmetric K_{ij} require symmetric $\boldsymbol{\nabla}$. Alternatively, $K_{ij}\equiv-\frac{1}{2}(\mathcal{L}_n h)_{ij}$, where $\mathcal{L}_n$ is the Lie derivative along the normal to $\Sigma(t)$.

(iii) The manner in which the coordinates are propagated. One defines the vector $(N,N^i) = (N,N^1,N^2,N^3)dt$ connecting $(x^i,t)=(\mathbf{x},t)$ with the point $(x^i,t+\delta t)=(\mathbf{x},t+\delta t)$.

Given the surface $x^0=t$ and surface $x^0=t+\delta t$, the quantity $N\ dt = d\tau$ defines a displacement normal to surface $x^0=t$. Here, N is the lapse function and N^i are the shift vectors. The $N^i\ dt$ yield the displacement from the point (t,x^i) to the base of the normal to $x^0=t$ through $(t+\delta t,x^i)$. The full metric can be expressed as

$$\begin{aligned} ds^2 &= g_{uv}\ dx^u\ dx^v \\ &= h_{ij}(dx^i+N^i\ dt)(dx^i+N^j\ dt)-N^2\ dt\ dt \\ &= -(N^2-N_iN^i)dt^2+N_i(dx^i\ dt+dt\ dx^i)+h_{ij}\ dx^i\ dx^j. \end{aligned} \tag{3.1}$$

Then $g_{00}=-(N^2-N_iN^i)$, $g_{i0}=g_{0i}=N^i$, $g_{ij}=h_{ij}$, and $K_{ij}=(-\partial_t h_{ij}+N_{i;j}+N_{j;i})$. The ''gauge choice'' $N^i=0$ eliminates the off-diagonal components, so that

$$ds^2=-N^2\ dt^2+h_{ij}\ dx^i\ dx^j. \tag{3.2}$$

The Cauchy evolution from $\Sigma(0)$ to $\Sigma(t)$ can be interpreted as the change of Riemmann 3-metric from $h_{ij}(\mathbf{x},0)$ to $h_{ij}(\mathbf{x},t)$ so that h_{ij} is the dynamical variable. This is the motivation for Hamiltonian general relativity and (pre-Ashtekhar) canonical quantum gravity. We are primarily interested in the induced Riemmann metric $h_{ij}(\mathbf{x},t)$ on $\Sigma(t)$ and set $N=$const. For the 3-geometry,

$$ds^2=h_{ij}(\mathrm{x},t)dx^i\ dx^j=h_{11}(x,y,z,t)dx^2+h_{22}(x,y,z,t)dy^2+h_{33}(x,y,z,t)dz^2. \tag{3.3}$$

Given Σ, one denotes[26] by *Riemm*(Σ) the space of Riemannian C^∞ metrics h_{ij} on Σ, and *Diff*(Σ), the group of diffeomorphisms on Σ. If $\phi\in Diff(\Sigma)$ then $\phi:\Sigma\to\Sigma$, and the group *Diff*(Σ) preserves the metric structure. If $h_{ij}\in Riemm(\Sigma)$ and $\phi\in Diff(\Sigma)$ then $\phi^*h(\phi(\mathbf{x}),t))=h(\mathbf{x},t)$ for $\mathbf{x}\in\Sigma$ and $t\in\mathbb{R}$, where ϕ^* is the pullback of ϕ. The group *Diff*(Σ) acts as a transformation group on *Riemm*(Σ). Its action maps (ϕ,h) to ϕ^*h for all $\beta\in Riemm(\Sigma)$ and $\phi\in Diff(\Sigma)$. Displacements ds^2 are preserved. The space of all orbits of *Diff*(Σ) is $\mathcal{S}(\Sigma)$:

$$\mathcal{S}(\Sigma)=Riemm(\Sigma)/Diff(\Sigma). \tag{3.4}$$

The space $\mathcal{S}(\Sigma)$ is (Wheeler's) superspace.[26] Hence, for all $h\in Riemm(\Sigma)$, one considers all metrics derived from h, by the group elements $\phi\in Diff(\Sigma)$. If two 3-metrics h and h' are on the same orbit, then there is ϕ of Σ such that $\phi^*h=h'$, so that h and h' are isometric. Superspace $\mathcal{S}(\Sigma)$ is then the set of geometries of Σ that are equivalence classes of isometric Riemannian metrics.

To make the analysis as clear and simple as possible, we can let $d=2$, giving a (1+1)-dimensional manifold with topology $\mathscr{M}^2=I(t)\times\mathbb{R}$, so that $I(t)$ for any $t\in\mathbb{R}$ is an interval and $x\in I(t)$. This still embodies the essential features we wish to elucidate. The (1+1) decomposition procedure of the metric $g_{ab}=(g_{00},\beta)$ is as before:

$$ds^2=\beta(x,t)dx^2-N^2\ dt^2, \tag{3.5}$$

so that along the one-dimensional spatial manifold $I(t)$, we have ${}^{(3)}ds^2=\beta(x,t)dx^2$. The $I(t)$ can be a closed interval or a circle S^1. For a bounded interval, $I(t)=[0,l]\supset\mathbb{R}$ so that for any $x,x'\in I(t)$; then $|x-x'|_{\max}=l$ or $|x-x'|<l$. The spatial metric β is then simply a scalar function $\beta(x,t)$ for any $x\in I(t)$ and $t\in\mathbb{R}$ and $\beta(x,t')=\beta(x,t)$ for all t', $t\in\mathbb{R}$. The group *Diff*(I) or *Diff*($I(t)$) is the group of diffeomorphisms on the interval $I(t)$ and *Riemm*(I) is the space of all metrics $\beta(x,t)$ on $I(t)$ for all $x\in I(t)$ and $t\in\mathbb{R}$, with $\phi\in Diff(I)$. As before, $\phi^*\beta(\phi(x),t)=\beta(x,t)$ and a ''superspace'' of 1-metrics $\mathscr{S}(I)=Riemm(I)/Diff(I)$.

All known physical systems possess noise at some critical length scale l with emergence of correlation–fluctuation behavior. For example, in Ref. 27, it is suggested that Einstein's equations are macroscopic ''state equations,'' arising from such a deeper microsubstructure of the classical geometry itself. In the same way, a homogeneous gas with macroscopic temperature and pressure is comprised of discrete atoms. Linearized Einstein gravity treated as a quantum field theory [see Eq. (2.10)], suggests that for $\mathbf{x},\mathbf{x}' \in I(t)$ with $|\mathbf{x}-\mathbf{x}'| <= l$ and $t \in \mathbb{R}$, $\langle \beta(\mathbf{x},t)\beta'(\mathbf{x}',t)\rangle \sim (l_*/l)^2$ at equal time. Correlations emerge as $l \to l_*$, so, depending on l, the 1-metric $\beta(x,t)$ can evolve from a fixed, deterministic variable β into a randomly fluctuating (stochastic) variable such that $\beta \to \langle \beta \rangle$. Essentially, a simple quantum $(1+1)$ space–time will be interpreted as a classical space–time subject to random stochastic fluctuations (in real time $t \in \mathbb{R}$). This first requires a reasonable metric fluctuation probability interpretation and measure on the bounded real interval $I(t)=[0,l]$.

IV. PARAMETRIZED STOCHASTIC FLUCTUATIONS OF THE RIEMANN 1-METRIC

The function $\beta(x,t)$ is redefined as a random, stochastic variable that appears or evolves to being deterministic under specific conditions. Following Refs. 27–29, we briefly address the general theory of random variables and then apply the formalism to the Riemmann 1-metric function $\beta(x,t)$ for the $(1+1)$ manifold $\mathscr{M}^2$ with $x \in I(t)$ and $t \in \mathbb{R}$. If Ω is a generic set then a σ-algebra $\mathcal{F}$ on Ω is a family $\mathcal{F}$ of subsets of Ω with the following properties: (i) $\emptyset \in \mathcal{F}$; (ii) if $A \in \mathcal{F}$ then $A^{\mathrm{com}} \in \mathcal{F}$, where $A^{\mathrm{com}} = \Omega/A$ is the complement of A in Ω; (iii) if $A_1, A_2, \ldots, \in \mathcal{F}$ then

$$A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}. \tag{4.1}$$

The pair $(\Omega,\mathcal{F})$ is the measurable space. A probability measure $\mathcal{P}$ on $(\Omega,\mathcal{F})$ is a function $\mathcal{P}:\mathcal{F} \to [0,1]$ such that (a) $\mathcal{P}(\emptyset)=0$ and $\mathcal{P}(\Omega)=1$ (b) for $A_1, A_2, \ldots, \in \mathcal{F}$ and for $\{A_i\}_{i\to 1}$ disjoint (i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$); then

$$\mathcal{P}(\bigcup_i A_i) = \sum_i \mathcal{P}(A_i). \tag{4.2}$$

The triplet $(\Omega,\mathcal{F},\mathcal{P})$ is the probability space. The subsets S of Ω that belong to $\mathcal{F}$ are $\mathcal{F}$-measurable sets. In the probability interpretation, these sets are called events, and we use the interpretation $\mathcal{P}(S)=$''probability that event S occurs.'' If $\mathcal{P}(S)=1$, then the probability is absolute or deterministic. A random variable Q is $Q:\Omega \to \mathbb{R}^n$. A generic, stochastic process in time $\{Q_t\}_{t} \in \mathbb{R}$ is a parametrized collection of random variables, defined on $(\Omega,\mathcal{F},\mathcal{P})$ and assuming values in $\mathbb{R}^n$. For example, the process may depend on two variables (t,ω) on $(\Omega \otimes \mathbb{R})$ so that $(t,\omega) \to Q(t,\omega)$ from $(\mathbb{R} \times \Omega)$ onto $\mathbb{R}^n$. As t evolves, then $Q(t,\omega)$ is a random Brownian path. For a random variable Q on $[0,\infty)$ or $\mathbb{R}$, the distribution function F of Q is $F(q)=\mathcal{P}[Q<=q]$, with $0<=F<=1$; $\lim_{q\to 0} F(q) \to 0$; $\lim_{q\to\infty} F(q)=1$. The variable Q has the density $p(q)$ if

$$F(q) = \int_0^q p(v)\,dv, \quad \forall q, \tag{4.3}$$

and the expectation $\mathscr{E}[q]$ is

$$\mathscr{E}[q] = \int_{\mathbb{R}} q p(q)\,dq. \tag{4.4}$$

If the stochastic process $\{Q(\omega,t)\}$ is identified with $\{\beta(x,t)\}$ for $\beta \in \mathit{Riemm}(I(t)) \subset \mathbb{R}$ and $x \in I(t)$ and $t \in \mathbb{R}$, then

$$\{\beta(x,t)\} = \{\beta_1(x_1,t_1), \beta_2(x_2,t_2), \ldots, \beta_k(x_k,t_k), \ldots\} \tag{4.5}$$

is the set of parametrized, random metric fluctuations or Brownian paths about an initial deterministic metric $\beta(x,t)$ with $x\in I(t)$, $t\in\mathbb{R}$, and $\{x_1,x_2,\ldots,x_k,\ldots\}\in I(t)=[0,l]$ and Cauchy development $t<t_1<t_2<\cdots<t_k<\cdots$. These metric fluctuations all lie on the same orbit $\mathcal{S}(I)=\mathit{Riemm}(I)/\mathit{Diff}(I)$. For the classical (1+1) manifold, $\beta(x,t)$ is deterministic so that $\beta(x,t)=\beta(x,t_1)=\beta(x,t_2)=\cdots$ is a deterministic sample path generated by the Cauchy development. The density for the fluctuating metric is a functional of $\beta(x,t)$, so that we identify $p(q)$ with $\mathsf{K}[\![\beta(x,t)]\!]$. The differential variation in $\beta(x,t)$ is then simply $d\beta(x,t)=\partial_k\partial_x\beta(x,t)\,\delta x+\partial_t\beta(x,t)\,\delta t$. For classical, determinstic geometry, the first term is zero so that $d\beta(x,t)=\partial_t\beta(x,t)\,\delta t$. The $\mathsf{K}[\![\beta(x,t)]\!]$ is a scalar on $[0,1]$ and normalized. Note the notation $[\![\cdots]\!]$, for functionals of $\beta(x,t)$. The density or kernel $\mathsf{K}[\![\beta(x,t)]\!]$ is normalized on $[0,1]$ as

$$\int_{\mathit{Riemm}(I)\supset\mathbb{R}}\mathsf{K}[\![\beta(x,t)]\!]d\beta(x,t)=1, \tag{4.6}$$

for all $x\in I(t)$ and $t\in\mathbb{R}$. The kernel $\mathsf{K}[\![\beta(x,t)]\!]$ and the normalization is invariant under $\phi\in\mathit{Diff}(I)$, in that $\phi^*\mathsf{K}[\![l:\beta(\phi(x),t)]\!]=\mathsf{K}[\![l:\beta(x,t)]\!]$. Since $I(t)=[0,l]$, we can include the dependence on l via the notation $\mathsf{K}[\![l:\beta(x,t)]\!]$ so that $\mathsf{K}\rightarrow 1$ and is deterministic for some $l\rightarrow l_c$. The normalization and expectations $\mathscr{E}[\beta(x,t)]\equiv\langle\beta(x,t)\rangle$ and $\mathscr{E}[\Psi[\![\beta(x,t)]\!]]\equiv\langle\Psi[\![\beta(x,t)]\!]\rangle$ for a functional $\Psi[\![\beta(x,t)]\!]$ are defined as (at a given $t\in\mathbb{R}$)

$$\int_{\mathit{Riemm}(I)}\mathsf{K}[\![l:\beta(x,t)]\!]d\beta(x,t)=1, \tag{4.7a}$$

$$\mathscr{E}[\beta(x,t)]=\int_{\mathit{Riemm}(I)}\beta(x,t)\mathsf{K}[\![l:\beta(x,t)]\!]d\beta(x,t), \tag{4.7b}$$

$$\mathscr{E}[\Psi[\![\beta(x,t)]\!]]=\int_{\mathit{Riemm}(I)}\Psi[\![\beta(x,t)]\!]\mathsf{K}[\![l:\beta(x,t)]\!]d\beta(x,t). \tag{4.7c}$$

The expectations are taken as invariant under the group action $\phi\in\mathit{Diff}(I)$. If the kernel $\mathsf{K}[\![l:\beta(x,t)]\!]$ is invariant under all time shifts or Cauchy development $I(t)\rightarrow I(t+\delta t)\rightarrow I(t+2\delta t)\rightarrow\cdots$, then K is stationary over the Cauchy evolution. If a fluctuating metric has value $\beta'(x,t)$ for $x'\in I(t')$ and $t'\in\mathbb{R}$ and value $\beta(x,t)$ for $x\in I(t)$ and $t\in\mathbb{R}$ with $t>t'$, then the transition kernel is $\mathsf{K}[\![l:\beta(x,t)|\beta'(x',t')]\!]$. Of course, $I(t')=I(t)=[0,l]$ and $|x-x'|<=l$. For the smooth, classical, deterministic (1+1) geometry, we always have $\mathsf{K}[\![l:\beta(x,t)|\beta'(x',t')]\!]=0$ and $\mathsf{K}[\![[l:\beta'(x',t)|\beta'(x',t')]\!]=1$. Given $\phi\in\mathit{Diff}(I)$, we take it to hold that K and expressions derived from K are invariant under the action of ϕ for all $x\in I(t)$ and $t\in\mathbb{R}$. At some microlength scale l, the transition kernal is defined on $[0,1]$ for the parametrized stochastic process $\{\beta(x,t)\}$ for all $x\in I(t)$ and $t\in\mathbb{R}$.

V. CAUCHY EVOLUTION OF THE KERNEL: DERIVATION OF A FOKKER–PLANCK EQUATION

The quantum fluctuations in $I(t)$ will be modeled as a stochastic Weiner–Markov process[30,31] in real time $t\in\mathbb{R}$ on the (1+1) Lorentzian manifold $\mathscr{M}^2$. Quantum fluctuations of conventional fields (scalars and gauge) are considered a Markov–Weiner process in Refs. 1–3 so this should be viable for the metric, at least within a linear or quadratic approximation. A Markov process has no memory of its past for a given present so that only the initial data is required for prediction of the future evolution. This fits in nicely with the notion of classical deterministic Cauchy development for a finite time into the future. For data on a general spacelike surface $\Sigma(0)$ at $t=0$, the state at $\Sigma(t)$ can be ascertained. For $\Sigma(t)=I(t)$ and for initial stochastic probability $\mathsf{K}[\![l:\beta(x,t)]\!]$ for $x\in I(t)$ and $t\in\mathbb{R}$, only the future probability can be deduced from an iterative Chapman–Kolmogorov integral. As $t\rightarrow\infty$, we are interested in the stationary or (thermal) equilibrium limit solution and its associated correlations.

Let $\beta'(x',t)$ be the scalar classical metric component at $x' \in I(t)$ and $t \in \mathbb{R}$, which fluctuates to $\beta(x) = \beta'(x',t) + \delta\beta = \beta(x',t+\delta t)$ at $t = t + \delta t$ for small Cauchy variation δt or $I(t) \to I(t + \delta t)$. At a length scale l, the transition probability kernel on $[0,1]$ would be denoted $\mathsf{K}[\![l:\beta(x,t + \delta t)|\beta'(x',t)]\!]$. The Cauchy evolution is also smooth since we take it that $\delta t > t_*$, the Planck time. Since the fluctuations are taken to be Markovian, then in the limit of infinitesimal Cauchy variation δt, the Kolmogorov integral is

$$\mathsf{K}[\![l:\beta(x,t+\delta t)]\!] = \int d\beta'(x')\mathsf{K}[\![l:\beta(x,t+\delta t)|\beta'(x',t)]\!]\mathsf{K}[\![l:\beta'(x',t)]\!]. \tag{5.1}$$

The integral is understood to be over the space of all metrics on I. Given that we have the initial data that the metric has the value $\beta'(x,t)$ with respect to $x' \in I(t)$, then we predict the future probability with respect to $I(t+\delta t)$, that the metric randomly fluctuates to some $\beta(x,t+\delta t)$; the past values with respect to $I(t-\delta t)$ are considered irrelevant to the future evolution. All Brownian sample paths or histories are equivalent. Classically, Eq. (4.1) reduces to $0=(0)\times(1)$, for a classical metric component $\beta'(x',t)$ that has no fluctuations. The purely deterministic Cauchy evolution or sample path is $\beta'(x',t) \to \beta'(x',t+\delta t) = \beta'(x',t)$, since the Cauchy development is built into the Riemann 1-metric $\beta(x,t)$. Of course, one cannot guarantee that both diffeomorphism invariance and the Markov assumption are not lost at the Planck length, however, it will be taken that they still hold very near the scale.

Given $\mathsf{K}[\![l:\beta(x,t+\delta t)]\!]$ in the limit of infinitesimal δt, one can define the σth-order (diffusion) functionals $\mathscr{D}_\sigma[\![l:\beta'(x',t)]\!]$. This is always possible[30,31] for a generic Weiner–Markov process $\{K\}$ on $[0,1]$ obeying (5.1):

$$\mathscr{D}_\sigma[\![l:\beta(x',t)]\!] = [\sigma(\sigma-1)(\sigma-2)\cdots]^{-1}(\delta t)^{-1}| \times \int \mathsf{K}[\![l:\beta(x,t+\delta t)|\beta'(x',t)]\!]\left(\prod_{j=1}^{\sigma} |\beta_j(x_j - \beta'(x')|\right) d\beta(x). \tag{5.2}$$

Let $G:\mathbb{R}\to\mathbb{R}$ or $G:Riemm(I)\to\mathbb{R}$, be a functional of $\beta(x,t)$ at some $t \in \mathbb{R}$ so that $G[\![\beta(x)]\!] = G[\![\beta(x,t)]\!]$. The parametrized, stochastic, random 1-metrics are $\beta_j(x_j,t)$, for all $x_j \in I(t)$ and $t \in \mathbb{R}$. The functional Taylor expansion of $G[\![\beta(x)]\!]$ with respect to $G[\![\beta'(x')]\!]$ is

$$G[\![\beta(x)]\!] = G[\![\beta'(x')]\!] + \sum_{\sigma=1}^{\infty} [\sigma!]^{-1}\left(\prod_{j}^{\sigma} |\beta_j(x_j) - \beta'(x')| \frac{\delta}{\delta\beta_j(x_j)}\right) G[\![\beta'(x')]\!]. \tag{5.3}$$

The functional derivatives on $I(t)$ for $x_j \in I(t)$ are defined as

$$\prod_{j=1}^{1} \frac{\delta}{\delta\beta_j(x_j)} = \frac{\delta}{\delta\beta_1(x_1)}, \tag{5.4a}$$

$$\prod_{j=1}^{2} \frac{\delta}{\delta\beta_j(x_j)} = \frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}, \tag{5.4b}$$

$$\prod_{j=1}^{3} \frac{\delta}{\delta\beta_j(x_j)} = \frac{\delta^3}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)\,\delta\beta_3(x_3)}. \tag{5.4c}$$

For example, second-order 3-metric functional derivatives $\partial^2/\partial h_{ij}\,\partial h_{kl}$ occur in the Wheeler–DeWitt equation for the wave function $\Psi[h,\phi]$ of the universe.[26,32] To second order, the expansion of $G[\![\beta(x)]\!]$ is

$$G[\![\beta(x)]\!]=G[\![\beta'(x')]\!]+(\beta_1(x_1)-\beta'(x'))\ \frac{\delta}{\delta\beta_1(x_1)}\ G[\![\beta'(x')]\!]$$
$$+\frac{1}{2}\left((\beta_1(x_1)-\beta'(x'))\right)\left((\beta_2(x_2)-\beta'(x'))\right)\ \frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}\ G[\![\beta'(x')]\!]+\cdots\ . \tag{5.5}$$

We can derive a stochastic integrodifferential equation for the functional expectation $\langle G[\![\beta(x)]\!]\rangle$ to all orders in σ, and from this equation derive moment relations to any order. Multiply Eq. (5.1) by $G[\![\beta(x)]\!]$ and integrate over $d\beta(x)$. The $G[\![\beta(x)]\!]$ can be placed under the integral $\int d\beta'(x')$ and then replace by the right-hand side of the Taylor series expansion (5.3) for $G[\beta(x)]$:

$$\begin{aligned}\langle G[\![\beta(x)]\!]\rangle_{t+\delta t}&\equiv\langle G[\![\beta(x,t+\delta t)]\!]\rangle\\ &=\int d\beta(x)G[\![\beta\langle x)]\!]\mathsf{K}[\![l:\beta(x,t+\delta t)]\!]\\ &=\int\int\left[G[\![\beta'(x')]\!]+\sum_{\sigma=1}^{\infty}\left(\left(\prod_{j=1}^{\sigma}|\beta_j(x_j)-\beta'(x',t)|\right)\right.\right.\\ &\quad\left.\left.\times\left(\prod_{j=1}^{\sigma}\frac{\delta}{\delta\beta_j(x_j)}\right)G[\![\beta'(x')]\!]\right)\right]\\ &\quad\times\mathsf{K}[\![l:\beta(x,t+\delta t)|\beta'(x',t)]\!]\mathsf{K}[\![l:\beta'(x',t)]\!]d\beta(x)d\beta'(x').\end{aligned} \tag{5.6}$$

Using the normalization $\int \mathsf{K}[\![l:\beta(x,t+\delta t)|\beta'(x',t)]\!]d\beta'(x')=1$ and Eq. (5.2) for $\mathscr{D}_\sigma[\![l:\beta(x,t)]\!]$ gives a simpler expression:

$$\begin{aligned}\langle G[\![\beta(x)]\!]\rangle_{t+\delta t}=\langle G[\![\beta(x)]\!]\rangle_t+[\delta t]\sum_{\sigma=1}^{\infty}\left[\int d\beta'(x')\mathsf{K}[\![l:\beta'(x',t)]\!]\right.\\ \left.\times\left(\prod_{j=1}^{\sigma}\frac{\delta}{\delta\beta_j(x_j)}\left[G[\![\beta'(x')]\!]\mathscr{D}_j[\![l:\beta'(x',t)]\!]\right]\right)\right].\end{aligned} \tag{5.7}$$

Note that $\langle G[\![\beta(x)]\!]\rangle_t\equiv\langle G[\![\beta(x,t)]\!]\rangle$. Since $\langle G[\![\beta(x)]\!]\rangle_{t+\delta t}-\langle G[\![\beta(x)]\!]\rangle_t=\delta t\ \partial_t\langle G[\![\beta(x)]\!]\rangle$, the differential equation satisfied by the functional expectation $\langle G[\![\beta(x)]\!]\rangle$ to all orders in σ over the Cauchy evolution of $\beta(x,t)=\beta(x)$ is

$$\begin{aligned}\frac{\partial\langle G[\![\beta(x)]\!]\rangle}{\partial t}&=\sum_{\sigma=1}^{\infty}\left\langle\mathscr{D}_\sigma(l)\left(\prod_{j=1}^{\sigma}\frac{\delta}{\delta\beta_j(x_j)}\ G[\![\beta(x)]\!]\right)\right\rangle\\ &=\left\langle\mathscr{D}_1(l)\ \frac{\delta}{\delta\beta_1(x_1)}\ G[\![\beta(x)]\!]\right\rangle+\left\langle\mathscr{D}_2(l)\ \frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}\ G[\![\beta(x)]\!]\right\rangle+\cdots.\end{aligned} \tag{5.8}$$

In the equilibrium limit $\partial_t\langle G[\![\beta(x)]\!]\rangle=0$. The functional $\mathscr{D}_\sigma[\![l:\beta(x,t)]\!]$ is taken as homogeneous over all $x\in I(t)$ and $t\in\mathbb{R}$. The fundamental nature of the fluctuations in geometry should be equivalent within any small region anywhere in the universe and for any time. Hence, $\mathscr{D}_\sigma(l)=\text{const}$ and $\mathscr{D}_\sigma(l)\to 0$ as l increases beyond the fluctuation–correlation length scale. The coefficients should depend only on the relative positions $l=x-x'$ or length scale of the fluctuations, and not on the location of the fluctuation itself.

Let $x_j=x_1,x_2,x_3\in I(t)$. If we set $G[\![\beta_1(x,t)]\!]=\beta_1(x_1,t)$ then $G[\![\beta(x,t)]\!]=\beta_1(x_1,t)\beta_2(x_2,t)$ and $G[\![\beta(x,t)]\!]=\beta_1(x_1,t)\beta_2(x_2,t)\beta_3(x_3,t)\cdots$ in Eq. (4.8) we obtain moment relations to any order σ:

$$\partial_t\langle\beta_1(x_1,t)\rangle=-\langle\mathscr{D}_1(l)\rangle=\mathscr{D}_1(l), \tag{5.9a}$$

$$\partial_t \langle \beta_1(x_1,t)\beta_2(x_2,t)\rangle = 2\langle \beta_2(x_2,t)\mathscr{D}_1(l)\rangle + \langle \mathscr{D}_2(l)\rangle, \tag{5.9b}$$

and so on. To second order, the coefficient $\mathscr{D}_1 = \mathscr{D}_1(l)$ is a "drift" while $\mathscr{D}_2(l)$ is a "diffusion."

It is now possible to derive differential equations to all orders for $\{\mathsf{K}[\![:l\beta(x,t)]\!]\}$ at length scale l, where $0<\mathsf{K}<1$, and deduce the functional dependence on l. If we take the expectation of the Dirac Δ functional $G[\beta(x_j)] = \Delta[\beta(x)-\beta_j(x_j)]$, then

$$\begin{aligned}\langle G[\![\beta_j(x_j)]\!]\rangle &= \langle \Delta[\beta(x)-\beta_j(x_j)]\rangle \\ &\equiv \langle \Delta[\beta_j(x_j)-\beta(x)]\rangle \\ &= \int G[\![\beta_j(x_j)]\!]\Delta[\beta(x)-\beta_j(x_j)]d\beta_j(x_j) = G[\![l:\beta(x,t)]\!].\end{aligned} \tag{5.10}$$

Note that we use the notation "$\Delta\,[\beta(x)-\beta_j(x_j)]$" here to denote the Dirac Δ function rather than the conventional "$\delta[\beta(x)-\beta_j(x_j)]$" in order to avoid potential confusion with differential metric variations $\delta\beta$. If this choice is inserted into Eq. (5.8),

$$\begin{aligned}&\frac{\partial}{\partial t}\langle[\Delta(\beta(x_j)-\beta(x))]\rangle \\ &\quad\equiv \partial_t \mathsf{K}[\![l:\beta(x,t)]\!] \\ &\quad= \sum_{\sigma=1}^{\infty}\left[\mathscr{D}_\sigma(l)\left(\prod_{j=1}^{\sigma}\int d\beta_j(x_j)\,\frac{\delta}{\delta\beta_j(x_j)}\,\Delta[\beta(x)-\beta_j(x_j)]\mathsf{K}[\![l:\beta(x,t)]\!]\right)\right] \\ &\quad= \int \mathscr{D}_1(l)\int d\beta_1(x_1)\,\frac{\delta}{\delta\beta_1(x_1)}\,\Delta[\beta(x)-\beta_1(x_1)]\mathsf{K}[\![l:\beta(x,t)] \\ &\qquad+\mathscr{D}_2(l)\int d\beta_1(x_1)\int d\beta_2(x_2)\,\frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)} \\ &\qquad\times\Delta[\beta(x)-\beta_1(x_1)]\Delta[\beta(x)-\beta_2(x_2)]\mathsf{K}[\![l:\beta(x,t)]\!]+\cdots\,.\end{aligned} \tag{5.11}$$

Integrating (by parts) and using the filter properties of $\Delta[\beta(x)-\beta_j(x_j)]$ gives a differential equation for $\mathsf{K}[\![l:\beta(x,t)]\!]$ to all perturbation orders in σ:

$$\frac{\partial}{\partial t}\mathsf{K}[\![l:\beta(x,t)]\!] = \sum_{\sigma=1}^{\infty}\mathscr{D}_\sigma(l)\left(\prod_{\sigma=1}^{\sigma}\frac{\delta}{\delta\beta_j(x_j)}\mathsf{K}[\![l:\beta(x,t)]\!]\right) = -f\mathsf{K}[\![l:\beta(x,t)]\!]. \tag{5.12}$$

The equation can be expressed as

$$\frac{\partial}{\partial t}\mathsf{K}[\![l:\beta(x,t)]\!] = f\mathsf{K}[\![l:\beta(x,t)]\!] = \sum_{\sigma=1}^{\infty} f_\sigma \mathsf{K}[\![l:\beta(x,t)]\!], \tag{5.13}$$

where the functional differential operator f is defined by the infinite expansion:

$$f = \sum_{\sigma=1}^{\infty}\mathscr{D}_\sigma(l)\left(\prod_{j=1}^{\sigma}\frac{\delta}{\delta\beta_j(x_j)}\right) = \sum_{\sigma=1}^{\infty} f_\sigma\,. \tag{5.14}$$

Expanding out Eq. (5.12) in powers of σ gives

$$\frac{\partial}{\partial t}\mathsf{K}[\![l:\beta(x,t)]\!]=\left(\sum_{\sigma=1}^{\infty}f_{\sigma}\right)\mathsf{K}[\![l:\beta(x,t)]\!]$$
$$=\mathscr{D}_1(l)\,\frac{\delta}{\delta\beta_1(x_1)}\,\mathsf{K}[\![:l:\beta(x,t)]\!]+\mathscr{D}_2(l)\,\frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}\,\mathsf{K}[\![l:\beta(x,t)]\!]+\cdots\,. \tag{5.15}$$

If we truncate the expansion at order $\sigma=2$ in (5.15), we then obtain linear (Fokker–Planck) equations for kernels $\{\mathsf{K}[\![l:\beta(x,t)]\!]\}$.

VI. THE EQUILIBRIUM LIMIT SOLUTION AND ITS CORRELATIONS

Stationary or equilibrium solutions $\mathsf{K}_E[\![l:\beta(x)]\!]$ remain invariant under Cauchy development $I(t')\rightarrow I(t)$ for all $t>t'\in\mathbb{R}$ and $x\in I$. We also require that $(\partial_t\mathsf{K})=0$. One can also find the full time-dependent solution and take the limit $t\rightarrow\infty$ to obtain the equilibrium, stationary solution K_E as $\mathsf{K}[\![l:\beta(x,t)]\!]=\exp(ft)\mathsf{K}_E[\![l:\beta(x)]\!]$. The main idea behind stochastic quantization is to identify the (thermal) equilibrium or stationary correlations of a stochastic model with the equal time n-point Green's functions or correlations of the corresponding quantum theory. The stationary Fokker–Planck equation for the Cauchy-invariant fluctuation processes $\{\mathsf{K}\}$ in the geometry $I(t)$ for all $t\in\mathbb{R}$ is then

$$\left(\mathscr{D}_1(l)\,\frac{\delta}{\delta\beta_1(x_1)}\,\mathsf{K}_E[\![l:\beta(x)]\!]+\mathscr{D}_2(l)\,\frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}\,\mathsf{K}_E[\![l:\beta(x)]\!]\right)=0. \tag{6.1}$$

In the equilibrium limit, any functional $G[\![\beta(x)]\!]$ satisfies Eq. (5.8), with $\partial_t G[\![\beta(x)]\!]=0$,

$$\left\langle\mathscr{D}_1(l)\,\frac{\delta}{\delta\beta_1(x_1)}\,G[\![l:\beta(x)]\!]\right\rangle+\left\langle\mathscr{D}_2(l)\,\frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}\,G[\![l:\beta(x)]\!]\right\rangle=0. \tag{6.2}$$

Given a generic functional derivative $\delta/\delta q(x)$, there exists the trivial vanishing integral $\int dq(x)[\delta/\delta q(x)]=0$. It will be useful to express the stationary Fokker–Planck equation (6.1) in the equivalent form

$$\int d\beta_1(x_1)\mathscr{D}_1(l)\left(\frac{\delta}{\delta\beta_1(x_1)}\right)\mathsf{K}_E[\![l:\beta(x)]\!]+\int d\beta_1(x_1)\int d\beta_2(x_2)\mathscr{D}_2(l)$$
$$\times\left(\frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}\right)\mathsf{K}_E[\![l:\beta(x)]\!]=0. \tag{6.3}$$

The equilibrium solution is easily found to be

$$\mathsf{K}_E[\![l:\beta(x)]\!]=\left(\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\right)\exp\left[-\left(\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\right)\beta(x)\right], \tag{6.4a}$$

$$\mathsf{K}_E[\![l:\delta\beta(x)]\!]=\left(\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\right)\exp\left[-\left(\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\right)\delta\beta(x)\right], \tag{6.4b}$$

where $\delta\beta(x)\equiv\beta(x)-\beta_j(x_j)$ for any $x\in I$ and $x_j\in I$ with $|x-x'|<=l$. As $\mathscr{D}_1(l)\rightarrow 0$ and $\mathscr{D}_2(l)\rightarrow 0$ by appropriately varying the resolution or interval boundary l since $I=[0,l]$, the fluctuation distribution vanishes, so there is zero probability of a fluctuation to any $\beta_j(x_j)$ away from the classical deterministic value $\beta(x)$. The fluctuations decay exponentially so the probability of large fluctuations should diminish rapidly, at a scale determined by the ratio $(\mathscr{D}_1(l)/\mathscr{D}_2(l))$. When the functional derivative of $\mathsf{K}_E[\![l:\beta(x)]\!]$ is taken with respect to $\delta/\delta\beta_1(x_1)$ or $\delta/\delta\beta_2(x_2)$ for $x,x_1,x_2\in I$, the exponent brings down delta functions, since

$$\left(\frac{\delta}{\delta\beta_1(x_1)}\right)\beta(x)=\Delta[\beta(x)-\beta_1(x_1)]=\Delta[\beta_1(x_1)-\beta(x)], \tag{6.5a}$$

$$\left(\frac{\delta}{\delta\beta_2(x_2)}\right)\beta(x)=\Delta[\beta(x)-\beta_2(x_2)]=\Delta[\beta_2(x_2)-\beta(x)], \tag{6.5b}$$

since the Dirac Δ function is symmetrical.

Hence,

$$\begin{aligned}\mathscr{D}_1(l)\,\frac{\delta}{\delta\beta_1(x_1)}\,\mathsf{K}_E[\![l:\beta(x,t)]\!]&=-\frac{[\mathscr{D}_1(l)]^3}{[\mathscr{D}_2(l)]^2}\left(\frac{\delta}{\delta\beta_1(x_1)}\right)\beta(x)\exp\left[-\frac{\mathscr{D}_1}{\mathscr{D}_2}\,\beta(x)\right]\\&=-\frac{[\mathscr{D}_1(l)]^3}{[\mathscr{D}_2(l)]^2}\,\Delta[\beta(x)-\beta_1(x_1)]\exp\left[-\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\,\beta(x)\right].\end{aligned} \tag{6.6}$$

Similarly,

$$\begin{aligned}&\mathscr{D}_2(l)\left(\frac{\delta^2}{\delta\beta_1(x_1)\,\delta\beta_2(x_2)}\right)\mathsf{K}_E[\![l:\beta(x)]\!]\\&\quad=\frac{[\mathscr{D}_1(l)]^3}{[\mathscr{D}_2(l)]^2}\,\Delta[\beta(x)-\beta_1(x_1)]\Delta[\beta(x)-\beta_2(x_2)]\exp\left[-\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\,\beta(x)\right].\end{aligned} \tag{6.7}$$

The equilibrium Fokker–Planck equation (6.3) then becomes

$$\begin{aligned}0=&\int d\beta_1(x_1)\Delta[\beta(x)-\beta_1(x_1)]\exp\left[-\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\,\beta(x)\right]\\&+\int d\beta_1(x_1)\int d\beta_2(x_2)\Delta[\beta(x)-\beta_1(x_1)]\Delta[\beta(x)-\beta_2(x_2)]\exp\left[-\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\,\beta(x)\right],\end{aligned} \tag{6.8}$$

which is satisfied since

$$\int d\beta_1(x_1)\Delta[\beta(x)-\beta_1(x_1)]=1, \tag{6.9a}$$

$$\int d\beta_1(x_2)\Delta[\beta(x)-\beta_2(x_2)]=1. \tag{6.9b}$$

This is compatible with Eq. (5.10). Using this Cauchy-invariant or equilibrium solution K_E, one can calculate the correlations $\langle\beta(x)\rangle$ and $\langle\beta_1(x_1)\beta_2(x_2)\rangle$ in order to deduce the drift and diffusion coefficients $\mathscr{D}_1(l)$ and $\mathscr{D}_2(l)$. A generic correlation $C_{12}=\langle\varphi_1\varphi_2\rangle$ in field theory or statistical mechanics, generally has a scale dependence and C_{12} falls off as a power law or exponential $C_{12}\sim x^{-v}\exp(-x/\xi)$ for distances large in comparison to the correlation length ξ. For linearized general relativity treated as a quantum field theory, Eq. (2.9) gives equal time correlations, $\sqrt{[\langle\beta(x_1)\beta(x_2)\rangle]}=(8/\pi)(l_*|x_2-x_1|^{-1})$ or $\langle\beta_1(x_1)\beta_2(x_2)\rangle\sim(l_*/l)^2$, where l_* is the Planck length. Using (6.4a), the expectations $\mathscr{E}[\beta(x)]\equiv\langle\beta(x)\rangle$ and $\mathscr{E}[\beta_1(x_1)\beta_2(x_2)]\equiv\langle\beta_1(x_1)\beta_2(x_2)\rangle$ are derived from the equilibrium solution

$$\langle\beta(x)\rangle=\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\int\beta(x)\exp\left[-\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\,\beta(x)\right]d\beta(x), \tag{6.10a}$$

$$\langle\beta_1(x_1)\beta_2(x_2)\rangle=\frac{[\mathscr{D}_1(l)]^2}{[\mathscr{D}_2(l)]^2}\int\int\beta_1(x_1)\beta_2(x_2)\exp\left[-\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\,\beta_1(x_1)\right]$$
$$\times\exp\left[-\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\,\beta_2(x_2)\right]d\beta_1(x_1)\,d\beta_2(x_2). \tag{6.10b}$$

The form of the integral is straightforward: $\int_D^\infty q(x)\exp[-aq(x)]dq(x)=(1/a^2)$ for fixed x, where $a=[\mathscr{D}_1(l)/\mathscr{D}_2(l)]$. Hence,

$$\langle\beta(x)\rangle=\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\left(\frac{\mathscr{D}_2(l)}{\mathscr{D}_1(l)}\right)^2=\left(\frac{\mathscr{D}_2(l)}{\mathscr{D}_1(l)}\right), \tag{6.11a}$$

$$\langle\beta_1(x_1)\beta_2(x_2)\rangle=\left(\frac{\mathscr{D}_1(l)}{\mathscr{D}_2(l)}\right)^2\left(\frac{\mathscr{D}_2(l)}{\mathscr{D}_1(l)}\right)^3=\left(\frac{\mathscr{D}_2(l)}{\mathscr{D}_1(l)}\right)^2. \tag{6.11b}$$

The correlations in the stationary, equilibrium limit of the stochastic model can be compared with the equal time correlations estimated from the linearized Einstein theory [see (2.8) and (2.9)] treated as a quantum field theory: $\langle\beta_1(x_1)\beta_2(x_2)\rangle\sim(l_*/l)^2$, where $l=x_1-x_2$. Both results are derived at second order, which in the stochastic model was tantamount to truncation at $\sigma=2$ to yield a Fokker–Planck equation. One has $(\mathscr{D}_2(l)/\mathscr{D}_1(l))^2=(l_*/l)^2$, which is satisfied if we have $\mathscr{D}_1(l)=(l_*/l)$ and $\mathscr{D}_2(l)=(l_*/l)^2$. The maximum allowed spatial fluctuations of order $|x-x'|=l$ within I diminish rapidly as the resolution l is increased and the corresponding diffusion coefficient $\mathscr{D}_2(l)\sim(l_*/l)^2$ vanishes along with the drift coefficent $\mathscr{D}_1(l)\sim(l^*/l)$. Note, even on the nuclear scale, that we have $(l_*/l)\sim10^{-20}$. The equilibrium kernel is

$$\mathsf{K}_E[\![l:\delta\beta(x)]\!]=\frac{l}{l_*}\exp\left[-\frac{l}{l_*}\,\delta\beta(x)\right]. \tag{6.12}$$

This rapidly damps to zero for $l\gg l_*$. For a real probability on [0,1] we require

$$0\leqslant\frac{l}{l_*}\exp\left[-\frac{l}{l_*}\,\delta\beta(x)\right]<1. \tag{6.13}$$

Clearly, $(l/l_*)>0$ is always satisfied and $\exp[-(l/l_*)\delta\beta(x)]<=(l_*/l)<=1$ if $l=l_*$ is a lower bound, such that $I\rightarrow I_*=[0,l_*]$. This gives the uncertainty relation $l\,\delta\beta(x)>=l_*$ or $\delta\beta(x)>=(l_*/l)$. The fluctuations rapidly vanish for $l\gg l_*$ and the kernel is defined on [0,1] only if $l>=l_*$. For $l\gg l_*$, the probabiity has a sharp peak at $\delta\beta=0$. The expansion or contraction of the spatial interval $[0,l]$ then damps or amplifies the fluctuations in the 1-metric. As l is reduced toward l_*, the smooth interval $I=[0,l]$ becomes increasingly ''fuzzy'' and the continuous, deterministic geometry is lost. Also, as l gets very near to l_*, higher-order correlations emerge between three $(x_1,x_2,x_3)\in I$ or more points $(\sigma>=2)$ within I, with $\mathscr{D}_\sigma(l)\sim(l_*/l)^\sigma$. Equation (5.8) is a linear approximation to order $\sigma=2$. This situation is somewhat analogous to the kinetic theory of a gas. As the density ρ increases, the probability for a larger number of molecules to collide at the same point increases. The expansion parameter is the density $\rho\sim(1/V)\sim(1/l^3)$ or $\rho\sim(1/l)$ for a one-dimensional gas. For the fluctuations in geometry, the expansion parameter is similarly $\mathscr{D}_\sigma(l)\sim(l_*/l)^\sigma$. However, in the classical limit as $l\gg1_p$ or $\hbar\rightarrow0$, then $l_*=(G\hbar)^{1/2}\rightarrow0$ and all $\mathscr{D}_\sigma(l)\rightarrow0$ and so $\mathsf{K}_E[\![\beta(x)]\!]\rightarrow0$. Correlations to all orders vanish and the geometry returns to being deterministic at all length scales.

Interestingly, a similar mathematical structure to (6.12) can be obtained for a fluctuation amplitude in standard Euclidean path integral quantum gravity. Padmanabhan[33,34] has deduced the existence of a minimum length within quantum gravity, where the conformal factor is quantized. Beginning with a conformally flat metric $g_{uv}=[1+\phi(x)^2]\eta_{uv}$, the Euclidean path integral over the conformal fluctuation ϕ is

$$\int \mathscr{D}\phi \exp\left(-\frac{i}{l_*^2}\int d^4x\ \eta_{uv}\ \nabla_u\phi\ \nabla_v\phi\right), \tag{6.14}$$

which can be evaluated in a closed form. The resulting amplitude for a measurement of the conformal fluctuation having value ϕ is

$$\mathcal{A}(\phi)=\left(\frac{l}{l_*}\right)^{1/4}\exp\left(-\left(\frac{l}{l_*}\right)^2\phi^2\right). \tag{6.15}$$

This amplitude behaves in a very similar way to the probability in (5.6) and damps out for $l \gg l_*$.

VII. CONCLUSIONS

A basic stochastic construction has been presented to describe Planck scale fluctuations in the geometry of a simple globally hyperbolic $(1+1)$ manifold. The emphasis has been on the conceptual issues and potential of such an approach rather than on technical subtleties, but the scheme is mathematically self-consistent and makes contact with basic physical features of metric fluctuations and quantum geometry. The second-order stationary equilibrium solution of the Fokker–Planck equation derived from the construction makes contact with the quantum field theoretic predictions for metric fluctuations and equal time correlations in linearized general relativity. It is predicted that probabilities for (linear) stochastic metric component fluctuations on the interval $[0,l]$ diminish exponentially with a scale dependence $l \gg l_*$. As l increases above l_*, the second-order solution illustrates that the spatial geometry evolves from being smooth and deterministic to stochastic or noisy, with a maxima at the Planck length. This is the so-called quantum ''foam'' structure. Away from the Planck length, the spatial geometry returns to a smooth and deterministic condition. Clearly, there are many highly nontrivial technical issues that arise in any approach to a quantum interpretation of space–time or gravitation. These include problems in defining general path measures and the potential for ill-defined expressions. However, the point of view presented is that a basic construction of this kind, with the emergence of a well-known and expected physical consequence of quantum gravity and geometry, is a necessary prerequisite to more detailed approaches, which may have the potential to extend the scope of quantum geometry.

ACKNOWLEDGMENTS

I would like to thank David Miller and Marla Manning.

[1] M. Namiki, *Stochastic Quantization* (Springer-Verlag, Berlin, 1993).
[2] *Stochastic Quantization*, edited by P. H. Damgaard and H. Huffel (World Scientific, Singapore, 1987).
[3] P. H. Damgaard and H. Huffel, Phys. Rep. **152**, 227 (1987).
[4] D. Bohm, Phys. Rev. **85**, 166 (1952).
[5] E. Neslon, Phys. Rev. **150**, 1079 (1966).
[6] H. Huffel and H. Rumpf, Phys. Lett. B **14**, 104 (1984).
[7] E. Gozzi, Phys. Lett. B **14**, 104 (1985).
[8] H. Nakazata and Y. Yamanak, Phys. Rev. D **34**, 492 (1986).
[9] B. S. DeWitt, Phys. Rev. **160**, 1113 (1967).
[10] A. Ashtekhar, Phys. Rev. D **36**, 1587 (1987).
[11] A. Ashtekhar, C. Rovelli, and L. Smolin, Phys. Rev. Lett. **69**, 237 (1992).
[12] C. Rovelli and L. Smolin, Nucl. Phys. B **331**, 80 (1990).
[13] *Euclidean Quantum Gravity*, edited by S. W. Hawking and G. W. Gibbons (World Scientific, Singapore, 1993).
[14] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, 1987).
[15] E. Witten, Nucl. Phys. B **443**, 85 (1995).
[16] P. Horava and E. Witten, Nucl. Phys. B **475**, 94 (1996).
[17] C. Isham, ''Structural issues in quantum gravity,'' gr-qc/9510063, 1996.
[18] L. J. Garay, Int. J. Mod. Phys. A **10**, 145 (1995).
[19] J. W. Moffat, gr-qc/961007, University of Toronto preprint, 1997; Phys. Rev. D **56**, 6264 (1997).
[20] J. A. Wheeler, Phys. Rev. **97**, 511 (1955).
[21] J. A. Wheeler, *Geometrodynamics* (Academic, New York, 1962).

[22] H. Rump, Phys. Rev. D **33**, 942 (1986).
[23] B. Hatfield, *Quantum Field Theory of Point Particles and Strings*, Frontiers in Physics, No. 75 (Addison-Wesley, Reading MA, 1990).
[24] S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).
[25] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Spacetime* (Cambridge University Press, Cambridge, 1973).
[26] E. Giampiero, *Quantum Gravity, Quantum Cosmology and Lorentzian Geometries* (Springer-Verlag, Berlin, 1993).
[27] T. Jacobson, Phys. Rev. Lett. **75**, 1260 (1995).
[28] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications* (Springer-Verlag, Berlin, 1993).
[29] D. Williams, *Probability with Martingales* (Cambridge University Press, Cambridge, 1991).
[30] J. L. Doob, *Stochastic Processes* (Wiley, New York, 1953).
[31] M. Lax, Rev. Mod. Phys. **32**, 25 (1960).
[32] J. B. Hartle and S. W. Hawking, Phys. Rev. D **28**, 2960 (1983).
[33] T. Padmanabhan, Gen. Relativ. Gravit. **17**, 215 (1985).
[34] T. Padmanabhan, Class. Quantum Grav. **3**, 911 (1986).

On rotational coherent states in molecular quantum dynamics

Jorge A. Morales, Erik Deumens, and Yngve Öhrn
Quantum Theory Project, Departments of Chemistry and Physics, University of Florida, Gainesville, Florida 32611-8435

(Received 5 June 1998; accepted for publication 16 November 1998)

Coherent states suitable for the description of molecular rotations are developed and their connection to similar coherent states in the literature are explored. In particular their quasiclassical properties are developed. The use of such coherent states in time-dependent electron nuclear dynamics studies of molecular collision processes is discussed.
[S0022-2488(99)04602-2]

I. INTRODUCTION

Coherent states (CS) are a set of elements $\{|\mu\rangle\}$ of a Hilbert space $\mathcal{H}$. All CS share two properties in common:[1]
(1) continuity, i.e., the states $|\mu\rangle$ are continuous functions[1] of a parameter set μ,

$$\lim_{\mu\to\mu_0}|\mu\rangle=|\mu_0\rangle, \tag{1}$$

(2) resolution of the identity, i.e., there exists a positive measure $d\mu_+\geqslant 0$ for which

$$1=\int d\mu_+|\mu\rangle\langle\mu|. \tag{2}$$

There exists a weaker formulation of the second property which will allow a larger class of CS:[1]
$(2')$ The closed linear span of $\{|\mu\rangle\}$ is the Hilbert space $\mathcal{H}$. This means that any state vector in the Hilbert space may be represented as a (possibly infinite) linear sum of CS.[1] Such CS may satisfy a resolution of the identity with an indefinite measure $d\mu_\pm$,

$$1=\int d\mu_\pm|\mu\rangle\langle\mu|. \tag{3}$$

Both in the stronger and the weaker definitions, the CS form a nonorthogonal overcomplete set.

There are a great variety of CS known and used in various areas of physics. For problems in molecular physics and in chemistry the canonical CS[2,1] also referred to as Glauber states[3] are commonly used.[4] These states $\{|\alpha\rangle\}$ are associated with the harmonic oscillator Hamiltonian $H_{\mathrm{vib}}=\hbar\omega(a^\dagger a+\frac{1}{2})$, where ω is the angular frequency. The harmonic oscillator creation operators can be expressed as

$$a^\dagger=\frac{1}{\sqrt{2}}\left(\sqrt{\frac{m\omega}{\hbar}}\hat{x}-\frac{i}{\sqrt{m\hbar\omega}}\hat{p}\right) \tag{4}$$

in terms of the self-adjoint operators of position $\hat{x}$ and momentum $\hat{p}$, where m is the oscillator mass. The complex parameter α can be expressed in terms of the real parameters of average position $x_\alpha=\langle\alpha|\hat{x}|\alpha\rangle$ and average momentum $p_\alpha=\langle\alpha|\hat{p}|\alpha\rangle$ as

$$\alpha=\frac{1}{\sqrt{2}}\left(\sqrt{\frac{m\omega}{\hbar}}x_\alpha-\frac{i}{\sqrt{m\hbar\omega}}p_\alpha\right). \tag{5}$$

An expansion in terms of harmonic oscillator stationary states $\{|n\rangle, n=0,1,...\}$ exists,

$$|\alpha\rangle=\exp\left(-\frac{1}{2}|\alpha|^2\right)\sum_{n=0}^{\infty}\frac{\alpha^n}{\sqrt{n!}}|n\rangle=\exp\left(-\frac{1}{2}|\alpha|^2\right)\exp(\alpha a^\dagger)|0\rangle, \tag{6}$$

from which the resolution of the identity is readily proven with the positive measure

$$d\mu_+(\alpha)=\frac{1}{\pi}d\ \mathrm{Re}\ \alpha d\ \mathrm{Im}\ \alpha. \tag{7}$$

The spin coherent states $\{|\beta\rangle\}$, with a complex parameter β, constitute another example of CS used in molecular physics.[1] These states are associated with the total spin angular momentum $\vec{S}=(S_x S_y S_z)$ and an expansion in terms of spin eigenstates $\{|SM\rangle, S=0,1/2,1,...; M=S, S-1,...-S\}$ exists,

$$|\beta\rangle=\sum_{M=-S}^{S}\sqrt{\frac{(2S)!}{(S-M)!(S+M)!}}\left[\frac{\beta^{S+M}}{(1+|\beta|^2)^s}\right]|SM\rangle=\frac{1}{(1+|\beta|^2)^s}\exp(\beta S_+)|S\ -S\rangle, \tag{8}$$

where $S_\pm=S_x\pm iS_y$. The resolution of the identity exists with the positive measure

$$d\mu_+(\beta)=\frac{2S+2}{\pi(1+|\beta|^2)^2}d\ \mathrm{Re}\ \beta d\ \mathrm{Im}\ \beta. \tag{9}$$

It suffices here to mention as a third example the fermion CS,[1] also known as the Thouless determinant.[5,6] These CS are used, e.g., in the description of many-electron systems.[4] For N electrons in a basis of rank $K\geqslant N$ the normalized Thouless CS $\{|z\rangle\}$ can be expressed as

$$|z\rangle=\det(I+z^\dagger z)^{-1/2}\exp\left[\sum_{h=1}^{N}\sum_{p=N+1}^{K}z_{ph}b_p^\dagger b_h\right]|\Psi_0\rangle, \tag{10}$$

where z denotes the set of complex parameters $\{z_{ph}\}$, the $b_i^\dagger$ and b_i are the fermion creation and annihilation operators, respectively, and where

$$|\Psi_0\rangle=\prod_{i=1}^{N}b_i^\dagger|\mathrm{vac}\rangle. \tag{11}$$

The resolution of the identity exists with the positive measure

$$d\mu_+(z)=\eta\det(I+z^\dagger z)^{-K}d^2z, \tag{12}$$

where

$$d^2z=\frac{1}{\pi}\prod_{ph}d\ \mathrm{Re}\ z_{ph}d\ \mathrm{Im}\ z_{ph}, \tag{13}$$

and

$$\eta=\frac{1!2!\cdots K!}{1!2!\cdots(K-J)!1!2!\cdots J!}. \tag{14}$$

A set of coherent states may be related to a particular Lie group. Rasetti[7] and Solomon[8] have made seminal contributions to the theory of group-related CS. Perelomov[9] introduced a systematic procedure for the construction of such group-related CS. For instance, the canonical CS of the harmonic oscillator are related to the Weyl group, the spin CS to the special unitary group SU(2), and the Thouless CS to the unitary group U(K). There are, however, important CS that are not group related. The construction of coherent states requires a portion of mathematical intuition.

II. QUASICLASSICAL COHERENT STATES

A prominent property of many CS is their quasiclassical dynamics. A state $|\psi\rangle$ is said to be quasiclassical when the evolution of average position, momenta, and energy,

$$x_{\mathrm{qc}}=\langle\psi|\hat{x}|\psi\rangle,\quad p_{\mathrm{qc}}=\langle\psi|\hat{p}|\psi\rangle,\quad H_{\mathrm{qc}}=\langle\psi|\hat{H}|\psi\rangle, \tag{15}$$

satisfy classical Hamilton equations, i.e.,

$$\dot{x}_{\mathrm{qc}}=\frac{\partial H_{\mathrm{qc}}}{\partial x_{\mathrm{qc}}},\quad \dot{p}_{\mathrm{qc}}=-\frac{\partial H_{\mathrm{qc}}}{\partial p_{\mathrm{qc}}}. \tag{16}$$

In other words, the average position and momentum of the quasiclassical state evolve in time as the position and momentum of their classical analogs. One should note that the definition of a quasiclassical state does not demand the semiclassical limit $\hbar\to 0$ to be invoked. Neither is there *a priori* any guarantee that a quasiclassical state even exists for a given Hamiltonian. Ehrenfest's theorem[10] offers a means to investigate the quasiclassical property, i.e., the equations

$$\begin{aligned} i\hbar\frac{d}{dt}\langle\psi|\hat{x}|\psi\rangle&=\langle\psi|[\hat{x},\hat{H}]|\psi\rangle,\\ i\hbar\frac{d}{dt}\langle\psi|\hat{p}|\psi\rangle&=\langle\psi|[\hat{p},\hat{H}]|\psi\rangle \end{aligned} \tag{17}$$

should reduce to the classical ones of Eq. (16) for the state $|\psi\rangle$ to be quasiclassical.

In this manner it is straightforward to show that the canonical CS of Eq. (6) are quasiclassical. In particular,

$$\begin{aligned} \langle\alpha|\hat{x}|\alpha\rangle&=x_{\alpha}(t)=\sqrt{\frac{2\hbar}{m\omega}}\,\mathrm{Re}[\alpha\exp(-i\omega t)],\\ \langle\alpha|\hat{p}|\alpha\rangle&=p_{\alpha}(t)=\sqrt{2m\hbar\omega}\,\mathrm{Im}[\alpha\exp(-i\omega t)] \end{aligned} \tag{18}$$

and the total energy using the harmonic oscillator Hamiltonian is

$$E_{\alpha}\equiv H_{\mathrm{qc}}=\langle\alpha|H_{\mathrm{vib}}|\alpha\rangle=\hbar\omega|\alpha|^2+\frac{\hbar\omega}{2}=E_c^{\alpha}+\frac{\hbar\omega}{2}, \tag{19}$$

where

$$E_c^{\alpha}=\frac{1}{2m}p_{\alpha}^2+\frac{1}{2}m\omega^2x_{\alpha}^2 \tag{20}$$

is the classical energy of the harmonic oscillator. This particular set of CS satisfies the minimum uncertainty relation

$$\Delta x_{\alpha}(t)\Delta p_{\alpha}(t)=\hbar/2, \tag{21}$$

where the widths $\Delta x_\alpha(t)$ and $\Delta p_\alpha(t)$ are

$$\Delta x_\alpha(t)=\Delta x_\alpha=\sqrt{\frac{\hbar}{2m\omega}},\quad \Delta p_\alpha(t)=\Delta p_\alpha=\sqrt{\frac{m\hbar\omega}{2}}. \tag{22}$$

The coordinate representation of the canonical CS is

$$\psi_\alpha(x,t)=\langle x|\alpha(t)\rangle=\exp(i\theta_\alpha(t))\left(\frac{m\omega}{\pi\hbar}\right)^{1/4}\exp(-i\omega t)\exp\left(-\left(\frac{x-x_\alpha}{2\Delta x_\alpha}\right)^2\right)\exp\left(\frac{ip_\alpha(t)x}{\hbar}\right), \tag{23}$$

where $\theta_\alpha(t)$ is a global phase. The spin CS, Eq. (8), are quasiclassical with respect to a Hamiltonian describing the spin dynamics under a time-dependent magnetic field.[11,1] Minimum uncertainty conditions are also known for this CS.[11,1] There are CS that do not exhibit the quasiclassical property. The Thouless CS is not a quasiclassical state. However, it is possible to obtain classical-like equations for the Thouless parameters via the time-dependent variational principle (TDVP).[4,12]

III. ELECTRON NUCLEAR DYNAMICS AND COHERENT STATES

In this section, we make the connection between the CS and the electron nuclear dynamics (END) theory.[13,4,12] The END wave function[4] is

$$\Psi_{\mathrm{END}}(t)=\frac{1}{N_{\mathrm{nucl}}}F_{\mathrm{nucl}}[\mathbf{R}(t),\mathbf{P}(t)]f_{\mathrm{el}}[\mathbf{z}(t),\mathbf{R}(t)]\exp\left[\frac{i}{\hbar}\gamma_{\mathrm{END}}(t)\right], \tag{24}$$

where $\gamma_{\mathrm{END}}(t)$ is the total phase. At the simplest level of approximation the nuclear part F_{nucl} is the product of Gaussian wave packets of positions $\mathbf{R}(t)$ and momenta $\mathbf{P}(t)$,

$$F_{\mathrm{nucl}}(\mathbf{X};\mathbf{R},\mathbf{P})=\prod_{k=1}^{\mathrm{nucl}}\exp(-a_k[\mathbf{X_k}-\mathbf{R_k}(t)]^2+i\mathbf{P_k}(t)\cdot[\mathbf{X_k}-\mathbf{R_k}(t)]), \tag{25}$$

and the electronic part $f_{\mathrm{el}}[z_{ph}(t),\mathbf{R}(t),]=|z\rangle$ is the fermion (Thouless) CS shown above. The very role of the fermion CS is to provide a nonredundant and continuous parametrization of the single-determinant electronic wave function. It should be noted that the total system END wave function is given in the laboratory frame and includes translational and overall rotational motion. Using this approximate $\Psi_{\mathrm{END}}(t)$ and the TDVP a set of classical Hamilton-like equations are obtained for the Thouless parameters $\mathbf{z}(t)$ and $\mathbf{z}^*(t)$.[13,14,4,12] In order to obtain the proper END equations for this level of approximation the quantum mechanical Lagrangian is first obtained in the limit of zero width nuclear Gaussian wave packets. This approach leads to a classical treatment of the nuclei that retains the nonadiabatic electron–nuclear coupling terms. This approximation may be described as full, nonlinear time-dependent Hartree–Fock (TDHF) for electrons and narrow wave packet nuclei.

The time propagation of a molecular system undergoing a reaction may produce a set of product fragments. One important aim of molecular reaction dynamics is to predict the distribution of products over rovibrational states. The treatment of such a reactive process at the simplest END level of approximation leads to a fragment in some electronic state with its system of nuclei vibrating and rotating as a classical object. It has been demonstrated how the END representation of the nuclear part of such a fragment under very general conditions factors as

$$F_{\mathrm{nucl}}[\mathbf{X};\mathbf{R}(t),\mathbf{P}(t)]\approx F_0F_{\mathrm{vib}}F_{\mathrm{rot}}. \tag{26}$$

Viewing this wave function, even in the narrow wave packet limit, as an evolving state and representing this state in terms of suitable CS makes possible an *a posteriori* quantum state resolution of the nuclear dynamics. In this way we adopt the approximate labeling of product states in terms of vibrational (harmonic oscillator) quantum numbers and rotational (rigid rotor) quantum numbers. Obviously, more ambitious CS could be attempted, but this level of sophistication seems reasonable for low energy reactive collisions.

The *a posteriori* vibrational analysis in terms of harmonic oscillator CS has been outlined (see Ref. 12) and applied to obtain vibrationally resolved differential cross sections for proton collisions with hydrogen molecules at 30 eV,[15] in excellent agreement with experiment. A corresponding analysis for a quasiclassical treatment of the rotational dynamics is an interesting prospect and the necessary development is discussed in this paper.

IV. ROTATIONAL COHERENT STATES

The term rotational CS denotes those CS which are quasiclassical with a field-free rotor Hamiltonian. It is important to emphasize that the previously discussed spin CS is not a rotational CS by virtue of the preceding definition. Most of the rotational CS known in the literature stem in some way from the spin CS. Except for one case discussed below,[16] the majority of the rotational CS concerns the description of the linear rotor.

The first known rotational CS were derived by Atkins and Dobson.[17] The Atkins–Dobson CS are group generated by the Schwinger boson operators of the angular momentum,[18] have a positive measure, and can in principle be applied to linear rotors. An interesting and more useful contribution to the theory of rotational CS was made by Janssen,[16,1] who constructed rotational CS for the general asymmetric rotor. Janssen CS $\{|xyz\rangle\}$ can be expressed as

$$|xyz\rangle=\sum_{IMK} J_{IMK}(x,y,z)|IMK\rangle, \tag{27}$$

where $|IMK\rangle$; $I=0,\frac{1}{2},1,...$; M, $K=0,\pm\frac{1}{2},\dots\pm I$ are the integer (boson) and half-integer (fermion) rotational states associated with the asymmetric rotor Hamiltonian, x, y, and z the CS parameters, and $J_{IMK}(x,y,z)$ a set of coefficients. Janssen CS are not group generated and have an indefinite measure. These CS satisfy quasiclassical dynamical equations when evolved by the asymmetric rotor Hamiltonian in the Hilbert space spanned by the states $|IMK\rangle$. An interesting feature is that the Janssen CS are identical to the Atkins–Dobson CS in the limit of the linear rotor. The proof of that identity involves a proper reparametrization of Janssen CS. Both sets of CS have the property of mixing half-integer and integer quantum numbers. Therefore, they are not directly useful for the discussion of molecular rotational spectra for which a corresponding development in terms of only integer quantum numbers is necessary.

Similar developments were published by Bhaumik *et al.*, but again for the case of the linear rotor states $|IMK=0\rangle$. A review of some rotational CS for linear rotors was published by Fonda *et al.*[19] This study discusses the Atkins–Dobson CS among many others but misses those by Janssen and Bhaumik *et al.* More importantly, new CS generated by the $SO(3)\otimes R^5$ group are developed there to study diatomic molecules in the presence of an electromagnetic field.

In this section we introduce a new set of rotational CS following closely the definition of Janssen. Only integer quantum numbers are used for these CS and their quasiclassical behavior is analyzed.

A. Rotor Hamiltonian

The pure rotor Hamiltonian for a molecular system can be written as[20]

$$\hat{H}_{\rm rot}=\frac{\hat{L}_x^2}{2A_x}+\frac{\hat{L}_y^2}{2A_y}+\frac{\hat{L}_z^2}{2A_z}, \tag{28}$$

where A_i $(i=x,y,z)$ are the moments of inertia and $\hat{L}_i$ the body-fixed components of the orbital angular momentum. Please note that from here on $\hbar=1$. The analogous space-fixed components of orbital angular momentum are $\hat{J}_i$ and the following relations hold $(\hat{J}^2=\hat{L}^2)$:

$$[\hat{J}_i,\hat{J}_j]=i\epsilon_{ijl}\hat{J}_l, \quad [\hat{L}_i,\hat{L}_j]=-i\epsilon_{ijl}\hat{L}_l \tag{29}$$

and

$$[\hat{J}_i,\hat{L}_k]=[\hat{L}^2,\hat{L}_z]=[\hat{J}^2,\hat{J}_z]=0, \tag{30}$$

where ϵ_{ikl} are the components of the Levi–Civita tensor. Note the so-called anomalous commutation relationship[20] of the $\hat{L}_i$ components. As a result of these commutation relationships, there exists a complete set of rotor eigenstates $|IMK\rangle$ so that

$$\begin{aligned}
\hat{L}^2|IMK\rangle &= I(I+1)|IMK\rangle, \quad I=0,1,2,...,\\
\hat{L}_z|IMK\rangle &= K|IMK\rangle, \quad K=0,\pm 1,...,\pm I,\\
\hat{J}_z|IMK\rangle &= M|IMK\rangle, \quad M=0,\pm 1,...,\pm I.
\end{aligned} \tag{31}$$

These rotor eigenstates in angular representation are

$$\langle \phi,\theta,\chi|IMK\rangle=\left[\frac{2I+1}{8\pi^2}\right]^{1/2} D^{I*}_{MK}(\phi,\theta,\chi), \tag{32}$$

where $D^I_{MK}(\phi,\theta,\chi)$ are elements of a rotation matrix (Wigner D functions).[20]

It follows from the above commutator relations that the rotational Hamiltonian satisfies

$$[\hat{H}_{\mathrm{rot}},\hat{J}_i]=0, \quad [\hat{H}_{\mathrm{rot}},\hat{J}^2]=0. \tag{33}$$

The Hamiltonian eigenfunctions Ψ^α_{IM} must satisfy

$$\begin{aligned}
\hat{H}_{\mathrm{rot}}\Psi^\alpha_{IM} &= E^\alpha_{\mathrm{rot}}\Psi^\alpha_{IM},\\
\hat{J}^2\Psi^\alpha_{IM} &= I(I+1)\Psi^\alpha_{IM}, \quad I=0,1,2,...,\\
\hat{J}_z\Psi^\alpha_{IM} &= M\Psi^\alpha_{IM}, \quad M=0,\pm 1,...,\pm I,
\end{aligned} \tag{34}$$

where the superscript α is an additional label of a particular rotational eigenstate. Another set of relations implied by the above commutation relations is

$$[\hat{H}_{\mathrm{rot}},\hat{L}_i]=i\sum_j \frac{\epsilon_{ijk}}{2A_k}(\hat{L}_j\hat{L}_k+\hat{L}_k\hat{L}_j). \tag{35}$$

The Ψ^α_{IM} eigenfunctions are expressed in the symmetric rotor basis as

$$\Psi^\alpha_{IM}=\sum_K c^{IM\alpha}_K|IMK\rangle, \tag{36}$$

where the coefficients $c^{IM\alpha}_K$ are to be determined.

In the special case of a spherical rotor $A=A_x=A_y=A_z$ (e.g., CH_4 and SF_6), the eigenvalue problem simplifies to

$$\hat{H}_{\mathrm{rot}}=\frac{\hat{L}^2}{2A}=\frac{\hat{J}^2}{2A}, \quad \Psi^\alpha_{IM}=\Psi_{IM}=|IMK\rangle, \quad E^{IM\alpha}_{\mathrm{rot}}=E^I_{\mathrm{rot}}=\frac{I(I+1)}{2A}. \tag{37}$$

In the case of a prolate symmetric rotor, the moments of inertia satisfy $A_x \leqslant A_y = A_z$ (e.g., CH_3Cl and PCl_5). The eigenvalue problem then becomes

$$\hat{H}_{\mathrm{rot}} = \frac{\hat{L}^2}{2A_z} + \left(\frac{1}{2A_x} - \frac{1}{2A_z} \right) \hat{L}_z^2,$$

$$\Psi_{IM}^{\alpha} = \Psi_{IM} = |IMK\rangle, \tag{38}$$

$$E_{\mathrm{rot}}^{IM\alpha} = E_{\mathrm{rot}}^{IM} = \frac{I(I+1)}{2A_z} + \left(\frac{1}{2A_x} - \frac{1}{2A_z} \right) K^2.$$

The equivalent expressions for the case of an oblate symmetric rotor $A_x = A_y \leqslant A_z$ (e.g., $CHCl_3$ and C_6H_6) can be obtained by interchanging the A_x with the A_z in the last equations. The case of the linear rotor (e.g., all diatomics, CO_2, and C_2H_2) is obtained as the $A_x = 0$ limit of the prolate symmetric case. Then

$$\hat{H}_{\mathrm{rot}} = \frac{\hat{J}^2}{2A}, \quad \Psi_{IM}^{\alpha} = \Psi_{IM} = |IM0\rangle,$$

$$E_{\mathrm{rot}}^{IM\alpha} = E_{\mathrm{rot}}^{I} = \frac{I(I+1)}{2A}, \quad \langle \theta, \phi | IM0 \rangle = Y_{IM}(\theta, \phi), \tag{39}$$

where the $Y_{IM}(\theta,\phi)$ are the spherical harmonics.[20] Finally, in the case of an asymmetric rotor, with the moments of inertia satisfying $A_x \leqslant A_y \leqslant A_z$ (e.g., CH_2H_2), the eigenfunctions Ψ_{IM}^{α} keep their linear combination form, and the $c_K^{IM\alpha}$ coefficients must be specifically calculated.

B. Groups

Although the set of CS under construction is not group related in the Perelomov sense, it is, of course, connected to the rotation groups. Specifically, the states $|IMK\rangle$ span the irreducible representations of the semidirect product of SO(3)⊗SO(3) with an Abelian group. The generators of the first SO(3) group are the $\hat{L}_i$, referring to the molecule-fixed frame, while those of the second one are the $\hat{J}_i$ referring to the space-fixed frame. The generators of the Abelian group $R^{(2\lambda+1)^2}$ belong to a family of tensor operators $\hat{T}_{\mu\nu}^{\lambda}$ ($\lambda = 0, \frac{1}{2}, 1, \frac{3}{2}, \ldots;\ \mu, \nu = 0, \pm\frac{1}{2}, \ldots, \pm\lambda$). We select the tensors with $\lambda = 1$ in order to limit the CS to integer rotational quantum numbers. These tensor operators commute among themselves, i.e.,

$$[\hat{T}_{\mu\nu}^{\lambda}, \hat{T}_{\mu'\nu'}^{\lambda}] = 0 \tag{40}$$

and satisfy the relations

$$\hat{T}_{\mu\nu}^{\lambda\dagger} = (-1)^{\nu-\mu} \hat{T}_{-\mu-\nu}^{\lambda}, \tag{41}$$

and

$$\sum_{\mu} \hat{T}_{\mu\nu}^{\lambda\dagger} \hat{T}_{\mu\nu'}^{\lambda} = \delta_{\nu\nu'}, \quad \sum_{\nu} \hat{T}_{\mu\nu}^{\lambda\dagger} \hat{T}_{\mu'\nu}^{\lambda} = \delta_{\mu\mu'}. \tag{42}$$

In addition to the commutation relations among these tensor operators we need the ones with the SO(3) generators in Eq. (29) and the relations (30),

$$[\hat{L}_z, \hat{T}_{\mu\nu}^{\lambda}] = \nu \hat{T}_{\mu\nu}^{\lambda}, \quad [\hat{J}_z, \hat{T}_{\mu\nu}^{\lambda}] = \mu \hat{T}_{\mu\nu}^{\lambda}, \tag{43}$$

as well as

$$
\begin{aligned}
&[\hat{L}_{\pm},\hat{T}^{\lambda}_{\mu\nu}]=[\lambda(\lambda+1)-\nu(\nu\mp1)]^{1/2}\hat{T}^{\lambda}_{\mu\nu\mp\lambda},\\
&[\hat{J}_{\pm},\hat{T}^{\lambda}_{\mu\nu}]=[\lambda(\lambda+1)-\mu(\mu\pm1)]^{1/2}\hat{T}^{\lambda}_{\mu\pm\lambda\nu},
\end{aligned}
\tag{44}
$$

where

$$
\hat{L}_{\pm}=\hat{L}_x\mp i\hat{L}_y,\quad \hat{J}_{\pm}=\hat{J}_x\pm i\hat{J}_y. \tag{45}
$$

The rotation matrix elements $D^{\lambda}_{\mu\nu}(\alpha,\beta,\gamma)$, $(\lambda=0,\frac{1}{2},1,\frac{3}{2},\ldots;\ \mu,\nu=0,\pm\frac{1}{2},\ldots,\pm\lambda)$ are a realization of the $\hat{T}^{\lambda}_{\mu\nu}$ operators and define their action on the states $|IMK\rangle$. Here it suffices to know that

$$
\hat{T}^{1}_{-1-1}|I\ -I\ -I\rangle=\left(\frac{2I+1}{2I+3}\right)^{1/2}|I+1\ -I-1\ -I-1\rangle. \tag{46}
$$

C. Construction of coherent states

The straightforward application of Perelemov's prescription[9] would make the set of rotational coherent states be

$$
X(x,\psi)Z(z,\omega)Y(y^{1}_{\mu\nu})|000\rangle, \tag{47}
$$

where the $X(x,\psi)$ and $Z(z,\omega)$ operators are two parametrizations of the SO(3) group

$$
\begin{aligned}
&X(x,\psi)=N_X e^{xJ_+}e^{-x^*J_-}e^{-i\psi J_z},\\
&Z(z,\omega)=N_Z e^{zL_+}e^{-z^*L_-}e^{-i\omega L_z},
\end{aligned}
\tag{48}
$$

with the real parameters $0\leqslant\psi\leqslant2\pi$, $0\leqslant\omega\leqslant2\pi$, and the complex parameters x $(-\infty\leqslant\operatorname{Re}x\leqslant\infty$, $-\infty\leqslant\operatorname{Im}x\leqslant\infty)$ and z $(-\infty\leqslant\operatorname{Re}z\leqslant\infty$, $-\infty\leqslant\operatorname{Im}z\leqslant\infty)$, respectively. N_X and N_Z denote normalization constants. The unitary operator $Y(y^{1}_{\mu\nu})$ is a general element of the Abelian group R^9 generated by the tensor operators $T^{1}_{\mu\nu}$, i.e.,

$$
Y(y^{1}_{\mu\nu})=\exp\left(\sum_{\mu,\nu}y^{1}_{\mu\upsilon}T^{1}_{\mu\nu}\right) \tag{49}
$$

with the complex parameters $y^{1}_{\mu\nu}$ $(-\infty\leqslant\operatorname{Re}y^{1}_{\mu\nu}\leqslant\infty$, $-\infty\leqslant\operatorname{Im}y^{1}_{\mu\nu}\leqslant\infty)$ satisfying $y^{1}_{\mu\nu}=(-1)^{\mu-\upsilon+1}y^{1*}_{-\mu-\nu}$. This mode of construction combines two spin CS with CS belonging to the abelian group R^9. This produces a set of CS of some complexity, which will not be further analyzed.

Instead, in analogy with the Janssen's approach,[16] we propose the simpler construction

$$
|xyz\rangle=\exp[-\tfrac{1}{2}yy^*(1+xx^*)^2(1+zz^*)^2]e^{xJ_+}e^{zL_+}\exp[yf(\hat{I})T^{1}_{-1-1}]|000\rangle, \tag{50}
$$

where the parameter y relates to the above discussion such that $y=y^{1}_{-1-1}$. The function $f(\hat{I})$ is

$$
f(\hat{I})=\sqrt{(2\hat{I}^2+\hat{I})/(2\hat{I}-1)}, \tag{51}
$$

where the operator $\hat{I}$ is defined by (compare Ref. 16)

$$
\hat{I}|IMK\rangle=I|IMK\rangle. \tag{52}
$$

The effect of the $f(\hat{I})$ function is to generate a desired Poisson distribution by canceling some factors occurring in Eq. (46). The operator $\hat{I}$ can be expressed in terms of the Schwinger boson operators, but in the present context it can be seen as a purely formal construct that serves to simplify some expressions. Note the subtle differences in the normalization factor and in the right exponential operator in comparison to those in Ref. 16. It follows straightforwardly that

$$\exp(yf(\hat{I})\hat{T}^{1}_{-1-1})|000\rangle=\sum_{I=0,1,\ldots}^{\infty}\frac{(yf(\hat{I})\hat{T}^{1}_{-1-1})^{I}}{I!}|000\rangle=\sum_{I=0,1,\ldots}^{\infty}\frac{y^{I}}{\sqrt{I!}}|I\ -I\ -I\rangle \tag{53}$$

and that

$$\begin{aligned} e^{z\hat{L}_+}|I\ -I\ -I\rangle &=\sum_{n=0}^{\infty}\frac{z^n}{n!}\hat{L}_+^n|I\ -I\ -I\rangle \\ &=\sum_{n=0}^{2I}\frac{z^n}{n!}\{2I(2I-1)\cdots[2I-(n-1)]\}^{1/2}(n!)^{1/2}|I\ -I\ -I+n\rangle \\ &=\sum_{K=-I}^{I}\frac{z^{(I+K)}}{[(I+K)!]^{1/2}}\left[\frac{(2I)!}{(I-K)!}\right]^{1/2}|I-IK\rangle, \end{aligned} \tag{54}$$

where $n=I+K$ has been used from the second to the third line. By changing M to K, and $\hat{L}_+$ to $\hat{J}_+$, the analogous expansion of $e^{x\hat{J}_+}|I-I-I\rangle$ is obtained. These results make possible the following expression for the CS:

$$\begin{aligned} |xyz\rangle &=\exp[-\tfrac{1}{2}yy^*(1+xx^*)^2(1+zz^*)^2] \\ &\quad\times\sum_{IMK}\left\{\frac{[(2I)!]^2}{(I+M)!(I-M)!(I+K)!(I-K)!}\right\}^{1/2}\frac{x^{(I+M)}y^{I}z^{(I+K)}}{(I!)^{1/2}}|IMK\rangle. \end{aligned} \tag{55}$$

Each member in this set of CS is normalized to unity since

$$\begin{aligned} \langle xyz|xyz\rangle &=\exp[-yy^*(1+xx^*)^2(1+zz^*)^2] \\ &\quad\times\sum_{IMK}\frac{[(2I)!]^2}{(I+M)!(I-M)!(I+K)!(I-K)!}\frac{(xx^*)^{(I+M)}(yy^*)^{I}(zz^*)^{(I+K)}}{I!} \\ &=\exp[-yy^*(1+xx^*)^2(1+zz^*)^2]\sum_{I=0}^{\infty}\frac{(yy^*)^{I}}{I!} \\ &\quad\times\sum_{M=-I}^{I}\frac{(2I)!(xx^*)^{(I+M)}}{(I+M)!(I-M)!}\sum_{K=-I}^{I}\frac{(2I)!(zz^*)^{(I+K)}}{(I+K)!(I-K)!} \\ &=\exp[-yy^*(1+xx^*)^2(1+zz^*)^2]\sum_{I=0}^{\infty}\frac{[(yy^*)(1+xx^*)^2(1+zz^*)^2]^{I}}{I!}=1, \end{aligned} \tag{56}$$

where the power expansions, such as

$$(1+zz^*)^{2I}=\sum_{K=-I}^{I}\frac{(2I)!}{(I+K)!(I-K)!}(zz^*)^{I+K} \tag{57}$$

have been used. It is noteworthy that we now have a Poisson distribution in the variable $yy^*(1+xx^*)^2(1+zz^*)^2$. In addition, the overlap $\langle x'y'z'|xyz\rangle$ between two different members of the set of CS is

$$\langle x'y'z'|xyz\rangle = \exp[-\tfrac{1}{2}yy^*(1+xx^*)^2(1+zz^*)^2]\exp[-\tfrac{1}{2}y'y'^*(1+x'x'^*)^2(1+z'z'^*)^2] \times \exp[yy'^*(1+xx'^*)^2(1+zz'^*)^2], \tag{58}$$

which can be obtained in an analogous manner showing the general nonorthogonality of these CS.

Because of their construction, the $\{|xyz\rangle\}$ states satisfy the condition Eq. (1) for CS. In order to verify whether these CS satisfy the stronger or the weaker formulation of the second condition, we need to construct a proper measure $d\mu$. Using the measure of Ref. 16 as a guide we obtain

$$d\mu_{\pm}(x,y,z) = \frac{1}{\pi^3}\{4[(1+xx^*)(1+zz^*)]^4(yy^*)^2 - 8[(1+xx^*)(1+zz^*)]^2yy^* + 1\}dx\, dy\, dz, \tag{59}$$

where

$$dx = d\,\mathrm{Re}\, x\, d\,\mathrm{Im}\, x, \quad dy = d\,\mathrm{Re}\, y\, d\,\mathrm{Im}\, y, \quad dz = d\,\mathrm{Re}\, z\, d\,\mathrm{Im}\, z \tag{60}$$

so that

$$\int d\mu_{\pm}(x,y,z)|xyz\rangle\langle xyz| = \sum_{IMK} |IMK\rangle\langle IMK| = 1. \tag{61}$$

Then, it follows immediately [see Eq. (55)] that

$$|IMK\rangle = \int d\mu_{\pm}(x,y,z)\exp\left[-\frac{1}{2}yy^*(1+xx^*)^2(1+zz^*)^2\right]x^{*I+M}y^{*I}z^{*I+K} \times \left\{\frac{[(2I)!]^2}{I!(I+M)!(I-M)!(I+K)!(I-K)!}\right\}^{1/2}|xyz\rangle. \tag{62}$$

Note that both the CS from Ref. 16 and the present ones satisfy the weaker version of the second condition for CS, because the measure of neither is positive.

D. Coherent state operator averages

In order to develop the dynamics related to the rotational CS certain operator averages need to be determined. Evaluation of the necessary integrals involves using some properties of the binomial power expansion and the Poisson distribution. The final results are as follows:

$$\langle \hat{I}\rangle = yy^*(1+xx^*)^2(1+zz^*)^2 = \zeta, \tag{63}$$

$$\langle \hat{L}_x\rangle = \frac{z+z^*}{1+zz^*}\zeta, \quad \langle \hat{L}_y\rangle = \frac{i(z^*-z)}{1+zz^*}\zeta, \quad \langle \hat{L}_z\rangle = \frac{(zz^*-1)}{1+zz^*}\zeta, \tag{64}$$

$$\langle \hat{L}_i^2\rangle = \begin{cases} \dfrac{\zeta}{2} + \langle \hat{L}_i\rangle^2\left[1+\dfrac{1}{2\zeta}\right] & \text{if } \zeta>0, \\ 0 & \text{if } \zeta=0, \end{cases} \tag{65}$$

$$\langle \hat{L}_i\hat{L}_j + \hat{L}_j\hat{L}_i\rangle = \begin{cases} 2\langle \hat{L}_i\rangle\langle \hat{L}_j\rangle\left[1+\dfrac{1}{2\zeta}\right] & \text{if } \zeta>0, \\ 0 & \text{if } \zeta=0, \end{cases} \tag{66}$$

where $i \neq j$, and the notation $\langle \cdots \rangle = \langle xyz | \cdots | xyz \rangle$ is used. By changing $\hat{L}_i$ to $\hat{J}_i$, and z to x^* in the above expressions, the averages of the components of $\hat{J}$ are obtained. The integral $\langle xyz | \hat{I} | xyz \rangle$ turns out to be slightly different from that of Ref. 16. However, the functionality of the first-order averages with respect to that basic integral remains essentially the same and real differences appear in the second-order averages.

Uncertainty relationships for the CS can be derived by combining the well-known relationship[10]

$$(\Delta \hat{L}_i)^2 (\Delta \hat{L}_j)^2 \geq \tfrac{1}{4} |\langle \hat{L}_k \rangle|^2 \quad (i \neq j \neq k), \tag{67}$$

where

$$(\Delta \hat{L}_i)^2 = \langle \hat{L}_i^2 \rangle - \langle \hat{L}_i \rangle^2, \tag{68}$$

with Eqs. (64) and (65) to obtain

$$(\Delta L_i)^2 (\Delta L_j)^2 = \begin{cases} \dfrac{1}{4} \left[1 + \dfrac{\langle \hat{L}_i \rangle^2}{\zeta^2} \right] \left[1 + \dfrac{\langle \hat{L}_j \rangle^2}{\zeta^2} \right] \zeta^2 & \text{if } \zeta < 0 \\ 0 & \text{if } \zeta = 0. \end{cases} \tag{69}$$

Note that in the special case of $\langle \hat{L}_i \rangle = \langle \hat{L}_j \rangle = 0$, $\zeta > 0$ the uncertainty relationship is minimized for that pair of components, i.e., $(\Delta \hat{L}_i)^2 (\Delta \hat{L}_j)^2 = \frac{1}{4} |\langle \hat{L}_k \rangle|^2$, $(i \neq j \neq k)$.

E. Reparametrization

For the purpose of physical interpretation a new parametrization of the CS is introduced. The parameters related to the spin CS, i.e., x and z, are changed by adopting the stereographic projection onto a plane[18]

$$x = e^{-i\alpha} \cot \frac{\beta}{2}, \quad z = e^{+i\gamma} \cot \frac{\delta}{2}, \tag{70}$$

where $0 \leq \alpha, \gamma \leq 2\pi$, and $0 \leq \beta, \delta \leq \pi$. The remaining parameter y is expressed as

$$y = \zeta^{1/2} \sin^2 \frac{\beta}{2} \sin^2 \frac{\delta}{2} e^{i(\alpha - \gamma - \epsilon)}, \tag{71}$$

where $0 \leq \zeta \leq \infty$ and $0 \leq \epsilon \leq 2\pi$. In terms of these new parameters one finds

$$\begin{gathered} xx^* = \cot^2 \frac{\beta}{2}, \quad zz^* = \cot^2 \frac{\delta}{2}, \quad 1 + xx^* = \operatorname{cosec}^2 \frac{\beta}{2}, \\ 1 + zz^* = \operatorname{cosec}^2 \frac{\delta}{2}, \quad yy^* = \zeta \sin^4 \frac{\beta}{2} \sin^4 \frac{\delta}{2}, \quad yy^*(1 + xx^*)^2 (1 + zz^*)^2 = \zeta, \end{gathered} \tag{72}$$

and

$$
\begin{aligned}
|xyz\rangle &= |\alpha\beta\gamma\delta\epsilon\zeta\rangle \\
&= \exp\left(-\frac{\zeta}{2}\right)\sum_{IMK}\left\{\frac{[(2I)!]^2}{(I+K)!(I-K)!(I+M)!(I-M)!}\right\}^{1/2}\left[e^{-iM\alpha}\cos^{I+M}\left(\frac{\beta}{2}\right)\sin^{I-M}\left(\frac{\beta}{2}\right)\right] \\
&\quad\times\left[e^{+iK\gamma}\cos^{I+M}\left(\frac{\delta}{2}\right)\sin^{I-K}\left(\frac{\delta}{2}\right)e^{-iI\epsilon}\right]\frac{\zeta^{1/2}}{\sqrt{I!}}|IMK\rangle \\
&= \exp\left(-\frac{\zeta}{2}\right)\sum_{IMK} D^I_{MI}(\alpha,\beta,0)D^I_{KI}(-\gamma,\delta,\epsilon)\frac{\zeta^{1/2}}{\sqrt{I!}}|IMK\rangle,
\end{aligned}
\tag{73}
$$

where the definition of the rotational matrices D^I_{MK}[18,20] has been used.

We write

$$
\langle\alpha\beta\gamma\delta\epsilon\zeta|\hat{I}|\alpha\beta\gamma\delta\epsilon\zeta\rangle \equiv \langle\hat{I}\rangle = \zeta \tag{74}
$$

and similarly

$$
\begin{aligned}
\langle\hat{L}_x\rangle &= \zeta\cos\gamma\sin\delta, \quad \langle\hat{J}_x\rangle = \zeta\cos\alpha\sin\beta, \\
\langle\hat{L}_y\rangle &= \zeta\sin\gamma\sin\delta, \quad \langle\hat{J}_y\rangle = \zeta\sin\alpha\sin\beta, \\
\langle\hat{L}_z\rangle &= \zeta\cos\delta, \quad \langle\hat{J}_z\rangle = \zeta\cos\beta,
\end{aligned}
\tag{75}
$$

and

$$
\langle\hat{L}^2\rangle = \langle\hat{J}^2\rangle = \zeta(\zeta+2). \tag{76}
$$

From these expressions, it follows that the parameter ζ is the angular momentum modulus, the pairs of angles γ,δ, and α,β are the azimuthal and the polar angle of the $\langle\hat{\mathbf{L}}\rangle$ and $\langle\hat{\mathbf{J}}\rangle$ vectors in the body-fixed and the space-fixed frame, respectively. The angle ϵ is associated with the relative orientation of the body-fixed and the space-fixed frames. Finally, the probability $P_{IMK}(\zeta,\beta,\delta)$ to find the rotational state $|IMK\rangle$ in the CS is

$$
\begin{aligned}
P_{IMK}(\zeta,\beta,\delta) &= \left\{\frac{[(2I)!]^2}{(I+K)!(I-K)!(I+M)!(I-M)!}\right\}\left[\cos^{2(I+M)}\frac{\beta}{2}\sin^{2(I-M)}\frac{\beta}{2}\right] \\
&\quad\times\left[\cos^{2(I+K)}\frac{\delta}{2}\sin^{2(I-K)}\frac{\delta}{2}\right]\exp(-\zeta)\frac{\zeta^I}{I!} \\
&= \left\{\frac{[(2I)!]^2}{(I+K)!(I-K)!(I+M)!(I-M)!}\right\}p^{(I+M)}(1-p)^{(I-M)} \\
&\quad\times q^{(I+K)}(1-q)^{(I-M)}\exp(-\zeta)\frac{\zeta^I}{I!},
\end{aligned}
\tag{77}
$$

where $p=\cos^2\delta$ and $q=\cos^2\beta$. Note that $P_{IMK}(\zeta,\beta,\delta)$ combines binomial distributions in p and q, and a Poisson distribution in ζ. The result

$$
\sum_{I=0,1,\ldots,}^{\infty}\sum_{M=-I}^{I}\sum_{K=-I}^{I}P_{IMK}(\zeta,\beta,\delta)=1 \tag{78}
$$

is readily obtained. Alternatively, one can write

$$P_{IMK}(\zeta,\beta,\delta)=|D^{I}_{MI}(\alpha,\beta,0)|^2|D^{I}_{KI}(-\gamma,\delta,\epsilon)|^2\frac{\zeta^I}{I!}\exp(-\zeta)$$

$$=[d^{I}_{MI}(\beta)]^2[d^{I}_{KI}(\delta)]^2\frac{\zeta^I}{I!}\exp(-\zeta). \tag{79}$$

This probability exhibits more detail than is commonly needed and when used in the calculation of suitable *S*-matrix elements averaging procedures will wash out the excessive details.

F. Coherent state dynamics

Applying Ehrenfest's theorem to the operators $\hat{L}_i$ for $i=x,y,z$, i.e.,

$$\frac{d}{dt}\langle\hat{L}_i\rangle=-i\langle[\hat{H}_{\rm rot},\hat{L}_i]\rangle=\left\langle\sum_j \varepsilon_{ijk}\frac{1}{2A_k}(\hat{L}_j\hat{L}_k+\hat{L}_k\hat{L}_j)\right\rangle, \tag{80}$$

for $\zeta>0$ results in [see Eq. (66)]

$$\begin{aligned}\frac{d}{dt}\langle\hat{L}_x\rangle&=\langle\hat{L}_y\rangle\langle\hat{L}_z\rangle\left(1+\frac{1}{2\zeta}\right)\left(\frac{1}{A_z}-\frac{1}{A_y}\right),\\ \frac{d}{dt}\langle\hat{L}_y\rangle&=\langle\hat{L}_z\rangle\langle\hat{L}_x\rangle\left(1+\frac{1}{2\zeta}\right)\left(\frac{1}{A_z}-\frac{1}{A_x}\right),\\ \frac{d}{dt}\langle\hat{L}_z\rangle&=\langle\hat{L}_x\rangle\langle\hat{L}_y\rangle\left(1+\frac{1}{2\zeta}\right)\left(\frac{1}{A_x}-\frac{1}{A_y}\right),\end{aligned} \tag{81}$$

and for $\zeta=0$,

$$\frac{d}{dt}\langle\hat{L}_i\rangle=0,\quad i=x,y,z. \tag{82}$$

The quasiclassical rotation vector or angular velocity $\boldsymbol{\omega}$ is then defined with components

$$\omega_i=\begin{cases}\dfrac{1}{A_i}\dfrac{\zeta+\frac{1}{2}}{\zeta}\langle\hat{L}_i\rangle, & \zeta>0\\ 0, \quad \zeta=0.\end{cases} \tag{83}$$

The quasiclassical nature of this rotation vector is evident from

$$\begin{aligned}\dot{\omega}_x&=\frac{\omega_y\omega_z}{A_x}(A_y-A_z),\\ \dot{\omega}_y&=\frac{\omega_x\omega_z}{A_y}(A_x-A_z),\\ \dot{\omega}_z&=\frac{\omega_x\omega_y}{A_z}(A_x-A_y),\end{aligned} \tag{84}$$

which are the classical Euler equations for the motion of a rigid body without torque.[21] It follows that the rotation vector $\vec{\omega}$ behaves exactly as that of a classical rigid body with the same moment of inertia when the CS are propagated by the Schrödinger equation with the Hamiltonian of Eq. (28). This is analogous to the classical motion Eq. (18) of the harmonic oscillator CS Eq. (6).

However, the definition of the rotation vector $\boldsymbol{\omega}$ in Eq. (83) differs from the definition

$$\omega_i = \frac{1}{A_i} \langle \hat{L}_i \rangle \tag{85}$$

by Janssen,[16] which corresponds exactly to the classical definition[21]

$$\omega_i = \frac{1}{A_i} L_i . \tag{86}$$

This means that for the Janssen all-spin rotational CS the expectation values of the angular momenta $\hat{L}_i$ are quasiclassical variables, whereas they are not for our integer-only rotational CS. Inspection of Eq. (83) reveals that the variables

$$\Lambda_i = \begin{cases} \dfrac{\zeta + \frac{1}{2}}{\zeta} \langle \hat{L}_i \rangle, & \zeta > 0 \\ 0, & \zeta = 0, \end{cases} \tag{87}$$

are quasiclassical. This difference requires some explanation, which, as will be shown, is to be found in the Heisenberg uncertainty principle.

The proportionality factor $(\zeta + \frac{1}{2})/\zeta$ between the quasiclassical variables and the expectation values of $\hat{L}_i$ is a constant of the motion related to the total angular momentum. The limit of total angular momentum tending to zero is equivalent to $\zeta \to 0$ [Eq. (76)]. This means that for the Janssen all-spin CS

$$\omega^2 = \omega_x^2 + \omega_y^2 + \omega_z^2 \leqslant \frac{1}{A_x} (\langle \hat{L}_x \rangle^2 + \langle \hat{L}_y \rangle^2 + \langle \hat{L}_z \rangle^2) \leqslant \frac{\zeta^2}{A_x}, \tag{88}$$

where we have used the convention $A_z \geqslant A_y \geqslant A_x$. From this it follows that $\omega^2 \to 0$ when the angular momentum is decreased to zero and the CS are "spinning down." In contrast, the same calculation for the integer-only CS shows that

$$\frac{1}{A_x} \left(\zeta + \frac{1}{2} \right)^2 \geqslant \omega^2 \geqslant \frac{1}{A_z} \left(\zeta + \frac{1}{2} \right)^2 \tag{89}$$

and, thus, as the angular momentum decreases toward zero these CS keep "spinning" with an angular momentum of $\frac{1}{2}$ in units of $\hbar$, i.e., the zero angular momentum state cannot be reached by continuously decreasing the value of the angular momentum.

This result is related to the uncertainty principle. The orientation of the rigid body in terms of the Euler angles with decreasing angular momentum cannot be determined precisely, neither can the rotation vector being the time derivative of the orientation. As a result the integer-only CS, which are the appropriate ones to describe rotating bodies, yield a finite lower bound to the value of the angular velocity as the angular momentum decreases. Thus, since the angular velocity cannot decrease below a certain value, one cannot assume that the orientation can change by an arbitrarily small amount in a given time interval. It is noteworthy that apart from the lower bound Eq. (89) in the rotation vector length the motion is completely classical. The lower bound tends to zero for increasing moments of inertia indicating that this effect is not present for macroscopic bodies.

The remaining question is why the Janssen all-spin CS do not exhibit this finite lower bound on the angular velocity vector. The answer lies in the fact that half-integer spin systems have angular momenta that do not correspond to any position coordinate. As a result they always possess an uncertainty of $\frac{1}{2}\hbar$ in the orientation part of the total angular momentum even when its value tends to zero. For such CS the attempt to precisely define the angular momentum orientation does not introduce additional constraints allowing the angular velocity $\boldsymbol{\omega}$ to decrease to zero when the total angular momentum does.

V. APPLICATIONS

The rotational CS can with advantage be applied to spectroscopic analysis of molecular processes. For instance, the harmonic oscillator CS have been employed to successfully predict the distribution of reaction products over vibrational quantum states.[22] The role of the CS in this analysis was to allow a detailed *a posteriori* quantum level analysis of calculations on a reactive molecular system that only employed essentially classical trajectories. This procedure results in great savings of computer time and thus increases the range of studies to quite complex chemical reactions.

Because rotations and vibrations are intimately coupled in molecular systems, such analysis is most successful when both kinds of motion are treated on an equal footing. This is accomplished by combining the canonical CS and the integer-only rotational CS developed here. The theory can be applied without difficulties to general molecular fragments and it is straightforward to study a general asymmetric rotor. However, in order to avoid inessential technicalities we describe in this section how the theory applies to a diatomic molecule, which can be thought of as a linear rotor.

A. Linear rotor CS

By selecting the body-fixed z axis of a diatomic molecule $(A_x = A_y,\ A_z = 0)$ as that of the molecular bond, the component of the angular momentum vector $\hat{\mathbf{L}}$ in that direction vanishes at all times $(K=0)$. In that case, the rotational states $|IMK\rangle$ become the eigenstates of the linear rotor, i.e., the spherical harmonics $Y_{IM}(\theta,\varphi)$, see Eq. (39). Furthermore, inserting $\omega_z = 0$ into Eq. (84) makes the other two components of $\boldsymbol{\omega}$ in the body-fixed frame constant as well. This implies that the z parameter does not vary in time, its specific value being dependent on the orientation of the x and the y axes in the body-fixed frame. The CS of Eq. (55) then become

$$\langle \theta\varphi | xy \rangle_{\text{diat}} = \exp[-\tfrac{1}{2} yy^*(1+xx^*)^2] \sum_{IM} \left\{ \frac{(2I)!}{(I+M)!(I-M)!} \right\}^{1/2} \frac{x^{(I+M)} y^I}{\sqrt{I!}} Y_{IM}(\theta,\varphi), \quad (90)$$

where the superfluous terms in z^{I+K} have been omitted. Alternatively, these CS can be reparametrized so that

$$\begin{aligned} \langle \theta\varphi | xy \rangle_{\text{diat}} &= \langle \theta\varphi | \alpha\beta\zeta \rangle_{\text{diat}} \\ &= \exp\left(-\frac{\zeta}{2}\right) \sum_{IM} \left\{ \frac{(2I)!}{(I+M)!(I-M)!} \right\}^{1/2} \left[e^{-iM\alpha} \cos^{I+M}\frac{\beta}{2} \sin^{I-M}\frac{\beta}{2} \right] \frac{\zeta^{I/2}}{\sqrt{I!}} Y_{IM}(\theta,\varphi) \\ &= \exp\left(-\frac{\zeta}{2}\right) \sum_{IM} D^I_{MI}(\alpha,\beta,0) \frac{\zeta^{I/2}}{\sqrt{I!}} Y_{IM}(\theta,\varphi). \end{aligned} \quad (91)$$

When the CS are prepared with the vector $\langle \alpha\beta\zeta | \hat{\mathbf{J}} | \alpha\beta\zeta \rangle$ normal to the xy plane (i.e., with $\langle \alpha\beta\zeta | J_x | \alpha\beta\zeta \rangle = \langle \alpha\beta\zeta | J_y | \alpha\beta\zeta \rangle = 0$) it holds that the initial value of β is 0, or π. Consider the case $\beta = 0$. Then,

$$\langle \theta\varphi | 00\zeta \rangle_{\text{diat}} = \exp\left(-\frac{\zeta}{2}\right) \sum_I \frac{\zeta^{I/2}}{\sqrt{I!}} Y_{II}(\theta,\varphi), \quad (92)$$

where the angle α has been arbitrarily set to zero. The CS time evolution of $|0\beta(t)\zeta\rangle_{\text{diat}}$ from that initial condition, is

$$\begin{aligned}\langle\theta\varphi|0\beta(t)\zeta\rangle_{\text{diat}}&=\langle\theta\varphi|\exp(-i\hat{H}t)|00\zeta\rangle_{\text{diat}}\\&=\langle\theta\varphi|\exp\left(-\frac{\zeta}{2}\right)\sum_I\frac{\zeta^{I/2}}{\sqrt{I!}}\exp\left[-i\frac{I(I+1)}{2A}t\right]Y_{II}(\theta,\varphi),\\&=\exp\left(-\frac{\zeta}{2}\right)\sum_I\frac{\zeta^{I/2}}{\sqrt{I!}}Y_{II}(\theta,\varphi-\omega_I t),\end{aligned}\tag{93}$$

where $\hat{H}$ is the diatomic Hamiltonian, A the moment of inertia, and

$$\omega_I=\frac{(I+1)}{2A}.\tag{94}$$

In the last line of Eq. (93), the explicit form[23]

$$Y_{I\pm I}(\theta,\varphi)=(-1)^I\frac{1}{2^I I!}\left(\frac{(2I+1)!}{4\pi}\right)^{1/2}\sin^I\theta\exp(iI\varphi)\tag{95}$$

has been used. From Eqs. (93) and (95), it is easy to see that the rotational CS peak symmetrically around $\theta=\pi/2$ (i.e., the maximum lies in the xy plane) throughout the evolution. The shape of the rotational CS with respect to the angle φ is more difficult to describe analytically. For large ζ values, the superposition of the spherical harmonics given by Eq. (93) is sharply peaked around $I=I_{\max}\sim\zeta$. When this holds, the sum over I can be approximated by an integration. The evaluation of that integral by the stationary phase method[24] reveals that the CS peak around the value $\phi=\phi_{\max}$ where

$$\phi_{\max}(t)\sim\frac{(I_{\max}+\frac{1}{2})}{A}t\sim\frac{(\zeta+\frac{1}{2})}{A}t\sim\omega_z t.\tag{96}$$

This implies that when the total angular momentum is high the peak's center moves with the constant angular velocity ω_z along the equator of the θ,φ sphere. The general properties of the diatomic rotational CS can be easily derived form this example prepared with $\beta=0$. If the rotation $\hat{R}(\alpha,\beta,0)$ is applied to $|00\zeta\rangle_{\text{diat}}$ then the general diatomic rotational CS, Eq. (91), are recovered

$$\begin{aligned}\langle\theta\varphi|\alpha\beta\zeta\rangle_{\text{diat}}&=\langle\theta\varphi|\hat{R}(\alpha,\beta,0)|00\zeta\rangle_{\text{diat}}\\&=\exp\left(-\frac{\zeta}{2}\right)\sum_{IM}D^I_{MI}(\alpha,\beta,0)\frac{\zeta^{I/2}}{\sqrt{I!}}Y_{IM}(\theta,\varphi)\\&=\exp\left(-\frac{\zeta}{2}\right)\sum_I\frac{\zeta^{I/2}}{\sqrt{I!}}Y_{II}(\theta',\varphi'),\end{aligned}\tag{97}$$

where the properties of the spherical harmonics have been applied and where the angles θ' and φ' are given with respect to the rotated space-fixed frame. The rotation transforms the vector $\langle\alpha\beta\zeta|\hat{\mathbf{J}}|\alpha\beta\zeta\rangle$ from the z direction of the space-fixed frame to the (α,β) direction, in accordance with Eq. (75). The final expression in Eq. (97) is formally identical to the nonrotated CS in Eq. (92). The properties of $|\alpha\beta\zeta\rangle_{\text{diat}}$ are the same as those of $|00\zeta\rangle_{\text{diat}}$ but now referred to a plane normal to the vector $\langle\alpha\beta\zeta|\hat{\mathbf{J}}|\alpha\beta\zeta\rangle$.

No closed form expression is known for the rotational CS Eq. (55) or Eq. (90). However, in Sec. V B a wave function for a diatomic molecule is constructed that clarifies the CS and makes their structure more explicit.

B. The diatom END wave function

For purposes of interpretation in terms of CS we rewrite the END wave function in the center-of-mass (c.m.) frame. The actual END propagation is always done in the space-fixed laboratory frame, but the analysis of reagents (at the initial time) and products (at the final time) is better done in the c.m. frame of each fragment. The END wave function for each fragment in the narrow wave packet limit for the nuclei and with a Thouless determinant[5] for the electrons takes the form

$$\Psi_{\mathrm{END}}(\mathbf{X},\mathbf{x},t)=F_{\mathrm{nucl}}[\mathbf{X};\mathbf{R}(t),\mathbf{P}(t)]f_{\mathrm{el}}[\mathbf{x};\mathbf{z}(t),\mathbf{R}(t)]\exp[i\gamma_{\mathrm{END}}(t)], \tag{98}$$

when there no longer exist any overlaps or exchange terms between fragments. For the case of a diatomic fragment the nuclear part consists of two generalized frozen Gaussian wave packets,

$$F_{\mathrm{nucl}}(\mathbf{X};\mathbf{R},\mathbf{P})=\frac{1}{N_{\mathrm{nucl}}}\prod_{k=1}^{2}\exp\left\{-a_k[\mathbf{X_k}-\mathbf{R_k}(t)]^2+\frac{i}{\hbar}\mathbf{P_k}(t)\cdot[\mathbf{X_k}-\mathbf{R_k}(t)]\right\}. \tag{99}$$

The electronic part $f_{\mathrm{el}}[\mathbf{z}(t),\mathbf{R}(t)]$ is a Thouless single determinant wave function. The total phase $\gamma_{\mathrm{END}}(t)$ is the quantum mechanical action

$$\gamma_{\mathrm{END}}(t)=\int_0^t ds\, L[\mathbf{R}(s),\mathbf{P}(s),\mathbf{z}(s)], \tag{100}$$

in terms of the END quantum Lagrangian $L(\mathbf{R},\mathbf{P},\mathbf{z})$.

The Gaussian wave packets have finite width, $1/a_k$, explicitly defined as those of canonical CS [compare Eqs. (22) and (23)]. The SC limit of zero width then yields the simplest END approximation.

The transformation to the c.m. coordinates is using Jacobi coordinates, i.e.,

$$\mathbf{X_0}=\frac{1}{M}(m_1\mathbf{X_1}+m_2\mathbf{X_2}),\quad \mathbf{x}=\mathbf{X_2}-\mathbf{X_1}, \tag{101}$$

with a similar transformation of the average nuclear positions

$$\begin{aligned}\mathbf{R_0}(t)&=\frac{1}{M}[m_1\mathbf{R_1}(t)+m_2\mathbf{R_2}(t)],\\ \mathbf{r}(t)&=\mathbf{R_2}(t)-\mathbf{R_1}(t),\end{aligned} \tag{102}$$

and with $M=m_1+m_2$ the total mass. This transformation results in the product of two independent Gaussian wave packets:

$$F_{\mathrm{nucl}}(\mathbf{X};\mathbf{R},\mathbf{P})=F_0(\mathbf{X_0};\mathbf{R_0},\mathbf{P_0})F_{\mathrm{int}}(\mathbf{x};\mathbf{r},\mathbf{p}) \tag{103}$$

only if the condition: $\omega_1=\omega_2=\omega$ is imposed. Then it follows that

$$F_0(\mathbf{X_0};\mathbf{R_0},\mathbf{P_0})=\frac{1}{N_0}\exp\{-a_0[\mathbf{X_0}-\mathbf{R_0}(t)]^2+i\mathbf{P_0}(t)\cdot[\mathbf{X_0}-\mathbf{R_0}(t)]\}, \tag{104}$$

and

$$F_{\text{int}}(\mathbf{x};\mathbf{r},\mathbf{p})=\frac{1}{N_{\text{int}}}\exp\{-a_{\mu}[\mathbf{x}-\mathbf{r}(t)]^2+i\mathbf{p}(t)\cdot[\mathbf{x}-\mathbf{r}(t)]\}. \tag{105}$$

Here, the new momentum parameters are

$$\mathbf{P_0}(t)=\mathbf{P_1}(t)+\mathbf{P_2}(t) \tag{106}$$

and

$$\mathbf{p}(t)=\mu(\dot{\mathbf{R}}_2-\dot{\mathbf{R}}_1)=\mu\dot{\mathbf{r}}, \tag{107}$$

respectively, where μ is the reduced mass. The transformed exponents turn out to be

$$a_0=\frac{M\omega}{2} \tag{108}$$

and

$$a_{\mu}=\frac{\mu\omega}{2}, \tag{109}$$

in units of $1/\hbar$. The electronic part of the wave function, being a function of relative nuclear positions only, is not be affected by this transformation. A similar partition of the total phase means that

$$\gamma_{\text{END}}(t)=\gamma_0(t)+\gamma_{\text{int}}(t), \tag{110}$$

where $\gamma_0(t)$ contains only the nuclear c.m. variables and $\gamma_{\text{int}}(t)$ the internal nuclear variables along with the electronic parameters.

The c.m. part is a Gaussian with a trivial time evolution. This part together with its phase $\gamma_0(t)$ can be totally separated from the rest leaving the internal END wave function

$$\Psi_{\text{int}}^{\text{END}}(t)=F_{\text{int}}[\mathbf{r}(t),\mathbf{p}(t)]f_{\text{el}}[\mathbf{z}(t),\mathbf{r}(t)]\exp[i\gamma_{\text{int}}(t)]. \tag{111}$$

The internal wave function is now split into a vibrational and a rotational part. Because of the coupling of rotations and vibrations in molecules this separation must be approximate. Define $\mathbf{x}=x\mathbf{n}$ and $\mathbf{r}=r\mathbf{m}$, where $\mathbf{n}$ and $\mathbf{m}$ are unit vectors. For a vibrating molecule rotating with rotation vector $\boldsymbol{\omega}$ we write the momentum

$$\mathbf{p}=\mu\dot{\mathbf{r}}=p_v\mathbf{m}+\mu b\boldsymbol{\omega}\times\mathbf{m}, \tag{112}$$

with p_v the vibrational part of the momentum parallel to the axis of the molecule and b the equilibrium bond length. The END evolution of the molecule is described through the parameters $\mathbf{r}(t)$, $\mathbf{p}(t)$, and $\mathbf{z}(t)$ only. When the coupling of rotation and vibration is neglected the rotation vector is constant yielding the angular momentum

$$\mathbf{L}=A\boldsymbol{\omega}=\mu b^2\boldsymbol{\omega}. \tag{113}$$

The exponent of $F_{\text{int}}(\mathbf{x};\mathbf{r},\mathbf{p})$ in Eq. (105) becomes

$$\begin{aligned}-a_{\mu}[\mathbf{x}-\mathbf{r}]^2+i\mathbf{p}\cdot[\mathbf{x}-\mathbf{r}]=&-a_{\mu}(x-r)^2-2a_{\mu}r(x-r)\mathbf{n}\cdot(\mathbf{n}-\mathbf{m})\\&-a_{\mu}r^2(\mathbf{n}-\mathbf{m})^2+i(x-r)p_v\mathbf{m}\cdot\mathbf{n}+i\mu b(x-r)(\boldsymbol{\omega}\times\mathbf{m})\cdot\mathbf{n}\\&+irp_v\mathbf{m}(\mathbf{n}-\mathbf{m})+i\mu br(\boldsymbol{\omega}\times\mathbf{m})\cdot(\mathbf{n}-\mathbf{m}).\end{aligned} \tag{114}$$

Introducing the angles η between $\mathbf{n}$ and $\mathbf{m}$ and ξ between $\boldsymbol{\omega}\times\mathbf{m}$ and $\mathbf{n}$ yields

$$-a_\mu(x-r)^2-2a_\mu r(x-r)(1-\cos\eta)-2a_\mu r^2(1-\cos\eta),$$
$$i(x-r)p_v\cos\eta+i\mu b\omega(x-r)\cos\xi-irp_v(1-\cos\eta)+i\mu b\omega r\cos\xi. \tag{115}$$

Since **x** and **r** are close the angle η is small and so is $\pi/2-\xi$. Using these facts in an order analysis considering $x-r$ as small, omitting terms of third order or higher in the real part, and of second order or higher in the imaginary part, and leaving out the coupling term result in

$$-a_\mu(x-r)^2-2a_\mu r^2(1-\cos\eta),$$
$$+i(x-r)p_v+i\mu\omega br\cos\xi. \tag{116}$$

Replacing r in the angular part by the equilibrium bond distance b, using Eq. (109) for a_μ, and expressing η and ξ in terms of the polar coordinates θ,φ of **x** in the laboratory frame using obvious right spherical triangles produces a vibrational and a rotational factor

$$F_{\text{int}}(x,\theta,\varphi;r,0,0)\approx F_{\text{vib}}(x;r)F_{\text{rot}}(\theta,\varphi;0,0), \tag{117}$$

with

$$F_{\text{vib}}(x;r)=\frac{1}{N_{\text{vib}}}\exp\left[-\frac{\mu\omega}{2}(x-r)^2+ip_v(x-r)\right] \tag{118}$$

and

$$F_{\text{rot}}(\theta,\varphi)=\frac{1}{N_{\text{rot}}}\exp[-\mu\omega b^2(1-\sin\theta\cos\varphi)+i\mu\omega b^2\sin\theta\sin\varphi]$$
$$=\frac{1}{N_{\text{rot}}}\exp[-\mu\omega b^2(1-\sin\theta e^{i\varphi})]. \tag{119}$$

The general case is obtained by rotating $\boldsymbol{\omega}$ to the orientation α,β in the laboratory frame.

Identification of the CS parameter ζ can be accomplished by evaluation of the expectation values of $\hat{\mathbf{J}}$ and its square with respect to the END wave function Eq. (99),

$$F_{\text{nucl}}=F_0F_{\text{int}}=F_0F_{\text{vib}}F_{\text{rot}}. \tag{120}$$

One finds easily that

$$\langle\hat{J}_x\rangle=\mu\omega b^2\sin\beta\cos\alpha,$$
$$\langle\hat{J}_y\rangle=\mu\omega b^2\sin\beta\sin\alpha, \tag{121}$$
$$\langle\hat{J}_z\rangle=\mu\omega b^2\cos\beta,$$

and that

$$\langle\hat{J}_x^2\rangle=\langle\hat{J}_x\rangle^2+A_{xx}\omega,\quad \langle\hat{J}_y^2\rangle=\langle\hat{J}_y\rangle^2+A_{yy}\omega,\quad \langle\hat{J}_z^2\rangle=\langle\hat{J}_z\rangle^2+A_{zz}\omega. \tag{122}$$

Using the result

$$\text{Tr}(A)=2\mu b^2, \tag{123}$$

it follows that

$$\langle \hat{\mathbf{J}}^2 \rangle = \langle \hat{\mathbf{J}} \rangle^2 + 2\mu b^2 \omega = \mu \omega b^2 (\mu \omega b^2 + 2) = \zeta(\zeta + 2), \tag{124}$$

such that the CS parameter ζ may be identified with the total nuclear angular momentum $\mu \omega b^2$ of the diatomic molecule in units of $\hbar$.

The expansion of the rotational factor of the END wave function Eq. (119)

$$F_{\text{rot}}(\theta, \varphi; 0,0) = \frac{1}{N_{\text{rot}}} \exp(-\zeta) \sum_I \frac{(\zeta \sin \theta e^{i\varphi})^I}{I!} \tag{125}$$

should be compared to the CS Eq. (93) using Eq. (95)

$$\langle \theta\varphi | 00\zeta \rangle_{\text{diat}} = \exp(-\zeta/2) \sum_I \left(\frac{(2I+1)!!}{4\pi(2I)!!} \right)^{1/2} \frac{(-\zeta^{1/2} \sin \theta e^{i\varphi})^I}{\sqrt{I!}}. \tag{126}$$

The two series have the same structure for the arguments θ and φ, but consecutive terms in the CS series decrease less rapidly as functions of I.

Even though the two functions are not identical the rotational CS are useful for final state analysis. That the vibrational CS match the vibrational part of the END wave function is, of course, not surprising as the END wave function is built from Gaussians with the appropriate widths. The CS, because of their nature, retain shape during evolution. Thus, the choice of fixed a Gaussian shape for the END wave function ensures that also the shape of the rotational factor persists during evolution. Our analysis has shown that the average orientation and the width match those of the rotational CS. Furthermore, the END wave function by the TDVP construction is limited to the dynamics of average values, which are the internal vibrational coordinates and orientation of the body fixed frame. It then makes sense to replace the factors of the END wave function by their CS counterparts with the matching parameters, including the time evolution generated by the END wave function. This procedure is the recommended one to use for *a posteriori* quantum state analysis even when the END evolution has been done for the case of zero width Gaussians, i.e., classical nuclei. Thus, since it is now established that the END nuclear wave function for low excitations can to a good approximation be represented as a product of vibrational and rotational CS, the expressions in Eqs. (6) and (79) can be used to compute probabilities for vibrational and rotational eigenstates, and, thus, provide *a posteriori* quantum vibrational and rotational resolution of cross sections obtained from classical trajectory calculations.

VI. CONCLUSION

The canonical CS have countless applications. First derived by Schrödinger[2] and later analyzed by many[1] these CS display remarkable properties. One of the most useful properties in the context of molecular processes is the quasiclassical evolution. Janssen[16] constructed all-spin rotational CS that evolve quasiclassically. Involving both integer and half-integer spins these CS have not seen much application to physical systems. Our work establishes that integer-only rotational CS can be constructed that exhibit quasiclassical dynamics. A notable property of the integer-only rotational CS is the nonzero minimum angular velocity attained as the angular momentum decreases to zero. The Janssen construction does not follow Perelomov's prescription of constructing group-related CS. The metric of these CS for the resolution of the identity is nondefinite. The facts that the construction of Atkins and Dobson[17] employs Perelomov's prescription for the case of the symmetric rotor CS, that it coincides with the Janssen construction restricted to symmetric rotors, and that these have a positive definite measure, make it plausible that there exists CS from a suitable Perelomov construction with a positive measure for both the all-spin CS and the integer-only CS.

It is known that the canonical CS remain quasiclassical in the presence of a time-dependent linear external field. We have not investigated whether this holds true also for the rotational CS when subjected to some external torque.

In this paper we have concentrated on extracting probabilities for rotational quantum levels of molecules from nuclear trajectory calculations of molecular processes. This application is valid both for dynamics involving predetermined potential energy surfaces as well as for direct nonadiabatic approaches such as END.

ACKNOWLEDGMENTS

The support of this research by of the Office of Naval Research is gratefully acknowledged. The authors thank Professor John Klauder for useful conversations about coherent states.

[1] J. R. Klauder and B.-S. Skagerstam, *Coherent States, Applications in Physics and Mathematical Physics* (World Scientific, Singapore, 1985).
[2] E. Schrödinger, Naturwissenschaften **14**, 664 (1926).
[3] R. J. Glauber, Phys. Rev. Lett. **10**, 84 (1963).
[4] E. Deumens, A. Diz, R. Longo, and Y. Öhrn, Rev. Mod. Phys. **66**, 917 (1994).
[5] D. J. Thouless, Nucl. Phys. **21**, 225 (1960).
[6] D. J. Thouless, *The Quantum Mechanics of Many-Body System* (Academic, New York, 1961).
[7] M. Rasetti, J. Theor. Phys. **13**, 425 (1973).
[8] A. I. Solomon, J. Math. Phys. **12**, 390 (1971).
[9] A. M. Perelomov, *Generalized Coherent States and Their Applications* (Springer, New York, 1986).
[10] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics* (Wiley, New York, 1977), Vol. 1.
[11] A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[12] E. Deumens and Y. Öhrn, J. Chem. Soc., Faraday Trans. **93**, 919 (1997).
[13] Y. Öhrn *et al.*, in *Time-Dependent Quantum Molecular Dynamics*, edited by J. Broeckhove and L. Lathouwers (Plenum, New York, 1992), pp. 279–292.
[14] E. Deumens, A. Diz, H. Taylor, and Y. Öhrn, J. Chem. Phys. **96**, 6820 (1992).
[15] J. A. Morales, A. C. Diz, E. Deumens, and Y. Öhrn, J. Chem. Phys. **103**, 9968 (1995).
[16] D. Janssen, Sov. J. Nucl. Phys. **25**, 479 (1977).
[17] P. W. Atkins and J. C. Dobson, Proc. R. Soc. London, Ser. A **321**, 321 (1971).
[18] J. D. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics*, Encyclopedia of Mathematics and its Applications, Vol. 8, 1st ed. (Addison–Wesley, Reading, MA, 1981).
[19] L. Fonda, N. Mankoc, and M. Rosina, Phys. Rep. **158**, 160 (1988).
[20] R. Zare, *Angular Momentum*, 1st ed. (Wiley, New York, 1988).
[21] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison–Wesley, Reading, MA, 1980).
[22] J. A. Morales, A. C. Diz, E. Deumens, and Y. Öhrn, Chem. Phys. Lett. **233**, 392 (1995).
[23] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, 4th ed. (USGPO, Washington, DC, 1964).
[24] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, 1st ed. (Dover, New York, 1986), reprint of the 1975 edition.

Gauge theories with graded differential Lie algebras

Raimar Wulkenhaar[a)]
Institut für Theoretische Physik, Universität Leipzig, Augustusplatz 10/11, D-04109 Leipzig, Germany

(Received 17 November 1997; accepted for publication 30 September 1998)

We present a mathematical framework of gauge theories that is based upon a skew-adjoint Lie algebra and a generalized Dirac operator, both acting on a Hilbert space. © *1999 American Institute of Physics.* [S0022-2488(99)02201-X]

I. INTRODUCTION

In this article the author's previous article,[1] in which we proposed a mathematical calculus towards gauge field theories based upon graded differential Lie algebras is made more precise. Given a skew-adjoint Lie algebra $\mathfrak{g}$, a representation π of $\mathfrak{g}$ on a Hilbert space h_0 as well as an (unbounded) operator D and a grading operator Γ on h_0, we developed a scheme providing connection and curvature forms to build physical actions. The general part of our exposition was on a very formal level; we worked with unbounded operators (even the splitting of a bounded in two unbounded operators) without specification of the domain.

In the present article, we correct this shortcoming. The idea is to introduce a second Hilbert space h_1, which is the domain of the unbounded operator D. Now, D is a linear continuous operator from h_1 to h_0, and the just mentioned splitting involves continuous operators only. Moreover, the awkward connection theory in the previous paper is resolved in a strict algebraic description in terms of normalizers of graded Lie algebras. Finally, our construction of the universal graded differential Lie algebra is considerably simplified (thanks to a hint by Schmüdgen).

The scope of our framework is the construction of Yang–Mills–Higgs models in noncommutative geometry.[2] The standard procedure[3,4] starts from spectral triples with real structure[5,6] and is limited to the standard model.[7] The hope is[8] that the replacement of the unital associative $*$-algebra in the prior Connes–Lott prescription[9] by a skew-adjoint Lie algebra admits representations general enough to construct grand unified theories. For a realization of this strategy see Refs. 10–12. We discuss the relation to the axiomatic formulation[6] of noncommutative geometry in the last section.

II. THE ALGEBRAIC SETTING

Let $\mathfrak{g}$ be a skew-adjoint Lie algebra, $a^*=-a$ for all $a\in\mathfrak{g}$. Let h_0,h_1 be Hilbert spaces, where h_1 is dense in h_0. Denoting by $\mathcal{B}(h_0)$ and $\mathcal{B}(h_1)$ the algebras of linear continuous operators on h_0 and h_1, respectively, we define $\mathcal{B}:=\mathcal{B}(h_0)\cap\mathcal{B}(h_1)$. The vector space of linear continuous mappings from h_1 to h_0 is denoted by $\mathcal{L}$. Let π be a representation of $\mathfrak{g}$ in $\mathcal{B}$. Let $D\in\mathcal{L}$ be a generalized Dirac operator with respect to $\pi(\mathfrak{g})$. This means that D has an extension to a self-adjoint operator on h_0, that $[D,\pi(a)]\in\mathcal{L}$ even belongs to $\mathcal{B}$ for any $a\in\mathfrak{g}$ and that the resolvent of D is compact. Finally, let $\Gamma\in\mathcal{B}$ be a grading operator, i.e., Γ^2 is the identity on both h_0 and h_1, $[\Gamma,\pi(a)]=0$ on both h_0,h_1 for any $a\in\mathfrak{g}$ and $D\Gamma+\Gamma D=0$ on h_1 extends to 0 on h_0. This setting was called *L-cycle* in Ref. 1, referring to a *Lie*-algebraic version of a *K-cycle*, the former name[2,9] for spectral triple.[5,6]

The standard example of this setting $(\mathfrak{g},h_0,h_1,D,\pi,\Gamma)$ is

[a)]Present address: Centre de Physique Théorique, CNRS-Luminy, Case 907, F-13288 Marseille, France. Electronic mail: raimar@cpt.univ-mrs.fr or raimar.wulkenhaar@itp.uni-leipzig.de

$$\mathfrak{g} = C^\infty(X) \otimes \mathfrak{a}, \quad h_0 = L^2(\mathcal{S}) \otimes \mathbb{C}^F,$$
$$h_1 = W_1^2(\mathcal{S}) \otimes \mathbb{C}^F, \quad D = \mathrm{i}\partial\!\!\!/ \otimes 1_F + \boldsymbol{\gamma} \otimes \mathcal{M},$$
$$\pi = \mathrm{id} \otimes \hat{\pi}, \quad \Gamma = \boldsymbol{\gamma} \otimes \hat{\Gamma}.$$

Here, $C^\infty(X)$ denotes the algebra of real-valued smooth functions on a compact Riemannian spin manifold X, $\mathfrak{a}$ is a skew-adjoint matrix Lie algebra, $L^2(\mathcal{S})$ denotes the Hilbert space of square integrable sections of the spinor bundle $\mathcal{S}$ over X, $W_1^2(\mathcal{S})$ denotes the Sobolev space of square integrable sections of $\mathcal{S}$ with a generalized first derivative, $\mathrm{i}\partial\!\!\!/$ is the Dirac operator of the spin connection, $\boldsymbol{\gamma}$ is the grading operator on $L^2(\mathcal{S})$ anti-commuting with $\mathrm{i}\partial\!\!\!/$, $\hat{\pi}$ is a representation of $\mathfrak{a}$ in $M_F\mathbb{C}$ and $\hat{\Gamma}$ a grading operator on $M_F\mathbb{C}$ commuting with $\hat{\pi}(\mathfrak{a})$ and anti-commuting with $\mathcal{M} \in M_F\mathbb{C}$.

III. THE UNIVERSAL GRADED DIFFERENTIAL LIE ALGEBRA Ω

For $\mathfrak{g}$ being a real Lie algebra we consider the real vector space $\mathfrak{g}^2 = \mathfrak{g} \times \mathfrak{g}$, with the linear operations given by $\lambda_1(a_1,a_2) + \lambda_2(a_3,a_4) = (\lambda_1 a_1 + \lambda_2 a_3, \lambda_1 a_2 + \lambda_2 a_4)$, for $a_i \in \mathfrak{g}$ and $\lambda_i \in \mathbb{R}$. Let T be the tensor algebra of $\mathfrak{g}^2$, equipped with the $\mathbb{N}$–grading structure $\deg((a,0)) = 0$ and $\deg((0,a)) = 1$, and linear extension to higher degrees, $\deg(t_1 \otimes t_2) = \deg(t_1) + \deg(t_2)$, for $t_i \in T$. Defining $T^n = \{t \in T : \deg(t) = n\}$, we have $T = \oplus_{n\in\mathbb{N}} T^n$ and $T^k \otimes T^l \subset T^{k+l}$. We regard T as a graded Lie algebra with graded commutator given by $[t^k, t^l] := t^k \otimes t^l - (-1)^{kl} t^l \otimes t^k$, for $t^n \in T^n$.

Let $\tilde{\Omega} = \oplus_{n\in\mathbb{N}} \tilde{\Omega}^n = \Sigma\, [\mathfrak{g}^2,[\cdots[\mathfrak{g}^2,\mathfrak{g}^2]\cdots]]$ be the $\mathbb{N}$-graded Lie subalgebra of T (due to the graded Jacobi identity) given by the set of sums of repeated commutators of elements of $\mathfrak{g}^2$. Let I' be the vector subspace of $\tilde{\Omega}$ of sums of elements of the following type:

$$[(a,0),(b,0)] - ([a,b],0), \quad [(a,0),(0,b)] + [(0,a),(b,0)] - (0,[a,b]), \tag{2}$$

for $a,b \in \mathfrak{g}$. The first part extends the Lie algebra structure of $\mathfrak{g}$ to the first component of $\mathfrak{g}^2$ and the second part plays the rôle of a Leibniz rule, see below. Obviously, $I := I' + [\mathfrak{g}^2, I'] + [\mathfrak{g}^2,[\mathfrak{g}^2,I']] + \cdots$ is an $\mathbb{N}$-graded ideal of $\tilde{\Omega}$, so that $\Omega := \oplus_{n\in\mathbb{N}} \Omega^n$, $\Omega^n := \tilde{\Omega}^n/(I \cap \Omega^n)$ is an $\mathbb{N}$-graded Lie algebra.

On T we define recursively a graded differential as an $\mathbb{R}$-linear map $d: T^n \to T^{n+1}$ by

$$d(a,0) = (0,a), \quad d(0,a) = 0, \tag{3}$$

$$d((a,0)\otimes t) = d(a,0) \otimes t + (a,0) \otimes dt, \quad d((0,a)\otimes t) = -(0,a) \otimes dt,$$

for $a \in \mathfrak{g}$ and $t \in T$. One easily verifies $d^2 = 0$ on T and the graded Leibniz rule $d(t^k \otimes t^l) = dt^k \otimes t^l + (-1)^k t^k \otimes dt^l$, for $t^n \in T^n$. Thus, d defined by (3) is a graded differential of the tensor algebra T and of the graded Lie algebra T as well, $d[t^k,t^l] = [dt^k,t^l] + (-1)^k [t^k, dt^l]$.

Due to $d\mathfrak{g}^2 \subset \mathfrak{g}^2$ we conclude that d is also a graded differential of the graded Lie subalgebra $\tilde{\Omega} \subset T$. Moreover, one easily checks $dI' \subset I'$, giving $dI \subset I$. Therefore, $(\Omega, [,], d)$ is a graded differential Lie algebra, with the graded differential d given by $d(\varpi + I) := d\varpi + I$, for $\varpi \in \tilde{\Omega}$.

We extend the involution $*: a \mapsto -a$ on $\mathfrak{g}$ to an involution of T by $(a,0)^* = -(a,0)$, $(0,a)^* = -(0,a)$ and $(t_1 \otimes t_2)^* = t_2^* \otimes t_1^*$, giving $[t^k,t^l]^* = -(-1)^{kl}[t^{k*}, t^{l*}]$. Clearly, this involution extends to Ω. The identity $a = -a^*$ yields $\omega^{k*} = -(-1)^{k(k-1)/2}\omega^k$, for any $\omega^k \in \Omega^k$.

The graded differential Lie algebra Ω is universal in the following sense.

Proposition 1: Let $\Lambda = \oplus_{n\in\mathbb{N}} \Lambda^n$ *be an* $\mathbb{N}$*-graded Lie algebra with graded differential* d *such that*

(i) $\Lambda^0 = \pi(\mathfrak{g})$ *for a surjective homomorphism* π *of Lie algebras,*

(ii) Λ *is generated by* $\pi(\mathfrak{g})$ *and* $\mathrm{d}\pi(\mathfrak{g})$ *as the set of repeated commutators. Then there exists a differential ideal* $I_\Lambda \subset \Omega$ *fulfilling* $\Lambda \cong \Omega/I_\Lambda$.

Proof: Define a linear surjective mapping $p: \Omega \to \Lambda$ by

$$p((a,0))=\pi(a), \quad p(d\omega)=\mathrm{d}(p(\omega)), \quad p([\omega,\tilde{\omega}]):=[p(\omega),p(\tilde{\omega})],$$

for $a\in\mathfrak{g}$ and $\omega,\tilde{\omega}\in\Omega$. Because of $d\ker p\subset\ker p$, $I_\Lambda=\ker p$ is the desired differential ideal of Ω. □

IV. THE GRADED DIFFERENTIAL LIE ALGEBRA Ω_D

Using the grading operator Γ we define on $\mathcal{L}$ and $\mathcal{B}$ a $\mathbb{Z}_2$-grading structure, the even subspaces carry the subscript 0 and the odd subspaces the subscript 1. Then, the graded commutator $[\,.\,,\,.\,]_g:\mathcal{L}_i\times\mathcal{B}_j\to\mathcal{L}_{(i+j)\bmod 2}$ is defined by

$$[A_i,B_j]_g:=A_i\circ B_j-(-1)^{ij}B_j\circ A_i\equiv-(-1)^{ij}[B_j,A_i]_g, \tag{4}$$

where $B_j\in\mathcal{B}_j$ and $A_i\in\mathcal{L}_i$. If $A_i\in\mathcal{B}_i$ then $[\,.\,,\,.\,]_g$ maps h_1 to h_1 and h_0 to h_0.

Using the elements π and D of our setting we define a linear mapping $\pi:\Omega\to\mathcal{B}$ by

$$\begin{gathered}\pi((a,0)):=\pi(a), \quad \pi((0,a)):=[-\mathrm{i}D,\pi(a)]_g,\\ \pi([\omega^k,\omega^l]):=[\pi(\omega^k),\pi(\omega^l)]_g,\end{gathered} \tag{5}$$

for $a\in\mathfrak{g}$ and $\omega^n\in\Omega^n$. The self-adjointness of D on h_0 implies that π is involutive, $\pi(\omega^*)=(\pi(\omega))^*$.

Note that $\pi(\Omega)$ is not a graded *differential* Lie algebra. The standard procedure to construct such an object is to define $\mathcal{J}=\ker\pi+d\ker\pi\subset\Omega$. It is easy to show that $\mathcal{J}$ is a graded differential ideal of Ω, providing the graded differential Lie algebra,

$$\Omega_D=\bigoplus_{n\in\mathbb{N}}\Omega_D^n, \quad \Omega_D^n:=\Omega^n/\mathcal{J}^n\cong\pi(\Omega^n)/\pi(\mathcal{J}^n), \tag{6}$$

where $\mathcal{J}^n=\mathcal{J}\cap\Omega^n$. One has $\Omega_D^0\cong\pi(\Omega^0)\equiv\pi(\mathfrak{g})$ and $\Omega_D^1\cong\pi(\Omega^1)$. By construction, the differential d on Ω_D is given by $d(\pi(\omega^n)+\pi(\mathcal{J}^n)):=\pi(d\omega^n)+\pi(\mathcal{J}^{n+1})$, for $\omega^n\in\Omega^n$.

It is very useful to consider an extension of the second formula of (5), $\pi(d(a,0)):=[-\mathrm{i}D,\pi((a,0))]_g$, *to higher degrees:*

$$\pi(d\omega^n)=[-\mathrm{i}D,\pi(\omega^n)]_g+\sigma(\omega^n), \quad \omega^n\in\Omega^n. \tag{7}$$

It turns out[1] that $\sigma:\Omega\to\mathcal{L}$ is a linear mapping recursively given by

$$\begin{gathered}\sigma((a,0))=0, \quad \sigma((0,a))=[D,[D,\pi(a)]_g]_g,\\ \sigma([\omega^k,\omega^l])=[\sigma(\omega^k),\pi(\omega^l)]_g+(-1)^k[\pi(\omega^k),\sigma(\omega^l)]_g.\end{gathered} \tag{8}$$

Equation (7) has an important consequence: Putting $\omega^n\in\ker\pi$ we get

$$\pi(\mathcal{J}^{n+1})=\{\sigma(\omega^n):\omega^n\in\Omega^n\cap\ker\pi\}. \tag{9}$$

The point is that $\sigma(\Omega)$ can be computed from the last equation (8) once $\sigma(\Omega^1)$ is known. Then one can compute $\pi(\mathcal{J})$ and obtains with (7) an explicit formula for the differential on Ω_D.

V. CONNECTIONS

We define the graded normalizer $N_{\mathcal{L}}(\pi(\Omega))$ of $\pi(\Omega)$ in $\mathcal{L}$, its vector subspace $\mathcal{H}$ compatible with $\pi(\mathcal{J})$ and the graded centralizer $\mathcal{C}$ of $\pi(\Omega)$ in $\mathcal{L}$ by

$$
\begin{aligned}
N^k_{\mathcal{L}}(\pi(\Omega)) = \{ \eta^k \in \mathcal{L}_{k \bmod 2} :\ \eta^k \ \text{has} \left\{ {\text{self-adjoint} \atop \text{skew-adjoint}} \right\} \text{extension for } \tfrac{k(k-1)}{2} \left\{ {\text{odd} \atop \text{even}} \right\}, \\
[\eta^k, \pi(\omega^n)]_g \in \pi(\Omega^{k+n}) \forall \omega^n \in \Omega^n, \forall n \in \mathbb{N} \}, \\
\mathcal{H}^k = \{ \eta^k \in N^k_{\mathcal{L}}(\pi(\Omega)) : [\eta^k, \pi(j^n)]_g \in \pi(\mathcal{J}^{k+n}) \forall j^n \in \mathcal{J}^n \}, \\
\mathcal{C}^k = \{ c^k \in N^k_{\mathcal{L}}(\pi(\Omega)) : [c^k, \pi(\omega)]_g = 0 \forall \omega \in \Omega \}.
\end{aligned} \tag{10}
$$

Here, the linear continuous operator $[\eta^k, \pi(\omega^n)]_g : h_1 \to h_0$ must have its image even in the subspace $h_1 \subset h_0$ and must have an extension to a linear continuous operator on h_0. For each degree n we have the following system of inclusions:

$$
\begin{array}{ccccccc}
\mathcal{L} & \supset & \mathcal{H}^n & \supset & \pi(\Omega^n) & \supset & \pi(\mathcal{J}^n) \\
 & & \cup & & \cap & & \\
 & & \mathcal{C}^n & & \mathcal{B} & \subset & \mathcal{L}
\end{array} . \tag{11}
$$

The graded Jacobi identity and Leibniz rule define the structure of a graded differential Lie algebra on $\hat{\mathcal{H}} = \oplus_{n \in \mathbb{N}} \hat{\mathcal{H}}^n$, with $\hat{\mathcal{H}}^n = \mathcal{H}^n / (\mathcal{C}^n + \pi(\mathcal{J}^n))$:

$$
\begin{aligned}
& [[\eta^k + \mathcal{C}^k + \pi(\mathcal{J}^k),\ \eta^l + \mathcal{C}^l + \pi(\mathcal{J}^l)]_g,\ \pi(\omega^n) + \pi(\mathcal{J}^n)]_g \\
& \quad := [\eta^k, [\eta^l, \pi(\omega^n)]_g]_g - (-1)^{kl} [\eta^l, [\eta^k, \pi(\omega^n)]_g]_g + \pi(\mathcal{J}^{k+l+n}), \\
& [d(\eta^k + \mathcal{C}^k + \pi(\mathcal{J}^k)),\ \pi(\omega^n) + \pi(\mathcal{J}^n)]_g \\
& \quad := \pi \circ d \circ \pi^{-1}([\eta^k, \pi(\omega^n)]_g) - (-1)^k [\eta^k, \pi(d\omega^n)]_g + \pi(\mathcal{J}^{k+n+1}),
\end{aligned} \tag{12}
$$

for $\eta^n \in \mathcal{H}^n$ and $\omega^n \in \Omega^n$.

The lesson is that $\pi(\Omega)$ and its ideal $\pi(\mathcal{J})$ give rise not only to the graded differential Lie algebra Ω_D but also to $\hat{\mathcal{H}}$, both being natural. It turns out that it is the differential Lie algebra $\hat{\mathcal{H}}$ which occurs in our connection theory

Definition 2: Within our setting, a connection ∇ *and its associated covariant derivative* $\mathcal{D}$ *are defined by*
(i) $\mathcal{D} \in \mathcal{L}_1$ *with a self-adjoint extension,*
(ii) $\nabla : \Omega^n_D \to \Omega^{n+1}_D$ *is linear,*
(iii) $\nabla(\pi(\omega^n) + \pi(\mathcal{J}^n)) = [-\mathrm{i}\mathcal{D}, \pi(\omega^n)]_g + \sigma(\omega^n) + \pi(\mathcal{J}^{n+1}),\quad \omega^n \in \Omega^n.$
The operator $\nabla^2 : \Omega^n_D \to \Omega^{n+2}_D$ *is called the curvature of the connection.*

This definition states that the covariant derivative $\mathcal{D}$ generalizes the operator D of the setting and the connection ∇ generalizes the differential d. In particular, both $\mathcal{D}$ and ∇ are related via the same equation (iii) as D and d are according to (7).

Proposition 3: Any connection/covariant derivative has the form $\nabla = d + [\rho + \mathcal{C}^1, \cdot]_g$ *and* $\mathcal{D} = D + \mathrm{i}\rho$, *for* $\rho \in \mathcal{H}^1$. *The curvature is* $\nabla^2 = [\theta, \cdot]$, *with* $\theta = d\hat{\rho} + \frac{1}{2}[\hat{\rho}, \hat{\rho}]_g \in \hat{\mathcal{H}}^2$, *where* $\hat{\rho} = \rho + \mathcal{C}^1 \in \hat{\mathcal{H}}^1$.

Proof: There is a canonical pair of connection/covariant derivative given by $\nabla = d$ and $\mathcal{D} = D$. If $(\nabla^{(1)}, \mathcal{D}^{(1)})$ and $(\nabla^{(2)}, \mathcal{D}^{(2)})$ are two pairs of connections/covariant derivatives, we get from (iii),

$$
(\nabla^{(1)} - \nabla^{(2)})(\pi(\omega^n) + \pi(\mathcal{J}^n)) = [\nabla^{(1)}_h - \nabla^{(2)}_h, \pi(\omega^n)]_g + \pi(\mathcal{J}^{n+1}).
$$

This means that $\rho := \nabla^{(1)}_h - \nabla^{(2)}_h \in \mathcal{H}^1$ is a concrete representative and $\nabla^{(1)} - \nabla^{(2)} = [\hat{\rho}, \cdot]_g$, where $\hat{\rho} = \rho + \mathcal{C}^1 \in \hat{\mathcal{H}}^1$. The formula for θ is a direct consequence of (12). □

VI. GAUGE TRANSFORMATIONS

The exponential mapping defines a unitary group

$$\mathcal{U}:=\left\{\prod_{\alpha=1}^{N}\exp(v_\alpha):\ \exp(v_\alpha):=1_{\mathcal{B}}+\sum_{k=1}^{\infty}\frac{1}{k!}(v_\alpha)^k,\ \ v_\alpha\in\mathcal{H}^0\cap\mathcal{B},dv_\alpha-[-\mathrm{i}D,v_\alpha]\in\mathcal{C}^1\right\}. \qquad (13)$$

Due to $\exp(v)A\exp(-v)=A+\Sigma_{k=1}^{\infty}\frac{1}{k!}([v,[v,\ldots,[v,A]\cdots]])_k$, where $()_k$ contains k commutators of $A\in\mathcal{L}$ with v, we have a natural degree-preserving representation Ad of $\mathcal{U}$ on $\mathcal{H}$, $\mathrm{Ad}_u(\eta^n)=u\,\eta^n u^*\in\mathcal{H}^n$, for $\eta^n\in\mathcal{H}^n$ and $u\in\mathcal{U}$.

Definition 4: In our setting, the gauge group is the group $\mathcal{U}$ defined in (13). Gauge transformations of the connection and the covariant derivative are given by

$$\nabla\mapsto\nabla':=\mathrm{Ad}_u\nabla\mathrm{Ad}_{u^*},\qquad \mathcal{D}\mapsto\mathcal{D}':=u\,\mathcal{D}u^*,\qquad u\in\mathcal{U}.$$

Note that the consistency relation (iii) in Definition 2 reduces on the infinitesimal level to the condition $dv_\alpha-[-\mathrm{i}D,v_\alpha]\in\mathcal{C}^1$ in (13). The gauge transformation of the curvature form reads as $\theta\mapsto\theta'=\mathrm{Ad}_u\theta$.

VII. PHYSICAL ACTION

We borrow the integration calculus introduced by Connes to noncommutative geometry[2,5] and summarize the main results. Let E_n be the eigenvalues of the compact operator (compactness was assumed in the setting) $|D|^{-1}=(DD^*)^{-1/2}$ on h_0, arranged in decreasing order. Here, the finite dimensional kernel of D is not relevant so that $E_1<\infty$. The K-cycle (h_0,D) over the C^*-algebra $\mathcal{B}(h_0)$ is called d^+-summable if $\Sigma_{n=1}^{N}E_n=O(\Sigma_{n=1}^{N}n^{-1/\mathrm{d}})$. Equivalently, the partial sum of the first N eigenvalues of $|D|^{-\mathrm{d}}$ has (at most) a logarithmic divergence as $N\to\infty$ so that $|D|^{-\mathrm{d}}$ belongs to the (two-sided) Dixmier trace class ideal $\mathcal{L}^{(1,\infty)}(h_0)$. Therefore, $f|D|^{-\mathrm{d}}\in\mathcal{L}^{(1,\infty)}(h_0)$ for any $f\in\mathcal{B}(h_0)$, and the Dixmier trace provides for $f>0$ a linear functional $f\mapsto\mathrm{Tr}_\omega(f|D|^{-\mathrm{d}})=\mathrm{Lim}_\omega(1/\ln N)\Sigma_{n=1}^{N}\mu_n\in\mathbb{R}^+$. Here, μ_n are the eigenvalues of $f|D|^{-\mathrm{d}}$ and the limit Lim_ω involves an appropriate limiting procedure ω. The Dixmier trace fulfills $\mathrm{Tr}_\omega(f|D|^{-\mathrm{d}})=\mathrm{Tr}_\omega(ufu^*|D|^{-\mathrm{d}})$, for unitary $u\in\mathcal{B}(h_0)$.

Let $\theta_0^*:h_0\to h_1$ be the uniquely determined adjoint of a representative $\theta_0:h_1\to h_0$ of the curvature form $\theta\in\hat{\mathcal{H}}^2$. It follows that $\theta_0\theta_0^*\in\mathcal{B}(h_0)$ so that we propose the following definition for the physical action.

Definition 5: The bosonic action S_B and the fermionic action S_F of the connection ∇ and covariant derivative $\mathcal{D}$ are given by

$$S_B(\nabla):=\min_{\mathrm{j}^2\in\mathcal{C}^2+\pi(\mathcal{J}^2)}\mathrm{Tr}_\omega((\theta_0+\mathrm{j}^2)(\theta_0+\mathrm{j}^2)^*|D|^{-\mathrm{d}}),$$
$$S_F(\psi,\mathcal{D}):=\langle\psi,\mathcal{D}\psi\rangle_{h_0},\qquad \psi\in h_1, \qquad (14)$$

where $\langle,\rangle_{h_0}$ is the scalar product on h_0.

The bosonic action S_B is independent of the choice of the representative θ_0. Thus, we can take the canonical dependence of the gauge potential ρ,

$$\theta_0=\{-\mathrm{i}D,\rho\}+\tfrac{1}{2}\{\rho,\rho\}+\sigma\circ\pi^{-1}(\rho),$$

where $\sigma\circ\pi^{-1}$ is supposed to be extended from $\pi(\Omega^1)$ to $\mathcal{H}^1$. It is unique up to elements of $\mathcal{C}^2+\pi(\mathcal{J}^2)$. Since the Dixmier trace is positive, the element $\mathrm{j}_0^2\in\mathcal{C}^2+\pi(\mathcal{J}^2)$ at which the minimum in (14) is attained is the unique solution of the equation

$$\mathrm{Tr}_\omega((\theta_0+j_0^2)(j^2)^*|D|^{-\mathrm{d}})=0,\qquad \forall\, j^2\in\mathcal{C}^2+\pi(\mathcal{J}^2).$$

It is clear that the action (14) is invariant under gauge transformations,

$$\nabla \mapsto \mathrm{Ad}_u \nabla \mathrm{Ad}_{u*}, \quad \mathcal{D} \to u\mathcal{D}u^*, \quad \psi \mapsto u\psi, \quad u \in \mathcal{U}. \tag{15}$$

Note that our gauge group as defined in (13) is always connected, which means that we have no access to ''big'' gauge transformations. Note further that there exist Lie groups having the same Lie algebra. In that case there will exist fermion multiplets ψ which can be regarded as multiplets of different Lie groups. For the bosonic sector only the Lie algebra is important, so one can have the pathological situation of a model with identical particle contents and identical interactions, but different gauge groups. We consider such gauge theories as identical.

VIII. REMARKS ON THE STANDARD EXAMPLE

Recall (5) that the general form of an element $\tau^1 \in \pi(\Omega^1)$ is

$$\tau^1 = \sum_{\alpha, z \geq 0} [\pi(a_\alpha^z), [\cdots [\pi(a_\alpha^1), [-\mathrm{i}D, \pi(a_\alpha^0)]] \cdots]], \quad a_\alpha^i \in \mathfrak{g}. \tag{16}$$

For $a_\alpha^i = f_\alpha^i \otimes \hat{a}_\alpha^i \in C^\infty(X) \otimes \mathfrak{a}$ we get with (1),

$$\tau^1 = \sum_{\alpha, z \geq 0} (f_\alpha^z \cdots f_\alpha^1 \not{\partial}(f_\alpha^0) \otimes \hat{\pi}([\hat{a}_\alpha^z, [\cdots [\hat{a}_\alpha^1, \hat{a}_\alpha^0] \cdots]])$$

$$+ f_\alpha^z \cdots f_\alpha^1 f_\alpha^0 \boldsymbol{\gamma} \otimes [\hat{\pi}(\hat{a}_\alpha^z), [\cdots [\hat{\pi}(\hat{a}_\alpha^1), [-\mathrm{i}\mathcal{M}, \hat{\pi}(\hat{a}_\alpha^0)]] \cdots]]). \tag{17}$$

Let us first assume that $\mathfrak{a}$ is semisimple. In this case the two lines in (17) are independent. The first line belongs to $\Lambda^1 \otimes \hat{\pi}(\mathfrak{a})$, because the gamma matrices occurring in $\not{\partial}$ provide a 1-form basis. In physical terminology, these Lie algebra-valued 1-forms are Yang–Mills fields acting via the representation $\mathrm{id} \otimes \hat{\pi}$ on the fermions. In the second line of (17) we split $\mathcal{M}$ into generators of irreducible representations of $\mathfrak{a}$, tensorized by generation matrices. Obviously, these irreducible representations are spanned after taking the commutators with $\hat{\pi}(\hat{a}_\alpha^i)$. Thus, the second line of (17) contains sums of function-valued representations of the matrix Lie algebra (times $\boldsymbol{\gamma}$ and generation matrices), which are physically interpreted as Higgs fields. In other words, the prototype τ^1 of a connection form (=gauge potential) describes representations of both Yang–Mills and Higgs fields on the fermionic Hilbert space.

From a physical point of view, this is a more satisfactory picture than the usual noncommutative geometrical construction of Yang–Mills–Higgs models.[3,4] Namely, descending from Connes' noncommutative geometry[2,5,6] there is only a limited set of Higgs multiplets possible:[13] Admissible Higgs multiplets are tensor products $\mathbf{n} \otimes \mathbf{m}^*$ of fundamental representations (and their complex conjugate) $\mathbf{n}, \mathbf{m}$ of simple gauge groups, where the adjoint representation never occurs. This rules out[7] the construction of interesting physical models. In our framework there are no such restrictions and — depending on the choice of $\mathcal{M}$ and h — Higgs fields in any representation of a Lie group are possible. Thus, a much larger class of physical models can be constructed.

The treatment of Abelian factors $\mathfrak{a}'' \subset \mathfrak{a}$ in our approach is somewhat tricky. One remarks that in the first line of (17) only the $(z=0)$-component of $\mathfrak{a}''$ survives. The consequence is that linear independence of the two lines in (17) is not automatic. Thus, to avoid pathologies, we need a condition[1] between $\mathcal{M}$ and the representations of $\mathfrak{a}$ to assure independence. The $u(1)$-part of the standard model is admissible in this sense.

The second consequence of the missing $(z>0)$-components in the first line is that the space–time 1-form part of Abelian factors in τ^1 is a total differential $\not{\partial}(f_0^0) \in d\Lambda^0 \subset \Lambda^1$. This seems to be a disaster at first sight for the description of Abelian Yang–Mills fields. However, our gauge potential lives in the bigger space $\mathcal{H}^1 \supset \pi(\Omega^1)$. Always, if there is a part $d\Lambda^0 \otimes \pi(\mathfrak{a}'')$ in $\pi(\Omega^1)$ there is a part $\Lambda^1 \otimes \pi(\mathfrak{a}'')$ in $\mathcal{H}^1$. There can be even further contributions from $\mathcal{H}^1$ to the gauge potential, which are difficult to control, in general. Fortunately, it turns out[1] that after imposing a locality condition for the connection (which is equivalent to saying that ρ commutes with functions), possible additional $\mathcal{H}^1$-degrees of freedom are either of the Yang–Mills type or the Higgs type.

This framework of gauge field theories was successfully applied to formulate the standard model,[10] the flipped SU(5)×U(1)-grand unification[11] and SO(10)-grand unification.[12] It is not possible to describe pure electrodynamics. The reason is that in the Abelian case the curvature form $\theta\neq 0$ commutes with all elements of $\pi(\Omega)$. Hence, it belongs to the graded centralizer $\mathcal{C}^2$ and is projected away in the bosonic action (14).

IX. DO THE AXIOMS OF NONCOMMUTATIVE GEOMETRY EXTEND TO THE LIE ALGEBRAIC SETTING?

The present status of noncommutative geometry is that this theory is governed by seven axioms.[6] In the commutative case, these axioms provide the algebraic description of classical spin manifolds. The question now is whether or not our Lie algebraic version, which is in close analogy with the prior Connes–Lott formulation[9] of noncommutative geometry, can also be brought into contact with Connes' axioms. We list and discuss below the axioms in their form they would have in terms of Lie algebras.

(1) *Dimension: $|D|^{-1}$ is an infinitesimal of order 1/d, i.e., the eigenvalues E_n of $|D|^{-1}$ grow as $n^{-1/\mathrm{d}}$, where d is an even natural number.*

(3) *Smoothness: For any $a\in\mathfrak{g}$, both a and $[D,a]$ belong to the domain of δ^m, where $\delta(.)$ $:=[|D|,\cdot]$.*

The axioms (1) and (3) can be directly transferred to the Lie algebraic setting. We cannot treat the odd-dimensional case as the grading operator Γ is essential to detect the sign for the graded commutator.

(4) *Orientability:* Connes requires the $\mathbb{Z}_2$-grading operator Γ to be the image under π of a Hochschild d-cycle. We are not going to touch the extension of Hochschild homology to Lie algebras, but even a requirement such as $\Gamma\in\pi(\Omega^{\mathrm{d}})$ is problematic. For the standard example we have the decomposition $\Gamma=\boldsymbol{\gamma}\otimes\hat{\Gamma}$, and the comparison with the general form[1] of $\pi(\Omega^{\mathrm{d}})$ yields that $\hat{\Gamma}$ has to be the image under $\hat{\pi}$ of the non-Abelian part of $\mathfrak{a}$. In all models we have studied so far this is not the case. It seems to be impossible to maintain orientability in our framework. The grading operator Γ, which commutes with $\pi(\mathfrak{g})$ and anti-commutes with D, is an extra piece which has no relation with orientability.

(7) *Reality: There exists an antilinear isometry $J:h_i\to h_i$ such that $[\pi(a),J\pi(b)J^{-1}]=0$ for all $a,b\in\mathfrak{g}$, $J^2=\epsilon$, $JD=DJ$ and $J\Gamma=\epsilon'\Gamma J$, with $\epsilon=(-1)^{\mathrm{d(d+2)/8}}$ and $\epsilon'=(-1)^{\mathrm{d/2}}$.*

(2) *First order: $[[D,\pi(a)],J\pi(b)J^{-1}]=0$ for all $a,b\in\mathfrak{g}$.*

Both axioms (7) and (2) can be trivially fulfilled as soon as an antilinear involution $\mathcal{I}$ on h_i is available. It suffices to define

$$J:=\begin{pmatrix}0 & \epsilon\mathcal{I}^{-1}\\ \mathcal{I} & 0\end{pmatrix},\quad h_i\mapsto\begin{pmatrix}h_i\\ h_i\end{pmatrix},\quad D\mapsto\begin{pmatrix}D & 0\\ 0 & \mathcal{I}D\mathcal{I}^{-1}\end{pmatrix},$$

$$\pi(a)\mapsto\begin{pmatrix}\pi(a) & 0\\ 0 & 0\end{pmatrix},\qquad \Gamma\mapsto\begin{pmatrix}\Gamma & 0\\ 0 & \epsilon'\mathcal{I}\Gamma\mathcal{I}^{-1}\end{pmatrix}.$$

The question is whether there are nontrivial real structures which also satisfy the other axioms. The existence of the real structure J (Tomita's involution) is a central piece of Connes' theory. It has proved very useful in understanding the commuting electroweak and strong sectors of the standard model. The same idea could be applied to our formulation of the standard model.[10] For other gauge theories,[11,12] however, a nontrivial real structure J seems to be rather disturbing as it requires the fermions to sit in (generalized) adjoint representations. To achieve this one had to add auxiliary $u(1)$-factors, which is in contradiction to the grand unification philosophy.

(5) *Finiteness and absolute continuity:* Connes requires $h_\infty = \cap_m \text{domain}(D^m)$ to be a finite projective module. Thus, our task would be to define the notion of a finite projective module over a Lie algebra $\mathfrak{g}$ and the Lie analogs of the *K*-groups. We are not aware of these structures, but without them it is impossible to talk about generalizations of the index pairing of *D* with the *K*-groups and of the following.

(6) *Poincaré duality*.

In conclusion, our Lie algebraic version of noncommutative geometry is not a possible generalization of classical spin manifolds, or at least there is a lot to do to derive the Lie analogs of standard algebraic structures. Our approach provides a powerful tool to build gauge field theories with spontaneous symmetry breaking; the price for this achievement is the lost of any contact with spin manifolds.

ACKNOWLEDGMENT

This 1998 work was supported by the German Academic Exchange Service (DAAD), Grant No. D/97/20386.

[1] R. Wulkenhaar, ''Non-commutative geometry with graded differential Lie algebras,'' J. Math. Phys. **38**, 3358–3390 (1997).
[2] A. Connes, *Non Commutative Geometry* (Academic, New York 1994).
[3] B. Iochum, D. Kastler, and T. Schücker, ''Fuzzy mass relations in the standard model,'' preprint hep-th/9507150.
[4] C. P. Martín, J. M. Garcia-Bondia, and J. C. Varilly, ''The standard model as a noncommutative geometry: the low energy regime,'' Phys. Rep. **294**, 363–406 (1998).
[5] A. Connes, ''Noncommutative geometry and reality,'' J. Math. Phys. **36**, 6194–6231 (1995).
[6] A. Connes, ''Gravity coupled with matter and the foundation of noncommutative geometry,'' Commun. Math. Phys. **182**, 155–176 (1996).
[7] F. Lizzi, G. Mangano, G. Miele, and G. Sparano, ''Constraints on unified gauge theories from noncommutative geometry, '' Mod. Phys. Lett. A **11**, 2561–2572 (1996).
[8] R. Wulkenhaar, ''Graded differential Lie algebras and model building,'' J. Geom. Phys. **25**, 305–325 (1998).
[9] A. Connes and J. Lott, ''The metric aspect of noncommutative geometry,'' in *Proceedings of the 1991 Cargèse Summer Conference,* edited by J. Fröhlich *et al.* (Plenum, New York, 1992), pp. 53–93.
[10] R. Wulkenhaar, ''The standard model within non-associative geometry, '' Phys. Lett. B **390**, 119–1276 (1997).
[11] R. Wulkenhaar, ''Graded differential Lie algebras and SU(5)×U(1)-grand unification,'' Int. J. Mod. Phys. A **13**, 2627–2692 (1998).
[12] R. Wulkenhaar, ''SO(10)-unification in noncommutative geometry revisited,'' Int. J. Mod. Phys. A (to be published), preprint hep-th/9804046.
[13] T. Krajewski, ''Classification of finite spectral triples,'' J. Geom. Phys. **28**, 1–30 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Reduction of constrained systems with symmetries

Frans Cantrijn[a)]
Theoretical Mechanics Division, University of Gent, Krijgslaan 281, B-9000 Gent, Belgium

Manuel de León[b)]
Instituto de Matemáticas y Física Fundamental, CSIC, Serrano 123, E-28006 Madrid, Spain

Juan Carlos Marrero[c)]
Departamento de Matemática Fundamental, Facultad de Matemáticas, Universidad de la Laguna, Tenerife, Canary Islands, Spain

David Martín de Diego[d)]
Departamento de Economía Aplicada (Matemáticas), Facultad de CC. Economicas y Empresariales, Avda Valle Esgueva 6, E-47011 Valladolid, Spain

(Received 13 July 1998; accepted for publication 21 September 1998)

A general model is proposed for constrained dynamical systems on a symplectic manifold which covers, among others, the description of Lagrangian and Hamiltonian systems with nonholonomic constraints and the canonical description of mechanical systems with a singular Lagrangian. The reduction properties of these systems in the presence of symmetry are investigated within this general framework. © *1999 American Institute of Physics.* [S0022-2488(99)00902-0]

I. INTRODUCTION

In this paper, we propose a general model for constrained Hamiltonian systems on a symplectic manifold, providing a unified setting for the description of various types of mechanical systems with constraints. Special attention is then paid to aspects concerning symmetry and reduction for the class of systems under consideration.

The constraints encountered in classical mechanics can be classified, roughly speaking, into two different categories, which may be labeled as ''internal constraints'' and''external constraints,'' respectively. Internal constraints are those that find their origin in the degeneracy of the Lagrangian describing a certain system, which prevents a straightforward transition to an equivalent Hamiltonian formulation. This type of constraint usually reflects the presence of ''gauge'' degrees of freedom, and is, in fact, more relevant to relativistic mechanics and field theory. The standard treatment of degenerate (or singular) Lagrangian systems is based on the Dirac–Bergmann constraint analysis (see, e.g., Ref. 1). An intrinsic geometric formulation and generalization of this theory is provided by the so-called presymplectic constraint algorithm, developed by Gotay and Nester.[2–4] A different approach to singular Lagrangian systems, advocated by Tulczyjew, consists in treating them as implicit dynamical systems (see, e.g., Ref. 5).

External constraints refer to those physical constraints which are imposed on a given system from outside. Here we can make a further distinction between (time-independent or time-dependent) holonomic and nonholonomic, one-sided and two-sided constraints. Holonomic constraints are restrictions on the admissible configurations (i.e., ''positions'') of the system under consideration, whereas nonholonomic constraints depend on the velocities in an essential way, i.e.,

[a)]Senior Research Associate at the Fund for Scientific Research—Flanders, Belgium. Electronic mail: frans.cantrijn@rug.ac.be
[b)]Electronic mail: mdeleon@fresno.csic.es
[c)]Electronic mail: jcmarrer@ull.es
[d)]Electronic mail: dmartin@esgueva.eco.uva.es

they can not be integrated to relations depending on the position variables only. The prototype of nonholonomic constraints are the conditions for ''rolling without slipping.'' For the purpose of this paper, we restrict our attention to the case of systems with two-sided, time-independent nonholonomic constraints. (For a geometric approach to systems with one-sided constraints, which are analytically expressed by inequalities, we refer to Refs. 6, 7.) The classical approach to nonholonomic mechanical systems is based on the method of Lagrange multipliers (see, e.g., Ref. 8 for a comprehensive treatment). The fundamental work of Vershik and Faddeev[9,10] has marked the beginning of a period of intensive research on nonholonomic systems within the realm of geometric mechanics: see, for instance, Refs. 11–14 for a more detailed bibliography. In particular, the relevance of these studies for the further development of control theory has recently attracted a lot of attention (see, e.g., Refs. 15, 16, and references therein). We note, in passing, that nonholonomic systems have also been treated as implicit dynamical systems by Ibort *et al.*[11]

In spite of the difference in the ''physical'' nature of the constraints, it turns out that the geometrical models adopted for describing systems with either internal or external constraints, have many aspects in common. Indeed, in the canonical treatment of degenerate systems as well as in the Lagrangian and Hamiltonian treatment of nonholonomic systems, the search for consistent equations of motion eventually leads to a framework consisting of the following ingredients: a symplectic manifold (P,ω), a smooth function H on P, a submanifold M of P, and a distribution F along M (i.e., a subbundle of the restricted tangent bundle $TP_{|M}$). Depending on the case, P hereby represents the velocity phase space TQ or the momentum phase space T^*Q of the system under consideration, with underlying configuration space Q, and ω is either the Poincaré–Cartan 2-form on TQ, induced by a regular Lagrangian, or the canonical symplectic form on T^*Q. In the case of degenerate systems, M is the ''final constraint submanifold'' generated by the appropriate constraint algorithm, and F coincides either with TM or with the tangent bundle of a larger submanifold containing M (the primary constraint submanifold). H denotes the energy function or the (extended) Hamiltonian. In the case of a nonholonomic system, M simply denotes the constraint submanifold defined by the given external constraints, and the distribution F is characterized by the property that its annihilator is the co-distribution generated by the reaction forces, induced by the constraints. The problem then consists in finding a vector field on P, generated by H, which is tangent to M and compatible, in an appropriate sense, with the distribution F.

In the present paper we will take the above ingredients as building stones for constructing a general model for constrained dynamical systems in a symplectic setting. This model can be seen, in particular, as a unifying model for the description of degenerate systems as well as of mechanical systems with nonholonomic constraints. Our main goal then is to study the geometry of such systems in the presence of symmetry. Guided by various recent treatments of nonholonomic systems with symmetry (cf. Refs. 17, 15, 12, 18, 19), we will discuss in some detail the reduction problem for general constrained Hamiltonian systems with symmetry.

The scheme of this paper is as follows. In the next section we briefly recall some aspects of the geometrical approach to singular Lagrangian systems and to systems with nonholonomic constraints. In Sec. III we then propose a general model for constrained systems and investigate the existence and uniqueness conditions for the dynamics. In Sec. IV we deal with the problem of solving the dynamics. In Sec. V, we introduce symmetry into our model and present some general reduction results. After putting forward a classification of constrained systems with symmetry, inspired on the one introduced by Bloch *et al.*[15] for nonholonomic systems, we describe some further reduction results for each class separately in Sec. VI. Finally, in Sec. VII we illustrate the obtained results on some particular cases.

Throughout this paper, we work in the category of smooth (i.e., C^∞) objects. For convenience, we will usually not make a notational distinction between a (vector) bundle over a manifold and the ring of its smooth sections, i.e., if F denotes a vector bundle over a manifold N (for instance, a subbundle of TN), then $X \in F$ simply means that $X:N \rightarrow F$ is a section of F. The sole exception to this rule will be the occasional use of the notation $\mathfrak{X}(N)$ for the ring of smooth vector fields on N.

II. EQUATIONS OF MOTION OF CONSTRAINED LAGRANGIAN SYSTEMS

Consider a smooth, finite dimensional manifold Q, with local coordinates denoted by (q^A). As is well known, the tangent bundle TQ of Q, with canonical projection $\tau_Q : TQ \to Q$, is equipped with a dilation vector field Δ, i.e., the so-called Liouville vector field, and a canonical type (1,1) tensor field S, called the vertical endomorphism, which determines the almost tangent structure of TQ. In the natural bundle coordinates (q^A, v^A) of TQ these objects read as

$$\Delta = v^A \frac{\partial}{\partial v^A}, \quad S = \frac{\partial}{\partial v^A} \otimes dq^A.$$

Given a Lagrangian on TQ, i.e., a smooth function $L: TQ \to \mathbb{R}$, one can define the corresponding Poincaré–Cartan 1- and 2-forms θ_L and ω_L, respectively, and the energy function E_L, according to

$$\theta_L = S^*(dL), \quad \omega_L = -d\theta_L, \quad E_L = \Delta(L) - L,$$

with S^* denoting the action of S on 1-forms. In geometrical terms, the equations of motion for the Lagrangian system with Lagrangian L can then be expressed by

$$i_Z \omega_L = dE_L. \tag{1}$$

If L is regular, that is, if the Hessian matrix $(\partial^2 L / \partial v^A \partial v^B)$ is nondegenerate everywhere, then ω_L is a symplectic form. In that case (1) admits a unique solution for Z, which we will denote by Γ_L, and which is usually called the Euler–Lagrange vector field corresponding to L. In particular, Γ_L is a second order differential equation field (SODE, for short), that is, $S(\Gamma_L) = \Delta$. The base integral curves $q^A(t)$ of Γ_L (i.e., the projections of its integral curves onto Q) verify the Euler–Lagrange equations,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^A}\right) - \frac{\partial L}{\partial q^A} = 0. \tag{2}$$

For later use, we recall that the symplectic form ω_L, corresponding to a regular Lagrangian, induces two bundle isomorphisms (''musical mappings'') $\flat_L : T(TQ) \to T^*(TQ)$ and $\sharp_L : T^*(TQ) \to T(TQ)$, where $\flat_L(X) = i_X \omega_L$ and $\sharp_L = \flat_L^{-1}$.

A. Singular Lagrangian systems

A Lagrangian (system) is called singular, or degenerate, if the Hessian matrix $(\partial^2 L / \partial v^A \partial v^B)$ is singular. In such a case, the equation of motion (1), in general, does not have a solution, and if a solution exists, it will not be unique. If the Poincaré–Cartan 2-form ω_L corresponding to a singular Lagrangian has constant rank and, hence, ω_L happens to be a presymplectic form, one can apply the so-called presymplectic constraint algorithm, developed by Gotay and Nester (see, e.g., Refs. 2, 3). This algorithm generates a descending sequence of constraint submanifolds which, under the appropriate conditions, converges to a closed immersed submanifold P_f of TQ (the ''final constraint submanifold'') on which there exist consistent equations of motion for the given system. More precisely, it follows by construction that the equation

$$(i_Z \omega_L - dE_L)_{|P_f} = 0 \tag{3}$$

admits at least one solution Z which is everywhere tangent to P_f. In addition, one can always find a submanifold of P_f on which there exists a unique solution Z which also verifies the SODE condition (cf. Ref. 4). For quantization purposes, however, it is more convenient to develop the analysis of a degenerate Lagrangian system on an ambient symplectic space. This can be achieved by passing to an appropriate Hamiltonian formulation.

Let us recall that, given a Lagrangian L, the Legendre map Leg:$TQ \to T^*Q$ is a fibered mapping over Q, which is locally written as

$$\mathrm{Leg}(q^A, v^A) = (q^A, p_A),$$

where $p_A = \partial L / \partial v^A$. If ω_Q denotes the canonical symplectic form on T^*Q, we have $\mathrm{Leg}^* \, \omega_Q = \omega_L$ (see, e.g., Ref. 20). Hence, L is regular if and only if Leg is a local diffeomorphism, and L is said to be hyperregular if Leg is a global diffeomorphism. For a singular Lagrangian system it is not possible, in general, to obtain a consistent Hamiltonian description. Let us assume, however, that L is almost-regular, i.e., $M_1 = \mathrm{Leg}(TQ)$ is a submanifold of T^*Q and Leg is a submersion onto M_1 with connected fibers. In that case, the energy function E_L projects onto a function $h_1 : M_1 \to \mathbb{R}$ which is uniquely determined by $h_1 \circ \mathrm{Leg}_1 = E_L$, where Leg_1 simply stands for the ''restriction'' of Leg, regarded as a mapping from TQ onto M_1. If we now denote by ω_1 the pull-back of ω_Q to M_1, then the equation

$$i_X \omega_1 = dh_1, \tag{4}$$

is precisely the Hamiltonian counterpart of Eq. (1).

Starting from (4), one can again apply the presymplectic constraint algorithm which, in case the given problem is consistent, leads to a nonempty final constraint submanifold M_f such that the equation

$$(i_X \omega_1 - dh_1)_{|M_f} = 0 \tag{5}$$

admits well-defined solutions. This approach yields a global version of the classical Dirac–Bergmann theory for constrained systems.[1] Let ''$\perp$'' denote the symplectic orthogonal with respect to the canonical symplectic form ω_Q. Then, if X is an arbitrary solution of (5), all other solutions will be of the form $X+Y$, with $Y \in TM_f \cap TM_1^{\perp}$.

A simple argument shows that, for almost regular Lagrangians, the Lagrangian and the Hamiltonian formulations are fully equivalent (see Refs. 2, 3). In particular, the final constraint submanifolds on both sides are connected via the Legendre transformation, in the sense that the latter induces a fibration $\mathrm{Leg}_f : P_f \to M_f$. Whenever Z is a projectable solution of (3), its projection onto M_f yields a solution of (5) and, conversely, given a solution X of (5), any vector field Z on P_f which projects onto X satisfies (3).

A geometric constraint algorithm, closely related to the Gotay–Nester approach, is the one developed by Hinds[21] (see also Ref. 2 for a brief discussion). Again starting from (4), this algorithm generates a descending sequence of constraint submanifolds. In the favorable case, the algorithm stabilizes at a final constraint submanifold which, for simplicity, we will denote again by M_f. It is important to point out that, in general, this M_f will be different from the final constraint submanifold obtained by the presymplectic constraint algorithm. In principle, both algorithms start to diverge from each other after the second step. This is due to the fact that in the Hinds algorithm, at each step, possibly new constraints are generated by imposing consistency conditions on the equations of motion induced on the previous constraint submanifold by a pull-back procedure. In the presymplectic constraint algorithm, on the other hand, the consistency conditions are imposed on the equations obtained by taking the restriction of (4) to the successive constraint submanifolds. The equations of motion obtained through Hinds' algorithm can be written as

$$i_X \omega_f = dh_f, \tag{6}$$

with ω_f and h_f denoting the pull-back to M_f of ω_1 and h_1, respectively. Given a solution X of this equation, it follows that $X+Y$ is also a solution for any $Y \in TM_f \cap TM_f^{\perp}$. Note that (6) is an equation induced on the final constraint submanifold, i.e., it expresses an equality of 1-forms on

M_f, whereas (5) represents an equality of 1-forms on M_1, restricted to points of the corresponding M_f. This indeed reflects the difference in spirit between both algorithms, as described above.

Following Dirac,[1] the constraints produced in the course of the constraint analysis can be classified in two different ways. On the one hand, depending on the order of appearance, there are primary, secondary, (tertiary, etc.,) constraints. On the other hand, there is the more significant distinction between first and second class constraints. In physics, it is customary to assume that all first class constraints (primary, secondary,...) are generators of gauge transformations, i.e., transformations that do not change the physical state of the system (see, e.g., Refs. 22, 23). This property is automatically verified when applying Hinds' algorithm. In the Gotay–Nester approach, all primary first class constraints generate gauge transformations but, in general, this need not be the case for all subsequent (secondary,...) first class constraints (see Ref. 2 for more details). From a physical point of view, therefore, it may be argued that (6) is in better agreement with the ''standard'' interpretation of gauge transformations than (5).

We will now recast the equations (5) and (6) into a form which better serves our purpose. Taking an arbitrary extension $H_1 : T^*Q \to \mathbb{R}$ of the Hamiltonian $h_1 : M_1 \to \mathbb{R}$, it follows that (5) is formally equivalent to

$$(i_X\omega_Q - dH_1)_{|M_f} \in (TM_1^o)_{|M_f}, \quad X_{|M_f} \in TM_f, \tag{7}$$

where TM_1^o is the annihilator of TM_1 in T^*T^*Q. Locally, these conditions precisely generate the equations of motion ensuing from the classical Dirac–Bergmann constraint analysis. Likewise, taking an arbitrary extension H_f of h_f to T^*Q, (6) can be rewritten in terms of the canonical symplectic form as follows:

$$(i_X\omega_Q - dH_f)_{|M_f} \in TM_f^o, \quad X_{|M_f} \in TM_f, \tag{8}$$

where it should be emphasized again that, for the same system, the constraint submanifolds M_f in (7) and (8), in general, need not be the same.

B. Nonholonomic Lagrangian systems

In this section, we start by considering a regular Lagrangian system with Lagrangian $L : TQ \to \mathbb{R}$, subjected to a set of nonholonomic constraints which are linear in the velocities, i.e., they can be (locally) represented by a set of independent functions of the form $\phi_i := \mu_{iA}(q)v^A$, for $1 \leqslant i \leqslant m$. We can describe this nonholonomic Lagrangian system in geometrical terms as follows. The constraint equations $\phi_i = 0$ define a $(n-m)$-dimensional distribution $\mathcal{D}$ on the n-dimensional configuration manifold Q. We denote its total space by D, which is a $(2n-m)$-dimensional submanifold of TQ: *the constraint submanifold*. For simplicity we always assume in the sequel that $\tau_Q(D) = Q$, i.e., the constraints are ''purely kinematical'' in the sense that they do not impose restrictions on the allowable positions. The motions of the system are forced to take place on D, and this requires the introduction of some (unknown) ''reaction forces.'' In de León *et al.*,[24] an intrinsic expression for the equations of motion was obtained, which we will describe below.

First of all, we define a distribution $\mathcal{D}^v$ on TQ by prescribing its annihilator as a subbundle of T^*TQ which, along the constraint submanifold D, represents the bundle of reaction forces. More precisely, given a set of independent 1-forms $\{\mu_i ; 1 \leqslant i \leqslant m\}$ on Q, which locally generate the annihilator $\mathcal{D}^o$ of $\mathcal{D}$, we put

$$(\mathcal{D}^v)^o = \langle \mu_i^v \rangle,$$

where μ_i^v denotes the vertical lift of the 1-form μ_i to TQ (see Ref. 20). A direct computation reveals that $\mathcal{D}^v$ is, in fact, globally defined. Note, in passing, that with $\mu_i = \mu_{iA}\, dq^A$, the given constraint functions ϕ_i are precisely the evaluation maps of these 1-forms.

Next, it can then be shown that the equations of motion for such a nonholonomic mechanical system are given by

$$(i_X\omega_L - dE_L)_{|D} \in (\mathcal{D}^v)^o, \quad X_{|D} \in TD. \tag{9}$$

It should be pointed out that each solution of (9) (if there exists one) is automatically a SODE along D. This implies that, in local coordinates, the integral curves of X on D are of the form $(q^A(t),\dot{q}^A(t))$, whereby the $q^A(t)$ are solutions of the system of differential equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^A}\right) - \frac{\partial L}{\partial q^A} = \lambda^i \mu_{iA}, \tag{10}$$

together with the constraint equations $\mu_{iA}(q)\dot{q}^A = 0$, and where the λ^i are Lagrange multipliers.

We will now describe a procedure which permits us to decide under what conditions (9) admits a solution and, if these conditions are fulfilled, to obtain a solution by projection of the Euler–Lagrange vector field of the corresponding unconstrained system.

Applying the isomorphism $\sharp_L$ to the co-distribution $(\mathcal{D}^v)^o$ we obtain the symplectic orthogonal complement (with respect to ω_L) of $\mathcal{D}^v$, i.e., $\sharp_L((\mathcal{D}^v)^o) = (\mathcal{D}^v)^\perp$. Obviously, $\dim(\mathcal{D}^v)^\perp$ is equal to the number of independent constraints. We will say that the given nonholonomic system satisfies the *compatibility condition* if $T_xD \cap (\mathcal{D}^v_x)^\perp = \{0\}$ at each point $x \in D$. In such a case, taking into account that $\dim(\mathcal{D}^v_x)^\perp = m$, we have a direct sum decomposition,

$$T_x(TQ) = T_xD \oplus (\mathcal{D}^v_x)^\perp, \quad x \in D,$$

which, in turn, gives rise to two complementary projectors, say

$$\mathcal{P}_x : T_x(TQ) \to T_xD, \quad \mathcal{Q}_x : T_x(TQ) \to (\mathcal{D}^v_x)^\perp.$$

A direct calculation shows that $\Gamma_{L,D} = \mathcal{P}(\Gamma_{L_{|D}})$ is a solution of (9). Moreover, one can easily show that this solution is necessarily unique.[24] The procedure just described is essentially equivalent to the classical one based on the use of Lagrange multipliers.

If the nonholonomic system does not verify the compatibility condition, that is, $T_xD \cap (\mathcal{D}^v_x)^\perp \neq \{0\}$ at some points $x \in D$, we can develop a constraint algorithm which is very similar to the one described above for singular Lagrangians (cf. Ref. 24). Under the appropriate conditions, this algorithm determines a final constraint submanifold D_f on which there exist consistent equations of motion for the given constrained problem. More precisely, the algorithm guarantees the existence of well-defined solutions X of the system

$$(i_X\omega_L - dE_L)_{|D_f} \in (\mathcal{D}^v)^o, \quad X_{|D_f} \in TD_f \tag{11}$$

(see Ref. 24 for details). Again, it turns out that a solution of (11) is a SODE along D_f.

The previous analysis of nonholonomic systems can be further extended to the case where, in addition, the Lagrangian happens to be singular (see Refs. 25, 26). To fix the ideas, let us assume that L is almost regular, and that $\ker(\mathrm{Leg})_* \subset \mathcal{D}^c$, where $\mathcal{D}^c$ denotes the tangent or complete lift of $\mathcal{D}$ to TQ, i.e., $\mathcal{D}^c$ is the distribution on TQ whose annihilator is given by $(\mathcal{D}^c)^o = \langle \mu_i^v, \mu_i^c \rangle$. The nonholonomic mechanical system (L,D) is then also said to be almost regular. Under these assumptions, the following is proved in Ref. 25 (using the notations of the previous subsection):

- $\overline{D} = \mathrm{Leg}(D)$ is a submanifold of $M_1 = \mathrm{Leg}(TQ)$, and the restriction $\mathrm{Leg}_{|D} : D \to \overline{D}$ is a surjective submersion whose fiber at a point $x \in \overline{D}$ is precisely given by $\mathrm{Leg}^{-1}(x)$.
- If $\{\mu_i\}$ is a basis of $\mathcal{D}^o$, then $(\overline{\mathcal{D}^v})^o = \langle \pi_Q^* \mu_i \rangle$ defines a distribution $\overline{\mathcal{D}^v}$ on T^*Q, where $\pi_Q : T^*Q \to Q$ is the canonical projection. By construction, the distributions $\mathcal{D}^v$ and $\overline{\mathcal{D}^v}$ are Leg-related.
- Let $\overline{\mathcal{D}_1^v}$ be the distribution on M_1, the annihilator of which is the co-distribution obtained by taking the pull-back to M_1 of the forms generating $(\overline{\mathcal{D}^v})^o$. The system

$$(i_Y\omega_1 - dh_1)_{|\overline{D}} \in (\overline{\mathcal{D}_1^v})^o, \quad Y_{|\overline{D}} \in T\overline{D}, \tag{12}$$

is then equivalent to the system

$$(i_X\omega_L - dE_L)_{|D} \in (\mathcal{D}^v)^o, \quad X_{|D} \in TD,$$

where h_1 is the projection of E_L onto M_1. Indeed, one can develop a constraint algorithm for both systems such that, at each stage, the respective constraint submanifolds D_k and $\overline{D}_k$ are Leg-related, that is, $\mathrm{Leg}(D_k) = \overline{D}_k$. Moreover, with D_f and $\overline{D}_f$ denoting the final constraint submanifolds, one can show that the Legendre map induces a surjective submersion $\mathrm{Leg}_f : D_f \to \overline{D}_f$ which projects solutions onto solutions.

In particular, when applying the constraint algorithm to (12) we end up with the dynamical equation,

$$(i_Y\omega_1 - dh_1)_{|\overline{D}_f} \in (\overline{\mathcal{D}_1^v})^o, \quad Y_{|\overline{D}_f} \in T\overline{D}_f, \tag{13}$$

which, by construction, admits well-defined solutions Y. Finally, as in the treatment of (free) singular Lagrangian systems, discussed in the previous subsection, one can prove the formal equivalence of (13) with

$$(i_Y\omega_Q - dH)_{|\overline{D}_f} \in (\overline{\mathcal{D}^v} \cap TM_1)^o, \quad Y_{|\overline{D}_f} \in T\overline{D}_f, \tag{14}$$

where $H: T^*Q \to \mathbb{R}$ is an arbitrary extension of h_1.

Remark II.1: In the previous discussion we have confined ourselves to the case of linear nonholonomic constraints. Much of the above, however, applies equally well to the case of affine or even nonlinear constraints. For instance, for a regular Lagrangian system, subjected to nonlinear, nonholonomic constraints, described by a submanifold M of the tangent bundle TQ, the equations of motion are again of the form (9), with the vector subbundle D being replaced by M and $(\mathcal{D}^v)^o$ by the co-distribution $S^*((TM)^o)$ (see, e.g., Refs. 27, 28).

III. A GENERAL FRAMEWORK FOR CONSTRAINED SYSTEMS

When looking at the systems (7), (8), (9), (11) and (14), we see that, in spite of the difference in (physical) origin and interpretation, they all have a similar geometrical structure. This prompts us to introduce the following general model for constrained dynamical systems within a symplectic setting.

Consider a symplectic manifold (P,ω), a smooth function $H: P \to \mathbb{R}$ (the Hamiltonian), an embedded submanifold M of P (the constraint submanifold) and a distribution F on P along M, i.e., F is a vector subbundle of $TP_{|M}$. We are then interested in the following problem: find a smooth section X of the restricted tangent bundle $TP_{|M} \to M$, such that

$$(i_X\omega - dH)_{|M} \in F^o, \quad X \in TM, \tag{15}$$

with F^o the annihilator of F in $T^*P_{|M}$. In particular, X then defines a vector field on M. It is clear that (7), (8), (9), (11) and (14) belong to the class of problems described by (15). (We thereby ignore the technicality that in the treatment of singular Lagrangian systems, the final constraint submanifold, in principle, may be an immersed rather than an embedded submanifold.)

In what follows we will denote by $\flat : TP \to T^*P, X \mapsto i_X\omega$ and $\sharp = \flat^{-1}: T^*P \to TP$, the bundle isomorphisms over P induced by the symplectic form ω.

We now first study the problem of the existence and uniqueness of solutions of the constrained system (15).

Proposition III.1: (i) (Existence) The system (15) admits a solution if and only if

$$dH(x) \in (F_x \cap T_x M^\perp)^o,$$

at each point $x \in M$.

(ii) (Uniqueness) If (15) has a solution, then it is unique if and only if

$$F^\perp \cap TM = 0.$$

Proof: (i) If X is a solution of (15) then, since $X(M) \subset TM$, we have $i_X \omega_{|M} \in (TM^\perp)^o$, from which it follows that

$$dH_{|M} \in F^o + (TM^\perp)^o = (F \cap TM^\perp)^o.$$

Conversely, assume that $dH_{|M} \in F^o + (TM^\perp)^o$. Then, $dH_{|M} - \beta \in (TM^\perp)^o$ for some $\beta \in F^o$. Since $\flat(TM) = (TM^\perp)^o$, we deduce that there exists a vector field X satisfying (15).

(ii) Now, let X and X' be two solutions of (15). Then

$$X - X' \in F^\perp \cap TM.$$

Hence, if a solution exists, it will be unique if and only if $F^\perp \cap TM = 0$. Q.E.D.

Note that the existence condition can be equivalently expressed as

$$X_{H|M} \in TM + F^\perp,$$

where X_H denotes the (unconstrained) Hamiltonian vector field on (P, ω) with Hamiltonian H. Hence, any solution X of (15) is of the form

$$X = X_{H|M} + Z, \tag{16}$$

for some $Z \in F^\perp$. An interesting special case occurs when rank $F = \dim M$ or, equivalently, $\dim F_x = \dim T_x M$ for all $x \in M$.

Corollary III.2: If rank $F = \dim M$, *then the condition* $F^\perp \cap TM = 0$ *implies both the existence and uniqueness of a solution of (15).*

Proof: A simple algebraic argument shows that, under the given assumptions, $TP_{|M} = F^\perp \oplus TM$. Taking the symplectic complements of both sides, we find that $0 = F \cap TM^\perp$ and, hence, $T^*P_{|M} = (F \cap TM^\perp)^o$. The result now readily follows from the previous Proposition. Q.E.D.

Under the conditions of the Corollary, (15) is a constrained Hamiltonian system in the sense of Marle,[19] who has studied such systems in the more general setting of Poisson manifolds.

Let us now check the existence and uniqueness conditions for the examples discussed in the previous section. For the nonholonomic system (9), with a regular Lagrangian, we have $(P, \omega) = (TQ, \omega_L)$, $M = D$ and $F = \mathcal{D}^v$. The compatibility condition introduced for such a system precisely coincides with the unicity condition from Proposition 3.1. Since a simple counting of dimensions shows that rank $\mathcal{D}^v = \dim D$, it follows from the above Corollary that a compatible nonholonomic system indeed admits a unique solution. For the other cases (7), (8), (11) and (14), we note that the equations of motion are obtained after applying a constraint algorithm. The latter is precisely conceived so as to guarantee the existence of a consistent solution, i.e., in these cases the existence condition of Proposition 3.1 holds by construction. The uniqueness condition, however, need not be satisfied: in general, there will be ''gauge degrees of freedom.''

Returning to the general model (15), it is important to point out that if the system admits a solution X, it need not be true, in general, that (the restriction of) H is a first integral of X. In classical mechanics, for instance, it is well known that imposing nonholonomic constraints on a conservative mechanical system may destroy the conservation of energy (see, e.g., Ref. 19). An additional assumption on the nature of the constraints therefore is needed to ensure the conservation of energy. For a Lagrangian system subject to general (i.e., not necessarily linear) nonholonomic constraints, a sufficient condition for the energy E_L to be conserved is that the constraints are ''homogeneous,'' which, in geometrical terms, means that the dilation vector field Δ should be tangent to the constraint submanifold (see Refs. 27, 28, where the less appropriate denomination ''ideal constraints'' was used instead of homogeneous constraints). In the case of linear constraints, this condition is always fulfilled.

Remark III.3: If (15) admits no solution, then it is possible to develop a constraint algorithm which, at least in case the given problem is consistent, will lead to a final constraint submanifold M_f on which there exist a well-defined dynamics. The system to be considered then reads as

$$(i_X\omega - dH)_{|M_f} \in F^o, \quad X_{|M_f} \in TM_f, \tag{17}$$

which is again of the same type as (15). By construction this system now has well-defined solutions. Therefore, without loss of generality, we will henceforth always assume that the existence condition of Proposition 3.1 is satisfied

IV. SOLVING THE DYNAMICS

Given a constrained system of the form (15) for which condition (i) of Proposition 3.1 holds, we will now indicate how one can explicitly construct a (local) solution for the dynamics.

Let X_H again denote the ''unconstrained'' Hamiltonian system on (P,ω), corresponding to the Hamiltonian H. Take a local basis $\{\mu_i; 1 \leqslant i \leqslant m\}$ of F^o, and let $\{\Phi_a; 1 \leqslant a \leqslant s\}$ be an independent set of constraint functions which locally define M. Denote by Z_i the symplectic gradient of μ_i, that is, $\flat(Z_i) = \mu_i$. Then, $F^\perp$ is locally generated by the vector fields Z_i and, according to (16), any solution X of (15) can be written as

$$X = X_H + \lambda^i Z_i,$$

where the λ^i are Lagrange multipliers which can be determined from the tangency condition:

$$0 = X(\Phi_a)_{|M} = X_H(\Phi_a)_{|M} + \lambda^i Z_i(\Phi_a)_{|M}, \quad \forall a.$$

Indeed, the existence condition for solutions of (15), in particular, implies that this system of equations can be solved for the λ^i, i.e., on M we have

$$\operatorname{rank}(Z_i(\Phi_a)) = \operatorname{rank}(Z_i(\Phi_a); -X_H(\Phi_a)).$$

Of course, the solution for the λ^i need not be unique.

Next, let us assume that both conditions of Proposition 3.1 are satisfied, so that the system admits a unique solution. Our goal now is to construct a projection operator which allows us to deduce the constrained dynamics from the unconstrained dynamics X_H.

From the assumption $F^\perp \cap TM = 0$ it readily follows that for each $x \in M$, $\dim F_x^\perp \leqslant \operatorname{codim} T_xM$, i.e., $m = \operatorname{corank} F \leqslant \operatorname{codim} M = s$. We can now distinguish the following two cases.

Assume $m = s$.

A simple dimensional argument shows that

$$TP_{|M} = TM \oplus F^\perp.$$

Therefore, there exist two complementary projectors $\mathcal{P}: TP_{|M} \to TM$ and $\mathcal{Q}: TP_{|M} \to F^\perp$ and it is straightforward to check that $\mathcal{P}(X_H)$ is a solution of (15). Using the above notations, a local expression for $\mathcal{P}$ is given by

$$\mathcal{P} = Id - \mathcal{C}^{ij} Z_i \otimes d\Phi_j,$$

where $(\mathcal{C}^{ij})$ is the inverse of the regular matrix $(\mathcal{C}_{ij})$, with $\mathcal{C}_{ij} = Z_j(\Phi_i)$. Hence we obtain

$$\mathcal{P}(X_H) = X_H - \mathcal{C}^{ij} X_H(\Phi_j) Z_i. \tag{18}$$

Assume $m<s$.

In this case, we have

$$TM\oplus F^{\perp}\subsetneq TP_{|M},$$

with two complementary projectors $\mathcal{P}:TM\oplus F^{\perp}\to TM$ and $\mathcal{Q}:TM\oplus F^{\perp}\to F^{\perp}$. From the existence condition it easily follows that $X_{H|M}\in TM\oplus F^{\perp}$. As above, the projection $\mathcal{P}(X_H)$ then provides the unique solution of the constrained dynamics.

The matrices $(Z_i(\Phi_a))$ and $(Z_i(\Phi_a);-X_H(\Phi_a))$, with $(1\leqslant i\leqslant m;1\leqslant a\leqslant s)$, both have maximal rank m. To obtain an explicit (local) description for $\mathcal{P}(X_H)$ we only need to select m independent rows from the matrix $(Z_i(\Phi_a))$. Without loss of generality, we may assume these to be the first m rows $(1\leqslant a\leqslant m)$, so that we recover (18).

Remark IV.1: Recently, various authors have pointed out that the dynamics of nonholonomic systems can be conveniently described in terms of a ''pseudo-Poisson'' bracket (see, e.g., Refs. 18, 29, 30). On the other hand, in Refs. 31, 25, a unified treatment of constrained systems has also been proposed in terms of Dirac brackets. The relation between these various bracket approaches has been discussed in Cantrijn *et al.*[32] It is rather straightforward to see that these bracket formulations of constrained dynamics can be extended to the general model for constrained systems considered in this paper, but we will not further enter into this matter here.

In the next three sections we wish to investigate the effect of symmetry on the dynamics of constrained systems of type (15). In particular, we will describe various reduction schemes for such systems. The subsequent analysis remains close in spirit to some related treatments of nonholonomic systems with symmetry (see, for instance, Refs. 33, 17, 15, 27, 34, 19, 35).

V. SYMMETRY AND REDUCTION

Consider a constrained system of the form (15) and let there be a given symplectic action $\Phi:G\times P\to P$ of a Lie group G on the symplectic manifold (P,ω), such that the submanifold M, the Hamiltonian function H and the vector subbundle F are G-invariant. For simplicity we will always assume that this action is free and proper. For each $g\in G$ and $x\in P$ we put $\Phi(g,x)=\Phi_g(x)=gx$. The infinitesimal generator (fundamental vector field) corresponding to $\xi\in\mathfrak{g}$, with $\mathfrak{g}$ the Lie algebra of G, will be denoted by ξ_P. By assumption we thus have for all $g\in G$,

- $\Phi_g^*(H)=H\circ\Phi_g=H$;
- $\Phi_g(M)\subseteq M$;
- $T\Phi_g(F_x)=F_{\Phi_g(x)}$, for all $x\in M$.

If (15) admits a solution X, it is routine to verify that Φ_g^*X will also be a solution for each $g\in G$. This still means that at each point $x\in M$, $\Phi_g^*X(x)-X(x)\in F_x^{\perp}\cap T_xM$. In particular, in case (15) has a unique solution, the latter will be G-invariant.

In discussing the reduction of a G-invariant solution of (15) we will proceed in two stages. First, we will show that the above assumptions already allow us to construct a Poisson reduction. Next, upon invoking an additional hypothesis, we will establish a kind of symplectic reduction, in the sense of the one derived by Bates and Śniatycki[17] for nonholonomic Hamiltonian systems.

(i) Poisson reduction. Since the action Φ is free and proper, the orbit space $\bar{P}=P/G$ is a differentiable manifold and $\rho:P\to\bar{P}$ is a principal bundle over $\bar{P}$ with structure group G, whereby ρ denotes the natural projection. Moreover, Φ being a symplectic action, it is, in particular, a Poisson action with respect to the natural Poisson structure induced by ω on P, i.e., it leaves the corresponding Poisson tensor field Λ on P invariant. It is known that the orbit space $\bar{P}$ then admits a unique Poisson structure such that the projection ρ becomes a Poisson map (see, e.g., Ref. 36). The corresponding Poisson tensor field $\bar{\Lambda}$ on $\bar{P}$ is unambiguously determined by

$$\bar{\Lambda}(d\bar{f},d\bar{g})(\bar{y})=\Lambda(\rho^*d\bar{f},\rho^*d\bar{g})(y),$$

for all $\overline{f},\overline{g} \in C^\infty(\overline{P})$ and $y \in \rho^{-1}(\overline{y})$. Let $\overline{\sharp}: T^*\overline{P} \to T\overline{P}$ be the linear bundle map induced by $\overline{\Lambda}$ according to

$$\langle \overline{\sharp}(\overline{\alpha}),\overline{\beta}\rangle = \overline{\Lambda}_{\overline{y}}(\overline{\alpha},\overline{\beta}),$$

for all $\overline{y} \in \overline{P}$ and $\overline{\alpha},\overline{\beta} \in T^*_{\overline{y}}\overline{P}$.

The Hamiltonian H being G-invariant, it induces a function $\overline{H}$ on $\overline{P}$. Moreover, M is also assumed to be G-invariant and, clearly, the G-action induced by Φ on M will still be free and proper. Thus, the quotient manifold $\overline{M} = M/G$ is a smooth submanifold of $\overline{P}$. Finally, we note that the G-invariance of F also implies the G-invariance of $F^\perp$. For each $\overline{x} \in \overline{M}$ we put $(\overline{F^\perp})_{\overline{x}} = T\rho(F_x^\perp)$ for some $x \in \rho^{-1}(\overline{x})(\subset M)$. This definition is independent of the choice of $x \in \rho^{-1}(\overline{x})$. We then put

$$\overline{F^\perp} = \cup_{\overline{x}\in\overline{M}}(\overline{F^\perp})_{\overline{x}},$$

which defines a generalized distribution on $\overline{P}$ along $\overline{M}$. In principle, the bundle $\overline{F^\perp}$ need not have constant rank. Assume now that there exists a G-invariant solution X of (15). As pointed out above, this will automatically be the case if the equation admits a unique solution. Then, X is projectable onto $\overline{M}$ and its projection $\overline{X}$ verifies

$$\overline{X} \in \overline{\sharp}(d\overline{H}) + \overline{F^\perp},$$

that is,

$$\overline{X} = X_{\overline{H}|\overline{M}} + \overline{Z},$$

for some $\overline{Z} \in \overline{F^\perp}$, with $X_{\overline{H}} = \overline{\sharp}(d\overline{H})$. Indeed, according to (16) we can always write X in the form $X = X_H + Z$, with $Z \in F^\perp$. The symmetry assumptions already guarantee the projectability of the Hamiltonian vector field X_H. Therefore, if X is G-invariant, Z is also G-invariant and its projection onto $\overline{M}$ is a section of $\overline{F^\perp}$.

Next, we will show that under an additional condition, the reduced dynamics $\overline{X}$ can be expressed in terms of a 2-form defined on a vector subbundle of $T\overline{P}_{|\overline{M}}$. The analysis closely follows the one developed in Ref. 17 (see also Ref. 34).

(ii) Bates–Śniatycki reduction. In what follows, we assume that there exists a G-invariant solution X of (15) such that $X \in F$. Recall that the latter assumption, in particular, implies that $X(H) = 0$.

Remark V.1: For the mechanical systems considered in Sec. II, the condition that the constrained dynamics should belong to the distribution F is not at all restrictive. Indeed, for (7) and (8) we have that every solution X automatically belongs to F since, in those cases, $TM \subset F$. In the case of (9) and (11), the property that $X \in F$ is a consequence of the fact that X is a SODE. Finally, for (14), the condition will be satisfied if the solution Y on $\overline{D}_f$ is the projection of a SODE along a submanifold of D_f. It is known that one can always find such a submanifold and such a solution.[25]

In the sequel, we will denote by $\mathcal{V}$ the subbundle of TP whose fibers are the tangent spaces to the G-orbits, i.e., $\mathcal{V}_x = T_x(Gx)$ or, equivalently, $\mathcal{V} = \ker T\rho$. Note that $\mathcal{V}_x \subset T_xM$ for all $x \in M$, i.e., $\mathcal{V}_{|M} \subset TM$. For simplicity, we will also usually write $\mathcal{V}$, instead of $\mathcal{V}_{|M}$, when referring to its restriction to M (the precise meaning should be clear from the context).

We now define a (generalized) vector subbundle U of $TP_{|M}$, whose fiber at $x \in M$ is given by

$$U_x = \{v \in F_x \cap T_xM / \omega(v,\tilde{\xi}) = 0, \quad \text{for all } \tilde{\xi} \in \mathcal{V}_x \cap F_x\}. \tag{19}$$

In general, this bundle need not be of constant rank, i.e., it determines a generalized distribution on P along M. In the sequel, however, we will always tacitly assume that U is a genuine vector bundle over M, although much of the analysis also holds in the more general situation. Note that $U=(F\cap TM)\cap(\mathcal{V}\cap F)^{\perp}$, where $(\mathcal{V}\cap F)^{\perp}$ is the ω-complement of $\mathcal{V}\cap F$ in $TP_{|M}$. It is readily seen that U is G-invariant and, hence, projects onto a subbundle $\bar{U}$ of $T\bar{P}_{|\bar{M}}$. Let us now denote by ω_U the restriction of ω to U. Clearly, ω_U is also G-invariant and since, moreover, $i_{\tilde{\xi}}\omega_U=0$ for all $\tilde{\xi}\in\mathcal{V}\cap U$, the 2-form ω_U pushes down to a 2-form $\omega_{\bar{U}}$ on $\bar{U}$ (i.e., $\omega_{\bar{U}}$ only acts on vectors belonging to $\bar{U}$). Similarly, the restriction of dH to U, denoted by d_UH, pushes down to a 1-form $d_{\bar{U}}\bar{H}$ on $\bar{U}$, which is simply the restriction of $d\bar{H}$ to $\bar{U}$. Note that neither $\omega_{\bar{U}}$ nor $d_{\bar{U}}\bar{H}$ are genuine differential forms on $\bar{M}$; they are exterior forms on a vector bundle over $\bar{M}$, with smooth dependence on the base point.

Proposition V.2: Let X be a G-invariant solution of (15) such that, in addition, X belongs to F. Then, the projection $\bar{X}$ of X onto $\bar{M}$ is a section of $\bar{U}$ satisfying the equation

$$i_{\bar{X}}\omega_{\bar{U}}=d_{\bar{U}}\bar{H}.$$

Proof: Essentially, all that remains to be checked is that X is a section of U. Along M, the given solution X verifies

$$i_X\omega=dH+\beta,$$

with $\beta\in F^o$. H being G-invariant, it follows that for any section $\tilde{\xi}$ of $\mathcal{V}\cap F$, $dH(\tilde{\xi})=0$. Since, obviously, we also have $\beta(\tilde{\xi})=0$, we may indeed conclude that $X\in U$. Consequently, the following relation holds along M:

$$i_X\omega_U=d_UH.$$

The remainder of the proof now readily follows from the symmetry assumptions and from the previous considerations. Q.E.D.

It is important to observe that, in general, the 2-form $\omega_{\bar{U}}$ may be degenerate. However, in the case of a mechanical system with linear nonholonomic constraints, for instance, one can prove that $\omega_{\bar{U}}$ is nondegenerate, such that $(\bar{U},\omega_{\bar{U}})$ becomes a symplectic vector bundle over $\bar{M}$ (see Ref. 17). The reduced dynamics is then uniquely determined by the equation mentioned in the previous Proposition.

In the next section, we will identify three distinguished classes of constrained systems with symmetry, which will be analyzed in some more detail.

VI. A CLASSIFICATION OF CONSTRAINED SYSTEMS WITH SYMMETRY

We again consider a constrained system (15) with symmetry, as described in the previous section. Recall that $\mathcal{V}=\ker T\rho$. For each infinitesimal generator ξ_P of the given group action on P, corresponding to some $\xi\in\mathfrak{g}$, the restriction to M is precisely the infinitesimal generator ξ_M of the induced action on M. If ξ_M is a section of $\mathcal{V}\cap F$, we will call it a **horizontal symmetry** of the given constrained system (see also Refs. 17, 15). The following classification, which is inspired on the one introduced by Bloch *et al.*[15] for mechanical systems with linear or affine nonholonomic constraints, reflects the various possible ways the subspaces $\mathcal{V}_x$ and F_x may intersect.

(i) *The purely kinematic case:* $\mathcal{V}_x\cap F_x=\{0\}$ and $T_xM=\mathcal{V}_x+(F_x\cap T_xM)$, for all $x\in M$.
(ii) *The case of horizontal symmetries:* $\mathcal{V}_x\cap F_x=\mathcal{V}_x$, for all $x\in M$, which is equivalent to $\mathcal{V}_x\subset F_x$, for all $x\in M$.
(iii) *The general case:* $\{0\}\subsetneq\mathcal{V}_x\cap F_x\subsetneq\mathcal{V}_x$, for all $x\in M$.

A. The purely kinematic case

Suppose that $\mathcal{V}_x \cap F_x = \{0\}$ and $T_x M = \mathcal{V}_x + (F_x \cap T_x M)$, for all $x \in M$. This implies that $T_x M = \mathcal{V}_x \oplus (F_x \cap T_x M)$. In other words, observing that in this case $U = F \cap TM$, we have $TM = \mathcal{V}_{|M} \oplus U$. Since U is G-invariant, this decomposition defines a principal connection Γ on the principal G-bundle $\rho_{|M}: M \to \overline{M}$, with horizontal subspace U_x at $x \in M$. Note, in passing, that U here represents a vector bundle of constant rank. In what follows we let X denote a fixed G-invariant solution of (15) which, moreover, belongs to F. In particular, this means that X is horizontal, i.e., $X \in U$.

Denote by $\mathbf{h}: TM \to U$ and $\mathbf{v}: TM \to \mathcal{V}$ the horizontal and vertical projectors associated with the decomposition $TM = \mathcal{V}_{|M} \oplus U$. The curvature of Γ is the tensor field of type (1,2) on M, given by

$$R = \tfrac{1}{2}[\mathbf{h}, \mathbf{h}],$$

where [,] denotes the Nijenhuis bracket of type (1,1) tensor fields. Taking into account that in the present case $\overline{U} = T\overline{M}$, and applying the method developed in Sec. V, we obtain on $\overline{M}$ a 2-form $\overline{\omega}$ (which is now a genuine differential form on $\overline{M}$) and a function $\overline{H}$ such that the projection $\overline{X}$ of X verifies

$$i_{\overline{X}} \overline{\omega} = d\overline{H}. \tag{20}$$

It should be pointed out that the reduced 2-form $\overline{\omega}$ in general need not be closed. We will show, however, that in case the given 2-form ω on P is exact, one can construct a reduced equation, equivalent to (20), but now in terms of a closed 2-form on $\overline{M}$.

Assume $\omega = d\theta$ for some 1-form θ on P. Denote by θ' the 1-form on M defined by $\theta' = j_M^* \theta$, where $j_M: M \hookrightarrow P$ is the canonical inclusion. By means of the given solution X of (15) we can construct a 1-form α_X on M as follows:

$$\alpha_X = i_X(\mathbf{h}^* d\theta' - d\mathbf{h}^* \theta'), \tag{21}$$

with the usual convention that, for an arbitrary p-form β, $\mathbf{h}^*\beta$ is the p-form defined by the prescription $\mathbf{h}^*\beta(X_1, \ldots, X_p) = \beta(\mathbf{h}(X_1), \ldots, \mathbf{h}(X_p))$.

Lemma VI.1: We have that

$$\alpha_X(Y) = \mathbf{v}(Y)(\theta'(X)) - \theta'(R(X,Y)) + \theta'(\mathbf{h}[X, \mathbf{v}(Y)]),$$

for all $Y \in \mathfrak{X}(M)$.

Proof: Indeed, for any $Y \in \mathfrak{X}(M)$ we easily find

$$\begin{aligned} \alpha_X(Y) &= i_X(\mathbf{h}^* d\theta' - d\mathbf{h}^*\theta')(Y) \\ &= \mathbf{h}X(\theta'(\mathbf{h}Y)) - \mathbf{h}Y(\theta'(\mathbf{h}X)) - \theta'[\mathbf{h}X, \mathbf{h}Y] - X(\theta'(\mathbf{h}Y)) + Y(\theta'(\mathbf{h}X)) + \theta'(\mathbf{h}[X,Y]) \\ &= \mathbf{v}Y(\theta'(X)) - \theta'(R(X,Y)) + \theta'(\mathbf{h}[X, \mathbf{v}Y]), \end{aligned}$$

taking into account that X is horizontal. Q.E.D.

Proposition VI.2: Assume, in addition, that the given action Φ *leaves* θ *invariant. Then, the 1-forms* $\mathbf{h}^*\theta'$ *and* α_X *are projectable. Moreover, the projection* $\overline{X}$ *of* X*, which is a solution of (20), also satisfies the equation*

$$i_{\overline{X}}\, d\overline{\theta'}_h = d\overline{H} - \overline{\alpha_X}, \tag{22}$$

where $\overline{\theta'}_h$ *and* $\overline{\alpha_X}$ *are the projections of the 1-forms* $\mathbf{h}^*\theta'$ *and* α_X*, respectively.*

Proof: We divide the proof in three parts: (i) the ρ-projectability of $\mathbf{h}^*\theta'$; (ii) the ρ-projectability of α_X; (iii) the derivation of the reduced equation of motion (22).

(i) Let ξ_M be the fundamental vector field on M induced by an arbitrary element $\xi \in \mathfrak{g}$. One can readily see that $i_{\xi_M}\mathbf{h}^*\theta' = 0$. We now show that also $i_{\xi_M}d(\mathbf{h}^*\theta') = 0$. Observe that for all $X' \in \mathfrak{X}(M)$ we have

$$i_{\xi_M}(d\mathbf{h}^*\theta')(X') = \xi_M(\theta'(\mathbf{h}X')) + \theta'(\mathbf{h}[\xi_M, X']).$$

Thus, for X' vertical, i.e., $X' \in \mathcal{V}$, we obtain $i_{\xi_M}(d\mathbf{h}^*\theta')(X') = 0$. Suppose now that X' is horizontal, i.e., $X' \in U$. Taking into account the G-invariance of θ' we deduce that

$$0 = \xi_M(\theta'(X')) - \theta'([\xi_M, X']).$$

Herewith we obtain

$$i_{\xi_M}(d\mathbf{h}^*\theta')(X') = \theta'([\xi_M, X'] - \mathbf{h}[\xi_M, X']) = \theta'(\mathbf{v}[\xi_M, X']) = 0,$$

since X' is horizontal.

Summarizing, we have shown that each fundamental vector field of the G-action on M is a characteristic vector field of $\mathbf{h}^*\theta'$ and, hence, the latter is a ρ-projectable 1-form.

(ii) To prove the projectability of α_X we first note that

$$i_{\xi_M}\alpha_X = i_{\xi_M} i_X(\mathbf{h}^*\, d\theta' - d\mathbf{h}^*\, \theta') = -i_{\xi_M} i_X(d\mathbf{h}^*\theta') = 0,$$

where the last equality follows by a similar argument as above, taking into account that the given X is horizontal.

Next, we prove that $i_{\xi_M} d\alpha_X = 0$. For this it suffices to show that $i_{\xi_M} d\alpha_X$ vanishes when acting on infinitesimal generators and on horizontal lifts of vector fields on $\bar{M}$. Using the previous property, i.e., $\alpha_X(\xi_M) = 0$, a straightforward calculation shows that for all $X' \in \mathfrak{X}(M)$:

$$(i_{\xi_M} d\alpha_X)(X') = \xi_M(\alpha_X(X')) - X'(\alpha_X(\xi_M)) - \alpha_X([\xi_M, X']) = \xi_M(\alpha_X(X')) - \alpha_X([\xi_M, X']).$$

From this we immediately deduce that if X' is a fundamental vector field of the group action, $i_{\xi_M}\, d\alpha_X(X') = 0$. On the other hand, if X' is the horizontal lift of a vector field Y on $\bar{M}$, i.e., $X' = Y^h$, we obtain, using Lemma VI.1 and the fact that the function $\theta'(R(X, Y^h))$ is G-invariant,

$$(i_{\xi_M}\, d\alpha_X)(Y^h) = \xi_M(\alpha_X(Y^h)) = -\xi_M(\theta'(R(X, Y^h))) = 0.$$

(iii) Recall that X satisfies an equation of the form $i_X\, d\theta = dH + \beta$, for some $\beta \in F^o$. Putting $H' = j_M^*(= H_{|M})$ and $\beta' = j_M^*\beta$, and taking into account that X is tangent to M, we can take the pull-back of this equation to M:

$$i_X\, d\theta' = dH' + \beta'.$$

Since X is horizontal, i.e., $\mathbf{h}X = X$, it follows that $\mathbf{h}^*(i_X\, d\theta') = i_X\mathbf{h}^* d\theta'$. Furthermore, H (and, hence H') being G-invariant, we have $\mathbf{h}^*\, dH' = dH'$ and, finally, it is also readily seen that $\mathbf{h}^*\beta' = 0$. The horizontal projection of the equation of motion on M therefore becomes

$$i_X\mathbf{h}^*\, d\theta' = dH'.$$

In view of the definition of the 1-form α_X, we then obtain

$$i_X\, d\mathbf{h}^*\theta' = i_X\mathbf{h}^*\, d\theta' - \alpha_X = dH' - \alpha_X.$$

All terms in this equation are projectable onto $\bar{M}$ and the reduced equation is indeed given by (22).

Q.E.D.

Proposition 6.2 describes a situation where a constrained Hamiltonian system (15) with symmetry, admits a reduction to an unconstrained system (22), but with an additional ''nonconservative force'' represented by $\overline{\alpha_X}$. It is interesting to observe that, by construction, the 1-form α_X satisfies

$$i_X \alpha_X = 0.$$

We now briefly comment on the problem of reconstructing the dynamics on M from the reduced dynamics on $\bar{M}$ in the case where (15) admits a unique solution X. Suppose the flow of the reduced system $\bar{X}$ is known. In order to recover flow of the constrained dynamics on M, one can first lift the integral curves of $\bar{X}$ to M by means of the horizontal lift operation associated with the principal connection Γ. The integral curves of X are then obtained by ''shifting'' these lifted curves along the fibres of $\rho_{|M}$. This second step can be implemented in the standard way.[37,38]

Finally, we can summarize the situation in the case of purely kinematic constraints in the following diagram:

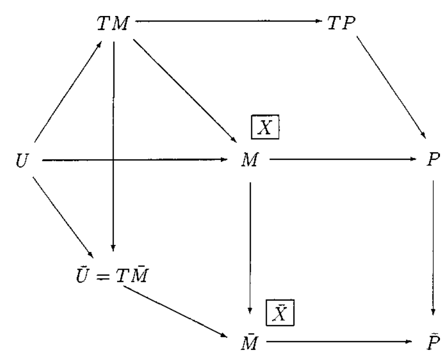

B. The case of horizontal symmetries

The assumption now is that $\mathcal{V}_x \cap F_x = \mathcal{V}_x$, for all $x \in M$ or, equivalently, $\mathcal{V}_{|M} \subset F$. In particular, every infinitesimal generator of the given group action then yields a horizontal symmetry as defined at the beginning of this section. Note also that an unconstrained Hamiltonian system with symmetry can be regarded as a special subcase of this case, since we then have $M = P$, $F = TP$ and, obviously, $\mathcal{V} \subset TP$.

For the further analysis of this case we assume, in addition, that the given symplectic action Φ on P is a Hamiltonian action, in the sense that it admits an Ad*-equivariant momentum map $J: P \to \mathfrak{g}^*$, such that for all $\xi \in \mathfrak{g}$, $i_{\xi_P}\omega = d\langle J, \xi \rangle$. Let $\mu \in \mathfrak{g}^*$ be a regular value of J, and suppose that the isotropy group G_μ acts freely and properly on the level set $J^{-1}(\mu)$. It is known (see Refs. 37, 36) that under these conditions $(P_\mu = J^{-1}(\mu)/G_\mu, \omega_\mu)$ is a symplectic manifold, where ω_μ is the 2-form defined by

$$\pi_\mu^* \omega_\mu = j_\mu^* \omega,$$

with $\pi_\mu : J^{-1}(\mu) \to P_\mu$ the canonical projection and $j_\mu : J^{-1}(\mu) \hookrightarrow P$ the natural inclusion.

With ξ_P again denoting the infinitesimal generator of the group action on P, corresponding to an element $\xi \in \mathfrak{g}$, it follows from the definition of the momentum mapping that $\xi_P = X_{J_\xi}$, where $J_\xi(x) = J(x)(\xi)$ for all $x \in P$. Taking into account that, by assumption, $\mathcal{V}_{|M} \subset F$, we find that for any solution X of (15), along the constraint submanifold M,

$$X(J_\xi) = 0,$$

i.e., the components of the momentum mapping are conserved quantities for the constrained dynamics. This is a version of Noether's theorem for constrained systems. (For the case of mechanical systems with nonholonomic constraints, see in this respect also Refs. 15, 39, 35.)

Imposing a condition of clean intersection of M and $J^{-1}(\mu)$, we have that $M' = M \cap J^{-1}(\mu)$ is a submanifold of $J^{-1}(\mu)$ which is G_μ-invariant. Passing to the quotient we then obtain a submanifold $M_\mu = M'/G_\mu$ of P_μ. Next, we can define a distribution F' on P along M' by putting

$$F'_{x'} = T_{x'}(J^{-1}(\mu)) \cap F_{x'}, \quad \forall x' \in M',$$

and we now make the further simplifying assumption that F' has constant rank. It is obvious that F' is a G_μ-invariant subbundle of $TP_{|M'}$ and, hence, it projects onto a subbundle F_μ of TP_μ along M_μ. Finally, since the restriction of the Hamiltonian H to $J^{-1}(\mu)$ is also G_μ-invariant, it induces a function H_μ on P_μ.

Theorem VI.3: *Suppose that X is a G-invariant solution of (15). Then, X induces a vector field X_μ on M_μ, such that*

$$(i_{X_\mu}\omega_\mu - dH_\mu)_{|M_\mu} \in F^o_\mu, \quad X_\mu \in TM_\mu. \tag{23}$$

Proof: First of all, notice that $X' = X_{|M'}$ is everywhere tangent to M', since both $J^{-1}(\mu)$ and M are invariant submanifolds of X. Pulling back (15) to $J^{-1}(\mu)$, we find that X' satisfies an equation of the form

$$(i_{X'} j^*_\mu \omega - d(H \circ j_\mu))_{|M'} = \beta$$

for some section β of F'^o. Since X is G-invariant, and taking account the other symmetry assumptions, it follows that both X' and β are G_μ-equivariant sections of TM' and F'^o, respectively. Moreover, from the fact that we are dealing with horizontal symmetries we may deduce, in particular, that for all $\xi \in \mathfrak{g}_\mu$ (= the Lie algebra of G_μ), $(\xi_P)_{|M'}$ is a section of F'. Therefore, β projects onto a section of F^o_μ. Using a standard argument, it now readily follows that X' projects onto a vector field on M_μ for which (23) holds. Q.E.D.

In the case of horizontal symmetries we have thus proved that, under the appropriate assumptions, the given constrained problem on (P,ω) reduces to a constrained problem on (P_μ, ω_μ).

As far as the reconstruction of the original constrained dynamics from the reduced dynamics is concerned, we observe that, unlike in the purely kinematic case, we now first have to select an arbitrary connection on the principal G_μ-bundle $M' \to M_\mu$. This connection will enable us to subsequently lift the integral curves of the reduced system from M_μ to M'. The reconstruction of the flow of X then further proceeds as in the previous case.

The following diagram illustrates the situation in the case of horizontal symmetries. Note in passing that, modulo the appropriate embeddings, one may identify M_μ with $\bar{M} \cap P_\mu$ where, as before, $\bar{M} = M/G$.

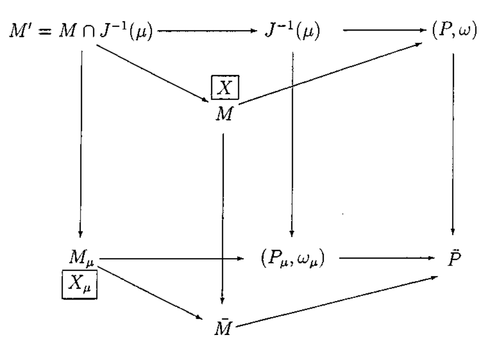

C. The general case

We now consider the case where, at $x\in M$, $\{0\}\neq\mathcal{V}_x\cap F_x\neq\mathcal{V}_x$. Assuming again that the given action of G on P is Hamiltonian, with momentum map J, it is no longer true that J is a conserved quantity for the constrained dynamics. However, extending a procedure developed by Bloch *et al.*[15] for nonholonomic mechanical systems (see also Ref. 40), we will derive an equation which describes the evolution of some components of the momentum map along the integral curves of the constrained system.

For each $x\in M$, we put

$$\mathfrak{g}^x=\{\xi\in\mathfrak{g}\,|\,\xi_M(x)\in F_x\},$$

and

$$S^x=\{\xi_M(x)\,|\,\xi\in\mathfrak{g}^x\},$$

i.e., $S^x=\mathcal{V}_x\cap F_x$. Recall that ξ_M is just the restriction of ξ_P to the G-invariant submanifold M. We have that $\mathfrak{g}^x$ and S^x are vector subspaces of $\mathfrak{g}$ and T_xM $(\subset T_xP)$, respectively. Putting

$$\mathfrak{g}^F=\coprod_{x\in M}\mathfrak{g}^x,\quad S^F=\coprod_{x\in M}S^x,$$

where we use the symbol ''$\coprod$'' to denote the disjoint union of the respective vector spaces, we obtain two (''generalized'') vector bundles over M, with corresponding natural projections $\mathfrak{g}^F\to M:\xi\in\mathfrak{g}^x\mapsto x$ and $S^F\to M:\xi_M(x)\mapsto x$. In general, these bundles need not have constant rank. However, for the subsequent discussion we make the simplifying assumption that $\mathfrak{g}^F$ and S^F are genuine vector bundles over M, the fibers of which have constant dimension (independent of the base point). The given action being a free action, the mapping $\mathfrak{g}^F\to S^F:\xi\in\mathfrak{g}^x\mapsto\xi_M(x)$ then defines a smooth vector bundle isomorphism.

Suppose now that the symplectic form ω is exact, say $\omega=d\theta$, and that the G-action leaves θ invariant. In such a case there always exists a well-defined momentum mapping $J:P\to\mathfrak{g}^*$ such that

$$\langle J(x),\xi\rangle=-(\theta_x)(\xi_P(x)),\quad\forall x\in P,\quad\forall\xi\in\mathfrak{g}$$

(see, e.g., Ref. 37). Herewith we can define a smooth section $J^{(c)}:M\to(\mathfrak{g}^F)^*$ of the dual bundle $(\mathfrak{g}^F)^*$ as follows:

$$J^{(c)}(x):\mathfrak{g}^x\to\mathbb{R},\quad J^{(c)}(x)(\xi)=\langle J(x),\xi\rangle.$$

We may call $J^{(c)}$ the ''constrained momentum map.'' In Ref. 15, which deals with nonholonomic mechanical systems, this map was denoted by J^{nhc}. Given a smooth section $\overline{\xi}$ of the vector bundle $\mathfrak{g}^F$, we can then define a smooth function $J^{(c)}_{\overline{\xi}}$ on M according to

$$J^{(c)}_{\overline{\xi}}=\langle J^{(c)},\overline{\xi}\rangle.$$

In addition, we can construct a vector field Ξ on M by putting

$$\Xi(x)=(\overline{\xi}(x))_M(x),\quad\forall x\in M.$$

Denoting the Lie derivative operator with respect to Ξ as $\mathcal{L}_\Xi$, we have the following interesting result.

Theorem VI.4: *Let X be an arbitrary solution of (15). For any smooth section $\overline{\xi}$ of $\mathfrak{g}^F$ we then have*

$$X(J_{\xi}^{(c)}) = -(\mathcal{L}_{\Xi}\theta)(X). \tag{24}$$

Proof: Since Ξ takes values in F, it follows from (15) that, along M,

$$i_{\Xi} i_X \omega - i_{\Xi}\, dH = 0.$$

From the above definitions we further deduce that $J_{\xi}^{(c)} = -i_{\Xi}(j_M^* \theta)$, with $j_M : M \hookrightarrow P$ again denoting the inclusion map. A straightforward computation then gives

$$X(J_{\xi}^{(c)}) = -i_X\, d i_{\Xi}(j_M^*\theta) = -i_X \mathcal{L}_{\Xi}(j_M^*\theta) + i_X i_{\Xi}(j_M^*\omega) = -\mathcal{L}_{\Xi} i_X(j_M^*\theta) + i_{[\Xi,X]}(j_M^*\theta) - \Xi(H \circ j_M).$$

Since H is G-invariant, it follows from the definition of Ξ that $\Xi(H \circ j_M) = 0$. Herewith, the previous relation immediately reduces to (24) (with a slight abuse of notation). Q.E.D.

Note that for the above result we do not have to require X to be G-invariant. Equation (24) is called *the momentum equation* for the given constrained system. In the case of linear nonholonomic constraints we precisely recover the result established by Bloch *et al.*[15]

Suppose again that X is a solution of (15) and let $\overline{\xi}$ be a constant section of $\mathfrak{g}^F$, i.e., $\overline{\xi}(x) = \xi^0 \in \mathfrak{g}$ for all $x \in M$. We may then identify the corresponding vector field Ξ with the infinitesimal generator ξ_M^0 and, clearly, $J_{\xi}^{(c)} = (J_{\xi^0})_{|M}$. Moreover, by construction, ξ_M^0 is a horizontal symmetry. The momentum equation (24) then leads to

$$X(J_{\xi}^{(c)}) = X(J_{\xi^0})_{|M} = 0,$$

i.e., we have obtained a conserved quantity of X associated with the horizontal symmetry ξ_M^0. This is again a manifestation of Noether's theorem for constrained systems (cf. the previous subsection).

In the next section we will apply some of the previous results to the case of a singular Lagrangian system and to a Lagrangian system with linear nonholonomic constraints induced by a principal connection.

VII. APPLICATIONS

A. Singular Lagrangian systems

Consider a system described by a singular Lagrangian function $L: TQ \to \mathbb{R}$ such that ω_L is presymplectic. We assume that a Lie group G acts freely and properly on the configuration manifold Q and that L is invariant under the lifted action of G on TQ. It then easily follows that both ω_L and E_L are also G-invariant. In addition, we know that the lifted action of G on T^*Q leaves invariant the Liouville 1-form θ_Q and, hence, also the canonical symplectic form $\omega_Q = -d\theta_Q$ (see, e.g., Ref. 37). From all this, one can subsequently deduce that the Legendre mapping is G-equivariant and that the constraint submanifolds generated by the presymplectic constraint algorithm, both on the Lagrangian and on the Hamiltonian side, are G-invariant. In particular, the final constraint submanifold M_f in T^*Q is G-invariant.

Let us now consider the constrained equations of motion (7) where, for simplicity, we write H instead of H_1, i.e.,

$$(i_X \omega_Q - dH)_{|M_f} \in TM_1^o, \quad X_{|M_f} \in TM_f.$$

Since M_f is G-invariant, it follows that $\mathcal{V}_{|M_f} \subset TM_f \subseteq TM_1 = F$ and, hence, we are in the case of horizontal symmetries. Moreover, the lifted symplectic action of G on T^*Q admits an equivariant momentum map J and so we can apply the reduction procedure described in subsection IV B. Given a regular value μ of J, it is easy to check that the reduced system then becomes

$$(i_{X_\mu}\omega_\mu - dH_\mu)_{|(M_f)_\mu} \in (T(M_1)_\mu)^o, \quad X_\mu \in T(M_f)_\mu,$$

whereby we observe that

$$(T(M_1)_\mu)^o \subset (T(M_f)_\mu)^o.$$

Suppose, on the other hand, we would have started from the description of the given constrained system in terms of (8), again denoting the extended Hamiltonian by H, i.e.,

$$(i_X\omega_Q - dH)_{|M_f} \in TM_f^o, \quad X_{|M_f} \in TM_f. \tag{25}$$

Under the given assumptions, the final constraint submanifold M_f, generated through Hinds' algorithm, will also be G-invariant such that $\mathcal{V}_{|M_f} \subset F = TM_f$, i.e., we are again in the case of horizontal symmetries. Given a regular value μ of the momentum map J, it is easy to check that we now have

$$(TM_f)_\mu^o = (T(M_f)_\mu)^o,$$

where, assuming clean intersection of M_f and $J^{-1}(\mu)$, $(M_f)_\mu = (M_f \cap J^{-1}(\mu))/G_\mu$ and $(TM_f)_\mu = (TJ^{-1}(\mu) \cap TM_f)/G_\mu$. If (25) admits a G-invariant solution X, it follows from Theorem 6.3 that the reduced dynamics will satisfy the constrained system

$$(i_{X_\mu}\omega_\mu - dH_\mu)_{|(M_f)_\mu} \in (T(M_f)_\mu)^o, \quad X_\mu \in T(M_f)_\mu. \tag{26}$$

We now have the following diagram:

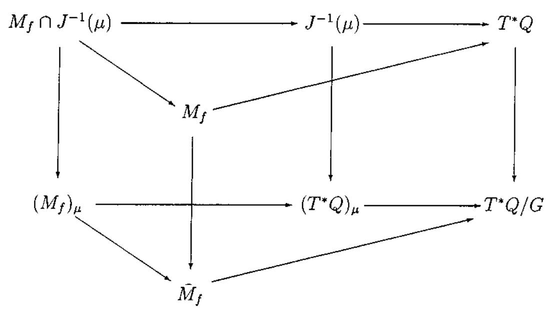

Notice that, according to Proposition 3.1, the reduced system (26) admits a unique solution if and only if $T(M_f)_\mu^\perp \cap T(M_f)_\mu = 0$, which implies that $(M_f)_\mu$ is a symplectic submanifold of $(T^*Q)_\mu$. In that case we have the direct sum decomposition

$$T((T^*Q)_\mu)_{|(M_f)_\mu} = T(M_f)_\mu \oplus T(M_f)_\mu^\perp$$

and we can construct the unique solution of (26) in the following way. Let X_{H_μ} denote the Hamiltonian vector field on $((T^*Q)_\mu, \omega_\mu)$, corresponding to H_μ. The reduction X_μ of X is then obtained by first taking the restriction of X_{H_μ} to $(M_f)_\mu$, and then projecting it onto $T(M_f)_\mu$.

Example: Consider the singular Lagrangian function $L: T\mathbb{R}^6 \to \mathbb{R}$ given by

$$L = m_2(\dot{x}_2^2 + \dot{y}_2^2) + m_3(\dot{x}_3^2 + \dot{y}_3^2) + \dot{y}_2 x_2 - \dot{x}_2 y_2 + \dot{y}_3 x_3 - \dot{x}_3 y_3 - x_1^2 - y_1^2 - x_2^2 - y_2^2 - x_3^2 - y_3^2,$$

with coordinates $(x_1, y_1, x_2, y_2, x_3, y_3, \dot{x}_1, \dot{y}_1, \dot{x}_2, \dot{y}_2, \dot{x}_3, \dot{y}_3)$ on $T\mathbb{R}^6$. Here m_2 and m_3 are constants. The above Lagrangian is a particular case of those considered by Capri and Kobayashi[41,42] (see also Ref. 43). This type of Lagrangian occurs in some models of field theories coupled to external fields.

When passing to the Hamiltonian side, we obtain the following two primary constraints: $\phi_1 = p_{x_1} = 0$ and $\phi_2 = p_{y_1} = 0$ which determine the constraint submanifold M_1. The 2-form ω_1 is given in local coordinates $(x^1, y^1, x^2, y^2, x^3, y^3, p_{x_2}, p_{y_2}, p_{x_3}, p_{y_3})$ on M_1 by

$$\omega_1 = dx_2 \wedge dp_{x^2} + dy_2 \wedge dp_{y^2} + dx_3 \wedge dp_{x^3} + dy_3 \wedge dp_{y^3},$$

which is presymplectic, with

$$\ker \omega_1 = \left\langle \frac{\partial}{\partial x_1}, \frac{\partial}{\partial y_1} \right\rangle.$$

The energy function E_L projects onto the function h_1 on M_1 given by

$$h_1 = \frac{1}{4m_2}((p_{x_2}+y_2)^2 + (p_{y_2}-x_2)^2) + \frac{1}{4m_3}((p_{x_3}+y_3)^2 + (p_{y_3}-x_3)^2) + (x_1)^2 + (y_1)^2 + (x_2)^2 + (y_2)^2 + (x_3)^2 + (y_3)^2.$$

Consistency of the constraints ϕ_1 and ϕ_2 leads to the secondary constraints

$$\phi_3 = x_1 = 0, \quad \phi_4 = y_1 = 0,$$

and the constraint submanifold M_2 determined by the vanishing of the constraints ϕ_i, $1 \leqslant i \leqslant 4$, turns out to be the final constraint submanifold, i.e., $M_2 = M_f$. We note, in passing, that in this case the final constraint submanifolds generated by the Gotay–Nester algorithm and the Hinds algorithm, coincide.

Consider the function H on T^*Q with coordinate expression equal to that of h_1. (In fact, one might take any arbitrary extension of h_1 which coincides with the latter on M_f; for instance: $\tilde{H} = H + \lambda_1 \phi_1 + \lambda_2 \phi_2$ for some arbitrary functions λ_i). Note also that

$$TM_f^{\perp} = \left\langle \frac{\partial}{\partial x_1}, \frac{\partial}{\partial y_1}, \frac{\partial}{\partial p_{x_1}}, \frac{\partial}{\partial p_{y_1}} \right\rangle.$$

The constrained equations of motion (8), i.e.,

$$i_X \omega_Q - dH \in TM_f^o, \quad X \in TM_f, \tag{27}$$

admit a unique solution since $TM_f \cap TM_f^{\perp} = 0$.

The initial system admits nongauge symmetries which are rotations on the configuration space. The action

$$\Phi : \mathbb{T}^2 \times \mathbb{R}^6 \to \mathbb{R}^6,$$

where $\mathbb{T}^2$ is the two-dimensional torus, is given by

$$\Phi((\theta_2,\theta_3)\times(x^1,y^1,x^2,y^2,x^3,y^3)) = (x_1, y_1, x_2 \sin\theta_2 + y_2 \cos\theta_2, x_2 \cos\theta_2 - y_2 \sin\theta_2, x_3 \sin\theta_3 + y_3 \cos\theta_3, x_3 \cos\theta_3 - y_3 \sin\theta_3).$$

The infinitesimal generators of this action are

$$\left\langle x_i \frac{\partial}{\partial y_i} - y_i \frac{\partial}{\partial x_i} \right\rangle, \quad 2 \leqslant i \leqslant 3.$$

The infinitesimal generators of the lifted action to T^*Q are

$$\left\langle x_i \frac{\partial}{\partial y_i} - y_i \frac{\partial}{\partial x_i} + p_{x_i} \frac{\partial}{\partial p_{y_i}} - p_{y_i} \frac{\partial}{\partial p_{x_i}} \right\rangle, \quad 2 \leqslant i \leqslant 3.$$

This lifted action is symplectic and preserves the Hamiltonian H. Moreover, the manifold M_f is also $\mathbb{T}^2$-invariant.

According to our classification of constrained systems, we are in the case of horizontal symmetries, since at each point $x \in M_f$,

$$\mathcal{V}_x \cap (F_x \cap T_x M_f) = \mathcal{V}_x,$$

because $F = TM_f$.

Consider the equivariant momentum map $J: T^*Q \to \mathbb{R}^2$ given by

$$J(x_1, y_1, x_2, y_2, x_3, y_3; p_{x_1}, p_{y_1}, p_{x_2}, p_{y_2}, p_{x_3}, p_{y_3}) = (x_2 p_{y_2} - y_2 p_{x_2}, x_3 p_{y_3} - y_3 p_{x_3}).$$

For any regular value $\mu = (\mu_1, \mu_2)$, applying the cotangent bundle reduction, we have that $J^{-1}(\mu)/(\mathbb{T}^2)_\mu$ is a differentiable manifold equipped with a symplectic 2-form ω_μ. Denote by H_μ the projection of $H_{|J^{-1}(\mu)}$, and by $(M_f)_\mu$ the projection of $M_f \cap J^{-1}(\mu)$, with projection map $\pi_\mu : J^{-1}\mu \to J^{-1}(\mu)/(\mathbb{T}^2)_\mu$. Then, from Theorem 6.3 it follows that the solution of system (27) projects onto the solution of the system

$$(i_{X_\mu}\omega_\mu - dH_\mu)_{|(M_f)_\mu} \in (T(M_f)_\mu)^o, \quad X_\mu \in T(M_f)_\mu.$$

By taking polar coordinates on M_f, i.e., $(r_2, \varphi_2, r_3, \varphi_3; p_{r_2}, p_{\varphi_2}, p_{r_3}, p_{\varphi_3})$, we have that $J^{-1}(\mu) \cap M_f$ is the $(\mathbb{T}^2)_\mu$-invariant submanifold of M_f determined by

$$p_{\varphi_2} = \mu_1 \quad \text{and} \quad p_{\varphi_3} = \mu_2.$$

Passing to the quotient we find that $(M_f)_\mu$ is a four-dimensional submanifold of $J^{-1}(\mu)/(\mathbb{T}^2)_\mu$, with induced coordinates $(r_2, r_3; p_{r_2}, p_{r_3})$ and equipped with the symplectic form

$$(\omega_{M_f})_\mu = dr_2 \wedge dp_{r_2} + dr_3 \wedge dp_{r_3}.$$

B. Nonholonomic Lagrangian systems

We again consider an action of a Lie group G on a manifold Q, and let $L: TQ \to \mathbb{R}$ be a regular Lagrangian which is G-invariant. The lifted action of G on the symplectic manifold (TQ, ω_L) is then Hamiltonian. We assume that the Lagrangian system is subjected to some linear nonholonomic constraints, described by a distribution $\mathcal{D}$ on Q, such that the resulting nonholonomic system verifies the compatibility condition (cf. Sec. II B) and such that, in addition, the vector subbundle D of TQ, spanned by $\mathcal{D}$, is G-invariant. The constrained equations then read (cf. (9))

$$i_X \omega_L - dE_L \in (\mathcal{D}^v)^o, \quad X_{|D} \in TD. \tag{28}$$

We now consider an interesting special subcase of the purely kinematic case, namely, a (generalized) *Čaplygin system*. For a system of Čaplygin type, the configuration manifold Q is a principal G-bundle $\pi: Q \to Q/G$, and the constraints are given by the horizontal subspaces of a principal connection Γ on π (see Refs. 12, 24).

Under the above conditions, one can easily see that there exists a well-defined Lagrangian function $L^*: T(Q/G) \to \mathbb{R}$, given by

$$L^*(Y) = L((Y^h)_q),$$

for any $Y \in T_y(Q/G)$, where $q \in Q$ is an arbitrary point in the fiber over $y \in Q/G$ and Y^h denotes the horizontal lift of Y with respect to Γ.

A direct computation shows that, with the notations introduced in Sec. V, $\mathcal{V} \cap \mathcal{D}^v = 0$. Moreover, we have $U = \mathcal{D}^v \cap TD$, and U is symplectic with respect to ω_L. Therefore we deduce that

$$TD=\mathcal{V}\oplus U.$$

Thus, a Čaplygin system fits indeed very nice in the purely kinematic case. Moreover, one can prove that $\overline{D}=D/G\cong T(Q/G)$ and $\overline{E}_L=E_{L*}$.

We have seen that the compatibility condition,

$$(\mathcal{D}^v)^{\perp}\cap TD=0,$$

ensures the existence of a unique solution $X=\Gamma_{L,D}$ of (28) which, moreover, is a SODE. Notice that $\Gamma_{L,D}$ can be obtained by projecting the unconstrained Euler–Lagrange vector field Γ_L by means of the first projector associated with the decomposition,

$$T(TQ)_{|D}=TD\oplus(\mathcal{D}^v)^{\perp}.$$

Since $\omega_L=-d\theta_L$, the reduced equation becomes

$$i_{\overline{X}}\omega_{L*}=dE_{L*}-\overline{\alpha_{\Gamma_{L,D}}},$$

where $\overline{\alpha_{\Gamma_{L,D}}}$ is the projection of the 1-form $\alpha_{\Gamma_{L,D}}$, defined by (21). Observe that

$$i_{\overline{\Gamma}}\overline{\alpha_{\Gamma_{L,D}}}=0,$$

for any SODE $\overline{\Gamma}$ on $T(Q/G)$. This implies that $\overline{\alpha_{\Gamma_{L,D}}}$ is a 1-form of gyroscopic type.

Example VII.1: The vertical rolling disk. Consider a rolling disk of radius R constrained to remain vertical on a horizontal plane. The standard coordinates of the configuration space $\mathbb{R}\times S^1\times S^1$ are: the Cartesian coordinates x,y of the center of mass, the angle θ_1 between the tangent of the disk at the point of contact and the x-axis and the angle θ_2 determined by some diameter of the disk and the vertical.

The dynamics of this mechanical system is described by the following:

(i) the regular Lagrangian:

$$L=\tfrac{1}{2}(m\dot{x}^2+m\dot{y}^2+I_1\dot{\theta}_1^2+I_2\dot{\theta}_2^2),$$

where m is the mass, and I_1,I_2 are moments of inertia;

(ii) the nonholonomic constraints:

$$\phi_1=\dot{x}-(R\cos\theta_1)\dot{\theta}_2=0,\quad \phi_2=\dot{y}-(R\sin\theta_1)\dot{\theta}_2=0.$$

The Poincaré–Cartan 2-form of the Lagrangian L is

$$\omega_L=m\ dx\wedge d\dot{x}+m\ dy\wedge d\dot{y}+I_1\ d\theta_1\wedge d\dot{\theta}_1+I_2\ d\theta_2\wedge d\dot{\theta}_2,$$

so that the Euler–Lagrange vector field of the free (i.e., unconstrained system) is

$$\Gamma_L=\dot{x}\frac{\partial}{\partial x}+\dot{y}\frac{\partial}{\partial y}+\dot{\theta}_1\frac{\partial}{\partial\theta_1}+\dot{\theta}_2\frac{\partial}{\partial\theta_2}.$$

Consider the group $G=\mathbb{R}^2$ and its trivial action by translations on Q:

$$\Phi:G\times Q\to Q$$

$$(r,s)\times(x,y,\theta_1,\theta_2)\mapsto(x+r,y+s,\theta_1,\theta_2).$$

If we consider the lifted action Φ^1 of Φ to TQ, given by $(\Phi^1)_g=T\Phi_g$, then the infinitesimal generators of this action are

$$\mathcal{V} = \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right\rangle.$$

One easily verifies that the constraint submanifold D, determined by ϕ_1, ϕ_2, is invariant with respect to Φ^1. Choose local coordinates $(x, y, \theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2)$ on D. In these coordinates we find that the distribution U on D is generated by the vector fields:

$$U = \left\langle \frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, R \cos \theta_1 \frac{\partial}{\partial x} + R \sin \theta_1 \frac{\partial}{\partial y} + \frac{\partial}{\partial \dot{\theta}_2}, \frac{\partial}{\partial \dot{\theta}_1} \right\rangle,$$

and we readily have that $\mathcal{V}_{|D} \subset TD$ and $\mathcal{V}_{|D} \oplus U = TD$, i.e., we are in the purely kinematic case. In fact, noting that $\rho: Q \to S^1 \times S^1$ is a principal bundle, with structure group $G = \mathbb{R}^2$, and D is the horizontal subbundle of a principal connection, we see that the given system is a Čaplygin system. Following the above analysis we then obtain

$$L^* = \tfrac{1}{2}(I_1 \dot{\theta}_1^2 + (mR^2 + I_2)\dot{\theta}_2^2),$$

$$\omega_{L^*} = I_1 \, d\theta_1 \wedge d\dot{\theta}_1 + (mR^2 + I_2) d\theta_2 \wedge d\dot{\theta}_2.$$

In this particular case the gyroscopic 1-form $\overline{\alpha_{\Gamma_{L,D}}} = 0$ and then the 2-form ω_U^- is closed and

$$\omega_U^- = \omega_{L^*}.$$

Example VII.2: The two-wheeled carriage. The configuration space of the two-wheeled carriage is $Q = \mathbb{R}^2 \times S^1 \times \mathbb{T}^2$ with coordinates $(x, y, \varphi, \Psi_1, \Psi_2)$ (see, e.g., Ref. 24 for more details).

This system is determined by the following data:

(i) A regular Lagrangian L,

$$L = \tfrac{1}{2} m(\dot{x}^2 + \dot{y}^2) + m_0 l \dot{\varphi}(\dot{y} \cos \varphi - \dot{x} \sin \varphi) + \tfrac{1}{2} I \dot{\varphi}^2 + \tfrac{1}{2} C(\dot{\Psi}_1^2 + \dot{\Psi}_2^2);$$

(i) and the nonholonomic constraints,

$$\phi_1 = \dot{x} + \frac{a \cos \varphi}{2} \dot{\Psi}_1 + \frac{a \cos \varphi}{2} \dot{\Psi}_2,$$

$$\phi_2 = \dot{y} + \frac{a \sin \varphi}{2} \dot{\Psi}_1 + \frac{a \sin \varphi}{2} \dot{\Psi}_2,$$

$$\phi_3 = \dot{\varphi} + \frac{a}{2r} \dot{\Psi}_1 - \frac{a}{2r} \dot{\Psi}_2.$$

These constraints are linear in the velocities and determine a distribution $\mathcal{D}$ on Q whose annihilator is generated by the 1-forms,

$$\mu_1 = dx + \frac{a \cos \varphi}{2} d\Psi_1 + \frac{a \cos \varphi}{2} d\Psi_2,$$

$$\mu_2 = dy + \frac{a \sin \varphi}{2} d\Psi_1 + \frac{a \sin \varphi}{2} d\Psi_2,$$

$$\phi_3 = d\varphi + \frac{a}{2r} d\Psi_1 - \frac{a}{2r} d\Psi_2.$$

Consider the group of Euclidean motions in the plane, $G = \mathbb{R}^2 \times S^1$, with its standard action on Q:

$$\Phi : G \times Q \to Q,$$

$$(r,s,\theta)\times(x,y,\varphi,\Psi_1,\Psi_2)\mapsto(r+x\cos\theta-y\sin\theta, s+x\sin\theta+y\cos\theta, \theta+\varphi,\Psi_1,\Psi_2),$$

whose infinitesimal generators are

$$\left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, -y\frac{\partial}{\partial x}+x\frac{\partial}{\partial y}+\frac{\partial}{\partial\varphi}\right\rangle.$$

Observe that the distribution $\mathcal{D}$ is G-invariant and, hence, also the constraint submanifold D is preserved, that is, for all $x\in D$ and $g\in G$, $T\Phi_g(x)\in D$.

Taking $(x,y,\varphi,\Psi_1,\Psi_2,\dot\Psi_1,\dot\Psi_2)$ as coordinates on D, we have that the pull-back ω_D of the Poincaré–Cartan 2-form ω_L to D is

$$\begin{aligned}
\omega_D=&\left(-\frac{am\cos\varphi}{2}+\frac{am\sin\varphi}{2}\dot\Psi_1+\frac{am_0 l\sin\varphi}{2r}\right)dx\wedge d\dot\Psi_1\\
&+\left(-\frac{am\cos\varphi}{2}+\frac{am\sin\varphi}{2}\dot\Psi_2-\frac{am_0 l\sin\varphi}{2r}\right)dx\wedge d\dot\Psi_2\\
&+\left(-\frac{am\sin\varphi}{2}-\frac{am\cos\varphi}{2}\dot\Psi_2-\frac{am_0 l\cos\varphi}{2r}\right)dy\wedge d\dot\Psi_1\\
&+\left(-\frac{am\sin\varphi}{2}-\frac{am\cos\varphi}{2}\dot\Psi_2+\frac{am_0 l\cos\varphi}{2r}\right)dy\wedge d\dot\Psi_2\\
&+\frac{am_0 l\sin\varphi}{2r}(\dot\Psi_1-\dot\Psi_2)dy\wedge d\varphi+\frac{am_0 l\cos\varphi}{2r}(\dot\Psi_1-\dot\Psi_2)dx\wedge d\varphi\\
&-\frac{aI}{2r}d\varphi\wedge d\dot\Psi_1+\frac{aI}{2r}d\varphi\wedge d\dot\Psi_2+Cd\Psi_1\wedge d\dot\Psi_1+Cd\Psi_2\wedge d\dot\Psi_2.
\end{aligned}$$

A basis of $U=\mathcal{D}^v\cap TD$ is given by the vectors fields

$$\begin{aligned}
&\left\langle\frac{\partial}{\partial\dot\Psi_1},\frac{\partial}{\partial\dot\Psi_2},\frac{\partial}{\partial\Psi_1}-\frac{a}{2}\cos\varphi\frac{\partial}{\partial x}-\frac{a}{2}\sin\varphi\frac{\partial}{\partial y}-\frac{a}{2r}\frac{\partial}{\partial\varphi},\right.\\
&\left.\frac{\partial}{\partial\Psi_2}-\frac{a}{2}\cos\varphi\frac{\partial}{\partial x}-\frac{a}{2}\sin\varphi\frac{\partial}{\partial y}+\frac{a}{2r}\frac{\partial}{\partial\varphi}\right\rangle.
\end{aligned}$$

Observe again that D/G can be identified with the space $T(Q/G)$. The projected 2-form $\omega_{\bar U}$ is

$$\begin{aligned}
\omega_{\bar U}=&-\frac{a^3}{4r^2}m_0 l(\dot\Psi_1-\dot\Psi_2)d\Psi_1\wedge d\Psi_2+\left(\frac{ma^2}{4}+\frac{Ia^2}{4r^2}+C\right)d\Psi_1\wedge d\dot\Psi_1\\
&+\left(\frac{ma^2}{4}+\frac{Ia^2}{4r^2}+C\right)d\Psi_2\wedge d\dot\Psi_2+\left(\frac{ma^2}{4}-\frac{Ia^2}{4r^2}\right)d\Psi_1\wedge d\dot\Psi_2\\
&+\left(\frac{ma^2}{4}-\frac{Ia^2}{4r^2}\right)d\Psi_2\wedge d\dot\Psi_1.
\end{aligned}$$

This 2-form $\omega_{\bar U}$ is an almost symplectic form, that is, it is nondegenerate but not closed. The solution of the dynamics after the reduction procedure is given by

$$i_{\bar{X}}\omega_{\bar{U}} = dE_{L*},$$

where $L^*:T(\mathbb{T}^2)\rightarrow\mathbb{R}$ is defined by

$$L^*(\Psi_1,\Psi_2,\dot{\Psi}_1,\dot{\Psi}_2)=\frac{1}{8}ma^2(\dot{\Psi}_1+\dot{\Psi}_2)^2+\frac{Ia^2}{8r^2}(\dot{\Psi}_2-\dot{\Psi}_1)^2+\frac{1}{2}C(\dot{\Psi}_1^2+\dot{\Psi}_2^2).$$

Alternatively, it is possible to find an equation in terms of a symplectic 2-form, but with an additional gyroscopic type 1-form:

$$i_{\bar{X}}\omega_{L*} = dE_{L*} + \overline{\alpha_{\Gamma_{L,D}}},$$

where

$$\overline{\alpha_{\Gamma_{L,D}}}=\frac{m_0la^3}{4r^2}(\dot{\Psi}_2-\dot{\Psi}_1)\dot{\Psi}_2\, d\Psi_1-\frac{m_0la^3}{4r^2}(\dot{\Psi}_2-\dot{\Psi}_1)\dot{\Psi}_1\, d\Psi_2.$$

ACKNOWLEDGMENTS

This work was partially supported by grants DGICYT (Spain) PB94-0106, Consejeria de Educacion y Cultura de la Comunidad de Madrid, Universidad de La Laguna and by the Fund for Scientific Research (Flanders, Belgium). D. Martín de Diego also wishes to thank the University of Gent (Belgium) for a Visiting Research Grant.

[1] P. A. M. Dirac, ''Lectures on quantum mechanics,'' Belfer Graduate School of Science, Yeshiva University, New York, 1964.
[2] M. J. Gotay, ''Presymplectic manifolds, geometric constraint theory and the Dirac–Bergmann theory of constraints,'' Ph.D. thesis, Center for Theoretical Physics, University of Maryland, 1979.
[3] M. J. Gotay and J. M. Nester, ''Presymplectic Lagrangian systems I: The constraint algorithm and the equivalence theorem,'' Ann. Inst. Henri Poincaré, Sect. A **30**, 129–142 (1979).
[4] M. J. Gotay and J. M. Nester, ''Presymplectic Lagrangian systems II: The second-order differential equation problem,'' Ann. Inst. Henri Poincaré, Sect. A **32**, 1–13 (1980).
[5] G. Marmo, G. Mendella, and W. M. Tulczyjew, ''Symmetries and constants of the motion for dynamics in implicit form,'' Ann. Inst. Henri Poincaré: Phys. Theor. **57**, 147–166 (1992).
[6] E. A. Lacomba and W. M. Tulczyjew, ''Geometric formulation of mechanical systems with one-sided constraints,'' J. Phys. A **23**, 2801–2813 (1990).
[7] A. Ibort, E. A. Lacomba, M. de León, D. Martín de Diego, and P. Pitanga, ''Mechanical systems subjected to impulsive constraints,'' J. Phys. A **30**, 5835–5854 (1997).
[8] J. Neimark and N. Fufaev, ''Dynamics of nonholonomic systems,'' *Transactions of Mathematical Monographs*, Vol. 33 (American Mathematical Society, Providence, 1972).
[9] A. M. Vershik and L. D. Faddeev, ''Differential geometry and Lagrangian mechanics with constraints,'' Sov. Phys. Dokl. **171**, 34–36 (1972).
[10] A. M. Vershik, ''Classical and non-classical dynamics with constraints,'' *Global Analysis—Studies and Applications I*, Lecture Notes in Mathematics No. 1108 (Springer-Verlag, Berlin, 1984), pp. 278–301.
[11] A. Ibort, M. de León, G. Marmo, D. and Martín de Diego, ''Non-holonomic constrained systems as implicit differential equations,'' in *Geometrical Structures for Physical Theories, I*, Rend. Sem. Mat. Univ. Pol. Torino, Vol. 54, No. 3 (Università e Politecnico, Torino, 1966), pp. 295–317.
[12] J. Koiller, ''Reduction of some classical non-holonomic systems with symmetry,'' Arch. Ration. Mech. Anal. **118**, 113–148 (1992).
[13] E. Massa and E. Pagani, ''Classical dynamics of non-holonomic systems: A geometric approach,'' Ann. Inst. Henri Poincaré: Phys. Theor. **55**, 511–544 (1991).
[14] W. Sarlet, F. Cantrijn, and D. J. Saunders, ''A differential geometric setting for mixed first- and second-order ordinary differential equations,'' J. Phys. A **30**, 4031–4052 (1997).
[15] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and R. M. Murray, ''Nonholonomic mechanical systems with symmetry,'' Arch. Ration. Mech. Anal. **136**, 21–99 (1996).
[16] A. D. Lewis, ''Aspects of geometric mechanics and control of mechanical systems,'' Ph.D. Thesis, California Institute of Technology, Pasadena, 1995.
[17] L. Bates and J. Śniatycki, ''Nonholonomic reduction,'' Rep. Math. Phys. **32**, 99–115 (1992).
[18] W. S. Koon and J. E. Marsden, ''Poisson reduction of nonholonomic mechanical systems with symmetry,'' in *Proceedings of the Workshop on Non-Holonomic Constraints in Dynamics*, Calgary, 26–29 August 1997, Rep. Math. Phys. **42**, 101–134 (1998).

[19] C.-M. Marle, ''Reduction of constrained mechanical systems and stability of relative equilibria,'' Commun. Math. Phys. **174**, 295–318 (1995).
[20] M. de León and P. R. Rodrigues, *Methods of Differential Geometry in Analytical Mechanics*, Math. Ser. 152 (North-Holland, Amsterdam, 1989).
[21] G. Hinds, ''Foliations and the Dirac theory of constraints,'' Ph.D. thesis, University of Maryland, 1965.
[22] J. F. Cariñena, ''Theory of singular Lagrangians,'' Fortschr. Phys. **88**, 641–679 (1990).
[23] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, NJ, 1992).
[24] M. de León and D. Martín de Diego, ''On the geometry of non-holonomic Lagrangian systems,'' J. Math. Phys. **37**, 3389–3414 (1996).
[25] A. Ibort, M. de León, J. C. Marrero, and D. Martín de Diego, ''Dirac brackets in constrained dynamics,'' Fortschr. Phys. (in press, 1998).
[26] M. de León and D. Martín de Diego, ''A constraint algorithm for singular Lagrangians subjected to nonholonomic constraints,'' J. Math. Phys. **38**, 3055–3062 (1997).
[27] F. Cantrijn, M. de León, J. C. Marrero, and D. Martín de Diego, ''Reduction of nonholonomic mechanical systems with symmetries,'' in Ref. 18, pp. 25–45.
[28] M. de León, J. C. Marrero, and D. Martín de Diego, ''Mechanical systems with non-linear constraints,'' Int. J. Theor. Phys. **36**, 973–989 (1997).
[29] C.-M. Marle, ''Various approaches to conservative and nonconservative nonholonomic systems,'' in Ref. 18, pp. 211–229.
[30] A. J. van der Schaft and B. M. Maschke, ''On the Hamiltonian formulation of nonholonomic mechanical systems,'' Rep. Math. Phys. **34**, 225–233 (1994).
[31] A. Ibort, M. de León, J. C. Marrero, and D. Martín de Diego, ''A Dirac-bracket for non-holonomic Lagrangian systems,'' in *Proceedings of the Vth Fall Workshop: Differential Geometry and its Applications*, Jaca, 23–25 September 1996.
[32] F. Cantrijn, M. de León, and D. Martín de Diego, ''On almost-Poisson structures in nonholonomic mechanics,'' preprint, 1998.
[33] L. Bates, H. Graumann, and C. MacDonnell, ''Examples of gauge conservation laws in nonholonomic systems,'' Rep. Math. Phys. **37**, 295–308 (1996).
[34] W. S. Koon and J. E. Marsden, ''The Hamiltonian and Lagrangian approaches to the dynamics of nonholonomic systems,'' Rep. Math. Phys. **40**, 21–62 (1997).
[35] J. Śniatycki, ''Non-holonomic Noether theorem and reduction of symmetries,'' in Ref. 18, pp. 5–23.
[36] P. Libermann and C.-M. Marle, *Symplectic Geometry and Analytical Mechanics* (Reidel, Dordrecht, 1987).
[37] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed. (Benjamin Cummings, Reading, MA, 1978).
[38] J. E. Marsden, R. Montgomery, and T. Ratiu, ''Reduction, symmetry, and phases in mechanics,'' Mem. Am. Math. Soc., Vol. 436 (American Mathematical Society, Providence, 1990).
[39] R. Cushman, D. Kemppainen, J. Śniatycki, and L. Bates, ''Geometry of nonholonomic constraints,'' Rep. Math. Phys. **36**, 275–286 (1995).
[40] F. Cantrijn, M. de León, and D. Martín de Diego, ''The momentum equation for non-holonomic systems with symmetry,'' in *Proceedings of the National Conference on Theoretical and Applied Mechanics*, Leuven, 22–23 May 1997, pp. 31–34.
[41] A. Capri and M. Kobayashi, ''A mechanical model with constraints,'' J. Math. Phys. **23**, 736–741 (1982).
[42] A. Capri and M. Kobayashi, ''The first-rank tensor field coupled to an electromagnetic field,'' J. Phys. A **26**, 6101–6112 (1987).
[43] A. Echeverría-Enríquez, M. C. Muñoz-Lecanda, and N. Román-Roy, ''Reduction of presymplectic manifolds with symmetry,'' preprint, 1997.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Gauge-invariant variationally trivial problems on T^*M

M. Castrillón López[a)]
Departamento de Geometría y Topología, Universidad Complutense de Madrid, Ciudad Universitaria, s/n, 28040-Madrid, Spain

J. Muñoz Masqué[b)]
Instituto de Física Aplicada, Consejo Superior de Investigaciones Científicas, Serrano 144, 28006 Madrid, Spain

(Received 29 June 1998; accepted for publication 29 October 1998)

A classification of variationally trivial Lagrangians on T^*M which are invariant under the Lie algebra of infinitesimal gauge transformations of the principal bundle $\pi: M\times \mathrm{U}(1)\to M$, is given. A characterization of Lagrangian densities on T^*M which are invariant under the Lie algebra of all infinitesimal automorphisms of $M\times \mathrm{U}(1)$ is also obtained.
[S0022-2488(99)02202-1]

I. INTRODUCTION

In the theory of the electromagnetic field, the notion of gauge invariance plays an important role. Given a space–time M, a gauge potential is a differential one-form ω on the manifold M. Moreover, the cotangent bundle T^*M can be identified in a natural way with the bundle of connections of the space of phase factors $\pi: P=M\times \mathrm{U}(1)\to M$. The automorphisms of the bundle P inducing the identity over M are the gauge transformations. The infinitesimal gauge transformations (that is, the gauge algebra of P, denoted $\mathrm{gau}\,P$) are the π-vertical U(1)-invariant vector fields on P. The flow of a vector field in $\mathrm{gau}\,P$ induces a one-parameter family of automorphisms of P. Automorphisms act on connections by pulling back connection forms. Taking the derivative of this representation we thus obtain a Lie algebra representation $X\mapsto \tilde{X}$, from the gauge algebra of P into the vector fields of T^*M. It is a well-known fact that gauge invariance corresponds to invariance under this representation. More precisely, a Lagrangian density $\mathcal{L}v_n$, where v_n is a volume form on M and $\mathcal{L}$ is a differentiable function on $J^1(T^*M)$, is said to be gauge invariant if for every $X\in \mathrm{gau}\,P$ we have $L_{\tilde{X}^{(1)}}(\mathcal{L}v_n)=0$, where $\tilde{X}^{(1)}$ is the natural lifting of $\tilde{X}$ to the one-jet bundle. As $\tilde{X}^{(1)}$ is π-vertical the above condition simply says that $\tilde{X}^{(1)}(\mathcal{L})=0$. Gauge-invariant Lagrangians are classified by the so-called geometric formulation of Utiyama's theorem (see, e.g., Refs. 1–4).

In this paper we first classify Lagrangian densities which are not only gauge invariant but also invariant under the full Lie algebra of infinitesimal automorphisms of P (not necessarily π-vertical), which is denoted by $\mathrm{aut}\,P$. Note that $\mathrm{gau}\,P$ is an Abelian ideal in $\mathrm{aut}\,P$ and $\mathrm{aut}\,P/\mathrm{gau}\,P\cong\mathfrak{X}(M)$. Theorem 3 states that there is no $\mathrm{aut}\,P$-invariant Lagrangian density if $n=\dim M$ is odd, and each $\mathrm{aut}\,P$-invariant Lagrangian density is proportional to the Pfaffian if $n=\dim M$ is even. An immediate consequence is that every $\mathrm{aut}\,P$-invariant Lagrangian density is variationally trivial. Hence invariance under the full Lie algebra of infinitesimal automorphism is a too strong condition in order to be of variational interest. This poses the problem of determining all gauge-invariant Lagrangians which are also variationally trivial. This is achieved in the second part of the paper. These Lagrangians admit a natural geometric interpretation in terms of multivector fields on the ground manifold (Theorem 6). In this way we also obtain a criterion for two gauge-invariant Lagrangians $\mathcal{L},\mathcal{L}'$ to be variationally equivalent.

[a)]Electronic mail: mcastri@mat.ucm.es
[b)]Electronic mail: jaime@iec.csic.es

II. PRELIMINARIES

A. The bundle $M\times U(1)$

Given a C^∞ manifold M of dimension n, let us consider the trivial bundle $\pi:P=M\times U(1)\to M$. The bundle of connections of π can be identified with the cotangent bundle $p:T^*M\to M$. In fact, every connection form on $M\times U(1)$ can be uniquely written as $\omega_\Gamma=(d\theta+\pi^*\omega)\otimes A$, where ω is an arbitrary one-form on M, θ is the angle coordinate on U(1), and A stands for the standard basis of the Lie algebra $\mathfrak{u}(1)$; that is, A is the invariant vector field on U(1) corresponding to the homomorphism $\mathbb{R}\to U(1)$, $t\mapsto\exp(tA)$. Let $(U;q^1,\dots,q^n)$ be a coordinate open domain in M. A vector field X on P is U(1)-invariant if and only if it can be written as

$$X=f^i(q^1,\dots,q^n)\frac{\partial}{\partial q^i}+g(q^1,\dots,q^n)V,\quad f^i,g\in C^\infty(U), \tag{1}$$

where V is the fundamental vector field on P defined by A. Note that each U(1)-invariant vector field X on P is π-projectable and its projection is given by

$$\pi_*X=f^i(q^1,\dots,q^n)\frac{\partial}{\partial q^i}.$$

In particular, X is π-vertical (and hence it is a gauge vector field) if and only if $f^i=0$. We think of the Lie algebra of U(1)-invariant vector fields as being the infinitesimal automorphisms of the principal bundle P and we denote it by $\operatorname{aut}P$. Gauge vector fields are an Abelian ideal in $\operatorname{aut}P$, and we have an exact sequence of Lie algebras

$$0\to\operatorname{gau}P\to\operatorname{aut}P\xrightarrow{\pi_*}\mathfrak{X}(M)\to 0. \tag{2}$$

In fact, $\operatorname{aut}P$ is the semidirect sum of $\operatorname{gau}P$ and $\mathfrak{X}(M)$. We denote by $(p^{-1}(U);q^1,\dots,q^n,p_1,\dots,p_n)$ the coordinate system induced on the cotangent bundle. That is, $w=p_i(w)(dq^i)_x$, for every covector $w\in T_x^*M$.

B. The basic representation

Automorphisms of a principal bundle act on connections in a natural way (Ref. 5, Sec. II, Preposition 6.1). The induced representation of the Lie algebra $\operatorname{aut}P$ into the vector fields of the cotangent bundle is then given by (Ref. 2, Sec. III B)

$$\tilde{X}=f^i\frac{\partial}{\partial q^i}-\left(\frac{\partial g}{\partial q^i}+\frac{\partial f^h}{\partial q^i}p_h\right)\frac{\partial}{\partial p_i}, \tag{3}$$

where the local expression of $X\in\operatorname{aut}P$ is given by formula (1). In particular, for gauge vector fields we have

$$\tilde{X}=-\frac{\partial g}{\partial q^i}\frac{\partial}{\partial p_i},\quad X\in\operatorname{gau}P. \tag{4}$$

We remark that in this case $\tilde{X}$ is the Hamiltonian vector field associated with the function g with respect to the canonical symplectic form $dq^i\wedge dp_i$.

C. Jet prolongation

Let $p_1:J^1(T^*M)\to M$ be the first jet prolongation of the cotangent bundle. Each coordinate open domain $(U;q^1,\dots,q^n)$ in M induces a system of coordinates $(q^i,p_i;p_{i,j})$, $i,j=1,...,n$, in $p_1^{-1}(U)\subseteq J^1(T^*M)$, where $(p^{-1}(U);q^i,p_i)$ is the coordinate system induced on the cotangent bundle, which are defined by

$$p_{i,j}(j_x^1\omega)=\frac{\partial(p_i\circ\omega)}{\partial q^j}(x), \quad i,j=1,...,n.$$

From the general formulas for jet prolongation for every $X\in \operatorname{aut} P$ we obtain

$$\tilde{X}^{(1)}=f^i\frac{\partial}{\partial q^i}-\left(\frac{\partial g}{\partial q^i}+\frac{f^h}{\partial q^i}p_h\right)\frac{\partial}{\partial p_i}-\left(\frac{\partial^2 g}{\partial q^i\partial q^j}+\frac{f^h}{\partial q^i}p_{h,j}+\frac{\partial f^h}{\partial q^j}p_{i,h}+\frac{\partial^2 f^h}{\partial q^i\partial q^j}p_h\right)\frac{\partial}{\partial p_{i,j}}, \quad (5)$$

which is the jet prolongation of the algebra of infinitesimal automorphisms.

D. Lie derivative with respect to a multivector field

We recall (cf. Ref. 6; Ref. 7—p. 79) that a decomposable multivector field $\chi=X_1\wedge\ldots\wedge X_k \in\wedge^k\mathfrak{X}(M)$, induces a $C^\infty(M)$-linear graded endomorphism $i_\chi:\Omega^\cdot(M)\to\Omega^\cdot(M)$ of degree $-k$ by setting

$$i_\chi\omega_r=i_{X_1}\circ\ldots\circ i_{X_k}\omega_r\in\Omega^{r-k}(M), \quad \omega_r\in\Omega^r(M). \quad (6)$$

For an arbitrary multivector field χ we can define i_χ by extending the above formula by linearity. Then the Lie derivative $L_\chi:\Omega^\cdot(M)\to\Omega^\cdot(M)$ with respect to $\chi\in\wedge^k\mathfrak{X}(M)$ is given by the formula

$$L_\chi=i_\chi\circ d-(-1)^k d\circ i_\chi=[i_\chi,d]. \quad (7)$$

We remark that L_χ is a graded operator of degree $-k+1$.

III. INVARIANT LAGRANGIANS

A. Gauge invariance

A Lagrangian function $\mathcal{L}:J^1(T^*M)\to\mathbb{R}$ is said to be gauge invariant if

$$\tilde{X}^{(1)}(\mathcal{L})=0, \quad \forall X\in \operatorname{gau} P.$$

As the structure group is Abelian the Utiyama theorem states that a Lagrangian is gauge invariant if and only if it can be written as

$$\mathcal{L}=\bar{\mathcal{L}}\circ\Omega,$$

where $\Omega:J^1(T^*M)\to\wedge^2T^*M$ is the mapping given by $\Omega(j_x^1\omega)=(d\omega)_x$ and $\bar{\mathcal{L}}:\wedge^2T^*M\to\mathbb{R}$ is an arbitrary differentiable function (cf. Ref. 1). Let us introduce coordinates on the bundle $p_2:\wedge^2T^*M\to M$. For every two-covector $w_2\in\wedge^2T_x^*M$ we set

$$w_2=p_{ij}(w_2)(dq^i\wedge dq^j)_x, \quad i<j.$$

Then, (q^i,p_{ij}) is a coordinate system on $p_2^{-1}(U)$. We set $p_{ij}=-p_{ji}$ for $i\geqslant j$. With respect to these coordinates the equations of Ω are

$$p_{ij}\circ\Omega=p_{i,j}-p_{j,i}. \quad (8)$$

Hence gauge-invariant Lagrangians can be locally written as

$$\mathcal{L}=\bar{\mathcal{L}}(q^h,p_{i,j}-p_{j,i}).$$

B. aut P-invariance

Let us assume that M is oriented by a volume form v_n. A Lagrangian density $\mathcal{L}v_n$, $\mathcal{L}\in C^\infty(J^1(T^*M))$, is said to be invariant under the Lie algebra of all infinitesimal automorphisms of P (in short, aut P-invariant) if

$$L_{\tilde{X}^{(1)}}(\mathcal{L}v_n)=0, \quad \forall X\in \mathrm{aut}\,P.$$

As $L_{\tilde{X}^{(1)}}(\mathcal{L}v_n)=(\tilde{X}^{(1)}(\mathcal{L}))v_n+\mathcal{L}\,\mathrm{div}(\pi_* X)v_n=0$, where the divergence is taken with respect to v_n, the invariance condition is equivalent to the following:

$$\tilde{X}^{(1)}(\mathcal{L})+\mathcal{L}\,\mathrm{div}(\pi_* X)=0, \quad \forall X\in \mathrm{aut}\,P.$$

Remark: Unlike the case of gauge-invariant Lagrangians, there are no aut P-invariant Lagrangians except for constant functions. This is easily seen from formula (5). In fact, assume that the Lagrangian $\mathcal{L}$ is aut P-invariant; i.e., $\tilde{X}^{(1)}(\mathcal{L})=0$, $\forall X\in \mathrm{aut}\,P$. Taking $f^i=1$, $g=0$ in (1) we obtain $\partial\mathcal{L}/\partial q^i=0$; taking $f^i=0$, $g=q^i$, we obtain $\partial\mathcal{L}/\partial p_i=0$; finally, the choice $f^i=0$, $g=q^iq^j$ yields $\partial\mathcal{L}/\partial p_{i,j}=0$. In the next theorem we classify the aut P-invariant Lagrangian densities.

Lemma 1: Let N be a differentiable manifold. If $f\in C^\infty(\mathbb{R}^k\times N)$ satisfies $f=\Sigma_{i=1}^k x_i(\partial f/\partial x_i)$, then $\partial f/\partial x_i\in C^\infty(N)$, $1\leqslant i\leqslant k$.

Proof: From the property in the statement we obtain $f(tx,y)=tf(x,y)$, $\forall t\in\mathbb{R}$, $\forall x\in\mathbb{R}^n$, $\forall y\in N$. Hence $tf(x,y)=f(tx,y)=\Sigma_{i=1}^k tx_i(\partial f/\partial x_i)(tx,y)$ and then, $f(x,y)=\Sigma_{i=1}^k x_i(\partial f/\partial x_i)(tx,y)$. Letting $t=0$, we have $f(x,y)=\Sigma_{i=1}^k x_i(\partial f/\partial x_i)(0,y)$ and taking derivatives with respect to x_j, we conclude. □

Lemma 2: If $f\in C^\infty(\wedge^2T^*\mathbb{R}^n)$ satisfies

$$f=\sum_{j=1}^n p_{ij}(\partial f/\partial p_{ij}), \quad 1\leqslant i\leqslant n, \tag{9}$$

then,

(1) *If n is odd, then $f=0$.*
(2) *If n is even, then f is a polynomial of the form $\lambda^{i_1i_2\cdots i_{n-1}i_n}p_{i_1i_2}\cdots p_{i_{n-1}i_n}$, with $\{i_1,i_2,\ldots,i_{n-1},i_n\}=\{1,2,\ldots,n-1,n\}$, where $\lambda^{i_1i_2\cdots i_{n-1}i_n}\in C^\infty(\mathbb{R}^n)$.*

Proof: We proceed by induction on n. If $n=2$, then $f=p_{12}(\partial f/\partial p_{12})$. Hence f is proportional to p_{12}. If $n=3$, then for $i=1$, we have $f=p_{12}(\partial f/\partial p_{12})+p_{13}(\partial f/\partial p_{13})$ and for $i=2$ we also have $f=p_{21}(\partial f/\partial p_{21})+p_{23}(\partial f/\partial p_{23})$. By virtue of Lemma 1 the functions $\partial f/\partial p_{12}$, $\partial f/\partial p_{13}$ do not depend on p_{12}, p_{13} and $\partial f/\partial p_{21}=-\partial f/\partial p_{12}$, $\partial f/\partial p_{23}$ do not depend on p_{23}. Accordingly, $\partial f/\partial p_{12}$ is a constant. Similarly, $\partial f/\partial p_{13}$ is also a constant. By comparing the above two expressions of f we conclude that $\partial f/\partial p_{23}=0$ and the same occurs for the other variables.

Assume $n>3$. We set $f_i^j=\partial f/\partial p_{ij}$. For $i=1,\ldots,n$, we have $f=p_{ij}f_i^j$ and from Lemma 1 we can conclude $\partial f_i^j/\partial p_{ik}=0$. In particular, $\partial f_1^j/\partial p_{1k}=0$ and $\partial f_1^j/\partial p_{jk}=-\partial f_1^j/\partial p_{jk}=0$. In other words, f_1^j does not depend on p_{ik} for $i=1,j$, $1\leqslant k\leqslant n$. Next, we prove that f_1^j also satisfies (9) in $\wedge^2T^*\mathbb{R}^{n-2}$. In fact, for $i>1$, we have

$$f=p_{ik}\frac{\partial f}{\partial p_{ik}}=p_{ik}\frac{\partial}{\partial p_{ik}}(p_{1j}f_1^j)=-p_{i1}f_1^i+p_{ik}p_{1j}\frac{\partial f_1^j}{\partial p_{ik}},$$

and taking derivatives with respect to p_{1l}, with $l\neq i$, we obtain

$$f_1^l=\sum_{k=1}^n p_{ik}\frac{\partial f_1^l}{\partial p_{ik}}.$$

Consequently, we can apply the induction hypothesis thus concluding. □

Theorem 3: *We have*

(1) *If the dimension of M is even, say* $\dim M = 2m$, *then up to a multiplicative constant, the unique* aut P*-invariant Lagrangian density on* T^*M *is* $(Pf\circ\Omega)v_{2m}$, *where* $\mathrm{Pf}:\wedge^2 T^*M \to \mathbb{R}$ *is the Pfaffian function,* i.e.,

$$\mathrm{Pf}(w_2)v_{2m} = w_2 \wedge \overset{(m}{\ldots} \wedge w_2 .$$

(2) *If the dimension of M is odd, then zero is the unique* aut P*-invariant Lagrangian density*.

Proof: Let $(U;q^1,\ldots,q^n)$ be a coordinate open domain such that $v_n = dq^1\wedge\cdots\wedge dq^n$. Let us consider an aut P-invariant Lagrangian density $\mathcal{L}v_n$. In particular, $\mathcal{L}$ is gauge invariant. Hence from Utiyama's theorem we have $\mathcal{L}=\overline{\mathcal{L}}\circ\Omega$, where $\overline{\mathcal{L}}:\wedge^2 T^*M\to\mathbb{R}$ is an arbitrary differentiable function, or else $\mathcal{L}=\overline{\mathcal{L}}(q^h,p_{i,j}-p_{j,i})$, locally. Consequently we only need to impose invariance under the vector fields such that $g=0$. We have

$$0=\tilde{X}^{(1)}\mathcal{L}+\mathcal{L}\ \mathrm{div}(\pi_* X)=f^i\,\frac{\partial\mathcal{L}}{\partial q^i}-\left(\frac{\partial f^h}{\partial q^i}\,p_{h,j}+\frac{\partial f^h}{\partial q^j}\,p_{i,h}+\frac{\partial^2 f^h}{\partial q^i\partial q^j}\,p_h\right)\frac{\partial\mathcal{L}}{\partial p_{i,j}}+\mathcal{L}\,\frac{\partial f^i}{\partial q^i}.$$

Let us fix an index $b=1,\ldots,n$. If we take $f^a=\delta^{ab}$, $1\leqslant a\leqslant n$, then we obtain

$$\frac{\partial\mathcal{L}}{\partial q^i}=0,\quad 1\leqslant i\leqslant n. \tag{10}$$

If we take $f^a=\delta^{ab}q^b$, $1\leqslant a\leqslant n$, then we obtain

$$\mathcal{L}=p_{i,j}\,\frac{\partial\mathcal{L}}{\partial\partial p_{i,j}}+p_{j,i}\,\frac{\partial\mathcal{L}}{\partial p_{j,i}},\quad 1\leqslant i\leqslant n. \tag{11}$$

Finally, if we fix another index $c\neq b$, and take $f^a=\delta^{ab}q^c$, $1\leqslant a\leqslant n$, we obtain

$$p_{i,j}\,\frac{\partial\mathcal{L}}{\partial p_{k,j}}+p_{j,i}\,\frac{\partial\mathcal{L}}{\partial p_{j,k}}=0,\quad i,k=1,\ldots,n,\quad i\neq k. \tag{12}$$

Taking into account that

$$\frac{\partial\mathcal{L}}{\partial p_{i,j}}=\frac{\partial\overline{\mathcal{L}}}{\partial p_{ij}}\circ\Omega,$$

and that Ω is surjective, Eqs. (10)–(12) can be rewritten as

$$\frac{\partial\overline{\mathcal{L}}}{\partial q^i}=0,\quad 1\leqslant i\leqslant n. \tag{13}$$

$$\overline{\mathcal{L}}\delta^{hi}=p_{hj}\,\frac{\partial\overline{\mathcal{L}}}{\partial p_{ij}},\quad h,i=1,\ldots,n. \tag{14}$$

Letting $h=i$ in (14), from Lemma 2 we deduce that $\overline{\mathcal{L}}$ vanishes if n is odd and if n is even then $\overline{\mathcal{L}}$ is a polynomial,

$$\overline{\mathcal{L}}=\lambda^{i_1 i_2\cdots i_{n-1} i_n}p_{i_1 i_2}\cdots p_{i_{n-1} i_n}, \tag{15}$$

where $i_1,\ldots,i_n$ is a permutation of $1,\ldots,n$, and $\lambda^{i_1 i_2 \cdots i_{n-1} i_n}$ are constants due to (13). If we take $h \neq i$ in (14) we obtain

$$p_{hj}\,\frac{\partial \overline{\mathcal{L}}}{\partial p_{ij}}=0. \tag{16}$$

Taking derivatives with respect to p_{hk} in (16) we have

$$\frac{\partial \overline{\mathcal{L}}}{\partial p_{ij}}+p_{hj}\,\frac{\partial^2 \overline{\mathcal{L}}}{\partial p_{hk}\partial p_{ij}}=0,$$

and taking derivatives with respect to p_{hl} in the above equation, we have

$$\frac{\partial^2 \overline{\mathcal{L}}}{\partial p_{hl}\partial p_{ik}}+\frac{\partial^2 \overline{\mathcal{L}}}{\partial p_{hk}\partial p_{il}}=0, \tag{17}$$

as the third partial derivative with respect to p_{hl}, p_{hk}, p_{il} vanishes by virtue of (15). Equation (17) tells us that $\lambda^{i_1,\ldots,k,\ldots,l,\ldots,i_n}+\lambda^{i_1,\ldots,l,\ldots,k,\ldots,i_n}=0$, and since the transpositions generate the group of permutations we obtain

$$\lambda^{\pi(1),\ldots,\pi(n)}=(\operatorname{sgn}\,\pi)\lambda^{1,\ldots,n},$$

for every permutation π of $1,\ldots,n$, and the result follows from the very definition of the Pfaffian [cf. Ref. 8 formula (4.12), p. 66; Ref. 5, Sec. XII]. □

Remark: Once the function $\mathcal{L}$ is known to be a polynomial of degree *m*, the above result can be also obtained from (Ref. 9, Theorem 2.6.2), taking into account the remark below.

Remark: Let $\pi: P \to M$ be an arbitrary principal U(1)-bundle and let us denote by $p:\mathcal{C}(P) \to M$ the bundle of connections of P (see, e.g., Ref. 10). As in the case of the trivial bundle, we can introduce coordinates (q^i, p_j) on the bundle $\mathcal{C}(P)$. Then, it is not difficult to see that the manifold $J^1(\mathcal{C}(P))$ is endowed with a canonical horizontal two-form locally given by $\eta=(p_{i,j}-p_{j,i})dq^i \wedge dq^j$. This form enjoys the following characteristic property. If $\sigma_\Gamma: M \to \mathcal{C}(P)$ is the section of p induced from a connection Γ on P, then $\pi^*(j^1\sigma_\Gamma)^*\eta=\Omega_\Gamma$, the curvature form of the given connection. Because of this, the form η can be called the universal Chern form of line bundles.

Taking into account the above observation, it is worth mentioning that the first item of Theorem 3 can be interpreted by saying that the only aut P-invariant Lagrangian densities for connections on P are constant multiples of the m-fold wedge product of the universal Chern form.

IV. VARIATIONALLY TRIVIAL LAGRANGIANS

A Lagrangian density $\mathcal{L}v_n$, $\mathcal{L}\in C^\infty(J^1(T^*M))$, is said to be variationally trivial if every differential one-form is a solution to the Euler–Lagrange equations of the given Lagrangian; that is, for every differential form ω defined on a neighborhood of a point $x \in M$ we have

$$\frac{\partial \mathcal{L}}{\partial p_i}\,(j_x^1\omega)-\frac{\partial}{\partial q^k}\left(\frac{\partial \mathcal{L}}{\partial p_{i,k}}\circ j^1\omega\right)(x)=0. \tag{18}$$

Proposition 4: A Lagrangian density $\mathcal{L}v_n$ on T^*M is variationally trivial if and only if the following partial differential equations hold true:

$$\frac{\partial^2 \mathcal{L}}{\partial p_{i,h}\partial p_{j,k}}+\frac{\partial^2 \mathcal{L}}{\partial p_{i,k}\partial p_{j,h}}=0, \quad \forall h,i,j,k=1,\ldots,n, \tag{19}$$

$$\frac{\partial \mathcal{L}}{\partial p_i} = \frac{\partial^2 \mathcal{L}}{\partial q^k \partial p_{i,k}} + \frac{\partial^2 \mathcal{L}}{\partial p_j \partial p_{i,k}} p_{j,k}, \quad \forall i = 1,...,n. \tag{20}$$

Proof: Expanding Eq. (18) we obtain

$$\frac{\partial \mathcal{L}}{\partial p_i}(j_x^1\omega) = \frac{\partial^2 \mathcal{L}}{\partial q^k \partial p_{i,k}}(j_x^1\omega) + \frac{\partial^2 \mathcal{L}}{\partial p_j \partial p_{i,k}}(j_x^1\omega)\frac{\partial f_j}{\partial q^k}(x) + \frac{\partial^2 \mathcal{L}}{\partial p_{j,h} \partial p_{i,k}}(j_x^1\omega)\frac{\partial^2 f_j}{\partial q^h \partial q^k}(x),$$

where $\omega = f_j dq^j$. Once the one-jet of ω at x has been fixed, the second partial derivatives $(\partial^2 f_j / \partial q^h \partial q^k)(x)$ can be arbitrarily taken. Hence Eq. (19) follows and also Eq. (20). □

Notation 5: Given a multivector field $\chi_k \in \Gamma(M, \wedge^{2k}TM)$ of even degree $2k$, we denote by $\overline{\chi_k} : \wedge^2 T^*M \to \mathbb{R}$ the function given by

$$\overline{\chi_k}(w_2) = i_{\chi_k}(w_2 \wedge \overset{(k}{\ldots} \wedge w_2), \quad w_2 \in \overset{2}{\wedge} T^*M,$$

where i_χ is the total insertion operator (cf. Sec. II D).

Theorem 6: *A gauge-invariant Lagrangian density $\mathcal{L}v_n$ on T^*M is variationally trivial if and only $\mathcal{L}$ can be written as follows:*

$$\mathcal{L} = \overline{\chi} \circ \Omega,$$

where $\Omega : J^1(T^*M) \to \wedge^2 T^*M$ is the curvature mapping (see Sec. III A) *and*

$$\overline{\chi} = \sum_{k=0}^{[n/2]} \overline{\chi_k},$$

χ_k being a multivector field of degree $2k$, such that $L_{\chi_k} v_n = 0$ (cf. Sec. II D).

Proof: As $\mathcal{L}v_n$ is gauge invariant according to Utiyama's theorem (Sec. III A) we have

$$\mathcal{L} = \overline{\mathcal{L}} \circ \Omega,$$

where $\overline{\mathcal{L}}$ is an arbitrary differentiable function on $\wedge^2 T^*M$. As $\mathcal{L}$ must also be variationally trivial, the above composition should satisfy Eqs. (19) and (20). Taking into account the equations of Ω [see formula (8)] we have

$$\frac{\partial^2 \overline{\mathcal{L}}}{\partial p_{ih} p_{jk}} \circ \Omega + \frac{\partial^2 \overline{\mathcal{L}}}{\partial p_{ik} \partial p_{jh}} \circ \Omega = 0, \quad \forall h,i,j,k = 1,...,n,$$

and

$$\frac{\partial^2 \overline{\mathcal{L}}}{\partial q^k \partial p_{ik}} \circ \Omega = 0, \quad \forall i = 1,...,n,$$

as $\partial \mathcal{L} / \partial p_i$, $\partial^2 \mathcal{L} / \partial p_j \partial p_{i,k}$ vanish for $\mathcal{L}$ does not depend on p_i. Since Ω is a surjective mapping we obtain the corresponding equations for $\overline{\mathcal{L}}$ on the bundle $\wedge^2 T^*M$, i.e.,

$$\frac{\partial^2 \overline{\mathcal{L}}}{\partial p_{ih} \partial p_{jk}} + \frac{\partial^2 \overline{\mathcal{L}}}{\partial p_{ik} \partial p_{jh}} = 0, \quad \forall h,i,j,k = 1,...,n, \tag{21}$$

$$\frac{\partial^2 \overline{\mathcal{L}}}{\partial q^k \partial p_{ik}} = 0, \quad \forall i = 1,...,n. \tag{22}$$

From (21) we obtain $\partial^2 \overline{\mathcal{L}}/\partial p_{ih}^2 = 0$. Hence for each variable p_{ih} we have $\overline{\mathcal{L}} = a + b p_{ih}$, where a,b are differentiable functions depending on q^j and the rest of the variables p_{kl}. By recurrence we conclude that $\overline{\mathcal{L}}$ is a polynomial in the variables p_{ih} with coefficients in $C^\infty(M)$; that is,

$$\overline{\mathcal{L}} = \sum_k f^{i_1 \cdots i_{2k}} p_{i_1 i_2} \cdots p_{i_{2k-1} i_{2k}}, \quad f^{i_1 \cdots i_{2k}} \in C^\infty(M). \tag{23}$$

For $k = h$, from (21) we conclude that the functions $f^{i_1 \cdots i_{2k}}$ with a repeated index, must vanish. For arbitrary indices, Eq. (21) tells us that $f^{i_1, \dots, j, \dots, l, \dots, i_{2k}} + f^{i_1, \dots, l, \dots, j, \dots, i_{2k}} = 0$. Hence

$$f^{i_{\pi(1)}, \dots, i_{\pi(2k)}} = (\operatorname{sgn} \pi) f^{i_1, \dots, i_{2k}},$$

for every permutation π of $1,...,2k$, and therefore

$$\overline{\mathcal{L}} = \sum_{k=0}^{[n/2]} \sum_{i_1 < \cdots < i_{2k}} \sum_{\pi \in \mathfrak{S}_{2k}} (\operatorname{sgn} \pi) f^{i_1, \dots, i_{2k}} p_{i_{\pi(1)} i_{\pi(2)}} \cdots p_{i_{\pi(2k-1)} i_{\pi(2k)}}.$$

Hence $\overline{\mathcal{L}} = \overline{\chi} = \Sigma_{k=0}^{[n/2]} \overline{\chi}_k$, where

$$\chi_k = (-1)^{\binom{2k}{2}} 2^k \sum_{i_1 < \cdots < i_{2k}} f^{i_1, \dots, i_{2k}} \frac{\partial}{\partial q^{i_1}} \wedge \cdots \wedge \frac{\partial}{\partial q^{i_{2k}}}.$$

Let us check that Eq. (22) implies that each multivector field χ_k is divergence free. Let $(U; q^1, \dots, q^n)$ be a coordinate open domain such that $v_n = dq^1 \wedge \cdots \wedge dq^n$. By using the formula (7) in Sec. II D, we have

$$\begin{aligned} c_k(L_{\chi_k} v_n) &= (-1)^{k+1} \sum_{i_1 < \cdots < i_{2k}} d(f^{i_1, \dots, i_{2k}} i_{\partial/\partial q^{i_1} \wedge \cdots \wedge \partial/\partial q^{i_{2k}}} (dq^1 \wedge \cdots \wedge dq^n)) \\ &= - \sum_{i_1 < \cdots < i_{2k}} (-1)^{i_1 + \cdots + i_{2k}} d(f^{i_1, \dots, i_{2k}} dq^{j_1} \wedge \cdots \wedge dq^{j_{n-2k}}), \end{aligned}$$

where $c_k = (-1)\binom{2k}{2} 2^{-k}$ and $j_1 < \cdots < j_{n-2k}$ is the complementary subset of $i_1 < \cdots < i_{2k}$ in $\{1,...,n\}$. Hence

$$c_k(L_{\chi_k} v_n) = - \sum_{i_1 < \cdots < i_{2k}} (-1)^{i_1 + \cdots + i_{2k}} \sum_\alpha \frac{\partial f^{i_1, \dots, i_{2k}}}{\partial q^\alpha} dq^\alpha \wedge dq^{j_1} \wedge \cdots \wedge dq^{j_{n-2k}}.$$

If $j_0 < \cdots < j_{n-2k}$ is the complementary subset of $i_2 < \cdots < i_{2k}$ in $\{1,...,n\}$, then the coefficient of $dq^{j_0} \wedge dq^{j_1} \wedge \cdots \wedge dq^{i_{n-2k}}$ in the above formula is

$$(-1)^{i_2 + \cdots + i_{2k}} \sum_\alpha \frac{\partial f^{\alpha, i_2, \dots, i_{2k}}}{\partial q^\alpha},$$

where for (not necessarily increasing) pairwise different indices $i_1, \dots, i_{2k}$ we have set $f^{i_1, \dots, i_{2k}} = (\operatorname{sgn} \pi) f^{\pi(i_1), \dots, \pi(i_{2k})}$, π being the unique permutation of the set $\{i_1, \dots, i_{2k}\}$ such that $\pi(i_1) < \cdots < \pi(i_{2k})$. Accordingly, $L_{\chi_k} v_n = 0$ if and only if

$$\sum_\alpha \frac{\partial f^{\alpha, i_2, \dots, i_{2k}}}{\partial q^\alpha} = 0. \tag{24}$$

Moreover, from the formula (23) we have

$$\frac{\partial^2 \overline{\mathcal{L}}}{\partial q^\alpha \partial p_{j\alpha}} = \sum_{k=0}^{[n/2]} \frac{\partial f^{i_1,\ldots,i_{2k}}}{\partial q^\alpha} \frac{\partial p_{i_1 i_2} \cdots p_{i_{2k-1} i_{2k}}}{\partial p_{j\alpha}} = \sum_{k=0}^{[n/2]} k \frac{\partial f^{j,\alpha,i_3 \cdots i_{2k}}}{\partial q^\alpha} p_{i_3 i_4} \cdots p_{i_{2k-1} i_{2k}}.$$

Hence the formula (22) holds true if and only if

$$\sum_\alpha \frac{\partial f^{j,\alpha,i_3 \cdots i_{2k}}}{\partial q^\alpha} = 0,$$

which is equivalent to (24), thus finishing the proof of the theorem. □

Corollary 7: Every aut P*-invariant Lagrangian density on* T^*M *is variationally trivial.*

Proof: If dim M is odd, the result is trivial as in odd dimensions every aut P-invariant Lagrangian density vanishes by virtue of Theorem 3. Assume dim $M = n = 2m$. In this case, on a coordinate system such that $v_n = dq^1 \wedge \cdots \wedge dq^n$, as a simple calculation shows we have $Pf = \overline{\chi_m}$, where

$$\chi_m = (-1)^{\binom{n}{2}} \frac{\partial}{\partial q^1} \wedge \cdots \wedge \frac{\partial}{\partial q^n},$$

and $L_{\chi_m} v_n = 0$ obviously. □

ACKNOWLEDGMENT

J.M.M. was supported by the DLES (Spain) under Grant No. PB 95-0124.

[1] D. Bleecker, *Gauge Theory and Variational Principles* (Addison–Wesley, Reading, MA, 1981).

[2] L. Hernández Encinas and J. Muñoz Masqué, ''Symplectic structure and gauge invariance on the cotangent bundle,'' J. Math. Phys. **35**, 426–434 (1994).

[3] K. B. Marathe and G. Martucci, ''The geometry of gauge fields'' J. Geom. Phys. **6**, 1–106 (1989).

[4] R. Utiyama, ''Invariant theoretical interpretation of interaction'' Phys. Rev. **101**, 1597–1607 (1956).

[5] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley Interscience, New York, 1963), Vol. I, Vol. II (1969).

[6] J. V. Beltrán and J. Monterde, ''Graded Poisson structures on the algebra of differential forms,'' Comment. Math. Helvetici **70**, 383–402 (1995).

[7] A. Cabras and A. M. Vinogradov, ''Extensions of the Poisson bracket to differential forms and multivector fields,'' J. Geom. Phys. **9**, 75–100 (1992).

[8] J. L. Dupont, *Curvature and Characteristic Classes* [Lect. Notes Math **640** (1978)].

[9] P. B. Gilkey, *Invariance Theory, the Heat Equation, and the Atiyah–Singer Index Theorem* (CRC, Boca Raton, FL, 1994).

[10] D. J. Eck, ''Gauge-natural bundles and generalized gauge theories,'' Mem. Am. Math. Soc. **247** (1981).

On the geometric structure of thermodynamics

M. Chen
Vanier College, 821 Ste. Croix Avenue, St. Laurent, Quebec H4L 3X9, Canada

(Received 5 October 1998; accepted for publication 28 October 1998)

In this paper we consider symmetry transformations of the generalized entropy function that preserve the Gibbs one-form (Gibbs relation). We show that this symmetry consideration naturally leads to the geometric structure of thermodynamics in terms of contact geometry. We also construct an example based on the van der Waals' fluid to illustrate the method discussed in the paper. © *1999 American Institute of Physics.* [S0022-2488(99)03102-3]

I. INTRODUCTION

It is well known that Legendre transformations play an important role in equilibrium thermodynamics (ET).[1] In 1973 Hermann[2] suggested that ET might be reformulated in a geometric setting in terms of a contact manifold M. Since then several geometric theories of ET have been proposed in the past.[3] Recently Mrugala *et al.*[4] further investigated the applications of contact geometry to ET, where thermodynamic processes were considered as the flows of a vector field in TM, the tangent bundle of M. The work of Mrugala *et al.* was modified and extended to irreversible thermodynamics by Chen and Tseng.[5] In this paper we reexamine the symmetry property of the second law of thermodynamics and show that this symmetry consideration naturally gives rise to the intrinsic geometric structure of thermodynamics.

Consider a system of molecules in r components contained in a volume V, where no chemical reactions take place. Let E be the total internal energy of the system, N_i be the number of molecules of species i, P be the hydrostatic pressure, and μ_i be the chemical potential corresponding to N_i. In ET the work one-form dW is given by $dW=-P\,dV+\sum_i\mu_i\,dN_i$. In order to consider irreversible processes we enlarge the set of conserved variables (E,V,N_i) by including the nonconserved variables Φ_i^α, where Φ_i^α denotes the various generalized fluxes such as the mass flux, heat flux, etc.[6] Thus, $\{E,V,N_i,\Phi_i^\alpha;\ i=1,2,...,r;\ \alpha=1,2,...,k\}$ is the set of global thermodynamic variables. Note that the index α represents the different kinds of fluxes. For simplicity we denote $\{E,V,N_i,\Phi_i^\alpha\}$ by $(x^1,x^2,\ldots,x^n)=(x^1,\hat{x})$, where x^i is a differentiable function of time t. Let B_n be the thermodynamic space with coordinate cover x. The work one-form for a dissipative system can be written as

$$dW=-P\,dV+\sum_i \mu_i\,dN_i-\sum_{i,\alpha} X_i^\alpha\,d\Phi_i^\alpha, \tag{1}$$

where X_i^α is the generalized potential conjugate to the generalized fluxes Φ_i^α.[6] Hereafter we assume that P, μ_i, and X_i^α are functions of x of class C^1 on an open set of B_n with compact support. These functions are the constitutive relations based on phenomenological considerations.

The global formulation of the first law can be expressed as

$$\Delta E=\Delta W+\Delta Q+\Delta Q_d, \tag{2}$$

while the second law can be expressed as[7]

$$\Omega\wedge d\Omega=0, \tag{3a}$$

$$\Delta Q_d\geqslant 0. \tag{3b}$$

Here $\Omega = dE - dW$ and ΔQ represents the net amount of heat exchanged between the system and its surroundings, while ΔQ_d represents the change of dissipative energy associated with irreversible processes. We assume that ΔQ_d is semipositive definite and vanishes only at thermodynamic equilibrium. It should be noticed that ΔQ and ΔQ_d cannot be considered as one-forms in the vector space $\wedge(B_n)$ of differential forms in B_n, while $\Omega = dE - dW$ is a one-form in $\wedge^1(B_n)$.[8]

Let γ be a closed path in B_n. It can be proved that the second law formulated in (3a) and (3b) is equivalent to Clausius' inequality[9]

$$\int_\gamma T^{-1}\, dQ \leqslant 0,$$

as well as Kelvin's principle.[10]

Denote $\{-P, \mu_i, -X_i^\alpha;\ i=1,2,...,r;\ \alpha=1,2,...,k\}$ by $(w_2, w_3, \ldots, w_n)$. Then the work one-form dW can be rewritten as

$$dW = \sum_{j \geqslant 2} w_j(x)\, dx^j, \tag{4}$$

and the one-form $\Omega = dE - dW$ as

$$\Omega = dE - dW = dx^1 - \sum_{j \geqslant 2} w_j(x)\, dx^j. \tag{5}$$

It can be shown that the Caratheodory inaccessibility condition (3a) (or Frobenius integrability condition[11]) is equivalent to the Maxwell relations $\partial_j w_k = \partial_k w_j$ for $j,k \geqslant 2$.[7] Consequently, there exists a C^2 function H defined in B_n such that $w_j = -\partial_j H$, $j \geqslant 2$. The function H is closely related to the generalized Helmholtz potential function[5] which can be obtained by solving the Pfaffian equation $\Omega = 0$. Hence we consider the following equation;

$$dx^1 = \sum_{j \geqslant 2} w_j(x)\, dx^j, \quad x^1(0) = u, \tag{6}$$

where u is a real number. Fix $\hat{a} = (a^2, \ldots, a^n)$ and set $\hat{x} = \hat{a}t$. By assumption w_j is a C^1 function in an open subset of B_n with compact support. According to the theory of ordinary differential equations, (6) has a unique solution curve $x^1 = \hat{F}(t,u,\hat{a})$ such that $x^1(0) = u$. Since $x^1 = \hat{F}(\alpha t, u, \alpha^{-1} a)$ also satisfies (6),

$$x^1 = \hat{F}(t,u,\hat{a}) = \hat{F}(\alpha t, u, \alpha^{-1} a)$$

is the unique solution of (6). Set $\alpha t = 1$ and denote $\hat{F}(1,u,\hat{a}t) = F(u,\hat{a}t)$. Then

$$x^1 = F(u,\hat{a}t), \quad x^1(0) = F(u,0) = u.$$

The solution of (6) depends continuously on the initial value of u. In order to obtain the solution surface of the Pfaffian equation $\Omega = 0$, we consider u as well as $\hat{a}t$ as a new set of coordinate variables in a neighborhood (nbd) U of the origin ($u=0$, $\hat{a}t=0$). Without loss of generality we denote the new variables $\hat{a}t$ by $\hat{x}^*$. Let $(u,\hat{x}^*)$ be an arbitrary point in U and let D be an open subset of B_n. Define $G: U \to D$ by $G(u,\hat{x}^*) = (x^1, \hat{x})$ where $x^1 = F(u,\hat{x})$, $\hat{x} = \hat{x}^*$. Since the determinant of the Jacobian matrix of G does not vanish at the origin, by the inverse function theorem there exists an open subset $U_1 \subset U$ containing the origin, such that $G|U_1$ has an inverse which is a C^1 function defined on $V = G(U_1)$ containing $G(0) = 0 \in V$. This inverse function is given by

$$G^{-1}(x) = (f(x), \hat{x}) = (u, \hat{x}),$$

that is, $u=f(x)$ if and only if $x^1=F(u,\hat{x})$. We now identify $H=-F$ so that $w_j=\partial_j F$, $j\geqslant 2$. Therefore

$$\Omega=dx^1-\sum_{j\geqslant 2} w_j(x)\,dx^j=(\partial_u F)du+\sum_{j\geqslant 2}(\partial_j F-w_j)\,dx^j=(\partial_u F)\,du,$$

where $\partial_u F=\partial F/\partial u$. Since $(\partial_u F)(x)\neq 0$, we have $du=[(\partial_u F)(x)]^{-1}\Omega$. If we identify $(\partial_u F)\times(x)$ as the thermodynamic temperature T in absolute temperature scale, and $u=S=f(x)$ as the generalized entropy function, then

$$\Omega=TdS=dE+PdV-\sum_i \mu_i\,dN_i+\sum_{i,\alpha} X_i^\alpha\,d\Phi_i^\alpha \tag{7}$$

is the generalized Gibbs one-form (Gibbs relation).

In the next section we consider symmetry transformations of the generalized entropy surface $u=S=f(x)$ that preserve the contact condition specified by the Gibbs one-form (7).

II. INVARIANCE OF THE SECOND LAW

From a mathematical point of view it is evident that Caratheodory's inaccessibility condition (3a) is equivalent to the Gibbs one-form (7). Thus we consider the symmetry transformations of the generalized entropy surface that preserve the one-form (7). Let $(x,u)\in B_n\times R^1$ and consider the transformations

$$x^*=F(x,u),\quad u^*=G(x,u), \tag{8}$$

where G and $F=(F_1,\ldots,F_n)$ are differentiable functions. Furthermore, we assume that the transformation $(x,u)\to(x^*,u^*)$ is one-to-one. By (8) $u=f(x)$ becomes $u^*=f^*(x^*,u^*)$, where f^* is a C^1 function. Denote $y=(y_1,\ldots,y_n)$, $y^*=(y_1^*,\ldots,y_n^*)$, and $y_i=\partial_i f$, $y_i^*=\partial_i^* f^*$, where ∂_i, ∂_u, ∂^i, ∂_i^*, and ∂_u^* are partial differentiation operators with respect to the variables x^i, u, y_i, $(x^*)^i$, and u^*, respectively. Now (8) preserves the contact condition

$$du^*=\sum_i y_i^*\,d(x^*)^i=y_i^*\,d(x^*)^i,$$

if and only if

$$(D_i F)u_j^*=D_i G,\quad D_i=\partial_i+y_i\partial_u\,. \tag{9}$$

Here we have adopted the summation convention over repeated indices. Let

$$A=\begin{bmatrix} D_1F_1 & \cdots & D_1F_n\\ \vdots & & \vdots\\ D_nF_1 & \cdots & D_nF_n\end{bmatrix}$$

and assume that A is invertible. Then (9) can be rewritten as

$$\begin{bmatrix} y_1^*\\ \vdots\\ y_n^*\end{bmatrix}=A^{-1}\begin{bmatrix} D_1G\\ \vdots\\ D_nG\end{bmatrix}. \tag{10}$$

Thus (8) and (10) lead to the following extended transformation $(x,u,y)\to(x^*,u^*,y^*)$ given by

$$x^*=F(x,u),\quad u^*=G(x,u),\quad y^*=H(x,u,y). \tag{11}$$

Suppose the transformations in (11) define a one-parameter Lie group of transformations

$$(x^*)^i = F^i(x,u;\varepsilon), \quad u^* = G(x,u;\varepsilon), \quad y_i^* = H_i(x,u,y;\varepsilon). \tag{12}$$

Consider the infinitesimal transformations of (12):

$$\begin{aligned} (x^*)^i &= x^i + \xi^i(x,u)\varepsilon + O(\varepsilon^2), \\ u^* &= u + \eta(x,u)\varepsilon + O(\varepsilon^2), \\ y_i^* &= y_i + \eta_i(x,u,y)\varepsilon + O(\varepsilon^2). \end{aligned} \tag{13}$$

According to (10) and (13) the infinitesimal generator X of (13) can be obtained easily as[12]

$$X = \xi^i(x,u)\partial_i + \eta(x,u)\partial_u + \eta_i(x,u,y)\partial^i \tag{14}$$

with

$$\eta_i = D_i\eta - (D_i\xi^j)y_j = (\partial_i + y_i\partial_u)g, \tag{15}$$

where $g = \eta - y_j\xi^j$ is a C^1 function of (x,u,y). Hence (3a) is invariant under (13) if and only if (15) is satisfied. We conclude this section with a discussion on Legendre transformations in ET.[1] Let $u = f(x)$ and consider the following transformations:

$$x_i^* = F^i(x,u,y), \quad u^* = G(x,u,y), \quad y_i^* = H_i(x,u,y). \tag{16}$$

In order to preserve the one-form $du^* = y_i^* d(x^*)^i$, y_i^* must satisfy the conditions

$$D_kG = (D_kF^i)y_i^*, \quad \partial^kG = (\partial^kF^i)y_i^*. \tag{17}$$

Now let

$$\begin{aligned} u^* &= G(x,u,y) = -u + \sum_{i=1}^{m} y_i x^i, \quad m \leqslant n, \\ x^* &= F(x,u,y) = (y_1,\ldots,y_m, x^{m+1},\ldots,x^n), \quad m \leqslant n. \end{aligned} \tag{18}$$

By (17) we have

$$y_i^* = \begin{cases} x^i, & i = 1,2,\ldots,m, \\ -y_{m+i}, & i = 1,2,\ldots,n-m. \end{cases} \tag{19}$$

Define the transformation ρ: $R^{2n+1} \to R^{2n+1}$ by

$$\rho(x,u,y) = (x^*,u^*,y^*),$$

where x^*, u^* and y^* are given by (18) and (19). Let $\omega = du - y_i dx^i$ and $\omega^* = du^* - y_i^* d(x^*)^i$. Furthermore, let ρ^* be the pull back of ρ. Then $\rho^*\omega^* = -\omega$ and $\rho^2 = I$ (identity). Hence ρ is a partial Legendre involution. Define

$$L = \{(x,u,y) \in R^{2n+1} | u = f(x), y_i = \partial_i f\}$$

and

$$L^* = \{(x^*,u^*,y^*) \in R^{2n+1} | u^* = f^*(x^*), y_i^* = \partial_i^* f^*\}.$$

Then $L^* = \rho L$. Therefore (3a) is invariant under ρ. Notice that L represents the fundamental equation of state in ET.

The discussions presented in this section can be reformulated in a more geometric setting in terms of contact geometry. This will be discussed in the next section.

III. CONTACT GEOMETRY IN THERMODYNAMICS

In the previous section we considered the symmetry transformations of the generalized entropy surface that preserved the inaccessibility condition of Caratheodory. We also discussed the partial Legendre involutions in ET. These discussions naturally lead to the basic concept of contact geometry we now elucidate. Although some of our discussions in the following overlap somewhat with the work of Mrugala *et al.*,[4] our approach is different. It further clarifies and supplements the work presented in Ref. 4.

Let B_n be the thermodynamic space with coordinate cover $x = (E, V, N_i, \Phi_i^\alpha) = (x^1, x^2, \ldots, x^n)$, and let $G = R \times B_n$ with coordinate cover $(u,x) = (S,x)$, where $u = S = f(x)$ is the generalized entropy function of class C^1 on B_n. Denote the conjugate variables of x by $y = (T^{-1}, PT^{-1}, -\mu_i T^{-1}, X_i^\alpha T^{-1}) = (y_1, y_2, \ldots, y_n)$ so that $y_i = \partial f$. Consider $M = G \times B_n$ with coordinate cover (u,x,y) where u,x,y are independent variables. Define the one-form $\omega = du - y_i dx^i$ in M. For every $x \in B_n$ the vector space $\Delta_x = \{ v \in T_x M \,|\, \langle w(x), v(x) \rangle = w_i v^i = 0; w(x) = w_i(x) dx^i, v(x) = v^i(x) \partial_i \}$ is called the contact hyperplane to M at x. Since $\omega \wedge (d\omega)^n \neq 0$ and $\omega \wedge (d\omega)^{n+1} = 0$, the one-form ω defines a nondegenerate hyperplane distribution $x \rightarrow \Delta_x$, where Δ_x is the kernel of ω. This distribution of the hyperplanes (field of hyperplanes) is called the contact structure of M. If λ is a nowhere vanishing function on M, then $\lambda\omega$ defines the same contact structure of M. The differentiable manifold M equipped with such a one-form ω is called a contact manifold.[13] Thus M is a $(2n+1)$-dimensional manifold which can be identified as $M = T^*(B_n) \times R$, where $T^*(B_n)$ is the contangent bundle of B_n.

Notice that M can also be identified with the one-jet space $J^1(B_n, R)$ from B_n into R, which is a vector bundle with base B_n. The fiber at $x \in B_n$ is $R \times T^*(B_n)$. The jet $j_x^1 f$ is the pair $(f(x), df(x))$ and the canonical projection $\pi: M \rightarrow B_n$ is the mapping $j_x^1 f \rightarrow x$. A section of M is a mapping $\sigma: B_n \rightarrow M$ such that $\pi \circ \sigma(x) = x$ for every $x \in B_n$. Hence the mapping $j^1 f: x \rightarrow j_x^1 f = (f(x), df(x))$ defines a global section of M such that $(j^1 f)^* \omega = 0$. The image of B_n under $j^1 f$ is the one-graph space L given by

$$L = \{ (u,x,y) \in M \,|\, u = f(x), y_i = \partial_i f \}, \tag{20}$$

which is an n-dimensional Legendre submanifold of M. Therefore, the generalized entropy surface $u = f(x)$ can be used to define a Legendre submanifold with $y = (T^{-1}, PT^{-1}, -\mu_i T^{-1}, X_i^a T^{-1})$ as the conjugate variables of the thermodynamics variables x, and Gibbs one-form (7) can be utilized to define the contact one-form ω in M.

Next we consider the transformations in (13) as transformations in the contact manifold M. To this end we let $\eta - y_i \xi^i = g$. Then $\xi^i = -\partial^i g$. In view of (14) and (15) the infinitesimal generator X of (13) can be expressed as

$$\begin{aligned} X &= \xi^i(x,u)\partial_i + \eta(x,u)\partial_u + \eta_i(x,u,y)\partial^i \\ &= g\partial_u + \{ -(y_j \partial^j g)\partial_u + (y_i \partial_u g)\partial^i + [(\partial_i g)\partial^i - (\partial^i g)\partial_i] \} \\ &= X_v + X_H, \end{aligned} \tag{21}$$

with $X_v = g\partial_u$ and X_H given in the curly bracket.

Suppose $\hat{X} = v^i \partial_i$ and $\theta = \theta_i dx^i$. An inner multiplication between $\hat{X}$ and θ is defined by $\hat{X} \lrcorner \theta = i(\hat{X})\theta = v^i \theta_i$. Thus $\partial_i \lrcorner dx^j = \delta_i^j$. The following results can be easily verified:

$$i(X_v)\omega = g, \quad i(X_v)d\omega = 0, \tag{22a}$$

$$i(X_H)\omega=0, \quad i(X_H)d\omega=i(x)d\omega=-Dg, \tag{22b}$$

where $Dg=dg-[i(X_u)dg]\omega$, $X_u=\partial_u$. Thus $X_v\in\ker\omega$ and $X_H\in\ker d\omega$ where $\ker d\omega$ is a vector bundle generated by X_u. Therefore X is a vector field in the tangent bundle TM and the decomposition of X into a vertical component X_v and a horizontal component X_H is unique.

The infinitesimal generator X generates a one-parameter group of transformations $\rho_\tau=T(\tau)=e^{iX\tau}:M\to M$ in (12) with $\varepsilon=\tau$. Let $L(X)$ be the Lie derivative associated with X. Then

$$L(X)\omega=i(X)d\omega+di(X)\omega=(\partial_u g)\omega=\lambda\omega. \tag{23}$$

Hence the contact ideal generated by ω is invariant under ρ_τ. Since $(f,df)\to(f^*,df^*)$ under ρ_τ and the graph of j^1f is a Legendre submanifold of M, thus a Legendre submanifold L is carried into another Legendre submanifold L^* by ρ_τ. By assumption the dissipative energy ΔQ_d is semipositive definite for any irreversible process (independent of ρ_τ); the second law formulated in (3a) and (3b) is therefore invariant under the contact transformations generated by X.

It should be noticed that the construction of the vector field X in (21) depends on an arbitrary function $g(x,u,y)$ defined on M, such that, $\xi^i(x,u)=-\partial^i g$, $\eta(x,u)=g+y_i\xi^i(x,u)$ and $\eta_i=(\partial_i+y_i\partial_u)g$. In fact, there is a one-to-one correspondence between a C^1 function g defined on K with the properties specified above and the vector field $X\in TM$. This one-to-one correspondence gives rise to an isomorphism ψ from the Lie algebra of vector fields in TM onto the vector space of real-valued functions on M, whose Lie algebra structure is defined by the Jacobi bracket $[g_1,g_2]=i([X_{g_1},X_{g_2}])\omega$.

Once X is determined, the flows of X are governed by the following equations:

$$\frac{du}{d\tau}=\eta(x,u), \quad \frac{dx^i}{d\tau}=\xi^i(x,u), \quad \frac{dy_i}{d\tau}=\eta_i(x,u,y). \tag{24}$$

Consider the projection of the flow equations on L:

$$\left.\frac{du}{d\tau}\right|_L=\eta(x,f(x))=(\partial_i f)\left.\frac{dx^i}{d\tau}\right|_L=(\partial_i f)\xi^i(x,f(x)).$$

This implies that $g=\eta-y_i\xi^i$ vanishes on L. Conversely if $g|_L=0$, then the flows of X lie on L. Thus $X_L=\xi^i(x,u)\partial_i+\eta(x,u)\partial_u$ is the projection of X on L, while X is the lift of X_L from TL to TM.

For an illustration of this argument we conclude this paper with an example on the van der Waals fluid with constant heat capacity in ET. Let $x^1=E$, $x^2=V$, $y_1=\partial_1 f=T^{-1}$, $y_2=\partial_2 f=PT^{-1}$, $x=(x^1,x^2)$, and $y=(y_1,y_2)$, where the entropy function S is given by

$$u=S=f(x^1,x^2)=\frac{3}{2}R\ln\left(x^1+\frac{a}{x^2}\right)+R\ln(x^2-b)-\left[\frac{3}{2}R\ln c_v+R\ln R\right]. \tag{25}$$

Here R is the gas constant and c_v is the specific heat at constant volume.

A vector field X_L on the entropy surface can be written as

$$X_L=\eta(x,u)\partial_u+\xi^1(x,u)\partial_1+\xi^2(x,u)\partial_2$$

with $\eta(x,u)=R/(x^2-b)$, $\xi^1=a/(x^2)^2$, and $\xi^2=1$, where a, b are constants. This vector field is constructed such that $i(X_L)\theta=0$, $\theta=du-y_i dx^i$. Next we lift X_L to the contact manifold M with coordinate cover (x^1,x^2,u,y_1,y_2). Define the scalar-valued function g by

$$g(x^1,x^2,u,y_1,y_2)=\eta(x,u)-y_1\xi^1(x,u)-y_2\xi^2(x,u).$$

Then

$$\eta_1=(\partial_1+y_1\partial_u)g=0$$

and

$$\eta_2=(\partial_2+y_2\partial_u)g=-\frac{R}{(x^2-b)^2}+\frac{2ay_1}{(x^2)^3}.$$

Consequently, the vector field X can be written as

$$X=\eta\partial_u+\xi^i\partial_i+\eta_i\partial^i,$$

which yields the following flow equations:

$$\frac{dx^1}{d\tau}=\xi^1(x,u)=\frac{a}{(x^2)^2},\quad \frac{dx^2}{d\tau}=1,$$

$$\frac{dy_1}{d\tau}=\eta_1(x,u,y)=0,\quad \frac{dy_2}{d\tau}=\eta_2(x,u,y)=-\frac{R}{(x^2-b)^2}+\frac{2ay_1}{(x^2)^3},$$

$$\frac{du}{d\tau}=\eta(x,u)=\frac{R}{x^2-b}.$$

We now solve these flow equations. First, $dx^2/d\tau=1$. We set $x^2=\tau$. Then $dx^1=[a/(x^2)^2]dx^2$. So $x^1=-a/x^2+c_1$ or $x^1+a/x^2=c_1$. Next, $dy_1/d\tau=0$. Thus, $y_1=T^{-1}=\text{const}$. Let $c_1=3RT/2$. Then $y_1=\partial_1 f=T^{-1}=\frac{3}{2}R(x^1+a/x^2)^{-1}$, which yields the constitutive relation

$$x^1=E=\frac{3RT}{2}-\frac{a}{V}.$$

The equation for y_2 has the solution $y_2=R/(x^2-b)-ay_1/(x^2)^2+c_2$. If we set $c_2=0$, we immediately obtain the second constitutive relation

$$y_2=PT^{-1}=\partial_2 f=\frac{R}{x^2-b}-\frac{3R}{2}\left(x^1+\frac{a}{x^2}\right)^{-1}\frac{a}{(x^2)^2}$$

or

$$\left(P+\frac{a}{V^2}\right)(V-b)=RT.$$

The equation for u can be rewritten as

$$\begin{aligned}du&=\frac{R}{x^2-b}dx^2=(\partial_1 f)dx^1+(\partial_2 f)dx^2\\&=\frac{3R}{2}\left(x^1+\frac{a}{x^2}\right)^{-1}dx^1+\left[\frac{R}{x^2-b}-\frac{3R}{2}\left(x^1+\frac{a}{x^2}\right)^{-1}\frac{a}{(x^2)^2}\right]dx^2,\end{aligned}\tag{26}$$

which yields the solution given by (25). Furthermore, we can show that g vanishes on the entropy surface $u=S=f(x^1,x^2)$.[14]

This example shows that the flows of X give rise to the constitutive relations. We remark that this example is different from the one given in Ref. 4, which describes an unphysical system since the results there depend on an unbounded continuous parameter. Further, it is incomprehensible that the equation of states for an ideal gas can be transformed into those of a real gas through a sequence of contact transformations ρ_τ as stated in Ref. 4.

Finally, it is interesting to note that for some model equations describing irreversible processes, the dynamical equations of the thermodynamic variables can be embedded in the flows of the vector field X.[15]

[1] H. Callen, *Thermodynamics* (Wiley, New York, 1960).
[2] R. Hermann, *Geometry, Physics, and Systems* (Dekker, New York, 1973).
[3] F. Weinhold, J. Chem. Phys. **63**, 2479 (1975); **63**, 2484 (1975); **63**, 2488 (1975); **63**, 2496 (1975); **65**, 559 (1976); G. Ruppeiner, Phys. Rev. A **20**, 1608 (1979); **27**, 1116 (1983).
[4] R. Mrugala, J. Nulton, J. Schön, and P. Salamon, Phys. Rev. A **41**, 3156 (1990); Rep. Math. Phys. **29**, 109 (1991).
[5] M. Chen and H. C. Tseng, J. Math. Phys. **39**, 329 (1998).
[6] D. Jou, N. Casse-Vasquez, and G. Lebon, *Extended Irreversible Thermodynamics* (Springer, Berlin, 1993); I. Müller and T. Ruggeri, *Extended Thermodynamics* (Springer, Heidelberg, 1993); B. C. Eu, *Kinetic Theory and Irreversible Thermodynamics* (Wiley, New York, 1992).
[7] M. Chen, Phys. Rev. E **56**, 285 (1997); J. Math. Phys. **38**, 3589 (1997); **38**, 5153 (1997).
[8] ΔQ and ΔQ_d can not be expressed only as functions of x.
[9] R. Clausius, Ann. Phys. (Leipzig) **125**, 313 (1965); B. C. Eu was the first to investigate Clausius' inequality based on the notion of uncompensated heat. B. C. Eu, Phys. Rev. E **51**, 768 (1995); also see Ref. 7.
[10] W. Thompson, in *On the Second Law of Thermodynamics*, edited by J. Kestin (Dowden, Hutchison, and Ross, Stroudburg, 1975).
[11] C. Caratheodory, Math. Ann. **67**, 355 (1905); D. G. B. Edelen, *Applied Exterior Calculus* (Wiley, New York, 1985).
[12] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations* (Springer-Verlag, New York, 1989).
[13] P. Liebermann and C. M. Marle, *Symplectic Geometry and Analytic Mechanics* (Reidel, Holland, 1987).
[14] If we make use of the relation $dx^1=[a/(x^2)^2]dx^2$, then (26) reduces to $du=\eta dx^2$.
[15] See Ref. 5. Also, Grmela and Jou have applied the notion of contact geometry in their study of extended kinetic theory. M. Grmela and D. Jou, J. Math. Phys. **34**, 2290 (1993).

Green's function of the Fokker–Planck equation: General formula of frequency expansion

Toru Miyazawa[a)]
Department of Physics, Gakushuin University, Tokyo 171-8588, Japan

(Received 28 April 1998; accepted for publication 12 November 1998)

The one-variable Fokker–Planck equation is studied in its general form by means of an algebraic method. An expression of the Green's function is derived as an expansion in powers of the square root of frequency. The expansion coefficient of arbitrary order is expressed as a functional of the potential in terms of integrals. © *1999 American Institute of Physics.* [S0022-2488(99)03402-7]

I. INTRODUCTION

In this paper we study the one-variable Fokker–Planck equation describing the Brownian motion of a particle in an external potential. It has the form[1]

$$\frac{\partial}{\partial t}P(x;t)=\frac{\partial^2}{\partial x^2}P(x;t)-2\frac{\partial}{\partial x}[f(x)P(x;t)], \tag{1.1}$$

$$f(x)=-\frac{1}{2}\frac{d}{dx}V(x), \tag{1.2}$$

where $P(x;t)$ denotes the probability density of the particle, and $V(x)$ is the external potential. This equation is a fundamental equation describing fluctuations in nonequilibrium phenomena, and it is of considerable importance in various fields of physics.

The Green's function plays the most important role in diffusion problems. We define the Green's function $G(x,x';t)$ as the solution of (1.1) satisfying the initial condition $P(x;t=0)=\delta(x-x')$. It can be interpreted as the transition probability, i.e., the probability density of finding the particle at position x at time t under the condition that it was initially at position x'. Let $G(x,x';\omega)$ denote the Fourier–Laplace transform of $G(x,x';t)$ with respect to time:

$$G(x,x';\omega)=\int_0^\infty e^{i\omega t}G(x,x';t)dt. \tag{1.3}$$

If the potential $V(x)$ is an x-independent constant, the Green's function has the well-known form

$$G(x,x';t)=\frac{1}{\sqrt{4\pi t}}e^{-(x-x')^2/(4t)}, \tag{1.4}$$

and, correspondingly,

$$G(x,x';\omega)=\frac{i}{2\kappa}e^{i\kappa|x-x'|}, \tag{1.5}$$

where κ is defined by

a)Electronic mail: toru.miyazawa@gakushuin.ac.jp

$$\kappa^2 = i\omega, \quad \operatorname{Im} \kappa \geqslant 0. \tag{1.6}$$

We are going to study $G(x,x';\omega)$ for general $V(x)$. For the time being, let us assume that both $V(+\infty) \equiv \lim_{x\to+\infty} V(x)$ and $V(-\infty) \equiv \lim_{x\to-\infty} V(x)$ are finite, and that $V(x)$ tends to these limits sufficiently fast. (We also discuss other cases later.) The Green's function can be expressed as a power series in terms of κ as

$$G(x,x';\omega) = \frac{i}{2\kappa} e^{-V(x)} [p_0 + i\kappa p_1 + (i\kappa)^2 p_2 + (i\kappa)^3 p_3 + \cdots]. \tag{1.7}$$

(The expansion coefficients p_i depend on x and x'. We have singled out the factor $e^{-V(x)}$ so as to make p_i more symmetric.) By Fourier transforming this expression, we can obtain a large-t expansion of $G(x,x';t)$. Thus, the power-series expression (1.7) is important for analyzing the long-time behavior. (See the Appendix for a more detailed description.) The main objective in this paper is to derive an expression of the expansion coefficient p_i for general order i. It is expressed as a functional of $V(x)$ in terms of integrals involving $V(x)$.

A method for calculating p_i was discussed in a previous work.[2] However, only an algorithm for calculation was given there, and, in practice, only the first few terms of expansion could be calculated by that method. In the present paper we use a different, more sophisticated method, and derive a more explicit formula for the expansion coefficient of arbitrary order.

It is well known that the Fokker–Planck equation (1.1) is equivalent to the Schrödinger equation with the potential $V_S(x) \equiv [f(x)]^2 + f'(x)$. It has been customary to study the Fokker–Planck equation by transforming it into the Schrödinger equation, for which various methods of solution are available. The method developed in this paper is different from any existing method for solving the Schrödinger equation. So we may apply this new method to problems described by the Schrödinger equation as well. The frequency expansion formula derived here provides a general and systematical method for the analysis of quantum-mechanical problems in the low-energy region.

We use as our starting point an algebraic expression of the Green's function, which was derived in a previous paper.[3] We review this expression in Sec. II. Sections III and IV are devoted to mathematical preliminaries. Section V contains the main result of the present paper. An expression for the coefficients of expansion is derived here. In Sec. VI we discuss some properties of these expansion coefficients. In Sec. VII the power series expansion is studied in relation to the behavior of the potential at infinity.

II. ALGEBRAIC EXPRESSION

Here we review the algebraic expression of the Green's function, which is the starting point for the analysis in the present paper. Only the result is shown here. The derivation is given in Ref. 3.

We consider a set of operators $J_\pm$, J_3, and $Q_\pm$ satisfying the following commutation relations:

$$[J_3, J_+] = J_+, \quad [J_3, J_-] = -J_-, \quad [J_+, J_-] = 2J_3, \tag{2.1a}$$

and

$$\begin{gathered} [J_3, Q_+] = \tfrac{1}{2} Q_+, \quad [J_+, Q_+] = 0, \quad [J_-, Q_+] = Q_-, \\ [J_3, Q_-] = -\tfrac{1}{2} Q_-, \quad [J_+, Q_-] = Q_+, \quad [J_-, Q_-] = 0, \\ \{Q_+, Q_+\} = -2J_+, \quad \{Q_-, Q_-\} = 2J_-, \quad \{Q_+, Q_-\} = 2J_3, \end{gathered} \tag{2.1b}$$

where $[A,B]=AB-BA$ and $\{A,B\}=AB+BA$. These operators constitute a Lie superalgebra[4] that is called OSp(1/2). (Note that this superalgebra has no relation to the well-known ''super-symmetiric'' structure[5] of the Fokker–Planck equation.) As can be seen from (2.1a), the operators J_3, J_+, and J_- are generators of SU(2) [or its complex form SL(2,**C**)]. As usual, we also use the notation

$$J_1=(J_+ + J_-)/2, \quad J_2=-i(J_+ - J_-)/2. \tag{2.2}$$

The operators $J_\pm$, J_3, and $Q_\pm$ act on vectors belonging to a certain representation space. We define an inner product in the representation space, which we denote by $(\cdots,\cdots)$. We require that it should have the following properties:

$$(\Phi,J_3\Psi)=(J_3\Phi,\Psi), \quad (\Phi,J_\pm\Psi)=-(J_\mp\Phi,\Psi), \quad (\Phi,Q_\pm\Psi)=(Q_\mp\Phi,\Psi), \tag{2.3}$$

where Ψ and Φ are arbitrary vectors. As an inner product, it should also satisfy

$$(\Psi,c_1\Phi_1+c_2\Phi_2)=c_1(\Psi,\Phi_1)+c_2(\Psi,\Phi_2), \quad (\Psi,\Phi)=(\Phi,\Psi)^*, \tag{2.4}$$

but the condition of positive definiteness is not necessary.

We let Ψ_0 denote the lowest state, which is the vector annihilated by Q_-:

$$Q_-\Psi_0=0. \tag{2.5}$$

We define the evolution operator $U(x,x')$ as the solution of

$$\frac{\partial}{\partial x}U(x,x')=[2i\kappa J_3-2f(x)J_1]U(x,x'), \tag{2.6}$$

with the initial condition $U(x=x')=1$ (identity operator). Here $f(x)$ is the function that appears in the Fokker–Planck equation (1.1), and κ is the complex number defined by (1.6). The operator $U(x,x')$ depends on κ, though we do not write it explicitly.

With these definitions, we can express the Green's function of the Fokker–Planck equation as

$$G(x,x';\omega)=\frac{g_0}{2}\frac{(\Psi_0,U(\infty,x)(Q_++Q_-)U(x,x')(Q_++Q_-)U(x',-\infty)\Psi_0)}{(\Psi_0,U(\infty,-\infty)J_3\Psi_0)}, \tag{2.7}$$

with

$$g_0\equiv\frac{i}{2\kappa}e^{-[V(x)-V(x')]/2}. \tag{2.8}$$

[In this paper we assume $x\geqslant x'$ without loss of generality. The expression for $x'\geqslant x$ is obtained by interchanging x and x' in (2.7).] Equation (2.7) holds in any representation as long as (2.1), (2.3), (2.5), and (2.6) are satisfied. We can obtain various expressions for the Green's function by writing (2.7) in specific representations.

III. REPRESENTATION OF SL(2,*C*)

Equation (2.7) is a purely algebraic expression; the operators J_3, $J_\pm$, and $Q_\pm$ in (2.6) and (2.7) can be any operators as long as the commutation relations (2.1) are satisfied. We can obtain a specific expression for the Green's function by choosing a specific form of these operators. We wish to obtain the power series expression (1.7) by writing (2.7) in an appropriate representation. Here and in the next section we make technical preparations for this purpose. (The method we use here is an extension of the techniques employed in Ref. 6 for studying the scattering coefficients.)

Let us first focus our attention on J_3 and $J_\pm$, leaving $Q_\pm$ aside for the time being. We consider a realization of the commutation relations (2.1a) in terms of differential operators of the form

$$J_3^{(\nu)}=\frac{\nu}{2}\cosh w+\sinh w\,\frac{\partial}{\partial w},$$
$$J_\pm^{(\nu)}=\mp\frac{\nu}{2}\sinh w+(1\mp\cosh w)\,\frac{\partial}{\partial w}, \tag{3.1}$$

where ν is a real parameter. These operators act on functions of w, which we assume to be a complex variable; namely, the representation space is the space of analytic functions of w. The background of (3.1), as well as the meaning of w, is explained in Ref. 6. However, for our present purpose it it not necessary to be concerned about the meaning of (3.1). Here it suffices to note only that the operators (3.1) with arbitrary ν indeed satisfy the commutation relations (2.1a). In terms of J_1 and J_2 these expressions read as

$$J_1^{(\nu)}=\frac{\partial}{\partial w},\quad J_2^{(\nu)}=i\,\frac{\nu}{2}\sinh w+i\cosh w\,\frac{\partial}{\partial w},\quad J_3^{(\nu)}=\frac{\nu}{2}\cosh w+\sinh w\,\frac{\partial}{\partial w}. \tag{3.2}$$

Also, let us remark that (3.1) can be rewritten as

$$J_3^{(\nu)}=2\left(\cosh\frac{w}{2}\sinh\frac{w}{2}\right)^{1-\nu/2}\frac{\partial}{\partial w}\left(\cosh\frac{w}{2}\sinh\frac{w}{2}\right)^{\nu/2}, \tag{3.3a}$$

$$J_+^{(\nu)}=-2\left(\sinh\frac{w}{2}\right)^{2-\nu}\frac{\partial}{\partial w}\left(\sinh\frac{w}{2}\right)^{\nu},\quad J_-^{(\nu)}=2\left(\cosh\frac{w}{2}\right)^{2-\nu}\frac{\partial}{\partial w}\left(\cosh\frac{w}{2}\right)^{\nu}. \tag{3.3b}$$

We define the function $u_0^{(\nu)}(w)$ as

$$u_0^{(\nu)}(w)\equiv\left(\cosh\frac{w}{2}\right)^{-\nu}. \tag{3.4}$$

From (3.3b) we find that $u_0^{(\nu)}$ is the lowest state in the representation space satisfying

$$J_-^{(\nu)}u_0^{(\nu)}=0. \tag{3.5}$$

Substituting (3.4) into (3.3a), we obtain

$$J_3^{(\nu)}u_0^{(\nu)}=\frac{\nu}{2}u_0^{(\nu)}. \tag{3.6}$$

This relation means that the quantity $-\nu/2$ corresponds to the spin. Thus, the number ν serves as a label of the representation. In this paper we deal with infinite-dimensional representations with $\nu\geqslant 0$.

We construct the basis $u_n^{(\nu)}$ ($n=0,1,2,...$) by repeatedly applying $J_+^{(\nu)}$ to the lowest state:

$$u_n^{(\nu)}(w)\equiv(-1)^n\frac{1}{n!}(J_+^{(\nu)})^n u_0^{(\nu)}(w)=\binom{\nu+n-1}{n}\left(\cosh\frac{w}{2}\right)^{-\nu}\left(\tanh\frac{w}{2}\right)^{n}, \tag{3.7}$$

where the binomial coefficient is defined as

$$\binom{\nu+n-1}{n}\equiv\frac{\nu(\nu+1)\cdots(\nu+n-1)}{n!}. \tag{3.8}$$

[For $n=0$ we have $\binom{\nu-1}{0}=1$.] We define the inner product between basis states as

$$(u_m^{(\nu)},u_n^{(\nu)})^{(\nu)}\equiv\binom{\nu+n-1}{n}\delta_{mn}. \tag{3.9}$$

[The meaning of the basis (3.7) and the inner product (3.9) is also explained in Ref. 6.] By rescaling the basis as

$$\psi_n^{(\nu)}\equiv\frac{1}{\sqrt{\binom{\nu+n-1}{n}}}u_n^{(\nu)}, \tag{3.10}$$

we can rewrite (3.9) as

$$(\psi_m^{(\nu)},\psi_n^{(\nu)})^{(\nu)}=\delta_{mn}. \tag{3.11}$$

The inner product of arbitrary two states is determined by (3.11) with the property of linearity (2.4).

We let $A^\dagger$ denote the adjoint of A, satisfying

$$(A^\dagger\phi_1,\phi_2)^{(\nu)}=(\phi_1,A\phi_2)^{(\nu)}, \tag{3.12}$$

for arbitrary ϕ_1 and ϕ_2. It is easy to show that the basis states satisfy

$$J_3^{(\nu)}\psi_n^{(\nu)}=\left(\frac{\nu}{2}+n\right)\psi_n^{(\nu)}, \tag{3.13a}$$

$$J_+^{(\nu)}\psi_n^{(\nu)}=-\sqrt{(\nu+n)(n+1)}\,\psi_{n+1}^{(\nu)},\quad J_-^{(\nu)}\psi_n^{(\nu)}=\sqrt{(\nu+n-1)n}\,\psi_{n-1}^{(\nu)}. \tag{3.13b}$$

From (3.13) and (3.11) it is obvious that

$$(J_3^{(\nu)})^\dagger=J_3^{(\nu)},\quad (J_\pm^{(\nu)})^\dagger=-J_\mp^{(\nu)}. \tag{3.14}$$

For an operator A we define its kernel A_K as

$$A_K(w_1,w_2^*)\equiv\sum_{n=0}^{\infty}[A\psi_n^{(\nu)}(w_1)][\psi_n^{(\nu)}(w_2)]^*. \tag{3.15}$$

[Strictly speaking, the expression $A\psi_n^{(\nu)}(w_1)$ should read as $(A\psi_n^{(\nu)})(w_1)$.] By using the completeness of the basis, we can rewrite (3.15) as

$$\begin{aligned}A_K(w_1,w_2^*)&=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\psi_m^{(\nu)}(w_1)(\psi_m^{(\nu)},A\psi_n^{(\nu)})^{(\nu)}[\psi_n^{(\nu)}(w_2)]^*\\&=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\psi_m^{(\nu)}(w_1)(A^\dagger\psi_m^{(\nu)},\psi_n^{(\nu)})^{(\nu)}[\psi_n^{(\nu)}(w_2)]^*\\&=\sum_{n=0}^{\infty}\psi_n^{(\nu)}(w_1)[A^\dagger\psi_n^{(\nu)}(w_2)]^*.\end{aligned} \tag{3.16}$$

The kernel of the identity operator is obtained as

$$I_K^{(\nu)}(w_1,w_2^*)=\sum_{n=0}^{\infty}\psi_n^{(\nu)}(w_1)[\psi_n^{(\nu)}(w_2)]^*=\left[\cosh\left(\frac{w_1-w_2^*}{2}\right)\right]^{-\nu}, \tag{3.17}$$

where we have used (3.10), (3.7), and the formula $\Sigma_n \binom{\nu+n-1}{n}x^n=(1-x)^{-\nu}$. We also define the operator $\overleftarrow{A}$, which operates to the left, as

$$\phi(w)\overleftarrow{A}\equiv(A^\dagger)^*\phi(w). \tag{3.18}$$

[Here the asterisk means that we take complex conjugate of scalars included in the operator. For example, from (3.2) we have $(J_2^{(\nu)})^*=-J_2^{(\nu)}$. Note that w is not replaced by w^*.]

For the convenience of notation, we shall hereafter write w and w' in place of w_1 and w_2^*. The kernel of identity is written as

$$I_K^{(\nu)}(w,w')=\left[\cosh\left(\frac{w-w'}{2}\right)\right]^{-\nu}. \tag{3.19}$$

From (3.15), (3.16), and (3.17) it follows that

$$A_K(w,w')=AI_K^{(\nu)}(w,w')=I_K^{(\nu)}(w,w')\overleftarrow{A}', \tag{3.20}$$

where A' is the operator obtained from A by replacing w with w'. Here we are assuming that the differential operator A acts on w and not on w'; similarly, $\overleftarrow{A}'$ acts on w' and not on w.

Matrix elements of an operator A can be easily obtained from its kernel A_K. In particular, since $\psi_0(0)=1$ and $\psi_n(0)=0$ $(n\neq 0)$, from the first line of (3.16) we find

$$(\psi_0^{(\nu)},A\psi_0^{(\nu)})^{(\nu)}=A_K(w=w'=0). \tag{3.21}$$

Corresponding to (2.6), we define the evolution operator restricted within SL(2,**C**) as the solution of

$$\frac{\partial}{\partial x}U^{(\nu)}(x,x')=[2i\kappa J_3^{(\nu)}-2f(x)J_1^{(\nu)}]U^{(\nu)}(x,x'),\quad U^{(\nu)}(x=x')=1, \tag{3.22}$$

with $J_1^{(\nu)}$ and $J_3^{(\nu)}$ given by (3.2). Just like $J_1^{(\nu)}$ and $J_3^{(\nu)}$, this $U^{(\nu)}(x,x')$ is a differential operator that acts on functions of w. [Note that the operator $U^{(\nu)}$ represents an element of the SL(2,**C**) Lie group.] Here we make a position-dependent shift of the variable,

$$w\rightarrow W\equiv w+V(x), \tag{3.23}$$

and express $U^{(\nu)}(x,x')$ in terms of W rather than w. Under (3.23) the partial derivatives change as

$$\frac{\partial}{\partial x}\rightarrow\frac{\partial}{\partial x}+\frac{dV}{dx}\frac{\partial}{\partial W}=\frac{\partial}{\partial x}-2f(x)\frac{\partial}{\partial W},\quad \frac{\partial}{\partial w}\rightarrow\frac{\partial}{\partial W}, \tag{3.24}$$

and so (3.22) becomes

$$\frac{\partial}{\partial x}U^{(\nu)}(x,x')=i\kappa[e^{-V(x)}\hat{J}_+^{(\nu)}+e^{V(x)}\hat{J}_-^{(\nu)}]U^{(\nu)}(x,x'),\quad U^{(\nu)}(x=x')=1, \tag{3.25}$$

where

$$\hat{J}_+^{(\nu)}\equiv e^{W}\left(\frac{\nu}{2}+\frac{\partial}{\partial W}\right),\quad \hat{J}_-^{(\nu)}\equiv e^{-W}\left(\frac{\nu}{2}-\frac{\partial}{\partial W}\right). \tag{3.26}$$

Let us remark that $\hat{J}_+$ and $\hat{J}_-$, together with $\hat{J}_3 \equiv \partial/\partial W$, satisfy the commutation relations (2.1a). Since $\hat{J}_\pm^{(\nu)} = e^{\pm V(x)}(J_3^{(\nu)} \mp iJ_2^{(\nu)})$, from (3.14) we find

$$(\hat{J}_\pm^{(\nu)})^\dagger = \hat{J}_\pm^{(\nu)}. \tag{3.27}$$

The solution of (3.25) is obtained as[6]

$$U^{(\nu)}(x,x') = \sum_{n=0}^{\infty} (i\kappa)^n \sum_{\{s_i=\pm 1\}} [s_1,\ldots,s_n]_{x'}^{x} \hat{J}_{-s_n}^{(\nu)} \cdots \hat{J}_{-s_1}^{(\nu)}, \tag{3.28}$$

where we have introduced the notation

$$[s_1,s_2,\ldots,s_n]_{x'}^{x} \equiv \int \cdots \int_{x' \leq y_1 \leq y_2 \leq \cdots \leq y_n \leq x} dy_1 \cdots dy_n \exp\left[\sum_{i=1}^{n} s_i V(y_i)\right] \quad (s_i = \pm 1). \tag{3.29}$$

The second summation in (3.28) is over $s_1 = \pm 1$, $s_2 = \pm 1$,...,$s_n = \pm 1$. Here $\hat{J}_{-s}^{(\nu)}$ stands for $\hat{J}_-^{(\nu)}$ and $\hat{J}_+^{(\nu)}$ for $s = +1$ and $s = -1$, respectively. (Note that $\hat{J}_{-s}^{(\nu)}$ with $s = -1$ is not $\hat{J}_1^{(\nu)}$.) The term of order $n = 0$ in (3.28) should be understood as the identity operator. The integrations in (3.29) are carried out in the interval (x',x) with the restriction $y_{i-1} \leq y_i \leq y_{i+1}$.

It is easy to check that the operator $U^{(\nu)}$ given by (3.28) indeed satisfies (3.25). The expression (3.28) is the expansion of the evolution operator in powers of κ. The coefficients of expansion are expressed in terms of integrals defined by (3.29). We shall see later that the expansion coefficients p_i of the Green's function [see Eq. (1.7)] can be expressed in terms of integrals of the form (3.29), too.

Let us consider the kernel of the evolution operator. According to (3.20) we may write

$$U_K^{(\nu)} = U^{(\nu)}(x,x') I_K^{(\nu)}(w,w') = U^{(\nu)}(x,x') \left[\cosh\left(\frac{w-w'}{2}\right)\right]^{-\nu}. \tag{3.30}$$

Since the expression (3.28) [with (3.26)] is written in terms of W, it is natural to express $U_K^{(\nu)}$ as a function of W rather than w. The evolution operator $U^{(\nu)}(x_2,x_1)$ transfers the position from x_1 to x_2; so the position-dependent shift (3.23) implies $W = w + V(x_1)$ before the operation of $U^{(\nu)}(x_2,x_1)$, and $W = w + V(x_2)$ after the operation. Therefore, the variable w in the expression $\cosh[(w-w')/2]$ in Eq. (3.30) is to be replaced by $W - V(x')$, since this w is a quantity before the operation of $U^{(\nu)}(x,x')$. If we define $W' \equiv w' + V(x')$, then (3.30) becomes

$$U_K^{(\nu)}(x,x';W,W') = U^{(\nu)}(x,x') \left[\cosh\left(\frac{W-W'}{2}\right)\right]^{-\nu}. \tag{3.31}$$

Substituting (3.28) into (3.31), we can express $U_K^{(\nu)}$ in terms of W and W', where

$$W = w + V(x), \quad W' = w' + V(x'). \tag{3.32}$$

Now we have $W = w + V(x)$ [not $W = w + V(x')$], since this W is a quantity obtained after the operation of $U^{(\nu)}(x,x')$. By using the last expression of (3.20), we can also write

$$U_K^{(\nu)}(x,x';W,W') = \left[\cosh\left(\frac{W-W'}{2}\right)\right]^{-\nu} \overleftarrow{U}'^{(\nu)}(x,x'), \tag{3.33}$$

where

$$\tilde{U}'^{(\nu)}(x,x')=\sum_{n=0}^{\infty}(i\kappa)^n\sum_{\{s_i=\pm 1\}}[s_1,\ldots,s_n]_{x'}^{x}\hat{\tilde{J}}'^{(\nu)}_{-s_n}\cdots\hat{\tilde{J}}'^{(\nu)}_{-s_1}. \tag{3.34}$$

From (3.27) and the definition (3.18), we can see that the operators $\hat{\tilde{J}}'^{(\nu)}_{\pm}$ act as

$$g(W')\hat{\tilde{J}}'^{(\nu)}_{\pm}=\hat{J}'^{(\nu)}_{\pm}g(W'), \tag{3.35}$$

for arbitrary $g(W')$.

Matrix elements of the evolution operator can be obtained from $U_K^{(\nu)}$. Since $U_K^{(\nu)}$ is now expressed as a function of W and W', Eq. (3.21) reads as

$$(\psi_0^{(\nu)},U^{(\nu)}(x,x')\psi_0^{(\nu)})^{(\nu)}=U_K^{(\nu)}(x,x';W=V(x),W'=V(x')). \tag{3.36}$$

IV. EXTENSION TO SUPERALGEBRA

Now let us extend the arguments of the previous section to the superalgebra, taking into account Q_+ and Q_- in addition to J_3 and $J_\pm$. The operators Q_+ and Q_- raises and lowers, respectively, the value of the spin by $\frac{1}{2}$. In the previous section we labeled the representation by a number ν, and, as explained below Eq. (3.6), the spin corresponds to $-\nu/2$. So $Q_\pm$ changes the value of ν by one. We can construct a representation of the superalgebra (2.1) by putting together two representations of the SL(2,**C**) algebra, with labels ν and $\nu+1$. The operators J_3, $J_\pm$, and $Q_\pm$ can be expressed in the form of 2×2 matrices as

$$\mathsf{J}_a^{(\nu,\nu+1)}=\begin{pmatrix} J_a^{(\nu)} & 0\\ 0 & J_a^{(\nu+1)}\end{pmatrix}\quad (a=1,2,3,\ \text{or}\ \pm), \tag{4.1a}$$

$$\mathsf{Q}_+^{(\nu,\nu+1)}=\begin{pmatrix} 0 & \nu^{1/2}\sinh(w/2)\\ -[\nu^{1/2}\sinh(w/2)]^{-1}J_+^{(\nu)} & 0\end{pmatrix}, \tag{4.1b}$$

$$\mathsf{Q}_-^{(\nu,\nu+1)}=\begin{pmatrix} 0 & \nu^{1/2}\cosh(w/2)\\ [\nu^{1/2}\cosh(w/2)]^{-1}J_-^{(\nu)} & 0\end{pmatrix}, \tag{4.1c}$$

where $J_a^{(\nu)}$ are the operators defined by (3.1). It is easy to check by direct calculation that the operators (4.1) satisfy the commutation relations (2.1). From now on we work with the representation given by (4.1). Although we shall hereafter drop the representation label $(\nu,\nu+1)$, it should not be forgotten that these expressions depend on ν. We write J and Q instead of J and Q in order to stress that they are 2×2 matrices of the form (4.1).

Now the representation space consists of two-component column vectors, where each component is an analytic function of w. The basis $\boldsymbol{\Psi}_n$ $(n=0,1,2,...)$ is given by

$$\boldsymbol{\Psi}_{2m}=\begin{pmatrix}\psi_m^{(\nu)}\\ 0\end{pmatrix},\quad \boldsymbol{\Psi}_{2m+1}=\begin{pmatrix}0\\ \psi_m^{(\nu+1)}\end{pmatrix}\quad (m=0,1,2,...), \tag{4.2}$$

with ψ_m defined by (3.10) and (3.7). From (3.5) it is obvious that $\boldsymbol{\Psi}_0$ is the lowest state satisfying $\mathsf{Q}_-\boldsymbol{\Psi}_0=0$.

We define the inner product between basis vectors as

$$(\boldsymbol{\Psi}_m,\boldsymbol{\Psi}_n)=\delta_{mn}. \tag{4.3}$$

Then it follows that the inner product of arbitrary two vectors $\boldsymbol{\Phi}_a=\binom{a_1}{a_2}$ and $\boldsymbol{\Phi}_b=\binom{b_1}{b_2}$ is

$$(\boldsymbol{\Phi}_a,\boldsymbol{\Phi}_b)=(a_1,b_1)^{(\nu)}+(a_2,b_2)^{(\nu+1)}, \tag{4.4}$$

with $(\cdots,\cdots)^{(\nu)}$ defined by (3.11). We let $\mathsf{A}^\dagger$ denote the adjoint of A, satisfying

$$(\mathsf{A}^\dagger \mathbf{\Phi}_a, \mathbf{\Phi}_b) = (\mathbf{\Phi}_a, \mathsf{A}\mathbf{\Phi}_b). \tag{4.5}$$

Since we have

$$\begin{aligned} &\mathsf{Q}_+ \mathbf{\Psi}_{2m} = \sqrt{\nu+m}\,\mathbf{\Psi}_{2m+1}, \quad \mathsf{Q}_+ \mathbf{\Psi}_{2m-1} = \sqrt{m}\,\mathbf{\Psi}_{2m}, \\ &\mathsf{Q}_- \mathbf{\Psi}_{2m+1} = \sqrt{\nu+m}\,\mathbf{\Psi}_{2m}, \quad \mathsf{Q}_- \mathbf{\Psi}_{2m} = \sqrt{m}\,\mathbf{\Psi}_{2m-1}, \\ &\mathsf{J}_3 \mathbf{\Psi}_{2m} = \left(\frac{\nu}{2}+m\right)\mathbf{\Psi}_{2m}, \quad \mathsf{J}_3 \mathbf{\Psi}_{2m+1} = \left(\frac{\nu+1}{2}+m\right)\mathbf{\Psi}_{2m+1}, \end{aligned} \tag{4.6}$$

it is obvious that

$$(\mathsf{J}_3)^\dagger = \mathsf{J}_3, \quad (\mathsf{J}_\pm)^\dagger = -\mathsf{J}_\mp, \quad (\mathsf{Q}_\pm)^\dagger = \mathsf{Q}_\mp, \tag{4.7}$$

where we have also used $(\mathsf{Q}_\pm)^2 = \mp \mathsf{J}_\pm$. Thus, the inner product defined by (4.3) satisfies the requirements (2.3).

Let us make some definitions parallel to the ones in the last section. The kernel of a 2×2 operator matrix A is defined as

$$\mathsf{A}_K(w_1, w_2^*) \equiv \sum_n [\mathsf{A}\mathbf{\Psi}_n(w_1)][\mathbf{\Psi}_n(w_2)]^{*T} = \sum_n \mathbf{\Psi}_n(w_1)[\mathsf{A}^\dagger \mathbf{\Psi}_n(w_2)]^{*T}. \tag{4.8}$$

Here T denotes transposition; $(\mathbf{\Psi}_n)^{*T}$ is a two-component row vector, and so A_K is a 2×2 matrix. As before, we shall write w and w' in place of w_1 and w_2^*. It is easy to see that the kernel of the identity operator is

$$\mathsf{I}_K(w,w') = \begin{pmatrix} I_K^{(\nu)}(w,w') & 0 \\ 0 & I_K^{(\nu+1)}(w,w') \end{pmatrix}, \tag{4.9}$$

with $I_K^{(\nu)}$ given by (3.19). We define $\overleftarrow{\mathsf{A}}$ by

$$\mathbf{\Phi}^T \overleftarrow{\mathsf{A}} \equiv [(\mathsf{A}^\dagger)^* \mathbf{\Phi}]^T. \tag{4.10}$$

[Note that $(\mathsf{A}^\dagger)^* \neq \mathsf{A}^T$, since the components of A are differential operators.] As an extension of (3.20), we have

$$\mathsf{A}_K(w,w') = \mathsf{A}\mathsf{I}_K(w,w') = \mathsf{I}_K(w,w')\overleftarrow{\mathsf{A}}', \tag{4.11}$$

where A' is the operator obtained from A by replacing w with w'. Once again, let us notice that the operator A acts on w and not on w', whereas $\overleftarrow{\mathsf{A}}'$ acts on w' and not on w. Corresponding to (3.21), we have

$$(\mathbf{\Psi}_0, \mathsf{A}\mathbf{\Psi}_0) = (\mathsf{A}_K)_{11}(w = w' = 0), \tag{4.12}$$

where $(\mathsf{A}_K)_{11}$ denotes the upper-left component of the 2×2 matrix A_K.

The evolution operator U, which is the solution of

$$\frac{\partial}{\partial x}\mathsf{U}(x,x') = [2i\kappa \mathsf{J}_3 - 2f(x)\mathsf{J}_1]\mathsf{U}(x,x'), \quad \mathsf{U}(x=x') = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{4.13}$$

is simply obtained as

$$\mathsf{U}=\begin{pmatrix} U^{(\nu)} & 0 \\ 0 & U^{(\nu+1)} \end{pmatrix}. \tag{4.14}$$

Now we are ready to write out the expression (2.7) in our representation given by (4.1).

V. POWER SERIES EXPRESSION

We define

$$\mathsf{S}(x,x';z,z')\equiv \mathsf{U}(z,x)(\mathsf{Q}_+ + \mathsf{Q}_-)\mathsf{U}(x,x')(\mathsf{Q}_+ + \mathsf{Q}_-)\mathsf{U}(x',z'), \tag{5.1}$$

where $\mathsf{Q}_\pm$ and U are the 2×2 operator matrices introduced in the last section. The relation (4.11) gives the kernel of S as

$$\mathsf{S}_K(x,x';z,z';w,w')=\mathsf{S}(x,x';z,z')\mathsf{I}_K(w,w'). \tag{5.2}$$

We let S_0 denote the upper-left component of the matrix S_K:

$$S_0\equiv(\mathsf{S}_K)_{11}. \tag{5.3}$$

Now let us write Eq. (2.7) in the representation (4.1). We can express it in terms of S as

$$G(x,x')=\frac{g_0}{2}\frac{(\mathbf{\Psi}_0,\mathsf{S}(x,x';\infty,-\infty)\mathbf{\Psi}_0)}{(\mathbf{\Psi}_0,\mathsf{U}(\infty,-\infty)\mathsf{J}_3\mathbf{\Psi}_0)}=\frac{g_0}{\nu}\frac{(\mathbf{\Psi}_0,\mathsf{S}(x,x';\infty,-\infty)\mathbf{\Psi}_0)}{(\mathbf{\Psi}_0,\mathsf{U}(\infty,-\infty)\mathbf{\Psi}_0)}, \tag{5.4}$$

where we have used $\mathbf{J}_3\mathbf{\Psi}_0=(\nu/2)\mathbf{\Psi}_0$. [Here and hereafter $G(x,x')$ stands for $G(x,x';\omega)$.] More precisely, Eq. (5.4) means

$$G(x,x')=\lim_{\substack{z\to\infty\\ z'\to-\infty}}\frac{g_0}{\nu}\frac{(\mathbf{\Psi}_0,\mathsf{S}(x,x';z,z')\mathbf{\Psi}_0)}{(\mathbf{\Psi}_0,\mathsf{U}(z,z')\mathbf{\Psi}_0)}. \tag{5.5}$$

From (4.14), (4.2), and (4.4) we have

$$(\mathbf{\Psi}_0,\mathsf{U}\mathbf{\Psi}_0)=(\psi_0^{(\nu)},U^{(\nu)}\psi_0^{(\nu)})^{(\nu)}. \tag{5.6}$$

It can be shown that[7]

$$(\psi_0^{(\nu)},U^{(\nu)}\psi_0^{(\nu)})^{(\nu)}=[(\psi_0^{(1)},U^{(1)}\psi_0^{(1)})^{(1)}]^{\nu}, \tag{5.7}$$

and hence

$$\lim_{\nu\to 0}(\mathbf{\Psi}_0,\mathsf{U}\mathbf{\Psi}_0)=\lim_{\nu\to 0}(\psi_0^{(\nu)},U^{(\nu)}\psi_0^{(\nu)})^{(\nu)}=1. \tag{5.8}$$

Since (5.5) holds for any ν, we may take the limit $\nu\to 0$ in (5.5). We obtain

$$G(x,x')=\lim_{\substack{z\to\infty\\ z'\to-\infty}}\lim_{\nu\to 0}\frac{g_0}{\nu}(\mathbf{\Psi}_0,\mathsf{S}(x,x';z,z')\mathbf{\Psi}_0)=\lim_{\substack{z\to\infty\\ z'\to-\infty}}\lim_{\nu\to 0}\frac{g_0}{\nu}S_0(x,x';z,z';w=w'=0), \tag{5.9}$$

where we have used (4.12) and (5.3). By using (4.11), we can write (5.2) as

$$\begin{aligned}\mathsf{S}_K&=\mathsf{U}(z,x)(\mathsf{Q}_+ + \mathsf{Q}_-)\mathsf{U}(x,x')(\mathsf{Q}_+ + \mathsf{Q}_-)\mathsf{U}(x',z')\mathsf{I}_K(w,w')\\ &=\mathsf{U}(z,x)(\mathsf{Q}_+ + \mathsf{Q}_-)\mathsf{U}(x,x')(\mathsf{Q}_+ + \mathsf{Q}_-)\mathsf{I}_K(w,w')\tilde{\mathsf{U}}'(x',z'),\end{aligned} \tag{5.10}$$

where

$$\overleftarrow{\mathsf{U}}' = \begin{pmatrix} \overleftarrow{U}'^{(\nu)} & 0 \\ 0 & \overleftarrow{U}'^{(\nu+1)} \end{pmatrix}. \tag{5.11}$$

Adding (4.1b) and (4.1c), we have

$$\mathsf{Q}_+ + \mathsf{Q}_- = \begin{pmatrix} 0 & \nu^{1/2} e^{w/2} \\ \nu^{-1/2} e^{w/2}\left(\nu + 2\dfrac{\partial}{\partial w}\right) & 0 \end{pmatrix}. \tag{5.12}$$

From (3.19), we can easily see that

$$\left(\nu + 2\frac{\partial}{\partial w}\right) I_K^{(\nu)} = \nu e^{(w'-w)/2} I_K^{(\nu+1)}. \tag{5.13}$$

Therefore, from (5.12) and (4.9) we find

$$(\mathsf{Q}_+ + \mathsf{Q}_-)\mathsf{I}_K = \begin{pmatrix} 0 & \nu^{1/2} e^{w/2} I_K^{(\nu+1)} \\ \nu^{1/2} e^{w'/2} I_K^{(\nu+1)} & 0 \end{pmatrix}. \tag{5.14}$$

Substituting (4.14), (5.11), (5.12), and (5.14) into the last expression of (5.10), we obtain the upper-left component of S_K as

$$S_0 = \nu U^{(\nu)}(z,x) e^{w/2} U^{(\nu+1)}(x,x') I_K^{(\nu+1)}(w,w') e^{w'/2} \overleftarrow{U}'^{(\nu)}(x',z'). \tag{5.15}$$

It is convenient to change the variables from (w,w') to (W,W'), as we did for U_K in Sec. III, and express everything in terms of W and W' rather than w and w'. Using (3.30) we can write (5.15) in the form

$$\begin{aligned} S_0 &= \nu U^{(\nu)}(z,x) e^{w/2} U_K^{(\nu+1)}(x,x';W,W') e^{w'/2} \overleftarrow{U}'^{(\nu)}(x',z') \\ &= \nu U^{(\nu)}(z,x) e^{[W-V(x)]/2} U_K^{(\nu+1)}(x,x';W,W') e^{[W'-V(x')]/2} \overleftarrow{U}'^{(\nu)}(x',z'). \end{aligned} \tag{5.16}$$

Substituting (3.31) and (3.33) gives

$$\begin{aligned} S_0 &= \nu e^{-[V(x)+V(x')]/2} U^{(\nu)}(z,x) e^{W/2} U^{(\nu+1)}(x,x') \left(\frac{1}{\cosh[(W-W')/2]}\right)^{\nu+1} e^{W'/2} \overleftarrow{U}'^{(\nu)}(x',z') \\ &= \nu e^{-[V(x)+V(x')]/2} U^{(\nu)}(z,x) e^{W/2} \left(\frac{1}{\cosh[(W-W')/2]}\right)^{\nu+1} \overleftarrow{U}'^{(\nu+1)}(x,x') e^{W'/2} \overleftarrow{U}'^{(\nu)}(x',z'). \end{aligned} \tag{5.17}$$

The operators U and $\overleftarrow{U}'$ are expressed in terms of W and W' as (3.28) and (3.34). Substituting them into (5.17), we obtain S_0 as a function of W and W'. As mentioned before, the operator $U^{(\nu)}(z,x)$ transfers the position from x to z. So we have $W = w + V(z)$ after the operation of $U^{(\nu)}(z,x)$. Similarly, we have $W' = w' + V(z')$ after the operation of $\overleftarrow{U}'^{(\nu)}(x',z')$. Therefore, in the final expression of $S_0(x,x';z,z')$, the relation between (W,W') and (w,w') is

$$W = w + V(z), \quad W' = w' + V(z'). \tag{5.18}$$

Now (5.9) reads as

$$G(x,x')=\lim_{\substack{z\to\infty\\ z'\to-\infty}}\lim_{\nu\to0}\frac{g_0}{\nu}S_0(x,x';z,z';W=V(z),W'=V(z')).\tag{5.19}$$

We define

$$\overline{G}(x,x';z,z';W,W')=\lim_{\nu\to0}\frac{g_0}{\nu}S_0(x,x';z,z';W,W').\tag{5.20}$$

This $\overline{G}$ can be interpreted as a generalized form of the Green's function $G(x,x')$, and hence we may deal with $\overline{G}$ instead of G. The original Green's function is recovered as

$$G(x,x')=\overline{G}(x,x';\infty,-\infty;V(\infty),V(-\infty)).\tag{5.21}$$

As a matter of fact, it can be shown that $\overline{G}(x,x',z,z';W,W')$ is the Green's function for the Fokker-Planck equation with the potential $V(x)$ replaced by

$$\overline{V}(x)=\begin{cases}W' & (x<z'),\\ V(x) & (z'<x<z),\\ W & (z<x).\end{cases}\tag{5.22}$$

For a detailed explanation on this interpretation, see Ref. 6.

We obtain $\overline{G}$ by substituting (5.17) into (5.20). The factor ν cancels out, and so we can set $\nu=0$ in the rest of the expression. As a result, we have

$$\begin{aligned}\overline{G}(x,x';z,z';W,W')&=\frac{i}{2\kappa}e^{-V(x)}U^{(0)}(z,x)e^{W/2}U^{(1)}(x,x')\frac{1}{\cosh[(W-W')/2]}e^{W'/2}\overleftarrow{U}'^{(0)}(x',z')\\ &=\frac{i}{2\kappa}e^{-V(x)}U^{(0)}(z,x)e^{W/2}\frac{1}{\cosh[(W-W')/2]}\overleftarrow{U}'^{(1)}(x,x')e^{W'/2}\overleftarrow{U}'^{(0)}(x',z').\end{aligned}\tag{5.23}$$

Substituting (3.28) and (3.34) into (5.23), we obtain

$$\begin{aligned}\overline{G}=\frac{i}{2\kappa}e^{-V(x)}&\sum_{n_1,n_2,n_3=0}^{\infty}\sum_{\{a_i,b_i,c_i=\pm1\}}(i\kappa)^{n_1+n_2+n_3}C(a_1,\dots,a_{n_1};b_1,\dots,b_{n_2};c_1,\dots,c_{n_3})\\ &\times[a_1,\dots,a_{n_1}]_{z'}^{x'}[b_1,\dots,b_{n_2}]_{x'}^{x}[c_1,\dots,c_{n_3}]_{x}^{z},\end{aligned}\tag{5.24}$$

where

$$\begin{aligned}&C(a_1,\dots,a_{n_1};b_1,\dots,b_{n_2};c_1,\dots,c_{n_3})\\ &=\hat{J}^{(0)}_{-c_{n_3}}\cdots\hat{J}^{(0)}_{-c_1}e^{W/2}\hat{J}^{(1)}_{-b_{n_2}}\cdots\hat{J}^{(1)}_{-b_1}\frac{1}{\cosh[(W-W')/2]}e^{W'/2}\overleftarrow{\hat{J}}'^{(0)}_{-a_{n_1}}\cdots\overleftarrow{\hat{J}}'^{(0)}_{-a_1}\\ &=\hat{J}^{(0)}_{-c_{n_3}}\cdots\hat{J}^{(0)}_{-c_1}e^{W/2}\frac{1}{\cosh[(W-W')/2]}\overleftarrow{\hat{J}}'^{(1)}_{-b_{n_2}}\cdots\overleftarrow{\hat{J}}'^{(1)}_{-b_1}e^{W'/2}\overleftarrow{\hat{J}}'^{(0)}_{-a_{n_1}}\cdots\overleftarrow{\hat{J}}'^{(0)}_{-a_1}.\end{aligned}\tag{5.25}$$

The second summation in (5.24) is over $a_1=\pm 1$, $a_2=\pm 1$,...,$a_{n_1}=\pm 1$, and so on. For $n_1=0$, the expression $[a_1,\ldots,a_{n_1}]_{z'}^{x'}$ in (5.24) is replaced by 1, and the corresponding product of operators in (5.25) is interpreted as the identity operator (and similarly for the cases $n_2=0$ and $n_3=0$). Using (3.35) we can rewrite (5.25) as

$$
\begin{aligned}
C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3})
&=\hat{J}^{(0)}_{-c_{n_3}}\cdots\hat{J}^{(0)}_{-c_1}e^{W/2}\hat{J}^{(1)}_{-b_{n_2}}\cdots\hat{J}^{(1)}_{-b_1}\hat{J}'^{(0)}_{-a_1}\cdots\hat{J}'^{(0)}_{-a_{n_1}}e^{W'/2}\frac{1}{\cosh[(W-W')/2]}\\
&=\hat{J}'^{(0)}_{-a_1}\cdots\hat{J}'^{(0)}_{-a_{n_1}}e^{W'/2}\hat{J}'^{(1)}_{-b_1}\cdots\hat{J}'^{(1)}_{-b_{n_2}}\hat{J}^{(0)}_{-c_{n_3}}\cdots\hat{J}^{(0)}_{-c_1}e^{W/2}\frac{1}{\cosh[(W-W')/2]},
\end{aligned}
\tag{5.26}
$$

although (5.25) is more transparent in structure. The expression (5.24) with (5.25) is the power series expansion of the Green's function in terms of κ; thus, we have achieved our main purpose.

It is also possible to express the coefficients C entirely in terms of $\hat{J}$, without $\overleftarrow{\hat{J}}'$ or $\hat{J}'$. Let us assume $n_1\neq 0$. It can be shown that[8]

$$
\begin{aligned}
&\hat{J}'^{(0)}_{-a_1}\hat{J}'^{(0)}_{-a_2}\cdots\hat{J}'^{(0)}_{-a_{n_1}}\tanh[(W-W')/2]\\
&\quad=\hat{J}^{(2)}_{-a_{n_1}}\cdots\hat{J}^{(2)}_{-a_2}\hat{J}^{(2)}_{-a_1}\tanh[(W-W')/2]+\hat{J}^{(2)}_{-a_{n_1}}\cdots\hat{J}^{(2)}_{-a_2}a_1e^{-a_1W}.
\end{aligned}
\tag{5.27}
$$

Since

$$
\hat{J}'^{(0)}_{\pm}\frac{e^{W'/2}}{\cosh[(W-W')/2]}=-e^{W/2}\hat{J}'^{(0)}_{\pm}\tanh[(W-W')/2],
\tag{5.28}
$$

from (5.26) and (5.27) we obtain

$$
\begin{aligned}
&C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3})\\
&\quad=\hat{J}^{(0)}_{-c_{n_3}}\cdots\hat{J}^{(0)}_{-c_1}e^{W/2}\hat{J}^{(1)}_{-b_{n_2}}\cdots\hat{J}^{(1)}_{-b_1}e^{W/2}\hat{J}^{(2)}_{-a_{n_1}}\cdots\hat{J}^{(2)}_{-a_2}\{-a_1e^{-a_1W}-\hat{J}^{(2)}_{-a_1}\tanh[(W-W')/2]\}.
\end{aligned}
\tag{5.29}
$$

For $n_1=0$ we have simply

$$
C(;b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3})=\hat{J}^{(0)}_{-c_{n_3}}\cdots\hat{J}^{(0)}_{-c_1}e^{W/2}\hat{J}^{(1)}_{-b_{n_2}}\cdots\hat{J}^{(1)}_{-b_1}e^{W'/2}\frac{1}{\cosh[(W-W')/2]}.
\tag{5.30}
$$

VI. PROPERTIES OF THE EXPANSION COEFFICIENTS

From (5.25), we can easily see that

$$
\begin{aligned}
C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3},+)&=-e^{-W}\frac{\partial}{\partial W}C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3}),\\
C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3},-)&=e^{W}\frac{\partial}{\partial W}C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3}),\\
C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2},+;)&=e^{-W}\left(1-\frac{\partial}{\partial W}\right)C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};),
\end{aligned}
\tag{6.1}
$$

$$C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2},-;)=e^W\frac{\partial}{\partial W}C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};),$$

where we have written "+" and "−" in place of "+1" and "−1" for simplicity. Here $C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};)$ denotes the coefficients for $n_3=0$. Similarly, we have

$$\begin{aligned}
C(+,\ldots;\ldots;\ldots)&=-e^{-W'}\frac{\partial}{\partial W'}C(\ldots;\ldots;\ldots),\\
C(-,\ldots;\ldots;\ldots)&=e^{W'}\frac{\partial}{\partial W'}C(\ldots;\ldots;\ldots),\\
C(;+,\ldots;\ldots)&=e^{-W'}\left(1-\frac{\partial}{\partial W'}\right)C(;\ldots;\ldots),\\
C(;-,\ldots;\ldots)&=e^{W'}\frac{\partial}{\partial W'}C(;\ldots;\ldots),
\end{aligned}\tag{6.2}$$

where $C(;\ldots;\ldots)$ stands for the coefficients with $n_1=0$. The coefficient for $n_1=n_2=n_3=0$ is

$$C(;;)=\frac{e^{(W+W')/2}}{\cosh[(W-W')/2]}=\frac{2}{e^{-W}+e^{-W'}}.\tag{6.3}$$

We can calculate the coefficients $C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3})$ by using the relations (6.1) and (6.2), starting from (6.3). It is evident that these coefficients have the left–right symmetry,

$$C(a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3};W,W')=C(c_{n_3},\ldots,c_1;b_{n_2},\ldots,b_1;a_{n_1},\ldots,a_1;W',W),\tag{6.4}$$

where we have written out the dependence on W and W' explicitly. It is also obvious that

$$C(\ldots;\ldots;\ldots,+)=-e^{-2W}C(\ldots;\ldots;\ldots,-)\tag{6.5a}$$

and

$$C(+,\ldots;\ldots;\ldots)=-e^{-2W'}C(-,\ldots;\ldots;\ldots).\tag{6.5b}$$

Finally, from (6.1) and (6.2) it follows that

$$C(\ldots,-;\ldots;\ldots)=C(\ldots;-,\ldots;\ldots),\quad C(\ldots;\ldots,-;\ldots)=C(\ldots;\ldots;-,\ldots).\tag{6.6}$$

Namely, a minus sign can pass through a semicolon. (This does not hold for a plus sign.)

Now let us write down the first few terms of the expansion (5.24) explicitly. We define the notation

$$[a_1,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3}]\equiv[a_1,\ldots,a_{n_1}]_{z'}^{x'}[b_1,\ldots,b_{n_2}]_{x'}^{x}[c_1,\ldots,c_{n_3}]_x^z.\tag{6.7}$$

If $n_1=0$, the quantity $[a_1,\ldots,a_{n_1}]_{z'}^{x'}$ on the right-hand side should be interpreted as unity, and similarly for $n_2=0$ or $n_3=0$; for example, $[+;;-]=[+]_{z'}^{x'}[-]_x^z$ and $[;+-;]=[+-]_{x'}^x$. Using this notation, we can write

$$\overline{G}(x,x';z,z';W,W')=\frac{i}{2\kappa}e^{-V(x)}[\overline{p_0}+i\kappa\overline{p_1}+(i\kappa)^2\overline{p_2}+(i\kappa)^3\overline{p_3}+\cdots],\tag{6.8}$$

with

$$\overline{p_0}=C(;;)=\frac{2}{e^{-W}+e^{-W'}}, \tag{6.9a}$$

$$\begin{aligned}\overline{p_1}&=C(-;;)[-;;]+C(;-;)[;-;]+C(;;-)[;;-]\\&\quad+C(+;;)[+;;]+C(;+;)[;+;]+C(;;+)[;;+]\\&=\frac{\overline{p_0}^2}{2}\{[-]_{z'}^{x'}+[-]_{x'}^{x}+[-]_{x}^{z}-e^{-2W'}[+]_{z'}^{x'}+e^{-W-W'}[+]_{x'}^{x}-e^{-2W}[+]_{x}^{z}\},\end{aligned} \tag{6.9b}$$

$$\begin{aligned}\overline{p_2}=\frac{\overline{p_0}^3}{2}\Bigg(&[--;;]+[-;-;]+[-;;-]+[;--;]+[;-;-]+[;;--]-e^{-2W'}\{[+-;;]\\&+[+;-;]+[+;;-]\}-e^{-2W}\{[;;-+]+[;-;+]+[-;;+]\}-\frac{e^{-W}-e^{-W'}}{2}e^{-W}\{[-;+;]\\&+[;-+;]\}+\frac{e^{-W}-e^{-W'}}{2}e^{-W'}\{[;+;-]+[;+-;]\}+e^{-W-W'}\{[-+;;]+[;;+-]\}\\&-e^{-W-3W'}[++;;]-e^{-W'-3W}[;;++]+\frac{e^{-W}-e^{-W'}}{2}e^{-W-2W'}[+;+;]\\&-\frac{e^{-W}-e^{-W'}}{2}e^{-W'-2W}[;+;+]+e^{-2W-2W'}\{[+;;+]+[;++;]\}\Bigg).\end{aligned} \tag{6.9c}$$

For example, the coefficient of $[;-+;]$ in (6.9c) has been calculated by using (6.1) as

$$\begin{aligned}C(;-+;)&=e^{-W}\left(1-\frac{\partial}{\partial W}\right)C(;-;)\\&=e^{-W}\left(1-\frac{\partial}{\partial W}\right)e^{W}\frac{\partial}{\partial W}C(;;)\\&=e^{-W}\left(1-\frac{\partial}{\partial W}\right)e^{W}\frac{\partial}{\partial W}\overline{p_0}=-\frac{1}{4}e^{-W}(e^{-W}-e^{-W'})\overline{p_0}^3.\end{aligned} \tag{6.10}$$

In such a calculation, it is convenient to use the property

$$\left(e^{W}\frac{\partial}{\partial W}\right)^{n}\overline{p_0}=\frac{n!}{2^n}\overline{p_0}^{n+1}, \tag{6.11}$$

which can be easily proved.

VII. THE LIMITING PROCESS

The original Green's function $G(x,x')$ is obtained from $\overline{G}(x,x';z,z';W,W')$ given by (5.24) by taking the limit $z\to\infty$, $z'\to-\infty$, $W\to V(\infty)$, and $W'\to V(-\infty)$. [We are assuming that both $V(\infty)$ and $V(-\infty)$ are finite.] Accordingly, the expansion coefficients p_0, p_1, $p_2,\ldots$ [Eq. (1.7)] are obtained from $\overline{p_0}$, $\overline{p_1}$, $\overline{p_2},\ldots$ [Eq. (6.8)] in this limit. This is not a trivial matter since the power series expansion is not necessarily commutable with the limiting process. We have to make sure that the coefficients $\overline{p_i}$ thus obtained are indeed finite.

Let us look at the coefficient $\overline{p_1}$ given by (6.9b). The integrals $[-]_{z'}^{x'}$, $[-]_x^z$, $[+]_{z'}^{x'}$, and $[+]_x^z$ all diverge in the limit $z'\to-\infty$, $z\to+\infty$. However, if we first set $W=V(\infty)$, $W'=V(-\infty)$ and then take the limit $z'\to-\infty$, $z\to+\infty$, the divergence is canceled. Indeed, we have

$$\lim_{z'\to-\infty}([-]_{z'}^{x'}-e^{-2V(-\infty)}[+]_{z'}^{x'})=2e^{-V(-\infty)}\int_{-\infty}^{x'}\sinh[V(-\infty)-V(y')]dy', \tag{7.1a}$$

$$\lim_{z\to+\infty}([-]_x^z-e^{-2V(\infty)}[+]_x^z)=2e^{-V(\infty)}\int_x^{\infty}\sinh[V(\infty)-V(y)]dy, \tag{7.1b}$$

and these integrals are finite as long as the potential $V(x)$ converges to $V(\infty)$ and $V(-\infty)$ sufficiently fast at $x\to\pm\infty$. This cancellation of divergence takes place at every order of expansion on account of the properties (6.5a) and (6.5b). We define

$$\begin{aligned}\langle -,a_2,a_3,\ldots,a_{n_1}]_{-\infty}^{x'}&\equiv[-,a_2,a_3,\ldots,a_{n_1}]_{-\infty}^{x'}-e^{-2V(-\infty)}[+,a_2,a_3\ldots,a_{n_1}]_{-\infty}^{x'}\\&=2e^{-V(-\infty)}\int\cdots\int_{-\infty\leq y_1\leq\cdots\leq y_{n_1}\leq x'}dy_1\cdots dy_{n_1}\\&\quad\times\sinh[V(-\infty)-V(y_1)]\exp\left(\sum_{i=2}^{n_1}a_iV(y_i)\right),\end{aligned} \tag{7.2}$$

and

$$\begin{aligned}[c_1,c_2,\ldots,c_{n_3-1},-\rangle_x^{\infty}&\equiv[c_1,c_2,\ldots,c_{n_3-1},-]_x^{\infty}-e^{-2V(\infty)}[c_1,c_2,\ldots,c_{n_3-1},+]_x^{\infty}\\&=2e^{-V(\infty)}\int\cdots\int_{x\leq y_1\leq\cdots\leq y_{n_3}\leq\infty}dy_1\cdots dy_{n_3}\\&\quad\times\sinh[V(\infty)-V(y_{n_3})]\exp\left(\sum_{i=1}^{n_3-1}c_iV(y_i)\right).\end{aligned} \tag{7.3}$$

[For $n_1=1$ and $n_3=1$, we let $\langle-]_{-\infty}^{x'}$ and $[-\rangle_x^{\infty}$ stand for the integrals (7.1a) and (7.1b), respectively.] We obtain the expansion coefficient p_n as

$$p_n(x,x')=\lim_{\substack{z\to+\infty\\z'\to-\infty}}\overline{p_n}(x,x';z,z';W=V(\infty),W'=V(-\infty)), \tag{7.4}$$

and p_n can be expressed in terms of finite integrals of the form (7.2) and (7.3). We can also write (7.4) as

$$p_n(x,x')=\lim_{\substack{z\to+\infty\\z'\to-\infty}}\overline{p_n}(x,x';z,z';W=V(z),W'=V(z')), \tag{7.5}$$

but we cannot let $z\to\infty$ and $z'\to-\infty$ before $W\to V(\infty)$ and $W'\to V(-\infty)$.

In the same way as (6.7), let us use the notation

$$\begin{aligned}&\langle -,a_2,\ldots,a_{n_1};b_1,\ldots,b_{n_2};c_1,\ldots,c_{n_3-1},-\rangle\\&\quad\equiv\langle -,a_2,\ldots,a_{n_1}]_{-\infty}^{x'}[b_1,\ldots,b_{n_2}]_{x'}^{x}[c_1,\ldots,c_{n_3-1},-\rangle_x^{\infty}.\end{aligned} \tag{7.6}$$

We also use this notation for $n_1=0$, where $\langle;\ldots;\ldots\rangle\equiv[\ldots]_{x'}^{x}[\ldots\rangle_x^{\infty}$, and similarly for $n_3=0$. For example, $\langle--;;\rangle=\langle--]_{-\infty}^{x'}$ and $\langle;+;-\rangle=[+]_{x'}^{x}[-\rangle_x^{\infty}$. With this notation, we can obtain p_0, p_1, $p_2,\ldots,$ from $\overline{p_0}$, $\overline{p_1}$, $\overline{p_2},\ldots,$ shown in (6.9) by replacing W and W' by $V(\infty)$ and $V(-\infty)$, removing the terms with $a_1=+1$ or $c_{n_3}=+1$, and replacing [...;...;...] by ⟨...;...;...⟩. Namely, we have

$$p_0=\frac{2}{e^{-V}+e^{-V'}},\tag{7.7a}$$

$$p_1=\frac{p_0^2}{2}\{\langle-;;\rangle+\langle;-;\rangle+\langle;;-\rangle+e^{-V-V'}\langle;+;\rangle\},\tag{7.7b}$$

$$\begin{aligned}p_2=\frac{p_0^3}{2}\Bigg(&\langle--;;\rangle+\langle-;-;\rangle+\langle-;;-\rangle+\langle;--;\rangle+\langle;-;-\rangle+\langle;;--\rangle\\&-\frac{e^{-V}-e^{-V'}}{2}e^{-V}\{\langle-;+;\rangle+\langle;-+;\rangle\}+\frac{e^{-V}-e^{-V'}}{2}e^{-V'}\{\langle;+;-\rangle+\langle;+-;\rangle\}\\&+e^{-V-V'}\{\langle-+;;\rangle+\langle;;+-\rangle\}+e^{-2V-2V'}\langle;++;\rangle\Bigg),\end{aligned}\tag{7.7c}$$

where $V\equiv V(\infty)$ and $V'\equiv V(-\infty)$.

Until now we have been assuming that both $V(+\infty)$ and $V(-\infty)$ are finite. However, this formalism is also applicable to some other cases. Let us first consider the case where $V(+\infty)$ is finite and $V(-\infty)=+\infty$. In this case we can also use (7.4) to obtain the coefficients p_i. Namely, we let $W'\to+\infty$ in $\overline{p_i}$ before taking the limit $z'\to-\infty$. Using the notation $[A;B;C\rangle\equiv[A]_{-\infty}^{x'}[B]_{x'}^{x}[C\rangle_x^{\infty}$, we can write

$$p_0=2e^{V(\infty)},\quad p_1=2e^{2V(\infty)}([-;;\rangle+[;-;\rangle+[;;-\rangle),\tag{7.8a}$$

$$\begin{aligned}p_2=&4e^{3V(\infty)}([--;;\rangle+[-;-;\rangle+[-;;-\rangle+[;--;\rangle+[;-;-\rangle+[;;--\rangle)\\&-2e^{V(\infty)}([-;+;\rangle+[;-+;\rangle).\end{aligned}\tag{7.8b}$$

The integrals in these expressions are finite if $V(x)$ diverges at $x\to-\infty$ faster than $\log|x|$. Similarly, if $V(+\infty)$ is finite and $V(-\infty)=-\infty$, we may first let $W'\to-\infty$ and then $z'\to-\infty$. We have

$$p_0=0,\quad p_1=-2[+;;\rangle,\tag{7.9a}$$

$$p_2=-4e^{-V(\infty)}[++;;\rangle-2e^{-V(\infty)}[+;+;\rangle.\tag{7.9b}$$

Expressions for the cases $V(+\infty)=\pm\infty$ with $V(-\infty)$ finite can be obtained in the same way.

The expressions become still simpler if $V(+\infty)=+\infty$ and $V(-\infty)=-\infty$. We can obtain p_i by taking the limit $W\to+\infty$, $W'\to-\infty$ in $\overline{p_i}$ and then letting $z\to\infty$, $z'\to-\infty$. It is easy to see that $p_n=0$ for n even. We have

$$p_1=-2[+;;],\tag{7.10a}$$

$$p_3=2[+;+;-]+2[+;+-;]+4[++;;-]+4[++;-;]+4[++-;;],\tag{7.10b}$$

and so on. [Here we are using the notation (6.7) with $z'=-\infty$ and $z=\infty$.] The case $V(+\infty)=-\infty$, $V(-\infty)=+\infty$ can be treated similarly.

TABLE I. Classification of the behavior of $V(x)$ at infinity.

	$V(+\infty)=$finite	$V(+\infty)=+\infty$	$V(+\infty)=-\infty$
$V(-\infty)=$finite	i	ii	iii
$V(-\infty)=+\infty$	ii′	iv	v
$V(-\infty)=-\infty$	iii′	v′	vi

The cases $V(+\infty)=+\infty$ with $V(-\infty)=+\infty$ and $V(+\infty)=-\infty$ with $V(-\infty)=-\infty$ need special care. The expansion coefficients p_i for these cases cannot be obtained simply by taking the limit of $\overline{p_i}$. In particular, for the case $V(+\infty)=+\infty$ with $V(-\infty)=+\infty$, the power series expansion of the Green's function begins with a term of order $1/\kappa^2$, not $1/\kappa$ as in (1.7). Extending the formalism of the present paper to such cases is an interesting problem, and it will be discussed elsewhere.

VIII. CONCLUSION

In this paper we derived an expression for the power series expansion of the Green's function. Our main result is (5.24) with (5.25). This is only a basic expression, and we may further study the structure of the expansion coefficients starting from (5.25). By doing so we can derive more practically useful expressions, and investigate various properties of the Green's function.

APPENDIX: LONG TIME BEHAVIOR OF THE GREEN'S FUNCTION

The small-ω behavior (and the long-time behavior) of the Green's function depends much on the behavior of the potential $V(x)$ at $x\to+\infty$ and $x\to-\infty$. We consider the following three cases for the behavior at $x\to+\infty$:

$$V(+\infty)=\text{finite},\quad V(+\infty)=+\infty,\quad V(+\infty)=-\infty,$$

where $V(+\infty)$ stands for $\lim_{x\to\infty}V(x)$. We assume that the potential either converges to a finite limit sufficiently fast or diverges sufficiently fast. (If the potential is a function that converges or diverges slowly, the Green's function may show a singular behavior.[9] However, here we do not discuss such cases.) Similarly, we consider the three cases for $x\to-\infty$:

$$V(-\infty)=\text{finite},\quad V(-\infty)=+\infty,\quad V(-\infty)=-\infty.$$

In all, there are nine cases, as shown in Table I. The Green's function behaves at small ω as[10]

$$G(x,x';\omega)=\frac{C_{-1}}{\kappa}+C_0+C_1\kappa+C_2\kappa^2+\cdots,\quad \text{cases (i), (ii), and (ii}'\text{)},\tag{A1}$$

$$=C_0+C_1\kappa+C_2\kappa^2+C_3\kappa^3+\cdots,\quad \text{cases (iii) and (iii}'\text{)},\tag{A2}$$

$$=\frac{C_{-2}}{\kappa^2}+C_0+C_2\kappa^2+C_4\kappa^4+\cdots,\quad \text{case (iv)},\tag{A3}$$

$$=C_0+C_2\kappa^2+C_4\kappa^4+C_6\kappa^6+\cdots,\quad \text{cases (v), (v}'\text{), and (vi)},\tag{A4}$$

where κ is defined by (1.6).

The long-time behavior is also classified according to (A1)–(A4) above. The expressions (A1) and (A2) correspond to the cases where the eigenvalue spectrum of the Fokker–Planck operator (or the ''energy'' spectrum) is continuous. By Fourier transforming (A1) or (A2), we obtain the

−rge-t expression of $G(x,x';t)$ as an expansion in powers of $1/t$. (The contribution comes only from odd powers of κ.) For (A1) the power series begins with the $t^{-1/2}$ term, and for (A2) it begins with $t^{-3/2}$. Thus, $G(x,x';t)$ shows a power-law decay for large t.

On the other hand, (A3) and (A4) correspond to the cases of a discrete spectrum (at least in the low-energy region). In these cases $G(x,x';\omega)$ does not include odd powers of κ, and hence $G(x,x';t)$ does not contain a part that falls off with a power law. The long-time behavior of the Green's function is determined by low-lying eigenvalues, and $G(x,x';t)$ decays exponentially for large t. The eigenvalues correspond to the poles of $G(x,x';\omega)$. (See Refs. 6 and 11 for a method for the calculation of the eigenvalues.)

[1] H. Risken, *The Fokker–Planck Equation* (Springer-Verlag, Berlin, 1984).
[2] T. Miyazawa, J. Math. Phys. **34**, 1587 (1993).
[3] T. Miyazawa, J. Math. Phys. **36**, 5643 (1995).
[4] R. Slansky, in *Les Houches, Session XLIV, 1985*, edited by P. Ramond and R. Stora (North-Holland, Amsterdam, 1987), p. 533; P. G. O. Freund, *Introduction to Supersymmetry* (Cambridge University Press, Cambridge, 1986).
[5] E. Witten, Nucl. Phys. B **185**, 513 (1981); M. Bernstein and L. S. Brown, Phys. Rev. Lett. **52**, 1933 (1984); H. R. Jauslin, J. Phys. A **21**, 2337 (1988).
[6] T. Miyazawa, J. Math. Phys. **39**, 2035 (1998).
[7] See Ref. 6. The quantity $\tau(z,z')=(\psi_0^{(1)},U^{(1)}(z,z')\psi_0^{(1)})^{(1)}$ is defined there as the transmission coefficient for the interval (z',z).
[8] This relation can be derived by comparing Eqs. (11.30c) and (11.30d) of Ref. 6. (The symbol $\hat{\mathscr{J}}$ in this reference corresponds to J in the present paper.)
[9] T. Miyazawa, J. Math. Phys. **33**, 2428 (1992).
[10] T. Miyazawa, Phys. Rev. A **39**, 1447 (1989).
[11] T. Miyazawa, J. Phys. A **25**, 2359 (1992).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

The Camassa–Holm equation as a geodesic flow on the diffeomorphism group

Shinar Kouranbaeva[a)]
Department of Mathematics, University of California, Santa Cruz, California 95064

(Received 23 July 1998; accepted for publication 2 November 1998)

Misiolek [J. Geom. Phys. **24**, 203–208 (1998)] has shown that the Camassa–Holm equation is a geodesic flow on the Bott–Virasoro group. In this paper it is shown that the Camassa–Holm equation for the case $\kappa=0$ is the geodesic spray of the weak Riemannian metric on the diffeomorphism group of the line or the circle obtained by right translating the H^1 inner product over the entire group. This paper uses the right-trivialization technique to rigorously verify that the Euler–Poincaré theory for Lie groups can be applied to diffeomorphism groups. The observation made in this paper has led to physically meaningful generalizations of the CH-equation to higher dimensional manifolds. © *1999 American Institute of Physics.* [S0022-2488(99)02102-7]

I. INTRODUCTION

Camassa and Holm[1,2] derived a new completely integrable dispersive shallow water equation that is bi-Hamiltonian and thus possesses an infinite number of conservation laws in involution. The Camassa–Holm (CH) equation is obtained by using an asymptotic expansion directly in the Hamiltonian for Euler's equations in the shallow water regime. Camassa and Holm[1] also have the formal Lie–Poisson derivation of the equation. Below, another remarkable property of the CH-equation is shown. Namely, the CH-equation can be realized as a geodesic equation on a Riemannian manifold on which the methods of infinite-dimensional geometry can be applied. The geodesic nature of the CH-equation enables one to transfer the problem from the equation to a problem of finding geodesics on the diffeomorphism group. This idea was first rigorously carried out by Ebin and Marsden[3] for the Euler equations. They have shown that the Euler equations are geodesic equations for the right-invariant L^2 metric on the group of volume-preserving diffeomorphisms.

Section II illustrates the main result by formally applying the Euler–Poincaré theory for Lie groups to a continuum mechanical system. Section III verifies the legitimacy of the application. In addition, Sec. III contains independent results on the Riemannian geometry of a C^1-manifold which is also a topological group with C^1 right translation. Using the right-trivialization technique, a global Christoffel map is introduced, and formulas are derived for the spray and the Levi–Civita connection similar to the finite-dimensional case. The method is inspired by the theory of affine connections on parallelizable manifolds developed by Marsden, Ratiu and Raugel.[4] At the end of Sec. III, a version of the Euler–Poincaré theorem for a diffeomorphism group is verified. Section IV utilizes the results of Sec. III to demonstrate that the CH-equation is a geodesic flow of the right-invariant metric on the diffeomorphism group of **R** or of the circle. Section V addresses uniqueness and existence issues for solutions of the CH-equation. Observations made in this paper have n-dimensional generalizations to the volume-preserving diffeomorphism group of a Riemannian manifold which lead to a new class of models for mean hydrodynamic motion. See Ref. 5 for application of this to numerous fluids models, such as those in geophysics, and see Ref. 6 for the development of the geometry and curvature of volume-preserving diffeomorphism groups with right-invariant H^1 metric. For Riemannian manifolds with

[a)]Electronic mail: shinar@cats.ucsc.edu

boundary, new subgroups of the diffeomorphism group have been established which give rise to remarkable theorems on the limit of zero viscosity. See Ref. 7 for a detailed account.

II. FORMAL DERIVATION

In this section we illustrate the main result of the paper by *formally* applying the pure Euler–Poincaré theorem on the right-invariant Lagrangians on Lie groups (see Ref. 8, and references therein) to the case of the diffeomorphism group of a certain Sobolev class H^s, $s>\frac{3}{2}$. The diffeomorphism group is *not* a Lie group (left translation and inversion are not smooth, only continuous, whereas right translation is smooth), and the pure Euler–Poincaré theorem strictly does not apply. However, we will demonstrate in the following sections that the formal derivation given in this section can be rigorously justified using standard trivialization techniques.

Let M be the flat circle S^1 or the real line $\mathbf{R}$. $\mathrm{Diff}^s(M)\equiv\mathcal{D}^s$ denotes the diffeomorphism group of M of some given Sobolev class. The case $M=S^1$ corresponds to periodic boundary conditions. For the case $M=\mathbf{R}$, the chosen Sobolev space automatically imposes appropriate decay hypotheses at infinity. Under these boundary conditions, $\mathrm{Diff}^s(M)$ is a smooth infinite-dimensional manifold and a topological group relative to the induced manifold topology. $\boldsymbol{\mathcal{X}}(M)$ denotes the vector fields on M of the same differentiability class. Formally, this is the *right* Lie algebra of $\mathrm{Diff}^s(M)$, e.g., the standard left Lie algebra bracket is *minus* the usual Lie bracket for vector fields. For $u,v\in\boldsymbol{\mathcal{X}}(M)$ the adjoint action of the Lie algebra on itself is given by

$$\mathrm{ad}_u v=[u,v].$$

Consider the H^1 inner product on $\boldsymbol{\mathcal{X}}(M)$ and define a weak Riemannian metric on the whole group $\mathcal{D}^s$ by right-translation of the given inner product on the Lie algebra. The corresponding quadratic form defines a right-invariant Lagrangian on $\mathrm{Diff}^s(M)$ whose restriction to the Lie algebra $\boldsymbol{\mathcal{X}}(M)$ is equal to the square of the H^1 norm:

$$l(u)=\frac{1}{2}\int_M (u^2+u_x^2)\,dx. \tag{1}$$

Next, one defines ad_u^* the adjoint of ad_u with respect to the H^1 inner product, that is, for $u,v,w \in\boldsymbol{\mathcal{X}}(M)$

$$\langle \mathrm{ad}_u^*\, w,v\rangle_{H^1}=\langle w,[u,v]\rangle_{H^1}.$$

Also, for a function $l:\boldsymbol{\mathcal{X}}(M)\to\mathbf{R}$, define the functional derivative $\delta l/\delta u\in\boldsymbol{\mathcal{X}}(M)$ with respect to the given metric by

$$\left\langle \frac{\delta l}{\delta u},v\right\rangle_{H^1}=\delta l(u)\cdot v \quad \text{for } v\in\boldsymbol{\mathcal{X}}(M).$$

Assuming the existence of ad_u^* for each $u\in\boldsymbol{\mathcal{X}}(M)$, we can formally write the Euler–Poincaré equations

$$\frac{d}{dt}\frac{\delta l}{\delta u}=-\mathrm{ad}_u^*\frac{\delta l}{\delta u}.$$

After computing ad_u^* and $\delta l/\delta u$ (for the computations see Sec. IV) the Euler–Poincaré equations yield the Camassa–Holm equation

$$u_t-u_{xxt}=-3uu_x+2u_xu_{xx}+uu_{xxx}. \tag{2}$$

The above observation motivates one to develop a theory to perform the procedure.

III. RIEMANNIAN GEOMETRY OF PARTIAL LIE GROUPS

A. Review of definitions

In what follows, the general structure needed is that of a C^1-manifold G which is also a topological group in the induced manifold topology, and it is assumed that only the right translation is C^1. We call such a group G a partial Lie group. Below we review the notations used in the paper. The proofs for the stated formulas can be found in the standard texts such as Spivak.[9] Let G be a manifold equipped with a metric $\langle\cdot,\cdot\rangle$. Let $\pi_G:TG\to G$ and $\pi_{TG}:TTG\to TG$ be the tangent bundle projections and denote by $V=\ker T\pi_G$ the vertical subbundle of TTG. The *connector* $K:TTG\to TG$ is given by

$$K(TY\cdot X)=\nabla_X Y,$$

for $X,Y\in\mathcal{X}(G)$ the Lie algebra of vector fields and ∇ the Levi-Civita connection coming from a metric.

A vector $U\in TTG$ is called *horizontal* if $U\in\ker K$; $H=\ker K$ is a subbundle of TTG called the *horizontal subbundle* of the connection and we have the decomposition $TTG=H\oplus V$ over TG with the projection π_{TG}. Then the *horizontal lift* of $w\in T_gG$ to $T_v(TG)$, $v\in T_gG$, is defined as

$$\mathrm{hor}_v\, w=(T_v\pi_G|H_v)^{-1}(w).$$

The horizontal lift operator $\mathrm{hor}_v:T_gG\to H_v$ is an isomorphism for all $v\in TG$ and locally,

$$\mathrm{hor}_v\, w=b^i\frac{\partial}{\partial x^i}-\Gamma^i_{jk}b^ja^k\frac{\partial}{\partial v^i},$$

where $v=a^i\partial/\partial x^i$ and $w=b^i\partial/\partial x^i$.

The *spray* $S:TG\to TTG$ is by definition the Lagrangian vector field of the energy function $E(v)=L(v)=\frac{1}{2}\langle v,v\rangle$, i.e.,

$$\mathbf{i}_S\omega_L=\mathbf{d}E,$$

where ω_L is the symplectic form on TG, and $\mathbf{i}_S$ denotes the interior product (for more details refer to *Foundations of Mechanics*[10]). Locally,

$$S(v)=a^i\frac{\partial}{\partial x^i}-\Gamma^i_{jk}a^ja^k\frac{\partial}{\partial v^i},$$

where $v=a^i\partial/\partial x^i$, and notice that by definition, we have the useful identity

$$S(v)=\mathrm{hor}_v\, v. \tag{3}$$

Let $g(t)$ be a smooth curve in G and let $\dot g(t)$ be its tangent vector field. If Y is another vector field, define the *covariant derivative* of Y along $g(t)$ by

$$\frac{DY}{dt}=\nabla_{\dot g(t)}Y.$$

If the covariant derivative of Y is zero, Y is said to be parallel along $g(t)$. It follows from the definition of a connector that $TY(\dot g(t))\in H$ if and only if $DY/dt=0$. Locally for a given curve $g(t)$ this equation becomes a linear system of ordinary differential equations

$$\frac{dY^i(t)}{dt}+\Gamma^i_{jk}\dot g^j(t)Y^k(t)=0.$$

A curve $g(t)$ is called the *geodesic* of a connection ∇, if $\dot{g}(t)$ is parallel along $g(t)$, i.e., if

$$\nabla_{\dot{g}(t)}\dot{g}(t)=0.$$

Locally, this is a second-order differential equation

$$\ddot{g}^i(t)+\Gamma^i_{jk}\dot{g}^j(t)\dot{g}^k(t)=0.$$

B. The Levi–Civita connection and the spray for a right-invariant metric on *G*

Let G be a C^1-manifold which is a topological group with C^1 right translation. Assume that G admits a right-invariant metric. There is a vector bundle isomorphism called the right trivialization map

$$\rho: TG \to G\times\mathcal{G},$$

$$v \mapsto (g, T_g R_{g^{-1}}\cdot v).$$

Then $T\rho: TTG\to T(G\times\mathcal{G})$ maps TTG isomorphically onto $TG\times\mathcal{G}\times\mathcal{G}$. We can further trivialize via $\rho\times id$

$$TTG \overset{(\rho\times id)\circ T\rho}{\longrightarrow} G\times\mathcal{G}\times\mathcal{G}\times\mathcal{G}.$$

Note the isomorphic image of the vertical subbundle of TTG is equal to $G\times O\times\mathcal{G}\times\mathcal{G}$, the projection being onto the first and third factors. To keep the base points in the first two factors, we apply the involution map

$$\sigma: G\times\mathcal{G}\times\mathcal{G}\times\mathcal{G}\to G\times\mathcal{G}\times\mathcal{G}\times\mathcal{G},$$

$$(g,X,Y,Z)\mapsto(g,Y,X,Z).$$

Then the image of the vertical bundle V of TTG equals $G\times\mathcal{G}\times O\times\mathcal{G}$ with the projection being on the first two factors. Therefore, the isomorphism we are working with is

$$TTG \overset{\sigma\circ(\rho\times id)\circ T\rho}{\longrightarrow} G\times\mathcal{G}\times\mathcal{G}\times\mathcal{G}.$$

A given metric gives rise to the Levi–Civita connection which determines the horizontal bundle. We wish to express it in the trivialization ρ, which in turn helps us to find the spray. Define the continuous $\mathbf{R}$-bilinear map $\gamma_g:\mathcal{G}\times\mathcal{G}\to\mathcal{G}$ depending smoothly on g by

$$\rho((\nabla_X Y)(g))=(g, d\overline{Y}(g)\cdot X(g)+\gamma_g(\overline{X}(g),\overline{Y}(g))), \tag{4}$$

where

$$\rho(X(g))=(g,\overline{X}(g)),\quad X,Y\in\boldsymbol{\mathcal{X}}(G).$$

Recall that a connection in a finite-dimensional case in coordinates is given by

$$S(X)=X^i\frac{\partial}{\partial g^i}-\Gamma^i_{jk}X^jX^k\frac{\partial}{\partial\dot{g}^i},$$

and

$$(\nabla_X Y)^i=X^jY^i_{,j}+\Gamma^i_{jk}X^jX^k.$$

Define $\gamma = \Gamma^i_{jk} X^j X^k \partial/\partial \dot{g}^i$, then

$$(\nabla_X Y)^i = X^j Y^i_{,j} + \gamma^i. \tag{5}$$

Observe that formulas (4) and (5) are analogous.

Let $g(t)$ be a curve in G with $g(0)=g$, $\dot{g}(0)=w$, and $v \in T_g G$. There exists a curve $v(t) \in TG$ such that

$$v(0)=v, \quad \pi_G(v(t))=g(t), \quad \frac{Dv(t)}{dt}=0. \tag{6}$$

Therefore, $Tv(t)\cdot \dot{g}(t)$ is horizontal; i.e. the tangent vector field dv/dt of $v(t)$ is always horizontal. Therefore $\dot{v}(0)=\mathrm{hor}_v\, w$, the horizontal lift of $w=\dot{g}(0)\in T_g G$ to $T_v(TG)$. If

$$\rho(v(t)) = (g(t),\xi(t)), \quad \rho(v)=(g,\xi),$$

$$\rho(\dot{g}(t)) = (g(t),\zeta(t)), \quad \rho(w)=(g,\zeta),$$

then $v(0)=v$ in the trivialization reads $\xi(0)=\xi$, and the base projection condition is automatically satisfied. By the definition of a covariant derivative and the chain rule, we find that

$$\rho\left(\frac{Dv}{dt}\right) = (g(t), d\xi(t)\cdot \dot{g}(t) + \gamma_{g(t)}(\zeta(t),\xi(t))) = \left(g(t), \frac{d\xi}{dt} + \gamma_{g(t)}(\zeta(t),\xi(t))\right),$$

so that $(g(t),\xi(t))$ is parallel along $g(t)$ in $G\times\mathcal{G}$ relative to the push-forward connection by ρ *if and only if*

$$\frac{d\xi}{dt} + \gamma_{g(t)}(\zeta(t),\xi(t)) = 0, \quad \xi(0)=\xi. \tag{7}$$

Equation (7) enables us to compute the horizontal lift of (g,ζ) to $T_{(g,\xi)}(G\times\mathcal{G})$ and hence the spray using (3). Let us compute $\mathrm{hor}_v\, w = \dot{v}(0)$ in the trivialization given by ρ. We have that

$$\begin{aligned}(\sigma\circ(\rho\times id)\circ T\rho)(\dot{v}(0)) &= \sigma\circ(\rho\times id)\left(\left.\frac{d}{dt}\right|_{t=0} (\rho\circ v)(t)\right)\\ &= \sigma\circ(\rho\times id)\left(\dot{g}(0), \xi(0), \left.\frac{d\xi(t)}{dt}\right|_{t=0}\right)\\ &= \sigma(g(0),\zeta(0),\xi(0), -\gamma_{g(0)}(\zeta(0),\xi(0))) = (g,\xi,\zeta,-\gamma_g(\zeta,\xi)).\end{aligned}$$

Therefore,

$$\mathrm{hor}_{(g,\xi)}(g,\zeta) = (g,\xi,\zeta,-\gamma_g(\zeta,\xi))$$

and the spray of the Levi–Civita connection in its right trivialization is given by

$$\overline{S}(g,\xi) = \mathrm{hor}_{(g,\xi)}(g,\xi) = (g,\xi,\xi,-\gamma_g(\xi,\xi)). \tag{8}$$

Applying $(\rho^{-1}\times id)\circ\sigma^{-1}$ we can express the right trivialization of the spray as

$$\overline{S}(g,\xi) = (T_e R_g \cdot \xi, \xi, -\gamma_g(\xi,\xi)), \tag{9}$$

where $X(g) = T_e R_g\cdot\xi$ is the right-invariant vector field on G associated with a Lie algebra element ξ. It follows that

$$S(v)=T\rho^{-1}\circ\overline{S}\circ\rho(v). \tag{10}$$

Given a vector bundle E over G, we shall denote by $\mathcal{E}$ the collection of all smooth sections $\sigma:G\rightarrow E$ such that $\pi\circ\sigma=id_G$. Let $E=G\times\mathcal{G}$, then the condition $\pi\circ\sigma=id_G$ implies that

$$\mathcal{E}=\{\sigma:G\rightarrow\mathcal{G}\mid\sigma\ \text{is smooth}\}.$$

If $\rho(X(g))=(g,\overline{X}(g))$, define the map $\overline{\nabla}:\boldsymbol{\mathcal{X}}(\mathrm{G})\times\mathcal{E}\rightarrow\mathcal{E}$ via

$$(\overline{\nabla}_X\sigma)(g)=T_g\sigma\cdot X(g)+\gamma_g(\overline{X}(g),\sigma(g)).$$

It is straightforward to check that $\overline{\nabla}$ is a vector bundle connection. Therefore, a bilinear map γ_g defines the vector bundle connection $\overline{\nabla}:\boldsymbol{\mathcal{X}}(G)\times\mathcal{E}\rightarrow\mathcal{E}$. Moreover when γ_g is defined as in (4), the push-forward by ρ of the Levi–Civita connection is equal to

$$\rho(\nabla_X Y(g))=(g,\overline{\nabla}_X Y(g)),$$

and therefore,

$$\nabla_X Y(g)=T_e R_g(\overline{\nabla}_X\overline{Y}(g)).$$

Conclusion 1: If for a given connection ∇ *we think of the map* $\gamma_g:T_eG\times T_eG\rightarrow T_eG$ *as a generalized Christoffel map of the push-forward connection* $\overline{\nabla}$ *under the right trivialization map* ρ, *then we have the formula*

$$\nabla_X Y(g)=\rho^{-1}(d\overline{Y}(g)\cdot X(g)+\gamma_g(\overline{X}(g),\overline{Y}(g))).$$

If in addition we restrict ourselves to the Levi–Civita connection coming from a given metric on G, *we have the formula for the spray using the Christoffel map*

$$S(X(G))=T\rho^{-1}(X(g),\overline{X}(g),-\gamma_g(\overline{X}(g),\overline{Y}(g))).$$

The above two formulas are analogous to the finite-dimensional formulas and they are globally defined on G.

Notice we have not used the right-invariance condition, these conclusions are true for any metric on G. However, our results below will depend heavily on the right-invariance of the metric.

Proposition III.1: Let G be a C^1-manifold which is a topological group with C^1 right translation. Suppose that G admits a right-invariant metric. Then the spray of the corresponding Lagrangian $L(v)=\frac{1}{2}\langle v,v\rangle_g$ is given by

$$S(v)=T\rho^{-1}\circ\overline{S}\circ\rho(v),$$

where

$$\overline{S}(g,\xi)=(T_eR_g\cdot\xi,\xi,-B(\xi,\xi)) \tag{11}$$

and

$$B:T_eG\times T_eG\rightarrow T_eG$$

is defined implicitly by

$$\langle B(\zeta,\xi),\eta\rangle=\langle\zeta,[\xi,\eta]\rangle,\quad \xi,\eta,\zeta\in T_eG. \tag{12}$$

Remark: There is an assumption in the proposition that the operator B exists.

Remark: For the case of Lie groups, the proof of this result can be found in Ref. 10. The operator B was introduced by Arnold.[11] The above proposition is more general as it covers diffeomorphism groups which are of a great interest in hydrodynamics.

Proof: To verify (11) we need to calculate the Christoffel map $\gamma_g(\xi,\xi)$ in (9).

Since ρ is a diffeomorphism, we can push-forward the symplectic form ω_L on TG to define the symplectic form $\omega^s=\rho_*\omega_L$ (the superscript s stands for ''spatial'' because the right trivialization gives rise to spatial coordinates in applications). It can be checked that the push-forward of the spray S on TG is the Lagrangian vector field expressed in space coordinates and $\rho_*S=\overline{S}$; consequently,

$$\mathbf{i}_{\overline{S}}\omega^s=\mathbf{d}(\rho_*E). \tag{13}$$

To calculate the left-hand side recall the following formula (see Ref. 10):

$$\omega^s(g,\xi)((v,\zeta),(w,\eta))=-\langle\zeta,T_gR_{g^{-1}}(w)\rangle_e+\langle\eta,T_gR_{g^{-1}}(v)\rangle_e-\langle\xi,[T_gR_{g^{-1}}(v),T_gR_{g^{-1}}(w)]\rangle_e.$$

By this formula we have that

$$\omega^s(g,\xi)((T_eR_g\cdot\xi,-\gamma_g(\xi,\xi)),(w,\eta))=-\langle-\gamma_g(\xi,\xi),T_gR_{g^{-1}}(w)\rangle_e+\langle\eta,\xi\rangle_e \\ -\langle\xi,[\xi,T_gR_{g^{-1}}(w)]\rangle_e. \tag{14}$$

Since the metric is right-invariant, it follows that

$$E\circ\rho^{-1}(g,\xi)=\tfrac{1}{2}\langle T_eR_g\cdot\xi,T_eR_g\cdot\xi\rangle_g=\tfrac{1}{2}\langle\xi,\xi\rangle_e.$$

Therefore the right-hand side of (13) is equal to

$$\mathbf{d}(E\circ\rho^{-1})(g,\xi)\cdot(w,\eta)=\langle\xi,\eta\rangle_e. \tag{15}$$

From (13), (14), and (15) we may conclude that the value of $\gamma_g(\xi,\xi)$ does not depend on the base point g. Moreover, its value is defined by the following relationship:

$$\langle\gamma(\xi,\xi),\zeta\rangle_e=\langle\xi,[\xi,\zeta]\rangle_e\quad\text{for }\ \xi,\zeta\in T_eG.$$

From the definition (22) of the operator B it follows that $\gamma(\xi,\xi)=B(\xi,\xi)$, and hence (11) is true.

It is known that $-B(\xi,\xi)=(\nabla_{X_\xi}X_\xi)(e)$, where $X_\xi(g)=T_eR_g\cdot\xi$ (see Arnold,[11] Bao and Ratiu[12]). Thus, for right-invariant vector fields, we also have that

$$\overline{S}(g,\xi)=(X_\xi(g),\xi,(\nabla_{X_\xi}X_\xi)(e)).$$

Conclusion 2: Given a right-invariant metric on G, we can find its geodesic equations by finding the spray. The above formulas show that the spray is completely defined by either the operator B or the value of the Levi–Civita connection at the identity.

See the remark in Sec. V.

C. The Euler–Poincaré equations

The Euler–Poincaré equations are the fundamental result about geodesic flow on an arbitrary Lie group. See, for example, Theorem 13.8.3 in Marsden and Ratiu[8] or Appendix 2 in Arnold.[13] Herein, this result is proven for diffeomorphism groups, the configuration space for ideal fluid dynamics. The idea of studying geodesics on diffeomorphism groups in order to do hydrodynamics is due to Arnold.[11]

In order to establish our notation, let us recall some results from Refs. 3 and 14. For $s > n/2$ and M a compact manifold without boundary, we may define the Sobolev H^s maps from M into M. Let $\mathcal{D}^s(M) = \{\eta \in H^s(M,M) | \eta \text{ is bijective and } \eta^{-1} \in H^s(M,M)\}$. If $s > n/2+1$, then $\mathcal{D}^s(M)$ is open in $H^s(M,M)$ and hence is a manifold. Note that $\mathcal{D}^s$ is not a Lie group, but rather a topological group. However, like a Lie group, $\mathcal{D}^s$ has an exponential map which associates to every tangent vector at the identity a one-parameter subgroup of $\mathcal{D}^s$. Such a tangent vector is an H^s vector field on M and the one-parameter subgroup is its flow. If $\pi: TM \to M$ is the canonical projection, one forms the Hilbert space

$$T_\eta \mathcal{D}^s = \{V: M \to TM | V \text{ is } H^s \quad \text{and} \quad \pi \circ V = \eta\},$$

the tangent space at $\eta \in \mathcal{D}^s$. An element V of the tangent space at $\eta \in \mathcal{D}^s$ is called a vector space over η.

Theorem III.1: *Assume that $\mathcal{D}^s(M)$ is equipped with a metric $\langle \cdot , \cdot \rangle$ that is invariant under right translations. Also assume that there exists the operator $B: \boldsymbol{\mathcal{X}}(M) \times \boldsymbol{\mathcal{X}}(M) \to \boldsymbol{\mathcal{X}}(M)$ defined by the identity*

$$\langle B(w,u),v \rangle = \langle w,[u,v] \rangle \quad \text{for } u,v,w \in \boldsymbol{\mathcal{X}}(M).$$

Then a curve $t \to \eta(t)$ in $\mathcal{D}^s$ is a geodesic of this metric if and only $u(t) = T_{\eta(t)} R_{\eta(t)^{-1}} \dot{\eta}(t) = \dot{\eta}(t) \circ \eta^{-1}(t)$ satisfies

$$\frac{du}{dt} = -B(u,u). \tag{16}$$

Proof: By proposition (III.1) the spray S is given by

$$S(V) = T\rho^{-1}(V, V\circ \eta^{-1}, -B(V \circ \eta^{-1}, V\circ \eta^{-1})) \quad \text{for } V \in T_\eta \mathcal{D}^s. \tag{17}$$

In this case the group $G = \mathcal{D}^s$ and the algebra $\mathcal{G} = \boldsymbol{\mathcal{X}}(M)$. Let us compute $T\rho^{-1}$:

$$\rho^{-1}: G \times \mathcal{G} \to TG,$$

$$(\eta, u) \mapsto u \circ \eta,$$

$$T_{(\eta,u)}\rho^{-1}: T_\eta \mathcal{D}^s \times \mathcal{G} \to T_{u\circ\eta}(TG).$$

Let $(\eta(t), u(t))$ be a curve in $\mathcal{D}^s \times \boldsymbol{\mathcal{X}}(M)$ such that $(\eta(0), u(0)) = (\eta, u)$ and $(\dot{\eta}(0), \dot{u}(0)) = (V, w)$, then

$$T_{(\eta,u)}\rho^{-1} \cdot (V,w) = \left.\frac{d}{dt}\right|_{t=0} \rho^{-1}(\eta(t), u(t)) = \left.\frac{d}{dt}\right|_{t=0} u(t) \circ \eta(t) = w \circ \eta + Tu \circ V.$$

Therefore,

$$S(V) = -B(V\circ \eta^{-1}, V \circ \eta^{-1}) \circ \eta + T(V \circ \eta^{-1}) \circ V,$$

or

$$S(V)\circ\eta^{-1}=-B(V\circ\eta^{-1},V\circ\eta^{-1})+T(V\circ\eta^{-1})\circ(V\circ\eta^{-1}). \tag{18}$$

For an integral curve $V_t\in T_{\eta(t)}\mathcal{D}^s(M)$, its pullback $u_t=V_t\circ\eta_t^{-1}$ is a curve in $T_e\mathcal{D}^s(M)$ that consists of H^s-vector fields on M. Then the spray equation,

$$\frac{dV_t}{dt}=S(V_t), \tag{19}$$

is equivalent to

$$\frac{d}{dt}(u_t\circ\eta_t)=S(V_t),$$

or

$$\frac{du_t}{dt}=S(V_t)\circ\eta_t^{-1}-Tu_t\circ u_t=-B(u_t,u_t)+Tu_t\circ u_t-Tu_t\circ u_t=-B(u_t,u_t). \tag{20}$$

□

IV. CAMASSA–HOLM EQUATION AS A GEODESIC FLOW

Henceforth, the subscript t will denote the partial derivative with respect to t. We shall apply the general results obtained in Sec. III to the CH-equation (2). For periodic boundary conditions, the configuration space is $G=\mathcal{D}^s(S^1)$ with Lie algebra $\mathcal{G}=\boldsymbol{\mathcal{X}}(S^1)$ of the same differentiability class. One may also consider (2) on R with the appropriate decay conditions at infinity guaranteed by the Sobolev class H^s. Formal computations are identical in both cases and, hence, may be treated simultaneously. Consider the H^1 inner product on $\mathcal{G}$ given by

$$\langle u,v\rangle_1=\int (uv+u_xv_x)dx,$$

where the integral may either be taken over S^1 or R. Camassa and Holm[2] have shown that the Lagrangian for the CH-equation (2) is given by the square of the H^1 norm (1).

Given an inner product on a Lie algebra, we can define a metric on all $T\mathcal{D}^s$ by right translation so that for $V,W\in T_\eta\mathcal{D}^s$,

$$\begin{aligned}\langle V,W\rangle_\eta&=\langle T_\eta R_{\eta^{-1}}\cdot V,T_\eta R_{\eta^{-1}}\cdot W\rangle_{id}\\&=\langle V\circ\eta^{-1},W\circ\eta^{-1}\rangle_1\\&=\int [(V\circ\eta^{-1})(W\circ\eta^{-1})+(V\circ\eta^{-1})_x(W\circ\eta^{-1})_x]dx.\end{aligned} \tag{21}$$

This metric is right-invariant by definition and it defines the extended Lagrangian

$$L(V)=\tfrac{1}{2}\langle V,V\rangle_\eta.$$

The main result of this section is

Theorem IV.1: *Let $t\to\eta(t)$ be a curve in the diffeomorphism group $\mathcal{D}^s$ starting at the identity. Then $\eta(t)$ is a geodesic of the metric (21) if and only if the time-dependent vector field $u(t)=\dot\eta(t)\circ\eta^{-1}(t)$ satisfies the CH-equation (2).*

Proof: By the theorem (III.1), the geodesic equations for the metric (21) are equivalent to (16). From the definition of the operator B, we have that

$$\begin{aligned}\langle B(w,u),v\rangle=\langle w,[u,v]\rangle&=\int(-uv_x+u_xv)w+(-uv_x+u_xv)_xw_x\,dx\\&=\int(u_xw+(uw)_x)v-w_{xx}(-uv_x+u_xv)\,dx\\&=\int(2u_xw-2u_xw_{xx}+uw_x-uw_{xxx})\,v\,dx\\&=\int(2u_x(1-\partial_x^2)w+u(1-\partial^2)w_x)v\,dx.\end{aligned}$$

Furthermore,

$$\langle B(w,u),v\rangle=\int(B(w,u)v+B(w,u)_xv_x)\,dx=\int((1-\partial^2)B(w,u))v\,dx,$$

and hence the formula for the operator B is

$$B(w,u)=(1-\partial^2)^{-1}(2u_x(1-\partial^2)w+u(1-\partial^2)w_x). \tag{22}$$

Observe that the map $v\mapsto(1-\partial^2)v$ is a smooth map from H^s to H^{s-2}. The product (not the composition) of H^s functions is H^s again. Also $(1-\partial^2)^{-1}$ is an isomorphism from H^{s-2} to H^s. Therefore (22) explicitly shows that the operator $B:T\mathcal{D}^s\times T\mathcal{D}^s\to T\mathcal{D}^s$ has no derivative loss. Now we obtain the Euler–Poincaré equation:

$$\begin{aligned}\frac{\partial u}{\partial t}=-B(u,u)&=-(1-\partial^2)^{-1}(2u_xu+uu_x-2u_xu_{xx}-uu_{xxx})\\&=-(1-\partial^2)^{-1}(3uu_x-2u_xu_{xx}-uu_{xxx}).\end{aligned}$$

This completes the proof that the geodesic equations for the metric coming from the H^1 inner product on the Lie algebra of vector fields $\mathcal{G}$ are equivalent to the CH-equation (2).

Remark: The Lie-algebra bracket $[u,v]$ on $\mathfrak{X}(M)$ is minus the Jacobi–Lie bracket (for an explanation refer to Marsden and Ratiu,[8] Chap. 9). □

(a) *Alternative derivation.* Below, we begin to compute the geodesic equations for the metric (21) by calculating the spray of the corresponding Lagrangian. Camassa and Holm[2] have shown that the CH-equation can be expressed in the integral form

$$u_t+uu_x=-(1-\partial^2)^{-1}\partial(u^2+\tfrac{1}{2}u_x^2)=-\int e^{-|x-y|}\left(uu_y+\frac{1}{2}u_yu_{yy}\right)dy.$$

Equation (18) together with the fact that $Tu\circ u$ is simply uu_x in one dimension shows that the spray is equal to

$$S(V)=-(1-\partial^2)^{-1}\partial((V\circ\eta^{-1})^2+\tfrac{1}{2}(V\circ\eta^{-1})_x^2)\circ\eta. \tag{23}$$

Letting $u=V\circ\eta^{-1}$ verifies the claim.

Remark. Herein, we have established that the spray (23) does not have derivative loss, and hence is a continuous operator from H^s into H^s. In fact, the spray (23) is actually smooth and this fact follows from arguments in Theorem 3.3 in Shkoller[6] as well as Theorem 4.2 in Ref. 7. See the remark in Sec. V.

V. DISCUSSION

We would like to emphasize that we built a right invariant metric on $\mathcal{D}^s$ by taking the H^1 inner product on the tangent space at the identity and right-translating it over the whole space. This does not coincide with the *usual* H^1 metric on each fiber $T_\eta \mathcal{D}^s$; see the remark after Theorem 4.1 in Ref. 7. To illustrate the difference of two approaches let us compare the geodesics of the L^2 metric with the right-invariant L^2 metric in the one-dimensional case.

A curve $\eta(t) \in \mathcal{D}^s$ is a geodesic of the L^2-metric if and only if the corresponding spatial velocity field

$$u = V \circ \eta^{-1}$$

satisfies Burger's equation:

$$u_t + u u_x = 0. \tag{24}$$

The corresponding Euler–Lagrange equations for the material velocity $V = \dot{\eta}$ are given by

$$V_t = 0.$$

The spray of this metric is equal to zero and hence smooth; however, as the metric is not right invariant, the Euler–Poincaré theorem does not apply.

For the right-invariant L^2 metric the Euler–Poincaré equations are given by

$$u_t + 3 u u_x = 0. \tag{25}$$

The corresponding Euler–Lagrange equations are given by

$$\eta_X \dot{V} + 2 V V_X = 0,$$

where η_X is the Jacobian of η, and X denotes the material coordinate of the fluid particle. The spray in this case is given by

$$S(\eta, V) = -\frac{2}{\eta_X} V V_X. \tag{26}$$

Since there is a loss of derivatives, the spray is not smooth (cf. Remark 3.5 in Shkoller[6]).

As we see from the above calculations, the two equations in the spatial velocities differ only in a scalar coefficient multiplying the derivative term, however, the corresponding sprays are completely different.

Remark: We note that Eq. (23) for the geodesic spray of the right-invariant H^1 metric on either S^1 or R has no derivative loss and hence shows that the CH-equation is an ordinary differential equation on the group $\mathcal{D}^s$. Thus, existence and uniqueness of solutions to (2) may be obtained by standard Picard iteration argument in the event that S is locally Lipschitz.

Lemmas 3.1 and 3.2 of Shkoller[6] show that S is C^1, and hence the result follows. See Refs. 7 and 6 for the well-posedness of the geodesic flow of the diffeomorphism groups on n-dimensional Riemannian manifolds.

It would be interesting to study the Lagrangian stability of the CH-equation, and this requires analysis of the curvature operator. Misiolek[15] has computed the sectional curvature of $\mathcal{D}^s(S^1)$. Shkoller[6] has obtained an explicit form for the H^1 covariant derivative $\tilde{\nabla}^1$ on volume-preserving diffeomorphism groups and has proved that the weak curvature tensor of $\tilde{\nabla}^1$ is a bounded trilinear operator in the H^s topology. We would like to explore these type of estimates on the full diffeomorphism group of the circle, as well as investigate the role of generalized flows in peakon dynamics. Regarding peakon dynamics, there is the paper of Alber *et al.*[16]

ACKNOWLEDGMENTS

I am thankful to Darryl D. Holm and Tudor Ratiu for their kind guidance during the course of this work. Their helpful suggestions, patient discussions, and sincere interest in the subject helped me to build confidence in this work. I also would like to thank Jerrold E. Marsden and Steve Shkoller for their comments on the paper. Their comments helped to make the paper more clear and detailed.

[1] R. Camassa and D. D. Holm, ''A completely integrable dispersive shallow water equation with peaked solutions,'' Phys. Rev. Lett. **71**, 1661–1664 (1993).

[2] R. Camassa, D. D. Holm, and J. M. Hyman, ''A new integrable shallow water equation with peaked solutions,'' Adv. Appl. Mech. **31**, 1–33 (1994).

[3] D. Ebin and J. E. Marsden, ''Groups of diffeomorphisms and the motion of an incompressible fluid,'' Ann. Math. **92**, 102–163 (1970).

[4] J. E. Marsden, T. Ratiu, and G. Raugel, ''Symplectic connections and the linearisation of Hamiltonian systems,'' Proc. R. Soc. Edinburgh, Sect. A: Math. **117**, 329–380 (1991).

[5] D. D. Holm, J. E. Marsden, and T. Ratiu, ''The Euler–Poincaré equations and semidirect products with applications to continuum theories,'' Adv. Math. **137**, 1–81 (1998).

[6] S. Shkoller, ''Geometry and curvature of diffeomorphism groups with H^1 metric and mean hydrodynamics,'' J. Funct. Anal. **160**, 337–365 (1998).

[7] D. D. Holm, S. Kouranbaeva, J. E. Marsden, T. Ratiu, and S. Shkoller, ''A nonlinear analysis of the averaged Euler equations,'' Arnol'dfest Proceedings, Communications of the Field Institute (to appear).

[8] J. E. Marsden and T. Ratiu, *Introduction to Mechanics and Symmetry* (Springer, New York, 1994).

[9] M. Spivak, *Differential Geometry* (Brandeis University, Waltham, MA, 1970), Vols. 1–5.

[10] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed. (Addison–Wesley, Reading, MA, 1978).

[11] V. Arnold, ''Sur la géometrie différentielle des groupes de Lie de dimension infinie et ses applications a l'hydrodynamique des fluides parfaits,'' Ann. Inst. Fourier Grenoble **16**, 319–361 (1966).

[12] D. Bao and T. Ratiu, ''On a Monge–Ampére equation and some geometry of diffeomorphism groups'' (unpublished).

[13] V. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).

[14] J. E. Marsden, D. Ebin, and A. Fischer, ''Diffeomorphism groups, hydrodynamics and relativity,'' in *Proceedings of the 13th Biennial Seminar of the Canadian Mathematical Congress on Differential Topology: Differential Geometry and Applications* (Vanstone, Montreal, 1972), pp. 135–279.

[15] G. Misiolek, ''A shallow water equation as a geodesic flow on the Bott–Virasoro group,'' J. Geom. Phys. **24**, 203–208 (1998).

[16] M. Alber, R. Camassa, D. D. Holm, and J. E. Marsden, ''The geometry of peaked solitons and billiard solutions of a class of integrable pde's,'' Lett. Math. Phys. **32**, 137–151 (1994).

Large smoke rings with concentrated vorticity

Carlo Marchioro[a)]
Dipartimento di Matematica, Università di Roma La Sapienza, Piazzale A. Moro 2, 00185 Roma, Italy

(Received 23 March 1998; accepted for publication 30 September 1998)

In this paper we study an incompressible inviscid fluid when the initial vorticity is sharply concentrated in N disjoint regions. This problem has been well studied when a planar symmetry is present, i.e., the fluid moves in R^2. In this case we know that, when the diameter σ of each region supporting the vorticity is very small, the time evolution of the fluid is quite well described by a dynamical system with finite degrees of freedom called the ''point vortex model.'' In particular the connection between this model and the Euler equation has been proved rigorously as $\sigma \to 0$. In the present paper we discuss the ''stability'' of the point vortex model with respect to a particular small perturbation of the planar symmetry. More precisely we consider a fluid moving in R^3 with a cylindrical symmetry without swirl in which each vortex is no longer a straight tube, but a vorticity ring. We prove that large annuli of radii $r \approx \sigma^{-\beta}$ for any $\beta > 0$ remain ''localized'' and hence we obtain the point vortex model as $\sigma \to 0$.
[S0022-2488(99)02701-2]

I. INTRODUCTION

In the present paper we study the time evolution of an incompressible inviscid fluid when the initial vorticity is concentrated in N small disjoint regions of space. This problem is well studied when a planar symmetry is present, i.e., the fluid moves in $\mathbb{R}^2$. In this case we can approximate the time evolution of the flow by using a model with finite degrees of freedom, called the ''point vortex system,'' which reads

$$\dot{\mathbf{x}}_i = -\frac{1}{2\pi} \sum_{j=1, j\neq i}^{N} a_j \nabla_i^{\perp} \ln|\mathbf{x}_i - \mathbf{x}_j|, \quad \mathbf{x}_i(0) = \mathbf{x}_i, \quad i = 1,...,N, \tag{1.1}$$

where

$$\nabla_i^{\perp} \equiv (\partial_{y_i}, -\partial_{x_i}), \quad \mathbf{x}_i \equiv (x_i, y_i) \in \mathbb{R}^2. \tag{1.2}$$

The N real constants a_i are called ''charges'' or ''intensities'' of the vortices.

This model was introduced in the 19th century[1–4] and widely studied in many papers.[5–12] One of the more interesting applications is related to the so-called ''vortex method,'' a mean field limit introduced by Chorin to investigate fluids with a weak viscosity[13] and generalized to an inviscid flow by many authors. (On the system (1.1), the vortex method and related topics there is a wide literature. We quote here some main papers[1–35] and we suggest for more detailed references to see, for instance, Ref. 33). This limit corresponds to N going to ∞ and each ''charge'' going to 0. Another possible connection exists; N and the ''charges'' remain fixed and the diameter of the region supporting the vorticity goes to zero. This limit has been rigorously studied in[27–35] and in particular the following result called ''localization''[33–35] has been proved. Let the initial vorticity be concentrated in N small disjoint regions $\Lambda_i(0)$ of diameter σ, then, for any fixed time, the time

[a)]Electronic mail: marchioro@axcasp.caspur.it

evolved vorticity remains concentrated in N small disjoint regions $\Lambda_i(t)$ of diameter d such that $d \to 0$ as $\sigma \to 0$. This property provides the main tool for proving the connection between the Euler flow and the point vortex model.

In the present paper we discuss the same problem in the presence of a small perturbation of the planar symmetry. It would be more interesting to deform each vortex tube independently. Asymptotic results have been obtained in this direction (see Ref. 36, and references quoted in), but the task to obtain rigorous results appears too hard. So we consider a common curvature perturbation. More precisely, we consider a fluid moving in $\mathbb{R}^3$ and with a cylindrical symmetry without swirl (see Sec. II). In this case the analog of the straight vortex tubes become rings of vorticity (the so-called ''smoke ring''). We consider N smoke rings of mean radius $r_i \approx r_0$, transversal section of diameter σ and nonzero intensity. When r_0 is bounded and $\sigma \to 0$, they move with an infinite speed proportional to ln σ. The case of one vortex ring of intensity proportional to $\ln^{-1} \sigma$ has been studied and it has been proven that as $\sigma \to 0$ the smoke ring moves with a constant speed.[37] (For more results on a single vortex ring see for instance Refs. 38–43, and the reference therein.) Heuristically in this limit many vortex rings do not interact. When r_0 grows the system can have some nontrivial limit. When $r_0 \approx |\ln \sigma|$ the fluid ''converges'' (formally) to a dynamical system studied in a recent paper.[44] When r_0 increases, it is reasonable to expect that with *some* dependence of r_0 on σ we obtain the point vortex model. The main result of the present paper is that this dependence is *very weak* in such a way that the two limits $(r_0 \to \infty, \sigma \to 0)$ are almost independent. In fact in the next section we prove that when $r_0 \approx \sigma^{-\beta}$ *for any* $\beta > 0$ the solutions of the Euler Equation remain ''localized'' and converge as a measure to the system (1.1). We remark that the validity of this result for any β means essentially that the point vortex model is ''stable'' with respect of such curvature perturbations.

It is interesting also to obtain the ''stability'' of the point vortex model with respect to a viscosity perturbation. This problem has been studied in the planar case Refs. 45,46.

In the next section we give the exact statement of the problem studied in the present paper. The proof is quite similar to the planar case, but it differs by an essential point: the interaction between the fluid particles differs from the one of the planar case because of a term in the Green function. This term is similar to an external field already studied in other papers but, unfortunately, it is *not Lipschitz,* while in the planar case the smoothness of the field was essential. So we need some improvements and for completeness we prefer to sketch here the whole proof.

II. STATEMENTS AND PROOFS

We consider an incompressible inviscid fluid of unitary density moving in $\mathbb{R}^3$. The Euler equation reads

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \boldsymbol{\nabla}) \mathbf{u} = -\boldsymbol{\nabla} p, \tag{2.1}$$

$$\boldsymbol{\nabla} \cdot \mathbf{u} = 0 \tag{2.2}$$

with the initial and the boundary conditions. Here $\mathbf{u} = \mathbf{u}(\mathbf{x},t)$ denotes the velocity field and $p = p(\mathbf{x},t)$ the pressure.

In the present paper we assume that the velocity $\mathbf{u}$ decays at infinity.

We introduce the vorticity $\boldsymbol{\omega}$ as

$$\boldsymbol{\omega} = \nabla \wedge \mathbf{u}. \tag{2.3}$$

We can reconstruct the velocity field by means on the vorticity $\boldsymbol{\omega}$ as

$$\mathbf{u}(\mathbf{x},t) = -\frac{1}{4\pi} \int_{R^3} d\mathbf{y} \, \frac{(\mathbf{x}-\mathbf{y}) \wedge \boldsymbol{\omega}(\mathbf{y},t)}{|\mathbf{x}-\mathbf{y}|^3}. \tag{2.4}$$

We introduce cylindrical coordinates (z, r, θ) and we suppose that the initial velocity field is axisymmetric without swirl, i.e.,

$$\mathbf{u}(\mathbf{x},t)=(u_z(z,r,t),u_r(z,r,t),0). \tag{2.5}$$

We observe that the time evolution conserves the symmetry and Eqs. (2.1), (2.2), (2.3) become

$$\boldsymbol{\omega}=\nabla\wedge\mathbf{u}=(0,0,\omega_\theta)=(0,0,\partial_z u_r-\partial_r u_z), \tag{2.6}$$

$$\partial_t\omega_\theta+(u_z\partial_z+u_r\partial_r)\alpha_\theta-\frac{u_r\omega_\theta}{r}=0, \tag{2.7}$$

$$\partial_z(ru_z)+\partial_r(ru_r)=0. \tag{2.8}$$

From now on we denote ω_θ by ω. We want to study the system when no strong property of regularity is assumed on the initial data, and so we consider a weak formulation of the Euler equation. A first one follows from the observation that Eq. (2.7) means the conservation during the motion of the quantity ω/r,

$$\frac{\omega(z(0),r(0),0)}{r(0)}=\frac{\omega(z(t),r(t),t)}{r(t)}, \tag{2.9}$$

where $z(t)$ and $r(t)$ are the time evolution of $r(0)$ and $r(0)$ according to the velocity field given by Eq. (2.4), i.e.,

$$\dot{z}=u_z,\quad \dot{r}=u_r. \tag{2.10}$$

Equations (2.4), (2.9), (2.10) give a weak formulation of the Euler equation. Another equivalent weak formulation is given by a (formal) integration by parts of (2.7),

$$\partial_t\omega_t[f]=\omega_t[u_z\partial_z f+u_r\partial_r f+\partial_t f], \tag{2.11}$$

where $f=f(z,r,t)$ is a bounded smooth test function and

$$\omega_t[f]=\int_{-\infty}^{\infty}dz\int_0^{\infty}dr\ f\omega. \tag{2.12}$$

From Eq. (2.4) we obtain

$$u_z(z,r,t)=-\frac{1}{2\pi}\int_{-\infty}^{\infty}dz'\int_0^{\infty}r'\ dr'\int_0^{\pi}d\theta\,\frac{\omega(z',r',t)(r\cos\theta-r')}{[(z-z')^2+(r-r')^2+2rr'(1-\cos\theta)]^{3/2}}, \tag{2.13}$$

$$u_r(z,r,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}dz'\int_0^{\infty}r'\ dr'\int_0^{\pi}d\theta\,\frac{\omega(z',r',t)(z-z')\cos\theta}{[(z-z')^2+(r-r')^2+2rr'(1-\cos\theta)]^{3/2}}. \tag{2.14}$$

We now introduce a change of variables

$$z=x,\quad r=r_0+y;\quad \mathbf{x}=(x,y), \tag{2.15}$$

where r_0 will be precise in the sequel.

In the proofs discussed later on it is useful to outline the difference between the velocity fields given by Eqs. (2.13), (2.14) and the planar case in which

$$\mathbf{u}_0(\mathbf{x},t)=\int\mathbf{K}(\mathbf{x}-\mathbf{x}')\omega(\mathbf{x}',t)d\mathbf{x}', \tag{2.16}$$

where

$$\mathbf{K}(\mathbf{x})=\left(-\frac{1}{2\pi}\frac{y}{|\mathbf{x}|^2},\ \frac{1}{2\pi}\frac{x}{|\mathbf{x}|^2}\right). \tag{2.17}$$

We write the velocity field as

$$\mathbf{u}(\mathbf{x},t)=\int(\mathbf{K}(\mathbf{x}-\mathbf{x}')+\mathbf{D}(\mathbf{x},\mathbf{x}'))\omega(\mathbf{x}',t)d\mathbf{x}', \tag{2.18}$$

where the previous equation is a definition of **D**.

We consider initial data of the form

$$\omega_\sigma(\mathbf{x},0)=\sum_{i=1}^{N}\omega_{\sigma,i}(\mathbf{x},0), \tag{2.19}$$

where $\omega_{\sigma,i}(\mathbf{x},0)$ is a function with a definite sign supported in a region $\Lambda_{\sigma,i}$ such that

$$\Lambda_{\sigma,i}\equiv\text{supp }\omega_{\sigma,i}\subset\Sigma(\mathbf{x}_i|\sigma), \tag{2.20}$$

where

$$(x_i,y_i)\neq(x_j,y_j)\quad\text{if } i\neq j$$
$$\sigma<\tfrac{1}{2}\min_{i,j\ \ i\neq j}((x_i-x_j)^2+(y_i-y_j)^2)^{1/2} \tag{2.21}$$

and consequently

$$\Sigma(\mathbf{x}_i|\sigma)\cap\Sigma(\mathbf{x}_j|\sigma)=\emptyset\quad\text{if } i\neq j. \tag{2.22}$$

Here $\Sigma(\mathbf{x}|\sigma)$ denotes a disk of center ($\mathbf{x}$) and radius σ.

We assume that initially

$$|\omega_{\sigma,i}(\mathbf{x},0)|\leq M\sigma^{-\gamma},\ \ 0<M<\infty,\ \ 2\leq\gamma<\infty, \tag{2.23}$$

and

$$\int_{-\infty}^{\infty}dx\int_{-r_0}^{\infty}dy\ \ \omega_{\sigma,i}(\mathbf{x},0)=a_i, \tag{2.24}$$

where $a_i\in\mathbb{R}$, $i=1,...,N$, and a_i is called the "intensity of the i-annulus of vorticity."

We study now the time evolution of the Euler equation with the initial conditions (2.19) when σ becomes very small and r_0 becomes very large,

$$r_0=\text{const }\sigma^{-\beta}, \tag{2.25}$$

where from now on const denotes a constant independent of σ.

Depending on the size of β, we have different behavior of the solutions as $\sigma\to 0$: if $r_0=\text{const}$ the velocity field (2.13) and (2.14) is unbounded and so the problem has no meaning, if $r_0=\text{const}\ |\log\sigma|$ we obtain, in a formal way, a dynamical system widely studied in Ref. 44, if r_0 is given by Eq. (2.25) we obtain the point vortex model as we will see in this section.

Theorem 2.1: *Let assumption (2.25) hold and denote by $\omega_\sigma(\mathbf{x},t)$ the time evolution of $\omega_\sigma(\mathbf{x},0)$ according to the Euler equation. Then, for any fixed $T>0$ and for any $\alpha<\min(\frac{1}{3},\beta/3)$:*

(1) There exists $C(\alpha,T)$ such that for $0\leq t\leq T$

$$\text{supp }\omega_{\sigma,i}(\mathbf{x},t)\subset\Sigma(\mathbf{x}_i(t)|d), \tag{2.26}$$

where

$$d=C(\alpha,T)\sigma^{\alpha}, \tag{2.27}$$

and $\mathbf{x}_i(t)$ *is the solution of the ordinary system (1.1) provided that such a solution exists up to the time T.*

(2) For any continuous bounded function $f(\mathbf{x})$,

$$\int d\mathbf{x}\ \omega_{\sigma,i}(\mathbf{x},t)f(\mathbf{x})\rightarrow\sum_{i=1}^{N} a_i f(\mathbf{x}_i(t)) \quad\text{as}\quad \sigma\rightarrow 0. \tag{2.28}$$

Proposition (1) states that the blobs of vorticity remain concentrated until time T. Proposition (2) states that

$$\omega_{\sigma}(\mathbf{x},t)\rightarrow\sum_{i=1}^{N} a_i\delta(\mathbf{x}_i(t)) \quad\text{as}\quad \sigma\rightarrow 0 \tag{2.29}$$

weakly in the sense of the measures, where $\delta(\cdot)$ denotes the Dirac measure. This last statement gives a rigorous justication of the point vortex model as a limit of large annuli.

We remark that the singular nature of the right-hand side of Eq. (1.1), diverging when two vortices are close, does not guarantee the existence of the solution of Eq. (1.1) for every time. In many cases (for instance, for all $a_i>0$) collapses are forbidden by the first integrals of motion, but there are cases in which singularities happen. However it can be proved that the collapses are exceptional.[28] We can say that Theorem 2.1 holds up to the time T for which the solution of Eq. (1.1) exists.

Proof: The proof is similar to that studied in Refs. 32,34, but the nonregularity of the interaction in the cylindrical case imposes some improvements and hence, for completeness, we sketch it wholly.

Initially the blobs of vorticity are disjoint. We follow the evolution of one of them. Its center of vorticity moves under the action of the curvature of the annulus and of the other vortices. We simulate the influence of the other vortices by an external field and we study in details the time evolution of a single annulus of vorticity under a suitable external field. We will prove (in the next Theorem) that it remains concentrated. Finally it is easy to obtain from this result the proof of Theorem 2.1.

We consider a single blob of unitary vorticity moving in an external, divergence-free, uniformly bounded, time dependent vector field satisfying the Lipschitz condition velocity field $\mathbf{F}(\mathbf{x},t)$,

$$|\mathbf{F}(\mathbf{x},t)-\mathbf{F}(\mathbf{y},t)|\leq L|\mathbf{x}-\mathbf{y}|, \quad L>0. \tag{2.30}$$

Equation (2.10) becomes

$$\dot{\mathbf{x}}(t)=\mathbf{u}(\mathbf{x},t)+\mathbf{F}(\mathbf{x},t) \tag{2.31}$$

while Eqs. (2.9), (2.4) remain unchanged. The Euler equation in weak form reads also

$$\dot{\omega}[f]=\omega[(\mathbf{u}+\mathbf{F})\cdot\nabla f]+\omega[\partial_t f]. \tag{2.32}$$

We prove proposition (1) of Theorem 2.1 for this particular evolution. Define the center of vorticity of the blob as

$$\mathbf{B}_{\sigma}(t)=\int \mathbf{x}\omega_{\sigma}(\mathbf{x},t)d\mathbf{x}. \tag{2.33}$$

Theorem 2.2: *Suppose that*

$$\int \omega_\sigma(\mathbf{x},0)d\mathbf{x}=1, \tag{2.34}$$

$$\text{supp } \omega_\sigma(\mathbf{x},0)\subset\Sigma(\mathbf{x}^*|\sigma), \tag{2.35}$$

and

$$|\omega_\sigma(\mathbf{x},0)|\leq \text{const } \sigma^{-\gamma} \quad 2\leq\gamma<\infty. \tag{2.36}$$

Then, for any fixed $T>0$ and for any $\alpha<\min(\frac{1}{3},\beta/3)$

(1) there exists $C(\alpha,T)$ such that for $0\leq t\leq T$,

$$\text{supp } \omega_\sigma(\mathbf{x},t)\subset\Sigma(\mathbf{B}(t)|d), \tag{2.37}$$

where

$$d=C(\alpha,T)\sigma^\alpha \tag{2.38}$$

and $\mathbf{B}(t)$ is the solution of the ordinary system

$$\dot{\mathbf{B}}(t)=\mathbf{F}(\mathbf{B}(t),t) \quad \mathbf{B}(0)=\mathbf{x}^*. \tag{2.39}$$

Moreover,

$$|\mathbf{B}_\sigma(t)-\mathbf{B}(t)|\to 0 \text{ as } \sigma\to 0 \textit{ uniformly in } t\in[0,T], \textit{ at least as } \sigma^{\beta'} \tag{2.40}$$

for any $\beta'<\beta$.

Proof: The difficulty of the proof arises from the singularity of the kernel $\mathbf{K}+\mathbf{D}$ which forces a fluid particle to rotate with a very large velocity around the center of vorticity. To overcome this difficulty we study the motion of the center of vorticity, which turn out to be more regular than the motion of a given fluid particle. Moreover we control the spreading of the vorticity around the center of vorticity by using the moment of inertia, which is almost conserved during the motion. However, as we will see, this control is not enough and it must be improved to acheive the proof.

Theorem 2.2 states an asymptotic result as $\sigma\to 0$ and hence without lack of generality we can suppose that $\sigma\ll 1$.

The more difficult case arises for β small and so in the proof we suppose, without lack of generality, that $\beta<1$.

We start supposing they exist y_M, y_m, independent of σ such that $|y_M|<\infty$, $|y_m|<\infty$, and up to time t, $O\leq t\leq T$, the support of the vorticity remains bounded,

$$\text{supp } \omega(\mathbf{x}(t),t)\subset\{\mathbf{x}|y_m\leq y\leq y_M\}. \tag{2.41}$$

At the end of this section we shall prove that this hypothesis is fulfilled.

We remark that the vorticity remains constant during the motion,

$$\int \omega_\sigma(\mathbf{x}(t),t)d\mathbf{x}=\int \omega_\sigma(\mathbf{x}(0),0)d\mathbf{x}=1 \tag{2.42}$$

as we can see by using Eq. (2.11).

We introduce the moment of inertia $I_\sigma(t)$ with respect of $\mathbf{B}_\sigma(t)$,

$$I_\sigma(t)=\int \omega_\sigma(\mathbf{x},t)(\mathbf{x}-\mathbf{B}_\sigma(t))^2 d\mathbf{x} \tag{2.43}$$

and we study the growth in time of $\mathbf{B}_\sigma(t)$ and $I_\sigma(t)$. By using the Euler equation, we have

$$\dot{\mathbf{B}}_\sigma(t)=\int d\mathbf{x}[\mathbf{u}(\mathbf{x},t)+\mathbf{F}(\mathbf{x},t)]\omega_\sigma(\mathbf{x},t), \tag{2.44}$$

$$\begin{aligned}\dot{I}_\sigma(t)&=2\int d\mathbf{x}\{(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot(\mathbf{u}(\mathbf{x},t)+\mathbf{F}(\mathbf{x},t))-(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\dot{\mathbf{B}}_\sigma(t)\}\omega_\sigma(\mathbf{x},t)\\&=2\int d\mathbf{x}\ \omega_\sigma(\mathbf{x},t)(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\left(\mathbf{u}(\mathbf{x},t)-\int d\mathbf{x}'\ \omega_\sigma(\mathbf{x}',t)\mathbf{u}(\mathbf{x}',t)\right)\\&\quad+2\int d\mathbf{x}\ \omega_\sigma(\mathbf{x},t)(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\left(\mathbf{F}(\mathbf{x},t)-\int d\mathbf{x}'\ \omega_\sigma(\mathbf{x}',t)\mathbf{F}(\mathbf{x}',t)\right).\end{aligned} \tag{2.45}$$

We study in detail Eq. (2.45). The terms with $\mathbf{F}$ can be easily bounded by using the Lipschitz condition, the Cauchy–Schwarz inequality and the remark that

$$\int d\mathbf{x}\ \omega_\sigma(\mathbf{x},t)(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot(\mathbf{F}(\mathbf{B}_\sigma(t),t)-\mathbf{F}(\mathbf{x}',t))=0. \tag{2.46}$$

We obtain

$$\begin{aligned}&\left|2\int d\mathbf{x}\ \omega_\sigma(\mathbf{x},t)(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\left(\mathbf{F}(\mathbf{x},t)-\int d\mathbf{x}'\ \omega_\sigma(\mathbf{x}',t)\mathbf{F}(\mathbf{x}',t)\right)\right|\\&\quad=\left|2\int d\mathbf{x}\ \omega_\sigma(\mathbf{x},t)(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\int d\mathbf{x}'\ \omega_\sigma(\mathbf{x}',t)[\mathbf{F}(\mathbf{x},t)-\mathbf{F}(\mathbf{x}',t)]\right|\leqslant 2LI_\sigma(t).\end{aligned} \tag{2.47}$$

To evaluate the terms in $\mathbf{u}$ we use the following two properties discussed in the Appendix:

$$\left|\int d\mathbf{x}\ \mathbf{u}(\mathbf{x},t)\omega_\sigma(\mathbf{x},t)\right|\leqslant\text{const }\sigma^\beta\ \ln|\sigma|\leqslant\text{const }\sigma^{\beta'}\quad\text{for any}\quad\beta'<\beta, \tag{2.48}$$

$$\left|\int d\mathbf{x}\ \omega_\sigma(\mathbf{x},t)(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\mathbf{u}(\mathbf{x},t)\right|\leqslant\text{const }\sigma^{\beta'}I_\sigma^{1/2}\quad\text{for any}\quad\beta'<\beta. \tag{2.49}$$

In conclusion,

$$|\dot{I}_\sigma(t)|\leqslant 2LI_\sigma(t)+\text{const }\sigma^{\beta'}I_\sigma^{1/2} \tag{2.50}$$

and then

$$I_\sigma(t)\leqslant\text{const }\sigma^{2\beta'}. \tag{2.51}$$

The next steps to obtain Eq. (2.40) are similar to that used in Ref. 32 to obtain Eq. (3.16) and we omit them.

We have obtained Eq. (2.51), which says that the main part of the vorticity is concentrated around the center of vorticity. *A priori* small filaments of vorticity could go far away. We want to prove that this not the case and the support of the vorticity remains concentrated around the center. For this purpose we study the radial part of the velocity field near the boundary of the support of the vorticity and we prove that the difference between this field and the velocity field acting on the center of the vortex vanishes as $\sigma\to 0$. So, the particle paths cannot go far apart from $\mathbf{B}_\sigma$. The radial field is essentially due to three terms; the velocity produced by the external field, the velocity produced by the particles near the center of the vortex and the velocity produced by the particle near the boundary. The first contribution is easily controlled by the Lipschitz condition, the second contribution vanishes as the vorticity is sharply concentrated, and the third contribution needs more care and vanishes after an iterative procedure, which is given in the sequel.

We study the growth in time of the distance of the particle in $\mathbf{x}\in \text{supp}\ \omega(\mathbf{x},t)$ farthest from $\mathbf{B}_\sigma$,

$$\left|(\mathbf{u}(\mathbf{x},t)+\mathbf{F}(\mathbf{x},t)-\dot{\mathbf{B}}_\sigma(t))\cdot \frac{\mathbf{x}-\mathbf{B}_\sigma(t)}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\right|$$
$$\leqslant\left|\mathbf{F}(\mathbf{x},t)-\int d\mathbf{y}\ \omega_\sigma(\mathbf{y},t)\mathbf{F}(\mathbf{y},t)\right|+\left|\left(\mathbf{u}(\mathbf{x},t)-\int d\mathbf{y}\ \omega_\sigma(\mathbf{y},t)\mathbf{u}(\mathbf{y},t)\right)\cdot \frac{\mathbf{x}-\mathbf{B}_\sigma(t)}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\right|$$
$$\leqslant\left|\int d\mathbf{y}\ \omega_\sigma(\mathbf{y},t)(\mathbf{F}(\mathbf{x},t)-\mathbf{F}(\mathbf{y},t))\right|+\left|\mathbf{u}(\mathbf{x},t)\cdot \frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\right|+\text{const}\ \sigma^{\beta'}. \tag{2.52}$$

The first contribution due to the external field can be controlled by using the Lipschitz condition,

$$\left|\int d\mathbf{y}\ \omega_\sigma(\mathbf{y},t)(\mathbf{F}(\mathbf{x},t)-\mathbf{F}(\mathbf{y},t)\right|\leqslant \text{const}\ R, \tag{2.53}$$

where

$$R\equiv|\mathbf{x}-\mathbf{B}_\sigma(t)|. \tag{2.54}$$

We study now the other terms,

$$\left|\frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\cdot\mathbf{u}(\mathbf{x},t)\right|=\left|\frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\cdot(\mathbf{u}_0(\mathbf{x},t)+(\mathbf{u}(\mathbf{x},t)-\mathbf{u}_0(\mathbf{x},t))\right|, \tag{2.55}$$

where $\mathbf{u}_0(\mathbf{x},t)$ is the planar velocity field defined in Eq. (2.16). The last term can be bounded by using the following equation:

$$|(\mathbf{u}(\mathbf{x},t)-\mathbf{u}_0(\mathbf{x},t))|\leqslant \text{const}\ \sigma^{\beta'}, \tag{2.56}$$

a consequence of the estimates of the Appendix.

We evaluate the term $|(\mathbf{x}-\mathbf{B}_\sigma(t))/|\mathbf{x}-\mathbf{B}_\sigma(t)|\cdot\mathbf{u}_0(\mathbf{x},t)|$. We divide the circle $\Sigma(\mathbf{B}_\sigma(t),R)$ into many different annuli,

$$\Sigma(\mathbf{B}_\sigma(t)|R)=\sum_{k=1}^{k*}[\Sigma(\mathbf{B}_\sigma(t)|a_k)-\Sigma(\mathbf{B}_\sigma(t)|a_{k-1})]\cup[\Sigma(\mathbf{B}_\sigma(t)|R)-\Sigma(\mathbf{B}_\sigma(t)|a_{k*})], \tag{2.57}$$

where

$$a_0=0,\ \ a_1=\sigma^{\beta'},\ \ a_k=2a_{k-1}. \tag{2.58}$$

We choose $k*$ such that $a_{k*+1}\leqslant R$ and $a_{k*+2}>R$.

The radial velocity can be expressed by the sum of the contribution obtained when the particles are contained in each annulus,

$$\frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\cdot\int_{[\Sigma(\mathbf{B}_\sigma(t)|a_k)-\Sigma(\mathbf{B}_\sigma(t)|a_{k-1})]}\mathbf{K}(\mathbf{x}-\mathbf{x}')\,\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'$$
$$=\frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\cdot\int_{[\Sigma(\mathbf{B}_\sigma(t)|a_k)-\Sigma(\mathbf{B}_\sigma(t)|a_{k-1})]}\mathbf{K}(\mathbf{x}-\mathbf{B}_\sigma(t))\,\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'$$
$$+\frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\cdot\int_{[\Sigma(\mathbf{B}_\sigma(t)|a_k)-\Sigma(\mathbf{B}_\sigma(t)|a_{k-1})]}(\mathbf{K}(\mathbf{x}-\mathbf{x}')-\mathbf{K}(\mathbf{x}-\mathbf{B}_\sigma(t)))\,\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'. \tag{2.59}$$

The first term in the right-hand side of Eq. (2.59) vanishes because of $\mathbf{x}\cdot\mathbf{K}(\mathbf{x})=0$. Moreover by the explicit form of $\mathbf{K}(\mathbf{x})$, we have

$$|\mathbf{K}(\mathbf{x}-\mathbf{x}')-\mathbf{K}(\mathbf{x})|<\text{const}\,\frac{\rho}{|\mathbf{x}|(|\mathbf{x}|-\rho)}\quad\text{if}\ \ |\mathbf{x}'|<\rho<|\mathbf{x}|. \tag{2.60}$$

Hence

$$\frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\cdot\int_{[\Sigma(\mathbf{B}_\sigma(t)|a_k)-\Sigma(\mathbf{B}_\sigma(t)|a_{k-1})]}(\mathbf{K}(\mathbf{x}-\mathbf{x}')-\mathbf{K}(\mathbf{x}-\mathbf{B}_\sigma(t)))\,\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'$$
$$<\text{const}\,\frac{a_k}{R(R-a_k)}\int_{[\Sigma(\mathbf{B}_\sigma(t)|a_k)-\Sigma(\mathbf{B}_\sigma(t)|a_{k-1})]}\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'. \tag{2.61}$$

The last integral can be bounded by $I_\sigma(t)$. In fact it is obvious that

$$I_\sigma(t)\geqslant r^2 m_t(r), \tag{2.62}$$

where

$$m_t(r)=\int_{R^2-\Sigma(\mathbf{B}_\sigma(t)|r)}\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'. \tag{2.63}$$

Equation (2.51) implies

$$m_t(r)\leqslant\text{const}\ \sigma^{2\beta'}r^{-2}. \tag{2.64}$$

Hence

$$\int_{[\Sigma(\mathbf{B}_\sigma(t)|a_k)-\Sigma(\mathbf{B}_\sigma(t)|a_{k-1})]}\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'<\text{const}\,\frac{\sigma^{2\beta'}}{a_{k-1}^2},\quad k>1. \tag{2.65}$$

We put Eq. (2.65) in Eq. (2.61) and we obtain

$$\frac{(\mathbf{x}-\mathbf{B}_\sigma(t))}{|\mathbf{x}-\mathbf{B}_\sigma(t)|}\cdot\int_{\Sigma(\mathbf{B}_\sigma(t)|a_{k*})}\mathbf{K}(\mathbf{x}-\mathbf{x}')\,\omega_\sigma(\mathbf{x}',t)\,d\mathbf{x}'<\text{const}\,\frac{\sigma^{2\beta'}}{R^2}. \tag{2.66}$$

Now we prove that the vorticity mass close to the boundary of the support is very small and so it can produce a very weak velocity field.

To control the vorticity flux we introduce for any $R>0$ the following non-negative function $G_R\in C^\infty(\mathbb{R}^2)$, $\mathbf{r}\rightarrow G_R(\mathbf{r})$ depending only on $|\mathbf{r}|$, defined as

$$G_R(\mathbf{r})=\begin{cases}1 & \text{if } |r|>2R\\ 0 & \text{if } |r|<R\end{cases}\tag{2.67}$$

such that for some $C_1>0$,

$$|\nabla G_R(\mathbf{r})|<\frac{C_1}{R}\tag{2.68}$$

$$|\nabla G_R(\mathbf{r})-\nabla G_R(\mathbf{r}')|<\frac{C_1}{R^2}\,|\mathbf{r}-\mathbf{r}'|.\tag{2.69}$$

We define the quantity

$$\mu_t(R)=\int G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\,\omega_\sigma(\mathbf{x},t)\,d\mathbf{x}.\tag{2.70}$$

We choose $\mu_t(R)$ as a measure of the localization of $\omega_\sigma(\mathbf{x},t)$ around $\mathbf{B}_\sigma(t)$. In fact if supp $\omega_\sigma(\mathbf{x},t)\subset\Sigma(\mathbf{B}_\sigma(t)|R)$, then $\mu_t(R)=0$. We evaluate its time derivative by using Eq. (2.11),

$$\begin{aligned}\dot{\mu}_t(R)&=\int d\mathbf{x}\;\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot(\mathbf{u}(\mathbf{x},t)+\mathbf{F}(\mathbf{x},t))-\dot{\mathbf{B}}_\sigma(t))\,\omega_\sigma(\mathbf{x},t)\\&=\int d\mathbf{x}\;\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\,\omega_\sigma(\mathbf{x},t)\cdot\int d\mathbf{x}'\;\mathbf{K}(\mathbf{x}-\mathbf{x}')\,\omega_\sigma(\mathbf{x}',t)\\&\quad+\int d\mathbf{x}\;\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\,\omega_\sigma(\mathbf{x},t)\cdot\int d\mathbf{x}'\;\mathbf{D}(\mathbf{x}-\mathbf{x}')\,\omega_\sigma(\mathbf{x}',t)\\&\quad-\int d\mathbf{x}\;\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\int d\mathbf{x}'\;\mathbf{u}(\mathbf{x}',t)\,\omega_\sigma(\mathbf{x}',t)\\&\quad+\int d\mathbf{x}\;\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\,\omega_\sigma(\mathbf{x},t)\cdot\left[\left(\mathbf{F}(\mathbf{x},t)-\int d\mathbf{x}'\;\mathbf{F}(\mathbf{x}',t)\right)\omega_\sigma(\mathbf{x}',t)\right].\end{aligned}\tag{2.71}$$

We estimate the first term in the right-hand side of Eq. (2.71). By the antisymmetry of **K**, it can be written as

$$\frac{1}{2}\int\int d\mathbf{x}\,d\mathbf{x}'(\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))-\nabla G_R(\mathbf{x}'-\mathbf{B}_\sigma(t))\cdot\mathbf{K}(\mathbf{x}-\mathbf{x}')\,\omega_\sigma(\mathbf{x},t)\,\omega_\sigma(\mathbf{x}',t).\tag{2.72}$$

To estimate this term for $R=\sigma^{\beta'}2^{n-1}$, we split the integration domain in the following sets:

$$T_h=\{(\mathbf{x},\mathbf{x}')|\mathbf{x}\notin\Sigma(\mathbf{B}_\sigma(t)|R),\;\;\mathbf{x}'\in(\Sigma(\mathbf{B}_\sigma(t)|a_h)-\Sigma(\mathbf{B}_\sigma(t)|a_{h-1}))\quad\text{if } h<n,\tag{2.73}$$

$$T_h=\{(\mathbf{x},\mathbf{x}')|\mathbf{x}\notin\Sigma(\mathbf{B}_\sigma(t)|R),\;\;\mathbf{x}'\notin\Sigma(\mathbf{B}_\sigma(t)|a_{h-1})\quad\text{if } h=n,\tag{2.74}$$

$$S_h=\{(\mathbf{x},\mathbf{x}')|\mathbf{x}'\notin\Sigma(\mathbf{B}_\sigma(t)|R),\;\;\mathbf{x}\in(\Sigma(\mathbf{B}_\sigma(t)|a_h)-\Sigma(\mathbf{B}_\sigma(t)|a_{h-1})\quad\text{if } h<n,\tag{2.75}$$

$$S_h=\{(\mathbf{x},\mathbf{x}')|\mathbf{x}'\notin\Sigma(\mathbf{B}_\sigma(t)|R),\;\;\mathbf{x}\notin\Sigma(\mathbf{B}_\sigma(t)|a_{h-1})\quad\text{if } h=n.\tag{2.76}$$

Remark that the integrand in Eq. (2.72) vanishes in the complement of $\cup_{h=1}^{n}(T_h\cup S_h)$.

Thanks to the identities $\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\cdot\mathbf{K}(\mathbf{x}-\mathbf{B}_\sigma(t))=0$ and $\nabla G_R(\mathbf{x}'-\mathbf{B}_\sigma(t))=0$ if $\mathbf{x}'\in(\Sigma(\mathbf{B}_\sigma(t)|a_h)-\Sigma(\mathbf{B}_\sigma(t)|a_{h-1})$, $h<n$, the contribution to the integral due to T_h, $h<n$ is bounded by

$$\left| \int d\mathbf{x} \int_{\Sigma(\mathbf{B}_\sigma(t)|a_h) - \Sigma(\mathbf{B}_\sigma(t)|a_{h-1})} d\mathbf{x}' \nabla G_R(\mathbf{x} - \mathbf{B}_\sigma(t)) \right.$$
$$\left. \cdot (\mathbf{K}(\mathbf{x}-\mathbf{x}') - \mathbf{K}(\mathbf{x} - \mathbf{B}_\sigma(t)))\, \omega_\sigma(\mathbf{x},t)\, \omega_\sigma(\mathbf{x}',t) \right|. \tag{2.77}$$

We now use Eq. (2.68), the fact that $\nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))=0 \quad \text{if } |\mathbf{x}-\mathbf{B}_\sigma(t)|<R$, and we obtain the bound

$$(2.72) < \text{const}\, \frac{m_t(R)}{R} \left(\text{const}\, \frac{\sigma^{2\beta'}}{R^2} + \sum_{h=2}^{n-1} \frac{a_h}{R(R-a_h)} \frac{\sigma^{2\beta'}}{a_{h-1}^2} \right) < \text{const}\, \frac{\sigma^{\beta'}}{R^3}\, m_t(R). \tag{2.78}$$

To estimate the contribution due to T_n, we use the obvious inequality $|\mathbf{K}(\mathbf{x})| \le |\mathbf{x}|^{-1}$, Eq. (2.69) and the bound

$$|(\nabla G_R(\mathbf{x}) - \nabla G_R(\mathbf{x}')) \cdot \mathbf{K}(\mathbf{x}-\mathbf{x}')| < \frac{\text{const}}{R^2}. \tag{2.79}$$

We obtain this contribution smaller than const $(\sigma^{2\beta'}/R^4)m_t(R)$. We can handle in the same way the term with S_h. We study now the second term in the right-hand side of Eq. (2.71) that is smaller than

$$\frac{\text{const}\ \sigma^{\beta'}}{R}\, m_t(R) \tag{2.80}$$

as a consequence of estimates of the Appendix.

By using Eq. (2.48) the third term is smaller than

$$\text{const}\, \frac{\sigma^{\beta'}}{R}\, m_t(R). \tag{2.81}$$

Finally we study the last term in Eq. (2.71). We consider two cases; either $|\mathbf{x}'-\mathbf{B}_\sigma(t)|>R$ or $|\mathbf{x}'-\mathbf{B}_\sigma(t)|\le R$. In the first case,

$$\left| \int d\mathbf{x}\ \nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\, \omega_\sigma(\mathbf{x},t) \int d\mathbf{x}' (\mathbf{F}(\mathbf{x},t) - \mathbf{F}(\mathbf{x}',t))\, \omega_\sigma(\mathbf{x}',t) \right| < \text{const} \|\mathbf{F}\|_\infty \frac{\sigma^{2\beta'}}{R^3}\, m_t(R). \tag{2.82}$$

In the second case, by using the Lipschitz condition,

$$\left| \int d\mathbf{x}\ \nabla G_R(\mathbf{x}-\mathbf{B}_\sigma(t))\, \omega_\sigma(\mathbf{x},t) \int d\mathbf{x}' (\mathbf{F}(\mathbf{x},t) - \mathbf{F}(\mathbf{x}',t))\, \omega_\sigma(\mathbf{x}',t) \right| < \text{const}\ m_t(R). \tag{2.83}$$

In conclusion, we have

$$\dot{\mu}_t(R) \le A(R)\, m_t(R), \tag{2.84}$$

where

$$A(R) = \text{const}\, \frac{\sigma^{\beta'}}{R^3} + \text{const}\, \frac{\sigma^{2\beta'}}{R^4} + \text{const}. \tag{2.85}$$

We observe now that

$$m_t(R) \leqslant \mu_t\left(\frac{R}{2}\right). \tag{2.86}$$

We put Eq. (2.86) in the integral form of Eq. (2.85) and we obtain

$$\mu_t(R) \leqslant \mu_0(R) + A(R) \int_0^t \mu_{t_1}\left(\frac{R}{2}\right) dt_1. \tag{2.87}$$

We start an iterative procedure

$$\begin{aligned} \mu_t(R) &\leqslant \mu_0(R) + A(R) \int_0^t \mu_{t_1}\left(\frac{R}{2}\right) dt_1 \\ &\leqslant \mu_0(R) + A(R)t + A(R)A(R/2) \int_0^t dt_1 \int_0^{t_1} \mu_{t_2}\left(\frac{R}{4}\right) dt_2 \end{aligned} \tag{2.88}$$

and so on.

We start from $R = \text{const } \sigma^{\alpha}$, where $\alpha < \min(\frac{1}{3}, \sigma^{\beta'}/3)$. We iterate Eq. (2.84) n times, where n is chosen such that $n \to \infty$ as $\sigma \to 0$ and in the same time $A(R2^{-k})$ is bounded for any positive integer $k \leqslant n$ and $\mu_0(R2^{-n}) = 0$. We choose

$$n = \text{Integer part of } \left[-\frac{1-3\alpha}{4} \log_2 \sigma^{\beta'} \right]. \tag{2.89}$$

Then

$$R2^{-n} = \text{const } \sigma^{[\beta'(1+\alpha)]/4} \tag{2.90}$$

and

$$A(R2^{-k}) \leqslant \text{const} \tag{2.91}$$

for any positive integer $k \leqslant n$. Hence after n iterations we have

$$m_t(R) \leqslant \frac{(\text{const})^n}{n!} \to 0 \text{ faster than any power in } \sigma. \tag{2.92}$$

In conclusion we have proven that the vorticity mass becomes very small near the boundary of the support. We bound the velocity field produced by it,

$$\begin{aligned} &\left| \int_{\Sigma(\mathbf{B}_\sigma(t)|R) - \Sigma(\mathbf{B}_\sigma(t)|a_{k^*})} d\mathbf{x}' (\mathbf{K}(\mathbf{x}-\mathbf{x}') + \mathbf{D}(\mathbf{x},\mathbf{x}')) \omega_\sigma(\mathbf{x}',t) \right| \\ &\qquad \leqslant \text{const} \left| \int_{\Sigma(\mathbf{B}_\sigma(t)|R) - \Sigma(\mathbf{B}_\sigma(t)|a_{k^*})} d\mathbf{x}' |\mathbf{x}-\mathbf{x}'|^{-1} \omega_\sigma(\mathbf{x}',t) \right|. \end{aligned} \tag{2.93}$$

The integrand is monotonically unbounded as $\mathbf{x} \to \mathbf{x}'$ and so the maximum of the integral is obtained when we rearrange the vorticity mass as closed as possible to the singularity,

$$\left| \int_{\Sigma(\mathbf{B}_\sigma(t)|R) - \Sigma(\mathbf{B}_\sigma(t)|a_{k^*})} d\mathbf{x}' |\mathbf{x}-\mathbf{x}'|^{-1} \omega_\sigma(\mathbf{x}',t) \right| \leqslant \text{const } \sigma^{-\gamma} \left| \int_{\Sigma(\mathbf{O}|\eta} d\mathbf{x}' |\mathbf{x}'|^{-1}, \tag{2.94}$$

where $\mathbf{O}$ denotes the origin and η is such that

$$M\sigma^{-\gamma}\pi\eta^2=m_t(a_{k*}). \tag{2.95}$$

By using Eq. (2.92) we have that

$$\left|\int_{\Sigma(\mathbf{B}_\sigma(t)|R)-\Sigma(\mathbf{B}_\sigma(t)|a_{k*})} d\mathbf{x}'|\mathbf{x}-\mathbf{x}'|^{-1}\omega_\sigma(\mathbf{x}',t)\right|\to 0 \text{ faster than any power.} \tag{2.96}$$

We are now able to bound the radial velocity of a particle at a distance R from $\mathbf{B}_\sigma(t)$. We sum the partial bounds we have obtained and we have

$$\dot{R}\leqslant \text{const } R+\text{const }\frac{\sigma^{\beta'}}{R^2}+\text{terms smaller than any power in } \sigma \text{ when } R>\text{const }\sigma^{\alpha}. \tag{2.97}$$

Hence for $R>\text{const }\sigma^\alpha$ the last two terms in the right-hand side of Eq. (2.97) are neglectable and this differential inequality gives bound (2.37) by using the Gronwall Lemma. □

We return to the proof of Theorem 2.1. It is similar to that of the planar case and we only sketch it. We denote R_m the minimal distance between point vortices evolving via Eq. (1.1) and we choose $\sigma\ll R_m$. Initially the vortices are separeted and we study one vortex. We simulate the influence of other vortices on it as an external fields. We observe that actually the other vortices produce a velocity field depending on σ but this dependence is very small. Moreover each vortex moves in a bounded region of the space and so assumption (2.41) is satisfied. Then, it is easy to prove the convergence stated in Theorem 2.1. □

ACKNOWLEDGMENT

Work performed under the auspices of the Italian Council of Researches (C.N.R.-G.N.F.M.) and the Ministry of the University (M.U.R.S.T).

APPENDIX

We will use the following two integrals I_1 and I_2:

$$I_1=\int_0^\pi d\theta\,\frac{\cos\,\theta}{[a^2+2(1-\cos\,\theta)]^{3/2}}=\int_0^\pi d\theta\,\frac{\cos\frac{\theta}{2}+\left(\cos\,\theta-\cos\frac{\theta}{2}\right)}{\left[a^2+4\left(\sin\frac{\theta}{2}\right)^2\right]^{3/2}}. \tag{A1}$$

The first term is easily computed,

$$\int_0^\pi d\theta\,\frac{\cos\frac{\theta}{2}}{\left[a^2+4\left(\sin\frac{\theta}{2}\right)^2\right]^{3/2}}=\frac{2}{a^2(a^2+4)^{1/2}}. \tag{A2}$$

The remaining terms are less divergent for small a (see the study of I_2) and so, for small a, $I_1\approx a^{-2}$,

$$I_2 = \int_0^\pi d\theta \, \frac{(1-\cos\,\theta)}{[a^2+2(1-\cos\,\theta)]^{3/2}}$$
$$= \int_0^\pi d\theta \, \frac{2\left(\sin\frac{\theta}{2}\right)^2 \cos\frac{\theta}{2} + \left[1-\cos\,\theta - 2\left(\sin\frac{\theta}{2}\right)^2 \cos\frac{\theta}{2}\right]}{\left[a^2+4\left(\sin\frac{\theta}{2}\right)^2\right]^{3/2}}. \tag{A3}$$

The first term is easily computed

$$\int_0^\pi d\theta \, \frac{2\left(\sin\frac{\theta}{2}\right)^2 \cos\frac{\theta}{2}}{\left[a^2+4\left(\sin\frac{\theta}{2}\right)^2\right]^{3/2}} = -\frac{1}{(a^2+4)^{1/2}} + \frac{1}{2}\ln[2+(a^2+4)^{1/2}] - \frac{1}{2}\ln a. \tag{A4}$$

The other terms for small a are bounded and so $I_2 \approx -\frac{1}{2}\ln a$.

We are now able to obtain the bound (2.48). We use definition (2.18) and we study separately terms with **K** and with **D**. The term with **K** vanishes as we can easily prove, writing explicitely Eq. (2.48), interchanging **x** and **x**′ and using the antisymmetry of **K**. We evaluate the term with **D**. We use the explicit form of **u**, we take the leading term in σ^{-1}, and we obtain by using the previous eximates on I_1 and I_2 a bound of the form,

$$(2.48) \leq \text{const } \sigma^\beta + \text{const } \sigma^\beta \int\int\int d\mathbf{x}\, d\mathbf{x}' \;\; \omega_\sigma(\mathbf{x},t)\,\omega_\sigma(\mathbf{x}',t)(\text{const } \sigma^\beta + |\ln|\mathbf{x}-\mathbf{x}'||). \tag{A5}$$

To give a bound to the last integral we use the assumption (2.36), we perform a symmetrical rearrangement on the vorticity around the point (x,y), that is we concentrate the vorticity in a disk centered in the singularity (x,y) having a radius η such that the total vorticity is preserved, i.e., $(M/\sigma^\gamma)\pi\eta^2 = 1$. We perform the integral in the polar coordinate and it is smaller than const ln $\sigma^{-\beta}$. We put this bound in Eq. (A5) and we obtain the bound (2.48) [and in the same time Eq. (2.56)].

Equation (2.49) is obtained in the same way. In addition, in this case in the last step we use the Cauchy-Schwarz inequality.

[1] H. Helmholtz, ''On the integrals of the hydrodynamical equations which express vortex motion,'' Philos. Mag. **33**, 485 (1867).
[2] G. Kirchhoff, *Vorlesungen uber Math. Phys.* (Leipzig, Teuber, 1876).
[3] H. Poincarè, *Thèories des Tourbillons* (George Carrè, Paris, 1983).
[4] J. Kelvin, *Mathematical and Physical Papers* (Cambridge University Press, Cambridge, 1910).
[5] C. C. Lin, *The Motion of the Vortices in Two Dimensions* (University of Toronto Press, Toronto, 1943).
[6] J. L. Synge, ''On the motion of three vortices,'' Can. J. Math. **1**, 257–270 (1949).
[7] H. Aref, ''Motion of three vortices,'' Phys. Fluids **22**, 393–400 (1979).
[8] E. A. Novikov, ''Dynamics and statistics of a system of vortices,'' Sov. Phys. JETP **41**, 937–943 (1976).
[9] S. Ziglin, ''Nonintegrability of a problem of four vortices,'' Sov. Math. Dokl. **1**, 296–299 (1980).
[10] K. M. Khanin, ''Quasi-periodic motions of vortex systems,'' Physica D **4**, 261–269 (1982).
[11] H. Aref, ''Integrable, chaotic, and turbulent vortex motion in two-dimensional flows,'' Annu. Rev. Fluid Mech. **15**, 345–389 (1983).
[12] M. S. A. Castilla, V. Moauro, P. Negrini, and W. M. Oliva, ''The four vortices problem: Regions of chaotic behavior and the nonintegrability,'' Ann. Inst. Henri Poincaré: Phys. Theor. **59**, 99–115 (1993).
[13] A. J. Chorin, ''Numerical study of slightly viscous flow,'' J. Fluid Mech. **57**, 785–796 (1973).
[14] T. E. Dushane, ''Convergence for a vortex method for solvings Euler's equation,'' Math. Comput. **27**, 719–728 (1973).
[15] V. M. Del Prete and O. H. Hald, ''Convergence of vortex methods for Euler's equations,'' Math. Comput. **32**, 791–809 (1978).
[16] J. T. Beale and A. Majda, ''Vortex method. I. Convergence in three dimensions,'' Math. Comput. **39**, 1–27 (1982).
[17] J. T. Beale and A. Majda, ''Vortex Method. II. High-order accuracy in two and three dimensions,'' Math. Comput. **29**, 29–52 (1982).

[18] C. Marchioro and M. Pulvirenti, ‘‘Hydrodynamics in two dimensions and vortex theory,’’ Commun. Math. Phys. **84**, 483–503 (1982).
[19] E. A. Novikov, ‘‘Generalized dynamics of three-dimensional vortical singularities (vortons),’’ Sov. Phys. JETP **57**, 566–569 (1983).
[20] C. R. Anderson and C. Greengard, ‘‘On the vortex method,’’ SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **22**, 413–440 (1985).
[21] M. J. Aksman, E. A. Novikov, and S. A. Orzag, ‘‘Vorton method in three-dimensional hydrodynamics,’’ Phys. Rev. Lett. **54**, 2410–2413 (1985).
[22] P. G. Saffman and D. I. Meiron, ‘‘Difficulties with three-dimensional weak solutions for inviscid incompressible flow,’’ Phys. Fluids **29**, 2373–2375 (1986).
[23] G. Winckelmans and A. Leonard, ‘‘Weak solutions of the three-dimensional vorticity equation with vortex singularity,’’ Phys. Fluids **31**, 1838–1839 (1988).
[24] C. Greengard and E. Thomann, ‘‘Singular vortex systems and weak solutions of the Euler equations,’’ Phys. Fluids **31**, 2810–2813 (1988).
[25] J. Goodman, T. Y. Hou, and J. Lowengrub, ‘‘Convergence of the point vortex method for 2D Euler equations,’’ Commun. Pure Appl. Math. **43**, 415–430 (1990).
[26] T. Y. Hou and J. Lowengrub, ‘‘Convergence of the point vortex mehtod for 3D Euler equations,’’ Commun. Pure Appl. Math. **43**, 965–981 (1990).
[27] C. Marchioro and M. Pulvirenti, ‘‘Euler evolution for singular data and vortex theory,’’ Commun. Math. Phys. **91**, 563 (1983).
[28] C. Marchioro and M. Pulvirenti, ‘‘Vortex method in two-dimensional fluid mechanics,’’ Lect. Notes Phys. **203**, 1–132 (1984).
[29] C. Marchioro and E. Pagani, ‘‘Evolution of two concentrated vortices in a two-dimensional bound domain,’’ Math. Methods Appl. Sci. **8**, 328–344 (1986).
[30] B. Turkington, ‘‘On the evolution of a concentrated vortex in an ideal fluid,’’ Arch. Ration. Mech. Anal. **97**, 75–87 (1987).
[31] C. Marchioro, ‘‘Euler evolution for singular data and vortex theory: A global solution,’’ Commun. Math. Phys. **116**, 45 (1988).
[32] C. Marchioro and M. Pulvirenti, ‘‘Vortices and localization in Euler flows,’’ Commun. Math. Phys. **154**, 49–61 (1993).
[33] C. Marchioro and M. Pulvirenti, ‘‘Mathematical theory of incompressible nonviscous fluids,’’ Appl. Math. Sci. **96**, 1–283 (Springer-Verlag, Berlin, 1994).
[34] C. Marchioro, ‘‘On the localization of the vortices,’’ Bollet. Unione Matematica Italiana (in press).
[35] P. Serfati, ‘‘Borne en temps des caractéristiques de l’équation d’Euler 2D à tour billion positif et localisation pour le modèle point-vortex (1998) (preprint).
[36] R. Klein, A. J. Majda, and K. Damodaran, ‘‘Simplified equations for the interaction of nearly parallel vortex filaments,’’ J. Fluid Mech. **288**, 201–248 (1995).
[37] D. Benedetto, E. Caglioti, and C. Marchioro, ‘‘On the motion of a vortex ring with a sharply concentrated vorticity,’’ University of Rome ‘‘La Sapienza,’’ Department of Math. (preprint) (1997), Math. Methods Appl. Sci. (submitted).
[38] L. E. Fraenkel, ‘‘On steady vortex rings of a small-cross section in an ideal fluid,’’ Proc. R. Soc. London, Ser. A **316**, 29–62 (1970).
[39] L. E. Fraenkel and M. S. Berger, ‘‘A global theory of steady vortex rings in an ideal fluid,’’ Acta Math. **132**, 13–51 (1974).
[40] L. E. Fraenkel and M. S. Berger, ‘‘Nonlinear Desingularization in Certen Free-Boundary Problems,’’ Commun. Math. Phys. **77**, 149–172 (1980).
[41] W.-M. Ni, ‘‘On the Existence of Global Vortex Rings,’’ J. Anal. Math. **37**, 208–247 (1980).
[42] A. Ambrosetti and M. Struwe, ‘‘Existence of steady rings in an ideal fluid,’’ Arch. Ration. Mech. Anal. **108**, 97–108 (1989).
[43] K. Shariff and A. Leonard, ‘‘Vortex Rings,’’ Annu. Rev. Fluid Mech. **24**, 235–279 (1992).
[44] C. Marchioro and P. Negrini, ‘‘On a dynamical system related to fluid dynamics,’’ Nota Interna Dipartimento di Matematica, Università ‘‘La Sapienza’’ 97/40(1998) (preprint), Nonlinear Diff. Eq. Appl. (submitted).
[45] C. Marchioro, ‘‘On the vanishing viscosity limit for two-dimensional Navier–Stokes equations with singular initial data,’’ Math. Methods Appl. Sci. **12**, 463–470 (1990).
[46] C. Marchioro, ‘‘On the inviscid limit for a fluid with a concentrated vorticity,’’ Commun. Math. Phys. **196**, 53–65 (1998).

Exact periodic solutions of the complex Ginzburg–Landau equation

A. V. Porubov
Instituto Pluridisciplinar, Universidad Complutense de Madrid, Madrid 28040, Spain and Institute of High-Performance Computing and Data Bases, Russian Ministry of Sciences, P.O. Box 71, St. Petersburg, 194291 Russia

M. G. Velarde
Instituto Pluridisciplinar, Universidad Complutense de Madrid, Paseo Juan XXIII, 1, Madrid 28040, Spain

(Received 29 December 1997; accepted for publication 26 October 1998)

Three new exact periodic solutions of the complex Ginzburg–Landau equation are obtained in terms of the Weierstrass elliptic function $\wp$. Furthermore, the new periodic solutions and other shock solutions appear as their bounded limits (along the real axis) for particular relationships between the coefficients in the equation. In the general case, bounded limits are nothing but the already known pulse, hole, and shock solutions. It is also shown that the shapes of the solutions are quite different from the shape of the usual envelope wave solution. In particular, the spatial structure of the new bounded periodic solutions varies in time, while the pulse solution may exhibit breather-like behavior. © *1999 American Institute of Physics.* [S0022-2488(99)02302-6]

I. INTRODUCTION

Most of the mathematical work in the realm of nonlinear phenomena refers to integrable equations and their exact solutions, particularly periodic. Among the recently developed general methods, the algebro-geometrical approach may be used in an efficient way to find such solutions. Not only the numerical realization and graphical representation of the solution is provided by this method, but also multiphase quasiperiodic solutions as well as purely periodic ones may be represented using the algebrogeometrical approach, as illustrated in Ref. 1.

It is only recently that due to significant progress achieved in the study of patterns and waves in nonequilibrium, dissipative systems the need appeared of the study of equations nonintegrable by means of the Inverse Scattering Transform Method (ISTM). A paradigmatic case is the complex Ginzburg–Landau equation (CGLE).

$$\iota u_t + p u_{xx} + q|u|^2 u = \iota \gamma u, \tag{1}$$

where the constant coefficients are $p = p_r + \iota p_i$, $q = q_r + \iota q_i$, with p_j, $q_j \neq 0$, $(u,p,q) \in \mathbb{C}$, $\gamma \in \mathbb{R}$. The subscripts t and x denote temporal and spatial derivatives, respectively. This equation appears in the description of a large variety of physical phenomena,[2,3] (and references therein). Here we concentrate on obtaining *exact* solutions of (1). Moreover, we shall not study any *degeneracies* of CGLE, corresponding to vanishing of any of p_j, q_j, or γ, because these cases are already studied in the literature (see Refs. 4 and 5 and references therein). Exact solutions usually appear as a result of some balance between the significant terms in the equation. In particular, CGLE (1), besides nonlinear ($q_r|u|^2u$) and dispersive ($p_r u_{xx}$) terms, contains nonlinear saturation ($q_i|u|^2u$) and diffusion ($p_i u_{xx}$) terms, as well as the linear growth term (γu); thus solitary wave and shocks are possible solutions of (1). Exact solutions may describe not only the propagation of nonlinear waves but also spatially localized structures of permanent shape that may be of interest to experimenters.[6–8] Finally, the knowledge of particular exact solutions may be useful when numerically exploring unsteady processes governed by Eq. (1).

As the CGLE (1) is nonintegrable, only particular *exact* solutions may be obtained. To find such solutions various *direct* methods seem promising. Indeed, the pulse solution of (1) with *nonzero coefficients* was obtained in Ref. 9 using a special ansatz, while the shock-wave solution was found in Ref. 10 by direct integration. A modified Hirota bilinear method has been applied in Ref. 11, and three exact solutions, pulse, hole, and shock solution, respectively, were found. Further, the development of a movable critical points analysis or Painlevé analysis[12] led to direct methods based on the nature of the poles of the solution (see, e.g., Refs. 5, 13, and references therein). However, neither the application of the truncated Painlevé expansions[13] nor another extension of the Painlevé analysis proposed in Ref. 5, have led to new exact solutions of the CGLE. Moreover, no bounded periodic solutions of (1) have been found other than usual harmonic wave solution or the so-called phase winding solution.[14]

Particular periodic solutions of nonintegrable equations may be obtained using direct methods based either on the application of a suitable change of variables bringing equations under study to those possessing already known exact periodic solutions or by using an appropriate ansatz for the solution. In order to construct the ansatz, a clue may come from the Painlevé analysis, which is based, in particular, on looking for solutions whose movable critical points are poles only. Therefore the use of elliptic functions in the ansatz is rather natural because they are the most general functions having such singular points. However, complications appear due to the need of using in the ansatz *four* theta functions; see, e.g., Ref. 15, or *three* Jacobi elliptic functions; see, e.g., Ref. 16. The possibility exists to look for solutions using just *one* Weierstrass elliptic function $\wp$.[17,18]

The paper is organized as follows. In Sec. II a procedure is proposed to obtain exact solutions using the Weierstrass elliptic $\wp$ function. A transformation of CGLE (1) into two coupled equations for the amplitude and the phase parts of the solution is provided in Sec. III, and possibilities of their decoupling are found. In Sec. IV, the procedure introduced in Sec. II is applied to obtain exact solutions in terms of the Weierstrass elliptic function for the amplitude and the phase equations. Three new solutions are obtained. It is shown in Sec. V that their periodic and localized limits bounded along the real axis may be represented in terms of Jacobi elliptic functions and hyperbolic functions. The profiles of these bounded solutions are studied in Sec. VI in order to clarify the role of the phase part of the solution on its qualitative behavior. In Sec. VII we summarize our results.

II. WEIERSTRASS ELLIPTIC FUNCTION AS A TOOL FOR OBTAINING EXACT SOLUTIONS

According to its definition,[19] the Weierstrass function is analytical in the complex plane other than in points where it has double poles. It obeys the equation

$$\{\wp'(\zeta)\}^2=4\wp^3-g_2\wp-g_3, \tag{2}$$

where g_2 and g_3 are constants. Any derivative of the function $\wp$ can be written by means of itself. Further, any *elliptic function* f may be expressed using $\wp$ and its first derivative as

$$f=A(\wp)+B(\wp)\wp', \tag{3}$$

where A and B are rational functions with respect to $\wp$.[19] Depending on the ratio between g_2 and g_3, the Weierstrass function may be bounded or unbounded inside the domain under study. The bounded periodic solutions are more conveniently expressed by writing them in terms of the Jacobi elliptic functions cn, sn, and dn, which are bounded along the real axis. For this purpose the relationship between the Weierstrass function and the Jacobi functions is used as a special case of (3). Indeed, the familiar link is obtained in Ref. 19, but using the singular function sn^{-2},

$$\wp(\zeta,g_2,g_3)=e_3+(e_1-e_3)\mathrm{sn}^{-2}(\sqrt{e_1-e_3}\zeta,k). \tag{4}$$

However, following the method introduced in Ref. 19, one can check that

$$\wp(\zeta,g_2,g_3)=e_2-(e_2-e_3)\mathrm{cn}^2(\sqrt{e_1-e_3}\zeta,k), \tag{5}$$

connects the Weierstrass function with the Jacobi function cn, regular along the real axis. Here $k=\sqrt{(e_2-e_3)/(e_1-e_3)}$ is the modulus of the Jacobian elliptic function, while $\tau=e_m$ ($m=1,2,3,e_3\leqslant e_2\leqslant e_1$) are the real roots of the cubic equation,

$$4\tau^3-g_2\tau-g_3=0. \tag{6}$$

Expressing these results in terms of an appropriate choice of parameters, the *wave number* $\kappa=\sqrt{e_1-e_3}$ and the Jacobian elliptic modulus k, we have

$$\begin{gathered} e_1=\frac{2-k^2}{3}\kappa^2, \quad e_2=\frac{2k^2-1}{3}\kappa^2, \quad e_3=-\frac{1+k^2}{3}\kappa^2, \\ g_2=\tfrac{8}{3}\kappa^4(1-k^2+k^4), \quad g_3=\frac{4}{27}\kappa^6(k^2+1)(2-k^2)(1-2k^2). \end{gathered} \tag{7}$$

The localized solutions appear in the limit $k\to 1$ of the Jacobi elliptic functions.

Recently, the Weierstrass function was used for finding periodic solutions by applying the spectral theory for the Láme equation with elliptic potentials,[20] and some exact solutions have been obtained for the coupled nonlinear Schrödinger equations (CNLSEs) at high symmetry. This method can only be applied to integrable nonlinear equations admitting Lax pairs representation.

As the CGLE is known to exhibit chaotic behavior, it is not integrable. However, exact solutions may be found, even for nonintegrable equations using the Weierstrass elliptic function $\wp$. For instance, Samsonov[17,18] obtained some exact solutions of the Korteweg-de Vries–Burgers equation, the Gardner equation, the Hunter–Shroulle equation, and the FitzHugh–Nagumo equation. He provided the relevant set of transformations bringing a wide class of second-order nonlinear ordinary differential equations to just Eq. (2). Moreover, the use of (3) was proposed in Ref. 18 to construct the ansatz for direct substitution in the equation under study. Recently various new exact periodic solutions have been obtained in Ref. 21 for CNLSEs for *arbitrary* values of the coefficients, when CNLSEs are nonintegrable, and for two dissipative–dispersive equations, appearing in convective problems.[22,23] Comparing these results with others, based on the use of theta functions[15] or Jacobi functions,[16] we see that the algebra is drastically simplified when the ansatz for the solution demands only *one* Weierstrass function instead of *four* theta functions or *three* Jacobi elliptic functions. Only self-similar or traveling wave solutions may be found using this method.

The ansatz for the possible solution may be constructed using the information about the poles of the solution. The Painlevé analysis,[12,13] as well as the nature of the poles of the Weierstrass function, ought to be taken into account.[21–23] However, any poles may be modeled by means of various expressions in terms of the Weierstrass function, and the problem of selecting the most ''efficient'' expression is far from being solved.

An alternative is to try to transform the equation under study to the Weierstrass equation (2) or, more generally, to Painlevé-type equations. This procedure is far from simple, especially for dissipative nonlinear equations. As mentioned, earlier some progress has been achieved in this direction in Ref. 18. However, the transformations proposed in Ref. 18 look rather inconvenient when searching for a solution in closed form. Here we propose a different approach. We restrict consideration to second-order ordinary differential equations. Consider the equation

$$y''+y'^2Q_1(y)+y'Q_2(y)+Q_3(y)=0, \tag{8}$$

where Q_i are rational functions with respect to y. The standard reduction of this autonomous equation, $y' = \sqrt{F(y)}$, yields an ''irrational'' and nonautonomous equation for F. Therefore, obtaining a solution of Eq. (8) is unlikely. In order to overcome this problem we propose the following alternative transformation:

$$y = A_1 v + A_2 v'/v + A_3, \quad v' = \sqrt{F(v)}, \tag{9}$$

where A_i are constants. After substitution of (9) into the equation (8), we get

$$R_1(v,F) + \sqrt{F(v)} R_2(v,F) = 0, \tag{10}$$

where both R_1 and R_2 are polynomials in v and F and its first- and second-order derivatives. Equating to zero R_1 and R_2 separately, one can find for this overdetermined system the solution for F, polynomial in v. Then the solution for y (9) may be obtained by direct integration.

III. TRANSFORMATIONS OF THE CGLE

First, we decompose the solution $u(x,t)$ in its amplitude, y, and phase, θ, both real,

$$u = y(\zeta) e^{\iota\theta}, \tag{11}$$

where $\zeta = x - ct$, $\theta = \theta(\zeta,t)$. Substituting (11) into (1) and equating to zero the real and imaginary parts, we get for y and $z \equiv \theta_\zeta$,

$$p_r y_{\zeta\zeta} - p_i(2 y_\zeta z + y z_\zeta) - p_r y z^2 + c y z - \beta y + q_r y^3 = 0, \tag{12}$$

$$p_i y_{\zeta\zeta} + p_r(2 y_\zeta z + y z_\zeta) - p_i y z^2 - c y_\zeta - \gamma y + q_i y^3 = 0, \tag{13}$$

where $\theta_t \equiv \beta = \text{const}(t)$. A further simplification of Eqs. (12), (13) may be achieved if we assume that $\Phi = y^2$ and $\Psi = y^2 z$. Then (12), (13) become

$$2\Phi\Phi_{\zeta\zeta} - \Phi_\zeta^2 - a_1 \Phi\Phi_\zeta + a_2 \Phi\Psi - 4\Psi^2 - a_3\Phi^2 + a_4 \Phi^3 = 0, \tag{14}$$

$$\Psi_\zeta - b_1 \Phi_\zeta - b_2 \Psi + b_3 \Phi + b_4 \Phi^2 = 0, \tag{15}$$

with

$$a_1 = \frac{2 p_i c}{l_1}, \quad a_2 = \frac{4 p_r c}{l_1}, \quad a_3 = \frac{4(p_r\beta + p_i\gamma)}{l_1}, \quad a_4 = \frac{4 l_3}{l_1},$$

$$b_1 = \frac{a_2}{8}, \quad b_2 = \frac{a_1}{2}, \quad b_3 = \frac{p_i\beta - p_r\gamma}{l_1}, \quad b_4 = \frac{l_2}{l_1},$$

$$l_1 = p_r^2 + p_i^2, \quad l_2 = p_r q_i - p_i q_r, \quad l_3 = p_r q_r + p_i q_i.$$

Equation (14) is *quadratic algebraic* for Ψ, but substitution of the expression for Ψ into Eq. (15) results in a very complicated equation for Φ. On the other hand, the form of Eq. (15) provides conditions for Eqs. (14) and (15) to be solved separately. Indeed, in the nonlinear Schrödinger equation (NLSE) limit, $p_i = q_i = \gamma = 0$, Eq. (15) is integrated, giving $\Psi = b_1\Phi + C$, $C = \text{const}$. Two possibilities exist when (15) is solved for the function Ψ if the coefficients of the CGLE do not vanish; either

$$(\text{A.1}) \quad \Psi = C, \quad \text{if } c = 0, \quad l_2 = 0, \quad b_3 = 0;$$

or

$$(\text{A.2}) \quad \Psi = b_1 \Phi, \quad \text{if } l_2 = 0, \quad b_1 b_2 = b_3 .$$

Otherwise we have to deal with Eqs. (14) and (15) together.

Let us now use Painlevé analysis to the equations to learn about the poles of the solutions in the general case. Assume that possible solutions of the equations (14), (15) have a pole at $\zeta = \zeta_0$. Then solutions can be sought near that point as a Laurent series:

$$\Phi = \frac{r_k}{(\zeta - \zeta_0)^k} + \frac{r_{k-1}}{(\zeta - \zeta_0)^{k-1}} + \cdots, \quad \Psi = \frac{h_m}{(\zeta - \zeta_0)^m} + \frac{h_{m-1}}{(\zeta - \zeta_0)^{m-1}} + \cdots, \tag{16}$$

with both k and m real. The order of the pole may be found by substituting the series (16) into the equations (14), (15) and comparing the leading-order terms. When the leading-order nonlinear term and the leading-order derivative term are in balance, then from Eq. (15) we have the condition $m+1=k$, which used in Eq. (14) yields $k=2$, $m=3$. Then, for subsequent action we assume

$$(\text{B}) \quad \Psi = A \Phi_\zeta + B \Phi + S,$$

with A, B, and S parameters yet to be determined.

In cases (A.1), (A.2) we deal with only one equation (14) for just one unknown Φ, while in case (B) Eq. (15) is not satisfied for arbitrary Φ, and we need to solve an overdetermined system of two equations for only one unknown.

IV. EXACT SOLUTIONS OF THE EQUATIONS (14) AND (15) IN TERMS OF THE WEIERSTRASS ELLIPTIC FUNCTION

A. Case (A.1)

Consider case (A.1) with Eq. (15) satisfied identically for an arbitrary function Φ. Substituting $\Psi = C$ into Eq. (14), we get.

$$2\Phi\Phi_{\zeta\zeta} - \Phi_\zeta^2 - 4C^2 - \frac{4\beta}{p_r}\Phi^2 + \frac{4q_r}{p_r}\Phi^3 = 0, \tag{17}$$

where the conditions $c=0$ and $\beta = p_r \gamma / p_i$ have been used. Using the procedure introduced in Sec. II, one obtains from Eq. (10) that $R_2 \equiv 0$, and the solution of equation $R_1 = 0$ is written through the Weierstrass elliptic function $\wp(\zeta, g_2, g_3)$ as

$$\Phi = Q\wp + G, \tag{18}$$

with $Q = -2p_r/q_r$, and $G = 2\gamma/q_i$. The parameter g_2 is free, while C and g_3 obey the relationship

$$\frac{4C^2}{Q^2} = 4\left(\frac{G}{Q}\right)^3 - g_2\left(\frac{G}{Q}\right) + g_3 . \tag{19}$$

Depending on the relationship between g_2 and g_3, the solution (18) describes either bounded or unbounded periodic solutions.[19] Localized limits yield, respectively, either a solitary wave solution or a localized discontinuous solution.

B. Case (A.2)

Equation (15) is again satisfied identically. Substituting $\Psi = b_1 \Phi$ into Eq. (14), we obtain

$$2\Phi\Phi_{\zeta\zeta} - \Phi_\zeta^2 - a_1 \Phi \Phi_\zeta + (a_2 b_1 - 4b_1^2 - a_3)\Phi^2 + a_4 \Phi^3 = 0. \tag{20}$$

The equation (20) belongs to the class of (8). Then, replacing in (9) y with Φ and substituting the transformation into our equation, we obtain an equation of the form (10). Then the conditions $R_1=0$, $R_2=0$ allow us to assume the polynomial functional form for F, and finally yields

$$\Phi = A\wp + \frac{B\wp_\zeta}{\wp + C} + S, \tag{21}$$

with

$$A=-\frac{2p_r}{q_r},\quad B=\frac{p_i p_r c}{3q_r l_1},\quad C=-\frac{p_i^2 c^2}{108 l_1^2},\quad S=-\frac{5p_i^2 p_r c^2}{54 q_r l_1^2},\quad c^2=\frac{36\gamma l_1^2}{p_i(p_i^2-9l_1)},$$
$$g_2=12C^2,\quad 4C^3-g_2C+g_3=0. \tag{22}$$

It follows from (22) that the solution (21) bounded along the real axis corresponds only to the solitary wave solution. Other possibilities are either localized or periodic discontinuities.

C. Case (B)

We have an overdetermined system of two equations. Using the ansatz (B) from Sec. IV B, Eq. (15) reduces to

$$\Phi_{\zeta\zeta} + \alpha_0\Phi_\zeta + \alpha_1\Phi + \alpha_2\Phi^2 + \alpha_3 = 0, \tag{23}$$

with

$$\alpha_0=\frac{B-b_1-Ab_2}{A},\quad \alpha_1=\frac{b_3-Bb_2}{A},\quad \alpha_2=\frac{b_4}{A},\quad \alpha_3=-\frac{Sb_2}{A}.$$

Equation (23) is the ODE reduction of the Korteweg–de Vries–Burgers equation (KdVBE), which is the dynamical system underlying a Helmholtz–Thompson oscillator.[24–26] Its exact solution is[17]

$$\Phi = 6\alpha_0^2/(25\alpha_2)\exp(-2\alpha_0\zeta/5)\wp[\exp(-\alpha_0\zeta/5)+z_0,0,g_3],$$

where z_0 and g_3 are free parameters. Unfortunately, additional conditions are needed for the existence of this solution, such as $\alpha_3=0$ and $\alpha_1=6\alpha_0^2/25$.

Another solution can be obtained using the procedure from Sec. II. Thus, using (9), we get from the conditions $R_1=0$ and $R_2=0$ that again the solution for F is a third-order polynomial, hence

$$\Phi = M\wp + \frac{N\wp_\zeta}{\wp + C} + R, \tag{24}$$

with the parameters defined by either

$$(\text{B.1})\quad M=-6/\alpha_2,\quad N=-3\alpha_0/(5\alpha_2),\quad R=\alpha_0^2/(50\alpha_2)-\alpha_1/(2\alpha_2),\quad g_2=12C^2,$$
$$g_3=8C^3,\quad C=-\alpha_0^2/300,\quad \text{if}\ \ \alpha_3=\alpha_1^2/(4\alpha_2)-9\alpha_0^4/(625\alpha_2);$$

or

$$(\text{B.2})\quad M=-6/\alpha_2,\quad N=0,\quad R=-\alpha_1/(2\alpha_2),\quad g_2=\alpha_1^2/12-\alpha_2\alpha_3/3,\quad g_3\text{-free},\quad \text{if}\ \ \alpha_0=0.$$

Using the ansatz (B) and Eq. (23), Eq. (14) becomes

$$\Phi_\zeta^2+\beta_0\Phi\Phi_\zeta+\beta_1\Phi_\zeta+\beta_2+\beta_3\Phi+\beta_4\Phi^2+\beta_5\Phi^3=0, \tag{25}$$

with

$$\beta_0=\frac{16B-a_2}{8A},\quad \beta_1=\frac{8AS}{1+4A^2},\quad \beta_2=\frac{4S^2}{1+4A^2},\quad \beta_3=\frac{S(8AB-a_1-a_2A)}{A(1+4A^2)},$$

$$\beta_4=\frac{2b_3+a_3A-a_2AB-a_1B+4B^2}{A(1+4A^2)},\quad \beta_5=\frac{2b_4-a_4}{A(1+4A^2)}.$$

Substituting (24) into (25) for case (B.1), we get

$$A=(-3l_3\pm\sqrt{D})/(4l_2),\quad D=8l_2^2+9l_3^2,\quad B=c(3p_r-4Ap_i)/(6l_1),$$

with either

(B.1.a)

$$S=0,\quad c^2=-\frac{36\gamma l_1^2}{p_i(8l_1+p_r^2)},\quad \beta=\frac{p_r\gamma(p_r^2-10l_1)}{p_i(8l_1+p_r^2)}; \tag{26}$$

or

(B.1.b)

$$S=\frac{6cp_i(l_2-2Al_3)}{l_2[l_2(2p_i^2+9p_r^2+24Ap_ip_r)+3Al_3(4p_i^2-9p_r^2)]},$$

$$c^2=\frac{36l_1^2(2Al_3-l_2)\gamma}{p_i[l_2(2p_i^2+9p_r^2+24Ap_ip_r)+3Al_3(4p_i^2-9p_r^2)]},$$

$$\beta=\frac{p_r}{p_i}\gamma+\frac{2p_i+3Ap_r}{6Al_1}c^2. \tag{27}$$

The second possibility is realized when the coefficients of the CGLE satisfy either

$$9(1-l_2)l_3^2=2l_2(3-l_2),$$

or

$$3l_2l_3[l_3(135p_r^2-36p_i^2)-88l_2p_ip_r]+A[8l_2p_ip_r(26l_2+135l_3)$$
$$+135l_3^2(4p_i^2-9p_r^2)+2l_2^2l_3(56p_i^2-99p_r^2)]=0.$$

For case (B.2), we get

$$A=(-3l_3\pm\sqrt{D})/(4l_2),\quad D=8l_2^2+9l_3^2,\quad B=0,\quad S=0,\quad c=0,\quad \alpha_3=0,\quad g_3=\alpha_1^3/212,$$

with

$$\beta=\frac{(p_r-4Ap_i-4A^2p_r)\gamma}{p_i+4Ap_r-4A^2p_i}.$$

These results have been obtained using the MATHEMATICA package.[27] Again, for both cases (B.1) and (B.2) the relationships between g_2 and g_3 ensure the existence of only a localized (not periodic) limit of the solution (24) bounded along the real axis.

V. BOUNDED SOLUTIONS OF THE CGLE

Using solutions (18), (21), and (24), one can obtain solutions (11) of the CGLE (1) in terms of the Weierstrass function. Let us concentrate on solutions u (11) *bounded along the real axis*. A representation of these particular cases is better done in terms of Jacobi elliptic functions using the relationship (5). Further, we need to change the parameter set according to (7). Thus, from (19) and (6) we obtain for the case (A.1),

$$C^2=-Q^2(H_1-e_1)(H_1-e_2)(H_1-e_3), \tag{28}$$

with $H_1\equiv -G/Q$. For $C=0$ the possibilities are $G=-Qe_m$ $(m=1,2,3)$. More generally, H_1 may lie in either $I_1\equiv(-\infty,e_3]$ or $I_2\equiv(e_2,e_1)$, where the right-hand side of Eq. (28) is positive. Then for I_1, we get $H_1=c_3-\delta^2$, $\delta=\delta(C)$, and

$$u=\sqrt{-\frac{2q_r}{p_r}[k^2\kappa^2\,\mathrm{sn}^2(\kappa\zeta,k)+\delta^2]}\exp\iota\theta, \tag{29}$$

with

$$\theta=\frac{p_r\gamma}{p_i}t-\frac{p_rC\Pi[\varphi,n,k]}{2q_r\delta^2\kappa}-\theta_0,\quad \kappa^2=-\frac{3}{1+k^2}\left(\frac{\gamma}{3p_i}+\delta^2\right), \tag{30}$$

$\Pi[\varphi,n,k]$ is the elliptic integral of the third kind, $\sin\varphi=\mathrm{sn}(\kappa\zeta,k)$, and $n=-\kappa^2k^2/\delta^2$.

Similarly, for I_2 we assume that $H_1=e_1-\delta_1^2$, $0\leqslant\delta_1^2\leqslant\kappa_1^2(1-k^2)$, $\delta_1=\delta_1(C)$. The solution is

$$u=\sqrt{\frac{2q_r}{p_r}\{\kappa_1^2\,\mathrm{dn}^2(\kappa_1\zeta,k)-\delta_1^2\}}\exp\iota\theta, \tag{31}$$

with

$$\theta=\frac{p_r\gamma}{p_i}t+\frac{p_rC\Pi[\varphi,n_1,k]}{2q_r(\kappa_1^2-\delta_1^2)\kappa_1}-\theta_0,\quad \kappa_1^2=\frac{3}{2-k^2}\left(\frac{\gamma}{3p_i}+\delta_1^2\right),\quad n_1=\frac{k^2\kappa_1^2}{\kappa_1^2-\delta_1^2}. \tag{32}$$

The only bounded limit corresponds to the case (A.2), namely, a shock-type solution,

$$u=\sqrt{-\frac{2p_r}{q_r}}\,\kappa[1+\tanh(\kappa\zeta)]\exp\iota(\beta t+b_1\zeta-\theta_0), \tag{33}$$

with

$$\beta=\frac{p_r\gamma(p_i^2-10l_1)}{p_i(26l_1+p_r^2)},\quad \kappa=\frac{p_ic}{6l_1},\quad c\ \text{ is defined by (22).}$$

We get for both possibilities (B.1.a) and (B.1.b) the only bounded shock-type solution:

$$u=\sqrt{-\frac{6Al_1}{l_2}}\,\kappa[1-\tanh(\kappa\zeta)]\exp\iota\theta, \tag{34}$$

with $\kappa=\alpha_0/10$, while for θ we obtain

$$\theta=\beta t+B\zeta+2A\ln\left\{\sqrt{-\frac{6Al_1}{l_2}}\,\kappa[1-\tanh(\kappa\zeta)]\right\}$$
$$+S\left(\frac{1}{4[1-\tanh(\kappa\zeta)]^2}+\frac{1}{4[1-\tanh(\kappa\zeta)]}+\frac{1}{8}\ln\left|\frac{1+\tanh(\kappa\zeta)}{1-\tanh(\kappa\zeta)}\right|\right),$$

where the parameters S, c, and β are defined by either (26) or (27). A solution for the case (B.1) yields the shock solution obtained in Ref. 11.

In case (B.2) the bounded limit of our general solution (24) is either the *pulse* solution,

$$u=\sqrt{\frac{6Al_1}{l_2}}\,\kappa\cosh^{-1}(\kappa\zeta)\exp\iota\theta, \tag{35}$$

with

$$\kappa^2=-\frac{\gamma}{2(p_i+4Ap_r-4A^2p_i)},$$

or the *hole* solution,

$$u=\sqrt{-\frac{6Al_1}{l_2}}\,\kappa\tanh(\kappa\zeta)\exp\iota\theta, \tag{36}$$

with

$$\kappa^2=\frac{\gamma}{p_i+4Ap_r-4A^2p_i}$$

in agreement with the results found in Ref. 11.

VI. ANALYSIS OF WAVE PROFILES

Strictly speaking, only the solution (33) for the case (A.2) is a ''genuine'' envelope wave similar to the *bright* soliton of the NLSE. Other solutions have an amplitude-dependent phase, θ. Let us begin by considering the new periodic wave solutions (29) and (31). Because of their similar behavior we shall study in detail (31) only. The first derivative of its real part with respect to x, $v=(\mathrm{Re}\,u)_x$ is

$$v=\frac{Z}{y}\sin\left[\theta+\arcsin\left(\frac{y\,y_x}{Z}\right)\right], \tag{37}$$

with

$$y=\sqrt{\frac{2q_r}{p_r}[\kappa_1^2\,\mathrm{dn}^2(\kappa_1\zeta,k)-\delta_1^2]},\quad Z=\sqrt{y^2y_x^2+C^2}. \tag{38}$$

When $C=0$, the zeros of the first derivative (37) are defined by the zeros of the function y_x and correspond to the zeros of the Jacobi functions cn and sn. Their positions do not change in time, and for u we have harmonic temporal oscillations of the spatially periodic state defined by the amplitude parts of the solutions (31). However, the situation changes dramatically when $C\neq 0$. In this case Z never vanishes, and the zeros of the first derivative are defined by the zeros of

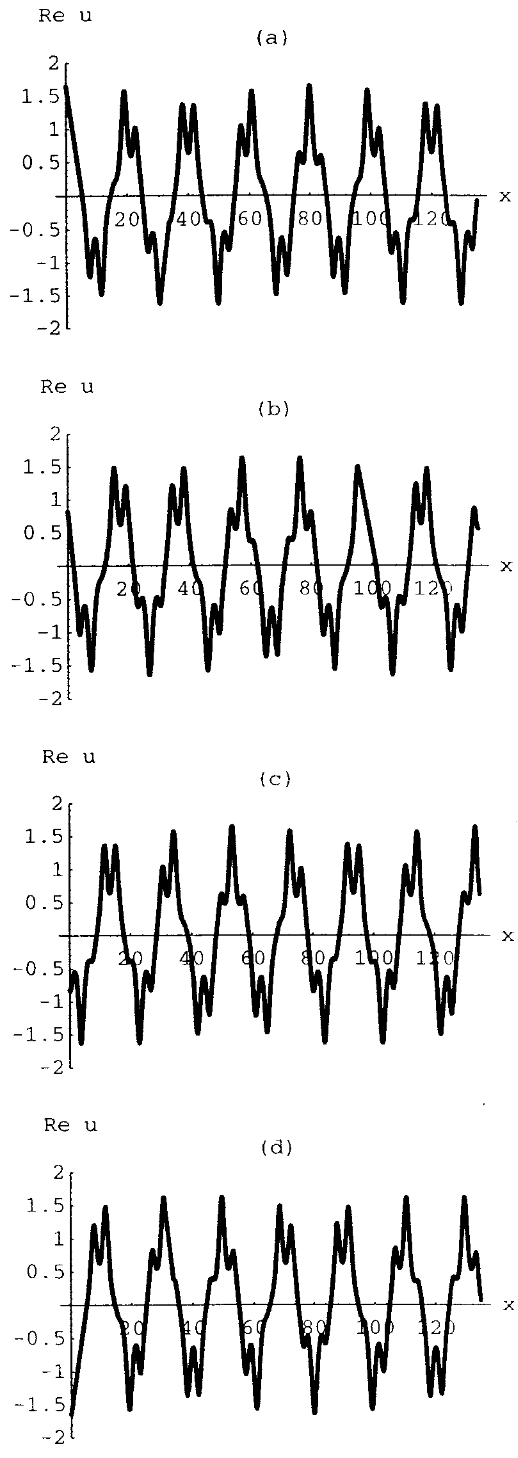

FIG. 1. Evolution of the periodic solution (31) Re u vs x for times $t=mp_i\pi/(3p_r\gamma)$, $0<m<3$ with $k=0.9$, $\delta_1=0.5$. (a) $m=0$, (b) $m=1$, (c) $m=2$, (d) $m=3$.

the sin function only, whose position varies in time. A typical situation is shown in Fig. 1. Figure 1(a) shows a structure with four spatially repeated parts. During half of the time period the shapes of these parts vary, and we get in Fig. 1(d) a profile that is practically the mirror image to Fig. 1(a). Qualitatively, this evolution does not depend upon the value of the modulus k of Jacobi functions.

Consider now the case (B.2) for pulse or hole solutions. In particular, the first derivative for the real part of (35) is

$$(\mathrm{Re}\, u)_x = \sqrt{\frac{6Al_1(1+4A^2)}{l_2}}\,\kappa^2 \cosh^{-1}(\kappa\zeta)\tanh(\kappa\zeta)\sin\left[\theta - \arctan\left(\frac{1}{2A}\right)\right]. \tag{39}$$

Thus, from (39) it follows that additional zeros of the first derivative may appear if

$$\kappa > \sqrt{\frac{l_2}{6Al_1}}\exp\left(\frac{\arctan[1/(2A)]}{2A}\right). \tag{40}$$

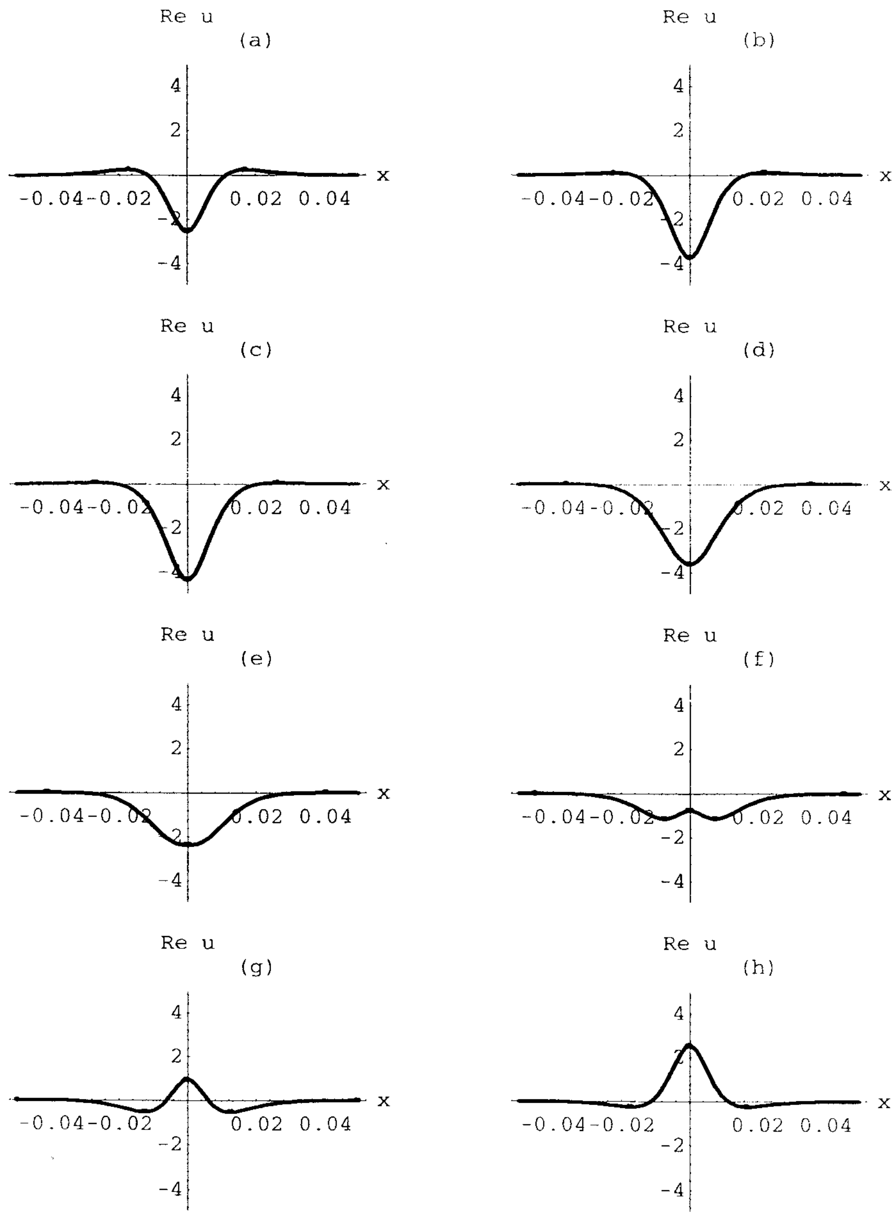

FIG. 2. Evolution of the pulse solution (35) Re u vs x for times $t = m\pi/\beta$, $0 < m < 8$. (a) $m=0$; (b) $m=1$, (c) $m=2$, (d) $m=4$, (e) $m=5$, (f) $m=6$, (g) $m=7$, (h) $m=8$.

The evolution of the real part of the solution (35) is illustrated in Fig. 2. Again, we see that two initial maxima in Fig. 2(a) disappear, Fig. 2(e); then an initial minimum at $\zeta=0$ is changed into a maximum, while two minima arise, Figs. 2(f)–2(h). Therefore, our solution is breather-like. If (40) is not satisfied, there is a pulse solution whose spatial behavior is determined by the function $\cosh^{-1}(\kappa\zeta)$, only with one extremum at $\zeta=0$. A similar situation occurs for the hole solution (36).

The evolution of the solution in case (B.1) depends upon the relationship between β, A, B, and S in the phase θ. However, due to the dependence upon the hyperbolic tangent function, any significant alterations can only occur in the narrow area, $|\kappa\zeta|\ll 1$. Thus, no qualitatively new solutions are found relative to the usual envelope wave.

VII. CONCLUSIONS

New exact general *periodic* solutions of the complex Ginzburg–Landau equation (1), (18), (21), and (24), are obtained in terms of the Weierstrass elliptic function. Among their special limits, we found new *exact periodic* solutions to the CGLE *bounded along the real axis*. Their existence requires additional but nontrivial restrictions on the coefficients of the equation. Accordingly, only a suitable balance between all significant terms in the CGLE is required in order to get a periodic solution. We also found a new shock solution (34), with parameters defined by (27). All bounded solutions early found in Ref. 11 by the modified Hirota method are found also as *particular* limits of the Weierstrass function solutions.

The role of the amplitude dependence in the phase of the solutions has been studied. The form of the solution may strongly differ from the usual envelope solution. In particular, a pulse solution exhibits breather-like behavior. Also, a temporal change in the spatial structure occurs for the bounded periodic solutions.

ACKNOWLEDGMENTS

The authors wish to thank Dr. P. Colinet for useful comments on the manuscript. They are also grateful to the anonymous referee for references and suggestions that helped improve the presentation of the results. This research was carried out while A.V.P. worked as a Visiting Scientist at Instituto Pluridisciplinar, UCM, supported by to the Spanish Ministry of Education and Science. This research has also been supported by DGICYT (Spain) under Grant No. PB 96-0599 and by the European Union under Network Grant No. FRB FM RX-CT96-10.

[1] E. Belokolos, A. Bobenko, V. Enol'skij, A. Its, and V. Matveeev, *Algebro-Geometrical Approach to Nonlinear Integrable Equations* (Springer-Verlag, Berlin, 1994).
[2] W. van Saarlos and P. C. Hohenberg, Physica D **56**, 303 (1992).
[3] M. C. Cross and P. C. Hohenberg, Rev. Mod. Phys. **65**, 851 (1993).
[4] D. Cai, A. R. Bishop, N. Grønbech-Jensen, and B. A. Malomed, Phys. Rev. Lett. **78**, 223 (1997).
[5] R. Conte and M. Musette, Physica D **69**, 1 (1993).
[6] C. O. Weiss, Phys. Rep. **219**, 311 (1992).
[7] S. Fauve and O. Thual, Phys. Rev. Lett. **64**, 282 (1990).
[8] J. J. Niemela, G. Ahlers, and D. S. Cannell, Phys. Rev. Lett. **64**, 1365 (1990).
[9] N. R. Pereira and L. Stenflo, Phys. Fluids **20**, 1733 (1977).
[10] A. V. Klyachkin, ''Modulational instability and autowaves in the active media described by the nonlinear equations of Ginzburg-Landau type,'' unpublished.
[11] K. Nozaki and N. Bekki, J. Phys. Soc. Jpn. **53**, 1581 (1984).
[12] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[13] F. Cariello and M. Tabor, Physica D **39**, 77 (1989).
[14] P. Manneville, *Dissipative Structures and Weak Turbulence* (Academic, New York, 1990).
[15] K. W. Chow, J. Math. Phys. **36**, 4125 (1995).
[16] N. A. Kostov and I. M. Uzunov, Opt. Commun. **89**, 389 (1992).
[17] A. M. Samsonov, in *Nonlinear Waves and Dissipative Effects*, edited by D. Fusco and A. Jeffrey (Longman, London, 1991), p. 123.
[18] A. M. Samsonov, Appl. Anal. **57**, 85 (1995).
[19] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, Cambridge, 1927).
[20] P. L. Christiansen, J. C. Eilbeck, V. Z. Enolskii, and N. A. Kostov, Proc. R. Soc. London, Ser. A **451**, 685 (1995).

[21] A. V. Porubov and D. F. Parker, Wave Motion **29**, 97 (1998).
[22] A. V. Porubov, J. Phys. A **29**, L797 (1993).
[23] A. V. Porubov, Phys. Lett. A **221**, 391 (1996).
[24] R. S. Johnson, Phys. Fluids **15**, 1639 (1972).
[25] E. del Rio, A. Rodriguez-Lozano, and M. G. Velarde, Rev. Sci. Instrum. **63**, 4208 (1992).
[26] C. I. Christov and M. G. Velarde, Physica D **86**, 323 (1995).
[27] S. Wolfram, MATHEMATICA, 2nd ed. (Addison-Wesley, New York, 1991).

Similarity reduction for a class of algebraically special perfect fluids

A. Rainer
Ulmgasse 11, A-8501 Lieboch, Austria

H. Stephani[a)]
Theoretisch-Physikalisches Institut der Universität Jena, Max-Wien-Platz 1, D-07743 Jena, Germany

(Received 9 December 1997; accepted for publication 14 July 1998)

For a class of perfect fluids first considered by Wainwright [Int. J. Theor. Phys. **10**, 39 (1974)], a complete symmetry analysis of the field equations is performed. The results are used for a symmetry reduction of the field equations and the construction of (new) similarity solutions.
[S0022-2488(99)02501-3]

I. INTRODUCTION

In 1974, Wainwright[1] considered a special class of algebraically special perfect fluids for which the multiple null eigenvector **k** is geodesic, twist free, shear free, but expanding. They thus generalize the vacuum Robinson–Trautman solutions and contain them as a limiting case. He was able to reduce the field equations to one or two partial differential equations in three independent variables (given in Sec. II). Up to now only very few solutions of these differential equations have been found; one has been given by Wainwright[1] himself, two others are contained in Kramer[2] and Drauschke.[3]

Here we give a complete symmetry analysis of these equations (Sec. III). The Lie-point-symmetries turn out to show a richer structure than one might have guessed. We then use them to perform a one- or twofold similarity reduction of the field equations (Sec. IV); the resulting ordinary differential equations are listed in the Appendix. In Sec. V, we solve some of these equations and find several new classes of solutions. The results are discussed in Sec. VI.

II. THE METRICS AND DIFFERENTIAL EQUATIONS

The metrics and field equations as given by Wainwright[1] divide into two classes. The first metric and the associated field equations read

$$\begin{gathered} ds^2 = 2\epsilon(r^2-1)P^{-2}(\zeta,\overline{\zeta},u)d\zeta\, d\overline{\zeta} - 2du\, dr - 2H(\zeta,\overline{\zeta},r,u)du^2, \\ H = -r\partial_u \ln P + \epsilon K/2 - \epsilon(r^2-1)\left(b - 3m\int \frac{dr}{(r^2-1)^2}\right) - \epsilon c, \\ P^2\partial_\zeta\partial_{\overline{\zeta}}K + P^2\partial_u\partial_u(P^{-2}) + 6\epsilon m\partial_u \ln P + 2c(K-2c) = 0, \\ \text{with } K = 2P^2\partial_\zeta\partial_{\overline{\zeta}}\ln P, \quad cm=0, \quad c(c+2b)=0, \quad \epsilon=\pm 1 \end{gathered} \tag{1}$$

(ϵ has to be chosen such that the metric has the correct signature), where four-velocity u_i, pressure p and mass density μ are given by

$$u_i\, dx^i = [2\epsilon c(r^2-1) - 2H]^{-1/2}(-dr - \epsilon c(r^2-1)du),$$

[a)]Electronic mail: ste@tpi.uni-jena.de

$$p=\frac{2}{(r^2-1)^2}\left(r\partial_u \ln P+\epsilon c-\frac{\epsilon}{2}K\right)-\frac{2\epsilon b}{r^2-1}(3r^2-2)$$
$$-\frac{2\epsilon m}{r^2-1}\left[\frac{r}{r^2-1}+(3r^2-2)\int\frac{dr}{(r^2-1)^2}\right], \tag{2}$$
$$\mu-p=\frac{4\epsilon b}{r^2-1}(3r^2-1)+\frac{4\epsilon c}{r^2-1}+\frac{4\epsilon m}{r^2-1}\left[\frac{r}{r^2-1}+(3r^2-1)\int\frac{dr}{(r^2-1)^2}\right].$$

For $c=0$, the constant b appearing (via H) in the line element does not enter the main field equation (the fourth-order equation for P); the addition of a term $2b\epsilon(r^2-1)du^2$ to the metric is always possible. This transformation is a generalized Kerr–Schild transformation, i.e., a transformation of the form $\tilde{g}_{ab}=g_{ab}+2H^0k_ak_b$. This property is in fact an extra symmetry of Einstein's equations.

If one performs a coordinate transformation

$$r=\tilde{r}/\lambda,\quad P=\tilde{P}/\lambda,\quad u=\tilde{u}\lambda, \tag{3}$$

in metric (1), followed by the limit $\lambda\rightarrow 0$ (and $b=0$), one arrives at the Robinson–Trautman class of vacuum solutions. For larger r, so that $\epsilon=1$, and regular P and K, pressure and mass density approach $\mu=-p=6b$, i.e., vacuum for $b=0$. In general there will be singularities at $r^2=1$.

The second metric and the associated field equations are given by

$$ds^2=2\epsilon rP^{-2}(\zeta,\bar{\zeta},u)d\zeta\, d\bar{\zeta}-2du\, dr-2H(\zeta,\bar{\zeta},r,u)du^2,$$
$$H=-r\partial_u \ln P-\epsilon ra-\epsilon br\ln\epsilon r-H^0(\zeta,\bar{\zeta},u), \tag{4}$$
$$2P^2\partial_\zeta\partial_{\bar{\zeta}}\ln P-\epsilon\partial_u \ln P=b,\quad 2\partial_\zeta\partial_{\bar{\zeta}}H^0+\epsilon\partial_u(H^0P^{-2})=0,$$

with

$$u_u\, dx^i=-(-2H)^{-1/2}\, dr,$$
$$p=r^{-1}(\tfrac{1}{2}\partial_u \ln P-2\epsilon b-\tfrac{1}{2}\epsilon a)+\tfrac{1}{2}r^{-2}H^0-\tfrac{1}{2}\epsilon br^{-1}\ln\epsilon r, \tag{5}$$
$$\mu-p=2\epsilon r^{-1}(2b+a)+2\epsilon br^{-1}\ln\epsilon r.$$

The terms $H^0+\epsilon ra$ appearing (via H) in the line element do not enter the main field equation for P; they again represent a generalized Kerr–Schild transformation. From the two differential equations the (nonlinear) equation for P is the more difficult one, and we will concentrate on its symmetries; the (linear) equation for H^0 may or may not share these symmetries. The extra symmetry due to linearity of the field equation for H^0 offers the possibility of adding arbitrary solutions H^0, i.e., performing Kerr–Schild transformation. For large r (and regular P and K), these metrics approach vacuum ($\mu=p=0$). Note that the four-velocity is timelike only for $H<0$, but then the coordinate r is necessarily timelike, too.

For both metrics, dust solutions and solutions of Petrov type *III* or *N* are not possible.

III. THE SYMMETRIES AND SUBALGEBRAS

A. The symmetries

The result of the symmetry analysis is that the following generators $\mathbf{X}_a$ can occur as symmetries of the differential equation(s) for the function P [$h(\zeta)$ is an arbitrary function].

Case I: Metric (1) with $c=0=m$, metric (4) with $b=0$:

$$\begin{aligned} \mathbf{X}_1 &= h(\zeta)\partial_\zeta + \overline{h(\zeta)}\partial_{\bar\zeta} + \tfrac{1}{2}[h' + \overline{h'}]P\partial_P, \\ \mathbf{X}_2 &= \partial_u, \quad \mathbf{X}_3^1 = u\partial_u - \tfrac{1}{2}P\partial_P. \end{aligned} \tag{6}$$

Case II: Metric (1) with $c=0$, $m\neq 0$, $\alpha\equiv 3\epsilon m$, metric (4) with $b\neq 0$, $\alpha\equiv 2\epsilon b$:

$$\begin{aligned} \mathbf{X}_1 &= h(\zeta)\partial_\zeta + \overline{h(\zeta)}\partial_{\bar\zeta} + \tfrac{1}{2}[h' + \overline{h'}]P\partial_P, \\ \mathbf{X}_2 &= \partial_u, \quad \mathbf{X}_3^2 = e^{\alpha u}\partial_u - \tfrac{1}{2}\alpha e^{\alpha u}P\partial_P. \end{aligned} \tag{7}$$

Case III: Metric (1) with $c\neq 0$, $m=0$:

$$\mathbf{X}_1 = h(\zeta)\partial_\zeta + \overline{h(\zeta)}\partial_{\bar\zeta} + \tfrac{1}{2}[h' + \overline{h'}]P\partial_P, \quad \mathbf{X}_2 = \partial_u. \tag{8}$$

[$\mathbf{X}_3^1$ is the limit $\alpha\to 0$ of $(\mathbf{X}_3^2 - \mathbf{X}_2)/\alpha$.]

Since it is known that the only symmetries of Einstein's field equations are those corresponding to coordinate transformations and scalings, one would in general also expect that the specialized field equations given above will admit only these symmetries—from experience in other subcases of Einstein's equations one knows that additional symmetries are rare. The symmetries $\mathbf{X}_1$ and $\mathbf{X}_2$ fit into this picture: they correspond to the still existing coordinate freedom in the metrics and lead to Killing vectors. The occurrence of the symmetries $\mathbf{X}_3^1$ and $\mathbf{X}_3^2$ is a surprise. For (4) with $b=0$, $\mathbf{X}_3^1$ corresponds to a scaling combined with a Kerr–Schild transformation; in the other cases they are new symmetries. Their existence will lead to a richer structure than one may have guessed.

The finite transformations belonging to the generators listed above are

$$\begin{aligned} &\mathbf{X}_1: \ \tilde\zeta = h(\zeta), \quad \tilde u = u, \quad \tilde P = |h_{,\zeta}|P, \\ &\mathbf{X}_2: \ \tilde\zeta = \zeta, \quad \tilde u = u + \varepsilon, \quad \tilde P = P, \\ &\mathbf{X}_3^1: \ \tilde\zeta = \zeta, \quad \tilde u = u e^{\varepsilon}, \quad \tilde P = P e^{-\varepsilon/2}, \\ &\mathbf{X}_3^2: \ \tilde\zeta = \zeta, \quad \tilde u = -\frac{1}{\alpha}\ln(e^{-\alpha u} - \alpha\varepsilon), \quad \tilde P = (1 - \alpha\varepsilon e^{\alpha u})^{1/2}P, \end{aligned} \tag{9}$$

ε being the group parameter. These transformations can—and will—be used to simplify generators and metrics.

Obviously the cases I and II are the most interesting ones. Since the symmetry reduction mainly depends on the structure of the Lie algebra, we will discuss the different subalgebras and the corresponding similarity variables for these two cases in the next sections, and only then will we return to the field equations and their solutions.

B. The nonequivalent one-dimensional subalgebras and their similarity variables

By choosing suitable coordinates, one can always make $h(\zeta)=i$. With that choice the following one-dimensional algebras can occur:

Case	Symmetry	Similarity variables
I	$\mathbf{Y}_1 = i(\partial_\zeta - {}_{\bar\zeta})$	$P, u, x=\zeta+\bar\zeta$
	$\mathbf{Y}_2 = \partial_u$	$P, \zeta, \bar\zeta$
	$\mathbf{Y}_3 = u\partial_u - \tfrac{1}{2}P\partial_P$	$w = P\sqrt{u}, \zeta, \bar\zeta$
	$\mathbf{Y}_4 = i(\partial_\zeta - \partial_{\bar\zeta}) + \partial_u$	$P, z=u+i\zeta, \bar z$

	$\mathbf{Y}_5=i(\partial_\zeta-\partial_{\bar\zeta})+u\partial_u-\frac{1}{2}P\partial_P$	$w=P\sqrt{u},\ z=ue^{i\zeta},\ \bar z$
II	$\mathbf{Y}_1=i(\partial_\zeta-\partial_{\bar\zeta})$	$P,u,x=\zeta+\bar\zeta$
	$\mathbf{Y}_2=\partial_u$	$P,\zeta,\bar\zeta$
	$\mathbf{Y}_3=e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P)$	$w=Pe^{\alpha u/2},\zeta,\bar\zeta$
	$\mathbf{Y}_4=i(\partial_\zeta-\partial_{\bar\zeta})+\partial_u$	$P,z=u+i\zeta,\bar z$
	$\mathbf{Y}_5=i(\partial_\zeta-\partial_{\bar\zeta})+e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P)$	$w=Pe^{\alpha u/2},\ z=i\zeta-e^{-\alpha u}/\alpha$
III	$\mathbf{Y}_1=i(\partial_\zeta-\partial_{\bar\zeta})$	$P,u,x=\zeta+\bar\zeta$
	$\mathbf{Y}_2=\partial_u$	$P,\zeta,\bar\zeta$
	$\mathbf{Y}_4=i(\partial_\zeta-\partial_{\bar\zeta})+\partial_u$	$P,z=u+i\zeta,\bar z$

These generators are essentially different in that they cannot be transformed into each other by group transformations.

For the two metrics (with their subcases), a similarity reduction will thus lead to 23 partial differential equations in two independent variables.

C. The nonequivalent two-dimensional subalgebras and their similarity variables

In each case, one can start with one of the normal forms $[\mathbf{X}_1,\mathbf{X}_2]=\mathbf{X}_1$ or $[\mathbf{X}_1,\mathbf{X}_2]=0$ of a two-dimensional algebra, then take for $\mathbf{X}_1$ one of the five different normal forms $\mathbf{Y}_\alpha$ given in the tables and determine the possible different structures of $\mathbf{X}_2$. The results is the following lists.

Case	Symmmetries	Similarity variables
I.1	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}$	P,u
I.2	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}-2\partial_u$	$P,u+\zeta+\bar\zeta$
I.3	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}+u\partial_u-\frac{1}{2}P\partial_P$	$P\sqrt{u},\ ue^{-(\zeta+\bar\zeta)/2}$
I.4	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P$	$P/(\zeta+\bar\zeta),u$
I.5	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P+c_1(u\partial_u-\frac{1}{2}P\partial_P)$	$Pu^{(A-2)/2A},\ (\zeta+\bar\zeta)u^{-1/A}$
I.6	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_u+\zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P$	$Pe^{-u},\ (\zeta+\bar\zeta)e^{-u}$
I.7	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_u$	$P,\zeta+\bar\zeta$
I.8	$i(\partial_\zeta-\partial_{\bar\zeta}),\ u\partial_u-\frac{1}{2}P\partial_P$	$P\sqrt{u},\ \zeta+\bar\zeta$
I.9	$\partial_u,\ u\partial_u-\frac{1}{2}P\partial_P$	($\zeta,\bar\zeta$; no P !)
I.10	$\partial_u,\ i(\partial_\zeta-\partial_{\bar\zeta})+u\partial_u-\frac{1}{2}P\partial_P$	$Pe^{-i(\zeta-\bar\zeta)/4},\ \zeta+\bar\zeta$
I.11	$i(\partial_\zeta-\partial_{\bar\zeta})+\partial_u,\ \zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+\frac{1}{2}P\partial_P+u\partial_u$	$P/\sqrt{\zeta+\bar\zeta},\ \dfrac{2u+i(\zeta-\bar\zeta)}{\zeta+\bar\zeta}$

Case	Symmetries	Similarity variable
II.1	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}$	P,u
II.2	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}-2\partial_u$	$P,u+\zeta+\bar\zeta$
II.3	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}+e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P)$	$Pe^{\alpha u/2},\ \zeta+\bar\zeta+2e^{-\alpha u}/\alpha$
II.4	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P$	$P/(\zeta+\bar\zeta),u$
II.5	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P+e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P)$	$P\ \exp[\alpha u/2+e^{-\alpha u}/\alpha]$, $\alpha\ \ln(\zeta+\bar\zeta)+e^{-\alpha u}$
II.6	$i(\partial_\zeta-\partial_{\bar\zeta}),\ A\partial_u+\zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P$	$Pe^{-u/A},\ (\zeta+\bar\zeta)e^{-u/A}$

II.7	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_u$	$P,\zeta+\bar\zeta$
II.8	$i(\partial_\zeta-\partial_{\bar\zeta}),\ e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P)$	$Pe^{\alpha u/2},\ \zeta+\bar\zeta$
II.9	$\partial_u,\ e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P)$	$(\zeta,\bar\zeta;$ no P !)
II.10	$e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P),\ i(\partial_\zeta-\partial_{\bar\zeta})+\partial_u$ $i(\partial_\zeta-\partial_{\bar\zeta})+e^{\alpha u}(\partial_u-(\alpha/2)\ P\partial_P),$	$Pe^{\alpha[2u+i(\zeta-\bar\zeta)]/4},\ \zeta+\zeta$
II.11	$\zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P-(1/\alpha)\ \partial_u$	$\dfrac{Pe^{\alpha u/2}}{\sqrt{\zeta+\bar\zeta}},\ \dfrac{i(\zeta-\bar\zeta)-2e^{-\alpha u}/\alpha}{\zeta+\bar\zeta}$

Case	Symmetries	Similarity variables
III.1	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}$	P,u
III.2	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_\zeta+\partial_{\bar\zeta}-2\partial_u$	$P,u+\zeta+\bar\zeta$
III.3	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P$	$P/(\zeta+\bar\zeta),u$
III.4	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \zeta\partial_{\bar\zeta}+\bar\zeta\partial_\zeta+P\partial_P+A\partial_u$	$Pe^{-u/A},\ (\zeta+\bar\zeta)e^{-u/A}$
III.5	$i(\partial_\zeta-\partial_{\bar\zeta}),\ \partial_u$	$P,\zeta+\bar\zeta$

D. The nonequivalent three-dimensional subalgebras with a G_3 acting on V_2 (in the space of variables)

Due to the generators $\mathbf{X}=h(\zeta)\partial_\zeta+\overline{h(\zeta)}\partial_{\bar\zeta}+\frac{1}{2}h'+\overline{h'}]P\partial_P$ there are also two three-dimensional algebras with two-dimensional orbits, namely $[i(\partial_\zeta-\partial_{\bar\zeta}),\partial_\zeta+\partial_{\bar\zeta},i(\zeta\partial_\zeta-\bar\zeta\partial_{\bar\zeta})]$, with similarity variables P and u, and $[i(\partial_\zeta-\partial_{\bar\zeta}),\zeta\partial_\zeta+\bar\zeta\partial_{\bar\zeta}+P\partial_P,i(\zeta^2\partial_\zeta-\bar\zeta^2\partial_{\bar\zeta})+i(\zeta-\bar\zeta)P\partial_P]$, with similarity variables $P/(\zeta+\bar\zeta)$ and u. Both have two-dimensional subalgebras and are thus special cases of the types already considered. The algebra $[i(\partial_\zeta-\partial_{\bar\zeta}),\cos\zeta\partial_\zeta+\cos\bar\zeta\partial_{\bar\zeta},i(\sin\zeta\partial_\zeta-\sin\bar\zeta\partial_{\bar\zeta})]$ (two-dimensional representation of the rotation group) is *not* a symmetry of the field equations.

IV. THE REDUCED FIELD EQUATIONS

For each of the five subcases of the two field equations (1) and (4) one can perform the standard symmetry reduction, using the appropriate five (or three) nonequivalent symmetries of Cases I–III given in Section III A (one assumes a functional relationship between the three similarity variables exists in each case, e.g., of the form $P=e^{-\alpha u/2}f(\zeta,\bar\zeta)$ for Case I, $\mathbf{Y}_3$). We will not give a list of all the resulting 23 partial differential equations in two independent variables: they are relatively easy to obtain (see Rainer[4])—but of course nonlinear, and in most cases a solution can be found only when a *twofold* similarity reduction has been performed. Using the nonequivalent two-dimensional subalgebras given in Sec. III C, such a twofold reduction will lead to a set of 45 ordinary differential equations (not all of them are different). They are given in the Appendix.

V. SOLUTIONS AND METRICS

Of course, not all the differential equations given in the Appendix could be solved. Some could, but then one has to be aware that different solutions P may give rise to the same metrics (metrics of the same constant curvature K may appear in different disguises, etc.). Since in all cases two of the (at most) three symmetries are related to Killing vectors, a solution obtained by a twofold reduction will admit at least one Killing vector, and the following classes of metrics will occur:

(1) Space–times with a G_3 on V_2 (spherical, pseudospherical and plane symmetry).

(2) Space–times with one or two Killing vectors.
(3) Space–times without symmetries (e.g., solutions of (4) with $H^0 \neq 0$ which are related by a generalized Kerr–Schild transformation to a background with $H^0=0$ admitting symmetries), see the remarks following Eqs. (2) and (5).

For each of the two metrics, we will treat the first two of these three classes now in turn. We do not intend to discuss the linear differential equation for H^0, i.e., $2\partial_\zeta \partial_{\bar\zeta} H^0 + \epsilon \partial_u (H^0 P^{-2}) = 0$; for most of the functions P given below, many solutions H^0 of this linear equation could be found.

A. Space–times with a G_3 on V_2

As we will show, this high symmetry will occur if the ζ-$\bar\zeta$ space has constant curvature $K = K(u)$; space–time necessarily inherits this symmetry from the ζ-$\bar\zeta$ space. This subcase covers all field equations obtained by reductions using the algebras I.1, I.4, II.1, II.4, III.1 and III.3 for metrics (1), and I.7, I.10, II.7, II.8 and II.10 for the metric (4), and is contained as a special case in several of the other reductions. We will not start from any of these particular cases, but attack the field equations directly.

For constant curvature $K = K(u)$ it is known that the metric function P is of the form

$$P = \eta(u)\zeta\bar\zeta + \beta(u)\zeta + \bar\beta(u)\bar\zeta + \delta(u), \quad K = 2(\eta\delta - \beta\bar\beta), \tag{10}$$

see Kramer *et al.*[5]

In case of the metric (1), the field equation reads

$$P^2 \partial_u \partial_u (P^{-2}) + 6\epsilon m \partial_u \ln P + 2c(K - 2c) = 0, \quad cm = 0. \tag{11}$$

Inserting here the above representation of P and equating to zero the coefficients of the different powers of ζ and $\bar\zeta$, one obtains

$$\begin{aligned}
&\eta\eta'' - 3\eta'^2 - 3\epsilon m \eta\eta' = c(K-2c)\eta^2,\\
&\beta\beta'' - 3\beta'^2 - 3\epsilon m \beta\beta' = c(K-2c)\beta^2,\\
&\delta\delta'' - 3\delta'^2 - 3\epsilon m \delta\delta' = c(K-2c)\delta^2,\\
&\beta\eta'' + \beta''\eta - 6\eta'\beta' - 3\epsilon m(\beta\eta)' = 2c(K-2c)\eta\beta,\\
&\beta\delta'' + \beta''\delta - 6\delta'\beta' - 3\epsilon m(\beta\delta)' = 2c(K-2c)\delta\eta,\\
&\eta\delta'' + \eta''\delta + \beta\bar\beta'' + \beta''\bar\beta - 6(\eta'\delta' + \beta'\bar\beta')\\
&\quad - 3\epsilon m(\eta\delta + \beta\bar\beta)' = 2c(K-2c)(\delta\eta + \beta\bar\beta).
\end{aligned} \tag{12}$$

Eliminating the second derivatives in last three equations by means of the first three gives

$$(\ln \eta)' = (\ln \beta)' = (\ln \delta)', \tag{13}$$

i.e., the coefficients η, β and δ depend on u only by a common factor, and P has the form

$$P^2 = h^{-1}(u)(1 + k\zeta\bar\zeta/2)^2, \quad K(u) = k/h(u), \quad k = 0, \pm 1 \tag{14}$$

[note that for this choice of P other normal forms of the symmetry generators may apply, e.g., $i(\zeta\partial_\zeta - \bar\zeta\partial_{\bar\zeta})$ instead of $i(\partial_\zeta - \partial_{\bar\zeta})$]. With this result, the field equation (11) yields

$$h^{-1}\partial_u\partial_u(h) - 3\epsilon m h'/h + 2c(k/h - 2c) = 0. \tag{15}$$

Depending on the different possible values of m and c, we thus arrive at the metrics

$$ds^2=2\epsilon(r^2-1)\,\frac{h(u)d\zeta\,d\overline{\zeta}}{(1+k\zeta\overline{\zeta}/2)^2}-2du\,dr-2H(r,u)du^2,$$

$$H=\frac{1}{2}\,rh'/h-\frac{1}{2}\,\epsilon k/h-\epsilon(r^2-1)\left(b-3m\int\frac{dr}{(r^2-1)^2}\right)-\epsilon c, \tag{16}$$

$$\text{with}\quad \begin{aligned} &h(u)=Au+B \quad \text{for } m=0,\ c=0,\\ &h(u)=A+Be^{3\epsilon mu} \quad \text{for } m\neq 0,\ c=0,\\ &h(u)=A\,\sinh(2cu+B)+k/2c \quad \text{for } m=0,\ c=-2b. \end{aligned}$$

In case of the metric (4), $K=K(u)$ leads to

$$K(u)-\epsilon\partial_u\,\ln P=b, \tag{17}$$

which together with (10) again gives (14). The corresponding metrics are

$$ds^2=2\epsilon r\,\frac{h(u)d\zeta\,d\overline{\zeta}}{(1+k\zeta\overline{\zeta}/2)^2}-2du\,dr-2H(\zeta,\overline{\zeta},r,u)du^2,$$

$$H=\tfrac{1}{2}\,rh'/h-\epsilon ra-\epsilon br\,\ln\,\epsilon r-H^0(\zeta,\overline{\zeta},u), \tag{18}$$

$$2\partial_\zeta\partial_{\overline{\zeta}}H^0+\epsilon(1+k\zeta\overline{\zeta}/2)^{-2}\partial_u(hH^0)=0,$$

$$\text{with}\quad \begin{aligned} &h(u)=2\epsilon ku+A \quad \text{for } b=0\\ &h(u)=Ae^{2b\epsilon u}+k/b \quad \text{for } b\neq 0. \end{aligned}$$

The metrics (16) and (18) exhaust the space–times with $K=K(u)$; they admit a G_3 on V_2 [for metric (18): if H^0 is chosen appropriately]: they are spherically, or pseudospherically, or plane symmetric. Contained here is the solution $k=0=b$ found by Wainwright.[1]

B. Space–times with at most two Killing vectors I: The case $2P^2\partial_\zeta\partial_{\overline{\zeta}}K=0$, $K_{,\zeta}\neq 0$

The only explicitly known class of functions $P(\zeta,\overline{\zeta},u)$ which obey the condition $2P^2\partial_\zeta\partial_{\overline{\zeta}}K=0$, $K_{,\zeta}\neq 0$ is given by

$$P=[f(\zeta)_{,\zeta}\overline{f(\zeta)}_{,\overline{\zeta}}]^{-1/2}[f(\zeta,u)+\overline{f(\zeta,u)}]^{3/2},\quad K=-3(f+\overline{f}), \tag{19}$$

where $f(\zeta,u)$ is an arbitrary function; it has been found in the search for Robinson–Trautman *vacuum* solutions of the Petrov type III (see Kramer *et al.*,[5] Chap. 24). Since the field equations obtained by a reduction of the metric (1) using the algebras I.7, I.8, II.7 and II.8 satisfy $2P^2\partial_\zeta\partial_{\overline{\zeta}}K=0$, $K_{,\zeta}\neq 0$, their available (new) solutions contain subcases of (19). They read:

$$ds^2=2\epsilon(r^2-1)\,\frac{h(u)d\zeta\, d\bar{\zeta}}{(\zeta+\bar{\zeta})^3}-2du\; dr-2H\; du^2,$$

$$2H=r\,\frac{h'(u)}{h(u)}-\frac{3\epsilon(\zeta+\bar{\zeta})}{h(u)}+2\epsilon(r^2-1)\left[b-3m\int\frac{dr}{(r^2-1)^2}\right],\qquad (20)$$

$$h=Au+B\quad\text{for}\quad m=0,\quad h=Ae^{3\epsilon mu}+B\quad\text{for}\quad m\neq 0.$$

These solutions admit a Killing vector $i(\partial_\zeta-\partial_{\bar{\zeta}})$ and may thus be called axially symmetric.

C. Space–times with at most two Killing vectors II: The case $2P^2\partial_\zeta\partial_{\bar{\zeta}}K\neq 0$

Of course we tried to solve as many of the reduced field equations as possible in this most general case, but because of the shear amount of equations we did not attack all of them with some obstinacy. We could not find solutions in the cases I.3, I.5, I.6, I.10, I.11, II.5, II.6, II.11, III.2 and III.6 of Eq. (1), and in the cases I.3, I.6, I.11, II.5, II.6 and II.11 of Eq. (4). The following solutions have been obtained (or rediscovered).

Metric (1) with $2P^2\partial_\zeta\partial_{\bar{\zeta}}K\neq 0$:

$$ds^2=\frac{(1-r^2)}{2\epsilon c}\left[\frac{dw^2}{4w}+we^w\; d\varphi^2\right]-2du\; dr-\epsilon c(r^2+2w+3)du^2,$$

Case III.5: $P^{-2}(\zeta+\bar{\zeta})=-e^{w+2}w/2c,\quad w'=\sqrt{2}we^{1+w/2}.$ [Kramer (1984)[2]] (21)

$$ds^2=2\epsilon\,\frac{r^2-1}{P^2(z)}\,d\zeta\; d\bar{\zeta}-2du\; dr-2H\; du^2,$$

Case I.2: $P^2(z)=(e^{2Az}+2Be^{Az}+B^2-A^2)e^{-Az}/4A^2,$

$z=\zeta+\bar{\zeta}+u,\quad c=0=m$ [Drauschke (1997)[4]], (22)

Case II.2: $P^2(z)=e^{3\epsilon mu}(e^{2Az}+2Be^{Az}+B^2-A^2)e^{-Az}/4A^2,$

$$z=\zeta+\bar{\zeta}+2e^{-3\epsilon mu}/3\epsilon m,\quad c=0,\quad m\neq 0.$$

Metric (4) with $2P^2\partial_\zeta\partial_{\bar{\zeta}}K\neq 0$:

$$ds^2=2\epsilon rP^{-2}(z)d\zeta\; d\bar{\zeta}-2du\; dr-2H(\zeta,\bar{\zeta},r,u)du^2,$$

Case I.2: $P^2=\exp(Az)+\epsilon/2,\quad z=\zeta+\bar{\zeta}+u,\quad b=0,$

Case III.3: $P^2=e^{-2\epsilon bu}[\exp(Az)-1]/A;\quad z=\zeta+\bar{\zeta}+\epsilon e^{-2\epsilon bu}/b,\quad b\neq 0,$

$$\text{Case II.11:} \quad P^2 = \tfrac{1}{2}\,[Be^{-4\epsilon bu} - b(\zeta+\bar{\zeta})^2], \quad b \neq 0. \tag{23}$$

Except for the indicated rediscoveries, these solutions are new.

VI. CONCLUSIONS

When we started, we were aware of the few known solutions of the class under consideration, which all had at least one Killing vector. We set a goal of making a thorough search for solutions with symmetries, using the standard techniques of the symmetry analysis of differential equations, and hoped to find one or the other new solution. What we finally found was a surprise in some aspects (not all of them being positive).

First of all, it is known that the only Lie-point-symmetries admitted by Einstein's field equations correspond to coordinate transformations and scalings (leading to Killing or homothetic vectors in the similarity reduction). A restriction to a special class of solutions (here: algebraically special, etc.) may enlarge the group of admitted symmetries. But the experience from other special classes, e.g., from the algebraically special vacuum solutions, or axially symmetric stationary perfect fluids, showed that this rarely happens. So it was rather unexpected that in our case additional symmetries also showed up. This result should encourage people to always try a symmetry analysis when asking for solutions! As a side result we found that these metrics admit generalized Kerr–Schild transformations.

Connected with these additional symmetries was the huge number of different ordinary differential equations we obtained by the symmetry reduction (23 partial, or 45 ordinary differential equations, the latter ones being given in the Appendix). This made it practically impossible to make a detailed analysis of each of them. So we concentrated on a few simple classes.

The first class are the solutions where the curvature K of the ζ-$\bar{\zeta}$ space is a (possibly time-dependent) constant. The field equations then imply that the metrics admit a G_3 on V_2, i.e., they are of spherical, pseudospherical, or plane symmetry. The general solution for this case is given in (16) and (18). Many solutions with this symmetry are known from other approaches, but they were usually found and given in comoving (and non-null) coordinates. Since an inspection shows that our solutions have nonzero shear, they do not belong to the best-known class of solutions with vanishing shear and are probably new. As in most spherically (pseudospherically, plane) symmetric perfect fluids, a surface of vanishing pressure will exist, but the four-velocity will not be tangent to this surface so that it cannot serve as the outer boundary of a finite perfect fluid sphere. But since for large r pressure and mass density approach zero (if $b=0$), these solutions may be interpreted as approximate models for outer regions of an extended source. On the other hand, the interior regions ($r^2<1$) may serve as inhomogeneous cosmological models.

The second class (20) has been constructed by using the only known solution of the corresponding vacuum Robinson–Trautman class. The rest of the solutions, given in (21)–(23), could be found because of the simplicity of the reduced field equations for these cases. Except for the subcases due to Refs. 3 and 4, they are new.

Unfortunately, none of these last new solutions admits a convincing physical interpretation. Technically this is due to the fact that we did not start from any physical assumption as, e.g., an equation of state or special properties of the four velocity, but made a mathematical study of the field equations and tried to fill in the physics at a later stage. More specifically, we were not able to find examples where, e.g., there is a regular surface with vanishing pressure p which could serve as a model of an isolated source. So the hope of finding some interior solutions for the Robinson–Trautman vacuum solutions was not fulfilled.

APPENDIX: LISTS OF REDUCED FIELD EQUATIONS

We here give the lists of the ordinary differential equations which result from the field equations (1) and (4) after a twofold similarity reduction. In each of the five lists, we give the original field equation and then label the resulting equations using the tables of nonequivalent two-dimensional subalgebras.

1. Field equation (1), $c=0=m$: $\partial_\zeta\partial_{\bar\zeta}K+\partial_u\partial_u(P^{-2})=0$, $K\equiv 2P^2\partial_\zeta\partial_{\bar\zeta}\ln P$

Case	P	K	Field equation
I.1	$P(u)$	0	$(P^{-2})''=0$
I.2	$P(\zeta+\bar\zeta+u)$	$2P^2(\ln P)''$	$K''+(P^{-2})''=0$
I.3	$f(z)/\sqrt{u}$, $z\equiv ue^{-(\zeta+\bar\zeta)/2}$	$zf^2[z(\ln f)']'/2u$	$2u(zK')'+16(f^{-2})'+8z(f^{-2})''=0$
I.4	$(\zeta+\bar\zeta)f(u)$	$-2f^2$	$(f^{-2})''=0$
I.5	$u^{(2-A)/2A}f(z)$, $z\equiv(\zeta+\bar\zeta)u^{-1/A}$	$2f^2(\ln f)''/u$	$A^2uK''+z^2(f^{-2})''+(5-A)z(f^{-2})'+2(2-A)f^{-2}=0$
I.6	$e^uf(z)$, $z\equiv(\zeta+\bar\zeta)e^{-u}$	$2f^2(\ln f)''$	$K''z^2(f^{-2})''+5z(f^{-2})'+4f^{-2}=0$
I.7	$P(\zeta+\bar\zeta)$	$2P^2(\ln P)''$	$K''=0$
I.8	$f(\zeta+\bar\zeta)/\sqrt{u}$	$2f^2(\ln f)''/u$	$K''=0$
I.10	$e^{i(\zeta-\bar\zeta)/4}f(\zeta+\bar\zeta)$	$2e^{i(\zeta-\bar\zeta)/2}w$, $w\equiv f^2(\ln f)''$	$4w''+w=0$
I.11	$[(\zeta+\bar\zeta)f(z)]^{1/2}$, $z\equiv\frac{u+i(\zeta-\bar\zeta)/2}{\zeta+\zeta}$	$(\zeta+\bar\zeta)^{-1}w$, $w\equiv f[2z(\ln f)'-1+(z^2+\frac{1}{4})(\ln f)'']$	$[2+4z\partial_z+(z^2+\frac{1}{4})\partial_z\partial_z]w+(f^{-1})''=0$

2. Field equation (1), $c=0$, $m\neq 0$: $\alpha=3\epsilon m$: $P^2\partial_\zeta\partial_{\bar\zeta}K+P^2\partial_u\partial_u(P^{-2})+6\epsilon m\partial_u\ln P=0$

Case	P	K	Field equation
II.1	$P(u)$	0	$P^2(P^{-2})''+2\alpha(\ln P)'=0$
II.2	$P(\zeta+\bar\zeta+u)$	$2P^2(\ln P)''$	$P^2[K''+(P^{-2})'']+2\alpha(\ln P)'=0$
II.3	$e^{-(\alpha/2)u}f\left(\zeta+\bar\zeta+\frac{2}{\alpha}e^{-\alpha u}\right)$	$2e^{-\alpha u}f^2(\ln f)''$	$f^2(\ln f)''+2(f^{-2})''=0$
II.4	$(\zeta+\bar\zeta)f(u)$	$-2f^2$	$f^2(f^{-2})''+2\alpha(\ln f)'=0$
II.5	$\exp\left(-\frac{\alpha}{2}u-\frac{1}{\alpha}e^{-\alpha u}\right)f(z)$, $z\equiv(\zeta+\bar\zeta)\exp(e^{-\alpha u}/\alpha)$	$2e^{-\alpha u}f^2(\ln f)''$	$e^{\alpha u}f^2K''+z^2f^2(f^{-2})''-10z(\ln f)'+4=0$
II.6	$e^{u/A}f(z)$, $z\equiv(\zeta+\bar\zeta)e^{-u/A}$	$2f^2(\ln f)''$	$K''+[(z\partial_z)^2+(4+\alpha A)z\partial_z+4+2\alpha A]f^{-2}=0$
II.7	$P(\zeta+\bar\zeta)$	$2P^2(\ln P)''$	$K''=0$
II.8	$e^{-\alpha u/2}f(\zeta+\bar\zeta)$	$2e^{-\alpha u}f^2(\ln f)''$	$K''=0$

II.10	$\exp\{-(\alpha/2)[u+(i/2)(\zeta-\bar{\zeta})]\}f(\zeta+\bar{\zeta})$	$2\exp(-\alpha\{[u+\frac{1}{2}(\zeta-\bar{\zeta})]\})w$, $w\equiv f^2(\ln f)''$	$w''+\alpha^2 w/2=0$
II.11	$[(\zeta+\bar{\zeta})e^{-\alpha u}f(\zeta)]^{1/2}$, $z\equiv\dfrac{i(\zeta-\bar{\zeta})-2e^{-\alpha u}/\alpha}{\zeta+\bar{\zeta}}$	$(\zeta+\bar{\zeta})^{-1}e^{-\alpha u}w$, $w\equiv f[2z(\ln f)'+(z^2+1)(\ln f)''-1]$	$[(z^2+1)w]''+4(f^{-1})''=0$

3. Field equation (1), $c\neq 0$, $m=0$: $P^2\partial_\zeta\partial_{\bar{\zeta}}K+P^2\partial_u\partial_u(P^{-2})+2c(K-2c)=0$

Case	P	K	Field equation
III.1	$P(u)$	0	$P^2(P^{-2})''=0$
III.2	$P(\zeta+\bar{\zeta}+u)$	$2P^2(\ln P)''$	$P^2[K''+(P^{-2})'']+2c(K-2c)=0$
III.3	$(\zeta+\bar{\zeta})f(u)$	$-2f^2$	$f^2(f^{-2})''-4c(f^2+c)=0$
III.4	$e^{u/A}f(z)$, $z\equiv(\zeta+\bar{\zeta})e^{-1/A}$	$2f^2(\ln f)''$	$A^2f^2K''+z^2f^2(f^{-2})''+5zf^2(f^{-2})'+4+2A^2c(K-2c)=0$
III.5	$P(\zeta+\bar{\zeta})$	$2P^2(\ln P)''$	$P^2K''+4c(K-2c)=0$

4. Field equation (4), $b=0$: $2P^2\partial_\zeta\partial_{\bar{\zeta}}\ln P=\epsilon\partial_u\ln P$

Case	P	K	Field equation
I.1	$P(u)$	0	$(\ln P)'=0$
I.2	$P(\zeta+\bar{\zeta}+u)$	$2P^2(\ln P)''$	$K-\epsilon(\ln P)'=0$
I.3	$f(z)/\sqrt{u}$, $z\equiv ue^{-(\zeta+\bar{\zeta})/2}$	$zf^2[z(\ln f)']'/2u$	$2uK-2\epsilon z(\ln f)'+\epsilon=0$
I.4	$(\zeta+\bar{\zeta})f(u)$	$-2f^2$	$K=\epsilon(\ln f)'$
I.5	$u^{(2-A)/2A}f(z)$, $z\equiv(\zeta+\bar{\zeta})u^{-1/A}$	$2f^2(\ln f)''/u$	$4Af^2(\ln f)''+2\epsilon z(\ln f)' = (2-A)\epsilon$
II.6	$e^u f(z), z\equiv(\zeta+\bar{\zeta})e^{-u}$	$2f^2(\ln f)''$	$K+\epsilon z(\ln f)'=\epsilon$
I.7	$P(\zeta+\bar{\zeta})$	$2P^2(\ln P)''$	$K=0$
I.8	$f(\zeta+\bar{\zeta})/\sqrt{u}$	$2f^2(\ln f)''/u$	$2uK+\epsilon=0$
I.10	$e^{i(\zeta-\bar{\zeta})/4}f(\zeta+\bar{\zeta})$	$2e^{i(\zeta-\bar{\zeta})/2}f^2(\ln f)''$	$K=0$
I.11	$[(\zeta+\bar{\zeta})f(z)]^{1/2}$, $z\equiv\dfrac{u+i(\zeta-\bar{\zeta})/2}{\zeta+\bar{\zeta}}$	$w/(\zeta+\bar{\zeta})$, $w\equiv f[2z(\ln f)' + (z^2+\frac{1}{4})(\ln f)''-1)]$	$2K(\zeta+\bar{\zeta})=\epsilon(\ln f)'$

5. Field equation (4), $b\neq 0$, $\alpha=2\epsilon b$: $2P^2\partial_\zeta\partial_{\bar\zeta}\ln P-\epsilon\partial_u \ln P=b$

Case	P	K	Field equation
II.1	$P(u)$	0	$(\ln P)'=-\epsilon b$
II.2	$P(\zeta+\bar\zeta+u)$	$2P^2(\ln P)''$	$K-\epsilon(\ln P)'=b$
II.3	$e^{-(\alpha/2)u}f(\zeta+\bar\zeta+(2/\alpha)\,e^{-\alpha u})$	$2e^{-\alpha u}f^2(\ln f)''$	$f^2(\ln f)''+\epsilon(\ln f)'=0$
II.4	$(\zeta+\bar\zeta)f(u)$	$-2f^2$	$K-\epsilon(\ln f)'=b$
II.5	$\exp(-(\alpha/2)\,u-(1/\alpha)\,e^{-\alpha u})f(z)$, $z\equiv(\zeta+\bar\zeta)\exp(1/\alpha\, e^{-\alpha u})$	$2e^{-\alpha u}f^2(\ln f)''$	$2f^2(\ln f)''+\epsilon z(\ln f)'=\epsilon$
II.6	$e^{u/A}f(z)$, $z\equiv(\zeta+\bar\zeta)e^{-u/A}$	$2f^2(\ln f)''$	$AK+\epsilon z(\ln f)'=Ab+\epsilon$
II.7	$P(\zeta+\bar\zeta)$	$2P^2(\ln P)''$	$K=b$
II.8	$e^{-\alpha u/2}f(\zeta+\bar\zeta)$	$2e^{-\alpha u}f^2(\ln f)''$	$K=0$
II.10	$\exp\{-(\alpha/2)\,[u+(i/2)\,(\zeta-\bar\zeta)]\}f(\zeta+\bar\zeta)$	$2e\{-\alpha[u+\frac{1}{2}(\zeta-\bar\zeta)]\}w$, $w\equiv f^2(\ln f)'$	$K=0$
II.11	$[(\zeta+\bar\zeta)e^{-\alpha u}f(z)]^{1/2}$, $z\equiv\dfrac{i(\zeta-\bar\zeta)-2e^{-\alpha u}/\alpha}{\zeta+\bar\zeta}$	$(\zeta+\bar\zeta)^{-1}e^{-\alpha u}w$, $w\equiv f[2z(\ln f)'+(z^2+1)\times(\ln f)''-1]$	$f[2z(\ln f)'-\epsilon(\ln f)'+(z^2+1)(\ln f)''-1]=0$

[1] J. Wainwright, ''Algebraically special fluid space–times with hypersurface-orthogonal shearfree rays,'' Int. J. Theor. Phys. **10**, 39 (1974).
[2] D. Kramer, ''A new solution for a rotating perfect fluid in general relativity,'' Class. Quantum Grav. **1**, L3 (1984).
[3] A. Drauschke, Diplomarbeit Jena, 1996.
[4] A. Rainer, Diplomarbeit Graz, 1997.
[5] D. Kramer, H. Stephani, M. A. H. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).

Einstein's equations with asymptotically stable constraint propagation

Othmar Brodbeck
Center for Gravitational Physics and Geometry, Department of Physics, The Pennsylvania State University, State College, Pennsylvania 16802 and Institute for Theoretical Physics, The University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Simonetta Frittelli[a)]
Physics Department, Duquesne University, Pittsburgh, Pennsylvania 15282

Peter Hübner
Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, Schlaatzweg 1, 14473 Potsdam, Germany

Oscar A. Reula
Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, Schlaatzweg 1, 14473 Potsdam, Germany and FaMAF, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

(Received 2 October 1998; accepted for publication 21 October 1998)

We introduce a proposal to modify Einstein's equations by embedding them in a larger symmetric hyperbolic system. The additional dynamical variables of the modified system are essentially first integrals of the original constraints. The extended system of equations reproduces the usual dynamics on the constraint surface of general relativity, and therefore naturally includes the solutions to Einstein gravity. The main feature of this extended system is that, at least for a linearized version of it, the constraint surface is an attractor of the time evolution. This feature suggests that this system may be a useful alternative to Einstein's equations when obtaining numerical solutions to full, nonlinear gravity. © *1999 American Institute of Physics.* [S0022-2488(99)03002-9]

I. INTRODUCTION

Over the past decade, computer power has increased to the point that simulations of two- and even three-dimensional general relativity are now feasible. These simulations, which assume little or no symmetry of their generic field configurations, at first seemed to represent straightforward generalizations of simpler one-dimensional calculations. However, attempts to perform the higher dimensional simulations have revealed a variety of unexpected features which limit accurate simulations to a rather short time interval. One such feature, which is believed to be a major source of numerical error, is that numerical time evolution generates a rapidly growing violation of the constraint equations. In this paper, we propose a system of dynamical equations wherein the evolution naturally remains close to the constraint surface. Although the most obvious application of this approach is to numerical simulations, it may prove useful in other branches of general relativity as well.

As is well known analytically, the time evolution predicted by the exact Einstein equations is such that the constraint equations are satisfied on each time slice when they are satisfied by the initial data. Geometrically, the evolution vector field is tangential to the constraint submanifold, implying that solutions to the complete set of equations are insensitive to properties of the evolution field in the vicinity of the constraint surface. In discrete approximations, on the other hand,

[a)]Electronic mail: simo@mayu.physics.duq.edu

the notion of tangency is approximate, as is that of the constraint surface itself. As a consequence, the numerical evolution becomes sensitive to possible instabilities of the constraint submanifold and numerical solutions are, in general, carried away from it exponentially fast with time. Even in case one were able to construct a code whose discretized vector field were exactly tangent to a discretized version of the constraint submanifold, the same problem would be likely to arise, as numerical errors on the initial data would prevent a start of the time integration exactly on the constraint submanifold.

As demonstrated in Ref. 1, evolution schemes can be constructed in such a way that the violation of the constraints has the same convergence order as the scheme itself. This property, which is in the meantime a standard requirement for evolution schemes, implies that the choice of an appropriately fine grid is sufficient to satisfy the constraint equations at any given time with arbitrary accuracy. However, since the violation of the constraints grows very quickly with time, the utility of grid refinements to reduce constraint violations is very limited, especially in two- and three-dimensional calculations.

In the so-called constrained evolution schemes one attempts to solve this problem by isolating two sets of variables in Einstein's equations. One uses evolution equations to evolve one set and determines the variables of the other set by solving constraint equations on each time slice. This method has frequently been used in one-dimensional simulations where, on the one hand, it is easy to split the variables into dynamical and longitudinal ones, and where, on the other hand, the constraint equations are ordinary differential equations along a spacelike direction. However, in two- and three-dimensional simulations with spacelike hypersurfaces as time surfaces, the elliptic character of the constraint equations makes it expensive in computer time to solve the constraint equations on each time slice (not to mention the problems arising in the treatment of the grid boundaries). Furthermore, this approach does not guarantee that the complete set of Einstein's equations is solved. Since only a subset of the variables is determined by evolution equations, some of these equations remain unused. The problem is, therefore, shifted to the preservation of the unused evolution equations, which, as shown in Ref. 2, is a problem of similar nature.

In order to guarantee a good approximation to the complete set of field equations, it is, therefore, necessary to analyze the behavior of the evolution vector field in a whole neighborhood of the constraint submanifold. Away from the constraint submanifold the evolution field is not uniquely determined as field configurations violating the constraint equations are physically not relevant. Hence, the evolution vector field can be modified in an arbitrary way, as long as its values on the constraint submanifold remain unchanged, and as long as the modified field continues to be strongly hyperbolic, so that the Cauchy problem is well posed in a whole neighborhood of the constraint surface.

Of particular interest are modified equations for which the constraint submanifold is asymptotically stable, because for equations with this feature, sufficiently accurate codes are expected to generate solutions which remain close to the constraint surface, and which, therefore, would represent improved approximations to Einstein's equations.

Modifications of the evolution vector field have previously been studied. However, in general, these preserve the time reversal symmetry of Einstein's equations, which implies that modifications of this type cannot have the desired properties. Time reversal symmetry implies that if the evolution field is such that a solution to some initial data in a neighborhood of the constraint submanifold approaches the constraint submanifold during time evolution, then the solution to the time reversed initial data will asymptotically be carried away from this submanifold. Thus, without a modification of Einstein's equations which breaks the time reversal symmetry, the best one can expect to achieve is a set of equations for which the constraint submanifold is stable, but not asymptotically stable. However, stability of the undiscretized equations is not sufficient for numerical simulations, since spurious solutions to the discretized equations can grow very rapidly even for stable systems. To eliminate the impact of such solutions, it is, therefore, necessary that the constraint manifold is an attractor for the time evolution.

The above-mentioned modifications of Einstein's equations are, in general, obtained by including dynamical quantities which are proportional to the constraint expressions. An alternative argument showing that extensions of this type cannot lead to an asymptotically stable constraint surface is the following: Since the constraint equations are of the same order as the evolution equations, their inclusion affects mainly the principal part of the evolution equations, whence the freedom remaining after requiring that these terms do not destroy strong hyperbolicity is very limited. Thus, such extensions ensure that the problem is the well posed, but not the asymptotic stability. This can only be obtained either via modifications of the lower order terms or the addition of higher (second) order terms, that is, by including damping or diffusion terms.

In Sec. II, we propose a modification of Einstein's equations which includes new dynamical terms proportional to certain first integrals of the constraint expressions, rather than to the constraints themselves. The dissipation, that is, the time asymmetry, is not of the diffusive type, and is built into the definition of these integrals. (One could also introduce diffusive dissipation, but this would significantly reduce the allowed maximal time step in explicit discretization schemes.)

We show that the Cauchy problem for the resulting new system, which we call the λ-system, is locally well posed. (The name is a remnant of the way the system was originally guessed by Brodbeck and Hübner, namely by a formal application of Lagrangian multiplier techniques.) We also prove that if the constraints are initially satisfied, and if their first integrals initially vanish, then the λ-system provides solutions to Einstein's equations. Moreover, for initial data sets for the λ-system, which are sufficiently close to the constraint submanifold and sufficiently close to zero, respectively, we suspect that the solutions asymptotically tend to solutions to Einstein's equations.

In Sec. III, we give support to our expectation by proving that the linearized extended system is asymptotically stable, thus showing that in the linearized case, the constraint submanifold is indeed an attractor for the λ-system. In Sec. IV, we discuss further expectations in connection with our proposal.

II. THE λ-SYSTEM

In this section we spell out our proposal for a modification of Einstein's equations with an asymptotically stable constraint submanifold. For definiteness, we choose the symmetric hyperbolic system introduced by Frittelli and Reula in Ref. 3, which corresponds to the parameters $\alpha=\beta=-1$ in Ref. 4. Although the full equations (with the nonprincipal part terms added) are given in Ref. 5, we repeat them for completeness.

In the version of Einstein's vacuum equations chosen, the system is given by the following set of dynamical equations (where Latin indices run from 1 to 3):

$$\dot{h}^{ij}=N^n h^{ij}{}_{,n}+Q\sqrt{h}(2P^{ij}-Ph^{ij})-2h^{n(i}N^{j)}{}_{,n}, \tag{1}$$

$$\begin{aligned}\dot{M}^{ij}{}_k=&N^n M^{ij}{}_{k,n}+Q\sqrt{h}(P^{ij}{}_{,k}-2\delta_k{}^{(i}P^{j)n}{}_{,n})+Q\sqrt{h}(\tfrac{3}{2}P^{ij}M_k-PM^{ij}{}_k+Q^{-1}P^{ij}Q_{,k}\\&-2\delta_k{}^{(i}[h^{j)q}h_{mr}h_{ns}P^{mn}M^{rs}{}_q-2M^{j)p}{}_n P^{mn}h_{prn}+\tfrac{3}{2}P^{j)n}M_n-\tfrac{1}{2}h^{j)n}PM_n])\\&+h^{ij}N^n{}_{,nk}-h^{n(i}N^{j)}{}_{,nk}-2N^{(i}{}_{,n}M^{j)n}{}_k+N^m{}_{,k}M^{ij}{}_m,\end{aligned} \tag{2}$$

$$\begin{aligned}\dot{P}^{ij}=&N^n P^{ij}{}_{,n}+Q\sqrt{h}(h^{mn}M^{ij}{}_{m,n}-2h^{n(i}M^{j)k}{}_{k,n})+Q\sqrt{h}(4h_{np}h^{m(i}M^{j)n}{}_k M^{kp}{}_m\\&-h^{ik}h^{jn}h_{rp}h_{sq}M^{rs}{}_k M^{pq}{}_n+\tfrac{1}{2}h^{ik}h^{jn}M_k M_n+2M^{nk}{}_k M^{ij}{}_n-2M^{ik}{}_n M^{jn}{}_k-2h_{mn}h^{kp}M^{im}{}_k M^{jn}{}_p\\&-2h^{n(i}M^{j)k}{}_k M_n+M^{ij}{}_n h^{nk}M_k-Q^{-1}[h^{ik}h^{jn}Q_{,kn}+2M^{k(i}{}_n h^{j)n}Q_{,k}-M^{ij}{}_m h^{mk}Q_{,k}\\&-h^{ij}(h^{kn}Q_{,kn}+2M^{km}{}_m Q_{,k})]+2P^{ik}h_{kn}P^{nj}-\tfrac{3}{2}PP^{ij}+h^{ij}(\tfrac{1}{2}P^2-h_{mr}h_{ns}P^{mn}P^{rs}))\\&-2P^{k(i}N^{j)}{}_{,k}.\end{aligned} \tag{3}$$

Here, h^{ij} is the inverse intrinsic metric of the spacelike hypersurfaces Σ_t, $P^{ij}:=k^{ij}-h^{ij}k$ denotes a linear combination of the extrinsic curvature k^{ij} of the slice and its trace k, and $M^{ij}{}_k:=\tfrac{1}{2}(h^{ij}{}_{,k}$

$-h^{ij}h_{rs}h^{rs}{}_{,k}$) represents a linear combination of spatial derivatives of the inverse intrinsic metric. The functions P and M_k are abbreviations for $h_{ij}P^{ij}$ and $h_{ij}M^{ij}{}_k$, respectively, and Q and N^i are arbitrarily given functions fixing the gauge degrees of freedom.

This evolution system is symmetric hyperbolic with respect to the inner product

$$\langle h_1^{ij},P_1^{ij},M_1^{ij}{}_k|H_e|h_2^{ij},P_2^{ij},M_2^{ij}{}_k\rangle := \int_{\Sigma_t}\{e_{im}e_{jn}\overline{h_1^{ij}}h_2^{mn}+e_{im}e_{jn}\overline{P_1^{ij}}P_2^{mn}+e_{im}e_{jn}e^{kl}\overline{M_1^{ij}{}_k}M_2^{mn}{}_l\}d\Sigma, \tag{4}$$

where e_{ij} denotes an Euclidean flat metric on the hypersurface Σ_t. It is completed by the following set of constraints equations:

$$\mathcal{C}=0, \quad \mathcal{C}^i=0, \quad \mathcal{C}^{ij}{}_k=0, \tag{5}$$

where

$$\mathcal{C}\equiv -M^{kn}{}_{n,k}+h_{pq}M^{pk}{}_nM^{qn}{}_k-M^{kq}{}_qM_k+\tfrac{1}{4}h^{kn}M_kM_n-\tfrac{1}{2}h_{mn}h_{rs}h^{pq}M^{mr}{}_pM^{ns}{}_q$$
$$-\tfrac{1}{2}h_{mn}h_{rs}P^{mr}P^{ns}+\tfrac{1}{4}P^2, \tag{6}$$

$$\mathcal{C}^i\equiv P^{ik}{}_{,k}-2h_{mn}M^{im}{}_kP^{nk}-\tfrac{1}{2}h^{ik}PM_k+h_{mn}h_{pq}h^{ik}M^{mp}{}_kP^{nq}+\tfrac{3}{2}P^{ik}M_k, \tag{7}$$

$$\mathcal{C}^{ij}{}_k\equiv 2M^{ij}{}_k-h^{ij}h_{pq}M^{pq}{}_k-h^{ij}{}_{,k}. \tag{8}$$

The first two constraints are the scalar and the vector constraint of Einstein's equations, that is the time–time and time–space components of the Einstein tensor for a given 3+1 decomposition of space–time. The third is the statement that the tensor $M^{ij}{}_k$ is a linear combination of spatial derivatives of the three-metric.

To solve the initial value problem of general relativity in this approach, one prescribes an initial data set $(h_0^{ij},P_0^{ij},M_0^{ij}{}_k)$ at $t=0$ which satisfies the constraints' equations and subsequently solves the above evolution equations. Symmetric hyperbolicity of the evolution system implies that a unique local solution does exist.

By taking a time derivative of Eqs. (6)–(8) and using (1)–(3) to eliminate time derivatives in favor of spatial derivatives, the following evolution equations for the constraints are obtained:

$$\dot{\mathcal{C}}=N^n\mathcal{C}_{,n}+3Q\sqrt{h}\mathcal{C}^k{}_{,k}+\cdots, \tag{9}$$

$$\dot{\mathcal{C}}^i=N^n\mathcal{C}^i{}_{,n}+Q\sqrt{h}(h^{ik}\mathcal{C}_{,k}+h^{rs}\mathcal{C}^{ik}{}_{[r,k]s}+h^{is}h^{kl}h_{mn}\mathcal{C}^{mn}{}_{[s,k]l})+\cdots, \tag{10}$$

$$\dot{\mathcal{C}}^{ij}{}_k=N^n\mathcal{C}^{ij}{}_{k,n}-2Q\sqrt{h}(2\delta_k{}^{(i}\mathcal{C}^{j)}-h^{ij}h_{kl}\mathcal{C}^l)+\cdots, \tag{11}$$

where the ellipses ''$\cdots$'' represent undifferentiated terms which are linear in the constraint quantities and at least linear in the variables P^{ij} and $M^{ij}{}_k$.

Since Eq. (10) is of second order in spatial derivatives, we introduce a further constraint by[4]

$$\mathcal{C}^{ij}{}_{kl}:=2M^{ij}{}_{[k,l]}+2M^{ij}{}_{[k}M_{l]}. \tag{12}$$

(One could also consider the constraint $\mathcal{C}^i{}_j\equiv\mathcal{C}^{ik}{}_{jk}$, which still makes the constraint system symmetric hyperbolic and produces a smaller number of extra fields.) By taking a time derivative of (12), we obtain

$$\dot{\mathcal{C}}^{ij}{}_{kl}=N^n\mathcal{C}^{ij}{}_{kl,n}-2Q\sqrt{h}(\delta_k{}^{(i}\mathcal{C}^{j)}{}_{,l}-\delta_l{}^{(i}\mathcal{C}^{j)}{}_{,k})+\cdots, \tag{13}$$

and by plugging (12) into (10), we see that the evolution equation for $\mathcal{C}^i$ can be rewritten as

$$\dot{\mathcal{C}}^i = N^n \mathcal{C}^i{}_{,n} + Q\sqrt{h}(h^{ik}\mathcal{C}_{,k} + h^{rs}\mathcal{C}^{ik}{}_{rk,s}) + \cdots . \tag{14}$$

The constraint quantities $\mathcal{C}$, $\mathcal{C}^i$, $\mathcal{C}^{ij}{}_k$, and $\mathcal{C}^{ij}{}_{kl}$ thus propagate according to the first-order system of equations consisting of (9), (11), and (13)–(14), which is symmetric hyperbolic with respect to the following inner product:

$$\langle C_1, C_1^i, C_1^{ij}{}_k, C_1^{ij}{}_{kl} | H_C | C_2, C_2^i, C_2^{ij}{}_k, C_2^{ij}{}_{kl} \rangle := \int_{\Sigma_t} \left\{ \frac{1}{3}\overline{C}_1 C_2 + e_{ij}\overline{C}_1^i C_2^j + e_{ij}e_{kl}e^{rs}\overline{C}_1^{ik}{}_r C_2^{jl}{}_s + \frac{1}{4} e_{im}e_{jn}e^{kp}e^{lq}\overline{C}_1^{ij}{}_{kl} C_2^{mn}{}_{pq} \right\} d\Sigma. \tag{15}$$

The uniqueness of the solutions to this system implies that if the constraints are initially satisfied, then the exact evolution equations preserve them. When, as in numerical simulations, the constraint variables initially are not precisely zero, then the corresponding solution is, in general, carried away from the constraint surface during time evolution. However, since the evolution equations for the constraint variables are symmetric hyperbolic, the violation of the constraints becomes smaller when the constraints initially are satisfied with better accuracy.

In order to obtain a system with an asymptotically stable constraint submanifold, we propose a modification of Einstein's equations, which is inspired by the behavior of dissipative systems, where a transient eventually is dissipated away as the system settles down. We extend the set of dynamical variables by considering the following ''time integrals'' of the constraint variables:

$$\dot{\lambda} = \alpha_0 \mathcal{C} - \beta_0 \lambda, \tag{16}$$

$$\dot{\lambda}^i = \alpha_1 \mathcal{C}^i - \beta_1 \lambda^i, \tag{17}$$

$$\dot{\lambda}^{ij}{}_k = \alpha_3 \mathcal{C}^{ij}{}_k - \beta_3 \lambda^{ij}{}_k, \tag{18}$$

$$\dot{\lambda}^{ij}{}_{kl} = \alpha_4 \mathcal{C}^{ij}{}_{kl} - \beta_4 \lambda^{ij}{}_{kl}, \tag{19}$$

where the tensor-valued λ-variables are assumed to have the same symmetries as the corresponding C-variables, and where $\alpha_i \neq 0$ and $\beta_i > 0$ are constants.

Equations (16)–(19) represent evolution equations for the λ-variables which in terms of the fundamental variables $(h^{ij}, P^{ij}, M^{ij}{}_k)$ are given by

$$\dot{\lambda} = \alpha_0(-M^{kn}{}_{n,k} + h_{pq}M^{pk}{}_n M^{qn}{}_k - M^{kq}{}_q M_k + \tfrac{1}{4}h^{kn}M_k M_n) - \beta_0 \lambda, \tag{20}$$

$$\dot{\lambda}^i = \alpha_1(P^{ik}{}_{,k} - 2h_{mn}M^{im}{}_k P^{nk} - \tfrac{1}{2}h^{ik}PM_k) - \beta_1 \lambda^i, \tag{21}$$

$$\dot{\lambda}^{ij}{}_k = \alpha_3(2M^{ij}{}_k - h^{ij}h_{rs}M^{rs}{}_k - h^{ij}{}_{,k}) - \beta_3 \lambda^{ij}{}_k, \tag{22}$$

$$\dot{\lambda}^{ij}{}_{kl} = \alpha_4(2M^{ij}{}_{[k,l]} + 2M^{ij}{}_{[k}M_{l]}) - \beta_4 \lambda^{ij}{}_{kl}. \tag{23}$$

In the present form, the combined system (1)–(3), (20)–(23) is not symmetric hyperbolic, since Eqs. (20)–(23) involve spatial derivatives of the variables $(h^{ij}, P^{ij}, M^{ij}{}_k)$, whereas Eqs. (1)–(3) do not contain λ-variables at all. However, by adding terms containing first derivatives of the λ-variables it is possible to bring the system (1)–(3), (20)–(23) into a symmetric hyperbolic form,

$$\dot{h}^{ij} = \alpha_3 h^{mn}\lambda^{ij}{}_{m,n} + N^n h^{ij}{}_{,n} + \text{source terms}, \tag{24}$$

$$\dot{M}^{ij}{}_k = 2\alpha_4 h^{lm}\lambda^{ij}{}_{kl,m} - \alpha_0 \delta_k{}^{(i}h^{j)l}\lambda_{,l} + N^n M^{ij}{}_{k,n} + Q\sqrt{h}(P^{ij}{}_{,k} - 2\delta_k{}^{(i}P^{j)n}{}_{,n}) + \text{source terms}, \tag{25}$$

$$\dot{P}^{ij}=\alpha_2 h^{l(i}\lambda^{j)}{}_{,l}+N^n P^{ij}{}_{,n}+Q\sqrt{h}(h^{mn}M^{ij}{}_{m,n}-2h^{n(i}M^{j)k}{}_{k,n})+\text{source terms}. \tag{26}$$

By construction, the ''λ-system'' (20)–(26) is symmetric hyperbolic with respect to the inner product

$$\langle h_1^{ij},P_1^{ij},M_{1\,k}^{ij},\lambda_1,\lambda_1^i,\lambda_{1\,k}^{ij},\lambda_{1\,kl}^{ij}|H_E^\lambda|h_2^{ij},P_2^{ij},M_{2\,k}^{ij},\lambda_2,\lambda_2^i,\lambda_{2\,k}^{ij},\lambda_{2\,kl}^{ij}\rangle$$

$$:=\int_{\Sigma_t}\{e_{im}e_{jn}\overline{h_1^{ij}}h_2^{mn}+e_{im}e_{jn}\overline{P_1^{ij}}P_2^{mn}+e_{im}e_{jn}e^{kl}\overline{M_{1\,k}^{ij}}M_{2\,\,1}^{mn}+\overline{\lambda_1}\lambda_2+e_{ij}\overline{\lambda_1^i}\lambda_2^j$$

$$+e_{ip}e_{jq}e^{kr}\overline{\lambda_{1\,k}^{ij}}\lambda_{2\,r}^{pq}+e_{ip}e_{jq}e^{kr}e^{ls}\overline{\lambda_{1\,kl}^{ij}}\lambda_{2\,rs}^{pq}\}d\Sigma. \tag{27}$$

The initial data for this purely dynamical set of equations consists of arbitrary functions

$$(h_0^{ij},P_0^{ij},M_{0\,k}^{ij},\lambda_0,\lambda_0^i,\lambda_{0\,k}^{ij},\lambda_{0\,kl}^{ij}). \tag{28}$$

However, the dynamical degrees of freedom are extended by 40 λ-variables.

Clearly, for an arbitrary solution to Einstein's equations, $(h_E^{ij},P_E^{ij},M_{E\,k}^{ij})$, the embedded field configuration $(h^{ij},P^{ij},M^{ij}{}_k,\lambda,\lambda^i,\lambda^{ij}{}_k,\lambda^{ij}{}_{kl}):=(h_E^{ij},P_E^{ij},M_{E\,k}^{ij},0,0,0,0)$ is a solution to the λ-system. Conversely, every solution to the λ-system with vanishing λ-variables is also a solution to Einstein's equations. Due to this property, and since the solutions to the λ-system are unique, the λ-system naturally reproduces the dynamics on the constraint submanifold of general relativity.

Note that if the constraints are initially not satisfied, then, even when the λ-variables initially vanish, the λ-variables would pick up a nonzero value during time evolution. Hence, solutions to the λ-system corresponding to such initial data sets would not represent solutions to the complete set of Einstein's equations. In fact, they would not even solve the evolution equations of general relativity. However, for constraint- and λ-variables which initially are sufficiently close to zero, we suspect that the solutions asymptotically approach solutions to the Einstein equations. In Sec. III, we give analytical evidence that this conjecture could be true.

The system is by no means uniquely ''extended,'' since one could still add nonprincipal (undifferentiated) terms, as long as they vanish when $\lambda=\lambda^i=\lambda^{ij}{}_k=\lambda^{ij}{}_{kl}=0$. Such terms might be useful in order to treat the strongly nonlinear regime. Of particular interest might be to choose the coefficients α_i and β_i, which control the damping in the λ-equations, to be quadratic functions of the basic variables $(h^{ij},P^{ij},M^{ij}{}_k)$, so that the damping becomes stronger at points where the nonlinearities intensify.

It is fairly easy to implement similar schemes for alternative symmetric hyperbolic systems for the Einstein equations, as well as for symmetric hyperbolic systems for other theories with constraints, like, for instance, Yang–Mills theories. The strategy is the same: One writes equations with damping for first integrals of the constraints and modifies the evolution equations such that the extended system becomes symmetric hyperbolic. This can always be achieved, because the inclusion of the new equations modifies an off-diagonal sector of the principal symbol matrix.

III. ASYMPTOTIC STABILITY OF THE CONSTRAINT PROPAGATION

The inclusion of the λ-terms into (1)–(3) affects, in turn, the evolution of the constraint quantities $\mathcal{C}$, $\mathcal{C}^i$, $\mathcal{C}^{ij}{}_k$, and $\mathcal{C}^{ij}{}_{kl}$. Recalculating the time derivative of these, and using (24)–(26), yields the constraint evolution equations for the new system,

$$\dot{\mathcal{C}}=N^n\mathcal{C}_{,n}+3Q\sqrt{h}\mathcal{C}^k{}_{,k}-2\alpha_4 h^{mn}\lambda^{kl}{}_{km,nl}+2\alpha_0 h^{mn}\lambda_{,mn}+\cdots, \tag{29}$$

$$\dot{\mathcal{C}}^i=N^n\mathcal{C}^i{}_{,n}+Q\sqrt{h}(h^{ik}\mathcal{C}_{,k}+h^{rs}\mathcal{C}^{ik}{}_{rk,s})+\alpha_1 h^{m(n}\lambda^{i)}{}_{,mn}+\cdots, \tag{30}$$

$$\dot{\mathcal{C}}^{ij}{}_k = N^n \mathcal{C}^{ij}{}_{k,n} - 2Q\sqrt{h}(2\delta_k{}^{(i}\mathcal{C}^{j)} - h^{ij}h_{kl}C^l) + 2\alpha_3 h^{mn}\lambda^{ij}{}_{m,nk} + 2\alpha_4 h^{mn}(2\lambda^{ij}{}_{km,n} - h^{ij}h_{rs}\lambda^{rs}{}_{km,n}) - \alpha_0(2\delta^{(i}{}_k h^{j)l}\lambda_{,l} - h^{ij}\lambda_{,k}) + \cdots, \tag{31}$$

$$\dot{\mathcal{C}}^{ij}{}_{kl} = N^n \mathcal{C}^{ij}{}_{kl,n} - 2Q\sqrt{h}(\delta_k{}^{(i}\mathcal{C}^{j)}{}_{,l} - \delta_l{}^{(i}\mathcal{C}^{j)}{}_{,k}) + 2\alpha_4(h^{mn}\lambda^{ij}{}_{km,nl} - h^{mn}\lambda^{ij}{}_{lm,nk}) - \alpha_0(\delta_k{}^{(i}h^{j)m}\lambda_{,ml} + \delta_l{}^{(i}h^{j)m}\lambda_{,mk}) + \cdots. \tag{32}$$

Again the ellipses ''...'' represent undifferentiated terms that are linear in the constraint quantities and at least linear in $(P^{ij}, M^{ij}{}_k)$.

The propagation of the constraints is ruled by the system of equations consisting of (16)–(19) and (29)–(32), which determines whether or not the constraints asymptotically ''decay'' to zero. The crucial feature of this system is that the right-hand side also contains nonprincipal terms. Roughly speaking, the operator on the right-hand side amplifies constraint violations if the matrix representing its action on periodic functions has any eigenvalue with a positive real part. On the other hand, if all the eigenvalues have a negative real part, the operator induces an asymptotic decay of these violations.

Instead of attacking the full nonlinear problem as stated, which represents a problem well beyond the scope of present analytical techniques, we consider the linear regime of general relativity. That is, we restrict attention to three-metrics of the form $h^{ij} = e^{ij} + \epsilon\gamma^{ij}$ with $e^{ij} = \delta^{ij}$ and $\epsilon \ll 1$. This implies that the variables $(P^{ij}, M^{ij}{}_k)$ are of first order in ϵ, as are the constraint quantities $(\mathcal{C}, \mathcal{C}^i, \mathcal{C}^{ij}{}_k, \mathcal{C}^{ij}{}_{kl})$ and the variables $(\lambda, \lambda^i, \lambda^{ij}{}_k, \lambda^{ij}{}_{kl})$. Thus, the terms represented by the ellipses ''$\cdots$'' in Eqs. (29)–(32) are of second order in ϵ and shall be neglected. Without loss of generality, we restrict the following arguments to the case where the gauge source functions Q and N^i are constant. All arguments that follow refer to this linearized regime.

Although we lack a proof for the nonlinear case, the following considerations provide analytical evidence for the asymptotic stability of the constraint propagation, in particular since the full evolution equations are quasilinear.

For, as we believe, purely technical reasons, we adopt the following choice of coefficients: $\beta_0 = \beta_1 = \beta_3 = \beta_4 \equiv \beta > 0$ and $\alpha_4 = (\sqrt{3}/2)\alpha_0$.

Theorem 1: *With the above assumptions, the constraint submanifold of the linearized Einstein equations is an asymptotically stable submanifold for the solutions to the linearized, λ-extended Einstein equations.*

We partition the proof of this theorem in several lemmas: We first show that the initial value problem is well posed and that the solutions stay bounded with time. Thus, it is possible to apply Laplace transformation techniques, which reduce the problem to the study of the eigenfrequencies of the system. For these frequencies, we show that the real part is nonpositive, only approaches zero as the wave number goes to zero, and does so quadratically. Then stability follows from estimates in Ref. 6.

Without loss of generality, we expand the linearized dynamical fields in Fourier integrals of the the following form:

$$\lambda(x,t) = \int \hat{\lambda}(k,t)\exp(ik\cdot x)d^3k, \tag{33}$$

$$\lambda^i(x,t) = \int \hat{\lambda}^i(k,t)\exp(ik\cdot x)d^3k, \tag{34}$$

$$\vdots \tag{35}$$

$$C^{ij}{}_k(x,t) = \int \hat{C}^{ij}{}_k(k,t)\exp(ik\cdot x)d^3k, \tag{36}$$

$$C^{ij}{}_{kl}(x,t)=\int \hat{C}^{ij}{}_{kl}(k,t)\exp(ik\cdot x)d^3k, \tag{37}$$

where $k\cdot x:=k_i x^i$.

In terms of the Fourier transformed variables, Eqs. (29)–(32) and (16)–(19) reduce to the system of ordinary differential equations given by

$$\dot{\hat{\lambda}}=-\beta\hat{\lambda}+\alpha_0\hat{\mathcal{C}}, \tag{38}$$

$$\dot{\hat{\lambda}}{}^i=-\beta\hat{\lambda}^i+\alpha_1\hat{\mathcal{C}}^i, \tag{39}$$

$$\dot{\hat{\lambda}}{}^{ij}{}_{kl}=-\beta\hat{\lambda}^{ij}{}_{kl}+\frac{\sqrt{3}}{2}\alpha_0\hat{\mathcal{C}}^{ij}{}_{kl}, \tag{40}$$

$$\dot{\hat{\mathcal{C}}}=ik_nN^n\hat{\mathcal{C}}+3iQk_m\hat{\mathcal{C}}^m+\sqrt{3}\alpha_0\hat{\lambda}^{rl}{}_{rm}k^mk_l-2\alpha_0\hat{\lambda}k^nk_n, \tag{41}$$

$$\dot{\hat{\mathcal{C}}}{}^i=ik_nN^n\hat{\mathcal{C}}^i+iQ(k^i\hat{\mathcal{C}}+k^r\hat{\mathcal{C}}^{in}{}_{rn})-\tfrac{1}{2}\alpha_1(k^nk_n\hat{\lambda}^i+k^ik_n\hat{\lambda}^n), \tag{42}$$

$$\begin{aligned}\dot{\hat{\mathcal{C}}}{}^{ij}{}_{kl}=&ik_nN^n\hat{\mathcal{C}}^{ij}{}_{kl}-2iQ(\delta_k{}^{(i}\hat{\mathcal{C}}^{j)}k_l-\delta_l{}^{(i}\hat{\mathcal{C}}^{j)}k_k)-\sqrt{3}\alpha_0(\hat{\lambda}^{ij}{}_{kr}k^rk_l-\hat{\lambda}^{ij}{}_{lr}k^rk_k)\\&+\alpha_0\hat{\lambda}(\delta_k{}^{(i}k^{j)}k_l-\delta_l{}^{(i}k^{j)}k_k),\end{aligned} \tag{43}$$

and

$$\dot{\hat{\lambda}}{}^{ij}{}_k=-\beta\hat{\lambda}^{ij}{}_k+\alpha_3\hat{\mathcal{C}}^{ij}{}_k, \tag{44}$$

$$\dot{\hat{\mathcal{C}}}{}^{ij}{}_k=ik_nN^n\hat{\mathcal{C}}^{ij}{}_k-\alpha_3\hat{\lambda}^{ij}{}_mk^mk_k+\hat{S}^{ij}{}_k, \tag{45}$$

where

$$\hat{S}^{ij}{}_k:=-2Q(2\delta_k{}^{(i}\hat{\mathcal{C}}^{j)}-\delta^{ij}\hat{\mathcal{C}}_k)-\alpha_0(2\delta^{(i}{}_kk^{j)}-h^{ij}k_k)\hat{\lambda}+\sqrt{3}\alpha_0k^m(2\hat{\lambda}^{ij}{}_{km}-h^{ij}h_{rs}\hat{\lambda}^{rs}{}_{km}). \tag{46}$$

This system of equations naturally splits up in two subsystems, since Eqs. (38)–(42) couple to Eqs. (44) and (45) only via the ''source'' term in (45). In the following, we will first establish that the solutions to the subsystem (38)–(42), and hence the coupling term in (45), asymptotically decay to zero. In a second step, we consider this coupling as a given, decaying source, and discuss the asymptotic behavior of solutions to the subsystem (44)–(45).

Lemma 1: Let $\mathcal{H}$ be the space of the Fourier transformed $\hat{\lambda}^{ij}{}_{kl}\in L^2$, and let $\mathcal{D}\subset\mathcal{H}$ be the subspace defined by $\hat{\lambda}^{ij}{}_{ks}k^s=0$. Then $\mathcal{D}$ is invariant under time evolution, and the trivial solution $\hat{\lambda}^{ij}{}_{kl}=0$ is asymptotically stable for the evolution restricted to $\mathcal{D}$.

Proof: Multiplying Eq. (40) by k_m, antisymmetrizing, and using that $\hat{C}^{ij}{}_{[kl}k_{m]}=\hat{M}^{ij}{}_{[k}k_lk_{m]}=0$, we obtain

$$\dot{\hat{\lambda}}{}^{ij}{}_{[kl}k_{m]}=-\beta\hat{\lambda}^{ij}{}_{[kl}k_{m]}. \tag{47}$$

Next we note that for a function $\hat{\lambda}^{ij}{}_{kl}$ in $\mathcal{H}$, the component $(\hat{\lambda}^{ij}{}_{kl})^{\parallel}$ in $\mathcal{D}$ is given by $(\hat{\lambda}^{ij}{}_{kl})^{\parallel}=\hat{\alpha}^{ij}k^r\epsilon_{rkl}$, where $\hat{\alpha}^{ij}=\hat{\lambda}^{ij}{}_{[kl}k_{r]}\epsilon^{klr}/(6k^2)$. Equation (47) is, therefore, equivalent to

$$(\dot{\hat{\lambda}}{}^{ij}{}_{kl})^{\parallel}=-\beta(\hat{\lambda}^{ij}{}_{kl})^{\parallel}, \tag{48}$$

which proves lemma 1.

By direct inspection of the evolution equations, it follows that the equation for the component of $\hat{\lambda}^{ij}{}_{kl}$ in the subspace $\mathcal{D}$ decouples. It is, therefore, sufficient to concentrate on the evolution in the space $CF_\lambda \oplus CF_C$ which comprises those functions $(\hat{\lambda},\hat{\lambda}^i,\hat{\lambda}^{ij}{}_{kl},\hat{\mathcal{C}},\hat{\mathcal{C}}^i,\hat{\mathcal{C}}^{ij}{}_{kl}) \in L^2$ for which $\hat{\lambda}^{ij}{}_{kl} \in \mathcal{D}^\perp$. Here, $\mathcal{D}^\perp$ denotes the L^2 complement of $\mathcal{D}$ in $\mathcal{H}$, which, as easily seen, is spanned by the elements $\hat{\lambda}^{ij}{}_{kl} \in L^2$ satisfying $\hat{\lambda}^{ij}{}_{[kl}k_{m]}=0$. (For functions $\hat{\lambda}^{ij}{}_{kl}$ in $\mathcal{D}$ only the components along k^m are nontrivial, $\hat{\lambda}^{ij}{}_{kl}=-2k_{[l}\hat{\lambda}^{ij}{}_{k]m}k^m/k^2$. This can be seen by solving $\hat{\lambda}^{ij}{}_{[kl}k_{m]}=0$, and by using the antisymmetry in the lower indices of $\hat{\lambda}^{ij}{}_{kl}$.) Since for the constraint variable $\hat{C}^{ij}{}_{kl}$, the same property is fulfilled, $\hat{C}^{ij}{}_{[kl}k_{m]}=0$, this shows that the spaces CF_λ and CF_C are naturally isomorphic, $CF_\lambda \approx CF_C =: CF$.

To simplify the notation, and to display the structure of the evolution equations considered more transparently, let us introduce the following operator $\mathbf{E}$ acting on functions $\mathbf{v}\equiv(v,v^i,v^{ij}{}_{kl})$ in CF:

$$\mathbf{E}(\mathbf{v})\equiv(E(\mathbf{v}),E^i(\mathbf{v}),E^{ij}{}_{kl}(\mathbf{v})), \tag{49}$$

where

$$E(\mathbf{v})\equiv\sqrt{3}\,\alpha_0 v^{rl}{}_{rm}k^m k_l - 2\alpha_0 v k^n k_n\,, \tag{50}$$

$$E^i(\mathbf{v})\equiv -\tfrac{1}{2}\alpha_1(v^i k^n k_n + v^n k^i k_n), \tag{51}$$

$$E^{ij}{}_{kl}(\mathbf{v})\equiv -\sqrt{3}\,\alpha_0(v^{ij}{}_{kn}k^n k_l - v^{ij}{}_{ln}k^n k_k) + \alpha_0 v(\delta_k{}^{(i}k^{j)}k_l - \delta_l{}^{(i}k^{j)}k_k). \tag{52}$$

Taking advantage of these definitions, the evolution system (38)–(43) restricted to the subspace $CF\oplus CF$ can be rewritten as

$$\frac{d}{dt}\begin{pmatrix}\boldsymbol{\lambda}\\ \mathbf{C}\end{pmatrix} = \begin{pmatrix}-\mathbf{S} & \boldsymbol{\Gamma}\\ \mathbf{E} & i\mathbf{A}\end{pmatrix}\begin{pmatrix}\boldsymbol{\lambda}\\ \mathbf{C}\end{pmatrix} =: \mathbf{P}\begin{pmatrix}\boldsymbol{\lambda}\\ \mathbf{C}\end{pmatrix}, \tag{53}$$

where $\mathbf{S}$ and $\boldsymbol{\Gamma}$ are diagonal matrices determined by the parameters β and α_i, respectively, and where $\mathbf{A}$ is an operator of the form $\mathbf{A}^m k_m$.

In a next step, we show that the operator $e^{\mathbf{P}t}$ is bounded with respect to a suitably chosen norm. To this end, we first establish the following:

Lemma 2: The operator $H_\lambda := -\boldsymbol{\Gamma}^{-1}H_c\mathbf{E}$ considered as a matrix-valued field on the Fourier space $\mathbf{R}^3$ is symmetric and coercive with respect to the inner product $\langle\mathbf{u},\mathbf{v}\rangle := \bar{u}v + e_{ij}\bar{u}^i v^j + e_{ip}e_{jq}e^{kr}e^{ls}\bar{u}^{ij}{}_{kl}v^{pq}{}_{rs}$. That is, $\langle\mathbf{u},H_\lambda\mathbf{v}\rangle = \langle H_\lambda\mathbf{u},\mathbf{v}\rangle$ for all $\mathbf{u},\mathbf{v}\in CF$, and there exists a constant $c>0$ such that $\langle\mathbf{u},H_\lambda\mathbf{u}\rangle \geqslant ck^2(\mathbf{u},\mathbf{u})$ for all $\mathbf{u}\in CF$.

Proof: We have

$$\begin{aligned}
&\langle u,\boldsymbol{\Gamma}^{-1}H_c\mathbf{E}(\mathbf{v})\rangle - \langle\boldsymbol{\Gamma}^{-1}H_c\mathbf{E}(\mathbf{u}),\mathbf{v}\rangle\\
&= \frac{1}{3\alpha_0}\bar{u}(-2\alpha_0 v k_n k^n + \sqrt{3}\,\alpha_0 v^{kl}{}_{km}k^m k_l) + \frac{1}{\alpha_1}\bar{u}^i\left(-\frac{\alpha_1}{2}(v^j k_n k^n + k^j v^l k_l)\right)e_{ij}\\
&+ \frac{1}{2\sqrt{3}\,\alpha_0}e_{im}e_{jn}e^{kp}e^{lq}\bar{u}^{ij}{}_{kl}(-2\sqrt{3}\,\alpha_0 v^{mn}{}_{ps}k^s k_q + 2\alpha_0 v\,\delta^m{}_p k^n k_q)\\
&- \frac{1}{3\alpha_0}v(-2\alpha_0\bar{u}k_n k^n + \sqrt{3}\,\alpha_0\bar{u}^{kl}{}_{km}k^m k_l) - \frac{1}{\alpha_1}v^i\left(-\frac{\alpha_1}{2}(\bar{u}^j k_n k^n + k^j\bar{u}^l k_l)\right)e_{ij}
\end{aligned}$$

$$+\frac{1}{2\sqrt{3}\alpha_0} e_{im} e_{jn} e^{kp} e^{lq} v^{ij}{}_{kl}(-2\sqrt{3}\alpha_0 \overline{u}^{mn}{}_{ps} k^s k_q + 2\alpha_0 \overline{u} \delta^m{}_p k^n k_q) = 0.$$

The remaining part of the proof is given in Appendix A, where we show that H_λ is coercitive with constant $c = 1/4$.

With the help of lemma 2, it is now easy to prove

Lemma 3: The matrix-valued fields $\mathbf{P}_\pm$,

$$\mathbf{P}_+ := \begin{pmatrix} \mathbf{S} & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{P}_- := \begin{pmatrix} 0 & \mathbf{\Gamma} \\ \mathbf{E} & i\mathbf{A} \end{pmatrix}, \tag{54}$$

are Hermitian, respectively, anti-Hermitian with respect to the inner product

$$\langle (\boldsymbol{\lambda}_1, \mathbf{C}_1), H_T(\boldsymbol{\lambda}_2, \mathbf{C}_2)\rangle := \langle \boldsymbol{\lambda}_1, H_\lambda \boldsymbol{\lambda}_2\rangle + \langle \mathbf{C}_1, H_c \mathbf{C}_2\rangle. \tag{55}$$

Proof: Since $\mathbf{S} = \beta \mathbf{I}$, the statement for $\mathbf{P}_+$ is trivially true. The antisymmetry of $\mathbf{P}$ follows directly from lemma 2, and the symmetry of $\mathbf{A}$ with respect to H_c.

Taking advantage of lemma 3, we now obtain the following important estimate for the operator $\mathbf{P} = \mathbf{P}_+ + \mathbf{P}_-$:

$$H_T \mathbf{P} + \mathbf{P}^\dagger H_T = H_\lambda \mathbf{S} + \mathbf{S} H_\lambda = -2\beta H_\lambda \leqslant -2\beta \tfrac{1}{4} k_n k^n \leqslant 0, \tag{56}$$

where, for any Hermitian matrix $\mathbf{M}$, the inequality $\mathbf{M} \leqslant 0$ means $\langle \mathbf{v}, \mathbf{M}\mathbf{v}\rangle \leqslant 0$ for all $\mathbf{v}$. (Clearly, there are other possible choices of the operators $\mathbf{S}$ which lead to the same inequality. Here we have restricted ourselves to the simplest possibility, but for practical applications, alternative choices might be better suited.)

The symmetry and coercivity of the operator H_λ imply that H_λ can be used to define a scalar product on a (dense) subspace $\mathcal{D}(CF)$ of the Hilbert space CF,[7] which, in turn, shows that the operator $H_T = H_c + H_\lambda$ gives rises to a scalar product on $CF \oplus \mathcal{D}(CF)$. (In physical space, the relevant function space equipped with the norm corresponding to the above scalar product is very similar to the Sobolev space H_0^1.)

As is well known (see, for instance, Ref. 7), the estimate (56) implies that for all $t > 0$, the operator $e^{\mathbf{P}t}$ is bounded with respect to the norm defined by H_T. Hence, the initial value problem for the system considered is well posed. Moreover, all solutions with initial data which are bounded with respect to this norm remain bounded for all positive times. Thus Laplace transformation techniques can be applied,[7] and the relevant questions are the *sign* of the real part of the eigenvalues of $\mathbf{P}$, and how fast they approach zero as the wave number $k = \sqrt{k^i k_i}$ goes to zero. Hence, the proof is reduced to the eigenvalue problem for the operator $\mathbf{P}$,

$$\mathbf{P}\begin{pmatrix} \boldsymbol{\lambda}_s \\ \mathbf{C}_s \end{pmatrix} = s \begin{pmatrix} \boldsymbol{\lambda}_s \\ \mathbf{C}_s \end{pmatrix}. \tag{57}$$

Then we have the following:

Lemma 4: The eigenvalues of the above system have a nonpositive real part and furthermore there exist positive constants c_1 *and* w_1 *such that*

$$\Re(s) \leqslant -c_1 \frac{k^2}{w_1 + k^2} \tag{58}$$

for all wave vectors k_i.

Proof: From the $\boldsymbol{\lambda}$-rows of the eigenvalue equation, we get

$$\mathbf{C}_s = (s + \beta)\mathbf{\Gamma}^{-1} \boldsymbol{\lambda}_s. \tag{59}$$

Using this in the **C**-rows, we next obtain

$$(\mathbf{E}+(s+\beta)(-s\mathbf{I}+i\mathbf{A})\mathbf{\Gamma}^{-1})\boldsymbol{\lambda}_s=0. \tag{60}$$

Multiplying Eq. (60) by $-(\mathbf{\Gamma}^{-1})^\dagger\mathbf{H}_c$ from the left and subsequently contracting with $\boldsymbol{\lambda}_s$, we find the following second-order equation for the eigenvalue s:

$$\langle\boldsymbol{\lambda}_s,H_\lambda\boldsymbol{\lambda}_s\rangle+(s+\beta)(s\langle\boldsymbol{\lambda}_s,(\mathbf{\Gamma}^{-1})^\dagger H_c\mathbf{\Gamma}^{-1}\boldsymbol{\lambda}_s\rangle-i\langle\boldsymbol{\lambda}_s,(\mathbf{\Gamma}^{-1})^\dagger H_c\mathbf{A}\mathbf{\Gamma}^{-1}\boldsymbol{\lambda}_s\rangle)=0. \tag{61}$$

The established properties of the involved operators imply that

$$c(k_i^0)k^2:=\frac{\langle\boldsymbol{\lambda}_s,H_\lambda\boldsymbol{\lambda}_s\rangle}{\langle\boldsymbol{\lambda}_s,(\mathbf{\Gamma}^{-1})^\dagger H_c\mathbf{\Gamma}^{-1}\boldsymbol{\lambda}_s\rangle} \tag{62}$$

is positive for $k_i\neq 0$, and that

$$b(k_i^0)k:=\frac{\langle\boldsymbol{\lambda}_s,(\mathbf{\Gamma}^{-1})^\dagger H_c\mathbf{A}\mathbf{\Gamma}^{-1}\boldsymbol{\lambda}_s\rangle}{\langle\boldsymbol{\lambda}_s,(\mathbf{\Gamma}^{-1})^\dagger H_c\mathbf{\Gamma}^{-1}\boldsymbol{\lambda}_s\rangle} \tag{63}$$

is real, where k_i^0 denotes the unit vector in the direction of k_i, and k is the norm of k_i. Thus we have for each direction of k^i

$$(s+\beta)(s-ibk)+ck^2=0 \tag{64}$$

with β, b, c real and β, c positive. For this equation we prove in Appendix B that the real part of the roots satisfies the desired inequality, which establishes the result for each direction of the wave vector k_i. Using the maximal values of $-c_1k^2/(w_1+k^2)$ on the two-sphere of directions of k_i, we obtain the final inequality.

With this bound on the decay constants, it is now easy to prove asymptotic stability for the subsystem (38)–(43). Splitting the set of solutions into a part with frequencies with $k<1$, and another with $k\geqslant 1$, the above bound tells us that the solutions of the higher frequency part decay faster than $\exp(-c_1t/(w_1+1))$, while the decay of the solutions of the low frequency part can be estimated as in Ref. 6, lemma 1 and 2 of Sec. III].

We now turn our attention to the second set of equations, given by (44) and (45), and establish the following:

Lemma 5: Let $\mathcal{H}_3$ *be the space of the Fourier transformed* $(\hat{\lambda}^{ij}{}_k,\hat{C}^{ij}{}_k)\in L^2$. *Then* $\mathcal{H}_3$ *is invariant under time evolution, and the trivial solution* $(\hat{\lambda}^{ij}{}_k,\hat{C}^{ij}{}_k)=0$ *is asymptotically stable for the evolution restricted to* $\mathcal{H}_3$.

Proof: In a first step, we discuss the equation for the component of a solution in the subspace

$$\mathcal{D}_3:=\{(\hat{\lambda}^{ij}{}_k,\hat{C}^{ij}{}_k)\in L^2|\hat{\lambda}^{ij}{}_mk^m=\hat{C}^{ij}{}_mk^m=0\}. \tag{65}$$

Taking advantage of Eqs. (8), (12), and (44), we obtain

$$\dot{\hat{\lambda}}^{ij}{}_{[k}k_{l]}=-\beta\hat{\lambda}^{ij}{}_{[k}k_{l]}+\alpha_3\hat{C}^{ij}{}_{[k}k_{l]}, \tag{66}$$

$$i\hat{C}^{ij}{}_{[k}k_{l]}=\hat{C}^{ij}{}_{kl}-\tfrac{1}{2}\delta^{ij}\delta_{mn}\hat{C}^{mn}{}_{kl}, \tag{67}$$

which implies that the space $\mathcal{D}_3$ is invariant under time evolution. As already shown, the constraint variable $\hat{C}^{ij}{}_{kl}$ asymptotically decays to zero. The dynamics in $\mathcal{D}_3$ is, therefore, described by a system of ordinary differential equations of the form $\dot{u}=-u+f$, where f is a given source with $f\to 0$ as $t\to\infty$. Since any solution to this system satisfies $u\to 0$ as $t\to\infty$, it follows that solutions in $\mathcal{D}_3$ decay with time. [For a proof, choose T such that $f(t)<\epsilon/2$ for all $t>T$. Since the general solution to the above system is given by $u(t)=e^{-t}(u(0)+\int_0^t e^{\tilde{t}} f(\tilde{t})d\tilde{t})$, it follows that $u(t) \leq e^{-t}(u(0)+\int_0^T e^{\tilde{t}} f(\tilde{t})d\tilde{t}-\epsilon e^T/2)+\epsilon/2$. Hence, for a sufficiently large time $t_0>T$, the absolute value of the first term becomes smaller than $\epsilon/2$, which implies $|u(t)|<\epsilon$ for all $t>t_0$.]

It remains to discuss the complementary subspace $\mathcal{D}_3^\perp$,

$$\mathcal{D}_3^\perp=\{(\lambda^{ij}{}_k,\hat{C}^{ij}{}_k)\in L^2|\hat{\lambda}^{ij}{}_{[k}k_{l]}=\hat{C}^{ij}{}_{[k}k_{l]}=0\}. \tag{68}$$

For the component of a solution in this subspace, we find

$$\frac{d}{dt}\begin{pmatrix}\boldsymbol{\lambda}_3\\ \mathbf{C}_3\end{pmatrix}=\begin{pmatrix}-\beta & \alpha_3\\ -\alpha_3 k^2 & ik_m N^m\end{pmatrix}\begin{pmatrix}\boldsymbol{\lambda}_3\\ \mathbf{C}_3\end{pmatrix}-\begin{pmatrix}0\\ F^{ij}k_k\end{pmatrix}, \tag{69}$$

where $(\boldsymbol{\lambda}_3,\mathbf{C}_3):=(\hat{\lambda}^{ij}{}_k,\hat{C}^{ij}{}_k)^\perp\in\mathcal{D}_3^\perp$, and where $\hat{F}^{ij}k_k$ is a shorthand for the perpendicular component of the source term $\hat{S}^{ij}{}_k$, $\hat{F}^{ij}=\hat{S}^{ij}{}_m k^m/k^2$. Thus, as expected, the subspace $\mathcal{D}_3$ is invariant as well.

Since $\hat{P}^{ij}$ and consequently $\hat{C}^i/|k|=\hat{P}^{im}k_m/|k|$ are contained in L^2, Eq. (46) implies that the same is true for $\hat{F}^{ij}$, $\hat{F}^{ij}\in L^2$. Furthermore, the real part of the eigenvalues of the system (69) can, as in lemma 4, be estimated by the inequality (58), albeit for different constants. Adopting a similar reasoning as in the previous discussion, and applying lemmas 1 and 2 of Ref. 6 to this system, it follows that solutions in $\mathcal{D}_3^\perp$ also decay with time.

This completes the proof of lemma 5 and hence the proof of our main result.

IV. CONCLUSIONS

In the present paper we have shown that an arbitrary system of symmetric hyperbolic evolution equations with constraints admits extensions to symmetric hyperbolic systems which reproduce the original dynamics on the embedded constraint submanifold. We have given analytical evidence that the class of extensions proposed is sufficiently rich to contain systems for which the embedded constraint submanifold is an attractor of the time evolution. For the Einstein equations, we have constructed an extended evolution system for which, at least in the linearized case, this property is fulfilled.

It is natural to expect that, by making use of techniques developed in Ref. 6, the results proven for the linearized Einstein equations can be generalized to the regime of nonlinear general relativity describing space–times in the vicinity of Minkowski space. However, to establish similar results for more extended regions of the phase space of general relativity is well beyond the scope of present analytic techniques.

Numerical experiences with the Navier–Stokes equations for incompressible fluids show that asymptotic stability of the constraint submanifold is essential for accurate results.[8] For this system, techniques with a very similar effect have been used to include the incompressibility constraint into the evolution equations. On the basis of this observation, and the results established for linearized gravity, we suspect that the extensions of Einstein's equations constructed could be of interest when obtaining numerical solutions to general relativity. Numerical experiments testing aspects of this conjecture are in progress.

ACKNOWLEDGMENTS

Work supported by CONICOR, CONICET, the Swiss NSF, SeCyT, UNC, and by NSF under Grant No. PHY-9803301. O.B. would like to thank Chris Beetle for discussions and interesting comments. S.F. and O.R. gratefully acknowledge the hospitality of the Albert Einstein Institut, Potsdam, where part of this work was carried out. P.H. would like to thank Bernd Schmidt for interesting discussions.

APPENDIX A: PROOF OF COERCIVITY

In this Appendix we show that

$$\langle \mathbf{u}, H_\lambda \mathbf{u} \rangle \geq \tfrac{1}{4} k_n k^n \tag{A1}$$

for unitary $\mathbf{u}$ satisfying $u^{ij}{}_{rs} = u^{(ij)}{}_{[rs]}$ and $u^{ij}{}_{[rs}k_{l]} = 0$, as needed for Lemma 2. We treat this as the problem of extremizing the quadratic function of $\mathbf{u}$ on the left-hand side of (A1) under the constraint condition $\langle \mathbf{u}, \mathbf{u} \rangle = 1$.

From (50)–(52), we obtain

$$\langle \mathbf{u}, H_\lambda \mathbf{u} \rangle + \tau k^n k_n (1 - \langle \mathbf{u}, \mathbf{u} \rangle) = \tfrac{2}{3} k^n k_n u \overline{u} + \tfrac{1}{2} k^n k_n u^i \overline{u_i} + \tfrac{1}{2} u^i k_i \overline{u_n} k^n + u^{ij}{}_{rs} k^s \overline{u_{ij}}{}^{rm} k_m - \frac{\sqrt{3}}{3} \overline{u} u^{ij}{}_{is} k^s k_j$$
$$- \frac{1}{\sqrt{3}} u \overline{u_{ij}}{}^{is} k_s k^j + \tau k^n k_n (1 - u\overline{u} - u^i \overline{u_i} - u^{ij}{}_{rs} \overline{u_{ij}}{}^{rs}), \tag{A2}$$

where τ is a Lagrange multiplier and where indices are raised and lowered with e_{ij}. To simplify the algebra, we choose a basis in which $k^n = (0,0,k)$. Then $u^{ij}{}_{rs} = 0$, except when $s = 3$. Hence,

$$F(\mathbf{u},\tau) \equiv \langle \mathbf{u}, H_\lambda \mathbf{u} \rangle + \tau k^n k_n (1 - \langle \mathbf{u}, \mathbf{u} \rangle)$$
$$= k^2 \left(\tfrac{2}{3} u \overline{u} + \tfrac{1}{2} u^i \overline{u_i} + \tfrac{1}{2} u^3 \overline{u_3} + u^{ij}{}_{r3} \overline{u_{ij}}{}^{r3} - \frac{1}{\sqrt{3}} \overline{u} u^{i3}{}_{i3} - \frac{1}{\sqrt{3}} u \overline{u_{i3}}{}^{i3} \right.$$
$$\left. + \tau (1 - u \overline{u} - u^i \overline{u_i} - u^{ij}{}_{rs} \overline{u_{ij}}{}^{rs}) \right). \tag{A3}$$

The function $F(\mathbf{u},\tau)$ is extremized at points $(\mathbf{u},\tau)$ where

$$\frac{\partial F}{\partial \mathbf{u}} + \frac{\partial F}{\partial \overline{\mathbf{u}}} = 0, \tag{A4}$$

$$\frac{\partial F}{\partial \mathbf{u}} - \frac{\partial F}{\partial \overline{\mathbf{u}}} = 0, \tag{A5}$$

$$\frac{\partial F}{\partial \tau} = 0. \tag{A6}$$

Equation (A6) is the requirement that $\mathbf{u}$ has unit length. Equations (A4) and (A5) constitute a homogeneous linear system of equations for the real and imaginary parts of $\mathbf{u}$. Since $\mathbf{u}$ cannot vanish, the determinant of the linear system has to vanish. Up to numerical factors, this is given by

$$(2\tau - 1)^{13} (\tau - 1)^2 (\tau - \tfrac{1}{6}). \tag{A7}$$

As easily verified, $\tau = 1$ yields the following minimal value of $F(\mathbf{u},\tau)$ [when evaluated at unit $\mathbf{u}$ such that (A4)–(A5) are satisfied]:

$$F(\mathbf{u}_{\min}, 1) = \tfrac{1}{4} k^2, \tag{A8}$$

from which (A1) follows. The other extreme values of $F(\mathbf{u},\tau)$ are $(5/3)k^2$ and $(1/2)k^2$ for $\tau = 1/6, 1/2$.

APPENDIX B: ON THE PROOF OF LEMMA 4

In this Appendix we prove that the roots $s_\pm$ of the polynomial

$$P(s)=s^2+s(\beta-ibk)+ck^2 \tag{B1}$$

are subject to the inequality

$$\Re(s_\pm)\leqslant -c_1\frac{k^2}{w_1+k^2}, \tag{B2}$$

where $c_1=\beta c/(b^2+4c)$ and $w_1=\beta^2/(b^2+4c)$. As in the body of the text, it is assumed that the parameters of P are real, and that β and c are strictly positive.

To begin with, let us rewrite the polynomial P, and the above estimate in terms of suitably rescaled parameter. Defining

$$\tilde{s}=s/\beta,\quad \tilde{k}=2k\sqrt{b^2+4c}/\beta,\quad \tilde{b}=b/\sqrt{b^2+4c}, \tag{B3}$$

and dropping tildes, we obtain for the polynomial

$$P/\beta^2=s^2+s(1-ibk)+(1-b^2)k^2/4. \tag{B4}$$

The estimate for the roots in terms of the scaled parameters assumes the form

$$\Re(s_\pm)\leqslant -\frac{\gamma^2}{4}\,\frac{k^2}{1+k^2}, \tag{B5}$$

where $\gamma^2:=1-b^2\in(0,1]$.

As easily verified, the roots of the scaled polynomial satisfy

$$\max\{\Re(2s_+),\Re(2s_-)\}=-1+|\Re\sqrt{1-k^2+2ibk}|. \tag{B6}$$

It is, therefore, sufficient to show that

$$|\Re\sqrt{1-k^2+2ibk}|\leqslant 1-\frac{\gamma^2}{2}\,\frac{k^2}{1+k^2}. \tag{B7}$$

To give a proof of this inequality, we first evaluate the identity

$$2|\Re\sqrt{z}|^2=|z|+\Re(z) \tag{B8}$$

for $z:=1-k^2+2ibk$,

$$2|\Re\sqrt{z}|^2=\sqrt{(1-k^2)^2+4b^2k^2}+(1-k^2)=\sqrt{(1+k^2)^2-4\gamma^2k^2}-(1+k^2)+2.$$

Hence,

$$|\Re\sqrt{z}|^2=1+(1+k^2)\{\sqrt{1-4\gamma^2k^2/(1+k^2)^2}-1\}/2\leqslant 1-\gamma^2\frac{k^2}{1+k^2}, \tag{B9}$$

where we have used the estimate $\sqrt{1-x}\leqslant 1-x/2$, which holds for $x\leqslant 1$. Taking advantage of the latter estimate once again, it follows that

$$|\Re\sqrt{z}|\leqslant 1-\frac{\gamma^2}{2}\,\frac{k^2}{1+k^2}, \tag{B10}$$

which completes the proof of our claim.

[1] M. W. Choptuik, ''Consistency of finite-difference solutions of Einstein's equations,'' Phys. Rev. D **44**, 3124 (1991).
[2] S. Detweiler, ''Evolution of the constraint equations in general relativity,'' Phys. Rev. D **35**, 1095 (1987).
[3] S. Frittelli and O. A. Reula, ''On the Newtonian limit of general relativity,'' Commun. Math. Phys. **166**, 221 (1994).
[4] S. Frittelli and O. A. Reula, ''First-order symmetric hyperbolic Einstein equations for arbitrary fixed gauge,'' Phys. Rev. Lett. **76**, 4667 (1996).
[5] J. M. Stewart, ''The Cauchy problem and the initial boundary value problem in numerical relativity,'' Class. Quantum Grav. **15**, 2865 (1998).
[6] G. Kreiss, H.-O. Kreiss, and J. Lorenz, ''On stability of conservation laws,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. (to be published).
[7] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations* (Academic, Boston, 1989).
[8] H.-O. Kreiss (private communication).

Gravitating brane systems: Some general theorems

K. A. Bronnikov[a)]
Centre for Gravitation and Fundamental Metrology,
VNIIMS, 3-1 M. Ulyanovoy St., Moscow 117313, Russia

(Received 4 August 1998; accepted for publication 9 September 1998)

Multidimensional gravity interacting with intersecting electric and magnetic *p*-branes is considered for fields depending on a single variable. Some general features of the system behavior are revealed without solving the field equations. Thus, essential asymptotic properties of isotropic cosmologies are indicated for different signs of spatial curvature; a no-hair-type theorem and a single-time theorem for black holes are proved (the latter makes sense in models with multiple time coordinates). The validity of the general observations is verified for a class of exact solutions known for the cases when certain vectors, built from the input parameters of the model, are either orthogonal in minisuperspace, or form mutually orthogonal subsystems. From the nonexistence of Lorentzian wormholes, a universal restriction is obtained, applicable to orthogonal or block-orthogonal subsystems of any *p*-brane system. © *1999 American Institute of Physics.* [S0022-2488(99)01402-4]

I. INTRODUCTION

In the weak field limits of the bosonic sectors of supergravities,[1] superstring and M-theory, their generalizations and modifications,[2–6] there naturally appear multiple self-gravitating scalar dilatonic fields and antisymmetric forms, associated with *p*-branes.

This paper continues the studies of such models on the basis of a general action, see (1), without fixing the total space-time dimension D or other input parameters,[7–16] thus to a large extent abstracting from the details of specific underlying models, but with a hope to predict some features of new models, unformulated by now. We will here deal with the one-variable sector of the model, where all fields depend on a single coordinate: time in cosmological models, a radial coordinate in spherically symmetric models, etc. In this case the model reduces to a Toda-type dynamical system in minisuperspace, see (15), (16).

Much work has been devoted to searches for exact solutions and their subsequent analysis. Thus, in Ref. 15 the most general one-variable solution was presented for the case when certain vectors Y_s in the target space, built from the input parameters of the model, form an orthogonal system (OS). This solution describes a set of intersecting electrically and magnetically charged *p*-branes and generalized many previously found ones, beginning with Schwarzschild and Reissner-Nordström and ending with dilatonic and some more special *p*-brane solutions (see Refs, 4, 16–19, and references therein). The OS solution was further generalized[20] to models where Y_s forms a block-orthogonal system (BOS). The OS solution is recovered when each block consists of a single vector. Other families of exact solutions have been found for cases when Y_s forms the bases of integrable Toda models, see Refs. 14, 19, 21, and references therein. Many solutions are known beyond the one-variable sector (see Ref. 22, and references therein).

The exact solutions have disclosed many features of interest of physically relevant configurations, such as cosmological models and black holes. The generality of these features remains, however, questionable, since the equations of motion can be solved exactly only for special (though numerous) choices of the input parameters. To have an idea of what can and what cannot be expected from yet unknown solutions, it makes sense to try to extract some information directly from the equations. Such an attempt is undertaken here.

a)Electronic mail: kb@rgs.mccme.ru

It appears possible to reveal some important properties of p-brane cosmologies, namely, the nature of asymptotics for different signs of spatial curvature. For spherically symmetric configurations, among other results, two theorems about black holes (BHs) are proved: a ''no-hair theorem,'' that a BH is incompatible with the so-called quasiscalar F-forms (see (5)), and a ''single-time theorem,'' that even in spaces with multiple times a black hole may only exist with its unique, one-dimensional (physical) time.

One more general observation is[20] the absence of spherically symmetric Lorentzian wormholes under the requirement that all the fields bear positive energy, just as in conventional general relativity. On the other hand, for the known families [OS and BOS (Ref. 20)] of exact solutions one can deduce necessary and sufficient conditions under which a specific solution describes a wormhole, no matter, Lorentzian or Euclidean. Combined, these results lead to a universal restriction upon the input parameters of the model, valid for any brane system which has an OS or BOS subsystem (Theorems 4 and 4a, already announced[20] in a slightly different form). Having been obtained on the basis of specific exact solutions, this restriction still applies to systems for which solutions are yet to be found.

The paper is organized as follows. The introductory Sec. II describes the model and a convenient Toda-type representation of its one-variable sector, in line with our previous papers. Section III is devoted to general properties of cosmological and spherically symmetric p-brane configurations. Section IV gives a brief description of the OS and BOS solutions, necessary for obtaining the above-mentioned universal restriction. The latter is formulated in Sec. V. Section VI contains some concluding remarks, in particular, on the use of different conformal frames.

II. THE MODEL: MINISUPERSPACE REPRESENTATION

The starting point is, as in Refs. 10–15, the model action for D-dimensional gravity with several scalar dilatonic fields φ^a and antisymmetric n_s-forms F_s:

$$S=\frac{1}{2\kappa^2}\int_{\mathcal{M}} d^D z\sqrt{|g|}\left\{R[g]-\delta_{ab}g^{MN}\partial_M\varphi^a\partial_N\varphi^b-\sum_{s\in\mathcal{S}}\frac{\eta_s}{n_s!}e^{2\lambda_{sa}\varphi^a}F_s^2\right\}, \tag{1}$$

in a (pseudo-)Riemannian manifold $\mathcal{M}=\mathbb{R}_u\times\mathcal{M}_0\times\cdots\times\mathcal{M}_n$ with the metric

$$ds^2=g_{MN}dz^M dz^N=we^{2\alpha(u)}du^2+\sum_{i=0}^{n} e^{2\beta^i(u)}ds_i^2, \quad w=\pm 1, \tag{2}$$

where u is a selected coordinate ranging in $\mathbb{R}_u\subseteq\mathbb{R}$; $g^i=ds_i^2$ are metrics on d_i-dimensional factor spaces $\mathcal{M}_i$ of arbitrary signatures $\varepsilon_i=\text{sign } g^i$; $|g|=|\det g_{MN}|$ and similarly for subspaces; $F_s^2 =F_{s,M_1\ldots M_{n_s}}F_s^{M_1\ldots M_{n_s}}$; λ_{sa} are coupling constants; $\eta_s=\pm 1$ (to be specified later); $s\in\mathcal{S}$, $a\in\mathcal{A}$, where $\mathcal{S}$ and $\mathcal{A}$ are some finite sets. All $\mathcal{M}_i, i>0$ are assumed to be Ricci-flat, while $\mathcal{M}_0$ is allowed to be a space of constant curvature $K_0=0, \pm 1$.

In the one-variable sector, $\varphi^a=\varphi^a(u)$. The set of indices $\mathcal{S}=\{s\}$ in (1) will be used to jointly describe essentially u-dependent electric (F_{eI}) and magnetic (F_{mI}) F-forms, to be associated with different subsets $I=\{i_1,\cdots,i_k\}$ $(i_1<\ldots<i_k)$ of the set of numbers labelling the factor spaces: $\{i\}=I_0=\{0,\ldots,n\}$. Thus one can write

$$\mathcal{S}=\{s\}=\{eI_s\}\cup\{mI_s\}. \tag{3}$$

A given F-form may have several essentially (nonpermutatively) different components, both electric and magnetic; such a situation is sometimes called ''composite p-branes.''[18,23]

For convenience, we will nevertheless treat essentially different components of the same F-form as individual (''elementary'') F-forms. A subsequent reformulation to the composite ansatz is straightforward.

So, by construction, nonzero components of F_{eI} carry coordinate indices of u and the subspaces $\mathcal{M}_i, i\in I$, those of F_{mI} — the indices of $\mathcal{M}_i, i\in \overline{I} \stackrel{\rm def}{=} I_0\backslash I$ since a magnetic form is built as a form dual to a possible electric one. Therefore

$$n_{eI}=\text{rank}F_{eI}=d(I)+1, \quad n_{mI}=\text{rank}F_{mI}=D-\text{rank}F_{eI}=d(\overline{I}), \tag{4}$$

where $d(I)=\Sigma_{i\in I}d_i$ are the dimensions of the subspaces $\mathcal{M}_I=\mathcal{M}_{i_1}\times\cdots\times\mathcal{M}_{i_k}$.

Several electric and/or magnetic forms (with maybe different coupling constants λ_{sa}) can be associated with the same I and are then labeled by different values of s. (The index s by I is, however, sometimes omitted when this cannot cause confusion.)

This problem setting covers various classes of models: isotropic and anisotropic cosmologies where u is a timelike coordinate and $w=-1$; static models with various spatial symmetries (spherical, planar, pseudospherical, cylindrical, toroidal), where u is a spatial coordinate, $w=+1$, and time is selected among $\mathcal{M}_i$; and Euclidean models with similar symmetries or models with a Euclidean ''external'' space-time, where also $w=+1$.

A simple analysis shows that a positive energy density $-T^t_t$ of the fields F_s is achieved in all Lorentzian models with the signature $(-++\cdots+)$ if one chooses in (1), as usual, $\eta_s=1$ for all s. In more general models, with arbitrary ε_i, the requirement $-T^t_t>0$ is fulfilled if

$$\eta_{eI}=-\varepsilon(I)\varepsilon_t(I), \quad \eta_{mI}=-\varepsilon(\overline{I})\varepsilon_t(\overline{I}) \tag{5}$$

$$\varepsilon(I)\stackrel{\rm def}{=}\prod_{i\in I}\varepsilon_i, \quad \varepsilon_t(I)=\begin{cases}1, & \mathbb{R}_t\subset\mathcal{M}_I,\\ -1 & \text{otherwise},\end{cases} \tag{6}$$

where $\mathbb{R}_t$ is the time axis. If $\varepsilon_t(I)=1$, we are dealing with a genuine electric or magnetic field, while otherwise the F-form behaves as an effective scalar or pseudoscalar in the physical subspace. The latter happens, in particular, in isotropic cosmologies and their Euclidean counterparts where the time coordinate is u and $\mathbb{R}_t=\mathbb{R}_u$, unrelated to any subset I. F-forms with $\varepsilon_t(I)=-1$ will be called *quasiscalar.*

Example: Consider a spherically symmetric configuration, with $D=6$, $\mathcal{M}=\mathbb{R}_0\times\mathbb{R}_1\times S^2\times\mathbb{R}_4\times\mathbb{R}_5$, where the coordinate indices 0, 1, 4, 5, refer to time, radius and two extra dimensions, 2 and 3 to the spherical angles, respectively; thus $\mathbb{R}_0=\mathbb{R}_t$ and $\mathbb{R}_1=\mathbb{R}_u$. Then, for rank $F=3$, the component F_{015} is electric, $I\mapsto(0,5)$; F_{234} is magnetic, $I\mapsto(0,5)$; F_{145} is electric quasiscalar, $I\mapsto(4,5)$; F_{023} is magnetic quasiscalar, $I\mapsto(4,5)$, where the figures in parentheses are coordinate indices of the respective subspaces $\mathcal{M}_I$.

Let us now, as in Ref. 24 and many later papers, choose the harmonic u coordinate $(\nabla^M\nabla_M u=0)$, such that

$$\alpha(u)=\sum_{i=0}^{n} d_i\beta^i\equiv d_0\beta^0+\sigma_1(u), \quad \sigma_1(u)\stackrel{\rm def}{=}\sum_{i=1}^{n} d_i\beta^i. \tag{7}$$

The Maxwell-type equations due to (1) for the F-forms are easily integrated, giving

$$F_{eI}^{uM_1\ldots M_{d(I)}}=Q_{eI}e^{-2\alpha-2\overline{\lambda}_{eI}\overline{\varphi}}\varepsilon^{M_1\ldots M_{d(I)}}/\sqrt{|g_I|}, \quad Q_{eI}=\text{const}, \tag{8}$$

$$F_{mI,M_1\ldots M_{d(\overline{I})}}=Q_{mI}\varepsilon_{M_1\ldots M_{d(\overline{I})}}\sqrt{|g_{\overline{I}}|}, \quad Q_{mI}=\text{const}, \tag{9}$$

where $|g_I|=\Pi_{i\in I}|g^i|$, Q_s are charges and overbars replace summing in a. In what follows we will restrict the set $\mathcal{S}=\{s\}$ to such s that the charges $Q_s\neq 0$.

Consequently, at the r.h.s. of the Einstein equations due to (1), $R^N_M-\frac{1}{2}\delta^N_M R=T^N_M$, the energy-momentum tensor (EMT) T^N_M takes the form

$$e^{2\alpha}T_M^N=-\frac{w}{2}\sum_s \epsilon_s Q_s^2 e^{2\sigma(I)-2\chi_s\overline{\lambda}_s\overline{\varphi}}\mathrm{diag}(+1,[1]_I,[-1]_{\overline{I}})+\frac{w}{2}(\dot{\varphi}^a)^2\mathrm{diag}(+1,[-1]_{I_0}), \quad (10)$$

where the first place on the diagonal belongs to u and the symbol $[f]_J$ means that the quantity f is repeated along the diagonal for all indices referring to $\mathcal{M}_j, j\in J$; $\sigma(I)\overset{\mathrm{def}}{=}\Sigma_{i\in I}d_i\beta^i$; the sign factors ϵ_s and χ_s are

$$\epsilon_{eI}=-\eta_{eI}\varepsilon(I), \quad \epsilon_{mI}=w\,\eta_{mI}\varepsilon(\overline{I}); \quad \chi_{eI}=+1, \quad \chi_{mI}=-1, \tag{11}$$

so that χ_s distinguishes electric and magnetic forms.

Let us suppose, as is usually (and reasonably) done in p-brane studies, that *neither of I_s such that $Q_s\neq 0$ contains the index 0*, that is, neither of the branes ''lives'' in the subspace $\mathcal{M}_0$, interpreted as the external space or its subspace. (This means that, e.g., in the spherically symmetric case there is no electric or magnetic field along a coordinate sphere $\mathcal{M}_0=S^{d_0}$.) Then each constituent EMT and hence the total EMT possess the property $T_u^u+T_z^z=0$ if z belongs to M_0. As a result, the corresponding combination of the Einstein equations has a Liouville form and is integrated:

$$\begin{gathered}\ddot{\alpha}-\ddot{\beta}^0=wK_0(d_0-1)^2e^{2\alpha-2\beta^0}=0,\\ e^{\beta^0-\alpha}=(d_0-1)S(wK_0,k,u),\end{gathered}\tag{12}$$

where k is an integration constant (IC) and we have introduced the notation

$$\begin{gathered}S(1,h,t)=\begin{cases}h^{-1}\sinh ht, & h>0,\\ t, & h=0,\\ h^{-1}\sin ht, & h<0;\end{cases}\\ S(-1,h,t)=h^{-1}\cosh ht, \quad h>0;\\ S(0,h,t)=e^{ht}, \quad h\in\mathbb{R}.\end{gathered}\tag{13}$$

Another IC is suppressed by properly choosing the origin of the u coordinate.

With (12) the D-dimensional line element may be written in the form ($\overline{d}\overset{\mathrm{def}}{=}d_0-1$)

$$ds^2=\frac{e^{-2\sigma_1/\overline{d}}}{[\overline{d}S(wK_0,k,u)]^{2/d}}\left[\frac{w\,du^2}{[\overline{d}S(wK_0,k,u)]^2}+ds_0^2\right]+\sum_{i=1}^n e^{2\beta^i}ds_i^2. \tag{14}$$

Let us treat the remaining set of unknowns $\beta^i(u),\varphi^a(u)$ as a real-valued vector function $x^A(u)$ (so that $\{A\}=\{1,\ldots,n\}\cup\mathcal{A}$) in an $(n+|\mathcal{A}|)$-dimensional vector space $\mathcal{V}$ (target space). The field equations for β^i and φ^a can be derived from the Toda-type Lagrangian

$$\begin{gathered}L=G_{AB}\dot{x}^A\dot{x}^B-V_Q(y)\equiv\sum_{i=1}^n(\dot{\beta}^i)^2+\frac{\dot{\sigma}_1^2}{d_0-1}+\delta_{ab}\dot{\varphi}^a\dot{\varphi}^b-V_Q(y),\\ V_Q(y)=-\sum_s \epsilon_s Q_s^2 e^{2y_s},\end{gathered}\tag{15}$$

with the ''energy'' constraint

TABLE I. Sign factors wK_0 and ϵ_s for different kinds of models.

		Cosmology $w=-1$	Static spaces $w=+1$	Euclidean $w=+1$
	wK_0	$-K_0$	K_0	K_0
	Electric	none	$+1$	none
ϵ_s	Magnetic	none	$+1$	none
	Electric quasiscalar	-1	-1	-1
	Magnetic quasiscalar	-1	-1	$+1$

$$E=G_{AB}\dot{x}^A\dot{x}^B+V_Q(y)=\frac{d_0}{d_0-1}K, \quad K=\begin{cases} k^2\text{sign } k, & wK_0=1; \\ k^2, & wK_0=0,-1, \end{cases} \tag{16}$$

where the IC k has appeared in (12). The nondegenerate symmetric matrix

$$(G_{AB})=\begin{pmatrix} d_i d_j/\bar{d}+d_i\delta_{ij} & 0 \\ 0 & \delta_{ab} \end{pmatrix} \tag{17}$$

defines a positive-definite metric in $\mathcal{V}$; the functions $y_s(u)$ are defined as scalar products:

$$y_s=\sigma(I_s)-\chi_s\overline{\lambda_s}\overline{\varphi}\equiv Y_{s,A}x^A, \quad (Y_{s,A})=(d_i\delta_{iI_s}, \quad -\chi_s\lambda_{sa}), \tag{18}$$

where $\delta_{iI}=1$ if $i\in I$ and $\delta_{iI}=0$ otherwise. The contravariant components and scalar products of the vectors $\mathbf{Y}_s$ are found using the matrix G^{AB} inverse to G_{AB}:

$$(G^{AB})=\begin{pmatrix} \delta^{ij}/d_i-1/\bar{D} & 0 \\ 0 & \delta^{ab} \end{pmatrix}, \quad (Y_s{}^A)=\left(\delta_{iI}-\frac{d(I)}{\bar{D}}, \quad -\chi_s\lambda_{sa}\right); \tag{19}$$

$$Y_{s,A}Y_{s'}{}^A\equiv\mathbf{Y}_s\mathbf{Y}_{s'}=d(I_s\cap I_{s'})-\frac{d(I_s)d(I_{s'})}{\bar{D}}+\chi_s\chi_{s'}\overline{\lambda_s}\overline{\lambda_{s'}}, \quad \bar{D}=D-2. \tag{20}$$

The equations of motion in terms of $\mathbf{Y}_s$ read

$$\ddot{x}^A=\sum_s q_s Y_s{}^A e^{2y_s}, \quad q_s\overset{\text{def}}{=}\epsilon_s Q_s^2. \tag{21}$$

III. GENERAL PROPERTIES OF THE BRANE SYSTEMS

The positive energy requirement (5) that fixes the input signs η_s, can be written as follows for Lorentzian models using the notations (11):

$$\epsilon_s=\varepsilon_t(I_s). \tag{22}$$

The corresponding requirement for Euclidean models is obtained by applying the conventional Wick rotation to Lorentzian cosmologies. This rotation of the time t changes w but preserves all η_s as well as $\varepsilon(I)$ since $\mathbb{R}_t\not\subset\mathcal{M}_I$, $\forall I$. Then by (11), ϵ_{eI} remain invariable while ϵ_{mI} change. This distinction between electric and magnetic forms is also connected with the property of the duality transformation to change the sign of the EMT in Euclidean models.[25,26]

Table I shows the sign factors wK_0 and $\epsilon_s=\text{sign } q_s$ for F-forms in different classes of models under the above positive energy requirement.

In what follows, we restrict ourselves to the model described in Sec. II with the sign factors specified in Table I, unless specially indicated.

One general statement, to be taken into account in the subsequent proofs, can be formulated as a lemma:

Lemma 1: At any regular point of the space-time, for all $a \in \mathcal{A}$ *and* $s \in \mathcal{S}$,

$$e^{-2\alpha}(\dot{\varphi}^a)^2 < \infty, \quad e^{-2\alpha + 2y_s} < \infty. \tag{23}$$

Indeed, regularity implies finite values of all curvature invariants, including R and $R_M^N R_N^M$; by virtue of the Einstein equations, one must have $T_M^N T_N^M < \infty$. Since T_M^N has a block-diagonal structure, the latter invariant can be written as a sum of squares, where each summand must thus be finite, including $(T_t^t)^2$. The component T_t^t is in turn, due to (22), a sum of negative-definite terms, corresponding to scalar fields φ^a and F-forms F_s. Therefore every such term must be finite, leading to (23).

A. Isotropic cosmology

Table I shows that in isotropic cosmologies, when u is a time coordinate and $\mathcal{M}_0$ is identified with the physical space (conventionally $d_0 = 3$), $\epsilon_s = -1$: there are only quasiscalar forms since a true electric or magnetic field would violate the spatial isotropy.

The logarithm of the extra-dimension volume factor, σ_1, by virtue of (21) obeys the equation

$$\ddot{\sigma}_1 = -\frac{d_0 - 1}{D-2} \sum_s d(I_s) Q_s^2 e^{2y_s}, \tag{24}$$

whence $\ddot{\sigma}_1 < 0$. So this volume factor cannot have a minimum and, moreover, if it tends to a finite value $e^{\sigma_{10}}$ as $u \to \pm\infty$, at other values of u it is smaller than $e^{\sigma_{10}}$. This feature is unfavorable for obtaining models with the so-called dynamical compactification, where the size of extra dimensions decreases to microscopic scales in the course of the evolution.

Next, due to $\epsilon_s = -1$, both terms in the expression (16) for E are positive-definite, so that nontrivial solutions correspond to $k > 0$. The range of u is $\mathbb{R}$ for $K_0 = 0, +1$ and (without loss of generality) $u > 0$ for $K_0 = -1$. By (13) and (14), the model asymptotics are characterized as follows.

For any K_0, at the asymptotic $u \to +\infty$ the total volume factor $e^{d_0\beta^0 + \sigma_1}$ (which, by (7), coincides with e^{α}) tends to zero. Although separately the physical scale factor $a(u) = e^{\beta_0}$ and the ''internal'' one, e^{σ_1}, may have various limits, the behavior of e^{α} indicates that this asymptotic is singular. Moreover, since asymptotically $\alpha \sim cu$, $c = \text{const} < 0$, the proper time $t = \int e^{\alpha} du < \infty$: the singularity occurs at a finite proper time.

For $K_0 = +1$ the other asymptotic $u \to -\infty$ is like the one just described, due to the symmetry of the function cosh ku in (13). Thus closed models evolve in a finite proper time interval between two singularities where the total volume of the Universe tends to zero.

For $K_0 = 0$, the asymptotic $u \to -\infty$ corresponds to an infinitely growing total volume factor e^{α} while the proper time t is also infinite. In the special case when $\sigma_1 \to \sigma_{10} = \text{const}$, the physical scale factor a obeys the law $a \sim |t|^{1/d_0}$.

For $K_0 = -1$ the second asymptotic is $u \to 0$, and this is a regular point of the equations of motion (21) determining x^A. So the metric behavior is (now in the general case) governed by the function $S(1,k,u) \approx u$ in (14), while all e^{β^i}, $i > 0$ and consequently σ_1 tend to finite limits. As $u \to 0$,

$$e^{\alpha} \sim u^{-1-1/\bar{d}}, \quad |t| \sim u^{-1/\bar{d}} \to \infty, \quad a(t) = e^{\beta^0} \sim |t|, \tag{25}$$

corresponding to linear expansion or contraction of the physical space.

Finally, Eq. (16) with $V_Q > 0$ implies that all $\dot{x}^A$ are bounded above, hence $x^A(u)$ are finite for all finite u and cannot create a singularity. Therefore the above description of the asymptotics is quite general and applies to all isotropic cosmologies in the field model under consideration.

It should be noted, however, that this discussion concerns the model behavior in the D-dimensional Einstein conformal frame, in which the action (1) was postulated. See further comments in Sec. VI. One can add that after reduction to $1+d_0$ physical dimensions, one obtains cosmology with a set of interacting massless scalar fields, so that the $(1+d_0)$-metric can be written explicitly: in the $(1+d_0)$-dimensional Einstein frame it corresponds to the well-known solutions with ultrastiff matter.

B. Static spherical symmetry: General observations

In static, spherically symmetric models, where u is a radial coordinate, $w=+1$, $\mathcal{M}_0=S^{d_0}$, $K_0=+1$, among other $\mathcal{M}_i$ there should be a one-dimensional subspace, say, $\mathcal{M}_1$, which may be identified with time: $\varepsilon_1=-1$. The sign factor wK_0 in (12) is $+1$, while ϵ_s is, due to (22), $+1$ for normal electric and magnetic forms F_I and -1 for quasiscalar ones.

By construction, see Eqs. (13), (14), spatial infinity corresponds to $u=0$ (where the usual ''area function'' $e^{\beta^0}\sim u^{1/\bar{d}}$) and, without loss of generality, the range of u is

$$0<u<u_{\max}, \tag{26}$$

where $u_{\max}$ is either $+\infty$, or the smallest value of u where the fields lose regularity.

The experience of dealing with particular models belonging to the class (1) indicates that a generic spherically symmetric solution exhibits a naked singularity. Possible exceptions can be (i) black holes (BHs), (ii) wormholes (WHs) or wormhole-like objects with a neck and a second nonsingular asymptotic, (iii) configurations with a regular center (a soliton-like object, which might be expected for an interacting field system) and, finally, (iv) a situation where the coordinate patch we use is incomplete, terminates at a regular sphere $u=u_{\max}$ (which may be even infinitely remote in our static frame of reference), and a possible continuation may reveal either a singularity, or one of the opportunities (i)–(iii).

One can show, however, that for our model only the BH opportunity is viable. Lorentzian WHs do not exist according to Ref. 20 (see also Sec. V), while variants (iii) and (iv) are ruled out by the following theorem:

Theorem 1: *The model specified in Sec. II does not admit solutions describing a static, spherically symmetric configuration (a) with a regular center or (b) where $u=u_{\max}$ corresponds to a regular surface such that $\mathcal{M}_0$ is a sphere of finite radius.*

Proof: (a) A regular center implies local flatness of the metric at some $u=u^*$, where $e^{\beta^0}=0$, while other β^i remain finite. One easily shows that with Ref. 14 it may happen only when $k=0$, $u^*=u_{\max}=\infty$ (otherwise the correct radius-to-circumference ratio for small circles around the center cannot be achieved).Then due to (7), since $|\sigma_1|<\infty$,

$$e^{\beta^0}\sim u^{1/(d_0-1)}, \quad e^{\alpha}\sim u^{-d_0/(d_0-1)} \quad \text{as} \quad u\to\infty. \tag{27}$$

On the other hand, the EMT regularity requirement[27] (see Lemma 1) leads to $|\varphi|<\infty$ as $u\to\infty$. Therefore at such a center the F-forms behave like free fields exhibiting (see (8)–(10)) a singularity, with infinite values of the EMT invariants. Item (a) is proved.

The assumption **(b)** means that both β_0 and σ_1 are finite at $u=u_{\max}$. This cannot happen at $u_{\max}<\infty$ since there would be no reason to stop at this value of u; and at $u_{\max}=\infty$ this means that $S(1,k,\infty)<\infty$, contrary to the definition (13).

C. Black holes: No-hair and single-time theorems

We see that the only positive-energy Lorentzian spherically symmetric configurations without naked singularities are BHs. BH solutions of various models belonging to the class (1) have been studied in numerous recent papers (see Refs. 3, 15, and 18, and references therein). However, exact solutions have been (and probably can be) only obtained for a small subset of the whole set of models (1), whereas some general properties of BH solutions may be discovered without solving the equations. Two such properties, having the form of restrictions generalizing the previously observed properties of specific solutions,[15,20] are proved here.

In what follows, we will call a *horizon* a nonsingular surface $u=u_*$ in $\mathcal{M}$ where some scale factors $e^{\beta^i}=0$ (corresponding to possibly multiple time coordinates), while other β^i remain finite. A *BH solution* is a static, spherically symmetric solution containing a horizon. These working definitions, though incomplete, are sufficient for our purposes.

An immediate observation is

Lemma 2: BH solutions can only exist for $k\geqslant 0$ *and the horizon is then at* $u=\infty$.

Indeed, at a horizon, the function σ_1 defined in (7) tends to $-\infty$ along with a part of its constituents, another part remaining finite. According to (14), to obtain a finite value of β^0, one has then to require that $S(1,k,u_*)=+\infty$, which by (13) is only possible when $k\geqslant 0$ and $u_*=\infty$.

Another result applies to BHs in manifolds $\mathcal{M}$ with several time coordinates, as suggested in some recent unification models (see Refs. 5, 28, and references therein). If there is another time coordinate, some branes can evolve with it. The following theorem shows, however, that in our framework, even in a space-time with multiple times, a BH can only exist with its unique preferred, physical time, while other times are not distinguished by the metric behavior from extra spatial coordinates.

Theorem 2 (Single-Time Theorem): *Any* BH *solution with* $k>0$ *contains precisely one coordinate* t *such that* $g_{tt}=0$ *at the horizon.*

Proof: Suppose that $u=\infty$ is a horizon where some $e^{\beta^i}\to 0$, $i\in I_t\subseteq(I_0\backslash 0)$. As follows from (7), at the asymptotic $u\to\infty$ one has $\alpha\to-\infty$ and, moreover, the finiteness of β^0 means (see (14)) that $\alpha\sim -ku$. On the other hand, the condition (24) holds only if for all F-forms, at most,

$$e^{2y_s}=O(e^{-2ku}). \tag{28}$$

The equations of motion (21) then show that, as $u\to\infty$,

$$\dot{x}^A=-c^A+o(1), \quad c^A=\text{const}, \tag{29}$$

where $c^i>0$ for $i\in I_t$ and $c^i=0$ for other A.

In the constraint (16), the potential $V_Q(u)\underset{u\to\infty}{\longrightarrow}0$ due to (29), therefore

$$G_{AB}c^Ac^B=\frac{d_0}{d_0-1}. \tag{30}$$

The asymptotic of α and the condition (7) show that, simultaneously,

$$\sum_{i\in I_t} d_ic^i=k, \tag{31}$$

so that $c_i\leqslant k$. From (30) with (31) and (17) it follows

$$\sum_{i\in I_t} d_ic^{i2}=k^2. \tag{32}$$

Combined, Eqs. (31) and (32) lead to

$$\sum_{i\in I_t} d_ic^i(k-c^i)=0, \tag{33}$$

which is compatible with (32) for $0\leqslant c^i\leqslant k$ only when the sum consists of one term, to be labeled $i=1$, such that $d_1=1$ and $c^1=k$. This proves the theorem.

One more theorem shows that BH solutions can contain only true electromagnetic F-forms rather than quasiscalar ones.

Theorem 3 (No-Hair Theorem): *All F-forms in a* BH *solution with $k>0$ possess the property $\delta_{1I_s}=1$, where the number $i=1$ refers to the time axis.*

The proof rests on Lemma 1, which, applied to F-forms, leads again to (28). Now, according to Theorem 2, at a horizon $(u\to\infty)$ only $\beta^1\to-\infty$, while other β^i are finite. As is directly verified, (28) holds in the case $\delta_{1I_s}=1$ (for true electromagnetic forms), while for quasiscalar ones one has finite limits for e^{y_s}, leading to infiniteness of the corresponding EMT constituent.

Remark 1: The regularity of the scalar fields, $x^A=\varphi^a$, at $u\to\infty$ was not required in the conditions of Theorems 2 and 3; for $k>0$ it follows from (23). Under the additional requirement $\varphi^a<\infty$ as $u\to\infty$, Theorem 3 is easily proved for $k=0$ as well.

Remark 2: For $k=0$ we have no Theorem 2; moreover, Theorem 3 is not proved for $k=0$ without assuming $\varphi^a<\infty$. Nevertheless, for BH solutions with $k=0$ obtainable as a limit of ones with $k>0$, the statements of both theorems remain valid. (For known exact BH solutions, $k=0$ corresponds to the extreme limit of minimal mass for given charges.) Meanwhile, the existence of exceptional BH solutions with $k=0$, nonzero quasiscalar forms and/or multi-time horizons is not ruled out by our study; such solutions may perhaps exist with infinite limits of scalar fields that balance the infinity of $e^{-\alpha}$ in the EMT of F-forms.

Remark 3: If there are BH solutions, there are also others, where the scale factor showing a zero value is associated, instead of physical time, with one of the extra coordinates (such solutions are obtained from BH ones by simple re-denoting). One thus finds the so-called T-holes, where crossing a horizon leads to changing the signature of the external, physical space from $(-+++\cdots)$ to $(--++\cdots)$. Possible properties of such objects are discussed in more detail elsewhere[17,29] within the frames of dilaton gravity, but the considerations thereof are valid as well for the more general model (1). Theorems 2 and 3 are valid for T-holes after proper reformulation.

IV. SOME EXACT SOLUTIONS

A. Orthogonal systems (OS)

The field equations are entirely integrated if all $\mathbf{Y}_s$ are mutually orthogonal in $\mathcal{V}$, that is,

$$\mathbf{Y}_s\mathbf{Y}_{s'}=\delta_{ss'}/N_s^2, \quad 1/N_s^2=d(I)[1-d(I)/\overline{D}]+\overline{\lambda_s^2}>0. \tag{34}$$

Then the functions $y_s(u)$ obey the decoupled Liouville equations $\ddot{y}_s=b_s e^{2y_s}$, with $b_s\overset{\text{def}}{=}\epsilon_s Q_s^2/N_s^2$, whence

$$e^{-y_s(u)}=\sqrt{|b_s|}S(\epsilon_s,h_s,u+u_s), \tag{35}$$

where h_s and u_s are ICs and the function $S(.,.,.)$ has been defined in (13). For the sought functions $x^A(u)=(\beta^i,\varphi^a)$ we then obtain:

$$x^A(u)=\sum_s N_s^2 Y_s{}^A y_s(u)+c^A u+\underline{c}^A, \tag{36}$$

where the vectors of ICs $\mathbf{c}$ and $\underline{\mathbf{c}}$ are orthogonal to all Y_s: $c^A Y_{s,A}=\underline{c}^A Y_{s,A}=0$, or

$$c^i d_i\delta_{iI_s}-c^a\chi_s\lambda_{sa}=0, \quad \underline{c}^i d_i\delta_{iI_s}-\underline{c}^a\chi_s\lambda_{sa}=0. \tag{37}$$

The solution is general for the properly chosen input parameters; the number of independent charges equals the number of F-forms.

B. Block-orthogonal systems (BOS)

Suppose now[20] that the set $\mathcal{S}$ splits into several nonintersecting nonempty subsets,

$$\mathcal{S}=\bigcup_{\omega}\mathcal{S}_\omega, \quad |\mathcal{S}_\omega|=m(\omega), \tag{38}$$

such that the vectors $\mathbf{Y}_{\mu(\omega)}$ $(\mu(\omega)\in S_\omega)$ form mutually orthogonal subspaces $\mathcal{V}_\omega$ in $\mathcal{V}$:

$$\mathbf{Y}_{\mu(\omega)}\mathbf{Y}_{\nu(\omega')}=0, \quad \omega\neq\omega'. \tag{39}$$

Suppose, further, that, for each fixed ω, all $\mathbf{Y}_\nu$ (where $\nu=\nu(\omega)$) are linearly independent and the charge factors $q_\nu=\epsilon_\nu Q_\nu^2\neq 0$ satisfy the set of linear algebraic equations

$$(\mathbf{Y}_\nu-\mathbf{Y}_{\nu'})\mathbf{Z}_\omega=0, \quad \mathbf{Z}_\omega\stackrel{\text{def}}{=}\sum_{\mu\in\mathcal{S}_\omega} q_\mu\mathbf{Y}_\mu, \tag{40}$$

for each pair (ν,ν'). Then the function $y_\omega(u)\stackrel{\text{def}}{=}Y_{\mu(\omega),A}x^A$ is the same for all $\mu\in\mathcal{S}_\omega$ and satisfies the Liouville equation $\ddot{y}_\omega=b_\omega e^{2y_\omega}$. As a result, we obtain a solution to the equations of motion, generalizing (35), (36):

$$e^{-y_\omega}=\sqrt{|b_\omega|}S(\text{sign } b_\omega, h_\omega, u+u_\omega), \tag{41}$$

$$x^A=\sum_\omega N_\omega^2 Y_\omega{}^A y_\omega(u)+c^A u+\underline{c}^A, \tag{42}$$

where h_ω and u_ω are ICs; the constants c^A and $\underline{c}^A$ satisfy the same orthogonality relations (38) as for OS, that is, the vectors $\mathbf{c}$ and $\underline{\mathbf{c}}$ are orthogonal to each individual $\mathbf{Y}_s$, even if it is a member of a BOS subsystem. We have used the notations

$$b_\omega=\mathbf{Y}_{\nu(\omega)}\mathbf{Z}_\omega; \quad \mathbf{Y}_\omega=\mathbf{Z}_\omega\hat{q}_\omega; \quad N_\omega^{-2}=\mathbf{Y}_\omega^2=\frac{b_\omega}{\hat{q}_\omega}; \quad \hat{q}_\omega=\sum_{\mu\in\mathcal{S}_\omega} q_\mu. \tag{43}$$

Here b_ω is nonzero and independent of $\nu(\omega)\in S_\omega$ due to (40); moreover, $\hat{q}_\omega\neq 0$ since $\hat{q}_\omega=\mathbf{Z}_\omega^2/b_\omega$ while the nonzero vector $\mathbf{Z}_\omega$ is determined up to extension by (40).

The linear independence of $\mathbf{Y}_{\mu(\omega)}$ thus guarantees that Eqs. (40) yield $q_{\mu(\omega)}$ for a given subsystem up to a common factor. Therefore, unlike the OS solution, the BOS one is special: the number of independent charges coincides with $|\{\omega\}|$, the number of subsystems; however, we thus gain exact solutions for more general sets of input parameters, e.g., a one-charge solution can be obtained for actually an arbitrary configuration of branes with linearly independent $\mathbf{Y}_\mu$ (except possible cases when the solution of (40) leads to at least one zero charge).

When $m=1$, we have a single vector $\mathbf{Y}_\omega=\mathbf{Y}_s$ orthogonal to all others, with the norm $N_\omega^{-2}=N_s^{-2}$, and the charge factor is $b_\omega=b_s$. Thus single branes and BOS subsystems are represented in a unified way, and the OS solution is a special case $(m(\omega)=1,\forall\omega)$ of the BOS one.

The metric has the form (14), where the function σ_1 is

$$\sigma_1=-\frac{d_0-1}{D-2}\sum_\omega N_\omega^2 y_\omega(u)\sum_{\mu\in\mathcal{S}_\omega}\frac{q_\mu}{\hat{q}_\omega}d(I_\mu)+u\sum_{i=1}^n c^i+\sum_{i=1}^n \underline{c}^i. \tag{44}$$

For OS $(\omega\mapsto s)$ the sum in μ reduces to $d(I_s)$. The "conserved energy" (16) is

$$E=\sum_\omega N_\omega^2 h_\omega^2 \text{sign } h_\omega+c_A c^A=\frac{d_0}{d_0-1}K. \tag{45}$$

In the special case $m=2$, $\mathbf{Y}_1^2=\mathbf{Y}_2^2$, one easily obtains $b_1=b_2$, as was shown in Ref. 13 for a single F-form. By definition of b_s that means not only $Q_1^2=Q_2^2$, but also a coincidence of the sign factors $\operatorname{sign} b_s=\epsilon_s$. For instance, in spherical symmetry, the F-fields must be either both true electric/magnetic ones ($\operatorname{sign} b_s=1$), or both quasiscalar ones ($\operatorname{sign} b_s=-1$).

C. On cosmological and black-hole solutions

There is a large number of exact cosmological solutions to special cases of the model (1), see Refs. 12, 21, 30, and references therein. It can be seen that the description of Sec. III A (which is certainly confirmed by exact solutions) actually exhausts all general features of the model, since other details, such as, e.g., the particular behavior of the physical scale factor $a(t)$, depend on the choice of integration constants.

BHs are obtained as special spherically symmetric solutions when $h_\omega>0$, $u_{\max}=\infty$. The functions β^i $(i=0,2,\ldots,n)$ and φ^a remain finite as $u\to\infty$ under the following constraints on the ICs:

$$h_\omega=k, \quad \forall\omega; \quad c^A=k\sum_\omega N_\omega^2 Y_\omega{}^A-k\delta^A_1, \tag{46}$$

where $A=1$ corresponds to $i=1$ (time), $d_1=1$ (according to Theorem 2). The constraint (45) then holds automatically.

The subfamily (46) exhausts all BH solutions under OS or BOS assumptions, except the extreme case $k=0$; extreme BHs are obtained by subsequently passing to the limit $k\to 0$. One can notice that exceptional extreme BH solutions, whose possibility was mentioned in Sec. III C, are not found in this way.

General explicit forms of the OS and BOS BH solutions have been presented in Refs. 15 and 20, respectively. The BH properties stated in Theorems 2 and 3 are confirmed for the OS and BOS solutions and, moreover, have been first observed[15,20] for these solutions.

V. WORMHOLES

A. Wormhole existence conditions

Wormhole-like configurations which can appear as special OS or BOS solutions, have an infinite ''external radius'' $e^{\beta^0(u)}$ at both ends $u_\pm$ of the u range and are regular between them; all $\beta^i(u_\pm)$ $(i>0)$ and $\varphi^a(u_\pm)$ are finite. This happens when $k<0$ and the solution behavior is governed by the function $\sin ku$ (so that $u_-=0$ and $u_+=\pi/|k|$) and is possible if the first positive zero of the function $\sin[|h_s|(u-u_s)]$ is greater than $\pi/|k|$ for any s such that $h_s<0$ (see Fig. 1).

In the cosmological setting, this behavior would correspond to nonsingular, bouncing models, which are, however, absent according to Sec. III A (due to $k>0$). The static and Euclidean cases are not *a priori* excluded.

As is evident from Fig. 1, any WH solution is characterized by $|k|>|h_\omega|$ for all h_ω which are negative. Due to (45), for $k<0$ at least some h_ω should be negative as well. Furthermore, for $k<0$ and $h_\omega<0$ it is necessary to have $wK_0=1$ and $b_\omega>0$, respectively.

Table I shows that WHs can exist in static or Euclidean models only with spherical rather than pseudospherical or planar symmetry. In cosmology we have no fields capable to give negative h_s or h_ω, which again confirms the absence of nonsingular ''bounced'' models. In static spherical symmetry the necessary F-forms are true electric and magnetic ones. In Euclidean models, magnetic quasiscalar forms are needed.

Suppose $k<0$. Since in (45) $\mathbf{c}^2\geqslant 0$, the requirement $|k|>|h_s|$ means that

$$\sum_{\{\omega:\, h_\omega<0\}} N_\omega^2>\frac{d_0}{d_0-1}. \tag{47}$$

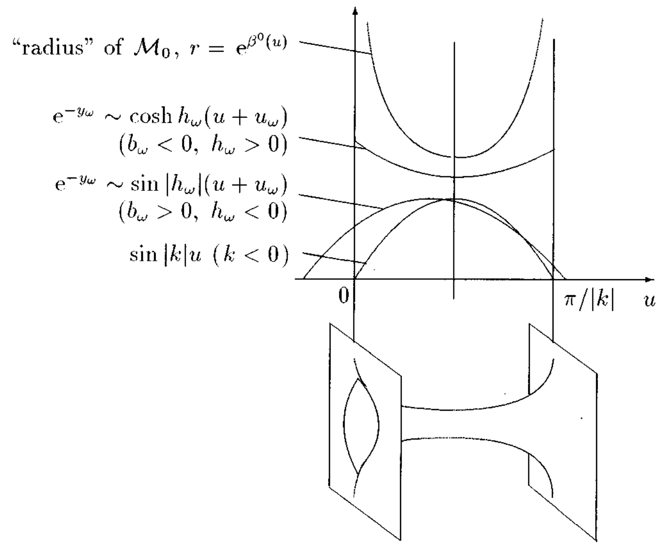

FIG. 1. A wormhole configuration; qualitative picture.

This inequality is not only *necessary*, but also *sufficient* for the existence of WHs with given input parameters: d_i and the vectors $\mathbf{Y}_s$. Indeed, put $c^A=0$ and turn to zero the charges $Q_{\mu(\omega)}$ in all subsystems with $\hat{q}_\omega<0$ (note that, by (43), sign $\hat{q}_\omega=\text{sign}b_\omega$.) Choose all h_ω to be negative and equal, then due to (47) $|h_\omega|<|k|$. It is now an easy matter to choose the ICs u_ω in such a way that $\sin[|h_\omega|(u+u_\omega)]>0$ on the whole segment $[0,\pi/|k|]$, and this results in a WH solution.

B. Lorentzian wormholes and a universal restriction for brane systems

In general relativity static[31] and even dynamic[32] traversable WHs are known to violate the null energy condition. It can be verified[20] that, under the present positive energy requirement, the model (1) after reduction to d_0+2 dimensions by integrating out all $\mathcal{M}_i, i>1$ and a transition to the Einstein conformal frame, reduces to general relativity with a set of material fields whose EMT satisfies the null energy condition, which rules out static WHs. On the other hand, given a static WH in D dimensions as described in the previous subsection, it would also appear as a static WH in (d_0+2)-dimensional Einstein frame since the relevant conformal factor (the volume factor of extra dimensions) is everywhere finite and nonzero. We have to conclude that static WHs are absent in our model.

This means in turn that the sufficient condition (47) must be violated, and a properly formulated opposite inequality must hold. We arrive at the following theorem for brane systems having an orthogonal subsystem:

Theorem 4: *Consider a vector space $\mathcal{V}$, with a scalar product defined by the metric* (17), *where $d_i\in\mathbb{N}$, $i=0,\ldots,n$, $d_0>1$, $d_1=1$, $\bar{D}=\Sigma_{i=0}^n d_i-1$, and a set of nonzero vectors $\mathbf{Y}_s, s\in\mathcal{S}$, defined in* (18) *($I_s\subseteq\{1,\ldots,n\}$, $\chi_s\lambda_{sa}\in\mathbb{R}$). Let there be a subset $\mathcal{S}_\perp\subset\mathcal{S}$ such that $\mathbf{Y}_s\mathbf{Y}_{s'}=0$ for $s\neq s'$, $s,s'\in\mathcal{S}_\perp$. Then the following inequality holds:*

$$\sum_{s\in\mathcal{S}_\perp}\delta_{1I_s}N_s^2\leqslant\frac{d_0}{d_0-1},\quad \textit{or for}\ \lambda_{sa}=0,$$
$$\sum_{s\in\mathcal{S}_\perp}\delta_{1I_s}\left[d(I_s)\left(1-\frac{d(I_s)}{D-2}\right)\right]^{-1}\leqslant\frac{d_0}{d_0-1}. \tag{48}$$

The factor δ_{1I_s} in (48) excludes quasiscalars. For $\mathcal{S}_\perp = \mathcal{S}$ the theorem has been already proved by the above reasoning. If there are $\mathbf{Y}_s \notin \mathcal{S}_\perp$, their influence can be ruled out by turning to zero the corresponding charges Q_s, and then, as before, assuming the contrary of (48), we immediately obtain a Lorentzian WH solution.

Comment: The formulation of Theorem 4 does not mention F-forms, time, or any other physical entities and is actually of purely geometric (or even combinatorial) nature. From the combinatorial viewpoint it is essential that in the set $I_0 = \{0, \ldots, n\}$ there is a distinguished number, in our case 1, with $d_1 = 1$, included in all subsets I_s entering into the sum. Our proof, however, rests on physically motivated analytical considerations.

A similar theorem for a brane system with a BOS subsystem is readily obtained:

Theorem 4a: *Consider the model described in Sec. II, under the conditions specified in the first sentence of Theorem 4. Let there be a subset* $\mathcal{S}' \subset \mathcal{S}$ *such that the vectors* $\mathbf{Y}_s, s \in \mathcal{S}'$ *form a block-orthogonal system with respect to the metric* (17). *Then the following inequality holds for* $s \in \mathcal{S}'$:

$$\sum_{\{\omega:\, \hat{q}_\omega > 0\}} N_\omega^2 \leqslant \frac{d_0}{d_0 - 1}, \tag{49}$$

where $\hat{q}_\omega$ *and* N_ω^2 *are defined in* (43) *and, for all* q_s *included in the sum,* $\epsilon_s = \operatorname{sign} q_s = -1 + 2\delta_{1I_s}$.

According to the latter condition, ϵ_s depends on the inclusion or noninclusion of the distinguished one-dimensional factor space $\mathcal{M}_1$ $(= \mathbb{R}_t$ in Lorentzian models) into the world volume of specific BOS members. Thus, unlike the OS case, the sum may include F-forms with different ϵ_s, but in such a way that the combined factor $\hat{q}_\omega = \Sigma_{\mu \in \omega} q_\mu$ be positive for each ω.

VI. CONCLUDING REMARKS

(1) Some general restrictions on the behavior of brane systems described by the action (1) have been obtained, independent of specific space-time symmetry and signature: cosmological asymptotics, some BH properties and a universal restriction on the parameters of possible orthogonal or block-orthogonal subsystems (Theorems 4 and 4a).

Throughout the paper, the D-dimensional Einstein (D-E) conformal frame was used, although in such a general setting of the problem there is no evident reason to prefer one frame or another. For any specific underlying theory that leads to (1) in a weak field limit, two conformal frames are physically distinguished: one where the theory is originally formulated and another, providing the validity of the weak equivalence principle (or geodesic motion) for ordinary matter in 4 dimensions; the latter depends on how fermions are introduced in the underlying theory.[17,33,34] The first one should be used when discussing such issues as singularities or topology of a model, etc. (*what happens as a matter of fact*), while the second one, the so-called atomic system of measurements, is necessary for formulating observational predictions (*what we see*). They are, generally speaking, different.

Among the present results, however, only cosmological ones are conformal frame-dependent if different frames are connected by exponentials of the internal scale factors β^i and dilatonic fields φ^a. Indeed, such factors, being regular everywhere including horizons and asymptotics, cannot change the BH or WH nature of a given metric. (The only exceptions are hypothetic exceptional extreme BH solutions mentioned in Remark 2.)

The conclusions of Sec. III A on cosmological asymptotics are directly applicable to theories formulated in the outset in the D-E frame, like the weak-field bosonic sector of $D = 11$ supergravity following from M-theory,[3] where the action (truncated by neglecting the Chern-Simons term) has the form (1) with a single antisymmetric 4-form and no scalar fields.

(2) Unlike Lorentzian ones, Euclidean WHs (EWHs) are not ruled out, and the reason is (taking, say, OS solutions as an example) that, when selecting the F-forms (branes) for WH

construction, in the Euclidean case we are no more restricted to I_s containing a distinguished number, connected with $\mathbb{R}_t$, while now $\mathbb{R}_t = \mathbb{R}_u$. So there is a wider choice of I_s able to give $h_s < 0$ and to fulfill the WH necessary and sufficient condition (47).

As seen from Table I, EWHs corresponding to (1), if any, may be built only with the aid of magnetic forms F_s, though the existence of electric forms in a WH solution is not excluded.

The situation is well exemplified for $D=11$ supergravity. Indeed, the orthogonality conditions (34) are satisfied by 2-branes, $d(I_s)=3$, and 5-branes, $d(I_s)=6$, if the intersection rules hold:

$$d(3\cap 3)=1, \quad d(3\cap 6)=2, \quad d(6\cap 6)=4. \tag{50}$$

(the notations are evident); for all F-forms $N_s^2 = 1/2$. In particular, with $d_0=2$ or $d_0=3$ and other $d_i=1$, there is a maximal OS of seven 2-branes,[12,15]

$$\begin{array}{ll} a:\ 123, & d:\ 345, \\ b:\ 147, & e:\ 246, \\ c:\ 156, & f:\ 257, \\ & g:\ 367, \end{array} \tag{51}$$

where the figures 1, . . . ,7 label one-dimensional factor spaces, and for static models ''1'' refers to the time axis $\mathbb{R}_t$. Only three of these I_s (a,b,c) have $\delta_{1I_s}=1$, i.e., describe true electric or magnetic fields in a static space-time. Lorentzian WHs are absent since (47) requires $\Sigma_s N_s^2 > 2$ for $d_0=2$ and $>3/2$ for $d_0=3$.

In the Euclidean case we can have as many as 7 magnetic 2-branes, each with $N_s^2=1/2$, and WHs are easily found. Though, the latter is true if one considers F_{mI} of rank 7. If one remains restricted, as usual, to F_s of rank 4,[3] then for magnetic forms $d(I)=6$, and EWHs cannot be obtained. In other words, OS and BOS solutions do not lead to EWHs in standard 11-dimensional supergravity. Examples of EWHs have been found[20] for $D=12$ theory.[6]

By construction, classical EWHs possess finite actions and are related to possible quantum tunneling processes. Explicit expressions for their action and throat radii in the case of symmetric WHs described by OS and BOS solutions, have been calculated[20] explicitly in a general form for WHs which are symmetric with respect to their throats.

(3) The present conclusions rest on the positive energy requirement that seems quite natural as long as we deal with classical fields. Thus, in particular, the well-known singularity theorems of general relativity actually work as well in multidimensional p-brane cosmology. Meanwhile, the low energy limit of the unification theories is believed to work at scales from Planckian to subatomic and in the early universe where quantum effects of both gravity and material fields must be of importance (e.g., the Casimir effect due to compactification of extra dimensions), and a classical treatment is only a tentative, though necessary, stage in studying such systems. One can mention some papers discussing the relevant quantum effects: the Wheeler-DeWitt equation for p-branes[14,21] and the Casimir effect in cosmology.[35] Some nonquantum effects able to prevent a cosmological singularity are discussed by Kaloper *et al.*[30] and Gasperini;[36] see also references therein. All such effects necessarily violate the usual energy requirements and can therefore create traversable Lorentzian wormholes.

ACKNOWLEDGMENTS

I am grateful to V. Ivashchuk, V. Melnikov, J. Fabris, S.-W. Kim and Y. Keam for helpful discussions, and to colleagues from DF-UFES, Vitória, Brazil, and Ewha Women's University, Seoul, Korea, where part of the work was done, for warm hospitality. The work was supported in part by CAPES (Brazil), by the Russian State Committee for Science and Technology, and by the Russian Basic Research Foundation.

[1] *Supergravities in Diverse Dimensions*, reprints in 2 volumes, edited by A. Salam and E. Sezgin (World Scientific, Singapore, 1989).
[2] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory*, in 2 volumes (Cambridge University Press, Cambridge, 1987).
[3] C. Hull and P. Townsend, ''Unity of superstring dualities,'' Nucl. Phys. B **438**, 109 (1995); P. Horava and E. Witten, *ibid.* **460**, 506 (1996); hep-th/9510209; hep-th/9603142; J. M. Schwarz, ''Lectures on Superstring and M-Theory Dualities,'' hep-th/9607201; K. S. Stelle, ''Lectures on Supergravity ρ-Branes,'' hep-th/9701088; M. J. Duff, ''M-theory (the Theory Formerly Known as Strings,'' hep-th/9608117.
[4] H. Lü, C. N. Pope, T. A. Tran, and K. W. Xu, hep-th/9608055.
[5] I. Bars and C. Kounnas, ''Theories with Two Times,'' hep-th/9703060; H. Nishino, ''Supergravity in 10+2 Dimensions as a Consistent Background for Superstring,'' hep-th/9703214.
[6] N. Khvengia, Z. Khvengia, and C. N. Pope, ''Toward Field Theory of F-theory,'' hep-th/9703012.
[7] V. D. Ivashchuk and V. N. Melnikov, ''Intersecting *p*-Brane Solutions in Multidimensional Gravity and *M*-Theory,'' Grav. Cosmol. **2**, 297 (1996); hep-th/9612089.
[8] V. D. Ivashchuk and V. N. Melnikov, Phys. Lett. B **384**, 58 (1996).
[9] V. D. Ivashchuk, V. N. Melnikov, and M. Rainer, ''Multidimensional Sigma-Models with Composite Electric *p*-Branes,'' Grav. Cosmol. **4**, 73 (1998).
[10] V. D. Ivashchuk and V. N. Melnikov, Phys. Lett. B **403**, 23 (1997).
[11] V. D. Ivashchuk and V. N. Melnikov, hep-th/9705036; Class. Quantum Grav. **14**, 3001 (1997).
[12] K. A. Bronnikov, M. A. Grebeniuk, V. D. Ivashchuk, and V. N. Melnikov, ''Integrable Multidimensional Cosmology with Intersecting *p*-Branes,'' Grav. Cosmol. **3**, 2 (1997); gr-qc/9709006.
[13] K. A. Bronnikov, U. Kasper, and M. Rainer, ''Interesting Electric and Magnetic *p*-Branes: Spherically Symmetric Solutions,'' gr-qc/9708058.
[14] V. D. Ivashchuk and V. N. Melnikov, ''Multidimensional classical and quantum cosmology with interesting *p*-branes,'' J. Math. Phys. **39**, 2866 (1998).
[15] K. A. Bronnikov, V. D. Ivashchuk, and V. N. Melnikov, ''The Reissner–Nordström Problem for Intersecting Electric and Magnetic *p*-Branes,'' Grav. Cosmol. **3**, 203 (1997); gr-qc/9710054.
[16] K. A. Bronnikov and G. N. Shikin, ''On Interacting Fields in General Relativity,'' Izv. Vuzov, Fiz. 1977, No. 9, 25.
[17] K. A. Bronnikov, ''Spherically Symmetric Solutions in *D*-Dimensional Dilaton Gravity,'' Grav. Cosmol. **1**, 67 (1995).
[18] I. Ya. Aref'eva, M. G. Ivanov, and O. A. Rytchkov, ''Properties of intersecting *p*-Branes in Various Dimensions,'' hep-th/9702077; I. Ya. Aref'eva, M. G. Ivanov, and I. V. Volovich, ''NonExtremal Intersecting *p*-Branes in Various Dimensions,'' hep-th/9702079.
[19] H. Lü, C. N. Pope, and K. W. Xu, ''Liouville and Toda Solitons in M-Theory,'' hep-th/9604058.
[20] K. A. Bronnikov, ''Block-Orthogonal Brane Systems, Black Holes and Wormholes,'' Grav. Cosmol. **4**, 49 (1998); hep-th/9710207.
[21] H. Lü, J. Maharana, S. Mukherji, and C. N. Pope, ''Cosmological Solutions, *p*-Branes, adn the Wheeler-DeWitt Equation,'' hep-th/9707182.
[22] V. D. Ivashchuk and V. N. Melnikov, ''Madjumdar-Papapetrou Solutions in Sigma Model and Intersecting *p*-Branes,'' hep-th/9802121.
[23] There is an exception: two components having only one noncoinciding index, cannot coexist since in this case there emerge nonzero off-block-diagonal components of the energy-momentum tensor (EMT), while the Einstein tensor in the l.h.s. of the Einstein equations is block-diagonal. See more details in Ref. 11.
[24] K. A. Bronnikov, Acta Phys. Pol. B **4**, 251 (1973).
[25] S. Giddings and A. Strominger, Nucl. Phys. B **306**, 890 (1988).
[26] Hongsu Kim, ''Axionic Wormholes: More on their Classical and Quantum Aspects,'' hep-th/9706015.
[27] One might just require $|\varphi|<\infty$ as part of the center regularity conditions. Our proof, however, also rules out a hypothetical situation when an infinite φ value, due to the factor $e^{2\lambda\varphi}$, modifies the behavior of *F*-form EMTs, leading to a regular geometry.
[28] V. D. Ivashchuk and V. N. Melnikov, Class. Quantum Grav. **11**, 1793 (1994); Int. J. Mod. Phys. D **4**, 167 (1995).
[29] K. A. Bronnikov, Int. J. Mod. Phys. D **4**, 491 (1995).
[30] N. Kaloper, I. I. Kogan, and K. A. Olive, ''Cos(M)ological Solutions in M- and String Theory,'' hep-th/9711027.
[31] D. Hochberg and M. Visser, ''Geometric Structure of the Generic Static Traversable Wormhole Throat,'' gr-qc/9704082.
[32] S. Hayward, ''Dynamic Wormholes,'' gr-qc/9805019.
[33] R. Dick, Gen. Relativ. Gravit. **30**, 435 (1998).
[34] K. P. Staniukovich and V. N. Melnikov, *Hydrodynamics, Fields and Constants in Theory of Gravity* (Energoatomizdat, Moscow, 1983) (in Russian).
[35] U. Günther, S. Kriskiv, and A. Zhuk, ''On Static Compactification with a Casimir-Like Potential,'' Grav. Cosmol. **4**, 1 (1998).
[36] M. Gasperini, ''Repulsive Gravity in the Very Early Universe,'' gr-qc/9805060.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Nonlinear realizations of the diffeomorphism group in metric-affine gauge theory of gravity

Giovanni Giachetta
Department of Mathematics and Physics, University of Camerino, 63032 Camerino MC, Italy

(Received 25 February 1998; accepted for publication 27 October 1998)

The action of diffeomorphisms on coupled metric and spinor fields on a world manifold X is interpreted in terms of nonlinear realizations of the group $\widetilde{\mathrm{GL}}^+(4,\mathbb{R})$, the universal twofold covering group of the general linear group $\mathrm{GL}^+(4,\mathbb{R})$, on the quotient manifold $(\widetilde{\mathrm{GL}}^+(4,\mathbb{R})\times V)/\mathrm{SL}(2,\mathbb{C})$, where $\mathrm{SL}(2,\mathbb{C})$ is the spin group and V is the spinor space. By using nonlinear realizations the connection of a metric-affine world manifold couples naturally to standard spinor fields. This enables us not to exceed the scope of usual spinor models as in the case in which infinite-dimensional representations of $\widetilde{\mathrm{GL}}^+(4,\mathbb{R})$ are considered. As an application, by starting from the familiar Lagrangian for spin-1/2 models and using the nonlinear realization method, a Lagrangian density for spinor fields which has $\widetilde{\mathrm{GL}}^+(4,\mathbb{R})$ as invariance group is constructed. The total Lagrangian density is obtained by adding the Lagrangian of the metric-affine gravity. The energy-momentum current associated with every vector field on the world manifold X is calculated explicitly. It turns out that spinor fields do not contribute to the corresponding superpotential, which takes a form similar to that obtained by Komar. © *1999 American Institute of Physics.* [S0022-2488(99)01902-7]

I. INTRODUCTION

As is well known,[1] the classical theory of general relativity, especially when the tetrad approach is used, can be reformulated in terms of nonlinear realizations[2] of the general linear group $\mathrm{GL}(4,\mathbb{R})$. This description is particularly well suited in gauge gravitation theory if one regards general covariance as a spontaneously violated symmetry because of the presence of spinor fields, the gravitational (metric) field being regarded as a Higgs field.[3]

More recently, a framework has been introduced to deal with dynamical gravitational fields in the presence of spinor fields.[4,5] For each gravitational field there is a distinct spinor bundle. Therefore the configuration space of coupled gravitational and spinor fields can be conveniently represented as the total space S of the composite fiber bundle

$$S\to\Sigma\to X, \tag{1}$$

where $\Sigma\to X$ is the configuration space of gravitational fields on the world manifold X and $S\to\Sigma$ is a vector bundle with structure group the spin group $\mathrm{SL}(2,\mathbb{C})$. Given a gravitational field $h:X\to\Sigma$, the restriction of $S\to\Sigma$ to $h(X)\subset\Sigma$ is exactly the space of spinors for h.

In this paper we show that this framework has the structure of a gauge theory with the structure group, the universal twofold covering group

$$\zeta:\widetilde{\mathrm{GL}}^+(4,\mathbb{R})\to\mathrm{GL}^+(4,\mathbb{R}) \tag{2}$$

of the general linear group $\mathrm{GL}^+(4,\mathbb{R})$, acting nonlinearly on the quotient

$$(\widetilde{\mathrm{GL}}^+(4,\mathbb{R})\times V)/\mathrm{SL}(2,\mathbb{C}). \tag{3}$$

Here V is the spinor space and the quotient is defined by choosing a spinor representation of $SL(2,\mathbb{C})$ in V. Note that (3) is the total space of the fiber bundle associated with the principal bundle

$$\widetilde{GL}^+(4,\mathbb{R})\to\widetilde{GL}^+(4,\mathbb{R})/SL(2,\mathbb{C}).$$

In particular, we prove that the composite fiber bundle (1) can be regarded as a fiber bundle associated with a $\widetilde{GL}^+(4,\mathbb{R})$-principal bundle, its typical fiber being exactly the quotient $(\widetilde{GL}^+(4,\mathbb{R})\times V)/SL(2,\mathbb{C})$ on which $\widetilde{GL}^+(4,\mathbb{R})$ acts nonlinearly on the left.

In the gauging of $\widetilde{GL}^+(4,\mathbb{R})$ by using nonlinear realizations, the connection of a metric-affine world manifold is naturally coupled to standard fermionic matter. Alternatively, one can consider spinor representations of $\widetilde{GL}^+(4,\mathbb{R})$[6,7] which, however, are infinite dimensional [the only finite-dimensional linear representations of $\widetilde{GL}^+(4,\mathbb{R})$ are those obtained from representations of $GL^+(4,\mathbb{R})$ via composition with the covering map (2)]. Elements of these representations are called world spinors and the corresponding field theory has been already developed (see Ref. 6 and references therein).

Under suitable topological conditions on X, every diffeomorphism $\varphi:X\to X$ can be lifted to a diffeomorphism $\varphi_S:S\to S$ which preserves the composite fiber bundle structure (1), but in general this action yields only a projective representation of the diffeomorphism group $D(X)$. Here we calculate explicitly the infinitesimal version of this action. More precisely, for every vector field $\tau:X\to TX$ on X we find the expression of its natural lift $\tau_S:S\to TS$ on S. This vector field projects onto a vector field τ_Σ on Σ and, of course, onto τ. Note that, since only the normal subgroup $D_0(X)$ of diffeomorphisms which are homotopic to the identity is involved at the infinitesimal level, the association $\tau\mapsto\tau_S$ yields a true representation, and not just a projective one.

As an application, starting with the familiar Lagrangian density for spin-1/2 models and using the nonlinear realization method, we construct a Lagrangian $\mathcal{L}_M$ which has $\widetilde{GL}^+(4,\mathbb{R})$ as invariance group and in which the spinor fields are coupled with the torsion of the linear connection on X. Then the total Lagrangian density $\mathcal{L}=\mathcal{L}_M+\mathcal{L}_{MAG}$, where $\mathcal{L}_{MAG}$ is the Lagrangian density of the metric-affine gravitation theory, is invariant under the action of the diffeomorphism group $D(X)$ on the total configuration space $C\times_X S$. Here $C\to X$ is the fiber bundle of linear connections on X. Hence, for every vector field τ on X, the Lie derivative of $\mathcal{L}$ with respect to the lift τ_{CS} of τ on $C\times_X S$ vanishes identically, and therefore the corresponding energy-momentum current is conserved. It turns out that spinor fields do not contribute at all to the energy-momentum superpotential, which takes the form of the generalized Komar superpotential in the metric-affine theory of gravity.[8]

The ideas and results are arranged as follows. In Sec. II we describe how the symmetry reduction scheme used in gauge gravitation theory can be extended to general gauge theories. This can be done by introducing fields (Higgs fields) transforming nonlinearly under the action of the gauge group. In Sec. III we set our notations and review the main concepts related to spin structures on a four-manifold. In Sec. IV we deal with the problems of constructing a configuration bundle of coupled metric and spinor fields and defining a covariant derivative of its sections. In Sec. V we concentrate on the action of the diffeomorphism group $D(X)$ on the coupled metric-spinor configuration space. Section VI is devoted to the example sketched above. An appendix is added to describe the general approach to differential conservation laws in Lagrangian field theory.

Throughout the paper manifolds are real, finite dimensional, Hausdorff, second countable (hence paracompact), and connected.

II. SPONTANEOUS SYMMETRY BREAKING

In this section we illustrate the general description of spontaneously broken symmetries in gauge theories where matter fields admit only exact symmetry transformations.[9,10]

In classical field theory the spontaneous symmetry breaking is described by classical Higgs fields. If the gauge theory is formulated on a principal bundle $P \to X$, the necessary condition for spontaneous symmetry breaking is the reduction of the structure group G (the group whose realizations are required) of this principal bundle to the closed subgroup H of exact symmetries. Higgs fields are then represented by global sections h of the fiber bundle

$$\Sigma = P/H \to X.$$

This is a P-associated bundle with the typical fiber G/H on which the structure group G acts naturally on the left.

As is well known (see Ref. 11, p. 57), the set of Higgs fields h is in one-to-one correspondence with the set of reduced H-principal subbundles P^h of P. Given such a subbundle P^h, let

$$Y^h = (P^h \times V)/H \to X \tag{4}$$

be an associated fiber bundle with a typical fiber V carrying a linear representation of the subgroup H. Its sections describe matter fields in the presence of the Higgs field h.

Let us now consider the composite fiber bundles

$$P \to \Sigma \to X \tag{5}$$

and

$$Y \to \Sigma \to X, \tag{6}$$

where

$$P \to \Sigma \tag{7}$$

is a principal bundle with the structure group H and

$$Y = (P \times V)/H \to \Sigma \tag{8}$$

is a fiber bundle associated with it. Given a Higgs field h and the corresponding fiber bundle (4), it is easily seen that this latter is canonically isomorphic with the subbundle of the composite fiber bundle (6) given by the restriction of $Y \to \Sigma$ to $h(X) \subset \Sigma$. It follows that there is one-to-one correspondence between pairs (h, ψ_h), formed by a matter field ψ_h in the presence of the Higgs field h, and sections $\psi : X \to Y$ of the composite fiber bundle (6).

We regard Y as the configuration space for coupled Higgs and matter fields in gauge theory with spontaneously broken symmetries.

Note that for different Higgs fields h and h' we have different bundles Y^h and $Y^{h'}$, and there is no natural way of identifying these bundles. In other words, the fiber bundle $Y \to \Sigma$ is not in general the fibered product of the configuration space of Higgs fields Σ and some typical matter bundle. Essentially, this is due to the fact that matter fields admit only exact symmetry transformations. More precisely, there is no representation of the group G in V that restricts to the given representation of H in V. In fact, the following result holds.

Proposition 1: Let the linear representation of H in V be the restriction of a representation of G in V. Then the fiber bundle (8) is canonically isomorphic with the pull-back bundle $\Sigma \times_X (P \times V)/G$.

Proof: It is readily verified that the map defined by

$$(p,v)H \mapsto (pH,(p,v)G), \quad \forall p \in P, \quad v \in V,$$

is an isomorphism of Y onto $\Sigma \times_X (P \times V)/G$, over Σ. □

Let us now specify the transformation laws for Higgs and matter fields. Every automorphism ϕ of the principal bundle $\pi : P \to X$ [a diffeomorphism $\phi : P \to P$ such that $\phi(pg) = \phi(p)g$ for every $p\in P$, $g\in G$] defines diffeomorphisms $\phi_\Sigma:\Sigma\to\Sigma$ and $\varphi:X\to X$ such that the following diagram commutes:

$$\begin{array}{ccc} P & \xrightarrow{\phi} & P \\ \downarrow & & \downarrow \\ \Sigma & \xrightarrow{\phi_\Sigma} & \Sigma \\ \downarrow & & \downarrow \\ X & \xrightarrow{\varphi} & X. \end{array}$$

Since ϕ is also an automorphism of the principal bundle $P\to\Sigma$, it defines an automorphism ϕ_Y of the associated bundle (8) according to the law

$$(p,v)H\mapsto(\phi(p),v)H, \quad \forall p\in P, \quad v\in V.$$

We have the following commutative diagram:

$$\begin{array}{ccc} Y & \xrightarrow{\phi_Y} & Y \\ \downarrow & & \downarrow \\ \Sigma & \xrightarrow{\phi_\Sigma} & \Sigma \\ \downarrow & & \downarrow \\ X & \xrightarrow{\varphi} & X. \end{array}$$

Remark 1: Given a pair (h,ψ_h) formed by a Higgs field h and a section ψ_h of the fiber bundle (4), the automorphism ϕ_Y defined above takes this pair into $(h',\psi_{h'})$, where $h'=\phi_\Sigma\circ h\circ\varphi^{-1}$ is the transformed Higgs field and $\psi_{h'}$ is the section of $Y^{h'}\to X$ uniquely determined by h' and the section $\psi'=\phi_Y\circ\psi\circ\varphi^{-1}$. Here $\psi:X\to Y$ is the section corresponding to (h,ψ_h) under the correspondence described above.

Having specified the transformation laws for the fields, we now have to define the covariant derivatives of field systems on composite fiber bundles. We assume that the Lie algebra $\mathfrak{g}$ of G is the direct sum

$$\mathfrak{g}=\mathfrak{h}\oplus\mathfrak{m}$$

of the Lie algebra $\mathfrak{h}$ of the subgroup H and a subspace $\mathfrak{m}\subset\mathfrak{g}$ such that

$$\mathrm{ad}(H)(\mathfrak{m})\subset\mathfrak{m},$$

where ad is the adjoint representation.

Let $\omega:P\to\mathfrak{g}$ be a connection form on P. Then the $\mathfrak{h}$-component $\omega_\mathfrak{h}$ of ω is a connection form on the H-principal bundle $P\to\Sigma$, whereas the $\mathfrak{m}$-component $\omega_\mathfrak{m}$ is a tensorial form of type (ad,$\mathfrak{m}$) on $P\to\Sigma$.[11] For every Higgs field h, the pull-back of $\omega_\mathfrak{h}$ on P^h is a connection form on the reduced subbundle $P^h\to X$. Moreover, the pull-back of $\omega_\mathfrak{m}$ on P^h coincides with the covariant derivative of h with respect to ω. It gives the obstruction for ω to reduce to a connection on P^h.

The connection form $\omega_\mathfrak{h}$ induces a connection $A:Y\to J^1Y_\Sigma$ on the associated bundle (8). Here J^1Y_Σ denotes the first jet prolongation of this bundle. If the total space Y is provided with adapted coordinates (x^μ,σ^p,y^i), where (x^μ,σ^p) are bundle coordinates on the fiber bundle $\Sigma\to X$, then A reads locally

$$A = dx^\mu \otimes \left(\frac{\partial}{\partial x^\mu} + I(\mathbf{A}_\mu)^i{}_j y^j \frac{\partial}{\partial y^i} \right) + d\sigma^p \otimes \left(\frac{\partial}{\partial \sigma^p} + I(\mathbf{A}_p)^i{}_j y^j \frac{\partial}{\partial y^i} \right),$$

where $\mathbf{A}_\mu$ and $\mathbf{A}_p$ are the connection parameters of $\omega_{\mathfrak{h}}$ (which are opposite in sign to the gauge potentials) and $I(\cdot)^i{}_j$ are the Lie algebra representation matrices of the given representation of H.

Let us consider the jet manifolds $J^1\Sigma$ of $\Sigma \to X$, $J^1 Y_\Sigma$ of $Y \to \Sigma$ and $J^1 Y$ of $Y \to X$, with the corresponding coordinates

$$(x^\mu, \sigma^p, \sigma^p_\mu), \quad (x^\mu, \sigma^p, y^i, \widetilde{y}^i_\mu, y^i_p), \quad (x^\mu, \sigma^p, y^i, \sigma^p_\mu, y^i_\mu).$$

Then there exists the canonical morphism

$$\rho: J^1\Sigma \underset{\Sigma}{\times} J^1 Y_\Sigma \to J^1 Y,$$

$$\rho(j^1_x h, j^1_{h(x)} g) = j^1_x (g \circ h), \quad y^i_\mu \circ \rho = \widetilde{y}^i_\mu + y^i_p \sigma^p_\mu,$$

where g and h are sections of the fiber bundles $Y \to \Sigma$ and $\Sigma \to X$, respectively. Using the composition

$$\rho \circ A : J^1\Sigma \times_\Sigma Y \to J^1 Y,$$

$$y^i_\mu \circ \rho \circ A = I(\mathbf{A}_\mu + \mathbf{A}_p \sigma^p_\mu)^i{}_j y^j,$$

and the affine structure of the fiber bundle $J^1 Y \to Y$, one obtains the first-order differential operator

$$D: J^1 Y \to T^*X \otimes VY_\Sigma,$$
$$D = dx^\lambda \otimes [y^i_\mu - I(\mathbf{A}_\mu + \mathbf{A}_p \sigma^p_\mu)^i{}_j y^j] \frac{\partial}{\partial y^i}. \tag{9}$$

Here VY_Σ denotes the vertical tangent bundle of $Y \to \Sigma$.

The operator (9) has the following property. Given a Higgs field h, the restriction

$$D_h : J^1 Y^h \to T^*X \otimes VY^h,$$

$$D_h = dx^\mu \otimes [y^i_\mu - I(\mathbf{A}_\mu + \mathbf{A}_p \partial_\mu h^p)^i{}_j y^j] \frac{\partial}{\partial y^i},$$

of D to Y^h is the familiar covariant derivative relative to the principal connection

$$A_h = dx^\mu \otimes \left[\frac{\partial}{\partial x^\mu} + I(\mathbf{A}_\mu + \mathbf{A}_p \partial_\mu h^p)^i{}_j y^j \frac{\partial}{\partial y^i} \right]$$

on the fiber bundle (4).

One can use D in order to construct Lagrangians on the jet manifold $J^1 Y$ of the configuration space Y which are invariant under automorphisms of the G-principal bundle $P \to X$, starting from any Lagrangian which is invariant under the subgroup H. A G-invariant quadratic term in ω_m can be added if the Higgs fields have to propagate like the matter fields.

Let us now consider the H-principal bundle $G \to G/H$ and its associated bundle $(G \times V)/H \to G/H$. The group G acts on $(G \times V)/H$ on the left as follows. For every $a \in G$, the left translation L_a of G (sending $g \in G$ into $ag \in G$) is an automorphism of the principal bundle $G \to G/H$, over the diffeomorphism $\overline{L_a}$ of the quotient space G/H which takes a coset gH into the coset agH. Hence L_a induces the automorphism

$$\Lambda_a:(g,v)H\mapsto(ag,v)H$$

of the associated bundle $(G\times V)/H\to G/H$, over the diffeomorphism $\overline{L_a}$.

The group action of G on the quotient space $(G\times V)/H$ can be regarded as a nonlinear realization of G on the objects of physical interest, which are used as parameters or coordinates on the quotient G/H, and on all other (matter) fields. To see this, note that by using a (local) gauge s of the principal bundle $G\to G/H$, one can write

$$as(gH)=s(agH)h(gH,a), \quad \forall a,g\in G,$$

where $h(\cdot,a)$ is a function of G/H into the subgroup H. This relation defines implicitly the group action $\overline{L_a}$ on G/H by the formula

$$s(agH)=as(gH)h(gH,a)^{-1},$$

but it also defines the function h. This functions governs the behavior of the matter fields under the action of G by the law:

$$v\in V\mapsto h(gH,a)v\in V.$$

Going back to the framework based on composite fiber bundles and their automorphisms, the key remark which relates this framework to the nonlinear realization method is formulated in the following proposition.

Proposition 2: The composite fiber bundle (6) can be regarded as a fiber bundle with structure group G and typical fiber $(G\times V)/H$. In fact, $(P\times V)/H$ can be identified with the total space of the fiber bundle $(P\times(G\times V)/H)/G\to X$ associated with $P\to X$ [the group action of G on the typical fiber $(G\times V)/H$ being that described above] as follows. An element

$$(p,(g,v)H)G\in(P\times(G\times V)/H)/G, \quad p\in P, \;\; g\in G, \;\; v\in V,$$

is mapped into the element

$$(pg,v)H\in(P\times V)/H.$$

Note that in addition to the fiber bundle structure over X, $(P\times(G\times V)/H)/G$ is provided also with a natural structure of fibration over Σ. It is easily seen that this structure is preserved under the above identification with $(P\times V)/H$.

III. SPIN STRUCTURES

Hereafter, by a world manifold X we mean a noncompact, parallelizable four-dimensional manifold with a given orientation.

Remark 2: In classical field theory, if cosmological models are not considered, the world manifold X is assumed to satisfy accepted rules of causality.[12] A compact manifold does not have these properties because every Lorentz metric on it generates closed timelike curves. On the other hand, on a noncompact four-dimensional manifold X a Lorentz metric can always be chosen as to give no closed timelike curves (see Ref. 13, p. 168). Moreover, given such a Lorentz metric X admits a spin structure if and only if it is parallelizable.[14]

A Lorentz metric h on the world manifold X is a section of the fiber bundle

$$\Sigma=LX/\mathrm{SO}^0(1,3)\to X, \tag{10}$$

where $\pi:LX\to X$ is the $\mathrm{GL}^+(4,\mathbb{R})$-principal bundle of oriented linear coframes on X and $\mathrm{SO}^0(1,3)$ denotes the proper Lorentz group. As is well known, (10) is a LX-associated fiber bundle with the typical fiber $\mathrm{GL}^+(4,\mathbb{R})/\mathrm{SO}^0(1,3)$.

Given a chart (U,x^μ) on X, every element $\xi=\{\xi^a\}\in\pi^{-1}(U)\subset LX$ takes the form

$$\xi^a = \xi^a{}_\mu dx^\mu,$$

where the matrix $\xi^a{}_\mu$ belongs to $\mathrm{GL}^+(4,\mathbb{R})$. Hence the coframe bundle LX can be provided with bundle coordinates

$$(x^\mu, \xi_a{}^\mu),$$

where $\xi_a{}^\mu$ denotes the inverse matrix of $\xi^a{}_\mu$. In these coordinates, the canonical right action of the structure group $\mathrm{GL}^+(4,\mathbb{R})$ on LX reads

$$\xi_a{}^\mu \mapsto \xi_b{}^\mu g^b{}_a, \quad \forall g^b{}_a \in \mathrm{GL}^+(4,\mathbb{R}).$$

For subsequent purposes, we also introduce coordinates on LX adapted to the composite fibration

$$LX \to \Sigma \to X \tag{11}$$

as follows. Every element $\xi_a{}^\mu$ of $\mathrm{GL}^+(4,\mathbb{R})$ in some suitable neighborhood of the identity element may be written as

$$\xi_a{}^\mu = \langle \xi \rangle_b{}^\mu [\xi]^b{}_a,$$

where $[\xi]^b{}_a$ is an element of $\mathrm{SO}^0(1,3)$ and $\langle \xi \rangle_b{}^\mu$ is pseudo-symmetric, that is,

$$\langle \xi \rangle_a{}^\mu \eta_{\mu b} = \langle \xi \rangle_b{}^\mu \eta_{\mu a}.$$

It follows that the pseudo-symmetric matrices $\sigma_a{}^\mu$ form a local coordinate system on the quotient manifold $\mathrm{GL}^+(4,\mathbb{R})/\mathrm{SO}^0(1,3)$, and hence

$$(x^\mu, \sigma_a{}^\mu, \lambda^a{}_b),$$

where $\lambda^a{}_b$ are elements of the proper Lorentz group, are suitable bundle coordinates on LX adapted to the composite fibration (11).

By the way, note that the $\mathrm{SO}^0(1,3)$-principal fiber bundle $LX \to \Sigma$ admits the (local) gauge

$$\overline{\xi} \to \langle \xi \rangle^a{}_\mu dx^\mu, \tag{12}$$

where $\overline{\xi}$ denotes the coset of ξ.

Every metric field h determines the tetrad functions

$$h_a{}^\mu(x) = \sigma_a{}^\mu \circ h(x),$$

which are related to the metric functions $g^{\mu\nu}(x)$ by the well-known relation

$$g^{\mu\nu}(x) = h_a{}^\mu(x) h_b{}^\nu(x) \eta^{ab}. \tag{13}$$

Here

$$\eta_{ab} = \mathrm{diag}\ (1,-1,-1,-1)$$

is the Minkowski metric written with respect to an orthonormal basis $\{e_a\}$ of the Minkowski space M.

Coming to spinors, let us briefly introduce our notations (see Refs. 15 and 16 for a general description of the Clifford algebra techniques). Let $\mathbb{C}_{1,3}$ be the complex Clifford algebra generated by elements of M. The spinor space V is defined to be a minimal left ideal of $\mathbb{C}_{1,3}$ on which this algebra acts on the left. We denote by

$$\gamma : M \otimes V \to V, \quad \gamma(e_a) = \gamma_a \tag{14}$$

the representation of elements of the Minkowski space M by γ-matrices in V.

The Clifford group $G_{1,3}$ comprises all invertible elements $\tilde{\lambda}$ of the real Clifford algebra $\mathbb{R}_{1,3}$ such that the corresponding inner automorphisms induce Lorentz transformations of the Minkowski space M, that is,

$$\tilde{\lambda} e \tilde{\lambda}^{-1} = \lambda(e), \quad \forall e \in M, \tag{15}$$

with $\lambda \in \mathrm{O}(1,3)$. The representation (14) satisfies the following equivariance property:

$$\gamma(\lambda e \otimes \tilde{\lambda} v) = \tilde{\lambda} \gamma(e \otimes v), \quad \forall \tilde{\lambda} \in G_{1,3}, \quad e \in M, \ v \in V.$$

Since the action (15) of the Clifford group on the Minkowski space M is not effective, one usually considers the pin subgroup Pin (1,3) of $G_{1,3}$. The even part of Pin (1,3) is the spin group Spin (1,3). The restriction of the map (15) to the identity component

$$\mathrm{Spin}^0(1,3) \simeq \mathrm{SL}(2,\mathbb{C})$$

of Spin(1,3) yields the well-known twofold universal covering group

$$\zeta_0 : \mathrm{SL}(2,\mathbb{C}) \to \mathrm{SO}^0(1,3)$$

of the proper Lorentz group. The spin group $\mathrm{SL}(2,\mathbb{C})$ acts on the spinor space V by the generators

$$I_{ab} = \tfrac{1}{4}[\gamma_a, \gamma_b], \tag{16}$$

induced by the generators

$$L_{ab}{}^c{}_d = \delta^c_a \eta_{db} - \delta^c_b \eta_{da} \tag{17}$$

of the Lie algebra of $\mathrm{SO}^0(1,3)$.

Let $L^h X$ denote the $\mathrm{SO}^0(1,3)$-principal subbundle of the coframe bundle LX corresponding to the metric field $h : X \to \Sigma$. In order to define spinor fields in the presence of h, we have first to give a h-spin structure on X, i.e., an $\mathrm{SL}(2,\mathbb{C})$-principal bundle $\widetilde{L^h X} \to X$ and a principal bundle morphism

$$z_h : \widetilde{L^h X} \to L^h X \tag{18}$$

over X.[16] Spinor fields in the presence of the metric field h are sections of the vector bundle

$$S^h = (\widetilde{L^h X} \times V)/\mathrm{SL}(2,\mathbb{C}) \to X$$

associated with $\widetilde{L^h X}$ via the spinor representation of $\mathrm{SL}(2,\mathbb{C})$ in the spinor space V.

IV. CONFIGURATION MANIFOLDS

If one considers theories in which the Lorentz metrics are dynamical variables as well as the spinor fields, one has to take into account deformations of the spin structure as a consequence of deformations of the metric. This dependence on the metric can be taken into account by introducing a double cover of the coframe bundle (see also Refs. 4 and 5).

As is well known, the group $\mathrm{GL}^+(4,\mathbb{R})$ is not simply connected. Its first homotopy group is[17]

$$\pi_1(\mathrm{GL}(4,\mathbb{R})) = \pi_1(\mathrm{SO}(4)) = \mathbb{Z}_2.$$

Therefore $\mathrm{GL}^+(4,\mathbb{R})$ admits the universal twofold covering group (2) as to make commutative the diagram[18]

$$\begin{array}{ccc} \widetilde{\mathrm{GL}}^+(4,\mathbb{R}) & \xrightarrow{\zeta} & \mathrm{GL}^+(4,\mathbb{R}) \\ {\scriptstyle \tilde{i}}\uparrow & & \uparrow{\scriptstyle i} \\ \mathrm{SL}(2,\mathbb{C}) & \xrightarrow{\zeta_0} & \mathrm{SO}^0(1,3) \end{array}$$

Here $\tilde{i}$ and i are the inclusion morphisms.

In order to apply the procedure outlined in Sec. II, the first step is that of choosing the subgroup H to $G=\widetilde{\mathrm{GL}}^+(4,\mathbb{R})$ from which the quotient G/H can be formed. We take H to be the spin group $\mathrm{SL}(2,\mathbb{C})$. The choice of this subgroup, along with its known representations, enable us to define group actions of $\widetilde{\mathrm{GL}}^+(4,\mathbb{R})$ on half-integer as well as integer spinor fields. As a second step we take a prolongation $(\widetilde{LX},z)$ of LX to $\widetilde{\mathrm{GL}}^+(4,\mathbb{R})$, i.e., a $\widetilde{\mathrm{GL}}^+(4,\mathbb{R})$-principal bundle $\widetilde{LX}\to X$ and a principal bundle morphism

$$z=\widetilde{LX}\to LX$$

over X.

Note that the quotient manifold $\widetilde{LX}/\mathrm{SL}(2,\mathbb{C})$ can be identified with $\Sigma=LX/\mathrm{SO}^0(1,3)$. This is a straightforward consequence of the diffeomorphism

$$\widetilde{\mathrm{GL}}^+(4,\mathbb{R})/\mathrm{SL}(2,\mathbb{C})\cong \mathrm{GL}^+(4,\mathbb{R})/\mathrm{SO}^0(1,3)$$

between their typical fibers. Therefore the following diagram commutes:

$$\begin{array}{ccc} \widetilde{LX} & \xrightarrow{z} & LX \\ & \searrow\swarrow & \\ & \Sigma & \end{array}$$

Let us consider the composite fiber bundle

$$\widetilde{LX}\to\Sigma\to X$$

and the composite spinor bundle

$$S\to\Sigma\to X, \tag{19}$$

where

$$S=(\widetilde{LX}\times V)/\mathrm{SL}(2,\mathbb{C})\to\Sigma$$

is the fiber bundle associated with the $\mathrm{SL}(2,\mathbb{C})$-principal bundle $\widetilde{LX}\to\Sigma$. In accordance with Sec. II, S can be regarded as the configuration space of coupled metric and spinor fields. Given a metric field h, the restriction of the spinor bundle $S\to\Sigma$ to $h(X)\subset\Sigma$ is canonically isomorphic with the spinor bundle S^h associated with the h-spin structure defined by the following commutative diagram:

$$\begin{array}{ccc} \widetilde{LX} & \xrightarrow{z} & LX \\ {\scriptstyle \tilde{i}}\uparrow & & \uparrow{\scriptstyle i} \\ \widetilde{L^hX} & \xrightarrow{z_h} & L^hX \end{array}$$

Remark 3: Given a metric field h, one can show[4] that the topological obstructions to the existence of a prolongation $(\widetilde{LX},z)$ are the same as the obstructions to the existence of a h-spin structure. Moreover, the set of equivalence classes of prolongations of LX is in one-to-one correspondence with the set of equivalence classes of h-spin structures.

Let us now define a covariant derivative on the configuration space S. To begin with, recall that there is one-to-one correspondence between principal connections on $\widetilde{LX}\to X$ and principal connections on the coframe bundle $LX\to X$ (linear connections on X). If $\omega:TLX\to\mathfrak{gl}(4,\mathbb{R})$ is a connection form on LX, the corresponding connection form $\tilde{\omega}:T\widetilde{LX}\to\tilde{\mathfrak{gl}}(4,\mathbb{R})$ on $\widetilde{LX}$ is defined by the composition

$$\tilde{\omega}=(\zeta')^{-1}\circ(z^*\omega),$$

where $\zeta':\tilde{\mathfrak{gl}}(4,\mathbb{R})\to\mathfrak{gl}(4,\mathbb{R})$ is the Lie algebra isomorphism induced by ζ.

As is readily seen, the Lie algebra of the general linear group $GL(4,\mathbb{R})$ is the direct sum

$$\mathfrak{gl}(4,\mathbb{R})=\mathfrak{so}^0(1,3)\oplus\mathfrak{m}$$

of the Lie algebra $\mathfrak{so}^0(1,3)$ of the Lorentz group and a subspace $\mathfrak{m}\subset\mathfrak{gl}(4,\mathbb{R})$ such that

$$\mathrm{ad}(SO^0(1,3))(\mathfrak{m})\subset\mathfrak{m},$$

where ad is the adjoint representation. Then (see Sec. II) the $\mathfrak{so}^0(1,3)$-component ω' of a connection form ω on LX is a principal connection on $\widetilde{LX}\to\Sigma$, whereas the $\mathfrak{m}$-component $\omega_{\mathfrak{m}}$ is a tensorial form of type $(\mathrm{ad},\mathfrak{m})$ on the same fiber bundle. The local expressions of ω, ω', and $\omega_{\mathfrak{m}}$ with respect to the bundle coordinates $(x^\mu,\xi_a{}^\mu)$ of LX are given respectively by

$$\omega^a{}_b=\xi^a{}_\mu(d\xi_b{}^\mu-K^\mu{}_{\nu\alpha}\xi_b{}^\nu dx^\alpha),$$

$$\omega^{[ab]}=\tfrac{1}{2}(\xi^a{}_\mu\eta^{bc}-\xi^b{}_\mu\eta^{ac})(d\xi_c{}^\mu-K^\mu{}_{\nu\alpha}\xi_c{}^\nu dx^\alpha),$$

$$\omega^{(ab)}=\tfrac{1}{2}(\xi^a{}_\mu\eta^{bc}+\xi^b{}_\mu\eta^{ac})(d\xi_c{}^\mu-K^\mu{}_{\nu\alpha}\xi_c{}^\nu dx^\alpha),$$

where the connection parameters $K^\mu{}_{\nu\alpha}$ are local functions on X.

The pull-back of ω' with respect to the gauge (12) yields the following expression for its connection parameters in the bundle coordinates $(x^\mu,\sigma_\alpha{}^\mu)$ on Σ:

$$A^{ab}{}_\alpha=\tfrac{1}{2}(\sigma^a{}_\mu\eta^{bc}-\sigma^b{}_\mu\eta^{ac})K^\mu{}_{\nu\alpha}\sigma_c{}^\nu,$$

$$A^{abc}{}_\mu=-\tfrac{1}{2}(\sigma^a{}_\mu\eta^{bc}-\sigma^b{}_\mu\eta^{ac}).$$

It follows that the differential operator (9) reads

$$D:J^1S\to T^*X\underset{\Sigma}{\otimes}S,$$

$$D=dx^\alpha\otimes\left[y^i_\alpha-\frac{1}{4}(\sigma^a{}_\nu\eta^{bc}-\sigma^b{}_\nu\eta^{ac})(K^\nu{}_{\mu\alpha}\sigma_c{}^\mu-\sigma_c{}^\nu{}_{,\alpha})I_{ab}{}^i{}_jy^j\right]\frac{\partial}{\partial y^i}, \tag{20}$$

where $(x^\mu,\sigma_a{}^\mu,y^i)$ denote bundle coordinates on S and $I_{ab}{}^i{}_j$ are the Lie algebra representation matrices (16) of the given representation of $SL(2,\mathbb{C})$.

Starting from any $SL(2,\mathbb{C})$-invariant Lagrangian and using the differential operator (20), one can construct Lagrangians which are invariant under automorphisms of the $\widetilde{GL}^+(4,\mathbb{R})$-principal bundle $\widetilde{LX}\to X$. For example, the familiar Dirac Lagrangian for spin-1/2 models leads to the following Lagrangian defined on the configuration space $C\times_XS$:

$$\mathcal{L}_M : C \underset{X}{\times} J^1 S \to \overset{4}{\wedge} T^*X, \quad \mathcal{L}_M = L_M d^4 x,$$

$$L_M = \frac{i}{2} \left\{ y_i^*(\gamma^0 \gamma^\alpha)^i{}_j \left[y^j_\alpha - \frac{1}{4} (\sigma^a{}_\nu \eta^{bc} - \sigma^b{}_\nu \eta^{ac})(k^\nu{}_{\mu\alpha} \sigma_c{}^\mu - \sigma_c{}^\nu{}_{,\alpha}) I_{ab}{}^j{}_k y^k \right] \right.$$
$$\left. - \left[y^*_{i\alpha} - \frac{1}{4} (\sigma^a{}_\nu \eta^{bc} - \sigma^b{}_\nu \eta^{ac})(k^\nu{}_{\mu\alpha} \sigma_c{}^\mu - \sigma_c{}^\nu{}_{,\alpha}) y^*_k I^*_{ab}{}^k{}_i \right] (\gamma^0 \gamma^\alpha)^i{}_j y^j \right\}$$
$$\times \det(\sigma^a{}_\mu) - m y_i^* (\gamma^0)^i{}_j y^j \det(\sigma^a{}_\mu). \tag{21}$$

Here $*$ denotes complex conjugation, $\gamma^\alpha = \gamma^a \sigma_a{}^\alpha$ and $C \to X$ is the fiber bundle of principal connections on LX. Its total space is the quotient[19]

$$C = J^1 LX / \mathrm{GL}^+(4,\mathbb{R}),$$

with bundle coordinates denoted by

$$(x^\mu, k^\mu{}_{\nu\alpha}). \tag{22}$$

It is easily verified that $\mathcal{L}_M$ satisfies the following relations:

$$\frac{\partial L_M}{\partial k^\alpha{}_{\beta\mu}} + \frac{\partial L_M}{\partial k^\alpha{}_{\mu\beta}} = 0, \tag{23}$$

$$\frac{\partial L_M}{\partial k^\alpha{}_{\nu\mu}} = \frac{\partial L_M}{\partial \sigma_a{}^\alpha{}_{,\nu}} \sigma_a{}^\mu. \tag{24}$$

Note that, by virtue of (23), the Lagrangian density (21) depends only on the torsion of the linear connection on X.

V. SPIN AND INFINITESIMAL DIFFEOMORPHISMS

In this section we define a morphism of the Lie algebra $\Xi(X)$ of vector fields on the world manifold X into the Lie algebra $\Xi(S)$ of vector fields on the configuration space S of coupled metric and spinor fields.

Every orientation-preserving diffeomorphism φ of the world manifold X lifts naturally to an automorphism $\hat{\varphi}$ of the coframe bundle LX defined by

$$\hat{\varphi}(\xi) = \{\xi^a \circ T\phi^{-1}\}, \quad \forall \xi \in LX. \tag{25}$$

We denote by φ_Σ the unique diffeomorphism of Σ such that the following diagram commutes:

$$\begin{array}{ccc} LX & \xrightarrow{\hat{\varphi}} & LX \\ \downarrow & & \downarrow \\ \Sigma & \xrightarrow{\varphi_\Sigma} & \Sigma \\ \downarrow & & \downarrow \\ X & \xrightarrow{\varphi} & X. \end{array}$$

It follows that, given a one-parameter group of orientation-preserving diffeomorphisms φ_t of X generated by a vector field $\tau : X \to TX$, the lift $\hat{\varphi}_t$ of φ_t to the coframe bundle generates a $\mathrm{GL}^+(4,\mathbb{R})$-invariant vector field $\hat{\tau}$ which is projectable onto a vector field τ_Σ on Σ and this, in turn, onto τ:

$$\begin{array}{ccc} LX & \xrightarrow{\hat{\tau}} & TLX \\ \downarrow & & \downarrow \\ \Sigma & \xrightarrow{\tau_\Sigma} & T\Sigma \\ \downarrow & & \downarrow \\ X & \xrightarrow{\tau} & TX. \end{array}$$

Since $z:\widetilde{LX}\to LX$ is a covering space, there exists a natural lift of $\hat{\varphi}_t$ to a one-parameter group of automorphisms $\widetilde{\varphi}_t$ of the principal bundle $\widetilde{LX}\to X$ and, hence, to a one-parameter group of automorphisms of the configuration space S of coupled metric and spinor fields given by

$$[\widetilde{\xi},v]\mapsto[\widetilde{\varphi}_t(\widetilde{\xi}),v], \quad \forall\widetilde{\xi}\in\widetilde{LX}, \ v\in V,$$

where $[\widetilde{\xi},v]$ is the coset of the element $(\widetilde{\xi},v)$. As a result, one obtains a vector field τ_S on S such that the following diagram commutes:

$$\begin{array}{ccc} S & \xrightarrow{\tau_S} & TS \\ \downarrow & & \downarrow \\ \Sigma & \xrightarrow{\tau_\Sigma} & T\Sigma \\ \downarrow & & \downarrow \\ X & \xrightarrow{\tau} & TX. \end{array}$$

Let us find the expression of the vector field τ_S in the bundle coordinates $(x^\mu,\sigma_a{}^\mu,y^i)$ of S. We start from the local expression of the automorphism (25) in the bundle chart $(x^\mu,\xi_a{}^\mu)$ of LX:

$$\hat{\varphi}:(x^\mu,\xi_a{}^\mu)\mapsto\left(\phi^\mu(x),\ \frac{\partial\phi^\mu}{\partial x^\nu}\ \xi_a{}^\nu\right).$$

Now, given a gauge $s:W\subset\Sigma\to LX$ of the principal bundle $LX\to\Sigma$, one can write

$$\hat{\varphi}\circ s(\overline{\xi})=(s\circ\varphi_\Sigma(\overline{\xi}))h(\overline{\xi}),$$

where $\overline{\xi}$ is the coset of ξ and h is a function of W into $\mathrm{SO}^0(1,3)$. In particular, if s is the gauge (12) the expression of $\hat{\varphi}$ reads

$$\hat{\varphi}:(x^\mu,\sigma_a{}^\mu,\lambda^a{}_b)\mapsto\left(\phi^\mu(x),\ \frac{\partial\phi^\mu}{\partial x^\nu}\ \sigma_b{}^\nu(h^{-1})^b{}_a,h^a{}_c\lambda^c{}_b\right). \tag{26}$$

Let $\tau=\tau^\mu\partial/\partial x^\mu$ be a vector field on X. Then, from (26), one finds the following expression for the vector field $\hat{\tau}$ on LX:

$$\hat{\tau}=\tau^\mu\ \frac{\partial}{\partial x^\mu}+\left(\frac{\partial\tau^\mu}{\partial x^\nu}\ \sigma_a{}^\nu-\sigma_b{}^\mu\tau^b{}_a\right)\ \frac{\partial}{\partial\sigma_a{}^\mu}+\frac{1}{2}\ \tau^{ab}\mathbf{e}_{ab}, \tag{27}$$

where $\tau^a{}_b(x^\mu,\sigma_a{}^\mu)$ is a function taking values into $\mathfrak{so}^0(1,3)$. Latin indices are lowered and raised with the Minkowski metric η_{ab} and $\mathbf{e}_{ab}=-\mathbf{e}_{ba}$ are the right invariant vector fields on $\mathrm{SO}^0(1,3)$ corresponding to the basis (17) of $\mathfrak{so}^0(1,3)$ given by

$$\mathbf{e}_{ab}=\frac{1}{2}\,(\delta^c_a\eta_{db}-\delta^c_b\eta_{da})\lambda^d{}_k\,\frac{\partial}{\partial\lambda^c{}_k}.$$

Consequently, the coordinate expression of the vector field τ_S on S reads

$$\tau_S=\tau^\mu\,\frac{\partial}{\partial x^\mu}+\left(\frac{\partial\tau^\mu}{\partial x^\nu}\,\sigma_a{}^\nu-\sigma_b{}^\mu\tau^b{}_a\right)\frac{\partial}{\partial\sigma_a{}^\mu}+\frac{1}{2}\,\tau^{ab}\left(I_{ab}{}^i{}_j y^j\,\frac{\partial}{\partial y^i}+y^*_j I^*_{ab}{}^j{}_i\,\frac{\partial}{\partial y^*_i}\right).$$

Given an orientation-preserving diffeomorphism φ of X, the equivariance of $\hat{\varphi}$ [with respect to the right action of $\mathrm{GL}^+(4,\mathbb{R})$ on LX] implies that the jet extension $j^1\hat{\varphi}:J^1LX\to J^1LX$ goes to the quotient $C=J^1LX/\mathrm{GL}^+(4,\mathbb{R})$, defining in this way an automorphism $\varphi_C:C\to C$ over φ. It follows that every vector field $\tau=\tau^\mu\partial_\mu$ on X can be naturally lifted to a vector field τ_C on C, whose expression with respect to the coordinates (22) is

$$\tau_C=\tau^\mu\partial_\mu+(\partial_{\beta\nu}\tau^\mu-k^\mu{}_{\beta\alpha}\partial_\nu\tau^\alpha-k^\mu{}_{\alpha\nu}\partial_\beta\tau^\alpha+k^\alpha{}_{\beta\nu}\partial_\alpha\tau^\mu)\,\frac{\partial}{\partial k^\mu{}_{\beta\nu}}$$

or, using the compact notation $y^A=k^\mu{}_{\beta\nu}$,

$$\tau_C=\tau^\mu\partial_\mu+(u^{A\beta\nu}_\mu\partial_{\beta\nu}\tau^\mu+u^{A\mu}_\beta\partial_\mu\tau^\beta)\,\frac{\partial}{\partial y^A}.$$

Later, we shall denote by

$$\tau_{CS}=\tau_C+\tau_S:C\underset{X}{\times}S\to T(C\underset{X}{\times}S)$$

the vector field on the configuration space $C\underset{X}{\times}S$ which is the sum of the vector fields τ_C and τ_S.

VI. APPLICATIONS

Let us consider the Lagrangian density

$$\mathcal{L}=\mathcal{L}_M+\mathcal{L}_{\mathrm{MAG}}:J^1(C\underset{X}{\times}S)\to\overset{4}{\wedge}T^*X,$$

where $\mathcal{L}_M$ is the Lagrangian density (21) and

$$\mathcal{L}_{\mathrm{MAG}}=L_{\mathrm{MAG}}d^4x,$$

$$L_{\mathrm{MAG}}=\sigma_a{}^\beta\sigma_b{}^\mu\eta^{ab}R^\alpha{}_{\beta\alpha\mu}\,\det(\sigma^a{}_\mu)$$

is the Lagrangian density of the metric-affine theory of gravity. Here

$$R^\alpha{}_{\beta\nu\mu}=k^\alpha{}_{\beta\nu,\mu}-k^\alpha{}_{\beta\mu,\nu}+k^\alpha{}_{\sigma\nu}k^\sigma{}_{\beta\mu}-k^\alpha{}_{\sigma\mu}k^\sigma{}_{\beta\nu}$$

is the curvature tensor. Note that

$$\partial L_{\mathrm{MAG}}/\partial k^\alpha{}_{\beta\mu,\nu}\equiv\pi_\alpha{}^{\beta\mu\nu}=-\pi_\alpha{}^{\beta\nu\mu},\tag{28}$$

$$\partial L_{\mathrm{MAG}}/\partial k^\alpha{}_{\beta\mu}=\pi_\sigma{}^{\beta\mu\nu}k^\sigma{}_{\alpha\nu}-\pi_\alpha{}^{\sigma\mu\nu}k^\beta{}_{\sigma\nu}.\tag{29}$$

By construction, the Lagrangian densities $\mathcal{L}_M$ and $\mathcal{L}_{\mathrm{MAG}}$ are invariant under the action of the diffeomorphism group $D(X)$. Hence they obey the relations

$$\mathbf{L}_{\tilde{\tau}_{CS}}\mathcal{L}_M=0, \quad \mathbf{L}_{\tilde{\tau}_{CS}}\mathcal{L}_{\mathrm{MAG}}=0, \quad \forall \tau \in \Xi(X), \tag{30}$$

which lead to conserved energy-momentum currents. Let us analyze these currents. For every vector field $\tau \in \Xi(X)$, one obtains the current (see the Appendix)

$$\begin{aligned} V^\mu(\mathcal{L},\tau)=&\frac{\partial L_{\mathrm{MAG}}}{\partial y^A_\mu}\,(u^{A\beta\gamma}_{\ \ \alpha}\partial_{\beta\gamma}\tau^\alpha+u^{A\alpha}_{\ \ \beta}\partial_\alpha\tau^\beta-y^A_\alpha\tau^\alpha)+\frac{\partial L_M}{\partial \sigma_c{}^\alpha{}_{,\mu}}\,(\sigma_c{}^\beta\partial_\beta\tau^\alpha-\sigma_a{}^\alpha\tau^a{}_c-\sigma_c{}^\alpha{}_{,\beta}\tau^\beta)\\ &+\frac{\partial L_M}{\partial y^i_\mu}\left(\frac{1}{2}\,\tau^{ab}I_{ab}{}^i{}_j y^j-y^i_\alpha\tau^\alpha\right)+\frac{\partial L_M}{\partial y^*_{i\mu}}\left(\frac{1}{2}\,\tau^{ab}y^*_j I^*{}_{ab}{}^j{}_i-y^*_{i\alpha}\tau^\alpha\right)+\tau^\mu(L_{\mathrm{MAG}}+L_M). \end{aligned} \tag{31}$$

By explicit calculation, one readily verifies that the term

$$-\frac{\partial L_M}{\partial \sigma_c{}^\alpha{}_{,\mu}}\,\sigma_a{}^\alpha\tau^a{}_c+\frac{1}{2}\,\frac{\partial L_M}{\partial y^i_\mu}\,\tau^{ab}I_{ab}{}^i{}_j y^j+\frac{1}{2}\,\frac{\partial L_M}{\partial y^*_{i\mu}}\,\tau^{ab}y^*_j I^*{}_{ab}{}^j{}_i\,,$$

corresponding to an ''internal'' transformation, vanishes identically. Therefore the current (31) reduces to

$$\begin{aligned} V^\mu(\mathcal{L},\tau)=&\frac{\partial L_{\mathrm{MAG}}}{\partial y^A_\mu}\,(u^{A\beta\gamma}_{\ \ \alpha}\partial_{\beta\gamma}\tau^\alpha+u^{A\alpha}_{\ \ \beta}\partial_\alpha\tau^\beta-y^A_\alpha\tau^\alpha)+\frac{\partial L_M}{\partial \sigma_c{}^\alpha{}_{,\mu}}\,(\sigma_c{}^\beta\partial_\beta\tau^\alpha-\sigma_c{}^\alpha{}_{,\beta}\tau^\beta)\\ &-\frac{\partial L_M}{\partial y^i_\mu}\,y^i_\alpha\tau^\alpha-\frac{\partial L_M}{\partial y^*_{i\mu}}\,y^*_{i\alpha}\tau^\alpha+\tau^\mu(L_{\mathrm{MAG}}+L_M). \end{aligned} \tag{32}$$

Due to the arbitrariness of the functions τ^α, the relations (30) imply the equalities

$$\delta^\alpha_\mu L_{\mathrm{MAG}}+\sigma_a{}^\alpha\,\frac{\partial L_{\mathrm{MAG}}}{\partial \sigma_a{}^\mu}+u^{A\alpha}_{\ \ \mu}\,\frac{\partial L_{\mathrm{MAG}}}{\partial y^A}+d_\beta u^{A\alpha}_{\ \ \mu}\,\frac{\partial L_{\mathrm{MAG}}}{\partial y^A_\beta}-y^A_\mu\,\frac{\partial L_{\mathrm{MAG}}}{\partial y^A_\alpha}=0, \tag{33}$$

and

$$\delta^\alpha_\mu L_M+\sigma_a{}^\alpha\,\frac{\partial L_M}{\partial \sigma_a{}^\mu}+\sigma_a{}^\alpha{}_{,\beta}\,\frac{\partial L_M}{\partial \sigma_a{}^\mu{}_{,\beta}}-\sigma_a{}^\beta{}_{,\mu}\,\frac{\partial L_M}{\partial \sigma_a{}^\beta{}_{,\alpha}}+u^{A\alpha}_{\ \ \mu}\,\frac{\partial L_M}{\partial y^A}-y^i_\mu\,\frac{\partial L_M}{\partial y^i_\alpha}-y^*_{i\mu}\,\frac{\partial L_M}{\partial y^*_{i\alpha}}=0. \tag{34}$$

Substituting the terms

$$\delta^\alpha_\mu L_{\mathrm{MAG}}-y^A_\mu\,\frac{\partial L_{\mathrm{MAG}}}{\partial y^A_\alpha}$$

and

$$\delta^\alpha_\mu L_M-\sigma_a{}^\beta{}_{,\mu}\,\frac{\partial L_M}{\partial \sigma_a{}^\beta{}_{,\alpha}}-y^i_\mu\,\frac{\partial L_M}{\partial y^i_\alpha}-y^*_{i\mu}\,\frac{\partial L_M}{\partial y^*_{i\alpha}}$$

from (33) and (34) into (32) and using the relations (23), (24), (28), and (29), the current $V^\mu(\mathcal{L},\tau)$ takes the superpotential form

$$V^\mu(\mathcal{L},\tau)\approx d_\beta U^{\beta\mu}(\mathcal{L},\tau),$$

where

$$U^{\beta\mu}(\mathcal{L},\tau)=\pi_\nu{}^{\kappa\beta\mu}(\nabla_\kappa\tau^\nu+T^\nu{}_{\alpha\kappa}\tau^\alpha). \tag{35}$$

Here ∇ is the covariant derivative with respect to the connection ω and T is its torsion.

Since

$$\pi_\nu{}^{\kappa\beta\mu}=(\sigma_a{}^\kappa\sigma_b{}^\mu\eta^{ab}\delta^\beta_\nu-\sigma_a{}^\kappa\sigma_b{}^\beta\eta^{ab}\delta^\mu_\nu)\ \det(\sigma^a{}_\mu),$$

(35) can be regarded as a generalized Komar superpotential, in the sense that if the connection ω has vanishing torsion and nonmetricity, then (35) coincides with the well-known Komar expression.[20]

ACKNOWLEDGMENTS

This work was supported by G.N.F.M.-C.N.R., by the National Research Project ‘‘Met. Geom. e Probab. in Fisica Matematica’’ of the M.U.R.S.T., and by the University of Camerino.

APPENDIX: LAGRANGIAN CONSERVATION LAWS

In this Appendix, the basic features of Lagrangian field theory in jet bundle terms are briefly introduced.[21,22]

Classical fields are represented by sections of a fiber bundle $Y\to X$ and their dynamics is described by means of jet manifolds. The first jet prolongation of $Y\to X$ is denoted by J^1Y. Given bundle coordinates (x^μ,y^i) on Y $(1\leqslant\mu\leqslant m=\dim X,\ 1\leqslant i\leqslant n, m+n=\dim Y)$, the induced coordinates on J^1Y are denoted by (x^λ,y^i,y^i_μ) and the vector fields along these coordinate directions are written (in compact notation) as $(\partial_\mu,\partial_i,\partial^\mu_i)$.

A (projectable) vector field

$$u:Y\to TY,\quad u=u^\mu(x)\partial_\mu+u^i(x,y)\partial_i\,,$$

on Y (an infinitesimal transformation of both the field and the base manifold variables y^i and x^μ, respectively) can be lifted to a (projectable) vector field $\overline{u}$ on J^1Y:

$$\overline{u}:J^1Y\to TJ^1Y,\quad \overline{u}=u^\mu\partial_\mu+u^i\partial_i+u^i_\mu\partial^\mu_i\,,\quad u^i_\mu=d_\mu u^i-y^i_\nu\partial_\mu u^\nu. \tag{A1}$$

Here

$$d_\mu=\partial_\mu+y^j_\mu\partial_j+\cdots$$

is the total derivative with respect to x^μ.

Let

$$\mathcal{L}:J^1Y\to\overset{m}{\wedge}T^*X,$$

$$\mathcal{L}=L(x^\mu,y^i,y^i_\mu)d^mx,\quad d^mx=dx^1\wedge\cdots\wedge dx^m,$$

be a first-order Lagrangian density. Given a vector field $u:Y\to TY$, the corresponding current is defined to be

$$V(\mathcal{L},u):J^1Y\to\overset{m-1}{\wedge}T^*X,\quad V(\mathcal{L},u)=V^\mu(\mathcal{L},u)\partial_\mu\rfloor d^mx,$$

$$V^\mu(\mathcal{L},u)=\partial^\mu_iL(u^i-y^i_\nu u^\nu)+u^\mu L.$$

By computing the Lie derivative $\mathbf{L}_{\overline{u}}\mathcal{L}$ of the Lagrangian density $\mathcal{L}$ with respect to the vector field (A1), one finds the relation

$$\mathbf{L}_{\bar{u}}\mathcal{L} \approx d_\mu V^\mu(\mathcal{L},u) d^m x, \tag{A2}$$

where the symbol $\approx$ stands for an equality valid on solutions of the field equations.

It follows that if the vector field u is an infinitesimal symmetry transformation of the Lagrangian density $\mathcal{L}$, i.e.,

$$\mathbf{L}_{\bar{u}}\mathcal{L} = 0,$$

then (A2) yields the differential conservation law

$$d_\mu V^\mu(\mathcal{L},u) \approx 0.$$

[1] C. J. Isham, A. Salam, and J. Strathdee, Ann. Phys. **62**, 98 (1971).
[2] S. Coleman, J. Wess, and B. Zumino, Phys. Rev. **177**, 965 (1969).
[3] G. Sardanashvily and O. Zakharov, *Gauge Gravitation Theory* (World Scientific, Singapore, 1992).
[4] L. Dabrowski and R. Percacci, Commun. Math. Phys. **106**, 691 (1986).
[5] S. Switt, J. Math. Phys. **34**, 3825 (1993).
[6] F. Hehl, J. McCrea, E. Mielke, and Y. Ne'eman, Phys. Rep. **258**, 1 (1995).
[7] Y. Ne'eman and Dj. Šijački, Ann. Phys. **120**, 292 (1979).
[8] G. Giachetta and G. Sardanashvily, Class. Quantum Grav. **13**, L67 (1996).
[9] G. Sardanashvily, J. Math. Phys. **33**, 1546 (1992).
[10] A. Trautman, *Differential Geometry for Physicists* (Bibliopolis, Naples, 1984).
[11] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Vol. I* (Wiley, New York, 1963).
[12] S. Hawking and G. Ellis, *The Large Scale Structure of a Space-Time* (Cambridge U.P., Cambridge, 1973).
[13] C. Dodson, *Categories, Bundles and Spacetime Topology* (Shiva, Orpington, 1980).
[14] R. Geroch, J. Math. Phys. **9**, 1739 (1968).
[15] P. Budinich and A. Trautman, *The Spinorial Chessboard* (Springer-Verlag, Berlin, 1988).
[16] H. Lawson and M-L. Michelson, *Spin Geometry* (Princeton U.P., Princeton, 1989).
[17] M. Greenberg, *Lectures on Algebraic Topology* (Benjamin, Menlo Park, 1971).
[18] R. O. Fulp, J. K. Lawson, and L. K. Norris, Int. J. Theor. Phys. **33**, 1011 (1994).
[19] G. Giachetta and L. Mangiarotti, Int. J. Theor. Phys. **29**, 789 (1990).
[20] A. Komar, Phys. Rev. **113**, 934 (1959).
[21] G. Giachetta, L. Mangiarotti, and G. Sardanashvily, *New Lagrangian and Hamiltonian Methods in Field Theory* (World Scientific, Singapore, 1997).
[22] G. Sardanashvily, *Gauge Theory in Jet Manifolds* (Hadronic, Palm Harbor, 1993).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Split structures in general relativity and the Kaluza–Klein theories

V. D. Gladush[a] and R. A. Konoplya
Department of Physics, Dnepropetrovsk State University, per. Nauchny 13, Dnepropetrovsk, 320625 Ukraine

(Received 4 May 1998; accepted for publication 26 October 1998)

We construct a general approach to the decomposition of the tangent bundle of pseudo-Riemannian manifolds into direct sums of subbundles, and the associated decomposition of geometric objects. An invariant structure $\mathcal{H}^r$ defined as a set of r projection operators is used to induce decomposition of the geometric objects into those of the corresponding subbundles. We define the main geometric objects characterizing decomposition. Invariant nonholonomic generalizations of the Gauss–Codazzi–Ricci's relations have been obtained. All the known types of decomposition (used in the theory of frames of reference, in the Hamiltonian formulation for gravity, in the Cauchy problem, in the theory of stationary spaces, and so on) follow from the present work as special cases when fixing a basis and dimensions of subbundles, and parametrization of a basis of decomposition. Various methods of decomposition have been applied here for the unified multidimensional Kaluza–Klein theory and for relativistic configurations of a perfect fluid. Discussing an invariant form of the equations of motion we have found the invariant equilibrium conditions and their 3+1 decomposed form. The formulation of the conservation law for the curl has been obtained in the invariant form. © *1999 American Institute of Physics.* [S0022-2488(99)01502-9]

I. INTRODUCTION

Most approaches and formalisms in General Relativity as well as in the multidimensional Unified Theories are connected with decomposition of spaces into direct sums of subspaces and the associated decomposition of geometrical objects. It means that in addition to the usual structures (differentiable structure, the metric structure, and so on) one should introduce a *split structure* which induces the decomposition of manifolds. This extra structure determines decomposition of all objects and structures defined on a manifold. Among the varieties of formalism of decomposition are the methods aimed to describe frames of reference and observable quantities in the theory of gravity. Similar methods have gained the wide acceptance in a great number of problems. Some of these problems are the canonical formalism for gravitational waves, the Cauchy problem in General Relativity, construction of the Unified Theory of interacting fields, quantization of the gravitational field, the tetradic formalism, the Newman–Penrose formalism, the theory of stationary and axisymmetric gravitational fields, the multidimensional and four-dimensional cosmologies.

Mathematicians and physicists developed methods of decomposition starting mainly from their intrinsic interests. It often took place independently and parallely, so that sometimes the same advances were overlooked and then refound.

The early stage of the development of mathematical technique for decomposition could be seen in the classical theory of hypersurfaces, and in the theory of n-dimensional surfaces imbedded in the n-dimensional Riemannian manifold.[1] Then, in the history of split methods, we can distinguish several ways. In mathematics, at the classical stage, there were constructed coordinate

[a]Electronic mail: gladush@ff.dsu.dp.ua

techniques for nonholonomic spaces and subspaces.[2,3] Owing to the use of the coordinate language such methods were rather cumbersome. In Physics split methods were induced by attempts to create the Unified Theory of fields, and, in particular, by appearance of the Kaluza–Klein theory.[4,5] This led to construction of a $4+1$ split method for a five-dimensional manifold and gave an impulse to study multidimensional Kaluza–Klein theories and multidimensional cosmologies[6–9] (see also references in Ref. 9).

Another physical branch of split methods was developed much more later than the original works on a $4+1$ split were. This branch has begun, apparently, with the work by Echart[10] and has been completed in the papers.[11–16] There have been constructed $(3+1)$, $(2+2)$, $(1+1+2)$ coordinate split methods and their special cases (see Ref. 17, and references therein).

The other independent direction to construct a split of spaces in General Relativity is connected with the questions of convenience of mathematical representation of the Einstein equations and with the study of these equations' structure. This branch is brought about by needs for construction of the canonical formalism,[18] of various projection formalisms in the theory of the stationary and axisymmetric gravitational fields[19–21] and for the positing of the Cauchy problem in General Relativity.[22]

Unfortunately, many of the works mentioned above, which have already become classical, use different and often inassociated approaches. Moreover a coordinate language applied there, especially in early works, makes it almost impossible to calculate the Einstein equations for some forms of the metric.

A new stage of development of split methods is based on modern differential geometry.[23,24] Its invariant language has become a working one in General Relativity.[25] It is not only a natural language for geometry in the whole, but also a convenient approach to calculations. Obtaining formulas is reduced, the formulas themselves become universal, and all the calculations can easily be made by computer.

The invariant split method was considered in Ref. 26 but without any connection with the previous works on a split. Objects introduced formally in this work have no clear geometrical meaning. One of us proposed the general invariant method of an $(n+m)$ split for pseudo-Riemannian manifolds.[27–30] Most of the approaches in this field were unified in the works,[27–30] and the objects introduced there have clear physical and geometrical meaning. For special cases of $(1+4)$, $(1+3)$, $(2+2)$, $(n+4)$ splits, in the coordinate representation, these objects reduce to known physical characteristics of a system.[8–16]

Multidimensional cosmologies and the Unified Theories imply that a manifold should be decomposed into more than two submanifolds. That is one reason why a split of a manifold requires the most general representation.

The theory of $(n_1+n_2+\cdots+n_r)$ decomposition of a pseudo-Riemannian manifold into the r nonholonomic orthogonal subbundles Σ_a of a dimension n_a $(a=1,2,\ldots,r)$ has been constructed in the present work $(n=n_1+n_2+\cdots+n_r$ is a dimension of a manifold). The $(n+m)$ and $(n+1)$ forms of invariant decomposition have been obtained as consequences. Choosing the projection operators and gauges of a basis of decomposition we construct various special cases. Some applications of this method are considered. Let us emphasize that we do not refer to problems of global geometry, but use its invariant formulations to construct decomposition of spaces into direct sums of subspaces.

Note, that the theory of structures mentioned above has found its further mathematical development, and now is widely known in differential geometry under the name ''almost product structure.''[26] The latter can be treated from a ''G-structure'' point of view[31] (see also Ref. 32, and references therein). We will follow a more natural approach and use the term ''a split structure,'' which is, in our opinion, more in the spirit of physical conceptions aroused in General Relativity when dealing with $(3+1)$, $(2+2)$, $(1+1+2)$ decomposition of space–time.

The plan of the paper is as follows. In Sec. II we introduce the necessary notations used in differential geometry, and the main definition of the theory, *a split structure* on a pseudo-Riemannian manifold. Then we introduce the metrics and connections induced on the corresponding subbundles as well as the main associated geometrical objects on the subbundles; the tensors of extrinsic curvature and of extrinsic torsion, analogies of the Ricci coefficients of rotation, and the curvature tensor. The invariant nonholonomic generalization of the Gauss–Codazzi–Ricci's relations has been found as various projections of the curvature tensor into every possible subbundle (see Appendix A).

In Sec. III the special case of the invariant split structure $r=2$ (when we deal with two subbundles only) is considered. In this case the generalized coefficients of rotation disappear from the curvature tensor, thereby the final formulas become much simpler. In Sec. IV the invariant formulas in the $(n+1)$ split form complete the scheme of the invariant split for a pseudo-Riemannian manifold. Further, for any concrete calculations, we must fix projection operators.

In Sec. V we, for illustration, briefly consider the $(n+m)$ and $(n+1)$ coordinate decomposition of the manifold with respect to the natural basis $\{\partial_\mu, dx^\nu\}$. In Sec. VI the method of $(n+m)$ decomposition is constructed with respect to an adopted basis. All the relations obtained in Sec. VI are basic ones for the other variants of decomposition in this paper. The final formulas (in Appendices B and C) can be used as an algorithm to compute the Ricci tensor, the Riemann tensor, the scalar curvature, and the corresponding Lagrangians.

In Sec. VII we define the canonical parametrization of a basis of decomposition. There have been obtained the main geometric objects with respect to this basis. Various well known special cases, which follow from this parametrization, are discussed in the section. Connections and relations among them are analyzed.

In Sec. VIII we obtain the decomposition induced by a given family of surfaces. In Sec. IX we consider the decomposition induced by a group of isometries. On the basis of this section's results we construct the Lagrangian of the unified multidimensional Kaluza–Klein theory (Sec. X). This decomposition, apart from everything else, serves as a methodical illustration of the possible use of the present method for physical theories.

Finally Sec. XI deals with the theory of configurations of a perfect fluid. Using the (3+1) canonical parametrization, one can define a one-form of the entalphy and a two-form of the curl. We have obtained the invariant equations of motion for a perfect fluid. The conservation law for the curl of an isentropic perfect fluid has been obtained in the invariant form. For this fluid, rotating in the stationary gravitational field, we have also deduced the equilibrium conditions and constructed its Lagrangian.

In this work we considered a torsion-free pseudo-Riemannian manifold only, nonetheless, our approach can be used without principal changes for theories of gravity with nonzero torsion. We see further development of the present theme in the possible expanding of the invariant decomposition to supergravity theories. We, mostly, used notations and definitions of the works.[23–25]

II. A SPLIT STRUCTURE ON A PSEUDO-RIEMANNIAN MANIFOLD

Let M be a pseudo-Riemannian manifold with the metric g; $T(M)=\cup_{p\in M}T_p$ and $T^*(M)=\cup_{p\in M}T_p^*$ are the tangent and cotangent bundles over M, where T_p and T_p^* are the corresponding fibers over a point p of M. The objects $X,Y,Z,\ldots\in T(M)$ and $\alpha,\beta,\omega,df\in T^*(M)$ denote contravariant and covariant vector fields (d is an exterior differential). We shall denote by $\omega(X)$ an inner product of a one-form ω and vector X. The scalar product of two vectors X,Y, and two forms α,β is determined by the metric g,

$$X\cdot\equiv(X,Y)\equiv g(X,Y);\ \ \langle\alpha,\beta\rangle\equiv g^{-1}(\alpha,\beta), \tag{2.1}$$

where g^{-1} is the inverse of the metric g.

We need to note that for each vector field $Y\in T(M)$ a dual one-form ω is determined uniquely by $\omega(X)=g(X,Y)$, $\forall X\in T(M)$. From now on we just will write $\omega=g(.,Y)$. Then the inverse of the metric g is given by

$$g^{-1}(\omega,\alpha)=g^{-1}(g(.,Y),\alpha)=\alpha(Y),\ \ \forall Y\in T(M),\ \ \forall\alpha\in T^*(M), \tag{2.2}$$

so that $Y=g^{-1}(.,w)$.

A linear operator L on $T(M)$ is a tensor of type (1,1) which acts according to the relation $L\cdot X \equiv L(X) \in T(M)$, $\forall X \in T(M)$. Then

$$(L^T\cdot\omega)(X)=(\omega\cdot L)(X)\equiv\omega(L(X)), \quad \forall X\in T(M), \tag{2.3}$$

where L^T is a transpose of an operator L.

The product of two linear operators $L\cdot H$ is defined by

$$(L\cdot H)\cdot X = L\cdot(H\cdot X)\in T(M), \quad \forall X\in T(M). \tag{2.4}$$

An operator H is called a symmetric one if

$$(H\cdot X,Y)=(X,H\cdot Y), \quad \forall X,Y\in T(M). \tag{2.5}$$

We have to introduce the new notation, *a split*, which denotes decomposition into direct sums. Therefore we shall say that *a split structure* $\mathcal{H}^r$ is introduced on M if the r linear symmetric operators H^a $(a=1,2,...,r)$ of a constant rank with the properties

$$H^a\cdot H^b=\delta^{ab}H^b; \quad \sum_{a=1}^{r} H^a = I, \tag{2.6}$$

where I is the unit operator ($I\cdot X=I$, $\forall X\in T(M)$), are defined on $T(M)$.

Now we introduce the notations

$$\Sigma_p^a \equiv \operatorname{Im} H_p^a; \ \ (\Sigma_a^*)_p\equiv \operatorname{Im}(H_p^a)^T; \ \ n_a=\dim\Sigma_p^a=\dim(\Sigma_a^*)_p\,, \tag{2.7}$$

where $\operatorname{Im} H_p^a$ is an image of an operator H^a at a point p of M, i.e., $\Sigma_p^a=\{X_p\in T_p | H^a\cdot X_p=X_p\}$. It is important that owing to constancy of a rank of the operator H^a, dimension n_a does not depend on a point p of M.

From the definitions presented here we can obtain the decomposition of the tangent and cotangent spaces,

$$T_p=\bigoplus_{a=1}^{r}\Sigma_p^a; \quad T_p^*=\bigoplus_{a=1}^{r}(\Sigma_a^*)_p; \quad \dim T_p=\dim T_p^*=\sum_{a=1}^{r} n_a\,, \tag{2.8}$$

where the sign $\oplus$ denotes the direct sum. Thus the tensors $\{H^a\}$ are the projection operators, which bring about decomposition of the fibers T_p, T_p^* into the r local subspaces Σ_p^a and $(\Sigma_a^*)_p$, respectively. By the same way, the bundles $T(M)$ and $T^*(M)$ are decomposed into the $(n_1+n_2+\cdots+n_r)$ subbundles Σ^a, Σ_a^*, so that

$$T(M)=\bigoplus_{a=1}^{r}\Sigma^a; \quad T^*(M)=\bigoplus_{a=1}^{r}\Sigma_a^*; \quad \Sigma^a=\bigcup_{p\in M}\Sigma_p^a; \quad \Sigma_a^*=\bigcup_{p\in M}(\Sigma_a^*)_p\,. \tag{2.9}$$

Then arbitrary vectors, covectors, and metrics are decomposed according to the scheme,

$$X=\sum_{a=1}^{r}X^a, \quad \alpha=\sum_{a=1}^{r}\alpha_a\,, \quad g=\sum_{a=1}^{r}g^a, \quad g^{-1}=\sum_{a=1}^{r}g_a^{-1}, \tag{2.10}$$

where

$$X^a=H^a\cdot X^a=H^a\cdot X; \quad H^b\cdot X^a=0; \quad X^a\cdot X^b=0; \quad (a\neq b), \tag{2.11}$$

$$\alpha_a=\alpha_a\cdot H^a=\alpha\cdot H^a; \quad \alpha_a\cdot H^b=0; \quad \alpha_a(X^b)=0 \quad (a\neq b),$$

$$g^a(X^a,Y^a)\equiv g(X^a,Y^a);\quad g_a^{-1}(\alpha_a,\beta_a)\equiv g^{-1}(\alpha_a,\beta_a),\tag{2.12}$$

$$\forall X^a,Y^a\in\Sigma^a,\quad \forall\alpha_a,\beta_a\in\Sigma_a^*.$$

In these relations $\{g^a\}$ are the metrics induced on the subbundles $\{\Sigma^a\}$ of the tangent bundle $T(M)$. Using this scheme we can obtain the decomposition of more complex tensors. We assume that all objects with indices $a,b,\dots$ are defined on the associated subbundles $\Sigma^a,\Sigma^b,\dots$.

Let ∇ be an affine (symmetric and compatible with g) connection such that

$$\nabla_X Y-\nabla_Y X=[X,Y],\quad X(Y\cdot Z)=Z\cdot\nabla_X Y+Y\cdot\nabla_X Z,\tag{2.13}$$

where $[X,Y]=XY-YX$ is the Lie bracket of two vector fields X and Y, $\nabla_X Y$ is the covariant derivative of Y in the direction X. A consequence of this is that

$$2Z\cdot\nabla_X Y=X(Y\cdot Z)+Y(Z\cdot X)-Z(X\cdot Y)+Z\cdot[X,Y]+Y\cdot[Z,X]-X\cdot[Y,Z].\tag{2.14}$$

Then the covariant derivative $\nabla_X T$ of a tensor T of type (s,r), where $s=0,1$ with respect to X is defined by

$$(\nabla_X T)(Y_1,\dots,Y_r)=\nabla_X(T(Y_1,\dots,Y_r))-\sum_{i=1}^{r}T(Y_1,\dots,Y_{i-1},\nabla_X Y_i,Y_{i+1},\dots,Y_r).\tag{2.15}$$

The Lie derivative $\mathcal{L}_X T$ of a tensor T with respect to a vector X and the exterior derivative of an r-form Ω are given by

$$(\mathcal{L}_X T)(Y_1,\dots,Y_r)=\mathcal{L}_X(T(Y_1,\dots,Y_r))-\sum_{i=1}^{r}T(Y_1,\dots,Y_{i-1},\mathcal{L}_X Y_i,Y_{i+1},\dots,Y_r),$$

$$(d\Omega)(Y_0,Y_1,\dots,Y_r)=\sum_{i=0}^{r}(-1)^i Y_i(\Omega(Y_0,\dots,\hat{Y}_i,\dots,Y_r))$$

$$+\sum_{0\le i<j\le r}(-1)^{i+j}\Omega([Y_i,Y_j],Y_0,\dots,\hat{Y}_i,\dots,\hat{Y}_j,\dots,Y_r),\tag{2.16}$$

where $\mathcal{L}_X Y=[X,Y]$, $\mathcal{L}_X\varphi=\nabla_X\varphi=(d\varphi)(X)=X\varphi$ for any scalar function φ; the symbol $\hat{}$ means that the associated term is omitted.

The curvature tensor is defined by the formula

$$R(X,Y)Z=(\nabla_X\nabla_Y-\nabla_Y\nabla_X-\nabla_{[X,Y]})Z.\tag{2.17}$$

Using a split structure $\mathcal{H}^r$, the decomposition of ∇ is easily set up,

$$\nabla_X Y=\sum_{a,b,c=1}^{r}\nabla^a_{X^b}Y^c,\quad \forall X,Y\in T(M).\tag{2.18}$$

In this sum we can distinguish the five different sorts of objects $\{\nabla^a_{X^a}Y^a,\ \nabla^a_{X^b}Y^b,\ \nabla^a_{X^a}Y^b,\ \nabla^a_{X^b}Y^a,\ \nabla^a_{X^b}Y^c\}$ $(a\neq b\neq c)$, which complete the whole of the projected connections. In particular, in this sum the objects

$$\nabla^a_{X^a}Y^a\equiv H^a\cdot\nabla_{X^a}Y^a,\quad \forall X^a,Y^a\in\Sigma^a\quad(a=1,2,...,r)\tag{2.19}$$

define connections $\{\nabla^a\}$ induced on the subbundles $\{\Sigma^a\}$. The object

$$\nabla^a_{X^b}Y^b\equiv H^a\cdot\nabla_{X^b}Y^b\equiv -B^a(X^b,Y^b),\quad \forall X^b,Y^b\in\Sigma^b\tag{2.20}$$

is the tensor of extrinsic nonholonomicity of the subbundle Σ^b. One can think that the objects

$$\nabla^a_{X^b} Y^c \equiv H^a \cdot \nabla_{X^b} Y^c \equiv -Q^a(X^b,Y^c) \equiv -Q^a_{bc}(X^b,Y^c), \quad \forall a \neq b \neq c \tag{2.21}$$

define the generalization of the Ricci's coefficients of rotation γ^a_{bc}.[34] In the general case they give the objects of rotation Q^a_{bc} of the subbundles Σ^b, Σ^c in the n_a-dimensional direction Σ^a. The other components can be expressed in terms of the introduced objects and the Lie derivative $\mathcal{L}_{X^b} Y^c$ projected into every possible subbundle Σ^a $(a \neq b \neq c)$. Thus, the components $\nabla^a_{X^a} Y^b \equiv H^a \cdot \nabla_{X^a} Y^b$ and $\nabla^a_{X^b} Y^a \equiv H^a \cdot \nabla_{X^b} Y^a$ satisfy the relations

$$Z^a \cdot \nabla^a_{X^a} Y^b = Y^b \cdot B^b(X^a,Z^a) \quad (a \neq b), \tag{2.22}$$

$$Z^a \cdot \nabla^a_{X^b} Y^a = Z^a \cdot \Lambda^a(X^b,Y^a) + X^b \cdot B^b(Y^a,Z^a), \tag{2.23}$$

where

$$\Lambda^a(X^b,Y^c) = [X^b,Y^c]^a \equiv H^a \cdot [X^b,Y^c] \equiv \mathcal{L}^a_{X^b} Y^c \quad (a \neq b \neq c). \tag{2.24}$$

Taking into account the relation (2.14) and the definition (2.21) we have

$$2Z^a \cdot Q^a(X^b,Y^c) = X^b \cdot \Lambda^b(Y^c,Z^a) - Z^a \cdot \Lambda^a(X^b,Y^c) - Y^c \cdot \Lambda^c(Z^a,X^b).$$

The tensor of extrinsic nonholonomicity can be expressed as the sum of symmetric and antisymmetric parts

$$B^a(X^b,Y^b) = S^a(X^b,Y^b) + A^a(X^b,Y^b), \tag{2.25}$$

where $S^a(X^b,Y^b)$ and $A^a(X^b,Y^b)$ define the tensors of extrinsic curvature and extrinsic torsion of subbundle Σ^b in the direction of the subbundle Σ^a. For these objects one can obtain the relations

$$2Z^a \cdot S^a(X^b,Y^b) = (\mathcal{L}_{Z^a} g^b)(X^b,Y^b), \tag{2.26}$$

$$2A^a(X^b,Y^b) = -[X^b,Y^b]^a \equiv -H^a \cdot [X^b,Y^b]. \tag{2.27}$$

It can easily be shown that a connection ∇^a induced on the subbundle Σ^a will be symmetric and compatible with the metric g^a. The projecting of the curvature tensor into the subbundles $\Sigma^a, \Sigma^b, \ldots$ gives us the nonholonomic generalizations of the Gauss–Codazzi–Ricci's equations. Using the definitions (2.10)–(2.12), (2.17)–(2.27) one can obtain all the necessary projections of the curvature tensor (for more details see Appendix A).

III. INVARIANT $n+m$ DECOMPOSITION OF A PSEUDO-REIMANNIAN MANIFOLD

If $r=2$, then there are only two subbundles Σ' and Σ'' of the tangent bundle $T(M)$ and the previous formulas become much simpler. Owing to the importance of this case it was deemed worthwhile to consider the split structure independently from Sec. II.[29,30]

Let H' be a linear idempotent symmetric operator of a constant rank with the property

$$H' \cdot H' = H'. \tag{3.1}$$

We shall say that H' defines *a* $(n+m)$ *split structure* on M if

$$\dim \operatorname{Im} H' = n; \quad \dim \operatorname{Ker} H' = m; \quad \dim M = n+m, \tag{3.2}$$

where $\operatorname{Ker} H'$ is a kernel of the operator H'. Since H' is defined, thereby we define the operator H'' such that

$$H''\cdot H''=H''; \quad H''\cdot H'=H'\cdot H''=0; \quad H'+H''=I. \tag{3.3}$$

Therefore H' and H'' are the projection operators which determine the split structure $\mathcal{H}^2$ on M due to the definition (2.6). We have

$$T(M)=\Sigma'\oplus\Sigma''; \quad \Sigma'=\bigcup_{p\in M}\Sigma'_p; \quad \Sigma'_p=\operatorname{Im}H'_p; \quad \Sigma''=\bigcup_{p\in M}\Sigma''_p; \quad \Sigma''_p=\operatorname{Ker}H'_p; \tag{3.4}$$

$$X=X'+X''; \quad g=g'+g''; \quad g^{-1}=(g')^{-1}+(g'')^{-1};$$

$$X'=H'\cdot X; \quad X''=H''\cdot X; \quad X'\cdot X''=0;$$

$$g'(X',Y')=g(X',Y'); \quad g''(X'',Y'')=g(X'',Y''). \tag{3.5}$$

A connection ∇ is decomposed into the following components: a connection on Σ', and the tensor of extrinsic nonholonomicity of the subbundle Σ', respectively,

$$\nabla'_{X'}Y'=H'\cdot\nabla_{X'}Y', \tag{3.6}$$

$$B''(X',Y')=-\nabla''_{X'}Y'=-H''\cdot\nabla_{X'}Y'. \tag{3.7}$$

Other components of ∇ can be expressed in terms of the components (3.6), (3.7) and the Lie derivatives of two vector fields

$$X'\cdot\nabla'_{Y'}Z''=Z''\cdot B''(Y',X'), \tag{3.8}$$

$$X'\cdot\nabla'_{Y''}Z'=X'\cdot\mathcal{L}_{Y''}Z'+Y''\cdot B''(Z',X'). \tag{3.9}$$

The rest of the components of ∇ $\{\nabla''_{X''}Y'', \nabla'_{X''}Y'', \nabla''_{X''}Y', \nabla''_{X'}Y''\}$ may be written out by substituting $X',Y',B',H',\ldots$ for $X'',Y'',B'',H'',\ldots$ and vice versa in formulas (3.6)–(3.9). This completes the set of all the eight possible projections of the connection.

The tensor B'' may be expressed as the sum of its symmetric and antisymmetric parts,

$$B''(X',Y')=S''(X',Y')+A''(X',Y'), \tag{3.10}$$

$$2Z''\cdot S''(X',Y')=(\mathcal{L}_{Z''}g')(X',Y'); \quad 2A''(X',Y')=-H''\cdot[X',Y'], \tag{3.11}$$

where S'' and A'' are the tensors of extrinsic curvature and torsion, respectively. If $A''=0$, the subbundle Σ' will be holonomic. It means that the subbundle Σ' is the union of the tangent bundles of an m-parameter family of n-dimensional surfaces $\{M^n(q)\subset M\}$, where $q=\{c^i\}\in D\subset R^m$ parametrizes the surfaces $M^n(q)$, and D is some range of parameters c^i $(i=1,2,...,m)$ in R^m, that is $\Sigma'=\cup_{q\in D}T(M^n(q))$. This implies that a covector basis of locally exact one-forms $\{dx^i\}$ exists on the dual subbundle $(\Sigma'')^*$, so that each of the surfaces of $\{M^n(q)\}$ is the intersection of hypersurfaces $x^i=c^i$ $(i=1,2,...,m)$ for some values of $c^i\in D$.

Using the definition of the curvature tensor (2.17) one can find every possible projection of the curvature tensor

$$\begin{aligned}R(X',Y')Z'\cdot V'={}&R'(X',Y')Z'\cdot V'+B''(X',Z')\cdot B''(Y',V')\\&-B''(Y',Z')\cdot B''(X',V')+2A''(X',Y')\cdot B''(Z',V');\end{aligned} \tag{3.12}$$

$$R(X',Y')Z'\cdot V''=V''\cdot\{(\nabla''_{Y'}B'')(X',Z')-(\nabla''_{X'}B'')(Y',Z')\}+2Z'\cdot B'(A''(X',Y'),V''); \tag{3.13}$$

$$R(X',Y'')Z'\cdot V''=(Z'\cdot(\nabla'_{X'}B')+\langle X'\cdot B',Z'\cdot B'\rangle)(Y'',V'')$$
$$+(V''\cdot(\nabla''_{Y''}B'')+\langle Y''\cdot B'',V''\cdot B''\rangle)(X',Z'); \tag{3.14}$$

R' is the curvature tensor of the subbundle Σ'

$$R'(X',Y')Z'\equiv\{\nabla'_{X'}\nabla'_{Y'}-\nabla'_{Y'}\nabla'_{X'}-\nabla'_{[X',Y']'}+2\mathcal{L}'_{A''(X',Y')}\}Z',\quad \forall X',Y',Z'\in\Sigma', \tag{3.15}$$

where $\mathcal{L}'$ is the Lie derivative projected into the subbundle Σ' $(\mathcal{L}'_X Y\equiv H'\cdot\mathcal{L}_X Y)$. This definition of the curvature tensor, introduced in the works,[28–30] is the invariant generalization of that introduced in coordinate form in Refs. 11–13 by analogy with Ref. 2. Note that the latter term in (3.15) is necessary in order that the differential curvature operator $R'(X',Y')$ on Σ' be a linear multiplicative one, or, in other words, R' be a tensor of type (1,3) on the nonholonomic subbundle Σ'. In a similar fashion this concerns the general case of $\mathcal{H}^r$ split structure [see Appendix A, (A6) for $R^a(X^a,Y^a)Z^a$].

The following expression, with the fixed vectors X',Z',Y'',V'',

$$(\langle Y''\cdot B'',V''\cdot B''\rangle)(X',Z')\equiv\langle Y''\cdot B''(X',.),V''\cdot B''(.,Z')\rangle$$

defines the scalar product of the two one-forms $\alpha\equiv Y''\cdot B''(X',.)$ and $\beta\equiv V''\cdot B''(.,Z')$ according to (2.1) by the metric $(g')^{-1}$. The covariant derivatives of the tensor B' are given by

$$(\nabla'_{X'}B')(Y'',Z'')=\nabla'_{X'}(B'(Y'',Z''))-B'(\nabla''_{X'}Y'',Z'')-B'(Y'',\nabla''_{X'}Z''), \tag{3.16}$$

$$(\nabla'_{X''}B')(Y'',Z'')=\nabla'_{X''}(B'(Y'',Z''))-B'(\nabla''_{X''}Y'',Z'')-B'(Y'',\nabla''_{X''}Z''). \tag{3.17}$$

The relations (3.12)–(3.14) are nonholonomic analogs of the well-known Gauss–Codazzi–Ricci's equations. Other nontrivial projections of the curvature tensor may be written out using the substitution "′" for "″" and vice versa.

In the special case of the coordinate representation of (3+1) and (2+2) decomposition, the objects introduced above give us the known tensors,[11–16] which have clear physical and geometrical meaning.

Let us note that the objects, presented in the work[26] may be expressed in terms of these tensors. For example, the torsion tensor introduced there as the Nijenhuis tensor[24] proved to be equal

$$S_{H'}(X,Y)\equiv[X,Y']'+[X',Y]'-[X',Y']-[X,Y]'=2A'(X'',Y'')+2A''(X',Y')$$

and the tensors T_XY and Q_XY of the work[26] are given by

$$T_XY\equiv\nabla''_{X'}Y'+\nabla'_{X'}Y''=-B''(X',Y')+g^{-1}(Y''\cdot B''(X',.),.),$$

$$Q_XY\equiv\nabla'_{X''}Y''+\nabla''_{X''}Y'=-B'(X'',Y'')+g^{-1}(Y'\cdot B'(X'',.),.).$$

They have not any simple interpretation even in the classical case of the hypersurfaces in M.

IV. AN INVARIANT $(n+1)$ SPLIT STRUCTURE ON A PSEUDO-RIEMANNIAN MANIFOLD

In this section we give the invariant generalization of $(n+1)$ decomposition of spaces (the monad method[13,15]) as a special case of $(n+m)$ decomposition when $m=1$.

Let u be a vector field (field of a monad) on M such that $u\cdot u=\epsilon=\pm 1$. It gives a one-form ω and projection operators uniquely by the formulas

$$\omega(X)=\epsilon u\cdot X,\quad \forall X\in T(M), \tag{4.1}$$

$$H''=u\otimes\omega;\quad H'=I-H''. \tag{4.2}$$

The operators satisfy all the necessary relations (3.1)–(3.3) when Σ'' is a one-dimensional subbundle $(m=1)$. The tensor product is denoted by ''$\otimes$.''

Thus, defining vector or covector fields, u or ω, respectively, we, thereby, induce an $(n+1)$ split structure on M. For any vector field X and metric g, this implies

$$X=X'+\omega(X)u,\quad g=g'+\epsilon\omega\otimes\omega,\quad g^{-1}=(g')^{-1}+\epsilon u\otimes u. \tag{4.3}$$

Hence it is apparent that $X''=\omega(X)u$ is collinear to u. The metrics $g''=\epsilon\omega\otimes\omega$, $(g'')^{-1}=\epsilon u\otimes u$ and g', $(g')^{-1}$ are the metrics on the subbundles $\Sigma'',\Sigma^{*}{}''$ and $\Sigma',\Sigma^{*}{}'$, correspondingly. A connection ∇ has the following components:

$$\nabla_{X'}Y'=\nabla'_{X'}Y'-B(X',Y')u;\quad \nabla_u u=\nabla'_u u=-B'(u,u)\equiv F, \tag{4.4}$$

where $B(X',Y')=\omega(B''(X',Y'))$. The latter equation in (4.4) follows from the formula $u\cdot u=\epsilon=\pm1$. If we consider a congruence of curves for which the vector u is the tangent vector, then F is the first curvature of this congruence. The tensor B of type (0,2) is the tensor of extrinsic nonholonomicity of the subbundle Σ' and can be written as the sum of its symmetric and antisymmetric parts,

$$B(X',Y')=-\omega(\nabla''_{X'}Y')=\epsilon S(X',Y')+A(X',Y'), \tag{4.5}$$

where $S(X',Y')=\epsilon\omega(S''(X',Y'))$, $A(X',Y')=\omega(A''(X',Y'))$, and

$$2S(X',Y')=(\mathcal{L}_u g')(X',Y');\quad 2A(X',Y')=(d\omega)(X',Y') \tag{4.6}$$

are the tensors of extrinsic curvature and extrinsic torsion of the subbundle Σ'.

The components of the curvature tensor in an $(n+1)$ decomposed form lead to the generalized Gauss–Codazzi–Ricci's equations,

$$R(X',Y')Z'\cdot V'=R'(X',Y')Z'\cdot V'+\epsilon[2A(X',Y')B(Z',V')+B(X',Z')B(Y',V')-B(Y',Z')(X',V')], \tag{4.7}$$

$$R(X',Y')Z'\cdot u=-2A(X',Y')F\cdot Z'+\epsilon[(\nabla_{Y'}B)(X',Z')-(\nabla_{X'}B)(Y',Z')], \tag{4.8}$$

$$R(X',u)Y'\cdot u=-Y'\cdot\nabla'_{X'}F+\epsilon(F\cdot X')(F\cdot Y')+(\epsilon\mathcal{L}_u B-\langle B,B^T\rangle)(X',Y'), \tag{4.9}$$

where the curvature tensor of the subbundle Σ' (see Ref. 27) is given by

$$R'(X',Y')Z'\equiv\{\nabla'_{X'}\nabla'_{Y'}-\nabla'_{Y'}\nabla'_{X'}-\nabla'_{[X',Y']'}+2A(X',Y')\mathcal{L}'_u\}Z'. \tag{4.10}$$

It is to be noted that for tensors of an arbitrary type the projection operators are constructed by the tensor product of the operators (4.2) and their transposes. If one does no more than $(n+1)$ decomposition of objects only from the Cartan algebra of exterior forms on M then the universal invariant construction of the projection operators is feasible (see, for example, Ref. 33 for the (3+1) decomposition).

V. $(n+m)$ DECOMPOSITION OF A PSEUDO-RIEMANNIAN MANIFOLD IN A COORDINATE FORM

In order to obtain a coordinate form of the invariant objects it is necessary to choose coordinate covector and vector bases $\{\partial_\mu=\partial/\partial x^\mu\}$, $\{dx^\mu\}$ in the domain U of some map x^μ $(\mu,\nu,\rho,\ldots=1,2,\ldots,n,n+1,\ldots,n+m)$. Then we can find all the relations given above with respect to this basis, i.e., in covariant form.

Thus in the case of an $(n+m)$ decomposition one has

$$H'=h'^{\nu}_{\mu}\partial_\nu\otimes dx^\mu=h'^{\nu}_{\mu}\partial'_\nu\otimes d'x^\mu;\quad \partial'_\mu\equiv h'^{\nu}_{\mu}\partial_\nu,\quad d'x^\mu\equiv h'^{\mu}_{\nu}dx^\nu,$$

$$H''=h''^{\nu}_{\mu}\partial_\nu\otimes dx^\mu=h''^{\nu}_{\mu}\partial''_\nu\otimes d''x^\mu;\quad \partial''_\mu\equiv h''^{\nu}_{\mu}\partial_\nu,\quad d''x^\mu\equiv h''^{\mu}_{\nu}dx^\nu,$$

$$h'^{\nu}_{\mu}h'^{\mu}_{\rho}=h'^{\nu}_{\rho};\quad h''^{\nu}_{\mu}h''^{\mu}_{\rho}=h''^{\nu}_{\rho};\quad h'^{\nu}_{\mu}h''^{\mu}_{\rho}=0;\quad h'^{\nu}_{\mu}+h''^{\nu}_{\mu}=\delta^\nu_\mu, \tag{5.1}$$

$$g=g'+g''=g'_{\mu\nu}d'x^\mu\otimes d'x^\nu+g''_{\mu\nu}d''x^\mu\otimes d''x^\nu,$$

$$g_{\mu\nu}\equiv\partial_\mu\cdot\partial_\nu=g'_{\mu\nu}+g''_{\mu\nu};\quad g'_{\mu\nu}=h'^{\rho}_{\mu}h'^{\sigma}_{\nu}g_{\rho\sigma}\quad g''_{\mu\nu}=h''^{\rho}_{\mu}h''^{\sigma}_{\nu}g_{\rho\sigma}. \tag{5.2}$$

Further, introducing the definitions

$$[\partial'_\mu,\partial'_\nu]'\equiv\lambda^{\rho'}_{\mu'\nu'}\partial'_\rho;\quad [\partial'_\mu,\partial''_\nu]'\equiv\lambda^{\rho'}_{\mu'\nu''}\partial'_\rho;\quad [\partial''_\mu,\partial''_\nu]'\equiv-2A^{\rho'}_{\mu''\nu''}\partial'_\rho, \tag{5.3}$$

$$\nabla'_{\partial'_\mu}\partial'_\nu\equiv L^{\rho'}_{\mu'\nu'}\partial'_\rho;\quad B'\equiv B^{\rho'}_{\mu''\nu''}\partial'_\rho\otimes d''x^\mu\otimes d''x^\nu; \tag{5.4}$$

one has

$$L^{\rho'}_{\mu'\nu'}=d'x^\rho(\nabla_{\partial'_\mu}\partial'_\nu);\quad B^{\rho'}_{\mu''\nu''}=dx'^{\rho}(\nabla_{\partial''_\mu}\partial''_\nu)=S^{\rho'}_{\mu''\nu''}+A^{\rho'}_{\mu''\nu''}, \tag{5.5}$$

$$2A^{\rho'}_{\mu''\nu''}=h''^{\omega}_{\mu}h''^{\gamma}_{\nu}(h'^{\rho}_{\gamma,\omega}-h'^{\rho}_{\omega,\gamma});\quad 2S_{\rho'\mu''\nu''}=\partial'_\rho g''_{\mu\nu}+g''_{\mu\sigma}\lambda^{\sigma''}_{\nu''\rho'}+g''_{\sigma\nu}\lambda^{\sigma''}_{\mu''\rho'}. \tag{5.6}$$

Here $h_\gamma\equiv\partial h/\partial x^\gamma$; $\mu',\nu',\rho',\ldots,\mu'',\nu'',\rho'',\ldots=1,2,...,n,n+1,...,n+m$. The indices "$'$" and "$''$" indicate that the corresponding objects are associated with the subbundles Σ' and Σ'', respectively. From the previous formulas it follows that there are the objects which are associated with both subbundles Σ' and Σ''. For instance, the tensor of extrinsic nonholonomicity $B^{\rho'}_{\mu''\nu''}$ is a contravariant vector on the subbundle Σ', and a covariant tensor of rank 2 on the subbundle Σ''.

The other necessary objects can be found by substituting "$'$" for "$''$" and vice versa. Using these formulas we can obtain the Gauss–Codazzi–Ricci's equations in terms of the introduced objects. If $n=m=2$, our treatment is reduced to the dyadic formalism (see Ref. 16).

In the case of an $(n+1)$ split structure (see Sec. IV), we have $u=u^\mu\partial_\mu$, $(\mu,\nu=1,2,...,n,n+1)$, and

$$u_\mu u^\mu=\epsilon=\pm1,\quad h''^{\nu}_{\mu}=\epsilon u_\mu u^\nu,\quad h'^{\nu}_{\mu}=\delta^\nu_\mu-\epsilon u_\mu u^\nu,$$

$$g_{\mu\nu}=\epsilon u_\mu u_\nu+g'_{\mu\nu};\quad g^{\mu\nu}=\epsilon u^\mu u^\nu+g'^{\mu\nu}, \tag{5.7}$$

$$g'_{\mu\nu}=h'^{\alpha}_{\mu}h'^{\beta}_{\nu}g_{\alpha\beta};\quad g'^{\mu\nu}=h'^{\mu}_{\alpha}h'^{\nu}_{\beta}g^{\alpha\beta}, \tag{5.8}$$

$$\partial'_\mu=h'^{\nu}_{\mu}\partial_\nu;\quad [\partial'_\mu,\partial'_\nu]=\epsilon A_{\mu\nu}u;\quad [\partial'_\mu,u]=-F_\mu u, \tag{5.9}$$

$$2A_{\mu\nu}=h'^{\rho}_{,\mu}h'^{\sigma}_{,\nu}(u_{\rho,\sigma}-u_{\sigma,\rho});\quad F_\mu=(u_{\mu,\nu}-u_{\nu,\mu})u^\nu;\quad 2S_{\mu\nu}=\mathcal{L}_u g'_{\mu\nu}. \tag{5.10}$$

Replacing all the objects in (4.7)–(4.9) by these relations we can find the components of the curvature tensor. Furthermore if we consider the $(3+1)$ decomposition of a relativistic space–time, our formalism is reduced to the monad method,[11–15] and to his special gauges. In this case abstract geometrical objects will have an explicit physical meaning. So, one can think of $A_{\mu\nu}$ as the local angular velocity tensor of the frame of reference. The first curvature vector of the congruence F_μ determines the acceleration of the reference body in a given point, and $S_{\mu\nu}$ is the rate of strain tensor.[17]

VI. $(n+m)$ DECOMPOSITION WITH RESPECT TO AN ADOPTED BASIS

To find all the relations considered above in an $(n+m)$ decomposed form for some fixed basis is a question of great significance for applications. The coordinate form of $(n+m)$ decomposition considered in Sec. V is rather cumbersome, and the objects themselves prove to be singular. One of the reasons of this is that the range of indices $\mu', \mu'', \dots$ is redundant. Therefore it is more convenient for applications to choose the adopted bases of $(n+m)$ decomposition which will eliminate such redundancy. One's choice of one basis or another is dictated by a physical situation, requirements of an interpretation of results, or just by the necessity to use the most comfortable way of calculation. We shall present here the invariant relations of Sec. III with respect to an adopted basis of decomposition. Note that in such a form the formulas will be quite feasible for any concrete basis of decomposition. All the known types of decomposition (for torsion free theories) can be obtained as special cases of the present formalism by choosing the corresponding bases. In an $(n+m)$ decomposed form our method is essentially useful for calculation of the Riemann tensor, the Ricci tensor, and the curvature scalar by computer.

We shall now consider two adopted dual bases of decomposition; a vector one $\{E_\mu\} = \{E_a, E_i\}$ on $T(M)$, and a covector basis $\{\theta^\mu\} = \{\theta^a, \theta^i\}$ on $T^*(M)$, where $E_b \in \Sigma' \equiv \Sigma^n$, $E_i \in \Sigma'' \equiv \Sigma^m$; $\theta^a \in \Sigma^{*\prime} \equiv \Sigma^{*n}$; $\theta^i \in \Sigma^{*\prime\prime} \equiv \Sigma^{*m}$. According to (3.4)–(3.5) we have

$$\theta^a(E_b) = \delta^a_b, \quad \theta^a(E_j) = 0; \quad \theta^i(E_b) = 0; \quad \theta^i(E_k) = \delta^i_k, \tag{6.1}$$

$$E_b \cdot E_k = 0, \quad \langle \theta^a, \theta^i \rangle = 0. \tag{6.2}$$

It should be emphasized that the indices $a, b, c, \dots$ and $i, j, k, \dots$ indicate the subbundles Σ^n, Σ^{*n} and Σ^m, Σ^{*m}, respectively. With respect to the basis $\{E_\mu\}$, $\{\theta^\mu\}$ one has

$$H' = E_a \otimes \theta^a; \quad H'' = E_i \otimes \theta^i; \quad g = g' + g'' = \gamma_{ab}\theta^a \otimes \theta^b + h_{ik}\theta^i \otimes \theta^k, \tag{6.3}$$

where $\gamma_{ab} = E_a \cdot E_b$ and $h_{ik} = E_i \cdot E_k$ are the components of the metrics g', g'' induced on the subbundles Σ^n and Σ^m.

Then we introduce the definitions

$$\begin{aligned} \nabla'_{E_a} E_b &= L^c_{ab} E_c; \quad \nabla''_{E_i} E_j = L^k_{ij} E_k; \\ B'(E_i, E_k) &= B^a_{ik} E_a; \quad B''(E_a, E_b) = B^i_{ab} E_i, \end{aligned} \tag{6.4}$$

$$\begin{aligned} [E_a, E_b]' &= \lambda^c_{ab} E_c; \quad [E_i, E_j]'' = \lambda^k_{ij} E_k; \\ [E_a, E_i]' &= \lambda^b_{ai} E_b; \quad [E_i, E_a]'' = \lambda^k_{ia} E_k, \end{aligned} \tag{6.5}$$

where L^c_{ab} and L^i_{jl} are the coefficients of connections ∇' induced on Σ^n and ∇'' induced on Σ^m. Similarly B^c_{ik} and B^i_{ab} are the coefficients of the tensors of extrinsic nonholonomicity of the subbundles Σ^m and Σ^n, respectively. Using the identity (2.14) one can find

$$\begin{aligned} L^c_{ab} &= \Delta^c_{ab} + \gamma^c_{ab}; \quad L^i_{jk} = \Delta^i_{jk} + \gamma^i_{jk}; \\ B^a_{ik} &= S^a_{ik} + A^a_{ik}; \quad B^i_{ab} = S^i_{ab} + A^i_{ab}, \end{aligned} \tag{6.6}$$

where

$$2\Delta_{cab} = E_a\gamma_{bc} + E_b\gamma_{ac} - E_c\gamma_{ab}; \quad 2\gamma_{cab} = \lambda_{cab} + \lambda_{bca} - \lambda_{abc}, \tag{6.7}$$

$$\begin{aligned} 2S_{aik} &= (\mathcal{L}_{E_a} g'')(E_i, E_k) = E_a h_{ik} + \lambda_{ika} + \lambda_{kia}; \\ 2A^a_{ik} &= (d\theta^a)(E_i, E_k); \quad 2A_{aik} = -E_a \cdot [E_i, E_k]. \end{aligned} \tag{6.8}$$

The coefficients $A_{iab}, S_{iab}, \gamma_{ijk}, \Delta_{ijk}$, unwritten here, can be obtained from (6.7)–(6.8) by the replacement $(a,b,c,\ldots \leftrightarrow i,j,k,\ldots)$. Adhering to this style here and below we shall write and discuss only those relations which can not be found by the change of indices. We should remind also that the indices $a,b,c,\ldots$ are raised and lowered by the metrics γ^{ab} and γ_{ab}. The curvature tensor and its contractions are presented in Appendix B.

In the special case of $(n+1)$ decomposition, i.e., when $m=1$ one has adopted bases $\{E_\mu\} = \{E_a, E\}$, $\{\theta^\mu\} = \{\theta^a, \theta\}$, $(a,b=1,2,...,n)$, so that

$$\begin{gathered}\theta^a(E_b) = \delta^b_a; \quad \theta_a(E) = 0 = \theta(E_a); \\ \theta(E) = 1; \quad E\cdot E_a = 0; \quad E\cdot E \equiv \epsilon N^2,\end{gathered} \tag{6.9}$$

where $\{E_a\} \in \Sigma^n$; $\theta^a \in \Sigma^{*n}$ and $E \in \Sigma^1$; $\theta \in \Sigma^{*1}$. In this case the projectors $H' = E_a \otimes \theta^a$ and $H'' = E \otimes \theta$ induce the decomposition of the metric

$$g = g' + g'' = \gamma_{ab}\theta^a \otimes \theta^b + \epsilon N^2 \theta \otimes \theta. \tag{6.10}$$

Then using the relations (6.4)–(6.8), (B1)–(B9), when $i=j=k=1$ or (4.4)–(5.1) when $u = N^{-1}E$, $\omega = N\theta$ we can find all the necessary relations in the $(n+1)$ decomposed form in an adopted basis. Thus, from (4.4) it follows that

$$F = N^{-2}(G - (E \log N)E); \quad G = \nabla_E E. \tag{6.11}$$

The tensor of extrinsic nonholonomicity of the subbundle Σ^n can be written in the form

$$B(E_a, E_b) = \epsilon S_{ab} + A_{ab} \equiv \epsilon N^{-1}\mathcal{B}_{ab}; \quad 2\mathcal{B}_{ab} = 2D_{ab} + F_{ab}, \tag{6.12}$$

where

$$\begin{gathered}S_{ab} = N^{-1}D_{ab}; \quad 2D_{ab} = (\mathcal{L}_E g')(E_a, E_b) = E\gamma_{ab} + E_a\cdot[E_b, E] + E_b\cdot[E_a, E]; \\ 2A_{ab} = \epsilon N^{-1}F_{ab}; \quad F_{ab} = \epsilon N^2 (d\theta)(E_a, E_b) = -E\cdot[E_a, E_b].\end{gathered} \tag{6.13}$$

Acting in the same way as in the previous sections we can find the generalized Gauss–Codazzi–Ricci's equations (see Appendix C).

VII. CANONICAL PARAMETRIZATION OF AN $(n+m)$ SPLIT STRUCTURE AND ITS SPECIAL CASES

The relations of Sec. VI are invariant under the transformation of adopted bases,

$$\theta^a = L^a_b e^b; \quad \theta^l = L^l_k e^k; \quad E_a = (L^{-1})^b_a e_b; \quad E_i = (L^{-1})^k_i e_k, \tag{7.1}$$

where $\{L^a_b\}$ and $\{L^k_i\}$ are $(n\times n)$ and $(m\times m)$ nonsingular matrices, and $\{(L^{-1})^b_a\}$ and $\{(L^{-1})^k_i\}$ are their inverse matrices. Using this property of invariance one can choose, without loss of generality, the simplest basis of decomposition which is useful for applications.

For this purpose we consider the expansion of the covector basis on Σ^{*m} in the domain U of definition of the map x^μ $(\mu = 1,2,...,n,n+1,...,n+m)$, i.e., $\theta^i = \theta^i_\mu dx^\mu$. Due to the fact that the rank of the $n\times(n+m)$ matrix $\{\theta^i_\mu\}$ is equal to n, there is an $(m\times m)$ nonsingular matrix $\{\theta^i_k\}$ as a box in $\{\theta^i_\mu\}$. Then the covectors θ^i can be written in the form, $\theta^i = \theta^i_k dx^k + \theta^i_a dx^a = L^i_k(dx^k + N^k_a dx^a) \equiv L^i_k e^k$, where $L^i_k = \theta^i_k$, $N^k_a = (L^{-1})^k_i \theta^i_a$. Thus the covector basis θ^i goes over into the new covector basis $e^k \in \Sigma^{*m}$. The vector basis on Σ^n can be written similarly as $E_a = E^\mu_a \partial_\mu$. From the condition of duality $e^k(E_a) = 0$ it follows that $E_a = (L^{-1})^b_a(\partial_b - N^k_b \partial_k) \equiv (L^{-1})^b_a e_b$, where $(L^{-1})^b_a = E^b_a$. Thereby we defined the new vector basis $e_b \in \Sigma^n$. The other vector and covector bases ($e^i \in \Sigma^m$ and $e^a \in \Sigma^{*n}$, respectively) are defined by the condition of duality up to $(n\cdot m)$ functions B^a_i. As a result one obtains the following parametrization of the basis of decomposition:

$$e^a = dx^a + B_i^a e^i \in \Sigma^{*n}; \quad e_a = \partial_a - N_a^i \partial_i \in \Sigma^n;$$
$$e^i = dx^i + N_a^i dx^a \in \Sigma^{*m}; \quad e_i = \partial_i - B_i^a e_a \in \Sigma^m. \tag{7.2}$$

We shall call this parametrization the canonical one.

If one follows similar procedure beginning with the covector basis $\theta^a \in \Sigma^{*n}$, one will obtain the other canonical parametrization of $(n+m)$ decomposition,

$$e^a = dx^a + A_i^a dx^i \in \Sigma^{*n}; \quad e_a = \partial_a - M_a^k e_k \in \Sigma^n;$$
$$e^i = dx^i + M_a^i e^a \in \Sigma^{*m}; \quad e_k = \partial_k - A_k^a \partial_a \in \Sigma^m. \tag{7.3}$$

When some metric g is fixed on M, the functions B_i^a (or M_a^i) can be found from the condition of orthogonality (6.2) in terms of $g_{\mu\nu}$ and N_a^i (or A_k^b). If, otherwise, we fix B_i^a (or M_a^i), then we can obtain the metric for both cases according to (6.3):

$$g = \gamma_{ab}(dx^a + B_i^a e^i) \otimes (dx^c + B_k^c e^k) + h_{ik} e^i \otimes e^k,$$
$$g = \gamma_{ab} e^a \otimes e^b + h_{ik}(e^i + M_a^i e^a) \otimes (e^k + M_b^k e^b). \tag{7.4}$$

With respect to the canonically parametrized basis (7.2), the objects (6.6)–(6.8) and the Lie bracket of the basic vector fields have the form

$$\lambda_{ab}^c = -2B_i^c A_{ab}^i; \quad \lambda_{ij}^k = (B_i^a e_j - B_j^a e_i) N_a^k;$$
$$\lambda_{ai}^c = -e_a B_i^c + 2A_{ab}^k B_i^b B_k^c + N_{a,i}^k B_k^c; \quad \lambda_{ia}^k = -2A_{ac}^k B_i^c - N_{a,i}^k,$$
$$2A_{ab}^i = e_b N_a^i - e_a N_b^i; \quad 2A_{ij}^a = e_i B_j^a - e_j B_i^a - \lambda_{ij}^k B_k^a, \tag{7.5}$$
$$2S_{aik} = (\mathcal{L}_{e_a} h)(e_i, e_k); \quad 2S_{iab} = (\mathcal{L}_{e_i} \gamma)(e_a, e_b),$$

where $\gamma = \gamma_{ab} e^a \otimes e^b$ and $h = h_{ik} e^i \otimes e^k$. Here all the geometrical characteristics are expressed in terms of the functions $h_{ij}, \gamma_{ab}, B_i^a, N_b^k$ and their derivatives. Substituting the objects (7.5) for those used in (B2)–(B8) we can obtain the Riemann tensor, the Ricci tensor and the scalar curvature in an $(n+m)$ decomposed form with respect to the canonically parametrized basis (7.2). All the relations for the parametrization (7.3) are found from (7.5) by the substitution $(a,b \leftrightarrow i,j;\ B_i^a \rightarrow M_a^i,\ N_a^i \rightarrow A_i^a)$.

In the case of $(n+1)$ decomposition both types of parametrizations should be considered independently. Thus for the (3+1) monad method there are two kinds of canonical parametrizations (with respect to local coordinates $\{x^\mu\} = \{t, x^i\}$) determined by

$$e_0 = \partial_t - N^i \partial_i = Nu; \quad e^0 = dt + B_i e^i = N^{-1}\omega;$$
$$e_i = \partial_i - B_i e_0; \quad e^i = dx^i + N^i dt, \tag{7.6}$$

and

$$e_0 = \partial_t - M^i e_i = Vu; \quad e^0 = dt + A_i dx^i = V^{-1}\omega;$$
$$e_i = \partial_i - A_i \partial_i; \quad e^k = dx^k + M^k e^0, \tag{7.7}$$

where u is a monad vector, ω is a one-form of time such that $\omega(u) = 1$.

The first set of bases (7.6) is the generalization of the well-known ADM parametrization.[18] In this case the metric has the form

$$ds^2 = N^2(dt + B_j e^j)^2 - h_{ik} e^i e^k, \quad (e^i = dx^i + N^i dt). \tag{7.8}$$

The second set of bases (7.7) implies that the metric is given by

$$ds^2 = V^2(e^0)^2 - h_{ik}(dx^i + M^i e^0)(dx^k + M^k e^0), \tag{7.9}$$

where $e^0 = dt + A_j dx^j$.

The latter parametrization is the generalization of those often used when describing stationary spaces. It is worth emphasizing that the redundant ''degrees of freedom'' of the metrics (7.8)–(7.9) may be used to fix a frame of reference or to simplify the Einstein equations. In the theory of stationary configurations, representation (7.9) is useful for examining of solutions, for which a flux of matter and the timelike Killing's vectors are noncollinear (so-called skew solutions[21]).

If B_j vanishes the metric (7.8) goes over into the standard ADM parametrization,

$$ds^2 = N^2 dt^2 - h_{ik}(dx^i + N^i dt)(dx^k + N^k dt). \tag{7.10}$$

When M^k vanishes, the metric (7.9) has the form

$$ds^2 = V^2(dt + A_j dx^j)^2 - h_{ik} dx^i dx^k. \tag{7.11}$$

This parametrization is often used when describing stationary spaces. If we take $N^i = 0$ or $A_j = 0$ for the metrics (7.10) and (7.11), respectively, then in both cases we have

$$ds^2 = g_{00} dt^2 - h_{ik} dx^i dx^k. \tag{7.12}$$

This kind of decomposition corresponds to a trivial case when Σ^3 is a family of hypersurfaces, where each of the hypersurfaces is orthogonal to the curves $x^i = \text{const}$. This decomposition is invariant under the transformations

$$t = t(t'), \quad x^i = x^i(x'^k). \tag{7.13}$$

The three-dimensional part of these transformations acts uniformly on all the hypersurfaces. Now we shall start, otherwise, from three-dimensional transformations (7.13) which can be extended to the gauge ones by supposing that they depend on time, i.e.,

$$t = t(t'), \quad x^i = x^i(t', x'^k). \tag{7.14}$$

These transformations, under which the hypersurfaces $t = \text{const}$ remain unchanged, have been called the kinemetric ones.[12] In order that the decomposition of the metric be invariant with respect to (7.14) we must ''make longer'' the time derivative $\partial_t \to \partial_t - N^i \partial_i$ (simultaneously we take $dx^i \to dx^i + N^i dt$) by using the gauge vector N^i. Thus it leads to the kinemetric method of decomposition,[12] which coincides with the ADM representation.[18]

Similarly extending the transformations of time we obtain the chronometric transformations[11]

$$t = t(t', x^{k\prime}), \quad x^i = x^i(x'^k). \tag{7.15}$$

It is obvious that the transformations (7.15) do not change the congruence of world lines $x^i = \text{const}$. According to Zelmanov[11] these transformations have been taken as a basis for the definition of the frame of reference. ''Making longer'' the time differential $dt \to dt + A_i dx^i$ (herewith $\partial_i \to \partial_i - A_i \partial_t$) we obtain the chronometric method of decomposition. The transformations (7.15) and (7.14) are the complements of one another and form together the general covariant transformations $x^\mu = x^\mu(x'^\nu)$.

Further generalizations of (7.10) and (7.11) lead to various parametrizations of the monad method. Thus, making longer dt, $dt \to dt + B_i e^i$ $(\partial_i \to \partial_i - B_i e_0)$ one has the canonical parametrization (7.6), (7.8). Making longer ∂_t, $\partial_t \to \partial_t - M^i e_i$ $(dx^k \to dx^k + M^k e_0)$ one obtains the other canonical parametrization (7.7), (7.9) of the monad method for M^4. ''Lengthening'' as referred to is connected with an extension of the admissible transformations, which are not coordinate but basic ones. The generalization (7.8) of the ADM parametrization is invariant under the transformations

$$\tilde{e}^k = \alpha_i^k e^i, \quad \tilde{h}_{ij} = \alpha_i^m \alpha_j^n h_{mn}, \quad \tilde{B}_j = \alpha_j^k B_k. \tag{7.16}$$

If we write the inverse of the metric (7.9),

$$(\partial_s)^2 = V^{-2}(\partial_t - M^i e_i)^2 - h^{ik} e_i e_k, \tag{7.17}$$

then it easily can be seen that the metric (7.9) is invariant under the transformations

$$\tilde{e}_i = \beta_i^k e_k, \quad \tilde{h}^{ij} = \beta_m^i \beta_n^j h^{mn}, \quad \tilde{M}^i = \beta_k^i M^k. \tag{7.18}$$

In (7.16), (7.18), $\{\alpha_k^i\}$ and $\{\beta_k^i\}$ are nonsingular matrices depending on a point p of M. From this we can clearly see the role of the parametrizations (7.8), (7.9) as such generalizations of the kinemetric and chronometric methods that the corresponding metrics admit nonholonomic transformations of spatial vector and covector bases (7.16) and (7.18), respectively.

VIII. DECOMPOSITION INDUCED BY A FAMILY OF SURFACES

Let $\{M^m \subset M\}$ be an n-parameter family of m-dimensional surfaces. One may think of these surfaces as intersections of the hypersurfaces $x^a = \text{const}$, i.e., $M^m = \cap_a \{x^a = \text{const}\}$. It is obvious that such a family induces $(n+m)$ decomposition of M. Indeed, there exists the vector basis $e_i = \partial_i$ on $T(M)$, $(i = n+1,...,n+m)$, because of holonomicity of the M^m itself. As a consequence of it, the covector basis on the dual to $T(M^m)$ subbundles Σ^{*n} is a set of one-forms $\{e^a = dx^a\}$. The corresponding dual bases to the bases $\{e_i\}$ and $\{e^a\}$ are determined up to $(n \cdot m)$ functions N_a^i such that

$$\begin{aligned} e^a &= dx^a \in \Sigma^{*n}, \quad e_a = \partial_a - N_a^i \partial_i \in \Sigma^n, \\ e^i &= dx^i + N_a^i dx^a \in \Sigma^{*m}, \quad e_i = \partial_i \in \Sigma^m. \end{aligned} \tag{8.1}$$

The functions N_a^i are expressed in terms of the components of the metric g by using the condition of orthogonality $e_a \cdot e_i = 0$. Thus the projection operators and the metric have the form

$$H' = (\partial_a - N_a^i \partial_i) \otimes dx^a, \quad H'' = \partial_k \otimes (dx^k + N_a^k dx^a), \tag{8.2}$$

$$g = \gamma_{ab} dx^a \otimes dx^b + h_{ik}(dx^i + N_c^i dx^c) \otimes (dx^k + N_d^k dx^d). \tag{8.3}$$

From the form of the metric (8.3) it can be seen that here we used the special case of canonical parametrization of $(n+m)$ decomposition (7.2) when B_i^a vanishes. In this case the formulas (7.5) become much simpler. Thus, one finds

$$\lambda_{ab}^c = 0; \quad \lambda_{ij}^k = 0; \quad \lambda_{ai}^c = 0; \quad \lambda_{ia}^k = -N_{a,i}^k; \quad A_{ij}^a = 0; \tag{8.4}$$

$$2S_{iab} = \gamma_{ab,i}; \quad 2A_{ab}^i = e_b N_a^i - e_a N_b^i, \tag{8.5}$$

$$2S_{aik} = h_{ik,a} - h_{ik,l} N_a^l - h_{lk} N_{a,i}^l - h_{il} N_{a,k}^l, \tag{8.6}$$

$$2L_{cab} = 2\Delta_{cab} = e_a \gamma_{bc} + e_b \gamma_{ca} + e_c \gamma_{ab}, \tag{8.7}$$

$$2L_{ijk} = 2\Delta_{ijk} = h_{ij,k} - h_{ik,j} - h_{jk,i}. \tag{8.8}$$

The partial derivatives with respect to coordinates x^i and x^a are denoted here by i and a, respectively. Then, according to (B2)–(B8), one can find the curvature tensor and its contractions.

IX. DECOMPOSITION INDUCED BY A GROUP OF ISOMETRIES

Let M assume to a nontransitive group of isometries G^n with the n linearly independent Killing's vectors $\{\xi_a\}$, which satisfy the relations

$$[\xi_a,\xi_b]=C^d_{ab}\xi_d \quad (a,b,d=1,2,...,n), \tag{9.1}$$

where the C^d_{ab} are the structure constants and obey the Jacobi identity $C^c_{[ab}C^f_{d]c}=0$ and the condition $C^c_{ab}+C^c_{ba}=0$. In addition, the metric g satisfies the Killing's equations,

$$(\mathcal{L}_{\xi_a}g)(X,Y)=\xi_a(X\cdot Y)-[\xi_a,X]\cdot Y-X\cdot[\xi_a,Y]=0, \quad \forall X,Y\in T(M). \tag{9.2}$$

The group G^n decomposes M into a family of m-codimensional surfaces $\{M^n\}\subset M$, on which G^n is simply transitive ($\{M^n\}$ are invariant manifolds). Thus, we can say that the group G^n induces $(n+m)$ decomposition of M into the m-parameter family of n-dimensional surfaces of transitivity. Then the subbundle $\Sigma^n=\cup T(M^n)$ is a union of the tangent bundles of the family $\{M^n\}$, and Σ^m is a union of all the m-dimensional directions, which are tangent to M and orthogonal to $T(M^n)$.

Now we shall start in the same way as in the previous section. Thus one may think of the surfaces M^n as an intersection of the invariant hypersurfaces $\{x^i=\text{const}\}$, i.e., $M^m=\cap_i\{x^i=\text{const}\}$, $(i=n+1,...,n+m)$. Moreover, one has $dx^i(\xi_a)=\xi_a x^i=0$. This is obvious that the invariant differential one-forms dx^i can be chosen as a covector basis on the subbundles Σ^{*m}. Then there exists the vector basis $\{\partial_a\}\in T(M^n)$, so that $dx^i(\partial_a)=0$ and $\xi_a=\xi^b_{(a)}\partial_b$. Having extended these bases to the "complete ones," $\{dx^i\}\rightarrow\{dx^\mu\}=\{dx^a,dx^i\}\in T^*(M)$ and $\{\partial_a\}\rightarrow\{\partial_\mu\}=\{\partial_a,\partial_i\}\in T(M)$, where $dx^\mu(\partial_\nu)=\delta^\mu_\nu$ and $[\xi_a,\partial_i]=0$, we can define one-forms ω^a such that

$$\omega^a(\xi_b)=\delta^a_b;\quad \omega^a(\partial_i)=0;\quad \mathcal{L}_{\partial_i}\omega^a=0, \tag{9.3}$$

$$\mathcal{L}_{\xi_a}\omega^b=-C^b_{ad}\omega^d;\quad 2d\omega^a=C^a_{bd}\omega^b\wedge\omega^d. \tag{9.4}$$

Let us now introduce an auxiliary definition. We shall say that a split structure $\mathcal{H}^2$ is compatible with a group of isometries if the conditions of invariance of $\mathcal{H}^2$ are satisfied, i.e., if

$$\mathcal{L}_{\xi_a}H'=0,\quad \mathcal{L}_{\xi_a}H''=0,\quad (a=1,2,...,n). \tag{9.5}$$

Using (6.3) and (9.1) one can easily verify that for the other vector and covector bases $\{E_k\}\in\Sigma^m$ and $\{\theta^a\}\in\Sigma^{*n}$ we have, respectively,

$$\mathcal{L}_{\xi_a}\theta^b=-C^b_{ad}\theta^d;\quad \mathcal{L}_{\xi_a}E_k=0. \tag{9.6}$$

To concretize the basis of decomposition we take $\theta^a=\theta^a_\mu dx^\mu$ and $E_i=E^\mu_i\partial_\mu$. Then the conditions of duality $\theta^a(\xi_b)=\delta^a_b$, $\theta^a(E_i)=0$, $dx^k(E_i)=\delta^k_i$ determine these bases up to $(n\cdot m)$ functions A^a_i. As a result the basis of $(n+m)$ decomposition has the form

$$\begin{gathered}\xi_a\in\Sigma^n;\quad e^a=\omega^a+A^a_i dx^i\in\Sigma^{*n},\\ e_i=\partial_i-A^a_i\xi_a\in\Sigma^m;\quad dx^k\in\Sigma^{*m},\quad [\xi_a,e_i]=0.\end{gathered} \tag{9.7}$$

The projection operators and the metric can be written as

$$H'=\xi_a\otimes(\omega^a+A^a_i dx^i);\quad H''=(\partial_i-A^a_i\xi_a)\otimes dx^i, \tag{9.8}$$

$$g=g'+g''=\gamma_{ab}(\omega^a+A_i^a dx^i)\otimes(\omega^b+A_j^b dx^j)+h_{kl}dx^k\otimes dx^l. \tag{9.9}$$

From the Killing's equations one finds

$$\xi_a\gamma_{bc}-C_{ab}^d\gamma_{dc}-C_{ac}^d\gamma_{bd}=0;\quad \xi_a A_i^b-C_{ad}^b A_i^d=0;\quad \xi_a h_{ik}=0. \tag{9.10}$$

Using these equations we obtain the main geometrical objects

$$\begin{gathered}A''(\xi_a,\xi_b)=0;\quad 2A'(e_i,e_k)\equiv F_{ik}^a\xi_a,\\ F_{ik}^a=A_{k,i}^a-A_{i,k}^a+C_{bd}^a A_k^b A_i^d,\\ S'(e_i,e_k)=0;\quad 2e_i\cdot S''(e_a,e_b)\equiv 2S_{iab}=e_i\gamma_{ab},\\ 2L_{abc}=C_{cab}+C_{bca}+C_{acb};\\ 2L_{ijk}=2\Delta_{ijk}=e_jh_{ik}+e_kh_{ij}-e_ih_{jk}.\end{gathered} \tag{9.11}$$

In the end, from relations (B2)–(B8) we can find the curvature tensor, the Ricci tensor, and scalar curvature (see Appendix D). When $m=0$ we come to the case of homogeneous spaces.

X. LAGRANGIANS OF THE UNIFIED MULTIDIMENSIONAL KALUZA–KLEIN THEORIES

The mathematical model we shall use for spaces of the unified theories is the totality of the following objects: (a) a connected $(4+n)$-dimensional pseudo-Riemannian C^∞ manifold M^{4+n} with a nonsingular metric g on it; (b) an n-parameter compact group of isometries G^n on M^{4+n} with linearly independent Killing's vectors $\xi_a\in T(M^{4+n})$ for which the structure constants C_{bd}^a satisfy the condition $C_{ad}^a=0$, $(a,b,d=4,5,...,n+3)$.

The physical space–time $V^4\equiv M^{4+n}/M^n$ is the quotient space M^{4+n} with respect to the invariant manifolds M^n of the group G^n. The V^4 is described by the components h_{ik} of the metric h, by the set of gauge fields A_i^b and by the multiplet $n(n+1)/2$ of scalar fields $\varphi_{ab}\equiv-\gamma_{ab}$. All these tensors are obtained under the $(4+n)$ decomposition of M^{4+n} (see Sec. IX). The true physical configuration is described not by a single set of fields $\{h_{ik},A_j^b,\varphi_{cd}\}$, but by a whole equivalence class of such sets; each of them corresponds to some point of the orbit G^n. The signature of the metric g is defined by two conditions; first, the metric h is a Lorentz one, and second, the energy density is positive for obtained Lagrangian of fields $\{A_j^b,\varphi_{cd}\}$. In addition, the metric g satisfies the $(4+n)$-dimensional variational Hilbert principle for the functional $S[g]$, i.e.,

$$\delta S[g]=\delta\left\{-\frac{1}{4\pi V}\int R^{(4+n)}\Omega^{(4+n)}\right\}=0, \tag{10.1}$$

where $R^{(4+n)}$ is the curvature scalar on $M^{(4+n)}$, the $(4+n)$-form $\Omega^{(4+n)}$ is the volume measure on M^{4+n}, and V is the n-dimensional invariant volume of M^n,

$$V=\int_{M^n}\omega^4\wedge\omega^4\wedge\cdots\wedge\omega^{4+n}\equiv\int_{M^n}\Omega^{(n)}. \tag{10.2}$$

The conditions $C_{ab}^a=0$ follow from the requirement that the volume measure $\Omega^{(n)}$ must be invariant. They are necessary for compatibility of the variational Hilbert principle and homogeneity of M^n with respect to the group of isometries G^n. This restricts the admissible variations of fields $\mathcal{L}_{\xi_a}\delta g=0$ in (10.1). (The similar situation may be found in the theory of homogeneous models of cosmology.[35,36])

Using the formulas of Sec. IX and Appendix D for the metric g in the $(n+4)$ decomposed form

$$g=h_{ik}dx^i\otimes dx^k-\varphi_{ab}(\omega^a+A^a_m dx^m)\otimes(\omega^b+A^b_n dx^n) \tag{10.3}$$

and omitting a divergence of some vector, we obtain

$$S^{(4+n)}[g]=S[\varphi_{ab},A^a_i,h_{jk}]=\int_{V^4}\sqrt{-h}L d^4x. \tag{10.4}$$

The Lagrangian density is

$$\sqrt{-h}L=-\frac{1}{4\pi}\sqrt{|h\varphi|}\left\{R^{(4)}+\frac{1}{4}\varphi_{ab}F^a_{ij}F^{bij}+(\varphi^{ab}\varphi^{cd}-\varphi^{ac}\varphi^{bd})h^{ik}D_i\varphi_{ab}D_k\varphi_{cd}+U(\varphi_{ab})\right\}, \tag{10.5}$$

where

$$U(\varphi_{ab})=\tfrac{1}{2}\varphi^{cd}C^a_{bc}(C^b_{ad}+\tfrac{1}{2}\varphi_{ap}\varphi^{bq}C^p_{qd}), \tag{10.6}$$

and

$$D_i\varphi_{ab}=\varphi_{ab,i}-T(A_i)^d_a\varphi_{db}-T(A_i)^d_b\varphi_{ad} \tag{10.7}$$

is the gauge-invariant derivative. The components $T(A)^a_b\equiv C^a_{bd}A^d$ of the matrix $T(A)$ realize the adjoint representation of the group G^n:$[T(A),T(B)]=T([A,B])$, $A=A^a\xi_a$, $B=B^a\xi_a$. The Lagrangian of this kind (but with the second derivatives of the fields φ_{ab}) has been obtained in Ref. 37.

When $n=1$ the Lagrangian (10.5) reduces to the Lagrangian of the five-dimensional Kaluza–Klein theory.[27] In the static case of spherical symmetry from $n=1$ it follows the Lagrangian of the simple dynamic system. Its equations can be integrated by the separation of variables of the corresponding Hamilton–Jacobi equation. In such a way the solution for the interacting scalar, electromagnetic, and gravitational fields was obtained in Ref. 38 within the framework of the united five-dimensional Kaluza–Klein theory.

XI. RELATIVISTIC CONFIGURATIONS OF A PERFECT FLUID

Let us consider space–time M^4 with the metric g in the $(3+1)$ decomposed form

$$g=V^2e^0\otimes e^0-h_{ik}e^i\otimes e^k,\quad g^{-1}=V^{-2}e_0\otimes e_0-h^{ik}e_i\otimes e_k, \tag{11.1}$$

where g^{-1} is the inverse of the metric g. For the time being, we require the basis of decomposition to be an adopted abstract one (i.e., not concretized). Let the source of the gravitational field described by the metric (11.1) be a perfect fluid with the field of 4-velocities $u=V^{-1}e_0=d/ds$ which is tangent to the flow lines $x^\mu=x^\mu(s)$. Herewith the mass density ρ obeys the conservation law,

$$\mathrm{div}(\rho u)\equiv(\nabla_{e_\mu}\rho u)(e_\mu)=V^{-1}h^{-1/2}\mathcal{L}''_{e_0}(\rho h^{1/2})=0, \tag{11.2}$$

where $\mathcal{L}''_{e_0}$ is the Lie derivative with respect to the basis $\{e_i\}$:$\mathcal{L}''_{e_0}\sqrt{h}=\frac{1}{2}\sqrt{h}h^{ik}(\mathcal{L}_{e_0}h)(e_i,e_k)$. The equation of motion for the fluid follows from the relation,

$$\mathrm{div}\, T\equiv(\nabla_{e_\mu}T)(e^\mu,.)=0. \tag{11.3}$$

The energy-momentum tensor T is

$$T=\mu V^{-2}e_0\otimes e_0+Ph^{ik}e_i\otimes e_k, \tag{11.4}$$

where μ is the energy density of the fluid, P is the pressure. Using the thermodynamic relations

$$d\mathcal{H}=Tds+\rho^{-1}dP,\quad \mathcal{H}=(\mu+P)\rho^{-1}, \tag{11.5}$$

one finds the equations of motion

$$(\operatorname{div} T)(e_0)=\rho TV^{-1}uS=-\rho V^{-1}dS/ds=0, \tag{11.6}$$

$$(\operatorname{div} T)(e_i)=h^{ik}(dP-\rho\mathcal{H}\mathcal{L}_u\omega)(e_k)=0. \tag{11.7}$$

Here we use the following notations: $\mathcal{H}$ is the enthalpy, S is the entropy, T is the temperature, and ω is the covector of the 4-velocity of the fluid ($\omega=Ve^0$, $\omega(u)=1$). We introduce ''the one-form of the enthalpy θ'' and ''the two-form of the curl Ω'' by

$$\theta=\mathcal{H}\omega=\mathcal{H}Ve^0,\quad \Omega=d\theta. \tag{11.8}$$

Then the equations of motion (11.6), (11.7) can be expressed as

$$\mathcal{L}_{e_0}\theta=d(\mathcal{H}V)-VTdS. \tag{11.9}$$

Using the formula $\mathcal{L}_{e_0}=i_{e_0}d+di_{e_0}$, where the operator i_{e_0} is defined by the relation $(i_{e_0}\Omega)(Y)=\Omega(e_0,Y)$, $\forall Y\in T(M^4)$, we obtain one more form of the equations of motion

$$i_{e_0}\Omega=-VTdS. \tag{11.10}$$

The condition of integrability of these relations leads to the equations of motion for the curl of a perfect fluid

$$\mathcal{L}_{e_0}\Omega=-d(TV)\wedge dS. \tag{11.11}$$

In the special case $S=\text{const}$ a perfect fluid is isentropic so that the equations for ''the one-form of the enthalpy'' (11.9) and ''the two-form of the curl'' (11.10), (11.11) are reduced to the relations

$$\mathcal{L}_{e_0}\theta=d(\mathcal{H}V), \tag{11.12}$$

$$i_{e_0}\Omega=0,\quad \mathcal{L}_{e_0}\Omega=0. \tag{11.13}$$

It is to be noted that the last equation in (11.13) is the condition of integrability of the equation (11.12). Moreover we may regard this condition as an invariant formulation of the theorem,[39] which states that the two-form of the curl Ω is constant along the world lines of particles of an isentropic perfect fluid. From the first relation in (11.13) it follows that Ω is singular, i.e., $\Omega(e_0,X)=0$, $\forall X\in T(M^4)$, and therefore ''completely spatial.'' This implies

$$\Omega=\sum_{i,j}\Omega_{ij}e^i\wedge e^j;\quad \Omega\wedge\Omega=d\theta\wedge d\theta=0. \tag{11.14}$$

Since in general case $\theta\wedge d\theta\neq 0$, then according to the theorem Darboux (see, for example, Ref. 23) it follows that there exists such functions ξ, η, ζ that $\theta=d\xi+\eta d\zeta$. This representation has been used in Ref. 40 to construct a number of families of solutions of the Einstein equations for an isentropic perfect fluid.

Now we shall consider the stationary spaces of General Relativity with a timelike Killing's vector ∂_t. Then the equations (11.6), (11.7), as well as their consequences (11.9)–(11.13), go over into the equilibrium conditions of a perfect fluid. For an isentropic stationary flow they admit completely three-dimensional formulation. Indeed, in this case one has

$$\mathcal{L}_{\partial_t}g=0,\quad \mathcal{L}_{\partial_t}e^\mu=0,\quad [\partial_t,e_\mu]=0. \tag{11.15}$$

Then using the parametrization of decomposition (7.7) we deduce that the functions V, A_i, M^k, h_{ik} as well as $\rho, \mu, P, \mathcal{H}$, do not depend on time. We define the vector $\mathbf{M}$ and covector A on the subbundles $\Sigma'' \equiv \Sigma^3$ by

$$\mathbf{M} = M^i \partial_i, \quad A = A_k dx^k. \tag{11.16}$$

In terms of $\mathbf{M}$ and A the conservation law for mass (11.2) is transformed into the three-dimensional equation of continuity of the flow lines

$$\mathrm{div}^{(3)}(\rho \mathbf{M}) \equiv (\nabla_{e_i} \rho \mathbf{M})(e^i) = h^{-1/2} \mathcal{L}_{\mathbf{M}}(\rho h^{1/2}) = 0. \tag{11.17}$$

When $S = \mathrm{const}$ the condition (11.9) may be rewritten in the three-dimensional form as well

$$i_{\mathbf{M}} dA = -d \log(\mathcal{H} V); \quad \mathbf{M}(\mathcal{H} V) = 0. \tag{11.18}$$

From now on the objects and operations are defined in the three-dimensional manifold $t = \mathrm{const}$ with respect to the bases $\{\partial_i\}$ and $\{dx^k\}$. For example, $dA = (1/2) \mathcal{F}_{ik} dx^i \wedge dx^k$, where $\mathcal{F}_{ik} = A_{k,i} - A_{i,k}$. The equilibrium condition (11.18) may be expressed in the form

$$\mathcal{L}_{\mathbf{M}} A = d\{A(\mathbf{M}) - \log(\mathcal{H} V)\} \tag{11.19}$$

showing that the one-form $\mathcal{L}_{\mathbf{M}} \mathbf{A}$ is exact. Hence, as the condition of integrability one obtains the conservation three-dimensional theorem for the curl dA along the three-dimensional flow lines, i.e.,

$$\mathcal{L}_{\mathbf{M}} dA = 0. \tag{11.20}$$

In the case of parametrization (7.6) for the stationary spaces the functions V, B_i, N^k, h_{jk} do not depend on time either. By analogy with (11.19) one has

$$\mathcal{L}_{\mathbf{N}} B = -d \log(\mathcal{H} V), \tag{11.21}$$

where

$$\mathbf{N} = N^i \partial_i, \quad B = B_k dx^k. \tag{11.22}$$

The condition of integrability gives the conservation theorem for the curl of B

$$\mathcal{L}_{\mathbf{N}} dB = 0. \tag{11.23}$$

If one of the two objects A and $\mathbf{M}$ in (11.19) (or $\mathbf{N}$ and B in (11.21)) vanishes then the equilibrium condition of an isentropic perfect fluid has the simple form

$$\mathcal{H} V = V(\mu + p)/\rho = k, \tag{11.24}$$

where k is the constant. Thus the Lagrangian of an isentropic perfect fluid in equilibrium is

$$L_m \equiv -V \sqrt{h} P = (k\rho - \mu V) \sqrt{h} = [k - (1+\epsilon) V] \rho \sqrt{h}, \tag{11.25}$$

where $\epsilon = \epsilon(\rho)$ is the internal energy of the fluid and $\mu = \rho(1+\epsilon)$.

As was noted above, the parametrizations (7.6), (7.7) have spurious degrees of freedom. It means that the vector $\mathbf{M}$ or covector A in (7.7) can be chosen arbitrarily, by using additional physical reasons. Therefore we have a right to introduce the potential of rotation Ψ_1 by the formula

$$\mathcal{L}_{\mathbf{M}} A = d(\log \Psi_1). \tag{11.26}$$

Then the equilibrium condition (11.19) is written as a relation for potentials

$$\Psi_1 HV = C_1 \exp(A_i M^i), \quad C_1 = \text{const} \tag{11.27}$$

and actually gives us the integral of motion. In another case of the parametrization the equilibrium condition (7.6) can be expressed in the form

$$\mathcal{L}_{\mathbf{N}} B = d(\log \Psi_2), \quad \Psi_2 HV = C_2 = \text{const}. \tag{11.28}$$

Thus the potentials Ψ_1 and Ψ_2 are different from each other by the exponential factor $\exp(A_i M^i)$.

ACKNOWLEDGMENTS

We would like to acknowledge M. Korkina and A. Sokolovsky for interesting and helpful discussions, and Yu. Vladimirov for encouragement and continual attention to the theme of this research. The authors wish also to thank M. Godina who was helpful in pointing out some inaccuracies and A. Boroviec for drawing our attention to several important works on the theory of ''almost-product structures.''

APPENDIX A: THE GENERALIZED GAUSS–CODAZZI–RICCI'S EQUATIONS

Replacing all the connections in the definition of the curvature tensor (2.17) by their ''split representatives'' (2.18)–(2.21) we have obtained the invariant nonholonomic generalizations of the Gauss–Codazzi–Ricci's equations,

$$R(X^a,Y^a)Z^a\cdot V^a = R^a(X^a,Y^a)Z^a\cdot V^a + \sum_{c\neq a}\{2A^c(X^a,Y^a)\cdot B^c(Z^a,V^a) + B^c(Y^a,V^a)\cdot B^c(X^a,Z^a) - B^c(X^a,V^a)\cdot B^c(Y^a,Z^a)\}, \tag{A1}$$

$$R(X^a,Y^a)Z^a\cdot V^b = V^b\cdot\{(\nabla^b_{Y^a}B^b)(X^a,Z^a) - (\nabla^b_{X^a}B^b)(Y^a,Z^a)\} + 2Z^a\cdot B^a(A^b(X^a,Y^a),V^b) + \sum_{c\neq a,b}\{2Z^a\cdot Q^a(A^c(X^a,Y^a),V^b) + B^c(X^a,Z^a)\cdot Q^c(Y^a,V^b) - B^c(Y^a,Z^a)\cdot Q^c(X^a,V^b)\}, \tag{A2}$$

$$R(X^a,Y^b)Z^a\cdot V^b = (Z^a\cdot(\nabla^a_{X^a}B^a) + \langle X^a\cdot B^a, Z^a\cdot B^a\rangle)(Y^b,V^b) + (V^b\cdot(\nabla^b_{Y^b}B^b) + \langle Y^b\cdot B^b, V^b\cdot B^b\rangle)(X^a,Z^a) + \sum_{c\neq a,b}\{B^c(X^a,Z^a)\cdot B^c(Y^b,V^b) - Q^c(X^a,V^b)\cdot Q^c(Y^b,Z^a) + V^b\cdot Q^b(\Lambda^c(X^a,Y^b),Z^a)\}, \tag{A3}$$

$$R(X^a,Y^b)Z^a\cdot V^d = V^d\cdot\{(\nabla^d_{Y^b}B^d)(X^a,Z^a) - (\nabla^d_{X^a}Q^d)(Y^b,Z^a) - B^d(Y^b,B^b(X^a_r,Z^a))\} + Z^a\cdot\{B^a(Y^b,Q^b(X^a,V^d)) - B^a(\Lambda^d(X^a,Y^b),V^d)\} + (\langle Y^b\cdot B^b, V^d\cdot B^d\rangle)(X^a,Z^a) - (\langle X^a, B^a, V^d\cdot Q^d\rangle)(Y^b,Z^a) - \sum_{c\neq a,b,d}\{Z^a\cdot Q^a(\Lambda^c(X^a,Y^b),V^d) - B^c(X^a,Z^a)\cdot Q^c(Y^b,V^d) + Q^c(X^a,V^d)\cdot Q^c(Y^b,Z^a)\}, \tag{A4}$$

$$\begin{aligned}R(X^a,Y^b)Z^c\cdot V^d=&V^d\cdot\{(\nabla^d_{Y^b}Q^d)(X^a,Z^c)-(\nabla^d_{X^a}Q^d)(Y^b,Z^c)\}\\&+B^d(X^a,Q^a(Y^b,Z^c))-B^d(Y^b,Q^b(X^a,Z^c))+B^d(\Lambda^c(X^a,Y^b),Z^c)\}\\&+Z^c\cdot\{B^c(Y^b,Q^b(X^a,V^d))-B^c(X^a,Q^a(Y^b,V^d))\\&-Q^c(\Lambda^c(X^a,Y^b),V^d)\}+(\langle Y^b\cdot B^b,V^d\cdot Q^d\rangle)(X^a,Z^c)\\&-(\langle X^a\cdot B^a,V^d\cdot Q^d\rangle)(Y^b,Z^c)+\sum_{f\neq a,b,c,d}\{Q^f(Y^b,V^d)\cdot Q^f(X^a,Z^c)\\&-Q^f(Y^b,Z^c)\cdot Q^f(X^a,V^d)+V^d\cdot Q^d(\Lambda^f(X^a,Y^b),Z^c)\}.\end{aligned}\tag{A5}$$

In the formula (A1) the curvature tensor R^a of the subbundle Σ^a, introduced in Ref. 30, is

$$R^a(X^a,Y^a)Z^a\equiv\left\{\nabla^a_{X^a}\nabla^a_{Y^a}-\nabla^a_{Y^a}\nabla^a_{X^a}-\nabla^a_{[X^a,Y^a]^a}+2\sum_{c\neq a}\mathcal{L}^a_{A^c(X^a,Y^a)}\right\}Z^a.\tag{A6}$$

The covariant derivatives of the values B^d and Q^d are given by

$$(\nabla^b_{X^a}B^b)(Y^a,Z^a)=\nabla^b_{X^a}(B^b(Y^a,Z^a))-B^b(\nabla^a_{X^a}Y^a,Z^a)-B^b(Y^a,\nabla^a_{X^a}Z^a),\tag{A7}$$

$$(\nabla^d_{X^b}B^b)(Y^a,Z^a)=\nabla^d_{X^b}(B^d(Y^a,Z^a))-B^d(\nabla^a_{X^b}Y^a,Z^a)-B^d(Y^a,\nabla^a_{X^b}Z^a),\tag{A8}$$

$$(\nabla^d_{X^a}Q^d)(Y^b,Z^c)=\nabla^d_{X^a}(Q^d(Y^b,Z^c))-Q^d(\nabla^b_{X^a}Y^b,Z^c)-Q^d(Y^b,\nabla^c_{X^a}Z^c).\tag{A9}$$

We also used the definition

$$(\langle Y^b\cdot B^b,V^d\cdot Q^d\rangle)(X^a,Z^c)\equiv\langle Y^b\cdot B^b(X^a,.),V^d\cdot Q^d(.,Z^c)\rangle.\tag{A10}$$

When fixing the vectors X^a, Y^b, V^d, Z^c, the definition (A10) gives us the scalar product $\langle\alpha^{ba},\beta^{dc}\rangle$ of one-forms

$$\alpha^{ba}\equiv Y^b\cdot B^b(X^a,.),\quad \beta^{dc}\equiv V^d\cdot Q^d(.,Z^c).$$

When $n_a=1$ $(a=1,2,...,r)$, i.e., when all the subbundles are one-dimensional, the relations obtained here reduce to the r-dimensional variant of the tetradic method's formulas.[34]

APPENDIX B: COMPONENTS OF THE CURVATURE TENSOR WITH RESPECT TO AN ADOPTED BASIS FOR $(n+m)$ DECOMPOSITION

Due to the definitions

$$\{E_\mu\}=\{E_a,E_i\};\quad R(E_\mu,E_\nu)E_\rho\cdot E_\sigma=R_{\sigma\rho\mu\nu};\quad R(E_\mu,E_\nu)E_\rho=R^\sigma_{\rho\mu\nu}E_\sigma,\tag{B1}$$

the generalized Gauss–Codazzi–Ricci's equations (3.12)–(3.15) have the form

$$R_{abcd}=R^{(n)}_{abcd}+2A^i_{.cd}B_{iba}+B^i_{.cb}B_{ida}+B^i_{.db}B_{ica},\tag{B2}$$

$$R_{ibcd}=B_{icb|d}-B_{idb|c}+2A^k_{.cd}B_{bki}+B^k_{.db}(B_{cik}-\lambda_{kic})-B^k_{.cb}(B_{dik}-\lambda_{kid}),\tag{B3}$$

$$R_{ibcj}=B_{bji|c}-B_{icb|j}-B_{bjk}B_{ci.}{}^k-B_{icd}B_{jb.}{}^d+B_{bki}\lambda^k_{.jc}+B_{bjk}\lambda^k_{.ic}+B_{idb}\lambda^d_{.cj}+B_{icd}\lambda^d_{.bj},\tag{B4}$$

where the curvature tensor of the subbundle Σ^n is defined by its components $R^{(n)}_{abcd}$ according to

$$R^{(n)a}{}_{bcd}=E_cL^a_{db}-E_dL^a_{cb}+L^f_{db}L^a_{cf}-L^f_{cb}L^a_{df}-\lambda^f_{cd}L^a_{fb}+2A^i_{.cd}\lambda^a_{bi}\tag{B5}$$

(and similarly for the replacement $n \to m$ and $a,b,c,\ldots \leftrightarrow i,j,k,\ldots$). Then the components of the Ricci tensor and the curvature scalar have the form

$$R_{bd} = R_{bd}^{(n)} - B_{db|i}^{i} - S_{b|d} + 2A_{ad}^{i} A_{ib.}{}^{a} + 2S_{iad} S^{ia}{}_{b} - S_{b}{}^{ij} S_{dij} + A_{bij} A_{d}{}^{ij} - S_{i} B_{.db}^{i} - B_{.da}^{i} \lambda_{.bi}^{a} - B_{.ab}^{i} \lambda_{.di}^{a}, \tag{B6}$$

$$R_{ia} = B_{ia.|b}^{b} + B_{ai.|k}^{k} - S_{i|a} - S_{a|i} - 2S_{ik}^{b} S_{ab}^{k} - 6A_{ik}^{b} A_{ab}^{k} + S^{k}(B_{aik} - \lambda_{kia}) + S^{b}(B_{iab} - \lambda_{bai}) + B_{ab}^{k} \lambda_{ki.}{}^{b} + B_{ik}^{b} \lambda_{ba.}{}^{k}, \tag{B7}$$

$$R = R^{(n)} = 2S^{i}{}_{|i} - S^{i} S_{i} - S_{ab}^{i} S_{i..}^{ab} - A_{ab}^{i} A_{i..}^{ab} + R^{(m)} - 2S_{|a}^{a} - S^{a} S_{a} - S_{ij}^{a} S_{a.}^{ij} - A_{ij}^{a} A_{a..}{}^{ij}, \tag{B8}$$

where $S^{i} = S_{ab}^{i} \gamma^{ab}$, $S^{a} = S_{ik}^{a} h^{ik}$. The signs ''$_{|i}$'' and ''$_{|a}$'' denote the covariant derivative with respect to the connections L_{mn}^{k} and L_{bc}^{a} in the directions of the vectors E_{i} and E_{a}, respectively. For example,

$$B_{icb|d} = E_{d} B_{icb} - B_{iab} L_{dc}^{a} - B_{ica} L_{db}^{a} \quad (a,b,c \leftrightarrow i,j,k). \tag{B9}$$

The other components of the Ricci tensor and the curvature tensor can be found from (B2)–(B7) by the formal substitution $a,b,c,\ldots$ for $i,j,k,\ldots$ and otherwise.

APPENDIX C: COMPONENTS OF THE CURVATURE TENSOR WITH RESPECT TO AN ADOPTED BASIS FOR $(n+1)$ DECOMPOSITION

The generalized Gauss–Codazzi–Ricci's equations for the metric (6.10) with respect to the basis (6.9) have the form,

$$R_{abcd} = R_{abcd}^{(n)} + \epsilon N^{-2} (\mathcal{B}_{cb} \mathcal{B}_{da} - \mathcal{B}_{db} \mathcal{B}_{ca} + F_{cd} \mathcal{B}_{ba}), \tag{C1}$$

$$R_{n+1,bcd} = N\{(N^{-1} \mathcal{B}_{cb})_{|d} - (N^{-1} \mathcal{B}_{db})_{|c}\} - \epsilon N^{-2} G_{b} F_{cd}, \tag{C2}$$

$$R_{n+1,bc,n+1} = N \mathcal{L}_{E} (N^{-1} \mathcal{B}_{cb}) - \mathcal{B}_{ca} \mathcal{B}_{b.}{}^{a} + \epsilon N^{-2} G_{b} G_{c} - N^{2} (N^{-2} G_{b})_{|c}, \tag{C3}$$

$$R_{bd} = R_{bd}^{(n)} - \epsilon N^{-2} [N \mathcal{L}_{E} (N^{-1} \mathcal{B}_{db}) + D \mathcal{B}_{db} + \tfrac{1}{2} F_{ba} F_{d.}{}^{a} - 2 D_{ba} D_{d.}{}^{a}] + \epsilon (N^{-2} G_{b})_{|d} - N^{-4} G_{b} G_{d}, \tag{C4}$$

$$R_{n+1,a} = N[(N^{-1} \mathcal{B}_{a.}{}^{b})_{|b} - E_{a}(N^{-1} D)] - \epsilon N^{-1} F_{ab} G^{b}, \tag{C5}$$

$$R_{n+1,n+1} = -N E(N^{-1} D) - D_{ab} D^{ab} + \tfrac{1}{4} F_{ab} F^{ab} + N^{2} (N^{-2} G^{a})_{|a} - \epsilon N^{-2} G_{a} G^{a}, \tag{C6}$$

$$R = R^{(n)} - 2\epsilon N^{-1} E(N^{-1} D) - \epsilon N^{-2} (D^{2} + D_{ab} D^{ab} + \tfrac{1}{4} F_{ab} F^{ab}) + 2\epsilon (N^{-2} G_{a})_{|a} - 2N^{-4} G_{a} G^{a}, \tag{C7}$$

$$R^{(n)a}{}_{bcd} = E_{c} L_{db}^{a} - E_{d} L_{cb}^{a} + L_{db}^{f} L_{cf}^{a} - L_{cb}^{f} L_{df}^{a} - \lambda_{cd}^{f} L_{fb}^{a} + \epsilon N^{-2} F_{cd} \lambda_{b}^{a}, \tag{C8}$$

where $\lambda_{b}^{a} = \theta^{a}([E_{b}, E])$ and $R^{(n)} = \gamma^{bd} R_{bd}^{(n)}$; $R_{bd}^{(n)} = R^{(n)a}{}_{bad}$.

APPENDIX D: COMPONENTS OF THE CURVATURE TENSOR FOR A DECOMPOSITION INDUCED BY A GROUP OF ISOMETRIES

The curvature tensor and its contractions with respect to the basis (9.7) for the metric (9.9) have the form,

$$R_{dcab}^{(m+n)} = R_{dcab}^{(n)} + S_{ic[a} S_{b]d}^{i}, \tag{D1}$$

$$R^{(m+n)}_{icab}=S^{k}_{c[a}F_{b]ki}+S_{icd}C^{d}_{.ba}+2S_{id[a}\gamma^{d}_{.b]c}\,, \tag{D2}$$

$$R^{(m+n)}_{ickb}=-S_{ibc;k}+S_{ibd}S_{kc.}{}^{d}+\tfrac{1}{4}F_{ckj}F_{bi.}{}^{j}-\tfrac{1}{2}\gamma^{d}_{.bc}F_{dki}\,, \tag{D3}$$

$$R^{(m+n)}_{ajkl}=F_{aj[l;k]}+F_{bj[k}S_{l]a}{}^{b}+F_{bkl}S^{b}_{ja}\,, \tag{D4}$$

$$R^{(m+n)}_{ijkl}=R^{(m)}_{ijkl}+\tfrac{1}{2}F_{ai[k}F^{a}_{.l]j}-\tfrac{1}{2}F_{aij}F^{a}_{kl}\,, \tag{D5}$$

$$R^{(m)i}_{jlk}=2e_{[k}\Delta^{i}_{l]j}+2\Delta^{m}_{j[l}\Delta^{i}_{k]m}\,,\quad R^{(n)d}_{.cab}=2\gamma^{d}_{q.[a}\gamma_{b]c.}{}^{q}-C^{qd}_{..c}\gamma_{aqb}\,, \tag{D6}$$

$$R^{(m+n)}_{ab}=R^{(n)}_{ab}-S^{i}_{ab;i}-S^{i}_{ab}S_{i}+2S^{i}{}_{ac}S^{\ c}_{ib}+\tfrac{1}{4}F_{aij}F_{b}{}^{ij}\,, \tag{D7}$$

$$R^{(m+n)}_{ai}=\tfrac{1}{2}F_{ai;k}{}^{k}+\tfrac{1}{2}F_{ail}S^{l}+C^{d}_{.ba}S^{b}_{id}-C^{b}_{bd}S^{d}_{ia}\,, \tag{D8}$$

$$R^{(m+n)}_{ik}=R^{(m)}_{ik}-S_{(i;k)}-S_{iab}S_{k}{}^{ab}+\tfrac{1}{2}F_{aij}F^{aj}_{..k}\,, \tag{D9}$$

$$R^{(m+n)}=R^{(n)}+R^{(m)}-2S^{i}_{;i}-S^{i}S_{i}-S^{iab}S_{iab}-\tfrac{1}{4}F^{a}_{ij}F_{a}{}^{ij}\,. \tag{D10}$$

Here $R^{(m)}=h^{ik}R^{(m)}_{ik}$; $R^{(m)}_{ik}=R^{(m)l}_{ilk}$, and $R^{(n)}=\gamma^{bd}R^{(n)}_{bd}$; $R^{(n)}_{bd}=R^{(n)a}_{bad}$. The covariant derivative in the direction of the vector e_k with respect to the connection Δ^{i}_{jk} is denoted by '';*k*.''

[1] L. P. Eisenhart, *Riemannian Geometry* (Princeton University Press, Princeton, New Jersey, 1949).
[2] J. A. Schouten and D. J. Struik, *1938 Einführung in die Neueren Methoden der Differentialgeometrie* (Zweiter band Geometrie von D. J. Struik, Groningen-Batavia, Noordhoff, Germany, 1938).
[3] D. Vagner, *Differential Geometry of Nonholonomic Manifolds, Proceedings on Vector and Tensor Analysis* (Moscow-Leningrad, OGIZ-GITTL, 1935), Issue 2-3 (in Russian).
[4] Th. Kaluza, Sitzungsber. K. Preus. Akad. Wiss. Berlin, Math. Phys. K1, 966 (1921).
[5] O. Klein, Z. Phys. **37**, 895 (1926).
[6] V. De Sabbeta and E. Schmutzer, *Unified Field Theories in More Than Four Dimensions* (World Scientific, Singapore, 1982).
[7] H. C. Lee, *An Introduction to Kaluza–Klein Theories* (World Scientific, Singapore, 1984).
[8] Yu. S. Vladimirov, *Physical Space–Time Dimensions and Unification of Interactions* (Moscow University Press, Moscow, 1987).
[9] V. D. Ivashchuk and V. N. Melnikov, Grav. Cosmol. **1**, 2 (1995); **1**, 133 (1995).
[10] C. Echart, Phys. Rev. **58**, 919 (1940).
[11] A. L. Zelmanov, Dokl. Akad. Nauk SSSR **107**, 815 (1956); in Proceedings of the 6th Conference on Cosmog (Izd. Akad. Nauk SSSR, Moscow, 1959), p. 144.
[12] A. L. Zelmanov, Dokl. Akad. Nauk SSSR **209**, 822 (1973).
[13] A. L. Zelmanov, Dokl. Akad. Nauk SSSR **227**, 78 (1976).
[14] A. L. Zelmanov and V. G. Agakov, *The Elements of General Relativity* (Nauka, Moscow, 1989) (in Russian).
[15] Yu. S. Vladimirov, *The Frames of Reference in General Relativity* (Energoizdat, Moscow, 1982) (in Russian).
[16] R. F. Polishchuk, Dokl. Akad. Nauk SSSR **209**, 76 (1973).
[17] V. I. Antonov, V. N. Efremov, and Yu. S. Vladimirov, Gen. Relativ. Gravit. **9**, 9 (1978).
[18] R. Arnowitt, S. Deser, and C. W. Misner, in *Gravitation, An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), p. 227.
[19] R. Geroch, J. Math. Phys. **12**, 918 (1971); P. Kundu, Phys. Rev. D **18**, 4471 (1978).
[20] F. J. Ernst, Phys. Rev. **167**, 1175 (1968); S. Kloster, M. M. Som, and A. Das, J. Math. Phys. **15**, 1096 (1974); R. H. Gowdy, Ann. Phys. (N.Y.) **83**, 1, 203 (1974).
[21] D. Kramer, M. Stephani, Maccalum, and E. Herlt, *Exact Solutions of the Einstein Field Equations*, edited by E. Schmutzer (Deutscher Verlagder Wissenschaften, Berlin, 1980).
[22] A. Lichnerowicz, *Théories Relativistes de la Gravitation et de L'Électromagnetisme* (Masson Cie, Paris, 1955).
[23] S. Sternberg, *Lectures on Differential Geometry* (Prentice–Hall, Englewood Cliffs, 1964); D. Gromol, W. Klingenberg, and W. Meyer, *Riemannishe Geometrie im Groben* (Springer, Berlin, 1968); V. A. Dubrovin, S. P. Novikov, and A. T. Fomenko, *Modern Geometry: Methods and Applications* (Nauka, Moscow, 1979) (in Russian).
[24] S. Koboyashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1969), Vols. 1 and 2.
[25] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973); C. W. Misner, K. S. Thorn, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[26] A. Gray, J. Math. Mech. **16**, 715 (1967).
[27] V. D. Gladush, Izv. Vyssh. Uchebn. Zaved. Fiz. **N11**, 58 (1979) [Sov. Phys. J. **22**, 1172 (1979)].
[28] V. D. Gladush, in *Theses of Rep. of V Sov. Grav. Conf.* (Moscow University Press, Moscow, 1981), p. 126; in *Theses of Rep. of VI Sov. Grav. Conf.* (Patrice Lumumba University Press, Moscow, 1984), p. 36.

[29] V. D. Gladush, in *Investigations on Classical and Quantum Gravity* (Dniepropetrovsk University Press, Dniepropetrovsk, 1983), pp. 28–36.
[30] V. D. Gladush, in *Abstracts of Contributed Papers of the 14th Int. Conf. on GRG* A104 (Florence, Italy, 1995).
[31] S. S. Chern, Bull. Am. Math. Soc. **72**, 167–219 (1966); K. Yano, *Differential Geometry on Complex and Almost Complex Spaces* (Pergamon, Oxford, 1965); S. Kobayashi, *Transformation Groups in Differential Geometry* (Springer, Berlin, 1972).
[32] A. Borowiec, M. Ferraris, M. Francaviglia, and I. Volovich, dg-ga/9612009, 1996.
[33] M. Fecko, J. Math. Phys. **38**, 4542 (1997).
[34] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, Oxford, 1971).
[35] M. P. Ryan, J. Math. Phys. **15**, 812 (1974).
[36] O. I. Bogoyavlensky, *The Methods of the Qualitative Theory for Dynamical Systems in Astrophysics and Gas Dynamics* (Nauka, Moscow, 1980) (in Russian).
[37] Y. M. Cho and G. O. Freund, Phys. Rev. D **12**, 1711 (1975).
[38] V. D. Gladush, Izv. Vyssh. Uchebn. Zaved. Fiz. **N3**, 74 (1980) [Sov. Phys. J. **32**, 240 (1980)].
[39] F. I. Frankel, Zh. Eksp. Teor. Fiz. **31**, 490 (1956); N. Sibgatullin, *Oscillations and Waves in Strong Gravitational and Electromagnetic Fields* (Nauka, Moscow, 1984).
[40] A. Krasiński, Acta Phys. Pol. B **5**, 411 (1974); **6**, 223 (1975).

Quantifying excitations of quasinormal mode systems

Hans-Peter Nollert
Astronomy and Astrophysics, University of Tübingen, 72076 Tübingen, Germany

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 10 September 1998; accepted for publication 7 October 1998)

Computations of the strong field generation of gravitational waves by black hole processes produces waveforms that are dominated by quasinormal (QN) ringing, a damped oscillation characteristic of the black hole. We describe here the mathematical problem of quantifying the QN content of the waveforms generated. This is done in several steps: (i) We develop the mathematics of QN systems that are complete (in a sense to be defined) and show that there is a quantity, the ''excitation coefficient,'' that appears to have the properties needed to quantify QN content. (ii) We show that incomplete systems can (at least sometimes) be converted to physically equivalent complete systems. Most notably, we give a rigorous proof of completeness for a specific modified model problem. (iii) We evaluate the excitation coefficient for the model problem, and demonstrate that the excitation coefficient is of limited utility. We finish by discussing the general question of quantification of QN excitations, and offer a few speculations about unavoidable differences between normal mode and QNM systems. © *1999 American Institute of Physics.* [S0022-2488(99)03201-6]

I. INTRODUCTION AND OUTLINE

Essentially all computations of the generation of gravitational waves by strong field black hole processes produce a gravitational wave with the shape of a damped sinusoid.[1] The oscillation period and damping time depend only on the parameters of the black hole, and not on the manner of excitation. The meaning of the complex frequency of this damped oscillation is now well understood. A single frequency perturbation outside the hole can satisfy the natural radiative boundary conditions (radiation into the black hole and outward to infinity) only if the frequency is one of the discrete set of frequencies, called quasinormal (hereafter QN) frequencies. The least damped of these complex frequencies is what dominates the appearance of computed waveforms.

QN excitations are relevant, in principle, to most or all systems with radiative boundary conditions. Stellar models, for example, have short periods for nonradial oscillations driven by fluid pressures, and long damping times of these fluid oscillations due to the weak emission of gravitational waves. The motions of the stellar fluid can be studied with radiation damping omitted (e.g., with the use of Newtonian gravitation theory, or post-Newtonian theory) and the weak radiation can be added, after the fact. When the radiative coupling is ''turned off'' the problem of the oscillation of a perfect fluid stellar model can be analyzed in normal modes[2,3] and one can find the radiated energy coming from each separate oscillation frequency, and can decompose the radiative power into that fraction assigned to each frequency.

The situation is dramatically different for black holes, which have only a single time scale. (For a nonrotating hole this is $2GM/c^3$ where G is the universal gravitational constant, c is the speed of light, and M is the mass of the hole.) The period and damping time are therefore of the same order and there is no meaningful way of turning off the damping for black hole oscillations; there is no underlying normal mode system. This suggests that there may be no clear way of specifying ''how much QN ringing'' of some particular black hole QN frequency is contained in an emitted waveform. This suggestion is made plausible by the mathematical origins of normal modes and QN modes. The properties and usefulness of normal modes are closely related to the fact that they are eigensolutions to a self adjoint problem. QN modes, on the other hand, are eigensolutions of a problem that is not self adjoint. But the dominance of QN frequencies in computed waveforms is so robust that it seems that the strength of QN ringing *must* be quantifiable, or at least that mathematical sense must be made of the question.

In this paper we try to make mathematical sense of quantification. In attempting this we draw upon parallels with normal modes systems. By the ''excitation'' of a mode we mean, in parallel to excitation in normal modes systems, an index of the contribution that each mode makes to the overall waveform and to the energy. To develop a description of QN excitation we start with a viewpoint that a meaningful and rigorous quantification is very implausible unless the QN system is, in some sense, complete. We then follow a three step process. First, in Sec. II, we define and posit the existence of QN systems that are complete (in a sense to be defined). We then point out difficulties in quantifying excitation in a complete QN system. We construct a particular measure, the ''excitation coefficient,'' that overcomes these difficulties, and is closely related to the description of the excitation of normal modes.

Our next step is to prove the existence of complete QN systems and relate the mathematics of black hole processes to complete systems. This step, carried out in Sec. III, requires a rather lengthy discussion of ''induced completeness.'' Though this discussion is not directly related to the problem of QN excitation, it is a necessary step (and is interesting in its own right). The discussion in Sec. III shows that completeness can be induced. That is, an incomplete QN system can be changed with a modification that satisfies two criteria: (i) The effect of the modification can be made arbitrarily weak. More specifically, the modification can be made small enough so that the waveform that evolves from any initial conditions is arbitrarily close to the waveform evolving with no modification. (ii) No matter how weak the modification is, the modified QN system is complete. Our demonstration in Sec. III does not consist of a general theory for such modifications; a conjecture about the general conditions has been given by Young *et al.*[4] (though their definition of completeness is somewhat different from ours). Here we will sacrifice generality and direct astrophysical relevance for specificity and rigor. We present the details of a specific model. We will start with a model, the ''TDP,'' with only a single conjugate pair of QN frequencies, and modify it to the ''spiked TDP,'' a model with an infinite QN spectrum. In the Appendix we give a rigorous proof of completeness of the spiked TDP (i.e., that under specified circumstances the outgoing waveform is a convergent sum of components at quasinormal frequencies). Numerical results are shown in Sec. III to demonstrate the negligible effect of the modification, and to demonstrate the pattern of convergence of sums of single frequency excitations.

Having established that completeness can be induced (at least in one model problem), we return, in Sec. IV, to the question of measuring the excitation of QN modes and, in particular, to the excitation coefficient, introduced in Sec. II. We demonstrate, with a few examples, that this formal measure of excitation does not generally give a useful quantification. The failure of this measure is discussed along with a broader discussion of differences between QN and normal mode systems, and conjectures about mathematical properties of QN systems.

II. COMPLETE QN SYSTEMS

A. Definition of QN frequencies

For definiteness we will limit considerations to solutions of the equation,

$$\frac{\partial^2\Psi}{\partial x^2}-\frac{\partial^2\Psi}{\partial t^2}-V(x)\Psi=0. \tag{1}$$

Such an equation describes the dynamics of many mechanical systems, and the evolution of multipole perturbations (scalar, electromagnetic, or gravitational) of spherically symmetric

(Schwarzschild) black holes.[5–7] Perturbations of rotating (Kerr) holes,[7] on the other hand, cannot be reduced to radial-time equations. QN oscillations are single frequency solutions of the form $\Psi(t,r)=\psi(x)\exp(i\omega t)$ and hence are solutions of the equation

$$\frac{\partial^2\Psi}{\partial x^2}+[\omega^2-V(x)]\Psi=0. \tag{2}$$

We assume that the domain of x includes $x=\infty$, and that the nature of the potential $V(x)$ is such that a ''radiative boundary condition'' can be defined at $x\rightarrow\infty$. A clear example is a potential with support of $V(x)$ only for x less than some x_{max}. In this case the boundary condition is that $\psi(x)\propto\exp(-i\omega x)$ for $x>x_{max}$.

For potentials that do not vanish, but fall off sufficiently fast as $x\rightarrow\infty$, the more general radiative boundary condition for real frequencies will be that for large x, the solution for $\psi(x)$ have the form $\exp(-i\omega x)F(\omega,x)$, with $F\rightarrow$constant, as $x\rightarrow\infty$. This condition is not quite sufficient if ω has a positive imaginary part; see Ref. 8 for a complete discussion. In short, the solution satisfying a radiative boundary condition for complex ω can be regarded as an analytical continuation of a solution satisfying a radiative boundary condition for real ω.

The boundary condition at the other end of of the x domain may be a standard Sturm–Liouville boundary condition (e.g., $\Psi=0$ at $x=0$) or may be a radiative boundary condition at $x\rightarrow-\infty$. For spherically symmetric black holes, the range of x extends from $-\infty$ to ∞, with $-\infty$ representing the black hole horizon.[8,9] The potentials fall off exponentially in x as $x\rightarrow-\infty$ and as const/x^2, as $x\rightarrow\infty$. Radiative boundary conditions are imposed both at $x\rightarrow-\infty$ and $x\rightarrow\infty$, corresponding to radiation moving inward through the black hole horizon, and outward towards spatial infinity.

QN frequencies are the eigenvalues $\omega=\omega_{QN}$ to the problem defined by (2) for radiative boundary conditions of the type just discussed. Due to the boundary conditions, this problem is generally not of the Sturm–Liouville type and the usual features of eigenvalues of a Sturm–Liouville problem are absent. In particular, the QN frequencies ω_{QN} are generally not real. A positive imaginary part indicates an exponential decrease with time. A negative imaginary part would indicate an instability; no frequencies with negative imaginary parts have been found for black hole QN systems.

It is clear that QN frequencies must occur in conjugate pairs. If ω_{QN} is a solution to the eigenproblem corresponding to ψ_{QN}, then $-\omega_{QN}^*$ is also a solution corresponding to ψ_{QN}^*. We will use a tilde (~) to denote the conjugate relationship of QN frequencies. Thus the conjugate to QN frequeny ω_7 is $\omega_{\tilde{7}}$, that is $\omega_{\tilde{7}}=-\omega_7^*$.

B. Definition of completeness

We will choose to give a rather specific meaning to a complete system of QN modes. The Cauchy data for (1) consists of ψ_0 and $\dot{\psi}_0$, the initial value of Ψ and of its time derivative,

$$\psi_0(x)\equiv\Psi(t,x)|_{t=0}, \quad \dot{\psi}_0(x)\equiv\partial\Psi(t,x)/\partial t|_{t=0}. \tag{3}$$

We consider an interval $x_2<x<x_1$ in the domain of x and we consider Cauchy data at $t=0$ for (1) which has support only in this interval. We then consider the solutions Ψ to (1) for such data, and we focus attention on the value of this solution at $x_{\rm obs}$, a particular value of x satisfying $x_{\rm obs}>x_1$. This corresponds to the physical situation of an observer at $x_{\rm obs}$ detecting radiation resulting from an initial disturbance located at some distance from her. The ''observed waveform'' that we focus on is then

$$f(t)\equiv\Psi(x_{\rm obs},t). \tag{4}$$

As shown in Fig. 1, there will, in general, be a minimum value $t_{\rm min}$ of t, such that the point $x_{\rm obs},t$ is influenced by the evolved Cauchy data. That is, for $t<t_{\rm min}$, the area between the past directed

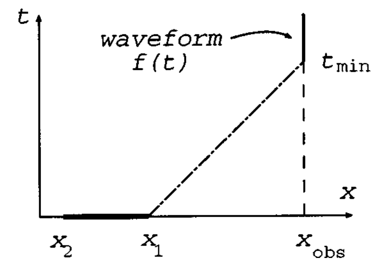

FIG. 1. Propagation of initial data to the observation location $x_{\rm obs}$ defining the waveform $f(t)$.

characteristics from $t,x_{\rm obs}$ intersects the $t=0$ hypersurface outside the support of the Cauchy data. We are interested in $f(t)$ only for $t\geqslant t_{\rm min}$. The physical interpretation of this is that we are considering only the waveform generated by the Cauchy data.

We take a complete QN system to be one which satisfies the following criteria.

- The solutions to the QN eigenvalue problem form a discrete spectrum and can be arranged in order of increasing $|\Re(\omega_n)|$.
- We consider only Cauchy data that
  –has support only within a compact region $[x_2,x_1]$,
  –belongs to a specific continuity class C^p, where p depends on the nature of the problem,
  –results in a waveform which is square integrable from $t=t_{\rm min}$ to ∞.
- For such Cauchy data, the waveform $f(t)$ that evolves from any such allowed Cauchy data can be written as

$$f(t)=\sum_n a^n e^{i\omega_n t}. \tag{5}$$

Here a^n is the nth coefficient in the sum over QN modes. Since $f(t)$ is a function of $x_{\rm obs}$, the a^n coefficients are also functions of $x_{\rm obs}$, but we shall not explicitly exhibit this dependence. The summation in (5) is in order of increasing $|\Re(\omega_n)|$, and the convergence is uniform for $t>t_{\rm min}$.

It is important to note that our view of completeness is rather different from other possible meanings of the term. In particular, our choice of the meaning of completeness has nothing directly to do with the x-dependence of the single frequency solutions and with the question of whether these solutions can be used to span acceptable Cauchy data. Our meaning of completeness, then, is rather different from that of Young *et al.*[4] It also differs from the concept of completeness used by Husain and Price[10] and by Beyer,[2] and Beyer and Schmidt,[3] but is closely related to the completeness used by Beyer[11] in his work on the Pöschl–Teller potential.[12]

C. Function space and inner product

In accordance with our definition of completeness, our vector space is the space of all functions $f(t)$, $t\geqslant t_{\rm min}$ that can evolve from acceptable Cauchy data. Our class of acceptable Cauchy data will always be chosen so that $f(t)$ is square integrable from $t=t_{\rm min}$ to ∞. On this space of functions we define an inner product to be

$$f\cdot g\equiv\int_{t_{\rm min}}^{\infty} f^*(t)g(t)dt. \tag{6}$$

We could, of course, include a weight function $W(t-t_{\rm min})$ in the integral defining the inner product, but the time translational symmetry of the background suggests that W should be constant. The choice in Eq. (6), furthermore, means that $f\cdot f$ is the time integral of the square of the wave function, a measure closely related to the energy content of a wave. (For black hole processes, the connection with gravitational wave power will be made explicit presently.) Our assumption of completeness above means that the functions $\exp(i\omega_{QN}t)$, while not elements of our function space themselves, span this function space in the sense of (5); we will therefore consider them a basis. For a function $f(t)$ in our space we can use the inner product to compute another set of coefficients a_n by

$$a_n \equiv (e^{i\omega_n t}) \cdot f(t) = \int_{t_{\min}}^{\infty} e^{-i\omega_n^* t} f(t) dt. \tag{7}$$

The following relations for the coefficients of conjugate modes are straightforward to verify:

$$a_{\tilde{k}} = (a_k)^*, \quad a^{\tilde{k}} = (a^k)^*. \tag{8}$$

Since the convergence is uniform by hypothesis, we can integrate term by term in the sum-of-modes expression for the norm of f to find

$$\int_{t_{\min}}^{\infty} |f(t)|^2 \, dt = \int_{t_{\min}}^{\infty} \left(\sum_n a^n e^{i\omega_n t} \right)^* f(t) dt = \sum_n (a^n)^* a_n . \tag{9}$$

The final sum in (9) is real, as it must be, since for any k, the sum $(a^k)^* a_k + (a^{\tilde{k}})^* a_{\tilde{k}}$ is real.

In most physical problems the radiated power is the square of the time derivative of the waveform. If at x_{obs} our wave function evolving from the initial data is $f(t) \equiv \psi(x_{\text{obs}}, t)$. If $f(t_{\min})$ vanishes (i.e., if the waveform starts continuously) then this type of energy can be evaluated as

$$\int_{t_{\min}}^{\infty} |f(\dot{t})|^2 \, dt = \sum_n (\omega_n^*)^2 (a^n)^* a_n . \tag{10}$$

As in (9) the reality of the sum is guaranteed by the relations of conjugate coefficients in (8).

Since we have an inner product, we have an equivalence between vectors and dual vectors in our function space and we can define a set of covariant basis functions $\phi^m(t)$ by the property $\phi^m(t) \cdot e^{i\omega_n t} = \delta_{nm}$. If follows from Eq. (7) that the a_n are the expansion coefficients for $f(t)$ with respect to the covariant basis functions ϕ^n. We shall henceforth refer to a^n and a_n, respectively, as the contravariant and covariant coefficients of $f(t)$. The components of the metric, in this function space, with respect to the QN basis, are

$$(e^{i\omega_n t}) \cdot (e^{i\omega_k t}) = \int_{t_{\min}}^{\infty} e^{it(\omega_k - \omega_n^*)} \, dt \equiv G_{nk} . \tag{11}$$

It should be noted that G is a Hermitian matrix, but it is not diagonal, i.e., the QN oscillations are not orthogonal. The metric coefficients can be used, in principle, to relate a^n and a_n. The expression for $f(t)$ in Eq. (5) can be substituted in Eq. (7). Since the convergence in Eq. (5) is uniform, we can integrate term by term and get

$$a_n = \sum_k G_{nk} a^k. \tag{12}$$

Since G_{nk} is not diagonal, the covariant basis vectors, i.e., the basis vectors dual to $e^{i\omega_n t}$, are mixtures of the $e^{i\omega_n t}$ functions (usually involving all of them). An indication of the unfamiliar problems this produces can be seen in the following rough argument. Let us suppose that we have a waveform that in some sense is ''almost pure'' (say) seventh mode. That is, suppose that $f(t) \approx a^7 e^{i\omega_7 t} + a^{\tilde{7}} e^{i\omega_{\tilde{7}} t}$. (We are supposing that this relationship is only approximate since it will, in general, be impossible to excite a truly pure single frequency mode with smooth, compact initial

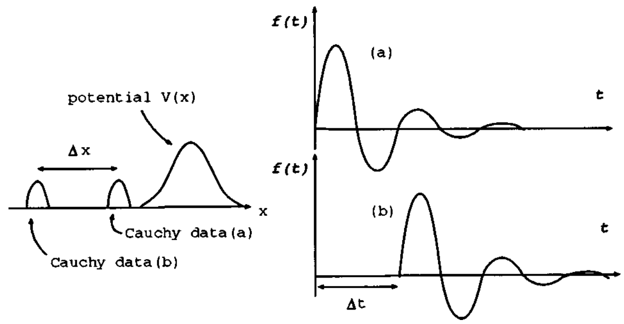

FIG. 2. Initial data shifted in location produces identical waveforms shifted in time.

data.) This waveform, for which (almost) the only contravariant coefficients are a^7 and $a^{\tilde{7}}$, will have contravariant coefficients a^n for all n. A waveform that is a pure (or almost pure) single mode excitation in one sense is therefore not a single mode excitation in another. This presages some of the problems in quantifying the excitation of a mode, and we will return to this point at the end of Sec. IV.

It is possible, of course, to use Gram–Schmidt orthogonalization to find basis functions which are orthogonal according to the inner product of (6). The resulting basis functions will not (except for one of them) correspond to single frequency excitations, and do not seem to be of interest.

D. Intuitive insights; the excitation coefficient

Some rough considerations of model problems suggest the intuitive basis for some of the mathematical difficulties to appear below, and point to a possible approach to quantifying excitation. The most obvious difficulty is the ''time shift'' problem. We imagine a configuration like that pictured in Fig. 2: A potential with compact support and two sets of Cauchy data (a) and (b). The two Cauchy data sets are localized and are identical except that set (b) is shifted to the left, to smaller x, by some finite displacement Δx. The support of neither Cauchy data set overlaps the support of the potential. In this case it is clear that the waveform generated from the two cases will be identical except that the waveform from (b) will be shifted to later times, relative to that for (a), by an amount $\Delta t = \Delta x$. Each contravariant coefficient will therefore be larger in the (b) waveform than in the (a) waveform. The conjugate pair of contravariant coefficients $\{a^7, a^{\tilde{7}}\}$ will, for example, be larger by $\exp[(\Im\omega_7)\Delta t]$ for (b) than for (a), though the excitation is physically identical. The analogous difficulty does not arise for normal modes; since they are not damped, the time delay only causes a phase shift. The trend is opposite for the covariant coefficients $\{a_7, a_{\tilde{7}}\}$. As defined in (7) $\{a_7, a_{\tilde{7}}\}$ will be smaller for (b), since $\exp(-i\omega_7^* t)$ (or $\exp(-i\omega_{\tilde{7}}^* t)$) are smaller by $\exp[(-\Im\omega_7)\Delta t]$ at the later times during which the (b) waveform has support. Therefore, neither the coefficients a^k nor a_k alone can provide a useful measure for the excitation of a QN mode.

Any useful measure of excitation *must* give the same result for the two waveforms in Fig. 2. We take advantage of the opposite tendencies of the contravariant and covariant coefficients under a time shift to define a quantity that is the same for both waveforms, the ''excitation coefficient'' A_k for the kth QN mode:

$$A_k \equiv (a_k)^* a^k + (a_{\tilde{k}})^* a^{\tilde{k}}. \tag{13}$$

We conjecture that these excitation coefficients, or quantities constructed from them (sums of excitation coefficients, functions of excitation coefficients, etc.), or quantities very closely related (see the energy excitation coefficient, below), are the only relevant mathematical objects in the vector space that are unaffected by the time shifts, and have the additional properties that we outline in the following.

In addition to the insensitivity of the excitation coefficient to time shifts of waveforms, the excitation coefficient has another important and relevant property [cf. Eq. (9)]:

$$\int_{t_{\min}}^{\infty}|f(t)|^2dt=\sum_k A_k, \tag{14}$$

where the sum is over conjugate pairs. The excitation coefficients of the complete set of QN modes sums to the norm of the waveform. We can define a quantity closely related to A_k:

$$E_k\equiv(\omega_k^*)^2(a_k)^*a^k+(\omega_k)^2(a_{\tilde{k}})^*a^{\tilde{k}}. \tag{15}$$

We shall refer to the E_k as the energy excitation coefficient. According to (10), the sum over conjugate pairs of these coefficients gives the norm of $\dot{f}$. The summation properties of the A_k and the E_k are important in clarifying what we mean by the ''excitation'' we are attempting to quantify. These coefficients appear to tell us the contribution made by each mode to a measure of the waveform. In the case of black hole perturbations it turns out that it is possible to make an even more direct connection. If $f(t)$ is a solution of the Zerilli equation[6] for even parity perturbations, then the integral on the left of (10) is proportional to the radiated energy and E_k has the appearance of the energy in the kth mode. If $f(t)$ is a solution of the Regge–Wheeler equation[5] for odd parity perturbations, then the integral in Eq. (9) is the energy and A_k has the appearance of the energy in the kth mode.

The possibility of quantification with the excitation coefficient (or energy excitation coefficient) will be a central focus, of Sec. IV, but before we begin specific computations, there are a few more possibly useful insights that can be found from intuitive considerations. For one thing, it is interesting that the two sets of coefficients can be related to two different aspects of an emitted wave. The contravariant coefficients, telling us how much of a certain mode must be added in order to get the waveform, can be considered a ''theoretical'' coefficient. For a given waveform, the projection operation on the waveform defined by (7) can be considered to give the ''experimental'' excitation coefficient.

Simple considerations can also produce insights into the nature of the metric matrix G_{nk} defined in (11). Consider the signal produced by some Cauchy data, and let the vector $\overline{a}$ denote the contravariant coefficients of the resulting waveform with respect to the QN mode basis. Let the vector of covariant coefficients be given by $\underline{a}$, so that $\underline{a}=G\overline{a}.$ Now consider the same Cauchy data, shifted to the left by Δx. The new contravariant and covariant coefficient vectors will be given by: $\tilde{\bar{a}}^k=\exp(-i\omega_k\Delta x)\overline{a}^k$ and $\tilde{\underline{a}}_k=\exp i\omega_k)\Delta x)\underline{a}_k$. Taking the standard linear algebra norm, we have $\|\tilde{\bar{a}}\|=\alpha\|\overline{a}\|$ and $\|\tilde{\underline{a}}\|=\beta\|\underline{a}\|$, where $\alpha>1$ and $\beta<1$.

The exact magnitude of α depends on the distribution of the contravariant coefficients and the shift Δx. In fact, we can always make the shift large enough so that $\alpha\gg1$. In the same way, we can ensure that $\beta\ll1$. Assuming that the metric matrix G has a minimum eigenvalue $e_{\min}$ and a maximum eigenvalue $e_{\max}$, we then find

$$e_{\min}\leqslant\frac{\|\tilde{\underline{a}}\|}{\|\tilde{\bar{a}}=\|}=\frac{\beta}{\alpha}\,\frac{\|\underline{a}\|}{\|\overline{a}\|}\ll\frac{\|\underline{a}\|}{\|\overline{a}\|}\leqslant e_{\max}. \tag{16}$$

Since G is an infinite dimensional matrix there need not be a maximum or a minimum eigenvalue. For the model problem introduced in Sec. III, we will give numerical evidence, in Sec. IV B, that the ratio of maximum and minimum eigenvalues diverges. The G matrix therefore is in some sense singular.

We now make the additional assumption that we have constructed Cauchy data such that the waveform consists almost exclusively of ringing in a single QN mode. (In the specific example discussed in the following sections, this will be possible for the fundamental mode using an arbitrarily short ''burst'' of initial data.) Let us modify the example of Fig. 2, by reducing the size of Cauchy data (b) by a factor $\exp(\Im\omega_1)\Delta x$, so that once the response to the (b) data starts it is

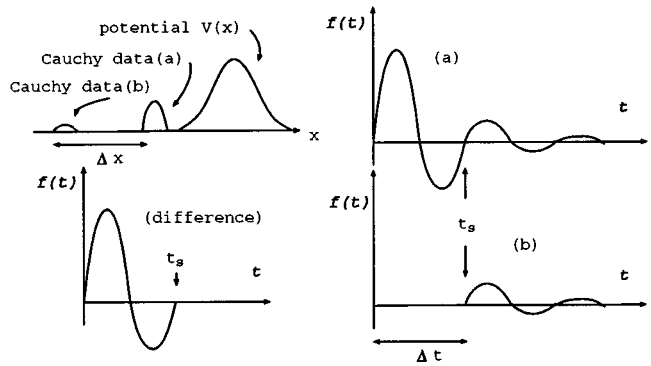

FIG. 3. Oscillations produced by shifting and scaling initial data.

identical to that of the (a) response at the same time. This situation is shown in Fig. 3 in which the response $f(t)$ is the same in both plots for $t \geqslant t_s = \Delta t$. Now let us suppose that the waveforms are dominated by ringing of the fundamental mode, and that $f(t)$ consists only of ringing of the fundamental mode. The difference between the two responses is a sum of QN modes which has the appearance of truncated fundamental mode ringing, but which has $a^1 = \tilde{a}^1 = 0$ in its mode sum. The time t_s can, in principle, be made arbitrarily long, so the truncation in the difference curve can be made arbitrarily late. We then have a sum of modes that looks arbitrarily close to fundamental mode ringing, but containing no fundamental mode. This suggests that the fundamental mode can ''almost'' be built as a superposition of the modes other than the fundamental mode, and that the modes are ''almost'' linearly dependent. We will discuss this further, based on numerical results, in Sec. IV C.

Again, this suggests that the infinite metric matrix G_{nk} is in some sense ''almost'' singular. A physical insight can be associated with this nature of the metric: Since G_{nk} is ''almost singular'' it is ''almost noninvertible.'' This means that there are difficulties in finding the contravariant coefficients from the covariant coefficients, since this requires the inverse of G_{nk}. The implication of this is possibly of pragmatic importance: the near singularity makes it difficult to find the ''theoretical'' coefficients from the ''experimental'' ones.

III. INDUCED COMPLETENESS

A. Small change

Our approach to quantification of excitation is based on the idea of complete QN systems. It is important to demonstrate that complete QN systems exist and to have an example of a system in which we can compute excitation coefficients. We must also deal with a more specific issue. The QN modes of potentials for black hole problems do not form a complete set; both the Regge–Wheeler potential, which describes perturbations of a Schwarzschild black hole, and the Pöschl–Teller potential,[12] which has been used as an approximation to the black hole potential,[13] have quasinormal mode sets which cannot completely describe the waveform resulting from an initial perturbation. An important question is what the relevance is to black hole processes, of any quantification based on an assumption of complete systems. The details of this section are somewhat disjoint from the main goal of the paper: quantification of QN excitation, but are a necessary step in developing our approach. (They are also rather interesting in their own right!)

The key idea in developing a demonstrably complete QN system, and in showing its relevance to black holes, is ''induced completeness.'' It has been argued[4] that for the problem defined by (2), the eigenmodes will be complete if there are two values of x at which $V(x)$ is not C^∞. If this is true, it appears that the black hole potentials can be modified in such a way that completeness of

QN oscillations is induced, while the effect on other aspects of the problem—in particular on the evolution of Cauchy data—is kept arbitrarily small. In this section we explore induced completeness.

It is not our purpose here to look into the generality of induced completeness. Rather, we want to have a specific example of a complete QN system with which to explore the question of quantifiability of QN excitations. We will therefore focus on a very specific model that contains most of the flavor of black hole potentials, but is simple enough to allow rather straightforward analysis and a proof of completeness.

We use the fact that (2) can be solved with elementary functions in the case that the potential has the form of a truncated centrifugal potential,

$$V_{\mathrm{TCP}}(x)=\begin{cases}0, & x<x_0,\\ l(l+1)/x^2, & x\geqslant x_0,\end{cases} \tag{17}$$

where l is an integer. It gives a good representation of the sharp cutoff of the black hole potential as $x\to-\infty$ and the approximate asymptotic form of the black hole potential at $x\to\infty$. It differs from the true black hole potential in the details of the $x\to\infty$ potential that give rise to the power-law late time tails of black hole waveforms. These tails almost certainly are an obstacle to an analysis of QN excitation, and a modification of the potential to eliminate these tails has already been made in work on QN excitation.

We will hereafter consider only the case $l=1$, to be called the ''truncated dipole potential'' (TDP). An example of the simplicity of the TDP problem is that it has only a single pair of QN frequencies: $\omega_1=(1+i)/2x_0$ and $\omega_{\tilde{1}}=(-1+i)/2x_0$ (see Sec. 1 of Appendix A). There are several ways in which we could try to induce completeness into this problem. We have studied both the addition of a small step (with discontinuities) to the TDP and the addition of a delta function. The delta function has the disadvantage of its distributional nature, but it offers the advantage of considerable simplicity as compared with the step. We have found no significant ''practical'' difference between the results (QN locations, convergence of QN sums) between the two examples, so we will describe here completeness induced with a delta function. The total potential in this case will be called the ''spiked truncated dipole potential'' (STDP) and is given by

$$V_{\mathrm{STDP}}(x)=V_{\mathrm{TDP}}(x)+V_\delta\delta(x-x_\delta), \tag{18}$$

where $V_{\mathrm{TDP}}(x)$ is the $l=1$ form of the potential in (17).

We first establish that the influence on the evolution of initial data can be made arbitrarily small. To do this we choose the Cauchy data, at $t=0$, to be given by

$$\psi_0=\sin\left(2\pi\frac{x-x_2}{x_1-x_2}\right)\quad(x_2\leqslant x'\leqslant x_1),$$
$$\dot{\psi}_0=-\partial\psi_0/\partial x, \tag{19}$$

with $x_1\leqslant x_0$, so that the initial data has the form of one full cycle of a sine wave, located to the left of the potential, traveling to the right. (This Cauchy data is chosen for convenience in demonstrating mathematical points; it has no justification as initial data for gravitational waves being produced in the neighborhood of a black hole.)

For the results to be shown here and in following sections, we choose $x_0=1$, $x_1=1$, $x_2=-5$, and $x_\delta=10$, and $x_{\mathrm{obs}}=120$. For this choice of x_1 and x_2, the sine has a wavelength, 6, that is roughly half that of the QN oscillation of the TDP. This allows us to see similar, but distinguishable signs in the evolved waveform of the propagation of the Cauchy data and of QN oscillation.

Figure 4 shows the time evolution of the Cauchy data for the original TDP and for the TDP with an added δ-function with different amplitudes ranging from $V_\delta=1$ to $V_\delta=10^{-6}$. The waveforms are followed out to times at which they have decreased in magnitude from the maximum by

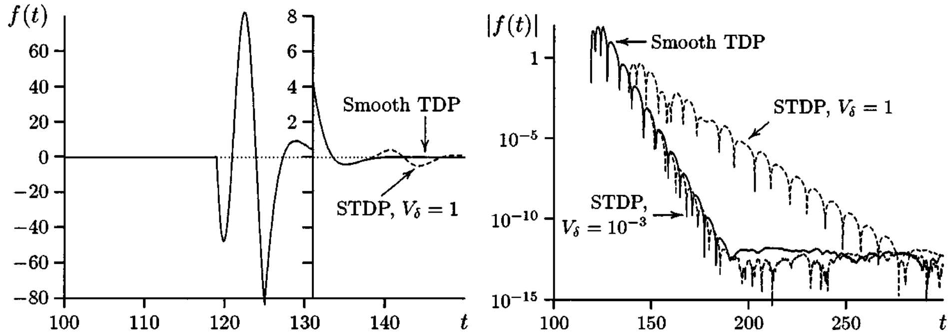

FIG. 4. Time evolution for initial data incident on the TDP and on the TDP with an added δ-function. Left: linear plot, right: logarithmic plot.

a factor of $\sim 10^{-15}$, at which point numerical error obscures the results. Even for the largest value of V_δ, the influence of the delta function on evolution is visible in the linear plot only after we change the scale for $f(t)$ at $t=131$. To see the effects more clearly we also plot the logarithm of $|f(t)|$ for a longer observation time, showing several interesting features. For all but the largest amplitude of the delta function ($V_\delta=1$) the waveform after the first few oscillations consists only of QN ringing, with the characteristic distance $\Delta t = |\Re\omega_7|\pi = 2\pi x_0 = 2\pi$ between zeros of $f(t)$. For $V_\delta=10^{-3}$, the effect of the delta function shows up only as a phase shift after the the first four, or so, full cycles of oscillation. The effect of the smallest delta function amplitude $V_\delta = 10^{-6}$ is smaller by two orders of magnitude than that for $V_\delta=10^{-3}$, and too small to be seen even in the logarithmic plot of Fig. 4.

These results make it clear that any reasonable measure of the influence of the delta function, such as the integral of the square deviation from the TDP waveform, is tiny and decreases with decreasing V_δ. We will accept these numerical results as a sufficient demonstration of this point, and will not attempt an analytic proof of this point.

B. QN spectra

Although the influence of a small delta function on the evolution of Cauchy data is negligible, the influence on the spectrum of QN frequencies is profound. The method of computing the QN frequencies for the STDP is outlined in Appendix A. The results of the computation are presented in Fig. 5 for the same potentials considered in Fig. 4.

The original TDP has only one pair of QN frequencies at $(\pm 1 + i)/2$. With an added δ-function, an entirely new set of modes appear. Note that the real parts of the additional frequen-

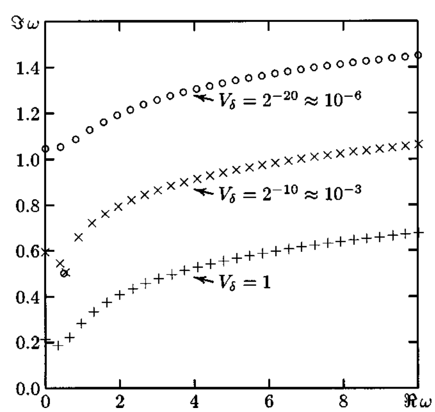

FIG. 5. Quasinormal frequencies for the TDP with an additional δ-function at $x_\delta = 10$.

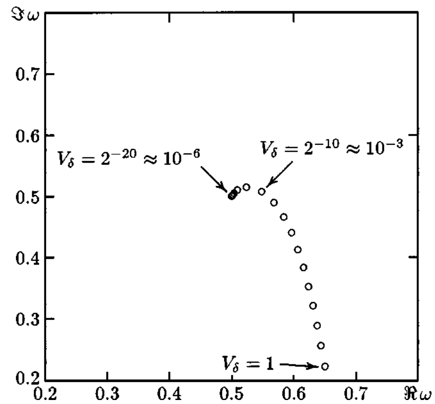

FIG. 6. ''Path'' of the original quasinormal frequency of the TDP as the amplitude of the additional δ-function is decreased.

cies seem to be unbounded. The asymptotic spacing $\Re(\omega_n) = 2\pi/10$ of the real parts of the frequencies is shown in Sec. 3 of Appendix A to correspond to the length of the ''cavity'' bounded by x_0 and the δ-function: $L \equiv x_\delta - x_0 = 10$. The imaginary parts are shown in Appendix A to increase as the logarithm of the real part.

When V_δ becomes sufficiently small, one of the QN frequencies approaches the value $(1 + i)/2x_0$ of the ''native'' QN frequency, that of the original, TDP. This can be seen more clearly in Fig. 6, which shows the ''path'' of this QN frequency in the complex plane for values of V_δ varying in steps of 1/2 from $V_\delta = 1$ to $V_\delta \sim 10^{-6}$. As V_δ decreases, the imaginary parts of the additional frequencies increase, moving them away from the native one, eventually leading to two very distinct subsets of QN frequencies.

It might seem that it is an obvious necessity for the QN spectrum to have a mode approximately at the location of the native mode, since the evolution of Cauchy data is only slightly affected. This turns out not to be true, however, for other ways of inducing apparent completeness.[14] Cutting off the TDP potential at some very large value of x, for example, has a negligible effect on the evolution of Cauchy data, and it also produces an unbounded set of additional QN frequencies. However, the spectrum of QN frequencies turns out to contain no frequency near the location of the native mode.

C. Complete sums

We now develop the connection between the Cauchy data ψ_0 and $\dot{\psi}_0$ and the coefficients of QN oscillations (i.e., the contravariant components a^n in the case of complete QN bases). Here we simply outline how coefficients associated with QN basis functions are computed in general. In Appendix B a proof of completeness of sums with these coefficients will be given for the case of the spiked TDP, and of Cauchy data meeting certain criteria. One of the criteria will be compact support for the Cauchy data, and we will assume from the outset that ψ_0 and $\dot{\psi}_0$ vanish outside the interval $x_2 < x < x_1$.

The QN coefficients are found by starting with a Laplace transform,

$$\hat{\psi}(s,x) = \int_0^\infty e^{-st} \Psi(t,x)\, dt. \tag{20}$$

(The connection between the Laplace variable s and the Fourier ω used in Appendix A and most of the paper is through the correspondence $s \leftrightarrow i\omega$.) The transformed wave equation reads as

$$\partial \hat{\psi}(s,x)/\partial x^2 - [s^2 + V(x)]\hat{\psi}(s,x) = J(s,x), \tag{21}$$

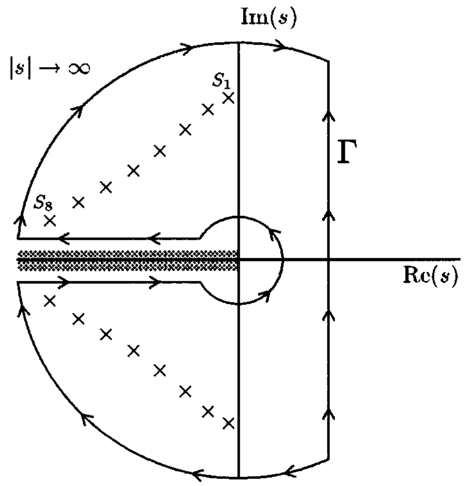

FIG. 7. Closing the contour for integration in the complex s plane. The curve Γ is the original contour of integration. It can be closed by the addition of an arc at infinity. If there is a branch point at the origin, as shown, then a branch cut can be drawn from $s=0$ to $s=-\infty$, and the contour shown can be drawn.

in which the source $J(s,x)$ is determined by the Cauchy data:

$$J(s,x)=-\dot{\psi}_0(x)-s\psi_0(x). \tag{22}$$

A solution can be found in the form

$$\hat{\psi}(s,x)=\int_{x_2}^{x_1}G(s,x,x')J(s,x')dx'. \tag{23}$$

Here the Green's function can be constructed from the homogeneous solutions $f_-(s,x)$ and $f_-(s,x)$ of the wave equation (21):

$$G(s,x,x')=\frac{1}{W(s)}\begin{cases} f_-(s,x')f_+(s,x) & (x'<x),\\ f_-(s,x)f_+(s,x') & (x'>x),\end{cases} \tag{24}$$

where $W(s)$, the Wronskian of f_- and f_+, is independent of x and x'.

We will assume that the potential falls off quickly enough at $x\to\pm\infty$ so that homogeneous solutions f_- and f_+ can be found with the property

$$f_-(s,x)\sim e^{sx}(1+\mathcal{O}(1/|x|)),\ \text{as}\ x\to-\infty \quad \text{and} \quad f_+(s,x)\sim e^{-sx}(1+\mathcal{O}(1/|x|)),\ \text{as}\ x\to\infty. \tag{25}$$

This condition is satisfied for black hole potentials and for the TDP.

Once we have found a solution of the Laplace transformed wave equation, we can reconstruct a solution of the time-dependent wave equation by applying the inverse Laplace transform:

$$\Psi(t,x)=\frac{1}{2\pi i}\int_\Gamma e^{st}\hat{\psi}(s,x)ds, \tag{26}$$

where Γ denotes the path of integration, which lies parallel to and just to the right of the imaginary axis.

Figure 7 illustrates features in the complex s plane. There may be poles and essential singularities, drawn as $\times$'s, and there may be a branch point, like that shown at $s=0$ in the figure. The contour Γ can be closed with the addition of a single arc to the left if $\hat{\psi}(s,x)$ is a single valued function of s to the left of Γ. If there are branch points then a path like that shown in Fig. 7 must be used. The Green's function for the Regge–Wheeler potential has a branch point at $s=0$, while the TDP and the STDP do not.

The integral for $\hat{\psi}$ may then be evaluated as

$$\Psi(t,x)=\sum \operatorname{Res}(e^{st}\hat{\psi}(s,x),s_j)+\frac{1}{2\pi i}\int_{\mathcal{C}} e^{st}\hat{\psi}(s,x)ds, \tag{27}$$

where the first term is the sum of residues at the singularities inside the completed contour. The second term represents contributions to the integral from arcs at ∞, and along branch cuts. In the case of the spiked TDP there is no branch cut, and the arc at ∞ makes no contributions, so we are left only with the first term, a sum of oscillations at discrete frequencies s_j that correspond to singularities of $\hat{\psi}(s,x)$. From (23) and (24) we see that singularities in $\hat{\psi}(s,x)$ must either be singularities of the homogeneous solutions f_+, f_- or zeros of the Wronskian. For any finite s we can find homogeneous solutions f_+ and f_- of (21); therefore, singularities in the Green's function can occur only at zeros of $W(s)$, and the residues in (27) must be taken at these zeros. But the vanishing of $W(s)$ at s_j means that $f_+ \propto f_-$ at s_j, and hence that s_j is a QN frequency.[8]

If, as in the case of the sTDP, the only contributions to (27) occur in the sum, then we are left with

$$\Psi(t,x)=\sum \operatorname{Res}(e^{st}\hat{\psi}(s,x),s_j), \tag{28}$$

where the sum is over QN frequencies.

We now assume that the zeros of $W(s)$ are all first order, so that

$$W(s)=\left.\frac{dW}{ds}\right|_{s=s_j}(s-s_j)+\mathcal{O}[(s-s_j)^2]. \tag{29}$$

For $x>x_1$ (i.e., for x to the right of the region in which the Cauchy data have support) we have, from (24), that

$$G(s,x,x')=\frac{f_-(s_j,x')f_+(s_j,x)}{dW/ds|_{s_j}(s-s_j)}+\mathcal{O}(s-s_j)^0, \tag{30}$$

and hence, from (23), the residues of $\hat{\psi}$ at s_j is

$$\operatorname{Res}\hat{\psi}(s,x)=\frac{e^{s_j t}f_+(s_j,x)}{dW/ds|_{s_j}}\int_{x_2}^{x_1} J(s_j,x')f_-(s_j,x')dx'. \tag{31}$$

Finally, the sum over the QN basis functions can be written as

$$\Psi(t,x)=\sum b^j u_j(t,x), \tag{32}$$

where

$$u_j(t,x)=f_+(s_j,x)e^{s_j t} \tag{33}$$

and

$$b^j=-\frac{1}{dW/ds|_{(s_j)}}\int (\dot{\psi}_0(x')+s_j\psi_0(x'))f_-(s_j,x')dx'. \tag{34}$$

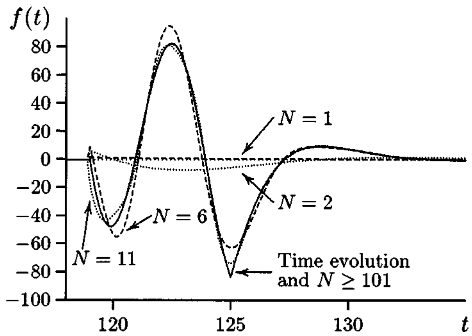

FIG. 8. Values of the mode sum for a different number of terms N, compared to the waveform resulting from integrating the time-dependent wave equation ($V_\delta = 1$).

The waveform at $x_{\text{obs}} > x_1$ is then given by (5) with

$$i\omega_j = s_j, \quad a^j = b^j f_+(s_j, x_{\text{obs}}). \tag{35}$$

This prescription has been applied to the spiked TDP, and the sine wave Cauchy data, with our standard choice of parameter values, $x_0 = 1$, $x_\delta = 10$. Notice that our choice $x_2 = -5$ in (19) satisfies the criterion in the convergence proof that the continuous, but nonsmooth, Cauchy data have support only for $x > x_0 - 2/3(x_\delta - x_0) = -5$. For this potential and these Cauchy data, the functions $\psi_0(x)$, $\dot{\psi}_0(x)$ and $f_-(s_j, x)$ are trigonometric or exponential functions, so the integral in (34) is elementary. Once the QN frequencies, ω_j, and the factors $dW/ds|_{\omega_j}$ have been computed (as described in Appendix A), the coefficients are easily evaluated. The figures below show the result of using these coefficients in sums of the form (5).

Figure 8 shows the computed result for time evolution of the Cauchy data in the case that $V_\delta = 1$. This is compared along with mode sums for an increasing number of terms, up to $N = 10\,001$ terms. Figure 9 shows the same plots in the case that $V_\delta = 10^{-6}$. In order to avoid cluttering the picture, we do not plot the values of the mode sums when they are far from having converged. These figures illustrate the fact that the mode sum converge faster for a larger δ-function. Also, for the smaller δ-function, the mode sum converges more rapidly at later times, with convergence ''sweeping down'' from late to early times.

For a better view of the differences between the mode sums and the time evolved wave functions, Fig. 10 shows the logarithmic difference (evolved waveform versus mode sum) for $V_\delta = 10^{-6}$.

Some of the systematics shown in these figures can be heuristically understood. The convergence in the $V_\delta = 1$ case is similar to that of a Fourier series (equally rapid at different times) since

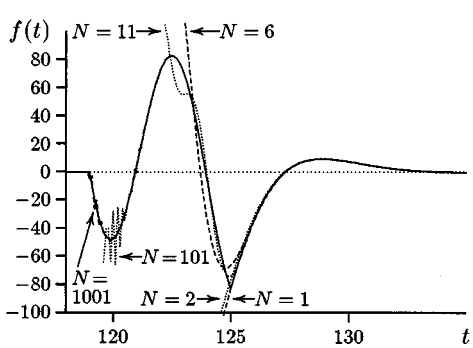

FIG. 9. Values of the mode sum for a different number of terms N, compared to the waveform resulting from integrating the time-dependent wave equation ($V_\delta = 10^{-6}$).

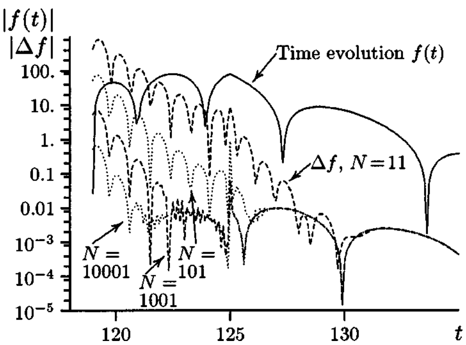

FIG. 10. Logarithmic difference Δf between the values of the mode sum for a different number of terms and the waveform resulting from integrating the time-dependent wave equation ($V_\delta = 10^{-6}$).

the QN frequencies, in this case, with small imaginary parts, are similar to the real frequencies of a Fourier series. On the other hand, for $V_\delta = 10^{-6}$, the damping of the additional (i.e., non-native) modes is much stronger and increases faster with N. Any error that originates from cutting off the mode sum after a finite number of terms can be regarded as being composed of modes with very strong damping; these modes are very large at early time.

We point out here an interesting technical feature of these results. The mode sums are so accurate at later times that the differences shown in Fig. 10 for $N = 10\,001$, is actually dominated, after $t \approx 122$, by the numerical truncation error in computing the evolved waveform. A smaller time step in numerical evolution can improve the numerical accuracy of the computed waveform, and can move to a slightly larger time the point at which the evolution versus sum difference is dominated by the numerical errors in the evolution.

IV. QUANTIFYING EXCITATION

A. The excitation coefficient

We now return to the question of how to quantify the excitation of QN oscillation. In Sec. II we defined the excitation coefficient A_k for a QN oscillation in a complete QN system. Due to the time shift problem, we argued that the excitation coefficient seems the only plausible indicator of the QN content of a waveform. In Sec. III we have seen that, at least in the particular example of the TDP, completeness can be induced in an incomplete system to create one that is ''physically equivalent,'' i.e., differs negligibly in the evolution of Cauchy data. We can now ask whether, at least for the model problem at hand, we can use the excitation coefficients of the completed system to quantify the QN oscillation in the original (''native'') system.

It is worth emphasizing that the TDP is a particularly simple starting point for these considerations, since it has only a single conjugate pair of native QN modes. The sine wave data we have used in the previous sections is also convenient since it produces a waveform (see Fig. 4) which clearly contains QN ringing, but contains significant oscillation at a different frequency ($\omega = 2\pi/6$).

The results of computations with this model are presented in Table I. Results are shown for the spiked TDP for different values of V_δ and for the standard Cauchy data, a right moving sine wave initially extending over the interval $[-5,1]$. For comparison, a shifted sine wave, initially at $[-8,-2]$, was also computed in the case $V_\delta = 10^{-6}$. Also included are results for the smooth TDP, and for the Zerilli equation with the Regge–Wheeler potential with initial data corresponding the the Close Limit technique for black hole collisions.[15]

The table presents values of excitation coefficients $A_1 \equiv a^1(a_1)^* + (a^1)^*(a_1)$ and the energy excitation coefficients $E_1 \equiv (\omega_1^*)^2 (a_1)^* a^1 + (\omega_1)^2 (\tilde{a_1})^* \tilde{a}^1$, introduced in (13) and (15). In each case the QN frequency ω_1 is taken to be that eigenfrequency which corresponds to the native mode in the limit $V_\delta \to 0$; that is, ω_1 always lies on the path in the complex plane shown in Fig.

TABLE I. Excitation coefficients and norms for the spiked TDP QN system and for black holes in the close limit approximation.

$V(x)$	V_δ	Initial data	A_1	E_1	$\|f\|^2$	$\|\dot{f}\|^2$
STDP	10^{-3}	Sine $[-5,1]$	$-40\,854\pm1$	$-23\,810\pm1$	$19\,246\pm7$	$27\,404\pm12$
STDP	10^{-6}	Sine $[-5,1]$	$-57\,933\pm1$	$5\,804.7\pm0.1$	$19\,246\pm7$	$27\,404\pm12$
STDP	10^{-6}	Sine $[-8,-2]$	$-57\,934\pm1$	$5\,804.5\pm0.3$	$19\,246\pm7$	$27\,404\pm12$
STDP	10^{-7}	Sine $[-5,1]$	$-57\,879\pm1$	$5\,811.9\pm0.1$		
STDP	10^{-8}	Sine $[-5,1]$	$-57\,873\pm1$	$5\,812.6\pm0.1$		
STDP	10^{-9}	Sine $[-5,1]$	$-57\,873\pm1$	$5\,812.6\pm0.1$		
TDP	–	Sine $[-5,1]$	$-57\,872.8\ldots$	$5\,812.64\ldots$	$19\,248.0\ldots$	$27\,415.5\ldots$
Zerilli	–	Close Limit	10.699×10^{-4}		5.566×10^{-4}	

6. For small values of V_δ this QN frequency is close to the native QN frequency $(\pm1+i)/2$. The waveform norm and the total energy are computed from waveforms obtained by explicit numerical integration of the wave equation (1). Also included are estimates for the numerical uncertainties of the results.

The norm and total energy can be computed from Eq. (9), or directly from the waveform. Similarly, the covariant coefficient may be computed from the contravariant ones, using Eq. (12), or from the waveform via Eq. (7). Even the contravariant coefficients can be obtained from the waveform itself, using an asymptotic fit as it approaches a pure QN oscillation at the least damped QN frequency, instead of the residues of the Green's function through Eqs. (34) and (35). The waveform itself, in turn, can be obtained by direct numerical integration of the wave equation, or by using the mode sum as in Eq. (5). We have checked these alternatives for obtaining the quantities listed in the table; they result in essentially the same values as the route we have taken here.

We first notice that the values for the excitation coefficients are identical, within the numerical errors, for the original initial data (sine wave over the interval $[-5,1]$) and the shifted initial data (over the interval $[-8,-2]$). This confirms our earlier argument that the excitation coefficients we have defined are independent of a translation of the initial data or, correspondingly, to a time shift of the evolved waveform.

It is important to realize that there are different ways to compute the quantities in the table, and several of them do not depend on the complete set of modes. The waveform and energy norm, of course, require only the waveform, but all coefficients can also be computed entirely from the waveform itself. The least damped QN frequency can be inferred from the asymptotic late time behavior, as can the contravariant coefficient for the least damped mode. [See (5).] Of the cases we study the least damped mode always corresponds to the native mode except for the $V_\delta=1$ model. (See Fig. 5.) Once the QN frequency is known the covariant coefficient can be computed directly from the integral in (7) and, with the contravariant coefficient known, the excitation and energy coefficients can then be found from (13) and (15). Since the quantities in the table can be computed from waveforms, we can compute them for the smooth TDP. In this case we do not have a complete set of modes, but that is irrelevant to the procedure for computation. (It turns out, in fact, that the simplicity of the smooth TDP and the sine wave data allows a closed form solution for the waveform, and for the norms and coefficients. This closed form solution has been used, and the values for the smooth TDP can be found for the table to an essentially arbitrary precision.)

The numbers in the table make it clear that the results for the smooth TDP are the $V_\delta\to0$ limit of the spiked TDP. This is obvious from a computational point of view, since all results can be computed from the waveform, and the $V_\delta\to0$ waveform approaches the smooth TDP waveform. From another point of view, however, this result is interesting and important. It means that we can compute the excitation coefficient independent of the method with which completeness is induced. Put another way, it means that we can compute the excitation coefficient for a small completeness-inducing change, independent of the nature of the change. This conclusion, in fact, is crucial to the possible importance of the excitation coefficient. We could not use the excitation coefficient to characterize an excitation of a physical system if the value of the coefficient depended on our choice of a modification of that physical system.

The table also gives a value for the excitation coefficient for the gravitational radiation produced, in the close limit approximation,[14] by the head on collision of two equal mass nonrotating holes. In this case the QN spectrum is infinite, but the QN oscillations are not complete. The excitation coefficient given is for the least damped of the QN oscillations, a frequency that appears to dominate the waveform produced.

Up to this point we have noted that computations with the model problem illustrate and confirm the features of the excitation coefficient that made it an attractive candidate for the quantification of the excitation of QN oscillations. Table I, however, also makes it unlikely that the excitation is a *useful* index of QN oscillation. For the small-δ spiked TDP the excitation coefficient is negative, and is larger (by a factor ~ 3) in magnitude than the norm. Note that we cannot ignore this ''wrong'' sign, and simply keep the large magnitude as an indication of a strong QN presence. Due to relation (9) we must conclude that the sum of all the other QN oscillations (those unrelated to the native QN mode) must be greater (by a factor ~ 4) than the norm. For the close limit waveform the excitation coefficient is roughly twice the size of the norm of f and hence the sum of excitation coefficients of all other QN oscillations (if the system were made complete somehow) would be negative.

As a possible alternative to the excitation coefficient we have also computed the energy excitation coefficient, as defined in (15). The results listed for E_1 and the norm of $\dot{f}$, however, do not make this any more attractive as a measure of QN excitation.

At this point, we also note that the excitation coefficient A_1 is not related to the norm of the QN mode contribution corresponding to the first pair of modes, i.e., of $b^1 u_1 + (b^1 u_1)^*$. Similarly, the energy excitation coefficient E_1 is no measure for the energy of this QN mode contribution.

It is, of course, impossible to prove that something like quantifying QN excitations cannot be done in a mathematically acceptable way. Despite this, and perhaps to provoke further work (by others!) we are tempted to offer the following very tentative conclusion: Consider the following three conditions, which allow a mathematically meaningful as well as a useful measure of excitation; they are all satisfied for normal mode systems.

- The measure of excitation is independent of a simple shift of the wave form (i.e., a shift in space of the initial data, corresponding to a shift in time for the time evolved waveform).
- Excitation strength can be quantified individually for any number of modes, with the individual excitations adding up to the total norm of the waveform.
- The measure of excitation is useful for quantifying the excitation in a comparative way; in particular, it always lies between 0% and 100%.

We conjecture: There is no quantification of QN oscillations of the waveform, based on the algebra of the function space of QN oscillations, which satisfies all three of these criteria.

The excitation coefficients we are defining in this paper satisfy the first two conditions, but not the third one. (In Table I we see that the excitation coefficient can be negative and can be greater than 100%.) One might regard these two conditions as related to mathematical properties of the QN mode system, while the third one is of a more practical significance. We are currently investigating another technique to quantify excitations, with a measure that satisfies the first and third criteria, but not the second one. This measure may turn out to be of some practical utility, but is not as closely related to the mathematics of the function space as are the present considerations. A description of this work will be published elsewhere.

Our conjecture is based specifically on the appearance of the excitation coefficient as the only quantity in that function space that solves the ''time shift'' problem, and on the observed failure of the excitation coefficient to be ''useful.'' It is also based, less specifically, on our belief that there are differences between normal mode and QN systems that cannot be bridged. For this

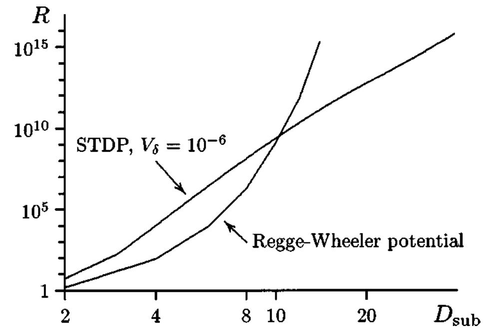

FIG. 11. The condition number R for the metric matrix for a truncated subspace of dimension $D_{\rm sub}$, spanned by the first $D_{\rm sub}$ QN basis functions. Results are shown for both the spiked TDP ($V_\delta = 10^{-6}$) and the Regge–Wheeler potential.

reason we speculate that inducing completeness, while it is mathematically interesting, is not likely to lead to useful tools for understanding the underlying native system. The reason is that the spectrum of added QN frequencies is unrelated to the native system and characterizes only the method used to induce the completeness. We, in fact, are willing to extrapolate in the direction of this speculation. We conjecture that in a complete QN spectrum which has a dense set of frequencies, and a small number of ''isolated'' frequencies, like the spectrum of the spiked TDPs in Fig. 5, it is useful to modify the system to *remove* completeness, in order to get a more useful understanding and a simplified method of analysis. The obvious example of this is removing the δ functions from the spiked TDP problem in order to get a physically equivalent, but simpler incomplete system.

A very similar point of view is that in a QN spectrum, not all frequencies are equally important. Some will actually be evident in waveforms produced in the evolution of generic Cauchy data; others will not. For the small-δ spectra of Fig. 5, the ''interesting'' QN frequencies are the isolated ones near $(1+i)/2$. In the $V_\delta = 1$ case, on the other hand, there is again the appearance of QN ringing in the waveform, but the spectrum contains no isolated frequency. A similar situation was found in a study where the Regge–Wheeler potential was replaced by a series of step potentials.[14] More generally, one should ask the following: If one has only the spectrum and the associated quantities (e.g., the metric matrix), is there a way of identifying which QN frequencies are ''important'' in the sense of really characterizing the evolution of Cauchy data? In this sense we are asking a question related to ''to what extent are (some) QN modes like normal modes?'' since normal modes *do* characterize the system in which they arise.

B. Condition number of the metric matrix

The metric matrix (11) would appear to be a likely place to find a way of characterizing a QN system without regard to specific waveforms produced by specific Cauchy data. We will now discuss some numerical results which confirm the intuitive insight we had gained by doing thought experiments on specific Cauchy data in Sec. II D. Studying the metric matrix directly, we do not need to refer to specific initial data any more, as we had to do before.

We first note the singular nature of the infinite metric matrix in two cases. To characterize this singular property we truncate the set of QN functions, keeping only the first D. We then compute the condition number R (ratio of maximum to minimum eigenvalue) for the D dimensional subspace. The condition number as a function of $D_{\rm sub}$ is plotted in Fig. 11 for two cases: the spiked TDP and the QN spectrum of Schwarzschild black holes (the QN modes of the Zerilli or Regge–Wheeler potentials). For the spiked TDP the approximately straight line in the log–log plot suggests that the condition number increases roughly as the twelfth power of the dimension $D_{\rm sub}$ of the subspace. For the black hole QN spectrum the increase is even more dramatic, suggesting perhaps a ''more singular'' metric for this incomplete QN spectrum.

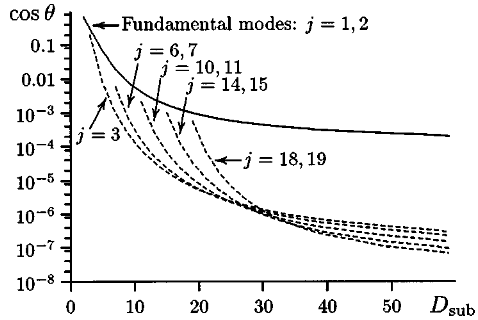

FIG. 12. Cosine of angles between several contravariant and covariant basis vectors, as a function of the dimension of a subspace spanned by $D_{\rm sub}$ QN modes of the spiked TDP with $V_\delta = 10^{-6}$.

C. Angles between basis functions

In Sec. II we introduced covariant basis functions $\phi^m(t)$ with the definition $\phi^m(t)\cdot e^{i\omega_n t} = \delta_{mn}$. If the basis functions $e^{i\omega_n t}$ were orthogonal we would have that $\phi^n(t)$ and $e^{i\omega_n t}$ are ''parallel,'' that is, $\phi^n(t) \propto e^{i\omega_n t}$. It is plausible that such statements as ''this wave is dominated by oscillation at the fundamental QN frequency'' are most meaningful if the covariant and contravariant basis vectors are nearly ''parallel.'' To measure the extent to which $\phi^j(t)$ and $e^{i\omega_j t}$ fail to be proportional we can introduce an angle α_j between them, defined by

$$\cos(\alpha_j) = \frac{\phi^j(t)\cdot e^{i\omega_j t}}{\|\phi^j(t)\|\ \|e^{i\omega_j t}\|} = \frac{1}{\sqrt{G_{jj}G^{jj}}}, \tag{36}$$

where G^{ij} is the matrix inverse of G_{ij}. The components of the infinite matrix G^{ij} cannot be computed, so again we truncate a subspace by keeping only the first $D_{\rm sub}$ vectors, and we compute the angles in that subspace with (36). Results are shown in Fig. 12 for the spiked TDP with $V_\delta = 10^{-6}$. The value of $\cos(\alpha_j)$ is shown for several QN modes as a function of $D_{\rm sub}$, the dimension at which the subspace is truncated. Figure 13 shows $\cos(\alpha_j)$ for the black hole QN spectrum. In Fig. 12, the decrease of $\cos(\alpha_1)$, corresponding to the native mode, is much slower than that of the additional modes. One might speculate that this is related to the fact that the native mode is more characteristic of the system than the additional ones. No such clear distinction can be seen in Fig. 13 for the case of the Regge–Wheeler potential. However, the angles increase much faster for the more highly damped modes as well, which might again indicate that the fundamental, least damped mode is more characteristic of the system than the more strongly damped ones.

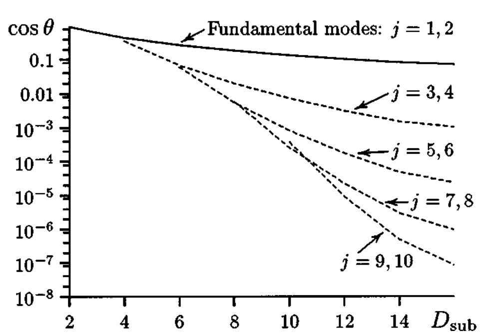

FIG. 13. Cosine of angles between several contravariant and covariant basis vectors, as a function of the dimension of a subspace spanned by $D_{\rm sub}$ QN modes of the Regge–Wheeler potential.

ACKNOWLEDGMENTS

We wish to thank Ted Newman for useful discussions and numerous helpful suggestions on the presentations of these results. We also thank Karel Kuchař, Wai-Mo Suen, and John Whelan for useful discussions during the development of this work. We thank the Deutsche Forschungsgemeinschaft for support through SFB 382 and NATO for support through Grant No. CRG.971092. One of us (R.H.P.) thanks the National Science Foundation for support under Grants No. PHY95-07719 and No. PHY97-34871.

APPENDIX A: FINDING QUASINORMAL FREQUENCIES OF THE TDP AND OF THE SPIKED TDP

1. The unmodified TDP

The TDP is defined as

$$V(x)=\begin{cases}0, & x<x_0,\\ l(l+1)/x^2, & x\geqslant x_0,\end{cases} \tag{A1}$$

where l is an integer. We look only at $l=1$, but the procedure can easily be extended for larger values of l. Also, we will generally let $x_0=1$.

The domain of the wave equation is naturally divided into two regions.

I. $x<x_0$. In this region, the potential vanishes, and therefore the wave equation has the two trivial solutions:

$$\psi_{\mathrm{I}1}(x)=e^{+i\omega x}, \tag{A2}$$

$$\psi_{\mathrm{I}2}(x)=e^{-i\omega x}. \tag{A3}$$

II. $x\geqslant x_0$. For integer values of l, the solutions are given by finite sums. For $l=1$, we have

$$\psi_{\mathrm{II}1}(x)=e^{+i\omega x}\left(1-\frac{1}{i\omega x}\right), \tag{A4}$$

$$\psi_{\mathrm{II}2}(x)=e^{-i\omega x}\left(1+\frac{1}{i\omega x}\right). \tag{A5}$$

Obviously, the solutions satisfying the required boundary conditions at negative and positive infinity are

$$\psi_-(x)=\psi_{\mathrm{I}1}(x)\quad (x<x_0), \tag{A6}$$

$$\psi_+(x)=\psi_{\mathrm{II}2}(x)\quad (x\geqslant x_0). \tag{A7}$$

In general, of course, $\psi_-(x)$ will be a linear combination of $\psi_{\mathrm{II}1}(x)$ and $\psi_{\mathrm{II}2}(x)$ for $x\geqslant x_0$, and $\psi_+(x)$ a combination of $\psi_{\mathrm{I}1}(x)$ and $\psi_{\mathrm{I}2}(x)$ for $x<x_0$. A quasinormal mode is a solution where both boundary conditions are satisfied simultaneously, i.e., $\psi_-(x)=\psi_+(x)$. The easiest way to find out if this is the case is to compare $\psi_-(x)$ and $\psi_+(x)$, as defined in (A6), at $x=x_0$. Strictly speaking, $\psi_-(x)$ is not defined at $x=x_0$. However, any solution of the wave equation will have to be continuous, and have a continuous first derivative, at $x=x_0$. It is therefore permissible to use the left limit of $\psi_-(x)$ and of $\psi'_-(x)$ as $x\to x_{0^-}$, and compare them with the values of $\psi_+(x)$ and $\psi'_+(x)$ at $x=x_0$.

The comparison is done using the Wronskian determinant of $\psi_-(x)$ and $\psi_+(x)$:

$$W[\psi_-,\psi_+](x_0)=\psi_-(x_0)\psi'_+(x_0)-\psi'_-(x_0)\psi_+(x_0)$$
$$=-2i\omega-\frac{2}{x_0}-\frac{1}{i\omega x_0^2}. \tag{A8}$$

Solving the equation $W[\psi_-,\psi_+](x_0)=0$ for ω yields the quasinormal frequencies

$$\omega=\frac{\pm 1+i}{2x_0}. \tag{A9}$$

Therefore, for $l=1$, there is only one pair of quasinormal frequencies.

2. QN frequencies of the spiked TDP

We define the ''spiked'' potential as

$$\overline{V}(x)=V(x)+V_\delta\delta(x-x_\delta). \tag{A10}$$

We now have to distinguish three areas.

I. $x<x_0$. This region is identical to region I for the unmodified potential, with the same set of two solutions.

IIa. $x_0\leqslant x\leqslant x_\delta$. Again, there are two linearly independent solutions:

$$\psi_{\mathrm{IIa1}}(x)=e^{+i\omega x}\left(1-\frac{1}{i\omega x}\right), \tag{A11}$$

$$\psi_{\mathrm{IIa2}}(x)=e^{-i\omega x}\left(1+\frac{1}{i\omega x}\right). \tag{A12}$$

IIb. $x>x_\delta$. The two linearly independent solutions are

$$\psi_{\mathrm{IIb1}}(x)=e^{+i\omega x}\left(1-\frac{1}{i\omega x}\right), \tag{A13}$$

$$\psi_{\mathrm{IIb2}}(x)=e^{-i\omega x}\left(1+\frac{1}{i\omega x}\right). \tag{A14}$$

In our notation the function $\psi_{\mathrm{IIb2}}(x)$, for example, refers to the solution in all regions that, in region IIb, has the functional form shown in Eq. (A14).

Due to the δ-function separating regions IIa and IIb, the solutions $\psi_{\mathrm{IIa2}}(x)$ and $\psi_{\mathrm{IIb2}}(x)$ are not the same. Rather, $\psi_{\mathrm{IIb2}}(x)$ will be a linear combination of $\psi_{\mathrm{IIa1}}(x)$ and $\psi_{\mathrm{IIa2}}(x)$ in region IIa:

$$\psi_{\mathrm{IIb2}}(x)=p_1\psi_{\mathrm{IIa1}}(x)+p_2\psi_{\mathrm{IIa2}}(x). \tag{A15}$$

Once again, the solutions

$$\psi_-(x)=\psi_{\mathrm{I1}}(x)\quad (x<x_0), \tag{A16}$$

$$\psi_+(x)=\psi_{\mathrm{IIb2}}(x)\quad (x\geqslant x_\delta), \tag{A17}$$

satisfy the boundary conditions at negative and positive infinity. However, in order to compare solutions at $x=x_0$, we now have to determine the representation of $\psi_{\mathrm{IIb2}}(x)$ in region IIa, i.e., we need to know the coefficients for the linear combination (A15).

These coefficients can be determined using the junction conditions for any solution $\psi(x)$ of the wave equation across the δ-function.

(1) $\psi(x)$ must be continuous at x_δ.
(2) The derivative $\psi'(x)$ must have a discontinuity given by

$$\psi'(x_\delta^+)-\psi'(x_\delta^-)=V_\delta\psi(x_\delta). \tag{A18}$$

The second condition is obtained by integrating the wave equation from $x=x_\delta-\epsilon$ to $x=x_\delta+\epsilon$, letting $\epsilon\rightarrow 0$, and using the first condition.

Using ψ_{IIb2} for $\psi'(x_\delta^+)$ and Eq. (A15) for $\psi'(x_\delta^-)$, we can solve these conditions for p_1 and p_2. We obtain

$$p_1=V_\delta\frac{[\psi_{\mathrm{IIa2}}(x_\delta)]^2}{W_{+12}}, \tag{A19}$$

$$p_2=1-V_\delta\frac{\psi_{\mathrm{IIa1}}(x_\delta)\,\psi_{\mathrm{IIa2}}(x_\delta)}{W_{+12}}, \tag{A20}$$

where $W_{+12}=W[\psi_{\mathrm{IIa1}},\psi_{\mathrm{IIa2}}]=-2i\omega$.

Therefore,

$$\begin{aligned}W[\psi_-,\psi_+](x_0)=W[\psi_{\mathrm{I1}},\psi_{\mathrm{IIb2}}](x_0)&=p_1W[\psi_{\mathrm{I1}},\psi_{\mathrm{IIa1}}](x_0)+p_2W[\psi_{\mathrm{I1}},\psi_{\mathrm{IIa2}}](x_0)\\&=R_1+V_\delta(R_2+R_3e^{-2i\omega L}),\end{aligned} \tag{A21}$$

where $L=x_\delta-x_0$, and $W[\psi_{\mathrm{I1}},\psi_{\mathrm{IIa1}}](x_0)=e^{2i\omega x_0}/(i\omega x_0^2)$, and thus

$$R_1=W[\psi_{\mathrm{I1}},\psi_{\mathrm{IIa2}}](x_0)=-2i\omega-\frac{2}{x_0}-\frac{1}{i\omega x_0^2}, \tag{A22}$$

$$R_2=-\frac{\psi_{\mathrm{IIa1}}(x_\delta)\,\psi_{\mathrm{IIa2}}(x_\delta)}{W_{+12}}W[\psi_{\mathrm{I1}},\psi_{\mathrm{IIa2}}](x_0)=-\left(1-\frac{1}{(i\omega x_\delta)^2}\right)\left(1+\frac{1}{i\omega x_0}+\frac{1}{2(i\omega x_0)^2}\right), \tag{A23}$$

$$R_3=e^{2i\omega L}\frac{\psi_{\mathrm{IIa2}}(x_\delta)^2}{W_{+12}}W[\psi_{\mathrm{I1}},\psi_{\mathrm{IIa1}}](x_0)=-\frac{1}{2(i\omega x_0)^2}\left(1+\frac{1}{i\omega x_\delta}\right)^2. \tag{A24}$$

The quasinormal frequencies of the spiked TDP can now be computed numerically by searching for roots of the equation

$$W[\psi_-,\psi_+](x_0)=0. \tag{A25}$$

3. Asymptotic approximation for the QN frequencies of the spiked TDP

It is possible to find an asymptotic formula for the QN frequencies under the assumption that the absolute value of the frequency becomes large. We start by assuming that in (A25), we have $|\omega x_0|\gg 1$ and $|\omega x_\delta|\gg 1$. The condition for QN frequencies can then be written as

$$2i\omega x_0+\mathcal{O}([\omega x)]^0)+V_\delta\left[1+\mathcal{O}([\omega x)]^{-1})+\frac{e^{-2i\omega L}}{2(i\omega x_0)^2}(1+\mathcal{O}([\omega x)]^{-1}))\right]=0, \tag{A26}$$

where x is the minimum of x_0 and x_δ. It is clear that for $|\omega x|\gg 1$ there can be solutions only if

$$2i\omega+V_\delta\frac{1}{2(i\omega x_0)^2}\,e^{-2i\omega L}\approx 0, \tag{A27}$$

and we use this approximation to find an iterative solution for the QN frequencies. We start by taking the cube root of

$$e^{-2i\omega L} = -\frac{4(i\omega)^3 x_0^2}{V_\delta} \tag{A28}$$

to write

$$e^{(-2/3)i\omega_R L} e^{(2/3)\omega_I L} e^{i\Delta} = -\left(\frac{4x_0^2}{V_\delta}\right)^{(1/3)} i\omega \equiv -Ai\omega, \tag{A29}$$

where $\Delta = j\frac{2}{3}\pi,\ j=0,1,2$.

Taking the absolute values on both sides gives

$$e^{(2/3)\omega_I L} = A|\omega|. \tag{A30}$$

With $\omega \equiv \omega_R + i\omega_I$ this last relation already shows that $\omega_I \ll |\omega| \approx \omega_R$ is required for a QN frequency.

Using (A30) to rewrite (A29) we find

$$\cos\left(\frac{2}{3}L\omega_R + \Delta\right) - i\,\sin\left(\frac{2}{3}L\omega_R + \Delta\right) = -i\frac{\omega}{|\omega|}. \tag{A31}$$

We now make the approximation $\omega_I \approx 0$, i.e., $\omega \approx \omega_R \approx |\omega|$, leading to

$$\cos\left(\frac{2}{3}L\omega_R + \Delta\right) = 0, \quad \sin\left(\frac{2}{3}L\omega_R + \Delta\right) = 1, \tag{A32}$$

$$\frac{2}{3}L\omega_R = \frac{\pi}{2} + 2N\pi - \frac{2}{3}j\pi, \quad N=0,1,\ldots;\ j=0,1,2, \tag{A33}$$

or, equivalently,

$$(\omega_R)_n = \frac{1}{L}\left(\frac{3\pi}{4} + n\pi\right), \quad n=0,1,\ldots. \tag{A34}$$

An approximation for ω_I is then obtained by using (A30):

$$(\omega_I)_n = \frac{3}{2}\frac{1}{L}(\ln A + \ln\,\omega_R). \tag{A35}$$

The approximate solutions of (A34) and (A35) can now be iteratively improved. We rewrite (A25) as

$$e^{-2i\omega L} = -\frac{R_1 + V_\delta R_2}{V_\delta R_3} =: R(\omega) = R(\omega_R, \omega_I), \tag{A36}$$

and take absolute values to get

$$e^{2\omega_I L} = |R|, \tag{A37}$$

$$\omega_I = \frac{\ln|R|}{2L}. \tag{A38}$$

With these inserted in (A36) we arrive at

$$\cos(2\omega_R L) = -\frac{\Re(R)}{|R|}, \quad \sin(2\omega_R L) = \frac{\Im(R)}{|R|}. \tag{A39}$$

This can be used as an iterative method to find the pth iteration from the $(p-1)$th approximation, as follows:

$$\cos(2(\omega_R)^p L) = -\frac{\Re(R((\omega_R)^{p-1},(\omega_I)^{p-1}))}{|R((\omega_R)^{p-1},(\omega_I)^{p-1})|}, \quad \sin(2(\omega_R)^p L) = \frac{\Im(R(\ldots))}{|R(\ldots)|}, \tag{A40}$$

and

$$(\omega_I)^p = \frac{\ln|R((\omega_R)^p,(\omega_I)^{p-1})|}{2L}. \tag{A41}$$

This iterative solution can be started with *any* value of n in the zeroth approximation of (A34) and (A35), and the iteration converges to the exact solution. The iteration cannot be used to find the native QN frequency of the smooth TDP since $V_\delta = 0$ in that case.

APPENDIX B: PROOF OF CONVERGENCE

A proof is given here of the convergence of the sum of quasinormal excitations for the spiked TDP under appropriate restrictions on the Cauchy data. To do this we start by defining the integral

$$\mathcal{I}(d_1) \equiv \int F(s)ds \equiv \frac{1}{2\pi i}\int \frac{s^2\phi(s)e^{s(d_1+d_2)}}{\Delta(s)}ds. \tag{B1}$$

Here $\Delta(s)$ is defined to be

$$\Delta(s) \equiv P_1(s) + e^{2sL}s^3 P_2(s), \tag{B2}$$

in which P_1 and P_2 are polynomials of finite order in $1/s$:

$$P_1(s) \equiv A_1 + B_1/s + \ldots, \quad P_2(s) \equiv A_2 + B_2/s + \ldots\ . \tag{B3}$$

The path of integration in the complex s plane is along the vertical axis, from $-i\infty$ to $+i\infty$. For $\Re(s) \leqslant 0$ the function $\phi(s)$ is required to have the property that for some real constants K_ϕ and p,

$$|\phi(s)| < \frac{K_\phi}{|s|^p}. \tag{B4}$$

(This condition will be related below to restrictions on acceptable Cauchy data.) The constants appearing in (B1) are taken to satisfy the following conditions: (i) The ratio A_1/A_2 is real and positive. (ii) L, d_1 and d_2 are real and nonnegative, and $2L > d_2$. (iii) The constant p appearing in (B4) must be large enough that

$$p + \frac{3d_2}{2L} - 2 > 0. \tag{B5}$$

The roots of Δ are denoted s_k. (They represent, of course, the QN frequencies according to the usual correspondence $s \leftrightarrow i\omega$.) Since they must occur in complex conjugate pairs it is convenient here to use the notation $s_1, s_{-1}, s_2, s_{-2}, \ldots$, with $s_{\pm k}$ indicating the corresponding roots with positive and negative imaginary parts.

What we will prove is that under the conditions stated above $\mathcal{I}(d_1)$ can be written as

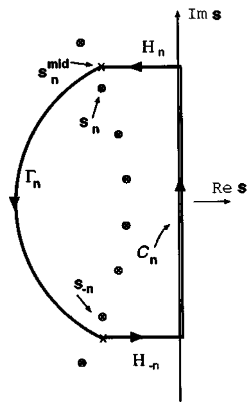

FIG. 14. Contour for proving convergence for the spiked TDP oscillations.

$$\mathcal{I}(d_1)=\sum_{k=-\infty}^{k=+\infty} a^{(k)}e^{s_k d_1}, \tag{B6}$$

where

$$a^{(k)}\equiv s_k^2\phi(s_k)e^{s_k d_2}/(d\Delta/ds)|_{s=s_k}. \tag{B7}$$

The convergence of the sum in (B6) is not uniform at $d_1\to 0$, but is uniform for any interval of d_1 bounded away from 0. We give the details of the relationship to the physical problem of Sec. III after we prove the main result above. The proof is organized with a set of lemmas.

Lemma 1: Let $s_n^{mid}=s_n+i\pi/2L$ and let H_n be the horizontal path (as shown in Fig. 14) in the s plane from the imaginary s axis to the point s_n^{mid}. Let

$$\mathcal{I}_{H_n}\equiv\int_{H_n}F(s)ds \tag{B8}$$

be the integral of $F(s)$ on this path, then N can be chosen so that, for $n>N$,

$$|\mathcal{I}_{H_n}|\leqslant \text{const}\times n^{2-p-(3d_2/2L)}. \tag{B9}$$

Proof: The discussion of the roots of the spiked TDP Wronskian in Appendix A can be applied to the roots of Δ. For large n, the n dependence of the roots takes the form

$$\Im(s_n)=n\pi/L+\text{const}+\ldots, \tag{B10}$$

$$\Re(s_n)=-(3/2L)\ln(n)+\text{const}+\ldots\ . \tag{B11}$$

It follows that a constant can be chosen so that $|s_n|>\text{const}\times n$, and hence $|\phi(s)|<2\pi K/n^p$, for some constant K independent of n. Since $\Re(s)\leqslant 0$ on H_n and d_1 is non-negative, we have that $|e^{sd_1}|\leqslant 1$. With $s=s_n+i\pi/2L+\sigma$ and with σ running from $-\Re(s_n)$ to 0 on the path H_n, the integral must satisfy

$$|\mathcal{I}_{H_n}|\leqslant\frac{K}{n^{p-2}}\left|\int_0^{-\Re(s_n)}\frac{d\sigma}{D(s)}\right|, \tag{B12}$$

where $D(s)\equiv e^{-sd_2}\Delta(s)$. On H_n we have that

$$e^{2Ls}=-e^{2Ls_n}e^{2L\sigma}, \tag{B13}$$

and $D(s)$ can be written as

$$D(s)=P_1(s_n)e^{-i\Im(s)d_2}[e^{-\Re(s)d_2}\mathcal{R}_1+e^{2L\sigma}e^{-\Re(s)d_2}\mathcal{R}_2], \tag{B14}$$

where

$$\mathcal{R}_1=\frac{P_1(s)}{P_1(s_n)}, \quad \mathcal{R}_2=\frac{s^3P_2(s)}{s_n^3P_2(s_n)}. \tag{B15}$$

For n larger than some N we can make both $\mathcal{R}_1$ and $\mathcal{R}_2$ arbitrarily close to unity. For $n>N$ it follows that the magnitude of the sum in square brackets in (B14) must be larger than the second term, and hence

$$|D(s)|>|P_1(s_n)|e^{2L\sigma}e^{-\Re(s)d_2}|\mathcal{R}_2|. \tag{B16}$$

By choosing N large enough we can make $|P_1(s_n)|$ and $|\mathcal{R}_2|$ larger than n-independent constants, so that

$$|D(s)|>\text{const}\times e^{2L\sigma}e^{-\Re(s)d_2}=\text{const}\times e^{2L\sigma}e^{-\Re(s_n)d_2}e^{-\sigma d_2}. \tag{B17}$$

We have then that

$$\begin{aligned}|\mathcal{I}_{H_n}|&\leqslant\frac{K'}{n^{p-2}}e^{\Re(s_n)d_2}\left|\int_0^{-\Re(s_n)}e^{\sigma(d_2-2L)}\,d\sigma\right|\\&<\frac{K'}{n^{p-2}}e^{\Re(s_n)d_2}\left|\frac{1}{2L-d_2}\right|.\end{aligned} \tag{B18}$$

It follows from (B10) that we can choose a constant so that $|e^{\Re(s_n)d_2}|<\text{const}/n^{(3d_2/2L)}$, and therefore that

$$|\mathcal{I}_{H_n}|\leqslant\frac{\text{const}}{n^{p+(3d_2/2L)-2}}, \tag{B19}$$

which was to be proven. A similar proof shows that the same result applies to $|\mathcal{I}_{H_{-n}}|$.

Lemma 2: On the arc Γ_n from s_n^{mid} to s_{-n}^{mid} with $|s|=|s_n^{mid}|$, the magnitude of $\Delta(s)$ is larger than some constant that is independent of n.

Proof: We write

$$\Delta(s)=P_1(s)[1+\Phi(s)], \tag{B20}$$

where

$$\Phi(s)\equiv e^{2sL}s^3P_2(s)/P_1(s). \tag{B21}$$

We must show that $|1+\Phi|$ is bounded from below by an n-independent constant. We recall that $s_n^{mid}=s_n+i\pi/2L$ and that $\Phi(s_n)=-1$, so that

$$\Phi(s_n^{mid})\equiv\left(\frac{s_n^{mid}}{s_n}\right)^3\frac{P_2(s_n^{mid})}{P_2(s_n)}\frac{P_1(s_n)}{P_1(s_n^{mid})}. \tag{B22}$$

For n larger than some $N_{\min}$ we have that $\Phi(s_n^{mid})$ is arbitrarily close to unity, so that

$$\Phi(s_n^{mid})=1+\lambda, \tag{B23}$$

and $N_{\min}$ can be chosen large enough to make $|\lambda|$ arbitrarily small.

On the arc Γ_n we write $\Phi(s)$ as

$$\Phi(s)=e^{2sL}s^3(A_2/A_1)[1+\rho(s)]. \tag{B24}$$

By choosing $N_{\min}$ sufficiently large, we can bound the magnitude of ρ to be less than an arbitrarily small n-independent constant. For s on the arc Γ_n, we now write s as

$$s=R_n ie^{i\theta}, \tag{B25}$$

where $R_n\equiv|s_n^{mid}|$ and θ is the counterclockwise angle from the positive imaginary s axis to the point s on Γ_n. We denote by θ_n the angle to s_n^{mid}, so that $s_n^{mid}\equiv R_n ie^{i\theta_n}$ and we write $\theta\equiv\theta_n+\delta\theta$. In terms of this notation we have

$$\Phi(s)=F_0F_1(\delta\theta)F_2(\delta\theta), \tag{B26}$$

where

$$F_0\equiv(1+\lambda)(1+\rho(s))/(1+\rho(s_n^{mid})), \tag{B27}$$

$$F_1\equiv e^{2iLR_n[\cos(\theta_n+\delta\theta)-\cos(\theta_n)]}e^{3i\delta\theta}, \tag{B28}$$

$$F_2\equiv e^{2LR_n[\sin(\theta_n)-\sin(\theta_n+\delta\theta)]}. \tag{B29}$$

The complex phase of $\Phi(s)$ is near zero at $s=s_n^{mid}$, and decreases as s moves counterclockwise along the arc Γ_n. We use the expressions above to find at what value $\delta\theta^*$ of $\delta\theta$ the phase of Φ first becomes $-\pi/2$. We note that $|F_1|=1$ and on the top half of the arc $F_2\leqslant 1$. The value of $\delta\theta^*$ is given by

$$2LR_n[\cos(\theta_n+\delta\theta^*)-\cos(\theta_n)]+3\delta\theta^*+\zeta(\theta_n+\delta\theta^*)=-\pi/2, \tag{B30}$$

where $\zeta(\theta)$ is the phase of F_0. We note that

$$|\cos(\theta_n+\delta\theta^*)-\cos(\theta_n)|>\sin(\theta_n)\delta\theta^*.$$

From (B10) we know that sin θ_n decreases with n as $\ln(n)/n$, and R_n increases as n. Thus, the first term in (B30) is larger than the second by a factor that increases as $\ln(n)$. The third term, the ζ term, decreases with increasing n, and we can bound it to be smaller than an arbitrarily small constant by choosing $N_{\min}$ sufficiently large. From these considerations it follows that we can choose $N_{\min}$ large enough that the magnitude of the first term in (B30) is larger than, say 2/3, of the magnitude of the left hand side, and hence $\delta\theta^*$ satisfies

$$2LR_n[\cos(\theta_n)-\cos(\theta_n+\delta\theta^*)]>\pi/3. \tag{B31}$$

Let us also take $N_{\min}$ large enough so that $\theta_n+\delta\theta^*<\pi/4$. In that case we have

$$\frac{\sin(\theta_n+\delta\theta^*)-\sin(\theta_n)}{\cos(\theta_n)-\cos(\theta_n+\delta\theta^*)}>1. \tag{B32}$$

From this it follows that $2LR_n[\sin(\theta_n+\delta\theta^*)-\sin(\theta_n)]>\pi/3$, and hence $F_2<e^{-\pi/3}$. Since the deviation of $|F_0|$ from unity is arbitrarily small, let us use $|F_0|<e^{\pi/12}$ and conclude that $|\Phi|<e^{-\pi/4}$ at the point along Γ_n at which Φ first becomes purely imaginary. As $\delta\theta$ increases, the magnitude of Φ continues to decrease. It follows that for every point along the top of the arc,

$$|1+\Phi|>1-e^{-\pi/4}. \tag{B33}$$

A similar analysis starting at s_{-n} shows that (B33) holds also for the bottom half of the arc.

Lemma 3: On the arc Γ_n from s_n^{mid} to s_{-n}^{mid} with $|s|=|s_n^{mid}|$, the integral,

$$\mathcal{I}_{\Gamma_n}(d_1)\equiv\frac{1}{2\pi i}\int\frac{s^2\phi(s)e^{s(d_1+d_2)}}{\Delta(s)}ds, \tag{B34}$$

satisfies

$$|\mathcal{I}_{\Gamma_n}(d_1)|<\frac{\text{const}}{n^{p+(3d_2/L)-2}}, \tag{B35}$$

where the constant is independent of n.

Proof: On Γ_n we have that $\Re(s)\leqslant\Re(s_n)$, so that

$$|e^{sd_2}|\leqslant|e^{\Re(s_n)d_2}|, \tag{B36}$$

and, for some constant, the right hand side of (B36) is less than const/$n^{(3d_2/2L)}$. We have seen that $|\Delta|$ is bounded from below on Γ_n. With the bound on $|\phi(s)|$ from (B4), we have then that

$$|\mathcal{I}_{\Gamma_n}(d_1)|<\frac{\text{const}}{n^{p+(3d_2/2L)-2}}\int_{\Gamma_n}e^{\Re(s)d_1}|ds|. \tag{B37}$$

Since the integrand is everywhere non-negative, we have that

$$\int_{\Gamma_n}e^{\Re(s)d_1}|ds|<\int_{\text{arc}}e^{\Re(s)d_1}|ds|, \tag{B38}$$

where the arc is the half circle with $|s|=|s_n^{mid}|=R_n$ in the left half plane. But, the integral along the half circle is

$$\int_{\text{arc}}e^{\Re(s)d_1}|ds|=R_n\int_0^{\pi}e^{-d_1R_n\sin\theta}d\theta=2R_n\int_0^{\pi/2}e^{-d_1R_n\sin\theta}d\theta$$
$$<R_n\int_0^{\pi/2}e^{-2d_1R_n\theta/\pi}d\theta=\frac{\pi}{d_1}(1-e^{-d_1R_n}). \tag{B39}$$

This completes the proof of the lemma.

Proof of main result: We define C_n as the integration path on the imaginary s axis from $i\Im(s_{-n})-i\pi/2L$ to $i\Im(s_n)+i\pi/2L$, and we define

$$\mathcal{I}_n(d_1)\equiv\int_{C_n}F(s)ds. \tag{B40}$$

We let $\mathcal{I}_{n,\text{closed}}(d_1)$ be the integral on the closed path consisting of C_n, of Γ_n, of H_n and of H_{-n} traced backwards. From the lemmas above we have

$$|\mathcal{I}_n(d_1)-\mathcal{I}_{n,\text{closed}}(d_1)|<|\mathcal{I}_{\Gamma_n}(d_1)|+|\mathcal{I}_{H_n}(d_1)|+|\mathcal{I}_{H_{-n}}(d_1)|<\frac{\text{const}}{n^{p+(3d_2/2L)-2}}. \tag{B41}$$

The integral on the closed path is $2i\pi$ times the sum of the residues inside the path,

$$\mathcal{I}_{n,\text{closed}}(d_1)=\sum_{-n}^{n}a^{(k)}e^{s_kd_1}, \tag{B42}$$

where $a^{(k)}$ is the residue of $s^2\phi(s)e^{sd_2}/\Delta$ at $s=s_k$. Since the only singularities of the integrand in the finite s plane are simple poles at the roots of $\Delta(s)$, these $a^{(k)}$ coefficients are those defined in (B7). We have then

$$\left|\mathcal{I}_n(d_1)-\sum_{-n}^{n} a^{(k)}e^{s_k d_1}\right|<\frac{\text{const}}{n^{p+(3d_2/2L)-2}}, \tag{B43}$$

and our main result follows from the fact that $\mathcal{I}(d_1)$ is the limit of $\mathcal{I}_n(d_1)$ as $n\to\infty$.

We now turn to the role played by the Cauchy data. In the Green's function solution for the waveform [see the discussion following (21)], a function of s appears representing the integral of the product of $f_-(s,x)$ and the combination $J(x,s)\equiv-\dot{\psi}_0(x)-s\psi_0(x)$. In the case of the TDP or spiked TDP, $f_-(s,x)=e^{sx}$. Let us suppose that the support of the Cauchy data is confined to the region $x_2<x<x_1$. The Cauchy data then enters the s integral through the function

$$\mathcal{J}(s)\equiv\int_{x_2}^{x_1}J(x,s)e^{sx}dx. \tag{B44}$$

If the initial waveform $\psi_0(x)$ satisfies

$$\left|\frac{d^{p+1}}{dx^{p+1}}\psi_0(x)\right|<b_0, \tag{B45}$$

then from integration by parts, we have

$$\int_{x_2}^{x_1}e^{sx}\psi_0(x)dx=-\frac{1}{s}\int_{x_2}^{x_1}e^{sx}\frac{d}{dx}[\psi_0(x)]dx \tag{B46}$$

$$=\ldots\pm\frac{1}{s^{p+1}}\int_{x_2}^{x_1}e^{sx}\frac{d^{p+1}}{dx^{p+1}}[\psi_0(x)]dx \tag{B47}$$

$$=\ldots\pm\frac{1}{s^{p+1}}e^{sx_2}\int_{x_2}^{x_1}e^{s(x-x_2)}\frac{d^{p+1}}{dx^{p+1}}[\psi_0(x)]dx. \tag{B48}$$

For $\Re(s)\leqslant 0$ the factor $e^{s(x-x_2)}$ in the last integral is $\leqslant 1$, so that

$$\left|e^{-sx_2}\int_{x_2}^{x_1}e^{sx}\psi_0(x)dx\right|<\frac{|b_0(x_1-x_2)|}{|s|^{p+1}}. \tag{B49}$$

If, in addition to the constraint in (B45) we have that the pth derivative of $\dot{\psi}_0(x)$ is bounded, then by a very similar argument we can show that $\phi(s)$, defined as $e^{-sx_2}\mathcal{J}(s)$, satisfies

$$|\phi(s)|<\frac{\text{const}}{|s|^p}. \tag{B50}$$

We can now apply the above mathematical results to the Green's function integral from Sec. III. The waveform is given by the following integral along the imaginary s axis:

$$\psi(t,x)=\frac{1}{2i\pi}\int\frac{e^{s(t-x)}\mathcal{J}(s)}{W(s)}ds, \tag{B51}$$

where $W(s)$ is given in (A21)–(A24) and has the form $W(s)=s^{-2}e^{-2sL}\Delta(s)$ in which $\Delta(s)$ is a special case of (B2) and (B3). We can therefore rewrite the solution as

$$\psi(t,x)=\frac{1}{2i\pi}\int\frac{s^2e^{s(t-x+x_2+2L)}\phi(s)}{\Delta(s)}ds. \tag{B52}$$

The above proof requires that p, d_1, d_2, and L satisfy the inequalities that follow (B4). The details of the proof show that the rate of convergence depends on these parameters. In particular, on the value of $\gamma\equiv p+(3d_2/2L)-2$. A small value of this parameter means slow convergence. We can now relate the details of the proof to the examples presented in Sec. III, and examine the interesting nature of the convergence of the series given there. We start by noting that a straightforward computation of $\phi(s)\equiv e^{-sx_2}\mathcal{J}(s)$, for the Cauchy data of (19), shows that $p=1$. Comparing (B1) and (B52) we see that

$$d_1+d_2=t-x+x_2+2L=t-x+x_1-x_1+x_2+2L. \tag{B53}$$

An ''obvious'' choice is to take $d_1=t-x+x_1$, the retarded time from the start of the reception of signals from the Cauchy data. At any x this is the equivalent of $t-t_{\min}$. With this choice we are left with $d_2=x_2-x_1+2L$. (Note that $2L-d_2=x_1-x_2$ is positive, as required in the proof.) The value of γ for this choice is given by

$$\gamma\equiv p+\frac{3d_2}{2L}-2=2-3\frac{x_1-x_2}{2L}. \tag{B54}$$

In our examples, we choose x_1-x_2, the range of support of the Cauchy data, to be 6 and we have $L=9$, so γ is unity. Suppose, though, that we had chosen $x_1=1$ (as in our standard sine wave data of Sec. III) but had decreased the value of x_2 below our standard choice, -5. The details of the proof show that convergence would require more terms for a given level of accuracy and that the series would fail to converge for $x_2\leqslant 11$. This limit can be extended if we use initial data with one or more continuous derivatives, i.e., if $p>1$. This rather unusual feature was, in fact, exactly what was observed in numerical experiments.

We point out next that Lemma 3 shows that convergence is not uniform in d_1. As d_1 gets smaller, more terms in the series are needed. With our choice of d_1 to be $t-t_{\min}$, this implies that the convergence of our QN series in (B6) is not uniform as $t-t_{\min}\to 0$, contradicting our claims of uniform convergence made following (5). But note that we can choose $d_1\equiv t-t_{\min}+1$, so that convergence *is* uniform for $t>t_{\min}$. In this case we have $\gamma=2-3(x_1-x_2+1)/2L$. For both our sine wave Cauchy data this γ has a numerical value of 5/6, and the series is convergent. It is clear that there is an interaction between the allowed range of the Cauchy data, and the range of t for which the QN series converges. By moving the left edge of the support of the Cauchy data to the left by some amount δ, we increase by $3\delta/(2L)$ the value of t at which the series first converges. It should also be noted that the dependence on d_1 explains why the QN series converge more quickly at early times than at late, a feature evident in Fig. 9.

[1] The following references contain examples of QN ringing in a variety of contexts, and contain references to further work: C. V. Vishveshwara, Nature (London) **227**, 936 (1970); W. H. Press, Astrophys. J. **170**, L105 (1971); S. L. Detweiler and E. Szedenits, Jr., *ibid.* **231**, 211 (1979); C. T. Cunningham, R. H. Price, and V. Moncrief, *ibid.* **236**, 674 (1980); K. Oohara, Prog. Theor. Phys. **71**, 738 (1984); R. F. Stark and T. Piran, Phys. Rev. Lett. **55**, 891 (1985); **56**, 97(E) (1986); A. M. Abrahams and C. R. Evans, Phys. Rev. D **46**, R4117 (1992); P. Anninos, D. Hobill, E. Seidel, L. Smarr, and W.-M. Suen, **71**, 2851 (1993); R. Gleiser, C. Nicasio, R. H. Price, and J. Pullin, Class. Quantum Grav. **13**, L117 (1996).
[2] H. R. Beyer, Max-Planck-Institut für Astrophsik preprint MPA **811**, 1994; J. Math. Phys. **36**, 4792–4814 (1995); **36**, 4815–4825 (1995).
[3] H. R. Beyer and B. G. Schmidt, Astron. Astrophys. **296**, 722–726 (1995).
[4] E. S. C. Ching, P. T. Leung, W.-M. Suen, and K. Young, Phys. Rev. Lett. **74**, 4588–4591 (1995); Phys. Rev. D **54**, 3778–3791 (1996).
[5] T. Regge and J. A. Wheeler, Phys. Rev. **108**, 1063 (1957).
[6] F. J. Zerilli, Phys. Rev. Lett. **24**, 737 (1970).
[7] S. Chandrasekhar, *The Mathematical Theory of Black Holes* (Clarendon, Oxford, 1983).
[8] H.-P. Nollert and B. G. Schmidt, Phys. Rev. D **45**, 2617 (1992).
[9] E. W. Leaver, Proc. R. Soc. London, Ser. A **402**, 285 (1985).
[10] V. Husain and R. H. Price, Phys. Rev. Lett. **68**, 1973 (1992).

[11] H. R. Beyer, Commun. Math. Phys. (submitted, 1998), gr-qc/9803034.
[12] G. Pöschl and E. Teller, Z. Phys. **83**, 143 (1933).
[13] V. Ferrari and B. Mashhoon, Phys. Rev. Lett. **16**, 1361 (1984).
[14] H.-P. Nollert, Phys. Rev. D **53**, 4397 (1996).
[15] R. H. Price and J. Pullin, Phys. Rev. Lett. **72**, 3297 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

A recursive method for the construction of irreducible representations of the orthogonal group O(n)

Nir Barnea[a)]
ECT, European Centre for Theoretical Studies in Nuclear Physics and Related Areas, Strada delle Tabarelle 286, 1-38050 Villazzano, Trento, Italy*

(Received 18 May 1998; accepted for publication 8 July 1998)

An algorithm for the construction of O(n) Gel'fand–Zetlin states in an invariant vector space is developed. The states are calculated recursively using a new type of coefficient of fractional parentage. These coefficients are the eigenvectors of the Gel'fand invariants. It is shown that the calculation of the Gel'fand invariants' matrix elements can be reduced to the evaluation of a single generator at each step. This algorithm provides a new approach to the calculation of the Clebsh–Gordon coefficients and isoscalar factors for the orthogonal group and can be applied to construct a basis function with well-defined orthogonal symmetry for physical systems where separation between collective motions and intrinsic motions, associated with the group O(n), is of interest.
[S0022-2488(99)01702-8]

I. INTRODUCTION

A few years ago Novoselsky, Katriel, and Gilmore (NKG)[1] developed a new algorithm for the construction of irreducible representations (irreps) of the symmetric group. In their approach they used the canonical nature of the group subgroup chain of the symmetric groups, $S_n \supset S_{n-1} \supset \cdots \supset S_2 \supset S_1$, and formulated a recursive algorithm in which one starts with basis functions that belong to the irreps of S_1, then transforms this basis into new basis functions that belong to well-defined irreps of S_2. These basis functions are then symmetrized with respect to S_3, and so on. After $n-1$ recurrence steps one gets basis functions that belong to well-defined irreps of the symmetric group S_n and are characterized by the Yamanouchi symbols. Assuming that basis functions with well-defined S_{n-1} symmetry have already been constructed, NKG have shown that the eigenvalues of the transposition class-sum operator uniquely identify the irreps of symmetric group S_n, and, therefore, the eigenvectors are the new basis functions which belong to well defined irreps of S_n. The key point in this algorithm is that the calculation of the matrix elements of the transposition class-sum operator can be reduced to the calculation of the matrix elements of the transposition $(n,n-1)$. The algorithm has been successfully applied for a number of mathematical and physical problems, such as the calculation of coefficients of fractional parentage in the $L-S$ coupling scheme[1] and the evaluation of the inner product and the outer product isoscalar factors of the symmetric group.[2] It was also applied for constructing harmonic oscillator nonspurious states[3] and hyperspherical states[4] with definite permutational symmetry. In all these applications the *same* basic algorithm has been employed using the appropriate realization of the transposition $(n,n-1)$ and the invariant subspaces. Comparing it with other symmetrization methods the NKG algorithm was proven to be very suitable for numerical calculations.

In the current work we present a generalization of the NKG algorithm for a canonical chain of Lie groups. This paper is devoted to the orthogonal group O(n) because of its importance in physics.[5] However, the results may be applied with minor necessary changes to the unitary group as well.

The orthogonal group O(n) is the set of all linear transformations in the n-dimensional real

[a)]Electronic mail: barnea@ect.unitn.it

Euclidean space which conserve the quadratic form, i.e., the sum over all the squares of the coordinates. This group is actually the set of all $n\times n$ real orthogonal matrices with determinant ±1 [For the special case where the determinant of these matrices is only $+1$ we obtain the special orthogonal group SO(n).] The irreducible representations of this group are well known. The analog of the Yamanouchi symbol for O(n) is the Gel'fand–Zetlin (GZ) pattern[6,7] which provides an elegant way for labeling the states in a definite O(n) irrep in terms of the canonical group subgroup chain $\mathrm{O}(n)\supset\mathrm{O}(n-1)\supset\cdots\supset\mathrm{O}(3)\supset\mathrm{O}(2)$.

The orthogonal group and the GZ basis play already an important role in the theoretical microscopic approach to nuclear collective models[8,9]. A problem of $3n$ degrees of freedom of n nucleons can be described by 9 collective degrees of freedom and $3n-9$ internal degrees of freedom which may be associated with the manifold $\mathrm{O}(n-1)/\mathrm{O}(n-4)$. Looking for wave functions with a definite irrep of a dynamical group Sp($6n$), where $\mathrm{Sp}(6)\times\mathrm{O}(n-1)$ is one of its subgroups, Moshinsky[10] has noted that the collective effects can be introduced by the constraint that the n-body wave function is restricted to a given irrep of the orthogonal group $\mathrm{O}(n-1)$. Moshinsky and Quesene have shown in Ref. 11 that the irrep of the group $\mathrm{O}(n-1)$ determines that of the group $Sp(6)$ and thus these two groups are ''complementary.'' Therefore, some of the microscopic collective models of the nuclei are related to the $Sp(6)$ group and others to the $\mathrm{O}(n-1)$ group.

Another physical application for the GZ states and the O(n) group-subgroup chain, arises in a few body calculations where one tries to construct n-body basis functions with definite permutational symmetry. The symmetric group S_n is a subgroup of the orthogonal group $\mathrm{O}(n-1)$. Knowing that, Surkov[12] suggested many years ago to use the orthogonal group as an intermediate step in the construction of symmetric hyperspherical states. In his work, Surkov constructed, at first, four-body hyperspherical states, of power 4 or less, that belong to well defined irreps of O(3) and then reduced these states into irreps of the symmetric group S_4. The main difficulty in applying these ideas to physics is the problem of constructing basis functions which belong to a definite GZ state. This problem was a great obstacle in the development of the orthogonal group approach to the n-body problem[10,13,14] and in further development of Surkov's approach.

Recently Barnea and Novoselsky[5] have extended the NKG method for the calculation of hyperspherical states with definite $\mathrm{O}(n-1)$ GZ pattern. This states where later symmetrized using the original NKG method for the evaluation of the $\mathrm{O}(n-1)\downarrow S_n$ coefficients of fractional parentage. The new computational algorithm was applied for the construction of states with arbitrary permutational symmetry and found to be substantially more efficient than the original one. In that work,[5] we used the second-order Casimir operator to separate the O(n) irreps. However, unlike the transposition class-sum operator the eigenvalues of the second-order Casimir operator of O(n) are degenerate, i.e., they have different irreps with the same eigenvalue, for instance, the irreps (4,0) and (3,3) of O(4) have the same eigenvalue 24. The usual method to remove this degeneracy is to use higher-order Casimir operators, also known as the Gel'fand invariants. In the case of the orthogonal group one should use only the forth-, sixth-, ..., i.e., the even-order operators since the eigenvalues of the $(2k+1)$th-order operator can be written as a function of the eigenvalues of the second-, forth-, $\ldots,2k$-th order operators. However, the Gel'fand operators involve a sum over products of generators and their evaluation is an unpleasant task in practice. Therefore, we were forced to use the Gram–Schmidt procedure in order to remove the degeneracy.

In this paper we present a method which overcome this difficulty, and reduce the evaluation of the Gel'fand invariants to the evaluation of a single generator at each recurrence step. In order to overcome this difficulty we follow Edwards[15] who has shown that the mth order Gel'fand invariant can be related with the trace of the mth power of a suitably defined operator P. Using this idea, the calculation of the Gel'fand invariants is reduced to evaluation of the matrix elements of a single generator, namely $X_{k,k-1}$ at each step. Unlike Ref. 5, which was dedicated for the specific case of the hyperspherical basis function, in this work we present general derivation of the recursive algorithm for constructing O(n) Gel'fand–Zetlin states in terms of orthogonal parentage coefficients, opcs. This algorithm can be applied for the efficient construction of n-body states as well as for the calculation of the Clebsh–Gordon coefficients and the outer product isoscalar factors of the orthogonal group. Its application for the unitary group is straightforward. The only differences that come along with the application of the method to various problems are the identification of the invariant subspaces and the realization of the generators.

The algorithm proceeds as follows:

(i) After $(k-1)$ recurrence steps the $O(n)$ invariant space is reduced into $O(k)$ invariant subspaces, each correspond to a single $O(k)$ irrep. Every subspace is spanned by the appropriate GZ states.

(ii) The even-order Gel'fand invariants, i.e., the second-order Casimir operator, the forth-order Casimir operator, etc., of the group $O(k+1)$ are diagonalized in this basis. The operators are diagonalized one by one until all the irreps are uniquely identified.

(iii) After the diagonalization of the appropriate matrices:

(1) The eigenvalues identify the irrep of the relevant group to which the eigenvector belong, and
(2) the eigenvectors are the $O(k)\rightarrow O(k+1)$ transformation coefficients. These are the orthogonal parentage coefficients, opcs.

(iv) If $k+1$ is an even number, we diagonalize the generator $X_{k,k+1}$ in the highest weight $O(k)$ state, in order to distinguish the two $SO(k+1)$ irreps in $O(k+1)$.

We start this paper by reviewing the concept of parentage coefficients in view of its role in our algorithm; we also define the orthogonal parentage coefficients, opcs, and their properties. The Gel'fand invariants and Edwards method for their evaluation are then presented in Sec. III. In Sec. IV we derive the recurrence relation for the calculation of the P operator and then present in Sec. V the evaluation of the Gel'fand invariants in terms of the generator $X_{n,n-1}$. The $O(k+1)$ Gel'fand invariants do not mix states from different $O(k)$ irreps since they commute with all the generators of $O(k)$. Therefore, different $O(k)$ states enter into the $O(k+1)$ irrep with arbitrary phase. The procedure used to solve this problem is also discussed in Sec. V. A full and systematic presentation of the algorithm presented above is given in Sec. VI.

II. THE ORTHOGONAL PARENTAGE COEFFICIENTS

The rank, r, of the Lie algebra associated with the Lie group $O(k)$ is given by $r=[k/2]$. Therefore the irreducible representations (irreps) of the orthogonal group $O(k)$, $\boldsymbol{\lambda}_k$, are labeled by $\boldsymbol{\lambda}_k=(\lambda_{k,1},\lambda_{k,2},\ldots,\lambda_{k,r})$. The values of $\lambda_{k,j}$ are positive and either all integers or half-integers. In the case $k=2r$, the $O(k)$ representations $(\lambda_{k,1},\lambda_{k,2},\ldots,\lambda_{k,r})$ splits into the two $SO(k)$ irreps $(\lambda_{k,1},\lambda_{k,2},\ldots,\lambda_{k,r})$ and $(\lambda_{k,1},\lambda_{k,2},\ldots,-\lambda_{k,r})$ if $\lambda_{k,r}\neq 0$, while for $\lambda_{k,r}=0$ or odd k the irreps of $O(k)$ are irreducible under $SO(k)$. In what follows we shall be working with the group $SO(k)$, remembering the connection between its irreps and the irreps of $O(k)$. The basis vectors of an irrep $\boldsymbol{\lambda}_k$ are completely specified by the Gel'fand-Zetlin[5] states, $\boldsymbol{\Lambda}_k$. These states, $\boldsymbol{\Lambda}_k$, are labeled by the canonical chain of subgroups $SO(k)\supset SO(k-1)\supset\cdots\supset SO(3)\supset SO(2)$, or $\boldsymbol{\Lambda}_k=[\boldsymbol{\lambda}_k\boldsymbol{\lambda}_{k-1}\cdots\boldsymbol{\lambda}_2]$, and are restricted according to the following rules, given by Gel'fand and Zetlin,[6] $\lambda_{k,j}\geqslant\lambda_{k-1,j}\geqslant|\lambda_{k,j+1}|$ and $\lambda_{k,j}\geqslant|\lambda_{k-1,j}|$.

The purpose of this paper is to present a recursive method to transform a given basis, $\{|a\rangle,a=1,\ldots,dim(V_p)\}$, of an $O(n)$ invariant vector space V_p into the Gel'fand–Zetlin basis, i.e., we are looking for the transformation

$$|\boldsymbol{\Lambda}_k^p a_k\rangle=\sum_{a_1=1}^{\dim(V_p)} U^{(k)}_{\boldsymbol{\Lambda}_k^p a_k,a_1}|a_1\rangle. \tag{1}$$

Here a_k is the degeneracy of the $SO(k)$ irrep $\boldsymbol{\lambda}_k$ in V_p, $a_1=a$, $\mathbf{U}^{(k)}$ is an unitary matrix and the superscript p is used to denote the vectors in V_p, the primary space (we are going to introduce another, secondary vector space in the next section). The main idea in the recursive method is to present $\mathbf{U}^{(k)}$ as a product of a sequence of block diagonal matrices $\{\mathbf{C}^{(m)},m=2,\ldots,k\}$, where $\mathbf{C}^{(m)}$, the step transformation matrix, is used to transform the $O(m-1)$ GZ states into $O(m)$ states,

$$|\mathbf{\Lambda}_m^p a_m\rangle=\sum_{a_{m-1}} C^{(m)\mathbf{\Lambda}_{m-1}^p}_{\boldsymbol{\lambda}_m^p a_m,a_{m-1}}|\mathbf{\Lambda}_{m-1}^p a_{m-1}\rangle. \tag{2}$$

The step transformation matrix $\mathbf{C}^{(m)}$ is block diagonal since $\mathbf{\Lambda}_m^p=[\boldsymbol{\lambda}_m^p,\mathbf{\Lambda}_{m-1}^p]$ and it doesn't alter the value of $\mathbf{\Lambda}_{m-1}^p$. Therefore, we can see that

$$U^{(k)}_{\mathbf{\Lambda}_k^p a_k,a_1}=\sum_{a_{k-1},a_{k-2},\ldots,a_2} C^{(k)\mathbf{\Lambda}_{k-1}^p}_{\boldsymbol{\lambda}_k^p a_k,a_{k-1}} C^{(k-1)\mathbf{\Lambda}_{k-2}^p}_{\boldsymbol{\lambda}_{k-2}^p a_{k-1},a_{k-2}}\cdots C^{(2)}_{\boldsymbol{\lambda}_2^p a_2,a_1}. \tag{3}$$

The orthogonal parentage coefficients, opcs, are the matrix elements of the step transformation matrix $C^{(m)\mathbf{\Lambda}_{m-1}^p}_{\boldsymbol{\lambda}_m^p a_m,a_{m-1}}$. The opcs, defined in Eq. (2) satisfy, the orthogonality relation

$$\sum_{a_{m-1}} C^{(m)\mathbf{\Lambda}_{m-1}^p}_{\boldsymbol{\lambda}_m^p a_m,a_{m-1}}{}^{*}C^{(m)\mathbf{\Lambda}_{m-1}^p}_{\boldsymbol{\lambda}'^p_m a'_m,a_{m-1}}=\delta_{\boldsymbol{\lambda}'^p_m,\boldsymbol{\lambda}_m^p}\delta_{a'_m,a_m}, \tag{4}$$

which follows from the orthogonality of the mth step basis states. The opcs also satisfy the completeness relation

$$\sum_{\boldsymbol{\lambda}_m^p,a_m} C^{(m)\mathbf{\Lambda}_{m-1}^p}_{\boldsymbol{\lambda}_m^p a_m,a_{m-1}}{}^{*}C^{(m)\mathbf{\Lambda}_{m-1}^p}_{\boldsymbol{\lambda}_m^p a_m,a'_{m-1}}=\delta_{a'_{m-1},a_{m-1}}. \tag{5}$$

As an example, let V_p be the invariant subspace created by the outer product of two irreps of O(n), i.e., $V_p=\boldsymbol{\lambda}_n^1\otimes\boldsymbol{\lambda}_n^2$. In this case the matrix elements of the transformation matrix $\mathbf{U}^{(k)}$, which transforms the outer product subspace $\boldsymbol{\lambda}_n^1\otimes\boldsymbol{\lambda}_n^2$ into the GZ states $\boldsymbol{\lambda}_n^p$ are just the O(n) Clebsh–Gordon coefficients, and the opcs are the O(n) isoscalar factors.

III. THE O(n) GENERATORS AND THE P OPERATOR

Let X_{ij} be the generators of the group O(n) in the primary invariant vector space, V_p, satisfying the following commutation relations:

$$[X_{ij},X_{kl}]=-i(\delta_{jk}X_{il}+\delta_{il}X_{jk}-\delta_{ik}X_{jl}-\delta_{jl}X_{ik}). \tag{6}$$

The set of operators, namely the Gel'fand invariants for O(n), defined by

$$I_k(X_{ij})=\sum_{i_1,i_2,\ldots,i_k}^{n} X_{i_1i_2}X_{i_2i_3}\ldots X_{i_ki_1}, \tag{7}$$

are a set of Casimir operators for O(n), which commute with the algebra generators, i.e., $[I_k(X_{ij}),X_{i_1i_2}]=0$ for $i_1,i_2=1,\ldots,n$. The eigenvalue of $I_k(X_{ij})$ corresponding to a given irrep of O(n) or SO(n), presented by $\boldsymbol{\lambda}_n$, can be written in the following way:[15]

$$I_k|_{\boldsymbol{\lambda}_n}=\sum_{i=1}^{n}(q_i)^k\Pi_{j=1 j\neq i}^{n}\frac{q_j-q_i+1+\epsilon_{ji}}{q_j-q_i}, \tag{8}$$

where

$$\epsilon_{ji}=\begin{cases}-\delta_{j,n+1-i} & \text{for } n=2r,\\ \delta_{j,n+1-i}-\delta_{i,r+1} & \text{for } n=2r+1,\end{cases} \tag{9}$$

and the q_i's are given by the following expressions: for $i=1,\ldots,r=[(n/2)]$,

$$q_i=\lambda_{n,i}+n-i-1, \tag{10}$$

for $i=n-r+1,\dots,n$,

$$q_i=-\lambda_{n,n+1-i}+n-i, \tag{11}$$

and if n is an odd number, then

$$q_{(n+1)/2}=r. \tag{12}$$

The spectra of the Gel'fand invariants can be used to uniquely identify the irreps of $\mathrm{O}(n)$. It turns out[16] that the invariant operators $I_k(X_{ij})$ with odd k are not independent and can be expressed in terms of the $I_{2j}(X_{ij})$ operators with $2j<k$. However, the Gel'fand operators $I_k(X_{ij})$ with even k are invariant to reflection and therefore they can't be used to split an $\mathrm{O}(n)$ irrep into the proper $\mathrm{SO}(n)$ irreps. Following Edwards[15] we introduce a second invariant vector space, V_s, taken to be the carrier space of the fundamental representation $1_n=(1,0,\dots,0)$ and denote by E_{ij} the generators of $\mathrm{O}(n)$ on V_s. We may now define a representation on the tensor product carrier space $V=V_s\otimes V_p$ by the generators $G_{ij}=E_{ij}\otimes 1+1\otimes X_{ij}$. Consider now the operator P_n in $V_s\otimes V_p$, defined by

$$P_n=\frac{1}{2}\sum_{ij}^{n} E_{ij}\otimes X_{ij}. \tag{13}$$

We shall see that the kth-order Gel'fand invariant can be identified with the trace of $(P_n)^k$. P_n commutes with the generators G_{ij} as can be seen from the identity

$$I_2(G_{ij})=(E_{ij}\otimes 1+1\otimes X_{ij})(E_{ij}\otimes 1+1\otimes X_{ij})=I_2(E_{ij})\otimes 1+4P_n+1\otimes I_2(X_{ij}). \tag{14}$$

We may choose a basis for V_s such that the generators of $\mathrm{O}(n)$ have matrix elements:

$$\langle\alpha|E_{ij}|\beta\rangle=\delta_{i\alpha}\delta_{j\beta}-\delta_{j\alpha}\delta_{i\beta}. \tag{15}$$

If we denote by $|a\rangle$ the set of basis vectors of V_p, then by writing $|\alpha;a\rangle$ for $|\alpha\rangle\otimes|a\rangle$ we have

$$\langle\alpha;a|P_n|\beta;b\rangle=\langle a|X_{\alpha\beta}|b\rangle. \tag{16}$$

Then

$$\begin{aligned}\langle\alpha;a|(P_n)^k|\beta;b\rangle&=\sum_{a_1\dots a_{k-1}}\sum_{\alpha_1\dots\alpha_{k-1}}\langle\alpha;a|P_n|\alpha_1;a_1\rangle\\&\quad\times\langle\alpha_1;a_1|P_n|\alpha_2;a_2\rangle\dots\langle\alpha_{k-1};a_{k-1}|P_n|\beta;b\rangle\\&=\sum_{a_1\dots a_{k-1}}\sum_{\alpha_1\dots\alpha_{k-1}}\langle a|X_{\alpha\alpha_1}|a_1\rangle\langle a_1|X_{\alpha 1\alpha_2}|a_2\rangle\dots\langle a_{k-1}|X_{\alpha_{k-1}\beta}|b\rangle\\&=\sum_{\alpha_1\dots\alpha_{k-1}}\langle a|X_{\alpha\alpha_1}X_{\alpha 1\alpha_2}\dots X_{\alpha_{k-1}\beta}|b\rangle.\end{aligned} \tag{17}$$

Thus,

$$\sum_{\alpha}\langle\alpha;a|(P_n)^k|\alpha;b\rangle=\sum_{\alpha\alpha_1\dots\alpha_{k-1}}\langle a|X_{\alpha\alpha_1}X_{\alpha_1\alpha_2}\dots X_{\alpha_{k-1}\alpha}|b\rangle=\langle a|I_k(X_{ij})|b\rangle. \tag{18}$$

It should be noted that this result is independent of the choice of basis for V_s, as we use the trace operator in this space.

IV. THE RECURRENCE RELATIONS FOR P_n

In this section the general algorithm for the construction of the GZ states is given. First let us assume the existence of a complete set of basis vectors, $|\boldsymbol{\Lambda}_{n-1}^p a_{n-1}\rangle$, for the primary space, V_p. These states are characterized by the SO$(n-1)$ GZ states and by a_{n-1}, the degeneracy of the SO$(n-1)$ irrep $\boldsymbol{\lambda}_{n-1}^p$ in V_p. Our aim is to transform this basis vectors into basis vectors that belong to well-defined irreps of the group SO(n). The actual transformation will take place by diagonalization of the Gel'fand invariants, Eq. (7), one by one until all the O(n) irreps are completely determined. The procedure for reducing the O(n) irreps into the proper SO(n) irreps will be presented below. The common eigenvectors of $I_k, \{k=2,4,\ldots,n\}$ are simply the step transformation matrix $\mathbf{C}^{(n)}$ for SO$(n-1)\rightarrow$O(n), whose elements are the opcs. As we have seen in the previous section, the calculation of the matrix elements of the Gel'fand invariants is equivalent to the calculation of the matrix elements of the operator P_n in the tensor product space V. The operator P_n presented in Sec. III, Eq. (13), can be rewritten as

$$P_n = P_{n-1} + \sum_{i=1}^{n-1} E_{ni} \otimes X_{ni}. \tag{19}$$

Since P_n commutes with the generators of O(n) in the tensor product carrier space V, the set $P_2, P_3, \ldots, P_n$ forms a set of mutually commuting operators. More than that, P_n is a block diagonal and does not mix vectors that belong to different irreps of SO(k) for $k<n$. Denoting by $|\boldsymbol{\Lambda}_{n-1}^s\rangle$ the basis vectors of V_s, we can use the SO$(n-1)$ Clebsch–Gordan coefficients to construct basis vectors of irreducible representations in the tensor product space:

$$|(\boldsymbol{\lambda}_{n-1}^s;\boldsymbol{\lambda}_{n-1}^p a_{n-1})\boldsymbol{\Lambda}_{n-1}\rangle = \sum_{\boldsymbol{\Lambda}_{n-2}^s \boldsymbol{\Lambda}_{n-2}^p} \left(\begin{matrix} \boldsymbol{\lambda}_{n-1}^s & \boldsymbol{\lambda}_{n-1}^p \\ \boldsymbol{\Lambda}_{n-2}^s & \boldsymbol{\Lambda}_{n-2}^p \end{matrix} \middle| \begin{matrix} \boldsymbol{\lambda}_{n-1} \\ \boldsymbol{\Lambda}_{n-2} \end{matrix} \right) |\boldsymbol{\lambda}_{n-1}^s \boldsymbol{\Lambda}_{n-2}^s\rangle \otimes |\boldsymbol{\lambda}_{n-1}^p \boldsymbol{\Lambda}_{n-2}^p a_{n-1}\rangle. \tag{20}$$

In principle there might be multiplicity of the representation $\boldsymbol{\lambda}_{n-1}$ in the tensor product space $1_{n-1}\otimes\boldsymbol{\lambda}_{n-1}^p$. However, as can be seen in Appendix A, as far as we are concerned there is no such problem. These vectors, Eq. (20), are eigenvectors of the second-order Casimir operators in V_p, V_s and in V. Therefore, as a result of Eq. (14) they are eigenvectors of P_{n-1} as well,

$$\begin{aligned} &\langle(\boldsymbol{\lambda}_{n-1}^s;\boldsymbol{\lambda}_{n-1}^p)\boldsymbol{\Lambda}_{n-1}a_{n-1}|P_{n-1}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\Lambda}'_{n-1}a'_{n-1}\rangle \\ &\quad = \delta_{\boldsymbol{\Lambda}_{n-1}\boldsymbol{\Lambda}'_{n-1}} \delta_{\boldsymbol{\lambda}_{n-1}^s,\boldsymbol{\lambda}'^s_{n-1}} \delta_{\boldsymbol{\lambda}_{n-1}^p,\boldsymbol{\lambda}'^p_{n-1}} \delta_{a_{n-1},a'_{n-1}} \tfrac{1}{4}[I_2|(\boldsymbol{\lambda}_{n-1}) - I_2|(\boldsymbol{\lambda}_{n-1}^p) - I_2|(\boldsymbol{\lambda}_{n-1}^s)]. \end{aligned} \tag{21}$$

The second term on the right-hand side of Eq. (19) can be evaluated recalling the fact that an element $g_{i,n-1}$ always exists in SO$(n-1)$ such that $E_{n,i}\otimes X_{n,i} = g_{i,n-1}^{-1}(E_{n,n-1}\otimes X_{n,n-1})g_{i,n-1}$. Therefore the matrix element of the second term in (19) is

$$\begin{aligned} &\sum_{i=1}^{n-1} \langle(\boldsymbol{\lambda}_{n-1}^s;\boldsymbol{\lambda}_{n-1}^p)\boldsymbol{\Lambda}_{n-1}a_{n-1}|E_{n,i}\otimes X_{n,i}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\Lambda}'_{n-1}a'_{n-1}\rangle \\ &\quad = \delta_{\boldsymbol{\Lambda}_{n-1}\boldsymbol{\Lambda}'_{n-1}} \sum_{i=1}^{n-1} \langle(\boldsymbol{\lambda}_{n-1}^s;\boldsymbol{\lambda}_{n-1}^p)\boldsymbol{\Lambda}_{n-1}a_{n-1}|g_{i,n-1}^{-1}(E_{n,n-1} \\ &\qquad \otimes X_{n,n-1})g_{i,n-1}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\Lambda}_{n-1}a'_{n-1}\rangle. \end{aligned} \tag{22}$$

The $\delta_{\boldsymbol{\Lambda}_{n-1}\boldsymbol{\Lambda}'_{n-1}}$ term on the rhs of Eq. (22) results from the fact that P_n is a scalar operator with respect to O$(n-1)$ in V. Replacing the $g_{i,n-1}$ operators by their representation matrices $D^{(\boldsymbol{\lambda}_{n-1})}_{\boldsymbol{\Lambda}_{n-2},\boldsymbol{\Lambda}'_{n-2}}(g_{i,n-1})$ we obtain the following expression for the matrix element (22):

$$\sum_{i=1}^{n-1} \langle(\boldsymbol{\lambda}^s_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\Lambda}_{n-1}a_{n-1}|E_{n,i}\otimes X_{n,i}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\Lambda}'_{n-1}a'_{n-1}\rangle$$
$$=\delta_{\boldsymbol{\Lambda}_{n-1}\boldsymbol{\Lambda}'_{n-1}}\sum_{i=1}^{n-1}\sum_{\boldsymbol{\Lambda}''_{n-2}} D^{(\boldsymbol{\lambda}_{n-1})}_{\boldsymbol{\Lambda}_{n-2},\boldsymbol{\Lambda}''_{n-2}}(g^{-1}_{i,n-1})D^{(\boldsymbol{\lambda}_{n-1})}_{\boldsymbol{\Lambda}''_{n-2},\boldsymbol{\Lambda}_{n-2}}(g_{i,n-1})$$
$$\times\langle(\boldsymbol{\lambda}^s_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\lambda}_{n-1}\boldsymbol{\Lambda}''_{n-2}a_{n-1}|E_{n,n-1}\otimes X_{n,n-1}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\lambda}_{n-1}\boldsymbol{\Lambda}''_{n-2}a'_{n-1}\rangle. \quad (23)$$

Deriving Eq. (23) we used the fact that the operators $X_{n,n-1}$ and $E_{n,n-1}$ commute with the generators of $O(n-2)$ and therefore the matrix element on the right-hand side is diagonal with respect to the $O(n-2)$ GZ states $\boldsymbol{\Lambda}''_{n-2}$. Since the matrix element is independent of $\boldsymbol{\Lambda}_{n-2}$, we can equivalently sum Eq. (23) on $\boldsymbol{\Lambda}_{n-2}$ and divide by $|\boldsymbol{\lambda}_{n-1}|$, the dimension of the $SO(n-1)$ irrep $\boldsymbol{\lambda}_{n-1}$, to get

$$\sum_{i=1}^{n-1} \langle(\boldsymbol{\lambda}^s_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\Lambda}_{n-1}a_{n-1}|E_{n,i}\otimes X_{n,i}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\Lambda}'_{n-1}a'_{n-1}\rangle$$
$$=\delta_{\boldsymbol{\Lambda}_{n-1}\boldsymbol{\Lambda}'_{n-1}}\frac{n-1}{|\boldsymbol{\lambda}_{n-1}|}\sum_{\boldsymbol{\Lambda}''_{n-2}}\langle(\boldsymbol{\lambda}^s_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\lambda}_{n-1}\boldsymbol{\Lambda}''_{n-2}a_{n-1}|E_{n,n-1}$$
$$\otimes X_{n,n-1}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\lambda}_{n-1}\boldsymbol{\Lambda}''_{n-2}a'_{n-1}\rangle. \quad (24)$$

Using Eqs. (15) and (20) and noting that in the fundamental representation

$$|n\rangle=[\boldsymbol{\lambda}_n=1_n,\boldsymbol{\lambda}_{n-1}=0_{n-1},\boldsymbol{\Lambda}_{n-2}=[0_{n-2}]],$$
$$|n-1\rangle=[\boldsymbol{\lambda}_n=1_n,\boldsymbol{\lambda}_{n-1}=1_{n-1},\boldsymbol{\Lambda}_{n-2}=[0_{n-2}]] \quad (25)$$

[here 1_n stands for the fundamental representation of $O(n)$ and 0_n stands for the scalar representation], the matrix element on the rhs of Eq. (24) takes the following form:

$$\langle(\boldsymbol{\lambda}^s_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\lambda}_{n-1}\boldsymbol{\Lambda}''_{n-2}a_{n-1}|E_{n,n-1}\otimes X_{n,n-1}|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\lambda}_{n-1}\boldsymbol{\Lambda}''_{n-2}a'_{n-1}\rangle$$
$$=(\delta_{\boldsymbol{\lambda}^s_{n-1},0_{n-1}}\delta_{\boldsymbol{\lambda}'^s_{n-1},1_{n-1}}-\delta_{\boldsymbol{\lambda}^s_{n-1},1_{n-1}}\delta_{\boldsymbol{\lambda}'^s_{n-1},0_{n-1}})$$
$$\times\left(\begin{array}{cc|c}\boldsymbol{\lambda}^s_{n-1} & \boldsymbol{\lambda}^p_{n-1} & \boldsymbol{\lambda}_{n-1}\\ [0_{n-2}] & \boldsymbol{\Lambda}^p_{n-2} & \boldsymbol{\Lambda}''_{n-2}\end{array}\right)^*\left(\begin{array}{cc|c}\boldsymbol{\lambda}'^s_{n-1} & \boldsymbol{\lambda}'^p_{n-1} & \boldsymbol{\lambda}_{n-1}\\ [0_{n-2}] & \boldsymbol{\Lambda}^p_{n-2} & \boldsymbol{\Lambda}''_{n-2}\end{array}\right)$$
$$\times\langle\boldsymbol{\lambda}^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}a_{n-1}|X_{n,n-1}|\boldsymbol{\lambda}'^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}a'_{n-1}\rangle. \quad (26)$$

Summing up the results of Eqs. (19) and (26) we get the expression for the matrix elements of P_n,

$$\langle(\boldsymbol{\lambda}^s_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\Lambda}_{n-1}a_{n-1}|P_n|(\boldsymbol{\lambda}'^s_{n-1};\boldsymbol{\lambda}'^p_{n-1})\boldsymbol{\Lambda}'_{n-1}a'_{n-1}\rangle$$
$$=\delta_{\boldsymbol{\Lambda}_{n-1}\boldsymbol{\Lambda}'_{n-1}}\Bigg\{\frac{1}{4}[I_2(\boldsymbol{\lambda}_{n-1})-I_2(\boldsymbol{\lambda}^p_{n-1})-I_2(\boldsymbol{\lambda}^s_{n-1})]\delta_{\boldsymbol{\lambda}^s_{n-1},\boldsymbol{\lambda}'^s_{n-1}}\delta_{\boldsymbol{\lambda}^p_{n-1},\boldsymbol{\lambda}'^p_{n-1}}\delta_{a_{n-1},a'_{n-1}}$$
$$+\frac{n-1}{|\boldsymbol{\lambda}_{n-1}|}\sum_{\boldsymbol{\Lambda}''_{n-2}}\langle\boldsymbol{\lambda}^p_{n-1}\boldsymbol{\Lambda}''_{n-2}a_{n-1}|X_{n,n-1}|\boldsymbol{\lambda}'^p_{n-1}\boldsymbol{\Lambda}''_{n-2}a'_{n-1}\rangle$$
$$\times\left[\left(\begin{array}{cc|c}1_{n-1} & \boldsymbol{\lambda}'^p_{n-1} & \boldsymbol{\lambda}_{n-1}\\ [0_{n-2}] & \boldsymbol{\Lambda}''_{n-2} & \boldsymbol{\Lambda}''_{n-2}\end{array}\right)\delta_{\boldsymbol{\lambda}^p_{n-1},\boldsymbol{\lambda}_{n-1}}-\left(\begin{array}{cc|c}1_{n-1} & \boldsymbol{\lambda}^p_{n-1} & \boldsymbol{\lambda}_{n-1}\\ [0_{n-2}] & \boldsymbol{\Lambda}''_{n-2} & \boldsymbol{\Lambda}''_{n-2}\end{array}\right)^*\delta_{\boldsymbol{\lambda}'^p_{n-1},\boldsymbol{\lambda}_{n-1}}\right]\Bigg\}. \quad (27)$$

The SO(n) CG coefficients needed in Eq. (27) were explicitly evaluated by Pang and Hecht[17] and are given in the appendix. Therefore the calculation of P_n is reduced to the calculation of $X_{n,n-1}$ at each recurrence step. It is assumed that we know how to evaluate the matrix elements of the generator $X_{n,n-1}$ between bra and ket states that belong to the original basis, i.e., $\{|a_1\rangle, a_1 = 1,\ldots,\dim(V_p)\}$, of the primary invariant subspace V_p. Then using the transformation matrix $\mathbf{U}^{(n-1)}$ defined in Eq. (1), we get the following expression

$$\langle \boldsymbol{\lambda}^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}a_{n-1}|X_{n,n-1}|\boldsymbol{\lambda}'^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}a'_{n-1}\rangle = \sum_{a_1,a'_1=1}^{\dim(V_p)} U^{(n-1)*}_{[\boldsymbol{\lambda}^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}]a_{n-1},a_1} U^{(n-1)}_{[\boldsymbol{\lambda}'^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}]a'_{n-1},a'_1}\langle a_1|X_{n,n-1}|a'_1\rangle, \tag{28}$$

for the matrix element of $X_{n,n-1}$. While applying this algorithm to the construction of hyperspherical or harmonic oscillator GZ functions we know how to evaluate the matrix elements of this generator between bra and ket states that belong to the set of basis states obtained before the last recurrence step. If this is the case we can save a lot of computation time and computer memory, skip the calculation of the transformation matrix $\mathbf{U}^{(n-1)}$, and write

$$\langle \boldsymbol{\lambda}^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}a_{n-1}|X_{n,n-1}|\boldsymbol{\lambda}'^p_{n-1}\boldsymbol{\Lambda}^p_{n-2}a'_{n-1}\rangle = \sum_{a_{n-2},a'_{n-2}} C^{(n-1)\boldsymbol{\Lambda}^p_{n-2}*}_{\boldsymbol{\lambda}^p_{n-1}a_{n-1},a_1} C^{(n-1)\boldsymbol{\Lambda}^p_{n-2}}_{\boldsymbol{\lambda}'^p_{n-1}a'_{n-1},a'_1}\langle \boldsymbol{\Lambda}^p_{n-2}a_{n-2}|X_{n,n-1}|\boldsymbol{\Lambda}^p_{n-2}a'_{n-2}\rangle. \tag{29}$$

As an example, let V_p be again the invariant subspace created by the outer product of two irreps of O(n), i.e., $V_p = \boldsymbol{\lambda}^1_n \otimes \boldsymbol{\lambda}^2_n$. The basis states obtained after $n-2$ recurrence steps are labeled in this example by $|\boldsymbol{\lambda}^1_n\boldsymbol{\lambda}^2_n(\boldsymbol{\lambda}^1_{n-1};\boldsymbol{\lambda}^2_{n-1})\boldsymbol{\Lambda}_{n-1}\rangle$. Since the generator $X_{n,n-1} = 1\otimes X^1_{n,n-1} + X^2_{n,n-1}\otimes 1$ commutes with the generators of O$(n-2)$ its matrix elements are independent of the O$(n-3)$ GZ states.

V. THE CALCULATION OF THE OPCS

So far we have seen that the calculation of the Gel'fand invariants can be reduced to the calculation of the operator P_n and that of P_n can be reduced to the calculation of the generator $X_{n,n-1}$ in each recurrence step. By simultaneous diagonalization of the Gel'fand invariants we can construct states which belong to a well defined O(n) irreps. The eigenvectors of these invariants are just the opcs of the SO$(n-1)\uparrow$O(n) induced representation up to a phase factor. The matrix elements of the Gel'fand invariants can be easily evaluated using Eqs. (18) and (20):

$$\begin{aligned}\langle \boldsymbol{\Lambda}^p_{n-1}a_{n-1}|I_k|\boldsymbol{\Lambda}'^p_{n-1}a'_{n-1}\rangle &= \delta_{\boldsymbol{\Lambda}^p_{n-1}\boldsymbol{\Lambda}'^p_{n-1}}\Bigg\{\langle(0_{n-1};\boldsymbol{\lambda}^p_{n-1})\Lambda^p_{n-1}a_{n-1}|(P_n)^k|(0_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\Lambda}^p_{n-1}a'_{n-1}\rangle \\ &+ \sum_{\boldsymbol{\Lambda}^s_{n-2}}\sum_{\Lambda_{n-1}}\left(\begin{array}{cc|c} 1_{n-1} & \lambda^p_{n-1} & \lambda_{n-1} \\ \Lambda^s_{n-1} & \Lambda^p_{n-2} & \Lambda_{n-2}\end{array}\right)^*\left(\begin{array}{cc|c} 1_{n-1} & \lambda^p_{n-1} & \lambda_{n-1} \\ \Lambda^s_{n-1} & \Lambda^p_{n-2} & \Lambda_{n-2}\end{array}\right) \\ &\times\langle(1_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\Lambda}_{n-1}a_{n-1}|(P_n)^k|(1_{n-1};\boldsymbol{\lambda}^p_{n-1})\boldsymbol{\Lambda}_{n-1}a'_{n-1}\rangle\Bigg\}.\end{aligned} \tag{30}$$

In deriving Eqs (30) we have used the fact that the fundamental representation $\boldsymbol{\lambda}_n = 1_n$ contains only the O$(n-1)$ irreps 0_{n-1} and 1_{n-1}. Furthermore, the Clebsch–Gordan coefficient for 0_{n-1} is 1. Since I_k commutes with $X_{i,j}$ it is diagonal with respect to $\boldsymbol{\Lambda}^p_{n-1}$, and we can sum Eq. (30) on $\boldsymbol{\Lambda}^p_{n-2}$ and use the orthogonality properties of the Clebsch–Gordan coefficients to obtain

$$\langle \mathbf{\Lambda}^p_{n-1} a_{n-1} | I_k | \mathbf{\Lambda}^p_{n-1} a'_{n-1} \rangle = \langle (0_{n-1}; \boldsymbol{\lambda}^p_{n-1}) \mathbf{\Lambda}^p_{n-1} a_{n-1} | (P_n)^k | (0_{n-1}; \boldsymbol{\lambda}^p_{n-1}) \mathbf{\Lambda}^p_{n-1} a'_{n-1} \rangle + \frac{1}{|\boldsymbol{\lambda}^p_{n-1}|} \times \sum_{\mathbf{\Lambda}_{n-1}} \langle (1_{n-1}; \boldsymbol{\lambda}^p_{n-1}) \mathbf{\Lambda}_{n-1} a_{n-1} | (P_n)^k | (1_{n-1}; \boldsymbol{\lambda}^p_{n-1}) \mathbf{\Lambda}_{n-1} a'_{n-1} \rangle. \tag{31}$$

Thus, we can see that at each recurrence step, once we know the matrix elements of the generator $X_{n,n-1}$, we can calculate P_n and then take its kth power to get the matrix elements of the Gel'fand invariants. The calculation of $(P_n)^k$ can be done by simply inserting a complete set of states between each power of P_n. After the diagonalization of $I_k\{k=2,4,\ldots,n\}$, we get states that belong to well-defined irreps of O(n).

When n is even we would like to split the O(n) irreps into SO(n) irreps. For this purpose, for each O(n) irrep we discard all the opcs but those originated from the highest weight SO($n-1$) irrep, $\boldsymbol{\lambda}^p_{n-1} = (\lambda^p_{n,1}, \lambda^p_{n,2}, \ldots, \lambda^p_{n,h})$, $(h=[(n-1)/2])$. The generator $X_{n,n-1}$ is then diagonalized in the space of the highest weight SO($n-2$) irrep $\boldsymbol{\lambda}^p_{n-2} = (\lambda^p_{n,1}, \lambda^p_{n,2}, \ldots, \lambda^p_{n,h'})$, $(h'=[(n-2)/2])$. The possible eigenvalues of $X_{n,n-1}$ are $\pm \lambda^p_{n\frac{n}{2}}$. The rest of the opcs will be recalculated using the procedure presented below for ensuring consistent phase of the opcs.

When states with a given SO(n) irrep $\boldsymbol{\lambda}^p_n$ can be obtained from more then one SO($n-1$) irrep $\boldsymbol{\lambda}^p_{n-1}$, our procedure leaves the relative phases of these states undetermined, since the diagonalization is performed for each SO($n-1$) irrep separately and the states $|\boldsymbol{\lambda}^p_{n-1} \mathbf{\Lambda}^p_{n-2} a_{n-1}\rangle$ enter the calculation in pairs. However from the works of Gel'fand and Zetlin[6] or Pang and Hecht[17] such states should be related as

$$\langle \boldsymbol{\lambda}^p_n \mathbf{\Lambda}'^p_{n-1} a_n | X_{n,n-1} | \boldsymbol{\lambda}^p_n \mathbf{\Lambda}^p_{n-1} a_n \rangle = \delta_{\mathbf{\Lambda}'_{n-2}, \mathbf{\Lambda}_{n-2}} F_n(\boldsymbol{\lambda}^p_n, \boldsymbol{\lambda}'^p_{n-1}, \boldsymbol{\lambda}^p_{n-1}, \boldsymbol{\lambda}^p_{n-2}), \tag{32}$$

where F_n, given in Appendix B, is zero unless $\boldsymbol{\lambda}'_{n-1} = \boldsymbol{\lambda}_{n-1}$ or $\lambda'_{n-1,i} = \lambda_{n-1,i} \pm \delta_{i,j}$. By expanding both sides of Eq. (32) in terms of the opcs and using the orthogonality of the opcs, Eq. (4), we obtain

$$C^{(n)\boldsymbol{\lambda}'^p_{n-1} \mathbf{\Lambda}^p_{n-2}}_{\boldsymbol{\lambda}^p_n a_n, a'_{n-1}} = \frac{1}{F_n(\boldsymbol{\lambda}^p_n, \boldsymbol{\lambda}'^p_{n-1}, \boldsymbol{\lambda}^p_{n-1}, \boldsymbol{\lambda}^p_{n-2})} \times \sum_{a_{n-1}} \langle \boldsymbol{\lambda}'^p_{n-1} \mathbf{\Lambda}^p_{n-2} a'_{n-1} | X_{n,n-1} | \boldsymbol{\lambda}^p_{n-1} \mathbf{\Lambda}^p_{n-2} a_{n-1} \rangle C^{(n)\boldsymbol{\lambda}^p_{n-1} \mathbf{\Lambda}^p_{n-2}}_{\boldsymbol{\lambda}^p_n a_n, a_{n-1}}. \tag{33}$$

This relation yields the orthogonal parentage coefficients $\boldsymbol{\lambda}'^p_{n-1} \to \boldsymbol{\lambda}^p_n$ in terms of the $\boldsymbol{\lambda}^p_{n-1} \to \boldsymbol{\lambda}^p_n$ orthogonal parentage coefficients. In conclusion, whenever the SO(n) irrep $\boldsymbol{\lambda}^p_n$ originates from more than one SO($n-1$) irrep we keep after diagonalization only the set of orthogonal parentage coefficients that originate from one particular irrep $\boldsymbol{\lambda}^p_{n-1}$ of SO($n-1$). The orthogonal parentage coefficients that originate from other SO($n-1$) irreps are constructed from the relation (33).

VI. THE COMPUTATIONAL ALGORITHM

In the previous sections we described the recursive method for the construction of O(n) GZ states by diagonalization of the Casimir operators I_k. Starting from basis states that belong to well-defined irreps of SO($n-1$) we may adopt the following procedure to carry out our algorithm systematically:

(i) Pick an irrep $\tilde{\boldsymbol{\lambda}}^p_{n-1}$ of SO(n), consider all the possible irreps $\boldsymbol{\lambda}_{n-1}$ that can be obtained from $\tilde{\boldsymbol{\lambda}}^p_{n-1}$ by the outer product $1_{n-1}\otimes\tilde{\boldsymbol{\lambda}}^p_{n-1}$. The possible $\boldsymbol{\lambda}_{n-1}$ irreps are $\lambda_{n-1,j}=\lambda^p_{n-1,j}$ and the irreps given by $\lambda_{n-1,j}=\lambda^p_{n-1,j}\pm\delta_{ij}$ that obey the GZ restrictions $\lambda_{n-1,j}\geqslant|\lambda_{n-1,j+1}|$.

(ii) For every irrep $\boldsymbol{\lambda}_{n-1}$ construct all the states (20) with $\boldsymbol{\lambda}^s_{n-1}=1_{n-1}$ or 0_{n-1} and the appropriate irreps $\boldsymbol{\lambda}^p_{n-1}$ of $SO(n-1)$.

(iii) Use Eq. (27) to evaluate the matrix elements of the operator P_n for each invariant subspace defined by $\boldsymbol{\lambda}_{n-1}$.

(iv) Calculate the matrix elements of the Gel'fand invariants $I_k, k=2,4,\ldots,n$ by taking the kth power of P_n, using Eq. (31).

(v) Evaluate the $SO(n-1)\uparrow O(n)$ opcs by diagonalizing of the Gel'fand invariants, $I_k, k=2,4,\ldots$, one by one until all the $O(n)$ irreps are completely determined.

(vi) For each $O(n)$ irrep $\boldsymbol{\lambda}^p_n$ keep the opcs for the highest weight $SO(n-1)$ irrep in $\boldsymbol{\lambda}^p_n$, i.e., $\boldsymbol{\lambda}^p_{n-1}=(\lambda^p_{n,1},\lambda^p_{n,2},\ldots,\lambda^p_{n,h})$, $(h=[(n-1)/2])$, and discard all the rest.

(vii) If n is even, split each $O(n)$ irrep into the appropriate $SO(n)$ irreps: for each $O(n)$ irrep diagonalize the generator $X_{n,n-1}$ in the space of highest weight states, taking $\boldsymbol{\lambda}^p_{n-2}=(\lambda^p_{n,1},\lambda^p_{n,2},\ldots,\lambda^p_{n,h'})$, $(h'=[(n-2)/2])$.

(viii) Regenerate the discarded opcs, with consistent phases, using Eq. (33) successively.

(ix) Pick another $SO(n-1)$ irrep and repeat the process. Continue for all the $SO(n-1)$ irreps.

It should be noted that if the invariant subspace is not too large, it is much more efficient to store the calculated matrix elements of P_n instead of regenerating them for each irrep $\tilde{\boldsymbol{\lambda}}^p_{n-1}$ of $SO(n)$.

VII. CONCLUSIONS

A new method for the construction of $O(n)$ Gel'fand–Zetlin states has been developed. The states are evaluated in terms of the given basis functions using the orthogonal parentage coefficients, opcs. The orthogonal parentage coefficients are calculated by diagonalization of the Gel'fand invariants. A crucial point is that it is sufficient to evaluate the matrix elements of the generator $X_{n,n-1}$ at each step. The construction of $O(n)$ Gel'fand Zetlin states is important primarily for systems with approximate $O(n)$ symmetry such as the general many-body theory where collective states are restricted to definite irrep of $O(n)$, or in the construction of states with definite permutational symmetry as the symmetry group S_n is a subgroup of $O(n-1)$. The recursive method described in this paper for the orthogonal group is closely related to the method developed few years ago by Novoselsky, Katriel, and Gilmore[1,3,4] for the calculation of the symmetry group cfps. This method may be applied to the harmonic oscillator or hyperspherical harmonic bases as well as for the calculation of outer product isoscalar factors of the orthogonal group.

ACKNOWLEDGMENTS

I wish to thank Professor Victor Mandelzweig for his comments and to Dr. Akiva Novoselsky for many inspiring discussions during this work.

APPENDIX A: CLEBSCH GORDON COEFFICIENTS

The $O(n)$ CG coefficients needed in Eq. (27) are the generalized CG coefficients for the Kronecker product $|1_n[0_{n-1}]\rangle\otimes|\boldsymbol{\lambda}_n\boldsymbol{\Lambda}_{n-1}\rangle$. This coefficients where explicitly evaluated by Pang and Hecht.[17] It is evident that these coefficients do not depend on the value of $\boldsymbol{\Lambda}_{n-2}$, thus we can write:

$$\left(\begin{matrix} 1_n & \boldsymbol{\lambda}'_n & \boldsymbol{\lambda}_n \\ [0_{n-1}] & \boldsymbol{\Lambda}_{n-1} & \boldsymbol{\Lambda}_{n-1} \end{matrix}\right)=\left(\begin{matrix} 1_n & \boldsymbol{\lambda}'_n & \boldsymbol{\lambda}_n \\ [0_{n-1}] & \boldsymbol{\lambda}_{n-1} & \boldsymbol{\lambda}_{n-1} \end{matrix}\right). \tag{A1}$$

The nonzero CG coefficients are given below by

$$\begin{pmatrix} 1_n & (\lambda_{n,1},\ldots,\lambda_{n,j}+1,\ldots,\lambda_{n,h},\lambda_{n,h+1}) & (\lambda_{n,1},\ldots,\lambda_{n,j},\ldots,\lambda_{n,h},\lambda_{n,h+1}) \\ [0_{n-1}] & (\lambda_{n-1,1},\ldots,\lambda_{n-1,j},\ldots,\lambda_{n-1,h}) & (\lambda_{n-1,1},\ldots,\lambda_{n-1,j},\ldots,\lambda_{n-1,h}) \end{pmatrix}$$

$$=\left|\prod_{i=1}^{h}\frac{(\lambda_{n,j}-\lambda_{n-1,i}+i-j+1)(\lambda_{n,j}+\lambda_{n-1,i}+n-i-j)}{(\lambda_{n,j}-\lambda_{n,i}+i-j+1)(\lambda_{n,j}+\lambda_{n,i}+n-i-j)}\right|^{\frac{1}{2}}, \tag{A2}$$

$$\left(\begin{matrix} 1_n & (\lambda_{n,1},\ldots,\lambda_{n,j},\ldots,\lambda_{n,h}) \\ [0_{n-1}] & (\lambda_{n-1,1},\ldots,\lambda_{n-1,j},\ldots,\lambda_{n-1,h}) \end{matrix}\middle| \begin{matrix} (\lambda_{n,1},\ldots,\lambda_{n,j},\ldots,\lambda_{n,h}) \\ (\lambda_{n-1,1},\ldots,\lambda_{n-1,j},\ldots,\lambda_{n-1,h}) \end{matrix}\right)$$

$$=(2h+2-n)\prod_{i=1}^{h}\frac{(\lambda_{n-1,i}+h-i)}{(\lambda_{n,i}+h-i)}. \tag{A3}$$

Here $h=[(n-1)/2]$, and for odd n $(n=2h+1)$ $\lambda_{n,h+1}=0$ in Eq. (A2). The expression for the CG coefficients for $\lambda_{n,j}-1$ are the same, replacing $\lambda_{n,j}$ by $\lambda_{n,j}-1$ on the rhs of Eq. (A2).

APPENDIX B: MATRIX ELEMENTS OF $X_{n,n-1}$

The matrix elements of the generator $X_{n,n-1}$ between GZ states have been given at first by Gel'fand and Zetlin[6] and were rederived later by Pang and Hecht.[17] For the sake of brevity we will denote the matrix elements of $X_{n,n-1}$ by F_n:

$$\langle \boldsymbol{\lambda}_n \boldsymbol{\Lambda}'_{n-1}|X_{n,n-1}|\boldsymbol{\lambda}_n\boldsymbol{\Lambda}_{n-1}\rangle=\delta_{\boldsymbol{\Lambda}'_{n-2},\boldsymbol{\Lambda}_{n-2}}F_n(\boldsymbol{\lambda}_n,\boldsymbol{\lambda}'_{n-1},\boldsymbol{\lambda}_{n-1},\boldsymbol{\lambda}_{n-2}). \tag{B1}$$

Here F_n is zero unless $\boldsymbol{\lambda}'_{n-1}=\boldsymbol{\lambda}_{n-1}$ or $\lambda'_{n-1,i}=\lambda_{n-1,i}\pm\delta_{i,j}$. Following the notation of Pang and Hecht[17] we shall define $t_{n,i}=\lambda_{n,i}+[(n+1)/2]-i$ and obtain for odd n $(=2k+1)$ the following expressions:

$$F_n(\boldsymbol{\lambda}_n,\boldsymbol{\lambda}'_{n-1}+\delta_{i,j},\boldsymbol{\lambda}_{n-1},\boldsymbol{\lambda}_{n-2})$$
$$=\frac{1}{2}\left|\frac{\Pi_{\alpha=1}^{k-1}(t_{2k-1,\alpha}-t_{2k,j}-1)(t_{2k-1,\alpha}+t_{2k,j})\Pi_{\beta=1}^{k}(t_{2k+1,\beta}-t_{2k,j}-1)(t_{2k+1,\beta}+t_{2k,j})}{\Pi_{\alpha\neq j}^{k}(t_{2k,\alpha}^2-t_{2k,j}^2)(t_{2k,\alpha}^2-(t_{2k,j}+1)^2)}\right|^{1/2}. \tag{B2}$$

For even n $(=2k)$ the diagonal matrix element is given by

$$F_n(\boldsymbol{\lambda}_n,\boldsymbol{\lambda}_{n-1},\boldsymbol{\lambda}_{n-1},\boldsymbol{\lambda}_{n-2})=i\,\frac{\Pi_{\alpha=1}^{k-1}t_{2k-2,\alpha}\Pi_{\beta=1}^{k}t_{2k,\beta}}{\Pi_{\alpha=1}^{k-1}t_{2k-1,\alpha}(t_{2k-1,\alpha}-1)} \tag{B3}$$

and the nondiagonal matrix element is

$$F_n(\boldsymbol{\lambda}_n,\boldsymbol{\lambda}_{n-1}+\delta_{i,j},\boldsymbol{\lambda}_{n-1},\boldsymbol{\lambda}_{n-2})$$
$$=\left|\frac{\Pi_{\alpha=1}^{k-1}(t_{2k-2,\alpha}^2-t_{2k-1,j}^2)\Pi_{\beta=1}^{k}(t_{2k,\beta}^2-t_{2k-1,j}^2)}{t_{2k-1,j}^2(4t_{2k-1,j}^2-1)\Pi_{\alpha\neq j}^{k-1}(t_{2k-1,\alpha}^2-t_{2k-1,j}^2)[(t_{2k-1,\alpha}-1)^2-t_{2k-1,j}^2]}\right|^{1/2}. \tag{B4}$$

Note that for $k=1$ there are undefined products in Eqs. (B2)–(B4) (for example, $\Pi_{\alpha=1}^{k-1}$). These products are equal to 1 in this case.

[1] A. Novoselsky, J. Katriel, and R. Gilmore, J. Math. Phys. **29**, 1368 (1988).
[2] J. Q. Chen, A. Novoselsky, M. Vallieres, and R. Gilmore, Phys. Rev. C **39**, 1088 (1989).
[3] A. Novoselsky and J. Katriel, Ann. Phys. (N.Y.) **196**, 135 (1989).
[4] A. Novoselsky and J. Katriel, Phys. Rev. A **49**, 833 (1994).
[5] N. Barnea and A. Novoselsky, Ann. Phys. (N.Y.) **256**, 192 (1997).

[6] I. M. Gel'fand and M. L. Zetlin, Dokl. Akad. Nauk SSSR **71**, 825 (1950).
[7] I. M. Gel'fand, R. A. Minols, and Z. Ya. Shapiro, *Representations of the Rotation and Lorentz Groups and Their Application* (Pergamon, New York, 1963), p. 353.
[8] D. J. Rowe and G. Rosensteel, Ann. Phys. (N.Y.) **126**, 198 (1980).
[9] G. F. Fillipov, Sov. J. Part. Nucl. **4**, 405 (1974).
[10] M. Moshinsky, J. Math. Phys. **25**, 1555 (1984).
[11] M. Moshinsky and C. Quesne, J. Math. Phys. **12**, 1772 (1971).
[12] E. L. Surkov, Yad. Fiz. **5**, 908 (1967) [Sov. J. Nucl. Phys. **5**, 644 (1967)].
[13] E. Chacon, P. Hess, and M. Moshinsky, J. Math. Phys. **25**, 1565 (1984).
[14] O. Castanos, E. Chacon, and M. Moshinsky, J. Math. Phys. **25**, 2815 (1984).
[15] S. A. Edwards, J. Math. Phys. **19**, 164 (1978).
[16] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (PWN—Polish Scientific, Warszawa, 1980).
[17] S. C. Pang and K. T. Hecht, J. Math. Phys. **29**, 1233 (1967).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Deformations of global symmetries in the extended antifield formalism

Friedemann Brandt[a)]
Institut für Theoretische Physik, Universität Hannover, Appelstraße 2, D-30167 Hannover, Germany

(Received 11 June 1998; accepted for publication 11 November 1998)

It is outlined how deformations of field theoretical rigid symmetries can be constructed and classified by cohomological means in the extended antifield formalism. Special attention is devoted to deformations referring only to a subset of the rigid symmetries of a given model and leading to a nontrivial extension of the graded Lie algebra associated with that subset. The method is illustrated for a $D=4$, $N=2$ supersymmetric model where the central extension of the supersymmetry algebra emerges via a deformation. Deformations of gauge fixed actions with a BRST symmetry are discussed too and illustrated by the Curci–Ferrari model. © *1999 American Institute of Physics.* [S0022-2488(99)02802-9]

I. INTRODUCTION

A problem often met in field theory is to what degree a given action functional can be nontrivially deformed while keeping some of its symmetries. A particularly interesting issue is whether the symmetry transformations themselves can be deformed in a nontrivial way, i.e., whether there are simultaneous deformations of the action and its symmetries.

Deformations of this sort can be studied systematically by cohomological methods in the spirit of Gerstenhaber's approach to deformation theory.[1] This was first described in Ref. 2 (see also Refs. 3 and 4) for gauge symmetries in the framework of the standard antifield formalism.[5–7] The inclusion of rigid (=global) symmetries was roughly sketched more recently in Ref. 8 within an extended antifield formalism. The aim of this work is to develop the latter approach more thoroughly, with special attention to deformations which are required to maintain only (a deformed version of) a *subset* of the rigid symmetries of a given model.

The restriction to a subset of the rigid symmetries is a typical situation, as often it is neither possible nor desirable to keep all the rigid symmetries when deforming a field theory because that may constrain the sought deformations too much. We shall thus base the deformation theory on an extended antifield formalism which involves only a ''closed'' subset of rigid symmetries. The ''closure'' of the subset requires that the graded commutator algebra of the rigid symmetries under study closes in the soft (field theoretical) sense, i.e., up to gauge transformations and on-shell trivial symmetries. (In order to set up the extended antifield formalism, it may be necessary to include also ''symmetries of higher order.''[8,9]) In other words, a closed subset of rigid symmetries forms a subalgebra (in the soft sense) of the graded commutator algebra of all the rigid symmetries.

When one applies the extended antifield formalism to study deformations of such a subset of rigid symmetries, one may encounter a ''subtlety.'' Namely, a deformation may turn a subset of rigid symmetries which is closed in the soft sense into an open one. That is, it can happen that the deformed commutator algebra involves symmetries which did not occur in the undeformed one. These additional symmetries are not ''new'' ones which are introduced through the deformation. Rather, they are present already in the original (undeformed) model. The subtlety is that usually it is not clear from the outset which additional symmetries of the original model can show up in the

[a)]Electronic mail: brandt@itp.uni-hannover.de

deformed commutator algebra. In particular, this may depend on the deformation itself.

Hence, the property of a subset of rigid symmetries to be a closed one is not necessarily preserved by deformations. This is actually an interesting phenomenon as it is related to extensions of the (graded) Lie algebra associated with the commutator algebra of the subset of rigid symmetries under study. Important examples are central extensions of extended supersymmetry algebras.[10] As an illustration, we shall discuss a simple four-dimensional $N=2$ supersymmetric model for a hypermultiplet[11,12] where the central extension of the supersymmetry algebra arises indeed via a deformation of the model.

The antifield formalism serves in this context as a tool that allows one to formulate the deformation theory conveniently in cohomological terms. Ghost fields are not dynamical in this approach (in particular, they are not paired with antighost fields), in contrast to their counterparts in the quantum field theoretical context. Nevertheless, the formalism applies also to gauge fixed action functionals which contain dynamical ghost and antighost fields. This application just requires a slight change of the point of view as compared to the one familiar from quantum field theory. Namely, the gauge fixed action simply takes the role of a classical action. Accordingly, the dynamical ghost and antighost fields occurring in the gauge fixed action are counted among the classical fields, and the Becchi–Rouet–Stora–Tyutin (BRST) symmetry of the gauge fixed action counts among the rigid symmetries. In particular this allows one to investigate deformations of the BRST symmetry after fixing the gauge. We shall discuss and illustrate this particular application in some detail in the Curci–Ferrari model.[13–15]

The paper has been organized as follows. Section II summarizes basic properties of global and local symmetries in Lagrangian field theory which are used later on. Then the extended antifield formalism and the construction and properties of the extended BRST differential are briefly reviewed in Secs. III and IV. The systematic approach to the deformation problem is described in Sec. V. Sections VI and VII contain the examples mentioned above, i.e., the hypermultiplet of $N=2$ supersymmetry and the Curci–Ferrari model. The paper is ended with some concluding remarks in Sec. VIII.

II. GLOBAL AND LOCAL SYMMETRIES

We shall first briefly summarize the definition and some properties of continuous rigid and gauge symmetries in Lagrangian field theories, following the presumably most popular approach based on the action (alternatively one can define rigid and gauge symmetries on the level of the field equations, via conserved currents and Noether identities respectively). We shall thus consider Lagrangian field theories which derive from an action functional for a set of fields $\phi^i(x)$,

$$S_{\text{class}}[\phi]=\int d^nxL(x,[\phi]), \tag{2.1}$$

where $L(x,[\phi])$ is a Lagrangian constructed of the fields and their partial derivatives. {Here and in the following, $[\phi]$ denotes collectively dependence on the fields and on their derivatives. In more precise mathematical terms, $\phi^i,\partial_\mu\phi^i,\partial_\mu\partial_\nu\phi^i,\ldots$ are to be understood as local coordinates of a jet space, and $\phi^i(x)$ as sections of the jet bundle over an n-dimensional base manifold (''space–time'') with local coordinates x^μ $(\mu=1,\ldots,n)$. The arguments of $L(x,[\phi])$ indicate that the Lagrangian may (but, of course, need not) depend explicitly on the x^μ.} The field equations (equations of motion) derive via the variational principle from S_{class}, i.e., they are the corresponding Euler–Lagrange equations.

A continuous rigid symmetry of an action (2.1) is generated by transformations of the fields with a constant infinitesimal parameter ε,

$$\phi^i\rightarrow\tilde{\phi}^i=\phi^i+\varepsilon\, G^i(x,[\phi]),\quad \varepsilon=\text{const}, \tag{2.2}$$

such that $L(x,[\tilde{\phi}])$ differs from $L(x,[\phi])$ to first order in ε at most by a total derivative,

$$L(x,[\tilde{\phi}])=L(x,[\phi])+\varepsilon\partial_\mu k^\mu(x,[\phi])+O(\varepsilon^2). \tag{2.3}$$

A gauge symmetry of an action (2.1) is defined similarly, with the important difference that it involves, instead of a constant parameter, an additional field $\lambda=\lambda(x)$ (i.e., a field which does not occur in the Lagrangian). It is generated by infinitesimal transformations of the form

$$\phi^i\to\tilde{\phi}^i=\phi^i+\sum_{k\geq 0} r^{i\mu_1\cdots\mu_k}(x,[\phi])\partial_{\mu_1}\cdots\partial_{\mu_k}\lambda \tag{2.4}$$

such that $L(x,[\tilde{\phi}])$ and $L(x,[\phi])$ differ to first order in λ at most by a total derivative,

$$L(x,[\tilde{\phi}])=L(x,[\phi])+\partial_\mu h^\mu(x,[\phi,\lambda])+O(\lambda^2). \tag{2.5}$$

This invariance condition must hold for an unconstrained field λ, i.e., it must neither impose a differential equation for λ, nor determine λ in terms of the fields ϕ^i and their derivatives [otherwise λ would turn into a function of the x^μ, ϕ^i and their derivatives and thus (2.4) would reduce to a rigid symmetry of the form (2.2)].

Now, the above standard definitions do not yet characterize symmetries satisfactorily for our purpose. An important ingredient, underplayed in many textbooks, is still missing: the distinction between trivial and nontrivial symmetries. For instance, consider transformations (2.2) and (2.4) with

$$G^i(x,[\phi])=E^{ij}(x,[\phi])\frac{\hat{\partial}L}{\hat{\partial}\phi^j}, \tag{2.6}$$

$$\sum_{k\geq 0} r^{i\mu_1\cdots\mu_k}(x,[\phi])\partial_{\mu_1}\cdots\partial_{\mu_k}\lambda=E^{ij}(x,[\phi,\lambda])\frac{\hat{\partial}L}{\hat{\partial}\phi^j}, \tag{2.7}$$

where $\hat{\partial}L/\hat{\partial}\phi^i$ is the Euler–Lagrange derivative of the Lagrangian with respect to ϕ^i,

$$\frac{\hat{\partial}L}{\hat{\partial}\phi^i}=\frac{\partial L}{\partial\phi^i}-\partial_\mu\frac{\partial L}{\partial(\partial_\mu\phi^i)}+\cdots, \tag{2.8}$$

and E^{ij} are any functions which are only required to be graded antisymmetric in their indices,

$$E^{ij}=(-)^{\epsilon_i\epsilon_j+1}E^{ji}, \tag{2.9}$$

where ϵ_i is the Grassmann parity of ϕ^i. It is easily verified that (2.6) and (2.7) give rigid and gauge symmetries, satisfying (2.3) and (2.5), respectively, for any choice of E^{ij} fulfilling (2.9). Such symmetries are examples of trivial symmetries which may be called ''on-shell trivial symmetries'' (the terminology reflects that the symmetry transformations vanish for every solution of the field equations, as the latter read $\hat{\partial}L/\hat{\partial}\phi^i=0$). More general trivial symmetries of this type are obtained from (2.6) and (2.7) when E^{ij} are differential operators of the form $E^{ij}=\sum e^{ij\,\mu_1\cdots\mu_k}\partial_{\mu_1}\cdots\partial_{\mu_k}$ with properties generalizing (2.9) appropriately.

In addition to on-shell trivial symmetries, there is a second type of trivial rigid symmetry whenever the action possesses a true gauge symmetry, i.e., a gauge symmetry which is not on-shell trivial. Indeed, in that case the action has automatically infinitely many further rigid symmetries which are to be considered as trivial too, even though they are not on-shell trivial. These additional trivial rigid symmetries arise from nontrivial gauge transformations (2.4) by replacing λ with $\varepsilon\, f(x,[\phi])$ there, where $f(x,[\phi])$ is any function of the fields and their derivatives. Indeed, as (2.5) holds for any λ, such a replacement results in a transformation (2.2) satisfying (2.3) with

$$G^i(x,[\phi])=\sum_{k\geqslant 0} r^{i\,\mu_1\cdots\mu_k}(x,[\phi])\,\partial_{\mu_1}\cdots\partial_{\mu_k}f(x,[\phi]). \tag{2.10}$$

Hence, every action has infinitely many trivial gauge and rigid symmetries. Gauge and rigid symmetries are therefore best defined as *equivalence classes* where two symmetries are called equivalent when they differ by a trivial symmetry (and by irrelevant redefinitions of the respective ε and λ, i.e., by multiplications of ε and λ with arbitrary constants and field dependent functions, respectively). [Clearly, two symmetries differing only through such redefinitions are to be identified, as (2.3) and (2.5) must hold for arbitrary constant parameters ε and unconstrained fields λ.] One can then introduce the concept of a basis of symmetries, containing one representative of each nontrivial equivalence class. We shall characterize such bases for the gauge and rigid symmetries through operations $\{\delta_\alpha\}$ and $\{\Delta_a\}$, respectively, which are related to symmetry transformations (2.2) and (2.4) according to

$$\delta_\alpha\phi^i=R^i_\alpha(x,[\phi])\equiv\sum_{k\geqslant 0} r_\alpha^{i\,\mu_1\cdots\mu_k}(x,[\phi])\,\partial_{\mu_1}\cdots\partial_{\mu_k}, \tag{2.11}$$

$$\Delta_a\phi^i=G^i_a(x,[\phi]). \tag{2.12}$$

As the graded commutator of two infinitesimal symmetry transformations is automatically again an infinitesimal symmetry transformation (due to the derivation property of infinitesimal transformations), there is always a graded commutator algebra associated with such bases. However, due to the presence of trivial symmetries, this graded commutator algebra is in general a quotient algebra because, in general, the graded commutator of two elements of the basis can be expressed in terms of elements of the same basis only up to trivial symmetries. In particular, the general form of the graded commutator of any two elements of a basis of infinitesimal rigid symmetry transformations is thus

$$[\Delta_a,\Delta_b]\phi^i=f_{ab}^c\Delta_c\phi^i+R^i_\alpha f^\alpha_{ab}(x,[\phi])+E^{ij}_{ab}(x,[\phi])\,\frac{\hat\partial L}{\hat\partial\phi^j}, \tag{2.13}$$

where the graded commutator of two objects A and B is defined by means of their Grassmann parities $\epsilon(A)$ and $\epsilon(B)$ through

$$[A,B]=AB-(-)^{\epsilon(A)\epsilon(B)}BA. \tag{2.14}$$

In (2.13), f_{ab}^c are constant coefficients which are the structure constants of a graded Lie algebra (as a consequence of $[[\Delta_a,\Delta_b],\Delta_c]+\text{cyclic}=0)$, while $f^\alpha_{ab}(x,[\phi])$ and $E^{ij}_{ab}(x,[\phi])$ are, in general, field-dependent functions and operators appearing in trivial rigid symmetries as described above, cf. (2.10) and (2.6).

III. EXTENDED ANTIFIELD FORMALISM

We shall now recall the basic features of the extended antifield formalism and fix our notation and conventions. For simplicity, we shall concentrate on the case that the gauge transformations (if any) are irreducible and that only ordinary rigid symmetries are present or needed, but no rigid symmetries of higher order in the terminology of Ref. 8. The general case is a straightforward extension of this one. As mentioned already, the extended antifield formalism can be established for any closed subset of rigid symmetries.[8] When higher-order rigid symmetries are absent, a closed subset is simply a subset $\{\Delta_{\underline{a}}\}$ of a basis $\{\Delta_a\}$ of nontrivial rigid symmetries such that, in the notation of the previous section,

$$\{\Delta_a\}=\{\Delta_{\underline{a}},\Delta_{\hat{a}}\},\quad f_{\underline{ab}}^{\hat{c}}=0\quad \forall \underline{a},\underline{b},\hat{c}. \tag{3.1}$$

Here $f_{\underline{ab}}^{\hat{c}}=0$ requires that the graded commutator algebra of the $\Delta_{\underline{a}}$ is a subalgebra of (2.13) in the "soft" sense, i.e., with respect to the quotient structure "modulo trivial rigid symmetries."

The fields and antifields of the standard antifield formalism are denoted by Φ^A and Φ_A^* where $\{\Phi^A\}$ contains the "classical" fields ϕ^i, i.e., the fields occurring in the "classical" action (2.1) under study, and the ghost fields C^α corresponding to the nontrivial gauge symmetries of this action. (As mentioned in the Introduction, the "classical" action may be actually a gauged fixed one. Ghost and antighost fields occurring in such an action count among the ϕ^i and must not be confused with the C^α. See Sec. VII for an example.) The extended antifield formalism, restricted to the subset $\{\Delta_{\underline{a}}\}$, contains in addition constant ("global") ghosts $\xi^{\underline{a}}$ for each $\Delta_{\underline{a}}$. These global ghosts have ghost number 1 and Grassmann parity opposite to the corresponding rigid symmetries. It is also very convenient (though not necessary in principle) to accompany each $\xi^{\underline{a}}$ with a constant antifield $\xi_{\underline{a}}^*$. The latter has ghost number (-2) and Grassmann parity opposite to $\xi^{\underline{a}}$. In particular, this allows one to set up the extended antifield formalism through an extended master equation of the form

$$(S,S)=0, \tag{3.2}$$

where $(\,,\,)$ is an extended antibracket defined by

$$(X,Y)=\frac{\partial^R X}{\partial \xi^{\underline{a}}}\frac{\partial^L Y}{\partial \xi_{\underline{a}}^*}-\frac{\partial^R X}{\partial \xi_{\underline{a}}^*}\frac{\partial^L Y}{\partial \xi^{\underline{a}}}+\int d^nx\left[\frac{\delta^R X}{\delta\Phi^A(x)}\frac{\delta^L Y}{\delta\Phi_A^*(x)}-\frac{\delta^R X}{\delta\Phi_A^*(x)}\frac{\delta^L Y}{\delta\Phi^A(x)}\right]. \tag{3.3}$$

Here superscripts R and L indicate right and left derivatives, respectively. The extended antibracket is defined in the space of local functionals of the form

$$\Gamma[\Phi,\Phi^*,\xi]+M^{\underline{a}}(\xi)\xi_{\underline{a}}^*, \tag{3.4}$$

where $\Gamma[\Phi,\Phi^*,\xi]$ is the space–time integral of a local function of the fields and antifields which may depend on the global ghosts but not on the global antifields, and $M^{\underline{a}}(\xi)$ is a polynomial in the global ghosts [note: $M^{\underline{a}}(\xi)\xi_{\underline{a}}^*$ does not involve a space–time integration]. The solution S of the extended master equation is a functional with ghost number 0 of the form (3.4). It contains the classical action, and encodes its gauge symmetries and the subset $\{\Delta_{\underline{a}}\}$ of its rigid symmetries, as well as the graded commutator algebra of these symmetries. In addition one often imposes that S be real. One then needs consistent conventions for complex conjugation. We denote complex conjugation by a bar, and use the convention (familiar from supersymmetry, see, e.g., Ref. 16) that complex conjugation of products involves a sign factor depending on the Grassmann parities,

$$\overline{(XY)}=(-)^{\epsilon_X\epsilon_Y}\overline{X}\,\overline{Y}. \tag{3.5}$$

The complex conjugate of an antifield Φ^* equals minus the antifield of the complex conjugate of Φ (independently of the Grassmann parity of Φ),

$$\overline{(\Phi^*)}=-(\overline{\Phi})^*\quad \forall \Phi\in\{\Phi^A,\xi^{\underline{a}}\}. \tag{3.6}$$

For instance, with these conventions, the antifield of a real field is purely imaginary.

To describe and compute S, it is useful to expand it in the antifield number (agh). The latter vanishes for the fields, and equals minus the ghost number for the antifields,

$$\begin{aligned}&\text{agh }\phi^i=\text{agh }C^\alpha=\text{agh }\xi^{\underline{a}}=0,\\ &\text{agh }\phi_i^*=1,\quad \text{agh }C_\alpha^*=\text{agh }\xi_{\underline{a}}^*=2.\end{aligned} \tag{3.7}$$

The expansion of S is denoted by

$$S=\sum_{k\geqslant 0} S_k, \quad \text{agh}\, S_k=k. \tag{3.8}$$

Here S_0 is the classical action,

$$S_0=S_{\text{class}}[\phi]. \tag{3.9}$$

S_1 encodes both the gauge transformations and the subset of the rigid symmetries under study,

$$S_1=-\int d^nx(R_\alpha^i C^\alpha+\xi^{\underline{a}}\Delta_{\underline{a}}\phi^i)\phi_i^*, \tag{3.10}$$

where we used the notation of the previous section. S_2 encodes the graded commutator algebra of the gauge symmetries and the subset of rigid symmetries under study, and thus, in particular, the subalgebra of (2.13) referring to $\{\Delta_{\underline{a}}\}$,

$$S_2=\frac{1}{2}\xi^{\underline{b}}\xi^{\underline{a}}\tilde{f}_{\underline{ab}}^{\underline{c}}\xi_{\underline{c}}^*+\int d^nx\, \xi^{\underline{b}}\xi^{\underline{a}}\left(\tfrac{1}{2}\tilde{f}_{\underline{ab}}^{\alpha}C_\alpha^*+\tfrac{1}{4}\phi_i^*\tilde{E}_{\underline{ab}}^{ij}\phi_j^*+\cdots\right), \tag{3.11}$$

where $\tilde{f}_{\underline{ab}}^{\underline{c}}$, $\tilde{f}_{\underline{ab}}^{\alpha}$, $\tilde{E}_{\underline{ab}}^{ij}$ coincide with $f_{\underline{ab}}^{\underline{c}}$, $f_{\underline{ab}}^{\alpha}$, $E_{\underline{ab}}^{ij}$ in (2.13) up to signs which follow from the formulas [e.g., $\tilde{f}_{\underline{ab}}^{\underline{c}}=(-)^{\epsilon_{\underline{b}}+1}f_{\underline{ab}}^{\underline{c}}$ where $\epsilon_{\underline{b}}$ is the Grassmann parity of $\Delta_{\underline{b}}$]. The nonwritten terms in (3.11) encode analogously the graded commutator algebra of the gauge symmetries, and of the gauge symmetries with the $\Delta_{\underline{a}}$. Higher terms S_k $(k>2)$ in the expansion of S reflect consistency relations following from the graded commutator algebra. The solution of the extended master equation encodes thus the complete algebraic structure of the gauge and rigid symmetries under study. In particular, the piece in $(S,S)=0$ which is linear in ξ^* yields

$$f_{[\underline{ab}}^{\underline{e}}f_{\underline{c}]\underline{e}}^{\underline{d}}=0, \tag{3.12}$$

where $[\cdots]$ indicates graded antisymmetrization. Equation (3.12) is the Jacobi identity for the structure constants of a graded Lie algebra and reflects again that the commutator algebra of the $\Delta_{\underline{a}}$ constitutes a subalgebra of (2.13) in the soft sense. Of course, in general this commutator algebra is not a true graded Lie algebra, but still a graded Lie algebra in the soft sense.

IV. EXTENDED BRST AND KOSZUL–TATE DIFFERENTIAL

The extended antifield formalism outlined in the previous section implies the existence of a nilpotent antiderivation which generalizes the standard BRST differential so as to incorporate rigid symmetries. We call this antiderivation extended BRST differential and denote it by s. It is defined in the space of local functionals of the form (3.4) via the extended antibracket through

$$sX=(S,X). \tag{4.1}$$

With this definition, s squares to zero (=is ‘‘nilpotent’’),

$$s(XY)=(sX)Y+(-)^{\epsilon_X}X(sY), \quad s^2=0. \tag{4.2}$$

Furthermore, s is a real differential if S is a real functional. As s is Grassmann odd, this means, due to (3.5),

$$\overline{(sX)} = (-)^{\epsilon_X} s\overline{X}. \tag{4.3}$$

It is useful to expand s in the antifield number. The structure of S implies that the expansion of s starts with a piece δ that has antifield number -1 (i.e., δ lowers the antifield number by one unit),

$$s = \delta + \gamma + \sum_{i \geqslant 1} s_i, \quad \text{agh}\,\delta = -1,, \quad \text{agh}\,\gamma = 0, \quad \text{agh}\,s_i = i. \tag{4.4}$$

The nilpotency of s implies anticommutation relations between the pieces in this decomposition,

$$\delta^2 = 0, \quad [\delta,\gamma] = 0, \quad \gamma^2 + [\delta, s_1] = 0, \ldots . \tag{4.5}$$

Here δ is the extension of the field theoretical Koszul–Tate differential.[6,17,18] It acts nontrivially only on the antifields, and coincides on Φ_A^* with the standard Koszul–Tate differential, while $\delta\xi_{\underline{a}}^*$ is an integrated local functional associated with the corresponding rigid symmetry,

$$\delta\Phi^A = \delta\xi^{\underline{a}} = 0, \quad \delta\phi_i^* = \frac{\hat{\partial}^R L}{\hat{\partial}\phi^i}$$
$$\delta C_\alpha^* = R_\alpha^{i\,\dagger}\phi_i^*, \quad \delta\xi_{\underline{a}}^* = (-)^{\epsilon_{\underline{a}}} \int d^n x (\Delta_{\underline{a}}\phi^i)\phi_i^*, \tag{4.6}$$

where $R_\alpha^{i\,\dagger}$ is the operator adjoint to R_α^i (its precise definition, which includes a sign depending on the Grassmann parity, follows from the formulas).

Here δ is a nilpotent antiderivation by (4.5). It therefore establishes the cohomological groups $H_k(\delta)$ at antifield number k in the space of local functionals (3.4). By construction, δ is acyclic at all positive antifield numbers $[H_k(\delta) \simeq 0 \ \ \forall k > 0]$ when S encodes *all* the gauge and rigid symmetries (of first and higher order).[8] In contrast, when only a subset of the rigid symmetries is included, $H_k(\delta)$ corresponds at positive antifield number k to the remaining rigid symmetries of order k and is represented by functionals that would be of the form $M^{\hat{a}}(\xi)\delta\xi_{\hat{a}}^*$ if all the rigid symmetries had been included. Hence, $H_1(\delta)$ is represented by functionals

$$M^{\hat{a}}(\xi) \int d^n x (\Delta_{\hat{a}}\phi^i)\phi_i^* . \tag{4.7}$$

V. DEFORMATION THEORY

The extended antifield formalism allows one to describe deformations of a given model and some of its symmetries as deformations of the solution of the extended master equation along the lines of Ref. 2. However, as anticipated in the Introduction, a deformation does not necessarily preserve the property that the selected subset of symmetries is a closed one. Therefore, the deformation itself may make it necessary to enlarge the subset of symmetries one has started with. In this section we describe how to cope with this phenomenon within a systematic approach to the deformation problem.

The starting point is a solution ${}^{(0)}S$ of the extended master equation which encodes the original (undeformed) classical action, its gauge symmetries and a closed subset $\{{}^{(0)}\Delta_{\underline{a}}\}$ of its rigid symmetries. The basic idea is to seek a continuous deformation of this solution of the form

$$S = \overset{(0)}{S} + g\overset{(1)}{S} + g^2\overset{(2)}{S} + \cdots , \tag{5.1}$$

where g is the deformation parameter. This problem is analyzed ‘‘perturbatively’’ by expanding $(S,S) = 0$ in g,

$$(\overset{(0)}{S},\overset{(0)}{S})=0, \tag{5.2}$$

$$(\overset{(0)}{S},\overset{(1)}{S})=0, \tag{5.3}$$

$$(\overset{(1)}{S},\overset{(1)}{S})+2(\overset{(0)}{S},\overset{(2)}{S})=0, \tag{5.4}$$

$$\vdots$$

Equation (5.2) is satisfied by assumption. In order to discuss the subsequent equations, one may now be tempted to adopt the arguments valid for deformations preserving only the gauge symmetries, as given in Ref. 2. One would then conclude from (5.3) that ${}^{(1)}S$ must be invariant under the undeformed extended BRST differential ${}^{(0)}s$, as the latter is generated by the antibracket with ${}^{(0)}S$ [see (4.1)]. Furthermore one can assume without loss of generality that ${}^{(1)}S$ is nontrivial in the cohomology of ${}^{(0)}s$, because otherwise it can be removed through local field redefinitions and/or redefinitions of the gauge and rigid symmetry transformations by adding trivial symmetries. This follows from standard arguments which parallel those for deformations of gauge symmetries (see, e.g., Ref. 3) and are not repeated here. In this way one would conclude that ${}^{(1)}S$ represents a nontrivial cohomology class of $H^0({}^{(0)}s)$, the cohomology of ${}^{(0)}s$ at ghost number 0 in the space of local functionals (3.4). However, this kind of reasoning overlooks that ${}^{(0)}s$ encodes only a subset of the rigid symmetries and may thus be extended, if necessary.

In order to discuss this possibility, we analyze (5.3) and the subsequent equations more carefully by expanding them in the antifield number. To this end we denote the decomposition of ${}^{(n)}S$ by

$$\overset{(n)}{S}=\sum_{k\geqslant 0}\overset{(n)}{S}_k,\quad \mathrm{agh}\,\overset{(n)}{S}_k=k. \tag{5.5}$$

The interpretation of the various terms in this expansion follows from the general discussion in Sec. III: ${}^{(n)}S_0$ is the deformation of the original classical action at order n in g, ${}^{(n)}S_1$ encodes the nth-order deformations of the symmetry transformations under study, ${}^{(n)}S_2$ yields the nth-order deformation of the graded commutator algebra of these symmetries, etc. Using the expansion of ${}^{(0)}s$ in the antifield number as in (4.4), Eq. (5.3) decomposes into

$$\overset{(0)}{\gamma}\overset{(1)}{S}_0+\overset{(0)}{\delta}\overset{(1)}{S}_1=0, \tag{5.6}$$

$$\overset{(0)}{s}_1\overset{(1)}{S}_0+\overset{(0)}{\gamma}\overset{(1)}{S}_1+\overset{(0)}{\delta}\overset{(1)}{S}_2=0, \tag{5.7}$$

$$\vdots$$

Equation (5.6) requires ${}^{(1)}S_0$ to be *invariant on-shell* under the undeformed gauge and rigid symmetries under study, where ''on-shell'' refers to the undeformed equations of motion. This is so because the undeformed symmetries under study and the original equations of motion are encoded in ${}^{(0)}\gamma$ and ${}^{(0)}\delta$, respectively. Let us assume we have found a solution to (5.6). The possible need for an enhancement of the subset of rigid symmetries under study arises for the first time in the next step, i.e., when seeking a solution of (5.7). To see this we act with ${}^{(0)}\gamma$ on (5.6). Using the anticommutation relations (4.5) for ${}^{(0)}s$, we infer that the functional W_1 defined by

$$W_1=\overset{(0)}{s}_1\overset{(1)}{S}_0+\overset{(0)}{\gamma}\overset{(1)}{S}_1 \tag{5.8}$$

is ${}^{(0)}\delta$-closed,

$$\overset{(0)}{\delta} W_1 = 0. \tag{5.9}$$

Now, (5.7) requires that W_1 be ${}^{(0)}\delta$-exact. Equation (5.9) is thus a necessary condition for the existence of a solution to (5.7). However, it is not sufficient in general when ${}^{(0)}s$ encodes only a subset of the rigid symmetries (see Sec. IV).

The question at this stage is therefore: can it happen that W_1 contains a rigid symmetry of the original action which is not contained in the closed subset of symmetries one has started with? The answer to this question is affirmative, as we shall illustrate explicitly in the next sections. Hence, as W_1 has antifield number 1, it may contain contributions of the form (4.7). Furthermore, W_1 has ghost number 1. Its general form is thus

$$W_1 = \frac{1}{2}(-)^{\epsilon_{\underline{a}}+\epsilon_{\hat{c}}} \overset{(1)}{f}{}_{\underline{ab}}^{\hat{c}} \xi^{\underline{b}} \xi^{\underline{a}} \int d^n x (\overset{(0)}{\Delta}_{\hat{c}} \phi^i) \phi_i^* - \overset{(0)}{\delta}(\cdots). \tag{5.10}$$

Recall that ${}^{(0)}\Delta_{\hat{c}}$ denotes a rigid symmetry of ${}^{(0)}S_0$ that is not contained in $\{{}^{(0)}\Delta_{\underline{a}}\}$. If such symmetries occur in W_1, i.e., if there are nonvanishing coefficients ${}^{(1)}f_{\underline{ab}}^{\hat{c}}$, the subset of rigid symmetries under study needs to be enlarged by including these symmetries in order to solve (5.7). Of course, this requires one first of all to construct a new solution ${}^{(0)}S$ of the extended master equation which incorporates the additional symmetries, too, and then to reexamine Eqs. (5.6) and (5.7) as ${}^{(0)}s$ gets extended.

Let us assume now that Eqs. (5.6) and (5.7) have been solved. Then there are no further obstructions to a solution of Eq. (5.3) if higher-order symmetries are absent, i.e., all the equations subsequent to (5.7) can be solved without further ado because then ${}^{(0)}\delta$ is acyclic at all antifield numbers exceeding 1. In contrast, if there are higher-order symmetries, it cannot be excluded in principle that some of them show up at a certain stage and must be included, too.

Once one has solved (5.3), one has to analyze (5.4) and the subsequent equations. Now, one has $({}^{(0)}S,({}^{(1)}S,{}^{(1)}S))=0$ as a consequence of (5.3), thanks to the Jacobi identity for the extended antibracket. $({}^{(1)}S,{}^{(1)}S)$ has ghost number 1 and is thus a cocycle in $H^1({}^{(0)}s)$. This is a necessary condition for the existence of a solution to (5.4) but, in general, it is not sufficient because (5.4) requires that $({}^{(1)}S,{}^{(1)}S)$ be ${}^{(0)}s$-exact. Therefore (5.4) may obstruct deformations through $H^1({}^{(0)}s)$. Note, however, that some of the cohomology classes in $H^1({}^{(0)}s)$ will originate from rigid symmetries that have not been included so far. These classes are represented by ${}^{(0)}s$-invariant extensions of functionals of the form (4.7) and their analogs for higher-order rigid symmetries (if any). Such classes can be removed by further extending the subset of rigid symmetries. We shall therefore refer to them as ''spurious anomalies,'' and call the other classes ''true anomalies.'' [The term ''anomaly'' is (ab)used here because these obstructions parallel those to the Slavnov–Taylor identity through gauge anomalies in quantum field theory. Indeed, the Slavnov–Taylor identity can be cast in the form of the master equation[19,20] and the gauge anomalies represent BRST cohomology classes at ghost number 1.[21]] These two kinds of anomalies show up at different antifield numbers. [I have not found an example where $({}^{(1)}S,{}^{(1)}S)$ contains spurious anomalies. On the other hand, I have neither found a general argument which excludes the occurrence of spurious anomalies. Hence, the question whether or not such anomalies can really occur in $({}^{(1)}S,{}^{(1)}S)$ is actually still open.] Using the expansion

$$(\overset{(1)}{S},\overset{(1)}{S}) = -2\sum_{k\geq 0} A_k, \quad \mathrm{agh}\, A_k = k, \tag{5.11}$$

(5.4) decomposes into

$$A_0 = \overset{(0)}{\gamma} \overset{(2)}{S}_0 + \overset{(0)}{\delta} \overset{(2)}{S}_1, \tag{5.12}$$

$$A_1 = \overset{(0)}{s}_1 \overset{(2)}{S}_0 + \overset{(0)}{\gamma} \overset{(2)}{S}_1 + \overset{(0)}{\delta} \overset{(2)}{S}_0, \tag{5.13}$$

$$\vdots$$

True anomalies can show up only in A_0 through contributions that are weakly (=on-shell) ${}^{(0)}\gamma$-closed but not weakly ${}^{(0)}\gamma$-exact. They can thus obstruct (5.12). In contrast, spurious anomalies would show up in the A_k with $k>0$, and thus in (5.13) and the equations subsequent to it. Thereby spurious anomalies stemming from rigid symmetries of order k would show up in A_k. In particular, when higher-order rigid symmetries are absent, actually only A_1 can give rise to spurious anomalies through terms of the form (4.7) with ghost number 1. Analogously one analyzes the equations subsequent to (5.4) and infers that they can obstruct the deformation in the same way through $H^1({}^{(0)}s)$.

To summarize, the extended antifield formalism permits a systematic analysis of deformations preserving certain rigid symmetries in addition to the gauge symmetries in a manner which is quite similar to the deformation theory[2] based on the standard antifield formalism. The main difference is that the deformation itself may force one to enlarge the subset of rigid symmetries one has started with. It should be clear from the above discussion that, in general, one cannot predict from the outset which symmetries need to be included in addition to those one has started with because that may depend on the solution to (5.6).

A deformation which requires the enlargement of an originally closed subset of symmetries results in a deformed symmetry algebra. For instance, (5.10) would yield

$$\overset{(1)}{S}_2 = \tfrac{1}{2}(-)^{\epsilon_{\underline{b}}+1} \xi^{\underline{b}} \xi^{\underline{a}} \overset{(1)}{f}{}_{\underline{ab}}^{\hat{c}} \xi^*_{\hat{c}} + \cdots . \tag{5.14}$$

This shows that the graded commutator algebra of the $\Delta_{\underline{a}}$ (i.e., of the deformed transformations) would not close anymore in the soft sense but involve the $\Delta_{\hat{c}}$. The ${}^{(1)}f_{\underline{ab}}^{\hat{c}}$ are the corresponding structure constants of the deformed graded Lie algebra to first order.

VI. CENTRAL CHARGE OF THE $N=2$ HYPERMULTIPLET

As an illustration, we shall now treat an $N=2$ supersymmetric model for a Fayet–Sohnius hypermultiplet[11,12] in flat four-dimensional spacetime. The multiplet contains two complex Lorentz-scalar fields φ^i ($i=1,2$) and two complex Weyl-spinor fields χ^α, ψ^α ($\alpha=1,2$). (We use conventions with a Minkowski metric $\eta_{\mu\nu}=\mathrm{diag}(+,-,-,-)$ as in Ref. 22 which differ only through signs from those in Ref. 16). As the basis of the classical fields we use these fields and their complex conjugates (equivalently we could have chosen, for instance, the real and imaginary parts of the fields),

$$\{\phi\} = \{\varphi^i, \overline{\varphi}_i, \chi^\alpha, \overline{\chi}^{\dot\alpha}, \psi^\alpha, \overline{\psi}^{\dot\alpha}\},$$

where $\overline{\varphi}_i$, $\overline{\chi}$ and $\overline{\psi}$ are complex conjugate to φ^i, χ and ψ, respectively,

$$\overline{\varphi}_i = \overline{\varphi^i}, \quad \overline{\chi}^{\dot\alpha} = \overline{\chi^\alpha}, \quad \overline{\psi}^{\dot\alpha} = \overline{\psi^\alpha}.$$

The position of the index of $\overline{\varphi}_i$ indicates that it transforms contragrediently to φ^i under the SU(2)-automorphism group of $N=2$ supersymmetry [i refers to the fundamental representation of this SU(2)]. Undotted and dotted spinor indices distinguish the (1/2,0) and (0,1/2) representations of the Lorentz group [resp., of its covering group SL(2,C)].

Our starting point is the action

$$\overset{(0)}{S}_0 = \int d^4x [\partial_\mu \varphi^i \partial^\mu \overline{\varphi}_i - \tfrac{1}{2}(\chi \partial\!\!\!/ \overline{\chi} + \overline{\chi} \partial\!\!\!/ \chi + \psi \partial\!\!\!/ \overline{\psi} + \overline{\psi} \partial\!\!\!/ \psi)], \tag{6.1}$$

where

$$\partial_{\alpha\dot\alpha}=\sigma^{\mu}_{\alpha\dot\alpha}\partial_{\mu}.$$

The action ${}^{(0)}S_0$ is among others invariant under rigid $N=2$ supersymmetry transformations ${}^{(0)}\Delta_{\alpha}{}^{i}$, ${}^{(0)}\Delta_{\dot\alpha i}$ given by

ϕ	φ^j	$\overline{\varphi_j}$	χ^β	$\overline{\chi}^{\dot\beta}$	ψ^β	$\overline{\psi}^{\dot\beta}$
$\overset{(0)}{\Delta}{}_{\alpha}{}^{i}\phi$	$\epsilon^{ij}\chi_\alpha$	$\delta^i_j\psi_\alpha$	0	$-\mathrm{i}\partial^{\dot\beta}_{\alpha}\overline{\varphi^i}$	0	$-\mathrm{i}\partial^{\dot\beta}_{\alpha}\varphi^i$
$\overset{(0)}{\Delta}{}_{\dot\alpha i}\phi$	$\delta^j_i\overline{\psi}_{\dot\alpha}$	$-\epsilon_{ij}\overline{\chi}_{\dot\alpha}$	$\mathrm{i}\partial^{\beta}_{\dot\alpha}\varphi_i$	0	$-\mathrm{i}\partial^{\beta}_{\dot\alpha}\overline{\varphi_i}$	0

(6.2)

where indices i are raised and lowered with the rules

$$X^i=\epsilon^{ij}X_j,\quad X_i=\epsilon_{ij}X^j,\quad \epsilon^{ij}=-\epsilon^{ji},\quad \epsilon_{ij}=-\epsilon_{ji},\quad \epsilon^{12}=\epsilon_{21}=1.$$

We consider the following subset of rigid symmetries, containing the supersymmetry transformations and the spacetime translations,

$$\{\overset{(0)}{\Delta}{}_{\underline{a}}\}\equiv\{\overset{(0)}{\Delta}{}_{\alpha}{}^{i},\overset{(0)}{\Delta}{}_{\dot\alpha i},\partial_\mu\}. \tag{6.3}$$

The graded commutator algebra of these symmetries reads

$$[\overset{(0)}{\Delta}{}_{\alpha}{}^{i},\overset{(0)}{\Delta}{}_{\dot\alpha j}]\approx-\mathrm{i}\delta^i_j\partial_{\alpha\dot\alpha},\quad [\overset{(0)}{\Delta}{}_{\underline{a}},\overset{(0)}{\Delta}{}_{\underline{b}}]\approx 0\quad \text{otherwise}, \tag{6.4}$$

where $\approx$ denotes equality up to on-shell trivial symmetries. Equation (6.4) is indeed the $N=2$ supersymmetry algebra without central charge (on-shell). The action has no gauge symmetries. Therefore, it has no higher-order symmetries either.[23] This implies the existence of a solution to the extended master equation which encodes only the symmetries (6.3) and their graded commutator algebra. This solution, which was computed first in Ref. 24, reads

$$\overset{(0)}{S}=\overset{(0)}{S}{}_0+\overset{(0)}{S}{}_1+\overset{(0)}{S}{}_2,$$

$$\overset{(0)}{S}{}_1=-\int d^4x\sum_{\phi}\,(\xi^\alpha_i\overset{(0)}{\Delta}{}_{\alpha}{}^{i}\phi+\overline{\xi}^{\dot\alpha i}\overset{(0)}{\Delta}{}_{\dot\alpha i}\phi+\xi^\mu\partial_\mu\phi)\phi^*, \tag{6.5}$$

$$\overset{(0)}{S}{}_2=-\mathrm{i}\xi_i\sigma^\mu\overline{\xi^i}\xi^*_\mu+\int d^4x[\overline{\chi}^*\overline{\xi^i}\xi_i\chi^*+\overline{\psi}^*\overline{\xi^i}\xi_i\psi^*+\tfrac{1}{2}\xi_i\xi^i\overline{\chi}^*\overline{\psi}^*+\tfrac{1}{2}\overline{\xi^i\xi_i}\psi^*\chi^*],$$

where

$$\{\phi^*\}=\{\varphi^*_i,\overline{\varphi}^{i*},\chi^*_\alpha,\overline{\chi}^*_{\dot\alpha},\psi^*_\alpha,\overline{\psi}^*_{\dot\alpha}\}.$$

The supersymmetry ghosts ξ^α_i and $\overline{\xi}^{\dot\alpha i}$ are Grassmann even and the translation ghosts ξ^μ are Grassmann odd. The ghosts and antifields have the reality properties

$$\overline{\xi}^{\dot\alpha i}=\overline{\xi^\alpha_i},\quad \xi^\mu=\overline{\xi^\mu},\quad \overline{\xi}^*_{\dot\alpha i}=-\overline{\xi^{i*}_\alpha},\quad \xi^*_\mu=-\overline{\xi^*_\mu},$$

$$\overline{\varphi}^{i*}=-\overline{\varphi^*_i},\quad \overline{\chi}^*_{\dot\alpha}=-\overline{\chi^*_\alpha},\quad \overline{\psi}^*_{\dot\alpha}=-\overline{\psi^*_\alpha}.$$

The first term in ${}^{(0)}S_2$ contains the structure constants of the supersymmetry algebra (6.4), while the contributions which are quadratic in the antifields reflect that the symmetry algebra closes only on-shell.

We now study deformations of the above model along the lines of the previous section. A solution to (5.6) which introduces mass terms for the fermions is easily found. Namely,

$$\overset{(1)}{S}_0=\int d^4x[m_1\chi\psi+\overline{m}_1\overline{\chi}\overline{\psi}+\tfrac{1}{2}(m_2\chi\chi+\overline{m}_2\overline{\chi\chi}+m_3\psi\psi+\overline{m}_3\overline{\psi\psi})] \tag{6.6}$$

is supersymmetric on-shell and translation invariant for any choice of complex mass parameters m_1, m_2 and m_3 and therefore yields a solution to (5.6). The corresponding functional ${}^{(1)}S_1$ is

$$\begin{aligned}\overset{(1)}{S}_1=\int d^4x[&m_1(\overline{\varphi_i}\overline{\chi}^*\overline{\xi^i}-\varphi_i\overline{\psi}^*\overline{\xi^i})-\overline{m}_1(\overline{\varphi}^i\xi_i\psi^*+\varphi^i\xi_i\chi^*)\\&-m_2\varphi_i\overline{\chi}^*\overline{\xi^i}-\overline{m}_2\overline{\varphi}^i\xi_i\chi^*+m_3\overline{\varphi_i}\overline{\psi}^*\overline{\xi^i}-\overline{m}_3\varphi^i\xi_i\psi^*].\end{aligned} \tag{6.7}$$

Next we calculate the functional W_1 in (5.8). The result is

$$\begin{aligned}\overset{(0)}{s}_1\overset{(1)}{S}_0+\overset{(0)(1)}{\gamma\, S}_1=\frac{1}{2}\overline{\xi^i}\overline{\xi_i}\int d^4x[&(-m_1\varphi^j+m_3\overline{\varphi}^j)\varphi_j^*+\overline{\varphi}^{j*}(m_1\overline{\varphi_j}-m_2\varphi_j)\\&-(m_1\chi+m_3\psi)\chi^*+\overline{\chi}^*(m_1\overline{\chi}-m_2\overline{\psi})\\&+(m_1\psi+m_2\chi)\psi^*+\overline{\psi}^*(-m_1\overline{\psi}+m_3\overline{\chi})]+\text{c.c.},\end{aligned} \tag{6.8}$$

where c.c. denotes complex conjugation. Equation (6.8) has the form of the first term in (5.10), i.e., it brings in an additional symmetry. This symmetry is part of a rigid SU(2)-invariance of the action (6.1). Indeed, as the functional (6.8) is ${}^{(0)}\delta$-invariant for any choice of m_1, m_2 and m_3, the parts in (6.8) involving m_1, m_2 and m_3, respectively, correspond to independent symmetries of the action (6.1). These symmetries form an SU(2) under which $(\varphi^1,\overline{\varphi}^1)$, $(\varphi^2,\overline{\varphi}^2)$, (χ,ψ) and $(\overline{\psi},\overline{\chi})$ transform as doublets (i.e., in the fundamental representation) and which commutes with the supersymmetry transformations (6.2). However, in contrast to the undeformed action, the first-order deformation (6.6) is not invariant under the full SU(2) but it is still invariant under a U(1) subgroup thereof generated by the transformations in (6.8). Hence, the deformation breaks the SU(2) but preserves this U(1) subgroup. (An analogous phenomenon was observed in Ref. 25 within the construction of supergravity couplings for hypermultiplets.)

We thus have to enlarge the subset of symmetries (6.3) by this U(1). It turns out that this suffices in order to construct a deformed solution of the extended master equation. We shall not further discuss the computation and spell out the solution only for the case $m_2=m_3=0$. Using $m=gm_1$ (g being the deformation parameter in the notation of the previous section), the deformed solution reads then

$$S=S_0+S_1+S_2, \tag{6.9}$$

$$S_0=\overset{(0)}{S}_0+\int d^4x(m\chi\psi+\overline{m\chi}\overline{\psi}-m\overline{m}\varphi^i\overline{\varphi_i}), \tag{6.10}$$

$$S_1=\int d^4x\Big[-\sum_{\phi}\,(\xi_i^{\alpha}\overset{(0)}{\Delta}{}_{\alpha}{}^{i}\phi+\overline{\xi}^{\dot\alpha i}\overset{(0)}{\Delta}{}_{\dot\alpha i}\phi+\xi^{\mu}\partial_{\mu}\phi)\phi^*$$
$$+\mathrm{i}\xi_{U(1)}(\varphi^i\varphi_i^*-\overline{\varphi}_i\overline{\varphi}^{i}{}^*+\chi\chi^*-\overline{\chi}^*\overline{\chi}-\psi\psi^*+\overline{\psi}^*\overline{\psi})$$
$$+m(\overline{\varphi}_i\overline{\chi}^*\overline{\xi}^i-\varphi_i\overline{\psi}^*\overline{\xi}^i)-\overline{m}(\overline{\varphi}^i\xi_i\psi^*+\varphi^i\xi_i\chi^*)\Big], \tag{6.11}$$

$$S_2=\overset{(0)}{S}{}_2+\frac{\mathrm{i}}{2}(m\overline{\xi^i\xi_i}+\overline{m}\xi_i\xi^i)\xi^*_{U(1)}, \tag{6.12}$$

with ${}^{(0)}S_2$ as in (6.5). Here $\xi_{U(1)}$ and $\xi^*_{U(1)}$ are the global ghost and antifield of the rigid U(1) symmetry obtained from (6.8) in the case $m_2=m_3=0$ ($\xi_{U(1)}$ is real and Grassmann odd, $\xi^*_{U(1)}$ is purely imaginary and Grassmann even).

Equation (6.10) is the deformed classical action. Apart from the original action (6.1) and its first-order deformation (6.6) (in the case $m_2=m_3=0$), it contains also a mass term for the Lorentz-scalar fields which arises at second order in the deformation parameter.

Equation (6.11) contains the deformed supersymmetry transformations, the rigid U(1) transformations, and the space–time translations. The deformed supersymmetry and the U(1) transformations are

ϕ	φ^j	$\overline{\varphi}_j$	χ^β	$\overline{\chi}^{\dot\beta}$	ψ^β	$\overline{\psi}^{\dot\beta}$
$\Delta_\alpha{}^i\phi$	$\epsilon^{ij}\chi_\alpha$	$\delta^i_j\psi_\alpha$	$\overline{m}\delta^\beta_\alpha\varphi^i$	$-\mathrm{i}\partial^{\dot\beta}_\alpha\overline{\varphi}^i$	$\overline{m}\delta^\beta_\alpha\overline{\varphi}^i$	$-\mathrm{i}\partial^{\dot\beta}_\alpha\varphi^i$
$\Delta_{\dot\alpha i}\phi$	$\delta^j_i\overline{\psi}_{\dot\alpha}$	$-\epsilon_{ij}\overline{\chi}_{\dot\alpha}$	$\mathrm{i}\partial^\beta_{\dot\alpha}\varphi_i$	$-m\delta^{\dot\beta}_{\dot\alpha}\overline{\varphi}_i$	$-\mathrm{i}\partial^\beta_{\dot\alpha}\overline{\varphi}_i$	$m\delta^{\dot\beta}_{\dot\alpha}\varphi_i$
$\Delta_{U(1)}\phi$	$-\mathrm{i}\varphi^j$	$\mathrm{i}\overline{\varphi}_j$	$-\mathrm{i}\chi^\beta$	$\mathrm{i}\overline{\chi}^{\dot\beta}$	$\mathrm{i}\psi^\beta$	$-\mathrm{i}\overline{\psi}^{\dot\beta}$

(6.13)

Equation (6.12) encodes the graded commutator algebra of the deformed symmetry transformations. The $N=2$ supersymmetry algebra has become extended by the U(1) through the deformation. The nonvanishing graded commutators are

$$[\Delta_\alpha{}^i,\Delta_{\dot\alpha j}]\approx-\mathrm{i}\,\delta^i_j\partial_{\alpha\dot\alpha},$$
$$[\Delta_\alpha{}^i,\quad\Delta_\beta{}^j]\approx\mathrm{i}\overline{m}\epsilon^{ij}\epsilon_{\alpha\beta}\Delta_{U(1)}, \tag{6.14}$$
$$[\Delta_{\dot\alpha i},\Delta_{\dot\beta j}]\approx-\mathrm{i}m\epsilon_{ij}\epsilon_{\dot\alpha\dot\beta}\Delta_{U(1)},$$

where $\approx$ now denotes equality up to transformations which are trivial on-shell in the deformed model (i.e., these transformations involve the deformed equations of motion).

Remark: The above results hold analogously in a formulation of the hypermultiplet with the standard auxiliary fields used already in Refs. 11 and 12. In that approach one sometimes introduces an ''off-shell central charge'' in order to close the commutator algebra of the supersymmetries, the central charge and the space–time translations off-shell. However, in the massless model that central charge is trivial on-shell and thus not to be accompanied by global ghosts. In contrast, the massive (deformed) model involves again a ''true'' central charge that does not vanish on-shell.

VII. CURCI–FERRARI MODEL

A particular case of a rigid symmetry is the BRST symmetry of a gauge fixed action constructed in the standard way from a solution to the usual master equation.[5–7] Deformations of a gauge fixed action may be obtained in two ways: (i) one constructs first consistent deformations of the underlying gauge theory along the lines of Ref. 2 and fixes the gauge afterwards, or (ii) one investigates directly deformations of the gauge fixed model and its BRST symmetry.

These two approaches are not equivalent in general. In particular, the first approach leads by construction to an on-shell nilpotent BRST symmetry of the standard type, whereas the second one may destroy the nilpotency property and is not physically acceptable in general. This is related to the different properties of the BRST cohomology before and after gauge fixing[26] (cf. also the remark at the end of this section), and is now to be discussed for the Curci–Ferrari model[13–15] in the framework of the extended antifield formalism. The loss of nilpotency emerges in this approach as a deformation of the BRST algebra along the lines of Sec. V. In the particular case of the Curci–Ferrari model, the deformed action has even the *same* BRST symmetry as the original one, but in the deformed model that symmetry does not square weakly to zero anymore (as the equations of motion change). Rather, it squares into a different nontrivial rigid symmetry.

We consider four-dimensional non-Abelian Yang–Mills theory with the following gauge fixed action,

$$\overset{(0)}{S}_0=\int d^4x\,\mathrm{Tr}\left[\frac{1}{4}F_{\mu\nu}F^{\mu\nu}+\frac{1}{2\alpha}(\partial_\mu A^\mu)^2-\frac{1}{2}B(\partial_\mu D^\mu+D_\mu\partial^\mu)C+\frac{\alpha e^2}{4}B^2C^2\right], \tag{7.1}$$

where α is the gauge fixing parameter, e is the gauge coupling constant, and $A_\mu=A^i_\mu t_i$, $F_{\mu\nu}=F^i_{\mu\nu}t_i$, $C=C^it_i$ and $B=B^it_i$ are the Lie algebra valued gauge fields, field strengths, ghost fields and antighost fields, respectively ($\{t_i\}$ denotes an appropriate matrix representation of the Lie algebra of the gauge group normalized such that $\mathrm{Tr}(t_it_j)=-\delta_{ij}$), and $D_\mu C$ is defined by

$$D_\mu C=\partial_\mu C+e(A_\mu C-CA_\mu). \tag{7.2}$$

{The gauge fixed action (7.1) arises in the standard manner from a ''minimal'' solution $S_{\min}$ of the usual master equation as follows. First one adds to $S_{\min}$ the ''nonminimal'' term $\int d^4x\,\mathrm{Tr}(-HB^*)$ where the H^i are Nakanishi–Lautrup auxiliary fields. Then one shifts the antifields by $\Phi^*_A\rightarrow\Phi^*_A+\delta^L\Psi/\delta\Phi^A$ where Ψ is the ''gauge fixing fermion'' $\Psi[\Phi]=\int d^4x\,\mathrm{Tr}[(\alpha/2)(BH+eB^2C)-B\partial_\mu A^\mu]$. Finally one eliminates the H^i by their algebraic equations of motion.} The action (7.1) is invariant under the rigid BRST transformations

$$\Delta_{\mathrm{brs}}A_\mu=D_\mu C,\quad \Delta_{\mathrm{brs}}C=-eC^2,\quad \Delta_{\mathrm{brs}}B=\frac{1}{\alpha}\partial_\mu A^\mu-\frac{e}{2}(BC+CB). \tag{7.3}$$

These transformations are nilpotent on-shell,

$$(\Delta_{\mathrm{brs}})^2=\tfrac{1}{2}[\Delta_{\mathrm{brs}},\Delta_{\mathrm{brs}}]\approx 0. \tag{7.4}$$

More precisely, Δ_{brs} is strictly nilpotent on A^i_μ and C^i, but squares into an on-shell trivial symmetry on B^i,

$$(\Delta_{\mathrm{brs}})^2A^i_\mu=(\Delta_{\mathrm{brs}})^2C^i=0,\quad (\Delta_{\mathrm{brs}})^2B^i=\frac{1}{\alpha}\delta^{ij}\frac{\delta^L\overset{(0)}{S}_0}{\delta B^j}. \tag{7.5}$$

We shall now apply the extended antifield formalism to the gauge fixed action (7.1) and the BRST symmetry (7.3). In this approach (7.1) plays the role of the classical action, i.e., the ghost and antighost fields C and B are viewed as Grassmann odd ''classical'' fields (C is real, B purely imaginary),

$$\{\phi\}=\{A^i_\mu,B^i,C^i\}.$$

Accordingly, we assign antifield number 1 to $A_i^{\mu *}$, B_i^* and C_i^*. Furthermore, we introduce a Grassmann even global ghost $\xi_{\rm brs}$ for the BRST symmetry (7.3).

That is, in this case we consider a subset of rigid symmetries containing only one element, namely $\Delta_{\rm brs}$:

$$\{\overset{(0)}{\Delta}_{\underline{a}}\}=\{\Delta_{\rm brs}\}. \tag{7.6}$$

The corresponding graded commutator algebra (2.13) is just (7.5). As the gauge fixed action (7.1) has no gauge symmetry and thus no higher-order rigid symmetry either, a corresponding solution of the extended master equation exists. This solution coincides, of course, with the gauge fixed solution of the master equation obtained in the standard antifield formalism, except that now the global ghost $\xi_{\rm brs}$ appears,

$$\overset{(0)}{S}=\overset{(0)}{S}_0+\overset{(0)}{S}_1+\overset{(0)}{S}_2,$$

$$\overset{(0)}{S}_1=\xi_{\rm brs}\int d^4x\,\mathrm{Tr}\left[A^{\mu *}D_\mu C+eC^2C^*-\left\{\frac{1}{\alpha}\partial_\mu A^\mu-\frac{e}{2}(BC+CB)\right\}B^*\right], \tag{7.7}$$

$$\overset{(0)}{S}_2=\frac{1}{2\alpha}\xi^2_{\rm brs}\int d^4x\,\mathrm{Tr}(B^*B^*),$$

where we have used $A^{\mu *}=-\delta^{ij}A_i^{\mu *}t_j$, etc. The presence of the term quadratic in B^* reflects that the algebra closes on B^i only on-shell [see (7.5)]. The ''extended'' BRST differential ${}^{(0)}s$, constructed from ${}^{(0)}S$ as in Eq. (4.1), coincides with the usual gauge fixed BRST operator for the action (7.1), except that now $\xi_{\rm brs}$ occurs. It is strictly nilpotent, in contrast to $\Delta_{\rm brs}$, and acts on the fields by

$$\overset{(0)}{s}A_\mu=\xi_{\rm brs}D_\mu C,\qquad \overset{(0)}{s}C=-\xi_{\rm brs}eC^2,$$
$$\overset{(0)}{s}B=\xi_{\rm brs}\left[\frac{1}{\alpha}\partial_\mu A^\mu-\frac{e}{2}(BC+CB)\right]-\frac{1}{\alpha}\xi^2_{\rm brs}B^*. \tag{7.8}$$

We shall now discuss the deformation of the action (7.1) through the Curci–Ferrari mass term

$$\overset{(1)}{S}_0=\int d^4x\,\mathrm{Tr}\left[\frac{1}{2}A_\mu A^\mu+\alpha BC\right]. \tag{7.9}$$

This term is off-shell invariant under the transformations (7.3) and thus yields a solution to Eq. (5.6) with

$$\overset{(1)}{S}_1=0. \tag{7.10}$$

The functional W_1 in Eq. (5.8) reads in this case

$$\overset{(0)}{s}_1\overset{(1)}{S}_0=-\xi^2_{\rm brs}\int d^4x\,\mathrm{Tr}(B^*C). \tag{7.11}$$

This has the form of the first term in Eq. (5.10) and contains an additional nontrivial rigid symmetry of the gauge fixed action (7.1), namely,

$$\Delta_{\mathrm{add}}B = C, \quad \Delta_{\mathrm{add}}A_\mu = \Delta_{\mathrm{add}}C = 0. \tag{7.12}$$

Hence, in order to construct a deformed solution of the extended master equation with ${}^{(1)}S_0$ as in (7.9), we *must* include this symmetry. It is straightforward to verify that this yields the following deformed solution of the extended master equation,

$$S = S_0 + S_1 + S_2, \quad S_0 = \overset{(0)}{S}{}_0 + g\,\overset{(1)}{S}{}_0,$$
$$S_1 = -\int d^4x \sum_\phi \,(\xi_{\mathrm{brs}}\Delta_{\mathrm{brs}}\phi + \xi_{\mathrm{add}}\Delta_{\mathrm{add}}\phi)\,\phi^*, \tag{7.13}$$
$$S_2 = \overset{(0)}{S}{}_2 + g\,\xi^2_{\mathrm{brs}}\xi^*_{\mathrm{add}}.$$

with ${}^{(0)}S_2$ as in (7.7). The last term reflects that the graded commutator algebra of Δ_{brs} and Δ_{add} reads in the deformed model

$$(\Delta_{\mathrm{brs}})^2 \approx g\Delta_{\mathrm{add}}, \quad [\Delta_{\mathrm{brs}}, \Delta_{\mathrm{add}}] = 0, \tag{7.14}$$

where $\approx$ now denotes on-shell equality in the deformed model, i.e., for the equations of motion following from ${}^{(0)}S_0 + g\,{}^{(1)}S_0$. Notice that Δ_{brs} is still a symmetry of the deformed model, without having been deformed. Nevertheless it is not nilpotent anymore on-shell because the equations of motion have changed.

Remarks: (a) In order to avoid possible confusion, we stress that the BRST cohomologies before and after gauge fixing are always isomorphic (provided all the antifields are kept). What changes, however, when the gauge fixed action is treated as a classical one, are the assignments of antifield numbers and the corresponding concept of weak (=on-shell) equality (as the ghost fields count now among the classical fields). As a consequence, it is *not* true that each local functional with vanishing antifield number which is on-shell Δ_{brs}-invariant can be extended to a cocycle of ${}^{(0)}s$ (this is just the phenomenon discussed in Ref. 26, but in the language used here). The Curci–Ferrari mass term (7.9) illustrates exactly this phenomenon: it *cannot* be extended so as to be ${}^{(0)}s$-closed, although it is Δ_{brs}-invariant. As a consequence, Δ_{brs} is not nilpotent anymore on-shell in the deformed model.

(b) The Curci–Ferrari model illustrates a general fact: a deformation of a gauge fixed action which destroys the on-shell nilpotency of Δ_{brs} (or a deformation thereof) *cannot* reflect a consistent deformation of the gauge symmetry in the sense of Ref. 2 because such consistent deformations result by their very construction in an on-shell nilpotent Δ_{brs} after gauge fixing.

(c) Of course, (7.13) yields via (4.1) a strictly nilpotent operator which incorporates both Δ_{brs} and Δ_{add}. However, this nilpotent operator cannot cure the unitarity problems of the Curci–Ferrari model discussed in Refs. 14 and 15 because Δ_{add} does not impose additional conditions that may select physical states. Indeed, as it is just the square of Δ_{brs} (on-shell), a state that is annihilated by Δ_{brs} (resp. by its quantum version) is automatically also annihilated by Δ_{add}. For the same reason, a state that is Δ_{add}-exact is also in the image of Δ_{brs}.

VIII. CONCLUSION

We have outlined how continuous deformations of an action functional, its gauge symmetries and a closed subset of its rigid symmetries can be analyzed systematically in the extended antifield formalism. The procedure is very similar to the study of continuous deformations of actions and their gauge symmetries described in Ref. 2. The main difference is that the deformation itself may make it necessary to enlarge the particular subset of rigid symmetries one has started with. This happens when the commutator algebra of the deformed version of the originally considered subset of symmetries does not close anymore in the soft sense (i.e., modulo gauge transformations and on-shell trivial symmetries) and thus results in a deformation of the symmetry algebra.

It is, however, not always clear from the outset which additional symmetries can occur in the deformed commutator algebra. This subtlety can be mastered when one proceeds as described in Sec. V, using an expansion in the antifield number. In this approach one first seeks functionals of the classical fields that are ''weakly'' (=on-shell) invariant under the symmetries under study. The method then provides automatically the additional symmetries which need to be included. This has been illustrated for the hypermultiplet of four-dimensional $N=2$ supersymmetry where the central extension of the $N=2$ supersymmetry algebra emerges via the deformation of a massless model to a massive one. The central extension turns out to be a surviving generator of an SU(2) symmetry of the massless action broken by the deformation. In this case it depends on the mass parameters, i.e., on the deformation itself, how the SU(2) is broken and which generator becomes the central extension.

We have also illustrated, for the Curci–Ferrari model, how deformations of a gauge fixed action and its BRST symmetry can be analyzed within this approach. The BRST symmetry is then treated in the same manner as other rigid symmetries, too, while the gauge fixed action is treated as a classical one. However, such deformations do not correspond necessarily to consistent deformations of the gauge symmetries in the sense of Ref. 2, and are therefore not always physically acceptable. In particular, it can happen that there are deformations of a gauge fixed action which are BRST invariant but nevertheless inconsistent because the BRST symmetry does not square to zero on-shell anymore in the deformed model. The Curci–Ferrari model illustrates exactly this phenomenon. Hence, a necessary condition for a deformation of a gauge fixed action to be a consistent one is the on-shell nilpotency of the BRST symmetry of the deformed action.

Finally, I remark that the procedure outlined in Sec. V can be extended analogously to the case that only a subset of the gauge symmetries is included. However, from the physical point of view this extension is mainly of academic interest and was therefore not discussed here.

ACKNOWLEDGMENTS

Discussions with Tobias Hurth and Kostas Skenderis about the Curci–Ferrari model are gratefully acknowledged. The author was supported by the Deutsche Forschungsgemeinschaft.

[1] M. Gerstenhaber, ''On the deformation of rings and algebras,'' Ann. Math. **79**, 59–103 (1964).
[2] G. Barnich and M. Henneaux, ''Consistent couplings between fields with a gauge freedom and deformations of the master equation,'' Phys. Lett. B **311**, 123–129 (1993) (hep-th/9304057).
[3] M. Henneaux, ''Consistent interactions between gauge fields: the cohomological approach,'' Contemp. Math. **219**, 93–109 (1998).
[4] J. D. Stasheff, ''Deformation theory and the Batalin-Vilkovisky master equation,'' in *Deformation Theory and Symplectic Geometry*, edited by D. Sternheimer *et al.* (Kluwer, Dordrecht 1997), pp. 271–284 (q-alg/9702012).
[5] I. A. Batalin and G. A. Vilkovisky, ''Gauge algebra and quantization,'' Phys. Lett. **102B**, 27–31 (1981); ''Quantization of gauge theories with linearly dependent generators,'' Phys. Rev. D **28**, 2567–2582 (1983); Erratum, *ibid.* **30**, 508 (1984).
[6] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton U.P., Princeton, NJ, 1992).
[7] J. Gomis, J. París, and S. Samuel, ''Antibracket, antifields and gauge theory quantization,'' Phys. Rev. **259**, 1–145 (1995) (hep-th/9412228).
[8] F. Brandt, M. Henneaux, and A. Wilch, ''Extended antifield formalism,'' Nucl. Phys. B **510**, 640–656 (1998) (hep-th/9705007).
[9] F. Brandt, M. Henneaux, and A. Wilch, ''Global symmetries in the antifield formalism,'' Phys. Lett. B **387**, 320–326 (1996) (hep-th/9606172).
[10] R. Haag, J. T. Lopuszanski, and M. Sohnius, ''All possible generators of supersymmetries of the S-matrix,'' Nucl. Phys. B **88**, 257–274 (1975).
[11] P. Fayet, ''Fermi-Bose hypersymmetry,'' Nucl. Phys. B **113**, 135–155 (1976).
[12] M. F. Sohnius, ''Supersymmetry and central charges,'' Nucl. Phys. B **138**, 109–121 (1978).
[13] G. Curci and R. Ferrari, ''On a class of Lagrangian models for massive and massless Yang-Mills fields,'' Nuovo Cimento A **32**, 151–168 (1976); ''The unitarity problem and the zero-mass limit for a model of massive Yang-Mills theory,'' *ibid.* **35**, 1–14 (1976).
[14] I. Ojima, ''Comments on massive and massless Yang-Mills Lagrangians with a quartic coupling of Faddeev-Popov ghosts,'' Z. Phys. C **13**, 173–177 (1982).
[15] J. de Boer, K. Skenderis, P. van Nieuwenhuizen, and A. Waldron, ''On the renormalizability and unitarity of the Curci-Ferrari model for massive vector bosons,'' Phys. Lett. B **367**, 175–182 (1996) (hep-th/9510167).
[16] J. Wess and J. Bagger, *Supersymmetry and Supergravity* (Princeton U.P., Princeton, NJ, 1983).

[17] J. M. L. Fisch and M. Henneaux, ''Homological perturbation theory and the algebraic structure of the antifield-antibracket formalism for gauge theories,'' Commun. Math. Phys. **128**, 627–640 (1990).
[18] M. Henneaux, ''Lectures on the antifield-BRST formalism for gauge theories,'' Nucl. Phys. B (Proc. Suppl.) **18A**, 47–106 (1990).
[19] J. Zinn-Justin, ''Renormalization of gauge theories,'' in *Trends in Elementary Particle Physics*, Lecture Notes in Physics, Vol. 37, edited by H. Rollnik and K. Dietz (Springer-Verlag, Berlin, 1975) pp. 2–39.
[20] B. W. Lee, ''Gauge theories,'' in *Methods in Field Theory*, edited by R. Balian and J. Zinn-Justin (North-Holland, Amsterdam, 1976), pp. 79–139.
[21] C. Becchi, A. Rouet, and R. Stora, ''Renormalization of the abelian Higgs-Kibble model,'' Commun. Math. Phys. **42**, 127–162 (1975); ''Renormalization of gauge theories,'' Ann. Phys. (N.Y.) **98**, 287–321 (1976).
[22] F. Brandt, ''Local BRST cohomology in minimal $D=4$, $N=1$ supergravity,'' Ann. Phys. (N.Y.) **259**, 253–312 (1997) (hep-th/9609192).
[23] G. Barnich, F. Brandt, and M. Henneaux, ''Local BRST cohomology in the antifield formalism: I. General theorems,'' Commun. Math. Phys. **174**, 57–91 (1995) (hep-th/9405109).
[24] L. Baulieu, M. Bellon, S. Ouvry, and J.-C. Wallet, ''Batalin-Vilkovisky analysis of supersymmetric systems,'' Phys. Lett. B **252**, 387–394 (1990).
[25] B. de Wit, J. W. van Holten, and A. Van Proeyen, ''Central charges and conformal supergravity,'' Phys. Lett. B **95**, 51–55 (1980).
[26] M. Henneaux, ''On the gauge-fixed BRST cohomology,'' Phys. Lett. B **367**, 163–169 (1996) (hep-th/9510116).

The Eikonal equation in asymptotically flat space–times

Simonetta Frittelli
Physics Department, Duquesne University, Pittsburgh, Pennsylvania 15282
and Physics Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Ezra T. Newman
Physics Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Gilberto Silva-Ortigoza
Physics Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260
and Facultad de Ciencias Físico Matemáticas de la Universidad Autónoma de Puebla,
Apartado Postal 1152, 72001 Puebla, Puebla, México

(Received 20 July 1998; accepted for publication 6 October 1998)

In an arbitrary Lorentzian manifold we provide a description for the construction of null surfaces and their associated singularities, via solutions of the Eikonal equation. In particular, we study the singularities of the past light-cones from points on null infinity, the future light-cones from arbitrary interior points and the intersection of these with null infinity and unifying relationships between the different singularities. The starting point for this work is the assumption of a known family of solutions to the Eikonal equation. The work is based on the standard theory of singularities of smooth maps by Arnold and his colleagues. Though the work is intended to stand on its own, it can be thought of as being closely related to the recently developed null surface reformulation of GR. © *1999 American Institute of Physics.* [S0022-2488(99)01302-X]

I. INTRODUCTION

In a recent work[1] we studied properties of solutions of the flat-space–time Eikonal equation, namely,

$$\eta^{ab}\partial_a S\partial_b S=0,$$

whose level "surfaces" [$S=S(x^a)=\text{const.}$], are, by definition, null (or characteristic) three-surfaces. These level surfaces, called by Arnold[2,3] "big wave fronts," can have self-intersections and need not be smooth everywhere. In particular we were concerned with finding the analytic form (described parametrically) of the general solution to the equation, studying its level surfaces and the "small (two-dimensional) wave fronts," (i.e., the intersection of a three surface with a big wave front) and then analyzing some of the resulting structures; the caustics of the full solution (three-dimensional), the singularities of the big wave front (two-dimensional), and the singularities of the "small wave fronts" (one-dimensional). These singularities are defined, respectively, by the intersection of a big wave front and the small wave front with the caustic surface. A special application of these ideas was to the study of 2-parameter families of solutions to the Eikonal equation from which it was possible to see an alternative analytic treatment of the structure of the singularities. This latter point of view plays an important role in a recent reformulation of GR known as the null surface formulation.[4,5]

In the present paper, we extend the ideas from the Minkowski case to,[1] first to arbitrary Lorentzian space–times and then specialize them to asymptotically flat space–times. There are two main reasons for doing this: (1) We want to understand in detail the structure of light-cones in the large, i.e., globally, in arbitrary space–times which are of great relevance to the general theory of gravitational lensing and (2) a recent reformulation of GR in terms of families of characteristic surfaces requires a deeper understanding of the singularities of the big and small wave fronts.

In Sec. II, we will show how from any arbitrary, but given, two parameter family of solutions of the curved-space Eikonal equation, any arbitrary characteristic surface can be constructed.

This construction will, in Sec. III, be specialized to an asymptotically flat space–time where the two-parameter family is chosen in a special way; namely, they are the family of past light cones from all the points on null infinity, $\mathfrak{I}^+$. Directly in terms of this fiducial family we can express any characteristic surface and in particular, we can express the light-cone, $\mathfrak{C}_x$, of any interior point x^a. Of particular interest is the singularity structure of the cones, $\mathfrak{C}_x$, which can be analyzed in terms of the variables of the fiducial family.

In Sec. IV, we will study the particular class of small wave fronts (two-dimensional) defined by the intersection of the three-dimensional cones, $\mathfrak{C}_x$, with the null surface $\mathfrak{I}^+$, i.e., the so-called light-cone cuts $\mathfrak{c}(x^a)$ of $\mathfrak{I}^+$. In particular we will be interested in finding (via Arnold's theory of Lagrange and Legendre maps[2,3,6,7]) the appropriate tools and variables to describe the singularities of these light-cone cuts.

Finally in Sec. V, we return to an issue that we deliberately postponed. We took, in Sec. III, a fiducial family of solutions of the Eikonal and used them to study the singularities of other characteristic surfaces but we avoided any discussion of the singularities of the fiducial family itself. The reason for the postponement is that this discussion is more complicated and difficult than the earlier ones and uses, in addition, different techniques; namely the equations of geodesic deviation.

The present work is partially intended to fill in the details of an earlier brief work in the Twistor Newsletter (TN43, 1997), where we anticipated some of these results.

II. SOLUTIONS OF THE EIKONAL EQUATION IN CURVED SPACE

In this section we will treat the Eikonal equation in a general curved Lorentzian space–time, $(g,\mathfrak{M})$, i.e.,

$$g^{ab}(x^a)\partial_a S\partial_b S=0 \tag{1}$$

and show how, if a special class of solutions is known, *any* solution can be easily constructed. An important special case of this will be the construction of any single characteristic surface, i.e., a level surface of *some* S, ''a big wave front.''

The difficult task (and it is very difficult, where perturbation techniques must be relied on) is to produce this special class. Specifically, the special class will be a two-parameter family of solutions, $S_0=Z(x^a,\zeta,\overline{\zeta})$ where the parameters are the complex stereographic coordinates on the sphere, S^2, and the null covector, $p_a=\partial_a Z(x^a,\zeta,\overline{\zeta})$ ranges over the entire light-cone at each point x^a as $(\zeta,\overline{\zeta})$ ranges over S^2. Later, in asymptotically flat space–times, we will make a unique choice of this family.

Now assuming that an allowable $Z(x^a,\zeta,\overline{\zeta})$ is known we can produce an arbitrary solution in the following fashion:[8] first, we rescale Z with a constant β and add to Z an arbitrary function, at least once differentiable, of $(\beta,\zeta,\overline{\zeta})$ and then extremize it with respect to the $(\beta,\zeta,\overline{\zeta})$, i.e., we have

$$S=\beta Z(x^a,\zeta,\overline{\zeta})-h(\beta,\zeta,\overline{\zeta}) \tag{2}$$

with

$$\beta\partial_\zeta Z-\partial_\zeta h=0,\quad \beta\partial_{\overline{\zeta}}Z-\partial_{\overline{\zeta}}h=0,\quad Z-\partial_\beta h=0. \tag{3}$$

For the simplicity of the immediate discussion (though the issue is an important one), we assume that the latter three equations can be solved for the $(\beta,\zeta,\overline{\zeta})$ as functions of the x^a, i.e., with

$$(\beta,\zeta,\bar{\zeta})=(B(x^a),\Upsilon(x^a),\bar{\Upsilon}(x^a)), \tag{4}$$

then on substitution back into Eq. (2), the resulting function of x^a also satisfies the Eikonal equation. To see this we have, from (3) that

$$\partial_a S=\beta\partial_a Z+(Z-\partial_\beta h)\partial\beta/\partial x^a+(\beta\partial_\zeta Z-\partial_\zeta h)\partial\zeta/\partial x^a+(\beta\partial_{\bar{\zeta}} Z-\partial_{\bar{\zeta}} h)\partial\bar{\zeta}/\partial x^a=\beta\partial_a Z, \tag{5}$$

which satisfies the Eikonal equation, by the assumption on Z.

The important issue of how to deal with the case when Eq. (3) cannot be solved for $(\beta,\zeta,\bar{\zeta})$ is discussed later in this section.

Though we will not go into the proof one can show that given arbitrary Cauchy data, $S_C(x^i)$ for the Eikonal equation, i.e., a function of three arguments, then it determines the function $h(\beta,\zeta,\bar{\zeta})$. The construction, thus, allows for the general solution to the Eikonal equation.

Remark 1: Since the function $h(\beta,\zeta,\bar{\zeta})$ determines a single solution $S^*(x^a)$ of the Eikonal equation (i.e., a one parameter family of characteristic surfaces) by replacing the $h(\beta,\zeta,\bar{\zeta})$ by the function $h(\beta,\zeta,\bar{\zeta};\eta,\bar{\eta})$ the above construction then produces a two-parameter family of solution of the Eikonal equation, $S^*=Z^*(x^a,\eta,\bar{\eta})$. We thus have that from any special two-parameter family, Z we can construct any other two parameter family that could also be used as the ''special'' family.

We now specialize the construction so that we can obtain any single characteristic surface to be given by $S=u=\text{constant}$; Eqs. (2) and (3) are replaced by the specialization, $\beta=1$,

$$S=Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta})=u, \tag{6}$$

$$\partial_\zeta Z-\partial_\zeta h=0, \quad \partial_{\bar{\zeta}} Z-\partial_{\bar{\zeta}} h=0. \tag{7}$$

Assuming that Eq. (7) can be solved for

$$(\zeta,\bar{\zeta})=(\Upsilon(x^a),\bar{\Upsilon}(x^a)), \tag{8}$$

that any characteristic surface can be obtained by a judicious choice of $h(\zeta,\bar{\zeta})$ can be seen from the argument that if we begin with any spacelike two-surface, $\mathfrak{G}$, parametrized by the same $(\zeta,\bar{\zeta})$, i.e., given by

$$x^a=x_0^a(\zeta,\bar{\zeta}), \tag{9}$$

we can choose

$$h(\zeta,\bar{\zeta})=Z(x_0^a(\zeta,\bar{\zeta}),\zeta,\bar{\zeta}). \tag{10}$$

The resulting characteristic surface. $S=0$, is formed by the null normals to $\mathfrak{G}$ and since any characteristic surface is formed by the null normals to some two-surface we have proven our contention.

Actually, in general, there are lower dimensional regions where Eqs. (7) cannot be solved for the $(\zeta,\bar{\zeta})$ pair. These regions (three dimensional) define the caustics of the solution. The intersection of these caustic regions with any particular level surface of S (big wave front), i.e., with $u=S=\text{const}$ defines the ''big wave front'' singularities[7] of Arnold. The intersection of $u=S$ with a generic three-surface, (e.g., a constant time surface) defines a ''small wave front'' while the intersection of the ''small wave front'' with the caustic three-surface, defines the ''small wave front'' singularities. Though for precise usage we should only refer to the full three dimensional caustic region as the ''caustics,'' we, however, will take the liberty of referring to the singularities of either the big or small wave front as the ''caustics.''

These caustic regions (or on the big or small wave front singularities) which are characterized by the inability to solve for the $(\zeta,\bar{\zeta})$ are simply determined from the implicit function theorem, by the condition

$$\hat{D}\equiv\begin{vmatrix}\frac{\partial^2(Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta}))}{\partial\zeta^2} & \frac{\partial^2(Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta}))}{\partial\zeta\partial\bar{\zeta}} \\ \frac{\partial^2(Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta}))}{\partial\zeta\partial\bar{\zeta}} & \frac{\partial^2(Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta}))}{\partial\bar{\zeta}^2}\end{vmatrix}=0, \tag{11}$$

a condition that will later play a basic role.

To determine the solution S, there is an alternative to solving Eqs. (7) for the $(\zeta,\bar{\zeta})$ that is often more desirable and can be used even when $\hat{D}=0$. Equations (6) and (7) can be considered as defining families of three different 3-surfaces parametrized by the $(\zeta,\bar{\zeta})$ pair. Their intersection defines a family of curves (parametrized by the $(\zeta,\bar{\zeta})$) that are the null geodesics that rule the level surfaces of $u=S$. The equations can always be solved in the following manner: of the four x^a there will be a subset of three of them (say x^i) and the fourth one, say x^* such that

$$x^i=X^i(x^*,u,\zeta,\bar{\zeta}), \tag{12}$$

which are the null geodesics themselves. They define, parametrically, the level surfaces of $u=S$.

An alternative treatment of the null geodesics, Eq. (12), is to introduce a geodesic parameter (not in general an affine parameter) by

$$r=(1+\zeta\bar{\zeta})^2\frac{\partial^2}{\partial\zeta\partial\bar{\zeta}}(Z-h), \tag{13}$$

which, with Eqs. (6) and (7), can be solved for

$$x^a=X^a(r,u,\zeta,\bar{\zeta}) \tag{14}$$

yielding the parametric description of the null geodesics ruling the level surfaces of S. Unfortunately this description can break down at the caustics of S where r sometimes becomes infinite. It nevertheless is a means of treating the geodesics almost everywhere.

Remark 2: The description we have given here for the construction of solutions to the Eikonal equation involves the construction of envelopes of tangent lines to the original two-parameter family of solutions $Z(x^a,\zeta,\bar{\zeta})$ to form the $S(x^a)$. This description and the treatment of the caustics is an example of V. I. Arnold's theory of generating families.[2,7]

III. EIKONAL EQUATION IN ASYMPTOTICALLY FLAT SPACE–TIMES

Before the introduction of a special or fiducial family of null surfaces $S_0=Z(x^a,\zeta,\bar{\zeta})$, we begin with a brief discussion of asymptotically flat space–times. These space–times allow a conformal rescaling of the space–time metric bringing null infinity (the end points of all future-directed null geodesics) into a finite region thereby defining a (null boundary) for the space–time. Though we will not be using the conformal rescaling explicitly, we will however use the language of the conformal boundary. The boundary, referred to as $\mathfrak{I}^+$, can be attained by limiting procedures in the unrescaled space–time. The boundary, which is a null three surface with topology

$R\times S^2$ can be given coordinates $(u,\zeta,\bar{\zeta})$, with u on the R part and the complex stereographic coordinates $(\zeta,\bar{\zeta})$ on the S^2 which label the null generators (geodesics) of $\mathfrak{I}^+$. It is this structure that we will use to obtain the fiducial family of null surfaces.

For *each* generator of $\mathfrak{I}^+$, i.e., for $(\zeta,\bar{\zeta})=(\zeta_0,\bar{\zeta}_0)=\text{constant}$, we choose the one parameter family of *past null cones* having their apexes on that generator. This yields the special solution of the Eikonal equation, $u=Z(x^a,\zeta_0,\bar{\zeta}_0)$. Doing the same for each generator defines for us our unique fiducial family of solutions. $S_0\equiv u=Z(x^a,\zeta,\bar{\zeta})$, the past cones of each point of $\mathfrak{I}^+$. We emphasize that we are describing these null surfaces in the language of the conformal compactification—in the language of the physical space–time they describe the family of all asymptotic plane waves and in the case of flat space they *are* the family of *all* plane waves. As the concepts described here are conformally invariant, the choice of language is at our discretion.

Our special family of solutions

$$u=Z(x^a,\zeta,\bar{\zeta}) \tag{15}$$

has the two important dual meanings: (1) As we just mentioned for fixed point. $(u,\zeta,\bar{\zeta})$, on $\mathfrak{I}^+$, as x^a varies, it defines the past cone of the point and (2) for a fixed value of the x^a, as $(\zeta,\bar{\zeta})$ are varied over the S^2, $u=Z$ defines a two-surface on $\mathfrak{I}^+$, the end points of all the null geodesics leaving x^a. This two-surface is referred to as the *light-cone cut* of the point x^a and is denoted by $\mathfrak{c}(x^a)$. The function $Z(x^a,\zeta,\bar{\zeta})$ will be referred to as the *light-cone cut function*.

Both meanings to $u=Z(x^a,\zeta,\bar{\zeta})$ play a fundamental role in the remainder of this work. The actual determination of $u=Z(x^a,\zeta,\bar{\zeta})$ is quite difficult and up to the present, depends on perturbation arguments that have not yet been completed. We nevertheless will assume that the function $Z(x^a,\zeta,\bar{\zeta})$ is known; we then study several consequences of this knowledge.

Remark 3: Though we will not be concerned with it here, we mention that the $Z(x^a,\zeta,\bar{\zeta})$ codes all conformal information of the space–time metric[4,5,9] *and in fact determines a conformal metric. Furthermore $Z(x^a,\zeta,\bar{\zeta})$, with a scalar function $\Omega(x^a,\zeta,\bar{\zeta})$ that acts as a conformal factor, can be used as the basic variables, replacing the metric, in a reformulation of the Einstein equations.*[4,5,9]

Our goal here is somewhat simpler (though some of the calculations themselves are not simple); we want to study the structure of the singular regions of different surfaces. First, we will show how to construct from the $Z(x^a,\zeta,\bar{\zeta})$, using the techniques of the previous section, the entire light cone $\mathfrak{C}_{x_0}$ of an arbitrary interior point x_0^a and then study its singularities. The light-cone cut $\mathfrak{c}(x_0^a)$ is the intersection of $\mathfrak{C}_{x_0}$ with $\mathfrak{I}^+$, defining a small wave front; its singularities will then be studied. Finally we return to and study the singular regions of the fiducial family of null surfaces, defined by the light-cone cut function, $Z(x^a,\zeta,\bar{\zeta})$ itself.

We first define, in the case of asymptotically flat spaces, several variables that play an important later role. Instead of using the notation of ∂_ζ and $\partial_{\bar{\zeta}}$ for the $(\zeta,\bar{\zeta})$ derivatives, we make use of the *edth* notation, e.g., $\eth Z=(1+\zeta\bar{\zeta})\partial_\zeta Z$. $\eth\bar{\eth}Z=(1+\zeta\bar{\zeta})^2\partial_\zeta\partial_{\bar{\zeta}}Z$ or $\eth^2 Z=\partial_\zeta\{(1+\zeta\bar{\zeta})^2\partial_\zeta Z\}$, etc. We then have by direct calculation from the $Z(x^a,\zeta,\bar{\zeta})$,

$$(1)\quad \omega\equiv\eth Z,\quad \bar{\omega}\equiv\eth Z,\quad \text{the tangent directions to the light-cone cuts,}$$

$$(2)\quad \Lambda\equiv\eth^2 Z,\quad \bar{\Lambda}\equiv\bar{\eth}^2 Z,\quad \text{``accelerations'' along the } (\bar{\zeta},\zeta) \text{ constant curves,}$$

$$(3)\quad R\equiv\eth\bar{\eth}Z \quad \text{extrinsic curvature of the light-cone cuts.} \tag{16}$$

Using this notation, the determination of the caustics, i.e., the vanishing of the determinant $\hat{D}$ from Eq. (11) is equivalent, using Eq. (7), to

$$D \equiv \begin{vmatrix} \eth^2[Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta})] & \eth\bar{\eth}[Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta})] \\ \eth\bar{\eth}[Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta})] & \bar{\eth}^2[Z(x^a,\zeta,\bar{\zeta})-h(\zeta,\bar{\zeta})] \end{vmatrix} = 0. \tag{17}$$

IV. LIGHT-CONES AND THEIR SINGULARITIES

In this section we consider the future light-cone of a point x_0^a, namely, the set of all (future directed) null geodesics that pass through x_0^a. As a three-surface in the four-dimensional space–time, the light-cones in general have singularities that are caused by the focusing effect of the space–time curvature. These singularities are characterized by the vanishing of the geodesic deviation vector associated with neighboring geodesics on the light-cone and are what we have been referring to as the caustics of the null surface. It is our purpose here to first find these light-cones and then describe their singularities in terms of the light-cone cut function, $Z(x^a,\zeta,\bar{\zeta})$.

As we pointed out earlier (Sec. II), given a two-parameter family of solutions to the Eikonal equation, $Z(x^a,\zeta,\bar{\zeta})$, any characteristic surface can be constructed by adding a term that depends only on the parameters $(\zeta,\bar{\zeta})$; i.e.,

$$S(x^a,\zeta,\bar{\zeta}) = Z(x^a,\zeta,\bar{\zeta}) - h(\zeta,\bar{\zeta})$$

and extremizing with respect to the two parameters. If we choose

$$h(\zeta,\bar{\zeta}) = Z(x_0^a,\zeta,\bar{\zeta}) \equiv Z_0(\zeta,\bar{\zeta}), \tag{18}$$

then the level surface obtained from

$$S = 0 = Z(x^a,\zeta,\bar{\zeta}) - Z(x_0^a,\zeta,\bar{\zeta}) \tag{19}$$

with the extremal conditions

$$\begin{aligned} &\eth[Z(x^a,\zeta,\bar{\zeta}) - Z(x_0^a,\zeta,\bar{\zeta})] = 0, \\ &\bar{\eth}[Z(x^a,\zeta,\bar{\zeta}) - Z(x_0^a,\zeta,\bar{\zeta})] = 0, \end{aligned} \tag{20}$$

describes the light-cone from the point x_0^a. To see this, we first remember that this construction yields characteristic surfaces, then we see that the surface does go through the point $x^a = x_0^a$ and coincides on $\mathfrak{I}^+$, with the light-cone cut of x_0^a. Finally if we take its gradient, i.e.,

$$\partial_a S|_{x=x_0} = \partial_a[Z(x^a,\zeta,\bar{\zeta}) - Z(x_0^a,\zeta,\bar{\zeta})]|_{x=x_0} = \partial_a Z(x^a,\zeta,\bar{\zeta})|_{x=x_0} \equiv p_a(x_0^a,\zeta,\bar{\zeta}), \tag{21}$$

we see that it ranges over the entire light-cone at x_0^a.

A Caveat: We have assumed that the cut function, $u = Z(x^a,\zeta,\bar{\zeta})$, for fixed x^a is a single valued function on $\mathfrak{I}^+$. In fact, in general, this is not true; most often there will be regions on $\mathfrak{I}^+$ where it will be multivalued and it must be given as several different ''sheets'' in different $(\zeta,\bar{\zeta})$ patches. Though this does not present obstacles in principle, it does present technical difficulties in implementation. Then Eqs. (19) and (20) must be repeated on the different sheets. A way to avoid this difficulty is to describe the light-cone cut function and the light-cone cut itself parametrically, i.e., to write it as $u = U(x^a,\lambda,\bar{\lambda})$ and $\zeta = \Gamma(x^a,\lambda,\bar{\lambda})$ with single-valued functions. For simplicity of presentation we will, for the moment, continue to treat the cut function as if it were single valued.

If, to the set of Eqs. (19) and (20), we add, from Eq. (13), the equation

$$r = (1+\zeta\bar{\zeta})^2 \frac{\partial^2}{\partial\zeta\partial\bar{\zeta}}(Z - Z_0) \equiv \eth\bar{\eth}(Z(x^a,\zeta,\bar{\zeta}) - Z(x_0^a,\zeta,\bar{\zeta})), \tag{22}$$

they implicitly define all the null geodesics of the light-cone $\mathfrak{C}_{x_0}$, i.e., they determine

$$x^a = X^a(x_0^a, r, \zeta, \overline{\zeta}). \tag{23}$$

If the geodesic goes from x_0^a to $\mathfrak{I}^+$ without encountering a caustic then r goes from 0 to infinity along that geodesic; if however it does encounter a caustic before $\mathfrak{I}^+$, r then becomes infinite before $\mathfrak{I}^+$.

The location of the caustics of $\mathfrak{C}_{x_0}$ (or the conjugate points to x_0^a) is given by the vanishing of D from Eq. (17), with $h = Z(x_0^a, \zeta, \overline{\zeta})$;

$$D \equiv \begin{vmatrix} \eth^2[Z(x^a,\zeta,\overline{\zeta}) - Z(x_0^a,\zeta,\overline{\zeta})] & \eth\bar{\eth}[Z(x^a,\zeta,\overline{\zeta}) - Z(x_0^a,\zeta,\overline{\zeta})] \\ \eth\bar{\eth}[Z(x^a,\zeta,\overline{\zeta}) - Z(x_0^a,\zeta,\overline{\zeta})] & \bar{\eth}^2[Z(x^a,\zeta,\overline{\zeta}) - Z(x_0^a,\zeta,\overline{\zeta})] \end{vmatrix} = 0, \tag{24}$$

or, with definitions (16),

$$D = (\Lambda - \Lambda_0)(\overline{\Lambda} - \overline{\Lambda}_0) - (R - R_0)^2 = 0. \tag{25}$$

We have thus been able to express the location of the caustics of an arbitrary light-cone in terms of derivatives of the cut function $Z(x^a, \zeta, \overline{\zeta})$. Given a fixed point x_0^a and a particular null geodesic (labeled by $(\zeta, \overline{\zeta})$), the curvature and ''acceleration'' of its light-cone cut is given by $(R_0(x_0^a, \zeta, \overline{\zeta}), \Lambda_0(x_0^a, \zeta, \overline{\zeta}))$ while for an arbitrary point along that geodesic it would be $(R(x^a, \zeta, \overline{\zeta}), \Lambda(x^a, \zeta, \overline{\zeta}))$. D which begins as zero at $r = 0$, does not vanish any other place along a geodesic that does not encounter a caustic but does go to zero at the caustic. There are special geodesics $(\zeta, \overline{\zeta})_c$ which meet the caustic on $\mathfrak{I}^+$. For this limiting case, it is difficult to study the behavior of Eq. (25) since $\Lambda \Rightarrow 0$ and $R \Rightarrow \infty$, the flat-space limits, which applies here since the points x^a near $\mathfrak{I}^+$ are in the very weak field region and R_0 and Λ_0 are infinite (see next section). Other techniques for this study are needed. (See Fig. 1, the light-cone with the crossovers and cusps.)

V. THE LIGHT-CONE CUTS AND THEIR SINGULARITIES

As we saw earlier, the cut function, $u = Z(x^a, \zeta, \overline{\zeta})$ has the dual meaning of being the past light-cones of the points $(u, \zeta, \overline{\zeta})$ of $\mathfrak{I}^+$ and representing the light-cone cut of an interior point, x^a. Fixing the interior point $x^a = x_0^a$, we studied, in the last section, its light-cone and saw that we could locate its caustics but as the caustics approached $\mathfrak{I}^+$ difficulties developed. We wish to study the singularities of the light-cone cuts by an alternative method.

First of all, if we assume that *all* the null geodesics leaving x_0^a arrive at $\mathfrak{I}^+$ without encountering a caustic then the cut function, $u = Z(x^a, \zeta, \overline{\zeta})$, will describe a single-valued smooth 2-surface on $\mathfrak{I}^+$. If however some did encounter caustics then the cut-surface will only be piecewise smooth and will have, in general, self-intersections. The appropriate way to describe the cut is not through the cut function but instead via the mapping of the space of null directions at x_0^a, i.e., at $S^2(x_0^a)$, coordinatized by $(\lambda, \overline{\lambda})$, onto $\mathfrak{I}^+$. It would be given by the relations

$$(u, \zeta, \overline{\zeta}) = (U(x_0^a, \lambda, \overline{\lambda}), \Gamma(x_0^a, \lambda, \overline{\lambda}), \overline{\Gamma}(x_0^a, \lambda, \overline{\lambda})), \tag{26}$$

which are just the ''end-points'' or boundary points of the null geodesics originating at x_0^a in the $(\lambda, \overline{\lambda})$ directions. (If the $(\lambda, \overline{\lambda}) \Rightarrow (\zeta, \overline{\zeta})$ were invertible, then one would have the smooth case, $u = U(x_0^a, \lambda, \overline{\lambda}) = Z(x_0^a, \zeta, \overline{\zeta})$.)

To obtain a clearer picture the light-cone cut can, in some sense, be thought of as an infinitely late ''small wave front.'' The ''early'' wave fronts on the future lightcone of x_0^a are smooth deformations of spheres, but they may become singular at sufficiently late times, from the focusing

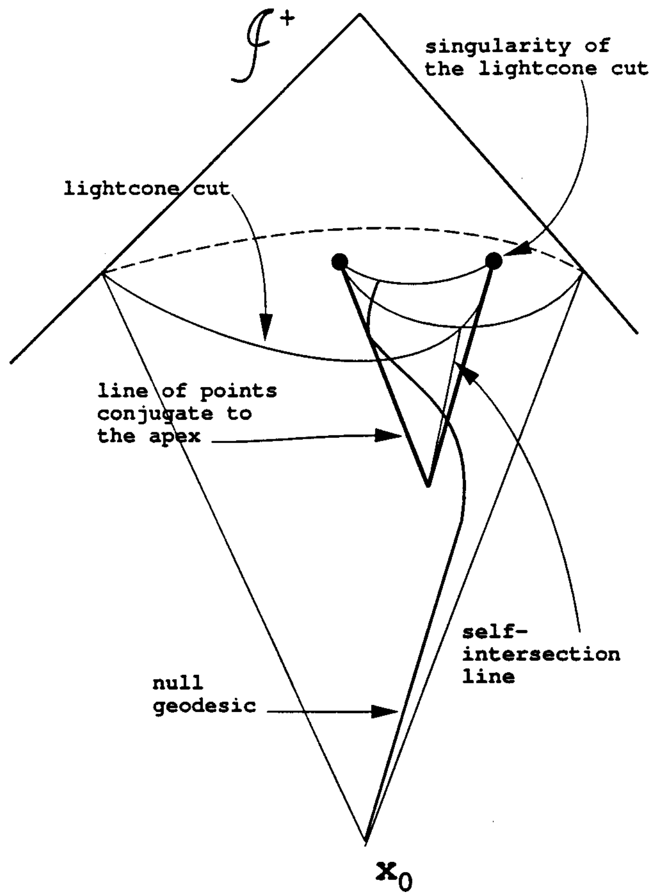

FIG. 1. The future light-cone and light-cone cut of a point x_0.

due to curvature. Therefore, light-cone cuts of (early) points in space–time are generically singular two-surfaces in three-dimensions, and they must exhibit the standard stable singularities, the cusp ridges, and swallowtails. In this view, for a fixed point x_0^a, a singularity in the light-cone cut would be a conjugate point to x_0^a. Generically, because singularities of two-surfaces lie on curves, the singularities of a light-cone cut would single out a one-parameter set $(\zeta(s),\overline{\zeta}(s))$ of null geodesics in the future light-cone for which the apex is a focal point.

Because Eq. (26) arises from the Hamiltonian evolutions (null geodesic flow) the map is a Legendre map and we can use the general theory of Legendre submanifolds and Legendre maps of Arnold and his colleagues in order to have a description of the location of the singularities of the light-cone cuts.[3,6] A two-dimensional surface in a three-manifold which is obtained by a Legendre map can always be represented as the projection of a smooth 2-surface (a Legendre submanifold) in a five-dimensional space, with the singularities located by the singularities of the projection. In other words, there exists a way to ''unfold'' a singular surface by adding two dimensions to the space where the surface lives. In this view, one of the three original dimensions (the u coordinate of $\mathfrak{I}^+$) is singled out from the remaining two-dimensional space; the two-dimensions, $(\zeta,\overline{\zeta})$, are to be considered as a configuration space. The two added dimensions consist of the two-dimensional cotangent space over the configuration space. Thus the enlarged five-dimensional space on which our surfaces ''unfold'' consists of points $(\zeta,\overline{\zeta},p_\zeta,p_{\overline{\zeta}},u)$, a contact bundle over the sphere. It is preferable to use *real* coordinates, and later translate the results in terms of our complex coordinates. Thus, in the following we assume that we have real coordinates (q^1,q^2) on the sphere, which can be taken to be the real and imaginary parts of ζ, and their corresponding momenta as (p_1,p_2).

A smooth ''unfolding'' is generically represented in terms of a smooth generating function $G(q^1,p_2)$. The points (q^1,q^2,p_1,p_2,u) that lie on such unfolding are given by

$$q^2=-\frac{\partial G(q^1,p_2)}{\partial p_2}, \tag{27a}$$

$$p_1=\frac{\partial G(q^1,p_2)}{\partial q^1}, \tag{27b}$$

$$u=G(q^1,p_2)+p_2q^2, \tag{27c}$$

and arbitrary values for q^1 and p_2. This is the expression of a two-dimensional surface within a five-dimensional space, parametrized by (q^1,p_2).

Remark 4: Note that from the general theory, there must be an invertible relationship between the parametrization (q^1,p_2) *of the Legendre submanifold and the directions* $(\lambda,\bar{\lambda})$.

A projection of this surface down to the space (q^1,q^2,u) is parametrically represented by a map $(q^1,p_2)\rightarrow(q^1,q^2(q^1,p_2),u(q^1,p_2))$ which breaks down at points where the Jacobian matrix

$$\left\|\begin{matrix} \dfrac{\partial q^1}{\partial q^1} & \dfrac{\partial q^1}{\partial p_2} \\ \dfrac{\partial q^2}{\partial q^1} & \dfrac{\partial q^2}{\partial p_2} \\ \dfrac{\partial u}{\partial q^1} & \dfrac{\partial u}{\partial p_2} \end{matrix}\right\| = \left\|\begin{matrix} 1 & 0 \\ -\dfrac{\partial^2 G}{\partial q^1\partial p_2} & -\dfrac{\partial^2 G}{\partial p_2^2} \\ \dfrac{\partial G}{\partial q^1}-p_2\dfrac{\partial^2 G}{\partial q^1\partial p_2} & -p_2\dfrac{\partial^2 G}{\partial p_2^2} \end{matrix}\right\| \tag{28}$$

drops rank, from 2 to 1 or 0. The drop in rank takes place where the three 2×2 determinants vanish, namely, where

$$\frac{\partial^2 G}{\partial p_2^2}=0, \tag{29}$$

$$p_2\frac{\partial^2 G}{\partial p_2^2}=0, \tag{30}$$

$$\frac{\partial G}{\partial q^1}\frac{\partial^2 G}{\partial p_2^2}=0. \tag{31}$$

Clearly all three equations can be satisfied if and only if

$$K(q^1,p_2)\equiv\frac{\partial^2 G(q^1,p_2)}{\partial p_2^2}=0. \tag{32}$$

Thus Eq. (32) locates the curve $K(q^1,p_2)=0$ in the (q^1,p_2) parameter space and hence, via Eqs. (27a) and (27c), it locates the singular points on the surface. Equation (32) also expresses the location of points where Eq. (27a) fails to be invertible; namely, if we think of Eq. (27a) as implicitly defining $p^2=h(q^1,q^2)$, then h fails to be differentiable there. (From the drop in rank, it is straightforward to see that $\partial h/\partial q^2$ blows up. See Eq. (39a) below.)

In order to translate this treatment into our complex notation, we pass from (q^1,q^2) to the complex coordinates $\zeta=\frac{1}{2}(q^1+iq^2)$ and reinterpret

$$u=G(x^a,q^1,h(q^1,q^2))+q^2h(q^1,q^2)$$

as our cut function $u=Z(x^a,\zeta,\overline{\zeta})$, where x^a are fixed parameters and play no role in the discussion of this section. We can then express Eq. (32) in terms of derivatives of Z in the following manner.

Beginning with the function $\Lambda\equiv\eth^2Z$, we express it parametrically in terms of (q^1,p_2),

$$\Lambda=2(1+\zeta\overline{\zeta})\overline{\zeta}\frac{\partial Z}{\partial\zeta}+(1+\zeta\overline{\zeta})^2\frac{\partial^2Z}{\partial\zeta^2}, \tag{33}$$

where

$$\frac{\partial}{\partial\zeta}=\left.\frac{\partial}{\partial q^1}\right|_{q^2}-i\left.\frac{\partial}{\partial q^2}\right|_{q^1}. \tag{34}$$

Carrying through the calculation, which involves several implicit differentiations, we first arrive at

$$\omega=(1+\zeta\overline{\zeta})\left(\frac{\partial G}{\partial q^1}-ip_2\right),\quad \overline{\omega}=(1+\zeta\overline{\zeta})\left(\frac{\partial G}{\partial q^1}+ip_2\right), \tag{35}$$

where

$$\zeta=\frac{1}{2}\left(q^1-i\frac{\partial G}{\partial p_2}\right),\quad \overline{\zeta}=\frac{1}{2}\left(q^1+i\frac{\partial G}{\partial p_2}\right). \tag{36}$$

Then

$$\Lambda=2\overline{\zeta}\omega+(1+\zeta\overline{\zeta})^2\left\{\frac{\partial^2G}{\partial(q^1)^2}-\left(\frac{\partial^2G}{\partial q^1\partial p_2}-i\right)^2\left(\frac{\partial^2G}{\partial p_2^2}\right)^{-1}\right\}. \tag{37}$$

Similarly, we obtain a parametric expression for $R\equiv\bar{\eth}\eth Z=(1+\zeta\overline{\zeta})^2(\partial^2Z/\partial\zeta\partial\overline{\zeta})$ in the form

$$R=(1+\zeta\overline{\zeta})^2\left\{\frac{\partial^2G}{\partial(q^1)^2}-\left(1+\left(\frac{\partial^2G}{\partial q^1\partial p_2}\right)^2\right)\left(\frac{\partial^2G}{\partial p_2^2}\right)^{-1}\right\}. \tag{38}$$

In deriving (37) and (38), the following were needed:

$$\frac{\partial h}{\partial q^1}=-\frac{\partial^2G}{\partial q^1\partial p_2}\left(\frac{\partial^2G}{\partial p_2^2}\right)^{-1},\quad \frac{\partial h}{\partial q^2}=-\left(\frac{\partial^2G}{\partial p_2^2}\right)^{-1}, \tag{39}$$

which are obtained by taking derivatives $\partial/\partial q^1|_{q^2}$ and $\partial/\partial q^2|_{q^1}$ of Eq. (27a).

From (37) and (38) we can see that both Λ and R diverge at points where Eq. (32) is satisfied, and only at those points, since G is assumed to be smooth. Therefore, we can locate the singular points (a curve), $(\zeta(s),\overline{\zeta}(s))$ of light-cone cuts by either of the conditions,

$$P(x_0^a,\zeta,\overline{\zeta})\equiv\frac{1}{\bar{\eth}\eth Z(x_0^a,\zeta,\overline{\zeta})}=0, \tag{40}$$

$$L(x_0^a,\zeta,\overline{\zeta})\equiv\frac{1}{\eth^2Z(x_0^a,\zeta,\overline{\zeta})}=0 \tag{41}$$

for given values of x_0^a.

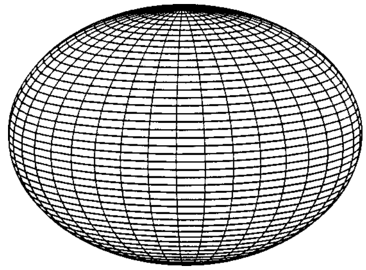

FIG. 2. A regular light-cone cut.

We interpret this result as follows. The light-cone cut represents a wave front that has progressed out to infinity, tracing the future light-cone of the point x_0^a. For a class of points x_0^a (at least sufficiently early), the wave front starts out spherical, but there is a time at which it becomes self-intersecting. Late wave fronts have singularities which represent the location of points conjugate to x_0^a. When the wave front reaches infinity, the points conjugate to the apex lie at infinity and form the singularities of the light-cone cut (See Fig. 2, for a smooth light-cone cut and Fig. 3, for a generic light-cone cut with cusp ridges and swallowtails.)

Finally note that the vanishing of $P(x_0^a,\zeta,\overline{\zeta})$ and $L(x_0^a,\zeta,\overline{\zeta})$ are not inconsistent with Eq. (25) of the previous section where as $\mathfrak{I}^+$ is approached. $\Lambda \to 0$. $R\to\infty$ and the $\Lambda_0\to\infty$ and $R_0\to\infty$.

VI. SINGULARITIES OF THE PAST LIGHT-CONES FROM $\mathfrak{I}^+$

Up to this point we have simply assumed that we had the three parameter family of null surfaces (or equivalently the two parameter family of solutions to the Eikonal equation) that we called the fiducial family or the light-cone cut function, namely, $u=Z(x^a,\zeta,\overline{\zeta})$, with $(u,\zeta,\overline{\zeta})$ constant. We never raised the issue of the location of their singularities until now. The reason was that, to locate them, requires a different technique, namely the use of pairs of geodesic deviation vectors (Jacobi fields) and their associated area element. It will be the vanishing of the area element (obtained from the Jacobi fields) along a geodesic that locates the singularities. We begin by returning to certain structures obtainable from the light-cone cut function $Z(x^a,\zeta,\overline{\zeta})$ that were defined earlier; namely,

$$u=Z(x^a,\zeta,\overline{\zeta}), \tag{42}$$

which represents the past light-cones from all points on $\mathfrak{I}^+$,

$$\omega=\eth Z(x^a,\zeta,\overline{\zeta}), \quad \overline{\omega}=\bar{\eth} Z(x^a,\zeta,\overline{\zeta}), \tag{43}$$

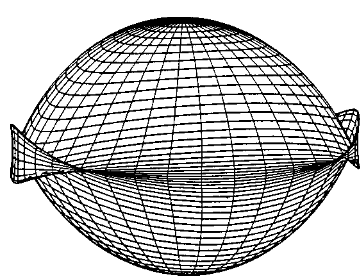

FIG. 3. A singular light-cone cut, showing cusp ridges and swallowtails.

which label the null geodesics leaving the point $(u,\zeta,\overline{\zeta})$ of $\mathfrak{I}^+$. From the compactified point of view they are the (stereographic) angles labeling the directions from the past light-cone while from the physical space point of view they are the ''distance'' between the asymptotic parallel geodesics,

$$R=\bar{\eth}\eth Z(x^a,\zeta,\overline{\zeta}), \tag{44}$$

which defines an ''optical distance'' or geodesic parameter (not affine) along the geodesics $(u,\omega,\overline{\omega})=\text{const}$. As we mentioned earlier, geometrically R is a curvature of the cut.

Using the notation

$$\theta^i=\theta^i(x^a,\zeta,\overline{\zeta})\equiv(u,\omega,\overline{\omega},R), \quad \text{with} \quad (i=0,+,-,1), \tag{45}$$

Eq. (45) can be interpreted either as a coordinate transformation, $\theta^i \Leftrightarrow x^a$, for every fixed value of the two parameters $(\zeta,\overline{\zeta})$ or simply as the introduction of four scalar functions parametrized by the $(\zeta,\overline{\zeta})$. We will make extensive use of the transformation interpretation though care must be taken in the regions where the Jacobian either vanishes or diverges. One might even expect that the troublesome region will be where the (big) wave front singularities develop.

In generic space–times, the presence of the curvature, Weyl or Ricci-type, has a *focusing effect* on parallel beams of light.[10] Thus, generically, two neighboring null geodesics in our asymptotically parallel congruences meet at some point, which means that our coordinate system breaks down by assigning two different labels to the same space–time point.

We will describe two alternative approaches to the region of breakdown.

(1) We can calculate the Jacobian of Eq. (45) most easily by returning to the description of the cut function Z by the generating function, $G(x^a,q^1,p_2)$ of the previous section, $Z=G+q^2p_2$. By a completely straightforward calculation (using MATHEMATICA to calculate the determinant) we find that

$$\left|\frac{\partial\theta^i}{\partial x^a}\right| \propto \left(\frac{\partial^2 G(x^a,q^1,p_2)}{\partial p_2^2}\right)^{-1}, \tag{46}$$

$$\left|\frac{\partial x^a}{\partial\theta^i}\right| \propto \left(\frac{\partial^2 G(x^a,q^1,p_2)}{p_2^2}\right) \tag{47}$$

so that the Jacobian breaks down precisely at the comparable point where the light-cone cut had its singularities.

Remark 5: In the previous section we saw that for fixed x^a, but varying the $(\zeta,\overline{\zeta})$, the functions $R(x^a,\zeta,\zeta)$ and $\Lambda(x^a,\zeta,\overline{\zeta})$ both diverged at the singularities of the light-cone cut. We can now see that for fixed $(\zeta,\overline{\zeta})$ but varying the point x^a along a null geodesic, the same functions diverge at the caustic of the past light-cone.

(2) In the second approach, we derive an explicit algebraic condition to locate these regions, by finding the points where a geodesic deviation vector vanishes. Our present derivation is in great measure a reinterpretation of an earlier derivation due to Kozameh and Newman,[11] reproduced here in current notation in order to maintain the unity of the present work.

By (in principle) inverting Eq. (45) we obtain

$$x^a=X^a(\theta^i,\zeta,\overline{\zeta})=X^a(u,\omega,\overline{\omega},R,\zeta,\overline{\zeta}), \tag{48}$$

which for fixed values of $(u,\zeta,\overline{\zeta})$ is the parametric form of the past cone of $\mathfrak{I}^+$ and for fixed values of $(u,\omega,\overline{\omega},\zeta,\overline{\zeta})$ it is the parametric form for the null geodesics on the cone each labeled by $(\omega,\overline{\omega})$.

Of prime importance to us are the connecting vectors to the null geodesics that are on the past null cone. Two connecting vectors (from which all others can be constructed) are given by

$$M^a = \frac{\partial X^a}{\partial \omega}, \quad \overline{M}^a = \frac{\partial X^a}{\partial \overline{\omega}}. \tag{49}$$

We are interested in the area A constructed from M^a and $\overline{M}^a$. Taking a pair of (complex) spacelike unit vectors m^a and $\overline{m}^a$ ($g_{ab}m^a m^b = 0$, $g_{ab}\overline{m}^a\overline{m}^b = 0$, $g_{ab}\overline{m}^a m^b = -1$), that are parallel transported along the null geodesics. M^a and $\overline{M}^a$ can be written as

$$M^a = \xi m^a + \overline{\eta m}^a, \quad \overline{M}^a = \overline{\xi}\overline{m}^a + \eta m^a, \tag{50}$$

so that the ''area'' form is

$$M^{[a}\overline{M}^{b]} = (\xi\overline{\xi} - \eta\overline{\eta}) m^{[a}\overline{m}^{b]} \equiv A m^{[a}\overline{m}^{b]}. \tag{51}$$

From this we see that

$$A^2 \equiv (g_{ab}M^a\overline{M}^b)^2 - (g_{ab}M^a M^b)(g_{ab}\overline{M}^a\overline{M}^b) \equiv (M\cdot\overline{M})^2 - (M\cdot M)(\overline{M}\cdot\overline{M}). \tag{52}$$

Our task (which requires a bit of preparation) is to express the $M\cdot\overline{M}$ and $M\cdot M$ in terms of $Z(x^a,\zeta,\overline{\zeta})$ and its derivatives. We choose the one-form basis

$$\theta^i_a \equiv \partial_a\theta^i = (\partial_a Z, \partial_a\omega, \partial_a\overline{\omega}, \partial_a R) \equiv (\theta^0_a, \theta^+_a, \theta^-_a, \theta^1_a) \tag{53}$$

and the dual vectors

$$\theta^a_i = (\theta^a_0, \theta^a_+, \theta^a_-, \theta^a_1) = (\partial X^a/\partial u, \partial X^a/\partial\omega, \partial X^a/\partial\overline{\omega}, \partial X^a/\partial R) \tag{54}$$

which satisfy

$$\theta^a_i\theta^j_a = \delta^j_i, \quad \theta^a_i\theta^i_c = \delta^a_c. \tag{55}$$

From the one-form basis θ^i_a, using the space–time metric, g^{ab}, one can express the dual basis set by

$$\theta^a_i = g^{ac}\theta^j_c\eta_{ji} \quad \text{or} \quad \theta^i_c = g_{ac}\theta^a_j\eta^{ij}, \tag{56}$$

where

$$\eta_{ij} = \theta^a_i\theta^c_j g_{ac}, \quad \eta^{ij} = \theta^i_a\theta^j_c g^{ac}. \tag{57}$$

Returning to the computation of the area, we have for the tangent vector to the geodesics,

$$\theta^a_1 \equiv L^a = \partial X^a/\partial R, \tag{58}$$

and from the geodesic deviation vectors, $M^a = \partial X^a/\partial\omega = \theta^a_+$ and $\overline{M}^a = \partial X^a/\partial\overline{\omega} = \theta^a_-$ that

$$M\cdot\overline{M} = \eta_{+-}, \quad M\cdot M = \eta_{++}, \quad \overline{M}\cdot\overline{M} = \eta_{--}. \tag{59}$$

Remark 6: Note that $l^a \equiv \Omega^2 L^a$ is the affine parametrized tangent vector to the geodesics. [See (Eq. (63) below for the definition of Ω.]

The calculation of the three η's though lengthy, is fairly straightforward; It is found from the inverse of η^{ij} [i.e., from the second version of Eq. (57)]. The components of η^{ij} are found by beginning with

$$\eta^{00}=\theta_a^0\theta_c^0 g^{ac}=g^{ac}\partial_a Z\partial_c Z=0 \tag{60}$$

which vanishes by definition. By applying the operators ð and $\bar{ð}$ several times to Eq. (60) one finds[4] for the relevant components of η^{ij} [see Eq. (16) for definitions]

$$\eta^{00}=0, \tag{61}$$

$$\eta^{0+}=0, \quad \eta^{0-}=0, \tag{62}$$

$$\eta^{01}\equiv\Omega^2=g^{ac}\partial_a Z ð\bar{ð}(\partial_c Z)=g^{ac}\partial_a Z\partial_c R, \tag{63}$$

$$\eta^{++}=-\Omega^2 L^a ð^2(\partial_a Z)=-\Omega^2 L^a\partial_a\Lambda=-\Omega^2\partial\Lambda/\partial R, \tag{64}$$

$$\eta^{--}=-\Omega^2 L^a \bar{ð}^2(\partial_a Z)=-\Omega^2 L^a\partial_a\overline{\Lambda}=-\Omega^2\partial\overline{\Lambda}/\partial R, \tag{65}$$

$$\eta^{-+}=-\Omega^2, \tag{66}$$

which in turn leads to

$$M\cdot M=\eta_{++}=-\frac{1}{\Omega^2\left(1-\frac{\partial\Lambda}{\partial R}\frac{\partial\overline{\Lambda}}{\partial R}\right)}, \tag{67}$$

$$M\cdot\overline{M}=\eta_{+-}=-\frac{\frac{\partial\overline{\Lambda}}{\partial R}}{\Omega^2\left(1-\frac{\partial\Lambda}{\partial R}\frac{\partial\overline{\Lambda}}{\partial R}\right)}. \tag{68}$$

The area then is

$$A^2=\frac{1}{\Omega^4\left(1-\frac{\partial\Lambda}{\partial R}\frac{\partial\overline{\Lambda}}{\partial R}\right)}. \tag{69}$$

This expression for A tells us several things; first of all to keep A real we must have the inequality

$$\left|\frac{\partial\Lambda}{\partial R}\right|\leqslant 1, \tag{70}$$

and we learn that Ω must diverge at the singularity given by $A=0$.

We have thus learned in this section that the singularities of the past light-cones from $\mathfrak{I}^+$ can be characterized by one of several methods:

(1) Using the generating function $G(x^a,q^1,p_2)$, the singularities are given by the vanishing of the Jacobian of the transformation (45), i.e., by

$$\frac{\partial^2 G(x^a,q^1,p_2)}{\partial p_2^2}=0. \tag{71}$$

(2) This, in turn, tells us (from the previous section) that both $R(x^a,\zeta,\overline{\zeta})$ and $\Lambda(x^a,\zeta,\overline{\zeta})$ diverge as the singularities are approached.

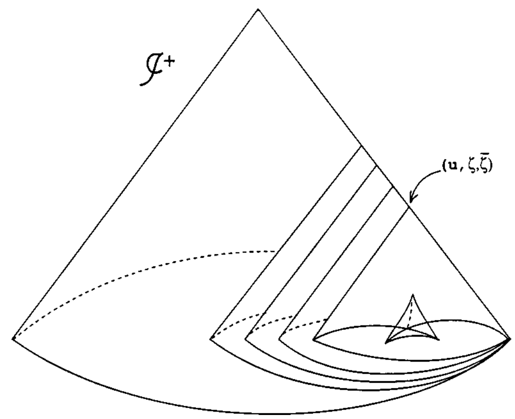

FIG. 4. The foliation of space–time by past light-cones from points at null infinity.

(3) From the geodesic deviation argument

$$\Omega \rightarrow \infty \tag{72}$$

as the singularity is approached.

(4) From Eqs. (69) and (70) we learn that $|\partial\Lambda/\partial R|$ must be bounded but we can not see what is its behavior is as the singularity is approached. However on the basis of several examples, e.g., Ref. 1, where $|\partial\Lambda/\partial R| \rightarrow 1$ it appears to be reasonable to expect that this result might be true in general. If so, then we would have that $[1-(\partial\Lambda/\partial R)(\partial\bar{\Lambda}/\partial R)] \rightarrow 0$ as the singularity is approached. In turn, from Eq. (69), we would gain some information about how fast both Ω and $[1-(\partial\Lambda/\partial R)(\partial\bar{\Lambda}/\partial R)]$ approach their limits. (See Fig. 4, a past light-cone from $\mathfrak{I}^+$.)

VII. SUMMARY AND CONCLUSIONS

In this work we have studied the kinematics or general structure of several different classes of surfaces (associated with surface forming null geodesic congruences) in asymptotically flat Lorentzian space–times, namely, the future light cones of interior points, $\mathfrak{C}_x$; the intersection of $\mathfrak{C}_x$ with $\mathfrak{I}^+$, i.e., the light-cone cuts, $\mathfrak{c}(x^a)$; and the past light-cones from points $(u,\zeta,\bar{\zeta})$ on $\mathfrak{I}^+$.

These surfaces, which in general have singularities, are closely related to each other; in particular there is a close association between their singularities. As was pointed out earlier, for the future light cones $\mathfrak{C}_{x_0}$ with an apex x_0^a that is sufficiently early in time, the small wave fronts begin spherical but as they evolve they become self-intersecting and develop singularities (the stable one being cusp ridges and swallowtails[6]) which represent the conjugate points to x_0^a. The limit, in the asymptotic future, of these small wave fronts is the light-cone cut $\mathfrak{c}(x_0^a)$; the singularities of $\mathfrak{c}(x_0^a)$ being the points conjugate to the apex. They are also the intersection of the singularities of $\mathfrak{C}_{x_0}$ with $\mathfrak{I}^+$ (see Figs. 1, 2, and 3.)

Alternatively (an example of the reciprocity theorem of Penrose and Sachs[11,12]), the singularities of light-cone cuts must be related to the singularities of the past light cones from points at infinity. The singularities of light-cone cuts are interpreted as singling out the null geodesics leaving $\mathfrak{I}^+$ which are conjugate to or focus at x_0^a. These null geodesic belong to two congruences of interest to us. First, they belong to the future light cone of the point x_0^a, and second, they belong to the past light cone of the point $(u,\zeta,\bar{\zeta})$ of $\mathfrak{I}^+$ reached by the first set. The light-cone cut function, with the vanishing of either $P(x^a,\zeta,\bar{\zeta})$ or $L(x^a,\zeta,\bar{\zeta})$, locate both the singularities of the light-cone cut and the interior points conjugate to points on $\mathfrak{I}^+$ (see Fig. 4).

Most of the kinematic issues raised here are, we believe, now reasonably well understood. {It still would be of considerable interest to determine the behavior of $[1-(\partial\Lambda/\partial R)(\partial\bar{\Lambda}/\partial R)]$, in the neighborhood of the caustics.} Our interest now is to apply these kinematic insights to the study of null surfaces (specifically, light-cones) in vacuum Einstein spaces. Though there is a formalism[4,5] in which the Einstein equations have been rewritten as differential equations for the cut function, $Z(x^a,\zeta,\overline{\zeta})$ and $\Omega(x^a,\zeta,\overline{\zeta})$ (aside from some very special cases), the equations have been difficult to deal with because of the difficulty of treating the caustics, which are ubiquitous. We feel that the situation has changed; we now know how to identify the presence of the caustics in terms of both R and Λ. {The reason for our interest in the term $[1-(\partial\Lambda/\partial R)(\partial\overline{\Lambda}/\partial R)]$ is that it arises frequently in denominators of the field equations and we would like to know if it always tends to zero at a caustic.} We have also realized that it probably will be very advantageous to use the representation of $Z(x^a,\zeta,\overline{\zeta})$ by

$$Z(x^a,\zeta,\overline{\zeta})=G(x^a,q^1,p_2)+q^2p_2 \tag{73}$$

with $q^2=-\partial G/\partial p_2$, $\zeta=1/2(q^1+iq^2)$ (see Secs. V and VI). Our immediate goals are first to find the behavior of $[1-(\partial\Lambda/\partial R)(\partial\overline{\Lambda}/\partial R)]$ near caustics and then rewrite the field equations in terms of the $G(x^a,q^1,p_2)$ rather than $Z(x^a,\zeta,\overline{\zeta})$.

ACKNOWLEDGMENTS

The authors thank Carlos Kozameh, Roger Penrose, Helmut Friedrich, and Juergen Ehlers for valuable discussions. E.N. and S.F. thank the NSF for support under Grant No. PHYS 92-05109. G.S.O. acknowledges the financial support from the Sistema Nacional de Investigoadores and from the Consejo nacional de Ciencia y Tecnologí́a (CONACyT), Mé́xico.

[1] S. Frittelli, E. T. Newman, and G. Silva-Ortigoza, ''The Eikonal Equation in Flat Space: Null Surfaces and Their Singularities. I'' (submitted).
[2] V. I. Arnold, S. M. Gusein-Zade, A. N. Varchenko, *Singularities of Differentiable Maps* (Birkhauser, Boston, 1985), Vol. I.
[3] V. I. Arnold, *Catastrophe Theory* (Springer, Berlin, 1986).
[4] S. Frittelli, C. Kozameh, and E. Newman, J. Math. Phys. **36**, 4984 (1995).
[5] S. Frittelli, C. Kozameh, and E. Newman, J. Math. Phys. **36**, 6397 (1995).
[6] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, Berlin, 1980).
[7] *Dynamical Systems*, edited by V. I. Arnold and S. P. Novikov (Springer, Berlin, 1990), Vol. IV.
[8] See the section on Hamilton–Jacobi Theory, in L. Landau and Lifschitz, *Classical Mechanics* (Pergamon, Headington Hill Hall, Oxford, 1960).
[9] S. Frittelli, C. Kozameh, and E. Newman, Phys. Rev. D **56**, 4729 (1997).
[10] S. Hawking and G. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973).
[11] C. Kozameh and E. Newman, *A Nonlocal Approach to the Vacuum Maxwell, Yang–Mills and Einstein Equations*, in *Topological Properties and Global Structure of Space–Time*, edited by P. Bergmann and V. de Sabbata (Plenum, New York, 1986).
[12] R. Penrose, in *Perspectives in Geometry and Relativity*, edited by B. Hoffman (Indiana University Press, Bloomington, 1966).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Cauchy noise and affiliated stochastic processes

Piotr Garbaczewski
Institute of Physics, Pedagogical University, PL 65-069, Zielona Góra, Poland

Robert Olkiewicz[a)]
Fakultät für Physik, Universität Bielefeld, D-33 615 Bielefeld, Germany

(Received 28 April 1998; accepted for publication 15 October 1998)

By departing from the previous attempt [Phys. Rev. E **51**, 4114 (1995)] we give a detailed construction of conditional and perturbed Markov processes, under the assumption that the Cauchy law of probability replaces the Gaussian law (appropriate for the Wiener process) as the model of primordial noise. All considered processes are regarded as probabilistic solutions of the so-called Schrödinger interpolation problem, whose validity is thus extended to the jump-type processes and their step process approximants. © *1999 American Institute of Physics.* [S0022-2488(99)00302-3]

I. INTRODUCTION

Probabilistic solutions of the so-called Schrödinger boundary data problem[1,2] are known to yield a unique Markovian interpolation between any two strictly positive probability densities designed to form the input–output statistics data (possibly phenomenological) for a certain dynamical process, taking place in a finite-time interval. The key problem, if one attempts to reconstruct the *most likely* (Markovian) dynamics, is to select the jointly continuous in space variables positive and contractive semigroup kernel. That issue was analyzed before in a number of publications.[1–8]

In the framework of the Schrödinger problem the choice of the integral kernel is arbitrary, except for the strict positivity (cf., however, Ref. 8) and continuity demand. It is thus rather natural to ask for the most general stochastic interpolation, that is admitted under the above premises.

First of all, the concept of Gaussian noise, regarded as a stochastic analog of the mechanical ''state of rest'' and traditionally linked with a Wiener process, can be extended to all infinitely divisible probability laws via the Lévy–Khintchine formula. It expands our framework from continuous diffusion processes to jump-type or combined diffusion–jump propagation scenarios[5] as appropriate mathematical models of the primordial ''free noise'' (this particular viewpoint is a novelty in the physics literature, where the Gaussian approach is dominant).

The next natural step in the analysis is to account for typical perturbations of any given process, according to the pattern of the Feynman–Kac formula, hence in terms of perturbed semigroups, where an appropriate generator (replacing the Laplacian) is additively modified by a suitable potential.[9,10] The additive perturbation choice stems from the fact[3,6,7] that basic stochastic processes of the nonequilibrium statistical physics (like, e.g., the Smoluchowski diffusion processes) involve perturbed Feynman–Kac kernels as building blocks of transition probability densities. We expect that an analogous feature (which is an artefact of the implicit Girsanov–Cameron–Martin measure change formula), is valid for non-Gaussian stochastic processes, see, e.g., Ref. 5 for explicit quantum mechanical motivations.

By referring to a physical terminology, let us consider Hamiltonians (semigroup generators) of the form $H=F(\hat{p})$, where $\hat{p}=-i\nabla$ stands for the momentum operator and for $-\infty<k<+\infty$, $F=F(k)$ is a real valued, bounded from below, locally integrable function. Here, $\hbar=c$

[a)]Permanent address: Institute of Theoretical Physics, University of Wrocław, PL 50-204 Wrocław, Poland.

$=1$. We simplify further discussion by considering processes in one spatial dimension (this limitation can be removed cf. the Remark closing Sec. III of the paper).

We easily learn that for times $t \geq 0$ there holds

$$[\exp(-tH)]f(x) = [\exp(-tF(p))\hat{f}(p)]^{\vee}(x), \tag{1}$$

where the superscript $\vee$ denotes the inverse Fourier transform and $\hat{f}$ stands for the Fourier transform of f.

Let us set $k_t = (1/\sqrt{2\pi})[\exp(-tF(p))]^{\vee}$, then the action of $\exp(-tH)$ can be given in terms of a convolution: $\exp(-tH)f = f * k_t$, where $(f*g)(x) := \int_R g(x-z)f(z)dz$.

We are interested in those $F(p)$ which give rise to positivity preserving semigroups: if $F(p)$ satisfies the Lévy–Khintchine formula, then k_t is a positive measure for all $t \geq 0$. Let us concentrate on the integral part of the Lévy–Khintchine formula, which is responsible for arbitrary stochastic jump features:

$$F(p) = -\int_{-\infty}^{+\infty} \left[\exp(ipy) - 1 - \frac{ipy}{1+y^2} \right] \nu(dy), \tag{2}$$

where $\nu(dy)$ stands for the so-called Lévy measure.

There are not many explicit examples (analytic formulas for probability densities) for processes governed by (2), except possibly for the so-called stable probability laws. Nowadays, their potential physical meaning gains recognition in various contexts ranging from deterministic (chaos) implementations of anomalous transport, anomalous diffusion studies in nonequilibrium statistical physics, through stochastic interpretation of nonlinear field equations and relativistic Hamiltonian problems in quantum theory, to investigations of the early evolution (inhomogeneity issue) of the Universe cf. Refs. 5, 11–20.

The best known example of the stable law is provided by the classic Cauchy density which will be our reference model below. Let us focus our attention on that selected choice for the characteristic exponent $F(p)$, namely: $F_0(p) = |p|$ which is the Cauchy process generator. The semigroup generator H_0 is a pseudodifferential operator. The associated kernel k_t in view of the ''free noise'' restriction (no potentials at the moment) is a transition density of the jump-type (Lévy) process, determined by the corresponding Lévy measure $\nu(dy) = (1/\pi)(dy/y^2)$. It is instructive to notice that a pseudodifferential analog of the Fokker–Planck equation holds true: $F_0(p) \Rightarrow \partial_t \overline{\rho}(x,t) = -|\nabla| \overline{\rho}(x,t)$. This evolution rule gives rise to the Cauchy process probability density $\rho(x,t) = (1/\pi)[t/(t^2+x^2)]$ and the corresponding space–time homogeneous transition density (e.g., the semigroup kernel in this free propagation case).

Our principal goal in the present paper is to generalize this observation to encompass the additive perturbations by physically motivated potentials and construct the related Markov processes. It is not *a priori* obvious that perturbed processes preserve the generic (jump-type) sample path properties of the unperturbed (the Cauchy ''free noise'') process. In addition, the physical intuition demands that an approximation of the jump-type process in terms of traditional jump processes should be generally possible which, is certainly not obvious. Those obstacles are overcome by giving a characterization of the affiliated Markovian jump-type processes in terms of approximating (convergent) families of *step* processes, that solve a suitable version of the Schrödinger interpolation problem. The construction is based on the Feynman–Kac formula for perturbed semigroups, with strictly positive and jointly continuous kernel functions.

Our demonstration explicitly pertains to the Cauchy process and its relatives, albeit the techniques and major statements may be extended to a broader class of Lévy processes and their perturbed versions cf. Refs. 5, 11–13, and 14–18 for related mathematical and physical connotations, including the numerical simulation issue, Ref. 20, with its inherent cutoffs (generic lower bound for the jump size).

II. THE CAUCHY PROCESS AND ITS CONDITIONAL RELATIVES

We consider Markovian propagation scenarios so remaining within the well established framework, where the input–output statistics data are provided in terms of two strictly positive boundary densities $\rho(x,0)$ and $\rho(x,T)$, $T>0$. In addition, a bivariate transition probability density is given in a specific factorized form: $m(x,y)=f(x)k(x,0,y,T)g(y)$, with marginals:

$$\int_R m(x,y)dy=\rho(x,0), \quad \int_R m(x,y)dx=\rho(y,T). \tag{3}$$

Here, $f(x)$, $g(y)$ are the *a priori* unknown functions, to come out as strictly positive solutions of the integral system of equations (3), provided that in addition to the density boundary data we have in hand *any* strictly positive, jointly continuous in space variables *function* $k(x,0,y,T)$. Additionally, we impose a restriction that $k(x,0,y,T)$ represents a certain strongly continuous dynamical semigroup kernel $k(y,s,x,t)$, $0\leqslant s\leqslant t<T$, while given at the time interval borders: It secures the Markov property of the sought for stochastic process.

Under those circumstances,[6] once we define functions

$$\theta(x,t)=\int dy\ k(x,t,y,T)g(y), \quad \theta_*(y,s)=\int dx\ k(x,0,y,s)f(x) \tag{4}$$

there exists a transition density

$$p(y,s,x,t)=k(y,s,x,t)\ \frac{\theta(x,t)}{\theta(y,s)}, \tag{5}$$

which implements a Markovian propagation of the probability density

$$\rho(x,t)=\theta(x,t)\,\theta_*(x,t), \qquad \rho(x,t)=\int p(y,s,x,t)\rho(y,s)dy \tag{6}$$

between the prescribed boundary data.

For a given semigroup which is characterized by its generator (Hamiltonian), the kernel $k(y,s,x,t)$ and the emerging transition probability density $p(y,s,x,t)$ are unique in view of the uniqueness of solutions $f(x)$, $g(y)$ (cf. Theorem 3.2 in Ref. 2). For Markov processes, the knowledge of the transition probability density $p(y,s,x,t)$ for all intermediate times $0\leqslant s<t\leqslant T$ suffices for the derivation of all other relevant characteristics.

At this point, let us make a definite choice of the kernel function, namely, that of the Cauchy kernel:

$$k(y,s,x,t)=\frac{1}{\pi}\,\frac{t-s}{(t-s)^2+(x-y)^2}. \tag{7}$$

We have:

Theorem 1:

(a) $p(y,s,x,t)$ defined by Eqs. (5) and (7) is a Markov transition kernel, that is (weak limit in below)

$$\int_R p(y,s,x,t)dx=1, \qquad \lim_{t\downarrow s}p(y,s,x,t)=\delta_y(x),$$

$$\int_R p(y,t_1,z,t_2)p(z,t_2,x,t_3)dz=p(y,t_1,x,t_3)$$

for all $0 \leq t_1 < t_2 < t_3 \leq T$, with δ_y standing for the Dirac delta.

(b) $\rho(x,t)$, Eq. (6), is a probability distribution interpolating between ρ_0 and ρ_T:

$$\int_R \rho(x,t)dx = 1, \qquad \rho(x,0) = \rho_0(x), \rho(x,T) = \rho_T(x).$$

(c) The process X_t having $p(y,s,x,t)$ as the transition kernel is a Markov interpolating process:

$$\int_R p(y,s,x,t)\rho(y,s)dy = \rho(x,t)$$

for all $0 \leq s < t \leq T$.

Proof: See, e.g., Refs. 5 and 6.

Let us notice that the process X_t is obtained from the Cauchy process X_t^C by means of a multiplicative transformation of transition function. Clearly, $\alpha_t^s = [\theta(X_t^C,t)/\theta(X_s^C,s)]$ is a multiplicative functional of X_t^C such that its average with respect to the Cauchy process reads $\int \alpha_t^s(\omega) P_x^C(d\omega) = 1$ for any $0 \leq s \leq t \leq T$ and any $x \in R$, see, e.g., Ref. 21. However α_t^s is not homogeneous and, even worse, not contracting (in fact, not even bounded). We cannot be *a priori* sure that the generic sample path properties of the Cauchy process can be attributed to X_t as well. In particular, an approximation of X_t in terms of jump processes with a finite number of jumps in a finite time interval is by no means obvious and needs a demonstration (to be given in below).

To this end, let us first notice that θ_* and θ satisfy the conjugate pseudodifferential equations:

$$\partial_t \theta_* = -|\nabla| \theta_*, \qquad \partial_t \theta = |\nabla| \theta, \tag{8}$$

where the operator $|\nabla|$ acts as follows:

$$|\nabla| f(x) = -\frac{1}{\pi} \int_R \left[f(x+y) - f(x) - \frac{y \nabla f(x)}{1+y^2} \right] \frac{dy}{y^2}. \tag{9}$$

Let us define a new operator $|\nabla|_\epsilon$ by

$$|\nabla|_\epsilon f(x) = -\frac{1}{\pi} \int_{|y|>\epsilon} [f(x+y) - f(x)] \frac{dy}{y^2} \tag{10}$$

and, accordingly,

$$\partial_t \theta_*^\epsilon = -|\nabla|_\epsilon \theta_*^\epsilon, \qquad \partial_t \theta^\epsilon = |\nabla|_\epsilon \theta^\epsilon, \tag{11}$$

with $\theta_*^\epsilon(x,0) = \theta_*(x,0)$, $\theta^\epsilon(x,T) = \theta(x,T)$ chosen to coincide with the respective initial and terminal data for solutions of Eq. (8).

Furthermore, let

$$q_\epsilon(x) = \frac{1}{\pi} \chi_{I_\epsilon^c}(x) \frac{1}{x^2}, \tag{12}$$

where $I_\epsilon^c = [-\epsilon, \epsilon]^c = \{x \in R : |x| > \epsilon\}$ and χ_A is an indicator function of a set A.

We have:

Theorem 2: Let us define the Poisson transition kernel corresponding to the measure $q_\epsilon(x)dx$:

$$k_\epsilon(x,t) = \left[\exp\left(-\frac{2t}{\epsilon\pi} \right) \right] \left[\delta_0(x) + t q_\epsilon(x) + \frac{t^2}{2!} (q_\epsilon * q_\epsilon)(x) + \cdots \right].$$

Then, functions,

$$\theta_*^\epsilon(x,t)=\int_R k_\epsilon(x-y,t)\,\theta_*(y,0)dy,$$

$$\theta^\epsilon(x,t)=\int_R k_\epsilon(x-y,T-t)\,\theta(y,T)dy,$$

solve Eq. (11).

Proof: The transition function in the above is called the Poisson transition kernel following the terminology of Ref. 21. We have $\theta_*^\epsilon(x,0)=\int_R \delta_0(x-y)\,\theta_*(y,0)dy=\theta_*(x,0)$ and

$$\partial_t\theta_*^\epsilon(x,t)=\int_R[\partial_t k_\epsilon(x-y,t)]\,\theta_*(y,0)dy,$$

where

$$\partial_t k_\epsilon(x,t)=-\frac{2}{\pi\epsilon}\,k_\epsilon(x,t)+\left[\exp\left(-\frac{2t}{\epsilon}\right)\right][q_\epsilon(x)+t(q_\epsilon*q_\epsilon)(x)+\cdots].$$

Consequently,

$$\begin{aligned}[\partial_t k_\epsilon(\cdot,t)*\theta_*(\cdot,0)](x)&=-\frac{2}{\pi\epsilon}\,\theta_*^\epsilon(x,t)+\left[\left[\exp\left(-\frac{2t}{\epsilon}\right)\right]q_\epsilon*\left(\delta_0+tq_\epsilon+\frac{t^2}{2!}\,q_\epsilon*q_\epsilon+\cdots\right)*\theta_*\right](x)\\&=-\frac{2}{\pi\epsilon}\,\theta_*^\epsilon(x,t)+[q_\epsilon*\theta_*^\epsilon(\cdot,t)](x)\\&=-\frac{2}{\pi\epsilon}\,\theta_*^\epsilon(x,t)+\int_R q_\epsilon(y)\,\theta_*^\epsilon(x-y,t)dy.\end{aligned}$$

But, there holds

$$\int_R q_\epsilon(y)\,\theta_*^\epsilon(x-y,t)dy=\frac{1}{\pi}\int_{|y|>\epsilon}\theta_*^\epsilon(x-y,t)\,\frac{dy}{y^2}=\frac{1}{\pi}\int_{|y|>\epsilon}\theta_*^\epsilon(x+y,t)\,\frac{dy}{y^2}$$

and, in view of the obvious identity

$$\frac{2}{\pi\epsilon}\,\theta_*^\epsilon(x,t)=\frac{1}{\pi}\int_{|y|>\epsilon}\theta_*^\epsilon(x,t)\,\frac{dy}{y^2},$$

we finally arrive at

$$\partial_t\theta_*^\epsilon(x,t)=\frac{1}{\pi}\int_{|y|>\epsilon}[\theta_*^\epsilon(x+y,t)-\theta_*^\epsilon(x,t)]\,\frac{dy}{y^2}=-|\nabla|_\epsilon\theta_*^\epsilon(x,t).$$

An analogous line of arguments follows with respect to $\theta^\epsilon(x,t)$, which completes the proof.

A random process with a Poisson transition function belongs to the class of so-called step processes,[21,22] that is jump processes with no accumulation points of jumps in a finite time interval: The number of jumps is finite on each finite time interval. We have:

Lemma 1: The Markov process Y_t^ϵ given by the transition function $k_\epsilon(x,t)$ is a step process with a characteristic function:

$$\Phi_\epsilon(p,t)=\exp(-t[\hat{q}_\epsilon(0)-\hat{q}_\epsilon(p)]),$$

where $\hat{q}_\epsilon(p)$ is the Fourier transform of $q_\epsilon(x)$.

Proof: We need to evaluate the characteristic function of the transition kernel, that is,

$$\Phi_\epsilon(p,t)=\exp\left(-\frac{2t}{\pi\epsilon}\right)\cdot\int_{-\infty}^{+\infty}[\exp(-ipx)]\left[\delta_0(x)+tq_\epsilon(x)+\frac{t^2}{2!}(q_\epsilon * q_\epsilon)(x)+\dots\right]dx$$

$$=\exp\left(-\frac{2t}{\pi\epsilon}\right)\cdot\left[1+t\hat{q}_\epsilon(p)+\frac{t^2}{2!}(\hat{q}(p))^2+\cdots\right]=\exp\left[-\frac{2t}{\pi\epsilon}+t\hat{q}_\epsilon(p)\right].$$

In view of $\hat{q}_\epsilon(0)=2/\pi\epsilon$, the Lemma holds true.

As a technical warm up we shall now prove that the Cauchy process is the limit (in distributions) of a one-parameter family of step processes Y_t^ϵ. We touch here an important issue of limits (convergence) of jump processes,[23–25] and there are many types of the pertinent convergence. For example, it is known that Y_t^ϵ tends to the Cauchy process in probability; Ref. 23, while major modern techniques refer to the weak convergence of probability measures, Ref. 25). Also, typical proofs refer only to processes with stationary independent increments, while we cannot respect this limitation in the presence of perturbations.

Lemma 2: There holds: $\lim_{\epsilon\to 0}\Phi_\epsilon(p,t)=\psi(p,t)$, where $\psi(p,t)$ is the Cauchy characteristic function $\psi(p,t)=\exp(-t|p|)$. Moreover, the limit is uniform for all $t\in[0,T]$.

Proof: Let us evaluate $\hat{q}_\epsilon(p)$:

$$\hat{q}_\epsilon(p)=\frac{1}{\pi}\int_{-\infty}^{+\infty}\exp(ipx)\cdot\hat{q}_\epsilon(x)dx=\frac{1}{\pi}\int_{|x|>\epsilon}\exp(ipx)\cdot\frac{dx}{x^2}=\frac{2}{\pi}\int_\epsilon^\infty\frac{\cos(px)-1}{x^2}dx+\frac{2}{\pi\epsilon}.$$

Consequently,

$$\Phi_\epsilon(p,t)=\exp\left[-\frac{2t}{\pi}\int_\epsilon^\infty\frac{1-\cos(px)}{x^2}dx\right].$$

In view of

$$\lim_{\epsilon\to 0}\int_\epsilon^\infty\frac{1-\cos(px)}{x^2}dx=\frac{|p|\pi}{2},$$

we arrive at

$$\lim_{\epsilon\to 0}\Phi_\epsilon(p,t)=\exp\left[-\frac{2t}{\pi}\cdot\frac{|p|\pi}{2}\right]=\exp(-t|p|).$$

The proof is completed.

Clearly, $|\nabla|_\epsilon$ is a well-defined semigroup generator for the step process Y_t^ϵ. Let us recall that sample paths of a step process have only a finite number of jumps in each finite time interval, and between jumps the sample path is constant.[22] The limiting Cauchy process belongs to the category of jump-type processes, where apart from the long jumps-tail (no fixed bound can be imposed on their length) that implies the nonexistence of moments of the probability measure, sample paths of the Cauchy process may have an infinite number of jumps of arbitrarily small size. By general arguments, pertaining to the space $D_E[0,\infty)$ of right continuous functions with left limits (cadlag), both in the finite and infinite time interval the number of jumps is at most countable.[23,26] It is also useful to recall that on a finite time interval there can be at most finitely many points $t\in[0,T]$ at which the jump size exceeds a given positive number. In view of that, $\sup_{t\in[0,T]}|Y_t^\epsilon|<\infty$. Obviously, there is no fixed upper bound for the size of jumps (except for being finite), since a stochastically continuous process with independent increments having, with probability 1, no jumps exceeding a certain constant C, would possess all moments.[22]

Now, we shall pass to a slightly more involved demonstration that a well-defined family of Markov processes X_t^ϵ (in fact, step ones) can be constructed, such that the process X_t of Theorem 1 can be approximated (in the sense of suitable convergence) to an arbitrary degree of accuracy.

Here, we are motivated by a heuristic analysis carried out in our earlier paper.[5] There, we found that after neglecting ''small jumps,'' the time evolution of the resultant probability density $\overline{\rho_\epsilon}=\theta^\epsilon\theta_*^\epsilon$ may be written as

$$\partial_t\overline{\rho_\epsilon}(A,t)=\int_R q_\epsilon(t,x,A)\overline{\rho_\epsilon}(x,t)dx+\langle v\rangle_A(t)\int_{|y|>\epsilon}\frac{y}{1+y^2}\,d\nu(y), \tag{13}$$

where $\langle v\rangle_A=\int_A\overline{\rho_\epsilon}(x,t)[\nabla\ \ln(\theta^\epsilon/\theta_*^\epsilon)(x,t)]dx$. The measure $d\nu$ is symmetric around the point $\{0\}$, hence the second term cancels, and we arrive at

$$\partial_t\overline{\rho_\epsilon}(A,t)=\int_R q_\epsilon(t,x,A)\overline{\rho_\epsilon}(x,t)dx, \tag{14}$$

where the so-called jump intensity reads

$$q_\epsilon(t,y,A)=\int_{|y|>\epsilon}\frac{\theta^\epsilon(y+x,t)}{\theta^\epsilon(y,t)}\,[\chi_A(x+y)-\chi_A(y)]d\nu(x) \tag{15}$$

and $\theta^\epsilon(x,t)$ comes out as a solution of the second pseudodifferential equation in the formula (11).

Let us define [cf. Eq. (12)]

$$h_\epsilon(t,y)=\int_{-\infty}^{+\infty}\frac{\theta^\epsilon(x+y,t)}{\theta^\epsilon(y,t)}\,q_\epsilon(x)dx \tag{16}$$

and

$$h_\epsilon(t,y,x)=\frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,t)}\,q_\epsilon(x-y). \tag{17}$$

Then, clearly the jump intensity (14) takes the form

$$q_\epsilon(t,y,A)=\int_A h_\epsilon(t,y,x)dx-h_\epsilon(t,y)\chi_A(y). \tag{18}$$

With those notations, we have:

Lemma 3: If the function $g(y)$ [cf. Eq. (4)] is uniformly bounded, then $h_\epsilon(t,y,x)$ is a density of a finite measure and $h_\epsilon(t,y)=\int_R h_\epsilon(t,y,x)dx$.

Proof: By our assumption, $g(y)\le M$ for all $y\in R$. Because of $\theta^\epsilon(x,t)=\int_R k_\epsilon(T-t,x-y)g(y)dy$, we have a bound

$$\theta^\epsilon(x,t)\le M\int_R k_\epsilon(T-t,x-y)dy=M.$$

Hence

$$\int_{-\infty}^{+\infty}h_\epsilon(t,y,x)dx=\int_{-\infty}^{+\infty}\frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,t)}\,q_\epsilon(x-y)dx=h_\epsilon(t,y)$$

and

$$h_\epsilon(t,y) \leqslant \frac{M}{\theta^\epsilon(y,t)} \frac{2}{\epsilon}.$$

It is also clear that $h_\epsilon(t,y,x) \geqslant 0$, which completes the proof.

Let us define $\overline{h}_\epsilon(t,y,A) = -h_\epsilon(t,y)\chi_A(y) + \int_A h_\epsilon(t,y,x)dx$. It is obvious that $\overline{h}_\epsilon$ is a charge (that is a real-valued measure with the property $\overline{h}_\epsilon(t,y,R)=0$).[22]

We shall show that there exists a *step* process corresponding to the charge $\overline{h}_\epsilon$.

To this end let us first prove:

Lemma 4: For any Borel set $A \subset R$, the function $t \rightarrow \int_A h_\epsilon(t,y,x)dx$ is continuous in t, uniformly in A.

Proof: We have the following estimate [cf. Eq. (18) and Lemma 3]:

$$\begin{aligned}
&\left| \int_A h_\epsilon(t,y,x)dx - \int_A h_\epsilon(t_0,y,x)dx \right| \\
&= \left| \int_A \frac{\theta^\epsilon(y+x,t)}{\theta^\epsilon(y,t)} q_\epsilon(x)dx - \int_A \frac{\theta^\epsilon(y+x,t_0)}{\theta^\epsilon(y,t_0)} q_\epsilon(x)dx \right| \\
&\leqslant \left| \int_{A\cap K^c} \left[\frac{\theta^\epsilon(y+x,t)}{\theta^\epsilon(y,t)} - \frac{\theta^\epsilon(y+x,t_0)}{\theta^\epsilon(y,t)} \right] q_\epsilon(x)dx \right| + \left| \int_{A\cap K} \left[\frac{\theta^\epsilon(y+x,t)}{\theta^\epsilon(y,t)} - \frac{\theta^\epsilon(y+x,t_0)}{\theta^\epsilon(y,t)} \right] q_\epsilon(x)dx \right| \\
&\quad + \left| \int_A \left[\frac{\theta^\epsilon(y+x,t_0)}{\theta^\epsilon(y,t)} - \frac{\theta^\epsilon(y+x,t_0)}{\theta^\epsilon(y,t_0)} \right] q_\epsilon(x)dx \right|,
\end{aligned}$$

where K is a compact set while K^c is its complement.

Let us denote the summands A_1, A_2, A_3, respectively. For the first summand we have

$$A_1 \leqslant \frac{1}{\theta^\epsilon(y,t)} \sup_{x\in R}(\theta^\epsilon(x,t) + \theta^\epsilon(x,t_0)) \int_{K^c} q_\epsilon(x)dx.$$

But

$$\sup_{x\in R}\theta^\epsilon(x,t) = \sup_{x\in R}\int_R k_\epsilon(T-t,x-y)g(y)dy \leqslant M \sup_{x\in R}\int_R k_\epsilon(T-t,x-y)dy = M.$$

By defining $N(y) = \sup_{t\in[t_0,t_0+1]}[1/\theta^\epsilon(y,t)]$ and adjusting the compact set K so that $\int_{K^c} q_\epsilon(x)dx \leqslant \delta/3MN(y)$, we arrive at $A_1 \leqslant \delta/3$.

With the second summand, A_2, we proceed as follows:

$$A_2 = \left| \int_{A\cap K} \left[\frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,t)} - \frac{\theta^\epsilon(x,t_0)}{\theta^\epsilon(y,t)} \right] q_\epsilon(y-x)dx \right| \leqslant N(y)\sup_{x\in K}|\theta^\epsilon(x,t) - \theta^\epsilon(x,t_0)| \frac{2}{\pi\epsilon}.$$

By choosing t so close to t_0 that $\sup_{x\in K}|\theta^\epsilon(x,t) - \theta^\epsilon(x,t_0)| \leqslant \pi\delta\epsilon/6N(y)$, we get $A_2 \leqslant \delta/3$. Analogously with A_3:

$$A_3 \leqslant \left| \frac{1}{\theta^\epsilon(y,t)} - \frac{1}{\theta^\epsilon(y,t_0)} \right| 2 \sup_{x\in R}\theta^\epsilon(x,t_0) \frac{2}{\pi\epsilon} \leqslant \frac{4}{\pi\epsilon} MN^2(y)|\theta^\epsilon(y,t_0) - \theta^\epsilon(y,t)|,$$

where by taking t such that $|\theta^\epsilon(y,t_0) - \theta^\epsilon(y,t)| \leqslant \pi\delta\epsilon/12MN^2(y)$ we shall get $A_3 \leqslant \delta/3$. The overall bound is thus δ, and the Lemma is proved.

As a byproduct of the above demonstration, we realize that the function $t \rightarrow h_\epsilon(t,x,A)$ is continuous in t uniformly on compact sets. As a consequence, see, e.g., Theorem 4 in Chap. 7, Sec. 7 of Ref. 22, there exists a stochastically continuous Markov process X_t^ϵ with continuous from the right sample paths. Moreover, for any $s \in [0,T]$, $y \in R$ and $A \subset R$, there holds

$$\lim_{t\downarrow s} \frac{p_\epsilon(y,s,A,t) - \chi_A(y)}{t-s} = \overline{h}_\epsilon(s,y,A), \tag{19}$$

where $p_\epsilon(y,s,A,t)$ is the transition kernel of the process X_t^ϵ.
There follows:

Theorem 3: The transition probability density of X_t^ϵ reads:

$$p_\epsilon(y,s,x,t) = k_\epsilon(t-s,x-y)\,\frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,s)}$$

and is a solution of the first Kolmogorov equation:

$$\partial_s p_\epsilon(y,s,x,t) = -\int_R p_\epsilon(z,s,x,t)\overline{h}_\epsilon(s,y,z)\,dz.$$

Proof: We must demonstrate that Eq. (19) is valid for the just introduced transition density (compare, e.g., also Theorem 1), i.e., there holds

$$\lim_{t\downarrow s} \frac{1}{t-s}\left[k_\epsilon(t-s,x-y)\,\frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,s)} - \delta_y(x)\right] = \overline{h}_\epsilon(s,y,x).$$

To this end, let us notice (adding and subtracting the same summand) that

$$\begin{aligned}\overline{h}_\epsilon(s,y,x) &= \frac{\theta^\epsilon(x,s)}{\theta^\epsilon(y,s)} \lim_{t\downarrow s} \frac{1}{t-s}\,[k_\epsilon(t-s,x-y) - \delta_y(x)]\\ &\quad + \frac{\delta_y(x)}{\theta^\epsilon(y,s)} \lim_{t\downarrow s} \frac{1}{t-s}\,[\theta^\epsilon(x,t) - \theta^\epsilon(y,s)]\\ &= \frac{\theta^\epsilon(x,s)}{\theta^\epsilon(y,s)}\left[q_\epsilon(x-y) - \frac{2}{\pi\epsilon}\,\delta_y(x)\right] + \frac{\delta_y(x)}{\theta^\epsilon(y,s)} \lim_{t\downarrow s} \frac{1}{t-s}\,[\theta^\epsilon(x,t) - \theta^\epsilon(y,s)].\end{aligned}$$

To evaluate the second term, let us take a continuous and bounded function $a(x)$ and consider

$$\begin{aligned}&\lim_{t\downarrow s} \int_R \frac{\delta_y(x)}{\theta_\epsilon(y,s)}\,\frac{1}{t-s}\,[\theta^\epsilon(x,t) - \theta^\epsilon(y,s)]a(x)\,dx\\ &\quad = \lim_{t\downarrow s} \frac{a(y)}{\theta^\epsilon(y,s)}\,\frac{1}{t-s}\,[\theta^\epsilon(y,t) - \theta^\epsilon(y,s)] = \frac{a(y)}{\theta^\epsilon(y,s)}\,\partial_s\theta^\epsilon(y,s).\end{aligned}$$

So, the second term converges weakly to

$$\frac{\delta_y(x)}{\theta^\epsilon(y,s)}\,\partial_s\theta^\epsilon(y,s).$$

We know that

$$\partial_s\theta^\epsilon(y,s) = |\nabla|_\epsilon\theta^\epsilon(y,s) = -\int_R [\theta^\epsilon(y+z,s) - \theta^\epsilon(y,s)]q_\epsilon(z)\,dz.$$

Consequently,

$$\frac{\partial_s \theta^\epsilon(y,s)}{\theta^\epsilon(y,s)} = -\int_R \frac{\partial^\epsilon(y+z,s)}{\theta^\epsilon(y,s)}\, q_\epsilon(z)dz + \frac{2}{\pi\epsilon} = \frac{2}{\pi\epsilon} - h_\epsilon(s,y)$$

and thus

$$\lim_{t\downarrow s} \frac{1}{t-s}\,[p_\epsilon(y,s,x,t) - \delta_y(x)] = \frac{\theta^\epsilon(x,s)}{\theta^\epsilon(y,s)}\, q_\epsilon(x-y) - \frac{2}{\pi\epsilon}\,\delta_y(x) + \frac{2}{\pi\epsilon}\,\delta_y(x) - h_\epsilon(s,y)\,\delta_y(x)$$
$$= h_\epsilon(s,y,x) - h_\epsilon(s,y)\,\delta_y(x)$$
$$= \overline{h}_\epsilon(s,y,x).$$

The first part of our Theorem is proved, and we can pass to its second part.

To check the validity of the Kolmogorov equation, we shall begin from

$$\partial_s p_\epsilon(y,s,x,t) = [\partial_s k_\epsilon(t-s,x-y)]\, \frac{\theta^\epsilon(x,t)}{\theta_\epsilon(y,s)} - p_\epsilon(y,s,x,t)\, \frac{\partial_s \theta^\epsilon(y,s)}{\theta^\epsilon(y,s)}.$$

But

$$\partial_s k_\epsilon(t-s,x-y) = -[q_\epsilon * k_\epsilon(t-s,\cdot)](x-y) + k_\epsilon(x-y)\, \frac{2}{\pi\epsilon}$$

and

$$\frac{\partial_s \theta^\epsilon(y,s)}{\theta_\epsilon(y,s)} = \frac{2}{\pi\epsilon} - h_\epsilon(s,y),$$

which leads to

$$\partial_s p(y,s,x,t) = -[q_\epsilon * k_\epsilon(t-s,\cdot)](x-y)\, \frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,s)} + \frac{2}{\pi\epsilon}\, p_\epsilon(y,s,x,t)$$
$$- \frac{2}{\pi\epsilon}\, p_\epsilon(y,s,x,t) + p_\epsilon(y,s,x,t) h_\epsilon(s,y)$$
$$= -\frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,s)} \int_R q_\epsilon(x-y-z) k_\epsilon(t-s,z)dz + p_\epsilon(y,s,x,t) \int_R \frac{\theta^\epsilon(x+y,s)}{\theta^\epsilon(y,s)}\, q_\epsilon(x)dx.$$

On the other hand,

$$-\int_R p_\epsilon(z,s,x,t)\overline{h}(s,y,z)dz = -\int_R k_\epsilon(t-s,x-z)\, \frac{\theta^\epsilon(x,t)}{\theta^\epsilon(z,s)} \left[\frac{\theta^\epsilon(z,s)}{\theta^\epsilon(y,s)}\, q_\epsilon(z-y) - \delta_y(z) h_\epsilon(s,y) \right] dz$$
$$= -\frac{\theta^\epsilon(x,t)}{\theta_\epsilon(y,s)} \int_R k_\epsilon(t-s,x-z) q_\epsilon(z-y)dz + p_\epsilon(y,s,x,t) h_\epsilon(s,y).$$

Since we know that $h_\epsilon(s,y) = \int_R [\theta^\epsilon(x+y,s)/\theta_\epsilon(y,s)] q_\epsilon(x)dx$, the assertion (e.g., the validity of the first Kolmogorov equation) follows.

Corollary 1: X_t^ϵ is a step process.

Proof: It suffices to check that $p_\epsilon(y,s,R,t) = 1$ (cf. Ref. 22). Since

$$p_\epsilon(y,s,R,t) = \int_R p_\epsilon(y,s,x,t)dx = \int_R k_\epsilon(t-s,x-y)\, \frac{\theta^\epsilon(x,t)}{\theta^\epsilon(y,s)}\, dx$$

and, by Theorem 2,

$$\int_R k_\epsilon(t-s,x-y)\,\theta^\epsilon(x,t)\,dx=\theta^\epsilon(y,s)$$

the Corollary holds true.

All previous considerations can be finally summarized by showing that the family X_l^ϵ of step processes consistently approximates (converges to) the process X_t. Indeed, we have:

Theorem 4: The limit

$$\lim_{\epsilon\downarrow 0}X_t^\epsilon=X_t$$

holds true in distributions and uniformly in $t\in[0,T]$. Moreover, the transition probability density p_ϵ converges pointwise to p when $\epsilon\downarrow 0$.

Proof: The probability density of the process X_t^ϵ equals $\rho_\epsilon(x,t)=\theta_*^\epsilon(x,t)\,\theta^\epsilon(x,t)$ and that of the process X_t is given by $\rho(x,t)=\theta_*(x,t)\,\theta(x,t)$. But, $\theta_*^\epsilon(x,t)=\int_R k_\epsilon(t,x-y)f(y)dy$ and $k_\epsilon(t,x-y)$ converges weakly to the Cauchy kernel $k(t,x-y)$, uniformly in t. Consequently $\lim_{\epsilon\downarrow 0}\ \theta_*^\epsilon(x,t)=\theta_*(x,t)$ also uniformly in $t\in[0,T]$. The same holds true for $\theta^\epsilon(x,t)$, and the first assertion follows.

The second statement follows from the fact that $k_\epsilon(t,x)$ tends to the Cauchy kernel $k(t,x)$ (see Lemma 2) when $\epsilon\downarrow 0$.

As stated before, considerations of the present section were mostly a preparation to the study of perturbed problems. However, it is useful to mention that the conditional Cauchy processes are covered by the developed scheme. In fact, we can here adjust to the Cauchy noise an observation previously utilized in the context of the Wiener noise.[2,6,5] The pertinent density can be given in the following form:

$$\rho(x,t)=\frac{k(y_0,t_0,x,t)k(x,t,z_T,T)}{k(y_0,t_0,z_T,T)} \tag{20}$$

with $y_0,z_T\in R$ and $0<t_0<t<T$. All previous considerations directly apply to the interpolating process supported by this density. See also a discussion of Lévy bridges (while specialized to the Cauchy context) in Ref. 27.

III. PERTURBATIONS OF THE CAUCHY NOISE

We are motivated by the strategy of Refs. 6 and 5 and the techniques developed in Sec. II. Let us address the problem analogous to that of Eq. (11), but now in reference to a perturbed semigroup:[11]

$$\partial_t\theta_*=-|\nabla|\,\theta_*-V\theta_*,\qquad \partial_t\theta=|\nabla|\,\theta+V\theta, \tag{21}$$

where V is a measurable function such that:

(a) for all $x\in R$, $V(x)\geqslant 0$,
(b) for each compact set $K\subset R$ there exists C_K such that for all $x\in K$, V is locally bounded $V(x)\leqslant C_K$.

Then V is locally integrable and for any compact K we have

$$\lim_{t\downarrow 0}\ \sup_{x\in R}\ E_x^C\left\{\int_0^t \chi_K(X_s^C)V(X_s^C)ds\right\}=0. \tag{22}$$

As a consequence, there holds

Lemma 5: If $1\leqslant r\leqslant p\leqslant\infty$ and $t>0$, then the operators T_t^V defined by

$$(T_t^V f)(x) = E_x^C\left\{ f(X_t^C)\exp\left[-\int_0^t V(X_s^C)ds\right]\right\}$$

are bounded from $L^r(R)$ into $L^p(R)$. Moreover, for each $r \in [1,\infty]$ and $f \in L^r(R)$, $T_t^V f$ is a bounded and continuous function.

Proof: See, e.g., Ref. 11, Proposition III.1.

Let us notice that for the ultimate construction of the (forward, see Refs. 3, 6, and 7) Markov process we utilize only the second equation (21) [cf. also Eq. (11)], although both equations (21) are indispensable for the Schrödinger problem solution.

We shall use another identity proved by Carmona,[11] namely:

Lemma 6: For any real-valued $f,g \in L^2(R)$ there holds

$$\int_R dx\, f(x)E_x^C\left\{ g(X_t^C)\exp\left[-\int_0^t V(X_s^C)ds\right]\right\} = \int_R dx\, g(x)E_x^C\left\{ f(X_t^C)\exp\left[-\int_0^t V(X_s^C)ds\right]\right\}.$$

Proof: See also Eq. (III.9) in Ref. 11.

We need to prove that T_t^V is an integral operator. To this end, a direct transfer of Simon's arguments, cf. Ref. 28, originally with respect to the Laplace differential operator, i.e., the usage of the Dunford–Pettis theorem (see pp. 450 in Ref. 28) and Lemma 5, gives rise to:

Lemma 7: For any $p \in [1,\infty]$ and $f \in L^p(R)$ there holds

$$(T_t^V f)(x) = \int_R k_t^V(x,y) f(y)dy,$$

where $k_t^V(x,y) \geqslant 0$ almost everywhere and, for q such that $1/q + 1/p = 1$, the kernel satisfies

$$\sup_{x\in R}\left[\int_R [k_t^V(x,y)]^q dy\right]^{1/q} < \infty.$$

Proof: See also Theorem A.1.1 and Corollary A.1.2 in Ref. 28.

Notice that by putting $p=1$ and thus $q=\infty$ we obtain that $k_t^V(x,y) \in L^\infty(R^2)$.

Our ultimate goal is to utilize $k_t^V(x,y)$ in the context of the Schrödinger boundary data and interpolation problem, Refs. 2 and 6, hence suitable properties of the kernel must be established. For our purposes, the joint continuity and positivity of the kernel is essential.

Lemma 8: $k_t^V(x,y)$ is jointly continuous in (x,y).

Proof: We begin from demonstrating that $k_t^V(x,y) = k_t^V(y,x)$ almost everywhere. By Lemma 6, we have

$$\int\int_{R^2} dx\, dy f(x) k_t^V(x,y) g(y) = \int\int_{R^2} dx\, dy g(x) k_t^V(x,y) f(y),$$

hence

$$\int\int_{R^2} dx\, dy f(x) g(y)[k_t^V(x,y) - k_t^V(y,x)] = 0$$

for all $f,g \in L^2(R)\cap L^1(R)$.

The same holds true for all finite combinations $\Sigma_{i,j}\, a_{ij} f_i(x) g_j(y)$. Therefore $\int\int_{R^2}[k_t^V(x,y) - k_t^V(y,x)]f(x,y)dx\, dy = 0$ for all $f(x,y)$ from a dense subset of $L^1(R^2)$. Because $L^\infty(R^2)$ is the dual space to $L^1(R^2)$, we conclude that $k_t^V(x,y) = k_t^V(y,x)$ almost everywhere.

Let us exploit the semigroup property of $k_t^V(x,y)$:

$$k_t^V(x,y)=\int_R k_{t/2}^V(x,w)k_{t/2}^V(w,y)dw.$$

For each y, $w\to k_{t/2}^V(w,y)\in L^\infty(R)$ so, by Lemma 5, $k_t^V(x,y)$ is continuous in x. By the symmetry, $k_t^V(x,y)$ is separately continuous in x and y.

Let us consider a sequence $(x_n,y_n)\to(x,y)$. Then:

$$\begin{aligned}|k_t^V(x_n,y_n)-k_t^V(x_0,y_0)|&\leqslant\left|\int\int_{R^2}dw\ dz[k_{t/3}^V(x_n,w)-k_{t/3}^V(x_0,w)]k_{t/3}^V(w,z)k_{t/3}^V(z,y_n)\right|\\&\quad+\left|\int\int_{R^2}dw\ dz k_{t/3}^V(x_0,w)k_{t/3}^V(w,z)[k_{t/3}^V(z,y_n)-k_{t/3}^V(x,y)]\right|\\&=\left|\int_R dw[k_{t/3}^V(x_n,w)-k_{t/3}^V(x_0,w)]k_{2t/3}^V(w,y_n)\right|\\&\quad+|k_t^V(x_0,y_n)-k_t^V(x_0,y_0)|.\end{aligned}$$

Because of

$$\|k_{2t/3}^V(\cdot,y_n)\|_{L^\infty}<C$$

for all y_n, knowing that $\sup_n k_{t/3}^V(x_n,w)$ exists and is integrable with respect to w, by the Lebesgue dominated convergence theorem the first summand tends to zero. Hence, $k_t^V(x,y)$ is jointly continuous in (x,y).

Lemma 9: $k_t^V(x,y)$ is strictly positive.

Proof: Because for the Cauchy process we have[20] (more general estimates of the growth of random walks and Lévy processes can be found in Ref. 29):

$$E_x^C\{\sup_{0\leqslant s\leqslant t}|X_s^C|>n\}\leqslant 3\ \sup_{0\leqslant s\leqslant t}E_x^C\left\{|X_s^C|>\frac{n}{3}\right\}$$

and

$$\sup_{0\leqslant s\leqslant t}E_x^C\left\{|X_s^C|>\frac{n}{3}\right\}=E_x^C\left\{|X_t^C|>\frac{n}{3}\right\}=1-\frac{2}{\pi}\ \arctan\left(\frac{n}{3t}\right)$$

there follows:

$$\lim_{n\to\infty}\ E_x^C\{\sup_{0\leqslant s\leqslant t}\ |X_s^C|>n\}=0.$$

This property will be used below.

Let $0<\delta<1$, then:

$$\int_{y-\delta}^{y+\delta}dy\ k_t^V(x,y)=E_x^C\left\{\chi_{[y-\delta,y+\delta]}(X_t^C)\exp\left[-\int_0^t V(X_s^C)ds\right]\right\}.$$

By the previously deduced property, for fixed x and y, we can choose a compact set $[-n,n]$ such that

$$E_x^C\{\Omega_{(t,[y-\delta,y+\delta])}^{(0,x)}(n)\}>\frac{1}{2}\int_{y-\delta}^{y+\delta}k_t(x,y)dy,$$

where

$$\Omega^{(0,x)}_{(t,[y-\delta,y+\delta])}(n)=\{\omega:\omega(0)=x,\omega(t)\in[y-\delta,y+\delta];s\in[0,t]\Rightarrow\omega(s)\in[-n,n]\}$$

and $k_t(x,y)$ is the Cauchy kernel. Hence

$$\int_{y-\delta}^{y+\delta}dy\ k_t^V(x,y)\geqslant\int_{\Omega(n)}\exp\left[-\int_0^t V(X_s^C)ds\right]dP_x^C(\omega)\geqslant\frac{1}{2}\exp(-c_nt)\cdot\int_{y-\delta}^{y+\delta}k_t(x,y)dy,$$

where $c_n=\sup_{x\in[-n,n]}V(x)$.

Because $k_t^V(x,y)$ is continuous and δ was arbitrary, we get

$$k_t^V(x,y)\geqslant\tfrac{1}{2}\exp(-c_nt)\cdot k_t(x,y).$$

The assertion of Lemma 9 is thus valid.

Lemmas 8 and 9 provide us with a strictly positive and jointly continuous in space variables kernel, which can be directly exploited for the analysis of the Schrödinger interpolation problem, as exemplified by Eqs. (3)–(6), see also Refs. 2, 3, and 6. Indeed, let $\rho_0(x)$ and $\rho_T(x)$ be strictly positive densities. Then, the Markov process X_t^V characterized by the transition probability density:

$$p^V(y,s,x,t)=k_{t-s}^V(x,y)\ \frac{\theta(x,t)}{\theta(y,s)} \tag{23}$$

and the density of distributions

$$\rho(x,t)=\theta_*(x,t)\,\theta(x,t),$$

where

$$\theta_*(x,t)=\int_R k_t^V(x,y)f(y)dy,\qquad \theta_*(y,t)=\int_R k_{T-t}^V(x,y)g(x)dx$$

is precisely that interpolating Markov process to which Theorem 1 extends its validity, when the perturbed semigroup kernel replaces the Cauchy kernel.

Clearly, for all $0\leqslant s\leqslant t\leqslant T$ we have

$$\theta_*(x,t)=\int_R k_{t-s}^V(x,y)\,\theta_*(y,s)dy,\qquad \theta(y,s)=\int_R k_{t-s}^V(x,y)\,\theta(x,t)dx \tag{24}$$

and that suffices for the Theorem 1 to hold true in the present case as well.

Following the strategy of Sec. II, we shall investigate an issue of approximating the perturbed Cauchy process (set by Lemmas 8 and 9 and Theorem 1) by means of step processes.

Let us first invoke the step process Y_t^ϵ of Lemma 1. It corresponds to the unperturbed generator $|\nabla|_\epsilon$. To account for a perturbation and the involved perturbed semigroup, let us consider a multiplicative, homogeneous, and contracting functional:

$$\alpha_t^s(\omega)=\exp\left[-\int_s^t V(Y_\tau^\epsilon(\omega))d\tau\right] \tag{25}$$

of the process Y_t^ϵ, for times $0\leqslant s\leqslant t\leqslant T$.

We recall that the process Y_t^ϵ is a step process obtained from the Cauchy process by neglecting ''small jumps'' (the ϵ cutoff).

We shall associate with the multiplicative functional (25) the process $Y_t^{\epsilon,V}$ and prove that under additional restrictions on the potential V, the pertinent perturbed process is also a step process.

Theorem 5: Let $0 \leqslant V(x) \leqslant M$ for all $x \in R$. The transition function:

$$p_{\epsilon,V}(t,x,\Gamma) = E_x^{\epsilon}\left\{ \chi_\Gamma(Y_t^{\epsilon}) \exp\left[-\int_0^t V(Y_s^{\epsilon}) ds \right] \right\}$$

determines the step process $Y_t^{\epsilon,V}$.

Proof: By Theorem 3.8 of Ref. 21 a sufficient condition for the existence of a Markovian step process $Y_t^{\epsilon,V}$ is that its transition function obeys

$$\lim_{t\downarrow 0} p_{\epsilon,V}(t,x,\{x\}) = 1$$

uniformly in $x \in R$.

Let us choose $t_1 > 0$ so that $1-\delta \leqslant \exp(-Mt_1)$ is secured. In view of

$$\exp(-Mt) \leqslant \exp\left[-\int_0^t V(Y_s^{\epsilon}(\omega)) ds \right] \leqslant 1$$

for all ω, we have for all $t < t_1$ the following estimate:

$$(1-\delta) p_{\epsilon}(t,x,\Gamma) \leqslant p_{\epsilon,V}(t,x,\Gamma) \leqslant p_{\epsilon}(t,x,\Gamma).$$

On the other hand, there exists t_2 such that for all $t < t_2$,

$$p_{\epsilon}(t,x,\{x\}) \geqslant 1-\delta$$

is valid for all $x \in R$.

Hence, for all $t < \min(t_1,t_2)$ we get

$$(1-\delta)^2 \leqslant p_{\epsilon,V}(t,x,\{x\}) \leqslant 1.$$

Because δ is arbitrary, after taking $\delta \to 0$, the assertion follows.

From the formula $p_{\epsilon,V}(t,x,\Gamma) \leqslant p_{\epsilon}(t,x,\Gamma)$ we conclude that the transition function $p_{\epsilon,V}(t,x,\Gamma)$ is absolutely continuous with respect to the Lebesgue measure, and hence posesses a density $k_{\epsilon,V}(t,x,y)$.

A new process $X_t^{\epsilon,V}$ can be defined by considering a multiplicative transformation of the process $Y_t^{\epsilon,V}$ by means of

$$\alpha_s^t = \frac{\theta^{\epsilon}(Y_t^{\epsilon,V},t)}{\theta^{\epsilon}(Y_s^{\epsilon,V},s)}, \tag{26}$$

where θ^{ϵ} is a positive solution of $\partial_t \theta^{\epsilon} = |\nabla|_{\epsilon} \theta^{\epsilon} + V\theta^{\epsilon}$.

The transition probability density of $X_t^{\epsilon,V}$ reads

$$p_{\epsilon,V}(s,y,t,x) = k_{\epsilon,V}(t-s,y,x) \frac{\theta^{\epsilon}(x,t)}{\theta^{\epsilon}(y,s)} \tag{27}$$

and by repeating arguments mimicking those of Sec. II, one can show that the perturbed step process $X_t^{\epsilon,V}$ converges in distribution to the perturbed Cauchy process X_t^V, when $\epsilon \to 0$, uniformly in $t \in [0,T]$.

A concise summary of all mathematical arguments of Secs. II and III, reads:

Corollary 2:

(a) The Schrödinger boundary-data and interpolation problem (3)–(6) admits a class of unique solutions in terms of Markov stochastic processes, for each concrete choice of the (Feynman–Kac) kernel function that is determined by the Cauchy generator plus a locally bounded, positive and measurable potential function.

(b) The pertinent processes are of the jump-type and arise as suitable limits of *step* processes. In particular, the uniform in time $t\in[0,T]$ convergence in distribution to the perturbed Cauchy process X_t^V is established, when the potential function is bounded.

Proof: To settle (a), the strictly positive and continuous Feynman–Kac kernel function is necessary. Theorem 1 refers to the Cauchy process, and its conditional relatives. Lemmas 8 and 9 refer to the perturbed Cauchy process.

The step-process approximations and their convergence to the corresponding jump-type processes are established in Theorems 2 through 4 in case of the Cauchy and conditional Cauchy processes. Theorem 5 together with Eqs. (26) and (27) are the key ingredients for the convergence-in-distribution argument.

Remark 1: The developed techniques can be used to investigate the existence issue (including that of the step process approximation) of more general jump-type processes, in particular those related to the quantum evolution with relativistic Hamiltonians.[5,30]

Remark 2: In the present paper, to simplify calculations and to make formulas more transparent, we have considered processes associated with the Cauchy generator (and thus with the α-stable symmetric process as a major tool) in space dimension 1. A glance at the construction of solutions of the Schrödinger problem makes clear that the previous limitations are inessential. In fact, we could consider any $\alpha\in(0,2)$-symmetric stable processes on R^n, for arbitrary $n\geqslant 1$, and secure the strict positivity and joint continuity in space variables of the corresponding transition density. Such properties for $n\geqslant 2$ and for potentials from the Kato class $K_{n,\alpha}$ were established in a very recent publication, Ref. 31, Theorems 3.3 and 3.5. However, an issue of sample path properties and of step-process approximations must be settled separately.

ACKNOWLEDGMENTS

The authors would like to thank Professor Ph. Blanchard, whose hospitality at the Bielefeld-Bonn-Stochastic Research Center at the University of Bielefeld made this collaboration possible. R.O. would like to thank the Alexander von Humboldt Foundation for financial support. P.G. would like to thank Professor R. Carmona and Professor W. Woyczynski for correspondence on the issue of Feynman–Kac kernels in non-Wiener contexts.

[1] E. Schrödinger, Ann. Inst. Henri Poincaré **2**, 269 (1932).
[2] B. Jamison, Z. Wahrscheinlichkeitstheor. Verwandte Geb. **30**, 65 (1974).
[3] J. C. Zambrini, J. Math. Phys. **27**, 2307 (1986).
[4] J. C. Zambrini, in *Chaos—The Interplay Between Stochastic and Deterministic Behaviour*, LNP Vol. 457, edited by P. Garbaczewski, M. Wolf, and A. Weron (Springer, Berlin, 1995).
[5] P. Garbaczewski, J. R. Klauder, and R. Olkiewicz, Phys. Rev. E **51**, 4114 (1995).
[6] P. Garbaczewski and R. Olkiewicz, J. Math. Phys. **37**, 732 (1996).
[7] P. Garbaczewski and G. Kondrat, Phys. Rev. Lett. **77**, 2608 (1996).
[8] Ph. Blanchard, P. Garbaczewski, and R. Olkiewicz, J. Math. Phys. **38**, 1 (1997).
[9] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1978), Vol. IV.
[10] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[11] R. Carmona, in *Schrödinger Operators*, LNP Vol. 345, edited by H. Holden and A. Jensen (Springer, Berlin, 1989), p. 65.
[12] R. Carmona, W. C. Masters, and B. Simon, J. Funct. Anal. **91**, 117 (1990).
[13] N. Jacob, *Pseudo-Differential Operators and Markov Processes* (Akademie, Verlag, Berlin, 1996).
[14] *Lévy Flights and Related Topics in Physics*, LNP Vol. 450, edited by M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch (Springer, Berlin, 1995).
[15] B. J. West and V. Seshadri, Physica A **113**, 203 (1982).
[16] F. Monti and H. R. Jauslin, J. Stat. Phys. **60**, 413 (1990).
[17] C. Tsallis *et al.*, Phys. Rev. Lett. **75**, 3589 (1995).
[18] G. Trefán *et al.*, Phys. Rev. E **50**, 2564 (1994).
[19] A. Posilicano, Lett. Math. Phys. **42**, 85 (1997).

[20] A. Janicki and A. Weron, *Simulation and Chaotic Behaviour of α-stable Stochastic Processes* (Dekker, New York, 1994).
[21] E. B. Dynkin, *Markov Processes* (Springer, Berlin, 1965), Vol. I.
[22] I. I. Gihman and A.-V. Skorohod, *Introduction to the Theory of Random Processes* (Saunders, Philadelphia, 1969).
[23] L. Breiman, *Probability* (Addison–Wesley, Reading, MA, 1968).
[24] P. Billingsley, *Probability and Measure* (Wiley, New York, 1979).
[25] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes* (Springer, Berlin, 1987).
[26] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence* (Wiley, New York, 1986).
[27] J. Bertoin, *Lévy Processes* (Cambridge University Press, Cambridge, 1996).
[28] B. Simon, Bull. Am. Math. Soc. **7**, 447 (1982).
[29] W. E. Pruitt, Ann. Prob. **9**, 948 (1981).
[30] G. F. De Angelis, J. Math. Phys. **31**, 1408 (1990).
[31] Z. Q. Chen and R. Song, J. Funct. Anal. **150**, 204 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Howe duality for the quantum groups $\mathcal{U}_q\mathrm{u}(m,n)$, $\mathcal{U}_q\mathrm{u}(M)$

Gavin Green
Department of Mathematics, School of Mathematics and Statistics, University of Newcastle, Newcastle upon Tyne NE1 7RU, England

(Received 30 March 1998; accepted for publication 8 July 1998)

A dual pair is defined to be a pair of mutually centralizing subgroups of the real symplectic group. Let (G,G') be a dual pair of reductive groups in which G is also compact and consider the decomposition of the metaplectic representation of the symplectic group into GG'-irreducibles. Each such irreducible is a tensor product of an irreducible of G with one of G', and it turns out there is a bijective correspondence between them, a particular irreducible of G only occurring with a particular one of G' and vice versa. This so-called Howe duality is here generalized to the quantum group case $(\mathcal{U}_q\mathrm{u}(m,n),\mathcal{U}_q\mathrm{u}(M))$ for q not a root of unity. A metaplectic representation for this dual pair is given in terms of the q-oscillator algebra and a $\mathcal{U}_q\mathrm{u}(m,n)\times\mathcal{U}_q\mathrm{u}(M)$-covariant Heisenberg–Weyl algebra is also constructed and realized on Fock space. The heart of the proof lies in showing that $\mathcal{U}_q\mathrm{u}(m,n)$ and $\mathcal{U}_q\mathrm{u}(M)$ are essentially mutual commutants on Fock space. The duality follows using the compactness of $\mathcal{U}_q\mathrm{u}(M)$. The proof is independent of the classical theory. As a consequence, given any two unitary highest weight representations of a quantum pseudo-unitary group (which arise from the restriction of a metaplectic representation), the decomposition of their tensor product is the same as in the classical case. [S0022-2488(98)02511-0]

I. INTRODUCTION

The metaplectic or oscillator representation of the real symplectic group first arose in the quantum field theory work of Shale and Segal.[1] Let $\mathcal{W}(n)$ denote the real Heisenberg–Weyl algebra generated by creation and annihilation operators, a_i^+, a_j $(i,j=1,...,n)$, satisfying the standard canonical commutation relations with $\hbar=1$,

$$[a_i,a_j]=[a_i^+,a_j^+]=0, \quad [a_i,a_j^+]=\delta_{ij}$$

and with real form $(a_i)^*=a_i^+$, $(a_i^+)^*=a_i$.

The complex symplectic Lie algebra $\mathrm{sp}(2n,\mathbb{C})$ may be realized as the complex span of the symmetrized quadratic expressions

$$a_ia_j, \quad \tfrac{1}{2}(a_ia_j^+ + a_j^+a_i), \quad a_i^+a_j^+,$$

with the Lie bracket defined as the commutator. The real form of $\mathcal{W}(n)$ picks out the real symplectic algebra, $\mathrm{sp}(2n,\mathbb{R})$. The Fock space action of $\mathcal{W}(n)$ then gives a unitary, infinite dimensional representation of the real symplectic Lie algebra, the infinitesimal version of the metaplectic representation.

It turns out that this exponentiates to a projective representation of the real symplectic group $\mathrm{Sp}(2n,\mathbb{R})$, only becoming single valued on the twofold covering group, the metaplectic group. It is an exact analog of the spin representation of the orthogonal group. At the group level there exists a cleaner abstract definition of the metaplectic representation as the set of intertwining operators of equivalent (by Stone–von Neumann theorem) representations of the Heisenberg group. We shall resist the temptation to give any details (see Refs. 2, 3), since we shall deal with quantum groups at the enveloping algebra level where the infinitesimal version is more suitable.

Further investigation of the metaplectic representation, its restriction to subgroups and the tensor products of representations so obtained, subsequently appeared in the literature, for example (Ref. 4). In particular, since we may identify $\mathbb{R}^{2(m+n)}\simeq\mathbb{C}^{m+n}$ such that the symplectic form is the imaginary part of a pseudo-Hermitian form on the complex space, we see that $\mathrm{U}(m,n)\subset \mathrm{Sp}(2(m+n),\mathbb{R})$. The pseudo-unitary group therefore has a metaplectic representation given by restriction. The M-fold tensor product of this was decomposed in Ref. 5, where the irreducibles were all shown to be unitary with highest weight. It was also conjectured and subsequently proved in Ref. 6 that in fact all such representations appear in such a decomposition for some M. Such results were encompassed and extended by Roger Howe's theory of dual pairs.[7–9]

A dual pair, (G,G'), is a pair of subgroups of the real symplectic group which are mutual centralizers. In the case that both groups are reductive we speak of a reductive dual pair. Howe proved the following theorem.

Theorem 1: *Let (G,G') be a reductive dual pair with G compact and let their corresponding Lie algebras be $\mathfrak{g}$, $\mathfrak{g}'$ respectively. The invariants in $\mathcal{W}(n)$ under the adjoint action of G (and hence which commute with G) are generated by (the image of) $\mathfrak{g}'$.*

It follows that we have the decomposition of the metaplectic μ into GG' irreducibles,

$$\mu|_{GG'} = \bigoplus \rho_i \otimes \rho_i' , \tag{1}$$

where ρ_i and ρ_i' are irreducible representations of G and G' respectively. Furthermore, there is a bijective correspondence, ρ_i only occurring with ρ_i' and vice versa. ■

The construction of dual pairs is straightforward by decomposing the underlying real symplectic vector space. Many families of dual pairs exist, the reductive ones having been classified.[8–10] Indeed every classical group arises as the member of some dual pair (ignoring centers and connected components).

For example, let $V\simeq \mathbb{C}^{m+n}$ be a pseudo-Hermitian of type (m,n) and $W\simeq \mathbb{C}^{M}$ be Hermitian. Then $V\otimes W$ is pseudo-Hermitian with respect to the natural product form, and thus $\mathrm{U}(m,n)$, $\mathrm{U}(M)\subset \mathrm{U}(mM,nM)$. Since $\mathrm{U}(mM,nM)\subset \mathrm{Sp}(2(m+n)M,\mathbb{R})$, as we saw earlier, a dual pair is obtained

$$\mathrm{U}(m,n),\mathrm{U}(M)\subset \mathrm{Sp}(2(m+n)M,\mathbb{R}). \tag{2}$$

Furthermore, the restriction of the action of $\mathrm{Sp}(2(m+n)M,\mathbb{R})$ to the subgroup $\mathrm{U}(m,n)$ is just the M-fold tensor product of its own metaplectic, that is, coming from restriction of the metaplectic of $\mathrm{Sp}(2(m+n),\mathbb{R})$. The duality theory then allows one to study the decomposition of this tensor product representation, (as studied in Refs. 5 and 6), in terms of the dually paired groups $\mathrm{U}(M)$. The precise theorem is given in Ref. 8, Theorem 4.6.

It is an analog of Theorem 1 for the quantum groups $\mathcal{U}_q\mathrm{u}(m,n),\mathcal{U}_q\mathrm{u}(M)$ which we shall prove. As a consequence we shall also deduce an analog of Theorem 4.6 of Ref. 8, referred to above.

Our motivations are several. First, from a mathematical point of view it would be desirable to extend the scope of the duality as far as it will go; by giving more and more examples we may shed light on a deeper underlying theory. There have already been several examples of duality which go beyond the scope of Theorem 1; for nonreductive groups,[10] super cases,[7] and modular representations.[11] In Refs. 12 and 13 independent proofs of duality were given for the pair of quantum unitary groups $(\mathcal{U}_q\mathrm{u}(2),\mathcal{U}_q\mathrm{u}(3))$. Reference 14 extends these results to the family of compact quantum unitary groups, $(\mathcal{U}_q\mathrm{u}(m),\mathcal{U}_q\mathrm{u}(M))$.

For quantum groups with q a root of unity some genuinely new representation theory occurs which appears to have similarities to the modular representations in the classical case. The example of duality for modular representations,[11] suggests that some remnant of duality may exist. Our goals presently are more modest, however, only dealing with q a nonroot of unity.

After some preliminary definitions and conventions, we recall the metaplectic representation of the quantum symplectic group and construct a new representation of the quantum pseudo-unitary group in Sec. IV. Using the fact that the classical action of $\mathrm{U}(m,n)\subset \mathrm{Sp}(2(m+n)M,\mathbb{R})$ is the M-fold tensor action, we construct in Sec. V the dual pair of quantum groups $\mathcal{U}_q\mathrm{u}(m,n),\mathcal{U}_q\mathrm{u}(M)$ and their action on Fock space.

This action of the dual pair gives an adjoint action on the endomorphism algebra of Fock space, and thus on the Heisenberg–Weyl algebra. Section VI is concerned with the construction of a deformed Heisenberg–Weyl algebra which transforms nicely under this action; a so-called covariant algebra. The heart of the proof of duality is to show that the dual pair are essentially mutual commutants in this algebra, which we do using the q analog of the Schur–Weyl duality between the general linear and symmetric groups. The duality correspondence is then deduced from the compactness of $\mathcal{U}_q\mathrm{u}(M)$.

From the duality, as in the classical case, the Clebsch–Gordan theory for the unitary highest weight representations which occur in tensor products of the metaplectic of $\mathcal{U}_q\mathrm{u}(m,n)$ (classically this is all unitary highest weight representations) is controlled by the theory for $\mathcal{U}_q\mathrm{u}(M)$ as M varies. Since the finite dimensional representation theory for quantum groups is the same as in the classical case (generically) we can conclude, as we might expect, that the decomposition of tensor products of the unitary highest weight representations of $\mathcal{U}_q\mathrm{u}(m,n)$ (occurring in tensor products of the metaplectic) is basically the same as in the classical case. For example, we shall get a handle on the ladder representations of $\mathcal{U}_q\mathrm{su}(2,2)$ which can be thought of as a quantum version of the conformal group.

It is worth noting that although one might expect that there is a duality for other pairs of quantum groups, the constructions used in the given proof do not always work; for example, a covariant Heisenberg–Weyl algebra for the dual pair $(\mathcal{U}_q\mathrm{o}(m),\mathcal{U}_q\mathrm{sp}(2n,\mathbb{R}))$ cannot be constructed in the same way.

The proof given is independent of the classical theory and thus may provide a model for a proof of some duality in the root of unity cases. Certainly some of the constructions can still be made.

II. QUANTUM GROUP CONVENTIONS

We shall assume throughout that q is a nonzero complex number which is not a root of unity. First some standard notation,

$$[n]_q=\frac{q^n-q^{-n}}{q-q^{-1}},\quad [n]_q!=[n]_q[n-1]_q\ldots[1]_q,\quad \begin{bmatrix} m \\ n \end{bmatrix}_q=\frac{[m]_q!}{[n]_q![m-n]_q!}.$$

Let a_{ij} be a symmetrizable Cartan matrix and put $q_i=q^{d_i}$, where d_i is the least positive integer such that d_ia_{ij} is symmetric.

$\mathcal{U}_q\mathfrak{g}$ is defined to be the unital, associative algebra over $\mathbb{C}$ generated by $\{e_{\pm i},k_i^{\pm 1/2}\}_{i=1}^{n-1}$ subject to

$$[k_i^{\pm 1},k_j^{\pm 1}]=[k_i^{\mp 1},k_j^{\mp 1}]=0,\quad k_ie_{\pm j}k_i^{-1}=q_i^{\pm a_{ij}}e_{\pm j},$$

$$[e_i,e_{-j}]=\delta_{ij}\,\frac{k_i-k_i^{-1}}{q_i-q_i^{-1}},$$

$$\sum_{k=0}^{1-a_{ij}}(-1)^k\begin{bmatrix} 1-a_{ij} \\ k \end{bmatrix}_q(e_{\pm i})^ke_{\pm j}(e_{\pm i})^{1-a_{ij}-k}=0,\quad i\neq j,$$

$$\Delta(e_{\pm i})=e_{\pm i}\otimes k_i^{(1/2)}+k_i^{-(1/2)}\otimes e_{\pm i},\quad \Delta(k_i^{\pm(1/2)})=k_i^{\pm(1/2)}\otimes k_i^{\pm(1/2)},$$

$$S(e_{\pm i})=-q_i^{\pm 1}e_{\pm i},\quad S(k_i^{\pm(1/2)})=k_i^{\mp(1/2)},$$

where S is the antipode and Δ the coproduct.

With the Cartan matrix of types A and C we obtain the quantum special linear and symplectic groups, respectively. For $q\in\mathbb{R}\backslash\{0,\pm 1\}$ the real forms (1) $\mathcal{U}_q\mathrm{su}(m,n)$ and (2) $\mathcal{U}_q\mathrm{sp}(2n,\mathbb{R})$ may be defined by the $*$-structures: (1) $k_i^*=k_i$ for all $i=1,...,n-1$, $e_{\pm i}^*=e_{\mp i}$ for $i=1,...,m-1,m+1,...,m+n-1$ and $e_{\pm m}^*=-e_{\mp m}$ and (2) $k_i^*=k_i$ for all $i=1,...,n$, $e_{\pm i}^*=e_{\mp i}$ for $i=1,...,n-1$, and $e_{\pm n}^*=-e_{\mp n}$. The quantum general linear group may be defined by adjoining group-like, central elements $k_\epsilon^{\pm 1}$ to the special linear group. To define $\mathcal{U}_q\mathrm{u}(m,n)$ we put $(k_\epsilon^{\pm 1})^*=k_\epsilon^{\pm 1}$.

If $\mathcal{R}\in\mathcal{U}_q\mathfrak{g}\otimes\mathcal{U}_q\mathfrak{g}$ is the universal R-matrix of a QUE algebra with fundamental representation ρ, then we let $R=(\rho\otimes\rho)\mathcal{R}$. Denote the flip operator by P, set $\hat{R}=PR$, $\check{R}=RP$ and use $\hat{R}^{-1}$ and

$\check{R}^{-1}$ as shorthand for $(\hat{R})^{-1}$ and $(\check{R})^{-1}$, respectively. The standard notation $\hat{R}_{ij}$ denotes an operator on a tensor product space acting as $\hat{R}$ in the ith and jth factors, and the identity elsewhere.

In the particular case that $\mathfrak{g}=\mathrm{sl}(n,\mathbb{C})$ (or $\mathrm{u}(n)$)

$$R=\sum_{i\neq j} E^i{}_i\otimes E^j{}_j+q\sum_i E^i{}_i\otimes E^i{}_i+(q-q^{-1})\sum_{i<j} E^i{}_j\otimes E^j{}_i,$$

where $E^i{}_j$ are the standard matrices with 1 in the ijth place and zeros elsewhere. The matrix $\hat{R}$ satisfies

$$(\hat{R}-q)(\hat{R}+q^{-1})=0. \tag{3}$$

The Artin braid group, B^p, is generated by $\{\tau_i\}_{i=1}^{p-1}$ subject to

$$\tau_i\tau_{i+1}\tau_i=\tau_{i+1}\tau_i\tau_{i+1},\quad \tau_i\tau_j=\tau_j\tau_i,\quad |i-j|\leq 2,$$

and the Hecke algebra, H_q^p, is the quotient by

$$(\tau_i-q)(\tau_i+q^{-1})=0\quad i=1,...,p-1.$$

The following analog of the Schur–Weyl duality,[15–17] will be used later.

Theorem 2: *The matrices $\hat{R}_{i(i+1)}$ for $i=1,...,p-1$ provide a representation of the Hecke algebra H_q^p by $\tau_i\to\hat{R}_{i(i+1)}$ on the space $\otimes^pV$, where V is the space on which the fundamental representation of $\mathcal{U}_q\mathrm{u}(n)$ is realized. Furthermore, in $\mathrm{End}(\otimes^pV)$ the Hecke algebra and $\mathcal{U}_q\mathrm{u}(n)$ generate mutual commutants.* ■

III. *q*-OSCILLATOR ALGEBRA

Define the deformed or q-oscillator algebra, $\mathcal{W}_q(n)$,[18–20] to be the unital, associative algebra over $\mathbb{C}$ generated by A_i, A_i^+ and $\nu_i^{\pm 1}$ for $i=1,...,n$, subject to

$$A_iA_i^+-q^{\pm 1}A_i^+A_i=\nu_i^{\mp 2},\quad \nu_i^{\pm 1}A_i^+=q^{\pm 1/2}A_i^+\nu_i^{\pm 1},\quad \nu_i^{\pm 1}A_i=q^{\mp 1/2}A_i\nu_i^{\pm 1},$$

and with $\{A_i,A_i^+,\nu_i^{\pm 1}\}$ commuting with $\{A_j,A_j^+,\nu_j^{\pm 1}\}$ for $i\neq j$. The real form is given for $q\in\mathbb{R}\backslash\{0,\pm 1\}$ by the $*$-structure $A^*=A^+$, $(A^+)^*=A$, and $(\nu_i^{\pm 1})^*=\nu_i^{\pm 1}$. (In some definitions only one of the relations between A and A^+ is assumed but both hold in the following Fock representation anyway.)

Let Ω be a vacuum vector, $A_i\Omega=0$ for all $i=1,...,n$. Introduce the shorthand notation

$$A^+(\mathbf{s})=(A_i^+)^{s_1}\ldots(A_n^+)^{s_n},\quad \mathbf{z}(\mathbf{s})=A^+(\mathbf{s})\Omega.$$

Proposition 3: There is a unitary action of $\mathcal{W}_q(n)$ on Fock space given by

$$A_i^+\mathbf{z}(\mathbf{s})=\mathbf{z}(\mathbf{s}+\mathbf{e}_i),\quad A_i\mathbf{z}(\mathbf{s})=[s_i]_q\mathbf{z}(\mathbf{s}-\mathbf{e}_i),\quad \nu_i^{\pm 1}\mathbf{z}(\mathbf{s})=q^{\pm(s_i/2)}\mathbf{z}(\mathbf{s}),$$

with the inner product, unique up to normalization,

$$\langle\mathbf{z}(\mathbf{s}),\mathbf{z}(\mathbf{t})\rangle=\delta_{\mathbf{s},\mathbf{t}}[s_1]_q!\ldots[s_n]_q!.$$

■

IV. METAPLECTIC REPRESENTATIONS

An explicit infinitesimal metaplectic or oscillator representation for the quantum symplectic group was given in Ref. 20. (A more implicit version was also given in Ref. 21 which gave the generators L of Ref. 22 in terms of a covariant Heisenberg–Weyl algebra. However, as was pointed out there, the approach did not allow one to give explicit expressions for the full generators $L^{\pm 1}$.)

Proposition 4: For $q\in\mathbb{R}\backslash\{0,\pm 1\}$ there is an algebra $*$-homomorphism

$$\mu:\mathcal{U}_q\mathrm{sp}(2n,\mathbb{R})\to\mathcal{W}_q(n),$$

determined for $i=1,...,n$ *by*

$$\mu(e_i)=\begin{cases}A_i^+A_{i+1}\\ \dfrac{-1}{q+q^{-1}}(A_n^+)^2\end{cases},\quad \mu(k_i^{1/2})=\begin{cases}\nu_i\nu_{i+1}^{-1} & i<n\\ q^{1/2}\nu_n^2 & i=n\end{cases}.$$

The Fock space representation of $\mathcal{W}_q(n)$ *therefore induces a unitary representation of* $\mathcal{U}_q\mathrm{sp}(2n,\mathbb{R})$. ■

Since classically $\mathrm{U}(m,n)$ is a subgroup of $\mathrm{Sp}(2(m+n),\mathbb{R})$ it inherits a metaplectic representation. In the quantum case this is still true for the compact quantum unitary group which sits nicely as a Hopf $*$-subalgebra in the quantum symplectic group. It is not as obvious how a general quantum pseudo-unitary group will sit inside the symplectic group so we give an independent definition of the representation, but one which coincides with the restriction in the classical limit. The proof is straightforward.

Proposition 5: For $q\in\mathbb{R}\backslash\{0,\pm1\}$ *there is a* $*$*-homomorphism,*

$$\mu:\mathcal{U}_q\mathrm{u}(m,n)\to\mathcal{W}_q(m+n),$$

determined for $i=1,...,m+n-1$ *by*

$$\mu(e_i)=\begin{cases}A_i^+A_{i+1}\\ -A_m^+A_{m+1}^+,\\ -A_iA_{i+1}^+\end{cases}\quad \mu(k_i^{1/2})=\begin{cases}\nu_i\nu_{i+1}^{-1} & i<m\\ q^{1/2}\nu_m\nu_{m+1} & i=m,\\ \nu_{i+1}\nu_i^{-1} & i>m\end{cases}$$

and with $\mu(k_\epsilon^{\pm1})=(\nu_1^2\ldots\nu_m^2\nu_{m+1}^{-2}\ldots\nu_{m+n}^{-2})^{\pm1}$.

The Fock space representation of $\mathcal{W}_q(n)$ *therefore induces a unitary representation of* $\mathcal{U}_q\mathrm{sp}(2n,\mathbb{R})$. ■

Note that the image of $\mathcal{U}_q\mathrm{u}(m,n)$ is still a $*$-subalgebra of the image of $\mathcal{U}_q\mathrm{sp}(2(m+n),\mathbb{R})$.

The dual representation μ^* is given by $\mu^*(h)=(\mu(S(h)))^*$ for all $h\in\mathcal{U}_q\mathrm{u}(m,n)$. In the case $n=0$ we normalize this to obtain the equivalent representation $\mu'\simeq\mu^*$ given by

$$\mu'(e_i)=A_iA_{i+1}^+,\quad \mu'(k_i^{1/2})=\nu_{i+1}\nu_i^{-1},\quad i=1,...,m-1.$$

V. DUAL PAIR $\mathcal{U}_q\mathrm{u}(m,n)$, $\mathcal{U}_q\mathrm{u}(M)$

For the construction of dual pairs in the classical case one can decompose the underlying symplectic vector space and so rely on the groups being defined by their fundamental representations. This is no longer true in the quantum case and there are considerable difficulties in trying to construct quantum dual pairs inside the quantum symplectic group (or alternatively starting with a deformed dual pair and trying to construct a deformed symplectic group around them). We shall not address these problems further here (see Refs. 12, 14), and so avoid giving a definition of a dual pair in the quantum case.

Instead we use the fact that the action of the pseudo-unitary group $\mathrm{U}(m,n)\subset\mathrm{Sp}(2(m+n)M,\mathbb{R})$ obtained by restriction of the metaplectic of $\mathrm{Sp}(2(m+n)M,\mathbb{R})$ is just the M-fold tensor action of its own metaplectic. The action of the dually paired $\mathrm{U}(M)\subset\mathrm{Sp}(2(m+n)M,\mathbb{R})$ is $\otimes^m\mu\otimes\otimes^n\mu^*$. We can therefore define the action of the dual pair without reference to the symplectic group at all. A slightly modified version of this works in the quantum case.

Denote the generators of $\mathcal{U}_q\mathrm{u}(m,n)$ by $\{e_{\pm i},k_i^{\pm(1/2)},k_\epsilon^{\pm1}\}_{i=1}^{m+n-1}$ and those of $\mathcal{U}_q\mathrm{u}(M)$ by $\{E_{\pm I},K_I^{\pm(1/2)},K_\epsilon^{\pm1}\}_{I=1}^{M-1}$.

Proposition 6: There exists a $*$*-algebra homomorphism,*

$$\mathcal{U}_q\mathrm{u}(m,n)\otimes\mathcal{U}_q\mathrm{u}(M)\to\mathcal{W}_q((m+n)M).$$

The unitary action of $\mathcal{W}_q((m+n)M)$ *on Fock space gives an action of the dual pair such that as* $q\to1$ *we obtain the classical metaplectic representation.*

Proof: Denote the metaplectic representations of $\mathcal{U}_q\mathrm{u}(m,n)$ and $\mathcal{U}_q\mathrm{u}(M)$ by $\mu_{(m,n)}$ and μ_M, respectively, so we have $*$-algebra homomorphisms into q-oscillator algebras

$$\mu_{(m,n)}:\mathcal{U}_q\mathrm{u}(m,n)\rightarrow\mathcal{W}_q(m+n),\quad \mu_M:\mathcal{U}_q\mathrm{u}(M)\rightarrow\mathcal{W}_q(M).$$

The tensor product representations $\otimes^M\mu_{(m,n)}$ and $\otimes^m\mu_M\otimes\otimes^n\mu'_M$ can be regarded as $*$-homomorphisms into $\mathcal{W}_q((m+n)M)$ by using the algebra isomorphisms

$$\mathcal{W}_q((m+n)M)\simeq\otimes^M\mathcal{W}_q(m+n)\simeq\otimes^{m+n}\mathcal{W}_q(M).$$

It remains to be proven that these actions of $\mathcal{U}_q\mathrm{u}(m,n)$ and $\mathcal{U}_q\mathrm{u}(M)$ commute. The relations between the Cartan subalgebra of $\mathcal{U}_q\mathrm{u}(m,n)$ (generated by the K's) with $\mathcal{U}_q\mathrm{u}(M)$ are trivial, and vice versa. The remaining relations can be proven directly for the generators. However, much work can be saved by noting that the actions of the maximal compact subgroups of each preserve the degree and are in fact subrepresentations of the tensor algebra of the fundamental representation. It is therefore sufficient to check that their actions commute on $V^{(m+n)}\otimes V^M$, the fundamental representation of the dual pair. Finally one may prove directly that

$$[e_m,E_{\pm I}]=[e_{-m},E_{\pm I}]=0,\quad I=1,...M-1.$$

■

Note that the image of the dual pair still forms a $*$-subalgebra of the image of $\mathcal{U}_q\mathrm{sp}(2(m+n)M,\mathbb{R})$. Also, the central $\mathcal{U}_q\mathrm{u}(1)$'s of each member are mapped to the same image; this is to be expected since in the classical case the central U(1) lies in the intersection of the dual pair inside the symplectic group.

VI. DUAL PAIR-COVARIANT HEISENBERG–WEYL ALGEBRAS

An algebra is said to be covariant with respect to the action of a Hopf algebra if the multiplication intertwines with this action. A deformation of the Heisenberg–Weyl algebra which is covariant with respect to the action of some quantum group will therefore be referred to as a covariant Heisenberg–Weyl algebra.

The most familiar examples are the $\mathcal{U}_q\mathrm{u}(n)$-covariance of Ref. 23 and $\mathcal{U}_q\mathrm{o}(n)$-covariance discussed in Ref. 24. In Ref. 14 the $\mathcal{U}_q\mathrm{u}(n)$-covariance is extended to the full $\mathcal{U}_q\mathrm{sp}(2n,\mathbb{R})$-covariance. Using the braided theory as presented in Ref. 25 we can regard the Pusz–Woronowicz algebra as a $\mathcal{U}_q\mathrm{sp}(2n,\mathbb{R})$ braided vector space, and thus take braided tensor products of it. In this way we could construct a Heisenberg–Weyl algebra which is covariant with respect to the action of one member of the dual pair. (We should also be able to use a braided exponential to obtain Weyl-type relations, that is, exponentiated canonical commutation relations.)

However, we can actually do slightly better for this dual pair and consider a deformation of the Heisenberg–Weyl algebra which is covariant under the joint action of the pair $\mathcal{U}_q\mathrm{u}(m,n)$, $\mathcal{U}_q\mathrm{u}(M)$. This is a generalization of the $\mathcal{U}_q\mathrm{u}(m)\otimes\mathcal{U}_q\mathrm{u}(M)$-covariance considered in Refs. 26 and 27 and also Ref. 14.

Proposition 7: The algebra $\mathcal{W}_{AA}((m,n),M)$ *with generators* $\{v_{iI},u^{iI}\}$ *for* $i=1,...,m+n$ *and* $I=1,...,M$ *satisfying the relations*

$$\mathbf{v}_1\mathbf{v}_2(\hat{r}^{-1}\hat{R}-1)=0,\quad(\check{r}^{-1}\check{R}-1)\mathbf{v}^*{}_1\mathbf{v}^*{}_2=0,\tag{4}$$

$$\mathbf{v}^*{}_1\mathbf{v}_2-\mathbf{v}_2r_{21}R_{21}\mathbf{v}^*{}_1=I,\tag{5}$$

is $\mathcal{U}_q\mathrm{gl}(m+n,\mathbb{C})\otimes\mathcal{U}_q\mathrm{gl}(M,\mathbb{C})$*-covariant, has the same Poincaré series as the classical Heisenberg–Weyl algebra, and is a* $*$*-algebra with respect to*

$$v_{iI}^*=\begin{cases}v^{iI} & i=1,...,m\\ -v^{iI} & i=m+1,...,m+n\end{cases}.$$

Before giving the proof we explain the notation. The subscript *AA* refers to the Cartan-type of $\mathcal{U}_q\mathrm{gl}(m+n,\mathbb{C})\otimes\mathcal{U}_q\mathrm{gl}(M,\mathbb{C})$ (which defines the braided or quasitensor category within which we work). The R-matrices of $\mathcal{U}_q\mathrm{gl}(m+n,\mathbb{C})$ and $\mathcal{U}_q\mathrm{gl}(M,\mathbb{C})$ (or their corresponding real forms) will be denoted by r and R, respectively, and thus that of the product will be $r_{13}R_{24}$, though we shall generally suppress the subscripts.

$((m,n),M)$ is used to distinguish the real form of $\mathcal{W}_{AA}$ we are considering, for compatability with $\mathcal{U}_q\mathrm{u}(m,n)\otimes\mathcal{U}_q\mathrm{u}(M)$. In the future we shall simply use the notation $\mathcal{W}_{AA}$ since the real form shall be understood.

Let $V=V^{(m+n)}\otimes V^M$ denote the fundamental representation (space) of the dual pair, and V^* its dual. Then the q-analogs of the symmetric algebras will be denoted $S_{AA}(V)$ and $S_{AA}(V^*)$, respectively.

The coordinate free notation in the relations is standard, see, for example, Ref. 25. $\mathbf{v}$ is thought of as a row vector with entries v_{iI}, $\mathbf{v}^*$ a column vector with entries v^{iI}; $\mathbf{v}_1=\mathbf{v}\otimes 1$ and $\mathbf{v}_2=1\otimes\mathbf{v}$. The R-matrices r and R lie in a tensor product of matrix spaces and the coordinate free notation is simply matrix and vector notation.

Proof: Let $\{v_{iI}\}$ be a basis for V with respect to which the representing matrices take the standard form, and let $\{v^{iI}\}$ be the dual basis of V^*. We wish to consider covariant algebras generated by these. The construction of such algebras is discussed for example in Ref. 25.

By analogy with the classical case, we would like the v_{iI} to generate a deformation of the symmetric algebra. Consideration of the eigenvalues and eigenspaces of $\hat{r}\hat{R}$, $\hat{r}\hat{R}^{-1}$ and $\hat{r}^{-1}\hat{R}$, using (3), leads to the relations (4). One can similarly construct a symmetric algebra from the dual basis v^{iI}.

It also follows from

$$(\hat{r}\hat{R}+1)(\hat{r}\hat{R}^{-1}-1)=0$$

that this algebra is a braided vector space as defined in Ref. 25, with respect to the braid matrix $\hat{r}\hat{R}$. This allows us to define the cross-relations (5) by braided differentiation; see also (Refs. 28, 14).

It is straightforward to check (using Bergmann's diamond lemma as applied in Ref. 29) that the Poincaré series (graded dimensions) of $S_{AA}(V)$ and $S_{AA}(V^*)$ are the same as the classical symmetric algebras. It then follows from the Wick ordering property given in the next proposition that the Poincaré series of $\mathcal{W}_{AA}$ is the same as that of the classical Heisenberg–Weyl algebra. ■

In order to be able to define a Fock representation of $\mathcal{W}_{AA}$ we need to check that it is a Wick algebra;[28] in other words the relations must allow us to order any expression such that the annihilation operators lie to the right of any creation operators. With foresight we define

$$a^+_{iI}=\begin{cases} v_{iI} \\ q^{(i+I-1-m)}v^{iI}, \end{cases} \quad a_{iI}=\begin{cases} v^{iI} & i\leqslant m \\ -q^{(i+I-1-m)}v_{iI} & i>m \end{cases}, \tag{6}$$

so that $a^*_{iI}=a^+_{iI}$ for all (iI).

Proposition 8: $\mathcal{W}_{AA}$ is a Wick algebra.

Proof: We wish to prove that any expression can be reordered so that all the a's lie to the right of all the a^+'s. By induction it is sufficient to prove that all the expressions $a_{iI}a^+_{jJ}$ may be so ordered.

There are four cases to consider, depending on the values of i, and j. In cases $i,j\leqslant m$ or $i\leqslant m,j>m$ or $i>m,j\leqslant m$, the explicit relations obtained from (4) and (5) may be used directly to give the result.

In the final case $i,j>m$ we need to invert the last relation by multiplying by the matrix $(r^{t_2}_{21})^{-1}(R^{t_2}_{21})^{-1}$, where t_2 denotes transposition in the second factor, and we use $R^{t_2}_{21}$ to mean $(R_{21})^{t_2}$. Explicitly $(R^{t_2}_{21})^{-1}$ is given by

$$\sum_{i\neq j} E^i{}_i\otimes E^j{}_j+q^{-1}\sum_i E^i{}_i\otimes E^i{}_i+(q^{-1}-q)\sum_{i>j} q^{-2(i-j)}E^i{}_j\otimes E^i{}_j\,.$$

Some of the terms obtained by this inversion will still not be ordered, but may then be ordered as previously, and the result follows.

Note that our choice of factors in defining the creation and annihilation operators was such that in the expressions for reordering $a_{iI}a_{iI}^{+}$, the constant term which appears will simply be 1.■

Now we can construct a Fock representation in the usual way by considering the existence of a vacuum vector Ω which is killed by the annihilation operators, $a_{iI}\Omega=0$ for all (iI).

Define an ordering of the double indices by $(iI)<(jJ)$ if $i<j$ or $(i=j$ and $I<J)$. The Fock space may then be built on the vacuum by the creation operators; we shall use the shorthand notation

$$a^{+}(\mathbf{s})=(a_{11}^{+})^{s_{11}}\dots(a_{m+nM}^{+})^{s_{m+nM}},\quad a(\mathbf{s})=(a^{+}(\mathbf{s}))^{*},$$

where the ordering on the right-hand side of the first equation is as given above.

Lemma 9: Let $s_{jJ}=0$ for all $(jJ)\leqslant(iI)$. *Then*

$$a_{iI}a^{+}(\mathbf{s})\Omega=0,$$

$$(a(\mathbf{s}))a^{+}(\mathbf{s})\Omega=(s_{11})_{q^{-1}}!\dots(s_{mM})_{q^{-1}}!(s_{m+11})_{q}!\dots(s_{m+nM})_{q}!\Omega,$$

where the second q-integer is defined to be

$$(n)_q=\frac{1-q^{-2n}}{1-q^{-2}}.$$

Proof: The first is a direct consequence of the relations. The second may be proved by an induction on $\Sigma_{iI}s_{iI}$ using the first. ■

Proposition 10: There is up to normalization a unique inner product such that the Fock representation is a $*$*-representation; explicitly*

$$(a^{+}(\mathbf{s})|a^{+}(\mathbf{t}))=\delta_{\mathbf{s},\mathbf{t}}(s_{11})_{q^{-1}}!\dots(s_{mM})_{q^{-1}}!(s_{m+11})_{q}!\dots(s_{m+nM})_{q}!.$$

It is irreducible, faithful, and unitary for $q\in\mathbb{R}\backslash\{0,\pm1\}$ (*i.e., the inner product is positive definite*).

Proof: The inner product follows from the previous lemma.

To prove that the representation is irreducible assume that W is an invariant subspace and let $\Sigma_{\mathbf{r}}b_{\mathbf{r}}\mathbf{x}(\mathbf{r})$ be any nonzero vector in W. Define an ordering on $\mathbb{N}^{(m+n)M}$ by $\mathbf{s}>\mathbf{r}$ if for some iI, $s_{iI}>r_{iI}$ and for all $jJ<iI$ we have $s_{jJ}=r_{jJ}$. Define $\mathbf{s}$ to be the maximum element of $\{\mathbf{r}|b_{\mathbf{r}}\neq0\}$. Then

$$a(\mathbf{s})\sum_{\mathbf{r}}b_{\mathbf{r}}\mathbf{x}(\mathbf{r})=b_{\mathbf{s}}(s_{11})_{q^{-1}}!\dots(s_{mM})_{q^{-1}}!(s_{m+11})_{q}!\dots(s_{m+nM})_{q}!\Omega.$$

This is nonzero if q is a nonroot of unity. The result then follows from the fact that Ω is a generating vector for the module.

For the representation to be faithful we consider

$$\sum_{\mathbf{s},\mathbf{t}}\alpha_{\mathbf{s},\mathbf{t}}a^{+}(\mathbf{s})a(\mathbf{t})\mathbf{x}(\mathbf{r})=0,$$

for all $\mathbf{r}$. If we let $\mathbf{t}_0$ be the least $\mathbf{t}$ such that $\alpha_{\mathbf{s},\mathbf{t}}\neq0$ then we deduce that $\alpha_{s,\mathbf{t}_0}=0$ and by an inductive argument that all $\alpha_{\mathbf{s},\mathbf{t}}=0$, hence the result.

For unitarity it is sufficient to note that for $q\in\mathbb{R}\backslash\{0,\pm1\}$, $(n)_q>0$ for all $n\in\mathbb{N}$. ■

In summary, we have an abstract $*$-algebra, $\mathcal{W}_{AA}$, which is acted on covariantly by the dual pair such that the generators transform as the fundamental and dual representations. In addition, we have actions of both $\mathcal{W}_{AA}$ and the dual pair on Fock space. There is therefore a second action of the dual pair on $\mathcal{W}_{AA}$ using the adjoint representation. Recall that if $\mathcal{H}$ is a Hopf algebra acting on a vector space V, then there is an associated adjoint action of $\mathcal{H}$ on $\mathrm{End}(V)$ given by

$$h(a)=\sum_{(h)}h^{(1)}aS(h^{(2)}),\quad \text{for all}\quad h\in\mathcal{H},a\in\mathrm{End}(V),\tag{7}$$

where $\Delta(h)=\Sigma_{(h)}h^{(1)}\otimes h^{(2)}$, and where we suppress the representation V.

We would like these actions to coincide, as in the classical case.

Proposition 11: Let $\mathcal{U}_q\mathrm{u}(m,n)\otimes\mathcal{U}_q\mathrm{u}(M)$ *act on the image of* $\mathcal{W}_{AA}$ *on Fock space using the adjoint action. Then the images of* v_{iI} *and* v^{iI} *still transform as the basis of the fundamental and dual representations (or in other words form the components of tensor operators for the fundamental and dual representations).*

Proof: First, note that $v_{11}=a_{11}^{+}=A_{11}^{+}q^{(1/2)N_{11}}$ on Fock space. One can also easily check directly that it is a highest weight vector for the fundamental representation under the adjoint action of the dual pair. Similarly $v^{11}=a_{11}=q^{(1/2)N_{11}}A_{11}$ is a lowest weight vector for the dual representation. We can now set $\tilde{v}_{11}=v_{11}$, $\tilde{v}^{11}=v^{11}$ and from this starting point construct operators $\tilde{v}_{iI}$ and $\tilde{v}^{iI}$ by applying the generators of the dual pair in the adjoint action. Bases for the fundamental and dual are assured by considering the classical limit and the finite dimensional representation theory of quantum groups at nonroots of unity. We may then define $\tilde{a}_{iI}^{+}$ and $\tilde{a}_{iI}$ by the same formulas as before, (6), only with tildes on each element.

The operators $\tilde{v}_{iI}$ must satisfy the same relations as v_{iI} since they are constructed to transform in the same way. In addition, an easy induction shows that

$$\tilde{a}^{+}(\mathbf{s})\Omega=c(\mathbf{s})A^{+}(\mathbf{s})\Omega,$$

for some constant $c(\mathbf{s})$. It follows by a dimensional argument that in fact $\tilde{v}_{iI}$ cannot satisfy any extra relations to those of v_{iI}. By applying the $*$-structure this is true of $\tilde{v}^{iI}$ also.

The action of $\tilde{a}_{iI}^{+}$ and $\tilde{a}_{iI}$ on $\tilde{a}^{+}(\mathbf{s})\Omega$ will be the same (under the obvious identification) as a_{iI}^{+} and a_{iI} on $a^{+}(\mathbf{s})\Omega$ since their Fock actions are completely determined by the relations and the fact that the vacuum is killed by the annihilation operators. By identifying the bases $a^{+}(\mathbf{s})\Omega$, $\tilde{a}^{+}(\mathbf{s})\Omega$ and $c(\mathbf{s})A^{+}(\mathbf{s})\Omega$ we can finally conclude that

$$\tilde{a}_{iI}^{+}=a_{iI}^{+},\quad \tilde{a}_{iI}=a_{iI}.$$

■

VII. DUALITY FOR $(\mathcal{U}_q\mathrm{u}(m,n),\mathcal{U}_q\mathrm{u}(M))$

The essential point to show is that the algebra of elements in $\mathcal{W}_{AA}$ which commute with the action of $\mathcal{U}_q\mathrm{u}(M)$ on Fock space, are actually generated by the image of elements of $\mathcal{U}_q\mathrm{u}(m,n)$. The proof of duality follows easily from this result by using the compactness of $\mathcal{U}_q\mathrm{u}(M)$ and the density of $\mathcal{W}_{AA}$ on Fock space.

Recall that an invariant, a, of a general Hopf algebra, $\mathcal{H}$, with counit ϵ, is an object which transforms according to

$$h(a)=\epsilon(h)a,\quad \forall h\in\mathcal{H}. \tag{8}$$

With this definition, we may begin with the following theorem as proved in Ref. 30.

Theorem 12: Let a Hopf algebra $\mathcal{H}$ *and an algebra* $\mathcal{A}$ *act on some vector space* V, *then* $\mathcal{H}$ *acts on* $\mathcal{A}$ *in the adjoint action. The* $\mathcal{H}$ *invariants are just those elements of* $\mathcal{A}$ *which commute with* $\mathcal{H}$ *on* V. ■

The elements in $\mathcal{W}_{AA}$ with which $\mathcal{U}_q\mathrm{u}(M)$ commutes are therefore just the invariants under the q-adjoint action. Let V denote the fundamental representation of the dual pair, V^* its dual. Now, there exists a $\mathcal{U}_q\mathrm{u}(M)$-module isomorphism

$$\mathcal{W}_{AA}\simeq S_{AA}(V)\otimes S_{AA}(V^*)\subset T(V)\otimes T(V^*),$$

with the q-adjoint action on $\mathcal{W}_{AA}$ and the tensor action on the others. So now we need only look for invariants in $T(V)\otimes T(V^*)$ and restrict them to $\mathcal{W}_{AA}$ using the identification with $S_{AA}(V)\otimes S_{AA}(V^*)$. Since $\mathcal{U}_q\mathrm{u}(M)$ preserves the natural grading, a general invariant will be a sum of homogeneous invariants and furthermore, by considering the action of the central $\mathcal{U}_q\mathrm{u}(1)$ we may restrict our search to $\otimes^{p}V\otimes\otimes^{p}V^*$. Now this question is answered by the q-Schur–Weyl theorem, Theorem 2. It then remains to show that these invariants (restricted to the Weyl algebra) are generated by the quadratic ones and that the quadratic invariants lie in the image of $\mathcal{U}_q\mathrm{u}(m,n)$.

Theorem 13: The $\mathcal{U}_q\mathrm{u}(M)$-invariants in $\mathcal{W}_{AA}$ are generated by the quadratic invariants. Furthermore, these quadratic invariants lie inside the image of $\mathcal{U}_q\mathrm{u}(m,n)$.

Proof: As outlined above, we need only look for invariants in $\otimes^p V \otimes \otimes^p V^* \simeq \mathrm{End}(\otimes^p V)$ and applying Theorem 12 we see that the required invariants are those elements of $\mathrm{End}(\otimes^p V)$ which commute with the action of $\mathcal{U}_q\mathrm{u}(M)$. However, these are given by Theorem 2,

$$\sum_I \tau(v_{i_1 I_1} \otimes v_{i_2 I_2} \otimes \cdots \otimes v_{i_p I_p}) \otimes (v^{j_p I_p} \otimes \cdots \otimes v^{j_1 I_1}),$$

where τ is any element of the Hecke algebra, $\{v_{iI}\}$ and $\{v^{iI}\}$ are bases for V and V^*, respectively. The restriction of these to $S_{AA}(V) \otimes S_{AA}(V^*)$ may then be identified with the elements

$$\sum_I \tau(v_{i_1 I_1} v_{i_2 I_2} \cdots v_{i_p I_p})(v^{j_p I_p} \cdots v^{j_1 I_1}),$$

of $\mathcal{W}_{AA}$ (abusing notation). Now using the fact that the indices $i_1,\ldots,i_p$ and $j_1,\ldots,j_p$ are arbitrary and the relations of the v_{iI}, (4), we see in fact that all order $2p$ invariants can be written as linear combinations of

$$\sum_{\mathbf{I}} (v_{i_1 I_1} v_{i_2 I_2} \cdots v_{i_p I_p})(v^{j_p I_p} \cdots v^{j_1 I_1}).$$

We now wish to show that all such invariants are generated by the quadratic invariants, $\Sigma_I v_{iI} v^{jI}$, for which we use induction on the order of the invariant. Trivially the quadratic invariants themselves are quadratic, and for the induction hypothesis we assume that all invariants of order $2(p-1)$ or less are generated by quadratic ones. Now the general invariant above can be suggestively rewritten

$$\sum_I \left(v_{i_1 I_1} v_{i_2 I_2} \cdots v_{i_{p-1} I_{p-1}} \left(\sum_{I_p} v_{i_p I_p} v^{j_p I_p} \right) v^{j_{p-1} I_{p-1}} \cdots v^{j_1 I_1} \right),$$

so that if we can drag the quadratic term in the middle through to the right hand side to produce a sum of products of order 2 and order $2(p-1)$ invariants plus a sum of lower order invariants, then we are done. This boils down to the following lemma.

Lemma 14: There exist constants α^{jkv}_{iub} (which may be determined), such that

$$\left(\sum_I v_{iI} v^{jI} \right) v^{kK} = \sum_{u,v,b} \alpha^{jkv}_{iub} v^{uK} \left(\sum_I v_{vI} v^{bI} \right) - \sum_b \alpha^{jku}_{iub} v^{bK}.$$

Proof:

$$\left(\sum_I v_{iI} v^{jI} \right) v^{kK} = \sum_{I,a,b,A,B} v_{iI}((\check{r})^{-1}{}^{jk}_{ab} \check{R}^{IK}_{AB} v^{aA} v^{bB})$$

$$= \sum_{I,a,b,A,B} (\check{r})^{-1}{}^{jk}_{ab} \hat{R}^{IK}_{AB} \left(\sum_{u,v,U,V} (r^{t_2}_{21})^{-1}{}^{ai}_{uv} (R^{t_2}_{21})^{-1}{}^{AI}_{UV} (v^{uU} v_{vV} - \delta^u_v \delta^U_V) \right) v^{bB}$$

$$= \sum_{a,b,B,u,v} (\check{r})^{-1}{}^{jk}_{ab} (r^{t_2}_{21})^{-1}{}^{ai}_{uv} (v^{uK} v_{vB} v^{bB} - \delta^u_v) v^{bK},$$

where the first two lines come from the relations of the algebra, (4) and (5), and the third uses $(\check{R})^{IK}_{AB} = (R^{t_2}_{21})^{KB}_{AI}$. The result follows. ∎

For the last part, we note that from the defining relations and the action on the basis, one can prove that

$$\sum_I v_{1I}v^{1I}=\sum_I q^{2\Sigma_{J<I}N_{1J}}(N_{1I})_{q^{-1}}=\left(\sum_I N_{1I}\right)_{q^{-1}},$$

where N_{iI} is the number operator, $N_{iI}\mathbf{x}(\mathbf{s})=s_{iI}\mathbf{x}(\mathbf{s})$, and since $(\Sigma_I N_{1I})_{q^{-1}}$ is in the image of $\mathcal{U}_q\mathrm{u}(m,n)$, so is $\Sigma_I v_{1I}v^{1I}$. It now follows by simple induction that all quadratic invariants lie in the image of $\mathcal{U}_q\mathrm{u}(m,n)$ since we can construct them from $\Sigma_I v_{1I}v^{1I}$ using the adjoint action of $\mathcal{U}_q\mathrm{u}(m,n)$.

This completes the proof of the theorem. ■

To conclude the duality theory, we need a technical lemma which uses the Haar measure on the dual quantum group. For general reference see Ref. 17.

Define $(\mathcal{F}_q^1 G,*)$ to be the Hopf $*$-subalgebra of the Hopf dual of $\mathcal{U}_q\mathfrak{g}$ generated by the matrix elements of finite dimensional representations of type 1. (A finite dimensional representation is said to be of type 1 if the weight $\mathbf{w}$ has $w_i=q^{(\alpha_i,\lambda)}$ for $\lambda\in P^+$.) Let $\int$ denote the Haar measure of $(\mathcal{F}_q^1 G,*)$. Let the coaction of $(\mathcal{F}_q^1 G,*)$ on $\mathcal{W}_{AA}$ be denoted δ so that $\delta:\mathcal{W}_{AA}\to\mathcal{W}_{AA}\otimes\mathcal{F}_q^1 G$.

Lemma 15: Let $\mathcal{A}$ denote those operators in $\mathcal{W}_{AA}$ which commute with the action of $\mathcal{U}_q\mathrm{u}(M)$ on $S_{AA}(V)$.

There exists a unique projection $\pi:\mathcal{W}_{AA}\to\mathcal{A}$ of $\mathcal{W}_{AA}$ onto $\mathcal{A}$, which satisfies the conditional expectation property,

$$(AB^\pi)^\pi=A^\pi B^\pi,\quad \forall A,B\in\mathcal{W}_{AA},$$

and such that if U is a finite dimensional $\mathcal{U}_q\mathrm{u}(M)$-invariant subspace of Fock space and $A\in\mathcal{W}_{AA}$ leaves U invariant and commutes with $\mathcal{U}_q\mathrm{u}(M)$ on U, then

$$A^\pi|_U=A|_U.$$

Proof: Classically we would define $A^\pi=\int_G dg\ \mathrm{ad}(g)(A)$. Since we have quantum analogs of the adjoint action and the Haar measure we can make essentially the same definition.

Define

$$A^\pi=\left(1\otimes\int\right)\delta A,\quad A\in\mathcal{W}_{AA}.$$

It is clear that the projection is onto since if A is invariant, $A^\pi=A$,

$$A^\pi=\left(1\otimes\int\right)\delta A=\left(1\otimes\int\right)(A\otimes 1)=A.$$

We also have (suppressing the unit map)

$$\delta A^\pi=\delta\left(\left(1\otimes\int\right)\delta A\right)=\left(1\otimes 1\otimes\int\right)(\delta\otimes 1)\delta A=\left(1\otimes 1\otimes\int\right)(1\otimes\Delta)\delta A=\left(1\otimes\int\right)\delta A=A^\pi\otimes 1.$$

Therefore, A^π is invariant by definition and so commutes with the action of $\mathcal{U}_q\mathrm{u}(M)$ by Theorem 12.

We also have the conditional expectation,

$$(AB^\pi)^\pi=\left(1\otimes\int\right)\delta(AB^\pi)=\left(1\otimes\int\right)(\delta A)(B^\pi\otimes 1)=A^\pi B^\pi.$$

Finally, assume U is a finite dimensional $\mathcal{U}_q\mathrm{u}(M)$ and A-invariant subspace on which the two commute. For any $\mathrm{u}\in U$,

$$(\delta A)\mathrm{u}=A\mathrm{u}\otimes 1,$$

since this is just the dual of the statement that A is invariant under the $\mathcal{U}_q\mathrm{u}(M)$-action when restricted to U, using the nondegeneracy of the Hopf pairing. It follows that

$$A^{\pi}\mathrm{u}=\left[\left(1\otimes\int\right)(\delta A)\right]\mathrm{u}=\left(1\otimes\int\right)(A\mathrm{u}\otimes 1)=A\mathrm{u}.$$

■

Theorem 16: Let q be a nonroot of unity. The decomposition of the joint metaplectic action of $\mathcal{U}_q\mathrm{u}(M)$, $\mathcal{U}_q\mathrm{u}(m,n)$ *is of the form*

$$\mu\simeq\bigoplus_i \rho_i\otimes\rho_i',$$

where ρ_i *and* ρ_i' *are irreducible representations of* $\mathcal{U}_q\mathrm{u}(M)$ *and* $\mathcal{U}_q\mathrm{u}(m,n)$, *respectively. Furthermore, there is bijection between these representations such that* ρ_i *only occurs with* ρ_i' *and vice versa; this is the duality correspondence.*

Proof: Let P denote the polynomial subspace on which the dual pair act. With respect to the $\mathcal{U}_q\mathrm{u}(M)$ action there will be an isotypic decomposition

$$P\simeq\bigoplus_{i=0}^{\infty} I_i.$$

By irreducibility, $\mathcal{W}_{AA}$ is dense on the representation space, so the restriction to any finite dimensional subspace will be the full endomorphism algebra. The isotypic components I_i are a direct sum of $\mathcal{U}_q\mathrm{u}(M)$-invariant finite dimensional spaces. On any invariant finite dimensional subspace $U\subset I_i$, the full commutant of $\mathcal{U}_q\mathrm{u}(M)$ in End(U) will just be the restriction of $\mathcal{U}_q\mathrm{u}(m,n)$ to U by Lemma 15. Then the double commutant theorem says that U is an irreducible representation under the action of $\mathcal{U}_q\mathrm{u}(M)\otimes\mathcal{U}_q\mathrm{u}(m,n)$. Since U was arbitrarily large, however, this must also hold for I_i, and so $I_i=V_i\otimes V_i'$, where V_i and V_i' are the spaces on which ρ_i and ρ_i' act, where these are irreducible representations of $\mathcal{U}_q\mathrm{u}(M)$ and $\mathcal{U}_q\mathrm{u}(m,n)$, respectively. It also follows by the lemma that for any $i\neq j$ there exists an element $X\in\mathcal{U}_q\mathrm{u}(m,n)$ which is 0 on I_i but nonzero on I_j. Thus we obtain the duality correspondence. ■

As an immediate consequence of the duality, we see that certain aspects of the Clebsch–Gordon theory of $\mathcal{U}_q\mathrm{u}(m,n)$ can be addressed by looking at $\mathcal{U}_q\mathrm{u}(M)$. For example, we give the quantum analog of Theorem 4.6 of Ref. 8. Let ρ_1 and ρ_2 be irreducible representations of $\mathcal{U}_q\mathrm{u}(m,n)$ which arise in the decomposition of $\otimes^M\mu_{(m,n)}$ and $\otimes^N\mu_{(m,n)}$, respectively, where $\mu_{(m,n)}$ is the metaplectic representation of $\mathcal{U}_q\mathrm{u}(m,n)$. We wish to consider the decomposition of $\rho_1\otimes\rho_2$ into irreducibles.

Theorem 17: Let ρ_1,ρ_2 *be as above then the duality theory allows one to associate to them the representations* ρ_1' *and* ρ_2' *of* $\mathcal{U}_q\mathrm{u}(M)$ *and* $\mathcal{U}_q\mathrm{u}(N)$, *respectively. If* ρ_3 *is an irreducible representation of* $\mathcal{U}_q\mathrm{u}(m,n)$ *occurring in* $\rho_1\otimes\rho_2$, *then its multiplicity in that space is the same as the multiplicity of the irreducible* $\mathcal{U}_q\mathrm{u}(M)\otimes\mathcal{U}_q\mathrm{u}(N)$*-module occurring in* ρ_3', *where* ρ_3', *is the irreducible representation of* $\mathcal{U}_q\mathrm{u}(M+N)$ *associated to* ρ_3 *via duality.*

Proof: Note first that the representation space of $\otimes^M\mu_{(m,n)}$ is acted on by $\mathcal{U}_q\mathrm{u}(M)$ and that furthermore, $\mathcal{U}_q\mathrm{u}(m,n)$ and $\mathcal{U}_q\mathrm{u}(M)$ form a dual pair; similarly for $\mathcal{U}_q\mathrm{u}(N)$ and $\mathcal{U}_q\mathrm{u}(M+N)$.

The result follows by applying the duality theory for $\mathcal{U}_q\mathrm{u}(m,n)$ when paired with $\mathcal{U}_q\mathrm{u}(M)$, $\mathcal{U}_q\mathrm{u}(N)$, and $\mathcal{U}_q\mathrm{u}(M+N)$. ■

ACKNOWLEDGMENTS

Much of this paper is derived from the doctoral thesis, (Ref. 14). I am endebted to Keith Hannabuss for suggesting the problem and numerous enlightening discussions. The financial support of the EPSRC for the thesis is gratefully acknowledged.

[1] D. Shale, ''Linear symmetries of free boson fields,'' Trans. Am. Math. Soc. **103**, 149–167 (1962).
[2] G. B. Folland, *Harmonic Analysis in Phase Space*, Number 122 in *Annals of Math. Studies* (Princeton University Press, Princeton, 1989).
[3] D. Bowes and K. C. Hannabuss, ''Weyl quantization and star products,'' J. Geom. Phys. **22**, 319–348 (1997).
[4] M. Moshinsky and C. Quesne, ''Linear canonical transformations and their unitary representations,'' J. Math. Phys. **12**, 1772–1780 (1971).

[5] M. Kashiwara and M. Vergne, ''On the Segal–Shale–Weil representations and harmonic polynomials,'' Invent. Math. **44**, 1–47 (1978).
[6] H. P. Jakobsen, ''On singular holomorphic representations,'' Invent. Math. **62**, 67–78 (1980).
[7] R. Howe, ''Remarks on classical invariant theory,'' Trans. Am. Math. Soc. **313**, 539–570 (1989).
[8] R. Howe, ''Dual pairs in physics: Harmonic oscillators, photons, electrons, and singletons,'' Lecture Appl. Math. **21**, 179–207 (1985).
[9] R. Howe, ''Transcending classical invariant theory,'' J. Am. Math. Soc. **2**, 535–552 (1989).
[10] T. Ton-That, *Dual Representations and Invariant Theory*, in *Representation Theory and Harmonic Analysis* (American Mathematical Society, Providence, RI, 1994), No. 191.
[11] K. C. Hannabuss (private communication).
[12] C. Quesne, ''Complementarity of $su_q(3)$ and $u_q(2)$ and q-boson realization of the $su_q(3)$ irreducible representations,'' J. Phys. A **25**, 5977–5998 (1992).
[13] Y. F. Smirnov and V. N. Tolstoy, ''On Complementary Relations Between the Unitary Quantum Algebras,'' in Group Theory and Special Symmetries in Nuclear Physics, Proceedings of the International Symposium in Honour of K. T. Hecht, edited by J. P. Draayer and J. Janecke (World Scientific, Singapore, 1992).
[14] G. Green, ''On Howe duality for quantum groups,'' Ph.D. thesis, Oxford University, 1996.
[15] M. Jimbo, ''A q-analog of $u(gl(n+1))$, Hecke algebra and the Yang–Baxter equation,'' Lett. Math. Phys. **11**, 247–252 (1986).
[16] T. Hayashi, Quantum Deformation of Classical Groups (RIMS, Kyoto University, 1992), Vol. 28, pp. 57–81.
[17] V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge University Press, Cambridge, 1994).
[18] L. C. Biedenharn, ''The quantum group $su_q(2)$ and a q-analog of the boson operators,'' J. Phys. A **22**, L873–L877 (1989).
[19] A. J. Macfarlane, ''On q-analog of the quantum harmonic oscillator and the quantum group $su_q(2)$,'' J. Phys. A **22**, 4581–4588 (1989).
[20] T. Hayashi, ''q-analogues of Clifford and Weyl algebras: Spinor and oscillator representations of quantum enveloping algebras,'' Commun. Math. Phys. **127**, 129–144 (1990).
[21] J. Ding and I. B. Frenkel, *Spinor and Oscillator Representations of Quantum Groups*, in *Lie Theory and Geometry* (Birkhauser, Boston, 1995), No. 123, pp. 127–165.
[22] L. Faddeev, N. Reshetikhin, and L. Takhtadzhyan, ''Quantization of Lie groups and Lie algebras,'' Leningrad Math. J. **1**, 193–225 (1990).
[23] W. Pusz and S. L. Woronowicz, ''Twisted second quantization,'' Rep. Math. Phys. **27**, 231–257 (1989).
[24] G. Fiore, ''$so_q(n)$-isotropic harmonic oscillator on the quantum Euclidean space $\mathbb{R}_q^n$,'' q-alg, 9501023, 1994.
[25] S. Majid, *Foundations of Quantum Group Theory* (Cambridge University Press, Cambridge, 1995).
[26] C. Quesne, ''q-bosonic operators as double irreducible tensors for $u_q(n)+u_q(m)$,'' Phys. Lett. B **298.3**, 344–350 (1993).
[27] C. Quesne, ''R-matrix formulation of $su_q(n)\times su_q(m)$-covariant q-bosonic algebras,'' Phys. Lett. B **322**, 344–348 (1994).
[28] P. E. T. Jorgensen, L. M. Schmitt, and R. F. Werner, ''Positive representations of general commutation relations allowing Wick ordering,'' J. Funct. Anal. **134**, 33–99 (1995).
[29] A. Schirrmacher, *Remarks on the Use of R-Matrices*, in *Quantum Groups and Related Topics* (Kluwer Academic, Dordrecht, 1992), pp. 55–65.
[30] J. R. Links and M. D. Gould, ''Casimir invariants for Hopf algebras,'' Rep. Math. Phys. **31**, 91–111 (1992).

Singularities at the tip of a plane angular sector

Joseph B. Keller[a)]
Departments of Mathematics and Mechanical Engineering, Stanford University, Stanford, California 94305-2125

(Received 27 April 1998; accepted for publication 16 October 1998)

Solutions of the Helmholtz and Laplace equations in three dimensions which vanish, or have vanishing normal derivative on an angular sector of opening angle β, are considered. The solutions are required to be functions of distance from the tip of the sector multiplied by functions of the angular coordinates. The angular functions are eigenfunctions of the Laplace–Beltrami operator on the unit sphere, which vanish or have vanishing normal derivative, on a great circle arc of length β. It is shown that the Dirichlet eigenvalues are nondecreasing functions of β, and the Neumann eigenvalues are nonincreasing. Furthermore, each Dirichlet eigenvalue of a sector of angle β is a Neumann eigenvalue of a sector of angle $2\pi-\beta$ and conversely. The eigenvalues for $\beta=0$, π, and 2π are found explicitly. These results lead to a qualitative description of the eigenvalues as functions of β. The eigenvalues determine the singular behavior of the solutions at the tip. © *1999 American Institute of Physics.* [S0022-2488(99)02002-2]

I. INTRODUCTION

We consider two boundary value problems in R^3 for the Helmholtz equation:

$$(\Delta+k^2)u=0, \quad x\notin S_\beta, \quad u(x)=0, \quad x\in S_\beta, \tag{1.1D}$$

$$(\Delta+k^2)u=0, \quad x\notin S_{\beta_1}, \quad \partial_n u(x)=0, \quad x\in S_\beta. \tag{1.1N}$$

Here S_β is a plane angular sector of opening angle β, $0\leq\beta\leq 2\pi$, and k is a real constant. We call (1.1D) the Dirichlet problem and (1.1N) the Neumann problem, and we seek solutions of the product form

$$\begin{aligned} u(x)&=k^{-\nu}j_\nu(kr)U(\theta,\varphi), \quad k\neq 0,\\ u(x)&=r^\nu U(\theta,\varphi), \quad k=0. \end{aligned} \tag{1.2}$$

The spherical Bessel function $j_\nu(kr)$ tends to $(kr)^\nu$ as kr tends to zero, so both expressions in (1.2) behave like r^ν near $r=0$. Thus the exponent ν determines the behavior of u at the tip of the plane angular sector, so it is important in scattering from objects with such tips. Therefore, we shall study the values of ν and their dependence upon β.

Upon using (1.2) in (1.1) we find that ν is an eigenvalue and U is the corresponding eigenfunction of one of the following problems:

$$BU+\nu(\nu+1)U=0, \quad (\theta,\varphi)\notin C_\beta, \quad U=0, \quad (\theta,\varphi)\in C_\beta, \tag{1.3D}$$

$$BU+\nu(\nu+1)U=0, \quad (\theta,\varphi)\notin C_\beta, \quad \partial_n U=0, \quad (\theta,\varphi)\in C_\beta. \tag{1.3N}$$

In (1.3), B is the Laplace–Beltrami operator on the unit sphere and C_β is the great circle arc of length β in which the plane angular sector S_β intersects the unit sphere.

[a)]Electronic mail: keller@math.stanford.edu

To describe our results we write $\nu_j = D_j(\beta)$ and $\nu_j = N_j(\beta)$ for the jth Dirichlet and Neumann eigenvalue, respectively, $j = 1,2,...$. In Sec. II we shall prove the following theorems:

Theorem 1:

(a) $D_j(\beta)$ is a nondecreasing function of $\beta, j = 1,2,...$.
(b) $N_j(\beta)$ is a nonincreasing function of $\beta, j = 1,2,...$.

Theorem 2: Every Dirichlet eigenvalue $D_j(\beta)$ is equal to some Neumann eigenvalue $N_{j'}(2\pi - \beta)$, and conversely: $D_j(\beta) = N_{j'}(2\pi - \beta)$.

Note that Theorem 2 and either part of Theorem 1 implies the other part of Theorem 1. In Sec. III, Theorem 3, we give $D_j(\beta)$ and $N_j(\beta)$ for all j and $\beta = 0$, π, and 2π. In Theorem $2^\sharp$ of Sec. IV, we give the relation between the values of j and j' which appear in Theorem 2.

The preceding results enable us to determine the qualitative behavior of all the eigenvalues as functions of β, as we shall show in Figs. 1 and 2. Therefore they supplement the numerical results obtained for the first few eigenvalues by Kraus and Levine,[1] Blume and Kirchner,[2] Blume and Kahl,[3] De Smedt and Van Bladel,[4,5] Boersma,[6] and Abawi *et al.*[7] Kraus and Levine[1] reduced (1.3) to two problems for two Lamé ordinary differential equations by separation of variables in spheroconal coordinates. Boersma[6] proved the special case of Theorem 2 for $j = 1$ and $j' = 2$, and we make use of his idea to prove it in general. Abawi *et al.*[7] proved that for $0 < \beta < \pi$, $D_j(\beta)$ lies between an integer n and $n + 1/2$ when the eigenfunction is even and $N_j(\beta)$ lies between $m + 1/2$ and m when the eigenfunction is odd, where m is an integer. These results follow from our Theorems 1 and 3. They also derived large j asymptotic formulas for $D_j(\beta)$ and $N_j(\beta)$.

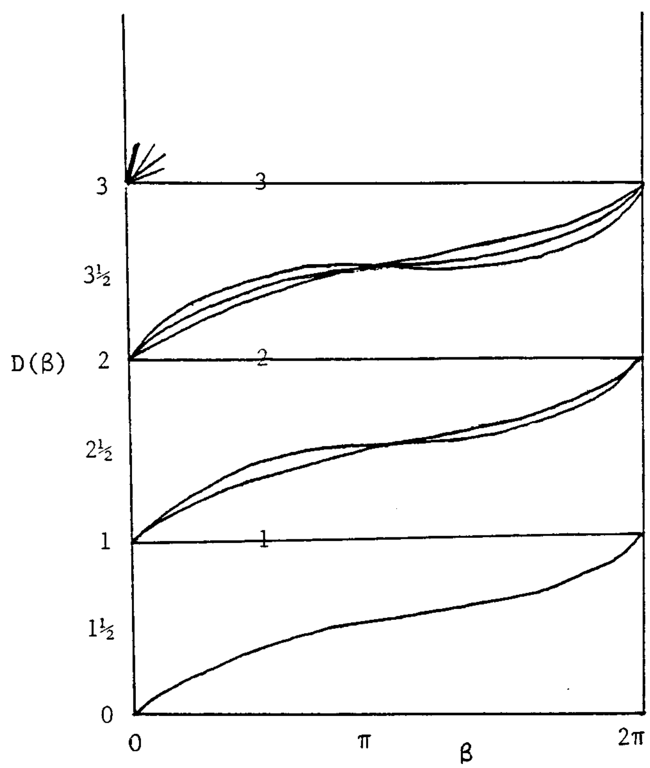

FIG. 1. Sketch of the Dirichlet eigenvalues $D_j(\beta)$ as functions of β. At $\beta = 0$, $D_j(0) = n$ with multiplicity $2n + 1$. Of them, n remain constant as β increases, and $n + 1$ increase. These $n + 1$ all equal $n + \frac{1}{2}$ at $\beta = \pi$ and all equal $n + 1$ at $\beta = 2\pi$.

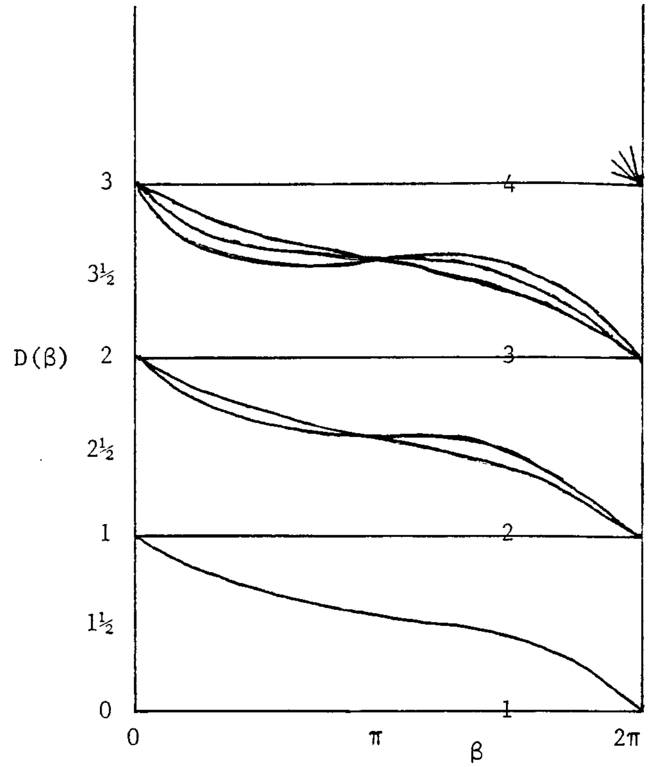

FIG. 2. Sketch of the Neumann eigenvalues $N_j(\beta)$ as functions of β. At $\beta=2\pi$, $N_j(2\pi)=n$ with multiplicity $2n+2$. Of them, $n+1$ remain constant and $n+1$ increase as β decreases. The increasing $n+1$ all equal $n+\frac{1}{2}$ at $\beta=\pi$ and all equal $n+1$ at $\beta=0$. There the multiplicity of $N_j(0)=n+1$ is $2(n+1)+1$. According to Theorem 2, Figs. 1 and 2 are mirror images of one another about the vertical line $\beta=\pi$. For each n, the numbers of constant eigenvalues differ by one.

II. PROOFS OF THEOREMS 1 AND 2

To prove Theorem 1 we recall that each eigenvalue is determined by the Courant minimax principle. For the Dirichlet problem, the admissible functions must vanish on the arc C_β. Since this arc increases with β, the class of admissible functions decreases as β increases. Therefore, the mimimax does not decrease as β increases, which proves part (a).

For the Neumann problem, the boundary condition is a natural condition, so it does not restrict the admissible functions. In fact, the admissible functions may be discontinuous across C_β, so as β increases, the class of admissible functions increases. Therefore, the minimax does not increase as β increases, which proves part (b), and completes the proof of Theorem 1.

To prove Theorem 2 we suppose that the plane angular sector lies in the plane $z=0$. The operator Δ in (1.1) is even in z. Therefore the even and odd parts of each solution u of (1.1) are also solutions of (1.1), so we can consider only even and odd solutions. Let $v_j^D(\beta)$ be an even eigenfunction of (1.1D) corresponding to the eigenvalue $D_j(\beta)$. Then the normal derivative of $v_j^D(\beta)$ vanishes in the complementary sector $S_{2\pi-\beta}$. Therefore the odd function $(\operatorname{sgn} z)v_j^D(\beta)$ has vanishing normal derivative on both sides of $S_{2\pi-\beta}$. Furthermore it is continuous across S_β because it vanishes there, and its normal derivative is continuous across S_β because $v_j^D(\beta)$ is even. Thus $(\operatorname{sgn} z)v_j^D(\beta)$ is a solution of (1.1N) with $S_{2\pi-\beta}$ instead of S_β and with the eigenvalue $\nu=D_j(\beta)$. Therefore $D_j(\beta)$ is also a Neumann eigenvalue of $S_{2\pi-\beta}$, say $N_{j'}(2\pi-\beta)$. Thus $D_j(\beta)=N_{j'}(2\pi-\beta)$. On the other hand, if $v_{j'}^N(2\pi-\beta)$ is an odd eigenfunction of (1.1N) with eigenvalue $N_{j'}(2\pi-\beta)$ then $(\operatorname{sgn} z)v_{j'}^N(2\pi-\beta)$ is a solution of (1.1D) with S_β instead of $S_{2\pi-\beta}$ and with eigenvalue $N_{j'}(2\pi-\beta)$. Therefore $N_{j'}(2\pi-\beta)$ is a Dirichlet eigenvalue of S_β, say $D_j(\beta)$, so $N_{j'}(2\pi-\beta)=D_j(\beta)$.

We have now shown that every Dirichlet eigenvalue $D_j(\beta)$ with an even eigenfunction is equal to some Neumann eigenvalue $N_j(2\pi-\beta)$ with an odd eigenfunction, and conversely. Next we consider a Dirichlet eigenvalue $D_j(\beta)$ with an odd eigenfunction. Such an eigenfunction must vanish on the entire plane $z=0$. Therefore it is a Dirichlet eigenfunction of S_β with eigenvalue $D_j(\beta)$ for every value of β in the interval $0\leqslant\beta\leqslant 2\pi$. In particular, $D_j(\beta)$ is a Dirichlet eigenvalue for $\beta=2\pi$. In Theorem 3, we shall show that every such eigenvalue is also a nonzero Neumann eigenvalue for $\beta=2\pi$, and also for every value of β in $0\leqslant\beta\leqslant 2\pi$, with the corresponding eigenfunction being even. In the same way we can show that every Neumann eigenvalue $N_{j'}(2\pi-\beta)$ with an even eigenfunction is also a Dirichlet eigenvalue $D_j(\beta)$ with an odd eigenfunction. This completes the proof of Theorem 2.

III. EIGENVALUES AND EIGENFUNCTIONS FOR $\beta=0,\pi,2\pi$

Now we shall determine the solutions of (1.3) explicitly by separation of variables in spherical polar coordinates for three special angles $\beta=0,\pi,2\pi$. For $\beta=0$, the boundary conditions are irrelevant, as can be shown by reformulating each problem in variational form. The consequence is that for both problems the eigenfunctions and eigenvalues are exactly those for the full sphere, which are just the spherical harmonics:

$$\begin{aligned} Y^e_{nm}(\theta,\varphi)&=P^m_n(\cos\theta)\cos m\varphi, \quad \nu=n, \quad m=0,1,...,n, \quad n=0,1,2,...,\\ Y^o_{nm}(\theta,\varphi)&=P^m_n(\cos\theta)\sin m\varphi, \quad \nu=n, \quad m=1,...,n, \quad n=1,2,... . \end{aligned} \tag{3.1}$$

Thus the multiplicity of the eigenvalue n is $2n+1$. To find the jth eigenvalues $D_j(0)$ and $N_j(0)$, we count the number of eigenvalues less than n with their multiplicities. This is just the sum of $2n'+1$ from $n'=0$ to $n'=n-1$, which is exactly n^2. Therefore we have

$$D_j(0)=N_j(0)=n, \quad j=n^2+m, \quad m=1,...,2n+1, \quad n=0,1,2,... . \tag{3.2}$$

To solve (1.3) for $\beta=2\pi$, we assume that the plane angular sector is the plane $z=0$ so that the great circle $C_{2\pi}$ is the circle $\theta=\pi/2$. The solutions of (1.3D) are those spherical harmonics (3.1) which vanish on $\theta=\pi/2$, and they are just the ones for which $P^m_n(\cos\theta)$ is an odd function of $\cos\theta$. Similarly the solutions of (1.3N) are those solutions (3.1) for which $P^m_n(\cos\theta)$ is an even function of $\cos\theta$. Since $P^m_n(-\cos\theta)=(-1)^{n+m}P^m_n(\cos\theta)$, the eigenfunction Y^e_{nm} or Y^o_{nm} is an odd or even function of $\cos\theta$ depending on whether $n+m$ is odd or even. It follows that n of the Y_{nm} are odd and $n+1$ are even. Therefore for (1.3D), $\nu=n$ with multiplicity n, while for (1.3N), $\nu=n$ with multiplicity $n+1$. Thus counting the eigenvalues less than n we find

$$D_j(2\pi)=n, \quad j=\frac{n^2-n}{2}+m, \quad m=1,2,...,n, \quad n=1,2,..., \tag{3.3}$$

$$N_j(2\pi)=n, \quad j=\frac{n^2+n}{2}+m, \quad m=1,...,n+1, \quad n=0,1,2,... . \tag{3.4}$$

When $\beta=\pi$ the plane angular sector is a half-plane that we choose to be the half-plane $\varphi=0$, which is also $\varphi=2\pi$. Then the Dirichlet eigenfunctions are spherical harmonics that vanish at $\varphi=0$ and $\varphi=2\pi$. They are

$$\left.\begin{aligned} \tilde{U}^o_{nm}(\theta,\varphi)&=P^{-m+1/2}_{n-1/2}(\cos\theta)\sin(m-\tfrac{1}{2})\varphi, \quad &\nu=n-\tfrac{1}{2}\\ U^o_{nm}(\theta,\varphi)&=P^m_n(\cos\theta)\sin m\varphi, \quad &\nu=n \end{aligned}\right\} \begin{aligned} &m=1,2,...,n,\\ &n=1,2,... . \end{aligned} \tag{3.5}$$

Thus each of the two eigenvalues $\nu=n$ and $\nu=n-\frac{1}{2}$ has multiplicity n for $n=1,2,...$. Upon counting the number of eigenvalues less than $n-\frac{1}{2}$ with their multiplicities, we get $n(n-1)$, while the number less than n is n^2. Thus we have

$$\left.\begin{array}{ll} D_j(\pi)=n-\tfrac{1}{2}, & j=n(n-1)+m \\ D_j(\pi)=n, & j=n^2+m, \end{array}\right\} \begin{array}{l} m=1,2,\ldots,n \\ n=1,2,\ldots . \end{array} \tag{3.6}$$

The Neumann eigenfunctions for $\beta=\pi$ are

$$\tilde{U}^e_{nm}(\theta,\varphi)=P^{-m+1/2}_{n-1/2}(\cos\theta)\cos(m-\tfrac{1}{2})\varphi, \quad \nu=n-\tfrac{1}{2}, \quad m=1,\ldots,n, \quad n=1,2,\ldots,$$

$$U^e_{nm}(\theta,\varphi)=P^m_n(\cos\theta)\cos m\varphi, \quad \nu=n, \quad m=0,1,\ldots,n, \quad n=0,1,\ldots . \tag{3.7}$$

For $n=0,1,2,\ldots$ there is an eigenvalue $\nu=n$ with multiplicity $n+1$, while for $n=1,2,\ldots$ there is an eigenvalue $\nu=n-\frac{1}{2}$ with multiplicity n. Counting leads to

$$N_j(\pi)=n-\tfrac{1}{2}, \quad j=n^2+m, \quad m=1,2,\ldots,n, \quad n=1,2,\ldots=n,$$

$$N_j(\pi)=n, \quad j=n(n+1)+m, \quad m=1,2,\ldots,n+1, \quad n=0,1,2,\ldots . \tag{3.8}$$

We shall summarize the preceding results as a theorem.

Theorem 3(a): The eigenvalues $D_j(0)$ and $N_j(0)$ are given by (3.2) and the corresponding eigenfunctions are given by (3.1).

(b): The eigenvalues $D_j(\pi)$ and $N_j(\pi)$ are given by (3.6) and (3.8), respectively, while the corresponding eigenfunctions are given by (3.5) and (3.7), respectively. The plane angular sector is the half-plane $\varphi=0$ and $\varphi=2\pi$.

(c): The eigenvalues $D_j(2\pi)$ and $N_j(2\pi)$ are given by (3.3) and (3.4), respectively, while the corresponding eigenfunctions are given by (3.1) with $n+m$ odd for $D_j(2\pi)$ and $n+m$ even for $N_j(2\pi)$. The plane angular sector is the plane $\theta=\pi/2$.

We note that the Dirichlet and Neumann eigenfunctions and eigenvalues for $\beta=2\pi$ are also eigenfunctions and eigenvalues for any value of β in $0\leqslant\beta\leqslant2\pi$. This is so because they vanish or have vanishing normal derivative on the whole plane $\theta=\pi/2$. Therefore they satisfy the differential equation and boundary conditions (1.3D) or (1.3N), respectively. Of course they are not all the eigenfunctions and eigenvalues. We shall state this as a corollary.

Corollary: Each eigenvalue $D_j(2\pi)=n$ with multiplicity n and $N_j(2\pi)=n$ with multiplicity $n+1$ is an eigenvalue with the same multiplicity, for every value of β in $0\leqslant\beta\leqslant2\pi$. It is not necessarily the jth one unless $\beta=2\pi$.

It is also possible to calculate the derivative with respect to β of each eigenvalue at $\beta=0,\pi$, and 2π. One way to do this is to use the Lamé ordinary differential equations obtained by Kraus and Levine,[1] which simplify at these values of β to the equations for Legendre functions and for trigonometric functions. Another way is to use the method of Ward and Keller,[8] which treats strong localized perturbations of eigenvalue problems, including perturbations of boundaries.

We have applied the latter method to calculate the derivative of the lowest Dirichlet eigenvalue at $\beta=\pi$, and we have obtained

$$D_1'(\pi)=\frac{1}{2\pi}. \tag{3.9}$$

This is just the slope of the straight line in the plane of β and $D_1(\beta)$ through the three points $[0,D_1(0)=0]$, $[\pi,D_1(\pi)=\frac{1}{2}]$, and $[2\pi,D_1(2\pi)=1]$.

We also find

$$D'_{m(n-1)+m}(\pi)\neq 0, \quad m=1$$

$$D'_{m(n-1)+m}(\pi)=0, \quad m=2,3,\ldots,n. \tag{3.10}$$

IV. CONCLUSION

Now we shall use our theorems and corollary to describe the overall behavior of the eigenvalues as functions of β. Let us begin with the Dirichlet eigenvalues $D_j(\beta)$ (See Fig. 1.) Theorem 3(a) shows via (3.2) that at $\beta=0$ each integer $n\geqslant 0$ is an eigenvalue with multiplicity $2n+1$. The corollary shows that of the $2n+1$ eigenvalues equal to n, n of them remain constant as β increases from 0 to 2π. The other $n+1$ are nondecreasing functions of β according to Theorem 1(a). At $\beta=\pi$ these $n+1$ eigenvalues have increased to the value $(n+1)-\frac{1}{2}=n+\frac{1}{2}$, as shown by Theorem 3(b) via (3.6) with n replaced by $n+1$. At $\beta=2\pi$ these $n+1$ eigenvalues have increased to the value $n+1$, as Theorem 3(c) shows, via (3.3) with n replaced by $n+1$.

Figure 2 displays the Neumann eigenvalues $N_j(\beta)$ as functions of β. Theorem 3(a) shows via (3.2) that at $\beta=0$ each integer $n\geqslant 0$ is an eigenvalue with multiplicity $2n+1$. The corollary shows that $n+1$ of them remain equal to n as β increases. Theorem 1(b) shows that the remaining n of them are nonincreasing functions of β. In Fig. 2, the $2n+1$ eigenvalues which equal $n+1$ at $\beta=0$ are shown. Of them, $n+2$ remain constant and $n+1$ decrease as β increases, reaching the value $n+\frac{1}{2}$ at $\beta=\pi$ and n at $\beta=2\pi$.

Theorem 2 implies that Figs. 1 and 2 are mirror images of one another about the vertical line $\beta=\pi$. The multiplicity of the constant eigenvalues, i.e., those independent of β, is greater by one in Fig. 2 than it is in Fig. 1. By using Figs. 1 and 2 we can determine the value of j' corresponding to j in Theorem 2. We can express the result as the following more precise form of Theorem 2:
Theorem 2#(a): For $n\geqslant 0$ and $s=1,2,...,n-1$,

$$D_{n^2+n+s}(\beta)=N_{(n+1)^2+s}(2\pi-\beta).$$

(b): For $n\geqslant 1$,

$$\begin{aligned} n&=D_{n^2+1}(\beta)=D_{n^2+2}(\beta)=\cdots=D_{n^2+n}(\beta)\\ &=N_{(n+1)^2-n}(2\pi-\beta)=\cdots=N_{(n+1)^2-1}(2\pi-\beta)\\ &=N_{(n+1)^2}(2\pi-\beta). \end{aligned}$$

ACKNOWLEDGMENT

This work was supported in part by the AFOSR.

[1] L. Kraus and L. Levine, ''Diffraction by an elliptic cone,'' Commun. Pure Appl. Math. **14**, 49–68 (1961).
[2] S. Blume and M. Kirchner, ''Das Beugungsnahfeld an der Spitze eines sektorförmigen Schlitzes in einem ebenen Schirm,'' Optik (Stuttgart) **29**, 185–203 (1969).
[3] S. Blume and G. Kahl, ''Field singularities at the tip of a cone with elliptic cross section,'' Optik (Stuttgart) **70**, 170–175 (1985).
[4] R. De Smedt and J. G. Van Bladel, ''Field singularities at the tip of a metallic cone of arbitrary cross section,'' IEEE Trans. Antennas Propag. **AP-34**, 865–870 (1986).
[5] R. De Smedt and J. G. Van Bladel, ''Field singularities near aperture corners,'' IEE Proc., Part A: Phys. Sci., Meas. Instrum., Manage. Educ. **134**, 694–698 (1987).
[6] J. Boersma, ''Singularity exponents for complementary sectors,'' Electron. Lett. **27**, 1484–1485 (1991).
[7] A. J. Abawi, R. F. Dashen, and H. Levine, ''The eigenvalues of the Laplacian on a sphere with boundary conditions specified on a segment of a great circle,'' J. Math. Phys. **38**, 1623–1649 (1997).
[8] M. J. Ward and J. B. Keller, ''Strong localized perturbations of eigenvalue problems,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **53**, 770–798 (1993).

Characteristic surface data for the eikonal equation

Ezra T. Newman and Alejandro Perez
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 14 September 1998; accepted for publication 24 September 1998)

A method of solving the eikonal equation, in either flat or curved space–times, with arbitrary Cauchy data, is extended to the case of data given on a characteristic surface. We find a beautiful relationship between the Cauchy and characteristic data for the same solution, namely they are related by a Legendre transformation. From the resulting solutions, we study and describe the wave-front singularities that are associated with their level surfaces (the characteristic surfaces or ''big wave fronts''). © *1999 American Institute of Physics.* [S0022-2488(99)03301-0]

I. INTRODUCTION

The high frequency limit of the wave equation is given by the eikonal equation, written in an arbitrary space–time as

$$g^{ab}(x^a)\partial_a S(x^a)\partial_b S(x^a)=0, \tag{1}$$

where the $x^a=(x^i,t)$ are any local coordinates, and $g^{ab}(x^a)$ is the metric of the given space–time. The level surfaces of a solution of Eq. (1), $S(x^a)=\text{const}$ (which need not be smooth every place and could have self-intersections), are three-dimensional characteristic surfaces (the ''big wave fronts'' in the terminology of Arnold[1]), and the sections $t=\text{constant}$ of these surfaces are the two-dimensional (''small'') wave fronts. The vector field $l^a=g^{ab}\partial_b S$ is tangent to the null geodesic that generate the characteristic surfaces.

In flat space–time the eikonal equation becomes

$$\eta^{ab}\partial_a S\partial_b S=(\partial_t S)^2-(\partial_x S)^2-(\partial_y S)^2-(\partial_z S)^2=0. \tag{2}$$

In Sec. II we review the method[2] to give a general solution of the eikonal equation in flat space–time adapted to appropriate Cauchy data given at $t=t_0$. In Sec. III we modify the method so that the eikonal equation is solved with arbitrary *characteristic* data given at null infinity. The relation between both methods is studied in Sec. IV where we find that the Cauchy and characteristic data are related by a Legendre transformation. The wave-front singularities of the level surfaces of the resulting solutions are described parametrically in Sec. V and finally the generalization of our results to asymptotically flat spaces–times is given in Sec. VI.

II. SOLUTIONS OF THE EIKONAL EQUATION

From the point of view of the theory of partial differential equation, the eikonal equation is a homogeneous first-order nonlinear partial differential equation; there exists a solution S^*, called the ''complete integral,'' depending on three arbitrary constants,[3] e.g., in flat space–time the function

$$S^*(x^i,t,\alpha_i)=x^i\alpha_i-t\sum\ (\alpha_i)^2 \tag{3}$$

is easily seen to satisfy (2).

Remark 1: The fact that the equation is homogeneous plays no role in this section, but will be crucial in the generalization to characteristic data.

Remark 2: From now on we will treat the problem of the eikonal in flat space; we leave the discussion of general space–times to the end.

It is possible to generate a ''general integral,'' i.e., a solution of the eikonal equation depending on an arbitrary function, by means of the following procedure: First, define the function $S^{**}(x^i,t,\alpha_i)$ of the coordinates and the free parameters α_i as

$$S^{**}(x^i,t,\alpha_i)\equiv S^*(x^i,t,\alpha_i)-H(\alpha_i), \tag{4}$$

where $H(\alpha_i)$ is any function of the α_i's.

Next, think of the α_i's as functions of the space–time points obtained from the following conditions:

$$\partial S^{**}/\partial\alpha_i=\partial S^*/\partial\alpha_i-\partial H/\partial\alpha_i=0. \tag{5}$$

Equation (5) determine the space–time dependence of the α_i's $[\alpha_i=A_i(x^i,t)]$ given that

$$\left|\frac{\partial^2 S^{**}(x^i,t,\alpha_i)}{\partial\alpha_i\partial\alpha_j}\right|\neq 0. \tag{6}$$

This condition can fail in lower dimensional regions called the caustics. This issue will be returned to in Sec. V.

Substituting $\alpha_i=A_i(x^i,t)$ into Eq. (4) we eliminate the α_i, and obtain

$$S^{**}(x^i,t)=S^*(x^i,t,A_i(x^i,t))-H(A_i(x^i,t)). \tag{7}$$

It is easy to verify that, because of the condition (5),

$$\partial_a S^{**}=\partial_a S^*, \tag{8}$$

which means that $S^{**}(x^i,t)$ is a new solution of the eikonal equation (1) determined by an arbitrary function H. We can determine the free function H so that the solution (7) satisfies initial Cauchy data at $t=t_0$. We denote the Cauchy data by $S_{\text{Cauchy}}(x^i)$. Conditions (5) imply that at $t=t_0$, $\alpha_i=\partial S_{\text{Cauchy}}/\partial x^i$. Inverting these relations we obtain $x^i=X^i(\alpha_i)$, and replacing them in (7) at $t=t_0$ we find the sought for relation:

$$H(\alpha_i)=S^*(X^i(\alpha_i),t_0,\alpha_i)-S_{\text{Cauchy}}(X^i(\alpha_i)). \tag{9}$$

This last equation relates the arbitrary function $H(\alpha_i)$ with the Cauchy data, $S_{\text{Cauchy}}(x^i)$, at $t=t_0$ and allows us to construct solutions of the eikonal equation in flat space–time for any initial data.

We now change the set of the α_i's for new parameters that are more appropriate to the study of asymptotically flat spaces and we rewrite our previous equations in terms of them. A complete integral of the eikonal equation, Eq. (1), can be written in terms of new parameters $(\beta,\zeta,\bar{\zeta})$ as

$$S^*(x^a,\beta,\zeta,\bar{\zeta})=\beta x^a l_a(\zeta,\bar{\zeta}), \tag{10}$$

where

$$l_a(\zeta,\bar{\zeta})=\frac{1}{\sqrt{2}(1+\zeta\bar{\zeta})}((1+\zeta\bar{\zeta}),(\zeta+\bar{\zeta}),-i(\zeta-\bar{\zeta}),(\zeta\bar{\zeta}-1)) \tag{11}$$

is the null covector pointing in the $(\zeta,\bar{\zeta})$ direction. The $(\zeta,\bar{\zeta})$ are the stereographic coordinates that parametrize the sphere of null directions.

From (10) and (11), we get the relations between the new parameters $(\beta,\zeta,\bar{\zeta})$, and the old α's:

$$\alpha_1 = \frac{\beta}{\sqrt{2}} \frac{\zeta + \bar{\zeta}}{1 + \zeta\bar{\zeta}}, \quad \alpha_2 = -i \frac{\beta}{\sqrt{2}} \frac{\zeta - \bar{\zeta}}{1 + \zeta\bar{\zeta}}, \quad \alpha_3 = \frac{\beta}{\sqrt{2}} \frac{\zeta\bar{\zeta} - 1}{1 + \zeta\bar{\zeta}^2}, \tag{12}$$

and

$$\beta = \sqrt{2 \sum \alpha_i^2}. \tag{13}$$

In terms of the new parameters Eq. (4) reads:

$$S^{**}(x^a, \beta, \zeta, \bar{\zeta}) = \beta x^a l_a(\zeta, \bar{\zeta}) - H(\beta, \zeta, \bar{\zeta}), \tag{14}$$

where $H(\beta,\zeta,\bar{\zeta})$ is an arbitrary function that will be determined by the initial conditions. Conditions (5) on $(\beta,\zeta,\bar{\zeta})$ become

$$x^a l_a(\zeta, \bar{\zeta}) - \frac{\partial H}{\partial \beta}(\beta, \zeta, \bar{\zeta})|_{\zeta,\bar{\zeta}} = 0, \tag{15}$$

$$\beta x^a m_a(\zeta, \bar{\zeta}) - \eth H(\beta, \zeta, \bar{\zeta})|_{\beta} = 0, \tag{16}$$

$$\beta x^a \bar{m}_a(\zeta, \bar{\zeta}) - \bar{\eth} H(\beta, \zeta, \bar{\zeta})|_{\beta} = 0. \tag{17}$$

Remark 3: We have replaced the derivatives with respect to ζ and $\bar{\zeta}$, respectively, by

$$\eth = (1 + \zeta\bar{\zeta}) \frac{\partial}{\partial \bar{\zeta}}, \quad \bar{\eth} = (1 + \zeta\bar{\zeta}) \frac{\partial}{\partial \zeta},$$

and used the fact that $\eth l_a(\zeta,\bar{\zeta}) = m_a$ and $\eth\bar{\eth} l_a = n_a - l_a$ where $(l_a, n_a, m_a, \bar{m}_a)$ form a null Minkowski space tetrad for each $(\zeta,\bar{\zeta})$.

The function $H(\beta,\zeta,\bar{\zeta})$ can be determined by means of the same procedure using the conditions $\alpha_i = \partial S_{\text{Cauchy}}/\partial x^i$ at $t = t_0$ and relations (12) to obtain the $x^i = X^i(\beta,\zeta,\bar{\zeta})$, and finally rewriting (9)

$$H(\beta,\zeta,\bar{\zeta}) = S^*(X^i(\beta,\zeta,\bar{\zeta}), t_0, \beta, \zeta, \bar{\zeta}) - S_{\text{Cauchy}}(X^i(\beta,\zeta,\bar{\zeta})).$$

III. CHARACTERISTIC DATA FOR THE EIKONAL EQUATION

The eikonal equation, being hyperbolic, admits a characteristic formulation. Even though the results of this section can be applied to any characteristic hypersurface in Minkowski, in flat space (as in any asymptotically simple space–time) there are two preferred characteristic surfaces, namely future and past null infinity, respectively. In the following we will formulate the characteristic problem in terms of data given at future null infinity, $\mathfrak{I}^+$. $\mathfrak{I}^+$ has the topology of $S^2 \times R$; we choose Bondi coordinates on it, namely $(\zeta,\bar{\zeta})$ on the S^2 and the retarded time u_B along R. In an analogous manner as for the Cauchy problem, the characteristic data at future null infinity will be defined by a function of $(u_B,\zeta,\bar{\zeta})$:

$$S_{\text{characteristic}} = L(u_B, \zeta, \bar{\zeta}). \tag{18}$$

The goal of this section is to develop a method to construct solutions of the eikonal equation geometrically adapted to the characteristic data, Eq. (18), at $\mathfrak{I}^+$.

Remark 4: In asymptotically flat space–times, in the neighborhood of future null infinity, $\mathfrak{I}^+$, there is a preferred class of coordinates referred to as Bondi coordinates. Given a Bondi system

$(u_B,\zeta,\bar\zeta)$ *at* $\mathfrak{I}^+$, *a new system*[4] $(u,\zeta,\bar\zeta)$ *can be defined by* $(u,\zeta,\bar\zeta)=(L(u_B,\zeta,\bar\zeta),\zeta,\bar\zeta)$. *The characteristic data, (18), can be thought of as being generated by this coordinate change representing a one parameter family of arbitrary* $u=\mathrm{const}$ *slices of* $\mathfrak{I}^+$.

In flat space–times we can define a two parameter family of null surfaces by

$$S^*=x^a l_a(\zeta,\bar\zeta). \tag{19}$$

As was pointed out in Ref. 2, Eq. (19) has a dual interpretation. For $(\zeta,\bar\zeta)$ kept constant, its level surfaces define null planes intersecting the time axis at a time equal the value of S^* and with its direction given by $(\zeta,\bar\zeta)$; on the other hand, for a fixed value of x^a it represents the light cone cut at $\mathfrak{I}^+$ of the space–time point x^a in the interior, i.e., it represents the intersection of the null cone from x^a with $\mathfrak{I}^+$.

We can think of the characteristic data (18) geometrically, as defining a one parameter family of cuts at $\mathfrak{I}^+$ in terms of $u=L(u_B,\zeta,\bar\zeta)=\mathrm{const}$. [It is assumed that this can be inverted so that the cuts are given by $u_B=L^{-1}(u=\mathrm{const},\zeta,\bar\zeta)$.] With this point of view, we construct a solution of the eikonal equation (1) [corresponding to the characteristic data $L(u_B,\zeta,\bar\zeta)$], such that the family of null surfaces in the interior are defined by the null geodesics normal to the family of cuts at infinity given by $L(u_B,\zeta,\bar\zeta)=\mathrm{const}$. In order to do so we will generalize the method of Sec. II.

Defining

$$S^{**}(x^a,\zeta,\bar\zeta)=L(S^*,\zeta,\bar\zeta)=L(x^a l_a(\zeta,\bar\zeta),\zeta,\bar\zeta), \tag{20}$$

we see immediately that it is a solution of the eikonal equation depending on two free parameters. [Note the duplication of notation which arises from the different meanings to the same object; $S^*=u_B=x^a l_a(\zeta,\bar\zeta)$.] By putting the requirement on ζ and $\bar\zeta$ that $\eth S^{**}=\bar\eth S^{**}=0$, i.e.,

$$\begin{aligned}\dot L(S^*,\zeta,\bar\zeta)x^a m_a(\zeta,\bar\zeta)+\eth L(S^*,\zeta,\bar\zeta)=0,\\ \dot L(S^*,\zeta,\bar\zeta)x^a \bar m_a(\zeta,\bar\zeta)+\bar\eth L(S^*,\zeta,\bar\zeta)=0,\end{aligned} \tag{21}$$

where $\dot L\equiv\partial_{S'}L$. we can solve for ζ and $\bar\zeta$ in terms of x^a, i.e., Eq. (21) gives us

$$\zeta=\Gamma(x^a) \tag{22}$$

and

$$\bar\zeta=\bar\Gamma(x^a), \tag{23}$$

except at the caustics when[5]

$$\begin{vmatrix}\eth^2 S^{**} & \eth\bar\eth S^{**}\\ \eth\bar\eth S^{**} & \bar\eth^2 S^{**}\end{vmatrix}=0. \tag{24}$$

This issue will be discussed in Sec. V.

Finally replacing $(\zeta,\bar\zeta)$ in (20) and differentiating we find

$$\partial_a S^{**}(x^a,\Gamma(x^a),\bar\Gamma(x^a))=\dot L l_a(\Gamma(x^a),\bar\Gamma(x^a)). \tag{25}$$

Therefore, the function

$$S^{**}(x^a,\Gamma(x^a),\bar\Gamma(x^a)) \tag{26}$$

satisfies the eikonal equation, and by construction (20) it is adapted to the characteristic data defined by the function $L(u_B,\zeta,\bar\zeta)$ at $\mathscr{I}^+$. The null normals to level surfaces

$$S^{**}(x^a,\Gamma(x^a),\overline{\Gamma}(x^a))=\text{const} \tag{27}$$

are normal to the cuts $L(u_B,\zeta,\overline{\zeta})=\text{const.}$ at $\mathscr{I}^+$. Note that the fact that $S^{**}(x^a)$ is a new solution of the eikonal equation is a consequence of property of the eikonal equation of being homogeneous in $\partial_a S$.

IV. RELATION BETWEEN THE CAUCHY AND CHARACTERISTIC CONSTRUCTIONS

In this section we give the connection between the two methods of construction. Earlier we showed how to relate the Cauchy data, $S_{\text{Cauchy}}(x^t)$, with the arbitrary function $H(\beta,\zeta,\overline{\zeta})$ of Sec. II, so that any solution of the eikonal equation can be cast in the form of Eq. (14). Therefore, there must be a relationship of the characteristic construction to the construction via Cauchy data and hence a relationship between the functions $L(x^a l_a(\zeta,\overline{\zeta}),\zeta,\overline{\zeta})$ and $H(\beta,\zeta,\overline{\zeta})$.

We first note that though in both methods there is an arbitrary function of three variables, in the characteristic method there appear only two parameters $(\zeta,\overline{\zeta})$ while in the Cauchy method there are the three $(\beta,\zeta,\overline{\zeta})$.

We can reduce the three to two by solving Eq. (15),

$$x^a l_a(\zeta,\overline{\zeta})-\frac{\partial H}{\partial\beta}(\beta,\zeta,\overline{\zeta})=0, \tag{28}$$

for $\beta=\beta(x^a l_a,\zeta,\overline{\zeta})$ or changing notation and using $u_B=x^a l_a$, we have $\beta=\beta(u_B,\zeta,\overline{\zeta})$. Now thinking of (28) as an implicit relation defining either $u_B=U(\beta,\zeta,\overline{\zeta})\equiv(\partial H/\partial\beta)(\beta,\zeta,\overline{\zeta})$ or $\beta=\beta(u_B,\zeta,\overline{\zeta})$. Note that if we treat H as a function of $(u_B,\zeta,\overline{\zeta})$, i.e., $H=H(\beta(u_B,\zeta,\overline{\zeta}),\zeta,\overline{\zeta})$ then

$$\dot{H}\equiv\partial_{u_B}H|_{\zeta,\overline{\zeta}}=\frac{\partial H}{\partial\beta}(\beta,\zeta,\overline{\zeta})\dot{\beta} \tag{29}$$

or

$$\dot{\beta}=\frac{\dot{H}}{\partial H/\partial\beta}. \tag{30}$$

We replace $\beta=\beta(u_B,\zeta,\overline{\zeta})$ into the two conditions, Eqs. (16) and (17), obtaining

$$\begin{aligned}\beta(u_B,\zeta,\overline{\zeta})x^a m_a(\zeta,\overline{\zeta})-\eth H(\beta(u_B,\zeta,\overline{\zeta}),\zeta,\overline{\zeta})|_\beta=0,\\ \beta(u_B,\zeta,\overline{\zeta})x^a \overline{m}_a(\zeta,\overline{\zeta})-\overline{\eth} H(\beta(u_B,\zeta,\overline{\zeta}),\zeta,\overline{\zeta})|_\beta=0,\end{aligned} \tag{31}$$

which appear similar to Eq. (21), namely:

$$\begin{aligned}\dot{L}(u_B,\zeta,\overline{\zeta})x^a m_a(\zeta,\overline{\zeta})+\eth L(u_B,\zeta,\overline{\zeta})|_{u_B}=0,\\ \dot{L}(u_B,\zeta,\overline{\zeta})x^a \overline{m}_a(\zeta,\overline{\zeta})+\overline{\eth} L(u_B,\zeta,\overline{\zeta})|_{u_B}=0.\end{aligned} \tag{32}$$

We explicitly write $|_\beta$ and $|_{u_B}$ in the ð operators to mean that the angular derivatives are taken keeping β or u_B constant, respectively; also $\dot{L}$ means $\partial_{u_B}L|_{\zeta,\overline{\zeta}}$ for any $L(u_B,\zeta,\overline{\zeta})$.

Remark 5: As we mentioned earlier, Eqs. (31) *or* (32) *implicitly define* $\zeta=\Gamma(x^a)$ *and* $\overline{\zeta}=\overline{\Gamma}(x^a)$ *everywhere except at the caustics. They can be approached in a limiting fashion.*

Given an arbitrary function $F(\beta,\zeta,\overline{\zeta})$ and $\beta(u_B,\zeta,\overline{\zeta})$ there is the following relation between differential operators,

$$\eth F(\beta,\zeta,\bar{\zeta})|_{\beta}=\eth F(\beta,\zeta,\bar{\zeta})|_{u_B}-(\partial F/\partial\beta)(\beta,\zeta,\bar{\zeta})\eth\beta|_{u_B},$$
$$\bar{\eth} F(\beta,\zeta,\bar{\zeta})|_{\beta}=\bar{\eth} F(\beta,\zeta,\bar{\zeta})|_{u_B}-(\partial F/\partial\beta)(\beta,\zeta,\bar{\zeta})\bar{\eth}\beta|_{u_B}. \tag{33}$$

Using these relations to replace the $\eth$ and $\bar{\eth}$ derivative operators at β constant by operators at u_B constant in (32) we obtain:

$$\beta x^a m_a(\zeta,\bar{\zeta})-\eth H(\beta,\zeta,\bar{\zeta})|_{u_B}+(\partial H/\partial\beta)(\beta,\zeta,\bar{\zeta})\eth\beta|_{u_B}=0,$$
$$\beta x^a \bar{m}_a(\zeta,\bar{\zeta})-\bar{\eth} H(\beta,\zeta,\bar{\zeta})|_{u_B}+(\partial H/\partial\beta)(\beta,\zeta,\bar{\zeta})\bar{\eth}\beta|_{u_B}=0. \tag{34}$$

Applying relations (33) to the function $F=u_B=x^a l_a(\zeta,\bar{\zeta})$, thought of as $u_B=U(\beta,\zeta,\bar{\zeta})$, via the following steps:

$$\eth F(\beta,\zeta,\bar{\zeta})|_{\beta}=\eth(x^a l_a)=x^a m_a\,, \tag{35}$$

$$\eth F(\beta,\zeta,\bar{\zeta})|_{u}=\eth u|_{u}=0, \tag{36}$$

$$(\partial F/\partial\beta)\eth\beta|_{u_B}=(\partial u/\partial\beta)\eth\beta|_{u_B}=\dot{\beta}^{-1}\eth\beta|_{u_B}, \tag{37}$$

we get the following important equation:

$$\dot{\beta}x^a m_a=-\eth\beta|_{u_B}. \tag{38}$$

Finally inserting this relation, with Eq. (30), into (34) we obtain

$$\beta x^a m_a(\zeta,\bar{\zeta})-\eth H(\beta,\zeta,\bar{\zeta})|_{u_B}-\dot{H}x^a m_a(\zeta,\bar{\zeta})=0,$$
$$\beta x^a \bar{m}_a(\zeta,\bar{\zeta})-\bar{\eth} H(\beta,\zeta,\bar{\zeta})|_{u_B}-\dot{H}x^a \bar{m}_a(\zeta,\bar{\zeta})=0, \tag{39}$$

which can be rewritten as

$$\eth(u_B\beta-H)|_{u_B}+\frac{\partial}{\partial u_B}(u_B\beta-H)x^a m_a(\zeta,\bar{\zeta})=0,$$
$$\bar{\eth}(u_B\beta-H)|_{u_B}+\frac{\partial}{\partial u_B}(u_B\beta-H)x^a \bar{m}_a(\zeta,\bar{\zeta})=0. \tag{40}$$

Comparing Eq. (40) with (32) we see that they are identical when we set

$$L(u_B,\zeta,\bar{\zeta})=u_B\beta(u_B,\zeta,\bar{\zeta})-H(\beta(u_B,\zeta,\bar{\zeta}),\zeta,\bar{\zeta}). \tag{41}$$

From Eq. (28) we also have that

$$u_B=\partial H/\partial\beta, \quad \beta=\partial L/\partial u_B\,. \tag{42}$$

We see that the two data functions $L(u_B,\zeta,\bar{\zeta})$ and $H(\beta,\zeta,\bar{\zeta})$ are related by the Legendre transformation, Eqs. (41) and (42). We have finally arrived at a very simple and beautiful relation between the two methods. An essential property of the eikonal equation for this relationship is that it is homogeneous in $\partial_a S$.

V. PARAMETRIC DESCRIPTION OF THE WAVE FRONTS

Using the methods described above we can construct a general solution of the eikonal equation either for the Cauchy or the (corresponding) characteristic data. In Sec. IV we showed that they are simply related by a Legendre transformation in the variables u_B and β. Once we have the solution of the eikonal equation we can study the geometry of its wave fronts. In particular we are interested in the description of the singularities developed by them, namely, its caustics.

A key step in both methods consists of expressing the free parameters, e.g., $(\beta,\zeta,\bar{\zeta})$, contained in the formalism as functions of the space–time points, x^a. In many cases the problem of inverting Eqs. (15)–(17) or Eq. (21) in order to get either $(\zeta,\bar{\zeta})$ or $(\beta,\zeta,\bar{\zeta})$ as functions of the space–time coordinates x^a, can be a formidable task and at times impossible. However, it is not absolutely necessary, since it is possible to give a parametric description of the null surfaces defined by Eq. (14) or Eq. (26), respectively. (We follow the path given in Ref. 2 for the case of the stationary eikonal equation).

In the Cauchy case we have three parameters $(\zeta,\bar{\zeta},\beta)$ in the initial data for the eikonal equation. See Eq. (14). We introduce the new parameter r together with $(\zeta,\bar{\zeta},\beta)$ by means of the following equation:

$$r=\beta^{-1}\eth\bar{\eth}S^{**}=x^a(n_a-l_a)-\beta^{-1}\eth\bar{\eth}H(\beta,\zeta,\bar{\zeta})|_{\beta} \tag{43}$$

and the previous equations;

$$x^a l_a(\zeta,\bar{\zeta})-\frac{\partial H}{\partial\beta}(\beta,\zeta,\bar{\zeta})|_{\zeta,\bar{\zeta}}=0,$$

$$\beta x^a m_a(\zeta,\bar{\zeta})-\eth H(\beta,\zeta,\bar{\zeta})=0,\quad \beta x^a \bar{m}_a(\zeta,\bar{\zeta})-\bar{\eth}H(\beta,\zeta,\bar{\zeta})=0. \tag{44}$$

The four equations (44) and (43) can be solved for the coordinates x^a in terms of $(\beta,r,\zeta,\bar{\zeta})$, using the orthonormality of the null tetrad:

$$x^a=\frac{\partial H}{\partial\beta}(l^a+n^a)-\left(r-\frac{\eth\bar{\eth}H}{\beta}\right)l^a-\frac{\bar{\eth}H}{\beta}m^a-\frac{\eth H}{\beta}\bar{m}^a. \tag{45}$$

Equation (45) is not very convenient for the analysis of the wave fronts because the parameter β does not have a simple geometric meaning related with the null surfaces. On the other hand, as we know, the level surfaces of $S^{**}=u=\text{const}$ in Eq. (14) define the null surfaces in which we are interested. Therefore, a sensible parametrization will be the one that replaces the β with the parameter u defined by

$$u=\beta x^a l_a-H(\beta,\zeta,\bar{\zeta})=L(x^a l_a,\zeta,\bar{\zeta}). \tag{46}$$

Constant values of u label the characteristic surfaces themselves and are different than $u_B=x^a l_a$. By changing the parameter β in favor of u we are switching to the characteristic description which provides a better framework to study the dynamics of the wave fronts.

Remark 6: Note that $r\equiv\beta^{-1}\eth\bar{\eth}S^{**}=\beta^{-1}\eth\bar{\eth}u$ *defines an affine parameter along the null geodesics that rule the characteristic surfaces* $u=\text{const}$.

Instead of performing the transformation from the "Cauchy parametrization" to the new set $(u,r,\zeta,\bar{\zeta})$ we take a shortcut, and start directly with the characteristic approach. Using the notation of the previous sections for the characteristic problem the new parameters are determined by the previous equations:

$$u=u(x^a)=L(x^a l_a,\zeta,\bar{\zeta}), \tag{47}$$

$$\begin{aligned}&\dot{L}(x^a l_a,\zeta,\bar{\zeta})x^a m_a(\zeta,\bar{\zeta})+\eth L(x^a l_a,\zeta,\bar{\zeta})=0\\&\dot{L}(x^a l_a,\zeta,\bar{\zeta})x^a \bar{m}_a(\zeta,\bar{\zeta})+\bar{\eth}L(x^a l_a,\zeta,\bar{\zeta})=0,\end{aligned} \tag{48}$$

and the new one defined by $r=\dot{L}^{-1}\eth\bar{\eth}S^{**}$ yielding

$$r=x^a(n_a-l_a)+\frac{\bar{\eth}\dot{L}}{\dot{L}}\,x^a m_a+\frac{\eth\dot{L}}{\dot{L}}\,x^a\bar{m}_a+\frac{\ddot{L}}{\dot{L}}\,x^a x^b m_a\bar{m}_b+\frac{\eth\bar{\eth}L}{\dot{L}}. \tag{49}$$

The coordinates x^a can be written in terms of the four parameters u, r, ζ, and $\bar{\zeta}$ as

$$x^a=u_B(l^a+n^a)+(r+\bar{\eth}\Phi+\bar{\Phi}\dot{\Phi})l^a-\bar{\Phi}m^a-\Phi\bar{m}^a, \tag{50}$$

where

$$\Phi\equiv-\eth L/\dot{L}, \tag{51}$$

and the function $u_B\equiv x^a l_a$ is written in terms of the parameters u, ζ, and $\bar{\zeta}$ implicitly by $u=L(x^a l_a,\zeta,\bar{\zeta})$, i.e., $x^a l_a=L^{-1}(u,\zeta,\bar{\zeta})$. ($L^{-1}$ denotes the inverse *function* of L.)

Treating Eq. (50) as a coordinate transformation between the natural coordinates associated with the solution, i.e., the $(u,\zeta,\bar{\zeta},r)$, and the standard space–time coordinates x^a, the transformation breaks down when its Jacobian vanishes. This is a three surface in the space–time; the caustic set associated with the solution.

After a lengthy calculation we find that this occurs when

$$J=\frac{\partial(t,x,y,z)}{\partial(u,r,\zeta,\bar{\zeta})}=r^2-\sigma^0\bar{\sigma}^0=0, \tag{52}$$

where

$$\sigma^0=\eth\Phi+\Phi\dot{\Phi}. \tag{53}$$

This is equivalent to Eq. (3.1) of Ref. 4.

There is a simple geometric interpretation of Eqs. (52) and (53); the shear function σ of the congruence of null geodesics that generate the surfaces $u=\text{constant}$, with the affine parameter r, is given by[6]

$$\sigma=\frac{\sigma^0}{r^2-\sigma^0\bar{\sigma}^0}.$$

Therefore, the vanishing of the Jacobian (52) implies that the shear of the congruence diverges. We regain the expression defining caustics from Ref. 2 in the stationary case, namely,

$$r^2-\eth^2L\bar{\eth}^2L=0, \tag{54}$$

since $\sigma^0=\eth^2L$.

The form of the metric tensor in the new coordinates is

$$\begin{aligned}ds^2&=\eta_{ab}\,dx^a\,dx^b\\&=2\,\frac{du}{\dot{L}}\left\{dr+du\left(\frac{1+\eth\overline{\dot{\Phi}}+\Phi\overline{\ddot{\Phi}}}{\dot{L}}\right)+d\bar{\zeta}\left(\frac{\eth\bar{\sigma}^0+\Phi\dot{\bar{\sigma}}^0-\dot{\Phi}\bar{\sigma}^0+\overline{\dot{\Phi}r}}{P}\right)\right.\\&\quad\left.+d\zeta\left(\frac{\bar{\eth}\sigma^0+\bar{\Phi}\dot{\sigma}^0-\overline{\dot{\Phi}}\sigma+\dot{\Phi}r}{P}\right)\right\}-\frac{2r}{P^2}\,(\sigma^0d\zeta^2+\bar{\sigma}^0d\bar{\zeta}^2)-2(r^2+\sigma^0\bar{\sigma}^0)\,\frac{d\zeta\,d\bar{\zeta}}{P^2},\end{aligned} \tag{55}$$

where $P=1+\zeta\bar{\zeta}$. This line element, corresponding to shearing nonstationary null coordinates, defined by Eq. (50) reduces to the one given in Ref. 2 in the stationary regime. As pointed out in this reference it might be of interest to use Eq. (55) as a background metric in linearized gravity for higher order pertubations in problems where gravitational radiation is important.

VI. THE EIKONAL EQUATION IN ASYMPTOTICALLY FLAT SPACE–TIMES

In a straightforward manner all our results can be generalized to the case of arbitrary curved space–times, and the proofs of all the relations above follow basically the same path. We will assume that there is given a system of local coordinates x^a in an arbitrary curved space–time and a two parameter family (sphere's worth) of solutions of the eikonal equation, i.e., $Z(x^a,\zeta,\bar{\zeta})$ such that

$$g^{ab}(x^a)\partial_a Z(x^a,\zeta,\bar{\zeta})\partial_b Z(x^a,\zeta,\bar{\zeta})=0, \tag{56}$$

such that its (null) gradient sweeps out the light cone at x^a as $(\zeta,\bar{\zeta})$ range over the sphere.

Such characteristic functions $S=Z(x^a,\zeta,\bar{\zeta})$ are one of the main variables of the null surface formulation of general relativity; they contain all the conformal information of the space–time. [7] In the special case of asymptotically flat space–times $Z(x^a,\zeta,\bar{\zeta})$ can be interpreted either as the light cone cut of $\mathscr{I}^+$ of the point with coordinates x^a, or as the past light cone of a point at $\mathscr{I}^+$ with coordinates $(u,\zeta,\bar{\zeta})$.[7]

We take the complete solution $\beta Z(x^a,\zeta,\bar{\zeta})$, and define, in an analogous manner to the flat–space construction,

$$u=S^{**}(x^a,\beta,\zeta,\bar{\zeta})=\beta Z(x^a,\zeta,\bar{\zeta})-H(\beta,\zeta,\bar{\zeta}). \tag{57}$$

On Eq. (57) we impose the conditions, equivalent to (15), (16), and (17), namely,

$$\frac{\partial S^{**}}{\partial\beta}=Z(x^a,\zeta,\bar{\zeta})-\frac{\partial H(\beta,\zeta,\bar{\zeta})}{\partial\beta}=0, \tag{58}$$

$$\eth S^{**}=\beta\eth Z(x^a,\zeta,\bar{\zeta})-\eth H(\beta,\zeta,\bar{\zeta})=0, \tag{59}$$

$$\bar{\eth} S^{**}=\beta\bar{\eth} Z(x^a,\zeta,\bar{\zeta})-\bar{\eth} H(\beta,\zeta,\bar{\zeta})=0, \tag{60}$$

and solve for $(\beta,\zeta,\bar{\zeta})$ [as noted earlier, this is always possible, aside from lower dimensional (caustic) regions which can be approached in a limiting fashion] in terms of the x^a.

When these are resubstituted into Eq. (57), S^{**} then becomes a new solution of the eikonal equation since

$$\partial_a S^{**}=\beta\partial_a Z. \tag{61}$$

As in the flat case, we can determine the arbitrary function $H(\beta,\zeta,\bar{\zeta})$ in terms of corresponding data given on a Cauchy surface Σ. Suppose that we are given a coordinate system (τ,x^i) such that $\tau=\tau_0$ corresponds to our Cauchy surface, together with suitable Cauchy data $S_{\text{Cauchy}}(x^i)$ on Σ. A needed generalization of the relationship $\alpha_i=\partial S_{\text{Cauchy}}/\partial x^i$ from Sec. II is

$$\frac{\partial S_{\text{Cauchy}}(x^i)}{\partial x^i}-\frac{\beta\partial Z(x^i,\tau_0,\zeta,\bar{\zeta})}{\partial x^i}=0, \tag{62}$$

which is to be considered as three equations for the determination of x^i in terms of $(\beta,\zeta,\bar{\zeta})$, i.e., $x^i=X^i(\beta,\zeta,\bar{\zeta})$. When these are inserted into Eq. (57) at $\tau=\tau_0$ we obtain

$$H(\beta,\zeta,\bar{\zeta})=\beta Z(X^i(\beta,\zeta,\bar{\zeta}),\tau_0,\zeta,\bar{\zeta})-S_{\text{Cauchy}}(X^i(\beta,\zeta,\bar{\zeta})),$$

in analogy to the results of Sec. II.

The characteristic formulation from Sec. III is even simpler. Starting with any function $L(u_B,\zeta,\bar{\zeta})$ defined on $\mathfrak{I}^+$, we obtain a solution to the eikonal equation with the characteristic data given by

$$u=S^{**}(x^a,\zeta,\bar{\zeta})=L(Z(x^a,\zeta,\bar{\zeta}),\zeta,\bar{\zeta}), \tag{63}$$

where $(\zeta,\bar{\zeta})$ are functions of the coordinates x^a such that the equivalent to Eq. (21) holds, i.e., when

$$\eth S^{**}=\dot{L}\eth Z(x^a,\zeta,\bar{\zeta})-\eth L(Z,\zeta,\bar{\zeta})=0, \tag{64}$$

$$\bar{\eth} S^{**}=\dot{L}\bar{\eth} Z(x^a,\zeta,\bar{\zeta})-\bar{\eth} L(Z,\zeta,\bar{\zeta})=0. \tag{65}$$

Again the relationship between $H(\beta,\zeta,\bar{\zeta})$ and $L(u,\zeta,\bar{\zeta})$ is given by the Legendre transformation

$$L(Z,\zeta,\bar{\zeta})=Z\beta(Z,\zeta,\bar{\zeta})-H(\beta(Z,\zeta,\bar{\zeta}),\zeta,\bar{\zeta}), \tag{66}$$

with

$$Z=\partial H/\partial\beta, \quad \beta=\partial L/\partial Z. \tag{67}$$

VII. CONCLUSION

We have generalized the results of Ref. 2 concerning solutions of the flat-space eikonal equation. We saw two different means of giving data and solving the eikonal equation: the Cauchy, and the characteristic formulation. Each one leads to different methods. The two methods are beautifully related by a Legendre transformation, Eqs. (41) and (42). Moreover, all our results can be generalized to the case of curved space–times. The characteristic formulation appears to be better for the study of the dynamics of the wave fronts. By means of a suitable parameterization we could describe the caustics in the wave fronts, and find a simple geometric interpretation in terms of the shear σ of the null congruence generating the wave fronts.

ACKNOWLEDGMENTS

E.T.N. thanks the NSF for support under Grant No. PHY 97-22049 and A.P. thanks FUNDACION YPF, Argentina, for its support. We extend our appreciation to Simonetta Frittelli for many stimulating conversations.

[1] V. I Arnold, *Catastrophe Theory* (Springer, Berlin, 1986).
[2] S. Frittelli, E. T. Newman, and G. Silva-Ortigoza, J. Math. Phys. (to be published).
[3] L. Landau and Lifschitz, *Classical Mechanics* (Pergamon, London, 1960).
[4] B. Aronson, R. Lind, J. Messmer, and E. T. Newman, J. Math. Phys. **12**, 2462 (1971).
[5] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, Berlin, 1980).
[6] R. W. Lind and E. T. Newman, J. Math. Phys. **15**, 1103 (1974).
[7] S. Fritelli, C. Kozameh, and E. T. Newman, J. Math. Phys. **36**, 4984 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 2 FEBRUARY 1999

Application of geometric probability techniques to the evaluation of interaction energies arising from a general radial potential

David Schleef,[a] Michelle Parry,[b] Shu-Ju Tu,
Brian Woodahl, and Ephraim Fischbach
Department of Physics, Purdue University, West Lafayette, Indiana 47907

(Received 10 August 1998; accepted for publication 15 October 1998)

A formalism is developed for using geometric probability techniques to evaluate interaction energies arising from a general radial potential $V(r_{12})$, where $r_{12}=|\mathbf{r}_2-\mathbf{r}_1|$. The integrals that arise in calculating these energies can be separated into a radial piece that depends on r_{12} and a nonradial piece that describes the geometry of the system, including the density distribution. We show that all geometric information can be encoded into a ''radial density function'' $G(r_{12};\rho_1,\rho_2)$, which depends on r_{12} and the densities ρ_1 and ρ_2 of two interacting regions. $G(r_{12};\rho_1,\rho_2)$ is calculated explicitly for several geometries and is then used to evaluate interaction energies for several cases of interest. Our results find application in elementary particle, nuclear, and atomic physics. © *1999 American Institute of Physics.* [S0022-2488(99)00102-4]

I. INTRODUCTION

In many areas of physics, integrals of the form

$$U=\int d^3r_1\ d^3r_2\ \rho(\mathbf{r}_1)\rho(\mathbf{r}_2)V(|\mathbf{r}_1-\mathbf{r}_2|) \tag{1.1}$$

are encountered, which typically describe the self-energy of a system with density profile $\rho(\mathbf{r})$ in the presence of a two-body central potential $V(|\mathbf{r}_1-\mathbf{r}_2|)$. A familiar example of such an integral arises in the calculation of the electrostatic self-energy of a spherical charge distribution (e.g., a nucleus) due to the Coulomb potential $V_C(|\mathbf{r}_1-\mathbf{r}_2|)$,

$$V_C(|\mathbf{r}_1-\mathbf{r}_2|)=\frac{e_0^2}{|\mathbf{r}_1-\mathbf{r}_2|}, \tag{1.2}$$

where e_0 is the electric charge ($e_0^2\cong\frac{1}{137}$). For a simple potential such as $V_C(|\mathbf{r}_1-\mathbf{r}_2|)$, the integral in Eq. (1.1) can be evaluated directly, by expanding $1/|\mathbf{r}_1-\mathbf{r}_2|$ in terms of Legendre polynomials. However, for some types of potentials, evaluating U in this way can be extremely tedious. An example of current interest[1,2] is the self-energy of a nucleus or a neutron star arising from neutrino–antineutrino ($\nu-\overline{\nu}$) exchange. In this case the analog of V_C in Eq. (1.2) for the neutron–neutron ($n-n$) potential in a neutron star arising from $\nu-\overline{\nu}$ exchange is[3–5]

$$V_{\nu\nu}(|\mathbf{r}_1-\mathbf{r}_2|)=\frac{G_F^2a_n^2}{4\pi^3|\mathbf{r}_1-\mathbf{r}_2|^5}, \tag{1.3}$$

where G_F is the weak Fermi constant, and $a_n=-\frac{1}{2}$ the coupling constant describing the strength of the $\nu-n$ interaction. One of the difficulties that arises in evaluating U, starting from Eq. (1.3),

---

[a]Present address: Department of Physics, University of California—Berkeley, Berkeley, California 94720.
[b]Present address: Department of Natural Sciences, Longwood College, Farmville, Virginia 23909.

is that the integral is well defined only if the neutron–neutron hard core interaction is used to cut off the lower limit of integration when $|\mathbf{r}_1-\mathbf{r}_2|<r_c\cong 0.5\times 10^{-13}$ cm. However, since this constraint applies to $r_{21}=|\mathbf{r}_1-\mathbf{r}_2|=r_{12}$, and not to $r_1=|\mathbf{r}_1|$ or $r_2=|\mathbf{r}_2|$ separately, the integration region in Eq. (1.1) implied by this constraint is somewhat complicated. As we discuss in detail below, the evaluation of integrals involving potentials such as $V_{\nu\nu}(r_{12})$, and other potentials as well, can be greatly facilitated using geometric probability techniques. By use of these techniques the six-dimensional integral in Eq. (1.1) can be replaced by a one-dimensional integral in the variable r_{12}, which can be easily integrated in all cases of interest. The geometric probability techniques are especially useful when $\rho(\mathbf{r}_1)$ is radially varying ($\rho(\mathbf{r}_1)=\rho(|\mathbf{r}_1|)$).

It is helpful to introduce the formalism of geometric probability by first considering the electrostatic (Coulomb) energy of a uniform spherical charge distribution of radius R. Direct evaluation of the six-dimensional integral in Eq. (1.1) yields

$$U_{\mathrm{C}}=\frac{6}{5}\,\frac{e_0^2}{R}. \tag{1.4}$$

For a spherically symmetric distribution containing Z charges there are $Z(Z-1)/2$ possible pairs, and hence the total Coulomb energy W_{C} of such a distribution is

$$W_{\mathrm{C}}=\frac{Z(Z-1)}{2}\,U_{\mathrm{C}}=\frac{3}{5}\,Z(Z-1)\,\frac{e_0^2}{R}, \tag{1.5}$$

which is the standard result.[6,7]

In contrast to the preceding derivation, which begins with a six-dimensional integral, the formalism of integral geometry expresses U_{C} immediately as a one-dimensional integral. For any function $g(r_{12})$, its average value $\langle g\rangle$ taken over a uniform spherical volume of radius R is

$$\langle g\rangle=\int_0^{2R}dr_{12}\,\mathcal{P}_3(r_{12})g(r_{12}), \tag{1.6}$$

where

$$\int_0^{2R}dr_{12}\,\mathcal{P}_3(r_{12})=1. \tag{1.7}$$

The function $\mathcal{P}_3(r_{12})$ denotes the normalized probability density for finding two points randomly chosen in a uniform three-dimensional sphere to be a distance r_{12} apart. The functional form of $\mathcal{P}_3(r_{12})$ has been obtained previously by a number of authors,[8–13]

$$\mathcal{P}_3(r_{12})=\frac{3r_{12}^2}{R^3}\left[1-\frac{3}{2}\left(\frac{r_{12}}{2R}\right)+\frac{1}{2}\left(\frac{r_{12}}{2R}\right)^3\right]. \tag{1.8}$$

Using Eq. (1.8), U_{C} is given by

$$U_{\mathrm{C}}=\langle e_0^2/r_{12}\rangle=\int_0^{2R}dr_{12}\left(\frac{3r_{12}^2}{R^3}\right)\left[1-\frac{3}{2}\left(\frac{r_{12}}{2R}\right)+\frac{1}{2}\left(\frac{r_{12}}{2R}\right)^3\right]\left(\frac{e_0^2}{r_{12}}\right)=\frac{6}{5}\,\frac{e_0^2}{R}, \tag{1.9}$$

in agreement with Eq. (1.4). The utility of the geometric probability formalism becomes more evident when one attempts to evaluate $U_{\nu\nu}=\langle V_{\nu\nu}(r_{12})\rangle$ using Eq. (1.3),

$$U_{\nu\nu}=\frac{G_{\mathrm{F}}^2a_n^2}{4\pi^3}\int_{r_c}^{2R}dr_{12}[\,\eta(r_c,R)\mathcal{P}_3(r_{12})]\,\frac{1}{r_{12}^5}. \tag{1.10}$$

In Eq. (1.10) $\eta(r_c,R)$ is a constant that ensures that $\mathcal{P}_3(r_{12})$ is appropriately normalized in the interval $2R \geqslant r_{12} \geqslant r_c$, and is given by

$$\eta(r_c,R)=(1-8s_c^3+8s_c^4-2s_c^6)^{-1}, \tag{1.11}$$

where $s_c=r_c/2R$. It follows from Eq. (1.11) that $\eta(0,R)=1$, as expected. Combining Eqs. (1.10) and (1.11) then gives immediately,

$$U_{\nu\nu}=\frac{3}{8\pi^3}\frac{G_F^2 a_n^2}{\hbar c}\frac{1}{R^3 r_c^2}\left(1-\frac{r_c}{2R}\right)^3\eta(r_c,R). \tag{1.12}$$

For a sphere containing N particles, the total energy $W_{\nu\nu}$ is then given by

$$W_{\nu\nu}=\frac{N(N-1)}{2}U_{\nu\nu}=\frac{3}{16\pi^3}N(N-1)\frac{G_F^2 a_n^2}{\hbar c}\frac{1}{R^3 r_c^2}\left(1-\frac{r_c}{2R}\right)^3\eta(r_c,R). \tag{1.13}$$

To evaluate $\langle g\rangle$ in Eq. (1.6) for a particular geometry, one must first determine the functional form of $\mathcal{P}_3(r_{12})$ appropriate to that geometry. In practice, it would be of great value to know $\mathcal{P}(r)\equiv\mathcal{P}_3(r_{12})$ for different (nonconstant) density distributions, as well as for other geometries. In this paper we address the former problem, by developing a general framework for calculating $\mathcal{P}(r)$ for geometries with variable density. We illustrate this approach in Sec. II by first rederiving (in a much simpler way) the result for a sphere of constant density given in Eq. (1.8). We then obtain $\mathcal{P}(r)$ for a sphere with a Gaussian density distribution. In Sec. III we apply our formalism to geometries that can be used to calculate the interaction energy between microscopic objects due to a generalized two-body interaction potential. One example is the van der Waals interaction.

II. GENERAL FORMALISM

A. The radial density function

Returning to Eq. (1.1), we introduce the change of variables,

$$\begin{aligned}\mathbf{r}_{12}&=\mathbf{r}_2-\mathbf{r}_1,\\ d^3r_{12}&=d^3r_2,\end{aligned} \tag{2.1}$$

so that

$$\begin{aligned}U&=\int dr_{12}\left[r_{12}^2\int d\Omega_{12}\int d^3r_1\ \rho_1(\mathbf{r}_1)\rho_2(\mathbf{r}_{12}+\mathbf{r}_1)\right]V(r_{12}),\\ &\equiv\int dr_{12}\ G(r_{12};\rho_1,\rho_2)V(r_{12}),\end{aligned} \tag{2.2}$$

where $r_{12}=|\mathbf{r}_2-\mathbf{r}_1|$. The ''radial density function'' $G(r_{12};\rho_1,\rho_2)$ is the generalization of the probability function $\mathcal{P}_3(r_{12})$ in Eq. (1.8). $G(r_{12};\rho_1,\rho_2)$ incorporates all the geometric information about the densities $\rho_1(\mathbf{r}_1)$ and $\rho_2(\mathbf{r}_2)$ and the geometry, but is independent of $V(r_{12})$.

B. Geometry with spherical symmetry

The first case we consider is when both $\rho_1(\mathbf{r}_1)$ and $\rho_2(\mathbf{r}_2)$ exhibit spherical symmetry about a common origin, so that $\rho_1=\rho_1(|\mathbf{r}_1|)$ and $\rho_2=\rho_2(|\mathbf{r}_2|)$ about this origin. From Eq. (2.2) we can then write

$$G(r_{12};\rho_1,\rho_2)=r_{12}^2\int d^3r_1\int d\Omega_{12}\ \rho_1(r_1)\rho_2(|\mathbf{r}_{12}+\mathbf{r}_1|). \tag{2.3}$$

Since $\rho(r_1)$ and $\rho_2(|\mathbf{r}_{12}+\mathbf{r}_1|)=\rho(r_2)$ are independent of $d\Omega_1$ and $d\phi_{12}$, we can integrate over these variables immediately to give

$$G(r_{12};\rho_1,\rho_2)=r_{12}^2\int_0^\infty dr_1\, 4\pi r_1^2\rho_1(r_1)\int_0^\pi d\theta_{12}\, 2\pi\, \sin\, \theta_{12}\rho_2(|\mathbf{r}_{12}+\mathbf{r}_1|). \tag{2.4}$$

Note that the upper limit of integration for r_1 can always be taken to be infinite, even for a finite spherical mass distribution, since $\rho_1(r_1)$ can be defined to be zero for $r_1>R$. Using the law of cosines, we have

$$-\cos\, \theta_{12}=\frac{r_{12}^2+r_1^2-r_2^2}{2r_1r_{12}}. \tag{2.5}$$

Since r_{12} and r_1 are the independent variables of integration in Eq. (2.3), it follows that cos θ_{12} depends only on r_2 for fixed values of r_{12} and r_1. Thus,

$$\sin\, \theta_{12}\, d\theta_{12}=\frac{-r_2}{r_1r_{12}}\, dr_2. \tag{2.6}$$

Combining Eqs. (2.6) and (2.4) then gives

$$G(r_{12};\rho_1,\rho_2)=8\pi^2 r_{12}\int_0^\infty dr_1\, r_1\rho_1(r_1)\int_{|r_{12}-r_1|}^{r_{12}+r_1} dr_2\, r_2\rho_2(r_2). \tag{2.7}$$

As an application of Eq. (2.7) we recalculate the Coulomb energy of a sphere of radius R and constant density $1/V$, where the density is normalized so that its integral over the spherical volume is unity. Since the integral in Eq. (2.7) is symmetric in the interchange of r_1 and r_2, we can write

$$G_{\text{sphere}}(r_{12};\rho_1,\rho_2)=16\pi^2 r_{12}\int_{r_{12}/2}^\infty dr_1\, r_1\rho(r_1)\int_{|r_{12}-r_1|}^{r_1} dr_2\, r_2\rho(r_2). \tag{2.8}$$

The lower limit on the r_1 integration follows by noting that when $r_2=r_1$ the triangle formed by the vectors $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_{12}$ is isosceles, and hence by the triangle inequality $2r_1>r_{12}$. From Eq. (2.8) we have

$$G_{\text{sphere}}(r_{12};\rho_1,\rho_2)=\frac{16\pi^2 r_{12}}{V^2}\int_{r_{12}/2}^R dr_1\, r_1\int_{|r_{12}-r_1|}^{r_1} dr_2\, r_2=\frac{3r_{12}^2}{R^3}\left[1-\frac{3}{2}\left(\frac{r_{12}}{2R}\right)+\frac{1}{2}\left(\frac{r_{12}}{2R}\right)^3\right], \tag{2.9}$$

in agreement with the expression for $\mathcal{P}_3(r_{12})$ in Eq. (1.8). The expression for the Coulomb energy of a sphere of charge then follows immediately from Eq. (1.9). Having demonstrated that the present formalism correctly reproduces the classical results for a sphere of constant density, we turn in the next section to a problem that has not been considered previously in the literature, the distribution of points in a sphere with a Gaussian density variation.

III. RADIAL DENSITY FUNCTION FOR A GAUSSIAN DISTRIBUTION

We derive in this section the radial density function for a spherically symmetric distribution of matter centered at the origin, whose density varies as

$$\rho(r)=Ae^{-r^2/R_0^2}, \tag{3.1}$$

where A and R_0 are constants, and r is measured from the origin. If we normalize $\rho(r)$ so that its integral over all space is unity, then

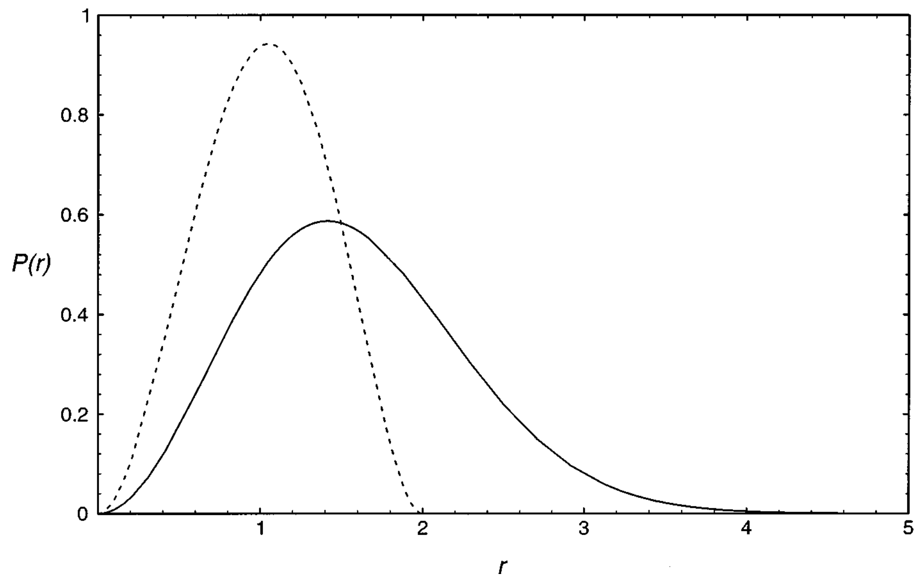

FIG. 1. The plot of $P(r)\equiv G(r;\rho)$ in Eq. (3.6) as a function of $r=r_{12}$ (solid line). For comparison the function $\mathcal{P}_3(r)$ in Eq. (1.8) is also shown (dashed line).

$$A=R_0^{-3}\pi^{-3/2}. \tag{3.2}$$

Combining Eqs. (3.1) and (2.8), we find

$$G(r_{12};\rho_1,\rho_2)\equiv G(r_{12};\rho)=8\pi^2A^2r_{12}\int_0^\infty dr_1r_1e^{-r_1^2/R_0^2}\int_{|r_{12}-r_1|}^{r_{12}+r_1}dr_2r_2e^{-r_2^2/R_0^2}. \tag{3.3}$$

Carrying out the integration with respect to r_2, we find

$$G(r_{12};\rho)=4\pi^2A^2R_0^2r_{12}e^{-r_{12}^2/R_0^2}\int_0^\infty dr_1r_1e^{-2r_1^2/R_0^2}[e^{-2r_{12}r_1/R_0^2}-e^{2r_{12}r_1/R_0^2}]. \tag{3.4}$$

The integration with respect to r_1 can then be performed by completing the square, which gives

$$G(r_{12};\rho)=4\pi^2A^2R_0^2r_{12}e^{-r_{12}^2/2R_0^2}\int_0^\infty dr_1\ r_1\{\exp[-2(r_1-r_{12}/2)^2/R_0^2]-\exp[-2(r_1+r_{12}/2)^2/R_0^2]\}$$

$$=4\pi^2A^2R_0^2r_{12}e^{-r_{12}/2R_0^2}\left[\frac{1}{2}\sqrt{\frac{\pi}{2}}\ r_{12}R_0\right]. \tag{3.5}$$

Combining Eqs. (3.2) and (3.5) yields the final result,

$$G(r_{12};\rho)=\sqrt{\frac{2}{\pi}}\frac{r_{12}^2}{R_0^3}\ e^{-r_{12}^2/2R_0^2}. \tag{3.6}$$

$G(r_{12};\rho)$ is shown in Fig. 1 and is normalized to unity over the interval $[0,\infty]$. When the lower limit of integration is replaced by r_c, $G(r_{12};\rho)$ must be divided by the constant $C(r_c,R_0)$ to be properly normalized, where

$$C(r_c,R_0)=\int_{r_c}^\infty dr_{12}\ G(r_{12};\rho)\cong 1-\sqrt{\frac{2}{\pi}}\frac{r_c^3}{3R_0^3}. \tag{3.7}$$

We note that for $r_{12}^2/R_0^2 \ll 1, G(r_{12};\rho)$ can be approximated by

$$G(r_{12};\rho) \cong \sqrt{\frac{2}{\pi}} \frac{r_{12}^2}{R_0^3}, \tag{3.8}$$

which agrees (up to an overall constant) with the results for a uniform sphere given in Eqs. (1.8) and (2.9). This agreement conforms to our intuition that when r_{12} is small compared to R_0, a spherically symmetric Gaussian distribution will look like that of a sphere with an approximately constant local density.

The result in Eq. (3.6) can be applied immediately to calculate both the Coulomb energy and the neutrino-exchange energy of a matter distribution with the Gaussian density profile given in Eq. (3.1). The Coulomb energy W_C is then given by

$$W_C = \frac{Z(Z-1)}{2} \left\langle \frac{e^2}{r_{12}} \right\rangle = \frac{Z(Z-1)}{2} \int_0^\infty dr_{12} \left(\frac{e^2}{r_{12}} \right) \times \frac{r_{12}^2}{R_0^3} \sqrt{\frac{2}{\pi}} e^{-r_{12}^2/2R_0^2} = \frac{1}{\sqrt{2\pi}} \frac{Z(Z-1)e^2}{R_0}. \tag{3.9}$$

As noted in the Introduction, geometric probability techniques are particularly useful when evaluating expressions where the nucleon–nucleon hard core radius r_c appears, as in the integral for $U_{\nu\nu}$ in Eq. (1.10). From Eq. (3.6) we have, for a Gaussian density distribution of N neutrons,

$$W_{\nu\nu} = \frac{N(N-1)}{2} \left\langle \frac{G_F^2 a_n^2}{4\pi^3 r_{12}^5} \right\rangle = \frac{N(N-1)}{2C(r_c, R_0)} \int_{r_c}^\infty dr_{12} \left(\frac{G_F^2 a_n^2}{4\pi^3 r_{12}^5} \right) \frac{r_{12}^2}{R_0^3} \sqrt{\frac{2}{\pi}} e^{-r_{12}^2/2R_0^2}. \tag{3.10}$$

Evaluation of the integral in Eq. (3.10) yields

$$W_{\nu\nu} = \frac{G_F^2 a_n^2}{8\pi^3} \frac{N(N-1)}{C(r_c, R_0)} \left\{ \frac{1}{\sqrt{2\pi}} \frac{e^{-r_c^2/R_0^2}}{r_c^2 R_0^3} + \frac{1}{2} \sqrt{\frac{\pi}{2}} \frac{1}{R_0^5} \left[-i + \frac{1}{\pi} Ei \left(\frac{-r_c^2}{2R_0^5} \right) \right] \right\}, \tag{3.11}$$

$$Ei(z) = P \int_{-z}^\infty \frac{(-1)}{te^t} dt,$$

where P denotes the principal value integration. We note that the quantity in square brackets in Eq. (3.11) is real, as hence $W_{\nu\nu}$ is real as well. As can be seen from Eq. (3.11), by using $G(r_{12};\rho)$ in Eq. (3.6) we obtain an exact closed-form expression for $W_{\nu\nu}$ for the case of a Gaussian density distribution. By way of contrast, the conventional approach would lead to an infinite series expression for $W_{\nu\nu}$. We complete this discussion by noting that for $r_c/R_0 \ll 1$ we can write

$$\int_{r_c}^\infty dr \frac{G(r;\rho)}{r^5} \cong \frac{2}{\sqrt{\pi}} \frac{1}{r_c^2 R_0^3}, \tag{3.12}$$

and, hence,

$$W_{\nu\nu} \cong \frac{G_F^2 a_n^2 N(N-1)}{8\sqrt{2}\pi^{7/2}} \frac{1}{r_c^2 R_0^3}. \tag{3.13}$$

As expected from Eq. (1.13), $W_{\nu\nu} \sim 1/r_c^2$ when $r_c/R_0 \ll 1$ for the Gaussian distribution, just as in the case of the uniform sphere.

IV. INTERACTION BETWEEN SOURCES

A. General formalism

In the previous section we have focused on calculating the radial density function $G(r_{12};\rho_1,\rho_2)$ needed to evaluate the self-energy of a spherically symmetric matter distribution. In this section we calculate the analogous expressions for $G(r_{12};\rho_1,\rho_2)$, which characterize the interaction of two different matter distributions in volumes ω_1 and ω_2, respectively. In particular, we generalize the calculations of Israelachvili[14] to allow any two-body radial potential. These results are of interest in the field of tribology, specifically in calculating interaction forces and energies due to van der Waals-type forces. This technique has been used to study the force of interaction between an Atomic Force Microscopy (AFM) probe tip and a flat sample.[15]

Returning to Eq. (2.3), we can rewrite the expression for $G(r_{12};\rho_1,\rho_2)$ in the form

$$G(r_{12};\rho_1,\rho_2)=r_{12}^2\int d\Omega_{12}\int d^3r_1\ \rho_1(\mathbf{r}_1)\rho_2(\mathbf{r}_{12}+\mathbf{r}_1)$$
$$=\int d^3r_1\ \rho_1(\mathbf{r}_1)\left\{r_{12}^2\int d\Omega_{12}\ \rho_2(\mathbf{r}_2)\right\}. \tag{4.1}$$

In Eq. (4.1) we have interchanged the order of the integrations, and have used Eq. (2.1) to replace $\mathbf{r}_{12}+\mathbf{r}_1$ by $\mathbf{r}_2$. In this section we deal with the situation in which $\rho_i(\mathbf{r}_i)$ are given by

$$\rho_i(\mathbf{r}_i)=\begin{cases}\rho_i, & \text{when } \mathbf{r}_i\in\omega_i,\\ 0, & \text{otherwise.}\end{cases} \tag{4.2}$$

For illustrative purposes we take ρ_1 and ρ_2 to be constants, so that

$$G(r_{12};\rho_1,\rho_2)=\rho_1\rho_2\int_{\omega_1}d^3r_1\left\{r_{12}^2\int_{4\pi}d\Omega_{12}\right\}\equiv\rho_1\rho_2\int_{\omega_1}d^3r_1\ S(r_{12},\mathbf{r}_1). \tag{4.3}$$

$S(r_{12},\mathbf{r}_1)$ can be viewed as the surface area formed by the intersection of a sphere centered at $\mathbf{r}_1=0$ (in the volume ω_1) and having radius r_{12}, with the second volume ω_2. Several examples will serve to clarify the application of Eq. (4.3).

B. Point to sphere

Here ω_1 is a point having an infinitesimal volume $d\tau$, so that Eq. (4.3) becomes

$$G(r_{12};\rho_1,\rho_2)=(\rho_1\ d\tau)\rho_2 S(r_{12},\mathbf{r}_1). \tag{4.4}$$

If ω_2 is a sphere of radius R, then, from Fig. 2,

$$R^2=r_{12}^2+r^2-2rr_{12}\cos\theta_0, \tag{4.5}$$

where r is the distance from the point to the center of the spherical distribution ω_2. It follows that

$$S(r_{12},\mathbf{r}_1)=2\pi r_{12}^2\int_0^{\theta_0}\sin\theta_{12}\,d\theta_{12}=\pi\frac{r_{12}}{r}[R^2-(r-r_{12})^2]. \tag{4.6}$$

Combining Eqs. (4.4) and (4.6) then gives

$$G(r_{12};\rho_1,\rho_2)=(\rho_1 d\tau)\rho_2\left\{\pi\frac{r_{12}}{r}[R^2-(r-r_{12})^2]\right\}. \tag{4.7}$$

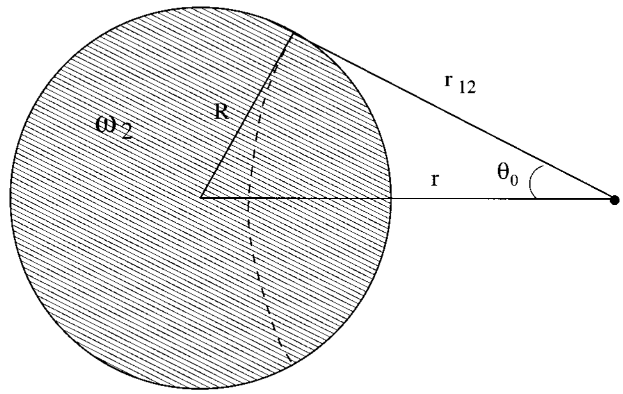

FIG. 2. The representation of the point-to-sphere geometry. R is the radius of the sphere, whose center is a distance r from the external point. $G(r_{12};\rho_1,\rho_2)$ is calculated as a function of the distance r_{12} between the external point and a point in the sphere.

Equation (4.7) can be checked by noting that when $r=R$, $G(r_{12};\rho_1,\rho_2)$ describes the distribution of distances between two points in a sphere, given that one of these points lies on the surface of the sphere. The latter distribution has been derived by Parry,[16] and it is straightforward to show that Eq. (4.7) agrees with this result when $r=R$. When combined with Eq. (2.3), Eq. (4.7) allows the interaction energy U to be calculated for an arbitrary two-body potential $V(r_{12})$ (e.g., Coulomb, Yukawa, van der Waals, etc.).

C. Point to half-space

This geometry is very similar to the point-to-sphere case, except that ω_2 is now an infinite half-space separated by a distance r from an external point. For this geometry, cos θ_0 is given by

$$\cos\ \theta_0=\frac{r}{r_{12}}, \tag{4.8}$$

and hence

$$S(r_{12},\mathbf{r}_1)=2\pi r_{12}^2(1-\cos\ \theta_0)=2\pi r_{12}(r_{12}-r). \tag{4.9}$$

Combining Eqs. (4.4) and (4.9), the radial density function is given by

$$G(r_{12};\rho_1,\rho_2)=(\rho_1\ d\tau)\rho_2\{2\pi r_{12}(r_{12}-r)\}. \tag{4.10}$$

As in the previous case, the expression in Eq. (4.10) can be checked by noting that when $r=0$, $G(r_{12};\rho_1,\rho_2)$ becomes proportional to r_{12}^2, which is the expected result for an infinite half-space.[16]

D. Arbitrary volume to half-space

We can apply the previous result to compute the radial density function for an arbitrary volume ω_1, in the presence of an infinite half-space. From Eq. (4.9) we see that $S(r_{12},\mathbf{r}_1)$ depends only on the distance x of a volume element from the boundary, and hence we need only specify the expression for the cross section $A(x)$ of ω_1 as a function of x. Then, from Eq. (4.10) we have

$$G(r_{12};\rho_1,\rho_2)=2\pi\rho_1\rho_2\int_0^{r_{12}}dx(r_{12}-x)A(x). \tag{4.11}$$

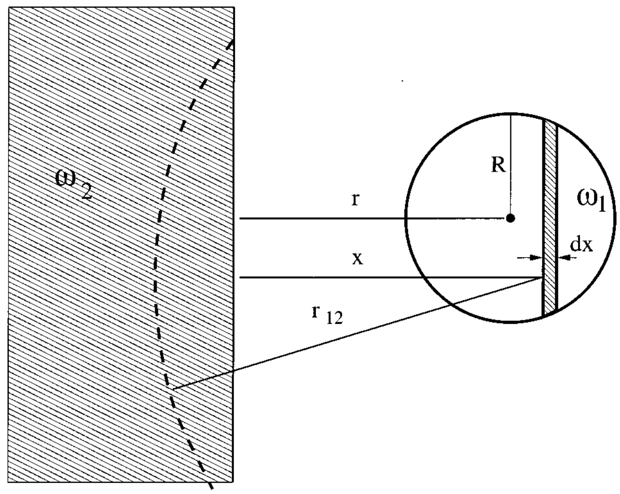

FIG. 3. Representation of the sphere-to-half-space geometry. See the text and the caption to Fig. 2 for details.

If ω_1 is a sphere of radius R whose center is a distance r from the boundary of the half-space, then, from Fig. 3,

$$A(x)=\begin{cases} 0, & x\leqslant r-R, \\ \pi[R^2-(r-x)^2], & r-R\leqslant x\leqslant r+R, \\ 0, & x\geqslant r+R. \end{cases} \tag{4.12}$$

Correspondingly, the density function is divided into three regions: $G(r_{12};\rho_1,\rho_2)=0$ if $r_{12}\leqslant(r-R)$, and

$$G(r_{12};\rho_1,\rho_2)=\begin{cases} \dfrac{\pi^2}{6}\,\rho_1\rho_2 r_{12}(r-R-r_{12})^3(r_{12}-r-3R), & r-R\leqslant r_{12}\leqslant r+R, \\ \dfrac{8\pi^2}{3}\,\rho_1\rho_2 r_{12}R^3(r_{12}-r), & r+R\leqslant r_{12}. \end{cases} \tag{4.13}$$

The results in Eq. (4.13) are useful in Atomic Force Microscopy since they can be used to analyze the interaction of a general AFM probe tip interacting with a flat sample.

V. CONCLUSIONS

The discussion in the Introduction illustrates the power of geometric probability techniques by demonstrating how a six-dimensional integral can be immediately reduced to a straightforward one-dimensional problem. In practice, this facilitates the evaluation of interaction energies such as $U_{\nu\nu}$ in Eq. (1.10), which would be extremely difficult to treat otherwise, due to the presence of r_c. We have extended the classical results of Refs. 8–13 to calculate for the first time the radial density functions for a Gaussian density profile, and for two regions of different shapes interacting with each other. These results can be applied to a wide variety of physical systems, as we will discuss elsewhere.

ACKNOWLEDGMENTS

One of the authors (E.F.) wishes to thank Albert Overhauser for introducing him to the subject of geometric probability, and for making available the unpublished results in Ref. 11. We also wish to thank Dennis Krause and Michael Yurko for helpful discussions. This work was supported in part by the U.S. Department of Energy under Contract No. DE-AC02-76ER01428.

[1] E. Fischbach, D. E. Krause, C. Talmadge, and D. Tadić, Phys. Rev. D **52**, 5417 (1995).
[2] E. Fischbach, Ann. Phys. (N.Y.) **247**, 213 (1996).
[3] G. Feinberg and J. Sucher, Phys. Rev. **166**, 1638 (1968).
[4] G. Feinberg, J. Sucher, and C.-K. Au, Phys. Rep. **180**, 83 (1989).
[5] S. D. H. Hsu and P. Sikivie, Phys. Rev. D **49**, 4951 (1994).
[6] G. F. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, 4th ed. (Academic, San Diego, 1995), p. 749.
[7] A. Bohr and B. R. Mottelson, *Nuclear Structure* (Benjamin, New York, 1969), Vol. I, p. 141.
[8] R. Deltheil, Ann. Fac. Sci. Univ. Toulouse **11**, 1 (1919).
[9] J. M. Hammersly, Ann. Math. Stat. **21**, 447 (1950).
[10] R. D. Lord, Ann. Math. Stat. **25**, 794 (1954).
[11] A. W. Overhauser (unpublished).
[12] M. G. Kendall and P. A. P. Moran, *Geometrical Probability* (Hafner, New York, 1963).
[13] L. A. Santaló, *Integral Geometry and Geometric Probability* (Addison-Wesley, Reading, MA, 1976), p. 212.
[14] J. N. Israelachvili, *Intermolecular and Surface Forces*, 2nd ed. (Academic, London, 1992).
[15] B. Gady, D. Schleef, R. Reifenberger, D. Rimai, and L. P. DeMejo, Phys. Rev. B **53**, 8065 (1996).
[16] M. Parry, Ph.D. thesis, Purdue University, 1998 (unpublished).

Erratum: ''On the dynamics of characteristic surfaces'' [J. Math. Phys. 36, 6397–6416 (1995)]

Simonetta Frittelli[a]
Department of Physics, Duquesne University, Pittsburgh, Pennsylvania 15282

Carlos N. Kozameh
FaMAF, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

Erza T. Newman
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 14 September 1998; accepted for publication 15 September 1998)

[S0022-2488(99)00402-8]

We list a series of corrections to equations appearing in Ref. 1. None of the following affect the content of the article. Equation (m_I) in the set of Eqs. (1) is missing a factor $\frac{1}{2}$. The correct equation reads as

$$(m_I) \quad \eth\Omega = \tfrac{1}{2} W(\Lambda,_1)\Omega. \tag{1}$$

On p. 6399, the expression for Q given in between Eqs. (1) and (2) is incorrect by an overall minus sign. The correct expression reads as

$$Q \equiv \frac{1}{4q}\overline{\Lambda},_{11}\Lambda,_{11} + \frac{3}{8q^2}(q,_1)^2 - \frac{1}{4q}q,_{11}\,. \tag{2}$$

There is a typo in the first line of Eq. (71) for the Green function. The correct expression reads as

$$\begin{aligned}\overline{K}^{+}_{2,0} \overset{\text{def}}{=} & -\frac{1}{4\pi}\frac{(1+\overline{\zeta}\eta)^2(\overline{\eta}-\overline{\zeta})}{(1+\zeta\overline{\zeta})(1+\eta\overline{\eta})(\eta-\zeta)} \\ = & -\frac{1}{4\pi}\frac{m\cdot\widehat{\overline{m}}\; m\cdot\hat{l}\; l\cdot\hat{m}}{l\cdot\hat{l}\; l\cdot\hat{n}}.\end{aligned} \tag{3}$$

On p. 6404, the expressions for σ and $\overline{\sigma}$ given by Eqs. (32) are mismatched. The correct expressions are

$$\overline{\sigma} = \frac{\overline{\eta}^2}{A}D\left(\frac{\overline{\xi}}{\overline{\eta}}\right), \quad \sigma = \frac{\eta^2}{A}D\left(\frac{\xi}{\eta}\right). \tag{4}$$

On p. 6405, the expression for ρ given by (44) is incorrect. The correct expression reads as

$$\rho = \tfrac{1}{2}D\;\ln(\Omega^2\sqrt{q}). \tag{5}$$

[1] S. Frittelli, E. T. Newman, and C. N. Kozameh, ''On the dynamics of characteristic surfaces,'' J. Math. Phys. **36**, 6397–6416 (1995).

[a] Electronic mail: simo@mayu.physics.duq.edu

Erratum: ''Lorentzian metrics from characteristic surfaces'' [J. Math. Phys. 36, 4975–4983 (1995)]

Simonetta Frittelli,[a)]
Department of Physics, Duquesne University, Pittsburgh, Pennsylvania 15282

Carlos N. Kozameh
FaMAF, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

Ezra T. Newman
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 14 September 1998; accepted for publication 15 September 1998)

[S0022-2488(99)00602-7]

We list a series of corrections to equations appearing in Ref. 1. None of the following affect the content of the article. Equations (14b) and (14d) have misprints. The correct equations read as

$$(1-\lambda_1\lambda_1^*)T_j^1=(\partial^*\lambda_j+\lambda_0\delta_j^-+\lambda_+\delta_j^1+\lambda_-\lambda_j^*)+\lambda_1(\partial\lambda_j^*+\lambda_0^*\delta_j^++\lambda_-^*\delta_j^1+\lambda_+^*\lambda_j), \tag{1}$$

$$(1-\lambda_1\lambda_1^*)T_j^{*1}=(\partial\lambda_j^*+\lambda_0^*\delta_j^++\lambda_-^*\delta_j^1+\lambda_+^*\lambda_j)+\lambda_1^*(\partial^*\lambda_j+\lambda_0\delta_j^-+\lambda_+\delta_j^1+\lambda_-\lambda_j^*). \tag{2}$$

Equations (24) have missing parentheses and misprints. The correct equations read as

$$\partial^* g^{0+}-\partial g^{0-}\equiv 0, \tag{3}$$

$$\partial^*(g^{++}+g^{01}\lambda_1)-\partial(g^{+-}+g^{01})\equiv 0, \quad \partial^*(g^{+-}+g^{01})-\partial(g^{--}+g^{01}\lambda_1^*)\equiv 0, \tag{4}$$

$$\partial^*(g^{1+}+g^{-i}\lambda_i)-\partial(g^{1-}+g^{+i}\lambda_i^*)\equiv 0. \tag{5}$$

Equations (26a) and (26b). A factor $\frac{1}{2}$ is missing from both. The correct equations read as

$$W(1-\tfrac{1}{4}\lambda_1\lambda_1^*)=-(\tfrac{1}{2}\partial^*\lambda_1-T_1^1-\tfrac{1}{2}\lambda_1\{\tfrac{1}{2}\partial\lambda_1^*-T_1^{*1}\}), \tag{6}$$

$$W^*(1-\tfrac{1}{4}\lambda_1\lambda_1^*)=-(\tfrac{1}{2}\partial\lambda_1^*-T_1^{*1}-\tfrac{1}{2}\lambda_1^*\{\tfrac{1}{2}\partial^*\lambda_1-T_1^1\}). \tag{7}$$

Equations (28). There are wrong signs and complex conjugations. The correct equations read as

$$g^{00}=0, \quad g^{01}=\Omega^2, \quad g^{0+}=g^{0-}=0, \tag{8}$$

$$g^{11}=\Omega^2(\partial^*(W-T_1^1)+W^*(W-T_1^1))-g^{+-}T_-^{*1}-g^{++}T_+^{*1}-g^{+1}T_1^{*1}, \tag{9}$$

$$g^{1+}=\Omega^2(W-T_1^1), \quad g^{1-}=\Omega^2(W^*-T_1^{*1}), \tag{10}$$

$$g^{++}=-\Omega^2\lambda_1, \quad g^{+-}=-\Omega^2, \quad g^{--}=-\Omega^2\lambda_1^*. \tag{11}$$

[1] S. Frittelli, C. N. Kozameh, and E. T. Newman, ''Lorentzian metrics from characteristic surfaces,'' J. Math. Phys. **36**, 4975–4983 (1995).

[a)] Electronic mail: simo@mayu.physics.duq.edu

Erratum: "GR via characteristic surfaces" [J. Math. Phys. 36, 4984–5004 (1995)]

Simonetta Frittelli[a)]
Department of Physics, Duquesne University, Pittsburgh, Pennsylvania 15282

Carlos N. Kozameh
FaMAF, Universidad Nacional de Cordoba, Ciudad Universitaria, 5000 Córdoba, Argentina

Ezra T. Newman
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 14 September 1998; accepted for publication 15 September 1998)

[S0022-2488(99)00702-1]

We list a series of corrections to equations appearing in Ref. 1. None of the following affect the content of the article. On p. 4991, the expression given for g^{11} in Eq. (2.22) contains two typos. The correct expression should read as

$$g^{11}=-2g^{01}+\bar{\eth}g^{1+}-g^{ab}\ \eth Z_{,a}\ \eth\bar{\Lambda}_{,b}. \tag{1}$$

On p. 4993, in the text of the *Remark*, one undisplayed equation is wrong. In the paper it reads as $\boldsymbol{\nabla}_b G=-4\kappa\boldsymbol{\nabla}_a(T^a{}_b-\frac{1}{4}\delta^a{}_b T)$. The correct equation should read as

$$\boldsymbol{\nabla}_b G=\kappa\ \boldsymbol{\nabla}_b T. \tag{2}$$

On p. 4996, there are incorrect signs in the expression of the vector ℓ_a. The correct expression should read as

$$\ell_a=\frac{1}{\sqrt{2}(1+\zeta\zeta)}((1+\zeta\bar{\zeta}),-(\zeta+\bar{\zeta}),i(\zeta-\bar{\zeta}),(1-\zeta\bar{\zeta})). \tag{3}$$

On p. 4998, the expression for the metric component g^{11} in Eqs. (A1) contains a typo. The correct expression reads as

$$g^{11}=-2-\tfrac{1}{2}\bar{\eth}^2\Lambda_{,1}+\eth\bar{\Lambda}_{,-}=-2-\tfrac{1}{2}\eth^2\bar{\Lambda}_{,1}+\bar{\eth}\Lambda_{,+}. \tag{4}$$

On p. 5000, the expression for Q given by (A8) is incorrect by an overall minus sign. The correct expression reads as

$$Q=\frac{1}{4q}D\bar{\Lambda}_1 D\Lambda_1+\frac{3}{8q^2}(Dq)^2-\frac{1}{4q}D^2 q. \tag{5}$$

In Appendix B, there is a typo on the T_1^i given by Eq. (B4). The correct expression for the T_i^1 reads as

$$qT_i^1=\{\Lambda_{,i}\bar{\Lambda}_{,+}+\eth\bar{\Lambda}_{,i}+\bar{\Lambda}_-\delta_i^1+\bar{\Lambda}_{,0}\delta_i^+-2\delta_i^-\}\Lambda_{,1}+\bar{\eth}\Lambda_{,i}+\Lambda_{,-}\bar{\Lambda}_{,i}+\Lambda_{,0}\delta_i^-+\Lambda_+\delta_i^1-2\delta_i^+. \tag{6}$$

[a)] Electronic mail: simo@mayu.physics.duq.edu

[1] S. Frittelli, C. M. Kozameh, and E. T. Newman, "GR via characteristic surfaces," J. Math. Phys. **36**, 4984–5004 (1995).

Erratum: ''Linearized Einstein theory via null surfaces'' [J. Math. Phys. 36, 5005–5022 (1995)]

Simonetta Frittelli[a)]
Department of Physics, Duquesne University, Pittsburgh, Pennsylvania 15282

Carlos N. Kozameh
FaMAF, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

Ezra T. Newman
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 14 September 1998; accepted for publication 15 September 1998)

[S0022-2488(99)00802-6]

We list a series of corrections to equations appearing in Ref. 1. None of the following affect the content of the article. On p. 5007 and 5018, Eq. (8) contains a sign mistake. The correct equation (8) should read as

$$\nu' = \nu + \tfrac{1}{4}\partial^a \xi_a . \tag{1}$$

On p. 5011, the expressions for the metric components f^{11} and f_{00} in Eqs. (24) and (25) contain a typo. The correct expressions read as

$$f^{11} = -\tfrac{1}{2}\bar{\eth}^2\Lambda_1 + \eth\overline{\Lambda_-} = -\tfrac{1}{2}\eth^2\overline{\Lambda_1} + \bar{\eth}\Lambda_+ , \tag{2}$$

$$f_{00} = -\tfrac{1}{2}\bar{\eth}^2\Lambda_1 + \eth\overline{\Lambda_-} = -\tfrac{1}{2}\eth^2\overline{\Lambda_1} + \bar{\eth}\Lambda_+ . \tag{3}$$

On p. 5015, in the text of the 18th line, one undisplayed expression is wrong. In the paper it reads as *Using (27) to eliminate* $\eth\Lambda_-$ *on the left* . . . . The correct expression should read as *Using (27) to eliminate* $\eth\Lambda_1$ *on the left* . . .

On p. 5016, a sequence of four displayed equations contain mistaken factors. The correct equation (48) should read as

$$-\frac{1}{4}\eth^3\overline{\Lambda_1} = \bar{\eth}\sigma,_0 + 3\bar{\eth}\int_\infty^R \partial_-\ \eth\nu\ d\tilde{R} - 3\eth\nu - \frac{3}{2}\,\eth^2\ \bar{\eth}\nu. \tag{4}$$

The displayed equation immediately below (48) should read as

$$\bar{\eth}^2\Lambda_0 = -\tfrac{1}{4}\,\bar{\eth}\ \eth^3\overline{\Lambda_1} - \tfrac{1}{4}\,\eth\bar{\eth}^3\Lambda_1 + 2\bar{\eth}\ \eth\nu + \eth\bar{\eth}\eth\ \bar{\eth}\nu. \tag{5}$$

The correct equation (49) should read as

$$\bar{\eth}^2\Lambda_0 = \bar{\eth}^2\sigma,_0 + \eth^2\overline{\sigma},_0 + 3\bar{\eth}^2\int_\infty^R \partial_-\ \eth\nu\ d\tilde{R} + 3\eth^2\int_\infty^R \partial_+\ \bar{\eth}\nu\ d\tilde{R} - 4\bar{\eth}\ \eth\nu - 2\bar{\eth}\eth^2\ \bar{\eth}\nu. \tag{6}$$

Finally, the correct lightcone cut equation (50) should read as

$$\bar{\eth}^2\ \eth^2 Z = \eth^2\sigma + \eth^2\overline{\sigma} + \int^u \left(3\bar{\eth}^2\int_\infty^R \partial_-\ \eth\nu\ d\tilde{R} + 3\eth^2\int_\infty^R \partial_+\bar{\eth}\nu\ d\tilde{R} - 4\bar{\eth}\eth\nu\ \ - 2\bar{\eth}\eth^2\ \bar{\eth}\nu \right) d\tilde{u}. \tag{7}$$

[1] S. Frittelli, C. N. Kozameh, and E. T. Newman, ''Linearized Einstein theory via null surfaces,'' J. Math. Phys. **36**, 5005–5022 (1995).

[a)] Electronic mail: simo@mayu.physics.duq.edu

Analytical solution of the relativistic Coulomb problem with a hard core interaction for a one-dimensional spinless Salpeter equation

F. Brau[a)]
Université de Mons-Hainaut, Place du Parc 20, B-7000 Mons, Belgium

(Received 25 March 1998; accepted for publication 11 December 1998)

In this paper, we construct an analytical solution of the one-dimensional spinless Salpeter equation with a Coulomb potential supplemented by a hard core interaction, which keeps the particle in the x positive region. © *1999 American Institute of Physics.* [S0022-2488(99)02003-4]

I. INTRODUCTION

A simple relativistic version of the Schrödinger equation is the Spinless Salpeter Equation (SSE). For the one-dimensional case we have

$$\sqrt{-d_x^2+m^2}\Psi(x)=(E-V(x))\Psi(x), \tag{1}$$

where m is the mass of the particle, $V(x)$ is the potential interaction, E the eigenenergy of the stationary state $\Psi(x), d_x^2=d^2/dx^2=-p^2$ and p is the relative momentum of the particle ($\hbar=c=1$). p and x are conjugate variables. The differential operator of the Schrödinger equation is well defined because it is a second derivative. To solve a physical problem, we must just solve an ordinary eigenvalue differential equation. The situation is more complicated with the SSE because the associated differential operator is a nonlocal one. Its action cannot be calculated directly from its operator form. Indeed, its action on a function $f(x)$ is known only if $f(x)$ is an eigenfunction of the operator d_x^2. In this case we obtain

$$\sqrt{-d_x^2+m^2}f(x)=\sqrt{-\alpha+m^2}f(x), \tag{2}$$

where α is the corresponding eigenvalue of d_x^2. That is why we need first to rewrite the SSE into a form easier to handle. Since the operator is a nonlocal one, this form could be an integral equation. This have been done for the three-dimensional case in Refs. 1, 2. We present the one-dimensional corresponding form in the next section. With the method used to obtain this form, it is possible to rewrite the SSE as an integro-differential equation (see Ref. 2 for the three-dimensional case). But the kernel is really complicated and the resulting equation seems to be very difficult to treat. We will use, here, another method to obtain the solution of the equation.

To solve the relativistic Coulomb problem we do not solve any differential equation. We calculate the action of the square-root operator on the functions $x^n e^{-\beta x}$. Because the result is analytical and because the wave functions of the Coulomb problem with a hard core interaction are an exponential multiplied by a polynomial, which is also the form of the Schrödinger and Klein–Gordon wave functions, a complete solution of Eq. (1), with $V(x)\propto 1/x$ and $x>0$, can be found.

The paper is organized as follow. In Sec. II, we give some useful mathematical results concerning the square-root operator. In Sec. III, we solve the one-dimensional Coulomb problem

[a)]Electronic mail: fabian.brau@umh.ac.be

with a hard core interaction. In Sec. IV, we compare our results to those obtained with the Schrödinger equation[3–6] and with the Klein–Gordon equation.[7,8] At last, we give our conclusion in Sec. V.

II. MATHEMATICAL FRAMEWORK

In this section we give some results concerning the square-root operator which we use to solve the Coulomb problem supplemented by a hard core interaction.

A. Integral representation of the square-root operator

To obtain the integral representation of the square-root operator we use the Fourier transform of the one-dimensional delta function. We have

$$\sqrt{-d_x^2+m^2}\Psi(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}dp\,dq\,\sqrt{p^2+m^2}e^{-i(q-x)p}\Psi(q). \tag{3}$$

Extracting the operator $-d_x^2+m^2$ and integrating over the momentum p (see Ref. 1) we obtain

$$\begin{aligned}\sqrt{-d_x^2+m^2}\Psi(x)&=\frac{1}{\pi}(-d_x^2+m^2)\int_{-\infty}^{+\infty}dq\,K_0(m|q-x|)\Psi(q)\\&=\frac{1}{\pi}(-d_x^2+m^2)\int_0^{+\infty}dq\,K_0(mq)[\Psi(x+q)+\Psi(x-q)],\end{aligned} \tag{4}$$

where $K_0(x)$ is the modified Bessel function of order 0 (Ref. 9, p. 952).

B. Invariant space functions of the one-dimensional square-root operator

In this section we calculate the action of the square-root operator on the functions $x^n e^{-\beta x}$. We obtain that the result is equal to a polynomial of order n, $M_n(m,\beta,x)$, multiplied by the same exponential. Thus, the space of functions $P_n(x)e^{-\beta x}$ is the invariant space functions of this operator. Using formula (4) we have

$$\sqrt{-d_x^2+m^2}x^n e^{-\beta x}=\frac{1}{\pi}(-d_x^2+m^2)e^{-\beta x}\int_0^{+\infty}dq\,K_0(mq)[(x+q)^n e^{-\beta q}+(x-q)^n e^{\beta q}]. \tag{5}$$

This leads to (Ref. 9, p. 712)

$$\sqrt{-d_x^2+m^2}x^n e^{-\beta x}=\frac{1}{\sqrt{\pi}}(-d_x^2+m^2)e^{-\beta x}\sum_{k=0}^{n}\binom{n}{k}G_k(m,\beta)x^{n-k}. \tag{6}$$

The coefficients $G_k(m,\beta)$ are given by

$$G_k(m,\beta)=\frac{\Gamma(k+1)^2}{\Gamma(k+3/2)}\left(\frac{1}{(m+\beta)^{k+1}}F\left(k+1,1/2;k+3/2;-\frac{m-\beta}{m+\beta}\right)+(-)^k(\beta\to-\beta)\right), \tag{7}$$

where $F(\alpha,\beta;\gamma;x)$ is the hypergeometric function (Ref. 9, p. 1039). Performing the derivation in Eq. (6) and rearranging the obtained relation, we have

$$\begin{aligned}\sqrt{-d_x^2+m^2}x^n e^{-\beta x}=\frac{1}{\sqrt{\pi}}e^{-\beta x}\sum_{k=0}^{n}\binom{n}{k}\{&(m^2-\beta^2)G_k(m,\beta)\\&+2\beta k G_{k-1}(m,\beta)-k(k-1)G_{k-2}(m,\beta)\}x^{n-k}.\end{aligned} \tag{8}$$

It is possible to write the coefficients $G_k(m,\beta)$ into a more useful form. This form will allow us to find a recursion relation between the coefficients $G_k(m,\beta)$ and to simplify the expression (8). We will be able to construct the polynomial, $M_n(m,\beta,x)$, for each value of n. We have the relation (Ref. 10, p. 562)

$$F(a,1/2;a+1/2;-x)=\Gamma(a+1/2)\,\frac{x^{(1-2a)/4}}{\sqrt{1+x}}\,P_{-1/2}^{1/2-a}\left(\frac{1-x}{1+x}\right), \tag{9}$$

with $x>0$. The functions $P_\nu^\mu(x)$ are the associated Legendre functions for x real and $|x|<1$. With this relation, we find that

$$G_{k-1}(m,\beta)=\frac{\Gamma(k)^2}{\sqrt{2m}(m^2-\beta^2)^{(2k-1)/4}}[P_{-1/2}^{1/2-k}(\beta/m)+(-)^{k-1}P_{-1/2}^{1/2-k}(-\beta/m)]. \tag{10}$$

Now, using the recursion relation of the associated Legendre functions (Ref. 9, p. 1005),

$$P_\nu^{\mu+2}(x)=-2(\mu+1)\,\frac{x}{\sqrt{1-x^2}}\,P_\nu^{\mu+1}(x)+(\mu-\nu)(\mu+\nu+1)P_\nu^\mu(x), \tag{11}$$

and the explicit expression of $P_{-1/2}^{-1/2}(x)$ and $P_{-1/2}^{1/2}(x)$ (Ref. 9, p. 1008), we can write the following relations:

$$G_{k+2}(m,\beta)=\frac{1}{m^2-\beta^2}[(k+1)^2G_k(m,\beta)-(2k+3)\beta G_{k+1}(m,\beta)], \tag{12}$$

with

$$G_0(m,\beta)=\sqrt{\frac{\pi}{m^2-\beta^2}}, \tag{13}$$

$$G_1(m,\beta)=-\frac{\sqrt{\pi}\beta}{(m^2-\beta^2)^{3/2}}. \tag{14}$$

At last, one can find, using Eq. (12), that the general coefficient of the sum of Eq. (8) becomes

$$F_{k,n}(m,\beta)=\frac{1}{\sqrt{\pi}}\binom{n}{k}[\beta G_{k-1}(m,\beta)-(k-1)G_{k-2}(m,\beta)], \quad \text{with } k\geqslant 1. \tag{15}$$

And with this form, a recursion relation for $F_{k,n}(m,\beta)$ can be easily found. Thus, to conclude this section, we are able now to rewrite Eq. (8) into a simple form:

$$\sqrt{-d_x^2+m^2}x^n e^{-\beta x}=M_n(m,\beta,x)e^{-\beta x}=\left[\sum_{k=0}^{n}F_{k,n}(m,\beta)x^{n-k}\right]e^{-\beta x}, \tag{16}$$

$$F_{0,n}(m,\beta)=\sqrt{m^2-\beta^2}, \tag{17}$$

$$F_{1,n}(m,\beta)=\frac{n\beta}{\sqrt{m^2-\beta^2}}, \tag{18}$$

$$F_{k+2,n}(m,\beta)=\frac{n-k-1}{(k+2)(m^2-\beta^2)}[(k-1)(n-k)F_{k,n}(m,\beta)-(2k+1)\beta F_{k+1,n}(m,\beta)], \tag{19}$$

$$F_{k,n+1}=\frac{n+1}{n+1-k}F_{k,n}. \tag{20}$$

We can see that we obtain the expected relation [from Eq. (2)] for $n=0$. And thus we see that we must have $\beta<m$. With the relations (17)–(20) the polynomial $M_n(m,\beta,x)$ is completely defined and we can construct it for each value of n. This result will allow us to find, with few calculations, the solution of the one-dimensional relativistic Coulomb problem with a hard core interaction. We give below the polynomials, as an example, for $n=0\rightarrow 4$,

$$M_0(m,\beta,x)=S, \tag{21}$$

$$M_1(m,\beta,x)=Sx+\frac{\beta}{S}, \tag{22}$$

$$M_2(m,\beta,x)=Sx^2+\frac{2\beta}{S}x-\frac{m^2}{S^3}, \tag{23}$$

$$M_3(m,\beta,x)=Sx^3+\frac{3\beta}{S}x^2-\frac{3m^2}{S^3}x+\frac{3m^2\beta}{S^5}, \tag{24}$$

$$M_4(m,\beta,x)=Sx^4+\frac{4\beta}{S}x^3-\frac{6m^2}{S^3}x^2+\frac{12m^2\beta}{S^5}x-\frac{3m^2}{S^7}(m^2+4\beta^2), \tag{25}$$

with

$$S=\sqrt{m^2-\beta^2}. \tag{26}$$

Note that these last relations can be simply checked by acting the square-root operator on each side of Eq. (16). For $n=1$, we see that we have an identity if we use the relation for $n=0$. Now, knowing these two relations we see that the relation for $n=2$ is also an identity, and so on for each value of n.

III. THE ONE-DIMENSIONAL RELATIVISTIC COULOMB PROBLEM WITH A HARD CORE INTERACTION

The equation to solve is

$$\sqrt{-d_x^2+m^2}\Psi(x)=\left(E+\frac{\kappa}{x}\right)\Psi(x). \tag{27}$$

We just consider here the case $x>0$ (we will discuss after the extension to the whole x axis). Physically this means that we have a hard core interaction for $x\leqslant 0$. Then the wave functions will possess the following asymptotic behavior: $\Psi(x)=0$ for $x\leqslant 0$ and for $x=+\infty$. Suppose that the wave functions have the following form:

$$\begin{aligned}\Psi(x)&\propto\sum_{k=1}^{n}\gamma_{k,n}x^k e^{-\beta x}, \quad \text{for} \quad x>0 \quad \text{and} \quad n=1,2,...,\\ \Psi(x)&=0 \quad \text{for} \quad x\leqslant 0.\end{aligned} \tag{28}$$

We do not consider the normalization of the functions here. Thus, replacing Eq. (28) into Eq. (27) and using Eq. (16), we obtain

$$\sum_{k=1}^{n} \gamma_{k,n} \sum_{p=0}^{k} F_{p,k}(m,\beta) x^{k-p} = E \sum_{k=1}^{n} \gamma_{k,n} x^k + \kappa \sum_{k=1}^{n} \gamma_{k,n} x^{k-1}. \tag{29}$$

Now equaling order by order we will determine the solution. The term of order n gives

$$E = F_{0,n}(m,\beta) = \sqrt{m^2 - \beta^2}. \tag{30}$$

From the term of order $n-1$, we have

$$\kappa = F_{1,n}(m,\beta), \tag{31}$$

which leads to

$$\beta = \frac{\kappa m}{n\sqrt{1+(\kappa/n)^2}}. \tag{32}$$

We can remark that we have as well the necessary relation $\beta < m$. We are now already able to determine the energy spectrum. Using Eq. (30) and Eq. (32) we have

$$E = \frac{m}{\sqrt{1+(\kappa/n)^2}}. \tag{33}$$

To obtain a complete solution, we must now find all the $\gamma_{k,n}$, and prove that the system of equations which gives these quantities always has a solution. Obviously we can fix $\gamma_{n,n}=1$. We see that the term of order $n-j$ determines the coefficient $\gamma_{n-j+1,n}$ if the previous $\gamma_{k,n}$ are known. Beginning with the term of order $n-2$, we obtain directly $\gamma_{n-1,n}$. And now we can get $\gamma_{n-2,n}$ from the term of order $n-3$. The independent term will fix the last factor $\gamma_{1,n}$. Thus, we have a triangular system of $n-1$ algebraic equations with $n-1$ unknowns. This system will always possess a solution if the determinant of the coefficient matrix is non-null. As this is a triangular matrix, the determinant is the product of the diagonal elements. The expression of these elements is $\kappa - F_{1,n-j}(m,\beta)$ which is equal to $j\beta/S$. These quantities are always non-null since $j>0$. The general form of $\gamma_{k,n}$ is obtained from the term of order $n-j-1$. We have

$$\gamma_{n-j,n} = \frac{S}{j\beta} \sum_{k=0}^{j-1} \gamma_{n-k,n} F_{j-k+1,n-k}(m,\beta). \tag{34}$$

We can inverse the summation to finally obtain

$$\gamma_{n-j,n} = \sum_{p_1=0}^{j-1} \sum_{p_2=0}^{p_1-1} \cdots \sum_{p_j=0}^{p_{j-1}-1} \tilde{F}(n,p_1,j)\tilde{F}(n,p_2,p_1)\ldots\tilde{F}(n,p_j,p_{j-1}), \tag{35}$$

with

$$\tilde{F}(n,k,j) = \frac{S}{j\beta} F_{j-k+1,n-k}(m,\beta). \tag{36}$$

For the summation in Eq. (35), we must use the following rule: If in a summation over p_α, α being arbitrary, the bound $p_{\alpha-1}-1$ is negative, all the $\tilde{F}(n,k,j)$ containing the indices $p_{\beta \geqslant \alpha}$ are equal to 1. With the formula (35), we are able to construct the wave functions for the Coulomb problem with a hard core interaction. As an example we give the three first wave functions:

$$\Psi(x) \propto x Q_n(m,\kappa,x) e^{-\beta x}, \tag{37}$$

with β given by Eq. (32), and

$$Q_1(m,\kappa,x)=1, \tag{38}$$

$$Q_2(m,\kappa,x)=x-\frac{m^2}{S^2\beta}, \tag{39}$$

$$Q_3(m,\kappa,x)=x^2-\frac{3m^2}{S^2\beta}x+\frac{3m^2}{2\beta^2S^4}(\beta^2+m^2), \tag{40}$$

with S defined by Eq. (26). Again, we can perform a simple verification by putting these solutions into Eq. (27) and using Eq. (16).

Contrary to the Schrödinger or Klein–Gordon equation, the extension of the solution to the whole x axis is really more complicated. We can try to use $\exp(-\beta|x|)$ instead of $\exp(-\beta x)$ in our solution. But the situation is quite more difficult. Indeed, the construction of the solution was based on the fact that $\exp(-\beta x)$ was an eigenfunction of the square-root operator and that the invariant space functions of this operator was $P_n(x)\exp(-\beta x)$, where $P_n(x)$ is a polynomial of order n. But it is easy to show, with Eq. (4), that

$$\sqrt{-d_x^2+m^2}\exp(-m|x|)=\frac{2m}{\pi}K_0(m|x|). \tag{41}$$

This is non-null, as this is the case in the Eq. (16). Thus, $\exp(-\beta|x|)$ is not an eigenfunction of the square-root operator and $P_n(x)\exp(-\beta|x|)$ is not an invariant space function of this operator. So it seems that the pure Coulomb problem has quite different solutions for the wave functions and certainly for the spectrum.

IV. DISCUSSION

The one-dimensional Coulomb problem has been treated by many authors, both nonrelativistically[3–6] and relativistically.[7,8] But in these works the whole x axis is considered. As a consequence, the ground state gives some difficulties.

In the nonrelativistic case the solution is

$$\Psi(x)=x\exp(-\kappa m|x|/n)L_{n-1}^{(1)}(2\kappa m|x|/n), \tag{42}$$

$$E=m-\frac{m\kappa^2}{2n^2}, \quad \text{with } n=1,2,\ldots. \tag{43}$$

But we see that for $n=1$, the wave function has a node at the origin. So this is not the wave function for the ground state. In fact, it is found to be infinitely bounded and the wave function is a delta function.[3,8]

In the Klein–Gordon case the solution is

$$\Psi(x)=x^S\exp(-\beta|x|/2)L_{n-1}^{(\gamma)}(\beta|x|), \tag{44}$$

$$E=m\Big/\sqrt{1+\frac{\kappa^2}{(n-1+S)^2}}, \tag{45}$$

with

$$\beta=2m\kappa/\sqrt{(n-1+S)^2+\kappa^2}, \quad \text{with } n=1,2,\ldots, \tag{46}$$

and

$$S=\tfrac{1}{2}(1+\gamma)=\tfrac{1}{2}(1\pm\sqrt{1-4\kappa^2}). \tag{47}$$

Thus, we see that we have two distinct solutions according the sign for S. Actually, the spectrum with the minus sign for S is not acceptable. Indeed, a reason is that, for $n=1$, when we perform the limit $\kappa\to 0$, we obtain $E=0$. This means that the particle is still bounded when the interaction vanishes. Thus, the problem for the ground state persists (see Ref. 8 for a complete discussion).

In this paper we do not consider the whole x axis and we have no problem with the ground state. We consider a hard core interaction, for $x\leqslant 0$, which gives $\Psi(x)=0$ in this region. Thus, $x\exp(-\beta x)$ is the wave function for the ground state. Actually the purpose of these work was to solve a particular kind of differential equations with a difficult to handle nonlocal operator. Indeed, any analytical solutions are known for the spinless Salpeter equation. Thus, we do not discuss the problem of the ground state of the one-dimensional Coulomb problem. In Sec. III, we have shown that the extension to the whole x axis is not easy. Moreover, the ground state problem could persist.

To compare our result to the results of previous works, we can consider the Schrödinger and the Klein–Gordon equation for the Coulomb potential supplemented by a hard core interaction. The spectra and the wave functions remain unchanged but the ground state problem has disappeared. In the three cases we have the same kind of wave functions: an exponential (with different arguments) multiplied by a polynomial (with different coefficients). For the spectrum we have, in the limit of small κ,

$$E_{\text{Sch}}=m\left(1-\frac{\kappa^2}{2n^2}\right), \tag{48}$$

$$E_{\text{KG}}=m\left(1-\frac{\kappa^2}{2n^2}-\frac{\kappa^4}{n^3}+\frac{3\kappa^4}{8n^4}\right), \tag{49}$$

$$E_{\text{Sal}}=m\left(1-\frac{\kappa^2}{2n^2}+\frac{3\kappa^4}{8n^4}\right). \tag{50}$$

Thus we see that in the expansion of the Salpeter spectrum the term in n^3 is missing compared to the Klein–Gordon spectrum. So the difference between these two spectra is rather important. For an electron in an electromagnetic Coulomb potential, the splitting is about 10^{-3} eV.

Another characteristic of the spinless Salpeter spectrum is that κ can grow up without limit. This could come from the fact that we have another kinetic operator than in the Klein–Gordon equation and that the result could be quite different. But the main explanation is certainly that we do not solve the real Coulomb problem and that this spectrum could be different contrary to the Klein–Gordon equation, which keeps the same spectrum in both cases. Indeed, for the SSE, there exists a limit value for κ in three dimensions.[11]

V. CONCLUSION

The purpose of this work was to find an analytical solution of a particular kind of differential equations containing a nonlocal differential operator. The equation considered in this paper, the one-dimensional spinless Salpeter equation (SSE), is a simple relativistic version of the one-dimensional Schrödinger equation. The SSE is not a marginal equation. For three dimensions, this equation comes from the Bethe–Salpeter equation (Refs. 12, 13, p. 297), which gives the correct description of bound states of two particles. Moreover, despite the presence of a so particular operator, the SSE is often used in the potential models (see, for instance, Refs. 14–20), which give a phenomenological description of hadrons.

To find this analytical solution, we calculate, in Sec. II B, the action of the square-root operator on a polynomial multiplied by an exponential and we show that this constitutes the invariant space functions of this operator. To be able to perform this calculation, we have constructed, in Sec. II A, an integral representation of the square-root operator. In Sec. III, we have obtained, without solving any differential equation, a complete solution of the SSE with a Coulomb potential and a hard core interaction. This last interaction is introduced to keep the particle in the *x* positive region. We remark that the SSE wave functions have the same form than the Schrödinger and the Klein–Gordon wave functions. We remark also that the splitting between the SSE and the Klein–Gordon spectrum is rather important. Indeed, it is of the same order of the first relativistic correction given by these two equations.

ACKNOWLEDGMENTS

We thank Dr. C. Semay, Professor Y. Brihaye, Professor J. Nuyts, and Professor F. Michel for useful discussions.

[1] L. J. Nickisch, L. Durand, and B. Durand, Phys. Rev. **30**, 660 (1984).
[2] F. Brau, J. Math. Phys. **39**, 2254 (1998).
[3] R. Loudon, Am. J. Phys. **27**, 649 (1959).
[4] M. Andrews, Am. J. Phys. **34**, 1194 (1966).
[5] L. K. Haines and D. H. Roberts, Am. J. Phys. **37**, 1145 (1969).
[6] J. F. Gomes and A. H. Zimmerman, Am. J. Phys. **48**, 579 (1980).
[7] H. N. Spector and J. Lee, Am. J. Phys. **53**, 248 (1985).
[8] R. E. Moss, Am. J. Phys. **55**, 397 (1987).
[9] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, corrected and enlarged edition (Academic, New York, 1980).
[10] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970).
[11] I. W. Herbst, Commun. Math. Phys. **53**, 285 (1977); **55**, 316 (1977) (addendum).
[12] E. E. Salpeter and H. A. Bethe, Phys. Rev. **84**, 1232 (1951).
[13] W. Greiner and J. Reinhardt, *Quantum Electrodynamics*, 2nd ed. (Springer-Verlag, Berlin, 1994).
[14] D. P. Stanley and D. Robson, Phys. Rev. D **21**, 3180 (1980).
[15] P. Cea, G. Nardulli, and G. Paiano, Phys. Rev. D **28**, 2291 (1983).
[16] S. N. Gupta, S. F. Radford, and W. W. Repko, Phys. Rev. D **31**, 160 (1985).
[17] S. Godfrey and N. Isgur, Phys. Rev. D **32**, 189 (1985), and references therein.
[18] D. LaCourse and M. G. Olson, Phys. Rev. D **39**, 2751 (1989), and references therein.
[19] L. P. Fulcher, Phys. Rev. D **50**, 447 (1994), and references therein.
[20] C. Semay and B. Silvestre-Brac, Nucl. Phys. A **618**, 455 (1997), and references therein.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

A new scaling property of the Casimir energy for a piecewise uniform string

I. Brevik[a)]
Division of Applied Mechanics, Norwegian University of Science and Technology, N-7034 Trondheim, Norway

E. Elizalde[b)]
CSIC/IEEC, Edifici Nexus 201, Gran Capità 2-4, E-08034 Barcelona, Spain and Department of ECM and IFAE, Faculty of Physics, Barcelona University, Diagonal 647, E-08028 Barcelona, Spain

R. Sollie[c)]
IKU, Sintef Group, N-7034 Trondheim, Norway

J. B. Aarseth[d)]
Division of Applied Mechanics, Norwegian University of Science and Technology, N-7034 Trondheim, Norway

(Received 28 July 1998; accepted for publication 11 November 1998)

An unexpected and very accurate scaling invariance of the Casimir energy of the piecewise uniform relativistic string is pointed out. The string consists of $2N$ pieces of equal length, of alternating type I/type II material, endowed with different tensions and mass densities but adjusted such that the velocity of transverse sound equals c. If $E_N(x)$ denotes the Casimir energy as a function of the tension ratio $x = T_{\mathrm{I}}/T_{\mathrm{II}}$, it turns out that the ratio $f_N(x)=E_N(x)/E_N(0)$, which lies between zero and one, will be practically *independent* of N for integers $N \geqslant 2$. Physical implications of this scaling invariance are discussed. Finite temperature theory is also considered. © *1999 American Institute of Physics.* [S0022-2488(99)02902-3]

I. INTRODUCTION

The purpose of the present paper is to discuss a rather unexpected scaling property of the Casimir energy for a piecewise uniform, relativistic string executing planar oscillations in its own plane. Both the zero temperature and the finite temperature theory will be considered.

First, some background material about the system: The string is taken to consist of $2N$ pieces of equal length, of alternating type I and type II material, and it is relativistic in the sense that the velocity of transverse sound is equal to the velocity of light. Figure 1 shows a sketch of the string in the case when $N=6$ (for clarity the thickness of the string is drawn finite in the figure). The center of mass of the string lies at rest. The total length of the string is L.

The piecewise uniform string model in its simplest version was introduced by Brevik and Nielsen in 1990,[1] and the model has since then been analyzed from various points of view.[2–10] The particular case of a $2N$-piece string was considered in Refs. 4, 7, and 10. In Ref. 7, a general formula for the Casimir energy $E_N(x)$ was derived, x meaning the ratio between the tensions T_{I} and T_{II} in the two kinds of material,

$$x=\frac{T_{\mathrm{I}}}{T_{\mathrm{II}}}. \tag{1}$$

[a)]Electronic mail: Iver.H.Brevik@mtf.ntnu.no
[b)]Electronic mail: eli@zeta.ecm.ub.es, elizalde@ieec.fcr.es
[c)]Electronic mail: Roger.Sollie@iku.sintef.no
[d)]Electronic mail: Jan.B.Aarseth@mtf.ntnu.no

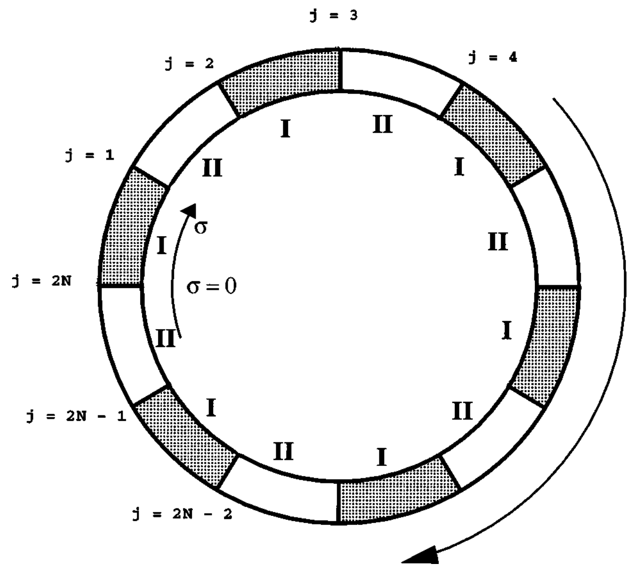

FIG. 1. Sketch of the composite $2N$ string, when $N=6$.

The formula reads, at zero temperature,

$$E_N(x)=\frac{N}{2\pi L}\int_0^\infty \ln\left|\frac{2(1-\alpha^2)^N-[\lambda_+^N(iq)+\lambda_-^N(iq)]}{4\sinh^2(Nq/2)}\right|dq. \tag{2}$$

Here α is defined as

$$\alpha=\frac{1-x}{1+x} \tag{3}$$

and $\lambda_\pm$ are the eigenvalues, for complex arguments iq, of the dispersion equation. Explicitly,

$$\lambda_\pm(iq)=\cosh q-\alpha^2\pm[(\cosh q-\alpha^2)^2-(1-\alpha^2)^2]^{1/2}. \tag{4}$$

We are now able to explain more closely the scaling property mentioned above: the Casimir energy, which is defined as the zero-point energy for the composite string minus the zero-point energy for the uniform string, turns out to be always nonpositive. And for a given value of the integer N, $E_N(x)$ becomes more negative the more the ratio x deviates from unity. [Note that the case $x=1$ corresponds to a uniform string, so that $E_N(1)=0$.]

Since the theory is invariant under the substitution $x\rightarrow 1/x$, we can limit ourselves to the case $x\leqslant 1$. The minimum energy for a given value of N is thus $E_N(0)$. In Ref. 7 the exact result in this case was calculated to be

$$E_N(0)=-\frac{\pi}{6L}(N^2-1). \tag{5}$$

Figure 2, which is equivalent to Fig. 4 in Ref. 7, shows how the ratio $f_N(x)$, defined as

$$f_N(x)=\frac{E_N(x)}{E_N(0)}, \tag{6}$$

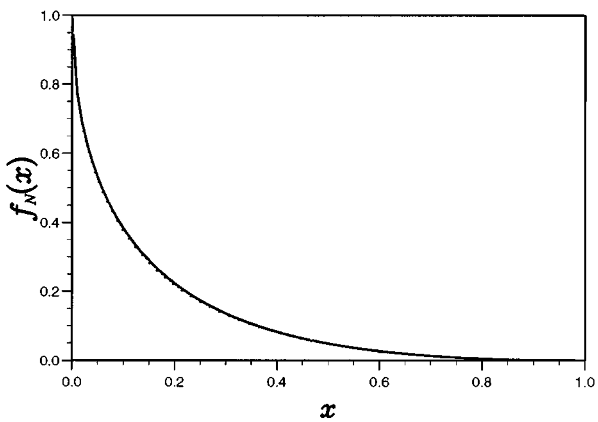

FIG. 2. The Casimir energy $f_N(x)=E_N(x)/E_N(0)$ versus tension ratio $x=T_{\rm I}/T_{\rm II}$, for $N=2,3,4,...,10$. The various curves collapse in practice into a single curve.

varies with x, for some different values of N $(N=2,3,4,...,10)$. It is seen that the various curves collapse in practice into a single curve. This is the scaling property. It means that for any x and $N\geqslant 2$ the Casimir energy, relative to the minimum obtainable energy, is to a very high accuracy independent of N.

This result—which is so far a property following from the numerics—is rather unexpected. It would be desirable to examine the issue more closely. This is the theme of the present paper. The problem is two-faceted. On the mathematical side one may ask: can we give at least a partial derivation of the scaling invariance in analytical terms? Moreover, can we give effective analytical expressions which reproduce the scaling invariance with satisfactory accuracy? And, on the physical side: what is the physical interpretation of the scaling invariance? Will it be possible to gain at least a partial physical understanding of *why* the scaling invariance exists? These issues are not merely of academic interest. The composite string model as such is quite an attractive model of two-dimensional quantum field theory in general. Moreover, the model may be of interest in connection with cosmology also, in string theories of the early universe.

II. POWER EXPANSIONS: PRELIMINARY REMARKS

From a mathematical viewpoint the simplest way of approach is to make some kind of power expansion of the integrand in Eq. (2). Numerically, it turns out that the scaling property becomes more pronounced the larger the value of N is. When N is large, obviously $e^{-Nq}\ll 1$ except in the limit $q\rightarrow 0$. It accordingly becomes natural, as a first approach, to adopt

$$z=e^{-q} \tag{7}$$

as an expansion parameter, keeping terms of order z^{N-1} while neglecting terms of order z^N and higher.

As a simple illustration, let us put $N=3$, and expand $\lambda_{\pm}$ up to relative order z^2. From Eq. (4) we have

$$\lambda_+=z^{-1}[1-2\alpha^2 z+\alpha^2(2-\alpha^2)z^2], \tag{8}$$

$$\lambda_-=(1-\alpha^2)^2 z[1+2\alpha^2 z+\alpha^2(5\alpha^2-2)z^2] \tag{9}$$

[it is useful to note that $\lambda_+\lambda_-=(1-\alpha^2)^2$]. That means, when expanding the integrand in Eq. (2) to order z^2 we can neglect λ_-^N. We get

$$\ln\left|\frac{2(1-\alpha^2)^N-[\lambda_+^N(iq)+\lambda_-^N(iq)]}{4\sinh^2(Nq/2)}\right|=-2N\alpha^2 z\left[1-\left(1-\frac{3}{2}\alpha^2\right)z+\mathcal{O}(z^2)\right]. \tag{10}$$

Using this expansion in Eq. (2) and replacing $E_N(0)$ with $-\pi N^2/(6L)$ we get as a first approximation

$$f_N(x)\simeq\frac{3\alpha^2}{\pi^2}\left(1+\frac{3}{2}\alpha^2\right). \tag{11}$$

This simple formula is actually quite useful. While derived for the case $N=3$, it may tentatively be applied for higher values of N also. This assumes, of course, that terms of order z^3 can be neglected also when $N>3$. We cannot expect high accuracy in this way, but note from Eq. (11) its most important property: it is *independent* of N. That is, we trace already on the present simple level of investigation the mentioned scaling invariance.

Consider some numerics of Eq. (11): for $x=1$ (uniform string) the formula gives $f_N(0)=1$ as it should, and for $x=0.5$, for instance, the formula gives $f_N(0.5)\simeq 0.039\,40$ instead of the ''exact'' result $f_3(0.5)=0.047\,72$ calculated numerically from Eqs. (2) and (5) with $N=3$. In the limit $x\to 0(\alpha\to 1)$, Eq. (11) yields $f_N(0)\simeq 0.760$ instead of $f_N(0)=1$ as it should. As expected it is the limit $x\to 0$ which is critical here; it is in this case that the inaccuracy of the formula (11) is largest.

It is to be observed that the case $N=1$ is an exceptional case: for a relativistic string composed of two pieces of equal length the Casimir energy is *zero*, irrespective of the value of x:[1]

$$E_1(x)=0. \tag{12}$$

One may imagine now to continue the expansion technique in z further, with the aim to increase the accuracy. We have actually done this, by means of the Mathematica and the Maple analytic programs, and have managed to expand the integrand of Eq. (2) up to the 12th order in z, within a reasonable computer time. The resulting expressions are complicated and will not be given here. It turns out, in fact, at least to moderate orders in z, that insufficient accuracy will be obtained in this way. For instance, if we choose $N=7$ and expand the integrand of Eq. (11) up to order $z^{N-1}=z^6$, we find that the power expansion for $f_7(x)$ yields the result 2.806 (instead of 1) in the limit $x\to 0$. That is, the accuracy actually becomes poorer than in the simple case considered above. Some reflection shows that this kind of behavior should not be so unexpected after all: our expansion procedure which employs $z=e^{-q}$ as the smallness parameter is good for large values of q, but becomes poorer in the region of small q. And it is precisely in the last-mentioned region that the magnitude of the integral of Eq. (2) is largest.

Although the accuracy can be improved by going to very high orders in z, it becomes clear that the present simple expansion procedure of the integrand of Eq. (2) in powers of z is not the most economic way of handling the integral. In the next section we shall turn to another mathematical method, which is numerically very accurate and which, moreover, helps us to construct a simple, *analytic*, expression for $f_N(x)$.

III. MORE ACCURATE NUMERICS. ANALYTICAL REPRESENTATIONS

It is both numerically and analytically advantageous to solve the integral (2) by an accurate numerical method, without relying simply on ''black box'' routines. To obtain a better control over the integrand in the whole integration domain, we shall first use the parameter $z=e^{-q}$ to express $f_N(x)$ (for $N\geqslant 2$) as an integral over z from 0 to 1:

$$f_N(x)=\int_0^1 I_N(x,z)\,dz. \tag{13}$$

Here

$$I_N(x,z)=\frac{-3N}{\pi^2(N^2-1)z}\ln\left|\frac{2(1-\alpha^2)^N z^N-[(z\lambda_+)^N+(z\lambda_-)^N]2^{-N}}{(1-z^N)^2}\right|. \tag{14}$$

It turns out that analytic solutions are, in fact, achievable, for small values of N and some special values of x. We give two examples, both evaluated by means of Maple: the first is for $x=3-2\sqrt{2}$ $(\alpha=1/\sqrt{2})$, the second is for $x=\frac{1}{3}$ $(\alpha=\frac{1}{2})$. The exact results are

$$f_2(3-2\sqrt{2})=\tfrac{1}{4},\quad f_2(\tfrac{1}{3})=\tfrac{1}{9},\quad f_4(3-2\sqrt{2})=\tfrac{23}{90}$$

(note that these cases correspond to a four-piece, and an eight-piece, string).

In looking for analytic solutions of (14), we may use the fact that $(z-1)^2$ is a factor both in the numerator and the denominator. In the denominator we get: $(z^N-1)^2/(z-1)^2=(z^{N-1}+z^{N-2}+\cdots+z+1)^2$ giving the following expression for the integral of the denominator:

$$\frac{3N}{\pi^2(N^2-1)}\int_0^1\frac{\ln(z^{N-1}+z^{N-2}+\cdots+z+1)^2}{z}\,dz=\frac{1}{N+1}. \tag{15}$$

We are then left with integrating the numerator of (14) divided by $(z-1)^2$. Note that the numerator simplifies considerably for $\alpha=1/\sqrt{N}$. We give two examples: first for $N=3$ and $\alpha=1/\sqrt{3}$ which means that $x=2-\sqrt{3}$, the second example is for $N=5$ and $\alpha=1/\sqrt{5}$ which means that $x=(3-\sqrt{5})/2$. The exact results are

$$f_3(2-\sqrt{3})=\tfrac{5}{32},$$

$$f_5\left(\frac{1}{2}(3-\sqrt{5})\right)=\frac{1}{8}\left[3+\frac{5}{\pi^2}\left(2\arctan\left(\frac{4}{3}\right)-2\pi\right)\arctan\left(\frac{1}{2}\right)-2\Re\,\mathrm{dilog}\left(\frac{-3+4i}{5}\right)\right]=0.0\,897\,261.$$

Consider now the numerics. It is important to get control over the integrand in Eq. (13) at the points $z=0$ and $z=1$. At first sight, one might conclude from the expression (14) that these points are singular points. However, a closer scrutiny shows that they are regular. Now, numerical integration routines are, in general, classified into closed and open routines, where the closed routines make use of the endpoints (here $z=0$ and 1) while the open routines approach these points but do not use them explicitly. In our case it turned out that even modern open FORTRAN routines (such as DQAGS from QUADPACK) ran into problems when approaching $z=0$ and $z=1$. By means of analytic Taylor expansions around these points we managed, however, to find the following exact expressions at $z=0$ and 1:

$$I_N(x,0)=\frac{6N^2\alpha^2}{\pi^2(N^2-1)}, \tag{16}$$

$$I_N(x,1)=-\frac{3N}{\pi^2(N+1)}\ln(1-\alpha^2). \tag{17}$$

In the critical limit $x\to 0(\alpha\to 1)$ it is seen that $I_N(x,1)$ diverges logarithmically; this is nevertheless compatible with the integrand itself being finite, $f_N(x\to 0)=1$. Moreover, the limit $N\to\infty$ is seen to correspond to the simple expressions

$$I_\infty(x,0)=\frac{6\alpha^2}{\pi^2},\quad I_\infty(x,1)=-\frac{3}{\pi^2}\ln(1-\alpha^2). \tag{18}$$

TABLE I. Numerical results for $f_N(x)$ versus x, for some representative values of N.

x	$N=2$	$N=10$	$N=100$
0.02	0.6742	0.6833	0.6837
0.10	0.3721	0.3814	0.3818
0.20	0.2158	0.2224	0.2227
0.40	0.0795	0.0823	0.0824
0.80	0.00 502	0.00 521	0.00 522
0.90	0.00 112	0.00 117	0.00 117

Armed with the expressions (16) and (17) at the endpoints it becomes easy now to calculate $f_N(x)$. Our calculations were done with the FORTRAN routine DQAGS (double precision). It turned out to be possible to continue the calculations up to very high N values (about 4000), which for practical purposes is equivalent to infinity.

Table I shows some results for $f_N(x)$ versus x, calculated for $N=2$, 10, and 100. Higher values of N, up to $N=4000$, gave the same answers as $N=100$, to the accuracy shown. The scaling invariance is seen to be so accurate that a usual graphical representation of $f_N(x)$, such as Fig. 2, is unable to distinguish between the various curves. As a further check of the numerical calculation, we wrote a separate program in Maple exploiting the possibility that this language offers to set $N=\infty$ directly. The values for $f_\infty(x)$ calculated in this way were compared with those calculated above for $N=4000$, and were found to agree to a very high accuracy (to 7, or 8, digits).

It is now of interest to ask: is it possible to represent the numerical results calculated above with analytic formulas, to a satisfactory accuracy? The answer turns out to be yes, and the formulas become actually surprisingly simple. Let us start from the ansatz

$$f_N(x) \to f(x) = (1-x^b)^c, \tag{19}$$

where b and c are constants. The expression (19) implies that the magnitudes of the logarithms of $f(x)$ and $(1-x^b)$ are predicted to be proportional. Figure 3 shows the outcome of one of our tests: the ordinate $|\ln f(x)|$, which is calculated numerically, is compared with the abscissa $|\ln(1-x^b)|$ for

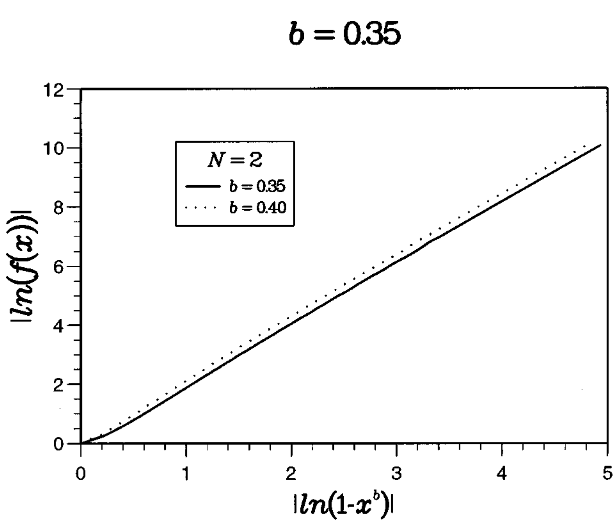

FIG. 3. Dependence of $\ln|f(x)|$, calculated numerically, upon $|\ln(1-x^b)|$ when $b=0.35$ and 0.40. The approximative linearity is evident.

$b=0.35$ and 0.40. The suggested linearity of the curve is seen to be verified quite accurately, although there is a weak change in slope around $x=0.45$. Numerical trials show that the simple form

$$f(x)=(1-\sqrt{x})^{5/2} \tag{20}$$

is a useful approximation when $0<x<0.45$. This form is intended to hold for all values of N.

We can analyze this behavior one step further, by considering the particular case $N=2$. Then we have from Eq. (13)

$$f_2(x)=-\frac{2}{\pi^2}\int_0^1 \ln\left|\frac{(1+z)^2-4\alpha^2 z}{(1+z)^2}\right|\frac{dz}{z}. \tag{21}$$

After some trials using Maple, we arrived at the following series expansion in x:

$$f_2(x)=1-\frac{8}{\pi}\sqrt{x}+\frac{16}{\pi^2}x+\frac{8}{3\pi}x^{3/2}-\frac{32}{3\pi^2}x^2-\frac{8}{5\pi}x^{5/2}+\frac{368}{45\pi^2}x^3$$
$$+\frac{8}{7\pi}x^{7/2}-\frac{704}{105\pi^2}x^4-\frac{8}{9\pi}x^{9/2}+\frac{9008}{1575\pi^2}x^5+\frac{8}{11\pi}x^{11/2}+\cdots. \tag{22}$$

This type of series is called a Puisseux series in mathematics. It is analogous to a Taylor series and is more accurate the smaller the value of x is. For $x=0.2$ the error is 0.015%; for $x=0.3$ it is 0.27%. For $x\geqslant 0.35$ the error becomes larger than 1%. Let us expand Eq. (22) around $x=0$:

$$f_2(x)=1-2.55\sqrt{x}+1.62\ x, \tag{23}$$

and similarly expand the empirical Eq. (20):

$$f_2(x)=1-\tfrac{5}{2}\sqrt{x}+\tfrac{15}{8}x\simeq 1-2.5\sqrt{x}+1.875\ x. \tag{24}$$

That is, we see from here the mathematical reason behind the structure of Eq. (20).

IV. ON THE PHYSICAL INTERPRETATION

The obvious question is now: what is the physical interpretation of the calculated scaling invariance?

Let us first remind ourselves about the main features of the string model: it is relativistic in the sense that the transverse sound velocity equals c; it consists of $2N$ pieces of alternating type I/II material but of equal length; and it is described by small amplitude wave theory since the condition that the transverse elastic force be continuous at the junctions is expressed as $T\partial\psi/\partial\sigma$ $=$continuous (ψ denotes the displacement, σ the length coordinate).

Assume now that there exists such a string somewhere in the universe (it is most natural to think about the early universe), and assume that it is possible to make use of zero temperature theory in the quantum mechanical sense. The Casimir energy of the string is caused entirely by its *inhomogeneity*. There are two factors contributing to this: the tension ratio x, and the integer N. It is natural to expect that a string, originally starting out with the Casimir energy $E_N(x)(<0)$, wishes to make this energy as low as possible. This can be done while maintaining N constant, if the tension ratio is made extreme, $x\to 0$. The string's energy becomes then lowered to $E_N(0)$ $=f^{-1}(x)E_N(x)=-(\pi/6L)(N^2-1)$, where the scaling invariance is being accounted for. Finally, the string may divide itself into a large number of pieces, implying a quadratic negative divergence when $N\to\infty$, $E_N(0)\to-(\pi/6L)N^2$.

The finite temperature case is more difficult to interpret (see Appendix), all the time that the Casimir energy in the case of extreme tension ratio is diverging, $E_N^T(0)\to-\infty$ for $N>1$, due to the classical contribution. It will not be further commented upon here.

Inhomogeneous strings of the type considered here have not been observed in nature. As mentioned, its most probable application seems to lie within cosmology. Perhaps particle physics can be a possible application also: it is rather striking to observe how closely the above relativistic theory is formally related to the electrodynamic (or gluo-dynamic) theory of a field propagating in a medium whose (color) permitivity ϵ and permeability μ satisfy the condition $\epsilon\mu=1$. See, for instance, Lee's model of the exterior hadronic vacuum.[11]

ACKNOWLEDGMENTS

This investigation has been supported by the Norwegian–Spanish ministerial scientific interchange program, by CIRIT (Generalitat de Catalunya), and by DGICYT (MEC, Spain), Project No. PB96-0925.

APPENDIX A: THE CASE OF FINITE TEMPERATURES

The formulation of the theory when the string is situated in a thermal bath of temperature T is done by replacing the integral over imaginary frequencies $\xi=qN/L$ by a sum over discrete Matsubara frequencies, $\xi\to\xi_n=2\pi nk_BT$, with n an integer and k_B Boltzmann's constant. Referring to Eq. (2) we make the replacement $q\to q_n=L\xi_n/N$. The general substitution to be made in finite temperature quantum field theory is

$$\int_0^\infty d\xi\to 2\pi k_BT\sum_{n=0}^{\infty}{}', \tag{A1}$$

where the prime means that the term $n=0$ is taken with half weight. The finite temperature Casimir energy becomes[7]

$$E_N^T(x)=k_BT\sum_{n=0}^{\infty}{}'\ln\left|\frac{2(1-\alpha^2)^N-[\lambda_+^N(iq_n)+\lambda_-^N(iq_n)]}{4\sinh^2(Nq_n/2)}\right|. \tag{A2}$$

This formal expression holds for all integers N.

In the special case $N=1$ we get, since

$$\lambda_+(iq_n)-\lambda_-(iq_n)=2(\cosh q_n-\alpha^2), \tag{A3}$$

that

$$E_1^T(x)=0 \tag{A4}$$

for all x. Thus, a two-piece string composed of two equal parts will also, for finite temperature, have zero Casimir energy. The behavior is just as for $T=0$; cf. Eq. (12).

Of particular interest is the limit $x\to 0$, i.e., the critical case corresponding to minimum Casimir energy. According to Eq. (1.4) we then have

$$\lambda_+(iq_n)\to 4\sinh^2\left(\frac{q_n}{2}\right),\quad \lambda_-(iq_n)\to 0, \tag{A5}$$

so that from Eq. (A2) we get

$$E_N^T(0)=2k_BT\sum_{n=0}^{\infty}{}'\ln\left|\frac{2^N\sinh^N(q_n/2)}{2\sinh(Nq_n/2)}\right|. \tag{A6}$$

For $N=1$ we recover the result (A4), but for $N>1$ the expression (A6) actually *diverges*. This result would not be easy to anticipate beforehand. The divergence is caused by the $n=0$ term, i.e., by the lowest of all Matsubara frequencies. Low Matsubara frequencies are generally associated with the classical limit and are not related to quantum mechanics. It is noticeable that it is just this limit that is responsible for the divergence of (A6).

Some further insight can be obtained if we separate out the $n=0$ contribution to the general expression (A2). To this end we treat q_n as a small analytic variable, and expand

$$\cosh q_n-\alpha^2=(1-\alpha^2)\left[1+\frac{\frac{1}{2}q_n^2}{1-\alpha^2}+\mathcal{O}(q_n^4)\right], \tag{A7}$$

implying

$$|2(1-\alpha^2)^N-\lambda_+^N(iq_n)-\lambda_-^N(iq_n)|=(1-\alpha^2)^{N-1}N^2q_n^2 \tag{A8}$$

to the leading order. From Eq. (A2) we then get the expression

$$E_N^T(x)=k_\mathrm{B}T\left[\frac{N-1}{2}\ln(1-\alpha^2)+\sum_{n=1}^{\infty}\ln\left|\frac{2(1-\alpha^2)^N-\lambda_+^N(iq_n)-\lambda_-^N(iq_n)}{4\sinh^2(Nq_n/2)}\right|\right], \tag{A9}$$

in which the classical, $n=0$, contribution is explicit in the first term. If $N=1$ and x arbitrary, we recover the result (A4). If $x\to 0$ and $N>1$, we obtain $E_N^T(0)\to-\infty$, in accordance with Eq. (A6).

It is helpful to compare the expressions (A9) and (17). In both cases, the term $\ln(1-\alpha^2)$ appears. Expression (17) holds for $z=1$, i.e., $q=-\ln z=0$, which means zero frequency. When $x\to 0$ the expression is diverging. As we have seen, this divergence does not lead to an infinite energy $E_N(0)$ because the divergence is suppressed by the *integration* over q in Eq. (2). In the finite temperature case the integration is replaced by a Matsubara *sum* in which the zero temperature term will have to appear explicit. This illustrates why $E_N^T(0)$ is diverging, while $E_N(0)$ stays finite.

Generally speaking, the precise meaning of the "high temperature" limit ought to be made clear. There are *two* characteristic frequencies in this system, viz. a thermal frequency $\omega_T=k_\mathrm{B}T$, and a geometric frequency which may be defined as $\omega_\mathrm{geom}=1/L$. The high-temperature limit should be defined by the inequality $\omega_T/\omega_\mathrm{geom}\gg 1$, which implies $Nq_n=L\xi_n\gg 1$ for all $N\geq 1$. In this case the classical $n=0$ term gives the dominant contribution to the Casimir energy, and we obtain from (A9)

$$E_N^{T\to\infty}(x)=\frac{N-1}{2}k_\mathrm{B}T\ln(1-\alpha^2). \tag{A10}$$

We mention finally for completeness that we have in this Appendix considered the Casimir energy itself, not the fractional energy relative to the case $x\to 0$. To construct a fractional quantity $f_N^T(x)$, analogously to Eq. (6), would be meaningless since $E_N^T(0)$ is infinite.

[1] I. Brevik and H. B. Nielsen, Phys. Rev. D **41**, 1185 (1990).
[2] X. Li, X. Shi, and J. Zhang, Phys. Rev. D **44**, 560 (1991).
[3] I. Brevik and E. Elizalde, Phys. Rev. D **49**, 5319 (1994).
[4] I. Brevik and H. B. Nielsen, Phys. Rev. D **51**, 1869 (1995).
[5] I. Brevik, H. B. Nielsen, and S. D. Odintsov, Phys. Rev. D **53**, 3224 (1996).
[6] S. S. Bayin, J. P. Krisch, and M. Ozcan, J. Math. Phys. **37**, 3662 (1997).
[7] I. Brevik and R. Sollie, J. Math. Phys. **38**, 2774 (1997).
[8] M. H. Berntsen, I. Brevik, and S. D. Odintsov, Ann. Phys. (N.Y.) **257**, 84 (1997).
[9] J. Lu and B. Huang, Phys. Rev. D **57**, 5280 (1998).
[10] I. Brevik, A. A. Bytsenko, and H. B. Nielsen, hep-th/9703089.
[11] T. D. Lee, Particle *Physics and Introduction to Field Theory* (Harwood Academic, New York, 1988).

Dyon-Skyrmion lumps

Y. Brihaye
Physique-Mathematique, Universite de Mons-Hainaut, Mons, Belgium

B. Kleihaus
Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland

D. H. Tchrakian
Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland and School of Theoretical Physics-DIAS, 10 Burlington Road, Dublin 4, Ireland

(Received 28 September 1998; accepted for publication 27 October 1998)

We make a numerical study of the classical solutions of the combined system consisting of the Georgi–Glashow model and the SO(3) gauged Skyrme model. Both monopole-Skyrmion and dyon-Skyrmion solutions are found. A new bifurcation is shown to occur in the gauged Skyrmion solution sector. © *1999 American Institute of Physics.* [S0022-2488(99)04302-9]

I. INTRODUCTION

Monopole-Skyrmion solitons can be important in a semiclassical study of the Callan–Rubakov[1,2] mechanism for catalyzing proton decay. Indeed, the first such approach was made by Callan and Witten[3] where the U(1) gauged Skyrmion was coupled to the electromagnetic field of a monopole. Here, we study the classical soliton solutions of a model which incorporates the full non-Abelian monopole field and the corresponding gauged Skyrmion, described by an SO(3) gauged Skyrme model interacting with the Georgi–Glashow (GG) model through the gauge field. The solutions we study include both monopole-Skyrmions and dyon-Skyrmions.

There are two $3+1$ dimensional SO(3) gauge field models which support static soliton solutions. One is the GG model which supports the well-known monopole,[4] and the other is the $SU(2)_{L+R}$, or vector, gauged Skyrme model[5,6] which also supports SO(3) gauged Skyrmions. In addition to the monopole, the GG model supports also dyon solutions[7] which in addition to the magnetic charge carry an electric charge as well. The topological stability of the monopole comes from the magnetic charge, which is descended from the second Chern–Pontryagin charge, while the topological charge of the gauged Skyrmion is the degree of the map.

Combining these two models, we have a new system whose topological charge is a sum of the respective charges, and it can reasonably be expected that this system also supports static finite energy solitons. Note that in this case the local SO(3) symmetry is broken down to U(1) via the Higgs mechanism, in contrast to the SO(3) gauged Skyrme model on its own, in which case the local SO(3) symmetry is not broken at all and three massless gauge bosons survive. In this preliminary investigation, this is precisely what we have done. Using numerical methods, we verify that such solutions exist. Moreover, we have sought and found both monopole-Skyrmion and dyon-Skyrmion solutions, and studied some of their properties. The combined system supports solutions also with zero monopole charge, unit baryon charge, as well as with unit monopole charge, zero baryon charge.

Even though this is a self-contained numerical study of the classical solutions alluded to above, it is in order to put it into context both in the background of previous work involving the gauging of the Skyrme model,[8] and, from the viewpoint of its potential physical relevance.

The Skyrme model was gauged by Witten in Ref. 9, and others, e.g., in Ref. 10. These works were carried out in the context of current algebra results, and were not concerned with the solitonic aspects of the gauged Skyrmion. That was done subsequently by many authors, see, e.g., Ref. 11, where gauged Skyrme solitons were studied with the aim of explaining the low energy properties of Hadrons. Also in the context of electroweak theory, which can be regarded as a gauged Skyrme model in the limit of very high Higgs mass, Rubakov[12] and Eilam *et al.*[13] considered the static classical solutions of the SU_L gauged Skyrme model. In all these cases, there is no topological lower bound and the classical solutions are metastable, but for certain values of the parameters in one of these models[12,13] a stable branch of solitons appears as a result of catastrophic behavior. The advantage of the gauging used in Refs. 12 and 13 is that the four divergence of the topological current does not vanish but equals the local chiral anomaly,[14,15] which can present itself as a mechanism for Baryon number violation as explained in Ref. 12.

In the context of Baryon number violation, there is an older mechanism suggested by Rubakov[1] and by Callan[2] where monopole-(left-handed massless) fermion interactions lead to fermion number nonconservation. The mechanism involves the fluctuations of the electric field, in the presence of the magnetic field of the monopole, giving rise to nonzero chiral anomaly and hence fermion number violation. This was shown for the case of massless (left-handed) fermions, by scattering with the monopole, which describes a high energy process. The approximation techniques employed[1,2] are neither perturbation theoretic nor semiclassical. To describe a low energy process such as a decay, it would be more appropriate to deal with a process that is susceptible to semiclassical analysis. To this end, Callan and Witten[3] replaced the massless fermions by the Skyrme soliton,[8] interacting with the (Abelian) magnetic field of the monopole. While they[3] did not seek to demonstrate the existence of a U(1) gauged Skyrmion, this is implicit in their work and has recently been verified numerically.[16] In the background of this it is hoped that the present work, which sets out to find the monopole-Skyrmion and dyon-Skyrmion solutions, would be of concrete usefulness to a semiclassical method of describing baryon number decay. In particular the dyon-Skyrmion excites a nonzero classical quantity for the chiral anomaly, which can lead to chirality breaking as pointed out long ago by Marciano and Pagels.[17]

In Sec. II we present the model and give the topological lower bounds on the static energy. In Sec. III we give the static spherically symmetric fields and the field equations in the static limit. Sections IV and V deal, respectively, with the results of the numerical analysis of the $\mathbf{A}_0=0$ and $\mathbf{A}_0\neq 0$ cases. Section V in particular, includes an in-depth analysis of the Julia-Zee dyon.[7] In Secs. IV and V, we also give an account of the SO(3) gauged Skyrmion studied previously in Ref. 6, because these solutions play a certain technical role in the construction of dyon-Skyrmion solutions in Sec. V. We summarize and discuss our results in Sec. VI.

II. THE MODEL

The model under consideration is the combination of the Georgi–Glashow (GG) model and of the SO(3) gauged O(4) (Skyrme) model studied previously in Refs. 5 and 6. We state the Lagrangian of each of these models separately, defined in four-dimensional Minkowski space, each being normalized properly so that the value of the energy of the static soliton in each case lies above its own topological lower bound. The static solutions in question satisfy the Euler–Lagrange equations of the static energy density functional, which is the static Hamiltonian in the temporal gauge. In the GG case, this is the 'tHooft–Polyakov[4] monopole, while in the latter case it is the soliton studied in Refs. 5 and 6.

The GG model is described by

$$\mathcal{L}_{\mathrm{GG}}=-\tfrac{1}{4}\lambda_0^4|F_{\mu\nu}^{\alpha}|^2+\tfrac{1}{2}\lambda_1^4|D_\mu\Phi^{\alpha}|^2-\tfrac{1}{4}\lambda_2^4(\eta^2-|\Phi^{\alpha}|^2)^2, \tag{1}$$

$$F_{\mu\nu}^{\alpha}=\partial_\mu A_\nu^{\alpha}-\partial_\nu A_\mu^{\alpha}+\epsilon^{\alpha\beta\gamma}A_\mu^{\beta}A_\nu^{\gamma},\quad D_\mu\Phi^{\alpha}=\partial_\mu\Phi^{\alpha}+\epsilon^{\alpha\beta\gamma}A_\mu^{\beta}\Phi^{\gamma}. \tag{2}$$

The late Greek indices μ, ν,... label the Minkowski space vectors, while the early Greek indices α, β, $\ldots=1,2,3$ label the elements of the algebra of the gauge group SO(3). The Latin letters a, b, $\ldots=1,2,3,4$ so that $a=(\alpha,4)$ are reserved for the O(4) Skyrme model. In Eq. (1) the constant η is the vacuum expectation value (VEV) of the Higgs field and like the latter has the inverse dimension of a length. The constants λ_0, λ_1, and λ_2 are all dimensionless.

The SO(3) gauged Skyrme model is described by

$$\mathcal{L}_{O(4)} = -\tfrac{1}{4}\kappa_0^4 |F_{\mu\nu}^{\alpha}|^2 + \tfrac{1}{2}\kappa_1^2 |D_\mu \phi^a|^2 - \tfrac{1}{8}\kappa_2^4 |D_{[\mu}\phi^a D_{\nu]}\phi^b|^2, \tag{3}$$

$$D_\mu \phi^\alpha = \partial_\mu \phi^\alpha + \epsilon^{\alpha\beta\gamma} A_\mu^\beta \phi^\gamma, \quad D_\mu \phi^4 = \partial_\mu \phi^4, \tag{4}$$

with $\phi^\alpha\phi^\alpha + \phi^4\phi^4 = 1$.

The constants κ_0 and κ_2 are dimensionless and the constant κ_1 has dimension of an inverse length. The reason we keep all the coupling constants arbitrary in Eqs. (1) and (3) will appear soon.

When we consider the static Hamiltonians corresponding to the Lagrangians above in the temporal gauge, i.e., $A_0 = 0$, we can write the following topological identities:

$$\int d\mathbf{r}\, \mathcal{H}_{GG} \geqslant 4\pi\eta\lambda_0^2\lambda_1^2 M, \tag{5}$$

$$\mathcal{H}_{GG} = \tfrac{1}{4}\lambda_0^4 |F_{jk}^{\alpha}|^2 + \tfrac{1}{2}\lambda_1^4 |D_j \Phi^\alpha|^2 + \tfrac{1}{4}\lambda_2^4 (\eta^2 - |\Phi^\alpha|^2)^2, \tag{6}$$

where the integer M, representing the index of the mapping $\Phi^\alpha(\vec{x})$, is the monopole topological charge. Similarly[6]

$$\int d\mathbf{r}\, \mathcal{H}_{O(4)} \geqslant 12\pi^2 \kappa_1 \kappa_2^2 \frac{1}{2\sqrt{1 + 9\left(\frac{\kappa_2}{\kappa_0}\right)^4}} T, \tag{7}$$

$$\mathcal{H}_{O(4)} = \tfrac{1}{4}\kappa_0^4 |F_{jk}^{\alpha}|^2 + \tfrac{1}{2}\kappa_1^2 |D_j \phi^a|^2 + \tfrac{1}{8}\kappa_2^4 |D_{[j}\phi^a D_{k]}\phi^b|^2, \tag{8}$$

where the integer T, representing the index of the mapping $\phi^a(\vec{x})$, is the Skyrmion topological charge. In the Skyrme description of hadrons, T is identified with the baryon number.

In the following we will consider also the equations resulting from the superposition of the two Lagrangians Eqs. (1) and (3)

$$\mathcal{L}_m = \mathcal{L}_{GG} + \mathcal{L}_{O(4)}. \tag{9}$$

The finite energy configurations of this mixed Lagrangian are characterized by the couple of integers M, T. Classical solutions corresponding to the two different topological excitations can then be constructed, they correspond to the configuration with minimal energy in a class M, T.

In order to normalize the fields conventionally, we have to choose

$$\lambda_0^2 = \frac{1}{e}\cos(\theta), \quad \kappa_0^2 = \frac{1}{e}\sin(\theta), \quad \lambda_1^4 = 1, \tag{10}$$

where e denotes the gauge coupling constant. With the choice $\theta = \pi/4$, the topological inequality relating $\mathcal{H}_m$ to the class of solutions of indexes M, T reads

$$\int d\mathbf{r}\, \mathcal{H}_m \geqslant \frac{4\pi\eta}{e}\left(\frac{1}{\sqrt{2}}M + \frac{3\pi}{2}\sqrt{\frac{\xi\kappa}{1+18\kappa}}T\right), \tag{11}$$

where

$$\lambda = \frac{\lambda_2^4}{e^2}, \quad \xi = \frac{1}{\eta^2}\kappa_1^2, \quad \kappa = e^2\kappa_2^4. \tag{12}$$

In Eq. (12), λ is defined for later convenience. Note that the topological lower bound Eq. (11) can be refined by an optimal value of the mixing angle θ, depending on the parameters λ_1, λ_2, κ_1, κ_2. To achieve this it is necessary to solve a complicated nonlinear equation, which we shall not pursue here.

III. STATIC SPHERICALLY SYMMETRIC EQUATIONS

The classical equations corresponding to Eqs. (1), (3), and (9) are in general intractable. We will restrict our search of solutions to the static and spherically symmetric case. If we choose to employ the temporal gauge in the static limit, the Euler–Lagrange equations will reduce to the variational equations arising from the static Hamiltonians pertaining to the Lagrangians Eqs. (1) and (3). The latter would be bounded from below by the monopole charge and the baryon number densities, respectively. Hence the solutions to the classical equations of each of these static Hamiltonians, separately, can describe the 'tHooft–Polyakov monopole[4] and the soliton of the SO(3) gauged Skyrme model.[5,6] The Euler–Lagrange equations of the Hamiltonian of the combined static system, i.e., GG-Skyrme, in the temporal gauge also supports soliton solutions since the Hamiltonian is again bounded from below by the two topological charges Eq. (11). This is one of the problems studied in the present work yielding the monopole-Skyrmion solitons.

If instead of employing the temporal gauge we proceed like Julia and Zee[7] and solve the Euler–Lagrange equations pertaining to the Lagrangian Eq. (9) defined on Minkowski space in the static limit, the resulting solutions of the GG-Skyrme system describe the dyon-Skyrmion. This is the other problem studied in this work. As in the case of the dyon[7] on its own, we shall restrict ourselves to the spherically symmetric solutions only. (In this case the classical equations simplify sufficiently to become tractable. To our knowledge the only dyon solutions known are the spherically symmetric Julia–Zee[7] dyons.)

The spherically symmetric ansatz employed is

$$A_i^\alpha=\frac{a(r)-1}{r}\,\epsilon_{i\alpha\beta}\hat{x}^\beta, \quad A_0^\alpha=e\,\eta g(r)\hat{x}^\alpha, \tag{13}$$

$$\Phi^\alpha=\eta h(r)\hat{x}^\alpha, \tag{14}$$

$$\phi^\alpha=\sin f(r)\hat{x}^\alpha, \quad \phi^4=\cos f(r). \tag{15}$$

Notice that the functions $a(r)$, $h(r)$, $g(r)$, and $f(r)$ are dimensionless. We find it useful to introduce a dimensionless radial variable

$$x=M_W r, \quad M_W\equiv e\,\eta. \tag{16}$$

Substituting the ansatz Eqs. (13)–(15) into the static limit of the Lagrangian Eq. (9), leads to the following one-dimensional (radial) Lagrangian density L_m, defined by

$$\int L_m\,dx=\int \mathcal{L}_m\,d\mathbf{r}=E_1-E_2 \tag{17}$$

with

$$E_p\equiv\frac{4\pi}{e}\,\eta\tilde{E}_p=\frac{4\pi}{e}\,\eta\int dx\,\mathcal{E}_p, \quad p=1,2, \tag{18}$$

$$\mathcal{E}_1=\tfrac{1}{2}x^2(g')^2+a^2g^2, \tag{19}$$

$$\mathcal{E}_2=(a')^2+\frac{(a^2-1)^2}{2x^2}+\frac{1}{2}x^2(h')^2+a^2h^2+\frac{\lambda}{4}x^2(h^2-1)^2$$
$$+\frac{\xi}{2}[x^2(f')^2+2a^2\sin^2 f]+\kappa a^2\sin^2 f\left[(f')^2+a^2\frac{\sin^2 f}{2x^2}\right], \tag{20}$$

where we have separated the contribution E_1 due to the electric field and the prime denotes the derivative with respect to x. The total energy is given by

$$E=\frac{4\pi}{e}\eta\tilde{E}=\frac{4\pi}{e}\eta(\tilde{E}_1+\tilde{E}_2). \tag{21}$$

The static classical equations corresponding to the Lagrangian density $\mathcal{L}_m$, in the spherically symmetric ansatz, turn out to be equivalent to the equations obtained by varying the effective one-dimensional density [see Eqs. (19) and (20)] $\mathcal{E}_1-\mathcal{E}_2$ with respect to the radial functions a, g, h, and f. These equations are obtained straightforwardly and we do not list them here. We note however, that for each function the corresponding variational equation can be solved trivially by setting this function to zero.

It will be useful to present their asymptotic forms in the $x\gg 1$ region, to facilitate subsequent explanations. They are, in order of the variations of a, g, h, and f:

$$a''=a\left(\frac{a^2-1}{x^2}+h^2-g^2+\xi\sin^2 f+\cdots\right), \tag{22}$$

$$(x^2g')'=2ga^2, \tag{23}$$

$$(x^2h')'=h(2a^2+\lambda x^2(h^2-1)), \tag{24}$$

$$(x^2f')'=2a^2\sin f\cos f+o(\kappa/\xi). \tag{25}$$

[Note that Eqs. (23), (24) are exact.]

Following Ref. 7, we define the energy of a configuration by $E=E_1+E_2$, which coincides with the volume integral of the static Hamiltonian obtained in the usual way from the gauge invariant stress tensor. The topological lower bound for E_2 follows immediately from Eqs. (5), (7), and (11).

IV. NUMERICAL RESULTS, CASE $A_0=0$

We first discuss the classical solutions in absence of the electric field, i.e., with $g(x)=0$. Equation (23) is trivially solved and we are left with a system of three nonlinear differential equations. Only the part E_2 of the action is relevant in this case. In the following, we will conveniently denote the value $\tilde{E}_2$ of the solution with given M and T by

$$E_{MT}(\lambda,\xi,\kappa). \tag{26}$$

We now describe the four cases with $M\leq 1$ and $T\leq 1$, namely $(M=0,\ T=0)$, $(M=1,\ T=0)$, $(M=0,\ T=1)$, and $(M=1,\ T=1)$. All but the third of these, namely that characterized by the topological charge $(M=0,\ T=1)$, are solutions of Eqs. (22), (24), and (25), arising from the variation of Eq. (20). The third one on the other hand is described by the solutions of the SO(3) gauged Skyrme model,[5,6] which are described by Eqs. (22) and (25) arising from the variation of the functional (20) with the Higgs function $h(r)$ and the coupling of the Higgs potential λ both set equal to zero. The reason for including this field configuration in Sec. IV C below is that it will become useful for the construction of some solutions in Sec. V, and, because we have given an enhanced numerical study of it.

TABLE I. The energies of the monopole, the monopole-Skyrmion, and the gauged Skyrmion for several values of the Higgs coupling constant λ.

λ	Monopole	Monopole-Skyrmion ($\xi=1,\ \kappa=0.4$)	Gauged Skyrmion ($\xi=1,\ \kappa=0.4$)
0.0	1.000	3.450	2.98
0.05	1.106	3.470	⋯
0.10	1.138	3.480	⋯
0.20	1.180	3.490	⋯
0.40	1.220	3.510	⋯
0.60	1.250	3.520	⋯
0.80	1.270	3.530	⋯
1.00	1.290	3.536	⋯

A. Case $M=0$, $T=0$

This corresponds to the class of the vacuum which is not spherically symmetric. It has a zero energy

$$E_{00}(\lambda,\xi,\kappa)=0. \tag{27}$$

B. Case $M=1$, $T=0$

This case corresponds to the celebrated SU(2) magnetic monopole.[4] Since $T=0$, it has $f(r)=0$; as a consequence, the parameters ξ and κ are irrelevant for this case. The boundary conditions and asymptotic behavior of the functions a,h read

$$a(0)=1, \quad h(0)=0, \tag{28}$$

$$a(x)\simeq Ae^{-x}, \quad h(x)\simeq 1-Be^{-\sqrt{2\lambda}x} \quad (x\to\infty), \tag{29}$$

where A,B,F are constants. The values of the energy of the monopole solution were computed long ago[18] (our numerics fully reproduces these values); the energy increases monotonically with λ as demonstrated in Table I.

In the Bogomol'nyi limit, $\lambda=0$, the energy coincides with the topological lower bound, i.e. (omitting the parameters ξ and κ),

$$E_{10}(\lambda)\geqslant E_{10}(0)=1. \tag{30}$$

The solution, the Prasad–Sommerfield monopole, is expressed in terms of elementary functions.[19] Its behavior near the origin is given by Eq. (28) but, for $x\to\infty$, we have

$$a(x)\simeq xe^{-x}, \quad h(x)\simeq 1-1/x, \tag{31}$$

instead of Eq. (29).

C. Case $M=0$, $T=1$

The classical solutions considered in this case excite only the gauge and Skyrme fields degrees of freedom; the Higgs field is identically vanishing. The static equations describe the gauged Skyrmion studied in Ref. 6; Eq. (23) is trivial since $h(r)=0$. The classical energy is computed from Eq. (20) and makes sense only if $\lambda=0$ (in fact, since the Higgs field is zero, the Higgs potential plays no role). The topological lower bound reads

$$E_{01}(\lambda=0,\xi,\kappa)\geqslant \frac{3\pi}{2}\frac{\sqrt{\xi\kappa}}{\sqrt{1+9\kappa}}. \tag{32}$$

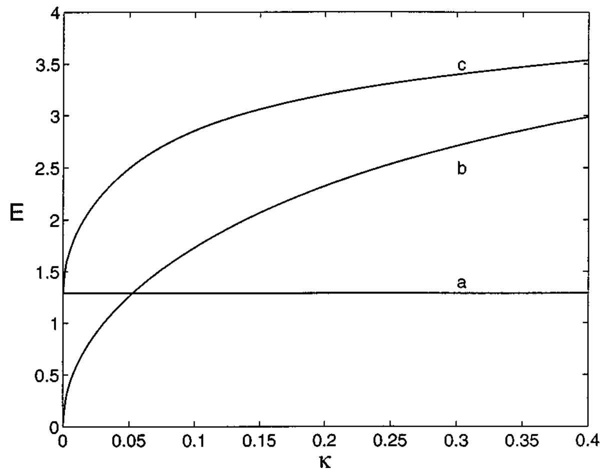

FIG. 1. The energies Eq. (26) of the monopole (line *a*), of the monopole-Skyrmion (line *c*) (for $\lambda = 1$, $\xi = 1$) and the energy of the gauged Skyrmion (line *b*) as functions of κ.

Due to the vanishing Higgs field, the parameter ξ can be changed by a rescaling of the radial variable x and it will be set equal to one. Comparison of the energies of the gauged Skyrmion and of the monopole is demonstrated in Table I and Fig. 1.

Let us now come to the detailed discussion of the solutions in the region $\kappa \approx 0.8$. For completeness, it is useful to summarize the possible boundary conditions available for the gauged Skyrmion. At the origin $x=0$ the behavior of the radial functions is uniquely determined by the condition of continuity of the fields at the origin:

$$a(x) = 1 + A_1 x^2 + o(x^3), \quad f(x) = \pi + F_1 x + o(x^2). \tag{33}$$

In contrast, in the $x \gg 1$ asymptotic region, several conditions are consistent with the finiteness of the energy. Classical solutions of the equations have been obtained[6] with the two following sets

$$\text{type A:} \quad a \approx 1 - \frac{A}{x}, \quad f \approx \frac{F}{x^2}, \tag{34}$$

$$\text{type B:} \quad a \approx \frac{A}{x^\alpha}, \quad f \approx \frac{F}{x}, \tag{35}$$

where $\alpha \equiv (\sqrt{4F^2 - 3} - 1)/2$.

The following results were obtained in Ref. 6. For small values of κ, the solution is of type A, its energy increases monotonically from $E=0$ (for $\kappa=0$) and the branch (say branch A) stops at a critical value $\kappa = \kappa_A^{cr} \approx 0.8091$. For large values of κ (in fact for $\kappa > \kappa_B^{cr} \approx 0.69122$) the solution is of type B. We call this branch B. By using arguments of catastrophe theory,[13] one can reasonably expect the occurrence of a third branch of solutions on the interval $\kappa \in [\kappa_B^{cr}, \kappa_A^{cr}]$, as was explained in Ref. 6.

A third branch indeed exists. The solutions on this branch obey the condition of type A and therefore we refer to it as branch $\tilde{A}$. The energies of the three branches of solutions are depicted in Fig. 2. The branches A and $\tilde{A}$ terminate at $\kappa = \kappa_A^{cr}$, forming a cusp catastrophy. The transitions of the profile of the solutions from branch A to branch $\tilde{A}$ is smooth.

In contrast, when the limit $\kappa \rightarrow \kappa_B^{cr}$ is considered, the solutions of the branch $\tilde{A}$ approach the limit of branch B in a subtle way. For instance, the value x_m for which the function $a(x)$ has a minimum (say a_m) tends to infinity, while a_m tends to zero. For values of κ close to κ_1, the solutions of branches $\tilde{A}$ and B coincide on a large interval of x (typically on $x \in [0, 10^7]$ for κ

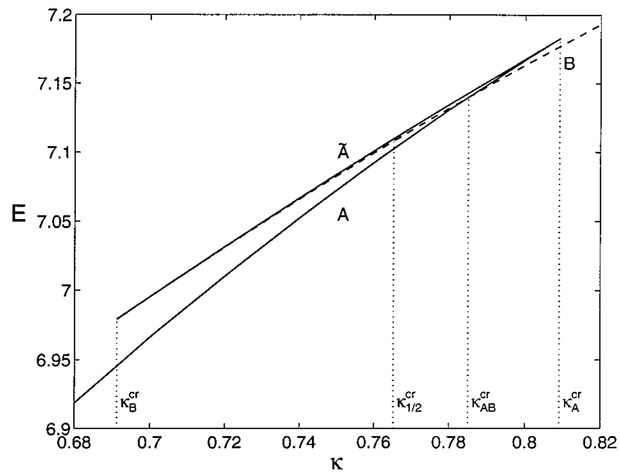

FIG. 2. The energy of the gauged Skyrmion as a function of κ in the region of the phase transition. The branches A, Ã are represented by the solid line and branch B by the dashed line.

$=0.6914$) and deviate from each other for larger values of x. In the limit $\kappa\to\kappa_B^{cr}$ this interval becomes infinitely large and the two solutions deviate at infinity. This can clearly be seen from Fig. 3. A similar demonstration can be made for the function $f(x)$, namely that near the critical point κ_B^{cr} these functions for the two solutions on branches B and Ã also coincide. We do not display the graphs analogous to Fig. 3 in this case. The behavior of the solutions is further illustrated by Figs. 4 and 5 where we plot respectively the value of F_1 [defined in Eq. (33)] for the three branches and the value of α as a function of κ.

Figure 2 furnishes a simple interpretation of the three solutions. To discuss it, we introduce κ_{AB}^{cr} as the value of κ where the energy of the branches A and B coincide ($\kappa_{AB}^{cr}\approx 0.785$). On the interval $\kappa\in[\kappa_B^{cr},\kappa_{AB}^{cr}]$ the solution on branch A constitutes the absolute minimum of the energy functional E_2, while the one on branch B is a local minimum. The solution on the branch Ã is a sphaleron corresponding to a saddle point which represents the energy barrier between the two minima. The situation is similar on the interval $\kappa\in[\kappa_{AB}^{cr},\kappa_A^{cr}]$; the absolute (respectively, local) minimum energy configuration is then on branch B (respectively, A). As κ approaches the critical value κ_B^{cr} the local minimum of branch B approaches the sattle point of branch Ã. At the critical

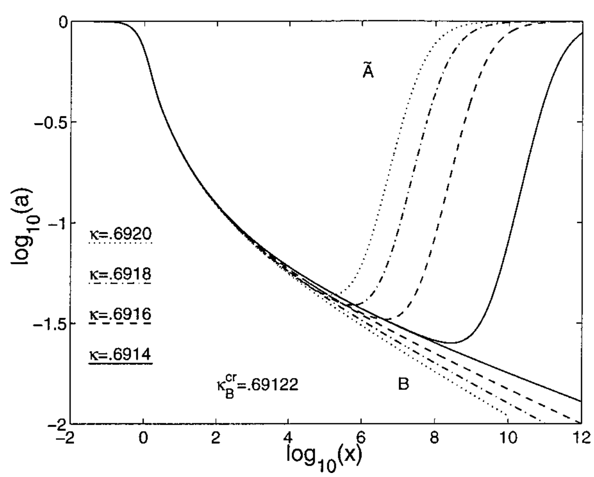

FIG. 3. The (logarithm of the) function $a(x)$ on the two branches B and Ã on a logarithmic scale for several values of κ approaching the critical value κ_B^{cr}.

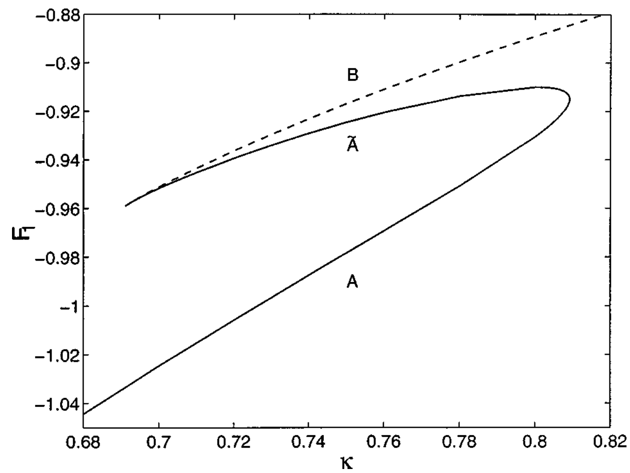

FIG. 4. The quantity F_1 defined in Eq. (33) is plotted as a function of κ for the branches A, $\tilde{\mathrm{A}}$ (the solid line) and for the branch B (the dashed line).

κ both coincide and form an inflection point. For $\kappa<\kappa_{\mathrm{B}}^{\mathrm{cr}}$ this point is no longer an extremum and the solutions of branches B and $\tilde{\mathrm{A}}$ cease to exist. The global minimum of branch A is then the only extremum and only branch A solutions exist. The same scenario applies at the other critical value $\kappa_{\mathrm{A}}^{\mathrm{cr}}$ where the solutions of branches A and $\tilde{\mathrm{A}}$ stop to exist and only branch B solutions exist.

D. Case $M=1$, $T=1$

It is natural to call this solution the ''monopole-Skyrmion.'' The three functions a, h, f are nontrivial and obey the following boundary conditions at $x=0$ and as $x\rightarrow\infty$, respectively,

$$a(0)=1,\quad h(0)=0,\quad f(0)=\pi, \tag{36}$$

$$a(x)\simeq Ae^{-x},\quad h(x)\simeq 1-Be^{-\sqrt{2\lambda}x},\quad f(x)\simeq\frac{F}{x}, \tag{37}$$

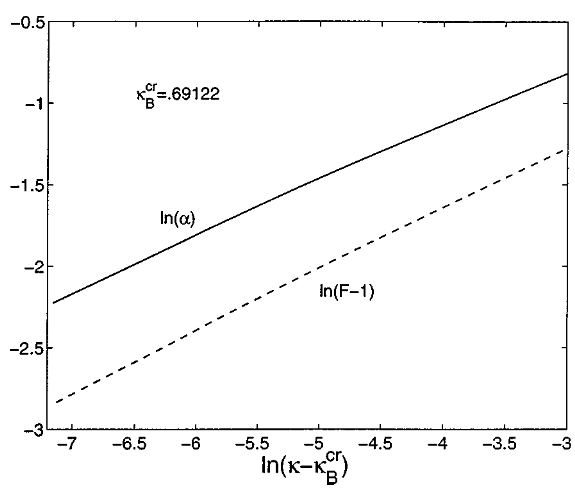

FIG. 5. The quantities $\ln(F-1)$ and $\ln(\alpha)$ [defined in Eq. (35)] are plotted as functions of the parameter $\ln(\kappa-\kappa_{\mathrm{B}}^{\mathrm{cr}})$.

where A, B and F are constants. In contrast to the case of the gauged Skyrmion solution,[6] the finite energy condition leads to a unique asymptotic behavior of the solution.

The energy of the solution is given in Table I for several values of λ (for $\xi=1$ and $\kappa=0.4$), indicating that the energy of the monopole-Skyrmion varies rather slightly with λ. The corresponding lower bound inequality reads

$$E_{11}(\lambda,\xi,\kappa) \geqslant \frac{1}{\sqrt{2}} + \frac{3\pi}{2}\frac{\sqrt{\xi\kappa}}{\sqrt{1+18\kappa}}. \tag{38}$$

Table I and Fig. 1 further exhibit a comparison between the energies of the monopole, the monopole-Skyrmion, and the gauged Skyrmion.

In Sec. IV C, we exposed the properties of the SO(3) gauged Skyrme model and depicted the bifurcations occurring in Fig. 2. The monopole-Skyrmion solution studied in this section is also expected to feature similar bifurcations, but this is not seen from Fig. 1. (Indeed, for that range of parameters, the $M=0$, $T=1$ solution also does not feature the bifurcations present in Fig. 2.) We have checked that for the value of the parameter $\xi=1$, there occur no bifurcations for all values of the parameter κ. On the other hand, if we change the values of the parameter ξ, then we would expect that similar bifurcations as in Fig. 2 will manifest themselves also for the monopole-Skyrmion. Since the main interest in the latter is the unwinding of the baryon number in the presence of the monopole, and since any bifurcations analogous to those in Fig. 2 would feature branches of the solution with the same ($T=1$) baryon number, we eschew a detailed discussion of these in the present work.

E. General properties

We have constructed numerically the three nontrivial topological solitons above for numerous values of the coupling constants λ, ξ, κ and computed their energies. In order to give an idea of the relative magnitudes for the different classes, let us choose $\lambda=0$, $\xi=1$, $\kappa=0.4$, then

$$E_{00}=0, \quad E_{10}=1.0, \quad E_{01}\approx 2.98, \quad E_{11}\approx 3.45. \tag{39}$$

Note that the monopole energy satisfies the topological lower bound.

The behavior of the solutions in the limit $\kappa\to 0$, with λ,ξ fixed, was carefully analyzed. Our numerical analysis strongly supports the following formula:

$$\lim_{\kappa\to 0} E_{M1}(\lambda,\xi,\kappa)=E_{M0}(\lambda,\xi,0) \quad \text{for } M=0,1 \tag{40}$$

as illustrated by Fig. 1. Indeed, in the limit $\kappa\to 0$, the functions $a(r),h(r)$ representing the solutions of the $M=T=1$ sector approach the profile of the monopole solution (i.e., $M=1$, $T=0$). At the same time, the function $f(r)$ is more and more peaked at $r=0$ (in particular $\lim_{\kappa\to 0} f'(0)=\infty$) and tends to zero if $r\neq 0$.

This result demonstrates in particular that the coupling of the Skyrmion to a monopole cannot stabilize the Skyrmion; the Skyrme term is necessary to guarantee a localized structure to the $T=1$ soliton.

The same phenomenon occurs with the branch of the gauged Skyrmion ($M=0$, $T=1$).[6] The energy in this limit tends to zero, namely to the energy of the vacuum ($M=T=0$).

A remark should be made concerning the interpretation of the monopole-Skyrmion as a bound system of a monopole with magnetic charge M and a gauged Skyrmion with baryon number T. Consider a monopole located in a region U_m centered at a point x_m and a gauged Skyrmion located in a region U_{Sk} centered at a point x_{Sk} far away from each other. Then the Skyrmion field and the corresponding gauge field will vanish outside the region U_{Sk}. Consequently, U_m contains a pure monopole, consisting of a gauge field and a Higgs field. Outside U_m the gauge field will vanish, however, the Higgs field does not vanish. Instead it will be equal to its VEV $\langle\Phi\rangle_{\mathrm{vac}}$. In the region U_{Sk} containing the Skyrmion the nonzero Higgs field is still present and we have to allow for interaction with the gauge field, $|D_j\langle\Phi\rangle^\alpha_{\mathrm{vac}}|^2$. The Higgs vacuum is a constant far away from the monopole and generates masses for the gauge fields. Furthermore, the Higgs vacuum breaks the rotational symmetry. Consequently, the gauged Skyrmion solutions in the presence of a constant Higgs field will no longer possess spherical symmetry. In addition, it might be expected that the electromagnetic flux will not vanish. If we impose the condition for the Higgs vacuum that the interaction with the gauge field has to vanish, $D_j\langle\Phi\rangle^\alpha_{\mathrm{vac}}=\epsilon^{\alpha\beta\gamma}A^\beta_j\langle\Phi\rangle^\gamma_{\mathrm{vac}}=0$, then we will find that the gauge field has to be parallel to the Higgs vacuum in isospace. This also breaks the spherical symmetry.

To conclude, the interpretation of the monopole-Skyrmions as a bound state of a spherically symmetric monopole and a spherically symmetric gauged Skyrmion seems to be misleading.

V. NUMERICAL RESULTS, CASE $A_0\neq 0$

In order to obtain a nontrivial function $g(x)$ from Eq. (23), a nonvanishing asymptotic value, say q, for this function has to be imposed.[7] In the asymptotic region $x\gg 1$ Eq. (23) is satisfied by

$$g(x)=q-\frac{c_1}{x}+o(x^{-2}), \tag{41}$$

where q,c_1 are constants and q plays a major role in the construction. The equations, together with the finite energy condition require $0\leqslant q\leqslant 1$, which can be seen as follows. In Eq. (22) the Higgs field and the dyon field contributions, $h^2(x)-g^2(x)$, generate asymptotically the mass term $m^2_{(a)}=1-q^2$ for the gauge field function $a(x)$. For $q>1$, $m^2_{(a)}$ becomes negative and leads to an oscillating function $a(x)$ in the asymptotic region. Consequently, the term a^2g^2 in Eq. (17) is not integrable and no dyon solution exists for $q>1$.

The electric charge, as defined in Ref. 7, is directly related to the constant c_1:

$$Q=\frac{1}{4\pi\eta}\int\vec{\Phi}\cdot\vec{F}_{0i}dS_i\equiv\frac{1}{e}\tilde{Q} \tag{42}$$

$$=\frac{1}{4\pi}\int\left(r^2\frac{dg}{dr}\right)\Big|_{r\to\infty}\sin\theta d\theta d\phi=\frac{1}{e}c_1, \tag{43}$$

having used the ansatz (13)–(14) and Eq. (41).

Another very interesting quantity is the chiral anomaly due to the dyon-Skyrmion soliton whose classical solutions will be studied numerically. The anomaly equation for the chiral charge is

$$\frac{dQ_5}{dt}=\frac{e^2}{8\pi^2}\int d\mathbf{x}\mathbf{E}_i\cdot\mathbf{B}_i=-\frac{e^2}{8\pi^2}4\pi[g(r)(a(r)-1)]^\infty_{r=0}=\frac{e^2}{2\pi}q, \tag{44}$$

having used the ansatz (13)–(14) and Eq. (41). We now discuss the solutions by adopting the same presentation as in Sec. IV. Again, in addition to the two types of solutions characterized by the nontrivial charges ($M=1$, $T=0$), and ($M=1$, $T=1$), we include the field configurations corresponding to ($M=0$, $T=0$), which is the dyonlike solution to the SO(3) gauged Skyrme model,[5,6] which as it happens in this case turns out to be the trivial field configuration.

A. Case $M=1$, $T=0$

The solutions are the dyons of Julia and Zee.[7] Here we present an in-depth analysis of this solution. The limit $\lambda=0$ corresponds to the Prasad–Sommerfield dyon[19] (PS dyon). It is worth analyzing this case separately because the solution can be computed analytically and it provides a good check of our numerical routines.

1. Case λ=0

The profile of the radial functions of the PS dyon reads[19]

$$a(x)=\frac{cx}{\sinh(cx)}, \tag{45}$$

$$g(x)=\frac{cq}{\sqrt{1-q^2}}\left(\coth cx-\frac{1}{cx}\right), \tag{46}$$

$$h(x)=\frac{c}{\sqrt{1-q^2}}\left(\coth cx-\frac{1}{cx}\right), \tag{47}$$

and the PS monopole is recovered for $q=0$. Our parameter q is related to γ of Ref. 19, by $q=\tanh(\gamma)$. We have chosen the arbitrary scale in the PS solution $c=\sqrt{1-q^2}$ so that the asymptotic value of the Higgs field function $h(x)$ of the PS solution given above be equal to 1, since we are also studying the dyons for the $\lambda>0$ case where the asymptotic value of $h(x)$ equals 1. The charge and energy of the PS dyon are given by

$$\tilde{Q}=\frac{q}{\sqrt{1-q^2}}, \quad \tilde{E}=\frac{1}{\sqrt{1-q^2}}, \tag{48}$$

$$\tilde{E}_1=\frac{q^2}{2\sqrt{1-q^2}}, \quad \tilde{E}_2=\left(1-\frac{1}{2}q^2\right)\frac{1}{\sqrt{1-q^2}}\approx 1+\frac{1}{8}q^4+o(q^6), \tag{49}$$

where the dimensionless quantities $\tilde{E}$, $\tilde{E}_p$ and $\tilde{Q}$ are defined in Eqs. (21), (18) and (42), respectively.

For small values of q the ''magnetic'' contribution to the energy, E_2, varies slightly with q, accounting for the feedback of the electric charge on the classical magnetic energy. We would like to stress that our numerical results are in full agreement with these exact formulas.

The dependence of the charge $\tilde{Q}$ of the PS dyon as a function of q is represented in Fig. 6 (curve a). Similarly we have reported in Fig. 7 (curve a) the energy $\tilde{E}$ of the PS dyon as a function of $\tilde{Q}$. Clearly the energy and the charge of the PS dyon can be arbitrarily large when $q\rightarrow 1$.

2. Case λ≠0

For the dyon solution, the boundary conditions for the function $g(x)$ can be read from Eq. (41), and those of the functions $a(x)$, $h(x)$ from Eqs. (28) and (29), with the exception of the behavior of the function $a(x)$ in the $x\gg 1$ region, which now takes the form

$$a(x)\simeq Ae^{-\sqrt{1-q^2}x}. \tag{50}$$

Our analysis demonstrates that the energy of the dyon obeys a Bogomol'nyi inequality $\tilde{E}(\lambda,Q)>\tilde{E}(0,0)=1$. The main distinguishing feature of the $\lambda\neq 0$ dyon versus the PS dyon is that its electric charge and its classical energy are bounded for $q\in[0,1]$.

This phenomenon appears clearly in Figs. 6 and 7, respectively, where the quantities $\tilde{Q}$ as a function of q, and $\tilde{E}$ as a function of $\tilde{Q}$, are plotted for $\lambda=0.5$. More generally, it appears that the electric charge of the dyon constructed with a given value of the parameter q decreases when λ increases. The three bullets on Fig. 7 represent the data given in Ref. 7; according to our numerical results they should lie on line *b*. Our numerical results therefore slightly disagree with Ref. 7.

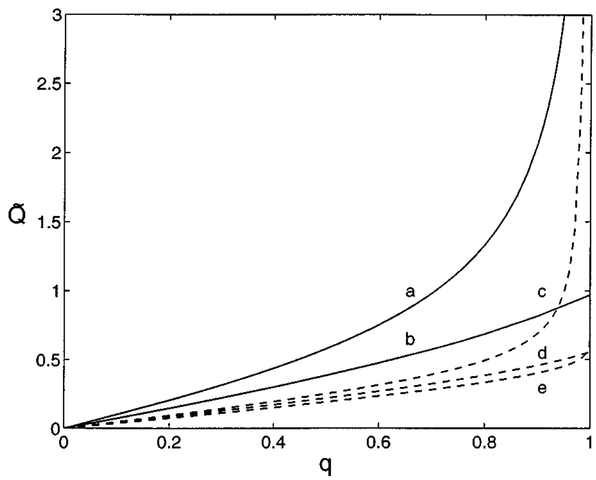

FIG. 6. The values of the electric charge $\tilde{Q}$ as a function of the parameter q. The solid lines represent the dyon for $\lambda=0$ (line a) and $\lambda=0.5$ (line b). The dashed lines represent the dyon-Skyrmion ($\xi=1$) for $\lambda=0$, $\kappa=0.4$ (line c), $\lambda=0.5$, $\kappa=0.4$ (line d) and $\lambda=0$, $\kappa=1$ (line e).

The star on line b of Fig. 7 indicates the maximal accessible charge of the dyon solutions for a fixed value of λ. This contrasts with line a which asymptotically tends to infinity, in agreement with Eqs. (48) and (49). The solutions with maximal electric charge and energy correspond to the case $q=1$ which we discuss next.

3. Case q=1

In the limit $q=1$ in Eq. (41) Eq. (22) ceases to impose the exponential decay Eq. (50) for the function $a(x)$; we have instead

$$a(x)\simeq Ae^{-\sqrt{8c_1x}} \quad \text{for } x\to\infty, \tag{51}$$

where c_1 is defined in Eq. (41).

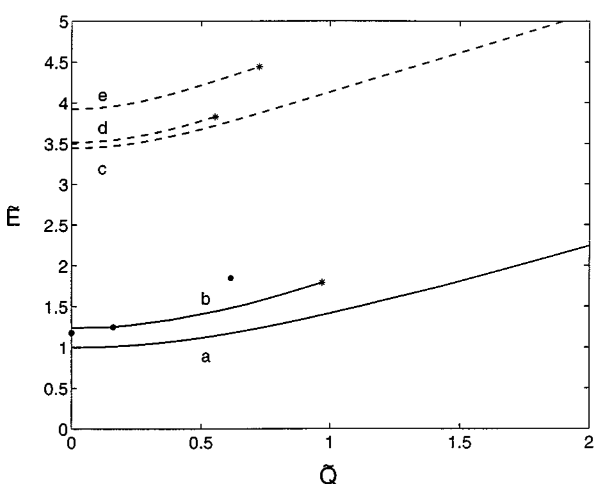

FIG. 7. The values of the energy $\tilde{E}$ as a function of the parameter $\tilde{Q}$. The solid lines represent the dyon for $\lambda=0$ (line a) and $\lambda=0.5$ (line b). The dashed lines represent the dyon-Skyrmion ($\xi=1$) for $\lambda=0$, $\kappa=0.4$ (line c), $\lambda=0.5$, $\kappa=0.4$ (line d) and $\lambda=0$, $\kappa=1$ (line e). The stars depict the points where the solution has maximal finite charge. The bullets correspond to the data of Ref. 7.

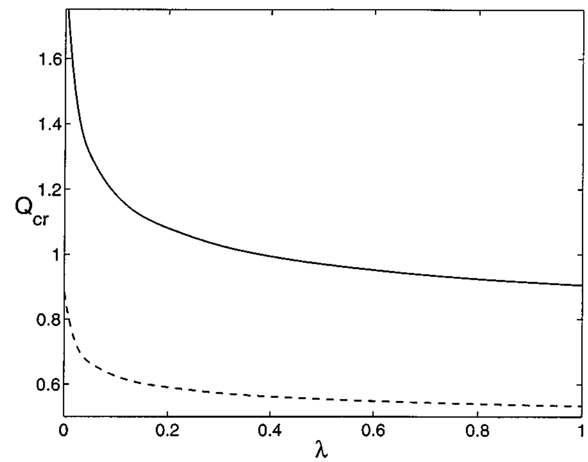

FIG. 8. The value of the critical charge Q_{cr} as a function of λ. The solid line refers to the dyon solution. The dashed line refers to the dyon-Skyrmion solution for $\xi=1$ and $\kappa=0.4$.

Fixing $\lambda \neq 0$, the electric charge (and similarly the classical energy) of the dyon cannot exceed a critical value, say $Q_{cr}(\lambda)$. This quantity is plotted against λ in Fig. 8 (solid line).

B. Case $M=0$, $T=1$

No finite energy dyonlike solutions supporting a nonvanishing (non-Abelian) electric field can be found in this case. Due to the absence of the Higgs field $(h=0)$, Eq. (22) leads to an oscillating asymptotic behavior of $a(x)$. The term a^2g^2 in the energy Eq. (17) cannot therefore be integrated.

C. Case $M=1$, $T=1$

The boundary conditions compatible with a finite energy solution in this case are identical to Eqs. (36), (37) and (41). It is possible to construct the dyon-Skyrmion solutions. The dyon-Skyrmion display many features of the dyons, discussed at length above. These features are illustrated by Figs. 6 and 7 (dashed curves c, d and e) and by Fig. 8 (the dashed line). In addition we illustrate the dependence of the energy on the Skyrme coupling constant κ in Fig. 9 (the dashed line) for $q=0.5$ and $\lambda=0$. The energy is an increasing function of κ. In the limit of vanishing κ

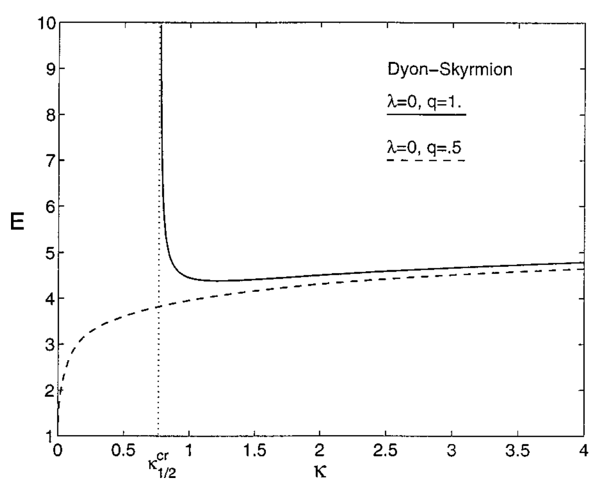

FIG. 9. The energies of the dyon-Skyrmions with $q=1$ (the solid line) and $q=0.5$ (the dashed line) as functions of $\kappa(\lambda=0, \xi=1)$. The vertical dotted line indicates the critical value of κ.

the energy of the dyon-Skyrmion converges to the energy of the dyon-monopole. This can be compared with the behavior of the energy of the monopole-Skyrmion, shown (for $\lambda=1$) in Fig. 1, where for vanishing κ the energy tends to the energy of the monopole. For our considerations leading to our conclusions in Fig. 9, we have chosen $q=0.5$ as a typical value in the allowed range $0\leqslant q\leqslant 1$. We expect that our results, summarized by the dashed curve in Fig. 9, are typical for any allowed value of q, and also for any value of the Higgs coupling constant λ, except in the important case of $\lambda=0$ and $q=1$. The dyon-Skyrmion characterized by the boundary value $q=1$ in the $\lambda=0$ model has peculiar and interesting properties which we analyze in the next paragraphs.

For $\lambda=0$ the solutions of Eqs. (23) and (24) are proportional to each other. Assuming that $h=1$ at infinity, the proportionality constant is given by q, Eq. (41). Thus for $q=1$ the functions $h(r)$ and $g(r)$ are identical. In this special case $h^2(r)$ and $g^2(r)$ cancel each other in Eq. (22). Consequently, Eqs. (22) and (25) reduce to the field equation of the gauged Skyrme model (Sec. IV C), and can be solved independently of $h(r)$, $g(r)$.

The solutions of these equations are now given by the branch B solutions of the gauged Skyrme model. Once a solution for the function $a(r)$ is found the equations for the functions $h(r)$ and $g(r)$ can be solved. Recalling that the branch B solutions exist for all $\kappa\geqslant\kappa_{\mathrm{B}}^{\mathrm{cr}}$, we expect the existence of the dyon-Skyrmion solution for the same range of coupling constants κ. However, not all of these solutions are finite energy solutions. This can be seen easily by inspecting the static Hamiltonian $\mathcal{E}=\mathcal{E}_1+\mathcal{E}_2$ given in Eqs. (19) and (20), where the contributions from the functions $h(r)$ and $g(r)$ do not cancel. The asymptotic behavior of these terms is dominated by $a^2(r)h^2(r)$, $a^2(r)g^2(r)$. Using the boundary conditions $h(\infty)=1$, $g(\infty)=1$, and the asymptotic form of the function $a(r)$, Eq. (35), these terms behave like $\approx A^2(\kappa)/x^{2\alpha(\kappa)}$ for large x, where $\alpha(\kappa)$ is a function determined numerically. Thus the integration of these terms will give finite contributions only if $\alpha(\kappa)>1/2$. This restricts the range of the coupling constant κ to $\kappa_{1/2}^{\mathrm{cr}}<\kappa<\infty$, where $\kappa_{1/2}^{\mathrm{cr}}$ is defined by $\alpha(\kappa_{1/2}^{\mathrm{cr}})=1/2$. For $\xi=1$ we find $\kappa_{1/2}^{\mathrm{cr}}=0.7652$.

In Fig. 9 we show the dependence of the energy on the coupling constant κ for $q=1$ (the solid line) and for $q=0.5$ (the dashed line). (For $q<1$, as stated above, solutions exist for all values of κ, the energy is a monotonic function of κ, and the limit $\kappa\to 0$ the energy approaches the energy of the dyon solution.) For $q=1$ the energy is an increasing function of κ for large values of κ only. It has a minimum at $\kappa=1.21$. As κ approaches its critical value $\kappa_{1/2}^{\mathrm{cr}}$ the energy becomes increasingly large and diverges at $\kappa=\kappa_{1/2}^{\mathrm{cr}}$.

The charge $\tilde{Q}$ of the solutions is determined by the asymptotic behavior of the function $g(r)$, Eq. (43). Solving Eq. (23) for large x we find the following expressions for the charge:

$$\tilde{Q}=\begin{cases}\lim\limits_{x\to\infty}\left(c_1-\dfrac{2A^2}{2\alpha-1}x^{-(2\alpha-1)}\right) & \text{for}\quad \alpha>\dfrac{1}{2}\\ \lim\limits_{x\to\infty}(2A^2\ln(x)) \quad \text{for}\quad \alpha=\tfrac{1}{2} & \\ \lim\limits_{x\to\infty}\left(-\dfrac{2A^2}{2\alpha-1}x^{-(2\alpha-1)}\right) & \text{for}\quad \alpha<\dfrac{1}{2}.\end{cases} \tag{52}$$

Thus solutions with finite charge exist only for $\alpha>1/2$, i.e., for the same range of the coupling constant κ where finite energy solutions exist.

In Fig. 6 we show the dependence of the charge on the parameter q for $\kappa=0.4$ and $\kappa=1.0$. For $\kappa=0.4$ ($<\kappa_{1/2}^{\mathrm{cr}}$) there is no finite charge solution for $q=1$. Consequently, the charge as a function of q diverges as q approaches the value 1. In contrast, for $\kappa=1$ ($>\kappa_{1/2}^{\mathrm{cr}}$) the solution with $q=1$ exists and the charge is finite for all values of $q\in[0,1]$.

In Fig. 7 the energy as a function of the charge is shown for $\kappa=0.4$ and $\kappa=1.0$. For $\kappa=0.4$ ($<\kappa_{1/2}^{\mathrm{cr}}$) the energy and the charge can take arbitrarily large values. In this case the energy is a monotonically increasing function of the charge with no end point. For $\kappa=1$ ($>\kappa_{1/2}^{\mathrm{cr}}$) the energy is again a monotonically increasing function of the charge. However, only finite energy and charge solutions exist for this value of κ. Thus the graph of the function $\tilde{E}(\tilde{Q})$ ends at the maximal value of the charge.

VI. SUMMARY AND DISCUSSION

We have found monopole-Skyrmion and dyon-Skyrmion solutions to an SO(3) gauged Higgs and O(4) sigma (Skyrme) model, in which both scalar matter fields interact with the gauge field but not with each other. The Higgs field is an isovector, like in the GG model, while the S^3 valued (sigma) field is gauged according to the prescription used in Refs. 5 and 6.

In the $A_0^\alpha = 0$ gauge the static Hamiltonian is bounded from below by the sum of the two topological charge densities, the monopole charge, and the degree of the map of the S^3 field on $\mathbb{R}_3$. Thus the imposition of spherical symmetry reduces the system to a one-dimensional subsystem, and the resulting differential equations are first integrated analytically in the asymptotic regions and then numerically. This yielded the monopole-Skyrmion. In addition to the monopole-Skyrmion solutions, we have made an enhanced numerical study of the SO(3) gauged Skyrmion, because these solutions are instrumental in constructing some of the dyon-Skyrmions given in the following paragraph.

In the $A_0^\alpha \neq 0$ gauge, the Euler–Lagrange equations arising from the variation of the static Hamiltonian density do not yield a soliton with nonvanishing A_0^α and hence have $E_0^\alpha = 0$. Instead, the variational equations arising from the (non-positive-definite) Lagrangian density in the static limit support spherically symmetric solutions with $E_i^\alpha \neq 0$. This is also what happens with the JZ dyon. There,[7] in spite of the nonpositive definiteness of the functional subjected to the variational principle, it happens that after taking the static limit and imposing spherical symmetry, these equations reduce to a set of consistent, i.e., not overdetermined, set of coupled ordinary differential equations. Their solutions support a nonvanishing A_0^α field. These ordinary differential equations also result from the variation of a certain one-dimensional (radial) functional which, in contrast to the one-dimensional energy functional, is not positive definite.

In the light of the surprisingly successful outcome for the JZ dyon, we were motivated to address the same question for the SO(3) gauged O(4) model.[5,6] Subjecting the Lagrangian to the variational principle and then taking the static limit and imposing spherical symmetry, we found that this also led to a consistent set of coupled ordinary differential equations. The same situation obtains with the composite model of this paper, and it is the dyonlike solitons of these last equations which yielded the dyon-Skyrmion. Concerning the SO(3) gauged Skyrme model on its own, while its equations of motion reduce to a consistent set of coupled ordinary differential equations, their solutions support only vanishing electric field.

As a byproduct of our study of the dyon-Skyrmion, we made a detailed reanalysis of the JZ dyon refining our understanding of the latter. Namely exploring the dependence of the energy of the dyon on its electric charge shows that the dyon energy is always higher than the energy of the PS monopole, extending the Bogomol'nyi identity available for monopole.

An important result of the numerical analysis of the monopole-Skyrmion solution is that, as the coupling strength of the Skyrme term is shrunk down to zero the monopole-Skyrmion reduces to the monopole, as depicted in Fig. 1. Thus the monopole does not stabilize the SO(3) gauged sigma model without a Skyrme term, something that is not prohibited by the Derrick scaling requirement.

Perhaps the most interesting aspect of the dyon-Skyrmion occurs for the model in the PS limit $(\lambda = 0)$ in the special case where the boundary value q of the function $g(r)$ parametrizing A_0^α equals 1. In this case, the equations governing the functions $a(r)$ (parametrizing the gauge field) and the function $f(r)$ (governing the Skyrme field) decouple from the fields $h(r)$ (parametrizing the Higgs field) and $g(r)$. As a consequence the solutions for the functions $a(r)$ and $f(r)$ are just the (branch B of the) gauged-Skyrmion solutions and exist only for values of the Skyrme coupling constant larger than a critical value $\kappa_{\mathrm{B}}^{\mathrm{cr}}$, as seen from Fig. 2. However, when the integrations of the Higgs field function $h(r)$ and of the dyon function $g(r)$ are taken into account, then finite energy solution only exists if the Skyrme coupling constant is larger than the critical value $\kappa_{1/2}^{\rm cr} > \kappa_{\rm B}^{\rm cr}$, see Fig. 9. The energy of the solution at this critical value is found to become infinite and for lower values of the Skyrme coupling constant no finite energy solution exists. The time rate of change of the chiral charge Eq. (44) is equal to the integer 1 (up to normalization) for all values of the Skyrme coupling constant κ down to the critical value $\kappa_{1/2}^{\rm cr}$, below which no finite energy solutions exist. We hope that this result may prove relevant to the semiclassical description of monopole catalysis of Fermion number nonconservation. If for example it could be argued that the dyon-Skyrmion favored by Nature is the solution to the system Eq. (9) in the PS limit, with the asymptotic constant $q=1$, i.e., for which the quantity dQ_5/dt is an integer (up to normalization), then it would follow that below the critical value $\kappa_{1/2}^{\rm cr}$ there will be no Q_5 violating rate. We intend to return to this question in the near future.

ACKNOWLEDGMENTS

We are grateful to W. Nahm and V. A. Rubakov for helpful discussions. B.K. was supported by Forbairt Grant No. SC 97-636, and Y.B. and D.H.T. acknowledge partial support from Forbairt Project No. IC/97/019.

[1] V. A. Rubakov, JETP Lett. **33**, 658 (1981); ''Problems of High Energy Physics and Quantum Field Theory,'' Proceedings of the Fourth International Seminar, Protvino, 1981, IHEP (unpublished) vol. 1, pp. 148–155; Nucl. Phys. B **203**, 311 (1982).
[2] C. G. Callan, Phys. Rev. D **25**, 2141 (1982); Phys. Rev. D **26**, 2058 (1982).
[3] C. G. Callan and E. Witten, Nucl. Phys. B **239**, 161 (1985).
[4] G. 'tHooft, Nucl. Phys. B **79**, 276 (1974); A. M. Polyakov, JETP Lett. **20**, 194 (1974).
[5] K. Arthur and D. H. Tchrakian, Phys. Lett. B **378**, 187 (1996).
[6] Y. Brihaye and D. H. Tchrakian, Nonlinearity **11**, 891 (1998).
[7] B. Julia and A. Zee, Phys. Rev. D **11**, 2227 (1975).
[8] T. H. R. Skyrme, Proc. R. Soc. London, Ser. A **260**, 127 (1961); Nucl. Phys. **31**, 556 (1962).
[9] E. Witten, Nucl. Phys. B **223**, 422 (1983).
[10] O. Kaymakcalan, S. Rajeev, and J. Schechter, Phys. Rev. D **30**, 2354 (1984); **30**, 434 (1984); Y. Brihaye, N. K. Pak, and P. Rossi, Phys. Lett. B **149**, 191 (1984); Nucl. Phys. **254**, 71 (1985).
[11] M. Chemtob, Nucl. Phys. A **466**, 509 (1987).
[12] V. A. Rubakov, Nucl. Phys. **256**, 509 (1985); J. Ambjorn and V. A. Rubakov, Nucl. Phys. B **256**, 594 (1985).
[13] G. Eilam, D. Klabucar, and A. Stern, Phys. Rev. Lett. **56**, 1331 (1986).
[14] J. Goldstone and F. Wilczek, Phys. Rev. Lett. **47**, 986 (1981).
[15] E. D'Hoker and E. Farhi, Nucl. Phys. B **241**, 109 (1984).
[16] B. M. A. G. Piette and D. H. Tchrakian, ''Topologically stable soliton in the U(1) gauged Skyrme model,'' hep-th/9709189.
[17] H. Pagels, Phys. Rev. D **13**, 343 (1976); W. Marciano and H. Pagels, *ibid.* **14**, 531 (1976).
[18] E. B. Bogomol'nyi and M. S. Marinov, Sov. J. Nucl. Phys. **23**, 535 (1976).
[19] M. K. Prasad and C. M. Sommerfield, Phys. Rev. Lett. **35**, 760 (1975).

Measure-valued solution to the strongly degenerate compressible Heisenberg chain equations

Shijin Ding
Department of Mathematics, South China Normal University, Guangzhou 510631, People's Republic of China

Boling Guo and Fengqiu Su
Center for Nonlinear Studies, Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, People's Republic of China

(Received 8 June 1998; accepted for publication 14 July 1998)

In this paper we are concerned with the existence of solutions to the compressible Heisenberg chain equations. By the vanishing viscosity method we prove that this system admits at least one measure-valued solution. [S0022-2488(98)03212-5]

I. INTRODUCTION

In 1982, Fivez[1] revisited the 1D classical compressible Heisenberg chain described by the Hamiltonian

$$\mathcal{H}=\sum_{i=1}^{N} P_i^2/2m+\frac{\alpha}{2}\sum_{i=1}^{N-1}(x_{i+1}-x_i)^2-J\sum_{i=1}^{N-1}\mathbf{Z}_i\cdot\mathbf{Z}_{i+1}-\varepsilon\sum_{i=1}^{N-1}(x_{i+1}-x_i)\mathbf{Z}_i\cdot\mathbf{Z}_{i+1}$$

considered earlier by Cieplak and Turski in 1980 where x_i is the displacement of the magnetic ion from equilibrium, without spin–phonon coupling, α is the spring constant and $\varepsilon=J_x$. In the continuum limit, which corresponds to long-wavelength excitations, the equations of motion deduced by Fivez read as

$$m\ddot{\eta}=\alpha\ddot{\eta}+\frac{\varepsilon}{2}\frac{\partial}{\partial x}(\mathbf{Z}')^2, \tag{I.1}$$

$$\dot{\mathbf{Z}}=\frac{\partial}{\partial x}\{(J+\varepsilon\eta')\mathbf{Z}\times\mathbf{Z}'\}, \tag{I.2}$$

where $\mathbf{Z}(x,t)=(Z^1(x,t),Z^2(x,t),Z^3(x,t))\in\mathbb{R}^3$ and the substitution $x_i\rightarrow\eta(x,t)$, $\mathbf{Z}_i\rightarrow\mathbf{Z}(x,t)$ has been made, a dot denotes derivation with respect to t, a prime with respect to x. Fivez tried the solution of the form $\mathbf{Z}=\{\sin\theta\cos\varphi,\sin\theta\sin\varphi,\cos\theta\}$, $\eta=ax+f(u)$ where $\theta(x,t)=\theta(u)$, $\varphi(x,t)=\overline{\varphi}(u)+\Omega t$ with $u=x-ct$ (the lattice and spin wave are assumed to travel at the same velocity c).

Equation (I.1) now becomes

$$(mc^2-\alpha)f''=\frac{\varepsilon}{2}\frac{\partial}{\partial x}(\mathbf{Z}')^2,$$

with boundary condition

$$f'(-\infty)=0,\quad \mathbf{Z}'(-\infty)=0,\quad Z^3(-\infty)=1.$$

By integration, one has

$$(mc^2-\alpha)f'=\frac{\varepsilon}{2}(\mathbf{Z}')^2,$$

$$(\mathbf{Z}')^2=\theta'^2+\sin^2\theta\varphi'^2.$$

Hence Fivez derived the following compressible Heisenberg chain equation:

$$\mathbf{Z}_t=(G(\mathbf{Z}_x)\mathbf{Z}\times\mathbf{Z}_x)_x, \tag{I.3}$$

in which $G(\xi)=A+B|\xi|^2$ with $A=J+\varepsilon\alpha$, $B=(\varepsilon^2/2)(mc^2-\alpha)$, where $mc^2>\alpha$.

The solitons of (I.3) were given by Magyari in Ref. 2. Equation (I.3) with $B=0$ called inhomogeneous Heisenberg chain equations was derived by Balakrishnan in 1982 in Ref. 3. When $B=0$, $A=g(x)$ is some given function, the existence and uniqueness of the smooth solution of (I.3) (the Cauchy problem) were obtained in Ref. 4 ($g(x)\equiv 1$) and in Ref. 5 ($g(x)\not\equiv$constant).

From the mathematical point of view, (I.3) is a strongly degenerate and strongly coupled parabolic system with strong nonlinearity; thus it is kind of important and hard to discuss evolutionary equations.

In this paper, for simplicity, we shall assume A,B are positive constants. We intend to establish the existence of measure-valued solution to the following periodic initial value problem:

$$\mathbf{Z}_t=(G(\mathbf{Z}_x)\mathbf{Z}\times\mathbf{Z}_x)_x, \quad x\in\mathbb{R}^1, \quad t\in\mathbb{R}_+, \tag{I.4}$$

$$\mathbf{Z}(x,0)=\mathbf{Z}_0(x), \quad \mathbf{Z}(x+D,t)=\mathbf{Z}(x-D,t), \quad |\mathbf{Z}_0(x)|\equiv 1, \quad x\in\mathbb{R}^1, \tag{I.5}$$

where $D>0$ is a constant.

To this aim, we use the vanishing viscosity method. Consider the viscosity problem

$$\mathbf{Z}_t=\varepsilon(G(\mathbf{Z}_x)\mathbf{Z}_x)_x+(G(\mathbf{Z}_x)\mathbf{Z}\times\mathbf{Z}_x)_x, \quad x\in\mathbb{R}^1, \quad t\in\mathbb{R}_+, \tag{I.6}$$

$$\mathbf{Z}(x,0)=\mathbf{Z}_0(x), \quad \mathbf{Z}(x+D,t)=\mathbf{Z}(x-D,t), \quad |\mathbf{Z}_0(x)|\equiv 1, \quad x\in\mathbb{R}^1. \tag{I.7}$$

We first prove that problem (I.6)–(I.7) admits at least one global weak solution, and then give the *a priori* estimates for such solutions uniformly in ε to get the existence of the measure-valued solution to (I.4)–(I.5) by letting $\varepsilon\to 0$.

Equations (I.4) and (I.6) are evolutionary p-Laplace like equations but with a ''$\times$'' term. This term gives rise to difficulties in the discussions. We first prove that problem (I.6)–(I.7) is solvable in the space $L^\infty(0,\infty;W^{1,4}(\Omega))\cap L^2(0,\infty;H^2(\Omega))$ when ε is fixed (by the difference method), then we give the *a priori* estimates uniformly in ε. We note that these uniform estimates do not allow us to get the weak solution to (I.4)–(I.5) because of the nonlinearity and the cross product. Hence we can only obtain the measure-valued solution.

II. $\varepsilon>0$: GLOBAL WEAK SOLUTION

To get the existence of the local solution of (I.6)–(I.7), we apply the difference method. We need the following well known lemmas. In the sequel, we denote $\Omega=(-D,D)$.

Lemma II.1 (Ref. 4): Let q, r be real numbers and j, m be integers such that $1\leqslant q,r\leqslant\infty$, $0\leqslant j<m$. *If* $u\in W^{m,r}(\Omega)\cap L^q(\Omega)$, *then*

$$\|D^j u\|_p\leqslant C\|u\|_q^{1-\alpha}\|D^m u\|_r^\alpha,$$

where $\|\cdot\|_p=\|\cdot\|_{L^p(\Omega)}$, $p\geqslant 1$, $j/m\leqslant\alpha\leqslant 1$ *and*

$$\frac{1}{p}-j=\frac{1-\alpha}{q}+\alpha\left(\frac{1}{r}-m\right), \quad \Omega\subset\mathbb{R}^1.$$

Lemma II.2 (Ref. 6): Let p be real number and j, m be integers such that $2 \leqslant p \leqslant \infty,\ 0 \leqslant j < m$. *Then*

$$\| \delta^j u_h \|_p \leqslant C \| u_h \|_2^{1-\alpha} \left(\| \delta^m u_h \|_2 + \frac{\| u_h \|_2}{(2D)^m} \right)^{\alpha},$$

where $u_h = \{ u_j = u(x_j) | j = 0,1,2, \ldots, J \},\ x_j = jh,\ h = 2D/J,\ \alpha = (1/m)(j + 1/2 - 1/p)$,

$$\| \delta^k u_h \|_p = \left(\sum_{i=0}^{J-k} \left| \frac{\Delta_+^k u_i}{h^k} \right|^p h \right)^{1/p}.$$

Lemma II.3 (Ref. 6): Let $u_h = \{ u_j | j = 0, \pm 1, \pm 2, \ldots, \pm J, \ldots \},\ v_h = \{ v_j | j = 0, \pm 1, \pm 2, \ldots, \pm J, \ldots \}$ *such that* $u_{j+J} = u_j,\ v_{j+J} = v_j$. *We have*

$$\text{(i)} \quad \sum_{j=0}^{J-1} u_j \Delta_+ v_j = -\sum_{j=1}^{J} v_j \Delta_- u_j,$$

$$\text{(ii)} \quad \sum_{j=1}^{J} u_j \Delta_+ \Delta_- v_j = -\sum_{j=0}^{J-1} \Delta_+ u_j \Delta_+ v_j,$$

$$\text{(iii)} \quad \Delta_+(u_j v_j) = u_{j+1} \Delta_+ v_j + v_j \Delta_+ u_j,$$

$$\text{(iv)} \quad \Delta_-(u_j v_j) = u_j \Delta_- v_j + v_{j-1} \Delta_- u_j,$$

$$\text{(v)} \quad \Delta_-(u_j \times v_j) = u_j \times \Delta_- v_j + \Delta_- u_j \times v_{j-1},$$

where Δ_+, Δ_- *denote the forward and backward differences respectively.*

We use the difference method to prove the local existence of solution of (I.6)–(I.7). For simplicity, we let $\varepsilon = 1$. We establish the following difference-differential equation:

$$\frac{d\mathbf{Z}_j}{dt} = \frac{\Delta_- \left(G\left(\frac{\Delta_+ \mathbf{Z}_j}{h} \right) \frac{\Delta_+ \mathbf{Z}_j}{h} \right)}{h} + \frac{\Delta_- \left(\mathbf{Z}_j \times G\left(\frac{\Delta_+ \mathbf{Z}_j}{h} \right) \frac{\Delta_+ \mathbf{Z}_j}{h} \right)}{h}, \tag{II.1}$$

$$\mathbf{Z}_j |_{t=0} = \mathbf{Z}_{0j} = \mathbf{Z}_0(jh), \tag{II.2}$$

$$\mathbf{Z}_{j+J} = \mathbf{Z}_j, \tag{II.3}$$

where $j = 0, \pm 1, \ldots, \pm J, \ldots,\ h = 2D/J,\ J > 0$.

It is clear that the initial value problem of ordinary differential equation (II.1)–(II.3) admits a local smooth solution. We shall give some estimates uniformly in h for such a solution, and then get the local solution to problems (I.6)–(I.7). In this section we always denote the solution of (II.1)–(II.3) by $\mathbf{Z}_j$.

Lemma II.4: If $\mathbf{Z}_0(x) \in W^{1,4}(\Omega)$, *there are constants* $T_0 > 0, C > 0$ *independent of h such that*

$$\sup_{0 \leqslant t \leqslant T_0} (\| \mathbf{Z}_h(t) \|_2 + \| \delta \mathbf{Z}_h(t) \|_2 + \| \delta \mathbf{Z}_h(t) \|_4) \leqslant C, \tag{II.4}$$

$$\int_0^t \int_\Omega \| \delta^2 \mathbf{Z}_h(t) \|_2 \leqslant C. \tag{II.5}$$

Proof: Multiplying (II.1) by $\mathbf{Z}_j h$ and summing from $j = 1$ to J, we have

$$\frac{1}{2}\frac{d}{dt}\sum_{j=1}^{J}|\mathbf{Z}_j|^2h=-\sum_{j=0}^{J-1}G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2h+\sum_{j=0}^{J-1}\left(\mathbf{Z}_j\times G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)\cdot\frac{\Delta_+\mathbf{Z}_j}{h}h$$
$$=-\sum_{j=0}^{J-1}\left(A+B\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2\right)\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2h.$$

Therefore, we have

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{Z}_h\|_2^2+A\|\delta\mathbf{Z}_h\|_2^2+B\|\delta\mathbf{Z}_h\|_4^4=0. \tag{II.6}$$

It follows from (II.1) that

$$\frac{d\Delta_+\mathbf{Z}_j}{dt}=\frac{\Delta_+\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}+\frac{\Delta_+\Delta_-\left(\mathbf{Z}_j\times G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}.$$

Multiply this equation by $G(\Delta_+\mathbf{Z}_j)/h)(\Delta_+\mathbf{Z}_j/h)$ and summing from $j=0$ to $j=J-1$ to give

$$\sum_{j=0}^{J-1}G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\frac{d\Delta_+\mathbf{Z}_j}{dt}=\sum_{j=0}^{J-1}\frac{\Delta_+\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}$$
$$+\sum_{j=0}^{J-1}\frac{\Delta_+\Delta_-\left(\mathbf{Z}_j\times G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}$$
$$=-\sum_{j=1}^{J}\left|\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}\right|^2h$$
$$-\sum_{j=1}^{J}\frac{\Delta_-\left(\mathbf{Z}_j\times G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}h. \tag{II.7}$$

We claim

$$\sum_{j=1}^{J}\frac{\Delta_-\left(\mathbf{Z}_j\times G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}h=0. \tag{II.8}$$

In fact, we have from Lemma II.3(v) that

$$\sum_{j=1}^{J}\frac{\Delta_-\left(\mathbf{Z}_j\times G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}h$$
$$=\sum_{j=1}^{J}\left(\frac{\Delta_-\mathbf{Z}_j}{h}\times G\left(\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right)\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right)\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}h$$

$$=\sum_{j=1}^{J}\left(\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\times G\left(\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right)\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right)\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}h=0,$$

since $\mathbf{a}\times\mathbf{a}=0$, $(\mathbf{a}\times\mathbf{b})\cdot\mathbf{b}=0$ and $\Delta_-\mathbf{Z}_j=\Delta_+\mathbf{Z}_{j-1}$. The claim is proved.

We have from (II.7)–(II.8) that

$$\frac{1}{2}A\frac{d}{dt}\sum_{j=0}^{J-1}\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2h+\frac{1}{4}B\frac{d}{dt}\sum_{j=0}^{J-1}\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^4h+\sum_{j=1}^{J}\left|\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}\right|^2h=0. \tag{II.9}$$

However, it follows from the definition of $G(\xi)$ and Lemma II.3 that

$$\begin{aligned}
\left|\frac{\Delta_-\left(G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\mathbf{Z}_j}{h}\right)}{h}\right|^2 &= \left|G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}+B\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\,\frac{\Delta_-\left(\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2\right)}{h}\right|^2\\
&=\Bigg|G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\\
&\quad+B\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\cdot\left(\frac{\Delta_+\mathbf{Z}_j}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}+\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right)\Bigg|^2\\
&=G^2\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2\\
&\quad+B^2\left|\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right|^2\left|\frac{\Delta_+\mathbf{Z}_j}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}+\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2\\
&\quad+2BG\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\\
&\quad\cdot\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\left(\frac{\Delta_+\mathbf{Z}_j}{h}\cdot\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}+\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\cdot\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right)\\
&=G^2\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2\\
&\quad+B^2\left|\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right|^2\left|\frac{\Delta_+\mathbf{Z}_j}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}+\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2\\
&\quad+2BG\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right|^2\\
&\quad+2BG\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left(\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right)\left(\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_j}{h}\right).
\end{aligned} \tag{II.10}$$

Since

$$G^2\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2=\left(A^2+2AB\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2+B^2\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^4\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2, \tag{II.11}$$

and

$$\begin{aligned}&\left|2BG\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left(\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right)\left(\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_j}{h}\right)\right|\\&\quad\leqslant BG\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left(\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right|^2+\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2\right)\\&\quad=BG\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\cdot\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right|^2+\left(AB\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2+B^2\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^4\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2,\end{aligned} \tag{II.12}$$

we have from (II.9)–(II.12) that

$$\begin{aligned}&A\frac{d}{dt}\sum_{j=0}^{J-1}\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2h+\frac{1}{2}B\frac{d}{dt}\sum_{j=0}^{J-1}\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^4h+A^2\sum_{j=1}^{J}\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2h\\&\quad+AB\sum_{j=1}^{J}\left|\frac{\Delta_+\mathbf{Z}_j}{h}\right|^2\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2h+B\sum_{j=1}^{J}G\left(\frac{\Delta_+\mathbf{Z}_j}{h}\right)\left|\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\,\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right|^2h\\&\quad+B^2\sum_{j=1}^{J}\left|\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\right|^2\left|\frac{\Delta_+\mathbf{Z}_j}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}+\frac{\Delta_+\mathbf{Z}_{j-1}}{h}\,\frac{\Delta_+\Delta_-\mathbf{Z}_j}{h^2}\right|^2h\leqslant 0.\end{aligned} \tag{II.13}$$

This inequality combined with (II.6) leads to

$$\frac{d}{dt}(\|\mathbf{Z}_h\|_2^2+\|\delta\mathbf{Z}_h\|_2^2+\|\delta\mathbf{Z}_h\|_4^4)+\|\mathbf{Z}_h\|_2^2+\|\delta\mathbf{Z}_h\|_2^2+\|\delta\mathbf{Z}_h\|_4^4+\|\delta^2\mathbf{Z}_h\|_2^2\leqslant 0. \tag{II.14}$$

Inequality (II.14) combined with Gronwall inequality implies that there exist constants $T_0,C>0$ independent of h such that

$$\|\mathbf{Z}_h(t)\|_2+\|\delta\mathbf{Z}_h(t)\|_2+\|\delta\mathbf{Z}_h(t)\|_4^4\leqslant C,\quad\forall t\in[0,T_0],$$

$$\int_0^{T_0}\|\delta^2\mathbf{Z}_h(t)\|_2^2\leqslant C.$$

Lemma II.4 is proved. □

Corollary II.1: Under the conditions in Lemma II.4, we have, for some constant C independent of h,

$$\sup_{0\leqslant t\leqslant T_0;1\leqslant j\leqslant J}|\mathbf{Z}_j|\leqslant C. \tag{II.15}$$

Now we have the local existence of the solution to (I.6)–(I.7).

Theorem II.1: *Let* $\varepsilon=1$, $\mathbf{Z}_0(x)\in W^{1,4}(\Omega)$. *Then (I.6)–(I.7) admits a local solution* $\mathbf{Z}(x,t)$ *in* $[0,T_0]$ *in the space*

$$\mathbf{Z}(x,t)\in L^\infty(0,T_0;W^{1,4}(\Omega))\cap L^2(0,T_0;H^2(\Omega)), \tag{II.16}$$

and the following estimates hold:

$$\sup_{0\leqslant t\leqslant T_0;x\in\Omega}|\mathbf{Z}|\leqslant C, \tag{II.17}$$

$$\|\mathbf{Z}(t)\|_2+\|\mathbf{Z}_x(t)\|_2+\|\mathbf{Z}_x\|_4^4\leqslant C,\quad \forall t\in[0,T_0], \tag{II.18}$$

$$\int_0^{T_0}\|\mathbf{Z}_{xx}(t)\|_2^2\leqslant C. \tag{II.19}$$

In order to prove the global existence, we need the following a priori estimates for the solution of (I.6)–(I.7).

Lemma II.5: Let $\varepsilon=1$, $\mathbf{Z}_0(x)\in W^{1,4}(\Omega)$, $T>0$ *and* $\mathbf{Z}(x,t)\in L^\infty(0,T;W^{1,4}(\Omega))\cap L^2(0,T;H^2(\Omega))$ *is a solution of (I.6)–(I.7). Then the following estimates hold*:

$$\sup_{0\leqslant t\leqslant T;1\leqslant j\leqslant J}|\mathbf{Z}|\leqslant C, \tag{II.20}$$

$$\|\mathbf{Z}(t)\|_2+\|\mathbf{Z}_x(t)\|_2+\|\mathbf{Z}_x\|_4^4\leqslant C,\quad \forall t\in[0,T], \tag{II.21}$$

$$\int_0^{T}\|\mathbf{Z}_{xx}(t)\|_2^2\leqslant C. \tag{II.22}$$

Proof: Multiplying (I.6) by $\mathbf{Z}(x,t)$ and integrating it over Ω, we get

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{Z}(\cdot,t)\|_2^2+A\|\mathbf{Z}_x\|_2^2+B\|\mathbf{Z}_x\|_4^4=0. \tag{II.23}$$

Differentiating (I.6) with respect to x and then testing it by $G(\mathbf{Z}_x)\mathbf{Z}_x$, one has

$$G(\mathbf{Z}_x)\mathbf{Z}_x\mathbf{Z}_{xt}=(G(\mathbf{Z}_x)\mathbf{Z}_x)_{xx}G(\mathbf{Z}_x)\mathbf{Z}_x+(G(\mathbf{Z}_x)\mathbf{Z}\times\mathbf{Z}_x)_{xx}G(\mathbf{Z}_x)\mathbf{Z}_x.$$

Integrating this equation by parts, we have

$$\frac{A}{2}\frac{d}{dt}\|\mathbf{Z}_x\|_2^2+\frac{B}{4}\frac{d}{dt}\|\mathbf{Z}_x\|_4^4+\int_\Omega|(G(\mathbf{Z}_x)\mathbf{Z}_x)_x|^2=-\int_\Omega(G(\mathbf{Z}_x)\mathbf{Z}\times\mathbf{Z}_x)_x(G(\mathbf{Z}_x)\mathbf{Z}_x)_x$$

which implies

$$\frac{1}{2}A\frac{d}{dt}\|\mathbf{Z}_x\|_2^2+\frac{1}{4}B\frac{d}{dt}\|\mathbf{Z}_x\|_4^4+\int_\Omega|(G(\mathbf{Z}_x)\mathbf{Z}_x)_x|^2=0. \tag{II.24}$$

Since

$$\begin{aligned}|(G(\mathbf{Z}_x)\mathbf{Z}_x)_x|^2&=|G(\mathbf{Z}_x)\mathbf{Z}_{xx}+2B(\mathbf{Z}_x\cdot\mathbf{Z}_{xx})\mathbf{Z}_x|^2=G^2(\mathbf{Z}_x)|\mathbf{Z}_{xx}|^2+4B^2|(\mathbf{Z}_x\cdot\mathbf{Z}_{xx})\mathbf{Z}_x|^2\\&\quad+4BG(\mathbf{Z}_x)|\mathbf{Z}_x\cdot\mathbf{Z}_{xx}|^2\geqslant A^2|\mathbf{Z}_{xx}|^2,\end{aligned}$$

where C depends only on $\|\mathbf{Z}_0\|_{W^{1,4}(\Omega)}$, it follows from (II.24) that

$$\frac{1}{2}A\frac{d}{dt}\|\mathbf{Z}_x\|_2^2+\frac{1}{4}B\frac{d}{dt}\|\mathbf{Z}_x\|_4^4+A^2\int_\Omega|\mathbf{Z}_{xx}|^2\leqslant 0.$$

Putting this inequality and (II.23) together, we get from the Gronwall inequality that (II.21) and (II.22) hold. (II.20) can be derived from these inequalities. Lemma II.5 is proved. □

Now, we can use the extension method to give a global solution. Repeating for general ε, we have the following.

Theorem II.2: *Let $\varepsilon>0$ be fixed and $\mathbf{Z}_0\in W^{1,4}(\Omega)$. Then problem (I.6)–(I.7) admits a global solution $\mathbf{Z}_\varepsilon(x,t)$ in the space*

$$\mathbf{Z}_\varepsilon(x,t)\in L^\infty(0,\infty;W^{1,4}(\Omega))\cap L^2(0,\infty;H^2(\Omega)), \tag{II.25}$$

and the following estimates hold:

$$\sup_{0\leqslant t<\infty;x\in\Omega}|\mathbf{Z}_\varepsilon(x,t)|\leqslant C_1, \tag{II.26}$$

$$\|\mathbf{Z}_\varepsilon(t)\|_2+\|\mathbf{Z}_{\varepsilon x}(t)\|_2+\|\mathbf{Z}_{\varepsilon x}(t)\|_4^4\leqslant C_1, \quad \forall t\in[0,\infty), \tag{II.27}$$

$$\|\mathbf{Z}_{\varepsilon t}\|_{L^{4/3}(0,\infty;W^{-1,4/3}(\Omega))}\leqslant C_1, \tag{II.28}$$

$$\int_0^\infty\|\mathbf{Z}_{\varepsilon xx}(t)\|_2^2\leqslant C_\varepsilon, \tag{II.29}$$

where C_1 depends only on $\|\mathbf{Z}_0\|_{W^{1,4}(\Omega)}$.

Remark: The fact that C_1 *is independent of* ε *can be seen from the proof of Lemma II. 5, but* C_ε *depends on* ε.

III. MEASURE–VALUED SOLUTION TO THE STRONGLY DEGENERATE EQUATIONS

Since we can only get the uniform estimates (in ε) (II.26)–(II.28) for the solutions of the viscosity equations (I.6) and these estimates are not enough to obtain the weak solution for (I.4)–(I.5), we apply the notion of the measure-valued solution as in Ref. 7.

Denote $M=C_1$ where C_1 is given in Theorem II.2 which depends only on $\|\mathbf{Z}_0\|_{W^{1,4}(\Omega)}$. Let $\mathcal{Z}^\varepsilon=(\mathbf{Z}_\varepsilon,\mathbf{Z}_{\varepsilon x})$, $\tau(\xi)=B|(\xi_4,\xi_5,\xi_6)|^2(\xi_1,\xi_2,\xi_3)\times(\xi_4,\xi_5,\xi_6):(\mathbb{R}^3\cap\{|(\xi_1,\xi_2,\xi_3)|\leqslant M\})\times\mathbb{R}^3\to\mathbb{R}^3$. Then $\mathcal{Z}^\varepsilon$ is uniformly bounded in $L^4(Q)^6$ where $Q=\Omega\times(0,T)\in\mathbb{R}^2$ and

$$|\tau(\xi)|\leqslant CM(1+|\xi|)^3,\forall\xi\in(\mathbb{R}^3\cap\{|(\xi_1,\xi_2,\xi_3)|\leqslant M\})\times\mathbb{R}^3.$$

The following Lemma can be proved by the same method as in Ref. 7.

Lemma III.1: Let $Q\subset\mathbb{R}^2$ be a bounded open set. Let $\mathcal{Z}^{\varepsilon_n}$ be uniformly bounded in $L^4(Q)^6$. Then there exists a subsequence, still denoted by $\mathcal{Z}^{\varepsilon_n}$, and a measure-valued function ν such that for all $\tau:(\mathbb{R}^3\cap\{|(\xi_1,\xi_2,\xi_3)|\leqslant M\})\times\mathbb{R}^3\to\mathbb{R}^3$ satisfying for some $q>0$ the growth condition

$$|\tau(\xi)|\leqslant C(1+|\xi|)^3, \quad \forall\xi\in(\mathbb{R}^3\cap\{|(\xi_1,\xi_2,\xi_3)|\leqslant M\})\times\mathbb{R}^3,$$

we have

$$\tau(\mathcal{Z}^{\varepsilon_n})\rightharpoonup\overline{\tau} \text{ weakly in } L^r(Q),$$

$$\overline{\tau}(y)=<\nu_y,\tau>a.e.y\in Q,$$

provided that

$$1<r\leqslant\frac{4}{3}.$$

Proof: It suffices to verify the condition (2.7) of Theorem 2.1 of Ref. 7. Taking the Young function $\Psi(u)=u^r$, we have

$$\int_Q \Psi(\tau|\mathcal{Z}^{\varepsilon_n}|) = \int_Q |\mathcal{Z}^{\varepsilon_n}|^r \leqslant C^r \int_Q (1+|\mathcal{Z}^{\varepsilon_n}|)^{3r},$$

and the last term is uniformly bounded (with respect to n) if $3r \leqslant 4$. The lower bound $r>1$ follows from the properties of Orlicz functions, namely from $\lim_{s\to\infty}\Psi(s)/s=\infty$. □

Definition: A pair $(\mathbf{Z},\nu)$ is called a measure-valued solution of (I.4)–(I.5) if

$$\mathbf{Z}\in L^{\infty}(0,\infty;W^{1,4}(\Omega)), \tag{III.1}$$

$$\nu\in L^{\infty}_{\omega}(Q;\mathrm{Prob}(\mathbb{R}^6)), \tag{III.2}$$

and if for any $\varphi\in\mathcal{D}(-\infty,T;C^{\infty}_{\mathrm{per}}(\Omega))$ there holds

$$\int_Q \mathbf{Z}_0\varphi = \int_Q \mathbf{Z}\varphi_t - A\int_Q \varphi_x \mathbf{Z}\times\mathbf{Z}_x - \int_Q \varphi_x \int_{\mathbf{R}^6} \tau(\lambda)\, d\nu_{t,x}(\lambda)\, dx\, dt, \tag{III.3}$$

where τ is defined as above, $Q=\Omega\times(0,T)$.

Theorem III.1: *Let $\mathbf{Z}_0\in W^{1,4}(\Omega)$. Then problem (I.4)–(I.5) admits a measure-valued solution.*

Proof: It follows from Lemma III.1 that there exists a subsequence $\mathcal{Z}^{\varepsilon_n}$ and a measure-valued function ν such that

$$\tau(\mathcal{Z}^{\varepsilon_n}) \rightharpoonup \overline{\tau} \text{ weakly in } L^r(Q), \quad 1<r\leqslant \frac{4}{3}, \tag{III.4}$$

$$\overline{\tau}(x,t) = \langle \nu_{t,x}, \tau\rangle, \quad a.e.\,(x,t)\in Q. \tag{III.5}$$

To finish the proof, we only need to prove, for some subsequence $\mathbf{Z}_{\varepsilon_n}$, that

$$\int_Q \mathbf{Z}_{\varepsilon_n t}\varphi \to -\int_Q \mathbf{Z}_t\varphi + \int_\Omega \mathbf{Z}_0\varphi(x,0), \tag{III.6}$$

$$\int_Q (\mathbf{Z}_{\varepsilon_n}\times\mathbf{Z}_{\varepsilon_n x})\varphi_x \to \int_Q (\mathbf{Z}\times\mathbf{Z}_x)\varphi_x, \tag{III.7}$$

$$Z^i_x = \int_{\mathbf{R}^6} \lambda_{i+3}\, d\nu_{t,x}(\lambda), \quad a.e.\,(x,t)\in Q, \quad i=1,2,3. \tag{III.8}$$

In view of (II.27), we have for some subsequence $\mathbf{Z}_{\varepsilon_n}$ that

$$\mathbf{Z}_{\varepsilon_n} \rightharpoonup \mathbf{Z} \text{ weakly in } L^r(0,T;L^2(\Omega)), \quad \forall r>1 \tag{III.9}$$

$$\mathbf{Z}_{\varepsilon_n x} \rightharpoonup \mathbf{Z}_x \text{ weakly in } L^r(0,T;L^4(\Omega)), \quad \forall r>1. \tag{III.10}$$

To prove (III.6), we take $\varphi\in\mathcal{D}(-\infty,T;C^{\infty}_{\mathrm{per}}(\Omega))$ to give

$$\int_0^T\int_\Omega \mathbf{Z}_{\varepsilon_n t}\varphi = -\int_0^T\int_\Omega \mathbf{Z}_{\varepsilon_n}\varphi_t + \int_\Omega \mathbf{Z}_0(x)\varphi(x,0) \to -\int_0^T\int_\Omega \mathbf{Z}\varphi_t + \int_\Omega \mathbf{Z}_0(x)\varphi(x,0);$$

this proves (III.6).

Now we prove (III.7). It follows from (II.27) and (II.28) that $\mathbf{Z}_{\varepsilon_n}$ is uniformly bounded in the space

$$\{v:v\in L^r(0,T;W^{1,4}_{\mathrm{per}}(\Omega)),v_t\in L^{4/3}(0,T;(W^{1,4}(\Omega))^*)\},\tag{III.11}$$

for any $r>1$. Since $W^{1,4}_{\mathrm{per}}(\Omega)\hookrightarrow\hookrightarrow L^r(\Omega)\hookrightarrow(W^{1,4}(\Omega))^*$, applying an Aubin–Lions Lemma[7] with $X_0=W^{1,4}_{\mathrm{per}}(\Omega)$, $X=L^r(\Omega)$, $X_1=(W^{1,4}(\Omega))^*$, $\alpha=r$ $(r>1)$, $\beta=4/3$, we know that the space defined in (III.11) is compactly imbedded into $L^r(0,T;L^r(\Omega))$, that is,

$$\mathbf{Z}_{\varepsilon_n}\to\mathbf{Z}\ \text{ strongly in }\ L^r(0,T;L^r(\Omega)).\tag{III.12}$$

Since

$$\int_Q((\mathbf{Z}_{\varepsilon_n}\times\mathbf{Z}_{\varepsilon_n x})\varphi_x-(\mathbf{Z}\times\mathbf{Z}_x)\varphi_x)=\int_Q((\mathbf{Z}_{\varepsilon_n}-\mathbf{Z})\times\mathbf{Z}_{\varepsilon_n x})\varphi_x+\int_Q(\mathbf{Z}\times(\mathbf{Z}_{\varepsilon_n x}-\mathbf{Z}_x))\varphi_x=I_1+I_2,$$

it follows from (II.26) and (III.10) that

$$I_2\to0\ \text{ as }\ n\to\infty,$$

and it follows from (II.27) and (III.12) that

$$|I_1|\le\left(\int_0^T\int_\Omega|\mathbf{Z}_{\varepsilon_n}-\mathbf{Z}|^4\right)^{1/4}\left(\int_0^T\int_\Omega|\varphi_x|^4\right)^{1/4}\left(\int_0^T\int_\Omega|\mathbf{Z}_{\varepsilon_n x}|^2\right)^{1/2}$$
$$\le C\left(\int_0^T\int_\Omega|\mathbf{Z}_{\varepsilon_n}-\mathbf{Z}|^4\right)^{1/4}\to0\ \text{ as }\ n\to\infty.$$

The proof of (III.7) is complete.

Since Lemma III.1 is true for all τ, if we let $\tau=\mathrm{Id}$, then for $r=4$, $q=1$, $\forall\psi\in L^{4/3}(Q)$, we have for $\mathcal{Z}^{\varepsilon_n}=(\mathbf{Z}_{\varepsilon_n},\mathbf{Z}_{\varepsilon_n x})$ that

$$\int_Q\mathcal{Z}^{\varepsilon_n}\psi\,dx\,dt\to\int_Q\psi\int_{\mathbf{R}^6}\lambda\,d\nu_{t,x}(\lambda)dx\,dt.$$

However, $\mathbf{Z}_{\varepsilon_n}\to\mathbf{Z}$ strongly in $L^r(Q)$ and $\mathbf{Z}_{\varepsilon_n x}\rightharpoonup\mathbf{Z}_x$ in $L^4(Q)$, we know

$$Z_x^i=\int_{\mathbf{R}^6}\lambda_{i+3}\nu_{t,x}(\lambda),\quad a.e.(x,t)\in Q,\quad i=1,2,3.$$

This verifies (III.8). □

Remark: Since the estimates obtained in section II are independent of D, we get by letting $D\to\infty$ that the Cauchy problem of (I.4) admits a solution in $L^\infty(\mathbb{R}_+;W^{1,4}(\mathbb{R}^1))\cap L^2(\mathbb{R}_+;H^2(\mathbb{R}^1))$.

[1] J. Fivez ''On the continuum limit of a classical compressible Heisenberg chain,'' Phys. C. Solid State Phys. **15**, L641–L643 (1982).

[2] E. Magyari, ''Solitary waves along the compressible Heisenberg chain,'' Phys. C. Solid State Phys. **15**, L1159–L1163 (1982).

[3] R. Balakrishnan, ''On the inhomogeneous Heisenberg chain,'' Phys. C: Solid State Phys. **15**, L1305–L1308 (1982).

[4] Y. L. Zhou, B. L. Guo, and S. Tan, ''Existence and uniqueness of smooth solution for system of ferromagnetic chain,'' Sci. Sinica A **34**, 157–166 (1991).

[5] S. Ding and B. L. Guo, ''Smooth solution for the 1D Heisenberg chain equations,'' Proc. R. Soc. Edinburgh A (to appear).

[6] Y. L. Zhou, J. C. M. **2**, 276–281 (1984).

[7] J. Malek, J. Necas, M. Rokyta, and M. Ruzicka, *Weak and Measure-Valued Solutions to Evolutionary PDEs* (Chapman & Hill, 1996).

Rectangular well as perturbation

Mariusz Dudek, Stefan Giller,[a] and Piotr Milczarski[b]
Theoretical Physics Department II, University of Łódź, Pomorska 149/153, 90-236 Łódź, Poland

(Received 13 March 1998; accepted for publication 10 November 1998)

We discuss a finite rectangular well of a depth λ^2 as a perturbation for the infinite one with λ as a perturbation parameter. In particular, we consider a behavior of energy levels in the well as functions of complex λ. It is found that all the levels of the same parity are defined on infinitely sheeted Riemann surfaces whose topological structures are described in detail. These structures differ considerably from those found in models investigated earlier. It is shown that perturbation series for all the levels converge what is in a contrast with the known results of Bender and Wu. The last property is shown to hold also for the infinite rectangular well with the Dirac delta barrier as a perturbation considered earlier by Ushveridze. © *1999 American Institute of Physics.* [S0022-2488(99)03103-5]

I. INTRODUCTION

Since the papers of Bender and Wu[1] we have known why the perturbation series were, in general, divergent. We have known also that in many cases investigations of perturbation series could be reduced to the investigations of the corresponding semiclassical series.[2] It was also realized that the divergent perturbation series could be summed and one of the summation methods applied here was very often the Borel one.[2,3]

One of the byproducts of these investigations was a discovery of so called level crossing, i.e., of the fact that in the case of confining polynomial potentials all the discrete energy levels they produce or only groups of them are no longer isolated of each other if considered as functions of a perturbation parameter.[1,3–6] Being a little bit more precise, the latter statement means that the levels inside each group appear as branches of ramified functions of the perturbation parameter considered as a complex variable. It means, in particular, that each energy level belonging to a group considered as a function of a real perturbation parameter can be analytically continued into the complex plane of the parameter so that any energy level of the group can be reached by the analytic continuation procedure of some arbitrary chosen level belonging to the group. This means also that the complex plane of the perturbation parameter converts rather into some (more or less) complicated Riemann surface.

It is also well known[7] that it is an existing symmetry group of the Hamiltonian considered which is completely responsible for a decay of the energy spectrum into disjoint (with respect to analytic continuation) groups of them. Therefore, a degree of complication of respective Riemann surfaces on which the energy levels are defined can give us information about an existence of the relevant symmetry group, i.e., the more levels appear as branches defined on respective sheets of the same Riemann surface the less rich a relevant symmetry group has to be. In particular, if the corresponding Riemann surfaces are all finitely sheeted (i.e., if there are finite numbers of energy levels attached to each of them) then they have to be defined by some algebraic conditions relating energies and a perturbation parameter, with the conditions being a clear sign of the existence of an underlying symmetry group.[7]

There are only a few examples of the analysis described above in the case when the relevant Riemann surface is infinitely sheeted.[1,6,8] The analysis performed is more or less numerical. This

[a]Electronic mail: sgiller@krysia.uni.lodz.pl
[b]Electronic mail: jezykmil@krysia.uni.lodz.pl

is mostly because it is very difficult to find an example of potential providing us with a closed functional form of the quantization condition being simultaneously sufficiently simple to perform an analysis in a ''classical,'' non-numerical way. According to these investigations a Riemann surface topology corresponding to a given perturbation parameter, i.e., loci of its branch points seems to be still mysterious and depending on a parameter chosen so needing still further studies.

In this paper we will consider a possibly simple but not a trivial example of a Hamiltonian provided by the familiar rectangular well of a finite height which allows us for such a classical analysis. As a perturbation parameter in this example we will choose its height $V=\lambda^2>0$; strictly speaking the square root of it. The perturbed potential is then the infinite rectangular well approached when $V^{-1}\rightarrow 0$.

There are several basic properties which differ the case of the finite rectangular well from the ones considered earlier. First it is just a finite number of energy levels existing for a given λ but varying with λ so that a potentially infinite number of levels can appear when $\lambda^{-1}\rightarrow 0$. The remaining properties are enumerated as points 2, 4 and 5. below.

The rectangular well is not an analytical potential and as such it provides us also with a nonanalytical quantization condition. However, an analytical extension of the latter into complex values of the quantities considered is possible and results of the relevant analysis are the following.

1. The system of energy levels of the well decays into two disjoint families (of different parities) with the levels inside each of the group being analytical continuations of each other with respect to the perturbation parameter.

2. The perturbation series for each level is convergent to the level itself, i.e., the property which is quite opposite to that of Bender and Wu for the unharmonic oscillator case. As such they are trivially Borel summable.

3. The Riemann surfaces for both the groups of levels are infinitely sheeted and their branch point structures can be understood by some simple properties of both the quantization conditions.

4. The energy level poles existing in the complex momentum plane corresponding to the case are accompanied by poles which do not represent discrete nor resonant parts of the energy spectrum [in particular, because resonances are absent in the case of the finite rectangular well independently of whether the latter is a real well (for real λ) or is a rectangular barrier (for imaginary λ)]. These second sort of poles we shall call pseudoenergy levels.

5. The level crossing which happens for *real* λ is not between two real energies but just between an energy and its pseudoenergy partner.

We will reconsider also an example of a numerical analysis performed earlier by Ushveridze[8] to show that its ''classical'' analysis is possible in its full size confirming the main results of the author mentioned but completed them with such an important conclusion as a convergence of both perturbation series corresponding to the weak and the strong couplings to the Dirac delta perturbation used in the example.

Another goal of our investigations was to look for rules governing distributions of energy level branch points on the perturbation parameter Riemann surface. It has been demonstrated by Ushveridze[8] that one can predict an existence of energy level crossing as well as arrange the crossing to join more than two levels but as in the case of the anharmonic potential[1] distributions of the corresponding branch points seemed to be unpredictable prior to direct calculations.

Our main conclusion in this respect is similar. Indeed, such distributions of branch points is rather strongly model (potential) dependent although it can happen that some properties of these distributions can be drawn just by an inspection of properties of investigated potentials. However, in this way only some crude procedure can be formulated to predict their presence or absence in some domains of the underlying Riemann surface (see also Ref. 2).

II. FINITE RECTANGULAR WELL AS A PERTURBATION

A. Analytic properties of energy levels as functions of perturbation parameter

A finite rectangular well as a perturbation sounds a little bit exotic but it can be considered as such for the infinite will in the same way as, for example, a finite potential well:

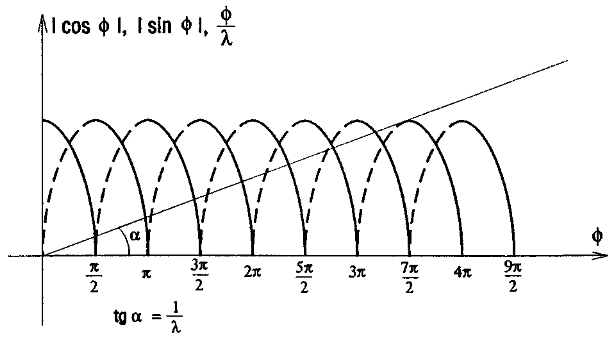

FIG. 1. The ϕ angles chosen by the quantization conditions (3).

$$\frac{1}{2}\frac{\omega^2}{\alpha^2}\left(1-\frac{1}{\cosh(\alpha x)}\right) \tag{1}$$

is a perturbation for the harmonic one, i.e., we get the latter from (1) for $\alpha\rightarrow 0$.

In the case of a finite rectangular potential given by

$$V(x)=\begin{cases} V(>0), & \text{for } |x|>1,\\ 0, & \text{for } |x|<1, \end{cases} \tag{2}$$

a perturbation parameter can be chosen to be $\lambda^2=V$ which together with the rescaled energy $E/V\equiv z^2$ leads us to the following quantization conditions:[9]

$$\begin{aligned} &\lambda|\cos\phi|=\phi, \quad z=|\cos\phi|, \quad \tan\phi>0, \quad \text{for positive parity levels;}\\ &\lambda|\sin\phi|=\phi, \quad z=|\sin\phi|, \quad \tan\phi<0; \quad \text{for negative parity levels.} \end{aligned} \tag{3}$$

All (positive) solutions $\phi_k(\lambda)$, $k=1,2,\ldots$, to Eqs. (3) are represented picturesquely in Fig. 1 as given by the points of intersections of the straight line ϕ/λ with the right arcs of $|\sin(\phi)|$ and $|\cos(\phi)|$ functions. The corresponding solutions for energy levels are then obtained as $E_k(\lambda)=\lambda^2 z^2(\phi_k(\lambda))$. It is just a dependence of E_k's on λ as a complex parameter which is the main interest of this paper.

To investigate this dependence it is, however, necessary to make an analytical continuation of the conditions (3) into the complex λ. Such a continuation of Eqs. (3) is possible when dropping the absolute value marks in (3), which provides us with following analytic conditions:

$$\begin{aligned} &\lambda\cos\phi=\phi, \quad z=\cos\phi, \quad \text{for } 0\leqslant\phi\leqslant\tfrac{1}{2}\pi \pmod{2\pi};\\ &\lambda\cos\phi=-\phi, \quad z=-\cos\phi, \quad \text{for } \pi\leqslant\phi\leqslant\tfrac{3}{2}\pi \pmod{2\pi};\\ &\lambda\sin\phi=\phi, \quad z=\sin\phi, \quad \text{for } \tfrac{1}{2}\pi\leqslant\phi\leqslant\pi \pmod{2\pi};\\ &\lambda\sin\phi=-\phi, \quad z=-\sin\phi, \quad \text{for } \tfrac{3}{2}\pi\leqslant\phi\leqslant 2\pi \pmod{2\pi} \end{aligned} \tag{4}$$

equivalent to (3) for positive ϕ.

According to (4) the energy spectrum for the potential (2) is formally divided into four groups. Each group contains every fourth member of the spectrum, starting from E_1^+, E_3^+, E_2^-, E_4^- energy levels in the corresponding group. The energy levels in the groups are defined by the conditions (4) in the order mentioned.

Dropping further the restrictions for the ranges of changing ϕ in the conditions (4) we get fully analytical quantization conditions but describing rather different spectra with respect to which our original ones are only parts of them. This is, however, the necessary price for investigating the complex analytical dependence of energy levels on λ.

Let us observe, however, that for both the parity levels it is enough to continue analytically only the first of the corresponding conditions (4) depriving them of the corresponding restrictions for ϕ (i.e., ϕ can now take any complex value in these conditions). This is because corresponding solutions $\phi^{\pm}(\lambda)$ to the first conditions generate solutions $\phi^{\pm}(-\lambda)$ to the second ones. [In fact, for the even parity both the solutions almost coincide since $\phi^{+}(-\lambda)=-\phi^{+}(\lambda)$].

Therefore we see that all energy levels of both parities: $E_k^{+}(\lambda)=\lambda^2\cos^2\phi_k^{+}(\lambda)=(\phi_k^{+}(\lambda))^2$ and $E_k^{-}(\lambda)=\lambda^2\sin^2\phi_k^{-}(\lambda)=(\phi_k^{-}(\lambda))^2$ can be obtained by solving only the first conditions (4) of the corresponding parities and performing analytic continuations in λ from its positive to its negative values.

We shall analyze first the even parity group determining the ground state energy level. An analysis of the odd parity case is quite similar.

1. Even parity energy spectrum case

Making in the first of the conditions (4) a change of variable $e^{i\phi}=\sigma$ we get instead

$$\lambda^{+}=-\frac{2i\sigma\ln\sigma}{\sigma^2+1},\qquad z^{+}=\frac{1}{2}\left(\sigma+\frac{1}{\sigma}\right). \tag{5}$$

Since a dependence of z^{+} on σ as given by (5) is rather simple then to get the corresponding dependence of z^{+} on λ it is necessary to invert the dependence of the latter variable on σ as given by (5). To do this one needs to know the following:

1^0 The Riemann surface structure for $\lambda^{+}(\sigma)$;

2^0 The loci of all zeros of $\lambda^{+'}(\sigma)$ on the surface; and

3^0 A pattern of lines $\mathrm{Re}\,\lambda^{+}=\mathrm{const}$ and $\mathrm{Im}\,\lambda^{+}=\mathrm{const}$ on the surface.

As it follows from (5) the Riemann surface structure for $\lambda^{+}(\sigma)$ is determined by

a. The logarithmic branch point at $\sigma=0$; and

b. A pair of simple poles at $\sigma=\pm i$ located on every sheet the latter being generated in an infinite number by the logarithm.

Zeros of $\lambda^{+'}(\sigma)$ are determined by the following equation:

$$\ln\sigma=\frac{\sigma^2+1}{\sigma^2-1}. \tag{6}$$

Putting $e^{i\phi+y}=\sigma$ (σ is now an arbitrary complex number on the surface) and assuming that $\phi,y\neq 0$ we transform (6) into

$$\frac{\sin 2\phi}{2\phi}=-\frac{\sinh 2y}{2y},\qquad \cos 2\phi=-\frac{\sinh 2y}{2y}+\cosh 2y, \tag{7}$$

$$\phi,\ \ y\neq 0,$$

from which it follows that all zeros of $\lambda^{+'}(\sigma)$ have to lie on the circle $|\sigma|=1$ and/or on the real half axis $\phi=0$. Therefore the corresponding conditions for them are

$$\cot\phi=-\phi,\qquad \ln\sigma=\frac{\sigma^2+1}{\sigma^2-1}, \tag{8}$$

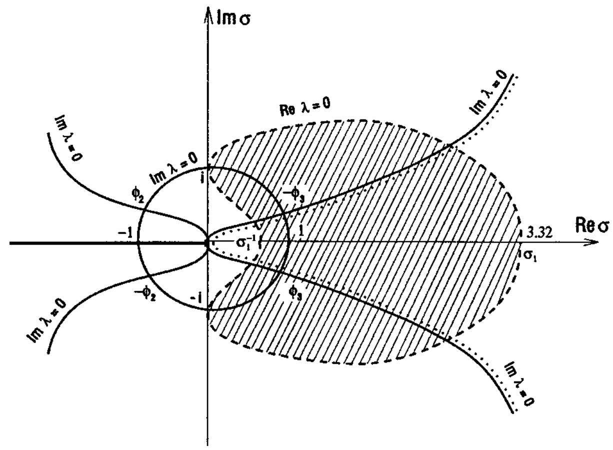

FIG. 2. The ''first'' sheet of the σ-Riemann surface corresponding to $\lambda^+(\sigma)$.

where in the last equation σ is real and positive.

Solutions to the first of Eqs. (8) (in fact, infinitely many of them) are given therefore as the intersection points of the functions $\cot\phi$ and $-\phi$. The form of the condition can be easily identified as the one for the straight line ϕ/λ of Fig. 1 to be tangent to $\cos\phi$. Therefore the points have to lie close to $\phi=2k\pi$ ones, $k=1,2,\ldots$, on the left to them, and close to $\phi=k\pi$, $k=-1,-3,\ldots$, on the right to the latter.

Two real solutions to the second of Eqs. (8) are placed on both the sides of $\sigma=1$ at $\sigma_1 \approx 3.32$ and at $\sigma_1^{-1}\approx 0.301$.

Let us note that the points on the unit circle at $\phi=k\pi, k=-1,+2,-3,+4,\ldots$, and the point $\sigma=1$ are all the physical ''thresholds'' for successive appearing of the corresponding even parity energy levels according to changing λ from zero to infinity. There is a temptation to understand the zeros provided by the conditions (8) as a shifting of these thresholds from their real physical positions mentioned to their actual ones because of the approximations which the analytical conditions (8) effectively are to our rectangular well quantization problem. This shifting of thresholds remains in a deep relation to analytical properties of reflection and transmission coefficients of the corresponding scattering problem arising when energy is higher than λ^2. We shall discuss this relation below.

Although the solutions to the condition (8) cannot be considered as thresholds for any real energy spectrum case (this is excluded by the absence of the energy level degeneracy in 1-dim SE) we shall consider them as such for convenience and call them pseudothresholds.

Thus the solutions to the first of the conditions (8) are the pseudothresholds for the energy levels lying inside the potential above the ground state one. These pseudothresholds as we have mentioned earlier coincide with results of demanding for the line ϕ/λ in Fig. 1 to be tangent to $\cos\phi$. The latter demand is just the lower limit for λ above which the real solutions for the higher energy levels can exist.

Both the solutions at σ_1 and σ_1^{-1} between which $\lambda^+(\sigma)$ is pure imaginary are then the corresponding limits for the ground state energy level E_1^+ to be a real quantity. However, for $E_1^+(\lambda)$ as well as for the higher energies the condition to be real is not enough to represent a real physical level. The sufficient conditions will be discussed below in Sec. II B.

A pattern of the lines $\mathrm{Re}\,\lambda^+=\mathrm{Im}\,\lambda^+=0$ is sketched on Fig. 2 together with the positions of all singular points of $\lambda^+(\sigma)$. Since all zeros of $\lambda^{+\prime}(\sigma)$ are simple they result as square root branch point singularities of $\sigma(\lambda)$ on its λ-Riemann surface. A consequence of this is a perpendicular crossing of two lines $\mathrm{Im}\,\lambda^+=0$ as well as the corresponding two lines $\mathrm{Re}\,\lambda^+=0$ in every such zero. All the lines $\mathrm{Im}\,\lambda^+=0$ emanating outside the unit circle run to infinity on the corresponding sheets of the σ-Riemann surface parallel to the line $\mathrm{Re}\,\sigma=0$. These emanating inside the circle have to cross the logarithmic branch point at $\sigma=0$ being tangent at this point to the line

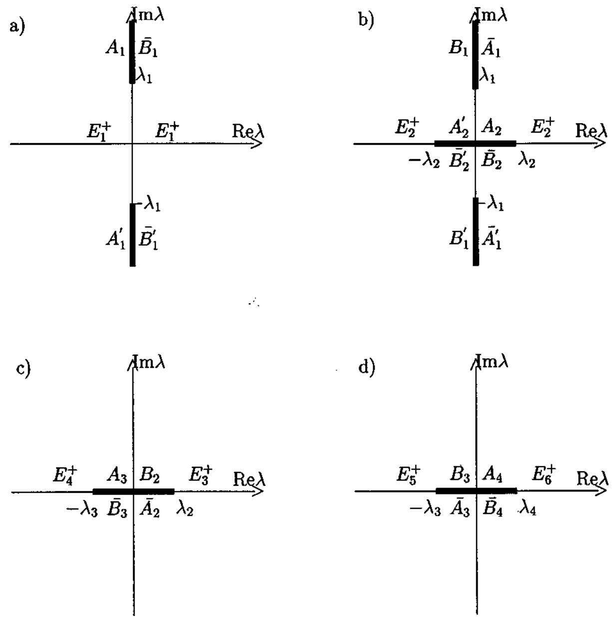

FIG. 3. The few first λ-Riemann surface sheets for $\sigma^+(\lambda)(E^+(\lambda))$.

Re $\sigma=0$. The remaining lines Im $\lambda^+=0$ coincide with all semicircles of the unit circle lying between the poles at $\sigma=\pm i$ on every logarithmic sheet of the σ-Riemann surface.

The λ-Riemann surface on which the inverse function $\sigma^+(\lambda)$ is defined is now easy to construct. It is also an infinitely sheeted surface. Its first sheet is a map of the dashed area in Fig. 2 and is shown in Fig. 3(a). This is the sheet on which the ground state energy level is defined, i.e., for λ changing along the real axis of the sheet or along a segment $[\lambda_1,-\lambda_1]\equiv[-\lambda^+(\sigma_1),\lambda^+(\sigma_1^{-1})]$ on the imaginary axis where the corresponding energy E_1^+ is real. On the rest of the sheet the energy is complex.

There are two cuts on the sheet emanating of the two complex conjugate imaginary branch points at $\lambda_1=\lambda^+(\sigma_1)=-\lambda^+(\sigma_1^{-1})$ and running to infinity along its imaginary axis. The function $\sigma^+(\lambda)$ is holomorphic on the sheet approaching $\pm i$ for λ escaping to the infinity on the right or on the left half planes of the sheet, respectively.

The second sheet corresponds to the second energy level. This sheet can be achieved by crossing (in any direction) one of the two cuts described above. Despite these two latter cuts there are another two square root branch points on the sheet lying on the real axis at $\lambda=\pm\lambda_2=\lambda^+(e^{\mp i\phi_2})=\mp\phi_2/\cos\phi_2$. The energy level E_2^+ is given by the values of λ changing on both the sides of the real axis from the real branch points mentioned up to the corresponding infinities. On the σ-Riemann surface of Fig. 2 these ranges of changing λ correspond to varying σ along the two left unit semicircles between the poles at $\sigma=\pm i$.

There is still an additional branch point on the discussed sheet at $\lambda=0$ which is also present at all the other sheets except the first one. It is a common picture of the point $\sigma=0$ and all the infinity points of the σ-Riemann surface of Fig. 2. Therefore this point gives rise to an infinite branching of $\sigma^+(\lambda)$ around it.

Starting from the third energy level, every one of the levels is represented on two different sheets which can communicate with themselves by the branch point at $\lambda=0$. Each of these two sheets belongs to two different families of them the latter being generated by the two branch points of the second sheet at $\lambda=\pm\lambda_2$ [see Fig. 3(b)]. One of these two families corresponds to mapping of the σ-Riemann surface of Fig. 2 into the λ one in the clockwise direction moving around the logarithmic branch point of $\lambda^+(\sigma)$ at $\sigma=0$. The second family of sheets in the corresponding mapping in the opposite direction.

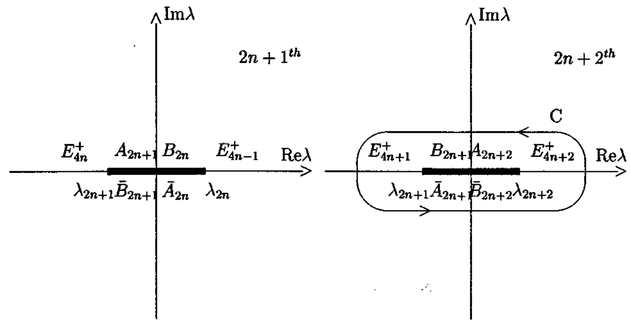

FIG. 4. The general structure of λ-Riemann surface sheets for $\sigma^+(\lambda)(E^+(\lambda))$.

Let us analyze the family generated by the branch point at $\lambda=\lambda_2$. The other family is a twin picture of this one obtained by a transformation $\lambda\to-\lambda$.

Going around λ_2 through the cut $A_2/\overline{B_2}$ we find ourselves on the sheet shown in Fig. 3(c). There is an additional cut on the sheet generated by the branch point at $\lambda=\lambda_3=\lambda^+(e^{-i\phi_3})=-\phi_3/\cos\phi_3$. The branch point opens the possibility for the fourth and the fifth energy levels to appear. The ranges of the energy levels E_3^+ and E_4^+ are shown on the figure.

To achieve a sheet corresponding to a pair E_5^+, E_6^+ of the levels we have to cross the cut $A_3/\overline{B_3}$ in Fig. 3(c) (in any direction). The sheet is shown in Fig. 3(d) together with corresponding ranges for the energies.

A full structure of the considered family of sheets of the λ-Riemann surface is now obvious. The n^{th} sheet, $n=3,4,\ldots,$ contains three branch points lying at $\lambda=0$, $\lambda=\lambda_{n-1}=-\phi_{n-1}/\cos\phi_{n-1}$ and at $\lambda=\lambda_n=-\phi_n/\cos\phi_n$. Each of the last two branch points opens a pair of energy levels: E^+_{2n-4}, E^+_{2n-3} and E^+_{2n-2}, E^+_{2n-1}, respectively, the levels in each pair lying on different sheets.

And inversely, the energy levels appear on every sheet in pairs. A $2n+1$th sheet corresponding to energy levels E^+_{4n-1}, E^+_{4n}, $n=1,2,\ldots$ (see Fig. 4) is cut by two cuts: $B_{2n}/\overline{A}_{2n}$ beginning at $\lambda=\lambda_{2n}=-\phi_{2n}/\cos\phi_{2n}>0$ and opening the level E^+_{4n-1} and $B_{2n+1}/\overline{A}_{2n+1}$ beginning at $\lambda=\lambda_{2n+1}=-\phi_{2n+1}/\cos\phi_{2n+1}<0$ and opening the level E^+_{4n}. Both the cuts end at $\lambda=0$. A corresponding sheet for the levels E^+_{4n+1}, E^+_{4n+2} is cut from λ_{2n+1} to 0 and from 0 to $\lambda=\lambda_{2n+2}=-\phi_{2n+2}/\cos\phi_{2n+2}>0$ with the latter branch point opening the second of the considered levels (see Fig. 4). Note also that a pair E^+_{4n-1}, E^+_{4n} is determined by the crossing line ϕ/λ with $4n-1$th and $4n$th arcs of $|\cos\phi|$ of Fig. 1, respectively, and a pair E^+_{4n+1}, E^+_{4n+2} with the arcs $4n+1$th, $4n+2$th correspondingly. This is why the members of each such a pair have to approach the same limit when $\lambda\to+\infty$.

The structure of the second family of sheets generated by the branch point at $\lambda=-\lambda_2$ is obtained by an inversion: $\lambda\to-\lambda$ from the first one discussed above.

2. Odd parity energy spectrum case

Making in the third of the conditions (4) a change of variable $\sigma=e^{i\phi}$ we get the condition in the form:

$$\lambda^-=\frac{2\sigma\ln\sigma}{\sigma^2-1},\quad z^-=\frac{1}{2i}\left(\sigma-\frac{1}{\sigma}\right). \tag{9}$$

It follows from (9) that $\lambda^-(\sigma)$ is a meromorphic function of σ on the σ-Riemann surface with the logarithmic branch point at $\sigma=0$ and with simple poles at $\sigma=\pm1$ on every logarithmic sheet except the first one where the pole at $\sigma=1$ is absent. The surface is shown in Fig. 5 where the lines $\mathrm{Im}\,\lambda^-(\sigma)=0$ are also shown schematically. Zeros of $\lambda^{-\prime}(\sigma)$ are distributed in this case only

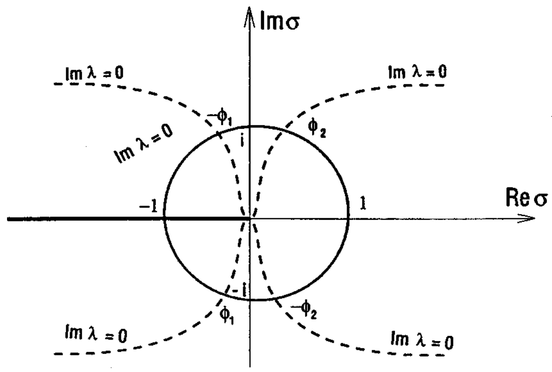

FIG. 5. The ''first'' sheet of the σ-Riemann surface corresponding to $\lambda^-(\sigma)$.

on the unit circle $|\sigma|=1$ at the point $\sigma=1$ on the first sheet and at points $\sigma=e^{i\phi_k}$ with $\phi_k(=-\phi_{-k}>0)$ satisfying the condition $\phi_{\pm k}=\tan\phi_{\pm k}$, $k=2,\ldots,$ on the remaining sheets so that the first two of the latter singular points lie on the second sheet of Fig. 5.

On the other hand, the inverse function $\sigma^-(\lambda)$ is holomorphic on the λ-Riemann surface with branch points at $\lambda=1$ and at $\lambda_k=\lambda^-(e^{i\phi_k})$ which are images of zeros of $\lambda^{-\prime}(\sigma)$.

The spectrum is opened with the level $E_2^-(\lambda)$ which belongs to the odd parity spectrum of the rectangular well contrary to the next level $E_3^-(\lambda)$ which does not.

An interesting property of $E_2^-(\lambda)$ is a singularity which it has to have at $\lambda=0$, i.e., the level $E_2^-(\lambda)$ as a function of λ is not bounded from below at this point. This conclusion follows from an observation that although the point $\lambda=1$ is a branch one for $\sigma^-(\lambda)$ [below which $\sigma^-(\lambda)$ has two values: σ and σ^{-1} for every λ, $0\leqslant\lambda\leqslant 1$] it is not as such for the level $E_2^-(\lambda)$. On the other hand, $E_2^-(\lambda)(=-\ln^2\sigma)$ for $0<\lambda<1$ $(0<\sigma<+\infty)$ is real and negative. Therefore a pseudothreshold for the latter appears to be at the point $\lambda=0$ at which $E_2^-(\lambda)$ becomes infinitely large and negative. Here again we want to stress, however, that the reality of $E_2^-(\lambda)$ does not mean automatically, its existence as a real physical energy level (see a discussion below, Sec. II C) and, in fact, in the case considered $E_2^-(\lambda)$ disappears as a physical level below $\lambda=\pi/2$ as it follows from Fig. 1.

Properties of the remaining levels are the following.

The level $E_3^-(\lambda)$ starts with a pseudothreshold $E_3^-(e^{i\phi_1})$. A sheet on which both the levels vary is shown in Fig. 6. It emerges as a map of an area of Fig. 5. lying between the real half line $\sigma\geqslant 0$ and the line $\mathrm{Im}\,\lambda^-(\sigma)=0$ crossing the point $\sigma=e^{i\phi_1}$. The map is provided by $\lambda^-(\sigma)$.

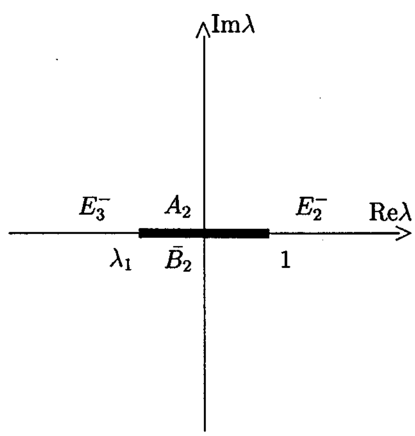

FIG. 6. The sheet of the λ-Riemann surface corresponding to the level $E_2^-(\lambda)$ and its pseudolevelcompartner $E_3^-(\lambda)$.

The second sheet and all the next ones are arranged in a very similar way to the corresponding sheets in the even parity energy level case. To achieve the second sheet on which the levels $E_4^-(\lambda)$ and $E_5^-(\lambda)$ are defined, we should cross a cut $A_2/\overline{B_2}$ of Fig. 6 in any direction close to the branch point $\lambda_1=\lambda^-(e^{i\phi_1})=\phi_1/\sin\phi_1<0$. The point opens the level $E_4^-(\lambda)$ on the sheet for $-\propto<\lambda<\lambda_1$. There is a second branch point on the sheet at $\lambda=\lambda_2=\phi_2/\sin\phi_2>0$, opening the level $E_5^-(\lambda)$ for $\lambda_2<\lambda<+\propto$. The sheet emerges as a map of an area of Fig. 5 lying between two successive lines $\mathrm{Im}\,\lambda^-(\sigma)=0$, the first one crossing the singular point $\sigma_1=e^{i\phi_1}$ and the next one crossing the unit circle at $\sigma_2=e^{i\phi_2}$.

All the next sheets appear as maps of successive areas bounded by the corresponding pairs of the lines $\mathrm{Im}\,\lambda^-(\sigma)=0$ crossing the successive singular points $\sigma_k=e^{i\phi_k}$, $k=2,3,\ldots,$ in the direction of an increasing argument of σ. However, contrary to the even parity case maps of the negative argument sheets of the σ-Riemann surface do not produce additional (twin in forms) sheets on the λ-Riemann surface, which corresponds to the absence of relevant right cut on the first sheet of the latter surface. It means that the complex conjugated positive and negative argument sheets of the σ-Riemann surface map into the same sheets of the λ-Riemann surface for the odd energy function $E^-(\lambda)$.

B. Relation between analytical properties of energy levels and analytical properties of transmission coefficient

It is a standard result of the 1-dim quantum mechanics that energy levels of bound states are simple poles for a transmission coefficient of the corresponding 1-dim scattering problem.[9] These poles have to occupy the positions on the positive imaginary axis of the complex momentum corresponding to an infinite motion. In fact, this last property is the main criterion for selecting the real bound state energies from the whole set of poles the transmission coefficient can have.

Since the analysis of the analytical quantization conditions of the previous section provided us with a variety of the solutions which all have to be poles for a transmission coefficient, then to select out those of them which are the physical bound state levels we have to discuss roles played by them in the corresponding transmission coefficient.

We shall give to our considerations a standard formulation shifting all energy levels by $-\lambda^2$ so that the infinite motion takes place for $E>0$ with the momentum $k=\sqrt{E}$ outside the well and with the momentum $k'=(k^2+\lambda^2)^{1/2}$ inside it (the bottom of the well is now at $V=-\lambda^2$, of course). Then the reflection (R) and transmission (T) coefficients for the case are the following:

$$T(k)=\frac{kk'e^{-2ik}}{(k\cos k'-ik'\sin k')(k'\cos k'-ik\sin k')},$$
$$R(k)=\frac{i\lambda^2\sin 2k'}{2kk'}T(k). \tag{10}$$

It is seen from (10) that T (and R as well) as a function of complex k is meromorphic with their poles coinciding with roots of the T-denominator. Clearly these poles occupy exactly the positions of energy levels in the k-plane we have found in the previous paragraphs. The denominator factorization in (10) occurs due to the reflection symmetry of the potential well. In particular, the first denominator factor in (10) corresponds to the odd parity levels and the second to the even ones.

We can separate the investigations of the levels of different parities (according to what we have done earlier) by considering instead of the coefficients (10) two pairs of the following ones:

$$T^+(k)=\frac{k}{k'\cos k'-ik\sin k'},\quad R^+(k)=\frac{i\lambda\cos k'}{k}T^+(k), \tag{11}$$

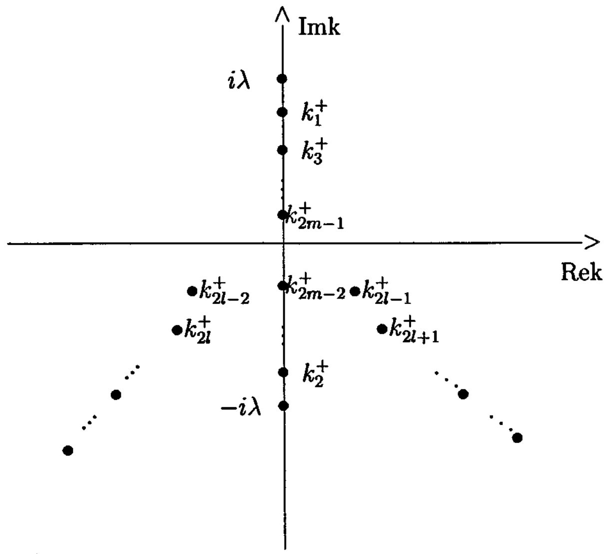

FIG. 7. The k-plane of the transmission coefficient $T^+(k)$.

$$T^-(k)=\frac{k}{k\cos k'-ik'\sin k'},\quad R^-(k)=\frac{i\lambda\sin k'}{k}T^-(k).$$

Of course, the two above pairs of coefficients correspond now to two different potential wells: $V^+(x,\lambda)$ and $V^-(x,\lambda)$, respectively, which can be reconstructed with the help of the inverse scattering method.[10]

Let us consider first a pole structure of the k-plane corresponding to the even parity case. The corresponding analysis of the odd one is analogous.

This structure is shown in Fig. 7 and follows from an observation that $\phi_l^+(\lambda)$ as an lth analytical solution to the first of the conditions (4) defines a pole of $T^+(k)$ on the k-plane at $k_l^+(\lambda)=i\lambda\sin\phi_l^+(\lambda)$.

According to our earlier observation together with solutions we have for a given $\lambda>0$, we have to consider also the ones we have for $-\lambda$.

There is a finite number m ($|\phi_m^+(\pm\lambda)|\leqslant m\pi\leqslant\lambda$) of solutions to (4) for which $\phi_l^+(\pm\lambda)$ are real and $\pm\lambda\sin\phi_m^+(\pm\lambda)>0$. Poles corresponding to these solutions [including the one for the ground state (for $l=1$)] are distributed on the segment $(0,i\lambda)$ of the positive imaginary axis and represent the physical bound states of the potential $V^+(x,\lambda)$.

There are also m-1 real solutions to (4), but with $\pm\lambda\sin\phi_m^+(\pm\lambda)<0$ generating in this way poles lying in the k-plane on the segment $(0,-i\lambda)$ (i.e., *below* the real axis of the plane). Therefore the latter which we call pseudoenergy levels cannot represent the bound states in $V^+(x,\lambda)$. These poles appear as branch partners of the previous ones (except the ground state partner which does not exist) which therefore can coincide with the latter at the corresponding pseudothresholds $\lambda=\lambda_l$ (see Figs. 3–4). However, loci of these pseudothresholds in the k-plane are just below the real axis not contradicting therefore the physical level nondegeneracy theorem. According to our previous notation (see Figs. 3–4) these $2m-1$ poles are $k_1^+(\lambda)$, $(k_2^+(\lambda),k_3^+(\lambda))$, $(k_4^+(\lambda),k_5^+(\lambda))$, ..., $(k_{2m-2}^+(\lambda),k_{2m-1}^+(\lambda))$ where the branch partners are paired and the poles with odd indices represent the physical levels (lie above the real axis).

Finally, there are infinitely many poles of $T^+(k,\lambda)$ lying below the real axis of the k-plane on both sides of the imaginary one and symmetrically with respect to it [due to the relation $k_l(\lambda)=-k_l^*(\lambda^*)$ considered for real λ] but outside of it with finite distances between any two of them and because of this distributed up to infinity. These are, of course, the poles $(k^+_{2l-2}(\lambda),k^+_{2l-1}(\lambda))$, $l=m+1,m+2,\ldots$.

If λ varies through real positive values increasing all the poles lying outside the segment $(-i\lambda,i\lambda)$ move towards it and a symmetric pair of them $(k^+_{2l-2}(\lambda),k^+_{2l-1}(\lambda))$ achieves the segment just for λ approaching λ_l a value being a corresponding pseudothreshold for the pair. While achieving the segment the pair disjoints again with its member $k^+_{2l-1}(\lambda)$ moving upwards of the imaginary axis and with the member $k^+_{2l-2}(\lambda)$ moving downwards. The first one crosses eventually the real axis becoming a bound state whilst the second one does it never, becoming a pseudoenergy level.

The above description works for every pair of the pole partners except the one corresponding to the ground state energy which is single. This for real positive λ [as well as for the negative one due to a relation $k_l(\lambda)=k_l(-\lambda)$ valid for any λ] is above the real k-axis and thus represents the ground state energy. However, for imaginary $\lambda(\sigma)$, when $0<\sigma<1$ or $1<\sigma<+\propto$, $E^+_1(\lambda)=\lambda^2$ is real (and, of course, negative) but $k^+_1(\lambda)$ has a negative imaginary part and therefore $E^+_1(\lambda)$ cannot represent a bound state. This fits well our intuition, since for imaginary λ the rectangular well become rather a rectangular barrier excluding, of course, any bound state.

A similar analysis of the odd parity levels corresponding to the potential $V^-(x,\lambda)$ leads us to the following conclusions.

For a given $\lambda>0$ and $m\pi\leqslant\lambda$ poles: $k^-_2(\lambda)$, $(k^-_3(-\lambda),k^-_4(-\lambda))$, $(k^-_5(\lambda),k^-_6(\lambda))$, $(k^-_7(-\lambda),k^-_8(-\lambda)),\ldots,(k^-_{2m-1}((-1)^{m-1}\lambda),k^-_{2m}((-1)^{m-1}\lambda))$ lie in the segment $(-i\lambda,i\lambda)$ whilst the remaining ones $(k^-_{2l-1}((-1)^{l-1}\lambda),k^-_{2l}((-1)^{l-1}\lambda))$, $l=m+1,m+2,\ldots$, lie outside the segment, below the real axis and symmetrically with respect to the imaginary axis.

A behavior of all the paired poles with λ varying along the positive real axis is exactly the same as in the even parity level case. Only the pole $k^-_2(\lambda)$ seems to differ with this respect in comparison with $k^-_1(\lambda)$ escaping to infinity along the negative imaginary axis when $\lambda\to 0_+$. However, if one considers a behavior of the levels rather as functions of σ for $\sigma\to 0_+$ (or for $\sigma\to+\propto$) on the corresponding first sheets of Figs. 5 and 2, respectively (this corresponds to $\lambda\to 0_+$ for the odd level but describes a little bit more complex path for the even one ending, however, at $\lambda=0$) then one can find that both the levels behave identically.

C. Perturbation series for energy levels and their summability

Since the finite rectangular well considered is a perturbation for the infinite rectangular one when $\lambda^{-2}\to 0$ then the corresponding energy levels of the former should approach in this limit the corresponding levels of the latter potential. It is interesting that with the previous considerations we are able to conclude that all the corresponding perturbation series expansions are convergent. The corresponding conclusions are, in fact, obvious for all the energy levels higher than the ground state one. To show this let us consider a pair of levels E^+_{4n+1}, E^+_{4n+2} lying on the $2n+2^{th}$ sheet shown in Fig. 4, with the level E^+_{4n+1} belonging to the even spectrum of the finite rectangular well. A function $E^+(\lambda)=\lambda^2z^2(\lambda)$ being a holomorphic extension of both the levels considered is holomorphic on the $2n+2$th sheet outside any closed contour C containing the branch points $\lambda=0$, λ_{2n+1} and λ_{2n+2} inside (see Fig. 4). For $\lambda\to\infty$ on the sheet $E^+(\lambda)\sim E^+_{4n+1}(\lambda)\sim\pi^2(8n+1)^2/4$ and therefore according to the Cauchy theorem we get for $E^+(\lambda)$,

$$E^+(\lambda)=\frac{((8n+1)\pi)^2}{4}-\frac{1}{2\pi i}\oint_C\frac{E^+(\lambda')}{\lambda'-\lambda}d\lambda'. \qquad (12)$$

It follows from (12) that the series

$$\frac{((8n+1)\pi)^2}{4}+\sum_{k\geqslant 0}\frac{a_{2n+1,k}}{\lambda^{k+1}},$$

with

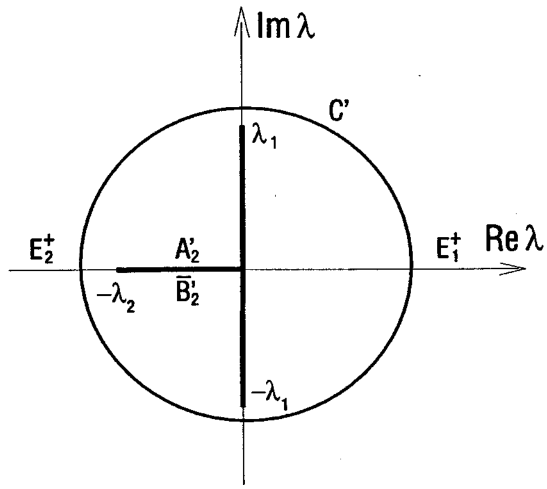

FIG. 8. The cut structure of the sheet on which the perturbative series for $E_1^+(\lambda)$ is studied.

$$a_{2n+1,k}=\frac{(-1)^k}{2\pi i}\oint_C E^+(\lambda)\lambda^k\,d\lambda, \tag{13}$$

which is the perturbative one for the $E_{4n+1}^+(\lambda)$ level is convergent to the level for $|\lambda|>|\lambda_{2n+2}|$, i.e., for all λ's just above its pseudothreshold.

For the ground state energy level $E_1^+(\lambda)$ an analogous conclusion can be obtained by considering instead the cut pattern shown in Fig. 3(a) the one when the upper and the lower left cut boundaries in the figure are rotated by $\pm\pi$, respectively, providing us with a sheet arranged according to Fig. 8. It follows from the figure that the perturbative series for $E_1^+(\lambda)$ is obtained from (12) by putting there $n=1$ and $C=C'$ with its convergence radius given by $\max(|\lambda_1|,|\lambda_2|)$.

The same conclusions can be drawn for the perturbative series constructed for the odd parity energy levels, applying exactly the same technique of considerations as we did in the case of the even parity levels. The series constructed for a level $E_n^-(\lambda)$ is therefore convergent for $|\lambda|$ sufficiently large on the sheet on which the level is defined.

III. DIRAC DELTA BARRIER AS PERTURBATION

A second example of a perturbation which can be analyzed analytically is provided by the Dirac delta barrier introduced into the infinite rectangular well, i.e., it is given by

$$V(x,g)=2g\,\delta(x), \quad |x|<1 \quad \psi(\pm 1)=0, \tag{14}$$

where $\psi(x)$ is a wave function for the case. A role of a perturbation parameter is played by g. For $g=0$ we get a problem of the energy spectrum in the infinite rectangular well, i.e., an asymptotic limit of any energy level $E_n(g)$ of the potential (14) is just a corresponding level E_n of the rectangular well. In the limit $g\to+\infty$ we get instead two infinite rectangular wells with the half of the size of the well we started with for which their energy spectra should coincide with the corresponding limit of $E_n(g)$. This is not unexpected. What is interesting in this example is the fact that every even parity level of the potential (14) is a branch of some ramified function $E(g)$ considered as a function of complex g (odd parity energy levels decouple of g and therefore coincide with the odd ones of the infinite rectangular well).[8]

The last property makes to some extent the considered case similar to the previous one. However, there is one considerable difference between them. This is that the existence of all the even energy levels $E_n(g)$ does not depend on g, i.e., each level exists for any g, including $g=0$. Therefore, we cannot expect on a g-Riemann surface $\mathbf{R}_g$ of $E(g)$ the real branch points to

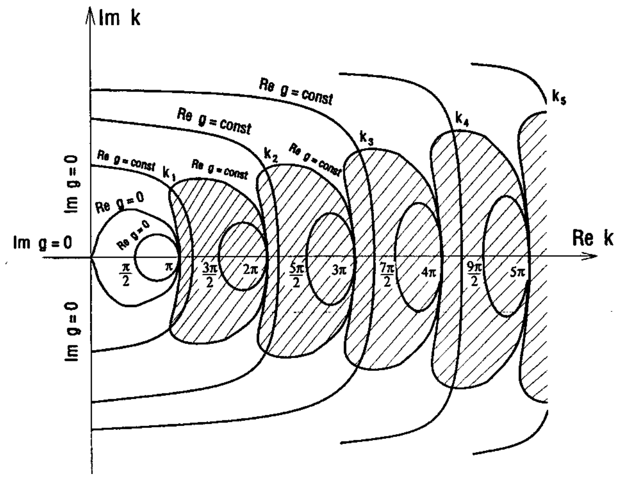

FIG. 9. The pattern of lines Reg(k) = const and Im $g(g)$ = const for $g(k) = -k \cot k$.

appear having meanings of pseudothresholds for the levels $E_n(g)$. On the other hand, there are branch points on $\mathbf{R}_g$ [as they have to be because $E(g)$ is a ramified function] which are complex but which physical meaning seems to be a puzzle.

Nevertheless, there is also a property of the considered case which is a copy of the corresponding one for the finite rectangular well. Namely, the perturbative series for even $E_n(g)$ are all convergent. All these follow from the quantization condition for even energy levels:

$$k \cot k = -g, \quad k = \sqrt{E}. \tag{15}$$

The condition (15) defines g as a meromorphic function of k with simple poles on the k-plane at $k = r\pi$, $r = \pm 1, \pm 2, \ldots$. Therefore, an inversion of the relation (15) will be easy if we know positions of zeros of $g'(k)$ in the k-plane. These positions are determined by (15) and by the following condition:

$$g^2 + g + k^2 = 0. \tag{16}$$

Both the last equations are equivalent to

$$\sin 2k = 2k, \quad g = -k \cot k. \tag{17}$$

The first of Eqs. (17) has a solution for $k = 0$ and an infinite number of complex solutions in the k-plane (i.e., $\mathrm{Im}\, k \neq 0$ for all these solutions) with the property that if k_l, $l = 1,2, \ldots$, is a solution to it then $-k_l$, k_l^* and $-k_l^*$ are also. It is easy to show that for large l $\mathrm{Im}\, k_l$ increases as $\ln l$ and $\mathrm{Re}\, k_l$ as $l\pi$.[8] A corresponding pattern of lines $\mathrm{Re}\, g = $ const, taking into account the distribution of singular points described above is shown in Fig. 9. The pattern allows for an easy construction of a g-Riemann surface $\mathbf{R}_g$ for an inverse function $k(g)$. Namely, we can arrange cuts on $\mathbf{R}_g$ in such a way as to map an lth dashed area of Fig. 9 into a sheet of $\mathbf{R}_g$ corresponding to E_lth even energy level of the potential considered. This is shown in Fig. 11.

Because $g(k)$ is a symmetric function of k, i.e., $g(k) = g(-k)$ with the singular point at $k = 0$ [i.e., $g'(0) = 0$] its inverse function $k(g)$ is defined on the Riemann surface $\mathbf{R}_g$ which consists of two twin systems of sheets joined by a single cut beginning at $g = -1$ and running to $-\infty$. These two systems of sheets are maps of the right and the left k-half planes of Fig. 9, respectively. The two sheets opening the systems and joined by the cut described above are the ones on which the ground state energy level is defined, taking the same values in the corresponding points of the

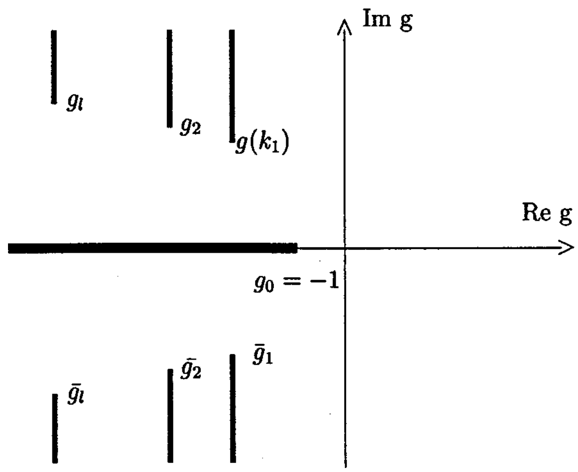

FIG. 10. The cut structure of the g-Riemann surface for k(g) determining the surface.

sheets. This is because $E(g)=k^2(g)$. Therefore, for $E(g)$ the point $g=-1$ on the first sheet considered is not a branch one and we can consider only one of the two systems of sheets corresponding, for example to $\mathrm{Re}\, k \geqslant 0$ what is assumed from now on.

In accordance with the results of Ushveridze,[8] all singular points of $E(g)$ lie on the first sheet corresponding to the ground state energy level of the potential (11). The sheet is an image of an area which completes the dashed ones of Fig. 9 to the full right k-half plane and is sketched in Fig. 10.

As it has been already explained earlier on the lth sheet on which the lth energy level $E_l(g)$ is defined, there is a unique pair of complex conjugated branch points at $g_l(=g(k_l))$ and $\overline{g_l}$ by which the level contacts with the ground state one (see Fig. 11). The level is holomorphic on its sheet at $g=0$ and approaches $[(l+1)\pi]^2$ or $(l\pi)^2$ for $g\rightarrow\infty$ in the half planes $\mathrm{Re}\, g>\mathrm{Re}\, g_l$ or $\mathrm{Re}\, g<\mathrm{Re}\, g_l$, respectively.

Because of the holomorphicity of $E_l(g)$ at $g=0$ its perturbative series with respect to g converges inside the circle $|g|=|g_l|$.

For large g, however, we have to consider together with $E_l(g)$ also its neighbor $E_{l+1}(g)$ for which its limit for g in the half plane $\mathrm{Re}\, g<\mathrm{Re}\, g_l$ is the same as for $E_l(g)$ in the half plane $\mathrm{Re}\, g>\mathrm{Re}\, g_l$. A common sheet for both the levels is shown in Fig. 12. It is obvious that we can apply here the Cauchy integral technique to show the convergence of the asymptotic series for $E_l(g)$ and for $E_{l+1}(g)$ for large g with $|g|>|g_{l+1}|$, $l=1,2,\ldots$.

The corresponding statement for the ground state energy level perturbation series is also true for $|g|>|g_1|$.

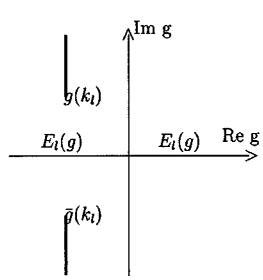

FIG. 11. The g-Riemann surface for lth energy level of the infinite rectangular well with the Dirac delta barrier.

IV. DISCUSSION AND CONCLUSIONS

The model of the finite rectangular well which we have considered provided us with several properties, differing it from the previous ones.

First it was a number of energy levels varying with a perturbation parameter thus changing a quality of the levels of being real to becoming only potential ones.

Secondly, it was the existence of the pseudoenergy levels in the model the main role of which was just to allow the level crossing phenomenon to appear.

It was also an easiness of predictions of the pseudothreshold distribution as a part of all the level crossing loci. The latter property followed, however, from the simplicity of the quantization conditions for the case as illustrated by Fig. 1. In fact, the remaining three branch points, the two for the ground state level at $\lambda = \pm i$ and the logarithmic one for all the levels at $\lambda = 0$, could be identified only by detailed calculations.

Finally, it was the convergence of the perturbation series for large λ and for all the levels. This property could not be deduced prior to the detailed knowledge of the λ-Riemann surface topology just because for imaginary λ (no matter small or large) the rectangular well became the rectangular barrier, suggesting rather a possibility for the perturbation series to be divergent in these directions because of a repelling character of the barrier. However, the repulsion of the rectangular barrier appears to be not enough strong to destroy the convergence of the perturbation series.

The property of the perturbation series to be convergent was shared also by the corresponding series constructed for the levels in the infinite rectangular well perturbed by the Dirac delta barrier.

In general, there is a question about the possibilities to predict some basic properties of analytical dependence of energy levels on a parameter considered as the perturbation one prior to any detailed calculation, i.e., relying only on some general properties of the potential given. Obviously, such qualitative predictions are possible in some simpler cases. There are, however, also many traps and puzzles which one can meet trying to proceed in this way.

Consider, for example, the case of the potential (1). Its properties remind many of those of the finite rectangular well if we consider $\lambda = \alpha^{-2}$ as a perturbation parameter (an obvious choice after a rescaling $x \rightarrow \alpha^{-1} x$ in the corresponding Schrödinger equation).

First, a corresponding quantization condition is given by a nonalgebraic dependence between energy E and λ which can be readily continued to complex λ.[11] We can expect therefore $E(\lambda)$ to be defined on infinitly sheeted Riemann surfaces for both the parities.

Secondly, a number of its energy level depends on its actual height (equal to $\lambda \omega^2/2$) and therefore the levels of this well should have real and positive thresholds as their branch points and therefore should be accompanied by the corresponding pseudolevels.

Third, its levels approach the ones of the harmonic oscillator when $\lambda \rightarrow +\infty$.

Next, changing $\lambda \rightarrow -\lambda$ in (1) we find that the spectrum of the finite well (1) is transformed into the corresponding spectrum of the infinite well defined by $\cos(\alpha x)$, the latter arising from $\cosh(\alpha x)$ after the transformation. The spectrum of this infinite well approaches again the harmonic one when $\lambda \rightarrow +\infty$ (after the above change of sign of λ). On the other hand, when $\lambda \rightarrow 0$ each level of the considered infinite well goes to infinity (since the well becomes infinitely narrow), i.e., $E(\lambda)$ has to have a singularity at $\lambda = 0$ for each level. However, such a singularity

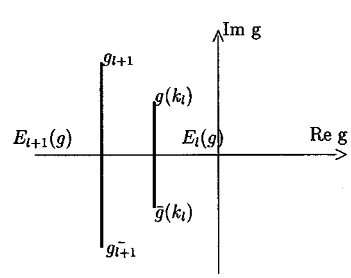

FIG. 12. The cut structure of the sheet on which the perturbative series for $E_l(g)$ is studied.

should not be detected on the sheet on which the ground state level of the finite well is represented since this level vanishes (amounts to zero) at $\lambda=0$. This suggests that both the families of levels are distributed on different sheets.

Finally, using the standard analysis[2] to establish the maximal sectors of analyticity of $E(\lambda)$, one can find that for each level sheet in the finite well case $|\arg\lambda|$ in such a sector does not exceed $\pi/2$ and for the level sheets in the infinite well case the latter value is not exceeded by $|\arg\lambda-\pi|$, which means that in both the cases there are singularities on both the positive and negative imaginary half axes for each level sheet.

It seems that the above qualitative analysis is all this that can be concluded not involving oneself into a further detailed analysis of the exact quantization condition corresponding to the case. The latter condition, however, is complicated enough[11] to allow only a numerical approach. In particular, having the above conclusions it is not possible to say whether the corresponding perturbation series both for the levels in the finite well and in the infinite one are convergent or asymptotic only. The last properties are determined uniquely in the case considered by a distribution of singularities on the imaginary axis of each level sheet and whether there is a finite number of them on the axis (the perturbation series on such a sheet converges then) or they are distributed on the axis up to infinity (the series is asymptotic) cannot be inferred without a detailed analysis of the quantization condition.

In the considered case the singularities lying on the imaginary axes of the level sheets are certainly the branch points for $E(\lambda)$ and describe the way the different levels can communicate with themselves by analytical continuations. If there is a finite number of such branch points on the axis then the level attached to the sheet can contact directly with only a finite number of other levels. (The remaining levels can be achieved in such a case via the former reached directly.) This is the case of the finite rectangular well and the Dirac delta barrier in the infinite well considered in this paper.

In the case of an infinite number of the branch points, the directly contacted levels are also infinite in number. This is the case of the anharmonic oscillator of Bender and Wu.

Unfortunately, we could not find some visible and general criterion allowing us to judge between any of these two possibilities prior to any detailed calculations of the distribution of such singularities.

The following ''small'' modification of the previous example show us, however, that the qualitative analysis as demonstrated above can be misleading if not supported by the detailed calculations.

Namely, let us substitute $\cosh(\alpha x)$ in the potential (1) by its square $\cosh^2(\alpha x)$. No doubt one can repeat all the previous reasonings also in this case with the identical conclusions. However, the case is known to be integrable,[11–14] i.e., its corresponding quantization condition, is obtained as an algebraical equation for $E(\lambda)$ which can be easily solved to give

$$E_n(\lambda)=\left(n+\frac{1}{2}\right)\sqrt{(\omega\hbar)^2+\frac{\hbar^4}{4\lambda^2}}-\left(n+\frac{1}{2}\right)^2\frac{\hbar^2}{2\lambda}-\frac{\hbar^2}{8\lambda},$$
$$\lambda\geqslant\lambda_n=\frac{\hbar}{\omega}\sqrt{n(n+1)},n=0,1,\ldots, \tag{18}$$

where λ_n, $n=0,1,\ldots,$ are thresholds below which the corresponding energies $E_n(\lambda)$ become unphysical.

It is seen from (18) that $E(\lambda)$ for the case considered enjoys all the properties postulated earlier for the energy levels of the potential (1) except the following two of them: 1^0 that it is defined on the infinitely sheeted λ-Riemann surface(s), and 2^0 that there are branch points (thresholds) on the positive real half axes and the corresponding pseudoenergy levels. The lack of the latter two properties follows, of course, as a result of decoupling from each other of all the level sheets so that each level is defined on a separate double sheeted Riemann surface.

It is still worth noting that as it follows from (18) for each n, $n=0,1,\ldots,$ the levels $E_n^{fin}(\lambda)$ and $E_n^{inf}(\lambda)$ corresponding to the finite and infinite wells, respectively, are defined on the same double sheeted Riemann surface R_n. The first of the levels is defined on the positive real half axis of the first sheet of R_n, whilst the second—on the negative real one of the second sheet of R_n. The level $E_n(\lambda)$ branches at the points $\lambda=\pm\imath\hbar/(2\omega)$. The ground state level of the finite well is finite (vanishes) at $\lambda=0$ whilst all the remaining ones (for both the wells) have a simple pole at this point on their sheets. Of course, the convergence of the corresponding perturbation series for the levels considered when λ is close to infinity is a trivial conclusion from (18).

We can conclude therefore that generally the analytical properties of energy levels considered above show that they depend strongly on the model used. Crossing of levels considered as the main cause for the presence of branch points in the dependence of levels on a chosen perturbation parameter seems to be as obvious as puzzled in most of the models. In particular, these are the distributions of branch points in investigated models which does not seem to be covered by some universal rules. Needless to say it is just this distribution which decides whether the corresponding pertubation series are convergent or only asymptotic. Both the latter statements are true also in both the cases of potentials investigated in this paper. In particular, the existence of pseudothreshold branch points cannot be considered as a universal rule as it is shown by the formula (18). On the other hand, the case of the Dirac δ-function barrier in the infinite rectangular well demonstrates the unpredictibility of the branch point distributions prior to the detailed calculations.

It is worth mentioning in this context, however, that there is a group of ''perturbed'' potentials (depending on a ''large'' perturbation parameter) for which one can predict with the certainty that for some families of their energy levels the corresponding perturbation series have to be convergent. Namely, these are so called quasi exactly solvable potentials mentioned in the Introduction.[7] For them, both the corresponding families of energy levels and the perturbation parameters satisfy algebraic equations so that these families are defined on finite sheeted Riemann surfaces of the perturbation parameter and therefore there can be only a finite number of singular points of the levels on the surface. Because of that the corresponding perturbation series have to be convergent.

In general, however, it seems that both a large variety of different parameters which were used as perturbation ones in the models considered so far and the models themselves do not allow us to formulate some common rules covering the distributions of branch points and giving them in this way some common sense.

Nevertheless, guided by a detailed knowledge of a considered potential and a corresponding quantization condition as well as by a meaning and a role some particular perturbation parameter of the potential can play in it, we can in many cases predict qualitatively much of the actual dependence of the levels on the perturbative parameter prior to detailed calculations.

ACKNOWLEDGMENTS

S.G. was supported by KBN 2PO3B 07610. P. M. was supported by the Łódź University Grant No. 580.

[1] C. M. Bender and T. T. Wu, Phys. Rev. **184**, 1231 (1969); Phys. Rev. D **7**, 1620 (1973).
[2] S. Giller, Acta Phys. Pol. B **23**, 457 (1992).
[3] B. Simon and A. Dicke, Ann. Phys. (N.Y.) **58**, 76 (1970).
[4] T. I. Banks and C. M. Bender, J. Math. Phys. **13**, 1320 (1972).
[5] C. M. Bender, H. J. Happ, and B. Svetitsky, Phys. Rev. D **8**, 2324 (1974).
[6] P. E. Shanley, Phys. Lett. A **117**, 161 (1986).
[7] A. G. Ushveridze, *Quasi-Exactly Solvable Problems in Quantum Mechanics* (IOP, Bristol, 1993).
[8] A. G. Ushveridze, J. Phys. A **21**, 955 (1988).
[9] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics. Nonrelativistic Theory* (Pergamon, Oxford, 1965).
[10] V. E. Zacharov, S. V. Manakov, S. P. Novikov, and L. P. Pitayevski, *Soliton Theory* (Nauka, Moscow, 1980) (in Russian).
[11] P. Milczarski, ''Exactness of conventional and supersymmetric JWKB formulae and global symmetries of stokes graphs,'' quant-ph/9807039 (1998).
[12] C. Rozenzweig and J. B. Krieger, J. Math. Phys. **9**, 849 (1968).
[13] N. Rosen and P. M. Morse, Phys. Rev. **42**, 210 (1932).
[14] G. Pöschl and E. Teller, Z. Phys. **83**, 143 (1933).

A braided interpretation of fractional supersymmetry in higher dimensions

R. S. Dunne[a)]
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 9EW, United Kingdom

(Received 19 March 1997; accepted for publication 10 March 1998)

A many variable q-calculus is introduced using the formalism of braided covector algebras. Its properties are discussed in detail and related to fractional supersymmetry when certain of its deformation parameters are roots of unity. The special cases of two-dimensional supersymmetry and fractional supersymmetry are developed in detail. [S0022-2488(98)01806-4]

I. INTRODUCTION

In four recent papers[1–4] the properties of the braided line,[5,6] when its deformation parameter is a root of unity, were discussed. Most notably, these studies led to a novel understanding of one-dimensional supersymmetry and fractional supersymmetry.[7–9] Our aim in the present paper is to extend these results to the many variable case. In order to do this we construct a many variable, generic q, generalized Grassmann algebra using the formalism of braided covector algebras[5,6] with suitable R and R' matrices. Within the framework provided by this formalism the construction of the corresponding many variable left q-calculus is straightforward. After a little further work the corresponding right q-calculus is also obtained.

In this many variable case it is convenient to generalize the graded brackets used in Ref. 2 to a pair of braided brackets (left and right), which we introduce in Sec. II. This change has several, in general useful, consequences. In particular, left and right differentiation and integration become truly distinct, rather than being the same thing induced by different algebraic operators as was the case in Ref. 2. There are well-defined and simple commutation relations between all of these operations. Another advantage over the approach of Ref. 2 is that the conditions which govern the commutation relations of noncommuting constants are built into the many variable algebra, so that they are no longer additional constraints. In contrast to the situation with graded brackets, the conditions necessary in order that left and right differentiation be induced are compatible. A consequence of this is that we can now work quite generally with both left and right differentiation/integration, rather than being restricted to one or the other, as was the case for graded-bracket-based q-calculus.

In Sec. III we take the $q_a \to \tilde{q}_a$ limit ($\tilde{q}_a$ is a root of unity) of the many variable q-calculus, and obtain many variable analogs of the structures and decompositions seen in Refs. 2–4. A full set of commutation relations between the different algebraic elements, derivatives, and integrals is also given. At the end of Sec. III the braided Hopf structure of both the coordinates and the derivatives is given, as well as the duality between them. Further details of this duality, as well as an alternative discussion of the braided line Hopf algebra (at generic q and at q a root of unity) are give in the Appendix.

Section IV deals with the case of two-dimensional supersymmetry. This plays an important role in superstring theory,[10] in which it corresponds to supersymmetry on the world sheet of the string. The full two-dimensional supersymmetry algebra and transformations are recovered, and all of the transformation properties of the bosonic spacetime variables x and t emerge as consequences of their definition as different combinations of the $q_a \to -1$ limits of two braided line coordinates $\{\theta_a\}$ $(a=1,2)$. Together these two braided lines make up a braided plane, but we note that this is not the braided/quantum plane which is usually encountered in the literature.[5,6] In the limit, translations within this braided plane induce supersymmetry transformations of x and t.

[a)]Electronic mail: r.s.dunne@damtp.cam.ac.uk

Furthermore, once the Lorentz transformations of θ_1 and θ_2 are specified, those of x and t follow automatically. The results are in agreement with the standard version of the two-dimensional super-Poincaré transformation.

Section V extends the results of Sec. IV to the case of *mixed* fractional supersymmetry in two dimensions. The word mixed is used to indicate that the deformation parameters of the two braided plane coordinates on which the fractional supersymmetry is based are not necessarily at the same root of unity. All of the algebraic and transformation properties are worked out, and as in the supersymmetric case, spacetime Lorentz transformations are induced by suitable transformations of the braided plane coordinates. We are thus able to introduce full two-dimensional mixed fractional super-Poincaré transformations. Finally we extend the arguments of Ref. 2 concerning the Berezin integral to the two-dimensional case.

II. *q*-CALCULUS FOR AN ARBITRARY NUMBER OF VARIABLES

In this section we develop the q-calculus associated with r generalized Grassmannian variables. This calculus can be viewed as a particular example of the braided differential calculus described in Refs. 5 and 6 and we present it from this point of view.

Given any matrix $R_{12}\in M_r\otimes M_r$ which satisfies the quantum Yang–Baxter equation,

$$R_{12}R_{13}R_{23}=R_{23}R_{13}R_{12}, \tag{1}$$

as well as an associated matrix $R'_{12}\in M_r\otimes M_r$ satisfying

$$R_{12}R_{13}R'_{23}=R'_{23}R_{13}R_{12}, \quad R_{23}R_{13}R'_{12}=R'_{12}R_{13}R_{23}, \tag{2}$$

$$R_{21}R'_{12}=R'_{12}R_{21}, \tag{3}$$

$$(PR+1)(PR'-1)=0, \tag{4}$$

where P is the permutation matrix, we can define a braided covector algebra[5,6] with elements $\{x_i,1\}$. This has a braided Hopf algebraic structure given by

$$\begin{gathered}\mathbf{x}_1\mathbf{x}_2=\mathbf{x}_2\mathbf{x}_1R'_{12}, \quad \text{i.e.,} \quad x_ix_j=x_cx_bR'^{bc}_{ij}, \\ \Delta x_i=x_i\otimes 1+1\otimes x_i, \quad \epsilon(x_i)=0, \\ S(x_i)=-x_i, \quad \Psi_{12}(\mathbf{x}_1\otimes\mathbf{x}_2)=\mathbf{x}_2\otimes\mathbf{x}_1R_{12},\end{gathered} \tag{5}$$

as well as $\Delta(1)=1\otimes 1$, $\epsilon(1)=1$, $S(1)=1$, $\Psi_{12}(1\otimes x_i)=x_i\otimes 1$, $\Psi_{12}(x_i\otimes 1)=1\otimes x_i$. It is convenient to use the notation $\mathbf{w}=\mathbf{x}\otimes 1$, $\mathbf{x}=1\otimes\mathbf{x}$ so that the $\{w_i\}$ satisfy the same algebra as the $\{x_i\}$. In this notation[6] the coproduct Δ above is just

$$\Delta\mathbf{x}=\mathbf{w}+\mathbf{x}, \tag{6}$$

and the braiding Ψ_{12} is equivalent to the following braid statistics between $\mathbf{x}$ and $\mathbf{w}$:

$$\mathbf{x}_1\mathbf{w}_2=\mathbf{w}_2\mathbf{x}_1R_{12}, \quad \text{i.e.,} \quad x_iw_j=w_cx_bR^{bc}_{ij}. \tag{7}$$

The notation (6) suggests that we regard the coproduct as generating left translations within the braided covector space, and motivates its alternative name *coaddition*.[5,6] No additional information is needed to construct the corresponding braided left differential calculus. The left derivatives form a braided vector algebra, with commutation relations given by

$$\partial_{l1}\partial_{l2}=R'_{12}\partial_{l2}\partial_{l1}, \tag{8}$$

and the cross relations giving their action on the covectors are

$$\partial_{l1}\mathbf{x}_2-\mathbf{x}_2R_{21}\partial_{l1}=\delta_{12}. \tag{9}$$

The reason for the form of the second of these relations is that along with (7) it implies that $[\mathbf{w}_1\partial_{l1},\mathbf{x}_2]=\delta_{12}\mathbf{w}_1$, so that $\mathbf{w}_1\partial_{l1}$ can be viewed as the generator of the left translation (6). Equation (9) can be viewed as giving the braiding between the covectors and their derivatives.[11] To see this we identify

$$\Psi_{12}^{-1}(\partial_{l1}\otimes\mathbf{x}_2)=\mathbf{x}_2\otimes R_{21}\partial_{l1}, \tag{10}$$

so that we can rewrite (9) as

$$\partial_{l1}\mathbf{x}_2-\cdot\Psi_{12}^{-1}(\partial_{l1}\otimes\mathbf{x}_2)=\delta_{12}. \tag{11}$$

It is not difficult to extend this formalism so that it includes right derivatives, and since these play an important role in supersymmetric and fractional supersymmetric theories, we will do so explicitly. Among themselves the right derivatives have the same commutation relations as the left derivatives, and thus also form a braided vector algebra so that

$$\partial_{r1}\partial_{r2}=R'_{12}\partial_{r2}\partial_{r1}. \tag{12}$$

To find the cross relations between these derivatives and the covectors we must first reinterpret Δ as the generator of a right shifts. We can do this by writing

$$\Delta\mathbf{x}=\mathbf{x}+\mathbf{y}, \tag{13}$$

where we have introduced the alternative notation $\mathbf{x}=\mathbf{x}\otimes 1$ and $\mathbf{y}=1\otimes\mathbf{x}$. From the braiding Ψ_{12} given by (5) we obtain the braid statistics

$$\mathbf{y}_1\mathbf{x}_2=\mathbf{x}_2\mathbf{y}_1R_{12}. \tag{14}$$

In order for the right derivatives to generate the translation (13) they must satisfy $[\mathbf{y}_1\partial_{r1},\mathbf{x}_2]=\delta_{12}$. In combination with the braid statistics (14) this implies that

$$\mathbf{y}_1\partial_{r1}\mathbf{x}_2-\mathbf{y}_1\mathbf{x}_2R_{12}^{-1}\partial_{r1}=\delta_{12}\mathbf{y}_1, \tag{15}$$

from which it is clear that suitable cross relations are

$$\partial_{r1}\mathbf{x}_2-\mathbf{x}_2R_{12}^{-1}\partial_{r1}=\delta_{12}. \tag{16}$$

As in the case of left derivatives we can interpret this as giving the braiding between the right derivatives and the covectors. Thus by identifying

$$\Psi_{21}(\partial_{r1}\otimes\mathbf{x}_2)=\mathbf{x}_2\otimes R_{12}^{-1}\partial_{r1}, \tag{17}$$

we can rewrite (16) as

$$\partial_{r1}\mathbf{x}_2-\cdot\Psi_{21}(\partial_{r1}\otimes\mathbf{x}_2)=\delta_{12}. \tag{18}$$

Relationships (11) and (18) motivate the introduction of bilinear left and right braided brackets,

$$\begin{aligned}[][A,B]_L&:=AB-\cdot\Psi_{12}^{-1}(A\otimes B),\\ [A,B]_R&:=AB-\cdot\Psi_{21}(A\otimes B).\end{aligned} \tag{19}$$

The bilinearity follows from the bilinearity of Ψ. This bracket is well defined on products as long as we remember the expansion rule for the braiding. From Ref. 12 this is

$$\begin{aligned}\Psi(AB\otimes C)&=(1\otimes\cdot)(\Psi\otimes 1)(A\otimes\Psi(B\otimes C)),\\ \Psi(A\otimes BC)&=(\cdot\otimes 1)(1\otimes\Psi)(\Psi(A\otimes B)\otimes C).\end{aligned} \tag{20}$$

Note that (as one would expect) Ψ_{12}^{-1} and Ψ_{21} expand in the same way. Using these brackets we can define left and right differentiation as follows:

$$\left(\frac{d}{d\mathbf{x}_1}\right)_L f := [\partial_{l1}, f]_L,$$
$$\left(\frac{d}{d\mathbf{x}_1}\right)_R f := [\partial_{r1}, f]_R. \tag{21}$$

Here $f=f\{x_i\}$. We will provide a specific example shortly. We are now able to introduce the generalized Grassmann algebra, which we define as the braided covector algebra in which R and R' are the following r dimensional matrices:

$$R'_{12}=W_{12}, \quad R_{12}=W_{12}+(Q_1-1)\delta_{12}=T_{12}, \tag{22}$$

the coordinate form of which is:

$$R'^{ij}_{ab}=\omega_{ab}\delta^i_a\delta^j_b, \quad R^{ij}_{ab}=(\omega_{ab}+(q_a-1)\delta_{ab})\delta^i_a\delta^j_b=t_{ab}\delta^i_a\delta^j_b. \tag{23}$$

Here $\omega_{ba}=\omega_{ab}^{-1}$ so that $\omega_{aa}=1$, $\omega_{ab}\neq 0$, and $q_a\neq 0$. It follows directly from the fact that R and R' are diagonal that (1)–(3) are satisfied. To show that (4) is also satisfied we expand it explicitly,

$$\begin{aligned}(PR+1)(PR'-1)&=R_{21}R'_{12}+R'_{21}+R_{21}-1\\&=(W_{21}+(Q_2-1)\delta_{21})W_{12}+W_{21}-W_{21}-(Q_2-1)\delta_{21}-1\\&=1+(Q_2-1)\delta_{21}-(Q_2-1)\delta_{21}-1=0.\end{aligned} \tag{24}$$

Putting this R' into (5) we obtain the defining algebra of r generalized Grassmann variables,

$$\theta_a\theta_b=\theta_j\theta_i\omega_{ab}\delta^i_a\delta^j_b, \tag{25}$$

which is equivalent to

$$[\theta_a,\theta_b]\omega_{ab}=0. \tag{26}$$

For left shifts we use the notation $\theta_a=1\otimes\theta_a$ and $\boldsymbol{\epsilon}_a=\theta_a\otimes 1$. From (7) we obtain

$$\theta_a\boldsymbol{\epsilon}_b-\boldsymbol{\epsilon}_j\theta_i\delta^i_a\delta^j_b t_{ij}=0, \tag{27}$$

which is equivalent to

$$[\boldsymbol{\epsilon}_a,\theta_b]_{t_{ba}^{-1}}=0. \tag{28}$$

For the corresponding derivatives $\mathscr{D}_{La}$ we obtain from (8) and (9) the following commutation and cross relations:

$$[\mathscr{D}_{La},\mathscr{D}_{Lb}]_{\omega_{ab}}=0,$$
$$[\mathscr{D}_{La},\theta_b]_{t_{ba}}=\delta_{ab}. \tag{29}$$

For right shifts we use the notation $\theta_a=\theta_a\otimes 1$ and $\eta_a=1\otimes\theta_a$. Then from (14) we obtain

$$[\eta_a,\theta_b]_{t_{ab}}=0, \tag{30}$$

while (12) and (16) give us

$$[\mathscr{D}_{Ra},\mathscr{D}_{Rb}]_{\omega_{ab}}=0,$$
$$[\mathscr{D}_{Ra},\theta_b]_{t_{ab}^{-1}}=\delta_{ab}. \tag{31}$$

These derivatives are generated by the left and right braided brackets, thus from (19),

$$\left(\frac{d}{d\theta_a}\right)_L\theta_b=[\mathscr{D}_{La},\theta_b]_L=[\mathscr{D}_{La},\theta_b]_{t_{ba}}=\delta_{ab},$$
$$\left(\frac{d}{d\theta_a}\right)_R\theta_b=[\mathscr{D}_{Ra},\theta_b]_R=[\mathscr{D}_{Ra},\theta_b]_{t_{ab}^{-1}}=\delta_{ab}. \tag{32}$$

As an example of differentiation induced by the braided brackets (19) we consider the case of $r=2$, and functions $f(\theta_1,\theta_2)$ which can be expanded as positive power series of the form

$$f(\theta_1,\theta_2)=\sum_{l,m=0}^{\infty}C_{l,m}\theta_1^l\theta_2^m. \tag{33}$$

Then using definitions (21) we find that

$$\left(\frac{d}{d\theta_1}\right)_L f(\theta_1,\theta_2)=[\mathscr{D}_{L1},f(\theta_1,\theta_2)]_L=\left[\mathscr{D}_{L1},\sum_{l,m=0}^{\infty}C_{l,m}\theta_1^l\theta_2^m\right]_L$$
$$=\sum_{l,m=0}^{\infty}C_{l,m}[\mathscr{D}_{L1},\theta_1^l\theta_2^m]_{q_1^l t_{21}^m}$$
$$=\sum_{l,m=0}^{\infty}[l+1]_{q_1}C_{l+1,m}\theta_1^l\theta_2^m. \tag{34}$$

Similarly, for right differentiation we find

$$\left(\frac{d}{d\theta_i}\right)_R f(\theta_1,\theta_2)=[\mathscr{D}_{R1},f(\theta_1,\theta_2)]_R=\sum_{l,m=0}^{\infty}[l+1]_{q_1^{-1}}C_{l+1,m}\theta_1^l\theta_2^m. \tag{35}$$

More generally the $C_{l,m}$ can be functions of θ_j where $j\neq 1,2$. This does not affect the result of the above differentiation, but of course the explicit form given in the third line of (34) is no longer valid. In fact, since for $a\neq b$, $t_{ab}=t_{ba}=\omega_{ab}$, we have for $C=C(\theta_j)$ with $j\neq i$,

$$[\mathscr{D}_{Li},C]_L=[\mathscr{D}_{Li},C]_{q_{ci}}=0,$$
$$[\mathscr{D}_{Ri},C]_R=[\mathscr{D}_{Ri},C]_{q_{ci}}=0, \tag{36}$$

which are the braided bracket analogs of (3.10) and (3.11). Note that unlike in the graded bracket case considered in Ref. 2 these conditions are not additional constraints on C, but instead follow directly from our definition of the many variable q-calculus. Note also that the conditions for induced left and right differentiation are compatible, so that in the many variable case, working with braided brackets, it is not necessary to choose between these. Another difference between the graded bracket induced derivatives of Ref. 2 and the braided bracket induced derivatives of the present paper is that in the latter case $\mathscr{D}_{Ra}$ appears on the left of the braided bracket. One consequence of this is that here $\mathscr{D}_{Ra}$ has a different normalization. In the many variable case, and with the new normalization, the number and shift operators are as follows:

$$N_a=\sum_{m=0}^{\infty}\frac{(1-(q_a)^{m-1})}{[m]_{q_a}}\theta_a^m\mathscr{D}_{La}^m=\sum_{m=0}^{\infty}\frac{(1-(q_a)^{1-m})}{[m]_{q_a^{-1}}}\theta_a^m\mathscr{D}_{Ra}^m, \tag{37}$$

$$q_a^{kN_a}=\sum_{m=0}^{\infty}\frac{1}{[m]_{q_a}}\left(\prod_{l=1}^{m-1}(q_a^k-q_a^l)\right)\theta_a^m\mathscr{D}_{La}^m,$$
$$q_a^{-kN_a}=\sum_{m=0}^{\infty}\frac{1}{[m]_{q_a^{-1}}}\left(\prod_{l=1}^{m-1}(q_a^{-k}-q_a^{-l})\right)\theta_a^m\mathscr{D}_{Ra}^m,\tag{38}$$

$$G_{La}=\exp_{q_a^{-1}}(\epsilon_a\mathscr{D}_{La}),\quad G_{Ra}=\exp_{q_a}(\eta_a\mathscr{D}_{Ra}).\tag{39}$$

These satisfy

$$[N_a,\theta_b]=\delta_{ab}\theta_a,\quad [N_a,\mathscr{D}_{Lb}]=-\delta_{ab}\mathscr{D}_{Lb},\quad [N_a,\mathscr{D}_{Rb}]=-\delta_{ab}\mathscr{D}_{Rb},\tag{40}$$

$$G_{La}\theta_bG_{La}^{-1}=\delta_{ab}\epsilon_a+\theta_b,\quad G_{Ra}\theta_bG_{Ra}^{-1}=\theta_b+\delta_{ab}\eta_a.\tag{41}$$

Using the identity $\mathscr{D}_{La}\theta_a-\theta_a\mathscr{D}_{La}=q_a^{N_a}$ which follows from (38), it is clear that with the braided bracket normalization the relationship between the left and right algebraic derivatives is

$$\mathscr{D}_{Ra}=q_a^{-N_a}\mathscr{D}_{La}.\tag{42}$$

It follows immediately from this and (29) or (31) that

$$[\mathscr{D}_{Ra},\mathscr{D}_{Lb}]_{t_{ab}}=0.\tag{43}$$

Another consequence of this change of normalization is that the Q_a and D_a are related by

$$Q_a=\mathscr{D}_{La},\quad D_a=\mathscr{D}_{Ra}.\tag{44}$$

Left and right integrals[13] can also be introduced. As in the one-dimensional case these are defined so as to invert the effect of the corresponding derivatives. Another important advantage of the switch to braided brackets is that the left and right integrals are truly distinct, and that there are simple and well-defined commutation relations among these as well as between them and the derivatives. Specifically, the left integrals are defined by

$$\int(d\theta_a)_L\theta_a^m=\frac{\theta_a^{m+1}}{[m+1]_{q_a}},\tag{45}$$

and the right integrals by

$$\int(d\theta_a)_R\theta_a^m=\frac{\theta_a^{m+1}}{[m+1]_{q_a^{-1}}}.\tag{46}$$

To integrate functions of many variables we also need the cross relations

$$\left[\int(d\theta_a)_L,\theta_b\right]_{\omega_{ab}}=\left[\int(d\theta_a)_R,\theta_b\right]_{\omega_{ab}}=0,\tag{47}$$

which hold for $a\neq b$. It is also straightforward to show, for example by comparing

$$\left(\frac{d}{d\theta_a}\right)_L\int(d\theta_a)_R\theta_a^m=\frac{[m+1]_{q_a}}{[m+1]_{q_a^{-1}}}\ \theta_a^m,\tag{48}$$

with

$$\int (d\theta_a)_R \left(\frac{d}{d\theta_a}\right)_L \theta_a^m = \frac{[m]_{q_a}}{[m]_{q_a^{-1}}} \theta_a^m, \tag{49}$$

that the commutation relations between differentiation and integration are as follows:

$$\begin{aligned} &\left[\left(\frac{d}{d\theta_a}\right)_L, \int (d\theta_b)_L\right]_{\omega_{ba}} = 0, \quad \left[\left(\frac{d}{d\theta_a}\right)_L, \int (d\theta_b)_R\right]_{t_{ba}} = 0, \\ &\left[\left(\frac{d}{d\theta_a}\right)_R, \int (d\theta_b)_R\right]_{\omega_{ba}} = 0, \quad \left[\left(\frac{d}{d\theta_a}\right)_R, \int (d\theta_b)_L\right]_{t_{ab}^{-1}} = 0. \end{aligned} \tag{50}$$

By similar methods we also find

$$\left[\int (d\theta_a)_L, \int (d\theta_b)_L\right]_{\omega_{ab}} = 0, \quad \left[\int (d\theta_a)_R, \int (d\theta_b)_R\right]_{\omega_{ab}} = 0, \quad \left[\int (d\theta_a)_R, \int (d\theta_b)_L\right]_{t_{ab}} = 0, \tag{51}$$

which are the integral analogs of (29), (31), and (43).

III. GENERALIZED GRASSMANN CALCULUS AT q_a A ROOT OF UNITY

One of the central results of Ref. 2 was that if $\tilde{q}_a$ is a primitive n_ath root of unity, and z_a, ∂_{z_a} are defined by

$$z_a = \lim_{q_a \to \tilde{q}_a} \frac{(\theta_a)^{n_a}}{[n_a]_{q_a}!}, \quad \partial_{z_a} = \mathscr{D}_{La}^{n_a} = -(-1)^{n_a} \mathscr{D}_{Ra}^{n_a}, \tag{52}$$

in which (42) has been used, and it is assumed that $(\theta_a)^{n_a} \to 0$ as $q_a \to \tilde{q}_a$ in such a way that z_a is well defined in this limit, then

$$[\partial_{z_a}, z_a] = 1. \tag{53}$$

Using these definitions and the results of Sec. IV B it is easy to establish the full commutation relations in the limit as $q_a \to \tilde{q}_a$ (note that this limit need not be taken for all a). When $q_a \to \tilde{q}_a$ and $q_b \to \tilde{q}_b$ we find from (26), (29), and (53) that

$$[\partial_{z_a}, z_b]_{(t_{ba})^{n_a n_b}} = \delta_{ab}, \quad [z_a, z_b]_{(\omega_{ab})^{n_a n_b}} = 0, \quad [\partial_{z_a}, \partial_{z_b}]_{(\omega_{ab})^{n_a n_b}} = 0. \tag{54}$$

This clearly reduces to ordinary calculus if $(\omega_{ba})^{n_a n_b} = 1$. It will often be sensible to make this choice. It also follows from (29), (31), and (52) that

$$\begin{aligned} [\mathscr{D}_{La}, z_b]_{(t_{ba})^{n_b}} &= \delta_{ab} \frac{(\theta_b)^{n_b - 1}}{[n_b - 1]_{q_b}!}, \\ [\mathscr{D}_{Ra}, z_b]_{(t_{ab})^{-n_b}} &= -(-1)^{n_b} \delta_{ab} \frac{(\theta_b)^{n_b - 1}}{[n_b - 1]_{q_b^{-1}}!}, \end{aligned} \tag{55}$$

and that

$$[\mathscr{D}_{La}, \partial_{z_b}]_{(\omega_{ab})^{n_b}} = 0, \quad [\mathscr{D}_{Ra}, \partial_{z_b}]_{(\omega_{ab})^{n_b}} = 0, \quad [\partial_{z_b}, \theta_a]_{(t_{ab})^{n_b}} = 0. \tag{56}$$

Note that (55) and (56) hold even when q_a is not a root of unity, as long as q_b is. Following Ref. 2 we can, when q_a is a root of unity, expand the algebraic *total* derivatives $\mathscr{D}_{La}$ and $\mathscr{D}_{Ra}$ by using the algebraic partial derivatives ∂_{θ_a} and δ_{θ_a}. These satisfy

$$(\partial_{\theta_a})^{n_a}=(\delta_{\theta_a})^{n_a}=0, \quad [\delta_{\theta_a},\partial_{\theta_b}]_{t_{ab}}=0,$$
$$[\partial_{\theta_a},\partial_{\theta_b}]_{\omega_{ab}}=0, \quad [\delta_{\theta_a},\delta_{\theta_b}]_{\omega_{ab}}=0, \tag{57}$$

as well as

$$[\partial_{\theta_a},\theta_b]_{t_{ba}}=\delta_{ab}, \quad [\partial_{\theta_a},z_b]_{(t_{ba})^{n_b}}=0, \quad [\partial_{\theta_a},\partial_{zb}]_{(t_{ba})^{-n_b}}=0,$$
$$[\delta_{\theta_a},\theta_b]_{(t_{ab})^{-1}}=\delta_{ab}, \quad [\delta_{\theta_a},z_b]_{(t_{ab})^{-n_b}}=0, \quad [\delta_{\theta_a},\partial_{z_b}]_{(t_{ab})^{n_b}}=0. \tag{58}$$

So that if we expand $\mathscr{D}_{La}$ and $\mathscr{D}_{Rb}$ as follows:

$$\mathscr{D}_{La}=\partial_{\theta_a}+\frac{\theta_a^{n_a-1}}{[n_a-1]_{q_a}!}\,\partial_{z_a},$$
$$\mathscr{D}_{Ra}=\delta_{\theta_a}-(-1)^{n_a}\frac{\theta_a^{n_a-1}}{[n_a-1]_{q_a^{-1}}!}\,\partial_{z_a}, \tag{59}$$

then (54) and (55) are implied by (57)–(59).
If we note the identity

$$\lim_{q_a\to\tilde{q}_a}\frac{\theta_a^{rn_a+p}}{[rn_a+p]_{q_a}}=\frac{z_a^r\theta^p}{r![p]!}, \tag{60}$$

then we can take the limit of (45) and (46) to obtain

$$\int (d\theta_a)_L z_a^r\theta_a^p=(1-\delta_{p,n-1})\frac{z_a^r\theta_a^{p+1}}{[p+1]_{q_a}}+\delta_{p,n-1}[n_a-1]_{q_a}!\,\frac{z_a^{r+1}\theta_a^p}{(r+1)},$$
$$\int (d\theta_a)_R z_a^r\theta_a^p=(1-\delta_{p,n-1})\frac{z_a^r\theta_a^{p+1}}{[p+1]_{q_a^{-1}}}-(-1)^{n_a}\delta_{p,n-1}[n_a-1]_{q_a^{-1}}!\,\frac{z_a^{r+1}\theta_a^p}{(r+1)}. \tag{61}$$

In analogy with the introduction of partial derivatives, we introduce the following ''partial'' integrals:

$$\int d\theta_a z_a^r\theta_a^p=(1-\delta_{p,n-1})\frac{z_a^r\theta_a^{p+1}}{[p+1]_{q_a}},$$
$$\int \delta\theta_a z_a^r\theta_a^p=(1-\delta_{p,n-1})\frac{z_a^r\theta_a^{p+1}}{[p+1]_{q_a^{-1}}}, \tag{62}$$
$$\int dz_a z_a^r\theta_a^p=\frac{z_a^{r+1}\theta_a^p}{(r+1)}.$$

Using these and (61) we can write

$$\int (d\theta_a)_L=\int d\theta_a+\frac{\partial^{n_a-1}}{\partial^{n_a-1}\theta_a}\int dz_a,$$
$$\int (d\theta_a)_R=\int \delta\theta_a-(-1)^{n_a}\frac{\delta^{n_a-1}}{\delta^{n_a-1}\theta_a}\int dz_a, \tag{63}$$

which are the integral analogs of (59). We also note the identities

$$\int (d\theta_a)_L^n = \int dz_a, \quad \int (d\theta_a)_R^n = -(-1)^{n_a} \int dz_a, \quad \int d\theta_a^n = \int \delta\theta_a^n = 0. \tag{64}$$

We conclude this section with some comments on the braided Hopf structure of the generalized Grassmann algebra and the dual algebra of derivatives as $q_a \to \tilde{q}_a$. For an alternative derivation of (52) as well as a derivation of the duality properties in the single variable case see the Appendix. The results of the Appendix are easily extended to the many variable case, and we give the results below. For generic q_a the braided Hopf structure of θ_a, which follows directly from (5) is as follows:

$$\Delta\theta_a = \theta_a \otimes 1 + 1 \otimes \theta_a, \quad \epsilon(\theta_a) = 0, \quad S(\theta_a^m) = q^{m(m-1)/2}(-\theta_a)^m. \tag{65}$$

When $q_a \to \tilde{q}_a$ it follows directly from this and (52) that in addition to (65), which holds as in the generic case, we have the following braided Hopf structure for z_a:

$$\Delta z_a = z_a \otimes 1 + 1 \otimes z_a + \sum_{m=1}^{n_a-1} \frac{\theta_a^m \otimes \theta_a^{n_a-m}}{[m]_{q_a}![n_a-m]_{q_a}!},$$

$$\epsilon(z_a) = 0, \quad S(z_a) = -z_a. \tag{66}$$

In the dual Hopf algebra with elements $\mathscr{D}_{La}$, the braided Hopf structure is as follows:

$$\Delta\mathscr{D}_{La} = \mathscr{D}_{La} \otimes 1 + 1 \otimes \mathscr{D}_{La},$$

$$\epsilon(\mathscr{D}_{La}) = 0, \quad S(\mathscr{D}_{La}) = q^{m(m-1)/2}(-\mathscr{D}_{La})^m. \tag{67}$$

The duality is given by the inner product

$$\langle \mathscr{D}_{La}, \theta_b \rangle = \delta_{ab}, \tag{68}$$

which satisfies/is extended to products by all of the usual identities (see the Appendix)—Eq. (A8). When $q_a \to \tilde{q}_a$ the Hopf structure is extended to include

$$\Delta\partial_{za} = \partial_{za} \otimes 1 + 1 \otimes \partial_{za}, \quad \epsilon(\partial_{za}) = 0, \quad S(\partial_{za}) = -\partial_{za}. \tag{69}$$

In this case the duality is given by

$$\langle \mathscr{D}_{La}, \theta_b \rangle = \delta_{ab}, \quad \langle \partial_{za}, z_b \rangle = \delta_{ab}, \quad \langle \partial_{za}, \theta_b \rangle = 0, \quad \langle \mathscr{D}_{La}, z_b \rangle = 0, \tag{70}$$

which follow directly from (52) and (68). Note that we could equally well have worked with $\mathscr{D}_{Ra}$, the only advantage of using $\mathscr{D}_{La}$ being that we avoid the factors of $(-1)^{n_a+1}$ which would arise due to (52).

IV. THE BRAIDED INTERPRETATION OF TWO-DIMENSIONAL SUSY

We can use the work in the previous sections to extend our new interpretation of supersymmetry (SUSY) to the two-dimensional case. This is of great interest in physics since it is related to world sheet supersymmetry in superstring theory. The most interesting new feature in two dimensions is the presence of Lorentz transformations. We consider a two-dimensional generalized Grassmann algebra $\{\theta_a\}$, $a=1,2$, and its associated calculus, examining first the case of $q_1 = q_2 = \omega_{12} = q$. We begin by *defining*

$$p_\mu = -\tfrac{1}{2}\mathscr{D}_{La}(\gamma_\mu\gamma_0)_{ab}\mathscr{D}_{Lb},$$

$$x_\mu = \lim_{q\to -1} \frac{i}{[2]_q}\, \theta_a(\gamma_0\gamma_\mu)_{ab}\theta_b. \tag{71}$$

Here $\mu=0,1$ and $\gamma_0=\sigma_2$, $\gamma_1=i\sigma_1$, where σ_a are the usual Pauli matrices, so that we are working in the Majorana–Weyl basis for the Dirac gamma matrices. Note that other than those implied by the right-hand side, no transformation properties are assigned to p_μ and x_μ. Since $\gamma_0\gamma_\mu$ is diagonal, the above can be written as

$$x_0=i(z_1+z_2), \quad p_0=-\tfrac{1}{2}(\partial_{z_1}+\partial_{z_2}),$$
$$x_1=i(z_1-z_2), \quad p_1=\tfrac{1}{2}(\partial_{z_1}-\partial_{z_2}), \tag{72}$$

from which, using (54), it is clear that

$$[p_\mu,x_\nu]=-ig_{\mu\nu}, \quad [p_\mu,p_\nu]=0, \quad [x_\mu,x_\nu]=0. \tag{73}$$

Here $g_{\mu\nu}=\text{diag}\{1,-1\}$ so that $\{p_\mu\}$ and $\{x_\mu\}$ behave just like the quantized momenta and coordinates of two-dimensional spacetime. To establish their transformation properties, we proceed as follows. Under a translation

$$\theta_a\to\epsilon_a+\theta_a, \tag{74}$$

we find from (71) that the coordinates $\{x_\mu\}$ transform as follows:

$$x_\mu\to\lim_{q\to-1}\frac{i}{[2]_q}(\epsilon_a+\theta_a)(\gamma_0\gamma_\mu)_{ab}(\epsilon_b+\theta_b)$$
$$=\lim_{q\to-1}\frac{i}{[2]_q}\theta_a(\gamma_0\gamma_\mu)_{ab}\theta_b$$
$$+\lim_{q\to-1}\frac{i}{[2]_q}\epsilon_a(\gamma_0\gamma_\mu)_{ab}\epsilon_b+i\epsilon_a(\gamma_0\gamma_\mu)_{ab}\theta_b=x_\mu+x'_\mu+i\epsilon_a(\gamma_0\gamma_\mu)_{ab}\theta_b. \tag{75}$$

Together (74) and (75) constitute the usual two-dimensional SUSY transformation,[10] only now we can see that just as the $\{x_\mu\}$ are defined by (71) in terms of the $\{\theta_a\}$, the $\{x'_\mu\}$ are defined by

$$x'_\mu=\lim_{q\to-1}\frac{i}{[2]_q}\epsilon_a(\gamma_0\gamma_\mu)_{ab}\epsilon_b, \tag{76}$$

which is the same as (71) but with θ_a replaced by ϵ_a. In the notation of generalized Grassmann calculus, the infinitesimal generators of the translation (74) are $\epsilon_a\mathscr{D}_{La}$. On the other hand, in the usual SUSY notation, this transformation is generated by the supercharge Q_a, and thus (as expected) we can make the identification

$$\mathscr{D}_{La}=Q_a. \tag{77}$$

Using this we can write the definition (71) of p_μ as

$$p_\mu=-\tfrac{1}{2}Q_a(\gamma_\mu\gamma_0)_{ab}Q_b, \tag{78}$$

which can easily be inverted to yield

$$\{Q_a,Q_b\}=-2(\gamma_0\gamma_\mu)_{ab}p^\mu, \tag{79}$$

where $p^\mu=g^{\mu\nu}p_\mu$. Along with $[p_\mu,p_\nu]=0$ from (73), this is just the two-dimensional supersymmetry algebra in its usual form. The usual superspace realization of this algebra can be obtained by using (59) and (71). We find

$$Q_a=\mathscr{D}_{La}=\partial_{\theta_a}+\theta_a\partial_{z_a}=\partial_{\theta_a}-(\gamma_0\gamma_\mu)_{ab}\theta_b p^\mu. \tag{80}$$

The covariant derivatives D_a from two-dimensional SUSY also arise naturally in the $q\to -1$ limit of two-dimensional q-calculus. To see this, we write down their usual superspace realization,

$$D_a = \partial_{\theta_a} + (\gamma_0\gamma_\mu)_{ab}\theta_b p^\mu. \tag{81}$$

Then, using (38) and (57)

$$\partial_{\theta_a} = (-1)^{N_a}\delta_{\theta_a} = (\mathscr{D}_{La}\theta_a - \theta_a\mathscr{D}_{La})\delta_{\theta_a} = (\partial_{\theta_a}\theta_a - \theta_a\partial_{\theta_a})\delta_{\theta_a} = \delta_{\theta_a}, \tag{82}$$

we can write this as

$$D_a = \delta_{\theta_a} + (\gamma_0\gamma_\mu)_{ab}\theta_b p^\mu, \tag{83}$$

and thus from (59) and (71) we have

$$D_a = \mathscr{D}_{Ra}. \tag{84}$$

These satisfy

$$\{D_a, D_b\} = 2(\gamma_0\gamma_\mu)_{ab}p^\mu. \tag{85}$$

The cross relations $\{D_a, Q_b\} = 0$ follow directly from (43). Thus the supercharges and covariant derivatives used in two-dimensional supersymmetry, correspond, respectively, to the left and right *total* derivatives in the $q_a \to -1$ limit of two-dimensional q-calculus.

In two-dimensional SUSY, the Grassmann variables θ_a transform as the components of a Lorentz spinor,

$$\theta_a \to S_{ab}\theta_b, \tag{86}$$

where

$$S_{ab} = \begin{pmatrix} \exp\left(\frac{\phi}{2}\right) & 0 \\ 0 & \exp\left(\frac{-\phi}{2}\right) \end{pmatrix}. \tag{87}$$

Due to (71), the transformation properties of the coordinates $\{x_\mu\}$ are entirely determined by those of the θ_a. To find these explicitly we first note that

$$\gamma_0^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \gamma_0\gamma_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{88}$$

so that

$$\begin{aligned} S\gamma_0^2 S^T &= \begin{pmatrix} \exp\phi & 0 \\ 0 & \exp(-\phi) \end{pmatrix} = \gamma_0^2 \cosh\phi + \gamma_0\gamma_1 \sinh\phi, \\ S\gamma_0\gamma_1 S^T &= \begin{pmatrix} \exp\phi & 0 \\ 0 & -\exp(-\phi) \end{pmatrix} = \gamma_0^2 \sinh\phi + \gamma_0\gamma_1 \cosh\phi. \end{aligned} \tag{89}$$

Now from (71) and (86), we find that under a Lorentz transformation the coordinates $\{x_\mu\}$ behave as follows:

$$x_\mu \to \lim_{q\to -1} \frac{i}{[2]_q}\theta_b S_{ab}(\gamma_0\gamma_\mu)_{ab}S_{cd}\theta_d = \lim_{q\to -1}\frac{i}{[2]_q}\theta_a \Lambda_\mu^\nu(\gamma_0\gamma_\nu)_{ab}\theta_b = \Lambda_\mu^\nu x_\nu, \tag{90}$$

in which from (89) Λ_μ^ν has the form

$$\Lambda_\mu^{\ \nu}=\begin{pmatrix}\cosh\phi & \sinh\phi\\ \sinh\phi & \cosh\phi\end{pmatrix}, \tag{91}$$

so we have shown that, as expected, the coordinates $\{x_\mu\}$ transform like the components of a covariant Lorentz vector. Note that $x^\mu:=g^{\mu\nu}x_\nu$ also has the expected transformation properties, i.e., the $\{x^\mu\}$ transform like the components of a contravariant Lorentz vector

$$x'^\mu=(\Lambda_\nu^\mu)^{-1}x^\nu, \tag{92}$$

so that the length $x_\mu x^\mu$ is invariant. Note also that from $\{D_{La},\theta_b\}=\delta_{ab}$ and (86) it follows that under a Lorentz transformation $D_{La}\rightarrow D_{Lb}S_{ba}^{-1}$, and that through definition (71), this leads to the correct transformation properties for the $\{p_\mu\}$. By combining the translation and Lorentz transformation above, we can consider the effect on the coordinates $\{x_\mu\}$ of a general super-Poincaré transformation of θ:

$$\theta_a\rightarrow\epsilon_a+S_{ab}\theta_b. \tag{93}$$

Under such a transformation we have, from (71)

$$\begin{aligned}x_\mu\rightarrow&\lim_{q\rightarrow-1}\frac{i}{[2]_q}(\epsilon_a+S_{ab}\theta_b)^2\\ =&\lim_{q\rightarrow-1}\frac{i}{[2]_q}\epsilon_a(\gamma_0\gamma_\mu)_{ab}\epsilon_b\\ &+\lim_{q\rightarrow-1}\frac{i}{[2]_q}\theta_bS_{ab}(\gamma_0\gamma_\mu)_{ac}S_{cd}\theta_d+i\epsilon_a(\gamma_0\gamma_\mu)_{ab}S_{bc}\theta_c,\end{aligned} \tag{94}$$

which by (75) and (90) reduces to

$$x_\mu\rightarrow x'_\mu+\Lambda_\mu^{\ \nu}x_\nu+i\epsilon_a(\gamma_0\gamma_\mu)_{ab}S_{bc}\theta_c, \tag{95}$$

which is in exact agreement with the usual super-Poincaré transformation of $\{x_\mu\}$.

Although it seems reasonable to expect that there is an analogous interpretation of super-Poincaré transformations in higher dimensions, based on (71) or some similar relationship, and it is indeed straightforward to construct higher dimensional algebras with supersymmetric properties using our techniques, the generalization of our work in this section to $d>2$ is a nontrivial problem, and at present it remains unsolved.

V. MIXED FSUSY IN TWO DIMENSIONS

Using (72) it is clear that in terms of $\{\theta_a\}$ and $\{z_a\}$ the general super-Poincaré transformation (95) of the coordinates $\{x_\mu\}$ which follows from (93) takes on the following simple form:

$$\begin{aligned}z_1&\rightarrow z_1\exp\phi+z'_1+\epsilon_1\theta_1\exp(\phi/2),\\ z_2&\rightarrow z_2\exp(-\phi)+z'_2+\epsilon_2\theta_2\exp(-\phi/2).\end{aligned} \tag{96}$$

The fact that the pairs $\{z_1,\theta_1\}$ and $\{z_2,\theta_2\}$ are not mixed by this transformation has the consequence that in this basis the generalization to fractional supersymmetry (FSUSY) is straightforward. To construct the most general two-dimensional FSUSY, we consider a two-dimensional q-calculus in the limit as $q_i\rightarrow\tilde{q}_i$, and choose ω_{12} so that $\omega_{12}^{n_1}=\omega_{12}^{n_2}=1$. We have included the $n_1\neq n_2$ case, and for this reason refer to our construction as *mixed* FSUSY. A suitable definition for S_{ab} in the Lorentz transformation $\theta_a\rightarrow S_{ab}\theta_b$ of mixed anyonic spinors, such as θ_a is

$$S_{ab}=\begin{pmatrix} \exp\left(\frac{\phi}{n_1}\right) & 0 \\ 0 & \exp\left(\frac{-\phi}{n_2}\right) \end{pmatrix}. \tag{97}$$

As we will see, this ensures that $\{x_\mu\}$ transform as the components of a Lorentz vector. Under a mixed anyonic Poincaré transformation

$$\theta_a \to \epsilon_a + S_{ab}\theta_b, \tag{98}$$

it follows from (52) that z_1 and z_2 transform as follows:

$$\begin{aligned} z_1 &\to z_1 \exp \phi + z_1' + \sum_{m=1}^{n_1-1} \frac{\epsilon_1^m \theta_1^{n_1-m}}{[n_1-m]_{q_1}![m]_{q_1}!} \exp\left(\frac{(n_1-m)\phi}{n_1}\right), \\ z_2 &\to z_2 \exp(-\phi) + z_1' + \sum_{m=1}^{n_2-1} \frac{\epsilon_2^m \theta_2^{n_2-m}}{[n_2-m]_{q_2}![m]_{q_2}!} \exp\left(\frac{(m-n_2)\phi}{n_2}\right). \end{aligned} \tag{99}$$

To make contact with the usual spacetime coordinates $\{x_\mu\}$ we note that as in the $n_a=n_b=2$ case z_1z_2 is invariant under a pure Lorentz transformation $(\epsilon_1=\epsilon_2=0)$. Thus we have $z_1z_2 \propto x_0^2-x_1^2$. In fact the definitions of x_0 and x_1 in terms of z_1 and z_2 are

$$\begin{aligned} x_0 &= F(z_1+z_2), \quad p_0=-\frac{i}{2F}(\partial_{z_1}+\partial_{z_2}), \\ x_1 &= F(z_1-z_2), \quad p_1=\frac{i}{2F}(\partial_{z_1}-\partial_{z_2}), \end{aligned} \tag{100}$$

with $F=i$ for even n as in (72) and $F=1$ for odd n. These factors correspond to those relating t to z in Refs. 1–4 and ensure the reality of p_μ and x_μ. From (54) it follows that the operators defined by (100) satisfy (73) as in the supersymmetric case covered in Sec. IV. After a little algebra we obtain the mixed anyonic transformation of the $\{x_\mu\}$ coordinates,

$$x_\mu \to x_\mu' + \Lambda_\mu^{\ \nu} x_\nu + \sum_{a,b=1}^{2} \sum_{m=1}^{n_a-1} \frac{F\epsilon_a^m (\gamma_0\gamma_\mu)_{ab}(S_{bc}\theta_c)^{n_a-m}}{[n_a-m]_{q_a}![m]_{q_a}!}. \tag{101}$$

Here $\Lambda_\mu^{\ \nu}$ is the same as in (91). The fractional supercharge and covariant derivative are also easy to deduce. From (59) and (100) we find

$$Q_a=\mathscr{D}_{La}=\partial_{\theta_a}+\frac{\theta_a^{n_a-1}}{[n_a-1]_{q_a}!}\partial_{za}=\partial_{\theta_a}+\frac{iF}{[n_b-1]_{q_b}!}(\gamma_0\gamma_\mu)_{ab}\theta_b^{n_b-1}p^\mu, \tag{102}$$

and

$$D_a=\mathscr{D}_{Ra}=\delta_{\theta_a}-(-1)^{n_a}\frac{\theta_a^{n_a-1}}{[n_a-1]_{q_a^{-1}}!}\partial_{za}=\delta_{\theta_a}-\frac{iF(-1)^{n_a}}{[n_b-1]_{q_b^{-1}}!}(\gamma_0\gamma_\mu)_{ab}\theta_b^{n_b-1}p^\mu. \tag{103}$$

The algebraic (left) integral of a function $f(z_1,z_2,\theta_1,\theta_2)$ on two-dimensional fractional superspace is

$$\int (d\theta_2)_L \int (d\theta_1)_L f(z_1,z_2,\theta_1,\theta_2). \tag{104}$$

Note that this integral would change by an overall multiplicative factor if we reversed the order of $\int(d\theta_2)_L$ and $\int(d\theta_1)_L$, so that in writing down (104), we have made a choice of convention. This integral can be expanded using (61) to yield

$$\int d\theta_2 \int d\theta_1 f + \int d\theta_2 \frac{\partial^{n_1-1}}{\partial^{n_1-1}\theta_1} \int dz_1 f + \frac{\partial^{n_2-1}}{\partial^{n_2-1}\theta_2} \int dz_2 \int d\theta_1 f + \frac{\partial^{n_2-1}}{\partial^{n_2-1}\theta_2} \frac{\partial^{n_1-1}}{\partial^{n_1-1}\theta_1} \int dz_2 \int dz_1 f. \tag{105}$$

To obtain a numerical measure from this algebraic integral we now make use of an argument similar to that given in Ref. 2. θ_1 and θ_2 are nilpotent and thus all of their eigenvalues are zero. On the other hand, the bosonic limits denoted by z_1 and z_2 are non-nilpotent and thus do have nonzero eigenvalues. After integration, the first three terms in (105) always involve θ_1 or θ_2 raised to some nonzero power, whereas the last term involves z_1 and z_2 only. Any numerical measure based on the integral (105) must be based on its eigenvalues in some representation. Consequently only the last term contributes and thus the first three can be dropped. It is convenient at this point to introduce a fractional Berezin integral as in Ref. 2,

$$\int (d\theta_a)_{\mathrm{Ber}} = \frac{\partial^{n_a-1}}{\partial^{n_a-1}\theta_a}. \tag{106}$$

The resulting numerical integral measure on two-dimensional fractional superspace can now be written as

$$I(f) = \int dz_2 dz_1 (d\theta_2)_{\mathrm{Ber}} (d\theta_1)_{\mathrm{Ber}} f(z_1, z_2, \theta_1, \theta_2). \tag{107}$$

If we expand f as a power series

$$f(z_1, z_2, \theta_1, \theta_2) = \sum_{m_1=0}^{n_1-1} \sum_{m_2=0}^{n_2-1} C_{m_1,m_2}(z_1, z_2) \frac{\theta_1^{m_1}}{[m_1]_{q_1}!} \frac{\theta_2^{m_2}}{[m_2]_{q_2}!}, \tag{108}$$

then (107) reduces to

$$I(f) = \int dz_2 dz_1 C_{n_1-1, n_2-1}(z_1, z_2). \tag{109}$$

Note that up to a constant Jacobian factor this is equal to

$$\int dx_0 dx_1 C_{n_1-1, n_2-1}(z_1, z_2), \tag{110}$$

which, for $n=2$, is just the integral which arises in supersymmetric field theories involving one space and one time dimension.

ACKNOWLEDGMENTS

Thanks to Alan Macfarlane, José de Azcárraga, and Carlos Pérez Bueno for many helpful discussions and communications. I gratefully acknowledge the financial support provided by the E.P.S.R.C. and St John's College, Cambridge during the preparation of this paper.

APPENDIX

The results of Refs. 1–4 can also be derived from a different and in some ways mathematically nicer point of view. Our work here uses a technique similar to that employed by G. I. Lusztig in his work on the properties of deformed universal enveloping algebras with deformation parameter equal to a root of unity.[14,15] To the best of our knowledge this is the first time such a technique has been applied to a braided object. Let us begin by introducing the braided Hopf algebra $\mathscr{A}$, which we define for all q. This has elements $\{\theta^{(m)}\}$, $m=0,1,2,...,\infty$ with $\theta^{(0)}=1$, and relations

$$\theta^{(m)}\theta^{(p)}=\frac{[m+p]_q!}{[m]_q![p]_q!}\,\theta^{(m+p)}, \tag{A1}$$

as well as

$$\Delta\theta^{(m)}=\sum_{r=0}^{m}\theta^{(m-r)}\otimes\theta^{(r)},\quad \epsilon(\theta^{(m)})=\delta_{m,0}, \tag{A2}$$

$$S(\theta^{(m)})=(-1)^m q^{m(m-1)/2}\theta^{(m)}.$$

The braiding is given by

$$\psi(\theta^{(m)}\otimes\theta^{(s)})=q^{ms}\theta^{(m)}\otimes\theta^{(s)}, \tag{A3}$$

so that

$$(\theta^{(r)}\otimes\theta^{(m)})(\theta^{(s)}\otimes\theta^{(s)})=q^{ms}\theta^{(r)}\theta^{(s)}\otimes\theta^{(m)}\theta^{(t)}. \tag{A4}$$

We also define $\mathscr{H}$ the braided Hopf algebra dual to $\mathscr{A}$ as follows. This has elements $\{\mathscr{D}_L^{(m)}\}$, $m=0,1,2,...,\infty$ with $\mathscr{D}_L^{(0)}=1$, and relations

$$\mathscr{D}_L^{(m)}\mathscr{D}_L^{(p)}=\mathscr{D}_L^{(m+p)}, \tag{A5}$$

as well as

$$\Delta\mathscr{D}_L^{(m)}=\sum_{r=0}^{m}\frac{[m]_q!}{[r]_q![m-r]_q!}\,\mathscr{D}_L^{(m-r)}\otimes\mathscr{D}_L^{(r)},$$

$$\epsilon(\mathscr{D}_L^{(m)})=\delta_{m,0}, \tag{A6}$$

$$S(\mathscr{D}_L^{(m)})=(-1)^m q^{m(m-1)/2}\mathscr{D}_L^{(m)}.$$

The braiding is given by

$$\psi(\mathscr{D}_L^{(m)}\otimes\mathscr{D}_L^{(s)})=q^{ms}\mathscr{D}_L^{(m)}\otimes\mathscr{D}_L^{(s)}. \tag{A7}$$

These two braided Hopf algebras are dual in the sense that there is a bilinear map $\langle,\rangle$: $\mathscr{A}\otimes\mathscr{H}\mapsto$ the complex plane, such that

$$\langle a,xy\rangle=\langle\Delta a,x\otimes y\rangle,\quad \langle ab,x\rangle=\langle a\otimes b,\Delta x\rangle,\quad \langle 1,x\rangle=\epsilon_{\mathscr{H}}(x),$$

$$\langle a,1\rangle=\epsilon_{\mathscr{A}}(a),\quad \langle S(a),x\rangle=\langle a,S(x)\rangle. \tag{A8}$$

Specifically, in this case we have

$$\langle\theta^{(m)},\mathscr{D}_L^{(p)}\rangle=\delta_{mp}, \tag{A9}$$

the compatibility of which with (A8) is easy to verify. We now consider the cases of generic q and q a root of unity separately.

(i) Generic q or $q=1$. From (A1) it follows that

$$\theta^{(m)}=\frac{\theta^{(1)}\theta^{(m-1)}}{[m]_q}=\frac{\theta^{(1)2}\theta^{(m-2)}}{[m]_q[m-1]_q}=\frac{(\theta^{(1)})^m}{[m]_q!}, \tag{A10}$$

and similarly

$$\mathscr{D}_L^{(m)}=(\mathscr{D}_L^{(1)})^m. \tag{A11}$$

Consequently, at generic q the braided Hopf algebra $\mathscr{A}$ and its dual $\mathscr{K}$ are both finite dimensional, each containing the identity, and only one other element. If we define $\theta=\theta^{(1)}$ and $\mathscr{D}_L=\mathscr{D}_L^{(1)}$ then we can write the generic q braided Hopf structure as follows. For $\mathscr{A}$ we have

$$\Delta\theta=\theta\otimes 1+1\otimes\theta,\quad \epsilon(\theta)=0,\quad S(\theta)=-\theta, \tag{A12}$$

which recovers the braided line at generic q, and for $\mathscr{K}$ we have

$$\Delta\mathscr{D}_L=\mathscr{D}_L\otimes 1+1\otimes\mathscr{D}_L,\quad \epsilon(\mathscr{D}_L)=0,\quad S(\mathscr{D}_L)=-\mathscr{D}_L. \tag{A13}$$

The duality simplifies to

$$\langle\theta,\mathscr{D}_L\rangle=1, \tag{A14}$$

and is extended to products via (A8). By comparing (A12) and (A13) with one of Refs. 5, 6, 16 we are able to identify both the braided Hopf algebra $\mathscr{A}$ and its dual $\mathscr{K}$ with the braided line when q is not a root of unity.

(ii) q a primitive nth root of unity. As in the generic q case we can use (A1) to obtain

$$\theta^{(p)}=\frac{(\theta^{(1)})^p}{[p]_q!}, \tag{A15}$$

but since $[n]_q=0$ this only works for $p=0,1,...,n-1$. However we are able to write

$$\theta^{(rn+p)}=\theta^{(rn)}\theta^{(p)}\lim_{q\to\epsilon}\frac{[rn]_q![p]_q!}{[rn+p]_q!}=\theta^{(rn)}\theta^{(p)}, \tag{A16}$$

where $r\geqslant 0$ and $0\leqslant p\leqslant n-1$. Also using (A1) we find that

$$\theta^{(rn)}=\theta^{(n)}\theta^{((r-1)n)}\lim_{q\to\epsilon}\frac{[(r-1)n]_q![n]_q!}{[rn]_q!}=\frac{\theta^{(n)}\theta^{((r-1)n)}}{r}. \tag{A17}$$

Iterating we finally obtain

$$\theta^{(rn)}=\frac{(\theta^{(n)})^r}{r!}, \tag{A18}$$

so that

$$\theta^{(rn+p)}=\frac{(\theta^{(n)})^r\theta^{(p)}}{r!}. \tag{A19}$$

Similarly, for the dual we find that

$$\mathscr{D}_L^{(rn+p)}=(\mathscr{D}_L^{(n)})^r\mathscr{D}_L^{(p)}. \tag{A20}$$

Thus when q is a root of unity ($q\neq 1$) the braided Hopf algebra $\mathscr{A}$ is finite dimensional, having two independent elements $\theta^{(1)}$ and $\theta^{(n)}$ besides the identity. The dual $\mathscr{K}$ is also finite dimensional, but it has only one independent element $\mathscr{D}_L^{(1)}$ besides the identity. It is convenient to define

$$\theta=\theta^{(1)},\quad z=\theta^{(n)},\quad \mathscr{D}_L=\mathscr{D}_L^{(1)},\quad \partial_z=\mathscr{D}_L^{(n)}. \tag{A21}$$

Using this notation, the algebraic relations (A1) reduce to $[\theta,z]=0$ and $\theta^n=0$, and the braided Hopf structure (A2) reduces to

$$\Delta\theta=\theta\otimes 1+1\otimes\theta,\quad \epsilon(\theta)=0,\quad S(\theta)=-\theta, \tag{A22}$$

and

$$\Delta z = z\otimes 1 + 1\otimes z + \sum_{m=1}^{n-1} \frac{\theta^m \otimes \theta^{n-m}}{[n-m]_q![m]_q!},$$

$$\epsilon(z)=0, \quad S(z)=-z. \tag{A23}$$

The braided Hopf structure of the dual $\mathscr{K}$ is given by

$$\Delta \mathscr{D}_L = \mathscr{D}_L \otimes 1 + 1 \otimes \mathscr{D}_L, \quad \epsilon(\mathscr{D}_L)=0, \quad S(\mathscr{D}_L) = -\mathscr{D}_L, \tag{A24}$$

which, using $\partial_z = \mathscr{D}_L^n$, implies the following braided Hopf structure for ∂_z;

$$\Delta \partial_z = \partial_z \otimes 1 + 1 \otimes \partial_z, \quad \epsilon(\partial_z)=0, \quad S(\partial_z) = -\partial_z.$$

The duality (A9) now takes on the form

$$\langle z^r \theta^p, \mathscr{D}_L^{r'n+p'} \rangle = \langle z^r \theta^p, \partial_z^r \mathscr{D}_L^{p'} \rangle = \delta_{r,r'} \delta_{p,p'} r! [p]_q!, \tag{A26}$$

so that in particular

$$\langle \theta, \mathscr{D}_L \rangle = 1, \quad \langle z, \partial_z \rangle = 1. \tag{A27}$$

Thus when q is a root of unity $\mathscr{A}$ coincides with the braided Hopf algebra which was associated in previous work with a limit of the braided line as its deformation parameter goes to a root of unity. Using this approach we have also obtained the braided Hopf structure of the dual and details of the duality when q is a root of unity (this is an alternative form of the braided line when q is a root of unity). Note also that the θ part of $\mathscr{A}$ forms a braided sub-Hopf algebra, but that the z part does not.

The advantage of the approach adopted here is that it enables us to restrict the taking of limits to purely numerical quantities, for which they are manifestly well defined. In this Appendix we have worked with left derivatives $\mathscr{D}_L$ only, but we could equally well have chosen right derivatives $\mathscr{D}_R$, for which a closely analogous treatment exists.

The relationship between the work in this Appendix and the work of Refs. 14 and 15 suggests that the latter might also have a physical interpretation in terms of supersymmetry and fractional supersymmetry. This idea will be developed further in Ref. 17.

[1] R. S. Dunne, A. J. Macfarlane, J. A. de Azcárraga, and J. C. Pérez Bueno, ''Supersymmetry from a braided point of view,'' Phys. Lett. B **387**, 294–299 (1996).
[2] R. S. Dunne, A. J. Macfarlane, J. A. de Azcárraga, and J. C. Pérez Bueno, ''Geometrical foundations of fractional supersymmetry,'' hep-th/9610087 [Int. J. Math. Phys. A].
[3] R. S. Dunne, A. J. Macfarlane, J. A. de Azcárraga, and J. C. Pérez Bueno, in ''Quantum groups and integrable systems,'' Czech. J. Phys. **46**, 1145–1152 (1996).
[4] J. A. de Azcárraga, R. S. Dunne, A. J. Macfarlane, and J. C. Pérez Bueno, in ''Quantum groups and integrable systems,'' Czech. J. Phys. **46**, 1235–1242 (1996).
[5] S. Majid, ''Introduction to braided geometry and q-Minkowski space,'' report hep-th/9410241, 1994.
[6] S. Majid, *Foundations of Quantum Group Theory* (Cambridge University Press, Cambridge, 1995).
[7] J. A. de Azcárraga and A. J. Macfarlane, J. Math. Phys. **37**, 1115 (1996).
[8] C. Ahn, D. Bernard, and A. Leclair, Nucl. Phys. B **346**, 409–439 (1990).
[9] R. Kerner, J. Math. Phys. **33**, 403–411 (1992); Z_3-*Grading and Ternary Algebraic Structures, in Symmetries in Science*, edited by B. Gruber (Plenum, New York, 1993), pp. 373–388.
[10] M. B. Green, J. H. Schartz, Witten, *Superstring Theory*, (Cambridge University Press, Cambridge, 1987), Vols. 1, 2.
[11] S. Majid, J. Math. Phys. **34**, 4843 (1993).
[12] S. Majid, J. Math. Phys. **32**, 3246–3253 (1991).
[13] A. Kempf and S. Majid, J. Math. Phys. **35**, 6802 (1994).
[14] G. Lusztig, Contemp. Math. **82**, 59–77 (1989).
[15] G. Lusztig, Geometrica Dedicata **35**, 89–114 (1990).
[16] S. Majid, in ''Spinors, Twistors, Clifford Algebras and Quantum Deformations, *Proceedings of the Second Max Born Symposium*, Wroclaw, Poland, 1992, edited by Z. Oziewicz *et al.* (Kluwer, Dordrecht, 1992).
[17] R. S. Dunne, ''Intrinsic anyonic spin through deformed geometry'' (unpublished).

On spectral properties of Harper-like models

D. J. L. Herrmann[a)] and T. Janssen
Theoretische Fysica, Katholieke Universiteit Nijmegen, Postbus 9010, 6500 GL Nijmegen, The Netherlands

(Received 16 March 1998; accepted for publication 12 August 1998)

We study spectral properties of Harper-like models by algebraic and combinatorial methods and derive sufficient conditions for the existence of spectral gaps with qualitative estimates. For this class the Chambers relation holds and we obtain an analytic expression for the representation dependent part. Models corresponding to the rectangular and triangular lattice are studied. In the second case we show that one class of spectral gaps is open for magnetic fields with ''rational magnetic flux per unit cell.'' A quantitative estimate for the gap widths is given for the anisotropic case and for ''irrational magnetic flux'' fulfilling some Liouville condition the spectrum is a Cantor set.
[S0022-2488(99)03401-5]

I. INTRODUCTION

Spectral properties of Hamiltonians describing charged particles on a two-dimensional (2D) lattice with an external magnetic field perpendicular to the lattice play a crucial role in various 2D problems of physical interest,[1] for the quantum Hall effect see also Ref. 2. In his famous work Hofstadter[3] found for the square lattice case a puzzling nesting structure of the spectra for rational values of the magnetic strength and gave some formal rules for the self-similarity of this so-called Hofstadter butterfly. This model has been investigated in several papers and rigorous results on spectral gaps, measures and other properties have been obtained, see Refs. 4–6 for more details. However, less is known about the analogous model on the triangular lattice, first discussed by Claro and Wannier.[7] It describes the situation in a crystal with hexagonal symmetry and similar behavior of the spectra has been observed.[8,9] We generalize a method used by Choi *et al.*[10] for the square lattice case and derive, under some mild extra conditions, a quantitative estimate for the gap widths of the class of Hamiltonians discussed below. In the case of the triangular lattice this leads to an explicit estimate for one class of spectral gaps.

The behavior of a Bloch electron in an external magnetic field is usually described by a tight-binding approximation. Without electron–electron interaction and with nearest-neighbor interaction only one obtains the following Hamiltonian for the square lattice:

$$H_{\mathrm{sq}}\Psi(k,l)=e^{i\eta_1(k,l)}\Psi(k+1,l)+e^{-i\eta_1(k-1,l)}\Psi(k-1,l)+e^{i\eta_2(k,l)}\Psi(k,l+1)+e^{-i\eta_2(k,l-1)}\Psi(k,l-1), \tag{1}$$

where $\eta_{1,2}(k,l)$ denotes the line integral of the vector potential of the external magnetic field from (k,l) to $(k+1,l)$ and from (k,l) to $(k,l+1)$, respectively (the compensating gauge transformation[11]). Since the magnetic field is uniform, we have the constraint

$$\eta_1(k,l)+\eta_2(k+1,l)-\eta_1(k,l+1)-\eta_2(k,l)=2\pi\frac{\Phi}{\Phi_0}\equiv\theta,$$

[a)]Electronic mail: danielh@sci.kun.nl

where Φ is the flux through the unit cell and $\Phi_0 = hc/e$. Using Landau's gauge $\eta_1(k,l)=0$, $\eta_2(k,l)=\theta k$ and the separation ansatz $\Psi(k,l)=e^{i\beta l}g(k)$ we derive the well-known Harper equation

$$g(k+1)+g(k-1)+2\cos(2\pi\theta k+\beta)g(k)=Eg(k).$$

In the mathematical literature the corresponding operator is called the (discrete) almost-Mathieu operator.

Let us describe H_{sq} in a more algebraic way by introducing the magnetic translation operators[12]

$$\begin{aligned} U\Psi(k,l) &= e^{i\eta_1(k,l)}\Psi(k+1,l), \\ V\Psi(k,l) &= e^{i\eta_2(k,l)}\Psi(k,l+1), \end{aligned} \tag{2}$$

which obey the Heisenberg commutation relation $UV=e^{2\pi i\theta}VU$. So we can write the Hamiltonian of (1) as $H_{\mathrm{sq}}=U+U^*+V+V^*$. In fact, (2) can be seen as a representation of the rotation algebra

$$A(\theta)=\left\{ \sum_{-\infty}^{\infty} a_{kl}v^k u^l \,\middle|\, a_{kl} \text{ rapidly decreasing} \right\},$$

generated by two abstract unitaries u and v satisfying

$$uv=\exp[2\pi i\theta]vu \tag{3}$$

first introduced for purely mathematical reasons.

Let us introduce some convenient notation for products of the unitaries u and v. These operators may be viewed as Weyl operators $w(m)$, $m\in\mathbb{Z}^2$ for discrete position and momentum, where the flux plays the role of the Planck constant.[13]

$$\begin{gathered} w(1,0)=u, \quad w(0,1)=v, \\ w(m+n)=\gamma^{-\sigma(m,n)}w(m)w(n) \quad (m,n\in\mathbb{Z}^2), \\ \gamma=\exp[i\pi\theta], \end{gathered} \tag{4}$$

where σ denotes the standard discrete symplectic form, $\sigma(m,n)=m_1n_2-m_2n_1$. The adjoint operation is simply expressed as

$$w(n)^*=w(n)^{-1}=w(-n). \tag{5}$$

We study the following family of Hamiltonians (self-adjoint operators):

$$\mathcal{H}_c=\sum_{m\in\mathcal{D}} c_m w(m) \tag{6}$$

with $\mathcal{D}=\{(0,\pm1),(\pm1,0),(\pm1,\pm1)\}$ and $c_m^*=c_{-m}\in\mathbb{C}$.

The article is organized as follows. Section II is devoted to the expansion coefficients of powers of $\mathcal{H}_c$ with respect to the Weyl operators (4), since they play a crucial role in our treatment. Actually, all interpretations and results of this section extend to more general Hamiltonians. In Sec. III we study $\mathcal{H}_c$ as element of the rotation algebra $A(\theta)$ and the behavior of spectral properties depending on θ. In Sec. IV we discuss the model with next nearest-neighbor interaction on the rectangular lattice and the model with nearest-neighbor interaction on the triangular lattice in a unified view. This is followed by some concluding remarks.

II. EXPANSION COEFFICIENTS

The expansion coefficients of powers of $\mathcal{H}_c$ with respect to the Weyl operators can be seen as a generalization of the binomial coefficients, leading to a geometric interpretation of them. We derive an analytic expression for the expansion coefficients.

A. Definition and symmetries

The well-known binomial coefficient has many different enumerative, geometric and algebraic meanings. Newton's binomial formula is one way to define it,

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k},$$

where a and b commute. Straightforward generalization yields to the multinomial coefficient. A ''q analog'' of the binomial coefficient first appeared in the literature in Ref. 14. Denote $\mathbb{Z}[u,v,\gamma]$ the ring generated by elements u, v and γ with relations[15]

$$uv=\gamma^2 vu, \quad \gamma u=u\gamma, \quad \gamma v=v\gamma,$$

then the ''q-binomial coefficient'' is uniquely defined by

$$(u+v)^n=\sum_{k=0}^{n}\binom{n}{k}_{\gamma}u^k v^{n-k},$$

where $\binom{n}{k}_\gamma$ is an element in $\mathbb{Z}[\gamma]$ (=the center of $\mathbb{Z}[u,v,\gamma]$). For the q-binomial coefficient a formula is known.[10,16] However, less is known about the q-multinomial coefficient or other sorts of generalizations.

The expansion coefficients for powers of $u+u^{-1}+v+v^{-1}$ lie also in $\mathbb{Z}[\gamma]$ and can be seen as a generalization of the q-binomial coefficients. This idea extends to any (self-adjoint) element $\mathcal{H}$ in the (finite) span of $\{w(n)|n\in\mathbb{Z}^2\}$. The expansion coefficient $(n;k)_{\mathcal{H}}$ for the kth power of $\mathcal{H}$ is uniquely defined by

$$\mathcal{H}^k=\left(\sum_{n\in\mathbb{Z}^2}c_n w(n)\right)^k=\sum_{n\in\mathbb{Z}^2}(n;k)_{\mathcal{H}}w(n), \tag{7}$$

where $(n;k)_{\mathcal{H}}$ again belongs to $\mathbb{Z}[\gamma]$. The second sum is also finite, since only a finite number of $(n;k)_{\mathcal{H}}$ are nonzero.

We are especially interested in the expansion coefficients for the powers of $\mathcal{H}_c$ defined in (6). In the following we restrict ourselves to this class and suppress the dependence on $\mathcal{H}_c$ whenever there is no ambiguity.

Writing ${\mathcal{H}_c}^{k+1}$ as ${\mathcal{H}_c}^k\mathcal{H}_c$ and $\mathcal{H}_c{\mathcal{H}_c}^k$ with coefficients $(n;k+1)$ and $(n;k)$ we get two ''conjugated'' recurrence relations

$$(n;k+1)=\sum_{m\in\mathcal{D}}c_m\gamma^{\sigma(n,m)}(n-m;k)=\sum_{m\in\mathcal{D}}c_m\gamma^{-\sigma(n,m)}(n-m;k) \tag{8}$$

with initial condition $(n;0)=\delta_n$.

Proposition 2.1: Let $\mathcal{H}_c$ be defined by (6). Then $(n;k)$ is a polynomial in c_m with real coefficients and therefore $(n;k)_{\mathcal{H}_c}=\overline{(-n;k)}_{\mathcal{H}_c}$ holds, particularly $(n;k)_{\mathcal{H}_c}=(-n;k)_{\mathcal{H}_c}\in\mathbb{R}$ for $c_{m\in\mathcal{D}}\in\mathbb{R}$.

Proof: Since we have the real initial condition $(n;0)=\delta_n$ and the recurrence relation

$$(n;k+1)=\sum_{m\in\mathcal{D}}c_m(\gamma^{\sigma(n,m)}+\gamma^{-\sigma(n,m)})/2(n-m;k), \tag{9}$$

obtained from (8), we get by induction that $(n;k)$ is a polynomial in c_m with real coefficients. The second property follows from (5) and the self-adjointness of $\mathcal{H}_c$,

$$\sum_{n\in\mathbb{Z}^2}(n;k)w(n)=\mathcal{H}_c{}^k=(\mathcal{H}_c{}^k)^*=\sum_{n\in\mathbb{Z}^2}\overline{(-n;k)}w(n). \qquad \square$$

Remark: The Hamiltonian $\mathcal{H}_c$ for a particle on a lattice in a homogeneous magnetic field has real coefficients c_m. For Hamiltonians $\mathcal{H}_c$ with additional symmetries in the parameters c_m we get further relations for the coefficients $(n;k)_{\mathcal{H}_c}$.

Proposition 2.2: Let $\Phi\in GL(2,\mathbb{Z})$. If c is such that for all $m\in\mathcal{D},c_m=s^{-1}c_{\Phi(m)}$ holds, then $(n;k)_{\mathcal{H}_c}=s^k(\Phi(n);k)_{\mathcal{H}_c}$.

A symmetry of $\mathcal{H}_c$ gives rise to such an automorphism. For example in the square lattice case the rotation by $\pi/2$ induces $((n_1,n_2);k)=((-n_2,n_1);k)$.

B. Geometric interpretation

Each coefficient $(n;k)$ may be identified with a weighted sum of all paths from (0,0) to (n_1,n_2) in $\mathbb{Z}^2$ in exactly k steps belonging to $\mathcal{D}$ — now regarded as a set of steps. This can be seen in the following way. If one expands the product $\mathcal{H}_c{}^k$ each summand may be mapped on a path of length k starting at (0,0) by identifying $w(m)$ with the step $m\in\mathcal{D}$. This is obviously a one to one mapping between the summands of $\mathcal{H}_c{}^k$ and the set $\mathcal{D}_k$={paths of length k starting at (0,0) with steps in $\mathcal{D}$ only}.

Further one can extend the map from $\mathcal{D}_k$ to the summands in the following way. Every path of length k with steps only in $\mathcal{D}$ is mapped on a summand of $\mathcal{H}_c$ such that two paths have the same image if and only if they differ by a translation. We index a summand by its (up to the translation) unique path v, i.e., S_v. Let the path v be composable with ω, i.e., the end point of ω agrees with the starting point of v and let us denote the composition by $v\circ\omega$. Then we have $S_vS_\omega=S_{v\circ\omega}$ and $S_v^{-1}=S_{v^{-1}}$. Therefore the map is a groupoid homomorphism with the composition of two paths as groupoid structure in the domain and the usual multiplication in the range. Next we define the standard path ρ_n from (0,0) to (n_1,n_2) as the unique path of length $|n_1|+|n_2|$ changing its direction only at the vertex $(n_1,0)$. Let $\mathcal{D}_k(n)=\{v\in\mathcal{D}_k|v$ and ρ_n^{-1} are composable}. Since $w(n)=\gamma^{-n_1n_2}S_{\rho_n}$, the coefficient $(n;k)$ can be written as

$$(n;k)=\gamma^{n_1n_2}\sum_{v\in\mathcal{D}_k(n)}S_{\rho_n^{-1}}S_v=\gamma^{n_1n_2}\sum_{v\in\mathcal{D}_k(n)}S_{\rho_n^{-1}\circ v}\,.$$

For a path v let $\tilde{c}_v=\Pi_{m\in\mathcal{D}}\,c_m^{m(v)}$, where $m(v)$ is the number of steps of sort m in v. Then we define the weight of a closed path σ by $\tilde{c}_\sigma\gamma^{2A(\sigma)}$, where the oriented area $A(\sigma)$ enclosed by σ is defined by $A(\sigma)=\int_\sigma x\,dy$. The motivation for this definition is, that every path from (0,0) to (n_1,n_2) can be transformed into the standard path ρ_n by interchanging steps, reducing consecutive forward and backward steps and splitting the steps $(\pm1,\pm1)$ into $(\pm1,0)$ and $(0,\pm1)$. Every action gives a factor, such that in the end $S_v=\tilde{c}_{\rho_n^{-1}\circ v}\gamma^{2A(\rho_n^{-1}\circ v)}S_{\rho_n}$ holds, i.e.,

Proposition 2.3: $(n;k)$ is equal to the weighted sum over all paths v in $\mathcal{D}_k(n)$ as follows

$$(n;k)=\tilde{c}_{\rho_n}\gamma^{n_1n_2}\sum_{v\in\mathcal{D}_k(n)}S_{v_n}S_{\rho_n^{-1}}=\gamma^{\,n_1n_2}\sum_{v\in\mathcal{D}_k(n)}\tilde{c}_v\,\gamma^{2A(\rho_n^{-1}\circ v)}.$$

Proof: For $k=0$ the agreement is obvious. Further one verifies that the weighted sum of paths defined above fulfill the same recurrence relation as $(n;k)$, see (8). Therefore, by induction over k the proof is complete. $\square$

C. An analytic expression for $(n;k)$

The recurrence relation (8) involves two degrees of freedom and has complex coefficients. It seems that there exists no canonical way in the literature to solve it. We make some kind of discrete Fourier transform in such a way that the new coefficients obey a recurrence relation in which one (out of two) degree of freedom is constant. This leads to a closed formula for the expansion coefficients.

Let $\mu\in\mathcal{D}$. We define the (partial) ''Fourier transform'' in direction μ by

$$[d;k]_{(\omega,\mu)}=\sum_{\langle\mu,n\rangle=d}\gamma^{\sigma(\mu,n)(\omega+d)/\langle\mu,\mu\rangle}(n;k), \tag{10}$$

where σ is the symplectic form of (4) and the sum runs over n. If γ is a qth root of unity, then the inverse of this transformation is given by

$$(n;k)=\frac{1}{2q}\int_0^{2q}\gamma^{-\sigma(\mu,n)(\omega+\langle\mu,n\rangle)/\langle\mu,\mu\rangle}[\langle\mu,n\rangle;k]_{(\omega,\mu)}d\omega \tag{11}$$

and by continuous prolongation otherwise. Actually, for a given k, the integral can be replaced by a finite sum. The ''Fourier-transformed'' coefficients fulfill a recurrence relation induced by (8). Fix $\omega\in\mathbb{Z}$, $\mu\in\mathcal{D}$, then we get

$$\begin{aligned}[d;k+1]&=\sum_{\langle\mu,n\rangle=d}\gamma^{\sigma(\mu,n)(\omega+d)/\langle\mu,\mu\rangle}\sum_{m\in\mathcal{D}}c_m\gamma^{-\sigma(m,n-m)}(n-m;k)\\
&=\sum_{\substack{m\in\mathcal{D}\\ \langle\mu,n\rangle=d}}\sum c_m\gamma^{[\sigma(\mu,n)(\omega+d)-\langle\mu,m\rangle\sigma(\mu,n-m)+\sigma(\mu,m)(d-\langle\mu,m\rangle)]/\langle\mu,\mu\rangle}(n-m;k)\\
&=\sum_{\substack{m\in\mathcal{D}\\ \langle\mu,n\rangle=d}}\sum c_m\gamma^{\sigma(\mu,m)[\omega+2d-\langle\mu,m\rangle]/\langle\mu,\mu\rangle+\sigma(\mu,n-m)[\omega+d-\langle\mu,m\rangle]/\langle\mu,\mu\rangle}(n-m;k)\\
&=\sum_{m\in\mathcal{D}}c_m\gamma^{\sigma(\mu,m)[\omega+2d-\langle\mu,m\rangle]/\langle\mu,\mu\rangle}[d-\langle\mu,m\rangle;k],\end{aligned} \tag{12}$$

where we first used the recurrence relation (8) and second the identity $\langle\mu,\mu\rangle m=\langle\mu,m\rangle\mu+\langle\mu_s,m\rangle\mu_s$ with $\mu_s=(\mu_2,-\mu_1)$ and therefore

$$\sigma(\langle\mu,\mu\rangle m,n-m)=\langle\mu,m\rangle\sigma(\mu,n-m)-\sigma(\mu,m)(d-\langle\mu,m\rangle).$$

Obviously, ω is constant in the recurrence relation (12). The initial condition transforms to $[d;0]=\delta_d$, hence it is independent of the ''frequency'' ω. Notice that $\langle\mu,m\rangle$ have values only in $\{0,\pm1,\pm2\}$. Therefore the recurrence relation has at most five terms and is solved by using the path picture. We will do this for $\mu=(1,0)$. Let $P_d^k:=\{v\in\{-1,0,1\}^k|\Sigma_l v_l=d\}$, then

$$[d;k]_{(\omega,\mu)}=\sum_{v\in P_d^k}\prod_{l=1}^{k}a_v(l)=\langle e_d,A^ke_0\rangle$$

with $a_v(l):=\tilde{a}(t,v_l):=c_{(v_l,1)}\gamma^{\omega+2t-v_l}+c_{(v_l,-1)}\gamma^{-(\omega+2t-v_l)}+c_{(v_l,0)}$ and $t=\Sigma_{s=1}^{l}v_s$, where all $c_m=0$ for $m\notin\mathcal{D}$. In the last expression A is the tridiagonal matrix $A_{ts}=\tilde{a}(t,t-s)$ and e_d denote the canonical basis vectors in $\ell^2(\mathbb{Z})$. The inverse Fourier transform (11) leads to an involved but analytic expression for $(n;k)$, see also the Appendix.

III. ON SPECTRAL PROPERTIES OF $\mathcal{H}_c$

The spectral properties of $\mathcal{H}_c$ depend strongly on θ (i.e., on the magnetic flux). For rational θ the spectrum is the union of a finite number of closed intervals. Whether these intervals are separated by a gap for arbitrary rational θ, is the subject of Sec. III A. We reformulate this question in algebraic words and give an answer in terms of the generalized binomial coefficients. Then we show that the quantitative estimate of the gap widths in the rational case implies the Cantor set property for some irrational values of θ.

A. Rational magnetic flux

Let us recall some elementary properties of the (C^*-) algebra $A(\theta)$, defined in Sec. I. Let $\mathbb{T}$ denote the unit circle in the complex plane. There exists a canonical action $z\rightarrow\phi_z$ of $\mathbb{T}^2$ on $A(\theta)$, such that $\phi_z(u)=z_1 u$ and $\phi_z(v)=z_2 v$. Any element of $A(\theta)$ fixed by ϕ is a scalar multiple of the identity id. There is a unique tracial state τ of $A(\theta)$ invariant under ϕ. This yields to a noncommutative differential structure and by analogy $A(\theta)$ is called the noncommutative torus, for further properties see Ref. 17.

If θ is rational, $\theta=p/q$ with p and q relative prime integers, then there is an irreducible representation Π of $A(p/q)$ on $\mathbb{C}^q$ such that Πv is the cyclic shift and $\Pi u=\mathrm{diag}(1,\gamma^2,\ldots,\gamma^{2q-2})$. It is not difficult to see, that every irreducible representation of $A(p/q)$ is unitarily equivalent to $\Pi_z=\Pi\phi_z$ for some $z\in\mathbb{T}^2$, and that two such representations Π_z and $\Pi_{\tilde{z}}$ are unitarily equivalent if, and only if, $\tilde{z}_i=\gamma^{2n_i}z_i$ for some $n\in\mathbf{I}_q=\{0,1,...,q-1\}^2$. [Since u^q and v^q are central in $A(p/q)$ if $\gamma^{2q}=1$, the image of $A(p/q)$ in any irreducible representation is linearly spanned by q^2 monomials $w(n)$, $n\in\mathbf{I}_q$ defined in Sec. II. Hence the representation is at most of dimension q. Because of the commutation relation (3) $vuv^{-1}=\gamma^2u$ we have the property, that if $z_1\in\mathbb{T}$ is an eigenvalue of u then so is γ^2z_1. Hence there must be a basis in which the image of u is $z_1\Pi u$, and then, after changing the basis element by a phase factor, the image of v is $z_2\Pi v$, as desired.]

The roles of u and v are completely symmetric, the corresponding bases are related by discrete Fourier transform.[13] Let $w(n)$ be defined by (4) in Sec. I.

Lemma 3.1: $\{\Pi_z w(n)\}_{n\in\mathbf{I}_q}$ *form a basis of* $\Pi_z A(p/q)\cong M_q(\mathbb{C})$.

Proof: Without loss of generality we may assume $z_1=z_2=1$. Since the representation $\Pi A(p/q)$ has dimension q, we have to prove the linear independence of $\{\Pi w(n)\}_{n\in\mathbf{I}_q}$ only.

Let $a_n\in\mathbb{C}$ such that $\Sigma_{n\in\mathbf{I}_q}a_n\Pi w(n)=0$ holds, then all its matrix elements $(d,d+s(\mathrm{mod}\ q))$ are zero $(0\leqslant d,s\leqslant q-1)$, i.e.,

$$0=\sum_{n\in\mathbf{I}_q}a_n\gamma^{-n_1n_2+2n_1d}\delta(s-n_2)=\sum_{n_1=0}^{q-1}a_{(n_1,s)}\gamma^{n_1(2d+s)}.$$

Hence we are left with

$$\begin{aligned}0&=\sum_{d=0}^{q-1}\gamma^{2dt}\sum_{n_1=0}^{q-1}a_{(n_1,s)}\gamma^{n_1(2d+s)}\\&=\sum_{n_1=0}^{q-1}a_{(n_1,s)}\gamma^{n_1s}\sum_{d=0}^{q-1}\gamma^{2d(n_1-t)}\\&=\sum_{n_1=0}^{q-1}a_{(n_1,s)}\gamma^{n_1s}q\,\delta(n_1-t)=a_{(t,s)}\gamma^{ts}q\end{aligned}$$

and therefore all a_n are zero. □

The spectrum of $\mathcal{H}_c$ [in $A(\theta)$] is of course the union of the images of $\mathcal{H}_c$ in a complete set of irreducible representations of $A(\theta)$. With the representations for rational θ above we have

$$\mathrm{Spec}(\mathcal{H}_c)=\bigcup_{z\in\mathbb{T}^2}\mathrm{Spec}(\Pi_z\mathcal{H}_c)$$

and for finite dimension, the spectra are determined by the characteristic polynomials. For $\mathcal{H}_c$ defined in (6) these polynomials behave nicely as a function of the irreducible representations. In the case of the square lattice Hamiltonian this was first observed in Ref. 18, and afterwards in several other cases similar behavior has also been observed (see Refs. 8 and 13).

Theorem 3.2: *Let $\theta=p/q$, p and q being relative prime integers. Then the characteristic polynomial* ch_z *of* $\Pi_z\mathcal{H}_c$ *as a function of* $z\in\mathbb{T}^2:z\to\mathrm{ch}_z(\cdot)$ *fulfills the Chambers relation*[18]

$$\mathrm{ch}_z(x)=f(x)-h(z), \tag{13}$$

where $f(x)=a_0+a_1x+\ldots+a_qx^q$, $h(z)=\Sigma_{m\in\mathcal{D}}b_mz^{qm}$ *and* $z^{qm}=z_1^{qm_1}z_2^{qm_2}$. *The coefficients are given by*

$$b_m=(-1)^{pq}(qm;q)=(-1)^{pq}c_m^q \quad \text{for } m\in\{-1,1\}^2,$$

$$b_m=(qm;q)=2\varepsilon^qT_q(c_m/2\varepsilon) \quad \text{for } m\in\{(\pm1,0),(0,\pm1)\}$$

with $\varepsilon=\sqrt{c_{(m_1+m_2,m_2+m_1)}c_{(m_1-m_2,m_2-m_1)}}$ *and T_q is the Chebyshev polynomial of degree q. Further the a_k fulfill*

$$\sum_{k=0}^{q}a_k(n;k)_{\mathcal{H}_c}=0 \quad \text{for } \|n\|_\infty<q. \tag{14}$$

Note that b_m is also well defined at $\varepsilon=0$ by continuous prolongation.

Proof: Using the properties of the rotation algebra $A(\theta)$, the characteristic polynomial can be written

$$\mathrm{ch}_z(x)=\det(x\ \mathrm{id}-\Pi_z\mathcal{H}_c)=\det\left(x\ \mathrm{id}-\sum_{m\in\mathcal{D}}c_mz^m\Pi w(m)\right).$$

First observe, because of $\mathrm{ch}_z(x)=\mathrm{ch}_{z\gamma^{2n}}(x)$, where $z\gamma^{2n}=(z_1\gamma^{2n_1},z_2\gamma^{2n_2})$, the characteristic polynomial has only powers of the form z^{qn}, $n\in\mathbb{Z}^2$. Since the determinant is homogeneous of degree q and any entry is of the form $t_0x+\Sigma_{m\in\mathcal{D}}\,t_mz^m$, $\mathrm{ch}_z(x)$ is a Laurent polynomial in z_1, z_2 and x with $\max\{\deg_{z_1},\deg_{z_2}\}+\deg_x\leq q$, where $\deg_*$ is the degree with respect to the variable. Hence the dependence on x and z splits

$$\mathrm{ch}_z(x)=f(x)-h(z)=\sum_{k=0}^{q}a_kx^k-\sum_{m\in\mathcal{D}}b_mz^{qm}.$$

Now we calculate the coefficients. The map $z\to\mathrm{ch}_z(\Pi_z\mathcal{H}_c)$ extends uniquely to an analytic function $\mathbb{C}^2\backslash\{0\}\to M_q(\mathbb{C}):z\to\Sigma_{n\in\mathbb{Z}^2}\,d_nz^n$, which is zero for $|z_1|=|z_2|=1$ and hence everywhere. In other words $d_n=0$, for every $n\in\mathbb{Z}^2$. Since

$$\begin{aligned}\mathrm{ch}_z(\Pi_z\mathcal{H}_c)&=\sum_{k=0}^{q}a_k\Pi_z\mathcal{H}_c^k-h(z)\Pi_z\mathrm{id}\\&=\sum_{k=0}^{q}a_k\sum_{n\in\mathbb{Z}^2}(n;k)\Pi_zw(n)-\sum_{m\in\mathcal{D}}b_mz^{qm}\Pi_z\mathrm{id}\\&=\sum_{n\in\mathbb{Z}^2}\sum_{k=0}^{q}a_k(n;k)z^n\Pi w(n)-\sum_{m\in\mathcal{D}}b_mz^{qm}\Pi\mathrm{id},\end{aligned}$$

the vanishing of the d_n's imply for $m\in\mathcal{D}$

$$0=\sum_{k=0}^{q} a_k(qm;k)\Pi w(qm)-b_m\Pi \mathrm{id}=(-1)^{pqm_1m_2}(qm;q)\Pi \mathrm{id}-b_m\Pi \mathrm{id},$$

and $0=\Sigma_{k=0}^{q}a_k(n;k)\ \Pi w(n)$, for $\|n\|_\infty<q$. Remember $\Pi w(qm)=(-1)^{pqm_1m_2}\Pi \mathrm{id}$, for $m\in\mathcal{D}$ and $a_q=1$ by the very definition of $\mathrm{ch}_z(x)$ and $(n;k)=0$ for $\|n\|_\infty>k$.

Since all $\Pi w(n)\neq 0$, we obtain (14). The value of $b_m=(-1)^{pqm_1m_2}(qm;k)$ is calculated in the Appendix. □

Remark: By the above argument, the coefficients $(n;q)$ vanish for $n\in\mathbb{Z}^2\{qm|m\in D\}$ with $\|n\|_\infty=q$ as already calculated in the Appendix.

For the doubly discrete quantum pendulum the above expression has been derived in Ref. 13 employing a different method. The Chambers relation (13) is crucial to prove the existence and estimate the size of the spectral gaps, since it reduces the problem to a finite-dimensional eigenvalue problem.

Theorem 3.3: *The spectrum of $\mathcal{H}_c\in A(p/q)$ consists of q disjoint bands if and only if*

$$\inf_{z\in\mathbb{T}^2}\{|a-b|\,\big|\,a\neq b\in \mathrm{Spec}(\Pi_z\mathcal{H}_c)\}>0. \tag{15}$$

Further this is a lower bound for the gap widths.

Proof: By the Chambers relation (13) we know that the spectrum of $\mathcal{H}_c$ is the preimage of the interval $[h_{\min},h_{\max}]=\{h(z)|z\in\mathbb{T}^2\}$ under the polynomial $f(x)$ of degree q, i.e., it is the union of q bands. They are disjoint if and only if the spectrum of $\Pi_z\mathcal{H}_c$ is not degenerate for the representation corresponding to $h_{\min}$ and $h_{\max}$. By continuity of $z\to\Pi_z$ these representations are not degenerate if and only if (15) holds. Obviously $\inf_{z\in\mathbb{T}^2}\{|a-b|\,\big|\,a\neq b\in \mathrm{Spec}(\Pi_z\mathcal{H}_c)\}$ is also an optimal lower bound for the gap widths. □

This also proves the following Corollary.

Corollary 3.4: The spectrum of $\mathcal{H}_c\in A(p/q)$ can only be degenerate in representations Π_z corresponding to extreme values of $h(z)$ and if it is not degenerate, then the infimum (15) is reached for every $z'\in\mathbb{T}^2$ with $h(z')=\max h(z)$ or for every $z''\in\mathbb{T}^2$ with $h(z'')=\min h(z)$.

For the almost Mathieu operator, Choi *et al.*[10] used an elegant method to obtain an estimate for (15). We employ and generalize it.

Lemma 3.5: Let B be a self-adjoint operator in a complex Hilbert space of dimension q. Then every eigenvalue of B is simple if and only if $\|g(B)\|\neq 0$ for every monic polynomial g of degree $q-1$. A lower bound for the difference of two (different) eigenvalues of B is given by

$$|a-b|\geq\frac{\inf\{\|g(B)\|\}}{(2\|B\|)^{q-2}},$$

where the infimum is taken over the set of monic polynomials g of degree $q-1$.

Proof: A monic polynomial of degree $q-1$ can be written as

$$g(x)=\prod_{j=1}^{q-1}(x-\varrho_j),$$

where $\varrho_j\in\mathbb{C}$ are the roots of g. Denote by $\kappa_1\leq\kappa_2\leq\cdots\leq\kappa_q$ the eigenvalues of B with each eigenvalue counted according to its multiplicity. As a consequence of the spectral theorem, the spectrum of $g(B)$ is given by the $1\leq k\leq q$ values $\Pi_{j=1}^{q-1}(\kappa_k-\varrho_j)$. Since $g(B)$ is normal we have

$$\|g(B)\|=\max_{1\leq k\leq q}\left\{\prod_{j=1}^{q-1}|\kappa_k-\varrho_j|\right\}$$

and therefore $\|g(B)\|=0$ if and only if $\{\kappa_1,\kappa_2,\ldots,\kappa_q\}\subset\{\varrho_1,\varrho_2,\ldots,\varrho_{q-1}\}$. Hence, if at least one eigenvalue of B is not simple then there exists one monic polynomial g of degree $q-1$ with $g(B)=0$. On the other hand, if all eigenvalues of B are simple, obviously $\{\kappa_1,\kappa_2,\ldots,\kappa_q\}\subset\{\varrho_1,\varrho_2,\ldots,\varrho_{q-1}\}$ can never happen.

Let $N\neq M\in\{1,...,q\}$. With $|\kappa_i-\kappa_j|\leqslant 2\|B\|$ we get for the monic polynomial $g_N(x)=\Pi_{j=1,j\neq N}^{q}(x-\kappa_j)$

$$\begin{aligned}\|g_N(B)\| &= \max_{1\leqslant k\leqslant q}\left\{\prod_{j=1,j\neq N}^{q}|\kappa_k-\kappa_j|\right\}\\ &=\prod_{j=1,j\neq N}^{q}|\kappa_N-\kappa_j|\\ &=|\kappa_N-\kappa_M|\prod_{\substack{j=1\\ j\neq N,M}}^{q}|\kappa_N-\kappa_j|\leqslant|\kappa_N-\kappa_M|(2\|B\|)^{q-2}\end{aligned}$$

by applying the spectral theorem and using that $g_N(B)$ is normal. Therefore $\inf\{\|g(B)\|\}$ over all monic polynomials of degree $q-1$ fulfill the inequality of the lemma.

Now we are left with the following situation. Let g be an arbitrary monic polynomial $g(x)=a_0+a_1x+\cdots+a_{q-1}x^{q-1}$ of degree $q-1$. Then a lower bound of $\|g(\Pi_z\mathcal{H}_c)\|$ leads to a lower bound for the spectral gaps of $\mathcal{H}_c$. By Corollary 3.4 we need only to consider representations Π_z, where $h(z)$ has a global extremum.

First we give an argument for $g(\Pi_z\mathcal{H}_c)\neq 0$. Suppose that there exists such a g with $g(\Pi_z\mathcal{H}_c)\equiv 0$. This means that the system of linear equations, arising from the coefficients of the basis $(w(n))_{n\in\mathbf{I}_q}$

$$\begin{aligned}0&=\sum_{m\in\mathbb{Z}^2}\sum_{k=0}^{q-1}z^{qm}\gamma^{q\sigma(n,m)+q^2m_1m_2}(n-qm;\kappa)a_k\\ &=\sum_{m\in\mathcal{D}_+}\sum_{k=0}^{q-1}z^{qm}(-1)^{p\sigma(n,m)+pqm_1m_2}(n-qm;\kappa)a_k,\end{aligned}\tag{16}$$

with $z^{qm}=z_1^{qm_1}z_2^{qm_2}$ and $\mathcal{D}_+=\{0,1\}^2$ has a nontrivial solution $(a_0,..,a_{q-1})$.[19] Whenever a nontrivial solution exists, there is at least one with $a_{q-1}=1$ [since $g(x)|_{x=\Pi_z\mathcal{H}_c}=0\Rightarrow g(x)(x-b)|_{x=\Pi_z\mathcal{H}_c}=0$].

Hence $g(\Pi_z\mathcal{H}_c)\neq 0$ if we find a linear combination (α_l) of the equations $(16)_n$ such that

$$\left|\sum_{l=1}^{s}\alpha_l\sum_{m\in\mathcal{D}_+}\sum_{k=0}^{q-1}z^{qm}(-1)^{p\sigma(n,m)-pqm_1m_2}(n_l-qm;k)a_k\right|\geqslant 1\tag{17}$$

holds. Moreover, such a linear combination already gives a lower bound for $\|g(\Pi_z\mathcal{H}_c)\|$.

Lemma 3.6: Let $\theta=p/q$ with p and q relative prime integers and $z\in\mathbb{T}^2$. Suppose there exists a linear combination $\alpha_l(z)$, such that (17) holds for any arbitrary monic polynomial g of degree $q-1$, then the eigenvalues of $\Pi_z\mathcal{H}_c$ are not degenerate and a lower bound for the distance between two of them is given by

$$(2^{q-2}s\|\Pi_z\mathcal{H}_c\|^{q-2}\|\alpha(z)\|_\infty)^{-1},$$

where s is the number of nonzero coefficients $\alpha_l(z)$ and $\|\alpha(z)\|_\infty$ its supremum.

Proof: We calculate a lower bound for $\|g(\Pi_z\mathcal{H}_c)\|$ and then apply Lemma 3.5. For a given linear combination (α_l) of Eq. (16) and fixed z we have

$$\left\| \sum_{l=1}^{s} \alpha_l w(-n_l) \right\| \leqslant \sum_{l=1}^{s} |\alpha_l| \|w(-n_l)\| \leqslant s \|\alpha\|_\infty .$$

Since $\|\Pi_z w(n)\| \geqslant \sqrt{\mathbf{Tr}}(\Pi_z w(n)) = \delta_n q^{-1}$ for every $n \in \mathbf{I}_q$, we have with $\tau := p\sigma(n,m) - pqm_1 m_2$,

$$\begin{aligned}
\|g(\Pi_z \mathcal{H}_c)\| \|\alpha\|_\infty s &\geqslant \|g(\Pi_z \mathcal{H}_c)\| \left\| \sum_{l=1}^{s} \alpha_l \Pi_z w(-n_l) \right\| \\
&\geqslant \left\| \sum_{n \in \mathbf{I}_q} \sum_{m \in D_+} \sum_{k=0}^{q-1} z^{qm}(-1)^\tau a_k(n-qm;k) \Pi_z w(n) \sum_{l=1}^{s} \alpha_l \Pi_z w(-n_l) \right\| \\
&\geqslant q^{-1} \left| \mathbf{Tr} \left(\sum_{k=0}^{q-1} \sum_{l=1}^{s} \alpha_l \sum_{m \in \mathcal{D}_+} z^{qm}(-1)^\tau a_k(n_l-qm;k) \Pi_z w(0) \right) \right| \\
&\geqslant \left| \sum_{k=0}^{q-1} \sum_{l=1}^{s} \alpha_l \sum_{m \in \mathcal{D}_+} z^{qm}(-1)^\tau a_k(n_l-qm;k) \right| .
\end{aligned}$$

Under the assumption of the Lemma the last term is greater than 1 and therefore we obtain an estimate for $\|g(\Pi_z \mathcal{H}_c)\|$. By Lemma 3.5 the spectrum of $\Pi_z \mathcal{H}_c$ is not degenerate and two eigenvalues are separated at least by □

$$(2^{q-2} s \|\Pi_z \mathcal{H}_c\|^{q-2} \|\alpha(z)\|_\infty)^{-1}.$$

Because of the Chambers relation (13) the spectrum of $\mathcal{H}_c$ is the union of the q bands $[\lambda_1,\mu_1]$, $[\mu_2,\lambda_2]$, $[\lambda_3,\mu_3]$,… where $\lambda_1 \leqslant \mu_1 \leqslant \mu_2 \leqslant \lambda_2 \leqslant \dots$ is the spectrum of $\Pi_z \mathcal{H}_c$ with $h(z)$ maximal and with $h(z)$ minimal, respectively. Thus the spectral gaps can be divided into two classes, one corresponding to the maximum and the other to the minimum of $h(z)$. An estimate for the gap widths in such a class can be calculated in the corresponding representation. Therefore by Lemma 3.6 the following Theorem holds.

Theorem 3.7: *Let $\theta = p/q$ with p and q relative prime integers. Suppose for $z \in \mathbb{T}^2$ with $h(z) = \min_{z'} h(z')$ or $h(z) = \max_{z'} h(z')$ there exists a linear combination α_l, such that (17) holds for z and any arbitrary monic polynomial g of degree $q-1$, then the spectral gaps of the corresponding class are open and are at least of size*

$$(2^{q-2} s \|\Pi_z \mathcal{H}_c\|^{q-2} \|\alpha(z)\|_\infty)^{-1}, \tag{18}$$

where s is the number of nonzero coefficients α_l and $\|\alpha\|_\infty$ its supremum.

At first sight it seems difficult to find such a linear combination $\alpha_l(z)$, since the a_n are arbitrary except for $a_{q\text{-}1} = 1$ and the ''behavior'' of coefficients $(n;k)$ is complex. Moreover, this has to be done for every ''periodic length'' $q \in \mathbb{N}$. A solution is provided in many cases as follows.

Theorem 3.8: *Let $\theta = p/q$ with p and q relative prime integers and $\mathcal{H}_c$ with $c_\mu \in \mathbb{R}$. Then for every $z \in \mathbb{T}^2$ with $z_m^{2q} \neq 1$, $(qm-m;q-1)_{\mathcal{H}_c} \neq 0$ and $m = (1,0)$ or $(0,1)$, there exists a linear combination $\alpha_l(z)$ of two equations such that (17) holds and*

$$\|\alpha(z)\|_\infty = |(1 - z_m^{2q})(qm-m;q-1)_{\mathcal{H}_c}|^{-1}.$$

Proof: First let $m = (1,0)$ and $z_1^{2q} \neq 1$. Because of Proposition 2.1 we have $(n;k) = (-n;k)$. Therefore Eq. (16) with $n = m$ and $n = (q-1)m$, respectively, leads to

$$0 = \sum_{k=0}^{q-1} (m;k) a_k + z_1^q (qm-m;q-1) a_{q-1},$$

$$0 = \sum_{k=0}^{q-1} z_1^q(m;k)a_k + (qm-m;q-1)a_{q-1}.$$

Taking $\alpha_{(q-1,0)} = 1/(1-z_1^{2q})(qm-m;q-1)$ and $\alpha_{(1,0)} = -z_1^q\alpha_{(q-1,0)}$, we obtain such a linear combination.

If $z_2^{2q} \neq 1$ we repeat the above argument for $m=(0,1)$. □

The coefficients $(qm-m;q-1)$ are given by a straightforward calculation. For $m=(1,0)$ the partial Fourier transformed coefficient $[q-1;q-1]_{(m,\omega)}$ is given by $[q;q]_{(m,\omega)}/(c_{(10)} + c_{(11)}\gamma^{\omega-1} + c_{(1-1)}\gamma^{-\omega+1})$, where $[q;q]_{(m,\omega)}$ has been calculated in the Appendix. The inverse Fourier transformation leads then to $(qm-m;q-1)$, thereby several cases depending on c have to be distinguished.

B. Quantitative continuity of the spectrum

The spectral gap boundaries of a normal element in $A(\theta)$ are continuous in the Hausdorff metric as a function of the magnetic field.[20] For a large class of them even more qualitative properties have been proved. It was first shown in Ref. 10 that the spectral gap boundaries of the almost Mathieu operator is Hölder continuous with exponent 1/3. This result was extended in Ref. 21 to a larger class and improved to the Hölder exponent 1/2, which is optimal in view of the semiclassical analysis as pointed out in Ref. 22. Therein Lipshitz continuity of the spectrum was proven for elements in certain ''Sobolev classes'' of $A(\theta)$. The Lipshitz constant depends on the width of the corresponding spectral gap and diverges near values of θ for which the gap closes. For the almost Mathieu operator more about the continuity and measure of its spectrum is known, see, e.g., Ref. 5 and references therein. Following Ref. 10, we use the quantitative continuity to conclude from spectral gaps for rational values of θ to a possibly large class of irrational values.

The result in Ref. 21 does not cover all Hamiltonians $\mathcal{H}_c$, though the extension to it is possible.[23] We will state the result only. Let $f:\mathbb{R}\to\mathbb{R}$ and $h:\mathbb{R}\to\mathbb{C}$ continuous differentiable 2π-periodic functions. For θ, $\beta\in\mathbb{R}$ and $h_n = h(2\pi\theta n-\beta)$, $f_n = f(2\pi\theta n-\beta)$ define a Jacobi matrix by

$$(H_{f,h}^{\theta,\beta}\phi)(n) = \overline{h_{n+1}}\phi(n+1) + f_n\phi(n) + h_n\phi(n-1)$$

and denote the union of spectra $\cup_{\beta\in\mathbb{R}}\mathrm{Spec}(H_{f,h}^{\theta,\beta})$ by $\sigma_{fh}(\theta)$.

Theorem 3.9: *Let f, h as above and $M_{fh} = 2\pi(2\|h'\|_\infty + \|f'\|_\infty)$. Then the Hausdorff distance between $\sigma_{fh}(\theta)$ and $\sigma_{fh}(\theta')$, with $\theta,\theta'\in\mathbb{R}$ is bounded by*

$$\mathrm{Dist}(\sigma_{fh}(\theta),\sigma_{fh}(\theta')) \leq 2\sqrt{5\|h\|_\infty M_{fh}|\theta-\theta'|} + M_{fh}|\theta-\theta'|.$$

For suitable f and h, the spectrum of $\mathcal{H}_c$ is given by $\sigma_{fh}(\theta)$ as follows. Let $\beta\in\mathbb{R}$ and Π_β be the representation of $A(\theta)$ in $l^2(\mathbb{Z})$ taking u into the twosided shift and v into the operator of multiplication by the function $s\to\exp 2\pi i(s\theta+\beta)$. The direct integral of these representations is faithful, and so

$$\mathrm{Spec}\ \mathcal{H}_c(\theta) = \bigcup_{\beta\in\mathbb{R}} \mathrm{Spec}\ \Pi_\beta\mathcal{H}_c(\theta).$$

For convenience write h_θ^β for $\Pi_\beta\mathcal{H}_c(\theta)$. [Since $A(\theta)$ is simple if θ is irrational, for any β, Π_β is faithful and $\mathrm{Spec}\ \mathcal{H}_c = \mathrm{Spec}\ h_\theta^\beta$, in that case. In any case, the set of representations $\{(\mathrm{Ad}\ M_y)\Pi_\beta; y\in\mathbb{T},\beta\in\mathbb{R}\}$, where M_y denotes the multiplication operator $s\to y^s$, is invariant under the canonical action ϕ of $\mathbb{T}^2$ on $A(\theta)$, and so the proper ideal $\cap_\beta \ker\Pi_\beta$ of $A(\theta)$ is invariant under this action. If this ideal were nonzero, it would contain a positive element fixed under Φ (since $\mathbb{T}^2$ is compact), which would be a nonzero scalar, and the ideal would be all of $A(\theta)$. This shows $\cap_\beta\ker\Pi_\beta = 0$ as desired.]

Corollary 3.10: For $\theta,\theta'\in\mathbb{R}$,

$$\mathrm{Dist}(\mathrm{Spec}\ \mathcal{H}_c(\theta),\mathrm{Spec}\ \mathcal{H}_c(\theta'))\leqslant 4\tilde{c}\sqrt{5\pi|\theta-\theta'|}+4\pi\tilde{c}|\theta-\theta'|,$$

with $\tilde{c}=|c_{(1,1)}|+|c_{(1,-1)}|+\max(|c_{(1,0)}|,|c_{(0,1)}|)$.

Proof: Obviously h_θ^β is a Jacobi matrix of the form $H_{f,h}^{\theta,\beta}$ with estimates $\|h\|_\infty\leqslant|c_{(1,1)}|+|c_{(1,-1)}|+|c_{(1,0)}|$ and $M_{fh}\leqslant 4\pi(|c_{(1,1)}|+|c_{(1,-1)}|+|c_{(0,1)}|)$. Substitution of these values into the inequality of Theorem 3.9 completes the proof. □

C. On spectral gaps for certain Liouville numbers

In this section we prove the existence of the spectral gaps for irrational θ under the condition that this is true for all rational numbers in a neighborhood of θ. This argument works only for irrational numbers, which are sufficiently well approximated by the rationales, i.e., some Liouville numbers.

Theorem 3.11: *Let I be an open interval, such that for every* $p/q\in I$ *all* $q-1$ *gaps in the spectrum of* $\mathcal{H}_c\in A(p/q)$ *are at least of size* $s_I(q)\in\mathbb{R}_+$ *with p and q relative prime integers. If for* $\theta\in I$ *an arbitrarily large q exists with* $|\theta-p/q|\leqslant\{3^4\sqrt{5\pi}\ \tilde{c}^{\ 2}\}^{-1}s_I(q)^2$, *then all gaps of the spectrum of* $\mathcal{H}_c\in A(\theta)$ *are open.*

Recall, $\tilde{c}=|c_{(1,1)}|+|c_{(1,-1)}|+\max(|c_{(1,0)}|,|c_{(0,1)}|)$.

Proof: Fix $\varepsilon>0$. Let be $p/q\in I$, such that $|\theta-p/q|\leqslant\min\{[s_\theta(q)^2/3^4 5\pi\tilde{c}^2];5\varepsilon^2/\pi\}$, then by Theorem 3.9 we have

$$\begin{aligned}\mathrm{Dist}(\mathrm{Spec}\ \mathcal{H}_c(\theta),\mathrm{Spec}\ \mathcal{H}_c(p/q))&\leqslant 4\tilde{c}\sqrt{5\pi|\theta-\theta'|}+4\pi\tilde{c}|\theta-\theta'|\\&\leqslant 4\tilde{c}\sqrt{5\pi}(1+\varepsilon)\sqrt{|\theta-\theta'|}\\&\leqslant 4/9(1+\varepsilon)s_\theta(q).\end{aligned}$$

Furthermore, $\mathrm{Dist}(\mathrm{Spec}\ \mathcal{H}_c(\theta),\mathrm{Spec}\ \mathcal{H}_c(\theta'))\leqslant 4/9(1+\varepsilon)s_\theta(q)$ for any θ' between θ and p/q. Therefore, for $|\theta-p/q|<\min\{[s_\theta(q)^2/3^4 5\pi\tilde{c}^2];5\varepsilon^2/\pi\}$ some part of a gap from $\mathcal{H}_c\in A(p/q)$ is contained in a gap from $\mathcal{H}_c\in A(\theta')$ and hence open. □

For those values θ we obtain a topological description for Spec $\mathcal{H}_c$.

Corollary 3.12: Suppose the conditions of Theorem 3.11 hold, then the spectrum of $\mathcal{H}_c(\theta)$ *is a Cantor set for those irrational numbers* θ.

Proof: By Theorem 3.11 we have to prove that the gaps are dense in Spec $\mathcal{H}_c(\theta)$. This would follow from knowing that for $\theta=p/q\in I$ with $(p,q)=1$, the gaps in Spec $\mathcal{H}_c(\theta)$ are at most a distance $12\pi c_s/q$ apart, where we used $c_s=\sup_{m\in\mathcal{D}}|c_m|$. With other words, the length of any interval in the spectrum is at most $12\pi c_s/q$.

By the Weyl spectral variation inequality for Hermitian matrices (see Ref. 24), the distance between corresponding eigenvalues of $\Pi_z\mathcal{H}_c$ and $\Pi_{(1,1)}\mathcal{H}_c$ (numbered in decreasing order) is at most

$$\begin{aligned}\|\Pi_z\mathcal{H}_c-\Pi_{(1,1)}\mathcal{H}_c\|&\leqslant 2c_s(|1-z_1|+|1-z_2|+|1-z_1z_2|+|1-z_1/z_2|)\\&\leqslant 4c_s(|1-\exp(\pi i/q)|+|1-\exp(2\pi i/q)|)\leqslant 12c_s\pi/q,\end{aligned}$$

where we used that every irreducible representation is unitarily equivalent to one representation Π_z with $0\leqslant\arg(z_1)$, $\arg(z_2)\leqslant\pi/q$. Since each interval in Spec $\mathcal{H}_c$ consists of a set of such corresponding eigenvalues, the length of any interval is at most $12c_s\pi/q$. □

IV. MODEL DISCUSSION

In this section we investigate classes of $\mathcal{H}_c$ corresponding to tight-binding models on different lattices. In principle it is possible to calculate, for arbitrary parameters c, the critical point of the representation depending on part $h(z)$ of the Chambers relation (13) for every $\theta=p/q$. However many cases have to be distinguished and therefore no good insight of the behavior is obtained. In the following we make the substitution $z_1 \to \exp(i\alpha/q)$, $z_2 \to \exp(i\beta/q)$ and understand h as a function of α and β.

A. Rectangular lattice

Consider a tight-binding model on a rectangular lattice with next-nearest-neighbor interaction. The symmetry of a rectangular lattice and a homogenous magnetic field yields to $H_{\mathrm{rec}} = Au + Bv + C(\overline{\gamma} uv + \gamma uv^*) + \text{h.c.}$ with $A, B, C \in \mathbb{R}$.

Without second neighbor interaction $(C=0)$, the model reduces to the well-known Harper model (see Sec. I). It has been extensively studied by many authors. Results on the spectral measure, continuity of the gaps and Anderson localization are obtained with various methods.[4,21,22,25,10,20,5,6,26] For ''rational values'' of the magnetic flux the existence and width of the spectral gaps have been proved, using different methods.[25,10,26]

We will derive an analytic expression for the characteristic polynomials (13) in terms of the coefficients $(n;k)$. Because of (14) we get

$$\sum_{k=s}^{q-1} (s,0;k) a_k = -(s,0;q)$$

and therefore

$$\begin{pmatrix} a_0 \\ \vdots \\ a_{q-1} \end{pmatrix} = -((s,0;k))_{sk}^{-1} \begin{pmatrix} (0,0;q) \\ \vdots \\ (q,0;q) \end{pmatrix}.$$

The inverse exists, since it is an upper triangular matrix with diagonal elements of the form $A^s \neq 0$. For lq odd all coefficients a_l vanishes, since $(s,0;k)$ vanish, if sk is odd. This also follows from symmetry considerations. Obviously, we can derive for any $\mathcal{H}_c$ such a formula.

This fact can be used to derive an analytic expression for the density of states of the almost Mathieu operator with $A=B=1$. The integrated density of states is given by

$$N(E) = \int_0^{2\pi} \int_0^{2\pi} d\alpha \; d\beta \; \mathbf{Tr}(\chi_{[-4,E]}(\pi_z \mathcal{H}_c)), \tag{19}$$

where χ_I is the characteristic function of $I \in \mathbb{R}$ and is expressed in terms of the density of states (DOS) $g(\mu)$ at energy μ as

$$N(E) = \int_{-4}^{E} d\mu \; g(\mu).$$

Wannier, Obermair and Ray[27] derived from (19) and the Chambers relation (13) that

$$g(\mu) = \begin{cases} \dfrac{1}{2\pi^2 q} \dfrac{d}{d\mu} f(\mu) K'(\tfrac{1}{4}|f(\mu)|) & \text{for } |f(\mu)| \leq 4 \\ 0 & \text{otherwise,} \end{cases}$$

where $f(\mu)$ is given in the Chambers relation (13) and

$$K'(x) = \mathcal{F}\left(\frac{\pi}{2}, \sqrt{1-x^2}\right) = \int_0^{\pi/2} (1-(1-x^2)\sin^2 y)^{-1/2} dy$$

is the complete elliptic integral of the first kind. Since we have computed the coefficients of $f(\mu)$ analytically, this gives an analytic expression for the density of states.

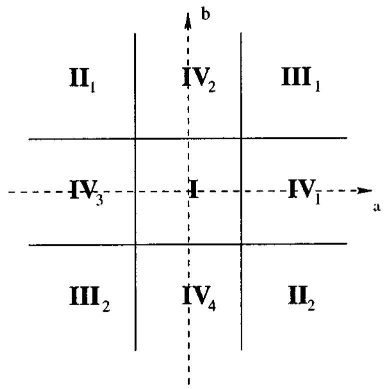

FIG. 1. Phase diagram for the rectangular lattice: $a=4C^qT_q(A/2C)$ and $b=4C^qT_q(B/2C)$ and region I $=[-4C^q,4C^q]^2$.

For nontrivial second neighbor interaction $(-1)^\rho C$ the representation dependent part of the Chambers relation is given by

$$\frac{1}{4C^q}\,h(\alpha,\beta)=T_q\left(\frac{A}{2C}\right)\cos\,\alpha+T_q\left(\frac{B}{2C}\right)\cos\,\beta+(-1)^{(p+\rho)q}\,\cos(\alpha)\cos(\beta).$$

Notice that changing the second neighbor coupling, the phase diagram is only rescaled and therefore we keep C in our discussion constant. The case $(p+\rho)q$ odd can be derived from the even one by variable transformation $\alpha\rightarrow\alpha+\pi,\beta\rightarrow\beta+\pi$. In the following we assume $(p+\rho)q$ even. Let us describe the phase diagram for the critical point of h with respect to A, B and C. First, the phase diagram of the critical points of h in terms of $a=4C^qT_q(A/2C)$ and $b=4C^qT_q(B/2C)$ is given by Fig. 1. Since a and b are the qth Chebyshev polynomials in A and B, respectively, the $A-B$ phase diagram is easily derived from this one.

Consider a path in the $A-B$ phase diagram with A running from $-\infty$ to $+\infty$ and B constant. The corresponding image of this path in the $a-b$ phase diagram lies again on a straight line parallel to the a axis and $b=4C^qT_q(B/2C)$. Recall, the $q-1$ extrema of the Chebyshev polynomials have the value ±1 in the interval $[-1,1]$. For q odd and $|B|<2C$ the path for $A<-2C$ lies in the region IV_3, then $(-2C<A<2C)$ oscillates exactly $q/2$ times in region I. The path touches the border of the neighbor region at the turning points of the oscillation. For $A>2C$ the path lies in IV_1. For $|B|>2C$ we have similar behavior. Therefore, up to this touching, the phase diagram in terms of A and B is a homeomorphic image of Fig. 1. For q even, the path lies on the same straight line and oscillates again exactly $q/2$ times in the corresponding region, but will disappear again to $+\infty$. Hence, for q even, regions $\Pi_{1/2}$, $\mathrm{IV}_{3/4}$ and III_2 do not appear in the $A-B$ phase diagram.

On such a path the width of the range of $h(\alpha,\beta)$ oscillates and because of the Chambers relation, this should be reflected on the total bandwidth of the spectrum. Numerical calculations confirm this behavior, the roots of $T_q(A/2C)$ correspond to the minima of the total bandwidth and the extrema to the maxima, respectively. Recognize, though $h(\alpha,\beta)$ scales with C^q, the spectrum does not.

The critical points are listed in Table I. Consider region I, there are two minima and maxima. The maximum at $(\alpha,\beta)=(0,0)$ is global for $T_q(A/2C)+T_q(B/2C)>0$ and local otherwise. Similarly, the minimum at $(\alpha,\beta)=(\pi,0)$ is global for $T_q(A/2C)-T_q(B/2C)>0$ and local otherwise.

B. Triangular lattice

The (anisotropic) tight-binding Hamiltonian for a particle on the triangular lattice can be expressed as

TABLE I. Critical points for the rectangular lattice model with next-nearest-neighbor interaction and pq even. $e_1 = T_q(A/2C) + T_q(B/2C)$, $e_2 = T_q(A/2C) - T_q(B/2C)$ and sad=saddle point.

($\cos\alpha,\cos\beta$)	Energy in $4C^q$	I	$\mathrm{II}_{1/2}$	$\mathrm{III}_{1/2}$	IV_1	IV_2	IV_3	IV_4
(1,1)	e_1+2	max	sad	max/min	max	max	sad	sad
$(-1,-1)$	$-e_1+2$	max	sad	min/max	sad	sad	max	max
$(-1,1)$	$-e_2-2$	min	max/min	sad	min	sad	sad	min
$(1,-1)$	e_2-2	min	min/max	sad	sad	min	min	sad
$(-\mathbf{b},-\mathbf{a})$	$-\mathbf{ab}$	sad	—	—	—	—	—	—

$$\mathcal{H}_{\text{tri}} = A(u+u^*) + B(v+v^*) + C\gamma(vu+v^*u^*). \tag{20}$$

The model can be generalized by introducing different fluxes in the two classes of triangles.[8] The representation dependent part of the Chambers relation (13) is given by

$$h(\alpha,\beta) = 2A^q \cos\alpha + 2B^q \cos\beta + (-1)^{pq} 2C^q \cos(\alpha+\beta).$$

We describe the critical points of $h(\alpha,\beta)$ only in the generic case $1 = (-1)^{pq} 2C^q \geqslant 2A^q$, $2B^q \geqslant 0$, all other cases can be derived from this one by suitable transformation of α and β and rescaling of $h(\alpha,\beta)$. The critical points are listed in Table II and the phase diagram is shown in Fig. 2. In region II an extra critical point of $h(\alpha,\beta)$ appears, not being multiples of π. The point (1,1) in the phase diagram correspond to the case of hexagonal symmetry, so that the different phases help to distinguish strongly and weakly anisotropic interaction for the triangular lattice model. We call region II the ''weakly anisotropic'' triangular phase. Obviously, for fixed $0 < A,B < 1$ the phase depends on the magnetic flux $\theta = p/q$. The sequence (A^q,B^q) converges to (0,0). For large enough q the phase becomes in any case — except $A = B \leqslant 1$ — strongly anisotropic.

C. Weakly anisotropic triangular phase

In this section, we analyze for the weakly anisotropic triangular phase the gap structure of the spectrum for the corresponding Hamiltonian. By applying the theory of Sec. III we show the existence of gaps for rational and certain irrational values of the magnetic flux per unit cell. For convenience let in the following $C \geqslant A,B > 0$. Since $\mathcal{H}_{\text{tri}}$ has real coefficients in the sense of Proposition 2.1, the generalized coefficients $(n;k)_{\mathcal{H}_{\text{tri}}}$ defined by (7) are real. Let g be an arbitrary monic polynomial of degree $q-1$ with real coefficients.[28] A lower bound for the spectral gaps of $\mathcal{H}_{\text{tri}} \in A(p/q)$ is given by Theorem 3.6, if we find a linear combination α_l, such that (17) holds.

Theorem 4.1: *For every $\theta = p/q$ with p and q relative prime integers and A,B,C in a weakly anisotropic phase the spectral gaps of $\mathcal{H}_{\text{tri}}$ corresponding to the minimum of $h(z)$ if pq even and to the maximum if pq odd, are open.*

TABLE II. Critical points $(\alpha,\beta) \in]-\pi,\pi]^2$ for the triangular lattice model with anisotropic nearest-neighbor interaction and $w_\alpha = \arccos(1/2a^2b^2)(b^3 - b^3a^2 - ba^2)$, $w_\beta = \arccos(1/2a^2b^2)(a^3 - a^3b^2 - ab^2)$.

(α,β)	Energy	I	II	III
(0,0)	$a+b+1$	max	max	max
(π,0)	$-a+b-1$	min	loc min	loc min
(0,π)	$a-b-1$	loc min	loc min	min
(π,π)	$-a-b+1$	loc max	loc max	loc max
$\pm(w_\alpha, w_\beta)$	$-\dfrac{a^2+a^2b^2+b^2}{2ab}$	—	min	—

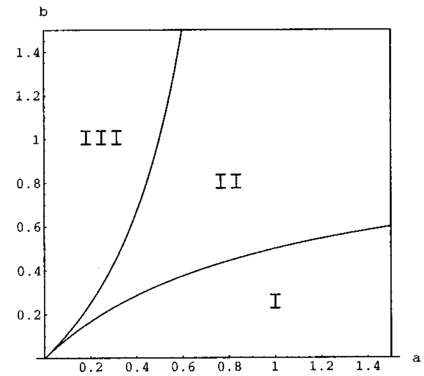

FIG. 2. Phase diagram for the triangular lattice, with $a = A^q$, $b = B^q$ and $b < a/(1+a)$ in region I, $a < b/(1+b)$ in region III.

Proof: For pq even and A,B,C in a weakly anisotropic phase, i.e., $A^q > B^q/(B^q + C^q)$ and $B^q > A^q/(A^q + C^q)$ the minimum of $h(z)$ is reached at $z = (\exp(iw_\alpha/q),\exp(iw_\beta/q))$, see Table II and hence not a $2q$th root of unity. Therefore one can apply Theorem 3.8 and gets with Theorem 3.7 the following estimate for the gap widths:

$$\max\left(\frac{A|\sin w_\alpha|}{2^{2q-4}(1+B/A+C/A)^{q-2}}, \frac{B|\sin \omega_\beta|}{2^{2q-4}(1+A/B+C/B)^{q-2}}\right) > 0.$$

For pq odd the proof is after rescaling of $h(z)$ analogous. □

Remark: Because of the Chambers relation (13) one of two consecutive gaps in the spectrum corresponds to $h_{\min}$ and the other to $h_{\max}$. Therefore Theorem 4.1 implies that at least one of two consecutive gaps is open.

As already mentioned above only those $\mathcal{H}_{\text{tri}}$ with $0 < A = B \leq C$ stay in the weakly anisotropic phase for any rational value of θ. In this case the following Corollary is a direct consequence of Theorem 4.1.

Corollary 4.2: For every $\theta = p/q$ with p and q relative prime integers and $0 < A = B \leq C$ the spectral gaps of $\mathcal{H}_{\text{tri}}$ corresponding to the minimum of $h(z)$ if pq even and to the maximum if pq odd, are open and have length at least

$$\frac{\sqrt{4-(A/C)^{2q}}A^{q-1}}{2^{2q-5}(2A+C)^{q-2}}.$$

Proof: Since $\sin w_\alpha = \sqrt{1-(A/C)^{2q}/4}$ the estimate follows from the one in the proof of Theorem 4.1. □

The analysis of the spectra for rational values of the magnetic flux per unit cell above does not yield all conditions needed in Theorem 3.11 and particularly Corollary 3.12. But looking at the proof again we see that nevertheless Corollary 3.12 holds.

Theorem 4.3: *Let $0 < A/C = \kappa \leq 1$, then for all $\theta \in \mathbb{R}$ with $|\theta - p/q| \leq \kappa^{2q-2}/\sqrt{5\pi}2^{4q-12}3^{2q}$, for arbitrarily large q, the spectrum of $\mathcal{H}_{\text{tri}}(\theta)$ is a Cantor set.*

Proof: Suppose θ fulfills the Liouville condition of the theorem. Take p and q. For pq even the gaps of Spec $\mathcal{H}_c(p/q)$ corresponding to the minimum of $h(z)$ are by Corollary 4.2 at least of size $\kappa^{q-2}A/2^{2q-5}3^{q-2}$. Since between such two gaps lies exactly one gap (possibly degenerated) corresponding to the maximum of $h(z)$, one gap of two consecutive gaps is at least of the size above. This is also true for pq odd. This fact induces analogous to the proof of Theorem 3.11 that there are at least $\lfloor (q-1)/2 \rfloor$ open gaps in the spectrum of $\mathcal{H}_{\text{tri}}(\theta)$ for every such p/q. For deducing that the gaps in Spec $\mathcal{H}_c(\theta)$ are dense the proof of Corollary 3.12 have to be altered at one point only. The estimate for the length of any interval in Spec $\mathcal{H}_{\text{tri}}$ for θ' rational is altered by a factor 2. □

V. CONCLUDING REMARKS

We have presented a derivation for an estimate of the gap widths in the spectrum of the Hamiltonians investigated here (6), using an algebraic and combinatorial approach. The Chambers relation extends to this class of Hamiltonians and its explicit form is described in terms of generalized binomial coefficients. Rather little is known about such kind of coefficients, especially the analytic formula for them (11) seems to be a new result. In the way the Chambers relation is derived, one easily gets conditions under which a general self-adjoint element in $A(\theta)$ fulfills the Chambers relation. So one recognizes that the class of Hamiltonians (6) contains essentially the most general type of Hamiltonians fulfilling the Chambers relation. A limitation of this method is given by the fact that the estimates derived for the spectral gaps vanish for critical points that are multiples of π. Numerical calculations suggest that, for arbitrary rational magnetic field and interaction, occasional ''gap closing'' occurs. So it is unlikely that the existence of a much more general estimate can be derived within this approach. However, we know from the derived estimate that this occasional ''gap closing'' can only occur if the relative location of the critical points are multiples of π or the generalized binomial coefficient in the estimate vanishes.

ACKNOWLEDGMENTS

We learned from P. van Mouche the idea to formulate the coexistence problem for the triangular lattice Hamiltonian in the algebraic setting and had stimulating discussions with him. This work was supported in part by the Studienstiftung des deutschen Volkes.

APPENDIX

Since the coefficients of the form $(qm;q)$ appear in the Chambers relation in Sec. III, we calculate them here using formula (2). Let $m=(1,0)$, then

$$
\begin{aligned}
[q;q]_{(m,\omega)} &= \prod_{k=1}^{q} \left(c_{(1,0)}+c_{(1,1)}\gamma^{\omega+2k-1}+c_{(1,-1)}\gamma^{-\omega-2k+1}\right) \\
&= \gamma^{-q(q+2-\omega)}c_{(1,1)}^{q}\prod_{k=1}^{q}(\lambda_1-\gamma^{2k})(\lambda_2-\gamma^{2k}) \\
&= \gamma^{-q(q+2-\omega)}c_{(1,1)}^{q}(\lambda_1^{q}-1)(\lambda_2^{q}-1) \\
&= \gamma^{q(q-\omega)}c_{(1,-1)}^{q}+2^{-q}\left(c_{(1,0)}+\sqrt{c_{(1,0)}^{2}-4c_{(1,1)}c_{(1,-1)}}\right)^{q} \\
&\quad+\gamma^{q(q+\omega)}c_{(1,1)}^{q}+2^{-q}\left(c_{(1,0)}-\sqrt{c_{(1,0)}^{2}-4c_{(1,1)}c_{(1,-1)}}\right)^{q}.
\end{aligned}
$$

The inverse transformation leads to

$$
(qm;q)=\begin{cases} 2\varepsilon^{q}T_q(c_m/2\varepsilon) & \text{for } \varepsilon=\sqrt{c_{(1,1)}c_{(1,-1)}}\neq 0 \\ c_m^{q} & \text{otherwise,} \end{cases}
$$

where $T_q(x)=\cos q \arccos x=1/2(x+\sqrt{x^2-1})^q+1/2(x-\sqrt{x^2-1})^q$ denotes the Chebyshev polynomial of degree q. Similarly we obtain an analogous expression for $(qm;q)$ with $m=(0,1)$. For $m=(\pm 1,\pm 1)$ the coefficients $(qm;q)$ are given by $(-1)^{pq}c_m^{q}$. Further we see that any other coefficient $(n;q)$ with $\max_l|n_l|=q$ is zero.

[1] R. Rammal and J. Bellissard, ''An algebraic semiclassical approach to Bloch electrons in a magnetic field,'' J. Phys. (France) **51**, 1803–1830 (1990).

[2] J. Bellissard, A. van Elst, and H. Schulz-Baldes, ''The noncommutative geometry of the quantum Hall effect,'' J. Math. Phys. **35**, 5373–5451 (1994).
[3] D. Hofstadter, ''The energy-levels of Bloch electrons in rational and irrational magnetic fields,'' Phys. Rev. B **14**, 2239–2249 (1976).
[4] Y. Last, ''Almost everything about the almost Mathieu operator. I, in Proceedings of the XIth International Conference of Mathematical Physics, Paris, edited by D. Iagolnitzer (Int. Press, Boston, 1995), pp. 366–372.
[5] S. Jitomirskaya and Y. Last, ''Anderson localization for the almost Mathieu equation. III. Semiuniform localization, continuity of gaps, and measure of the spectrum,'' mp-arc 97-426, 1997.
[6] B. Helffer and J. Sjöstrand, ''Semiclassical analysis for Harpers equation. III,'' Mém. SMF **39**, 117:4 (1989).
[7] F. Claro and G. Wannier, ''Magnetic subband structure of electrons in hexagonal lattices,'' Phys. Rev. B **19**, 6068–6074 (1979).
[8] J. Bellissard, Ch. Kreft, and R. Seiler, ''Analysis of the spectrum of a particle on a triangular lattice with two magnetic fluxes by algebraic and numerical methods,'' J. Phys. A **24**, 2329–2353 (1991).
[9] P. Kerdelhué, ''Equation de Schrödinger Magnétique Périodique avec Symétrie d'Ordre Six: Mesure du spectre. II,'' Ann. Inst. Henri Poincaré: Phys. Theor. **62**, 181–209 (1995).
[10] M. Choi, G. Elliott, and N. Yui, ''Gauss polynomials and the rotation algebra,'' Invent. Math. **99**, 225–246 (1990).
[11] A. Janner and T. Janssen, ''Electromagnetic compensating gauge transformations,'' Physica (Amsterdam) **53**, 1–27 (1971).
[12] J. Zak, ''Magnetic translation group,'' Phys. Rev. A **134**, 1602–1606 (1964); see also J. Schwinger, ''Unitary operator bases,'' Proc. Natl. Acad. Sci. **46**, 570–573 (1960).
[13] Ch. Kreft and R. Seiler, ''Models of the Hofstadter-type,'' J. Math. Phys. **37**, 5207–5243 (1996).
[14] M. P. Schützenberg, ''Une interprétation de certaines solutions de l'equation fonctionelle: $F(x+y)=f(x)f(y)$,'' C.R. Acad. Sci. Paris **236**, 352–353 (1953).
[15] In this context q is a complex number and one is usually interested in values between 0 and 1. Since q denoted here an integer, we ''γ^2 deform'' to prevent equivocalness.
[16] R. Askey, ''Extensions of hermite polynomials and other orthogonal polynomials,'' SIAM news **5**, 14 (1994).
[17] J. Bellissard, ''Noncommutative methods in semiclassical analysis,'' in *Transition to Chaos in Classical and Quantum Mechanics*, Montecatini Terme, Italy, 6–13 July 1991, edited by S. Graffi (Springer, Berlin, 1994).
[18] W. G. Chambers, ''Linear-network model for magnetic breakdown in two dimensions,'' Phys. Rev. **140**, A135–A143 (1965).
[19] The factor $q\sigma(n,m)+q^2 m_1 m_2$ has a geometric interpretation (Sec. II), it is twice the area of the rectangle with vertices $0,n,n+(qm_1,0),n+qm$.
[20] G. A. Elliott, ''Gaps in the spectrum of an almost periodic Schrödinger operator,'' C. R. Math. Rep. Acad. Sci. Canada **4**, 255–259 (1982).
[21] J. Avron, P. van Mouche, and B. Simon, ''On the measure of the spectrum for the almost Mathieu operator,'' Commun. Math. Phys. **132**, 103–118 (1990).
[22] J. Bellissard, ''Lipshitz continuity of gap boundaries for Hofstadter-like spectra,'' Commun. Math. Phys. **160**, 599–613 (1993).
[23] P. van Mouche (private communication).
[24] R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Pitman Research Notes in Mathematics Vol. 162 (Longman, London, 1987).
[25] J. Bellissard and B. Simon, ''Cantor spectrum for the almost Mathieu equation,'' J. Funct. Anal. **48**, 408–419 (1982).
[26] P. van Mouche, ''The coexistence problem for the discrete Mathieu operator,'' Commun. Math. Phys. **122**, 23–33 (1989).
[27] G. H. Wannier, G. M. Obermair, and R. Ray, ''Magnetoelectronic density of states for a model crystal,'' Phys. Status Solidi (b) **93**, 337 (1979).
[28] Since $\Pi_z \mathcal{H}_{\text{tri}}$ is self-adjoint, $g(\mathcal{H}_{\text{tri}})=0$ if and only if $g(\mathcal{H}_{\text{tri}}{}^*)^*=0$, hence it is no restriction to consider polynomials with real coefficients.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Weak coupling limit and removing an ultraviolet cutoff for a Hamiltonian of particles interacting with a quantized scalar field

Fumio Hiroshima[a),b)]
Institute of Applied Mathematics, University of Bonn, Wegeler str. 6, D53115 Bonn, Germany

(Received 6 April 1998; accepted for publication 4 September 1998)

An interaction system consisting of particles and a quantized scalar field is considered. The Hamiltonian of the system is defined as a self-adjoint operator in a Hilbert space. An ultraviolet cutoff is imposed on the Hamiltonian. A renormalized Hamiltonian is defined by subtracting a renormalization term from the Hamiltonian. Our aim in this paper is to remove the ultraviolet cutoff and take the weak coupling limit *simultaneously* for the renormalized Hamiltonian. By using a functional integral that contains a vector-valued stochastic integral, a Schrödinger Hamiltonian with a many-body Coulomb potential (resp., Yukawa potential) is derived, if the mass of the quantized scalar field is zero (resp., positive). © *1999 American Institute of Physics.* [S0022-2488(99)00202-9]

I. INTRODUCTION

In this paper, we pursue the study of an interaction system consisting of an arbitrary but conserved number of particles and a quantized scalar field with *non-negative mass*. An ultraviolet cutoff is imposed on the quantized scalar field. The Hamiltonian of the system is defined as a self-adjoint operator in a Hilbert space. A renormalized Hamiltonian is defined by subtracting a renormalization term from the Hamiltonian. Our aim in this paper is ''to remove the ultraviolet cutoff and take weak coupling limit *simultaneously*'' (we call it ''WCL-RUV'' for short) for the renormalized Hamiltonian. Then we derive a Schrödinger Hamiltonian with a many-body Coulomb potential (or Yukawa potential) in its WCL-RUV.

In Ref. 1, the author elaborates WCL-RUV for a model[2] with a *massive* quantized scalar field and shows that a Schrödinger Hamiltonian with a many-body Yukawa potential appears in its WCL-RUV (also see Refs. 3–5). In Ref. 1, it is crucial that the quantized scalar field has positive mass. Our main purpose in this paper is to extend the result in Ref. 1 to the case where the quantized scalar field has non-negative mass.

A mathematical formulation of the physical description of the interaction system is reduced to the theory of self-adjoint operators acting in the tensor product $\tilde{\mathcal{L}}$ of two Hilbert spaces. The statistics of the particles does not play any role. However, in this paper, we assume that the particles are fermions. (Naturally all the results extend to the case where the particles are bosons.) Let $\mathcal{F}_b$ and $\mathcal{F}_a$ be the Boson Fock space and the Fermion Fock space over $L^2(\mathbb{R}^3)$, respectively,

$$\mathcal{F}_b \equiv \bigoplus_{N=0}^{\infty} L_s^2(\mathbb{R}^{3N}), \quad \mathcal{F}_a \equiv \bigoplus_{N=0}^{\infty} L_{as}^2(\mathbb{R}^{3N}),$$

where $L_s(\mathbb{R}^{3N})$ [resp., $L_{as}(\mathbb{R}^{3N})$], $N \geqslant 1$, denotes the set of symmetric (resp., antisymmetric) functions in $L^2(\mathbb{R}^3)$ and $L_s^2(\mathbb{R}^0) = L_{as}^2(\mathbb{R}^0) \equiv \mathbb{C}$. Then

a)Electronic mail: hiro@mailcip.iam.uni-bonn.de
b)Permanent address: Department of Mathematics, Faculty of Science, Hokkaido University, Sapporo 060 Japan; Electronic mail: f-hirosh@math.sci.hokudai.ac.jp

$$\tilde{\mathcal{L}} \equiv \mathcal{F}_a \otimes \mathcal{F}_b \cong \bigoplus_{Z=0}^{\infty} \tilde{\mathcal{H}}_Z, \quad \tilde{\mathcal{H}}_Z = L^2_{as}(\mathbb{R}^{3Z}) \otimes \mathcal{F}_b.$$

The number Z counts the number of the particles. Let $a(k)$, $a^\dagger(k)$, $b(k)$, and $b^\dagger(k)$ be the formal annihilation and creation operators in $\mathcal{F}_b$ and the formal annihilation and creation operators in $\mathcal{F}_a$, respectively. Formally, we define

$$\Psi(x) = \frac{1}{\sqrt{(2\pi)^3}} \int b(k) e^{ikx}\, dk, \quad \Psi^\dagger(x) = \frac{1}{\sqrt{(2\pi)^3}} \int b^\dagger(k) e^{-ikx}\, dk.$$

The Hamiltonian $\tilde{H}$ of the system is explained as an operator acting in the Hilbert space $\tilde{\mathcal{L}}$. Formally, $\tilde{H}$ has the following form [see (4.16) for rigorous definition]:

$$\tilde{H} = \left(\int \Psi^\dagger(x) \left(-\frac{1}{2m} \Delta \right) \Psi(x) dx \right) \otimes I - g \int \Psi^\dagger(x) \phi_F(x) \Psi(x) dx + I \otimes \tilde{H}_b,$$

where Δ is the Laplacian in $L^2(\mathbb{R}^3)$, $g \in \mathbb{R}$ a coupling constant, $m>0$ the mass of the particle, $\omega(k) = \sqrt{k^2+\mu^2}$, $\mu \geqslant 0$, $\phi_F(x)$ a time-zero quantized scalar field in $\mathcal{F}_b$ and

$$\tilde{H}_b = \int \omega(k) a^\dagger(k) a(k) dk.$$

The restriction of $\tilde{H}$ to the sector $\tilde{\mathcal{H}}_Z$, $Z \geqslant 1$, has the form

$$\tilde{H}|_{\tilde{\mathcal{H}}_Z} = \frac{1}{2m} \mathbf{p}_Z^2 \otimes I - g \sum_{j=1}^{Z} \phi_F(x_j) + I \otimes \tilde{H}_b \Big|_{\tilde{\mathcal{H}}_Z}.$$

Here $\mathbf{p}_Z = (\mathbf{p}^1, \ldots, \mathbf{p}^Z)$, $\mathbf{p}^j = (-i(\partial/\partial x_1^j), -i(\partial/\partial x_2^j), -i(\partial/\partial x_3^j))$, $j=1,\ldots,Z$. In Ref. 1, for each fixed $Z \geqslant 1$, the author defines a scaling Hamiltonian $\tilde{H}^Z(\Lambda)$ in $\tilde{\mathcal{L}}_Z = L^2(\mathbb{R}^{3Z}) \otimes \mathcal{F}_b$ by

$$\tilde{H}^Z(\Lambda) = \frac{1}{2m} \mathbf{p}_Z^2 \otimes I - \Lambda g \tilde{H}_I^Z(\Lambda^\alpha) + \Lambda^2 I \otimes \tilde{H}_b, \quad \Lambda > 0, \quad \alpha > 0, \tag{1.1}$$

where $\tilde{H}_I^Z(\Lambda^\alpha)$ is defined by introducing an ultraviolet cutoff, which is parametrized by a parameter $\Lambda^\alpha > 0$, in $\Sigma_{j=1}^Z \phi_F(x_j)$. In (1.1), in the case where we make Λ's in the coefficients of $I \otimes \tilde{H}_b$ and $\tilde{H}_I^Z(\Lambda^\alpha)$ tend to infinity with Λ^α in $\tilde{H}_I^Z(\Lambda^\alpha)$ replaced by a fixed parameter, we call the limit ''weak coupling limit.'' [6,7] Conversely, the case where Λ^α in $\tilde{H}_I^Z(\Lambda^\alpha)$ tends to infinity with the other Λ's replaced by fixed parameters corresponds to removing the ultraviolet cutoff.[2] In Ref. 1, only in the case of *positive mass* $\mu > 0$, the author obtains the following (WCL-RUV):

$$s-\lim_{\Lambda \to \infty} e^{-t(\tilde{H}^Z(\Lambda) + V \otimes I - g^2 Z E(\Lambda^\alpha))} = e^{-t(H_{\text{eff}}^Z + V)} \otimes P_b, \quad 0 < \alpha < \tfrac{1}{2}, \tag{1.2}$$

where the potential V is infinitesimally small with respect to $\mathbf{p}_Z^2$, $E(\Lambda^\alpha)$ a renormalization term that goes to minus infinity as $\Lambda \to \infty$, P_b the projection operator onto the closed subspace generated by the vacuum vector in $\mathcal{F}_b$ and H_{eff}^Z is as follows:

$$H_{\text{eff}}^Z = \frac{1}{2m} \mathbf{p}_Z^2 - \frac{g^2}{4\pi} \sum_{1 \leqslant i < j \leqslant Z} \frac{e^{-\mu|x^i - x^j|}}{|x^i - x^j|}, \quad Z \geqslant 2, \quad \mu > 0.$$

[In Ref. 1, actually, the strong limit of the resolvent of $\tilde{H}^Z(\Lambda)+V_Z\otimes I-g^2ZE(\Lambda^\alpha)$ is considered.] In this paper, we consider WCL-RUV for the case of non-negative mass $\mu\geqslant 0$. We take the Schrödinger representation, $\mathcal{F}$, of $\mathcal{F}_b$.[8–10] The operators and spaces in the Schrödinger representation corresponding to $\tilde{H}^Z(\Lambda)$, $\tilde{H}_b$, $\tilde{H}_I^Z(\Lambda)$, $\tilde{\mathcal{L}}$, $\tilde{\mathcal{L}}_Z$, and $\tilde{\mathcal{H}}_Z$ are denoted by $H^Z(\Lambda)$, H_b, $H_I^Z(\Lambda)$, $\mathcal{L}$, $\mathcal{L}_Z$, and $\mathcal{H}_Z$, respectively, in what follows.

The basic method presented here is as follows: In the massive case, $\mu>0$, the author in Ref. 1 introduces a unitary operator that separates a renormalization term and transforms the Hamiltonian $H^Z(\Lambda)$ to an operator handled easily. Nevertheless, in the massless case, $\mu=0$, we cannot define such a unitary operator. To avoid this difficulty, we introduce a unitary operator with an infrared cutoff $K>0$, which transforms the renormalized Hamiltonian $H^Z(\Lambda)-gZE(\Lambda^\alpha)$ to an operator $H^Z(K,\Lambda)$. We shall prove that, for sufficiently large $\Lambda>0$, it is possible to construct a functional integral representation of $e^{-tH^Z(K,\Lambda)}$. The functional integral representation gives a good way to analyze the strong limit of $e^{-tH^Z(K,\Lambda)}$ as $\Lambda\to\infty$.

We organize this paper as follows. In Sec. II, we present the basic notation and facts. Section III is devoted to constructing a functional integral representation of $e^{-tH_{\mathrm{formal}}}$,

$$H_{\mathrm{formal}}=\frac{1}{2m}(\mathbf{p}_Z\otimes I-A)^2+U+V\otimes I+I\otimes H_b, \tag{1.3}$$

where A and U are defined in Sec. III. The operator H_{formal} is an abstract version of $H^Z(K,\Lambda)$. Section IV is the main section of this paper. In this section, we give the definition of the Hamiltonian $H(V,\Lambda)$ of the system in the Schrödinger representation and analyze its WCL-RUV. Theorem 4.11 is the main theorem in this paper. In Sec. V, we give some remarks.

II. SCHRÖDINGER REPRESENTATION

In this section we define basic notation and prepare some concepts on Schrödinger representation of a Boson Fock space and on a functional integral representation of a heat semigroup. For a Hilbert space $\mathcal{H}$ over $\mathbb{C}$, we denote the scalar product by $\langle f,g\rangle_{\mathcal{H}}$ and the associated norm by $\|f\|_{\mathcal{H}}$, where the scalar product is linear in g and antilinear in f. For a tempered distribution f, $\overline{f}$ denotes the complex conjugate of f and $\hat{f}$ (resp., $\check{f}$) the Fourier transform of f (resp., the inverse Fourier transform of f). We denote the domain of an operator A by $D(A)$. We denote by $C_b^n(\mathbb{R}^d;\mathcal{H})$ the set of n times strongly continuously differentiable functions, together with bounded up to n times derivative, from $\mathbb{R}^d$ to a Hilbert space $\mathcal{H}$ and $C_b(\mathbb{R}^d)$ the set of bounded continuous functions on $\mathbb{R}^d$. Let $\mathcal{S}_r'(\mathbb{R}^m)$ be the set of real tempered distributions on $\mathbb{R}^m$ and define

$$\mathcal{H}_{-1/2}=\left\{ f\in\mathcal{S}_r'(\mathbb{R}^3)\,\middle|\, \|f\|^2_{\mathcal{H}_{-1/2}}\equiv\int_{\mathbb{R}^3}\frac{|\hat{f}(k)|^2}{\omega(k)}\,dk<\infty\right\},$$

$$\mathcal{W}=\left\{ f\in\mathcal{S}_r'(\mathbb{R}^{3+1})\,\middle|\, \|f\|^2_{\mathcal{W}}\equiv 2\int_{\mathbb{R}^{3+1}}\frac{|\hat{f}(k,k_0)|^2}{\omega(k)^2+k_0^2}\,dk\,dk_0<\infty\right\}.$$

For simplicity, we put $\|f\|_{\mathcal{H}_{-1/2}}=\|f\|_{-1/2}$. Let $\{\phi(f)\,|\,f\in\mathcal{H}_{-1/2}\}$ be the Gaussian mean zero random process indexed by $\mathcal{H}_{-1/2}$ so that

$$\int_Q e^{i\phi(f)}\,d\mu=e^{-(1/4)\|f\|^2_{-1/2}},\quad f\in\mathcal{H}_{-1/2},$$

where (Q,μ) denotes a probability measure space. We regard ϕ as the variable of the Gaussian random process $\phi(f)$. Similarly, let $\{\Phi(f)\,|\,f\in\mathcal{W}\}$ be the Gaussian mean zero random process indexed by $\mathcal{W}$ on a probability measure space (Q_E,μ_E) with

$$\int_{Q_E} e^{i\Phi(f)}\, d\mu_E = e^{-(1/4)\|f\|^2_{\mathcal{W}}}, \quad f\in\mathcal{W}.$$

We set $\mathcal{F}=L^2(Q,d\mu)$ and $\mathcal{E}=L^2(Q_E,d\mu_E)$. The "Wick product" $:\phi(f_1)\cdots\phi(f_n):$ in $\mathcal{F}$ is defined by recurrences as follows:

$$:\phi(f):=\phi(f),$$

$$:\phi(f_1)\cdots\phi(f_n):=\phi(f_1):\phi(f_2)\cdots\phi(f_n):-\sum_{j=2}^{n}\frac{1}{2}\langle f_1,f_j\rangle_{-1/2}:\phi(f_2)\cdots\widehat{\phi(f_j)}\cdots\phi(f_n): \quad n\geqslant 2,$$

where $\widehat{\phi(f)}$ means the omission of the term $\phi(f)$. Set

$$\Gamma_0(\mathcal{F})=\mathbb{C},$$

$$\Gamma_n(\mathcal{F})=\overline{\mathbf{L}\{:\phi(f_1)\cdots\phi(f_n):|f_j\in\mathcal{H}_{-1/2},\ j=1,...,n\}}, \quad n\geqslant 1.$$

Here $\mathbf{L}\{\cdots\}$ denotes the complex linear hull of the vectors in $\{\cdots\}$ and $\overline{\{\cdots\}}$ the closure in $\mathcal{F}$. Then one sees that

$$\mathcal{F}=\bigoplus_{n=0}^{\infty}\Gamma_n(\mathcal{F}).$$

"The finite particle subspace" in $\mathcal{F}$ is defined by

$$\mathcal{F}_0=\bigcup_{N=0}^{\infty}\Big[\bigoplus_{n=0}^{N}\Gamma_n(\mathcal{F})\bigoplus_{n=N+1}^{\infty}\{0\}\Big],$$

which is dense in $\mathcal{F}$. Let T be a contraction linear operator from $\mathcal{H}_{-1/2}$ to $\mathcal{W}$ and h a non-negative self-adjoint operator in $\mathcal{H}_{-1/2}$. Then a linear operator $\Gamma(T)$ from $\mathcal{F}$ to $\mathcal{E}$ and a linear operator $d\Gamma(h)$ in $\mathcal{F}$ are defined by

$$\Gamma(T)\Omega_{\mathcal{F}}=\Omega_{\mathcal{E}},$$

$$\Gamma(T):\phi(f_1)\cdots\phi(f_n):=:\Phi(Tf_1)\ldots\Phi(Tf_n):, \quad f_1,\ldots,f_n\in\mathcal{H}_{-1/2}, \quad n\geqslant 1,$$

$$d\Gamma(h)\Omega_{\mathcal{F}}=0,$$

$$d\Gamma(h):\phi(f_1)\cdots\phi(f_n):=\sum_{j=1}^{n}:\phi(f_1)\cdots\phi(hf_j)\cdots\phi(f_n):, \quad f_1,\ldots,f_n\in D(h), \quad n\geqslant 1.$$

Here $\Omega_{\mathcal{F}}\equiv 1\in\mathcal{F}$, $\Omega_{\mathcal{E}}\equiv 1\in\mathcal{E}$. It is checked that $\Gamma(T)$ uniquely extends to a contraction linear operator from $\mathcal{F}$ to $\mathcal{E}$. We denote its extension by the same symbol. Moreover, we see that $d\Gamma(h)$ is essentially self-adjoint on $\mathcal{F}_0$. We also denote its self-adjoint extension by the same symbol. We define a non-negative self-adjoint operator $\tilde{\omega}$ in $\mathcal{H}_{-1/2}$ by

$$\widehat{\tilde{\omega}f}(k)=\omega(k)\hat{f}(k), \quad f\in\mathcal{H}_{-1/2},$$

with $D(\tilde{\omega})=\{f\in\mathcal{H}_{-1/2}|\sqrt{\omega}\hat{f}\in L^2(\mathbb{R}^3)\}$. The family of operators j_t: $\mathcal{H}_{-1/2}\to\mathcal{W}$, $t\geqslant 0$, is defined by

$$j_t f=\delta_t\otimes f, \quad f\in\mathcal{H}_{-1/2},$$

where δ_t is the delta function with mass at $t \in \mathbb{R}$. It is well known that j_t is isometry. We define the family of isometries J_t and a non-negative self-adjoint operator H_b by

$$J_t = \Gamma(j_t), \quad t \geqslant 0,$$

$$H_b = d\Gamma(\tilde{\omega}).$$

The following relation is crucial in the next section:

$$J_t^* J_s = e^{-|t-s|H_b}, \quad t,s > 0. \tag{2.1}$$

We shall give a unitary equivalence between $\mathcal{F}$ and $\mathcal{F}_b$. Define $\Omega = \{1,0,0,0,...\} \in \mathcal{F}_b$. Let the annihilation operator and the creation operator in $\mathcal{F}_b$ denote by $a(f), f \in L^2(\mathbb{R}^3)$, and $a^\dagger(g), g \in L^2(\mathbb{R}^3)$, respectively, which satisfy the canonical commutation relations on a dense domain:

$$[a(f), a^\dagger(g)] = \langle \overline{f}, g \rangle_{L^2(\mathbb{R}^3)}, \quad [a^\dagger(f), a^\dagger(g)] = [a(f), a(g)] = 0.$$

We define

$$\phi_F(\hat{f}) = \frac{1}{\sqrt{2}} \left\{ a^\dagger\left(\frac{\tilde{\hat{f}}}{\sqrt{\omega}} \right) + a\left(\frac{\hat{f}}{\sqrt{\omega}} \right) \right\}, \quad \frac{\hat{f}}{\sqrt{\omega}} \in L^2(\mathbb{R}^3),$$

$$\pi_F(\hat{f}) = \frac{i}{\sqrt{2}} \{ a^\dagger(\sqrt{\omega}\tilde{\hat{f}}) - a(\sqrt{\omega}\hat{f}) \}, \quad \sqrt{\omega}\hat{f} \in L^2(\mathbb{R}^3).$$

Here $\tilde{g}(k) = g(-k)$. Define a map $\mathcal{T}$ from $\mathcal{F}_b$ to $\mathcal{F}$ by

$$\mathcal{T}\Omega = \Omega_{\mathcal{F}},$$

$$\mathcal{T}: \phi_F(\hat{f}_1) \cdots \phi_F(\hat{f}_n): \Omega = :\phi(f_1) \cdots \phi(f_n):, \quad f_1, \ldots, f_n \in \mathcal{H}_{-1/2},$$

and extend $\mathcal{T}$ by linearity. Here the "Wick product" $:\phi_F(f_1) \cdots \phi_F(f_n):$ in $\mathcal{F}_b$ is defined by moving all the creation operators to the left side and all the annihilation operators to the right side without commutation relations. The operator $\mathcal{T}$ uniquely extends to a unitary operator from $\mathcal{F}_b$ to $\mathcal{F}$. We denote its extension by the same symbol. We see that

$$\mathcal{T}^{-1}\phi(\hat{f})\mathcal{T} = \phi_F(f), \quad f \in \mathcal{H}_{-1/2}, \quad \mathcal{T}^{-1} H_b \mathcal{T} = \tilde{H}_b. \tag{2.2}$$

Moreover, $\mathcal{T}$ implements the following unitary equivalence:

$$\tilde{\mathcal{L}} \cong \mathcal{F}_a \otimes \mathcal{F} \equiv \mathcal{L}.$$

The Hilbert space $\mathcal{L}$ can be decomposed as follows:

$$\mathcal{L} = \bigoplus_{Z=0}^{\infty} \mathcal{H}_Z, \quad \mathcal{H}_Z = L^2_{as}(\mathbb{R}^{3Z}) \otimes \mathcal{F}.$$

III. FUNCTIONAL INTEGRAL

In this section, we derive a functional integral representation of the heat semigroup generated by self-adjoint operators in (1.3) acting in the Hilbert space,

$$\mathcal{L}_Z \equiv L^2(\mathbb{R}^{3Z})\otimes\mathcal{F}\cong\int_{\mathbb{R}^{3Z}}^{\oplus}\mathcal{F}\,dx.$$

Throughout this section, we suppose that $A(x)=(A_1(x),\ldots,A_{3Z}(x))$ has the form

$$A_\mu(x)=P_\mu(\phi(f_{\mu,1}(x)),\ldots,\phi(f_{\mu,M_\mu}(x)))+A_\mu^0(x)\otimes I,\quad \mu=1,\ldots,3Z, \tag{3.1}$$

where $P_\mu(y_1,\ldots,y_{M_\mu})$ is a real polynomial, $f_{\mu,j}\in C_b^2(\mathbb{R}^{3Z};\mathcal{H}_{-1/2})$, and $A_\mu^0\in C_b^2(\mathbb{R}^{3Z};\mathbb{R})$. The operator H_{formal} in (1.3) is well defined on $\mathcal{L}_Z^\infty=C_0^\infty(\mathbb{R}^{3Z})\hat{\otimes}[\mathcal{F}_0\cap D(H_b)]$, but it is not known whether $\mathcal{L}_Z^\infty$ is a core for H_{formal} or not. Here, $\hat{\otimes}$ denotes algebraic tensor product. Thus, it is possible for $H_{\text{formal}}|_{\mathcal{L}_Z^\infty}$ not to have a unique self-adjoint extension. Then we have to make it clear which self-adjoint extensions we choose in consideration. We shall take the following strategy: At the beginning, we construct the family of contraction self-adjoint operators, Q_s, $s\geqslant 0$, from which we derive a strongly continuous symmetric one-parameter contraction semigroup, G_t, $t\geqslant 0$. Second, we show that its generator, $H_{00}(A)$, i.e., $e^{-tH_{00}(A)}=G_t$, which is a non-negative self-adjoint operator, has the same action on the domain $\mathcal{L}_Z^\infty$ as that of

$$\tilde{H}(A)=\frac{1}{2m}(\mathbf{p}_Z\otimes I-A)^2.$$

Next, we define $H_0(A)$ by the quadratic form sum of $H_{00}(A)$ and $I\otimes H_b$:

$$H_0(A)=H_{00}(A)\dot{+}I\otimes H_b. \tag{3.2}$$

Finally, by using a diamagnetic inequality[8,11] for $e^{-tH_0(A)}$, we shall show that the following self-adjoint operator can be defined:

$$H(A)=H_0(A)\dot{+}U\dot{+}(V_+\otimes I)\dot{-}(V_-\otimes I), \tag{3.3}$$

where U is a relatively form bounded operator with respect to

$$H_F^Z=\frac{1}{2m}\mathbf{p}_Z^2\otimes I+I\otimes H_b,$$

with relative bound <1 so that $U(\cdot)$ is an $\mathcal{F}$-valued continuous function on $\mathbb{R}^{3Z}$, $V_+\in L^1_{\text{loc}}(\mathbb{R}^{3Z})$ and V_- a relatively $\mathbf{p}_Z^2$ form bounded real multiplication operator. We adopt $H(A)$ as a mathematically rigorous definition of the formally defined Hamiltonian H_{formal} in (1.3). We shall give a functional integral representation of $e^{-tH(A)}$.

In what follows, for simplicity, we put $m=1$. For each $x,y\in\mathbb{R}^{3Z}$, we define a unitary operator on $\mathcal{F}$ by

$$U(x,y)=\exp\left\{\frac{i}{2}(A(x)+A(y))(x-y)\right\}.$$

Let $p_s(x)$ be a heat kernel, i.e., the integral kernel of $e^{-s(-(1/2)\Delta)}$ in $L^2(\mathbb{R}^{3Z})$:

$$p_s(x)=(2\pi s)^{-(3Z/2)}\exp\left(-\frac{1}{2s}|x|^2\right),\quad s>0,\quad x\in\mathbb{R}^{3Z}.$$

We define the family of contraction self-adjoint operators $\{Q_s\}_{s\geqslant 0}$ by Bochner integrals,

$$(Q_sF)(x)=\int_{\mathbb{R}^{3Z}}p_s(x-y)U(x,y)F(y)dy,$$

$$(Q_0F)(x)=F(x),\quad F\in\mathcal{L}_Z,\quad x\in\mathbb{R}^{3Z}.$$

Note that Q_sF is weakly right continuous at $s=0$, i.e.,

$$\lim_{s\downarrow 0}\langle Q_sF,G\rangle_{\mathcal{L}_Z}=\langle F,G\rangle_{\mathcal{L}_Z},\quad F,G\in\mathcal{L}_Z. \tag{3.4}$$

We show the following key lemma.

Lemma 3.1: Let $G\in\mathcal{L}_Z$ *and* $F\in\mathcal{L}_Z^\infty$. *Then* $\langle Q_sF,G\rangle$ *is differentiable at* $s>0$ *and right differentiable at* $s=0$, *with*

$$\lim_{s\downarrow 0}\left\langle \frac{Q_s-Q_0}{s}F,G\right\rangle_{\mathcal{L}_Z}=-\langle\tilde{H}(A)F,G\rangle_{\mathcal{L}_Z}. \tag{3.5}$$

Proof: We show an outline of a proof. Put

$$A_\mu^{(1)}(x,y)=\sum_{n=1}^{3Z}\left(\frac{\partial A_n(y)}{\partial y_\mu}(x_n-y_n)-(A_\mu(x)+A_\mu(y))\right),$$

$$A_\mu^{(2)}(x,y)=\sum_{n=1}^{3Z}\left(\frac{\partial^2 A_n(y)}{\partial y_\mu^2}(x_n-y_n)-2\frac{\partial A_\mu(y)}{\partial y_\mu}\right),\quad \mu=1,...,3Z.$$

By the Fubini theorem, one sees that

$$\frac{d}{ds}\langle Q_s,F,G\rangle_{\mathcal{L}_Z}=\int_{\mathbb{R}^{3Z}}p_s(X)dX\int_{\mathbb{R}^{3Z}}\Gamma(x,x-X)dx,$$

where

$$\Gamma(x,y)=\left\langle U(x,y)\frac{\partial^2F(y)}{\partial y_\mu^2},G(x)\right\rangle_{\mathcal{F}}+2\left\langle U(x,y)A_\mu^{(1)}(x,y)\frac{\partial F(y)}{\partial y_\mu},G(x)\right\rangle_{\mathcal{F}}$$
$$+\langle U(x,y)\{A_\mu^{(1)}(x,y)^2+A_\mu^{(2)}(x,y)\}F(y),G(x)\rangle_{\mathcal{F}}.$$

By a direct calculation, we show that

$$\left|\int_{\mathbb{R}^{3Z}}\Gamma(x,x-X)dx\right|\leqslant\epsilon_1+\epsilon_2|X|+\epsilon_3|X|^2.$$

Here $\epsilon_1,\epsilon_2,\epsilon_3$ are positive constants (see Ref. 8, Lemma 4.1 for details). Hence, we have

$$\lim_{s\downarrow 0}\int_{\mathbb{R}^{3Z}}p_s(X)dX\int_{\mathbb{R}^{3Z}}\Gamma(x,x-X)dx=\int_{\mathbb{R}^{3Z}}\Gamma(x,x)dx=-\langle\tilde{H}(A)F,G\rangle_{\mathcal{L}_Z},$$

which, together with (3.4), implies (3.5). The proof is complete. □

We fix probabilistic notation. Let (Ω,db) be a probability space for the $3Z$-dimensional Brownian motion $(b(t))_{t\geqslant 0}=(b_\mu(t))_{t\geqslant 0,\mu=1,...,3Z}$. Put $x_\mu+b_\mu(t)=\omega_\mu(t)$, $\mu=1,...,3Z$, and $\omega(t)=(\omega_1(t),\ldots,\omega_{3Z}(t))$. We define measure spaces $(\tilde{Q},\nu)$ and $(\tilde{Q}_E,\nu_E)$ by

$$\widetilde{Q}=\mathbb{R}^{3Z}\times\Omega\times Q,\quad d\nu=dx\otimes db\otimes d\mu,$$

$$\widetilde{Q}_E=\mathbb{R}^{3Z}\times\Omega\times Q_E,\quad d\nu_E=dx\otimes db\otimes d\mu_E.$$

Lemma 3.2: For all $t\geqslant 0$, the strong limit,

$$s-\lim_{n\to\infty}(Q_{t/2^n})^{2^n}\equiv G_t,$$

exists. Moreover, G_t is a strongly continuous symmetric one-parameter contraction semigroup on $\mathcal{L}_Z$ and has the following functional integral representation:

$$\langle F,G_tG\rangle_{\mathcal{F}}=\int_{\widetilde{Q}}d\nu\,\overline{F(\omega(t))}G(\omega(0))e^{i\widetilde{K}(A,t)},\quad F,G\in\mathcal{L}_Z,\tag{3.6}$$

where

$$\widetilde{K}(A,t)=\sum_{\mu=t}^{3Z}\int_0^t A_\mu(\omega(s))db_\mu(s)+\frac{1}{2}\sum_{\mu=1}^{3Z}\int_0^t\frac{\partial A_\mu}{\partial x_\mu}(\omega(s))ds\equiv\int_0^t A(\omega(s))\circ db(s),$$

where $\int\cdots db_\mu(s)$ is an $\mathcal{F}$-valued stochastic integral.

Proof: By the definition of Q_t, one sees that

$$\langle F,(Q_{t/2^n})^{2^n}(Q_{s/2^m})^{2^m}G\rangle_{\mathcal{L}_Z}=\int_{\mathbb{R}^{3Z}}dx\langle F(\omega(t+s)),e^{i\widetilde{K}_{m,n}(A,t,s)}G(\omega(0))\rangle_{L^2(\Omega;\mathcal{F})},$$

where $L^2(\Omega;\mathcal{F})$ is the set of $\mathcal{F}$-valued L^2 functions on Ω and

$$2\widetilde{K}_{m,n}(A,t,s)=\sum_{k=1}^{2^m}\left\{A\left(\omega\left(\frac{sk}{2^m}\right)\right)+A\left(\omega\left(\frac{s(k-1)}{2^m}\right)\right)\right\}\left\{\omega\left(\frac{sk}{2^m}\right)-\omega\left(\frac{s(k-1)}{2^m}\right)\right\}$$
$$+\sum_{k=1}^{2^n}\left\{A\left(\omega\left(\frac{tk}{2^n}+s\right)\right)+A\left(\omega\left(\frac{t(k-1)}{2^n}+s\right)\right)\right\}\left\{\omega\left(\frac{tk}{2^n}+s\right)-\omega\left(\frac{t(k-1)}{2^n}+s\right)\right\}.$$

Since, by (3.1), $A_\mu\in C_b^1(\mathbb{R}^{3Z};\mathcal{F})$, $\mu=1,...,3Z$, it is seen that

$$s-\lim_{m\to\infty}\lim_{n\to\infty}\widetilde{K}_{m,n}(A,t,s)=\widetilde{K}(A,t+s),$$

in $L^2(\Omega;\mathcal{F})$. Then we have, by the Lebesgue dominated convergence theorem,

$$\lim_{m\to\infty}\lim_{n\to\infty}\langle F,(Q_{t/2^n})^{2^n}(Q_{s/2^m})^{2^m}G\rangle_{\mathcal{L}_Z}=\int_{\widetilde{Q}}d\nu\,\overline{F(\omega(t+s))}G(\omega(0))e^{i\widetilde{K}(A,t+s)}.$$

Hence, it follows that $(Q_{t/2^n})^{2^n}$ is a Cauchy sequence in $\mathcal{L}_Z$ and $s-\lim_{n\to\infty}(Q_{t/2^n})^{2^n}$ has the functional integral representation (3.6). The semigroup property, and the strong continuity of G_t in t can be checked by (3.6). Thus, the proof is complete. □

Lemma 3.2 and Stone's theorem yield that there exists a non-negative self-adjoint operator $H_{00}(A)$, so that

$$G_t=e^{-tH_{00}(A)},\quad t\geqslant 0.$$

Lemma 3.3: The self-adjoint operator $H_{00}(A)$ is a self-adjoint extension of $\widetilde{H}(A)|_{\mathcal{L}_Z^\infty}$.

Proof: Let $F\in D(H_{00}(A))$ and $G\in\mathcal{L}_Z^\infty$. Then we see that, by Lemma 3.1,

$$\left\langle \frac{1}{t}\,(e^{-tH_{00}(A)}-I)G,F\right\rangle_{\mathcal{L}_Z}=\lim_{n\to\infty}\left\langle \frac{1}{t}\,\{(Q_{t/2^n})^{2^n}-I\}G,F\right\rangle_{\mathcal{L}_Z}$$

$$=\lim_{n\to\infty}\sum_{j=0}^{2^n-1}\frac{1}{2^n}\left\langle \frac{2^n}{t}\,\{Q_{t/2^n}-I\}G,(Q_{t/2^n})^jF\right\rangle_{\mathcal{L}_Z}$$

$$=-\int_0^1\langle\tilde{H}(A)G,e^{-tsH_{00}(A)}F\rangle_{\mathcal{L}_Z}\,ds.$$

As $t\to 0$, we get

$$\langle G,H_{00}(A)F\rangle_{\mathcal{L}_Z}=\langle\tilde{H}(A)G,F\rangle_{\mathcal{L}_Z}.$$

Thus, the proof is complete. □

Remark 3.4: Since the operator H_{formal} is defined on a larger domain than $\mathcal{L}_Z^\infty$, one can extend Lemma 3.3 to a larger domain. Actually, putting

$$\mathcal{M}_Z^\infty=\{u\in C_b^1(\mathbb{R}^{3Z};\mathbb{R})\,|\,\|\partial^k u\|_{L^2(\mathbb{R}^{3Z})}<\infty,|k|\leqslant 2\}\hat{\otimes}\mathcal{F}_0,$$

we also show that Lemmas 3.1 and 3.3 hold with $\mathcal{L}_Z^\infty$ replaced by $\mathcal{M}_Z^\infty$.[8]

Define $H_0(A)$ as (3.2). We state the main theorem in this section.

Theorem 3.5: *Let $A_\mu(x)\in D(H_b)$, $\mu=1,...,3Z$, for all $x\in\mathbb{R}^{3Z}$, and*

$$\sup_{x\in\mathbb{R}^{3Z}}\|H_bA_\mu(x)\|_{\mathcal{F}}<\infty. \tag{3.7}$$

Furthermore, we suppose that a multiplication operator U is a bounded operator on $\mathcal{L}_Z$ and $U(\cdot)$ is an $\mathcal{F}$-valued continuous function on $\mathbb{R}^{3Z}$. Then, for F, $G\in\mathcal{L}_Z$,

$$\langle F,e^{-t(H_0(A)+U)}G\rangle_{\mathcal{L}_Z}=\int_{\tilde{Q}_E}d\nu_E\,\overline{J_tF(\omega(t))}J_0G(\omega(0))e^{i\mathcal{K}(A,t)}e^{-E(U,t)}, \tag{3.8}$$

where

$$\mathcal{K}(A,t)=\sum_{\mu=1}^{3Z}\int_0^t J_sA_\mu(\omega(s))db_\mu(s)+\sum_{\mu=1}^{3Z}\frac{1}{2}\int_0^t J_s\,\frac{\partial A_\mu}{\partial x_\mu}\,(\omega(s))ds\equiv\int_0^t J_sA(\omega(s))\circ db(s),$$

$$E(U,t)=\int_0^t J_sU(\omega(s))ds,$$

where $\int_0^t\ldots db_\mu(s)$ is an $\mathcal{E}$-valued stochastic integral.

Proof: By the strong Trotter product formula,[12] we see that

$$\langle F,e^{-t(H_0(A)+U)}G\rangle_{\mathcal{L}_Z}=\lim_{n\to\infty}\langle F,(e^{-(t/2^n)H_{00}(A)}e^{-(t/2^n)I\otimes H_b}e^{-(t/2^n)U})^{2^n}G\rangle_{\mathcal{L}_Z}\equiv\lim_{n\to\infty}S_{2^n}.$$

Put $t/2^n=\epsilon$. From the definition of $H_{00}(A)$ and (2.1), it follows that

$$S_{2^n} = \lim_{k\to\infty} \langle F, J_t^*(J_t(Q_{\epsilon/2^k})^{2^k}J_t^*)(J_{t-\epsilon}e^{-\epsilon U}J_{t-\epsilon}^*)(J_{t-\epsilon}(Q_{\epsilon/2^k})^{2^k}J_{t-\epsilon}^*)(J_{t-2\epsilon}e^{-\epsilon U}J_{t-2\epsilon}^*)$$
$$\cdots(J_\epsilon e^{-\epsilon U}J_\epsilon^*)(J_\epsilon(Q_{\epsilon/2^k})^{2^k}J_\epsilon^*)(J_0 e^{-\epsilon U}J_0^*)J_0 G\rangle_{\mathcal{L}_Z}$$
$$= \lim_{k\to\infty} \int_{\mathbb{R}^{3Z}\times\underbrace{\mathbb{R}^{3Z\cdot 2^k}\times\cdots\times\mathbb{R}^{3Z\cdot 2^k}}_{2^n}} dx\, d\vec{x}_1\cdots d\vec{x}_{2^n}\, P_\epsilon(\vec{x}_1)\cdots P_\epsilon(\vec{x}_{2^n})$$
$$\times\langle F(x), J_t^*(E_t W_t(\vec{x}_1)E_t)(E_{t-\epsilon}U_{t-\epsilon}(x_1^{2^k})E_{t-\epsilon})(E_{t-\epsilon}W_{t-\epsilon}(\vec{x}_2)E_{t-\epsilon})$$
$$\cdots(E_\epsilon U_\epsilon(x_{2^n-1}^{2^k})E_\epsilon)(E_\epsilon W_\epsilon(\vec{x}_{2^n})E_\epsilon)(E_0 U_0(x_{2^n}^{2^k})E_0)J_0 G(x_{2^n}^{2^k})\rangle_{\mathcal{F}}$$
$$\equiv \lim_{k\to\infty} S_{2^n,2^k}, \tag{3.9}$$

where $J_sJ_s^* = E_s$ and $\vec{x}_j = (x_j^1,\ldots,x_j^{2^k}) \in \mathbb{R}^{3Z\cdot 2^k}$, $j=1,\ldots,2^n$,

$$P_\epsilon(\vec{x}_j) = p_\epsilon(x_{j-1}^{2^k} - x_j^1)p_\epsilon(x_j^1 - x_j^2)\cdots p_\epsilon(x_j^{2^k-1} - x_j^{2^k}),$$

$$W_s(\vec{x}_j) = \exp\left\{\frac{i}{2} J_s \sum_{l=1}^{2^k} (A(x_j^{l-1}) + A(x_j^l))(x_j^{l-1} - x_j^l)\right\}, \quad x_j^0 = x_{j-1}^{2^k}, \quad x_0^{2^k} \equiv x,$$

$$U_s(x_j^i) = \exp(-\epsilon J_s U(x_j^i)).$$

By the Markov property of E_s,[10] one can neglect E_s's in (3.9). Then it holds that

$$S_{2^n,2^k} = \int_{\mathbb{R}^{3Z}} dx \langle J_t F(\omega(t)), e^{\{i\Sigma_{j=0}^{2^n-1} J_{jt/2^n}\mathcal{K}_k(jt/2^n) - (t/2^n)\Sigma_{j=1}^{2^n} J_{jt/2^n}U(\omega(jt/2^n))\}} J_0 G(x)\rangle_{L^2(\Omega;\mathcal{E})},$$

where

$$\mathcal{K}_k(T) = \frac{1}{2}\sum_{m=1}^{2^k}\left\{A\left(\omega\left(\frac{m}{2^k}\frac{t}{2^n} + T\right)\right) + A\left(\omega\left(\frac{(m-1)}{2^k}\frac{t}{2^n} + T\right)\right)\right\}$$
$$\times\left\{\omega\left(\frac{m}{2^k}\frac{t}{2^n} + T\right) - \omega\left(\frac{(m-1)}{2^k}\frac{t}{2^n} + T\right)\right\}, \quad j = 0,\ldots,2^n - 1.$$

As is seen in the proof of Lemma 3.2, since $A_\mu(\cdot) \in C_b^2(\mathbb{R}^{3Z},\mathcal{F})$, we have

$$\lim_{k\to\infty} S_{2^n,2^k} = \int_{\mathbb{R}^{3Z}} dx \langle J_t F(\omega(t)), e^{\{i\Sigma_{j=0}^{2^n-1}\mathcal{K}(jt/2^n) - (t/2^n)\Sigma_{j=1}^{2^n} J_{jt/2^n}U(\omega(jt/2^n))\}} J_0 G(x)\rangle_{L^2(\Omega;\mathcal{E})},$$

where

$$\mathcal{K}(T) = \int_T^{T+t/2^n} J_T A(\omega(s))\circ db(s).$$

Since J_s is isometry, (2.1) and (3.7) implies that

$$\|J_sA(x) - J_{s'}A(x)\|_{\mathcal{F}}^2 \leq 2|s-s'| \sup_{x\in\mathbb{R}^{3Z}} \|A(x)\|_{\mathcal{F}} \|H_bA(x)\|_{\mathcal{F}}, \quad s,s' \in \mathbb{R}.$$

Thus one sees that (see Ref. 8, Theorem 2.5 for details)

$$s-\lim_{n\to\infty}\sum_{j=0}^{2^n-1}\mathcal{K}\left(\frac{jt}{2^n}\right)=\int_0^t J_sA(\omega(s))\circ db(s)$$

in $L^2(\Omega;\mathcal{E})$. Since $J.U(\cdot)$ is an $\mathcal{E}$-valued continuous function on $\mathbb{R}\times\mathbb{R}^{3Z}$, we have

$$\lim_{n\to\infty}\frac{t}{2^n}\sum_{j=0}^{2^n-1}J_{jt/2^n}U\left(\omega\left(\frac{jt}{2^n}\right)\right)=\int_0^t J_sU(\omega(s))ds,$$

in $L^2(\Omega;\mathcal{E})$. Hence, again by the Lebesgue dominated convergence theorem, we get the desired result. □

Remark 3.6: It is enough to assume that $A_\mu(\cdot)\in C_b^1(\mathbb{R}^{3Z};\mathcal{F})$, $\mu=1,...,3Z$ *and (3.7) to define the* $\mathcal{E}$*-valued stochastic integral* $\int_0^t J_sA(\omega(s))\circ db(s)$. *However, it is difficult to prove (3.8) under the above conditions. One of the reasons is that, since we do not know a concrete core for* H_{formal}, *we cannot use a limiting argument. Thus, we need the additional condition that* $f_{\mu,j}\in C_b^2(\mathbb{R}^{3Z};\mathcal{H}_{-1/2})$ *and* $A_\mu^0\in C_b^2(\mathbb{R}^{3Z};\mathbb{R})$, $\mu=1,\ldots,3Z, j=1,\ldots,M_\mu$, which implies that $A_\mu(\cdot)\in C_b^2(\mathbb{R}^{3Z};\mathcal{F})$ *to verify (3.8).*

We consider an extension of Theorem 3.5 to a much more general multiplication operator U. From Theorem 3.5 and the fact that J_t is positivity preserving, the following inequality follows:

$$|\langle F,e^{-t(H_0(A)+U)}G\rangle_{\mathcal{L}_Z}|\leqslant\langle|F|,e^{-t(H_F^Z+U)}|G|\rangle_{\mathcal{L}_Z}. \tag{3.10}$$

Let us define for $G\in\mathcal{L}_Z$,

$$\text{sgn } G(x)=\begin{cases}\dfrac{G(x)}{|G(x)|}, & |G(x)|\neq 0,\\ 0, & \text{otherwise.}\end{cases}$$

Lemma 3.7: Let $|U|$ *be a multiplication operator that is* H_F^Z*-form bounded with relative bound* ϵ. *Then* $|U|$ *is* $H_0(A)$*-form bounded with relative bound* $\leqslant\epsilon$.

Proof: Set $\tilde{G}=\text{sgn}(e^{-tH_0(A)}G)$. Putting $F=\tilde{G}K$, $K\geqslant 0$, and $U=0$ in (3.10), we have

$$\langle K,|e^{-tH_0(A)}G|\rangle_{\mathcal{L}_Z}\leqslant\langle K,e^{-tH_F^Z}|G|\rangle_{\mathcal{L}_Z}.$$

Hence, it follows that for a.e. $(x,\phi)\in\mathbb{R}^{3Z}\times Q$, $|e^{-tH_0(A)}G|(x,\phi)\leqslant(e^{-tH_F^Z}|G|)(x,\phi)$. Then a fundamental calculation shows that, for $E>0$,

$$|(H_0(A)+E)^{-1/2}G|(x,\phi)\leqslant((H_F^Z+E)^{-1/2}|G|)(x,\phi),\quad \text{a.e.}\quad (x,\phi)\in\mathbb{R}^{3Z}\times Q.$$

Hence, we have

$$\sup_{\|G\|_{\mathcal{L}_Z}=1}\||U|^{1/2}(H_0(A)+E)^{-1/2}G\|_{\mathcal{L}_Z}\leqslant\sup_{\|G\|_{\mathcal{L}_Z}=1}\||U|^{1/2}(H_F^Z+E)^{-1/2}|G|\,\|_{\mathcal{L}_Z}.$$

Thus, the lemma follows. □

We define a class of multiplication operators in $L^2(\mathbb{R}^{3Z})$.

Definition 3.8: We say that a multiplication operator $W\in\mathcal{M}_\pm(Z)$, $Z\geqslant 1$, *if* W_- *is relatively form bounded with respect to* $\mathbf{p}_Z^2$ *with relative bound* <1 *and* $W_+\in L^1_{loc}(\mathbb{R}^{3Z})$, *where* $W=W_+-W_-$ *(*W_+ *is the non-negative part of W and* $-W_-$ *is the negative part of W). Moreover, we say that* $W\in\mathcal{M}_\pm(Z)_r$ *if* $W\in\mathcal{M}_\pm(Z)$ *and* W_- *and* W_+ *are reduced by the closed subspace* $L^2_{as}(\mathbb{R}^{3Z})$ *in* $L^2(\mathbb{R}^{3Z})$, *respectively.*

Theorem 3.9: *Let U be an H_F^Z-form bounded multiplication operator with relative bound* <1 *so that* $U(\cdot)$ *is an* $\mathcal{F}$*-valued continuous function on* $\mathbb{R}^{3Z}$*, and* $V\in\mathcal{M}_\pm(Z)$*. Then Theorem 3.5 holds with* $H_0(A)+U$ *replaced by* $H_0(A)\dot{+}U\dot{+}(V_+\otimes I)\dot{-}(V_-\otimes I)$*.*

Proof: First, we assume that $V\in L^\infty(\mathbb{R}^{3Z})$. Then, by virtue of Lemma 3.7, $H_0(A)\dot{+}U+V\otimes I$ is a well-defined self-adjoint operator (Ref. 13 KLMN Theorem). The strong Trotter product formula for forms and a limiting argument with respect to V yield (3.8) with $H_0(A)+U$ replaced by $H_0(A)\dot{+}U+V\otimes I$ in the same way as the proof of Theorem 3.5. Next, for any $V\in\mathcal{M}_\pm(Z)$, defining

$$V_{+n}=\begin{cases} V_+, & V_+<n,\\ n, & \text{otherwise,}\end{cases}\quad V_{-m}\begin{cases} V_{-1}, & V_-<m,\\ m, & \text{otherwise,}\end{cases}$$

one sees that (3.8) holds with $H_0(A)+U$ replaced by $H_0(A)\dot{+}U+(V_{+n}\otimes I)-(V_{-m}\otimes I)$. By using convergence theorems for forms as $n,m\to\infty$ (Ref. 8, Theorem 4.13; Ref. 14, Theorem 6.2), we get the desired result. □

We conclude this section with giving a typical example of functional integral representations. Let $f=(f_1,\ldots,f_{3Z})$ satisfy

$$\frac{f_\mu}{\sqrt{\omega}}\in L^2(\mathbb{R}^3),\quad \overline{f_\mu(k)}=f_\mu(-k),\quad \mu=1,...,3Z. \tag{3.11}$$

We introduce a notation $\tilde{f}_\mu(x)$ for $x=(x^1,\ldots,x^N)\in\mathbb{R}^{3Z}$ as follows:

$$\tilde{f}_\mu(x)=\frac{1}{\sqrt{(2\pi)^3}}\,(f_\mu(k)e^{-ikx^{[(\mu/3)+1]}})^\vee,\quad \mu=1,..,3Z,$$

where $[s]$ is the integer part of s, the inverse Fourier transformation $\vee$ is taken with respect to the variable $k\in\mathbb{R}^3$. Because of (3.11), for each $x\in\mathbb{R}^{3Z}$, $\tilde{f}_\mu(x)\in\mathcal{H}_{-1/2}$. Thus, the multiplication operator $\phi(\tilde{f}_\mu(x))$ is well defined and set

$$A_\mu(f)=A_\mu(f(x))\equiv\phi(\tilde{f}_\mu(x)).$$

Put $A(f)=(A_1(f),\ldots,A_{3Z}(f))$. Let $\hat{\varrho}=(\hat{\varrho}_1,\ldots,\hat{\varrho}_{3Z})$, $\hat{\tau}=(\hat{\tau}_1,\ldots,\hat{\tau}_{3Z})$, and $\hat{\eta}=(\hat{\eta}_1,\ldots,\hat{\eta}_{3Z})$ satisfy that $\overline{\hat{\varrho}_\mu(k)}=\hat{\varrho}_\mu(-k)$, $\overline{\hat{\tau}_\mu(k)}=\hat{\tau}_\mu(-k)$ and $\overline{\hat{\eta}_\mu(k)}=\hat{\eta}_\mu(-k)$, and $\hat{\varrho}_\mu/(\sqrt{\omega})^n$, $\hat{\eta}_\mu/(\sqrt{\omega})^m$, $\hat{\tau}_\mu/(\sqrt{\omega})^l\in L^2(\mathbb{R}^3)$, $n,m=-1,0,1,2$, $l=1,2$, $\mu=1,...,3Z$. We fix $\hat{\varrho}$, $\hat{\tau}$, and $\hat{\eta}$. Define

$$H_{\hat{\varrho},\hat{\tau},\hat{\eta}}=\frac{1}{2m}(\mathbf{p}_Z\otimes I-\gamma A(\hat{\varrho}))^2+\alpha\sum_{\mu=1}^{3Z}A_\mu(\hat{\tau})+\beta\sum_{\mu=1}^{3Z}A_\mu(\hat{\eta})^2+I\otimes H_b,$$

where α, β, and γ are real constants.

Theorem 3.10: *Suppose that the absolute value of coupling constants* γ *and* β *are sufficiently small. Then* $H_{\hat{\varrho},\hat{\tau},\hat{\eta}}$ *is self-adjoint on* $D(H_F^Z)$*, bounded from below and essentially self-adjoint on any core for* H_F^Z*. Moreover we have, for* $V\in\mathcal{M}_\pm(Z)$*,*

$$\langle F,e^{-t(H_{\hat{\varrho},\hat{\tau},\hat{\eta}}\dot{+}(V_+\otimes I)\dot{-}(V_-\otimes I))}G\rangle_{\mathcal{L}_Z}=\int_{\tilde{Q}_E}d\nu_E\,\overline{J_tF(\omega(t/m))}J_0G(\omega(0))e^{\mathcal{K}_{\hat{\varrho},\hat{\tau},\hat{\eta}}(t)}e^{-\int_0^tV(\omega(s))ds}, \tag{3.12}$$

where

$$\mathcal{K}_{\hat{\varrho},\hat{\tau},\hat{\eta}}(t)=\sum_{\mu=1}^{3Z}\left\{i\gamma\Phi\left(\int_0^{t/m}j_{ms}\tilde{\hat{\varrho}}_\mu(\omega(s))\circ db_\mu(s)\right)\right.$$
$$\left.-\alpha\Phi\left(\int_0^t j_s\tilde{\hat{\tau}}_\mu(\omega(s))ds\right)-\beta\int_0^t J_s\phi(\tilde{\hat{\eta}}_\mu(\omega(s)))^2 ds\right\}.$$

Proof: For simplicity, we put $\|\cdot\|_{L^2(\mathbb{R}^3)}=\|\cdot\|_2$ and $H_{\hat{\varrho},\hat{\tau},\hat{\eta}}=H_F^Z+H_R+\alpha H_I+\beta H_{II}$ in this proof. Suppose that $V\in C_b(\mathbb{R}^{3Z})$. Note that, for $\Psi\in\mathcal{F}_0\cap D(H_b^{1/2})$,

$$\|\phi(f)\Psi\|_{\mathcal{F}}\leqslant\frac{1}{\sqrt{2}}\left(2\left\|\frac{\hat{f}}{\omega}\right\|_2\|H_b^{1/2}\Psi\|_{\mathcal{F}}+\left\|\frac{\hat{f}}{\sqrt{\omega}}\right\|_2\|\Psi\|_{\mathcal{F}}\right),$$

$$\|[(H_b+I)^{1/2},\phi(f)]\Psi\|_{\mathcal{F}}\leqslant\frac{1}{\sqrt{2}}\left(\frac{1}{2\pi}\int_0^\infty\frac{\sqrt{\lambda}}{(\lambda+1)^2}\,d\lambda\right)(\|\hat{f}\|_2\|H_b^{1/2}\Psi\|_{\mathcal{F}}+\|\sqrt{\omega}\hat{f}\|_2\|\Psi\|_{\mathcal{F}}).$$

By assumptions $\hat{\tau}_\mu/\omega,\ \hat{\tau}_\mu/\sqrt{\omega}\in L^2(\mathbb{R}^3)$, one sees that H_I is infinitesimally small with respect to $I\otimes H_b$. Moreover, we see that, by the above inequalities, for $\Psi\in D(H_F^Z)$,

$$\|H_R\Psi\|_{\mathcal{L}_Z}\leqslant\sum_{\mu=1}^{3Z}\left\{\gamma^2\left(\left\|\frac{\hat{\varrho}_\mu}{\omega}\right\|_2\sum_{n=-1}^{2}\left\|\frac{\hat{\varrho}_\mu}{(\sqrt{\omega})^n}\right\|_2+\left\|\frac{\hat{\varrho}_\mu}{\sqrt{\omega}}\right\|_2\sum_{n=1}^{2}\left\|\frac{\hat{\varrho}_\mu}{(\sqrt{\omega})^n}\right\|_2\right)+|\gamma|\sum_{n=-1}^{2}\left\|\frac{\hat{\varrho}_\mu}{(\sqrt{\omega})^n}\right\|_2\right\}$$
$$\times\|(H_F^Z+I)\Psi\|_{\mathcal{L}_Z}\times C_1,$$
$$\|H_{II}\Psi\|_{\mathcal{L}_Z}\leqslant\sum_{\mu=1}^{3Z}|\beta|\left(\left\|\frac{\hat{\eta}_\mu}{\omega}\right\|_2\sum_{n=-1}^{2}\left\|\frac{\hat{\eta}_\mu}{(\sqrt{\omega})^n}\right\|_2+\left\|\frac{\hat{\eta}_\mu}{\sqrt{\omega}}\right\|_2\sum_{n=1}^{2}\left\|\frac{\hat{\eta}_\mu}{(\sqrt{\omega})^n}\right\|_2\right)\|(H_F^Z+I)\Psi\|_{\mathcal{L}_Z}\times C_2,$$

where C_1 and C_2 are positive constants. Since γ and β are sufficiently small, it holds that $H_R+\alpha H_I+\beta H_{II}$ is relatively bounded with respect to H_F^Z with relative bound <1. Hence, the Kato–Rellich theorem[13] yields that $H_{\hat{\varrho},\hat{\tau},\hat{\eta}}$ is self-adjoint on $D(H_F^Z)$, bounded from below and essentially self-adjoint on any core for H_F^Z. Next we prove (3.12). We suppose that $\omega\sqrt{\omega}\hat{\varrho}_\mu\in L^2(\mathbb{R}^3)$. Then $\hat{\varrho}_\mu/\sqrt{\omega}$, $\sqrt{\omega}\hat{\varrho}_\mu$, $\omega\sqrt{\omega}\hat{\varrho}_\mu\in L^2(\mathbb{R}^3)$ implies that $A_\mu(\hat{\varrho}(\cdot))\in C_b^2(\mathbb{R}^{3Z};\mathcal{F})$, and $A_\mu(\hat{\varrho}(x))\in D(H_b)$ for $x\in\mathbb{R}^{3Z}$, with $\sup_{x\in\mathbb{R}^{3Z}}\|H_bA_\mu(\hat{\varrho}(x))\|_{\mathcal{F}}<\infty$, $\mu=1,2,\ldots,3Z$. Hence, noting that H_I and H_{II} are regarded as $\mathcal{F}$-valued continuous functions on $\mathbb{R}^{3Z}$, by Theorem 3.5, one sees that

$$\text{the right-hand side of (3.12)}=\langle F,e^{-t(H_0(\gamma A(\hat{\varrho}))\dot{+}(\alpha H_I+\beta H_{II})+V\otimes I)}G\rangle_{\mathcal{L}_Z}.$$

Since $\mathcal{L}_Z^\infty$ is a core for $H_{\hat{\varrho},\hat{\tau},\hat{\eta}}$, by Lemma 3.3, we can see that

$$H_0(\gamma A(\hat{\varrho}))\dot{+}(\alpha H_I+\beta H_{II})+V\otimes I=H_{\hat{\varrho},\hat{\tau},\hat{\eta}}+V\otimes I,$$

as self-adjoint operators in $\mathcal{L}_Z$. Hence, (3.12) follows for such $\hat{\varrho}$'s. Let $\omega\sqrt{\omega}\hat{\varrho}_\mu\notin L^2(\mathbb{R}^3)$. We find sequences $\hat{\varrho}^{(n)}=(\hat{\varrho}_1^{(n)},\ldots,\hat{\varrho}_{3Z}^{(n)})$ so that $\omega\sqrt{\omega}\hat{\varrho}_\mu^{(n)}\in L^2(\mathbb{R}^3)$ and

$$\lim_{n\to\infty}\left\|\frac{\hat{\varrho}_\mu}{(\sqrt{\omega})^m}-\frac{\hat{\varrho}_\mu^{(n)}}{(\sqrt{\omega})^m}\right\|_2=0,\quad \mu=1,2,\ldots,3Z,\quad m=2,1,0,-1.$$

For sufficiently small β and γ, $\mathcal{L}_Z^\infty$ is a common core for $H_{\hat{\varrho}^{(n)},\hat{\tau},\hat{\eta}}+V\otimes I$ and

$$s-\lim_{n\to\infty}(H_{\hat{\varrho}^{(n)},\hat{\tau},\hat{\eta}}+V\otimes I)\Psi=(H_{\hat{\varrho},\hat{\tau},\hat{\eta}}+V\otimes I)\Psi,\quad \Psi\in\mathcal{L}_Z^\infty,$$

which implies that, on $\mathcal{L}_Z$,

$$s-\lim_{n\to\infty} e^{-t(H_{\hat{\varrho}^{(n)},\hat{\tau},\hat{\eta}}+V\otimes I)}=e^{-t(H_{\hat{\varrho},\hat{\tau},\hat{\eta}}+V\otimes I)}. \tag{3.13}$$

Since one sees that

$$\left\|\Phi\left(\int_0^t j_s\tilde{\hat{\varrho}}_\mu(\omega(s))db_\mu(s)-\int_0^t j_s\widetilde{\hat{\varrho}_\mu^{(n)}}(\omega(s))db_\mu(s)\right)\right\|^2_{L^2(\Omega;\mathcal{E})}\leqslant C_3\left\|\frac{\hat{\varrho}_\mu}{\sqrt{\omega}}-\frac{\hat{\varrho}_\mu^{(n)}}{\sqrt{\omega}}\right\|_2^2,$$

$$\left\|\Phi\left(\int_0^t j_s\frac{\partial\tilde{\hat{\varrho}}_\mu}{\partial x_\mu}(\omega(s))ds-\int_0^t j_s\frac{\partial\widetilde{\hat{\varrho}_\mu^{(n)}}}{\partial x_\mu}(\omega(s))ds\right)\right\|^2_{L^2(\Omega;\mathcal{E})}\leqslant C_4\|\sqrt{\omega}\hat{\varrho}_\mu-\sqrt{\omega}\hat{\varrho}_\mu^{(n)}\|_2^2,$$

where C_3 and C_4 are positive constants, we have

$$s-\lim_{n\to\infty}\mathcal{K}_{\hat{\varrho}^{(n)},\hat{\tau},\hat{\eta}}(t)=\mathcal{K}_{\hat{\varrho},\hat{\tau},\hat{\eta}}(t),$$

in $L^2(\Omega;\mathcal{E})$. Putting the right-hand side of (3.12) with $\hat{\rho}$ replaced by $\hat{\varrho}^{(n)}$ by $I(n)$, we show that $I(n)$ converges to the right-hand side of (3.12) as $n\to\infty$. Hence, we get (3.12) with $V\in C_b(\mathbb{R}^{3Z})$. For any $V\in\mathcal{M}_\pm(Z)$, by the same limiting argument as in Theorem 3.9, we get the desired result. □

IV. WCL-RUV FOR THE MODEL

In this section, our task is to define the Hamiltonian $H(V,\Lambda)$ of the system with non-negative mass $\mu\geqslant 0$ and a scaling parameter $\Lambda>0$ in the Schrödinger representation in a rigorous manner. Moreover, we investigate the WCL-RUV for the Hamiltonian $H(V,\Lambda)$. Set

$$\Xi_\Lambda(k)=\begin{cases}1, & |k|<\Lambda,\\ 0, & \text{otherwise.}\end{cases}$$

For each $x\in\mathbb{R}^{3Z}$ and $K>0$, put

$$\tilde{\Xi}_K^Z(x)=\frac{1}{\sqrt{(2\pi)^3}}\sum_{j=1}^Z(\Xi_K(k)e^{-ikx^j})^\vee,$$

$$(\Xi_{\Lambda^\alpha})^j_{K,\mu}(x)=\frac{1}{\sqrt{(2\pi)^3}}\left(\frac{\Xi_{\Lambda^\alpha}(k)(I-\Xi_K(k))e^{-ikx^j}k_\mu}{\omega(k)}\right)^\vee,\quad \mu=1,2,3,j=1,\ldots,Z,$$

where the inverse Fourier transformation $\vee$ is taken with respect to the variable $k\in\mathbb{R}^3$. An operator $H^Z(\Lambda)$, $Z\geqslant 1$, acting in $\mathcal{L}_Z$ is defined by

$$H^Z(\Lambda)=\frac{1}{2m}\mathbf{p}_Z^2\otimes I-\Lambda g H_I^Z(\Lambda^\alpha)+\Lambda^2 I\otimes H_b,$$

where

$$H_I^Z(\Lambda^\alpha)=\int_{\mathbb{R}^{3Z}}^\oplus \phi(\tilde{\Xi}^Z_{\Lambda^\alpha}(x))dx,\quad Z\geqslant 1,\quad \alpha>0.$$

Proposition 4.1 (Ref. 2): For any $\Lambda>0$, the operator $H^Z(\Lambda)$ is self-adjoint on $D(H_F^Z)$ and bounded from below. Moreover, it is essentially self-adjoint on any core for H_F^Z.

Set

$$E(\Lambda^\alpha) = -\frac{1}{2(2\pi)^3}\int_{\mathbb{R}^3}\frac{\Xi_{\Lambda^\alpha}(k)}{\omega^2(k)}\,dk,$$

$$\pi_K = \int_{\mathbb{R}^{3Z}}^{\oplus}\mathcal{T}\left\{\pi_F\left(\frac{1}{\sqrt{(2\pi)^3}}\sum_{j=1}^{Z}\frac{\Xi_{\Lambda^\alpha}(I-\Xi_K)e^{ikx^j}}{\omega^2}\right)\right\}\mathcal{T}^{-1}\,dx.$$

We define a unitary operator $\mathcal{U}_K(\Lambda)$ on $\mathcal{L}_Z$ with an infrared cutoff $K>0$ by

$$\mathcal{U}_K(\Lambda) = \exp\left(-i\,\frac{g}{\Lambda}\,\pi_K\right).$$

Here the symbols π_F and $\mathcal{T}$ are defined in Sec. II.

Lemma 4.2: The unitary operator $\mathcal{U}_K(\Lambda)$ maps $\mathcal{L}_Z^\infty$ to $D(H_F^Z)$ with

$$\mathcal{U}_K(\Lambda)^{-1}(H^Z(\Lambda) - g^2 ZE(\Lambda^\alpha))\mathcal{U}_K(\Lambda)|_{\mathcal{L}_Z^\infty}$$

$$= \frac{1}{2m}\left(\mathbf{p}_Z\otimes I - \frac{g}{\Lambda}\,A_K(\Lambda)\right)^2 + U_K^Z(\Lambda) + \Lambda^2 I\otimes H_b\Big|_{\mathcal{L}_Z^\infty}, \tag{4.1}$$

where

$$A_K(\Lambda) = \left(\int_{\mathbb{R}^{3Z}}^{\oplus}\phi((\Xi_{\Lambda^\alpha})_{K,1}^1(x))dx, \int_{\mathbb{R}^{3Z}}^{\oplus}\phi((\Xi_{\Lambda^\alpha})_{K,2}^1(x))dx,\dots,\int_{\mathbb{R}^{3Z}}^{\oplus}\phi((\Xi_{\Lambda^\alpha})_{K,3}^Z(x))dx\right),$$

$$U_K^Z(\Lambda) = -g\Lambda H_I^Z(K) + g^2 V_K^Z(\Lambda)\otimes I - g^2 ZE(K),$$

$$V_K^Z(\Lambda) = V_K^Z(\Lambda,x) = -\frac{1}{2(2\pi)^3}\sum_{i\neq j}^{Z}\int_{\mathbb{R}^3}\frac{\Xi_{\Lambda^\alpha}(k)(1-\Xi_K(k))e^{-ik(x^i-x^j)}}{\omega^2(k)}\,dk,\quad Z\geqslant 2,\quad V_K^1(\Lambda)=0.$$

Proof: Note that $[[-i(\partial/\partial x_\mu^j)\otimes I,\pi_K],\pi_K]=0$, $\mu=1,2,3$, $j=1,\dots,N$. Then (4.1) follows from a direct calculation. □

Remark 4.3: If $\mu>0$, we can put $K=0$ in π_K. Since $\Xi_{\Lambda^\alpha}/\omega\sqrt{\omega}\notin L^2(\mathbb{R}^3)$, in the case $\mu=0$, we need to introduce the infrared cutoff $K>0$.

Since the right-hand side of (4.1) is closable, we denote the closed extension of the right-hand side of (4.1) by $H^Z(K,\Lambda)$ and

$$H_F^Z(\Lambda) = \frac{1}{2m}\,\mathbf{p}_Z^2\otimes I + \Lambda^2 I\otimes H_b.$$

Define a symmetric operator $R(\Lambda,K)$ by

$$H^Z(K,\Lambda) = H_F^Z(\Lambda) + U_K^Z(\Lambda) + R(\Lambda,K).$$

Let

$$V_\mu^Z(x) = -\frac{1}{4\pi}\sum_{1\leqslant i<j\leqslant Z}\frac{e^{-\mu|x^i-x^j|}}{|x^i-x^j|},\quad V_K^Z(x) = -\frac{1}{2(2\pi)^3}\sum_{i\neq j}^{Z}\int_{\mathbb{R}^3}\frac{\Xi_K(k)e^{-ik(x^i-x^j)}}{\omega^2(k)}\,dk,$$

$$V_K^{Zc}(x) = -\frac{1}{2(2\pi)^3}\sum_{i\neq j}^{Z}\int_{\mathbb{R}^3}\frac{(I-\Xi_K(k))e^{-ik(x^i-x^j)}}{\omega^2(k)}\,dk,\quad Z\geqslant 2,$$

$$V_\mu^1(x) = V_K^1(x) = V_K^{1c}(x) = 0.$$

Note that $V_K^Z + V_K^{Zc} = V_\mu^Z$ and V_K^Z is bounded.

Lemma 4.4: For any $\epsilon>0$, there exist Λ_0 and $b(\epsilon)>0$ so that for all $\Lambda>\Lambda_0$, $D(H_F^Z)\subset D(U_K^Z(\Lambda)+R(\Lambda,K))$, and for $\Psi\in D(H_F^Z)$,

$$\|(U_K^Z(\Lambda)+R(\Lambda,K))\Psi\|_{\mathcal{L}_Z}\leqslant\epsilon\|H_F^Z(\Lambda)\Psi\|_{\mathcal{L}_Z}+b(\epsilon)\|\Psi\|_{\mathcal{L}_Z}.$$

Proof: From Ref. 1, Theorem 3.8(1), it follows that, with constant $b_1(\epsilon)\geqslant 0$ and $b_2(\epsilon)\geqslant 0$,

$$\|(g^2V_K^Z(\Lambda)\otimes I+R(\Lambda,K))\Psi\|_{\mathcal{L}_Z}\leqslant\epsilon\|H_F^Z\Psi\|_{\mathcal{L}_Z}+b_1(\epsilon)\|\Psi\|_{\mathcal{L}_Z},$$

$$\|-g\Lambda H_I^Z(K)\Psi\|_{\mathcal{L}_Z}\leqslant\epsilon\|(I\otimes\Lambda^2H_b)\Psi\|_{\mathcal{L}_Z}+b_2(\epsilon)\|\Psi\|_{\mathcal{L}_Z}.$$

Combining above two inequalities, we get the desired result. □

It follows from Lemma 4.4 that, for sufficiently large $\Lambda>0$ and $0<\alpha<\frac{1}{2}$, $H^Z(K,\Lambda)$ is self-adjoint on $D(H_F^Z)$ and essentially self-adjoint on any core for H_F^Z.

Lemma 4.5: Suppose that $0<\alpha<\frac{1}{2}$ and Λ is sufficiently large. Then $\mathcal{U}_K(\Lambda)$ maps $D(H^Z(K,\Lambda))$ onto $D(H^Z(\Lambda))$, with

$$\mathcal{U}_K(\Lambda)^{-1}e^{-tH^Z(\Lambda)}\mathcal{U}_K(\Lambda)=e^{-tH^Z(K,\Lambda)},\quad t\geqslant 0. \tag{4.2}$$

Proof: Since Λ is sufficiently large, $\mathcal{L}_Z^\infty$ is a core for $H^Z(K,\Lambda)$. Then the equality (4.1) extends to the equality on the domain $D(H^Z(K,\Lambda))$. Then (4.2) follows. □

Lemma 4.6: For sufficiently large Λ, $e^{-tH^Z(K,\Lambda)}$ has a functional integral representation:

$$\langle F,e^{-tH^Z(K,\Lambda)}G\rangle_{\mathcal{L}_Z}=\int_{\tilde{Q}_E}d\nu_E\,\overline{J_{\Lambda^2t}F(\omega(t/m))}J_0G(\omega(0))e^{\mathcal{K}_\Lambda^Z(t)}e^{-E_\Lambda^Z(t)}, \tag{4.3}$$

where

$$\mathcal{K}_\Lambda^Z(t)=g\left\{\frac{i}{\Lambda}\sum_{j=1}^{Z}\sum_{\mu=1}^{3}\Phi\left(\int_0^{t/m}j_{m\Lambda^2s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))\circ db_\mu^j(s)\right)-\Lambda\Phi\left(\int_0^t j_{\Lambda^2s}\tilde{\tilde{\Xi}}_K(\omega(s))ds\right)\right\},$$

$$E_\Lambda^Z(t)=g^2\int_0^t(V_K^Z(\Lambda,\omega(s))-ZE(K))ds.$$

Proof: From Theorem 3.10, (4.3) follows. □

Theorem 4.7: *Let $0<\alpha<\frac{1}{2}$. Then*

$$s-\lim_{\Lambda\to\infty}e^{-t(H^Z(\Lambda)-g^2ZE(\Lambda^\alpha))}=e^{-t((1/2m)\mathbf{p}_Z^2+g^2V_\mu^Z)}\otimes P_b,$$

where P_b denotes the projection operator onto the subspace $\{z\Omega_{\mathcal{F}}|z\in\mathbb{C}\}\subset\mathcal{F}$.

Before proving Theorem 4.7, we show some lemmas. We denote by $\mathcal{S}(\mathbb{R}^m)$ the set of rapidly decreasing infinitely many times differentiable functions on $\mathbb{R}^m$ and define

$$\mathcal{F}^\infty=\{F(\phi(f_1),\dots,\phi(f_m))|F\in\mathcal{S}(\mathbb{R}^m),f_j\in\mathcal{H}_{-1/2},j=1,...,m,m\geqslant 1\}.$$

Lemma 4.8: Let $0<\alpha<\frac{1}{2}$ and $F_l=e^{i\phi(f_l)}$, $f_l\in\mathcal{H}_{-1/2}$, $l=1,...,M$. Suppose that $0\leqslant t_1<t_2<\cdots<t_M\leqslant t$ and put $F_\Lambda^E(t)=e^{i\Sigma_{l=1}^M\Phi(j_{\Lambda^2t_l}f_l)}$. Then for *u* so that

$$\int_\Omega u(b)^2e^{-2\int_0^tg^2V_K^Z(\omega(s))ds}db<\infty,$$

we have

$$\lim_{\Lambda\to\infty}\int_\Omega db\; u(b)\int_{Q_E} d\mu_E\; F_\Lambda^E(t)e^{\mathcal{K}_\Lambda^Z(t)}=\Pi_{l=1}^M\langle\Omega_{\mathcal{F}},F_l\rangle_{\mathcal{F}}\int_\Omega db\; u(b)e^{-\int_0^t g^2 V_K^Z(\omega(s))ds}e^{-g^2 ZE(K)}. \tag{4.4}$$

Proof: For simplicity, we put $m=1$ in this proof. Let $Z\geqslant 2$. The case of $Z=1$ is similarly proved. The integral on the left-hand side (lhs) of (4.4) is calculated as follows:

$$\text{lhs of (4.4)}=\lim_{\Lambda\to\infty}\int_\Omega u(b)\exp\{-\tfrac{1}{4}(I_\Lambda-II_\Lambda+2iIII_\Lambda)\}db,$$

where

$$I_\Lambda=\left\|\sum_{l=1}^M j_{\Lambda^2 t_l}f_l+\sum_{j=1}^Z\sum_{\mu=1}^3\frac{g}{\Lambda}\int_0^t j_{\Lambda^2 s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))\circ db_\mu^j(s)\right\|_{\mathcal{W}}^2,$$

$$II_\Lambda=\left\|\Lambda g\int_0^t j_{\Lambda^2 s}\tilde{\Xi}_K^Z(\omega(s))ds\right\|_{\mathcal{W}}^2,$$

$$III_\Lambda=\left\langle\sum_{l=1}^M j_{\Lambda^2 t_l}f_l+\sum_{j=1}^Z\sum_{\mu=1}^3\frac{g}{\Lambda}\int_0^t j_{\Lambda^2 s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))\circ db_\mu^j(s),-\Lambda g\int_0^t j_{\Lambda^2 s}\tilde{\Xi}_K^Z(\omega(s))ds\right\rangle_{\mathcal{W}}.$$

We shall estimate $I_\Lambda, II_\Lambda, III_\Lambda$ separately. We see that

$$\begin{aligned}I_\Lambda&=\left\|\sum_{l=1}^M j_{\Lambda^2 t_l}f_l\right\|_{\mathcal{W}}^2+\left\|\sum_{j=1}^Z\sum_{\mu=1}^3\frac{g}{\Lambda}\int_0^t j_{\Lambda^2 s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))\circ db_\mu^j(s)\right\|_{\mathcal{W}}^2\\&\quad+2\Re\left\langle\sum_{l=1}^M j_{\Lambda^2 t_l}f_l,\sum_{j=1}^Z\sum_{\mu=1}^3\frac{g}{\Lambda}\int_0^t j_{\Lambda^2 s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))\circ db_\mu^j(s)\right\rangle_{\mathcal{W}}^2\\&=I_\Lambda^{(1)}+I_\Lambda^{(2)}+2\Re I_\Lambda^{(3)}.\end{aligned}$$

Then one can see that, by the definition of the operator j_t,

$$\lim_{\Lambda\to\infty}I_\Lambda^{(1)}=\lim_{\Lambda\to\infty}\sum_{i,j=1}^M\langle f_i,e^{-\Lambda^2|t_i-t_j|\omega}f_j\rangle_{-1/2}=\sum_{i=1}^M\|f_i\|_{-1/2}^2. \tag{4.5}$$

Moreover, we have

$$\|j_{\Lambda^2 s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))\|_{\mathcal{W}}^2\leqslant\frac{\mathbf{S}^2}{(2\pi)^3}\int_K^{\Lambda^\alpha}r\,dr,$$

$$\left\|j_{\Lambda^2 s}\frac{\partial(\Xi_{\Lambda^\alpha})_{K,\mu}^j}{\partial x_\mu^j}(\omega(s))\right\|_{\mathcal{W}}^2\leqslant\frac{\mathbf{S}^2}{(2\pi)^3}\int_K^{\Lambda^\alpha}r^3\,dr,$$

where $\mathbf{S}^2$ denotes the volume of the two-dimensional sphere. Since $0<\alpha<\frac{1}{2}$, it holds that

$$\begin{aligned}&\lim_{\Lambda\to\infty}\int_\Omega d\mu\left\|\sum_{j=1}^Z\sum_{\mu=1}^3\frac{g}{\Lambda}\int_0^t j_{\Lambda^2 s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))db_\mu^j(s)\right\|_{\mathcal{W}}^2\\&\quad=\lim_{\Lambda\to\infty}\int_\Omega d\mu\sum_{j=1}^Z\sum_{\mu=1}^3\frac{g^2}{\Lambda^2}\int_0^t ds\|j_{\Lambda^2 s}(\Xi_{\Lambda^\alpha})_{K,\mu}^j(\omega(s))\|_{\mathcal{W}}^2\leqslant\lim_{\Lambda\to\infty}\frac{3Ztg^2\mathbf{S}^2}{(2\pi)^3}\frac{1}{\Lambda^2}\int_K^{\Lambda^\alpha}r\,dr=0,\end{aligned}$$

moreover

$$\lim_{\Lambda\to\infty}\left\|\sum_{j=1}^{Z}\sum_{\mu=1}^{3}\frac{g}{\Lambda}\int_0^t j_{\Lambda^2 s}\frac{\partial(\Xi_{\Lambda^\alpha})_{K,\mu}^j}{\partial x_\mu^j}(\omega(s))ds\right\|_{\mathcal{W}}^2\leqslant\lim_{\Lambda\to\infty}\frac{3Ztg^2\mathbf{S}^2}{(2\pi)^3}\frac{1}{\Lambda^2}\int_K^{\Lambda^\alpha}r^3\,dr=0.$$

Hence we have

$$\lim_{\Lambda\to\infty}\int_\Omega|e^{-(1/4)I_\Lambda^{(2)}}-1|^2db=0. \tag{4.6}$$

By (4.5) and (4.6), we have

$$\lim_{\Lambda\to\infty}\int_\Omega|e^{-(1/4)I_\Lambda^{(3)}}-I|^2db=0.$$

Consequently,

$$\lim_{\Lambda\to\infty}\int_\Omega e^{-(1/4)I_\Lambda}db=\int_\Omega e^{-(1/4)\Sigma_{t=1}^M\|f_l\|_{-1/2}^2}db. \tag{4.7}$$

Next, we consider II_Λ, which is essential to derive a Z-body Coulomb (or Yukawa) potential. Note that $\widehat{j_t f}(k,k_0)=(1/\sqrt{2\pi})e^{-ik_0t}\hat{f}(k)$. Then

$$\begin{aligned}II_\Lambda&=\Lambda^2\frac{2g^2}{(2\pi)^3}\int_{\mathbb{R}^3\times\mathbb{R}}\frac{dk\,dk_0}{\omega^2(k)+k_0^2}\left|\frac{1}{\sqrt{2\pi}}\int_0^t e^{-ik_0\Lambda^2 s}\sum_{j=1}^{Z}\Xi_K(k)e^{-ik\omega^j(s)}\,ds\right|^2\\&=\frac{2g^2}{(2\pi)^3}\int_{\mathbb{R}^3}\Xi_K(k)dk\int_{\mathbb{R}}\frac{dk_0}{\omega^2(k)+\dfrac{k_0^2}{\Lambda^4}}\left|\frac{1}{\sqrt{2\pi}}\int_0^t e^{-ik_0 s}\sum_{j=1}^{Z}e^{-ik\omega^j(s)}\,ds\right|^2.\end{aligned}$$

By the Lebesgue dominated convergence theorem, one sees that

$$\begin{aligned}\lim_{\Lambda\to\infty}II_\Lambda&=\frac{2g^2}{(2\pi)^3}\int_{\mathbb{R}^3}\frac{\Xi_K(k)dk}{\omega^2(k)}\int_{\mathbb{R}}dk_0\left|\frac{1}{\sqrt{2\pi}}\int_0^t e^{-ik_0 s}\sum_{j=1}^{Z}e^{-ik\omega^j(s)}\,ds\right|^2\\&=\frac{2g^2}{(2\pi)^3}\int_{\mathbb{R}^3}\frac{\Xi_K(k)dk}{\omega^2(k)}\int_0^t\left|\sum_{j=1}^{Z}e^{-ik\omega^j(s)}\right|^2ds\\&=4\left\{\int_0^t ds\,\frac{g^2}{2(2\pi)^3}\sum_{i\neq j}^{Z}\int_{\mathbb{R}^3}\frac{\Xi_K(k)e^{-ik(\omega^i(s)-\omega^j(s))}}{\omega^2(k)}\,dk-g^2Z\int_0^t ds\,E(K)\right\}\\&=-4\int_0^t(g^2V_K^Z(\omega(s))+g^2ZE(K))ds.\end{aligned} \tag{4.8}$$

Here, in the second equality in (4.8) we use that the Fourier transformation with respect to k_0 is unitary on $L^2(\mathbb{R}^3)$. Note that II_Λ is monotonously increasing as $\Lambda\to\infty$. Finally, we consider III_Λ. From (4.6) and (4.8), it follows that

$$\lim_{\Lambda\to\infty} III_\Lambda = \lim_{\Lambda\to\infty}\left\langle \sum_{l=1}^{M} j_{\Lambda^2 t_l} f_l, -\Lambda g\int_0^t j_{\Lambda^2 s}\tilde{\Xi}_K(\omega(s))ds\right\rangle_{\mathcal{W}}$$

$$= \lim_{\Lambda\to\infty}\frac{-2\Lambda g}{\sqrt{(2\pi)^3}}\int_{\mathbb{R}^3\times\mathbb{R}}\frac{dk\ dk_0}{\sqrt{2\pi}}\sum_{l=1}^{M}\frac{e^{ik_0\Lambda^2 t_l}\overline{\hat{f}_l}(k)}{\omega^2(k)+k_0^2}\int_0^t\frac{ds}{\sqrt{2\pi}}e^{-i\Lambda^2 sk_0}\Xi_K(k)\sum_{j=1}^{Z}e^{-ik\omega^j(s)}$$

$$= \lim_{\Lambda\to\infty}\frac{-1}{\sqrt{(2\pi)^3}}\frac{g}{\Lambda}\int_{\mathbb{R}^3}dk\sum_{l=1}^{M}\frac{e^{-\Lambda^2|t_l-s|\omega(k)}\overline{\hat{f}_l}(k)\Xi_K(k)}{\omega(k)}\int_0^t\sum_{j=1}^{Z}e^{-ik\omega^j(s)}\ ds.$$

Hence, by the Schwartz inequality,

$$\lim_{\Lambda\to\infty}|III_\Lambda|\leqslant\lim_{\Lambda\to\infty}\frac{Zt}{\sqrt{(2\pi)^3}}\sum_{l=1}^{M}\|f_l\|_{-1/2}\frac{g}{\Lambda}\left(\int_{\mathbb{R}^3}e^{-2\Lambda^2|t_l-s|\omega(k)}\frac{\Xi_K(k)}{\omega(k)^2}dk\right)^{1/2}=0. \qquad (4.9)$$

Then by (4.7), (4.8), and (4.9) and the assumption of u, it holds that

$$\lim_{\Lambda\to\infty}\int_\Omega u(b)e^{-(1/4)(I_\Lambda-II_\Lambda+2iIII_\Lambda)}db=\int_\Omega u(b)e^{-\int_0^t g^2V_K^Z(\omega(s))ds}e^{-g^2ZE(K)}e^{-(1/4)\Sigma_{l=1}^M\|f_l\|_{-1/2}^2}db.$$

Thus, the proof is complete. □

Lemma 4.9: Let $0<\alpha<\frac{1}{2}$ *and* u *as in Lemma 4.8. Then, for* $F,G\in\mathcal{F}$, *we have*

$$\lim_{\Lambda\to\infty}\int_\Omega db\ u(b)\int_{Q_E}d\mu_E(\overline{J_{\Lambda^2 t}F})(J_0G)e^{\mathcal{K}_\Lambda(t)}=\langle F,P_bG\rangle_{\mathcal{F}}\int_\Omega db\ u(b)e^{-\int_0^t g^2V_K^Z(\omega(s))ds}e^{-g^2ZE(K)}. \qquad (4.10)$$

Proof: First we assume that $F,G\in\mathcal{F}^\infty$ so that, with $f\in\mathcal{S}(\mathbb{R}^n)$, $g\in\mathcal{S}(\mathbb{R}^m)$,

$$F=\frac{1}{\sqrt{(2\pi)^n}}\int_{\mathbb{R}^n}\check{f}(t_1,\ldots,t_n)e^{i\Sigma_{j=1}^n\phi(f_j)t_j}\ dt_1\cdots dt_n,$$

$$G=\frac{1}{\sqrt{(2\pi)^m}}\int_{\mathbb{R}^m}\check{g}(s_1,\ldots,s_m)e^{i\Sigma_{j=1}^m\phi(g_j)s_j}\ ds_1\cdots ds_m.$$

Then from Lemma 4.8 (4.10) follows. Since $\mathcal{F}^\infty$ is dense in $\mathcal{F}$, one sees (4.10) for any $F,G\in\mathcal{F}$ by an approximation argument. □

Proof of Theorem 4.7: Let $Z\geqslant 2$. The case of $Z=1$ is similarly proved. It is seen that

$$s-\lim_{\Lambda\to\infty}\mathcal{U}_K(\Lambda)=I.$$

Hence, by Lemma 4.5, and uniform boundedness of $e^{-tH^Z(K,\Lambda)}$ and $e^{-t((1/2m)\mathbf{p}_Z^2+g^2V_\mu^Z)}\otimes P_b$, it is enough to show that for any Θ, $\Psi\in C_0^\infty(\mathbb{R}^{3Z})\hat{\otimes}\mathcal{F}^\infty$,

$$\lim_{\Lambda\to\infty}\langle\Theta,e^{-tH^Z(K,\Lambda)}\Psi\rangle_{\mathcal{L}_Z}=\langle\Theta,e^{-t((1/2m)\mathbf{p}_Z^2+g^2V_\mu^Z)}\otimes P_b\Psi\rangle_{\mathcal{L}_Z}.$$

Let $\Theta=u\otimes F$, $\Psi=v\otimes G$, where F and G have the same forms as those in the proof of Lemma 4.9 and $u,v\in C_0^\infty(\mathbb{R}^{3Z})$. Then

$$\langle\Theta,e^{-tH^Z(K,\Lambda)}\Psi\rangle_{\mathcal{L}_Z}=\int_{\mathbb{R}^{3Z}\times\Omega}dx\ db\ \overline{u(\omega(t/m))}v(\omega(0))e^{-E_\Lambda^Z(t)}\int_{Q_E}d\mu_E(\overline{J_{\Lambda^2 t}F})(J_0G)e^{\mathcal{K}_\Lambda^Z(t)}.$$

First we estimate $E_\Lambda^Z(t)$. Since

$$V_K^Z(\Lambda,x^1,\ldots,x^Z)=-\frac{1}{2\pi^2}\sum_{1\leqslant i<j\leqslant Z}\frac{1}{|x^i-x^j|}\int_K^{\Lambda^\alpha}\frac{r}{r^2+\mu^2}\sin(r|x^i-x^j|)dr,$$

using a contour integral, one can check that there exist $\delta_1>0$ and $\delta_2>0$, which are independent of $\Lambda>0$, so that

$$|V_K^Z(\Lambda,\omega(s))|\leqslant\delta_1|V_\mu^Z(\omega(s))|+\delta_2,\quad \mu\geqslant 0. \tag{4.11}$$

Set

$$\Omega_0=\left\{(x,b)\in\mathbb{R}^{3Z}\times\Omega\;\middle|\;\int_0^t\frac{e^{-\mu|\omega^i(s)-\omega^j(s)|}}{|\omega^i(s)-\omega^j(s)|}ds=\infty,\quad i\neq j,i,j=1,\ldots,Z\right\}.$$

The measure of Ω_0 is zero. One can check that for $(x,b)\in[\mathbb{R}^{3Z}\times\Omega]\backslash\Omega_0$,

$$\lim_{\Lambda\to\infty}V_K^Z(\Lambda,\omega(s))=V_K^{Zc}(\omega(s))<\infty,\quad a.e.s\in[0,t]. \tag{4.12}$$

Combining (4.11) and (4.12), one can see that, for $(x,b)\in[\mathbb{R}^{3Z}\times\Omega]\backslash\Omega_0$, by the Lebesgue dominated convergence theorem,

$$\lim_{\Lambda\to\infty}\int_0^t ds\ V_K^Z(\Lambda,\omega(s))=\int_0^t ds\ V_K^{Zc}(\omega(s)).$$

Hence, for almost everywhere $(x,b)\in\mathbb{R}^{3Z}\times\Omega$,

$$\lim_{\Lambda\to\infty}e^{-E_\Lambda^Z(t)}=e^{-\int_0^t g^2V_K^{Zc}(\omega(s))ds}e^{g^2ZE(K)}.$$

On the other hand, under the notation in the proof of Lemma 4.8, we see that

$$|e^{\mathcal{K}_\Lambda^Z(t)}|\leqslant e^{(1/4)II_\Lambda}. \tag{4.13}$$

As is seen in the proof of Lemma 4.8, the right-hand side of (4.13) is monotonously increasing as $\Lambda\to\infty$. Hence we have, from the definition of F and G,

$$\left|\int_{Q_E}d\mu_E(\overline{J_{\Lambda^2 t}F})(J_0G)e^{\mathcal{K}_\Lambda^Z(t)}\right|\leqslant\left|\int_{\mathbb{R}^{n+m}}\overline{\check{f}(t)}\check{g}(s)dt\ ds\right|e^{-\int_0^t g^2V_K^Z(\omega(s))ds}e^{-g^2ZE(K)}.$$

Then, by (4.11), we have for almost everywhere $(x,b)\in\mathbb{R}^{3Z}\times\Omega$,

$$\left|\overline{u(\omega(t/m))}v(\omega(0))e^{-E_\Lambda^Z(t)}\int_{Q_E}d\mu_E(\overline{J_{\Lambda^2 t}F})(J_0G)e^{\mathcal{K}_\Lambda^Z(t)}\right|$$
$$\leqslant|\overline{u(\omega(t/m))}v(\omega(0))|C_{fg}e^{-\int_0^t(g^2V_K^Z(\omega(s))-\delta_1|V_\mu^Z(\omega(s))|-\delta_2)ds}, \tag{4.14}$$

where $C_{fg}=|\int_{\mathbb{R}^{n+m}}\overline{\check{f}(t)}\check{g}(s)dt\ ds|$. The right-hand side of (4.14) is integrable. By (4.10) and the Lebesgue dominated convergence theorem, we have

$$\lim_{\Lambda\to\infty}\int_{\mathbb{R}^{3Z}\times\Omega}dx\ db\ \overline{u(\omega(t/m))}v(\omega(0))|e^{-E_\Lambda^Z(t)}-e^{-\int_0^t g^2V_K^{Zc}(\omega(s))ds}e^{g^2ZE(K)}|$$
$$\times\int_{Q_E}d\mu_E(\overline{J_{\Lambda^2 t}F})(J_0G)e^{K_\Lambda^Z(t)}=0.$$

While, by Lemma 4.9, we have

$$\lim_{\Lambda\to\infty}\int_{\mathbb{R}^{3Z}\times\Omega}dx\ db\,\overline{u(\omega(t/m))}v(\omega(0))e^{-\int_0^t g^2V_K^{Zc}(\omega(s))ds}e^{g^2ZE(K)}\int_{Q_E}d\mu_E(\overline{J_{\Lambda^2 t}F})(J_0G)e^{K_\Lambda^Z(t)}$$

$$=\int_{\mathbb{R}^{3Z}\times\Omega}dx\ db\overline{u(\omega(t/m))}v(\omega(0))e^{-\int_0^t g^2V_\mu^Z(\omega(s))ds}\langle F,P_bG\rangle_{\mathcal{F}}.$$

Thus

$$\lim_{\Lambda\to\infty}\langle\Theta,e^{-tH^Z(K,\Lambda)}\Psi\rangle_{\mathcal{L}_Z}=\langle u,e^{-t((1/2m)\mathbf{p}_Z^2+g^2V_\mu^Z)}v\rangle_{L^2(\mathbb{R}^{3Z})}\langle F,P_bG\rangle_{\mathcal{F}}.$$

The proof is complete. □

Next, we study the interaction system consisting of the arbitrary but conserved number of particles and the quantized scalar field.

Lemma 4.10: Suppose that $V_Z\in\mathcal{M}_\pm(Z)_r$, $Z\geqslant 1$. *Then* $H^Z(\Lambda)\dot{+}(V_{Z+}\otimes I)\dot{-}(V_{Z-}\otimes I)$ *is reduced by* $\mathcal{H}_Z$.

Proof: Let S be the projection operator from $L^2(\mathbb{R}^{3Z})$ onto $L^2_{as}(\mathbb{R}^{3Z})$. It is seen that

$$H^Z(\Lambda)(S\otimes I)\Psi=(S\otimes I)H^Z(\Lambda)\Psi,\tag{4.15}$$

where $\Psi\in C_0^\infty(\mathbb{R}^{3Z})\hat{\otimes}\mathcal{F}_0$. Since $C_0^\infty(\mathbb{R}^{3Z})\hat{\otimes}\mathcal{F}_0$ is a core for $H^Z(\Lambda)$, (4.15) also holds on $D(H^Z(\Lambda))$. Thus $H^Z(\Lambda)$ is reduced by $\mathcal{H}_Z$. Hence $H^Z(\Lambda)\dot{+}(V_{Z+}\otimes I)\dot{-}(V_{Z-}\otimes I)$ is also reduced by $\mathcal{H}_Z$. □

For the set of potentials $V=\{V_Z\}_{Z=1}^\infty$ with $V_Z\in\mathcal{M}_\pm(Z)_r$, $Z\geqslant 1$, a scaling Hamiltonian $H(V,\Lambda)$ of the system and a self-adjoint operator H_{∞,V,V_μ} in $\otimes_{Z=1}^\infty\mathcal{H}_Z$ are defined by

$$H(V,\Lambda)\equiv\bigoplus_{Z=1}^\infty[H^Z(\Lambda)\dot{+}(V_{Z+}\otimes I)\dot{-}(V_{Z-}\otimes I)|_{\mathcal{H}_Z}],$$

$$H_{\infty,V,V_\mu}\equiv\bigoplus_{Z=1}^\infty\left[\frac{1}{2m}\mathbf{p}_Z^2\dot{+}V_{Z+}\dot{-}V_{Z-}+g^2V_\mu^Z\Big|_{L^2_{as}(\mathbb{R}^{3Z})}\right].\tag{4.16}$$

Let N_a be the number operator in $\mathcal{F}_a$, i.e.,

$$(N_a\Psi)_Z=Z\Psi_Z,\quad D(N_a)=\left\{\Psi=\{\Psi_Z\}_{Z=0}^\infty\,\middle|\,\sum_{Z=0}^\infty Z^2\|\Psi_Z\|^2_{L_{as}(\mathbb{R}^{3Z})}<\infty\right\}.$$

Since $-g^2E(\Lambda^\alpha)N_a\otimes I$ is a non-negative self-adjoint operator and reduced by $\otimes_{Z=1}^\infty L^2_{as}(\mathbf{R}^{3Z})\equiv\mathcal{F}_p$, $H(V,\Lambda)\dot{+}(-g^2E(\Lambda^\alpha)N_a|_{\mathcal{F}_p}\otimes I)$ is well defined. We state the main theorem in this paper.

Theorem 4.11: *Let* $0<\alpha<\frac{1}{2}$ *and* $V=\{V_Z\}_{Z=1}^\infty$ *with* $V_Z\in\mathcal{M}_\pm(Z)_r$, $Z\geqslant 1$. *Then*

$$s-\lim_{\Lambda\to\infty}e^{-t(H(V,\Lambda)\dot{+}(-g^2E(\Lambda^\alpha)N_a|_{\mathcal{F}_p}\otimes I))}=e^{-tH_{\infty,V,V_\mu}}\otimes P_b.$$

Proof: For $\Psi=(\Psi_0,\Psi_1,\ldots.)\in\mathcal{L}$, it follows that, by Theorem 4.7 and the Lebesgue dominated convergence theorem,

$$s-\lim_{\Lambda\to\infty}e^{-t(H(V,\Lambda)\dot{+}(-g^2E(\Lambda^\alpha)N_a|_{\mathcal{F}_p}\otimes I))}\Psi=s-\lim_{\Lambda\to\infty}\bigoplus_{Z=1}^\infty e^{-t(H^Z(\Lambda)\dot{+}(V_{Z+}\otimes I)\dot{-}(V_{Z-}\otimes I)-g^2ZE(\Lambda^\alpha))}\Psi_Z$$

$$=\bigoplus_{Z=1}^\infty[e^{-t((1/2m)\mathbf{p}_Z^2\dot{+}V_{Z+}\dot{-}V_{Z-}+g^2V_\mu^Z)}\otimes P_b]\Psi_Z$$

$$=e^{-tH_{\infty,V,V_\mu}}\otimes P_b\Psi_Z.$$

Thus, the proof is complete. □

TABLE I.

	$V_\mu^Z(Z\geqslant 2)$	$E(\Lambda^\alpha)$	μ	α
$d=3$	$-\frac{1}{4\pi}\Sigma_{1\leqslant i<j\leqslant Z}\frac{e^{-\mu\lvert x^i-x^j\rvert}}{\lvert x^i-x^j\rvert}$	$-\frac{1}{2(2\pi)^3}\int_{\mathbb{R}^3}\frac{\Xi_\Lambda{}^\alpha(k)}{k^2+\mu^2}\,dk$	$\mu\geqslant 0$	$0<\alpha<\frac{1}{2}$
$d=2$	$-\frac{1}{4\pi}\Sigma_{1\leqslant i\leqslant j\leqslant Z}\int_0^\infty\frac{rJ_0(r\lvert x^i-x^j\rvert)}{r^2+\mu^2}\,dr$	$-\frac{1}{2(2\pi)^2}\int_{\mathbb{R}^2}\frac{\Xi_\Lambda{}^\alpha(k)}{k^2+\mu^2}\,dk$	$\mu>0$	$0<\alpha<\frac{2}{3}$
$d=1$	$-\frac{1}{4}\Sigma_{i,j=1}^Z\frac{e^{-\mu\lvert x^i-x^j\rvert}}{\mu}$	0	$\mu>0$	$0<\alpha<1$

V. CONCLUDING REMARKS

(1) Let $V_Z=0$, $Z\geqslant 1$, in Theorem 4.11. Formally, we may write

$$H_{\infty,0,V_\mu}=\int\Psi^\dagger(x)\left(-\frac{1}{2m}\Delta\right)\Psi(x)dx-\frac{g^2}{4\pi}\int\Psi^\dagger(x)\Psi^\dagger(y)\frac{e^{-\mu\lvert x-y\rvert}}{\lvert x-y\rvert}\Psi(y)\Psi(x)dx\,dy.$$

(2) We can also investigate the WCL-RUV for models in the space dimension $d=1,2$. Without proofs, we only show results in the space dimension $d=1,2$ in Table I.
Here J_0 is the Bessel function: $J_0(x)=\Sigma_{n=0}^\infty[(-1)^n/n!\Gamma(n+1)](x/2)^{2n}$. Since, in the case where $d=1,2$, each potential V_μ^Z does not converge as $\mu\to 0$, we cannot expect to get their WCL-RUV with $\mu=0$.

ACKNOWLEDGMENTS

The author is grateful to Professor Asao Arai for useful discussions. This work is supported in part by the Japan Society for the Promotion of Science (JSPS).

[1] F. Hiroshima, ''Weak coupling limit with a removal of an ultraviolet cut-off for a Hamiltonian of particles interacting with a massive scalar field,'' Inf. Dimen. Anal. Quantum Prob. Relat. Top. **1**, 407–423 (1998).
[2] E. Nelson, ''Interaction of nonrelativistic particles with a quantized scalar field,'' J. Math. Phys. **5**, 1190–1197 (1964).
[3] F. Hiroshima, ''Scaling limit of a model of quantum electrodynamics,'' J. Math. Phys. **34**, 4478–4518 (1993).
[4] F. Hiroshima, ''Scaling limit of a model of quantum electrodynamics with many nonrelativistic particles,'' Rev. Math. Phys. **9**, 201–225 (1997).
[5] F. Hiroshima, ''Asymptotic behaviors of an interaction Hamiltonian,'' Nonlinear Anal. Theory Meth. Appl. **30**, 4863–4874 (1997).
[6] A. Arai, ''Scaling limit for quantum systems of nonrelativistic particles interacting with a bose field,'' Hokkaido University preprint Series #59.
[7] E. B. Davies, ''Particle–boson interaction and the weak coupling limit,'' J. Math. Phys. **20**, 345–351 (1979).
[8] F. Hiroshima, ''Functional integral representation of a model in quantum electrodynamics,'' Rev. Math. Phys. **9**, 489–530 (1997).
[9] J. Glimm and A. Jaffe, *Quantum Physics* (Springer-Verlag, New York, 1981).
[10] B. Simon, *The $P(\phi)_2$ Euclidean (Quantum) Field Theory* (Princeton University Press, Princeton, NJ, 1974).
[11] F. Hiroshima, ''Diamagnetic inequalities for systems of nonrelativistic particles with a quantized field,'' Rev. Math. Phys. **8**, 185–203 (1996).
[12] T. Kato and K. Masuda, ''Trotter's product formula for nonlinear semigroups generated by the subdifferentiables of convex functionals,'' J. Math. Soc. Jpn. **30**, 169–178 (1978).
[13] M. Reed and B. Simon, *Method of Modern Mathematical Physics II; Fourier Analysis, Self-Adjointness* (Academic, New York, 1978).
[14] B. Simon, *Functional Integral and Quantum Physics* (Academic, New York, 1979).

Uncertainty principle for proper time and mass

Shoju Kudaka
Department of Physics, University of the Ryukyus, Okinawa, Japan

Shuichi Matsumoto[a)]
Department of Mathematics, University of the Ryukyus, Okinawa, Japan

(Received 6 July 1998; accepted for publication 19 November 1998)

We review Bohr's reasoning in the Bohr–Einstein debate on the photon box experiment. The essential point of his reasoning leads us to an uncertainty relation between the proper time and the rest mass of the clock. It is shown that this uncertainty relation can be derived if only we take the fundamental point of view that the proper time should be included as a dynamic variable in the Lagrangian describing the system of the clock. Some problems and some positive aspects of our approach are then discussed. © *1999 American Institute of Physics.*
[S0022-2488(99)02103-9]

I. INTRODUCTION

In various arguments about time, perhaps the most spectacular is the Einstein–Bohr debate on the photon box experiment.[1,2] Their concern in the debate was Heisenberg's time–energy uncertainty relation. However, Bohr's reasoning reveals, as shown in the following, an uncertainty relation between the proper time and the rest mass of a clock. In fact, his essential point was simply that the very act of weighing a clock, according to general relativity, interferes with the rate of the clock.

In order to review Bohr's reasoning, we consider an experiment in which we measure the rest mass of a clock. We assume, of course, that the clock keeps its own proper time.

Following Einstein's stratagem, we try to weigh the clock by suspending it with a spring. That is to say, if the spring stretches by the length l, we can calculate the mass m of the clock from the relation

$$kl = mg,$$

where g is the gravitational acceleration and k is a constant characterizing the spring.

Assume that a scale is fixed to the spring support, and that we read the length l on it with an accuracy Δq. Then the determination of l involves a minimum latitude Δp in the momentum of the clock, related to Δq by the equation $\Delta q\, \Delta p \approx h$. Let t be the time interval in which we read the length l. (We should note that t is measured by a clock other than the suspended clock.) Then we cannot determine the force exerted by the gravitational field on the clock to a finer accuracy than $\Delta p/t$. Therefore we cannot determine the mass m to a finer accuracy than Δm given by the relation

$$\frac{\Delta p}{t} \approx g\, \Delta m. \qquad (1)$$

Now, according to general relativity theory, a clock, when displaced in the direction of the gravitational force by an amount Δq, changes its rate in such a way that its reading in the course of a time interval t differs by an amount $\Delta \tau$ given by the relation

[a)]Electronic mail: shuichi@edu.u-ryukyu.ac.jp

$$\frac{\Delta\tau}{t}=\frac{g\,\Delta q}{c^2}. \tag{2}$$

By combining (1), (2) and the relation $\Delta q\,\Delta p \approx h$, we see, therefore, that there is an uncertainty relation,

$$c^2\,\Delta m\,\Delta\tau \approx h \tag{3}$$

between the rest mass m and the proper time τ of the clock.

The relativistic redshift formula (2) was, of course, essential in Bohr's reasoning above. The more essential it seems to be, however, the stronger the apprehension we feel that the uncertainty relation (3) may fail if we can think of a weighing procedure not resorting to any interaction between the clock and the gravitational field. We check one such case in the following.

Assume that the clock has been brought to rest after being charged with an electric charge e, and that a uniform electric field $\mathcal{E}$ is then switched on. After a short time t, we measure the distance the clock has moved. (Again t is the time measured by a clock other than our clock in the electric field.) Then we can know the average velocity v of the clock by dividing the distance by the value of t, and we can determine the mass m of the clock by virtue of the formula

$$e\mathcal{E}=m\frac{v}{t}.$$

Assume that the determination of the distance is made with a given accuracy Δq. Then it implies a minimum latitude Δp in the momentum of the clock, where $\Delta q\,\Delta p \approx h$. Hence, we cannot determine the force exerted by the electric field on the clock to a finer accuracy than $\Delta p/t$. Therefore, even when the velocity v is obtained, we cannot determine the mass m to a finer accuracy than Δm given by the relation

$$\frac{\Delta p}{t}\approx \Delta m\frac{v}{t}, \quad \text{i.e.}\ \ \Delta p \approx v\,\Delta m. \tag{4}$$

Now, according to special relativity theory, when a clock has a speed v, its rate τ in the course of a time interval t is given by the relation

$$\tau = t\sqrt{1-(v/c)^2}. \tag{5}$$

On the other hand, the average velocity v has an uncertainty Δv given by the relation

$$t\,\Delta v \approx \Delta q.$$

Correspondingly, the clock has an uncertainty in its rate τ of the order $\Delta\tau$ given by

$$\Delta\tau = t\cdot\Delta\sqrt{1-(v/c)^2}\approx \frac{v}{c^2}t\Delta v \approx \frac{v}{c^2}\Delta q. \tag{6}$$

By combining (4), (6), and the relation $\Delta q\,\Delta p \approx h$, we arrive, therefore, at the same uncertainty relation,

$$c^2\,\Delta m\,\Delta\tau \approx h,$$

as (3) obtained by Bohr's reasoning.

Thus, the uncertainty relation (3) has been confirmed for a weighing procedure that does not rely on any gravitational interaction. Moreover, in this case, the time-shift formula (5) played an essential role in place of the relativistic redshift formula.

Each of these formulas is, of course, one of the deepest and most important results in relativistic theory. The fact that these important formulas play essential roles in deriving the uncertainty relation (3) lends some confidence as to its universality.

Our objective in this article is to show the following: The uncertainty relation (3) can be derived satisfactorily only if we describe the system of the clock by using a Lagrangian that includes the proper time as a dynamic variable.

In the next section, selecting the simplest Lagrangian that is in accord with the above approach, we examine the Hamiltonian formalism of the clock. Our conclusion is that the rest energy can be considered the momentum conjugate to the proper time. In the third section, following Dirac's procedure, we quantize the system of the clock, and we obtain the same uncertainty relation as (3). Some comments then follow on our quantization.

II. LAGRANGIAN AND HAMILTONIAN FORMALISM

A gravitational field $g_{\mu\nu}$ and an electromagnetic field A_μ are assumed to be given, and we consider our clock to be one material particle moving in those fields with electric charge e.

The Lagrangian that is generally used in such a case is the following:

$$L_0 = -mc\sqrt{-g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu} + eA_\mu(x)\dot{x}^\mu,$$

where x^μ ($\mu = 0,1,2,3$) are the variables and the overdot denotes the differential with respect to an arbitrary parameter λ. It goes without saying that m is the rest mass of the clock and that c is the speed of light.

We, however, cannot consider the proper time τ a physical quantity if we describe the system by using the Lagrangian L_0. On the other hand, it is clear that the proper time of a clock is a measurable physical quantity. (It is why a clock is so named.) Hence, we have to find another Lagrangian that is in accord with the system of the clock.

Our first purpose in this section is to find a Lagrangian L that satisfies the following conditions: (1) The Lagrangian L has the proper time τ as a new variable, in addition to x^μ. (2) The motion equations for the variables x^μ are invariant between L and L_0.

As a candidate, we consider the Lagrangian defined by

$$L = M(\dot{\tau} - \sqrt{-g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu}/c) + eA_\mu(x)\dot{x}^\mu,$$

where the dynamic variables are τ, M, and x^μ.

The Lagrange's equations of motion are as follows:

$$\dot{M} = 0, \tag{7}$$

$$\dot{\tau} = \sqrt{-g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu}/c, \tag{8}$$

$$\frac{d}{d\lambda}\left[\frac{M}{c}\frac{g_{\rho\mu}\dot{x}^\mu}{\sqrt{-g_{\mu\nu}(x)\dot{x}^\mu \dot{x}^\nu}} + eA_\rho(x)\right] - \frac{M}{c}\frac{g_{\mu\nu,\rho}\dot{x}^\mu \dot{x}^\nu}{2\sqrt{-g_{\mu\nu}\dot{x}^\mu \dot{x}^\nu}} - eA_{\mu,\rho}(x)\dot{x}^\mu = 0. \tag{9}$$

The second equation (8) means that we can identify the variable τ with the proper time of this clock. Moreover, we have $d\tau/d\lambda > 0$, and therefore it is possible to change the differential with respect to λ to one with respect to τ in the third equation (9). As a result, we find that

$$\frac{d}{d\tau}\left[\frac{M}{c^2}g_{\rho\mu}\dot{x}^\mu + eA_\rho(x)\right] - \frac{M}{2c^2}g_{\mu\nu,\rho}\dot{x}^\mu \dot{x}^\nu - eA_{\mu,\rho}(x)\dot{x}^\mu = 0,$$

where the overdot denotes the differential with respect to τ. Rewriting this equation, we get

$$\frac{M}{c^2}[\ddot{x}^\rho+\Gamma^\rho{}_{\mu\nu}\dot{x}^\mu\dot{x}^\nu]=ef^{\rho\mu}\dot{x}_\mu, \tag{10}$$

where $\Gamma^\rho_{\mu\nu}$ and $f_{\mu\nu}$ are defined by

$$\Gamma^\rho{}_{\mu\nu}=\tfrac{1}{2}g^{\rho\sigma}(-g_{\mu\nu,\sigma}+g_{\nu\sigma,\mu}+g_{\sigma\mu,\nu}),\quad f_{\mu\nu}=A_{\nu,\mu}-A_{\mu,\nu}.$$

On the other hand, the motion equation derived from the original Lagrangian L_0 is

$$m[\ddot{x}^\rho+\Gamma^\rho{}_{\mu\nu}\dot{x}^\mu\dot{x}^\nu]=ef^{\rho\mu}\dot{x}_\mu. \tag{11}$$

Equation (10) is just the same as Eq. (11) if we identify M with the constant mc^2. Equation (7) indicates that this identification is possible.

Thus, our first purpose has been achieved. Moreover, this Lagrangian L is the simplest of those that satisfy the above two conditions.

The second purpose in this section is to investigate, by using the Lagrangian L, the consequences of our assertion that the proper time should be considered a dynamic variable.

We note that it is possible to propose an argument without imposing any limitation on the fields $g_{\mu\nu}$ and A_μ. In such an argument, however, we have to handle the coordinate time $x^0=ct$ as a dynamic variable, and then determine certain constraint conditions for the variables. A discussion of such constraints is not essential for our purpose. We therefore assume for simplicity hereafter that the fields $g_{\mu\nu}$ and A_μ are so-called static in the following sense: (1) The functions $g_{\mu\nu}$ and A_μ depend on only x^1, x^2, x^3. (2) For $i=1,2,3$, we have $g_{i0}(=g_{0i})=0$.

Assuming the above conditions, we get

$$L=M(\dot{\tau}-\sqrt{f(x)^2-g_{ij}(x)\dot{x}^i\dot{x}^j/c^2})+ceA_0(x)+eA_i(x)\dot{x}^i,$$

where f is defined by $g_{00}=-f^2$ $(f>0)$. The dynamic variables are τ, M, x^i $(i=1,2,3)$, and the overdot denotes the differential with respect to t.

The momentum conjugate to those variables are given by

$$p_\tau\equiv\frac{\partial L}{\partial\dot{\tau}}=M,\quad p_M\equiv\frac{\partial L}{\partial\dot{M}}=0$$

and

$$p_i\equiv\frac{\partial L}{\partial\dot{x}^i}=\frac{M}{c^2}\frac{g_{ij}\dot{x}^j}{\sqrt{f^2-g_{jk}\dot{x}^j\dot{x}^k/c^2}}+eA_i.$$

We have

$$H_0=p_\tau\dot{\tau}+p_M\dot{M}+p_i\dot{x}^i-L=f\sqrt{M^2+c^2g^{ij}(p_i-eA_i)(p_j-eA_j)}-ceA_0.$$

If M is replaced by mc^2, then H_0 is identical with the Hamiltonian that is derived from the original Lagrangian L_0. In our case, however, there exist two constraints:

$$\phi_1\equiv M-p_\tau=0,\quad \phi_2\equiv p_M=0.$$

Taking account of these constraints, we have to consider the total Hamiltonian:

$$H\equiv H_0+u_1\phi_1+u_2\phi_2,$$

where u_1 and u_2 are Lagrange's undetermined multipliers.

The multipliers u_1 and u_2 are determined in the following manner:[3] Poisson's bracket of ϕ_1 and ϕ_2 is

$$\{\phi_1, \phi_2\} = 1,$$

and therefore we have

$$\dot{\phi}_1 = \{\phi_1, H\} \approx u_2,$$

$$\dot{\phi}_2 = \{\phi_2, H\} \approx -u_1 - \frac{fM}{\sqrt{M^2 + c^2 g^{ij}(p_i - eA_i)(p_j - eA_j)}},$$

where the symbol ''≈'' denotes the weak equality defined by the constraints $\phi_1 = \phi_2 = 0$. Hence, the consistency conditions

$$\dot{\phi}_1 \approx 0 \quad \text{and} \quad \dot{\phi}_2 \approx 0$$

require the multipliers u_1 and u_2 to be

$$u_1 = -\frac{fM}{\sqrt{M^2 + c^2 g^{ij}(p_i - eA_i)(p_j - eA_j)}} \quad \text{and} \quad u_2 = 0,$$

which give

$$H = H_0 - \frac{fM(M - p_\tau)}{\sqrt{M^2 + c^2 g^{ij}(p_i - eA_i)(p_j - eA_j)}}. \tag{12}$$

Hamilton's canonical equations of motion are as follows:

$$\dot{\tau} = \frac{\partial H}{\partial p_\tau} = \frac{fM}{\sqrt{M^2 + c^2 g^{ij}(p_i - eA_i)(p_j - eA_j)}},$$

$$\dot{p}_\tau = -\frac{\partial H}{\partial \tau} = 0,$$

$$\dot{M} = \frac{\partial H}{\partial p_M} = 0,$$

$$\dot{p}_M = -\frac{\partial H}{\partial M} \approx 0,$$

$$\dot{x}^i = \frac{\partial H}{\partial p_i} \approx \frac{\partial H_0}{\partial p_i},$$

$$\dot{p}_i = -\frac{\partial H}{\partial x^i} \approx -\frac{\partial H_0}{\partial x^i}.$$

Defining a matrix W_{ij} by

$$W_{ij} \equiv \{\phi_i, \phi_j\} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

we can write Dirac's bracket:

$$\{A,B\}_{\mathrm{D}}=\{A,B\}-\sum_{i,j=1}^{2}\{A,\phi_i\}W_{ij}^{-1}\{\phi_j,B\}=\{A,B\}+\{A,\phi_1\}\{\phi_2,B\}-\{A,\phi_2\}\{\phi_1,B\}.$$

We can easily calculate Dirac's brackets between the canonical variables:

$$\{\tau,p_\tau\}_{\mathrm{D}}=\{\tau,M\}_{\mathrm{D}}=1,\quad \{x^i,p_j\}_{\mathrm{D}}=\delta^i{}_j,\quad \text{the others}=0.$$

We are now in a position to be able to state our conclusions in this section.

It is easily shown that

$$\phi_1,\ \ \phi_2,\ \ T\equiv\tau-p_M,\quad E\equiv p_\tau,\ \ x^i,\ \ p_i\quad (i=1,2,3)$$

are canonical variables, and, therefore, the variables T,E,x^i,p_i $(i=1,2,3)$ can be interpreted as canonical variables on the submanifold defined by the constraints $\phi_1=\phi_2=0$. We can show also that

$$\{A,B\}_{\mathrm{D}}=\frac{\partial A}{\partial T}\frac{\partial B}{\partial E}-\frac{\partial A}{\partial E}\frac{\partial B}{\partial T}+\sum_{i=1}^{3}\left(\frac{\partial A}{\partial x^i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial x^i}\right),$$

on the submanifold.

Since we have that

$$T=\tau\quad\text{and}\quad E=M(=mc^2)$$

on the submanifold defined by $\phi_1=\phi_2=0$, it follows from the above that the rest energy mc^2 is considered the momentum conjugate to the proper time τ.

III. QUANTIZATION AND DISCUSSIONS

Thus, we have arrived at the following conclusion: If we accept the view that we should describe a clock by using a Lagrangian, which includes the proper time as a dynamic variable like the positions x^i, then we find that the rest energy $E=mc^2$ turns out to be the general momentum conjugate to the proper time, and that τ, E, x^i, and p_i are canonical variables of the system.

Since τ,E,x^i,p_i are the canonical variables, if we quantize the system by Dirac's procedure, there are corresponding operators:

$$\hat\tau,\quad \hat E,\quad \hat x^i,\quad \hat p_i\quad (i=1,2,3),$$

which satisfy the commutation relations

$$[\hat\tau,\hat E]=[\hat x^i,\hat p_i]=i\hbar. \tag{13}$$

The relation $[\hat\tau,\hat E]=i\hbar$ in (13) leads us to the uncertainty relation,

$$c^2\,\Delta m\,\Delta\tau\geqslant\hbar/2, \tag{14}$$

which was argued in the Introduction to this article.

Our quantization leads to some desirable results besides the uncertainty relation (14), but at the same time gives rise to some problems.

First, we should make some comment on the problems. In our quantization, the operators $\hat\tau$, $\hat E$, $\hat x^i$ and $\hat p_i$ $(i=1,2,3)$ can be represented in the Hilbert space composed of square integrable functions of τ, x^1, x^2, and x^3. In particular, the operator $\hat E$ is represented by the differential operator $-i\hbar\partial/\partial\tau$, and, therefore, the rest energy $\hat{E}$ cannot have any discrete spectrum. Furthermore, this Hilbert space includes some states in which the mean values of $\hat{E}$ are negative.

The problems of the continuous mass spectrum and of the negative mass are inevitable in our formulation. The authors cannot, at present, judge whether these characteristics are desirable or not. These problems will be discussed in a subsequent paper from a rather different viewpoint.

Second, we focus our attention on some positive aspects of our quantization. We restrict ourselves, for simplicity, to the case in which the space–time metric is flat and $A_\mu=0$. Then the Hamiltonian in (12) is rather simple and the Hamiltonian operator has the form

$$\hat{H}\equiv\sqrt{\hat{E}^2+c^2\hat{\mathbf{p}}^2}.$$

(We omit, hereafter, the overcarets representing the operators, since there is no possibility of misunderstanding.)

For the Heisenberg representation of the operator τ,

$$\tau(t)=e^{itH/\hbar}\tau e^{-itH/\hbar},$$

we find that

$$\frac{d}{d\tau}\tau(t)=\frac{i}{\hbar}e^{itH/\hbar}[H,\tau]e^{-itH/\hbar}=\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}, \tag{15}$$

by virtue of

$$[\tau,H]=i\hbar\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}.$$

Hence, we have

$$\tau(t)=\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}t+\tau. \tag{16}$$

We note that the last term of (15) is the operator that represents the time delay of the moving clock.

We can, moreover, show that

$$\frac{d}{dt}\tau(t)^2=\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}\tau(t)+\tau(t)\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}=2\frac{E^2}{E^2+c^2\mathbf{p}^2}t+\left[\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}},\tau\right]_+,$$

where we have used Eq. (16), and where $[A,B]_+$ denotes the anticommutator of operators A and B. Integrating this, we have

$$\tau(t)^2=\frac{E^2}{E^2+c^2\mathbf{p}^2}t^2+\left[\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}},\tau\right]_+t+\tau^2.$$

Hence, the standard deviation $\Delta\tau(t)$ in a state ψ is given by

$$(\Delta\tau(t))^2\equiv\langle\tau(t)^2\rangle-\langle\tau(t)\rangle^2=\left(\left\langle\frac{E^2}{E^2+c^2\mathbf{p}^2}\right\rangle-\left\langle\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}\right\rangle^2\right)t^2+\left(\left\langle\left[\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}},\tau\right]_+\right\rangle\right.$$
$$\left.-2\left\langle\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}\right\rangle\langle\tau\rangle\right)t+(\langle\tau^2\rangle-\langle\tau\rangle^2), \tag{17}$$

where $\langle A\rangle$ denotes the mean value of an operator A in the state ψ.

Here we must introduce some approximations: We assume that the Hamiltonian operator has a very sharp value (say $\mathcal{E}$) in the state ψ. This assumption seems to be natural since the clock is moving as a free particle. Under this assumption, we can approximately estimate the two terms in (17) in the following manner;

$$\left\langle \frac{E^2}{E^2+c^2\mathbf{p}^2}\right\rangle - \left\langle \frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}\right\rangle^2 \approx \frac{1}{\mathcal{E}^2}(\langle E^2\rangle - \langle E\rangle^2),$$

$$\left\langle \left[\frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}},\tau\right]_+\right\rangle - 2\left\langle \frac{E}{\sqrt{E^2+c^2\mathbf{p}^2}}\right\rangle\langle\tau\rangle \approx \frac{1}{\mathcal{E}}(\langle[E,\tau]_+\rangle - 2\langle E\rangle\langle\tau\rangle). \tag{18}$$

On the other hand, the term $\langle[E,\tau]_+\rangle - 2\langle E\rangle\langle\tau\rangle$ in (18) often vanishes, as it does in the case of all optimal simultaneous measurements of E and τ. (We can easily check it by setting, for example, $\tau = i\hbar\,\partial/\partial E$ and ψ=a Gaussian function of E.) Taking this cancellation into account, we neglect the second term in (17).

Thus, we have arrived at

$$(\Delta\tau(t))^2 \approx \frac{1}{\mathcal{E}^2}(\Delta E)^2 t^2 + (\Delta\tau)^2,$$

and, by virtue of the inequality

$$\frac{1}{\mathcal{E}^2}(\Delta E)^2 t^2 + (\Delta\tau)^2 \geqslant \frac{2}{\mathcal{E}}\Delta\tau\,\Delta E\, t,$$

we finally have

$$(\Delta\tau(t))^2 \geqslant \frac{\hbar}{\mathcal{E}}t, \tag{19}$$

where we have used the uncertainty relation $\Delta\tau\,\Delta E \geqslant \hbar/2$ of (14).

When the motion of the clock is so slow that the value of $\mathcal{E}$ is approximately equal to mc^2, then our inequality (19) has the form

$$(\Delta\tau(t))^2 \geqslant \frac{\hbar}{mc^2}t, \tag{20}$$

which exactly coincides with an inequality derived by Salecker and Wigner from another point of view [see Eq. (6) in Ref. 4].

In conclusion, we should make some comment on the meaning of our results to physics.

Bohr and Rosenfeld stressed the principle that every proper theory should provide in and by itself its own means for defining the quantities with which it deals. One of the key points this principle makes is that we should analyze the means of measuring those quantities in order to argue the consistency of a physical theory. In their case, they succeeded in showing that the definition of the standard quantization of electromagnetic field is consistent in the above sense by discussing the means of measuring the classical electromagnetic field.[5,6]

Several authors have applied this principle to the theory of relativity to find a consistent quantization of the space–time geometry. The theory deals with such quantities as the metric tensor, the curvature tensor, the covariant derivative, and connection coefficients. The measurement of the distance between two events is most fundamental in the procedures by which we measure these quantities. For this we require the concept of a clock,[7,8] and the clock cannot be independent of the various physical laws. Thus, if the above principle should be a general feature of physical theory, a consistent formulation of the quantization of the space–time geometry should have some inherent relation with various limitations on the accuracy of the clock resulting from the physical laws.

Various gedanken experiments on such limitations have been proposed and elaborated on for some 50 years.[4,7–16] In many of them, however, the clock is assumed to have some structure, from which starting point the argument is developed. It seems uncertain therefore whether their results are universal or not. Moreover, different studies sometimes reach different conclusions. Our objective in the present paper was to propose an attempt to dispose of this ambiguity. We showed the following: (a) There is an uncertainty relation between the proper time and the rest mass of a clock independent of its structure [see Eq. (3)]. (b) A limitation on the accuracy of the clock is derived from the uncertainty relation in a natural way [see Eqs. (19) and (20)].

The subject raised here has been argued, despite its importance, only at the level of thought experiments. The authors are uneasy with this situation, and think that the time has come to argue it at a more positive level. We hope that the importance of this subject is recognized and that, for example, the relation (20) is verified by experiment in the near future.

[1] A. Pais, *'Subtle is the Lord...' The Science and the Life of Albert Einstein* (Oxford University Press, Oxford, 1982).
[2] M. Jammer, *The Philosophy of Quantum Mechanics* (Wiley, New York, 1974).
[3] P. A. M. Dirac, Can. J. Math. **2**, 129 (1950); Proc. R. Soc. London, Ser. A **246**, 326 (1958).
[4] H. Salecker and E. P. Wigner, Phys. Rev. **109**, 571 (1958).
[5] N. Bohr and L. Rosenfeld, Mat. Fys. Medd. K. Dan. Vidensk. Selsk. **12**, XX (1933); Phys. Rev. **78**, 794 (1950).
[6] L. D. Landau and R. Peierls, Z. Phys. **69**, 56 (1931); in *Collected Papers of Landau*, edited by D. ter Haar (Gordon and Breach, New York, 1965), pp. 40–51.
[7] E. P. Wigner, Rev. Mod. Phys. **29**, 255 (1957).
[8] R. F. Marzke and J. A. Wheeler, in *Gravitation and Relativity*, edited by H. Y. Chiu and W. F. Hoffman (Benjamin, New York, 1964).
[9] A. Peres and N. Rosen, Phys. Rev. **118**, 335 (1960).
[10] C. A. Mead, Phys. Rev. **135**, B849 (1964).
[11] F. Károlyházy, A. Frenkel, and B. Lukács, in *Quantum Concepts in Space and Time*, edited by R. Penrose and C. J. Isham (Clarendon, Oxford, 1986).
[12] A. Charlesby, Radiat. Phys. Chem. **33**, 487 (1989).
[13] L. Diósi and B. Lukács, Phys. Lett. A **142**, 331 (1989).
[14] F. Károlyházy, in *Sixty-Two Years of Uncertainty*, edited by A. I. Miller (Plenum, New York, 1990).
[15] M. Maggiore, Phys. Lett. B **304**, 65 (1993).
[16] S. Doplicher, Ann. Inst. Henri Poincaré Phys. Theor. **64**, 543 (1996).

Path integral for the relativistic three-dimensional Aharonov–Bohm–Coulomb system

De-Hone Lin[a)]
Department of Physics, National Tsing Hua University, Hsinchu 30043, Taiwan

(Received 2 December 1997; accepted for publication 19 October 1998)

The path integral for the relativistic three-dimensional spinless Aharonov–Bohm–Coulomb system is solved, and the energy spectra are extracted from the resulting amplitude. [S0022-2488(99)00403-X]

I. INTRODUCTION

With the help of Duru and Kleinert's path-dependent time transformation,[1] the list of solvable path integrals has been extended to essentially all potential problems that possess a solvable Schrödinger equation.[2,3] Only recently has the technique been extended to relativistic potential problems.[2,4–9] In this paper, we perform the path integral of the relativistic particle in three dimensions in the presence of both an infinitely thin Aharonov–Bohm magnetic field along the z axis[10] and a $1/r$ Coulomb potential centered at the origin (ABC system). The energy spectra of the system are extracted from the resulting amplitude.

II. THE RELATIVISTIC PATH INTEGRAL

Adding a vector potential $\mathbf{A}(\mathbf{x})$ to Kleinert's path integral for a relativistic particle in a potential $V(\mathbf{x})$,[4] we find that the expression of the fixed-energy amplitude of a relativistic particle in external static electromagnetic fields is given by[6]

$$G(\mathbf{x}_b,\mathbf{x}_a;E)=\frac{i\hbar}{2Mc}\int_0^\infty dL\int \mathcal{D}\rho(\lambda)\Phi[\rho(\lambda)]\int \mathcal{D}^D x(\lambda)e^{-A_E/\hbar}, \tag{2.1}$$

with the action

$$A_E=\int_{\lambda_a}^{\lambda_b} d\lambda\left[\frac{M}{2\rho(\lambda)}\mathbf{x}'^2(\lambda)-i(e/c)\mathbf{A}(\mathbf{x})\cdot\mathbf{x}'(\lambda)-\rho(\lambda)\frac{(E-V(\mathbf{x}))^2}{2Mc^2}+\rho(\lambda)\frac{Mc^2}{2}\right]. \tag{2.2}$$

For the ABC system under consideration, the potential is

$$V(r)=-e^2/r, \tag{2.3}$$

and the vector potential reads as

$$A_i=2g\,\partial_i\varphi, \tag{2.4}$$

where e is the charge and φ is the azimuthal angle around the tube:

$$\varphi(\mathbf{x})=\arctan(x_2/x_1). \tag{2.5}$$

The associated magnetic field lines are confined to an infinitely thin tube along the z axis:

$$B_3=2g\epsilon_{3jk}\partial_j\,\partial_k\varphi=2g2\pi\delta^{(2)}(\mathbf{x}_\perp), \tag{2.6}$$

[a)]Electronic mail: d793314@phys.nthu.edu.tw

where $\mathbf{x}_\perp$ is the transverse vector $\mathbf{x}_\perp \equiv (x_1, x_2)$.

To obtain a tractable path integral for the potential $V(\mathbf{x})$, we have to regularize it via a so-called f transformation,[2,5] which exchanges the path parameter λ by a new one s:

$$d\lambda = ds\, f_l(\mathbf{x}_n) f_r(\mathbf{x}_{n-1}), \tag{2.7}$$

where $f_l(\mathbf{x})$ and $f_r(\mathbf{x})$ are invertible functions whose product is positive. The freedom in choosing $f_{l,r}$ amounts to an invariance under path-dependent reparametrizations of the path parameter λ in the fixed-energy amplitude of Eq. (2.1).[2] By this transformation, the $(D+1)$-dimensional relativistic fixed-energy amplitude for arbitrary time-independent potential turns into[2,5,6]

$$G(\mathbf{x}_b, \mathbf{x}_a; E) \approx \frac{i\hbar}{2Mc} \int_0^\infty dS \prod_{n=1}^{N+1} \left[\int d\rho_n\, \Phi(\rho_n) \right] \frac{f_l(\mathbf{x}_a) f_r(\mathbf{x}_b)}{(2\pi\hbar \epsilon_b^s \rho_b f_l(\mathbf{x}_b) f_r(\mathbf{x}_a)/M)^{D/2}} \times \prod_{n=1}^{N} \left[\int_{-\infty}^{\infty} \frac{d^D x_n}{(2\pi\hbar \epsilon_n^s \rho_n f(\mathbf{x}_n)/M)^{D/2}} \right] \exp\left\{ -\frac{1}{\hbar} A^N \right\}, \tag{2.8}$$

with the s-sliced action

$$A^N = \sum_{n=1}^{N+1} \left[\frac{M(\mathbf{x}_n - \mathbf{x}_{n-1})^2}{2\epsilon_n^s \rho_n f_l(\mathbf{x}_n) f_r(\mathbf{x}_{n-1})} - i\frac{e}{c} \mathbf{A}(\mathbf{x}_n) \cdot (\mathbf{x}_n - \mathbf{x}_{n-1}) - \epsilon_n^s \rho_n f_l(\mathbf{x}_n) f_r(\mathbf{x}_{n-1}) \frac{(E - V(\mathbf{x}_n))^2}{2Mc^2} + \epsilon_n^s \rho_n f_l(\mathbf{x}_n) f_r(\mathbf{x}_{n-1}) \frac{Mc^2}{2} \right]. \tag{2.9}$$

A family function that regulates the ABC system is

$$f_l(\mathbf{x}) = f(\mathbf{x}), \quad f_r(\mathbf{x}) = 1, \tag{2.10}$$

whose product satisfies $f_l(\mathbf{x}) f_r(\mathbf{x}) = f(\mathbf{x}) = r$. Thus, arrive at the amplitude

$$G(\mathbf{x}_b, \mathbf{x}_a; E) \approx \frac{i\hbar}{2Mc} \int_0^\infty dS \prod_{n=1}^{N+1} \left[\int d\rho_n\, \Phi(\rho_n) \right] \frac{r_a}{(2\pi\hbar \epsilon_b^s \rho_b r_b/M)^{3/2}} \times \prod_{n=2}^{N+1} \left[\int_{-\infty}^{\infty} \frac{d^3 \triangle x_n}{(2\pi\hbar \epsilon_n^s \rho_n r_{n-1}/M)^{3/2}} \right] \exp\left\{ -\frac{1}{\hbar} A^N \right\}, \tag{2.11}$$

where the action is

$$A^N = \sum_{n=1}^{N+1} \left[\frac{M(\mathbf{x}_n - \mathbf{x}_{n-1})^2}{2\epsilon_n^s \rho_n r_n} - i(e/c) \mathbf{A}(\mathbf{x}_n) \cdot (\mathbf{x}_n - \mathbf{x}_{n-1}) - \epsilon_n^s \rho_n r_n \frac{(E - V(\mathbf{x}_n))^2}{2Mc^2} + \epsilon_n^s \rho_n r_n \frac{Mc^2}{2} \right]. \tag{2.12}$$

In Eq. (2.11), we have changed the notation of the measure of integration, since x_n are Cartesian coordinates and are certainly identical in the time-sliced expressions:[2]

$$\prod_{n=1}^{N} \left[\int_{-\infty}^{\infty} d\mathbf{x}_n \right] = \prod_{n=2}^{N+1} \left[\int_{-\infty}^{\infty} d\triangle \mathbf{x}_n \right], \tag{2.13}$$

where the integrals over $\triangle \mathbf{x}_n$ may be performed successively from $n=N$ down to $n=1$. To apply the Kustaanheimo–Stiefel (KS) transformation (e.g., Ref. 2), we now incorporate the dummy fourth dimension into the action by replacing $\mathbf{x}$ in the kinetic term by the four-vector $\vec{x}$ and extending the kinetic action to

$$A_{\rm kin}^{N}=\sum_{n=1}^{N+1}\frac{M}{2}\frac{(\vec{x}_n-\vec{x}_{n-1})}{\epsilon_n^s\rho_n r_n}. \tag{2.14}$$

This is achieved by inserting the following trivial identity:

$$\prod_{n=1}^{N+1}\left[\int\frac{d(\triangle x^4)_n}{(2\pi\hbar\epsilon_n^s\rho_n r_n/M)^{1/2}}\right]\exp\left\{-\frac{1}{\hbar}\sum_{n=1}^{N+1}\frac{M}{2}\frac{(\triangle x_n^4)^2}{\epsilon_n^s\rho_n r_n}\right\}=1. \tag{2.15}$$

Hence the fixed-energy amplitude of the ABC system in three dimensions can be rewritten as the four-dimensional path integral,

$$G(\mathbf{x}_b,\mathbf{x}_a;E)\approx\frac{i\hbar}{2Mc}\int_0^\infty dS\prod_{n=1}^{N+1}\left[\int d\rho_n\,\Phi(\rho_n)\right]\int dx_a^4\frac{r_a}{(2\pi\hbar\epsilon_b^s\rho_b r_b/M)^2}$$
$$\times\prod_{n=2}^{N+1}\left[\int_{-\infty}^{\infty}\frac{d^4\triangle x_n}{(2\pi\hbar\epsilon_n^s\rho_n r_{n-1}/M)^2}\right]\exp\left\{-\frac{1}{\hbar}A^N\right\}, \tag{2.16}$$

where A^N is the action of Eq. (2.12) in which the three-vectors $\mathbf{x}_n$ of the kinetic term are replaced by the four-vectors $\vec{x}_n$. With the help of the following approximation:

$$\frac{r_a}{(2\pi\hbar\epsilon_b^s\rho_b r_b/M)^2}\prod_{n=2}^{N+1}\left[\int_{-\infty}^{\infty}\frac{d^4\triangle x_n}{(2\pi\hbar\epsilon_n^s\rho_n r_{n-1}/M)^2}\right]$$
$$\approx\frac{1}{r_a}\frac{1}{\left(2\pi\hbar\epsilon_b^s\rho_b/M\right)^2}\prod_{n=2}^{N+1}\left[\int_{-\infty}^{\infty}\frac{d^4\triangle x_n}{(2\pi\hbar\epsilon_n^s\rho_n r_n/M)^2}\right], \tag{2.17}$$

we arrive at

$$G(\mathbf{x}_b,\mathbf{x}_a;E)\approx\frac{i\hbar}{2Mc}\int_0^\infty dS\prod_{n=1}^{N+1}\left[\int d\rho_n\,\Phi(\rho_n)\right]\int dx_a^4\frac{1}{r_a}\frac{1}{(2\pi\hbar\epsilon_b^s\rho_b/M)^2}$$
$$\times\prod_{n=2}^{N+1}\left[\int_{-\infty}^{\infty}\frac{d^4\triangle x_n}{(2\pi\hbar\epsilon_n^s\rho_n r_n/M)^2}\right]\exp\left\{-\frac{1}{\hbar}\sum_{n=1}^{N+1}A^N\right\}. \tag{2.18}$$

We now solve the ABC system by introducing the KS transformation,

$$\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=A(\vec{u})\begin{pmatrix}u^1\\u^2\\u^3\\u^4\end{pmatrix}, \tag{2.19}$$

with the 3×4 matrix

$$A(\vec{u})=\begin{pmatrix} u^3 & u^4 & u^1 & u^2 \\ u^4 & -u^3 & -u^2 & u^1 \\ u^1 & u^2 & -u^3 & -u^4 \end{pmatrix}. \tag{2.20}$$

The transformation maps $R^4 \rightarrow R^3$, which means that the transformation is not one to one and no Jacobian exists. Since

$$r=(u^1)^2+(u^2)^2+(u^3)^2+(u^4)^2 \equiv (\vec{u})^2, \tag{2.21}$$

after replacing the Coulomb potential $V(r)$ with $1/r$ in Eq. (2.18), the KS transformation certainly transforms the potential terms r and $1/r$ in Eq. (2.18) into harmonic in $\vec{u}$ and $1/\vec{u}^2$, respectively. The mapping of the tangent vectors du^μ into dx^i is given by

$$\begin{pmatrix} dx_1 \\ dx_2 \\ dx_3 \end{pmatrix} = 2\begin{pmatrix} u^3 & u^4 & u^1 & u^2 \\ u^4 & -u^3 & -u^2 & u^1 \\ u^1 & u^2 & -u^3 & -u^4 \end{pmatrix}\begin{pmatrix} du^1 \\ du^2 \\ du^3 \\ du^4 \end{pmatrix}. \tag{2.22}$$

To make the mapping unique, let us embed the tangent vector (dx_1, dx_2, dx_3) into a fictitious four-dimensional space and define a new, fourth component dx_4 by an additional fourth row in the matrix, thereby extending Eq. (2.22) to the four-vector equation,

$$d\vec{x}=2A(\vec{u})d\vec{u}. \tag{2.23}$$

The arrow on the top of the x indicates that x has become a four-vector. For symmetry reasons, the 4×4 matrix $A(\vec{u})$ is chosen as

$$A(\vec{u})=\begin{pmatrix} u^3 & u^4 & u^1 & u^2 \\ u^4 & -u^3 & -u^2 & u^1 \\ u^1 & u^2 & -u^3 & -u^4 \\ u^2 & -u^1 & u^4 & -u^3 \end{pmatrix}. \tag{2.24}$$

The transformation of the coordinate difference is given by

$$(\triangle \mathbf{x}_n^i)^2 = 4\overline{\mathbf{u}_n^2}(\triangle \mathbf{u}_n^i)^2, \tag{2.25}$$

where $\overline{\mathbf{u}_n} \equiv (\mathbf{u}_n + \mathbf{u}_{n-1})/2$. In the continuum limit, this amounts to

$$d^4x_n = 16\mathbf{u}_n^2\, d^4u_n = 16 r_n\, d^4u_n, \tag{2.26}$$

$$\vec{x}'^2 = 4\vec{u}^2\vec{u}'^2 = 4r\vec{u}'^2. \tag{2.27}$$

The magnetic interaction under the KS transformation turns into

$$\mathbf{A}(\mathbf{x}_n)\cdot(\mathbf{x}_n-\mathbf{x}_{n-1}) = -2g\,\frac{x_n^2\triangle x_n^1 - x_n^1\triangle x_n^2}{r_n^2} = -2g\left[\frac{u_n^1\triangle u_n^2 - u_n^2\triangle u_n^1}{(u_n^1)^2+(u_n^1)^2} + \frac{u_n^4\triangle u_n^3 - u_n^3\triangle u_n^4}{(u_n^3)^2+(u_n^4)^2}\right]. \tag{2.28}$$

We obtain a path integral equivalent to Eq. (2.16),

$$G(\mathbf{x}_b, \mathbf{x}_a; E) = \frac{i\hbar}{2Mc}\int_0^\infty dS\, e^{SEe^2/\hbar Mc^2} G(\vec{u}_b, \vec{u}_a; S), \tag{2.29}$$

where $G(\vec{u}_b,\vec{u}_a;S)$ denotes the s-sliced amplitude,

$$\prod_{n=1}^{N+1}\left[\int d\rho_n\,\Phi(\rho_n)\right]\frac{1}{16}\int\frac{dx_a^4}{r_a}\frac{1}{(2\pi\hbar\epsilon_b^s\rho_b/m)^2}\prod_{n=1}^{N}\left[\int_{-\infty}^{\infty}\frac{d^4u_n}{(2\pi\hbar\epsilon_n^s\rho_n/m)^2}\right]\exp\left\{-\frac{1}{\hbar}A^N\right\}, \tag{2.30}$$

with the action

$$A^N=\sum_{n=1}^{N+1}\left\{\frac{m(\triangle\vec{u}_n)^2}{2\epsilon_n^s\rho_n}-i(e/c)(\vec{A}(u_n)\cdot\triangle\vec{u}_n)+\epsilon_n^s\rho_n\frac{m\omega^2\vec{u}_n^2}{2}-\epsilon_n^s\rho_n\frac{\hbar^2 4\alpha^2}{2m\vec{u}_n^2}\right\}. \tag{2.31}$$

Here

$$m=4M,\quad \omega^2=\frac{M^2c^4-E^2}{4M^2c^2}, \tag{2.32}$$

and

$$\vec{A}(u_n)\cdot\triangle\vec{u}_n=-2g\left[\frac{u_n^1\triangle u_n^2-u_n^2\triangle u_n^1}{(u_n^1)^2+(u_n^1)^2}+\frac{u_n^4\triangle u_n^3-u_n^3\triangle u_n^4}{(u_n^3)^2+(u_n^4)^2}\right]. \tag{2.33}$$

We now choose the gauge $\rho(s)=1$ in Eq. (2.30). This leads to the Duru–Kleinert transformed action,

$$A=\int_0^S ds\left[\frac{m\vec{u}'^2}{2}-2i(e/c)(\vec{A}\cdot\vec{u}')+\frac{m\omega^2\vec{u}^2}{2}-\frac{4\hbar^2\alpha^2}{2m\vec{u}^2}\right]. \tag{2.34}$$

It describes a particle, forgetting the magnetic interaction term for awhile, of mass $m=4M$ moving as a function of the ''pseudotime'' s in a four-dimensional harmonic oscillator potential of frequency,

$$\omega^2=\frac{M^2c^4-E^2}{4M^2c^2}. \tag{2.35}$$

The oscillator possesses an additional attractive potential $-4\hbar^2\alpha^2/2m\vec{u}^2$, which is conveniently parametrized in the form of a centrifugal barrier,

$$V_{\text{extra}}=\hbar^2\frac{l_{\text{extra}}^2}{2m\vec{u}^2}, \tag{2.36}$$

whose squared angular momentum has the negative value $l_{\text{extra}}^2\equiv-4\alpha^2$, where α denotes the fine-structure constant $\alpha\equiv e^2/\hbar c\approx\frac{1}{137}$.

There are no s-slicing corrections. This is ensured by the affine connection of the KS transformation, satisfying

$$\Gamma_\mu{}^{\mu\lambda}=g^{\mu\nu}e_i{}^\lambda\partial_\mu e^i{}_\nu=0, \tag{2.37}$$

and the transverse gauge $\partial_\mu A^\mu=0$.[2,5] We now analyze the effect coming from the magnetic interaction. Note that the system becomes separable like $R^4\to R^2\times R^2$ if the centrifugal term is not considered for awhile. Therefore the path integral in u space becomes two independent two-dimensional AB plus a harmonic oscillator. This makes the path integral calculation of

$G(\vec{u}_b,\vec{u}_a;S)$ extremely simple. For each two-dimensional system, the derivatives in front of φ in Eq. (2.6) commute everywhere, except at the origin, where Stokes' theorem yields

$$\int d^2u(\partial_1\partial_2-\partial_2\partial_1)\varphi=\oint d\varphi=2\pi. \tag{2.38}$$

The magnetic flux through the tube is defined by the integral

$$\Phi=\int d^2u\, B_3. \tag{2.39}$$

A comparison with Eq. (2.6) shows that the coupling constant g in Eq. (2.4) is related to the magnetic flux by

$$g=\frac{\Phi}{4\pi}. \tag{2.40}$$

When inserting $A_i=2g\,\partial_i\varphi$ into Eq. (2.34), the interaction takes the form

$$\mathcal{A}_{\mathrm{mag}}=-\hbar\beta_0\int_0^S ds\,\varphi', \tag{2.41}$$

where β_0 is the dimensionless number,

$$\beta_0\equiv-\frac{2eg}{\hbar c}. \tag{2.42}$$

The minus sign is a matter of convention. Since the particle orbits are present at all times, their world lines in space–time can be considered as being closed at infinity, and the integral

$$n=\frac{1}{2\pi}\int_0^S ds\,\varphi' \tag{2.43}$$

is the topological invariant with integer values of the winding number n. The magnetic interaction is therefore a purely topological one, its value being

$$\mathcal{A}_{\mathrm{mag}}=-\hbar\beta_0 2\pi n. \tag{2.44}$$

After adding this to the action of Eq. (2.34) in the radial decomposition of the relativistic path integral,[2,5,6] we rewrite the sum over the azimuthal quantum numbers m via Poisson's summation formula, and obtain

$$G(\mathbf{u}_b,\mathbf{u}_a;S)=\int_{-\infty}^{\infty}d\beta\frac{1}{\sqrt{u_bu_a}}G(u_b,u_a;S)_\beta\sum_{n=-\infty}^{\infty}\frac{1}{2\pi}e^{i(\beta-\beta_0)(\varphi_b+2n\pi-\varphi_a)}. \tag{2.45}$$

Since the winding number n is often not easy to measure experimentally, let us extract observable consequences that are independent of n. The sum over all n forces β to be equal to β_0 modulo an arbitrary integer number. The result, for each R^2, is

$$G(\mathbf{u}_b,\mathbf{u}_a;S)=\sum_{k=-\infty}^{\infty}\frac{1}{\sqrt{u_bu_a}}G(u_b,u_a;S)_{k+\beta_0}\frac{1}{2\pi}e^{ik(\varphi_b-\varphi_a)}. \tag{2.46}$$

Therefore we obtain the fixed-energy amplitude,

$$G(\mathbf{x}_b,\mathbf{x}_a;E)=\frac{i\hbar}{2Mc}\int_0^\infty dS\, e^{SEe^2/\hbar Mc^2}\frac{1}{16}\int\frac{dx_a^4}{r_a}\left(\frac{m\omega}{\hbar\sinh\omega s}\right)^2$$
$$\times\sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty}e^{ik_1(\varphi_{1,b}-\varphi_{1,a})}e^{ik_2(\varphi_{2,b}-\varphi_{2,a})}\exp\left\{-\frac{m\omega}{2\hbar}(\sigma_{1,b}^2+\sigma_{1,a}^2+\sigma_{2,b}^2\right.$$
$$\left.+\sigma_{2,a}^2)\coth\omega s\right\}I_{|k_1+\beta_0|}\left(\frac{m}{\hbar}\frac{\omega\sigma_{1,b}\sigma_{1,a}}{\sinh\omega s}\right)I_{|k_2+\beta_0|}\left(\frac{m}{\hbar}\frac{\omega\sigma_{2,b}\sigma_{2,a}}{\sinh\omega s}\right),\tag{2.47}$$

where (σ_1,φ_1) and (σ_2,φ_2) are defined by

$$u^1=\sigma_1\sin\varphi_1,\quad u^2=\sigma_1\cos\varphi_1,\quad u^3=\sigma_2\cos\varphi_2,\quad u^4=\sigma_2\sin\varphi_2.\tag{2.48}$$

In order to perform the x_a^4 integration, we express $(\sigma_1,\varphi_1,\sigma_2,\varphi_2)$ in terms of a three-dimensional spherical coordinate with an auxiliary angle γ:

$$\begin{array}{l}u^1=\sqrt{r}\cos(\theta/2)\cos[(\varphi+\gamma)/2]\\u^2=\sqrt{r}\cos(\theta/2)\sin[(\varphi+\gamma)/2]\\u^3=\sqrt{r}\sin(\theta/2)\cos[(\varphi-\gamma)/2]\\u^4=\sqrt{r}\sin(\theta/2)\sin[(\varphi-\gamma)/2]\end{array}\left(\begin{array}{c}0\leqslant\theta\leqslant\pi\\0\leqslant\varphi\leqslant2\pi\\0\leqslant\gamma\leqslant4\pi\end{array}\right),\tag{2.49}$$

and identify

$$\sigma_1=\sqrt{r}\cos(\theta/2),\quad\varphi_1=(\varphi+\gamma+\pi)/2,\quad\sigma_2=\sqrt{r}\sin(\theta/2),\quad\varphi_2=(\varphi-\gamma)/2.\tag{2.50}$$

Then one can change the x_a^4 integration into the γ_a integration, whose result is easily represented as the Kronecker delta δ_{k_1,k_2}. Hence, one can carry out k_2 summation and it finally becomes

$$G(\mathbf{x}_b,\mathbf{x}_a;E)=\frac{i\hbar}{2Mc}\frac{M^2\omega}{\pi\hbar^2}\sum_{k=-\infty}^{\infty}e^{ik(\varphi_b-\varphi_a)}\int_0^\infty dy\, e^{2(Ee^2/2\omega\hbar Mc^2)y}\frac{1}{\sinh^2 y}e^{-(m\omega/2\hbar)(r_b+r_a)\coth y}$$
$$\times I_{|k+\beta_0|}\left(\frac{m\omega\sqrt{r_br_a}}{\hbar\sinh y}\cos\theta_b/2\cos\theta_a/2\right)I_{|k+\beta_0|}\left(\frac{m\omega\sqrt{r_br_a}}{\hbar\sinh y}\sin\theta_b/2\sin\theta_a/2\right),\tag{2.51}$$

where we have defined the new variable $y=\omega s$. We now make use of the addition theorem, Ref. 11, Vol. II, p. 99:

$$\frac{z}{2}J_\nu(z\sin\alpha\sin\beta)J_\mu(z\cos\alpha\cos\beta)$$
$$=(\sin\alpha\sin\beta)^\nu(\cos\alpha\cos\beta)^\mu\sum_{l=0}^\infty(-1)^l(\mu+\nu+2l+1)\frac{\Gamma(\mu+\nu+l+1)\Gamma(\nu+l+1)}{l!\Gamma(\mu+l+1)\Gamma^2(\nu+1)}$$
$$\times J_{\mu+\nu+l+1}(z)\,{}_2F_1(-l,\mu+\nu+l+1,\nu+1;\sin^2\alpha)\,{}_2F_1(-l,\mu+\nu+l+1,\nu+1;\sin^2\beta),\tag{2.52}$$

and the relation between hypergeometric function and Jacobi polynomials,

$$P_l^{(\alpha,\beta)}(z)=\frac{\Gamma(\alpha+l+1)}{\Gamma(\alpha+1)l!}\,{}_2F_1\left(\alpha+\beta+l+1,-l;1+\alpha;\frac{1-z}{2}\right).\tag{2.53}$$

We arrive at

$$G(\mathbf{x}_b,\mathbf{x}_a;E)=\frac{i\hbar}{2Mc}\,\frac{M}{2\pi\hbar\sqrt{r_b r_a}}\sum_{k=-\infty}^{\infty}\sum_{n=0}^{\infty}e^{ik(\varphi_b-\varphi_a)}$$
$$\times(\cos\theta_b/2\cos\theta_a/2)^{|k+\beta_0|}(\sin\theta_b/2\sin\theta_a/2)^{|k+\beta_0|}$$
$$\times\frac{n!\,\Gamma(n+2|k+\beta_0|+1)(2n+2|k+\beta_0|+1)}{\Gamma^2(n+2|k+\beta_0|+1)}$$
$$\times\left\{\int_0^\infty dy\, e^{2(Ee^2/2\omega\hbar Mc^2)y}\frac{1}{\sinh y}e^{-m\omega/2\hbar\,(r_b+r_a)\coth y}I_{2n+2|k+\beta_0|+1}\left(\frac{m\omega\sqrt{r_b r_a}}{\hbar\sinh y}\right)\right\}$$
$$\times P_n^{(|k+\beta_0|,|k+\beta_0|)}(\cos\theta_b)P_n^{(|k+\beta_0|,|k+\beta_0|)}(\cos\theta_a). \tag{2.54}$$

At this place, the additional centrifugal barrier (2.36) is incorporated via the replacement[12]

$$(2n+2|k+\beta_0|+1)\to\sqrt{(2n+2|k+\beta_0|+1)^2-4\alpha^2}. \tag{2.55}$$

This integral can be calculated by employing the formula

$$\int_0^\infty dy\,\frac{e^{2\nu y}}{\sinh y}\exp\left[-\frac{t}{2}(\zeta_a+\zeta_b)\coth y\right]I_\mu\left(\frac{t\sqrt{\zeta_b\zeta_a}}{\sinh y}\right)=\frac{\Gamma((1+\mu)/2-\nu)}{t\sqrt{\zeta_b\zeta_a}\,\Gamma(\mu+1)}W_{\nu,\mu/2}(t\zeta_b)M_{\nu,\mu/2}(t\zeta_b), \tag{2.56}$$

with the range of validity

$$\zeta_b>\zeta_a>0,\quad \mathrm{Re}[(1+\mu)/2-\nu]>0,\quad \mathrm{Re}(t)>0,|\arg t|<\pi,$$

where $M_{\mu,\nu}$ and $W_{\mu,\nu}$ are the Whittaker functions, we complete the integration of Eq. (2.54), and find the amplitude for $r_b>r_a$ in the closed form

$$G(\mathbf{x}_b,\mathbf{x}_a;E)=\frac{i\hbar}{2Mc}\,\frac{Mc}{4\pi r_b r_a\sqrt{M^2c^4-E^2}}\sum_{k=-\infty}^{\infty}\sum_{n=0}^{\infty}e^{ik(\varphi_b-\varphi_a)}$$
$$\times(\cos\theta_b/2\cos\theta_a/2)^{|k+\beta_0|}(\sin\theta_b/2\sin\theta_a/2)^{|k+\beta_0|}\,\frac{n!(2n+2|k+\beta_0|+1)}{\Gamma(n+2|k+\beta_0|+1)}$$
$$\times\frac{\Gamma((1/2)+(1/2)\sqrt{(2n+2|k+\beta_0|+1)^2-4\alpha^2}-(E\alpha/\sqrt{M^2c^4-E^2}))}{\Gamma(\sqrt{(2n+2|k+\beta_0|+1)^2-4\alpha^2}+1)}$$
$$\times W_{E\alpha/\sqrt{M^2c^4-E^2},\sqrt{(2n+2|k+\beta_0|+1)^2-4\alpha^2}/2}\left(\frac{2}{\hbar c}\sqrt{M^2c^4-E^2}\,r_b\right)$$
$$\times M_{E\alpha/\sqrt{M^2c^4-E^2},\sqrt{(2n+2|k+\beta_0|+1)^2-4\alpha^2}/2}\left(\frac{2}{\hbar c}\sqrt{M^2c^4-E^2}\,r_a\right)$$
$$\times P_n^{(|k+\beta_0|,|k+\beta_0|)}(\cos\theta_b)P_n^{(|k+\beta_0|,|k+\beta_0|)}(\cos\theta_a). \tag{2.57}$$

The energy spectra can be extracted from the poles. They are determined by

$$\frac{1}{2}+\frac{1}{2}\sqrt{(2n+2|k+\beta_0|+1)^2-4\alpha^2}-\frac{E\alpha}{\sqrt{M^2c^4-E^2}}=-n_r,\quad n_r=0,1,2,\ldots. \tag{2.58}$$

Expanding this equation into a power of α, we get

$$E_{n_r,n,k}=\pm Mc^2\left\{1-\frac{1}{2}\left(\frac{\alpha}{n_r+n+|k+\beta_0|+1}\right)^2-\frac{\alpha^4}{(n_r+n+|k+\beta_0|+1)^3}\right.$$
$$\left.\times\left[\frac{1}{2n+2|k+\beta_0|+1}-\frac{3}{8(n_r+n+|k+\beta_0|+1)}\right]+O(\alpha^6)\right\},$$
$$n_r,n=0,1,2,3,\ldots. \tag{2.59}$$

In the nonrelativistic limit, the spectra is in agreement with the result in Refs. 13–15. It is worth noting that if the flux is quantized, i.e., $4\pi g=2\pi\hbar c/e\times$ integer, $|k+\beta_0|$ is an integer and the spectrum is that of the relativistic hydrogen atom. In this case, there is no Aharonov–Bohm effect.

ACKNOWLEDGMENT

The author is grateful to Professor H. Kleinert for his critical reading and correcting of the manuscript.

[1] H. Duru and H. Kleinert, Phys. Lett. B **84**, 185 (1979); Fortschr. Phys. **30**, 401 (1982).
[2] For reviews, see H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics and Polymer Physics*, 2nd ed. (World Scientific, Singapore, 1995), and references therein.
[3] For reviews, see A. Inomata, H. Kuratsuji, and C. C. Gerry, *Path Integrals and Coherent States of SU(2) and SU(1,1)* (World Scientific, Singapore, 1992), and references therein.
[4] H. Kleinert, Phys. Lett. A **212**, 15 (1996).
[5] D. H. Lin, J. Phys. A **30**, 3201 (1997); **30**, 4365 (1997); **31**, 7577 (1998); hep-th/9709152.
[6] D. H. Lin, J. Phys. A **31**, 4785 (1998); hep-th/9708144.
[7] B. Bentag, L. Chetouani, L. Guechi, and T. F. Hammann, Nuovo Cimento B **111**, 99 (1996).
[8] R. Ho and A. Inomata, Phys. Rev. Lett. **53**, 107 (1984).
[9] J. Gamboa and V. O. Rivelles, J. Phys. A **24**, L659 (1991).
[10] Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).
[11] A. Erdelyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1985), Vol. II.
[12] H. Kleinert, Phys. Lett. A **116**, 201 (1986).
[13] D. K. Park, S. K. Yoo, S. Y. Lee, J. R. Kahug, C. S. Park, E. S. Yim, and C. H. Lee, quant-ph/9705007v2.
[14] L. Chetouani, L. Guechi, and T. F. Hammann, J. Math. Phys. **30**, 655 (1989).
[15] M. Kibler and T. Negadi, Phys. Lett. A **124**, 42 (1987); Gh. E. Draganascu, C. Campigotto, and M. Kibler, *ibid.* **170**, 339 (1992).

Metric symmetries and spin asymmetries of Ricci-flat Riemannian manifolds

Brett McInnes,[a]
Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Republic of Singapore

(Received 31 March 1998; accepted for publication 29 October 1998)

The Calabi–Yau and Joyce manifolds used in string and *M*-theory compactifications have no continuous groups of isometries, but they often have nontrivial discrete (actually finite) isometry groups. Discrete isometries of nonsimply connected Riemannian manifolds do not necessarily map spin structures into themselves, however; thus, inconsistencies are possible if a spin connection is used to construct the gauge vacuum. We consider this problem in detail and show how it may be avoided. [S0022-2488(99)00103-6]

I. INTRODUCTION

The ''duality revolution'' in string and *M* theories[1] has revealed that the large discrete symmetry groups of these theories impose extremely powerful consistency conditions. The status of some of these symmetries remains conjectural, however, and much remains to be clarified regarding the *meaning* of symmetry in the string context.

There is no doubt that the most comprehensible symmetries of any higher-dimensional theory are the *geometric* symmetries of the underlying Riemannian and semi-Riemannian manifolds. As we shall see, even these most familiar symmetries can give rise to surprises; it is beneficial to have these complications under control before embarking on the analysis of the far more general, but much less well understood, duality symmetries. More generally, the Ricci-flat Riemannian manifolds that are our primary concern here are of continuing interest in various applications (see, for example, Ref. 2), and one wishes to understand their geometry, particularly their spin geometry,[3] as well as possible.

String and brane theories are often formulated on background manifolds of the form $M_4 \times M$, where M_4 is Minkowski space and M is a compact, locally irreducible, Riemannian, Ricci-flat, henceforth CLIRRF, manifold. (Recall that M is said to be locally irreducible if its universal Riemannian cover is irreducible, that is, not isometric to a product manifold.) Orientation-preserving isometries of $M_4 \times M$ are symmetries of the corresponding physical theory.[4] (Notice, however, that an orientation-preserving isometry of $M_4 \times M$ need *not* be orientation-preserving when restricted to M_4 or to M. Since Minkowski space admits orientation-reversing isometries, this means that we must consider all isometries of M, whether orientation preserving or orientation reversing.) The Poincaré group of symmetries of a string or brane theory is, of course, incorporated through the isometry group of M_4. It is natural to ask whether the isometry group of the internal manifold M has a similarly profound significance. Here we encounter one of the basic properties of CLIRRF manifolds: their isometry groups are necessarily *finite*.[5] Thus, there is a remarkable contrast between the external manifold M_4, which has a large group of isometries, and the internal space, which has only a discrete symmetry group—despite being a vacuum manifold. The message of ''duality,'' however, is that discrete symmetries are not to be neglected. In fact, we shall see that, precisely because they are discrete, the symmetries of CLIRRF manifolds can behave in ways that are not possible for continuous groups of isometries.

[a]Electronic mail: matmcinn@nus.edu.sg

In Riemannian geometry, the metric tensor is the basic object. One expects its symmetries to be communicated to the other structures defined by it, and this is indeed true of the Levi-Civitá connection and its curvature. We wish to ask, however, whether it is also true of the *spin connection*, the connection used to define derivatives of spinor fields. (This is a completely different object to the Levi-Civitá connection, of course, though for the purposes of strictly local calculations one often pretends that the spin connection is just the Levi-Civitá connection referred to a local orthonormal basis. See below). This is an important question, not merely because spinor fields are so important in physics, but because the spin connection itself often plays a direct role in applications. Most notably, in heterotic string compactifications[4] on Calabi–Yau manifolds, the spin connection defines the gauge vacuum. This procedure (''embedding the spin connection in the gauge group'') still plays an important role in string phenomenology.[6] Our question can be phrased as follows. Must a *metric symmetry* of a Riemannian manifold also be a *spin symmetry*? That is, do isometries necessarily induce maps that preserve spin connections?

The answer to this question is usually yes. In particular, if the isometry group in question is continuous (that is, associated with Killing vector fields in the familiar manner) or if the manifold itself is simply connected, then (apart from minor technicalities to be discussed below) every metric symmetry is a spin symmetry. The CLIRRF manifolds often used in string and brane theories—the Calabi–Yau[4] and Joyce[7] manifolds—often violate *both* of these conditions, however, and so a further investigation is needed. We find that, in general, one must expect that metric symmetries of nonsimply connected CLIRRF manifolds will *not* be spin symmetries. The reason for this is that while an isometry must necessarily induce a map from the bundle of orthonormal frames to itself, there is no reason to expect that a (discrete) isometry group will map a given spin structure[3] to itself. For, in general, a nonsimply connected Riemannian manifold will have *many* spin structures, and a discrete isometry group will tend to permute these. (In this discussion, we have been tacitly assuming that the isometries in question are orientation preserving. The orientation-reversing case is similar, though technically more complicated.)

The consequences of having a spin asymmetry in a string compactification are potentially dire. For example, it would introduce a subtle inconsistency into the anomaly cancellation mechanism: it would no longer make sense to speak of embedding ''the'' spin connection in the gauge group. Again, some Calabi–Yau manifolds have isometries with a definite physical meaning, connected, for example, with CP invariance.[8] A spin asymmetry in this case would mean that CP is not well defined when acting on fermionic fields.

The purpose of this work is to explain the methods of global differential geometry relevant to the spin asymmetry problem, and to discuss how this problem may be avoided. We begin with a rapid survey of the spin geometry of the known CLIRRF manifolds.

II. THE SPIN GEOMETRY OF CLIRRF MANIFOLDS OF NONGENERIC HOLONOMY

The only known way—and perhaps the only possible way—to obtain examples of CLIRRF manifolds is to constrain the linear holonomy group.[9] Some caution is required here, because we are interested in nonsimply connected Riemannian manifolds, and the holonomy theory of these spaces is considerably more intricate than that of their simply connected counterparts.[10,11] The spin geometry is particularly delicate in this case, so we shall briefly summarize the main points here.

Recall[5] that the *restricted* linear holonomy group of a Riemannian manifold is defined by parallel transport of vectors around contractible loops. This group, a subgroup of the full linear holonomy group, is the one classified by the classical Berger theorem.[9] A classification of the full holonomy group, which is typically a *disconnected* Lie group, has yet to be given, except in special cases. We shall say that a Riemannian manifold M is of *nongeneric* linear holonomy if its restricted linear holonomy group is a proper subgroup of the special orthogonal group $\mathrm{SO}(n)$, where here and henceforth n denotes the (real) dimension of M. Notice that such a manifold need not be orientable, since the full holonomy group may not be contained in $\mathrm{SO}(n)$. [Of course, it is always a subgroup of $\mathrm{O}(n)$.]

Now it is easy to prove that every CLIRRF manifold of nongeneric holonomy is a spin manifold,[3] *provided that it is simply connected*. It is important to note, however, that the corresponding statement is certainly *false* in the nonsimply connected case: for example, Yau's theorem[9] implies that there are Ricci-flat Kähler metrics (which are therefore of nongeneric holonomy) on the Enriques surfaces,[12] but Rochlin's theorem[3] shows at once that these are not spin manifolds. In fact, it is possible to prove that, in all dimensions that are a multiple of four, a CLIRRF manifold of nongeneric holonomy can be spin only if its fundamental group satisfies certain extremely restrictive conditions. In other dimensions, however, the restrictions are far milder.

Theorem 1: Let M be a compact, locally irreducible, Riemannian, Ricci-flat manifold of nongeneric linear holonomy. Suppose that the real dimension of M is *not* a multiple of four. Then M is spin if and only if it is orientable.

Henceforth we shall concentrate on manifolds such that n is not a multiple of four, and particularly on the cases $n=6$ and $n=7$. The full linear holonomy groups of *orientable* CLIRRF manifolds in these cases are, respectively, SU(3) and the exceptional group G_2. By the above theorem, these manifolds are always spin manifolds, whether they be simply connected or not. In fact, if they are not simply connected, they will typically (but not invariably) have *several* spin structures. The spin structures are counted[3] by the cohomology group $H^1(M,\mathbb{Z}_2)$. For example, Joyce[13] gives examples of compact seven-dimensional manifolds of linear holonomy G_2 and fundamental group isomorphic to $\mathbb{Z}_2$. Here $H^1(M,\mathbb{Z}_2)=\mathbb{Z}_2$, and so these manifolds have two distinct spin structures. In general, we denote the spin structures over M by $\mathrm{Spin}^{(i)}(M)$, where the superscript runs from one to the order of $H^1(M,\mathbb{Z}_2)$.

Each $\mathrm{Spin}^{(i)}(M)$ is a $\mathbb{Z}_2$ principal bundle over $\mathrm{SO}(M)$, a bundle of oriented orthonormal frames over M. Now because M is, in the six-dimensional case, a Kähler manifold, SO(M) reduces[5] to a sub-bundle, the bundle of unitary frames $\mathrm{U}(M)$, with structural group U(3). For each i, $\mathrm{Spin}^{(i)}(M)$ has a sub-bundle $\mathrm{Spin}^{(i)}\mathrm{U}(M)$, which projects onto $\mathrm{U}(M)$. The structural group of $\mathrm{Spin}^{(i)}\mathrm{U}(M)$ is isomorphic to U(3), since this is the subgroup of Spin(6) which projects onto the U(3) subgroup of SO(6). The existence of the bundles $\mathrm{Spin}^{(i)}\mathrm{U}(M)$ is the spin-geometric expression of the fact that M is a Kähler manifold.

For six-dimensional Calabi–Yau manifolds, the holonomy reduction theorem[5] implies a further reduction of $\mathrm{U}(M)$ to an SU(3) sub-bundle, $\mathrm{SU}(M)$. For each i, $\mathrm{Spin}^{(i)}\mathrm{U}(M)$ has a sub-bundle $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$ that projects onto $\mathrm{SU}(M)$. The structural group of $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$ is the disconnected group $\mathbb{Z}_2\times\mathrm{SU}(3)$, this being the subgroup of Spin(6) that projects onto the SU(3) subgroup of SO(6). The existence of the bundles $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$ is the spin-geometric expression of the fact that M is a *Ricci-flat* Kähler manifold.

The Levi-Civitá connection of such a manifold can be regarded as a one-form ω_L on $\mathrm{SU}(M)$. The projection $\mathrm{Spin}^{(i)}\mathrm{SU}(M)\rightarrow\mathrm{SU}(M)$ allows us to pull ω_L back to a one-form $\omega_D^{(i)}$ on $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$, and, because $\mathbb{Z}_2\times\mathrm{SU}(3)$ and SU(3) have isomorphic Lie algebras, $\omega_D^{(i)}$ defines a connection on $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$. This, in turn, defines connections, also denoted by $\omega_D^{(i)}$, on $\mathrm{Spin}^{(i)}\mathrm{U}(M)$ and $\mathrm{Spin}^{(i)}(M)$. These connections are called the Dirac or *spin connections* on M: notice that there is one spin connection for each spin structure. In seven dimensions the situation is similar: $\mathrm{SO}(M)$ reduces to a G_2 sub-bundle $G_2(M)$, and each spin structure $\mathrm{Spin}^{(i)}(M)$ reduces to a sub-bundle $\mathrm{Spin}^{(i)}G_2(M)$ with structural group $\mathbb{Z}_2\times G_2$; the Levi-Civitá connection of the Joyce metric defines a spin connection on each spin structure.

In heterotic string compactifications,[4] a spin connection is used to define an E_8 gauge configuration. (Strictly, the gauge group is $E_8\times E_8$, but we can ignore the second E_8.) The natural way to interpret this is to suppose that the bundle representing the E_8 vacuum configuration is constructed by *extending* the structural group[14] of $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$, for some fixed i, from $\mathbb{Z}_2\times\mathrm{SU}(3)$ to E_8. In this way we represent $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$ as a sub-bundle of the vacuum E_8 bundle E, and then $\omega_D^{(i)}$ defines an E_8 connection on E (by the push-forward). This is the obvious global formulation of "embedding the spin connection in the gauge group." But now we arrive at the crux: if this construction of the gauge vacuum is to be consistent with the known symmetries of string theory, we must show that isometries of M induce mappings of $\mathrm{Spin}^{(i)}\mathrm{SU}(M)$ to itself.

However, this is, in general, not the case. In order to address this problem, we need techniques for analyzing isometries of CLIRRF manifolds, and we now turn to these.

III. GLOBAL GEOMETRY OF ISOMETRIES OF CLIRRF MANIFOLDS

The most striking property of metric symmetries of CLIRRF manifolds is the fact that the corresponding groups are necessarily *finite*, like the symmetry groups of the Platonic solids. (One can see the reason for this in a very rough, intuitive way by recalling that the Ricci tensor is a kind of average sectional curvature, and that the latter controls geodesic deviation. The geodesics emanating from a typical point of a CLIRRF manifold therefore cannot deviate in a consistent way, indicating that continuous symmetries are unlikely to be found.) A familiar consequence of this absence of Killing vectors is the fact that the metrics of these spaces are not known explicitly. Surprisingly, however, this does not prevent an analysis of the isometry groups. For example, let M be a compact Kähler manifold with a vanishing first Chern class,[9] and let Γ be a finite group of holomorphic or antiholomorphic diffeomorphisms of M. Then it is possible to prove that there exists a Ricci-flat Kähler metric on M with respect to which Γ acts isometrically.[9] The isometries of these manifolds can thus be discussed quite explicitly, if need be.

CLIRRF manifolds of nongeneric holonomy have a second peculiarity, of equal importance to us. Let M be any Riemannian manifold, and let $f:M\rightarrow M$ be an isometry. If $\mathrm{O}(M)$ is the bundle of orthonormal frames[5] over M, then $\tilde{f}$, the *natural lift* of f, is an $\mathrm{O}(M)$ bundle automorphism defined by

$$\tilde{f}(u)\xi=f_*(u\xi),$$

where $u\in\mathrm{O}(M)$, $\xi\in\mathbb{R}^n$ (with $n=\dim M$), $u\xi$ is the tangent vector with components ξ with respect to u, and f_* is the differential of f. For a generic (not necessarily orientable) M, $\mathrm{O}(M)$ is not reducible, and nothing more can be said. But if M, for example, is a six-dimensional Calabi–Yau manifold, then $\mathrm{O}(M)$ reduces to the SU(3) bundle $\mathrm{SU}(M)$ and the Levi-Civitá connection ω_L may be regarded as a connection on $\mathrm{SU}(M)$. It is natural to ask whether $\tilde{f}$ restricts to an automorphism of $\mathrm{SU}(M)$—this is related to the question raised at the end of the preceding section. The answer, in general, is no. For while it is true that symmetries of the metric are also symmetries of ω_L, this merely implies that $\tilde{f}$ maps holonomy bundles[5] to *other* holonomy bundles, not necessarily to themselves. Thus $\mathrm{SU}(M)$ does not respect (all of) the symmetries of M.

We can deal with this problem as follows. Let M be any Riemannian manifold, let Γ be a group of isometries, and let P be a sub-bundle of $\mathrm{O}(M)$ with a connection that induces the Levi-Civitá connection on $\mathrm{O}(M)$. We shall say that P is *minimal* for Γ if Γ induces (through the natural lifts) bundle automorphisms of P, but not of any proper sub-bundle of P. Our task is to construct, for CLIRRF manifolds of nongeneric holonomy, the sub-bundles of $\mathrm{O}(M)$ that are minimal for a given group of isometries. Let us do this explicitly for the six- and seven-dimensional cases.

Let U(3) be regarded as a subgroup of SO(6) as usual, and let $\mathrm{SU}_k(3)$, for any positive integer k, be the subgroup of U(3) consisting of all 3×3 unitary matrices A such that

$$[\det A]^k=1.$$

This is, for all $k>1$, a disconnected group with SU(3) as an identity component. Next, let, for any even n, $\theta_{n/2}$ denote the diagonal $n\times n$ matrix with the first $n/2$ entries equal to $+1$, and the remainder equal to -1. Conjugation by $\theta_{n/2}$ maps $\mathrm{U}(n/2)$, as a subgroup of $\mathrm{SO}(n)$, to itself, by complex conjugation; therefore it is possible to define a group $\mathrm{SU}_k^*(3)$ by

$$\mathrm{SU}_k^*(3)=\mathrm{SU}_k(3)\cup\mathrm{SU}_k(3)\cdot\theta_3.$$

This is a disconnected subgroup of O(6) with SU(3) as an identity component. Abstractly it is $[\mathbb{Z}_{3k}\cdot\mathrm{SU}(3)]\rtimes\mathbb{Z}_2$, where the dot denotes the local direct product, and the product with $\mathbb{Z}_2$ is semidirect. Thus, $SU_k^*(3)$ has $2k$ connected components. Now, for a Calabi–Yau manifold, $O(M)$ is reducible to $SU(M)$, and so the structural group O(6) is reducible to any subgroup of O(6) containing SU(3). Hence we can find, for any k, a sub-bundle of $O(M)$ with $SU_k(3)$ or $SU_k^*(3)$ as a structural group, and the Levi-Civitá connection can be regarded as a connection on any of these bundles, which we denote by $SU_k(M)$ and $SU_k^*(M)$, respectively. Similarly, for seven-manifolds of linear holonomy G_2, we set

$$G_2^* = \mathbb{Z}_2 \times G_2,$$

where $\mathbb{Z}_2$, the center of O(7), is generated by $-I_7$, where I_7 is the 7×7 identity matrix; and we define $G_2^*(M)$ as a G_2^* sub-bundle of $O(M)$ in the obvious way. The result we need is as follows.

Theorem 2: Let M be a compact Riemannian six-manifold of linear holonomy SU(3), or a compact Riemannian seven-manifold of linear holonomy G_2, and let Γ be any group of isometries of M. In the six-dimensional case, let Γ_0 be the subgroup of Γ generated by holomorphic isometries. If $\Gamma = \Gamma_0$, then there exists an integer k such that $SU_k(M)$ is minimal for Γ, and there is a homomorphism from Γ_0 onto $\mathbb{Z}_k$. If $\Gamma \neq \Gamma_0$, then there exists an integer k such that $SU_k^*(M)$ is minimal for Γ, and again there is a homomorphism from Γ_0 onto $\mathbb{Z}_k$. In the seven-dimensional case, the holonomy bundle $G_2(M)$ is minimal for any orientation-preserving Γ, while $G_2^*(M)$ is minimal for any Γ containing an orientation-reversing isometry.

The proof will not be given here; see Ref. 15 for the relevant techniques.

This theorem marks the first step toward solving the problem raised at the end of Sec. II, namely, the fact that metric symmetries do not, in general, induce maps from $\mathrm{Spin}^{(i)}\,SU(M)$ to itself. For example, in the six-dimensional case, suppose that Γ is generated by a single antiholomorphic involution f. (That is, f_* anticommutes with the complex structure, and $f^2 = 1$, the identity map.) Then Γ_0 is trivial, so $k=1$, and $SU_1^*(M)$ is minimal for Γ. Thus, if we replace $SU(M)$ by $SU_1^*(M)$, we obtain a bundle that *is* mapped into itself by Γ. But now we encounter another problem, when we go to the level of spin and pin structures.

IV. SPIN SYMMETRIES AND ASYMMETRIES

In this section we introduce the concept of *spin asymmetries:* that is, maps $f: M \to M$ on a Riemannian manifold that preserve the metric but not the spin structure. To see how this is possible, let M be *any* six-dimensional orientable spin manifold. Then M is also a pin manifold, that is, the full bundle of orthonormal frames $O(M)$ has at least one nontrivial double cover that is also a Pin(6) bundle over M. [Here Pin(6) is the usual[3] double cover of O(6).] The pin structures can be expressed in terms of the spin structures in the following useful way. Let $\{e_i\}$, $i = 1\cdots 6$ generate the Clifford algebra, and set

$$\hat{\theta}_3 = e_4 e_5 e_6,$$

so that $\hat{\theta}_3$ projects onto the O(6) matrix θ_3 introduced earlier. Clearly, $\hat{\theta}_3$ is an element of Pin(6) but not of Spin(6), and $(\hat{\theta}_3)^2 = 1$. [In ten dimensions, we have $(\hat{\theta}_5)^2 = -1$, so the pin element corresponding to $\theta_{n/2}$ need not be of order two. This point will arise again, below.] Now we have

$$\mathrm{Pin}(6) = \mathrm{Spin}(6) \cup \mathrm{Spin}(6) \cdot \hat{\theta}_3,$$

and one can prove that the pin structures over M can be expressed as

$$\mathrm{Pin}^{(i)}(M) = \mathrm{Spin}^{(i)}(M) \cup \mathrm{Spin}^{(i)}(M) \cdot \hat{\theta}_3.$$

Spinors of type i on M are sections of associated bundles of $\mathrm{Spin}^{(i)}(M)$, the standard fibre being the representation space V corresponding to a specified representation ρ of Spin(6). A spinor may therefore also be regarded as a V-valued function ψ on $\mathrm{Spin}^{(i)}(M)$ satisfying $\psi(sg) = \rho(g^{-1})\psi(s)$ for each s in $\mathrm{Spin}^{(i)}(M)$ and each g in Spin(6). ''Pinors'' are defined in the obvious way. This way of thinking about spinors and pinors [as objects defined directly on $\mathrm{Spin}^{(i)}(M)$ or $\mathrm{Pin}^{(i)}(M)$] makes it clear that an isometry of M has a well-defined action on fermionic fields if and only if it induces a bundle automorphism of $\mathrm{Spin}^{(i)}(M)$ or, in the orientation-reversing case, $\mathrm{Pin}^{(i)}(M)$. Now for each i, $\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3$ is a principal bundle over $\mathrm{SO}(M)\cdot\theta_3$ with structural group $\mathbb{Z}_2$. The assumed orientability of M means that $\mathrm{O}(M)$ is the disjoint union of $\mathrm{SO}(M)$ and $\mathrm{SO}(M)\cdot\theta_3$; if f is an orientation-reversing isometry of M, then the natural lift $\tilde{f}$ maps $\mathrm{SO}(M)$ to $\mathrm{SO}(M)\cdot\theta_3$. Now the *induced bundle construction*[5] allows us to pull $\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3$ back to a $\mathbb{Z}_2$ bundle, conventionally denoted $\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3)$, over $\mathrm{SO}(M)$. The construction also supplies us with a $\mathbb{Z}_2$ bundle homomorphism $\hat{f}$,

$$\hat{f}:\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3)\rightarrow\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3,$$

covering $\tilde{f}:\mathrm{SO}(M)\rightarrow\mathrm{SO}(M)\cdot\theta_3$. Clearly $\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3)$ is just a spin structure over M. If it were possible to deform f continuously (through isometries) to the identity map, then we could try to argue that $\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3)$ must be none other than $\mathrm{Spin}^{(i)}(M)$ itself. (In fact, even in the continuous case, there can be a slight complication if the group is not simply connected; but this problem can always be resolved by replacing the group by a suitable double cover.) But if the isometry group of M is finite, we cannot argue in this way: we have

$$\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3)=\mathrm{Spin}^{(j)}(M),$$

for some j that may or may not be equal to i. The map $\hat{f}$ therefore sends $\mathrm{Spin}^{(j)}(M)$ to $\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3$, and so it does not map $\mathrm{Pin}^{(j)}(M)$ into itself unless it so happens that $i=j$. We conclude that, if $i\neq j$, then f disrupts the structure of $\mathrm{Pin}^{(i)}(M)$ and $\mathrm{Pin}^{(j)}(M)$: the very existence of f is inconsistent with any gauge vacuum built from $\mathrm{Pin}^{(i)}(M)$ or $\mathrm{Pin}^{(j)}(M)$, and f has no well-defined action on fermions of type i or j. Similarly, if f is an orientation-preserving isometry that cannot be deformed through isometries to the identity, then we expect it to permute spin structures rather than map them into themselves. This is the phenomenon of a *spin asymmetry*. Such asymmetries have arisen, at least implicitly, in the spin geometry literature: for example, the G-spin theorem (see Ref. 3, p. 267) can be made to work only if one explicitly *assumes* that the isometry in question maps a specific spin structure into itself. The question then arises as to how one can verify this assumption.

In fact, for a generic Riemannian manifold, it is very difficult to do this; both the metric and the topology of $\mathrm{O}(M)$ would have to be specified in great detail. *The possible existence of spin asymmetries should be considered very carefully in any physical theory involving nonsimply connected Riemannian manifolds*. Spin asymmetries are particularly troublesome for CLIRRF manifolds, since the metric cannot be given explicitly. We shall return to this problem in Sec. VI. Let us first, however, consider the consequences if a metric symmetry *does* happen to be a spin (or pin) symmetry.

V. EMBEDDING THE SPIN CONNECTION IN THE GAUGE GROUP

As usual, we shall discuss the more complicated, orientation-reversing case. Suppose, then, that f is an orientation-reversing isometric involution of a six-dimensional Riemannian manifold M, and suppose further that, for some i, we have

$$\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3)=\mathrm{Spin}^{(i)}(M).$$

Thus, f can be a pin symmetry as well as a metric symmetry. As above, we have a $\mathbb{Z}_2$ bundle homomorphism $\hat{f}$,

$$\hat{f}:\mathrm{Spin}^{(i)}(M)\rightarrow\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3,$$

covering $\tilde{f}\colon \mathrm{SO}(M) \to \mathrm{SO}(M)\cdot\theta_3$. Now $\tilde{f}$ also defines a map from $\mathrm{SO}(M)\cdot\theta_3$ to $\mathrm{SO}(M)$, and if f is to be a pin symmetry we must have

$$\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)) = \mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3.$$

This gives us another $\mathbb{Z}_2$ bundle homomorphism, in this case from $\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3$ to $\mathrm{Spin}^{(i)}(M)$. Combining these, we obtain an automorphism $\hat{f}\colon \mathrm{Pin}^{(i)}(M) \to \mathrm{Pin}^{(i)}(M)$. As defined, $\hat{f}$ is only an automorphism of $\mathrm{Pin}^{(i)}(M)$ as a $\mathbb{Z}_2$ bundle over $\mathrm{O}(M)$, not as a Pin(6) bundle over M. However, using the fact that $\tilde{f}$ is an automorphism of $\mathrm{O}(M)$ as an O(6) bundle over M, we see that $\hat{f}(sg) = \pm\hat{f}(s)g$ for all s in $\mathrm{Pin}^{(i)}(M)$ and g in Pin(6). If g is actually in Spin(6), a continuity argument rules out the minus sign, while if $g = \hat{\theta}_3$, the minus can be consistently absorbed into the definition of $\hat{f}$. Thus, f does have a well-defined action on $\mathrm{Pin}^{(i)}(M)$ if there is no pin asymmetry. Finally, we have the commuting diagram shown, valid for this particular value of i:

$$\begin{array}{ccc} \mathrm{Pin}^{(i)}(M) & \xrightarrow{\hat{f}} & \mathrm{Pin}^{(i)}(M) \\ \downarrow & & \downarrow \\ \mathrm{O}(M) & \xrightarrow{\tilde{f}} & \mathrm{O}(M) \\ \downarrow & & \downarrow \\ M & \xrightarrow{f} & M \end{array}.$$

Note that since f is an involution, so (by the chain rule) is $\tilde{f}$; but from this we can deduce only that $(\hat{f})^2 = \pm 1$, where, as usual, -1 denotes the automorphism of $\mathrm{Pin}^{(i)}(M)$ induced by the right action of -1, the generator of the center of Pin(6). If f is an antiholomorphic involution of a Calabi–Yau manifold, corresponding[4] to the CP operator in a heterotic string compactification, then having $(\hat{f})^2 = -1$ would mean that CP is represented by an operator of order four rather than two. For generic M it is difficult to resolve this ambiguity, but more can be said precisely when M is a CLIRRF manifold; see Theorem 3 below.

As f is a symmetry of the metric, it is also a symmetry of the Levi-Civitá connection: we have $\tilde{f}^*\omega_L = \omega_L$, where we regard[5] the Levi-Civitá connection as a one-form on $\mathrm{O}(M)$. Now let $p^{(i)}$ be the canonical projection $p^{(i)}\colon \mathrm{Pin}^{(i)}(M) \to \mathrm{O}(M)$; then the spin connection is $p^{(i)*}\omega_L = \omega_D^{(i)}$, and we have

$$\hat{f}^*\omega_D^{(i)} = (p^{(i)}\circ\hat{f})^*\omega_L = (\tilde{f}\circ p^{(i)})^*\omega_L = p^{(i)*}\tilde{f}^*\omega_L = \omega_D^{(i)}.$$

Thus, $\hat{f}$ is indeed a symmetry of the spin connection.

We shall now use these ideas to give a rigorous formulation of ''embedding the spin connection in the gauge group.''[4] It would be easy to ''embed the *linear* connection in the gauge group''—that is, to use ω_L to construct an E_8 gauge vacuum. The only complication is that, if M is a Calabi–Yau manifold and f is an antiholomorphic isometric involution on M, then $\tilde{f}$ is not an automorphism of $\mathrm{SU}(M)$. But we saw in Sec. III that this problem can always be solved—in this case, by using the bundle $\mathrm{SU}_1^*(M)$ instead of $\mathrm{SU}(M)$. But here we wish to use a *spin* connection to construct the gauge vacuum. [In heterotic string compactifications, when the ''holonomy group'' SU(3) is embedded in E_8, this embedding is through a Spin(6) subgroup of E_8, *not* through SO(6): therefore the E_8 vacuum configuration is indeed constructed from some Dirac connection $\omega_D^{(i)}$, and *not* from the Levi-Civitá connection directly.] Hence, we need to study a pin bundle over $\mathrm{SU}_1^*(M)$.

Recalling that $\mathrm{SU}_1^*(3)$ was defined as the subgroup of O(6) given by $\mathrm{SU}(3)\cup\mathrm{SU}(3)\cdot\theta_3$, we can define, for each spin structure, a sub-bundle of $\mathrm{Pin}^{(i)}(M)$ by

$$\mathrm{Spin}^{(i)}\mathrm{SU}_1^*(M)=\mathrm{Spin}^{(i)}\,\mathrm{SU}(M)\cup\mathrm{Spin}^{(i)}\,\mathrm{SU}(M)\cdot\hat{\theta}_3.$$

This is a $\mathbb{Z}_2\times\mathrm{SU}(3)\rtimes\mathbb{Z}_2$ bundle over M, where the first $\mathbb{Z}_2$ is $\{\pm 1\}$, and the second is $\{1,\hat{\theta}_3\}$. Just as $\mathrm{SU}_1^*(M)$ is the smallest sub-bundle of $\mathrm{O}(M)$ on which f induces an automorphism, so $\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)$ is the smallest sub-bundle of $\mathrm{Pin}^{(i)}(M)$ on which f induces an automorphism—*provided*, of course, that the spin asymmetry problem does not arise for this pin structure. That is, *if* we have $\tilde{f}^{-1}(\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3)=\mathrm{Spin}^{(i)}(M)$ for some i, then the pin automorphism $\hat{f}\cdot\mathrm{Pin}^{(i)}(M)\to\mathrm{Pin}^{(i)}(M)$ restricts to an automorphism $\hat{f}:\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)\to\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)$. Regarding the spin connection $\omega_D^{(i)}$ as a connection on $\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)$ [which we can do, since the latter has $\mathrm{Spin}^{(i)}\,\mathrm{SU}(M)$ as a sub-bundle], we have $\hat{f}^*\omega_D^{(i)}=\omega_D^{(i)}$. Thus, we have found a bundle on which the spin connection is well defined, and that is mapped into itself by f. [By considering $\mathrm{Spin}^{(i)}\,\mathrm{SU}_k^*(M)$ for suitable k, we can deal with isometry groups containing both holomorphic and antiholomorphic maps.] Clearly, we must use $\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)$ to construct a gauge vacuum consistent with the existence of f.

To do this, note that E_8 contains a disconnected group[16] of the form $[\mathrm{SU}(3)\cdot E_6]\rtimes\mathbb{Z}_2$, where $\mathrm{SU}(3)\cdot E_6=[\mathrm{SU}(3)\times E_6]/\mathbb{Z}_3$, and $\mathbb{Z}_2$ is generated by γ such that $\mathrm{Ad}(\gamma)$ induces complex conjugation on SU(3), together with an outer automorphism of E_6. By choosing a $\mathbb{Z}_2$ subgroup of the fixed point set of $\mathrm{Ad}(\gamma)$ in E_6 (this subgroup is of nonzero rank, so this can be done) we obtain $\mathbb{Z}_2\times\mathrm{SU}(3)\rtimes\mathbb{Z}_2$ as a closed subgroup of E_8. It follows that $\mathrm{Spin}^{(i)}\,SU_1^*(M)$ can be extended[14] to an E_8 bundle E over M. It is easy to see that $\hat{f}$ extends to an automorphism of E, that the spin connection extends to an E_8 gauge connection on E, and that the extended objects satisfy $\hat{f}^*\omega_D^{(i)}=\omega_D^{(i)}$. In short, the spin connection has been embedded in the gauge group *in a way that is consistent with the symmetries of the theory*. Similarly, if f has a physical interpretation in terms of the CP operator, then the latter has a well-defined action on fermionic fields.

In summary, then, the existence of isometries of M poses no problems, *provided* that there are no spin asymmetries. Let us now consider how to ensure this.

VI. AVOIDING SPIN ASYMMETRIES ON CLIRRF MANIFOLDS

Let M be any CLIRRF manifold of nongeneric linear holonomy. There are, in general, three simple, practical ways of ensuring that every metric symmetry of M is a spin symmetry.

The first approach is to apply nonisometric diffeomorphisms to M so that, in fact, M has no nontrivial isometries. For a Calabi–Yau manifold represented as a projective variety, this is simply a matter of adjusting the coefficients of the defining equations in such a way that M admits no holomorphic or antiholomorphic self-maps.

A second, less brutal approach involves verifying that M has only one spin structure: clearly, spin asymmetries are impossible in this case. It is quite easy to compute the number of spin structures on M if the fundamental group, $\pi_1(M)$, is known. First, $\pi_1(M)$ must be Abelianized. In the decomposition of the resulting group into finite cyclic factors, discard the factors of odd order and replace each of the others by $\mathbb{Z}_2$. The result is $H^1(M,\mathbb{Z}_2)$, which counts the spin structures over M. Clearly the spin structure is unique if $\pi_1(M)$ is of odd order; sometimes it is also unique when $\pi_1(M)$ is of even order (because a group of even order can have an Abelianization of odd order). Usually, however, M will have several spin structures if $\pi_1(M)$ is of even order.

If M does have several spin structures, we have a third approach based on the following observation. One of the most important properties of spin CLIRRF manifolds of nongeneric linear holonomy is that they always satisfy, *locally*, the integrability conditions for the existence of a parallel (covariant constant) spinor field. The existence of a *local* parallel spinor, serving as a supersymmetry generator, is essential in heterotic string phenomenology.[4] While it can be argued that it is not physically necessary to extend the local parallel spinor to a *global* field, it is natural to ask whether this can indeed be done. One finds that local parallel spinors can always be extended globally *if M* is simply connected; but if M is not simply connected, one must ask *which* spin structure is being used to make the extension. For, in fact, if *M* has more than one spin structure, there will always be spin structures that do not allow the extension to be made. On a given *M*, we shall describe a spin structure as a GPS spin structure if it admits a Global Parallel Spinor. The existence of GPS spin structures is of interest in its own right,[17] but for our purposes their importance derives from the fact that they behave well under the action of metric symmetries.

Theorem 3: Let *M* be any CLIRRF spin manifold of nongeneric linear holonomy, and let Γ be any group of isometries of *M*. There exists a GPS spin structure on *M* such that each $f\in\Gamma$ induces an automorphism $\hat{f}$ either on that spin structure or on the corresponding pin structure. Furthermore, the group generated by these automorphisms is isomorphic to Γ.

Proof: We give the proof in the case where *M* is a six-dimensional Calabi–Yau manifold and *f* is an antiholomorphic involution; the other cases are similar or easier.

Given any six-dimensional Kähler spin manifold *M*, we define

$$\mathrm{U}^*(M)=\mathrm{U}(M)\cup\mathrm{U}(M)\cdot\theta_3,$$

a sub-bundle of $\mathrm{O}(M)$ with structural group $\mathrm{U}^*(3)$ defined in the obvious way. For each spin structure, we have a spin bundle $\mathrm{Spin}^{(i)}\,\mathrm{U}(M)$ over $\mathrm{U}(M)$, and similarly we set

$$\mathrm{Spin}^{(i)}\,\mathrm{U}^*(M)=\mathrm{Spin}^{(i)}\,\mathrm{U}(M)\cup\mathrm{Spin}^{(i)}\,\mathrm{U}(M)\cdot\hat{\theta}_3.$$

Now SU(3) is a normal subgroup of U(3) [regarded as a subgroup of SO(6)], so we can define a canonical U(1) bundle over *M* by

$$K(M)=\mathrm{U}(M)/\mathrm{SU}(3).$$

Similarly, SU(3) is normal in U(3) regarded as a subgroup of Spin(6), so we can define, for each spin structure, a *spin canonical bundle*, by

$$\mathrm{Spin}^{(i)}\,K(M)=\mathrm{Spin}^{(i)}\,\mathrm{U}(M)/\mathrm{SU}(3).$$

Again, SU(3) is actually normal in $\mathrm{U}^*(3)$, so we can define

$$K^*(M)=\mathrm{U}^*(M)/\mathrm{SU}(3).$$

$$\mathrm{Spin}^{(i)}\,K^*(M)=\mathrm{Spin}^{(i)}\,U^*(M)/\mathrm{SU}(3).$$

These bundles decompose into connected components as follows:

$$K^*(M)=K(M)\cup K(M)\cdot\sigma_3,$$

where σ_3 is the SU(3) projection of θ_3, and

$$\mathrm{Spin}^{(i)}\,K^*(M)=\mathrm{Spin}^{(i)}\,K(M)\cup\mathrm{Spin}^{(i)}\,K(M)\cdot\hat{\sigma}_3,$$

where $\hat{\sigma}_3$ is the SU(3) projection of $\hat{\theta}_3$. Note that $(\sigma_3)^2=(\hat{\sigma}_3)^2=1$.

The point of passing to the SU(3) quotients of all of these structures is that when *M* is a Calabi–Yau manifold, $K(M)$ [unlike $\mathrm{U}(M)$] is actually *trivial*.[9] This allows us to describe the spin canonical bundles and hence to analyze the behavior of the spin structures.

Now *f* is an antiholomorphic, orientation-reversing isometry, so its natural lift $\tilde{f}$ is a homomorphism of U(3) bundles, $\tilde{f}:\mathrm{U}(M)\to\mathrm{U}(M)\cdot\theta_3$. Hence, it factors through the SU(3) projection and defines a U(1) homomorphism, which we also denote by $\tilde{f}$, from $K(M)$ to $K(M)\cdot\sigma_3$. Similarly, the bundle homomorphism $\hat{f}:\mathrm{Spin}^{(j)}\mathrm{U}(M)\to\mathrm{Spin}^{(i)}\,\mathrm{U}(M)\cdot\hat{\theta}_3$ defines a homomorphism

$\hat{f}$:$\mathrm{Spin}^{(j)}K(M)\rightarrow\mathrm{Spin}^{(i)}K(M)\cdot\hat{\sigma}_3$, with $j\neq i$, in general, of course. We obtain the commuting diagram shown, where the vertical arrows correspond to the $\mathbb{Z}_2$ projection π, and $\mathrm{Spin}^{(j)}K(M)$ may be regarded as the pull-back $\tilde{f}^{-1}(\mathrm{Spin}^{(i)}K(M)\cdot\hat{\sigma}_3)$,

$$\begin{array}{ccc} \mathrm{Spin}^{(j)}K(M) & \overset{\hat{f}}{\rightarrow} & \mathrm{Spin}^{(i)}K(M)\cdot\hat{\sigma}_3 \\ \downarrow & & \downarrow\pi \\ K(M) & \overset{\tilde{f}}{\rightarrow} & K(M)\cdot\sigma_3 \end{array}.$$

Now, while $K(M)$ is trivial over M, the same is *not* true, in general, of the spin canonical bundles. These last are constructed by recalling that they satisfy $\mathrm{Spin}^{(i)}K(M)/\mathbb{Z}_2=K(M)$. Typically, therefore, $\mathrm{Spin}^{(i)}K(M)$ will take the form $(\overline{M}\times\mathrm{U}(1))/\mathbb{Z}_2$, where $\overline{M}$ is some nontrivial double cover of M and $\mathbb{Z}_2$ acts diagonally. In such a case, the holonomy group[5] of the spin connection $\omega_D^{(i)}$ will be isomorphic to $\mathbb{Z}_2\times\mathrm{SU}(3)$. [This is possible because this holonomy group must project onto the *linear* holonomy group, which is SU(3), of course.] The presence of the $\mathbb{Z}_2$ factor means that there are noncontractible loops on M such that a locally constant spinor reverses sign when parallel transported around those loops. Clearly, global parallel spinors cannot be defined in such a case. Conversely, the spin structure with a trivial spin canonical bundle has SU(3) as holonomy group, and so it is a GPS spin structure. Fix i as the label for this spin structure, and let ϕ be a global cross section of $\mathrm{Spin}^{(i)}K(M)$. Consider the map $\overline{\phi}:M\rightarrow\mathrm{Spin}^{(j)}K(M)$, defined by

$$\overline{\phi}:x\rightarrow[\tilde{f}(\pi(\phi(f(x))\cdot\hat{\sigma}_3)),\phi(f(x))\cdot\hat{\sigma}_3],$$

for each $x\in M$. Here we regard $\mathrm{Spin}^{(j)}K(M)$ as the pull-back of $\mathrm{Spin}^{(i)}K(M)\cdot\hat{\sigma}_3$, so that elements of $\mathrm{Spin}^{(j)}K(M)$ are represented[5] as pairs $[k,s]$ in $K(M)\times\mathrm{Spin}^{(i)}K(M)\cdot\hat{\sigma}_3$, such that $\tilde{f}(k)=\pi(s)$. The projection for the pull-back, as a bundle over $K(M)$, is defined by $[k,s]\rightarrow k$, and since the map $x\rightarrow\tilde{f}(\pi(\phi(f(x))\cdot\hat{\sigma}_3))$ is a global cross section of $K(M)$, we see that $\overline{\phi}$ is a global cross section of $\mathrm{Spin}^{(j)}K(M)$. Thus $\mathrm{Spin}^{(j)}(M)$ is the spin structure with a trivial spin canonical bundle: that is, $j=i$, and so f induces a spin symmetry, $\hat{f}$:$\mathrm{Spin}^{(i)}(M)\rightarrow\mathrm{Spin}^{(i)}(M)\cdot\hat{\theta}_3$.

Our final task is to show that the group generated by $\hat{f}$ is isomorphic to the group generated by f: that is, we must show that $(\hat{f})^2=1$ rather than -1. Recall from Sec. V that $\hat{f}$ restricts to a map $\hat{f}$:$\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)\rightarrow\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)$. In general, $\mathrm{Spin}^{(i)}\,\mathrm{SU}(M)$ is a $\mathbb{Z}_2\times\mathrm{SU}(3)$ bundle that is not further reducible; but in the present case, since the spin holonomy group is SU(3), $\mathrm{Spin}^{(i)}\,\mathrm{SU}(M)$ decomposes as a disjoint union $P\cup(-P)$, where P is a holonomy bundle.[5] Therefore we have

$$\hat{f}(P)=\pm P\cdot\hat{\theta}_3,$$

and so, since $(\hat{\theta}_3)^2=+1$, we find that $(\hat{f})^2P=P$. That is, $(\hat{f})^2$ is an automorphism of P. But then if $(\hat{f})^2=-1$, we conclude that SU(3), the structural group of P, contains a central element of order two. Since this is not the case, we conclude that $(\hat{f})^2=+1$. This completes the proof of Theorem 3.

Evidently Theorem 3 provides us, *in the case where M* is a CLIRRF *manifold of nongeneric linear holonomy*, with another way of avoiding the spin asymmetry problem: if we insist that the local parallel spinors (corresponding, in the applications to string theory, to generators of local supersymmetries) on such manifolds must be extended to global parallel spinors, then a spin structure is automatically selected such that all metric symmetries become spin symmetries. In practice, this is perhaps the easiest way to avoid spin asymmetries; but one should bear in mind the possibility that the other spin structures may play some physical role (perhaps in a path integral in the quantum gravity context), in which case spin asymmetries will have to be considered anew.

VII. IS CP A GAUGE SYMMETRY?

The realization[18,19] that wormhole fluctuations can violate global discrete symmetries leads one to ask how these symmetries have survived into the relatively low-energy era. The most satisfactory answer[20,21] is that discrete symmetries are, in fact, *gauge* symmetries. This works very well for internal symmetries, but it is not easy to see what this proposal means in the case of the charge-parity operator CP, which is partly a geometric, space–time symmetry. In heterotic string compactifications, CP can be completely geometric, since it corresponds to a combination of the usual orientation-reversing isometry of Minkowski space with an antiholomorphic isometric involution of a Calabi–Yau manifold.[4] Nevertheless, Choi *et al.*[8] have proposed that CP is a gauge symmetry, precisely in this context. As an application of the formalism developed above, let us analyze the meaning of this proposal.

Let M be a six-dimensional Calabi–Yau manifold, and let $\mathrm{Pin}^{(i)}(M)$ be the GPS pin structure discussed in the proof of Theorem 3, so that the antiholomorphic involution f corresponding to CP induces an automorphism $\hat{f}:\mathrm{Pin}^{(i)}(M)\rightarrow\mathrm{Pin}^{(i)}(M)$ with $(\hat{f})^2=1$. As in Sec. V, we use $\mathrm{Pin}^{(i)}(M)$ to construct an E_8 bundle E over M, and we regard $\hat{f}$ as an automorphism of E. Recall that the spin connection defines a gauge connection $\omega_D^{(i)}$ on E, satisfying $\hat{f}^*\omega_D^{(i)}=\omega_D^{(i)}$. The point to be emphasized here is that, because of the special way in which the gauge bundle E is constructed, CP has a natural interpretation as *an automorphism of E that preserves the (vacuum) gauge connection*.

As is well known, the nontriviality of the vacuum gauge connection breaks E_8 to a subgroup. The precise meaning of this statement is as follows. Let M be any Riemannian manifold, let Γ by any group of isometries of M, and let $Q(M,G,\omega)$ be any principal G bundle over M with a connection ω. Following Fischer,[22] we define the *generalized gauge group* of this system as the group $A(Q,\Gamma,\omega)$ consisting of all those automorphisms $F:Q\rightarrow Q$ such that F induces an element of Γ and $F^*\omega=\omega$. The usual gauge group is the subgroup consisting of those F that cover the identity isometry of M. Notice that the gauge group is not, strictly speaking, a subgroup of the structural group G; however, it is naturally isomorphic to such a subgroup, namely, the centralizer of the holonomy group of ω. [In our case, with $G=E_8$, this centralizer is isomorphic either to $[\mathrm{U}(1)\times\mathrm{Spin}(10)]/\mathbb{Z}_4$ if the holonomy group of the spin connection is $\mathbb{Z}_2\times\mathrm{SU}(3)$, or to E_6 if—as is the case for the GPS spin structure—the spin holonomy group is SU(3).]

The *generalized* gauge group construction extends the usual gauge group by allowing F to have some ''horizontal'' as well as ''vertical'' action on the gauge bundle. As long as the induced action on M is isometric, this is a reasonable and natural extension of the gauge symmetry concept. Furthermore, like the usual gauge group, $A(Q,\Gamma,\omega)$ is naturally associated with a subgroup of G. Let P be any holonomy bundle[5] of ω in Q, and let F be any element of $A(Q,\Gamma,\omega)$; then $F^*\omega=\omega$ implies that $F(P)$ is another holonomy bundle, and so we have

$$F(P)=P\alpha(F),$$

for some $\alpha(F)\in G$. Let H_A be the subgroup of G generated by all elements of the form $\alpha(F)$: we shall say that H_A is the subgroup of G associated with $A(Q,\Gamma,\omega)$. It is not difficult to show[23] that the subgroup of G associated in this way with $V(Q,\omega)$ [the normal subgroup of $A(Q,\Gamma,\omega)$ consisting of ω-preserving, purely vertical automorphisms of Q] contains the centralizer of the holonomy group of ω, so this construction generalizes the usual isomorphism of $V(Q,\omega)$ with a subgroup of G.

The generalized gauge group construction seems ideally suited to the problem of ''gauging'' CP. Unfortunately, for a generic bundle Q, this does not work well. Note first that $A(Q,\Gamma,\omega)$ is defined as a group of automorphisms that induce some element of Γ—but this is not to say that *every* element of Γ is covered by some element of $A(Q,\Gamma,\omega)$. In general, this is not so. Even if

Γ is completely covered by $A(Q,\Gamma,\omega)$, the group structure of $A(Q,\Gamma,\omega)$ may not be reasonable from a physical point of view. For example, it would not be physically acceptable, in the string case, for an element of $A(Q,\Gamma,\omega)$ representing CP to commute with every element of E_6: in fact, we should find that $A(Q,\Gamma,\omega)$ is a semidirect product of E_6 with a $\mathbb{Z}_2$ generated by an automorphism covering CP. However, this will *not* be the case for all Q. Finally, the subgroup of G associated with $A(Q,\Gamma,\omega)$ may not contain an isomorphic copy of it, so that CP can fail to correspond to any element of the structural group.

Happily, ''embedding the spin connection in the gauge group'' eliminates all of these objections. This is the content of the following result, the proof of which is a straightforward application of Theorem 3 and will not be given here.

Theorem 4: Let M be a six-dimensional Calabi–Yau manifold, let f be an antiholomorphic isometric involution on M, let $\mathrm{Pin}^{(i)}(M)$ be a GPS pin structure on M such that f induces an automorphism $\hat{f}$ on $\mathrm{Pin}^{(i)}(M)$, and let E be the E_8 bundle over M constructed by extending $\mathrm{Spin}^{(i)}\,\mathrm{SU}_1^*(M)$ through $[\mathrm{SU}(3)\cdot E_6]\rtimes\mathbb{Z}_2$. If $\omega_D^{(i)}$ is the extension of the $\mathrm{Pin}^{(i)}(M)$ spin connection to E, then the generalized gauge group $A(E,\{1,f\},\omega_D^{(i)})$ covers $\{1,f\}$, and it is naturally isomorphic to the $E_6\rtimes\mathbb{Z}_2$ subgroup of E_8.

In simple terms, $\hat{f}$ *is the same kind of mathematical object* as the usual gauge transformations—namely, an automorphism of E that preserves the vacuum gauge fields. It is therefore perfectly natural to unify it with the gauge group, and Theorem 4 asserts that it then behaves in a physically acceptable manner. In this sense, CP is indeed a gauge symmetry, *provided* that the spin connection is embedded in the gauge group, and *provided* that we avoid spin asymmetries.

VIII. CONCLUSION

Spin connections are deceptively simple objects. Locally, they can be regarded as nothing more than the Levi-Civitá connection referred to an orthonormal basis. Globally, however, they can give rise to several interesting pathologies: in particular, they can misbehave under the action of isometries. In view of their importance for applications, it is desirable to have techniques available for dealing with their global properties. A few such techniques have been introduced here, in the context of CLIRRF manifolds.

One of the surprising byproducts of our investigation has been the discovery that it does not always make sense to speak of ''the'' spin holonomy group of a CLIRRF spin manifold: a given, fixed Calabi–Yau or Joyce manifold can have two distinct spin holonomy groups. This leads us to ask whether it is feasible to classify the spin holonomy groups of compact, locally irreducible Riemannian manifolds, after the manner of Berger's classification of linear holonomy in the simply connected case.[9] Our results in that direction will be reported elsewhere.

[1] S. Ferrara, ''Duality, strings and supergravity: a status report,'' hep-th/9708164.
[2] M. R. Douglas, A. Kato, and H. Ooguri, ''D-brane actions on Kähler manifolds,'' hep-th/9708012.
[3] H. B. Lawson and M. L. Michelsohn, *Spin Geometry* (Princeton University Press, Princeton, NJ, 1989).
[4] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, 1987).
[5] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. I.
[6] G. B. Cleaver, ''Advances in old-fashioned heterotic string model building,'' hep-th/9708023.
[7] B. S. Acharya, Nucl. Phys. B **475**, 579 (1996).
[8] K. Choi, D. B. Kaplan, and A. E. Nelson, Nucl. Phys. B **391**, 515 (1992).
[9] A. L. Besse, *Einstein Manifolds* (Springer-Verlag, Berlin, 1987).
[10] B. McInnes, J. Math. Phys. **34**, 4287 (1993).
[11] B. McInnes, J. Math. Phys. **36**, 4450 (1995).
[12] A. Beauville, *Comple Algebraic Surfaces* (Cambridge University Press, Cambridge, 1983).
[13] D. D. Joyce, J. Diff. Geom. **43**, 291 (1996).
[14] D. Husemoller, Fibre Bundles (Springer-Verlag, New York, 1975).
[15] B. McInnes, Commun. Math. Phys. **154**, 307 (1993).
[16] B. McInnes, Commun. Math. Phys. **138**, 107 (1991).
[17] B. McInnes, J. Math. Phys. **39**, 2362 (1998).
[18] S. W. Hawking, Phys. Rev. D **37**, 904 (1988).
[19] S. Coleman, Nucl. Phys. B **310**, 643 (1988).

[20] L. M. Krauss and F. Wilczek, Phys. Rev. Lett. **62**, 1221 (1989).
[21] J. Preskill and L. M. Krauss, Nucl. Phys. B **341**, 50 (1990).
[22] A. E. Fischer, Commun. Math. Phys. **113**, 231 (1987).
[23] B. McInnes, J. Math. Phys. **36**, 5414 (1995).

Nonextensive Bose–Einstein condensation model

T. Michoel[a] and A. Verbeure[b]
Instituut voor Theoretische Fysica, Katholieke Universiteit Leuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium

(Received 26 November 1997; accepted for publication 3 November 1998)

The imperfect Boson gas supplemented with a gentle repulsive interaction is completely solved. In particular, it is proved that it has nonextensive Bose–Einstein condensation, i.e., there is condensation without macroscopic occupation of the ground ($k=0$) state level. © *1999 American Institute of Physics.* [S0022-2488(99)03902-X]

I. INTRODUCTION

The search for microscopic models of interacting bosons showing Bose–Einstein condensation is an ever challenging problem. It is known that the phenomenon only appears for space dimensions $d \geqslant 3$.[1] A general two-body interacting Bose system in a finite centered cubic box $\Lambda \subset \mathbb{R}^d$, with volume $V = L^d$, is given by a Hamiltonian,

$$H_\Lambda = T_\Lambda + U_\Lambda, \tag{1}$$

where

$$T_\Lambda = \sum_{k \in \Lambda^*} \epsilon_k a_k^* a_k, \quad \epsilon_k = \frac{|k|^2}{2m},$$

$$U_\Lambda = \frac{1}{2V} \sum_{q,k,k' \in \Lambda^*} v(q) a_{k+q}^* a_{k'-q}^* a_{k'} a_k, \quad a(x) = \frac{1}{\sqrt{V}} \sum_{k \in \Lambda^*} a_k e^{ik \cdot x}.$$

The $a^\sharp(x)$ are the Boson operators satisfying the commutation rules

$$[a(x), a^*(y)] = \delta(x-y), \quad [a(x), a(y)] = 0,$$

and

$$\Lambda^* = \left\{ k : k = \frac{2\pi}{L} n, n \in \mathbb{Z}^d \right\}.$$

We limit ourself to periodic boundary conditions.

Rigorous results on the existence of Bose–Einstein condensation are known for very special potentials v in (1), in particular, of course, for $v=0$, the free Bose gas, and for v in the δ-function limit[2] or in the van der Waals limit.[3] Another class of models that are treatable is this for which the Hamiltonian is a function of the number operators $N_k = a_k^* a_k$ only. These models are called the diagonal models.[4] The Hamiltonian is a function of a set of mutually commuting operators with a spectrum consisting of the integers. The operators can be considered as random variables taking values in the integers. The equilibrium states are looked for among the measures minimizing the

[a]Aspirant van het Fonds voor Wetenschappelijk Onderzoek—Vlaanderen.
Electronic mail: tom.michoel@fys.kuleuven.ac.be
[b]Electronic mail: andre.verbeure@fys.kuleuven.ac.be

free energy. This method, developed in a series of papers (Ref. 4, and references therein), opened the possibility to derive rigorous results for so far unsolved interacting Bose gas models. The method is a powerful application of the large deviation principle for quantum systems.

In this paper we derive some rigorous results for another diagonal model, inspired by Ref. 5, where the pressure is computed. We are not using the large deviation technique of Ref. 4, but the full quantum mechanical technology, in particular, correlation inequalities, in order to prove the existence of Bose–Einstein condensation. In Sec. II, we first rederive the result of Ref. 5, and give a concise, rigorous, and direct proof of the pressure formula. Some arguments of Ref. 6 are translated into our situation. Our main contribution is in Sec. III, where we prove the occurrence of Bose–Einstein condensation, and where we study in detail the type of condensation.

There exist different types of condensation. The best known is macroscopic occupation of the ground state, but there is also so-called generalized condensation, when the number of particles distributed over a set of arbitrary small energies above the lowest energy level becomes macroscopic, proportional to the volume. This notion has been put into a rigorous and workable form in Ref. 7.

As far as our results are concerned, this notion of generalized condensation is crucial. We prove that in our model generalized condensation occurs without macroscopic occupation of the ground state. As far as we know, this is the first model of an interacting Bose gas for which this type of condensation is found. The only existing result is for the free Bose gas, considering a special thermodynamic limit, not of the type of increasing, absorbing cubes.[8,9]

The result of Sec. III also allows us, using the technique of Ref. 10, to give an explicit form of the equilibrium states in the thermodynamic limit. One verifies that they are of the same type as the equilibrium state of the imperfect Bose gas.

II. THE MODEL

In Ref. 5 Schröder considers a Bose gas contained in a d-dimensional $(d\geqslant 3)$ cubic box with Dirichlet boundary conditions on two opposite faces and periodic boundary conditions on the remaining surface. This can be interpreted as the model of a Bose gas enclosed between two hard walls at a macroscopic distance. An interaction term is introduced that behaves locally like the mean field interaction. This gives rise to the following Hamiltonian:

$$H_\Lambda=\sum_{k\in\Lambda^*}\epsilon_k N_{k,\Lambda}+\frac{\lambda}{V}\left(N_\Lambda^2+\frac{1}{2}\sum_{j\in\mathbb{N}}\tilde{N}_{j,\Lambda}^2\right), \tag{2}$$

where

$$\Lambda=\left\{x\in\mathbb{R}^d:-\frac{L}{2}\leqslant x_i\leqslant\frac{L}{2}, i=1,\dots,d-1; 0\leqslant x_d\leqslant L\right\}; V=L^d,$$

$$\Lambda^*=\frac{2\pi}{L}\mathbb{Z}^{d-1}\times\frac{\pi}{L}\mathbb{N},\quad N_{k,\Lambda}=a^*(f_{k,\Lambda})a(f_{k,\Lambda}),$$

$$f_{k,\Lambda}=\left(\frac{2}{V}\right)^{1/2}\exp[i(k_1x_1+\cdots+k_{d-1}x_{d-1})]\sin(k_dx_d),$$

$$\lambda\in\mathbb{R}^+,\quad \tilde{N}_{j,\Lambda}=\sum_{\{k\in\Lambda^*:k_d=(\pi/L)j\}}N_{k,\Lambda},$$

$$N_\Lambda=\sum_{k\in\Lambda^*}N_{k,\Lambda}.$$

Schröder shows that the grand-canonical pressure of this so-called local mean field model coincides with the grand-canonical pressure of the usual mean field model, or imperfect Bose gas, with Hamiltonian

$$H_\Lambda^{\mathrm{MF}} = \sum_{k\in\Lambda^*} \epsilon_k N_{k,\Lambda} + \frac{\lambda}{V} N_\Lambda^2, \tag{3}$$

which is a soluble model.

From this result, Schröder concludes that his model exhibits a phase transition with the same critical behavior as the imperfect Bose gas, although macroscopic occupation of the ground state may not occur, and opens the question of whether generalized condensation, as defined in Ref. 7, does take place.

We study a model of an interacting Bose gas that is inspired by Schröder's model, but that contains a nontrivial part of the self-interaction terms appearing in the general two-body repulsive interaction (1). More precisely, we consider a system of identical bosons in a centered cubic box $\Lambda \in \mathbb{R}^d$, $d \geqslant 3$, with volume $V = L^d$, with periodic boundary conditions for the wave functions, and described by the Hamiltonian

$$H_\Lambda = \sum_{k\in\Lambda^*} \epsilon_k N_{k,\Lambda} + \frac{\lambda}{V}\left(N_\Lambda^2 + \frac{1}{2}\sum_{k\in\Lambda^*} N_{k,\Lambda}^2 \right), \tag{4}$$

where now

$$\Lambda^* = \frac{2\pi}{L}\mathbb{Z}^d, \quad N_{k,\Lambda} = a_{k,\Lambda}^* a_{k,\Lambda},$$

$$a_{k,\Lambda}^* = \frac{1}{\sqrt{V}} \int_\Lambda dx\, e^{ik\cdot x} a^*(x),$$

$$\lambda \in \mathbb{R}^+, \quad N_\Lambda = \sum_{k\in\Lambda^*} N_{k,\Lambda}.$$

Our model can also be compared to the Huang–Yang–Luttinger model, rigorously studied in Ref. 11. Compared to our model, here the interaction terms $N_{k,\Lambda}^2$ appear with a minus sign and are therefore attractive perturbations of the imperfect Bose gas. The attractive character enhances (see Ref. 11) the condensation in the zero mode. The repulsive character of these terms in our model should make condensation in the zero mode more difficult. Heuristically one might expect that our model is a candidate for nonextensive Bose–Einstein condensation.

First we give a new proof, inspired by a proof in Ref. 6, of the main result of Schröder, i.e., the equality of the grand-canonical pressure of this model and the grand-canonical pressure of the imperfect Bose gas. From this we can immediately prove that there is no macroscopic occupation of any single-particle state.

For every μ in $\mathbb{R}$, denote

$$H_\Lambda(\mu) = \sum_{k\in\Lambda^*} \epsilon_k N_{k,\Lambda} + \frac{\lambda}{V}\left(N_\Lambda^2 + \frac{1}{2}\sum_{k\in\Lambda^*} N_{k,\Lambda}^2 \right) - \mu N_\Lambda, \tag{5}$$

and

$$H_\Lambda^{\mathrm{MF}}(\mu) = \sum_{k\in\Lambda^*} \epsilon_k N_{k,\Lambda} + \frac{\lambda}{V} N_\Lambda^2 - \mu N_\Lambda. \tag{6}$$

For $\alpha \leqslant 0$, let

$$\mathcal{C}^{\alpha} = \{ t \in \mathcal{C}^{b}(\mathbb{R}^{d}) : \inf_{k \in \mathbb{R}^{d}} (\epsilon_k - t_k - \alpha) > 0 \},$$

with $\mathcal{C}^{b}(\mathbb{R}^{d})$ the space of continuous bounded functions on $\mathbb{R}^{d}$. For $t \in \mathcal{C}^{\alpha}$, let

$$H_{\Lambda}^{t+\alpha} = \sum_{k \in \Lambda^*} (\epsilon_k - t_k - \alpha) N_{k,\Lambda}.$$

First, we prove the following.

Lemma 1:

$$\frac{1}{\beta V} \ln \operatorname{tr} e^{-\beta H_{\Lambda}(\mu)} \geqslant \frac{1}{\beta V} \ln \operatorname{tr} e^{-\beta H_{\Lambda}^{t+\alpha}} - \frac{1}{V} \omega_{\Lambda}^{t+\alpha}(H_{\Lambda}(\mu) - H_{\Lambda}^{t+\alpha}), \tag{7}$$

with

$$\omega_{\Lambda}^{t+\alpha}(A) = \frac{\operatorname{tr} e^{-\beta H_{\Lambda}^{t+\alpha}} A}{\operatorname{tr} e^{-\beta H_{\Lambda}^{t+\alpha}}}.$$

Proof: The function $x \in [0,1] \mapsto \ln \operatorname{tr} e^{C+xD}$, for C and D self-adjoint is convex. Hence, define the convex function f on $[0,1]$ by

$$f(x) = \ln \operatorname{tr} e^{-\beta(x H_{\Lambda}(\mu) + (1-x) H_{\Lambda}^{t+\alpha})}.$$

For all a,b in $[0,1]$, $f(a) - f(b) - (a-b) f'(b) \geqslant 0$, in particular,

$$f(1) \geqslant f(0) + f'(0),$$

which immediately yields the stated inequality. □

We can now prove a first result.

Theorem 1: *The grand-canonical pressure at chemical potential μ,*

$$\tilde{p}(\mu) = \lim_{V \to \infty} \tilde{p}_{\Lambda}(\mu) = \lim_{V \to \infty} \frac{1}{\beta V} \ln \operatorname{tr} e^{-\beta H_{\Lambda}(\mu)},$$

exists for every μ in $\mathbb{R}$ and is given by

$$\tilde{p}(\mu) = p^{\mathrm{MF}}(\mu) = \inf_{\alpha \leqslant 0} \left(p(\alpha) + \frac{(\mu - \alpha)^2}{4\lambda} \right),$$

with $p^{MF}(\mu)$ the grand-canonical pressure of the imperfect Bose gas at chemical potential μ and $p(\alpha)$ the free-gas grand-canonical pressure at chemical potential α.

[The expression for $p^{\mathrm{MF}}(\mu)$ is computed in Ref. 3.]

Proof: Since for every $\mu \in \mathbb{R}$, $H_{\Lambda}(\mu) \geqslant H_{\Lambda}^{\mathrm{MF}}(\mu)$, we have

$$\tilde{p}_{\Lambda}(\mu) \leqslant p_{\Lambda}^{\mathrm{MF}}(\mu),$$

and hence

$$\limsup_{V \to \infty} \tilde{p}_{\Lambda}(\mu) \leqslant \lim_{V \to \infty} p_{\Lambda}^{\mathrm{MF}}(\mu) = p^{\mathrm{MF}}(\mu).$$

To prove the lower bound, we make use of Lemma 1. For $\alpha \leqslant 0$ and $t \in \mathcal{C}^{\alpha}$, let

$$\rho(k;t,\alpha)=\frac{1}{e^{\beta(\epsilon_k-t_k-\alpha)}-1}.$$

Then

$$\omega_\Lambda^{t+\alpha}(N_{k,\Lambda})=\rho(k;t,\alpha),$$

$$\omega_\Lambda^{t+\alpha}(N_{k,\Lambda}N_{k',\Lambda})=\rho(k;t,\alpha)\rho(k';t,\alpha),\quad \text{if } k\neq k',$$

$$\omega_\Lambda^{t+\alpha}(N_{k,\Lambda}^2)=\rho(k;t,\alpha)(2\rho(k;t,\alpha)+1).$$

We calculate the rhs of (7). The first term gives

$$\frac{1}{\beta V}\ln \operatorname{tr} e^{-\beta H_\Lambda^{t+\alpha}}=-\frac{1}{\beta V}\sum_{k\in\Lambda^*}\ln(1-e^{-\beta(\epsilon_k-t_k-\alpha)}).$$

To calculate $(1/V)\omega_\Lambda^{t+\alpha}(H_\Lambda(\mu))$, we write

$$H_\Lambda(\mu)=\sum_{k\in\Lambda^*}(\epsilon_k-\mu)N_{k,\Lambda}+\frac{\lambda}{V}\sum_{k\in\Lambda^*}\sum_{k'\neq k\in\Lambda^*}N_{k,\Lambda}N_{k',\Lambda}+\frac{3\lambda}{2V}\sum_{k\in\Lambda^*}N_{k,\Lambda}^2,$$

hence

$$\frac{1}{V}\omega_\Lambda^{t+\alpha}(H_\Lambda(\mu))=\frac{1}{V}\sum_{k\in\Lambda^*}(\epsilon_k-\mu)\rho(k;t,\alpha)+\frac{\lambda}{V^2}\sum_{k\in\Lambda^*}\sum_{k'\neq k\in\Lambda^*}\rho(k;t,\alpha)\rho(k';t,\alpha)+\frac{c_V}{V},$$

where

$$c_V=\frac{3\lambda}{2V}\sum_{k\in\Lambda^*}\rho(k;t,\alpha)(2\rho(k;t,\alpha)+1).$$

Also,

$$\frac{1}{V}\omega_\Lambda^{t+\alpha}(H_\Lambda^{t+\alpha})=\frac{1}{V}\sum_{k\in\Lambda^*}(\epsilon_k-t_k-\alpha)\rho(k;t,\alpha).$$

Substituting all this in (7), we get

$$\begin{aligned}\tilde{p}_\Lambda(\mu)\geq&-\frac{1}{\beta V}\sum_{k\in\Lambda^*}\ln(1-e^{-\beta(\epsilon_k-t_k-\alpha)})+\frac{1}{V}\sum_{k\in\Lambda^*}(\mu-t_k-\alpha)\rho(k;t,\alpha)\\&-\frac{\lambda}{V^2}\sum_{k\in\Lambda^*}\sum_{k'\neq k\in\Lambda^*}\rho(k;t,\alpha)\rho(k';t,\alpha)-\frac{c_V}{V}.\end{aligned}$$

Since $\rho(k;t,\alpha)$ and c_V, for V large enough, are bounded

$$\begin{aligned}\liminf_{V\to\infty}\tilde{p}_\Lambda(\mu)\geq&-\beta^{-1}\int_{\mathbb{R}^d}\frac{dk}{(2\pi)^d}\ln(1-e^{-\beta(\epsilon_k-t_k-\alpha)})\\&+\int_{\mathbb{R}^d}\frac{dk}{(2\pi)^d}(\mu-t_k-\alpha)\rho(k;t,\alpha)-\lambda\left(\int_{\mathbb{R}^d}\frac{dk}{(2\pi)^d}\rho(k;t,\alpha)\right)^2.\end{aligned}\tag{8}$$

For $\alpha\leq 0$ the free-gas pressure is given by

$$p(\alpha) = -\beta^{-1}\int_{\mathbb{R}^d} \frac{dk}{(2\pi)^d} \ln(1 - e^{-\beta(\epsilon_k - \alpha)})$$

and

$$p'(\alpha) = \int_{\mathbb{R}^d} \frac{dk}{(2\pi)^d} \frac{1}{e^{\beta(\epsilon_k-\alpha)} - 1}.$$

Also, let $\rho_c = p'(0)$ as usual.

First, consider the case $\mu < 2\lambda\rho_c$. Taking $\alpha < 0$ and $t = 0$ in (8) we get

$$\liminf_{V\to\infty} \tilde{p}_\Lambda(\mu) \geq p(\alpha) + (\mu - \alpha)p'(\alpha) - \lambda(p'(\alpha))^2. \tag{9}$$

For $\mu < 2\lambda\rho_c$, since $p'(\alpha)$ is increasing and $p'(0) = \rho_c$, the equation

$$p'(\alpha) = \frac{\mu - \alpha}{2\lambda}$$

has a unique solution $\alpha^* < 0$. Taking $\alpha = \alpha^*$ in (9), we get

$$\liminf_{V\to\infty} \tilde{p}_\Lambda(\mu) \geq p(\alpha^*) + \frac{(\mu - \alpha^*)^2}{4\lambda} = \inf_{\alpha \leq 0}\left\{ p(\alpha) + \frac{(\mu-\alpha)^2}{4\lambda}\right\} = p^{\mathrm{MF}}(\mu),$$

which proves the theorem for $\mu < 2\lambda\rho_c$.

Consider now the case $\mu \geq 2\lambda\rho_c$. Take $\alpha = 0$ and an appropriate t in (8):

$$\liminf_{V\to\infty} \tilde{p}_\Lambda(\mu) \geq -\beta^{-1}\int_{\mathbb{R}^d} \frac{dk}{(2\pi)^d} \ln(1 - e^{-\beta(\epsilon_k - t_k)}) + \int_{\mathbb{R}^d} \frac{dk}{(2\pi)^d} (\mu - t_k)\rho(k;t) - \lambda\left(\int_{\mathbb{R}^d} \frac{dk}{(2\pi)^d} \rho(k;t)\right)^2, \tag{10}$$

with

$$\rho(k;t) = \frac{1}{e^{\beta(\epsilon_k - t_k)} - 1}.$$

For all $\delta > 0$, take $t_\delta \in \mathcal{C}^0$ such that

$$t_\delta(k) = 0, \quad |k| > \delta.$$

Then

$$\int_{\mathbb{R}^d} \frac{dk}{(2\pi)^d} \rho(k;t_\delta) = \int_{|k|\leq\delta} \frac{dk}{(2\pi)^d} \rho(k;t_\delta) + \int_{|k|>\delta} \frac{dk}{(2\pi)^d} \frac{1}{e^{\beta\epsilon_k} - 1}.$$

Letting $\delta \to 0$, the second term on the rhs converges to ρ_c. Take t_δ such that the first term on the rhs converges to $\mu/2\lambda - \rho_c$ as $\delta \to 0$. Such a sequence of t_δ's can be constructed rigorously by using the Approximation theorem proved in Ref. 12. It certainly means that $t_\delta \to 0$ as $\delta \to 0$. Hence we get

$$\int_{\mathbb{R}^d} \frac{dk}{(2\pi)^d} \rho(k;t_\delta) \to \frac{\mu}{2\lambda},$$

as $\delta\rightarrow 0$, and thus

$$\liminf_{V\rightarrow\infty}\tilde{p}_\Lambda(\mu)\geqslant p(0)+\frac{\mu^2}{4\lambda}\geqslant\inf_{\alpha\leqslant 0}\left\{p(\alpha)+\frac{(\mu-\alpha)^2}{4\lambda}\right\}=p^{\mathrm{MF}}(\mu),$$

so that the theorem is proved for $\mu\geqslant 2\lambda\rho_c$ as well. □

From Theorem 1 we can immediately derive that there is no macroscopic occupation of any single-particle state, in particular, the following.

Theorem 2: *For every* $\epsilon>0$ *and for* V *large enough, we have, for every* $k\in\Lambda^*$:

$$\frac{1}{V}\omega_\Lambda(N_{k,\Lambda})<\epsilon,$$

where ω_Λ *is the finite-volume Gibbs state of* $H_\Lambda(\mu)$.

Proof: We have

$$e^{\beta V p_\Lambda^{\mathrm{MF}}(\mu)}=\mathrm{tr}\, e^{-\beta H_\Lambda^{\mathrm{MF}}(\mu)}=\mathrm{tr}(e^{-\beta H_\Lambda(\mu)}e^{(\beta\lambda/2V)\Sigma_{k\in\Lambda^*}N_{k,\Lambda}^2})=\omega_\Lambda(e^{(\beta\lambda/2V)\Sigma_{k\in\Lambda^*}N_{k,\Lambda}^2})e^{\beta V\tilde{p}_\Lambda(\mu)}.$$

Hence,

$$p_\Lambda^{\mathrm{MF}}=\tilde{p}_\Lambda(\mu)+\frac{1}{\beta V}\ln\omega_\Lambda(e^{(\beta\lambda/2V)\Sigma_{k\in\Lambda^*}N_{k,\Lambda}^2}).$$

By Theorem 1 we get

$$\lim_{V\rightarrow\infty}\frac{1}{\beta V}\ln\omega_\Lambda(e^{(\beta\lambda/2V)\Sigma_{k\in\Lambda^*}N_{k,\Lambda}^2})=0.$$

From the Jensen inequality, i.e., for F a convex function and ω a normal state,

$$\omega(F(X))\geqslant F(\omega(X)),$$

we get

$$\omega_\Lambda(e^{(\beta\lambda/2V)\Sigma_{k\in\Lambda^*}N_{k,\Lambda}^2})\geqslant e^{(\beta\lambda/2V)\Sigma_{k\in\Lambda^*}\omega_\Lambda(N_{k,\Lambda}^2)},$$

or

$$0\leqslant\frac{\lambda}{2V^2}\sum_{k\in\Lambda^*}\omega_\Lambda(N_{k,\Lambda}^2)\leqslant\frac{1}{\beta V}\ln\omega_\Lambda(e^{(\beta\lambda/2V)\Sigma_{k\in\Lambda^*}N_{k,\Lambda}^2}).$$

Hence

$$\lim_{V\rightarrow\infty}\frac{1}{V^2}\sum_{k\in\Lambda^*}\omega_\Lambda(N_{k,\Lambda}^2)=0.$$

Since for each $k\in\Lambda^*$,

$$0\leqslant\left(\frac{1}{V}\omega_\Lambda(N_{k,\Lambda})\right)^2\leqslant\frac{1}{V^2}\omega_\Lambda(N_{k,\Lambda}^2)\leqslant\frac{1}{V^2}\sum_{k'\in\Lambda^*}\omega_\Lambda(N_{k',\Lambda}^2),$$

we get the Theorem. □

III. BOSE–EINSTEIN CONDENSATION

In Ref. 7 it is stressed that Bose condensation does not necessarily manifest itself through a macroscopic occupation of a single-particle state (the ground state usually), but that there are, in fact, two good candidates for the concept of macroscopic occupation of the zero-kinetic energy state. Macroscopic occupation of the ground state is said to occur when the number of particles in the ground state becomes proportional to the volume; generalized condensation is said to occur when the number of particles whose energy levels lie in an arbitrary small band above zero becomes proportional to the volume. Obviously, the first implies the second. However, the second can occur without the first; this is called nonextensive condensation. The concept of generalized condensation was first introduced in Ref. 13. More precisely, we have the following.

(i) Macroscopic occupation of the ground state if the limit

$$\lim_{V\to\infty}\frac{1}{V}\,\omega_\Lambda(N_{0,\Lambda})$$

exists and is strictly positive; (ii) generalized condensation if the limit

$$\lim_{\delta\to 0}\lim_{V\to\infty}\frac{1}{V}\sum_{\{k\in\Lambda^*:\epsilon_k<\delta\}}\omega_\Lambda(N_{k,\Lambda})$$

exists and is strictly positive; (iii) nonextensive condensation if the limit

$$\lim_{V\to\infty}\frac{1}{V}\,\omega_\Lambda(N_{0,\Lambda})=0,$$

but nevertheless the limit

$$\lim_{\delta\to 0}\lim_{V\to\infty}\frac{1}{V}\sum_{\{k\in\Lambda^*:\epsilon_k<\delta\}}\omega_\Lambda(N_{k,\Lambda})$$

exists and is strictly positive.

Examples of these different occurrences of Bose condensation in the free Bose gas, depending on how the bulk limit is taken, can be found in Refs. 7–9.

As is proved in Theorem 2, there is no macroscopic occupation of the ground state in our system. However, as we will show, there is generalized condensation. In other words, we have a model for an interacting Bose gas displaying nonextensive condensation.

Our approach is based on Ref. 10, where the imperfect Bose gas is treated. The system is given by the local Hamiltonian H_Λ, with periodic boundary conditions

$$H_\Lambda=\sum_{k\in\Lambda^*}\epsilon_k N_{k,\Lambda}+\frac{\lambda}{V}\left(N_\Lambda^2+\frac{1}{2}\sum_{k\in\Lambda^*}N_{k,\Lambda}^2\right)-\mu_\Lambda N_\Lambda, \tag{11}$$

as specified before, and μ_Λ is determined by the constant density $\rho>0$ equation:

$$\frac{1}{V}\,\omega_\Lambda(N_\Lambda)=\rho.$$

We study the equilibrium state of this system in the grand-canonical ensemble. The key technique is the equivalence of the equilibrium condition or Gibbs state ω_Λ with the correlation inequalities[14,15]

$$\beta\omega_\Lambda(X^*[H_\Lambda,X])\geqslant\omega_\Lambda(X^*X)\ln\frac{\omega_\Lambda(X^*X)}{\omega_\Lambda(XX^*)},\tag{12}$$

for all local observables X belonging to the domain of $[H_\Lambda,\cdot]$. In particular, we take for X polynomials in the creation and annihilation operators. We prove the occurrence of nonextensive condensation in this model, and follow closely the method used in Ref. 10.

Lemma 2: $\forall k,\ j\in\Lambda^*$:

(i)

$$\beta\omega_\Lambda\left(-\epsilon_k N_{k,\Lambda}+\left(\mu_\Lambda-\frac{2\lambda}{V}N_\Lambda\right)N_{k,\Lambda}-\frac{\lambda}{V}N_{k,\Lambda}^2+\frac{3\lambda}{2V}N_{k,\Lambda}\right)\geqslant\omega_\Lambda(N_{k,\Lambda})\ln\frac{\omega_\Lambda(N_{k,\Lambda})}{\omega_\Lambda(N_{k,\Lambda})+1};\tag{13}$$

(ii)

$$\omega_\Lambda\left(\left(\mu_\Lambda-\frac{2\lambda}{V}N_\Lambda\right)N_{k,\Lambda}\right)\leqslant\omega_\Lambda\left(\epsilon_j N_{k,\Lambda}+\frac{4\lambda}{V}N_{j,\Lambda}N_{k,\Lambda}+\frac{3\lambda}{2V}N_{k,\Lambda}\right).\tag{14}$$

Proof: For *(i)*, the result follows by taking $X=a_k$ in the correlation inequality (12). One gets *(ii)* by taking

$$X=a_j N_{k,\Lambda}^{1/2},$$

in the inequality

$$\omega_\Lambda([X^*,[H_\Lambda,X]])\geqslant 0,$$

which follows immediately from (12) by adding the correlation inequality for X and the complex conjugate of the correlation inequality for X^*. □

Lemma 3: For every $\delta>0$, *for every* V *and for every* $k\in\Lambda^*$, $|k|\geqslant\delta$,

$$\omega_\Lambda(N_{k,\Lambda})\leqslant\frac{1}{e^{c_k(\Lambda)}-1}+\frac{4\lambda}{V}\omega_\Lambda(N_{j,\Lambda}N_{k,\Lambda})\frac{1}{1-e^{-c_\delta(\Lambda)}},$$

with

$$c_k(\Lambda)=\beta\left(\epsilon_k-\frac{\delta^2}{8m}-\frac{3\lambda}{V},\right),$$

$c_\delta(\Lambda)=c_k(\Lambda)|_{|k|=\delta}$ *and* $j\in\Lambda^*$, $|j|\leqslant\delta/2$.

Proof: Substitution of (14) in (13), changing the sign, and using the trivial bound $\omega_\Lambda(N_{k,\Lambda}^2)\geqslant 0$ we get

$$\beta\left(\epsilon_k-\epsilon_j-\frac{3\lambda}{V}\right)\omega_\Lambda(N_{k,\Lambda})-\frac{4\lambda}{V}\omega_\Lambda(N_{j,\Lambda}N_{k,\Lambda})\leqslant\omega_\Lambda(N_{k,\Lambda})\ln\frac{\omega_\Lambda(N_{k,\Lambda})+1}{\omega_\Lambda(N_{k,\Lambda})}.\tag{15}$$

Take $\delta>0$ arbitrary, $|k|\geqslant\delta$, and $|j|\leqslant\delta/2$.

Using $\epsilon_j\leqslant\delta^2/8m$, (15) becomes

$$c_k(\Lambda)\omega_\Lambda(N_{k,\Lambda})-\frac{4\lambda}{V}\omega_\Lambda(N_{j,\Lambda}N_{k,\Lambda})\leqslant\omega_\Lambda(N_{k,\Lambda})\ln\frac{\omega_\Lambda(N_{k,\Lambda})+1}{\omega_\Lambda(N_{k,\Lambda})}.$$

The lemma now follows from convexity arguments on the rhs: we want to solve for t the inequality

$$ct-b\leqslant t\ln\frac{t+1}{t},$$

with c and b positive constants and $t\in\mathbb{R}^+$. It follows that $t\leqslant t_2$, with t_2, satisfying

$$ct_2-b=t_2\ln\frac{t_2+1}{t_2}.$$

One can write this as $t\leqslant t_1+(t_2-t_1)$, with

$$t_1=\frac{1}{e^c-1}.$$

Let $f(t)=t\ln(t+1)/t$, f is concave, hence

$$f(t_2)-f(t_1)-(t_2-t_1)f'(t_1)\leqslant 0,$$

and

$$t_2-t_1\leqslant b\,\frac{1}{t-e^{-c}}.$$

Substitute this into the inequality $t\leqslant t_1+(t_2-t_1)$, one gets

$$t\leqslant\frac{1}{e^c-1}+b\,\frac{1}{1-e^{-c}}.$$

Finally, use $|k|\geqslant\delta$ in the second term on the rhs to prove the lemma. □

Lemma 4: For every $\epsilon>0$, *V large enough and* $j\in\Lambda^*$:

$$\frac{1}{V^2}\,\omega_\Lambda(N_\Lambda N_{j,\Lambda})<\epsilon.$$

Proof: (13) gives

$$\beta\omega_\Lambda\left(-\epsilon_j N_{j,\Lambda}+\left(\mu_\Lambda-\frac{2\lambda}{V}N_\Lambda\right)N_{j,\Lambda}+\frac{3\lambda}{2V}N_{j,\Lambda}\right)\geqslant\omega_\Lambda(N_{j,\Lambda})\ln\frac{\omega_\Lambda(N_{j,\Lambda})}{\omega_\Lambda(N_{j,\Lambda})+1}\geqslant-1.$$

This can be rewritten in the form

$$\frac{2\lambda}{V^2}\,\omega_\Lambda(N_\Lambda N_{j,\Lambda})\leqslant\frac{1}{\beta V}+\left(\mu_\Lambda+\frac{3\lambda}{2V}-\epsilon_j\right)\frac{1}{V}\,\omega_\Lambda(N_{j,\Lambda}).\tag{16}$$

Taking $X=a_j$ in the inequality

$$\omega_\Lambda([X^*,[H_\Lambda,X]])\geqslant 0,$$

gives

$$\mu_\Lambda\leqslant 2\lambda\rho+\epsilon_j+\frac{4\lambda}{V}\,\omega_\Lambda(N_{j,\Lambda})+\frac{3\lambda}{2V}.$$

Putting this into (16) gives

$$\frac{2\lambda}{V^2}\,\omega_\Lambda(N_\Lambda N_{j,\Lambda})\le \frac{1}{\beta V}+\left(2\lambda\rho+\frac{3\lambda}{V}\right)\frac{1}{V}\,\omega_\Lambda(N_{j,\Lambda})+\frac{4\lambda}{V^2}\,\omega_\Lambda(N_{j,\Lambda})^2.$$

Using Theorem 2 proves the lemma. □

We now prove the existence of generalized condensation in the thermodynamic limit $V\to\infty$, taken with constant particle density ρ.

Theorem 3: *One has (i)*

$$\lim_{\delta\to 0}\lim_{V\to\infty}\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|<\delta\}}\omega_\Lambda(N_{k,\Lambda})\ge\rho-\int_{\mathbb{R}^d}\frac{dk}{(2\pi)^d}\,\frac{1}{e^{\beta\epsilon_k}-1};$$

(ii) for every $\rho>0$, there is a β_c such that for all $\beta>\beta_c$:

$$0<\lim_{\delta\to 0}\liminf_{V\to\infty}\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|<\delta\}}\omega_\Lambda(N_{k,\Lambda})\le\lim_{\delta\to 0}\limsup_{V\to\infty}\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|<\delta\}}\omega_\Lambda(N_{k,\Lambda})\le\rho.$$

Proof: We have clearly

$$\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|<\delta\}}\omega_\Lambda(N_{k,\Lambda})=\rho-\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|\ge\delta\}}\omega_\Lambda(N_{k,\Lambda}).$$

Applying Lemma 3 gives

$$\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|<\delta\}}\omega_\Lambda(N_{k,\Lambda})\ge\rho-\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|\ge\delta\}}\frac{1}{e^{c_k(\Lambda)}-1}-\frac{4\lambda}{V^2}\sum_{\{k\in\Lambda^*:|k|\ge\delta\}}\omega_\Lambda(N_{j,\Lambda}N_{k,\Lambda})\,\frac{1}{1-e^{-c_\delta(\Lambda)}}. \tag{17}$$

Take $\epsilon>0$ arbitrary, and V large enough such that Lemma 4 is satisfied. This implies that

$$\frac{1}{V^2}\sum_{\{k\in\Lambda^*:|k|\ge\delta\}}\omega_\Lambda(N_{j,\Lambda}N_{k,\Lambda})\le\frac{1}{V^2}\,\omega_\Lambda(N_\Lambda N_{j,\Lambda})<\epsilon.$$

Hence taking V large enough, the second term on the rhs of (17) can be made arbitrarily close to

$$\int_{|k|\ge\delta}\frac{dk}{(2\pi)^d}\,\frac{1}{e^{\beta(\epsilon_k-\delta^2/8m)}-1},$$

whereas the third term is made arbitrarily small.

Hence in the limit $V\to\infty$, one gets

$$\lim_{V\to\infty}\frac{1}{V}\sum_{\{k\in\Lambda^*:|k|<\delta\}}\omega_\Lambda(N_{k,\Lambda})\ge\rho-\int_{|k|\ge\delta}\frac{dk}{(2\pi)^d}\,\frac{1}{e^{\beta(\epsilon_k-\delta^2/8m)}-1}.$$

Now take the limit $\delta\to 0$ to get *(i)*.

The function

$$\beta\mapsto f(\beta)=\int_{\mathbb{R}^d}\frac{dk}{(2\pi)^d}\,\frac{1}{e^{\beta\epsilon_k}-1}$$

is clearly decreasing in $\beta>0$ and, furthermore, $f(\beta)\to\infty$ for $\beta\to 0$, and $f(\beta)\to 0$ for $\beta\to\infty$. Hence, for every $\rho>0$ there exists $\beta_c>0$, defined by

$$\rho=\int_{\mathbb{R}^d}\frac{dk}{(2\pi)^d}\,\frac{1}{e^{\beta_c\epsilon_k}-1}.$$

Together with *(i)* this proves *(ii)*. □

Theorem 1 proves that the model (11) has the same pressure as the imperfect Bose gas. Theorems 2 and 3 prove that the model (11) shows a Bose–Einstein condensation exactly as the imperfect Bose gas, be it that the nature of the condensation is different. One aspect of this is that the ground state $(k=0)$ condensation of the imperfect Bose gas is unstable against any arbitrary small repulsive perturbation of the type $(\gamma/V)\Sigma_{k\in\Lambda*}N_{k,\Lambda}^2$, for any $\gamma>0$. The condensation becomes nonextensive. However, on the level of the thermodynamics the models are similar.

The natural question to ask is, whether the equilibrium states of the two models coincide. For the imperfect Bose gas, this problem is solved, e.g., in Ref. 10. We are not going into the details, but the technique of Ref. 10 can also be used in order to solve rigorously the equilibrium—or KMS—equations of our model. The result is that all equilibrium states are of the same type as the ones of the imperfect Bose gas. In particular, the equilibrium states are also integrals over a set of quasifree or generalized free states.

On the other hand, it is interesting to remark the following. Given this result, one might ask whether the variational principle of statistical mechanics, formulated in the thermodynamic limit, but restricted to the set of quasifree states, does also give the results of this paper, namely, the existence of condensation and the equilibrium states. Performing this program, one remarks that the particular type of condensation is not recovered by this method. Hence, for the time being, the only way to keep track of it is to follow closely the details of the thermodynamic limit, as is done above. In this work we illustrate clearly that care must be taken of this limit and that statistical mechanics remains the theory of really handling the thermodynamic limit.

ACKNOWLEDGMENT

The authors thank the referee for his careful reading of the paper, leading to a better version of it.

[1] P. C. Hohenberg, Phys. Rev. **158**, 383 (1967).
[2] V. E. Korepin, Commun. Math. Phys. **94**, 93 (1984).
[3] M. van den Berg, J. T. Lewis, and P. de Smedt, J. Stat. Phys. **37**, 697 (1984).
[4] T. C. Dorlas, J. T. Lewis, and J. V. Pulè, Commun. Math. Phys. **156**, 37 (1993).
[5] M. Schröder, J. Stat. Phys. **58**, 1151 (1990).
[6] J. V. Pulè and V. A. Zagrebnov, Ann. Inst. Henri Poincaré **59**, 421 (1993).
[7] M. van den Berg, J. T. Lewis, and J. V. Pulè, Helv. Phys. Acta **59**, 1271 (1986).
[8] M. van den Berg, J. T. Lewis, and M. Lunn, Helv. Phys. Acta **59**, 1289 (1986).
[9] M. van den Berg and J. T. Lewis, Physica A **110**, 550 (1982).
[10] M. Fannes and A. Verbeure, J. Math. Phys. **21**, 1809 (1980).
[11] M. van den Berg, J. T. Lewis, and J. V. Pulè, Commun. Math. Phys. **118**, 61 (1988).
[12] M. van den Berg, T. C. Dorlas, J. T. Lewis, and J. V. Pulè, Commun. Math. Phys. **127**, 41 (1990).
[13] M. Girardeau, J. Math. Phys. **1**, 516 (1960).
[14] M. Fannes and A. Verbeure, Commun. Math. Phys. **55**, 125 (1977).
[15] M. Fannes and A. Verbeure, Commun. Math. Phys. **57**, 165 (1977).

Extended supersymmetries for the Schrödinger–Pauli equation

J. Niederle[a)]
Institute of Physics, Academy of Sciences of the Czech Republic, Na Slovance 2, Prague 8, Czech Republic

A. G. Nikitin
Institute of Mathematics, National Academy of Sciences of the Ukraine, Tereshchenkivs'ka Street 3, Kiev-4, Ukraine

(Received 5 September 1997; accepted for publication 16 April 1998)

It is argued that extended, reducible, and generalized supersymmetry (SUSY) are common in many systems of standard nonrelativistic quantum mechanics. For example, it is proved that a well-studied quantum mechanical system of a spin-$\frac{1}{2}$ particle interacting with constant and homogeneous magnetic field admits the $N=4$ SUSY and has the internal symmetry so(3,3). Then an approach of energy spectra of a SUSY nature is presented and developed. It is applied to a wide class of systems described by the Schrödinger–Pauli equation admitting $N=3$, $N=4$, and $N=5$ SUSY. Some of these supersymmetries have a very peculiar property—their supercharges are realized without usual fermionic variables. It is shown that for them, the usual extension $N=3$ to $N=4$ SUSY is no longer guaranteed.
 [S0022-2488(99)00203-0]

I. INTRODUCTION

A beautiful and rich concept of supersymmetry (SUSY) has been introduced by several authors in various contexts (see Refs. 1 but also Refs. 2, where the idea of SUSY was formulated in a somewhat rudimental form). Since that time it has played a more and more important role in physics and mathematics, in general, and in modern particle physics and quantum mechanics,[3] in particular. This is due to the fact that SUSY presents a powerful tool for transforming bosons to fermions, and vice versa, for formulating theories with nontrivial unification of space–time and internal symmetries, for formulating string theories and their most powerful dualities (refer, e.g., to Refs. 4–6), for understanding the relations between spectra of different Hamiltonians as well as for explaining degeneracy of their spectra, for constructing exactly or quasiexactly solvable systems, for justifying formulations of initial and bound problems, etc.; see, e.g., surveys.[4,7,8]

In this work we shall concentrate on quantum mechanical systems since they provide a ground for testing the principal question: whether SUSY is realized in nature or not, free of the complexities of field theories. Examples of such systems (like interaction of a spin-$\frac{1}{2}$ particle with the Coulomb or constant and inhomogeneous magnetic field), which admit exact $N=2$ SUSY, are well known[9,10] (see also Refs. 7, 8 and references therein). Here we search for problems with *extended* ($N=3$ and $N=4$) SUSY.

In this connection let us remind you that the quantum mechanical models that include $N>2$ supercharges were investigated, e.g., in Refs. 11, and examples of quantum mechanical systems with extended SUSY were discussed in Refs. 12–16. In Refs. 17 the so-called ''generalized SUSY'' was proposed; it includes extended SUSY as a particular case. It was pointed out in Refs. 12–14 that some quantum mechanical models are invariant wrt reducible representations of SUSY algebra; we will refer to the related symmetry as ''reducible SUSY.''

Of course it is interesting to search for physical systems that admit exact (especially extended,

[a)]Electronic mail: niederle@fzu.cz

reducible, or generalized) SUSY. First, they bring additional indications that SUSY is indeed the symmetry of nature, and second, for such systems we have standard methods for their analysis at our disposal.

In fact, it will be shown in the present paper that the extended, reducible, and generalized SUSYs are common in many problems of standard nonrelativistic quantum mechanics. For example, we prove that the well-studied system of a spin-$\frac{1}{2}$ particle interacting with a constant and homogeneous magnetic field, which can be described by the Schrödinger–Pauli equation, admits $N=4$ SUSY and $N=2$ reducible SUSY as well.

In Sec. II we show that the extended, reducible, and generalized SUSY appear naturally in a wide class of problems of standard one-dimensional SUSY quantum mechanics. In Sec. III we consider the quantum mechanical system of a spin-$\frac{1}{2}$ particle interacting with a constant and homogeneous magnetic field and prove that it has $N=4$ extended SUSY. The reducible SUSY and so(3,3) symmetry of this model are discussed in Sec. III C.

In Sec. IV we search for extended and reducible SUSY of the Schrödinger–Pauli equation for a particle interacting with a static inhomogeneous magnetic field. We find a wide class of systems admitting these supersymmetries and discuss briefly their physical consequences.

II. ADDITIONAL EXTENDED AND REDUCIBLE SUSY OF SUPERSYMMETRIC QUANTUM MECHANICS

Supersymmetric quantum mechanical systems are described by the Schrödinger equation with a matrix potential,[3]

$$\hat{H}\psi \equiv \tfrac{1}{2}(p^2+W^2+\sigma_3 W')\psi = E\psi, \tag{2.1}$$

where $p=-i(\partial/\partial x)$, $W=W(x)$ is a superpotential, $W'=\partial W/\partial x$, and σ_3 is the Pauli matrix of the form

$$\sigma_3=\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

It is well known that Eq. (2.1) admits the following specific supersymmetries (supercharges):

$$Q_1=\frac{1}{\sqrt{2}}(\sigma_1 p+\sigma_2 W), \quad Q_2=\frac{1}{\sqrt{2}}(\sigma_2 p-\sigma_1 W), \tag{2.2}$$

which satisfy the superalgebra

$$\{Q_a,Q_b\}=2\delta_{ab}H, \quad [Q_a,H]=0, \tag{2.3}$$

where $a,b=1,2$ and $[\cdot,\cdot]$ and $\{\cdot,\cdot\}$ denote a commutator and anticommutator, respectively.

Let us demonstrate that, in addition to the transparent $N=2$ SUSY, Eq. (2.1) admits $N=3$ extended SUSY provided the corresponding superpotential $W(x)$ is an even function of x.

Proposition 1 (*Ref. 12*): Let $W(-x)=W(x)$, then there exists the third supercharge,

$$Q_3=i\sigma_1 R Q_1, \tag{2.4}$$

satisfying relations (2.3) for $a=1,2,3$, together with operators (2.2). Here R is defined by

$$R\psi(x)=\psi(-x). \tag{2.5}$$

Proof: The proposition can be proved by a simple direct calculation, taking into account the relations

$$[\sigma_1 R,Q_2]=\{\sigma_1 R,Q_1\}=0, \quad (\sigma_1 R)^2=1.$$

Thus, even the simplest SUSY model (2.1) can admit the extended SUSY generated by three supercharges.

Another interesting possibility is connected with the fact that the representation of superalgebra (2.2), (2.3) can be reducible. This occurs for the systems described by Eq. (2.1) with odd superpotentials.

Proposition 2 (*Refs. 12,13*): Let $W(-x)=-W(x)$; then the representation (2.2) of superalgebra (2.3) is reducible.

Proof: For odd $W(x)$ there exists an invariant operator, namely,

$$I=\sigma_3 R, \tag{2.6}$$

which commutes with any element of algebra (2.3).

Using the mapping $I\rightarrow I'=UIU^\dagger$, where

$$U=R_+-i\sigma_2 R_-, \quad R_\pm=\tfrac{1}{2}(1\pm R), \tag{2.7}$$

the operator (2.6) is transformed to the diagonal matrix,

$$I_3'=\sigma_3. \tag{2.8}$$

The corresponding transformed supercharges $Q_\alpha'=UQ_\alpha U'$ and the Hamiltonian $H'=UHU^\dagger$ commute with σ_3 and thus are diagonal too:

$$Q_\alpha'=\begin{pmatrix} q_+^\alpha & 0\\ 0 & q_-^\alpha\end{pmatrix}, \quad \hat{H}'=\begin{pmatrix}\hat{H}_+ & 0\\ 0 & \hat{H}_-\end{pmatrix}, \quad \alpha=1,2. \tag{2.9}$$

Here

$$q_\pm^1=iRP\pm W, \quad q_\pm^2=\pm p-iRW, \quad H_\pm=\tfrac{1}{2}(p^2+W^2\pm W'R). \tag{2.10}$$

Thus, supercharges $Q_{\alpha,\alpha}'=1,2$, are expressed as the direct sum of q_+^α and q_-^α Q.E.D.

Propositions 1 and 2 indicate how to find extended and reducible SUSY of realistic three-dimensional systems by first determining and then applying the appropriate involutive *discrete symmetries* (e.g., parities) of the system.

It is easy to see that the above-obtained extended and reducible supersymmetries imply the existence of some generalized ones,[17] i.e., supersymmetries satisfying the relations[17]

$$Q^2=\hat{H}, \quad \{I_a,Q\}=0, \quad I_a^2=1, \quad I_aI_b=\pm I_bI_a, \quad a=1,2,..., \tag{2.11}$$

where all involutions I_a either commute among themselves *or* anticommute.

Indeed, for even superpotentials there exist the anticommuting involutions $I_1=\sigma_3$ and $I_2=\sigma_1 R$, which, together with $Q=Q_1$, satisfy relations (2.11). In the case of odd superpotentials, relations (2.11) are satisfied by supercharge Q equal to Q_1, together with the commuting involutions $I_1=\sigma_3$ and $I_2=R$ (compare with Refs. 17).

In the systems analyzed later on we shall find their extended and reducible SUSY too. However, in contradistinction to the systems studied in the present section, the existence of these SUSYs will not imply the existence of the corresponding generalized SUSY.

III. SPIN-$\frac{1}{2}$ PARTICLE IN CONSTANT, HOMOGENEOUS MAGNETIC FIELD

A. Degeneracy of the spectrum of energy

Consider a quantum mechanical system consisting of the spin-$\frac{1}{2}$ charged particle interacting with a constant and homogeneous magnetic field. In a nonrelativistic approximation, this system is described by the Schrödinger–Pauli equation,

$$2mE\psi=\hat{H}\psi, \quad \hat{H}=\boldsymbol{\pi}^2-\tfrac{1}{2}eg\boldsymbol{\sigma}\cdot\mathbf{H}, \tag{3.1}$$

with

$$\pi_a=-i\frac{\partial}{\partial x_a}-eA_a, \quad a=1,2,3, \quad \boldsymbol{\sigma}=(\sigma_1,\sigma_2,\sigma_3), \quad g=2, \tag{3.2a}$$

and

$$A_1=eHx_2, \quad A_2=A_3=0, \quad e\mathbf{H}=i\boldsymbol{\pi}\times\boldsymbol{\pi}=e(0,0,H). \tag{3.2b}$$

Here σ_a are the Pauli matrices, and H is a constant characterizing the strength of the magnetic field $\mathbf{H}$.

The system (3.1) is exactly solvable.[18] The corresponding eigenvalues of energy E (Landau levels) are given by

$$2mE=2neH+p_3^2, \quad n=0,1,2,\dots. \tag{3.3}$$

For any $n\neq 0$ there exist two independent eigenfunctions (see, e.g., Ref. 19):

$$\begin{aligned}\psi_{1,p_1,p_3}&=\exp(ip_1x_1+ip_3x_3)\exp(-y^2/2)\begin{pmatrix}H_n(y)\\ H_{n-1}(y)\end{pmatrix},\\ \psi_{2,p_1,p_3}&=\exp(ip_1x_1+ip_3x_3)\exp(-y^2/2)\begin{pmatrix}H_n(y)\\ -H_{n-1}(y)\end{pmatrix},\end{aligned} \tag{3.4}$$

with H_n being Hermite polynomials, $H_{-1}=0$ and

$$y=\sqrt{eH}x_2-\frac{p_1}{\sqrt{eH}}. \tag{3.5}$$

For $n=0$ the eigenfunctions ψ_1 and ψ_2 coincide.

Thus, any energy levels (except the ground one) are two-fold degenerate due to the $N=2$ SUSY of Eq. (3.1). Moreover, there exists the infinite degeneracy of any energy level due to independence of E on p_1.[18]

In spite of the fact that symmetries and supersymmetries of Eq. (3.1) have been studied quite intensively (see, e.g., Refs. 6–8, 20, 21), we shall find a new, additional (extended) SUSY for this equation.

Starting with (3.4) and taking into account the quadratic dependence of energy E on p_3 and independence of E on p_1, we can write, for instance, six additional solutions corresponding to the same energy (3.3), namely

$$\begin{aligned}&\psi_{3,p_1,p_3}=\psi_{1,-p_1,p_3}, \quad \psi_{4,p_1,p_3}=\psi_{1,-p_1,-p_3}, \quad \psi_{5,p_1,p_3}=\psi_{1,p_1,-p_3},\\ &\psi_{6,p_1,p_3}=\psi_{2,-p_1,p_3}, \quad \psi_{7,p_1,p_3}=\psi_{2,-p_1,-p_3}, \quad \psi_{8,p_1,p_3}=\psi_{1,p_1,-p_3}.\end{aligned} \tag{3.6}$$

A bit surprisingly, the corresponding eight-fold degeneracy of energy levels can be interpreted as caused by $N=4$ *extended SUSY* of the system (3.1).

B. Extended SUSY

It is well known that whenever the gyromagnetic ratio g of the particle is equal to 2, Eq. (3.1) admits $N=2$ SUSY.[10] Here we demonstrate that this SUSY is reducible and that there exists a more extended, namely, $N=4$ SUSY for (3.1) in addition.

A standard supercharge for Eq. (3.1) has the form[4]

$$Q_1 = \boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_1^2 = \hat{H}. \tag{3.7}$$

The remaining three additional supercharges can be constructed using the fact that (3.1) is invariant wrt the following three discrete transformations:

$$\psi \to iR_3\psi, \quad \psi \to CR_1\psi, \quad \psi \to CR_2\psi, \tag{3.8}$$

where R_a $(a=1,2,3)$ are the space reflection transformations,

$$R_a\psi(\mathbf{x}) = \sigma_a\theta_a\psi(\mathbf{x}), \quad \theta_a\psi(\mathbf{x}) = \psi(r_a\mathbf{x}). \tag{3.9a}$$

Here

$$r_1\mathbf{x} = (-x_1, x_2, x_3), \quad r_2\mathbf{x} = (x_1, -x_2, x_3), \quad r_3\mathbf{x} = (x_1, x_2, -x_3), \tag{3.9b}$$

and $C = i\sigma_2 c$, where c is the operator of complex conjugation,

$$c\psi(\mathbf{x}) = \psi^*(\mathbf{x}). \tag{3.10}$$

Note that operators defined in (3.8)–(3.10) satisfy the following relations:

$$\begin{gathered}\{R_a, \boldsymbol{\sigma}\cdot\boldsymbol{\pi}\} = \{R_a, C\} = \{CR_1, \boldsymbol{\sigma}\cdot\boldsymbol{\pi}\} = \{CR_2, \boldsymbol{\sigma}\cdot\boldsymbol{\pi}\} = 0,\\ R_a^2 = -C^2 = 1, \quad a = 1,2,3.\end{gathered} \tag{3.11}$$

Thus, using (3.7), (3.11) we can see that the operators

$$Q_1 = \boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2 = iR_3\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_3 = CR_1\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_4 = CR_2\boldsymbol{\sigma}\cdot\boldsymbol{\pi} \tag{3.12}$$

fulfill the following relations:

$$\{Q_k, Q_l\} = 2g_{kl}\hat{H}, \quad [Q_k, \hat{H}] = 0, \tag{3.13}$$

where $k,l = 1,2,3,4$, $g_{11} = g_{22} = -g_{33} = -g_{44} = 1$; $g_{kl} = 0$, $k \neq l$. In other words, operators (3.12) are supercharges generating the $N=4$ extended SUSY of Eq. (3.1).

We notice that choosing the basis

$$\hat{Q}_1 = \frac{1}{\sqrt{2}}(Q_1 + Q_3), \quad \hat{Q}_2 = \frac{1}{\sqrt{2}}(Q_2 + Q_4), \quad \hat{Q}_1^\dagger = \frac{1}{\sqrt{2}}(Q_1 - Q_3), \quad \hat{Q}_2^\dagger = \frac{1}{\sqrt{2}}(Q_2 - Q_4), \tag{3.14}$$

it is possible to represent the commutation and anticommutation relations (3.13) in a more familiar form,

$$\{\hat{Q}_\alpha, \hat{Q}_\beta\} = 0, \quad \{\hat{Q}_\alpha, \hat{Q}_\beta^\dagger\} = 2\delta_{\alpha\beta}\hat{H},$$

$$[\hat{Q}_\alpha, \hat{H}] = 0, \quad \alpha, \beta, = 1,2.$$

Thus, we have proved that the well-known $N=2$ SUSY of Eq. (3.1) can be extended to $N=4$ SUSY, taking into account involutive symmetries (3.8). Acting by supercharges (3.12) on standard solutions (3.4) we obtain the set of eight linearly independent solutions (3.4), (3.6). The interpretation of the corresponding eight-fold degeneracy is given in the next section.

C. Internal symmetries and reducible SUSY

A direct consequence of the $N=4$ SUSY is a specific four-fold degeneracy of the corresponding nonground states.[11] However, we have shown that system (3.1) has eight-fold degeneracy. Let us demonstrate that this is due to the existence of a special internal symmetry algebra. This algebra appears as follows.

First, for any nonzero eigenvalue E of Hamiltonian (3.1) we can choose the following set of symmetry operators:

$$S_{6k}=\frac{1}{2\sqrt{E}}Q_k, \quad S_{65}=\frac{1}{2}R_3, \quad S_{mn}=[S_{6m},S_{6n}]. \tag{3.15}$$

Here Q_k and R_3 are operators defined in (3.9), (3.12), $k=1,2,3,4$ and $m,n=1,2,3,4,5$. However, there exists an additional symmetry operator, namely, the operator

$$I_1=i(\sigma_1\pi_2-\sigma_2\pi_1)p_3R_3, \tag{3.16}$$

which commutes with any of the operators (3.15).

Thus, taking into account that operators S_{6n} form the Clifford algebra

$$\{2S_{6n},2S_{6m}\}=2g_{mn}, \tag{3.17}$$

with nonzero components of g_{mn} being $g_{11}=g_{22}=-g_{33}-g_{44}=g_{55}=1$, we can easily find that the symmetry operators (3.15) and (3.16) satisfy the following commutation relations:

$$[S_{kl},S_{mn}]=g_{kl}S_{ln}+g_{ln}S_{kl}-g_{kn}S_{lm}-g_{lm}S_{kn}, \tag{3.18a}$$

$$[S_{kl},I_1]=0 \tag{3.18b}$$

(with $k,l,m,n=1,2,3,4,5,6$ and $g_{66}=-1$), i.e., they form the central extension of Lie algebra $so(3,3)$ by I_1.[22] Its invariant operators are given by

$$C_1=\tfrac{1}{2}S_{kl}S^{kl}\equiv\tfrac{15}{4}, \quad C_2=\tfrac{1}{2}S_{kl}S^{ln}S_{nf}S^{fk}\equiv\tfrac{315}{16},$$
$$C_3=\frac{1}{6!}\epsilon_{mnrslk}S^{mn}S^{rs}S^{lk}\equiv\frac{1}{8}R_1R_2, \quad C_4=I_1. \tag{3.19}$$

Using (3.11), (3.12) it is easy to show that the eigenvalues of operators C_3 and C_4 are $\pm\frac{1}{8}$ and $\pm p_3\sqrt{2neH}$, respectively. Four possible combinations of these eigenvalues specify four orthogonal subspaces of the solutions of Eq. (3.1) for p_3 and n different from zero. Thus, operators (3.15) appear to realize a reducible representation of $so(3,3)$, namely, the direct sum of irreducible representations $2D(\frac{1}{2}\,\frac{1}{2}\,\frac{1}{2})\oplus 2D(\frac{1}{2}\,\frac{1}{2}\,-\frac{1}{2})$ of the algebra so(3,3).[23] The corresponding representation space is 16-dimensional over $\mathbb{R}$, so effectively we have the eight-fold degeneracy over the field of complex numbers.

Restricting ourselves to linear symmetries (i.e., to those including no antilinear operator of complex conjugation), $N=4$ SUSY is reduced to $N=2$ SUSY, which is generated by supercharges Q_1 and Q_2 specified in (3.12). However, this SUSY is reducible since there exist two linear symmetries for (3.1), namely, C_3 and C_4 (3.19), which are involutive up to constant factors and commute with supercharges Q_1 and Q_2:

$$[C_3,Q_a]=[I_1,Q_a]=0, \quad [I_1,C_3]=0, \quad a=1,2. \tag{3.20}$$

Analogously, as in the proof of Proposition 2 we can diagonalize C_3 and C_4 and reduce each of supercharges Q_1, Q_2 to a direct sum of four supercharges. This yields four invariant spaces $\Phi^{(\alpha)}$ ($\alpha=1,2,3,4$) of supercharges Q_1 and Q_2 with basis elements $\Phi_1^{(\alpha)}$, $\Phi_2^{(\alpha)}$, where

$$\Phi_1^{(1)}=\psi_{1,p_1,p_3}+\psi_{1,-p_1,p_3}+i\psi_{1,p_1,-p_3}+i\psi_{1,-p_1,-p_3},$$
$$\Phi_2^{(1)}=\psi_{2,p_1,p_3}+\psi_{2,-p_1,p_3}-i\psi_{2,p_1,-p_3}-i\psi_{2,-p_1,-p_3}; \tag{3.21}$$

$$\Phi_1^{(2)}=\psi_{1,p_1,-p_3}+\psi_{1,-p_1,-p_3}+i\psi_{1,p_1,p_3}+i\psi_{1,-p_1,p_3},$$
$$\Phi_2^{(2)}=\psi_{2,p_1,-p_3}+\psi_{2,-p_1,-p_3}-i\psi_{2,p_1,p_3}-i\psi_{2,-p_1,p_3}; \tag{3.22}$$

$$\Phi_1^{(3)}=-\psi_{1,p_1,p_3}+\psi_{1,-p_1,p_3}-i\psi_{1,p_1,-p_3}+i\psi_{1,-p_1,-p_3},$$
$$\Phi_2^{(3)}=-\psi_{2,p_1,p_3}+\psi_{2,-p_1,p_3}+i\psi_{2,p_1,-p_3}-i\psi_{2,-p_1,-p_3}; \tag{3.23}$$

$$\Phi_1^{(4)}=-\psi_{1,p_1,-p_3}+\psi_{1,-p_1,-p_3}-i\psi_{1,p_1,p_3}-i\psi_{1,-p_1,p_3},$$
$$\Phi_1^{(4)}=-\psi_{2,p_1,-p_3}+\psi_{2,-p_1,-p_3}+i\psi_{2,p_1,p_3}-i\psi_{1,-p_1,p_3}; \tag{3.24}$$

where ψ_{1,p_1,p_3} and ψ_{2,p_1,p_3} are functions defined in (3.4).

IV. SPIN-$\frac{1}{2}$ PARTICLE IN A STATIC, NONHOMOGENEOUS MAGNETIC FIELD

A. Extended SUSY

In this section we shall show that the system of a spin-$\frac{1}{2}$ particle interacting with various magnetic fields has extended SUSY too, provided the external magnetic field has definite parities.

Starting with reflections (3.9b) we find that the corresponding parity properties of vector function $A(x)$ (3.2b) are of the form

$$\mathbf{A}(r_1\mathbf{x})=-r_1A(\mathbf{x}),\quad \mathbf{A}(r_2\mathbf{x})=-r_2\mathbf{A}(\mathbf{x}),\quad \text{and}\ \mathbf{A}(r_3\mathbf{x})=r_3\mathbf{A}(\mathbf{x}). \tag{4.1}$$

Relations (4.1) are satisfied by a large class of potentials, which includes (3.2b) as a particular case. For all such potentials the corresponding equation (3.1) is invariant wrt involutions (3.8), and so admits the extended SUSY generated by supercharges (3.12). Moreover, Eq. (3.1) for $g=2$ and an arbitrary uniform magnetic field, i.e., the field

$$A_1=A_1(x_1,x_2),\quad A_2=A_2(x_1,x_2),\quad A_3=0, \tag{4.2}$$

admits all internal symmetries described in Sec. III B, provided $\mathbf{A}(\mathbf{x})$ satisfies relations (4.1).

Other systems with extended SUSY can be found by extending reflections (3.9b) to the eight-dimensional group of involutions, i.e., by adding the transformations

$$r_{12}\mathbf{x}=(-x_1,-x_2,x_3),\quad r_{31}\mathbf{x}=(-x_1,x_2,-x_3),\quad r_{23}\mathbf{x}=(x_1,-x_2,-x_3),$$
$$r_{123}\mathbf{x}=(-x_1,-x_2,-x_3),\quad I\mathbf{x}=\mathbf{x}, \tag{4.3}$$

to reflections (3.9b).

We notice that r_a $(a=1,2,3)$ and r_{123} are reflections while r_{ab} $(a,b=1,2,3)$ are rotations.

Let us suppose now that the vector potential $\mathbf{A}(\mathbf{x})$ has definite parities wrt a subset of transformations (3.9b) and (4.3). All possible transformations for the vector potential with definite parities wrt (9.3b) and (4.3), which are compatible with the Lorentz gauge $\mathbf{p}\cdot\mathbf{A}=0$, can be expressed as follows:

$$\mathbf{A}(r_{ab}\mathbf{x})=r_{ab}\mathbf{A}(\mathbf{x}),\quad a,b,=1,2,3, \tag{4.4a}$$

$$\mathbf{A}(r_a\mathbf{x})=r_a\mathbf{A}(\mathbf{x}), \tag{4.4b}$$

$$\mathbf{A}(r_{123}\mathbf{x}) = -\mathbf{A}(\mathbf{x}), \tag{4.4c}$$

and

$$\mathbf{A}(r_{ab}\mathbf{x}) = -r_{ab}\mathbf{A}(\mathbf{x}), \tag{4.5a}$$

$$\mathbf{A}(r_a\mathbf{x}) = -r_a\mathbf{A}(\mathbf{x}), \tag{4.5b}$$

$$\mathbf{A}(r_{123}\mathbf{x}) = \mathbf{A}(\mathbf{x}). \tag{4.5c}$$

It is easy to see that whenever $\mathbf{A}(\mathbf{x})$ transforms according to one of the relations (4.4a)–(4.4c) or (4.5a)–(4.5c) (for fixed values of indices a,b) the equation (3.1) remains invariant wrt this transformation provided $\psi(\mathbf{x})$ cotransforms accordingly, i.e., via the relations

$$\psi(\mathbf{x}) \to iR_aR_b\psi(\mathbf{x}) \equiv R_{ab}\psi(\mathbf{x}), \tag{4.6a}$$

$$\psi(\mathbf{x}) \to R_a\psi(\mathbf{x}), \tag{4.6b}$$

$$\psi(\mathbf{x}) \to R_1R_2R_3\psi(\mathbf{x}) \equiv R\psi(\mathbf{x}), \tag{4.6c}$$

or

$$\psi(\mathbf{x}) \to i\sigma_2 cR_a\psi(\mathbf{x}) \equiv CR_{ab}\psi(\mathbf{x}), \tag{4.7a}$$

$$\psi(\mathbf{x}) \to i\sigma_2 cR_a\psi(\mathbf{x}) \equiv CR_a\psi(\mathbf{x}), \tag{4.7b}$$

$$\psi(\mathbf{x}) \to i\sigma_2 cR\psi(\mathbf{x}) \equiv CR\psi(\mathbf{x}), \tag{4.7c}$$

respectively. Here R_a and c are the operators introduced in (3.9), (3.10).

Transformations (4.6b)–(4.7c) are involutions anticommuting with $\boldsymbol{\sigma}\cdot\boldsymbol{\pi}$, so yielding $N=2$ SUSY with supercharges given by

$$Q_1 = \boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2 = i\hat{R}\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \tag{4.8}$$

where $\hat{R}$ denotes the relevant operators from (4.6b)–(4.7c) [i.e., for the symmetry (4.4b) the operator $\hat{R} = R_{a'}$, for (4.4c) the operator $\hat{R} = R$, etc.].

More complicated cases, in which the vector $\mathbf{A}(\mathbf{x})$ has the definite transformation properties wrt combined parities, can be discussed analogously. First, using the group properties of involutions (3.9b), (4.3), it is easy to show that whenever $\mathbf{A}(\mathbf{x})$ has definite parities wrt two of these involutions, it has also the definite parity wrt their product. Requiring definite parities wrt various triplets of involutions enumerated in (3.9b), (4.3), we receive cases that are either equivalent to those with the definite transformation properties wrt doublets of parities or wrt all eight involutions (3.9b), (4.3).

If $\mathbf{A}(\mathbf{x})$ satisfies two compatible relations from (4.4) and (4.5) simultaneously, then Eq. (3.1) with $g=2$ admits $N=2$ or $N=3$ SUSY. All these nonequivalent possibilities are enumerated in the Appendix. Here we consider only one example, namely, when the vector potential has the property

$$\mathbf{A}(r_1\mathbf{x}) = r_1\mathbf{A}(\mathbf{x}), \quad \mathbf{A}(r_2\mathbf{x}) = r_2\mathbf{A}(\mathbf{x}), \tag{4.9}$$

but has no definite parity wrt reflection r_3. The related supercharges are of the form

$$Q_1 = \boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2 = iR_1\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \qquad \text{and} \quad Q_3 = iR_2\boldsymbol{\sigma}\cdot\boldsymbol{\pi}. \tag{4.10}$$

They satisfy relations (2.3) [where $\hat{H}$ is Hamiltonian (3.1), $a,b=1,2,3$] and thus generate the symmetry algebra equal to $N=3$ SUSY for the system. This SUSY causes a four-fold degeneracy of the corresponding (nonground) energy levels, since for any nonzero E it yields the four-dimensional representation $D(\frac{1}{2}\,\frac{1}{2})\oplus D(\frac{1}{2}-\frac{1}{2})$ of the Lie algebra $so(4)$ generated by

$$S_{4a}=\frac{1}{2\sqrt{E}}Q_a, \quad S_{ab}=-i[S_{4a},S_{4b}]. \tag{4.11}$$

The most extended, $N=4$ and $N=5$ SUSY appears in the cases for which the vector potential has definite parities wrt all involutions (3.9b), (4.3). In addition to (4.1), there are three more possible transformation properties of $\mathbf{A}(\mathbf{x})$:

$$\mathbf{A}(r_a\mathbf{x})=\mathbf{A}(\mathbf{x}), \quad a=1,2,3, \tag{4.12}$$

$$\begin{gathered}\mathbf{A}(r_a\mathbf{x})=\mathbf{A}(\mathbf{x}),\ \ \mathbf{A}(r_b\mathbf{x})=\mathbf{A}(\mathbf{x}),\ \ \mathbf{A}(r_c\mathbf{x})=-\mathbf{A}(\mathbf{x}),\\ a\neq b, \quad b\neq c, \quad c\neq a, \quad c \ \text{ is fixed,}\end{gathered} \tag{4.13}$$

and

$$\mathbf{A}(r_a\mathbf{x})=-\mathbf{A}(\mathbf{x}), \quad a=1,2,3. \tag{4.14}$$

They allow us to construct the corresponding supercharges, namely,

$$Q_1=iR_1\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2=iR_2\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_3=iR_3\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_4=\boldsymbol{\sigma}\cdot\boldsymbol{\pi}; \tag{4.15}$$

$$Q_0=CR_c\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_1=\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2=iR_a\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_3=iR_b\boldsymbol{\sigma}\cdot\boldsymbol{\pi}; \tag{4.16}$$

and

$$Q_1=CR_1\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2=CR_2\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_3=CR_3\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_4=CR\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_5=\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \tag{4.17}$$

for the cases (4.12), (4.13), and (4.14), respectively.

Operators (4.15) and Hamiltonian (3.1) satisfy relations (2.3) for $a,b=1,2,3,4$ and thus generate $N=4$ extended SUSY. The corresponding internal symmetries reduce to the algebra $so(5)$ whose basis elements [constructed analogously to (4.11)] generate the four-dimensional irreducible representation $D(1/2\,1/2\,1/2)$. Thus, for the system (3.1), (3.2a) we can expect a four-fold degeneracy of nonground energy levels whenever the vector potential of an external field satisfies the relations (4.12).

The operators in (4.16) and (4.17) satisfy the relations (3.13) for $g_{00}=-g_{11}=-g_{22}=-g_{33}=1$ and $-g_{11}=-g_{22}=-g_{33}=g_{44}=g_{55}=1$, respectively, and thus generate $N=4$ and $N=5$ SUSY.

B. Reducible SUSY

In this section the involutions (4.6), (4.7) are used to find out reducible SUSY for systems described by the equations (3.1) with $g=2$ and vector potentials $\mathbf{A}(\mathbf{x})$.

First, let us assume that the parity properties of the vector potential are specified by relations (4.12) and (4.13). Then the corresponding equation (3.1) admits $N=4$ SUSY. Moreover, there exist the involutions

$$I=R_{23} \tag{4.18}$$

and

$$I^{(c)}=CR_{ab}, \quad a,b\neq c, \tag{4.19}$$

commuting with pairs of supercharges Q_1, Q_2 from (4.15) and Q_0, Q_1 from (4.16), respectively, so that the corresponding $N=2$ SUSY is reducible.

If parities of the vector potential are specified by relations (4.14), then there exists the involution

$$I=R_{23}, \tag{4.20}$$

which commutes with a triplet of supercharges, namely, with supercharges Q_1, Q_2, and Q_3 of (4.16). Consequently, the related equation (3.1) admits $N=3$ reducible SUSY.

If the vector potential satisfies all relations (4.1) then there exists the involution

$$I=R_{12}, \tag{4.21}$$

commuting with all four supercharges (3.12), and so the corresponding system has $N=4$ reducible SUSY.

Indeed, diagonalizing involutions (4.18)–(4.21), the corresponding supercharges are transformed to block diagonal forms. For instance, for involutions (4.18) and supercharges (4.15), we obtain

$$I\rightarrow UIU^{\dagger}=\sigma_3, \quad Q_a\rightarrow UQ_aU^{\dagger}=\tfrac{1}{2}(1+\sigma_3)Q_a^{+}+\tfrac{1}{2}(1-\sigma_3)Q_a^{-}, \quad a=1,2, \tag{4.22}$$

where

$$U=\frac{1}{\sqrt{2}}(1+\sigma_3 I), \quad U^{\dagger}=\frac{1}{\sqrt{2}}(1-\sigma_3 I), \tag{4.23}$$

and

$$Q_1^{+}=(\pi_1-i\pi_2)\theta_{23}+\pi_3, \quad Q_2^{+}=(i\pi_1+\pi_2)\theta_1+i\pi_3\theta_{123}, \tag{4.24}$$

$$Q_1^{-}=(-\pi_1-i\pi_2)\theta_{23}-\pi_3, \quad Q_2^{+}=(i\pi_1-\pi_2)\theta_1-i\pi_3\theta_{123}, \tag{4.25}$$

with $\theta_{ab}=\theta_a\theta_b$, $\theta_{123}=\theta_1\theta_2\theta_3$, and operators θ_a defined in (3.9a).

The operators (4.23), together with

$$\hat{H}=\hat{H}^{+}=\pi^2-2e[H_3+(iH_2-H_1)\theta_{23}], \tag{4.26}$$

form superalgebra (2.3), while operators (4.24) satisfy (2.3) with the Hamiltonian of the form

$$\hat{H}=\hat{H}^{-}=\pi^2+2e[H_3-(H_1+iH_2)\theta_{23}]. \tag{4.27}$$

Here H_1, H_2, and H_3 denote the components of the magnetic field strength.

The supercharges generating reducible SUSY for other systems described by (3.1) can be diagonalized in a similar way. The explicit form of the corresponding transformation operators is given in the Appendix.

Let us note that supercharges (4.23), (4.24) depend on three variables x_1,x_2,x_3 and have a very peculiar property: they include no fermionic variables.

V. DISCUSSION

In this article we have described the approach for a systematical study of quantum systems whose symmetry group includes extended SUSY and whose degeneracy of energy spectra is of a SUSY nature.

In Sec. IV, requiring definite parity properties of the vector potential, we find a number of quantum mechanical systems with $N=3$, $N=4$, and $N=5$ SUSY.

It is necessary to stress that there exist a lot of realistic physical systems whose parities satisfy required relations (4.1), (4.12)–(4.14). In addition to the vector potential of the constant magnetic field, given by relations (3.2b), we present here as examples the potential of an infinite straight conductor with the constant current I directed along the third coordinate axis,

$$A_1=A_2=0, \quad A_3=-\frac{I}{2\pi}\ln(x_1^2+x_2^2), \tag{5.1}$$

superpositions of potentials (5.1) that are generated by two or four infinite straight conductors shifted by distance $2b$ (two neighbor currents have opposite directions),

$$A_1=A_2=0, \quad A_3=-\frac{I}{4\pi}\ln\left[\frac{(x_1-b)^2+x_2^2}{(x_1+b)^2+x_2^2}\right], \tag{5.2}$$

$$A_1=A_2=0, \quad A_3=-\frac{I}{4\pi}\ln\frac{[(x_1+b)^2+(x_2+b)^2][(x_1-b)+(x_2-b)^2]}{[(x_1+b)^2+(x_2-b)^2][(x_1-b)^2+(x_2+b)^2]}, \tag{5.3}$$

and the magnetic octopole potential,[16]

$$A_1=\frac{a^2m}{4\pi}\,\frac{x_1(x^2-x_2^2)}{x^7}, \quad A_2=\frac{a^2m}{4\pi}\,\frac{x_2(x^2-x_1^2)}{x^7}, \quad A_3=0. \tag{5.4}$$

Parities of potentials (5.1), (5.2), (5.3), and (5.4) are given by relations (4.1), (4.13), (4.14), and (4.12), respectively.

Moreover, analyzing various superpositions of magnetic dipole and straight conductor potentials, it is possible to generate models of physical systems with any parity properties enumerated in (A1)–(A3), (A8)–(A12).

The very existence of such systems presents a strong indication that the extended SUSY is indeed realized in nature. Moreover, knowledge of extended SUSY for systems described by the Schrödinger–Pauli equation enables us to predict the specific degeneracy of the corresponding energy levels. This degeneracy can be removed by adding a small symmetry-breaking term corresponding, e.g., to the interaction with a weak external electric field and thus experimentally verified.

We did not discuss the question of whether the found extended SUSY is exact or broken. To this end it is necessary to analyze degeneracy of the ground state of the considered systems. For two-dimensional quantum systems, such an analysis was made in Refs. 24 and 25.

Our approach to extended SUSY can be compared with that using generalized SUSY[17] whenever all supercharges of the considered systems are linear operators (i.e., not including complex conjugation) and can be constructed starting with involutions satisfying (2.11). Since for some of our systems the corresponding supercharges include the antilinear operator of complex conjugation, the above-mentioned correspondence does not exist [in this case (2.11) does not hold]. Consequently, our approach covers more general situations than the approach proposed in Ref. 17.

It is well known that $N=3$ SUSY can always be extended to that of $N=4$[26] for systems in which SUSY is realized by Grassmanian variables. In our paper we show that such an extension of odd N SUSY to even one is not guaranteed in general.

Analogous to the above-mentioned cases with time-independent magnetic fields, it is possible to search for systems with extended SUSY described by the Schrödinger–Pauli equation with a time-dependent magnetic field. The case with $N=2$ SUSY was discussed in Ref. 27.

We notice that our approach can be extended to the relativistic Dirac equation with a similar result (for particular examples, see Refs. 13 and 14). However, Dirac's equation admits an extended SUSY also for the cases with external electric fields and scalar potentials.[15]

Another intriguing problem is to generalize the above results for particles with spin $s > \frac{1}{2}$. This can be done, e.g., in the framework of the weak SUSY approach.[28]

ACKNOWLEDGMENTS

This publication is based on work sponsored by the Grant Agency of the Academy of Sciences of the Czech Republic under Project No. A1010711. One of us (AGN) was supported by the Ukrainian DFFD foundations, Project No. 1.4/356.

APPENDIX: COMBINED PARITIES, SUPERCHARGES, AND EXPLICIT REDUCTIONS

Here explicit forms of supercharges are presented for the cases when $\mathbf{A}(\mathbf{x})$ satisfies all possible combinations of relations (4.6), (4.7).

First, we shall consider systems with $N=2$ SUSY. They correspond to the following parity properties of the electromagnetic field:

$$\mathbf{A}(r_a\mathbf{x}) = r_a\mathbf{A}(\mathbf{x}), \quad \mathbf{A}(r_{bc}\mathbf{x}) = r_{bc}\mathbf{A}(\mathbf{x}), \tag{A1}$$

$$\mathbf{A}(r_a\mathbf{x}) = -r_a\mathbf{A}(\mathbf{x}), \quad \mathbf{A}(r_{bc}\mathbf{x}) = r_{bc}\mathbf{A}(\mathbf{x}), \tag{A2}$$

and

$$\mathbf{A}(r_a\mathbf{x}) = -r_a\mathbf{A}(\mathbf{x}), \quad \mathbf{A}(r_{bc}\mathbf{x}) = -r_{bc}\mathbf{A}(\mathbf{x}), \tag{A3}$$

where $b,c \neq a$, a is fixed.

The related supercharges have the form

$$Q_1 = \boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2 = iR_a\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \tag{A4}$$

for parity properties (A1), and

$$Q_1 = \boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \quad Q_2 = CR_a\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, \tag{A5}$$

for the cases when $\mathbf{A}(\mathbf{x})$ satisfies (A2) or (A3).

In all these cases the corresponding $N=2$ SUSY is reducible. The involutions commuting with supercharge (A4) and (A5) have the form

$$I_1 = R_{bc} \quad \text{and} \quad I_2 = R_{bc}, \quad I_3 = CR_{bc},$$

respectively. Particular cases of these involutions are expressed in the formulas (4.18)–(4.21).

The operators diagonalizing both I_1 and I_2 have the form

$$U = \frac{1}{2}(1 - i\sigma_2)(1 + i\sigma_2\theta_{12}), \quad \text{for } a=3; \quad U = \frac{1}{\sqrt{2}}(1 + \sigma_3 I_1), \quad \text{for } a \neq 3, \tag{A6}$$

whereas the expressions for the operators diagonalizing I_3 are

$$\begin{aligned} U = U_1 &= \tfrac{1}{2}(1 - i\sigma_2)(1 + i\sigma_2\theta_{23}), \quad \text{for } a=1, \\ U = U_2 &= \tfrac{1}{2}(1 + C)(1 - i\sigma_1\theta_{31}), \quad \text{for } a=2, \\ U = U_3 &= \frac{1}{\sqrt{2}}(1 + \sigma_3 I_3), \quad \text{for } a=3. \end{aligned} \tag{A7}$$

Now we shall present systems with $N=3$ SUSY. In addition to (4.11), we have the following nonequivalent combinations of parity properties:

$$\mathbf{A}(r_{12}\mathbf{x})=r_{12}\mathbf{A}(\mathbf{x}),\quad \mathbf{A}(r_{23}\mathbf{x})=r_{23}\mathbf{A}(\mathbf{x}),$$
$$Q_1=R_{23}\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_2=R_{31}\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_3=R_{12}\boldsymbol{\sigma}\cdot\boldsymbol{\pi}; \tag{A8}$$

$$\mathbf{A}(r_a\mathbf{x})=r_a\mathbf{A}(\mathbf{x}),\quad \mathbf{A}(r_{bc}\mathbf{x})=-r_{bc}\mathbf{A}(\mathbf{x}),$$
$$Q_1=i\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_2=iR_a\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_3=CR_{bc}\boldsymbol{\sigma}\cdot\boldsymbol{\pi}; \tag{A9}$$

$$\mathbf{A}(r_a\mathbf{x})=-r_a\mathbf{A}(\mathbf{x}),\quad \mathbf{A}(r_b\mathbf{x})=-r_b\mathbf{A}(\mathbf{x}),$$
$$Q_0=\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_1=CR_a\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_2=CR_b\boldsymbol{\sigma}\cdot\boldsymbol{\pi}; \tag{A10}$$

$$\mathbf{A}(r_a\mathbf{x})=r_a\mathbf{A}(\mathbf{x}),\quad \mathbf{A}(r_b\mathbf{x})=-r_b\mathbf{A}(\mathbf{x}),$$
$$Q_0=CR_b\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_1=\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_2=iR_a\boldsymbol{\sigma}\cdot\boldsymbol{\pi}; \tag{A11}$$

$$\mathbf{A}(r_{ab}\mathbf{x})=r_{ab}\mathbf{A}(\mathbf{x});\quad \mathbf{A}(r_{bc}\mathbf{x})=-r_{bc}\mathbf{A}(\mathbf{x}),$$
$$Q_0=iR_{ab}\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_1=CR_{bc}\boldsymbol{\sigma}\cdot\boldsymbol{\pi},\quad Q_2=CR_{ac}\boldsymbol{\sigma}\cdot\boldsymbol{\pi}. \tag{A12}$$

The supercharges in (A8) satisfy the relations (2.3) for $a,b=1,2,3$; the supercharges in (A9), (A10) and (A11), (A12) satisfy the relations (3.13) for $g_{11}=-g_{22}=-g_{33}=1$ and $-g_{11}=g_{22}=g_{33}=1$, respectively.

The supercharges in (A10) commute with the involution R_{ab} and thus generate the $N=3$ reducible SUSY. The supercharges Q_1 and Q_2 in (A12) commute with this involution too and generate the $N=2$ reducible SUSY. The remaining supercharges, i.e., those in (A8), (A9), and (A11) are irreducible.

[1] Yu. A. Gol'fand and E. P. Lichtman, JETP Lett. **13**, 452 (1971); D. V. Volkov and V. P. Akulov, Phys. Lett. B **46**, 109 (1973); J. Wess and B. Zumino, Nucl. Phys. B **70**, 39 (1974).

[2] H. J. Lipkin, Phys. Lett. **9**, 203 (1964); J. Schwinger, Phys. Rev. **152**, 1219 (1966); G. L. Stawraki, in *High Energy Physics and the Theory of Elementary Particles*, Naukova Dumka, Kiev, 1966, p. 296 (in Russian); preprint ITP 67-21, Kiev, 1967; H. Migazawa, Prog. Theor. Phys. **36**, 1266 (); Phys. Rev. **170**, 1586 (1968); M. Flato and P. Hillion, Phys. Rev. D **1**, 1667 (1970); A. Neveu and J. M. Schwartz, Nucl. Phys. B **31**, 86 (1971); J. L. Gervais and B. Sakita, Nucl. Phys. B **34**, 633 (1971); A. Joseph, Nuovo Cimento A **8**, 217 (1972); Y. Aharonov, A. Casher, and L. Susskind, Phys. Lett. B **35**, 512 (1974).

[3] E. Witten, Nucl. Phys. B **188**, 513 (1981).

[4] V. A. Kostelecky and D. K. Campbell, Physica D **15**, 3 (1985).

[5] M. B. Green, J. H. Schwartz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, 1987), Vols. 1 and 2; M. Kaku, Strings, *Conformal Field Theory and Topology* (Springer-Verlag, New York, 1989).

[6] N. Seiberg and E. Witten, Nucl. Phys. B **426**, 19 (1994); N. Seiberg, Phys. Rev. D **49**, 6857 (1994); A. Khare and J. Maharana, Nucl. Phys. B **244**, 409 (1984).

[7] L. E. Gendenshtein and I. V. Krive, Usp. Fiz. Nauk **146**, 553 (1985).

[8] F. Cooper, A. Khare, and U. Sukhatme, Phys. Rep. **211**, 268 (1995).

[9] S. V. Sukhumar, J. Phys. A **18**, L697 (1985).

[10] F. Ravndal, Phys. Rev. D **21**, 2461 (1980).

[11] N. de Grumbrughe and V. Rittenberg, Ann. Phys. (N.Y.) **151**, 99 (1983); V. P. Berezovoi and A. I. Pashnev, Teor. Mat. Fiz. **78**, 289 (1989).

[12] J. Beckers, N. Debergh, and A. G. Nikitin, Int. J. Theor. Phys. **36**, 1991 (1997).

[13] A. G. Nikitin, ''On extended supersymmetries and parasupersymmetries,'' in *The X International Conference ''Problems of Quantum Field Theory,''* JINR E2-96-369, Dubna, 1996.

[14] J. Niederle and A. G. Nikitin, J. Phys. A **30**, 999 (1997); also see J. Niederle, ''Discrete involutions and their applications in mathematical physics,'' a Plenary Talk at the *2nd International Conference ''Symmetries in Nonlinear Mathematical Physics,''* Kiev, 8–13 July 1997.

[15] A. G. Nikitin, ''On extended supersymmetries in quantum mechanics,'' in *GROUP21: Physical applications and mathematical aspects of geometry, groups and algebras*, edited by H.-D. Doebner, W. Sherer, and P. Natterman (World Scientific, Singapore, 1997), pp. 509–514; J. Niederle, ''Discrete symmetries and supersymmetries—Powerful tools for studying quantum mechanical problems,'' an Invited Talk at the *XVIth Workshop on Geometric Methods in Physics*, Barcelona, 30 June–6 July 1997.

[16] V. M. Tkachuk and S. I. Vakarchuk, Phys. Lett. A **228**, 141 (1997).

[17] N. B. Borisov, K. N. Illinski, and V. M. Uzdin, Phys. Lett. A **169**, 422 (1992); Teor. Mat. Fiz. **94**, 418 (1993); A. D. Dolgallo and K. N. Illinski, Ann. Phys. (N.Y.) **236**, 219 (1994).
[18] L. D. Landau and E. M. Livshits, *Quantum Mechanics*, 3rd. ed. (Pergamon, New York, 1977).
[19] W. I. Fushchich and A. G. Nikitin, *Symmetries of Equations of Quantum Mechanics* (Allerton, New York, 1994).
[20] I. A. Malkin and B. A. Man'ko, *Dynamical Symmetries and Coherent States of Quantum Systems* (Nauka, Moscow, 1979).
[21] R. J. Huges, V. A. Kostelecky, and M. M. Nieto, Phys. Rev. D **34**, 1100 (1986); F. Cooper, A. Khare, R. Masto, and A. Wapf, Ann. Phys. (N.Y.) **187**, 73 (1988).
[22] This algebra is defined over the field of real numbers $\mathbb{R}$ due to the presence of antilinear operators of complex conjugation.
[23] We use the Gelfand–Tsetlin indices (Ref. 24) (n_1, n_2, n_3) to enumerate finite-dimensional IRs of algebras so(k,m). In our case $n_1 = n_2 = n_3 = \frac{1}{2}$.
[24] I. M. Gel'fand, A. P. Minlos, and Z. Ya. Shapiro, *Representations of the Rotation and Lorentz Groups* (Fizmatgiz, Moscow, 1960).
[25] Y. Aharonov and A. Casher, Phys. Rev. A **19**, 1461 (1986).
[26] H. Baake, P. Recnicke, and V. Rittenberg, J. Math. Phys. **26**, 1070 (1985); N. Debergh and A. G. Nikitin, Helv. Phys. Acta **68**, 21 (1995).
[27] V. A. Kostelecky, V. I. Man'ko, M. M. Nieto, and D. R. Truax, Phys. Rev. A **48**, 951 (1993).
[28] A. G. Nikitin, J. Nonlinear Math. Phys. **1**, 201 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Determination of Wigner distribution function for the *d*-dimensional Coulomb problem

Saeid Nouri
The Center for Theoretical Physics and Mathematics, A.E.O.I., P.O. Box 11365-8486, Tehran, Iran and Department of Physics, Amir-Kabir University, Tehran, Iran

(Received 23 July 1998; accepted for publication 12 November 1998)

In this work we present a theoretical study of the *d*-dimensional Coulomb problem in quantum phase space. A coordinate transformation in hyperspherical space is used that maps the *d*-dimensional Coulomb problem into the *D*-dimensional harmonic oscillator and the Wigner distribution function for the *d*-dimensional Coulomb problem is then obtained. This exactly soluble model can shed some light on finite-size features of Wigner's distribution, which will be a vital experience for various dynamic problems.
[S0022-2488(99)03502-1]

I. INTRODUCTION

The problems associated with the Coulomb and harmonic oscillator problems,[1,2] together with the connection between the two in arbitrary dimensions, which has been studied from various viewpoints,[3–11] have been discussed in detail by many authors. The purpose of this paper is to take advantage of the above connection in order to determine the Wigner distribution function for the *d*-dimensional Coulomb problem in quantum phase space.

The Wigner distribution function plays the central role in a reformulation of Schrödinger quantum mechanics, the phase space picture of quantum mechanics, which describes states by functions in configuration space. In this picture both the position and momentum variables are *c* numbers. In Sec. II the Schrödinger equation for the *d*-dimensional Coulomb problem and the *D*-dimensional harmonic oscillator in hyperspherical coordinates are solved and their energy eigenvalues and eigenfunctions are obtained. In Sec. III the Schrödinger equation for the *d*-dimensional Coulomb problem is mapped onto the *D*-dimensional harmonic oscillator by a coordinate transformation in hyperspherical space and then the connection between energy eigenfunctions of these two systems are obtained. In Sec. IV by using the above connection, the explicit Wigner phase space distribution function for the *d*-dimensional Coulomb problem is calculated.

II. SOLUTION OF THE SCHRÖDINGER EQUATION FOR COULOMB AND HARMONIC OSCILLATOR PROBLEMS IN ARBITRARY DIMENSIONS

The Schrödinger equation for the *d*-dimensional Coulomb problem is

$$\left(-\frac{\hbar^2}{2m}\nabla_d^2-\frac{e^2}{r}\right)\psi(\mathbf{r})=E\psi(\mathbf{r}), \tag{1}$$

where $\mathbf{r}$ is a *d*-dimensional position vector having Cartesian components $x_1, x_2, \ldots, x_d$ with magnitude $r=(\Sigma_{j=1}^{d} x_j^2)^{1/2}$ and the Laplacian ∇_d^2 given by

$$\nabla_d^2=\sum_{j=1}^{d}\frac{\partial^2}{\partial x_j^2}. \tag{2}$$

Because of the spherical symmetry of the problem it is convenient to introduce the hyperspherical coordinates, which are defined as follows:[12]

$$
\begin{aligned}
x_1 &= r\cos\theta_1 \sin\theta_2 \cdots \sin\theta_{d-1},\\
x_2 &= r\sin\theta_1 \sin\theta_2 \cdots \sin\theta_{d-1},\\
x_3 &= r\cos\theta_2 \sin\theta_3 \cdots \sin\theta_{d-1},\\
&\vdots\\
x_j &= r\cos\theta_{j-1} \sin\theta_j \cdots \sin\theta_{d-1},\\
&\vdots\\
x_{d-1} &= r\cos\theta_{d-2} \sin\theta_{d-1},\\
x_d &= r\cos\theta_{d-1},
\end{aligned}
\tag{3}
$$

where $d=2,3,...,$ $0\leqslant r\leqslant\infty$, $0\leqslant\theta_1\leqslant 2\pi$, $0\leqslant\theta_j\leqslant\pi$, and $j=1,2,...,d-1$. As in three dimensions, we substitute the following in Eq. (1):

$$\psi(\mathbf{r})=\mathcal{R}_{nl}(r)Y_{l_1,l_2,...,l_{d-1}}(\theta_1,\theta_2,\ldots,\theta_{d-1}), \tag{4}$$

where $\mathcal{R}_{nl}(r)$ is the radial wave function and $Y_{l_1,l_2,...,l_{d-1}}(\theta_1,\theta_2,\ldots,\theta_{d-1})$ is the generalized spherical harmonics, in which $l_{d-1}=0,1,2,...;$ $l_{d-2}=0,1,2,...,l_{d-1};$ $...;$ $l_2=0,1,2,...,l_3;$ $l_1=-l_2,$ $-l_2+1,...,l_2-1,l_2$. We obtain the radial part of Schrödinger equation as

$$\left\{-\frac{\hbar^2}{2m}\left[\frac{d^2}{dr^2}+\frac{d-1}{r}\frac{d}{dr}-\frac{l(l+d-2)}{r^2}\right]-\frac{e^2}{r}\right\}\mathcal{R}_{nl}(r)=E\mathcal{R}_{nl}(r). \tag{5}$$

Equation (5) can be written as[13]

$$\left[\frac{d^2}{du^2}+\frac{d-1}{u}\frac{d}{du}-\frac{l(l+d-2)}{u^2}+\frac{k}{u}-\frac{1}{4}\right]\phi(u,d,n,l)=0, \tag{6}$$

where $u=r/kr_0$, $r_0=\hbar^2/2me^2$, $k=n+\frac{1}{2}(d-3)$, $l=0,1,2,...,n-1$, and $n\geqslant l+1$.

The energy eigenvalues ϵ_n and their corresponding eigenfunctions $\phi(u,d,n,l)$ are given by

$$\epsilon_n=-\frac{\epsilon_0}{[n+\frac{1}{2}(d-3)]^2}, \tag{7}$$

where $\epsilon_0=me^4/2\hbar^2$, and principle quantum number $n=1,2,3,...,$ and

$$\phi(u;d,n,l)=c(d,n,l)e^{-u/2}u^l L_{n-l-1}^{(2l+d-2)}(u), \tag{8}$$

with the normalization constant

$$c(d,n,l)=r_0^{-d/2}[n+\tfrac{1}{2}(d-3)]^{-(d+1)/2}[\Gamma(n-1)]^{1/2}[2\Gamma(n+l+d-2)]^{-1/2}. \tag{9}$$

Note that the Laguerre polynomials $L_n^{(\alpha)}$ are those defined in handbooks on mathematical functions and are not the more limited $L_{n+\alpha}^{\alpha}$ often used in discussions of the hydrogen atom eigenfunctions.

In a similar way, the radial equation of the D-dimensional harmonic oscillator is given by[12]

$$\left\{-\frac{\hbar^2}{2m}\left[\frac{d^2}{dR^2}+\frac{D-1}{R}\frac{d}{dR}-\frac{L(L+D-2)}{R^2}\right]+\frac{1}{2}m\omega^2R^2\right\}\mathcal{R}_{nl}(R)=E\mathcal{R}_{nl}(R), \tag{10}$$

which can be written as[14]

$$\left[\frac{d^2}{dU^2}+\frac{D-1}{U}-\frac{L(L+D-2)}{U^2}-U^2+K\right]\Phi(U,D,N,L)=0, \tag{11}$$

where $U=R/R_0$, $R_0=(m\omega/\hbar)^2$, $K=2N+D$, and $N\geqslant L$.
The energy eigenvalues E_N and their corresponding eigenfunctions $\Phi(U,D,N,L)$ are given by

$$E_N=\tfrac{1}{2}\hbar\omega(2N+D), \tag{12}$$

$$\Phi(U,D,N,L)=C(D,N,L)e^{-U^2/2}U^L L_{N/2-L/2}^{(L+D/2-1)}(U^2), \tag{13}$$

with the normalization constant

$$C(D,N,L)=R_0^{-D/2}\left[2\Gamma\left(\frac{N}{2}-\frac{L}{2}+1\right)\right]^{1/2}\left[\Gamma\left(\frac{N}{2}+\frac{L}{2}+\frac{D}{2}\right)\right]^{-1/2}. \tag{14}$$

Having obtained the eigenfunctions for Coulomb and harmonic oscillator problems in arbitrary dimensions, Eqs. (8) and (13), we will, in Sec. III, set to link the two cases by writing the *d*-dimensional Coulomb problem eigenfunctions in terms of the *D*-dimensional harmonic oscillator eigenfunctions.

III. MAPPING OF THE COULOMB PROBLEM ONTO THE HARMONIC OSCILLATOR IN ARBITRARY DIMENSIONS

The connection between the Coulomb and harmonic oscillator problems has been studied from various viewpoints and has been discussed in detail by many authors. The main point in this section is the mapping of the *d*-dimensional Coulomb problem onto the *D*-dimensional harmonic oscillator. The map taking Eq. (6) into Eq. (11) is $u=U^2$. The appropriate relation between solutions of Eqs. (8) and (13) when restricting *D, N*, and *L* to integers is[9]

$$\phi(u,d,n,l)=\Lambda\Phi(U,2d-2,2n-2,2l), \tag{15}$$

where

$$\Lambda=\left\{\tfrac{1}{2}R_0^{2d-2}/r_0^d\left[n+\tfrac{1}{2}(d-3)\right]^{d+1}\right\}^{1/2}. \tag{16}$$

The *d*- and *n*-dependent constant Λ arises because $\phi(u,d,n,l)$ and $\Phi(U,D,N,L)$ are normalized to unity in *d* and *D* dimensions, respectively. The identification Eq. (15) yields the solution

$$D=2d-2,\quad N=2n-2,\quad L=2l. \tag{17}$$

It is a general feature of this mapping that the spectrum of the *d*-dimensional Coulomb problem is related to half the spectrum of the *D*-dimensional harmonic oscillator for any even integer *D*. However, the quantities in Eq. (17) have parameter spaces that are further restricted by the properties chosen for this mapping. From Eq. (17), we find that all states of the *d*-dimensional Coulomb problem with $n\geqslant 1$ and $l\geqslant 0$, can be mapped onto the appropriate harmonic oscillator with $N\geqslant 0$ and $L\geqslant 0$, except for $d=1$.

Now by using coordinates (3) and ignoring the constant Λ, we can write Eq. (15) in Cartesian space as

$$\phi_{N_1,\ldots,N_D}=\prod_{j=1}^{D}\,(\alpha/\sqrt{\pi}2^{N_j}N_j!)^{1/2}e^{-(\alpha^2/2)x_j^2}H_{N_j}(\alpha x_j), \tag{18}$$

where $H_N(\alpha x)$ is the Hermite polynomials of order N, and $\alpha=(m\omega/\hbar)^{1/2}$. Thus the d-dimensional Coulomb problem wave function is expanded as a linear combination of simple harmonic oscillator wave functions in Hermite polynomials.

IV. THE WIGNER DISTRIBUTION FUNCTION FOR THE *d*-DIMENSIONAL COULOMB PROBLEM

As the wave function plays the central role in the Schrödinger picture, the phase-space distribution function introduced by Wigner is the starting point in the phase-space picture of quantum mechanics. This distribution function is widely known as the Wigner function.

The Wigner function is constructed from the Schrödinger wave function. It is a function of both position and momentum variables and in d-dimensional phase space is defined as[15,16]

$$W(x_1,\ldots,x_d,p_1,\ldots,p_d;t)=(\pi\hbar)^{-d}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}dy_1\cdots dy_d\exp[2i(p_1y_1+\cdots+p_dy_d)/\hbar]$$
$$\times\psi^*(x_1+y_1,\ldots,x_d+y_d;t)\psi(x_1-y_1,\ldots,x_d-y_d;t), \tag{19}$$

where $x_1,\ldots,x_d$ are independent coordinates and $p_1,\ldots,p_d$ are conjugate momentum variables in d-dimensional phase space. By using the coordinates (3), in the time-independent d-dimensional Coulomb problem which is equivalent to the D-dimensional harmonic oscillator we have

$$W(x_1,\ldots,x_D,p_1,\ldots,p_D)=(\pi\hbar)^{-D}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}dy_1\cdots dy_D\exp[2i(p_1y_1+\cdots+p_Dy_D)/\hbar]$$
$$\times\phi^*(x_1+y_1,\ldots,x_D+y_D)\phi(x_1-y_1,\ldots,x_D-y_D), \tag{20}$$

where $p_1,\ldots,p_D$ are conjugate momenta of $x_1,\ldots,x_D$, which satisfy the commutation relations $[x_j,p_k]=i\hbar\delta_{jk}$, where $p_j=-i\hbar\partial/\partial x_j$ and $j,k=1,2,\ldots,D$.

Now the wave function (18) is substituted in Eq. (20) and after some manipulations and using $H_N(-x)=(-1)^N H_N(x)$ we find

$$W(x_1,\ldots,x_D,p_1,\ldots,p_D)=(\pi\hbar)^{-D}\pi^{-D/2}\prod_{j=1}^{D}\frac{(-1)^{N_j}}{2^{N_j}N_j!}e^{-\alpha^2x_j^2+\beta_j^2}\int_{-\infty}^{\infty}dz_j\,e^{-z_j^2}$$
$$\times H_{N_j}(z_j+\beta_j+\alpha x_j)H_{N_j}(z_j+\beta_j-\alpha x_j), \tag{21}$$

where $z=\alpha(y-ip/\alpha^2\hbar)$ and $\beta=ip/\alpha\hbar$. By taking advantage of the expression[17]

$$\int_{-\infty}^{\infty}dz\,e^{-z^2}H_N(z+\beta+\alpha x)H_N(z+\beta-\alpha x)=\sqrt{\pi}2^N N!L_N(2(\alpha^2x^2-\beta^2)), \tag{22}$$

where L_N is Laguerre polynomials, we have

$$W_{N_1,\ldots,N_D}(\rho_1,\ldots,\rho_D)=(\pi\hbar)^{-D}\prod_{j=1}^{D}\,(-1)^{N_j}e^{-\rho_j^2/2}L_{N_j}(\rho_j^2), \tag{23}$$

where dimensionless quantity ρ is defined by $\rho=[2(\alpha^2x^2+p^2/\alpha^2\hbar^2)]^{1/2}$.

Since

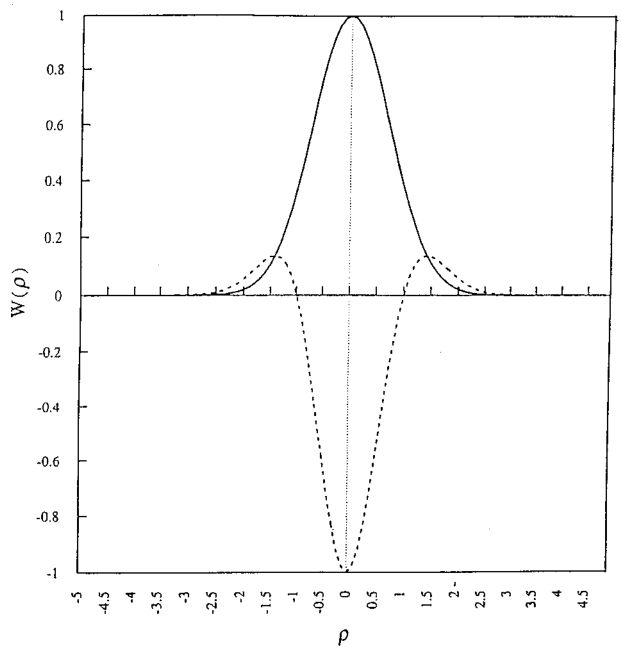

FIG. 1. Cross section of the Wigner distribution function $W(\rho)$ against ρ. Solid and dashed lines correspond to $W_{0,\ldots,0}(\rho_1,\ldots,\rho_D)$ and $W_{1,0,\ldots,0}(\rho_1,\ldots,\rho_D)$, respectively. $W(\rho)$ is given in units of $(\pi\hbar)^{-D}$.

$$\frac{\rho^2}{2}=\frac{2}{\hbar\omega}\left(\frac{p^2}{2m}+\frac{1}{2}m\omega^2x^2\right)=\frac{2}{\hbar\omega}\mathcal{H}(x,p),$$

where $\mathcal{H}(x,p)$ is the Hamiltonian of the harmonic oscillator,

$$W_{N_1,\ldots,N_D}(\rho_1,\ldots,\rho_D)=(\pi\hbar)^{-D}\prod_{j=1}^{D}(-1)^{N_j}\exp\left[\frac{-2}{\hbar\omega}\mathcal{H}(\rho_j)\right]L_{N_j}\left(\frac{4}{\hbar\omega}\mathcal{H}(\rho_j)\right), \tag{24}$$

which is the desired result. The three-dimensional representation of this result has been obtained by using a special mapping.[18]

For the ground state we have

$$W_{0,\ldots,0}(\rho_1,\ldots,\rho_D)=(\pi\hbar)^{-D}\exp[-\tfrac{1}{2}(\rho_1+\cdots+\rho_D)] \tag{25}$$

and for first excited state

$$W_{1,0,\ldots,0}(\rho_1,\ldots,\rho_D)=(\pi\hbar)^{-D}(\rho_1^2-1)\exp[-\tfrac{1}{2}(\rho_1^2+\cdots+\rho_D^2)]. \tag{26}$$

The ground state is positive everywhere in phase space. The first excited state is negative at the origin, but positive for sufficiently large values of $\rho_1+\cdots+\rho_D$ and both become vanishingly small for very large values of $\rho_1+\cdots,\rho_D$. Even though the first excited state is negative around the origin, the probability density in x is always positive. The cross section of variations of the ground and first excited states against ρ is shown in Fig. 1.

In view of the fact that the expression represented in Eq. (24) is a distribution function, then it is possible to calculate some physical and chemical quantities for the hydrogen atom as a special case of the d-dimensional Coulomb problem, using this function.

ACKNOWLEDGMENT

I would like to thank Professor M. Golshani for fruitful discussions in the preparation of this work.

[1] L. I. Schiff, *Quantum Mechanics* (McGraw–Hill, New York, 1968).
[2] E. Merzbacher, *Quantum Mechanics* (Wiley, New York, 1970).
[3] E. Schrödinger, Proc. R. Ir. Acad. Sect. A, Math. Phys. Sci. **46**, 183 (1941).
[4] C. Quesne and M. Moshinsky, J. Math. Phys. **12**, 1780 (1971).
[5] A. O. Barut, C. K. E. Schneider, and R. Wilson, J. Math. Phys. **20**, 2244 (1979).
[6] A. C. Chen, Phys. Rev. A **22**, 333 (1980); **22**, 2901(E) (1980).
[7] F. H. J. Cornish, J. Phys. A **17**, 323 (1984).
[8] D. Bhaumik, B. D. Roy, and G. Ghosh, J. Phys. A **19**, 1365 (1986).
[9] V. A. Kostelecky, M. M. Nieto, and D. R. Truax, Phys. Rev. D **32**, 2627 (1985).
[10] M. Kibler, A. Ronveaux, and T. Negadi, J. Math. Phys. **27**, 1541 (1986).
[11] G. J. Zeng, K. L. Su, and M. Li, Phys. Rev. A **50**, 4373 (1994).
[12] A. Chatterjee, Phys. Rep. **186**, 249 (1990).
[13] M. M. Nieto, Am. J. Phys. **47**, 1067 (1979).
[14] J. D. Louck and W. H. Schaffer, J. Mol. Spectrosc. **4**, 285 (1960).
[15] E. P. Wigner, Phys. Rev. **40**, 749 (1932).
[16] M. Hillery, R. F. O'Connell, M. O. Scully, and E. P. Wigner, Phys. Rep. **106**, 121 (1984).
[17] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Product* (Academic, New York, 1980).
[18] S. Nouri, Phys. Rev. A **57**, 1526 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Dynamical semigroups for interacting quantum and classical systems

R. Olkiewicz
Institute of Theoretical Physics, University of Wrocław, Pl. Maxa Borna 9, PL-50-204 Wrocław, Poland

(Received 22 October 1997; accepted for publication 4 December 1998)

A mathematical framework for the completely positive semigroup coupling between classical and quantum systems is proposed. The coupling ensures a flow of information from the quantum system to the classical one and the influence of the classics on the dynamics of the quantum system in a dissipative way. The classical evolution on average is modified by the expectation value of some quantum operator. Examples of a classical particle moving along a geodesic line in a curved space interacting with the quantum system, and the coupling of a two state quantum system to all pure states, are discussed.
[S0022-2488(99)00803-8]

I. INTRODUCTION

It is well known that using quantum mechanics one can analyze the behavior of electrons in atoms, molecules, and solids. It is also successful in investigating properties of chemical reactions, conductors, and many others. But the quantum dynamics based on the Schrödinger equation makes it difficult to describe irreversible processes like measurements or interactions with a classical environment. Moreover, it seems to be impossible to apply quantum mechanics to explain the occurrence of quantum macroscopic effects, which are expressed solely in classical terms. But the enthusiasm due to successes of quantum mechanics appeared to cover the lack of its completeness for many years.

The situation started to change when technological progress made it possible to make experiments with individual quantum systems. Because experimentalists see not the averages but individual samples, which are the next subject to averaging, the standard interpretation has become insufficient. The importance of the concept of an event and an intrinsic incapability of quantum mechanics to deal with this concept have been stressed by many authors (Ref. 1 and references therein). Haag[2,3] suggested the discreteness and irreversibility of an event in quantum theory and stressed that ''transformation of possibilities into facts must be an essential ingredient which must be included in the fundamental formulation of the theory.''

Recently a mathematically consistent description of the interaction between classical and quantum systems that permits the construction of a new model of quantum measurements has been proposed[4–8] (see also Refs. 9 and 10). From the structural and mathematical point of view the three most essential ingredients of the Blanchard and Jadczyk model are

(i) tensoring of a noncommutative quantum algebra of operators with a commutative classical algebra of functions,
(ii) renouncing pure states for density matrices and replacing Schrödinger unitary dynamics by a completely positive one, and
(iii) interpreting the continuous time evolution of statistical states in terms of a piecewise-deterministic random process on pure states.

In this model classical quantities become elements of the center of the total algebra. Because automorphisms of an algebra leave its center invariant, it was necessary to use completely positive semigroups to enable the transfer of information between the classical and quantum systems. Thus the evolution of the quantum object becomes dissipative and the modification of the dynamics of the classical system through some appropriate expectation value appears. With a given dynamical semigroup T_t we can associate a Markov–Feller process with values in the pure state space of the total system in such a way that $T_t(P_x)=\int P(t,x,dy)P_y$ is satisfied. Here P_x is a one-dimensional projector representing pure state x and $P(t,x,dy)$ is the transition probability function of a desired process. It consists of a mixture of deterministic motion with random jumps. In the case when a discrete classical system (a measuring apparatus) is coupled to a finite quantum system described by a matrix algebra it was shown in Ref. 11 that such a process exists and, moreover, contrary to the pure quantum case, that it is unique.

Clearly, the key point in the coupling is to construct a generator of a dynamical semigroup of the total system. Recently, an example of such a generator of the interaction between a classical system and a quantum one has been introduced in Ref. 12. The classical system was represented by an algebra of continuous functions on some symplectic manifold M while the quantum system was described by a von Neumann algebra acting in some separable Hilbert space. The generator has been built out of the following data:

(a) a self-adjoint quantum operator $\hat{P}$,
(b) a connection between the points of the spectrum of $\hat{P}$ and shifts on a classical phase space M,
(c) a function $f:M\times \mathrm{sp}(\hat{P})\to \mathbf{R}$ responsible for a junction between the classical system and the quantum one.

Such a generator turned out to be suitable for a rigorous discussion of the superconducting quantum interference device (SQUID) -tank model, which consists of an electric oscillatory circuit coupled via a mutual inductance to a superconducting ring containing a weak link constriction. In that system the oscillatory circuit acts as an external flux source for the SQUID ring, which induces a screening current in the ring. This screening current is coupled back to the classical circuit due to the mutual inductance. It results in the modification of the differential equation for the classical harmonic oscillator by the expectation value of the superconducting screening current operator.

However, quite often, a quantum system is characterized by a semispectral measure on some homogeneous space, like in the generalized coherent state approach,[13] or when we use the generalized systems of imprimitivity.[14] In that case there is no particular self-adjoint operator which could be responsible for the coupling, but all quantum states can affect the classical system. The purpose of this paper is to provide a mathematical framework, which could be used to describe such a coupling. Although it is only a mathematical model, we hope the proposed scheme can be applied to a large class of physical phenomena. The paper is organized as follows. In Sec. II the mathematical apparatus is presented. In Sec. III two examples are discussed. Concluding remarks are given in Sec. IV.

II. THE FRAMEWORK

At first we present the formal scheme of the classical-quantum coupling. Let us consider a classical system C with a finite number of degrees of freedom. Its phase space is a symplectic manifold (M,ω). The C^*-algebra $C_0(M)$ of continuous and vanishing at infinity functions represents complex observables of the system. Because it will be more convenient to consider von Neumann algebras we pass to the representation in the Hilbert space $\mathcal{H}_c=L^2(M,\mathcal{B},\mu)$, where $\mathcal{B}$ is the Borel σ-algebra and $d\mu$ is the unique Borel measure determined by the volume form ω^n, $n=\dim M/2$. We assume that the classical algebra $\mathcal{A}_c$ equals to $C_0(M)''=L^\infty(M,\mathcal{B},\mu)$. Statistical states of C are then normed and positive elements of $L^1(M,\mathcal{B},\mu)$. The time evolution of C is described by a flow on M, i.e., a mapping $g:(t,x)\to g_t(x)$ such that

(a) $g:\mathbf{R}\times M\to M$ is smooth,
(b) for any t, g_t is a diffeomorphism of M, and

(c) $t \to g_t$ is a group homomorphism.

Its generator is a complete vector field X on M. It gives an ultraweakly continuous one-parameter group of automorphisms of $\mathcal{A}_c : f(x) \to f(g_t^{-1}x)$, $x \in M$. Its generator we denote by δ_c.

Now we come to the quantum system. Let us consider a quantum particle moving on a homogeneous configuration space $Q = G/K$, where G is a locally compact topological group which satisfies the second axiom of countability and K is a closed subgroup. We assume moreover that G and K are both unimodular. The quantum theory of such a system may be introduced by the concept of generalized coherent states.[13] Let $(\pi, \mathcal{H}_q)$ be a unitary irreducible representation of G, such that for every $k \in K, \pi(k)\psi_0 = e^{i\alpha(k)}\psi_0$ for some unit vector $\psi_0 \in \mathcal{H}_q$. It follows that for each $q \in Q$ we have a one-dimensional projector $P_q = |\pi(g)\psi_0\rangle\langle\psi_0\pi(g)|$, where $[g] = q$. The quantum algebra $\mathcal{A}_q$ is defined as

$$\mathcal{A}_q = \left\{ \int f(q) P_q \, d\alpha(q), \quad f \in C_c(Q) \right\}'' = \{P_q, \quad q \in Q\}''.$$

If for any q' the reproducing kernel $q \to K(q',q)$ vanishes only on a set of α-measure zero, we show that $\mathcal{A}_q = L(\mathcal{H}_q)$. To see this it is enough to prove that every one-dimensional projector P on $\mathcal{H}_q$ belongs to $\mathcal{A}_q$. Let

$$D = \left\{ \sum_{i=1}^{m} z_i \pi(g_i)\psi_0 : \quad m \in \mathbf{N}, \ z_i \in \mathbf{C}, \quad g_i \in G \right\}.$$

Because the representation π is irreducible, ψ_0 is cyclic and D is dense in $\mathcal{H}_q$. Let $\psi_n \to \psi$, where $\psi_n \in D$ with $\|\psi_n\| = 1$ and $P = |\psi\rangle\langle\psi|$, $\|\psi\| = 1$. Then $P_n = |\psi_n\rangle\langle\psi_n|$ tends to P in the weak topology. But for any $n \in \mathbf{N}$ there is

$$P_n = \sum_{k,l}^{\text{finite}} \overline{z_k} z_l |\pi(g_k)\psi_0\rangle\langle\psi_0\pi(g_l)|.$$

Because $P_{qk}P_{ql} = K(q_k, q_l)|q_k\rangle\langle q_l|$, then $P_n \in \mathcal{A}_q$. Hence $P \in \mathcal{A}_q$, too, and $\mathcal{A}_q = L(\mathcal{H}_q)$.

In this case we can define a semispectral measure on Q by $E(B) = \int_B P_q \, d\alpha(q)$, where $B \in \mathcal{B}(Q)$, the Borel σ-algebra on Q, and $d\alpha$ is a unique (up to a positive constant) G-invariant and σ-finite Borel measure on Q. By a semispectral measure we understand a map $E : \mathcal{B}(Q) \to L(\mathcal{H}_q)_+$ such that for every $\psi \in \mathcal{H}_q$ and $B \in \mathcal{B}(Q)$ the map $B \to \langle\psi, E(B)\psi\rangle$ is a positive measure. Although the description of a quantum system by coherent states is sufficient in many cases, for further applications *we generalize this scheme and assume that the quantum algebra is generated (as a von Neumann algebra) by an arbitrary semispectral measure dE on Q*. Statistical states of the quantum system are given by non-negative density matrices $\rho \in \mathcal{A}_q$ with $\mathrm{Tr}(\rho) = 1$. The time evolution of a quantum observable A is given by $A \to e^{itH}Ae^{-itH}$, where H is a self-adjoint operator affiliated to $\mathcal{A}_q$. Its generator $i[H, \cdot]$ we denote by δ_q. The requirement that H is affiliated to the algebra $\mathcal{A}_q$, which represents the quantum system, follows from the need that all its spectral projectors should belong to $\mathcal{A}_q$.

Let us now consider the joint system. For the total algebra $\mathcal{A}_T$ we take the tensor product $\mathcal{A}_T = \mathcal{A}_c \otimes \mathcal{A}_q$ as von Neumann algebras on $\tilde{\mathcal{H}} = \mathcal{H}_c \otimes \mathcal{H}_q$. The set of states is equal to

$$\mathcal{S}_T = \left\{ \tilde{\rho} \in \mathcal{A}_{T*} : \tilde{\rho}(x) \in \mathrm{Tr}(\mathcal{H}_q)_+ \text{a.e. and } \int_M \mathrm{Tr}(\tilde{\rho}(x)) d\mu(x) = 1 \right\}.$$

The mean value of $\tilde{A} \in \mathcal{A}_T$ in a state $\tilde{\rho} \in \mathcal{S}_T$ is given by

$$\langle \tilde{A} \rangle_{\tilde{\rho}} = \int_M d\mu(x) \mathrm{Tr}[\tilde{A}(x)\tilde{\rho}(x)].$$

Now let us discuss the evolution of the total system. The total generator consists of three parts: $\delta_c \otimes \mathrm{id}$, $\mathrm{id} \otimes \delta_q$, and a superoperator L which describes the interaction between the classical and the quantum system.

To construct L we assume the following:

(a) Let to every point $x \in M$ correspond a finite and positive measure $\nu_x \in M(G) = C_0(G)^*$ such that $x \to \nu_x(G)$ is uniformly bounded and for every Borel set $B \subset G$ the map $x \to \nu_x(B)$ is Borel. The influence of the classical system on the quantum one is reflected by a map $dE \to d(\nu_x * E)$ which changes a quantum observable $\hat{f} = \int f(q) dE(q)$ to $\int f(q) d(\nu_x * E)(q)$. We assume that

$$\exists C > 0 \forall x \in M \forall B \in \mathcal{B}(Q) \text{ and } \alpha(B) < \infty \, \|(\nu_x * E)(B)\| \leqslant C\alpha(B).$$

(b) Let to every point $q \in Q$ corresponds a shift on the phase space. It will be responsible for an action of the quantum system on the classical one. By the shift we mean a morphism of $(M, \mathcal{B}, \mu)$, i.e., a bijective map $h_q : M \to M$ such that h_q and h_q^{-1} are measurable and leave the measure $d\mu$ invariant. Moreover, we demand that for any $f \in C_c(M)$ and any $x \in M$ a mapping $q \to f(h_q^{-1}x)$ is Borel measurable. Then $U_q : \mathcal{H}_c \to \mathcal{H}_c, (U_q \Psi)(x) = \Psi(h_q^{-1}x)$ is a unitary operator. Moreover, $q \to \langle f_1, U_q f_2 \rangle$ is measurable for any $f_1, f_2 \in C_c(M)$ and so $q \to U_q$ is weakly measurable.

At first we show that $\nu_x * E$ is a semispectral measure on Q and so the condition from the point (a) is well defined. Let $\psi_1, \psi_2 \in \mathcal{H}_q$ and

$$dE_{\psi_1, \psi_2}(q) = d\langle E(q)\psi_1, \psi_2 \rangle.$$

Let for $f \in C_0(Q)$

$$\phi(f) = \int_Q \int_G f(gq) d\nu_x(g) dE_{\psi_1, \psi_2}(q).$$

It is a continuous functional and so it determines the unique complex measure with finite variation $\nu_x * E_{\psi_1, \psi_2}$ defined on the Borel σ-algebra $\mathcal{B}(Q)$. Because for every $B \in \mathcal{B}(Q)(\psi_1, \psi_2) \to (\nu_x * E_{\psi_1, \psi_2})(B)$ is a positive and continuous sesquilinear form there is a bounded and positive operator $(\nu_x * E)(B)$ such that $B \to \langle \psi, (\nu_x * E)(B)\psi \rangle$ is a positive measure.

In the following two propositions we introduce situations in which assumption $\|(\nu_x * E) \times (B)\| \leqslant C\alpha(B)$ is satisfied.

Proposition 2.1: Let $Q = G$ and $d\alpha$ is a left-invariant Haar measure on G. If $d\nu_x = f(x,g) d\alpha(g)$, where $f: M \times G \to \mathbf{R}_+$ is a Borel function such that for every $x \in M f(x, \cdot) \in L^1(G, d\alpha)$ and $\sup_{x,g} f(x,g) \leqslant C$, then for every semispectral measure dE on G we have that $\|(\nu_x * E)(B) \leqslant C\|E(G)\|\alpha(B)$, where B is an arbitrary but α-finite Borel subset of G.

Proof: Because $(\nu_x * E)(B)$ is a positive operator we have that

$$\|(\nu_x * E)(B)\| = \|(\nu_x * E)^{1/2}(B)\|^2 = \sup_{\|\psi\|=1} \langle \psi, (\nu_x * E)(B)\psi \rangle = \sup_{\|\psi\|=1} (\nu_x * E_{\psi,\psi})(B).$$

However,

$$(\nu_x * E_{\psi,\psi})(B) = \int_G \nu_x(Bg^{-1}) dE_{\psi,\psi}(g) = \int_G \int_B f(x, hg^{-1}) d\alpha(h) dE_{\psi,\psi}(g) \leqslant C\|E(G)\|\alpha(B)\|\psi\|^2.$$

□

Proposition 2.2: Let $Q = G/K$ and $d\alpha$ is the G-invariant measure on Q. If $E(B)$ is given by $E(B) = \int_B P(q)\, d\alpha(q)$, where $P(q)$ is a projector and $q \to P(q)$ is weakly Borel measurable and weakly α-integrable, then $\|(\nu_x * E)(B)\| \leqslant C\alpha(B)$ for every $\nu_x \in M(G)$ such that $\sup_x \nu_x(G) \leqslant C$.

Proof: Now we have

$$(\nu_x * E_{\psi,\psi})(B) = \int_G E_{\psi,\psi}(g^{-1}B)\, d\nu_x(g) = \int_G \int_{g^{-1}B} \langle \psi, P(q)\psi\rangle\, d\alpha(q)\, d\nu_x(g)$$

$$\leq \|\psi\|^2 \int_G \alpha(g^{-1}B)\, d\nu_x(g) \leq C\|\psi\|^2 \alpha(B),$$

and so the assertion follows. □

In the coherent state case we can describe more precisely the influence of the classical system onto the quantum one. Let $\hat{f} = \int f(q)\, dE(q)$ be a quantum operator corresponding to a function *f*. A point $x \in M$ changes it to

$$\hat{f}_x = \int_Q f(q)\, d(\nu_x * E)(q).$$

Because

$$(\nu_x * E)(B) = \int_G E(g^{-1}B)\, d\nu_x(g)$$

and

$$E(g^{-1}B) = \int_B P_{gq}\, d\alpha(q),$$

then

$$d(\nu_x * E)(q) = \left(\int_G P_{gq}\, d\nu_x(g) \right) d\alpha(q).$$

However, $P_{gq} = \pi(g) P_q \pi(g)^*$, hence

$$\hat{f}_x = \int_G \pi(g) \hat{f} \pi(g)^* \; d\nu_x(g).$$

If ν_x is a point measure, then we get just a unitary automorphism of $\hat{f}$.

Let us begin the construction of the generator *L*. We will need the following:

Lemma 2.3: The function $x \to (\nu_x * E)(B)$ is weakly Borel measurable for every $B \in \mathcal{B}(Q)$.

Proof: Let $\psi \in \mathcal{H}_q$. Because of the polarization formula it is enough to consider the function $x \to \langle \psi, (\nu_x * E)(B)\psi\rangle$. However,

$$\langle \psi, (\nu_x * E)(B)\psi\rangle = (\nu_x * E_{\psi,\psi})(B) = \int_G E_{\psi,\psi}(g^{-1}B)\, d\nu_x(g) = \lim_{n\to\infty} \sum_{i=1}^{n} E_{\psi,\psi}(g_i^{-1}B)\, \nu_x(G_i).$$

Because G_i is a Borel subset of *G* and the function $x \to \nu_x(G_i)$ is Borel, the assertion follows. □

It is known that if a semispectral measure *dE* has the property $\|E(B)\| \leq C\alpha(B)$ for a σ-finite measure $d\alpha$, then *dE* possesses an operator-valued density.[14] Now we generalize this fact to a family of semispectral measures $d(\nu_x * E)$ and show the essential boundedness of the corresponding operator-valued density.

Theorem 2.4: There is a weakly Borel measurable function $\tilde{V}: M \times Q \to L(\mathcal{H}_q)_+$ such that $\sup_{x,q} \|\tilde{V}(x,q)\| \leq \infty$ and for every $x \in M \setminus N$, where *N* is a Borel subset of μ-measure zero,

$$(\nu_x * E)(B) = \int_B \tilde{V}(x,q)\, d\alpha(q)$$

in the weak sense. Moreover, any other such $\tilde{V}'$ agrees with $\tilde{V}$ $\mu \times \alpha$ almost everywhere.

Proof: See Appendix A.

Corollary 2.5: From the Fubini theorem there is a Borel set $Q_0 \subset Q$ such that $\alpha(Q_0) = 0$ and the map $x \to \tilde{V}_q(x) = \tilde{V}(x,q)$ is weakly Borel measurable for all $q \in Q \backslash Q_0$. Putting $\tilde{V}_q = 0$ for $q \in Q_0$ we get a well-defined map $q \to \tilde{V}_q \in \mathcal{A}_T$ for every $q \in Q$.

Proposition 2.6: $\sup_q \|\tilde{V}_q\| \leqslant \infty$ and $q \to \tilde{V}_q$ is weakly Borel measurable.

Proof: Only the second statement needs a proof. Because $\tilde{V}_q$ is uniformly bounded, it is enough to check it for simple tensors $\tilde{\Psi} = \psi \otimes \Psi$, where $\psi \in \mathcal{H}_q$ and $\Psi \in \mathcal{H}_c$. If $q \in Q \backslash Q_0$, then

$$\langle \psi \otimes \Psi, \tilde{V}_q(\psi \otimes \Psi) \rangle = \int_M d\mu(x) |\Psi(x)|^2 \langle \psi, \tilde{V}(x,q) \psi \rangle.$$

Again from the Fubini theorem, $q \to \langle \psi, \tilde{V}(x,q)\psi \rangle$ is Borel for a.e. $x \in M$, and so $q \to \int_M d\mu(x) |\Psi(x)|^2 \langle \psi, \tilde{V}(x,q)\psi \rangle$ is Borel, too. Because Q_0 is a Borel set and $\tilde{V}_q = 0$ for $q \in Q_0$, so $q \to \tilde{V}_q$ is weakly Borel measurable for all q. □

Let $\tilde{U}_q = U_q \otimes \mathbf{1}$ and let $\tilde{W}_q = \tilde{U}_q \cdot \tilde{V}_q$. Clearly $\tilde{W}_q^* \tilde{A} \tilde{W}_q \in \mathcal{A}_T$ and the map $q \to \tilde{W}_q^* \tilde{A} \tilde{W}_q$ is weakly Borel measurable.

Now let $W_n(\tilde{A})$ be defined by

$$\langle \tilde{\Psi}_1, W_n(\tilde{A}) \tilde{\Psi}_2 \rangle = \int_{Q^n} d\alpha(q) \langle \tilde{\Psi}_1, \tilde{W}_q^* \tilde{A} \tilde{W}_q \tilde{\Psi}_2 \rangle, \quad \tilde{\Psi}_1, \tilde{\Psi}_2 \in \tilde{\mathcal{H}},$$

where $Q^n := \cup_{i=1}^{i=n} Q_i$ and Q_i are α-finite sets which covers Q. Q^n is also α-finite. Because the function $q \to \langle \tilde{\Psi}_1, \tilde{W}_q^* \tilde{A} \tilde{W}_q \tilde{\Psi}_2 \rangle$ is Borel and uniformly bounded, the integral on the right-hand side exists and defines a linear and bounded operator on $\tilde{\mathcal{H}}$. As the weak limit of operators from $\mathcal{A}_T$, $W_n(\tilde{A})$ also belongs to $\mathcal{A}_T$.

Proposition 2.7: $W(\tilde{A}) = \lim_{n \to \infty} W_n(\tilde{A})$ exists and belongs to $\mathcal{A}_T$.

Proof: Let $\tilde{A}$ be a positive operator. Then

$$\tilde{W}_q^* \tilde{A} \tilde{W}_q \leqslant \|\tilde{A}\| \tilde{W}_q^* \tilde{W}_q = \|\tilde{A}\| \tilde{V}_q^2.$$

However, $\tilde{V}_q$ is positive so $\tilde{V}_q^2 \leqslant \|\tilde{V}_q\| \tilde{V}_q$ and

$$\begin{aligned}
\int_{Q^n} d\alpha(q) \langle \tilde{\Psi}, \tilde{W}_q^* \tilde{A} \tilde{W}_q \tilde{\Psi} \rangle &\leqslant \|\tilde{A}\| (\sup_q \|\tilde{V}_q\|) \int_{Q^n} d\alpha(q) \langle \tilde{\Psi}, \tilde{V}_q \tilde{\Psi} \rangle \\
&\leqslant C \|\tilde{A}\| \int_{Q^n} d\alpha(q) \int_M d\mu(x) \langle \tilde{\Psi}(x), \tilde{V}(x,q) \tilde{\Psi}(x) \rangle \\
&= C \|\tilde{A}\| \int_M d\mu(x) (\nu_x * E_{\tilde{\Psi}(x), \tilde{\Psi}(x)})(Q^n) \\
&\leqslant C \|\tilde{A}\| \int_M d\mu(x) \|(\nu_x * E)(Q)\| \cdot \|\tilde{\Psi}(x)\|^2 \\
&\leqslant C \|\tilde{A}\| (\sup_x \|(\nu_x * E)(Q)\|) \|\tilde{\Psi}\|^2.
\end{aligned}$$

However,

$$\|(\nu_x * E)(Q)\| = \sup \lim_{\|\psi\|=1} (\nu_x * E_{\psi,\psi})(Q) = \sup \lim_{\|\psi\|=1} \int_G E_{\psi,\psi}(Q)\, d\nu_x(g) = \|E(Q)\| \, \nu_x(G),$$

and $\sup_x \nu_x(G) = C_1$. Thus

$$\int_{Q^n} d\alpha(q) \langle \tilde{\Psi}, \tilde{W}_q^* \tilde{A} \tilde{W}_q \tilde{\Psi} \rangle \leqslant C C_1 \|\tilde{A}\| \cdot \|E(Q)\| \cdot \|\tilde{\Psi}\|^2.$$

So from the polarization formula we have that $\forall \tilde{\Psi}_1, \ \tilde{\Psi}_2 \in \tilde{H}$ and $\forall \tilde{A} \in \mathcal{A}_T$, $\int_Q d\alpha(q) \times \langle \tilde{\Psi}_1, \tilde{W}_q^* \tilde{A} \tilde{W}_q \tilde{\Psi}_2 \rangle$ exists. This means that $W_n(\tilde{A})$ is weakly convergent to $W(\tilde{A})$ and thus $W(\tilde{A}) \in \mathcal{A}_T$. □

Now let us consider a*-linear map $W: \mathcal{A}_T \to \mathcal{A}_T$.

Lemma 2.8: W is a completely positive and normal map.

Proof: See Appendix B.

Thus we have proved the following.

Theorem 2.9: Let $L(\tilde{A}) = W(\tilde{A}) - \frac{1}{2}\{W(\tilde{\mathbf{1}}), \tilde{A}\}$, where $\{\cdot,\cdot\}$ stands for the anticommutator. Then L is a bounded complete dissipation and $L(\tilde{\mathbf{1}}) = 0$, where $\tilde{\mathbf{1}}$ is the unit in $\mathcal{A}_T$. □

Remark: The adjoint of L is a generator of a one-parameter semigroup of completely positive and conservative maps of $\mathcal{S}_T$. Formally it can be written as

$$L_*(\tilde{\rho})(x) = \int_Q d\alpha(q) \tilde{V}_q(x) \tilde{\rho}(h_q^{-1}(x)) \tilde{V}_q(x) - \frac{1}{2} \left\{ \int_Q d\alpha(q) \tilde{V}_q^2(x), \tilde{\rho}(x) \right\}.$$

III. APPLICATIONS

In this section we present two examples of a possible classical–quantum coupling and discuss some of its properties.

Example 1: The first area of possible applications of the above formalism is the semiclassical theory of gravity, in which a classical gravitational field interacts with quantum matter. In this approach one studies the generalized Einstein's equation which is modified by regularized vacuum expectation value of the energy-momentum tensor of matter-field operators. In the so-called back-reaction problem the influence of particle creation on the dynamics of gravitational field is analyzed. In Ref. 15 the effective equation for the evolution of a Bianchi type I metric with such back reaction was derived. It was also shown there that there is the dissipation of anisotropy in the Bianchi type I universe through quantum effects. For a general discussion of the dissipation in semiclassical gravity, see Ref. 16. Here we present a simpler situation, in which a nonrelativistic classical particle moving along a geodesic line in a curved space interact with a quantum system. As the result we obtain a dissipative behavior of the quantum density matrix and the modification of the classical evolution on average through the expectation value of the quantum position operator.

At first we describe the classical system. Let us consider a classical particle moving freely on the Lobatchevsky space:

$$Q = \mathbf{R} \times \mathbf{R}_+ = \{(x_1, x_2): \quad x_2 > 0\}.$$

The phase space is the cotangent space $M = T^*Q$ with the canonical symplectic form $\omega = dp_1 \wedge dx_1 + dp_2 \wedge dx_2$. It leads to the Lebesgue'a measure $d\mu = dx_1 dx_2 dp_1 dp_2$ on M. The time evolution is governed by a complete vector field X on M:

$$X(x_i, p_i) = \sum_{k=1}^{2} \left[f_k(x_i, p_i) \frac{\partial}{\partial x_k} + g_k(x_i, p_i) \frac{\partial}{\partial p_k} \right],$$

where $f_k(x_i,p_i)=p_k$ and $g_1(x_i,p_i)=2p_1p_2/x_2$, $g_2(x_i,p_i)=(p_2^2-p_1^2)/x_2$. For simplicity we denote four parameters (x_1,x_2,p_1,p_2) by (x_i,p_i). It gives the following second-order differential equations for the position coordinates:

$$\ddot{x}_1=\frac{2}{x_2}\dot{x}_1\dot{x}_2,$$

$$\ddot{x}_2=\frac{1}{x_2}[(\dot{x}_2)^2-(\dot{x}_1)^2].$$

Hence the classical particle moves on a geodesic line with respect to the metric

$$g=\frac{1}{x_2^2}(dx_1\otimes dx_1+dx_2\otimes dx_2).$$

To describe the quantum system we use the system of generalized coherent states.[13] Let us consider a quantum particle on the Lobatchevsky space Q. It is a homogeneous space $Q=\mathrm{SL}(2,\mathbf{R})/\mathrm{U}(1)$. For simplicity we take the first representation from the series $(\mathcal{H}_k,\pi_k)$, where $k=1,\frac{3}{2},2,\ldots$. That is,

$$\mathcal{H}_q=\left\{f:\|f\|^2=\int d\mu_1(z)|f(z)|^2<\infty\right\},$$

where f is a holomorphic function in the unit complex disc $|z|<1$ and $d\mu_1=(1/\pi)dzd\overline{z}.$ For $q=(x_1,x_2)\in Q$ we have one-dimensional projectors $P_q=|\zeta\rangle\langle\zeta|$, where

$$|\zeta\rangle=\frac{1-|\zeta|^2}{(1-\overline{\zeta}z)^2}\quad\text{and}\quad\zeta=\frac{1-x_2+ix_1}{1+x_2-ix_1}.$$

The semispectral measure $E(B)$, $B\in\mathcal{B}(Q)$, is given by $E(B)=\int_B P_q\,d\alpha(q)$, where $d\alpha$ is the unique $\mathrm{SL}(2,\mathbf{R})$ invariant measure on Q normalized in such a way that $\int_Q P_q\,d\alpha(q)=I$, the identity operator. The quantum operator corresponding to a function f on Q reads

$$\hat{f}=\int_Q f(q)dE(q).$$

A self-adjoint operator H on $\mathcal{H}_q$ determines the time evolution in the standard way.

To define a generator L of the total system we assume the following:

(i) $\forall m=(q,p)\in M$ $\nu_{(q,p)}=\delta_e$, where δ_e denotes the point measure concentrated in the neutral element of the group $\mathrm{SL}(2,\mathbf{R})$, and
(ii) $\forall q\in Q$ $h_q:M\to M$ is given by $(q',p)\to(q',p+q)$.

Then $\tilde{V}(m,q)=P_q$ and so

$$L(\tilde{A})(m)=\int_Q d\alpha(q)P_q\tilde{A}(h_q(m))P_q-\tilde{A}(m),$$

$$L_*(\tilde{\rho})(m)=\int_Q d\alpha(q)P_q\tilde{\rho}(h_q^{-1}(m))P_q-\tilde{\rho}(m).$$

Let us derive the equation for the quantum density operator $\rho_t=\int_M\tilde{\rho}_t(x_i,p_i)d\mu$:

$$\dot{\rho}_t = \int_M \dot{\tilde{\rho}}_t(x_i,p_i)\,d\mu = \int_M [(\delta_c^{ad}\otimes \mathbf{1})\tilde{\rho}_t(x_i,p_i) + (\mathbf{1}\otimes \delta_q^{ad})\tilde{\rho}_t(x_i,p_i) + L_*(\tilde{\rho}_t)(x_i,p_i)]\,d\mu.$$

Here again we denote four parameters (x_1,x_2,p_1,p_2) by (x_i,p_i). Because for simple tensors $\phi \otimes \rho$, $\phi \in L^1(M,d\mu)$ and $\rho \in \mathrm{Tr}(\mathcal{H}_q)$, there is

$$\int_m (\delta_c^{ad}\otimes \mathbf{1})(\phi\otimes\rho)(x_i,p_i)\,d\mu = \rho\int_M (\delta_c^{ad}\phi)(x_i,p_i)\,d\mu = \rho\int_M \phi(x_i,p_i)\,\delta_c(id)(x_i,p_i)\,d\mu = 0$$

and

$$\int_M (\mathbf{1}\otimes \delta_q^{ad})(\phi\otimes\rho)(x_i,p_i)\,d\mu = \delta_q^{ad}\left(\rho\int_M \phi(x_i,p_i)\,d\mu\right)$$

and

$$\tilde{\rho}_t = \lim_{m\to\infty}\sum_{i=1}^{m} \phi_i(t)\otimes\rho_i(t)$$

so

$$\int_M (\delta_c^{ad}\otimes \mathbf{1})\tilde{\rho}_t(x_i,p_i)\,d\mu = 0$$

and

$$\int_M (\mathbf{1}\otimes \delta_q^{ad})\tilde{\rho}_t(x_i,p_i)\,d\mu = \delta_q^{ad}(\rho_t) = -i[H,\,\rho_t].$$

Finally

$$\int_M (L_*\tilde{\rho}_t)(x_i,p_i)\,d\mu = \int_M d\mu \int_Q d\alpha(q)\,P_q\tilde{\rho}_t(h_q^{-1}(x_i,p_i))P_q - \rho_t\,.$$

Because h_q^{-1} leaves the measure $d\mu$ invariant, by changing the order of integrals we get

$$\int_M (L_*\tilde{\rho}_t)(x_i,p_i)\,d\mu = \int_Q d\alpha(q)\,P_q\rho_t P_q - \rho_t\,.$$

Hence the time evolution equation for ρ_t is modified by a dissipative factor

$$\dot{\rho}_t = -i[H,\rho_t] + \int_Q d\alpha(q)\,P_q\rho_t P_q - \rho_t\,.$$

To see how this dissipative part acts we assume now that $H=0$. Thus

$$L_q(\rho) = \int_Q P_q\rho P_q\,d\alpha(q) - \rho.$$

Let $T_q(t)$ be the semigroup on $\mathrm{Tr}(\mathcal{H}_q)$ generated by L_q.

Theorem 3.1: Let $P_n = |n\rangle\langle n|$, where $|n\rangle(z) = \sqrt{n+1}z^n$, $n\in\mathbf{N}\cup\{0\}$, is an orthonormal base in $\mathcal{H}_q$. Then

$$T_q(t)(P_n)=\sum_{m=0}^{\infty} P(t,n,\{m\})P_m\,,$$

where

$$P(t,n,\{m\})=e^{t(\Pi-I)}=e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\Pi_k(n,\{m\})$$

is the Poisson transition kernel with the probabilistic measure on $\mathbf{N}\cup\{0\}$ given by

$$\Pi(n,\{m\})=\frac{2(n+1)(m+1)}{(n+m+1)(n+m+2)(n+m+3)}.$$

Here $I(n,\{m\})=\delta_{nm}$, $\Pi_0(n,\{m\})=\delta_{nm}$, and

$$\Pi_k(n,\{m\})=\sum_{l=0}^{\infty}\Pi(n,\{l\})\Pi_{k-1}(l,\{m\})$$

for $k\geqslant 1$.

Proof: Because $T_q(t)=e^{tL_q}$, it is enough to show that

$$L_q(P_n)=\sum_{m=0}^{\infty}\Pi(n,\{m\})P_m-P_n\,.$$

To calculate the integral $\int P_qP_nP_q\,d\alpha(q)$ we use the representation of the Lobatchevski space as the Poincaré disc with SL(2,**R**) replaced by SU(1,1). Thus

$$d\alpha(q)\rightarrow d\alpha(\zeta)=\frac{1}{\pi}\frac{d\zeta d\bar{\zeta}}{(1-|\zeta|^2)^2}.$$

Let us observe that

$$P_qP_nP_q=|\langle\zeta,n\rangle|^2|\zeta\rangle\langle\zeta|=(n+1)(1-|\zeta|^2)^2|\zeta|^{2n}|\zeta\rangle\langle\zeta|.$$

In the above we use the formula $\langle\zeta,n\rangle=(1-|\zeta|^2)|n\rangle(\zeta)$.[13] We show that $\int P_qP_nP_q\,d\alpha(\zeta)$ is also diagonal in the $\{|n\rangle\}_0^\infty$ base. Indeed

$$\left\langle n_1,\int P_qP_nP_q d\alpha(\zeta)n_2\right\rangle=\int|\langle\zeta,n\rangle|^2\langle n_1,\zeta\rangle\langle\zeta,n_2\rangle d\alpha(\zeta)$$
$$=\frac{(n+1)\sqrt{(n_1+1)(n_2+1)}}{\pi}\int(1-|\zeta|^2)^2|\zeta|^{2n}\zeta^{n_2}\bar{\zeta}^{n_1}\,d\zeta d\bar{\zeta}=0.$$

On the other hand,

$$\left\langle m,\int P_qP_nP_q d\alpha(\zeta)m\right\rangle=\frac{(n+1)(m+1)}{\pi}\int(1-|\zeta|^2)^2|\zeta|^{2(n+m)}d\zeta d\bar{\zeta}$$
$$=\frac{(n+1)(m+1)}{\pi}\int_0^1\int_0^{2\pi}(1-r^2)^2r^{2(n+m)}r\,d\phi\,dr$$
$$=\frac{2(n+1)(m+1)}{(n+m+1)(n+m+2)(n+m+3)}.$$

Hence

$$\int P_q P_n P_q \, d\alpha(\zeta) = \sum_{m=0}^{\infty} \Pi(n,\{m\}) P_m ,$$

so the assertion follows. □

Now let us choose $\tilde{\rho}$ from some dense subset of $\mathcal{S}_T$ {which is compactly supported in the classical part and with the property that the commutator $[H,\tilde{\rho}(x_i,p_i)]$ belongs to the trace class}. Let us define so-called collective classical coordinates

$$\overline{x_k} = \widetilde{\mathrm{Tr}}(\tilde{\rho}(x_k \otimes \mathbf{1})) = \int_M x_k (\mathrm{Tr}\, \tilde{\rho}(x_i,p_i)) d\mu$$

and

$$\overline{p_k} = \widetilde{\mathrm{Tr}}(\tilde{\rho}(p_k \otimes \mathbf{1})) = \int_M p_k (\mathrm{Tr}\, \tilde{\rho}(x_i,p_i)) d\mu$$

for $k=1,2$. Their time evolution is given by

$$\dot{\bar{x}}_k = \int_M x_k (\mathrm{Tr}\, \dot{\tilde{\rho}}_t(x_i,p_i)) d\mu = \int_M x_k (\mathrm{Tr}(\delta_c^{ad} \otimes \mathbf{1} + \mathbf{1} \otimes \delta_q^{ad} + L_*) \tilde{\rho}_t(x_i,p_i)) d\mu.$$

However,

$$\int_M x_k \,\mathrm{Tr}(\delta_c^{ad} \otimes \mathbf{1}) \tilde{\rho}_t(x_i,p_i) d\mu = \int_M \delta_c(x_k) \mathrm{Tr}\, \tilde{\rho}_t(x_i,p_i) d\mu = \int_M p_k \,\mathrm{Tr}\, \tilde{\rho}_t(x_i,p_i) d\mu = \overline{p_k} ,$$

$$\int_M x_k \,\mathrm{Tr}(\mathbf{1} \otimes \delta_q^{ad}) \tilde{\rho}_t(x_i,p_i) d\mu = -i \int_M x_k \,\mathrm{Tr}[H, \tilde{\rho}_t(x_i,p_i)] d\mu = 0,$$

and

$$\int_M d\mu(x_i,p_i) x_k \,\mathrm{Tr}\left(\int_Q d\alpha(q) P_q \tilde{\rho}_t(x_i,p_i - q_i) P_q \right)$$
$$= \int_M d\mu(x_i,p_i') x_k \,\mathrm{Tr}\left(\int_Q d\alpha(q) P_q \right) \tilde{\rho}_t(x_i,p_i) = \overline{x_k} ,$$

so $\dot{\bar{x}}_k = \overline{p_k}$. In the same way we get that

$$\int_M p_k \,\mathrm{Tr}(\delta_c^{ad} \otimes \mathbf{1}) \tilde{\rho}_t(x_i,p_i) d\mu = \overline{\delta_c(p_k)} = \overline{g_k(x_i,p_i)},$$

$$\int_M p_k \,\mathrm{Tr}(\mathbf{1} \otimes \delta_q^{ad}) \tilde{\rho}_t(x_i,p_i) d\mu = 0,$$

and

$$\int_M d\mu(x_i,p_i)p_k \operatorname{Tr}\left(\int_Q d\alpha(q)P_q\tilde{\rho}_t(x_i,p_i-q_i)P_q\right)$$

$$=\int_M d\mu(x_i,p_i')(p_k+q_k)\operatorname{Tr}\left(\int_Q d\alpha(q)P_q\tilde{\rho}_t(x_i,p_i')\right)$$

$$=\overline{p_k}+\int_M d\mu(x_i,p_i)\operatorname{Tr}\left(\int_Q q_kP_q\,d\alpha(q)\right)\tilde{\rho}_t(x_i,p_i).$$

Let us notice that the operator $\int_Q q_kP_q\,d\alpha(q)$ is the quantum position operator $\hat{q}_k$. Thus

$$\dot{\bar{p}}_k=\overline{g_k(x_i,p_i)}+\operatorname{Tr}(\hat{q}_k\rho_t),$$

where $\rho_t=\int_M\tilde{\rho}_t(x_i,p_i)d\mu$ is the quantum density operator. Hence the classical evolution is perturbed by the expectation value of the quantum position operator.

Example 2: In this example we describe a quantum system coupled to all one-dimensional projectors considered as a classical device. Such a model was discussed in Ref. 17 in connection with the question: how do we determine the state of an individual quantum system?

Let us consider pure spin 1/2 system. Then $\mathcal{H}_q=\mathbf{C}^2$ and $\mathcal{A}_q=M_{2\times 2}(\mathbf{C})$. As the classical phase space we take the space of all one-dimensional projectors on $\mathcal{H}_q$, that is, $M=\mathbf{C}P^1=S^2$, the two-dimensional unit sphere. For any $\mathbf{n}\in S^2$, that is, $\mathbf{n}=(n_1,n_2,n_3)$, $\Sigma n_i^2=1$, let us define

$$e(\mathbf{n})=\frac{(I+\sigma\cdot\mathbf{n})}{2},$$

where $\sigma=\{\sigma_i\}$, $i=1,2,3$, denote the Pauli matrices. It is clear that $e(\mathbf{n})$ is a one-dimensional projector corresponding to a point $\mathbf{n}\in S^2$. On M there is the unique U(2) invariant measure $d\mu$ normalized to $\mu(M)=1$. Let us notice that $\mathcal{A}_q=\{e(\mathbf{n}):\mathbf{n}\in S^2\}''$ and thus a semispectral measure connected with the quantum system is given by $E(B)=\int_B e(\mathbf{n})d\mu(\mathbf{n})$ with the property $E(S^2)=I/2$. To present the coupling we assume that:

(i) $\forall\mathbf{n}\in S^2\quad \nu_{\mathbf{n}}=\delta_e$, the point measure at the neutral element of U(2), and
(ii) $\forall\mathbf{n}\in S^2\quad h_{\mathbf{n}}:S^2\to S^2$ is the geodesic symmetry on the symmetric space $S^2=\mathrm{U}(2)/\mathrm{U}(1)\times\mathrm{U}(1)$.

Then $\tilde{V}_{\mathbf{n}'}(\mathbf{n})=e(\mathbf{n}')$ and so

$$L_*(\tilde{\rho})(\mathbf{n})=\int_{S^2}d\mu(\mathbf{n}')e(\mathbf{n}')\tilde{\rho}(\mathbf{n}'(\mathbf{n}))e(\mathbf{n}')-\frac{1}{2}\tilde{\rho}(\mathbf{n}).$$

Since there is no classical evolution, the total generator reads

$$\dot{\tilde{\rho}}_t=(\mathbf{1}\otimes\delta_q^{ad})\tilde{\rho}_t+L_*(\tilde{\rho}_t).$$

Let us derive the equation for the time evolution of the quantum density operator $\rho_t=\int\tilde{\rho}_t(\mathbf{n})d\mu(\mathbf{n})$. At first we calculate the integral $\int e(\mathbf{n})e(\mathbf{n}')e(\mathbf{n})d\mu(\mathbf{n})$ for fixed projector $e(\mathbf{n}')$. Because

$$e(\mathbf{n})e(\mathbf{n}')e(\mathbf{n})=\tfrac{1}{2}(1+\mathbf{n}'\cdot\mathbf{n})e(\mathbf{n}),$$

it is enough to compute

$$\int_{S^2}n_ie(\mathbf{n})d\mu(\mathbf{n})=\int_{S^2}n_i\left(\frac{I}{2}+\frac{\sigma\cdot\mathbf{n}}{2}\right)d\mu(\mathbf{n}).$$

However, $\int n_i \, d\mu(\mathbf{n}) = 0$ for all $i = 1, 2, 3$ and $\int n_i n_j \, d\mu(\mathbf{n}) = 0$ if $i \neq j$. Thus

$$\int_{S^2} n_i e(\mathbf{n}) d\mu(\mathbf{n}) = \frac{\sigma_i}{2} \int_{S^2} n_i^2 \, d\mu(\mathbf{n}) = \frac{\sigma_i}{6},$$

and so

$$\int_{S^2} e(\mathbf{n}) e(\mathbf{n}') e(\mathbf{n}) d\mu(\mathbf{n}) = \frac{1}{6}(e(\mathbf{n}') + I).$$

By linearity, we obtain that for every $A \in M_{2\times 2}$ there is

$$\int_{S^2} e(\mathbf{n}) A e(\mathbf{n}) d\mu(\mathbf{n}) = \frac{1}{6}(A + \mathrm{Tr}(A)).$$

So,

$$\begin{aligned} &\int_{S^2} d\mu(\mathbf{n}) \int_{S^2} d\mu(\mathbf{n}') e(\mathbf{n}') \tilde{\rho}_t(\mathbf{n}'(\mathbf{n})) e(\mathbf{n}') - \frac{1}{2} \int_{S^2} d\mu(\mathbf{n}) \tilde{\rho}_t(\mathbf{n}) \\ &= \int_{S^2} d\mu(\mathbf{n}') e(\mathbf{n}') \rho_t e(\mathbf{n}') - \frac{1}{2} \rho_t \\ &= \frac{1}{6}(I - 2\rho_t). \end{aligned}$$

In the above we use the fact that $\mathbf{n}'(\mathbf{n})$ does not change the measure $d\mu$. Thus

$$\dot{\rho}_t = L_q(\rho_t) = -i[H, \rho_t] + \tfrac{1}{6}(\mathrm{Tr}\, \rho_t I - 2\rho_t).$$

We show that every density operator dissipates to the totally mixed state $I/2$. At first we describe E_r the reversible and E_o the irreversible part of the generator L_q.

Proposition 3.2: Let $A \in M_{2\times 2}$. Then

$$L_q(A) = 0 \Leftrightarrow A = zI, \quad z \in \mathbf{C},$$

$$(L_q(A) = i\lambda A, \quad \lambda \in \mathbf{R} \setminus \{0\}) \Leftrightarrow A = 0.$$

Proof: Direct calculations. □

Hence $E_r = zI$. Because $E_r \cap E_o = 0$ and $E_r \oplus E_o = M_{2\times 2}$, therefore

$$E_0 = \begin{pmatrix} z & z_1 \\ z_2 & -z \end{pmatrix}.$$

However, (see Lemma 4.2 in Ref. 18) for $A \in E_o$ there is $\lim_{t\to\infty} \|T_t A\| = 0$, where T_t denotes the semigroup generated by L_q. Thus T_∞ is the projection from $M_{2\times 2}$ onto E_r, and so for any density matrix ρ, $T_\infty(\rho) = I/2$.

IV. CONCLUDING REMARKS

The presented framework starts with a phenomenological assumption. At the very beginning we divide the world into two parts: a classical and a quantum one, which are next assumed to interact. The coupling ensures a flow of information from the quantum system to the classical system and the influence of the classics on the dynamics of the quantum system in a dissipative way. The classical evolution on average is perturbed by the expectation value of some quantum operator.

In each practical use we can guess what constitutes events and forms a classical system. Here we discuss two concrete models. However, this approach seems to be flexible enough to describe more general situations. It was shown in Ref. 19 that the spacelike asymptotic classical electromagnetic field is a superselection rule of the theory of a quantum particle carrying an electric charge, that is, each vector state from one superselection sector has the same asymptotic electromagnetic field. Also, in Ref. 20 the question of whether the radiation field can generate molecular superselection rules was discussed. Hence some classical observables can be thought of as a fundamental part of the quantum theory. Although dealing with infinite classical systems requires some elements of the scheme to be clarified, we believe that the proposed model can be also applied to such cases.

Finally, let us point out that the evolution equation for statistical states given by the dynamical semigroup encodes the behavior of individual samples. In the first example we showed that (assuming the quantum Hamiltonian is equal to zero) the evolution of a pure state is given by the Poisson probability distribution. The stochastic process describing the evolution from the second example was discussed in Ref. 17. A general construction of a piecewise deterministic process associated to a dynamical semigroup of the coupled quantum and classical continuous systems was presented in Ref. 21. Those processes can provide a practical way for calculating numbers that are needed in experiments. Such an algorithm, which generalize that of the quantum Monte Carlo method, was given in Ref. 22.

ACKNOWLEDGMENTS

I would like to thank Professor A. Jadczyk for helpful discussions and Professor Ph. Blanchard for his hospitality in Bielefeld, where the paper was finished. The financial support from A. von Humboldt Foundation is also acknowledged.

APPENDIX A: PROOF OF THEOREM 2.4

Proof. Let $\{e_k\}_1^\infty$ be an orthonormal basis in $\mathcal{H}_q$. Because for every $x\in M$

$$\alpha(B)=0 \Rightarrow (\nu_x * E_{k,l})(B)=\langle e_k,(\nu_x * E)(B)e_l\rangle=0,$$

there is a Borel and α-integrable function $f_{k,l}(x,q)$ such that $(\nu_x{}^*E_{k,l})(B)=\int_B f_{k,l}(x,q)\,d\alpha(q)$ for every $B\in\mathcal{B}(Q)$. At first we modify functions $f_{k,l}$. We have that $f_{k,l}\in L^1(Q,\mathcal{B}(Q),d\alpha)$ and from Lemma 3.5 the map $x\rightarrow\int_B f_{k,l}(x,q)\,d\alpha(q)$ is Borel for every $B\in\mathcal{B}(Q)$. Moreover, Q, as a space which satisfies the second axiom of countability, is a separable Borel space. So from the selection of representative lemma[23] there is a Borel function $f^*_{k,l}:M\times Q\rightarrow\mathbf{C}$ and a Borel set $N_{k,l}\subset M$ of μ-measure zero such that $\forall x\in M\setminus N_{k,l}$ $f^*_{k,l}(x,q)=f_{k,l}(x,q)$ α almost everywhere. Let $N=\cup_{k,l}N_{k,l}$. Then $\mu(N)=0$ and $\forall x\in M\setminus N$ $f^*_{k,l}(x,q)=f_{k,l}(x,q)$ α a.e. for every k, $l\in\mathbf{N}$.

Let U be a subspace of $\mathcal{H}_q$ generated by finite linear combinations of e_k with coefficients z_k such that $\operatorname{Re} z_k$ and $\operatorname{Im} z_k$ are rational numbers. It is clear that U is dense and countable. Let $\phi_{x,q}$ be a sesquilinear form on U defined by

$$\phi_{x,q}(v_1,v_2)=\sum_{k,l}\overline{y_k}z_l f^*_{k,l}(x,q),\quad v_1=\sum_k y_k e_k,\quad v_2=\sum_l z_l e_l.$$

Because for fixed $v\in U$ the function $(x,q)\rightarrow\phi_{x,q}(v,v)$ is Borel, sets

$$\Theta_v=\{(x,q):|\phi_{x,q}(v,v)|\leqslant C\|v\|^2\},$$

$$\Xi_v=\{(x,q):\phi_{x,q}(v,v)\geqslant 0\}$$

are Borel subsets of $M\times Q$. Thus $\Omega:=(\cap_v\Theta_v)\cap(\cap_v\Xi_v)$ is also Borel. On this set the form $\phi_{x,q}$ may be extended to a positive and bounded form $\overline{\phi}_{x,q}$ on the whole $\mathcal{H}_q$. It means that there is a function $V:\Omega\to L(\mathcal{H}_q)_+$ such that $\forall\psi_1,\psi_2\in\mathcal{H}_q\,\langle\psi_1,V(x,q)\psi_2\rangle=\overline{\phi_{x,q}}(\psi_1,\psi_2)$ when $(x,q)\in\Omega$. Now we show that Ω is big enough, i.e., that $\mu\times\alpha(M\times Q\backslash\Omega)=0$. Let $v\in U$, i.e., $v=\sum_{k=1}^{n}z_k e_k$. Then for every $x\in M\backslash N$

$$\int_Q\phi_{x,q}(v,v)d\alpha(q)=\sum_{k,l=1}^{n}\overline{z_k}z_l\int_Q f_{k,l}^{*}(x,q)d\alpha(q)$$
$$=\sum_{k,l=1}^{n}\overline{z_k}z_l\int_Q f_{k,l}(x,q)d\alpha(q)=(\nu_x*E_{v,v,})(Q)\geqslant 0.$$

So there is a Borel subset $B_{x,v}\subset Q$ such that $\alpha(B_{x,v})=0$ and $\forall q\in Q\backslash B_{x,v}\ \phi_{x,q}(v,v)\geqslant 0$. Moreover, for α-finite $B\in\mathcal{B}(Q)$ we have that

$$\int_B\phi_{x,q}(v,v)d\alpha(q)=\langle v,(\nu_x*E)(B)v\rangle\leqslant C\|v\|^2\alpha(B).$$

Thus for every Q_n $[Q=\cup_{n=1}^{\infty}Q_n$ and $\alpha(Q_n)\leqslant\infty]$ there is a Borel subset $C_{x,n,v}\subset Q_n$ such that $\alpha(C_{x,n,v})=0$ and $\forall q\in Q_n\backslash C_{x,n,v}\ \phi_{x,q}(v,v)\leqslant C\|v\|^2$. Let

$$D_x=\left(\bigcup_v B_{x,v}\right)\cup\left(\bigcup_n\bigcup_v C_{x,n,v}\right).$$

Then $\alpha(D_x)=0$ and $\forall x\in M\backslash N\forall q\in Q\backslash D_x$, $\phi_{x,q}$ is positive and bounded by constant C. Let $\Omega_0=\{(x,q):x\in M\backslash N,q\in Q\backslash D_x\}$. Because $\Omega_0\subset\Omega$, then $M\times Q\backslash\Omega\subset M\times Q\backslash\Omega_0$. However, $M\times Q\backslash\Omega$ is Borel, so from the Fubini theorem

$$\mu\times\alpha(M\times Q\backslash\Omega)=\int_M\alpha[(M\times Q\backslash\Omega)_x]d\mu(x)=\int_{M\backslash N}\alpha[(M\times Q\backslash\Omega)_x]d\mu(x).$$

Because $\forall x\in M\backslash N$

$$(M\times Q\backslash\Omega)_x=\{q\in Q:(x,q)M\times Q\backslash\Omega\}\subset D_x,$$

we have that $\alpha[(M\times Q\backslash\Omega)_x]\leqslant\alpha(D_x)=0$ and so $\mu\times\alpha(M\times Q\backslash\Omega)=0$. Let us define

$$\tilde{V}(x,q)=\begin{cases}V(x,q), & (x,q)\in\Omega,\\ 0\quad, & \text{elsewhere}\end{cases}.$$

It is clear that $(x,q)\to\tilde{V}(x,q)$ is weakly Borel measurable and $\sup_{x,q}\|\tilde{V}(x,q)\|\leqslant C$. Moreover, for every $x\in M\backslash N$, $B\in\mathcal{B}(Q)$, $\psi_1,\psi_2\in\mathcal{H}_q$,

$$\int_B\langle\psi_1,\tilde{V}(x,q)\psi_2\rangle d\alpha(q)=\int_B\overline{\phi_{x,q}}(\psi_1,\psi_2)d\alpha(q)=\langle\psi_1,(\nu_x*E)(B)\psi_2\rangle.$$

Now we check the uniqueness of $\tilde{V}$ up to a set of $\mu\times\alpha$ measure zero. Let $\tilde{V}'$ be another map which satisfies the thesis. So there is a Borel subset $N'\subset M$ such that $\mu(N')=0$ and for every $x\in M\backslash N'$

$$\langle\psi_1,(\nu_x*E)(B)\psi_2\rangle=\int_B\langle\psi_1,\tilde{V}'(x,q)\psi_2\rangle d\alpha(q).$$

Because $\tilde{V}$ and $\tilde{V}'$ are both weakly Borel measurable, a set

$$\Gamma_v=\{(x,q):\langle v,\tilde{V}(x,q)v\rangle\neq\langle v,\tilde{V}'(x,q)v\rangle\}$$

is Borel for every $v\in V$. Let $\Gamma=\cup_v\Gamma_v$. It follows that quadratic forms $v\to\langle v,\tilde V(x,q)v\rangle$ and $v\to\langle v,\tilde V'(x,q)v\rangle$ agree on $M\times Q\backslash\Gamma$. As they are both bounded the operator functions $\tilde V$ and $\tilde V'$ differ from each other only on Γ. Because $\mu(N\cup N')=0$, we have that

$$\mu\times\alpha(\Gamma)=\int_M \alpha(\Gamma_x)d\mu(x)=\int_{M\backslash(N\cup N')}\alpha(\Gamma_x)d\mu(x).$$

However, $\forall x\in M\backslash(N\cup N')$, $\tilde V(x,q)=\tilde V'(x,q)$ α a.e. so $\alpha(\Gamma_x)=0$ and $\mu\times\alpha(\Gamma)=0$. □

APPENDIX B: PROOF OF LEMMA 2.8

Proof. W_n is the weak limit of completely positive maps and W is the weak limit of W_n. Because $\forall k\in\mathbf{N}$, $W\otimes\mathbf{1}_k$, $\mathbf{1}_k\in M_{k\times k}$, is the weak limit of positive maps, for any positive operator $\hat A$ from $\mathcal{A}_T\otimes M_{k\times k}$ $W\otimes 1_k(\hat A)$ is positive. To prove that W is normal we have to show that W is an adjoint map of some bounded $T:\mathrm{Tr}(\tilde{\mathcal{H}})\to\mathrm{Tr}(\tilde{\mathcal{H}})$. Let $\tilde\rho\in\mathrm{Tr}(\tilde{\mathcal{H}})_+$. There exists a basis $\{e_m\}$ in $\tilde{\mathcal{H}}$ such that $\tilde\rho=\Sigma_{m=1}^{\infty}c_mP_m$, where P_m is a projector onto e_m. At first we show that

$$T_n(\tilde\rho)=\int_{Q^n}d\alpha(q)\tilde W_q\tilde\rho\tilde W_q^*$$

exists in $\mathrm{Tr}(\tilde{\mathcal{H}})$.

Because it exists in $L(\tilde{\mathcal{H}})_+$ and

$$S_k=\sum_{m=1}^{m=k}\langle e_m,T_n(\tilde\rho)e_m\rangle=\sum_{m=1}^{m=k}\int_{Q^n}d\alpha(q)\langle e_m,\tilde W_q\tilde\rho\tilde W_q^*e_m\rangle$$
$$\leqslant\int_{Q^n}d\alpha(q)\|\tilde W_q\tilde\rho\tilde W_q^*\|_{\mathrm{Tr}}\leqslant\alpha(Q^n)(\sup_q\|\tilde W_q\|^2)\|\tilde\rho\|_{\mathrm{Tr}},$$

so $T_n(\tilde\rho)\in\mathrm{Tr}(\tilde{\mathcal{H}})_+$. Moreover,

$$\int_Q d\alpha(q)\|\tilde W_q\tilde\rho\tilde W_q^*\|_{\mathrm{Tr}}=\int_Q d\alpha(q)\sum_{m=1}^{\infty}\langle e_m,\tilde W_q\tilde\rho\tilde W_q^*e_m\rangle$$
$$=\int_Q\sum_{m=1}^{\infty}\|\tilde\rho^{1/2}\tilde W_q^*e_m\|^2$$
$$=\int_Q d\alpha(q)\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}|\langle\tilde\rho^{1/2}\tilde W_q^*e_m,e_n\rangle|^2.$$

However, $\tilde\rho^{1/2}\tilde W_q^*\in HS(\tilde{\mathcal{H}})$. So, we get

$$\int_Q d\alpha(q)\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}c_n|\langle e_m,\tilde W_qe_n\rangle|^2=\int_Q d\alpha(q)\sum_{n=1}^{\infty}c_n\|\tilde W_qe_n\|^2$$
$$=\sum_{n=1}^{\infty}c_n\int_Q d\alpha(q)\|\tilde V_qe_n\|^2$$
$$=\sum_{n=1}^{\infty}c_n\int_Q d\alpha(q)\langle e_n,\tilde V_q^2e_n\rangle$$

$$\leqslant C\sum_{n=1}^{\infty} c_n\int_Q d\alpha(q)\int_M d\mu(x)\langle e_n(x),\tilde{V}(x,q)e_n(x)\rangle$$

$$=C\sum_{n=1}^{\infty} c_n\int_M d\mu(x)\langle e_n(x),(\nu_x * E)(Q)e_n(x)\rangle$$

$$\leqslant C(\sup_x\|(\nu_x * E)(Q)\|)\|\tilde{\rho}\|_{\mathrm{Tr}}$$

$$=C\|E(q)\|(\sup_x \nu_x(G))\|\tilde{\rho}\|_{\mathrm{Tr}}=CC_1\|E(Q)\|\cdot\|\tilde{\rho}\|_{\mathrm{Tr}}.$$

Thus $T(\tilde{\rho})=\int_Q d\alpha(q)\tilde{W}_q\tilde{\rho}\tilde{W}_q^*\in \mathrm{Tr}(\tilde{\mathcal{H}})_+$ and $\mathrm{Tr}(T(\tilde{\rho})\tilde{A})=\mathrm{Tr}(\tilde{\rho}W(\tilde{A}))$ for any $\tilde{A}\in\mathcal{A}_T$. ■

[1] H. P. Stapp, *Mind, Matter and Quantum Mechanics* (Springer Verlag, Berlin, 1993).
[2] R. Haag, in *Quantum Control and Measurement*, Proc. ISQM Satellite Workshop, edited by H. Ezawa and Y. Murayama (North Holland, Amsterdam, 1995).
[3] R. Haag, Commun. Math. Phys. **132**, 245 (1990).
[4] Ph. Blanchard and A. Jadczyk, Phys. Lett. A **175**, 157 (1993).
[5] Ph. Blanchard and A. Jadczyk, Phys. Lett. A **183**, 272 (1993).
[6] Ph. Blanchard and A. Jadczyk, in *Proceedings of the Symposium on Foundations of Modern Physics*, Cologne, edited by P. Mittelstaedt (World Scientific, Singapore, 1994).
[7] Ph. Blanchard and A. Jadczyk, in *Advances in Dynamical Systems and Quantum Physics*, Capri, edited by R. Figari (World Scientific, Singapore, 1994).
[8] Ph. Blanchard and A. Jadczyk, in *Proceedings of the Third Max Born Symposium, Stochasticity and Quantum Physics*, Sobotka, edited by Z. Haba *et al.* (Kluwer Academic, Dordrecht, 1995).
[9] A. Jadczyk, Prog. Theor. Phys. **93**, 631 (1995).
[10] A. Jadczyk, Found. Phys. **25**, 743 (1995).
[11] A. Jadczyk, G. Kondrat, and R. Olkiewicz, J. Phys. A **30**, 1863 (1997).
[12] R. Olkiewicz, Rev. Math. Phys. **9**, 719 (1997).
[13] A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[14] A. C. Holevo, *Quantum Probability and Quantum Statistics*, Itogi Nauki i Techniki Vol. 83, (Winiti, Moscow, 1991) (in Russian).
[15] E. Calzetta and B. L. Hu, Phys. Rev. D **35**, 495 (1987).
[16] B. L. Hu, Physica A **158**, 399 (1989).
[17] A. Jadczyk, in *Infinite Dimensional Geometry, Noncommutative Geometry, Operator Algebras and Fundamental Interactions*, Proceedings of the First Caribbean School of Mathematicas and Theoretical Physics, Saint Francois, Guadeloupe, edited by R. Coquereaux (World Scientific, Singapore, 1995).
[18] R. Nagel, in *Semigroups of Linear and Nonlinear Operations and Applications*, edited by G. R. Goldstein and J. A. Goldstein (Kluwer Academic, Dordrecht, 1993).
[19] D. Buchholz, Phys. Lett. B **174**, 331 (1986).
[20] A. Amann, J. Math. Chem. **6**, 1 (1991).
[21] Ph. Blanchard and R. Olkiewicz, ''Interacting Quantum and Classical Continuous Systems I. The Piecewise Deterministic Dynamics,'' J. Stat. Phys. **94**, (1999) (in press).
[22] Ph. Blanchard and A. Jadczyk, Ann. Phys. **4**, 583 (1995).
[23] V. S. Varadarajan, *Geometry of Quantum Theory*, (Van Nostrand Reinhold, New York, 1970), Vol. 2.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Arithmetic properties of spectra produced by Farey hierarchies of approximants

O. Radulescu[a] and T. Janssen
Institute of Theoretical Physics, Nijmegen University, Postbus 9010, 6500 GL Nijmegen, The Netherlands

(Received 12 June 1998; accepted for publication 30 November 1998)

We discuss the consequences of the hierarchical nature of series of approximants of aperiodic crystals on their diffraction patterns and spectra of elementary excitations. We show how a linear form defined on $\mathbb{Z}^3$ can be used to order Bragg reflections in diffraction patterns according to their amplitudes, and gaps in spectra of elementary excitations according to their widths, for all the structures in the hierarchy. Bragg peaks amplitudes and gap widths are projective functions on $\mathbb{P}(\mathbb{Z}^3)$, recursively defined on 2D Farey sets (generalization of Farey series).
 [S0022-2488(99)00603-9]

I. INTRODUCTION

Several aperiodic ordered systems occurring in solid state physics, such as incommensurate modulated structures, composite compounds, and quasicrystals, have complex phase diagrams containing also series of commensurate structures called approximants. It is possible to pass from one approximant to another in a series, by altering the chemical nature of the building units, or by changing the thermodynamic equilibrium parameters (temperature, composition, pressure). In certain cases, approximants are dynamically selected nonequilibrium phases.

The series of approximants is hierarchical, members of this series having strong mutual structural relations. The universality of the structural relations between approximants and the consequences of these relations on their physical properties can be understood if Farey series are used to index approximants. A short summary of the properties of the Farey series is given in Appendix A. In this paper we discuss only one-dimensional (1D) structures, but our approach could in principle be generalized to higher dimension as well.

Different approximants can be embedded in a Farey hierarchy (see Appendix A for precise definitions). For each rational number p/q the corresponding approximant is called (p,q) commensurate structure and the integer numbers (p,q) are called commensurability indices. The elementary step in constructing a Farey hierarchy is the mediant construction (well known for Farey series): the unit cell of an approximant with commensurability indices (p_1+p_2,q_1+q_2) is obtained by concatenating the unit cells of two approximants with commensurability indices (p_1,q_1) and (p_2,q_2), respectively.

The periods of (p,q) commensurate structures obtained by concatenation have the general form $L_{p,q}=a(\beta p+\alpha q)$ (see Appendix A), and therefore their diffraction patterns consist of equidistant Bragg reflections with a distance $2\pi/L_{p,q}$ between consecutive reflections situated at

$$k_{p,r,q}=\frac{2\pi r}{a(\beta p+\alpha q)}. \tag{1}$$

The structure factor $h(p,r,q)$ for a reflection of wave vector $k_{p,r,q}$ is defined as the complex amplitude of the reflection divided by the total number of atoms in the unit cell, which for

[a]Present address: IRC in Polymer Science and Technology, Department of Physics, University of Leeds, Leeds LS2 9JT, UK. Electronic mail: phyor@irc.leeds.ac.uk

approximants generated by concatenation is $n_{p,q}=sp+tq$ (Appendix A). If for simplicity the atomic scattering factor is considered equal to one for all atoms, then $h(p,r,q) =(1/n_{p,q})\Sigma_{i=1}^{n_{p,q}}\exp(ikx_i)$, x_i being the atomic positions in the unit cell.

Let us define the family of complex functions $F_{p,q}(z)=\Sigma_{i=1}^{n_{p,q}}z^{x_i/a}$. It follows

$$h(p,r,q)=\frac{1}{sp+tq}F_{p,q}[\exp(ik_{p,r,q}a)],$$
$$F_{p_1+p_2,q_1+q_2}(z)=F_{p_1,q_1}(z)+z^{\beta p_1+\alpha q_1}F_{p_2,q_2}(z), \tag{2}$$

whenever $p_1/q_1,p_2/q_2$ are Farey-consecutive.

$F_{p,q}$ is uniquely defined by Eq. (2) for all p,q relatively prime and $0\leqslant p/q\leqslant 1$, once $F_{0,1}(z)$ and $F_{1,1}(z)$ are given.

At this level we may notice the occurrence of $\mathbb{Z}^3$ in the problem. Bragg peaks for all the structures in the hierarchy can be labeled with three integers, two being the commensurability indices and the third one indicating the position of the Bragg reflection along the reciprocal line. Equation (2) illustrates the special two indices recursion on Farey series satisfied by the structure factor. It is straightforward from Eq. (1) that positions of Bragg reflections are only depending on the ratios p/q, r/q. From Eq. (2) it follows that the same holds for the structure factor, i.e., for the amplitude of Bragg peaks. Thus, Bragg reflections for the entire Farey hierarchy of approximants are in correspondence with directions of $\mathbb{Z}^3$, i.e., with points in the projective module $\mathbb{P}(\mathbb{Z}^3)$.

It is known that for crystals there is a connection between Bragg reflections and gaps in the spectra of elementary excitations (electrons, phonons). This is true more generally for almost periodic structures generated by substitutions.[1] We shall show that when the structure belongs to a Farey hierarchy, then gap widths obey a recursion having the same form as the one obeyed by the structure factor [Eq. (2)], with coefficients α,β that may be different. But first of all let us clear up the relation between Bragg reflections and gaps in 1D models for phonons and electrons in crystals (periodic and aperiodic). A review of these models can be found elsewhere.[2]

Energy of electrons in a 1D periodic crystal follows from the Schrödinger equation with periodic potential:

$$(H_o+V_{p,q})\Phi=E\Phi,$$
$$H_o=-\frac{\hbar^2\nabla^2}{2m}, \tag{3}$$
$$V_{p,q}(x+L_{p,q})=V_{p,q}(x).$$

In the nearly free electron approximation the electron energy spectrum is obtained by using degenerate first-order Rayleigh–Schrödinger perturbation theory in the space spanned by the two planar waves $\exp(\pm ikx)$. A gap of width $\Delta_{p,r,q}=2\tilde{V}_{p,q}(k_{p,r,q})$ ($\tilde{V}_{p,q}$ is the Fourier transform of $V_{p,q}$) opens if $2k=k_{p,r,q}$, for some Bragg wave vector $k_{p,r,q}$. The energetic positions of the gaps are given, like the positions of the Bragg reflection, by a function of p/q and r/q, $E_{p,r,q}=\hbar^2k_{p,r,q}^2/8m$. The correspondence between Bragg reflections and gaps in the electron energy spectrum (within nearly free electron approximation) is 1–1 and, supposing that the periodic potential $V_{p,q}$ mimics the distribution of atoms, strong Bragg peaks correspond to wide gaps. The origin of gaps is the Bragg reflection of planar waves on the periodic potential $V_{p,q}$.

Furthermore, for a Farey hierarchy of structures $V_{p,q}$ obey

$$V_{p_1+p_2,q_1+q_2}(x)=\begin{cases}V_{p_1,q_1}(x), & x\in U_{p_1,q_1},\\ V_{p_2,q_2}(x-L_{p_1,q_1}), & x\in U_{p_2,q_2}+L_{p_1,q_1}.\end{cases}$$

If we define the following family of complex functions $F_{p,q}(z)=\tilde{V}_{p,q}(\log(z)/ia)$, then the absolute value of $F_{p,q}$ is half the gap width, i.e., $|F_{p,q}(\exp(ik_{p,r,q}a))|=|\tilde{V}_{p,q}(k_{p,r,q})|$. The functions $F_{p,q}(z)$ obey exactly the same recursion (with the same values of the coefficients α,β) as the similar functions giving the structure factor [Eq. (2)].

Of course, the 1–1 correspondence between Bragg reflections and gaps in the spectrum of nearly free electrons is only an approximation. Actually, for periodic structures the number of gaps is finite and the number of Bragg reflections is infinite (the correspondence can be 1–1 only for aperiodic, almost periodic structures).

A different approximation is the discretized version of Eq. (3):

$$H_o\Phi_n+V_n^{p,q}\Phi_n=E\Phi_n,$$

$$H_o\Phi_n=-\Phi_{n-1}-\Phi_{n+1}+2\Phi_n, \tag{4}$$

$$V_{n+q}^{p,q}=V_n^{p,q}.$$

This equation describes tight-binding electrons with modulated onsite potential[2] but also phonon excitations for harmonic chains under influence of a periodic substrate potential (Frenkel–Kontorova models[3]).

In this case degenerate perturbation theory applies to the Bloch waves $\exp(\pm ikn)$ and the origin of gaps is the Bragg reflection of Bloch waves on the superlattice potential $V^{p,q}$ (notice that if one defines the translation operators $T_q\Phi_n=\Phi_{n+q}$, $[H_o,T_1]=0, [V^{p,q},T_q]=0$, the above Bloch waves are eigenvectors of T_1). There are $q-1$ gaps, appearing at wave vectors obeying $k=-k+(2\pi r/q)(\mathrm{mod}\,2\pi)$. Each gap is related to an infinite family $\{2\pi r/q+2\pi n\}_{n\in\mathbb{Z}}$ of Bragg reflections. The energetic positions of the gaps are $E_{p,r,q}=4\sin^2(2\pi r/q)$, while their widths are $\Delta_{p,r,q}=(1/q)|\tilde{V}^{p,q}(2\pi r/q)|$, where $\tilde{V}^{p,q}(k)=\Sigma_{n=0}^{q-1}V_n^{p,q}\exp(ikn)$. We may notice the absence of a characteristic length a in the expression of the gap widths. The Euclidean metric occurring in the calculation of the structure factor becomes the graph metric along the chain of atomic sites. Tight-binding calculations for electrons or chain models for phonons assume instantaneous propagation of waves from one site to another. In this approximation the relation between amplitudes of Bragg peaks and widths of gaps may be nonmonotonic.

The hierarchical properties of gap widths are a consequence of

$$V_n^{p_1+p_2,q_1+q_2}=\begin{cases} V_n^{p_1,q_1}, & n=0,\ldots,q_1-1,\\ V_{n-q_1}^{p_2,q_2}, & n=q_1,\ldots,q_1+q_2.\end{cases}$$

Defining $F_{p,q}(z)=\tilde{V}^{p,q}((1/i)\log(z))$, then $\Delta_{p,r,q}=(1/q)|F_{p,q}(\exp(2\pi r/q)i)|$ and the functions $F_{p,q}(z)$ obey similar recursions as the functions giving the structure factor [Eqs. (2), with $\alpha=1$, $\beta=0$, $s=0,t=1$]. α and β are now the coefficients characterizing the linear dependence of the graph length of the unit cell (number of successive chain atoms in the unit cell) on the commensurability indices (p,q). These are generally different from the values expressing the dependence of the Euclidean length on (p,q), occurring in the structure factor problem. For Frenkel–Kontorova models the substrate is rigid with a fixed period a, therefore the length of the unit cell is pa $(\alpha=0,\beta=1)$, while the graph length scales as q $(\alpha=1,\beta=0)$.

A more complex Farey hierarchy is described by the double chain model. This model was designed[4] for the study of phonons in composite structures. The modulation is no longer diagonal because there are two interacting harmonic subsystems. The double chain infinite-dimensional dynamical matrix (second derivative of the configuration potential energy) acts as

$$D(y_n^{(1)},y_m^{(2)})=\left(-(y_{n+1}^{(1)}+y_{n-1}^{(1)})+y_n^{(1)}\left(2+\sum_m \chi_{n,m}\right)-\sum_m \chi_{n,m}y_m^{(2)},\right.$$
$$\left.-(y_{m+1}^{(2)}+y_{m-1}^{(2)})+y_m^{(2)}\left(2+\sum_n \chi_{n,m}\right)-\sum_n \chi_{n,m}y_n^{(1)}\right). \tag{5}$$

The eigenvalues of the dynamical matrix form the phonon spectrum. We have shown[4] that in the case when the interaction between the two chains is harmonic and short range represented by

$$\chi_{n,m}=\begin{cases}\chi, & m=\left\lfloor \dfrac{nq}{p}+\dfrac{1}{2}\right\rfloor,\\ 0, & \text{elsewhere},\end{cases}$$

the gap widths are given by the same type of recursion relations as in Eq. (2), but with $\alpha=1,\beta=1$. The ''effective'' graph length of the unit cell scales like $p+q$.

According to these introductory remarks, amplitudes of Bragg peaks in the diffraction spectra and widths of gaps in the spectra of the elementary excitations of approximant structures belonging to a Farey hierarchy are functions defined on the projective module $\mathbb{P}(\mathbb{Z}^3)$, i.e., on the set of fractions having common denominator $(p/q,r/q)$. These functions obey special recursion relationships, involving Farey series. The purpose of this paper is to study the properties of these kinds of functions and to analyze their consequences for the spectral properties of approximants.

The structure of this paper is the following. In the next section we introduce and discuss the properties of Farey sets that are the 2D analog of Farey series. The Farey sets were used by Kim and Oslund[5] for the problem of simultaneous approximation of pairs of real numbers by pairs of rational numbers of common denominator. We prove here a set of properties of Farey sets that are direct extensions of properties of Farey series, and that are useful for our purpose and could be also used in other applications. In the third section we discuss a representation of rational numbers that comes naturally out of the previous construction and has heuristic utility for many of the reasonings involving Farey series. In the fourth section we introduce functions that are hierarchically defined on Farey sets, and prove a theorem revealing their arithmetical structure. For Farey hierarchies of structures, widths of gaps in the spectra of elementary excitations and amplitudes of Bragg peaks are functions hierarchically defined on Farey sets. In the last section we discuss the consequences of this on the structure of spectra generated by Farey hierarchies of structures. The proofs of properties and theorems can be found in Appendix B.

II. TWO-DIMENSIONAL FAREY SETS

The Farey series $\mathcal{F}_n$ are included in the one-dimensional projective module $\mathbb{P}(\mathbb{Z}^2)$ that is composed of all directions in $\mathbb{Z}^2$. The mapping $(p,q)\rightarrow p/q$ identifies $\mathbb{P}(\mathbb{Z}^2)$ with the set of rational numbers extended by the point at infinity $\hat{\mathbb{Q}}:=\mathbb{Q}\cup\infty$.

Similarly, one may define a two-dimensional projective module $\mathbb{P}(\mathbb{Z}^3)$ consisting of all directions in $\mathbb{Z}^3$, and the mapping $(p,r,q)\rightarrow(p/q,r/q)$ allows us to identify $\mathbb{P}(\mathbb{Z}^3)$ with the set of pairs of fractions $\{(p/q,r/q)|gcd[p,q,r]=1,q\geqslant 0\}$ having a common denominator and such that q is positive and p,r,q are relatively prime. As *(p,r,q)* and *(-p,-r,-q)* represent the same direction in $\mathbb{Z}^3$ we may conventionally impose $q\geqslant 0$. The logic of this convention will show up later. The above application also maps the surface of a sphere onto the plane (Fig. 1) and is known in cartography under the name of gnomonic projection. This projection has the remarkable property of transforming geodesics (great circles) on the sphere into straight lines and proved its utility in airline navigation, meteors maps in astronomy, and indexing of the Laue diagrams in x-ray crystallography. It has the disadvantage of not being conformal (it distorts angles). The gnomonic projection embeds $\mathbb{P}(\mathbb{Z}^3)$ in a Euclidean model of the projective plane,[6] provided that the line at infinity is considered as well. For $x=(p/q,r/q)\in\mathbb{P}(\mathbb{Z}^3)$ we denote the common denominator by $\mathrm{den}(x)=q\geqslant 0$.

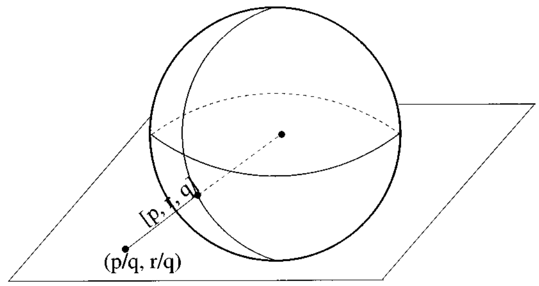

FIG. 1. Gnomonic or ''central'' projection: a direction of indices $[p,r,q]$ of the lattice with origin in the center of the sphere cuts the tangent plane in a point of coordinates $(p/q,r/q)$.

A partial projection $(p,r,q)\to(p/q,r)$ has as image $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ that can be identified to the subset $\{(p/q,r/q)|gcd[p,q]=1\}$ of $\mathbb{P}(\mathbb{Z}^3)$ via the injection $(p/q,r)\to(p/q,r/q)$. By ''abus de langage'' we shall denote this subset with $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ whenever confusion is not dangerous.

Definition 1: A direction in $\mathbb{P}(\mathbb{Z}^3)$ *is the 1D projective module* $\Delta_{h,k,l}:=\{(p/q,r/q)|ph+rk+ql=0\}$. *It is the image via the gnomonic projection of a plane of* $\mathbb{Z}^3$ *containing the origin, and whose normal has indices (h,k,l). We shall refer to the indices (h,k,l), gcd[h,k,l]=1, as the Miller indices of the direction. An orientation is conventionally chosen, by imposing* $k\geqslant 0$.

Definition 2: The mediant of two points $x_1=(p_1/q_1,r_1/q_1),x_2=(p_2/q_2,r_2/q_2)$ *is defined as* $x_1\oplus x_2:=((p_1+p_2)/(q_1+q_2),(r_1+r_2)/(q_1+q_2))$ *(see also Ref. 5).*

Example 1: In Fig. 2 $(\frac{1}{2},\frac{1}{2})=(\frac{0}{1},\frac{1}{1})\oplus(\frac{1}{1},\frac{0}{1})$, $(\frac{2}{3},\frac{1}{3})=(\frac{1}{2},\frac{1}{2})\oplus(\frac{1}{1},\frac{0}{1})$.

Definition 3: The external product of two points $x_1=(p_1/q_1,r_1/q_1),x_2=(p_2/q_2,r_2/q_2)$, $gcd[p_1,r_1,q_1]=1$, $gcd[p_2,r_2,q_2]=1$, $q_1,q_2>0$, *is defined as the following vector of* $\mathbb{Z}^3$: $x_1\wedge x_2:=(r_1q_2-q_1r_2,q_1p_2-p_1q_2,p_1r_2-r_1p_2)$.

Remark 1: The external product of two different points along the same direction is an integral multiple of (h,k,l), the set of Miller indices of the direction. This clears up the conventional choice $q\geqslant 0$ that was made in order to have $q_1p_2-p_1q_2>0$ and thus $x_1\wedge x_2$ and (h,k,l) to be in the same direction as soon as $p_1/q_1<p_2/q_2$.

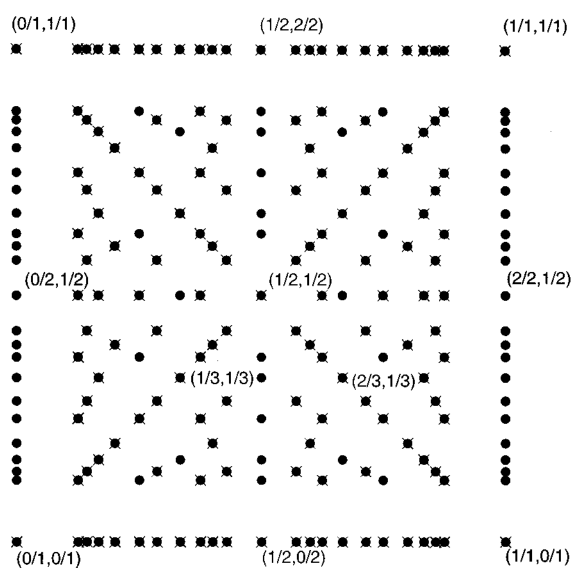

FIG. 2. The 2D Farey set of order 8 in the Euclidean plane: dots are points of $\mathcal{F}_8^{(2,0)}$, while crosses mark the subset $\mathcal{F}_8^{(1,1)}$.

The following theorem shows that the external product of two different points along the same direction of $\mathbb{P}(\mathbb{Z}^3)$ takes all values integral multiples of (h,k,l).

Theorem 4: *For any $x \in \mathbb{P}(\mathbb{Z}^3)$ along the direction $\Delta_{h,k,l}$ of Miller indices (h,k,l), there is $y \in \Delta_{h,k,l}$ such that $x \wedge y = (h,k,l)$.*

Definition 5: We call Farey set of order n in $\mathbb{P}(\mathbb{Z}^3)$ the set $\mathcal{F}_n^{(2,0)} = \{(p/q, r/q) | gcd[p,r,q] = 1, 0 \leq p/q \leq 1, 0 \leq r/q \leq 1, 0 < q \leq n\}$

Definition 6: We call Farey set of order n in $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ the set $\mathcal{F}_n^{(1,1)} := \mathcal{F}_n^{(2,0)} \cap [\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}] = \{(p/q, r/q) | gcd[p,q] = 1, 0 \leq p/q \leq 1, 0 \leq r/q \leq 1, 0 < q \leq n\}$

Remark 2: $\mathcal{F}_n^{(1,1)}$ is a subset of $\mathcal{F}_n^{(2,0)}$ (see Fig. 2). $\mathcal{F}_n^{(1,1)}$ and $\mathcal{F}_n^{(2,0)}$ coincide along some special directions, as shown by the following property.

Property 7: $\Delta_{h,k,l} = \Delta_{h,k,l} \cap [\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}]$ iff $k=1$. This is valid iff $\Delta_{h,k,l}$ contains two points $(p/q, r/q)$, $(p_1/q_1, r_1/q_1)$ such that

(i) $gcd[p,q] = 1$, $gcd[p_1,q_1] = 1$, and
(ii) *p_1/q_1 and p/q are Farey consecutive, i.e., $q_1 p - p_1 q = 1$.*

Definition 8: We call two points $x_1 = (p_1/q_1, r_1/q_1), x_2 = (p_2/q_2, r_2/q_2)$ in $\mathbb{P}(\mathbb{Z}^3) \cap [0,1]^2$ *Farey-consecutive if they belong to a direction $\Delta_{h,k,l}$, being oriented such that $x_1 \wedge x_2 = \lambda(h,k,l), \lambda > 0$, and all the points $x = (p/q, r/q)$ between them on $\Delta_{h,k,l}$ obey $q > \max(q_1, q_2)$.*

This definition is equivalent to the following:

Property 9: Two points $x_1, x_2 \in \mathbb{P}(\mathbb{Z}^3) \cap [0,1]^2$ are Farey-consecutive iff for some n and for some direction $\Delta_{h,k,l}$, $x_1, x_2 \in [\Delta_{h,k,l} \cap \mathcal{F}_n^{(2,0)}]$ and *$x_1 \wedge x_2 = \lambda(h,k,l), \lambda > 0$ and there are no other points of $\mathcal{F}_n^{(2,0)}$ between them along $\Delta_{h,k,l}$.*

The following properties are extensions of similar properties of Farey series.

Property 10: $x_1 = (p_1/q_1, r_1/q_1)$ and $x_2 = (p_2/q_2, r_2/q_2)$, $x_1, x_2 \in \mathbb{P}(\mathbb{Z}^3) \cap [0,1]^2$ are Farey-consecutive iff $x_1 \wedge x_2 = (h,k,l)$, where h,k,l are the Miller indices of the direction passing through them. In this case, $r_2/(\beta p_2 + \alpha q_2) - r_1/(\beta p_1 + \alpha q_1) = (\beta l - \alpha h)/[(\beta p_1 + \alpha q_1)(\beta p_2 + \alpha q_2)]$, where $\alpha, \beta \in \mathbb{R}$.

Remark 3: The similar property of Farey series is: $p_1/q_1 < p_2/q_2$ are Farey consecutive iff $q_1 p_2 - q_2 p_1 = 1$. (Th. 2 of Neville,[7] Th. 28 of Hardy and Wright,[8] and Th. 10.2 of Hua.[9])

Property 11: Two points $x_1 = (p_1/q_1, r_1/q_1), x_2 = (p_2/q_2, r_2/q_2)$ on a segment $\Delta_{h,k,l} \cap [0,1]^2$ with $k=1$ are Farey-consecutive iff p_1/q_1, p_2/q_2 are Farey-consecutive.

Property 12: If $x_1, x_2, x_3 \in \mathbb{P}(\mathbb{Z}^3) \cap [0,1]^2$ are three collinear, Farey-consecutive points, x_2 being between x_1, x_3 (i.e., $x_1 \wedge x_2 = \lambda(x_2 \wedge x_3), \lambda > 0$), *then $x_2 = x_1 \oplus x_3$.*

Remark 4: The similar property of Farey series is the following.

For three Farey-consecutive rationals, the one in the middle is the mediant of the other two. (Th. 29 in Hardy and Wright,[8] and Th. 10.3 in Hua.[9])

Property 13: The equation $x = x_1 \oplus x_2$ has a unique solution in $\Delta_{h,k,l} \cap [0,1]^2$, with *$x_1, x_2$ Farey-consecutive. x_1, x_2 are called right and left ascendants of x along the direction $\Delta_{h,k,l}$,* respectively, and they satisfy $x_1 \wedge x = (h,k,l)$, $x \wedge x_2 = (h,k,l)$.

Remark 5: The similar property of Farey series is the following.

The equation $p_1/q_1 \oplus p_2/q_2 = p/q$, with $p_1/q_1 < p_2/q_2$ Farey-consecutive has an unique solution. We call p_1/q_1 and p_2/q_2 left and right ascendant of p/q, respectively. (This follows from Th. 5 in Neville[7] and Remarks 3,4.)

Example 2: In Fig. 2 $(\frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{1}, \frac{0}{1})$ are the ascendants of $(\frac{2}{3}, \frac{1}{3})$ along the direction $(1,1,-1)$.

III. FAREY GRAPH

Let us consider the subset $\{(p/q, 1/q) | gcd[p,q] = 1\}$ of $\mathbb{P}(\mathbb{Z}^3)$, which can be identified to $\hat{\mathbb{Q}}$ via the mapping $(p/q, 1/q) \to p/q$. This representation of rational numbers[10] has several useful properties that were the heuristic source of many of the results in this paper (see Fig. 3):

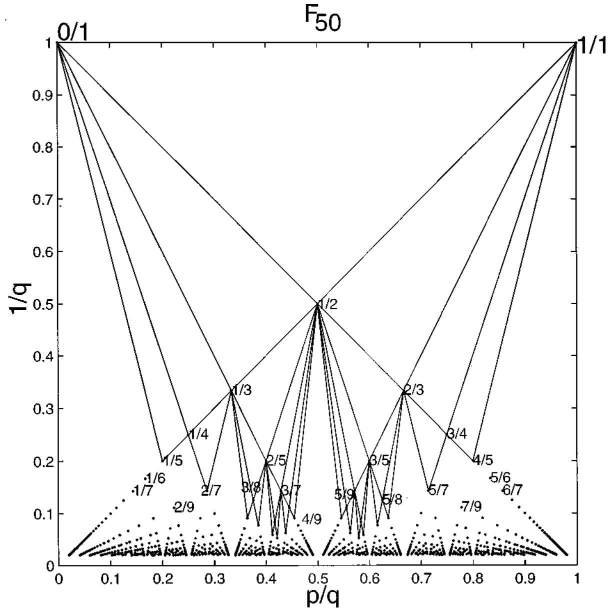

FIG. 3. Farey undirected graph (upper part of it) in the representation $(p/q,1/q)$: two vertices are connected by a segment iff one is the ascendant of the other. The triangles made by edges of the graph perform the so-called Farey-tessellation.

(i) The points representing $\mathcal{F}_n$ are all those having coordinates $y \geqslant 1/n$ in the plane of the projection. Low denominator rationals have high y coordinate.

(ii) The mediant construction has a simple geometrical interpretation. $p_1/q_1 \oplus p_2/q_2$ is represented by the intersection of two lines, one connecting the point $(p_1/q_1, 0/q_1)$ to the point $(p_2/q_2, 1/q_2)$, the second connecting the point $(p_2/q_2, 0/q_2)$ to the point $(p_1/q_1, 1/q_1)$.

(iii) There is a simple way to see if two rationals are Farey-consecutive: check if all points with x-coordinate between the x-coordinates of the two tested points have lower y-coordinate than any of the y-coordinates of the tested points.

(iv) The left and right ascendants of a rational p/q are the higher points, Farey-consecutive with p/q to its left and right, respectively. For instance, the left ascendant of 1/2, or of 1/3 is 0/1. Descendants of two Farey-consecutive rationals are all the lower points in the diagram with x-coordinate in between.

(v) If p_1/q_1 is the left ascendant of p/q, then $xl_n := p_n^l/q_n^l = (p_1+np)/(q_1+nq)$ converges to p/q in $\hat{\mathbb{Q}}$. This series is made of the left ascendant of p/q and of all rational numbers that have as right ascendant p/q, all on a straight line in the diagram. We call this the series of left low convergents of p/q in order to distinguish it from the finite series of convergents to a rational number coming from the continuous fraction expansion. Low convergents still obey the property of convergents of being best approximants (i.e., $p/q - p_n^l/q_n^l < C/(q_n^l)^2$ with lowest C), but they have denominators that can be greater than q (continuous fraction expansion convergents have all denominators smaller than q). For instance the series $0/1, 1/3, 2/5, 3/7, 4/9, \ldots$ converges to 1/2 from the left. In the same way, if p_2/q_2 is the right ascendant of p/q, the series of right low convergents $xr_n := (p_2+np)/(q_2+nq)$ converges to p/q from the right. As an example, $1/1, 2/3, 3/5, 4/7, 5/9, \ldots$ converges to 1/2 from the right. Any other series converging to p/q, and not containing p/q, are made of descendants of xl_n, xl_{n+1} or xr_n, xr_{n+1}, performing worse approximation of p/q than the low convergents.

The last remark inspired a special reasoning scheme that we shall use throughout the next sections. Let us consider $x = (p/q, r/q) \in \Delta_{h,k,l} \cap (\mathbb{P}(\mathbb{Z}^2) \times \mathbb{Z})$, thus $gcd[p,q] = 1, q > 0$. $(p_1/q_1, r_1/q_1)$ and $(p_2/q_2, r_2/q_2)$ are the unique right and left ascendants of x along the direction $\Delta_{h,k,l}$, respectively.

Let us endow $\mathbb{P}(\mathbb{Z}^3)$ with the topology induced by the gnomonic projection. i.e., the weakest topology that makes the gnomonic projection continuous from $\mathbb{P}(\mathbb{Z}^3)$ to $\mathbb{R}^2$ (with the Euclidean metric topology).

The series $xl_n^{[h,k,l]} := (p_n^l/q_n^l, r_n^l/q_n^l), p_n^l := p_1 + np, q_n^l := q_1 + nq, r_n^l := r_1 + nr$, $xr_n^{[h,k,l]} := (p_n^r/q_n^r, r_n^r/q_n^r), p_n^r := p_2 + np, q_n^r := q_2 + nq, r_n^r := r_2 + nr$, that we call series of left and right low convergents of x along $\Delta_{h,k,l}$, converge to x from the left and from the right along the direction $\Delta_{h,k,l}$, respectively. It is easy to check that all other series converging to x along $\Delta_{h,k,l}$ are made of descendants of $xl_n^{[h,k,l]}, xl_{n+1}^{[h,k,l]}$, or $xr_n^{[h,k,l]}, xr_{n+1}^{[h,k,l]}$. The following induction argument allows us to extend results valid for $xl_n^{[h,k,l]}$ and $xr_n^{[h,k,l]}$ to results valid for any series converging to x along a direction with Miller index $k=1$.

Lemma 14: Let $x_n := p_n/q_n \to p/q$ where x_n, x_{n+1} or x_{n+1}, x_n are Farey-consecutive rational numbers for any n. Define $\mathcal{I}_r := \mathcal{F}_r \cap [x_n, x_{n+1}]$. Let (P) be the property: $f(x) = y + \mathcal{O}(1/q_n)$, where f is an arbitrary real function and $y \in \mathbb{R}$. Suppose that

(i) *x_n, x_{n+1} obey (P).*
(ii) *If all $x \in \mathcal{I}_r$ obey (P), then all $x \in \mathcal{I}_{r+1}$ obey (P).*

Then all $x \in [x_n, x_{n+1}] \cap \mathbb{Q}$ obey (P).

Remark 6: We shall usually apply Lemma 14 by first checking (i), then checking the implication: if $x', x'' \in [x_n, x_{n+1}]$ *are Farey-consecutive and obey (P)*, then $x' \oplus x''$ obeys *(P)*. According to the hierarchical structure of Farey series this implies (ii).

IV. PROJECTIVE FUNCTIONS AND RECURSION ON FAREY SETS

Definition 15: We call a function $g: \mathbb{Z}^3 \to \mathbb{R}$ projective if it is constant on directions containing the origin $g(p,q,r) = g(np,nq,nr)$. A projective function induces an application $g: \mathbb{P}(\mathbb{Z}^3) \to \mathbb{R}$.

Let $\{F_{p,q}(z) | p,q \in \mathbb{Z}, gcd[p,q] = 1, 0 \leq p/q \leq 1\}$ be a family of complex functions defined as following:

$$F_{p_1+p_2,q_1+q_2}(z) = F_{p_1,q_1}(z) + z^{\beta p_1 + \alpha q_1} F_{p_2,q_2}(z), \tag{6}$$

whenever p_1/q_1, p_2/q_2 are Farey- consecutive in $\mathbb{Q}$. Equation (6) defines $F_{p,q}(z)$ for all rational numbers $0 \leq p/q \leq 1$, $gcd[p,q] = 1$, provided $F_{0,1}(z)$ and $F_{1,1}(z)$ are known.

Using this family we define the following projective function on $\mathbb{P}(\mathbb{Z}^3)$:

$$h(p,r,q) = \frac{|F_{p^*,q^*}(\exp[2\pi i r^*/(\beta p^* + \alpha q^*)]|}{sp^* + tq^*}, \tag{7}$$

where $p^* = p/gcd[p,q], q^* = q/gcd[p,q], r^* = r/gcd[p,q]$ ($r^* \in \mathbb{Q}$ is not necessarily an integer).

We have seen in the Introduction that for a Farey hierarchy of structures the structure factor at Bragg reflections and the widths of gaps in the spectra of elementary excitations are projective functions defined on $\mathbb{P}(\mathbb{Z}^2) \times \mathbb{Z}$ (see also Sec. V), and obeying relation (2) [relation (7) represents the natural extension of relation (2) to $\mathbb{P}(\mathbb{Z}^3)$].

The following theorem represents our main result, stating an important property of $h(p,r,q)$.

Theorem 16: *For any projective function h obeying Eqs. (6) and (7) and for each direction $\Delta_{h,k,l} = \{(p/q, r/q) | ph + rk + ql = 0\}$ with Miller index $k=1$, there is a projective function $\epsilon_{h,k,l}(x)$ and a projective function $f_{h,k,l}(x)$ such that.*

(i) $h(x) = (1 + \epsilon_{h,k,l}(x)) f_{h,k,l}(x), x \in \Delta_{h,k,l}$.

(ii) $f_{h,k,l}(x)$ is continuous on $\Delta_{h,k,l}$.

(iii) $\lim_{n\to\infty} \sup_{x \in [\mathbb{P}(\mathbb{Z}^2) \times \mathbb{Z} \setminus \mathcal{F}_n^{(1,1)}] \cap \Delta_{h,k,l} \cap [0,1]^2} |\epsilon_{h,k,l}(x)| = 0$.

(iv) *The limit in (iii) is more rapid for directions $\Delta_{h,k,l}$ for which $\mathcal{I} := |\beta l - \alpha h|$ is smaller. Precisely, if $|\beta l' - \alpha h'| > |\beta l - \alpha h|$, then*

$$0 \leq \lim_{n\to\infty} \frac{\sup_{x\in[\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}\setminus\mathcal{F}_n^{(1,1)}]\cap\Delta_{h,k,l}\cap[0,1]^2} \epsilon_{h,k,l}(x)}{\sup_{x\in[\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}\setminus\mathcal{F}_n^{(1,1)}]\cap\Delta_{h',k',l'}\cap[0,1]^2} \epsilon_{h',k',l'}(x)} < 1.$$

In particular, $\epsilon_{h,k,l}(x)=0$ *for directions with Miller indices satisfying* $\beta l-\alpha h=0$.

(v) *There is* $\epsilon_o>0$, *such that for any* $x=(p/q\,,r/q)\in\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$, *and for any two directions passing through x of Miller indices* (h,k,l) *and* (h',k',l'), *such that* $|\beta l-\alpha h|<|\beta l'-\alpha h'|$ *and* $\epsilon_{h,k,l}(x)<\epsilon_o$ *we have* $\epsilon_{h,k,l}(x)<\epsilon_{h',k',l'}(x)$, *i.e.,* $f_{h,k,l}(x)>f_{h',k',l'}(x)$.

The proof of Theorem 16 uses Lemma 14 and the following two lemmas:

Lemma 17: With the notations of Section III, if $xl_n^{[h,k,l]}$, $xr_n^{[h,k,l]}$ *are the series of left and right low convergents of x along* $\Delta_{h,k,l}$ *with* $k=1$, *then*

$$h(xl_n^{[h,k,l]})=h(x)/[1+\epsilon_{h,k,l}(x)]+\mathcal{O}\left(\frac{1}{q_n^l}\right),$$

$$h(xr_n^{[h,k,l]})=h(x)/[1+\epsilon_{h,k,l}(x)]+\mathcal{O}\left(\frac{1}{q_n^l}\right),$$

where

$$\epsilon_{h,k,l}(x)=\begin{cases}\left|\dfrac{\pi(\beta l-\alpha h)}{\beta p+\alpha q}\Big/\sin\left[\dfrac{\pi(\beta l-\alpha h)}{\beta p+\alpha q}\right]\right|-1, & if\ |\beta l-\alpha h|>0,\\ 0, & if\ \beta l-\alpha h=0,\end{cases}$$

and $x=(p/q\,,r/q)$ *is the unique representation of* $x\in\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ *with* $gcd[p,q]=1, q>0$.

Lemma 18: Consider the series of left and right low convergents of x along $\Delta_{h,k,l}$, $k=1$. *Consider the proposition (P1):* $h(z)=h(x)/[1+\epsilon_{h,k,l}(x)]+\mathcal{O}(1/q_n^l)$. *The following implication (and the analogous statement for left low convergents) is valid: if (P1) is satisfied when* $z=x',x''$ where x',x'' are any two Farey-consecutive points between $xl_n^{[h,k,l]},xl_{n+1}^{[h,k,l]}$ along $\Delta_{h,k,l}$, *then (P1) is also satisfied by* $z:=x'\oplus x''$.

V. EXAMPLES AND CONCLUSIONS

We have shown that the amplitudes of the structure factor at Bragg reflections and the widths of gaps (in first order of Rayleigh-Schrödinger perturbation theory) in spectra of elementary excitations of approximants forming a Farey hierarchy are given by projective functions recursively defined on $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$.

Theorem 16 states that these functions are approximated with arbitrary precision by smooth functions of the commensurability ratio p/q along special directions in $\mathbb{P}(\mathbb{Z}^3)$ [for which $k=1$, therefore common to $\mathbb{P}(\mathbb{Z}^3)$ and to $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$], when restricted to points that are ''not very rational'' i.e., they are outside $\mathcal{F}_n^{(1,1)}$, with n big enough.

Furthermore, Lemma 17 provides an analytical formula of a projective function performing the smoothing of the projective functions giving the values of gap widths (reflection amplitudes).

From point (iv) of Theorem 16 it follows that the lines with $k=1$ and smaller values of $\mathcal{I}=|\beta l-\alpha h|$ correspond to weaker fluctuations of the gap widths (amplitudes of reflections) with respect to the smooth approximating function. The smaller $\mathcal{I}$ is, the weaker the fluctuations are.

Point (v) of Theorem 16 implies that when several directions intersect, the values of the smooth approximating functions are higher for the lines with smallest value of $|\beta l-\alpha h|$ (if we ignore the tail oscillations of $\sin\Theta/\Theta$). Thus smoother lines are also thicker.

These results are very much related to what in dynamical system theory is called modular smoothing.[11] When applied to spectral properties of Farey hierarchies of structures, it reveals a simple arithmetic structure, that is, the consequence of the hierarchical construction rules.

Let us discuss here two examples:

The double-chain model.[4]

For a *(p,q)* commensurate double chain, and for small interchain coupling, the spectrum of phonon excitations has $[(p+q-1)/2]$ gaps at the positions

$$E(p,r,q)=f_1\left(\frac{p}{q},\frac{r}{q}\right),\quad r=1,2,\ldots,\left\lfloor\frac{p+q-1}{2}\right\rfloor,$$

$$f_1(x,y)=4\sin^2\left(\pi\frac{y}{x+1}\right)+\chi\frac{x}{x+1},$$

and the gap widths in first order of the interchain coupling constant χ are

$$\Delta(p,r,q)=\chi\phi_1\left(\frac{p}{q}\right)\frac{F_{p,q}(\exp[2\pi ir/(p+q)])}{p+q},$$

$$\phi_1(x)=x+1/x,$$

where $F_{p,q}$ obey Eq. (6) with $\alpha=1,\beta=1$.

The values of the projective function $\Delta(p,r,q)$ has physical meaning only for $gcd[p,q]=1$ [thus, on $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$] and for $r=1,2,\ldots,[(p+q-1)/2]$, where one has gaps.

It is customary to represent the spectra for different commensurability ratios as a plot versus p/q made of black vertical segments for bands and white vertical segments for gaps. This reveals self-similar ''Hofstadter-type'' diagrams. In Fig. 4 black and white are reversed.

Figure 4 shows first (a) the values of the function $F_{p,q}(\exp[2\pi ir/(p+q)])/(p+q)$ on $\mathcal{F}_{20}^{(1,1)}$, then (b) restricts them to the set $\{(p/q\,,\,r/q)|gcd[p,q]=1,p\leqslant q,r=1,2,\ldots,\lfloor(p+q-1)/2\rfloor\}$, and applies the continuous transformations necessary to get the real positions and widths of the gaps. The values of $\mathcal{I}=|\beta l-\alpha h|$ are represented for several directions $\Delta_{h,k,l}$ with $k=1$. As generally stated by Theorem 16 (v), when two lines of gaps intersect, the line with smaller value of n is thicker. Also, as stated at point (iv) of Theorem 16, lines with smaller values of n are smoother.

The Aubry repeated parabolas model.[2] This model corresponds to the Hamiltonian in Eq. (4) with $V_n^{p,q}=\chi\lfloor np/q+\frac{1}{2}\rfloor$.

Positions of gaps in the phonon spectrum are

$$E(p,r,q)=f_2\left(\frac{p}{q},\frac{r}{q}\right),\quad r=1,2,\ldots,q-1,$$

$$f_2(x,y)=4\sin^2(\pi y)+\chi x,$$

and the gap widths in first order of χ are

$$\Delta(p,r,q)=\omega_2^2\chi\frac{G_{p,q}(\exp[2\pi ir/q])}{q},$$

where $G_{p,q}$ obey Eq. (6) with $\alpha=1,\beta=0$.

As for the double chain model, the function $\Delta(p,r,q)$ has physical meaning only for $gcd[p,q]=1$ and for $r=1,2,\ldots,q-1$.

Figure 5 shows first (a) the values of the function $G_{p,q}(\exp[2\pi ir/q])/q$ on $\mathcal{F}_{20}^{(1,1)}$, then (b) restricts them to the set $\{(p/q\,,\,r/q)|gcd[p,q]=1,p\leqslant q,r=1,2,\ldots,q-1\}$, and applies the continuous transformations necessary to get the real positions and widths of the gaps.

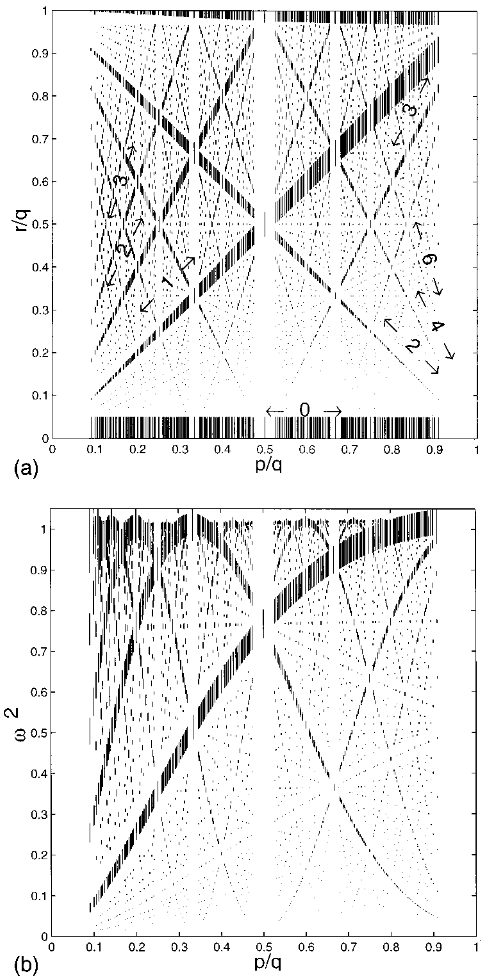

FIG. 4. Positions and widths of gaps for the DCM model: a) having $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ as support; b) real gap positions. The values of $\mathcal{I}=|\beta l-\alpha h|$ are indicated for different lines of gaps.

The properties stated at points *(iv)* and *(v)* of Theorem 16 hold in this case as well as shown by the values of $\mathcal{I}=|\beta l-\alpha h|$.

The transformation $(p/q\,,r/q)\rightarrow(1-p/q\,,r/q)$ transforms (h,k,l) to $(h,-k,-l)$ and $|\beta l-\alpha h|$ to $|\beta l+\alpha h|$. This explains the symmetrical aspect of Fig. 5 (when $\beta=0$ and $|\beta l-\alpha h|=|\beta l+\alpha h|$) and the absence of symmetry in Fig. 4 (when $\beta=1$ and $|\beta l-\alpha h|\neq|\beta l+\alpha h|$).

To summarize, here are the consequences of our results:

(i) For hierarchically defined 1D approximants, gaps in the phonon spectrum (or Bragg reflections in the diffraction pattern) are in one-to-one relation to subsets of $\mathbb{P}(\mathbb{Z}^3)$ (the set of directions in a 3D lattice). The gnomonic projection of $\mathbb{Z}^3$ onto the plane maps the directions of $\mathbb{Z}^3$ to gaps (or Bragg reflections) positions in the plot $(p/q\,,r/q)$, where r is the integer index giving the position of gaps in the spectrum at fixed commensurability ratio p/q when arranged in increasing order of energy.

(ii) The 2D planes of $\mathbb{Z}^3$ (containing the origin) with Miller indices (h,k,l) project onto lines connecting gaps (or Bragg reflections) that belong to spectra of different approximants and whose widths (amplitudes) are not very different and can be approximated by a smooth function of the commensurability ratio.

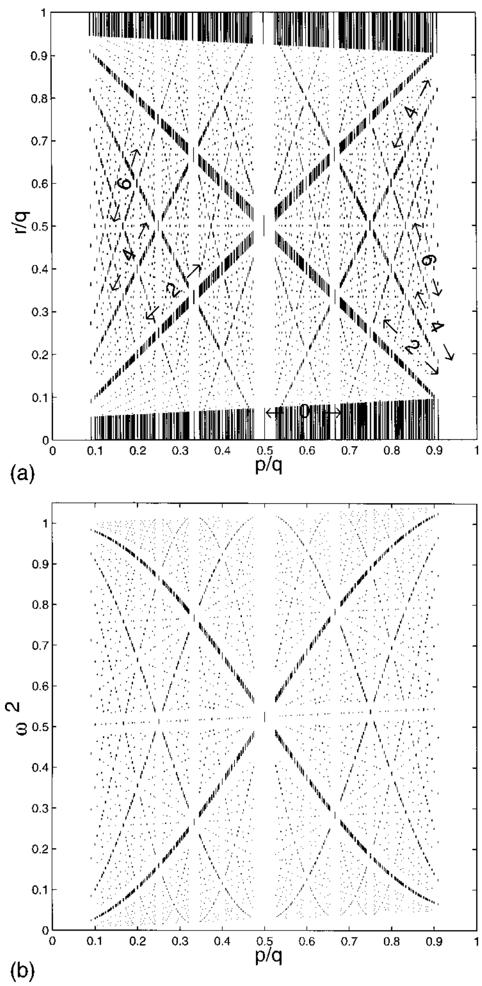

FIG. 5. Positions and widths of gaps for the Aubry model: (a) having $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ as support; (b) real gap positions. The values of $\mathcal{I}=|\beta l-\alpha h|$ are indicated for different lines of gaps.

(iii) The planes (lines in projection) with Miller indices $k=1$ and smallest values of $\mathcal{I}:=|\beta l-\alpha h|$ correspond to gaps (or Bragg reflections) whose widths (amplitudes) have weakest fluctuations with respect to the smooth approximating function.

(iv) Along a given line, the worst approximation is for the most ''rational'' points $(p/q\,,r/q)$ of $\mathbb{Z}^3$, i.e., the ones having low denominator q; the deviations relative to the smooth approximating function vanish for very irrational points $(q\rightarrow\infty)$.

(v) If we compare different lines, the smaller $\mathcal{I}$ is, the weaker the deviations with respect to the approximating smooth function are. Smoother (weakly oscillating gaps widths, or Bragg reflections amplitudes) lines are also thicker (large gaps, intense reflections).

(vi) The aspect of the ''Hofstadter-type'' diagrams is dictated by the values of α and β, that say which lines contain thicker and less fluctuating gaps (Bragg reflections), by the restriction on the possible values of r (only for the gap problem), and by the smooth function f that deforms the lines. For the diffraction problem the values of α and β are given by the way the Euclidean length of the unit cell scales with the commensurability indices p,q. For the gap problem the same values depend on the topology of the couplings between atoms and express the way the ''effective'' graph length of the unit cell scales with p,q.

The above results show the universal, simple arithmetical structure of spectra of Farey hierarchies of approximants and emphasize the importance of the geometry of numbers for spectral properties of hierarchical structures. Because the results were obtained in first order of the Rayleigh–Schrödinger perturbation theory and by using rather special recursions the range of their strict applicability is limited to 1D models of structures obtained by concatenation, in the limit of weak coupling. For strong coupling, nonlinear effects may destroy the universal structure of the spectra in the following way: lines of narrow gaps with big values of $\mathcal{I}=|\beta l-\alpha h|$ are strongly sheared and may lose their individuality each time they cross lines of wider gaps that open more rapidly with the value of the coupling. The lines of wider gaps are more robust and they preserve their individuality and structure for higher values of the coupling. Nonperturbative approaches are needed to solve the more complex details of the spectra in the case of strong coupling.

Our approach can be extended in several ways. One can imagine other types of Farey hierarchy than the one considered in this paper. For instance, approximant ground states of the Frenkel–Kontorova model can be constructed by concatenation, but continuous deformation of the unit cells should be applied at each step. In our definition of a Farey hierarchy we ignored the effect of this deformation. Also, the concatenation may be replaced by more general binary operations combining more than one copy of the unit cells. It would be also interesting to look for higher dimensional extensions of these results. The generalization of the diffraction problem to higher dimension is almost straightforward, but the spectral problem is not trivial. Already for one-dimensional models the possibility of obtaining simple recursions for the widths of the gaps relies on the topology of the couplings. It is possible to obtain these relations in all cases for diagonal tight-binding models for electrons (obtained by concatenation, therefore with a finite set of values of the onsite potential) or for the problem of phonons in the repeated parabolas Frenkel–Kontorova model, but for nondiagonal models like the double-chain model the dynamical matrix obeys simple recursions only in the special case of nearest-neighbor truncated coupling.[4]

The results presented here are also meant to be applied in material science. They provide a simple way to compare dynamical and spectral properties of approximants in homologous series. For instance, urea inclusion compounds having channels filled with different types of chainlike molecules were systematically studied for a long time.[12] It has been shown that, depending on the length of the included molecule, compounds with different host/guest ratio of periods can be produced. Because the inclusion channels are linear, 1D hierarchical models and the above analysis could be useful in the study of the variety of dynamical properties of these compounds. Another situation is represented by quasicrystalline alloys. Changing the stoichiometry or the temperature, or simply changing the nonequilibrium preparation method, produce samples that are quasiperiodic or contain a rich variety of periodic approximants having hierarchical structural relations between them. We would like to compare spectral properties of different approximants. Although quasicrystals are not 1D, 1D hierarchical models could be used to explain the very basic features of spectra.[13]

ACKNOWLEDGMENTS

This work was financed by the EEC Grant No. ERBCHBGCT940690. We are particularly grateful to Professor A. Janner for stimulating discussions. O.R. thanks T. Misirpashaev for help in the proof of Lemma 17.

APPENDIX A: FAREY SERIES AND FAREY-HIERARCHIES OF STRUCTURES

The Farey series of order n is the set $\mathcal{F}_n=\{p/q\,|\,p,q\in\mathbb{Z},gcd[p,q]=1,0\leqslant p\leqslant q\leqslant n\}$ of fractions in their lowest terms, arranged in an increasing order. Two rationals that are consecutive in $\mathcal{F}_n$, for some order n, are called Farey-consecutive. Equivalently, two rationals $0\leqslant p_1/q_1<p_2/q_2\leqslant 1$ are Farey-consecutive iff for all p/q between them $q>\max(q_1,q_2)$.

The construction of Farey series provides an automatic way to order all rational numbers into a hierarchy of increasingly dense discrete dissections of the continuum. A nice description of this hierarchy can be found in Ref. 14. The hierarchical nature of the Farey series is emphasized by the mediant construction. The mediant of two rationals p_1/q_1 and p_2/q_2 is $p_1/q_1 \oplus p_2/q_2 := (p_1 + p_2)/(q_1+q_2)$. In order to obtain the elements of $\mathcal{F}_{n+1}$ one should add to $\mathcal{F}_n$ all the mediants of Farey-consecutive rationals (we shall call these mediants descendants) $p_1/q_1, p_2/q_2 \in \mathcal{F}_n$ in $\mathcal{F}_n$ such that $q_1+q_2=n+1$. In such a way, all rational numbers in the interval [0,1] can be obtained from the two elements of $\mathcal{F}_1$, i.e., $\frac{0}{1}$ and $\frac{1}{1}$.

The structures of different approximants are related in a similar way. The longer period and more complex unit cells of approximants that are at the bottom of the hierarchy can be obtained by concatenation of unit cells of approximants situated at previous levels of the hierarchy.

Let us call the Farey-hierarchy of structures a family $\{\mathcal{S}\}_{0 \leqslant p/q \leqslant 1}$ of sets of atomic coordinates that have the following properties:

(i) $\mathcal{S}_{p/q}$ is obtained by the periodic repeat of the primitive unit cell $U_{p,q}$. $\mathcal{S}_{p/q} = \cup_{n\in\mathbb{Z}}(U_{p,q} + nL_{p,q})$, where $L_{p,q}$ is the period. (p,q) and p/q are called commensurability indices and commensurability ratio, respectively.

(ii) The unit cells are obtained by concatenation, i.e.,

$$U_{p_1+p_2,q_1+q_2} = U_{p_1,q_1} \cup (U_{p_2,q_2} + L_{p_1,q_1}), \tag{A1}$$

where $p_1/q_1, p_2/q_2$ are Farey-consecutive.

Like the Farey series, the entire family $\mathcal{S}_{p/q}$ can be obtained by consecutive application of Eq. (A1) once $\mathcal{S}_{0/1}$ and $\mathcal{S}_{1/1}$ are known.

It is easy to check that the unit cell $U_{p,q}$ contains $n_{p,q} := \# U_{p,q} = sp + tq$ discrete atomic positions, and the period is of the form $L_{p,q} = a(\beta p + \alpha q)$, where $s = n_{1,1} - n_{0,1}, t = n_{0,1}$, $a\beta = L_{1,1} - L_{0,1}, a\alpha = L_{0,1}$.

APPENDIX B: PROOFS

Proof of Theorem 4. Consider $\{a_1,a_2\}$ a basis of the plane $\Pi_{h,k,l}$ of Miller indices (h,k,l) of $\mathbb{Z}^3$, $\Pi_{h,k,l} := \{(p,r,q) | ph + rk + ql = 0\}$. $x=(p,r,q), gcd[p,r,q]=1$ is in $\Pi_{h,k,l}$, hence $x = x_1 a_1 + x_2 a_2$, with $gcd[x_1,x_2]=1$ (because otherwise $gcd[p,r,q]>1$). Then we can choose $y = y_1 a_1 + y_2 a_2$, with y_1, y_2 solution of the equation $x_1 y_2 - y_1 x_2 = 1$. $\{x,y\}$ is obtained from $\{a_1,a_2\}$ by a unimodular transformation, therefore it is another basis of $\Pi_{h,k,l}$. Because $gcd[h,k,l]=1$ there is a vector $z=(z_1,z_2,z_3) \in \mathbb{Z}^3$ such that $z_1 h + z_2 k + z_3 l = 1$. Let us show that $\{x,y,z\}$ is a basis of $\mathbb{Z}^3$. For any vector $v=(v_1,v_2,v_3) \in \mathbb{Z}^3$ we have $v_1 h + v_2 k + v_3 l = n_3 \in \mathbb{Z}$. Let $u=(u_1,u_2,u_3) = v - n_3 z$. It is easy to check that $u_1 h + u_2 k + u_3 l = 0$, therefore $u \in \Pi_{h,k,l}$ and thus v can be expressed as $v = n_1 x + n_2 y + n_3 z$. Because $\{x,y,z\}$ is a basis of $\mathbb{Z}^3$ the determinant made of the components of the three vectors is equal to one (or can be made equal to one by changing y into $-y$). As $x \wedge y = n(h,k,l)$, one can check that this means $n=1$. ■

Proof of Property 9: If $x_1, x_2 \in \mathcal{F}_n^{(2,0)}$ and any $x=(p/q, r/q)$ between x_1, x_2 obeys $x \notin \mathcal{F}_n^{(2,0)}$, then $q > n \geqslant \max(q_1,q_2)$. Conversely, let $n = \max(q_1,q_2)$, thus $x_1, x_2 \in \mathcal{F}_n^{(2,0)}$. Then for any $x=(p/q, r/q)$ between x_1, x_2 such that $q > \max(q_1,q_2) = n$ it follows that $x \notin \mathcal{F}_n^{(2,0)}$. ■

Proof of Property 10: At least one of the indices (h,k,l) is nonzero. Without restricting generality we may suppose $k>0$ (the proof can be adapted to strictly negative indices as well). Because $\lambda, k, q_1, q_2 > 0$, $x_1 \wedge x_2 = \lambda(h,k,l)$ implies $p_1/q_1 < p_2/q_2$. From Theorem 4 there is a point $x=(p/q, r/q)$ such that $x_1 \wedge x = (h,k,l)$, and such that $n - q_1 < q \leqslant n$, i.e.,

$$\frac{p}{q} - \frac{p_1}{q_1} = \frac{k}{q_1 q}.$$

We intend to show that $x_2 = x$. Suppose that $x_2 \neq x$. Then, as x_1, x_2 are consecutive in $\mathcal{F}_n^{(2,0)}$, and $x \in \mathcal{F}_n^{(2,0)}$, $p_2/q_2 < p/q$. By adding term by term the inequalities $p/q - p_2/q_2 = (pq_2 - p_{2q})/qq_2 \geqslant k/qq_2$, and $p_2/q_2 - p_1/q_1 = (p_2 q_1 - p_1 q_2)/q_1 q_2 \geqslant k/q_1 q_2$ we obtain $k/q_1 q = p/q - p_1/q_1 \geqslant k(1/qq_2 + 1/q_1 q_2) = k(q+q_1)/qq_1 q_2 > kn/q_1 q_2 q \geqslant k/q_1 q$, contradiction. The proof of the reciprocal property is like the proof of Theorem 1 of Neville.[7]

Using $x_1\wedge x_2=(h,k,l)$, we find $r_2/(\beta p_2+\alpha q_2)-r_1/(\beta p_1+\alpha q_1)=(\beta l-\alpha h)/[(\beta p_1+\alpha q_1)(\beta p_2+\alpha q_2)]$. ■

Proof of Property 12: The equivalence between Properties 10 and 12 is entirely analogous to the known proof[8,9] for the Farey series. ■

Proof of Property 13: Using Properties 12 and 10 it follows that $x=y_1\oplus y_2$ with y_1,y_2 Farey-consecutive is equivalent to y_1,x and x,y_2 and y_1,y_2 being Farey-consecutive and therefore to having the following equations simultaneously satisfied:

$$x\wedge y_2=(h,k,l), \tag{B1}$$

$$y_1\wedge x=(h,k,l), \tag{B2}$$

$$y_1\wedge y_2=(h,k,l). \tag{B3}$$

The general solution of Eq. (B1) is of the form $y_2=y_0+nx$ where y_0 is a particular solution (its existence follows from Theorem 4), that can be chosen to satisfy (i) $0<den(y_0)\leqslant den(x)$. Also n should satisfy (ii) $den(y_0)+nden(x)>0$, otherwise the external product will have to change sign on readjusting $den(y_0+nx)$ to be positive. Similarly, Eq. (B2) has the general solution $y_1=-y_0+mx$, where m satisfies (iii) $-den(y_0)+mden(x)>0$. Then Eq. (B3) is satisfied provided $m+n=1$, which implies $n\leqslant 1$. From (i) and (ii) it follows that $n\geqslant 0$, from (iii) it follows that $m>0$, therefore $n=0,m=1$ corresponds to the unique solution we are looking for. ■

Proof of Property 7: Suppose $k=1$. Consider any p_1/q_1, p/q Farey-consecutive. Then $pq_1-p_1q=1$, hence any rational number $p_2/q_2, gcd[p_2,q_2]=1$ can be expressed as $p_2/q_2=(mp_1+np)/(mq_1+nq)$. The line $\Delta_{h,k,l}$ passes through the points $(p_1/q_1, r_1/q_1),(p/q, r/q)$ and intersects the vertical line $x=p_2/q_2$ at $((mp_1+np)/(mq_1+nq),(mr_1+nr)/(mq_1+mq))$, being made entirely of points of $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$.

Conversely, if $\Delta_{h,k,l}=\Delta_{h,k,l}\cap[\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}]$ then necessarily $k\neq 0$, because otherwise the direction would be of the form $\Delta_{-q,0,p}=\{(op/oq, r/oq)|gcd[p,q]=1\}$ with fixed p,q and arbitrary r,o and it would contain points outside $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ (those with $o\neq 1$). Let us now consider any two Farey-consecutive rationals p_1/q_1, p/q. Because $k\neq 0$, $\Delta_{h,k,l}$ intersects the lines $x=p_1/q_1$ and $x=p/q$ at points that are in $\mathbb{P}(\mathbb{Z}^2)\times\mathbb{Z}$ by hypothesis, thus of the form $(p/q, r/q)$, $(p_1/q_1, r_1/q_1)$. From Remark 3 and Property 10 it follows that $k=1$. ■

Proof of Property 11: This follows from the fact that $(p_1/q_1, r_1/q_1)\wedge(p_2/q_2, r_2/q_2)=(h,k,l)$ with $gcd[h,k,l]=1$, and $k=1$, as soon as $p_2q_1-p_1q_2=1$, i.e., $p_1/q_1, p_2/q_2$ are Farey-consecutive. ■

Proof of Lemma 17: Because p_n^l/q_n^l and p/q are Farey-consecutive for any n, then $F_{p_n^l,q_n^l}(z)=F_{p_{n-1}^l,q_{n-1}^l}(z)+z^{\beta p_{n-1}^l+\alpha q_{n-1}^l}F_{p,q}(z)$. It follows

$$F_{p_n^l,q_n^l}=F_{p_1,q_1}+z^{\beta p_1+\alpha q_1}\frac{z^{n(\beta p+\alpha q)}-1}{z^{\beta p+\alpha q}-1}F_{p,q}. \tag{B4}$$

Let $\Phi_n^l:=2\pi r_n^l/(\beta p_n^l+\alpha q_n^l)$. Using (B4) it is easy to check that

$$h(xl_n^{[h,k,l]})=\frac{|\sin n(\beta p+\alpha q)\Phi_n^l/2|}{(\beta p+\alpha q)\Phi_n^l/2}\,\frac{|F_{p,q}(\exp(i\Phi_n^l))|}{sp_n^l+tq_n^l}+\mathcal{O}\left(\frac{1}{q_n^l}\right)$$

$$=\frac{h(x)}{1+\epsilon_{h,k,l}(x)}+\mathcal{O}\left(\frac{1}{q_n^l}\right). \tag{B5}$$

For the last equality we used $(\beta p+\alpha q)\Phi_n^l/2=\pi r+\pi(\beta l-\alpha h)/(\beta p_n^l+\alpha q_n^l)$, which is a consequence of Property 10.

The part of the proof concerning $xr_n^{[h,k,l]}$ is along the same lines. ■

Proof of Lemma 18: To simplify the proof, we assumed that $\alpha,\beta\geqslant 0$, but the other situations may be analyzed in the same way.

The function $F_{p,q}$ has the following form:

$$F_{p,q}(z)=\sum_{r=1}^{\mathcal{N}_{p,q}} a_r^{(p,q)} z^{\zeta_r^{(p,q)}}, \tag{B6}$$

where $\zeta_1^{(p,q)}<\zeta_2^{(p,q)}<\cdots<\zeta_{\max}^{(p,q)}$.

Let us show that for all p'/q', $p_n^l/q_n^l<p'/q'<p_{n+1}^l/q_{n+1}^l$ we have

$$\zeta_{\max}^{(p',q')}=(\beta p'+\alpha q')\left[1+\mathcal{O}\left(\frac{1}{q_n^l}\right)\right]. \tag{B7}$$

From Eq. (B4) it follows

$$\zeta_{\max}^{(p_n^l,q_n^l)}=\max[\zeta_{\max}^{(p_1,q_1)},\zeta_{\max}^{(p,q)}+\beta(p_1-p)+\alpha(q_1-q)+n(\beta p+\alpha q)]=n(\beta p+\alpha q)+\mathcal{O}(1), \tag{B8}$$

meaning that (B7) holds for p_n^l/q_n^l ($p_n^l=p_1+np$, $q_n^l=q_1nq$).

From Eq. (6) it follows, for any $p_n^l/q_n^l\leqslant p'/q'<p''/q''\leqslant p_{n+1}^l/q_{n+1}^l$,

$$\frac{\zeta_{\max}^{(p'+p'',q'+q'')}}{\beta(p'+p'')+\alpha(q'+q'')}=\max\left[\frac{\zeta_{\max}^{(p',q')}}{\beta(p'+p'')+\alpha(q'+q'')},\chi_1\frac{\zeta_{\max}^{(p'',q'')}}{\beta p''+\alpha q''}+\chi_2\right]=1+\mathcal{O}\left(\frac{1}{q_n^l}\right),$$

where $\chi_1=(\beta p''+\alpha q'')/[\beta(p'+p'')+\alpha(q'+q'')],\chi_2=(\beta p'+\alpha q')/[\beta(p'+p'')+\alpha(q'+q'')]$, $\chi_1+\chi_2=1$.

The last equality follows if we suppose that $\zeta_{\max}^{(p',q')}$ and $\zeta_{\max}^{(p'',q'')}$ obey (B7). Lemma 14 and Remark 6 end this part of the proof.

Let $\Phi_{p',q'}:=2\pi r'/(\beta p'+\alpha q'),\Phi_{p'',q''}:=2\pi r''/(\beta p''+\alpha q'')$. From Property 10 it follows

$$\Phi_{p',q'}-\Phi_{p'',q''}=\frac{2\pi(\beta l-\alpha h)}{(\beta p'+\alpha q')(\beta p''+\alpha q'')}. \tag{B9}$$

Using Equations (B7), (B6), and (B9) it is easy to check that

$$|F_{p'',q''}(\exp i\Phi_{p',q'})-F_{p'',q''}(\exp i\Phi_{p'',q''})|\leqslant\left(\sum_{r=1}^{\mathcal{N}_{p'',q''}}|a_r^{p'',q''}|\right)\frac{2\pi(\beta l-\alpha h)}{\beta p''+\alpha q''}\left[1+\mathcal{O}\left(\frac{1}{q_n^l}\right)\right]. \tag{B10}$$

Let us show that

$$\sum_{r=1}^{\mathcal{N}_{p',q'}}|a_r^{p',q'}|<C(\beta p'+\alpha q')\left[1+\mathcal{O}\left(\frac{1}{q_n^l}\right)\right] \tag{B11}$$

for any $p_n^l/q_n^l\leqslant p'/q'\leqslant p_{n+1}^l/q_{n+1}^l$.

From Eq. (B4) it follows that all $|a_r^{p_n^l,q_n^l}|$ are less than a constant A and that $\mathcal{N}_{p_n^l,q_n^l}\leqslant\mathcal{N}_{p_1,q_1}+n\mathcal{N}_{p,q}$. It follows

$$\frac{\sum_{r=1}^{\mathcal{N}_{p_n^l,q_n^l}}|a_r^{p_n^l,q_n^l}|}{\beta p_n^l+\alpha q_n^l}<A\frac{\mathcal{N}_{p_n^l,q_n^l}}{\beta p_n^l+\alpha q_n^l}\leqslant\frac{A\mathcal{N}_{p,q}}{\beta p+\alpha q}\left[1+\mathcal{O}\left(\frac{1}{q_n^l}\right)\right]. \tag{B12}$$

In a similar way we use Eq. (6) to show that if Eq. (B11) is fulfilled by both (p',q') and (p'',q'') with $p_n^l/q_n^l \leqslant p'/q' < p''/q'' \leqslant p_{n+1}^l/q_{n+1}^l$ then it is also fulfilled by $(p'+p'',q'+q'')$. Lemma 14 and Remark 6 conclude this part of the proof.

Using Eqs. (B10) and (B11) we show that

$$|F_{p'',q''}(\exp i\Phi_{p',q'}) - F_{p'',q''}(\exp i\Phi_{p'',q''})| \leqslant C'\left[1+\mathcal{O}\left(\frac{1}{q_n^l}\right)\right] \tag{B13}$$

and, in a similar way,

$$|F_{p'+p'',q'+q''}(\exp i\Phi_{p',q'}) - F_{p'+p'',q'+q''}(\exp i\Phi_{p'+p'',q'+q''})| \leqslant C''\left[1+\mathcal{O}\left(\frac{1}{q_n^l}\right)\right]. \tag{B14}$$

From Eqs. (6), (B13), (B14) it follows that

$$h\left(\frac{p'+p''}{q'+q''},\frac{r'+r''}{q'+q''}\right) = h\left(\frac{p'}{q'},\frac{r'}{q'}\right)\eta_1 + h\left(\frac{p''}{q''},\frac{r''}{q''}\right)\eta_2 + \mathcal{O}\left(\frac{1}{q_n^l}\right), \tag{B15}$$

where $\eta_1 = (sp'+tq')/[s(p'+p'')+t(q'+q'')]$, $\eta_2 = (sp''+tq'')/[s(p'+p'')+t(q'+q'')]$, $\eta_1 + \eta_2 = 1$.

Because x',x'' satisfy (P1) from Eq. (B15) it follows that $x' \oplus x''$ satisfies (P1).

The proof for $xr_n^{[h,k,l]}$ is along the same lines. ■

Proof of Theorem 16: Along a direction $\Delta_{h,k,l}$ with $k=1$, $h(x)$ $[x=(p/q\,,r/q),gcd[p,q]=1]$ is a function of p/q only and Lemma 14 applies in this case as well, proving together with Lemmas 17 and 18 that $h(x_n) \to h(x)/[1+\epsilon_{h,k,l}(x)]$ for any series $x_n := (p_n/q_n\,,r_n/q_n)$ converging to x along $\Delta_{h,k,l}$. In this case $\epsilon_{h,k,l}(x_n) \to 0$ because $p_n\,,q_n \to \infty$. Thus, $f(x_n) \to f(x)$, where $f(x) := h(x)/[1+\epsilon_{h,k,l}(x)]$, for any series converging to x along $\Delta_{h,k,l}$, f being a continuous function along $\Delta_{h,k,l}$.

It is easy to check that $\epsilon_{h,k,l}(x)$ satisfies the conditions iii–v with $\epsilon_o = 1/y_o - 1$, y_o being the highest local maximum different from 1 of the function $|\sin\Theta/\Theta|$. ■

[1] J. M. Luck, ''Cantor spectra scaling of gaps widths in deterministic aperiodic systems,'' Phys. Rev. B **39**, 5834–5849 (1989).
[2] T. Janssen, ''Aperiodic Schrödinger operators; electrons and phonons in aperiodic crystals,'' in *Mathematics of Long-range Aperiodic Order*, edited by R. V. Moody (Kluwer, Amsterdam, 1997).
[3] S. Aubry, ''Exact models with a complete Devil's staircase,'' J. Phys. C **16**, 2497–2508 (1982).
[4] O. Radulescu and T. Janssen, ''Dynamics of lattice vibrations of one-dimensional commensurate and incommensurate composites with harmonic interaction,'' J. Phys. A **30**, 4199–4214 (1997).
[5] S.-H. Kim and S. Ostlund, ''Simultaneous rational approximation in the study of dynamical systems,'' Phys. Rev. A **34**, 3426–3434 (1986).
[6] C. E. Springer, *Geometry and Analysis of Projective Spaces* (Freeman, San Francisco, 1964).
[7] E. Neville ''The structure of Farey series,'' Proc. London Math. Soc. **51**, 132–144 (1949).
[8] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers* (Clarendon, Oxford, 1979).
[9] L. K. Hua, *Introduction to Number Theory* (Springer-Verlag, Berlin, 1982).
[10] O. Radulescu and M. Kléman, ''Periodicity and quasiperiodicity of the structures with local icosahedral order,'' C. R. Acad. Sci., Ser. II; Mec., Phys., Chim., Sci. Terre Univers. **315**, 1629–1634 (1992).
[11] N. Burić and I. C. Percival, ''Modular smoothing and KAM tori,'' Physica D **71**, 39–54 (1994).
[12] M. D. Hollingsworth and K. D. M. Harris, *Comprehensive Supramolecular Chemistry*, edited by D. D. MacNicol, F. Toda, and R. Bishop (Pergamon, Oxford, 1996), Vol. 6, Chap. 7, pp. 177–237.
[13] M. Quilichini and T. Janssen, ''Phonon excitations in quasicrystals,'' Rev. Mod. Phys. **69**, 277–314 (1997).
[14] J. C. Lagarias and C. P. Tresser, ''A walk along the branches of the extended Farey tree,'' IBM J. Res. Dev. **39**, 283–294 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

On a trace formula of the Buslaev–Faddeev type for a long-range potential

Alexei Rybkin[a)]
Department of Mathematics, The University of Alaska, Fairbanks, P.O. Box 6660, Fairbanks, Alaska 99775

(Received 20 February 1997; accepted for publication 14 December 1998)

We propose an approach to obtaining new trace formulas of the Gel'fand–Levitan–Buslaev–Faddeev type, valid for Hilbert–Schmidt perturbations. In this way we obtain a new trace formula for Schrödinger operators on the half-line with long-range potentials.
[S0022-2488(99)02703-6]

I. INTRODUCTION

The formulas we are concerned with first appeared in 1953 in the classic paper[1] by Gel'fand and Levitan, where the authors obtained some identities for the eigenvalues of a regular Sturm–Liouville operator. A detailed account of trace formulas on finite intervals can be found in Dikii's paper.[2] These nice formulas relating the spectrum of an operator to certain characteristics of the operator itself called trace formulas became a pattern to follow; and in 1957, Faddeev considered a generalization for the case of a singular Sturm–Liouville operator, which obtained a full exposition by Buslaev and Faddeev[3] in 1960. Namely, they studied trace formulas for a self-adjoint Schrödinger operator H in $L_2(0,\infty)$:

$$H=H_0+v(x),\quad H_0=-\frac{d^2}{dx^2},\quad u(0)=0, \tag{1.1}$$

where the real-valued potential $v(x)$ is assumed to be continuous and short range [i.e., having the first moment on $(0,\infty)$]. In particular, if, in addition, $v'(x)$ is also continuous and $v'(x)$ has a finite limit as $x\to\infty$, the following formula is valid:

$$\sum_{n\geqslant 0}\lambda_n+\int_0^\infty\left(\zeta(t)-\frac{1}{2\pi\sqrt{t}}\int_0^\infty v(x)dx\right)dt=-\frac{1}{4}v(0), \tag{1.2}$$

where $\{\lambda_n\}$ are eigenvalues of H and $\zeta(t)$ is defined by the limiting phase θ by the formula $\zeta(t)=\pi^{-1}\theta(\sqrt{t})$, $t\geqslant 0$. Formula (1.2) is, in fact, the first formula in the chain of relations that can be interpreted as expressions of regularized spectral traces of integer powers of H in terms of the potential $v(x)$, and its proof is based on Krein's trace formula that for a pair of resolvent comparable abstract operators H,H_0 reads (see, e.g., Ref. 4) as

$$\mathrm{tr}\{(H-z)^{-1}-(H_0-z)^{-1}\}=\int_{-\infty}^\infty\frac{d}{dt}(t-z)^{-1}\zeta(t)dt,\quad \mathrm{Im}\,z\neq 0. \tag{1.3}$$

The real-valued function $\zeta(t)$ is called the Krein spectral shift function of the pair H,H_0. It is summable on $(-\infty,\infty)$ with the weight $(1+t^2)^{-1}$, and for a pair (1.1) with a short-range potential $v(x)$ the following relation holds:[2]

a)Electronic mail: ffavr@uaf.edu

$$\zeta(t)=-N(t),\quad t<0;\quad \zeta(t)=\pi^{-1}\theta(\sqrt{t}),\quad t>0, \tag{1.4}$$

where $N(t)$ is the number of eigenvalues of H lying to the left from t. Note that the Gel'fand and Levitan original proof of their trace relations does not make a use of (1.3), and the connection between the Gel'fand–Levitan and Krein trace formulas was rigorously justified only quite recently in Ref. 5.

The Krein trace formula (1.3) gave a rise to a large number of papers dealing with key words: trace formulas, Krein's spectral shift function, the scattering matrix, etc.; too many to be mentioned. We refer the interested reader to the survey[4] and recent papers,[6–10] and the literature therein. Note especially only papers[6–8] by Gesztesy and Simon with coauthors, where the authors obtain new trace formulas and systematically apply them to solving the inverse problem for Schrödinger operators.

It is quite natural to ask what kind of trace formulas are available when we go over from short-range potentials to long-range ones. For such pairs H,H_0 the spectral shift function, in general, does not exist and formula (1.3) is no longer valid. However, there are quite a few substantial trace formulas[11–14] serving long-range potentials, but none of them are quite similar to (1.2). Our goal will just be finding a direct extension of (1.3) to the case of a long-range potential. We obtain such a formula basing upon a generalization of (1.3) due to L. S. Koplienko,[15] valid for self-adjoint operators H,H_0 $(H=H_0+V)$, subject to the condition $V|H_0-i|^{-1/2}$ is a Hilbert–Schmidt operator. Namely, for such a pair H,H_0, there is a unique (up to an additive constant) real-valued function $\eta(t)$ summable on $(-\infty,\infty)$ with the weight $(1+t^2)^{-\delta}$, $\delta>\frac{1}{2}$, such that $(\operatorname{Im} z\neq 0)$

$$\operatorname{tr}\{(H-z)^{-1}-(H_0-z)^{-1}+V(H_0-z)^{-2}\}=-\int_{-\infty}^{\infty}\frac{d^2}{dt^2}(t-z)^{-1}\eta(t)dt. \tag{1.5}$$

We call the function $\eta(t)$ the regularized spectral shift function. If a pair H,H_0 is defined by (1.1) with $v(x)$ twice differentiable, satisfying the condition

$$\left|\frac{d^n v(x)}{dx^n}\right|\leqslant Q(1+x)^{-\alpha-n};\quad \alpha>1/2,\quad n=0,1,2, \tag{1.6}$$

then the function $\eta(t)$ is differentiable and

$$\eta'(t)=-N(t),\quad t<0;\quad \eta'(t)=\pi^{-1}\theta(\sqrt{t}),\quad t>0, \tag{1.7}$$

where $N(t)$ is as above and θ is the modified limiting phase defined from the asymptotics $(\operatorname{Im} k=0)$:

$$\psi(x,k)=A(k)\sin\left(kx-\int_0^x\frac{\sin^2 ks}{k}v(s)ds-\theta(k)\right)+o(1),\quad x\to\infty, \tag{1.8}$$

of the solution to the following Cauchy problem:

$$H\psi=k^2\psi,\quad \psi(0,k)=0,\quad \psi'(0,k)=1.$$

II. SOME ASYMPTOTIC FORMULAS

In the sequel we will always assume that H,H_0 are defined by (1.1), and for the time being suppose that the potential $v(x)$ is four times continuously differentiable and finitely supported.

Proposition 2.1: Let R_z,R_z^0 be the resolvents of H,H_0 respectively, then

$$-\frac{1}{2\pi i}\oint_{|z|=R} z^2\,\mathrm{tr}\{R_z-R_z^0+v(R_z^0)^2\}dz=\frac{3}{8}\sqrt{R}\cdot\int_0^\infty v^2(x)dx-\frac{v^2(0)}{4}+o\left(\frac{1}{\sqrt{R}}\right),\quad R\to\infty. \tag{2.1}$$

Proof: Due to our condition on the potential, $R_z-R_z^0$ is of the trace class, and we have

$$\begin{aligned}\mathrm{tr}\{R_z-R_z^0+v(R_z^0)^2\}&=\mathrm{tr}\{R_z-R_z^0\}+\mathrm{tr}\{v(R_z^0)^2\}\\ &=-\frac{d}{dz}\log M(\sqrt{z})+\frac{d}{dz}\mathrm{tr}\{v(R_z^0)\}\\ &=\frac{d}{dz}\{\mathrm{tr}\{v(R_z^0)\}-\log M(\sqrt{z})\},\quad \mathrm{Im}\sqrt{z}\geqslant 0,\end{aligned} \tag{2.2}$$

where $M(\sqrt{z})=\det(I+vR_z^0)$ is the perturbation determinant of the pair H,H_0. Fix the branch of $\log M(\sqrt{z})$ so that $\log M(\sqrt{z})\to 1$, $|z|\to\infty$; then,[3] uniformly in z, $\mathrm{Im}\sqrt{z}\geqslant 0$, we have

$$\log M(\sqrt{z})=-\sum_{k=1}^{4}\frac{q_k}{(2i\sqrt{z})^k}+o\left(\frac{1}{|z|^2}\right),\quad |z|\to\infty, \tag{2.3}$$

where

$$q_1=\int_0^\infty v(x)dx,\quad q_2=v(0),\quad q_3=-v'(0)-\int_0^\infty v^2(x)dx,\quad q_4=v''(0)-2v^2(0).$$

Since R_z^0 is an integral operator with the kernel

$$G(x,y;z)=-\frac{1}{i\sqrt{z}}(e^{i\sqrt{z}|x-y|}-e^{i\sqrt{z}(x+y)}), \tag{2.4}$$

one can easily see that

$$\mathrm{tr}\{v(R_z^0)\}=-\frac{1}{2i\sqrt{z}}\int_0^\infty v(x)(1-e^{2i\sqrt{z}x})dx,\quad \mathrm{Im}\sqrt{z}\geqslant 0,$$

and integration by parts three times gives

$$\mathrm{tr}\{v(R_z^0)\}=-\sum_{k=1}^{4}\frac{p_k}{(2i\sqrt{z})^k}+o\left(\frac{1}{|z|^2}\right),\quad |z|\to\infty, \tag{2.5}$$

where $p_1=q_1$, $p_k=(-1)^k v^{(k-2)}(0)$, $k=2,3,4$. Asymptotics (2.5) is obviously uniform in z, $\mathrm{Im}\sqrt{z}\geqslant 0$, and, hence, plugging (2.3) and (2.5) into (2.2), we obtain

$$z^2\,\mathrm{tr}\{R_z-R_z^0+v(R_z^0)^2\}=\frac{1}{8}\sum_{k=0}^{2}\frac{(k+2)(q_{k+2}-p_{k+2})}{(2i\sqrt{z})^k}+o\left(\frac{1}{|z|^3}\right),\quad |z|\to\infty.$$

This asymptotics is uniform in z, $\mathrm{Im}\sqrt{z}\geqslant 0$, and a straightforward computation leads to (2.1).

Proposition 2.2: For the regularized spectral shift function we have

$$\eta(R)=-\frac{1}{4\pi}\int_0^\infty v^2(x)dx\cdot\frac{1}{\sqrt{R}}+O\left(\frac{1}{\sqrt{R^3}}\right),\quad R\to\infty. \tag{2.6}$$

Proof: As it follows from (1.3), (1.5),

$$\eta(R)=-\operatorname{tr}\{vE_0(R)\}+\int_{-\infty}^{R}\zeta(t)dt, \tag{2.7}$$

where E_0 is the resolution of identity for H_0 and $\zeta(t)$ is the spectral shift function of H,H_0. We calculate the asymptotics for each term of the right side of (2.7). For the kernel of E_0 we have

$$E_0(x,y;R)=\frac{1}{\pi}\left(\frac{\sin\sqrt{R}(x-y)}{(x-y)}-\frac{\sin\sqrt{R}(x+y)}{(x+y)}\right).$$

Hence,

$$\begin{aligned}\operatorname{tr}\{vE_0(R)\}&=\frac{1}{\pi}\int_0^\infty v(x)\left(\sqrt{R}-\frac{\sin 2\sqrt{R}x}{2x}\right)dx\\&=\frac{\sqrt{R}}{\pi}\int_0^\infty v(x)dx-\frac{1}{\pi}\int_0^\infty v(x)\frac{\sin 2\sqrt{R}x}{2x}dx\\&=\frac{\sqrt{R}}{\pi}\int_0^\infty v(x)dx-\frac{1}{2\pi}\int_0^\infty\frac{v(x)-v(0)}{x}\sin 2\sqrt{R}x\,dx-\frac{v(0)}{2\pi}\int_0^\infty\frac{\sin 2\sqrt{R}x}{x}dx.\end{aligned} \tag{2.8}$$

For the second integral we have

$$\int_0^\infty\frac{v(x)-v(0)}{x}\sin 2\sqrt{R}x\,dx=\frac{v'(0)}{2}\frac{1}{\sqrt{R}}+O\left(\frac{1}{\sqrt{R^3}}\right),\quad R\to\infty. \tag{2.9}$$

Taking into account that

$$\int_0^\infty\frac{\sin 2\sqrt{R}x}{x}dx=\frac{\pi}{2},$$

for (2.8) we get

$$\operatorname{tr}\{vE_0(R)\}=\frac{\sqrt{R}}{\pi}\int_0^\infty v(x)dx-\frac{v(0)}{4}-\frac{v'(0)}{4\pi}\frac{1}{\sqrt{R}}+O\left(\frac{1}{\sqrt{R^3}}\right),\quad R\to\infty.$$

Represent the second term of (2.7) as follows:

$$\int_{-\infty}^{R}\zeta(t)dt=\int_{-\infty}^{0}\zeta(t)dt+\int_0^R\zeta(t)dt. \tag{2.10}$$

Using the asymptotic expansion for $\zeta(t)$,[3]

$$\zeta(t)=\sum_{k=0}^{1}\frac{(-1)^k q_{2k+1}}{(2\sqrt{t})^k}+O\left(\frac{1}{\sqrt{t^3}}\right),\quad t\to\infty, \tag{2.11}$$

where q's are as in Proposition 2.1, and since by (1.4) $\zeta(t)=-N(t)$, $t<0$, (2.10) can be continued,

$$=\sum_n \lambda_n+\int_0^R\left(\zeta(t)-\frac{q_1}{2\pi\sqrt{t}}\right)dt+\frac{q_1}{\pi}\int_0^R\frac{dt}{2\sqrt{t}}$$

$$=\sum_n \lambda_n+\frac{q_1}{\pi}\sqrt{R}+\int_0^\infty\left(\zeta(t)-\frac{q_1}{2\pi\sqrt{t}}\right)dt-\int_R^\infty\left(\zeta(t)-\frac{q_1}{2\pi\sqrt{t}}\right)dt. \tag{2.12}$$

From the Buslaev–Faddeev trace formula (1.2),

$$\int_0^\infty\left(\zeta(t)-\frac{q_1}{2\pi\sqrt{t}}\right)dt=-\sum_n \lambda_n-\frac{v(0)}{4},$$

and (2.12) can be continued,

$$=\frac{q_1}{\pi}\sqrt{R}-\frac{v(0)}{4}-\int_R^\infty\left(\zeta(t)-\frac{q_1}{2\pi\sqrt{t}}\right)dt. \tag{2.13}$$

Estimate the integral in (2.13). Taking into account (2.11), we have

$$\int_R^\infty\left(\zeta(t)-\frac{q_1}{2\pi\sqrt{t}}\right)dt=\int_R^\infty\left(\zeta(t)-\frac{q_1}{2\pi\sqrt{t}}+\frac{q_3}{8\pi t\sqrt{t}}\right)dt-\frac{q_3}{8\pi}\int_R^\infty\frac{dt}{t\sqrt{t}}$$

$$=-\frac{q_3}{4\pi\sqrt{R}}+O\left(\frac{1}{\sqrt{R^3}}\right). \tag{2.14}$$

Therefore

$$\int_{-\infty}^R\zeta(t)dt=\frac{q_1}{\pi}\sqrt{R}-\frac{v(0)}{4}-\frac{q_3}{4\pi\sqrt{R}}+O\left(\frac{1}{\sqrt{R^3}}\right)$$

$$=\frac{\sqrt{R}}{\pi}\int_0^\infty v(x)dx-\frac{v(0)}{4}-\frac{v'(0)+\int_0^\infty v^2(x)dx}{4\pi}\frac{1}{\sqrt{R}}+O\left(\frac{1}{\sqrt{R^3}}\right),\quad R\to\infty,$$

and, combining (2.9) and (2.14), we arrive at (2.6).

III. THE TRACE FORMULA

In this section we state and prove the main result.

Theorem 3.1: *Let a potential $v(x)$ be long range, such that*

$$\left|\frac{d^n v(x)}{dx^n}\right|\leqslant Q(1+x)^{-\alpha-n};\quad \alpha>1/2,\quad n=0,1,2,3,4. \tag{3.1}$$

Then the following trace formula holds:

$$-\sum_n \lambda_n^2+2\int_0^\infty\left(\eta(t)+\frac{1}{4\pi\sqrt{t}}\cdot\int_0^\infty v^2(x)dx\right)dt=\frac{v^2(0)}{4}. \tag{3.2}$$

Proof: We prove (3.2) first for potentials as in the previous section. Multiply (1.5) by $-(1/2\pi i)z^2$ and integrate the result along the contour $C_R^\epsilon=\{z:|z|=R,\epsilon\leqslant\arg z\leqslant 2\pi-\epsilon\}$ with a sufficiently large R:

$$-\frac{1}{2\pi i}\int_{C_R^\epsilon} z^2\,\mathrm{tr}\{R_z - R_z^0 + v(R_z^0)^2\}dz = \frac{1}{2\pi i}\int_{C_R^\epsilon} dz\cdot z^2\int_{-\infty}^{\infty}\frac{d^2}{dt^2}(t-z)^{-1}\eta(t)dt. \qquad (3.3)$$

Let $\epsilon\to 0$ and apply Proposition 2.1 and Lemma 4.1 to (3.2). Taking into account (1.7), we get

$$\frac{3}{8}\sqrt{R}\int_0^\infty v^2(x)dx - \frac{v^2(0)}{4} + O\left(\frac{1}{\sqrt{R}}\right) = \sum_n \lambda_n^2 - 2\int_0^R \eta(t)dt + 2R\eta(R) - R^2\eta'(R). \qquad (3.4)$$

Apply now Proposition 2.2 to the last two terms of (3.4). Performing trivial simplifications and letting $R\to\infty$, one has

$$-\sum_n \lambda_n^2 + 2\int_0^\infty\left(\eta(t) + \frac{1}{4\pi\sqrt{t}}\cdot\int_0^\infty v^2(x)dx\right)dt = \frac{v^2(0)}{4}. \qquad (3.5)$$

Now we need to drop the requirement of boundedness of $\mathrm{supp}\, v(x)$. Following Koplienko's proof of (1.5), let $v(x)$ be subject to the condition of the theorem; we construct $v_n(x)$ such that $v_n(x)=v(x)$, $0\leqslant x\leqslant n$; $v_n(x)=0$, $x\geqslant n$, and for $n\leqslant x\leqslant n+1$ we set $v_n(x)$ subject to

$$\left|\frac{d^m v_n(x)}{dx^m}\right| \leqslant Qn^{-\alpha}; \quad m=0,1,2,3,4.$$

For such $v_n(x)$ formula (3.5) clearly holds. It can be derived from Ref. 15 that $\eta_n(t)\to\eta(t)$ in $L_1((1+x^2)^{-\alpha})$ and hence in $L_1(0,a)$ for any $a<\infty$, say $a=2\max|v(x)|$. But for $t\geqslant 2\max|v(x)|$, it follows from Lemma 4.2, Proposition 2.2, and the relation $\eta'(t)=-\pi^{-1}\theta(\sqrt{t})$, $t>0$, that

$$\left|\eta_n(t) - (4\pi)^{-1}t^{-1/2}\int_0^\infty v_n^2(x)dx\right| \leqslant C\cdot t^{-3/2},$$

with some C dependent only on Q, α in (3.1). The Lebesgue dominated convergence theorem lets us pass to the limit in (3.5) as $n\to\infty$ and the theorem is proven.

Remark 3.2: Formula (3.1) can be included into a chain of trace formulas of higher order. To obtain these formulas we need to know more terms in asymptotics (2.6). We chose to present a recipe of treating long-range potentials rather than completeness of the results.

Remark 3.3: We note that Theorem 3.1 can be proved under condition (1.6) with an additional condition on $v''(x)$ of a Lipschitz's type. This additional condition is essential and it is possible to construct a potential $v(x)$ subject to (1.6) such that the integral in (3.2) is absolutely divergent. We plan to discuss this matter in detail elsewhere.

Remark 3.4: The way we obtained trace formulas of the Buslaev–Faddeev type for long-range potentials can be easily adjusted to some other settings such as a Schrodinger operator on the whole line or the three-dimensional case. We hope to return to this point elsewhere.

APPENDIX: AUXILIARY STATEMENTS

In proving Theorem 3.1 we made a use of the following two lemmas.

Lemma 4.1: Let $\eta(t)$ be a real-valued function belonging to $L_1((1+t^2)^{-\delta}dt;(0,\infty))$, $\delta>\frac{1}{2}$, and let $C_R^\epsilon=\{z:|z|=R, \epsilon\leqslant\arg z\leqslant 2\pi-\epsilon\}$. Then, for almost all $R>0$,

$$-\frac{1}{2\pi i}\lim_{\epsilon\to 0}\int_{C_R^\epsilon} dz\cdot z^2\int_0^\infty \frac{d^2}{dt^2}(t-z)^{-1}\eta(t)dt = 2\int_0^R \eta(t)dt - 2R\eta(R) + R^2\eta'(R). \qquad (A1)$$

Proof: Let $a>R$ and consider the following expression:

$$\int_{C_R^\epsilon} dz\cdot z^2\int_a^\infty (t-z)^{-3}\cdot \eta(t)dt=\int_a^\infty dt\cdot \eta(t)\int_{C_R^\epsilon} dz\cdot z^2(t-z)^{-3}. \tag{A2}$$

Since

$$\left|\int_{C_R^\epsilon} dz\cdot z^2(t-z)^{-3}\right|\leqslant R^2\int_{C_R^\epsilon}|dz|\cdot|t-z|^{-3}\leqslant\frac{2\pi R^3}{(a-R)^3},$$

by the Lebesgue theorem, we can pass to the limit in (A2) as $\epsilon\to 0$:

$$\lim_{\epsilon\to 0}\int_{C_R^\epsilon} dz\cdot z^2\int_a^\infty (t-z)^{-3}\cdot \eta(t)dt=\int_a^\infty dt\cdot \eta(t)\int_{C_R^\epsilon} dz\cdot z^2(t-z)^{-3}=0,\quad t\in \operatorname{Ext} C_R .$$

Denote

$$F(z)=\int_0^a (t-z)^{-1}\cdot \eta(t)dt.$$

The function $F(z)$ is clearly finite, and we have

$$\int_{C_R^\epsilon} dz\cdot z^2\int_0^a \frac{d^2}{dt^2}(t-z)^{-1}\cdot \eta(t)dt=\int_{C_R^\epsilon} dz\cdot z^2F''(z)=[z^2F'(z)-2zF(z)]_{C_R^\epsilon}+2\int_{C_R^\epsilon}F(z)dz. \tag{A3}$$

Representing $F'(z)$ in the form

$$F'(z)=-\frac{\eta(a)}{a-z}-\frac{\eta(0)}{z}+\int_0^a (t-z)^{-1}\cdot \eta'(t)dt,$$

we have

$$\begin{aligned}\lim_{\epsilon\to 0}[z^2F'(z)-2zF(z)]_{C_R^\epsilon}&=R^2\left\{\int_0^a (t-z+i0)^{-1}\cdot \eta'(t)dt-\int_0^a (t-z-i0)^{-1}\cdot \eta'(t)dt\right\}\\&\quad-2R\left\{\int_0^a (t-z+i0)^{-1}\cdot \eta(t)dt-\int_0^a (t-z-i0)^{-1}\cdot \eta(t)dt\right\}\\&=-2\pi i(R^2\eta'(R)-2R\eta(R))\end{aligned} \tag{A4}$$

For the second term of the right side of (A3), we get

$$\begin{aligned}2\int_{C_R^\epsilon}F(z)dz&=2\int_{C_R^\epsilon}dz\int_0^a (t-z)^{-1}\cdot \eta(t)dt\\&=2\int_0^a dt\,\eta(t)\int_{C_R^\epsilon}(t-z)^{-1}\,dz\\&=-2\int_0^a dt\ \eta(t)[\log(\mathrm{Re}^{i\epsilon}-t)]_{C_R^\epsilon}.\end{aligned}$$

Since $|[\log(\mathrm{Re}^{i\epsilon}-t)]_{C_R^\epsilon}|=2|\arctan R\sin\epsilon/(R\cos\epsilon-1)|\leqslant\pi$ and $\lim_{\epsilon\to 0}[\log(\mathrm{Re}^{i\epsilon}-t)]_{C_R^\epsilon}=2\pi i$, if $t<R$, and 0, if $t>R$, we obtain

$$2\lim_{\epsilon\to 0}\int_{C_R^\epsilon}F(z)dz=-2\int_0^a dt\ \eta(t)\lim_{\epsilon\to 0}[\log(\mathrm{Re}^{i\epsilon}-t)]_{C_R^\epsilon}=-4\pi i\int_0^R\eta(t)dt. \tag{A5}$$

Let $\epsilon\to 0$. Plugging (A4), (A5) into (A3), we arrive at the conclusion of the lemma.

Lemma 4.2: Let a potential $v(x)$ satisfy

$$\left|\frac{d^n v(x)}{dx^n}\right|\leqslant Q(1+x)^{-\alpha-n};\quad \alpha>1/2,\quad n=0,1,2,3,4. \tag{A6}$$

Then, for the limiting phase $\theta(k)$ [defined by (1.8)], the following estimate holds:

$$\left|\theta(k)-\frac{1}{8k^3}\int v^2(x)dx\right|\leqslant Ck^{-5},\quad k\geqslant\sqrt{2\max|v(x)|}, \tag{A7}$$

where C depends only on Q and α.

Proof: Throughout the proof we agree to denote by $O(k^{-n})$ any expression whose absolute value does not exceed Ck^{-n} with some constant C depending only on Q and α in condition (4.6). It is well known,[15,16] that the equation $-y''+v(x)y=k^2y$, $x\geqslant 0$, has a solution subject to

$$\lim e^{i\tau(x,k)}f(x,k)=1,\quad x\to\infty,\quad \tau(x,k)=kx-\int_0^x\frac{\sin^2 kt}{k}v(t)dt.$$

For $k\geqslant\sqrt{2\max|v(x)|}=k_0$, $f(x,k)$ admits the representation[17]

$$f(x,k)=\frac{\sqrt{k}}{\sqrt[4]{k^2-v(x)}}e^{-\alpha(k)}g(x,k), \tag{A8}$$

where $\alpha(k)=\lim_{x\to\infty}(u(x,k)-\tau(x,k))$, $u(x,x)=\int_0^x\sqrt{k^2-v(t)}dt$, and $g(x,k)$ is the solution to the integral equation,

$$g(x,k)=e^{iu(x,k)}+\int_x^\infty\sin\{u(t,k)-u(x,k)\}R(t,k)g(t,k)dt, \tag{A9}$$

where

$$R(t,k)=-\frac{v''(t)}{4\{k^2-v(t)\}^{3/2}}-\frac{5}{16}\frac{(v'(t))^2}{\{k^2-v(t)\}^{5/2}}.$$

Under condition (A6) [even (1.6)], Eq. (A9) can be solved by iteration. We need two first iterations:

$$g_0(x,k)=e^{iu(x,k)},$$

$$g_1(x,k)=e^{iu(x,k)}+\int_x^\infty\sin\{u(t,k)-u(x,k)\}R(t,k)g_0(t,k)dt. \tag{A10}$$

Due to condition (A6), for $x\geqslant 0$, $k\geqslant k_0$, we have

$$|g(x,k)-g_1(x,k)|\leqslant\left(\int_0^\infty|R(t,k)|dt\right)^2=O(k^{-6}),$$

and, in particular, for $x=0$,

$$|g(0,k)-g_1(0,k)|=O(k^{-6}), \tag{A11}$$

$$g_1(0,k)=1+\int_0^\infty \sin u(t,k)R(t,k)e^{iu(t,k)}\,dt. \tag{A12}$$

Since (Refs. 16 and 17)

$$\theta(k)=\arg f(0,k)=\arg g(0,k)-\alpha(k), \tag{A13}$$

it is enough to take care of

$$\operatorname{Im} g_1(0,k)=\int_0^\infty \frac{1-\cos 2u(t,k)}{2}R(t)dt$$
$$=-\frac{1}{8}\int_0^\infty \frac{1-\cos 2u(t,k)}{\{k^2-v(t)\}^{3/2}}v''(t)dt-\frac{5}{16}\int_0^\infty \frac{1-\cos 2u(t,k)}{\{k^2-v(t)\}^{5/2}}(v'(t))^2\,dt.$$

The second integral is clearly $O(k^{-5})$; for the first one, we have

$$-\frac{1}{8k^3}\int_0^\infty v''(t)dt+\frac{1}{8k^3}\int_0^\infty v''(t)\cos 2u(t,k)dt+O(k^{-5})$$
$$=\frac{v'(0)}{8k^3}+\frac{1}{16k^4}\int_0^\infty \frac{v''(t)}{\sqrt{k^2-v(t)}}d\sin 2u(t,k)+O(k^{-5}).$$

Since $(1-k^{-2}v(t))^{-1/2}=O(k^{-2})$, $k\geqslant k_0$, after integration by parts twice the last equation becomes

$$\frac{v'(0)}{8k^3}-\frac{1}{16k^4}\int_0^\infty v'''(t)\cos 2u(t,k)dt+O(k^{-5})$$
$$=\frac{v'(0)}{8k^3}+\frac{v'''(0)}{32k^5}+\frac{1}{32k^5}\int_0^\infty v^{(4)}(t)\sin 2u(t,k)dt+O(k^{-5})=\frac{v'(0)}{8k^3}+O(k^{-5}).$$

Therefore

$$\arg g_1(0,k)=\operatorname{Im}\log g_1(0,k)=\frac{v'(0)}{8k^3}+O(k^{-5}). \tag{A14}$$

Let us now estimate $\alpha(k)$:

$$\alpha(k)=\lim_{x\to\infty}(u(x,k)-\tau(x,k))=\lim_{x\to\infty}\left\{\int_0^x\sqrt{k^2-v(t)}dt-kx+\int_0^x\frac{\sin^2 kt}{k}v(t)dt\right\}.$$

As one can easily observe, that for $k\geqslant k_0$,

$$\int_0^x\sqrt{k^2-v(t)}dt=kx-\frac{1}{2k}\int_0^x v(t)dt-\frac{1}{8k^3}\int_0^\infty v^2(t)dt+O(k^{-5}),$$

$$\int_0^x\frac{\sin^2 kt}{k}v(t)dt=\frac{1}{2k}\int_0^x v(t)dt+\frac{v'(0)}{8k^3}+O(k^{-5}).$$

Therefore

$$\alpha(k)=\frac{v'(0)-\int_0^x v^2(t)dt}{8k^3}+O(k^{-5}). \tag{A15}$$

Now, combining (A13), (A11), (A14), and (A15), we arrive at (A7), and the lemma is proved.

[1] I. M. Gel'fand and B. M. Levitan, ''On a simple identity for eigenvalues of a second order differential operator,'' Dokl. Akad. Nauk SSSR **88**, 593–596 (1953) (in Russian); English translation in I. M. Gel'fand, *Collected Papers* (Springer-Verlag, Berlin, 1987), Vol. 1, pp. 457–461.

[2] L. A. Dikii, ''Trace formulas for Sturm-Liouville differential operators,'' Am. Math. Soc. Trans. **18**, 81–115 (1961).

[3] B. S. Buslaev and L. D. Faddeev, ''Formulas for traces for a singular Sturm–Liouville differential operator,'' Dokl. Akad. Nauk SSSR **132**, 13–16 (1960) (in Russian); English translation in Sov. Math. Dokl. **1**, 451–454 (1960).

[4] M. Sh. Birman and D. R. Yafaev, ''The spectral shift function: The papers of M. G. Krein and their further development,'' St. Petersburg Math. J. **4**, 833–870 (1993).

[5] V. A. Lyubishkin, ''On the trace formulas of Gelfand–Levitan and Krein,'' Mat. Sb. **74**, 531–540 (1993).

[6] F. Gesztesy and B. Simon, ''The xi function,'' Acta Math. **176**, 49–71 (1996).

[7] F. Gesztesy, H. Holden, and B. Simon, ''Absolute summability of the trace relation for certain Schrödinger operators,'' Commun. Math. Phys. **168**, 137–161 (1995).

[8] F. Gesztesy, H. Holden, B. Simon, and Z. Zhao, ''Higher order trace relations for Schrödinger operators,'' Rev. Math. Phys. **7**, 893–922 (1995).

[9] A. V. Rybkin, ''The spectral shift function, the characteristic function of a contraction and a generalized integral,'' Russ. Acad. Sci. Sb. Math. **83**, 237–281 (1995).

[10] A. V. Rybkin, ''On A-integrability of the spectral shift function of unitary operators arising in the Lax–Phillips scattering theory,'' Duke Math. J. **83**, 683–699 (1996).

[11] D. R. Yafaev, ''Trace formulas for charged particles in non-relativistic quantum mechanics,'' Teor. Mat. Fiz. **11**, 78–92 (1972) (in Russian); English translation in Theor. Math. Phys. **11**, 358–369 (1972).

[12] A. Kvitsinsky, ''Analog of Levinson's formula for a Schroedinger operator with a long-range potential,'' Teor. Mat. Fiz. **68**, 244–254 (1986) (in Russian); English translation in Theor. Math. Phys. **68**, 801–808 (1986).

[13] A. Kvitsinsky, ''Trace formula for Schroedinger operator with a Coulomb potential in three-dimensional space,'' Teor. Mat. Fiz. **70**, 104–114 (1987) (in Russian); English translation in Theor. Math. Phys. **70**, 72–79 (1987).

[14] A. Melin, Trace distributions associated to the Schrödinger operator, J. Anal. Math. **59**, 133–160 (1992).

[15] L. S. Koplienko, ''A regularized spectral shift function for a one dimensional Schrödinger operator with slowly decreasing potential,'' Sibirsk. Mat. Zh. **26**, 72–77 (1985) (in Russian); English translation in Siberian Math. J. **26**, 365–369 (1985).

[16] V. B. Matveev and M. M. Skriganov, ''Scattering problem for the radial Schrödinger equation with slowly decreasing potential,'' Teor. Mat. Fiz. **10**, 238–248 (1972) (in Russian); English translation in Theor. Math. Phys. **10** (1972).

[17] E. C. Titchmarsh, Eigenfunction expansions associated with second-order differential equations (Oxford University Press, Oxford, 1958).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Toward quantum mathematics. I. From quantum set theory to universal quantum mechanics

Karl-Georg Schlesinger[a)]
Erwin Schrödinger Institute for Mathematical Physics, Boltzmanngasse 9, 1090 Vienna, Austria

(Received 10 August 1998; accepted for publication 27 October 1998)

We develop the old idea of von Neumann of a set theory with an internal quantum logic in a modern categorical guise [i.e., taking the objects of the category **H** of (pre-)Hilbert spaces and linear maps as the sets of the basic level]. We will see that in this way it is possible to clarify the relationship between categorification and quantization and besides this to understand that in some sense a categorificational approach to quantization is a discretized version of the one taken by noncommutative geometry. The tower of higher categorifications will appear as the analog of the von Neumann hierarchy of classical set theory (where by classical set theory, we will understand the usual Zermelo–Fraenkel system). Finally, we make a suggestion of how to understand all the different categorifications as different realizations of one and the same abstract structure by viewing quantum mechanics as universal in the sense of category theory. This gives the possibility to view extended topological quantum field theories purely as involving an abstract notion of quantum mechanics plus representation theory without the need to enlarge the class of kinematic structures of quantum systems on each step of categorification. In a future part of the work we will apply the language developed here to deal especially with the question of a categorification of the manifold notion. [S0022-2488(99)03003-0]

I. INTRODUCTION

In recent years, the word *quantum*, originating from physics, has found a rapidly increasing number of occurrences in the mathematics literature. In this paper, we start from the historically first use of *quantum* to denote a mathematical structure, the quantum logic of Birkhoff and von Neumann presented in Ref. 1. But our aims are the modern quantum structures that we hope to understand in a more unified way by taking a logic-based perspective. Beyond this, our goal is to be able to extend quantization to more involved mathematical structures, especially to the nonlinear graviton construction of twistor theory (Ref. 2). Let us now be a little bit more precise about the aim and the content of the paper.

In 1936, Birkhoff and von Neumann suggested in Ref. 1 that the lattice of closed linear subspaces of a Hilbert space determines a nonclassical propositional calculus for quantum objects, in the same way the Boolean lattice on two symbols determines the propositional calculus of classical logic. Later von Neumann proposed that one should consider a quantum set theory, corresponding to quantum logic, as does conventional set theory to classical logic (see Ref. 3 and the literature cited therein). This idea is very intriguing for several reasons: First of all, set theory is the mathematical theory of the notion of pure objects. So, the question for a quantum set theory is the question for a theory of pure quantum objects. This may sound strange to some because quantum mechanics is believed to be just this theory. But there is no contradiction here. We will see that quantum set theory is in some sense just another view on quantum mechanics, albeit one where the machinery of set theory helps us to ask more refined questions on the object notion. If

[a)]Electronic mail: schles@math.uni-wuppertal.de

one is strongly convinced that there is a notion of mathematical object beyond any physical experience and that set theory captures this notion, then there is, of course, no need to change set theory on the advent of quantum mechanics. But if one is prepared to accept that even the abstract object notion of set theory might just be distilled from our experience of physical objects (as von Neumann believed), we have to ask which object notion (and therefore set theory) the fundamental theories of physics really determine. Just as we had to learn that geometry is as well a part of physics as of mathematics, it might well turn out that logic and set theory are, too (a similar standpoint is also taken by Deutsch in the closely related field of quantum computation; see Ref. 4).

Second, the idea that even set theory should be quantized—and not only geometry or topology—has been around already for quite some time in the field of quantum gravity (see, e.g., Refs. 5, 6). There is a simple argument supporting this view: In general relativity the set theoretic structure is of immediate physical relevance. The points of space–time (i.e., the elements of the underlying set) represent the physical events. In a diffeomorphism invariant theory this has no meaning in the trivial sense of a point having four coordinate values, but a point has to be identified as the intersection of the world lines of particles or by values of fields (e.g., by the null cone structure and a conformal factor, as is done in twistor theory). So, by the diffeomorphism invariance of the theory, the set theoretic structure is intimately linked to something like collision processes of particles. For this reason, we should expect that this structure might become smeared in a full fledged theory of quantum gravity.

We will use the language of category theory to deal with nonclassical set theory. In some sense categories can be thought of as generalizations of the classical set universe **Set** (the category of sets and set theoretic functions). For special categories—topoi and certain generalizations thereof—this has been made precise, i.e., they have been shown to correspond to set theories with a nonclassical internal logic (see Ref. 7, which also gives an easy introduction to general category theory). We consider the category **H** of complex (pre-)Hilbert spaces and linear maps as the (basic) quantum set universe. (Because there appear unbounded operators in quantum mechanics, we normally do not assume continuity of the maps in **H**. This sometimes causes technical problems when dealing with certain subspaces defined by linear operators, because we do not easily get closure properties then. In order to avoid these difficulties in a first approach to the general structures involved, we work with pre-Hilbert spaces instead of Hilbert spaces here.) Though **H** is not a topos, we will see that it has properties similar to one. In spite of this, in doing mathematics in quantum set theory, we are not going to follow strictly the approach of topos theory of doing everything inside **H**, i.e., in terms of objects of **H**. We will allow ourselves the freedom to consider structures in quantum set theory as objects outside of **H**, too, because this is just the way ordinary mathematics is done in practice. Of course, we can formulate classical mathematics largely in categorical language inside **Set**, but for the working mathematician this is not always the most convenient way to conceive of the structure he is dealing with (unless he wants to transport the structures to another category). The question if at least in principle a reformulation totally inside **H** is possible in quantum mathematics, too, is certainly an interesting one from the foundational viewpoint, but since our aim is to apply quantum set theory to problems in mathematical physics, we will not deal with it here (the differences of **H** to a topos should one make ready to the possibility that a reformulation might not be possible). But we will use the idea that set theory is universal in the opposite way, namely, as a requirement saying that all structures we use can also be regarded as sets. This leads to an enlargement of quantum set theory beyond **H**, i.e., we regard the objects of **H** as some kind of basic level of set theory with other levels present, too. This approach allows us to keep the universality of set theory and the possibility to stay close to the spirit of the working mathematician (and it will give some interesting results).

Remark 1: Obviously, Hilbert spaces are not the true state spaces of quantum systems but contain redundant information. To avoid this, one one would have to go over to the projective case. On the one hand, this would make some structural elements more straightforward (e.g., there would exist a true terminal object then; see below). On the other hand, Hilbert spaces have established themselves as the language of the working physicist because their linear structure is *very convenient in other respects. Since our main aim is to develop quantum set theory as a tool for mathematical physics, we stay to the convention used in this field and take the category* **H** *as our basic quantum sets.*

Besides this, a reader comparing to other approaches to nonclassical set theories, should keep in mind that—by using a categorical approach—like in topos theory, the subobject (instead of the membership) relation will become the fundamental one.

As we mentioned already at the beginning, our first aim is to understand the quantum structures appearing in the modern literature from a more unified perspective. For the modern approaches, the idea of categorification is central. (A categorification of an algebraic structure is a category with a similar structure. For example, the categorification of a module is a module category, i.e., a category with a functorially given structure satisfying the module axioms up to isomorphism. In addition, one normally needs so-called *coherence conditions* for an iterative application of the axioms. For example, for a structure with associativity up to an isomorphism, one needs an additional axiom stating that even four—not only three—factors can be rebracketed, which then suffices to guarantee that an arbitrary number of factors can be. See Ref. 8 for the details.) There is especially the idea that categorification is linked to quantization (see again Refs. 8 and 9, 10). We will see that we can make this correspondence precise in our logic-based approach.

A more advanced goal is then to quantize more involved (nonalgebraic) mathematical structures like manifolds. Especially, we are interested in a quantization of Penrose's nonlinear gravitons. In Ref. 2, Penrose suggested to consider nonlinear gravitons as the one-particle states of a future quantum theory of gravity. Since the standard symmetric Fock space construction is not applicable in this case, the proposal has not been put into a proper theory since then. It seems therefore to be natural to explore the possibility if a different approach to quantization can lead to a mathematical structure that is suitable for a second quantized (i.e., many-particle) theory of nonlinear gravitons.

This work consists of two (possibly three) parts. In this paper we deal with the fundamentals, clarifying the relationship between categorification and quantization. The more involved applications will follow in the forthcoming part(s), where we will also compare the approach presented here to the results in Ref. 11 on a quantization of the category of topological spaces and continuous injections.

As the title indicates, we consider this as an approach toward a quantum mathematics, i.e., a mathematical theory where all the ingredients (like logic and set theory) adhere to the rules of quantum mechanics. We inserted the word *toward* in the title because we surely do not consider the following results as the final word on the structure of a theory of quantum mathematics. It is only a try to explore the pathway that finally could lead to such a theory.

In closing the Introduction, let us now give a short overview of the content of the individual parts of the paper. In Sec. II, we investigate in which aspects the category **H** is similar to a topos. We will find that there is no terminal object in a technical sense, but that the one-dimensional Hilbert space $\mathbb{C}$ for several purposes takes the role of it. We have some kind of exponentiation and we have pullbacks though—if we use the tensor product to construct pullbacks (which is the favorable approach from the standpoint of quantum mechanics)—we do not get universality but a superposition of pullback structures. The same holds true for the case of the subobject classifiers. We will see that this is linked to the fact that quantum set theory may be understood as describing the observation of a quantum system by a quantum system and the iteration of this process.

In Sec. III, we go on to introduce the analogs of number systems in quantum set theory. We start with the natural numbers that can be formulated in two different ways: Either internal (as an object in **H**) as the universal infinite-dimensional Hilbert space $\mathcal{N}$ or external as the category **Hilb** of finite-dimensional Hilbert spaces and linear maps. Reals and complex numbers are then found as the self-adjoint operators, respectively, the whole operator algebra, on $\mathcal{N}$.

Having introduced number systems, in the first part of Sec. IV, we deal with the analog of module structures. We then discuss the relevance of this for the categorifications appearing in TQFTs. The rest of this section is devoted to presenting the quantum version of the von Neumann hierarchy of classical set theory up to the level of the first infinite ordinal. Actually, we will see that there are two different but related hierarchies because there are two different set concepts present in quantum set theory.

A unified view on all the different levels of quantum set theory is presented in Sec. V. By considering quantum mechanics as universal in the sense of category theory, we understand all the (higher) categorifications as different realizations of one and the same abstract structure. Section VI contains some concluding remarks. Finally, we should remark that often details of the structures presented are not worked out precisely. For example, we use the concept of weak n and weak ω category, though the existence of a single concept of this form is still in the stadium of proof (but there is hope that it will be established in the near future; see the literature cited below). We think that in a new subject like quantum set theory it is justified first to explore the general territory emerging and try to get a feeling for its general power and possible limits before sticking to the technical details.

II. THE TOPOS-LIKE STRUCTURE

In this section we will explore in which respect the category **H** of complex (pre-)Hilbert spaces and linear maps resembles a topos. Recall that a topos is a category that has a terminal object, pullbacks, exponentiation, and a subobject classifier (see Ref. 7). We will consider the existence of all of these structures in **H**, in turn, now.

A. Terminal objects

A terminal object C of a category $\mathcal{C}$ is an object such that for any other object D of $\mathcal{C}$ there is one and only one arrow in $\mathcal{C}$ from D to C. Since there is no homomorphism set with one element in **H** (because all the homomorphism sets have a linear structure themselves), we do not have a terminal object.

One could have the idea that maybe the question is wrongly posed. Since only the projective (i.e., subspace) structure is relevant in quantum mechanics, we should really not look for single element homomorphism sets but for one-dimensional ones. But even in this sense, we do not have a terminal object in **H**. Nevertheless, we will see below that the one-dimensional Hilbert space $\mathbb{C}$ (or, of course, any other one-dimensional Hilbert space) in some respects plays the role of a terminal object in **H**, though it is not one in a technical sense. (The reason why this works is that the true state spaces in quantum mechanics are the projective Hilbert spaces. In the category of these spaces $\mathbb{C}$ is terminal.)

B. Pullbacks

A pullback for a pair of arrows,

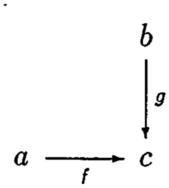

is an object d together with arrows f', g', making

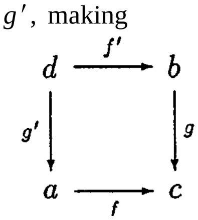

commutative and satisfying that for any other object e and arrows h,j with this property there exists precisely one arrow,

$$k : e \rightarrow d,$$

making

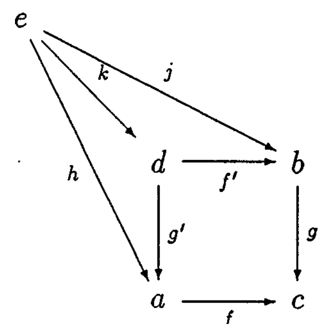

commutative.

Now, suppose

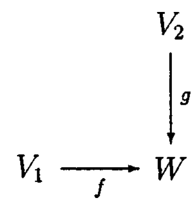

is given in **H**. Consider the direct sum and product Hilbert spaces $V_1 \oplus V_2$ and $V_1 \otimes V_2$ and denote in both cases by pr_1, pr_2 ''projections'' on the first and second component, respectively. Here, we put ''projections'' into quotation marks because in the tensor product case there exist no projections in the technical sense, of course. In this case, we mean the following: Fix a product basis in the product space and use the linear extension of the projection onto components existing for this basis. So, the construction we use is dependent on the additional structure of a fixed basis. But this dependence is not relevant from the standpoint of quantum logics, since there is redundant information in the Hilbert space structure (see above), and the remaining nonuniqueness is precisely described by the superposition structure below. Then

$$V_+ = \{ v \in V_1 \oplus V_2, \quad (f \circ \mathrm{pr}_1)(v) = (g \circ \mathrm{pr}_2)(v) \}$$

and

$$V_\times = \{ v \in V_1 \otimes V_2, \quad (f \circ \mathrm{pr}_1)(v) = (g \circ \mathrm{pr}_2)(v) \}$$

are pre-Hilbert spaces and

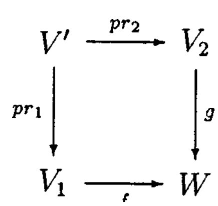

(where V' stands for either V_+ or $V_\times$) is commutative. For the case of V_+ the universality follows directly from the properties of the direct sum. If we use $V_\times$ instead, it is easy to show that a map k exists, too, but it is not unique. We only get uniqueness on product elements because for a general linear combination of product elements the application of pr_1 and pr_2 gives only the sum of the projections of the individual product elements. As a consequence, the pullback structure is not universal in this case—we could shift the uniqueness from the product elements to other ones by using more general projections instead of pr_1 and pr_2—but we get a kind of superposition of pullback structures. The structure of this superposition—i.e., the nonuniqueness of the pullback—is determined by the projective structure of $V_\times$. We will call such a superposition of pullback structures a *quantum pullback*.

Why do we care for these superposed structures at all and not simply restrict ourselves to the true pullbacks defined via the direct sum. The reason is that from the standpoint of quantum mechanics, of course, the tensor product is the appropriate product to be used in the category **H**. Indeed, the fact that superpositions of structures occur in this case is a fundamental feature of quantum set theory. From the standpoint of the category theorist, the pullback is one of the most fundamental structures leading to a level raising in set theory, i.e., on applying a pullback we climb up the ladder of the von Neumann hierarchy. We will understand below why this is related to an occurrence of superpositions in a set theory with an internal quantum logic.

C. Exponentiation

A category with a product $\times$ has exponentiation if for any pair of objects a and b there is an object b^a and an arrow,

$$ev: b^a \times a \to b$$

(called evaluation), such that for any object c and arrow

$$g: c \times a \to b,$$

there is a unique arrow

$$\hat{g}: c \to b^a,$$

making

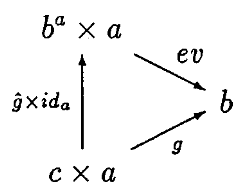

commutative.

In the case of **H**, take as a product the tensor product (the direct sum works correspondingly) and define

$$W^V = \mathrm{Hom}_{\mathbf{H}}(V, W)$$

and

$$ev: W^V \otimes V \to W,$$

as the evaluation map.

Now, consider the following diagram:

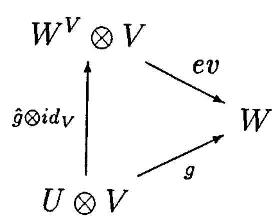

where U and g are supposed to be given. For $u \in U$ define $g_u: V \to W$ by

$$g_u(v) = g(u \otimes v),$$

and let $\hat{g}(u)=g_u$. Then clearly the diagram above commutes. On the other hand, commutativity of the diagram implies

$$ev(\hat{g}(u)\otimes v)=g_u(v),$$

i.e.,

$$\hat{g}(u)=g_u,$$

and therefore uniqueness of $\hat{g}$. So, **H** has a kind of exponentiation but not a true one in the technical sense of topos theory because $\mathrm{Hom}_{\mathbf{H}}(V,W)$ is not an object of **H**. [This is only true if we restrict to finite-dimensional spaces where $\mathrm{Hom}_{\mathbf{H}}(V,W)$ can be identified with $W\otimes V^*$.] But for the reason explained above, we do not consider this as a serious drawback.

D. Subobject classifier

This is the most interesting part of the topos structure since it refers to the internal logic.

In a category with a terminal object 1, a subobject classifier is an object Ω together with an arrow,

$$\mathrm{true}:1\rightarrow\Omega$$

(the name refers to the fact that it represents the truth value ‘‘true’’ while Ω is the generalization of the Boolen algebra on two symbols of ordinary set theory, i.e., giving all the possible truth values) having the property that for each monomorphism,

$$f:a\rightarrowtail d,$$

there is precisely one arrow,

$$\chi_f:d\rightarrow\Omega,$$

making

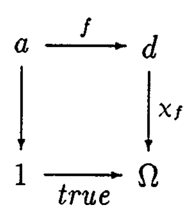

into a pullback square (observe that $a\rightarrow 1$ is unique by the fact that 1 is a terminal object). Since monomorphisms define subobjects, we can regard χ_f as the characteristic function of

$$f:a\rightarrowtail d$$

in d.

Since the definition of a subobject classifier involves the terminal object in an essential way, we cannot directly use the approach of topos theory.

In a topos the object Ω has the role of representing the internal logic of the topos. Though **H** is not a topos, one suspects that the two-dimensional Hilbert space $\mathbb{C}^2$ should play a role similar to that of Ω in a topos by representing the binary quantum logic. Since the Birkhoff and von Neumann paper,[1] it has become customary to name as quantum logic every lattice of closed subspaces of a Hilbert space (and there are even more general definitions used in the literature). But considering a general Hilbert subspace lattice as some kind of logic is just the quantum counterpart to interpreting a general Boolean algebra as logic, i.e., in doing this, one mixes some kind of classically incomplete knowledge with the linear subspace structure imposed by quantum mechanics. So, the quantum analog of the classical binary logic—where one has the maximal obtainable information—is the quantum binary logic represented by the subspaces of $\mathbb{C}^2$ (physically speaking, representing a quantum bit of information). So, in considering **H** in comparison to **Set**, we should look at $\mathbb{C}^2$ as the analog of $2=\{0;1\}$.

As in **Set**, we choose a linear map,

$$\text{true}:\mathbb{C}\to\mathbb{C}^2,$$

as representing the truth value true. Let U be a subspace of V in **H** with inclusion map

$$f:U\rightarrowtail V.$$

If the following diagram gives one of the possibilities in the superpositions of a quantum pullback square,

$$\begin{array}{ccc} U & \xrightarrow{f} & V \\ \downarrow & & \downarrow \chi \\ \mathbb{C} & \longrightarrow & \mathbb{C}^2 \end{array}$$

we call χ *a quantum characteristic function (qcf)*. Here

$$\mathbb{C}\to\mathbb{C}^2$$

is the arrow true and the arrow

$$U\to\mathbb{C}$$

needs some extra comment: In a projective sense this arrow is unique, so, we may without loss of generality assume that for every space U a fixed choice of a map to $\mathbb{C}$ has been made.

As a consequence of the quantum pullback structure, we have again a superposition structure in the characteristic functions. Let us now discuss the relevance of this. Consider a quantum object described by a (pre-)Hilbert space V. The subobjects of it are then described by the linear subspaces U of V if we take an external view, i.e., if we describe the quantum object V as seen from outside by a classical observer. This is the view normally taken in quantum mechanics. The superposition structure of quantum characteristic functions means that we now do not describe the subobjects by the subspace lattice but by a kind of lattice of lattices where the structure of the outer lattice is given by the projective structure of U. This means we view the subobject question as seen from inside by the quantum object U itself. From the quantum perspective, the question of the extension of a subobject (which is the one, one asks when building a characteristic function) again has a superposition of answers. If we would apply a pullback construction twice, we would get a lattice of lattices of lattices, and so on, on iterated application. So, in this sense quantum set theory describes the observation of a quantum object by a quantum object and the iteration of this process. The higher levels of quantum set theory therefore can be seen as refining the notion of a quantum object beyond the level of description from the perspective of a classical one. To deal with questions in atomic physics, the higher levels of quantum set theory are surely not needed, but in other fields like quantum cosmology or the theory of quantum computation these structures could well be of interest. For instance, we can imagine a quantum computer observing a quantum system or the self-measurement of a quantum mechanical system. Questions of this type are beginning to be touched in the field of quantum computation (see the article[4] of Deutsch and the work mentioned there).

Finally, we should mention that the superposition structure of quantum set theory is also captured by the model theoretic approach of Takeuti (see Ref. 12), but we prefer the category theoretic approach because it makes the general scheme very transparent and allows us to connect it to results in modern mathematical physics.

In conclusion, **H** closely resembles a topos, the main difference being the appearance of superpositions of structures (if one uses the tensor product to define pullbacks). We could now use the topos-like structure to develop propositional calculus and elementary set algebra in **H**, along the lines this is done in topos theory (see Ref. 7). We refrain here from doing so explicitly because the results closely resemble the ones for set theory in the category of topological quantum field theories (though they are here often considerably more simple), which is presented in detail in Ref. 13. The only point on elementary set theory we would like shortly to mention is the question of extensionality.

The axiom of extensionality states that sets are described by their elements and nothing more, i.e., two sets are equal if and only if they have the same elements (this is maybe the most basic property of conventional set theory). Stated for functions—and in this way leading to a generalization to category theory—this means that two functions are equal if they always give the same value. Since extensionality can already be violated in topos theory (see Ref. 7), we have to ask the question if it holds true in **H**. Suppose two parallel arrows,

$$f,g:V\rightrightarrows W,$$

in **H** are given that are not equal, even in the sense of maps between projective spaces (which is surely the degree of unequalness we should require). Deciding them on an element in categorical terms means we have to give a map j from the terminal object (i.e., here $\mathbb{C}$) to V, such that

$$f\circ j\neq g\circ j.$$

But this is possible by taking j to be a map that assigns to $\mathbb{C}$ a subspace of V, where f and g differ. The other direction is trivial anyway, so we have extensionality for the basic level of quantum set theory.

III. NUMBER SYSTEMS

Having seen that **H** may be considered as a quantum set universe, it is natural next to ask if we can find analogs of the usual number systems in quantum set theory. We start by developing a concept of natural numbers.

Observe that we can equip **Set** with some additional structure: Taking the Cartesian product and the disjoint union of two sets as the product and sum on **Set**, respectively, makes **Set** into a rig category (see Ref. 8, the word *rig* is alluding to ring, the difference being that there need not be inverses with respect to the operation of addition). Then the usual arithmetic on the natural numbers is just the one induced by this category arithmetic, and the natural numbers themselves may be defined by successively taking sums of a terminal object in **Set** with itself. Especially, by taking the sum of a terminal object with itself one time, we get an object isomorphic to the subobject classifier $\{0;1\}$, determining the internal logic of **Set**.

Now, **H** may be considered a rig category, too, by equipping it with the tensor product and direct sum of (pre-)Hilbert spaces. As we noticed, in **H** the space $\mathbb{C}$ plays the role of a terminal object, i.e., we get the space $\mathbb{C}^n$ as the analog of the natural number $n\in\mathbb{N}$. In **Set**, $\mathbb{N}$ is just defined as being the target of an inclusion arrow for every natural number and by being minimal with respect to this property (minimality being defined by a factorization property for arrows). In **H** these requirements are fulfilled by an infinite-dimensional separable Hilbert space $\mathcal{N}$.

Remark 2: Since all infinite-dimensional separable Hilbert spaces are unitarily equivalent, $\mathcal{N}$ *may be considered as the universal infinite-dimensional separable Hilbert space.*

Alternatively, we could identify the natural numbers with the finite sets, i.e., taking the category **Hilb** of finite-dimensional Hilbert spaces as the analog of $\mathbb{N}$. Since finite sets are distinguished from natural numbers only by isomorphisms, it is natural to identify the two concepts in a categorical approach, i.e., to call every finite set a natural number. Both approaches—considering the universal infinite-dimensional separable Hilbert space or the category **Hilb** as the natural numbers—are fruitful, as we will see. To distinguish them, we will speak of the *internal* and *external natural numbers*, respectively, because in the latter case they are a subcategory of **H** but not an object in **H**.

Remark 3: There is still another approach to natural numbers in **H**, *namely trying to mimick the definition of a so-called natural numbers object of topos theory (see Ref. 7). We do not follow this approach explicitly here because in the case of* **H** *it is a little bit clumsy and in the end leads to considering* $\mathcal{N}$ *as the natural numbers, too.*

Let us now go on to consider an analog of the complex numbers (we use the internal natural numbers for this). Considered purely as a set, the complex numbers are isomorphic to the set of functions from $\mathbb{N}$ to $\mathbb{N}$ (because they have the same cardinality). In **H** this means that the purely set theoretic structure should be given by the space of linear operators on $\mathcal{N}$. Now, the complex numbers are, of course, not only a set but they carry an algebraic structure of addition and multiplication. On

$$\mathrm{Hom}_{\mathbf{H}}(\mathcal{N},\mathcal{N}),$$

such a structure is naturally given by the sum and product of operators, so we take this. To be able to talk sensibly of complex numbers, one more structural element is needed, namely, an operation of taking real and imaginary parts (and thereby defining an operation of complex conjugation). But this again is possible in the operator algebra on $\mathcal{N}$ by taking the self-adjoint, respectively, anti-self-adjoint, part of an operator. So, the algebra of linear operators on $\mathcal{N}$ suggests itself as the counterpart of the complex numbers. As a consequence, the self-adjoint operators on $\mathcal{N}$ take the role of the reals.

Remark 4: The fact that the self-adjoint operators take the role of the reals in quantum set theory has also been noted in Refs. 12 and 14, but here it is a direct consequence of very simple set theoretic and algebraic arguments.

That the self-adjoint operators appear as the reals of **H** gives some justification for the practice of physicists to put an overcaret on the symbols when they quantize the corresponding theory. It is therefore a first very slight confirmation to the hope that quantum set theory might be useful as a tool for the quantization of physical systems. But only a very slight one, because we do not get a prescription for which equations between real numbers translate to the corresponding ones between operators (and in which form, i.e., the factor ordering problem). That not all equations can translate is clear from the fact that quantization is proved not to be a functor (see Ref. 15).

Remark 5: We can, of course, identify the natural numbers with the projection operators and the sum of two natural numbers with the sum of disjoint projection operators. In this way, we have an embedding of the natural numbers into the real ones, as we are used to from classical mathematics.

Observe that the separability of the Hilbert space is nearly automatic in this approach.

Remark 6: We did not care for analogs of negative numbers above, but these can, of course, be gained by applying the usual algebraic construction (see Ref. 16).

Remark 7: Considering set theory as universal, we should require that numbers are sets. This is a first example, where we get an enlargement of quantum set theory beyond **H** *to operator algebras (see Ref. 17 for a reference on their theory).*

Considering the operator algebra on a separable Hilbert space as the analog of the complex numbers is just what is done in noncommutative geometry, where in this way one regards noncommutative C^* algebras as the algebras of complex-valued functions on some quantum space (see Refs. 18, 19, the latter reference really gives a representation theorem for a special class of C^* algebras in terms of operator-valued functions on some discrete space). In the spirit of algebraic geometry, one can then develop geometry on these algebras.

There is one more number concept in quantum set theory, namely, the algebraic operations on **H** inducing the ones on **Hilb**. This is the analog of Cantor's cardinal arithmetic in classical set theory.

IV. MODULES, TOPOLOGICAL QUANTUM FIELD THEORY, AND THE VON NEUMANN HIERARCHY

Having introduced analogs of the number systems, the next step is to consider modules over these. We follow the approach based on the external natural numbers in this section because it proves to be especially fruitful here. Let us start with finite-dimensional modules over the natural numbers. This is just a collection of tuples of natural numbers subject to the laws of addition and scalar multiplication, i.e., in the case of quantum set theory we get tuples of finite-dimensional Hilbert spaces subject to the mentioned laws. And since we use a categorical approach, we should only require these laws up to an isomorphism (so, we have to introduce coherence conditions, in addition, to ensure that the iterative application of the laws is possible). There immediately is a concept of a map between tuples of finite-dimensional Hilbert spaces, namely, a tuple of linear operators. So, a finite-dimensional module over the natural numbers in quantum set theory can be understood as just what is called a finite-dimensional module category over **Hilb** (or a 2-vector space in the terminology of Refs. 8 and 20). The analog of general (not necessarily finite-dimensional) modules over the natural numbers are then general module categories over **Hilb**. In the same spirit we can consider module categories over **H**.

The fact that modules over the natural numbers turn into module categories over **Hilb** gives a precise meaning to the connection between categorification and quantization discussed in Refs. 8–10. Categorification just means looking at the structure from the perspective of quantum logic. We have explicitly seen this for modules and for rings, but it extends to other algebraic structures, too. The reason is that having a module concept, we can easily discover the group (and then monoid) concept by forgetting part of the structure or, alternatively, by considering symmetry actions on the module. It turns out this way that monoids in quantum set theory are just monoidal categories. In the same way we proceed to discover other algebraic structures. There is one case that should be specially mentioned here, namely, forgetting all of the operations of an algebraic structure. This leads to the underlying sets in the classical case. Here we get general categories in this way, which is somewhat surprising because we already introduced a quantum set notion. This is a feature of quantum set theory we already encountered in the case of number systems: Starting from different classical versions of one and the same concept, we may arrive at two different and not necessarily equivalent concepts in quantum set theory. (Observe that even in topos theory this already occurs, e.g., there the Cauchy and Dedekind reals do, in general, not agree; see Ref. 7). The quantum sets as objects of **H** constitute the basic level of logically defined sets, i.e., sets are understood as realizing the logic in an extensional sense. The quantum sets as general categories give the basic level of sets understood as carriers of algebraic structures in quantum set theory. Observe that the structures in the second case followed from the logically conceived sets because quantization by categorification was a consequence of our logically minded approach. In the past, e.g., in the early works of von Neumann, quantum sets were only considered from the logical perspective. We feel that it is a decisive prerequisite for applications of quantum set theory to decide clearly between the two different set versions. In the sequel, we will term them *quantum logic sets* and *quantum algebraic sets,* respectively, if it is not clear from the context which concept we refer to. We will see below that the link between categorification and quantization carries through even to the higher levels (in the sense of the von Neumann hierarchy) of quantum set theory.

In the same way as for the case of categorifications, we see that the approach of noncommutative geometry (where one uses the operator algebra on $\mathcal{N}$ as complex numbers) again means viewing structures from the perspective of quantum logic but this time structures involving the real or complex numbers instead of the natural numbers (module categories over **Hilb**) or set algebra (module categories over **H**). So, quantum set theory clarifies the relationship between noncommutative geometry and categorifications. The latter in spirit gives a discretized version of the former one. This puts, especially, stress on the question for categorifications of manifolds because manifolds in the case of noncommutative geometry are already known.

Categorifications are of special interest in the realm of (extended) topological quantum field theories (see Ref. 8 and the literature cited therein). (Extended) n-dimensional topological quantum field theories (TQFTs) are representations of the (extended) n-dimensional cobordism category in **Hilb** (or a possibly higher categorification thereof in the extended case). In quantum mechanics, studying the quantum counterpart of a classical system means studying the Hilbert space representations of its algebraic description. TQFT shows that for very large systems—like the cobordism categories—this scheme has to be enlarged: We have to allow for representations in a module category (**Hilb** is the prototype of a module category) or even higher categorification thereof (the question of giving a Hilbert space structure on such higher modules is dealt with in Refs. 8, 16 and 21). Quantum set theory sheds new light on this enlargement of the scheme: We have to allow for modules in quantum set theory instead of simple vector space structures only, i.e., for the quantized version of the structures. We will now see that this view carries through, even to the higher levels of categorification.

We have mentioned above that quantum set theory can be seen as describing the observation of a quantum system by a quantum system and the iteration of this process. Since we have seen that categorification means quantization of a structure, one suspects that higher categorifications should be linked to iterated quantization, i.e., to higher levels in the von Neumann hierarchy of quantum set theory. This is indeed true.

Climbing up the von Neumann hierarchy in classical set theory involves the application of the power set operation. But there is another possibility of viewing the von Neumann hierarchy, namely, as incorporating the universality requirement of set theory: Higher-order structures should themselves be interpretable as sets. From a quantum perspective the most important structure is surely the module one, since it defines the superposition structure of quantum theory itself. So, the next higher level should be seen as given by quantized module structures, i.e., by module categories. Remember that in Sec. I we have seen that in an internal view of quantum set theory the subspaces of a module are given by superpositions of subspace lattices. Therefore the power set approach, too, leads to a consideration of superposed module structures, as we do in a module category. It is then obvious how to proceed. We have to consider module structures over the weak 2-category of module categories (or better 2-Hilbert spaces) in the next step, i.e., weak module 2-categories. (For the notions of higher category theory see, e.g., Ref. 22. The notion of a weak n-category for $n \geqslant 4$ has long not been precise, but there are now different approaches available, see Refs. 23–26 and there is hope that they can be proved to be equivalent. We therefore feel free to proceed as if there were a single coherent concept.) In general, sets of the next higher level correspond to the module structures of the foregoing one, i.e., the finite part of the quantum von Neumann hierarchy is given by the tower of weak n-categories (with a module and Hilbert space structure). At the ω level (i.e., the level of the first infinite ordinal of classical set theory) we then reach weak ω categories.

Remark 8: The tower of weak ω categories gives the von Neumann hierarchy of the quantum logic sets. The von Neumann hierarchy of the quantum algebraic sets is given by considering categories in **Cat** *that are (weak) double categories (see Ref. 22), and proceeding in this way,* i.e., *we get the tower of weak n-tuple categories (we term the more general cubic version of higher categories as n-tuple categories and the spherical version as n-categories). At the ω level the monoidal globular categories (MGCs) of Batanin appear (see Refs. 25 and 27). From the perspective of quantum set theory, the difference between the two types of higher categorifications is therefore due to the two different set concepts. It would therefore be interesting to see applications of the cubic version in mathematical physics.*

We could now develop analogs of number systems on each level of the von Neumann hierarchy, e.g., on the second level we could consider the weak 2-category of weak 2-Hilbert spaces as natural numbers. The module structure preserving functors on an infinite-dimensional 2-Hilbert space with a countable base are then a candidate for complex numbers, and we could go on to consider categorifications of the algebraic structures appearing in noncommutative geometry (this is highly nontrivial in detail, of course, but the general direction is clear). We see a double role for categorification here: As we remarked above, on the first level it can be seen as a discrete version of the approach of noncommutative geometry. But since it is also the operation for climbing up the von Neumann hierarchy, on the higher levels the two approaches can be mixed. Nevertheless, the whole von Neumann hierarchy of quantum logic sets is constructed in a discretized version of quantum theory because we always consider analogs of modules over discrete rigs.

It is certainly an interesting task for future work to try to understand the analogs of the higher transfinite levels of the von Neumann hierarchy in quantum set theory. But for the moment we stop at the ω level and in the next section try to understand what we have reached so far from a still more unified perspective.

V. UNIVERSAL QUANTUM MECHANICS

We have seen in the last section that on quantizing the structure of quantum mechanics again (i.e., categorifying it), we get module categories instead of vector spaces, in the next step we come to weak module 2-categories, and so on. But the notion of a module can be given an abstract sense by formulating it in purely arrow language. For example, the notion of an object with an associative addition can be formulated in every monoidal category $\mathcal{C}$ with product $\otimes$ as an object C of $\mathcal{C}$ together with a morphism,

$$f: C\otimes C \rightarrow C,$$

making

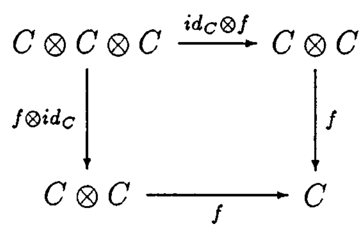

commutative. To formulate the existence of a unit with respect to f, we have to use the unit with respect to $\otimes$. In this way we can define the notion of a ring in $\mathcal{C}$ and then a module structure over it. A module in the category **Cat** then is just a module category over some ring category while a usual module is one in **Set**. A 2-category with a module structure turns out to be a module in **2-Cat**, the category of 2-categories, and so on. There is one problem concerning this scheme: We get only the strict versions of the structures in this way (i.e., in the categorification the axioms of the algebraic structure are satisfied precisely and not only up to an isomorphism), e.g., we get a 2-category with a module structure and not a weak one. One can remedy this problem by taking the following standpoint: Together with the category $\mathcal{C}$ one should specify what commutativity of a diagram in $\mathcal{C}$ means. In the category **Cat**, which is actually a 2-category, it should mean commutativity up to natural transformations. In this way we could include the weak versions of the structures if there would not be the problem to specify the correct coherence conditions to be attached. From an abstract point of view, we can see the coherence conditions as the requirement that the laws of the structure—here the module axioms—should be iteratively applicable. The precise formulation of the conditions in a category $\mathcal{C}$ can then be a very nontrivial matter (as is known from higher-order categorifications), but it can be seen as belonging to precisely working out the concrete structure realized in $\mathcal{C}$ and not to the abstract concept of the structure (where only the existence of conditions giving coherence is required).

We can therefore subsume all the different iteratively categorified structures (i.e., all the different levels of quantum set theory) into one abstract structure concept by adding one additional postulate to conventional quantum mechanics:

Postulate (universality of quantum mechanics). Quantum mechanics is universal in the sense of category theory, i.e., we conceive of it as an abstract structure formulated in arrow language that can be concretely realized in every category $\mathcal{C}$ with a monoidal structure.

Remark 9: The work in Refs. 8, 16, and 21 on the categorification of the full Hilbert space structure shows also a way to formulate this structure in the abstract setting.

The postulate above may need some additional restriction in the sense that one has to be more specific about the ring (or rig) of scalars allowed. Certainly one should allow for the usual number systems and their categorifications, but one is less sure, e.g., concerning the p-adic numbers, though these now turn up from time to time in the physics literature.

A universality postulate for quantum mechanics frees one from the need to extend quantum mechanics to include axiomatic TQFT, i.e., we do not have to enlarge the class of kinematic structures appearing for quantum systems with each level of categorification. If we accept a universality postulate, we have once and for all one fixed abstract structure that has realizations in different categories $\mathcal{C}$. Just as we conclude from special properties of a group that we should, e.g., study its representations on an infinite-dimensional Hilbert space, we then have to conclude from the fact that we want to represent a weak n-category that we have to choose $\mathcal{C}$ as the weak $(n+1)$-category of weak n-categories in this case. We want to call the abstract structure referred to by the postulate above *universal quantum mechanics*.

Besides the concrete arguments mentioned, the postulate is certainly appealing from a principle perspective: Why should quantum mechanics be bound to the category **Set** of classical transfinite set theory? Do we not anyway believe that its principles are of an abstract algebraic nature independent of a set-theoretic base?

From the perspective of quantum set theory the universality postulate means that we can describe all levels of it in a single unified way.

VI. CONCLUSION

In this paper, we introduced the language and developed the fundamentals of a quantum set theory. We have seen that in this way we can understand the relationship between quantization and categorification and see that, indeed, categorification is quantization in the sense of shifting to a nonclassical set theory. We also understand the relationship to the view taken by noncommutative geometry in this way. Finally, we presented a suggestion of how to view the different levels of quantum set theory (i.e., the different categorifications) from a unified perspective.

In the second part of the work we will use the now established language to deal especially with the question of a categorification of the manifold notion.

ACKNOWLEDGMENTS

I want to thank Professor Dr. H. Grosse, Professor Dr. M. Reeken, and Professor Dr. E. Rosinger for discussions on the subjects involved. This work has been done during a stay at the Institute of Theoretical Physics of the University of Vienna and the Erwin Schrödinger Institute for Mathematical Physics, Vienna. The hospitality of both institutions is gratefully acknowledged. This work was made possible by a research grant under the *Gemeinsames Hochschulsonderprogramm III von Bund und Ländern* from the DAAD (German Academic Exchange Service).

[1] G. Birkhoff and J. von Neumann, ''The logic of quantum mechanics,'' Ann. Math. **37**, 823–843 (1936).
[2] R. Penrose, ''Nonlinear gravitons and curved twistor theory,'' Gen. Relativ. Gravit. **7**, 31–52 (1976).
[3] J. Bub, ''What does quantum logic explain?,'' in *Current Issues in Quantum Logic*, edited by E. G. Beltrametti and B. C. Van Fraassen (Plenum, New York, 1981), pp. 89–100.
[4] D. Deutsch, ''Quantum computation,'' Phys. World **5**, 57–61 (1992).
[5] C. J. Isham, ''An introduction to quantum gravity,'' in *Quantum Gravity*, edited by C. J. Isham, R. Penrose, and D. W. Sciama (Oxford University Press, Oxford, 1975), pp. 1–77.
[6] C. J. Isham, ''Conceptual and geometrical problems in quantum gravity,'' in *Recent Aspects of Quantum Fields*, Lecture Notes in Physics 396, edited by H. Mitter and H. Gausterer (Springer-Verlag, Berlin, 1991) pp. 123–229.
[7] R. Goldblatt, *Topoi*, 2nd ed. (North-Holland, Amsterdam, 1984).
[8] J. C. Baez and J. Dolan, ''Higher dimensional algebra and topological quantum field theory,'' J. Math. Phys. **36**, 6073–6105 (1995).
[9] L. Crane, ''Clock and category: Is quantum gravity algebraic?,'' J. Math. Phys. **36**, 6180–6193 (1995).
[10] L. Crane and I. B. Frenkel, ''Four-dimensional topological quantum field theory, Hopf categories and the canonical bases,'' J. Math. Phys. **35**, 5136–5154 (1994).
[11] K.-G. Schlesinger, ''On topology and quantum mechanics,'' J. Math. Phys. **39**, 1441–1450 (1998).
[12] G. Takeuti, ''Quantum set theory,'' in *Current Issues in Quantum Logic*, edited by E. G. Beltrametti and B. C. van Fraassen (Plenum, New York, 1981), pp. 303–322.

[13] K.-G. Schlesinger, ''The set theory of topological quantum field theories,'' preprint.
[14] M. Davis, ''A relativity principle in quantum mechanics,'' Int. J. Theor. Phys. **16**, 867 (1977).
[15] V. Chari and A. Pressley, *Quantum Groups* (Cambridge University Press, Cambridge, 1994).
[16] D. S. Freed, ''Higher algebraic structures and quantization,'' Commun. Math. Phys. **159**, 343–398 (1994).
[17] R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras* (Academic, Orlando, 1986), Vol. II.
[18] A. Connes, *Noncommutative Geometry* (Academic, Orlando, 1994).
[19] G. Landi, ''An introduction to noncommutative spaces and their geometry,'' available as hep-th/9701078.
[20] M. M. Kapranov and V. A. Voevodsky, ''2-categories and Zamolodchikov tetrahedra equations,'' in *Algebraic Groups and Their Generalizations*, Proceedings of Symposia in Pure Mathematics, edited by W. J. Haboush and B. J. Parshall (American Mathematical Society, Providence, 1994), Vol. 56, Part 2, pp. 177–259.
[21] J. C. Baez, ''Higher-dimensional algebra II: 2-Hilbert spaces,'' preprint, q-alg/9609018v2.
[22] J. W. Gray, *Formal Category Theory*, Lectures Notes in Mathematics 391 (Springer-Verlag, Berlin, 1974).
[23] J. C. Baez, ''An introduction to *n*-categories,'' preprint, q-alg/9705009.
[24] J. C. Baez and J. Dolan, ''Higher-dimensional algebra III: *n*-categories and the algebra of opetopes,'' preprint, q-alg/9702014.
[25] M. A. Batanin, ''Monoidal globular categories as a natural environment for the theory of weak *n*-categories,'' preprint.
[26] Z. Tamsamani, ''Sur de notion de *n*-categorie et *n*-groupoide non-stricte via des ensembles multi-simpliciaux,'' Thesis, Universite Paul Sabatier, Toulouse, 1996 (available as alg-geom 95-12 and 96-07).
[27] R. Street, ''The role of Michael Batanin's monoidal globular categories,'' Notes of Lectures at the *Conference on Higher Category Theory and Mathematical Physics* (Northwestern University, Evanston, IL, 1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

On the anisotropic Manev problem

Scott Craig and Florin Diacu
Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, V8W 3P4, Canada

Ernesto A. Lacomba and Ernesto Perez
Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa, Apdo. 55534, México, D.F., México

(Received 8 October 1997; accepted for publication 11 December 1998)

We consider the Manev potential, given by the sum between the inverse and the inverse square of the distance, in an anisotropic space, i.e., such that the force acts differently in each direction. Using McGehee coordinates, we blow up the collision singularity, paste a collision manifold to the phase space, study the flow on and near the collision manifold, and find a positive-measure set of collision orbits. Besides *frontal homothetic*, *frontal nonhomothetic*, and *spiraling* collisions and ejections, we put into the evidence the surprising class of *oscillatory* collision and ejection orbits. Using the infinity manifold, we further tackle capture and escape solutions in the zero-energy case. By finding the connection orbits between equilibria and/or cycles at impact and at infinity, we describe a large class of capture-collision and ejection-escape solutions.
[S0022-2488(99)01903-9]

I. INTRODUCTION

The type of anisotropic problems we tackle in this paper have been defined by Gutzwiller[1] in the 1970s to find connections between classical and quantum mechanics. Gutzwiller considered the anisotropic Kepler problem, which was later extensively analyzed by Devaney[2] and by Casasayas and Llibre.[3] Here we add to these problems the anisotropic Manev two-body problem, which we call for short the *anisotropic Manev problem*. As we will see, the system of differential equations describing it has some surprising properties, unlike any kind of isotropic or anisotropic problems studied up to now.

The name Manev (or Maneff in French and German spelling) is connected to a gravitational model defined by a potential of the form $\alpha/r+\beta/r^2$, where r is the distance between particles and $\alpha,\beta>0$ are constants.[4] But this potential goes back to Newton, who first tackled it in *Principia*. In Book I, Article IX, Proposition XLIV, Theorem XIV, Corollary 2, Newton claims that it leads to a *precesionally elliptic* orbit. He introduced this potential to explain the apsidal motion of the moon, for which he found no reasonable argument in the framework of the inverse-square force law. It seems that Newton was more interested in this type of potential than it has been previously believed. The *Portsmouth Collection* of unpublished manuscripts contains several papers, written long after the publication of *Principia*, dedicated to the understanding of this attraction force.

In terms of a central-force problem, a precessionally elliptic orbit is one in which the particle moves on an ellipse that rotates in its plane of motion. The determination of this trajectory occurs as a problem in Goldstein's *Classical Mechanics* text. A formula for the solution is easy to obtain and has been known for a long time, but its complete physical picture was only recently understood[5] by using McGehee transformations and the qualitative theory of dynamical systems. The advantage of Manev's model over the Newtonian one is that it explains the perihelion advance of the inner planets with the same accuracy as relativity.[6]

Combining Gutzwiller's anisotropy with Manev's potential, we were led to the anisotropic Manev problem, described by a nonintegrable system of differential equations. In this paper, though far from obtaining a complete picture of the global flow, we settle some local and global questions and point out the main differences between this and the anisotropic Kepler problem. The object of our endeavors is to describe the flow near collision in the general case, the main features of the global flow in the zero-energy case, and to provide the physical interpretation of the solutions we encounter.

In Sec. II we define the problem and obtain the equations of motion. Then, in Sec. III, we put into the evidence the symmetries and note that they are similar to those of the anisotropic Kepler problem. In Sec. IV we blow up the singularities by using McGehee transformations and paste to the phase-space the so-called collision manifold, which is homeomorphic to a torus. Then we find out that the flow on the collision manifold is formed only by periodic orbits, except for the eight equilibria and the eight heteroclinic orbits that connect certain equilibria (see Fig. 1). This shows an important difference between this flow and that of the Kepler problems (isotropic and nonisotropic), in which the orbits are always increasing with respect to one of the variables, giving rise to heteroclinic orbits that connect the lower and upper equilibria of the collision manifold. Such transitions vanish naturally both in the isotropic and nonisotropic Manev problems.

In Sec. V we study the flow near the collision manifold and obtain the first main result that shows a sensitive difference between the physical motion in the anisotropic Manev and Kepler problems. Using a first-return-map argument we prove that for each periodic orbit belonging to the upper (lower) part of the collision manifold, there is a local two-dimensional analytic manifold of orbits ejecting from (tending to) it. The only periodic orbit for which both types of manifolds occur is the middle one, which separates the upper and lower set of periodic orbits. Physically, these manifolds correspond to *spiraling collisions*, i.e., solutions that eject from (tend to) a binary collision such that the particles spiral around each other infinitely many times after (before) contact; for these solutions the angular momentum is different from zero. This is like a ''black-whole effect'' (or an ''inverse black-whole effect'' in case of ejections), when the bodies do not simply collide in a straightforward manner, but one is absorbed by the other towards collision as in whirlpool in whose center stands one of the bodies [see Fig. 3(b)].

Other types of collision are given by those orbits that eject from (tend to) the equilibria. For each upper (lower) equilibrium there is a local one-dimensional unstable (stable) analytic manifold outside the collision manifold. Physically, these manifolds correspond to *frontal collisions*, i.e., solutions that eject from (tend to) collision such that the orbits of the two bodies have a common tangent. But even in this class of orbits we distinguish between *homothetic* and *nonhomothetic* solutions. The homothetic ones move on straight lines, whereas the nonhomothetic ones do not [see Fig. 3(a)].

But the most interesting solutions are those that eject from (tend to) the periodic orbits around the ''bumps'' of the collision manifold. Physically they correspond to *oscillatory collisions*; the orbit oscillates with smaller and smaller amplitudes when tending to collision, without tending to a definite direction, but remaining contained in a cone [see Fig. 3(c)]. This type of motion is unlike any other one encountered in the up-to-now studied two-body problems, isotropic or anisotropic. It obviously arises due to the combination between the Manev potential and the anisotropy of the space. It would be interesting to know if other types of potentials lead to such motions in an anisotropic space.

In Sec. VI we study the flow of the zero-energy level of the phase space. To determine the asymptotic behavior of the motion at infinity we define the so-called *infinity manifold*, which extends the phase space to contain infinity. Then we see that the flows on the zero-energy manifold and on the infinity manifold are *gradientlike*, which roughly means that they increase with respect to one of the variables. We find out that the infinity manifold has also eight equilibria (see Fig. 4) and show that there exist eight homothetic orbits connecting pairwise the lower and the upper equilibria of the collision manifold and the infinity manifold, respectively (see Fig. 5). Finally we describe the main features of the flow on the zero-energy manifold. We find connecting orbits between equilibria and/or cycles and give their physical interpretation in terms of capture-collision and ejection-escape orbits.

II. EQUATIONS OF MOTION

Consider the two-degrees-of-freedom Hamiltonian system of ordinary differential equations given by

$$\begin{cases} \dot{\mathbf{q}} = \mathbf{p} \\ \dot{\mathbf{p}} = \nabla W(\mathbf{q}), \end{cases} \tag{1}$$

where $\mathbf{q}=(q_1,q_2)$ and $\mathbf{p}=(p_1,p_2)$ denote the *configuration* and the *momentum* coordinates of a physical system of two particles. W is a *quasihomogeneous anisotropic potential*, given by

$$W(\mathbf{q}) = \frac{1}{\sqrt{\mu q_1^2 + q_2^2}} + \frac{b}{\mu q_1^2 + q_2^2},$$

where $\mu>0$ and $b>0$ are parameters. For example, in the astronomical applications of the classical Manev problem the parameter b is considered very small; of the order 10^{-10}. Equations (1) define the motion of two particles of unit mass in an anisotropic space, i.e., a space in which the attraction forces act differently in every direction. The above potential defines the anisotropy of the space as a function of the parameter μ. If $\mu<1$, the attraction is the weakest in the direction of the q_1-axis and the strongest in that of the q_2-axis, the situation being reversed if $\mu>1$. If $\mu=1$, the space is isotropic and we are in the case of the classical Manev problem, whose global phase-space structure was completely described in Ref. 5; therefore we will not deal with it here. Since both remaining cases have a weakest-force and a strongest-force direction, we can assume, without loss of generality, that $\mu>1$.

The Hamiltonian function of the system (1) is given by

$$H(\mathbf{p}(t),\mathbf{q}(t)) = (1/2)\|\mathbf{p}(t)\|^2 - W(\mathbf{q}(t)),$$

the sum of the kinetic and potential energies, which yields the integral of energy

$$H(\mathbf{p}(t),\mathbf{q}(t)) = h. \tag{2}$$

However, since the force ∇W is not central, the angular momentum $L(t)=\|\mathbf{p}(t)\times\mathbf{q}(t)\|$ is *not* an integral of the system, as it is in the classical (isotropic) Manev problem (see Ref. 5).

Since $W:\mathbb{R}\backslash\{\mathbf{0}\}\to\mathbb{R}$ is real analytic, standard results of differential-equation theory guarantee, for any initial data $(\mathbf{q}(0),\mathbf{p}(0))\in(\mathbb{R}\backslash\{\mathbf{0}\})\times\mathbb{R}$, the existence and uniqueness of an analytic solution defined on a maximal interval $[0,t^*)$, where $0<t^*\leqslant\infty$. If $t^*<\infty$, the solution is said to experience a *singularity*.

A particular type of singularity, called *collision*, occurs when $\mathbf{q}(t)\to\mathbf{0}$ as $t\to t^*$. In fact, by imitating the proof used in the classical Kepler problem,[7] we can show that in the anisotropic Manev problem all singularities are collisions. Solutions leading to collisions as well as those coming close to collisions are of particular interest because the whole qualitative structure of the phase space depends on their behavior. We will study these solutions starting with Sec. IV. The next section is devoted to the study of symmetries.

III. SYMMETRIES

The symmetries in the anisotropic Manev problem are the same as in the anisotropic Kepler problem (see Ref. 3). The elements of the group $\langle S_0,S_1,S_2\rangle$, generated by S_0, S_1, and S_2, map solutions of the anisotropic Manev problem into solutions. The generating elements of this group of symmetries are given by the formulas

$$S_0(q_1,q_2,p_1,p_2,t) = (q_1,q_2,-p_1,-p_2,-t),$$

$$S_1(q_1,q_2,p_1,p_2,t) = (q_1,-q_2,-p_1,p_2,-t),$$

$$S_2(q_1,q_2,p_1,p_2,t)=(-q_1,q_2,p_1,-p_2,-t).$$

Notice that the symmetry S_0 implies the reversibility of the flow.

Invariant sets are those which remain invariant under the flow, i.e., if the initial condition is in an invariant set, then the whole solution is in this set. Like in Ref. 3, we can prove that the above group of symmetries defines exactly two invariant planes for the anisotropic Manev problem. These planes are

$$\Pi_1=\{(q_1,0,p_1,0)|(q_1,p_1)\in(\mathbb{R}\backslash\{0\})\times\mathbb{R}\},$$

$$\Pi_2=\{(0,q_2,0,p_2)|(q_2,p_2)\in(\mathbb{R}\backslash\{0\})\times\mathbb{R}\}.$$

In each of these invariant planes the flow is given by a Hamiltonian system with one degree of freedom. For Π_1 the Hamiltonian function is $H_1(q_1,p_1)=p_1^2/2-1/\sqrt{\mu}|q_1|-b/\mu q_1^2$, and for Π_2 it is $H_2(q_2,p_2)=p_2^2/2-1/|q_2|-b/q_2^2$. The qualitative structure of the flow in each of these invariant planes is the same.

The phase plane is divided in two regions by the curves $p_1^2/2-1/\sqrt{\mu}|q_1|-b/\mu q_1^2=0$ and $p_2^2/2-1/|q_2|-b/q_2^2=0$, which represent the case $h=0$ for Π_1 and Π_2, respectively, where h is the energy constant. The outside region consists of solutions with $h>0$, whereas the inside region is filled with curves representing solutions with $h<0$. Each curve in the positive-energy region follows an asymptote; this is either $p_i=\sqrt{2h}$, if the curve belongs to the half-plane $p_i>0$, or $p_i=-\sqrt{2h}$, if the curve is in the half-plane $p_i<0$.

IV. THE COLLISION MANIFOLD

In the study of collision and near-collision solutions it is helpful to transform the system (1) using a method fully developed by McGehee.[8] The idea of the method is to ''blow-up'' the collision singularity, paste instead a manifold and extend the phase space to it. Of course, this manifold is fictitious, in the sense that the flow on it does not represent orbits that have correspondent in the physical reality. However, due to the continuity of the solutions with respect to initial data, knowing the flow on the collision manifold means to have information on nearby solutions, i.e., to know what the motion looks like near collision.

We first define the transformations of the dependent variables (phase space coordinates),

$$\begin{cases} r=\|\mathbf{q}\| \\ \theta=\arctan(q_2/q_1) \\ y=\dot{r}=(q_1p_1+q_2p_2)/\|\mathbf{q}\| \\ x=r\dot{\theta}=(q_1p_2-q_2p_1)/\|\mathbf{q}\|, \end{cases}$$

and

$$\begin{cases} v=ry \\ u=rx, \end{cases}$$

and then consider a transformation of the independent variable (time),

$$d\tau=r^{-2}dt.$$

Composing these transformations, which are analytic diffeomorphisms in their respective domains, the energy relation (2) becomes

$$u^2+v^2-2r\Delta^{-1/2}-2b\Delta^{-1}=2r^2h, \tag{3}$$

and the equations of motion (1) take the form

$$\begin{cases} r' = rv \\ v' = 2r^2h + r\Delta^{-1/2} \\ \theta' = u \\ u' = (1/2)(\mu-1)(r\Delta^{-3/2} + 2b\Delta^{-2})\sin 2\theta, \end{cases} \tag{4}$$

where $\Delta = \mu\cos^2\theta + \sin^2\theta$. The new variables $(r,v,\theta,u) \in (0,\infty)\times\mathbb{R}\times S^1\times\mathbb{R}$ depend on the fictitious time τ, so the prime here denotes differentiation with respect to the new independent variable τ. Note that Eqs. (4) extend analytically to $r=0$.

The symmetries S_0, S_1, S_2 in the new coordinates are changed into $\overline{S_0}$, $\overline{S_1}$, $\overline{S_2}$, where

$$\overline{S_0}(r,v,\theta,u,\tau) = (r,-v,\theta,-u,-\tau),$$

$$\overline{S_1}(r,v,\theta,u,\tau) = (r,-v,-\theta,u,-\tau),$$

$$\overline{S_2}(r,v,\theta,u,\tau) = (r,-v,\pi-\theta,u,-\tau).$$

Notice that the sets $\{(r,v,\theta,u)|r=0\}$ and $\{(r,v,\theta,u)|r>0\}$ are invariant manifolds for the Eqs. (4). The set

$$C = \{(r,v\,\theta,u)|r=0 \text{ and the energy relation (3) holds}\}$$

is called the collision-ejection manifold or simply the collision manifold. It replaces the set of singularities $\{(\mathbf{q},\mathbf{p})|\mathbf{q}=\mathbf{0}\}$ of the original system (1), with a two-dimensional manifold in the space of the new variables. This two-manifold is embedded in $\mathbb{R}^3\times S^1$ and is given by the equations

$$r=0 \quad \text{and} \quad u^2+v^2 = 2b\Delta^{-1}. \tag{5}$$

This shows that C is homeomorphic to a torus.

From now on we will work on a fixed *constant energy surface*,

$$\mathcal{E} = \{(r,v,\theta,u)|r>0 \text{ and the energy relation (3) holds}\}.$$

The system (4) does not have singularities on $\mathcal{E}_h \cup C$. The flow on C is fictitious, in the sense that it has no physical meaning, but—due to the continuity of the solutions with respect to the initial data—its structure gives information about the behavior of the nearby flow on constant energy manifolds, that is, information about collision and near-collision solutions.

The restriction of Eqs. (4) to C yields the system

$$\begin{cases} v' = 0 \\ \theta' = u \\ u' = b(\mu-1)\Delta^{-2}\sin 2\theta. \end{cases} \tag{6}$$

Since $v'=0$, the solutions of (6) lie on the level curves $v=$constant of the torus C. There are eight equilibrium points (*equilibria*) for the system (4). In the variables (r,v,θ,u), the first four of these equilibria are $A_0^{\pm} = (0,\pm\sqrt{2b/\mu},0,0)$ and $A_\pi^{\pm}(0,\pm\sqrt{2b/\mu},\pi,0)$. At these points the linearized system has the matrix

$$\begin{bmatrix} \pm\sqrt{2b/\mu} & 0 & 0 & 0 \\ 1/\sqrt{\mu} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 2b(\mu-1)/\mu^2 & 0 \end{bmatrix},$$

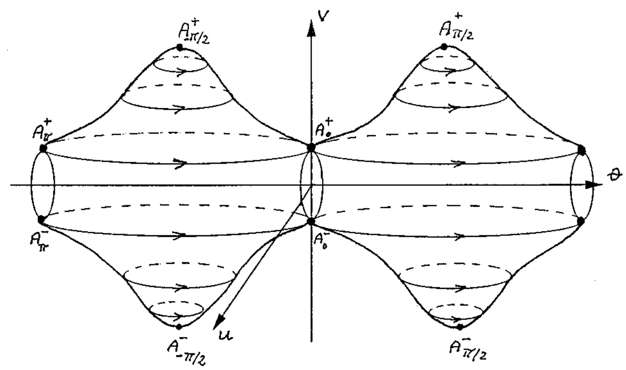

FIG. 1. The flow on the collision manifold, which shows that unlike the anisotropic Kepler problem, whose flow is gradient-like, the flow in the anisotropic Manev problem is formed by periodic orbits, four equilibria, and eight heteroclinic orbits.

the corresponding eigenvalues being real and taking the values, $\pm\sqrt{2b/\mu}$, 0, $\sqrt{2b(\mu-1)/\mu}$, and $-\sqrt{2b(\mu-1)/\mu}$. The other four equilibria are $A^{\pm}_{\pm\pi/2}=(0,\pm\sqrt{2b},\pm\pi/2,0)$ and the linearized system at these points is given by the matrix

$$\begin{bmatrix} \pm\sqrt{2b} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 2b(1-\mu) & 0 \end{bmatrix},$$

the corresponding eigenvalues being $\pm\sqrt{2b}$, 0, $\sqrt{2b(1-\mu)}$, and $-\sqrt{2b(1-\mu)}$, where the $\pm$ sign corresponds to the upper index of *A*. Since $\mu>1$, the last two eigenvalues are purely imaginary.

There are eight *heteroclinic* orbits (i.e., orbits connecting two distinct equilibria) which lie in the level sets $v=\pm\sqrt{2b/\mu}$. All the other solutions are periodic (see Fig. 1). Hence the structure of the flow on the collision manifold is fairly simple but it differs from that of the anisotropic Kepler problem (compare with Ref. 3).

Our next goal is to understand the flow and the physical behavior of the motion near the collision manifold. We will see that the flow outside the collision manifold also differs drastically from the flow of the anisotropic Kepler problem.

V. THE FLOW NEAR THE COLLISION MANIFOLD

For a fixed value of *h*, the constant-energy surface $\mathcal{E}_h$ is a three-dimensional manifold, invariant under the flow of the system (4) and whose boundary is the two-dimensional collision manifold *C*. All of this is embedded in the four-dimensional (r,θ,u,v)-space. Let us denote by P_η the periodic orbit on *C* having $v=\eta$, i.e.,

$$P_\eta=\{(r,v,\theta,u)|r=,v=\eta\}.$$

Notice that for each $\eta\in(\sqrt{2b/\mu},\sqrt{2b})\cup(-\sqrt{2b},-\sqrt{2b/\mu})$ there are two periodic orbits whose angular coordinate θ varies in different domains (see Fig. 1). However, as long as there is no danger for confusion, we will denote each of them by the same P_η.

Using this notation we can now prove the following result which summarizes the behavior close to the total collision manifold.

Theorem 5.1: *On the collision manifold C the equilibria $A^{\pm}_0$ and $A^{\pm}_\pi$ are saddles, whereas the equilibria $A^{\pm}_{\pm\pi/2}$ are centers. Outside the collision manifold the equilibria A^+_0, $A^+_{\pm\pi/2}$, and A^+_π have a one-dimensional unstable analytic manifold, whereas the equilibria A^-_0, $A^-_{\pm\pi/2}$, and A^-_π*

have a one-dimensional stable analytic manifold. Each periodic orbit P_η on C with $v=\eta>0(v=\eta<0)$ has a two-dimensional local unstable (stable) analytic manifold, while the periodic orbit $v=0$ has both a two-dimensional local unstable and a two-dimensional local stable manifold.

Physical interpretation. Before proceeding with the proof, we will give the physical interpretation of the solutions described above. As we mentioned earlier, the orbits on the collision manifold have no physical meaning, so we will deal only with those existing outside the collision manifold. The solutions tending to (ejecting from) the equilibria represent collision (ejection) orbits which have a common tangent at collision, here the limiting angular momentum of the solution is zero; we will call them *frontal collisions (ejections)* or just *frontal collisions*, for short.

It is important to distinguish here between two types of frontal collisions (ejections): the *homothetic* and the *nonhomothetic* ones. In the physical plane (q_1,q_2), the homothetic orbits move on straight lines; on the q_1 axis for $A_\pi^\pm$ and on the q_2 axis for $A_0^\pm$. The nonhomothetic ones have a different behavior. For example, there exist orbits ending at A_π^- that will pass first close to A_0^- (see Fig. 2). In physical space such an orbit comes close to a collision [see Fig. 3(a)], departs from it, then returns to a collision from the negative part of the q_1 axis, and such that the axis is tangent to the orbit at the collision point (the origin of the q_1q_2 frame). For $h=0$ this kind of solutions form a large class of collision orbits as we will see in Sec. VI.

Let us now describe the physical interpretation of orbits tending to (ejecting from) the cycles of the collision manifold. Here we have to distinguish between two classes of solutions. First are those concerning the cycles for which $v\in(-\sqrt{2b/\mu},\sqrt{2b/\mu})$. They represent collision (ejection) orbits that spiral infinitely many times without tending to a definite direction. We will call them *spiraling collisions (ejections)*. The respective angular momentum is always different from zero. In physical space they look like Fig. 3(b).

The orbits concerning cycles on the ''bumps,'' for which $v\in(-\sqrt{2b},-\sqrt{2b/\mu})$ or $v\in(\sqrt{2b/\mu},\sqrt{2b})$, have an oscillatory behavior, therefore we will call them *oscillatory collisions (ejections)*. To understand this class of orbits, consider a cycle with $v\in(-\sqrt{2b},-\sqrt{2b/\mu})$ and $\theta>0$. In Fig. 2, the intersection of the plane $Ov\theta$ (i.e., $u=0$) with this cycle corresponds to two values of the angular momentum (say θ_1 and θ_2, symmetric with respect to the q_2 axis). The angular momentum θ of an orbit tending to this cycle will have, at $u=0$, values smaller than θ_1 and larger than θ_2. Since when the particle tends to collision, r tends to 0, the physical orbit will oscillate, as in Fig. 3(c) intersecting the lines *ON* and *OM* corresponding to θ_1 and θ_2. After infinitely many oscillations outside the cone *MON*, the particle will collide with the origin. It is remarkable to mention that oscillatory collisions do not occur in any of the Manev problem, Kepler problem, or anisotropic Kepler problem, so this unintuitive type of motion is characteristic to the anisotropic Manev problem. It would be interesting to know if any other potentials lead to such collisions in an anisotropic space.

Proof of Theorem 5.1: The part of the theorem concerning the equilibria is obvious from the

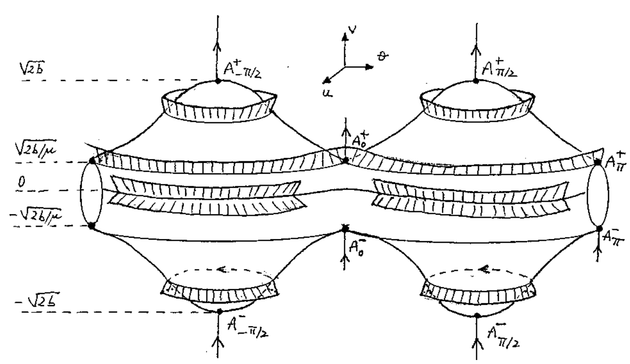

FIG. 2. The flow of the anisotropic Manev problem can reach the collision manifold at the equilibria or at any of the periodic orbits, unlike in the anisotropic Kepler problem in which the collision manifold can be reached only at equilibria.

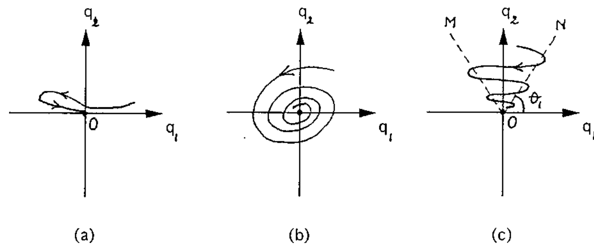

FIG. 3. Different types of collision orbits. (a) Nonhomothetic frontal collision; (b) spiralling collision; (c) oscillatory collision.

study of the eigenvalues done at the end of the previous section, so we need to deal now only with the periodic solutions P_η. For this we will distinguish between two different cases. We will first deal with the periodic orbits that go around the whole collision-manifold torus, i.e., those periodic orbits with $|\eta|<\sqrt{2b/\mu}$, and then with the orbits circling only the ''bumps'' of the collision manifold, i.e., those for which $\sqrt{2b/\mu}<|\eta|<\sqrt{2b}$. In each case we will construct the *first return map* and determine its qualitative behavior.

The first and second equations of (4) show that for small values of r the variable v is increasing. Also notice that r is decreasing in the region $v<0$ and is increasing in the region $v>0$.

We first consider the case when $0<\eta<\sqrt{2b/\mu}$. Let us fix such an η and an initial value for $\theta(0)=\theta_0$. Then, by the continuity of the flow, there is a neighborhood V of $(r,v)=(0,\eta)$ in $[0,\infty)\times\mathbb{R}$ such that every solution with initial conditions $\theta(0)=\theta_0$, $(r_0,v_0)\in V$, satisfies $\eta/2<v(t)<\frac{1}{2}(\eta+\sqrt{2b/\mu})$. For these solutions, $\theta(t)$ can be treated as the independent variable, since $\theta'=u$ and $u\neq 0$ by shrinking the neighborhood V if necessary. In this case system (4) is equivalent to the *nonautonomous* system,

$$\begin{cases} \dfrac{dr}{d\theta}=\dfrac{r'}{\theta'}=\dfrac{rv}{\sqrt{2r^2h+2r\Delta^{-1/2}+2b\Delta^{-1}-v^2}}, \\ \dfrac{dv}{d\theta}=\dfrac{v'}{\theta'}=\dfrac{2r^2h+r\Delta^{-1/2}}{\sqrt{2r^2h+2r\Delta^{-1/2}+2b\Delta^{-1}-v^2}}, \end{cases} \tag{7}$$

where $u(\theta)$ is recovered by using the energy relation (3).

The solutions to (7) form an analytic function

$$\Psi: V\times(\theta_0-\epsilon,\theta_0+2\pi+\epsilon)\to\mathbb{R}^2, \quad (r_0,v_0,\theta)\mapsto(r(\theta),v(\theta)),$$

where $r(\theta)$ and $v(\theta)$ are the solutions determined by r_0, v_0, and θ_0.

The first return map ψ is analytic and can be written as

$$\psi(r_0,v_0)=\begin{bmatrix}\psi_1(r_0,v_0)\\ \psi_2(r_0,v_0)\end{bmatrix}. \tag{8}$$

Of interest are the eigenvalues of $D\psi(r_0,v_0)$. This matrix can be calculated from the variational equations of system (7) along the periodic orbit. After some tedious computations we get the matrix

$$D\psi(0,\eta)=\begin{bmatrix}\dfrac{\partial\psi_1}{\partial r}(0,\eta) & 0\\ \dfrac{\partial\psi_2}{\partial r}(0,\eta) & 1\end{bmatrix}, \tag{9}$$

where $(\partial\psi_1/\partial r)(0,\eta)$ and $(\partial\psi_2/\partial r)(0,\eta)$ remain to be determined. The zero entry ensures that the eigenvalues of $D\psi|_{(0,\eta)}$ are 1 and $(\partial\psi_1/\partial r)(0,\eta)$. To draw the desired conclusion, we need an estimate on $(\partial\psi_1/\partial r)(0,\eta)$.

Now, the denominators in (7) are bounded on $V\times(\theta_0-\epsilon,\theta_0+2\pi+\epsilon)$, so there exists an $M>0$ such that

$$\frac{dr}{d\theta}\geqslant Mr(\theta)v(\theta)\geqslant\frac{M\eta}{2}r(\theta), \tag{10}$$

for $\theta\in[0,2\pi]$ and for all solutions $(r(\theta),v(\theta))$ starting in V. Integration with respect to θ from 0 to 2π yields the inequality

$$\psi_1(r_0,v_0)\geqslant r_0e^{M\pi v_0},$$

for all $(r_0,v_0)\in V$.

We further need the following Tauberian lemma: If $f(x)\geqslant Kx$, *where K is a constant, and* $f(0)=0$, *then* $f'(0)\geqslant K$. The proof of this lemma is obvious since $f'(0)=\lim_{h\to0^+}[f(h)-f(0)]/h=\lim_{h\to0^+}[f(h)/h]\geqslant K$.

Fix now v_0. Then, since $\psi_1(0,v_0)=0=0\cdot e^{M\pi v_0}$, by the above Tauberian lemma the slope of the curve $r_0\mapsto\psi_1(r_0,v_0)$ at $r_0=0$ is greater than that of $r_0\mapsto e^{M\pi v_0}\cdot r_0$. This is equivalent to

$$\frac{\partial}{\partial r}\psi_1\bigg|_{(0,v_0)}\geqslant e^{M\pi v_0}>1, \tag{11}$$

the last inequality being true for all $(r,v)\in V$. Thus, the periodic orbit P_η has a two-dimensional analytical unstable manifold of orbits that eject outside C. The analyticity of these invariant submanifolds follows if using a recent result by Cabré and Fontich (Theorem 4.1 in Ref. 9). This shows that the limiting behavior at collision depends analytically on the initial conditions. With a similar argument we can prove that for $-\sqrt{2b/\mu}<\eta<0$,

$$\frac{\partial\psi_1}{\partial r}(0,\eta)\leqslant e^{M\eta\pi}<1, \tag{12}$$

and so the corresponding periodic orbit P_η has a stable analytic manifold of orbits approaching it. These orbits represent solutions in which a collision (or ejection) occurs as the particles spin around each other, in contrast to the classical Newtonian case in which collisions are frontal, i.e., the particles asymptotically approach each other radially, following ultimately a definite direction.

We will now see what happens in the case $\eta=0$. For this we will apply a generalization of a theorem due to Casasayas, Fontich, and Nunes.[10]

Let $F=(F_1,F_2)$ *be an analytic function from a neighborhood of* $(0,0)$ *in* $\mathbb{R}^2$ *to* $\mathbb{R}^2$, *such that*

(*i*) $F(0,v)=(0,v)$,
(*ii*) $DF(0,0)=\begin{pmatrix}1&0\\c&1\end{pmatrix}$ *with* $c>0$,
(*iii*) $\alpha=D_rD_vF_1(0,0)>0$.

Then there exists stable and unstable manifolds of $(0,0)$ *which are, locally, graphs of analytic functions, that is,* $W^s_{loc}(\delta)=\{(\varphi^s(v),v)|v\in(-\delta,0)\}$ *and* $W^u_{loc}(\delta)=\{(\varphi^u(v),v)|v\in(0,\delta)\}$, *where* $\varphi^{s,u}\sim(\alpha/2c)v^2$. (By W^s_{loc} and W^u_{loc} we have denoted the *local stable manifold* and the *local unstable manifold*, respectively.) At $v=0$ *the functions* $\varphi^{s,u}$ *are only Lipschitz in general*.

The proof of this result follows identically the one in Ref. 10, the only difference being that in that paper the value of the constant c equals 1. This more general result is obtained by a linear transformation of coordinates. We will prove now that the conditions of this theorem are fulfilled by our first return map ψ, which is also defined by $v_0=0$ with a conveniently chosen domain.

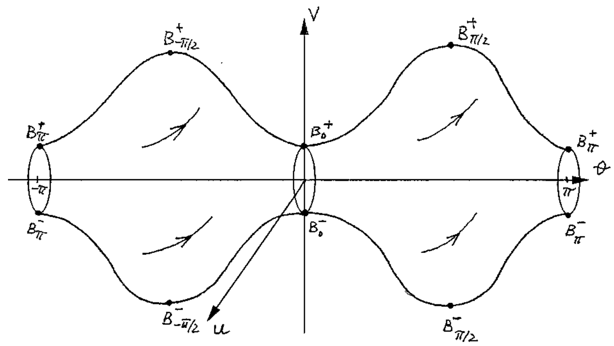

FIG. 4. The flow on the infinity manifold.

Now, the solutions starting at $(r,v)=(0,v_0)$ are periodic, so (i) is clearly satisfied. Straightforward computations in the variational equations of (7) show that

$$c=\frac{\partial}{\partial r}\psi_2\bigg|_{(0,0)}>0,$$

and from (11) and (12) we know that

$$\lim_{v_0\to 0^+}\frac{\partial}{\partial r}\psi_1\bigg|_{(0,v_0)}\geqslant 1$$

and that

$$\lim_{v_0\to 0^-}\frac{\partial}{\partial r}\psi_1\bigg|_{(0,v_0)}\leqslant 1.$$

The analyticity of ψ_1 implies that the common limit should be 1, so (ii) is fulfilled. For (iii) we have

$$\alpha=D_rD_v\psi_1(0,0)=\lim_{v_0\to 0^+}\frac{D_r\psi_1(0,v_0)-D_r\psi_1(0,-v_0)}{2v_0}\geqslant\lim_{v_0\to 0^+}\frac{e^{M\pi v_0}-e^{-M\pi v_0}}{2v_0}=M\pi>0.$$

The second equality follows again by the analyticity of ψ_1. This means that each of the two periodic orbits P_η with $\eta=0$ have both a stable manifold of approaching orbits and an unstable manifold of ejecting orbits.

To complete the proof, let us now describe the flow near a periodic solution P_η with $\sqrt{2b/\mu}<\eta<\sqrt{2b}$ and $\theta>0$. (The case with $\theta<0$ or $-\sqrt{2b}<\eta<-\sqrt{2b/\mu}$ is similar.) We start by shifting the origin of the frame from 0 to $\pi/2$ and change the variable θ to a variable ϕ. When moving from 0 to 2π, the new angular variable ϕ rotates around the v-axis, which now goes through $A^+_{\pi/2}$ instead of A_0 (see Fig. 4), such that it allows us to describe only the periodic orbits P_η with $v>\sqrt{2b/\mu}$ and $\theta>0$. This is done by defining the transformation

$$\begin{cases} u=w\sin\phi,\\ \theta=\dfrac{\pi}{2}+w\cos\phi. \end{cases}$$

With this transformation the system (4) changes to

$$\begin{cases} r' = rv, \\ v' = 2r^2 h + r\tilde{\Delta}^{-1/2}, \\ \phi' = -\sin^2\phi + (1/2)(\mu-1)w^{-1}(r\tilde{\Delta}^{-3/2} + 2b\tilde{\Delta}^{-2})\cos\phi\sin(2w\cos\phi), \\ w' = (1/2)(\mu-1)(r\tilde{\Delta}^{-3/2} + 2b\tilde{\Delta}^{-2})\sin\phi\sin(2w\cos\phi) + w\sin\phi\cos\phi, \end{cases} \tag{13}$$

and the energy relation becomes

$$w^2\sin^2\phi + v^2 - 2r\tilde{\Delta}^{-1/2} - 2b\tilde{\Delta}^{-1} = 2r^2 h, \tag{14}$$

where $\tilde{\Delta} = \mu\sin^2(w\cos\phi) + \cos^2(w\cos\phi)$.

Now, proceed as in the other case. In an appropriate neighborhood of the periodic orbit, we have $\phi' \neq 0$, so the nonautonomous system

$$\begin{cases} \dfrac{dr}{d\phi} = \dfrac{r'}{\phi'}, \\ \dfrac{dw}{d\phi} = \dfrac{v'}{\phi'}, \end{cases}$$

is analytic in that neighborhood, where the energy relation (14) is used to recover w (the positive root). Using the same methods as before, the matrix

$$D\Phi(0,\eta) = \begin{bmatrix} \dfrac{\partial\Phi_1}{\partial r}(0,\eta) & 0 \\ \dfrac{\partial\Phi_2}{\partial r}(0,\eta) & 1 \end{bmatrix} \tag{15}$$

is the derivative of the Poincaré map $\Phi(r,v)$ on the section $\phi=0$. The right column is the same as in (9) because again r' and v' are multiples of r. Since ϕ' is bounded, the same inequalities as the ones used for $(dr/d\theta)$ show that $(\partial\Phi_1/\partial r)(0,\eta) > 1$ for $\sqrt{2b} < \eta < \sqrt{2b/\mu}$. So each periodic solution with $\sqrt{2b} < \eta < \sqrt{2b/\mu}$ has an unstable manifold of ejecting solutions. Similarly each solution with $-\sqrt{2b/\mu} < \eta < -\sqrt{2b}$ has a stable manifold of approaching solutions, in which the angular momentum oscillates, and the particles stay within an acute angle from the weak axis of the force. This completes the proof of the theorem.

An obvious consequence of Theorem 5.1 is the following:

Corollary 5.2: The set of initial data leading to collisions has positive measure. More precisely, the set of initial data leading to frontal collisions has zero measure, whereas each of the sets of initial data leading to spiraling and oscillatory collisions, respectively, have positive measure.

We will further consider the zero-energy manifold and in what remains of this paper will give a qualitative description of the flow in this particular case.

VI. THE ZERO-ENERGY MANIFOLD

In this section we will study the flow on $\mathcal{E}_0$, i.e., the case of the zero energy level, $h=0$, and compare the flow on the collision manifold with the one on the so-called *infinity manifold*, which we define below. In order to understand the global flow we will analyze the invariant submanifolds associated to the equilibrium points and to the periodic orbits on the collision manifold, as well as their connection orbits.

A. Infinity manifold and homothetic orbits

Let us start by describing a characteristic property of the flow on $\mathcal{E}_0$. For this, we need the following definition:[11]

Definition 6.1: A flow is called *gradientlike* with respect to one of the coordinates, if every nonequilibrium solution increases on that coordinate.

With this definition we can now prove the following:

Lemma 6.2: The flow on $\mathcal{E}_0$ is gradientlike with respect to the v-coordinate.

Proof: For $h=0$, the second equation in (4) shows that $v'>0$ for all values of θ.

To study the asymptotic behavior at infinity, we will apply a suitable blow-up transformation. Since the potential is a *quasihomogeneous function* (i.e., the sum of homogeneous functions of different degrees[12]), this transformation is slightly different from the one used in the case of the collision. This is because the term of degree -1 predominates when $t\to\infty$.

Taking $h=0$ and $\rho=1/r$, Eqs. (4) become

$$\begin{cases} \rho' = -\rho v, \\ v' = \rho^{-1}\Delta^{-1/2}, \\ \theta' = u, \\ u' = [(\mu-1)/2](\rho^{-1}\Delta^{-3/2} + 2b\Delta^{-2})\sin 2\theta, \end{cases} \tag{16}$$

and the energy relation takes the form

$$\rho(u^2+v^2) - 2\Delta^{-1/2} - 2b\rho\Delta^{-1} = 0. \tag{17}$$

Rescaling the velocities by using the transformations $\overline{v}=\rho^{1/2}v$, $\overline{u}=\rho^{1/2}u$, and rescaling the (independent) time variable by defining the transformation $d\tau=\rho^{1/2}ds$, Eqs. (16) take the form

$$\begin{cases} \dot{\rho} = -\rho\overline{v}, \\ \dot{\overline{v}} = -(1/2)\overline{v}^2 + \Delta^{-1/2}, \\ \dot{\theta} = \overline{u}, \\ \dot{\overline{u}} = -(1/2)\overline{v}\,\overline{u} + [(\mu-1)/2](\Delta^{-3/2} + 2b\rho\Delta^{-2})\sin 2\theta, \end{cases} \tag{18}$$

where the dot denotes differentiation with respect to the new (fictitious) time variable s. In the new coordinates the energy relation becomes

$$\overline{u}^2 + \overline{v}^2 - 2\Delta^{-1/2} - 2b\rho\Delta^{-1} = 0. \tag{19}$$

Analogously to the collision manifold, we define the *infinity manifold* I,

$$I\{(\rho,\overline{v},\theta,\overline{u}) \mid \rho=0 \text{ and } \overline{u}^2+\overline{v}^2 = 2\Delta^{-1/2}\},$$

which is also homeomorphic to a torus (see Fig. 4).

The flow on I is given by

$$\begin{cases} \dot{\overline{v}} = (1/2)\overline{u}^2, \\ \dot{\theta} = \overline{u}, \\ \dot{\overline{u}} = -(1/2)\overline{v}\,\overline{u} + [(\mu-1)/2]\Delta^{-3/2}\sin 2\theta. \end{cases} \tag{20}$$

As in the case of the collision manifold, the flow on the infinity manifold is fictitious in the sense that it has no physical meaning. But once again, using the continuity of the solutions with respect to the initial data, the structure of the flow on the infinity manifold will allow us to draw conclusions about the behavior of the flow near infinity. Therefore, let us now study the flow on I.

There are eight equilibrium points on the infinity manifold I, which, in the new $(\overline{v},\theta,\overline{u})$-coordinates, have the form $B_0^{\pm}=(\pm\sqrt{2\mu^{-1/2}},0,0)$, $B_\pi^{\pm}=(\pm\sqrt{2\mu^{-1/2}},\pi,0)$, and $B_{\pm\pi/2}^{\pm}=(\pm\sqrt{2},\pm\pi/2,0)$. We can now show that the flow on I is fairly simple.

Lemma 6.3: The flow on I is gradient like with respect to the $\overline{v}$-coordinate.

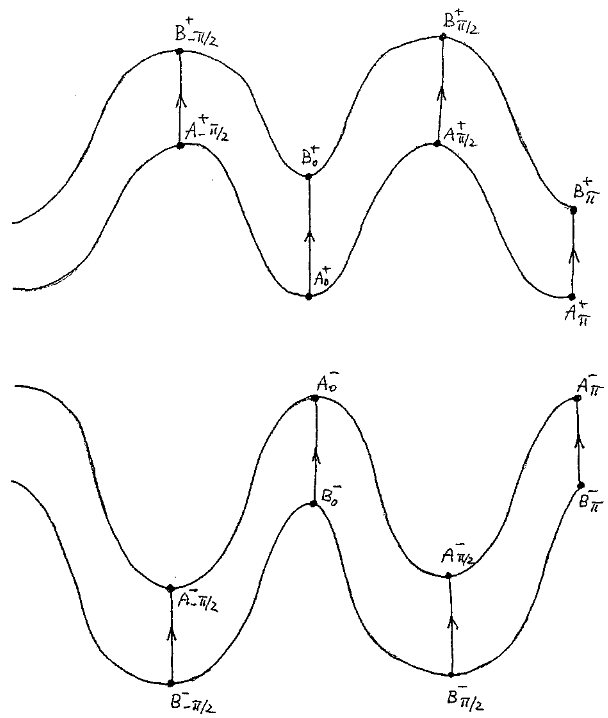

FIG. 5. Homothetic orbits connecting the collision and the infinity manifolds.

Proof: From the first equation of system (20) we see that $\dot{\bar{v}}\geqslant 0$. If $\dot{\bar{v}}=0$, then $\bar{u}=0$, so $\ddot{\bar{v}}=\bar{u}\dot{\bar{u}}=0$. But $\dddot{\bar{v}}=\dot{\bar{u}}^2+\bar{u}\ddot{\bar{u}}=\dot{\bar{u}}^2$, which is 0 only at the equilibria and is positive otherwise. This completes the proof.

We now define the notion of a *central configuration*, a concept that plays an important role both in the Newtonian case as well as in the one considered here.

Definition 6.4: The configuration q_0 is called central if $\nabla W(q_0)$ is parallel to q_0. An orbit such that the position is a homothety to q_0 is called a homothetic orbit.

The following proposition shows the existence of homothetic orbits in the anisotropic Manev problem.

Proposition 6.5: There exist eight orbits connecting the respective equilibrium points on the collision manifold C to the ones on the infinity manifold I.

Proof: Recall that the equilibrium points are defined in each of the blow-up coordinate systems, at collision and at infinity. This defines two different charts, carrying in each case the corresponding time scales for the differential equations (see Fig. 5).

In the chart containing the infinity manifold *I*, the homothetic orbits are given by the equations,

$$\dot{\rho}=-\rho\bar{v},\quad \dot{\bar{v}}=-\tfrac{1}{2}\bar{v}^2+\mu^{-(1/2)},\quad \text{if}\quad \theta=0\quad \text{or}\quad \theta=\pi,\quad \bar{u}=0,$$

$$\dot{\rho}=-\rho\bar{v},\quad \dot{\bar{v}}=-\tfrac{1}{2}\bar{v}^2+1,\quad \text{if}\quad \theta=\pm\pi/2,\quad \bar{u}=0.$$

In the chart containing the collision manifold *C*, the homothetic orbits are given by the equations

$$r'=rv,\quad v'=v^2-\mu^{-(1/2)}r-2b\mu^{-1},\quad \text{if}\quad \theta=0\quad \text{or}\quad \theta=\pi,\quad u=0,$$

$$r'=rv,\quad v'=v^2-r-2b,\quad \text{if}\quad \theta=\pm\pi/2,\quad u=0.$$

Using the energy relation and correspondingly changing the time scales, by straightforward computation we see that the set of eight orbits described in the infinity-manifold chart is identical to the eight orbits described in the collision-manifold chart. This completes the proof.

Remark: Notice that the α- and ω-limits of the four homothetic orbits described above are equilibrium points on C and I.

To describe the global flow in the zero-energy case, notice that, on $\mathcal{E}_0$, the flow is gradient-like with respect to the v coordinate, but it is gradient-like with respect to $\overline{v}$ only on the infinity manifold I, not on $\mathcal{E}_0$. Moreover, the so-called *zero velocity curve* is empty. This is easy to see, for the zero velocity curve is defined as the set of phase-space points for which the momentum coordinate p is zero in the energy relation (2). In McGehee coordinates this corresponds to taking $v=0$ and $u=0$ in the energy relation (3).

Recall from Theorem 5.1 that for each $\eta \neq \pm\sqrt{2b/\mu}$, the periodic orbit P_η has, at least locally, a two-dimensional stable (unstable) manifold if $\eta<0$ ($\eta>0$); if $\eta=0$, then P_0 has both a stable and an unstable two-dimensional submanifold. Also recall that, without loss of generality, we can take $\mu>1$.

Due to the gradient-like structure of the global flow on $\mathcal{E}_0$, the invariant manifolds corresponding to periodic orbits P_η with $\eta\leqslant 0$ cannot intersect invariant manifolds corresponding to periodic orbits P_η with $\eta\geqslant 0$.

B. The local structure

Before going deeper into the global structure of the flow on $\mathcal{E}_0$, we have to analyze the hyperbolic character of the equilibrium points. The computations are set in McGehee coordinates for the equilibrium points belonging to the total collision manifold C and in infinity-blow-up coordinates for those belonging to the infinity manifold I.

We begin with McGehee coordinates. From the energy relation (3), for $h=0$ we obtain

$$r=(1/2)(u^2+v^2)\Delta^{1/2}-b\Delta^{-1/2}.$$

Substituting r in the equations of motion (4) in McGehee coordinates, we obtain the system

$$\begin{cases} v'=(1/2)(u^2+v^2)-b\Delta^{-1}, \\ \theta'=u, \\ u'=[(\mu-1)/4][(u^2+v^2)\Delta^{-1}+2b\Delta^{-2}]\sin 2\theta. \end{cases} \tag{21}$$

The matrix of the attached linear system of variables (v,θ,u), at the equilibrium points $A_0^\pm=(\pm\sqrt{2b/\mu},0,0)$ and $A_\pi^\pm=(\pm\sqrt{2b/\mu},\pi,0)$, is

$$\begin{bmatrix} \pm\sqrt{2b/\mu} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 2b(\mu-1)/\mu^2 & 0 \end{bmatrix},$$

and has the eigenvalues $\lambda_1=\pm\sqrt{2b/\mu}$, $\lambda_2=\sqrt{2b(\mu-1)}/\mu$, $\lambda_3=-\sqrt{2b(\mu-1)}/\mu$, which, since $\mu>1$, shows that these equilibrium points are hyperbolic. The equilibria A_0^- and A_π^- have a two-dimensional stable manifold and a one-dimensional unstable manifold, whereas the equilibria A_0^+ and A_π^+ have a one-dimensional stable manifold and a two-dimensional unstable manifold.

The matrix of the attached linear system of variables (v,θ,u), at the equilibrium points $A^\pm_{\pm\pi/2}=(\pm\sqrt{2b},\pm\pi/2,0)$ is

$$\begin{bmatrix} \pm\sqrt{2b/\mu} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -2b(\mu-1) & 0 \end{bmatrix},$$

with eigenvalues $\lambda_1=\pm\sqrt{2b}$, $\lambda_2=i\sqrt{2b(\mu-1)}$, and $\lambda_3=-i\sqrt{2b(\mu-1)}$. This is in agreement with the fact that these equilibria are α- or ω-limits of the corresponding homothetic orbits and that the structure of the flow restricted to the collision manifold proves them to be centers. So these equilibrium points are *not* hyperbolic.

We pass now to the infinity-blow-up coordinates. For $h=0$, the corresponding energy relation (19) is equivalent to

$$2\rho b=(\overline{u}^2+\overline{v}^2)\Delta-2\Delta^{1/2}.$$

Substitution of ρ in Eqs. (18) gives

$$\begin{cases} \dot{\overline{v}}=-(1/2)\overline{v}^2+\Delta^{-1/2}, \\ \dot{\theta}=\overline{u}, \\ \dot{\overline{u}}=-(1/2)\overline{v}\,\overline{u}+[(\mu-1)/2][(\overline{v}^2+\overline{u}^2)\Delta^{-1}-\Delta^{-3/2}]\sin 2\theta. \end{cases} \tag{22}$$

The matrix of the attached linear system of variables $(\overline{v},\theta,\overline{u})$ at the equilibrium points $B_0^{\pm}=(\pm\sqrt{2\mu^{-1/2}},0,0)$ and $B_\pi^{\pm}=(\pm\sqrt{2\mu^{-1/2}},\pi,0)$ is

$$\begin{bmatrix} \mp\sqrt{2\mu^{-1/2}} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & (\mu-1)/\mu^{3/2} & \mp(1/2)\sqrt{2\mu^{-1/2}} \end{bmatrix}.$$

The eigenvalues of this matrix are

$$\lambda_1=\mp\sqrt{2\mu^{-1/2}},$$

$$\lambda_2=\mp(4\mu)^{-1/4}/2+(1/2)\sqrt{(4\mu)^{-1/2}+4(\mu-1)\mu^{-3/2}},$$

$$\lambda_3=\mp(4\mu)^{-1/4}/2-(1/2)\sqrt{(4\mu)^{-1/2}+4(\mu-1)\mu^{-3/2}},$$

so the equilibria are hyperbolic. The equilibria B_0^- and B_π^- have a one-dimensional stable manifold and a two-dimensional unstable manifold, whereas the equilibria B_0^+ and B_π^+ have a two-dimensional stable manifold and a one-dimensional unstable manifold.

Finally, the corresponding matrix of the linear system in variables $(\overline{v},\theta,\overline{u})$, at the equilibria $B_{\pm\pi/2}^{\pm}=(\pm\sqrt{2},\pm\pi/2,0)$, is

$$\begin{bmatrix} \mp\sqrt{2} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1-\mu & \mp\sqrt{2}/2 \end{bmatrix},$$

with eigenvalues

$$\lambda_1=\mp\sqrt{2},$$

$$\lambda_2=\mp\sqrt{2}/4+(1/2)\sqrt{9/2-4\mu},$$

$$\lambda_3=\mp\sqrt{2}/4-(1/2)\sqrt{9/2-4\mu},$$

the sign $\mp$ being chosen with respect to the upper $\pm$ sign of B. This shows that the equilibria are also hyperbolic. In fact $B_{\pm\pi/2}^{+}$ are sinks, whereas $B_{\pm\pi/2}^{-}$ are sources. Like in the previous case, the eigenvalue λ_1 corresponds to the homothetic orbit, while the others correspond to the flow restricted to the infinity manifold.

Thus, the behavior of our flow for $\rho=0$ is in agreement with what we know about the flow on the total collision manifold for the anisotropic Kepler problem (see Ref. 3).

Now that the local structure of the flow on $\mathcal{E}_0$ at the equilibria is understood, we can go further towards enhancing the global picture of the zero-energy case.

C. Qualitative aspects of the global flow

Using all the above information, we are now in a position to state and prove the main result in this section, which describes connecting orbits on $\mathcal{E}_0$, offering the main qualitative features of the flow on the zero-energy manifold. More precisely, we find all possible connections between periodic orbits on C and equilibrium points.

Theorem 6.6: *All the orbits tending to every lower cycle of the collision manifold (including the median one), must eject from a lower equilibrium of the infinity manifold; all the orbits ejecting from every upper cycle of the collision manifold (including the median one) must tend to an upper equilibrium of the infinity manifold. There does not exist orbits in $\mathcal{E}_0$ connecting cycles, or cycles and equilibria, of the collision manifold. There does exist noncollision orbits in $\mathcal{E}_0$ connecting lower and upper equilibria of the infinity manifold*.

Physical interpretation. Any zero-energy solution tending to a lower cycle must eject from a lower equilibrium of the infinity manifold. These solutions, called *capture-collision* orbits, are unbounded at time $-\infty$ and end in a spiraling collision in finite time. Symmetrically, solutions ejecting form an upper cycle must tend to an upper equilibrium of the infinity manifold. They are called *ejection-escape* orbits, start from a spiraling ejection at a finite time and become unbounded at time $+\infty$. The next statement tells that there does not exist solutions that start from a collision (spiraling or frontal) at a finite time and end in a collision (spiraling or frontal) at a later finite time. The last sentence states the existence of collisionless orbits that are unbounded at times $-\infty$ and $+\infty$.

Proof: Recall that we have obtained a three-dimensional flow by eliminating the coordinate r from the energy relation and that the stable manifolds of the periodic orbits P_η are contained in this compact three-dimensional manifold with boundary $C\cup\mathcal{E}_0\cup I$. We will prove the existence of connecting orbits between each cycle P_η of the collision manifold and equilibria of the infinity manifold. For this let P_η, $\eta\leqslant 0$, be a cycle and take any orbit belonging to the two-dimensional local manifold of orbits tending to P_η. Since the flow is gradient-like with respect to the variable v on $\mathcal{E}_0\cup I$ (outside the collision manifold), and there are no other equilibria or cycles below P_η having an unstable manifold of positive dimension, the chosen orbit must come from one of the lower equilibria of the infinity manifold. Using the reversibility of the flow, we can prove that any orbit starting asymptotically at a cycle P_η, $\eta\geqslant 0$, connects with the upper equilibria of the infinity manifold.

The nonexistence of orbits connecting cycles of the collision manifold and of orbits connecting cycles with equilibria of the collision manifold, follows again from the gradient-like property of the flow and the nonexistence of unstable manifolds for the lower equilibria and the cycles P_η with $\eta<0$ and from the nonexistence of stable manifolds for the upper equilibria and the cycles P_η with $\eta>0$. Also, because of the gradient-like property, there are no homoclinic connections for the cycles or the equilibria.

To prove the existence of orbits connecting lower and upper equilibria of the infinity manifold, take an initial condition in the plane $v=0$ of $\mathcal{E}_0$, close to, but outside the collision manifold. Due to the gradient-like property and the fact that the upper equilibria and cycles of the collision manifold have no stable manifolds, the corresponding orbit has to end at one of the upper equilibria of the infinity manifold. Using the same arguments, we can check that this orbit starts at one of the lower equilibria at infinity. This completes the proof of the theorem.

We have thus described the main features of the flow in the zero-energy case. Unfortunately at this point we have only a vague understanding about the behavior of orbits coming from or tending to heteroclinic connections, so we still miss a complete foliation of the global flow on the zero-energy manifold.

ACKNOWLEDGMENTS

Scott Craig did part of this research as a Master's student at the University of Victoria under the supervision of Florin Diacu.

Florin Diacu was supported in part by the NSERC Grant No. OGP0122045 of Canada. Ernesto Lacomba and Ernesto Perez received partial support from the CONACYT Grant No. 1772-E9210 of Mexico. Ernesto Lacomba was also partially supported by the sabbatical fellowship SAB95-0252 of the DGICYT of Spain.

[1] M. C. Gutzwiller, ''Periodic orbits and classical quantization conditions,'' J. Math. Phys. **12**, 343–358 (1971); ''The anisotropic Kepler problem in two dimensions,'' *ibid.* **14**, 139–152 (1973); ''Bernoulli sequences and trajectories in the anisotropic Kepler problem,'' *ibid.* **18**, 806–823 (1977).

[2] R. L. Devaney, ''Collision orbits in the anisotropic Kepler problem,'' Invent. Math. **45**, 221–251 (1978); ''Singularities in classical mechanical systems,'' in Ergodic Theory and Dynamical Systems I (Birkhäuser, Boston, 1981), pp. 211–333.

[3] J. Casasayas and J. Llibre, ''Qualitative analysis of the anisotropic Kepler problem,'' Memoirs of the American Mathematical Society (American Mathematical Society, Providence, RI, 1984), Vol. 52, No. 312.

[4] G. Maneff, ''La gravitation et le principe de l'égalité de l'action et de la réaction,'' Comptes Rendus **178**, 2159–2161 (1924); ''Die Gravitation und das Prinzip von Wirkung und Gegenwirkung,'' Z. Phys. **31**, 786–802 (1925); ''Le principe de la moindre action et la gravitation,'' Comptes Rendus **190**, 963–965 (1930); ''La gravitation et l'énergie au zéro,'' *ibid.* **190**, 1374–1377 (1930).

[5] F. N. Diacu, A. Mingarelli, V. Mioc, and C. Stoica, ''The Manev two-body problem: Quantitative and qualitative theory,'' in *Dynamical Systems and Applications*, World Scientific Series in Applicable Analysis, edited by R. P. Agarwal (World Scientific, Singapore, 1995), Vol. 4, pp. 213–227; J. Delgado, F. N. Diacu, E. A. Lacomba, A. Mingarelli, V. Mioc, E. Perez, and C. Stoica, ''The global flow of the Manev problem,'' J. Math. Phys. **37**, 2748–2761 (1996); F. N. Diacu, V. Mioc, and C. Stoica, ''Phase-space structure and regularization of Manev-type problems,'' Nonlinear Analysis (to appear).

[6] Y. Hagihara, *Celestial Mechanics* (MIT Press, Cambridge, MA, 1975), Vol. II, Part I.

[7] A. Wintner, *The Analytical Foundations of Celestial Mechanics* (Princeton University Press, Princeton, NJ, 1941).

[8] R. McGehee, ''Triple collision in the collinear three-body problem,'' Invent. Math. **27**, 191–227 (1974).

[9] X. Cabré and E. Fontich, ''Regularity and uniqueness of one dimensional invariant manifolds'' (preprint, 1997).

[10] J. Casasayas, E. Fontich, and A. Nunes, ''Invariant manifolds for a class of parabolic points,'' Nonlinearity **5**, 1193–1210 (1992).

[11] F. N. Diacu, *Singularities of the N-Body Problem: An Introduction to Celestial Mechanics* (Les Publications CRM, Montréal, 1992).

[12] F. N. Diacu, ''Near-collision dynamics for particle systems with quasihomogeneous potentials,'' J. Diff. Eqns. **128**, 58–77 (1996).

Nonholonomic constraints in time-dependent mechanics

G. Giachetta and L. Mangiarotti[a)]
Department of Mathematics and Physics, University of Camerino, 62032 Camerino (MC), Italy

G. Sardanashvily[b)]
Department of Theoretical Physics, Moscow State University, 117234 Moscow, Russia

(Received 21 July 1998; accepted for publication 20 November 1998)

The constraint reaction force of ideal nonholonomic constraints in time-dependent mechanics on a configuration bundle $Q\to\mathbf{R}$ is obtained. Using the vertical extension of Hamiltonian formalism to the vertical tangent bundle VQ of $Q\to\mathbf{R}$, the Hamiltonian of a nonholonomic constrained system is constructed. The present setting is more general than the one usually considered in the literature on nonholonomic mechanics. © *1999 American Institute of Physics.*
[S0022-2488(99)01403-6]

I. INTRODUCTION

This work addresses the geometric theory of nonholonomic constraints in time-dependent mechanics. We refer the reader to Refs. 1–7 for the autonomous case. We follow the approach based on the D'Alembert principle because the variational methods with Lagrange multipliers are not always appropriate to nonholonomic constraints (see Refs. 2, 5, 6, and 8).

Let the jet manifold J^1Q be a velocity phase space of time-dependent mechanics on a configuration bundle $Q\to\mathbf{R}$. When studying mechanical systems with constraints, one usually represents nonholonomic constraints as distributions on Q or submanifolds of the jet manifold J^1Q.[8–10] In this paper the notion of nonholonomic constraint is generalized in such a way to include codistributions $\mathbf{S}$ or, accordingly, distributions $\mathrm{Ann}(\mathbf{S})$ on the jet manifold J^1Q.[11,12] We study the following problem. Let ξ be a second-order dynamic equation on Q and $\mathbf{S}$ a codistribution on J^1Q whose annihilator $\mathrm{Ann}(\mathbf{S})$ is treated as a nonholonomic constraint. The goal is to find a decomposition,

$$\xi=\tilde{\xi}+r, \tag{1}$$

where $\tilde{\xi}$ is a second-order dynamic equation obeying the condition

$$\tilde{\xi}\subset\mathrm{Ann}(\mathbf{S}). \tag{2}$$

One can think of $\tilde{\xi}$ as describing a mechanical system subject to the nonholonomic constraint $\mathbf{S}$, while $(-r)$ is the constraint reaction acceleration. The decomposition (1), however, is not unique. In the case of Newtonian systems, including nondegenerate Lagrangian systems, we obtain the decomposition (1) which satisfies the D'Alembert principle for ideal nonholonomic constraints. We construct the Hamiltonian counterpart of the constrained equation of motion (2). We show that this can be seen as Hamilton equations in the framework of the vertical extension of Hamiltonian formalism to the configuration space VQ which is the vertical tangent bundle of $Q\to\mathbf{R}$. This may be a step towards the functional integral formulation of nonholonomic time-dependent mechanics and its further quantization.

[a)]Electronic mail: mangiaro@camserv.unicam.it
[b)]Electronic mail: sard@grav.phys.msu.su

II. GEOMETRIC INTERLUDE

All manifolds throughout the paper are real, finite dimensional, second countable (hence, paracompact), and connected.

We refer the reader to Refs. 8, 9, and 11–15 for the geometric formulation of Lagrangian and Hamiltonian time-dependent mechanics. In accordance with this formulation, a configuration space of time-dependent mechanics is an $(m+1)$-dimensional fiber bundle $Q\to\mathbf{R}$, coordinated by (t,q^i). Its base $\mathbf{R}$ is treated as a time axis provided with the Cartesian coordinate t. With this coordinate, $\mathbf{R}$ is equipped with the standard vector field ∂_t and the standard one-form dt. For the sake of convenience, we will also utilize the compact notation q^λ, where $q^0=t$. Obviously, any fiber bundle $Q\to\mathbf{R}$ is trivial, but it cannot be canonically identified to a product $\mathbf{R}\times M$ in general. Different trivializations $Q\cong\mathbf{R}\times M$ correspond to different reference frames.

The velocity phase space of time-dependent mechanics is the first-order jet manifold J^1Q of $Q\to\mathbf{R}$, coordinated by (t,q^i,q^i_t). There is the canonical imbedding,

$$\lambda: J^1Q\hookrightarrow TQ, \quad (t,q^i,q^i_t)\mapsto(t,q^i,\dot t=1,\dot q^i=q^i_t), \tag{3}$$

of J^1Q onto the affine subbundle of the tangent bundle TQ of Q which is modeled over the vertical tangent bundle VQ of $Q\to\mathbf{R}$. From now on we will identify the jet manifold J^1Q with its image in TQ.

Similarly, we have the imbeddings,

$$J^2Q\hookrightarrow J^1J^1Q\hookrightarrow TJ^1Q,$$

$$(t,q^i,q^i_t,q^i_{tt})\mapsto(t,q^i,q^i_t,\dot t=1,\dot q^i=q^i_t,\dot q^i_t=q^i_{tt}),$$

where J^2Q, coordinated by $(q^\lambda,q^i_t,q^i_{tt})$, is the second-order jet manifold of the fiber bundle $Q\to\mathbf{R}$. The affine bundle $J^2Q\to J^1Q$ is modeled over the vertical tangent bundle,

$$V_QJ^1Q\cong J^1Q\underset{Q}{\times}VQ, \tag{4}$$

of the affine jet bundle $J^1Q\to Q$.

The jet manifold J^1Q is provided with the canonical tangent-valued form,

$$\hat v=\theta^i\otimes\partial^t_i,$$

where $\theta^i=dq^i-q^i_t dt$ are the contact forms. We have the corresponding endomorphism,

$$\hat v(\partial_t)=-q^i_t\partial^t_i, \quad \hat v(\partial_i)=\partial^t_i, \quad \hat v(\partial^t_i)=0,$$

of the tangent bundle TJ^1Q and that

$$\hat v(dt)=0, \quad \hat v(dq^i)=0, \quad \hat v(dq^i_t)=\theta^i,$$

of the cotangent bundle T^*J^1Q of J^1Q. The nilpotent rule $\hat v^2=0$ holds.

Due to the imbeddings (3), any connection,

$$\Gamma=dt\otimes(\partial_t+\Gamma^i\partial_i),$$

on a fiber bundle $Q\to\mathbf{R}$ can be identified with a nowhere vanishing horizontal vector field,

$$\Gamma=\partial_t+\Gamma^i\partial_i, \tag{5}$$

on Q which is the horizontal lift of the standard vector field ∂_t on $\mathbf{R}$ by means of Γ. Conversely, any vector field Γ on Q such that $dt\rfloor\Gamma=1$ defines a connection on $Q\to\mathbf{R}$. Accordingly, the covariant differential,

$$D_\Gamma : J^1Q \underset{Q}{\to} VQ, \quad \dot{q}^i\circ D_\Gamma = q^i_t - \Gamma^i,$$

associated with a connection Γ on $Q\to\mathbf{R}$, takes its values into the vertical tangent bundle VQ of $Q\to\mathbf{R}$.

Remark: From the physical viewpoint, a connection (5) sets a reference frame. There is one-to-one correspondence between these connections and the equivalence classes of atlases of local constant trivializations of the fiber bundle $Q\to\mathbf{R}$, i.e., such that transition functions $q^i \to q'^i$ of the corresponding bundle coordinates are independent of t, and $\Gamma=\partial_t$ with respect to these coordinates.[13–15] In particular, every trivialization of Q defines a complete connection Γ on $Q\to\mathbf{R}$, and vice versa.

A connection ξ on the jet bundle $J^1Q\to\mathbf{R}$ is said to be holonomic if it is a section,

$$\begin{aligned} &\xi = \partial_t + q^i_t\partial_i + \xi^i\partial^t_i, \\ &dt\rfloor\xi = 1, \quad \xi\rfloor\hat{\upsilon} = 0, \end{aligned} \tag{6}$$

of the holonomic subbundle $J^2Q\to J^1Q$ of the affine jet bundle $J^1J^1Q\to J^1Q$. Holonomic connections (6) make up an affine space modeled over the linear space of vertical vector fields on the affine jet bundle $J^1Q\to Q$, i.e., which live in V_QJ^1Q. Every holonomic connection ξ defines the corresponding covariant differential on the jet manifold J^1Q:

$$\begin{aligned} &D_\xi : J^2Q \underset{J^1Q}{\longrightarrow} V_QJ^1Q \subset VJ^1Q, \\ &\dot{q}^i\circ D_\xi = 0, \quad \dot{q}^i_t\circ D_\xi = q^i_{tt} - \xi^i, \end{aligned} \tag{7}$$

which takes its values into the vertical tangent bundle V_QJ^1Q of the affine jet bundle $J^1Q\to Q$. Any integral section $\bar{c}:\mathbf{R}\supset(\,)\to J^1Q$ for a holonomic connection ξ is holonomic, i.e., $\bar{c}=\dot{c}$ where c is a curve in Q.

A second-order dynamic equation (or simply a dynamic equation) on a configuration bundle $Q\to\mathbf{R}$ is defined as the kernel,

$$q^i_{tt} = \xi^i(t,q^j,q^j_t), \tag{8}$$

of the covariant differential (7) for some holonomic connection ξ on the jet bundle $J^1Q\to\mathbf{R}$. Therefore, holonomic connections are also called dynamic equations. By a solution of the dynamic equation (8) is meant a curve c in Q whose second-order jet prolongation $\ddot{c}$ lives in (8). Any integral section $\bar{c}$ for the holonomic connection ξ is the jet prolongation $\dot{c}$ of a solution c of the dynamic equation (8), and vice versa.

III. NONHOLONOMIC CONSTRAINTS

Let $\mathbf{S}$ be an n-dimensional codistribution on the velocity phase space J^1Q. Its annihilator Ann$(\mathbf{S})$ is treated as a nonholonomic constraint. Let the codistribution $\mathbf{S}$ be locally spanned by the one-forms,

$$s^a = s^a_0 dt + s^a_i dq^i + \dot{s}^a_i dq^i_t,$$

on the jet manifold J^1Q. Then a dynamic equation $\tilde{\xi}$ on the configuration bundle $Q\to\mathbf{R}$ is said to be compatible with the nonholonomic constraint $\mathbf{S}$ if

$$s^a(\tilde{\xi}) = \tilde{\xi}\rfloor s^a = s^a_0 + s^a_i q^i_t + \dot{s}^a_i \tilde{\xi}^i = 0.$$

This equation is algebraically solvable for n components of $\tilde{\xi}$ iff the $n\times m$ matrix $\dot{s}^a_i(q^\lambda, q^i_t)$ has everywhere maximal rank $n \leq m$. Therefore, we restrict our consideration to the nonholonomic constraints, called admissible, such that $\dim \mathbf{S} = \dim \hat{v}(\mathbf{S})$.

If a nonholonomic constraint is admissible, there exists a local $m\times n$ matrix $\dot{s}^i_a(q^\lambda, q^i_t)$ such that

$$\dot{s}^i_a \dot{s}^b_i = \delta^b_a .$$

Then the local decomposition (1) of a dynamic equation ξ can be written in the form

$$\xi^i = \tilde{\xi}^i + \dot{s}^i_a s^a(\xi). \tag{9}$$

The global decomposition (1) exists by virtue of the following lemma.

Lemma 1: The intersection

$$W = J^2Q \cap \mathrm{Ann}(\mathbf{S})$$

is an affine bundle over J^1Q, modeled over the vector bundle

$$\overline{W} = V_Q J^1 Q \cap \mathrm{Ann}(\mathbf{S}).$$

Proof: $\overline{W}$ consists of the vertical vectors $v^i \partial^t_i \in V_Q J^1 Q$ which fulfill the conditions

$$\dot{s}^a_i(q^\lambda, q^j_t) v^i = 0.$$

Since the nonholonomic constraint $\mathbf{S}$ is admissible, every fiber of $\overline{W}$ is of dimension $m-n$, i.e., $\overline{W}$ is a vector bundle, while W is an affine bundle.

The affine structure of $W \to J^1Q$ implies that it has a global section $\tilde{\xi}$.

To construct the global decomposition (1), one usually performs a splitting of the vertical tangent bundle,

$$V_Q J^1 Q = \overline{W} \underset{J^1Q}{\oplus} \mathcal{V}, \tag{10}$$

and obtain the corresponding splitting of the second-order jet manifold,

$$J^2 Q = W \underset{J^1Q}{\oplus} \mathcal{V}. \tag{11}$$

Here $\mathcal{V} \to J^1Q$ should be interpreted as the bundle of possible constraint reaction accelerations.

If an admissible nonholonomic constraint $\mathbf{S}$ is of dimension $n = m$, a dynamic equation ξ is decomposed in a unique fashion. If $n < m$, the decomposition (1) is not unique. Different variants of this decomposition lead to different constraint reaction forces which, from the physical viewpoint, characterize different types of nonholonomic constraints. In next section, we will construct the decomposition of dynamic equations of Newtonian systems which corresponds to ideal nonholonomic constraints.

Now, let us consider some important examples of nonholonomic constraints.

Let N be a closed imbedded submanifold of the velocity phase space J^1Q, defined locally by the equations

$$f^a(q^\lambda, q^i_t) = 0, \quad a = 1,...,n < m.$$

One can treat N as a nonholonomic constraint given by the codistribution $\mathbf{S}=\mathrm{Ann}(TN)$ on $J^1Q|_N$. This codistribution is locally spanned by the one-forms

$$s^a=df^a=\partial_t f^a dt+\partial_j f^a dq^j+\partial_j^t f^a dq_t^j.$$

The nonholonomic constraint N is admissible iff the matrix $(\partial_j^t f^a)$ is of maximal rank n. It follows that N is a fibered submanifold of the affine jet bundle $J^1Q\to Q$.

A nonholonomic constraint N is said to be linear if it is an affine subbundle of the affine jet bundle $J^1Q\to Q$. Locally, a linear constraint N is given by the equations

$$f^a=f_0^a(q^\lambda)+f_i^a(q^\lambda)q_t^i=0, \tag{12}$$

where the matrix f_i^a is of maximal rank. A linear constraint is always admissible. Since N is an affine subbundle of $J^1Q\to Q$, it has a global section Γ (5) which is a connection on the configuration bundle $Q\to\mathbf{R}$, called the constraint reference frame. Then, the connection coefficients Γ^i satisfy the equations

$$f_0^a(q^\lambda)+f_i^a(q^\lambda)\Gamma^i=0,$$

and hence the constraint equations (12) take the form

$$f_i^a(q^\lambda)(q_t^i-\Gamma^i)=0. \tag{13}$$

One can think of $\dot q_\Gamma^i=q_t^i-\Gamma^i$, satisfying the equation (13), as virtual velocities relative to the linear constraint N.

Let now a configuration space Q admit a composite fibration $Q\to\Sigma\to\mathbf{R}$, where

$$\pi_{Q\Sigma}:Q\to\Sigma$$

is a fiber bundle, and let (t,σ^r,q^a) be coordinates on Q, compatible with this fibration. Given a connection,

$$B=dt\otimes(\partial_t+B^a\partial_a)+d\sigma^r\otimes(\partial_r+B_r^a\partial_a), \tag{14}$$

on the fiber bundle $Q\to\Sigma$, we have the corresponding horizontal splitting of the tangent bundle TQ. Restricted to the jet manifold $J^1Q\subset TQ$, this splitting reads

$$J^1Q=B(\pi_{Q\Sigma}^*J^1\Sigma)\underset{Q}{\oplus}V_\Sigma Q,$$

$$\partial_t+\sigma_t^r\partial_r+q_t^a\partial_a=[(\partial_t+B^a\partial_a)+\sigma_t^r(\partial_r+B_r^a\partial_a)]+[q_t^a-B^a-\sigma_t^rB_r^a]\partial_a,$$

where $\pi_{Q\Sigma}^*J^1\Sigma$ is the pull-back of the affine jet bundle $J^1\Sigma\to\Sigma$ onto Q. It is readily observed that

$$N=B(\pi_{Q\Sigma}^*J^1\Sigma)$$

is an affine subbundle of the affine jet bundle $J^1Q\to Q$, defined locally by the equations

$$q_t^a-\sigma_t^rB_r^a(q^\lambda)-B^a(q^\lambda)=0.$$

This subbundle yields a linear nonholonomic constraint.[16,17] The corresponding codistribution $\mathbf{S}=\mathrm{Ann}(TN)$ is locally spanned by the one-forms,

$$s^a=-(\partial_tB^a+\sigma_t^r\partial_tB_r^a)dt-(\partial_sB^a+\sigma_t^r\partial_sB_r^a)d\sigma^s-(\partial_bB^a+\sigma_t^r\partial_bB_r^a)dq^b+dq_t^a-B_r^a d\sigma_t^r. \tag{15}$$

With the connection (14), we also have the splitting of the vertical tangent bundle VQ of $Q\to\mathbf{R}$ and the corresponding splitting of the vertical tangent bundle V_QJ^1Q which reads

$$V_Q J^1 Q = \overline{W} \underset{J^1Q}{\oplus} \mathcal{V},$$
$$\dot\sigma^r_t \partial^t_r + \dot q^a_t \partial^t_a = \dot\sigma^r_t (\partial^t_r + B^a_r \partial^t_a) + (\dot q^a_t - B^a_r \dot\sigma^r_t)\partial^t_a . \tag{16}$$

It is readily observed that $\overline{W}|_N$ consists of vertical vectors which are the annihilators of the codistribution (15). The splitting (16) yields the corresponding splitting (11) of the second-order jet manifold J^2Q. Then we obtain the decomposition (1) of every dynamic equation ξ on J^1Q as

$$\widetilde{\xi}^r = \xi^r, \quad \widetilde{\xi}^a = \xi^a - s^a(\xi).$$

IV. NEWTONIAN SYSTEMS WITH NONHOLONOMIC CONSTRAINTS

Let $Q \to \mathbf{R}$ be a fiber bundle together with (i) a nondegenerate fiber metric,

$$\hat m : J^1Q \to V^*Q \underset{Q}{\otimes} V^*Q, \quad \hat m = \tfrac{1}{2} m_{ij} \overline{d}q^i \vee \overline{d}q^j,$$

in the fiber bundle $V_Q J^1 Q \to J^1 Q$ which satisfies the symmetry condition,

$$\partial^t_k m_{ij} = \partial^t_j m_{ik}, \tag{17}$$

and (ii) a dynamic equation ξ (6) on the jet bundle $J^1Q \to \mathbf{R}$, related to the fiber metric $\hat m$ by the compatibility condition,

$$2\xi \rfloor dm_{ij} + m_{ik}\partial^t_j \xi^k + m_{jk}\partial^t_i \xi^k = 0. \tag{18}$$

The triple $(Q, \hat m, \xi)$ is called a Newtonian system and $\hat m$ is named a mass metric.[15] A Newtonian system is said to be standard if $\hat m$ is the pull-back of a fiber metric in the vertical tangent bundle VQ in accordance with the isomorphism (4). In this case, $\hat m$ is independent of the velocity coordinates q^i_t.

There are two main reasons in order to consider Newtonian systems. From the physical viewpoint, with a mass metric, we can introduce the notion of an external force, defined as a section of the vertical cotangent bundle $V^*_Q J^1 Q \to J^1 Q$. Let $(Q, \hat m, \xi)$ be a Newtonian system and F an external force. Then

$$\xi^i_F = \xi^i + (m^{-1})^{ik} F_k$$

is a dynamic equation, but the triple $(Q, \hat m, \xi_F)$ is a Newtonian system only if F possesses the property

$$\partial^t_i F_j + \partial^t_j F_i = 0. \tag{19}$$

From the mathematical viewpoint, the equation

$$m_{ik}(q^k_{tt} - \xi^k) = 0 \tag{20}$$

is the kernel of an Euler–Lagrange-type operator. By an appropriate choice of a mass metric, one may hope to bring it into Lagrange equations. This is the well-known inverse problem in time-dependent mechanics.

Here, we consider Newtonian systems because they provide the vertical tangent bundle $V_Q J^1 Q$ with a nondegenerate fiber metric $\hat m$. Let us assume that $\hat m$ is a Riemannian metric. With this metric, we immediately obtain the splitting (10), where $\mathcal{V}$ is the orthocomplement of $\overline{W}$. Then the corresponding decomposition (9) takes the form[11]

$$\xi^i = \tilde{\xi}^i + \tilde{m}_{ab} m^{ij} \dot{s}_j^a s^b(\xi), \tag{21}$$

where $\tilde{m}_{ab}$ is the inverse matrix of

$$\tilde{m}^{ab} = \dot{s}_i^a \dot{s}_j^b m^{ij}.$$

It is readily observed that the decomposition (21) satisfies the generalized D'Alembert principle. The constraint reaction acceleration,

$$-r^i = -\tilde{m}_{ab} m^{ij} \dot{s}_j^a s^b(\xi), \tag{22}$$

is orthogonal to every element of $V_Q J^1 Q \cap \mathrm{Ann}(\mathbf{S})$ with respect to the mass metric $\hat{m}$. Since elements of $V_Q J^1 Q \cap \mathrm{Ann}(\mathbf{S})$ can be treated as the virtual accelerations relative to the nonholonomic constraint **S**, the constraint reaction acceleration (22) characterizes **S** as an ideal constraint.

The Gauss principle is also fulfilled as follows. Given a dynamic equation ξ and the abovementioned fiber metric $\hat{m}$, let us define a positive function $G(w)$ on $J^2 Q$ as

$$G(w) = \hat{m}(\xi(\pi_1^2(w)) - w, \xi(\pi_1^2(w)) - w),$$

$$G(q^\lambda, q_t^i, q_{tt}^i) = m_{ij}(q^\lambda, q_t^k)(\xi^i(q^\lambda, q_t^k) - q_{tt}^i)(\xi^j(q^\lambda, q_t^k) - q_{tt}^j).$$

We say that $\|w\| = G(w)^{1/2}$ is a norm of $w \in J^2 Q$.

Proposition 2: Among all dynamic equations compatible with a nonholonomic constraint, the dynamic equation $\tilde{\xi}$ defined by the decomposition (21) is that of least norm.

Proof: Let ζ be another dynamic equation which takes its values into W. Then $\tilde{\xi} - \zeta \subset \overline{W}$ and

$$\hat{m}(\tilde{\xi} - \zeta, \xi - \tilde{\xi}) = 0.$$

Hence, we obtain

$$\|\zeta\| = \hat{m}(\xi - \tilde{\xi} + \tilde{\xi} - \zeta, \xi - \tilde{\xi} + \tilde{\xi} - \zeta) = \|\tilde{\xi}\| + \hat{m}(\tilde{\xi} - \zeta, \tilde{\xi} - \zeta).$$

In the next section, we will show that, in the case of nondegenerate Lagrangian systems and linear nonholonomic constraints, the decomposition (21) satisfies the traditional D'Alembert principle.

V. LAGRANGIAN SYSTEMS WITH NONHOLONOMIC CONSTRAINTS

Nondegenerate Lagrangian systems are particular Newtonian systems.

A Lagrangian is defined as a horizontal density,

$$L = \mathcal{L} dt, \quad \mathcal{L}: J^1 Q \to \mathbf{R}, \tag{23}$$

on the velocity phase space $J^1 Q$. Here, we apply in a straightforward manner the first variational formula.[13,15]

Let us consider a projectable vector field

$$u = u^t \partial_t + u^i \partial_i, \quad u^t = 0,1,$$

on the configuration bundle $Q \to \mathbf{R}$ and calculate the Lie derivative of the Lagrangian (23) along the jet prolongation,

$$\overline{u} = u^t \partial_t + u^i \partial_i + d_t u^i \partial_i^t,$$

of u, where $d_t = \partial_t + q_t^i \partial_i + \cdots$ is the operator of formal derivative. We obtain

$$\mathbf{L}_{\bar{u}}L=(\overline{u}\rfloor d\mathcal{L})dt=(u^t\partial_t+u^i\partial_i+d_tu^i\partial_i^t)\mathcal{L}dt. \tag{24}$$

The first variational formula provides the following canonical decomposition of the Lie derivative (24) in accordance with the variational problem:

$$\overline{u}\rfloor d\mathcal{L}=(u^i-u^tq_t^i)\mathcal{E}_i+d_t(u\rfloor H_L), \tag{25}$$

where

$$H_L=\hat{v}(d\mathcal{L})+L=\pi_i dq^i-(\pi_i q_t^i-\mathcal{L})dt \tag{26}$$

is the Poincaré–Cartan form and

$$\mathcal{E}_L:J^2Q\to V^*Q,$$

$$\mathcal{E}_L=\mathcal{E}_i\theta^i=(\partial_i-d_t\partial_i^t)\mathcal{L}\theta^i \tag{27}$$

is the Euler–Lagrange operator for L. We will use the notation

$$\pi_i=\partial_i^t\mathcal{L},\quad \pi_{ji}=\partial_j^t\partial_i^t\mathcal{L}.$$

A Lagrangian L is called nondegenerate if det $\pi_{ji}\neq 0$ everywhere on the velocity phase space J^1Q.

The kernel $\operatorname{Ker}\mathcal{E}_L\subset J^2Q$ of the Euler–Lagrange operator (27) defines the system of second-order differential equations,

$$(\partial_i-d_t\partial_i^t)\mathcal{L}=0, \tag{28}$$

on Q, called the Lagrange equations. Their solutions are (local) section c of the fiber bundle $Q\to\mathbf{R}$ whose second-order jet prolongations $\ddot{c}$ live in (28).

A holonomic connection on the jet bundle $J^1Q\to\mathbf{R}$ is said to be a Lagrangian connection ξ_L for the Lagrangian L if it takes its values in the kernel (28) of the Euler–Lagrange operator $\mathcal{E}_L$. Every Lagrangian connection ξ_L defines a dynamic equation on the configuration space Q whose solutions are also solutions of the Lagrange equations (28). If L is nondegenerate, the Lagrange equation (28) can be algebraically solved for the second-order derivatives, and they are equivalent to the dynamic equation,

$$q_{tt}^i=\xi_L^i,\quad \xi_L^i=(\pi^{-1})^{ij}\mathcal{E}_j+q_{tt}^i, \tag{29}$$

called the Lagrange dynamic equation.

Every Lagrangian L on the jet manifold J^1Q yields the Legendre map,

$$\hat{L}:J^1Q\to V^*Q,\quad p_i\circ\hat{L}=\pi_i, \tag{30}$$

where (t,q^i,p_i) are holonomic coordinates on the vertical cotangent bundle V^*Q. As is well known, the Legendre map (30) is a local diffeomorphism iff L is nondegenerate. A Lagrangian L is called hyperregular if the Legendre map $\hat{L}$ is a diffeomorphism.

The vertical tangent map $V\hat{L}$ to the Legendre map $\hat{L}$ reads

$$V\hat{L}:V_QJ^1Q\to VV^*Q\cong V^*Q\underset{Q}{\times}V^*Q.$$

It yields the linear fibred morphism $V_QJ^1Q\to V_Q^*J^1Q$ and the corresponding mapping,

$$J^1Q\to V_Q^*J^1Q\underset{J^1Q}{\otimes}V_Q^*J^1Q,\quad m_{ij}=\pi_{ij}. \tag{31}$$

If a Lagrangian L is nondegenerate, then (31) is a mass metric which satisfies the symmetry condition (17) and the compatibility condition (18) for the Lagrange dynamic equation (29).

Thus, every nondegenerate Lagrangian L defines a Newtonian system. Moreover, a nondegenerate Lagrangian system plus an external force which fulfills the condition (19) is also a Newtonian system. Conversely, every standard Newtonian system can be seen as a Lagrangian system with the Lagrangian

$$L=\tfrac{1}{2}m_{ij}(q_t^i-\Gamma^i)(q_t^j-\Gamma^j)dt, \tag{32}$$

where Γ is a reference frame, plus an external force.

Given a nondegenerate Lagrangian L with a Riemannian mass metric $m_{ij}=\pi_{ij}$, let now $\mathbf{S}$ be an admissible nonholonomic constraint on the velocity phase space J^1Q. Since this is a particular Newtonian system, we obtain the dynamic equation

$$\begin{aligned} q_{tt}^i&=\xi_L^i-\widetilde{m}_{ab}m^{ij}\dot{s}_j^a(\dot{s}_k^b\xi_L^k+s_k^bq_t^k+s_0^b),\\ \xi_L^i&=m^{ij}(-\partial_t\pi_j-\partial_k\pi_jq_t^k+\partial_j\mathcal{L}), \end{aligned} \tag{33}$$

which is compatible with the constraint $\mathbf{S}$, treated as an ideal nonholonomic constraint. This is the Lagrange dynamic equation in the presence of the additional constraint reaction force

$$F_i=-\widetilde{m}_{ab}\dot{s}_i^a s^b(\xi_L). \tag{34}$$

Let us consider the energy conservation law in the presence of this force.

The energy conservation law in Lagrangian time-dependent mechanics is deduced from the first variational formula (25) when the vector field $u=\Gamma$ is a reference frame. On the shell $\mathcal{E}_i=0$ (28), this formula leads to the weak identity,

$$\mathbf{L}_{\bar{\Gamma}}L\approx-d_t(\pi_i\dot{q}_\Gamma^i-\mathcal{L}), \tag{35}$$

where $\dot{q}_\Gamma^i=q_t^i-\Gamma^i$ is a relative velocity and

$$T_\Gamma=\pi_i\dot{q}_\Gamma^i-\mathcal{L} \tag{36}$$

is the energy function with respect to the reference frame Γ.[14,15,18] In the presence of an external force F, i.e., on the shell $\mathcal{E}_i=-F_i$, the weak identity (35) is modified as

$$\mathbf{L}_{\bar{\Gamma}}L-\dot{q}_\Gamma^iF_i=-d_tT_\Gamma.$$

It is readily observed that, if a nonholonomic constraint is linear and Γ is a constraint reference frame, the constraint reaction force (34) does not contribute to the energy conservation law. It follows that, in this case, the standard D'Alembert principle holds, while the equation (33) describes a motion in the presence of an ideal nonholonomic constraint in the spirit of this principle.

The constrained equation of motion (33) is neither Lagrange equations nor a dynamic equation of a Newtonian system. In Sec. VI, we aim to show that it can be seen as a part of Hamilton equations in the framework of the Hamiltonian formalism extended to the configuration space VQ.

VI. VERTICAL EXTENSION OF HAMILTONIAN FORMALISM

This section provides a brief exposition of Hamiltonian formalism of time-dependent mechanics on a configuration bundle $Q\to\mathbf{R}$ and its extension to the vertical configuration space VQ. We consider this extension because any first-order dynamic equation on the momentum phase space V^*Q can be seen as a Hamilton equation in the framework of the extended Hamiltonian formalism. This extension is also of interest in the path-integral formulation of mechanics.[19,20]

Given a mechanical system on a configuration bundle $Q\to\mathbf{R}$, its momentum phase space is the vertical cotangent bundle V^*Q of $Q\to\mathbf{R}$, equipped with the holonomic coordinates $(t,q^i,p_i=\dot q_i)$.[13–15] The momentum phase space V^*Q is endowed with the canonical exterior three-form,

$$\mathbf{\Omega}=dp_i\wedge dq^i\wedge dt.$$

Let us consider the cotangent bundle T^*Q of Q with the holonomic coordinates (t,q^i,p,p_i). It admits the canonical Liouville form

$$\Xi=p\,dt+p_i dq^i. \tag{37}$$

An exterior one-form H on the momentum phase space V^*Q is called a Hamiltonian form if it is the pull-back

$$H=h^*\Xi=p_i dq^i-\mathcal{H}dt \tag{38}$$

of the Liouville form Ξ (37) by a section h of the fiber bundle

$$\zeta:T^*Q\to V^*Q. \tag{39}$$

Remark: With respect to a trivialization $Q\cong\mathbf{R}\times M$, the Hamiltonian form (38) is the well-known integral invariant of Poincaré–Cartan, where $\mathcal{H}$ is a Hamiltonian. The peculiarity of Hamiltonian time-dependent mechanics issues from the fact that Hamiltonians are not scalar functions under time-dependent transformations, but make up an affine space modeled over the linear space of functions on V^*Q.

For instance, every connection Γ on a configuration bundle $Q\to\mathbf{R}$ is an affine section.

$$p\circ\Gamma=-p_i\Gamma^i,$$

of the fiber bundle (39), and defines the Hamiltonian form

$$H_\Gamma=p_i dq^i-p_i\Gamma^i dt.$$

It follows that any Hamiltonian form on the momentum phase space V^*Q admits the splitting,

$$H=H_\Gamma-\tilde{\mathcal{H}}_\Gamma dt=p_i dq^i-(p_i\Gamma^i+\tilde{\mathcal{H}}_\Gamma)dt,$$

where Γ is a connection on $Q\to\mathbf{R}$ and $\tilde{\mathcal{H}}_\Gamma$ is a real function on V^*Q, called the Hamiltonian function. The following assertions are basic facts in the Hamiltonian formulation of time-dependent mechanics.[14,15]

Proposition 3: Every Hamiltonian form H on the momentum phase space V^*Q defines the associated Hamiltonian map,

$$\hat H:V^*Q\to J^1Q, \quad q^i_t\circ\hat H=\partial^i\mathcal{H}.$$

Proposition 4: Given a Hamiltonian form H on the momentum phase space V^*Q there exists a unique connection

$$\gamma_H=\partial_t+\partial^i\mathcal{H}\partial_i-\partial_i\mathcal{H}\partial^i \tag{40}$$

on $V^*Q\to\mathbf{R}$, called a Hamiltonian connection, such that

$$\gamma_H\rfloor\mathbf{\Omega}=dH.$$

The kernel of the covariant differential of the Hamiltonian connection (40) defines the Hamilton equations,

$$q^i_t = \partial^i \mathcal{H}, \tag{41a}$$

$$p_{ti} = -\partial_i \mathcal{H}, \tag{41b}$$

for the Hamiltonian form H. Their solutions are integral curves for the Hamiltonian connection γ_H (40).

Now let us consider the vertical tangent bundle VQ of the fiber bundle $Q \to \mathbf{R}$, coordinated by $(t,q^i,\dot{q}^i)$. It can be seen as a new configuration space, called the vertical configuration space. The corresponding vertical momentum phase space is the vertical cotangent bundle V^*VQ of $VQ \to \mathbf{R}$. The vertical momentum phase space V^*VQ is canonically isomorphic to the vertical tangent bundle VV^*Q of the ordinary momentum phase space $V^*Q \to \mathbf{R}$, coordinated by $(t,q^i,p_i,\dot{q}^i,\dot{p}_i)$. It is easily seen from the transformation laws that $(q^i,\dot{p}_i)$ and $(\dot{q}^i,p_i)$ are canonically conjugate pairs.

The vertical momentum phase space VV^*Q is endowed with the canonical three-form,

$$\mathbf{\Omega}_V = [d\dot{p}_i \wedge dq^i + dp_i \wedge d\dot{q}^i] \wedge dt.$$

For the sake of brevity, one can write $\mathbf{\Omega}_V = \partial_V \mathbf{\Omega}$, where $\partial_V = \dot{q}^i \partial_i + \dot{p}_i \partial^i$ is the vertical derivative.

The notions of a Hamiltonian connection, a Hamiltonian form, etc., on the vertical momentum phase space $VV^*Q \cong V^*VQ$ are introduced similarly to those on the ordinary momentum phase space V^*Q. In particular, a Hamiltonian form on VV^*Q reads

$$H_V = \dot{p}_i dq^i + p_i d\dot{q}^i - \mathcal{H}_V dt.$$

Since Hamiltonian forms are determined modulo exact forms and the function $p_i \dot{q}^i$ is globally defined on VV^*Q, we will write

$$H_V = \dot{p}_i dq^i - \dot{q}^i dp_i - \mathcal{H}_V dt. \tag{42}$$

The corresponding Hamilton equations read

$$\gamma^i = q^i_t = \dot{\partial}^i \mathcal{H}_V, \tag{43a}$$

$$\gamma_i = p_{ti} = -\dot{\partial}_i \mathcal{H}_V, \tag{43b}$$

$$\overline{\gamma}^i = \dot{q}^i_t = \partial^i \mathcal{H}_V, \tag{43c}$$

$$\overline{\gamma}_i = \dot{p}_{ti} = -\partial_i \mathcal{H}_V, \tag{43d}$$

where

$$\overline{\gamma} = \partial_t + \gamma^i \partial_i + \gamma_i \partial^i + \overline{\gamma}^i \dot{\partial}_i + \overline{\gamma}_i \dot{\partial}^i$$

is a Hamiltonian connection on the vertical momentum phase space $VV^*Q \to \mathbf{R}$.

There is the following relation between Hamiltonian formalisms on V^*Q and VV^*Q.[13,15] Let VT^*Q be the vertical tangent bundle of the cotangent bundle $T^*Q \to \mathbf{R}$ equipped with holonomic coordinates $(t,q^i,p_i,p,\dot{q}^i,\dot{p}_i,\dot{p})$ and endowed with the canonical form,

$$\Xi_V = \dot{p}\,dt + \dot{p}_i dq^i - \dot{q}^i dp_i.$$

Proposition 5: Let γ_H be a Hamiltonian connection on the ordinary momentum phase space $V^*Q \to \mathbf{R}$ for a Hamiltonian form,

$$H = h^* \Xi = p_i dq^i - \mathcal{H} dt. \tag{44}$$

Then the connection

$$V\gamma_H : VV^*Q \to VJ^1V^*Q \cong J^1VV^*Q,$$
$$V\gamma_H = \partial_t + \gamma^i\partial_i + \gamma_i\partial^i + \partial_V\gamma^i\dot\partial_i + \partial_V\gamma_i\dot\partial^i \tag{45}$$

on the vertical momentum phase space $VV^*Q \to \mathbf{R}$ is a Hamiltonian connection for the Hamiltonian form,

$$H_V = (Vh)^*\Xi_V = \dot p_i dq^i - \dot q^i dp_i - \partial_V\mathcal{H}dt, \tag{46}$$

$$\partial_V\mathcal{H} = (\dot q^i\partial_i + \dot p_i\partial^i)\mathcal{H}, \tag{47}$$

where $Vh: VV^*Q \to VT^*Q$ is the vertical tangent map to h.

The corresponding Hamilton equations read

$$\gamma^i = \dot\partial^i\mathcal{H}_V = \partial^i\mathcal{H}, \tag{48a}$$

$$\gamma_i = -\dot\partial_i\mathcal{H}_V = -\partial_i\mathcal{H}, \tag{48b}$$

$$\overline{\gamma}^i = \partial^i\mathcal{H}_V = \partial_V\partial^i\mathcal{H}, \tag{48c}$$

$$\overline{\gamma}_i = -\partial_i\mathcal{H}_V = -\partial_V\partial_i\mathcal{H}. \tag{48d}$$

It is easily seen that the equations (48a) and (48b) are exactly the Hamilton equations (41a) and (41b) for the Hamiltonian form H.

Remark: In order to clarify the physical meaning of the Hamilton equations (48c) and (48d) let $r(t)$ be a solution of the Hamilton equations (48a) and (48b). Let $\dot r(t)$ be a Jacobi field, i.e., $r(t) + \varepsilon\dot r(t)$ is also a solution of the same Hamilton equations modulo terms of order two in ε. Then it is readily observed that the Jacobi field $\dot r(t)$ fulfills the Hamilton equations (48c) and (48d).

The following assertion plays a prominent role in the sequel.[13,15]

Proposition 6: Any connection γ on the momentum phase space $V^*Q \to \mathbf{R}$ gives rise to the Hamiltonian connection,

$$\gamma^i = \gamma^i, \quad \gamma_i = \gamma_i, \quad \overline{\gamma}^i = \dot p_j\partial^i\gamma^j - \dot q^j\partial^i\gamma_j, \quad \overline{\gamma}_i = -\dot p_j\partial_i\gamma^j + \dot q^j\partial_i\gamma_j, \tag{49}$$

for the Hamiltonian form,

$$H_V = \dot p_i(dq^i - \gamma^i dt) - \dot q^i(dp_i - \gamma_i dt) = \dot p_i dq^i - \dot q^i dp_i - (\dot p_i\gamma^i - \dot q^i\gamma_i)dt,$$

on the vertical momentum phase space VV^*Q.

In particular, if γ is a Hamiltonian connection on the fiber bundle $V^*Q \to \mathbf{R}$, then (49) is exactly the connection $V\gamma$ (45).

It follows that every first-order dynamic equation on the momentum phase space V^*Q can be seen as the Hamilton equations (43a) and (43b) for a suitable Hamiltonian form on the vertical momentum phase space.

VII. HAMILTONIAN SYSTEMS WITH NONHOLONOMIC CONSTRAINTS

Let L be a hyperregular Lagrangian with a Riemannian mass metric $\hat m$. In this case, Hamiltonian and Lagrangian formalisms of time-dependent mechanics are equivalent. There exists a unique associated Hamiltonian form H (38) on V^*Q such that

$$\hat H = \hat L^{-1}, \quad p_i \equiv \pi_i(q^\lambda, \partial^j\mathcal{H}(q^\lambda, p_k)), \quad q^i_t \equiv \partial^i\mathcal{H}(q^\lambda, \pi_j(q^\lambda, q^k_t)), \tag{50a}$$

$$\mathcal{L}\circ\hat{H}\equiv\gamma_H\rfloor H=p_i\partial^i\mathcal{H}-\mathcal{H}. \tag{50b}$$

As an immediate consequence of (50a), we have $J^1\hat{H}=(J^1\hat{L})^{-1}$, where the jet prolongations of the Hamiltonian and Legendre maps read

$$J^1\hat{H}:J^1V^*Q\to J^1J^1Q,\quad (q^\lambda,q^i_t,q^i_{(t)},q^i_{tt})\circ J^1\hat{H}=(q^\lambda,\partial^i\mathcal{H},q^i_t,d_t\partial^i\mathcal{H}),$$

$$J^1\hat{L}:J^1J^1Q\to J^1V^*Q,\quad (q^\lambda,p_i,q^i_t,p_{ti})\circ J^1\hat{L}=(q^\lambda,\pi_i,q^i_{(t)},d_t\pi_i).$$

Then, using (50a) and (50b), we obtain

$$\gamma_H=J^1\hat{L}\circ\xi_L\circ\hat{H}.$$

Let introduce the notation $M^{ij}=\partial^i\partial^j\mathcal{H}$. There are the relations

$$M^{ik}(m_{kj}\circ\hat{H})=\delta^i_j,\quad m_{kj}(M^{ik}\circ\hat{L})=\delta^i_j,\quad m_{ij}=\pi_{ij}.$$

It follows that M is a fiber metric in the vertical tangent bundle V_QV^*Q of the fiber bundle $V^*Q\to Q$.

Given a codistribution $\mathbf{S}$ on J^1Q, let us consider the pull-back codistribution $\hat{H}^*\mathbf{S}$ on V^*Q, spanned locally by one-forms

$$\beta^a=\hat{H}^*s^a=(s^a_0+\dot{s}^a_j\partial_t\partial^j\mathcal{H})dt+(s^a_i+\dot{s}^a_j\partial_i\partial^j\mathcal{H})dq^i+\dot{s}^a_iM^{ij}dp_j=\beta^a_0dt+\beta^a_idq^i+\dot{\beta}^{ai}dp_i.$$

This codistribution defines a nonholonomic constraint on the momentum phase space V^*Q.

Given a Hamiltonian connection γ_H (40), let us find its splitting

$$\gamma_H=\tilde{\gamma}+\vartheta, \tag{51}$$

where $\tilde{\gamma}$ is a connection on $V^*Q\to\mathbf{R}$ which satisfies the condition

$$\tilde{\gamma}\subset\mathrm{Ann}(\hat{H}^*\mathbf{S}). \tag{52}$$

The connection $\tilde{\gamma}$ (52) obviously defines a first-order dynamic equation on the momentum phase space V^*Q which is compatible with the nonholonomic constraint $\hat{H}^*\mathbf{S}$. The decomposition (51) is not unique. Let us construct it as follows.

Given a Hamiltonian connection γ_H, we consider the codistribution S_H on V^*Q, spanned locally by the one-forms $dq^i-\gamma^i_Hdt$. Its annihilator $\mathrm{Ann}(S_H)$ is an affine subbundle of the affine jet bundle $J^1V^*Q\to V^*Q$, modeled over the vertical tangent bundle V_QV^*Q. The Hamiltonian connection γ_H is a section of this subbundle. Let us take the intersection

$$W=\mathrm{Ann}(S_H)\cap\mathrm{Ann}(\hat{H}^*\mathbf{S}).$$

Lemma 7: W is an affine bundle over V^*Q, modeled over the vector bundle

$$\overline{W}=V_QV^*Q\cap\mathrm{Ann}(\hat{H}^*\mathbf{S}).$$

Proof: The intersection $\overline{W}$ consists of elements $v=v_i\partial^i$ of V_QV^*Q which fulfill the conditions

$$v_i\dot{\beta}^{ai}=0.$$

Since the nonholonomic constraint $\mathbf{S}$ is admissible and the matrix M^{ij} is nondegenerate, every fiber of $\overline{W}$ is of dimension $m-n$, i.e., $\overline{W}$ is a vector bundle, while W is an affine bundle.

Then, using the fiber metric M in V_QV^*Q, we obtain the splitting

$$V_Q V^*Q = \overline{W} \oplus \mathcal{V},$$

where $\mathcal{V}$ is the orthocomplement of $\overline{W}$, and the associated splitting

$$\mathrm{Ann}(S_H) = W \oplus \mathcal{V}.$$

The corresponding decomposition (51) reads

$$\tilde{\gamma} = \gamma_H - \tilde{M}_{ab} M_{ij} \dot{\beta}^{ai} \beta^b(\gamma_H) \partial^j, \tag{53}$$

where $\tilde{M}_{ab}$ is the inverse matrix of

$$\tilde{M}^{ab} = \dot{\beta}^{ai} \dot{\beta}^{bj} M_{ij}.$$

The splitting (53) is the Hamiltonian counterpart of the splitting (21). We have the relations

$$\tilde{m}^{ab} = \tilde{M}^{ab} \circ \hat{H}, \quad \beta^a(\gamma_H) = s^a(\xi_L) \circ \hat{H},$$

and as a consequence

$$\tilde{\gamma} = J^1 L \circ \tilde{\xi} \circ \hat{H}.$$

Remark: The above procedure can be extended in a straightforward manner to any standard Newtonian system, seen as a Lagrangian system with the Lagrangian (32) and an external force. Following this procedure, one may also study a nonholonomic Hamiltonian system without appealing to its Lagrangian counterpart.

The connection (53) defines the system of first-order dynamic equations,

$$q^i_t = \partial^i \mathcal{H}, \quad p_{ti} = -\partial_i \mathcal{H} - \tilde{M}_{ab} M_{ij} \dot{\beta}^{ai} \beta^b(\gamma_H), \tag{54}$$

on the momentum phase space V^*Q, which are not Hamilton equations. Nevertheless, in accordance with Proposition 6, one can restate the constrained equations of motion (54) as the Hamilton equations (48a) and (48b) for the Hamiltonian form,

$$H_V = \dot{p}_i dq^i - \dot{q}^i dp_i - \partial_V \mathcal{H} dt - \dot{q}^i \tilde{M}_{ab} M_{ij} \dot{\beta}^{ai} \beta^b(\gamma_H) dt,$$

on the vertical momentum phase space VV^*Q, where the last term can be written in brief as $(-\partial_V \rfloor \vartheta \rfloor \mathbf{\Omega})$.

The Hamiltonian form of the constrained equations of motion may be important in connection with the following speculations.

Given a Hamiltonian form H_V (42) on the vertical momentum phase space VV^*Q, let us consider the Lagrangian

$$L_H = \dot{p}_i q^i_t - \dot{\tilde{q}}^i p_{ti} - \mathcal{H}_V \tag{55}$$

on the first-order jet manifold $J^1 VV^*Q$ of the fiber bundle $VV^*Q \to \mathbf{R}$. It is readily observed that the corresponding Lagrange equations are exactly the Hamilton equations (43a)–(43d) for the Hamiltonian form H_V. In particular, let H be a Hamiltonian form on an ordinary momentum phase space V^*Q and $\mathcal{H}_V = \partial_V \mathcal{H}$. In this case, the Lagrangian (55) reads

$$L_H = \dot{p}_i (q^i_t - \partial^i \mathcal{H}) - \dot{q}^i (p_{ti} + \partial_i \mathcal{H}).$$

It is easily seen that this Lagrangian vanishes on solutions of the Hamilton equations for the Hamiltonian form H. By this reason, it is applied to the functional integral formulation of mechanics.[19,20]

[1] A. M. Vershik, in *Global Analysis—Studies and Applications 1*, Lecture Notes in Mathematics (Springer-Verlag, Berlin, 1984), Vol. 1108, p. 278.
[2] J. Cariñena and M. Rañada, J. Phys. A **26**, 1335 (1993).
[3] A. M. Vershik and V. Ya. Gershkovich, in *Dynamical Systems VII*, edited by V. I. Arnol'd and S. P. Novikov (Springer-Verlag, Berlin, 1994), p. 1.
[4] C.-M. Marle, Commun. Math. Phys. **174**, 295 (1995).
[5] F. Cardin and M. Favretti, J. Geom. Phys. **18**, 295 (1996).
[6] M. de León and D. Martín de Diego, J. Math. Phys. **37**, 3389 (1996).
[7] M. de León, J. Marrero, and D. Martín de Diego, Int. J. Theor. Phys. **36**, 979 (1997).
[8] M. de León, J. Marrero, and D. Martín de Diego, J. Phys. A **30**, 1167 (1997).
[9] O. Krupkova, J. Math. Phys. **38**, 5098 (1997).
[10] A. Ibort, M. de León, E. Lacomba, J. Marrero, D. Martín de Diego, and P. Pitanga, J. Phys. A **31**, 2655 (1998).
[11] G. Giachetta, J. Math. Phys. **33**, 1652 (1992).
[12] E. Massa and E. Pagani, Ann. Inst. Henri Poincaré Phys. Theor. **61**, 17 (1994).
[13] G. Giachetta, L. Mangiarotti, and G. Sardanashvily, *New Lagrangian and Hamiltonian Methods in Field Theory* (World Scientific, Singapore, 1997).
[14] G. Sardanashvily, J. Math. Phys. **39**, 2714 (1998).
[15] L. Mangiarotti and G. Sardanashvily, *Gauge Mechanics* (World Scientific, Singapore, 1998).
[16] W. Sarlet, F. Cantrijn, and D. Saunders, J. Phys. A **28**, 3253 (1995).
[17] W. Sarlet, F. Cantrijn, and D. Saunders, J. Phys. A **29**, 4265 (1996).
[18] A. Echeverría Enríquez, M. Muñoz Lecanda, and N. Romń Roy, J. Phys. A **28**, 5553 (1995).
[19] E. Gozzi, M. Reuter, and W. Thacker, Phys. Rev. D **40**, 3363 (1989).
[20] E. Gozzi, M. Reuter, and W. Thacker, Phys. Rev. D **46**, 757 (1992).

An evolutionary equation in thermoelasticity of dipolar bodies

Marin Marin[a)]
Faculty of Mathematics, University of Brasov Str. Iuliu Mariu No. 50, 2200 Brasov, Romania

(Received 6 April 1998; accepted for publication 10 December 1998)

In this paper we apply the theory of semigroups of operators in order to obtain the existence and uniqueness of solutions for the mixed initial-boundary value problems in thermoelasticity of dipolar bodies. The continuous dependence of the solutions upon initial data and supply terms is also proved. © *1999 American Institute of Physics.* [S0022-2488(99)03203-X]

I. INTRODUCTION

In recent years new continuous models of elastic bodies have been intensively studied. The departure from classical theories begins with polar theories (see Refs. 1–3, for instance).

The nonclassical (which include the theory of dipolar bodies) have found important applications in a variety of fields. Crystals, composites, polymers, suspensions, blood, grids and multibar systems can by considered as examples of media with microstructure. The domain of applicability of different non-classical theories of elastic media has been investigated in the paper.[4] Many aspects of these theories we can find in many new papers (see Refs. 5, 4, for instance).

The deformation of a dipolar medium is described by the variables

$$u_i = u_i(X,t), \varphi_{jk} = \varphi_{jk}(X,t), (X,t) \in B \times [0,t_0),$$

where u_i is the displacement field and φ_{jk} is the dipolar displacement field.

The theories of dipolar bodies are quite sufficient for a large number of solid mechanics applications. Because the system of governing equations and conditions for the thermoelasticity of dipolar bodies is more complicated, it is necessarily a new approach for the boundary value problem in this context. In this paper we establish an existence and uniqueness result for the solutions of the initial-boundary value problem in the context of the thermoelasticity of dipolar bodies. In this paper we also investigate the continuous dependence upon the initial data and supply terms of the solutions of the above problem. An inhomogeneous and anisotropic elastic material is considered and the initial-boundary value problem is transformed in an abstract temporally homogeneous evolutionary equation in a Hilbert space. By using the results of the semigroups theory of linear operators, the existence, uniqueness and continuous dependence results are derived. The proof is given for the first boundary value problem, but the results are the same if the boundary conditions are replaced by those from the second or the third problem.

II. NOTATIONS AND BASIC EQUATIONS

Let B be an open region of three-dimensional Euclidian space occupied by the reference configuration of a dipolar body. We assume that B is regular and we denote the closure of B by $\overline{B}$. The boundary ∂B of B is closed and bounded. We use a fixed system of rectangular Cartesian axes and adopt Cartesian tensor notation. Points in B are denoted by x_j and $t \in [0,\infty)$ is time. Also, the spatial argument and the time argument of a function will be omitted when there is no likelihood of confusion. The usual summation and differentiation convection are employed: Latin subscripts

[a)]Electronic mail: m.marin@unitbv.ro

are understood to range over the integers (1,2,3); a summation over repeated subscripts is implied and a subscript preceded by a comma denotes partial differentiation with respect to the corresponding Cartesian coordinate. We also use a superposed dot to denote the partial differentiation with respect to time, t. In the following we consider the theory of thermoelasticity of dipolar bodies as it is established in the paper.[6] For convention, the notations and terminology chosen are almost identical to those of Refs. 6 and 7.

The basic equations in that theory are as follows: the equations of motion:

$$(\tau_{ij}+\sigma_{ij})_{,j}+\varrho F_i=\varrho \ddot{u}_i,$$

$$\mu_{ijk,i}+\sigma_{jk}+\varrho G_{jk}=I_{kr}\ddot{\varphi}_{jr}; \tag{1}$$

the equation of energy:

$$\varrho T_0\dot{\eta}=q_{i,i}+\varrho r;(x,t)\in B\times[0,\infty) \tag{2}$$

the constitutive equations:

$$\tau_{ij}=C_{ijmn}\varepsilon_{mn}+G_{ijmn}\gamma_{mn}+F_{mnrij}\chi_{mnr}-E_{ij}\theta,$$

$$\sigma_{ij}=G_{ijmn}\varepsilon_{mn}+B_{ijmn}\gamma_{mn}+D_{ijmnr}\chi_{mnr}-D_{ij}\theta,$$

$$\mu_{ijk}=F_{ijkmn}\varepsilon_{mn}+D_{mnijk}\gamma_{mn}+A_{ijkmnr}\chi_{mnr}-F_{ijk}\theta, \tag{3}$$

$$\varrho\,\eta=a\,\theta+E_{ij}\varepsilon_{ij}+D_{ij}\gamma_{ij}+F_{ijk}\chi_{ijk},$$

$$q_i=k_{ij}\theta_{,j}(x,t)\in B\times[0,\infty)$$

the geometric equations:

$$2\varepsilon_{ij}=u_{j,i}+u_{i,j},\gamma_{ij}=u_{j,i}-\varphi_{ij},\chi_{ijk}=\varphi_{ij,k}. \tag{4}$$

In the above equations we have used the following notations: ϱ—the constant reference density; u_i—the components of displacement; φ_{jk}—the components of dipolar displacement; $\tau_{ij},\sigma_{ij},\mu_{ijk}$—the components of stress tensors; $\varepsilon_{ij},\gamma_{ij},\chi_{ijk}$—the components of strain tensors; q_i—the components of the heat conduction vector; η—the specific entropy; T_0—the constant reference temperature; θ—the temperature variation measured from the reference temperature T_0; F_i—the components of body force per unit mass; G_{jk}—the components of body couple force per unit mass; r—the heat supply per unit mass and unit time; I_{ij}—the components of inertia; $C_{ijmn},B_{ijmn},\ldots,a$—the characteristic constants of the material and they are subject to the symmetry conditions

$$C_{ijmn}=C_{mnij}=C_{ijnm},B_{ijmn}=B_{mnij},G_{ijmn}=G_{ijnm},$$

$$F_{ijkmn}=F_{ijknm},A_{ijkmnr}=A_{mnrijk},E_{ij}=E_{ji},k_{ij}=k_{ji}. \tag{5}$$

The entropy production inequality implies that

$$k_{ij}\theta_i\theta_j\geqslant 0. \tag{6}$$

To the equations (1)–(5) we adjoin the following prescribed boundary conditions:

$$u_i(x_k,t)=0,\varphi_{jk}(x_k,t)=0,\theta(x_k,t)=0,(x_k,t)\in\partial B\times[0,\infty), \tag{7}$$

and the initial conditions

$$u_i(x_k,0)=a_i(x_k),\dot{u}_i(x_k,0)=b_i(x_k),\varphi_{jk}(x_k,0)=c_{jk}(x_k),$$
$$\dot{\varphi}_{jk}(x_k,0)=d_{jk}(x_k),\theta(x_k,0)=\theta^0(x_k),(x_k)\in B, \tag{8}$$

where a_i,b_i,c_{jk},d_{jk} and θ^0 are prescribed functions. Introducing (3) and (4) in (1) and (2), we obtain the following system:

$$\varrho\ddot{u}_i=[(C_{ijmn}+G_{ijmn})u_{n,m}+(G_{mnij}+B_{ijmn})(u_{n,m}-\varphi_{mn})$$
$$+(F_{mnrij}+D_{ijmnr})\varphi_{nr,m}-(E_{ij}+D_{ij})]\theta]_{,j}+\varrho F_i,$$

$$I_{kr}\ddot{\varphi}_{jr}=[F_{ijkmn}u_{n,m}+D_{mnijk}(u_{n,m}-\varphi_{mn})+A_{ijkmnr}\varphi_{nr,m}-F_{ijk}\theta]_{,i}$$
$$+G_{jkmn}u_{m,n}+B_{jkmn}(u_{n,m}-\varphi_{mn})+D_{jkmnr}\varphi_{nr,m}-D_{jk}\theta]+\varrho G_{jk}, \tag{9}$$

$$aT_0\dot{\theta}=-T_0[E_{ij}v_{j,i}+D_{ij}(v_{j,i}-\psi_{ij})+F_{ijk}\psi_{jk,i}]+(k_{ij}\theta_{,j})_{,i}+\varrho r,$$

where $v_i=\dot{u}_i,\psi_{jk}=\dot{\varphi}_{jk}$.

By a solution of the mixed initial boundary value problem of the dipolar thermoelasticity in the cylinder $\Omega_0=B\times[0,\infty)$ we mean an ordered array $(u_i,\varphi_{jk},\theta)$ which satisfies the system (9) for all $(x,t)\in\Omega_0$, the boundary conditions (7) and the initial conditions (8).

We shall use the following assumptions on the material properties:

(i) $\varrho>0,T_0>0,I_{ij}>0,a>0;$

(ii) $k_{ij}\xi_i\xi_j\geqslant k_0\xi_i\xi_i,k_0>0,\forall\xi_i;$

(iii) $C_{ijmn}\xi_{ij}\xi_{mn}+2G_{ijmn}\xi_{ij}\eta_{mn}+2F_{mnrij}\xi_{ij}\delta_{mnr}+B_{ijmn}\eta_{ij}\eta_{mn}$
$+2D_{ijmnr}\eta_{ij}\delta_{mnr}+A_{ijkmnr}\delta_{ijk}\delta_{mnr}$
$\geqslant a_0(\xi_{ij}\xi_{ij}+\eta_{ij}\eta_{ij}+\delta_{ijk}\delta_{ijk}),\forall\eta_{ij},\forall\delta_{ijk},\forall\xi_{ij}=\xi_{ji}$, where $a_0>0$.

The above assumptions are in agreement with the usual restrictions imposed in the mechanics of continua in order to obtain the existence and uniqueness of solutions. For instance, the condition (ii) represents a considerable strengthening of the consequence (6) of the entropy production inequality.

We shall use the vectorial notations

$$\mathbf{u}=(u_i),\mathbf{v}=(v_i),\varphi=(\varphi_{ij}),\psi=(\psi_{ij}),i,j=1,2,3.$$

Let us define

$$X=\{\mathbf{W}=(u,\mathbf{v},\varphi,\psi,\theta):\mathbf{u}\in\mathbf{H}_0^{1,3}(B),\mathbf{v}\in\mathbf{H}^{0,3}(B),$$
$$\varphi\in\mathbf{H}_0^{1,9}(B),\psi\in\mathbf{H}^{0,9}(B),\theta\in H^0(B)\}, \tag{10}$$

where $H_0^m(B)$ and $H^m(B)$ are the familiar Sobolev spaces, (see Ref. 8), and we used the notations $\mathbf{H}^{m,n}(B)=[H^m(B)]^n,\mathbf{H}_0^{m,n}(B)=[H_0^m(B)]^n$.

We wish to transform our initial-boundary value problem, defined by (9), (7) and (8) into a temporally homogeneous abstract equation in the Hilbert space X. Thus, we define the operators

$$
\begin{aligned}
A_i\mathbf{W}&=v_i,\\
B_i\mathbf{W}&=\frac{1}{\varrho}[(C_{ijmn}+G_{ijmn})u_{m,n}+(G_{mnij}+B_{ijmn})\gamma_{mn}\\
&\quad+(F_{mnrij}+D_{ijmnr})\varphi_{nr,m}-(E_{ij}+D_{ij})\theta]_{,j},\\
C_{ij}\mathbf{W}&=\psi_{ij},\\
D_{jk}\mathbf{W}&=(I_{ks})^{-1}[F_{ijsmn}u_{m,n}+D_{mnijs}\gamma_{mn}+A_{ijsmnr}\varphi_{nr,m}-F_{ijs}\theta]_{,i}\\
&\quad+G_{jkmn}u_{m,n}+B_{jkmn}\gamma_{mn}+D_{jkmnr}\varphi_{nr,m}-D_{jk}\theta,\\
E\mathbf{W}&=-\frac{1}{a}[E_{ij}v_{j,i}+D_{ij}(v_{j,i}-\psi_{ij})+F_{ijk}\psi_{jk,i}]+\frac{1}{aT_0}(k_{ij}\theta_{,j})_{,i}.
\end{aligned}
\tag{11}
$$

Let $\mathcal{L}$ be the operator

$$\mathcal{L}=(\mathbf{AW},\mathbf{BW},\mathbf{CW},\mathbf{DW},E\mathbf{W}) \tag{12}$$

where $\mathbf{A}=(A_i),\mathbf{B}=(B_i),\mathbf{C}=(C_{ij}),\mathbf{D}=(D_{ij}),i,j=1,2,3,$ with the domain

$$D=D(\mathcal{L})=\{\mathbf{W}\in X:\mathcal{L}\mathbf{W}\in X,\mathbf{v}=0,\psi=0 \ \text{ on } \ \partial B\}. \tag{13}$$

The closure of $D(\mathcal{L})$ is obviously the space X and hence $D(\mathcal{L})$ is dense in X. $D(\mathcal{L})$ is not empty; it contains at least $[C_0^\infty(B)]^{25}$. Thus, we reduce the initial-boundary value problem (9), (7), (8) to the abstract initial value problem on the Hilbert space X,

$$\frac{d\mathbf{W}}{dt}=\mathcal{L}\mathbf{W}+\mathcal{F}(t),\mathbf{W}(0)=\mathbf{W}_0,0\leqslant t\leqslant t_0, \tag{14}$$

where $\mathcal{F}(t)=(\mathbf{0},\mathbf{F},\mathbf{0},M,r),\mathbf{W}_0=(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d},\theta^0),\mathbf{F}=(F_i),\mathbf{G}=(G_{jk}),$ $\mathbf{a}=(a_i),\mathbf{b}=(b_i),\mathbf{c}=(c_{ij}),\mathbf{d}=(d_{ij})$.

III. BASIC RESULTS

Let X_* be the Hilbert space X equipped with the norm induced by the inner product,

$$
\begin{aligned}
\langle\mathbf{W},\overline{\mathbf{W}}\rangle_*=\int_B[&\varrho v_i\overline{v_i}+I_{ks}\psi_{js}\overline{\psi_{jk}}+a\theta\overline{\theta}+C_{ijmn}\varepsilon_{ij}\overline{\varepsilon_{mn}}\\
&+G_{ijmn}(\gamma_{ij}\overline{\varepsilon_{mn}}+\overline{\gamma_{ij}}\varepsilon_{mn})+F_{mnrij}(\varepsilon_{ij}\overline{\chi_{mnr}}+\overline{\varepsilon_{ij}}\chi_{mn})\\
&+B_{ijmn}\gamma_{ij}\overline{\gamma_{mn}}+D_{ijmnr}(\gamma_{ij}\overline{\chi_{mni}}+\overline{\gamma_{ij}}\chi_{mnr})+A_{ijkmnr}\chi_{ij}\overline{\chi_{mnr}}]dV.
\end{aligned}
\tag{15}
$$

By taking into account the hypotheses (i), (ii), (iii) we obtain

$$
\begin{aligned}
|\mathbf{W}|_*^2=\langle\mathbf{W},\mathbf{W}\rangle_*&=\int_B[\varrho v_iv_i+I_{ks}\psi_{js}\psi_{jk}+a\theta^2+C_{ijmn}\varepsilon_{ij}\varepsilon_{mn}+2G_{ijmn}\varepsilon_{ij}\gamma_{mn}\\
&\quad+2F_{mnrij}\varepsilon_{ij}\chi_{mnr}+B_{ijmn}\gamma_{ij}\gamma_{mn}+2D_{ijmnr}\gamma_{ij}\chi_{mnr}+A_{ijkmnr}\chi_{ijk}\chi_{mnr}]dV\\
&\geqslant\int_B[\varrho v_iv_i+I_{ks}\psi_{js}\psi_{jk}+a\theta^2+a_0(\varepsilon_{ij}\varepsilon_{ij}+\gamma_{ij}\gamma_{ij}+\chi_{ijk}\chi_{ijk})]dV\\
&\geqslant c_1|\mathbf{W}|_X^2.
\end{aligned}
\tag{16}
$$

On the other hand, by using the first Korn inequality, (see Ref. 2), and (15), we can prove that $|\mathbf{W}|_*^2 \leqslant c_2 |\mathbf{W}|_X^2$ such that, in view of (16), we have

$$c_1 |\mathbf{W}|_X^2 \leqslant |\mathbf{W}|_*^2 \leqslant c_2 |\mathbf{W}|_X^2;$$

hence the norm $|\cdot|_*$ is a norm equivalent to the original norm in X.

Lemma 1: The operator $\mathcal{L}$ is dissipative, that is,

$$\langle \mathcal{L}\mathbf{W}, \mathbf{W} \rangle_* \leqslant 0, \forall \mathbf{W} \in D(\mathcal{L}).$$

Proof: According to the relations (11) we have

$$\begin{aligned}\langle \mathcal{L}\mathbf{W}, \mathbf{W} \rangle_* = \int_B \Big\{ & v_i [(C_{ijmn} + G_{ijmn}) u_{m,n} + (G_{mnij} + B_{ijmn}) \gamma_{mn} \\ & + (F_{mnrij} + D_{ijmnr}) \varphi_{nr,m} - (E_{ij} + D_{ij}) \theta]_{,j} + \psi_{jk} [F_{ijkmn} u_{m,n} + D_{mnijk} \gamma_{mn} \\ & + A_{ijkmnr} \varphi_{nr,m} - F_{ijk} \theta]_{,i} + \psi_{jk} [G_{jkmn} u_{m,n} + B_{jkmn} \gamma_{mn} \\ & + D_{jkmnr} \varphi_{nr,m} - D_{jk} \theta] + \theta \left[\frac{1}{T_0} (k_{ij} \theta_{,j})_{,i} - E_{ij} v_{j,i} - D_{ij} (v_{j,i} - \psi_{ij}) - F_{ijk} \psi_{jk,i} \right] \\ & + C_{ijmn} u_{i,j} v_{m,n} + G_{ijmn} [u_{i,j} (v_{n,m} - \psi_{mn}) + v_{n,m} (u_{j,i} - \varphi_{ij})] \\ & + F_{mnrij} (u_{i,j} \psi_{nr,m} + v_{i,j} \varphi_{nr,m}) + B_{ijmn} (u_{j,i} - \varphi_{ij}) (v_{n,m} - \varphi_{mn}) \\ & + D_{ijmnr} [(u_{j,i} - \varphi_{ij}) \psi_{nr,m} + (v_{j,i} - \psi_{ij}) \varphi_{nr,m}] + A_{ijkmnr} \varphi_{jk,i} \psi_{nr,m} \Big\} dV.\end{aligned}$$

We now make use of the Green–Gauss formula and the boundary conditions (7), such that it results in

$$\langle \mathcal{L}\mathbf{W}, \mathbf{W} \rangle_* = -\frac{1}{T_0} \int_B k_{ij} \theta_{,i} \theta_{,j} \, dV. \tag{17}$$

On the basis of the inequality (iii), from (17) we obtain

$$\langle \mathcal{L}\mathbf{W}, \mathbf{W} \rangle_* \leqslant -\frac{k_0}{T_0} \int_B \theta_{,i} \theta_{,j} \, dV, \tag{18}$$

such that the proof of Lemma 1 is complete.

Lemma 2: The operator $\mathcal{L}$ satisfies the range condition, that is,

$$R(\lambda I - \mathcal{L}) = X, \lambda > 0. \tag{19}$$

Proof: Assume that $\hat{\mathbf{W}} = (\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\varphi}, \hat{\psi}, \hat{\theta}) \in X$. Then we must show that for all $\hat{\mathbf{W}} \in X$ the equation

$$\lambda \mathbf{W} - \mathcal{L}\mathbf{W} = \hat{\mathbf{W}} \tag{20}$$

has at least a solution $\mathbf{W}$ in $D(\mathcal{L})$. By eliminating the functions v_i and ψ_i in (20), we obtain the following system of equations in the variables u_i, v_i and θ:

$$\mathcal{L}_i\omega=\lambda^2u_i-\frac{1}{\varrho}[(C_{ijmn}+G_{ijmn})u_{m,n}+(G_{mnij}+B_{ijmn})\gamma_{mn}$$
$$+(F_{mnrij}+D_{ijmnr})\varphi_{nr,m}-(E_{ij}+D_{ij})\theta]_{,j}=g_i\,,$$
$$\mathcal{L}_{j+k+3}\omega=\lambda^2\varphi_{jk}-(I_{ks})^{-1}[F_{ijsmn}u_{m,n}+D_{mnijs}\gamma_{mn}+A_{ijsmnr}\varphi_{ns,m}-F_{ijs}\theta]_{,i}$$
$$+B_{jkmn}\gamma_{mn}+D_{jkmnr}\varphi_{nr,m}-D_{jk}\theta=g_{j+k+3}\,, \tag{21}$$
$$\mathcal{L}_{13}\omega=\lambda^2\theta-\frac{1}{aT_0}(k_{ij}\theta_{,j})_{,i}-\frac{1}{a}[E_{ij}v_{j,i}+D_{ij}(v_{j,i}-\psi_{ij})+F_{ijk}\psi_{jk,i}]=g_{13}\,,$$

where

$$\omega=(\mathbf{u},\varphi,\theta),g_i=\lambda\hat{u}_i+\hat{v}_i\,,g_{j+k+3}=\lambda\hat{\varphi}_{jk}+\hat{\psi}_{jk}\,,$$
$$g_{13}=\hat{\theta}+\frac{1}{a}[E_{ij}\hat{v}_{j,i}+D_{ij}(\hat{v}_{j,i}-\hat{\psi}_{ij})+F_{ijk}\hat{\psi}_{jk,i}]. \tag{22}$$

Let $\langle\cdot,\cdot\rangle$ denote the conveniently weighted $[L_2(B)]^{13}$ inner product and consider the bilinear form

$$Q[\omega,\overline{\omega}]=\langle\mathbf{L}\omega,\overline{\omega}\rangle=\langle(\mathcal{L}_i\omega,\mathcal{L}_{j+k+3}\omega,\mathcal{L}_{13}\omega),(\overline{u_i},\overline{\varphi_{jk}},\overline{\theta})\rangle$$
$$=\int_B\left[\varrho\overline{u_i}\mathcal{L}_i\omega+I_{ks}\overline{\psi_{js}}\mathcal{L}_{j+k+3}\omega+\frac{a}{\lambda}\overline{\theta}\mathcal{L}_{13}\omega\right]dV. \tag{23}$$

Using the Green–Gauss formula and the boundary conditions (7), it results that

$$Q[\omega,\omega]=\int_B[\varrho v_iv_i+I_{ks}\psi_{js}\psi_{jk}+a\theta^2+C_{ijmn}u_{i,j}u_{m,n}$$
$$+2G_{ijmn}(u_{j,i}-\varphi_{ij})u_{m,n}+2F_{mnrij}(u_{j,i}-\varphi_{ij})\varphi_{nr,m}$$
$$+B_{ijmn}(u_{j,i}-\varphi_{ij})(u_{n,m}-\varphi_{mn})+2D_{ijmnr}(u_{j,i}-\varphi_{ij})\varphi_{nr,m}$$
$$+A_{ijkmnr}\varphi_{jk,i}\varphi_{nr,m}]dV+\frac{1}{\lambda T_0}\int_B k_{ij}\theta_{,i}\theta_{,j}dV, \tag{24}$$

for any $\omega=(\mathbf{u},\varphi,\theta)\in Y, Y\equiv\mathbf{H}_0^{1,3}(B)\times\mathbf{H}_0^{1,9}(B)\times H_0^1(B)$.

Due to the hypotheses (i), (ii), (iii) and the first Korn inequality, it follows that

$$Q[\omega,\omega]\geqslant C_1|\omega|_Y^2,\quad \text{for all}\ \ \omega=(\mathbf{u},\varphi,\theta)\in Y, \tag{25}$$

C_1 is a positive, conveniently chosen, constant and the norm $|\omega|_Y$ is defined by

$$|\omega|_Y=|(\mathbf{u},\varphi,\theta)|_Y=|\mathbf{u}|_{\mathbf{H}^1(B)}+|\varphi|_{\mathbf{H}^1(B)}+|\theta|_{H^1(B)}\,.$$

In the usual way, we can prove that $Q[\omega,\omega]\leqslant C_1|\omega|_Y^2$, hence the bilinear form $Q[\omega,\overline{\omega}]$ determines a norm equivalent to the original norm in Y. Since the bilinear form $Q[\omega,\overline{\omega}]$ is continuous in $Y\times Y$ we deduce that there exists a linear bounded transformation T from Y into itself such that we have

$$Q[\omega,\overline{\omega}]=\langle\omega,T\omega\rangle_Y,\quad \text{for any}\ \ (\omega,\overline{\omega})\in Y\times Y. \tag{26}$$

Since

$$\langle \omega, T\omega\rangle_Y = Q[\omega,\omega] \geqslant C_1 |\omega|_Y^2, \tag{27}$$

we deduce that

$$|T\omega|_Y \geqslant C_1 |\omega|_Y, \omega \in Y. \tag{28}$$

Let $R(T)$ be the range of T. The linear transformation T is one to one. We need to prove that $T\omega = 0$ implies that $\omega = 0$. Indeed, if there is $\omega_0 \in Y$ such $T\omega_0 = 0$, then (26) implies $Q[\omega_0, \omega_0] = 0$, and then the inequality (25) proves that $\omega_0 = 0$. Therefore, there exists $T^{-1}: R(T) \to Y$. Now we prove that $R(T)$ is dense in Y. We assume to the contrary that there is $\omega_0 \in Y \backslash R(T)$, $0\omega_0 \neq 0$ such that $\langle \omega_0, T\overline{\omega}\rangle_Y = 0$ for any $\overline{\omega} \in Y$. But from (26) we deduce that $Q[\omega_0, \omega_0] = 0$ such that with the aid of (25) we deduce that $\omega_0 = 0$. This contradicts the initial assumptions and therefore we obtain that $R(T)$ is dense in Y. So we can continue T^{-1} to Y, such that

$$T^{-1}: Y \to Y \quad \text{and} \quad |T^{-1}| \leqslant C_1^{-1}.$$

Let $\mathbf{z}$ be in $R(T)$ and ω the only function in Y such that $\mathbf{z} = T\omega$. We define the functional $\mathcal{K}$ by $\mathcal{K}(\mathbf{z}) = \langle \mathbf{g}, \omega\rangle$. Obviously, we have

$$|\mathcal{K}(\mathbf{z})| \leqslant |\mathbf{g}|_{\mathbf{H}_0^{-1}(B)} |\omega|_Y \leqslant C_1^{-1} |\mathbf{g}|_{\mathbf{H}_0^{-1}(B)} |\mathbf{z}|_Y,$$

and then we deduce that $\mathcal{K}$ is a linear bounded functional defined over $R(T)$ such that

$$|\mathcal{K}| \leqslant C_1^{-1} |\mathbf{g}|_{\mathbf{H}_0^{-1}(B)}.$$

We can continue $\mathcal{K}$ in the whole space Y, in such a way that the continued functional $\mathcal{K}$ shall have the same norm. On the other hand, since Y is a Hilbert space, the Riesz–Frechet theorem, (see Ref. 8), proves that there exists a unique $\omega \in Y$ such that

$$\mathcal{K}(\tilde{\omega}) = \langle \omega, \tilde{\omega}\rangle_Y, \quad \text{for any } \tilde{\omega} \in Y; \ |\omega|_Y = |\mathcal{K}| \leqslant C_1^{-1} |\mathbf{g}|_{\mathbf{H}_0^{-1}(B)}. \tag{29}$$

If we choose $\tilde{\omega} = T\overline{\omega}$, then from (26) and (29), it follows that the unique $\omega \in Y$ satisfies the equation

$$Q[\omega, \overline{\omega}] = \langle \mathbf{g}, \tilde{\omega}\rangle, \quad \text{for all } \overline{\omega} \in Y. \tag{30}$$

From the relations $\lambda u_i - \hat{u}_i = v_i, \lambda\varphi_i - \hat{\varphi}_{jk} = \psi_{jk}$ and $\lambda\theta - \hat{\theta} = \tau$, it follows that $\mathbf{v} \in \mathbf{H}_0^{1,3}(B), \psi \in \mathbf{H}_0^{1,9}(B)$ and $\tau \in H_0^1(B)$. Therefore we deduce that $\mathbf{W} = (\mathbf{u}, \mathbf{v}, \varphi, \psi, \theta)$ is in $D(\mathcal{L})$ and the proof of Lemma 2 is complete.

Theorem 1: The operator $\mathcal{L}$ defined by the relations (12) *generates a C_0-semigroup of contractions on X.*

Proof: This result follows immediately from the Lummer–Phillips theorem, (see Ref. 9, for instance).

In order to study the existence and uniqueness of the solution for the inhomogeneous equation (14), we use the following result.

Theorem 2: Let $\mathcal{L}$ be the infinitesimal generator of a C_0-contractive semigroup $T(t)$ on X. If *$\mathcal{F}(s)$ is continuously differentiable on $[0,t_0]$, then the initial value problem* (14) *has, for every $\mathbf{W}_0 \in D(\mathcal{L})$, the unique solution*

$$\mathbf{W}(t) = T(t)\mathbf{W}_0 + \int_0^1 T(t-s)\mathcal{F}(s)\,ds, t \in [0,t_0], \tag{31}$$

such that $\mathbf{W}(t)\in C^1([0,t_0];X)\cap C^0([0,t_0];D(\mathcal{L}))$.

On the basis of the above theorem, we deduce the following.

Theorem 3: Let us assume that the thermoelastic coefficients, which are continuously differentiable, satisfy the conditions (i), (ii), (iii). Moreover, we assume that $\mathbf{F}\in C^1([0,t_0];\mathbf{L}_2(B))$, $\mathbf{M}\in C^1([0,t_0];\mathbf{L}_2(B))$, $r\in C^1([0,t_0];L_2(B))$ *and* $\mathbf{W}_0=(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d},\theta^0)\in D(\mathcal{L})$.

Then there exists an unique solution of the initial-boundary value problem (9), (7), (8) *such that*

$$(\mathbf{u},\dot{\mathbf{u}},\varphi,\dot{\varphi},\theta)\in[C^1([0,t_0];X)\cap C^0([0,t_0];D(\mathcal{L}))]^{13}.$$

The following theorem establish the continuous dependence of the solution of our problem upon the initial data and supply terms. Let (u_i,φ_i,θ) be the difference of two solutions of the problem defined by (9), (7), (8) but corresponding to the differences of the initial data and to the differences of body forces, body couples and heat supplies, $\mathbf{W}_0=(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d},\theta^0)$, $(\mathbf{F},\mathbf{M},r)$, respectively.

Theorem 4: Let us assume that the thermoelastic coefficients, which are continuously differentiable, satisfy the conditions (i), (ii), (iii). Moreover, we assume that $\mathbf{F}\in C^1([0,t_0];\mathbf{L}_2(B))$, $\mathbf{M}\in C^1([0,t_0];\mathbf{L}_2(B))$, $r\in C^1([0,t_0];L_2(B))$ *and* $\mathbf{a}\in\mathbf{H}^1(B),\mathbf{b}\in\mathbf{H}^0(B)$, $\mathbf{c}\in\mathbf{H}^1(B),\mathbf{d}\in\mathbf{H}^0(B)\,\theta\in H^1(B)$.

If $(\mathbf{u},\varphi,\theta)$ *is the difference of two solutions of the problem* (9), (7), (8) , *then there exists a positive constant M such that*

$$|\mathbf{u}|_{\mathbf{H}^{1,3}(B)}+|\dot{\mathbf{u}}|_{\mathbf{H}^{0,3}(B)}+|\varphi|_{\mathbf{H}^{1,9}(B)}+|\bar{\varphi}|_{\mathbf{H}^{0,9}(B)}+|\theta|_{H^0(B)}$$
$$\leq M\Bigg\{|\mathbf{a}|_{\mathbf{H}^{1,3}(B)}+|\mathbf{b}|_{\mathbf{H}^{0,3}(B)}+|\mathbf{c}|_{\mathbf{H}^{1,9}(B)}+|\mathbf{d}|_{\mathbf{H}^{0,9}(B)}+|\theta^0|_{H^0(B)}$$
$$+\int_0^t[|\mathbf{F}(s)|_{\mathbf{H}^{0,3}(B)}+|\mathbf{G}(s)|_{\mathbf{H}^{0,9}(B)}+|r(s)|_{H^0(B)}]ds\Bigg\}. \tag{32}$$

Proof: On the basis of the equations (9), (7), (8) we can deduce the following identity:

$$\int_B[\varrho\dot{u}_i\dot{u}_i+I_{ks}\dot{\varphi}_{js}\dot{\varphi}_{jk}+a\theta^2+C_{ijmn}u_{i,j}u_{m,n}+2G_{ijnm}(u_{j,i}-\varphi_{ij})u_{m,n}$$
$$+F_{mnrij}(u_{j,i}-\varphi_{ij})\varphi_{nr,m}+B_{ijmn}(u_{j,i}-\varphi_{ij})(u_{n,m}-\varphi_{mn})$$
$$+2D_{ijmnr}(u_{j,i}-\varphi_{ij})\varphi_{nr,m}+A_{ijkmnr}\varphi_{jk,i}\varphi_{nr,m}]dV+\int_B\frac{1}{\lambda T_0}k_{ij}\theta_{,i}\theta_{,j}dV$$
$$=\int_B[\varrho\dot{a}_i\dot{a}_i+I_{ks}\dot{c}_{js}\dot{c}_{jk}+a(\theta^0)^2+C_{ijmn}a_{i,j}a_{m,n}+2G_{ijnm}(a_{j,i}-c_{ij})a_{m,n}$$
$$+2F_{mnrij}(a_{j,i}-c_{ij})c_{nr,m}+2B_{ijmn}(a_{j,i}-c_{ij})(a_{n,m}-c_{mn})$$
$$+2D_{ijmnr}(a_{j,i}-c_{ij})c_{nr,m}+A_{ijkmnr}c_{jk,i}c_{nr,m}]dV$$
$$+2\int_0^t\int_B\left[F_iu_i+G_{jk}\varphi_{jk}+\frac{1}{T_0}r\theta\right]dVds,s\in[0,t_0]. \tag{33}$$

By using the Schwarz's inequality, the hypotheses (i), (ii), (iii) and the first Korn's inequality, from the identity (33) we deduce a Gronwall inequality that demonstrates the estimate (32).

Remark: A similar procedure can be used in the case when the boundary conditions (7) are replaced by the other boundary conditions and the above results are still valid.

[1] A. C. Eringen, *Continuum Physics* (Academic, New York, 1976).
[2] I. Hlavacek and J. Necas, Arch. Ration. Mech. Anal. **36**, 305 (1970).
[3] R. D. Mindlin, Arch. Ration. Mech. Anal. **16**, 51 (1964).
[4] R. Quintanilla, J. Therm. Stresses **15**, 393 (1992).
[5] M. Ciarletta and A. Scalia, J. Elast. **25**, 149 (1991).
[6] M. Marin, Int. J. Eng. Sci. **8**, 1229 (1994).
[7] M. Marin, C. R. Acad. Sci., Ser. IIb: Mec., Phys., Chim., Astron. **321**, 475 (1995).
[8] K. Yosida, *Functional Analysis* (Springer-Verlag, Berlin, 1978).
[9] A. Pazy, Semigroups of Linear Operators (Springer-Verlag, Berlin, 1983).

Transfer matrices for scalar fields on curved spaces

E. Prodan
University of Houston, 4800 Calhoun Road, Houston, Texas 77204-5508

(Received 13 August 1998; accepted for publication 23 November 1998)

We apply Nelson's technique of constructing Euclidean fields to the case of classical scalar fields on curved spaces. It is shown how to construct a transfer matrix and, for a class of metrics, the basic spectral properties of its generator are investigated. An application concerning the decoupling of two non-convex disjoint regions is given. © *1999 American Institute of Physics.* [S0022-2488(99)00703-3]

I. INTRODUCTION

We start our construction from the ideas comprised in Nelson's axioms[1] for scalar Euclidean–Markoff quantum fields. Here, the Markoff property of certain projectors is one of the basic ingredients in defining the transfer matrix in which the generator is identified with the Hamiltonian of Wightman quantum scalar field. We found that these ideas can be used in the same way at the nonquantum level. In the case of the scalar fields on Riemannian manifolds, for an arbitrary direction, we construct a propagator by using the Markoff property. In the stationary case it becomes a semigroup which can be considered as the transfer matrix of the system and, further, it can be used in introducing a Hamiltonian. We will show that the propagator is exponentially bounded by using Agmon's[2] results in exponential decay of solutions of second-order elliptic equations. An application concerning the decoupling (in the sense of Ref. 3) of two disjoint nonconvex regions is given.

II. INTRODUCTORY DEFINITIONS AND RESULTS

Let us consider the Riemannian manifold (R^{n+1},g) and the Laplace–Beltrami operator on it, Δ. For a point in R^{n+1} we use the notation (t,x). Let $E_m(t,x;s,y)$ be the kernel of $(\Delta+m^2)^{-1}$ on $L^2(R^{n+1},\sqrt{g}\,dt\,dx)$. As in Ref. 4, we will not consider the additional term $\frac{1}{6}\rho$. One defines the space $N\subset\mathcal{D}'(R^{n+1})$, $f\in N$ if:

$$\|f\|_N^2=\int_{R^{n+1}}\int_{R^{n+1}}\overline{f}(t,x)E_m(t,x;s,y)f(s,y)\sqrt{g(t,x)}\sqrt{g(s,y)}dt\,dx\,ds\,dy<\infty, \tag{1}$$

and, for each $\sigma\in R$, let $N_\sigma\subset D'(R^n)$ be the space: $g\in N_\sigma$ if

$$\|g\|_{N_\sigma}^2=\int_{R^n}\overline{g}(x)E_m(\sigma,x;\sigma,y)g(y)\sqrt{g(\sigma,x)}\sqrt{g(\sigma,y)}dx\,dy<\infty. \tag{2}$$

We will consider that, as in the Euclidean case, the space $L^2(R^n,d\mu_\sigma)\subset N_\sigma$, where $d\mu_\sigma(x)=\sqrt{g(\sigma,x)}\,d^nx$ and that it is dense in N_σ for each $\sigma\in R$. Now, let $\hat{E}_\sigma:N_\sigma\to L^2(R^n,d\mu_\sigma)$ be the operator corresponding to the kernel $E_m(\sigma,x;\sigma,y)$. Then $\hat{E}_\sigma^{1/2}$ defines an isometry from N_σ to $L^2(R^n,d\mu_\sigma)$ and let $(\hat{E}_\sigma^{1/2})^\dagger:L^2(R^n,d\mu_\sigma)\to N_\sigma$ be its adjoint. The following are true:

$$\hat{E}_\sigma^{1/2}\circ(\hat{E}_\sigma^{1/2})^\dagger=1_{L^2(R^n,d\mu_\sigma)}\quad\text{and}\quad(\hat{E}_\sigma^{1/2})^\dagger\circ\hat{E}_\sigma^{1/2}=1_{N_\sigma}. \tag{3}$$

With our assumptions, $\hat{E}_\sigma^{1/2}(N_\sigma)=L^2(R^n,d\mu_\sigma)\subset N_\sigma$, the operator $\hat{E}_\sigma^{1/2}$ is bounded on N_σ. Moreover, one can view $(\hat{E}_\sigma)^\dagger$ as a dense defined unbounded operator on N_σ, in fact, it is the inverse operator of $\hat{E}_\sigma$.

For $\sigma\in R$, let j_σ be the operator $j_\sigma:N_\sigma\to N$, $(j_\sigma\psi)(t,x)=\psi(x)\delta(t-\sigma)$ and j_σ^* be its adjoint. If Λ is a closed subset of R^{n+1} we denote by N_Λ the subspace of N which comprises all distributions with support in Λ. The orthogonal projection of N in N_Λ will be denoted by e_Λ. Following Ref. 5 we have:

Proposition 1: The operators j_σ are isometries and $j_\sigma^* j_\sigma=1_{N_\sigma}$, $j_\sigma j_\sigma^*=e_\sigma$, where e_σ denotes the projector corresponding to the subset of R^{n+1}, $t=\sigma$.

Then we define the operators:

$$U_{\sigma,\sigma'}:N_{\sigma'}\to N_\sigma, \quad U_{\sigma,\sigma'}=j_\sigma^*\circ j_{\sigma'}. \tag{4}$$

We will derive in the following that $U_{\sigma,\sigma'}$ are propagators in the sense of Ref. 6. This will follow from the Markoff property of the projectors e_σ.

Lemma 2: Let A, B, and C be closed subsets in R^{n+1} such that C separates A and B. Then $e_A\circ e_C\circ e_B=e_A\circ e_B$.

Solution 3: This is the consequence of the fact that E_m is the kernel of a local operator. The proof is identical with that of Ref. 5.

The basics properties of $U_{\sigma,\sigma'}$ operators are stated in the following proposition.

Proposition 4: The family of operators $U_{\sigma,\sigma'}$, $\sigma,\sigma'\in R$ has the following properties:

(1) $U_{\sigma,\sigma'}\circ U_{\sigma',\sigma''}=U_{\sigma,\sigma''}$,

(2) $U_{\sigma,\sigma}=1_{N_\sigma}$,

(3) $\|U_{\sigma,\sigma'}\|\leqslant 1$.

Solution 5: (1) *Using the Markoff property we have:*

$$e_\sigma\circ e_{\sigma'}\circ e_{\sigma''}=e_\sigma e_{\sigma''}\Leftrightarrow j_\sigma\circ j_\sigma^*\circ j_{\sigma'}\circ j_{\sigma'}^*\circ j_{\sigma''}\circ j_{\sigma''}^*=j_\sigma\circ j_\sigma^*\circ j_{\sigma''}\circ j_{\sigma''}^*. \tag{5}$$

By composition with $j_{\sigma''}$ at the right, we have

$$j_\sigma\circ(j_\sigma^*\circ j_{\sigma'}\circ j_{\sigma'}^*\circ j_{\sigma''}-j_\sigma^*\circ j_{\sigma''})=0. \tag{6}$$

From the definition of $U_{\sigma,\sigma'}$ and since j_σ are isometries, we conclude $U_{\sigma,\sigma'}U_{\sigma',\sigma''}=U_{\sigma,\sigma''}$.

(2) *It follows from proposition 1 and definition of $U_{\sigma,\sigma'}$.*

(3) *Because j_σ^* and j_σ are isometries, the property results immediately.*

III. EXPONENTIAL BOUNDS ON PROPAGATORS

To improve our estimates on the propagators $U_{\sigma,\sigma'}$ we need a supplementary condition on the metric g. We say that an application $Q:R^{n+1}\to M(n+1,n+1)$ has stable positivity if there exists $\epsilon>0$ such that for any application $\delta:R^{n+1}\to M(n+1,n+1)$ with $|\delta(x)^{ij}|\leqslant\epsilon$ the matrices $Q(x)-\delta(x)$ are positively defined for any $x\in R^{n+1}$. The following result is a direct application of Agmon theory[2] of exponentially decay of solutions of elliptic second-order operators.

Proposition 6: If the metric g has stable positivity then for any $f\in N_{\sigma'}$:

$$\int_{T_0}^\infty d\sigma\{e^{\omega\sigma}\|\hat{E}_\sigma^{1/2}\circ U_{\sigma,\sigma'}f\|_{N_\sigma}\}^2<\infty, \tag{7}$$

provided $\omega<m/\sqrt{\sup g^{11}}$.

Solution 7: Starting from

$$\begin{aligned}\langle u,U_{\sigma,\sigma'}f\rangle_{N_\sigma}&=\langle u,\hat{E}_\sigma\circ U_{\sigma,\sigma'}f\rangle_{L^2(R^n,d\mu_\sigma)}\\&=\int_{R^n}\overline{u}(x)\left[\int_{R^n}E_m(\sigma,x;\sigma',y)f(y)d\mu_{\sigma'}(y)\right]d\mu_\sigma(x)\end{aligned} \tag{8}$$

for $u\in N_\sigma$ and $f\in N_{\sigma'}$, it follows that $\varphi(\sigma,x)=(\hat{E}_\sigma\circ U_{\sigma,\sigma'}f)(x)$ is a solution of

$$(\Delta+m^2)\varphi(\sigma,x)=0 \tag{9}$$

for $\sigma>\sigma'$. *Let* $\rho_m\,(\cdot;\cdot)$ *denote the distance corresponding to the metric* $g_m=mg$. *The metric* g *has stable positivity so, there is an* $\epsilon\in R_+$ *such that* $\rho_m(\sigma_0,x_0;\sigma,x)>(\epsilon/m)|\sigma-\sigma_0|$. *For* $\Omega=\{(\sigma,x):\sigma>T_0\}$, $T_0\in R_+$ *and for some positive* λ:

$$\begin{aligned}\int_\Omega|\varphi(\sigma,x)|^2e^{-\lambda\rho_m(T_0,x_0;\sigma,x)}\sqrt{g(\sigma,x)}d\sigma\,d^nx\\
&=\int_{T_0}^\infty d\sigma\langle\hat{E}_\sigma\circ U_{\sigma,\sigma'}f,\hat{E}_\sigma\circ U_{\sigma,\sigma'}f\rangle_{L^2(R^n,d\mu_\sigma)}e^{-\lambda(\epsilon/m)(\sigma-T_0)}\\
&<ct.\int_{T_0}^\infty d\sigma\langle U_{\sigma,\sigma'}f,\hat{E}_\sigma\circ U_{\sigma,\sigma'}f\rangle_{L^2(R^n,d\mu_\sigma)}e^{-\lambda(\epsilon/m)(\sigma-T_0)}\\
&=ct.\int_{T_0}^\infty d\sigma\|U_{\sigma,\sigma'}f\|^2_{N_\sigma}e^{-\lambda(\epsilon/m)(\sigma-T_0)}<\infty.\end{aligned} \tag{10}$$

So we are in the conditions of the main theorem of Ref. 2. It follows that:

$$\begin{aligned}&\int_\Omega d\sigma\,d^nx\sqrt{g(\sigma,x)}|\varphi(\sigma,x)|^2(m^2-g(\nabla h(\sigma,x),\nabla h(\sigma,x)))e^{2h(\sigma,x)}\\
&\leqslant\frac{2(1+2d)}{d^2}m^2\int_{\Omega\backslash\Omega_d}|\varphi(\sigma,x)|^2e^{2h(\sigma,x)}\sqrt{g(\sigma,x)}dx,\end{aligned} \tag{11}$$

where d *is a positive number and* $\Omega_d=\{(\sigma,x)\in\Omega:\rho_m((\sigma,x),\{\infty\})>d\}$. *Here*

$$\rho_m((\sigma,x),\{\infty\})=\sup\{\rho_m((\sigma,x),\Omega\backslash K):K\text{ is a compact subset of }\Omega\}. \tag{12}$$

The function h *is any function which satisfies the condition* $g(\nabla h(\sigma,x),\nabla h(\sigma,x))<m^2$. *We choose* $h(\sigma,x)=\omega\sigma$ *with* $\omega<m/\sqrt{\sup g^{11}}$. *The above inequality becomes*

$$\int_\Omega d\sigma\,d^nx\sqrt{g(\sigma,x)}|\varphi(\sigma,x)|^2e^{2\omega\sigma}<\frac{2(1+2d)}{d^2}\,\frac{m^2}{m^2-g^{11}\omega^2}\int_{\Omega\backslash\Omega_d}d\sigma\,dx\sqrt{g(\sigma,x)}|\varphi(\sigma,x)|^2e^{2\omega\sigma}. \tag{13}$$

If for any point $(\sigma,x)\in\Omega$ *there is a geodesic which starts in* (σ,x) *and ends in the hyperplane* $\sigma=T_0$ *then* $\Omega\backslash\Omega_d\subset\{(\tau,x):0<\sigma\leqslant T\}$ *with* T *sufficiently large but finite. In conclusion,*

$$\int_\Omega d\sigma\,d^nx\sqrt{g(\sigma,x)}|\varphi(\tau,x)|^2e^{2\omega\sigma}=\int_{T_0}^\infty d\sigma\,e^{2\omega\sigma}\langle\hat{E}_\sigma\circ U_{\sigma,\sigma'}f,\hat{E}_\sigma\circ U_{\sigma,\sigma'}f\rangle_{L^2(R^n,\mu_\sigma)}<\infty, \tag{14}$$

or

$$\int_{T_0}^\infty d\sigma\,e^{2\omega\sigma}\langle\hat{E}_\sigma\circ U_{\sigma,\sigma'}f,\hat{E}_\sigma\circ U_{\sigma,\sigma'}f\rangle_{L^2(R^n,\mu_\sigma)}<\infty, \tag{15}$$

which implies

$$\int_{T_0}^\infty d\sigma\{e^{\omega\sigma}\|\hat{E}_\sigma^{1/2}\circ U_{\sigma,\sigma'}f\|_{N_\sigma}\}^2<\infty. \tag{16}$$

IV. THE STATIONARY CASE

We consider in this section that there is a coordinate system such that the metric g is independent of the first coordinate. In this case, the spaces N_σ and the operators $\hat{E}_\sigma^{1/2}$ are identical and will be denoted by N_0 and $\hat{E}_0^{1/2}$, respectively. Thus, the operators $U_{\sigma,\sigma'}$ are defined on the same Hilbert space and depend only on the difference $\sigma-\sigma' : U_{\sigma,\sigma'}=U_{\sigma-\sigma'}$. The family of operators $\{U_\tau\}_{\tau\in R_+}$ forms a semigroup. Using the results about existence and properties of the generators of semigroups,[7] we can obtain bounds directly on the transfer matrix U_τ.

Proposition 8: The semigroup $\{U_\tau\}_{\tau\in R_+}$ *is exponentially bounded:* $\|U_\tau\|_{N_0}<e^{-\tau\omega}$ *provided* $\omega<m/\sqrt{\sup g}$[11].

Solution 9: Because we have found estimates on $\hat{E}_0^{1/2}\circ U_\tau$, *we will consider the operators* $\tilde{U}_\tau=\hat{E}_0^{1/2}\circ U_\tau\circ(\hat{E}_0^{1/2})^\dagger$, *well defined on* $L^2(R^n,d\mu_0)$. *Using the fact that* $L^2(R^n,d\mu_0)$ *is dense in* N_0 *we can extend these operators by continuity on the space* N_0. *In this way we have built the semigroup* $\{\tilde{U}_\tau\}_{\tau\in R_+}$ *which satisfies the estimates of the precedent section:*

$$\int_{T_0}^{\infty} d\tau\{e^{\omega\tau}\|\tilde{U}_\tau\|_{N_0}\}^2<\infty, \tag{17}$$

for some $T_0>0$. *So* $\{\tilde{U}_\tau\}_{\tau\in R_+}$ *is exponentially bounded and in consequence,*[7] *if* $\tilde{K}$ *is its generator* $(\tilde{U}_\tau=e^{-\tau\tilde{K}})$ *the resolvent set of* $\tilde{K}$ *satisfies:*

$$\{z\in C|\mathrm{Re}\, z\in(-\infty,\omega)\}\subset\rho(\tilde{K}). \tag{18}$$

If K *is the generator of* $\{U_\tau\}_{\tau\in R_+}$ *then, on* $\mathcal{D}(K)$ *we have:*

$$K=(\hat{E}_0^{1/2})^\dagger\circ\tilde{K}\circ\hat{E}_0^{1/2} \tag{19}$$

by using the reciprocal formula

$$U_\tau=(\hat{E}_0^{1/2})^\dagger\circ\tilde{U}_\tau\circ\hat{E}_0^{1/2}, \tag{20}$$

valid on N_0. *If the operator*

$$(\hat{E}_0^{1/2})^\dagger\circ(\tilde{K}-z)^{-1}\circ\hat{E}_0^{1/2} \tag{21}$$

is well defined, even on a dense subset of N_0, *then* $K-z$ *is inversable.*

From (20) it follows that, if $(\tilde{K}-z)^{-1}$ exists, then

$$(\tilde{K}-z)^{-1}(L^2(R^n,d\mu_0))\subset L^2(R^n,d\mu_0), \tag{22}$$

and in consequence $(\hat{E}_0^{1/2})^\dagger\circ(\tilde{K}-z)^{-1}\circ\hat{E}_0^{1/2}$ is well defined on the entire N_0. Will follow that $\rho(\tilde{K})\subset\rho(K)$ and this ends the proof.

If the metric is symmetric at transformation $x^1\to -x^1$, the transfer matrix generator is self-adjoint and it can be considered as the Hamiltonian of the scalar field.

V. APPLICATION

Our application is for the Euclidean case. The results concerning decoupling of different regions in quantum Euclidean fields are based primarily on estimates of $\|e_{\Lambda_1}e_{\Lambda_2}\|_N$, where Λ_1,Λ_2 are two disjoint regions. Let us consider the two dimensional case. The most difficult case is when Λ_1,Λ_2 are not convex and there is no possibility of drawing a straight line between the two subsets. We can sharpen the existent estimates[5] for these cases by using the previous results. The idea is to make a change of coordinates such that for the new coordinates, lines like $\sigma=ct$.

separate the two sets and they are as close as possible to the boundaries of Λ_1,Λ_2. Then we can use the exponential bounds of the previous section to evaluate $\|e_{\Lambda_1}e_{\Lambda_2}\|_N$. More precisely:

Proposition 10: Let Λ_1,Λ_2 two regions in R^2 such that the construction of the coordinates 26 to be possible (after a rotation if necessary). Then

$$\|e_{\lambda_1}\circ e_{\Lambda_2}\|_N \leqslant e^{-m|\beta-\alpha|\min|\cos\theta|}, \tag{23}$$

where θ and $|\beta-\alpha|$ will be defined during the proof.

Solution 11: Let (t,x) denote the original coordinates in which the metric is diagonal. Let $\gamma{:}R\rightarrow R^2$ be a curve which separates Λ_1,Λ_2 and $\gamma(0)=(t=0,x=0)$. We define a new coordinate system (σ,ξ) by

$$\begin{aligned} t(\sigma,\xi)&=\sigma+\gamma^1(\xi),\\ x(\sigma,\xi)&=\gamma^2(\xi). \end{aligned} \tag{24}$$

In the new coordinates, the metric is

$$g'(\sigma,\xi)=\begin{pmatrix} 1 & \dfrac{d\gamma^1}{d\xi} \\ \dfrac{d\gamma^1}{d\xi} & \left(\dfrac{d\gamma^1}{d\xi}\right)^2+\left(\dfrac{d\gamma^2}{d\xi}\right)^2 \end{pmatrix} \tag{25}$$

so we are in the conditions of the last section. Using the Markoff property,

$$\|e_{\Lambda_1}\circ e_{\Lambda_2}\|_N=\|e_{\Lambda_1}\circ e_\alpha\circ e_\beta\circ e_{\Lambda_2}\|_N \leqslant \|e_\alpha\circ e_\beta\|_N, \tag{26}$$

where the lines $\sigma=\alpha$, $\sigma=\beta$ separate Λ_1 and Λ_2 exactly in the order they appear in the above relation (in the sense that $\sigma=\alpha$ separates Λ_1 by $\sigma=\beta$, etc.). Further,

$$\|j_\alpha\circ j_\alpha^\dagger\circ j_\beta\circ j_\beta^\dagger\|_N=\|j_\alpha\circ U_{\alpha-\beta}\circ j_\beta^\dagger\|_N=\|U_{\alpha-\beta}\|_{N_0}. \tag{27}$$

The element $(g')^{11}$ is given by $(g')^{11}=1/\cos^2\theta$, where θ is the angle between the tangent to the curve γ and the x axis. Using the bounds of the last section we have

$$\|e_{\Lambda_1}\circ e_{\Lambda_2}\|_N \leqslant e^{-m|\beta-\alpha|\min|\cos\theta|}. \tag{28}$$

Performing first a rotation, one can choose the best values for $|\beta-\alpha|$ and $\min|\cos\theta|$.

VI. CONCLUSIONS

Our primary goal was to define the transfer matrix for scalar fields on curved spaces and to investigate the basic spectral properties of its generator. Even though the generator is not self-adjoint in the general case, this approach allows us to investigate this problem by using at least two new tools besides the methods of Green's functions. One is the perturbations of hypercontractive semigroups[8] and the other is the adiabatic theorem.

Now it is straightforward to quantize the field by defining the Markoff field over the space *N*. For the stationary, symmetric at time reflection case (static), we think that one now has all elements to construct the physical field (for example that proposed in Ref. 4) by following the Nelson reconstruction method and the holomorphic continuation of the transfer matrix. Note that, according to the results of Ref. 6, the holomorphic continuation of the transfer matrix to real time is still possible in the stationary case without symmetry at time reflection, as long as the spectrum of the generator belongs to the real axis. Of course, one has to check that the results of Ref. 9 (systematized in Ref. 5), which are the core of the reconstruction theorem, are still valid. For the general case, we think that the adiabatic theorem, especially the adiabatic reduction theory,[10] may play an important role in defining the physical quantum field by following Nelson's approach.

[1] E. Nelson, J. Funct. Anal. **12**, 97 (1973).
[2] S. Agmon, *Lectures on Exponential Decay of Solutions of Second-order Elliptic Equations* (Princeton University Press, Princeton, 1982).
[3] F. Guerra, L. Rosen, and B. Simon, Ann. Math. **101**, 111 (1975).
[4] J. Dimock, J. Math. Phys. **20**, 2549 (1979).
[5] B. Simon, *The* $P(\phi)_2$ *Euclidean (Quantum) Field Theory* (Princeton University Press, Princeton, 1974).
[6] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975), Vol. 2.
[7] J. Neerven, *The Asymptotic Behavior of Semigroups of Linear Operators* (Birkhauser, Boston, 1996).
[8] B. Simon and R. Hoegh-Krohn, J. Funct. Anal. **9**, 121 (1972).
[9] E. Nelson, J. Funct. Anal. **11**, 211 (1972).
[10] G. Nenciu, Commun. Math. Phys. **152**, 479 (1993).

Darboux transformation and solutions for an equation in 2+1 dimensions

P. G. Estévez[a)]
Area de Física Teórica, Facultad de Física, Universidad de Salamanca, 37008 Salamanca, Spain

(Received 5 May 1998; accepted for publication 13 October 1998)

Painlevé analysis and the singular manifold method are the tools used in this paper to perform a complete study of an equation in 2+1 dimensions. This procedure has allowed us to obtain the Lax pair, Darboux transformation and τ functions in such a way that a plethora of different solutions with solitonic behavior can be constructed iteratively. © *1999 American Institute of Physics.*
[S0022-2488(99)04202-4]

I. INTRODUCTION

Among the various approaches followed to study the behavior of nonlinear partial differential equations (PDEs), Painlevé analysis has proved to be one of the most fruitful, providing an algorithmic procedure that affords a systematic way to deal with nonlinear PDEs. Despite this, it has often been used merely as a test of integrability while other methods, as Hirota's method or inverse scattering, have been used to obtain explicit solutions.

Our aim here is to show, for an equation in 2+1, that an approach based on Painlevé techniques, such as the singular manifold method (SMM), can be successful in identifying many of the properties of nonlinear PDEs (Bäcklund and Darboux transformations, τ-functions, etc.) as well as in constructing an iterative procedure to obtain multisolitonic solutions.

The subject of our study is the 2+1 PDE

$$\begin{aligned} 0&=V_y-(u\omega)_x,\\ 0&=\lambda u_t+u_{xx}-2uV, \qquad (1.1)\\ 0&=\lambda\omega_t-\omega_{xx}+2\omega V. \end{aligned}$$

The real version of this equation was obtained in Ref. 1 as a reduction of self-dual Yang–Mills equations while the complex version appears in Ref. 2. The equation has the Painlevé property (PP) as it has been shown by Radha and Lakshmanan[3] (real version) and Porsezian[4] (complex version). The bilinear method was applied in Ref. 3 to obtain some soliton and dromion solutions.

For $\lambda=i$ and $\omega=u^*$, Eq. (1.1) is the expression proposed by Fokas in Ref. 5. This case contains the nonlinear Schrödinger equation when $x=y$.[6]

Recently,[6] the author and her co-worker have shown that there is a Miura transformation between Eq. (1.1) and the generalized dispersive long wave equation.[7,8]

The plan of this paper is as follows: In Sec. II we shall apply the singular manifold method to Eq. (1.1). Sections III, IV, and V are devoted to showing how the SMM allows us to determine algorithmically the Lax pair as well as Darboux transformations and τ-functions. In Sec. VI, several solutions are constructed explicitly. We close with a list of conclusions.

[a)]Electronic mail: pilar@sonia.usal.es

II. THE SINGULAR MANIFOLD METHOD

The equation under study in this paper is the real version of Eq. (1.1), which reads

$$
\begin{aligned}
0 &= m_y + u\omega, \\
0 &= u_t + u_{xx} + 2um_x, \qquad (2.1) \\
0 &= \omega_t - \omega_{xx} - 2\omega m_x,
\end{aligned}
$$

where we have set $\lambda = 1$ and $V = -m_x$.

A. Leading term analysis

To check the Painlevé property[9] for Eq. (2.1), we require a generalized Laurent expansion of the fields in terms of an arbitrary singularity manifold (depending on the initial data) $\chi(x,y,t) = 0$. This expansion should be of the form[10]

$$u = \sum_{j=0}^{\infty} u_j \chi^{j-a}, \quad \omega = \sum_{j=0}^{\infty} \omega_j \chi^{j-b}, \quad m = \sum_{j=0}^{\infty} m_j \chi^{j-c}. \qquad (2.2)$$

By substituting Eq. (2.2) into Eq. (2.1), we have for the leading terms

$$a = b = c = 1, \quad m_0 = \chi_x, \quad u_0 \omega_0 = \chi_x \chi_y. \qquad (2.3)$$

Leading analysis is able to determine the product of the dominant terms $u_0\omega_0$ but not each one independently, which means that u and ω are not good fields in which to apply the singularity analysis because their dominant behavior is not well defined. However, for the field m, the leading term m_0 is well defined. This suggests that the ''good field,'' from the point of view of the Painlevé analysis, is m. Accordingly, our first aim will be to write Eq. (2.1) as a partial differential equation only for m. It is not difficult to check (see Appendix) that if we identify

$$m_t = n_x, \qquad (2.4)$$

we can remove u and ω from Eq. (2.1) to obtain the PDE

$$0 = m_y^2(n_{yt} - m_{xxxy}) + m_{xy}(n_y^2 - m_{xy}^2) + 2m_y(m_{xy}m_{xxy} - n_y n_{xy}) - 4m_y^3 m_{xx}. \qquad (2.5)$$

In Ref. 6, it has been shown that there is a Miura transformation between Eq. (2.5) and the generalized long dispersive wave equation.[7,8] This is why below we shall be referring, in the next, to Eq. (2.5) as MGLDW (modified generalized long dispersive wave equation). The study of this equation for the field m will be the subject of the rest of this paper. Furthermore, u and ω can easily be obtained from m as

$$u = \sqrt{m_y} e^{\int (n_y/2m_y)dx}, \quad \omega = -\sqrt{m_y} e^{-\int (n_y/2m_y)dx}, \qquad (2.6)$$

as we show in a detailed manner in the Appendix

B. Truncated expansion: Auto-Bäcklund transformations

As stated above, the singularity manifold χ is an arbitrary function depending on the initial data. The SMM requires us to restrict ourselves to the particular cases of the singularity manifold for which the expansion (2.2) truncates at the constant level.[11] In this case the singularity manifold is not longer an arbitrary function because it is ''determined'' by the condition of truncation. We call it ''singular manifold'' and we shall use ϕ to refer to it. Thus, the truncation of Eq. (2.2) is

$$m' = m + \frac{\phi_x}{\phi} \Rightarrow n' = n + \frac{\phi_t}{\phi}, \tag{2.7}$$

where both m and m' are solutions of Eq. (2.5). Accordingly, truncation of the Painlevé series adopts the form of an auto-Bäcklund transformation between two solutions of Eq. (2.5).

C. Expression of the solutions in terms of the singular manifold

Substitution of Eq. (2.7) into Eq. (2.5) provides a polynomial in ϕ. The way to proceed in the SMM is to require that all the coefficients of this polynomial should be zero. The result should be

a) The expression of m in terms of ϕ.
b) The equations to be satisfied by ϕ.

For Eq. (2.5), the polynomial in ϕ is rather complicated. We used MAPLE V to handle the calculation. This allows us to obtain the derivatives of m in terms of the singular manifold. The result is

$$4m_x = p_t - v_x - \frac{v^2 + w^2}{2}, \tag{2.8}$$

$$4n_y = 2\frac{(q_x + qv)_x p_y - (q_x + qv)p_{xy}}{q} + 4\frac{(p_y + qp_x)}{q} m_y, \tag{2.9}$$

$$4m_y = \frac{p_y^2 - (q_x + qv)^2}{q}, \tag{2.10}$$

where p,q,wyv are defined from the singular manifold as

$$v = \frac{\phi_{xx}}{\phi_x}, \quad w = \frac{\phi_t}{\phi_x} = p_x, \quad q = \frac{\phi_y}{\phi_x}. \tag{2.11}$$

D. Singular manifold equations

The equations to be satisfied by the singular manifold ϕ are not difficult to obtain:

• On one hand, there are some generic equations arising from the compatibility of the definitions (2.11). These are

$$\begin{gathered} \phi_{xxt} = \phi_{txx} \Rightarrow v_t = (w_x + vw)_x, \\ \phi_{xxy} = \phi_{yxx} \Rightarrow v_y = (q_x + vq)_x, \\ \phi_{yt} = \phi_{ty} \Rightarrow q_t = w_y + wq_x - qw_x. \end{gathered} \tag{2.12}$$

• Also, there is an equation that is specific for Eq. (2.5) that can be determined by taking the cross derivatives in Eqs. (2.8)–(2.10). This equation is

$$p_{yt} = q_{xxx} + q(v_{xx} - vv_x) + p_x p_{xy} + \left(\frac{p_y^2 - q_x^2}{q}\right)_x. \tag{2.13}$$

The set (2.12) and (2.13) forms the singular manifold equations.

III. LAX PAIR AND SMM

It is unnecessary to talk about the importance of determining the Lax pair of a nonlinear PDE. Nevertheless, in most cases it is determined by inspection. We shall see here that a nontrivial advantage of Painlevé analysis is that it allows us to determine the Lax pair in an algorithmic way.[12,13]

A. Dominant terms in singular manifold equations

Returning to the singular manifold Eqs. (2.12) and (2.13), these can be considered to be a system of coupled nonlinear PDEs. We can, therefore, analyze their leading terms. This requires us to set

$$w \sim w_0 \chi^a, \quad v \sim v_0 \chi^b, \quad q \sim q_0 \chi^c. \tag{3.1}$$

The balance of leading powers yields:

$$a = b = -1, \quad c = 0, \tag{3.2}$$

which means that only w and v have an expansion in negative powers of χ. Thus, the Painlevé expansion is only pertinent for them but not for q. Moreover, the leading analysis provides the leading coefficients $w_0 y v_0$

$$w_0 = \pm \chi_x, \quad v_0 = \chi_x. \tag{3.3}$$

The $\pm$ sign of w_0 means that there are two possible Painlevé expansions: The problem of systems with two Painlevé branches has been extensively discussed in Refs. 12–15. The suggestion of the author and co-worker is that, for this class of systems, it is necessary to consider both branches simultaneously by using two singular manifolds; one for each branch.

B. Eigenfunctions and the singular manifold

With this idea in mind, for the dominant terms of w and v we can write

$$v = \frac{\psi_x^+}{\psi^+} + \frac{\psi_x^-}{\psi^-}, \quad w = \frac{\psi_x^+}{\psi^+} - \frac{\psi_x^-}{\psi^-}, \tag{3.4}$$

where we have used ψ^+ for the singular manifold of the positive branch and ψ^- for the negative one. As we will seen later on, ψ^+ and ψ^- will be the eigenfunctions of the Lax pair and hereafter we will designate them as eigenfunctions.

• Taking the derivatives of Eq. (3.4) with respect to t and y and using Eq. (2.12) to integrate them in x, we can write

$$w_x + wv = \frac{\psi_t^+}{\psi^+} + \frac{\psi_t^-}{\psi^-}, \quad p_t = \frac{\psi_t^+}{\psi^+} - \frac{\psi_t^-}{\psi^-} \tag{3.5}$$

and

$$q_x + qv = \frac{\psi_y^+}{\psi^+} + \frac{\psi_y^-}{\psi^-},$$

$$p_y = \frac{\psi_y^+}{\psi^+} - \frac{\psi_y^-}{\psi^-}. \tag{3.6}$$

Expressions (3.4)–(3.6) allow us *to write the logarithmic derivatives of the eigenfunctions* ψ^+ *and* ψ^- *in terms of the singular manifold as*

$$\alpha^+=\frac{\psi_x^+}{\psi^+}=\frac{v+w}{2},\quad \alpha^-=\frac{\psi_x^-}{\psi^-}=\frac{v-w}{2},\tag{3.7}$$

$$\beta^+=\frac{\psi_y^+}{\psi^+}=\frac{q_x+qv+p_y}{2},\quad \beta^-=\frac{\psi_y^-}{\psi^-}=\frac{q_x+qv-p_y}{2},\tag{3.8}$$

$$\gamma^+=\frac{\psi_t^+}{\psi^+}=\frac{w_x+wv+p_t}{2},\quad \gamma^-=\frac{\psi_t^-}{\psi^-}=\frac{w_x+wv-p_t}{2},\tag{3.9}$$

$\alpha,\beta y\,\gamma$ have been introduced with the single purpose of simplifying later calculations.

• Conversely, the determination of ϕ from $\psi^+ y\,\psi^-$ is not difficult taking into account Eq. (2.11), which allows us integrate Eq. (3.4) with respect to x, which yields

$$\phi_x=\psi^+\psi^-,\tag{3.10}$$

where the integration constant has been set at zero with no loss of generality (because the singular manifold is defined except for a multiplicative constant). The t derivative of ϕ can be obtained by combining Eqs. (2.11), (3.4), and (3.10) to obtain

$$\phi_t=\psi^-\psi_x^+-\psi^+\psi_x^-,\tag{3.11}$$

and, similarly, ϕ_y arises from Eqs. (2.10), (2.11), and (3.6) as

$$\phi_y=-\frac{\psi_y^+\psi_y^-}{m_y}.\tag{3.12}$$

Equations (3.10)–(3.12) allow us to construct ϕ from ψ^+ and ψ^-. Accordingly, *the total correspondence between singular manifolds and eigenfunctions is explicitly constructed*.

C. Linearization of the singular manifold equations: The Lax pair

We return to Eqs. (2.8)–(2.10). These equations are the expression of the seminal solution m in terms of the singular manifold. At the same time, Eqs. (3.7)–(3.12) relate the singular manifold to the eigenfunctions. The question is now: How can we express m in terms of ψ^+ and ψ^-?

• As a previous step, it is easy to see that Eq. (2.10) can be combined with Eq. (3.8), yielding

$$m_y=-\frac{\beta^+\beta^-}{q}\Rightarrow\frac{\psi_y^+\psi_y^-}{\psi^+\psi^-}=-qm_y,\tag{3.13}$$

which shows the coupling between ψ^+ and ψ^-.

• Let us return to Eq. (2.8). To write this in terms of the eigenfunctions, we need to substitute v and p_t from Eqs. (3.7) and (3.9)

$$4m_x=2\gamma^+-2\alpha_x^+-2(\alpha^+)^2,\quad \text{or}\quad 4m_x=-2\gamma^--2\alpha_x^--2(\alpha^-)^2.$$

Now, by substituting α and γ

$$0=\psi_t^+-\psi_{xx}^+-2m_x\psi^+,\quad \text{or}\quad 0=\psi_t^-+\psi_{xx}^-+2m_x\psi^-,\tag{3.14}$$

and this can be considered as the temporal part of the Lax pair.

• Finally, by combining it with Eqs. (3.7) and (3.8), Eq. (2.9) can be written as

$$qn_y=[\beta^+\beta_x^--\beta^-\beta_x^+]+m_y[(\beta^+-\beta^-)+q(\alpha^+-\alpha^-)].\tag{3.15}$$

If we use $\beta^+\beta^-=-qm_y$ and $\alpha^++\alpha^-=v$ to remove from Eq. (3.15) (β^-,α^-) or (β^+,α^+).

$$n_y = -m_{xy} + 2m_y\left(\frac{\beta_x^+}{\beta^+} + \alpha^+ + \frac{m_y}{\beta^+}\right), \quad n_y = m_{xy} - 2m_y\left(\frac{\beta_x^-}{\beta^-} + \alpha^- + \frac{m_y}{\beta^+}\right),$$

or

$$(n_y + m_{xy})\psi_y^+ = 2m_y(\psi_{xy}^+ + m_y\psi^+), \quad (-n_y + m_{xy})\psi_y^- = 2m_y(\psi_{xy}^- + m_y\psi^-), \tag{3.16}$$

and this can be considered the spatial part of the Lax pair.

Thus, the SMM allows us to define two eigenfunctions, ψ^+ and ψ^-, such that *the expression of the truncated solutions in terms of these eigenfunctions is precisely the Lax pair* Eqs. (3.15) and (3.16).

IV. DARBOUX TRANSFORMATIONS

This section will be devoted to determining an algorithmic procedure for constructing solutions.

• We shall summarize the results obtained in the previous section: Let m be a solution of Eq. (2.5), and ϕ_1 a singular manifold for it. This singular manifold can be constructed by means of two eigenfunctions ψ_1^+ and ψ_1^- through

$$\phi_{1x} = \psi_1^+\psi_1^-, \quad m_y\phi_{1y} = -\psi_{1y}^+\psi_{1y}^-, \quad \phi_{1t} = \psi_1^-\psi_{1x}^+ - \psi_1^+\psi_{1x}^-, \tag{4.1}$$

where ψ_1^+ and ψ_1^- satisfy the Lax pairs

$$0 = \psi_{1t}^+ - \psi_{1xx}^+ - 2m_x\psi_1^+, \quad 0 = \psi_{1t}^- + \psi_{1xx}^- + 2m_x\psi_1^-,$$
$$0 = 2m_y\psi_{1xy}^+ - (m_{xy} + n_y)\psi_{1y}^+ + 2m_y^2\psi_1^+, \quad 0 = 2m_y\psi_{1xy}^- - (m_{xy} - n_y)\psi_{1y}^- + 2m_y^2\psi_1^-. \tag{4.2}$$

• According to Eq. (2.7), the singular manifold ϕ_1 allows us to define a new solution m'

$$m' = m + \frac{\phi_{1x}}{\phi_1} \Rightarrow n' = n + \frac{\phi_{1t}}{\phi_1}, \tag{4.3}$$

whose Lax pairs will be

$$0 = \psi_t'^+ - \psi_{xx}'^+ - 2m_x'\psi'^+, \quad 0 = \psi_t'^- + \psi_{xx}'^- + 2m_x'\psi'^-,$$
$$0 = 2m_y'\psi_{xy}'^+ - (m_{xy}' + n_y')\psi_y'^+ + 2m_y'^2\psi'^+, \quad 0 = 2m_y'\psi_{xy}'^- - (m_{xy}' - n_y')\psi_y'^- + 2m_y'^2\psi'^-, \tag{4.4}$$

and ψ'^+ and ψ'^- can be used to construct, for m', a singular manifold ϕ' defined as

$$\phi_x' = \psi'^+\psi'^-, \quad m_y'\phi_y' = -\psi_y'^+\psi_y'^-, \quad \phi_t' = \psi'^-\psi_x'^+ - \psi'^+\psi_x'^-. \tag{4.5}$$

A. Truncated expansion in the Lax pair

A Lax pair, such as Eq. (4.4), is usually considered to be a linear system for ψ', where m' is the potential and hence the inverse scattering method can be applied.

A different interpretation[12,16] of Eq. (4.4) is to consider it as a coupled ''nonlinear'' system of PDEs for the fields m', n', ψ'^+, ψ'^-. In this case, the singular manifold method can be applied to the Lax pair itself and the truncated expansion (4.3) for m and n should be combined with a similar expansion for ψ'^+ and ψ'^-. In fact, this expansion could be written as

$$\psi'^+ = \psi_2^+ + \frac{\psi_0^+}{\phi_1}, \quad \psi'^+ = \psi_2^- + \frac{\psi_0^-}{\phi_1},$$

where ψ_0^+, ψ_0^- are the dominant terms. It is useful. For later calculations, it is useful to set $\psi_0^+ = -\psi_1^+\Omega^+$, and $\psi_0^- = -\psi_1^-\Omega^-$. Therefore

$$\psi'^+ = \psi_2^+ - \psi_1^+ \frac{\Omega^+}{\phi_1}, \quad \psi'^- = \psi_2^- - \psi_1^- \frac{\Omega^-}{\phi_1}. \tag{4.6}$$

Substitution of the truncated expansions (4.3) and (4.6) in the Lax pairs (4.4) provides the following results (we used MAPLE for the calculations):

• (1) ψ_2^+ and ψ_2^- are eigenfunctions for m. Consequently, they satisfy Lax pairs such as

$$0 = \psi_{2t}^+ - \psi_{2xx}^+ - 2m_x\psi_2^+, \quad 0 = \psi_{2t}^- + \psi_{2xx}^- + 2m_x\psi_2^-,$$

$$0 = 2m_y\psi_{2xy}^+ - (m_{xy}+n_y)\psi_{2y}^+ + 2m_y^2\psi_2^+, \quad 0 = 2m_y\psi_{2xy}^- - (m_{xy}-n_y)\psi_{2y}^- + 2m_y^2\psi_2^-. \tag{4.7}$$

• (2) Ω^+ and Ω^- are related to the eigenfunctions in the following way:

$$\Omega_x^- = \psi_1^+\psi_2^-, \quad m_y\Omega_y^- = -\psi_{1y}^+\psi_{2y}^-, \quad \Omega_t^- = \psi_{1x}^+\psi_2^- - \psi_1^+\psi_{2x}^-, \tag{4.8}$$

$$\Omega_x^+ = \psi_2^+\psi_1^-, \quad m_y\Omega_y^+ = -\psi_{2y}^+\psi_{1y}^-, \quad \Omega_t^+ = \psi_{2x}^+\psi_1^- - \psi_2^+\psi_{1x}^-. \tag{4.9}$$

• To summarize: Two pairs of eigenfunctions $(\psi_1^+, \psi_1^-), (\psi_2^+, \psi_2^-)$ for a solution (m,n) are sufficient to construct the following transformation:,

$$\begin{aligned} m' &= m + \frac{\phi_{1x}}{\phi_1}, \quad n' = n + \frac{\phi_{1t}}{\phi_1}, \\ \psi'^+ &= \psi_2^+ - \psi_1^+ \frac{\Omega^+}{\phi_1}, \quad \psi'^- = \psi_2^- - \psi_1^- \frac{\Omega^-}{\phi_1}, \end{aligned} \tag{4.10}$$

where ϕ_1, Ω^+, and Ω^- are related to the eigenfunction through Eqs. (4.1), (4.8), and (4.9).

Equation (4.10) is a transformation of potentials and eigenfunctions that leaves invariant the Lax pairs. It should, therefore, be considered a *Darboux transformation*.[17]

V. ITERATION OF THE SINGULAR MANIFOLD: τ-FUNCTIONS

A well known method for obtaining multisolitonic solutions of PDEs is the bilinear Hirota method. Indeed, some solutions of Eq. (2.1) have been identified with this method.[3] Let us address ourselves to the task of establishing, by explicit construction, the relationship between the singular manifold and the τ-functions of Hirota's method.

• Equation (4.5) could be considered as a nonlinear system among ϕ', ψ'^+ and ψ'^-. For this system we can use the same criterion used in the previous section. It requires that the expansion

$$\psi'^+ = \psi_2^+ - \psi_1^+ \frac{\Omega^+}{\phi_1}, \quad \psi'^- = \psi_2^- - \psi_1^- \frac{\Omega^-}{\phi_1}, \tag{5.1}$$

for ψ'^+ and ψ'^- should be combined with a truncated expansion for ϕ'.

$$\phi' = \phi_2 + \frac{\Delta}{\phi_1}. \tag{5.2}$$

It is not difficult to prove that the substitution of Eqs. (5.1) and (5.2) into Eq. (4.5) gives

$$\Delta = -\Omega^+\Omega^-, \tag{5.3}$$

while ϕ_2 is the singular manifold for m related to ψ_2^+ and ψ_2^-, which means

$$\phi_{2x}=\psi_2^+\psi_2^-,$$
$$m_y\phi_{2y}=-\psi_{2y}^+\psi_{2y}^-, \tag{5.4}$$
$$\phi_{2t}=\psi_2^-\psi_{2x}^+-\psi_2^+\psi_{2x}^-.$$

• As far as Eq. (5.2) defines a singular manifold for m', it can be used to obtain a new solution

$$m''=m'+\frac{\phi_x'}{\phi'},\quad n''=n'+\frac{\phi_t'}{\phi'}, \tag{5.5}$$

which, combined with Eq. (4.3), is

$$m''=m+\frac{\tau_x}{\tau},\quad n''=n+\frac{\tau_t}{\tau}, \tag{5.6}$$

where

$$\tau=\phi'\phi_1=\phi_1\phi_2-\Omega^+\Omega^-. \tag{5.7}$$

In the previous section we have shown that $\phi_1,\phi_2,\Omega^+,\Omega^-$ are obtained from the eigenfunctions $(\psi_1^+,\psi_1^-),(\psi_2^+,\psi_2^-)$. Therefore, *Eq. (5.7) affords the relationship between* τ*-functions, on one hand, and singular manifolds*, on the other hand.

VI. SOLUTIONS

From the previous results we can derive an iterative procedure to construct solutions. It can be summarized as follows:

(1) Starting with a seminal solution *m*, solve the Lax pairs (4.2) and (4.8) to obtain $\psi_1^+,\psi_1^-,\psi_2^+,\psi_2^-$.
(2) Perform the integration of Eqs. (4.1), (4.8), (4.9), and (5.4) to get ϕ_1, Ω^+, Ω^- and ϕ_2: Use Eq. (5.7) to construct τ.
(3) Use Eq. (4.3) to obtain the solution m' for the first iteration and Eq. (5.5) for the second one m''.

The easiest way to obtain explicit solutions is to apply the above explained procedure, starting with a trivial seminal solution. We shall use as seminal solutions $m=m_0y$ and $m=0$. From the dependence on y of Eqs. (4.1), (4.9), (4.10), and (5.4) it is that the behavior is totally different, depending on whether m_y is zero or not and giving rise to line-soliton or dromion behavior, respectively.

A. Line solitons $m=\omega_0 y$

• The easiest solutions of Eqs. (4.2) and (4.7) are

$$\psi_1^+=\exp\left[a_1^+x-\frac{\omega_0}{a_1^+}y+a_1^{+2}t\right],\quad \psi_2^+=\exp\left[a_2^+x-\frac{\omega_0}{a_2^+}y+a_2^{+2}t\right],$$
$$\psi_1^-=\exp\left[a_1^-x-\frac{\omega_0}{a_1^-}y-a_1^{-2}t\right],\quad \psi_2^-=\exp\left[a_2^-x-\frac{\omega_0}{a_2^-}y-a_2^{-2}t\right], \tag{6.1}$$

where $a_1^+,a_1^-,a_2^+ y a_2^-$ are arbitrary constants.
• Integration of Eqs. (4.1), (4.8), (4.9), and (5.4) affords

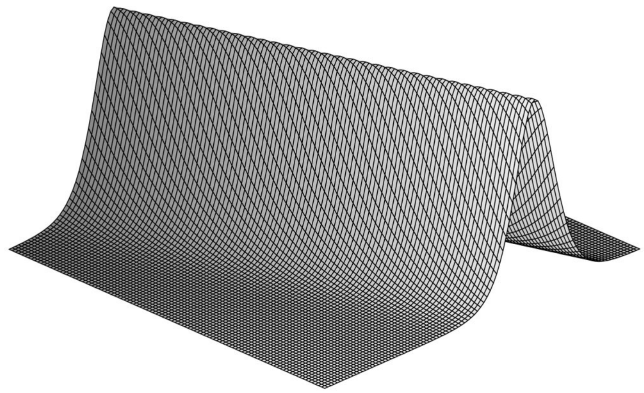

FIG. 1. Line soliton.

$$\phi_1=\frac{1}{a_1^+ + a_1^-}(b_1+\psi_1^+\psi_1^-),\quad \Omega^+=\frac{1}{a_2^+ + a_1^-}(c^+ + \psi_1^+\psi_2^-),$$
$$\phi_2=\frac{1}{a_2^+ + a_2^-}(b_2+\psi_2^+\psi_2^-),\quad \Omega^-=\frac{1}{a_1^+ + a_2^-}(c^- + \psi_2^+\psi_1^-), \tag{6.2}$$

where $b_1, b_2, c^+ y c^-$ are arbitrary constants.

• The first iteration provides the solution (Fig. 1)

$$m_y'=\psi_0+\partial_{xy}[\ln \phi_1], \tag{6.3}$$

and the second (Fig. 2)

$$m_y''=\omega_0+\partial_{xy}[\ln \tau], \tag{6.4}$$

where

$$\phi_1=\frac{b_1}{a_1^+ + a_1^-}(1+F_1). \tag{6.5}$$

$$\tau=\phi_1\phi_2-\Omega^+\Omega^-=\frac{b_1 b_2}{(a_1^+ + a_1^-)(a_2^+ + a_2^-)}[1+F_1+F_2+AF_1F_2], \tag{6.6}$$

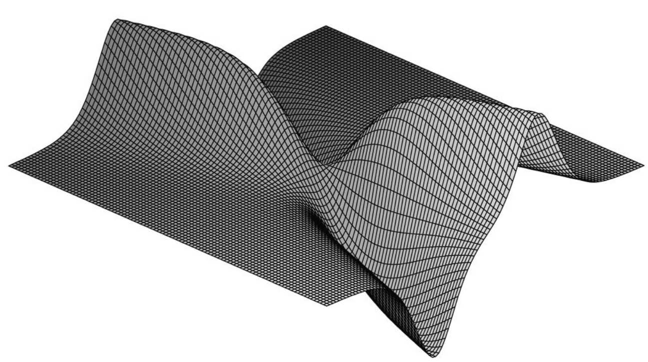

FIG. 2. Interaction of two line solitons.

and

$$F_i(x,y,t)=\exp\left[(a_i^+ + a_i^-)\left\{x-\frac{\omega_0}{a_i^+ a_i^-}y+(a_i^+ - a_i^-)t\right\}+\varphi_i\right], \tag{6.7}$$

$$A=\frac{(a_2^+ - a_1^+)(a_2^- - a_1^-)}{(a_2^+ + a_1^-)(a_2^- + a_1^+)}, \tag{6.8}$$

and b_i has been redefined as: $b_i = e^{-\varphi_i}$.

Particular case: When $a_2^+ = a_1^+$, or $a_2^- = a_1^-$, $A=0$ and this is said to be *resonant state*.[18]

B. Dromions $m=0$

In this case Eqs. (4.1), (4.8), (4.9), and (5.4) require that

$$\psi_{1y}^+\psi_{1y}^- = \psi_{2y}^+\psi_{2y}^- = \psi_{1y}^+\psi_{2y}^- = \psi_{2y}^+\psi_{1y}^- = 0.$$

Therefore, it is compulsory that $\psi_{1y}^- = \psi_{2y}^- = 0$, or $\psi_{1y}^+ = \psi_{2y}^+ = 0$.

- If we choose the possibility $\psi_{1y}^- = \psi_{2y}^- = 0$, then simple solutions of Eqs. (4.2) and (4.7) are

$$\begin{aligned}\psi_1^- &= e^{a_1^- x - a_1^{-2} t}, \quad \psi_1^+ = (e^{a_1^+ x + a_1^{+2} t})E_1(y) = Q_1^+(x,t)E_1(y),\\ \psi_2^- &= e^{a_2^- x - a_2^{-2} t}, \quad \psi_2^+ = (e^{a_2^+ x + a_2^{+2} t})E_2(y) = Q_2^+(x,t)E_2(y),\end{aligned} \tag{6.9}$$

where $a_1^+, a_1^-, a_2^+, a_2^-$ are arbitrary constants while E_i are arbitrary functions of y.

- We can now perform now the integration of Eqs. (4.1), (4.8), (4.9), and (5.4) to obtain

$$\begin{aligned}\phi_1 &= \frac{1}{a_1^+ + a_1^-}(E_1 Q_1^+ \psi_1^- + M_1(y)), \quad \Omega^+ = \frac{1}{a_2^+ + a_1^-}(E_2 \psi_1^- Q_2^+ + N^+(y)),\\ \phi_2 &= \frac{1}{a_2^+ + a_2^-}(E_2 \psi_2^- Q_2^+ + M_2(y)), \quad \Omega^- = \frac{1}{a_1^+ + a_2^-}(E_1 \psi_2^- Q_1^+ + N^-(y)).\end{aligned} \tag{6.10}$$

N^+, N^-, and M_i^+ are arbitrary functions of y. The arbitrariness of the six functions E_i, M_i, N^j, and the four constants a_i^+, a_i^- implies that there are many particular cases. We list some of them.

1. 1+1 *dromions*

These can be obtained by choosing

$$E_i(y) = 1 + b_i e^{c_i y}, \quad M_i(y) = 1 + e^{c_i y}, \quad N^* = N^- = 0, \tag{6.11}$$

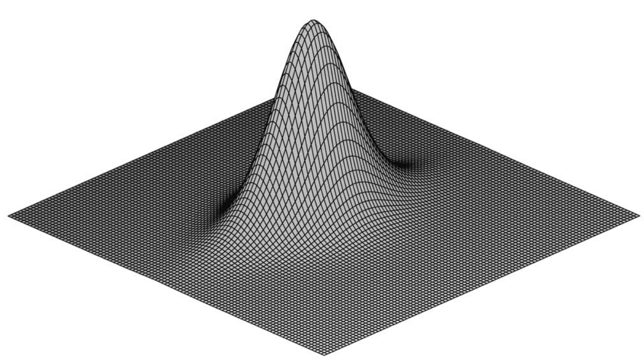

FIG. 3. One dromion.

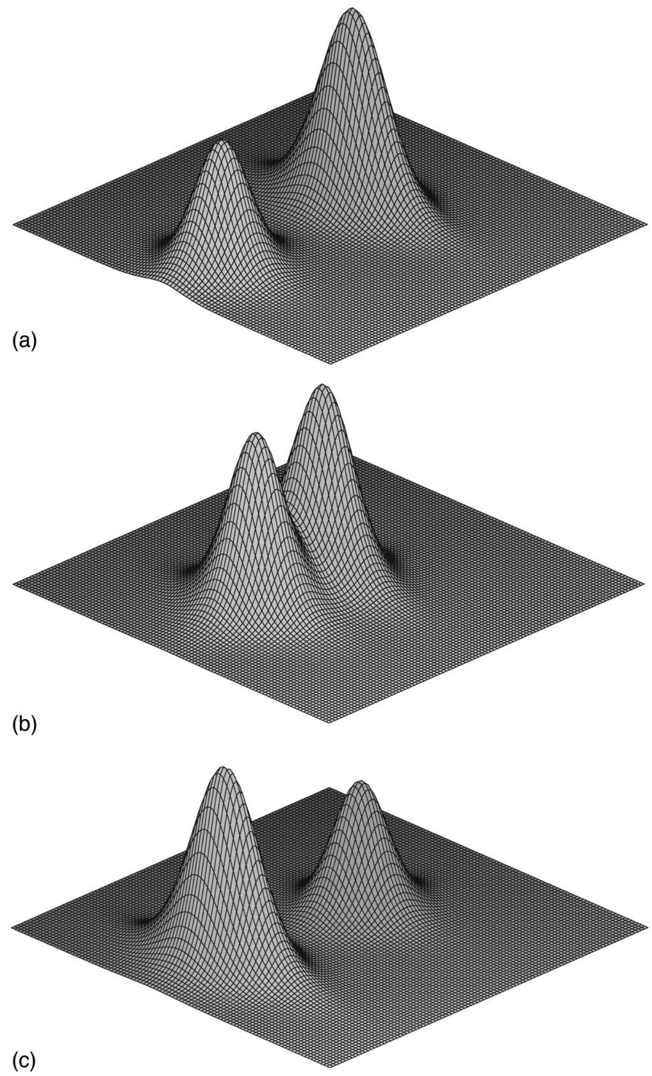

FIG. 4. (a) Interaction of two dromions, $t<0$. (b) Interaction of two dromions, $t=0$. (c) Interaction of two dromions, $t>0$.

where b_i and c_i are arbitrary constants.

• The first and second iteration yield

$$m_y' = \partial_{xy}[\ln \phi_1], \tag{6.12}$$

$$m_y'' = \partial_{xy}[\ln \tau], \tag{6.13}$$

where

$$\phi_1 = \frac{1}{a_1^+ + a_1^-}(M_1 + E_1 F_1), \tag{6.14}$$

$$\tau = \phi_1 \phi_2 - \Omega^+ \Omega^- = \frac{1}{(a_1^+ + a_1^-)(a_2^+ + a_2^-)}[M_1 M_2 + M_1 E_2 F_2 + M_2 E_1 F_1 + A E_1 E_2 F_1 F_2], \tag{6.15}$$

and

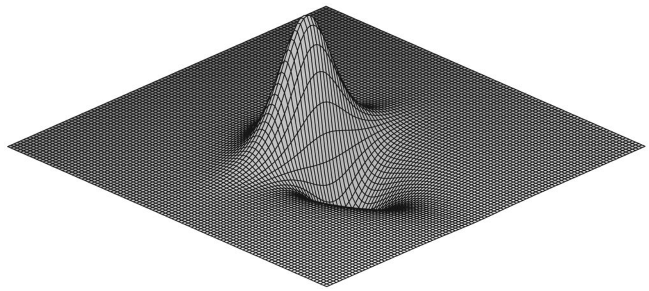

FIG. 5. $2+1$ dromion.

$$F_i(x,t)=\exp[(a_i^+ + a_i^-)\{x+(a_i^+ - a_i^-)t\}], \tag{6.16}$$

$$A=\frac{(a_2^+ - a_1^+)(a_2^- - a_1^-)}{(a_2^+ + a_1^-)(a_2^- + a_1^+)}. \tag{6.17}$$

The behavior of Eqs. (6.12) and (6.13) are represented in Figs. 3 and 4, respectively.

2. $1+n$ *dromion*

Dromions with several jumps in the y direction can be obtained by choosing

$$E_i=1+\sum_{j=1}^{n} b_{ij}e^{c_{ij}y}, \quad M_i=1+\sum_{j=1}^{n} e^{c_{ij}y}.$$

The first iteration

$$m_y'=\partial_{xy}(\ln \phi_1),$$

describes a structure with n jumps located along the y-direction, moving in the x-direction with velocity $a_1^+ - a_1^-$.

Figure 5 represents one of these structures with $n=2$ and $c_{11}>0$, $c_{12}<0$.

Figure 6 corresponds to $n=3$ and $c_{11}>0$, $c_{12}>0$, $c_{13}>0$.

The solution that we have obtained in this section generalizes the solutions found in Ref. 3 by means of the bilinear method.

VII. CONCLUSIONS

• A system of nonlinear PDEs proposed by different authors as one of the simplest equations in $2+1$ dimensions is studied from the point of view of the Painlevé property. The dominant

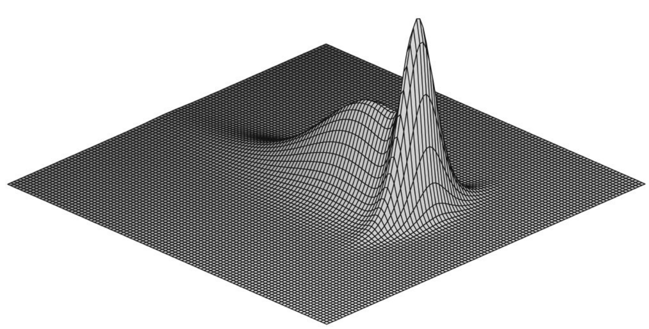

FIG. 6. $3+1$ dromion.

behavior indicates the best field to use Painlevé analysis. On this basis, we rewrite the system as a PDE [Eq. (2.5)] with only one field. This equation[6] can be considered as the modified version of the generalized long dispersive wave equation.[7] This is why we have call it MGLDW (modified generalized long dispersive wave equation).

• The singular manifold method was applied to MGLDW in Sec. II. The singular manifold equations, as well as the expression of the seminal field in terms of the singular manifold were obtained.

• In Sec. III, we linearized the singular manifold equations to obtain the Lax pair. The relation between the singular manifold and the eigenfunctions of the Lax pair is constructed explicitly.

• In Sec. IV the Lax pair was considered as a system of nonlinear coupled PDE. We applied the singular manifold method to the Lax pair itself. The bonus is the construction of Darboux transformations for MGLDW. Its transformations allow us to determine an iterative method for obtaining solutions. The relation between this method and the Hirota τ-functions is shown in Sec. V.

• Finally Sec. VI is devoted to the construction of solitonic solutions of MGLDW. A rich collection of solutions with different solitonic behavior appear depending on the seminal solution that we have chosen.

• We believe that the equation analyzed in depth in this paper is a good example of how to obtain maximum information about the equation using Painlevé analysis and the singular manifold method as the only tools.

ACKNOWLEDGMENTS

This research has been supported in part by DGICYT under contract PB95-0947. I would like to thank Professor J. M. Cerveró for encouragement and a careful reading of the manuscript.

APPENDIX: COMPLEMENTARY CALCULATION

We first attempt to write

$$0=u_t+u_{xx}+2um_x, \tag{A1}$$

$$0=\omega_t-\omega_{xx}-2\omega m_x, \tag{A2}$$

$$0=u\omega+m_y, \tag{A3}$$

as an equation for only one field

Taking u out of Eq. (A3) and substituting it into Eq. (A1), we obtain

$$0=\frac{m_{yt}+m_{xxy}}{\omega}-2m_{xy}\frac{\omega_x}{\omega^2}+m_y\left(\frac{2m_x}{\omega}-\frac{\omega_t}{\omega^2}-\frac{\omega_{xx}}{\omega^2}-2\frac{\omega_x^2}{\omega^3}\right). \tag{A4}$$

Using Eq. (A2) into Eq. (A4), we also obtain

$$0=m_{xxy}+m_{yt}-\left(2m_y\frac{\omega_x}{\omega}\right)_x, \tag{A5}$$

which can easily be integrated by setting $m_t=n_x$, which yields

$$\frac{\omega_x}{\omega}=\frac{m_{xy}+n_y}{2m_y}. \tag{A6}$$

Substituting Eq. (A6) into Eq. (A2), we obtain

$$\frac{\omega_t}{\omega}=2m_x+\frac{m_{xxy}+n_{xy}}{2m_y}-\frac{m_{xy}^2-n_y^2}{4m_y^2}. \tag{A7}$$

Next, we calculate the identity $(\omega_t/\omega)_x=(\omega_x/\omega)_t$ using Eqs. (A6) and (A7), and finally we obtain for m the equation

$$0=m_t-n_x \tag{A8}$$

$$0=m_y^2(n_{yt}-m_{xxxy})+m_{xy}(n_y^2-m_{xy}^2)+2m_y(m_{xy}m_{xxy}-n_yn_{xy})-4m_y^3m_{xx}. \tag{A9}$$

The integration of Eq. (A6) is

$$u=\sqrt{m_y}e^{\int(n_y/2m_y)dx}. \tag{A10}$$

And from Eq. (3.4) we finally obtain

$$\omega=-\sqrt{m_y}e^{\int-(n_y/2m_y)dx}. \tag{A11}$$

[1] S. Chakravarty, S. L. Kent, and E. T. Newman, J. Math. Phys. **36**, 763–772 (1995).
[2] A. Maccari, J. Math. Phys. **37**, 6207–6212 (1996).
[3] R. Radha and M. Lakshmanan, J. Math. Phys. **38**, 292–299 (1997).
[4] K. Porsezian, J. Math. Phys. **38**, 4675–4679 (1997).
[5] A. Fokas, Inverse Probl. **10**, L19–L22 (1994).
[6] J. M. Cerveró and P. G. Estévez, J. Math. Phys. **39**, 2800–2807 (1998).
[7] M. Boiti, J. Leon, and F. Pempinelli, Inverse Probl. **3**, 37–45 (1987).
[8] M. J. Ablowitz and P. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, London Mathematical Society, Lecture Note Series (Cambridge University Press, Cambridge 1991), Vol. 149.
[9] P. Painlevé, Acta Math. **25**, 1–86 (1901).
[10] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522–526 (1983).
[11] J. Weiss, J. Math. Phys. **24**, 1405–14013 (1983).
[12] P. G. Estévez and P. R. Gordoa, Inverse Probl. **13**, 939–957 (1997).
[13] P. G. Estévez, E. Conde-Calvo, and P. R. Gordoa, J. Nonlinear. Math. Phys. **5**, 82–114 (1998).
[14] R. Conte, M. Musette, and A. Pickering, J. Phys. A **27**, 2831–2836 (1994).
[15] P. G. Estévez, P. R. Gordoa, L. Martinez-Alonso, and E. Medina-Reus, J. Phys. A **26**, 1915–1925 (1993).
[16] B. G. Konopelchenko and W. Stramp, J. Math. Phys. **24**, 40–49 (1991).
[17] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons*, Springer Series in Nonlinear Dynamics (Springer-Verlag, Berlin, 1991).
[18] P. G. Estévez and S. B. Leble, Inverse Probl. **11**, 925–937 (1995).

Symmetry reductions of the BKP hierarchy

Ignace Loris[a)]
Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, B-1050 Brussel, Belgium

Ralph Willox
Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, B-1050 Brussel, Belgium, and Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8914, Japan

(Received 21 October 1998; accepted for publication 9 December 1998)

A general symmetry of the bilinear BKP hierarchy is studied in terms of tau functions. We use this symmetry to define reductions of the BKP hierarchy, among which new integrable systems can be found. The reductions are connected to constraints on the Lax operator as well as on the bilinear formulation. A class of solutions for the reduced equations is derived. © *1999 American Institute of Physics.* [S0022-2488(99)02603-1]

I. INTRODUCTION

Many integrable nonlinear partial differential equations may be obtained as dimensional reductions of hierarchies of the KP (Kadomtsev–Petviashvili) type.[1–4] In this paper, we turn our attention to symmetry reductions of the $2+1$-dimensional BKP hierarchy. Although reductions of this hierarchy were already introduced,[1,2] it is important to realize that these authors did not achieve a complete picture of those systems obtainable as reductions of the BKP hierarchy. We wish to argue that their inability to do so results from failing to identify a ''general'' symmetry of the BKP hierarchy.

Reductions may be defined in many different ways: by imposing a constraint on the Lax operator of the system, by imposing a symmetry constraint on the soliton field,... .[1,5,6,4] Here we shall reduce the BKP hierarchy by imposing a symmetry constraint on the bilinear equations, i.e., we shall implement a symmetry reduction on the level of the BKP tau function.

Typically, the KP hierarchy of $2+1$-dimensional nlpde's is defined with the help of a general (pseudodifferential) Lax operator, with which Lax equations and Zakharov–Shabat equations are written. The BKP hierarchy is obtained from this construction by imposing a certain condition on this Lax operator, while at the same time fixing a subset of the time variables.[7,8] The resulting hierarchy of integrable $2+1$-dimensional equations allows for its own tau function and for its own accompanying bilinear formulation: the BKP bilinear identity. This identity is essential in the study of this hierarchy, as it provides its most concise formulation.

An important part of this paper is devoted to the study of certain ''eigenfunction symmetries'' of the BKP bilinear equations. Eigenfunctions are defined as the fields that solve the BKP linear (Zakharov–Shabat) problem (without necessarily having to solve the Lax eigenvalue problem). Eigenfunctions are useful in the study of the BKP hierarchy, as they can be used to construct a potential closely related to the symmetries of this hierarchy. The appropriate potential for the BKP hierarchy is, however, quite different from the one used in the KP hierarchy.[9,10,4] It can, nevertheless, be defined using the KP eigenfunction potential. Making use of the BKP bilinear identity, it will be shown that any BKP eigenfunction potential generates a symmetry for the BKP hierarchy and that this potential itself can be expressed as the ratio of two BKP tau functions.

The actual symmetry reductions are defined by imposing a relation between an elementary symmetry τ_{t_k} and this eigenfunction potential symmetry. Examples of the resulting systems in-

[a)]igloris@vub.ac.be

clude equations, which, as far as the authors were able to verify, did not yet appear in the literature. In particular, as opposed to the discussion of Refs. 1, 2, in the present analysis the KdV equation is a mere special case of the 1-constrained BKP hierarchy.

Next, we investigate how such a symmetry reduction can be connected to a constraint on the Lax operator. An additional bilinear identity for the constrained BKP hierarchy is derived and an alternative identity for the *k*-constrained BKP tau function is also found.

The last section of this paper is devoted to the solutions of the reduced BKP hierarchies. First we briefly show how one may obtain the ''Pfaffian''-type tau functions[11,12] for the (unreduced) BKP hierarchy. We investigate the form of the accompanying eigenfunctions and eigenfunction potentials. Next we decide which supplementary conditions need to be imposed on the (arbitrary) functions appearing in these Pfaffian expressions in order for them to solve the constrained BKP equations. In this way solutions to these hierarchies are derived. Some examples of solitons and rational-type solutions are given.

II. BKP HIERARCHY AND EIGENFUNCTION SYMMETRIES

The Kadomtsev–Petviashvili (KP) hierarchy is defined[7,13] in terms of the pseudodifferential gauge operator $P\equiv 1+w_1\partial^{-1}+w_2\partial^{-2}+\cdots$ and the Lax operator $L\equiv P\,\partial P^{-1}\equiv\partial+u_2\partial^{-1}+u_3\partial^{-2}+\cdots$. On the gauge operator one imposes Sato's equation $P_{t_n}=-(L^n)_-P$, implying

$$L_{t_n}=[B_n,L]\quad\text{and}\quad B_{n,t_m}-B_{m,t_n}=[B_m,B_n],$$

for differential operators $B_n\equiv(L^n)_+(=\Sigma_{j=0}^n b_{n,j}\partial_x^j)$. These nonlinear partial differential equations [for the fields $u_2(t_1,t_2,\dots),u_3(t_1,t_2,\dots),\dots$] are the compatibility conditions of the linear equations

$$L\psi=\lambda\psi,\quad \psi_{t_n}=B_n\psi,$$

satisfied by the wave function $\psi(t_1,t_2,\dots;\lambda)\equiv P(\partial)\exp(\lambda t_1+\lambda^2t_2+\cdots)$.

The *BKP hierarchy*[7] is obtained from this construction by imposing the condition

$$P^*=\partial P^{-1}\partial^{-1},\quad\text{and hence}\quad L^*\partial+\partial L=0, \tag{1}$$

at the expense of having to supress the evolutions with respect to $t_2,t_4,\dots$ (i.e., fix $t_2=t_4=\cdots=0$); the condition implies that $u_3=-u_{2,x}$, $u_5=-2u_{4,x}-u_{2,3x},\dots$, and hence $b_{n,0}=0$ and $B_n1=0$ (for n odd).

The defining relation (1) implies $P\,\partial^{-1}P^*=\partial^{-1}$ and hence that one has $\mathrm{Res}_\partial[P\,\partial^{-1}P^*\,\partial^m]=\delta_{0m}$ for $m\in\{0,1,2,...\}$. Since, in general, one has (e.g., Ref. 14, p. 82)

$$\mathrm{Res}_\lambda[\lambda^{k-1}\psi(\mathbf{t},\lambda)(-\partial)^{m+1}\psi(\mathbf{t},-\lambda)]=\mathrm{Res}_\partial[P\,\partial^{k-1}P^*\,\partial^{m+1}],\quad\forall k,m; \tag{2}$$

this relation leads to the BKP bilinear identity $(\mathbf{t}=(t_1=x,t_3,t_5,\dots))$:[7]

$$\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)]=1,\quad\forall\mathbf{t},\mathbf{t}', \tag{3}$$

for BKP wave functions. This bilinear identity (3) is equivalent to the entire BKP hierarchy of nonlinear partial differential equations, and as such is an important relation for our further discussion.

One can show that there exists a ''tau function'' $\tau(\mathbf{t})$, such that[7]

$$\psi(\mathbf{t},\lambda)=\frac{\tau(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))}{\tau(\mathbf{t})}\exp\xi(\mathbf{t},\lambda), \tag{4}$$

with $\xi(\mathbf{t},\lambda)=\Sigma_{n=0}^\infty t_{2n+1}\lambda^{2n+1}$ and $\boldsymbol{\epsilon}(\lambda)=2(\lambda^{-1},\lambda^{-3}/3,\lambda^{-5}/5,...)$. The bilinear identity (3) can now be written as

$$\mathrm{Res}_\lambda[\lambda^{-1}\tau(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\tau(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))\exp\xi(\mathbf{t}-\mathbf{t}',\lambda)]=\tau(\mathbf{t})\tau(\mathbf{t}'). \tag{5}$$

This equation contains the Hirota bilinear equations for all the nonlinear partial differential equations in the BKP hierarchy. The lowest-order equation is

$$(9D_1D_5-5D_1^3D_3-5D_3^2+D_1^6)\tau\cdot\tau=0, \tag{6}$$

where D_i the Hirota D operator with respect to t_i.[15]

It is important to remark that the BKP tau function differs from the KP tau function:[7] $\tau_{\mathrm{BKP}}=\tau_{\mathrm{KP}}(t_2=t_4=\cdots=0)^{1/2}$.

In the rest of this section we shall concentrate on the notion of eigenfunctions and their connection to the symmetries of the BKP hierarchy. Eigenfunctions are solutions of the linear equations: $\Phi_{t_n}=B_n\Phi$ (for n:1,3,5,...). Adjoint eigenfunctions satisfy the adjoint linear equations $\Phi^*_{t_n}=-B_n^*\Phi^*$ (with n:1,3,5,...). For example,

$$\Phi_{t_3}=B_3\Phi=\Phi_{3x}+3u_2\Phi_x, \quad \Phi^*_{t_3}=\Phi^*_{3x}+3(u_2\Phi^*)_x, \tag{7}$$

with $u_2=2\partial_x^2\log\tau$. Remark that the wave functions $\psi(\mathbf{t},\lambda)$ are special instances of eigenfunctions. Furthermore, a constant is always a BKP eigenfunction.

In the BKP hierarchy, eigenfunctions and adjoint eigenfunctions are related: As $L^*=-\partial L\partial^{-1}$, we find that $B_n^*\equiv(L^{*n})_+=(-1)^n(\partial L^n\partial^{-1})_+$. Hence $B_n^*\partial=-B_n\partial-B_{n,x}$ (n: odd). This relation implies that any eigenfunction Φ gives rise to an adjoint eigenfunction $\Phi^*=\Phi_x$.

In Refs. 16, 9 a KP eigenfunction potential $S(\Phi,\Phi^*)=\int^x\Phi\Phi^*$ was introduced for any pair of eigenfunction and adjoint eigenfunction (see also Refs. 4, 10). The derivatives of this potential can be expressed as a (bilinear) differential operator working on Φ and Φ^*: $\partial_{t_n}S(\Phi,\Phi^*)=\Sigma_{j=1}^n\Sigma_{i=1}^j(-1)^{i+1}(\Phi^*b_{n,j})_{(i-1)x}\Phi_{(j-i)x}$. In the case of the KP hierarchy, this potential generates a symmetry of this hierarchy.[4] Here, we shall use it to derive some results concerning BKP eigenfunctions. The first property is the following.

Property: If Φ is a BKP eigenfunction and $\Phi^*=\Phi_x$, then one has the following relation:

$$\mathrm{Res}_\lambda[\lambda^{-1}S(\psi(\mathbf{t},\lambda),\Phi^*(\mathbf{t}))\psi(\mathbf{t}',-\lambda)]=\Phi(\mathbf{t})-\Phi(\mathbf{t}'). \tag{8}$$

Proof: Denote the residue in (8) by $I(\mathbf{t},\mathbf{t}')$. From the definition of the potential S and by using the BKP bilinear identity (3), one finds that $\partial_{t_n}\partial_{t'_m}I(\mathbf{t},\mathbf{t}')=0$. Hence $I(\mathbf{t},\mathbf{t}')=f(\mathbf{t})+g(\mathbf{t}')$. Since $S(\psi(\mathbf{t},\lambda),\Phi^*(\mathbf{t}))=\mathcal{O}(\lambda^{-1})\exp\xi(\mathbf{t},\lambda)$, we have that $I(\mathbf{t},\mathbf{t}'=\mathbf{t})=0$; hence, $I(\mathbf{t},\mathbf{t}')=f(\mathbf{t})-f(\mathbf{t}')$. As $f_x(\mathbf{t})=\partial_xI(\mathbf{t},\mathbf{t}')=\Phi^*(\mathbf{t})\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)]=\Phi^*(\mathbf{t})=\Phi_x(\mathbf{t})$ [in fact, $f_{t_n}(\mathbf{t})=\partial_{t_n}I(\mathbf{t},\mathbf{t}')=\Phi_{t_n}(\mathbf{t})$], one finds $f=\Phi$. □

Representing $S(\psi(\mathbf{t},\lambda),\Phi^*(\mathbf{t}))$ as $K(\mathbf{t},\lambda)\exp\xi(\mathbf{t},\lambda)$ [with $K(\mathbf{t},\lambda)=\mathcal{O}(\lambda^{-1})$] and choosing $\mathbf{t}-\mathbf{t}'=\boldsymbol{\epsilon}(k)$, the relation (8) becomes

$$\mathrm{Res}_\lambda\left[\lambda^{-1}K(\mathbf{t},\lambda)\frac{\tau(\mathbf{t}-\boldsymbol{\epsilon}(k)+\boldsymbol{\epsilon}(\lambda))}{\tau(\mathbf{t}-\boldsymbol{\epsilon}(k))}\left(-1+\frac{2}{1-\lambda/k}\right)\right]=\Phi(\mathbf{t})-\Phi(\mathbf{t}-\boldsymbol{\epsilon}(k)),$$

or $2K(\mathbf{t},k)\tau(\mathbf{t})/\tau(\mathbf{t}-\boldsymbol{\epsilon}(k))=\Phi(\mathbf{t})-\Phi(\mathbf{t}-\boldsymbol{\epsilon}(k))$. Hence, one finds the following expression for the eigenfunction potential $S(\psi(\mathbf{t},\lambda),\Phi^*(\mathbf{t})\equiv\Phi_x(\mathbf{t}))$:

$$S(\psi(\mathbf{t},\lambda),\Phi^*(\mathbf{t}))=\tfrac{1}{2}(\Phi(\mathbf{t})-\Phi(\mathbf{t}-\boldsymbol{\epsilon}(\lambda)))\psi(\mathbf{t},\lambda). \tag{9}$$

The x derivative of this expression yields the relation

$$\Phi_x(\mathbf{t})+\Phi_x(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))=(\Phi(\mathbf{t})-\Phi(\mathbf{t}-\boldsymbol{\epsilon}(\lambda)))\left(\partial_x\log\frac{\tau(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))}{\tau(\mathbf{t})}+\lambda\right), \tag{10}$$

which is, in fact, another way of expressing the linear problem $\Phi_{t_n}=B_n\Phi\;\;(n:1,3,5,...)$; the reader may easily verify that relation (10) at $\mathcal{O}(\lambda^{-2})$ corresponds to the linear equation (7).

One may substitute relation (9) in Eq. (8) and use the BKP bilinear identity (3) to find

$$\mathrm{Res}_\lambda[\lambda^{-1}\Phi(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)]=2\Phi(\mathbf{t}')-\Phi(\mathbf{t}). \tag{11}$$

This identity generates bilinear equations that represent the BKP linear equations ($\Phi=\rho/\tau$):

$$\mathrm{Res}_\lambda[\lambda^{-1}\rho(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\tau(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))\exp\xi(\mathbf{t}-\mathbf{t}',\lambda)]=2\rho(\mathbf{t}')\tau(\mathbf{t})-\rho(\mathbf{t})\tau(\mathbf{t}'), \tag{12}$$

or

$$\sum_{j=0}^{\infty} p_j(-2\mathbf{y})p_j(2\tilde{D})e^{\Sigma_i y_i D_i}\tau\cdot\rho=2e^{\Sigma_i y_i D_i}\rho\cdot\tau-e^{\Sigma_i y_i D_i}\tau\cdot\rho,$$

with $\Sigma_{n=0}^{\infty}p_n(\mathbf{t})\lambda^n=\exp\xi(\mathbf{t},\lambda)$ and $\tilde{D}=(D_1,D_3/3,...)$. The simplest equation in this expression is the bilinear representation of Eq. (7):

$$(D_3-D_1^3)\rho\cdot\tau=0.$$

For the BKP hierarchy we introduce a new eigenfunction potential in terms of a pair of BKP eigenfunctions Φ_1 and Φ_2:

$$\Omega(\Phi_1,\Phi_2)=S(\Phi_2,\Phi_{1,x})-S(\Phi_1,\Phi_{2,x}), \tag{13}$$

i.e.,

$$\begin{aligned}\mathrm{d}\Omega(\Phi_1,\Phi_2)=(\Phi_{1,x}\Phi_2-\Phi_1\Phi_{2,x})\mathrm{d}x+[\Phi_{1,3x}\Phi_2-\Phi_1\Phi_{2,3x}-2\Phi_{1,2x}\Phi_{2,x}\\+2\Phi_{1,x}\Phi_{2,2x}+3u_2(\Phi_{1,x}\Phi_2-\Phi_1\Phi_{2,x})]\mathrm{d}t_3+\cdots.\end{aligned} \tag{14}$$

The potential is only defined up to a possible constant of integration. It is clear that $\Omega(\Phi_1,\Phi_2)=-\Omega(\Phi_2,\Phi_1)$ and that $\Omega(\Phi,1)=\Phi$ (remember that 1 is an eigenfunction), up to constants. In particular, all properties that will be proven for the eigenfunction potential Ω will also apply to the eigenfunctions themselves.

As wave functions are special eigenfunctions, we may compute the corresponding eigenfunction potential: from expression (9), representation (4), and definition (13), one finds

$$\Omega(\psi(\mathbf{t},\lambda),\psi(\mathbf{t},\mu))=\frac{\lambda-\mu}{\lambda+\mu}\,\frac{\tau(\mathbf{t}-\boldsymbol{\epsilon}(\lambda)-\boldsymbol{\epsilon}(\mu))}{\tau}\exp\xi(\mathbf{t},\lambda)+\xi(\mathbf{t},\mu)+C \tag{15}$$

(i.e. the BKP vertex operator[7] acting on τ) and taking $\lim_{\mu\to-\lambda}$ [with an appropriate choice of integration constant $C=\tilde{C}-2\lambda/(\mu+\lambda)$], one has

$$\Omega(\psi(\mathbf{t},\lambda),\psi(\mathbf{t},-\lambda))=4\sum_{n:1,3,...}\partial_{t_n}\log\tau\lambda^{-n}+2\sum_{n:1,3,...}nt_n\lambda^n+\tilde{C}. \tag{16}$$

For further reference, it is important to consider the effect of the shift $\mathbf{t}\to\mathbf{t}\pm\boldsymbol{\epsilon}(\lambda)$ on a BKP eigenfunction potential.

Property: Let $\Omega(\mathbf{t})\equiv\Omega(\Phi_1,\Phi_2)$; then

$$\Omega(\mathbf{t}\pm\boldsymbol{\epsilon}(\lambda))=\Omega(\mathbf{t})+\Phi_1(\mathbf{t}\pm\boldsymbol{\epsilon}(\lambda))\Phi_2(\mathbf{t})-\Phi_1(\mathbf{t})\Phi_2(\mathbf{t}\pm\boldsymbol{\epsilon}(\lambda)). \tag{17}$$

Proof: By definition (14), we have

$$\Omega_x(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))=\Phi_{1,x}(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\Phi_2(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))-\Phi_1(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\Phi_{2,x}(\mathbf{t}-\boldsymbol{\epsilon}(\lambda)).$$

The use of formula (10) in order to eliminate $\Phi_{1,x}(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))$ and $\Phi_{2,x}(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))$ yields

$$\left[-\Phi_{1,x}(\mathbf{t})+\Phi_1(\mathbf{t})\left(\partial_x \log\frac{\tau(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))}{\tau(\mathbf{t})}+\lambda\right)\right]\Phi_2(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))-\Phi_1(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))$$
$$\times\left[-\Phi_{2,x}(\mathbf{t})+\Phi_2(\mathbf{t})\left(\partial_x \log\frac{\tau(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))}{\tau(\mathbf{t})}+\lambda\right)\right].$$

Using formula (10) once more to eliminate the terms included in round brackets, we obtain

$$\Phi_{1,x}(\mathbf{t})\Phi_2(\mathbf{t})-\Phi_1(\mathbf{t})\Phi_{2,x}(\mathbf{t})+[\Phi_1(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\Phi_2(\mathbf{t})-\Phi_1(\mathbf{t})\Phi_2(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))]_x,$$

which proves expression (17). □

Next, we show that a BKP eigenfunction potential can always be completely expressed in terms of BKP tau functions.

Property: $\hat{\tau}\equiv\tau\Omega(\Phi_1,\Phi_2)$ is a BKP tau function:

$$\mathrm{Res}_\lambda[\lambda^{-1}\hat{\tau}(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\hat{\tau}(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))\exp\xi(\mathbf{t}-\mathbf{t}',\lambda)]=\hat{\tau}(\mathbf{t})\hat{\tau}(\mathbf{t}'). \tag{18}$$

Proof: Using representation (4), Eq. (18) is equivalent to

$$\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)\Omega(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\Omega(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))]=\Omega(\mathbf{t})\Omega(\mathbf{t}'),\quad \forall \mathbf{t},\mathbf{t}'$$

This relation may be proven by making use of relation (17), the BKP bilinear identity (3), and of relation (11) and its companion ($\lambda\to-\lambda$ and $\mathbf{t}\leftrightarrow\mathbf{t}'$):

$$\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)\Phi(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))]=2\Phi(\mathbf{t})-\Phi(\mathbf{t}'),$$

and the additional relation

$$\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)\Phi_i(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\Phi_j(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))]$$
$$=\Phi_i(\mathbf{t})\Phi_j(\mathbf{t}')-2\Omega(\Phi_i(\mathbf{t}),\Phi_j(\mathbf{t}))+2\Omega(\Phi_i(\mathbf{t}'),\Phi_j(\mathbf{t}')), \tag{19}$$

which is proven in the Appendix. □

This property shows that any eigenfunction potential can be written as the ratio of two tau functions, namely, $\Omega(\Phi_1,\Phi_2)=\hat{\tau}/\tau$.

As the eigenfunction potential $\Omega(\Phi_1,\Phi_2)$ is only defined up to an additive constant, we have as a consequence that $\tau\Omega(\Phi_1,\Phi_2)+C\tau$ also satisfies the BKP bilinear identity. It then follows from the above property that $\tau\Omega(\Phi_1,\Phi_2)$ is a *symmetry* for Eq. (5), meaning that $\tau+C^{-1}\tau\Omega$ satisfies the BKP bilinear identity (5) up to first order in C^{-1}. For example, the eigenfunction potential (16) generates the translation symmetries τ_{t_n} and the gauge symmetries $t_n\tau$.

Since any eigenfunction Φ can be written as the eigenfunction potential $\Omega(\Phi,1)=\Phi$, it follows from this property that the eigenfunction Φ is also expressible as $\Phi=\rho/\tau$, where ρ is a BKP tau function. It thus follows that the bilinear identity (12) constitutes a bilinear Bäcklund transformation for the bilinear BKP hierarchy. It should be clear that $\tau\Phi$ is a symmetry for the bilinear BKP equations.

III. REDUCTIONS

In this section we explore certain symmetry reductions of the BKP hierarchy. We then investigate the connection between these reductions and constraints on the BKP Lax operator L. We shall also derive a bilinear formulation of these constraints.

The k-constrained BKP hierarchy is defined by coupling the linear equations for a pair of eigenfunctions Φ_1 and Φ_2 to the nonlinear equations of the BKP hierarchy by means of the constraint

$$\tau_{t_k} = \tau\Omega(\Phi_1,\Phi_2) \tag{20}$$

(where k is a fixed odd integer). Since both the left-hand side and the right-hand side of this condition are symmetries of the BKP bilinear identity, it induces a so-called symmetry reduction on this hierarchy.

As an example, let us consider the case $k=1$. Denoting $q=\Phi_1$, $r=\Phi_2$, it follows from Eq. (7), constraint (20) and the definition (14) that (q,r) solves the system:

$$\begin{aligned} q_{t_3} &= q_{3x} + 6(q_x r - q r_x) q_x, \\ r_{t_3} &= r_{3x} + 6(q_x r - q r_x) r_x. \end{aligned} \tag{21}$$

Setting $r=\frac{1}{2}$ (a genuine eigenfunction), one finds that the field q_x satisfies the KdV equation $(q_{x,t_3} = q_{4x} + 6q_x q_{2x})$. This last equation is the result that was already derived,[1,2] but it should be clear that it is merely a special case of the 1-constrained BKP system (21). As a second example we take $k=3$ at which the constraint (20) yields the following system:

$$\begin{aligned} q_{t_3} &= q_{3x} + 3u q_x, \\ r_{t_3} &= r_{3x} + 3u r_x, \\ u_{t_3} &= 2(q_{2x} r - q r_{2x}). \end{aligned} \tag{22}$$

The special case $r=\frac{1}{2}$ $(q_{t_3} = q_{3x} + 3uq_x,\ u_{t_3} = q_{2x})$ was already given in Ref. 2. As far as the authors could verify, neither of the systems (21) or (22) has already been described in the literature.

In order to establish a connection between the symmetry reduction (20) and constraints on the Lax operator L, we consider $\Omega(\mathbf{t}) = \Omega(\Phi_1,\Phi_2)$ and $\tilde{\Omega}(\mathbf{t}) = \Omega(\psi(\mathbf{t},\lambda),\psi(\mathbf{t},-\lambda))$. By virtue of expression (16) for $\tilde{\Omega}(\mathbf{t})$, we may write the condition (20) as

$$\mathrm{Res}_\lambda[\lambda^{k-1}\tilde{\Omega}(\mathbf{t})] = 4\Omega(\mathbf{t}),$$

or, equivalently,

$$\mathrm{Res}_\lambda[\lambda^{k-1}\,\partial_{t_n}\tilde{\Omega}(\mathbf{t})] = 4\,\partial_{t_n}\Omega(\mathbf{t}), \quad \forall n{:}1,3,5,\ldots. \tag{23}$$

At $n=1$, this yields the relation

$$\mathrm{Res}_\lambda[\lambda^{k-1}(\psi_x(\mathbf{t},\lambda)\psi(\mathbf{t},-\lambda) - \psi(\mathbf{t},\lambda)\psi_x(\mathbf{t},-\lambda))] = 4(\Phi_{1,x}\Phi_2 - \Phi_1\Phi_{2,x}). \tag{24}$$

Since $\mathrm{Res}_\lambda[f(\lambda)] = \frac{1}{2}\ \mathrm{Res}_\lambda[f(\lambda) - f(-\lambda)]$, this can be written as

$$\mathrm{Res}_\lambda[\lambda^{k-1}\psi(\mathbf{t},\lambda)\psi_x(\mathbf{t},-\lambda)] = 2(\Phi_1\Phi_{2,x} - \Phi_{1,x}\Phi_2).$$

The x derivative of relation (24) similarly yields:

$$\mathrm{Res}_\lambda[\lambda^{k-1}\psi(\mathbf{t},\lambda)\psi_{2x}(\mathbf{t},-\lambda)] = 2(\Phi_1\Phi_{2,2x} - \Phi_{1,2x}\Phi_2).$$

Combining relation (23) at $n=3$ [bearing definition (14) in mind] and the second x derivative of relation (24), one finds

$$\mathrm{Res}_\lambda[\lambda^{k-1}\psi(\mathbf{t},\lambda)\psi_{3x}(\mathbf{t},-\lambda)] = 2(\Phi_1\Phi_{2,3x} - \Phi_{1,3x}\Phi_2).$$

In this way, it can be seen that (23) is equivalent to

$$\mathrm{Res}_\lambda[\lambda^{k-1}\psi(\mathbf{t},\lambda)\partial^{m+1}\psi(\mathbf{t},-\lambda)]=2(\Phi_1\Phi_{2,(m+1)x}-\Phi_2\Phi_{1,(m+1)x}), \tag{25}$$

for $m\geqslant -1$. [The case $m=-1$ follows from the definition (1): $(L^k)_+=B_k$ and $B_k 1=0$.] Keeping relations (1) and (2) in mind, one finds

$$\mathrm{Res}_\partial[L^k\partial^m]=2(-1)^m(\Phi_2\Phi_{1,(m+1)x}-\Phi_1\Phi_{2,(m+1)x}).$$

Hence, the constraint (20) is equivalent to requiring that

$$L^k=B_k+2\sum_{n\geqslant 1}(-1)^n(\Phi_1\Phi_{2,nx}-\Phi_2\Phi_{1,nx})\partial^{-n},$$

or to the following condition the BKP Lax operator:

$$L^k=B_k+2\Phi_2\,\partial^{-1}\Phi_{1,x}-2\Phi_1\,\partial^{-1}\Phi_{2,x}. \tag{26}$$

This form of reduction of the BKP hierarchy was already introduced in Refs. 1, 2. It was, however, limited to the case $\Phi_2=\frac{1}{2}$. It is our feeling that imposing a constraint on the Lax operator L to reduce the BKP equation is conceptually less clear than using a symmetry constraint. Furthermore, in the former case one must make sure that any constraints on the Lax operator do not violate the condition (1).

We may now easily derive a bilinear form for the constrained BKP hierarchies. Since wave functions and eigenfunctions obey the same evolution, $\psi_{t_n}=B_n\psi$ and $\Phi_{i,t_n}=B_n\Phi_i$ $(i:1,2)$, relation (25) implies

$$\mathrm{Res}_\lambda[\lambda^{k-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)]=2(\Phi_1(\mathbf{t})\Phi_2(\mathbf{t}')-\Phi_1(\mathbf{t}')\Phi_2(\mathbf{t})),\quad \forall \mathbf{t},\mathbf{t}'. \tag{27}$$

If we write the eigenfunctions and the wave functions in terms of tau functions ($\Phi_1=\rho_1/\tau$ and $\Phi_2=\rho_2/\tau$), we find the following bilinear formulation of the constraint (20):

$$\mathrm{Res}_\lambda[\lambda^{k-1}\tau(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\tau(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))\exp\xi(\mathbf{t}-\mathbf{t}',\lambda)]=2(\rho_1(\mathbf{t})\rho_2(\mathbf{t}')-\rho_1(\mathbf{t}')\rho_2(\mathbf{t})).$$

In terms of Hirota operators, this yields

$$\sum_{j=0}^{\infty}p_j(2\mathbf{y})p_{j+k}(-2\tilde{D})e^{\Sigma_i y_i D_i}\tau\cdot\tau=2e^{\Sigma_i y_i D_i}(\rho_1\cdot\rho_2-\rho_2\cdot\rho_1),$$

which implies, e.g., that $D_1\,D_k\tau\cdot\tau=2D_1\rho_1\cdot\rho_2$ [the x-derivative of (20)]. Hence, the bilinear form of the system (21) is

$$(D_3-D_1^3)\rho_1\cdot\tau=0,\quad (D_3-D_1^3)\rho_2\cdot\tau=0,\quad D_1^2\tau\cdot\tau=2D_1\rho_1\cdot\rho_2,$$

with $q=\rho_1/\tau$ and $r=\rho_2/\tau$.

A last important result is the following: as the product $\tau\Omega(\Phi_1,\Phi_2)$ is a BKP tau function, the symmetry reduction (20) implies the auxiliary bilinear identity

$$\mathrm{Res}_\lambda[\lambda^{-1}\tau_{t_k}(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\tau_{t_k}(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))\exp\xi(\mathbf{t}-\mathbf{t}',\lambda)]=\tau_{t_k}(\mathbf{t})\tau_{t_k}(\mathbf{t}'),$$

for a constrained BKP tau function: i.e., the t_k derivative of a constrained BKP tau function is itself a BKP tau function.

IV. SOLUTIONS

In this section we discuss solutions of the constrained BKP hierarchies. The tau functions corresponding to these solutions are expressed as ''Pfaffians.'' A Pfaffian is defined as the square root of the determinant of an antisymmetric matrix of even order; it is a polynomial in the elements of this matrix. For example, the Pfaffians of a 2×2 and 4×4 antisymmetric matrix A (with elements $a_{ij}=-a_{ji}$) are $\mathrm{Pf}(A)=a_{12}$ and $\mathrm{Pf}(A)=a_{12}a_{34}-a_{13}a_{24}+a_{14}a_{23}$, respectively.

Solutions in Pfaffian form for the bilinear BKP equation (6) were first derived in Ref. 11. These solutions may be obtained for all the equations in the bilinear BKP hierarchy (5) by starting from the well-known Grammian determinant solutions $\tau_{\mathrm{KP}}=\det[S(\varphi_i,\varphi_j^*)]_{1\le i,j\le 2N}$ [with $\varphi_{i,t_n}=\varphi_{i,nx}$ and $\varphi^*_{i,t_n}=(-1)^{n+1}\varphi^*_{i,nx}$ for $n:1,2,3,...$] of the KP hierarchy[17,18] and by using the fact that $\tau_{\mathrm{BKP}}=\tau_{\mathrm{KP}}\ (t_2=t_4=\cdots=0)^{1/2}$: Fixing $t_2=t_4=\cdots=0$ and choosing $\varphi_i=f_i$ and $\varphi_j^*=f_{j,x}$, one finds that $S(\varphi_i,\varphi_j^*)=\frac{1}{2}[f_if_j-\Omega(f_i,f_j)]$.[12] The determinant τ_{KP} is now easily seen to equal $\det[\Omega(f_i,f_j)]_{1\le i,j\le 2N}$ (up to a multiplicative constant, which is irrelevant in our discussion). The square root of this determinant yields the following solution to the BKP bilinear hierarchy:

$$\tau_{\mathrm{BKP}}=\sqrt{\det[\Omega(f_i,f_j)]_{1\le i,j\le 2N}}\equiv \mathrm{Pf}(f_1,f_2,\ldots,f_{2N}) \tag{28}$$

(with $f_{i,t_n}=f_{i,nx}$ for $n:1,3,5,...$). The expression $\mathrm{Pf}(f_1,\ldots,f_{2N})$ has the form $\Omega(f_1,f_2)\Omega(f_3,f_4)\cdots\Omega(f_{2N-1},f_{2N})-\cdots$ [only containing potentials of the type $\Omega(f_i,f_{j>i})$]. For example, for $N=1$ one finds the BKP tau function $\tau=\Omega(f_1,f_2)$, for $N=2$ one finds the tau function $\tau=\Omega(f_1,f_2)\Omega(f_3,f_4)-\Omega(f_1,f_3)\Omega(f_2,f_4)+\Omega(f_1,f_4)\Omega(f_2,f_3)$.

The notation introduced in (28) allows us to express the properties of such Pfaffians more easily: expression (28) is multilinear [i.e., $\mathrm{Pf}(\ldots,f+g,\ldots)=\mathrm{Pf}(\ldots,f,\ldots)+\mathrm{Pf}(\ldots,g,\ldots)$] and allows for the existence of expansion rules of the form $\mathrm{Pf}(f_1,\ldots,f_{2N})=\Sigma_{i=2}^{2N}(-1)^i\times \mathrm{Pf}(f_1,f_i)\mathrm{Pf}(f_2,\ldots,\hat{f}_i,\ldots,f_{2N})$.[11]

It also follows from the multilinearity that the derivative (w.r.t. any variable) is given by

$$\frac{\partial}{\partial t}\mathrm{Pf}(f_1,\ldots,f_{2N})=\sum_{i=1}^{2N}\mathrm{Pf}\left(f_1,\ldots,\frac{\partial}{\partial t}f_i,\ldots,f_{2N}\right). \tag{29}$$

Before discussing the solutions of the constrained BKP equations, we need expressions for eigenfunctions and eigenfunction potentials that correspond to the above BKP tau functions. We start from the representation $\Phi=\rho_{\mathrm{KP}}/\tau_{\mathrm{KP}}$ of eigenfunctions in terms of KP tau functions. We use the following expression for a ρ_{KP}, which can be seen to correspond to the above τ_{KP}:[17]

$$\rho_{\mathrm{KP}}=\det\begin{bmatrix} S(\varphi_1,\varphi_1^*) & \cdots & S(\varphi_1,\varphi_{2N}^*) & \varphi_1\\ \vdots & \ddots & \vdots & \vdots\\ S(\varphi_{2N},\varphi_1^*) & & S(\varphi_{2N},\varphi_{2N}^*) & \varphi_{2N}\\ -S(\varphi,\varphi_1^*) & \cdots & -S(\varphi,\varphi_{2N}^*) & -\varphi\end{bmatrix}.$$

Let A be the antisymmetric matrix corresponding to the BKP tau function (28): $A=[\Omega(f_i,f_j)]_{1\le i,j\le 2N}$, and introduce the two column vectors $\mathbf{f}=(f_1,f_2,\ldots,f_{2N})^t$ and $\mathbf{\Omega}=(\Omega(f_1,f_{2N+1}),\ldots,\Omega(f_{2N},f_{2N+1}))^t$. Choosing again $\varphi_i=f_i$, $\varphi_j^*=f_{j,x}$ and $\varphi=f_{2N+1}$, we find

$$\rho_{\mathrm{KP}}=\det\begin{bmatrix} A & \mathbf{f}\\ \mathbf{\Omega}^t & -f_{2N+1}\end{bmatrix}=\mathrm{Pf}(A)\mathrm{Pf}\begin{bmatrix} A & \mathbf{\Omega} & \mathbf{f}\\ -\mathbf{\Omega}^t & 0 & f_{2N+1}\\ -\mathbf{f}^t & -f_{2N+1} & 0\end{bmatrix}$$

(where we have ignored an unimportant multiplicative factor and used a Jacobi determinant identity). Hence, there exist eigenfunctions of the form

$$\Phi=\frac{\rho_{\mathrm{KP}}}{\tau_{\mathrm{KP}}}=\frac{\mathrm{Pf}(f_1,f_2,\ldots,f_{2N},f_{2N+1},1)}{\mathrm{Pf}(f_1,f_2,\ldots,f_{2N})}=\frac{\rho_{\mathrm{BKP}}}{\tau_{\mathrm{BKP}}},$$

with the understanding that here $\Omega(f_i,1)\equiv f_i$ (no integration constants allowed) and $f_{i,t_n}=f_{i,nx}$ with n:1,3,5,... . With this convention of Pf(...,1) one has, e.g., at $N=1$: $\Phi=\mathrm{Pf}(f_1,f_2,f_3,1)/\mathrm{Pf}(f_1,f_2)\equiv(\Omega(f_1,f_2)f_3-\Omega(f_1,f_3)f_2+f_1\Omega(f_2,f_3))/\Omega(f_1,f_2)$.

Special choices of f_{2N+1} include $f_{2N+1}=1$, for which one recovers $\Phi=1$ and $f_{2N+1}=0$, for which one finds

$$\Phi=\frac{\mathrm{Pf}(f_1,f_2,\ldots,f_{2N-1},1)}{\mathrm{Pf}(f_1,f_2,\ldots,f_{2N})} \tag{30}$$

[in the latter case by choosing $\Omega(f_{2N},f_{2N+1}=0)=1$ and $\Omega(f_i,f_{2N+1}=0)=0$ for $1\leqslant i<2N$].

The eigenfunction potential corresponding to the eigenfunctions $\Phi_1=\mathrm{Pf}(f_1,\ldots,f_{2N},f_{2N+1},1)/\mathrm{Pf}(f_1,\ldots,f_{2N})$ and $\Phi_2=\mathrm{Pf}(f_1,\ldots,f_{2N},f_{2N+2},1)/\mathrm{Pf}(f_1,\ldots,f_{2N})$ can be shown to take on the form

$$\Omega(\Phi_1,\Phi_2)=\frac{\mathrm{Pf}(f_1,\ldots,f_{2N},f_{2N+1},f_{2N+2})}{\mathrm{Pf}(f_1,\ldots,f_{2N})}. \tag{31}$$

This expression is obtained by proving that $\Omega_x=\Phi_{1,x}\Phi_2-\Phi_1\Phi_{2,x}$, with the help of the techniques used in Ref. 11 [in particular, by using relation (2.25) of Pfaffians in this reference]. The expression (31) agrees with the fact that such a potential is always the ratio of two BKP tau functions. Specifically, the eigenfunction potential corresponding to the eigenfunction Φ_1 in (30) is

$$\Omega(\Phi_1,\Phi_2)=\frac{\mathrm{Pf}(f_1,\ldots,f_{2N-1},f_{2N+2})}{\mathrm{Pf}(f_1,\ldots,f_{2N})}. \tag{32}$$

In order to find a solution of the constrained BKP equations, we need to find a BKP tau function and two eigenfunctions such that condition (20) is satisfied. In particular, we need to find a tau function such that its t_k derivative is again a BKP tau function. From property (29), we find that

$$\frac{\partial}{\partial t_k}\mathrm{Pf}(f,f_{t_k},\ldots,f_{(2N-1)t_k})=\mathrm{Pf}(f,f_{t_k},\ldots,f_{(2N-2)t_k},f_{2Nt_k}). \tag{33}$$

In the light of formula (32) and definition (20), this formula tells us that we may choose the k-constrained tau function as $\tau=\mathrm{Pf}(f,f_{t_k},\ldots,f_{(2N-1)t_k})$ with the corresponding eigenfunctions $\Phi_1=\rho_1/\tau$ $\Phi_2=\rho_2/\tau$:

$$\begin{aligned}\Phi_1&=\frac{\rho_1}{\tau}=\frac{\mathrm{Pf}(f,f_{t_k}\cdots,f_{(2N-2)t_k},1)}{\mathrm{Pf}(f,f_{t_k}\cdots,f_{(2N-1)t_k})},\\ \Phi_2&=\frac{\rho_2}{\tau}=\frac{\mathrm{Pf}(f,f_{t_k}\cdots,f_{(2N-1)t_k},f_{2Nt_k},1)}{\mathrm{Pf}(f,f_{t_k}\cdots,f_{(2N-1)t_k})}\end{aligned} \tag{34}$$

($f_{t_n}=f_{nx}$ n:1,3,5,...). Here, some care should be taken with respect to the integration constants in the eigenfunction potentials: there are $N(2N-1)$ eigenfunction potentials $\Omega(f_{it_k},f_{jt_k})$ in the above constrained tau function τ. The potentials with $i+j$ even should contain no integration constants [as formula (33) assumes, e.g., that $f_x^2-ff_{2x}\equiv\partial_x\Omega(f,f_x)=\Omega(f,f_{2x})$]. All potentials $\Omega(f_{it_k},f_{jt_k})$ for fixed $i+j$ odd are (for the same reason) connected. Hence, there are only $2N-1$ independent integration constants in the above constrained BKP tau function τ (i

$+j$:1,3,...,$4N-3)$. In ρ_2 there is one eigenfunction potential that does not appear in τ: $\Omega(f_{(2N-1)t_k},f_{2Nt_k})$; it may contain an integration constant, but this only amounts to $\Phi_2\to\Phi_2+c\Phi_1$ [a transformation that leaves the constraint (20) invariant].

For example, for $k=1$ (i.e., $t_k=x$), one may choose $N=1$, $f=x$ and find

$$\Phi_1=\frac{x}{x+c},\quad \Phi_2=\frac{-1}{x+c},$$

as an example of rational solutions of the systems (21). Another rational solution can be found by choosing $f=x^3/6+t_3$:

$$\Phi_1=\frac{10(x^3+6t_3)}{x^5-30x^2t_3+c},\quad \Phi_2=\frac{x^6-60x^3t_3+6cx-720t_3^2}{6(x^5-30x^2t_3+c)}.$$

The one-soliton solution of this system may be obtained by choosing $N=1$ and $f=\exp\xi(\mathbf{t},p_1)+\exp\xi(\mathbf{t},p_2)$. After some straightforward manipulations one finds ($p_1=k_1+ik_2$, $p_2=k_1-ik_2$):

$$\Phi_1=\frac{-k_1}{\sqrt{k_2}}\frac{\cos\theta_2}{\cosh\theta_1},\quad \Phi_2=\frac{k_1}{\sqrt{k_2}}\frac{\sin\theta_2}{\cosh\theta_1},$$

with $\theta_1=k_1x+k_1(k_1^2-3k_2^2)t_3+\cdots$ and $\theta_2=k_2x+k_2(3k_1^2-k_2^2)t_3+\cdots$. One has $\Phi_{1,x}\Phi_2-\Phi_2\Phi_{2,x}=\partial_x^2\log\tau=k_1^2\,\mathrm{sech}^2\,\theta_1$.

V. CONCLUSIONS

In this paper we deal with a novel class of reductions of the BKP hierarchy and as such these reductions allow us to identify certain new integrable systems.

The central object in our discussion is the BKP eigenfunction potential. It was shown to be expressible as a ratio of tau functions and to correspond to a (general) symmetry of the bilinear BKP equations, i.e., for the equations the BKP tau functions satisfy.

The eigenfunction symmetry was then used to define dimensional reductions of the $2+1$-dimensional BKP equations. A link with constraints on the Lax operator was established and a simple bilinear formulation was derived. We wish to remark that in analogy to the reductions of the KP hierarchy one could consider a generalization of these reductions by using a sum of eigenfunction potential symmetries in the symmetry reduction.

Finally, we believe it is worth emphasizing that our tau-function approach to symmetry reductions has, besides a certain conceptual clarity, the additional virtue of allowing for the straightforward derivation of a class of solutions to the reduced systems. In the present case these solutions were easily expressed in terms of Pfaffians.

ACKNOWLEDGMENTS

I. L. and R. W. are postdoctoral fellows at the F.W.O.-Vlaanderen (Belgium). The authors acknowledge the financial support extended within the framework of the ''Interuniversity poles of attraction programme,'' Contract No. P4/08—Belgian State, Prime Minister's Office—Federal Office for Scientific, Technical and Cultural affairs. R. W. is the beneficiary of a JSPS postdoctoral fellowship for foreign researchers in Japan.

APPENDIX: PROOF OF RELATION (19)

To prove relation (19) (e.g., denoting Φ_i as Φ_1 and Φ_j as Φ_2), we use the formula (9) for both $S(\psi(\mathbf{t},\lambda),\Phi_1^*(\mathbf{t}))$ and $S(\psi(\mathbf{t}',-\lambda),\Phi_2^*(\mathbf{t}'))$, together with the equation (8):

$$\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)\Phi_i(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\Phi_j(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))]$$
$$=4\,\mathrm{Res}_\lambda[\lambda^{-1}S(\psi(\mathbf{t},\lambda),\Phi_1^*(\mathbf{t}))S(\psi(\mathbf{t}',-\lambda),\Phi_2^*(\mathbf{t}'))]-3\Phi_1(\mathbf{t})\Phi_2(\mathbf{t}')$$
$$+2\Phi_1(\mathbf{t})\Phi_2(\mathbf{t})+2\Phi_1(\mathbf{t}')\Phi_2(\mathbf{t}'). \tag{A1}$$

Let us denote the residue on the right-hand side by $I(\mathbf{t},\mathbf{t}')$. We shall now try to compute $\partial_{t_n}\partial_{t'_m}I(\mathbf{t},\mathbf{t}')$. As mentioned in Sec. II, the derivative of $S(\psi,\Phi^*)$ is given by $(B_n=\Sigma_{j=0}^n b_{n,j}\partial_x^j)$:

$$\partial_{t_n}S(\psi,\Phi^*)=\sum_{j=1}^{n}\sum_{i=1}^{j}(-1)^{i+1}(\Phi^* b_{n,j})_{(i-1)x}\psi_{(j-i)x}. \tag{A2}$$

Hence, $\partial_{t_n}\partial_{t'_m}I(\mathbf{t},\mathbf{t}')$ will have the form of an operator in the variables x and x' working on the residue $\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)]$. But, by virtue of the BKP bilinear identity (3), all terms will vanish except the ones without derivatives on $\psi(\mathbf{t},\lambda)$ and $\psi(\mathbf{t}',-\lambda)$. These correspond to the terms with $i=j$ in expression (A2):

$$\sum_{j=1}^{n}(-1)^{j+1}(\Phi^* b_{n,j})_{(j-1)x}\psi=\psi(B_n\partial_x^{-1})^*\Phi^*. \tag{A3}$$

As $(B_n\partial_x^{-1})^*=-\partial_x^{-1}B_n^*=\partial^{-1}(B_n+B_{n,x}\partial^{-1})$ and $\Phi^*=\Phi_x$, we find that this equals

$$\psi\partial^{-1}(B_n+B_{n,x}\partial^{-1})\partial_x\Phi=\psi B_n\Phi=\psi\Phi_{t_n}. \tag{A4}$$

In this way we find that

$$\partial_{t_n}\partial_{t'_m}I(\mathbf{t},\mathbf{t}')=\Phi_{1,t_n}(\mathbf{t})\Phi_{2,t'_m}(\mathbf{t}').$$

It follows that

$$I(\mathbf{t},\mathbf{t}')=\Phi_1(\mathbf{t})\Phi_2(\mathbf{t}')+f(\mathbf{t})+g(\mathbf{t}'),$$

for some f and g. As $I(\mathbf{t},\mathbf{t}'=\mathbf{t})=0$ [since $S(\psi(\mathbf{t},\lambda),\Phi^*)=\mathcal{O}(\lambda^{-1})\exp\xi(\mathbf{t},\lambda)$, we have $g=-f-\Phi_1\Phi_2$ and hence $I(\mathbf{t},\mathbf{t}')=\Phi_1(\mathbf{t})\Phi_2(\mathbf{t}')+f(\mathbf{t})-f(\mathbf{t}')-\Phi_1(\mathbf{t}')\Phi_2(\mathbf{t}')$]. To find an expression for f, we compute $\partial_{z_n}I(\mathbf{t}-\mathbf{z},\mathbf{t}+\mathbf{z})_{\mathbf{z}=\mathbf{0}}=-2f_{t_n}(\mathbf{t})-2\Phi_{1,t_n}(\mathbf{t})\Phi_2(\mathbf{t})$ or, equivalently,

$$-\mathrm{Res}_\lambda[\lambda^{-1}S_{t_n}(\psi(\mathbf{t},\lambda),\Phi_1^*(\mathbf{t}))S(\psi(\mathbf{t},-\lambda),\Phi_2^*(\mathbf{t}))]$$
$$+\mathrm{Res}_\lambda[\lambda^{-1}S(\psi(\mathbf{t},\lambda),\Phi_1^*(\mathbf{t}))S_{t_n}(\psi(\mathbf{t},-\lambda),\Phi_2^*(\mathbf{t}))].$$

Representing $\partial_{t_n}S(\psi,\Phi^*)=\Sigma_{i,j=0}^{n-1}a_{ij}\psi_{ix}\Phi_{jx}^*$, we have

$$=-\sum_{i,j=0}^{n-1}a_{ij}\Phi_{1,jx}^*(\mathbf{t})\mathrm{Res}_\lambda[\lambda^{-1}\,\partial_x^i\psi(\mathbf{t},\lambda)S(\psi(\mathbf{t},-\lambda),\Phi_2^*(\mathbf{t}))]$$
$$+\sum_{i,j=0}^{n-1}a_{ij}\Phi_{2,jx}^*(\mathbf{t})\mathrm{Res}_\lambda[\lambda^{-1}S(\psi(\mathbf{t},\lambda),\Phi_1^*(\mathbf{t}))\partial_x^i\psi(\mathbf{t},-\lambda)].$$

These last residues equal $(\partial_x^j[\Phi_2(\mathbf{t}')-\Phi_2(\mathbf{t})])_{\mathbf{t}=\mathbf{t}'}=-\Phi_{2,jx}(\mathbf{t})+\delta_{j0}\Phi_2(\mathbf{t})$ and $(\partial_{x'}^j[\Phi_1(\mathbf{t})-\Phi_1(\mathbf{t}')])_{\mathbf{t}=\mathbf{t}'}=-\Phi_{1,jx}(\mathbf{t})+\delta_{j0}\Phi_1(\mathbf{t})$, respectively, by virtue of the relation (8). Hence

$$= -\sum_{i,j=0}^{n-1} a_{ij}\Phi^*_{1,jx}(\mathbf{t})(-\Phi_{2,ix}(\mathbf{t})+\delta_{i0}\Phi_2(\mathbf{t}))+\sum_{i,j=0}^{n-1} a_{ij}\Phi^*_{2,jx}(\mathbf{t})(-\Phi_{1,ix}(\mathbf{t})+\delta_{i0}\Phi_1(\mathbf{t})),$$

which—due to the reasoning that leads from (A3) to (A4)—equals

$$S_{t_n}(\Phi_2,\Phi_1^*)-\Phi_2\Phi_{1,t_n}-S_{t_n}(\Phi_1,\Phi_2^*)+\Phi_1\Phi_{2,t_n}=\partial_{t_n}\Omega(\Phi_1,\Phi_2)-\Phi_2\Phi_{1,t_n}+\Phi_1\Phi_{2,t_n}.$$

Hence, we find that $-2f_{t_n}(\mathbf{t})-2\Phi_{1,t_n}(\mathbf{t})\Phi_2(\mathbf{t})=\partial_{t_n}\Omega(\Phi_1,\Phi_2)-\Phi_2\Phi_{1,t_n}+\Phi_1\Phi_{2,t_n}$ or

$$f=-\tfrac{1}{2}\Omega(\Phi_1,\Phi_2)-\tfrac{1}{2}\Phi_1\Phi_2.$$

It then follows that

$$I(\mathbf{t},\mathbf{t}')=\Phi_1(\mathbf{t})\Phi_2(\mathbf{t}')-\tfrac{1}{2}\Omega(\Phi_1(\mathbf{t}),\Phi_2(\mathbf{t}))+\tfrac{1}{2}\Omega(\Phi_1(\mathbf{t}'),\Phi_2(\mathbf{t}'))-\tfrac{1}{2}(\Phi_1(\mathbf{t})\Phi_2(\mathbf{t})+\Phi_1(\mathbf{t}')\Phi_2(\mathbf{t}')),$$

and thus (A1) becomes

$$\mathrm{Res}_\lambda[\lambda^{-1}\psi(\mathbf{t},\lambda)\psi(\mathbf{t}',-\lambda)\Phi_i(\mathbf{t}-\boldsymbol{\epsilon}(\lambda))\Phi_j(\mathbf{t}'+\boldsymbol{\epsilon}(\lambda))]=\Phi_i(\mathbf{t})\Phi_j(\mathbf{t}')-2\Omega(\Phi_i(\mathbf{t}),\Phi_j(\mathbf{t}))+2\Omega(\Phi_i(\mathbf{t}'),\Phi_j(\mathbf{t}')).$$

[1] A. Y. Orlov, ''Volterra operator algebra for zero curvature representation. Universality of KP,'' in *Nonlinear Processes in Physics*, edited by A. S. Fokas, D. J. Kaup, A. C. Newell, and V. E. Zakharov (Springer-Verlag, Berlin 1991), pp. 126–131.

[2] Y. Cheng, ''Constraints of the Kadomtsev–Petviashvili hierarchy,'' J. Math. Phys. **33**, 3774–3782 (1992).

[3] H. Aratyn, E. Nissimov, and S. Pacheva, ''Constrained KP hierarchies: additional symmetries, Darboux–Bäcklund solutions and relations to multi-matrix models,'' Int. J. Mod. Phys. A **12**, 1265–1340 (1997).

[4] I. Loris and R. Willox, ''KP symmetry reductions and a generalized constraint,'' J. Phys. A **30**, 6925–6938 (1997).

[5] B. Konopelchenko and W. Strampp, ''New reductions of the Kadomtsev–Petviashvili and two-dimensional Toda lattice hieararachies via symmetry constraints,'' J. Math. Phys. **33**, 3676–3686 (1992).

[6] G. Helminck and J. W. van de Leur, ''An analytic description of the vector constrained KP hierarchy,'' Commun. Math. Phys. **193**, 627–641 (1998).

[7] E. Date, M. Kashiwara, M. Jimbo, and T. Miwa, ''Transformation groups for soliton equations,'' in *Proceedings of the RIMS Symposium on Non-Linear Integrable Systems—Classical Theory and Quantum Theory*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983), pp. 39–119.

[8] J. J. C. Nimmo, ''Darboux transformations from reductions of the KP hierarchy,'' in *Nonlinear Evolution Equations and Dynamical Systems*, edited by V. G. Makhankov, A. R. Bishop, and D. D. Holm (World Scientific, Singapore, 1995), pp. 168–177.

[9] W. Oevel and W. Schief, ''Darboux theorems and the KP hierarchy,'' in *Applications of Analytic and Geometric Methods to Nonlinear Differential Equations*, edited by P. A. Clarkson (Kluwer Academic, Dordrecht, 1993), pp. 193–206.

[10] H. Aratyn, E. Nissimov, and S. Pacheva, ''Method of squared eigenfunction potentials in integrable hierarchies of KP type,'' Commun. Math. Phys. **193**, 493–525 (1998).

[11] R. Hirota, ''Soliton solutions to the BKP equations. I. the Pfaffian technique,'' J. Phys. Soc. Jpn. **58**, 2285–2296 (1989).

[12] J. J. C. Nimmo and R. Willox, ''Darboux transformations for the two-dimensional Toda lattice,'' Proc. R. Soc. London, Ser. A **453**, 2497–2525 (1997).

[13] M. Jimbo and T. Miwa, ''Solitons and infinite dimensional Lie algebras,'' Publ. RIMS Kyoto Univ. **19**, 943–1001 (1983).

[14] L. Dickey, ''Soliton equations and Hamiltonian systems,'' Vol. 12 of *Advanced Series in Mathematical Physics* (World Scientific, Singapore, 1991).

[15] R. Hirota, ''Direct methods in soliton theory,'' in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer-Verlag, Berlin, 1980), pp. 157–176.

[16] W. Oevel, ''Darboux theorems and Wronskian formulas for integrable systems I. Constrained KP flows,'' Physica A **195**, 533–576 (1993).

[17] I. Loris and R. Willox, ''On solutions of constrained Kadomtsev–Petviashvili equations: Grammians,'' J. Math. Phys. **38**, 5190–5197 (1997).

[18] A. Nakamura, ''A bilinear *N*-soliton formula for the KP equation,'' J. Phys. Soc. Jpn. **58**, 412–422 (1989).

On dimension of the global attractor for damped nonlinear wave equations

Zhou Shengfan[a)]
Department of Mathematics, Sichuan Union University Chengdu, 610064, People's Republic of China

(Received 10 August 1998; accepted for publication 21 September 1998)

In this paper, we obtain a more precise estimate of upper bound of the Hausdorff dimension of the global attractor for damped nonlinear wave equations with the Dirichlet boundary condition. The obtained Hausdorff dimension decreases as the damping grows and is uniformly bounded for large damping, which conforms to physical intuition. © *1999 American Institute of Physics.*
[S0022-2488(99)04502-8]

I. INTRODUCTION

Consider the initial-boundary value problem of the damped nonlinear wave equation

$$u_{tt}+\alpha u_t-\Delta u+f(u)=g, \quad x\in\Omega, \ \ t>0,$$

$$u(x,t)|_{x\in\partial\Omega}=0, \quad t>0, \tag{1}$$

$$u(x,0)=u_0(x), \quad \frac{\partial u}{\partial t}(x,0)=u_1(x), \quad x\in\Omega,$$

where $u=u(x,t)$ is a real-valued function on $\Omega\times[0,+\infty)$, Ω is an open bounded set of R^n with a smooth boundary $\partial\Omega$, $g\in L^2(\Omega)$, $f(u)\in C^1(R;R),\alpha>0$.

Let $G(s)=\int_0^s f(r)dr$. We make the following assumptions on the functions $G(s)$ and $f(s)$:

(i)

$$\lim_{|s|\to\infty}\inf\frac{G(s)}{s^2}\geqslant 0. \tag{2}$$

(ii) There exist two positive constants $c_1>0$, $c_2>0$ such that

$$\lim_{|s|\to\infty}\inf\frac{sf(s)-c_1G(s)}{s^2}\geqslant 0, \tag{3}$$

and

$$|f'(s)|\leqslant c_2(1+|s|^\gamma) \ \text{ with } \begin{cases} 0\leqslant\gamma<\infty & \text{when } n=1,2\\ 0\leqslant\gamma<2 & \text{when } n=3, \ \forall s\in R\\ \gamma=0 & \text{when } n\geqslant 4.\end{cases} \tag{4}$$

(iii) There exists $\delta_1>0$ and for every $M>0$ there exists $c'=c'(M)$ such that

---

[a)]Electronic mail: nic2601@pop.scuu.edu.cn

$$|f'(u_1)-f'(u_2)|\leqslant c'\|u_1-u_2\|^{\delta_1},\forall u_1,u_2\in H_0^1(\Omega),\|u_1\|\leqslant M,\|u_2\|\leqslant M, \tag{5}$$

where $|\cdot|$ and $\|\cdot\|$ denote the norms of $L^2(\Omega)$, $H_0^1(\Omega)$, respectively.

Two examples of Eq. (1) are the sine-Gordon equation [i.e., $f(u)=\sin u$] and the equation of relativistic quantum mechanics ($f(u)=|u|^{\gamma}u$).

For the system (1) with conditions (2)–(5), Temam[1] showed that the continuous semigroup of mapping $S(t):\{u_0,u_1\}\mapsto\{u,u_t\}$, for $t\geqslant 0$ from $E=H_0^1(\Omega)\times L^2(\Omega)$ into itself, defined by system (1) possesses a global attractor in E and gave an upper bound of the Hausdorff dimension of attractor, but this upper bound is directly proportional to the coefficient α of damping for $\alpha\geqslant\sqrt{2\lambda_1}$ and tends to infinity as $\alpha\to+\infty$, which is obviously not precise in the physical sense, where $\lambda_1>0$ is the first eigenvalue of operator $-\Delta$.

In this paper, we obtain a more strict upper bound of the Hausdorff dimension for the global attractor by carefully estimating and splitting the positivity of the linear operator in the corresponding evolution equation of the first order in time. The obtained Hausdorff dimension decreases as the damping α grows and is uniformly bounded for large α, which conforms to physical intuition. The idea of using such a technique originates from Wang and Zhu.[2] The main result is the following theorem.

Theorem 1: *If the function $f(u)$ satisfies conditions (2)–(5), then for any $\alpha\geqslant\alpha_0>0$, the Hausdorff dimension d_H of the global attractor for system (1) satisfies:*

$$\begin{aligned} d_H&\leqslant\min\left\{l\,\middle|\,l\in N,\frac{1}{l}\sum_{j=1}^{l}\lambda_j^{-1}\leqslant\frac{2\lambda_1\alpha^2}{k^2\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}\right\}\\ &\leqslant\min\left\{l\,\middle|\,l\in N,\frac{1}{l}\sum_{j=1}^{l}\lambda_j^{-1}\leqslant\frac{2\lambda_1\alpha_0^2}{k^2\sqrt{\alpha_0^2+4\lambda_1}(\alpha_0+\sqrt{\alpha_0^2+4\lambda_1})}\right\}, \end{aligned}\tag{6}$$

where $\{\lambda_j\}_{j\in N}:0<\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_m\leqslant\cdots$, are the eigenvalues of operator $-\Delta$ with the Dirichlet boundary condition on Ω and $k=k(\alpha_0)$ is a positive constant.

Particularly, if the condition (4) is $|f'(u)|\leqslant k_0$, then for any $\alpha>0$,

$$d_H\leqslant\min\left\{l\,\middle|\,l\in N,\frac{1}{l}\sum_{j=1}^{l}\lambda_j^{-1}\leqslant\frac{2\lambda_1\alpha^2}{k_0^2\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}\right\}. \tag{7}$$

Obviously, the upper bound of d_H in (6) is a decreasing function of α and remains small for large damping α because

$$h(\alpha)=\frac{2\lambda_1\alpha^2}{k^2\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}$$

increases as α grows and

$$\lim_{l\to+\infty}\frac{1}{l}\sum_{j=1}^{l}\lambda_j^{-1}=0,\qquad \lim_{\alpha\to+\infty}\frac{2\lambda_1\alpha^2}{k^2\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}=\frac{\lambda_1}{k^2}.$$

II. PRELIMINARIES

For convenience, we omit statements of the existence and uniqueness of the solution of (1) which define a continuous semigroup of mapping

$$S(t):\{u_0,u_1\}\mapsto\{u,u_t\}\quad\text{for } t\geqslant 0 \tag{8}$$

from $H_0^1(\Omega)\times L^2(\Omega)$ (or $(H_0^1(\Omega)\cap H^2(\Omega))\times H_0^1(\Omega))$ into itself. We also omite statements of the existence of the global attractor for the semigroup $S(t)$, $t\geqslant 0$ (see, e.g., Ref. 1 for details).

It is known that the operator $A=-\Delta:D(A)=H_0^1(\Omega)\cap H^2(\Omega)\to L^2(\Omega)$ is a self-adjoint positive linear operator and its eigenvalues $\{\lambda_i\}_{i\in N}$ satisfy

$$0<\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_m\leqslant\cdots,\quad \lambda_m\to+\infty \quad \text{as} \quad m\to+\infty.$$

Let

$$E=H_0^1(\Omega)\times L^2(\Omega),\quad E_0=D(A)\times H_0^1(\Omega),$$

$$(u,v)=\int_\Omega uv\,dx,\quad |u|=(u,u)^{1/2},\quad \forall u,v\in L^2(\Omega),$$

$$((u,v))=\int_\Omega \nabla u\cdot\nabla v\,dx,\quad \|u\|=((u,u))^{1/2},\quad \forall u,v\in H_0^1(\Omega),$$

$$(y_1,y_2)_E=((u_1,u_2))+(v_1,v_2),\quad \forall y_i=(u_i,v_i)^T\in E,\quad i=1,2,$$

$$|y|_E=(y,y)_E^{1/2},\quad \forall y=(u,v)^T\in E$$

and

$$(y_1,y_2)_{E_0}=(Au_1,Au_2)+(v_1,v_2),\quad \forall y_i=(u_i,v_i)^T\in E_0,\quad i=1,2,$$

$$|y|_{E_0}=(y,y)_{E_0}^{1/2},\quad \forall y=(u,v)^T\in E_0$$

denote the usual inner products and norms in $L^2(\Omega)$, $H_0^1(\Omega)$, E, and E_0, respectively.

It is convenient to reduce (1) to an evolution equation of the first order in time. Let $\varphi=(u,v)^T$, $v=\dot{u}+\epsilon u$, where ϵ is chosen as

$$\epsilon=\frac{\lambda_1\alpha}{\alpha^2+4\lambda_1}, \tag{9}$$

then (1) can be written as

$$\dot{\varphi}+\Lambda\varphi=F(\varphi),\quad \varphi(0)=(u_0,u_1+\epsilon u_0), \tag{10}$$

where

$$F(\varphi)=\begin{pmatrix}0\\ -f(u)+g\end{pmatrix},\quad \Lambda=\begin{pmatrix}\epsilon I & -I\\ A-\epsilon(\alpha-\epsilon)I & (\alpha-\epsilon)I\end{pmatrix}. \tag{11}$$

It is easy to see that the semigroup

$$S_\epsilon(t):(u_0,u_1+\epsilon u_0)^T\to(u(t),u_t(t)+\epsilon u(t))^T,\quad E\to E(\text{or } E_0\to E_0) \tag{12}$$

defined by (10) has the following relation with $S(t)$:

$$S_\epsilon(t)=R_\epsilon S(t)R_{-\epsilon}, \tag{13}$$

where R_ϵ is an isomorphism of E (or E_0):

$$R_\epsilon:\{u,v\}\to\{u,v+\epsilon u\}.$$

Since the semigroup $\{S(t),t\geqslant 0\}$ defined by (1)–(8) possesses a global attractor $\beta_0\subset E_0\subset E$,[1] by (13), $\{S_\epsilon(t),t\geqslant 0\}$ also possesses a global attractor $\beta=R_\epsilon\beta_0$. Moreover, β and β_0 have the same dimension. So we need consider the equivalent system (10) only. First, we give a property concerning the positivity of the linear operator Λ, which plays the center role in this article.

Lemma 1: For any $\varphi=(u,v)^T\in E$,

$$(\Lambda\varphi,\varphi)_E\geqslant\sigma|\varphi|_E^2+\frac{\alpha}{2}|v|^2, \tag{14}$$

where

$$\sigma=\frac{\lambda_1\alpha}{\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}. \tag{15}$$

Proof: Since

$$\begin{aligned}(\Lambda\varphi,\varphi)_E-\sigma|\varphi|_E^2-\frac{\alpha}{2}|v|^2&=(\epsilon-\sigma)\|u\|^2+\left(\frac{\alpha}{2}-\epsilon-\sigma\right)|v|^2-\epsilon(\alpha-\epsilon)(u,v)\\&\geqslant(\epsilon-\sigma)\|u\|^2+\left(\frac{\alpha}{2}-\epsilon-\sigma\right)|v|^2-\frac{\epsilon\alpha}{\sqrt{\lambda_1}}\|u\|\cdot|v|\end{aligned}$$

and simple computation by (9) and (15) shows

$$4(\epsilon-\sigma)\left(\frac{\alpha}{2}-\epsilon-\sigma\right)=\frac{\epsilon^2\alpha^2}{\lambda_1}.$$

Thus, the proof is completed.

In this section, we suppose $\alpha\geqslant\alpha_0>0$, and we will show that the bounds of the global attractor of system (10) in the spaces E and E_0 depend on the constant α_0 only.

We write $\overline{G}(u)=\int_\Omega G(u)dx$. Taking the inner product $(\cdot,\cdot)_E$ of (10) with $\varphi=(u,v)^T$ in which $v=u_t+\epsilon u$, we find

$$\frac{1}{2}\frac{d}{dt}[|\varphi|_E^2+2\overline{G}(u)]+(\Lambda\varphi,\varphi)_E+\epsilon(f(u),u)=(g,v). \tag{16}$$

By (2), (3) and the Poincaré inequality, there exist two positive constants $k_1,k_2\geqslant 0$ such that

$$\overline{G}(u)+k_1\geqslant-\tfrac{1}{4}\|u\|^2,\quad \forall u\in H_0^1(\Omega), \tag{17}$$

$$(u,f(u))\geqslant c_1\overline{G}(u)-\tfrac{1}{4}\|u\|^2-k_2,\quad \forall u\in H_0^1(\Omega). \tag{18}$$

It is easy to see from (9) and (15) that

$$\sigma=\frac{\sqrt{\alpha^2+4\lambda_1}}{\alpha+\sqrt{\alpha^2+4\lambda_1}}\epsilon,$$

i.e.,

$$\tfrac{1}{2}\epsilon<\sigma<\epsilon. \tag{19}$$

Let

$$y = |\varphi|_E^2 + 2\overline{G}(u) + 2k_1 \geqslant \text{by } (17) \geqslant \tfrac{1}{2}|\varphi|_E^2 \geqslant 0. \tag{20}$$

By (14), (16), (17) and (18),

$$\frac{d}{dt}y + \beta_1 y \leqslant \frac{1}{\alpha}|g|^2 + 2\epsilon(c_1k_1 + k_2), \tag{21}$$

where

$$\beta_1 = \tfrac{1}{2}\epsilon\theta, \theta = \min(1,2c_1) = \begin{cases} 1 & \text{when } c_1 \geqslant \tfrac{1}{2} \\ 2c_1 & \text{when } 0 < c_1 < \tfrac{1}{2}. \end{cases} \tag{22}$$

By the Gronwall inequality, (20) and (21),

$$|\varphi|_E^2 \leqslant 2y(t) \leqslant 2y(0)\exp(-\beta_1 t) + \left(\frac{4}{\alpha\epsilon\theta}|g|^2 + \frac{8(c_1k_1+k_2)}{\theta}\right)[1 - \exp(-\beta_1 t)] \tag{23}$$

and

$$\lim_{t\to+\infty} \sup|\varphi|_E^2 \leqslant \frac{4}{\alpha\epsilon\theta}|g|^2 + \frac{8(c_1k_1+k_2)}{\theta} = M(\alpha). \tag{24}$$

By (24) and (9),

$$M(\alpha) = \frac{4}{\theta}\left(\frac{\alpha^2 + 4\lambda_1}{\lambda_1\alpha^2} + 2(c_1k_1+k_2)\right).$$

Since $M(\alpha)$ is a decreasing continuous function of α on $[\alpha_0 + \infty)$ and

$$\lim_{\alpha\to+\infty} M(\alpha) = \frac{4}{\theta}\left(\frac{1}{\lambda_1} + 2(c_1k_1+k_2)\right) < +\infty,$$

so, there exists a positive constant

$$M_0 = M_0(\alpha_0) = \frac{4}{\theta}\left(\frac{\alpha_0^2 + 4\lambda_1}{\lambda_1\alpha_0^2} + 2(c_1k_1+k_2)\right)$$

such that

$$M(\alpha) \leqslant M_0, \quad \forall \alpha \in [\alpha_0, +\infty). \tag{25}$$

We infer from (24) and (25) that the ball of E, $B_0 = B_E(0,2M_0)$, centered at 0 of radius $2M_0$, is an absorbing set in E for the semigroup $S_\epsilon(t)$, $t \geqslant 0$. So, the global attractor β is included in the bounded ball B_0. We also knew from Ref. 1 that if the set of all continuous and bounded functions from R_+ into the Hilbert space X is denoted by $C_b(R_+,X)$, then any solution $\varphi = (u,v)^T$ in $C_b(R_+,E)$ of system (10) belongs to $C_b(R_+,E_0)$, and the norm of φ in $C_b(R_+,E_0)$ is majorized by a bounded function of $|f| + |\varphi|_E$ independent of the parameter α. Hence, the global attractor β is included in a bounded ball in E_0 which depends on α_0 only, that is, there exists a constant $M_1 = M_1(\alpha_0) > 0$ such that

$$|\varphi|_{E_0} = (|Au|^2 + |v|^2)^{1/2} \leqslant M_1, \quad \forall \varphi = (u,v)^T \in \beta. \tag{26}$$

III. PROOF OF THEOREM 1

To estimate the Hausdorff dimension of the global attractor β for (10) in E, we consider the first variation equation of (10),

$$\Psi' = -\Lambda\Psi + F'(\varphi)\Psi, \quad \Psi(0) = (\xi,\eta)^T \in E, \tag{27}$$

where $\Psi = (U,V)^T \in E$, and $\varphi = (u,v)^T$ is a solution of (10) and

$$F'(\varphi) = \begin{pmatrix} 0 & 0 \\ -f'(u) & 0 \end{pmatrix}. \tag{28}$$

Lemma 2: The system (27) is a well-posed problem in E, the mapping $S_\epsilon(t)$ defined by (10) is Fréchet differentiable on E for any $t>0$, its differential at $\varphi = (u_0, u_1 + \epsilon u_0)^T$ is the linear operator on E, $(\xi,\eta)^T \mapsto (U(t),V(t))^T$, where $(U,V)^T$ is the solution of (27).

Proof: It is a direct consequence of (13) and lemma VI.6.1 in Ref. 1.

Lemma 3: For any orthonormal family of elements of E, $\{(\xi_j,\eta_j)^T\}_{j=1}^{l}$, we have

$$\sum_{j=1}^{l} |\xi_j|^2 \leq \sum_{j=1}^{l} \lambda_j^{-1}. \tag{29}$$

Proof: See lemma VI.6.3. in Ref. 1.

Lemma 4: Consider the system (10). Let Φ denote a set of l vectors $\{\Phi_1, \Phi_2, \ldots, \Phi_l\}$ which are orthonormal in E. If

$$\sup_{\Phi \subset E}\sup_{\varphi\in\beta}\sum_{i=1}^{l}((-\Lambda + F'(\varphi))\Phi_i, \Phi_i)_E \leq 0, \tag{30}$$

then the Hausdorff dimension of the global attractor β is less than or equal to l.

Proof: This is a direct consequence of theorem V.3.3, Eqs. (V.3.47)–(V.3.49) and identity (VI.6.24) of Ref. 1.

Lemma 5: If the function $f(u)$ satisfies the conditions (2), (3), (4), (5), then for any $\alpha \geq \alpha_0 > 0$, the Hausdorff dimension $d_H(\beta)$ of β for system (10) in E satisfies

$$d_H(\beta) \leq \min\left\{ l \,\middle|\, l\in N, \frac{1}{l}\sum_{j=1}^{l}\lambda_j^{-1} \leq \frac{2\alpha\sigma}{k^2}\right\}, \tag{31}$$

where $k = k(\alpha_0)$ is a positive constant depending α_0 only.

Proof: Let $l \in N$ be fixed. Consider l solutions $\Psi_1, \Psi_2, \ldots, \Psi_l$ of (27). At a given time τ, let $Q_l(\tau)$ denote the orthogonal projection in E onto the space spanned by $\Psi_1, \Psi_2, \ldots, \Psi_l$. Let $\Phi_j(\tau) = (\xi_j, \eta_j)^T \in E$, $j = 1,2,...,l$, be an orthonormal basis of

$$Q_l(\tau)E = \mathrm{span}\{\Psi_1(\tau), \Psi_2(\tau), \ldots, \Psi_l(\tau)\}.$$

From (14) and $|\Phi_j|_E = 1$, we have

$$-(\Lambda\Phi_j, \Phi_j)_E \leq -\sigma - \frac{\alpha}{2}|\eta_j|^2. \tag{32}$$

By (26),

$$|Au| \leq M_1, \quad \forall u \in D(A)\cap\beta = (H^2(\Omega)\cap H_0^1(\Omega))\cap\beta. \tag{33}$$

By (4), (33) and Sobolev embedding theorem, there exists a constant $k = k(\alpha_0) > 0$ (which is independent of α) such that

$$\sup_{u\in D(A)\cap\beta}|f'(u)|_{C(H_0^1(\Omega),L^2(\Omega))}\leqslant k<+\infty. \tag{34}$$

Thus by (28) and (34),

$$(F'(\varphi)\Phi_j,\Phi_j)_E=(-f'(u)\xi_j,\eta_j)\leqslant k|\xi_j|\cdot|\eta_j|\leqslant\frac{k^2}{2\alpha}|\xi_j|^2+\frac{\alpha}{2}|\eta_j|^2,\quad \forall\varphi=(u,v)\in\beta. \tag{35}$$

Hence,

$$\begin{aligned}\sup_{\varphi\in\beta}\sum_{i=1}^{l}((-\Lambda+F'(\varphi))\Phi_i,\Phi_i)_E&\leqslant \text{by (32) and (35)}\\ &\leqslant -l\sigma+\frac{k^2}{2\alpha}\sum_{j=1}^{l}|\xi_j|^2\\ &\leqslant \text{by (29)}\\ &\leqslant -\frac{lk^2}{2\alpha}\left(\frac{2\alpha\sigma}{k^2}-\frac{1}{l}\sum_{j=1}^{l}\lambda_j^{-1}\right).\end{aligned} \tag{36}$$

If

$$\frac{1}{l}\sum_{j=1}^{l}\lambda_j^{-1}\leqslant\frac{2\alpha\sigma}{k^2},$$

then by (36),

$$\sup_{\varphi\in\beta}\sum_{i=1}^{l}((-\Lambda+F'(\varphi))\Phi_i,\Phi_i)_E\leqslant 0.$$

By lemma 4, the proof is completed.

Combining with lemma 5 and (15), we complete the proof of theorem 1.

[1] R. Témam, ''Infinite-dimensional dynamical systems in mechanics and physics,'' *Applied Mathematical Sciences*, Vol. 68 (Springer, New York, 1988).
[2] G. Wang and S. Zhu, ''On dimension of the global attractor for damped sine-Gordon equation,'' J. Math. Phys. (to be published).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Analytical study of fractionally charged solitons in a one-dimensional trimerized electron–phonon system

Ryōen Shirasaki
Department of Physics, Faculty of Engineering, Yokohama National University, 79-5 Tokiwadai Hodogaya-ku, Yokohama 240-8501, Japan

Kaoru Iwano
Institute of Materials Structure Science, High Energy Accelerator Research Organization, 1-1 Oho, Tsukuba 305-0801, Japan

(Received 5 June 1998; accepted for publication 8 December 1998)

The fractionally charged solitons in a trimerized electron–phonon system are studied by both analytical and numerical methods. To make those methods possible, an effective Lagrangian is derived by diagrammatic calculations. It is found that this Lagrangian is similar to that derived by the so-called derivative expansion. Both the phase and amplitude parts of the complex order parameter are included in the Lagrangian, and here we particularly focus on the nonlinear coupling between them. When the electron–phonon coupling is very weak, the aforementioned coupling is also weak, and so the soliton is almost considered to be a pure phase soliton with a constant amplitude. While, in the intermediate electron–phonon coupling case, the two parts of the order parameter are nonlinearly coupled. As a result, the soliton changes its pattern from that of a phase soliton to a strongly amplitude-deformed one. Our both methods, i.e., the analytical and numerical ones, succeed in giving such changes as gradual ones. Moreover, the coincidence of the two results is also good at a quantitative level. © *1999 American Institute of Physics.* [S0022-2488(99)01703-X]

I. INTRODUCTION

The one-dimensional electron–phonon system is an interesting system to which many theoretical and the experimental investigations are devoted. This is one of the most ideal systems which bear nonlinear excitations such as solitons, polarons, and so on. They can be described very well, for example, by the SSH model that was proposed by Su *et al.*[1] This model contains the electron–phonon interaction in addition to electron hopping and lattice vibrations. Takayama *et al.* proposed the TLM model, which is a continuum version of the SSH model in the half-filling case and makes the analytical study of nonlinear excitations possible.[2]

A simpler model for analytical study of the one-dimensional electron–phonon system is the Fröhlich model.[3] The difference from the SSH model is in the momentum dependence of the electron–phonon coupling. While the coupling depends on both the transfer momentum and the initial momentum in the SSH model, it is simply constant in this model. The model has been analyzed by many authors. For example, Horovitz *et al.* studied the dynamical equation of both the phase and the amplitude of the complex order parameter in the commensurate case, using the self-consistent equation between electron- and phonon-degrees of freedom given by the Fröhlich model.[4] They derived an effective Lagrangian which can be considered as the sine-Gordon model, and obtained the phase soliton solution. Grabowski *et al.* solved the effective Lagrangian numerically.[5] The soliton of the sine-Gordon model was investigated in various systems.[6] The soliton appears as the spatial connection between the degenerate energy minimum points of the cosine potential. In the electron–phonon system, the local density of electron charge is represented by the spatial variation of the phase variable. The phase soliton in the trimerized system has fractionally charges $(\pm 2/3e)$.

Iwano *et al.* also investigated the Fröhlich model in the trimerized case using the Eilenberger equation which had been formulated in the study of the superconductivity in strong-coupling cases.[7] At a particular value for the coupling constant λ, the Eilenberger equation has an exact solution. They found that the coupling between the phase and the amplitude is non-negligible to cause a large distortion in the latter. Their result showed that the soliton in the commensurate and away from the half-filling system should be considered as a complex in which the two degrees of freedom are entangled rather than a simple phase soliton.

In the present paper, we study the fractionally charged soliton using the effective Lagrangian of the trimerized system. We apply another method to derive the effective Lagrangian, which has some differences from that given in Ref. 4. The solution is investigated classical mechanically using both an analytical perturbational calculation and a numerical calculation introducing a variational functional. We will show that the fractionally charged soliton has a complicated behavior which is consistent with the result given by the preceding work.

II. THE FRÖLICH MODEL

A. General formalism for a trimerized electron–phonon system

The Hamiltonian of the one-dimensional Fröhlich system of length $L=Na$ is given by

$$K=H-\mu N_e=\sum_{k,s}(\epsilon_k-\mu)C^\dagger_{k,s}C_{k,s}-\frac{\mathrm{i}}{\sqrt{N}}\sum_q\sum_{k,s}g(b_q+b^\dagger_{-q})C^\dagger_{k+q,s}C_{k,s}+\sum_q\omega_Q b^\dagger_q b_q. \quad (1)$$

The first term is the free electron part, where μ is the chemical potential, and N_e is the total number of electrons. Here we consider a tight-binding band of electrons as

$$\epsilon_k=-W\cos ka. \quad (2)$$

This band is $\frac{1}{3}$ filled with electrons. The second term represents the electron–phonon coupling. Since this is the Frölich model, the coupling factor g is now constant with no momentum dependence. The last term is the free phonon term. ω_Q is the phonon frequency for the wave number of $2k_F$, where k_F is the wave number of electron at the Fermi level.

Introducing the three component electronic field $\psi_s(x)$ and the lattice displacement operator $u(x)$ by

$$\psi_s(x)=\begin{pmatrix}\phi_s^{(1)}(x)\\ \phi_s^{(2)}(x)\\ \phi_s^{(3)}(x)\end{pmatrix}=\frac{1}{\sqrt{L}}\sum_{-k_F\leq k<k_F}\begin{pmatrix}C_{k_F+k,s}\\ C_{-k_F+k,s}\\ C_{3k_F+k,s}\end{pmatrix}e^{\mathrm{i}kx}, \quad (3)$$

and

$$u(x)=-\frac{\mathrm{i}}{\sqrt{N}}\sum_{-k_F\leq q<k_F}(b_{2k_{F+q}}+b^\dagger_{-2k_{F-q}})e^{\mathrm{i}qx}. \quad (4)$$

Equation (1) is rewritten as follows:

$$K=H_u+H_\psi-(\tfrac{1}{2}+\mu)N_e, \quad (5)$$

with

$$H_u=\frac{g^2}{2\lambda\pi v_F\omega_Q^2}\int(|\dot{u}(x)|^2+\omega_Q^2|u(x)|^2), \quad (6)$$

and

$$H_\psi=\sum_s \int \mathrm{d}x\,\psi_s^\dagger(x)H_e\psi_s(x), \tag{7}$$

where

$$H_e=\begin{pmatrix} -\mathrm{i}v_F\partial_x & 0 & 0 \\ 0 & \mathrm{i}v_F\partial_x & 0 \\ 0 & 0 & W\left(\frac{3}{2}+\frac{a^2}{2}\partial_x^2\right)\end{pmatrix}+g\begin{pmatrix} 0 & u(x) & u^*(x) \\ u^*(x) & 0 & u(x) \\ u(x) & u^*(x) & 0\end{pmatrix}. \tag{8}$$

H_u in Eq. (5) is the contribution from the lattice elastic energy given by the last term of Eq. (1). In the case of the trimerized system, the Fermi level is nearly equal to $-W/2$. We take μ as $-W/2$. The parameters are defined by

$$\lambda=\frac{ag^2}{\pi v_F\omega_Q}, \tag{9}$$

and the Fermi velocity of electron is given by

$$v_F=\frac{\sqrt{3}}{2}Wa. \tag{10}$$

In the following of this section, we concentrate our mind to obtain an effective potential of the lattice dynamics, performing renormalizations of the three component electronic field. First, introducing a two-component electronic field $\phi_s(x)$, and the amplitude $\Delta(x)$ and the phase $\theta(x)$ of the phonon field by

$$\phi_s(x)=\begin{pmatrix}\phi_s^{(1)}(x)\\ \phi_s^{(2)}(x)\end{pmatrix}, \tag{11}$$

and

$$\Delta(x)e^{\mathrm{i}\theta(x)}=gu(x), \tag{12}$$

respectively. H_ψ is rewritten as

$$\begin{aligned} H_\psi=&-\sum_s \mathrm{i}v_F\int \mathrm{d}x\,\phi_s^\dagger(x)\partial_x\sigma_3\phi_s(x)\\ &+\int \mathrm{d}x\,\phi_s^\dagger(x)\left\{\Delta(x)\begin{pmatrix}0 & e^{\mathrm{i}\theta(x)}\\ e^{-\mathrm{i}\theta(x)} & 0\end{pmatrix}-\frac{2}{3W}\Delta(x)^2\begin{pmatrix}1 & e^{-2\mathrm{i}\theta(x)}\\ e^{2\mathrm{i}\theta(x)} & 1\end{pmatrix}\right\}\phi_s(x)\\ &+\int \mathrm{d}x\left(\frac{2}{3W}\left|\frac{3W}{2}\phi_s^{(3)}(x)+\Delta(x)(\phi_s^{(1)}(x)e^{\mathrm{i}\theta(x)}+\phi_s^{(2)}(x)e^{-\mathrm{i}\theta(x)})\right|^2+\phi_s^{(3)*}\frac{Wa^2}{2}\partial_x^2\phi_s^{(3)}(x)\right). \end{aligned} \tag{13}$$

Replacing $\phi_s^{(3)}(x)$ by $\varphi_s(x)=\phi_s^{(3)}(x)-[2\Delta(x)/3W](\phi_s^{(1)}(x)e^{\mathrm{i}\theta(x)}+\phi_s^{(2)}(x)e^{-\mathrm{i}\theta(x)})$, and using that $\phi_s^{(3)}(x)$ varies slowly in the scale of the correlation length $\xi=v_F/|\Delta|$, the third and the fourth lines in the right-hand side of Eq. (13) are rewritten again as

$$\int \mathrm{d}x\,\varphi(x)\left(\frac{3W}{2}+\frac{Wa^2}{2}\partial_x^2\right)\varphi(x)+O\left(\Delta\left(\frac{\alpha}{\xi}\right)^2\left\langle\frac{\partial^2}{\partial(x/\xi)^2}\right\rangle\right). \tag{14}$$

Equation (14) means that the influence of Δ and θ on $\phi_s^{(3)}(x)$ is sufficiently small. Then, we totally neglect the contribution of $\phi_s^{(3)}(x)$ in the following.

To estimate the effective potential from the Hamiltonian in Eq. (13), we consider the partition function, which is defined by

$$Z=\int Du_r Du_i\left(\prod_s D\phi_s^\dagger D\phi_s\right)e^{-\int L\mathrm{d}\tau}, \tag{15}$$

where u_r and u_i are the real and imaginary parts of the field $u(x)$, respectively. Then, L is written as

$$L=L_\Delta^0+L_\phi, \tag{16}$$

where

$$L_\Delta^0=\frac{1}{2\pi v_F\lambda\omega_Q{}^2}\int \mathrm{d}x\left\{\left(\frac{\partial\Delta(x)}{\partial\tau}\right)^2+\Delta(x)^2\left(\frac{\partial\theta(x)}{\partial\tau}\right)^2+\omega_Q^2\Delta^2(x)\right\}, \tag{17}$$

and

$$L_\phi=\sum_s\int \mathrm{d}x\,\phi_s^\dagger(x)\frac{\partial\phi_s(x)}{\partial\tau}+\sum_s\int \mathrm{d}x\,\phi_s^\dagger(x)[-\mathrm{i}v_F\sigma_3\partial_x]\phi_s(x)$$
$$+\sum_s\int \mathrm{d}x\,\phi_s^\dagger(x)\left\{-\frac{2\Delta(x)^2}{3W}\begin{pmatrix}1&0\\0&1\end{pmatrix}+\begin{pmatrix}0&D(x)\\D^*(x)&0\end{pmatrix}\right\}\phi_s(x). \tag{18}$$

Here $\partial/(\partial\tau)$ is the derivatives with respect to the imaginary time τ. A new field $D(x)$ is introduced by

$$D(x)=\Delta(x)\mathrm{e}^{\mathrm{i}\theta(x)}-\frac{2}{3W}\Delta(x)^2\mathrm{e}^{-2\mathrm{i}\theta(x)}. \tag{19}$$

Using the phase of $D(x)$, we can introduce the chiral transformation[8]

$$\tilde{\phi}_s(x)=e^{-(\mathrm{i}/2)\chi(x)\sigma_3}\phi_s(x), \tag{20}$$

where $\chi(x)$ is the phase of the complex variable $D(x)$, namely,

$$D(x)=A(x)e^{\mathrm{i}\chi(x)}, \tag{21}$$

with

$$A(x)=|D(x)|=\sqrt{\Delta(x)^2+\left(\frac{2\Delta(x)^2}{3W}\right)^2-\frac{4}{3W}\Delta(x)^3\cos 3\theta(x)}. \tag{22}$$

Substituting Eq. (20) into Eq. (18), the Lagrangian for the scalar fields $\Delta(x)$, $\theta(x)$, and the electronic field $\tilde{\phi}_s$ is given by

$$L=L_\Delta^0+L_{\tilde{\phi}}^0+H_I, \tag{23}$$

with

$$L_{\tilde{\phi}}^0=\sum_s\int \mathrm{d}x\,\tilde{\phi}_s^\dagger(x)\frac{\partial\tilde{\phi}_s(x)}{\partial\tau}+\sum_s\int \mathrm{d}x\,\tilde{\phi}_s^\dagger(x)[-\mathrm{i}v_F\sigma_3\partial_x]\tilde{\phi}_s(x), \tag{24}$$

and

$$H_I=\sum_s \int \mathrm{d}x\,\tilde{\phi}_s^\dagger(x)\left\{\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}I+\mathrm{i}\frac{\partial\chi(x)}{\partial\tau}\frac{\sigma_3}{2}-\frac{2}{3W}\Delta(x)^2 I+A(x)\sigma_1\right\}\tilde{\phi}_s(x). \tag{25}$$

In the path integration, the Jacobian $\partial(\phi_s^\dagger,\phi_s)/\partial(\tilde{\phi}_s^\dagger,\tilde{\phi}_s)$ appears by the chiral transformation. This factor is explained physically by Fujikawa.[9] Let us consider the path integration at the zero temperature. Introducing the external electromagnetic field $(\varphi,v_F A_x)=(-\mathrm{i}A_0,A_1)$, and the Euclidian space-time $(x^0,x^1)=(\mathrm{i}t,x/v_F)$, the space and time derivatives are rewritten in the covariant forms as[10]

$$-\mathrm{i}v_F\frac{\partial}{\partial x}\to-\mathrm{i}(\partial_1+\mathrm{i}eA_1), \tag{26}$$

and

$$\mathrm{i}\frac{\partial}{\partial t}\to-(\partial_0+\mathrm{i}eA_0), \tag{27}$$

where ∂_i denotes the differentiation with respect to x^i. The Jacobian is obtained by the gauge invariant regularization of the divergent integrals.[9,11] It is given by

$$\frac{\partial(\phi_s^\dagger,\phi_s)}{\partial(\tilde{\phi}_s^\dagger,\tilde{\phi}_s)}=\exp\left(-\mathrm{i}\frac{e}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\chi(x)F_{01}(x)\mathrm{d}x^1\mathrm{d}x^0\right), \tag{28}$$

where

$$F_{01}=\frac{\partial A_1}{\partial x^0}-\frac{\partial A_0}{\partial x^1}=\mathrm{i}v_F E(x), \tag{29}$$

is the electric field.

Therefore, taking care of the spin degeneracy, we should examine the partition function at finite temperatures,

$$\begin{aligned}Z&=\int Du_r Du_i D\tilde{\phi}_s^\dagger D\tilde{\phi}_s\exp\left\{\frac{e}{\pi}\int_0^\beta\int\chi(x)E(x)\mathrm{d}x\mathrm{d}\tau-\int_0^\beta L\mathrm{d}\tau\right\}\\&=Z_0\left\langle T_\tau\exp\left(\frac{e}{\pi}\int_0^\beta\int\chi(x)E(x)\mathrm{d}x\mathrm{d}\tau-\int_0^\beta H_I\mathrm{d}\tau\right)\right\rangle,\end{aligned} \tag{30}$$

where $\langle\cdots\rangle$ means a thermal average with respect to the unperturbed Lagrangian. Moreover, H_I is the perturbation term given by Eq. (25), and Z_0 is the partition function of the unperturbed Lagrangian in which the influences of the external electromagnetic field and the electron–phonon coupling are excluded.

Equation (30) suggests that $-(e/\pi)(\partial\chi/\partial x)$ corresponds to the local charge density, and the dynamics of the order parameter involves variation of the charge distribution. This result was first derived by the study of the charge- and spin-density-waves in a one-dimensional electron system.[12]

B. The effective potential

Equation (25) has some analogy with the Takayama–Lin–Liu–Maki (TLM) model. The σ_1 coupling term in Eq. (25) appears in the same manner with the electron-lattice coupling term in the TLM model, whereas the σ_3 and I coupling terms are peculiar to this model. The effective

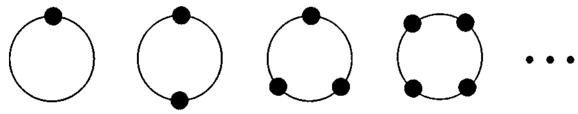

FIG. 1. The loop diagrams which contribute to the effective potential for the lattice distortion. The continuous line represents an electron Green function and the solid circle corresponds to the vertex.

potential for the TLM model was derived by the derivative expansion of the electron Green function by Horovits *et al.*[4] While, it is by our method derived by estimating the partition function of the TLM model, which comes back to the calculation of the connected loop diagrams with respect to the electron–phonon coupling.[13] By a renormalization, the effective Lagrangian at the zero temperature limit is given by

$$L_R=\frac{1}{2\pi v_F\lambda\omega_Q^2}\int dx(\dot{\Delta}(x)^2+\omega_Q^2\Delta(x)^2)+\frac{1}{2\pi v_F}\int dx\Delta(x)^2\ \ln\left(\left|\frac{\Delta(x)}{2\Lambda e^{1/2}}\right|^2\right)$$
$$+\frac{1}{\pi v_F}\int dx\Delta(x)\left(\frac{\sqrt{1+\gamma_1{}^2}}{\gamma_1}\sinh^{-1}\gamma_1-1\right)\Delta(x)$$
$$+\sum_{n=3}^{\infty}\frac{1}{n}\int dx\int dx_1\cdots\int dx_{n-1}\left(\prod_{j=1}^{n-1}\delta(x-x_j)\right)$$
$$\times D_n(x,x_1,\ldots,x_{n-1})\,\eta(x)\left(\prod_{i=1}^{n-1}\eta(x_i)\right), \tag{31}$$

where

$$\eta(x)=\Delta(x)-\Delta_0, \tag{32}$$

and

$$\gamma_1=-i\frac{v_F}{2\Delta_0}\frac{\partial}{\partial x}. \tag{33}$$

Here $D_n(x,x_1,\ldots,x_{n-1})$ is a function of derivatives with respect to $x,\ldots,x_{n-1}$, and Λ is the cut off energy of the electron band. Δ_0 is a half-width of the electron band gap at the Fermi level.

Comparing Eq. (25) and the TLM model, we find that the term $\tilde{\phi}_s^\dagger(x)A(x)\sigma_1\tilde{\phi}_s(x)$ has a similar contribution to the coupling term between the phonon field $\Delta(x)$ and the electronic one in the TLM model. Now, let us estimate the effective potential which is given by the renormalization of the electronic part Eqs. (24) and (25). The terms in which the differentials of scalar fields $\Delta(x)$ and $\theta(x)$ do not appear are given by replacing $\Delta(x)$ with $A(x)$, as

$$-\frac{1}{2\pi v_F}\int dxA(x)^2\left(2\ln\left|\frac{A(x)}{2\overline{W}}\right|-1\right), \tag{34}$$

where $\overline{W}$ is the energy cut off of the lowest electron band. The remainder in which the differentials directly appear, for example, the contribution which come from the second-order loop diagram (Figs. 1 and 2) is given by

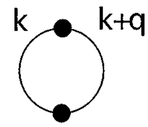

FIG. 2. The second-order loop diagram. This diagram has a momentum transfer q between the electron and the lattice.

$$-\frac{1}{\pi v_F}\int \mathrm{d}x A(x)\left(\frac{\sqrt{1+\gamma^2}}{\gamma}\sinh^{-1}\gamma-1\right)A(x)$$
$$-\frac{1}{\pi v_F}\int \mathrm{d}x\left(\frac{2\Delta(x)^2}{3W}-\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}\right)\left(\frac{\sinh^{-1}\gamma}{\gamma\sqrt{1+\gamma^2}}\right)\left(\frac{2\Delta(x)^2}{3W}-\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}\right), \tag{35}$$

as shown in Appendix A. In the calculation, we drop the time-dependency of the momentum transfer, which corresponds to an adiabatic approximation. We omit small terms of the order of Δ^2/W^2 and contributions which come from higher-order derivatives with respect to the space and the imaginary time than the second-order ones. Totally, the renormalized form of Lagrangian at the zero temperature limit is

$$L_{\mathrm{eff}}=L_\Delta^0-U, \tag{36}$$

where

$$U=-\frac{v_F}{4\pi}\int \mathrm{d}x\left(\frac{\partial\chi(x)}{\partial x}\right)^2+\frac{1}{\pi}\int \mathrm{d}x\,\frac{2\Delta(x)^2}{3W}\frac{\partial\chi(x)}{\partial x}$$
$$-\frac{1}{2\pi v_F}\int \mathrm{d}x A(x)^2\left(2\ln\left|\frac{A(x)}{2\overline{W}}\right|-1\right)-\frac{1}{\pi v_F}\int \mathrm{d}x A(x)\left(\frac{\gamma^2}{3}\right)A(x). \tag{37}$$

The operator γ is given by Eq. (A19). The partition function is rewritten as

$$Z=\int Du_r Du_i\,\exp\left\{\frac{e}{\pi}\int_0^\beta\int \chi(x)E(x)\mathrm{d}x\mathrm{d}\tau-\int_0^\beta L_{\mathrm{eff}}\,\mathrm{d}\tau\right\}. \tag{38}$$

In Eq. (38), we take $T\to 0$, i.e., $\beta=1/T\to\infty$.

III. CHARGE-DENSITY-WAVE STATE

First, let us consider the charge-density-wave ground state. This is the case where

$$\Delta(x)=\Delta_1=\mathrm{const}, \tag{39}$$

and

$$\theta(x)=\theta_1=\mathrm{const}. \tag{40}$$

Substituting Eqs. (39) and (40) into Eq. (37), we can see that the energy minimum point (Δ_1,θ_1) satisfies the equation

$$0=\frac{1}{\lambda}+\ln\left|\left(\frac{A_0}{W}\right)^2\right|\left[1+\frac{8\Delta_1^2}{9W^2}-\frac{2\Delta_1}{W}\cos 3\theta_1\right], \tag{41}$$

and

$$0=\sin 3\theta_1, \tag{42}$$

where

$$A_0^2=\Delta_1^2+\left(\frac{2\Delta_1^2}{3W}\right)^2-\frac{4\Delta_1^3}{3W}\cos 3\theta_1. \tag{43}$$

Then, using the solution t of the equation

$$\frac{1}{\lambda}+\left(1+2t+\frac{8}{9}t^2\right)\ln\left|t^2\left(1+\frac{2}{3}t\right)^2\right|=0, \tag{44}$$

θ_1 and Δ_1 at the minimum point are given by

$$\theta_1=\frac{n\pi}{3},\quad (n=\pm 1,\pm 3,...,\pm(2n+1),\ldots),$$

and

$$\Delta_1=Wt. \tag{45}$$

In weak coupling cases, namely, at small λ, t is much smaller than 1. For further analysis, we assume the following two conditions for the model:

(1) $\Delta(x)$ is much smaller than W.
(2) The differentiation of $\Delta(x)$ and $\theta(x)$ with respect to the space variable x are quantities of the order of $O(\Delta_1/\xi_0)$ and $O(1/\xi_0)$, respectively.

ξ_0 is a correlation length which is written as

$$\xi_0=\frac{v_F}{\Delta_1}. \tag{46}$$

Then, we can make approximations, that are,

$$A(x)^2\simeq\Delta(x)^2-\frac{4\Delta(x)^3}{3W}\cos 3\theta(x), \tag{47}$$

$$\frac{\sqrt{1+\gamma^2}}{\gamma}\sinh^{-1}\gamma-1\simeq\frac{1}{3}\gamma^2=-\frac{1}{12}\left(\frac{v_F}{\Delta_1}\right)^2\frac{\partial^2}{\partial x^2}, \tag{48}$$

$$\ln\left|\frac{A(x)}{W}\right|^2\simeq\ln\left|\frac{\Delta(x)^2}{W^2}\right|-\frac{4\Delta(x)}{3W}\cos 3\theta(x), \tag{49}$$

and

$$\frac{\partial\chi(x)}{\partial x}=\frac{\theta(x)'\left(1+\frac{2\Delta(x)}{3W}\cos 3\theta(x)-\frac{8\Delta(x)^2}{9W^2}\right)+\frac{2\Delta(x)'}{3W}\sin 3\theta(x)}{\left(1-\frac{2\Delta(x)}{3W}\right)^2-\frac{4\Delta(x)}{3W}(\cos 3\theta(x)-1)}$$

$$\simeq\theta(x)'\left(1+\frac{2\Delta(x)}{W}\cos 3\theta(x)\right)+\frac{2\Delta(x)'}{3W}\sin\theta(x). \tag{50}$$

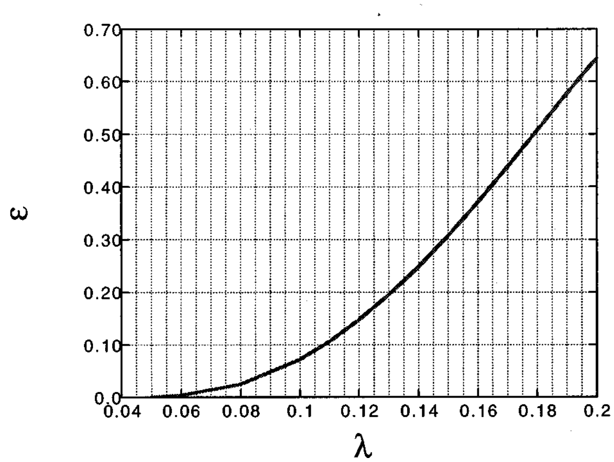

FIG. 3. The λ dependence of ϵ. It varies monotonically from 0 to 0.6 when λ changes from 0 to 0.2.

Using Eq. (45), Eq. (44) is rewritten as,

$$\ln|t| = -\frac{1}{2\lambda}\left(1+2t+\frac{8}{9}t^2\right)^{-1} - \ln\left(1+\frac{2}{3}t\right) \simeq \epsilon - \frac{1}{2\lambda} - \frac{2}{3}t + O(t^2). \tag{51}$$

Here, we introduce ϵ by

$$\epsilon = \frac{t}{\lambda} = \frac{\Delta_1}{W\lambda}. \tag{52}$$

In Fig. 3, we plot ϵ as the function of λ. In the weak coupling limit, t is nearly equal to $\exp(-1/2\lambda)$, and the parameter ϵ approaches to zero as

$$\epsilon \simeq \frac{e^{-(1/2\lambda)}}{\lambda} \to 0. \tag{53}$$

As is seen in Fig. 3, ϵ varies from 0 to 0.6 as λ changes from 0 to 0.19.

Using Eq. (37) and Eqs. (47)–(52), the total Lagrangian is given by

$$L_{\text{eff}} = L^0_{\text{eff}} + tL^1_{\text{eff}} + O(t^2), \tag{54}$$

where

$$\begin{aligned} L^0_{\text{eff}} = {} & \frac{1}{2\pi v_F \lambda \omega_Q^2} \int dx\, \dot{\Delta}(x)^2 + \frac{1}{2\pi v_F \lambda \omega_Q^2} \int dx \left\{ \Delta(x)^2 \dot{\theta}(x)^2 + \frac{\lambda \omega_Q^2 v_F^2}{2} \left(\frac{\partial \theta(x)}{\partial x} \right)^2 \right\} \\ & + \frac{1}{12\pi v_F} \left(\frac{v_F}{\Delta_1} \right)^2 \int dx \left(\frac{\partial \Delta(x)}{\partial x} \right)^2 + \frac{1}{2\pi v_F} \int dx \left[\Delta(x)^2 \left(\ln\left| \frac{\Delta(x)^2}{\Delta_1^2} \right| - 1 \right) \right] \\ & + \frac{\epsilon}{\pi v_F} \int dx\, \Delta(x)^2 \left(1 + \frac{2\Delta(x)}{3\Delta_1} \cos(3\theta(x)) \right), \end{aligned} \tag{55}$$

and

$$
\begin{aligned}
L_{\text{eff}}^{1}=&\frac{1}{2\pi v_F}\int \mathrm{d}x\left(\frac{\Delta(x)}{\Delta_1}\dot{\theta}(x)^2\cos(3\theta(x))+\frac{1}{3\Delta_1}\sin(3\theta(x))\dot{\Delta}(x)\dot{\theta}(x)\right)\\
&+\frac{1}{3\pi v_F}\left(\frac{v_F}{\Delta_1}\right)^2\int \mathrm{d}x\left\{\frac{\Delta(x)^2}{\Delta_1}\sin(3\theta(x))\left(\frac{\partial\Delta(x)}{\partial x}\right)\left(\frac{\partial\theta(x)}{\partial x}\right)-\frac{2\Delta(x)\cos(3\theta(x))}{3\Delta_1}\left(\frac{\partial\Delta(x)}{\partial x}\right)^2\right\}\\
&+\frac{v_F}{\pi}\int \mathrm{d}x\left\{\frac{\Delta(x)}{\Delta_1}\cos(3\theta(x))\left(\frac{\partial\theta(x)}{\partial x}\right)^2+\frac{\sin(3\theta(x))}{3\Delta_1}\left(\frac{\partial\theta(x)}{\partial x}\right)\left(\frac{\partial\Delta(x)}{\partial x}\right)\right\}\\
&-\frac{1}{\pi}\int \mathrm{d}x\,\frac{2\Delta(x)^2}{3\Delta_1}\left(\frac{\partial\theta(x)}{\partial x}\right)+\frac{2}{3\pi v_F}\int \mathrm{d}x\,\Delta(x)^2\\
&-\frac{2}{3\pi v_F}\int \mathrm{d}x\,\frac{\Delta(x)^3}{\Delta_1}\cos(3\theta(x))\ln\left|\frac{\Delta(x)^2}{\Delta_1^2}\right|.
\end{aligned}
\tag{56}
$$

Here it is expanded up to the first-order with respect to $t=\Delta_1/W$. The derivatives with respect to the imaginary time $\partial/\partial\tau$ are denoted by dots. The part of the second and fifth lines in Eq. (55) is the same as the sine-Gordon model. While, the first, the third, and the fourth lines in Eq. (55) give a part which is related to the dynamics of the amplitude. The whole part of L_{eff}^1 and the last line in Eq. (55) can be considered as the interaction terms between the phase $\theta(x)$ and the amplitude $\Delta(x)$. We can therefore say that Eq. (54) is an extended version of the sine-Gordon model.

Since t in Eq. (45) is much smaller than the unity, the variable $\chi(x)$, which is given by Eqs. (19) and (21), is nearly equal to $\theta(x)$. At the energy minimum point, the variable $\chi(x)$ gets the same value as $\theta(x)$.

The effective potential in Eq. (55) is also derived by Horovitz *et al.* In that derivation, they considered a self-consistent equation of the order parameter including the lattice dynamics. Our model has the same interaction term as theirs except for the difference in the multiplying factor. The other discrepancy which is found remarkably is that in the form of the derivative terms. The model of Horovitz *et al.* contained terms with the derivatives $(\partial\Delta)^2/\Delta(x)^2$, which is given by the expansion calculation of the electron Green function with respect to $(\partial\Delta(x))/\Delta(x)$. On the other hand, we have used perturbation calculation with respect to $\eta(x)/\Delta_1$, $\partial\eta(x)/\Delta_1$, $\partial^2\eta(x)/\Delta_1,\ldots,$ and so on. Then, the derivatives appear in the form of $(\partial\Delta(x))/\Delta_1$. This difference is important when $\Delta(x)$ deviates substantially from Δ_1.

IV. SOLITON IN THE TRIMERIZED SYSTEM

A. Soliton solution

The Lagrangian in Eq. (55) bears soliton solutions connecting the degenerated minimum points of the cosine potential, namely, $\theta(x)=(2n+1)\pi/3$ for integer n. This corresponds to the fractionally charged soliton. The soliton charge is determined by

$$\int_{-\infty}^{\infty}\mathrm{d}x\left(-\frac{e}{\pi}\frac{\partial\chi(x)}{\partial x}\right)=-\frac{e}{\pi}(\chi(\infty)-\chi(-\infty)). \tag{57}$$

When $\theta(x)$ varies from $-\pi/3$ to $\pi/3$, or vice versa the variable $\chi(x)$ varies by $2\pi/3$ or $-2\pi/3$. Then, these soliton carry $\mp 2/3e$ charge.

Soliton solutions in electron–phonon systems were investigated using the phase Hamiltonian.[14] The solution discussed by that Hamiltonian contained only the dynamics of the phase. However, in the numerical calculation, it was shown that the fractionally charged soliton had both the phase and amplitude degrees of freedom. For example, the soliton solution in the SSH model was numerically discussed by Ono *et al.*[15] The fractionally charged soliton in the trimerized case was analyzed, and it was found that the distortion of the amplitude was also accompanying. The soliton in the Frölich model was also investigated both analytically and numerically.[7,16] At a certain finite electron–phonon coupling, the soliton in the Frölich model showed a trajectory almost like a straight line in the complex plane of the variable $u(x)$.

In the present section, the analytical form of the fractionally charged soliton is studied for rather general values of the electron–phonon coupling. We propose the particular method to make the reductive perturbational calculation possible.

Replacing the imaginary time with the real time, the effective Lagrangian in Eq. (55) gives straightforwardly the dynamical equation

$$\frac{1}{\lambda\omega_Q^2}(\ddot{\Delta}(x)-\Delta(x)\dot{\theta}(x)^2)=\frac{\xi_0{}^2}{6}\Delta''(x)-2\Delta(x)\ln\left|\frac{\Delta(x)}{\Delta_1}\right|-2\Delta(x)\left(\frac{\Delta(x)}{\Delta_1}\cos 3\theta(x)+1\right)\epsilon, \tag{58}$$

and

$$\frac{1}{\lambda\omega_Q^2}\left(2\frac{\dot{\Delta}(x)\Delta(x)\dot{\theta}(x)}{\Delta_1^2}+\frac{\Delta(x)^2}{\Delta_1^2}\ddot{\theta}(x)\right)=\frac{\xi_0{}^2}{2}\theta''(x)+2\epsilon\left(\frac{\Delta(x)}{\Delta_1}\right)^3\sin 3\theta(x). \tag{59}$$

Equation (59) is the same as that of the sine-Gordon model, which has nonlinear excitations such as a soliton. The last term in the r.h.s. of Eq. (58) is considered as the coupling term between $\theta(x)$ and $\Delta(x)$, which causes the local distortion in $\Delta(x)$ influenced by the kink distortion in the phase mode.

We will study the effect of the interaction term for various strengths of the electron–phonon coupling. Consider the stationally case, $\dot{\Delta}(x)=0$, $\dot{\theta}(x)=0$. Introducing the space variable X instead of x as

$$X=\frac{2\sqrt{3\epsilon}}{\xi_0}x, \tag{60}$$

the equations are rewritten as follows:

$$0=-\Delta(x)\ln\left|\frac{\Delta(x)}{\Delta_1}\right|+\epsilon\left\{\frac{\partial^2\Delta(x)}{\partial X^2}-\Delta(x)\left(\frac{\Delta(x)}{\Delta_1}\cos 3\theta(x)+1\right)\right\}, \tag{61}$$

and

$$0=\frac{\partial^2(3\theta(x))}{\partial X^2}+\frac{\Delta(x)^3}{\Delta_1^3}\sin 3\theta(x). \tag{62}$$

Since $\epsilon\to 0$ as λ becomes 0, ϵ can be considered as the expansion parameter in the weak coupling case. The unperturbed equation for $\Delta(x)$ gives

$$\Delta(x)=\Delta_1=\text{const.} \tag{63}$$

Thus, the unperturbed equation for $\theta(x)$ gives the sine-Gordon equation,

$$0=\frac{\partial^2(3\theta(x))}{\partial X^2}+\sin 3\theta(x). \tag{64}$$

Equation (64) has the soliton solution,

$$3\theta_0(x)=\pi+4\tan^{-1}(e^{\pm X}). \tag{65}$$

This represents a soliton connecting the energy minimum points $\theta(x)=\pi/3$ and $\theta(x)=\pi$.

B. Perturbational expansion in the weak coupling limit

Starting from Eq. (65), we make a perturbational calculation. From now on, we take the plus sign in Eq. (65). Since $\Delta(x)>0$, we introduce $\delta(x)$ by

$$\Delta(x)=\Delta_1 e^{\delta(x)}, \tag{66}$$

the function $\delta(x)$ is zero in the weak-coupling limit, namely, $\epsilon=0$.

We introduce the deviation of $\theta(x)$ from $\theta_0(x)$ by

$$3\theta(x)=3\theta_0(x)+2\tan^{-1} v(x), \tag{67}$$

where $v(x)$ is a function which is a zero at $\epsilon=0$. Using Eq. (67), the cosine and the sine of $\theta(x)$ are given by

$$\cos 3\theta(x)=\cos 3\theta_0(x)\frac{1-v^2(x)}{1+v^2(x)}-\sin 3\theta_0(x)\frac{2v(x)}{1+v^2(x)}, \tag{68}$$

and

$$\sin 3\theta(x)=\sin 3\theta_0(x)\frac{1-v^2(x)}{1+v^2(x)}+\cos 3\theta_0(x)\frac{2v(x)}{1+v^2(x)}. \tag{69}$$

Substituting Eqs. (66)–(69) into Eqs. (61) and (62), we obtain

$$\begin{aligned}0&=\epsilon(\partial_X^2\delta(x)+(\partial_X\delta(x))^2)-\delta(x)\\&\quad-\epsilon\left(e^{\delta(x)}\cos(3\theta_0(x))\frac{1-v^2(x)}{1+v^2(x)}-e^{\delta(x)}\sin(3\theta_0(x))\frac{2v(x)}{1+v^2(x)}+1\right),\end{aligned} \tag{70}$$

and

$$\begin{aligned}0&=\partial_X^2(3\theta_0(x))+2\frac{\partial_X^2 v(x)}{1+v(x)^2}-\frac{4v(x)}{(1+v(x)^2)^2}(\partial_X v(x))^2+\cos 3\theta_0(x)\frac{2v(x)}{1+v(x)^2}e^{3\delta(x)}\\&\quad+\sin 3\theta_0(x)\frac{1-v(x)^2}{1+v(x)^2}e^{3\delta(x)},\end{aligned} \tag{71}$$

where ∂_X denotes the differential with respect to X. In the weak coupling limit, Eq. (71) becomes Eq. (64).

In the weak-coupling cases, with $\lambda<0.13$, we have

$$\epsilon<0.2. \tag{72}$$

Then, the ordinary perturbational method is applicable. We make expansions of the functions $\delta(x)$ and $v(x)$ as

$$\delta(x)=\epsilon\delta_1(x)+\epsilon^2\delta_2(x)+\cdots, \tag{73}$$

and

$$v(x)=\epsilon v_1(x)+\epsilon^2 v_2(x)+\cdots. \tag{74}$$

Comparing the first-order terms with respect to ϵ, we obtain the first-order perturbational equations,

$$0=-\delta_1(x)-(1+\cos 3\theta_0(x)), \tag{75}$$

and

$$0=2(\partial_X^2 v_1(x)+v_1(x)\cos 3\theta_0(x))+3\delta_1(x)\sin 3\theta_0(x). \tag{76}$$

The second-order terms with respect to ϵ give the second-order equations,

$$0=\partial_X^2\delta_1(x)-\delta_2(x)-(\delta_1(x)\cos 3\theta_0(x)-2v_1(x)\sin 3\theta_0(x)), \tag{77}$$

and

$$0=\sin 3\theta_0(x)(3\delta_2(x)+\tfrac{9}{2}\delta_1(x)^2-2v_1(x)^2)+2\partial_X^2 v_2(x)+2(v_2(x)+3\delta_1(x)v_1(x))\cos 3\theta_0(x). \tag{78}$$

Substituting the unperturbed solution in Eq. (65) into Eq. (75) and, using the relations

$$\cos 3\theta_0(x)=\frac{2}{\cosh^2 X}-1, \tag{79}$$

and

$$\sin 3\theta_0(x)=2\frac{\tanh X}{\cosh X}, \tag{80}$$

we obtain

$$\delta_1(x)=-\frac{2}{\cosh^2 X}. \tag{81}$$

Using Eq. (81), we can confirm from the direct substitution that the solution of Eq. (76) is given by (Appendix B)

$$v_1(x)=-\frac{3}{2}\frac{\tanh X}{\cosh X}. \tag{82}$$

Next, substituting Eqs. (79), (80), (81), and (82) into Eq. (77), $\delta_2(x)$ is given by

$$\delta_2(x)=-\frac{16}{\cosh^2 X}+\frac{22}{\cosh^4 X}. \tag{83}$$

Using Eqs. (79)–(82), we confirm from the direct substitution that the solution of Eq. (78) is given by (Appendix B)

$$v_2(x)=\frac{\tanh X}{\cosh X}\left(\frac{71}{12}\frac{1}{\cosh^2 X}-\frac{85}{24}\right). \tag{84}$$

Totally, $v(x)$ is explicitly given up to the second order with respect to ϵ by

$$\delta(x)=-\frac{2\epsilon+16\epsilon^2}{\cosh^2 X}+\frac{22\epsilon^2}{\cosh^4 X}+O(\epsilon^3), \tag{85}$$

and

$$v(x)=\frac{\tanh X}{\cosh X}\left(\left(-\frac{3}{2}\epsilon-\frac{85}{24}\epsilon^2\right)+\frac{71}{12}\epsilon^2\frac{1}{\cosh^2 X}\right)+O(\epsilon^3). \tag{86}$$

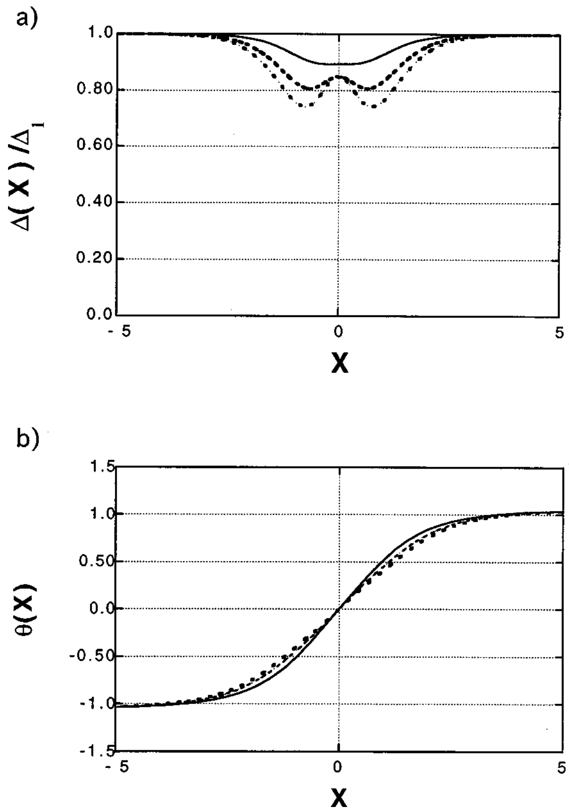

FIG. 4. The spatial change of the order parameter in the soliton configuration obtained by the reductive perturbational method. (a) is $\Delta(x)$ and (b) is $\theta(x)$. They are obtained at $\lambda=0.1$ (the solid line), $\lambda=0.12$ (the dotted line), and $\lambda=0.13$ (the rough dotted line).

In Fig. 4, we show the behaviors of $\Delta(x)$ and $\theta(x)$ as functions of the space variable X. For $\lambda \neq 0$, the amplitude is slightly deformed around the soliton center. As λ becomes larger, the distortion in the amplitude becomes larger, whereas the kink distortion of the phase $\theta(X)$ varies little. The soliton width varies as $\xi \propto 1/(t\sqrt{\epsilon})$ at small λ.

We plot trajectories of the soliton in the complex plane of $gu(x)$ in Fig. 5. When λ is 0.1, the distortion form of $u(x)$ is unusual. Since $\lambda=0.15$ corresponds to $\epsilon=0.3$, more than second-order terms give the same contribution as the first-order terms because the exponent $3\delta \simeq 3\epsilon$ in $\exp(3\delta)$ of Eq. (71) becomes the order of unity. It belongs to a case where the ordinary perturbation is not proper for analyzing a soliton.

C. Solution for $\epsilon>0.2$

Let us consider the case where

$$0.2<\epsilon<0.6, \tag{87}$$

which corresponds to

$$0.13<\lambda<0.19. \tag{88}$$

Because the ordinary reductive perturbational method is not suitable for the condition in Eq. (87), a particular analytical method is necessary.

Considering Eqs. (67) and (82), $\theta(x)$ is rewritten as follows:

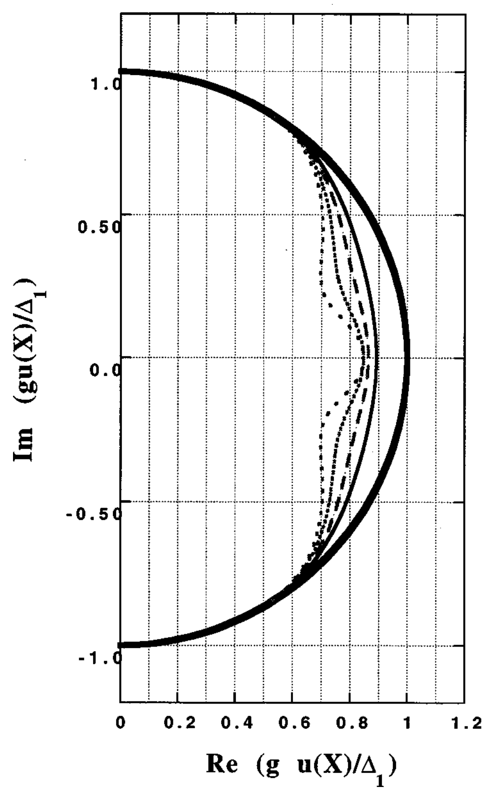

FIG. 5. The soliton configuration obtained by the reductive perturbational method. The order parameter is drawn as the trajectory in a complex plane of $\Delta(x)\exp(i\theta(x))$. The coupling constants are selected as $\lambda=0.1$ (the solid line), $\lambda=0.11$ (the broken line), $\lambda=0.12$ (the dotted line), and $\lambda=0.13$ (the rough dotted line).

$$3\theta(x)\simeq\pi+2\tan^{-1}(e^{X+c})+2\tan^{-1}(e^{X-c})+O(\epsilon^2), \tag{89}$$

where we introduce c as the parameter which satisfies the condition,

$$c^2=3\epsilon+O(\epsilon^2). \tag{90}$$

We apply a perturbational method using Eq. (89) as the unperturbed part. The deviation $t(X)$ should be introduced by

$$3\theta(x)=\frac{3\theta_0(X+c)+3\theta_0(X-c)}{2}+2\tan^{-1}t(X)=3\theta_1(X)+2\tan^{-1}t(X). \tag{91}$$

Then, the equation is rewritten as follows:

$$\begin{aligned}0=&\epsilon(\partial_X^2\delta(X)+(\partial_X\delta(X))^2)\\&-\delta(X)-\epsilon\left\{1+e^{\delta(X)}\left(\cos(3\theta_1(X))\frac{1-t^2(X)}{1+t^2(X)}-\sin(3\theta_1(X))\frac{2t(X)}{1+t^2(X)}\right)\right\}\end{aligned} \tag{92}$$

and

$$0=\partial_X^2(3\theta_1(X))+2\frac{\partial_X^2 t(X)}{1+t(X)^2}-\frac{4t(X)}{(1+t(X)^2)^2}(\partial_X t(X))^2$$
$$+\cos(3\theta_1(X))\frac{2t(X)}{1+t(X)^2}e^{3\delta(X)}+\sin 3\theta_1(X)\frac{1-t(X)^2}{1+t(X)^2}e^{3\delta(X)}. \tag{93}$$

When ϵ becomes larger than 0.3, $3\delta(X)$ in the exponential in Eq. (93) becomes larger than unity. Thus, in stead of ϵ, we introduce a new expansion parameter q by

$$q=\epsilon \mathrm{e}^{-\epsilon}. \tag{94}$$

Because $3q<1$ for $0<\lambda\leqslant 0.19$, q is suitable as an expansion parameter. The previous parameter ϵ is given by the relation,

$$\epsilon=q+q^2+O(q^3). \tag{95}$$

Consider Eq. (92), neglecting $t(X)$. The first-order terms in Eq. (92) give the relation,

$$\delta(X)\simeq-\epsilon(1+\cos 3\theta_1(X))=E. \tag{96}$$

Substituting Eq. (96) into Eq. (92), this relation is improved as

$$\delta(X)\simeq\epsilon(\partial_X^2 E)-\epsilon\{1+\mathrm{e}^E\cos 3\theta_1(X)\}. \tag{97}$$

Then, introducing a function $E(X)$ by

$$\delta(X)=-\epsilon+\epsilon e^{E(X)}(\partial_X^2 E(X)-\cos 3\theta_1(X)), \tag{98}$$

that of $\delta(X)$, i.e., Eq. (92), comes back to the equation of $E(X)$. Expanding $E(X)$ and $t(X)$ by q as

$$E(X)=qE_1(X)+q^2E_2(X)+\cdots, \tag{99}$$

and

$$t(X)=qt_1(X)+q^2t_2(X)+\cdots, \tag{100}$$

respectively. Moreover, using Eq. (98), Eq. (92) is also rewritten as follows:

$$0=q^2[-\partial_X{}^2(\cos 3\theta_1(X))+E_1(X)\cos 3\theta_1(X)-\partial_X{}^2E_1(X)+(1+\cos 3\theta_1(X))\cos 3\theta_1(X)$$
$$+2t_1(X)\cos(3\theta_1(X))\sin(3\theta_1(X))]+O(q^3). \tag{101}$$

The second-order terms with respect to q in Eq. (101) give the equation,

$$(-\partial_X{}^2+\cos 3\theta_1(X))(E_1(X)+\cos 3\theta_1(X)+1)+2t_1(X)\cos 3\theta_1(X)\sin 3\theta_1(X)=0. \tag{102}$$

Substituting Eqs. (99) and (100) into Eq. (93), and making an expansion with respect to q, the equation of $t(X)$ is rewritten as follows:

$$0=\partial_X{}^2(3\theta_1(X))+\sin 3\theta_1(X)+q(2\partial_X{}^2t_1(X)+2t_1(X)\cos 3\theta_1(X)$$
$$-3(1+\cos 3\theta_1(X))\sin 3\theta_1(X))+q^2[2\partial_X{}^2t_2(X)+2t_2(X)\cos 3\theta_1(X)$$
$$+\{-3E_1(X)\cos 3\theta_1(X)+2t_1^2(X)\partial_X{}^2(3\theta_1(X))-6t_1(X)(1+\cos 3\theta_1(X))+3\partial_X{}^2E_1(X)$$
$$+\tfrac{9}{2}(1+\cos 3\theta_1(X))^2-3(1+\cos 3\theta_1(X))\}\sin 3\theta_1(X)]+O(q^3). \tag{103}$$

On the other hand, using Eqs. (64) and (91), we can see that $\theta_1(X)$ satisfies the relations,

$$\begin{aligned}0&=\partial_X{}^2(3\theta_1(X))+\tfrac{1}{2}\sin(3\theta_0(X+c))+\tfrac{1}{2}\sin(3\theta_0(X-c))\\&=\partial_X{}^2(3\theta_1(X))+\sin 3\theta_1(X)\{1-\tanh^2 c(1+\cos(3\theta_1(X)))\}.\end{aligned}\tag{104}$$

Substituting Eq. (104) into Eq. (103), we can see that the zeroth- and the first-order terms in Eq. (103) give the equation,

$$0=q(2\partial_X{}^2 t_1(X)+2t_1(X)\cos 3\theta_1(X))+(\tanh^2 c-3q)(1+\cos 3\theta_1(X))\sin 3\theta_1(X).\tag{105}$$

To keep the consistency between Eq. (89) and Eq. (105), the first- and second-order terms with respect to $q=O(\epsilon)$ in Eq. (105) should cancel among themselves. Namely,

$$t_1(X)\equiv 0,\tag{106}$$

and

$$\tanh^2 c=3q.\tag{107}$$

Equation (107) is consistent with Eq. (90). The second-order terms with respect to q give the equation,

$$\begin{aligned}0&=2\partial_X{}^2 t_2(X)+2t_2(X)\cos(3\theta_1(X))+\{-3E_1(X)\cos(3\theta_1(X))+3\partial_X{}^2E_1(X)\\&\quad+\tfrac{9}{2}(1+\cos(3\theta_1(X)))^2-3(1+\cos(3\theta_1(X)))\}\sin(3\theta_1(X)).\end{aligned}\tag{108}$$

Using Eq. (106), Eq. (102) becomes

$$E_1(X)+\cos 3\theta_1(X)+1=0.\tag{109}$$

Then, $E_1(X)$ is given by

$$E_1(X)=-(1+\cos(3\theta_1(X)))=-\frac{2\cosh^2 c}{\cosh^2 X+\sinh^2 c}.\tag{110}$$

Using Eq. (109), Eq. (108) is rewritten as follows:

$$\begin{aligned}0&=2\partial_X^2 t_2(X)+2t_2(X)\cos(3\theta_1(X))\\&\quad+[\tfrac{33}{2}(1+\cos(3\theta_1(X)))^2-18(1+\cos(3\theta_1(X)))]\sin(3\theta_1(X)),\end{aligned}\tag{111}$$

where we use

$$\partial_X^2E_1(X)\simeq-4(1+\cos 3\theta_1(X))+3(1+\cos 3\theta_1(X))^2+O(3q).\tag{112}$$

Now we are interested in a solution that describes the soliton configuration. Therefore $t(X)$ should be odd with respect to the soliton center. Since $\sin 3\theta_1(X)$ is odd and $\cos 3\theta_1(X)$ is even, we assume that the solution takes a form of

$$t_2(X)=\sum_{n=1}^{\infty}a_n(\cos 3\theta_1(X))^n\sin 3\theta_1(X).\tag{113}$$

Substituting Eq. (113) into Eq. (111) and equating the coefficients of $(\cos(3\theta_1(X)))^n\sin(3\theta_1(X))$ on both sides, the coefficients a_n $(n=0,1,2,...)$ are determined as

$$a_0 = \tfrac{1}{12}, \quad a_1 = \tfrac{11}{12}, \tag{114}$$

and

$$a_n = 0 \quad \text{for} \quad n \geqslant 2, \tag{115}$$

where we use the relation

$$\begin{aligned}\partial_X^2[(\cos 3\theta_1(X))^n \sin 3\theta_1(X)] = (\cos 3\theta_1(X))^{n-1}[&2n(n-1) + n(2n-1)\cos 3\theta_1(X) \\ &-2(n+1)^2 \cos^2 3\theta_1(X) \\ &-(n+1)(2n+3)\cos^3 3\theta_1(X)]\sin 3\theta_1(X) + O(3q).\end{aligned} \tag{116}$$

After all, $\Delta(X)$ and $\theta(X)$ are solved correctly up to the second-order with respect to q, namely,

$$\Delta(X) = \Delta_1 e^{\delta(X)}, \tag{117}$$

and

$$3\theta(X) = 3\theta_1(X) + 2\tan^{-1} t(X). \tag{118}$$

Here

$$\delta(X) = -\epsilon e^{-q(1+\cos 3\theta_1(X))}(\cos 3\theta_1(X) + q\partial_X^2 \cos 3\theta_1(X)) - \epsilon, \tag{119}$$

and

$$t(X) = q^2\left(\tfrac{1}{12} + \tfrac{11}{12}\cos 3\theta_1(X)\right)\sin 3\theta_1(X). \tag{120}$$

The phase $\theta_1(X)$ is given by

$$3\theta_1(X) = \pi + 2\tan^{-1}(\exp(X + \sqrt{\tanh^{-1} 3q})) + 2\tan^{-1}(\exp(X - \sqrt{\tanh^{-1} 3q})). \tag{121}$$

In Fig. 6, we show the trajectory of $gu(x)$ for various values of electron–phonon coupling. At $\lambda \simeq 0.175$, trajectory is almost like a straight line, $\mathrm{Re}(g \cdot u(x)/\Delta_1) = 0.5$. The trajectory enters into the inner side of the straight line at $\lambda = 0.18$.

Lastly, let us examine the soliton width. Introducing a trial function,

$$3\theta(x) = 3\tan^{-1}(\sqrt{3}\tanh(x/\xi_1)) = 3\tan^{-1}\left(\sqrt{3}\tanh\left(\frac{\xi_0}{\xi 1}\frac{X}{2\sqrt{3\epsilon}}\right)\right), \tag{122}$$

we determine ξ_1/ξ_0 by applying the least square method to Eq. (118). Using ξ_1/ξ_0, the soliton width is given by

$$\xi_1 = \frac{\xi_1}{\xi_0} \cdot \frac{v_F}{\Delta_1} = \frac{\xi_1}{\xi_0} \cdot \frac{\sqrt{3}}{2t} a. \tag{123}$$

In Fig. 7, ξ_1/ξ_0 is plotted as a function of λ. ξ_1/ξ_0 becomes minimum nearly at $\lambda \simeq 0.18$, whereas the soliton width Eq. (123) decreases monotonically when $0 < \lambda < 0.19$. The soliton width at $\lambda = 0.175$ is about $1.45 v_F/\Delta_1$.

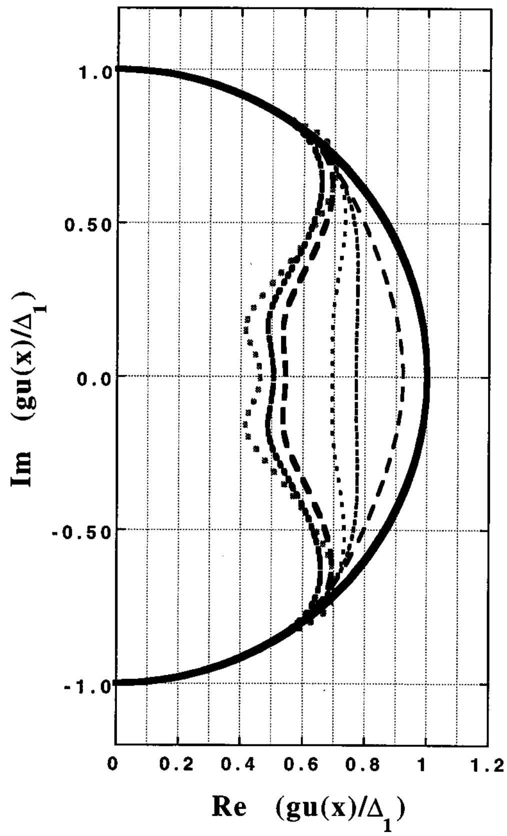

FIG. 6. The soliton configuration obtained by the perturbational calculation with Eq. (121) as the unperturbed part. The thick circular line is $\Delta(x)=\Delta_1$. The coupling constants are selected as $\lambda=0.1$ (the broken line), $\lambda=0.12$ (the dotted line), $\lambda=0.14$ (the rough dotted line), $\lambda=0.16$ (the thick broken line), $\lambda=0.175$ (the big dotted line), and $\lambda=0.19$ (the rough big dotted line).

V. NUMERICAL SOLUTIONS

In this section, we try finding numerical solutions for the effective Lagrangian derived in Sec. III. Since the space variable is continuous in this Lagrangian, we must discretize it in such a way

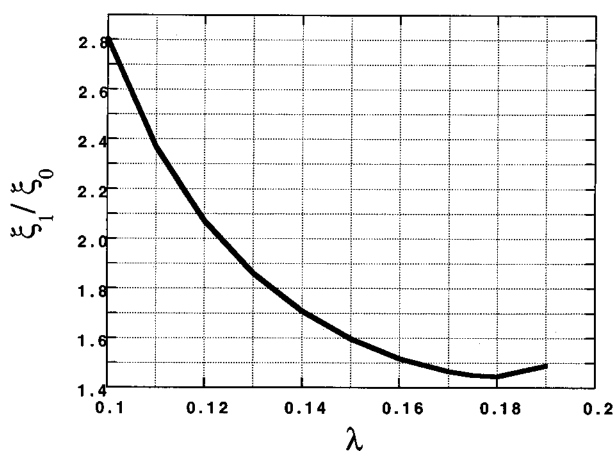

FIG. 7. The quantity ξ_1 defined by Eq. (122). It is plotted as a function of the coupling constant λ. The length ξ_0 is defined by Eq. (46).

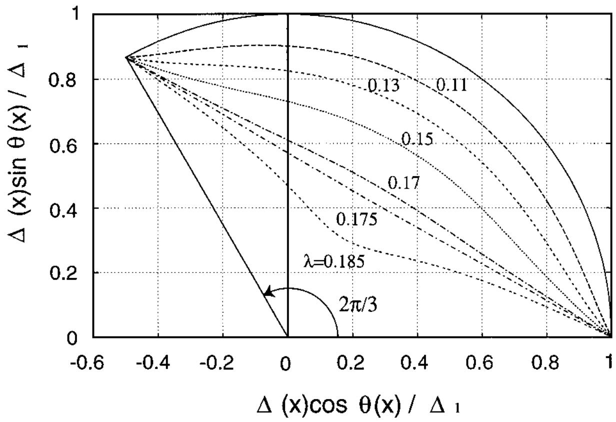

FIG. 8. The trajectories of the soliton. They are numerically obtained using the variational function Eqs. (124) and (125).

as $x=\Delta_m l$. In the following calculation, the mesh Δ_m is chosen to be small enough compared with the soliton half-width ξ. Actually, it is less than one-hundredth of ξ, even in the case with the shortest soliton width. Moreover, the system size should be large enough compared with the same length. Here we set it to be four times the largest ξ treated here.

We assume variational functional forms for the order parameters as

$$\Delta(x)=\Delta_1\sqrt{1-\tfrac{3}{4}\delta\,\mathrm{sech}(x/\xi_2)^2}, \tag{124}$$

and

$$3\theta(x)=3\tan^{-1}(\sqrt{3}\tanh(x/\xi_2))+\pi, \tag{125}$$

where ξ_2 and δ are determined so as to minimize the Hamiltonian. It should be remarked that these forms allow both types of solitons, namely, phase solitons and straight-line solitons, depending on ξ_2 and δ.

In Fig. 8, we show the optimized trajectories of the order parameter in the complex plane. As is seen clearly, the order parameter changes its shape from a phase soliton type via a straight-line to the one deformed in the inner direction. More specifically, the straight-line behavior is realized around $\lambda=0.175$, which is very close to that obtained in the exact solution.[7]

Next, we turn attention to the spatial change of the above order parameter. In Fig. 9, we plot the phase part in a weak-coupling case. Here two configurations are compared. One is the optimized solution at $\lambda=0.11$ and the other is the pure phase soliton described by a sine-Gordon solution. They are very similar, and hence this means that our variational function is correctly chosen in the weak-coupling limit.

While, beyond the weak-coupling limit, we find some discrepancy. In Fig. 10 we show the relationship between the coupling constant λ and the soliton width in the unit of v_F/Δ_1. At $\lambda=0.175$, which is the aforementioned value that gives the straight-line shape, the soliton width is about $1.54v_F/\Delta_1$. This value of λ is consistent with the result given by the analytical solution. However the width is longer than the exact value, i.e., $(2/\sqrt{3})v_F/\Delta_1\simeq 1.15v_F/\Delta_1$.[7] We think that this discrepancy comes from the neglected terms in deriving the effective potential.

In the rest of this section, we briefly check whether the above solutions are realistic. First, we discuss the straight-line soliton. At $\lambda=0.175$, the soliton width is about $1.54v_F/\Delta_1$. Since v_F is $(\sqrt{3}/2)Wa$ in the present definition and Δ_1 itself is calculated to be $0.104W$,

$$\xi_2=1.54v_F/\Delta_1=1.54(\sqrt{3}/2)Wa/0.104W=12.8a. \tag{126}$$

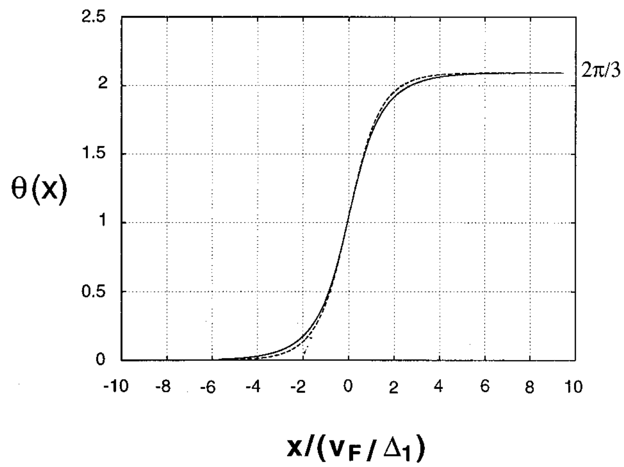

FIG. 9. The spatial pattern of the order parameter. The broken line is a pure phase soliton and the solid line is the soliton configuration at $\lambda = 0.11$.

Thus, this soliton width is much larger than the lattice constant a and is reasonable as a width. While, if the coupling is very small, the width is, for example, $2.38 v_F/\Delta_1$ at $\lambda = 0.11$. Since Δ_1 is $0.012W$ in this case, ξ_2 is calculated in the same way to be about $174a$. This value will be too large to observe such a soliton as an intrinsic object.

VI. SUMMARY

We have investigated the fractionally charged soliton in the trimerized electron–phonon system. We have found by both analytical and numerical methods that, in addition to the kink-type distortion in the phase mode, the soliton shows a peculiar distortion in the amplitude. The latter is new in the sense that the ordinary sine-Gordon model does not describe it.

We have used the Fröhlich model. Renormalizing the electronic part by the loop expansion of the partition function, we have constructed the effective Lagrangian of the order parameter. This can be applied to the investigation of the dynamics of both the amplitude and phase modes. It should be remarked that this Lagrangian is the almost same as that derived by the analysis of the

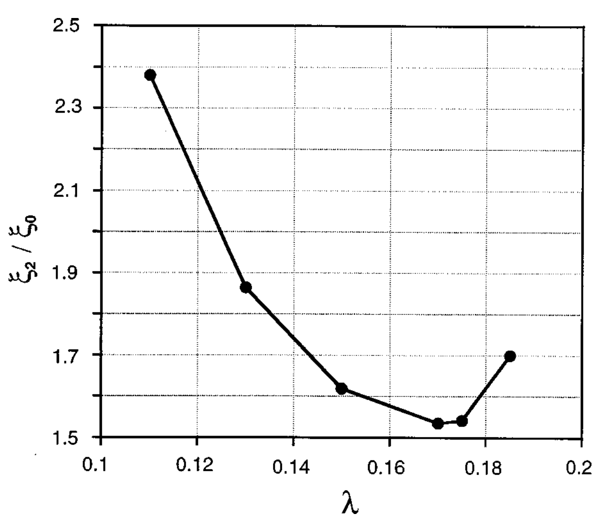

FIG. 10. The quantity ξ_2 defined by Eq. (125). It is plotted as a function of the coupling constant λ.

self-consistent equation for the Fröhlich model.[4] However, there are some discrepancies in the form of the derivative terms of the order parameter, and in the strength of the interaction term between the amplitude and phase modes.

Investigating the equations of the order parameter, we have found the analytical solution corresponding to the fractionally charged soliton. The solution is given by Eqs. (117)–(120). In the derivation, we have used the reductive perturbational method using the particular form of the unperturbed solution Eq. (121). The amplitude decreases around the soliton center. As the electron–phonon coupling becomes stronger, the distortion grows more and more. At $\lambda=0.175$, roughly speaking, the lattice distortion draws a straight-line in the complex plane. This behavior is the same with the exact solution given by Iwano *et al.*[7]

We have also analyzed the soliton solution numerically. In the calculation, the space variable is discretized and the variational functional forms are introduced. We have obtained the optimized trajectories of the order parameter. It has been shown that the order parameter changes its shape from a phase-soliton type via a straight line to the one deformed in the inner direction. Again the straight-line shape is realized nearly at $\lambda=0.175$, which is very close to that obtained in the exact solution. As for the soliton width, both the methods give similar tendencies. The width itself monotonically shortens as λ decreases. However, if it is scaled by ξ_0, it shows a saturation around $\lambda=0.17-0.18$. We think that this is characteristic of the straight-line solution.

The above results show that, in the trimerized electron-phonon system, the interaction between the phase and amplitude modes induces the distortion in the amplitude of the soliton. Thus, the spatial form of the soliton deviates substantially from that of the phase soliton when the electron–phonon coupling is not very weak.

It is important that the charge of the soliton is always the same, namely, fractional ones as $\pm\frac{2}{3}e$, irrespective of the spatial form. Thus, if we conceptually make it clear, we have two types of fractionally charged solitons, namely, a pure phase soliton and a strongly amplitude-deformed soliton, although both are related to each other by gradual changes. It of course depends on the parameter of each realistic system which type is observed experimentally. We, however, add that the former might be difficult to observed, because it has a very large width more than one hundred times the lattice constant.

ACKNOWLEDGMENT

One of the authors (R.S.) would like to express his special thanks to Professor K. Sasaki for fruitful discussions.

APPENDIX A: THE EFFECTIVE POTENTIAL

Here treat the main part of the Lagrangian concerning the lattice dynamics and study the influence of the lattice distortion on the electrons. The expression of the Lagrangian is given in Eq. (13). The partition function is given in Eq. (30).

By the path integration over $\tilde{\phi}_s$, the contribution from the electrons is renormalized to an effective potential. This estimation can be performed using a diagrammatic calculation. We take the summation over connected loop diagrams which are shown in Fig. 1.[17,18] The solid line is associated with the electron Green function. The solid circle corresponds to the vertex given by H_I. In the momentum space, the electron Green function is given by

$$G^{(0)}(k;\mathrm{i}\omega)\,\delta_{s,s'}=-\int_0^{\beta}\int\langle T_\tau\tilde{\phi}_s(x,\tau)\tilde{\phi}^{\dagger}_{s'}(x',0)\rangle e^{-\mathrm{i}k(x-x')-\mathrm{i}\omega\tau}\mathrm{d}x\,\mathrm{d}\tau=\frac{\mathrm{i}\omega I+v_F k\sigma_3}{(\mathrm{i}\omega)^2-(v_F k)^2}\,\delta_{s,s'}, \tag{A1}$$

where s denotes up or down of the spin orientation, and $\tilde{\phi}_s(x,\tau)$ is in the Heisenberg representation,

$$\tilde{\phi}_s(x,\tau)=\exp(H^0_{\tilde{\phi}}\tau)\cdot\tilde{\phi}_s(x)\cdot\exp(-H^0_{\tilde{\phi}}\tau). \tag{A2}$$

$H^0_{\tilde{\phi}}$ is the electronic part of the unperturbed Hamiltonian. The vertex is given by, in the momentum space,

$$D(q)=\frac{\mathrm{i}v_F q}{2}\chi(q)I-B(q)I+A(q)\sigma_1+\mathrm{i}\frac{1}{2}\frac{\partial\chi(q)}{\partial\tau}\sigma_3, \tag{A3}$$

where $A(q)$ and $B(q)$ are the Fourier transforms of $A(x)$ and $2\Delta(x)^2/(3W)$, respectively. Representing the summation of the diagrams in Fig. 1 as U, the partition function is approximately given by

$$\frac{Z}{Z_0}=\frac{\int(Du_rDu_i)\exp\left\{\frac{e}{\pi}\int\int\chi(x)E(x)\mathrm{d}x\mathrm{d}\tau-\int\mathrm{d}\tau(L^0_\Delta-U)\right\}}{\int(Du_rDu_i)\exp\left\{-\int\mathrm{d}\tau L^0_\Delta\right\}}. \tag{A4}$$

First, let us consider the contribution from the static lattice displacement. The lattice displacement $u(x)$ is given by

$$u_m=u_0e^{\mathrm{i}\theta_1},$$

and

$$\Delta(x)=gu_0=\Delta_1, \tag{A5}$$

where u_0 and θ_1 is constant. $D(x)$ is constant and takes the value,

$$D_0=-\frac{2\Delta_1^2}{3W}I+A_0{}^2\sigma_1,$$

with

$$A_0^2=\Delta_1^2\left(1+\frac{4\Delta_1^2}{9W^2}-\frac{4\Delta_1}{3W}\cos 3\theta_1\right). \tag{A6}$$

Contribution to the effective potential U coming from the static lattice displacement is given by the summation,

$$\frac{1}{\beta}\mathrm{Tr}\sum_{n,s,\omega}\int\mathrm{d}x\int\mathrm{d}k\frac{1}{2\pi n}[D_0G^{(0)}(k;\mathrm{i}\omega)]^n$$
$$=\frac{1}{\pi\beta}\sum_{\omega}\int\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}k\ln\left(\frac{\left(\mathrm{i}\omega-\frac{2\Delta_1^2}{3W}\right)^2-(v_Fk)^2-A_0^2}{(\mathrm{i}\omega)^2-(v_Fk)^2}\right), \tag{A7}$$

where the summation with respect to ω is taken over the Matsubara frequencies $\omega=T(2m+1)\pi$, m being an integer. In the calculation, we use the relation,

$$\mathrm{Tr}\ln(aI-b\sigma_1-c\sigma_3)=\ln\det(aI-b\sigma_1-c\sigma_3)=\ln(a^2-b^2-c^2). \tag{A8}$$

Taking the zero-temperature limit, the summation over ω changes to the integration, i.e.,

$$\frac{1}{\beta}\sum_{\omega,\mathrm{odd}}f(\mathrm{i}\omega)\to\frac{1}{2\pi}\int\mathrm{d}zf(\mathrm{i}z), \tag{A9}$$

where $f(\alpha)$ is a function of a complex variable α. Using Eq. (A9) and the relation

$$\frac{v_F}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}k\,\ln\left(\frac{\mathrm{i}v_Fk-a}{\mathrm{i}v_Fk-b}\right)=\frac{1}{2}(a-b),$$

for

$$0\leqslant \mathrm{Arg}(a),\quad \mathrm{Arg}(b)\leqslant\frac{\pi}{2}, \tag{A10}$$

Eq. (A7) is rewritten as

$$\begin{aligned}&\frac{1}{\pi}\int \mathrm{d}x\int_{-\overline{W}}^{W}\mathrm{d}z\left(\sqrt{\left(z+\mathrm{i}\frac{2\Delta_1^2}{3W}\right)^2+A_0^2}-|z|\right)\\&=-\frac{1}{2\pi v_F}\int \mathrm{d}x\left\{A_0^2\left(\ln\left|\frac{A_0^2}{(2\overline{W})^2}\right|-1\right)+\frac{1}{2}\left(\frac{4\Delta_1^2}{3W}\right)^2\right\},\end{aligned} \tag{A11}$$

where $\overline{W}$ is the energy cut off of the lowest electron band. In the trimerized case, it is $\overline{W}=W/2$.

The deviations of $\Delta(x)$ and $\theta(x)$ from $\Delta(x)=\Delta_1$ and $\theta(x)=\theta_0$, respectively, give an extra contribution to the effective potential. We define them as $\eta(x)=D(x)-D_0$ and assume that the time dependencies of $\Delta(x)$ and $\theta(x)$ are small enough. The contribution is given by,

$$\begin{aligned}-\mathrm{Tr}\frac{1}{\beta}\sum_{n=1}^{\infty}\sum_{\omega,s}\frac{1}{(2\pi)^n n}\int \mathrm{d}k\int \mathrm{d}q_1\cdots\int \mathrm{d}q_{n-1}\\ \times G(k;\mathrm{i}\omega)\,\eta(q_1)G(k+q_1;\mathrm{i}\omega)\,\eta(q_2-q_1)G(k+q_2;\mathrm{i}\omega)\,\eta(q_3-q_2)\\ \times\cdots G(k+q_{n-1};\mathrm{i}\omega)\,\eta(-q_{n-1}),\end{aligned} \tag{A12}$$

where $G(k;\mathrm{i}\omega)$ is the electron Green function in the stationary lattice distortion, which is

$$G(k;\mathrm{i}\omega)=G^{(0)}(k;\mathrm{i}\omega)[1-D_0G^{(0)}(k;\mathrm{i}\omega)]^{-1}=\frac{\left(\mathrm{i}\omega-\frac{2\Delta_1^2}{3W}\right)I+v_Fk\sigma_3+A_0\sigma_1}{\left(\mathrm{i}\omega-\frac{2\Delta_1^2}{3W}\right)^2-(v_Fk)^2-A_0^2}, \tag{A13}$$

and $\eta(q)$ is the Fourier transform of $\eta(x)$, namely,

$$\eta(q)=\int \mathrm{d}x\,\mathrm{e}^{-\mathrm{i}qx}\,\eta(x). \tag{A14}$$

The diagrams corresponding to Eq. (A12) are the same as those in Fig. 1, where the solid line and the solid circle correspond to the Green function Eq. (A13) and the vertex $\eta(q)$, respectively.

Now, we introduce the notation,

$$\eta(q)=\delta A(q)\sigma_1-\delta B(q)I+\mathrm{i}\frac{v_Fq}{2}\chi(q)I+\mathrm{i}\frac{1}{2}\frac{\partial\chi(q)}{\partial\tau}\sigma_3, \tag{A15}$$

where

$$\delta A(q)=\int \mathrm{d}x\,\mathrm{e}^{-\mathrm{i}qx}(A(x)-A_0),$$

and

$$\delta B(q)=\int \mathrm{d}x e^{-\mathrm{i}qx}\left(\frac{2\Delta(x)^2}{3W}-\frac{2\Delta_1^2}{3W}\right). \tag{A16}$$

Then, substituting Eq. (A14) into Eq. (A12), it becomes

$$\begin{aligned}
&-\sum_{n,\omega,s}\frac{1}{2\pi n\beta}\int \mathrm{d}x\int \mathrm{d}k\, \mathrm{Tr}[G(k;\mathrm{i}\omega)\eta(x)]^n\\
&=\frac{1}{\pi\beta}\sum_{\omega}\int \mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}k\\
&\quad\times\ln\left[\frac{\left(\mathrm{i}\omega+\frac{2\Delta(x)^2}{3W}-\frac{v_F}{2}\frac{\partial\chi}{\partial x}\right)^2-\left(v_F k+\mathrm{i}\frac{1}{2}\frac{\partial\chi}{\partial\tau}\right)^2-A(x)^2}{\left(\mathrm{i}\omega+\frac{2\Delta_1^2}{3W}\right)^2-(v_F k)^2-A_0^2}\right]\\
&=\frac{1}{2\pi v_F}\int \mathrm{d}x\left[\frac{1}{2}\left(\frac{4\Delta_1{}^2}{3W}\right)^2-\frac{1}{2}\left(\frac{4\Delta(x)^2}{3W}-v_F\frac{\partial\chi(x)}{\partial x}\right)^2\right.\\
&\quad\left.+A(x)^2\left(\ln\left|\frac{(2\overline{W})^2}{A(x)^2}\right|+1\right)-A_0^2\left(\ln\left|\frac{(2\overline{W})^2}{A_0^2}\right|+1\right)\right],
\end{aligned} \tag{A17}$$

where we use Eq. (A10). In the calculation of Eq. (A17), we drop the q_i dependency of the Green function $G(k+q_i;\mathrm{i}\omega)$.

We can make the calculation of the loop diagrams more exactly. Consider the second-order loop diagram with respect $\eta(q)$ that is shown in Fig. 2. The second-order term is given as follows:

$$\begin{aligned}
&-\frac{1}{2\beta(2\pi)^2}\mathrm{Tr}\sum_{\omega,s}\int_{-\infty}^{\infty}\mathrm{d}k\int_{-\infty}^{\infty}\mathrm{d}q\,G(k+q;\mathrm{i}\omega)\eta(q)G(k;\mathrm{i}\omega)\eta(-q)\\
&=-\frac{1}{4\pi^3 v_F}\int_{-\infty}^{\infty}\mathrm{d}(v_F k)\int_{-\infty}^{\infty}\mathrm{d}q\int_{-\overline{W}}^{\overline{W}}\mathrm{d}z\left\{\left(A_0^2+\left(\mathrm{i}z+\frac{2\Delta_1^2}{3W}\right)^2-v_F^2 k(k+q)\right)\delta A(q)\,\delta A(-q)\right.\\
&\quad+\left(A_0^2+\left(\mathrm{i}z+\frac{2\Delta_1^2}{3W}\right)^2+v_F^2 k(k+q)\right)\left(\delta B(q)-\frac{\mathrm{i}v_F q}{2}\chi(q)\right)\left(\delta B(-q)+\frac{\mathrm{i}v_F q}{2}\chi(-q)\right)\\
&\quad\left.+\left(-A_0^2+\left(\mathrm{i}z+\frac{2\Delta_1^2}{3W}\right)^2+v_F^2 k(k+q)\right)\left(\mathrm{i}\frac{1}{2}\frac{\partial\chi(q)}{\partial\tau}\right)\left(\mathrm{i}\frac{1}{2}\frac{\partial\chi(-q)}{\partial\tau}\right)\right\}\\
&\quad\times\left[\left(\mathrm{i}z+\frac{2\Delta_1^2}{3W}\right)^2-(v_F k)^2-A_0^2\right]^{-1}\left[\left(\mathrm{i}z+\frac{2\Delta_1^2}{3W}\right)^2-(v_F(k+q))^2-A_0^2\right]^{-1}\\
&=\frac{1}{\pi v_F}\int \mathrm{d}x\,\delta A(x)\left(\ln\left|\frac{2\overline{W}}{A_0}\right|-\frac{\sqrt{1+\gamma^2}}{\gamma}\sinh^{-1}\gamma\right)\delta A(x)\\
&\quad-\frac{1}{\pi v_F}\int \mathrm{d}x\left(\delta B(x)-\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}\right)\frac{\sinh^{-1}\gamma}{\gamma\sqrt{1+\gamma^2}}\left(\delta B(x)-\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}\right),
\end{aligned} \tag{A18}$$

where γ is defined by

$$\gamma=-\mathrm{i}\frac{v_F}{2\Delta_1}\frac{\partial}{\partial x}. \tag{A19}$$

In the calculation, the time dependency of the momentum transfer q is ignored. It corresponds to the adiabatic approximation. We can see easily that the contributions

$$\frac{1}{2\pi v_F}\int \mathrm{d}x(\delta A(x))^2\left(\ln\left|\frac{2\overline{W}}{A_0}\right|-1\right), \tag{A20}$$

and

$$-\frac{1}{\pi v_F}\int \mathrm{d}x\left(\delta B(x)-\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}\right)^2, \tag{A21}$$

in the r.h.s. of Eq. (A18) are already included in Eq. (A17). Then, the difference between Eq. (A18) and Eqs. (A20)–(A21) contributes to the effective potential. Adding it to Eq. (A17), and considering the total contribution to the effective potential, U is estimated as

$$\begin{aligned} U=&-\frac{1}{2\pi v_F}\int \mathrm{d}xA^2(x)\left(\ln\left|\frac{A^2(x)}{W^2}\right|-1\right)\\ &-\frac{1}{\pi v_F}\int \mathrm{d}xA(x)\left(\frac{\sqrt{1+\gamma^2}}{\gamma}\sinh^{-1}\gamma-1\right)A(x)\\ &-\frac{1}{\pi v_F}\int \mathrm{d}x\left(B(x)-\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}\right)\left(\frac{\sinh^{-1}\gamma}{\gamma\sqrt{1+\gamma^2}}\right)\left(B(x)-\frac{v_F}{2}\frac{\partial\chi(x)}{\partial x}\right). \end{aligned} \tag{A22}$$

Lastly, expanding $\sqrt{1+\gamma^2}$ and $\sinh^{-1}\gamma$ around $\gamma=0$ and neglecting higher-order terms with respect to γ than the second-order ones, the effective potential Eq. (37) is obtained.

APPENDIX B: SOLUTIONS OF EQS. (76) AND (78)

Substituting Eq. (79) into Eq. (76), the equation of $v_1(x)$ is rewritten as

$$\partial_X^2 v_1(x)+v_1(x)\left(\frac{2}{\cosh^2 X}-1\right)-\frac{3\tanh X}{\cosh^3 X}=0. \tag{B1}$$

Because $\tanh X/\cosh^3 X$ is the odd function, $v_1(x)$ should be odd with respect to the replacement $x\to -x$. Then we can introduce parameters $t_{1,n}$ ($n=0,1,2,...$) by

$$v_1(x)=\sum_{m=1}^{\infty} t_{1,m}\frac{\tanh X}{\cosh^{2m+1} X}. \tag{B2}$$

Substituting Eq. (B2) into Eq. (B1) and equating the coefficients of the functions $\tanh X/\cosh^{2m+1} X$ on both sides, the parameters are determined by

$$t_{1,0}=-\tfrac{3}{2}, \tag{B3}$$

and

$$t_{1,m}=0, \quad \text{for positive integer } m. \tag{B4}$$

The function $v_1(x)$ is therefore given by

$$v_1(x)=-3\tanh X/(2\cosh X). \tag{B5}$$

On the other hand, using Eqs. (81) and (B5), the equation of $v_2(x)$, namely, Eq. (78) is rewritten as,

$$\frac{\partial^2 v_2(X)}{\partial X^2}+v_2(X)\left(\frac{2}{\cosh^2 X}-1\right)+\frac{\tanh X}{2\cosh^3 X}\left(\frac{213}{\cosh^2 X}-123\right)=0. \tag{B6}$$

Introducing parameters $t_{2,m}$, we write $v_2(X)$ by

$$v_2(x)=\sum_{m=1}^{\infty} t_{2,m}\frac{\tanh X}{\cosh^{2m+1} X}. \tag{B7}$$

Substituting Eq. (B7) into Eq. (B6), and setting each coefficient of the functions $\tanh X/\cosh^{2m+1} X$ to zero, the parameters $t_{2,m}$ are determined as

$$t_{2,0}=-\tfrac{85}{24}, \tag{B8}$$

$$t_{2,1}=\tfrac{71}{12}, \tag{B9}$$

and

$$t_{2,m}=0, \quad \text{for } m\geqslant 2. \tag{B10}$$

After all, the function $v_2(x)$ is solved as

$$v_2(x)=-\frac{85}{24}\frac{\tanh X}{\cosh X}+\frac{71}{12}\frac{\tanh X}{\cosh^3 X}. \tag{B11}$$

[1] W. P. Su, J. R. Schrieffer, and A. J. Heeger, Phys. Rev. Lett. **42**, 1698 (1979); Phys. Rev. B **22**, 2099 (1980).
[2] H. Takayama, Y. R. Lin-Liu, and K. Maki, Phys. Rev. B **21**, 2388 (1980).
[3] H. Frölich, Proc. R. Soc. A **223**, 296 (1954).
[4] B. Horovitz and J. A. Krumhansl, Phys. Rev. B **29**, 2109 (1984).
[5] M. Grabowski and K. R. Subbaswamy, Solid State Commun. **34**, 911 (1980).
[6] A. C. Scott, F. Y. F. Chu, and D. W. Mclaughlin, Proc. IEEE **61**, 1443 (1973).
[7] K. Iwano and Y. Wada, J. Phys. Soc. Jpn. **56**, 3970 (1987).
[8] S. A. Brazovskii, JETP Lett. **28**, 606 (1978); Sov. Phys. JETP **51**, 342 (1980).
[9] K. Fujikawa, Phys. Rev. D **21**, 2848 (1980).
[10] B. Sakita and K. Shizuya, Phys. Rev. B **42**, 5586 (1990).
[11] N. Nagaosa and M. Oshikawa, J. Phys. Soc. Jpn. **65**, 2241 (1996).
[12] T. L. Ho, Phys. Rev. Lett. **48**, 1416 (1982); B. Horovitz, *ibid.* **48**, 1416 (1982); J. Hara, and H. Fukuyama, J. Phys. Soc. Jpn. **52**, 2128 (1983).
[13] R. Shirasaki, Synth. Met. **69**, 677 (1995).
[14] H. Fukuyama and H. Takayama, ''Electronic properties of inorganic quasi-one-dimensional compounds,'' edited by P. Monceau (Reidel, Boston, 1985), Part I, pp. 41–104.
[15] Y. Ono, Y. Ohfuti, and A. Terai, J. Phys. Soc. Jpn. **54**, 2641 (1985).
[16] Y. Ono, J. Phys. Soc. Jpn. **55**, 4164 (1986).
[17] J. M. Luttinger and J. C. Ward, Phys. Rev. **118**, 1417 (1960).
[18] S. Coleman and E. Weinberg, Phys. Rev. D **7**, 1888 (1973).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Classification and conformal symmetry in three-dimensional space–times

G. S. Hall[a] and M. S. Capocci[b]
Department of Mathematical Sciences, Meston Building, University of Aberdeen, Aberdeen AB24 3UE, Scotland, United Kingdom

(Received 28 August 1998; accepted for publication 26 October 1998)

Three-dimensional manifolds admitting Lorentz metrics are studied. The first part of the paper gives a classification of the Ricci and curvature tensors and also of the conformal (Schouten–Cotton–York) tensor. The second part of the paper investigates Killing and conformal symmetry and also the nature of the zeros of the associated vector fields. The maximum dimension of the Killing and conformal algebras is calculated. A theorem regarding the reduction of the conformal algebra to a Killing algebra of a conformally related metric is given. © *1999 American Institute of Physics.* [S0022-2488(99)01103-2]

I. INTRODUCTION

The recent interest in three-dimensional space–times together with the known usefulness and elegance of the various classification schemes in the four-dimensional space-time of general relativity suggest that a discussion of such schemes in the former may be useful.

Perhaps the most useful classification scheme in general relativity is the Petrov classification of the Weyl (conformal) tensor. The ''equivalent'' Weyl tensor in three-dimensions is zero and must be replaced by a tensor introduced by Schouten[1] (but often referred to as the *Cotton–York tensor*) which, like the Weyl tensor in four or more dimensions, is conformally invariant and which vanishes identically if and only if the metric is locally conformal to a flat metric. The classification of this tensor will be dealt with in Sec. IV.

The vanishing of the Weyl tensor in three dimensions leads to a close relationship between the Ricci and Riemann tensors. This link is explored in Sec. III where these tensors are classified by Segre type.

The final three sections of the paper deal with the study of conformal (including homothetic and Killing) vector fields on a three-dimensional space–time. Some results regarding the nature of zeros of such vectors fields are given and these are used to prove theorems on the maximum dimension of the conformal and related Lie algebras. A theorem concerning the reduction of the conformal algebra to a Killing algebra with respect to a conformally related metric is obtained. Section II contains some preliminary geometrical remarks about three-dimensional space–times.

II. THREE-DIMENSIONAL SPACE–TIMES

A three-dimensional space–time is a three-dimensional smooth paracompact connected manifold M admitting a global smooth Lorentz metric g of signature $(-,+,+)$. The associated Riemann and Ricci tensors are denoted by their respective components R_{abcd} and $R_{ab}(\equiv R^c_{acb})$ and a covariant derivative arising from the Levi-Civitá connection associated with g is denoted by a semicolon. Round and square brackets denote the usual symmetrization and skew symmetrization, respectively.

Let $m\in M$ and let T_mM denote the tangent space to M at m. Denote by Λ_m the three-dimensional vector space of skew symmetric tensors (bivectors) of either type (0,2) or (2,0) at m (there being no need to distinguish them for the present purposes because of the natural isomor-

[a]Electronic mail: gsh@maths.aberdeen.ac.uk
[b]Present address: 9 Victoria Street, Aylesbury, Buckinghamshire, HP20 1LZ, U.K.; electronic mail: michael@coexis.com

phism between them which arises from the metric $g(m)$ at m and in components is $F_{ab} \mapsto F^{ab}$). A triplet (l,n,x) of members of T_mM is called a *null triad* if the only nonvanishing inner products between them are $l^a n_a = x^a x_a = 1$ (so that l and n are null and x spacelike). A triplet $x,y,t \in T_mM$ is called an *orthogonal triad* if the only nonvanishing inner products between them are $x^a x_a = y^a y_a = -t^a t_a = 1$ (so that x and y are spacelike and t is timelike). If $0 \neq F \in \Lambda_m$ then its rank is even and hence equal to 2. Thus all bivectors in Λ_m are *simple*, that is, $F \in \Lambda_m$ can be written as $F^{ab} = 2p^{[a}q^{b]}$ for $p,q \in T_mM$. The two-dimensional subspace of T_mM spanned by p and q is uniquely determined by F and called the *blade* of F. A nonzero member of Λ_m can be classified as *timelike*, *null*, or *spacelike* according to whether its blade contains exactly two, exactly one, or no *null* one-dimensional subspaces (directions).

Let S be a nonzero symmetric type (0,2) tensor at m with components S_{ab}. The eigenvector–eigenvalue problem $(S_{ab} - \lambda g_{ab})k^b = 0$ at m for $k \in T_mM$ and $\lambda \in \mathbb{C}$ leads to a classification for S in terms of its Segre type. The details were given in Ref. 2 and based on the four-dimensional situation.[3,4] All four Segre types {1,11}, {21}, {3}, and $\{z\bar{z}1\}$ are possible and their canonical expressions in terms of a null triad (l,n,x) are, respectively,

$$S_{ab} = 2\alpha l_{(a}n_{b)} + \beta(l_a l_b + n_a n_b) + \gamma x_a x_b\,, \tag{1}$$

$$S_{ab} = 2\alpha l_{(a}n_{b)} + \lambda l_a l_b + \beta x_a x_b\,, \tag{2}$$

$$S_{ab} = 2\alpha l_{(a}n_{b)} + l_{(a}x_{b)} + \alpha x_a x_b\,, \tag{3}$$

$$S_{ab} = 2\alpha l_{(a}n_{b)} + \beta(l_a l_b - n_a n_b) + \gamma x_a x_b\,. \tag{4}$$

Here $\alpha,\beta,\gamma, \in \mathbb{R}$ and $\gamma = \pm 1$. In (1) the eigenvectors are $l \pm n$ and x with corresponding eigenvalues $\alpha \pm \beta$ and γ. In (2) the eigenvectors are l and x with respective eigenvalues α and β while in (3) the eigenvector is l with eigenvalue α. In (4), which is the only case to admit nonreal eigenvalues, $\beta \neq 0$ and the eigenvectors are $l \pm in$ and x with respective eigenvalues $\alpha \pm i\beta$ and γ. The type {1,11} is the only one where timelike eigenvectors can occur and the associated (unique) eigenvalue is represented by the first digit in the Segre symbol and is separated from the others by a comma. The degeneracies of these Segre types (denoted by enclosing the appropriate digits inside round brackets) are also possible. An alternative form for (1) in an orthogonal triad is

$$S_{ab} = \rho z_a z_b + \gamma x_a x_b - \delta t_a t_b\,, \tag{5}$$

where $\rho, \delta \in \mathbb{R}$, $\sqrt{2}z = l + n$, $\sqrt{2}t = l - n$ and so $\alpha + \beta = \rho$ and $\alpha - \beta = \delta$.

In (2) and (3) l spans the unique null eigendirection and in (1) there are no null eigendirections unless $\beta = 0$ (type {(1,1)1}) in which case there are exactly two independent ones spanned by l and n. There are no null eigendirections in (4).

The bivectors in Λ_m may also be classified according to Segre type. In fact in terms of an appropriate null triad a null bivector may be written as $2l_{[a}x_{b]}$ and a timelike bivector as $2l_{[a}n_{b]}$ and their respective Segre types are {3} and {111} while, in terms of an appropriate orthogonal triad a spacelike bivector may be written as $2x_{[a}y_{b]}$ and its Segre type is then $\{z\bar{z}1\}$. These canonical forms (but not the Segre types) will be required later.

There is a convenient ''duality'' between T_mM and Λ_m in a three-dimensional space–time achieved by using the alternating tensor ϵ_{abc}. Associated with $k \in T_mM$ the bivector $F \in \Lambda_m$ where $F_{ab} = \epsilon_{abc}k^c$ (and then k and F will be referred to as the *duals* of each other). Thus a bijective duality relation is established between T_mM and Λ_m. From this correspondence one finds $F_{ab}k^b = 0$ and so this duality relation may be geometrically regarded as associating a (necessarily simple) bivector with a vector (naturally scaled by ϵ) perpendicular to its blade. The vector k lies in the blade of F if and only if k (and hence F) is null. From the mathematical viewpoint it can be interpreted as a means of ''identifying'' a single tensor index with a skew symmetric pair of tensor indices. Like the duality operation in general relativity it will turn out to be rather useful algebraically and in this respect one notes the relation

$$\epsilon^{abc}\epsilon_{ade} = -(\delta^b_d\delta^c_e - \delta^c_d\delta^b_e). \tag{6}$$

It is remarked here that the above duality map between T_mM and Λ_m extends naturally to the complexification of these vector spaces taking a complex vector k^a to the complex bivector $\epsilon_{abc}k^c$. This will be required in dealing with the possible occurrence of complex eigenvectors in a later section.

The Lorentz group $\mathcal{L}$ in three-dimensions is the subgroup of GL(3,$\mathbb{R}$) given by $\mathcal{L}=\{A \in \mathrm{GL}(3,\mathbb{R}): A^T\eta A=\eta\}$ where $\eta=\mathrm{diag}(-1,1,1)$ and A^T is the transpose of A. The group $\mathcal{L}$ can, in a natural way, be given the structure of a three-dimensional Lie group which is a Lie subgroup of GL(3,$\mathbb{R}$). Let $\mathcal{L}_0$ be the identity component of $\mathcal{L}$ and let $l \in T_mM$ be null. The subgroup $N(l)$ of $\mathcal{L}_0$ of *null rotations about l* is the subgroup of $\mathcal{L}_0$ which preserves the direction of l. It is a two-dimensional Lie subgroup of $\mathcal{L}_0$ described by its effect on a null triad (l,n,x) by

$$l \mapsto l' = Al, \quad x \mapsto x' = x + bl, \quad n \mapsto n' = A^{-1}\left(n - bx - \frac{b^2}{2}l\right), \tag{7}$$

where $A,b \in \mathbb{R}$, $A>0$. Under such a null rotation n may be transformed to any other null vector at m except a null vector proportional to l.

III. THE RIEMANN AND RICCI TENSORS

The vanishing of the Weyl tensor on a three-dimensional space–time means that the Ricci and Riemann tensors are related by

$$R_{abcd} = 2R_{a[c}g_{d]b} + 2R_{b[d}g_{c]a} + Rg_{a[d}g_{c]b}, \tag{8}$$

where $R \equiv R_{ab}g^{ab}$ is the Ricci scalar. As a consequence the Ricci tensor vanishes at $m \in M$ if and only if the Riemann tensor vanishes at m.

Now construct the ''double dual'' T of the Riemann tensor where, in components,

$$T^{ef} = \epsilon^{abe}R_{abcd}\epsilon^{cdf}. \tag{9}$$

Thus T is a second order tensor and is easily seen to be symmetric. The inverse relation to (9) can be found using (6) and is

$$R_{abcd} = \tfrac{1}{4}\epsilon_{abe}T^{ef}\epsilon_{cdf}. \tag{10}$$

Theorem 1: *A vector k in T_mM (or its complexification) is an eigenvector of T with eigenvalue $\alpha \in \mathbb{C}$ if and only if the bivector $F_{ab}=\epsilon_{abc}k^c$ in Λ_m (or its complexification) is an eigenbivector of the Riemann tensor with eigenvalue $-\alpha/2 \in \mathbb{C}$.*

Proof: Suppose that $T^{ef}k_f = \alpha g^{ef}k_f = \alpha k^e$. Then (9) on premultiplying by ϵ_{emn} gives

$$\epsilon_{emn}\epsilon^{abe}R_{abcd}F^{cd} = \alpha\epsilon_{emn}k^e. \tag{11}$$

Use of (6) then reveals

$$R_{abcd}F^{cd} = -\frac{\alpha}{2}F_{ab} \tag{12}$$

and so F is an eigenbivector of the Riemann tensor with eigenvalue $-\alpha/2$. The converse result is similar and this completes the proof. □

Because of the close relationship between T and the Riemann tensor expressed in theorem 1, one is tempted to enquire about the relationship between T and the Ricci tensor. The answer is easily obtained by computing T in (9) using (8) and (6). One finds

$$T_{ab} = 4(R_{ab} - \tfrac{1}{2}Rg_{ab}) \tag{13}$$

and so T is, up to a numerical factor, the Einstein tensor. The Bianchi identities for the Riemann tensor then reappear, through duality, in the identities

$$T_a{}^b{}_{;b}=0. \tag{14}$$

It follows from (13) that $k\in T_mM$ (or its complexification) is an eigenvector of T if and only if it is an eigenvector of the Ricci tensor and that, although the eigenvalues will in general differ, any eigenvalue degeneracies will be preserved. From this remark and theorem 1 it is clear that the algebraic structure of T, of the Ricci tensor and of the Riemann tensor (regarded in the obvious way as a linear map from Λ_m to itself) are the same in the sense that their Segre type (including degeneracies) are identical.

Theorem 2: *If M is a three-dimensional space–time and $m\in M$ then M may be algebraically classified at m according to the Segre type of the Riemann tensor, the Ricci tensor, or the double dual of the Riemann tensor (the tensor T) at $m\in M$. The Segre type obtained is independent of whichever of these tensors is used and is either* $\{1,11\}$, $\{21\}$, $\{3\}$, $\{z\bar{z}1\}$, *or a degeneracy of one of these types.*

Proof: The proof follows from the preceding remarks and the material in Sec. II. □

Thus, as a corollary, one sees that every bivector in Λ_m is an eigenbivector of the Riemann tensor with the same eigenvalue if and only if every vector in T_mM is an eigenvector of the Ricci tensor with the same eigenvalue. Hence M is of constant curvature at m if and only if it is an Einstein space at m. The link between the algebraic structure of the Riemann and Ricci tensors was given earlier.[5]

Recalling the elegant and useful Bel criteria which can be used in an alternative formulation of the Petrov classification in general relativity[6,7] it is interesting to ask if one can reformulate the classification in theorem 2 in terms of ''canonical'' null directions in T_mM.

Theorem 3: *Let M be a three-dimensional space–time, let $m\in M$ and let $l\in T_mM$ be null. Then at m*:

(1) $R_{ab}l^al^b=0 \Leftrightarrow l_{[e}R_{a]bc[d}l_{f]}l^bl^c=0$.
(2) $R_{ab}l^a=\alpha l^b(\alpha\in\mathbb{R}) \Leftrightarrow R_{abcd}l^bl^c=\lambda l_al_d(\lambda\in\mathbb{R})$.
(3) $R_{abcd}l^d=F_{ab}l_c\neq 0$ *(where F is a null bivector in Λ_m satisfying $F_{ab}l^b=0$)* $\Leftrightarrow$ *the Ricci tensor has Segre type* $\{3\}$ *with eigenvector l and eigenvalue zero*.
(4) $R_{abcd}l^d=0$ $(R_{abcd}\neq 0)\Leftrightarrow$ *the Ricci tensor has Segre type* $\{(21)\}$ *with null eigenvector l and eigenvalue zero*.

Proof:

(1) If $R_{ab}l^al^b=0$ then a contraction of (8) with l^bl^c reveals the desired condition on the Riemann tensor. Conversely, this latter condition means that $R_{abcd}l^bl^c=2l_{(a}p_{d)}$ where the Riemann symmetries imply that p satisfies $l^ap_a=0$. A contraction over the indices a and b then gives $R_{ab}l^al^b=0$.
(2) If $R_{ab}l^b=\alpha l_a$ then a contraction of (8) with l^bl^c reveals the given condition on the Riemann tensor. Conversely, this latter condition implies $R_{ab}l^al^b=0$ and then the same contraction of (8) yields $R_{ab}l^b=\alpha l_a$.
(3) The condition on the Riemann tensor immediately implies that $R_{ab}l^b=0$ and that $R_{abcd}F^{cd}=0$ (since F is null). This latter condition implies that $T_{ab}l^b=0$ (theorem 1) and, since $R_{ab}l^b=0$, Eq. (13) yields $R=0$. Now construct a null triad (l,n,x) with $F_{ab}=2l_{[a}x_{b]}$ and contract (8) with l^dn^b (using $R=0$) to get $R_{ab}=2l_{(a}q_{b)}$ with $l^aq_a=0$ and $q\not\propto l$. A comparison with (3) with $\alpha=0$ reveals the stated Segre type for the Ricci tensor. Conversely, writing the Ricci tensor as in (3) with $\alpha=0$ one finds that $R=0$ and a contraction of (8) with l^d completes the proof.
(4) Using the real null triad employed in the proof of part 3 the conditions on the Riemann tensor show that $R_{abcd}F^{cd}=0$ where $F_{ab}=2l_{[a}x_{b]}$ or $F_{ab}=2l_{[a}n_{b]}$. Hence, from theorem 1, $T_{ab}l^b=0$ and $T_{ab}x^b=0$. Because $R_{abcd}\neq 0$ at m, $T\neq 0$ at m and so from (1)–(4) T must have Segre

type {(21)} with zero eigenvalue since if it had Segre type {1,11} [the only other type admitting (at least) two independent real eigenvectors] it would vanish. Thus $T_{ab}=\mu l_a l_b$ ($\mu\in\mathbb{R}$) and (13) reveals that $R_{ab}=(\mu/4)l_a l_b$ and the result follows. Conversely, the latter expression for R_{ab} used in (8) shows that $R_{abcd}l^d=0$.

It is remarked here that, from the Segre theory and Eqs. (1)–(4) in Sec. II, the null direction l in parts 3 and 4 of theorem 3 is unique and that at most two null directions could satisfy the conditions of part 2. What is less obvious is that there are at most four null directions satisfying the conditions of part 1. To see this suppose that at least one such direction exists and is spanned by l and construct a null triad (l,n,x) at m. Under a null rotation (7) n may be rotated to n' and, by a suitable choice of b, to span any null direction at m other than that spanned by l. The equation $R_{ab}n'^a n'^b=0$ is then a polynomial in b of order less than or equal to three and the result follows. There may, of course, be no null directions satisfying condition 1 of theorem 3.

IV. THE SCHOUTEN–COTTON–YORK TENSORS

For a three-dimensional space–time M consider the tensor $\bar{R}$ with components R_{abc} given in any coordinate system by

$$R_{abc}=2R_{a[b;c]}+\tfrac{1}{2}R_{;[b}g_{c]a}=R^d{}_{abc;d}+\tfrac{1}{2}R_{;[b}g_{c]a}\,. \tag{15}$$

This tensor has the properties

(1) $R_{abc}=-R_{acb}$,
(2) $R^a{}_{ac}=0$,
(3) $R_{[abc]}=0$,

and also the property that it is *conformally invariant* in the sense that it is unchanged if the metric g on M is changed to the metric $e^\sigma g$ on M for a smooth function $\sigma:M\to\mathbb{R}$. Further $\bar{R}$ vanishes on some open neighborhood U of m if and only if for some open neighborhood $V\subseteq U$ of m there exists a flat metric γ on V and a smooth function $\rho:V\to\mathbb{R}$ such that, on $V, g=e^\rho\gamma$.[1,8]

To algebraically classify $\bar{R}$ at $m\in M$ one recalls the ''duality identification'' made in Sec. II between T_mM and Λ_m. Now the tensor $\bar{R}$ may be thought of as a linear map $T_mM\to\Lambda_m$ given by $k^a\mapsto R_{abc}k^a$. Thus with the above identification $\bar{R}$ gives rise to a linear map $T_mM\to T_mM$ given by

$$k^a\mapsto R_{abc}k^a\mapsto -\tfrac{1}{2}\epsilon^{bca}(R_{dbc}k^d)=Y^a{}_d k^d, \tag{16}$$

where

$$Y^{ab}=-\tfrac{1}{2}\epsilon^{acd}R^b{}_{cd} \tag{17}$$

is the *Cotton–York tensor*.[8] The tensor Y is easily seen to be trace-free from the third property of $\bar{R}$ listed above and can also be shown to be symmetric by use of the Bianchi identity.[8] Thus $Y_{ab}=Y_{ba}$, $Y^a{}_a=0$ and (17) inverts to give

$$R^a{}_{bc}=\epsilon_{dbc}Y^{da}. \tag{18}$$

This bijective correspondence between $\bar{R}$ and Y suggests that a classification of $\bar{R}$ may be achieved by a classification of Y similar to that given in Sec. II. It is noted here that, while Y vanishes at m if and only if $\bar{R}$ does, the tensor Y is not conformally invariant. M will be called *conformally flat* if $\bar{R}\equiv 0\ (\Leftrightarrow Y\equiv 0)$ on M.

Theorem 4: *If M is a three-dimensional space–time the tensor Y may be classified at each $m\in M$ into four Segre types together with their degeneracies. They are* {1,11}, {(1,1)1}, {1,(11)}, {(1,11)}, {21}, {(21)}, {3}, *and* $\{z\bar{z}1\}$. *The trace-free condition means that if the type is* {(1,11)} *then* $Y=0$ *at m and that if the type is* {(21)} *or* {3} *the eigenvalue is zero.*

Proof: This follows immediately from (1)–(4) after imposing the trace-free condition. □

An alternative approach is to directly classify the tensor $\bar{R}$ according to its canonical null directions and in the spirit of theorem 3 for the Riemann tensor. To achieve this first note that if $k\in T_mM$ then

$$Y^a{}_b k^b=0 \Leftrightarrow R_{abc}k^a=0, \tag{19a}$$

$$Y^a{}_b k^b=\alpha k^a(\alpha\in\mathbb{R}) \Leftrightarrow R_{abc}k^ak^b=0(\Leftrightarrow R_{abc}k^ak^c=0). \tag{19b}$$

The proof of (19a) is immediate from (16). To prove (19b) let $G_{ab}=R^c{}_{ab}k_c$. Then $Y^a{}_bk^b=\alpha k^a$ and (17) imply that $\epsilon^{abc}G_{bc}\propto k^a$ and hence that $G_{ab}k^b=0$. Hence $R_{abc}k^ak^b=0$. Conversely, $R_{abc}k^ak^b=0$ implies that $G_{ab}k^b=0$ and hence that $\epsilon^{abc}G_{bc}\propto k^a$. Thus $Y^a{}_bk^b\propto k^a$. The results (19a) and (19b) are independent of whether *k* is null and are also true if *k* is complex. Hence they may be used as a reformulation of the classification in theorem 4. But it is interesting to consider the classification of $\bar{R}$ according to real *null* vectors *k* satisfying either (19a) or (19b) and to bear in mind [Eqs. (1)–(4)] that there are at most two independent real null solutions to each of Eqs. (19a) and (19b). The following theorem results directly from theorem 4.

Theorem 5: *Let M be a three-dimensional space–time, let* $m\in M$ *and suppose that* $\bar{R}(m)\neq 0$. *Then*

(1) *There are exactly two independent null solutions of (19b) if and only if Y has Segre type* {(1,1)1}.
(2) *There is exactly one independent null solution of (19b) and no non-trivial null solutions of (19a) if and only if Y has Segre type* {21}.
(3) *There is exactly one independent null solution of each of (19a) and (19b) if and only if Y has Segre type* {(21)} *or* {3} *(with eigenvalue necessarily equal to zero in each case).*
(4) *There is at most one nontrivial null solution of (19a).*

Again one sees a similarity to the well known Bel criteria for the Weyl tensor of general relativity. The number of independent *null* solutions for $\bar{R}(m)\neq 0$ of (19a) is 0 or 1 and of (19b) is 0, 1, or 2 and each can actually occur as (1)–(4) show. Also Eq. (16) and theorem 4 suggest that one thinks of the eigenvalues of the tensor Y at $m\in M$ as being the equivalent, in a three-dimensional space–time, of the well known Petrov scalars in a four-dimensional space–time. At this point one may suggest a classification scheme for $\bar{R}$ by using the same symbols as in the Petrov case but with a prime attached (and an extra distinguishing suffix needed in two of the types). The types at *m* are, from theorem 4:

Type I_1' (when $Y(m)$ has Segre type {1,11} and, recalling the trace-free condition, two independent eigenvalues),
Type I_2' ($\{z\bar{z}1\}$ and two independent eigenvalues),
Type D_1' ({(1,1)1} and one independent eigenvalue),
Type D_2' ({1,(11)} and one independent eigenvalue),
Type II' ({21} and one independent eigenvalue),
Type N' ({(21)} and all eigenvalues zero),
Type III' ({3} and all eigenvalues zero).

The relationship between each of these types and the corresponding Petrov type is clear. It is also important to note the geometry of this classification. If the type at *m* is I_1' then a unique timelike direction and a unique pair of spacelike directions are determined at *m* by the eigenvector structure. If the type at *m* is I_2' then a unique pair of null directions and a unique spacelike direction is determined at *m*. For type D_1' (respectively, D_2') a unique timelike (respectively, spacelike) two-space and a unique spacelike (respectively, timelike) direction are determined at *m*.

For types II', N', and III' a unique null and a unique spacelike direction are determined at m and they are orthogonal. A similar geometric discussion can be given for the earlier classifications of the Ricci and Riemann tensors.

V. CONFORMAL SYMMETRY ON A THREE-DIMENSIONAL SPACE–TIME

In this section conformal symmetry will be studied in the usual way by a discussion of conformal (including homothetic and Killing) vector fields on the three-dimensional space–time M. Further technical details may be found in Ref. 9.

A smooth vector field X on M is called *conformal* if in any chart of M

$$X_{a;b}=\phi g_{ab}+F_{ab}(\Leftrightarrow \mathcal{L}_X g=2\phi g), \tag{20}$$

where $F_{ab}(=-F_{ba})$ is the (conformal) bivector of X, $\phi: M\rightarrow \mathbb{R}$ is smooth and $\mathcal{L}$ denotes a Lie derivative. It is, in fact, sufficient to assume that X is C^3 for this implies that X is smooth.[10] A conformal vector field is called *homothetic* if ϕ is constant on M and *Killing* if $\phi\equiv 0$ on M. It is called *proper homothetic* if $\phi\equiv constant\neq 0$ and *proper conformal* if X is not homothetic. From (20) it follows that

$$F_{ab;c}=R_{abcd}X^d-2g_{c[b}\phi_{a]}, \tag{21}$$

$$\phi_{;ab}=2R_{c(a}F_{b)}{}^c-2\phi L_{ab}-L_{ab;c}X^c, \tag{22}$$

where $\phi_a=\phi_{;a}$ and $L_{ab}=R_{ab}-\frac{1}{4}Rg_{ab}$. The sets of all conformal (respectively, homothetic, Killing) vector fields on M will be denoted by C (respectively, H,K) so that $K\subseteq H\subseteq C$ and each of K, H, and C is a Lie algebra of smooth vector fields on M. If $m\in M$ the subset of C (respectively, H,K) consisting of vector fields which vanish at m is denoted by C_m (respectively, H_m, K_m) and is a subalgebra of C (respectively, H,K). If $X\in C$ and if the local diffeomorphisms of M associated with X in the usual way (see, e.g., Refs. 11 and 12) are denoted by ψ_t then, for each $X\in C$, $X(m)=0$ if and only if $\psi_t(m)=m$ for appropriate t and such a point m is referred to as either a *zero* or a *fixed point* of X. Such a zero of X is called *isometric* if $\phi(m)=0$ or *homothetic* if $\phi(m)\neq 0$. The subset of C_m consisting of all conformal vector fields on M for which m is an isometric zero is denoted I_m. Thus $I_m=\{X\in C_m : \phi(m)=0\}$ and is a subalgebra of C_m and C. Any conformal vector field X on M is uniquely determined by the values $X^a(m)$, $X^a_{;b}(m)$, and $X^a_{;bc}(m)$ [or equivalently by $X^a(m)$, $F_{ab}(m)$, $\phi(m)$, and $\phi_a(m)$] at any $m\in M$.

The study of conformal symmetry in the four-dimensional space–time of general relativity is facilitated by knowledge of what goes on at zeros of conformal vector fields[13–16] and a similar approach will be adopted here.

Theorem 6: *Let M be a three-dimensional space–time and X a conformal vector field on M. Then*

(1) $\mathcal{L}_X R_{abc}=0$,
(2) $\mathcal{L}_X Y^a{}_b=-3\phi Y^a{}_b$ *(or, equivalently,* $\mathcal{L}_X Y_{ab}=-\phi Y_{ab}$*)*,
(3) *If Y does not vanish over a nonempty open subset of M and l is a null eigenvector field of Y then* $\mathcal{L}_X l=\alpha l$ *for some smooth function* $\alpha: M\rightarrow\mathbb{R}$.

Proof: If one considers the smooth map ψ_t associated with X between open coordinate domains U and $\psi_t U$ then a consideration of the pullback ψ_t^{-1*} shows that this latter map preserves g up to a scalar $\lambda(t)$ (since X is conformal). The conformal invariance of $\overline{R}$ now reveals that $\psi_t^{-1*}\overline{R}=\overline{R}$ so that 1 follows. To establish 2, use of the standard Lie derivative formula together with the result that $\epsilon^{abc}_{;d}=0$ gives

$$\mathcal{L}_X \epsilon^{abc}=-3\phi\epsilon^{abc}+H^{abc}, \tag{23}$$

where H is a three-form on M satisfying $H^{abc}\epsilon_{abc}=0$. Hence $H=0$ and so

$$\mathcal{L}_X Y^a{}_b = \tfrac{1}{2}\mathcal{L}_X(\epsilon^{cad}R_{bcd}) = -3\phi Y^a{}_b\,. \tag{24}$$

For 3, if l is a null eigenvector field of Y then, from 2, so also is $\psi_t^{-1*}l$ since ψ_t^{-1*} preserves Y up to a scalar $\mu(t)$. Since Y has finitely many independent null eigenvectors at any point $m \in M$ where $Y(m)\neq 0$ it follows that $\psi_t^{-1*}l \propto l$ and 3 is proved. □

Theorem 7: *Let M be a three-dimensional space–time, $X(\not\equiv 0)$ a conformal vector field on M and let $m\in M$ be an isometric zero of X. Then*

(1) *If the conformal bivector F also vanishes at m (so that X, ϕ, and F all vanish at m) then $\overline{R}$ (and hence Y) vanishes at m.*
If on the other hand $F(m)\neq 0$ then the blade of F is, for $Y(m)\neq 0$, an eigenspace of $Y(m)$ and hence coincides with the two-space geometrically determined at m by the type of $Y(m)$ as discussed at the end of Sec. IV. In detail,
(2) *If $F(m)$ is spacelike, $Y(m)$ is either zero or of Segre type $\{1,(11)\}$ (type D_2');*
(3) *If $F(m)$ is timelike, $Y(m)$ is either zero or of Segre type $\{(1,1)1\}$ (type D_1');*
(4) *If $F(m)$ is null, $Y(m)$ is either zero or of Segre type $\{(21)\}$ with zero eigenvalue (type N').*

Proof: Under the conditions of 1, one has $X^a(m)=0$, $X^a_{;b}(m)=0$. Then if one writes out (24) in terms of covariant derivatives, covariantly differentiates, eliminates the bivector covariant derivative using (21) and evaluates at m one finds

$$\phi_b Y_{ad} - (Y_{ac}\phi^c)g_{bd} - (Y^c_b\phi_c)g_{ad} + Y_{bd}\phi_a = -3\phi Y_{ab}\,. \tag{25}$$

The skew part of (25) over the indices b and d when contracted with ϕ^a then gives $\phi^a Y_{a[d}\phi_{b]}=0$ at m. Now $\phi_a(m)\neq 0$ (otherwise the conditions of part 1 of the theorem would imply that $X\equiv 0$ on M) and so $\phi^a(M)$ is an eigenvector of $Y(m)$ with eigenvalue $\lambda\in\mathbb{R}$, $Y^a_b\phi^b = \lambda\phi^a$ at m. On substituting this into (25) and contracting with ϕ^d at m one obtains $(\phi^c\phi_c)Y_{ab}=0$ and so either $Y(m)=0$ or $\phi_a(m)$ is null. If $Y(m)\neq 0$, $\phi_a(m)$ is null and the same substitution but now followed by a contraction with ϕ^b yields $\lambda\phi_a\phi_b=0$ at m and so $\lambda=0$. Equation (25) then gives at m

$$\phi_b Y_{ad} + \phi_a Y_{bd} + 3\phi_d Y_{ab} = 0. \tag{26}$$

Now (26) implies first that $Y_{ab}=\mu\phi_a\phi_b$ at m with $\mu\in\mathbb{R}$. A final substitution into (26) then shows that $\mu=0$ and so $Y(m)=0$.

For the remainder of the proof $F(m)\neq 0$ and it is noted that (24), when evaluated at m, gives the ''commuting'' relation

$$F^a{}_c Y^c{}_b = Y^a{}_c F^c{}_b\,. \tag{27}$$

For (2) suppose that $F(m)$ is spacelike and so there exists an orthogonal triad x,y,t at m such that, at m, $F_{ab}=2\nu x_{[a}y_{b]}$ $(0\neq\nu\in\mathbb{R})$ and $F_{ab}t^b=0$. Then a contraction of (27) with t^b and using the fact that t is the unique solution, up to a scaling, of the equation $F_{ab}t^b=0$ shows that $Y^a_b t^b = \alpha t^a$ $(\alpha\in\mathbb{R})$ at m. So regarding Y as a linear map $T_mM\to T_mM$ in the usual way, this map preserves the one-dimensional subspace of T_mM spanned by t and hence preserves its orthogonal complement spanned by x and y (since $x^a t_a=0$ implies $Y^a{}_b x^b t_a - \alpha t_b x^b=0$ and similarly for y). This latter subspace is the blade of F and is positive definite. Since Y is symmetric this induced linear action is diagonalizable over $\mathbb{R}$ and one may thus assume that x and y above are eigenvectors of Y with respective eigenvalues $\beta,\gamma\in\mathbb{R}$. A contraction of (27) with x_a then reveals that $\beta=\gamma$ and so the blade of $F(m)$ is an eigenspace of Y. Thus $Y(m)$, if nonzero, had Segre type $\{1,(11)\}$ (i.e., type D_2') and canonical form $Y_{ab}=\beta(x_ax_b+y_ay_b+2t_at_b)$ at m (since $Y^a{}_a=0\Rightarrow\alpha=-2\beta$).

For (3) suppose $F(m)$ is timelike so there exists a null triad l,n,x at m such that, at m, $F_{ab}=2\nu l_{[a}n_{b]}$ $(0\neq\nu\in\mathbb{R})$ and $F_{ab}x^b=0$. A contraction of (27) with x^b then shows that $Y^a{}_b x^b = \alpha x^a$ $(\alpha\in\mathbb{R})$ at m and a contraction of (27) with l^b and use of the fact that the eigendirection of

$F(m)$ (contracted over the second index of F) spanned by l is the only one with eigenvalue ν reveals that $Y^a_b l^b = \beta l^a$ ($\beta \in \mathbb{R}$) at m. Similarly one finds $Y^a{}_b n^b = \gamma n^a$ ($\gamma \in \mathbb{R}$) at m. Finally one then sees that $\beta = Y^a{}_b l^b n_a = \gamma$ and so the blade of $F(m)$ is an eigenspace of Y. Thus if $Y(m)$ is nonzero it has Segre type {(1,1)1} (type D_1') and canonical form $Y_{ab} = \alpha(l_{(a}n_{b)} - x_a x_b)$ at m.

For (4) if $F(m)$ is null a null triad l,n,x may be chosen at m such that, at m, $F_{ab} = 2\nu l_{[a}x_{b]}$, $(0 \neq \nu \in \mathbb{R})$. A contraction of (27) with l^b shows that $Y^a_b l^b = \alpha l^a$ ($\alpha \in \mathbb{R}$). Now regarding Y as a linear map on $T_m M$ one sees that Y preserves the null direction spanned by l and hence, as before, its orthogonal subspace which is the blade of $F(m)$. Thus $Y^a{}_b x^b = \beta l^a + \gamma x^a$ for $\beta, \gamma \in \mathbb{R}$. On substituting the information so far obtained into (27) one finds $\alpha = \beta = \gamma = 0$. Hence $Y_{ab} l^b = Y_{ab} x^b = 0$ and so either $Y(m)$ is zero or of Segre type {(21)} (type N') with canonical form $Y_{ab} \propto l_a l_b$. □

This theorem together with the classification scheme given in Sec. IV is a direct analog of a similar theorem[17,14,15] in general relativity.

Theorem 8: *Let M be a three-dimensional space–time,* $X(\not\equiv 0)$ *a conformal vector field on M and* $m \in M$ *a homothetic zero of X. Then all the eigenvalues of* $Y(m)$ *vanish and either*

(1) $Y(m) = 0$; *or*
(2) $Y(m)$ *has Segre type* {(21)} *(type* N'*) and a null triad l,n,x may be chosen at m such that* $Y_{ab} = \pm l_a l_b$ *and* $F_{ab} = 3\phi(m) l_{[a}n_{b]}$; *or*
(3) $Y(m)$ *has Segre type* {3} *(type* III'*) and a null triad l,n,x may be chosen at m such that* $Y_{ab} = 2 l_{(a}x_{b)}$ *and* $F_{ab} = 6\phi(m) l_{[a}n_{b]}$.

Proof: Regarding the eigenvalues of $Y(m)$ the proof proceeds along similar lines to that in Refs. 13 and 14 (see Ref. 18). In fact, since $\mathcal{L}_X g_{ab} = 2\phi g_{ab}$ and $\mathcal{L}_X Y_{ab} = -\phi Y_{ab}$ at m, it is easily seen that if $v \in T_m M$ is an eigenvector of $Y(m)$ with eigenvalue $\alpha \in \mathbb{R}$ (so that $Y_{ab} v^b = \alpha g_{ab} v^b$ at m) then $\psi_t^* v$ is an eigenvector of $Y(m)$ with eigenvalue $\alpha \exp(3\phi(m)t)$. Since there are finitely many such eigenvalues, $\alpha = 0$. A slight modification of this argument deals with the case when $\alpha \in \mathbb{C}$ and the first result in theorem 8 follows.

Thus $Y(m)$ is either zero or has Segre type {(21)} or {3}. If $Y(m)$ has Segre type {(21)} then there exists $l \in T_m M$ with l null and $Y_{ab} = \pm l_a l_b$. Evaluating (24) at m gives

$$Y^a{}_c F^c{}_b - F^a{}_c Y^c{}_b = -3\phi(m) Y^a{}_b \tag{28}$$

and substituting for $Y^a{}_b$ at m gives $F_{ab} l^b = \alpha l_a$ ($\alpha \in \mathbb{R}$) at m. Substituting back into (28) then gives $\alpha = \frac{3}{2}\phi(m) \neq 0$. It follows that $F(m)$ is timelike and l may be extended to a null triad with the desired conditions of part 2 of the theorem. For part 3 there exists a null triad l,n,x such that, at m, $Y_{ab} = l_{(a}x_{b)}$. A substitution into (28) and contractions first with l_a and then with x_a show that $F_{ab} l^b = \mu l_a$ with $\mu = 3\phi(m) \neq 0$ so that $F(m)$ is timelike. Another back substitution and appropriate contractions show that $F_{ab} x^b = 0$ and hence that l and n span the blade of $F(m)$. This completes the proof. □

It is of interest to note that in part 3 the triad l,n,x yielding the given expression for Y_{ab} at m is easily checked to be uniquely determined by $Y(m)$. In (2), however, there is a potential freedom in the blade of $F(m)$ represented by the null rotation $l \mapsto l$, $n \mapsto n - bx - (b^2/2) l$ and hence $l_{[a}n_{b]} \mapsto l_{[a}n_{b]} - b l_{[a}x_{b]}$. This allows for the possibility (theorem 7) of taking a linear combination of X with a Killing vector field on M which vanishes at m. No such possibility exists in (3) as theorem 7 shows.

Again one notes that this theorem together with the classification scheme in Sec. IV is the direct analog of the corresponding result in general relativity.[13,14] There is one final result (again with a general relativistic analog)[13,14] which is conveniently given here.

Theorem 9: *Let M be a three-dimensional space–time,* $X(\not\equiv 0)$ *a homothetic vector field on M and* $m \in M$ *a zero of X.*

If X is proper homothetic (and so m is necessarily a homothetic zero of X) all eigenvalues of the Ricci tensor at m are zero and, at m, either

(1) *The Ricci tensor vanishes; or*
(2) *The Ricci tensor has Segre type* $\{(21)\}$ *with eigenvalue zero and a null triad l,n,x may be chosen at m such that* $R_{ab}=\pm l_a l_b$ *and* $F_{ab}=2\phi l_{[a}n_{b]}$; *or*
(3) *The Ricci tensor has Segre type* $\{3\}$ *with eigenvalue zero and a null triad l,n,x may be chosen at m such that* $R_{ab}=2l_{(a}x_{b)}$ *and* $F_{ab}=4\phi l_{a[}n_{b]}$.
If X is Killing (and so m is necessarily an isometric zero of X) then, at m, either
(4) *The Ricci tensor is a multiple of the metric tensor (possibly zero); or*
(5) *The Ricci tensor has Segre type* $\{1,(11)\}$ *and an orthogonal triad t,x,y may be chosen at m such that* $R_{ab}=\alpha(x_a x_b+y_a y_b)-\beta t_a t_b$ *and* $F_{ab}=2\mu x_{[a}y_{b]}$, $(\alpha,\beta,\mu\in\mathbb{R})$; *or*
(6) *The Ricci tensor has Segre type* $\{(1,1)1\}$ *and a null triad l,n,x may be chosen at m such that* $R_{ab}=2\alpha l_{(a}n_{b)}+\beta x_a x_b$ *and* $F_{ab}=2\mu l_{[a}n_{b]}$, $(\alpha,\beta,\mu\in\mathbb{R})$; *or*
(7) *The Ricci tensor has Segre type* $\{(21)\}$ *and a null triad l,n,x may be chosen at m such that* $R_{ab}=\alpha(2l_{(a}n_{b)}+x_a x_b)\pm l_a l_b$ *and* $F_{ab}=2\mu l_{[a}x_{b]}$, $(\alpha,\beta,\mu\in\mathbb{R})$.

Proof: The proof follows in a way sufficiently similar to those for theorems 7 and 8 for it only to be necessary to sketch it briefly. If X is homothetic then $\mathcal{L}_X R_{ab}=0$ and when computed at m this gives

$$2\phi R_{ab}+R_{ac}F^c{}_b+R_{bc}F^c{}_a=0. \tag{29}$$

If X is proper homothetic then $\phi\neq 0$ and the condition $\mathcal{L}_X R_{ab}=0$ reveals, as for Y in theorem 8, that each Ricci eigenvalue is zero at m. The proofs of parts 1–3 now follow from (29). If X is Killing, $\phi=0$, then one considers the separate cases when $F(m)$ is spacelike, timelike and null substituting the appropriate canonical forms for $F(m)$ into (29) to obtain parts 4–7. Again one notes that when X is Killing (and so m is an isometric zero of X) the blade of $F(m)$ is an eigenspace of the Ricci tensor when the latter is nonzero at m (cf. theorem 7) and hence one links the geometry of F at m with that of the canonical Ricci tensor type at m.

In the case when $X(\not\equiv 0)$ is Killing or homothetic and $X(m)=0$ the possibilities for the tensor Y at m are covered by theorems 7 (parts 2–4) and 8. □

Theorem 10: *Let M be a three-dimensional space–time, X a proper homothetic vector field on M and* $m\in M$ *a zero of X. Then either* $Y(m)$ [*and hence* $\bar{R}(m)$] *is zero or the Ricci tensor (and hence the Riemann tensor) vanishes at m.*

Proof: The proof is immediate from theorem 8 and theorem 9 parts 1–3 because of the incompatibility of the values for $F(m)$ if $Y(m)$ and the Ricci tensor at m are nonzero. □

Let X be a conformal vector field on a three-dimensional space–time M with a *homothetic* zero at m. Then if $Y(m)\neq 0$ it follows from theorem 8 that rank $X^a{}_{;b}(m)=3$. Hence, by the implicit function theorem, the zero of m at X is isolated. This result is not true for four-dimensional space–times, a counterexample being the well known plane wave metric.[19,13] It is also not true for three-dimensional space–times if the condition $Y(m)\neq 0$ is dropped. This can be seen by the following construction[20] (see also Refs. 13 and 19).

With X as above, the only possibility of m not being isolated [that is the rank of $X^a{}_{;b}(m)$ is less that 3] is, from theorems 8 and 9, when $Y(m)=0$ and when the conditions of theorem 9 (part 2) hold at m. In this case rank $X^a{}_{;b}=2$ and the condition $X^a{}_{;b}(m)n^b=0$ leads to a null geodesic of zeros of X with tangent n at m.[20] The metric, locally about m, can then be written as[20]

$$ds^2=dx^2+2\,du\,dv+f(u)x^2\,du^2 \tag{30}$$

for an arbitrary smooth function f. This metric admits the covariantly constant covector field $l_a=u_{;a}$ and the Ricci tensor satisfies $R_{ab}=-f(u)l_a l_b$. It follows from (15) that Y (and $\bar{R}$) are identically zero and the curvature tensor vanishes identically if and only if f does.[20] The vector field with components $(2v,0,x)$ is proper homothetic and its zeros are the null geodesic $x=v=0$. The bivector F and constant ϕ associated with X satisfy $F_{ab}=2\phi l_{[a}n_{b]}$ at these zeros and where n is a tangent to the corresponding null geodesic. A comparison with the four-dimensional case suggests referring to (30) as a three-dimensional plane wave and it serves as an ''example'' of theorem 10 where $Y(m)$ and $\bar{R}(m)$ vanish. Another ''example'' of this theorem arises from the metric given in a global chart v,u,x on $\mathbb{R}^3$ by[21] (see also Ref. 14)

$$ds^2 = e^{ux}\,du\,dv + dx^2. \tag{31}$$

This metric admits the proper homothetic vector field X with components $(3v,-u,x)$ and a Killing vector field $\partial/\partial v$ (and in fact $\dim C=2$). The coordinate origin is an isolated zero of X and the only nonvanishing curvature and Ricci tensor components (up to symmetries) are

$$\begin{aligned} R_{1212} &= \tfrac{1}{16}u^2 e^{2ux}, \quad R_{1223} = -\tfrac{1}{4}e^{ux}, \quad R_{1323} = -\tfrac{1}{8}u^2 e^{ux}, \\ R_{12} &= -\tfrac{1}{4}u^2 e^{ux}, \quad R_{23} = -\tfrac{1}{2}, \quad R_{33} = -\tfrac{1}{2}u^2. \end{aligned} \tag{32}$$

The Ricci tensor has Segre type {3} with zero eigenvalue at the zero (0,0,0) of X [see theorem 9 (part 3)] and it is easily checked from (15) that $\bar{R}$ (and hence Y) vanishes at this zero.

VI. THE DIMENSION OF THE CONFORMAL ALGEBRA

The Lie algebras C, H, and K of conformal, homothetic, and Killing vector fields, respectively, on M described in the last section satisfy the well known results that $\dim C \leqslant 10$, $\dim H \leqslant 7$, and $\dim K \leqslant 6$ with the optimum cases $\dim C=10$ (respectively, $\dim H=7$, $\dim K=6$) arising if $\bar{R}\equiv 0$, or equivalently $Y\equiv 0$ on M (respectively, if M is flat or M has constant curvature). In this section the values of the dimensions of these algebras will be explored under less restrictive conditions. The technique used is based on a similar one in general relativity[14,15] and depends heavily on theorems 7–9. The general idea is that when $\dim C$ exceeds the dimension of any orbit generated by C, zeros of members of C occur (in that orbit). The following theorem summarizes the present situation.

Theorem 11: *Let M be a three-dimensional space–time and let $m\in M$. Then*

(1) *If $Y(m)\neq 0 (\Leftrightarrow \bar{R}(m)\neq 0)$ then $\dim I_m \leqslant 1$ and $\dim C_m \leqslant 2$.*
(2) *If Y is of type {(21)} (N') at any point of M then $\dim C\leqslant 5$.*
(3) *If Y is of type {(1,1)1} (D_1'), type {1,(11)} (D_2') or type {3} (III') at any point of M then $\dim C \leqslant 4$.*
(4) *If Y is of type {21} (II'), type {1,11} (I_1') or type $\{z\bar{z}1\}$ (I_2') at any point of M then $\dim C \leqslant 3$.*
(5) *If Y is of type {(1,1)1} (D_1'), type {1,(11)} (D_2') or type {(21)} (N') at any point of M then $\dim K \leqslant 4$.*
(6) *If $\dim K \geqslant 5$ then M has constant curvature.*

Proof: For $Y(m)\neq 0$ the set $C(m)$, if not trivial, can (by taking appropriate linear combinations) be spanned by k independent conformal vector fields such that at least $k-1$ of them have m as an isometric zero. If $k-1>1$ then again by taking linear combinations and using theorem 7 one can find $X\in C_m$ with $\phi_m=0$ and $F(m)=0$ since all possible values of $F(m)$ are multiples of a particular bivector. The same theorem then shows that $Y(m)=0$. Thus if $Y(m)\neq 0$ then $\dim I_m \leqslant 1$ and $k=\dim C_m \leqslant 2$ since $\dim M=3$. It follows that $\dim C\leqslant 5$ or $\dim C \leqslant 4$ according as m is a homothetic zero of at least one member of C or not. Thus part 1 is proved and 2, 3, and 4 follow from theorems 7 and 8. Next, since $\dim I_m \leqslant 1$ and again taking linear combinations of members of K [and recalling that if X is Killing and $X(m)=0$, $F(m)=0$ then $X\equiv 0$ on M] part 5 follows from theorems 7 and 8. Finally if $\dim K\geqslant 5$ then theorem 7 shows that $Y\equiv 0$ on M and theorem 9 (parts 4–7) show that $R_{ab}\propto g_{ab}$ at each $m\in M$ and hence (see after theorem 2) M has constant curvature. This completes the proof. □

The above theorem shows that the conformal algebra C satisfies $\dim C \leqslant 5$ provided that Y is not identically zero on M. This result can, with a little more effort, be improved. The calculations are lengthy in places and so the proof will only be sketched with further details being available in Ref. 18.

Theorem 12: *Let M be a three-dimensional space–time which is not conformally flat. Then the conformal algebra C satisfies* $\dim C \leqslant 4$.

Proof: The previous theorem shows that $\dim C \leqslant 5$, so suppose that $\dim C = 5$. The algebra C leads to a ''generalized'' distribution Δ on M which associates with $m \in M$ the subspace $\Delta(m) = \{X(m) : X \in C\}$ of $T_p M$. Since the dimension of $\Delta(m)$ may vary with m one appeals to Hermann's generalization[22] (see also Ref. 9) of the Fröbenius theorem to obtain the existence of integral manifolds (orbits) associated with C through each $m \in M$ and whose dimension equals that of Δ_m at any m on the orbit. Now $\dim \Delta(m) + \dim C_m = \dim C = 5$ and, from theorem 10 (part 1), $\dim C_m \leqslant 2$ and so $\dim \Delta(m) = 3$ for each $m \in M$. A lengthy argument using the commutators and Jacobi relations associated with C then shows that $Y \equiv 0$ on M[18] and this contradiction completes the proof. □

The results of this section can be taken a step further. The study of the conformal algebra in any dimension possesses inherent difficulties due to its ''nonlinearity'' (see Ref. 14 for a discussion of this point). These difficulties do not arise for affine algebras (e.g., the Killing and homothetic algebras). An interesting approach to this problem was given in Ref. 23 and extended further in Ref. 24 and later in Ref. 15. These papers concerned general relativity and the essential idea, roughly speaking, was to show that under certain conditions the conformal algebra could (locally) be regarded as a Killing algebra with respect to a new metric which was conformally related to the original one. In the case of a three-dimensional space–time a similar theorem can now be stated. One interesting feature of it is the reduced number of restricting clauses compared with the corresponding four-dimensional result.[15] The proof is based on the work in the last two sections and further details can be found in Ref. 18.

Theorem 13: *Let M be a three-dimensional space–time with metric g. Suppose that the orbits associated with the conformal algebra C of M have the same dimension n at each* $m \in M$ *and, if* $n < 3$, *the same type at each* $m \in M$ [*i.e.* $\Delta(m)$ *is always spacelike, always timelike, or always null for all* $m \in M$]. *Then for* $m \in M$ *either*

(1) *There exists an open neighborhood U of m and a smooth function* $\sigma : U \to \mathbb{R}$ *such that the restrictions of members of C to U constitute a Lie algebra of Killing vector fields with respect to the metric* $e^{\sigma} g$; *or*

(2) *There exists an open neighborhood V of m on which Y vanishes; or*

(3) *The orbits are two-dimensional and null,* $C_m = I_m$, $\dim C_m = 1$, *and* $Y(m) = 0$.

1 J. A. Schouten, Math. Zeit. **11**, 58 (1921).
2 G. S. Hall, T. W. Morgan, and Z. Perjés, ''Three-dimensional spacetimes,'' Gen. Relativ. Gravit. **19**, 1137 (1987).
3 G. S. Hall, J. Phys. A **9**, 541 (1976).
4 J. Plebanski, Acta Phys. Pol. **26**, 964 (1964).
5 W. Kay, Ph.D. thesis, University of Aberdeen (1986).
6 L. Bel, Cah. Phys. **16**, 59 (1962).
7 R. K. Sachs, Proc. R. Soc. London, Ser. A **264**, 309 (1961).
8 J. W. York, Phys. Rev. Lett. **26**, 1656 (1971).
9 G. S. Hall, Gen. Relativ. Gravit. **30**, 1099 (1998).
10 G. S. Hall and J. da Costa, J. Math. Phys. **32**, 2848 (1991).
11 S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley Interscience, New York, 1963), Vol. 1.
12 R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications* (Springer Verlag, Berlin, 1988).
13 G. S. Hall, Gen. Relativ. Gravit. **20**, 671 (1988).
14 G. S. Hall, J. Math. Phys. **31**, 1198 (1990).
15 G. S. Hall and J. D. Steele, J. Math. Phys. **32**, 1847 (1991).
16 G. S. Hall, M. S. Capocci, and R. Beig, Class. Quantum Grav. **14**, L49 (1997).
17 J. Ehlers and W. Kundt, ''Exact solutions of the gravitational field equations,'' in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
18 M. S. Capocci, Ph.D. thesis, University of Aberdeen (1994).
19 D. V. Alexeevski, Ann. Global Anal. Geom. **3**, 59 (1985).

[20] G. S. Hall, D. J. Low, and J. R. Pulham, J. Math. Phys. **35**, 5930 (1994).
[21] J. K. Beem, Lett. Math. Phys. **2**, 317 (1978).
[22] R. Hermann, Contrib. Diff. Eq. **3**, 199 (1964).
[23] R. F. Bilyalov, Sov. Phys. Dokl. **8**, 878 (1964).
[24] L. Defrise-Carter, Commun. Math. Phys. **40**, 273 (1975).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

A rigorous solution concept for geodesic and geodesic deviation equations in impulsive gravitational waves

M. Kunzinger[a)] and R. Steinbauer[b)]
Department of Mathematics, University of Vienna, Strudlhofg. 4,
Institute for Theoretical Physics, University of Vienna, Boltzmanng. 5,
A-1090 Wien, Austria

(Received 9 June 1998; accepted for publication 18 November 1998)

The geodesic as well as the geodesic deviation equation for impulsive gravitational waves involve highly singular products of distributions $(\theta\delta,\theta^2\delta,\delta^2)$. A solution concept for these equations based on embedding the distributional metric into the Colombeau algebra of generalized functions is presented. Using a universal regularization procedure we prove existence and uniqueness results and calculate the distributional limits of these solutions explicitly. The obtained limits are regularization independent and display the physically expected behavior. © *1999 American Institute of Physics.* [S0022-2488(99)00903-2]

I. INTRODUCTION

Impulsive *pp*-waves (plane fronted gravitational waves with parallel rays) can be described by a metric of the form[1]

$$ds^2=\delta(u)f(x,y)du^2-du\,dv+dx^2+dy^2, \tag{1}$$

where (u,v) and (x,y) are a pair of null and (transverse) Cartesian coordinates, respectively, and f denotes the profile function subject to the field equations. Hence the space–time is flat everywhere except for the null hypersurface $u=0$, where it has a δ-like pulse modeling a gravitational shock wave. Such geometries arise most naturally as ultrarelativistic limits of boosted black hole space–times of the Kerr–Newman family (as shown by various authors[2–4]) and multipole solutions of the Weyl family.[5] Also, they play an important role in particle scattering at the Planck scale (see Ref. 6 and references therein).

There have also been intrinsic descriptions of impulsive *pp*-waves, viz. by Penrose[1] and by Dray and t'Hooft,[7] which essentially consist in glueing together two copies of Minkowski space–time with a warp across the null hypersurface $u=0$. Penrose also introduced a different coordinate system in which the components of the metric tensor are actually continuous. However, the transformation relating the coordinates used in (1) to these new ones is discontinuous (for the general form of the transformation see Ref. 8) and therefore—strictly speaking—the differential structure of the manifold is changed. In this paper we stick to the original distributional form of the metric, motivated by the fact that physically, i.e., in the ultrarelativistic limit, the space–time arises that way (cf. the approaches of Refs. 9 and 10). For recent work on *pp*-waves using the continuous form of the metric, see Ref. 11.

We describe the geometry of impulsive *pp*-waves entirely in the distributional picture using the framework of Colombeau's generalized functions, thereby generalizing previous work.[12] As discussed there in detail, the geodesic as well as the geodesic deviation equation for impulsive *pp*-waves involve formally ill-defined products of distributions, due to the nonlinearity of the equations and the presence of the Dirac δ-function in the space–time metric. However, as was also shown in Ref. 12, one can overcome these difficulties using a careful regularization procedure

[a)]Electronic mail: Michael.Kunzinger@univie.ac.at
[b)]Electronic mail: stein@doppler.thp.univie.ac.at

which, while mathematically sound, corresponds to the physical idea of viewing the impulsive wave as the limiting case of a sandwich wave of ever decreasing support but constant (integrated) strength. More precisely, regularizing the δ-distribution by a ''model δ-net'' [i.e., a net $\rho_\epsilon(x) := \epsilon^{-1}\rho(x\epsilon^{-1})$, where ρ is a smooth function with support contained in the interval $[-1,1]$ satisfying $\int\rho = 1$], it was shown that the solutions to the smoothened equations possess regularization-independent weak limits. These distributional ''solutions'' fit perfectly into the physically expected picture showing that the geometry of impulsive *pp*-waves can be described consistently using the distributional form of the metric. The reliability of the results is guaranteed by making use of regularization techniques instead of introducing ''multiplication rules'' into Schwartz linear distribution theory (cf. the discussion at the end of Sec. 2 in Ref. 12 or Ref. 13 for general remarks).

However, the ''solutions'' obtained by this naive regularization procedure exhibit a mathematically highly unsatisfactory feature. *They do not obey the original distributional equations* (unless, again, one is willing to impose certain ''multiplication rules''), as is common to such situations. Hence—strictly speaking—this approach does not provide a reasonable solution concept for the equations under consideration. Such a notion *is* available in the nonlinear theory of generalized functions[14–16] due to J. F. Colombeau, where one has—loosely speaking—a rigorous system of bookkeeping on the regularizing sequences. Recently Hermann and Oberguggenberger[17] (see also Ref. 18) studied systems of singular, nonlinear ordinary differential equations (ODEs) in the Colombeau algebra. In this work we are going to use similar techniques to treat the geodesic and geodesic deviation equation for impulsive *pp*-waves in the Colombeau algebra. Despite the nonlinearities involved in these equations (which in principle could lead to trapping, blow-up or reflection of solutions at the shock, cf. Ref. 17), we are able to prove existence and uniqueness of geodesics crossing the shock hypersurface. We derive the (regularization independent) distributional limits of these solutions, making use of the notion of association (see Sec. II below) in the algebra, thereby significantly generalizing the results of Ref. 12. In particular, the regularization of the δ-like wave profile will no longer be restricted to a ''model δ-net'' but belong to the largest ''reasonable'' class (cf. Definition 1 below). Moreover, note that the regularization idependence of the results has the following important physical consequence: in the impulsive limit the geodesics are totally independent of the particular shape of the sandwich wave. Hence the impulsive wave ''totally forgets its seed'' (cf. also the results in Ref. 19).

Finally, we discuss the case of a nonsmooth wave profile f and give an outlook to current research which allows us to fit our previous calculations into a manifestly covariant concept of Colomebeau algebras on manifolds.

II. MATHEMATICAL FRAMEWORK

A framework that allows consistent treatment of nonlinear operations with distributions and at the same time offers a well-developed theory of (linear and nonlinear) partial differential equations is provided by Colombeau's theory of algebras of generalized functions (cf., e.g., Refs. 14–16 and 20). To begin with, we give a short description of the algebra we are going to use in the sequel. Let

$$\mathcal{A}_0(\mathbb{R}^n) = \left\{ \varphi \in \mathcal{D}(\mathbb{R}^n) : \int \varphi(x)\,dx = 1 \right\},$$

$$\mathcal{A}_q(\mathbb{R}^n) = \left\{ \varphi \in \mathcal{A}_0(\mathbb{R}^n) : \int \varphi(x) x^\alpha dx = 0, 1 \leq |\alpha| \leq q \right\} (q \in \mathbb{N}),$$

and set (for any $\Omega \subseteq \mathbb{R}^n$ open)

$$\mathcal{E}(\Omega)=\{R:\mathcal{A}_0(\mathbb{R}^n)\times\Omega\to\mathbb{C}:x\to R(\varphi,x)\in\mathcal{C}^\infty(\Omega)\forall\varphi\in\mathcal{A}_0(\mathbb{R}^n)\},$$

$$\mathcal{E}_M(\Omega)=\{u\in\mathcal{E}(\Omega):\forall K\subset\subset\Omega\forall\alpha\in\mathbb{N}_0^n\exists p\in\mathbb{N}_0\forall\varphi\in\mathcal{A}_p(\mathbb{R}^n)$$
$$\exists c>0\exists\eta>0\sup_{x\in K}|\partial^\alpha u(\varphi_\varepsilon,x)|\leqslant c\varepsilon^{-p}(0<\varepsilon<\eta)\},$$

$$\mathcal{N}(\Omega)=\{u\in\mathcal{E}(\Omega):\forall K\subset\subset\Omega\forall\alpha\in\mathbb{N}_0^n\exists p\in\mathbb{N}_0\exists\gamma\in\Gamma\forall q\geqslant p\forall\varphi\in\mathcal{A}_q(\mathbb{R}^n)$$
$$\exists c>0\exists\eta>0\sup_{x\in K}|\partial^\alpha u(\varphi_\varepsilon,x)|\leqslant c\varepsilon^{\gamma(q)-p}(0<\varepsilon<\eta)\},$$

where $\Gamma=\{\gamma:\mathbb{N}_0\to\mathbb{R}_+ : \gamma \text{ strictly increasing}, \lim_{n\to\infty}\gamma(n)=\infty\}$. Derivation ∂^α is carried out with respect to x, while the φ are treated as parameters. Also, for $\varphi\in\mathcal{D}(\mathbb{R}^n)$, $\varphi_\varepsilon(x)=\varepsilon^{-n}\varphi(x/\varepsilon)$. Note that $\varphi_\varepsilon\to\delta$ in $\mathcal{D}'(\mathbb{R}^n)$.

Elements of $\mathcal{E}_M(\Omega)$ are called of *moderate growth*. With pointwise operations $\mathcal{E}_M(\Omega)$ is a differential algebra and $\mathcal{N}(\Omega)$ is an ideal in $\mathcal{E}_M(\Omega)$. The quotient algebra

$$\mathcal{G}(\Omega)=\mathcal{E}_M(\Omega)/\mathcal{N}(\Omega)$$

is called the *Colombeau algebra* over $\Omega\subseteq\mathbb{R}^n$. Elements of $\mathcal{G}(\Omega)$ will be denoted by $R=\mathrm{cl}[(R(\varphi,.))_{\varphi\in\mathcal{A}_0}]$ where $(R(\varphi,.))_{\varphi\in\mathcal{A}_0}$ is an arbitrary representative of R (again emphasizing the fact that the φ's are viewed as parameters).

For $\Omega=\mathbb{R}^n$ the map

$$\iota:\mathcal{E}'(\Omega)\to\mathcal{G}(\Omega),$$

$$w\to\mathrm{cl}[(w*\varphi)_{\varphi\in\mathcal{A}_0}]$$

(where $*$ denotes convolution) is a linear embedding commuting with partial derivatives and coinciding with the identical embedding $f\to\mathrm{cl}[(f)_{\varphi\in\mathcal{A}_0}]$ on $\mathcal{D}(\mathbb{R}^n)$.

Here $\mathcal{G}$ is a fine sheaf of differential algebras on $\mathbb{R}^n$ and there is a unique sheaf morphism $\hat{\iota}:\mathcal{D}'\to\mathcal{G}$ coinciding with ι on every $\mathcal{E}'(\Omega)$ and rendering $\mathcal{C}^\infty(\Omega)$ a faithful subalgebra of $\mathcal{G}(\Omega)$. From the definitions it is clear that any element of $\mathcal{G}(\Omega)$ is uniquely determined by the values of any representative on φ_ε for $\varphi\in\mathcal{A}_p$ with p arbitrarily large and ε arbitrarily small (i.e., by its ''germ''), a fact that turns out to be very helpful, e.g., in constructing solutions to differential equations in $\mathcal{G}$.

Inserting points into elements of $\mathcal{G}(\mathbb{R}^n)$ gives elements of the ring of generalized numbers $\overline{\mathbb{C}}(n)$, defined as $\overline{\mathbb{C}}(n)=\mathcal{E}(n)/\mathcal{N}(n)$, where

$$\mathcal{E}(n)=\{u:\mathcal{A}_0(\mathbb{R}^n)\to\mathbb{C}:\exists p\in\mathbb{N}_0\forall\varphi\in\mathcal{A}_p(\mathbb{R}^n)\exists c>0\exists\eta>0|u(\varphi_\varepsilon)|\leqslant c\varepsilon^{-p}(0<\varepsilon<\eta)\},$$

$$\mathcal{N}(n)=\{u:\mathcal{A}_0(\mathbb{R}^n)\to\mathbb{C}:\exists p\in\mathbb{N}_0\exists\gamma\in\Gamma\forall q\geqslant p\forall\varphi\in\mathcal{A}_q(\mathbb{R}^n)$$
$$\exists c>0\exists\eta>0|u(\varphi_\varepsilon)|\leqslant c\varepsilon^{\gamma(q)-p}(0<\varepsilon<\eta)\}.$$

Thus elements of $\mathcal{G}(\mathbb{R}^n)$ take values in $\overline{\mathbb{C}}(n)$. Explicit dependence of the ring of constants on n can be avoided by a more refined construction of the sets $\mathcal{A}_q$ in the definition of $\mathcal{G}$ (see Ref. 20). Clearly, $\mathbb{C}\hookrightarrow\overline{\mathbb{C}}$ via the canonical embedding $c\to\mathrm{cl}[(c)_{\varphi\in\mathcal{A}_0}]$.

Componentwise insertion of $R\in\mathcal{G}$ into a smooth function f yields a well-defined element $f(R)$ of $\mathcal{G}$ if f is *slowly increasing*, i.e., if all derivatives of f are polynomially bounded. Moreover, if R is *locally bounded*, i.e., if it possesses a representative such that $R(\varphi_\varepsilon,.)$ is bounded uniformly in ε on compact sets [for $\varphi\in\mathcal{A}_p(\mathbb{R}^n)$, p large], then $f\circ R$ exists for any smooth f.

Finally, we mention the notion of *association* in $\mathcal{G}(\Omega)$:$R_1, R_2 \in \mathcal{G}(\Omega)$ are called associated to each other ($R_1 \approx R_2$) if there exists some $p \in \mathbb{N}$ such that $R_1(\varphi_\varepsilon,.) - R_2(\varphi_\varepsilon,.) \to 0$ in $\mathcal{D}'(\Omega)$ as $\varepsilon \to 0$ for all $\varphi \in \mathcal{A}_p(\mathbb{R}^n)$. In particular, if $R_2 \in \mathcal{D}'(\Omega)$, then R_2 is called the macroscopic aspect (or *distributional shadow*) of R_1. Equality in $\mathcal{D}'$ is reflected as equality in the sense of association in $\mathcal{G}$, while equality in $\mathcal{G}$ is a stricter concept (for example, all powers of the Heaviside function are distinct in the Colombeau algebra although they are associated with each other).

III. EXACT SOLUTIONS OF GEODESIC AND GEODESIC DEVIATION EQUATIONS

As in Ref. 12 we consider the impulsive *pp*-wave metric

$$ds^2 = f(x^i)\,\delta(u)\,du^2 - du\,dv + (dx^i)^2, \tag{2}$$

where f is a smooth function of the transverse coordinates $x^i (i=1,2)$. Our aim is to derive solutions to the corresponding geodesic and geodesic deviation equations in the Colombeau algebra.

The general strategy for solving differential equations in $\mathcal{G}$ is to embed singularities (in our case: δ) into $\mathcal{G}$ which amounts to a regularization and then solve the corresponding regularized equations. In order to obtain general results we are therefore interested in imposing as few restrictions as possible on the regularization of δ. The largest ''reasonable'' class of smooth regularizations of δ is given by nets $(\rho_\varepsilon)_{\varepsilon \in (0,1)}$ of smooth functions ρ_ε satisfying

$$\text{(a)} \quad \text{supp}(\rho_\varepsilon) \to \{0\} \quad (\varepsilon \to 0),$$

$$\text{(b)} \quad \int \rho_\varepsilon(x)\,dx \to 1 \quad (\varepsilon \to 0), \ \text{and}$$

$$\text{(c)} \quad \exists\, \eta > 0\, \exists\, C \geqslant 0 : \int |\rho_\varepsilon(x)|\,dx \leqslant C\, \forall \varepsilon \in (0,\eta)$$

(cf. the definition of *strict delta nets* in Ref. 16, Chap. 2.7). [Note that since $\mathcal{D}$ is dense in L^1 practically even discontinuous regularizations (e.g., boxes) are included.] Obviously any such net converges to δ in distributions as $\varepsilon \to 0$. To simplify notations it is often convenient to replace (a) by

$$\text{(a}'\text{)} \quad \text{supp}(\rho_\varepsilon) \subseteq [-\varepsilon, \varepsilon] \quad \forall \varepsilon \in (0,1).$$

This motivates the following (cf. Ref. 17)

Definition 1: A generalized delta function *is an element D of* $\mathcal{G}(\mathbb{R}^n)$ *possessing a representative* $(D(\varphi,.))_{\varphi \in \mathcal{A}_0}$ *such that* $\exists\, p \in \mathbb{N}_0 \forall \varphi \in \mathcal{A}_p(\mathbb{R}^n) \exists\, \eta = \eta(\varphi) > 0$:

$$\text{(i)} \quad \text{supp}(D(\varphi_\varepsilon,.)) \subseteq [-\varepsilon, \varepsilon]\, \forall \varepsilon \in (0,\eta),$$

$$\text{(ii)} \quad \int D(\varphi_\varepsilon, x)\,dx \to 1 (\varepsilon \to 0),$$

$$\text{(iii)} \quad \exists\, C = C(\varphi) > 0 \ \text{such that} \ \int |D(\varphi_\varepsilon, x)|\,dx \leqslant C\, \forall \varepsilon \in (0,\eta).$$

The canonical embedding $R = \iota(\delta)$ of course falls into this class, but clearly there are many generalized delta functions that do not correspond to any distribution via ι. Moreover, every generalized delta function is associated to δ, i.e., all generalized delta functions equal δ on the distributional level. In a sense, they may be viewed as ''delta distributions with a more refined microstructure'' (fixing the additional nonlinear properties of the singularity).

Again, condition (i) in Definition 1 has been chosen in order to avoid technicalities in the proofs of the following results, which, however, remain true if (i) is replaced by

$$(i') \quad \mathrm{supp}(R(\varphi_\varepsilon,.))\to\{0\} \quad (\varepsilon\to 0).$$

Finally, we need the following technical preparation (which is actually a generalization of appendix A of Ref. 12).

Lemma 1: Let $g:\mathbb{R}^n\to\mathbb{R}^n$, $h:\mathbb{R}\to\mathbb{R}^n$ *smooth and let* $(\rho_\varepsilon)_{\varepsilon\in(0,1)}$ *be a net of smooth functions satisfying* (a') *and (c). For any* x_0, $\dot{x}_0\in\mathbb{R}^n$ *and any* $\varepsilon\in(0,1)$ *consider the system*

$$\begin{aligned}\ddot{x}_\varepsilon(t)&=g(x_\varepsilon(t))\rho_\varepsilon(t)+h(t),\\ x_\varepsilon(-1)&=x_0,\\ \dot{x}_\varepsilon(-1)&=\dot{x}_0.\end{aligned} \tag{3}$$

Let $b>0$, $M=\int_{-1}^{1}\int_{-1}^{s}|h(r)|drds$, $I=\{x\in\mathbb{R}^n:|x-x_0|\le b+|\dot{x}_0|+M\}$ *and* $\alpha=\min\{b/(C\|g\|_{L^\infty(I)}+|\dot{x}_0|),1/2LC,1\}$ *with L a Lipschitz constant for g on I. Then (3) has a unique solution on* $J_\varepsilon=[-1,\alpha-\varepsilon]$. *Consequently, for* ε *sufficiently small,* x_ε *is globally defined and both* x_ε *and* $\dot{x}_\varepsilon$ *are bounded, uniformly in* ε, *on compact sets.*

Proof: The operator $f\to Af$,

$$Af(t)=x_0+\dot{x}_0(t+1)+\int_{-1}^{t}\int_{-1}^{s}g(f(r))\rho_\varepsilon(r)drds+\int_{-1}^{t}\int_{-1}^{s}h(r)drds,$$

is a contraction on the complete metric space

$$\{f\in C(J_\varepsilon,\mathbb{R}^n):|f(t)-x_0|\le b+M+|\dot{x}_0|\}.$$

□

Let us now turn to the geodesic equation for the *pp*-wave metric (2). Using u as an affine parameter (which excludes trivial geodesics parallel to the shock) we obtain (cf. Ref. 12)

$$\begin{aligned}\ddot{v}(u)&=f(x^i(u))\,\dot{\delta}(u)+2\,\partial_i f(x^i(u))\dot{x}^i(u)\,\delta(u),\\ \ddot{x}^i(u)&=\tfrac{1}{2}\partial_i f(x^i(u))\,\delta(u).\end{aligned} \tag{4}$$

Since all operations appearing in (4) are well defined in $\mathcal{G}$ (cf. the remarks following Theorem 1 below), we may seek solutions of the corresponding initial value problem in the Colombeau algebra by embedding $\delta(u)$ into $\mathcal{G}$. In fact, it turns out that for *any* generalized delta function there exists a unique solution. Denoting the generalized functions corresponding to x^i and v by capital letters we state the following.

Theorem 1: *Let* $D\in\mathcal{G}(\mathbb{R})$ *be a generalized delta function,* $f\in\mathcal{C}^\infty(\mathbb{R}^2)$ *and let* v_0, $\dot{v}_0$, x_0^i, $\dot{x}_0^i\in\mathbb{R}$ $(i=1,2)$. *The initial value problem,*

$$\begin{aligned}\ddot{V}(u)&=f(X^i(u))\dot{D}(u)+2\,\partial_i f(X^i(u))\dot{X}^i(u)D(u),\\ \ddot{X}^i(u)&=\tfrac{1}{2}\partial_i f(X^i(u))D(u),\\ V(-1)&=v_0, \quad X^i(-1)=x_0^i,\\ \dot{V}(-1)&=\dot{v}_0, \quad \dot{X}^i(-1)=\dot{x}_0^i,\end{aligned} \tag{5}$$

has a unique locally bounded solution $(V,X^1,X^2)\in\mathcal{G}(\mathbb{R})^3$.

Note that we impose initial conditions in $u=-1$, i.e., ''long before'' the shock. Choosing initial conditions at $u=0$ would mean to start ''at the shock,'' which inevitably leads to regularization-dependent weak limits.

Proof: Existence: Choose $p\in\mathbb{N}$ as in Definition 1, fix $\varphi\in\mathcal{A}_p(\mathbb{R}^n)$ and let $\varepsilon<\eta(\varphi)$. Then componentwise we obtain the equations

$$\begin{aligned}
\ddot{V}(\varphi_\varepsilon,u)&=f(X^i(\varphi_\varepsilon,u))\dot{D}(\varphi_\varepsilon,u)+2\partial_i f(X^i(\varphi_\varepsilon,u))\dot{X}^i(\varphi_\varepsilon,u)D(\varphi_\varepsilon,u),\\
\ddot{X}^i(\varphi_\varepsilon,u)&=\tfrac{1}{2}\partial_i f(X^i(\varphi_\varepsilon,u))D(\varphi_\varepsilon,u),\\
V(\varphi_\varepsilon,-1)&=v_0,\quad X^i(\varphi_\varepsilon,-1)=x_0^i,\\
\dot{V}(\varphi_\varepsilon,-1)&=\dot{v}_0,\quad \dot{X}^i(\varphi_\varepsilon,-1)=x_0^i.
\end{aligned}\tag{6}$$

According to Lemma 1, the second line of (6) has a unique globally defined solution $X^i(\varphi_\varepsilon,.)$ with the specified initial values. Inserting this into the first line and integrating we also obtain a solution $V(\varphi_\varepsilon,.)$. From the boundedness properties of $X^i(\varphi_\varepsilon,.)$ established in Lemma 1 and the fact that $(D(\varphi,.))_{\varphi\in\mathcal{A}_0}\in\mathcal{E}_M(\mathbb{R})$ it follows easily by induction that $(X^i(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}$ and $(V(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}$ are moderate as well. Hence their respective classes in $\mathcal{G}(\mathbb{R})$ define solutions to (5).

Uniqueness: Suppose that $V_1=\mathrm{cl}[(V_1(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}]$ and $X_1^i=\mathrm{cl}[(X^i(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}]$ are locally bounded solutions of (5) as well. On the level of representatives this means that there exist $M=\mathrm{cl}[(M(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}]$, $N^i=\mathrm{cl}[(N^i(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}]\in\mathcal{N}(\mathbb{R})$ and n_{x^i}, $n_{\dot{x}^i}$, n_v, $n_{\dot{v}}\in\mathcal{N}(1)$ with

$$\begin{aligned}
\ddot{V}_1(\varphi_\varepsilon,u)&=f(X_1^i(\varphi_\varepsilon,u))\dot{D}(\varphi_\varepsilon,u)+2\partial_i f(X_1^i(\varphi_\varepsilon,u))\dot{X}_1^i(\varphi_\varepsilon,u)D(\varphi_\varepsilon,u)+M(\varphi_\varepsilon,u),\\
\ddot{X}_1^i(\varphi_\varepsilon,u)&=\tfrac{1}{2}\partial_i f(X_1^i(\varphi_\varepsilon,u))D(\varphi_\varepsilon,u)+N^i(\varphi_\varepsilon,u),\\
V_1(\varphi_\varepsilon,-1)&=v_0+n_v(\varphi_\varepsilon),\quad X_1^i(\varphi_\varepsilon,-1)=x_0^i+n_{x^i}(\varphi_\varepsilon),\\
\dot{V}_1(\varphi_\varepsilon,-1)&=\dot{v}_0+n_{\dot{v}}(\varphi_\varepsilon),\quad \dot{X}_1^i(\varphi_\varepsilon,-1)=\dot{x}_0^i+n_{\dot{x}^i}(\varphi_\varepsilon).
\end{aligned}\tag{7}$$

We have to show that $((V-V_1)(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}$ and $((X^i-X_1^i)(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}$ belong to the ideal $\mathcal{N}(\mathbb{R})$. Since $N^i\in\mathcal{N}(\mathbb{R})$ it follows that for p sufficiently large, ε small and $\varphi\in\mathcal{A}_p(\mathbb{R})$, $N^i(\varphi_\varepsilon,.)$ is bounded on compact sets, uniformly in ε. Thus by Lemma 1 the same holds true for $X_1^i(\varphi_\varepsilon,.)$ and its first derivative. From (7) we conclude

$$\begin{aligned}
(X^i-X_1^i)(\varphi_\varepsilon,u)=&-n_{x^i}(\varphi_\varepsilon)-(u+1)n_{\dot{x}^i}(\varphi_\varepsilon)+\frac{1}{2}\int_{-1}^{u}\int_{-1}^{s}D(\varphi_\varepsilon,r)[\partial_i f(X^i(\varphi_\varepsilon,r))\\
&-\partial_i f(X_1^i(\varphi_\varepsilon,r))]drds-\int_{-1}^{u}\int_{-1}^{s}N(\varphi_\varepsilon,r)drds.
\end{aligned}$$

Hence $\forall T>0\,\exists p\in\mathbb{N}_0\,\exists\gamma\in\Gamma\,\forall q\geqslant p\,\forall\varphi\in\mathcal{A}_q(\mathbb{R})\,\exists C>0\,\exists\eta>0\,\forall\varepsilon\in(0,\eta)\,\forall u\in[-T,T]$:

$$\begin{aligned}
|(X^i-X_1^i)(\varphi_\varepsilon,u)|\leqslant C\varepsilon^{\gamma(q)-p}+\frac{1}{2}\int_{-1}^{u}\int_{-r}^{u}\int_0^1|\nabla\partial_i f(\sigma X^i(\varphi_\varepsilon,r))\\
+(1-\sigma)X_1^i(\varphi_\varepsilon,r))|d\sigma|(X^i-X_1^i)(\varphi_\varepsilon,r)||D(\varphi_\varepsilon,r)|dsdr.
\end{aligned}\tag{8}$$

By the boundedness properties of X^i and X_1^i and by (*iii*), an application of Gronwall's lemma to the above inequality yields the $\mathcal{N}$-estimates of order 0 for $(X^i-X_1^i)$. A similar argument applies to the first derivatives. The estimates of higher order then follow inductively from the differential equation, so $((X^i - X^i_1)(\varphi_\varepsilon,.))_{\varphi\in\mathcal{A}_0}\in\mathcal{N}(\mathbb{R})$. Inserting this into the integral equation for $(V - V_1)$, the $\mathcal{N}$-estimates for $(V-V_1)$ also follow inductively. □

In the proof of Theorem 1 we have only made use of properties (*i*) and (*iii*) of the generalized delta function *D*. On the other hand, property (*ii*) will be essential for the explicit calculation of distributional limits of the unique solution constructed in Theorem 1, cf. Sec. IV. Also, note that we did not have to impose any growth restrictions on *f* to obtain a well-defined element $f(X^i)$ of $\mathcal{G}$. This is of course due to the fact that any componentwise solution of the initial value problem necessarily is bounded, uniformly in ε, on compact sets (for ε small). Our next goal is an analysis of the Jacobi equation for impulsive *pp*-waves in the framework of algebras of generalized functions. As in Ref. 12 to keep formulas more transparent we make some simplifying assumptions concerning geometry (namely axisymmetry) and initial conditions. Writing $x=x^1$ and $y=x^2$ we suppose that *f* depends exclusively on the two-radius $\sqrt{x^2+y^2}$ and work within the hypersurface $y=0$ (corresponding to initial conditions $y_0=0=\dot{y}_0$). Furthermore, we demand $v_0=0=\dot{x}_0$. As was shown in Ref. 12, in this situation the Jacobi equation

$$\frac{D^2N^\alpha}{dt^2}=-R^a_{bcd}T^bT^dN^c,$$

where $N^a(u)=(N^u(u),N^v(u),N^x(u),N^y(u))$ denotes the deviation vector field, takes the form

$$\ddot{N}^v=2[N^xf'(x)\delta]^{\cdot}-N^xf'(x)\dot{\delta}+[N^uf(x)\delta]^{\cdot\cdot}-N^uf''(x)\dot{x}^2\delta-N^uf'(x)\ddot{x}\delta,$$

$$\ddot{N}^x=[\dot{N}^uf'(x)+\tfrac{1}{2}N^xf''(x)]\delta+\tfrac{1}{2}f'(x)N^u\dot{\delta}, \tag{9}$$

$$\ddot{N}^y=\ddot{N}^u=0,$$

where *x* is determined by (4). Existence and uniqueness of solutions to the corresponding initial value problem in the Colombeau algebra is established in the following result where, for the sake of brevity, we denote the $\mathcal{G}$-functions corresponding to N^a again by N^a.

Theorem 2: *Let $D\in\mathcal{G}(\mathbb{R})$ be a generalized delta function, $f\in\mathcal{C}^\infty(\mathbb{R})$, n^a, $\dot{n}^a\in\mathbb{R}^4$, and let X denote the (unique) solution to system (5) with initial conditions and simplifications as discussed above. The initial value problem*

$$\ddot{N}^v=2[N^xf'(X)D]^{\cdot}-N^xf'(X)\dot{D}+[N^uf(X)D]^{\cdot\cdot}-N^uf''(X)\dot{X}^2D-N^uf'(X)\ddot{X}D,$$

$$\ddot{N}^x=[\dot{N}^uf'(X)+\tfrac{1}{2}N^xf''(X)]D+\tfrac{1}{2}f'(X)N^u\dot{D},$$

$$\ddot{N}^y=\ddot{N}^u=0, \tag{10}$$

$$N^a(-1)=n^a,\quad \dot{N}^a(-1)=\dot{n}^a.$$

has a unique solution $N^a\in\mathcal{G}(\mathbb{R})^4$.

Proof: Since the equations are linear in the components of the deviation field we are provided with globally defined solutions on the level of representatives. The last two equations are actually trivial and so is the first one once we know that its right-hand side belongs to $\mathcal{G}(\mathbb{R})$. Hence we are left with the equation for N^x which is of the form $\ddot{N}(t)=f''(X(t))D(t)N(t)+H(t)$ with *H* in $\mathcal{G}(\mathbb{R})$. Using the boundedness properties of *X* established in Lemma 1 the $\mathcal{E}_M$-bounds for N^x easily follow from Gronwall's lemma.

Uniqueness is established along the same lines again using Gronwall-type arguments. □

In the above proof we have again only used properties (*i*) and (*iii*) of the generalized delta function *D*.

To conclude this section we remark that unique solvability of the geodesic and geodesic deviation equation for (2) is not confined to the case where the profile function *f* is smooth. Indeed, it turns out that for a large class of generalized profile functions (those that are not ''too singular'') Theorems 1 and 2 retain their validity. More precisely, we have to demand that f belongs to the algebra of *tempered* generalized functions[15] to make sure that the composition $f(X)$ is well defined and that $\nabla\nabla f$ is of L^∞-log-type[17,18] to ensure existence and uniqueness of solutions to (5) and (10). However, to include many physically interesting examples (cf. Ref. 21), one has to cut out the worldline of the ultrarelativistic particle, i.e., the v-axis from the domain of definition (cf. Ref. 22).

IV. DISTRIBUTIONAL LIMITS

In this section we are going to calculate the distributional limits (or, in the terminology of Colombeau theory, the associated distributions) of the unique solutions to the geodesic and geodesic deviation equation constructed in Theorems 1 and 2. In Ref. 12 distributional limits for regularized versions of these equations have been calculated using a model delta net regularization. Translated into our current setting this amounts to using the particular generalized delta function $D=\iota(\delta)$. Our aim is to extend the validity of the limit relations derived there to the case of solutions in the Colombeau algebra and to generalized delta functions. At the same time we will be able to prove stronger convergence results in some cases.

Theorem 3: *The unique solution* (V,X^i) *of the geodesic equation (5) satisfies the following association relations:*

$$X^i\approx x_0^i+\dot{x}_0^i(1+u)+\tfrac{1}{2}\partial_i f(x_0^i+\dot{x}_0^i)u_+\,, \tag{11}$$

$$V\approx v_0+\dot{v}_0(1+u)+f(x_0^i+\dot{x}_0^i)\,\theta(u)+\partial_i f(x_0^i+\dot{x}_0^i)(\dot{x}_0^i+\tfrac{1}{4}\partial^i f(x_0^i+\dot{x}_0^i))u_+\,. \tag{12}$$

In addition, if $X^i=\mathrm{cl}[(X^i(\varphi,.))_{\varphi\in\mathcal{A}_0}]$, *then* $\exists p\in\mathbb{N}_0$ *such that* $\forall\varphi\in\mathcal{A}_p$

$$X^i(\varphi_\varepsilon,u)\rightarrow x_0^i+\dot{x}_0^i(1+u)+\tfrac{1}{2}\partial_i f(x_0^i+\dot{x}_0^i)u_+ \tag{13}$$

for $\varepsilon\rightarrow 0$, *uniformly on compact subsets of* $\mathbb{R}$.

Proof: Choose $p\in\mathbb{N}_0$ as in Definition 1 for D. Let $\varphi\in\mathcal{A}_p$ and $\varepsilon<\eta(\varphi)$. Since integrating amounts to convolution with the Heaviside function, which is a continuous operation on the convolution algebra of distributions supported in a cone, in order to prove (11) it suffices to show that

$$\ddot{X}^i(\varphi_\varepsilon,.)=\tfrac{1}{2}\partial_i f(X^i(\varphi_\varepsilon,.))D(\varphi_\varepsilon,.)\rightarrow\tfrac{1}{2}\partial_i f(x_0^i+\dot{x}_0^i)\,\delta$$

in distributions. We first note that $X(\varphi_\varepsilon,\varepsilon t)\rightarrow x_0^i+\dot{x}_0^i$ uniformly as can be seen from the integral equation for X^i [cf. (11) in Ref. 12]. Now if $\psi\in\mathcal{D}(\mathbb{R})$, then

$$\left|\int_{-\epsilon}^{\epsilon}\psi(t)\partial_i f(X^i(\varphi_\varepsilon,t))D(\varphi_\varepsilon,t)dt-\partial_i f(x_0^i+\dot{x}_0^i)\psi(0)\right|$$

$$\leqslant\sup_{-\epsilon\leqslant t\leqslant\epsilon}|\psi(t)\partial_i f(X^i(\varphi_\epsilon,t))-\partial_i f(x_0^i+\dot{x}_0^i)\psi(0)|\int_{-\epsilon}^{\epsilon}|D(\varphi_\epsilon,t)|dt$$

$$+\int_{-\epsilon}^{\epsilon}|D(\varphi_\epsilon,t)dt-1|\partial_i f(x_0^i+\dot{x}_0^i)\psi(0).$$

So the claim follows from properties (*iii*) and (*ii*) of the generalized delta function D. Since $\dot{X}^i(\varphi_\varepsilon,t)$ is bounded on compact sets, uniformly in ε, it follows that the family $\{X^i(\varphi_\varepsilon,t):\varepsilon\in(0,1)\}$ is locally equicontinuous. Hence Ascoli's Theorem implies (13). Concerning (12), as above it suffices to calculate the limit of

$$\ddot{V}(\varphi_\varepsilon,u)=[f(X^i(\varphi_\varepsilon,u))D(\varphi_\varepsilon,u)]^{\cdot}+\partial_i f(X^i(\varphi_\varepsilon,u))\dot{X}^i(\varphi_\varepsilon,u)D(\varphi_\varepsilon,u),$$

whose first summand converges to $f(x_0^i+\dot{x}_0^i)\dot{\delta}$ by an argument similar to the one above. For the second summand we have

$$\partial_i f(X^i(\varphi_\varepsilon,u))\dot{X}^i(\varphi_\varepsilon,u)D(\varphi_\varepsilon,u)=\underbrace{\partial_i f(X^i(\varphi_\varepsilon,u))D(\varphi_\varepsilon u)\dot{x}_0^i}_{(*)}$$
$$+\frac{1}{2}\partial_i f(X^i(\varphi_\varepsilon,u))D(\varphi_\varepsilon,u)\int_{-\varepsilon}^{t}\partial_i f(X^i(\varphi_\varepsilon,s))D(\varphi_\varepsilon,s)ds$$

and $(*)\to\partial_i f(x_0^i+\dot{x}_0^i)\dot{x}_0^i\delta.$ Finally, since

$$\int_{-\varepsilon}^{\varepsilon}\psi(t)\partial_i f(X^i(\varphi_\varepsilon,t))D(\varphi_\varepsilon,t)\int_{-\varepsilon}^{t}\partial_i f(X^i(\varphi_\varepsilon,s))D(\varphi_\varepsilon,s)ds\,dt$$
$$-\frac{1}{2}\partial_i f(x_0^i+\dot{x}_0^i)^2\psi(0)\int_{-\varepsilon}^{\varepsilon}D(\varphi_\varepsilon,t)dt\to 0,$$

the claim follows. □

In calculating distributional limits for the solution of the Jacobi equation to maintain the clarity of formulas, we shall make simplifying assumptions on the initial conditions, i.e.,

$$N^a(-1)=(0,0,0,0),$$
$$\dot{N}^a(-1)=(a,b,0,0). \tag{14}$$

Then we have the following.

Theorem 4: *The unique solution of the geodesic deviation equation (10) satisfies the following association relations:*

$$N^x\approx\tfrac{1}{2}af'(x_0)(u_+ +\theta(u)), \tag{15}$$

$$N^v\approx b(1+u)+a[f(x_0)\delta(u)+\tfrac{1}{4}f'(x_0)^2(\theta(u)+u_+)]. \tag{16}$$

Proof: The general structure of this proof is ''isomorphic'' to the calculation of distributional limits for the regularized Jacobi equation in Ref. 12. The main difference is that, for representatives of generalized delta functions, dominated convergence arguments are not applicable, which makes the calculations more tedious. Nevertheless, using the uniform convergence of $X^i(\varphi_\varepsilon,.)$ established above, all steps carried out in Ref. 12 can be adapted to the present situation as demonstrated in the proof of Theorem 3. □

V. DISCUSSION AND OUTLOOK

In the previous section we have shown that the unique solutions to the geodesic and geodesic deviation equation in the Colombeau algebra possess a physically reasonable macroscopic (i.e., distributional) aspect: even within the natural maximal class of delta-regularizations (namely the class of all generalized delta functions) the regularity of the equations is sufficiently high to ensure distributional limits corresponding to physical expectations. More precisely, from the distributional point of view, the geodesics correspond to refracted, broken straight lines as suggested by the form of the metric. The scale of the jump and kink is given by the values of f and its first derivatives at the shock hypersurface, which can be traced back to the values at the initial point $(u=-1)$, thereby precisely reproducing Penrose's junction conditions.[1] The distributional limit of the Jacobi field suffers a kink and jump in the x-direction as well as an additional δ-pulse in the v-direction, which may be understood from the form of the geodesics. For a more detailed discussion, see Ref. 12.

Finally we make some comments on diffeomorphism invariance of our results. Whereas the fine sheaf of Colombeau algebras can be lifted to manifolds in a straightforward manner, the action of a diffeomorphism does not commute with the canonical embedding $\mathcal{D}' \hookrightarrow \mathcal{G}$. The reason for this is that convolution relies on the additive group structure of $\mathbb{R}^n$ and is therefore not invariant under the action of diffeomorphisms. Note, however, that our calculations did not use the embedding via convolution and therefore are not affected by this defect.

A solution to the above mentioned problem was proposed in Ref. 23 using a modified definition of the mollifier spaces $\mathcal{A}_q$. A key ingredient of this construction is that diffeomorphisms act on the φ's, introducing an implicit x dependence into the first slot of the Colombeau functions $R(\varphi,x)$. Hence, to retain smooth x dependence of R the concept of Silva differentiability was used. Future work will be concerned with a simplified concept of Colombeau algebras on manifolds using calculus in convenient vector spaces.[24] A main goal of this line of research is to provide a workable solution concept for singular differential equations on manifolds.

ACKNOWLEDGMENTS

The authors wish to thank M. Grosser, M. Oberguggenberger and H. Urbantke for helpful discussions. M. Kunzinger was supported by Research Grant No. P10472-MAT of the Austrian Science Foundation. R. Steinbauer was supported by Austrian Academy of Science, Ph.D. programme, Grant No. 338 and by Research Grant No. P12023-MAT of the Austrian Science Foundation.

[1] R. Penrose, ''The geometry of impulsive gravitational waves,'' in *General Relativity, Papers in Honour of J. L. Synge*, edited by L. O'Raifeartaigh (Clarendon, Oxford, 1972), pp. 101–115.
[2] P. C. Aichelburg and R. U. Sexl, ''On the gravitational field of a massless particle,'' Gen. Relativ. Gravit. **2**, 303–312 (1971).
[3] H. Balasin and H. Nachbagauer, ''The ultrarelativistic Kerr-geometry and its energy-momentum tensor,'' Class. Quantum Grav. **12**, 707–713 (1995).
[4] C. O. Loustó and N. Sánchez, ''The ultrarelativistic limit of the Kerr-Newman geometry and particle scattering at the Planck scale,'' Phys. Lett. B **232**, 462–466 (1989).
[5] J. Podolosky and J. B. Griffiths, ''Boosted static multipole particles as sources of impulsive gravitational waves,'' to appear in Phys. Rev. D, gr-qc/9809003 (1998).
[6] E. Verlinde and H. Verlinde, ''High-energy scattering in Quantum Gravity,'' Class. Quantum Grav. **10**, 175–184 (1993).
[7] T. Dray and G. 't Hooft, ''The gravitational shockwave of a massless particle,'' Nucl. Phys. B **232**, 173–188 (1985).
[8] P. C. Aichelburg and H. Balasin, ''Generalized symmetries of impulsive gravitational waves,'' Class. Quantum Grav. **14**, A31–A41 (1997).
[9] H. Balasin, ''Geodesics for impulsive gravitational waves and the multiplication of distributions,'' Class. Quantum Grav. **14**, 455–462 (1997).
[10] V. Ferrari, P. Pendenza, and G. Veneziano, ''Beam like gravitational waves and their geodesics,'' Gen. Relativ. Gravit. **20**, 1185–1191 (1988).
[11] J. Podolsky and K. Vesely, ''Continuous coordinates for all impulsive *pp*-waves,'' Phys. Lett. A **241**, 145–147 (1998).
[12] R. Steinbauer, ''Geodesics and geodesic deviation for impulsive gravitational waves,'' J. Math. Phys. **39**, 2201–2212 (1998).
[13] O. Hájek, Bull. Am. Math. Soc. **12**, 272–279 (1985).
[14] J. F. Colombeau, *New Generalized Functions and Multiplication of Distributions* (North Holland, Amsterdam, 1984).
[15] J. F. Colombeau, *Elementary Introduction to New Generalized Functions* (North Holland, Amsterdam, 1985).
[16] M. Oberguggenberger, *Multiplication of Distributions and Applications to Partial Differential Equations* (Longman Scientific and Technical, New York, 1992).
[17] R. Hermann and M. Oberguggenberger, ''Generalized functions, calculus of variations and nonlinear ODEs,'' to appear in *Nonlinear Theory of Generalized Functions*, edited by M. Grosser, G. Hörmann, M. Kunzinger, and M. Oberguggenberger (Pitman Research Notes in Mathematics, 1999).
[18] M. Kunzinger, ''Lie Transformation Groups in Colombeau Algebras,'' Ph.D.-thesis, University of Vienna, 1996.
[19] J. Podolsky and K. Vesely, ''New examples of sandwich gravitational waves and their impulsive limit,'' Czech. J. Phys. **48**, 871 (1998).
[20] J. Aragona and H. A. Biagioni, ''Intrinsic definition of the Colombeau algebra of generalized functions,'' Analysis Mathematica **17**, 75–132 (1991).
[21] C. O. Loustó and N. Sánchez, ''Gravitational shock waves generated by extended sources: Ultrarelativistic cosmic strings, monopoles and domain walls,'' Nucl. Phys. B **355**, 231–249 (1991).

[22] R. Steinbauer, ''The ultrarelativistic Reissner-Nordstrøm field in the Colombeau algebra,'' J. Math. Phys. **38**, 1614–1622 (1997).
[23] J. F. Colombeau and A. Meril, ''Generalized functions and multiplication of distributions on manifolds,'' J. Math. Anal. Appl. **186**, 357–364 (1994).
[24] A. Kriegl and P. W. Michor, *The Convenient Setting of Global Analysis* (Mathematical Surveys and Monographs Vol. 53 (AMS, Providence, RI, 1997).

Static Bondi energy in the teleparallel equivalent of general relativity

J. W. Maluf[a)] and J. F. da Rocha-Neto
Departamento de Física, Universidade de Brasília, C.P. 04385, 70.919-970 Brasília, DF, Brazil

(Received 7 October 1997; accepted for publication 10 November 1998)

We consider Bondi's radiating metric in the context of the teleparallel equivalent of general relativity (TEGR). This metric describes the asymptotic form of a radiating solution of Einstein's equations. The total gravitational energy for this solution can be calculated by means of pseudotensors in the static case. In the nonstatic case, Bondi defines the *mass aspect* $m(u)$, which describes the mass of an isolated system. In this paper we express Bondi's solution in asymptotically spherical 3 $+1$ coordinates, not in radiation coordinates, and obtain Bondi's energy in the static limit by means of the expression for the gravitational energy in the framework of the TEGR. We can either obtain the total energy or the energy inside a large (but finite) portion of a three-dimensional spacelike hypersurface, whose boundary is far from the source. © *1999 American Institute of Physics.*
[S0022-2488(99)03602-6]

I. INTRODUCTION

The concept of energy in general relativity is considerably more intricate than in any other branch of physics. Any physical phenomena, except gravitation, is defined and described on a specific space–time, which is usually the flat space–time. For these phenomena the concept of energy can be intuitively conceived and mathematically realized. Generically, energy is an attribute of some physical system whose dynamics takes place on the space–time. Gravitation, however, acquires a distinct status because the dynamics of the gravitational field is the dynamics of the space–time itself. Consequently, the definition of the gravitational energy is not straightforward.

The several attempts at defining the gravitational energy (pseudotensors, quasilocal energy, actions, and Hamiltonians with surface terms) all agree in predicting the *total* energy of asymptotically flat gravitational fields. Moreover, there seems to exist a predominant point of view according to which the gravitational energy is not localizable, i.e., that there does not exist a gravitational energy *density*. These are probably the only two features shared by the various approaches, which are mostly based on the metric tensor. However, the very concept of a black hole lends support to the idea that gravitational energy is localizable. There is no process by means of which the gravitational mass inside a black hole can be made to vanish.

A detailed analysis of the structure of the pseudotensors shows that the (covariant) gravitational energy–momentum tensor would have to be defined by means of the first derivative of the metric tensor. But it is well known that it is not possible to write down a nontrivial covariant expression involving the first derivative of the metric, which captures the energy content of the field. However, it is possible to write down such covariant expressions with tetrads, and Møller noticed this fact long ago.[1–3]

The question of localizability of the gravitational energy can be discussed in the framework of the teleparallel equivalent of general relativity (TEGR),[3–6] which is an alternative geometrical formulation of Einstein's general relativity. The action integral of the TEGR is constructed en-

[a)]Electronic mail: wadih@fis.unb.br

tirely out of the torsion tensor. The analysis of the canonical structure of the TEGR[7] indicates the existence of a perfectly well-defined gravitational energy density. Such existence is possible, in principle, because the TEGR is defined in terms of tetrad fields. The torsion tensor allows the construction of a total divergence that transforms as a scalar (energy) density. In the $3+1$ formulation of the TEGR, the integral form of the Hamiltonian constraint equation $C=0$ can be written as an energy equation of the type[8]

$$C=H-E=0,$$

where E is the gravitational energy defined by

$$E_g=\frac{1}{8\pi G}\int_V d^3x\;\partial_i(eT^i), \tag{1}$$

where $e=\det(e_{(k)i})$, $\{e_{(k)i}\}$ are triads restricted to a three-dimensional spacelike hypersurface Σ, and T^i is the trace of the torsion tensor: $T^i=g^{ik}T_k=g^{ik}e^{(l)j}T_{(l)jk}$, $T_{(l)jk}=\partial_j e_{(l)k}-\partial_k e_{(l)j}$. V is an arbitrary three-dimensional volume of integration and G is the gravitational constant. This expression is simple and powerful. It has been successfully applied to rotating black holes,[9] de Sitter space,[10] and conical space–times.[11] The definition of gravitational energy in the TEGR may not be intuitively clear, but it is supported by its mathematical simplicity and by the applications to the space–times listed above. The use of (1) requires only the construction of the triads $\{e_{(k)i}\}$ with the appropriate boundary conditions, and that transform under the *global* SO(3) group.

The torsion tensor that appears in the Hamiltonian formulation of the TEGR is related to the antisymmetric component of the connection $\Gamma^i_{jk}=e^{(m)i}\,\partial_j e_{(m)k}$, whose curvature tensor is identically vanishing. Such a connection defines a space with teleparallelism, or absolute parallelism, or else *fernparallelismus*, according to Schouten.[12]

In this paper we investigate the energy of asymptotically flat gravitational waves, described by Bondi's radiating metric.[13] Since the metric describes an isolated system, the application of (1) is possible as it stands, provided we consider the metric in the $3+1$ spherical coordinates (t,r,θ,ϕ) at spacelike infinity, for which $t=$const defines a spacelike hypersurface. We note that the use of Cartesian (rectangular) coordinates in the asymptotic limit is necessary for the evaluation of pseudotensors out of this metric.

We recall that the Arnowitt–Deser–Misner (ADM) energy[14] is not suitable for the analysis of gravitational radiation, because it gives the *total* energy of the space–time, both from the source and from the emitted radiation, whereas the Bondi energy evaluated at null infinity furnishes only the energy of the source, from which it is possible to derive the well-known formula for the loss of mass.

The relevance of the definition (1) resides precisely in the fact that we can evaluate it on a large but *finite* volume V of the three-dimensional spacelike hypersurface, thereby not including the emitted radiation outside V. In view of the field equations (which are not considered here), the energy inside V turns out to be a decreasing function of time.

It is important to remark at this point that Bondi energy has been calculated in several geometrical frameworks, by different approaches.[15–19] A common feature of these approaches is that they yield the *total* energy of the field. In contrast, we will consider finite volumes of spacelike surfaces and obtain the energy contained within large spherical surfaces of radius r_0 up to the $1/r_0$ term.

In the next section we briefly describe the Lagrangian and Hamiltonian formulations of the TEGR. In Sec. III we compare our energy expression with Møller's expression. We show that both expressions agree for the *total* gravitational energy, but in spite of similarities they disagree when applied to finite volumes of the three-dimensional space. In Sec. IV we write Bondi's metric in (t,r,θ,ϕ) coordinates at infinity and proceed to carry out the construction of triads for the spacelike hypersurfaces Σ. There exists an infinite number of triads that lead to the metric restricted to the three-dimensional hypersurface. However, only two of them will be considered in detail. In

Sec. V we calculate both the total energy of the field and the energy contained within a large sphere of radius r_0. The total energy obtained by means of (1), in which case the integration is made over the whole Σ, agrees with the known result for the Bondi energy in the *static* case. We also obtain the expression for the energy contained within a surface of constant radius r_0 in the asymptotic region, where the metric coefficients may be determined.

Notation: spacetime indices $\mu,\nu,...$, and local Lorentz indices $a,b,...$, run from 0 to 3. In the $3+1$ decomposition Latin indices from the middle of the alphabet indicate space indices according to $\mu=0,i$, $a=(0),(i)$. The tetrad field $e^a\,_\mu$ and the spin connection $\omega_{\mu ab}$ yield the usual definitions of the torsion and curvature tensors: $R^a\,_{b\mu\nu}=\partial_\mu\omega_\nu\,^a\,_b+\omega_\mu\,^a\,_c\omega_\nu\,^c\,_b-\cdots$, $T^a\,_{\mu\nu}=\partial_\mu e^a\,_\nu+\omega_\mu\,^a\,_b e^b\,_\nu-\cdots$. The flat space–time metric is fixed by $\eta_{(0)(0)}=-1$.

II. THE TEGR

The Lagrangian density of the TEGR in empty spacetime is given by

$$L(e,\omega,\lambda)=-ke(\tfrac{1}{4}T^{abc}T_{abc}+\tfrac{1}{2}T^{abc}T_{bac}-T^aT_a)+e\lambda^{ab\mu\nu}R_{ab\mu\nu}(\omega), \tag{2}$$

where $k=1/16\pi G$, G is the gravitational constant; $e=\det(e^a\,_\mu)$, $\lambda^{ab\mu\nu}$ are Lagrange multipliers, and T_a is the trace of the torsion tensor defined by $T_a=T^b\,_{ba}$. The tetrad field $e_{a\mu}$ and the spin connection $\omega_{\mu ab}$ are completely independent field variables. The latter is enforced to satisfy the condition of zero curvature. Therefore this Lagrangian formulation is in no way similar to the usual Palatini formulation, in which the spin connection is related to the tetrad field via field equations. Later on, we will introduce the tensor Σ_{abc}, defined by

$$\tfrac{1}{4}T^{abc}T_{abc}+\tfrac{1}{2}T^{abc}T_{bac}-T^aT_a\equiv T^{abc}\Sigma_{abc}.$$

The equivalence of the TEGR with Einstein's general relativity is based on the identity

$$eR(e,\omega)=eR(e)+e(\tfrac{1}{4}T^{abc}T_{abc}+\tfrac{1}{2}T^{abc}T_{bac}-T^aT_a)-2\partial_\mu(eT^\mu), \tag{3}$$

which is obtained by just substituting the arbitrary spin connection $\omega_{\mu ab}={}^0\omega_{\mu ab}(e)+K_{\mu ab}$ in the scalar curvature tensor $R(e,\omega)$ on the left-hand side; ${}^0\omega_{\mu ab}(e)$ is the Levi-Civita connection and $K_{\mu ab}=\tfrac{1}{2}e_a\,^\lambda e_b\,^\nu(T_{\lambda\mu\nu}+T_{\nu\lambda\mu}-T_{\mu\nu\lambda})$ is the contorsion tensor. The vanishing of $R^a\,_{b\mu\nu}(\omega)$, which is one of the field equations derived from (2), implies the equivalence of the scalar curvature $R(e)$, constructed out of $e^a\,_\mu$ only, and the quadratic combination of the torsion tensor. It also ensures that the field equation arising from the variation of L with respect to $e^a\,_\mu$ is strictly equivalent to Einstein's equations in tetrad form. Let $\delta L/\delta e^{a\mu}=0$ denote the field equations satisfied by $e^{a\mu}$. It can be shown by explicit calculations that

$$\frac{\delta L}{\delta e^{a\mu}}=\frac{1}{2}e\left\{R_{a\mu}(e)-\frac{1}{2}e_{a\mu}R(e)\right\}. \tag{4}$$

We refer the reader to Refs. 7 and 8 for additional details.

Throughout this section we will be interested in asymptotically flat space–times. The Hamiltonian formulation of the TEGR can be successfully implemented if we fix the gauge $\omega_{0ab}=0$ from the outset, since in this case the constraints (to be shown below) constitute a *first class* set.[7] The condition $\omega_{0ab}=0$ is achieved by breaking the local Lorentz symmetry of (2). We still make use of the residual time-dependent gauge symmetry to fix the usual time gauge condition $e_{(k)}\,^0=e_{(0)i}=0$. Because of $\omega_{0ab}=0$, H does not depend on P^{kab}, the momentum canonically conjugated to ω_{kab}. Therefore arbitrary variations of $L=p\dot{q}-H$ with respect to P^{kab} yields $\dot{\omega}_{kab}=0$. Thus, in view of $\omega_{0ab}=0$, ω_{kab} drops out from our considerations. The above gauge fixing can be understood as the fixation of a reference frame.

As a consequence of the above gauge fixing, the canonical action integral obtained from (2) becomes[8]

$$A_{TL}=\int d^4x\{\Pi^{(j)k}\dot{e}_{(j)k}-H\}, \tag{5}$$

$$H=NC+N^iC_i+\Sigma_{mn}\Pi^{mn}+\frac{1}{8\pi G}\partial_k(NeT^k)+\partial_k(\Pi^{jk}N_j). \tag{6}$$

N and N^i are the lapse and shift functions, and $\Sigma_{mn}=-\Sigma_{nm}$ are Lagrange multipliers. The constraints are defined by

$$C=\partial_j(2keT^j)-ke\Sigma^{kij}T_{kij}-\frac{1}{4ke}\left(\Pi^{ij}\Pi_{ji}-\frac{1}{2}\Pi^2\right), \tag{7a}$$

$$C_k=-e_{(j)k}\,\partial_i\Pi^{(j)i}-\Pi^{(j)i}T_{(j)ik}, \tag{7b}$$

with $e=\det(e_{(j)k})$ and $T^i=g^{ik}e^{(j)l}T_{(j)lk}$. We remark that (5) and (6) are invariant under global SO(3) and general coordinate transformations.

If we assume the asymptotic behavior,

$$e_{(j)k}\approx\eta_{jk}+\tfrac{1}{2}h_{jk}(\tfrac{1}{r}), \tag{8}$$

for $r\rightarrow\infty$, then, in view of the relation

$$\frac{1}{8\pi G}\int d^3x\,\partial_j(eT^j)=\frac{1}{16\pi G}\int_S dS_k(\partial_i h_{ik}-\partial_k h_{ii})\equiv E_{ADM}, \tag{9}$$

where the surface integral is evaluated for $r\rightarrow\infty$, the integral form of the Hamiltonian constraint $C=0$ may be rewritten as

$$\int d^3x\left\{ke\Sigma^{kij}T_{kij}+\frac{1}{4ke}\left(\Pi^{ij}\Pi_{ji}-\frac{1}{2}\Pi^2\right)\right\}=E_{ADM}. \tag{10}$$

The integration is over the whole three-dimensional space. Given that $\partial_j(eT^j)$ is a scalar density, from (9) and (10) we define the gravitational energy density enclosed by a volume V of the space as

$$E=\frac{1}{8\pi G}\int_V d^3x\,\partial_j(eT^j). \tag{1}$$

It must be noted that E depends only on the triads $e_{(k)i}$ restricted to a three-dimensional spacelike hypersurface; the inverse quantities $e^{(k)i}$ can be written in terms of $e_{(k)i}$. From the identity (4) we observe that the dynamics of the triads does not depend on $\omega_{\mu ab}$. Therefore, E_g given above does not depend on the fixation of any gauge for $\omega_{\mu ab}$. We briefly remark that the reference space that defines the zero of energy has been discussed in Ref. 9.

We make now the important assumption that the general form of the canonical structure of the TEGR is the same for any class of space–times, irrespective of the peculiarities of the latter (for the de Sitter space,[10] for example, there is an *additional* term in the Hamiltonian constraint C). Therefore we assert that the integral form of the Hamiltonian constraint equation can be written as $C=H-E=0$ for *any* space–time, and that (1) represents the gravitational energy for arbitrary space–times with any topology.

Before closing this section, let us recall that Müller–Hoissen and Nitsch[20] and Kopczyński[21] have shown that, in general, the theory defined by (2) faces difficulties with respect to the Cauchy problem. They have shown that, in general, six components of the torsion tensor are not determined from the evolution of the initial data. On the other hand, the constraints of the theory constitute a first class set provided we fix the six quantities $\omega_{0ab}=0$ *before varying the action*.[7] This condition is mandatory and does not merely represent one particular gauge fixing of the theory. Since the fixing of ω_{0ab} yields a well-defined theory with first class constraints, we cannot assert that the field configurations of the latter are gauge equivalent to configurations whose time evolution is not precisely determined. The requirement of local SO(3,1) symmetry plus the addition of $\lambda^{ab\mu\nu}R_{ab\mu\nu}(\omega)$ in (2) has the ultimate effect of discarding the connection. Although we have no proof, we believe that the two properties above (the failure of the Cauchy problem and the fixation of $\omega_{0ab}=0$) are related to each other.

Constant rotations constitute a basic feature of the teleparallel geometry. According to Møller,[2] in the framework of the absolute parallelism tetrad fields, together with the boundary conditions, uniquely determine a *tetrad lattice*, apart from an arbitrary *constant rotation of the tetrads in the lattice*.

III. MØLLER'S ENERGY EXPRESSION

Møller carried out several investigations regarding the localizability of the gravitational energy. He faced difficulties in establishing a covariant expression using the metric tensor,[2] and because of this he arrived at an expression through the use of tetrads.[2,3] According to Møller, this latter expression still has a difficulty in that it is not invariant under *local* Lorentz transformations. It is very instructive to compare expression (1) with Møller's expression. For the sake of this comparison, we will put aside the difficulty regarding the noninvariance with respect to local Lorentz transformations.

Møller presents an expression for the energy–momentum of the gravitational field. However, we will only consider the energy expression. Translating into our notation, Møller's energy reads as

$$E=-\int d^3x\, \partial_\lambda U_0{}^{0\lambda}, \tag{11}$$

where the potential in the integrand is given by

$$U_\mu{}^{\nu\lambda}=\frac{1}{8\pi G}e[e^{a\nu}\nabla_\mu e_a{}^\lambda+(\delta^\nu_\mu e^{a\lambda}-\delta^\lambda_\mu e^{a\nu})\nabla_\sigma e_a{}^\sigma]. \tag{12}$$

In contrast with the notation of the previous section, all geometrical quantities in Eqs. (11)–(15) are *four-dimensional* quantities. In (12), ∇ represents the covariant derivative with respect to the Christoffel symbols $\Gamma^\lambda_{\mu\nu}$.

Møller's energy can be first rewritten as

$$E=\frac{1}{8\pi G}\int d^3x\, \partial_i(ee^{ak}\nabla_k e_a{}^i). \tag{13}$$

By means of the identity

$$\nabla_k e_{aj}=\partial_k e_{aj}-\Gamma^\sigma_{kj}e_{a\sigma}\equiv -{}^0\omega_k{}^b{}_a e_{bj}, \tag{14}$$

where ${}^0\omega_{\mu ab}$ is the Levi-Civita connection,

$${}^0\omega_{\mu ab}=-\tfrac{1}{2}e^c{}_\mu(\Omega_{abc}-\Omega_{bac}-\Omega_{cab}),$$

$$\Omega_{abc}=e_{a\nu}(e_b{}^\mu\,\partial_\mu e_c{}^\nu-e_c{}^\mu\,\partial_\mu e_b{}^\nu),$$

we can further rewrite expression (13) as

$$E=\frac{1}{8\pi G}\int d^3x\,\partial_i(ee^{ai}e^{bj}\,{}^0\omega_{jab}). \tag{15}$$

Up to this point $\{e_{a\mu}\}$ are tetrads of the four-dimensional space–time. In order to compare (15) with (1) let us impose the time gauge $e^{(0)}{}_k=e_{(j)}{}^0=0$ and establish the $3+1$ decomposition of the tetrads as in Refs. 7 and 22. Then the integrand on the right-hand side of (15) can be rewritten as

$${}^4e\,{}^4e^{ai}\,{}^4e^{bj}\,{}^0\omega_{jab}({}^4e)=Nee^{(m)i}e^{(n)j}\,{}^0\omega_{j(m)(n)}(e)-e(e^{ai}N^j-e^{aj}N^i)^0\omega_{j(0)(m)}.$$

N and N^i are the lapse and shift functions and $\{{}^4e^{a\mu}\}$ are tetrads of the four-dimensional space–time. In terms of triads restricted to a three-dimensional spacelike surface we have

$$E=\frac{1}{8\pi G}\int d^3x\,\partial_i[Nee^{(m)i}e^{(n)j}\,{}^0\omega_{j(m)(n)}-e(e^{(m)i}N^j-e^{(m)j}N^i)^0\omega_{j(0)(m)}]. \tag{16a}$$

A comparison with (1) can now be made if we make use of the *identity*[23]

$$\partial_i(ee^{(m)i}e^{(n)j}\,{}^0\omega_{j(m)(n)})\equiv\partial_i(eT^i),$$

where the right-hand side above is the same as in (1). Mǿller energy can be finally written as

$$E=\frac{1}{8\pi G}\int d^3x\,\partial_i[NeT^i-e(e^{(m)i}N^j-e^{(m)j}N^i)^0\omega_{j(0)(m)})]. \tag{16b}$$

Recall that we are ignoring *local* Lorentz transformations.

Besides the appearance of extra terms involving ${}^0\omega_{j(0)(m)}$ on the right-hand side of (16a) and (16b), there is also the crucial presence of the lapse function N multiplying eT^i. Therefore, even for configurations of the gravitational field for which the second term on the right-hand side of (16) does not contribute (if, say, $N^i=0$, as for the Schwarzschild solution) expressions (1) and (16) will lead to different results when applied to finite volumes of the three-dimensional space. Moreover, because of the presence of the lapse function, (16) is not invariant under time reparametrizations: $N'(x'^0)=(\partial x'^0/\partial x^0)N(x^0)$. Thus, for a finite volume of integration (16b) does not remain invariant under this reparametrization.

In the Einstein–Cartan theory the connection ${}^0\omega_{j(0)(m)}$ can be expressed in terms of the momenta canonically conjugated to $e_{(m)i}$. In the notation of Ref. 22 it is given by ${}^0\omega_{j(0)(m)}=(1/2e)(\pi_{(m)j}-\frac{1}{2}e_{(m)j}\pi)$ [see Eq. (12) of Ref. 22]. In this context (16b) can be rewritten as

$$E=\frac{1}{8\pi G}\int d^3x\,\partial_i\left[NeT^i-\frac{1}{2}e^{(m)i}N^j\pi_{(m)j}\right].$$

The expression above is exactly the energy expression for the Einstein–Cartan theory [see Eq. (21) of Ref. 22], assuming that the gravitational energy is obtained from the integration of surface terms of the total Hamiltonian. This expression is also very similar to (i) the integral of the surface terms in Eq. (6), and (ii) the energy expression considered by Nester[24] in the analysis of the positivity of the gravitational energy [Eq. (3.15) of Ref. 24]. All definitions of gravitational energy considered above agree for the total gravitational energy.

IV. BONDI'S RADIATING METRIC AND THE ASSOCIATED TRIADS

Bondi's metric is a not an exact solution of Einstein's equations. It describes the asymptotic form of a radiating solution. In terms of radiation coordinates (u,r,θ,ϕ), where u is the retarded time and r is the luminosity distance, Bondi's radiating metric is written as

$$ds^2=-\left(\frac{V}{r}e^{2\beta}-U^2r^2e^{2\gamma}\right)du^2-2e^{2\beta}\,du\,dr-2Ur^2e^{2\gamma}\,du\,d\theta+r^2(e^{2\gamma}\,d\theta^2+e^{-2\gamma}\sin^2\theta\,d\phi^2). \tag{17}$$

The metric above is such that the surfaces for which $u=\text{const}$ are null hypersurfaces. Each null radial (light) ray is labeled by particular values of u, θ, and ϕ. At spacelike infinity u takes the standard form $u=t-r$. The four quantities appearing in (17), V, U, β, and γ, are functions of u, r, and θ. Thus (17) displays axial symmetry. A more general form of this metric has been given by Sachs,[25] who showed that the most general metric tensor describing asymptotically flat gravitational waves depends on six functions of the coordinates.

The functions in (17) satisfy the following asymptotic behavior:

$$\beta=-\frac{c^2}{4r^2}+\cdots,\qquad \gamma=\frac{c}{r}+O\left(\frac{1}{r^3}\right)+\cdots,$$

$$\frac{V}{r}=1-\frac{2M}{r}-\frac{1}{r^2}\left[\frac{\partial d}{\partial\theta}+d\cot\theta-\left(\frac{\partial c}{\partial\theta}\right)^2-4c\left(\frac{\partial c}{\partial\theta}\right)\cot\theta-\frac{1}{2}c^2(1+8\cot^2\theta)\right]+\cdots,$$

$$U=\frac{1}{r^2}\left(\frac{\partial c}{\partial\theta}+2c\cot\theta\right)+\frac{1}{r^3}\left(2d+3c\frac{\partial c}{\partial\theta}+4c^2\cot\theta\right)+\cdots,$$

where $M=M(u,\theta)$ and $d=d(u,\theta)$ are related to the mass aspect and the dipole aspect, respectively, and from the function $c(u,\theta)$ we define the news function $\partial c(u,\theta)/\partial u$.

The application of (1) to Bondi's metric requires transforming it to coordinates t, r, θ and ϕ for which $t=\text{const}$ defines a spacelike hypersurface. Before proceeding, we recall that the analysis of (17) in t, x, y, z coordinates has already been performed by Goldberg,[26] in the investigation of the asymptotic invariants of gravitational radiation fields. Therefore, we carry out a coordinate transformation such that the new timelike coordinate is given by $t=u+r$. We arrive at

$$\begin{aligned}ds^2=&-\left(\frac{V}{r}e^{2\beta}-U^2r^2e^{2\gamma}\right)dt^2-2Ur^2e^{2\gamma}\,dt\,d\theta+2\left[e^{2\beta}\left(\frac{V}{r}-1\right)-U^2r^2e^{2\gamma}\right]dr\,dt\\&+\left[e^{2\beta}\left(2-\frac{V}{r}\right)+U^2r^2e^{2\gamma}\right]dr^2+2Ur^2e^{2\gamma}\,dr\,d\theta+r^2(e^{2\gamma}\,d\theta^2+e^{-2\gamma}\sin^2\theta\,d\phi^2).\end{aligned} \tag{18}$$

Therefore the metric restricted to a three-dimensional spacelike hypersurface is given by

$$ds^2=\left[e^{2\beta}\left(2-\frac{V}{r}\right)+U^2r^2e^{2\gamma}\right]dr^2+2Ur^2e^{2\gamma}\,dr\,d\theta+r^2(e^{2\gamma}\,d\theta^2+e^{-2\gamma}\sin^2\theta\,d\phi^2). \tag{19}$$

We must consider triads that correspond to the metric above. The construction of triads, in general, is a nontrivial step. If in a given coordinate system the metric tensor is diagonal, then the construction of triads is a relatively simple procedure. One must only make sure that the triads satisfy the appropriate boundary conditions at infinity. Recall that in order to obtain expression (9) for the ADM energy the triads must have the appropriate asymptotic behavior given by Eq. (8).

The metric tensor (19) has an off-diagonal element, and this fact adds a bit of complication in the construction of triads. Nevertheless, we can immediately write down two sets of triads that lead to this metric. They are given by

$$e_{(k)i}=\begin{pmatrix} A\sin\theta\cos\phi+B\cos\theta\cos\phi & rC\cos\theta\cos\phi & -rD\sin\theta\sin\phi \\ A\sin\theta\sin\phi+B\cos\theta\sin\phi & rC\cos\theta\sin\phi & rD\sin\theta\cos\phi \\ A\cos\theta-B\sin\theta & -rC\sin\theta & 0 \end{pmatrix}, \tag{20}$$

where

$$A=e^{\beta}\sqrt{2-\frac{V}{r}}, \tag{21a}$$

$$B=rUe^{\gamma}, \tag{21b}$$

$$C=e^{\gamma}, \tag{21c}$$

$$D=e^{-\gamma}, \tag{21d}$$

and

$$e_{(k)i}=\begin{pmatrix} A'\sin\theta\cos\phi & rB'\cos\theta\cos\phi+rC'\sin\theta\cos\phi & -rD'\sin\theta\sin\phi \\ A'\sin\theta\sin\phi & rB'\cos\theta\sin\phi+rC'\sin\theta\sin\phi & rD'\sin\theta\cos\phi \\ A'\cos\theta & -rB'\sin\theta+rC'\cos\theta & 0 \end{pmatrix}, \tag{22}$$

where

$$A'=\left[e^{2\beta}\left(2-\frac{V}{r}\right)+U^2r^2e^{2\gamma}\right]^{1/2}, \tag{23a}$$

$$B'=\frac{1}{A'}e^{\beta+\gamma}\sqrt{2-\frac{V}{r}}, \tag{23b}$$

$$C'=\frac{1}{A'}Ure^{2\gamma}, \tag{23c}$$

$$D'=e^{-\gamma}. \tag{23d}$$

It is easy to see that both (20) and (22) yield the metric tensor (19) through the relation $e_{(i)j}e_{(i)k}=g_{jk}$. They are related by a *local* SO(3) transformation.

Triads given by (20) and (22) are the *simplest* sets of triads that satisfy the two basic requirements: (i) the triads must have the asymptotic behavior given by (8); (ii) by making the physical parameters of the metric vanish we must have $T_{(k)ij}=0$ everywhere. In the present case if we make $M=d=c=0$, both (20) and (22) acquire the form

$$e_{(k)i}=\begin{pmatrix} \sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ \sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ \cos\theta & -r\sin\theta & 0 \end{pmatrix}. \tag{24}$$

In Cartesian coordinates the expression above can be reduced to the diagonal form $e_{(k)i}(x,y,z)=\delta_{ik}$. The requirement (ii) above is essentially equivalent to the establishment of a reference space, as discussed in Ref. 9. Without the notion of a reference space we cannot define gravitational energy. Note that by a suitable choice of a local SO(3) rotation we can make the flat space triad (24) satisfy the requirement (i), but not (ii).

It is possible to show that (20) and (22) are particular cases of an infinite set of triads that satisfy both requirements above. It is given by

$$e_{(k)i}=\begin{pmatrix} \mathcal{A}\sin\theta\cos\phi+\mathcal{B}\cos\theta\cos\phi & r\mathcal{C}\cos\theta\cos\phi+r\mathcal{D}\sin\theta\cos\phi & -re^{-\gamma}\sin\theta\sin\phi \\ \mathcal{A}\sin\theta\sin\phi+\mathcal{B}\cos\theta\sin\phi & r\mathcal{C}\cos\theta\sin\phi+r\mathcal{D}\sin\theta\sin\phi & re^{-\gamma}\sin\theta\cos\phi \\ \mathcal{A}\cos\theta-\mathcal{B}\sin\theta & -r\mathcal{C}\sin\theta+r\mathcal{D}\cos\theta & 0 \end{pmatrix}, \tag{25}$$

with the following definitions:

$$\mathcal{A}=\sqrt{e^{2\beta}(2-V/r)+U^2r^2(e^{2\gamma}-\mathbf{b}^2)}, \quad \mathcal{B}=\mathbf{b}Ur,$$

$$\mathcal{C}=\sqrt{e^{2\gamma}-\mathbf{d}^2U^2r^2}, \quad \mathcal{D}=\mathbf{d}Ur,$$

where $\mathbf{b}$ and $\mathbf{d}$ are arbitrary, dimensionless functions that must satisfy

$$\mathbf{b}=\sqrt{e^{2\gamma}-\mathbf{d}^2U^2r^2}+\mathbf{d}e^{-\gamma+\beta}\sqrt{2-V/r}.$$

By making $\mathbf{d}=0$ we obtain (20), and $\mathbf{b}=0$ leads to (22).

From the point of view of the TEGR, triads given by (20) and (22) are physically inequivalent (that is, they are not gauge equivalent), because we have seen that the Hamiltonian formulation established by Eq. (5) is not invariant under the local SO(3) group, but rather under the global SO(3). In the TEGR, the torsion tensor describes the way in which the space–time is deformed. The latter is thus considered as a continuum with microstructure.[5] Therefore, the same space, defined uniquely by the metric tensor, may be deformed in several ways, according to the manner in which one defines the triads. This is essentially the geometrical meaning of the noninvariance of the TEGR under local SO(3) transformations.

In the Hamiltonian formulation of the TEGR, the basic geometrical field variable is the triad, not the metric tensor. Any set of triads should be ruled out on physical grounds, i.e., if they lead to incorrect physical statements concerning the energy content of the gravitational field.

In the next section we will obtain the expressions for the gravitational energy arising from (20) and (22). These expressions are quite different. Although the expression corresponding to (20) is simpler, as we will see, we have no definite experimental evidence in favor of it.

V. GRAVITATIONAL RADIATION ENERGY

In this section we will apply expression (1) both to (20) and (22). Since the two triads display distinct geometrical properties, we expect to obtain different expressions for the energy density $(1/8\pi)\partial_i(eT^i)$ (we will make the gravitational constant $G=1$). Our analysis is meaningful only in the asymptotic region of large values of the radial distance. However, we have no reason to expect (20) and (22) to yield different expressions for the *total* energy of the field. In fact, as we will see, they yield the same (expected) expression.

As we mentioned earlier, the significance of the present approach to the analysis of gravitational radiation fields resides in the fact that we can evaluate the gravitational energy inside a large but finite portion of a three-dimensional spacelike surface. In other words, by means of Gauss' law (1) can be evaluated over a surface far from the source of radiation, in the asymptotic limit where the metric components are precisely determined. Specifically, we will evaluate (1) inside a large sphere of radius r_0. The time evolution of the metric field will determine the time dependence of this energy, and consequently the energy radiated out of it.

In order to calculate $T^i=g^{ik}g^{mj}e_{(l)m}T_{(l)jk}$ we need the inverse metric g^{ij}. In terms of the definitions (21), it is given by

$$g^{ij}=\begin{pmatrix} \frac{1}{A^2} & -\frac{B}{re^{\gamma}A^2} & 0 \\ -\frac{B}{re^{\gamma}A^2} & \frac{1}{r^2e^{2\gamma}}\left(1+\frac{B^2}{A^2}\right) & 0 \\ 0 & 0 & \frac{1}{r^2e^{-2\gamma}\sin^2\theta} \end{pmatrix}. \tag{26}$$

We will initially consider the set of triads given by (20). In the following, a comma after a field quantity indicates a derivative: $A_{,1}$ and $A_{,2}$ indicative derivative of A with respect to r and θ, respectively. The torsion components for (20) are given by

$$T_{(1)12}=\cos\theta\cos\phi(C+rC_{,1}-A-B_{,2})+\sin\theta\cos\phi(-A_{,2}+B),$$

$$T_{(1)13}=\sin\theta\sin\phi(A-D-rD_{,1})+\cos\theta\sin\phi B,$$

$$T_{(1)23}=\sin\theta\sin\phi(-rD_{,2})+\cos\theta\sin\phi(rC-rD),$$

$$T_{(2)12}=\cos\theta\sin\phi(C+rC_{,1}-A-B_{,2})+\sin\theta\sin\phi(-A_{,2}+B),$$

$$T_{(2)13}=\sin\theta\cos\phi(-A+D+rD_{,1})+\cos\theta\cos\phi(-B),$$

$$T_{(2)23}=\sin\theta\cos\phi(rD_{,2})+\cos\theta\cos\phi(-rC+rD),$$

$$T_{(3)12}=\sin\theta(A+B_{,2}-C-rC_{,1})+\cos\theta(-A_{,2}+B),$$

$$T_{(3)13}=T_{(3)23}=0.$$

Since we are interested in calculating the energy inside a surface of constant radius, only T^1 will be considered. By Gauss' law, the expression of this energy is given by

$$E=\frac{1}{8\pi}\int_S d\theta\, d\phi\, eT^1, \tag{27}$$

where S is a surface of fixed radius r_0, assumed to be large as compared with the dimension of the source, and the determinant e is given by $e=r^2A\sin\theta$. After a long but otherwise straightforward calculation, we arrive at

$$E_I=\frac{r_0}{4}\int_0^{\pi}d\theta\left\{\sin\theta\left[e^{\gamma}+e^{-\gamma}-\frac{2}{A}\right]+\frac{1}{A}\frac{\partial}{\partial\theta}(Ur\sin\theta)\right\}, \tag{28}$$

with A defined by (21a). Let us note that the field quantities appearing in (28) are functions of $u=t-r$: $M=M(t-r,\theta)$, $c=c(t-r,\theta)$, etc. Therefore, once these functions are known, one can explicitly calculate the variation of E_I with respect to the time t, however, only in the limit where the metric components are precisely determined.

Unfortunately, the expansion of E_I up to terms in $1/r_0$, making use of the asymptotic behavior of U, V, β, and γ, yields no simple expression. It is given by

$$E_I=\frac{1}{2}\int_0^{\pi}d\theta\sin\theta M-\frac{1}{4r_0}\int_0^{\pi}d\theta\sin\theta\left[\left(\frac{\partial c}{\partial\theta}\right)^2+4c\left(\frac{\partial c}{\partial\theta}\right)\cot\theta+4c^2\cot^2\theta\right]$$
$$-\frac{1}{4r_0}\int_0^{\pi}d\theta\, M\frac{\partial}{\partial\theta}\left[\sin\theta\left(\frac{\partial c}{\partial\theta}+2c\cot\theta\right)\right]. \tag{29}$$

In the calculation above we have assumed that $U(\theta)\sin\theta = d(\theta)\sin\theta = 0$ when $\theta = 0$ or π.

We observe that E_I yields Bondi energy in the limit $r\to\infty$ in the *static* case (i.e., when M is a function of θ only), so that E_I is the total gravitational energy. However, in the nonstatic case an expression for the loss of mass due to gravitational radiation can be obtained from (29). This is one major result of our analysis.

Let us recall that Bondi's *mass aspect* $m(u)$,

$$m(u) = \frac{1}{2}\int_0^\pi d\theta \sin\theta M(u,\theta), \tag{30}$$

depends on the null foliation used. The mass aspect $m(u)$ can be understood as a mass associated to each null cone determined by the equation $u=$const. Since the limit $r\to\infty$ corresponds to $u\to-\infty$ for finite t, we see once again that in this limit E_I gives the total energy because it corresponds to the initial value of the Bondi energy.

We will consider next the second set of triads, Eq. (22). The components of the torsion tensor resulting from the latter are given by

$$T_{(1)12} = \cos\theta\cos\phi(B' + rB,'_1 - A') + \sin\theta\cos\phi(C' + rC,'_1 - A,'_2),$$

$$T_{(1)13} = \sin\theta\sin\phi(A' - D' - rD,'_1),$$

$$T_{(1)23} = \cos\theta\sin\phi(rB' - rD') + \sin\theta\sin\phi(rC' - rD,'_2),$$

$$T_{(2)12} = \cos\theta\sin\phi(B' + rB,'_1 - A') + \sin\theta\sin\phi(C' + rC,'_1 - A,'_2),$$

$$T_{(2)13} = -\sin\theta\cos\phi(A' - D' - rD,'_1),$$

$$T_{(2)23} = -\cos\theta\cos\phi(rB' - rD') - \sin\theta\cos\phi(rC' - rD,'_2),$$

$$T_{(3)12} = -\sin\theta(B' + rB,'_1 - A') + \cos\theta(C' + rC,'_1 - A,'_2),$$

$$T_{(3)13} = T_{(3)23} = 0.$$

As in the previous case, we are interested in calculating the energy in the interior of a surface of constant radius r_0. Therefore, only the knowledge of T^1 will be necessary. After a long calculation, we first arrive at

$$\begin{aligned} eT^1 = &-\frac{r\sin\theta}{A}\left\{ e^{-2\gamma}\left[B'\frac{\partial}{\partial r}(rB') + C'\frac{\partial}{\partial r}(rC') - A'B' - C'\frac{\partial A'}{\partial\theta}\right]\right. \\ &\left. + e^{2\gamma}\left[-A'e^{-\gamma} - r\frac{\partial\gamma}{\partial r}e^{-2\gamma} + e^{-2\gamma}\right]\right\} - \frac{rB\sin\theta}{A}\left[C' + \frac{\partial\gamma}{\partial\theta}e^{-\gamma}\right] - \frac{rB\cos\theta}{A}[B' - e^{-\gamma}], \end{aligned} \tag{31}$$

where the primed quantities are given by (23). It is not straightforward to put the expression above in a simplified form. After some rearrangements we can finally write the energy expression (27) as

$$\begin{aligned} E_{II} = \frac{r_0}{4}\int_0^\pi d\theta \frac{1}{A}&\left\{ \sin\theta\left[e^{\gamma}A' + e^{-2\gamma}A'B' - 2 + e^{-2\gamma}\frac{\partial A'}{\partial\theta}C' - BC' - Be^{-\gamma}\frac{\partial\gamma}{\partial\theta}\right]\right. \\ &\left. - B\cos\theta[B' - e^{-\gamma}]\right\}. \end{aligned} \tag{32}$$

Like Eq. (28), this expression represents the energy enclosed by a large spherical surface of radius r_0. Expanding the expression above up to the first power of $1/r_0$, we find

$$E_{II}=\frac{1}{2}\int_0^\pi d\theta\, M\sin\theta-\frac{1}{4r_0}\int_0^\pi d\theta\sin\theta\bigg[3M^2+\frac{5}{2}\left(\frac{\partial c}{\partial\theta}\right)^2+10c\left(\frac{\partial c}{\partial\theta}\right)\cot\theta+8c^2\cot^2\theta -\left(\frac{\partial M}{\partial\theta}\right)\left(\frac{\partial c}{\partial\theta}+2c\cot\theta\right)\bigg]-\frac{1}{4r_0}\int_0^\pi d\theta\cos\theta\left[2c\left(\frac{\partial c}{\partial\theta}\right)+4c^2\cot\theta\right]. \tag{33}$$

We are again assuming $U(\theta)\sin\theta=d(\theta)\sin\theta=0$ for $\theta=0,\ \pi$.

We observe that in the limit $r\to\infty$, E_{II} also gives the total energy. As before, for a finite (but sufficiently large) value of r_0 we can compute the loss of mass due to gravitational radiation, once the functions M and c are known in the asymptotic region.

VI. THE SELECTION OF TRIADS

In Sec. IV we obtained an infinite set of triads that yield the three-dimensional spacelike section of Bondi's metric, and in the previous section we considered in detail only the simplest constructions. Of course, simplicity is a major feature of physical systems, but we are really in need of experimental evidence that leads to a definite description. We need actual realizations of the quantities $M(r-t,\theta)$, $c(r-t,\theta)$, $d(r-t,\theta)$ and experimental evidence on how the energy is radiated away in order to arrive simultaneously at the correct energy expression arising from (27) and at the definite expression of $e_{(k)i}$.

However, we can envisage two possible types of conditions on the triads that associate a unique triad with the three-dimensional metric tensor.

The first condition regards the energy content of the gravitational field. If we stick to the point of view according to which physical systems in nature have a tendency to be in states of minimum energy, then the correct triad for the spacelike section of Bondi's metric is the one that minimizes expression (27) for all possible constructions of $e_{(k)i}$. By means of this criterium we consider triads given by (25), or any further construction that complies with the two conditions stated in Sec. IV, and ask which one yields the smaller value of energy contained within a surface of constant radius, in similarity with the calculations of (29) and (33). Unfortunately, this analysis cannot be carried out unless M, c, and d are known.

Certainly one can ask whether only (27) should be minimized or the energy density should be everywhere a minimum. In the context of Bondi's metric the latter possibility cannot be considered, because the metric is valid only in the asymptotic limit; but in the general case it is an open question that must be carefully addressed.

The second condition takes into account Eq. (8): we require the triads to have the asymptotic behavior determined by (8) with a *symmetric* tensor $h_{jk}=h_{kj}$. Again, one has to find out of (25), which realization of $e_{(i)k}$ in Cartesian coordinates complies with this criterium. This condition may be understood as a *rotational gauge condition*. Note that, as it stands, h_{jk} in Eq. (8) is not required to be symmetric [in Eq. (9) only the symmetrical part contributes]. By explicit calculations we observe that neither (20) nor (22) satisfy this second condition.

The two conditions above may not be mutually excluding. On the contrary, they may lead to the same triad. The determination of the correct triad is certainly an essential and crucial issue of the theory and will be further investigated elsewhere in the general case, with special attention to Bondi's metric, in light of the conditions above.

We observed that both (29) and (33) yield the same total energy. This is also the case if we carry out the calculations with a more complicated triad, whether belonging to (25) or not, which is related to (20) or (22) by a local SO(3) transformation with an appropriate asymptotic behavior. Let us consider a local SO(3) transformation, given by

$$\tilde{e}^{(k)}{}_i(x)=\Lambda^{(k)}{}_{(l)}(x)e^{(l)}{}_i(x). \tag{34}$$

Under (34) the energy expression (1) transforms as

$$\tilde{E}=E+\frac{1}{8\pi}\int_V d^3x\,\partial_i[eg^{ik}\Lambda^{(l)}{}_{(m)}e^{(m)j}(e_{(n)k}\,\partial_j\Lambda_{(l)}{}^{(n)}-e_{(n)j}\,\partial_k\Lambda_{(l)}{}^{n})]. \tag{35}$$

Expression (35) can be best analyzed if we consider an infinitesimal rotation. We assume that in the limit $r\rightarrow\infty$, the SO(3) elements have the asymptotic behavior

$$\Lambda^{(k)}{}_{(l)}\approx\delta^{(k)}_{(l)}+{}^0\omega^{(k)}{}_{(l)}+{}^1\omega^{(k)}{}_{(l)}\left(\frac{1}{r}\right),$$

such that ${}^{0,1}\omega_{(k)(l)}=-{}^{0,1}\omega_{(l)(k)}$, and $\{{}^0\omega_{(k)(l)}\}$ are constants. Taking into account (8), it is easy to see that when integrated over the whole three-dimensional space the integral on the right-hand side of (35) reduces to a vanishing expression:

$$\frac{1}{8\pi}\int_{V\rightarrow\infty}d^3x(\partial_i\,\partial_j\,{}^1\omega_{(i)(j)}-\partial_i\,\partial_i\,{}^1\omega_{(j)(j)})=0.$$

Therefore we expect to find the same result for the total energy if we evaluate (27) out of any element of the set of triads (25).

VII. DISCUSSION

The application of the energy definition (1) for a given solution of Einstein's equations requires considering a foliation of the space–time in three-dimensional spacelike surfaces. The metric for the spacelike section of Bondi's radiating metric admits an infinite set of triads related by local SO(3) transformations. In the present case, from this infinite set of triads we singled out two of them. We have considered in detail the two ones that exhibit the simplest structures in spherical coordinates, and that (i) satisfy the asymptotic conditions given by (8); (ii) reduce to the reference space ($T_{(k)ij}=0$ everywhere) if we make the physical parameters vanish: $M=c=d=0$.

The two sets of triads, (20) and (22), describe the spacelike section of Bondi's metric given by (19) and lead to the energy expressions (28) and (32), respectively. These expressions establish distinct and quite definite physical predictions. They allow us to compute the energy radiated from the interior of a spherical surface of constant radius r_0. It would be a remarkable achievement of the TEGR if, on physical grounds, we could decide for one of them or even for an arbitrary element of (25). In the TEGR the space (space–time) geometry is fundamentally described by triads (tetrads). Unfortunately, we do not dispose of experimental information for taking such a decision.

It is a very important result that the *total* energy due to both sets of triads [as well as from any element of (25)] agrees exactly with the static Bondi energy, in which case the energy arises from the integration of $M=M(\theta)$. In fact, the definition of Bondi's *mass aspect* is basically motivated by the fact that in the static case, and by investigating the asymptotic properties of the gravitational field, $M(\theta)$ arises as the mass of an isolated system.

The final expressions (29) and (33) support the consistency of the definition (1), and the relevance of the TEGR as a fundamental description of general relativity. However the present analysis, which was developed on spacelike surfaces, has to be compared with the one recently carried out on null surfaces.[27] Let us recall that in order to obtain the Hamiltonian formulation given by (5)–(7) we imposed the time gauge condition. Therefore, the resulting geometry may be understood as a *three-dimensional* teleparallel geometry, since the teleparallelism is restricted to the three-dimensional spacelike surface. On the other hand, in the Hamiltonian formulation developed in Ref. 27, we have not fixed any particular tetrad component, and consequently the teleparallel geometry is really *four dimensional*.

In Ref. 27 the constraints also contain a total divergence, in similarity with (7a), and may be taken likewise to define the gravitational energy–momentum. Although it appears that the geometrical framework of Ref. 27 is better suited to the analysis of the Bondi–Sachs metric, we note that the energy expression arising there is considerably more complicated than (1). Moreover, we do not know yet whether the constraint algebra leads to a consistent Hamiltonian formulation (either on null or spacelike surfaces). We also note that one Hamiltonian formulation cannot be reduced to the other by means of gauge fixing. Nature admits only one correct physical description, and therefore either the three-dimensional or the four-dimensional teleparallel geometry is the correct candidate for describing the energy properties of the gravitational field. All these issues will be considered in the near future.

ACKNOWLEDGMENTS

This work was supported in part by CNPQ. J.F.R.N. is supported by CAPES, Brazil.

[1] C. Møller, Mat. Fys. Skr. Dan. Vidensk. Selsk. **1**, No. 10 (1961).
[2] C. Møller, Ann. Phys. (NY) **12**, 118 (1961).
[3] C. Møller, *Tetrad Fields and Conservation Laws in General Relativity*, Proceedings of the International School of Physics ''Enrico Fermi,'' edited by C. Møller (Academic, London, 1962).
[4] F. W. Hehl, in *Proceedings of the 6th School of Cosmology and Gravitation on Spin, Torsion, Rotation and Supergravity*, Erice, 1979, edited by P. G. Bergmann and V. de Sabbata (Plenum, New York, 1980).
[5] F. W. Hehl, J. D. McCrea, E. W. Milke, and Y. Ne'eman, Phys. Rep. **258**, 1 (1995).
[6] K. Hayashi and T. Shirafuji, Phys. Rev. D **19**, 3524 (1979).
[7] J. W. Maluf, J. Math. Phys. **35**, 335 (1994).
[8] J. W. Maluf, J. Math. Phys. **36**, 4242 (1995).
[9] J. W. Maluf, E. F. Martins, and A. Kneip, J. Math. Phys. **37**, 6302 (1996).
[10] J. W. Maluf, J. Math. Phys. **37**, 6293 (1996).
[11] J. W. Maluf and A. Kneip, J. Math. Phys. **38**, 458 (1997).
[12] J. A. Schouten, *Ricci Calculus*, 2nd ed. (Springer-Verlag, London, 1954), p. 142.
[13] H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, Proc. R. Soc. London, Ser. A **269**, 21 (1962).
[14] R. Arnowitt, S. Deser, and C. W. Misner, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[15] C. Møller, Phys. Lett. **3**, 329 (1963).
[16] S. A. Hayward, Phys. Rev. D **49**, 831 (1994).
[17] S. A. Hayward, Class. Quantum Grav. **11**, 3037 (1994).
[18] R. D. Hecht and J. M. Nester, Phys. Lett. A **217**, 81 (1996).
[19] J. D. Brown, S. R. Lau, and J. W. York, Jr., Phys. Rev. D **55**, 1977 (1997).
[20] F. Müller-Hoissen and J. Nitsch, Phys. Rev. D **28**, 718 (1983).
[21] W. Kopczyński, J. Phys. A **15**, 493 (1982).
[22] J. W. Maluf, Class. Quantum Grav. **8**, 287 (1991).
[23] J. W. Maluf, Gen. Relativ. Gravit. **30**, 413 (1998).
[24] J. M. Nester, Int. J. Mod. Phys. A **4**, 1755 (1989).
[25] R. K. Sachs, Proc. R. Soc. London, Ser. A **270**, 103 (1962).
[26] J. N. Goldberg, Phys. Rev. **131**, 1367 (1963).
[27] J. W. Maluf and J. F. da Rocha-Neto, ''General relativity on a null surface: Hamiltonian formulation in the teleparallel geometry,'' Gen. Relativ. Grav. **31** (1999), gr-qc/9808001.

Real sector of the nonminimally coupled scalar field to self-dual gravity

Merced Montesinos[a)]
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260 and Departamento de Física, Centro de Investigación y de Estudios Avanzados del I.P.N., Av. I.P.N. No. 2508, 07000 México D.F., México

Hugo A. Morales-Técotl[b)]
Departamento de Física, Universidad Autónoma Metropolitana-Iztapalapa, Apartado Postal 55-534, 09340, México D.F., México

Luis F. Urrutia[c)] and J. David Vergara[d)]
Departamento de Física de Altas Energías, Instituto de Ciencias Nucleares, Universidad Nacional Autonoma de México, Apartado Postal 70-543, 04510, México D.F., México

(Received 3 June 1998; accepted for publication 9 September 1998)

A scalar field nonminimally coupled to gravity is studied in the canonical framework, using self-dual variables. The corresponding constraints are first class and polynomial. To identify the real sector of the theory, reality conditions are implemented as second class constraints, leading to three real configurational degrees of freedom per space point. Nevertheless, this realization makes nonpolynomial some of the constraints. The original complex symplectic structure reduces to the expected real one, by using the appropriate Dirac brackets. For the sake of preserving the simplicity of the constraints, an alternative method preventing the use of Dirac brackets, is discussed. It consists of converting all second class constraints into first class by adding extra variables. This strategy is implemented for the pure gravity case. © *1999 American Institute of Physics.* [S0022-2488(99)01102-0]

I. INTRODUCTION

The failure of using perturbation theory to quantize General Relativity (GR) has led to different approaches, like string theory and canonical quantization, which intend to define quantum gravity.[1] Instead of incorporating the remaining fundamental interactions as the former, the latter approach consists of just adopting standard quantum theory and classical GR. The space–time structure, on the other hand, seems to be sensible to the nonperturbative aspects modeling it.[2,3] This is in itself a major motivation for studying canonical GR. Some success has been achieved within canonical quantum gravity since the late eighties, after the introduction, by Ashtekar,[4] of a set of complex canonical variables which greatly simplify the form of the constraints to which the theory is reduced. The kernel of this algebra of constraints defines the space of physical states in the theory and some formal elements of it have been constructed.[4]

In spite of its simplicity, the constraints of GR in terms of Ashtekar variables describe complex gravity. Ashtekar himself considered reducing to the real sector, through the introduction of an inner product designed to make Hermitian the physical operators. However, this strategy has not worked up to now, except for some particular cases. Other alternatives have been presented to avoid the use of reality conditions, at the price of a more cumbersome form of the constraints. Yet, quantum mechanically, the situation seems tractable.[3]

[a)]Electronic mail: merced@fis.cinvestav.mx
[b)]Associate member of Abdus Salam-ICTP, Trieste, Italy. Electronic mail: hugo@xanum.uam.mx
[c)]Electronic mail: me@aurora.nuclecu.unam.mx
[d)]Electronic mail: vergara@nuclecu.unam.mx

There still remains the possibility of keeping the self-dual canonical formalism and trying to envisage how to select the real sector of the theory. Indeed this is possible, as it has been shown for pure gravity in Ref. 5, at the classical level. Reality conditions are implemented as second class constraints. The present work is devoted to show how this result is also valid in the case of a scalar field nonminimally coupled to gravity.

The strategy followed in this paper is an extension of the pure gravity analysis of Ref. 5, to incorporate a nonminimally coupled scalar field . It consists of five steps: (i) each complex canonical variable is splitted into one real and one imaginary degrees of freedom; (ii) then, every real and imaginary part is regarded as a configuration variable. An extended phase space is hence defined, where the corresponding momenta will arise directly from the action. Notice that the definition of such momenta will produce primary constraints; (iii) next, the reality conditions are implemented upon the splitting, as additional primary constraints. The criterion being to restrict to a real three-metric together with a real scalar field, during the whole evolution of the system; (iv) after imposing the conservation of the constraints, which amounts to obtaining all possible secondary constraints, a final classification of the full set into first and second class is performed. This may require the redefinition of some of them; and finally, (v) the issue of how to deal with the resulting set of second class constraints should be addressed. Since the use of Dirac brackets take the first class constraints back to its Palatini canonical form, which is highly nonpolynomial, some alternatives should be tried.

A review of how to select the real sector in phase space of pure self-dual gravity is made in Sec. II. It includes the result of converting all second class into first class constraints, by introducing extra canonical variables. Thus, Dirac brackets are not used. Next, the extension of the method to the case of the scalar field nonminimally coupled to gravity is described in Sec. III. The polynomial form of the constraints for the complex theory is exhibited, as well as the whole set of first and second class constraints describing the real sector of the theory. Finally, the last section contains some conclusions and perspectives. A possibility regarding how to avoid the use of Dirac brackets to eliminate the second class constraints for the scalar field case is briefly discussed. A point on notation: $\tilde{E}^{ai}$ is considered here as a complex density inverse triad of weight one, whilst $\tilde{e}^{ai}$ is its real part. This is just the opposite convention of Ref. 5, but it has the advantage that adjusts to the rest of the literature.[4] The number of over and under tildes represent the positive or negative density weight (± 1, ± 2, etc.) respectively, unless it is obvious from the definition of the different variables. In the case of the $\delta(\tilde{x},y)$ the tilde indicates it is a density of weight one in the argument x.

II. PURE GRAVITY

The Ashtekar complex canonical variables are: (i) $\tilde{E}^{ai} := \tilde{E} E^{ai}$, with E^{ai} being the triad ($E^{ai}E^b_i := q^{ab}$, q^{ab} is the spatial three-metric), and $a,b,\ldots = 1,2,3$ are spatial indices, whereas $i,j,\ldots = 1,2,3$ are SO(3) internal indices. Also, $\tilde{E} := \det E_{bj}$ with E_{bj} being the inverse of E^{ai}. (ii) A_{ai} is the three-dimensional projection of the self-dual connection,[4] with associated covariant derivative $\mathcal{D}_a \lambda_i = \partial_a \lambda_i + \epsilon_{ijk} A^j_a \lambda^k$ and $F^i_{ab} := \partial_a A^i_b - \partial_b A^i_a + \epsilon^i{}_{jk} A^a_j A^k_b$ is the corresponding curvature. In terms of these variables, the self-dual action of canonical GR is given by

$$S = \int dt\, d^3x \{ -i \tilde{E}^{ai} \dot{A}_{ai} - \underset{\sim}{N} \tilde{\tilde{\mathcal{S}}} - N^a \tilde{\mathcal{V}}_a - N^i \tilde{\mathcal{G}}_i \}, \tag{1}$$

where

$$\tilde{\tilde{\mathcal{S}}} := \epsilon_{ijk} \tilde{E}^{ai} \tilde{E}^{bj} F^k_{ab}, \quad \tilde{\mathcal{V}}_a := \tilde{E}^b_j F^j_{ab}, \quad \tilde{\mathcal{G}}_i := \mathcal{D}_a \tilde{E}^a_i, \tag{2}$$

are the constraints of the theory and $\underset{\sim}{N}, N^a, N^i$ are Lagrange multipliers. Such constraints are first class and polynomial in the phase space variables. Let us denote by $\mathcal{R}$ any of them and by $\{\mathcal{R}\}$ the full set.

Notice that having 18 complex phase space variables $(A_{ai},\tilde{E}^{bj})$, together with 7 complex first class constraints, $\{\mathcal{R}\}$, leaves us with 2 complex configurational degrees of freedom. Then, in order to recover the 2 real configurational degrees of freedom per point, further constraints are necessary. To this end, let us introduce the splitting

$$\tilde{E}^{ai}=\tilde{e}^{ai}+i\tilde{\mathcal{E}}^{ai}, \quad A_{bj}=\gamma_{bj}-iK_{bj}. \tag{3}$$

From now on, all the above thirty six degrees of freedom are taken as configuration variables in the action (1). Hence, the associated canonical momenta Π lead to the primary constraints $\Phi_{\mathcal{E}ai}=\Pi_{\mathcal{E}ai}, \Phi_{\gamma}^{ai}=\Pi_{\gamma}^{ai}+i\tilde{e}^{ai}$, $\Phi_K^{ai}=\Pi_K^{ai}+\tilde{e}^{ai}$, $\Phi_{eai}=\Pi_{eai}$, which as a set is denoted by $\{\Phi\}$. The coordinates of the total phase space are $Y_A=(\tilde{e}^{ai},\tilde{\mathcal{E}}_{ai},\gamma^{ai},K_{ai},\Pi_{eai},\Pi_{\mathcal{E}ai},\Pi_{\gamma}^{ai},\Pi_{ai}^{K})$.

The reduction of the complex phase space $(A_{ai},\tilde{E}^{bj})$ to a real one is achieved by means of the following reality conditions

$$\psi^{ai}:=\tilde{\mathcal{E}}^{ai}=0, \quad \chi_{ai}:=\gamma_{ai}-f_{ai}(\tilde{e})=0, \tag{4}$$

which are subsequently taken as additional primary constraints. The constraints ψ^{ai} enforces the $\tilde{E}^{ai}$ to be real, and hence the corresponding three-metric. The constraints χ_{ai} ensures that, upon evolution, $\tilde{E}^{ai}$ keeps being real. Using the compatibility condition between a real torsion-free connection and the triad, the form of f_{ai} is chosen as $f_{ai}=\frac{1}{2}[\underset{\sim}{e}_{ai}\underset{\sim}{e}_c^j\epsilon_{jrs}-2\underset{\sim}{e}_{aj}\underset{\sim}{e}_c^j\epsilon_{irs}]\tilde{e}^{dr}\partial_d\tilde{e}^{cs}$. Let us observe that χ_{ai} is not polynomial in $\tilde{e}^{bj}$.

The full set of primary constraints is $\{\{\mathcal{R}\},\ \psi,\ \chi,\{\Phi\}\}$, written in terms of the real variables Y_A. The evolution of the primary constraints does not introduce additional constraints. After redefining $\Phi_{eai}\rightarrow\Phi'_{eai}=\Phi_{eai}+\alpha_{aibj}\Phi_{\gamma}^{bj}+\beta_{ai}^{bj}\chi_{bj}+\eta_{aibj}\Phi_K^{bj}$, the Poisson brackets matrix for the subset $\{\Xi\}=\{\{\Phi'\},\ \psi,\ \chi\}$ reveals them as second class constraints. Besides having a simple form, it is a phase space independent, block diagonal matrix with non zero determinant.

To keep $\{\mathcal{R}\}$ as a first class set it is enough to redefine each element as $\mathcal{R}'=\mathcal{R}+\{\Phi_{\mathcal{E}bj},\mathcal{R}\}\psi^{bj}+\{\Phi_{\gamma}^{bj},\mathcal{R}\}\chi_{bj}+\{\Phi'_{ebj},\mathcal{R}\}\Phi_K^{bj}-\{\Phi_K^{bj},\mathcal{R}\}\Phi'_{ebj}$, so that they have zero Poisson brackets with the previous second class subset. This redefinition preserves the property $\{\mathcal{R}',\mathcal{Q}'\}\approx 0$, for any pair of constraints in $\{\mathcal{R}'\}$. In this way, there are no additional contributions to the set of primary constraints $\{\Upsilon\}:=\{\mathcal{R}'\}\cup\{\Xi\}$, which includes the reality conditions. Counting the independent variables gives 2 real configurational degrees of freedom per space point, as it should be for real GR.[5]

At this point Dirac's program calls for the elimination of the second class constraints through the use of Dirac brackets. This, however, would yield a cumbersome form for the remaining constraints. One might avoid such treatment of the second class constraints by transforming them into first class constraints. To achieve this, by means of the Batalin-Tyutin procedure,[6,7] one adds a new canonical pair, $\{Q^{ai},P_{bj}\}=\delta_a^b\delta_i^j\delta^{(3)}$, per original couple of second class constraints, i.e., the phase space is further enlarged with the new variables $\Psi_{\Xi}=(Q_{\mathcal{E}}^{ai},Q_{\gamma ai},Q_e^{ai},P_{\mathcal{E}ai},P_{\gamma}^{ai},P_{eai})$. In the present case, the set of first class constraints replacing the former second class set is

$$\overline{\psi}^{ai}:=\tilde{\mathcal{E}}^{ai}+Q_{\mathcal{E}}^{ai}, \quad \overline{\Phi}_{\mathcal{E}ai}:=\Pi_{\mathcal{E}ai}-P_{\mathcal{E}ai},$$

$$\overline{\chi}_{ai}:=\gamma_{ai}-f_{ai}(\tilde{e})+Q_{\gamma ai}, \quad \overline{\Phi}_{\gamma}^{ai}:=\Pi_{\gamma}^{ai}+i\tilde{e}^{ai}-P_{\gamma}^{ai}, \tag{5}$$

$$\overline{\Phi}_K^{ai}:=\Pi_K^{ai}+\tilde{e}^{ai}+Q_e^{ai}, \quad \overline{\Phi}'_{eai}:=\Phi_{eai}+\alpha_{aibj}\Phi_{\gamma}^{bj}+\beta_{ai}^{bj}\chi_{bj}+\eta_{aibj}\Phi_K^{bj}-P_{eai},$$

which reduces to the original set by setting $Q^{ai}=0=P_{bj}$. Let us denote any of the constraints in (5) by $\overline{\Xi}_{\Lambda}$. Any pair satisfies $\{\overline{\Xi}_{\Lambda},\overline{\Xi}_{\Lambda'}\}=0$; i.e., the set (5) is first class. Next, it is necessary to keep the set $\{\mathcal{R}'\}$ first class. This can be done by recalling that the Poisson brackets matrix among the constraints $\{\Xi\}$ is independent of the phase space variables and by following the method of Ref. 7. Thus, one redefines $\mathcal{R}'$ as

$$\overline{\mathcal{R}'} \equiv \mathcal{R}'(Y-\overline{Y}), \tag{6}$$

where

$$Y_A-\overline{Y_A} := \Bigg\{ \tilde{e}^{ai}-Q_e^{ai}, \tilde{\mathcal{E}}^{ai}+Q_{\mathcal{E}}^{ai}, \gamma_{ai}+Q_{\gamma ai}+Q_e^{bj}\frac{\delta f_{ai}}{\delta\tilde{e}^{bj}}, K_{ai}+P_{eai}+iQ_e^{bj}\frac{\delta f_{bj}}{\delta\tilde{e}^{ai}},$$
$$\Pi_{eai}-P_{eai}-iQ_{\gamma ai}+P_\gamma^{bj}\frac{\delta f_{bj}}{\delta\tilde{e}^{ai}}+Q_e^{bj}\left(\frac{\delta^2 f_{ck}}{\delta\tilde{e}^{ai}\delta\tilde{e}^{bj}}\Phi_\gamma^{ck}+i\frac{\delta^2 f_{ai}}{\delta\tilde{e}^{ck}\delta\tilde{e}^{bj}}\Phi_K^{ck}+i\frac{\delta f_{bj}}{\delta\tilde{e}^{ai}}\right),$$
$$\Pi_{\mathcal{E}ai}-P_{\mathcal{E}ai}, \Pi_\gamma^{ai}-P_\gamma^{ai}-iQ_e^{ai}, \Pi_K^{ai}\Bigg\}. \tag{7}$$

The set (6) is in involution with $\{\overline{\Xi}_\Lambda\}$, i.e., $\{\overline{\mathcal{R}'},\overline{\Xi}_\Lambda\}=0$. Hence, the final whole set of constraints is first class and contains an Abelian ideal: $\{\overline{\Xi}\}$. The non-Abelian sector is just given by $\{\overline{\mathcal{R}'}\}$. By construction, this sector preserves the structure of the first class algebra among the elements of $\{\mathcal{R}'\}$. Notice that the set $\{\mathcal{R}\}$ depends only on the configurational variables $(\tilde{e},\mathcal{E},\gamma,K)$. In this way, the most involved modifications to $\{\mathcal{R}'\}$, via Eq. (6), come from the terms that are proportional to the second class constraints. It is worth emphasizing that all the nonpolynomiality of the constraints $\{\{\overline{\mathcal{R}'}\}\cup\{\overline{\Xi}\}\}$ arises only through one function, which is $f_{ai}(\tilde{e})$, appearing in the reality conditions (4). Thus, one might think that a choice of (4) in a polynomial form would solve this undesirable feature. However, as shown in Ref. 5, this not the case and one should look for alternative approaches.

III. NONMINIMAL SELF-DUAL ECKG THEORY

The second order action with scalar field nonminimally coupled to gravity is given by

$$S[g^{ab},\phi]=\int_M d^4x\left\{\sqrt{-g}\mathcal{R}-\frac{1}{2}\sqrt{-g}(g^{ab}\partial_a\phi\partial_b\phi+(m^2+\xi\mathcal{R})\phi^2)\right\}, \tag{8}$$

where ξ is a dimensionless constant. The canonical analysis of this action has been developed in Ref. 8. In a first order formalism one can adopt instead

$$S[\omega',e,\phi]=\int_M d^4x\left\{\frac{1}{2}ee_I^ae_J^b\Omega^2R'^{IJ}_{ab}[\omega']-\frac{1}{2}e(e_I^ae^{bI}\partial_a\phi\partial_b\phi+m^2\phi^2)\right\},$$
$$\Omega^2:=1-\xi\phi^2. \tag{9}$$

As opposed to (8), action (9) gives, upon variation with respect to ω'^{IJ}_a,

$$\omega'^{IJ}_a=\omega_a^{IJ}(e)+K_a^{IJ}, \quad K_a^{IJ}=\frac{1}{2}(e_a^Ie^{bJ}-e_a^Je^{bI})\frac{1}{\Omega^2}\partial_b\Omega^2, \tag{10}$$

where $\omega_a^{IJ}(e)$ is the spin connection for pure gravity and K_a^{IJ} is the contorsion supported by the matter field.[9] To construct a first order action equivalent to the second order action in (8), it is necessary to add the term $-(3/4)(1/\Omega^2)ee_I^ae^{bI}\partial_a\Omega^2\partial_b\Omega^2$ to the action (9). The corresponding canonical analysis of the self-dual part of this modified action was studied in Ref. 10. The conclusion in that paper, at the Hamiltonian level, is that the resulting constraints are nonpolynomial in the phase space variables. In the present work, the self-dual part of the action (9) is studied. Our result is that polynomial constraints are obtained, as opposed to the case in Ref. 10.

The system to be considered is described by the so called nonminimal self-dual Einstein–Cartan–Klein–Gordon (ECKG) action

$$S[e,{}^{+4}A,\phi]=\int d^4xe\left\{\Omega^{2+}\Sigma_{IJ}^{ab+4}F_{ab}^{IJ}({}^{+4}A)+\frac{\alpha}{2}(e_I^ae^{bI}\partial_a\phi\partial_b\phi+m^2\phi^2)\right\},$$

$$\Sigma_{IJ}^{ab}:=\tfrac{1}{2}(e_I^ae_J^b-e_J^ae_I^b),\qquad \Omega^2:=1+\alpha\xi\phi^2,\tag{11}$$

where the parameter α is introduced only to allow a rescaling of the scalar field ϕ. The value $\alpha=-1$ corresponds to the case usually found in the literature.[8,10] ${}^{+4}F_{ab}^{IJ}({}^{+4}A)$ is the curvature of the self-dual connection ${}^{+4}A$.

Proceeding with the canonical analysis, the 3+1 decomposition of space-time gives

$$S=\int dt\int_\Sigma d^3x\left\{\underset{\sim}{N}\Omega^2\left[-\frac{1}{2}\epsilon^{ijk}\tilde{E}_i^a\tilde{E}_j^bF_{abk}\right]+(-i\Omega^2\tilde{E}_i^a)\mathcal{L}_tA_a^i\right\}$$

$$+\int dt\int_\Sigma d^3x\{({}^{+4}A\cdot t)^i\mathcal{D}_a(-i\Omega^2\tilde{E}_i^a)+N^b(-i\Omega^2\tilde{E}_i^a)F_{ab}^i\}$$

$$+\frac{\alpha}{2}\int dt\int_\Sigma d^3x\left\{\underset{\sim}{N}\tilde{E}_i^a\tilde{E}^{bi}\partial_a\phi\partial_b\phi-\frac{1}{\underset{\sim}{N}}[\mathcal{L}_t\phi-\mathcal{L}_N\phi]^2+\underset{\sim}{N}(\sigma)^2m^2\phi^2\right\}.\tag{12}$$

In order to get (12) one performs the usual steps.[4] $\tilde{E}_i^a$ is the densitizied inverse triad field, A_a^i is the three-dimensional projection of the self-dual full connection (gravity +matter) and F_{ab}^i is the corresponding curvature given by $F_{ab}^i=\partial_aA_b^i-\partial_bA_a^i+\epsilon_{jk}^iA_a^jA_b^k$. $(\sigma)^2$ is the determinant of the three-dimensional spatial metric q_{ab}, expressed as a function of $\tilde{E}_i^a$ and ϵ^{ijk} is the volume element of the three-dimensional internal metric $q_{ij}=\mathrm{diag}(+1,+1,+1)$. Both sets of indices run from 1 to 3.

From (12) one finds the momentum variables associated with the gravitational and scalar fields. They are given by $\tilde{\pi}_i^a$ and $\tilde{\pi}_\phi$, where

$$\tilde{\pi}_i^a:=-i\Omega^2\tilde{E}_i^a,\qquad \tilde{\pi}_\phi:=\frac{\delta S}{\delta(\mathcal{L}_t\phi)}=-\frac{\alpha}{\underset{\sim}{N}}[\mathcal{L}_t\phi-\mathcal{L}_N\phi].\tag{13}$$

Then, plugging (13) into (12), one arrives at

$$S=\int dt\int_\Sigma d^3x\{\tilde{\pi}_i^a\mathcal{L}_tA_a^i+\tilde{\pi}_\phi\mathcal{L}_t\phi-(\underset{\sim}{N}^*\tilde{\tilde{C}}+({}^{+4}A\cdot t)^i\tilde{\mathcal{G}}_i+N^a\tilde{\mathcal{V}}_a)\},\tag{14}$$

where

$$\tilde{\tilde{C}}:=-\frac{1}{2}(\Omega^2)^2\epsilon^{ijk}\tilde{\pi}_i^a\tilde{\pi}_j^bF_{abk}+\frac{\alpha}{2}\Omega^2\tilde{\pi}_i^a\tilde{\pi}^{bi}\partial_a\phi\partial_b\phi-\frac{1}{2\alpha}(\Omega^2)^3(\tilde{\pi}_\phi)^2+\frac{\alpha}{2}i(\det\tilde{\pi}_i^a)m^2\phi^2,$$

$$\tilde{\mathcal{G}}_i:=-\mathcal{D}_a\tilde{\pi}_i^a,\qquad \tilde{\mathcal{V}}_a:=-\tilde{\pi}_i^bF_{ba}^i+\tilde{\pi}_\phi\partial_a\phi,\tag{15}$$

are the scalar, vector and Gauss constraints, respectively. In (14), the Lagrange multiplier $\underset{\sim}{N}$ was redefined to $\underset{\sim}{N}^*=\underset{\sim}{N}/(\Omega^2)^3$. Note that the set of constraints (15) is *polynomial* in the phase space variables A_a^i, $\tilde{\pi}_i^a$, ϕ and $\tilde{\pi}_\phi$, where the symplectic structure is given by

$$\{A_a^i(x),\tilde{\pi}_j^b(y)\}=\delta_a^b\delta_j^i\delta^3(x,\tilde{y}),$$

$$\{\phi(x),\tilde{\pi}_\phi(y)\}=\delta^3(x,\tilde{y}).\tag{16}$$

Let us compare the actions (1) and (14). We find the same number of constraints which, nevertheless, have extra terms containing the scalar field. Besides, the new momenta have contributions arising both from the scalar field and the gravity sector.

In order to count the number of degrees of freedom of the theory it is necessary to classify the constraints in terms of their algebra. We find it convenient to use the smeared form of the constraints on Σ. If $\underset{\sim}{N}$, N^a and v^i are arbitrary tensor fields, then one defines

$$C(\underset{\sim}{N}):=\int_\Sigma d^3x\underset{\sim}{N}\tilde{\tilde{C}}(x)$$

$$=\int_\Sigma d^3x\underset{\sim}{N}\left\{-\frac{1}{2}(\Omega^2)^2\epsilon^{ijk}\tilde{\pi}^a_i\tilde{\pi}^b_jF_{abk}+\frac{\alpha}{2}\Omega^2\tilde{\pi}^a_i\tilde{\pi}^{bi}\partial_a\phi\partial_b\phi\right.$$

$$\left.-\frac{1}{2\alpha}(\Omega^2)^3(\tilde{\pi}_\phi)^2+\frac{\alpha}{2}i(\det\tilde{\pi}^a_i)m^2\phi^2\right\},$$

$$C(\mathbf{N}):=\int_\Sigma d^3xN^a\tilde{\mathcal{V}}_a(x)=\int_\Sigma d^3xN^a\{-\tilde{\pi}^b_iF^i_{ba}+\tilde{\pi}_\phi\partial_a\phi\},\tag{17}$$

$$G(v):=\int_\Sigma d^3xv^i\tilde{\mathcal{G}}_i(x)=\int_\Sigma d^3xv^i\{-\mathcal{D}_a\tilde{\pi}^a_i\}.$$

A combination of the vector and Gauss constraints yields the so called *diffeomorphisms* constraint. In terms of the vector field **N**, this constraint has the form

$$D(\mathbf{N}):=\int_\Sigma d^3xN^a[\tilde{\mathcal{V}}_a+A^i_a\tilde{\mathcal{G}}_i](x)=\int_\Sigma d^3xN^a[-\tilde{\pi}^b_iF^i_{ba}+A^i_a(-\mathcal{D}_b\tilde{\pi}^b_i)].\tag{18}$$

The following results are useful in dealing with the algebra of constraints

$$\frac{\delta C(\underset{\sim}{N})}{\delta A^l_c}=-\epsilon_l^{\ ij}\mathcal{D}_b(\underset{\sim}{N}(\Omega^2)^2\tilde{\pi}^c_i\tilde{\pi}^b_j),$$

$$\frac{\delta C(\underset{\sim}{N})}{\delta\tilde{\pi}^c_l}=\underset{\sim}{N}\left[-(\Omega^2)^2\epsilon^{ljk}\tilde{\pi}^a_jF_{cak}+\alpha\Omega^2\tilde{\pi}^{al}\partial_a\phi\partial_c\phi+\left(\frac{\alpha}{2}im^2\phi^2\right)\right.$$

$$\left.\times\left(\frac{3}{3!}\underset{\sim}{\eta}_{abc}\tilde{\pi}^a_j\tilde{\pi}^b_k\epsilon^{jkl}\right),\right]$$

$$\frac{\delta G(v)}{\delta A^l_c}=-v^i\epsilon^k_{il}\tilde{\pi}^c_k,\quad\frac{\delta G(v)}{\delta\tilde{\pi}^c_l}=\mathcal{D}_cv^l,$$

$$\frac{\delta D(\mathbf{N})}{\delta A^l_c}=-\mathcal{L}_{\mathbf{N}}\tilde{\pi}^c_l,\quad\frac{\delta D(\mathbf{N})}{\delta\tilde{\pi}^c_l}=+\mathcal{L}_{\mathbf{N}}A^l_c.\tag{19}$$

Then, by using (19), the algebra of constraints turns out to be

$$\{C(\underset{\sim}{N}),C(\underset{\sim}{M})\}=-D(\mathbf{K})+G(K^aA_a),$$

$$\{C(\underset{\sim}{N}),D(\mathbf{M})\}=-C(\mathcal{L}_{\vec{M}}\underset{\sim}{N}),$$

$$\{C(\underset{\sim}{N}),G(v)\}=0,$$

$$\{D(\mathbf{N}),D(\mathbf{M})\}=D([\mathbf{N},\mathbf{M}]),$$

$$\{D(\mathbf{N}),G(v)\}=G(\mathcal{L}_{\mathbf{N}}v),$$

$$\{G(w),G(v)\}=D(-[w,v]). \tag{20}$$

In (20), the vector field $\mathbf{K}$ is defined by $K^a:=(\Omega^2)^4(\tilde{\pi}_i^a\tilde{\pi}^{bi})(\underset{\sim}{N}\partial_b\underset{\sim}{M}-\underset{\sim}{M}\partial_b\underset{\sim}{N})$, while the commutator of internal vectors is $[w,v]^i:=\epsilon^i_{jk}w^jv^k$. Also, the commutator of spatial vectors is defined by $[\mathbf{N},\mathbf{M}]^a:=\mathcal{L}_{\mathbf{N}}M^a$, as usual.

The set of constraints (20) is *first class*. The counting of degrees of freedom leads to: $2(9)+2(1)-2(7)=6$ phase space variables per point on Σ, which implies 3 *complex* degrees of freedom: two for the gravitational field and one for the scalar field. To recover the real sector of the theory, i.e., three *real* degrees of freedom per point, one has to supply additional constraints on the phase space variables A_a^i, $\tilde{\pi}_i^a$, ϕ, $\tilde{\pi}_\phi$. This is the subject of the following section.

IV. REAL DEGREES OF FREEDOM

The real sector of the theory is recovered by extending the corresponding steps developed for pure gravity in Ref. 5. To begin with, let us consider the action (14) in the explicit form

$$S=\int dt\int_\Sigma d^3x\{(-i\Omega^2\tilde{E}_i^a)\mathcal{L}_tA_a^i+\tilde{\pi}_\phi\mathcal{L}_t\phi-(\underset{\sim}{N}^*\tilde{\tilde{C}}+({}^{+4}A\cdot t)^i\tilde{\mathcal{G}}_i+N^a\tilde{\mathcal{V}}_a)\}, \tag{21}$$

where

$$\tilde{\tilde{C}}:=(\Omega^2)^2\left(-\frac{1}{2}\epsilon^{ijk}\right)(-i\Omega^2\tilde{E}_i^a)(-i\Omega^2\tilde{E}_j^b)F_{abk}+\frac{\alpha}{2}\Omega^2(-i\Omega^2\tilde{E}_i^a)(-i\Omega^2\tilde{E}^{bi})\partial_a\phi\partial_b\phi$$
$$-\frac{1}{2\alpha}(\Omega^2)^3(\tilde{\pi}_\phi)^2+\frac{\alpha}{2}i(-i\Omega^2)^3(\det\tilde{E}_i^a)m^2\phi^2,$$
$$\tilde{\mathcal{G}}_i:=-\mathcal{D}_a(-i\Omega^2\tilde{E}_i^a), \tag{22}$$
$$\tilde{\mathcal{V}}_a:=-(-i\Omega^2\tilde{E}_i^b)F^i_{ba}+\tilde{\pi}_\phi\partial_a\phi.$$

Step (i) in the construction consists of splitting each one of the fields involved in (22) into their real and imaginary parts,

$$\tilde{E}^{ai}=\tilde{e}^{ai}+i\tilde{\mathcal{E}}^{ai},\quad A_{ai}=M_{ai}+iV_{ai},\quad \phi=\phi_1+i\phi_2,\quad \tilde{\pi}_\phi=\tilde{\pi}_1+i\tilde{\pi}_2. \tag{23}$$

Now the key point, implemented as part of step (ii), is to promote each one of the real and imaginary parts to *independent* variables, which implies that the enlarged phase space has $2[4\times9+4]=80$ degrees of freedom per point on Σ. Next, one has to determine the corresponding momenta, which results in the following constraints:

$$\Phi_{\mathcal{E}ai}:=\Pi_{\mathcal{E}ai},\quad \Phi_{eai}:=\Pi_{eai},\quad \Phi_M^{ai}:=\Pi_M^{ai}+i\Omega_1^2\tilde{e}^{ai},\quad \Phi_V^{ai}:=\Pi_V^{ai}-\Omega_1^2\tilde{e}^{ai},$$
$$\Phi_{\pi_1}:=\Pi_{\pi_1},\quad \Phi_{\pi_2}:=\Pi_{\pi_2},\quad \Phi_{\phi_1}:=\Pi_{\phi_1}-\tilde{\pi}_1,\quad \Phi_{\phi_2}:=\Pi_{\phi_2}-i\tilde{\pi}_1, \tag{24}$$

denoted generically by Π. Note that Π_M^{ai} and Π_{ϕ_2} are purely imaginary, i.e., there are 40 primary constraints arising from the definition of momenta. The corresponding symplectic structure is

$$\{\tilde{\mathcal{E}}^{ai}(x),\Pi_{\mathcal{E}bj}(y)\}:=\delta_b^a\delta_j^i\delta^3(\tilde{x},y),\quad \{\tilde{e}^{ai}(x),\Pi_{ebj}(y)\}:=\delta_b^a\delta_j^i\delta^3(\tilde{x},y),$$

$$\{M_{ai}(x),\Pi_M^{bj}(y)\}:=\delta_a^b\delta_i^j\delta^3(x,\tilde{y}),\quad \{V_{ai}(x),\Pi_V^{bj}(y)\}:=\delta_a^b\delta_i^j\delta^3(x,\tilde{y}),$$

$$\{\phi_1(x),\Pi_{\phi_1}(y)\}:=\delta^3(x,\tilde{y}),\quad \{\phi_2(x),\Pi_{\phi_2}(y)\}:=\delta^3(x,\tilde{y}),$$

$$\{\tilde{\pi}_1(x),\Pi_{\pi_1}(y)\}:=\delta^3(\tilde{x},y),\quad \{\tilde{\pi}_2(x),\Pi_{\pi_2}(y)\}:=\delta^3(\tilde{x},y). \tag{25}$$

As for step (iii), the reality conditions are chosen here as a generalization of (4), i.e.,

$$\Psi_{\mathcal{E}}^{ai}:=\tilde{\mathcal{E}}^{ai},\quad \Psi_{Mai}:=M_{ai}-\Gamma_{ai}(\tilde{e})+\frac{1}{2}\epsilon_{ij}^{k}\underset{\sim}{e}_a^j\tilde{e}_k^c\frac{1}{\Omega_1^2}\partial_c\Omega_1^2,$$

$$\Psi_{\phi_2}:=\phi_2,\quad \Psi_{\pi_2}:=\tilde{\pi}_2, \tag{26}$$

where $\Gamma_{ai}(e)$ is the three-dimensional spin connection. Again, $\Psi_{\mathcal{E}}^{ai}$ plays the role of enforcing $\tilde{E}^{ai}$ to be real and Ψ_{Mai} keeps $\tilde{E}^{ai}$ real upon evolution. The term $\frac{1}{2}\epsilon_{ij}^{k}\underset{\sim}{e}_a^j\tilde{e}_k^c(1/\Omega_1^2)\partial_c\Omega_1^2$ is the *real* contribution of matter to the full connection A_a^i: it is determined as the real term of the matter contribution to the spatial part of ${}^+A$, upon variation of (11). Also, $\Psi_{\phi_2},\Psi_{\pi_2}$ constrain the scalar field, ϕ, to the real sector. At this stage there are 60 primary constraints, 20 of which arise from the reality conditions (26).

In the next step (iv), it is necessary to preserve upon evolution the full set of primary constraints (24) and (26). Before doing so, it is convenient to redefine some of them, Φ_{eai},Φ_{ϕ_1} and Φ_{π_1}, as

$$\Phi'_{eai}(x):=\Phi_{eai}(x)+\lambda_{aibj}(x,z)\Phi_M^{bj}(z)+\varepsilon_{ai}^{bj}(x,z)\Psi_{Mbj}(z)+\vartheta_{aibj}(x,z)\Phi_V^{bj}(z),$$

$$\Phi'_{\pi_1}:=\Phi_{\pi_1}-i\Psi_{\phi_2}, \tag{27}$$

$$\Phi'_{\phi_1}(x):=\Phi_{\phi_1}(x)+\mathcal{A}_{ai}(x,z)\Phi_M^{ai}(z)+\mathcal{B}^{ai}(x,z)\Psi_{Mai}(z)+\mathcal{C}_{ai}(x,z)\Phi_V^{ai}(z),$$

where

$$\lambda_{aibj}(x,y)=-\frac{\delta\Psi_{Mbj}(y)}{\delta\tilde{e}^{ai}(x)},\quad \varepsilon_{ai}^{bj}(x,y)=i\Omega_1^2(y)\delta_a^b\delta_i^j\delta^3(x,\tilde{y}),$$

$$\vartheta_{aibj}(x,y)=-i\frac{\delta\Psi_{Mbj}(y)}{\delta\tilde{e}^{ai}(x)},$$

$$\mathcal{A}_{ai}(x,y)=-\frac{\delta\Psi_{Mai}(y)}{\delta\phi_1(x)},\quad \mathcal{B}^{ai}(x,y)=i\frac{\delta\Omega_1^2(y)}{\delta\phi_1(x)}\tilde{e}^{ai}(y),$$

$$\mathcal{C}_{ai}(x,y)=-i\frac{\delta\Psi_{Mai}(y)}{\delta\phi_1(x)}. \tag{28}$$

Here, the Einstein summation convention for dummy indices is extended to the continuous case in such a way that it includes an implicit three-dimensional integral for the repeated z variable. Thus, for instance, the term $\lambda_{aibj}(x,z)\Phi_M^{bj}(z)$ means $\int_\Sigma d^3z\lambda_{aibj}(x,z)\Phi_M^{bj}(z)$.

As the first step in classifying the constraints into first and second class one computes the Poisson brackets of the constraints $\Phi_{\mathcal{E}ai}(x)$, $\Psi^{ai}_{\mathcal{E}}(x)$, $\Phi^{ai}_{M}(x)$, $\Psi_{Mai}(x)$, $\Phi_{\phi_2}(x)$, $\Psi_{\phi_2}(x)$, $\Phi_{\pi_2}(x)$, $\Psi_{\pi_2}(x)$, $\Phi'_{\phi_1}(x)$, $\Phi^{ai}_{V}(x)$, $\Phi'_{eai}(x)$ and $\Phi'_{\pi_1}(x)$. In the standard notation, the result can be expressed as

$$D=\begin{pmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{pmatrix}, \tag{29}$$

where

$$A=\begin{array}{c|cccc} & \Phi_{\mathcal{E}bj}(y) & \Psi^{bj}_{\mathcal{E}}(y) & \Phi^{bj}_{M}(y) & \Psi_{Mbj}(y) \\ \Phi_{\mathcal{E}ai}(x) & 0 & -\delta^b_a\delta^j_i\delta^3(x,\tilde{y}) & 0 & 0 \\ \Psi^{ai}_{\mathcal{E}}(x) & \delta^a_b\delta^i_j\delta^3(\tilde{x},y) & 0 & 0 & 0 \\ \Phi^{ai}_{M}(x) & 0 & 0 & 0 & -\delta^a_b\delta^i_j\delta^3(\tilde{x},y) \\ \Psi_{Mai}(x) & 0 & 0 & \delta^b_a\delta^j_i\delta^3(x,\tilde{y}) & 0 \end{array}, \tag{30}$$

$$B=\begin{array}{c|cccc} & \Phi_{\phi_2}(y) & \Psi_{\phi_2}(y) & \Phi_{\pi_2}(y) & \Psi_{\pi_2}(y) \\ \Phi_{\phi_2}(x) & 0 & -\delta^3(\tilde{x},y) & 0 & 0 \\ \Psi_{\phi_2}(x) & \delta^3(x,\tilde{y}) & 0 & 0 & 0 \\ \Phi_{\pi_2}(x) & 0 & 0 & 0 & -\delta^3(x,\tilde{y}) \\ \Psi_{\pi_2}(x) & 0 & 0 & \delta^3(\tilde{x},y) & 0 \end{array}, \tag{31}$$

$$C=\begin{array}{c|cccc} & \Phi'_{\phi_1}(y) & \Phi^{bj}_{V}(y) & \Phi'_{ebj}(y) & \Phi'_{\pi_1}(y) \\ \Phi'_{\phi_1}(x) & 0 & \dfrac{\delta\Omega^2_1(y)}{\delta\phi_1(x)}\tilde{e}^{bj}(y) & 0 & -\delta^3(\tilde{x},y) \\ \Phi^{ai}_{V}(x) & -\dfrac{\delta\Omega^2_1(x)}{\delta\phi_1(y)}\tilde{e}^{ai}(x) & 0 & -\Omega^2_1(x)\,\delta^a_b\delta^i_j\delta^3(\tilde{x},y) & 0 \\ \Phi'_{eai}(x) & 0 & \Omega^2_1(y)\,\delta^b_a\delta^j_i\delta^3(x,\tilde{y}) & 0 & 0 \\ \Phi'_{\pi_1}(x) & \delta^3(x,\tilde{y}) & 0 & 0 & 0 \end{array}. \tag{32}$$

From the structure of the matrices (29)–(32) one concludes that the set of constraints arising from the definition of the momenta, together with those which come from the reality conditions are *second class*. On the other hand, the original set of constraints $\tilde{\tilde{C}}$, $\tilde{\mathcal{G}}_i$, $\tilde{\mathcal{V}}_a$, in the complex theory, were the generators of the gauge symmetry of the system. To show that this property remains so in the present real theory, one begins by redefining them in such a way that they have zero Poisson brackets with the second class constraints. Then, it should be verified that this redefinition preserves the first class character of the algebra among them. Recall that $\tilde{\tilde{C}}$, $\tilde{\mathcal{G}}_i$ and $\tilde{\mathcal{V}}_a$ depend only upon configuration variables (See (15) and (23)). Denoting any of them by $\mathcal{R}$, we find that the appropriate redefinition is

$$\mathcal{R}'(x) := \mathcal{R}(x) + I^{bj}(x,z)\Psi_{Mbj}(z) + N_{bj}(x,z)\Phi_V^{bj}(z) + O^{bj}(x,z)\Phi'_{ebj}(z) + H_{bj}(x,z)\Psi_{\mathcal{E}}^{bj}(z) + P(x,z)\Psi_{\pi_2}(z) + Q(x,z)\Psi_{\phi_2}(z) + R(x,z)\Phi'_{\phi_1}(z) + S(x,z)\Phi'_{\pi_1}(z), \tag{33}$$

where

$$I^{ai}(x,y) = \{\Phi_M^{ai}(y), \mathcal{R}(x)\}, \quad H_{ai}(x,y) = \{\Phi_{\mathcal{E}ai}(y), \mathcal{R}(x)\},$$

$$N_{ai}(x,y) = -\frac{1}{\Omega_1^2(y)}\{\Phi'_{eai}(y), \mathcal{R}(x)\}, \quad R(x,y) = -\{\Phi'_{\pi_1}(y), \mathcal{R}(x)\},$$

$$O^{ai}(x,y) = \frac{1}{\Omega_1^2(y)}\left[\{\Phi_V^{ai}(y), \mathcal{R}(x)\} + \frac{\delta\Omega_1^2(y)}{\delta\phi_1(z)}\tilde{e}^{ai}(y)\{\Phi'_{\pi_1}(z), \mathcal{R}(x)\}\right],$$

$$P(x,y) = \{\Phi_{\pi_2}(y), \mathcal{R}(x)\}, \quad Q(x,y) = \{\Phi_{\phi_2}(y), \mathcal{R}(x)\},$$

$$S(x,y) = \{\Phi'_{\phi_1}(y), \mathcal{R}(x)\} - \frac{1}{\Omega_1^2(z)} \frac{\delta\Omega_1^2(z)}{\delta\phi_1(y)}\tilde{e}^{bj}(z)\{\Phi'_{ebj}(z), \mathcal{R}(x)\}. \tag{34}$$

The Poisson brackets between any pair $\mathcal{Q}'$ and $\mathcal{R}'$ is weakly zero. This can be shown by calculating

$$\begin{aligned}\{\mathcal{Q}'(x), \mathcal{R}'(y)\} &\approx \{\mathcal{Q}'(x), \mathcal{R}(y)\} \\ &= -\frac{1}{\Omega_1^2(z)}\{\Phi'_{eai}(z), \mathcal{Q}(x)\}\{\Phi_V^{ai}(z), \mathcal{R}(y)\} + \frac{1}{\Omega_1^2(z)}\{\Phi_V^{ai}(z), \mathcal{Q}(x)\} \\ &\quad \times\{\Phi'_{eai}(z), \mathcal{R}(y)\} - \{\Phi'_{\pi_1}(z), \mathcal{Q}(x)\}\{\Phi'_{\phi_1}(z), \mathcal{R}(y)\} + \{\Phi'_{\phi_1}(z), \mathcal{Q}(x)\} \\ &\quad \times\{\Phi'_{\pi_1}(z), \mathcal{R}(y)\} + \frac{1}{\Omega_1^2(z)} \frac{\delta\Omega_1^2(z)}{\delta\phi_1(\omega)}\tilde{e}^{ai}(z)\{\Phi'_{\pi_1}(\omega), \mathcal{Q}(x)\}\{\Phi'_{eai}(z), \mathcal{R}(y)\} \\ &\quad - \frac{1}{\Omega_1^2(\omega)} \frac{\delta\Omega_1^2(\omega)}{\delta\phi_1(z)}\tilde{e}^{ai}(\omega)\{\Phi'_{eai}(\omega), \mathcal{Q}(x)\}\{\Phi'_{\pi_1}(z), \mathcal{R}(y)\}. \end{aligned} \tag{35}$$

The above result was obtained by substituting (33) together with the fact that $\mathcal{Q}(x)$, $\mathcal{R}(y)$, $\Psi_M, \Psi_{\mathcal{E}}, \Psi_{\pi_2}$ and Ψ_{ϕ_2} are independent of the momenta. Now, by using the explicit form of $\Phi'_{eai}, \Phi'_{\phi_1}$ and Φ'_{π_1} one finds, after a long (but otherwise direct) calculation

$$\begin{aligned}\{\mathcal{Q}'(x), \mathcal{R}'(y)\} \approx &-\frac{1}{\Omega_1^2(z)}\{\Phi_{eai}(z), \mathcal{Q}(x)\}\{\Phi_V^{ai}(z), \mathcal{R}(y)\} \\ &+ \frac{1}{\Omega_1^2(z)}\{\Phi_V^{ai}(z), \mathcal{Q}(x)\}\{\Phi_{eai}(z), \mathcal{R}(y)\} \\ &- \{\Phi_{\pi_1}(z), \mathcal{Q}(x)\}\{\Phi_{\phi_1}(z), \mathcal{R}(y)\} + \{\Phi_{\phi_1}(z), \mathcal{Q}(x)\}\{\Phi_{\pi_1}(z), \mathcal{R}(y)\} \\ &+ \frac{1}{\Omega_1^2(z)} \frac{\delta\Omega_1^2(z)}{\delta\phi_1(\omega)}\tilde{e}^{ai}(z)\{\Phi_{\pi_1}(\omega), \mathcal{Q}(x)\}\{\Phi_{eai}(z), \mathcal{R}(y)\}\end{aligned}$$

$$-\frac{1}{\Omega_1^2(\omega)}\frac{\delta\Omega_1^2(\omega)}{\delta\phi_1(z)}\widetilde{e}^{ai}(\omega)\{\Phi_{eai}(\omega),\mathcal{Q}(x)\}\{\Phi_{\pi_1}(z),\mathcal{R}(y)\}. \tag{36}$$

The above expression is most suitably calculated in terms of the *original* complex phase space variables. The symplectic structure (16) yields

$$\{A_a^i(x),\widetilde{E}_j^b(y)\}=\left\{A_a^i(x),\frac{i\widetilde{\pi}_j^b(y)}{\Omega^2(y)}\right\}=\frac{i}{\Omega^2(y)}\delta_a^b\delta_j^i\delta^3(x,\widetilde{y}), \tag{37}$$

$$\{\widetilde{E}_i^a(x),\widetilde{\pi}_\phi(y)\}=\left\{\frac{i\widetilde{\pi}_i^a(x)}{\Omega^2(x)},\widetilde{\pi}_\phi(y)\right\}=-\frac{1}{\Omega^2(x)}\widetilde{E}_i^a(x)\frac{\delta\Omega^2(x)}{\delta\phi(y)}. \tag{38}$$

Upon substitution of these expressions in (36) one obtains

$$\begin{aligned}\{\mathcal{Q}'(x),\mathcal{R}'(y)\}\approx&\{A_a^i(z),\widetilde{E}_j^b(\omega)\}\left(\frac{\delta\mathcal{Q}(x)}{\delta A_a^i(z)}\frac{\delta\mathcal{R}(y)}{\delta\widetilde{E}_i^b(z)}-\frac{\delta\mathcal{Q}(x)}{\delta\widetilde{E}_j^b(z)}\frac{\delta\mathcal{R}(y)}{\delta A_a^i(z)}\right)\\&+\{\phi(z),\widetilde{\pi}_\phi(\omega)\}\left(\frac{\delta\mathcal{Q}(x)}{\delta\phi(z)}\frac{\delta\mathcal{R}(y)}{\delta\widetilde{\pi}_\phi(z)}-\frac{\delta\mathcal{Q}(x)}{\delta\widetilde{\pi}_\phi(z)}\frac{\delta\mathcal{R}(y)}{\delta\phi(z)}\right)\\&+\left\{\widetilde{E}_i^a(z),\{\widetilde{\pi}_\phi(\omega)\}\left(\frac{\delta\mathcal{Q}(x)}{\delta\widetilde{E}_i^a(z)}\frac{\delta\mathcal{R}(y)}{\delta\widetilde{\pi}_\phi(\omega)}-\frac{\delta\mathcal{Q}(x)}{\delta\widetilde{\pi}_\phi(\omega)}\frac{\delta\mathcal{R}(y)}{\delta\widetilde{E}_i^a(z)}\right)\right.\\&\approx\{\mathcal{Q}(x),\mathcal{R}(y)\}_{(A,\pi),(\phi,\pi_\phi)}\approx 0.\end{aligned} \tag{39}$$

In the last line, the Poisson brackets are taken with respect to the original complex symplectic structure (16). Therefore, it has been shown that the Poisson brackets between any pair of constraints $\tilde{\tilde{\mathrm{C}}}'$, $\widetilde{\mathcal{V}}'$, $\widetilde{\mathcal{G}}'$ are weakly zero.

Thus, the system is described by the following set of primary constraints: $\Phi_{\mathcal{E}ai}(x)$, $\Psi_{\mathcal{E}}^{ai}(x)$, $\Phi_M^{ai}(x)$, $\Psi_{Mai}(x)$, $\Phi_{\phi_2}(x)$, $\Psi_{\phi_2}(x)$, $\Phi_{\pi_2}(x)$, $\Psi_{\pi_2}(x)$, $\Phi'_{\phi_1}(x)$, $\Phi_V^{ai}(x)$, $\Phi'_{eai}(x)$, $\Phi'_{\pi_1}(x)$, $\tilde{\tilde{\mathrm{C}}}'$, $\widetilde{\mathcal{V}}'$ and $\widetilde{\mathcal{G}}'$. Now, following the Dirac method, the time conservation of the constraints is imposed using the Hamiltonian density

$$\begin{aligned}\mathcal{H}=&\mu^{\mathcal{E}ai}\Phi_{\mathcal{E}ai}+\mu_{ai}^{\mathcal{E}}\Psi_{\mathcal{E}}^{ai}+\mu_{ai}^{M}\Phi_M^{ai}+\mu^{Mai}\Psi_{Mai}+\mu^{\phi_2}\Phi_{\phi_2}+\nu^{\phi_2}\Psi_{\phi_2}\\&+\mu^{\pi_2}\Phi_{\pi_2}+\nu^{\pi_2}\Psi_{\pi_2}+\mu^{\phi_1}\Phi'_{\phi_1}+\mu_{ai}^{V}\Phi_V^{ai}+\mu^{eai}\Phi'_{eai}+\mu^{\pi_1}\Phi'_{\pi_1}\\&+\underset{\sim}{N}\tilde{\tilde{C}}'+N^a\widetilde{\mathcal{V}'_a}+N^i\widetilde{\mathcal{G}'_i},\end{aligned} \tag{40}$$

where no three-dimensional integral is involved.

From Eqs. (33) and (39) one finds that the Poisson brackets between $\tilde{\tilde{\mathrm{C}}}'$, $\widetilde{\mathcal{V}'}$, $\widetilde{\mathcal{G}'}$ and $H=\int_\Sigma d^3x\mathcal{H}(x)$ are weakly zero. Finally, since the set of constraints $\Phi_{\mathcal{E}ai}(x)$, $\Psi_{\mathcal{E}}^{ai}(x)$, $\Phi_M^{ai}(x)$, $\Psi_{Mai}(x)$, $\Phi_{\phi_2}(x)$, $\Psi_{\phi_2}(x)$, $\Phi_{\pi_2}(x)$, $\Psi_{\pi_2}(x)$, $\Phi'_{\phi_1}(x)$, $\Phi_V^{ai}(x)$, $\Phi'_{eai}(x)$ and $\Phi'_{\pi_1}(x)$ is second class, the Lagrange multipliers $\mu^{\mathcal{E}^{ai}}$, $\mu_{ai}^{\mathcal{E}}$, μ_{ai}^{M}, $\mu^{M^{ai}}$, μ^{ϕ_2}, ν^{ϕ_2}, μ^{π_2}, ν^{π_2}, μ^{ϕ_1}, $\mu^{V}{}_{ai}$, μ^{eai}, μ^{π_1} are determined, and shown to be zero. In other words, there are no secondary constraints and the total Hamiltonian density is given by

$$\mathcal{H}_{\mathrm{Total}}=\underset{\sim}{N}\tilde{\tilde{C}}'+N^a\widetilde{\mathcal{V}'}_a+N^i\widetilde{\mathcal{G}'}_i, \tag{41}$$

which is a combination of the first class constraints only.

Now, let us count the physical degrees of freedom in terms of the real variables that we have introduced. Recall that the enlarged phase space, with configuration variables (23), has dimension $2\times9\times4+2\times4=80$ per space point. Since there are $6\times9+6\times1=60$ second class constraints (24) and (26), the partially reduced phase space has dimension 20. The corresponding partially reduced symplectic structure can be obtained by using Dirac brackets. To this end, the inverse of the second class constraints Poisson brackets matrix (29) is needed. Its calculation produces

$$D^{-1}=\begin{pmatrix} A^{-1} & 0 & 0 \\ 0 & B^{-1} & 0 \\ 0 & 0 & C^{-1} \end{pmatrix}, \tag{42}$$

where

$$A^{-1}=\begin{pmatrix} 0 & \delta^a_b\delta^i_j\delta^3(\tilde{x},y) & 0 & 0 \\ -\delta^b_a\delta^j_i\delta^3(x,\tilde{y}) & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta^b_a\delta^j_i\delta^3(x,\tilde{y}) \\ 0 & 0 & -\delta^a_b\delta^i_j\delta^3(\tilde{x},y) & 0 \end{pmatrix}, \tag{43}$$

$$B^{-1}=\begin{pmatrix} 0 & \delta^3(x,\tilde{y}) & 0 & 0 \\ -\delta^3(\tilde{x},y) & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta^3(\tilde{x},y) \\ 0 & 0 & -\delta^3(x,\tilde{y}) & 0 \end{pmatrix}, \tag{44}$$

$$C^{-1}=\begin{pmatrix} 0 & 0 & 0 & \delta^3(x,\tilde{y}) \\ 0 & 0 & \frac{1}{\Omega_1^2(y)}\delta^a_b\delta^i_j\delta^3(x,\tilde{y}) & 0 \\ 0 & -\frac{1}{\Omega_1^2(x)}\delta^b_a\delta^j_i\delta^3(\tilde{x},y) & 0 & -\frac{1}{\Omega_1^2(x)}\frac{\delta\Omega_1^2(x)}{\delta\phi_1(y)}e^{bj}(x) \\ -\delta^3(\tilde{x},y) & 0 & \frac{1}{\Omega_1^2(y)}\frac{\delta\Omega_1^2(y)}{\delta\phi_1(x)}e^{ai}(y) & 0 \end{pmatrix}. \tag{45}$$

The use of (42)–(45), leads to

$$\begin{aligned} \{f(x),g(y)\}^* = {} & +\{f(x),g(y)\} \\ & -\{f(x),\Phi_{\mathcal{E}ai}(z)\}\{\Psi^{ai}_{\mathcal{E}}(z),g(y)\}+\{f(x),\Psi^{ai}_{\mathcal{E}}(z)\}\{\Phi_{\mathcal{E}ai}(z),g(y)\} \\ & -\{f(x),\Phi^{ai}_M(z)\}\{\Psi_{Mai}(z),g(y)\}+\{f(x),\Psi_{Mai}(z)\}\{\Phi^{ai}_M(z),g(y)\} \\ & -\{f(x),\Phi_{\phi_2}(z)\}\{\Psi_{\phi_2}(z),g(y)\}+\{f(x),\Psi_{\phi_2}(z)\}\{\Phi_{\phi_2}(z),g(y)\} \\ & -\{f(x),\Phi_{\pi_2}(z)\}\{\Psi_{\pi_2}(z),g(y)\}+\{f(x),\Psi_{\pi_2}(z)\}\{\Phi_{\pi_2}(z),g(y)\} \\ & -\{f(x),\Phi'_{\phi_1}(z)\}\{\Phi'_{\pi_1}(z),g(y)\}+\{f(x),\Phi'_{\pi_1}(z)\}\{\Phi'_{\phi_1}(z),g(y)\} \end{aligned}$$

$$-\{f(x),\Phi_V^{ai}(z)\}\frac{1}{\Omega_1^2(z)}\{\Phi'_{eai}(z),g(y)\}+\{f(x),\Phi'_{eai}(z)\}\frac{1}{\Omega_1^2(z)}\{\Phi_V^{ai}(z),g(y)\}$$
$$-\{f(x),\Phi'_{\pi_1}(z)\}\left[\frac{1}{\Omega_1^2(w)}\ \frac{\delta\Omega_1^2(\omega)}{\delta\phi_1(z)}e^{ai}(w)\right]\{\Phi'_{eai}(w),g(y)\}$$
$$+\{f(x),\Phi'_{eai}(z)\}\left[\frac{1}{\Omega_1^2(z)}\ \frac{\delta\Omega_1^2(z)}{\delta\phi_1(\omega)}e^{ai}(z)\right]\{\Phi'_{\pi_1}(\omega),g(y)\}, \tag{46}$$

for the Dirac brackets of any two functions f and g on the enlarged phase space. Finally, upon partial reduction, the canonical variables are V_a^i, $\pi_j^b=-\Omega_1^2 e_j^b$, ϕ_1, π_1 and the reduced symplectic structure is just

$$\{V_a^i(x),\pi_j^b(y)\}^*=\delta_a^b\delta^i_j\delta^3(x,\tilde{y}),$$
$$\{\phi_1(x),\pi_1(y)\}^*=\delta^3(x,\tilde{y}). \tag{47}$$

In the same way as in the pure gravity case,[5] the first class constraints (15) turn out to be either purely real or purely imaginary in the above partially reduced phase space. Then, the physical phase space has dimension $20-2\times7=6$, as expected for a real scalar field coupled to real gravity.

V. CONCLUSIONS AND PERSPECTIVES

Previous results yielding the identification in phase space of the real sector of pure complex gravity[5] have been successfully extended in this work to incorporate the case of a scalar field nonminimally coupled to gravity, starting from (complex) Ashtekar variables. This provides further support for the general validity of the method proposed.

The procedure is as follows: the complex canonical variables are splitted into real and imaginary parts, each of which is taken as an independent new configuration variable. The corresponding momenta are subsequently defined from the action, leading to primary constraints. The real sector of the theory is next defined by introducing appropriate reality conditions in the form of additional primary constraints. This is possible because the original phase space has been extended. The whole set of constraints is next classified into first and second class, after imposing the conservation of the primary constraints upon evolution. Finally, one faces the problem of how to conveniently deal with the resulting second class constraints, which include the reality conditions.

The advantages of our approach are: (i) Reality conditions are incorporated as true second class constraints within the canonical description of an extended phase space, uniquely associated to each physical system. (ii) It leads to the standard Dirac's method of counting the real physical degrees of freedom arising from an originally complex theory. (iii) Although we start from a pair of reality conditions (4), only the first is truly an input, because the second condition appears as the consequence of demanding the conservation of the former upon evolution.

As opposed to Ref. 10, we have presented here a theory for a scalar field nonminimally coupled to gravity, leading to polynomial constraints, using Ashtekar variables. Unfortunately, the nonpolynomiality shows up after implementing the reality conditions, in the process of identifying the real sector of the theory. This happens either for the nonpolynomial form of the reality conditions (26), or for their polynomial realization. Recently, however, certain nonpolynomial constraints have been shown to be tractable in the quantum theory.[3,11,12] Interestingly enough, in our case the whole nonpolynomiality is encoded in the single function Γ_{ai} appearing in (26). It certainly remains an open problem to determine whether or not the results presented here may provide a tractable alternative to deal with the quantum situation.

The use of Dirac brackets, which is the standard way of eliminating the second class constraints, yields the expected real nonpolynomial form of the theory. For example, it leads to the Palatini canonical form in the case of pure gravity.[4] To explore an alternative preventing the use of Dirac brackets in the pure gravity case, we have implemented the conversion of the full set of second class constraints into a first class set, following the method of Ref. 7. Thus, we have rewritten pure real gravity as a theory involving an alternative set of first class constraints, which, for example, has not been previously done starting from the Palatini formulation with second class constraints. However, their physical meaning, together with their usefulness in a quantum theory still needs to be clarified. The method of Ref. 7 works whenever the Poisson brackets matrix of the original second class constraints is independent of the phase space variables. This is indeed the case for pure gravity, but not for the scalar field nonminimally coupled to gravity considered in this work. Hence, the application of the same strategy to the latter theory would first require an extension of the method in Ref. 7 .

ACKNOWLEDGMENTS

Partial support is acknowledged from CONACyT Grant No. 3141P–E9608 and UNAM–DGAPA–IN100397. M.M. thanks all the members of the relativity group of the Department of Physics and Astronomy of the University of Pittsburgh for their warm hospitality. M.M.'s post-doctoral fellowship is funded through the CONACyT of México, fellow number 91825. Also M.M. thanks support from the *Sistema Nacional de Investigadores*.

[1] C. Rovelli, living reviews in relativity, http://www.livingreviews.org/Articles/Volume1/1998-1rovelli (1998); Strings, loops and others: a critical survey of the present approaches to quantum gravity, Plenary lecture on quantum gravity at the GR15 conference, Pune, India; gr-qc archives No. 9803024.
[2] A. Ashtekar, C. Rovelli, and L. Smolin, Phys. Rev. Lett. **69**, 237 (1992); C. Rovelli and L. Smolin, Nucl. Phys. B **442**, 593 (1995); **456**, 753(E) (1995); A. Ashtekar and J. Lewandowski, Class. Quantum Grav. **14**, A55 (1997).
[3] T. Thiemann, Phys. Lett. B **380**, 257 (1996); Class. Quantum Grav. **15**, 839 (1998).
[4] A. Ashtekar, Phys. Rev. Lett. **77**, 3288 (1986); Phys. Rev. D **36**, 1587 (1987); *Lectures on Nonperturbative Canonical Gravity (Notes prepared in collaboration with R.S. Tate)* (World Scientific, Singapore, 1991).
[5] H. A. Morales-Técotl, L. F. Urrutia, and J. D. Vergara, Class. Quantum Grav. **13**, 2933 (1996).
[6] I. A. Batalin and I. V. Tyutin, Int. J. Mod. Phys. A **6**, 3255 (1991).
[7] R. Amorim and A. Das, Mod. Phys. Lett. A **9**, 3543 (1994).
[8] C. Kiefer, Phys. Lett. B **225**, 227 (1989).
[9] F. W. Hehl, J. D. McCrea, E. W. Mielke, and Y. Ne'eman, Phys. Rep. **258**, 1 (1995).
[10] R. Capovilla, Phys. Rev. D **46**, 1450 (1992).
[11] J. F. Barbero, Phys. Rev. D **51**, 5507 (1995).
[12] J. F. Barbero, Phys. Rev. D **49**, 6935 (1994); **51**, 5498 (1995).

Discrete Riemannian geometry

Aristophanes Dimakis
Department of Mathematics, University of the Aegean,
GR-83200 Karlovasi, Samos, Greece

Folkert Müller-Hoissen[a)]
Max-Planck-Institut für Strömungsforschung, Bunsenstrasse 10,
D-37073 Göttingen, Germany

(Received 16 September 1998; accepted 23 November 1998)

Within a framework of noncommutative geometry, we develop an analog of (pseudo-) Riemannian geometry on finite and discrete sets. On a finite set, there is a counterpart of the continuum metric tensor with a simple geometric interpretation. The latter is based on a correspondence between first order differential calculi and digraphs (the vertices of the latter are given by the elements of the finite set). Arrows originating from a vertex span its (co)tangent space. If the metric is to measure length and angles at some point, it has to be taken as an element of the *left-linear* tensor product of the space of 1-forms with itself, and not as an element of the (nonlocal) tensor product over the algebra of functions, as considered previously by several authors. It turns out that linear connections can always be extended to this left tensor product, so that metric compatibility can be defined in the same way as in continuum Riemannian geometry. In particular, in the case of the universal differential calculus on a finite set, the Euclidean geometry of polyhedra is recovered from conditions of metric compatibility and vanishing torsion. In our rather general framework (which also comprises structures which are far away from continuum differential geometry), there is, in general, nothing like a Ricci tensor or a curvature scalar. Because of the nonlocality of tensor products (over the algebra of functions) of forms, corresponding components (with respect to some module basis) turn out to be rather nonlocal objects. But one can make use of the parallel transport associated with a connection to ''localize'' such objects, and in certain cases there is a distinguished way to achieve this. In particular, this leads to covariant components of the curvature tensor which allow a contraction to a Ricci tensor. Several examples are worked out to illustrate the procedure. Furthermore, in the case of a differential calculus associated with a hypercubic lattice we propose a new discrete analogue of the (vacuum) Einstein equations. © *1999 American Institute of Physics.* [S0022-2488(99)00303-5]

I. INTRODUCTION

In a series of papers[1–5] we have developed a formalism of differential geometry on finite and discrete sets with applications, in particular, to lattice gauge theory[6] and discrete completely integrable models.[7]

The most basic ''differential geometric'' structure on a discrete set $\mathcal{M}$ is a *differential calculus* $(\Omega(\mathcal{M}),\mathrm{d})$, where $\Omega(\mathcal{M})=\oplus_{r\geqslant 0}\Omega^r(\mathcal{M})$ is an analog of the algebra of differential forms on a differentiable manifold and the $\mathbb{C}$-linear map $\mathrm{d}:\Omega^r(\mathcal{M})\rightarrow\Omega^{r+1}(\mathcal{M})$ generalizes the exterior derivative. Here $\mathcal{A}:=\Omega^0(\mathcal{M})$ is the algebra of $\mathbb{C}$-valued functions on $\mathcal{M}$ and noncommutativity enters the stage via nontrivial commutation relations between functions and differentials [which are elements of $\Omega^1(\mathcal{M})$]. On a discrete set there are many choices of a (first order) differential

[a)]Electronic mail: fmuelle@gwdg.de

calculus and it turned out[3] that these amount to the selection of a digraph structure and thus neighborhood relations on the discrete set.

Whereas the concept of a connection seems to be well understood in the framework of noncommutative geometry, this is not quite so for the concept of a metric. In Connes' approach to noncommutative geometry,[8] Riemannian geometry is encoded in a self-adjoint operator on a Hilbert space and recovered from it via a formula for the distance of two points. The distance formula is then generalized to a more abstract setting, including the case of discrete sets (see also Ref. 9 and references therein). A major problem with this approach is that it is bound to (generalizations of) positive definite metrics and thus at least not directly applicable to space–time geometry. The underlying philosophy of ''spectral geometry,'' namely that all geometrical data should be encoded in the spectrum of certain self-adjoint operators on a Hilbert space, is certainly very interesting but by no means compulsive.

In several papers (see Refs. 5, 10–12, for example) a metric in noncommutative geometry has been taken to be an element of the tensor product space $\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})$ with certain properties. Here $\Omega^1(\mathcal{A})$ is the space of 1-forms of a differential calculus $(\Omega(\mathcal{A}),\mathrm{d})$ over an associative algebra $\mathcal{A}$. This has just been a formal generalization of one of several, in classical differential geometry equivalent, definitions of a metric tensor field, motivated by simplicity of the mathematical structure, but without a deeper, e.g., physical, substantiation. Even on the technical level a serious problem showed up, namely, the extensibility of a (linear) connection on $\Omega^1(\mathcal{A})$ to a connection on $\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})$, which is necessary in order to define metric compatibility of a linear connection (see Refs. 5, 13 for discussions and related references).

Needless to say, generalizing another — classically equivalent — metric concept, one does not, in general, arrive at equivalent structures in the noncommutative geometric setting. In fact, motivated by previous work[6,7] we recently investigated in more detail generalizations of the Hodge $\star$-operator.[14] The metric is recovered from $(\alpha,\beta)=\star^{-1}(\alpha\star\beta)$ where α,β are differential 1-forms. For a symmetric Hodge operator on a (*non*commutative) differential calculus over a *commutative* algebra $\mathcal{A}$, contact was made with a metric defined as an element,

$$g\in\Omega^1(\mathcal{A})\otimes_L\Omega^1(\mathcal{A}), \tag{1.1}$$

and not as an element of the space $\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})$. The tensor product $\otimes_L$ satisfies

$$(f\alpha)\otimes_L(h\beta)=fh(\alpha\otimes_L\beta), \quad \forall f,h\in\mathcal{A},\ \alpha,\beta\in\Omega(\mathcal{A}). \tag{1.2}$$

In the following we show that it is precisely the latter metric definition which directly reproduces some familiar results in *discrete* geometry and which allows us to develop discrete noncommutative geometry to a more satisfactory level. It should be noticed, however, that the tensor product $\otimes_L$ and therefore the metric definition (1.1) does not generalize in an obvious way to *non*commutative algebras $\mathcal{A}$, at least as far as we can see. But in Ref. 14 we have generalized the associated Hodge operator to the general noncommutative framework.

In Sec. II we recall some basic definitions of noncommutative geometry. In Sec. III we concentrate on finite sets and introduce metrics and compatible linear connections on them. In Sec. IV we deal with a technical problem which has its origin in the nonlocality of the tensor product over $\mathcal{A}$. In particular, the construction of a Ricci tensor is addressed in our framework. As an example of particular interest, the geometry of a hypercubic lattice is treated in Sec. V. In Sec. VI we deal with discrete surfaces of revolution. Some conclusions are collected in Sec. VII. In particular, we propose a new discrete version of the Einstein equations on a hypercubic lattice.

II. PRELIMINARIES

In the first subsection we recall the definition of a differential calculus over an associative algebra. The second subsection contains the general definitions of linear connections, torsion and curvature in the framework of noncommutative geometry.

A. Differential calculi on associative algebras

Let $\mathcal{A}$ be an associative algebra over $\mathbb{C}$ with unit $\mathbf{1}$. A *differential calculus* over $\mathcal{A}$ is a $\mathbb{Z}$-graded associative algebra (over $\mathbb{C}$),

$$\Omega(\mathcal{A}) = \bigoplus_{r \geqslant 0} \Omega^r(\mathcal{A}), \tag{2.1}$$

where the spaces $\Omega^r(\mathcal{A})$ are $\mathcal{A}$-bimodules and $\Omega^0(\mathcal{A}) = \mathcal{A}$. There is a $\mathbb{C}$-linear map,

$$\mathrm{d}: \quad \Omega^r(\mathcal{A}) \to \Omega^{r+1}(\mathcal{A}) \tag{2.2}$$

with the following properties:

$$\mathrm{d}^2 = 0, \tag{2.3}$$

$$\mathrm{d}(ww') = (\mathrm{d}w)w' + (-1)^r w \mathrm{d}w', \tag{2.4}$$

where $w \in \Omega^r(\mathcal{A})$ and $w' \in \Omega(\mathcal{A})$. The last relation is known as the (generalized) *Leibniz rule*. One also requires $\mathbf{1}w = w\mathbf{1} = w$ for all elements $w \in \Omega(\mathcal{A})$. The identity $\mathbf{11} = \mathbf{1}$ then implies

$$\mathrm{d}\mathbf{1} = 0. \tag{2.5}$$

Furthermore, we require that d generates the spaces $\Omega^r(\mathcal{A})$ for $r > 0$ in the sense that $\Omega^r(\mathcal{A}) = \mathcal{A}\, \mathrm{d}\Omega^{r-1}(\mathcal{A})\mathcal{A}$. The space $\Omega^r(\mathcal{A})$, $r > 1$ can then be identified with a quotient of the r-fold tensor product $\Omega^1(\mathcal{A}) \otimes_{\mathcal{A}} \cdots \otimes_{\mathcal{A}} \Omega^1(\mathcal{A})$ by some sub-bimodule.

B. Linear connections, torsion and curvature

Let $(\Omega(\mathcal{A}), \mathrm{d})$ be a differential calculus over an associative algebra $\mathcal{A}$. A linear (left $\mathcal{A}$-module) connection is a $\mathbb{C}$-linear map $\boldsymbol{\nabla} : \Omega^1(\mathcal{A}) \to \Omega^1(\mathcal{A}) \otimes_{\mathcal{A}} \Omega^1(\mathcal{A})$ such that

$$\boldsymbol{\nabla}(f\alpha) = \mathrm{d}f \otimes_{\mathcal{A}} \alpha + f \boldsymbol{\nabla} \alpha. \tag{2.6}$$

A linear connection extends to a map $\nabla: \Omega(\mathcal{A}) \otimes_{\mathcal{A}} \Omega^1(\mathcal{A}) \to \Omega(\mathcal{A}) \otimes_{\mathcal{A}} \Omega^1(\mathcal{A})$ via

$$\boldsymbol{\nabla}(w \otimes_{\mathcal{A}} \alpha) = \mathrm{d}w \otimes_{\mathcal{A}} \alpha + (-1)^r w \boldsymbol{\nabla} \alpha, \quad \forall w \in \Omega^r(\mathcal{A}),\ \alpha \in \Omega^1(\mathcal{A}). \tag{2.7}$$

The *torsion* of a linear connection $\boldsymbol{\nabla}$ is the map $\Theta : \Omega^1(\mathcal{A}) \to \Omega^2(\mathcal{A})$, given by

$$\Theta(\alpha) := \mathrm{d}\alpha - \pi \circ \boldsymbol{\nabla} \alpha, \tag{2.8}$$

where π is the natural projection $\Omega^1(\mathcal{A}) \otimes_{\mathcal{A}} \Omega^1(\mathcal{A}) \to \Omega^2(\mathcal{A})$. It satisfies

$$\Theta(f\alpha) = f\Theta(\alpha). \tag{2.9}$$

The torsion extends to a map $\Theta : \Omega(\mathcal{A}) \otimes_{\mathcal{A}} \Omega^1(\mathcal{A}) \to \Omega(\mathcal{A})$ via

$$\Theta(w \otimes_{\mathcal{A}} \alpha) := \mathrm{d}(w\alpha) - \pi \circ \boldsymbol{\nabla}(w \otimes_{\mathcal{A}} \alpha), \quad \forall w \in \Omega(\mathcal{A}), \quad \alpha \in \Omega^1(\mathcal{A}), \tag{2.10}$$

where π now denotes more generally the projection $\Omega(\mathcal{A}) \otimes_{\mathcal{A}} \Omega^1(\mathcal{A}) \to \Omega(\mathcal{A})$. Then

$$\Theta(\boldsymbol{\nabla}\alpha) = \mathrm{d}\pi \circ \boldsymbol{\nabla}(\alpha) - \pi \circ \nabla^2(\alpha) = \mathrm{d}(\mathrm{d}\alpha - \Theta(\alpha)) + \pi \circ \mathfrak{R}(\alpha), \tag{2.11}$$

where we have introduced the *curvature* $\mathfrak{R}$ of $\boldsymbol{\nabla}$ as the map

$$\mathfrak{R} := -\boldsymbol{\nabla}^2, \tag{2.12}$$

which satisfies

$$\mathfrak{R}(f\alpha)=f\mathfrak{R}(\alpha). \tag{2.13}$$

We arrive at the *first Bianchi identity*,

$$\mathrm{d}\circ\Theta+\Theta\circ\boldsymbol{\nabla}=\pi\circ\mathfrak{R}. \tag{2.14}$$

The *second Bianchi identity* is

$$(\boldsymbol{\nabla}\mathfrak{R})(\alpha):=\boldsymbol{\nabla}(\mathfrak{R}(\alpha))-\mathfrak{R}(\boldsymbol{\nabla}\alpha)=-\nabla^3\alpha+\nabla^3\alpha=0. \tag{2.15}$$

Example. For the universal differential calculus, we have $\pi=\mathrm{id}$ on $\Omega^1\otimes_{\mathcal{A}}\Omega^1$ and there is a unique linear connection with vanishing torsion given by $\boldsymbol{\nabla}=\mathrm{d}$ according to (2.8). The curvature of this linear connection vanishes. ■

III. DIFFERENTIAL GEOMETRY ON FINITE SETS

In this section we collect some facts about differential calculi, vector fields and linear connections on finite sets (see also Refs. 2–5, 15, 16). We then consider metrics and elaborate on the metric compatibility condition for a linear connection.

A. First order differential calculi on a finite set

Let $\mathcal{M}$ be a finite set of N elements and $\mathcal{A}$ the algebra of all $\mathbb{C}$-valued functions on it. $\mathcal{A}$ is a complex linear space with basis e^i, $i=1,\ldots,N$, where $e^i(j)=\delta^i_j$ for $i,j\in\mathcal{M}$. These functions satisfy the two identities,

$$e^ie^j=\delta^{ij}e^j,\quad \sum_i e^i=\mathbf{1}, \tag{3.1}$$

where $\mathbf{1}$ is the constant function on $\mathcal{M}$ with value 1. In Ref. 3 it has been shown that first order differential calculi on a finite set $\mathcal{M}$ are in bijective correspondence with digraph structures on $\mathcal{M}$. Given a digraph with a set of vertices $\mathcal{M}$, we associate with an arrow from some point i to another point j, denoted as $i\to j$ in the following, an algebraic object e^{ij} and define[17]

$$\Omega^1:=\mathrm{span}_{\mathbb{C}}\{e^{ij}|i\to j\}. \tag{3.2}$$

This is turned into an $\mathcal{A}$-bimodule via

$$e^ie^{kl}=\delta^{ik}e^{kl},\quad e^{kl}e^i=\delta^{li}e^{kl}. \tag{3.3}$$

Let us introduce

$$\rho=\sum_{k,l} e^{kl}, \tag{3.4}$$

where the summation has to be restricted to those k,l for which there is an arrow from k to l in the digraph. Then

$$\mathrm{d}f=[\rho,f],\quad f\in\mathcal{A} \tag{3.5}$$

defines a $\mathbb{C}$-linear map $\mathrm{d}:\mathcal{A}\to\Omega^1$ which satisfies the Leibniz rule. If there is an arrow from i to j in the digraph, then $e^i\rho e^j=e^{ij}$, otherwise $e^i\rho e^j=0$.

The subspace

$$\Omega^1_i:=e^i\Omega^1 \tag{3.6}$$

is generated by the 1-forms e^{ij} corresponding to the arrows originating from i in the digraph. It may be regarded as the cotangent space at $i \in \mathcal{M}$. We have

$$\Omega^1 = \bigoplus_{i \in \mathcal{M}} \Omega^1_i . \tag{3.7}$$

The complete digraph where all pairs of points in $\mathcal{M}$ are connected by a pair of antiparallel arrows corresponds to the largest first order differential calculus on $\mathcal{M}$, also known as the *universal* first order differential calculus since all other calculi can be obtained from it as a quotient with respect to some sub-bimodule.

There is a canonical commutative product in Ω^1 which satisfies

$$\alpha \bullet \mathrm{d}f = [\alpha, f] \tag{3.8}$$

and

$$(f\alpha f') \bullet (h\beta h') = fh(\alpha \bullet \beta) f'h', \quad \forall f,f',h,h' \in \mathcal{A}, \quad \alpha,\beta \in \Omega^1. \tag{3.9}$$

More generally, this product exists for every first order differential calculus over a commutative algebra.[18] In the case under consideration, it is given by

$$e^{ij} \bullet e^{kl} = \delta^{ik} \delta^{jl} e^{ij}. \tag{3.10}$$

The space of 1-forms Ω^1 is free as a (left or right) $\mathcal{A}$-module. A special left $\mathcal{A}$-module basis is given by

$$\rho^i := \sum_j e^{ji}, \quad \text{if } \rho e^i \neq 0, \tag{3.11}$$

since an arbitrary 1-form A can be written as

$$A = \sum_{i,j} A_{ij}\, e^{ij} = \sum_i A_i\, \rho^i, \tag{3.12}$$

where $A_i = \Sigma_j A_{ji} e^j$. Furthermore, $\Sigma_i A_i \rho^i = 0$ implies, via multiplication with e^j from the left, that $A_{ji} = 0$ and thus $A_i = 0$.

B. Higher order differential forms on a finite set

Concatenation of the 1-forms e^{ij} leads to the r-forms

$$e^{i_0 \cdots i_r} := e^{i_0 i_1} e^{i_1 i_2} \cdots e^{i_{r-1} i_r} \quad (r>0), \tag{3.13}$$

which can also be expressed as follows:

$$e^{i_0 \cdots i_r} = e^{i_0} \rho e^{i_1} \rho \cdots \rho e^{i_r}. \tag{3.14}$$

They satisfy the simple relations

$$e^{i_0 \cdots i_r} e^{j_0 \cdots j_s} = \delta^{i_r j_0} e^{i_0 \cdots i_{r-1} j_0 \cdots j_s} \tag{3.15}$$

and span Ω^r as a vector space over $\mathbb{C}$. Using (3.3), this space is turned into an $\mathcal{A}$-bimodule. The exterior derivative d extends to higher orders via

$$\mathrm{d}e^i = \rho e^i - e^i \rho, \tag{3.16}$$

$$\mathrm{d}\rho = \rho^2 + \sum_i e^i \rho^2 e^i, \tag{3.17}$$

and the (graded) Leibniz rule (2.4). In particular, this leads to

$$\mathrm{d}e^{ij} = \rho e^i \rho e^j - e^i \rho^2 e^j + e^i \rho e^j \rho, \tag{3.18}$$

$$\mathrm{d}e^{ijk} = \rho e^i \rho e^j \rho e^k - e^i \rho^2 e^j \rho e^k + e^i \rho e^j \rho^2 e^k - e^i \rho e^j \rho e^k \rho. \tag{3.19}$$

Starting with the universal first order differential calculus on $\mathcal{M}$, these formulas generate the *universal differential calculus* (which is also known as the *universal differential envelope* of $\mathcal{A}$). A smaller first order differential calculus (where some of the e^{ij} are missing) induces restrictions on the spaces of higher order forms. A missing arrow from *i* to some other point *j* (in the complete digraph on $\mathcal{M}$) means $e^i \rho e^j = 0$. Acting with d on this equation, using (3.16) and (3.17), leads to

$$i \not\to j \;\Rightarrow\; e^i \rho^2 e^j = 0. \tag{3.20}$$

Each differential calculus is obtained from the universal one as a quotient with respect to some differential ideal. If the differential ideal is generated by ''basic forms'' (3.13) only,[19] then the differential calculus is called *basic*.[16] This class of differential calculi has been associated with polyhedral representations of simplicial complexes.[16]

C. Vector fields on a finite set

Let $\mathfrak{X}$ denote the dual of Ω^1 as a complex vector space. Let $\{\partial_{ji}\}$ be the basis of $\mathfrak{X}$ dual to $\{e^{ij}\}$. If $\langle , \rangle_0$ denotes the duality contraction, then

$$\langle e^{ij}, \partial_{kl} \rangle_0 = \delta^i_l \delta^j_k . \tag{3.21}$$

$\mathfrak{X}$ is turned into an $\mathcal{A}$-bimodule by introducing the left and right actions

$$\langle \alpha, f \cdot X \rangle_0 := \langle \alpha f, X \rangle_0 , \quad \langle \alpha, X \cdot f \rangle_0 := \langle f \alpha, X \rangle_0 . \tag{3.22}$$

As a consequence,

$$e^k \cdot \partial_{ji} = \delta^k_j \partial_{ji} , \quad \partial_{ji} \cdot e^k = \delta^k_i \partial_{ji} . \tag{3.23}$$

An element $X \in \mathfrak{X}$ can be uniquely decomposed as follows:

$$X = \sum_{i \to j} X(i)^j \partial_{ji} \tag{3.24}$$

(where the summation runs over all $i,j \in \mathcal{M}$ for which there is an arrow from *i* to *j* in the digraph associated with Ω^1). Now we introduce a duality contraction $\langle , \rangle$ of Ω^1 as a right $\mathcal{A}$-module and $\mathfrak{X}$ as a left $\mathcal{A}$-module by setting

$$\langle e^{ij}, X \rangle := e^i \langle e^{ij}, X \rangle_0 , \tag{3.25}$$

for all $X \in \mathfrak{X}$. Then we have

$$\langle f\alpha, X \cdot h \rangle = f \langle \alpha, X \rangle h, \quad \langle \alpha, f \cdot X \rangle = \langle \alpha f, X \rangle. \tag{3.26}$$

The elements of $\mathfrak{X}$ become operators on $\mathcal{A}$ via

$$X(f) := \langle \mathrm{d}f, X \rangle. \tag{3.27}$$

Using the Leibniz rule for d, one proves

$$X(fh)=fX(h)+(h\cdot X)(f), \quad \forall f,h\in\mathcal{A}. \tag{3.28}$$

Furthermore,

$$(X\cdot f)(g)=X(g)f. \tag{3.29}$$

The duality contraction extends to the pair of spaces $\Omega\otimes_{\mathcal{A}}\Omega^1$ and $\mathfrak{X}\otimes_{\mathcal{A}}\Omega$ via

$$\langle w\otimes_{\mathcal{A}}\alpha, X\otimes_{\mathcal{A}}w'\rangle=w\langle\alpha,X\rangle w'. \tag{3.30}$$

The space

$$\mathfrak{X}_i:=\mathfrak{X}e^i=\{X\cdot e^i|X\in\mathfrak{X}\} \tag{3.31}$$

may be regarded as the tangent space at $i\in\mathcal{M}$. It is dual to Ω^1_i with respect to the duality contraction $\langle,\rangle_0$. The set $\{\partial_{ji}|j\in\mathcal{M}$ such that $i\rightarrow j\}$ is a basis of $\mathfrak{X}_i$ which is dual to the basis $\{e^{ij}|j\in\mathcal{M}$ such that $i\rightarrow j\}$ of Ω^1_i.

D. Linear connections on a finite set

Let ∇: $\Omega^1\rightarrow\Omega^1\otimes_{\mathcal{A}}\Omega^1$ be a (left $\mathcal{A}$-module) linear connection. Using (2.6) and the properties of ρ, one finds that

$$\mathbb{U}(\alpha):=\rho\otimes_{\mathcal{A}}\alpha-\nabla\alpha \tag{3.32}$$

is a left $\mathcal{A}$-homomorphism $\mathbb{U}:\Omega^1\rightarrow\Omega^1\otimes_{\mathcal{A}}\Omega^1$, i.e.,

$$\mathbb{U}(f\alpha)=f\mathbb{U}(\alpha), \quad \forall f\in\mathcal{A},\ \alpha\in\Omega^1. \tag{3.33}$$

We call $\mathbb{U}$ the *parallel transport* associated with the linear connection ∇. In particular, (3.33) implies $\mathbb{U}(e^{ij})=e^i\mathbb{U}(e^{ij})$, and thus we have an expansion,

$$\mathbb{U}(e^{ij})=\sum_{k,l}U(i)^j{}_{kl}\,e^{ik}\otimes_{\mathcal{A}}e^{kl}=\sum_k e^{ik}\otimes_{\mathcal{A}}\sum_l U(i)^j{}_{kl}\,e^{kl}, \tag{3.34}$$

with constants $U(i)^j{}_{kl}$. Via

$$e^{ik}\mapsto(e^{ik})\mathbb{U}^{ij}:=\sum_l U(i)^k{}_{jl}\,e^{jl}, \tag{3.35}$$

for fixed i and j, the parallel transport defines a linear map $\Omega^1_i\rightarrow\Omega^1_j$ with associated matrix $\mathbb{U}^{ij}$. Then we have

$$\mathbb{U}(\alpha)=\sum_{i,j}e^{ij}\otimes_{\mathcal{A}}[(e^i\alpha)\mathbb{U}^{ij}]. \tag{3.36}$$

Given a linear connection on Ω^1, there is a dual connection[20] ∇: $\mathfrak{X}\rightarrow\mathfrak{X}\otimes_{\mathcal{A}}\Omega^1$, such that

$$\mathrm{d}\langle\alpha,X\rangle=\langle\nabla\alpha,X\rangle+\langle\alpha,\nabla X\rangle \tag{3.37}$$

(cf. Ref. 5, appendix B). Using $\mathrm{d}\langle\alpha,X\rangle=[\rho,\langle\alpha,X\rangle]$ one proves that the dual parallel transport defined by

$$\langle\alpha,\mathbb{U}(X)\rangle=\langle\mathbb{U}(\alpha),X\rangle, \tag{3.38}$$

acts as follows on $\mathfrak{X}$:

$$\mathbb{U}(X) := X \otimes_{\mathcal{A}} \rho + \nabla X, \tag{3.39}$$

and satisfies

$$\mathbb{U}(X \cdot f) = \mathbb{U}(X) f. \tag{3.40}$$

(3.34) leads to

$$\mathbb{U}(\partial_{ji}) = \sum_{k,l} U(k)^l{}_{ij} \; \partial_{lk} \otimes_{\mathcal{A}} e^{ki}. \tag{3.41}$$

The parallel transport (and thus also the connection) extends in an obvious way to $\Omega \otimes_{\mathcal{A}} \Omega^1$ and $\mathfrak{X} \otimes_{\mathcal{A}} \Omega$ as graded left, respectively, right Ω-homomorphisms, i.e.,

$$\mathbb{U}(w \otimes_{\mathcal{A}} \alpha) = (-1)^r w \otimes_{\mathcal{A}} \mathbb{U}(\alpha), \quad \mathbb{U}(X \otimes_{\mathcal{A}} w) = (-1)^r \mathbb{U}(X) \otimes_{\mathcal{A}} w, \tag{3.42}$$

where $w \in \Omega^r$.

The map $\mathfrak{X}_j \to \mathfrak{X}_i$ dual to the parallel transport map with matrix $\mathbb{U}^{ij}$ defined in (3.35) is given by

$$\partial_{ki} \mapsto \sum_l U(j)^l_{ik} \; \partial_{lj} = \mathbb{U}^{ij}(\partial_{ki}). \tag{3.43}$$

Now (3.41) extends to

$$\mathbb{U}(X) = \sum_{i,j} \mathbb{U}^{ij}(X \cdot e^i) \otimes_{\mathcal{A}} e^{ij}. \tag{3.44}$$

We may introduce the curvature as the right Ω-homomorphism $\mathfrak{R}' : \mathfrak{X} \otimes_{\mathcal{A}} \Omega \to \mathfrak{X} \otimes_{\mathcal{A}} \Omega$ defined by

$$\mathfrak{R}' = \nabla^2. \tag{3.45}$$

Its dual $\mathfrak{R} : \Omega \otimes_{\mathcal{A}} \Omega^1 \to \Omega \otimes_{\mathcal{A}} \Omega^1$ is then given by $\mathfrak{R} = -\nabla^2$ in accordance with our general definition (2.12). We obtain

$$\mathfrak{R}(e^{ij}) =: \sum_{k,l,m} R(i)^j{}_{klm} \; e^{ikl} \otimes_{\mathcal{A}} e^{lm} = \sum_{k,l,m} \left(\sum_n U(i)^j{}_{kn} U(k)^n{}_{lm} - U(i)^j{}_{lm} \right) e^{ikl} \otimes_{\mathcal{A}} e^{lm}, \tag{3.46}$$

where it has been convenient to set

$$U(i)^j{}_{ik} := \delta^j_k. \tag{3.47}$$

We also have the following expression for the curvature:

$$\mathfrak{R}(\alpha) = \sum_{i,j,k} e^{ijk} \otimes_{\mathcal{A}} \{(e^i \alpha)[\mathbb{U}^{ij} \mathbb{U}^{jk} - \mathbb{U}^{ik}]\}, \tag{3.48}$$

where $\mathbb{U}^{ii} := \mathrm{id}_{\Omega^1_i}$.

For the torsion we find

$$\Theta(e^{ij}) = -e^i \rho^2 e^j + e^{ij} \rho + \sum_{k,l} U(i)^j{}_{kl} \; e^{ikl} = \sum_{k,l} \left(\delta^j_k - \delta^j_l + U(i)^j{}_{kl} \right) e^{ikl}. \tag{3.49}$$

Example. In case of the universal differential calculus, the condition of vanishing torsion leads to

$$U(i)^j{}_{kl} = \delta^j_l - \delta^j_k , \tag{3.50}$$

and thus fixes the linear connection completely.[21] As mentioned in more generality in the example in Sec. II B, this connection is given by $\boldsymbol{\nabla} = \mathrm{d}$ and its curvature vanishes. ■

E. Metrics and compatible linear connections on finite sets

Using

$$e^{ij} \otimes_L e^{kl} = e^{ij} \otimes_L e^k e^{kl} = e^k e^{ij} \otimes_L e^{kl} = \delta^{ki} e^{ij} \otimes_L e^{il}, \tag{3.51}$$

one finds that an element $g \in \Omega^1 \otimes_L \Omega^1$ can be expressed as

$$g = \sum_{i,j,k} g(i)_{jk} \; e^{ij} \otimes_L e^{ik}, \tag{3.52}$$

with constants $g(i)_{jk}$. This will be our candidate for a *metric* on $\mathcal{M}$.[22]

Example 1: Consider a digraph embedded in Euclidean space such that the arrows are straight lines of Euclidean length l_{ij}. Let ϑ_{jik} denote the angle between arrows from i to j and from i to k. Define[23]

$$g(i)_{jj} = l_{ij}^2 , \quad g(i)_{jk} = l_{ij} l_{ik} \cos \vartheta_{jik} . \tag{3.53}$$

In order to describe the geometry of a polygon (without orientation of its lines) embedded in Euclidean space completely, in general, we need to associate it with a *symmetric* digraph. A line between two points i and j is then represented by a pair of antiparallel arrows, so that e^{ij} and e^{ji} are both present. Of course, we should impose $l_{ij} = l_{ji}$.[24] ■

In order to define the compatibility of a linear connection and a metric, we have to extend the connection, respectively, the map $\mathbb{U}$, to a map from $\Omega^1 \otimes_L \Omega^1$ to $\Omega^1 \otimes_{\mathcal{A}} \Omega^1 \otimes_L \Omega^1$. Let us define

$$\mathbb{U}(\alpha \otimes_L \beta) := \bullet(\mathbb{U}(\alpha) \otimes_L \mathbb{U}(\beta)), \tag{3.54}$$

where a map

$$\bullet \; : \; (\Omega^1 \otimes_{\mathcal{A}} \Omega^1) \otimes_L (\Omega^1 \otimes_{\mathcal{A}} \Omega^1) \to \Omega^1 \otimes_{\mathcal{A}} (\Omega^1 \otimes_L \Omega^1) \tag{3.55}$$

is needed. Using the canonical product (3.10) in the space of 1-forms, such a map is given by

$$\bullet((\alpha \otimes_{\mathcal{A}} \beta) \otimes_L (\alpha' \otimes_{\mathcal{A}} \beta')) := (\alpha \bullet \alpha') \otimes_{\mathcal{A}} (\beta \otimes_L \beta'), \tag{3.56}$$

and, using (3.9), we have

$$\mathbb{U}(f(\alpha \otimes_L \beta)) = f\,\mathbb{U}(\alpha \otimes_L \beta). \tag{3.57}$$

As a consequence,

$$\nabla(\alpha \otimes_L \beta) := \rho \otimes_{\mathcal{A}} (\alpha \otimes_L \beta) - \mathbb{U}(\alpha \otimes_L \beta) \tag{3.58}$$

defines a (left $\mathcal{A}$-module) connection on $\Omega^1 \otimes_L \Omega^1$. The *metric compatibility* condition $\boldsymbol{\nabla} g = 0$ now amounts to

$$\rho \otimes_{\mathcal{A}} g = \mathbb{U}(g). \tag{3.59}$$

In terms of the matrices $\mathbb{U}^{ij}$ introduced in Sec. III D, we have

$$\mathbb{U}(\alpha\otimes_L\beta)=\sum_{i,j} e^{ij}\otimes_{\mathcal{A}}\{[(e^i\alpha)\mathbb{U}^{ij}]\otimes_L(e^i\beta)\mathbb{U}^{ij}]\}. \tag{3.60}$$

Lemma: Expressed in components, $\nabla g=0$ becomes

$$g(i)_{jk}=\sum_{m,n} g(l)_{mn}U(l)^m{}_{ij}U(l)^n{}_{ik}, \tag{3.61}$$

for all $i,l\in\mathcal{M}$ such that $l\to i$ (i.e., there is an arrow from l to i in the digraph associated with Ω^1).

Proof:

$$\mathbb{U}(g)=\sum_{l,m,n} g(l)_{mn}\,\bullet(\mathbb{U}(e^{lm})\otimes_L\mathbb{U}(e^{ln}))$$

$$=\sum_{l,m,n} g(l)_{mn}\sum_{i,j,k,p} U(l)^m{}_{ij}U(l)^n{}_{pk}\,\bullet((e^{li}\otimes_{\mathcal{A}}e^{ij})\otimes_L(e^{lp}\otimes_{\mathcal{A}}e^{pk})).$$

With

$$\bullet((e^{li}\otimes_{\mathcal{A}}e^{ij})\otimes_L(e^{lp}\otimes_{\mathcal{A}}e^{pk}))=(e^{li}\bullet e^{lp})\otimes_{\mathcal{A}}(e^{ij}\otimes_L e^{pk})=\delta^{ip}e^{li}\otimes_{\mathcal{A}}(e^{ij}\otimes_L e^{pk}),$$

this becomes

$$\mathbb{U}(g)=\sum_{i,j,k,l,m,n} g(l)_{mn}U(l)^m{}_{ij}U(l)^n{}_{ik}\,e^{li}\otimes_{\mathcal{A}}(e^{ij}\otimes_L e^{ik}).$$

Using (3.59), the last expression must be equal to

$$\rho\otimes_{\mathcal{A}}g=\sum_{i,j,k,l} g(i)_{jk}\,e^{li}\otimes_{\mathcal{A}}(e^{ij}\otimes_L e^{ik}).$$

A comparison of the coefficients on both sides now leads to our formula. ■

Example 2: Again, we consider the universal differential calculus on $\mathcal{M}$. With the unique torsion-free linear connection (3.50), the metric compatibility condition reads as[25]

$$g(i)_{kl}=g(j)_{kl}+g(j)_{ii}-g(j)_{ki}-g(j)_{il},\quad i,j,k,l\in\mathcal{M}. \tag{3.62}$$

Setting $k=j$ and $l=j$, respectively, we get

$$g(i)_{jk}=g(j)_{ii}-g(j)_{ik},\quad g(i)_{kj}=g(j)_{ii}-g(j)_{ki}, \tag{3.63}$$

which, in turn, implies

$$g(i)_{jk}-g(i)_{kj}=g(j)_{ki}-g(j)_{ik} \tag{3.64}$$

and

$$g(i)_{jj}=g(j)_{ii}. \tag{3.65}$$

Furthermore, the last equation together with (3.62) leads to

$$2g(i)_{kl}-g(i)_{kj}-g(i)_{jl}=2g(j)_{kl}-g(j)_{ki}-g(j)_{il}, \tag{3.66}$$

which for $k=l$ becomes

$$2g(i)_{kk}-g(i)_{kj}-g(i)_{jk}=2g(j)_{kk}-g(j)_{ki}-g(j)_{ik}. \tag{3.67}$$

Let us now consider the special case where all the components $g(i)_{jj}$ are equal. Then (3.64) and (3.67) lead to

$$g(i)_{kj}=g(j)_{ik}. \tag{3.68}$$

With the help of (3.63) and (3.65) we now obtain

$$g(i)_{jj}=g(i)_{jk}+g(i)_{kj}. \tag{3.69}$$

Assuming in addition that the metric is symmetric [i.e., $g(i)_{jk}=g(i)_{kj}$], we have

$$g(i)_{jj}=2g(i)_{jk}, \tag{3.70}$$

and we end up with a constant metric,

$$g(i)=\begin{pmatrix} a & a/2 & \dots & a/2 \\ a/2 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a/2 & \cdots & a/2 & a \end{pmatrix}, \quad \forall i\in\mathcal{M}. \tag{3.71}$$

Hence, there is a unique symmetric g for the universal differential calculus (associated with the complete digraph) on $\mathcal{M}$ which is compatible with the (unique) torsion-free linear connection and which has the property that all $g(i)_{jj}$ are equal. If $g(i)_{jj}$ is positive, we let it represent the square of the distance between i and j. The above requirement then means that all points are at equal distance $l=\sqrt{a}$ and from the metric compatibility condition we recover the Euclidean geometry of the regular polyhedron.

More generally, specializing to the ''Euclidean metric'' (3.53), our metric compatibility conditions (3.62) become

$$l_{ik}^2=l_{jk}^2+l_{ji}^2-2l_{ji}l_{jk}\cos(\vartheta_{ijk}), \tag{3.72}$$

$$l_{ik}l_{il}\cos(\vartheta_{kil})=l_{jk}l_{jl}\cos(\vartheta_{kjl})+l_{ji}^2-l_{ji}l_{jk}\cos(\vartheta_{ijk})-l_{ji}l_{jl}\cos(\vartheta_{ijl}), \tag{3.73}$$

which, in fact, reproduce well-known relations of Euclidean geometry. ■

In terms of the matrices

$$\mathbf{g}(i):=(g(i)_{jk}), \tag{3.74}$$

the metric compatibility condition takes the simple form

$$\mathbf{g}(j)=(\mathbb{U}^{ij})^t\mathbf{g}(i)\mathbb{U}^{ij}, \tag{3.75}$$

where $(\mathbb{U}^{ij})^t$ denotes the transpose of the matrix $\mathbb{U}^{ij}$. Hence, if there is an arrow from i to j in the digraph (i.e., $i\to j$), then $\mathbf{g}(i)$ determines $\mathbf{g}(j)$ via the parallel transport of a metric compatible linear connection.

The metric compatibility condition implies that, for any closed path $i_0\to i_1\to\dots\to i_r\to i_0$ in the digraph, the matrix $\mathbb{H}^{i_0\cdots i_r}:=\mathbb{U}^{i_0i_1}\mathbb{U}^{i_1i_2}\cdots\mathbb{U}^{i_ri_0}$ must be in the orthogonal group of $\mathbf{g}(i_0)$. The set of all matrices $\mathbb{H}^{i_0\cdots i_r}$, $r\geqslant 1$, forms the holonomy group $G_H(i_0)$ at $i_0\in\mathcal{M}$.

Example 3: The three point complete digraph.
Let $\mathcal{M}=\{1,2,3\}$ with $\rho=e_{12}+e_{13}+e_{21}+e_{23}+e_{31}+e_{32}$. We are dealing again with the universal differential calculus so that there are no 2-form relations. Then $\rho^2=e_{121}+e_{123}+e_{131}+e_{132}+e_{212}+e_{213}+e_{231}+e_{232}+e_{312}+e_{313}+e_{321}+e_{323}$. The condition of vanishing torsion determines the connection completely. We find

$$\mathbb{U}^{12}=\begin{pmatrix}-1 & -1\\ 0 & 1\end{pmatrix},\quad \mathbb{U}^{13}=\begin{pmatrix}0 & 1\\ -1 & -1\end{pmatrix},\quad \mathbb{U}^{23}=\begin{pmatrix}1 & 0\\ -1 & -1\end{pmatrix},$$
$$\mathbb{U}^{21}=\begin{pmatrix}-1 & -1\\ 0 & 1\end{pmatrix},\quad \mathbb{U}^{31}=\begin{pmatrix}-1 & -1\\ 1 & 0\end{pmatrix},\quad \mathbb{U}^{32}=\begin{pmatrix}1 & 0\\ -1 & -1\end{pmatrix}. \tag{3.76}$$

It follows that $\mathbb{H}^{ij}=I$, the unit matrix, for all $i\to j\to i$. Furthermore, for all permutations i,j,k of 1,2,3 we find $\mathbb{H}^{ijk}=\mathbb{U}^{ij}\mathbb{U}^{jk}\mathbb{U}^{ki}=I$. This means that parallel transport does not depend on the path which is related to the fact that the curvature vanishes. If we choose metric components at one point, then the metric components at the other points are determined via the metric compatibility condition. We find

$$g(1)=\begin{pmatrix}a & b\\ b & c\end{pmatrix},\quad g(2)=\begin{pmatrix}a & a-b\\ a-b & a-2b+c\end{pmatrix},\quad g(3)=\begin{pmatrix}c & c-b\\ c-b & a-2b+c\end{pmatrix}. \tag{3.77}$$

In particular, if $g(1)=g(2)=g(3)$ we are led to

$$g(i)=b\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix} \tag{3.78}$$

[in accordance with (3.71)] which (for $b>0$) describes an equilateral triangle. This may be considered as a simple model of a piece of a two-dimensional surface. ■

Thinking about an inverse (or dual) of a metric tensor, as defined above, one is led to elements $h\in\mathfrak{X}\otimes_R\mathfrak{X}$ where $\otimes_R$ denotes the *right*-linear tensor product. h can be expressed as

$$h=\sum_{i,j,k} h(i)^{jk}\ \partial_{ji}\otimes_R\partial_{ki}, \tag{3.79}$$

with constants $h(i)^{jk}$. The parallel transport (and thus also the connection) extends to $\mathfrak{X}\otimes_R\mathfrak{X}$ via

$$\mathbb{U}(X\otimes_R Y):=\bullet(\mathbb{U}(X)\otimes_R\mathbb{U}(Y)) \tag{3.80}$$

and

$$\bullet((X\otimes_{\mathcal{A}}\alpha)\otimes_R(Y\otimes_{\mathcal{A}}\beta)):=(X\otimes_R Y)\otimes_{\mathcal{A}}(\alpha\bullet\beta). \tag{3.81}$$

Compatibility of h with a linear connection, i.e., $\boldsymbol{\nabla}h=0$, now reads as

$$\mathbb{U}(h)=h\otimes_{\mathcal{A}}\rho, \tag{3.82}$$

and, in components,

$$h(i)^{rs}=\sum_{j,k} h(l)^{jk}U(i)^r{}_{lj}U(i)^s{}_{lk}, \tag{3.83}$$

provided that $i\to l$. In terms of the matrices $h(i):=(h(i)^{jk})$, the metric compatibility condition reads as

$$h(i)=\mathbb{U}^{ij}h(j)(\mathbb{U}^{ij})^t. \tag{3.84}$$

Remark: Consider a differential calculus, associated with a symmetric digraph, a metric g and a compatible linear connection. If $g(i_0)$ is invertible at some point i_0, setting $h(i_0):=g(i_0)^{-1}$ defines h via (3.84) on the connected component of the digraph containing i_0. Of course, h need not be inverse to g at other points. ■

F. Metrics and compatible linear connections on a finite set with a basic differential calculus

We consider a *basic* differential calculus (cf. Sec. III B). The general torsion-free connection is then given by

$$U(i)^j{}_{kl} = \delta^j_l - \delta^j_k + u(ikl)^j, \tag{3.85}$$

where $u(ikl)^j \neq 0$ only if $e^{ikl}=0$.[26] The metric compatibility condition now becomes

$$\begin{aligned} g(j)_{kl} = {} & g(i)_{kl} - g(i)_{kj} - g(i)_{jl} + g(i)_{jj} \\ & + \sum_{m,n} g(i)_{mn} \, [\delta^m_k \, u(ijl)^n + \delta^n_l \, u(ijk)^m + u(ijk)^m \, u(ijl)^n] \\ & - \sum_m \, [g(i)_{jm} \, u(ijl)^m + g(i)_{mj} \, u(ijk)^m], \end{aligned} \tag{3.86}$$

for all i,j with $i \to j$.

Remark: Let us consider again the case of a Euclidean embedding space (cf. example 1 in Sec. III E). If all $u(ijk)^l$ vanish, then (3.72) holds which is a familiar relation between the lengths and angles of a Euclidean triangle. As shown in Ref. 27, in the triangulation of a curved space by means of geodesic segments and in Riemann normal coordinates one has

$$2 l_{ij} l_{ik} \cos \vartheta_{jik} = l^2_{ik} + l^2_{ij} - l^2_{jk} - \tfrac{1}{3} R_{\mu\alpha\nu\beta} \, \Delta x^\mu_{ij} \, \Delta x^\nu_{ij} \, \Delta x^\alpha_{ik} \, \Delta x^\beta_{ik} + \mathcal{O}(\epsilon^5), \tag{3.87}$$

where ϵ is a typical length scale of the neighborhood in which the Riemann normal coordinates are defined, and x^μ_i are the Riemann normal coordinates of the vertex i. Obviously, from (3.86) we can expect to get additional terms in (3.72), related to curvature, only if we have nonvanishing $u(ijk)^l$, that is if we have 2-form relations as in our next example. ■

Example: A refined model for a piece of a two-dimensional surface is obtained from that considered in example 3 of Sec. III E by adding a fourth point to the triangle and joining it with all the vertices of the latter, but then discard the 2-forms corresponding to the base of the resulting tetrahedron (or a pyramid with a triangle base). Hence, we consider the complete digraph on $\mathcal{M} = \{1,2,3,4\}$, but not the universal differential calculus since we impose the 2-form relations

$$e^{123} = e^{132} = e^{213} = e^{231} = e^{312} = e^{321} = 0. \tag{3.88}$$

We assume that the matrices $\mathbb{U}^{ij}$ have maximal rank and that

$$\mathbb{H}^{ij} = \mathbb{U}^{ij} \mathbb{U}^{ji} = I. \tag{3.89}$$

The condition of vanishing torsion now leads to

$$\mathbb{U}^{12} = \begin{pmatrix} -1 & -1+u_1 & -1 \\ 0 & 1+u_2 & 0 \\ 0 & u_3 & 0 \end{pmatrix}, \quad \mathbb{U}^{13} = \begin{pmatrix} 0 & 1+v_1 & 0 \\ -1 & -1+v_2 & -1 \\ 0 & v_3 & 1 \end{pmatrix},$$

$$\mathbb{U}^{23} = \begin{pmatrix} 1+w_1 & 0 & 0 \\ -1+w_2 & -1 & -1 \\ w_3 & 0 & 1 \end{pmatrix}, \quad \mathbb{U}^{14} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}, \tag{3.90}$$

$$\mathbb{U}^{24}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}, \quad \mathbb{U}^{34}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & -1 \end{pmatrix},$$

and for $i<j$ we have $\mathbb{U}^{ji}=(\mathbb{U}^{ij})^{-1}$ according to (3.89). Setting

$$g(4)=l^2\begin{pmatrix} 1 & c & b \\ c & 1 & a \\ b & a & 1 \end{pmatrix} \tag{3.91}$$

means that the edges of the triangles 4-1-2, 4-1-3, 4-2-3 have equal length $l_{41}=l_{42}=l_{43}=l$ but possibly different angles $\cos\vartheta_{142}=c$, $\cos\vartheta_{143}=b$, $\cos\vartheta_{243}=a$. Via $g(i)=(\mathbb{U}^{4i})^t g(4)\,\mathbb{U}^{4i}$ for $i=1,2,3$ we obtain

$$g(1)=l^2\begin{pmatrix} 2(1-c) & 1+a-b-c & 1-c \\ 1+a-b-c & 2(1-b) & 1-b \\ 1-c & 1-b & 1 \end{pmatrix},$$

$$g(2)=l^2\begin{pmatrix} 2(1-c) & 1-a+b-c & 1-c \\ 1-a+b-c & 2(1-a) & 1-a \\ 1-c & 1-a & 1 \end{pmatrix}, \tag{3.92}$$

$$g(3)=l^2\begin{pmatrix} 2(1-b) & 1-a-b+c & 1-b \\ 1-a-b+c & 2(1-a) & 1-a \\ 1-b & 1-a & 1 \end{pmatrix}.$$

The remaining metric compatibility conditions now demand that

$$u_2=v_1=w_1=-2, \quad u_1=2\frac{bc-a}{c^2-1}, \quad v_2=2\frac{bc-a}{b^2-1}, \quad w_2=2\frac{ac-b}{a^2-1} \tag{3.93}$$

and

$$u_3=2\frac{1-a-b+c}{1+c}, \quad v_3=2\frac{1-a+b-c}{1+b}, \quad w_3=2\frac{1+a-b-c}{1+a}, \tag{3.94}$$

where we assumed that $a,b,c\neq\pm1$. We should mention here that $u_1=\cdots=w_3=0$ is also a solution. This parallel transport, which corresponds to the unique torsion-free connection on the universal differential calculus on the set of four points, has vanishing curvature. This shows that there is *a priori* no relation with the Regge curvature[28] which is given at point 4 by $2\pi-\vartheta_{142}-\vartheta_{143}-\vartheta_{243}$. We will return to this example in the next section (see example 5 there). ∎

IV. TRANSFORMATIONS TO ''LOCAL'' TENSOR PRODUCTS AND COVARIANT TENSOR COMPONENTS

As in the preceding section, we consider a finite set $\mathcal{M}$ and a differential calculus Ω (over the algebra of functions) on $\mathcal{M}$. In ordinary (continuum) differential geometry, the tensor product $\otimes_{\mathcal{A}}$ and the graded product in the space of differential forms are operations which take place over the same point. This is not so in the discrete framework under consideration. For example, in $e^{ij}\otimes_{\mathcal{A}} e^{jk}$ the first factor is an element of Ω^1_i while the second factor belongs to Ω^1_j. In contrast, in $e^{ij}\otimes_L e^{ik}$ both factors belong to the same cotangent space. As a consequence, the left components of an element of $\Omega^1\otimes_L\Omega^1$ transform covariantly under a change of module basis in Ω^1 (in contrast to the left, middle or right components of an element of $\Omega^1 \otimes_{\mathcal{A}} \Omega^1$). Covariant tensor components are of particular interest because of the possibility to construct new tensors from them via contraction. For example, we would like to build a kind of Ricci tensor from the curvature components $R(i)^j{}_{klm}$ in (3.46). The latter are not covariant, however. The indices j and l (or m) live in different (co)tangent spaces. In this section, we shall consider ways to modify or, more precisely, to ''localize'' expressions in order to provide a remedy for this problem. What we need is tensor products which act over the same point and furthermore suitable transformations from tensor products over $\mathcal{A}$ to these ''local'' tensor products. Given a connection, we have the parallel transports which enable us to move from one (co)tangent space to another and these should be expected as natural ingredients of the transformations we are looking for.

A map $\Omega^1 \otimes_L \Omega^1 \to \Omega^1 \otimes_{\mathcal{A}} \Omega^1$ is given by

$$\kappa(\alpha \otimes_L \beta) := \sum_{i,j} (e^i \alpha e^j) \otimes_{\mathcal{A}} [(e^i \beta) \mathbb{U}^{ij}]. \tag{4.1}$$

In particular,

$$\kappa(e^{ij} \otimes_L e^{ik}) = \sum_l U(i)^k{}_{jl} \, e^{ij} \otimes_{\mathcal{A}} e^{jl}. \tag{4.2}$$

κ is a left $\mathcal{A}$-homomorphism and has the property[29]

$$\kappa(\rho \otimes_L \beta) = \mathbb{U}(\beta). \tag{4.3}$$

A map,

$$\lambda_1 : \Omega^1 \otimes_{\mathcal{A}} \Omega^1 \to \Omega^1 \otimes_L \Omega^1, \tag{4.4}$$

in the opposite direction is not so easily at hand in an explicit form, except in some special cases like those listed below.

(i) If for all $i \to j$ the transport $\mathbb{U}^{ij}$ is invertible, we can define

$$\lambda_1(\alpha \otimes_{\mathcal{A}} \beta) := \sum_{i,j} (e^i \alpha e^j) \otimes_L [(e^j \beta)(\mathbb{U}^{ij})^{-1}]. \tag{4.5}$$

Then $\lambda_1 = \kappa^{-1}$. This choice is considered in case of the oriented lattice structures treated in Secs. V and VI.

(ii) If the digraph associated with Ω^1 is symmetric (i.e., a digraph where $i \to j \Leftrightarrow j \to i$) then we may define[30]

$$\lambda_1(\alpha \otimes_{\mathcal{A}} \beta) := \sum_{i,j} (e^i \alpha e^j) \otimes_L [(e^j \beta) \mathbb{U}^{ji}]. \tag{4.6}$$

In the following we assume that a map λ_1 is given, having the above examples in mind. Moreover, we will also need a similar map,

$$\lambda_2 : \Omega^2 \otimes_{\mathcal{A}} \Omega^1 \to \Omega^2 \otimes_L \Omega^1 \tag{4.7}$$

(and furthermore a way to ''localize'' 2-forms; see below). In our examples considered in Secs. V and VI, λ_1 induces such a map λ_2 in a natural way.

Example 1: Let $i \to j \to k \to l$ and $k \to i$. For $e^{ijk} \neq 0$ we may define

$$\lambda_2(e^{ijk} \otimes_{\mathcal{A}} e^{kl}) := e^{ijk} \otimes_L [(e^{kl}) \mathbb{U}^{ki}]. \tag{4.8}$$

If also $k \to j \to i$, another choice is

$$\lambda_2'(e^{ijk}\otimes_{\mathcal{A}} e^{kl}) := e^{ijk}\otimes_L[(e^{kl})\mathbb{U}^{kj}\mathbb{U}^{ji}]. \tag{4.9}$$

The two choices for λ_2 can be different as long as the holonomy of the connection is not trivial. Hence, in general, there are many different choices for λ_2. ■

Example 2: Let us now consider a differential calculus where the space of 1-forms is associated with a symmetric digraph and let us, moreover, assume that the differential calculus is basic (cf. Sec. III B). In this case, $e^{i_0\cdots i_r}\neq 0$ implies that $i_k\rightarrow i_l$ for all $0\leqslant k,l\leqslant r$ (cf. Ref. 16). A natural choice for λ_1, λ_2 and generalizations thereof is then given by

$$\lambda(e^{i_0\cdots i_r}\otimes_{\mathcal{A}} e^{i_r j}) := e^{i_0\cdots i_r}\otimes_{\mathcal{A}}[(e^{i_r j})\mathbb{U}^{i_r i_0}]. \tag{4.10}$$

■

In the following we simply write λ instead of λ_1 or λ_2.

Combining κ and π,

$$\alpha\cap\beta := \pi\circ\kappa(\alpha\otimes_L\beta) \tag{4.11}$$

determines a product $\Omega^1\otimes_L\Omega^1\rightarrow\Omega^2$ which is left $\mathcal{A}$-linear and therefore satisfies

$$e^i(\alpha\cap\beta) = (e^i\alpha)\cap(e^i\beta), \tag{4.12}$$

so that $\cap$ preserves "locality." If $(\kappa\circ\lambda)(\ker\pi)\subset\ker\pi$, the map

$$\mu := \pi\circ\kappa\circ\lambda\circ\pi^{-1}:\ \Omega^2\rightarrow\Omega^2 \tag{4.13}$$

is well-defined and can be used to transform usual products of 1-forms (i.e., elements of Ω^2) to $\cap$-products.

Example 3: Let us again consider the case of a differential calculus associated with a *symmetric* digraph. Using (4.6), we get

$$\kappa\circ\lambda_1(\alpha\otimes_{\mathcal{A}}\beta) = \sum_{i,j}(e^i\alpha e^j)\otimes_{\mathcal{A}}[(e^j\beta)\mathbb{H}^{ji}], \tag{4.14}$$

$$\lambda_1\circ\kappa(\alpha\otimes_L\beta) = \sum_{i,j}(e^i\alpha e^j)\otimes_L[(e^i\beta)\mathbb{H}^{ij}], \tag{4.15}$$

with the holonomies $\Omega^1_i\rightarrow\Omega^1_i$ given by $\mathbb{H}^{ij}$. Then

$$\mu(\alpha\beta) = \sum_{i,j}(e^i\alpha e^j)\cap[(e^j\beta)\mathbb{U}^{ji}] = \sum_{i,j}(e^i\alpha)[(e^j\beta)\mathbb{H}^{ji}]. \tag{4.16}$$

The 2-form relations are of the form

$$\sum_k e^{ikj} = 0, \quad \text{if } i\nrightarrow j \tag{4.17}$$

(where k runs over a subset of $\mathcal{M}$) and must be mapped to 0 by μ. In terms of the $\cap$-product the 2-form relations then read as

$$\sum_{k,l} U(k)^j{}_{il}\ e^{ik}\cap e^{il} = 0, \quad \text{if } i\nrightarrow j. \tag{4.18}$$

Using $(e^{kj})\mathbb{H}^{ki} =: \Sigma_l(\mathbb{H}^{ki})^j{}_l\ e^{kl}$, the condition $(\kappa\circ\lambda)(\ker\pi)\subset\ker\pi$ amounts to

$$\sum_k (\mathbb{H}^{ki})^j{}_l \, e^{ikl} = 0, \quad \forall l, \tag{4.19}$$

and thus induces restrictions on the connection, in general. ■

Lemma: For a basic differential calculus (Ω, d) and a torsion-free linear connection, we have

$$\begin{aligned} e^{ij} \cap e^{ij} &= -\sum_k e^{ijk}, \\ e^{ij} \cap e^{ik} &= e^{ijk}, \quad \text{if } j \neq k, \end{aligned} \tag{4.20}$$

and the map μ defined in (4.13) with λ from (4.6) satisfies

$$\begin{aligned} \mu(e^{iji}) &= -\sum_k e^{ij} \cap e^{ik}, \\ \mu(e^{ijk}) &= e^{ij} \cap e^{ik}, \quad \text{if } i \neq k. \end{aligned} \tag{4.21}$$

Proof: (4.20) follows from

$$e^{ij} \cap e^{ik} = \sum_m U(i)^k{}_{jm} \, e^{ijm},$$

together with (3.85). (4.21) results from

$$\mu(e^{ijk}) = e^{ij} \cap [e^{jk} \mathbb{U}^{ji}] = e^{ij} \cap \sum_m U(j)^k{}_{im} \, e^{im} = e^{ij} \cap \sum_m (\delta^k_m - \delta^k_i) e^{im},$$

using again (3.85). ■

Now we have everything at hand to ''localize'' torsion and curvature and to define corresponding covariant components as follows:

$$\mu \circ \Theta(e^{ij}) =: \sum_{k,l} Q(i)^j{}_{kl} \, e^{ik} \cap e^{il}, \tag{4.22}$$

$$(\mu \otimes_L \mathrm{id}) \circ \lambda \circ \mathfrak{R}(e^{ij}) =: \sum_{k,l,m} \hat{R}(i)^j{}_{klm} \, (e^{il} \cap e^{im}) \otimes_L e^{ik}. \tag{4.23}$$

As in ordinary differential geometry, a *Ricci tensor* can now be defined,

$$\mathrm{Ric}(i)_{jk} := \sum_l \hat{R}(i)^l{}_{jlk}, \quad \overline{\mathrm{Ric}}(i)_{jk} := \sum_l \hat{R}(i)^l{}_{jkl}. \tag{4.24}$$

There is also the contraction $\Sigma_l \hat{R}(i)^l{}_{ljk}$ which in classical Riemannian geometry vanishes identically. In the present framework its significance has still to be explored.

In order to construct a curvature scalar, we need an inverse of $g(i)$. This need not exist at all vertices of the digraph. There are examples where $g(i)$ is even degenerate at all vertices.

Example 4: We continue our example 2. With the assumptions made there, there are no conditions on the connection (cf. example 3). For the curvature we obtain

$$(\mu \otimes_L \mathrm{id}) \circ \lambda \circ \mathfrak{R}(e^{im}) = \sum_{i,j,k} e^{ij} \cap [e^{jk} \mathbb{U}^{ji}] \otimes_L \{(e^{im})[\mathbb{U}^{ij} \mathbb{U}^{jk} \mathbb{U}^{ki} - \mathbb{H}^{ik}]\}, \tag{4.25}$$

which for $e^{ij}\cap e^{ik}\neq 0$ yields

$$\hat{R}(i)^{m}{}_{njk}=\sum_{l} U(j)^{l}{}_{ik}\,[\mathbb{U}^{ij}\mathbb{U}^{jl}\mathbb{U}^{li}-\mathbb{H}^{il}]^{m}{}_{n}\,. \tag{4.26}$$

■

Example 5: We continue our example of Sec. III F and choose λ as in (4.10). The relations between the usual graded and the ∩-product are obtained from the above Lemma. In particular,

$$e^{41}\cap e^{41}=-e^{412}-e^{413}-e^{414},\quad e^{41}\cap e^{42}=e^{412} \tag{4.27}$$

and

$$e^{12}\cap e^{13}=0,\quad e^{12}\cap e^{12}=-e^{121}-e^{124},\quad e^{12}\cap e^{14}=e^{124}. \tag{4.28}$$

Since $\mathbb{H}^{ij}=I$, the map μ is well-defined. Then

$$\mu(e^{414})=-e^{41}\cap e^{41}-e^{41}\cap e^{42}-e^{41}\cap e^{43},\quad \mu(e^{412})=e^{41}\cap e^{42} \tag{4.29}$$

and

$$\mu(e^{123})=e^{12}\cap e^{13}=0,\quad \mu(e^{121})=-e^{12}\cap e^{12}-e^{12}\cap e^{14},\quad \mu(e^{124})=e^{12}\cap e^{14}. \tag{4.30}$$

The curvature $\hat{R}(i)_{jk}:=(\hat{R}(i)^{m}{}_{njk})$ at point 4 is given by

$$\hat{R}(4)_{11}=\hat{R}(4)_{22}=\hat{R}(4)_{33}=0 \tag{4.31}$$

and

$$\hat{R}(4)_{12}=\hat{R}(4)_{21}=\begin{pmatrix} 0 & 0 & 2(ac-b)/(c^2-1)\\ 0 & 0 & 2(bc-a)/(c^2-1)\\ 0 & 0 & -2\end{pmatrix},$$

$$\hat{R}(4)_{13}=\hat{R}(4)_{31}=\begin{pmatrix} 0 & 2(ab-c)/(b^2-1) & 0\\ 0 & -2 & 0\\ 0 & 2(bc-a)/(b^2-1) & 0\end{pmatrix}, \tag{4.32}$$

$$\hat{R}(4)_{23}=\hat{R}(4)_{32}=\begin{pmatrix} -2 & 0 & 0\\ 2(ab-c)/(a^2-1) & 0 & 0\\ 2(ac-b)/(a^2-1) & 0 & 0\end{pmatrix}.$$

Furthermore, we have $\hat{R}(1)_{22}=\hat{R}(1)_{33}=\hat{R}(1)_{44}=0$,

$$\hat{R}(1)_{24}=\hat{R}(1)_{42}=\begin{pmatrix} 0 & 2(bc-a)/(c^2-1) & 0\\ 0 & -2 & 0\\ 0 & 2(1-a-b+c)/(c+1) & 0\end{pmatrix}, \tag{4.33}$$

etc. and corresponding expressions for the curvature at the vertices 2 and 3. For the Ricci tensors, we find $\mathrm{Ric}(i)=\overline{\mathrm{Ric}}(i)$,

$$\mathrm{Ric}(4)=2\begin{pmatrix} 0 & (ac-b)/(a^2-1) & (ab-c)/(a^2-1)\\ (bc-a)/(b^2-1) & 0 & (ab-c)/(b^2-1)\\ (bc-a)/(c^2-1) & (ac-b)/(c^2-1) & 0\end{pmatrix}, \tag{4.34}$$

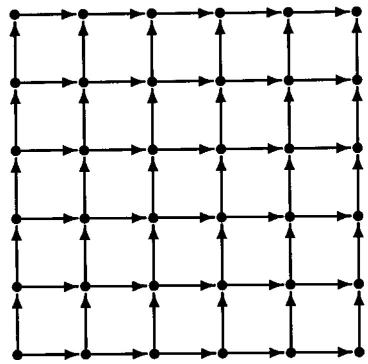

FIG. 1. A finite part of the oriented lattice graph.

$$\mathrm{Ric}(1)=2\begin{pmatrix} 0 & (1-a+b-c)/(b+1) & (bc-a)/(b^2-1) \\ (1-a-b+c)/(c+1) & 0 & (bc-a)/(c^2-1) \\ 0 & 0 & 0 \end{pmatrix}, \quad (4.35)$$

and corresponding expressions for $\mathrm{Ric}(j)$, $j=2,3$. The resulting expression for the curvature scalar turns out to be rather complicated. In the special case $a=b=c$, we obtain

$$R(4)=\sum_{i,j} g(4)^{ij}\ \mathrm{Ric}(4)_{ij}=\frac{1}{l^2}\frac{12a^2}{(a-1)(a+1)(2a+1)} \quad (4.36)$$

and

$$R(1)=R(2)=R(3)=-\frac{1}{l^2}\frac{8a}{(a+1)(2a+1)}. \quad (4.37)$$

■

The structures introduced in this section will also be exploited in the examples presented in the following two sections.

V. GEOMETRY OF THE ORIENTED LATTICE

In this section we choose $\mathcal{M}=\mathbb{Z}^n=\{a=(a^\mu)|a^\mu\in\mathbb{Z},\mu=1,\ldots,n\}$ and consider the differential calculus with

$$e^{ab}\neq 0 \Leftrightarrow b=a+\hat{\mu}, \quad \text{for some } \mu, \quad (5.1)$$

where $\hat{\mu}:=(\delta^\nu_\mu)\in\mathcal{M}$. The corresponding graph is an oriented lattice in n dimensions, a finite part of it is drawn in Fig. 1. Note that here we are dealing with an *infinite* set $\mathcal{M}$ for which in the formalism presented in the previous section in general technical problems associated with infinite sums arise. In the example under consideration we now sketch a transition to a formulation which then only makes reference to finitely generated $\mathcal{A}$-modules so that only finite sums appear and it is safe working on a purely algebraic level (see also Ref. 3).

Each $f\in\mathcal{A}$ can be written as a function of

$$x^\mu:=l_\mu\sum_a a^\mu e^a, \quad (5.2)$$

with $l_\mu\in\mathbb{R}$. Its differential is then given by

$$\mathrm{d}f = \sum_\mu \partial_{+\mu} f \ \mathrm{d}x^\mu, \tag{5.3}$$

where

$$(\partial_{+\mu} f)(x) := \frac{1}{l_\mu}[f(x+\boldsymbol{\mu}) - f(x)] \tag{5.4}$$

with $\boldsymbol{\mu} = l_\mu \hat{\mu}$. The 1-forms $\mathrm{d}x^\mu$ constitute a basis of Ω^1 as a left (or right) $\mathcal{A}$-module and satisfy the following commutation relations with a function of x^μ:

$$\mathrm{d}x^\mu \ f(x) = f(x+\boldsymbol{\mu}) \ \mathrm{d}x^\mu. \tag{5.5}$$

In particular, this implies

$$\mathrm{d}x^\nu \bullet \mathrm{d}x^\mu = [\mathrm{d}x^\mu, x^\nu] = l_\mu \delta^{\mu\nu} \mathrm{d}x^\mu \tag{5.6}$$

(cf. also Ref. 18) and, acting with d on the latter equation, leads to

$$\mathrm{d}x^\mu \ \mathrm{d}x^\nu + \mathrm{d}x^\nu \ \mathrm{d}x^\mu = 0. \tag{5.7}$$

The 1-form ρ introduced in (3.4) becomes

$$\rho = \sum_\mu \frac{1}{l_\mu} \mathrm{d}x^\mu. \tag{5.8}$$

It satisfies $\mathrm{d}\rho = 0$ and $\rho^2 = 0$. Moreover, for $w \in \Omega^r$ we have

$$\mathrm{d}w = \rho w - (-1)^r w \rho. \tag{5.9}$$

For a linear (left $\mathcal{A}$-module) connection on Ω^1 we write

$$\boldsymbol{\nabla} \mathrm{d}x^\mu = -\sum_\nu \Gamma^\mu{}_\nu \otimes_{\mathcal{A}} \mathrm{d}x^\nu, \quad \mathbb{U}(\mathrm{d}x^\mu) = \sum_\nu U^\mu{}_\nu \otimes_{\mathcal{A}} \mathrm{d}x^\nu. \tag{5.10}$$

Using (3.32), this leads to

$$U^\mu{}_\nu = \rho \delta^\mu_\nu + \Gamma^\mu{}_\nu =: \sum_\sigma \frac{1}{l_\sigma} U^\mu{}_{\sigma\nu} \ \mathrm{d}x^\sigma. \tag{5.11}$$

We shall require that $\lim_{\{l_\kappa\}\to 0} U^\mu_{\sigma\nu} = \delta^\mu_\nu$. This assumption will be used below where we work out continuum limits of curvature expressions.

The map κ introduced in Sec. IV is given by

$$\kappa(\mathrm{d}x^\mu \otimes_L \mathrm{d}x^\nu) = \sum_\sigma U^\nu{}_{\mu\sigma} \ \mathrm{d}x^\mu \otimes_{\mathcal{A}} \mathrm{d}x^\sigma. \tag{5.12}$$

For the left $\mathcal{A}$-linear $\cap$-product in Ω^2 we now obtain

$$\mathrm{d}x^\mu \cap \mathrm{d}x^\nu = \sum_\sigma U^\nu{}_{\mu\sigma} \ \mathrm{d}x^\mu \mathrm{d}x^\sigma. \tag{5.13}$$

Under a change of coordinates, $dx^\mu \cap dx^\nu$ transforms covariantly while $dx^\mu dx^\nu$ does not. Not all of the 2-forms $dx^\mu \cap dx^\nu$ are independent, in particular, as a consequence of (5.7). In the following we derive the relations which they satisfy under the assumption that κ has an inverse which means that $U^\mu{}_\nu$ has an inverse $V^\mu{}_\nu = \Sigma_\sigma (1/l_\sigma) V^\mu{}_{\sigma\nu}\, dx^\sigma$ in the sense that

$$\sum_\sigma U^\mu{}_\sigma \bullet V^\sigma{}_\nu = \rho\, \delta^\mu_\nu = \sum_\sigma V^\mu{}_\sigma \bullet U^\sigma{}_\nu . \tag{5.14}$$

In terms of components this becomes

$$\sum_\sigma U^\mu{}_{\alpha\sigma}\, V^\sigma{}_{\alpha\nu} = \delta^\mu_\nu = \sum_\sigma V^\mu{}_{\alpha\sigma}\, U^\sigma{}_{\alpha\nu}, \tag{5.15}$$

for all α. Now we have

$$dx^\mu dx^\nu = \sum_\sigma V^\nu{}_{\mu\sigma}\, dx^\mu \cap dx^\sigma. \tag{5.16}$$

We introduce

$$W^{\mu\nu}_{\rho\sigma} := U^\nu{}_{\mu\rho}\, V^\mu{}_{\rho\sigma}, \tag{5.17}$$

which satisfies $\lim_{\{l_\kappa\}\to 0} W^{\mu\nu}_{\rho\sigma} = \delta^\mu_\sigma \delta^\nu_\rho$ and

$$\sum_{\kappa,\lambda} W^{\mu\nu}_{\kappa\lambda}\, W^{\kappa\lambda}_{\rho\sigma} = \delta^\mu_\rho \delta^\nu_\sigma . \tag{5.18}$$

As a consequence,

$$(P^\pm)^{\mu\nu}_{\rho\sigma} := \frac{1}{2}(\delta^\mu_\rho \delta^\nu_\sigma \pm W^{\mu\nu}_{\rho\sigma}) \tag{5.19}$$

are projectors. In terms of the $\cap$-product, the 2-form relations (5.7) can now be expressed as follows:

$$\sum_{\kappa,\sigma} (P^+)^{\mu\nu}_{\kappa\sigma}\, dx^\kappa \cap dx^\sigma = 0. \tag{5.20}$$

This much more complicated form of the 2-form relations, as compared with (5.7), is the price we have to pay for the covariance. For a 2-form $A = \Sigma_{\mu,\nu} A_{\mu\nu}\, dx^\mu dx^\nu = \Sigma_{\mu,\nu} \hat{A}_{\mu\nu}\, dx^\mu \cap dx^\nu$ we obtain the implications

$$A = 0 \Leftrightarrow \sum_{\kappa,\sigma} (P^-)^{\kappa\sigma}_{\mu\nu}\, \hat{A}_{\kappa\sigma} = 0 \tag{5.21}$$

and

$$A_{\mu\nu} + A_{\nu\mu} = 0 \Leftrightarrow \sum_{\kappa,\sigma} (P^+)^{\kappa\sigma}_{\mu\nu}\, \hat{A}_{\kappa\sigma} = 0 \tag{5.22}$$

(since $A_{\mu\nu} = \Sigma_\rho \hat{A}_{\mu\rho} U^\rho{}_{\mu\nu}$).

With the help of (5.11), our general expression (2.8) for the torsion of a linear connection leads to

$$\Theta^\mu := \Theta(\mathrm{d}x^\mu) = \sum_{\nu,\rho} \frac{1}{l_\nu}(U^\mu{}_{\nu\rho} - \delta^\mu_\rho)\, \mathrm{d}x^\nu \mathrm{d}x^\rho = \sum_{\nu,\rho,\sigma} \frac{1}{l_\nu}(U^\mu{}_{\nu\rho} - \delta^\mu_\rho) V^\rho{}_{\nu\sigma}\, \mathrm{d}x^\nu \cap \mathrm{d}x^\sigma. \quad (5.23)$$

Writing

$$\Theta^\mu = \frac{1}{2}\sum_{\nu,\rho} Q^\mu{}_{\nu\rho}\, \mathrm{d}x^\nu \cap \mathrm{d}x^\rho, \quad (5.24)$$

where the coefficients $Q^\mu{}_{\nu\rho}$ are subject to

$$Q^\mu{}_{\nu\rho} = -\sum_{\kappa,\lambda} W^{\kappa\lambda}_{\nu\rho}\, Q^\mu{}_{\kappa\lambda}, \quad (5.25)$$

we are led to

$$Q^\mu{}_{\nu\rho} = \sum_{\kappa,\lambda,\sigma} \frac{1}{l_\kappa}(\delta^\kappa_\nu \delta^\lambda_\rho - W^{\kappa\lambda}_{\nu\rho})(U^\mu{}_{\kappa\sigma} - \delta^\mu_\sigma) V^\sigma{}_{\kappa\lambda}. \quad (5.26)$$

Example: If the torsion vanishes, we obtain

$$\frac{1}{l_\nu}(U^\mu{}_{\nu\rho} - \delta^\mu_\rho) = \frac{1}{l_\rho}(U^\mu{}_{\rho\nu} - \delta^\mu_\nu). \quad (5.27)$$

This is equivalent to the condition

$$\Gamma^\mu{}_{\nu\rho} = \Gamma^\mu{}_{\rho\nu}, \quad (5.28)$$

which is familar from continuum differential geometry. ■

A metric tensor (in the sense of Sec. III) is given by

$$g = \sum_{\mu,\nu} g_{\mu\nu}\, \mathrm{d}x^\mu \otimes_L \mathrm{d}x^\nu, \quad (5.29)$$

where $g_{\mu\nu}$ is now assumed to be a nondegenerate symmetric matrix. The metric compatibility condition $\boldsymbol{\nabla} g = 0$ with a linear connection $\boldsymbol{\nabla}$ leads to

$$g(x+\boldsymbol{\lambda})_{\mu\nu} = \sum_{\rho,\sigma} U(x)^\rho{}_{\lambda\mu}\, g(x)_{\rho\sigma}\, U(x)^\sigma{}_{\lambda\nu}, \quad (5.30)$$

for all λ. In matrix notation, this takes the form

$$g(x+\boldsymbol{\lambda}) = U(x)^t_\lambda\, g(x)\, U(x)_\lambda. \quad (5.31)$$

The continuum limit of this equation is obtained from the expansion

$$\begin{aligned} \tilde{g}_{\mu\nu} + l_\lambda(\partial_\lambda \tilde{g}_{\mu\nu} + b_{\mu\nu}) + \mathcal{O}(l^2_\mu) &= \sum_{\rho,\sigma} (\delta^\rho_\mu + l_\lambda \Gamma^\rho{}_{\lambda\mu}) g_{\rho\sigma} (\delta^\sigma_\nu + l_\lambda \Gamma^\sigma{}_{\lambda\nu}) \\ &= \tilde{g}_{\mu\nu} + l_\lambda\left(\sum_\rho (\tilde{\Gamma}^\rho{}_{\lambda\mu}\, \tilde{g}_{\rho\nu} + \tilde{g}_{\mu\rho}\, \tilde{\Gamma}^\rho{}_{\lambda\nu}) + b_{\mu\nu}\right) + \mathcal{O}(l^2_\mu), \end{aligned} \quad (5.32)$$

where

$$\tilde{\Gamma}^{\mu}{}_{\sigma\nu} := \lim_{\{l_\lambda\}\to 0} \Gamma^{\mu}{}_{\sigma\nu}, \quad \tilde{g}_{\mu\nu} := \lim_{\{l_\lambda\}\to 0} g_{\mu\nu}, \quad b_{\mu\nu} := \lim_{\{l_\lambda\}\to 0} \frac{\partial g_{\mu\nu}}{\partial l_\mu}, \tag{5.33}$$

which we assume to exist.

Remark: The vector fields $\partial_{+\mu} \in \mathfrak{X}$ are dual to the 1-forms dx^μ, i.e.,

$$\langle dx^\mu, \partial_{+\nu} \rangle = \delta^\mu_\nu. \tag{5.34}$$

The action of $X = \Sigma_\mu \partial_{+\mu} \cdot X^\mu$ on functions is given by

$$X(f) = \langle df, X \rangle = \sum_\mu X^\mu \, (\partial_{+\mu} f). \tag{5.35}$$

For the connection we have $\mathbb{U}(X) = X \otimes_{\mathcal{A}} \rho + \nabla X$, and thus

$$\mathbb{U}(\partial_{+\mu}) = \sum_\nu \partial_{+\nu} \otimes_{\mathcal{A}} U^\nu{}_\mu. \tag{5.36}$$

A dual metric tensor (cf. Sec. III) can be expressed as

$$h = \sum_{\mu,\nu} \partial_{+\mu} \otimes_R \partial_{+\nu} \cdot h^{\mu\nu}, \tag{5.37}$$

with components $h^{\mu\nu} \in \mathcal{A}$. The metric compatibility condition for a linear connection takes the form $\mathbb{U}(h) = h \otimes_{\mathcal{A}} \rho$. The latter leads to

$$h(x+\boldsymbol{\lambda})^{\mu\nu} = \sum_{\rho,\sigma} V(x)^\mu{}_{\lambda\rho} \, V(x)^\nu{}_{\lambda\sigma} \, h(x)^{\rho\sigma}. \tag{5.38}$$

With $h^{\mu\nu} = g^{\mu\nu}$, where $g^{\mu\nu}$ are the components of the matrix inverse to $(g_{\mu\nu})$, we obtain the metric tensor inverse to g. ■

Let us now turn to the calculation of the curvature of a linear connection. We have

$$\begin{aligned}
\mathfrak{R}(dx^\mu) &= \sum_\nu \left(d\Gamma^\mu{}_\nu + \sum_\rho \Gamma^\mu{}_\rho \Gamma^\rho{}_\nu \right) \otimes_{\mathcal{A}} dx^\nu \\
&= \sum_{\rho,\nu} U^\mu{}_\rho \, U^\rho{}_\nu \otimes_{\mathcal{A}} dx^\nu \\
&= \sum_{\rho,\kappa,\lambda,\nu} \frac{1}{l_\kappa l_\lambda} U(x)^\mu{}_{\kappa\rho} \, U(x+\boldsymbol{\kappa})^\rho{}_{\lambda\nu} \, dx^\kappa dx^\lambda \otimes_{\mathcal{A}} dx^\nu \\
&= \frac{1}{2} \sum_{\kappa,\lambda,\nu} \frac{1}{l_\kappa l_\lambda} [U(x)_\kappa \, U(x+\boldsymbol{\kappa})_\lambda - U(x)_\lambda \, U(x+\boldsymbol{\lambda})_\kappa]^\mu{}_\nu \, dx^\kappa dx^\lambda \otimes_{\mathcal{A}} dx^\nu.
\end{aligned} \tag{5.39}$$

With

$$\mathfrak{R}(dx^\mu) =: \frac{1}{2} \sum_{\kappa,\lambda,\nu} R^\mu{}_{\nu\kappa\lambda} \, dx^\kappa dx^\lambda \otimes_{\mathcal{A}} dx^\nu, \tag{5.40}$$

where $R^\mu{}_{\nu\kappa\lambda} = -R^\mu{}_{\nu\lambda\kappa}$, we thus have

$$R^\mu{}_{\nu\kappa\lambda} = \frac{1}{l_\kappa l_\lambda} [U(x)_\kappa \, U(x+\boldsymbol{\kappa})_\lambda - U(x)_\lambda \, U(x+\boldsymbol{\lambda})_\kappa]^\mu{}_\nu. \tag{5.41}$$

To obtain the tensorial components of the curvature, we need to transform $\otimes_{\mathcal{A}}$ into $\otimes_L$ and the $dx^\kappa dx^\lambda$ into $dx^\kappa \cap dx^\lambda$. We achieve this with $\lambda = \boldsymbol{\kappa}^{-1}$. First we note that

$$\lambda(dx^\mu \otimes_{\mathcal{A}} dx^\nu) = \sum_\rho V(x)^\nu{}_{\mu\rho}\, dx^\mu \otimes_L dx^\rho, \tag{5.42}$$

and therefore[31]

$$\begin{aligned}\lambda(dx^\mu dx^\nu \otimes_{\mathcal{A}} dx^\rho) &= \frac{1}{2}\, dx^\mu\left(\sum_\lambda V(x)^\rho{}_{\nu\lambda}\, dx^\nu \otimes_L dx^\lambda\right) - \frac{1}{2}\, dx^\nu\left(\sum_\lambda V(x)^\rho{}_{\mu\lambda}\, dx^\mu \otimes_L dx^\lambda\right)\\ &= \frac{1}{2}\sum_{\lambda,\sigma}\{V(x)^\lambda{}_{\mu\sigma}\, dx^\mu(V(x)^\rho{}_{\nu\lambda} dx^\nu) - V(x)^\lambda{}_{\nu\sigma}\, dx^\nu(V(x)^\rho{}_{\mu\lambda} dx^\mu)\}\otimes_L dx^\sigma\\ &= \frac{1}{2}\sum_\sigma [V(x+\boldsymbol{\mu})_\nu\, V(x)_\mu + V(x+\boldsymbol{\nu})_\mu\, V(x)_\nu]^\rho{}_\sigma\, (dx^\mu dx^\nu)\otimes_L dx^\sigma. \end{aligned} \tag{5.43}$$

Applying this formula, we find

$$\begin{aligned}\lambda\circ\mathfrak{R}(dx^\mu) = \frac{1}{4}\sum_{\kappa,\lambda,\nu}\frac{1}{l_\kappa l_\lambda}[(U(x)_\kappa\, U(x+\boldsymbol{\kappa})_\lambda - U(x)_\lambda\, U(x+\boldsymbol{\lambda})_\kappa)\\ \times(V(x+\boldsymbol{\kappa})_\lambda\, V(x)_\kappa + V(x+\boldsymbol{\lambda})_\kappa\, V(x)_\lambda)]^\mu{}_\nu\, dx^\kappa dx^\lambda \otimes_L dx^\nu.\end{aligned} \tag{5.44}$$

With

$$\lambda\circ\mathfrak{R}(dx^\mu) =: \sum_\nu \hat{R}^\mu{}_\nu \otimes_L dx^\nu, \tag{5.45}$$

this leads to

$$\hat{R}^\mu{}_\nu = \frac{1}{4}\sum_{\kappa,\lambda}\frac{1}{l_\kappa l_\lambda}[H(x)_{\kappa\lambda} - H(x)_{\lambda\kappa}]^\mu{}_\nu\, dx^\kappa dx^\lambda, \tag{5.46}$$

where

$$H(x)_{\kappa\lambda} := U(x)_\kappa\, U(x+\boldsymbol{\kappa})_\lambda\, V(x+\boldsymbol{\lambda})_\kappa\, V(x)_\lambda\,. \tag{5.47}$$

Expressing the 2-forms $\hat{R}^\mu{}_\nu$ as follows:

$$\hat{R}^\mu{}_\nu = \frac{1}{2}\sum_{\rho,\sigma}\hat{R}^\mu{}_{\nu\rho\sigma}\, dx^\rho \cap dx^\sigma, \tag{5.48}$$

with tensorial coefficients subject to

$$\hat{R}^\mu{}_{\nu\rho\sigma} = -\sum_{\kappa,\lambda} W^{\kappa\lambda}_{\rho\sigma}\, \hat{R}^\mu{}_{\nu\kappa\lambda}\,, \tag{5.49}$$

we get

$$\hat{R}^\mu{}_{\nu\kappa\lambda} = \frac{1}{2}\sum_\alpha \frac{1}{l_\kappa l_\alpha}[H(x)_{\kappa\alpha} - H(x)_{\alpha\kappa}]^\mu{}_\nu\, V(x)^\alpha{}_{\kappa\lambda}\,. \tag{5.50}$$

The resulting Ricci tensors are

$$\mathrm{Ric}_{\mu\nu}=\frac{1}{2}\sum_{\alpha,\beta}\frac{1}{l_\alpha l_\beta}[H(x)_{\beta\alpha}-H(x)_{\alpha\beta}]^\beta{}_\mu\ V(x)^\alpha{}_{\beta\nu}, \tag{5.51}$$

$$\overline{\mathrm{Ric}_{\mu\nu}}=\frac{1}{2}\sum_{\alpha,\beta}\frac{1}{l_\alpha l_\nu}[H(x)_{\nu\alpha}-H(x)_{\alpha\nu}]^\beta{}_\mu\ V(x)^\alpha{}_{\nu\beta}, \tag{5.52}$$

from which one obtains the curvature scalars $\hat{R}=g^{\mu\nu}\ \mathrm{Ric}_{\mu\nu}$ and $\overline{\hat{R}}=g^{\mu\nu}\ \overline{\mathrm{Ric}}_{\mu\nu}$ with the help of the inverse $g^{\mu\nu}$ of $g_{\mu\nu}$.

In order to elaborate the continuum limit of the curvature tensor, we use the expansions

$$U(x)_\kappa=I+l_\kappa\tilde{\Gamma}_\kappa+\frac{l_\kappa^2}{2}[(\tilde{\Gamma}_\kappa)^2+B_\kappa]+\mathcal{O}(l^3), \tag{5.53}$$

$$U(x+\boldsymbol{\lambda})_\kappa=I+l_\kappa\tilde{\Gamma}_\kappa+l_\kappa l_\lambda\partial_\lambda\tilde{\Gamma}_\kappa+\frac{l_\kappa^2}{2}[(\tilde{\Gamma}_\kappa)^2+B_\kappa]+\mathcal{O}(l^3), \tag{5.54}$$

$$V(x)_\kappa=I-l_\kappa\tilde{\Gamma}_\kappa+\frac{l_\kappa^2}{2}[(\tilde{\Gamma}_\kappa)^2-B_\kappa]+\mathcal{O}(l^3), \tag{5.55}$$

$$V(x+\boldsymbol{\lambda})_\kappa=I-l_\kappa\tilde{\Gamma}_\kappa-l_\kappa l_\lambda\partial_\lambda\tilde{\Gamma}_\kappa+\frac{l_\kappa^2}{2}[(\tilde{\Gamma}_\kappa)^2-B_\kappa]+\mathcal{O}(l^3). \tag{5.56}$$

This leads to

$$H(x)_{\kappa\lambda}=I+l_\kappa l_\lambda\ [\partial_\kappa\tilde{\Gamma}_\lambda+\tilde{\Gamma}_\kappa\ \tilde{\Gamma}_\lambda-\partial_\lambda\tilde{\Gamma}_\kappa+\tilde{\Gamma}_\lambda\ \tilde{\Gamma}_\kappa]+\mathcal{O}(l^3), \tag{5.57}$$

so that

$$\hat{R}^\mu{}_{\nu\kappa\lambda}=\partial_\kappa\tilde{\Gamma}^\mu{}_{\lambda\nu}-\partial_\lambda\tilde{\Gamma}^\mu{}_{\kappa\nu}+\tilde{\Gamma}^\mu{}_{\kappa\rho}\ \tilde{\Gamma}^\rho{}_{\lambda\nu}-\tilde{\Gamma}^\mu{}_{\lambda\rho}\ \tilde{\Gamma}^\rho{}_{\kappa\nu}+\mathcal{O}(l). \tag{5.58}$$

In this way we recover the continuum Riemann tensor in the limit $\{l_\alpha\}\to 0$.

We have set up a formalism which assigns geometrical notions like metric, curvature and Ricci tensor to a hypercubic lattice. In particular, one obtains a discrete counterpart of the Einstein (vacuum) equations in this way. Actually, there are several discrete Einstein equations depending on our choice of Ricci tensor. The results of the following section suggest that the difference $\mathrm{Ric}-\overline{\mathrm{Ric}}$ is the appropriate object.

Remark: The maps κ and λ extend to an arbitrary number of factors of the corresponding tensor products. We define

$$\kappa(\alpha_1\otimes_L\cdots\otimes_L\alpha_r):=(\mathrm{id}\otimes_{\mathcal{A}}\ \kappa)[\alpha_1\bullet\mathbb{U}(\alpha_2\otimes_L\cdots\otimes_L\alpha_r)], \tag{5.59}$$

and correspondingly for λ. These maps allow us to introduce covariant components of higher order forms by expressing them in terms of

$$\alpha_1\cap\cdots\cap\alpha_r:=\pi\circ\kappa(\alpha_1\otimes_L\cdots\otimes_L\alpha_r). \tag{5.60}$$

These r-forms satisfy very complicated relations which generalize (5.20) and involve the curvature, in general. ■

VI. DISCRETE SURFACES OF REVOLUTION

In terms of coordinates ϑ,φ we consider the differential calculus determined by

$$\mathrm{d}\vartheta\ f(\vartheta,\varphi)=f(\vartheta+l,\varphi)\ \mathrm{d}\vartheta,\quad \mathrm{d}\varphi\ f(\vartheta,\varphi)=f(\vartheta,\varphi+l)\ \mathrm{d}\varphi. \tag{6.1}$$

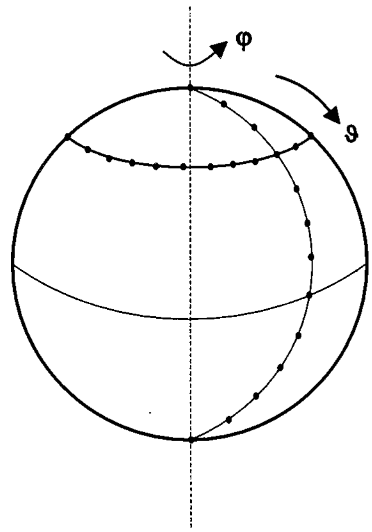

FIG. 2. Discretization of a sphere.

This is just a special case of (5.5). Via the rules of differential calculus it leads to

$$\mathrm{d}\vartheta \mathrm{d}\vartheta = 0, \quad \mathrm{d}\vartheta \mathrm{d}\varphi + \mathrm{d}\varphi \mathrm{d}\vartheta = 0, \quad \mathrm{d}\varphi \mathrm{d}\varphi = 0. \tag{6.2}$$

In contrast to the previous section, we interpret the coordinates as spherical coordinates where $\vartheta \in [0,\pi)$ and $\varphi \in [0,2\pi)$. With $l = \pi/n$, $n \in \mathbb{N}$, we obtain a discretization of the surface by fixing one point on the surface and moving in steps of coordinate length l in ϑ- and φ-directions. For the metric we make an ansatz,

$$g(\vartheta,\varphi) = \begin{pmatrix} 1 & 0 \\ 0 & b^2 \end{pmatrix}, \tag{6.3}$$

where b is a function of ϑ only. This models a surface of revolution (for example, a sphere as in Fig. 2).

Using $B := \mathrm{diag}(1,b)$, we have $g = B^t B$ and the metric compatibility condition for the parallel transport takes the form

$$(BU_\vartheta \tilde{B}^{-1})^t (BU_\vartheta \tilde{B}^{-1}) = I, \quad (BU_\varphi B^{-1})^t (BU_\varphi B^{-1}) = I, \tag{6.4}$$

where $\tilde{B} := \mathrm{diag}(1,\tilde{b})$ and $\tilde{b}(\vartheta) := b(\vartheta + l)$. As a consequence of these equations, $\tilde{B} U_\vartheta B^{-1}$ and $BU_\varphi B^{-1}$ are elements of the orthogonal group O(2). In order to obtain the correct continuum limit, we restrict them to be elements of SO(2), the component of O(2) which contains the identity. Then we have expressions

$$U_\vartheta = B^{-1} T(u) \tilde{B}, \quad U_\varphi = B^{-1} T(v) B, \tag{6.5}$$

where u,v are arbitrary functions of ϑ and φ and

$$T(\chi) = \begin{pmatrix} \cos\chi & -\sin\chi \\ \sin\chi & \cos\chi \end{pmatrix}. \tag{6.6}$$

The metric compatibility condition now leads to

$$U_\vartheta=\begin{pmatrix} \cos u & -\tilde{b}\sin u \\ (1/b)\sin u & (\tilde{b}/b)\cos u \end{pmatrix}, \quad U_\varphi=\begin{pmatrix} \cos v & -b\sin v \\ (1/b)\sin v & \cos v \end{pmatrix}, \tag{6.7}$$

and the condition of vanishing torsion becomes

$$\tilde{b}\sin u+\cos v=1, \quad \tilde{b}\cos u-\sin v=b. \tag{6.8}$$

These equations determine u and v completely in terms of b and $\tilde{b}$. We find

$$\cos u=\frac{1-p^2}{1+p^2}, \quad \sin u=\frac{2p}{1+p^2}, \quad \cos v=\frac{1+p^2-2\tilde{b}p}{1+p^2}, \quad \sin v=\frac{\tilde{b}-b-(\tilde{b}+b)p^2}{1+p^2}, \tag{6.9}$$

with

$$p=(2\tilde{b}\pm\sqrt{4\tilde{b}^2-(\tilde{b}^2-b^2)^2})/(b+\tilde{b})^2. \tag{6.10}$$

Only with the minus sign in the last expression we obtain the correct continuum limit (cf. Sec. V) where $\lim_{l\to 0}U_\vartheta=\lim_{l\to 0}U_\varphi=I$ (the unit matrix). This choice will be made in the following. The inverse parallel transport matrices are given by $V_\vartheta=\tilde{B}^{-1}T(-u)B$ and $V_\varphi=B^{-1}T(-v)B$, so that

$$V_\vartheta=\begin{pmatrix} \cos u & b\sin u \\ -(1/\tilde{b})\sin u & (b/\tilde{b})\cos u \end{pmatrix}, \quad V_\varphi=\begin{pmatrix} \cos v & b\sin v \\ -(1/b)\sin v & \cos v \end{pmatrix}. \tag{6.11}$$

With $\lambda=\kappa^{-1}$ (see Sec. IV), we obtain, for the curvature,

$$\lambda\circ\mathfrak{R}(\mathrm{d}x^\mu)=\sum_\nu r^\mu{}_\nu\,\mathrm{d}\vartheta\mathrm{d}\varphi\otimes_L\mathrm{d}x^\nu, \tag{6.12}$$

where $x^1=\vartheta, x^2=\varphi$ and

$$\begin{aligned} r&:=\frac{1}{2l^2}[U_\vartheta(\vartheta,\varphi)U_\varphi(\vartheta+l,\varphi)V_\vartheta(\vartheta,\varphi+l)V_\varphi(\vartheta,\varphi) \\ &\quad -U_\varphi(\vartheta,\varphi)U_\vartheta(\vartheta,\varphi+l)V_\varphi(\vartheta+l,\varphi)V_\vartheta(\vartheta,\varphi)] \\ &=\frac{1}{2l^2}B^{-1}[T(u)T(\tilde{v})T(-u)T(-v)-T(v)T(u)T(-\tilde{v})T(-u)]B \\ &=\frac{1}{2l^2}B^{-1}[T(\tilde{v}-v)-T(v-\tilde{v})]B=\frac{1}{l^2}\begin{pmatrix} 0 & -b\sin(\tilde{v}-v) \\ (1/b)\sin(\tilde{v}-v) & 0 \end{pmatrix}, \end{aligned} \tag{6.13}$$

with $\tilde{v}(\vartheta):=v(\vartheta+l)$. Since u and v are functions of b and $\tilde{b}$, they are functions of ϑ only. Using

$$\mathrm{d}\vartheta\mathrm{d}\varphi=V^\varphi_{\vartheta\vartheta}\,\mathrm{d}\vartheta\cap\mathrm{d}\vartheta+V^\varphi_{\vartheta\varphi}\,\mathrm{d}\vartheta\cap\mathrm{d}\varphi, \quad \mathrm{d}\varphi\mathrm{d}\vartheta=V^\vartheta_{\varphi\vartheta}\,\mathrm{d}\varphi\cap\mathrm{d}\vartheta+V^\vartheta_{\varphi\varphi}\,\mathrm{d}\vartheta\cap\mathrm{d}\varphi \tag{6.14}$$

and $r\mathrm{d}\vartheta\mathrm{d}\varphi=\frac{1}{2}r(\mathrm{d}\vartheta\mathrm{d}\varphi-\mathrm{d}\varphi\mathrm{d}\vartheta)$, we find the curvature components

$$\hat{R}_{\vartheta\vartheta}=-\frac{\sin u}{\tilde{b}}r, \quad \hat{R}_{\vartheta\varphi}=\frac{b\cos u}{\tilde{b}}r, \quad \hat{R}_{\varphi\vartheta}=-(\cos v)r, \quad \hat{R}_{\varphi\varphi}=-b(\sin v)r, \tag{6.15}$$

where $\hat{R}_{\kappa\lambda}=(\hat{R}^\mu{}_{\nu\kappa\lambda})$. We have the two Ricci tensors:

$$\mathrm{Ric} = \frac{1}{l^2}\begin{pmatrix} -(1/b)\cos v & -\sin v \\ (b/\tilde{b})\sin u & -(b^2/\tilde{b})\cos u \end{pmatrix} \sin(\tilde{v}-v), \tag{6.16}$$

$$\overline{\mathrm{Ric}} = \frac{1}{l^2}\begin{pmatrix} (1/\tilde{b})\cos u & -\sin v \\ (b/\tilde{b})\sin u & b\cos v \end{pmatrix} \sin(\tilde{v}-v), \tag{6.17}$$

and the combination

$$\widetilde{\mathrm{Ric}} := \frac{1}{2}(\mathrm{Ric} - \overline{\mathrm{Ric}}) = -\frac{1}{2l^2}\left(\frac{\cos u}{\tilde{b}} + \frac{\cos v}{b}\right)\sin(\tilde{v}-v)g, \tag{6.18}$$

from which we obtain the curvature scalars[32]

$$\hat{R} := g^{\mu\nu}\,\mathrm{Ric}_{\mu\nu} = -\frac{1}{l^2}\left(\frac{\cos u}{\tilde{b}} + \frac{\cos v}{b}\right)\sin(\tilde{v}-v), \tag{6.19}$$

$$\bar{\hat{R}} := g^{\mu\nu}\,\overline{\mathrm{Ric}}_{\mu\nu} = -\hat{R}, \tag{6.20}$$

$$\tilde{R} := g^{\mu\nu}\widetilde{\mathrm{Ric}}_{\mu\nu} = \hat{R}. \tag{6.21}$$

Now (6.18) becomes

$$\widetilde{\mathrm{Ric}}_{\mu\nu} = \frac{1}{2}\tilde{R}\, g_{\mu\nu}. \tag{6.22}$$

This result clearly distinguishes the particular linear combination (6.18) of Ricci tensors.

In the following, we present expansions in powers of l and consider the continuum limit $l \to 0$. We shall allow an explicit dependence of b on l, i.e., $b(\vartheta,l) = b_0(\vartheta) + b_1(\vartheta)l + \mathcal{O}(l^2)$. Then

$$\Gamma_\vartheta = \frac{1}{l}(U_\vartheta - I) = \begin{pmatrix} 0 & 0 \\ 0 & b_0'/b_0 \end{pmatrix} + \begin{pmatrix} 0 & -(b_0'^2/2) \\ b_0'^2/2b_0^2 & (2b_1' + b_0'')/2b_0 - b_1 b_0'/b_0^2 \end{pmatrix} l + \mathcal{O}(l^2), \tag{6.23}$$

$$\begin{aligned}\Gamma_\varphi &= \frac{1}{l}(U_\varphi - I) \\ &= \begin{pmatrix} 0 & -b_0 b_0' \\ b_0'/b_0 & 0 \end{pmatrix} + \begin{pmatrix} -b_0'^2/2 & -[b_1 b_0' + b_0 b_1' + b_0 b_0''/2] \\ (2b_1' + b_0'')/2b_0 - b_1 b_0'/b_0^2 & -b_0'^2/2 \end{pmatrix} l + \mathcal{O}(l^2),\end{aligned}$$

where b' denotes the derivative of b with respect to ϑ. For the curvature, we find $\hat{R}_{\vartheta\vartheta} = \mathcal{O}(l^2)$ and

$$\begin{aligned}\hat{R}_{\vartheta\varphi} &= \begin{pmatrix} 0 & -b_0 b_0'' \\ b_0''/b_0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & (-b_1 + b_0')b_0'' - b_0(b_1'' + b_0''') \\ -[(b_1 + b_0')b_0''/b_0^2 - (b_1'' + b_0''')/b_0] & 0 \end{pmatrix} l \\ &\quad + \mathcal{O}(l^2),\end{aligned}$$

$$\hat{R}_{\varphi\vartheta} = \begin{pmatrix} 0 & b_0 b_0' \\ -b_0''/b_0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & b_1 b_0'' + b_0(b_1'' + b_0''') \\ -[b_1 b_0'' - b_0(b_1'' + b_0''')]/b_0^2 & 0 \end{pmatrix} l + \mathcal{O}(l^2), \tag{6.24}$$

$$\hat{R}_{\varphi\varphi}=\begin{pmatrix} 0 & b_0^2 b_0' b_0'' \\ -b_0' b_0'' & 0 \end{pmatrix} l+\mathcal{O}(l^2).$$

The Ricci tensors have the following expansions:

$$\mathrm{Ric}=\begin{pmatrix} -b_0''/b_0 & 0 \\ 0 & -b_0 b_0'' \end{pmatrix}+\begin{pmatrix} [b_1 b_0''-b_0(b_1''+b_0''')]/b_0^2 & -b_0' b_0'' \\ 0 & (-b_1+b_0')b_0''-b_0(b_1''+b_0''') \end{pmatrix} l+\mathcal{O}(l^2), \tag{6.25}$$

$$\overline{\mathrm{Ric}}=\begin{pmatrix} b_0''/b_0 & 0 \\ 0 & b_0 b_0'' \end{pmatrix}+\begin{pmatrix} -(b_1+b_0')b_0''/b_0^2+(b_1''+b_0''')/b_0 & -b_0' b_0'' \\ 0 & b_1 b_0''+b_0(b_1''+b_0''') \end{pmatrix} l+\mathcal{O}(l^2), \tag{6.26}$$

$$\widetilde{\mathrm{Ric}}=\left[-\frac{b_0''}{b_0}+\frac{2b_1 b_0''+b_0' b_0''-2b_0(b_1''+b_0''')}{2b_0^2} l+\mathcal{O}(l^2)\right] g_0, \tag{6.27}$$

where $g_0:=\mathrm{diag}(1,b_0^2)$. For the curvature scalar we obtain

$$\hat{R}=-\frac{2b_0''}{b_0}+\frac{2b_1 b_0''+b_0' b_0''-2b_0(b_1''+b_0''')}{b_0^2} l+\mathcal{O}(l^2). \tag{6.28}$$

Example: In ordinary continuum differential geometry, the standard geometry of the unit sphere is obtained with $b(\vartheta)=\sin\vartheta$. With this choice, we get

$$\hat{R}=2+l\cot\vartheta+\mathcal{O}(l^2) \tag{6.29}$$

in the discrete framework and in the limit $l\to 0$ we recover the continuum result $\hat{R}=2$. To first order, there is a dependence of the curvature scalar on ϑ. With the refined choice $b(\vartheta,l)=[1+\vartheta l/4+\mathcal{O}(l^2)]\sin\vartheta$, we get

$$\hat{R}=2+\mathcal{O}(l^2). \tag{6.30}$$

■

Our discrete version of curvature describes *finite* distances on a space in contrast to infinitesimal distances as expressed by tangent vectors in continuum differential geometry. This means that the metric components in the case under consideration have to be expected to depend on the discretization (which should be regarded as a discretization of a chart), i.e., on l in the case under consideration. We still have to understand how, for example, spherical symmetry can be formulated in our framework. Then, we should be able to determine a spherically symmetric metric as a suitable discrete counterpart of the Riemannian metric of the (continuum) sphere. Furthermore, it remains to be seen how this is related to the metric with constant curvature scalar, approximated in the above example.

VII. CONCLUSIONS

Within a framework of noncommutative geometry, we have presented a formalism of discrete Riemannian geometry which is very much analogous to continuum Riemannian geometry.

Whereas the general formalism of noncommutative geometry suggests to consider a (generalized) metric tensor as an element of $\Omega^1\otimes_{\mathcal{A}}\Omega^1$, in this paper it was taken to be an element of $\Omega^1\otimes_L\Omega^1$ since a simple geometric meaning can be assigned to its components (with respect to the canonical basis e^{ij} of Ω^1; cf. Sec. III).[33]

The compatibility condition $\nabla g = 0$ for a metric and a linear connection on a finite set, when expressed in terms of parallel transport matrices, leads to relations (cf. Sec. III E) which are in complete accordance with what one should expect on the basis of a reformulation of metric compatibility in terms of parallel transport in (continuum) differential geometry.

An important role in ordinary differential geometry and especially in General Relativity is played by the Ricci tensor and the curvature scalar. There is no generalization of these tensors to the general framework of noncommutative geometry. In the case of a discrete set, we considered this problem in some detail in Sec. IV and showed that, with certain restrictions on the differential calculus (and thus the links between the points of the set), satisfactory candidates for discrete counterparts of the continuum Ricci tensor and curvature scalar do exist. The examples treated in Secs. IV–VI demonstrate how our definitions work. It should be quite evident by now that general definitions can hardly be expected since in noncommutative geometry, and already with a commutative algebra $\mathcal{A}$, we are dealing with a huge variety of structures of which only few should be expected to be close (in some sense) to continuum differential geometry.

In the last two sections we have developed discrete differential geometry on a hypercubic lattice. Since we were able to construct a Ricci tensor and a curvature scalar in this case, discrete counterparts of the (vacuum) Einstein equations are obtained. The results of the last section suggest to choose the following version:

$$\widetilde{\mathrm{Ric}}_{\mu\nu} - \tfrac{1}{2}\tilde{R} g_{\mu\nu} = 0. \tag{7.1}$$

On the left hand side we have tensor components in the sense that they transform covariantly under a change of module basis in the space of 1-forms. It is straightforward to include matter fields in this scheme. The ''discrete gravity'' theory which we propose here is very different from earlier approaches which were either based on Regge calculus,[28] other simplicial complex structures,[34] or on a certain reformulation of gravity as a gauge theory.[35] The correspondence between first order differential calculi on discrete sets and digraphs relates our formalism to the spin network approach to (quantum) gravity (see Ref. 36, in particular), at least on a basic level.

ACKNOWLEDGMENTS

A.D. is grateful to Professor Theo Geisel for financial support during a stay at the Max-Planck-Institut für Strömungsforschung. F. M.-H. would like to thank John Madore for a discussion.

[1] A. Dimakis and F. Müller-Hoissen, ''Quantum mechanics on a lattice and q-deformations,'' Phys. Lett. B **295**, 242 (1992).
[2] A. Dimakis and F. Müller-Hoissen, ''Differential calculus and gauge theory on finite sets,'' J. Phys. A **27**, 3159 (1994).
[3] A. Dimakis and F. Müller-Hoissen, ''Discrete differential calculus, graphs, topologies and gauge theory,'' J. Math. Phys. **35**, 6703 (1994).
[4] A. Dimakis, F. Müller-Hoissen, and F. Vanderseypen, ''Discrete differential manifolds and dynamics on networks,'' J. Math. Phys. **36**, 3771 (1995).
[5] K. Bresser, A. Dimakis, F. Müller-Hoissen, and A. Sitarz, ''Non-commutative geometry of finite groups,'' J. Phys. A **29**, 2705 (1996).
[6] A. Dimakis, F. Müller-Hoissen, and T. Striker, ''Noncommutative differential calculus and lattice gauge theory,'' J. Phys. A **26**, 1927 (1993).
[7] A. Dimakis and F. Müller-Hoissen, ''Integrable discretizations of chiral models via deformation of the differential calculus,'' J. Phys. A **29**, 5007 (1996).
[8] A. Connes, *Noncommutative Geometry* (Academic, San Diego, 1994).
[9] A. Dimakis and F. Müller-Hoissen, ''Connes' distance function on one-dimensional lattices,'' Int. J. Theor. Phys. **37**, 907 (1998).
[10] A. Sitarz, ''Gravity from noncommutative geometry,'' Class. Quantum Grav. **11**, 2127 (1994); ''On some aspects of linear connections in noncommutative geometry,'' hep-th/9503103.
[11] M. Dubois-Violette, J. Madore, T. Masson, and J. Mourad, ''Linear connections on the quantum plane,'' Lett. Math. Phys. **35**, 351 (1995).
[12] I. Heckenberger and K. Schmüdgen, ''Levi-Civita connections on the quantum groups $SL_q(N)$, $O_q(N)$ and $Sp_q(N)$,'' Commun. Math. Phys. **185**, 177 (1997).
[13] A. Dimakis, ''A note on connections and bimodules,'' q-alg/9603001.

[14] A. Dimakis and F. Müller-Hoissen, ''Deformations of classical geometries and integrable systems,'' physics/9712002; ''Noncommutative geometry and a class of completely integrable models,'' math-ph/9809023.

[15] S. Cho and K.S. Park, ''Linear connections on graphs,'' J. Math. Phys. **38**, 5889 (1997).

[16] R. R. Zapatrin, ''Polyhedral representations of discrete differential manifolds,'' J. Math. Phys. **38**, 2741 (1997).

[17] Instead of $\Omega(\mathcal{A})$ we simply write Ω in the following.

[18] H. C. Baehr, A. Dimakis, and F. Müller-Hoissen, ''Differential calculi on commutative algebras,'' J. Phys. A **28**, 3197 (1995).

[19] In general, a differential ideal is generated by linear combinations of basic forms.

[20] We use the same symbol $\boldsymbol{\nabla}$ for the connection and its dual.

[21] This is no longer so when Ω^2 is smaller than $\Omega^1 \otimes_{\mathcal{A}} \Omega^1$.

[22] At this point it is worth not to impose additional conditions. Finally, we will be interested in g being *real* and *symmetric* [i.e., $g(i)_{jk} = g(i)_{kj}$], or *Hermitian*. We refer to $g(i)_{jk}$ as the components of a ''metric'' at i in order to emphasize a certain analogy with a metric tensor in continuum differential geometry. However, a better name would be *distance matrix* of the digraph at i. In general, $g(i)$ will be degenerate.

[23] More generally, let us consider a graph embedded in some affine space $\mathbb{R}^d$, $d \in \mathbb{N}$, with inner product $(,)$. Hence, there is a map $\mathbf{x}: \mathcal{M} \to \mathbb{R}^d$ with $\mathbf{x} = \sum_{i \in \mathcal{M}} \mathbf{x}_i e^i$. Given a (first order) differential calculus on $\mathcal{M}$, we have $d\mathbf{x} = \sum_{i,j} (\mathbf{x}_j - \mathbf{x}_i) e^{ij}$. The inner product then induces a metric on $\mathcal{M}$ via $g(i)_{jk} = (\mathbf{x}_j - \mathbf{x}_i, \mathbf{x}_k - \mathbf{x}_i)$. If the inner product is the Euclidean one, then we have (3.53).

[24] Our formalism admits nonstandard geometries, however. For example, measuring the (not necessarily spatial) ''distances'' from i to j and from j to i in some (in a generalized sense) anisotropic space may lead to different results. This can be taken into account by dropping the restriction $l_{ij} = l_{ji}$.

[25] Note that $g(i)_{ik}$ and $g(i)_{ki}$ do not appear in (3.52) and have to be interpreted as 0 in the following formulas.

[26] Here ''if $e^{ikl} = 0$'' should be interpreted as ''if e^{ikl} is not present in the differential calculus.'' This abuse of notation has the great advantage of being much more concise and will therefore be repeatedly used in the following.

[27] L. Brewin, ''Riemann normal coordinates, smooth lattices and numerical relativity,'' gr-qc/9701057.

[28] T. Regge, ''General relativity without coordinates,'' Nuovo Cimento A **19**, 558 (1961); R. M. Williams and P. A. Tuckey, ''Regge calculus: a brief review and bibliography,'' Class. Quantum Grav. **9**, 1409 (1992).

[29] This shows that left $\mathcal{A}$-homomorphisms $\Omega^1 \otimes_L \Omega^1 \to \Omega^1 \otimes_{\mathcal{A}} \Omega^1$ are in one-to-one correspondence with left $\mathcal{A}$-module linear connections.

[30] If $e^i \alpha e^j \neq 0$, then $i \to j$ with which the parallel transport $\mathbb{U}^{ij}$ is associated. But instead, $\mathbb{U}^{ji}$ enters the definition (4.6) of λ_1. Therefore the symmetry condition is needed.

[31] The intermediate result in the second line is not well-defined, but helps to understand how the final formula is obtained.

[32] The geometrically interesting condition of a *constant* curvature scalar translates into a complicated difference equation for $b(\vartheta)$, $[\cos u(\vartheta)/b(\vartheta+l) + \cos v(\vartheta)/b(\vartheta)] \sin[v(\vartheta+l) - v(\vartheta)] = \text{const}$, where $b(\vartheta+l) \sin u(\vartheta) + \cos v(\vartheta) = 1$ and $b(\vartheta+l) \cos u(\vartheta) - \sin v(\vartheta) = b(\vartheta)$.

[33] In the case of a commutative algebra $\mathcal{A}$, one can think of replacing more generally $\otimes_{\mathcal{A}}$ by $\otimes_L$ in basic definitions like that of a connection. For a noncommutative differential calculus, this turns out to be inconsistent with the Leibniz rule, however. Also, it should be clear that the connection must be a nonlocal object, in contrast to something like a metric tensor.

[34] D. Weingarten, ''Geometric formulation of electrodynamics and general relativity in discrete space-time,'' J. Math. Phys. **18**, 165 (1977); A. N. Jourjine, ''Discrete gravity without coordinates,'' Phys. Rev. D **35**, 2983 (1987).

[35] A. Das, M. Kaku, and P. K. Townsend, ''Lattice formulation of general relativity,'' Phys. Lett. B **81**, 11 (1979); L. Smolin, ''Quantum gravity on a lattice,'' Nucl. Phys. B **148**, 333 (1979); K. I. Macrea, ''Rotationally invariant field theory on lattices. III. Quantizing gravity by means of lattices,'' Phys. Rev. D **23**, 900 (1981); C. L. T. Mannion and J. G. Taylor, ''General relativity on a flat lattice,'' Phys. Lett. B **100**, 261 (1981); K. Kondo, ''Euclidean quantum gravity on a flat lattice,'' Prog. Theor. Phys. **72**, 841 (1984); M. Caselle, A. D'Adda, and L. Magnea, ''Lattice gravity and supergravity as spontaneously broken gauge theories of the (super) Poincaré group,'' Phys. Lett. B **192**, 406 (1987); ''Doubling of all matter fields coupled to gravity on a lattice,'' *ibid.* **192**, 411 (1987); P. Renteln and L. Smolin, ''A lattice approach to spinorial quantum gravity,'' Class. Quantum Grav. **6**, 275 (1989); O. Boström, M. Miller, and L. Smolin, ''A new discretization of classical and quantum general relativity,'' preprint CGPG 94-3-3; R. Loll, ''Discrete approaches to quantum gravity in four dimensions,'' gr-qc/9805049.

[36] A. Ashtekar and J. Lewandowski, ''Quantum theory of geometry I: area operators,'' Class. Quantum Grav. **14**, A55 (1997); F. Markopoulou and L. Smolin, ''Causal evolution of spin networks,'' Nucl. Phys. B **508**, 409 (1997); M. P. Reisenberger, ''A lattice worldsheet sum for 4-d Euclidean general relativity,'' gr-qc/9711052.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Contractions of Lie algebras and separation of variables. The *n*-dimensional sphere

A. A. Izmest'ev, G. S. Pogosyan, and A. N. Sissakian
Joint Institute for Nuclear Research, Dubna, Moscow Region 141980, Russia

P. Winternitz[a)]
Centre de Recherches Mathématiques, Université de Montréal, C. P. 6128, succ. Centre Ville, Montréal, Québec H3C 3J7, Canada

(Received 1 April 1998; accepted for publication 24 September 1998)

Inönü–Wigner contractions from the rotation group $O(n+1)$ to the Euclidean group $E(n)$ are used to relate the separation of variables in Laplace–Beltrami operators on n-dimensional spheres and Euclidean spaces, respectively. In this article we consider all subgroup type coordinates corresponding to different chains of subgroups of $O(n+1)$ and $E(n)$, respectively. In particular, the contractions relate the graphical formalism of ''trees'' on spheres to the ''clusters'' on Euclidean spaces (introduced in this article). The contractions are considered analytically on several levels: the vector fields realizing the Lie algebras, the complete sets of commuting operators characterizing separable coordinate systems, the coordinate systems themselves and the separated eigenfunctions. © *1999 American Institute of Physics.* [S0022-2488(99)04102-X]

I. INTRODUCTION

Our purpose in this article is to use Lie algebra contractions to relate the separation of variables in Helmholtz equations on n-dimensional spheres S_n and on the Euclidean spaces E_n. An earlier article[1] was devoted to the case $n=2$. It was shown that spherical coordinates on S_2 can be contracted either to polar or Cartesian ones on E_2. Elliptic coordinates on S_2 were contracted to elliptic, parabolic and Cartesian ones on E_2.

The more complicated case of contractions from a two-dimensional Lorentzian hyperboloid H_2 to E_2 has also been studied.[2]

Here we are interested in the case of S_n for arbitrary n, but will only consider the simplest types of coordinates, the so-called subgroup type coordinates.[3–8] For S_n these are polyspherical coordinates introduced by Vilenkin[9,10] and described by the ''method of trees.''[9–13] Trees, or ''clusters'' can, of course, also be introduced to describe subgroup type coordinates in E_n, and we shall show how ''trees'' on S_n are related to ''clusters'' on E_n via the group contraction $O(n+1)\rightarrow E(n)$.

At least two definitions of Lie algebra contractions exist in the literature. The original Inönü–Wigner contractions[14–16] can be viewed as singular changes of bases. The more recent ''graded contractions''[17–23] are obtained as deformations of the original Lie algebra via modifications of the commutation relations, preserving a given grading of the Lie algebra. In many cases, though not all, the two concepts are equivalent.[23] In particular, the contractions considered in this article are simultaneously Inönü–Wigner and Z_2-graded ones.

Our main tool for dealing with contractions is the concept of ''analytic contractions,'' already introduced in Ref. 1. The generators of the original Lie algebra, in our case $o(n+1)$, are written as differential operators, involving the contraction parameters, in our case the radius R of the sphere. The parametrization must be such that in the contraction limit, in which the $o(n+1)$

[a)]Electronic mail: WINTERN@CRM.UMONTREAL.CA

algebra contracts to the $e(n)$ one, the generators themselves as differential operators, contract into generators of $e(n)$.

As a motivation for this study we mention, first of all, the theory of special functions. Indeed contractions relate two different groups and their homogeneous spaces. They relate separable coordinates in these two spaces, the separated equations and their solutions. The contractions will thus, in particular, provide asymptotic formulas and other relations between special functions.

Other applications concern the relations between integrable systems in different spaces, in particular, on spheres S_n and Euclidean spaces E_n. Indeed, each separable system can be extended by adding a potential that allows separation. The corresponding Hamiltonian systems will be integrable both on S_n and E_n, since they will also have n integrals of motion in involution. Again, the contractions relate the S_n and E_n integrable systems and their solutions.

In Sec. II we review some known results on the method of trees for S_n.[9–13] We introduce $O(n)$ subgroup diagrams and relate them to the tree diagrams. Section III is devoted to the separation of variables in Euclidean spaces E_n. We introduce $E(n)$ subgroup diagrams, E_n ''cluster'' diagrams, and relate the two. Beltrami coordinates are used in Sec. IV to introduce the radius of the sphere into the expressions for the elements of the $o(n+1)$ Lie algebras. This provides the tools for an analytical realization of the Lie algebra contraction $o(n+1)\rightarrow e(n)$. The contraction of the coordinate systems and the complete sets of commuting operators is presented in Sec. V. Finally, the asymptotic formulas representing contractions of the solutions of the Laplace–Beltrami equation on S_n to those of the Helmholtz equation on E_n are presented in Sec. VI.

II. SUBGROUP TYPE COORDINATES AND THE METHOD OF TREES

A. Subgroups of Lie groups and separable coordinates

We shall make use of an algebraic approach to separating variables in Helmholtz (and Hamilton–Jacobi) equations in Riemannian and pseudo-Riemannian spaces that are homogeneous spaces for some Lie group G.[3–8]

The equation that we are interested in can be written as

$$\Delta_{LB}\Psi = E\Psi, \quad \Delta_{LB} = \frac{1}{\sqrt{g}}\frac{\partial}{\partial \xi^i}\sqrt{g}g^{ij}\frac{\partial}{\partial \xi^j}, \quad g = |\det g_{ij}|, \tag{2.1}$$

where g_{ij} is the metric tensor written in the considered coordinates ξ_i. The space M can be identified with some factor space $M\sim G/G_0$, where G_0 is the isotropy group of the origin.

The separated solutions of Eq. (2.1) are simultaneous eigenfunctions of some complete set of n commuting operators Y_a (including the Laplace–Beltrami operator). We thus have

$$Y_a\Psi = \lambda_a\Psi, \quad \Psi = \prod_{i=1}^{n} \Psi_i(\zeta_i;\lambda_1,\lambda_2,\ldots,\lambda_n). \tag{2.2}$$

The operators Y_a are second order operators in the enveloping algebra of the Lie algebra of the isometry group G. Thus we have a Lie algebra L with basis $L\sim\{X_1,\ldots,X_N\}$ and put

$$Y_a = A^a_{ik}X_iX_k, \quad [Y_a,Y_b]=0, \quad A^a_{ik}=A^a_{ki}; \quad a=1,2,\ldots,n. \tag{2.3}$$

The commuting sets of operators $\{Y_1,\ldots,Y_n\}$ can be classified into conjugacy classes under the action of the group G. Mutually conjugate sets provide equivalent systems of coordinates, transformed amongst each other by the group G.

A classification of the sets $\{Y_a\}$ provides a classification of coordinate systems. The essential properties of the coordinate systems are related to properties of the operators Y_a. In particular, ignorable coordinates[24] ξ_j (i.e., coordinates that do not figure in the metric tensor g_{ik}) are associated with operators Y_j that are squares of elements of the Lie algebra,

$$Y_j = \left\{ \sum_{k=1}^{N} a_{jk} X_k \right\}^2 = \frac{\partial^2}{\partial \alpha_j^2}. \tag{2.4}$$

Hence, maximal Abelian subalgebras[25–31] of the algebra L will provide maximal sets of ignorable variables.

Particularly simple coordinate systems are obtained if all operators Y_a in a given set are either squares of elements in the Lie algebra L, as in Eq. (2.4), or Casimir operators of subalgebras of L. Such coordinate systems have been called *subgroup type coordinates*.[6] Thus, consider a chain of subalgebras,

$$L \supset L_1 \supset L_2 \supset \cdots \supset L_M, \tag{2.5}$$

such that each subalgebra L_j has at least one second order Casimir operator (second order operator in the center of the enveloping algebra of L_j). Subgroup type coordinates are obtained if the chain of subalgebras provides n linearly independent second order operators. They will automatically commute amongst each other.

In this article we restrict our attention to subgroup type coordinates on spheres S_n and Euclidean spaces E_n. We mention that on S_2 precisely two types of separable coordinates exist. Spherical coordinates are subgroup type, the subgroup chain being $O(3) \supset O(2)$. Elliptic coordinates are not of the subgroup type. On S_3, six separable coordinate systems exist,[6,32,33] two of them of the subgroup type, corresponding to the chain $O(4) \supset O(3) \supset O(2)$ and $O(4) \supset O(2) \otimes O(2)$, respectively. For E_3, three out of eleven separable coordinate systems are of the subgroup type: Cartesian, cylindrical and spherical.

A graphical method, called the ''method of trees,'' has been developed to treat subgroup type coordinates on real and complex spheres.[9–13] We will reproduce some of the relevant results for *real* spheres S_n in the following subsection, and then extend them to analyze subgroup type coordinates on E_n. Moreover, we will connect the tree diagrams with subgroup diagrams, introduced below.

B. Subgroup type coordinates on S_n and the method of trees

Let us consider the Lie algebra $o(n+1)$ and use the standard basis of operators on S_n:

$$L_{ik} = (u_i \partial_k - u_k \partial_i);$$

$$[L_{ij}, L_{rs}] = -g_{js} L_{ir} - g_{ir} L_{js} + g_{jr} L_{is} + g_{is} L_{jr}, \quad 0 \leqslant i,k,j,r,s \leqslant n. \tag{2.6}$$

Let us now consider the defining representation of $o(n+1)$ by matrices

$$X \in \mathcal{R}^{(n+1)(n+1)}, \quad X^T + X = 0, \tag{2.7}$$

acting on the space $\mathcal{R}^{(n+1)}$. Maximal reducibly imbedded subalgebras of $o(n+1)$ will leave some vector subspace of $\mathcal{R}^n$ invariant. All subalgebras of this type have the form

$$o(n+1) \supset o(n_1) \oplus o(n_2), \quad n_1 + n_2 = n+1, \quad n_1 \geqslant n_2 \geqslant 2, \quad \text{or} \quad o(n+1) \supset o(n). \tag{2.8}$$

Maximal irreducibly imbedded subalgebras also exist, e.g., $u(n) \subset o(2n)$ or $g_2 \subset o(7)$, but they will not be needed here.

Chains of mutually maximally imbedded subalgebras are obtained by further splitting $o(n_1)$ and $o(n_2)$ into pairs of algebras, until we end the chain with one-dimensional subalgebras $o(2)$ [we drop all the $o(1) \sim \{0\}$ algebras]. We shall describe subalgebra chains by subalgebra diagrams (or equivalently subgroup diagrams). Each $O(k)$ subgroup is represented by a circle with the corresponding number k in it. All subgroup diagrams of this type are shown in Fig. 1 for $n \leqslant 5$. Their recursive character is obvious: different subgroup diagrams for a given $O(n)$ correspond to different flags of invariant subspaces of $\mathcal{R}$.

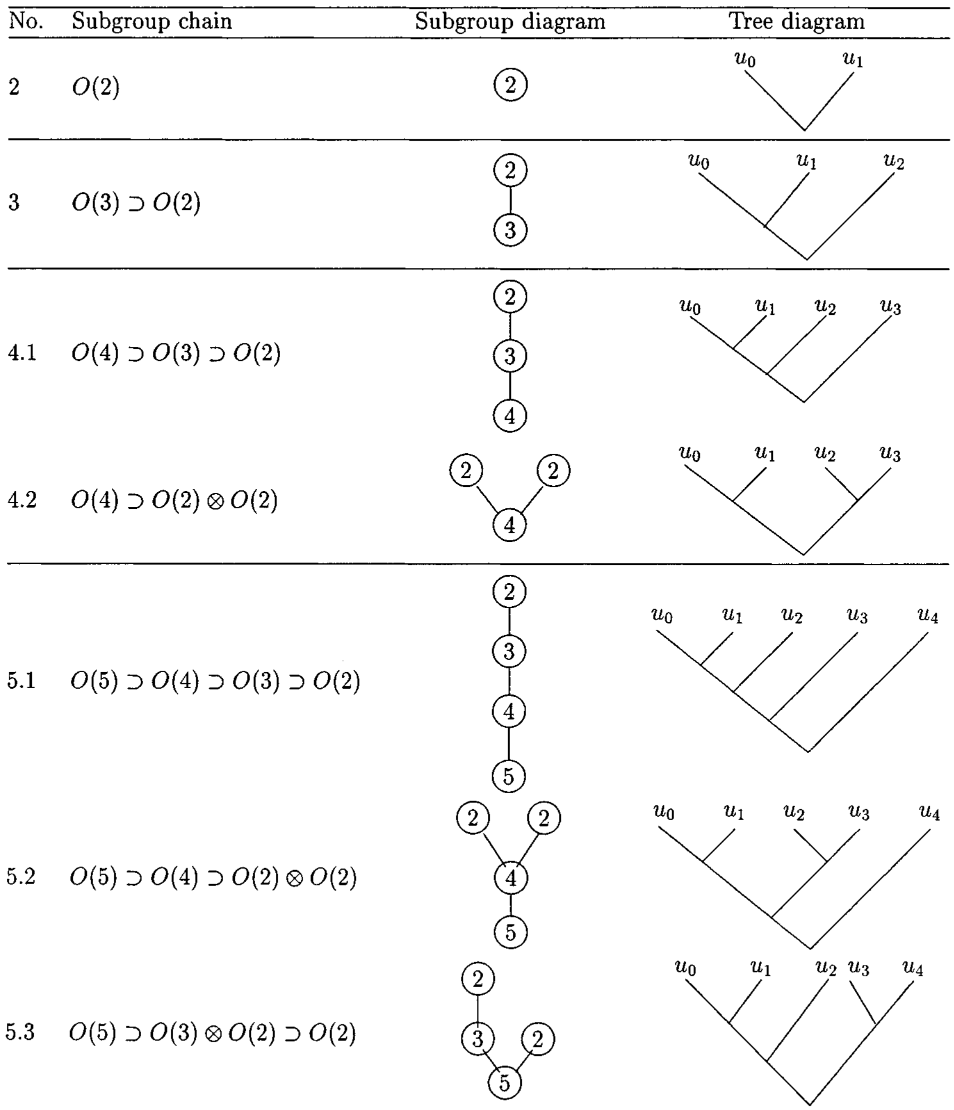

FIG. 1. Subgroup and tree diagrams for S_n.

The subgroup diagrams are closely related to the tree diagrams of Vilenkin,[9,10] describing polyspherical coordinates on S_n. In Fig. 1 we associate a tree diagram with each subgroup diagram for $2 \leqslant n \leqslant 5$. Families of different, but topologically equivalent, trees are associated with the same subgroup diagram. They are obtained either by permuting the end points, corresponding to the coordinates, or, equivalently, by rotating branches around branching points on the tree. All different trees, including equivalent ones, are shown for S_2, S_3, S_4 in Fig. 2.

The tree diagrams are best described in the original article[9] and the book.[13] Together with the subgroup diagrams described above, they provide a tool for writing coordinates on S_n, complete sets of commuting operators and their eigenvalues and separated solutions of the Helmholtz equation.

Let us recall some basic facts here, using the example of a specific tree, namely that in Fig. 3 for S_7. In Fig. 3(a) we give the corresponding $O(8)$ subgroup diagram. The actual S_7 tree is in Fig. 3(b). Figures 3(c) and 3(d) refer to the $E(7)$ group and E_7 space (after contraction) and will be used below.

Each end point on the tree of Fig. 3(b) corresponds to a Cartesian coordinate in the ambient

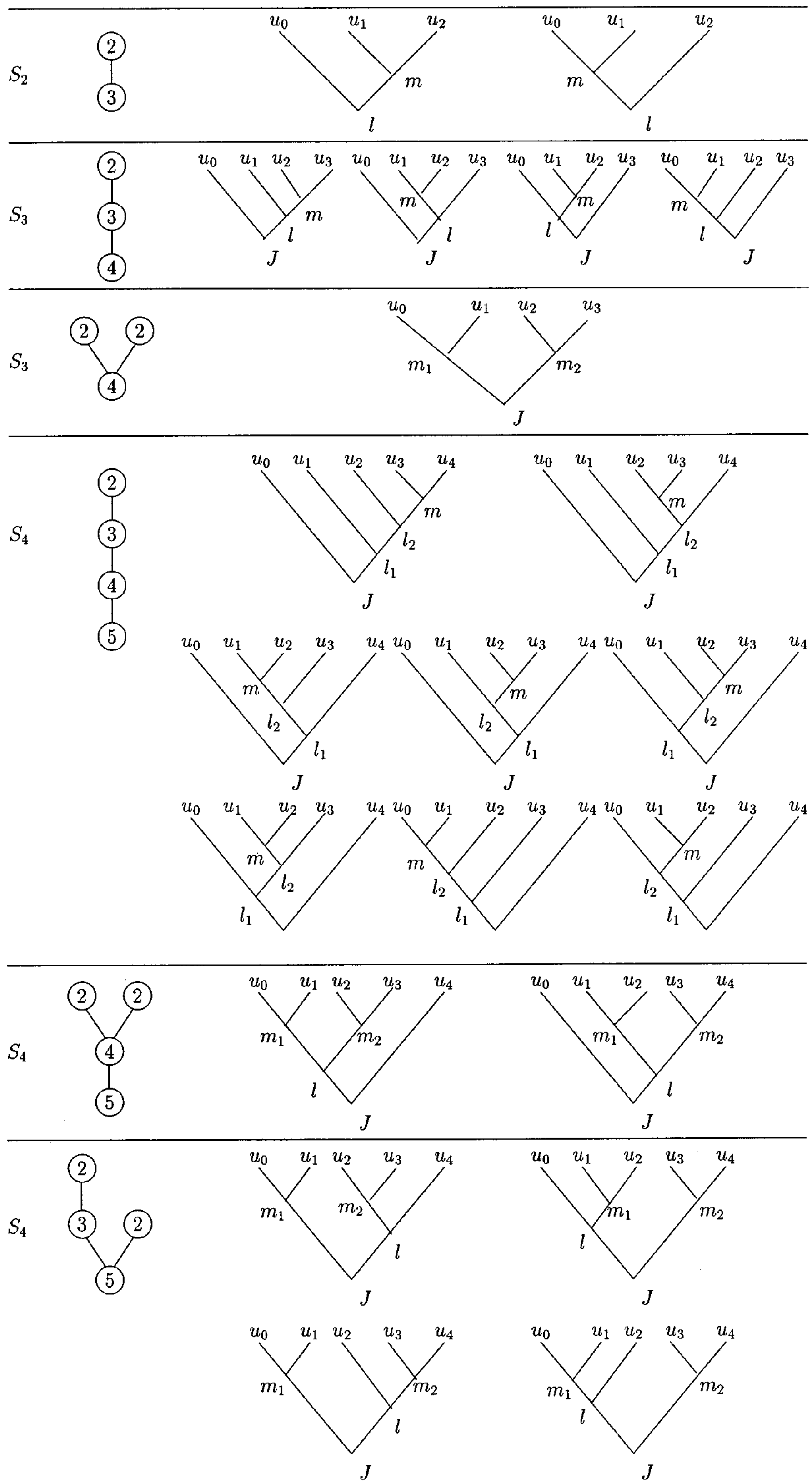

FIG. 2. Equivalent tree diagrams corresponding to one subgroup diagram for S_n.

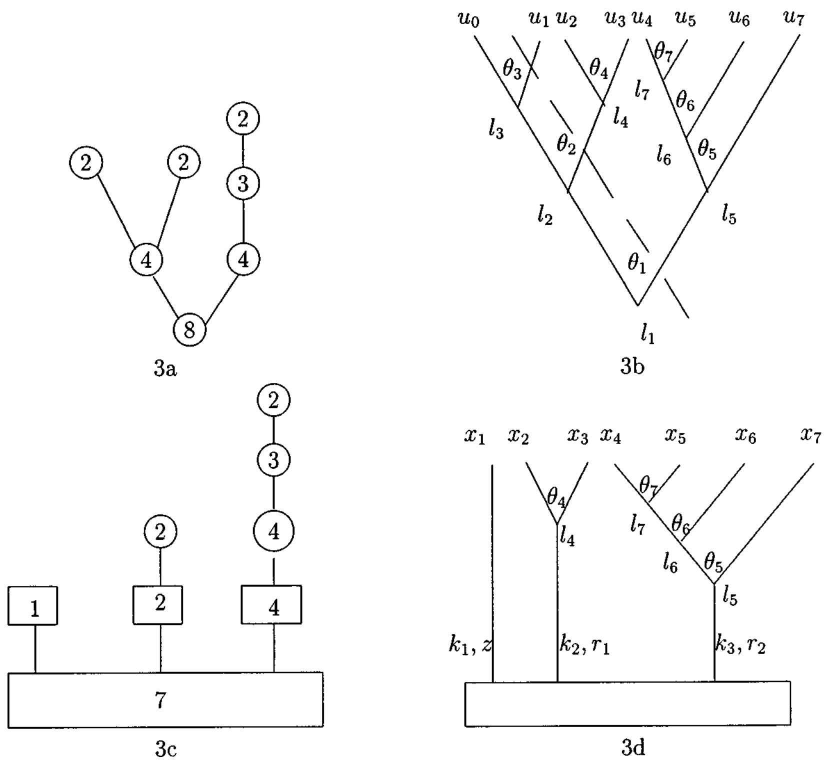

FIG. 3. Examples: An $O(8)$ subgroup diagram (3a) and the corresponding S_7 tree diagram (3b). An $E(7)$ subgroup diagram (3c) and the corresponding E_7 cluster diagram.

space $\mathcal{R}^8$. At each branching point we introduce an angle θ_i. We move along the tree from the ground upwards to a specific coordinate u_i. At each branching point we write $\cos\theta_a$ if we go to the left, $\sin\theta_a$ if we go to the right. The polyspherical coordinates corresponding to Fig. 3(b) hence are

$$\begin{aligned}
&u_0=R\cos\theta_1\cos\theta_2\cos\theta_3, \quad && u_4=R\sin\theta_1\cos\theta_5\cos\theta_6\cos\theta_7,\\
&u_1=R\cos\theta_1\cos\theta_2\sin\theta_3, \quad && u_5=R\sin\theta_1\cos\theta_5\cos\theta_6,\sin\theta_7,\\
&u_2=R\cos\theta_1\sin\theta_2\cos\theta_4, \quad && u_6=R\sin\theta_1\cos\theta_5\sin\theta_6,\\
&u_3=R\cos\theta_1\sin\theta_2\sin\theta_4, \quad && u_7=R\sin\theta_1\sin\theta_5.
\end{aligned} \tag{2.9}$$

The complete set of 7 commuting operators is also read off from the tree diagram, or from the subgroup one. We have

$$\begin{gathered}
Y_3=L_{01}^2, \quad Y_4=L_{23}^2, \quad Y_7=L_{45}^2, \quad Y_6=L_{45}^2+L_{56}^2+L_{46}^2,\\
Y_2=\sum_{0\leqslant i<k\leqslant 3} L_{ik}^2, \quad Y_5=\sum_{4\leqslant i<k\leqslant 7} L_{ik}^2, \quad Y_1=\sum_{0\leqslant i<k\leqslant 7} L_{ik}^2.
\end{gathered} \tag{2.10}$$

We see that Y_3, Y_4 and Y_7 are Casimir operators of $o(2)$ algebras, Y_6 of an $o(3)$ one, Y_2 and Y_5 correspond to $o(4)$ algebras and Y_1 is the original $o(8)$ Casimir operator. More generally, each circle in the subgroup chain provides the Casimir operator of the corresponding $o(k)$ to the set $\{Y_a\}$.

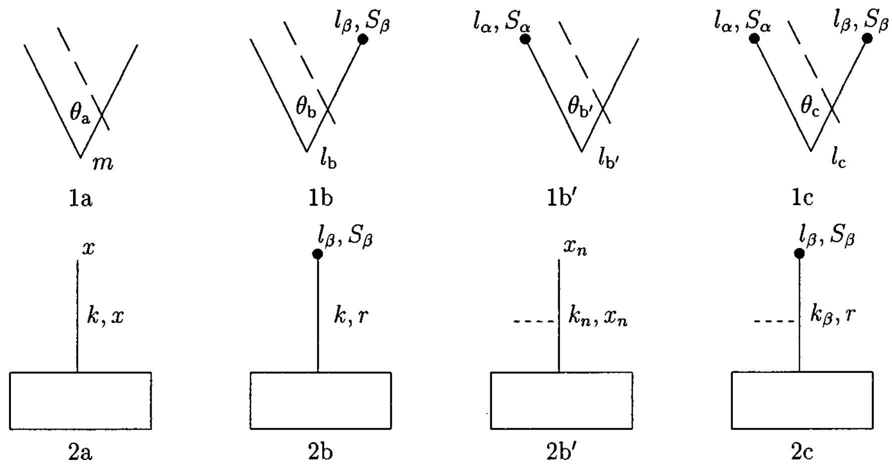

FIG. 4. Elementary cells for S_n (diagrams 1a,...,1c) and their contractions to E_n ones (diagrams 2a,...,2c).

To each branching point on the tree diagram, or each circle on the subgroup diagram, we also associate a quantum number l_j [see Fig. 3(b) for a specific case]. It will determine the eigenvalue λ of the corresponding $o(k)$ invariant operator according to the formula

$$Y_j\Psi=\Delta_{LB}\Psi=-l_j(l_j+k-2)\Psi, \tag{2.11}$$

where k is the dimension of the ambient space above the corresponding vertex on the tree [the same k as in $O(k)$]. The numbers l_j are non-negative integers, labeling irreducible representations of $O(k)$ for $k\geqslant 3$. For $k=2$, i.e., the group $O(2)$, we have $l_j=0,\pm 1,\pm 2,\ldots$.

C. The separated eigenfunctions for S_n

To specify the separated wave function,

$$\Psi=\Pi_{k=1}^{n}\Psi_k(\theta_k), \tag{2.12}$$

on S_n, we follow Refs. 9–13 and introduce four types of vertices, or ''cells'' on a tree, as illustrated in Fig. 4. The first row, diagrams 1a,...,1c, contains elementary S_n cells. The second row, 2a,...,2c contains E_n cells, obtained after a contraction, and will be discussed below in Sec. VI A. The dashed lines in row 1 will also be explained below. A circle on diagrams 1a,...1c denotes a ''closed'' end, i.e., one that leads to further branches. An open end (no circle) leads directly to a coordinate. For example, in Fig. 3 angles θ_3, θ_4 and θ_7 correspond to cells of type ''a,'' θ_5 and θ_6 to cells of type ''b′,'' and θ_1, θ_2 to cells of type ''c.'' The angles in the polyspheric coordinate systems satisfy

$$0\leqslant\theta_a<2\pi,\quad 0\leqslant\theta_b\leqslant\pi,\quad -\pi/2\leqslant\theta_{b'}\leqslant\pi/2,\quad 0\leqslant\theta_c\leqslant\pi/2. \tag{2.13}$$

The following numbers are associated with each cell: m, l, l_β, l_α are related to the separation constant corresponding to each vertex, S_α= number of vertices above vertex l_α, S_β= number of vertices above vertex l_β. The numbers m, l, l_β, l_α are all integers, labeling representations of the corresponding rotation subgroup in the chain, i.e., angular momentum type quantum numbers. We have

$$\Delta+c=n'-2, \tag{2.14}$$

where n' is the number of end points u_i connected to the vertex θ_j and c is the number of vertices above and to the left of vertex $\theta_{b'}$ or θ_c.

Each vertex and each angle θ_i provides a ''building block'' $\Psi_i(\theta_i)$ for the wave function $\Psi(\theta_1,\ldots,\theta_n)$ of Eq. (2.12). Specifically, we have the following.

(1) Cell of type a:

$$\Psi_m(\theta_a)=\frac{1}{\sqrt{2\pi}}e^{im\theta_a};\quad m=0,\pm1,\pm2,\ldots,\quad 0\leqslant\theta_a<2\pi. \tag{2.15}$$

(2) Cell of type b:

$$\Psi^{\alpha}_{n,l_\beta}(\theta_b)=N_n^{\alpha,\alpha}(\sin\theta_b)^{l_\beta}P_n^{(\alpha,\alpha)}(\cos\theta_b),$$

$$n=l-l_\beta,\quad \alpha=l_\beta+\frac{S_\beta}{2},\quad n=0,1,2,\ldots,\quad 0\leqslant\theta_b\leqslant\pi, \tag{2.16}$$

where $P_n^{(\alpha,\beta)}(x)$ is a Jacobi polynomial.

(3) Cell of type b′:

$$\Psi^{\beta}_{n,l_\alpha}(\theta_{b'})=N_n^{\beta,\beta}(\cos\theta_{b'})^{l_\alpha}P_n^{(\beta,\beta)}(\sin\theta_{b'});$$

$$n=l-l_\alpha,\quad \beta=l_\alpha+\frac{S_\alpha}{2},\quad n=0,1,2,\ldots,\quad -\pi/2\leqslant\theta_{b'}\leqslant\pi/2. \tag{2.17}$$

(4) Cell of type c:

$$\Psi^{\alpha,\beta}_{n,l_\beta,l_\alpha}(\theta_c)=2^{(\alpha+\beta)/2+1}N_n^{\alpha,\beta}(\sin\theta_c)^{l_\beta}(\cos\theta_c)^{l_\alpha}P_n^{(\alpha,\beta)}(\cos 2\theta_c);$$

$$n=\frac{l-l_\alpha-l_\beta}{2},\quad \alpha=l_\beta+\frac{S_\beta}{2},\quad \beta=l_\alpha+\frac{S_\alpha}{2},\quad n=0,1,2,\ldots,\quad 0\leqslant\theta_c\leqslant\pi/2. \tag{2.18}$$

The normalization constants are

$$N_n^{\alpha,\beta}=\sqrt{\frac{(2n+\alpha+\beta+1)\Gamma(n+\alpha+\beta+1)n!}{2^{\alpha+\beta+1}\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}}. \tag{2.19}$$

We mention that the wave functions (2.16) and (2.17) can also be expressed in terms of Gegenbauer polynomials, using the formula[34]

$$C_n^{\lambda}(x)=\frac{\Gamma(2\lambda+n)\Gamma(\lambda+1/2)}{\Gamma(2\lambda)\Gamma(\lambda+n+1/2)}P_n^{(\lambda-1/2,\lambda-1/2)}(x). \tag{2.20}$$

III. SUBGROUP TYPE COORDINATES ON E_n AND CLUSTER DIAGRAMS

Let us now consider the Euclidean Lie algebra $e(n)$, with a basis

$$L_{ik}=x_i\partial_{x_k}-x_k\partial_{x_i},\quad p_i=\partial_{x_i},\quad i,k=1,2,\ldots,n. \tag{3.1}$$

The commutation relations are, as in Eq. (2.6), together with

$$[p_j,L_{ik}]=\delta_{ji}p_k-\delta_{jk}p_i,\quad [p_i,p_k]=0. \tag{3.2}$$

Subalgebra chains (2.5) will include Euclidean subalgebras $e(k)$ and rotation subalgebras $o(k)$. A possible link in a subalgebra chain is

$$e(n)\supset e(n_1)\oplus e(n_2),\quad n_1+n_2=n,\quad n_1\geqslant n_2\geqslant 1. \tag{3.3}$$

The Casimir operator of $e(n)$ is

$$\Delta_n=p_1^2+p_2^2+\cdots+p_n^2. \tag{3.4}$$

Hence, we have $\Delta_n=\Delta_{n_1}+\Delta_{n_2}$ in the chain and only one of the Euclidean subalgebras (3.3) provides a new invariant operator, say $e(n_2)$. Alternatively, Δ_{n_1} and Δ_{n_2} can replace Δ_n. A further possible link in a chain is

$$e(n)\supset o(n),\quad n\geqslant 2, \tag{3.5}$$

where $o(n)$ will provide a new [with respect to $e(n)$] invariant operator.

As in the case of the $O(n)$ group we will introduce diagrams for the $E(n)$ group to illustrate subgroup chains and subgroup type coordinate systems on E_n Euclidean spaces. We shall use rectangles ("boxes") to denote $E(k)$ groups [or $e(k)$ algebras] and circles to denote $O(k)$ groups [or $o(k)$ algebras]. As an example, we give all subgroup chains for $E(n)$, $1\leqslant n\leqslant 4$ in Fig. 5. Maximality requires that as we go from one level to a higher one, we obey the following rules

(1) From a rectangle representing $e(n)$, we can go to two rectangles [see Eq. (3.3)], representing $e(n_1)\oplus e(n_2)$, with $n_1+n_2=n$, $n_1\geqslant n_2\geqslant 1$, or to a circle [see Eq. (3.5)], representing $o(n)$ (the same n as in the rectangle).

(2) From a circle representing $o(n)$ we can go to two circles, representing $o(n_1)\oplus o(n_2)$, $n_1+n_2=n$, $n_1\geqslant n_2\geqslant 2$, or to one circle, representing $o(n-1)$, $n\geqslant 3$.

Now let us consider subgroup type coordinates on the Euclidean space E_n and introduce diagrams to represent them. We shall call them "cluster diagrams" and they will consist of individual trees of the $O(k)$ type with a tree "trunk" added, or isolated "trunks," or of clusters of trees with trunks and isolated trunks. The E_n cluster diagrams are simpler than the $E(n)$ subgroup diagrams, since $E(k)$ subgroups that do not contribute new invariant operators will be omitted.

All clusters for E_n, $1\leqslant n\leqslant 4$, are also shown in Fig. 5. An isolated trunk corresponds to a Cartesian coordinate. A trunk with further branches above it corresponds to a radial coordinate r satisfying $0\leqslant r<\infty$. The tree above the trunk is treated exactly as in the case of polyspheric coordinates on S_n spheres.

As an example let us consider the diagrams in Fig. 3(d); the coordinates in E_7 are

$$\begin{aligned}
&x_1=z, && x_4=r_2\cos\theta_5\cos\theta_6\cos\theta_7, \\
&x_2=r_1\cos\theta_4, && x_5=r_2\cos\theta_5\cos\theta_6\sin\theta_7, \\
&x_3=r_1\sin\theta_4, && x_6=r_2\cos\theta_5\sin\theta_6, \quad x_7=r_2\sin\theta_5.
\end{aligned} \tag{3.6}$$

The prescriptions for writing the complete sets of commuting operators, eigenvalues and eigenfunctions are now quite simple.

To each tree trunk we associate an M-dimensional Laplace operator, where M is the number of end points (Cartesian coordinates) above the trunk. We also associate a number $k\in\mathbf{R}>0$ with each trunk. The corresponding radial eigenfunction [normalized to the delta function: $\delta(k'-k)$] is

$$\begin{gathered}
\Psi_{kl}(r)=\sqrt{\frac{k}{r^{M-2}}}J_{l+(M-2)/2}(kr),\quad M\geqslant 2, \\
\Psi_k(z)=\frac{e^{ikz}}{\sqrt{2\pi}},\quad M=1.
\end{gathered} \tag{3.7}$$

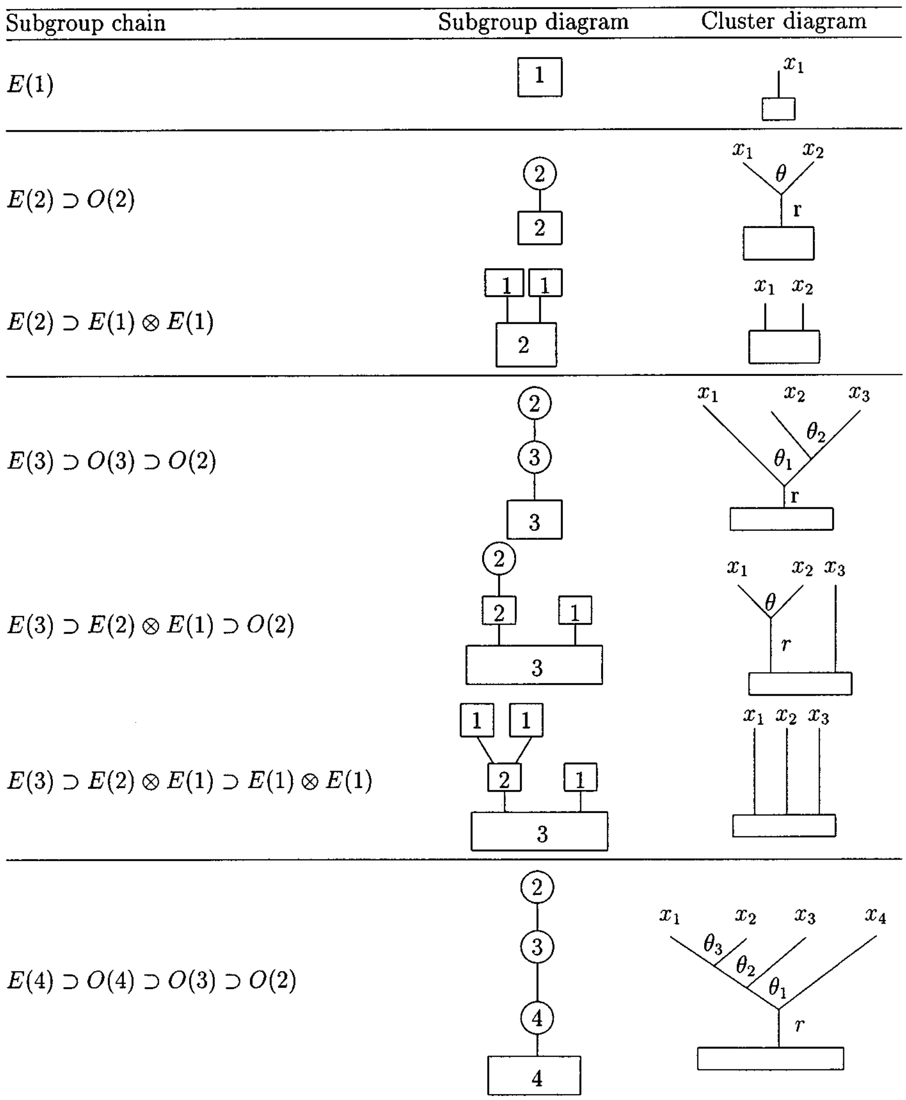

FIG. 5. Subgroup chains for $E(n)$ and cluster diagrams for E_n.

The angular part of the eigenfunctions is written following the rules for S_n spheres, as are the invariant operators and their eigenvalues.

For the example of Figs. 3(c), 3(d), the invariant operators are

$$Y_1=p_1^2,\quad Y_2=p_2^2+p_3^2,\quad Y_3=p_4^2+p_5^2+p_6^2+p_7^2,\quad Y_4=L_{23}^2,$$
$$Y_5=L_{45}^2,\quad Y_6=L_{45}^2+L_{56}^2+L_{46}^2,\quad Y_7=\sum_{4\leqslant i<k\leqslant 7} L_{ik}^2. \tag{3.8}$$

We note that the Laplace operator on E_7 does not figure explicitly; it is equal to

$$\Delta=\sum_{i=1}^{7} p_i^2=Y_1+Y_2+Y_3. \tag{3.9}$$

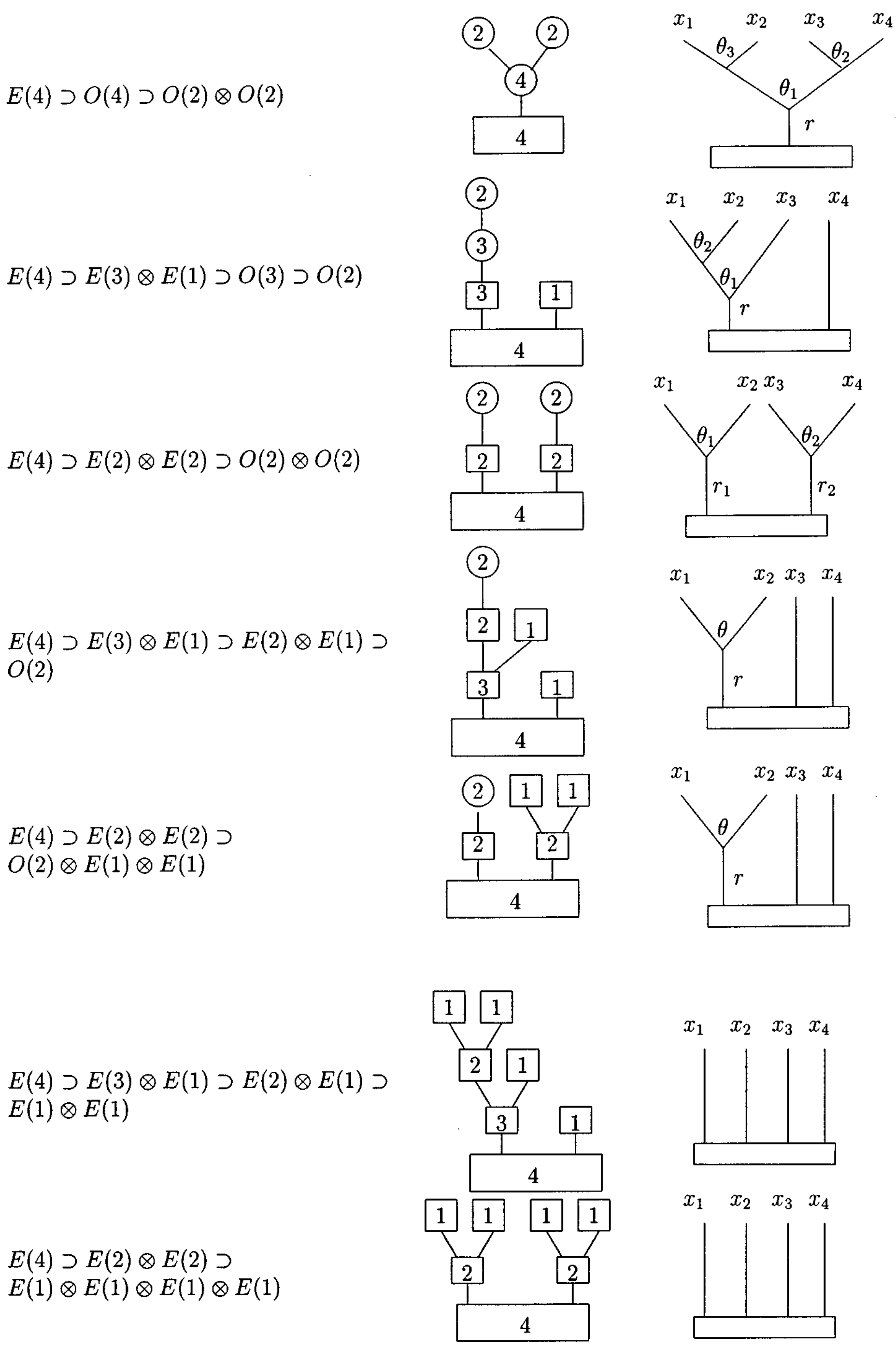

FIG. 5. (*Continued.*)

IV. CONTRACTIONS OF THE LIE ALGEBRA AND CASIMIR OPERATOR

Let us consider the n-dimensional sphere S_n:

$$u_0^2+\sum_{\nu=1}^{n} u_\nu^2=\sum_{\mu,\nu=0}^{n} g_{\mu\nu}u_\mu u_\nu=R^2, \quad R^2>0, \tag{4.1}$$

where u_μ are Cartesian coordinates in the Euclidean ambient space E_{n+1} and the metric tensor in this case has the form $g_{\mu\nu}=\mathrm{diag}(1,1,\ldots,1)$. The isometry group is $O(n+1)$. We choose a standard basis $L_{\mu,\nu}$ for the Lie algebra $o(n+1)$ as in Eq. (2.6).

The Laplace–Beltrami operator on S_n is

$$\Delta_{LB}=\frac{1}{R^2}\sum_{0\leqslant\mu<\nu\leqslant n} L^2_{\mu\nu}. \tag{4.2}$$

We shall use R^{-1} as the contraction parameter. To realize the contraction explicitly, let us introduce Beltrami coordinates on the sphere S_n, putting

$$y_i=R\frac{u_i}{u_0}=u_i\left(1-\frac{1}{R^2}\sum_{k=1}^{n} u_k^2\right)^{-1/2}, \quad i=1,2,3,\ldots,n. \tag{4.3}$$

The $O(n+1)$ generators then can be expressed as

$$\frac{L_{0i}}{R}\equiv\pi_i=p_i+\frac{y_i}{R^2}\sum_{k=1}^{n}(y_k p_k), \tag{4.4}$$

$$L_{ik}\equiv y_i p_k-y_k p_i=y_i\pi_k-y_k\pi_i; \quad i,k=1,2,\ldots,n, \tag{4.5}$$

where $p_i=\partial/\partial y_i$. The commutation relations now are

$$[L_{ik},L_{mn}]=\delta_{km}L_{in}+\delta_{in}L_{km}-\delta_{im}L_{kn}-\delta_{kn}L_{im}, \tag{4.6}$$

$$[\pi_i,L_{kj}]=\delta_{ik}\pi_j-\delta_{ij}\pi_k, \quad [\pi_i,\pi_k]=\frac{L_{ik}}{R^2}, \tag{4.7}$$

so that for $R\to\infty$ the $o(n+1)$ algebra contracts to the Euclidean $e(n)$ one. The Beltrami coordinates y_i (4.3) contract to Cartesian coordinates on E_n, and we have

$$y_i\to x_i, \quad \pi_i\to p_i=\frac{\partial}{\partial x_i}, \tag{4.8}$$

so that the rotation generators L_{0i} go into the translations p_i.

The $o(n+1)$ Laplace–Beltrami operator (2.1) contracts to the $e(n)$ one:

$$\Delta_{\mathrm{LB}}=\sum_{i=1}^{n}\pi_i^2+\sum_{i,k=1}^{n}\frac{L_{ik}^2}{2R^2}\to\Delta=p_1^2+p_2^2+\cdots+p_n^2. \tag{4.9}$$

V. CONTRACTION AND COORDINATE SYSTEMS. THE GRAPHICAL METHOD

A. General formulation

We have seen that all subgroup type coordinates on a sphere S_n can be characterized by tree diagrams. Similarly, there is a one-to-one correspondence between subgroup type coordinates in a Euclidean space E_n and the cluster diagrams of Sec. IV.

We shall now introduce a *graphical method* for connecting the subgroup type coordinate systems on S_n and E_n and give the rules relating the coordinates, invariant operators, eigenvalues and basis functions. The relations are asymptotic ones for the radius of the sphere satisfying $R\to\infty$ and one, or more, of the angles θ_i satisfying $\theta_i\to 0$.

A general S_n tree diagram can be represented by Fig. 6(a). One principal branch of the tree goes from the ground to the point representing the coordinate u_0. The branches growing from this one can lead directly to a coordinate u_i, or they can branch further and lead to sets of coordinates,

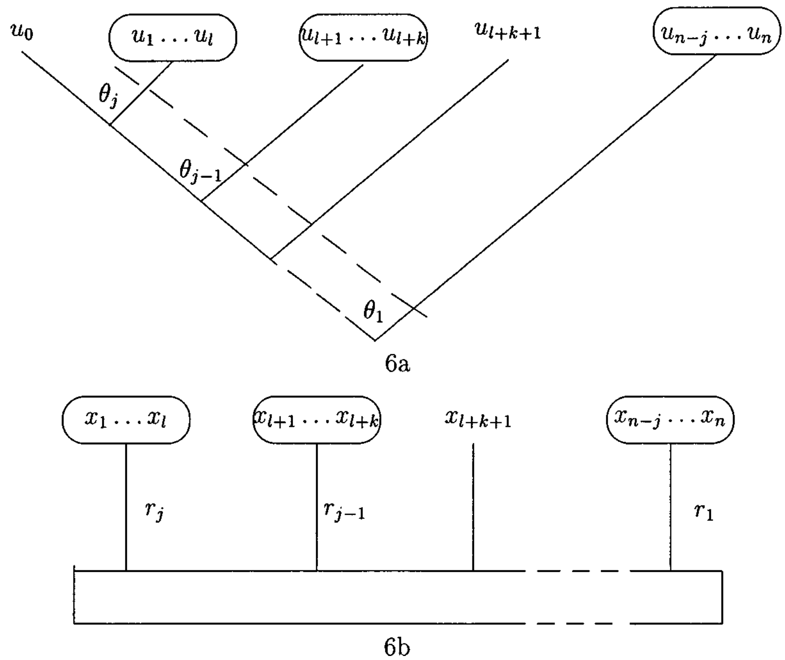

FIG. 6. Contractions of tree diagrams into cluster ones for $S_n \to E_n$.

e.g., $\{u_{l+1}, u_{l+2}, \dots, u_{l+k}\}$. Graphically the contraction $R \to \infty$ corresponds to the fact that we cut off the ground to u_0 branch by the dashed line in Fig. 6(a). The dashed line then becomes the ground for the corresponding cluster E_n diagram of Fig. 6(b) and the ambient space coordinates $(u_0, u_1, \dots, u_n)$ for S_n are replaced by the Cartesian coordinates $(x_1, x_2, \dots, x_n)$. The angles $\theta_1, \theta_2, \dots, \theta_j$ that lead to branches cut-off by the dotted line satisfy $\theta_i \to 0$ in the contraction and are replaced by radial coordinates r_i, or Cartesian coordinates x_m (if the surviving branch leads directly to a single coordinate on S_n and E_n). We have

$$R \to \infty, \quad \theta_i \to 0, \quad R \tan \theta_i \sim R \sin \theta_i \sim R \theta_i \to r_i. \tag{5.1}$$

The individual trees in an E_n cluster correspond to $O(k)$ subgroups of $O(n)$ that survive the contraction.

All contractions of coordinate systems for S_1, S_2, and S_3 are illustrated in Fig. 7. Let us run through the individual cases.

B. Contractions from S_1 to E_1

In the case of a one-dimensional sphere, i.e., a circle, we have only one diagram, namely No. 1 of Fig. 7. In the original ambient space we have polar coordinates

$$u_0 = R \cos \theta, \quad u_1 = R \sin \theta, \tag{5.2}$$

with $0 \leq \theta < 2\pi$. The Beltrami coordinate satisfies

$$y_1 = R \tan \theta \to x, \tag{5.3}$$

where x is a Cartesian coordinate on E_1.

C. Contractions from S_2 to E_2

In the case of the two-dimensional sphere S_2 we have two tree configurations and two types of coordinate contractions to consider, namely, No. 2 and No. 3 of Fig. 7.

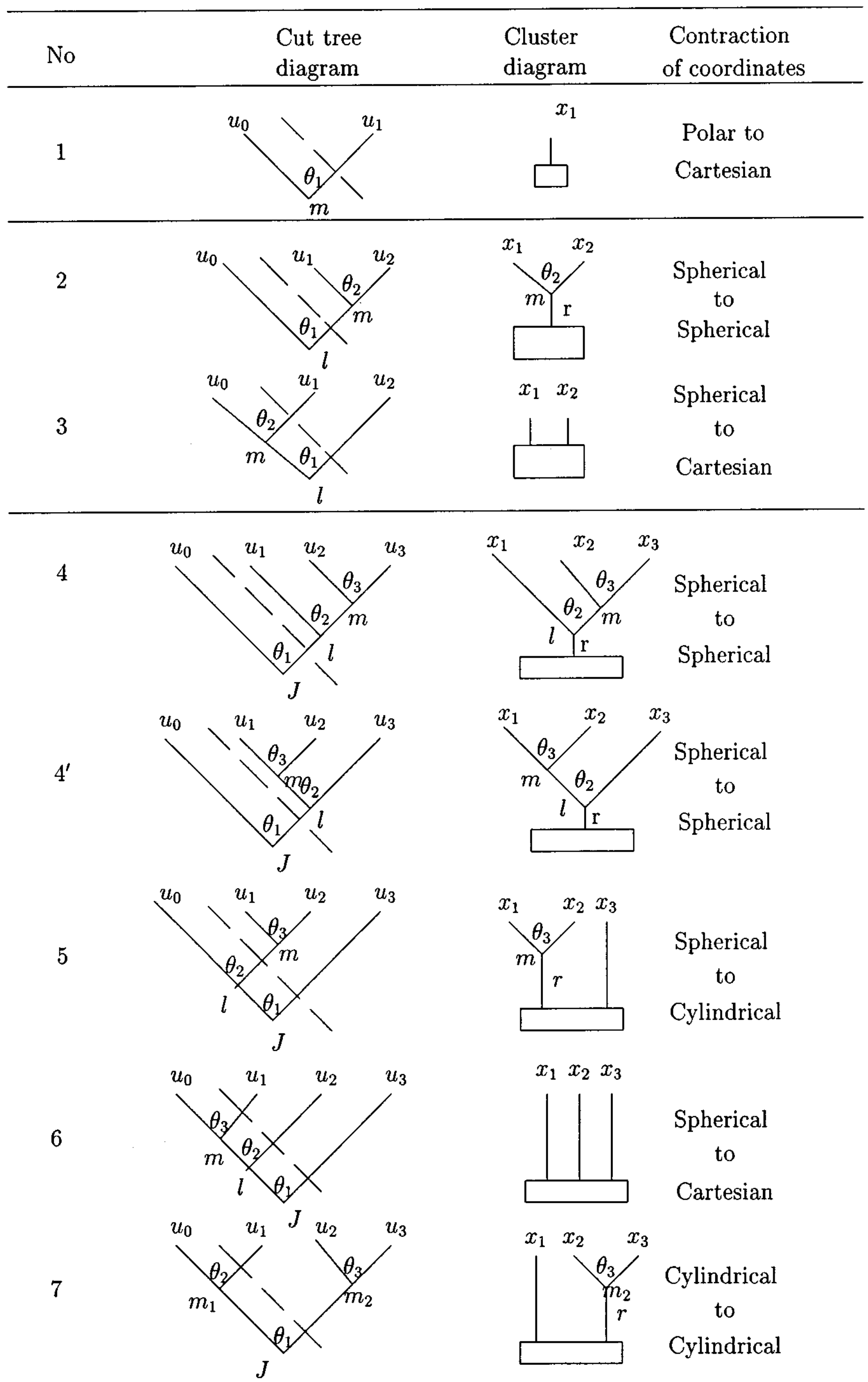

FIG. 7. Contractions of tree diagrams on S_n into cluster ones on E_n for $1 \leqslant n \leqslant 4$.

For *diagram* No. 2 we have

$$u_0 = R \cos\theta_1, \quad u_1 = R \sin\theta_1 \cos\theta_2, \quad u_2 = R \sin\theta_1 \sin\theta_2, \tag{5.4}$$

where $0 \leqslant \theta_1 < \pi$, $0 \leqslant \theta_2 < 2\pi$. Introducing Beltrami coordinates and taking the appropriate limits $R \to \infty$, $\theta_1 \sim r/R$, we have

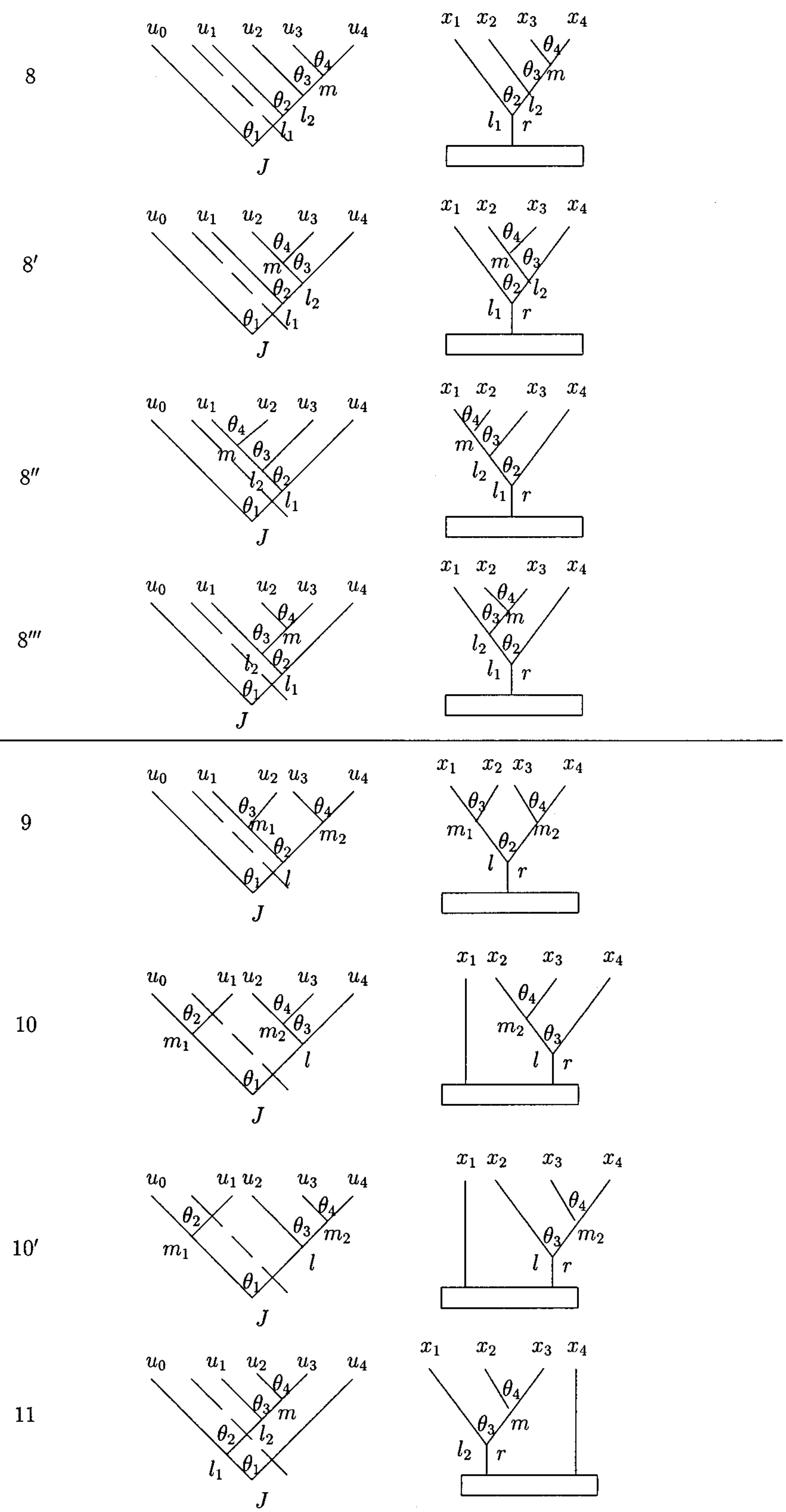

FIG. 7. (*Continued.*)

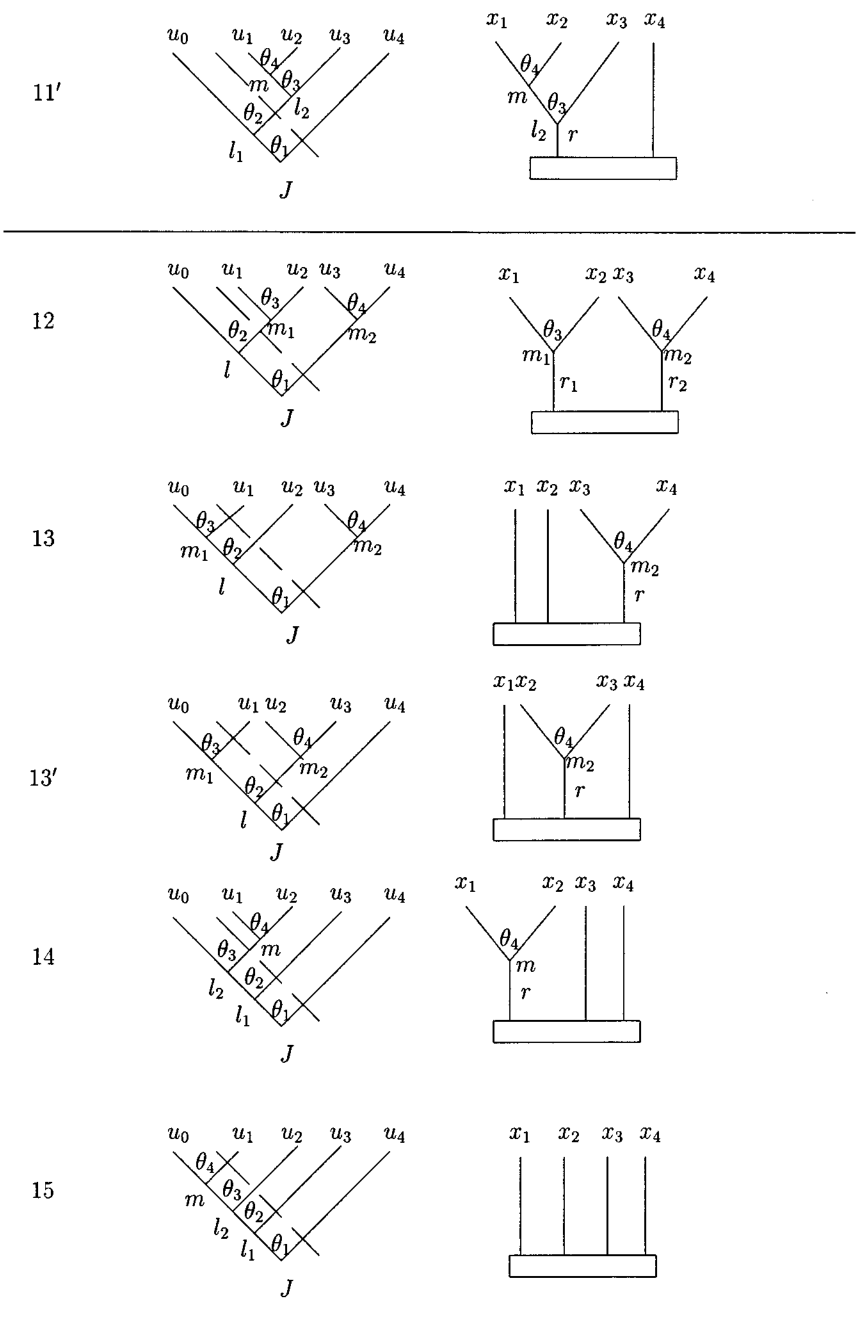

FIG. 7. (*Continued.*)

$$y_1 = R \tan\theta_1 \cos\theta_2 \rightarrow x_1 = r\cos\theta_2, \quad y_2 = R\tan\theta_1 \sin\theta_2 \rightarrow x_2 = r\sin\theta_2. \tag{5.5}$$

The subgroup chain $O(3)\supset O(2)$ contracts to the Euclidean one: $E(2)\supset O(2)$; the $O(2)$ invariant and its eigenvalues m survive the contraction $L_{12}^2 \rightarrow L_{12}^2$, $m\rightarrow m$.

For *diagram* No. 3 in Fig. 7 we have

$$u_0 = R\cos\theta_1\cos\theta_2, \quad u_1 = R\cos\theta_1 \sin\theta_2, \quad u_2 = R\sin\theta_1, \tag{5.6}$$

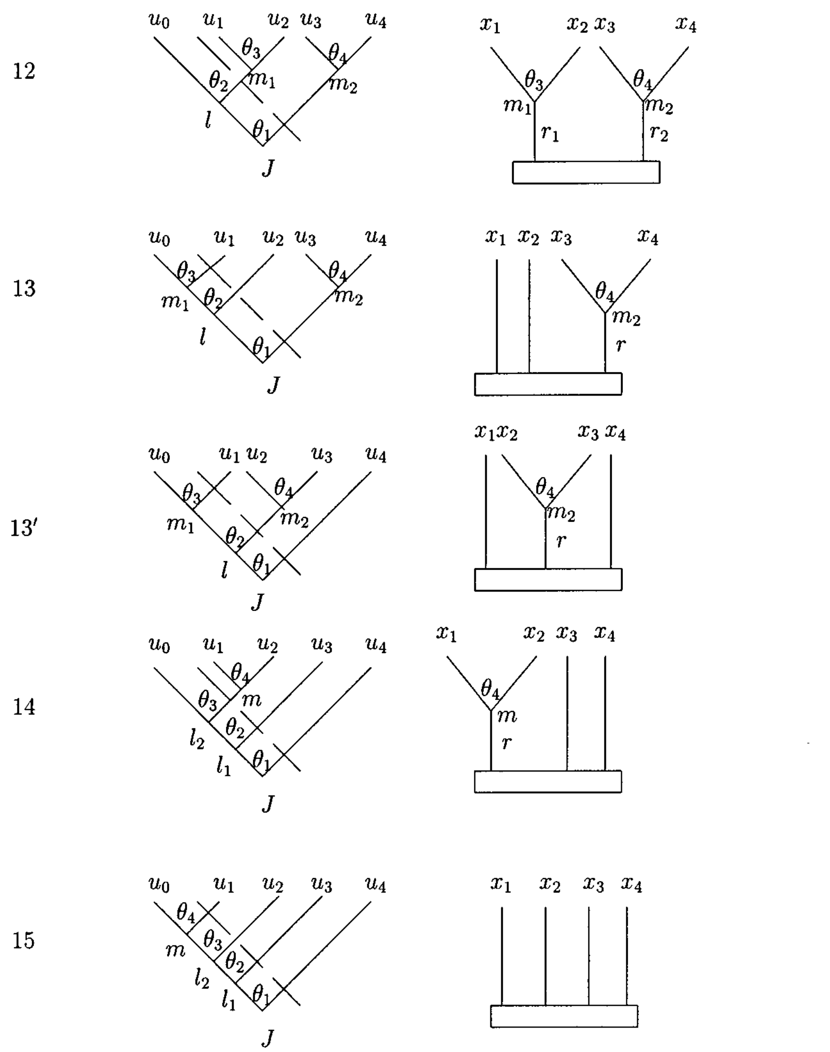

FIG. 7. (*Continued.*)

and the Beltrami coordinates satisfy ($R \to \infty$, $\theta_1 \sim x_1/R$, $\theta_2 \sim x_2/R$)

$$y_1 = R \tan \theta_2 \to x_1, \quad y_2 = R \frac{\tan \theta_1}{\cos \theta_2} \to x_2. \tag{5.7}$$

The subgroup chain $O(3) \supset O(2)$ contracts to $E(2) \supset E(1) \otimes E(1)$ and the $O(2)$ subgroup invariant undergoes a contraction,

$$\frac{Y_1}{R^2} = \frac{L_{01}^2}{R^2} = \pi_1^2 \to p_1^2. \tag{5.8}$$

D. Contractions from S_3 to E_3

Five types of $O(4)$ tree diagrams exist, but only four of them give different contractions.

The *diagrams* No. 4 and 4′ on Fig. 7 correspond to spherical coordinates on S_3 going into spherical coordinates on E_3. For No. 4 the polyspherical coordinates are

$$u_0 = R\cos\theta_1,\quad u_1 = R\sin\theta_1\cos\theta_2,$$
$$u_2 = R\sin\theta_1\sin\theta_2\cos\theta_3,\quad u_3 = R\sin\theta_1\sin\theta_2\sin\theta_3. \tag{5.9}$$

The Beltrami coordinates satisfy $(R\to\infty,\ \theta_1 \sim r/R)$

$$y_1 = R\tan\theta_1\cos\theta_2 \to x_1 = r\cos\theta_2,$$
$$y_2 = R\tan\theta_1\sin\theta_2\cos\theta_3 \to x_2 = r\sin\theta_2\cos\theta_3, \tag{5.10}$$
$$y_3 = R\tan\theta_1\sin\theta_2\sin\theta_3 \to x_3 = r\sin\theta_2\sin\theta_3.$$

We have $O(4)\supset O(3)\supset O(2)\to E(3)\supset O(3)\supset O(2)$ so that the $O(3)\supset O(2)$ subgroups and their invariants survive:

$$Y_1 = \mathbf{L}^2 = L_{12}^2 + L_{13}^2 + L_{23}^2 \to \mathbf{L}^2,\quad Y_2 = L_{23}^2 \to L_{23}^2. \tag{5.11}$$

The situation for diagram No. 4′ is quite analogous.

The case No. 5 in Fig. 7 corresponds to spherical coordinates contracting to cylindrical ones. We have

$$u_0 = R\cos\theta_1\cos\theta_2,\quad u_1 = R\cos\theta_1\sin\theta_2\cos\theta_3,$$
$$u_2 = R\cos\theta_1\sin\theta_2\sin\theta_3,\quad u_3 = R\sin\theta_1. \tag{5.12}$$

For Beltrami coordinates $(R\to\infty,\ \theta_2\sim r/R,\ \theta_1\sim x_3/R)$ we obtain

$$y_1 = R\tan\theta_2\cos\theta_3 \to x_1 = r\cos\theta_3,$$
$$y_2 = R\tan\theta_2\sin\theta_3 \to x_2 = r\sin\theta_3, \tag{5.13}$$
$$y_3 = R\frac{\tan\theta_1}{\cos\theta_2} \to x_3 = z.$$

The subgroup chain contraction is $O(4)\supset O(3)\supset O(2)\to E(3)\supset E(2)\otimes E(1)\supset O(2)$ and the subgroup invariants contract as

$$\frac{Y_1}{R^2} = \frac{1}{R^2}(L_{01}^2 + L_{02}^2 + L_{12}^2) = \pi_1^2 + \pi_2^2 + \frac{L_{12}^2}{R^2} \to p_1^2 + p_2^2,\quad Y_2 = L_{12}^2 \to L_{12}^2. \tag{5.14}$$

The diagram No. 6 in Fig. 7 corresponds to the contraction of spherical coordinates to Cartesian ones. We have

$$u_0 = R\cos\theta_1\cos\theta_2\cos\theta_3,\quad u_1 = R\cos\theta_1\cos\theta_2\sin\theta_3,$$
$$u_2 = R\cos\theta_1\sin\theta_2,\quad u_3 = R\sin\theta_1. \tag{5.15}$$

For Beltrami coordinates after the contraction $R\to\infty,\ \theta_3\sim x_1/R,\ \theta_2\sim x_2/R,\ \theta_1\sim x_3/R$, we have

$$y_1 = R\tan\theta_3 \to x_1,\quad y_2 = R\frac{\tan\theta_2}{\cos\theta_3} \to x_2,\quad y_3 = R\frac{\tan\theta_1}{\cos\theta_2\cos\theta_3} \to x_3. \tag{5.16}$$

The subgroup chain undergoes the contraction $O(4)\supset O(3)\supset O(2)\to E(3)\supset E(1)\otimes E(1)\otimes E(1)$ and the subgroup invariants satisfy

$$\frac{Y_1}{R^2}=\frac{1}{R^2}(L_{01}^2+L_{02}^2+L_{12}^2)=\pi_1^2+\pi_2^2=\frac{L_{12}^2}{R^2}\rightarrow p_1^2+p_2^2,\quad \frac{Y_2}{R^2}=\frac{L_{01}^2}{R^2}=\pi_1^2\rightarrow p_1^2. \tag{5.17}$$

Finally the diagram No. 7 of Fig. 7 corresponds to polyspherical (or cylindrical) coordinates on S_3 contracting to cylindrical ones on E_3. We have

$$\begin{aligned} u_0&=R\cos\theta_1\cos\theta_2,\quad u_1=R\cos\theta_1\sin\theta_2,\\ u_2&=R\sin\theta_1\cos\theta_3,\quad u_3=R\sin\theta_1\sin\theta_3. \end{aligned} \tag{5.18}$$

For the Beltrami coordinates after the contraction $R\rightarrow\infty$, $\theta_2\sim x_1/R$, $\theta_1\sim r/R$, we obtain

$$\begin{aligned} y_1&=R\tan\theta_2\rightarrow x_1,\\ y_2&=R\tan\theta_1\frac{\cos\theta_3}{\cos\theta_2}\rightarrow x_2=r\cos\theta_3,\\ y_3&=R\tan\theta_1\frac{\sin\theta_3}{\cos\theta_2}\rightarrow x_3=r\sin\theta_3. \end{aligned} \tag{5.19}$$

The subgroup chain satisfies $O(4)\supset O(2)\oplus O(2)\rightarrow E(3)\supset E(2)\oplus E(1)\supset O(2)$ so that for the subgroup invariants we have

$$\frac{Y_1}{R^2}=\frac{L_{01}^2}{R^2}=\pi_1^2\rightarrow p_1^2,\quad Y_2=L_{23}^2\rightarrow L_{23}^2. \tag{5.20}$$

VI. CONTRACTIONS OF BASIS FUNCTIONS

A. Contractions of functions corresponding to elementary cells

When we cut off the branches of a tree as in Fig. 6, the cutting line intersects an elementary cell (see Fig. 4) at each branch. Each elementary $O(n+1)$ cell then goes into an elementary trunk for $E(n)$, as indicated by the lower row of diagrams in Fig. 4.

Let us now discuss the four cases in Fig. 4. The limiting procedure is always the same, namely,

$$\theta_j\sim\frac{r_j}{R},\quad l_j\sim kR,\quad R\rightarrow\infty,\quad j=\mathrm{a,b,b',c}, \tag{6.1}$$

where r_j is the radius of the sphere that survives the contraction, i.e., corresponds to the circle on the right hand side of the $O(n+1)$ cell and on top of the $E(n)$ trunk. Thus, for $j=\mathrm{a}$ and $j=\mathrm{b'}$ we have $r_j=x$, a Cartesian coordinate. Similarly, we have $l_\alpha=m\in\mathbf{Z}$ and also $l_\beta=m\in\mathbf{Z}$.

Let us now run through the individual cells in Fig. 4.

1. Cell 1a to 2a

Using Eqs. (2.15) and (6.1) we have $(R\rightarrow\infty, m\sim kR, \theta\sim x/R)$

$$\lim_{R\rightarrow\infty}\frac{1}{\sqrt{2\pi}}e^{im\theta_{\mathrm{a}}}=\frac{1}{\sqrt{2\pi}}e^{ikx}. \tag{6.2}$$

2. Cell 1b to 2b

The contribution to the separated $O(n+1)$ basis function is given in Eq. (2.16). Using the formula for Jacobi polynomials in terms of the hypergeometric functions,[34] we have

$$N_{l-l_\beta}^{l_\beta+S_\beta/2,l_\beta+S_\beta/2}(\sin\theta_{\rm b})^{l_\beta}P_{l-l_\beta}^{(l_\beta+S_\beta/2,l_\beta+S_\beta/2)}(\cos\theta_{\rm b})$$
$$=\sqrt{\frac{(2l+S_\beta+1)(l+l_\beta+S_\beta)!}{2(l-l_\beta)!}}\cdot\frac{(\sin\theta_{\rm b})^{l_\beta}}{2^{l_\beta+S_\beta/2}\Gamma(l_\beta+S_\beta/2+1)}\,{}_2F_1\left(-l+l_\beta,l+l_\beta+S_\beta+1;l_\beta+\frac{S_\beta}{2}+1;\sin^2\frac{\theta_{\rm b}}{2}\right). \quad (6.3)$$

Now, using the asymptotic formulas for the hypergeometric and Γ functions ($l\sim kR,\ \theta_{\rm b}\sim r/R$),

$$\lim_{R\to\infty}{}_2F_1\left(-l+l_\beta,l+l_\beta+S_\beta+1;l_\beta+\frac{S_\beta}{2}+1;\sin^2\frac{\theta_{\rm b}}{2}\right)={}_0F_1\left(l_\beta+\frac{S_\beta}{2}+1;-\frac{k^2r^2}{4}\right), \quad (6.4)$$

$$\lim_{|z|\to\infty}\frac{\Gamma(z+\alpha)}{\Gamma(z+\beta)}=z^{\alpha-\beta}\left(1+\frac{1}{2z}(\alpha-\beta)(\alpha+\beta-1)+O(z^{-2})\right), \quad (6.5)$$

and the formula for the Bessel function,

$$J_\nu(z)=\left(\frac{z}{2}\right)^\nu\frac{1}{\Gamma(\nu+1)}\,{}_0F_1\left(\nu+1;-\frac{z^2}{4}\right), \quad (6.6)$$

we obtain

$$\lim_{R\to\infty}\frac{1}{\sqrt{R^{S_\beta+1}}}N_{l-l_\beta}^{l_\beta+S_\beta/2,l_\beta+S_\beta/2}(\sin\theta_{\rm b})^{l_\beta}P_{l-l_\beta}^{(l_\beta+S_\beta/2,l_\beta+S_\beta/2)}(\cos\theta_{\rm b})=\sqrt{\frac{k}{r^{S_\beta}}}J_{l_\beta+S_\beta/2}(kr). \quad (6.7)$$

3. Cell 1b′ to 2b′

The contribution of this cell to the $O(n+1)$ separated basis function is given in Eq. (2.17). In order to take the contraction limit (6.1) we express the Jacobi polynomials in terms of hypergeometric functions:[34]

$$P_n^{(\alpha,\alpha)}(x)=\frac{2^{2\alpha}}{\sqrt{\pi}}\frac{\Gamma(n+\alpha+1)}{\Gamma(n+2\alpha+1)}\times\begin{cases}(-1)^{n/2}\dfrac{\Gamma([n+1]/2+\alpha)}{\Gamma(n/2+1)}\,{}_2F_1\left(-\dfrac{n}{2},\dfrac{n+1}{2}+\alpha;\dfrac{1}{2};x^2\right), & n\ \text{even},\\ (-1)^{(n-1)/2}\dfrac{\Gamma(n/2+\alpha+1)}{\Gamma([n+1]/2)}2x\,{}_2F_1\left(-\dfrac{n-1}{2},\dfrac{n+2}{2}+\alpha;\dfrac{3}{2};x^2\right), & n\ \text{odd}.\end{cases} \quad (6.8)$$

In the limit $R\to\infty$ and $\theta_{\rm b'}\sim x_n/R, l\sim kR, l_\alpha\sim pR$, we have

$$\lim_{R\to\infty}(-1)^{(l-l_\alpha)/2}N_{l-l_\alpha}^{l_\alpha+S_\alpha/2,l_\alpha+S_\alpha/2}(\cos\theta_{\rm b'})^{l_\alpha}P_{l-l_\alpha}^{(l_\alpha+S_\alpha/2,l_\alpha+S_\alpha/2)}(\sin\theta_{\rm b'})$$
$$=\sqrt{\frac{2k}{\pi k_n}}\times\begin{cases}{}_0F_1\left(\dfrac{1}{2};\dfrac{-k_n^2x_n^2}{4}\right),\\ -i(k_nx_n)\,{}_0F_1\left(\dfrac{3}{2};\dfrac{-k_n^2x_n^2}{4}\right),\end{cases} \quad (6.9)$$

where $k^2 = p^2 + k_n^2$. The ${}_0F_1(x)$ hypergeometric functions in this case are expressible in terms of $\sin k_n x_n$ and $\cos k_n x_n$ functions,[34]

$$ {}_0F_1\left(\frac{1}{2}; \frac{-k_n^2 x_n^2}{4}\right) = \cos k_n x_n , \quad {}_0F_1\left(\frac{3}{2}; \frac{-k_n^2 x_n^2}{4}\right) = \sin k_n x_n , \tag{6.10} $$

and we finally have

$$ \lim_{R\to\infty} (-1)^{(l-l_\alpha)/2} N_{l-l_\alpha}^{l_\alpha + S_\alpha/2, l_\alpha + S_\alpha/2} (\cos\theta_{b'})^{l_\alpha} P_{l-l_\alpha}^{(l_\alpha + S_\alpha/2, l_\alpha + S_\alpha/2)} (\sin\theta_{b'}) = \sqrt{\frac{2k}{\pi k_n}} \begin{Bmatrix} \cos(k_n x_n) \\ -i\sin(k_n x_n) \end{Bmatrix}. \tag{6.11} $$

4. Cell 1c to 2c

The relevant basis function is given in Eq. (2.18). To take the limit (6.1), $l \sim kR$, $l_\alpha \sim k_\alpha R$ and $\theta \sim r/R$, we use the equation expressing Jacobi polynomials in terms of hypergeometric functions, and take the limit leading to Bessel functions:

$$ \begin{aligned} &\lim_{R\to\infty} \frac{\Gamma([l-l_\alpha-l_\beta]/2+1)}{\Gamma([l-l_\alpha+l_\beta+S_\beta]/2+1)} P_{(l-l_\alpha-l_\beta)/2}^{(l_\beta+S_\beta/2, l_\alpha+S_\alpha/2)}(\cos 2\theta_c) \\ &\quad = \lim_{R\to\infty} \frac{1}{\Gamma(l_\beta+S_\beta/2+1)} {}_2F_1\left(-\frac{l-l_\alpha-l_\beta}{2}, \frac{l+l_\alpha+l_\beta+S_\alpha+S_\beta}{2}+1; l_\beta+\frac{S_\beta}{2}+1; \sin^2\theta_c \right) \\ &\quad = \frac{1}{\Gamma(l_\beta+S_\beta/2+1)} {}_0F_1\left(l_\beta+\frac{S_\beta}{2}+1; -\frac{k_\beta^2 r^2}{4}\right) = \left(\frac{2}{k_\beta r}\right)^{l_\beta+S_\beta/2} J_{l_\beta+S_\beta/2}(k_\beta r), \end{aligned} \tag{6.12} $$

where $k_\alpha^2 + k_\beta^2 = k^2$. The final result is

$$ \begin{aligned} &\lim_{R\to\infty} \frac{2^{(l_\alpha+S_\alpha/2+l_\beta+S_\beta/2)/2+1}}{\sqrt{R^{S_\beta+1}}} N_{(l-l_\alpha-l_\beta)/2}^{l_\beta+S_\beta/2, l_\alpha+S_\alpha/2} (\sin\theta_c)^{l_\beta} (\cos\theta_c)^{l_\alpha} P_{(l-l_\alpha-l_\beta)/2}^{(l_\beta+S_\beta/2, l_\alpha+S_\alpha/2)}(\cos 2\theta_c) \\ &\quad = \sqrt{\frac{2k}{r^{S_\beta}}} J_{l_\beta+S_\beta/2}(k_\beta r). \end{aligned} \tag{6.13} $$

These contractions for basis functions of the elementary cells (1a,...,1c) determine the general contractions for hyperspherical functions corresponding to any tree for the sphere S_n.

B. Examples

The contraction formulas for basis functions of $O(3)$ were given in Ref. 1. Here we apply the general rules to give all different S_3 and S_4 contraction diagrams in Fig. 7.

1. The S_3 sphere

(1) Polyspherical to spherical coordinates [see Figs. 7(4)–7(4′)] ($R\to\infty$, $J\sim kR$),

$$ \lim_{R\to\infty} \frac{1}{R} \Psi_{Jlm}(\theta_1, \theta_2, \theta_3) = \sqrt{\frac{k}{r}} J_{l+1/2}(kr) Y_{lm}(\theta_2, \theta_3), \tag{6.14} $$

where $Y_{lm}(\theta_2, \theta_3)$ is a spherical function on S_2.

(2) Polyspherical to cylindrical coordinates [see Fig. 7(5)] ($R\to\infty$, $J\sim kR$, $l\sim k_3/R$),

$$ \lim_{R\to\infty} \frac{(-1)^{(J-l)/2}}{\sqrt{R}} \Psi_{Jlm}(\theta_1, \theta_2, \theta_3) = \sqrt{\frac{kp}{\pi k_3}} J_{|m|}(pr) \frac{e^{im\theta_3}}{\sqrt{2\pi}} \begin{Bmatrix} \cos k_3 z \\ -i\sin k_3 z \end{Bmatrix}, \tag{6.15} $$

where $k^2 = k_3^2 + p^2$.

(3) Polyspherical to Cartesian coordinates [see Fig. 7(6)] ($R \to \infty$, $J \sim k_1 R$, $l \sim k_2/R$, $m \sim k_3 R$)

$$\lim_{R\to\infty} (-1)^{(J-m)/2} \Psi_{Jlm}(\theta_1,\theta_2,\theta_3) = \sqrt{\frac{2}{\pi k_1 k_3}} \frac{e^{ik_1 x}}{\pi} \left\{ \begin{matrix} \cos k_2 y \cos k_3 z \\ -i \sin k_2 y \cos k_3 z \\ -i \cos k_2 y \sin k_3 z \\ -\sin k_2 y \sin k_3 z \end{matrix} \right\}, \tag{6.16}$$

where $k^2 = k_1^2 + k_2^2 + k_3^2$.

(4) Polyspherical (cylindrical) to cylindrical coordinates [see Fig. 7(7)] ($R \to \infty$, $J \sim kR$, $m \sim k_3 R$)

$$\lim_{R\to\infty} \frac{1}{\sqrt{R}} \Psi_{Jm_1m_2}(\theta_1,\theta_2,\theta_3) = \sqrt{\frac{k}{\pi}} J_{|m_2|}(pr) e^{ik_3 z} \frac{e^{im_2\theta_3}}{\sqrt{2\pi}}, \tag{6.17}$$

where $k^2 = k_3^2 + p^2$.

2. The S_4 sphere

(1) Polyspherical to polyspherical coordinates [see Figs. 7(8)–7(8‴)] ($R \to \infty$, $J \sim kR$),

$$\lim_{R\to\infty} \frac{1}{\sqrt{R^3}} \Psi_{Jl_1l_2m}(\theta_1,\theta_2,\theta_3,\theta_4) = \frac{\sqrt{k}}{r} J_{l_1+1}(kr) \Psi_{l_1l_2m}(\theta_2,\theta_3,\theta_4), \tag{6.18}$$

where $\Psi_{l_1l_2m}(\theta_2,\theta_3,\theta_4)$ is a hyperspherical function on S_3.

(2) Polyspherical to cylindrical coordinates [see Fig. 7(9)] ($R \to \infty$, $J \sim kR$),

$$\lim_{R\to\infty} \frac{1}{\sqrt{R^3}} \Psi_{Jlm_1m_2}(\theta_1,\theta_2,\theta_3,\theta_4) = \frac{\sqrt{k}}{r} J_{l+1}(kr) \Psi_{lm_1m_2}(\theta_2,\theta_3,\theta_4), \tag{6.19}$$

where $\Psi_{lm_1m_2}(\theta_2,\theta_3,\theta_4)$ is a hyperspherical function on S_3.

(3) Polyspherical to four-dimensional cylindrical coordinates in Fig. 7(10) [see also Fig. 7(10′)] ($R \to \infty$, $J \sim kR$, $m_1 \sim k_1 R$),

$$\lim_{R\to\infty} \frac{1}{R} \Psi_{Jlm_1m_2}(\theta_1,\theta_2,\theta_3,\theta_4) = \sqrt{\frac{k}{\pi r}} e^{ik_1x_1} J_{l+1/2}(pr) Y_{lm_2}(\pi/2-\theta_3,\theta_4), \tag{6.20}$$

where $Y_{lm_2}(\pi/2-\theta_3,\theta_4)$ is a spherical function on S_2 and $k^2 = k_1^2 + p^2$.

(4) Polyspherical to four-dimensional cylindrical coordinates in Fig. 7(11) [see also Fig. 7(11′)] ($R \to \infty$, $J \sim kR$, $l \sim k_4 R$),

$$\lim_{R\to\infty} \frac{(-1)^{(J-l_1)/2}}{R} \Psi_{Jl_1l_2m}(\theta_1,\theta_2,\theta_3,\theta_4) = \sqrt{\frac{2pk}{\pi k_4 r}} J_{l_2+1/2}(pr) Y_{l_2m}(\theta_3,\theta_4) \left\{ \begin{matrix} \cos(k_4x_4) \\ -i \sin(k_4x_4) \end{matrix} \right\}, \tag{6.21}$$

where $Y_{l_2m}(\theta_3,\theta_4)$ is a spherical function on S_2 and $k^2 = k_4^2 + p^2$.

(5) Polyspherical to bipolar coordinates in Fig. 7(12) ($R \to \infty$, $J \sim kR$, $l \sim k_1 R$),

$$\lim_{R\to\infty} \frac{1}{R} \Psi_{Jlm_1m_2}(\theta_1,\theta_2,\theta_3,\theta_4) = \frac{\sqrt{2kk_1}}{2\pi} J_{m_1}(k_1r_1) J_{m_2}(k_2r_2) e^{im_1\theta_3 + im_2\theta_4}, \tag{6.22}$$

where $k_1^2+k_2^2=k^2$.

(6) Polyspherical to double cylindrical coordinates in Fig. 7(13) [see also Fig. 7(13′)] ($R \to\infty$, $J\sim kR$, $l\sim k_1R$, $m\sim k_2R$),

$$\lim_{R\to\infty}\frac{(-1)^{(l-m_1)/2}}{\sqrt{R}}\Psi_{Jlm_1m_2}(\theta_1,\theta_2,\theta_3,\theta_4)=\sqrt{\frac{k\sqrt{k_1^2+k_2^2}}{\pi^3k_2}}e^{ik_1x_1}\begin{Bmatrix}\cos(k_2x_2)\\ -i\sin(k_2x_2)\end{Bmatrix}J_{|m_2|}(k_3r)\frac{e^{im_2\theta_4}}{\sqrt{2\pi}}, \tag{6.23}$$

where $k_1^2+k_2^2+k_3^2=k^2$.

(7) Polyspherical to double cylindrical coordinates in Fig. 7(14) ($R\to\infty$, $J\sim kR$, $l_1\sim k_3R$, $l_2\sim k_1R$),

$$\lim_{R\to\infty}\frac{(-1)^{(J-l_2)/2}}{\sqrt{R}}\Psi_{Jl_1l_2m}(\theta_1,\theta_2,\theta_3,\theta_4)$$
$$=\sqrt{\frac{2k_1k\sqrt{k_1^2+k_2^2}}{\pi^3k_2k_3}}J_{|m|}(k_1r)e^{im_2\theta_4}\begin{Bmatrix}\cos k_3x_3\cos k_4x_4\\ -i\sin k_3x_3\cos k_4x_4\\ -i\cos k_3x_3\sin k_4x_4\\ -\sin k_3x_3\sin k_4x_4\end{Bmatrix}, \tag{6.24}$$

where $k_1^2+k_2^2+k_3^2=k^2$.

(8) Polyspherical to Cartesian coordinates in Fig. 7(15) ($R\to\infty$, $J\sim kR$, $m\sim k_1R$, $l_2\sim k_2R$, $l_1\sim k_3R$),

$$\lim_{R\to\infty}(-1)^{(J-m)/2}\Psi_{Jl_1l_2m}(\theta_1,\theta_2,\theta_3,\theta_4)$$
$$=\sqrt{\frac{8k\sqrt{k_1^2+k_2^2}\sqrt{k_1^2+k_2^2+k_3^2}}{\pi^4k_2k_3k_4}}e^{ik_1x_1}$$
$$\times\begin{Bmatrix}\cos k_2x_2\cos k_3x_3\cos k_4x_4; & -i\sin k_2x_2\cos k_3x_3\cos k_4x_4;\\ -i\cos k_2x_2\sin k_3x_3\cos k_4x_4; & -\sin k_2x_2\sin k_3x_3\cos k_4x_4;\\ -i\cos k_2x_2\cos k_3x_3\sin k_4x_4; & -\sin k_2x_2\cos k_3x_3\sin k_4x_4;\\ -\cos k_2x_2\sin k_3x_3\sin k_4x_4; & -i\sin k_2x_2\sin k_3x_3\sin k_4x_4;\end{Bmatrix}, \tag{6.25}$$

where $k_1^2+k_2^2+k_3^2+k_4^2=k^2$.

As a final example, let us consider the contraction $O(8)\to E(7)$ for the coordinate systems of Fig. 3. The contraction of the $O(8)$ basis to the $E(7)$ one in this case is ($R\to\infty$, $l_1\sim kR$, $l_2 \sim k_2R$, $l_3\sim k_1R$):

$$\lim_{R\to\infty}\frac{1}{R^2}\Psi_{l_1l_2l_3l_4l_5l_6l_7}(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5,\theta_6,\theta_7)$$
$$=\frac{\sqrt{k_2k_3}}{2\pi r_2}J_{|l_4|}(k_2r_1)J_{l_5+1}(k_3r_2)e^{ik_1x_1}e^{il_4\theta_4}Y_{l_5l_6l_7}(\theta_5,\theta_6,\theta_7). \tag{6.26}$$

VII. CONCLUSION

In our previous paper[1] we studied contractions of all (i.e., both) coordinate systems on S_2 to all (four) coordinate systems on E_2. Here we have presented all possible contractions of subgroup type coordinate systems on S_n to subgroup type ones on E_n for n arbitrary. Moreover, we have developed a graphical formalism illustrating these contractions.

Contractions of ellipsoidal and parabaloidal coordinate systems will relate more ''exotic'' special functions amongst each other. For instance, Lamé polynomials and their generalizations will go into Mathieu functions, parabolic cylinder functions, spheroidal functions, etc. Work in this direction is in progress.

ACKNOWLEDGMENTS

We thank Professor E. G. Kalnins, Professor W. Miller, Jr., and Professor V. M. Ter-Antonyan for their interest and helpful discussions. We also thank Ye. M. Hakobyan for her help in the preparation of the diagrams. Two of the authors (G.P. and P.W.) thank each others' institutions for hospitality during mutual visits that made the research leading to the present article possible.

The research of P.W. was partially supported by research grants from NSERC of Canada and FCAR du Québec, that of G.P. by the RFBI grant N98-01-00330.

[1] A. A. Izmest'ev, G. S. Pogosyan, A. N. Sissakian, and P. Winternitz, ''Contractions of Lie algebras and separation of variables,'' J. Phys. A **29**, 5940–5962 (1996).
[2] A. A. Izmest'ev, G. S. Pogosyan, A. N. Sissakian, and P. Winternitz, ''Contractions of Lie algebras and separation of variables. Two-dimensional hyperboloid,'' Int. J. Mod. Phys. A **12**, 53–61 (1997).
[3] P. Winternitz and I. Fris, ''Invariant expansions of relativistic amplitudes and subgroups of the proper Lorentz group,'' Sov. J. Nucl. Phys. **1**, 636–643 (1965); Yad. Fiz. **1**, 889–901 (1965).
[4] P. Winternitz, I. Lukac, and Ya. A. Smorodinskii, ''Quantum numbers in the little groups of the Poincaré group,'' Sov. J. Nucl. Phys. **7**, 139–145 (1968); Yad. Fiz. **7**, 192–201 (1968).
[5] J. Patera and P. Winternitz, ''A new basis for representations of the rotation group. Lamé and Heun polynomials,'' J. Math. Phys. **14**, 1130–1139 (1973).
[6] W. Miller, Jr., J. Patera, and P. Winternitz, ''Subgroups of Lie groups and separation of variables,'' J. Math. Phys. **22**, 251–260 (1991).
[7] W. Miller, Jr., *Symmetry and Separation of Variables* (Addison-Wesley, Reading, MA, 1997).
[8] E. G. Kalnins, *Separation of Variables for Riemannian Spaces of Constant Curvature* (Longman, Burnt Mill, 1986).
[9] N. Ya. Vilenkin, ''Polyspherical and orispherical functions,'' Mat. Sb. **68**, 432–443 (1965).
[10] N. Ya. Vilenkin, *Special Functions and the Theory of Group Representations* (American Mathematical Society, Providence, RI, 1968).
[11] N. Ya. Vilenkin, G. I. Kuznetsov, and Ya. A. Smorodinskii, ''Eigenfunctions of the Laplace operator realizing representations of the groups $U(2)$, $SU(2)$, $SO(3)$, $U(3)$ and $SU(3)$ and the symbolic method,'' Sov. J. Nucl. Phys. **2**, 645–655 (1965); Yad. Fiz. **2**, 906–917 (1965).
[12] M. S. Kildyushov, ''Hyperspherical functions of the ''tree'' type in the n-particle problem,'' Sov. J. Nucl. Phys. **15**, 113–123 (1972); Yad. Fiz. **15**, 197–208 (1972).
[13] G. I. Kuznetsov, S. S. Moskalyuk, Yu. F. Smirnov, and V. P. Shelest, *A Graphical Theory of Representations of Orthogonal and Unitary Groups and its Physical Applications* (Naukova Dumka, Kiev, 1992) (in Russian).
[14] E. Inönü and E. P. Wigner, ''On the contraction of groups and their representations,'' Proc. Natl. Acad. Sci. USA **39**, 510–524 (1953).
[15] E. J. Saletan, ''Contraction of Lie groups,'' J. Math. Phys. **2**, 1–21 (1961).
[16] R. Gilmore, *Lie Groups, Lie Algebras and Some of their Applications* (Wiley, New York, 1974).
[17] R. V. Moody and J. Patera, ''Discrete and continuous graded contractions of representations of Lie algebras,'' J. Phys. A **24**, 2227–2258 (1991).
[18] X. Leng and J. Patera, ''Graded contractions of representations of $sl(n,C)$ with respect to the maximal parabolic subalgebras,'' J. Phys. A **27**, 1233–1250 (1987).
[19] M. Couture, J. Patera, R. T. Sharp, and P. Winternitz, ''Graded contractions of $sl(3,C)$,'' J. Math. Phys. **32**, 2310–2318 (1991).
[20] M. Ait Abdelmalek, X. Leng, J. Patera, and P. Winternitz, ''Grading refinements in the contractions of Lie algebras and their invariants,'' J. Phys. A **29**, 7519–7543 (1996).
[21] P. Winternitz, ''Successive refinements of gradings and graded contractions of $sl(3,C)$,'' Int. J. Mod. Phys. A **12**, 109–115 (1997).
[22] J. Patera, G. Pogosyan, and P. Winternitz, ''Graded contractions of the Lie algebra $e(2,1)$,'' CRM-2484, 1997.
[23] E. Weimar-Woods, ''Contractions of Lie algebras: generalized Inönü–Wigner contractions versus graded contractions,'' J. Math. Phys. **36**, 4519–4548 (1995).
[24] L. P. Eisenhart, ''Separable systems of Stäckel,'' Ann. Math. **35**, 284–305 (1934).
[25] D. A. Suprunenko and R. I. Tyshkevich, *Commutative Matrices* (Academic, New York, 1968).
[26] J. Patera, P. Winternitz, and H. Zassenhaus, ''Maximal Abelian subalgebras of real and complex symplectic Lie algebras,'' J. Math. Phys. **24**, 1973–1985 (1983).
[27] M. A. del Olmo, M. A. Rodriguez, P. Winternitz, and H. Zassenhaus, ''Maximal Abelian subalgebras of pseudounitary Lie algebras,'' Linear Algebr. Appl. **135**, 79–151 (1990).

[28] V. Hussin, P. Winternitz, and H. Zassenhaus, ''Maximal Abelian subalgebras of complex orthogonal Lie algebras,'' Linear Algebr. Appl. **141**, 183–220 (1990).
[29] V. Hussin, P. Winternitz, and H. Zassenhaus, ''Maximal Abelian subalgebras of pseudoorthogonal Lie algebras,'' Linear Algebr. Appl. **173**, 125–163 (1992).
[30] E. G. Kalnins and P. Winternitz, ''Maximal Abelian subalgebras of complex Euclidean Lie algebras,'' Can. J. Phys. **72**, 389–404 (1994).
[31] Z. Thomova and P. Winternitz, ''Maximal Abelian subgroups of the isometry and conformal groups of Euclidean and Minkowski spaces,'' J. Phys. A **31**, 1831–1859 (1998).
[32] M. N. Olevskii, ''Orthogonal systems in spaces of constant curvature in which the equation $\Delta_2 u + \lambda u = 0$ allows a complete separation of variables,'' Mat. Sb. **27**, 379–426 (1950) (in Russian).
[33] E. Kalnins, W. Miller, Jr., and P. Winternitz, ''The group $O(4)$, separation of variables and the hydrogen atom,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **30**, 630–664 (1976).
[34] G. Bateman and A. Erdélyi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953).

Highest weight irreducible representations of the Lie superalgebra $gl(1|\infty)$

T. D. Palev[a),b)] and N. I. Stoilova[a),c)]
Abdus Salam International Centre for Theoretical Physics, 34100 Trieste, Italy

(Received 27 May 1998; accepted for publication 27 August 1998)

Two classes of irreducible highest weight modules of the general linear Lie superalgebra $gl(1|\infty)$ are constructed. Within each module a basis is introduced and the transformation relations of the basis under the action of the algebra generators are written down. © *1999 American Institute of Physics.* [S0022-2488(99)04402-3]

I. INTRODUCTION

We construct two classes of irreducible representations of the infinite-dimensional general linear Lie superalgebra $gl(1|\infty)$. Both of them are classes of highest weight representations, corresponding to two different orderings of the basis in the Cartan subalgebra. Related to this it is convenient to define $gl(1|\infty)$ in two different, but certainly equivalent ways. We denote them as $gl_0(1|\infty)$ and $gl(\infty|1|\infty)$ (see the end of Sec. I for the notation that follows).

Definition 1: The Lie superalgebra $gl_0(1|\infty)$ *is a complex linear space with a basis* $\{e_{ij}\}_{i,j\in\mathbf{N}}$. *The* $\mathbf{Z}_2$*-grading on* $gl_0(1|\infty)$ *is defined from the requirement that* e_{1j}, e_{j1}, $j=2,3,...$ *are odd generators, whereas all other generators are even. The multiplication* ($\equiv$*the supercommutator*) $[\![,]\!]$ *on* $gl_0(1|\infty)$ *is a linear extension of the relations*:

$$[\![e_{ij},e_{kl}]\!]=\delta_{jk}e_{il}-(-1)^{\deg(e_{ij})\deg(e_{kl})}\delta_{il}e_{kj}, \quad i,j,k,l\in\mathbf{N}. \tag{1}$$

As a basis in the Cartan subalgebra $\mathcal{H}_0$ we choose $\{e_{ii}\}_{i\in\mathbf{N}}$ with a natural order between the generators: $e_{ii}<e_{jj}$, if $i<j$. Then $\mathcal{E}_0^+=\{e_{ij}\}_{i<j\in\mathbf{N}}$ (respectively, $\mathcal{E}_0^-=\{e_{ij}\}_{i>j\in\mathbf{N}}$) are the positive (respectively, the negative) root vectors and $\{e_{i,i+1}\}_{i\in\mathbf{N}}$ are the simple root vectors.

Definition 2: The Lie superalgebra $gl(\infty|1|\infty)$ *is a complex linear space with a basis* $\{E_{ij}\}_{i,j\in\mathbf{Z}}$. *The* $\mathbf{Z}_2$*-grading on* $gl(\infty|1|\infty)$ *is defined from the requirement that* E_{0j}, E_{j0}, $0\neq j\in\mathbf{Z}$ *are odd generators, whereas all other generators are even. The supercommutator on* $gl(\infty|1|\infty)$ *is a linear extension of the relations*:

$$[\![E_{ij},E_{kl}]\!]=\delta_{jk}E_{il}-(-1)^{\deg(E_{ij})\deg(E_{kl})}\delta_{il}E_{kj}, \quad i,j,k,l\in\mathbf{Z}. \tag{2}$$

As a basis in the Cartan subalgebra $\mathcal{H}$ we choose $\{E_{ii}\}_{i\in\mathbf{Z}}$ with a natural order between the generators: $E_{ii}<E_{jj}$, if $i<j$. $\mathcal{E}^+=\{E_{ij}\}_{i<j\in\mathbf{Z}}$ (respectively, $\mathcal{E}^-=\{E_{ij}\}_{i>j\in\mathbf{Z}}$) are the positive (respectively, the negative) root vectors in $gl(\infty|1|\infty)$ and $\{E_{i,i+1}\}_{i\in\mathbf{Z}}$ are the simple root vectors.

Both algebras are isomorphic. In order to see this let g: $\mathbf{Z}\rightarrow\mathbf{N}$ be a bijective map, defined as

$$g(z)=2|z|+\theta(z)\in\mathbf{N}, \quad \forall z\in\mathbf{Z}. \tag{3}$$

Then it is easy to verify that the map φ, which is a linear extension of the relations

$$\varphi(E_{ij})=e_{g(i),g(j)}, \quad i,j\in\mathbf{Z}, \tag{4}$$

[a)]Permanent address: Institute for Nuclear Research and Nuclear Energy, 1784 Sofia, Bulgaria.
[b)]Electronic mail: tpalev@inrne.bas.bg
[c)]Electronic mail: stoilova@inrne.bas.bg

is an isomorphism of $gl(\infty|1|\infty)$ on $gl_0(1|\infty)$. Therefore both $gl_0(1|\infty)$ and $gl(\infty|1|\infty)$ are two different realizations of one and the same algebra, namely $gl(1|\infty)$. Note that φ is a map of $\mathcal{H}$ onto $\mathcal{H}_0$; it is not however a map of $\mathcal{E}^+$ into $\mathcal{E}_0^+$. For instance take $E_{-1,0}\in\mathcal{E}^+$. Then $\varphi(E_{-1,0}) = e_{21}\in\mathcal{E}_0^-$. Hence a highest weight representation of $gl(\infty|1|\infty)$ may not be a highest weight representation of $gl_0(1|\infty)$.

The reasons for studying representations of this particular superalgebra, namely $gl(1|\infty)$, stem from physical considerations. Our motivation originates from an attempt to introduce new quantum statistics both in quantum mechanics[1,2] (in this case the superalgebras are finite-dimensional) and in quantum field theory (QFT).[3,4] In order to see where the connection to the statistics comes from, we recall shortly the origin of the Lie superstatistics.

The starting point is based on the observation that any n pairs $b_1^\pm,\ldots,b_n^\pm$ of Bose creation and annihilation operators (CAOs), namely (below and throughout $[x,y]=xy-yx$, $\{x,y\}=xy+yx$)

$$[b_i^-,b_j^+]=\delta_{ij},\quad [b_i^-,b_j^-]=[b_i^+,b_j^+]=0, \tag{5}$$

considered as odd elements, generate a representation, the Bose representation ρ, of the Lie superalgebra $osp(1|2n)\equiv B(0|n)$.[5] Denote by $B_1^\pm,\ldots,B_n^\pm$ those generators of $B(0|n)$, which in the Bose representation coincide with the Bose operators, $\rho(B_i^\pm)=b_i^\pm$. Similarly as the Chevalley generators do, the operators $B_1^\pm,\ldots,B_n^\pm$ and the relations they satisfy, namely

$$[\{B_i^\xi,B_j^\eta\},B_k^\epsilon]=(\epsilon-\xi)\delta_{ik}B_j^\eta+(\epsilon-\eta)\delta_{jk}B_i^\xi,\quad \xi,\eta,\epsilon=\pm \ \text{ or } \ \pm 1, \tag{6}$$

define uniquely the LS $B(0|n)$.[5] The operators $B_i^\pm$ are odd root vectors of $B(0|n)$, whereas $\{B_j^+,B_j^-\}$ belong to the Cartan subalgebra. The operators (6) are known in quantum field theory: these are the *para*-Bose operators, generalizing the statistics of the tensor fields.[6] The important conclusion is that the representation theory of n pairs of *para*-Bose (pB) operators is equivalent to the representation theory of the Lie superalgebra $B(0|n)$. Certainly in QFT the algebra is $B(0|\infty)$, it is infinite-dimensional.

The identification of the *para*-Bose statistics with a well-known algebraic structure provides a natural background for further generalizations. In QFT the commutation relations between the CAOs are determined from the translation invariance of the field under consideration.[7] In momentum space the translation invariance of a scalar (or tensor) field $\Psi(x)$ is expressed as a commutator between the energy–momentum P^m, $m=0,1,2,3$ and the CAOs $a_i^\pm$ of $\Psi(x)$:

$$[P^m,a_i^\pm]=\pm k_i^m a_i^\pm, \tag{7}$$

where the index i replaces all (continuous and discrete) indices of the field and

$$P^m=\frac{1}{2}\sum_j k_j^m\{a_j^+,a_j^-\}. \tag{8}$$

To quantize the field means, loosely speaking, to find solutions of Eqs. (7) and (8), where the unknowns are the CAOs $a_i^\pm$. The Bose operators (5) and their generalization, the pB operators (6), certainly satisfy (7). By no means however they do not exhaust the set of the possible solutions.

The first possibility for finding new solutions and hence for further generalization of the statistics stems from the observation that the commutation relations between the Cartan elements and the root vectors, in particular Eq. (7), remain unaltered upon q deformations. The deformations of the parastatistics along this line were studied in Refs. 8–11 and more generally in Ref. 12.

Another opportunity, closely related to the present paper, is based on the observation that $B(0|n)$ belongs to the class B superalgebras in the classification of Kac.[13] Therefore it is natural to try to satisfy the quantization equations (7) and (8) with CAOs, generating superalgebras from the classes A, C, and D or generating other superalgebras from the class B. In Refs. 3 and 4 it was shown that this is possible indeed. For charged tensor fields the main quantization condition (7) can be satisfied with CAOs, which generate the LS $gl(\infty|1|\infty)$, namely a LS from the class *A*. Up to now, however, this new statistics, the *A* superstatistics, did not achieve any further development. The reason is that so far the Fock spaces corresponding to the *A* superstatistics were not constructed. Here we come to the relation between the *A* superstatistics and the present investigation. The Fock spaces are representation spaces of $gl(1|\infty)$. In order to study the physical consequences of the *A* superstatistics in QFT one has to develop first the representation theory of $gl(\infty|1|\infty)$ (for charged scalar fields) and of $gl_0(1|\infty)$ (for neutral fields). This is what we do in the present paper. The reason to study only highest weight representations reflects the fact that there should exist a state with a lowest energy, a vacuum, which turns out to be the highest weight vector in the corresponding $gl(1|\infty)$ module.

So far the *A* superstatistics was tested only in finite-dimensional cases, namely in the frame of a (noncanonical) quantum mechanics. We have in mind the Wigner quantum systems, introduced in Refs. 1 and 2, which recently attracted some attention from different points of view.[14–16] These systems possess quite unconventional physical features, properties which cannot be achieved in the frame of the canonical quantum mechanics. The $(n+1)$-particle WQS, based on $sl(1/3n)$,[17] exhibits a quarklike structure: the composite system occupies a small volume around the center of mass and within it the geometry is noncommutative. The underlying statistics is a Haldane exclusion statistics,[18] a subject of considerable interest in condensed matter physics. The osp(3/2) WQS, studied in Ref. 19, leads to a picture where two spinless point particles, curling around each other, produce an orbital (internal angular) momentum 1/2. One can expect that also in QFT the Lie superstatistics could lead to new features.

In the literature one does not find many papers dealing with representations of infinite-dimensional simple Lie superalgebras.[20,21] Implicitly, however, such algebras and their representations were used in theoretical physics since the QFT was created. In the first place we have in mind the ordinary Fock space W_1 of infinitely many pairs of Bose CAOs $\{b_i^\pm\}_{i\in\mathbf{Z}}$. As mentioned above, the Bose operators are (representatives of) the odd generators of $B(0|\infty)$ and their Fock space W_1 is one particular irreducible $B(0|\infty)$ module. The Fock spaces W_p of *para*-Bose operators $\{B_i^\pm\}_{i\in\mathbf{Z}}$, corresponding to order of the parastatistics $p\in\mathbf{N}$,[6] are also irreducible and inequivalent to each other $B(0|\infty)$ modules. The Clifford construction in Ref. 21 is a generalization to the case when both bosons $\{b_i^\pm\}_{i\in\mathbf{Z}}$, considered as odd variables, and fermions $\{f_i^\pm\}_{i\in\mathbf{Z}}$, which are even generators, are involved. The assumption is that the bosons anticommute with the fermions. Then any *n* pairs of Bose CAOs and *m* pairs of Fermi CAOs generate (a representation of) the Lie superalgebra $B(m|n)$.[22] Therefore the Fock representation of $\{b_i^\pm, f_i^\pm\}_{i\in\mathbf{Z}}$ is an irreducible $B(\infty|\infty)$ module. Its restriction to $gl(\infty|\infty)$ leads to a set of irreducible representations of this superalgebra.

In the paper we essentially use results from the representation theory of $gl(1|n)$. The finite-dimensional irreducible modules (fidirmods) of the latter are, one can say, well understood. A character formula for all typical[13] and atypical[23] modules has been constructed. The dimensions of all fidirmods are known.[24,25] A basis, similar to the GZ basis of $gl(n)$, was defined and its transformation under the action of the Chevalley generators was written down.[26,27] The results were even generalized to the quantum algebra $U_q[gl(1|n)]$.[28] This is in contrast to the more general case of $gl(m|n)$ and $U_q[gl(m/n)]$, where most of the above problems are still waiting to be solved although partial results do exist.[29–33]

The irreducible highest weight representations of $gl_0(1|\infty)$, which we consider, are a generalization to the infinite-dimensional case of the finite-dimensional essentially typical representations of $gl(1|n)$ in the Gel'fand–Zetlin basis (GZ basis). In order to see where the possibility for a generalization comes from we recall (Sec. II A) the way the Gel'fand–Zetlin basis was introduced.[31] This basis is, however, inappropriate for a generalization to the case of highest weight $gl(\infty|1|\infty)$ modules. Therefore in Sec. II B we modify it, introducing a new basis, which we call a *C* basis. It is an analog of the *C* basis for gl_∞.[34,35] Section III is devoted to the irreducible $gl(1|\infty)$ modules. In Sec. III A we extend the Gel'fand–Zetlin basis to the infinite-dimensional case and apply it to $gl_0(1|\infty)$. The highest weight irreducible $gl(\infty|1|\infty)$ representations are defined in Sec. III B. They appear as a generalization of the essentially typical representations of $gl(1|n)$ in the C basis. The transformations of the basis under the action of the algebra generators are explicitly written down.

Throughout the paper we use the following notation:

LS, LSs	Lie superalgebra, Lie superalgebras;
CAOs	creation and annihilation operators;
fidirmod(s)	finite-dimensional irreducible module(s);
GZ basis	Gel'fand–Zetlin basis;
$\mathbf{N}$	all positive integers;
$\mathbf{Z}$	all integers;
$\mathbf{Z}_+$	all non-negative integers;
$\mathbf{Z}_2=\{0,\overline{1}\}$	the ring of all integers modulo 2;
$\mathbf{C}$	all complex numbers;

$$[p;q]=\{p,p+1,p+2,...,q-1,q\}, \text{ if } q-p\in\mathbf{Z}_+ \text{ and } [p;q]=\varnothing \text{ otherwise}; \tag{9}$$

$$[m]_k=[m_{1k},m_{2k},\dots,m_{kk}], \text{ where } m_{ik}\in\mathbf{C}; \tag{10}$$

$$[M]_{2k+\theta}=[M_{-k,2k+\theta},M_{-k+1,2k+\theta},\dots,M_{k-1+\theta,2k+\theta}], \quad \theta\in\{0,1\}, \quad k\in\mathbf{N}; \tag{11}$$

$$l_{1j}=m_{1j}+1, \quad l_{ij}=-m_{ij}+i-1, \quad i\in[2;j]; \tag{12}$$

$$\begin{aligned} &L_{0,2k+\theta}=M_{0,2k+\theta}, \quad \theta\in\{0,1\},\\ &L_{i,2k+\theta}=-M_{i,2k+\theta}+i+1, \quad \theta\in\{0,1\}, \quad i\in[-k;-1],\\ &L_{j,2k+\theta}=-M_{j,2k+\theta}+j-1, \quad \theta\in\{0,1\}, \quad j\in[1;k-1+\theta]; \end{aligned} \tag{13}$$

$$[m]\equiv[m_1,m_2,\dots,m_k,\dots]=\{m_i|m_i\in\mathbf{C},i\in\mathbf{N}\}; \tag{14}$$

$$[M]\equiv[\dots,M_{-p},\dots,M_{-1},M_0,M_1,\dots,M_q,\dots]=\{M_i|M_i\in\mathbf{C},i\in\mathbf{Z}\}; \tag{15}$$

$$P(j,l)=\begin{cases}1 & \text{for } j\geqslant l\\ -1 & \text{for } j<l\end{cases}; \tag{16}$$

$$Q(j,l)=\begin{cases}1 & \text{for } j>l\\ -1 & \text{for } j\leqslant l\end{cases}; \tag{17}$$

$$\theta(i)=\begin{cases}1 & \text{for } i\geqslant 0\\ 0 & \text{for } i<0\end{cases}. \tag{18}$$

II. FINITE-DIMENSIONAL ESSENTIALLY TYPICAL REPRESENTATIONS OF $gl(1|2n)$

As in the case of $gl(1|\infty)$ it is convenient to use two different notation for the finite-dimensional superalgebras from this class. In the first notation $gl_0(1|N)$ is the same as in Definition 1, but the indices i,j run from 1 to $N+1$.

Then $e_{11},e_{22},\dots,e_{N+1,N+1}$ is a basis in the Cartan subalgebra $\mathcal{H}_0$. Denote by $\epsilon^1,\dots,\epsilon^{N+1}$ the dual basis, $\epsilon^i(e_{jj})=\delta^i_j$. The correspondence root vector $\leftrightarrow$ root reads: $e_{ij}\leftrightarrow\epsilon^i-\epsilon^j$, $i\neq j=1,...,N+1$; $\Delta^0=\{\epsilon^i-\epsilon^j\}_{i\neq j\in[1;N+1]}$ is the root system; $\Delta^0_+=\{\epsilon^i-\epsilon^j\}_{i<j\in[1;N+1]}$ and

$$\pi^0=\{\epsilon^1-\epsilon^2,\epsilon^2-\epsilon^3,\dots,\epsilon^N-\epsilon^{N+1}\} \tag{19}$$

are the standard systems of positive roots and simple roots, respectively. The special linear superalgebra $sl_0(1|N)$ is a subalgebra of $gl_0(1|N)$ spanned by all $gl_0(1|N)$ root vectors and the Cartan elements $e_{11}+e_{ii}$ for all $i\neq 1$.

Similarly, $gl(M|1|N)$ is the same as in Definition 2, but $i,j=-M,-M+1,...,N$ and $M,N \in \mathbf{Z}_+$. In particular $\{E_{ii}\}_{i\in[-M;N]}$ is a basis in the Cartan subalgebra $\mathcal{H}$ with $\{\mathcal{E}^i\}_{i\in[-M;N]}$ its dual. The simple root vectors are $\{E_{i,i+1}\}_{i\in[-M;N-1]}$. Hence

$$\pi=\{\mathcal{E}^{-M}-\mathcal{E}^{-M+1},\mathcal{E}^{-M+1}-\mathcal{E}^{-M+2},\ldots,\mathcal{E}^{-1}-\mathcal{E}^0,\mathcal{E}^0-\mathcal{E}^1,\ldots,\mathcal{E}^{N-1}-\mathcal{E}^N\} \tag{20}$$

is the system of simple roots.

We have written explicitly the systems (19) and (20) in order to underline that they contain different number of odd roots: π^0 has only one, $\epsilon^1-\epsilon^2$, whereas the odd roots in π are $\mathcal{E}^{-1}-\mathcal{E}^0,\mathcal{E}^0-\mathcal{E}^1$. Therefore the systems of the simple roots of $sl_0(1|2n)$ and $sl(n|1|n)$ are different, despite the fact that these algebras are isomorphic. This property demonstrates one of the essential differences between the Lie algebras and the Lie superalgebras. For each simple Lie algebra there exists (up to a transformation from the Weyl group) only one system of simple roots. This is not the case for the basic Lie superalgebras, where several inequivalent simple root systems can be in general defined (for more details see Refs. 36–38). As a result one and the same irreducible $gl(1|2n)$ module can be described with different signatures. We shall have to take this into account in the definition of the C basis.

A. GZ basis (Ref. 31)

Let $V([m]_{N+1})$ be a highest weight finite-dimensional irreducible $gl_0(1|N)$ module (fidirmod) with a highest weight

$$[m]_{N+1}\equiv[m_{1,N+1},m_{2,N+1},\ldots,m_{N+1,N+1}]\equiv\sum_{i=1}^{N+1} m_{i,N+1}\epsilon^i, \tag{21}$$

where

$$m_{j,N+1}\in\mathbf{C},\quad j=1,...,N+1,\quad m_{i,N+1}-m_{i+1,N+1}\in\mathbf{Z}_+,\quad i=2,3,...,N. \tag{22}$$

If x_{N+1} is the highest weight vector in $V([m]_{N+1})$, then $e_{ii}x_{N+1}=m_{i,N+1}x_{N+1}$.

Consider the chain of subalgebras

$$gl_0(1|N)\supset gl_0(1|N-1)\supset gl_0(1|N-2)\supset\cdots\supset gl_0(1|2)\supset gl_0(1|1)\supset gl_0(1|0)\equiv gl_0(1). \tag{23}$$

Then $V([m]_{N+1})$ is said to be essentially typical, if it is completely reducible with respect to any one of the subalgebras in the chain (23). Each essentially typical module $V([m]_{N+1})$ carries a typical representation[13] of the special linear superalgebra $sl_0(1|n)$, but the inverse is in general not true.

Set

$$l_{1,N+1}=m_{1,N+1}+1,\quad l_{i,N+1}=-m_{i,N+1}+i-1,\quad i=2,3,...,N+1. \tag{24}$$

Proposition 1: (Ref. 31) The $gl_0(1|N)$ module $V([m]_{N+1})$ is essentially typical if and only if

$$l_{1,N+1}\notin[l_{2,N+1};l_{N+1,N+1}]. \tag{25}$$

Let $V([m]_{N+1})$ be an essentially typical $gl_0(1|N)$ module and let

$$V([m]_{N+1})\supset V([m]_N)\supset V([m]_{N-1})\supset\cdots\supset V([m]_{k+1})\supset\cdots V([m]_2)\supset V(m_{11}) \tag{26}$$

be a flag of $gl_0(1|k)$ fidirmods $V([m]_{k+1})$, $k=0,1,2,...,N$, where

$$[m]_{k+1}\equiv[m_{1,k+1},m_{2,k+1},\ldots,m_{k+1,k+1}]\equiv\sum_{i=1}^{k+1} m_{i,k+1}\epsilon^i \tag{27}$$

is the signature of $V([m]_{k+1})$. In the ordered basis

$$e_{11}, e_{22}, \ldots, e_{k+1,k+1} \tag{28}$$

of the Cartan subalgebra of $gl_0(1|k)$, $m_{i,k+1}$ is the eigenvalue of e_{ii} on the highest weight vector $x_{k+1} \in V([m]_{k+1})$,

$$e_{ii}x_{k+1} = m_{i,k+1}x_{k+1}, \quad i=1,...,k+1. \tag{29}$$

Since we consider only essentially typical modules and the fidirmods of $gl_0(1)$ are one dimensional, the flag (26) defines a vector $|m\rangle$ in $V([m]_{N+1})$. It turns out this vector is uniquely defined (up to, certainly, a multiplicative constant) by the signatures $[m]_{N+1}$, $[m]_N, \ldots, [m]_2, m_{11}$. Therefore one can set

$$|m\rangle \equiv \begin{bmatrix} [m]_{N+1} \\ [m]_N \\ \vdots \\ [m]_2 \\ m_{11} \end{bmatrix} \equiv \begin{bmatrix} m_{1,N+1} & m_{2,N+1} & \cdots & m_{N,N+1} & m_{N+1,N+1} \\ m_{1,N} & m_{2,N} & \cdots & m_{N,N} & \\ \cdots & \cdots & \cdots & & \\ m_{12} & m_{22} & & & \\ m_{11} & & & & \end{bmatrix}. \tag{30}$$

The vectors (30), corresponding to all possible flags (26), constitute a basis $\Gamma([m]_{N+1})$ in the $gl_0(1|N)$ fidirmod $V([m]_{N+1})$. This is the GZ basis introduced in Ref. 31 [for the more general case of $gl(M/N)$].

Proposition 2: (Ref. 31) The GZ basis $\Gamma([m]_{N+1})$ *in the essentially typical module* $V([m]_{N+1})$ *is given by all tables (30) for which*

(1) the numbers $m_{i,N+1}$, $i=1,2,...,N+1$ *are fixed for all tables and satisfy (22), (24), (25),*

$$(2) \quad m_{1i} - m_{1,i-1} \equiv \theta_{i-1} \in \{0,1\}, \quad i=2,3,...,N+1, \tag{31}$$

$$(3) \quad m_{i,j+1} - m_{ij} \in \mathbf{Z}_+, \quad m_{ij} - m_{i+1,j+1} \in \mathbf{Z}_+, \quad 2 \leqslant i \leqslant j \leqslant N. \tag{32}$$

The transformations of the basis $\Gamma([m]_{N+1})$ *under* $gl_0(1|N)$ *are completely defined from the action of the Chevalley generators*

$$e_{ii}|m\rangle = \left(\sum_{k=1}^{i} m_{ki} - \sum_{k=1}^{i-1} m_{k,i-1} \right) |m\rangle, \quad i=1,2,...,N+1, \tag{33}$$

$$e_{12}|m\rangle = \theta_1 |m\rangle_{(11)}, \quad e_{21}|m\rangle = (1-\theta_1)(l_{12}-l_{22})|m\rangle_{-(1,1)}, \tag{34}$$

$$\begin{aligned} e_{i,i+1}|m\rangle = {} & \theta_i(1-\theta_{i-1})|m\rangle_{(1i)} \\ & + \sum_{j=2}^{i} \left(- \frac{\Pi_{k=2}^{i-1}(l_{k,i-1}-l_{ji}+1)\Pi_{k=2}^{i+1}(l_{k,i+1}-l_{ji})}{\Pi_{k\neq j=2}^{i}(l_{ki}-l_{ji})(l_{ki}-l_{ji}+1)} \right)^{1/2} \\ & \times \frac{(l_{1i}-l_{ji})(l_{1i}-l_{ji}+1)}{(l_{1,i+1}-l_{ji})(l_{1,i-1}-l_{ji}+1)} |m\rangle_{(ji)}, \quad i=2,...,N, \end{aligned} \tag{35}$$

$$\begin{aligned} e_{i+1,i}|m\rangle = {} & \theta_{i-1}(1-\theta_i) \frac{\Pi_{k=2}^{i-1}(l_{1,i+1}-l_{k,i-1}-1)\Pi_{k=2}^{i+1}(l_{1,i+1}-l_{k,i+1})}{\Pi_{k=2}^{i}(l_{1,i+1}-l_{ki}-1)(l_{1,i+1}-l_{ki})} |m\rangle_{-(1,i)} \\ & + \sum_{j=2}^{i} \left(- \frac{\Pi_{k=2}^{i-1}(l_{k,i-1}-l_{ji})\Pi_{k=2}^{i+1}(l_{k,i+1}-l_{ji}-1)}{\Pi_{k\neq j=2}^{i}(l_{ki}-l_{ji}-1)(l_{ki}-l_{ji})} \right)^{1/2} |m\rangle_{-(ji)}, \quad i=2,...,N, \end{aligned} \tag{36}$$

where $l_{1j}=m_{1j}+1$, $l_{ij}=-m_{ij}+i-1$, $i\neq 1$ *and the table* $|m\rangle_{\pm(i,j)}$ *is obtained from the table* $|m\rangle$ *by the replacement* $m_{ij}\to m_{ij}\pm 1$.

If a vector from the right-hand side of (35) or (36) does not belong to the module under consideration, then the corresponding term is zero even if the coefficient in front is undefined; if an equal number of factors in numerator and denominator are simultaneously equal to zero, they should be canceled out.

The $gl_0(1|N)$ highest weight vector x_{N+1} in $V([m]_{N+1})$ is a vector from the GZ basis

$$x_{N+1}=|\hat{m}\rangle \quad \text{for which} \quad m_{ii}=m_{i,i+1}=\cdots=m_{i,N+1}, \quad i=1,2,...,N, \tag{37}$$

i.e.,

$$|\hat{m}\rangle=\left[\begin{array}{lllll} m_{1,N+1} & m_{2,N+1} & \cdots & m_{N,N+1} & m_{N+1,N+1} \\ m_{1,N+1} & m_{2,N+1} & \cdots & m_{N,N+1} & \\ \cdots & \cdots & \cdots & & \\ m_{1,N+1} & m_{2,N+1} & & & \\ m_{1,N+1} & & & & \end{array}\right]. \tag{38}$$

In this case

$$e_{ii}|\hat{m}\rangle=m_{i,N+1}|\hat{m}\rangle, \quad i=1,2,...,N+1, \quad e_{k,k+1}|\hat{m}\rangle=0, \quad k=1,2,...,N. \tag{39}$$

B. *C* basis

Let E_{ij}, $i,j=-n,-n+1,...,n$ be the generators of $gl(n|1|n)$. Define a sequence of subalgebras

$$gl(k|1|k-1+\theta)=\text{lin env}\{E_{ij}|i,j\in[-k;k-1+\theta]\}\forall\,\theta\in\{0,1\}, \quad k\in[1-\theta;n]. \tag{40}$$

As an ordered basis in the Cartan subalgebra of $gl(k|1|k-1+\theta)$ take

$$E_{-k,-k},E_{-k+1,-k+1},\ldots,E_{k-1+\theta,k-1+\theta}. \tag{41}$$

Proposition 3: The map φ, *which is a linear extension of the relations*

$$\varphi(E_{ij})=e_{g(i),g(j)}, \quad i,j=-n,-n+1,...,n, \tag{42}$$

is an isomorphism of $gl(n|1|n)$ *on* $gl_0(1|2n)$. *Its restriction on* $gl(k|1|k-1+\theta)$ *is an isomorphism of* $gl(k|1|k-1+\theta)$ *on* $gl_0(1|2k-1+\theta)$ *for each* $\theta\in\{0,1\}$ *and* $k\in[1-\theta;n]$. *The chain of subalgebras*

$$\begin{aligned} &gl(n|1|n)\supset gl(n|1|n-1)\supset gl(n-1|1|n-1)\supset gl(n-1|1|n-2)\supset\cdots \\ &\quad\supset gl(1|1|1)\supset gl(1|1)\supset gl(1), \end{aligned} \tag{43}$$

($gl(1|1|0)\equiv gl(1|1)$, $gl(0|1|0)\equiv gl(1)$) *is transformed by* φ *into the chain (23)*

$$gl_0(1|2n)\supset gl_0(1|2n-1)\supset gl_0(1|2n-2)\supset\cdots\supset gl_0(1|2)\supset gl_0(1|1)\supset gl_0(1). \tag{44}$$

The proof is straightforward.

The isomorphism φ allows one to turn any $gl_0(1|2k-1+\theta)$ irreducible module $V([m]_{2k+\theta})$ into a $gl(k|1|k-1+\theta)$ module:

$$\varphi(E_{ij})x=e_{g(i),g(j)}x, \quad \forall x\in V([m]_{2k+\theta}). \tag{45}$$

The relevant point for us is that each $V([m]_{2k+\theta})$ can be labeled also with its highest weight with respect to $gl(k|1|k-1+\theta)$. By definition it consists of the eigenvalues of the representatives of the Cartan generators (41), namely

$$\begin{gathered}\varphi(E_{-k,-k}),\quad \varphi(E_{-k+1,-k+1}),\ldots,\quad \varphi(E_{-2,-2}),\quad \varphi(E_{-1,-1}),\\ \varphi(E_{0,0}),\quad \varphi(E_{1,1}),\ldots,\quad \varphi(E_{k-1+\theta,k-1+\theta})\end{gathered} \tag{46}$$

on the $gl(k|1|k-1+\theta)$ highest weight vector $y_{2k+\theta}\in V([m]_{2k+\theta})$. The latter is defined from the requirements

$$\varphi(E_{ij})y_{2k+\theta}=0,\quad i<j=-k,-k+1,...,k-1+\theta, \tag{47}$$

$$\varphi(E_{ii})y_{2k+\theta}=M_{i,2k+\theta}y_{2k+\theta},\quad i=-k,-k+1,...,k-1+\theta. \tag{48}$$

Set

$$[M]_{2k+\theta}\equiv[M_{-k,2k+\theta},M_{-k+1,2k+\theta},\ldots,M_{k-1+\theta,2k+\theta}]. \tag{49}$$

The new signature $[M]_{2k+\theta}$ defines, as mentioned above, uniquely $V([m]_{2k+\theta})$. Hence

$$V([m]_{2k+\theta})=V([M]_{2k+\theta}). \tag{50}$$

Consider now a GZ basis vector $|m\rangle$ corresponding to the flag

$$V([m]_{2n+1})\supset V([m]_{2n})\supset V([m]_{2n-1})\supset\cdots\supset V([m]_{2k+\theta})\supset\cdots V([m]_2)\supset V(m_{11})\leftrightarrow|m\rangle, \tag{51}$$

namely the vector (30) with $N=2n$. In view of (50) the same flag can be written as

$$V([M]_{2n+1})\supset V([M]_{2n})\supset V([M]_{2n-1})\supset\cdots\supset V([M]_{2k+\theta})\supset\cdots\supset V([M]_2)\supset V(M_{11}) \tag{52}$$

and therefore the vector $|m\rangle$ is completely defined by the signatures $[M]_{2n+1},[M]_{2n},\ldots,[M]_2,M_{11}$. Therefore we can write any GZ basis vector (30) also in the form

$$|M\rangle\equiv\left[\begin{array}{cccccccc} M_{-n,2n+1} & M_{-n+1,2n+1} & \cdots & M_{-1,2n+1} & M_{0,2n+1} & M_{1,2n+1} & \cdots & M_{n-1,2n+1} & M_{n,2n+1}\\ M_{-n,2n} & M_{-n+1,2n} & \cdots & M_{-1,2n} & M_{0,2n} & M_{1,2n} & \cdots & M_{n-1,2n} & \\ & M_{-n+1,2n-1} & \cdots & M_{-1,2n-1} & M_{0,2n-1} & M_{1,2n-1} & \cdots & M_{n-1,2n-1} & \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & \\ & & \cdots & \cdots & \cdots & \cdots & \cdots & & \\ & & & M_{-1,3} & M_{03} & M_{13} & & & \\ & & & M_{-1,2} & M_{02} & & & & \\ & & & & M_{01} & & & & \end{array}\right]. \tag{53}$$

Obviously (30) (with $N=2n$) and (53) are two different labelings for one and the same vector $|m\rangle\equiv|M\rangle$. We call the basis, written in the notation (53), a C basis in $V([M]_{2n+1})\equiv V([m]_{2n+1})$ and denote it as $\Gamma([M]_{2n+1})$.

In order to use effectively the basis $\Gamma([M]_{2n+1})$ we need to determine all signatures $[M]_{2k+\theta}$, namely to find the values of the entries in (53). To this end we have to determine as a first step the highest weight vector $y_{2k+\theta}$ within each $gl(k|1|k-1+\theta)$ module $V([m]_{2k+\theta})$ in the chain (51) and subsequently, using (48), to compute its $gl(k|1|k-1+\theta)$ signature $[M]_{2k+\theta}$.

Proposition 4: The $gl(k|1|k-1+\theta)$ highest weight vector $y_{2k+\theta}$ in $V([m]_{2k+\theta})$ [from the chain (51)] is the GZ vector $|m\rangle_{2k+\theta}$, for which

$$m_{1,2r+\tau}+k-r=m_{1,2k+\theta},\quad \forall\tau\in\{0,1\},\quad r\in[1-\tau;k-\tau], \tag{54}$$

$$m_{r-j,2k-2j+\tau}=m_{r,2k+\theta}, \quad \forall r\in[3-\theta;k+1], \quad \tau\in\{0,1\}, \quad j\in[1-\theta;r-2], \tag{55a}$$

$$m_{r-j,2k-2j+\tau}=m_{r,2k+\theta}, \quad \forall r\in[k+2;2k], \quad \tau\in\{0,1\}, \quad j\in[1-\theta;2k-r+\tau]. \tag{55b}$$

Proof: It is easy to verify that the conditions (54) are equivalent to

$$\theta_{2i-1}=1, \quad i\in[1;k], \tag{56a}$$

$$\theta_{2i}=0, \quad i\in[1;k-1+\theta], \tag{56b}$$

whereas the conditions (55) can be replaced by

$$l_{s,2i+1}-l_{s,2i}=0, \quad i\in[1;k-1+\theta], \quad s\in[2;2i], \tag{57a}$$

$$l_{s+1,2i}-l_{s,2i-1}-1=0, \quad i\in[2;k], \quad s\in[2;2i-1]. \tag{57b}$$

We need to show that (47) holds for $y_{2k+\theta}=|m\rangle_{2k+\theta}$. It certainly suffices to verify it only for the $gl(k|1|k-1+\theta)$ simple root vectors, namely to prove that

$$\varphi(E_{-i,-i+1})|m\rangle_{2k+\theta}=0, \quad i\in[1;k], \tag{58}$$

$$\varphi(E_{i,i+1})|m\rangle_{2k+\theta}=0, \quad i\in[0;k-2+\theta]. \tag{59}$$

The validity of the latter follows from the observation that $\varphi(E_{-1,0})=e_{21}$, $\varphi(E_{01})=[e_{12},e_{23}]$, $\varphi(E_{-i,-i+1})=[e_{2i,2i-1},e_{2i-1,2i-2}]$, $i\in[2;k]$, $\varphi(E_{i-1,i})=[e_{2i-1,2i},e_{2i,2i+1}]$, $i\in[2;k-1+\theta]$ and Eqs. (34)–(36). This completes the proof.

We are now ready to determine the $gl(k|1|k-1+\theta)$ signature of $V([m]_{2k+\theta})$ for any $\theta\in\{0,1\}$ and $k\in[1;n]$. Taking into account (54), (55), and (45) and using the transformation relation (33), one obtains

$$\varphi(E_{ii})|m\rangle_{2k+\theta}=e_{2|i|,2|i|}|m\rangle_{2k+\theta}=(m_{i+k+2,2k+\theta}+1)|m\rangle_{2k+\theta}, \quad i\in[-k;-1], \tag{60a}$$

$$\varphi(E_{00})|m\rangle_{2k+\theta}=e_{11}|m\rangle_{2k+\theta}=(m_{1,2k+\theta}-k)|m\rangle_{2k+\theta}, \tag{60b}$$

$$\varphi(E_{ii})|m\rangle_{2k+\theta}=e_{2i+1,2i+1}|m\rangle_{2k+\theta}=m_{i+k+1,2k+\theta}|m\rangle_{2k+\theta}, \quad i\in[1;k-1+\theta]. \tag{60c}$$

Comparing (60) with the definition (48) we obtain the $gl(k|1|k-1+\theta)$ signature $[M]_{2k+\theta}$ of $V([m]_{2k+\theta})$:

$$M_{i,2k+\theta}=m_{i+k+2,2k+\theta}+1, \quad i\in[-k;-1], \tag{61a}$$

$$M_{0,2k+\theta}=m_{1,2k+\theta}-k, \tag{61b}$$

$$M_{i,2k+\theta}=m_{i+k+1,2k+\theta}, \quad i\in[1,k-1+\theta], \tag{61c}$$

$$M_{01}=m_{11}. \tag{61d}$$

We have added the evident relation (61d) for completeness, since it is not contained in (61a)–(61c). The above relations hold for any $\theta\in\{0,1\}$ and $k\in[1;n]$. In particular,

$$M_{i,2n+1}=m_{i+n+2,2n+1}+1, \quad i\in[-n;-1], \tag{62a}$$

$$M_{0,2n+1}=m_{1,2n+1}-n, \tag{62b}$$

$$M_{i,2n+1}=m_{i+n+1,2n+1}, \quad i\in[1,n]. \tag{62c}$$

The $gl(n|1|n)$ highest weight vector $y_{2n+1}\equiv|\hat{M}\rangle$ is the one from (53), for which $M_{i,j}=M_{i,2n+1}$ for any admissible i and j:

$$\left[\begin{array}{ccccccccc} M_{-n,2n+1} & M_{-n+1,2n+1} & \cdots & M_{-1,2n+1} & M_{0,2n+1} & M_{1,2n+1} & \cdots & M_{n-1,2n+1} & M_{n,2n+1} \\ M_{-n,2n+1} & M_{-n+1,2n+1} & \cdots & M_{-1,2n+1} & M_{0,2n+1} & M_{1,2n+1} & \cdots & M_{n-1,2n+1} & \\ & M_{-n+1,2n+1} & \cdots & M_{-1,2n+1} & M_{0,2n+1} & M_{1,2n+1} & \cdots & M_{n-1,2n+1} & \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & \\ & & \cdots & \cdots & \cdots & \cdots & \cdots & & \\ & & & M_{-1,2n+1} & M_{0,2n+1} & M_{1,2n+1} & & & \\ & & & M_{-1,2n+1} & M_{0,2n+1} & & & & \\ & & & & M_{0,2n+1} & & & & \end{array}\right]. \tag{63}$$

From (31) and (32) one derives the ''in-betweenness conditions,'' which define completely the new basis (53). The transformations of the C basis are most easily written in terms of the following variables:

$$\begin{aligned} &L_{0,2k+\theta}=M_{0,2k+\theta},\\ &L_{i,2k+\theta}=-M_{i,2k+\theta}+i+1, \quad i\in[-k;-1],\\ &L_{i,2k+\theta}=-M_{i,2k+\theta}+i-1, \quad i\in[1;k-1+\theta]. \end{aligned} \tag{64}$$

We formulate the result as a proposition.

Proposition 5: The $2n+1$*-tuple* $[M]_{2n+1}=[M_{-n,2n+1},M_{-n+1,2n+1},\ldots,M_{n,2n+1}]$ *is a signature of an essentially typical* $gl(n|1|n)$ *module* $V([M]_{2n+1})$ *if and only if*

$$M_{i,2n+1}\in\mathbf{C}, \quad i\in[-n;n], \tag{65a}$$

$$M_{i,2n+1}-M_{i+1,2n+1}\in\mathbf{Z}_+, \quad i\in[-n;-2]\cup[1;n-1], \tag{65b}$$

$$M_{-1,2n+1}-M_{1,2n+1}\in\mathbf{N}, \tag{65c}$$

$$M_{0,2n+1}=L_{0,2n+1}\notin[L_{-n,2n+1};L_{n,2n+1}]. \tag{65d}$$

The C basis $\Gamma([M]_{2n+1})$ *in* $V([M]_{2n+1})$ *consists of all tables (53) for which the labels*

$$M_{i,2k+\theta}, \quad \theta\in\{0,1\}, \quad k\in[1-\theta;n], \quad i\in[-k;k-1+\theta], \tag{66}$$

take all possible values consistent with the ''in-betweenness conditions''

$$M_{i,2k+1}-M_{i,2k}\in\mathbf{Z}_+, \quad k\in[1;n], \quad i\in[-k;-1]\cup[1;k-1], \tag{67a}$$

$$M_{i,2k-1}-M_{i,2k}\in\mathbf{Z}_+, \quad k\in[2;n], \quad i\in[-k+1;-1]\cup[1;k-1], \tag{67b}$$

$$M_{i-1,2k}-M_{i,2k-1}\in\mathbf{Z}_+, \quad k\in[2;n], \quad i\in[-k+1;-1]\cup[2;k-1], \tag{67c}$$

$$M_{i-1,2k}-M_{i,2k+1}\in\mathbf{Z}_+, \quad k\in[1;n], \quad i\in[-k+1;-1]\cup[2;k], \tag{67d}$$

$$M_{-1,2k}-M_{1,2k-1}\in\mathbf{N}, \quad k\in[2;n], \tag{67e}$$

$$M_{-1,2k}-M_{1,2k+1}\in\mathbf{N}, \quad k\in[1;n], \tag{67f}$$

$$M_{0,2k+1}-M_{0,2k}\equiv\psi_{2k}\in\{0,1\},\quad k\in[1;n],\tag{67g}$$

$$M_{0,2k}-M_{0,2k-1}\equiv\psi_{2k-1}\in\{0,-1\},\quad k\in[1;n].\tag{67h}$$

The transformations of the C basis under the action of the inverse images $\varphi^{-1}(e_{ii})$, $\varphi^{-1}(e_{i,i+1})$, and $\varphi^{-1}(e_{i+1,i})$ of the $gl_0(1|2n)$ Chevalley generators follow from (33) to (36) and (61), (62). The result reads [we write E_{ij} instead of $\varphi(E_{ij})$]:

$$E_{ii}|M\rangle=\left(\sum_{j=-|i|}^{|i|+\theta(i)-1}M_{j,2|i|+\theta(i)}-\sum_{j=-|i|+1-\theta(i)}^{|i|-1}M_{j,2|i|+\theta(i)-1}\right)|M\rangle,\quad i\in[-n;n],\tag{68}$$

$$E_{0,-1}|M\rangle=(1+\psi_1)|M\rangle_{(0,1)},\tag{69}$$

$$E_{-1,0}|M\rangle=-\psi_1(L_{0,2}-L_{-1,2})|M\rangle_{-(0,1)},\tag{70}$$

$$\begin{aligned}E_{i-1,-i}|M\rangle&=(1+\psi_{2i-1})(1-\psi_{2i-2})|M\rangle_{(0,2i-1)}\\&+\sum_{j\neq 0=-i+1}^{i-1}\left(-\frac{\prod_{k\neq 0=-i+1}^{i-2}(L_{k,2i-2}-L_{j,2i-1}+1)\prod_{k\neq 0=-i}^{i-1}(L_{k,2i}-L_{j,2i-1}+1)}{\prod_{k\neq 0,j;k=-i+1}^{i-1}(L_{k,2i-1}-L_{j,2i-1})(L_{k,2i-1}-L_{j,2i-1}+1)}\right)^{1/2}\\&\times\frac{(L_{0,2i-1}-L_{j,2i-1})(L_{0,2i-1}-L_{j,2i-1}+1)}{(L_{0,2i}-L_{j,2i-1}+1)(L_{0,2i-2}-L_{j,2i-1}+1)}|M\rangle_{(j,2i-1)},\quad i\in[2;n],\end{aligned}\tag{71}$$

$$\begin{aligned}E_{-i,i}|M\rangle&=-\psi_{2i}\psi_{2i-1}|M\rangle_{(0,2i)}\\&+\sum_{j\neq 0=-i}^{i-1}\left(-\frac{\prod_{k\neq 0=-i+1}^{i-1}(L_{k,2i-1}-L_{j,2i})\prod_{k\neq 0=-i}^{i}(L_{k,2i+1}-L_{j,2i})}{\prod_{k\neq 0,j;k=-i}^{i-1}(L_{k,2i}-L_{j,2i})(L_{k,2i}-L_{j,2i}+1)}\right)^{1/2}\\&\times\frac{(L_{0,2i}-L_{j,2i})(L_{0,2i}-L_{j,2i}+1)}{(L_{0,2i+1}-L_{j,2i})(L_{0,2i-1}-L_{j,2i})}|M\rangle_{(j,2i)},\quad i\in[1;n],\end{aligned}\tag{72}$$

$$\begin{aligned}E_{i,-i}|M\rangle&=(1+\psi_{2i-1})(1-\psi_{2i})\\&\times\frac{\prod_{k\neq 0=-i+1}^{i-1}(L_{0,2i+1}-L_{k,2i-1})\prod_{k\neq 0=-i}^{i}(L_{0,2i+1}-L_{k,2i+1})}{\prod_{k\neq 0=-i}^{i-1}(L_{0,2i+1}-L_{k,2i}-1)(L_{0,2i+1}-L_{k,2i})}|M\rangle_{-(0,2i)}\\&+\sum_{j\neq 0=-i}^{i-1}\left(-\frac{\prod_{k\neq 0=-i+1}^{i-1}(L_{k,2i-1}-L_{j,2i}-1)\prod_{k\neq 0=-i}^{i}(L_{k,2i+1}-L_{j,2i}-1)}{\prod_{k\neq 0,j;k=-i}^{i-1}(L_{k,2i}-L_{j,2i}-1)(L_{k,2i}-L_{j,2i})}\right)^{1/2}\\&\times|M\rangle_{-(j,2i)},\quad i\in[1;n],\end{aligned}\tag{73}$$

$$\begin{aligned}E_{-i,i-1}|M\rangle&=-\psi_{2i-2}\psi_{2i-1}\frac{\prod_{k\neq 0=-i+1}^{i-2}(L_{0,2i}-L_{k,2i-2})\prod_{k\neq 0=-i}^{i-1}(L_{0,2i}-L_{k,2i})}{\prod_{k\neq 0=-i+1}^{i-1}(L_{0,2i}-L_{k,2i-1})(L_{0,2i}-L_{k,2i-1}+1)}|M\rangle_{-(0,2i-1)}\\&+\sum_{j\neq 0=-i+1}^{i-1}\left(-\frac{\prod_{k\neq 0=-i+1}^{i-2}(L_{k,2i-2}-L_{j,2i-1})\prod_{k\neq 0=-i}^{i-1}(L_{k,2i}-L_{j,2i-1})}{\prod_{k\neq 0,j;k=-i+1}^{i-1}(L_{k,2i-1}-L_{j,2i-1}-1)(L_{k,2i-1}-L_{j,2i-1})}\right)^{1/2}\\&\times|M\rangle_{-(j,2i-1)},\quad i\in[2;n].\end{aligned}\tag{74}$$

We have written the transformation relations of the *C* basis under the action of generators, which are different from the $gl(n|1|n)$ Chevalley elements. These generators, however, define completely all other generators. In this sense Eqs. (68)–(74) are complete. We shall use them in order to derive the transformations of the $gl(\infty|1|\infty)$ irreducible modules under the action of the Chevalley generators.

Remark: We are thankful to the referee for pointing out that *Proposition 4* can be proved also without using the transformation relations (34)–(36). To this end note [see Eqs. (58) and (59)] that the $gl(k|1|k-1+\theta)$ highest weight vector $y_{2k+\theta}\equiv|m\rangle_{2k+\theta}\in V([m]_{2k+\theta})$ is determined from the requirement to be annihilated by the generators $\{\varphi(E_{-i,-i+1})|i\in[1;k]\}\cup\{\varphi(E_{i,i+1})|i\in[0;k-2+\theta]\}$, i.e., by $\{e_{2k,2k-2},e_{2k-2,2k-4},\dots,e_{42},e_{21},e_{13},e_{35},\dots,e_{2k+2\theta-3,2k+2\theta-1}\}$. The roots, corresponding to the above root vectors, namely

$$\hat{\pi}_{2k+\theta}=\{\epsilon^{2k}-\epsilon^{2k-2},\epsilon^{2k-2}-\epsilon^{2k-4},\dots,\epsilon^{4}-\epsilon^{2},\epsilon^{2}-\epsilon^{1},\epsilon^{1}-\epsilon^{3},\epsilon^{3}-\epsilon^{5},\dots,\epsilon^{2k+2\theta-3}-\epsilon^{2k+2\theta-1}\} \tag{75}$$

can be taken as a new system of simple roots of $gl(1|2k+\theta-1)$ with a system of positive roots $\hat{\Delta}_{+}^{2k+\theta}$.

Let $\Lambda_{2k+\theta}\equiv[m]_{2k+\theta}\equiv\Sigma_{i=1}^{2k+\theta}m_{i,2k+\theta}\epsilon^{i}$ be the standard signature (=the highest weight) of $V([m]_{2k+\theta})$, namely the signature corresponding to the choice of simple roots

$$\pi_{2k+\theta}=\{\epsilon^{1}-\epsilon^{2},\epsilon^{2}-\epsilon^{3},\dots,\epsilon^{2k-1+\theta}-\epsilon^{2k+\theta}\}. \tag{76}$$

Denote by $\Delta_{+}^{2k+\theta}$ the system of positive roots corresponding to it. The problem is to determine the signature (=the highest weight) $\hat{\Lambda}_{2k+\theta}$ of $V([m]_{2k+\theta})$ with respect to $\hat{\Delta}_{+}^{2k+\theta}$. This problem can be solved on the ground of results from Refs. 39 and 40. Given a subset of positive roots Δ'_{+} of $gl(1|2k+\theta-1)$ and a simple root $\alpha\in\Delta'_{+}$, one constructs a new system of positive roots Δ''_{+} by a simple α reflection $\langle\alpha\rangle$:[39,40]

$$\Delta''_{+}=\langle\alpha\rangle(\Delta'_{+})=\begin{cases} r_{\alpha}(\Delta'_{+}) & \text{if } \alpha \text{ is even}\\ (\Delta'_{+}\cup\{-\alpha\})\backslash\{\alpha\} & \text{if } \alpha \text{ is odd,}\end{cases} \tag{77}$$

where r_{α} is an element from the Weyl group of $gl(1|2k+\theta-1)$, corresponding to α.

If $V_{2k+\theta}$ is an essentially typical $gl(1|2k+\theta-1)$ module with a highest weight λ', corresponding to Δ'_{+}, then the highest weight with respect to Δ''_{+} is

$$\lambda''=r_{\alpha}(\lambda') \quad \text{if } \alpha \text{ is an even root,} \quad \lambda''=\lambda'-\alpha \quad \text{if } \alpha \text{ is an odd root.} \tag{78}$$

Let $\Pi_{i=1}^{N}\langle\alpha_i\rangle=\langle\alpha_1\rangle\langle\alpha_2\rangle\cdots\langle\alpha_N\rangle$. Then

$$\hat{\Delta}_{+}^{2k+\theta}=\prod_{i=1}^{k}\prod_{j=1}^{2i-1}\langle\epsilon_j-\epsilon_{2i}\rangle\Delta_{+}^{2k+\theta}. \tag{79}$$

From (77) to (79) one derives that

$$\hat{\Lambda}_{2k+\theta}=\sum_{j=2}^{k+1}(m_{j,2k+\theta}+1)\epsilon^{2k-2j+4}+(m_{1,2k+\theta}-k)\epsilon^{1}+\sum_{j=k+2}^{2k+\theta}m_{j,2k+\theta}\epsilon^{2j-2k-1}, \tag{80}$$

i.e.,

$$e_{2k-2i+4,2k-2i+4}|m\rangle_{2k+\theta}=(m_{i,2k+\theta}+1)|m\rangle_{2k+\theta}, \quad i\in[2;k+1], \tag{81a}$$

$$e_{11}|m\rangle_{2k+\theta}=(m_{1,2k+\theta}-k)|m\rangle_{2k+\theta}, \tag{81b}$$

$$e_{2i-2k-1,2i-2k-1}|m\rangle_{2k+\theta}=m_{i,2k+\theta}|m\rangle_{2k+\theta}, \quad i\in[k+2;2k+\theta]. \tag{81c}$$

Equations (81) are the same as (60) (written in somewhat different notation). Hence one obtains the $gl(k|1|k+\theta-1)$ signature as given in (61) and the highest weight $|m\rangle_{2k+\theta}$ corresponding to it (*Proposition 4*).

III. IRREDUCIBLE REPRESENTATIONS OF $gl(1|\infty)$

Here we construct representations of $gl_0(1|\infty)$ and $gl(\infty|1|\infty)$, which appear as a generalization to the case $n\to\infty$ of the results obtained in Sec. III. In both cases the representations (or the corresponding modules) are labeled with infinite sequences of (in general different) complex numbers. Due to the isomorphism φ [see (4)] each $gl_0(1|\infty)$ module is also a $gl(\infty|1|\infty)$ module and vice versa. Therefore we can also say that we describe below two classes of representations of the ''abstract'' Lie superalgebra $gl(1|\infty)$. For definiteness we refer to the class of representations of $gl_0(1|\infty)$ as the Gel'fand–Zetlin (GZ) representations (Sec. III A), whereas the representations of $gl(\infty|1|\infty)$ are said to be C representations.

A. Gel'fand–Zetlin representations

The extension of the results of Sec. II to the case $n\to\infty$ is rather evident. We collect the results in a proposition.

Proposition 6: To each sequence of complex numbers

$$[m]\equiv[m_1,m_2,\dots,m_k,\dots]\equiv\{m_i|m_i\in\mathbf{C},i\in\mathbf{N}\}, \tag{82}$$

such that

$$m_i-m_{i+1}\in\mathbf{Z}_+,\quad i=2,3,...,$$
$$l_1\notin\{l_2,l_2+1,l_2+2,...\}, \tag{83}$$

where

$$l_1=m_1+1,\quad l_i=-m_i+i-1,\quad i=2,3,..., \tag{84}$$

there corresponds an irreducible highest weight $gl_0(1|\infty)$ module $V([m])$ with a signature (82). The basis $\Gamma([m])$ in $V([m])$, which we call a GZ basis, consists of all tables

$$|m)\equiv\left[\begin{array}{ccccccc} m_1 & m_2 & \cdots & m_j & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & & \\ m_{1j} & m_{2j} & \cdots & m_{jj} & & & \\ \cdots & \cdots & \cdots & & & & \\ \cdots & \cdots & \cdots & & & & \\ m_{12} & m_{22} & & & & & \\ m_{11} & & & & & & \end{array}\right]\equiv\left[\begin{array}{c}[m]\\ \vdots \\ [m]_j \\ \vdots \\ [m]_2 \\ m_{11}\end{array}\right], \tag{85}$$

characterized by an infinite number of coordinates

$$m_{ij},\quad \forall j\in\mathbf{N},\quad i=1,2,...,j, \tag{86}$$

which are consistent with the conditions:

(1) for each table $|m)$ there exists a positive (depending on $|m)$) integer $N[|m)]\in\mathbf{N}$ such that

$$m_{ij}=m_i,\quad \forall j>N[|m)],\quad i=1,...,j, \tag{87}$$

$$\text{(2)}\quad m_{1i}-m_{1,i-1}\equiv\theta_{i-1}\in\{0,1\},\quad i=2,3,..., \tag{88}$$

$$\text{(3)}\quad m_{i,j+1}-m_{ij}\in\mathbf{Z}_+,\quad m_{ij}-m_{i+1,j+1}\in\mathbf{Z}_+,\quad 2\le i\le j\in\mathbf{N}. \tag{89}$$

The transformation of the basis (85) is determined from the action of the Chevalley generators

$$e_{ii}|m)=\left(\sum_{k=1}^{i} m_{ki}-\sum_{k=1}^{i-1} m_{k,i-1}\right)|m), \quad i\in\mathbf{N}, \tag{90}$$

$$e_{12}|m)=\theta_1|m)_{11}, \quad e_{21}|m)=(1-\theta_1)(l_{12}-l_{22})|m)_{-(1,1)}, \tag{91}$$

$$\begin{aligned} e_{i,i+1}|m)=&\theta_i(1-\theta_{i-1})|m)_{(1i)}\\ &+\sum_{j=2}^{i}\left(-\frac{\Pi_{k=2}^{i-1}(l_{k,i-1}-l_{ji}+1)\Pi_{k=2}^{i+1}(l_{k,i+1}-l_{ji})}{\Pi_{k\neq j=2}^{i}(l_{ki}-l_{ji})(l_{ki}-l_{ji}+1)}\right)^{1/2}\\ &\times\frac{(l_{1i}-l_{ji})(l_{1i}-l_{ji}+1)}{(l_{1,i+1}-l_{ji})(l_{1,i-1}-l_{ji}+1)}|m)_{(ji)}, \quad i=2,3,..., \end{aligned} \tag{92}$$

$$\begin{aligned} e_{i+1,i}|m)=&\theta_{i-1}(1-\theta_i)\frac{\Pi_{k=2}^{i-1}(l_{1,i+1}-l_{k,i-1}-1)\Pi_{k=2}^{i+1}(l_{1,i+1}-l_{k,i+1})}{\Pi_{k=2}^{i}(l_{1,i+1}-l_{ki}-1)(l_{1,i+1}-l_{ki})}|m)_{-(1,i)}\\ &+\sum_{j=2}^{i}\left(-\frac{\Pi_{k=2}^{i-1}(l_{k,i-1}-l_{ji})\Pi_{k=2}^{i+1}(l_{k,i+1}-l_{ji}-1)}{\Pi_{k\neq j=2}^{i}(l_{ki}-l_{ji}-1)(l_{ki}-l_{ji})}\right)^{1/2}|m)_{-(ji)}, \quad i=2,3,.... \end{aligned} \tag{93}$$

The highest weight vector $|\hat{m})$ *is the one from (85) for which*

$$m_{ij}=m_i, \quad \forall j\in\mathbf{N}, \quad i\in 1,2,...,j. \tag{94}$$

Proof: Let

$$|m)\equiv\begin{bmatrix}[m]\\ \vdots\\ [m]_{N+1}\\ \vdots\\ [m]_2\\ m_{11}\end{bmatrix}\in\Gamma([m]). \tag{95}$$

Then

(i) $[m]_{N+1}\equiv[m_{1,N+1},m_{2,N+1},\dots,m_{N+1,N+1}]$, $N=1,2,...$, is said to be the $(N+1)$th signature of $|m)$;

(ii)

$$|m)^{\text{up}(N+1)}\equiv\begin{bmatrix}[m]\\ \vdots\\ [m]_j\\ \vdots\\ [m]_{N+2}\end{bmatrix}, \quad |m)^{\text{low}(N+1)}\equiv\begin{bmatrix}[m]_{N+1}\\ \vdots\\ [m]_i\\ \vdots\\ [m]_2\\ m_{11}\end{bmatrix} \tag{96}$$

are said to be the $(N+1)$th upper and the $(N+1)$th lower part of $|m)$, respectively. Consider the subalgebra

$$gl_0(1|N)=\{e_{ij}|i,j=1,...,N+1\}\subset gl_0(1|\infty). \tag{97}$$

Observation 1: Let e be a $gl_0(1|N)$ generator or any polynomial of $gl_0(1|N)$ generators. Then, for any $|m)\in\Gamma([m])$, $e|m)$ is a linear combination of vectors from $\Gamma([m])$ with one and same $(N+1)$th upper part $|m)^{\text{up}(N+1)}$.

Denote by

$$\Gamma([m]_i|i\geqslant N+1)\subset\Gamma([m]) \tag{98}$$

the set of all vectors (85), that have one and the same $[m]_i$ signatures, for all $i\geqslant N+1$. Let

$$V([m]_i|i\geqslant N+1)\subset V([m]) \tag{99}$$

be the linear span of $\Gamma([m]_i|i\geqslant N+1)$. From (90) to (93) it follows that $V([m]_i|i\geqslant N+1)$ is invariant with respect to $gl_0(1|N)$. To each vector $|m)\in\Gamma([m]_i|i\geqslant N+1)$ put in correspondence its $(N+1)$th lower part:

$$f(|m))=|m)^{\text{low}(N+1)},\quad \forall|m)\in\Gamma([m]_i|i\geqslant N+1). \tag{100}$$

Let

$$\Gamma([m]_{N+1})=\{f(|m))||m)\in\Gamma([m]_i|i\geqslant N+1)\}. \tag{101}$$

Then f maps bijectively $\Gamma([m]_i)|i\geqslant N+1)$ on $\Gamma([m]_{N+1})$. Obviously $\Gamma([m]_{N+1})$ consists of all GZ tables of an essentially typical $gl_0(1|N)$ module with a signature $[m]_{N+1}$. Define an action of $gl_0(1|N)$ on $|m)\in\Gamma([m]_{N+1})$ with the relations (33)–(36). Then the linear envelope $V([m]_{N+1})$ of $\Gamma([m]_{N+1})$ is an essentially typical $gl_0(1|N)$ module with a signature $[m]_{N+1}$. After comparing the relations (90)–(93) with (33)–(36) and having in mind *Observation 1* we have:

Observation 2: The subspace $V([m]_i|i\geqslant N+1)\subset V([m])$ is an essentially typical finite-dimensional $gl_0(1|N)$ module with a signature $[m]_{N+1}$ and a GZ basis $\Gamma([m]_i|i\geqslant N+1)$.

Let e_{ij},e_{kl} be any two generators from $gl_0(1|\infty)$ and $|m)$ be an arbitrary vector from $\Gamma([m])$. Consider e_{ij},e_{kl} as elements from $gl_0(1|N)\subset gl_0(1|\infty)$, where $N+1\geqslant\max(i,j,k,l)$. Then $|m)$ is a vector from the $gl_0(1|N)$ fidirmod $V([m]_i|i\geqslant N+1)\subset V([m])$ and therefore (*Observation 2*)

$$(e_{ij}e_{kl}-(-1)^{\deg(e_{ij})\deg(e_{kl})}e_{kl}e_{ij})|m)=(\delta_{jk}e_{il}-(-1)^{\deg(e_{ij})\deg(e_{kl})}\delta_{li}e_{kj})|m). \tag{102}$$

Therefore the linear space $V([m])$ is a $gl_0(1|\infty)$ module.

Consider any two vectors $x,y\in V([m])$,

$$x=\sum_{i=1}^{p}\alpha_i|m^i),\quad y=\sum_{i=p+1}^{q}\alpha_i|m^i),\quad |m^i)\in\Gamma([m]),$$

$$\alpha_i\in\mathbf{C},\quad i=1,...,q. \tag{103}$$

Let

$$\tilde{N}=\max\{N[|m^i)]|i=1,...,q\}. \tag{104}$$

According to (87) all vectors $|m^i)$, $i=1,...,q$, have one and the same $k-1$ signatures, for every $k-1\geqslant\tilde{N}$. Therefore $|m^i)\in V([m]_{k-1}|k-1\geqslant\tilde{N})\subset V([m])$. Hence $x,y\in V([m]_{k-1}|k-1\geqslant\tilde{N})$. The space $V([m]_{k-1}|k-1\geqslant\tilde{N})$ is a $gl_0(1|\tilde{N})$ fidirmod (*Observation 2*) and, therefore, there exist a polynomial P of the $gl_0(1|\tilde{N})$ generators such that $y=Px$. Hence $V([m])$ is an irreducible $gl_0(1|\infty)$ module.

Consider the vector $|\hat{m})\in\Gamma([m])$ [see (91)]. From Eqs. (90) to (93) we have

$$e_{ii}|\hat{m})=m_i|\hat{m}),\quad \forall i\in\mathbf{N}, \tag{105}$$

and

$$e_{k,k+1}|\hat{m})=0,\quad \forall k\in\mathbf{N}. \tag{106}$$

Therefore the irreducible $gl_0(1|\infty)$ module $V([m])$ is a highest weight module with a signature

$$[m] \equiv [m_1, m_2, \ldots, m_k, \ldots] \tag{107}$$

and a highest weight vector $|\hat{m})$. This completes the proof.

B. *C* representations

Most of the preliminary work for constructing the representations of $gl(\infty|1|\infty)$ was done in Sec. II B. It remains to give a precise definition of the C basis in the infinite-dimensional case and to write down the transformation of the basis under the action of the Chevalley generators.

Let

$$[M] \equiv [\ldots, M_{-p}, \ldots, M_{-1}, M_0, M_1, M_2, \ldots] \equiv \{M_i\}_{i\in\mathbf{Z}} \tag{108}$$

be a sequence of complex numbers such that

$$M_i - M_{i+1} \in \mathbf{Z}_+, \quad i \in [-\infty; -2] \cup [1; \infty], \tag{109a}$$

$$M_{-1} - M_1 \in \mathbf{N}, \tag{109b}$$

$$M_0 + M_1 \notin \mathbf{Z}. \tag{109c}$$

Here and throughout

$$[-\infty; a] = \{a, a-1, a-2, \ldots, a-i, \ldots\} \equiv \{a-i\}_{i\in\mathbf{Z}_+}, \tag{110}$$

$$[b; \infty] = \{b, b+1, b+2, \ldots, b+i, \ldots\} \equiv \{b+i\}_{i\in\mathbf{Z}_+}. \tag{111}$$

A table $|M)$, consisting of infinitely many complex numbers

$$M_{i,2k+\theta-1}, \quad \forall k \in \mathbf{N}, \quad \theta \in \{0,1\}, \quad i = [-k-\theta+1; k-1], \tag{112}$$

will be called a C table, provided the following conditions hold.

(1) There exists a positive, depending on $|M)$, integer $N[|M)]$ such that

$$M_{i,2k+\theta-1} = M_i, \quad \forall k > N[|M)], \quad \theta \in \{0,1\}, \quad i \in [1-\theta-k; k-1]. \tag{113}$$

(2) The coordinates $M_{i,2k+\theta-1}$, $\theta \in \{0,1\}$, take all possible values

$$M_{i,2k+1-2\theta} - M_{i,2k} \in \mathbf{Z}_+, \quad k \in [1+\theta; \infty], \quad i \in [-k+\theta; -1] \cup [1; k-1], \tag{114a}$$

$$M_{i-1,2k} - M_{i,2k+1-2\theta} \in \mathbf{Z}_+, \quad k \in [1+\theta; \infty], \quad i \in [-k+1; -1] \cup [2; k-\theta], \tag{114b}$$

$$M_{-1,2k} - M_{1,2k+1-2\theta} \in \mathbf{N}, \quad k \in [1+\theta; \infty], \tag{114c}$$

$$M_{0,2k+1-\theta} - M_{0,2k-\theta} \equiv \psi_{2k-\theta} \in \{0, 1-2\theta\}, \quad k \in [1; \infty]. \tag{114d}$$

Order the complex numbers $M_{i,2k+\theta-1}$, $k \in \mathbf{N}$, $\theta \in \{0,1\}$, as in the table below

$$|M)\equiv\left[\begin{array}{cccccccc} \cdots & M_{1-\theta-k} & \cdots & M_{-1} & M_0 & M_1 & \cdots & M_{k-1\cdots} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & M_{1-\theta-k,2k+\theta-1} & \cdots & M_{-1,2k+\theta-1} & M_{0,2k+\theta-1} & M_{1,2k+\theta-1} & \cdots & M_{k-1,2k+\theta-1} \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ & & & M_{-1,3} & M_{03} & M_{13} & & \\ & & & M_{-1,2} & M_{02} & & & \\ & & & & M_{01} & & & \end{array}\right]. \tag{115}$$

We are ready now to state our main and final result.

Proposition 7: To each sequence (108) [see also (109)] there corresponds an irreducible highest weight $gl(\infty|1|\infty)$ module $V([M])$ with a signature $[M]$. The basis $\Gamma([M])$ in $V([M])$ consists of all C tables (115). The transformations of the basis under the action of the $gl(\infty|1|\infty)$ Chevalley generators read:

$$E_{kk}|M)=\left(\sum_{i=-|k|}^{|k|+\theta(k)-1} M_{i,2|k|+\theta(k)}-\sum_{i=-|k|+1-\theta(k)}^{|k|-1} M_{i,2|k|+\theta(k)-1}\right)|M),\quad k\in\mathbf{Z}, \tag{116}$$

$$E_{0,-1}|M)=(1+\psi_1)|M)_{(01)}, \tag{117}$$

$$E_{-1,0}|M)=-\psi_1(L_{0,2}-L_{-1,2})|M)_{-(01)}, \tag{118}$$

$$\begin{aligned} E_{01}|M)=&-\psi_2(1+2\psi_1)|M)^{(01)}_{(0,2)}+(1+\psi_1)(-(L_{-1,3}-L_{-1,2})(L_{13}-L_{-1,2}))^{1/2}\\ &\times\frac{(L_{02}-L_{-1,2})(L_{02}-L_{-1,2}+1)}{(L_{03}-L_{-1,2})(L_{01}-L_{-1,2})(L_{01}-L_{-1,2}+1)}|M)^{(01)}_{(-1,2)}, \end{aligned} \tag{119}$$

$$\begin{aligned} E_{10}|M)=&-(-1)^{\psi_1}(1-\psi_2)\frac{(L_{02}-L_{-1,2}-\psi_1-1)(L_{03}-L_{-1,3})(L_{03}-L_{13})}{(L_{03}-L_{-1,2}-1)(L_{03}-L_{-1,2})}|M)^{-(01)}_{-(02)}\\ &-\psi_1(-(L_{-1,3}-L_{-1,2}-1)(L_{13}-L_{-1,2}-1))^{1/2}|M)^{-(01)}_{-(-1,2)}, \end{aligned} \tag{120}$$

$$\begin{aligned} E_{k,k+1}|M)=&-\psi_{2k+2}(1-\psi_{2k})(1+2\psi_{2k+1})|M)^{(0,2k+1)}_{(0,2k+2)}+\sum_{j\neq 0=-k}^{k}\psi_{2k+2}\psi_{2k+1}\\ &\times\left(-\frac{\Pi_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{j,2k+1}+1)\Pi_{i\neq 0=-k-1}^{k}(L_{i,2k+2}-L_{j,2k+1}+1)}{\Pi_{i\neq 0,j;i=-k}^{k}(L_{i,2k+1}-L_{j,2k+1})(L_{i,2k+1}-L_{j,2k+1}+1)}\right)^{1/2}\\ &\times\frac{(L_{0,2k+1}-L_{j,2k+1})(L_{0,2k+1}-L_{j,2k+1}+1)}{(L_{0,2k+2}-L_{j,2k+1}+2)(L_{0,2k+2}-L_{j,2k+1}+1)(L_{0,2k}-L_{j,2k+1}+1)}|M)^{(j,2k+1)}_{(0,2k+2)}\\ &+\sum_{j\neq 0=-k-1}^{k}(1+\psi_{2k+1})(1-\psi_{2k})\\ &\times\left(-\frac{\Pi_{i\neq 0=-k}^{k}(L_{i,2k+1}-L_{j,2k+2})\Pi_{i\neq 0=-k-1}^{k+1}(L_{i,2k+3}-L_{j,2k+2})}{\Pi_{i\neq 0,j;i=-k-1}^{k}(L_{i,2k+2}-L_{j,2k+2})(L_{i,2k+2}-L_{j,2k+2}+1)}\right)^{1/2} \end{aligned}$$

$$\times\frac{(L_{0,2k+2}-L_{j,2k+2})(L_{0,2k+2}-L_{j,2k+2}+1)}{(L_{0,2k+3}-L_{j,2k+2})(L_{0,2k+1}-L_{j,2k+2})(L_{0,2k+1}-L_{j,2k+2}+1)}|M)^{(0,2k+1)}_{(j,2k+2)}$$

$$+\sum_{l\neq 0=-k-1}^{k}\sum_{j\neq 0=-k}^{k}Q(j,l)$$

$$\times\left(-\frac{\Pi_{i\neq 0,j;i=-k}^{k}(L_{i,2k+1}-L_{l,2k+2})\Pi_{i\neq 0=-k-1}^{k+1}(L_{i,2k+3}-L_{l,2k+2})}{\Pi_{i\neq 0,l;i=-k-1}^{k}(L_{i,2k+2}-L_{l,2k+2})(L_{i,2k+2}-L_{l,2k+2}+1)}\right)^{1/2}$$

$$\times\left(\frac{\Pi_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{j,2k+1}+1)\Pi_{i\neq 0,l;i=-k-1}^{k}(L_{i,2k+2}-L_{j,2k+1}+1)}{\Pi_{i\neq 0,j;i=-k}^{k}(L_{i,2k+1}-L_{j,2k+1})(L_{i,2k+1}-L_{j,2k+1}+1)}\right)^{1/2}$$

$$\times\frac{(L_{0,2k+2}-L_{l,2k+2})(L_{0,2k+2}-L_{l,2k+2}+1)(L_{0,2k+1}-L_{j,2k+1})(L_{0,2k+1}-L_{j,2k+1}+1)}{(L_{0,2k+3}-L_{l,2k+2})(L_{0,2k+1}-L_{l,2k+2})(L_{0,2k+2}-L_{j,2k+1}+1)(L_{0,2k}-L_{j,2k+1}+1)}$$

$$\times|M)^{(j,2k+1)}_{(l,2k+2)},\quad k\in[1;\infty], \tag{121}$$

$$E_{-k+1,-k}|M)$$

$$=-(1+\psi_{2k-1})\psi_{2k-3}(1-2\psi_{2k-2})|M)^{(0,2k-2)}_{(0,2k-1)}-\sum_{j\neq 0=-k+1}^{k-2}(1+\psi_{2k-1})(1-\psi_{2k-2})$$

$$\times\left(-\frac{\Pi_{i\neq 0=-k+2}^{k-2}(L_{i,2k-3}-L_{j,2k-2})\Pi_{i\neq 0=-k+1}^{k-1}(L_{i,2k-1}-L_{j,2k-2})}{\Pi_{i\neq 0,j;i=-k+1}^{k-2}(L_{i,2k-2}-L_{j,2k-2})(L_{i,2k-2}-L_{j,2k-2}+1)}\right)^{1/2}$$

$$\times\frac{(L_{0,2k-2}-L_{j,2k-2})(L_{0,2k-2}-L_{j,2k-2}+1)}{(L_{0,2k-1}-L_{j,2k-2})(L_{0,2k-1}-L_{j,2k-2}+1)(L_{0,2k-3}-L_{j,2k-2})}|M)^{(j,2k-2)}_{(0,2k-1)}$$

$$-\sum_{j\neq 0=-k+1}^{k-1}\psi_{2k-2}\psi_{2k-3}$$

$$\times\left(-\frac{\Pi_{i\neq 0=-k+1}^{k-2}(L_{i,2k-2}-L_{j,2k-1}+1)\Pi_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{j,2k-1}+1)}{\Pi_{i\neq 0,j;i=-k+1}^{k-1}(L_{i,2k-1}-L_{j,2k-1})(L_{i,2k-1}-L_{j,2k-1}+1)}\right)^{1/2}$$

$$\times\frac{(L_{0,2k-1}-L_{j,2k-1})(L_{0,2k-1}-L_{j,2k-1}+1)}{(L_{0,2k}-L_{j,2k-1}+1)(L_{0,2k-2}-L_{j,2k-1}+1)(L_{0,2k-2}-L_{j,2k-1}+2)}|M)^{(0,2k-2)}_{(j,2k-1)}$$

$$+\sum_{l\neq 0=-k+1}^{k-1}\sum_{j\neq 0=-k+1}^{k-2}P(j,l)$$

$$\times\left(-\frac{\Pi_{i\neq 0,j;i=-k+1}^{k-2}(L_{i,2k-2}-L_{l,2k-1}+1)\Pi_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{l,2k-1}+1)}{\Pi_{i\neq 0,l;i=-k+1}^{k-1}(L_{i,2k-1}-L_{l,2k-1})(L_{i,2k-1}-L_{l,2k-1}+1)}\right)^{1/2}$$

$$\times\left(\frac{\Pi_{i\neq 0=-k+2}^{k-2}(L_{i,2k-3}-L_{j,2k-2})\Pi_{i\neq 0,l;i=-k+1}^{k-1}(L_{i,2k-1}-L_{j,2k-2})}{\Pi_{i\neq 0,j;i=-k+1}^{k-2}(L_{i,2k-2}-L_{j,2k-2})(L_{i,2k-2}-L_{j,2k-2}+1)}\right)^{1/2}$$

$$\times\frac{(L_{0,2k-1}-L_{l,2k-1})(L_{0,2k-1}-L_{l,2k-1}+1)(L_{0,2k-2}-L_{j,2k-2})(L_{0,2k-2}-L_{j,2k-2}+1)}{(L_{0,2k}-L_{l,2k-1}+1)(L_{0,2k-2}-L_{l,2k-1}+1)(L_{0,2k-1}-L_{j,2k-2})(L_{0,2k-3}-L_{j,2k-2})}$$

$$\times|M)^{(j,2k-2)}_{(l,2k-1)},\quad k\in[2;\infty], \tag{122}$$

$$
\begin{aligned}
E_{k+1,k}|M) = & -(-1)^{\psi_{2k+1}}\psi_{2k}(1-\psi_{2k+2}) \\
& \times \frac{\prod_{i\neq 0=-k}^{k-1}(L_{0,2k+2}-L_{i,2k}-\psi_{2k+1}-1)\prod_{i\neq 0=-k-1}^{k}(L_{0,2k+2}-L_{i,2k+2}-\psi_{2k+1}-1)}{\prod_{i\neq 0=-k}^{k}(L_{0,2k+2}-L_{i,2k+1}-\psi_{2k+1}-1)(L_{0,2k+2}-L_{i,2k+1}-\psi_{2k+1})} \\
& \times \frac{\prod_{i\neq 0=-k}^{k}(L_{0,2k+3}-L_{i,2k+1})\prod_{i\neq 0=-k-1}^{k+1}(L_{0,2k+3}-L_{i,2k+3})}{\prod_{i\neq 0=-k-1}^{k}(L_{0,2k+3}-L_{i,2k+2}-1)(L_{0,2k+3}-L_{i,2k+2})}|M)^{-(0,2k+1)}_{-(0,2k+2)} \\
& -\sum_{j\neq 0=-k}^{k}(1+\psi_{2k+1})(1-\psi_{2k+2}) \\
& \times\left(-\frac{\prod_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{j,2k+1})\prod_{i\neq 0=-k-1}^{k}(L_{i,2k+2}-L_{j,2k+1})}{\prod_{i\neq 0,j;i=-k}^{k}(L_{i,2k+1}-L_{j,2k+1}-1)(L_{i,2k+1}-L_{j,2k+1})}\right)^{1/2} \\
& \times \frac{\prod_{i\neq 0,j;i=-k}^{k}(L_{0,2k+3}-L_{i,2k+1})\prod_{i\neq 0=-k-1}^{k+1}(L_{0,2k+3}-L_{i,2k+3})}{\prod_{i\neq 0=-k-1}^{k}(L_{0,2k+3}-L_{i,2k+2}-1)(L_{0,2k+3}-L_{i,2k+2})}|M)^{-(j,2k+1)}_{-(0,2k+2)} \\
& -\sum_{j\neq 0=-k-1}^{k}\psi_{2k}\psi_{2k+1} \\
& \times\left(-\frac{\prod_{i\neq 0=-k}^{k}(L_{i,2k+1}-L_{j,2k+2}-1)\prod_{i\neq 0=-k-1}^{k+1}(L_{i,2k+3}-L_{j,2k+2}-1)}{\prod_{i\neq 0,j;i=-k-1}^{k}(L_{i,2k+2}-L_{j,2k+2}-1)(L_{i,2k+2}-L_{j,2k+2})}\right)^{1/2} \\
& \times \frac{\prod_{i\neq 0,j;i=-k-1}^{k}(L_{0,2k+2}-L_{i,2k+2})\prod_{i\neq 0=-k}^{k-1}(L_{0,2k+2}-L_{i,2k})}{\prod_{i\neq 0=-k}^{k}(L_{0,2k+2}-L_{i,2k+1})(L_{0,2k+2}-L_{i,2k+1}+1)}|M)^{-(0,2k+1)}_{-(j,2k+2)} \\
& +\sum_{l\neq 0=-k-1}^{k}\sum_{j\neq 0=-k}^{k}Q(j,l) \\
& \times\left(-\frac{\prod_{i\neq 0,j;i=-k}^{k}(L_{i,2k+1}-L_{l,2k+2}-1)\prod_{i\neq 0=-k-1}^{k+1}(L_{i,2k+3}-L_{l,2k+2}-1)}{\prod_{i\neq 0,l;i=-k-1}^{k}(L_{i,2k+2}-L_{l,2k+2}-1)(L_{i,2k+2}-L_{l,2k+2})}\right)^{1/2} \\
& \times\left(\frac{\prod_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{j,2k+1})\prod_{i\neq 0,l;i=-k-1}^{k}(L_{i,2k+2}-L_{j,2k+1})}{\prod_{i\neq 0,j;i=-k}^{k}(L_{i,2k+1}-L_{j,2k+1}-1)(L_{i,2k+1}-L_{j,2k+1})}\right)^{1/2} \\
& \times|M)^{-(l,2k+2)}_{-(j,2k+1)}, \quad k\in[1;\infty],
\end{aligned}
\tag{123}
$$

$$
\begin{aligned}
E_{-k,-k+1}|M) = & -(-1)^{\psi_{2k-2}}(1+\psi_{2k-3})\psi_{2k-1} \\
& \times \frac{\prod_{i\neq 0=-k+2}^{k-2}(L_{0,2k-1}-L_{i,2k-3}-\psi_{2k-2})\prod_{i\neq 0=-k+1}^{k-1}(L_{0,2k-1}-L_{i,2k-1}-\psi_{2k-2})}{\prod_{i\neq 0=-k+1}^{k-2}(L_{0,2k-1}-L_{i,2k-2}-1)(L_{0,2k-1}-L_{i,2k-2}-2\psi_{2k-2})} \\
& \times \frac{\prod_{i\neq 0=-k+1}^{k-2}(L_{0,2k}-L_{i,2k-2})\prod_{i\neq 0=-k}^{k-1}(L_{0,2k}-L_{i,2k})}{\prod_{i\neq 0=-k+1}^{k-1}(L_{0,2k}-L_{i,2k-1})(L_{0,2k}-L_{i,2k-1}+1)}|M)^{-(0,2k-2)}_{-(0,2k-1)} \\
& +\sum_{j\neq 0=-k+1}^{k-2}\psi_{2k-2}\psi_{2k-1} \\
& \times\left(-\frac{\prod_{i\neq 0=-k+2}^{k-2}(L_{i,2k-3}-L_{j,2k-2}-1)\prod_{i\neq 0=-k+1}^{k-1}(L_{i,2k-1}-L_{j,2k-2}-1)}{\prod_{i\neq 0,j;i=-k+1}^{k-2}(L_{i,2k-2}-L_{j,2k-2}-1)(L_{i,2k-2}-L_{j,2k-2})}\right)^{1/2} \\
& \times \frac{\prod_{i\neq 0,j;i=-k+1}^{k-2}(L_{0,2k}-L_{i,2k-2})\prod_{i\neq 0=-k}^{k-1}(L_{0,2k}-L_{i,2k})}{\prod_{i\neq 0=-k+1}^{k-1}(L_{0,2k}-L_{i,2k-1})(L_{0,2k}-L_{i,2k-1}+1)}|M)^{-(j,2k-2)}_{-(0,2k-1)}
\end{aligned}
$$

$$+\sum_{j\neq 0=-k+1}^{k-1}(1+\psi_{2k-3})(1-\psi_{2k-2})$$
$$\times\left(-\frac{\Pi_{i\neq 0=-k+1}^{k-2}(L_{i,2k-2}-L_{j,2k-1})\Pi_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{j,2k-1})}{\Pi_{i\neq 0,j;i=-k+1}^{k-1}(L_{i,2k-1}-L_{j,2k-1}-1)(L_{i,2k-1}-L_{j,2k-1})}\right)^{1/2}$$
$$\times\frac{\Pi_{i\neq 0,j;i=-k+1}^{k-1}(L_{0,2k-1}-L_{i,2k-1})\Pi_{i\neq 0=-k+2}^{k-2}(L_{0,2k-1}-L_{i,2k-3})}{\Pi_{i\neq 0=-k+1}^{k-2}(L_{0,2k-1}-L_{i,2k-2}-1)(L_{0,2k-1}-L_{i,2k-2})}|M)^{-(0,2k-2)}_{-(j,2k-1)}$$
$$+\sum_{l\neq 0=-k+1}^{k-1}\sum_{j\neq 0=-k+1}^{k-2}P(j,l)$$
$$\times\left(-\frac{\Pi_{i\neq 0,j;i=-k+1}^{k-2}(L_{i,2k-2}-L_{l,2k-1})\Pi_{i\neq 0=-k}^{k-1}(L_{i,2k}-L_{l,2k-1})}{\Pi_{i\neq 0,l;i=-k+1}^{k-1}(L_{i,2k-1}-L_{l,2k-1}-1)(L_{i,2k-1}-L_{l,2k-1})}\right)^{1/2}$$
$$\times\left(\frac{\Pi_{i\neq 0=-k+2}^{k-2}(L_{i,2k-3}-L_{j,2k-2}-1)\Pi_{i\neq 0,l;i=-k+1}^{k-1}(L_{i,2k-1}-L_{j,2k-2}-1)}{\Pi_{i\neq 0,j;i=-k+1}^{k-2}(L_{i,2k-2}-L_{j,2k-2}-1)(L_{i,2k-2}-L_{j,2k-2})}\right)^{1/2}$$
$$\times|M)^{-(j,2k-2)}_{-(l,2k-1)},\quad k\in[2;\infty]. \tag{124}$$

The above transformation relations (116)–(124) were derived first for $gl(n|1|n)$ from (68) to (74) and the supercommutation relations. Therefore they give a representation of $gl(n|1|n)$ for any n. An essential requirement, when passing to $n\to\infty$, is given with the condition (113). It is straightforward to check that $V([M])$ is invariant under the action of the generators. The rest of the proof, which we skip, is rather similar to that of Proposition 6, although technically it is more involved.

IV. CONCLUDING REMARKS

We have constructed two classes of highest weight irreps of the infinite-dimensional Lie superalgebra $gl(1|\infty)$. It should be noted that the GZ representations are inequivalent to the C representations. More than that: the C representations, being highest weight irreps of $gl(\infty|1|\infty)$, are not highest weight representations of $gl_0(1|\infty)$ and vice versa. Indeed, assume that the $gl_0(1|\infty)$ module $V([m])$ is also a highest weight $gl(\infty|1|\infty)$ module with a highest weight vector y. Then y has to be a highest weight vector of any of the subalgebras $gl(k|1|k-1+\theta)$. Hence Eqs. (54) and (55) have to hold for any $\theta=0,1,k\in[1-\theta;\infty]$. Therefore $y\notin V([m])$ [see (87)].

Our primary interest in the present investigation is related to its eventual applications in a generalization of the statistics in quantum field theory. From this point of view our results are, however, very preliminary. The first observation in this respect is that the algebra (for definiteness) $gl(\infty|1|\infty)$ is not large enough. It does not contain important physical observables [like the energy-momentum of the field P^m, see (8)], which are infinite linear combinations of the generators of $gl(\infty|1|\infty)$. In order to incorporate them one has to go to the completed central extension $a(\infty|1|\infty)$ of $gl(\infty|1|\infty)$ in a way similar to that for the Lie algebra gl_∞ (Ref. 41) or the Lie superalgebra $gl_{\infty|\infty}$.[20] This is only the first step. The next one will be to determine those $gl(\infty|1|\infty)$ modules $V([M])$, which can be extended to $a(\infty|1|\infty)$ modules.

The most important and perhaps the most difficult step will be to express the transformations of the $gl(\infty|1|\infty)$ modules in terms of natural for the QFT variables, namely via the creation and the annihilation operators $a_i^\pm$ of $gl(\infty|1|\infty)$, which are just its odd generators.[4] This is, however, not simple and, may not even be necessary in the general context of the representation theory. The physical state spaces, the Fock spaces, have to satisfy several additional physical requirements.[42] In particular any such space has to be generated from the vacuum (the highest weight vector) by polynomials of the creation operators, which are only a part of the negative root vectors. This imposes considerable restriction on the physically admissible modules. Hence in the applications one has to select first the Fock spaces from all $gl(\infty|1|\infty)$ modules and then study their transformation properties under the action of the physically relevant operators, in particular of the CAOs.

An additional problem is related to the circumstance that in QFT the indices of the CAOs are not elements form a countable set. Therefore as a test model one can try to consider first the $gl(\infty|1|\infty)$ statistics in the frame of a lattice quantum field theory or locking the field in a finite volume.

ACKNOWLEDGMENTS

N.I.S. is grateful to Professor M. D. Gould for the invitation to work in his group at the Department of Mathematics in University of Queensland. T.D.P. is thankful to Professor S. Okubo for the kind invitation to conduct research under the Fulbright Program in the Department of Physics and Astronomy, University of Rochester. We wish to thank Professor Randjbar-Daemi for the kind hospitality at the High Energy Section of ICTP. This work was supported by the Australian Research Council, by the Fulbright Program of U.S.A., Grant No. 21857, and by the Contract No. Φ-416 of the Bulgarian Foundation for Scientific Research.

[1] T. D. Palev, J. Math. Phys. **23**, 1778 (1982).
[2] T. D. Palev, Czech. J. Phys., Sect. B **32**, 680 (1982).
[3] T. D. Palev, Czech. J. Phys., Sect. B **29**, 91 (1979).
[4] T. D. Palev, Communication No. JINR E2-11942, 1978.
[5] A. Ch. Ganchev and T. D. Palev, J. Math. Phys. **21**, 797 (1980).
[6] H. S. Green, Phys. Rev. **90**, 270 (1953).
[7] N. N. Bogoljubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Nauka, Moscow, 1957); english ed. (Interscience, New York, 1959).
[8] E. Celeghini, T. D. Palev, and M. Tarlini, Mod. Phys. Lett. B **5**, 187 (1991).
[9] T. D. Palev, J. Phys. A **26**, L1111 (1993); and hep-th/9306016.
[10] L. K. Hadjiivanov, J. Math. Phys. **34**, 5476 (1993).
[11] T. D. Palev and J. Van der Jeugt, J. Phys. A **28**, 2605 (1995); and q-alg/9501020.
[12] T. D. Palev, Commun. Math. Phys. **196**, 429 (1998); q-alg/9709003.
[13] V. G. Kac, Lect. Notes Math. **676**, 597 (1979).
[14] S. Okubo, J. Math. Phys. **35**, 2785 (1994).
[15] J. Van der Jeugt, *New Trends in Quantum Field Theory* (Heron, Sofia, 1996).
[16] S. Meljanac, M. Milekovic, and M. Stojic, q-alg/9712017.
[17] T. D. Palev and N. I. Stoilova, J. Math. Phys. **38**, 2506 (1997); hep-th/9606011.
[18] F. D. M. Haldane, Phys. Rev. Lett. **67**, 937 (1991).
[19] T. D. Palev and N. I. Stoilova, J. Phys. A **27**, 977 (1994); **27**, 7387 (1994); hep-th/9307102, hep-th/9405125.
[20] V. G. Kac and J. W. van der Leur, Ann. Inst. Fourier, Grenoble **37**, 99 (1987).
[21] V. G. Kac and J. W. van der Leur, Adv. Series Math. Phys. **7**, 369 (1988).
[22] T. D. Palev, J. Math. Phys. **22**, 2127 (1981).
[23] J. Van der Jeugt, J. W. B. Hughes, R. C. King, and J. Thierry-Mieg, Commun. Algebra **18**, 3453 (1990).
[24] H. Schlosser, Seminar Sophus Lie **3**, 15, TH Darmstadt (1993).
[25] H. Schlosser, Contribut. Algebra Geometry **35**, 193 (1994).
[26] T. D. Palev, Funkt. Anal. Prilozh. **21**, 85 (1987); Funct. Anal. Appl. **21**, 245 (1987) (English translation).
[27] T. D. Palev, J. Math. Phys. **30**, 1433 (1989).
[28] T. D. Palev and V. N. Tolstoy, Commun. Math. Phys. **141**, 549 (1991).
[29] J. Van der Jeugt, J. W. B. Hughes, R. C. King, and J. Thierry-Mieg, J. Math. Phys. **31**, 2278 (1990).
[30] J. W. B. Hughes, R. C. King, and J. Van der Jeugt, J. Math. Phys. **33**, 470 (1992).
[31] T. D. Palev, Funkt. Anal. Prilozh. **23**, 69 (1989); Funct. Anal. Appl. **23**, 141 (1989) (English translation).
[32] J. Van der Jeugt, J. Math. Phys. **36**, 605 (1995).
[33] T. D. Palev, N. I. Stoilova, and J. Van der Jeugt, Commun. Math. Phys. **166**, 367 (1994).
[34] T. D. Palev, Funkt. Anal. Prilozh. **24**, 69 (1990); Funct. Anal. Appl. **24**, 72 (1990) (English translation).
[35] T. D. Palev, J. Math. Phys. **31**, 579 (1990); **31**, 1078 (1990).
[36] V. V. Serganova, Math. USSR Izv. **24**, 359 (1985).
[37] D. A. Leites, M. V. Saveliev, and V. V. Serganova (unpublished).
[38] J. W. Van der Leur, Utrecht, thesis, 1985.
[39] I. Penkov and V. Serganova, Indag. Math. **3**, 419 (1992).
[40] V. G. Kac and M. Wakimoto, Progress Math. **123**, 415 (1994).
[41] V. G. Kac and V. G. Peterson, Proc. Natl. Acad. Sci. USA **78**, 3308 (1981).
[42] T. D. Palev, J. Math. Phys. **21**, 1293 (1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Generalized Lamé functions. I. The elliptic case

S. N. M. Ruijsenaars
Centre for Mathematics and Computer Science, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

(Received 24 April 1998; accepted for publication 17 September 1998)

We present and study a class of functions associated with the two-particle quantum relativistic Calogero–Moser system with elliptic interactions. The functions may be viewed as joint eigenfunctions of two independent commuting analytic difference operators, one of which is the defining quantum dynamics; The second one is obtained by interchanging the step size and the imaginary period. The functions depend on parameters that are dense in the natural parameter domain. In essence, they consist of products of Weierstrass σ-functions and plane waves. The zeros of the σ-functions satisfy a constraint system encoding both Schrödinger equations at once. © *1999 American Institute of Physics.* [S0022-2488(99)02402-0]

I. INTRODUCTION

This paper is concerned with eigenfunctions of an analytic difference operator generalizing the Lamé differential operator,

$$H_{\rm nr}=-\frac{d^2}{dx^2}+g(g-1)\wp(x), \tag{1.1}$$

where $\wp$ is the Weierstrass $\wp$-function. The pertinent analytic difference operator (henceforth AΔO) can be taken to be

$$H_{\rm rel}=\left(\frac{\sigma(x-i\beta g)}{\sigma(x)}\right)^{1/2}T_{i\beta}\left(\frac{\sigma(x+i\beta g)}{\sigma(x)}\right)^{1/2}+(\beta\to-\beta), \tag{1.2}$$

where σ is the Weierstrass σ-function. Here, the shift operator T_α is defined by

$$(T_\alpha f)(x)=f(x-\alpha),\quad \alpha\in\mathbb{C}, \tag{1.3}$$

so that one has

$$H_{\rm rel}=2+\beta^2H_{\rm nr}+O(\beta^4),\quad \beta\to 0. \tag{1.4}$$

The subscripts ''nr'' and ''rel'' in these formulas stand for ''nonrelativistic'' and ''relativistic.'' Indeed, the parameter β in the AΔO (1.2) may be viewed as $1/c$, with c the speed of light. Thus, (1.4) encodes the nonrelativistic limit $c\to\infty$; cf. Ref. 1.

In our survey paper, Ref. 2, and lecture notes, Ref. 3, we have announced and described $H_{\rm rel}$ eigenfunctions for integer coupling g. These functions generalize the $H_{\rm nr}$ eigenfunctions for integer g in the form presented on pp. 572–574 of Ref. 4. In this paper we shall not only elaborate on the $g=2,3,\ldots$ $H_{\rm rel}$ eigenfunctions from Subsection 6.3 in Ref. 3, but also obtain eigenfunctions for a dense set in the relevant parameter space. As we will detail, these functions are, in fact, *joint* eigenfunctions of *three* commuting independent AΔOs—a feature that generalizes symmetry properties of the hyperbolic specialization described in Subsection 6.3 of Ref. 3. (In the elliptic case, however, we were unable to find useful ''dual operators''—operators D acting on the spectral variable in the eigenfunctions in such a way that the latter are also D-eigenfunctions with x-dependent eigenvalues.)

In this paper we shall not dwell on the integrable system context in which the above operators arise. This setting is discussed at length in Refs. 2, 3, and is not necessary for understanding the following. On the other hand, the ancestry of the operators at issue makes itself felt in the aspects we emphasize: We are principally interested in real eigenvalues, and, more generally, in those features of the eigenfunctions that are important in promoting them to kernels of unitary operators that serve to redefine the AΔOs involved as bonafide self-adjoint Hilbert space operators. In this connection we recall that AΔOs have highly nonunique eigenfunctions (compared to differential and discrete difference operators), so that quite novel problems and features arise in their rigorous definition as quantum dynamics.

Next, we mention that a close relative S_0 of the AΔO $H_{\rm rel}$ was already introduced by Sklyanin,[5,6] together with AΔOs of a similar form. He obtained finite-dimensional representation spaces for the resulting AΔO algebra, spanned by very special eigenfunctions of S_0. General integer-g eigenfunctions of S_0 were recently presented in a related context by Krichever and Zabrodin.[7] (Roughly speaking, Sklyanin's functions correspond to eigenfunctions at the band edges in the finite-gap integration picture expounded in Ref. 7—a viewpoint that is far removed from our concerns in this paper.) For $g>2$ these functions have a rather different appearance from the eigenfunctions already detailed in Ref. 3.

Subsequent to Refs. 3 and 7, Felder and Varchenko obtained integer-g eigenfunctions in a form substantially equivalent to ours. They arrived at these eigenfunctions via their comprehensive study of representations of elliptic quantum groups, tying them in with the ''algebraic Bethe Ansatz'' of the Russian school, and with Baxter's work on the XYZ model. Their work—inasmuch as it concerns the operator $H_{\rm rel}$ and its eigenfunctions[8,9]—has a quite different perspective, emphasizing representation theoretic and algebro-geometric features. (See also a recent paper by Hasegawa[10] for yet another approach.)

Before summarizing our results, we would like to mention three forthcoming papers that are closely related to the present one. First, we point out that hyperbolic and trigonometric specializations are studied in a sequel to this paper.[11] In the latter regimes we can proceed much further, since a second, far more explicit representation of the relevant eigenfunctions exists. The results obtained in these special contexts also illuminate various issues pertaining to the elliptic regime, to which we restrict attention in this paper.

Second, the simplest nontrivial parameter choice $g=2$ will be reconsidered elsewhere.[12] This case admits an in-depth treatment that is independent of (and considerably simpler than) the present paper and its sequel. Moreover, as a striking feature of this special case we demonstrate that in a certain scaling limit its eigenfunctions give rise to the well-known eigenfunctions of the quantized nonlinear Schrödinger equation (alias the delta-function gas).

Third, our forthcoming conference contribution Ref. 13 reviews our findings regarding generalized Lamé functions and their specializations.

In order to sketch the results of the present paper, it is expedient to trade the Weierstrass σ-function $\sigma(z;\omega,\omega')$ for the function

$$s(r,a;z)\equiv\sigma\left(z;\frac{\pi}{2r},\frac{ia}{2}\right)\exp(-\eta z^2 r/\pi). \tag{1.5}$$

(Here and below, we use the elliptic function notation of Whittaker and Watson;[4] we also use some of the elliptic function lore collected in this reference.) The function $s(z)$ is an entire odd function with simple zeros in the lattice points $\mathbb{Z}\pi/r+i\mathbb{Z}a$. It is π/r-antiperiodic and obeys the analytic difference equation (henceforth AΔE)

$$\frac{s(z+ia/2)}{s(z-ia/2)}=-\exp(-2irz). \tag{1.6}$$

Moreover, it satisfies

$$\lim_{a\to\infty} s(r,a;z) = \frac{\sin rz}{r} \quad \text{(uniformly on compacts)}, \tag{1.7}$$

$$\lim_{r\to 0} s(r,a;z) = \frac{\sinh \pi z/a}{\pi/a} \quad \text{(uniformly on compacts)}, \tag{1.8}$$

and the scaling relation

$$s(r/\lambda,\lambda a;\lambda z) = \lambda s(r,a;z). \tag{1.9}$$

For later use we also note that iteration of the AΔE (1.6) yields

$$\frac{s(r,a;z+iLa)}{s(r,a;z)} = (-)^L \exp(arL^2 - 2irLz), \quad L \in \mathbb{Z}. \tag{1.10}$$

As a matter of fact, we have occasion to use two s-functions,

$$s_\delta(z) \equiv s(r,a_\delta;z), \quad \delta = +,-. \tag{1.11}$$

This is because the functions we define and study are actually joint eigenfunctions of the two AΔOs

$$H_\delta \equiv e^{-br}\left(\frac{s_\delta(x-ib)}{s_\delta(x)}\right)^{1/2} T_{ia_{-\delta}}\left(\frac{s_\delta(x+ib)}{s_\delta(x)}\right)^{1/2} + (i\to -i), \quad \delta = +,-. \tag{1.12}$$

In view of (1.5), each of these may be regarded as a multiple of the AΔO $H_{\rm rel}$ (1.2) when one sets $b = a_{-\delta}g$. The constant up front is chosen such that we have the symmetry property

$$H_\delta(a_+,a_-,b) = H_\delta(a_+,a_-,a_+ + a_- - b). \tag{1.13}$$

[Use (1.6) to verify this.] Here and below, it is understood that the parameters belong to the elliptic parameter domain

$$\mathcal{E} \equiv \{(r,a_+,a_-,b) \mid r,a_+,a_- > 0, b \in \mathbb{R}\}. \tag{1.14}$$

We begin by transforming H_δ to the form

$$A_\delta = e^{-br}\frac{s_\delta(x-ib)}{s_\delta(x)} T_{ia_{-\delta}} + (i\to -i), \quad \delta = +,-, \tag{1.15}$$

where

$$A_\delta \equiv w(x)^{-1/2} H_\delta w(x)^{1/2}. \tag{1.16}$$

The weight function $w(r,a_+,a_-,b;x)$ occurring here was introduced and studied in Ref. 14. It is a meromorphic solution to the two AΔEs

$$\frac{w(x+ia_\delta/2)}{w(x-ia_\delta/2)} = \frac{s_{-\delta}(x+ib-ia_\delta/2)}{s_{-\delta}(x-ib+ia_\delta/2)} \cdot \frac{s_{-\delta}(x+ia_\delta/2)}{s_{-\delta}(x-ia_\delta/2)}, \quad \delta = +,-, \tag{1.17}$$

which is why (1.16) entails (1.15).

The point of the similarity transformation (1.16) is that the AΔOs A_δ (1.15) have meromorphic coefficients. Thus, we may and will view them first as linear operators leaving the vector space

$$\mathcal{M} \equiv \{F(x) \mid F \ \text{meromorphic}\} \tag{1.18}$$

invariant. (We shall discuss Hilbert space aspects shortly.) In view of the AΔE (1.6), the maps $A_+, A_-: \mathcal{M} \to \mathcal{M}$ commute. Now at this point it should be emphasized that there are a great many functionally independent AΔOs commuting with A_+. (For instance, when one multiplies the two coefficients of A_- by distinct meromorphic functions with period ia_-, one obtains an AΔO that also commutes with A_+.) However, we are not aware of any general arguments guaranteeing the existence of nontrivial joint eigenspaces for A_+ and an independent AΔO in its commutant.

Even so, we have found two linearly independent joint eigenfunctions $\Psi(\pm x, y)$ of the $(a_+ \leftrightarrow a_-)$-symmetric pair (A_+, A_-), provided the parameters belong to a dense subset $\mathcal{D}$ of the parameter space $\mathcal{E}$ (1.14). (Note both AΔOs commute with parity.) The spectral variable y takes values in an interval (K, ∞), where K depends on the parameters. The eigenvalues E_+ and E_- are real-valued, real-analytic functions, satisfying

$$E_\delta(y) \sim \exp(a_{-\delta} y), \quad E'_\delta(y) \sim a_{-\delta} \exp(a_{-\delta} y), \quad \delta = +, -, \quad y \to \infty, \tag{1.19}$$

and separating points on (K, ∞):

$$K < y_1 < y_2 \Rightarrow (E_+(y_1), E_-(y_1)) \neq (E_+(y_2), E_-(y_2)). \tag{1.20}$$

The dense subset $\mathcal{D}$ is defined by (3.33)–(3.35) below. For expository simplicity, however, we shall summarize our results for a subset of $\mathcal{D}$, namely,

$$\mathcal{D}_{\text{irr}} \equiv \{(r, a_+, a_-, (N_+ + 1)a_+ - N_- a_-) \in \mathcal{E} | a_+/a_- \notin \mathbb{Q}, N_+, N_- \in \mathbb{N}\}. \tag{1.21}$$

Since the b-values allowed here are dense in $\mathbb{R}$ for a_+/a_- irrational, $\mathcal{D}_{\text{irr}}$ is already dense in $\mathcal{E}$.

Fixing $(r, a_+, a_-, b) \in \mathcal{D}_{\text{irr}}$, any joint eigenfunction of A_+ and A_- with eigenvalues $E_+(y)$ and $E_-(y)$, resp., is a linear combination of $\Psi(x, y)$ and $\Psi(-x, y)$. [More precisely, we prove that this holds true for all sufficiently large y; cf. Appendix B.] The latter are explicitly given by

$$\Psi(x, y) = \mathcal{N} \prod_{j=-N_+}^{N_+} \frac{1}{s_-(x + ija_+)} \cdot \prod_{\delta=+,-} \prod_{j=1}^{N_\delta} s_{-\delta}(x + z_j^\delta(y))$$
$$\times \exp[irx(2N_+N_- + N_+ + N_- + 1) + ixy], \quad y \in (K, \infty). \tag{1.22}$$

The normalizing factor $\mathcal{N}$ depends on the parameters, but not on x and y. The ''zero functions'' $z_1^+, \ldots, z_{N_+}^+$ and $z_1^-, \ldots, z_{N_-}^-$ are functions from (K, ∞) to $i(0, \infty)$ that are real-analytic and such that

$$\lim_{y \to \infty} z_j^\delta(y) = ija_\delta, \quad j = 1, \ldots, N_\delta, \quad \delta = +, -. \tag{1.23}$$

These functions are determined as solutions to a certain constraint system. This system depends on the parameters in a quite complicated fashion, and for brevity we do not describe it here. [It is given by (3.5), (3.10), and (3.11); cf. also Appendix A.]

In view of (1.23), the function $\Psi(x, y)$ has asymptotics

$$\Psi(x, y) \sim c(x) \exp(ixy), \quad y \to \infty. \tag{1.24}$$

Here, the c-function reads

$$c(r, a_+, a_-, (N_+ + 1)a_+ - N_- a_-; x) = \mathcal{N} \frac{\prod_{k=1}^{N_-} s_+(x + ika_-)}{\prod_{j=0}^{N_+} s_-(x - ija_+)} \exp irx(2N_+N_- + N_+ + N_- + 1). \tag{1.25}$$

It is not obvious, but true, that the normalization constant $\mathcal{N}$ can be chosen such that

$$c(r,a_+,a_-,b;x)=\frac{G(r,a_+,a_-;x-ib+i(a_++a_-)/2)}{G(r,a_+,a_-;x+i(a_++a_-)/2)}, \tag{1.26}$$

for parameters in $\mathcal{E}$ (1.14). Here, the function $G(r,a_+,a_-;z)$ is the generalized elliptic gamma function from Ref. 14, which is meromorphic in r, a_+, a_-, and z as long as a_+r and a_-r stay in the right half plane. The weight and scattering functions introduced in Ref. 14 can be written as

$$w(x)=\frac{1}{c(x)c(-x)}, \tag{1.27}$$

$$u(x)=-\exp(-2irx)\frac{c(x)}{c(-x)}. \tag{1.28}$$

[This easily follows from their definition in terms of the G-function; to check (1.28) one also needs the defining AΔEs of the latter; cf. Proposition III.8 in Ref. 14.] Thus, the joint eigenfunction

$$\mathcal{F}(x,y)=w(x)^{1/2}\Psi(x,y) \tag{1.29}$$

of the AΔOs H_+ and H_- has plane wave asymptotics

$$\mathcal{F}(x,y)\sim w(x)^{1/2}c(x)\exp(ixy)=[-\exp(2irx)u(x)]^{1/2}\exp(ixy), \quad y\to\infty. \tag{1.30}$$

The obvious question arising from these results is now the following: Are there meromorphic joint eigenfunctions $\Psi(x,y)$ for *arbitrary* parameters in $\mathcal{E}$ (1.14) that depend continuously on the parameters and are proportional to $\Psi(x,y)$ (1.22) for parameters in $\mathcal{D}_{\mathrm{irr}}$? (Since $\mathcal{D}_{\mathrm{irr}}$ is dense in $\mathcal{E}$, such an interpolation is unique up to scale factors depending on y and the parameters.) The point is that the same question has an affirmative answer for the c-function, as we have just seen.

The answer to the joint eigenfunction question, however, may well be ''No.'' To see why, one need only note that when a sequence $(r,a_{+,n},a_{-,n},b_n)\in\mathcal{D}_{\mathrm{irr}}$ converges to a point in $\mathcal{E}$, then the integers $N_{\delta,n}$ [cf. (1.21)] typically go to ∞. Thus, the poles of $\Psi(x,y)$ due to the first product in (1.22) become *dense* on the lines $\mathrm{Re}\,x=k\pi/r, k\in\mathbb{Z}$.

From this perspective our next result is quite surprising. To state it, we introduce the even function,

$$\chi(x,y)\equiv\Psi(x,y)+\Psi(-x,y). \tag{1.31}$$

Now consider a rectangle $|\mathrm{Re}\,x|<\pi/r, |\mathrm{Im}\,x|<L$. Fixing a_+, a_-, and a compact b-interval I, the number of poles of $\Psi(x,y)$ in the rectangle can be made arbitrarily large by choosing suitable $b\in I$; cf. the previous paragraph. By contrast, the number of poles of $\chi(x,y)$ in the rectangle is bounded above by a finite number that depends only on L and I!

To explain why this is true, we write

$$\chi(x,y)=\mathcal{N}\prod_{j=-N_+}^{N_+}\frac{1}{s_-(x+ija_+)}[\mathcal{H}(x,y)-\mathcal{H}(-x,y)], \tag{1.32}$$

where $\mathcal{H}$ is the holomorphic function

$$\mathcal{H}(x,y)\equiv\prod_{\delta=+,-}\prod_{j=1}^{N_\delta}s_{-\delta}(x+z_j^\delta(y))\cdot\exp[irx(2N_+N_-+N_++N_-+1)+ixy]. \tag{1.33}$$

The crux is now that one has the identities

$$\mathcal{H}(ik_+a_+ + ik_-a_-, y) = \mathcal{H}(-ik_+a_+ - ik_-a_-, y), \quad k_\delta \in \{-N_\delta, \ldots, 0, \ldots, N_\delta\}, \quad \delta = +, -. \tag{1.34}$$

Thus, poles of $\chi(x,y)$ on the imaginary axis can only occur for

$$x = ik_+a_+ + ik_-a_-, \quad k_+ \in \{-N_+, \ldots, 0, \ldots, N_+\}, \quad |k_-| > N_-, \tag{1.35}$$

and so the assertion in the previous paragraph readily follows.

Now it is clear from (1.22) that we have

$$\Psi(x \pm \pi/r, y) = \exp(\pm i\pi y/r)\Psi(x,y). \tag{1.36}$$

Thus, $\Psi(x,y)$ is π/r-periodic or π/r-antiperiodic when y/r is an integer. Defining

$$\chi_n(x) \equiv \chi(x, nr), \quad nr > K, \quad n \in \mathbb{N}, \tag{1.37}$$

we therefore obtain a function that is π/r-periodic/antiperiodic for n even/odd.

As a consequence, there is no apparent obstruction to the existence of a meromorphic interpolation for the functions $\chi_n(x)$. However, if an explicit representation for an interpolation exists, it is most likely vastly different from (1.22). Indeed, this is the case for the hyperbolic specialization, where we have found an interpolation in terms of a natural generalization of Euler's hypergeometric function; cf. Ref. 3 and papers to appear.

Before turning to orthogonality issues, we point out a consequence of the quasi-periodicity relations (1.36) that is of interest in itself, and that will be invoked later on. Recalling (1.3), we can rephrase (1.36) by saying that $\Psi(x,y)$ is an eigenfunction of the AΔOs $T_{\pm\pi/r}$. Now this is true for $\Psi(-x,y)$, too, but then we obtain different eigenvalues. On the other hand, when we introduce the ''quasi-periodicity AΔO,''

$$Q \equiv T_{\pi/r} + T_{-\pi/r}, \tag{1.38}$$

then we obtain

$$(Q\Psi)(\pm x, y) = E_Q \Psi(\pm x, y), \quad E_Q \equiv 2\cos(\pi y/r). \tag{1.39}$$

Thus, as already mentioned above, the functions $\Psi(\pm x,y)$ are, in fact, joint eigenfunctions of three commuting independent AΔOs A_+, A_-, and Q.

Next, we describe results concerning orthogonality of the functions

$$\Phi_n(x) \equiv w(x)^{1/2}\chi_n(x), \quad nr > K, \quad n \in \mathbb{N}, \quad x \in (0, \pi/r), \tag{1.40}$$

in the Hilbert space $L^2((0,\pi/r),dx)$. The weight function $w(x)$ is given by

$$w(r, a_+, a_-, (N_+ + 1)a_+ - N_-a_-; x) = (-)^{N_+ + N_- + 1}\mathcal{N}^{-2} s_-(x)^2 \frac{\prod_{\pm j=1}^{N_+} s_-(x - ija_+)}{\prod_{\pm k=1}^{N_-} s_+(x - ika_-)}; \tag{1.41}$$

cf. (1.25) and (1.27). It is non-negative for real x, and we take the positive square root in (1.40). Then the function $w(x)^{1/2}, x \in (0,\pi/r)$, has an analytic continuation to an odd, π/r-antiperiodic function that has no singularities for real x. The factor $s_-(x)$ in the latter function cancels the factor $1/s_-(x)$ in (1.22), and the remaining poles do not meet the real axis (since a_+/a_- is irrational). Hence $\Phi_n(x)$ (1.40) extends to an odd function without singularities for real x, which is π/r-periodic/-antiperiodic for n odd/even.

As a consequence, the functions $\Phi_n(x)$ are square-integrable on $(0,\pi/r)$. One of the principal results of this paper is now that these functions are *pairwise orthogonal*, provided the parameters belong to the region

$$\mathcal{C} \equiv \{(r,a_+,a_-,b) \in \mathcal{E} | 0 < b < a_+ + a_-\}. \tag{1.42}$$

To provide more perspective on this parameter restriction, we mention that $\mathcal{C}$ coincides with the parameter region for which the u-function has winding number 0 as x goes from 0 to π/r. Correspondingly, its logarithm has a rapidly convergent Fourier series. To be specific, we have [cf. Ref. 14 (4.87)]

$$u(r,a_+,a_-,b;x;.) = \exp\left(2i \sum_{j=1}^{\infty} \frac{\sinh(a_+ - b)nr \sinh(a_- - b)nr}{n \sinh a_+ nr \sinh a_- nr} \sin 2nrx \right). \tag{1.43}$$

For convergence of the series one needs $|\mathrm{Im}\, x| < d/2$, with

$$d \equiv a_+ + a_- - |a_+ - b| - |a_- - b|. \tag{1.44}$$

Thus, one gets $d > 0$ iff the parameters belong to $\mathcal{C}$ (1.42). [Note that (1.43) exhibits symmetry under $a_+ \leftrightarrow a_-$ and $b \to a_+ + a_- - b$.]

Provided the parameters belong to $\mathcal{C}$, the w-function admits a similar representation. It is expedient to write, first of all,

$$w(r,a_+,a_-,b;x) = C^2 s_-(x) s_+(x) w_r(x). \tag{1.45}$$

Here we have introduced the reduced weight function w_r, and the positive constant C reads

$$C(r,a_+,a_-) = 2r \prod_{k=1}^{\infty} [1 - \exp(-2kra_+)][a_+ \to a_-]; \tag{1.46}$$

cf. Ref. 14 (5.41). Then one has, from Ref. 14 (5.54),

$$w_r(r,a_+,a_-,b;x) = \exp\left(\sum_{j=1}^{\infty} \frac{\sinh(a_+ + a_- - 2b)nr}{n \sinh a_+ nr \sinh a_- nr} \cos 2nrx \right). \tag{1.47}$$

Just as for (1.43), the series converges in a strip containing the real axis iff the parameters belong to $\mathcal{C}$. (Note the strip width is larger than for the u-function when $a_+ < b < a_-$, say. Note also that w is symmetric under $a_+ \leftrightarrow a_-$, but not under $b \to a_+ + a_- - b$.)

Returning now to the Hilbert space $L^2((0,\pi/r),dx)$, we continue by pointing out that the restriction to $\mathcal{C}$ appears to be essential: For parameters outside $\mathcal{C}$, orthogonality is most likely violated, in general. Whenever this is the case, there exists no reinterpretation of the AΔOs H_δ (1.12) as symmetric Hilbert space operators whose domains include the eigenfunctions Φ_n and whose action equals the obvious one.

By contrast, when we restrict the parameters to $\mathcal{C}$, then we obtain self-adjoint operators (denoted again H_+, H_-) on the closed subspace

$$\mathcal{H}_1(K) \subset \mathcal{H}_1 \equiv L^2((0,\pi/r),dx), \tag{1.48}$$

spanned by the functions $\Phi_n, n > K/r$, by proceeding in the obvious way: We define

$$H_\delta \Phi_n \equiv E_\delta(nr) \Phi_n, \quad n > K/r, \quad \delta = +,-, \tag{1.49}$$

extend linearly, and then take the closure. Save for some special cases, we have not been able to prove our expectation that the orthocomplement of $\mathcal{H}_1(K)$ is spanned by joint eigenfunctions $\Phi_n, n = 0,1,\ldots,[K/r]$, of the AΔOs H_δ with real eigenvalues $E_\delta(nr)$.

The functional-analytic problems involved in the above were discussed already in Refs. 2, 3; briefly, the Hilbert space theory of analytic difference operators (as opposed to *discrete* difference operators) is virtually nonexistent. Indeed, from the concrete examples we study here and elsewhere it is likely that no straightforward generalization of the standard lore concerning self-adjointness and eigenfunction expansions for ordinary differential operators exists. Roughly speaking, our strategy is instead to exploit the properties of the explicit eigenfunctions to solve the orthogonality and self-adjointness problems simultaneously—with the above restrictions and provisos, however.

The results surveyed above are detailed in Secs. II–IV. Specifically, in Sec. II we restrict attention to the case $b=ga_+, g=2,3,\ldots,$ which we already briefly considered at the end of our lecture notes.[3] In this special case we need only appeal to the constraint system studied in Appendix A. This case has several other distinctive features compared to the general case. In particular, the ''nonrelativistic limit'' $a_+\downarrow 0$ can be handled, which gives rise to the integer g eigenfunctions of the Lamé operator (1.1).

The general case studied in Sec. III is more involved. Roughly speaking, we wind up with two constraint systems of the type studied in Appendix A: one corresponding to N_+ and the other to N_-; these two systems are coupled via the spectral variable y. In this way we can handle a dense subset $\mathcal{D}$ of $\mathcal{E}$ (1.14) [which contains (1.21)], but our knowledge about the analytic properties of the eigenfunctions neither suffices to deduce the existence of an interpolation (as discussed above) nor enables us to say anything about the eigenfunctions of the Lamé operator for g not equal to an integer. [Observe that the latter can be *formally* obtained already via sequences in $D_{\rm irr}$ (1.21). Since N_+ and/or N_- must go to ∞ in this limit, we are losing control of the eigenfunction limit, however.]

Section IV is mainly devoted to a study of self-adjointness and orthogonality questions. The principal results have already been summarized above. Here we add that we find it convenient to perform a second similarity transformation to AΔOs B_+ and B_- whose structure is quite close to those of the AΔOs A_+ and A_- given by (1.15). By contrast to the similarity (1.16), however, this second similarity does not admit an interpolation to all of the elliptic parameter domain $\mathcal{E}$ (1.14). Even so, A_δ and B_δ are sufficiently close to enable us to enlarge the parameter set for which joint (A_+,A_-)-eigenfunctions [and hence (H_+,H_-)-eigenfunctions] can be found. The pertinent enlargement is somewhat involved; furthermore, the relation between A_δ and B_δ may be quite confusing on the first acquaintance. The last part of Sec. IV, where we detail the extension, can be more easily understood for the hyperbolic specialization; cf. Ref. 11.

In Appendix A we handle the constraint system associated with the zero representation when one of the integers N_+,N_- vanishes. More precisely, we study a more general system that exhibits most (but not all) of the relevant features of the former system.

In Appendix B we first collect some results on second order analytic difference equations, associated with the notion of a Casorati determinant. These well-known results are used to prove Theorem B.1. Roughly speaking, this theorem says that the meromorphic functions $\Psi(\pm x,y)$ are a basis for the joint A_δ-eigenspace, provided the quotient a_+/a_- is irrational and the spectral variable y is large enough. [In fact, we work with the similarity transforms B_δ and their holomorphic eigenfunctions $\mathcal{H}(\pm x,y)$.]

II. EIGENFUNCTIONS FOR THE INTEGER-*g* CASE

In this section we choose

$$g=N+1=2,3,\ldots, \tag{2.1}$$

in the AΔO $H_{\rm rel}$ (1.2), and accordingly obtain eigenfunctions reducing to the Lamé functions for $\beta\to 0$. [Pushing the shifts to the right in (1.2), one sees that the $g=1$ case is trivial, just as for $H_{\rm nr}$ (1.1).] To ease the notation in this section, it is convenient to trade the parameters a_+, a_-, and b in the AΔOs (1.12)–(1.16) for

$$a_+=-iv, \quad a_-=a, \quad b=-i(N+1)v, \tag{2.2}$$

and to work with suitable positive multiples of the AΔOs (1.15).

Specifically, we start from the A_--multiple

$$A\equiv \frac{s(x-(N+1)v)}{s(x)}T_v+(v\rightarrow -v),\quad s(x)\equiv s(r,a;x), \tag{2.3}$$

and at first restrict v by requiring

$$2Nv\in i(0,a). \tag{2.4}$$

Substituting (2.2) in (1.22), we obtain (note $N_-=0$)

$$\Psi(x,y)=\mathcal{N}\prod_{j=-N}^{N}\frac{1}{s(x+jv)}\cdot\prod_{j=1}^{N}s(x+z_j(y))\cdot\exp(x\Sigma), \tag{2.5}$$

with

$$\Sigma\equiv ir(N+1)+iy. \tag{2.6}$$

In the hyperbolic case $(r=0)$ the existence of eigenfunctions of this form can be deduced for arbitrary y. This is because a second far more explicit form of the eigenfunctions exists in that case (cf. Ref. 2 and Sec. II in Ref. 11), from which the existence of the factorized representation (2.5) is readily deduced.

Staying with the elliptic case, one may view (2.5) as an Ansatz for the eigenvalue equation $A\Psi=E\Psi$. Doing so, one readily verifies the identity

$$\Psi(x,y)^{-1}(A\Psi)(x,y)=E(\Sigma,z_1,\ldots,z_N;x), \tag{2.7}$$

where

$$E\equiv\frac{1}{s(x)}\left(s(x+Nv)e^{-v\Sigma}\prod_{j=1}^{N}\frac{s(x-v+z_j)}{s(x+z_j)}+(v\rightarrow -v)\right). \tag{2.8}$$

Obviously, the function E is elliptic in x with periods $\pi/r,ia$, independently of the choice of $\Sigma,z_1,\ldots,z_N\in\mathbb{C}$. Choosing from now on the numbers $z_1,\ldots,z_N$ pairwise incongruent and incongruent to 0 (modulo the period lattice), the two summands of E have simple poles at $x\equiv 0$, $-z_1,\ldots,-z_N$.

As a consequence, we obtain an eigenfunction whenever the residues at all of these poles cancel. (Indeed, this entails that E is constant.) Now for $x=0$ we need

$$e^{-v\Sigma}\prod_{j=1}^{N}s(-v+z_j)-(v\rightarrow -v)=0, \tag{2.9}$$

so that we must have

$$\Sigma=\frac{1}{2v}\ln\left(\prod_{j=1}^{N}\frac{s(z_j-v)}{s(z_j+v)}\right). \tag{2.10}$$

Substituting this in (2.8), we now study whether the residues at $x=-z_k$ can be made to cancel. For this we clearly need

$$s(z_k-Nv)\prod_{\substack{j=1\\ j\neq k}}^{N}s(z_j-z_k-v)\prod_{j=1}^{N}s(z_j+v)-(v\rightarrow -v)=0, \tag{2.11}$$

where $k=1,\ldots,N$.

The system (2.11) of N equations for N unknowns $z_1,\dots,z_N$ is a special case of the constraint system mentioned in Sec. I. Introducing the function

$$f(w)=s(vw)/v, \quad s(z)\equiv s(r,a;z), \tag{2.12}$$

it can be rewritten as a concrete form of the system (A2)–(A3) studied in Appendix A. Thus we obtain solutions

$$z_1=v+tv, \quad z_l=lv+O(t^2), \quad dz_l/dt=O(t), \quad l=2,\dots,N, \quad t\to 0, \tag{2.13}$$

to the equations (2.11) with $k=2,\dots,N$.

Now for a general function $f(w)$ it would not follow that the solution (A8) to the system (A7) also solves the larger system (A3). Due to the ancestry of the special case (2.12) of the system, however, we may deduce that one also has $F_1(W(t))=0$. Indeed, inserting (2.13) and (2.10) in the elliptic function $E(\Sigma,z_1,\dots,z_N;x)$, we obtain vanishing residues at $x\equiv 0,-z_2,\dots,-z_N$, so the residues at $x\equiv -z_1$ must vanish, too. (Recall that a nonconstant elliptic function must have more than one pole in a period cell.)

Consequently, the system (2.11) with $k=1,\dots,N$ admits a holomorphic solution curve $z_1(t),\dots,z_N(t)$ of the form (2.13) for t near 0. Moreover, we may and will choose $\epsilon>0$ small enough so that we have

$$z_j(t)\in i(0,\infty), \quad s(z_j(t)\pm v)\neq 0, \quad s(z_j(t)+Nv)\neq 0, \quad j=1,\dots,N, \tag{2.14}$$

for all $t\in(0,\epsilon)$. This ensures that

$$y(t)=-(N+1)r-\frac{i}{2v}\ln\left(\prod_{j=1}^{N}\frac{s(z_j(t)-v)}{s(z_j(t)+v)}\right) \tag{2.15}$$

is a real-valued, real-analytic function on $(0,\epsilon)$. Moreover, since the function E (2.8) is x-independent, we may choose $x=Nv$, yielding

$$E=\mu\exp(-irv(N+1)-iyv), \quad \mu\equiv\frac{s(2Nv)}{s(Nv)}\prod_{j=1}^{N}\frac{s((N-1)v+z_j)}{s(Nv+z_j)}. \tag{2.16}$$

Clearly, $\mu(t)$ is holomorphic at $t=0$ and satisfies

$$\mu(t)=1+tv\left(\frac{s'(Nv)}{s(Nv)}-\frac{s'((N+1)v)}{s((N+1)v)}\right)+O(t^2), \quad t\to 0. \tag{2.17}$$

Next, we observe that $y'(t)$ is analytic in a neighborhood of $t=0$ but for a simple pole at the origin [cf. (2.15)]:

$$y'(t)=-\frac{i}{2vt}+O(1), \quad t\to 0. \tag{2.18}$$

Eventually decreasing ϵ, we can therefore ensure

$$y'(t)<0, \quad t\in(0,\epsilon). \tag{2.19}$$

Then y decreases monotonically from ∞ to $L_\epsilon-N-1\equiv K$ as t goes from 0 to ϵ. Thus we may and will trade the parameter t for y.

To proceed, we observe that we have

$$\frac{d}{dy}\mu(t(y))=\mu'(t)(y'(t))^{-1}\to 0, \quad y\to\infty; \tag{2.20}$$

cf. (2.17) and (2.18). In view of (2.16), this entails

$$\frac{d}{dy}E\sim -iv\exp(-irv(N+1)-iyv),\quad y\to\infty. \tag{2.21}$$

(Recall that $f\sim g$ stands for $f/g\to 1$.) To obtain the asymptotics of $E=E(y)$, we note that (2.15) yields

$$y(t)=-(N+1)r-\frac{i}{2v}\ln\left(\frac{tvs(v)}{s(Nv)s((N+1)v)}\right)+O(t),\quad t\to 0. \tag{2.22}$$

Combining this with (2.16) and (2.17), we readily deduce

$$E=\exp(-ivr(N+1)-ivy)+c_N\exp(ivr(N+1)+ivy)+O(\exp(3ivy)),\quad y\to\infty, \tag{2.23}$$

with

$$c_N\equiv[s'(Nv)s((N+1)v)-s'((N+1)v)s(Nv)]/s(v). \tag{2.24}$$

The results obtained thus far hold true when v satisfies (2.4). Indeed, this restriction guarantees first of all that the hypothesis (A6) in Theorem A.1 is satisfied; cf. (2.12). But it also enables us to ensure a well-defined eigenvalue formula (2.16).

Let us now require, more generally,

$$v\in i(0,\infty),\quad kv\notin i\mathbb{N}a,\quad k=1,\ldots,2N. \tag{2.25}$$

Then we arrive at the same results as before, but for a subtle change: To guarantee the reality of y for t near 0 we may have to choose t in an interval $(-\epsilon,0)$. This eventual sign change depends on the sign of the product in (2.15) for t near 0. In view of (2.13) this sign equals the sign of $vs(v)/s(Nv)s((N+1)v)$; cf. also the asymptotics (2.22).

Next, we introduce the AΔO

$$\tilde{A}\equiv(-)^{N+1}e^{2i(N+1)rx}T_{ia}+(i\to -i). \tag{2.26}$$

Using the iterated AΔE (1.10), one infers that $\tilde{A}$ equals a positive multiple of the AΔO A_+ (1.15); cf. (2.2). It is readily verified that $\Psi(x,y)$ (2.5) is an eigenfunction of $\tilde{A}$, with the eigenvalue

$$\tilde{E}=\exp\left(2ir\sum_{j=1}^{N}z_j(y)\right)e^{ay}+\exp\left(-2ir\sum_{j=1}^{N}z_j(y)\right)e^{-2(N+1)ar}e^{-ay}. \tag{2.27}$$

(Notice that this actually holds true for an *arbitrary* dependence of z_j on y.) Clearly, we have

$$\tilde{E}\sim\exp(iN(N+1)rv)e^{ay},\quad d\tilde{E}/dy\sim a\exp(iN(N+1)rv)e^{ay},\quad y\to\infty. \tag{2.28}$$

In summary, we have arrived at joint eigenfunctions,

$$\Psi(x,y)=\mathcal{N}\prod_{j=-N}^{N}\frac{1}{s(x+jv)}\cdot\prod_{j=1}^{N}s(x+z_j)\cdot\exp\left(\frac{x}{2v}\ln\left(\prod_{j=1}^{N}\frac{s(z_j-v)}{s(z_j+v)}\right)\right), \tag{2.29}$$

of the two AΔOs A (2.3) and $\tilde{A}$ (2.26), with eigenvalues E (2.16) and $\tilde{E}$ (2.27), respectively, and with $y\in(K,\infty)$ and parameters restricted solely by (2.25). The functions $z_1,\ldots,z_N$ are solutions to the constraint system (2.11) of the form (2.13), and the solution curve parameters t and y are related via (2.15). Since the y-derivatives of both eigenvalues are positive for y large, an eventual increase of K ensures that the eigenvalue pair $(E(y),\tilde{E}(y))$ separates points on (K,∞).

We proceed by obtaining the ''nonrelativistic limit'' $v\to 0$ of the constraints (2.10), (2.11) and eigenfunctions (2.29). First, (2.10) gives rise to

$$\Sigma=-\sum_{j=1}^{N}\frac{s'(z_j)}{s(z_j)}. \tag{2.30}$$

Second, dividing (2.11) by v and taking v to 0 yields

$$N\frac{s'(z_k)}{s(z_k)}+\sum_{\substack{j=1\\ j\neq k}}^{N}\frac{s'(z_j-z_k)}{s(z_j-z_k)}-\sum_{j=1}^{N}\frac{s'(z_j)}{s(z_j)}=0,\quad k=1,\ldots,N. \tag{2.31}$$

Proceeding formally, we can also take v to 0 in the eigenfunction (2.29), yielding the limit function

$$\Psi_0(x,y)=\mathcal{N}_0 s(x)^{-2N-1}\prod_{j=1}^{N}s(x+z_j)\exp[-xs'(z_j)/s(z_j)],\quad y=-(N+1)r+i\sum_{j=1}^{N}\frac{s'(z_j)}{s(z_j)}. \tag{2.32}$$

Now (1.5) entails

$$s'(x)/s(x)=\zeta(x)-2\eta xr/\pi, \tag{2.33}$$

where ζ is the Weierstrass ζ-function. Therefore, the functions (2.32) and constraints (2.30), (2.31) amount to the Lamé functions, and associated constraints, as specified by Whittaker and Watson; cf. p. 572 and p. 574, respectively, of Ref. 4. We do not have sufficient information on the solution curve to rigorously control the above limits, though.

Next, we derive a crucial property of the holomorphic function

$$\mathcal{H}_N(x)\equiv\prod_{j=1}^{N}s(x+z_j)\cdot\exp\left(\frac{x}{2v}\ln\left(\prod_{j=1}^{N}\frac{s(z_j-v)}{s(z_j+v)}\right)\right), \tag{2.34}$$

with $z_1,\ldots,z_N$ the above solutions to the constraint system (2.11). We have suppressed the y-dependence to prevent ambiguities in the next section. We can do so, since the property holds for arbitrary y; It reads

$$\mathcal{H}_N(kv)=\mathcal{H}_N(-kv),\quad k=1,\ldots,N. \tag{2.35}$$

(As will become clear later on, this algebraic property is a key ingredient in our orthogonality analysis.)

The only k-value for which (2.35) is immediate from (2.34) is $k=1$. Indeed, in that case it holds true for $z_1,\ldots,z_N$ having arbitrary y-dependence. In order to prove (2.35), we exploit the AΔE

$$B\mathcal{H}_N=E\mathcal{H}_N,\quad B\equiv\left(\prod_{j=-N}^{N}s(x+jv)\right)A\left(\prod_{j=-N}^{N}s(x+jv)\right)^{-1}, \tag{2.36}$$

satisfied by $\mathcal{H}_N$; cf. (2.29). Recalling (2.3), we obtain

$$B=\frac{s(x+Nv)}{s(x)}T_v+(v\to -v), \tag{2.37}$$

so this AΔE can be written as

$$s(x+Nv)\mathcal{H}_N(x-v)+s(x-Nv)\mathcal{H}_N(x+v)=Es(x)\mathcal{H}_N(x). \tag{2.38}$$

First, we put $x=0$ in (2.38). Using the oddness of $s(x)$ and the restriction (2.25), this yields $\mathcal{H}_N(-v)=\mathcal{H}_N(v)$. Now we proceed recursively. Assuming (2.35) for $k=1,\ldots,l$ with $l<N$, we first substitute $x=lv$ in (2.38) to obtain

$$s((l+N)v)\mathcal{H}_N((l-1)v)+s((l-N)v)\mathcal{H}_N((l+1)v)=Es(lv)\mathcal{H}_N(lv). \tag{2.39}$$

Next, we put $x=-lv$ and use the assumption and the oddness of $s(x)$ to get

$$s((l-N)v)\mathcal{H}_N(-(l+1)v)+s((l+N)v)\mathcal{H}_N((l-1)v)=Es(lv)\mathcal{H}_N(lv). \tag{2.40}$$

Comparing (2.39) and (2.40), we obtain $\mathcal{H}_N(kv)=\mathcal{H}_N(-kv)$ for $k=l+1$, since $s((l-N)v)\neq 0$; cf. (2.25). Thus, the asserted identities (2.35) readily follow.

Besides their use in the orthogonality problem, the identities (2.35) have two further striking consequences. First, consider the function (Casorati determinant)

$$C_N(x)\equiv\mathcal{H}_N(x+v/2)\mathcal{H}_N(-x+v/2)-\mathcal{H}_N(x-v/2)\mathcal{H}_N(-x-v/2). \tag{2.41}$$

Due to (2.35), it satisfies

$$C_N(nv)=0, \quad n=-N+1/2,-N+3/2,\ldots,N-1/2. \tag{2.42}$$

This is easily seen to entail the remarkably simple result

$$C_N(x)=\alpha_N\prod_{n=-N+1/2}^{N-1/2} s(x-nv), \tag{2.43}$$

where α_N does not depend on x. [Indeed, the quotient of $C_N(x)$ and the product on the rhs is elliptic with periods $\pi/r, ia$, and pole-free in view of (2.42).]

Second, combining (2.35) and (2.34), one deduces

$$\prod_{j=1}^{N}\frac{s(z_j-kv)}{s(z_j+kv)}=\prod_{j=1}^{N}\left(\frac{s(z_j-v)}{s(z_j+v)}\right)^k, \quad k=2,\ldots,N. \tag{2.44}$$

Thus, the asymptotics (2.13) can be rendered far more precise. Indeed, from (2.13) and (2.44) one readily obtains

$$z_l=lv+d_lvt^l+O(t^{l+2}), \quad l=2,\ldots,N, \quad t\to 0, \tag{2.45}$$

where

$$d_l\equiv v^{l-1}\prod_{j=1}^{N}\frac{s(jv+lv)}{s(jv+v)^l}\cdot\prod_{j=2}^{N}s(jv-v)^l\cdot\prod_{\substack{j=l\\ j\neq 1}}^{N}\frac{1}{s(jv-lv)}, \quad l=2,\ldots,N. \tag{2.46}$$

Note that these coefficients are real and nonzero due to (2.25). Moreover, one has

$$\lim_{v\to 0}d_l=\prod_{j=1}^{N}\frac{j+l}{(j+1)^l}\cdot\prod_{j=2}^{N}(j-1)^l\cdot\prod_{\substack{j=1\\ j\neq l}}^{N}\frac{1}{(j-l)}, \quad l=2,\ldots,N. \tag{2.47}$$

III. EIGENFUNCTIONS FOR A DENSE PARAMETER SET

In this section we take the AΔOs A_+, A_- (1.15) with

$$b=(N_+ +1)a_+ - N_- a_-, \quad N_+, N_- \in \mathbb{N}^*, \tag{3.1}$$

as our starting point. Using the AΔE (1.10), they can be rewritten as

$$A_+ = (-)^{N_+ +1} \exp[a_+ r(N_+ +1)N_+ - a_- r(2N_+ +1)N_-] \times \left(e^{2ir(N_+ +1)x} \frac{s_+(x+iN_- a_-)}{s_+(x)} T_{ia_-} + (i \to -i) \right), \tag{3.2}$$

$$A_- = (-)^{N_-} \exp[a_- r(N_- +1)N_- - a_+ r(2N_- +1)(N_+ +1)] \times \left(e^{-2irN_- x} \frac{s_-(x-i(N_+ +1)a_+)}{s_-(x)} T_{ia_+} + (i \to -i) \right). \tag{3.3}$$

In the hyperbolic case, the existence of joint eigenfunctions of the form

$$\Psi(x,y) = \mathcal{N} \prod_{j=-N_+}^{N_+} \frac{1}{s_-(x+ija_+)} \cdot \prod_{j=1}^{N_+} s_-(x+z_j^+(y)) \cdot \prod_{j=1}^{N_-} s_+(x+z_j^-(y)) \cdot e^{x\Sigma}, \tag{3.4}$$

with

$$\Sigma \equiv ir(2N_+ N_- + N_+ + N_- +1) + iy, \tag{3.5}$$

can be deduced for arbitrary y; cf. Sec. III in Ref. 11. In the elliptic case we view (3.4) as an Ansatz for solving the AΔEs

$$A_\delta \Psi = E_\delta \Psi, \quad \delta = +,-. \tag{3.6}$$

Correspondingly, we calculate the functions

$$\Psi(x,y)^{-1}(A_\delta \Psi)(x,y) = E_\delta(\Sigma, z^+, z^-; x), \quad \delta = +,-. \tag{3.7}$$

Using (1.6), this readily yields

$$E_\delta = e_\delta \frac{1}{s_\delta(x)} \left(\exp\left(2ir \sum_{j=1}^{N_\delta} z_j^\delta \right) s_\delta(x+iN_{-\delta} a_{-\delta}) \exp(-ia_{-\delta}\Sigma) \times \prod_{j=1}^{N_{-\delta}} \frac{s_\delta(x-ia_{-\delta}+z_j^{-\delta})}{s_\delta(x+z_j^{-\delta})} + (i \to -i) \right), \quad \delta = +,-, \tag{3.8}$$

with

$$e_\delta \equiv \exp[a_\delta r(N_\delta +1)N_\delta - a_{-\delta} r(2N_+ N_- + N_+ + N_- +1)]. \tag{3.9}$$

Clearly, E_δ is elliptic in x with periods $\pi/r, ia_\delta$. From now on we choose the numbers $z_1^{-\delta},\ldots,z_{N_{-\delta}}^{-\delta}$ pairwise incongruent and incongruent to 0 modulo the period lattice $\pi r^{-1}\mathbb{Z} + ia_\delta \mathbb{Z}$, so that the summands have only simple poles.

It is expedient to study first one of these two elliptic functions. To minimize signs, we concentrate on the function $E_{-\delta}$ and study if and when the residues at all of its poles can be made to vanish, so as to obtain an $A_{-\delta}$-eigenfunction with eigenvalue $E_{-\delta}$. For $x=0$ it suffices to require

$$\Sigma=\frac{2r}{a_\delta}\sum_{j=1}^{N_{-\delta}} z_j^{-\delta}+\frac{1}{2ia_\delta}\ln\left(\prod_{j=1}^{N_\delta}\frac{s_{-\delta}(z_j^\delta-ia_\delta)}{s_{-\delta}(z_j^\delta+ia_\delta)}\right). \tag{3.10}$$

Substituting this in $E_{-\delta}$, we require next

$$s_{-\delta}(z_k^\delta-iN_\delta a_\delta)\prod_{\substack{j=1\\ j\neq k}}^{N_\delta} s_{-\delta}(z_j^\delta-z_k^\delta-ia_\delta)\prod_{j=1}^{N_\delta} s_{-\delta}(z_j^\delta+ia_\delta)-(i\to -i)=0, \tag{3.11}$$

where $k=1,\ldots,N_{-\delta}$. Whenever these requirements are met, we obtain vanishing residues at all poles. Accordingly, the elliptic function $E_{-\delta}$ reduces to a constant, and so we obtain an $A_{-\delta}$-eigenfunction.

Introducing the function

$$f(w)=s_{-\delta}(ia_\delta w)/ia_\delta, \tag{3.12}$$

the constraint system (3.11) turns into a special case of the system studied in Appendix A. Now this special case arose already in the previous section. Requiring henceforth

$$ka_\delta\notin\mathbb{N}a_{-\delta},\quad k=1,\ldots,2N_\delta, \tag{3.13}$$

we can therefore deduce the existence of solutions $z_j^\delta(t_\delta)$ to (3.11) that are holomorphic at $t=0$ and satisfy

$$z_1^\delta=ia_\delta(1+t_\delta),\quad z_l^\delta=ia_\delta(l+d_{\delta,l}t_\delta^l)+O(t_\delta^{l+2}),\quad l=2,\ldots,N_\delta,\quad t_\delta\to 0, \tag{3.14}$$

where

$$d_{\delta,l}=(ia_\delta)^{l-1}\prod_{j=1}^{N_\delta}\frac{s_{-\delta}(i(j+l)a_\delta)}{s_{-\delta}(i(j+1)a_\delta)^l}\cdot\prod_{j=2}^{N_\delta}s_{-\delta}(i(j-1)a_\delta)^l\cdot\prod_{\substack{j=1\\ j\neq l}}^{N_\delta}\frac{1}{s_{-\delta}(i(j-l)a_\delta)}. \tag{3.15}$$

Substituting these solutions, we deduce as before that $E_{-\delta}$ does not depend on x. Taking, for example, $\delta=+$ in (3.10)–(3.14), we therefore obtain an A_--eigenfunction with eigenvalue E_-, independently of the choice of $z_1^-,\ldots,z_{N_-}^-$. But in order to obtain a *joint* eigenfunction of A_+ and A_-, the requirements (3.10)–(3.11) must be met *simultaneously* for $\delta=+$ and $\delta=-$.

This can be achieved as follows. Consider the functions

$$g_\delta(t_\delta)\equiv\frac{2ir}{a_{-\delta}}\sum_{j=1}^{N_\delta}z_j^\delta-\frac{1}{2a_\delta}\ln\left(\prod_{j=1}^{N_\delta}\frac{s_{-\delta}(z_j^\delta-ia_\delta)}{s_{-\delta}(z_j^\delta+ia_\delta)}\right),\quad \delta=+,-, \tag{3.16}$$

where $z^\delta=z^\delta(t_\delta)$ is the above solution to (3.11). Letting t_δ vary over $(-\epsilon_\delta,0)$ or $(0,\epsilon_\delta)$ (the choice being determined by positivity of the product), the functions $g_\delta:t_\delta\mapsto u$ are real-valued, real-analytic, and monotone for ϵ_δ small enough, and u goes to ∞ for $t_\delta\to 0$. Thus the inverse functions $h_\delta:u\mapsto t_\delta$ are well defined for u varying over an interval $I_\delta=(\rho_\delta,\infty)$, and they are real-analytic and monotone on I_δ.

Letting now $\rho=\max(\rho_+,\rho_-)$, we may view t_δ as a function $h_\delta(u)$ on (ρ,∞). Doing so, we define [cf. (3.5) and (3.10)]

$$\begin{aligned}y(u)&\equiv -r(2N_+N_-+N_++N_-+1)-\frac{2ir}{a_\delta}\sum_{j=1}^{N_{-\delta}}z_j^{-\delta}-\frac{1}{2a_\delta}\ln\left(\prod_{j=1}^{N_\delta}\frac{s_{-\delta}(z_j^\delta-ia_\delta)}{s_{-\delta}(z_j^\delta+ia_\delta)}\right)\\&=-r(2N_+N_-+N_++N_-+1)-2ir\sum_{\delta=+,-}\frac{1}{a_\delta}\sum_{j=1}^{N_{-\delta}}z_j^{-\delta}+u,\end{aligned} \tag{3.17}$$

where $z^{\delta}=z^{\delta}(h_{\delta}(u))$ and $u\in(\rho,\infty)$. Eventually increasing ρ, we deduce from (3.14) and holomorphy in t_{δ} that $y(u)$ is a real-analytic, increasing function on (ρ,∞), taking values in some interval (K,∞). Thus, we may and will view u as a real-analytic function of y on (K,∞).

The upshot is that there exist real-analytic functions

$$(K,\infty)\to i(0,\infty),\quad y\mapsto z_l^{\delta}(h_{\delta}(u(y))),\quad l=1,\ldots,N_{\delta},\quad \delta=+,- \tag{3.18}$$

[denoted once more by $z_l^{\delta}(y)$], such that (3.10) and (3.11) with $k=1,\ldots,N_{-\delta}$ are satisfied both for $\delta=+$ and for $\delta=-$. As a result, we obtain a joint eigenfunction $\Psi(x,y), y\in(K,\infty)$, of the AΔOs A_+ and A_-, as advertised. Eventually increasing K, we can ensure

$$s_{\delta}(iN_{-\delta}a_{-\delta}+z_j^{-\delta}(y))\neq 0,\quad j=1,\ldots,N_{\delta},\quad \delta=+,-, \tag{3.19}$$

for all $y\in(K,\infty)$. Then we choose $x=iN_{-\delta}a_{-\delta}$ in (3.8), yielding

$$E_{\delta}(y)=\exp\left[a_{\delta}r(N_{\delta}+1)N_{\delta}+2ir\sum_{j=1}^{N_{\delta}} z_j^{\delta}(y)+a_{-\delta}y\right]$$
$$\times\frac{s(2iN_{-\delta}a_{-\delta})}{s(iN_{-\delta}a_{-\delta})}\prod_{j=1}^{N_{-\delta}}\frac{s_{\delta}(i(N_{-\delta}-1)a_{-\delta}+z_j^{-\delta}(y))}{s_{\delta}(iN_{-\delta}a_{-\delta}+z_j^{-\delta}(y))},\quad \delta=+,-. \tag{3.20}$$

Clearly, these functions are real-valued and real-analytic on (K,∞).

The $y\to\infty$ asymptotics of E_{δ} is readily determined from the above. First, let us note that (3.5) and (3.10) entail

$$|t_{\delta}|=O(\exp(-2a_{\delta}y)),\quad y\to\infty. \tag{3.21}$$

Now from (3.14), we deduce

$$E_{\delta}=\exp(a_{-\delta}y)(1-2a_{\delta}rt_{\delta}+ia_{-\delta}d_{-\delta}t_{-\delta}+O(t_+^2)+O(t_-^2)),\quad y\to\infty, \tag{3.22}$$

$$d_{-\delta}\equiv\frac{s_{\delta}'(iN_{-\delta}a_{-\delta})}{s_{\delta}(iN_{-\delta}a_{-\delta})}-\frac{s_{\delta}'(i(N_{-\delta}+1)a_{-\delta})}{s_{\delta}(i(N_{-\delta}+1)a_{-\delta})}. \tag{3.23}$$

Thus, we conclude

$$E_{\delta}(y)=\exp(a_{-\delta}y)(1+O(\exp[-2\min(a_+,a_-)y])),\quad y\to\infty. \tag{3.24}$$

Moreover, one readily verifies that

$$\frac{d}{dy}z_l^{\delta}(y)\to 0,\quad l=1,\ldots,N_{\delta},\quad \delta=+,-,\quad y\to\infty, \tag{3.25}$$

so that (3.20) entails

$$\frac{d}{dy}E_{\delta}(y)\sim a_{-\delta}\exp(a_{-\delta}y),\quad y\to\infty. \tag{3.26}$$

From this large-y asymptotics we see that $E_{\delta}(y)$ is an increasing function of y for y sufficiently large. Thus, eventually increasing K, we may and will assume that the eigenvalue pair separates points on (K,∞). [I.e., (1.20) holds true.]

Summarizing, we have proved the existence of joint eigenfunctions when the parameter b is given by (3.1) and a_+,a_- are restricted by (3.13) with $\delta=+,-$. Whenever a_+/a_- is irrational, the restrictions (3.13) are obviously satisfied for all $N_+,N_-\in\mathbb{N}$. But (3.13) is also compatible with $a_+/a_-\in\mathbb{Q}$ and a *finite* number of $(N_+,N_-)\in\mathbb{N}^2$. To be specific, letting

$$a_+/a_-=n_-/n_+, \quad n_+,n_-\in\mathbb{N}^*, \quad n_+,n_- \text{ coprime}, \tag{3.27}$$

one easily verifies that (3.13) with $\delta=+,-$ is satisfied if and only if

$$N_\delta\in\mathbb{N}, \quad N_\delta<n_\delta/2, \quad \delta=+,-. \tag{3.28}$$

Of course, for rational a_+/a_- there are infinitely many distinct pairs $(N_+,N_-)\in\mathbb{N}^2$ yielding the same b. But the conditions (3.13) cannot be satisfied for more than one pair. (Indeed, assuming, for instance, that

$$(M+1)a_+-Na_-=(M'+1)a_+-N'a_-, \quad M>M'>0, \tag{3.29}$$

one gets $(M-M')a_+=(N-N')a_-$, so that $ka_+\in\mathbb{N}a_-$ for some $k\in\{1,\ldots,M-1\}$.) This entails, in particular, that all of the numbers $(N_++1)a_+-N_-a_-$ arising from (3.27) and (3.28) are *distinct*.

Next, we consider the case

$$b=(N_-+1)a_--N_+a_+, \quad N_+,N_-\in\mathbb{N}^*. \tag{3.30}$$

Clearly, this case can be handled in the same way as the case (3.1). Specifically, we need only interchange all subscripts $+$ and $-$ in various formulas, for example, in (3.2)–(3.4). Combining (3.4), (1.25), (1.27), and their obvious counterparts for (3.30), we obtain in both cases the *same* function $\mathcal{F}(x,y)$ (1.29), namely,

$$\mathcal{F}(x,y)=\phi(\mathcal{N})\prod_{\delta=+,-}\prod_{j=1}^{N_\delta}\frac{s_{-\delta}(x+z_j^\delta(y))}{[s_{-\delta}(x+ija_\delta)s_{-\delta}(x-ija_\delta)]^{1/2}}$$
$$\times\exp[irx(2N_+N_-+N_++N_-+1)]\exp(ixy). \tag{3.31}$$

[Here, $\phi(\mathcal{N})$ is the phase of the normalization constant $\mathcal{N}$; cf. (1.24)–(1.29); these quantities can be explicitly calculated from Ref. 14, but we do not need them in the elliptic case—by contrast to the hyperbolic case; cf. Sec. III in Ref. 11.] This coincidence is in agreement with the symmetry property (1.13) of the AΔOs H_+ and H_- (1.12). Note also that in the rational case the b-values obtained from (3.1) and (3.30) are distinct, save for one special case, viz., n_+,n_- odd; $N_+=[n_+/2],N_-=[n_-/2]$, both in (3.1) and in (3.30).

In this section we have thus far excluded the special cases $N_-=0$ and/or $N_+=0$. But these cases can be easily handled, too. Indeed, when N_+ and N_- both vanish, we may and will take

$$\Psi(x,y)=\mathcal{N}\frac{1}{s_{-\delta}(x)}\exp(irx+ixy), \quad b=a_\delta, \quad \delta=+,-, \tag{3.32}$$

and when one of N_+ and N_- equals 0, we can proceed just as in Sec. II. Clearly, the resulting function $\mathcal{F}(x,y)$ (1.29) is given by (3.31) in these special cases, too.

We now summarize and extend the above findings in the following theorem.

Theorem III.1: *Fix parameters in the set $\mathcal{D}\subset\mathcal{E}$ (1.14) defined by*

$$b=(N_\alpha+1)a_\alpha-N_{-\alpha}a_{-\alpha}, \quad \alpha\in\{+,-\}, \quad N_+,N_-\in\mathbb{N}, \tag{3.33}$$

$$ka_+\notin\mathbb{N}a_-, \quad k=1,\ldots,2N_+ \quad (N_+>0), \tag{3.34}$$

$$ka_-\notin\mathbb{N}a_+, \quad k=1,\ldots,2N_- \quad (N_->0). \tag{3.35}$$

Then there exists $K\in\mathbb{R}$ such that for all $y\in(K,\infty)$ the following holds true.

(i) *The above functions*

$$z_1^+,\ldots,z_{N_+}^+,z_1^-,\ldots,z_{N_-}^-:(K,\infty)\to i(0,\infty), \tag{3.36}$$

are real-analytic solutions to the system of equations (3.5), (3.10), and (3.11) (*where* $\delta=+,-$ *and* $k=1,\ldots,N_{-\delta}$); *they satisfy (3.19) and have large-*y *asymptotics*

$$z_l^\delta(y)=ila_\delta+O(\exp(-2la_\delta y)),\quad l=1,\ldots,N_\delta,\quad \delta=+,-,\quad y\to\infty. \tag{3.37}$$

(ii) *The A*Δ*Os* H_+ *and* H_- *(1.12) have joint eigenfunctions* $\mathcal{F}(\pm x,y)$ *given by (3.31), with eigenvalues* $E_+(y)$ *and* $E_-(y)$ *given by (3.20).*

(iii) *The eigenvalues are real-valued, real-analytic functions on* (K,∞) *satisfying (3.24), (3.26), and (1.20).*

(iv) *The A*Δ*Os* A_+ *and* A_- (1.15) *have joint eigenfunctions* $\Psi(\pm x,y)$ *with eigenvalues* $E_+(y)$ *and* $E_-(y)$; *explicitly,*

$$\Psi(x,y)=\mathcal{N}\prod_{j=-N_\alpha}^{N_\alpha}\frac{1}{s_{-\alpha}(x+ija_\alpha)}\cdot\mathcal{H}(x,y), \tag{3.38}$$

where

$$\mathcal{H}(x,y)\equiv\prod_{\delta=+,-}\prod_{j=1}^{N_\delta}s_{-\delta}(x+z_j^\delta(y))\cdot\exp[irx(2N_+N_-+N_++N_-+1)+ixy]. \tag{3.39}$$

(v) *Setting*

$$\mathcal{H}^{(\infty)}(x,y)\equiv\prod_{\delta=+,-}\prod_{j=1}^{N_\delta}s_{-\delta}(x+ija_\delta)\cdot\exp[irx(2N_+N_-+N_++N_-+1)+ixy], \tag{3.40}$$

one has

$$\mathcal{H}(x,y)=\mathcal{H}^{(\infty)}(x,y)+O(\exp(-2\min(a_+,a_-)y)),\quad y\to\infty, \tag{3.41}$$

where the bound is uniform on x*-compacts.*

(vi) *The poles on the imaginary axis of the function*

$$\chi(x,y)\equiv\Psi(x,y)+\Psi(-x,y) \tag{3.42}$$

are simple and can be located only at the points

$$x=ik_\alpha a_\alpha+ik_{-\alpha}a_{-\alpha},\quad \pm k_\alpha\in\{0,1,\ldots,N_\alpha\},\quad \pm k_{-\alpha}\in\{N_{-\alpha}+1,N_{-\alpha}+2,\ldots\}. \tag{3.43}$$

Proof: We have already proved (i)–(iv). The uniform large-y asymptotics (3.41) easily follows from (3.37). Thus, it remains to prove (vi). To this end we begin by generalizing the identities (2.35). Specifically, we claim that the function $\mathcal{H}(x,y)$ satisfies

$$\mathcal{H}(ik_+a_++ik_-a_-,y)=\mathcal{H}(-ik_+a_+-ik_-a_-,y),\quad \pm k_\delta\in\{0,\ldots,N_\delta\},\quad \delta=+,-. \tag{3.44}$$

To prove this claim, we use (1.10) and (3.10) to write

$$\mathcal{H}(ik_+a_++ik_-a_-,y)=p(k_+,k_-)\prod_{\delta=+,-}\mathcal{H}_{N_\delta}^\delta(ik_\delta a_\delta),\quad k_+,k_-\in\mathbb{Z}. \tag{3.45}$$

Here, we have introduced

$$\mathcal{H}^{\delta}_{N_\delta}(x) \equiv \prod_{j=1}^{N_\delta} s_{-\delta}(x+z_j^\delta) \cdot \exp\left(\frac{x}{2ia_\delta} \ln\left(\prod_{j=1}^{N_\delta} \frac{s_{-\delta}(z_j^\delta - ia_\delta)}{s_{-\delta}(z_j^\delta + ia_\delta)} \right) \right), \tag{3.46}$$

and the prefactor reads

$$p(k_+,k_-) = \prod_{\delta=+,-} (-)^{N_\delta k_{-\delta}} \exp rN_\delta (k_{-\delta}^2 a_{-\delta} + 2k_+k_-a_\delta). \tag{3.47}$$

Now in (3.46) the y-dependence occurs via the t_δ-dependence of z_j^δ; cf. (3.18). We may therefore invoke our previous result (2.35) to deduce that our claim (3.44) holds true. (Note that p is invariant under taking $k_+,k_- \to -k_+,-k_-$.)

We now exploit the identities (3.44) to locate the poles of $\chi(x,y)$ on the imaginary axis. The product in (3.38) gives rise to poles at

$$x = ik_\alpha a_\alpha + ik_{-\alpha} a_{-\alpha}, \quad \pm k_\alpha \in \{0,\ldots,N_\alpha\}, \quad k_{-\alpha} \in \mathbb{Z}. \tag{3.48}$$

In view of (3.34) and (3.35), all of these poles are simple. Now for $\pm k_{-\alpha} \in \{0,\ldots,N_{-\alpha}\}$, the poles are matched by zeros of $\mathcal{H}(x,y) - \mathcal{H}(-x,y)$ due to (3.44). Therefore, poles of χ on the imaginary axis must be located at (3.43). □

It is quite likely that $\chi(x,y)$ does have poles at the points (3.43), i.e., no further cancellation takes place. It is illuminating to rewrite these points as [cf. (3.33)]

$$\pm ix = b - a_+ - a_- - k_\alpha a_\alpha - k_{-\alpha} a_{-\alpha}, \quad k_\alpha \in \{0,\ldots,N_\alpha\}, \quad k_{-\alpha} \in \mathbb{N}, \tag{3.49}$$

$$\pm ix = k_\alpha a_\alpha + (N_{-\alpha} + k_{-\alpha}) a_{-\alpha}, \quad k_\alpha \in \{0,\ldots,N_\alpha\}, \quad k_{-\alpha} \in \mathbb{N}^*. \tag{3.50}$$

Indeed, from this representation it is clear that when b takes values in a bounded subset of $\mathbb{R}$, then the number of poles in a rectangle $|\mathrm{Re}\, x| < \pi/r, |\mathrm{Im}\, x| < L$ is bounded above. Moreover, assuming $a_+/a_- \notin \mathbb{Q}$, the restrictions (3.34) and (3.35) hold for arbitrary $N_+,N_- \in \mathbb{N}^*$. Thus we may let $N_+,N_- \to \infty$, whilst keeping b bounded. Doing so, the points (3.50) diverge away, whereas the points (3.49) become

$$\pm ix = b - a_+ - a_- - k_+ a_+ - k_- a_-, \quad k_+,k_- \in \mathbb{N}. \tag{3.51}$$

The latter limits illuminate the issue of arbitrary-b interpolations discussed already in Sec. I, but, of course, they do not imply that an interpolation exists. For one thing, the two summands of $\chi(x,y)$ have different Floquet multipliers $\exp(\pm i\pi y/r)$ under $x \to x + \pi/r$ unless y equals $nr, n \in \mathbb{Z}$, so that generically no pole/zero cancellation occurs on the lines $\mathrm{Re}\, x = k\pi/r, k \in \mathbb{Z}^*$. Thus poles get dense on these lines as $N_+,N_- \to \infty$, and so a meromorphic interpolation is not likely to exist for generic y.

Specializing, however, to

$$\chi_n(x) \equiv \chi(x,nr), \quad nr > K, \quad n \in \mathbb{N}, \tag{3.52}$$

we deduce

$$\chi_n(x + \pi/r) = (-)^n \chi_n(x), \tag{3.53}$$

so no such obstruction occurs for these functions. But we have neither information concerning parameter continuity nor any uniform bounds available that would help in proving the existence of a meromorphic interpolation.

Before studying orthogonality properties of the functions $\chi_n(x)$, we should consider the contingency that the functions $\Psi(x,y)$ and $\Psi(-x,y)$ are linearly dependent. Now they are manifestly not identically zero, and they have different Floquet multipliers unless y equals $nr, n \in \mathbb{Z}$, so we need only study whether one can have

$$\prod_{\delta=+,-}\prod_{j=1}^{N_\delta}\frac{s_{-\delta}(x+z_j^\delta(nr))}{s_{-\delta}(x-z_j^\delta(nr))}=C\exp[-2irx(2N_+N_-+N_++N_-+1+n)], \tag{3.54}$$

where C is a nonzero constant. Clearly, this equality can only hold if the poles and zeros on the lhs cancel. Recalling the asymptotics (3.34) and (3.35), one readily infers that no cancellation takes place for n large enough.

For arbitrary $nr\in(K,\infty)$, however, we only know that we have $z_j^\delta(nr)\in i(0,\infty)$. (Recall that we have restricted K such that this is the case.) Since the solution curve gets quite inaccessible when y moves away from ∞, it appears hard to exclude pole/zero cancellation, in general. In view of the AΔE (1.10), such a cancellation might be compatible with the rhs of (3.54).

Now, even without this difficulty, we know next to nothing about the *minimal* K compatible with the various restrictions we have imposed. Thus, we may just as well assume that K is chosen large enough so that for all $y\in(K,\infty)$ the functions $\Psi(x,y)$ and $\Psi(-x,y)$ are linearly independent, and we will do so from now on. Observe that this entails, in particular, that the functions $\chi_n(x)$ (3.52) do not vanish identically.

We conclude this section with some observations concerning parameter symmetries. To this end we fix $r,a_+,a_->0$ and $N_+,N_-\in\mathbb{N}$ such that (3.34) and (3.35) hold true. Defining $b_+\equiv(N_++1)a_+-N_-a_-$, we then obtain a point $(r,a_+,a_-,b_+)\in\mathcal{D}$. Since the representation of b_+ is unique [recall the paragraph containing (3.29)], we may define $\mathcal{F}(r,a_+,a_-,b_+;x,y)$ by the rhs of (3.31). Likewise, setting $b_-\equiv(N_-+1)a_--N_+a_+$ yields a point in $\mathcal{D}$, and we may once more define $\mathcal{F}(r,a_+,a_-,b_-;x,y)$ by the rhs of (3.31).

Proceeding in this way, we therefore obtain a well-defined function $\mathcal{F}(r,a_+,a_-,b;x,y)$ for all points in $\mathcal{D}$. A moment's thought shows that this function obeys

$$\mathcal{F}(r,a_+,a_-,b;x,y)=\mathcal{F}(r,a_+,a_-,a_++a_--b;x,y), \tag{3.55}$$

$$\mathcal{F}(r,a_+,a_-,b;x,y)=\mathcal{F}(r,a_-,a_+,b;x,y). \tag{3.56}$$

Now the w-function $w(r,a_+,a_-,b;x)$ is symmetric under $a_+\leftrightarrow a_-$, but not under $b\to a_++a_--b$. Thus, setting

$$\Psi(r,a_+,a_-,b;x,y)=w(r,a_+,a_-,b;x)^{-1/2}\mathcal{F}(r,a_+,a_-,b;x,y) \tag{3.57}$$

[cf. (1.29)], we are left with the symmetry

$$\Psi(r,a_+,a_-,b;x,y)=\Psi(r,a_-,a_+,b;x,y). \tag{3.58}$$

Note that these symmetry properties for parameters in $\mathcal{D}$ have a bearing on eventual interpolations for parameters in $\mathcal{E}$.

IV. ORTHOGONALITY AND SELF-ADJOINTNESS ISSUES

The functions χ_n (3.52) span an infinite-dimensional closed subspace

$$\mathcal{H}_w(K)\subset\mathcal{H}_w\equiv L^2((0,\pi/r),w(x)dx). \tag{4.1}$$

The AΔOs A_δ give rise to densely defined operators in $\mathcal{H}_w(K)$ (denoted again by A_δ) by setting

$$A_\delta\chi_n\equiv E_\delta(nr)\chi_n,\quad n>K/r,\quad \delta=+,-, \tag{4.2}$$

and extending linearly. The question of whether the operators thus obtained are symmetric amounts to the question of whether the functions χ_n are pairwise orthogonal. Indeed, orthogonality obviously entails symmetry, and symmetry entails orthogonality, since the eigenvalues are real and satisfy (1.20).

We are now going to prove symmetry of the operators A_δ (and hence pairwise orthogonality of the functions χ_n), assuming that the parameters belong to the convergence region $\mathcal{C}$ (1.42). This restriction is equivalent to

$$(N_\delta+1)a_\delta>N_{-\delta}a_{-\delta}, \quad \delta=+,-, \tag{4.3}$$

both when $\alpha=+$ and when $\alpha=-$ in (3.33).

Now when we work with the AΔOs A_δ (1.15), then the cases $\alpha=+$ and $\alpha=-$ cannot be simultaneously handled, since A_+ and A_- lack invariance under taking $b\to a_++a_--b$; cf. also (3.2) and (3.3). This asymmetry arises from the similarity transformation (1.16). Indeed, the AΔOs H_δ (1.12) *are* invariant; cf. (1.13), but the w-function is not; cf. (1.45)–(1.47).

On the other hand, we may just as well work with the AΔOs H_+ and H_- and eigenfunctions $\Phi_n(x)$ (1.40), since (1.48) and (1.49) are unitarily equivalent to (4.1) and (4.2). But this choice has the drawback that square root branch points occur.

We shall therefore opt for a third unitarily equivalent setting that can be associated to parameters in $\mathcal{D}$. It gives rise to economic notation and meromorphic functions, and yields the same objects for the two choices of α. Specifically, in view of (3.38) we may also consider the functions

$$\psi_n(x)\equiv\mathcal{H}(x,nr)-\mathcal{H}(-x,nr), \tag{4.4}$$

which yield a closed subspace

$$\mathcal{H}_{\hat{w}}(K)\subset\mathcal{H}_{\hat{w}}\equiv L^2((0,\pi/r),\hat{w}(x)dx), \tag{4.5}$$

with

$$\begin{aligned}\hat{w}(x)&\equiv(-)^{N_++N_-+1}\mathcal{N}^2\prod_{j=-N_\alpha}^{N_\alpha}s_{-\alpha}(x+ija_\alpha)^{-2}\cdot w((N_\alpha+1)a_\alpha-N_{-\alpha}a_{-\alpha};x)\\ &=1\Big/\prod_{\delta=+,-}\prod_{\pm k=1}^{N_\delta}s_{-\delta}(x+ika_\delta);\end{aligned} \tag{4.6}$$

cf. (1.41). The associated AΔOs/operators on $\mathcal{H}_{\hat{w}}(K)$ are then given by

$$B_\delta=p_\delta\left(\exp(-2irN_\delta x)\frac{s_\delta(x+iN_{-\delta}a_{-\delta})}{s_\delta(x)}T_{ia_{-\delta}}+(i\to-i)\right), \quad \delta=+,-, \tag{4.7}$$

where the prefactor reads

$$p_\delta\equiv(-)^{N_\delta}\exp[a_\delta r(N_\delta+1)N_\delta-a_{-\delta}r(2N_\delta+1)(N_{-\delta}+1)]. \tag{4.8}$$

[Recall that (3.2) and (3.3) hold for $\alpha=+$; cf. (3.1); the $\alpha=-$ counterparts are obtained by interchanging all subscripts $+$ and $-$.]

Though the simplicity of this choice is already apparent, it should be emphasized that the weight function $\hat{w}(x)$ does not have a continuous extension to parameters in $\mathcal{E}$ (1.14), by contrast to $w(x)$. This can be seen, e.g., as follows. Fix $N_+,N_-\in\mathbb{N}^*$, and $a_+=a\in(0,\infty)$, and choose $b=(N_++1)a-N_-a_-$. Now let $a_-\to qa$ with q a positive rational number. Then $\hat{w}(x)$ (4.6) clearly has a well-defined limit. But there are infinitely many *distinct* pairs $k,l\in\mathbb{N}^*$ yielding the *same* b for $a_-\to qa$ [i.e., such that $ka-lqa$ equals $(N_++1)a-N_-qa$]. Evidently, each of these pairs yields a different limiting $\hat{w}(x)$.

Likewise, an interpolation obstruction is present for the AΔOs B_δ. [Choose, e.g., $N_-=N_++1$ and $q=1$ in the previous paragraph. From (4.7) one then sees that the limiting $b=0$ AΔOs depend on N_+; cf. (1.10).] Of course, this leads one to expect that the joint eigenfunctions $\mathcal{H}(\pm x,y)$ cannot be interpolated either. In the hyperbolic context we show that this expectation is indeed fulfilled; cf. Sec. III in Ref. 11.

As long as we restrict attention to a *fixed* choice of parameters in $\mathcal{D}$ though, the third setting just detailed is the simplest to study. We shall also use it in Appendix B, where we prove that any meromorphic joint eigenfunction $M(x)$ with B_δ-eigenvalues $E_\delta(y), \delta=+,-$, must be a linear combination of $\mathcal{H}(x,y)$ and $\mathcal{H}(-x,y)$, provided a_+/a_- is irrational and $y\in(L,\infty)$ for some $L\geqslant K$.

Returning now to the symmetry question, we begin by observing that the functions $\psi_n(x)$ (4.4) are entire, odd, and $2\pi/r$-periodic. Moreover, provided $|k|\leqslant N_+,|l|\leqslant N_-$, they have zeros in the points

$$z_{kl}\equiv ika_+ + ila_-, \quad k,l\in\mathbb{Z}, \tag{4.9}$$

[due to (3.44)], and in the points $z_{kl}+\pi/r$ (since they are either π/r-periodic or π/r-antiperiodic).

Let us now define the vector spaces

$$\mathcal{O}_0\equiv\{F(x) \text{ entire, odd,} \quad 2\pi/r\text{-periodic}\}, \tag{4.10}$$

$$\mathcal{O}_1\equiv\{F\in\mathcal{O}_0|F(z_{kl})=0, \quad F(z_{kl}+\pi/r)=0, \quad |k|\leqslant N_+,|l|\leqslant N_-\}, \tag{4.11}$$

$$\mathcal{O}_2\equiv\{F\in\mathcal{O}_0|F(z_{kl})=0, \quad F(z_{kl}+\pi/r)=0, \quad k\in\mathbb{Z},|l|\leqslant N_-, \quad \text{and } |k|\leqslant N_+,l\in\mathbb{Z}\}. \tag{4.12}$$

Clearly, we have

$$\mathcal{O}_2\subset\mathcal{O}_1\subset\mathcal{O}_0\subset\mathcal{H}_{\hat{w}}, \tag{4.13}$$

and

$$\psi_n\in\mathcal{O}_1, \quad n>K/r. \tag{4.14}$$

Next, we fix $F\in\mathcal{O}_0$ and consider B_+F. For $N_-=0$ we have $B_+F\in\mathcal{O}_0$, but for $N_->0$ we get

$$\lim_{x\to 0} s_+(x)(B_+F)(x)=2p_+s_+(iN_-a_-)F(-ia_-), \tag{4.15}$$

where we used (4.7) and oddness of F. Now the rhs does not vanish unless $F(ia_-)=0$, so, in general, $(B_+F)(x)$ has a pole at $x=0$, entailing $B_+F\notin\mathcal{H}_{\hat{w}}$. Assuming $F\in\mathcal{O}_1$, however, we have $F(ia_-)=0$ and $F(\pi/r+ia_-)=0$, so that $B_+F\in\mathcal{H}_{\hat{w}}$. More generally, this argument yields the conclusion

$$B_\delta\mathcal{O}_1\subset\mathcal{H}_{\hat{w}}, \quad \delta=+,-. \tag{4.16}$$

Therefore, the AΔOs B_δ give rise to Hilbert space operators

$$B_\delta^{(j)}:\mathcal{O}_j\to\mathcal{H}_{\hat{w}}, \quad F(x)\mapsto(B_\delta F)(x), \tag{4.17}$$

where $\delta=+,-$ and $j=1,2$.

It is not hard to see that these operators are densely defined. Indeed, $\mathcal{O}_2$ contains the subspace

$$\mathcal{O}_\delta\equiv\frac{s_\delta(x)s_{-\delta}(x)^2}{\hat{w}(x)}\mathrm{Pol}(\cos rx), \quad \delta\in\{+,-\}, \tag{4.18}$$

where $\mathrm{Pol}(t)$ denotes the space of polynomials in t, and $\mathcal{O}_\delta$ is clearly dense in $\mathcal{H}_{\hat{w}}$. The following theorem makes clear why it is important to distinguish between the operators $B_\delta^{(1)}$ and $B_\delta^{(2)}$.

Theorem IV.1: *The operators* $B_\delta^{(2)}$ *are symmetric for all parameters in* $\mathcal{D}$; *moreover, their adjoints extend the operators* $B_\delta^{(1)}$. *The operators* $B_\delta^{(1)}$ *are not symmetric for parameters in* $\mathcal{D}\backslash\mathcal{C}$, *whereas they are symmetric for parameters in* $\mathcal{D}\cap\mathcal{C}$.

Proof: For convenience we choose $\delta=-$; the case $\delta=+$ can then be handled by interchanging the subscripts $+$ and $-$ in the following. To prove symmetry of $B_-^{(2)}$, it suffices to show $I_L=I_R$, with

$$I_L\equiv\int_0^{\pi/r}\left(\exp(2irN_-x)\frac{s_-(x-iN_+a_+)}{s_-(x)}F^*(x+ia_+)+(i\to -i)\right)G(x)\hat{w}(x)dx, \quad (4.19)$$

$$I_R\equiv\int_0^{\pi/r}F^*(x)\left(\exp(-2irN_-x)\frac{s_-(x+iN_+a_+)}{s_-(x)}G(x-ia_+)+(i\to -i)\right)\hat{w}(x)dx, \quad (4.20)$$

where we take $F,G\in\mathcal{O}_2$, and where we employ the notation

$$F^*(x)\equiv\overline{F(\overline{x})}. \quad (4.21)$$

In order to prove equality of these integrals, we introduce the function

$$I(x)\equiv\hat{w}(x-e)\exp(2irN_-(x-e))\frac{s_-(x-e-iN_+a_+)}{s_-(x-e)}F^*(x+e)G(x-e), \quad e\equiv\frac{ia_+}{2}. \quad (4.22)$$

From the definition (4.6) of $\hat{w}(x)$ we deduce that we may rewrite $I(x)$ as

$$I(x)=\hat{w}(x+e)\exp(-2irN_-(x+e))\frac{s_-(x+e+iN_+a_+)}{s_-(x+e)}F^*(x+e)G(x-e). \quad (4.23)$$

Now $\hat{w}(x)$ is even and $s_-(x),F^*(x)$ and $G(x)$ are odd, so we have

$$\begin{aligned}I_L-I_R&=\int_0^{\pi/r}(I(x+e)+I(-x+e)-I(x-e)-I(-x-e))dx\\&=\int_{-\pi/r}^{\pi/r}(I(x+e)-I(x-e))dx, \quad e\equiv ia_+/2.\end{aligned} \quad (4.24)$$

Recalling F^* and G are $2\pi/r$-periodic and noting $\hat{w}$ is π/r-periodic, it follows that $I(x)$ is $2\pi/r$-periodic. Thus the integral (4.24) vanishes (by Cauchy's theorem) whenever $I(x)$ has no poles in the strip $|\mathrm{Im}\,x|\leq a_+/2$. Now since we assumed $F,G\in\mathcal{O}_2$, the function $I(x)$ is, in fact, entire. Thus we obtain $I_L=I_R$, and so $B_-^{(2)}$ is indeed symmetric.

Choosing next $F,G\in\mathcal{O}_1$, we can proceed in the same way as before, but now $I(x)$ has poles, in general. But when one of F,G belongs to $\mathcal{O}_2$, then one easily sees that $I(x)$ is still entire. Thus, the domain of $B_-^{(2)*}$ contains $\mathcal{O}_1$, and the action of $B_-^{(2)*}$ on $\mathcal{O}_1$ coincides with the action of the AΔO B_-. A moment's thought now shows that this state of affairs holds true on a larger subspace than $\mathcal{O}_1$, so that the adjoint of $B_-^{(2)}$ is a proper extension of $B_-^{(1)}$. (One need not require entireness, for instance.)

To prove the second assertion of the theorem, we determine the location of eventual poles of $I(x)$ for $F,G\in\mathcal{O}_1$. From (4.22) and (4.6) we infer that the poles of $I(x)$ are equal to the poles of the function

$$J(x)\equiv\left(\prod_{\pm k=1}^{N_-}s_+(x-e+ika_-)\right)^{-1}\left(\prod_{n=-N_++1/2}^{N_+-1/2}s_-(x+ina_+)\right)^{-1}F^*(x+e)G(x-e). \quad (4.25)$$

Let us first analyze the poles of $J(x)$ on the imaginary axis, using the notation

$$p_{kn}\equiv ika_-+ina_+, \quad k\in\mathbb{Z}, \quad n\in\mathbb{Z}+1/2. \tag{4.26}$$

The first product yields poles at the points

$$p_{kn}, \quad \pm k\in\{1,\ldots,N_-\}, \quad n\in\mathbb{Z}+1/2, \tag{4.27}$$

and the second one at

$$p_{kn}, \quad k\in\mathbb{Z}, \quad \pm n\in\{1/2,\ldots,N_+-1/2\}. \tag{4.28}$$

Thus, the products yield double poles at

$$p_{kn}, \quad \pm k\in\{1,\ldots,N_-\}, \quad \pm n\in\{1/2,\ldots,N_+-1/2\}, \tag{4.29}$$

and simple poles at

$$p_{kn}, \quad \pm k\in\{0,N_-+1,N_-+2,\ldots\}, \quad \pm n\in\{1/2,\ldots,N_+-1/2\}, \tag{4.30}$$

$$p_{kn}, \quad \pm k\in\{1,\ldots,N_-\}, \quad \pm n\in\{N_++1/2,N_++3/2,\ldots\}. \tag{4.31}$$

Now the function $F^*(x+e)G(x-e)$ has double zeros at

$$p_{kn}, \quad \pm k\in\{0,\ldots,N_-\}, \quad \pm n\in\{1/2,\ldots,N_+-1/2\}, \tag{4.32}$$

and simple zeros at

$$p_{kn}, \quad \pm k\in\{0,\ldots,N_-\}, \quad \pm n=N_++1/2. \tag{4.33}$$

Therefore, poles of $J(x)$ can be located solely at the points

$$p_{kn}, \quad \pm k\in\{N_-+1,N_-+2,\ldots\}, \quad \pm n\in\{1/2,\ldots,N_+-1/2\}, \tag{4.34}$$

$$p_{kn}, \quad \pm k\in\{1,\ldots,N_-\}, \quad \pm n\in\{N_++3/2,N_++5/2,\ldots\}. \tag{4.35}$$

We proceed by proving that for parameters in $\mathcal{D}\cap\mathcal{C}$ the latter points lie outside the strip $|\mathrm{Im}\, x|\leqslant a_+/2$. Consider first (4.34). When k and n have the same sign, it is immediate that these points are outside the critical strip. Now let $k>0$ and $n<0$. Then we get

$$ka_-+na_+\geqslant(N_-+1)a_--(N_+-1/2)a_+>a_+/2, \tag{4.36}$$

due to (4.3). Similarly, we have $ka_-+na_+<-a_+/2$ for $k<0$ and $n>0$. Next, consider (4.35). Taking $k>0$ and $n<0$, we now have

$$ka_-+na_+\leqslant N_-a_--(N_++3/2)a_+<-a_+/2, \tag{4.37}$$

due to (4.3); the other cases are then clear.

The upshot is that eventual poles of $J(x)$ on the imaginary axis lie outside $|\mathrm{Im}\, x|\leqslant a_+/2$. The above analysis can be repeated for poles on the line $\mathrm{Re}\, x=\pi/r$, yielding the same conclusion. Since $J(x)$ is $2\pi/r$-periodic, we deduce the absence of poles in the critical strip. Thus $I(x)$ has no poles in the strip either, and so the integral (4.24) vanishes. Hence, $B_-^{(1)}$ is symmetric when (4.3) holds true.

Finally, we choose parameters in $\mathcal{D}\backslash\mathcal{C}$, so that (4.3) is violated. Thus, we either have $(N_-+1)a_-<N_+a_+$ or $(N_++1)a_+<N_-a_-$. In the first case we have

$$(N_-+1)a_--(N_+-1/2)a_+<a_+/2, \quad (N_-+1)a_--a_+/2>-a_+/2, \tag{4.38}$$

so at least one of the points $p_{N_-+1,n}$ (4.34) is in the critical strip. In the second case we have

$$N_-a_- - (N_+ + 3/2)a_+ > -a_+/2, \tag{4.39}$$

so at least one of the points $p_{N_-,n}$ (4.35) is in the strip. In either case, the integral (4.24) does not vanish, in general, since we are free to choose the values of F and G in the pertinent points. Therefore, $B_-^{(1)}$ is not symmetric for parameters outside the convergence region $\mathcal{C}$. □

Taking $F=\psi_n, G=\psi_m$ in the proof of this theorem, we clearly have

$$I_L - I_R = [E_-(nr) - E_-(mr)](\psi_n, \psi_m). \tag{4.40}$$

For parameters in $\mathcal{C}$ we therefore conclude that [using (1.20)]

$$(\psi_n, \psi_m) = 0, \quad n > m > K/r. \tag{4.41}$$

But for parameters outside $\mathcal{C}$ we cannot prove that (4.41) is violated. The point is that the relevant residue sum(s) might vanish.

We conjecture that this does not happen in general. More precisely, fixing parameters in $\mathcal{D}\backslash\mathcal{C}$, we expect that one can find a pair $n \neq m$ such that $(\psi_n, \psi_m) \neq 0$. Choosing $b = 2a_\alpha$, this conjectured orthogonality breakdown can be explicitly verified for $a_\alpha > a_{-\alpha}$ (with $a_\alpha \notin \mathbb{N}a_{-\alpha}/2$) and *all* pairs $n \neq m$ with $n-m$ even. Indeed, in this special case the integral (4.24) with $F=\psi_n, G=\psi_m$ equals a nonzero residue sum. [The resulting formula for (ψ_n, ψ_m) amounts to the formula obtained by more direct means in Ref. 12, so we skip the details.]

Since the AΔOs B_δ and domains $\mathcal{O}_j$ are invariant under complex conjugation, the operators $B_\delta^{(j)}$ admit self-adjoint extensions whenever they are symmetric. Fixing parameters in $\mathcal{C}$, the operators $B_\delta^{(1)}$ are most likely essentially self-adjoint, but the state of affairs for $B_\delta^{(2)}$ is quite opaque to us. We add one observation on the self-adjoint extensions of the latter, however. Whenever one chooses parameters outside $\mathcal{C}$ and a pair $n \neq m$ with $(\psi_n, \psi_m) \neq 0$, any self-adjoint extension of $B_\delta^{(2)}$ has a domain to which ψ_n and ψ_m may or may not belong, but if both functions belong to the domain, then the action of the extension on at least one of them cannot coincide with the B_δ-action. [If it did coincide, one would deduce $(\psi_n, \psi_m) = 0$, a contradiction.]

Let us now return to the subspace $\mathcal{H}_w(K)$ (4.1) and operators A_δ (4.2). For parameters outside $\mathcal{C}$, the operators A_δ are not symmetric whenever a pair $n \neq m$ exists for which $(\chi_n, \chi_m) \neq 0$. As mentioned above, we believe that this is always the case. Choosing parameters in $\mathcal{C}$, however, (4.41) amounts to pairwise orthogonality of the functions $\chi_n, n > K/r$, so that the operators A_δ are symmetric, as announced. We also expect that for parameters in $\mathcal{C}$ the orthocomplement of $\mathcal{H}_w(K)$ is spanned by joint eigenfunctions $\chi_0, \ldots, \chi_M, M = [K/r]$, of the AΔOs A_δ with real eigenvalues. (If so, the AΔOs A_δ are essentially self-adjoint on the linear span of $\chi_0, \chi_1, \ldots,$ of course.)

We conclude this section by exploiting the AΔOs B_δ (4.7) and their eigenfunctions $\mathcal{H}(\pm x, y)$ (3.39) in yet another way. Specifically, we use them to obtain and study joint eigenfunctions of the AΔOs $A_\delta(b)$ (1.15) for $b = -N_+a_+ - N_-a_-$ and for $b = (N_+ + 1)a_+ + (N_- + 1)a_-$. Here we have $N_+, N_- \in \mathbb{N}$, and a_+, a_- are restricted by (3.34) and (3.35).

Let us recall first that both for $b = (N_+ + 1)a_+ - N_-a_-$ and for $b = (N_- + 1)a_- - N_+a_+$ we obtain the *same* AΔOs B_δ and eigenfunctions $\mathcal{H}(\pm x, y)$. Thus, in both cases we may denote the AΔOs by $B_\delta(N_+, N_-)$, and their eigenvalues and eigenfunctions by $E_\delta(N_+, N_-; y)$ and $\mathcal{H}_{N_+,N_-}(\pm x, y)$. The key observation is now that we have the identities

$$A_\delta(-N_+a_+ - N_-a_-) = r_\delta B_\delta(N_+, N_-), \quad \delta = +, -, \tag{4.42}$$

$$A_\delta((N_+ + 1)a_+ + (N_- + 1)a_-) = P_{N_+,N_-}(x)^{-1} r_\delta B_\delta(N_+, N_-) P_{N_+,N_-}(x), \quad \delta = +, -, \tag{4.43}$$

where we have introduced

$$r_\delta \equiv \exp a_{-\delta} r(2N_+ + 1)(2N_- + 1), \quad \delta = +,-, \tag{4.44}$$

$$P_{N_+,N_-}(x) \equiv \prod_{\delta=+,-} \prod_{j=-N_\delta}^{N_\delta} s_{-\delta}(x + ija_\delta), \quad N_+, N_- \in \mathbb{N}. \tag{4.45}$$

[Indeed, this can be verified directly from (1.15) and (4.7) by using the AΔE (1.10).]

As a result, we deduce

$$A_\delta(-N_+a_+ - N_-a_-)\mathcal{H}_{N_+,N_-}(\pm x,y) = r_\delta E_\delta(N_+,N_-;y)\mathcal{H}_{N_+,N_-}(\pm x,y), \tag{4.46}$$

$$\begin{aligned} A_\delta((N_+ + 1)a_+ + (N_- + 1)a_-)P_{N_+,N_-}(x)^{-1}\mathcal{H}_{N_+,N_-}(\pm x,y) \\ = r_\delta E_\delta(N_+,N_-;y)P_{N_+,N_-}(x)^{-1}\mathcal{H}_{N_+,N_-}(\pm x,y). \end{aligned} \tag{4.47}$$

Thus, we obtain the joint eigenfunctions announced above. We claim that the $y \to \infty$ asymptotics of these new eigenfunctions and eigenvalues ties in with the asymptotics for the dense parameter set $\mathcal{D}$. (Notice that the new parameters do not belong to $\mathcal{D}$.) More precisely, we claim that this holds true when we set

$$\Psi(x,y) \equiv \mathcal{N} H_{N_+,N_-}(x,y(N_+,N_-)), \quad b = -N_+a_+ - N_-a_-, \tag{4.48}$$

$$\Psi(x,y) \equiv \mathcal{N} P_{N_+,N_-}(x)^{-1}\mathcal{H}_{N_+,N_-}(x,y(N_+,N_-)), \quad b = (N_+ + 1)a_+ + (N_- + 1)a_-, \tag{4.49}$$

where

$$y(N_+,N_-) \equiv y - (2N_+ + 1)(2N_- + 1)r, \tag{4.50}$$

so that the associated eigenvalues read

$$E_\delta(y) = r_\delta E_\delta(N_+,N_-;y(N_+,N_-)), \quad \delta = +,-. \tag{4.51}$$

To prove this claim, we recall the $\mathcal{H}$-asymptotics given by (3.40) and (3.41). It entails that $\Psi(x,y)$ as just defined satisfies (1.24), where the c-function reads

$$c(-N_+a_+ - N_-a_-;x) = \mathcal{N} \prod_{\delta=+,-} \prod_{j=1}^{N_\delta} s_{-\delta}(x + ija_\delta) \cdot \exp - irx(2N_+N_- + N_+ + N_-), \tag{4.52}$$

$$c((N_+ + 1)a_+ + (N_- + 1)a_-;x) = \mathcal{N} \prod_{\delta=+,-} \prod_{j=0}^{N_\delta} \frac{1}{s_{-\delta}(x - ija_\delta)} \cdot \exp - irx(2N_+N_- + N_+ + N_-). \tag{4.53}$$

The point is now that this agrees with the interpolation (1.26) for a suitable choice of $\mathcal{N} = \mathcal{N}(r,a_+,a_-,b)$. (This readily follows from Proposition III.8 in Ref. 14.) Similarly, the eigenvalues (4.51) have once more the y-asymptotics (1.19), as is clear from the asymptotics of $E_\delta(N_+,N_-;y)$ and the definition (4.44) of r_δ.

It is easily checked that the new eigenfunctions $\Psi(x,y)$ (4.48) and (4.49) also satisfy the quasiperiodicity relations (1.36). Thus, they are eigenfunctions of the AΔO Q (1.38) with eigenvalue $2\cos(\pi y/r)$; cf. (1.39). Furthermore, choosing a_+/a_- irrational, the uniqueness result in Appendix B applies. All of these findings are consistent with the existence of interpolating meromorphic joint (A_+,A_-,Q)-eigenfunctions $\Psi(x,y)$ for parameters in $\mathcal{E}$ (1.14), but they show once more that such an interpolation must have striking analyticity properties.

For instance, taking $N_+ = N_- = 0$ in (4.48), we obtain

$$\Psi(x,y)=\exp ixy,\quad b=0; \tag{4.54}$$

cf. (3.39), (1.24), and (1.26) with $b=0$. Now when we fix a_+ and a_- with a_+/a_- irrational, and let the number $b\equiv(N_++1)a_+-N_-a_-$ converge to 0 (by taking $N_+,N_-\to\infty$ in a suitable way), then the poles of the associated functions $\Psi(x,y)$ (1.22) become dense on the lines $\mathrm{Im}\,x=k\pi/r, k\in\mathbb{Z}$. It is fully unclear to us whether and how (suitable y-dependent multiples of) these functions converge to the entire function (4.54) as $b\to 0$. But if a continuous interpolation can be found, then the existence of this limit would be a corollary.

It should be noted that the b-values in (4.48) and (4.49) are outside the orthogonality region $\mathcal{C}$ (1.42). Of course, ''orthogonality'' refers to the Hilbert space $L^2((0,\pi/r),w(x)dx)$, with $w(x)=w(r,a_+,a_-,b;x)$. Now from (1.27), (4.52), and (4.53) we have

$$w(-N_+a_+-N_-a_-;x)=(-)^{N_++N_-}\mathcal{N}^{-2}\hat{w}_{N_+,N_-}(x), \tag{4.55}$$

$$w((N_++1)a_++(N_-+1)a_-;x)=(-)^{N_++N_-}\mathcal{N}^{-2}P_{N_+,N_-}(x)^2\hat{w}_{N_+,N_-}(x), \tag{4.56}$$

where $\hat{w}_{N_+,N_-}(x)$ is given by (4.6). Therefore, the analysis embodied in Theorem IV.1 can be applied to the *odd* linear combinations,

$$\zeta_n(x)\equiv\Psi(x,nr)-\Psi(-x,nr),\quad n\in\mathbb{N},\quad nr>K+(2N_++1)(2N_-+1)r, \tag{4.57}$$

to deduce orthogonality whenever (4.3) is satisfied.

For the even combinations $\chi_n(x)$, though, this analysis renders it quite unlikely that orthogonality holds true. In fact, for the trigonometric specialization with $b=-N_+a_+, N_+>0$, we prove in Sec. IV of Ref. 11 that orthogonality is indeed violated. Thus, in the elliptic case this must also be generically true. [Of course, the two cases where $N_+=N_-=0$, namely $b=0$ and $b=a_++a_-$, are exceptional in this regard; cf. (4.54); observe that they correspond to the boundary of $\mathcal{C}$.]

Finally, we point out that the functions $\tilde{\mathcal{F}}(x,y)$ (1.29) for the new b-values $(N_++1)a_++(N_-+1)a_-$ and $-N_+a_+-N_-a_-$ are in essence equal to the functions $\mathcal{F}(x,y)$ (3.31) for the b-values $(N_++1)a_+-N_-a_-$ and $(N_-+1)a_--N_+a_+$. More precisely, we have

$$\tilde{\mathcal{F}}(x,y)=\chi\mathcal{F}(x,y(N_+,N_-)), \tag{4.58}$$

where χ is a normalizing phase. [Indeed, this is readily verified by combining (4.48), (4.49) with (4.55), (4.56).]

This intimate relation [and also the formulas (4.42)–(4.51), for that matter] can be understood from a consideration of the AΔOs $H_\delta(b)$ given by (1.12). Indeed, it is straightforward to verify that one has the identity

$$H_\delta(-N_+a_+-N_-a_-)=r_\delta H_\delta((N_++1)a_+-N_-a_-),\quad \delta=+,-, \tag{4.59}$$

so the symmetry (1.13) also entails the identity

$$H_\delta((N_++1)a_++(N_-+1)a_-)=r_\delta H_\delta((N_++1)a_+-N_-a_-),\quad \delta=+,-. \tag{4.60}$$

This explains why (4.58) holds: the relevant AΔOs are *proportional*.

More generally, a consideration of the zeros of the coefficients of $H_\delta(b)$ shows that proportionality of $H_\delta(b_1)$ and $H_\delta(b_2)$ [for arbitrary $a_+,a_-\in(0,\infty)$] not only holds for $b_2=b_1$ (trivially) and $b_2=a_++a_--b_1$ [cf. (1.13)], but also when b_1 is of the quite special form

$$2b_1=ka_++la_-,\quad k,l\in\mathbb{Z}, \tag{4.61}$$

and b_2 satisfies

$$2b_2 \in \{ka_+ + la_-, ka_+ + (-l+2)a_-, (-k+2)a_+ + la_-, (-k+2)a_+ + (-l+2)a_-\}. \tag{4.62}$$

Thus we are dealing with the case $k,l \in 2\mathbb{Z}$ for the b-values at issue.

ACKNOWLEDGMENT

This work was partly supported by the Netherlands Organisation for the Advancement of Research (NWO).

APPENDIX A: THE CONSTRAINT SYSTEM

Let $f(w)$ be an entire, odd function satisfying

$$f(w) = w + O(w^3), \quad w \to 0. \tag{A1}$$

Fixing $N \geqslant 2$, define functions $F_1, \ldots, F_N$ by

$$F_k : \mathbb{C}^N \to \mathbb{C}, \quad W \mapsto f(w_k - N) \prod_{\substack{j=1 \\ j \neq k}}^{N} f(w_k - w_j + 1) \prod_{j=1}^{N} f(w_j + 1) - (W \to -W). \tag{A2}$$

Then we have the following result concerning the system of N equations:

$$F_k(W) = 0, \quad k = 1, \ldots, N, \tag{A3}$$

for N unknowns $w_1, \ldots, w_N$.

Theorem A.1: *The system (A3) admits the solution*

$$W_0 \equiv (1,2,\ldots,N), \tag{A4}$$

and the determinant of the functional matrix

$$(DF)_{kl} \equiv \partial_l F_k, \quad k,l = 1,\ldots,N, \tag{A5}$$

vanishes for $W = W_0$. Assuming

$$f(k) \neq 0, \quad k = 1,2,\ldots,N+1, \tag{A6}$$

the system of $N-1$ equations

$$F_k(W) = 0, \quad k = 2,\ldots,N, \tag{A7}$$

for N unknowns $w_1, \ldots, w_N$ admits a unique solution of the form

$$W(t) = (1+t, w_2(t), \ldots, w_N(t)), \tag{A8}$$

near W_0, with $w_k(t)$ holomorphic at $t=0$ and such that

$$w_k(t) = k + O(t^2), \quad w_k'(t) = O(t), \quad t \to 0, \quad k = 2,\ldots,N. \tag{A9}$$

Moreover, assuming that $f(w)$ is real-valued for real w, the functions $w_k(t)$ are real-valued for $t \in (-\epsilon, \epsilon)$ and ϵ small enough.

Proof: Letting $W = W_0$, the second term on the rhs of (A2) vanishes, since $W_{0,1} = 1$. The first term vanishes for $k = N$, since $W_{0,N} = N$. For $k < N$ the first term vanishes too, since $f(w_k - w_j + 1)$ yields a zero when $j = k+1$. Thus, W_0 solves the system (A3).

Next, we calculate the functional matrix $(DF)(W_0)$. Due to the factor $f(-w_1+1)$ in the second term on the rhs of (A2), this term can only yield a nonvanishing contribution to $\partial_l F_k(W_0)$

for $l=1$, and then the partial ∂_1 must act on $f(-w_1+1)$. Since $f(-w_k+w_j+1)$ yields a zero for $k=2,\ldots,N$ and $j=k-1$, the second term only contributes to $(\partial_1 F_1)(W_0)$. Specifically, using $f'(0)=1$ [cf. (A1)] we get

$$(\partial_1 F_1)(W_0)=f(1-N)\prod_{j=3}^{N} f(-j+2)\prod_{j=1}^{N} f(j+1)+f(-1-N)\prod_{j=2}^{N} f(j)\prod_{j=2}^{N} f(-j+1)$$

$$=\prod_{j=2}^{N} f(-j+1)\prod_{j=2}^{N} f(j)\cdot[f(N+1)+f(-1-N)]=0, \tag{A10}$$

since f is odd.

To calculate the remaining partials, we need only take the first term into account. Taking first $k=N$, the factor $f(w_N-N)$ yields a zero unless $l=N$. Thus, we get

$$(\partial_l F_N)(W_0)=0, \quad l=1,\ldots,N-1, \tag{A11}$$

$$(\partial_N F_N)(W_0)=\prod_{j=1}^{N-1} f(N+1-j)\prod_{j=1}^{N} f(j+1)=f(N+1)\prod_{k=2}^{N} f(k)^2. \tag{A12}$$

Taking next $k<N$, we get a zero for $j=k+1$ unless the pertinent factor is differentiated. Hence, we obtain

$$(\partial_l F_k)(W_0)=0, \quad k<N, \quad l\neq k,k+1, \tag{A13}$$

$$(\partial_{k+1} F_k)(W_0)=-f(k-N)\prod_{j\neq k,k+1} f(k-j+1)\prod_{j=1}^{N} f(j+1), \quad k<N, \tag{A14}$$

$$(\partial_k F_k)(W_0)=-(\partial_{k+1} F_k)(W_0), \quad 1<k<N. \tag{A15}$$

Summarizing, the functional matrix is of the form

$$(DF)(W_0)=\begin{pmatrix} 0 & -a_1 & 0 & \cdots & 0 & 0 \\ 0 & a_2 & -a_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{N-1} & -a_{N-1} \\ 0 & 0 & 0 & \cdots & 0 & a_N \end{pmatrix}, \tag{A16}$$

so its determinant vanishes.

From now on we assume (A6) holds true. Then we deduce that

$$a_2,\ldots,a_N\neq 0, \tag{A17}$$

so the principal minor with indices $2,\ldots,N$ is nonzero. Therefore, the implicit function theorem guarantees a solution to the system (A7) with the asserted properties. [Note that (A9) amounts to $w_k'(0)=0$, $k=2,\ldots,N$; these relations follow from the explicit formula (A16) via implicit differentiation.] □

For a general function f the solution $W(t)$ to (A7) need not be a solution to (A3), i.e., one has $F_1(W(t))\neq 0$ for t near 0. On the other hand, whenever f is such that $F_1(W(t))=0$ for t near 0, one readily deduces from the inverse function theorem that one must have $|DF(W(t))|=0$ for t near 0.

APPENDIX B: UNIQUENESS OF JOINT EIGENFUNCTIONS

As we have seen in Sec. IV, the joint eigenspace

$$\mathcal{V}(y) \equiv \{ M \in \mathcal{M} | B_\delta M = E_\delta(y) M, \delta = +,- \}, \tag{B1}$$

with $\mathcal{M}$ given by (1.18), the AΔOs B_δ by (4.7), and their eigenvalues $E_\delta(y)$ by (3.20), contains the holomorphic functions $\mathcal{H}(x,y)$ and $\mathcal{H}(-x,y)$, and hence all of their linear combinations. But when a_+/a_- is rational, then $\mathcal{V}(y)$ is infinite-dimensional. Indeed, letting $a_+ = pa$ and $a_- = qa$ with p and q coprime integers, all $m \in \mathcal{M}$ with period ia have periods ia_+ and ia_-, too. Thus, for any two multipliers $m_+, m_- \in \mathcal{M}$ with period ia, we have

$$m_+(x)\mathcal{H}(x,y) + m_-(x)\mathcal{H}(-x,y) \in \mathcal{V}(y). \tag{B2}$$

This entails dim $(\mathcal{V}(y)) = \infty$, as asserted.

By contrast, when a_+/a_- is irrational, then $\mathcal{V}(y)$ is 2-dimensional for y sufficiently large. This is the content of Theorem B.1 below. As a preparation for this theorem and its proof we recall some well-known general features of the second order AΔEs at issue in this paper; cf., for example, Nörlund's monograph.[15]

We start from an AΔE of the form

$$f_+(x)M(x+c) + f_-(x)M(x-c) = g(x)M(x), \quad c \in \mathbb{C}^*, \tag{B3}$$

where $f_+, f_-, g \in \mathcal{M}$ with $f_+, f_- \neq 0$, and where only solutions $M \in \mathcal{M}$ are considered. Let M_1, M_2 be two solutions to (B3). Then the Casorati determinant,

$$C(M_1, M_2; x) \equiv M_1(x+c/2)M_2(x-c/2) - M_1(x-c/2)M_2(x+c/2), \tag{B4}$$

vanishes identically iff M_1/M_2 belongs to the field F_c of c-periodic meromorphic functions. Assuming from now on $M_1/M_2 \notin F_c$, the function (B4) is a solution to the first order AΔE

$$\frac{C(x+c/2)}{C(x-c/2)} = \frac{f_-(x)}{f_+(x)}, \tag{B5}$$

as is readily verified.

Next, assume $M_3(x)$ is a third solution to (B3). Then one easily verifies the identity

$$M_3(x) = m_1(x)M_2(x) - m_2(x)M_1(x), \tag{B6}$$

with

$$m_j(x) \equiv C(M_j, M_3; x+c/2)/C(M_1, M_2; x+c/2), \quad j = 1,2. \tag{B7}$$

Now quotients of Casorati determinants are c-periodic in view of the AΔE (B5), so one has $m_1, m_2 \in F_c$. Conversely, any function of the form (B6) with $m_1, m_2 \in F_c$ solves (B3). Whenever two solutions exist whose Casorati determinant is not identically zero, the solution space is, therefore, 2-dimensional over the field F_c of c-periodic meromorphic functions.

Consider now two AΔEs of the form (B3), with shift parameters

$$c_1 = ia_+, \quad c_2 = ia_-, \quad a_+, a_- > 0. \tag{B8}$$

Assume that two joint solutions exist whose Casorati determinants w.r.t. c_1 and c_2 are nonzero. When $a_+/a_- \in \mathbb{Q}$, then the joint solution space is infinite-dimensional, as we have already seen above. (Here and from now on, the field of scalars is again $\mathbb{C}$.) It may well be that for a_+/a_- irrational one can show that the assumptions just stated imply that the joint solution space is 2-dimensional, but we are not aware of a proof.

For the concrete situation encountered in the main text, however, we have *explicit* solutions available. We shall now exploit this to prove 2-dimensionality for the case at hand.

Theorem B.1: *Assume $a_+/a_- \notin \mathbb{Q}$. Then there exists $L \geqslant K$ such that for all $y \in (L,\infty)$ the joint eigenspace $\mathcal{V}(y)$ (B1) is 2-dimensional with basis vectors $\mathcal{H}(x,y)$ and $\mathcal{H}(-x,y)$.*

Proof: Consider the quotient function

$$Q(x) \equiv \mathcal{H}(x,y)/\mathcal{H}(-x,y), \quad y \in (K,\infty). \tag{B9}$$

In view of (3.39) and (3.5), it reads

$$Q(x) = \prod_{\delta=+,-} \prod_{j=1}^{N_\delta} \frac{s_{-\delta}(x+z_j^\delta)}{s_{-\delta}(-x+z_j^\delta)} \cdot e^{2x\Sigma}, \quad \Sigma = ir(2N_+N_- + N_+ + N_- + 1) + iy. \tag{B10}$$

Since $z_j^\delta \in i(0,\infty)$, the function $Q(x)$ is analytic for $\mathrm{Re}\, x \in (0,\pi/r)$. We claim that there exists $L \geqslant K$ such that for all $y \in (L,\infty)$ we have

$$\lim_{\mathrm{Im}\, x \to \infty} Q(x) = 0, \quad \mathrm{Re}\, x \in (0,\pi/r). \tag{B11}$$

To prove this claim, we first note that the AΔE (1.10) entails the bound

$$\left| \frac{s(r,a;x+z)}{s(r,a;x-z)} \right| \leqslant C \exp\left(4r|z| \frac{\mathrm{Im}\, x}{a} \right), \quad \mathrm{Re}\, x \in (0,\pi/r), \quad z \in i(0,\infty), \quad \mathrm{Im}\, x \to \infty. \tag{B12}$$

Therefore we have

$$|Q(x)| = O(\exp[2(r\eta - y)\mathrm{Im}\, x]), \quad \mathrm{Re}\, x \in (0,\pi/r), \quad \mathrm{Im}\, x \to \infty, \tag{B13}$$

with

$$\eta \equiv 2\left(\frac{1}{a_+} \sum_{j=1}^{N_-} |z_j^-| + \frac{1}{a_-} \sum_{j=1}^{N_+} |z_j^+| \right) - 2N_+N_- - N_+ - N_- - 1. \tag{B14}$$

Now the sums have finite limits as $y \to \infty$ [recall (3.14)], so there exists $L \geqslant K$ such that $r\eta < y$ for all $y \in (L,\infty)$. Hence, our claim follows.

Fixing $y \in (L,\infty)$, it now follows from (B11) that $Q(x)$ is neither ia_+-periodic nor ia_--periodic. Therefore, the Casorati determinants of $\mathcal{H}(x,y)$ and $\mathcal{H}(-x,y)$ w.r.t. ia_+ and ia_- are nonzero. Letting $M(x) \in \mathcal{V}(y)$ (B1), we then have both

$$M(x) = \lambda_+(x)\mathcal{H}(x,y) + \lambda_-(x)\mathcal{H}(-x,y), \tag{B15}$$

with $\lambda_+, \lambda_- \in F_{ia_+}$, and

$$M(x) = \mu_+(x)\mathcal{H}(x,y) + \mu_-(x)\mathcal{H}(-x,y), \tag{B16}$$

with $\mu_+, \mu_- \in F_{ia_-}$.

Next, we combine (B15) and (B16) to obtain

$$\lambda_-(x) - \mu_-(x) = (\mu_+(x) - \lambda_+(x))Q(x). \tag{B17}$$

Since $\lambda_+(x)$ and $\lambda_-(x)$ are ia_+-periodic meromorphic functions, they are analytic on the lines $\mathrm{Re}\, x = \rho \in [0,\pi/r]$, save for finitely many ρ. Similarly, $\mu_+(x)$ and $\mu_-(x)$ have this property. Now let $\rho_0 \in (0,\pi/r)$ be such that λ_+, λ_-, μ_+, and μ_- are analytic on $\mathrm{Re}\, x = \rho_0$. By periodicity, λ_+ and μ_+ are bounded on this line, so (B11) and (B17) entail

$$\lim_{\mathrm{Im}\, x \to \infty} (\lambda_-(x) - \mu_-(x)) = 0, \quad \mathrm{Re}\, x = \rho_0 \in (0, \pi/r). \tag{B18}$$

In particular, this implies that

$$\lim_{k \to \infty} \lambda_-(\rho_0 + ika_-) = \mu_-(\rho_0), \quad k \in \mathbb{N}. \tag{B19}$$

Thus far, we have not used our assumption that a_+/a_- is irrational. But now we can combine this assumption with (B19) to deduce that $\lambda_-(x)$ equals $\mu_-(\rho_0)$ for $\mathrm{Re}\, x = \rho_0$ and so for all x. [Indeed, the numbers $\rho_0 + ika_-$, $k > N$, are dense $(\mathrm{mod}\, ia_+)$ in the interval $\rho_0 + i[0, a_+)$ for arbitrary $N \in \mathbb{N}$.] Consequently, we must have $\lambda_-(x) = \mu_-(x) = c_-$ and $\lambda_+(x) = \mu_+(x) = c_+$. □

[1] S. N. M. Ruijsenaars, ''Complete integrability of relativistic Calogero–Moser systems and elliptic function identities,'' Commun. Math. Phys. **110**, 191–213 (1987).

[2] S. N. M. Ruijsenaars, ''Finite-dimensional soliton systems,'' in *Integrable and Superintegrable Systems*, edited by B. Kupershmidt (World Scientific, Singapore, 1990), pp. 165–206.

[3] S. N. M. Ruijsenaars, ''Systems of Calogero–Moser type,'' to be published in *Proceedings of the 1994 Banff Summer School 'Particles and fields,'* edited by G. Semenoff and L. Vinet.

[4] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, Cambridge, 1973).

[5] E. K. Sklyanin, ''Some algebraic structures connected with the Yang–Baxter equation,'' Funct. Anal. Appl. **16**, 27–34 (1982).

[6] E. K. Sklyanin, ''Some algebraic structures connected with the Yang–Baxter equation. Representations of quantum algebras,'' Funct. Anal. Appl. **17**, 273–284 (1983).

[7] I. Krichever and A. Zabrodin, ''Spin generalization of the Ruijsenaars–Schneider model, non-Abelian 2D Toda chain and representations of Sklyanin algebra,'' Russ. Math. Surveys **50**, 1101–1150 (1995).

[8] G. Felder and A. Varchenko, ''Algebraic Bethe Ansatz for the elliptic quantum group $E_{\tau,\eta}(sl_2)$,'' Nucl. Phys. B **480**, 485–503 (1996).

[9] G. Felder and A. Varchenko, ''Algebraic integrability of the two-body Ruijsenaars operator,'' preprint.

[10] K. Hasegawa, ''Ruijsenaars' commuting difference operators as commuting transfer matrices,'' Commun. Math. Phys. **187**, 289–325 (1997).

[11] S. N. M. Ruijsenaars, ''Generalized Lamé functions. II. Hyperbolic and trigonometric specializations,'' J. Math. Phys. **40**, 1627–1663 (1999).

[12] S. N. M. Ruijsenaars, ''Relativistic Lamé functions: The special case $g=2$,'' to appear in J. Phys. A.

[13] S. N. M. Ruijsenaars, ''On relativistic Lamé functions,'' to be published in *Proceedings of the 1997 Montreal Workshop on Calogero-Moser-Sutherland systems*, edited by J. F. van Diejen and L. Vinet.

[14] S. N. M. Ruijsenaars, ''First order analytic difference equations and integrable quantum systems,'' J. Math. Phys. **38**, 1069–1146 (1997).

[15] N. E. Nörlund, *Vorlesungen über Differenzenrechnung* (Springer-Verlag, Berlin, 1924).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

Generalized Lamé functions. II. Hyperbolic and trigonometric specializations

S. N. M. Ruijsenaars
Centre for Mathematics and Computer Science, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

(Received 24 April 1998; accepted for publication 17 September 1998)

In Part I [J. Math. Phys. **40**, 1595 (1999)] we studied eigenfunctions of the quantum dynamics that defines the two-particle relativistic Calogero–Moser system with elliptic interaction. In the present paper we consider the same system with hyperbolic and trigonometric interactions. In these special regimes the eigenfunctions are shown to admit an elementary representation that is far more explicit than the ''zero representation'' of Part I. In particular, the new representation can be exploited to prove that the hyperbolic eigenfunctions can be chosen to be symmetric under interchanging position and momentum variables (self-duality). In the trigonometric case duality properties are derived, too, and several orthogonality and completeness results are obtained. © *1999 American Institute of Physics.* [S0022-2488(99)02502-5]

I. INTRODUCTION

In the preceding paper[1] (henceforth denoted by I) we obtained and studied joint eigenfunctions of two commuting analytic difference operators (AΔOs) A_+ and A_- [given by Eq. (1.15) of I or I(1.15)]. The coefficients of these AΔOs are, in essence, elliptic. More precisely, both AΔOs A_δ have meromorphic coefficients with real period π/r, $r>0$, and imaginary quasi-period ia_δ, $a_\delta>0$, $\delta=+,-$. In the present paper we study hyperbolic and trigonometric specializations of the operators and functions introduced in I, referring the reader to the Introduction of I for a description of the context from which the pertinent operators arise, their connection to the Lamé operator, and literature dealing with the subject area involved.

On the one hand, the results obtained in this paper illuminate the elliptic regime, inasmuch as various questions left open in I can be answered for the hyperbolic and trigonometric regimes. On the other hand, the special cases are of independent interest, and have some remarkable features no longer present at the elliptic level. We study the hyperbolic case in Secs. II and III, the trigonometric one in Sec. IV. Though we begin each section by indicating how the zero representation of the eigenfunctions obtained in I can be adapted, we need not and will not use these results. Indeed, we reobtain the zero representation from a second one that is quite elementary and explicit. More generally, this paper is largely independent of I, especially as concerns the hyperbolic case.

We proceed by sketching our hyperbolic results, turning to trigonometric results towards the end of this Introduction. For $r=0$ the commuting AΔOs I(1.15) reduce to

$$A_\delta(b)=\frac{s_\delta(x-ib)}{s_\delta(x)}T_{ia_{-\delta}}+(i\to -i),\quad \delta=+,-, \tag{1.1}$$

where

$$(T_\alpha f)(x)=f(x-\alpha),\quad \alpha\in\mathbb{C}. \tag{1.2}$$

Here and below, we use the notation

$$s_\delta(x)=\sinh(\pi x/a_\delta),\quad c_\delta(x)=\cosh(\pi x/a_\delta),\quad e_\delta(x)=\exp(\pi x/a_\delta),\quad \delta=+,-. \tag{1.3}$$

0022-2488/99/40(3)/1627/37/$15.00

[We should point out that our hyperbolic s_δ-function differs from the hyperbolic specialization of our elliptic s_δ-function by a factor a_δ/π; cf. I(1.11) and I(1.8). Though this may give rise to confusion, we are opting for this abuse of notation in order to minimize clutter from constants.]

Unless explicitly mentioned otherwise, we choose the parameters occurring here in the hyperbolic domain

$$\mathcal{H}\equiv\{(a_+,a_-,b)|a_+,a_->0,b\in\mathbb{R}\}. \tag{1.4}$$

This ensures that the Hamiltonians

$$H_\delta(b)\equiv\left(\frac{s_\delta(x-ib)}{s_\delta(x)}\right)^{1/2}T_{ia_{-\delta}}\left(\frac{s_\delta(x+ib)}{s_\delta(x)}\right)^{1/2}+(i\to-i),\quad \delta=+,-, \tag{1.5}$$

are formally self-adjoint. The latter are related to the AΔOs A_δ by the similarity

$$H_\delta=w(x)^{1/2}A_\delta w(x)^{-1/2}. \tag{1.6}$$

Here, $w(a_+,a_-,b;x)$ is the hyperbolic weight function studied in Ref. 2; cf. also I(1.16), I(1.17).

Save for the functional-analytic results in Sec. IV of I, it is straightforward to specialize the results in I to the hyperbolic regime. As a matter of fact, considerable simplification occurs at several places, in particular, in Appendix B of I, where uniqueness of joint eigenfunctions is studied.

The latter uniqueness results are the only ones needed, however. Indeed, we start from a representation of the joint eigenfunctions that looks quite different from the zero representation obtained in I. This new representation holds true for the dense subset of $\mathcal{H}$ (1.4) given by

$$\mathcal{D}_{\text{hyp}}\equiv\{(a_+,a_-,b)|a_+,a_->0,b=ka_++la_-,k,l\in\mathbb{Z}\}. \tag{1.7}$$

Thus no (k,l)-dependent restriction on (a_+,a_-) occurs, by contrast to the dense subset $\mathcal{D}\subset\mathcal{D}_{\text{hyp}}$ that arises upon specializing the zero representation in I.

More is true: We could even allow a_+ and a_- to be arbitrary numbers in $\mathbb{C}^*$. Similarly, x and the spectral variable p may be chosen complex. Indeed, for a fixed b of the form ka_++la_-, $k,l\in\mathbb{Z}$, we obtain functions $M_{k,l}(a_+,a_-;\pm x,p)$ that are one-valued analytic functions in all of their four arguments, and that satisfy the joint eigenfunction equations

$$A_\delta M=2c_\delta(p)M,\quad \delta=+,-. \tag{1.8}$$

The variable p is related to the variable y used in I via

$$p=a_+a_-y/\pi. \tag{1.9}$$

This rescaling ensures that the eigenfunctions are symmetric under interchanging x and p (self-duality). To be sure, this property is by no means evident from the explicit formulas—it is the quantum translation of a classical self-duality property that is not manifest either, cf. Ref. 3. As it turns out, quantum self-duality boils down to some novel ''q-identities'' [viz., symmetry of the coefficients $c_{kl}^{(N)}(q)$ given by (2.2)–(2.5)].

To provide more perspective on the b-restriction in $\mathcal{D}_{\text{hyp}}$ (1.7), we would like to mention that the even linear combination,

$$R_{k,l}(a_+,a_-;x,p)=M_{k,l}(a_+,a_-;x,p)+(x\to-x), \tag{1.10}$$

admits an interpolation to all parameters in $\mathcal{H}$ (1.4). To be specific, there exists a joint A_δ-eigenfunction $R(a_+,a_-,b;x,p)$ that reduces to $R_{k,l}$ for $b=ka_++la_-$; it is meromorphic in x and p for fixed $(a_+,a_-,b)\in\mathcal{H}$ and real-analytic on $\mathcal{H}$ for fixed x,p with $\operatorname{Re}x$, $\operatorname{Re}p\neq0$. We already detailed this function in Subsection 6.3 of our lecture notes, Ref. 4, and it will be further studied elsewhere. It is quite unclear whether the odd combination admits interpolation, too. (If so, it presumably has a quite different structure and weaker analyticity properties; cf. the pertinent discussion in I.)

Let us now describe the contents of Secs. II and III in some more detail. In Sec. II we study the special case where the coupling constant

$$g \equiv b/a_+ \tag{1.11}$$

takes integer values. More precisely, we only study the choices

$$b = (N+1)a_+, \quad N \in \mathbb{N}. \tag{1.12}$$

For this special case the hyperbolic eigenfunctions and several salient features thereof were already presented in our survey, Ref. 5, but detailed proofs were not given there. In Sec. II we demonstrate various properties of an algebraic nature, but we relegate an account of orthogonality and completeness properties to another occasion.

Specifically, the joint A_δ-eigenfunctions read

$$M_{N+1,0}(a_+,a_-;x,p) \equiv (-i)^{N+1}[P_N(x)P_N(p)]^{-1}K_N(x,p), \tag{1.13}$$

with

$$P_N(x) \equiv \prod_{j=-N}^{N} [2s_-(x+ija_+)], \tag{1.14}$$

$$K_N(x,p) \equiv \exp(i\pi xp/a_+a_-) \sum_{k,l=0}^{N} c_{kl}^{(N)}(q) e_-((N-2k)x+(N-2l)p). \tag{1.15}$$

Here, the coefficients c_{kl} depend only on N and the phase factor

$$q \equiv \exp(i\pi a_+/a_-). \tag{1.16}$$

Explicitly, they are Laurent polynomials in q with integer coefficients, given by (2.2)–(2.5).

Equivalently, the function $K_N(x,p)$ is a joint eigenfunction of the auxiliary AΔOs

$$B_- = \frac{s_-(x+iNa_+)}{s_-(x)} T_{ia_+} + (i \to -i), \tag{1.17}$$

$$B_+ = (-)^N T_{ia_-} + (i \to -i), \tag{1.18}$$

obtained by similarity transforming the AΔOs $A_\delta((N+1)a_+)$ with $P_N(x)$. Observe that one of the two eigenvalue equations, viz.,

$$B_+K_N(x,p) = 2c_+(p)K_N(x,p), \tag{1.19}$$

is immediate from the structure (1.15) of K_N, independently of the choice of c_{kl}. With (2.2)–(2.5) in force, the second one (2.1) is proved in Theorem II.1, together with various other features of $K_N(x,p)$.

With these results at our disposal, we are in the position to make the connection to the seemingly different joint eigenfunctions arising upon hyperbolic specialization of Sec. II in I. Moreover, several uniqueness aspects can be clarified by adapting Theorem B.1 in I to the case at hand. Subsequently, we study the even combination $R_{N+1,0}(a_+,a_-;x,p)$ (1.10) in Theorem II.2. [It is denoted $R_N(x,p)$ for brevity.] In particular, we show that this joint A_δ-eigenfunction specializes to a polynomial in $c_-(x)$ for certain values of p. These results will be exploited for the trigonometric regime (Sec. IV).

The third and last theorem of Sec. II concerns the case of a rational quotient a_+/a_-. It throws new light on the zero representation, and is also a crucial input for Sec. III.

In the latter section we obtain joint A_δ- and H_δ-eigenfunctions for arbitrary parameters in $\mathcal{D}_{\rm hyp}$ (1.7), but just as in Sec. II it is convenient to use an auxiliary pair of AΔOs B_+,B_- as a starting point. These AΔOs are defined for b of the form

$$b_{+-}\equiv(N_++1)a_+-N_-a_-,\quad N_+,N_-\in\mathbb{N},\tag{1.20}$$

by similarity transforming A_+,A_- (1.1) with $P_{N_+}(x)$ (1.14). Explicitly, this yields

$$B_\delta=(-)^{N_\delta}\frac{s_\delta(x+iN_{-\delta}a_{-\delta})}{s_\delta(x)}T_{ia_{-\delta}}+(i\to -i),\quad \delta=+,-.\tag{1.21}$$

[Note this reduces to (1.17), (1.18) for $N_+=N$, $N_-=0$, and $\delta=-,+$, as should be the case, of course.]

Using the functions $K_N(x,p)$ from Sec. II as building blocks, joint B_δ-eigenfunctions are readily constructed. By virtue of (the hyperbolic specialization of) Theorem B.1 in I, the joint B_δ-eigenspace associated with eigenvalues $2c_\delta(p)$ is two-dimensional for a_+/a_- *irrational*. Now it is clear that the b-values (1.20) with a_+/a_- irrational already give rise to a dense subset of the hyperbolic parameter domain $\mathcal{H}$ (1.4). Moreover, the AΔOs B_δ (1.21) may be reinterpreted as specializations of the AΔOs

$$B_\delta(b)=\frac{s_\delta(x-i\delta(b-a_+))}{s_\delta(x)}T_{ia_{-\delta}}+(i\to -i),\quad \delta=+,-,\tag{1.22}$$

which are defined for all of $\mathcal{H}$. [By contrast, the elliptic generalizations I(4.7) do not admit a continuous interpolation to the whole elliptic parameter domain.]

On the other hand, the joint B_δ-eigenfunctions exhibit an infinite-dimensional ambiguity already for the b-values b_{+-} (1.20) and *rational* a_+/a_-. This provides an example demonstrating that the absence of interpolation ambiguities cannot follow from general arguments (as one might believe). But the ambiguity exhibited by the joint B_δ-eigenfunctions does not occur for the joint H_δ- and A_δ-eigenfunctions. Indeed, we show that for rational a_+/a_- the infinity of distinct $(k,l)\in\mathbb{Z}^2$ yielding the same $b=ka_++la_-$ gives rise to an infinity of distinct representations for the *same* function.

In order to arrive at the latter conclusions, we need as technical input Theorem III.1, which deals with the case of rational a_+/a_-. The joint H_δ- and A_δ-eigenfunctions $F(\Xi;x,p)$ and $M(\Xi;x,p)$ for arbitrary $\Xi\in\mathcal{D}_{\rm hyp}$ (1.7) are further studied in Theorems III.2 and III.3, respectively; the meromorphic functions $M_{k,l}(a_+,a_-;\pm x,p)$ mentioned above are equal to $M(a_+,a_-,b;\pm x,p)$ for $b=ka_++la_-$.

Let us now turn to the trigonometric regime, studied in Sec. IV. This arises from the elliptic regime by sending one of the two imaginary periods ia_+,ia_- [cf. I(1.11)] to $i\infty$. We will take a_- to ∞ and trade a_+ for a new parameter β. Of course, the real period π/r is kept fixed. Thus, we arrive at the trigonometric parameter domain

$$\mathcal{T}\equiv\{(r,\beta,b)\,|\,r,\beta>0,b\in\mathbb{R}\}.\tag{1.23}$$

Obviously, the elliptic AΔOs H_+ and A_+ have no limits for $a_-\to\infty$. Therefore we are left with

$$A(b)=\frac{\sin r(x-ib)}{\sin rx}T_{i\beta}+(i\to -i)\tag{1.24}$$

and

$$H(b)=\left(\frac{\sin r(x-ib)}{\sin rx}\right)^{1/2}T_{i\beta}\left(\frac{\sin r(x+ib)}{\sin rx}\right)^{1/2}+(i\to -i). \tag{1.25}$$

These AΔOs are related by the similarity

$$A=w(x)^{-1/2}Hw(x)^{1/2}, \tag{1.26}$$

with $w(r,\beta,b;x)$ the trigonometric weight function from Ref. 2. Now the parameter b is of the form $ka_+ + la_-$, $k,l\in\mathbb{Z}$, for all of the eigenfunctions in I. Thus, we need to choose $l=0$ for b to remain finite as $a_-\to\infty$. Accordingly, we only obtain eigenfunctions for the $\mathcal{T}$-subset

$$\mathcal{D}_{\mathrm{trig}}\equiv\{(r,\beta,b)\,|\,r,\beta>0, b=k\beta, k\in\mathbb{Z}\}, \tag{1.27}$$

which is no longer dense. Just as in I, all of the pertinent functions are also eigenfunctions of the quasi-periodicity AΔO

$$Q\equiv T_{\pi/r}+T_{-\pi/r}. \tag{1.28}$$

Our trigonometric joint (A,Q)-eigenfunctions are obtained via analytic continuation of their hyperbolic counterparts from Sec. II. Besides the zero representation obtained by specializing Sec. II in paper I to the trigonometric regime, we therefore get a second, far more accessible, representation.

We begin Sec. IV by detailing the latter, and then clarify its relation to the zero representation. In the remainder of Sec. IV we deal with various functional-analytic aspects. Correspondingly, the spectral variable is discretized, and we wind up with Hilbert space eigenfunctions that are essentially q-Gegenbauer polynomials, with $q=\exp(-2\beta r)$. To our knowledge, our two representations are new even in this well-studied case.

By contrast to Secs. II, III, and the first part of Sec. IV, which are largely self-contained, the remainder of Sec. IV involves various features and issues already encountered in Sec. IV of I. In particular, the drastic simplification arising in the trigonometric case allows us to answer some questions that we left open in the elliptic setting. These questions can be studied by choosing k negative in (1.27).

II. THE HYPERBOLIC INTEGER-*g* CASE

The results of this section have already been summarized in some detail in the Introduction, and we will freely use the notation and operators introduced there.

We begin by recalling that in Sec. II of I we also restricted attention to the integer g case (1.12). Now when one replaces the function $s(r,a;x)$ from I by its hyperbolic counterpart $(a/\pi)\sinh(\pi x/a)$, then it is straightforward to adapt the arguments and results that can be found in Sec. II of I. There is only one minor snag in the reasoning below I(2.13): A nonconstant hyperbolic function may have one or no pole in a period strip; cf. the functions cotanh(x) and cosh(x). The pertinent hyperbolic function $E(x)$ I(2.8), however, has *finite* and *equal* limits for $\mathrm{Re}\, x\to\pm\infty$. Therefore, the usual residue argument for elliptic functions can be easily adapted to exclude the presence of only one pole in the period strip.

The results of this section go far beyond those of Sec. II in I, however. The crux is that the eigenfunctions and eigenvalues admit a simpler and much more explicit form in the hyperbolic setting, without restrictions on the spectral variable y and the pertinent parameters. In particular, this enables us to shed more light on the "zero representation" I(2.34) of the eigenfunctions. As will be shown, the latter structure of the eigenfunctions is a consequence of the eigenfunction representation employed in this section, but various features obtained below are invisible from I(2.34). For example, the spectral variable y appears to be on a very different footing from the variable x, whereas it will turn out that x and the rescaled spectral variable p (1.9) play symmetric roles.

We proceed by detailing the joint eigenfunctions $K_N(x,p)$ (1.15) of the AΔOs B_- (1.17) and B_+ (1.18) with eigenvalues $2c_\delta(p)$, $\delta=+,-$. As already pointed out, the eigenvalue equation (1.19) is satisfied irrespective of the choice of c_{kl}. It will be shown later on, however, that the coefficients are uniquely determined up to an overall q-dependent scale factor by requiring

$$B_-K_N(x,p)=2c_-(p)K_N(x,p), \tag{2.1}$$

and continuity in q.

In order to specify c_{kl}, we introduce N-element subsets $I_k^{(N)}$ of the $2N$-element set $\{-N,\ldots,-1,1,\ldots,N\}$, as follows:

$$I_k^{(N)}\equiv\{-N,\ldots,-N+k-1,\ldots,k+1,\ldots,N\},\quad k=0,\ldots,N. \tag{2.2}$$

Now we put

$$s_{kl}^{(N)}(w)\equiv\sum_{\substack{i_1<\ldots<i_k\\ i_m\in I_l^{(N)}}} w^{i_1+\cdots+i_k}, \tag{2.3}$$

$$c_l^{(N)}(w)\equiv\sum_{\substack{i_1<\ldots<i_l\\ i_m\in I_0^{(N)}}} w^{i_1+\cdots+i_l}, \tag{2.4}$$

$$c_{kl}^{(N)}(q)\equiv(-)^{k+l}q^{N(N+1)/2}s_{kl}^{(N)}(q^{-2})c_l^{(N)}(q^{-2}). \tag{2.5}$$

(Here, empty sums equal 1 by definition.) For later use we also introduce polynomials

$$Q_l^{(N)}(u)\equiv\sum_{k=0}^{N}(-)^k s_{kl}^{(N)}(w)u^k=\prod_{i\in I_l^{(N)}}(1-w^iu). \tag{2.6}$$

With these definitions in place, we are going to prove that (2.1) holds true. Before doing so, however, we specify the cases $N=0,\ldots,3$, exemplifying the above notation:

$$(N=0)\quad c_{00}=1, \tag{2.7}$$

$$(N=1)\quad c_{00}=c_{11}=q,\quad c_{01}=c_{10}=-q^{-1}, \tag{2.8}$$

$$\begin{gathered}(N=2)\quad c_{00}=c_{22}=q^3,\quad c_{02}=c_{20}=q^{-3},\\ c_{01}=c_{10}=c_{12}=c_{21}=-q-q^{-1},\quad c_{11}=q^5+q^3+q^{-3}+q^{-5},\end{gathered} \tag{2.9}$$

$$\begin{gathered}(N=3)\quad c_{00}=c_{33}=q^6,\quad c_{03}=c_{30}=-q^{-6},\\ c_{02}=c_{20}=c_{13}=c_{31}=-\overline{c_{32}}=-\overline{c_{23}}=-\overline{c_{01}}=-\overline{c_{10}}=1+q^{-2}+q^{-4},\\ c_{11}=c_{22}=-\overline{c_{12}}=-\overline{c_{21}}=q^{10}+q^8+q^6+1+2q^{-2}+2q^{-4}+q^{-6}.\end{gathered} \tag{2.10}$$

Note that, more generally, the coefficients are Laurent polynomials in q with integer coefficients for arbitrary $N\in\mathbb{N}$. The symmetry properties

$$c_{kl}=c_{lk}=c_{N-k,N-l}=(-)^N\overline{c_{N-k,l}},\quad k,l=0,\ldots,N, \tag{2.11}$$

exhibited by these special cases are, in fact, valid for arbitrary N; they are equivalent to the symmetry properties (2.13)–(2.15) in the following theorem.

Theorem II.1: *With c_{kl} defined by (2.2)–(2.5), the function $K_N(x,p)$ (1.15) satisfies the AΔE*

$$s_-(x+iNa_+)F(x-ia_+)+s_-(x-iNa_+)F(x+ia_+)=2s_-(x)c_-(p)F(x). \tag{2.12}$$

It has the symmetry properties

$$K_N(x,p)=K_N(p,x), \tag{2.13}$$

$$K_N(x,p)=K_N(-x,-p), \tag{2.14}$$

$$K_N(x,p)=(-)^N\overline{K_N}(-x,p), \quad x,p\in\mathbb{R}, \tag{2.15}$$

and satisfies

$$K_N(x,i\delta Na_+)=(2i)^N\prod_{k=N+1}^{2N}\sin(\pi ka_+/a_-), \quad \delta=+,-. \tag{2.16}$$

Now assume

$$ka_+\notin\mathbb{N}a_-, \quad k=1,\ldots,2N. \tag{2.17}$$

Then one has

$$K_N(x,i\delta(N-l)a_+)=i^NB_l^{(N)}(c_-(x)), \quad l=0,\ldots,N, \quad \delta=+,-, \tag{2.18}$$

where $B_l^{(N)}(u)$ is a polynomial of degree l and parity $(-)^l$ with real coefficients.

Proof: For $N=0$ we have

$$K_0(x,p)=\exp(i\pi xp/a_+a_-), \tag{2.19}$$

and all assertions are immediate. Assuming $N\in\mathbb{N}^*$ from now on, we find it convenient to rewrite $K_N(x,p)$ as

$$K_N(x,p)=K_0(x,p)e_-(Nx+Np)S_N(q;e_-(-2x),e_-(-2p)), \tag{2.20}$$

with

$$S_N(q;r,t)=\sum_{k,l=0}^{N}c_{kl}^{(N)}(q)r^kt^l. \tag{2.21}$$

Now we fix $N\in\mathbb{N}^*$ and suppress the dependence on N wherever this does not give rise to confusion.

We first view the general form (2.20)–(2.21) of $K(x,p)$ as an Ansatz for solving the AΔE (2.12), so as to arrive at a system of equations for the coefficients c_{kl}. We then study this system in its own right before proving that it is satisfied by the above coefficients (2.5). The general insights thus obtained will be crucial for later purposes.

Accordingly, we plug (2.20) into (2.12), and cancel factors to obtain

$$\begin{aligned}&[q^Ne_-(x)-q^{-N}e_-(-x)]e_-(p)q^{-N}S(q^2e_-(-2x),e_-(-2p))\\&\quad+[q^{-N}e_-(x)-q^Ne_-(-x)]e_-(-p)q^NS(q^{-2}e_-(-2x),e_-(-2p))\\&\quad=[e_-(x)-e_-(-x)][e_-(p)+e_-(-p)]S(e_-(-2x),e_-(-2p)).\end{aligned} \tag{2.22}$$

Multiplying by $e_-(-x-p)$ and using (2.21), this can be rewritten as

$$(1-w^N r)\sum_{k,l=0}^{N} c_{kl}w^{-k}r^k t^l+(1-w^{-N}r)t\sum_{k,l=0}^{N} c_{kl}w^k r^k t^l-(1-r)(1+t)\sum_{k,l=0}^{N} c_{kl}r^k t^l=0,$$
$$w\equiv q^{-2}=\exp(-2i\pi a_+/a_-). \tag{2.23}$$

Clearly, this is satisfied iff the coefficients d_{mn} of the monomials $r^m t^n$, $m,n=0,\ldots,N+1$, vanish. The latter read

$$d_{mn}=(1-w^{-N+m-1})c_{m-1,n-1}+(1-w^{N-m+1})c_{m-1,n}-(1-w^m)c_{m,n-1}-(1-w^{-m})c_{mn}. \tag{2.24}$$

We now study the system of equations $d_{mn}=0$ for the unknowns c_{kl} under the side conditions

$$c_{kl}=0,\quad k<0,\quad k>N,\quad l<0. \tag{2.25}$$

These conditions are obviously satisfied for the above coefficients, and they entail $d_{mn}=0$ for $m\leqslant 0$, $m\geqslant N+1$, and $n\leqslant -1$. (This is because the first two terms in brackets vanish for $m=N+1$ and the last two for $m=0$.) Thus, we wind up with the system $d_{mn}=0$, where $m=1,\ldots,N$, $n\in\mathbb{N}$, for unknowns c_{kl} in the vertical half-strip $k=0,\ldots,N$, $l\in\mathbb{N}$.

To avoid degeneracies, we now fix $a_+,a_-\in(0,\infty)$ such that $a_+/a_-\notin\mathbb{Q}$. We claim that the solution to the system is then uniquely determined, provided we prescribe the numbers $c_{0n}\equiv b_n$, $n\in\mathbb{N}$, at the left boundary of the half-strip. To explain this, we observe that the system involves four lattice points on a plaquette. Thus, we can calculate successively $c_{mn}=c_{10},c_{20},\ldots,c_{N0},c_{11},\ldots c_{N1},c_{21},\ldots$, etc. [Indeed, since $w^m\neq 1$ for $m\in\mathbb{Z}^*$, the term $(1-w^{-m})$ in (2.24) is nonzero.] Hence our claim follows. In particular, there exists a uniquely determined solution to the system when we choose boundary coefficients

$$b_n\equiv\begin{cases}(-)^n w^{-N(N+1)/4}\displaystyle\sum_{1\leqslant i_1<\cdots<i_n\leqslant N} w^{i_1+\cdots+i_n}, & n=0,\ldots,N,\\ 0, \quad n>N, & \end{cases} \tag{2.26}$$

in accordance with (2.2)–(2.5).

The unicity of this solution will be crucial shortly, but we first prove that the unique solution is actually given by (2.5). Though this can be seen directly, it is somewhat simpler to recall that the solution property is equivalent to (2.23), and to observe that (2.23) holds iff the coefficients of the monomials t^n, $n=0,\ldots,N+1$, vanish. With c_{kl} given by (2.5), the latter conditions can be written as

$$(1-w^N r)b_n Q_n(w^{-1}r)+(1-w^{-N}r)b_{n-1}Q_{n-1}(wr)-(1-r)[b_n Q_n(r)+b_{n-1}Q_{n-1}(r)]=0,$$
$$n=0,\ldots,N+1, \tag{2.27}$$

since we have

$$b_n Q_n(r)=\sum_{m=0}^{N} c_{mn}r^m,\quad b_n=c_{0n}; \tag{2.28}$$

cf. (2.6). The crux is that we may now cancel common factors in (2.27), which yields a recurrence relation for the boundary coefficients.

Specifically, taking first $n=0$ in (2.27) and noting

$$Q_0(r)=\prod_{j=1}^{N}(1-w^j r), \tag{2.29}$$

we deduce that (2.27) is satisfied. Similarly, (2.27) is satisfied for $n=N+1$. For $n\in\{1,\ldots,N\}$ we can cancel factors to obtain

$$b_n(1-w^n r)(1-w^{-N-1}r)+b_{n-1}(1-w^{N+1}r)(1-w^{-N+n-1}r)$$
$$-b_n(1-r)(1-w^{-N+n-1}r)-b_{n-1}(1-r)(1-w^n r)=0. \tag{2.30}$$

Simplifying this, we can divide by $w^{-N+n-1}-w^n$ to obtain

$$(1-w^{-n})b_n=(1-w^{N-n+1})b_{n-1}. \tag{2.31}$$

Thus, it remains to show that the boundary coefficients (2.26) satisfy this recurrence relation.

In order to prove this, we write the recurrence as

$$(1-w^{-n})\sum_{1\le i_1<\cdots<i_n\le N} w^{i_1+\cdots+i_n}+(1-w^{N-n+1})\sum_{1\le i_1<\cdots<i_{n-1}\le N} w^{i_1+\cdots+i_{n-1}}=0,$$
$$n=1,\ldots,N. \tag{2.32}$$

Now we first handle the special case $n=N$. Then (2.32) reads

$$(1-w^{-N})w^{1+\cdots+N}+(1-w)w^{1+\cdots+N}(w^{-1}+w^{-2}+\cdots+w^{-N})=0, \tag{2.33}$$

which is clearly true. Next, we use induction on N. Thus we assume (2.32) holds when N is replaced by $N-1$. Then we need only prove (2.32) for $n\in\{1,\ldots,N-1\}$. To this end we rewrite the first term on the lhs, using the induction hypothesis:

$$(1-w^{-n})\left(\sum_{1\le i_1<\cdots<i_n\le N-1} w^{i_1+\cdots+i_n}+w^N\sum_{1\le i_1<\cdots<i_{n-1}\le N-1} w^{i_1+\cdots+i_{n-1}}\right)$$
$$=-(1-w^{N-n})\sum_{1\le i_1<\cdots<i_{n-1}\le N-1} w^{i_1+\cdots+i_{n-1}}+(w^N-w^{N-n})$$
$$\times\sum_{1\le i_1<\cdots<i_{n-1}\le N-1} w^{i_1+\cdots+i_{n-1}}$$
$$=(w^N-1)\sum_{1\le i_1<\cdots<i_{n-1}\le N-1} w^{i_1+\cdots+i_{n-1}}. \tag{2.34}$$

Adding the second term, we obtain

$$w^N\left(\sum_{1\le i_1<\cdots<i_{n-1}\le N-1}-\sum_{0\le i_1<\cdots<i_{n-1}\le N-1}\right)w^{i_1+\cdots+i_{n-1}}$$
$$+\left(\sum_{1\le i_1<\cdots<i_{n-1}\le N}-\sum_{1\le i_1<\cdots<i_{n-1}\le N-1}\right)w^{i_1+\cdots+i_{n-1}}$$
$$=-w^N\sum_{1\le i_2<\cdots<i_{n-1}\le N-1} w^{i_2+\cdots+i_{n-1}}+w^N\sum_{1\le i_1<\cdots<i_{n-2}\le N-1} w^{i_1+\cdots+i_{n-2}}=0, \tag{2.35}$$

and so (2.32) follows.

The upshot is that $K(x,p)$ satisfies the AΔE (2.12). To prove the symmetry properties (2.13)–(2.15), we exploit the uniqueness of the solution to the system $d_{mn}=0$ with side conditions (2.25) and boundary condition (2.26). First, let us note that (2.13) is equivalent to symmetry of the coefficient matrix; cf. (1.15). Now it is clear from (2.2)–(2.5) that we have $c_{m0}=c_{0m}$ for $m=0,\ldots,N$, so by uniqueness it suffices to show that the transposed matrix solves the system $d_{mn}=0$, too.

In order to prove this, we use (2.26) to write the pertinent numbers d_{mn} as

$$b_{m-1}((-)^{n-1}s_{n-1,m-1}(1-w^{-N+m-1})+(-)^n s_{n,m-1}(1-w^{N-m+1}))$$
$$-b_m((-)^{n-1}s_{n-1,m}(1-w^m)+(-)^n s_{nm}(1-w^{-m})). \tag{2.36}$$

Now we deduce from the recurrence relation (2.31) that this expression vanishes iff

$$w^{-N+m-1}s_{n-1,m-1}+s_{n,m-1}=s_{nm}+w^m s_{n-1,m}. \tag{2.37}$$

From (2.3) we see that this amounts to

$$w^{-N+m-1}\sum_{\substack{i_1<\ldots<i_{n-1}\\ i_l\in I_{m-1}}} w^{i_1+\cdots+i_{n-1}}+\sum_{\substack{i_1<\ldots<i_n\\ i_l\in I_{m-1}}} w^{i_1+\cdots+i_n}$$
$$=\sum_{\substack{i_1<\ldots<i_n\\ i_l\in I_m}} w^{i_1+\cdots+i_n}+w^m\sum_{\substack{i_1<\ldots<i_{n-1}\\ i_l\in I_m}} w^{i_1+\cdots+i_{n-1}}. \tag{2.38}$$

A moment's thought reveals that this is indeed true: both the lhs and rhs are equal to the sum

$$\sum_{\substack{i_1<\ldots<i_n\\ i_l\in\{-N,\ldots,-N+m-1,m,\ldots,N\}}} w^{i_1+\cdots+i_n}. \tag{2.39}$$

Therefore, the self-duality relation (2.13) is now proved.

Next, we demonstrate (2.14) and (2.15). Since (2.14) follows by combining (2.15) with the already proved symmetry property (2.13), it suffices to show (2.15). In view of (1.15) this amounts to c_{kl} being equal to $(-)^N\overline{c_{N-k,l}}$, and since the coefficient matrix is symmetric we need only show

$$c_{kl}=(-)^N\overline{c_{k,N-l}}. \tag{2.40}$$

Now from (2.3) we deduce $\overline{s_{k,N-l}}=s_{kl}$, and from (2.4) we have

$$\overline{c_{N-l}}=\sum_{1\le i_1<\cdots<i_{N-l}\le N} w^{-(i_1+\cdots+i_{N-l})}=w^{-(1+2+\cdots+N)}\sum_{1\le j_1<\cdots<j_l\le N} w^{j_1+\cdots j_l}=q^{N(N+1)}c_l. \tag{2.41}$$

Therefore, (2.40) is clear from (2.5).

In summary, we have now proved that $K(x,p)$ (1.15) satisfies (2.12)–(2.15), provided $a_+/a_-\notin\mathbb{Q}$. (Recall that the restriction was needed to ensure uniqueness of the solution to the coefficient system. To see why uniqueness breaks down otherwise, one need only inspect the special case $a_+=a_-$.) Since the coefficients $c_{kl}(q)$ are Laurent polynomials in $q=\exp(i\pi a_+/a_-)$, the function $K(x,p)$ is well defined and continuous for all $a_+,a_-\in(0,\infty)$. Hence, it satisfies (2.12)–(2.15) for rational a_+/a_-, too.

We continue by proving (2.16). From (2.19)–(2.21) we have

$$K(x,-iNa_+)=e_-(Nx)e_-(Nx)q^{-N^2}\sum_{k,l=0}^{N} c_{kl}e_-(-2kx)q^{2lN}. \tag{2.42}$$

Using $c_{kl}=c_{lk}$ and recalling (2.2)–(2.6), this can be rewritten as

$$K(x,-iNa_+)=e_-(2Nx)q^{-N^2+N(N+1)/2}\sum_{k=0}^{N}(-)^k c_k e_-(-2kx)Q_k(q^{2N}). \quad (2.43)$$

The key point is now that $Q_k(u)$ vanishes for $u=q^{2N}=w^{-N}$, unless $k=N$. [Indeed, N belongs to $I_k^{(N)}$, save for $k=N$; cf. (2.2).] Hence we get

$$K(x,-iNa_+)=(-)^N q^{-N^2-N(N+1)/2}Q_N(q^{2N})=\prod_{k=N+1}^{2N}(q^k-q^{-k}). \quad (2.44)$$

This amounts to (2.16) with $\delta=-$. For $\delta=+$ we use (2.14) to obtain

$$K(x,iNa_+)=K(-x,-iNa_+)=K(x,-iNa_+). \quad (2.45)$$

To prove the last assertion of the theorem, we note that by virtue of (2.13), $K(x,p)$ satisfies the dual AΔE

$$s_-(p+iNa_+)K(x,p-ia_+)+s_-(p-iNa_+)K(x,p+ia_+)=2s_-(p)c_-(x)K(x,p). \quad (2.46)$$

Substituting $p=iNa_+$, this yields

$$s_-(2iNa_+)K(x,i(N-1)a_+)=2s_-(iNa_+)c_-(x)K(x,iNa_+). \quad (2.47)$$

Assuming (2.17) from now on, we have $s_-(ika_+)\neq 0$ for $k=1,\ldots,2N$. From (2.16) we then deduce that $K(x,i(N-1)a_+)$ is a nonzero multiple of $c_-(x)$. Putting next $p=i(N-1)a_+$ in the dual AΔE, we infer that $K(x,i(N-2)a_+)$ is of the form $Ac_-(x)^2+B$, with $A\neq 0$, etc. This yields (2.18) for $\delta=-$, and then the $\delta=+$ case follows from the evenness relation (2.14). □

It should be noted that the self-duality property (2.13) entails that we have

$$\hat{B}_\delta K_N(x,p)=2c_\delta(x)K_N(x,p), \quad \delta=+,-, \quad (2.48)$$

where $\hat{B}_\delta$ are the dual AΔOs

$$\hat{B}_-\equiv\frac{s_-(p+iNa_+)}{s_-(p)}\hat{T}_{ia_+}+(i\to -i), \quad (2.49)$$

$$\hat{B}_+\equiv(-)^N\hat{T}_{ia_-}+(i\to -i), \quad (2.50)$$

with

$$(\hat{T}_\alpha G)(p)\equiv G(p-\alpha), \quad \alpha\in\mathbb{C}. \quad (2.51)$$

Thus $K_N(x,p)$ is a joint eigenfunction of four independent AΔOs. [In fact, we already exploited (2.48) with $\delta=-$ in the above proof; cf. (2.46).]

We continue by detailing the relation between $K_N(x,p)$ and the function $\mathcal{H}_N(x,y)$ from Sec. II in I [cf. I(2.34)], specialized to the hyperbolic context. Consider the two-variable polynomial $S_N(r,t)$ (2.21). The coefficient of t^N reads

$$\sum_{k=0}^{N}c_{kN}r^k=(-)^N q^{-N(N+1)/2}Q_N(r)=(-)^N q^{-N(N+1)/2}\prod_{j=1}^{N}(1-q^{2j}r); \quad (2.52)$$

cf. (2.6). Similarly, the coefficient of t^0 reads

$$S(r,0)=q^{N(N+1)/2}Q_0(r)=q^{N(N+1)/2}\prod_{j=1}^{N}(1-q^{-2j}r). \tag{2.53}$$

Now we view $S(r,t)$ as a polynomial in r with t-dependent coefficients, recalling $S(r,t)=S(t,r)$. The coefficient of r^N is therefore given by the rhs of (2.52) with $r\to t$. Assuming $t\neq q^{-2j}$, $j=1,\ldots,N$, from now on, it follows that $S(r,t)$ is of degree N in r and can be written as

$$S(r,t)=q^{-N(N+1)/2}\prod_{j=1}^{N}(1-q^{2j}t)(\rho_j-r), \tag{2.54}$$

where the roots ρ_j depend on q and t. Likewise, (2.53) entails

$$S(0,t)=q^{N(N+1)/2}\prod_{j=1}^{N}(1-q^{-2j}t). \tag{2.55}$$

Hence, putting $r=0$ in (2.54), we deduce

$$\prod_{j=1}^{N}\rho_j=q^{N(N+1)}\prod_{j=1}^{N}\frac{1-q^{-2j}t}{1-q^{2j}t}. \tag{2.56}$$

In particular, none of the roots vanishes, provided $t\neq q^{2j}$, $j=1,\ldots,N$. Moreover, from (2.53) we infer that the root ρ_j may be chosen equal to q^{2j} for $t=0$.

We now rewrite t as $e_-(-2p)$, so that (2.56) becomes

$$\prod_{j=1}^{N}\rho_j=\prod_{j=1}^{N}\frac{s_-(p+ija_+)}{s_-(p-ija_+)}. \tag{2.57}$$

Restricting attention to $\{\mathrm{Re}\, p>0\}$, we may introduce (continuous) functions $z_j(p)$ by requiring

$$\rho_j=e_-(2z_j),\quad z_j(p)\to ija_+,\quad p\to\infty,\quad j=1,\ldots,N. \tag{2.58}$$

Then a routine calculation [using (2.54)] yields

$$e_-(Nx+Np)S(e_-(-2x),e_-(-2p))=2^{2N}\prod_{j=1}^{N}[s_-(p+ija_+)s_-(p-ija_+)]^{1/2}s_-(x+z_j(p)). \tag{2.59}$$

It should be emphasized that the above holds true for all positive a_+,a_-. To establish contact with Sec. II in I, however, we should require (2.17); cf. I(2.25). Then it easily follows that the zeros $z_j(p)$ may be identified with the zeros $z_j(y)$ in *loc. cit.*, with p and y related via (1.9), and that the relation to $\mathcal{H}_N$ I(2.34), reads

$$K_N(x,p)=(4\pi/a_-)^N\prod_{j=1}^{N}[s_-(p+ija_+)s_-(p-ija_+)]^{1/2}\cdot\mathcal{H}_N(x,\pi p/a_+a_-). \tag{2.60}$$

Moreover, (2.17) entails *nonconstancy* in p for all of the zeros $z_j(p)$. [Indeed, the coefficients d_l I(2.46) in the asymptotics I(2.45) are nonzero.] We will show later on that p-independent zeros do occur when (2.17) is violated; equivalently, the polynomial $S_N(r,t)$ is not irreducible in that case.

It is a remarkable consequence of (the hyperbolic specialization of) Sec. II in I that all of the roots ρ_j lie on the unit circle for $t\in(0,\epsilon)$ and ϵ small enough. For $N=1$ this remains true for all $t\in(0,1]$ and all $a_+,a_-\in(0,\infty)$; cf. (2.56). But already for $N=2,3$ and suitable a_+,a_-, the roots do not stay on the unit circle as t goes to 1. Hence, the functions $z_j(p)$ move off the imaginary axis as p decreases from ∞ to 0. This entails that the parameter K is indispensable when one requires the z_j to belong to $i(0,\infty)$—as we do in *loc. cit.*

To see the roots move off the unit circle for $N=2$, one need only use (2.9) to calculate

$$S_2(r,1)=(q^3-q-q^{-1}+q^{-3})(1+C_2r+r^2), \quad C_2\equiv q^2+2+q^{-2}. \tag{2.61}$$

Since we have $C_2=4\cos^2(\pi a_+/a_-)$, we get $C_2\in(2,4)$ for $a_+\in(0,a_-/4)$ (say). Thus, the roots $[-C_2\pm(C_2^2-4)^{1/2}]/2$ do not lie on the unit circle for $a_+/a_-<1/4$. Likewise, for $N=3$ one readily calculates from (2.10),

$$S_3(r,1)=[q^6-q^4-q^2-(q\to q^{-1})][1+C_3(r+r^2)+r^3], \quad C_3\equiv q^4+2q^2+3+2q^{-2}+q^{-4}. \tag{2.62}$$

For $q\to 1$ the roots therefore converge to those of the polynomial $(1+r)(1+8r+r^2)$. From this it easily follows that for a_+/a_- small enough (at least) two roots move off the unit circle as $p\downarrow 0$.

Next, we reconsider the formula I(2.43) for the Casorati determinant I(2.41). In view of (2.60) and (2.14) we may as well study

$$D_N^+(x)\equiv K_N(x+ia_+/2,p)K_N(x-ia_+/2,-p)-(i\to -i). \tag{2.63}$$

Adapting the reasoning in *loc. cit.* to the present context, we obtain

$$D_N^+(x)=\beta_N(p)\prod_{n=-N+1/2}^{N-1/2} s_-(x-ina_+). \tag{2.64}$$

Indeed, the quotient of $D_N^+(x)$ and the product on the rhs is hyperbolic with period ia_- and pole-free. Since the quotient has finite limits for $\mathrm{Re}\,x\to\pm\infty$, it is x-independent.

Now the limit $\beta_N(p)$ of the quotient for $\mathrm{Re}\,x\to\infty$ (say) can be determined explicitly from (1.15); it reads

$$\beta_N(p)=2^{2N}[e_-(-p)-e_-(p)]\left(\sum_{l=0}^{N} c_{0l}e_-((N-2l)p)\right)(p\to -p). \tag{2.65}$$

Using $c_{0l}=c_{l0}$ and (2.4)–(2.6), this can be rewritten as

$$\begin{aligned}\beta_N(p)&=2^{2N+1}s_-(-p)q^{N(N+1)}Q_0(e_-(-2p))Q_0(e_-(2p))\\&=(-)^{N+1}2^{4N+1}\prod_{j=-N}^{N} s_-(p+ija_+).\end{aligned} \tag{2.66}$$

Recalling (2.60), we deduce that α_N in I(2.43) specializes to

$$\alpha_N=(-)^{N+1}2\sinh(a_+y), \quad y=\pi p/a_+a_-. \tag{2.67}$$

From (2.66) we read off that the Casorati determinant of the solutions $K_N(x,p)$ and $K_N(x,-p)$ to the AΔE (2.12) vanishes identically iff p equals $p_{jk}\equiv ija_++ika_-$ with $j=-N,\ldots,N$ and $k\in\mathbb{Z}$. For other p-values it then follows that the meromorphic quotient function $K_N(x,p)/K_N(x,-p)$ is not ia_+-periodic (cf. Appendix B in I). Moreover, from (2.18) we obtain

$$K_N(x,p_{jk})=e_+(-2kx)K_N(x,-p_{jk}), \quad j=-N,\ldots,N, \quad k\in\mathbb{Z}. \tag{2.68}$$

As should be the case, this yields an ia_+-periodic quotient $e_+(-2kx)$ whenever the lhs does not vanish identically.

Consider next the Casorati determinant

$$D_N^-(x) \equiv K_N(x+ia_-/2,p)K_N(x-ia_-/2,-p)-(i\to -i), \tag{2.69}$$

corresponding to the AΔE (1.19). From (2.19)–(2.21), we obtain

$$D_N^-(x) = -2s_+(p)e_-(2Nx)S_N(q;-e_-(-2x),e_-(-2p))S_N(q;-e_-(-2x),e_-(2p)). \tag{2.70}$$

Thus, $D_N^-(x)$ vanishes for $p=ija_+$, $j\in\mathbb{Z}$, and for p such that $K_N(x,p)=0$ identically, while for other p-values the quotient $K_N(x,p)/K_N(x,-p)$ is not ia_--periodic. [Note that the functions $K_N(x,\pm ija_+)$ are manifestly either ia_--periodic or ia_--antiperiodic, depending on the parity of j.]

Restricting attention to $\mathrm{Re}\, p\neq 0$, both $D_N^+(x)$ and $D_N^-(x)$ are nonzero. Then the reasoning in the proof of Theorem B.1 in I applies with various simplifications. It leads to the conclusion that for $a_+/a_-\notin\mathbb{Q}$ and $\mathrm{Re}\, p>0$ the joint eigenspace of the AΔO pair (B_+,B_-) corresponding to eigenvalues $(2c_+(p),2c_-(p))$ is two-dimensional, and spanned by the functions $K_N(\pm x,p)$.

The result just arrived at amounts to a sharpening of Theorem B.1 in I for the hyperbolic integer g case. It entails, in particular, that for a_+/a_- irrational the coefficients in (2.21) must be proportional to (2.5) whenever (2.1) holds true. Hence, the assertion in the sentence containing (2.1) easily follows.

It is of interest to point out a second, closely related corollary. Recall that we showed in the proof of Theorem II.1 that the system $d_{mn}=0$ with side conditions (2.25) and irrational a_+/a_- has a unique solution c_{kl} for *arbitrary* boundary coefficients c_{0n}. We are now in the position to deduce that this solution does not vanish for all $l>N$ *unless* the boundary coefficients are proportional to b_n (2.26)—a surprising fact that we are unable to establish directly.

We continue by deriving some features of the joint eigenfunction

$$R_N(x,p) \equiv (-i)^{N+1}[K_N(x,p)-K_N(x,-p)][P_N(x)P_N(p)]^{-1}, \tag{2.71}$$

of the AΔOs A_+ and A_-. Notice that this definition entails, in particular,

$$R_0(x,p) = \frac{\sin(\pi xp/a_+a_-)}{2s_-(x)s_-(p)}; \tag{2.72}$$

cf. (2.19).

Theorem II.2: *The function $R_N(x,p)$ satisfies the AΔE*

$$s_-(x-i(N+1)a_+)F(x-ia_+)+s_-(x+i(N+1)a_+)F(x+ia_+) = 2s_-(x)c_-(p)F(x). \tag{2.73}$$

It has the symmetry properties

$$R_N(x,p)=R_N(p,x), \tag{2.74}$$

$$R_N(x,p)=R_N(-x,p)=R_N(x,-p), \tag{2.75}$$

$$R_N(x,p)=\overline{R_N}(x,p), \quad x,p\in\mathbb{R}. \tag{2.76}$$

Now assume

$$a_+/a_-\notin\mathbb{Q}. \tag{2.77}$$

Then one has

$$R_N(x,\pm i(N+1)a_+) = \prod_{k=N+1}^{2N+1}[2\sin(\pi ka_+/a_-)]^{-1}. \tag{2.78}$$

Moreover, one has

$$R_N(x,i\delta(N+1+l)a_+)=G_l^{(N)}(c_-(x)),\quad l\in\mathbb{N},\quad \delta=+,-,\tag{2.79}$$

where $G_l^{(N)}(u)$ is a polynomial of degree l and parity $(-)^l$ with real coefficients.

Proof: The features (2.73)–(2.76) readily follow from Theorem II.1. Combining (2.73) and (2.74) yields the dual AΔE

$$s_-(p-i(N+1)a_+)R_N(x,p-ia_+)+(i\to -i)=2s_-(p)c_-(x)R_N(x,p).\tag{2.80}$$

Substituting $p=i(N+1)a_+$, this reads

$$s_-(2i(N+1)a_+)R_N(x,i(N+2)a_+)=2s_-(i(N+1)a_+)c_-(x)R_N(x,i(N+1)a_+).\tag{2.81}$$

Assuming (2.77) from now on, let us first take (2.78) for granted. Then (2.81) entails that $R_N(x,i(N+2)a_+)$ is a nonzero real multiple of $c_-(x)$. Taking next $p=i(N+2)a_+$ in (2.80), we infer that $R_N(x,i(N+3)a_+)$ is of the form $Cc_-(x)^2+D$, with $C\in\mathbb{R}^*$, $D\in\mathbb{R}$. More generally, putting $p=i(N+l)a_+$, $l\in\mathbb{N}^*$, yields a three-term recurrence relation with coefficients in $i\mathbb{R}^*$, and so the last assertion of the theorem easily follows.

It remains to prove (2.78). Due to (2.71) this identity amounts to

$$K_N(x,-i(N+1)a_+)-K_N(x,i(N+1)a_+)=(-)^N\prod_{j=-N}^{N}(q^je_-(x)-q^{-j}e_-(-x))\prod_{l=1}^{N}(q^l-q^{-l}).\tag{2.82}$$

In view of (1.15) and (2.14), the lhs can be written as

$$2\sum_{k=0}^{N}L_ks_-((2k+1)x),\tag{2.83}$$

with

$$\begin{aligned}L_k&\equiv q^{N(N+1)}\sum_{l=0}^{N}c_{kl}^{(N)}(q)q^{-2l(N+1)}\\&=(-)^kq^{3N(N+1)/2}c_k^{(N)}(q^{-2})\prod_{j\in I_k}(1-q^{-2j}\cdot q^{-2(N+1)}),\quad k=0,\dots,N.\end{aligned}\tag{2.84}$$

[Here we used $c_{kl}=c_{lk}$, (2.5) and (2.6).]

On the other hand, we have

$$\begin{aligned}\prod_{j=-N}^{N}(q^je_-(x)-q^{-j}e_-(-x))&=e_-((2N+1)x)\prod_{j=-N}^{N}(1-q^{-2j}e_-(-2x))\\&=e_-((2N+1)x)\sum_{m=0}^{2N+1}(-)^ms_me_-(-2mx),\end{aligned}\tag{2.85}$$

where

$$s_m\equiv\sum_{-N\le j_1<\cdots<j_m\le N}q^{-2(j_1+\cdots+j_m)},\quad m=0,\dots,2N+1.\tag{2.86}$$

Now one easily sees that $s_{2N+1-m}=s_m$, so the rhs of (2.85) can be written as

$$2\sum_{k=0}^{N} R_k s_-((2k+1)x), \quad R_k \equiv (-)^{N-k} s_{N-k}, \quad k=0,\ldots,N. \tag{2.87}$$

Comparing, we deduce that (2.82) is equivalent to the identities

$$L_k = (-)^N R_k \prod_{l=1}^{N} (q^l - q^{-l}), \quad k=0,\ldots,N. \tag{2.88}$$

We proceed by proving (2.88). First, we take $k=N$. Then (2.84) yields [cf. (2.2) and (2.4)]

$$L_N = (-)^N q^{N(N+1)/2} \prod_{l=1}^{N} (1-q^{-2l}), \tag{2.89}$$

whereas (2.87) and (2.86) imply $R_N=1$. Hence (2.88) holds true for $k=N$.

Next, we note that the recurrence relation (2.31) obtained in the proof of Theorem II.1 can be rewritten as

$$(1-w^{-k}) c_k^{(N)}(w) = (w^{N-k+1}-1) c_{k-1}^{(N)}(w), \quad k=1,\ldots,N; \tag{2.90}$$

cf. (2.26) and (2.4). In view of (2.84) and (2.2), this entails

$$\frac{L_k}{L_{k-1}} = \frac{1-w^{N-k+1}}{1-w^{-k}} \cdot \frac{1-w^{-N+k-1}\cdot w^{N+1}}{1-w^{k}\cdot w^{N+1}} = \frac{w^k - w^{N+1}}{w^{k+N+1}-1}, \quad k=1,\ldots,N. \tag{2.91}$$

To conclude the proof of the theorem, it is therefore sufficient to show that the coefficients R_k satisfy the recurrence relation (2.91), too. Due to (2.87) this amounts to the recurrence

$$\frac{s_{N-k}}{s_{N-k+1}} = \frac{w^k - w^{N+1}}{1-w^{k+N+1}}, \quad k=1,\ldots,N. \tag{2.92}$$

To prove that (2.92) indeed holds, we observe that we may write (2.86) as

$$s_m = w^{-m(N+1)} \sum_{1\leqslant i_1<\cdots<i_m\leqslant 2N+1} w^{i_1+\cdots+i_m} = w^{-m(N+1)} c_m^{(2N+1)}(w). \tag{2.93}$$

Using (2.90) with $N \to 2N+1$, we therefore have

$$\frac{s_{m-1}}{s_m} = w^{N+1} \frac{1-w^{-m}}{w^{2N-m+2}-1}, \quad m=1,\ldots,2N+1. \tag{2.94}$$

Putting $m=N-k+1$, this yields (2.92), completing the proof. □

The polynomials $G_l^{(N)}$ (2.79) may be viewed as analytic continuations of q_t^2-Gegenbauer polynomials with $q_t \in (0,1)$ to q on the unit circle; cf. (1.16). This will become clear from our study of the trigonometric setting, which we undertake in Sec. IV. Indeed, the results embodied in Theorems II.1 and II.2 have trigonometric corollaries that can be obtained rather easily.

Our next and last theorem in this section has no bearing on the trigonometric case. Rather, it throws new light on the zero representation (2.59) and the restriction (2.17) corresponding to I(2.25). Moreover, the theorem plays a crucial role in Sec. III, where we handle the general hyperbolic case. It concerns the case of rational a_+/a_-, which we encode here as

$$a_+/a_- = s/r, \quad s,r \in \mathbb{N}^*, \quad s,r \text{ coprime}. \tag{2.95}$$

Assuming (2.95), the restriction (2.17) is satisfied iff $N<r/2$. Hence for $N<r/2$ all of the zeros $z_j(p)$ on the rhs of (2.59) are p-dependent. [Recall the paragraph containing (2.60).] Now (2.59) was derived without restrictions on a_+, a_-, and N. In particular, it holds true for

$$L=M+mr, \quad M\in\mathbb{N}, \quad M\leq r-1, \quad m\in\mathbb{N}^*. \tag{2.96}$$

Our next result entails that in (2.59) we then have

$$z_j(p)=ija_+, \quad j=M+1,\ldots,L. \tag{2.97}$$

That is, these zeros are p-independent and therefore equal to their limits for $p\to\infty$; cf. (2.58). Moreover, for $r>1$ and $M\in[r/2,r-1]$, one also has

$$z_j(p)=ija_+, \quad j=r-M,\ldots,M. \tag{2.98}$$

The following theorem contains far more information than its easy corollaries just mentioned. Note, however, that the prefactors in the formulas (2.99) and (2.100) can be independently checked when one takes (2.97) and (2.98) for granted and uses (2.59).

Theorem II.3: *Fix $a_+,a_->0$ such that (2.95) holds true, and assume (2.96). Then one has*

$$K_L(x,p)=q^{L(L+1)/2}q^{-M(M+1)/2}[4s_-(rx)s_-(rp)]^m K_M(x,p). \tag{2.99}$$

Next, assume $r>1$ and $M\in[r/2,r-1]$. Then one has

$$K_M(x,p)=\prod_{j=r-M}^{M}[4q^{-j}s_-(x+ija_+)s_-(p+ija_+)]\cdot K_N(x,p), \quad N\equiv r-1-M. \tag{2.100}$$

Proof: Since the variables

$$q=\exp(i\pi s/r), \quad w=\exp(-2i\pi s/r) \tag{2.101}$$

are fixed, we may as well suppress them. Our starting point is the identity

$$S_N(u,t)=q^{N(N+1)/2}\sum_{n=0}^{N}(-t)^n c_n^{(N)}Q_n^{(N)}(u), \tag{2.102}$$

which easily follows from the above definitions [cf. (2.21) and (2.2)–(2.6)]. It entails that (2.99) is equivalent to the relation

$$\sum_{l=0}^{L}(-t)^l c_l^{(L)}Q_l^{(L)}(u)=(1-u^r)^m(1-t^r)^m\sum_{k=0}^{M}(-t)^k c_k^{(M)}Q_k^{(M)}(u). \tag{2.103}$$

We prove (2.103) in several steps. First, we note the identity

$$\prod_{l=\rho+1}^{\rho+r}(1-w^l u)=1-u^r, \quad \rho\in\mathbb{Z}. \tag{2.104}$$

Indeed, since s and r are coprime, the numbers ls, with r consecutive integers l, are distinct $\bmod\, r$. Thus (2.104) is a consequence of the identity

$$\prod_{j=1}^{r}(1-\zeta^j u)=1-u^r, \quad \zeta\equiv\exp(-2i\pi/r), \tag{2.105}$$

whose proof is immediate.

Second, we exploit (2.104) to relate $Q_l^{(L)}$ for l of the form $k+jr$ with $k=0,\ldots,M$ and $j=0,\ldots,m$ to $Q_k^{(M)}$. Specifically, from the definitions (2.6) and (2.2), we obtain

$$\begin{aligned} Q_{k+jr}^{(L)}(u) &= \prod_{l=-L}^{-L+k+jr-1} (1-w^l u)\cdot \prod_{l=k+jr+1}^{L} (1-w^l u) \\ &= (1-u^r)^j \prod_{l=-M}^{-M+k-1} (1-w^l u)\cdot (1-u^r)^{m-j} \prod_{l=k+1}^{M} (1-w^l u) \\ &= (1-u^r)^m Q_k^{(M)}(u), \quad k=0,\ldots,M, \quad j=0,\ldots,m. \end{aligned} \tag{2.106}$$

Third, we combine the special case $k,j=0$ of (2.106), which we rewrite as

$$Q_0^{(L)}(u) = Q_0^{(M)}(u) \sum_{j=0}^{m} (-)^j \binom{m}{j} u^{jr}, \tag{2.107}$$

with the expansions

$$Q_0^{(L)}(u) = \sum_{l=0}^{L} c_l^{(L)} (-u)^l, \quad Q_0^{(M)}(u) = \sum_{k=0}^{M} c_k^{(M)} (-u)^k, \tag{2.108}$$

which follow from (2.3)–(2.6). Since $Q_0^{(M)}(u)$ has degree $M<r$, this yields

$$c_l^{(L)} = 0, \quad l = M+1,\ldots,r-1 \pmod r, \tag{2.109}$$

$$c_{k+jr}^{(L)} = (-)^{jr+j} \binom{m}{j} c_k^{(M)}, \quad k=0,\ldots,M, \quad j=0,\ldots,m. \tag{2.110}$$

Fourth, we use (2.109), (2.110), and (2.106) to write

$$\begin{aligned} \sum_{l=0}^{L} (-t)^l c_l^{(L)} Q_l^{(L)}(u) &= \sum_{k=0}^{M} \sum_{j=0}^{m} (-t)^{k+jr} c_{k+jr}^{(L)} Q_{k+jr}^{(L)}(u) \\ &= (1-u^r)^m \sum_{k=0}^{M} (-t)^k c_k^{(M)} Q_k^{(M)}(u) \sum_{j=0}^{m} (-t^r)^j \binom{m}{j} \\ &= (1-u^r)^m (1-t^r)^m \sum_{k=0}^{M} (-t)^k c_k^{(M)} Q_k^{(M)}(u). \end{aligned} \tag{2.111}$$

This equals (2.103), so (2.99) follows.

To prove (2.100), we begin by noting that when we write

$$Q_j^{(M)}(u) = P^{(M)}(u) R_j^{(M)}(u), \quad P^{(M)}(u) \equiv \prod_{k=r-M}^{M} (1-w^k u), \tag{2.112}$$

then $R_j^{(M)}(u)$ is a polynomial of degree N. Of course, this is plain from (2.6) for $j=0,\ldots,N$, independently of the value of w. Since we have $w^r=1$ in the present case however, the remainder term $R_j^{(M)}(u)$ is still a polynomial for $j=r-M,\ldots,M$.

From (2.102) with $N\to M$ we now deduce that $S_M(u,t)$ is the product of $P^{(M)}(u)$ and a polynomial in u and t. By self-duality (symmetry under $u\leftrightarrow t$) we then must have

$$S_M(u,t) = P^{(M)}(u) P^{(M)}(t) P_N(u,t), \tag{2.113}$$

where $P_N(u,t)$ is a polynomial of degree N in u and t, symmetric under the interchange of u and t. Using (2.113) to rewrite the lhs of (2.100), it now follows from a straightforward calculation that (2.100) amounts to

$$P_N(u,t)=\prod_{j=r-M}^{M} q^j \cdot S_N(u,t). \tag{2.114}$$

Next, we observe that (2.114) holds true for $u=t=0$. [To check this, use $S_K(0,0)=q^{K(K+1)/2}$ and $P^{(M)}(0)=1$.] Thus, we need only show that the polynomials P_N and S_N are proportional. Switching back, this amounts to the quotient function

$$Q_N(x,p)\equiv K_M(x,p) \Big/ \prod_{j=r-M}^{M} [s_-(x+ija_+)s_-(p+ija_+)] \tag{2.115}$$

being proportional to $K_N(x,p)$. We proceed by proving this, making suitable use of the first part of the proof of Theorem II.1.

First, we note that since $K_M(x,p)$ satisfies the AΔE (2.12) with $N\to M$, we must have

$$\begin{aligned} s_-(x+iMa_+)&\prod_{j=r-M}^{M} s_-(x+i(j-1)a_+)\cdot Q_N(x-ia_+,p)\\ &+s_-(x-iMa_+)\prod_{j=r-M}^{M} s_-(x+i(j+1)a_+)\cdot Q_N(x-ia_+,p)\\ &=2s_-(x)c_-(p)\prod_{j=r-M}^{M} s_-(x+ija_+)\cdot Q_N(x,p). \end{aligned} \tag{2.116}$$

When we now divide this by the product on the rhs and use the identity

$$s_-(x-iMa_+)s_-(x+i(M+1)a_+)=s_-(x+i(r-M)a_+)s_-(x+i(M+1-r)a_+), \tag{2.117}$$

then we obtain

$$s_-(x+iNa_+)Q_N(x-ia_+,p)+(i\to -i)=2s_-(x)c_-(p)Q_N(x,p). \tag{2.118}$$

Second, we recall that $K_N(x,p)$ also satisfies the AΔE (2.118). Indeed, we used the general form (2.20)–(2.21) of $K_N(x,p)$ as an Ansatz to arrive at the system of equations $d_{mn}=0$ with side conditions (2.25), and then showed that the coefficients (2.5) solve this system. Now in view of (2.113) $Q_N(x,p)$ has the same general form as $K_N(x,p)$, except that the coefficients of the monomials in $P_N(u,t)$ are as yet unknown. We do know, however, that the coefficient matrix is *symmetric*.

Third, we reconsider the paragraph containing (2.26). Choosing a_+/a_- irrational guaranteed a unique solution for each set of boundary coefficients b_n, $n\in\mathbb{N}$. In the present case, however, a_+/a_- is rational, and we have a symmetric solution $\tilde{c}_{kl}$ arising from $Q_N(x,p)$ on hand. The remaining problem, then, is to show that the latter coefficients equal the symmetric coefficients $c_{kl}^{(N)}$ occurring in $K_N(x,p)$, up to a common factor.

It is not hard to see that this is true. The key point is that we still have $w^m\neq 1$ for $m=1,\ldots,N$. Hence a *symmetric* solution to the system is uniquely determined up to an overall factor. Indeed, starting from a given c_{00}, we can calculate successively $c_{10},c_{20},\ldots,c_{N0}$, since $w^m\neq 1$. But then the boundary coefficients c_{0n} are determined by symmetry. Therefore, the remaining coefficients can be successively calculated (again because $w^m\neq 1$), entailing uniqueness. □

III. THE GENERAL HYPERBOLIC CASE

Just as in the special integer g case studied in Sec. II, it is easy to adapt our results for the general elliptic case (cf. Sec. III in I) to the hyperbolic regime. But the results from Sec. II can actually be exploited to proceed considerably beyond the hyperbolic specialization of Sec. III in I. Indeed, we are going to obtain joint eigenfunctions for all parameters in the space $\mathcal{D}_{\rm hyp}$ (1.7) and for all $p\in\mathbb{C}$. Moreover, for parameters in the subset $\mathcal{D}$ [defined by I(3.33)–I(3.35)], the representation derived below is far more explicit than the zero representation I(3.39).

We have occasion to make extensive use of the results obtained in Sec. II. To prevent ambiguous notation, the function $K_N(x,p)$ (2.20) is henceforth denoted by $K_N(a_+,a_-;x,p)$. We also need *two* ''q-variables,'' viz.,

$$q_+\equiv\exp(i\pi a_+/a_-),\quad q_-\equiv\exp(i\pi a_-/a_+). \tag{3.1}$$

Thus q (1.16) is, from now on, denoted by q_+.

To ease the exposition, we restrict attention to b-values of the form (1.20) until further notice, and, accordingly, study the auxiliary AΔOs B_δ (1.21). We now claim that the functions

$$K_{N_+,N_-}(a_+,a_-;x,p)\equiv\exp(i\pi xp/a_+a_-)\prod_{\delta=+,-}e_\delta(N_\delta[x+p])S_{N_\delta}(q_\delta;e_{-\delta}(-2x),e_{-\delta}(-2p)) \tag{3.2}$$

are joint B_δ-eigenfunctions with eigenvalues $2c_\delta(p)$. Given Theorem II.1, this is quite easily verified: For B_- we can use the identity

$$K_{N_+,N_-}(a_+,a_-;x,p)=K_{N_+}(a_+,a_-;x,p)e_+(N_-[x+p])S_{N_-}(q_-;e_+(-2x),e_+(-2p)), \tag{3.3}$$

whereas for B_+ we can use

$$K_{N_+,N_-}(a_+,a_-;x,p)=K_{N_-}(a_-,a_+;x,p)e_-(N_+[x+p])S_{N_+}(q_+;e_-(-2x),e_-(-2p)). \tag{3.4}$$

The joint eigenfunction property just demonstrated holds true for arbitrary $a_+,a_->0$. Restricting a_+ and a_- by I(3.34) and I(3.35), respectively, we also obtain a joint B_δ-eigenfunction $\mathcal{H}(x,y)$ I(3.39) in a quite different guise. Again, from Sec. II the connection between the two representations is easily established: One has

$$K_{N_+,N_-}(a_+,a_-;x,p)=\prod_{\delta=+,-}\left(\prod_{\pm j=1}^{N_\delta}\frac{4\pi}{a_{-\delta}}\sinh\frac{\pi}{a_{-\delta}}(p+ija_\delta)\right)^{1/2}\cdot\mathcal{H}(x,\pi p/a_+a_-). \tag{3.5}$$

[To see this, note first of all that I(3.17) becomes $y=u$ in the hyperbolic case. Canceling the plane wave $K_0(x,p)$ in the relation (2.60), the resulting formula readily yields (3.5).]

Next, we observe that the Casorati determinants

$$D^\delta_{N_+,N_-}(x)\equiv K_{N_+,N_-}(a_+,a_-;x+ia_\delta/2,p)K_{N_+,N_-}(a_+,a_-;x-ia_\delta/2,-p)-(i\to -i) \tag{3.6}$$

can be explicitly determined from (2.63)–(2.66) by using (3.3)/(3.4) for $\delta=+/-$. This yields

$$D^\delta_{N_+,N_-}(x)=(-)^{N_\delta+1}\prod_{k=-N_\delta}^{N_\delta}2s_{-\delta}(p-ika_\delta)\cdot\prod_{l=-N_\delta+1/2}^{N_\delta-1/2}2s_{-\delta}(x-ila_\delta)$$
$$\cdot e_\delta(2N_{-\delta}x)\prod_{\alpha=+,-}S_{N_{-\delta}}(q_{-\delta};-e_\delta(-2x),e_\delta(2\alpha p)),\quad \delta=+,-. \tag{3.7}$$

Furthermore, it follows as before that the determinants do not vanish identically for $\mathrm{Re}\, p \neq 0$. Adapting Theorem B.1 in I, we infer that for a_+/a_- irrational and $\mathrm{Re}\, p > 0$ the functions $K_{N_+,N_-}(a_+,a_-;x,\pm p)$ form a basis for the joint eigenspace of B_+ and B_- corresponding to eigenvalues $2c_+(p)$ and $2c_-(p)$, respectively.

Next, we recall from Sec. III in I that for points $(a_+,a_-,b) \in \mathcal{D}$ there is at most one way to write b as $(N_+ + 1)a_+ - N_- a_-$ with $N_+, N_- \in \mathbb{N}$ [cf. the paragraph containing I(3.29)]. Returning to the general case $a_+,a_- > 0$, this is no longer true, of course. In particular, let us choose

$$\frac{a_+}{a_-} = \frac{n_-}{n_+}, \quad n_+, n_- \in \mathbb{N}^*, \quad n_+, n_- \text{ coprime.} \tag{3.8}$$

Then we may rewrite (1.20) as

$$b_{+-} = (N_+ + 1 + m n_+)a_+ - (N_- + m n_-)a_-, \tag{3.9}$$

where m is an arbitrary integer.

Choosing $m \in \mathbb{N}$, we now deduce from the identity (2.99) that we have

$$K_{N_+ + mn_+, N_- + mn_-}(a_+,a_-;x,p) = \eta_{N_+,N_-}(m) \prod_{\delta=+,-} [4 s_{-\delta}(n_\delta x) s_{-\delta}(n_\delta p)]^m \cdot K_{N_+,N_-}(a_+,a_-;x,p), \tag{3.10}$$

where $\eta \in \{\pm 1, \pm i\}$ is given by

$$\eta_{N_+,N_-}(m) \equiv (-)^{mN_+n_- + mN_-n_+ + m^2 n_+ n_-} \cdot i^{mn_- + mn_+}. \tag{3.11}$$

All of the functions on the rhs of (3.10) are manifestly independent, so we wind up with an infinity of joint eigenfunctions for the same b-value!

We proceed by connecting the ambiguity just uncovered to the interpolation question discussed below I(4.8). As we have seen there, we get distinct weight functions $\hat{w}(x)$ for distinct $m \in \mathbb{N}$; cf. I(4.6). Moreover, in the elliptic case the AΔOs B_δ also depend on the choice of m. But as we have already detailed in the Introduction, the hyperbolic counterparts (1.21) do admit the continuous interpolation $B_\delta(b)$ (1.22). [A caveat is in order at this point: For b of the form $(N_- + 1)a_- - N_+ a_+$ one would need a *different* interpolation. Specifically, one must take $b \to a_+ + a_- - b$ on the rhs of (1.22) in that case.]

This fact leads to a remarkable conclusion of a general character that we wish to emphasize before we discard the auxiliary AΔOs B_δ in favor of the AΔOs $A_\delta(b)$ (1.1) and $H_\delta(b)$ (1.5), which are defined for arbitrary real b to begin with. Indeed, since the functions (3.10) are independent for different $m \in \mathbb{N}$, we may deduce that the commuting AΔO pair $B_\delta(b)$ (1.22) *does not admit* joint eigenfunctions depending continuously on the parameters, already for parameters $a_+,a_- > 0$ and b of the form (1.20). [In virtue of the specialization of Theorem B.1 in I, the ambiguity (3.10) is inescapable.]

This shows by example that the existence of interpolations cannot follow from general arguments. It is all the more remarkable that for the AΔOs $H_\delta(b)$ (1.5) [and hence for $A_\delta(b)$ (1.1), too] the interpolation ambiguity disappears: The ambiguity in the joint B_δ-eigenfunctions is canceled by the ambiguity in the auxiliary weight function $\hat{w}(x)$.

To detail this, we first introduce the renormalized weight function

$$\hat{w}_{N_+,N_-}(a_+,a_-;x) \equiv 1 \Big/ \prod_{\delta=+,-} \prod_{\pm j=1}^{N_\delta} \left[2 \sinh \frac{\pi}{a_{-\delta}}(x + i j a_\delta) \right]. \tag{3.12}$$

[It differs from the hyperbolic specialization of $\hat{w}(x)$ I(4.6) by a multiplicative constant.] With the rationality assumption (3.8) in effect, it satisfies

$$\hat{w}_{N_++mn_+,N_-+mn_-}(a_+,a_-;x)=\prod_{\delta=+,-}[2s_{-\delta}(n_\delta x)]^{-2m}\cdot\hat{w}_{N_+,N_-}(a_+,a_-;x). \tag{3.13}$$

[Indeed, this comes down to the identity (2.104).]

Consider now the functions

$$F_{N_+,N_-}(a_+,a_-;x,p)$$
$$\equiv\phi_{N_+,N_-}[\hat{w}_{N_+,N_-}(a_+,a_-;x)\hat{w}_{N_+,N_-}(a_+,a_-;p)]^{1/2}K_{N_+,N_-}(a_+,a_-;x,p), \tag{3.14}$$

where ϕ is the phase

$$\phi_{N_+,N_-}\equiv(-i)^{2N_+N_-+N_++N_-+1}. \tag{3.15}$$

By construction, they are joint eigenfunctions of the AΔOs $H_\delta((N_++1)a_+-N_-a_-)$ with eigenvalues $2c_\delta(p)$. The phase satisfies

$$\phi_{N_++mn_+,N_-+mn_-}=\phi_{N_+,N_-}\overline{\eta}_{N_+,N_-}(m), \tag{3.16}$$

so with (3.8) in force one deduces the equality

$$F_{N_++mn_+,N_-+mn_-}(a_+,a_-;x,p)=F_{N_+,N_-}(a_+,a_-;x,p),\quad m\in\mathbb{N}. \tag{3.17}$$

Hence the ambiguities cancel out, as announced.

It should be noted that the definition (3.14) preserves the symmetry under $x\leftrightarrow p$. Moreover, it entails that we have

$$F_{N_+,N_-}(a_+,a_-;x,p)=\mathcal{F}(x,\pi p/a_+a_-), \tag{3.18}$$

where $\mathcal{F}(x,y)$ is the hyperbolic specialization of I(3.31). Indeed, equality up to phase follows via (3.5), so we need only verify that the phase of the normalization constant $\mathcal{N}$ in the c-function I(1.25) equals ϕ_{N_+,N_-} (3.15). Now from Proposition III.8 in Ref. 2 we easily calculate

$$c(a_+,a_-,(N_++1)a_+-N_-a_-;x)=\phi_{N_+,N_-}\frac{\Pi_{k=1}^{N_-}2s_+(x+ika_-)}{\Pi_{j=0}^{N_+}2s_-(x-ija_+)}. \tag{3.19}$$

Hence the phase $\phi(\mathcal{N})$ in I(3.31) indeed equals (3.15) in the hyperbolic case. (In fact, it is not hard to see that this is still true in the elliptic case.)

Thus far we have assumed b-values of the form (1.20). Let us next assume b-values of the form

$$b_{-+}=-N_+a_++(N_-+1)a_-,\quad N_+,N_-\in\mathbb{N}. \tag{3.20}$$

Rewriting $H_\delta(b)$ (1.5) as

$$H_\delta(b)=\left(\frac{s_\delta(x-ib)s_\delta(x+ib-ia_{-\delta})}{s_\delta(x)s_\delta(x-ia_{-\delta})}\right)^{1/2}T_{ia_{-\delta}}+(i\rightarrow-i),\quad\delta=+,-, \tag{3.21}$$

we read off the symmetry property

$$H_\delta(b)=H_\delta(a_++a_--b). \tag{3.22}$$

Thus, we may and will choose as joint eigenfunctions of $H_+(b_{-+})$ and $H_-(b_{-+})$ the functions $F_{N_+,N_-}(a_+,a_-;x,p)$ just defined.

More generally, we obtain the same AΔO pair $H_\delta(b)$ for the four b-values in the set

$$B_{N_+,N_-}\equiv\{b_{+-},b_{-+},b_{--},b_{++}\}, \tag{3.23}$$

where we use the notation (1.20), (3.20), and

$$b_{--}\equiv -N_+a_+-N_-a_-, \quad b_{++}\equiv(N_++1)a_++(N_-+1)a_-. \tag{3.24}$$

[Once more, this can be read off from (3.21).] But then we have

$$H_\delta(b)F_{N_+,N_-}(a_+,a_-;x,p)=2c_\delta(p)F_{N_+,N_-}(a_+,a_-;x,p), \quad b\in B_{N_+,N_-}. \tag{3.25}$$

Hence, we have now constructed joint eigenfunctions for all parameters in $\mathcal{D}_{\mathrm{hyp}}$ (1.7), as advertised in the Introduction.

But more can and should be said. In particular, for the rational case (3.8) we have shown the absence of ambiguity for positive m in (1.20), but, of course, we can just as well choose m equal to a negative integer. As long as N_++mn_+ and N_-+mn_- are non-negative, it is clear one still obtains (3.17). But when one of these integers becomes negative, the state of affairs is quite unclear at this stage. The next theorem supplies, in particular, the information that will enable us to unambiguously define a joint H_δ-eigenfunction $F(a_+,a_-,b;x,p)$ for all (a_+,a_-,b) in $\mathcal{D}_{\mathrm{hyp}}$. But it also yields additional information about the rational case (3.8) that is of interest in itself.

Theorem III.1: *The function* $F_{N_+,N_-}(a_+,a_-;x,p)$ *(3.14) satisfies*

$$F_{N_+,N_-}(a_+,a_-;x,p)=F_{N_-,N_+}(a_-,a_+;x,p). \tag{3.26}$$

Now assume (3.8). Fixing $N_+,N_-\in\mathbb{N}$, *one has*

$$F_{N_++m_+n_+,N_-+m_-n_-}(a_+,a_-;x,p)=\zeta_{N_+,N_-}(m_+,m_-)F_{N_+,N_-}(a_+,a_-;x,p), \tag{3.27}$$

$$\zeta_{N_+,N_-}(m_+,m_-)\equiv(-)^{(m_+-m_-)(N_-n_+-N_+n_-)}\cdot i^{(m_+-m_-)[n_--n_++(m_+-m_-)n_+n_-]}, \tag{3.28}$$

where m_+ *and* m_- *are integers such that* $N_\delta+m_\delta n_\delta\geqslant 0$, $\delta=+,-$. *Moreover, choosing* $N_+\in[0,n_+/2)$, *one has*

$$F_{M_+,N_-}(a_+,a_-;x,p)=\xi_{N_+,N_-}F_{N_+,N_-}(a_+,a_-;x,p), \quad M_+\equiv n_+-1-N_+, \tag{3.29}$$

$$\xi_{N_+,N_-}\equiv(-)^{(n_+-1)N_-+(n_--1)N_+}\cdot i^{n_+n_--n_--n_++1}. \tag{3.30}$$

Proof: The symmetry property (3.26) can be read off from the definitions (3.14), (3.15), (3.12), and (3.2). To prove (3.27), we first note that (3.13) generalizes as

$$\hat{w}_{N_++m_+n_+,N_-+m_-n_-}(a_+,a_-;x)=\prod_{\delta=+,-}[2s_{-\delta}(n_\delta x)]^{-2m_\delta}\cdot\hat{w}_{N_+,N_-}(a_+,a_-;x). \tag{3.31}$$

Second, we can use the identity (2.99) once more to generalize (3.10). A straightforward calculation yields

$$K_{N_++m_+n_+,N_-+m_-n_-}(a_+,a_-;x,p)$$
$$=\eta_{N_+,N_-}(m_+,m_-)\prod_{\delta=+,-}[4s_{-\delta}(n_\delta x)s_{-\delta}(n_\delta p)]^{m_\delta}\cdot K_{N_+,N_-}(a_+,a_-;x,p), \tag{3.32}$$

$$\eta_{N_+,N_-}(m_+,m_-)\equiv(-)^{m_+N_+n_-+m_-N_-n_+}\cdot i^{[m_+n_-+m_-n_++(m_+^2+m_-^2)n_+n_-]}. \tag{3.33}$$

Third, (3.16) generalizes to

$$\phi_{N_++m_+n_+,N_-+m_-n_-}=\phi_{N_+,N_-}\zeta_{N_+,N_-}(m_+,m_-)\overline{\eta}_{N_+,N_-}(m_+,m_-). \tag{3.34}$$

Combining these relations, we obtain (3.27).

In order to prove (3.29), we note first

$$\hat{w}_{M_+,N_-}(a_+,a_-;x)=\prod_{\pm j=N_++1}^{n_+-1-N_+}[2s_-(x+ija_+)]^{-1}\cdot\hat{w}_{N_+,N_-}(a_+,a_-;x). \tag{3.35}$$

Second, we exploit (2.100) to write

$$K_{M_+,N_-}(a_+,a_-;x,p)=\prod_{j=N_++1}^{n_+-1-N_+}[4q_+^{-j}s_-(x+ija_+)s_-(p+ija_+)]\cdot K_{N_+,N_-}(a_+,a_-;x,p). \tag{3.36}$$

Consider now the function

$$Q(x)\equiv\frac{\Pi_{j=N_++1}^{n_+-1-N_+}q_+^{-j}s_-(x+ija_+)}{[\Pi_{\pm j=N_++1}^{n_+-1-N_+}s_-(x+ija_+)]^{1/2}}. \tag{3.37}$$

From the identity

$$s_-(x-ija_+)=(-)^{n_-}s_-(x+i(n_+-j)a_+), \tag{3.38}$$

we deduce that $Q(x)$ equals a phase, so taking $x\to\infty$ we obtain $Q(x)=1$. Hence (3.14) yields

$$F_{M_+,N_-}(a_+,a_-;x,p)=\phi_{M_+,N_-}\overline{\phi}_{N_+,N_-}\prod_{j=N_++1}^{n_+-1-N_+}q_+^j\cdot F_{N_+,N_-}(a_+,a_-;x,p). \tag{3.39}$$

Calculating the phase yields the rhs of (3.30), so (3.29) follows. □

Still assuming (3.8), this theorem shows that the vector space spanned by the functions F_{M_+,M_-}, $M_+,M_-\in\mathbb{N}$, is finite-dimensional: It is already spanned by the functions F_{N_+,N_-} with $N_\delta\in[0,n_\delta/2)$, $\delta=+,-$. Indeed, all of the former functions are phase multiples of the latter, as follows by combining (3.27), (3.29), and (3.26). This fact is in accordance with (but not implied by) the relation

$$H_\delta(b+mn_+a_+)=H_\delta(b+mn_-a_-)=H_\delta(b),\quad m\in\mathbb{Z},\quad \delta=+,-, \tag{3.40}$$

whose validity is clear from (3.21).

More importantly, the theorem enables us to dispose of the $m\in\mathbb{Z}$ ambiguity in (3.9) and its b_{-+}-analog. Specifically, taking $N_+,N_-\in\mathbb{N}$, we set [recall (1.20), (3.20), and (3.24)]

$$F(a_+,a_-,b_{+-};x,p)\equiv F_{N_+,N_-}(a_+,a_-;x,p), \tag{3.41}$$

$$F(a_+,a_-,b_{-+};x,p)\equiv F_{N_+,N_-}(a_+,a_-;x,p), \tag{3.42}$$

$$F(a_+,a_-,b_{--};x,p)\equiv\tilde{F}_{N_+,N_-}(a_+,a_-;x,p), \tag{3.43}$$

$$F(a_+,a_-,b_{++};x,p)\equiv\tilde{F}_{N_+,N_-}(a_+,a_-;x,p), \tag{3.44}$$

where

$$\tilde{F}_{N_+,N_-}(a_+,a_-;x,p) \equiv \chi_{N_+,N_-} F_{N_+,N_-}(a_+,a_-;x,p), \tag{3.45}$$

$$\chi_{N_+,N_-} \equiv (-)^{N_+ + N_-} \cdot i. \tag{3.46}$$

Of course, we are free to do so for $a_+/a_- \notin \mathbb{Q}$, since then all b-values $ka_+ + la_-$, $k,l \in \mathbb{Z}$, are distinct. But our task is now to show that for the rational case (3.8) the function $F(a_+,a_-,b;x,p)$ is still well defined.

Now we have already seen that (3.41) by itself is a legitimate definition; cf. (3.17). In view of the symmetry property (3.26), this is true for (3.42) as well. For (3.43) and (3.44) to be well defined by themselves, we should have

$$\tilde{F}_{N_+ + mn_+,N_-}(a_+,a_-;x,p) = \tilde{F}_{N_+,N_- + mn_-}(a_+,a_-;x,p), \quad m \in \mathbb{N}. \tag{3.47}$$

Recalling (3.27), we see that this amounts to

$$\zeta_{N_+,N_-}(m,0)\chi_{N_+ + mn_+,N_-} = \zeta_{N_+,N_-}(0,m)\chi_{N_+,N_- + mn_-}, \tag{3.48}$$

which is easily verified. To prove the compatibility of (3.41) and (3.43), we need to show that when $M_+ \in [n_+/2, n_+ - 1]$, then we have

$$F_{M_+,N_- + n_-}(a_+,a_-;x,p) = \tilde{F}_{N_+,N_-}(a_+,a_-;x,p), \quad N_+ \equiv n_+ - 1 - M_+. \tag{3.49}$$

Combining (3.27) and (3.29), we deduce that this amounts to the relation

$$\zeta_{M_+,N_-}(0,1)\xi_{N_+,N_-} = \chi_{N_+,N_-}, \quad M_+ = n_+ - 1 - N_+. \tag{3.50}$$

The phase χ_{N_+,N_-} obeys this relation (indeed, it is defined such that it does), so (3.49) follows. The remaining compatibilities can now be handled by using (3.26). Thus, the function $F(\Xi;x,p)$ is well defined for all parameters $\Xi = (a_+,a_-,b)$ in $\mathcal{D}_{\text{hyp}}$ (1.7).

We proceed by summarizing some salient features of the function $F(\Xi;x,p)$.

Theorem III.2: *For all $\Xi \in \mathcal{D}_{hyp}$ the definition (3.41)–(3.44) gives rise to a well-defined, generically two-valued, analytic function $F(\Xi;x,p)$ with a meromorphic square. It satisfies*

$$H_\delta F(\Xi;x,p) = 2c_\delta(p)F(\Xi;x,p), \quad \delta = +,-, \tag{3.51}$$

and has parameter and variable symmetries

$$F(a_+,a_-,b;x,p) = F(a_+,a_-,a_+ + a_- - b;x,p), \tag{3.52}$$

$$F(a_+,a_-,b;x,p) = F(a_-,a_+,b;x,p), \tag{3.53}$$

$$F(\Xi;x,p) = F(\Xi;p,x), \tag{3.54}$$

$$F(\Xi;x,p) = F(\Xi;-x,-p). \tag{3.55}$$

Now denote by F_r the function defined for $x,p > 0$ by taking positive square roots in (3.14). For $a_+/a_- \notin \mathbb{Q}$ this function has a real-analytic extension F_r to $x,p \in \mathbb{R}$, which satisfies

$$F_r(a_+,a_-,b;x,p) = -\overline{F_r}(a_+,a_-,b;-x,p), \quad b = b_{+-}, b_{-+}, \tag{3.56}$$

$$F_r(a_+,a_-,b;x,p) = \overline{F_r}(a_+,a_-,b;-x,p), \quad b = b_{--}, b_{++}. \tag{3.57}$$

Proof: It remains to prove (3.54)–(3.57). By virtue of (3.3) and (3.4), the holomorphic function (3.2) satisfies

$$K_{N_+,N_-}(a_+,a_-;x,p)=K_{N_+,N_-}(a_+,a_-;p,x), \tag{3.58}$$

$$K_{N_+,N_-}(a_+,a_-;x,p)=K_{N_+,N_-}(a_+,a_-;-x,-p), \tag{3.59}$$

$$K_{N_+,N_-}(a_+,a_-;x,p)=(-)^{N_++N_-}\overline{K}_{N_+,N_-}(a_+,a_-;-x,p), \quad x,p\in\mathbb{R}. \tag{3.60}$$

[Recall (2.13)–(2.15).] In view of (3.14), this entails

$$F_{N_+,N_-}(a_+,a_-;x,p)=F_{N_+,N_-}(a_+,a_-;p,x), \tag{3.61}$$

$$F_{N_+,N_-}(a_+,a_-;x,p)=F_{N_+,N_-}(a_+,a_-;-x,-p), \tag{3.62}$$

so (3.54) and (3.55) follow. For a_+/a_- irrational, the auxiliary weight function (3.12) has a real-analytic, positive, and even restriction to $\mathbb{R}$, so (3.56) and (3.57) follow from (3.60) and the phase definitions (3.15) and (3.46). □

Of course, for a_+/a_- rational, the restriction F_r is still real-analytic for $x,p>0$. But in that case the weight function (3.12) may have poles at the origin, so that ambiguities can arise for $x<0$. (Taking a real-analytic restriction to $\mathbb{R}$ and taking parameter limits need not commute; we mention the function $x\mapsto(x^2+\epsilon^2)^{-1/2}$ to exemplify this difficulty.)

Such square-root subtleties are not present for the meromorphic joint A_δ-eigenfunction

$$M(\Xi;x,p)\equiv[w(\Xi;x)w(\Xi;p)]^{-1/2}F(\Xi;x,p), \quad \Xi\in\mathcal{D}_{\text{hyp}}, \tag{3.63}$$

which we study next. From Ref. 2 Eq. (5.21), we have

$$w(a_+,a_-,k_+a_++k_-a_-;x)$$
$$=\prod_{\delta=+,-}\prod_{j=1}^{|k_\delta|}\left(\left[2\sinh\frac{\pi}{a_{-\delta}}(x+ia_\delta(j_\delta-\theta(k_\delta)))\right][i\to -i]\right)^{k_\delta/|k_\delta|}, \quad k_\delta\in\mathbb{Z}. \tag{3.64}$$

[Here, $\theta(k)=1$ for $k>0$ and $\theta(k)=0$ for $k<0$.] Using (3.14) and (3.41)–(3.46), this yields the explicit formulas

$$M(a_+,a_-,b_{\alpha,-\alpha};x,p)=(-i)^{2N_+N_-+N_++N_-+1}[P_{N_\alpha}(a_\alpha,a_{-\alpha};x)P_{N_\alpha}(a_\alpha,a_{-\alpha};p)]^{-1}$$
$$\cdot K_{N_+,N_-}(a_+,a_-;x,p), \quad \alpha=+,-, \tag{3.65}$$

$$M(a_+,a_-,b_{--};x,p)=i^{2N_+N_-+N_++N_-}K_{N_+,N_-}(a_+,a_-;x,p), \tag{3.66}$$

$$M(a_+,a_-,b_{++};x,p)=i^{2N_+N_-+N_++N_-}\left[\prod_{\alpha=+,-}P_{N_\alpha}(a_\alpha,a_{-\alpha};x)P_{N_\alpha}(a_\alpha,a_{-\alpha};p)\right]^{-1}$$
$$\cdot K_{N_+,N_-}(a_+,a_-;x,p), \tag{3.67}$$

where

$$P_N(a_+,a_-;x)\equiv\prod_{j=-N}^{N}\left[2\sinh\frac{\pi}{a_-}(x+ija_+)\right]. \tag{3.68}$$

Theorem III.3: *The meromorphic function* $M(\Xi;x,p)$, $\Xi\in\mathcal{D}_{\text{hyp}}$ (1.7), *satisfies*

$$A_\delta M(\Xi;x,p)=2c_\delta(p)M(\Xi;x,p), \quad \delta=+,-, \tag{3.69}$$

and has parameter and variable symmetries

$$M(a_+,a_-,b;x,p)=M(a_-,a_+,b;x,p), \tag{3.70}$$

$$M(\Xi;x,p)=M(\Xi;p,x), \tag{3.71}$$

$$M(\Xi;x,p)=M(\Xi;-x,-p), \tag{3.72}$$

$$M(\Xi;x,p)=\overline{M}(\Xi;-x,p), \quad x,p\in\mathbb{R}. \tag{3.73}$$

Proof: The asserted properties are clear from the definition of M and from (3.58)–(3.60). [Recall that $w(a_+,a_-,b;x)$ is symmetric under $a_+\leftrightarrow a_-$.] □

We do not know whether the function $M(\Xi;x,p)$ admits an interpolation to all of the hyperbolic parameter domain $\mathcal{H}$ (1.4). But for the even joint A_δ-eigenfunction,

$$R(\Xi;x,p)\equiv M(\Xi;x,p)+M(\Xi;-x,p), \quad \Xi\in\mathcal{D}_{\rm hyp}, \tag{3.74}$$

this is the case (see Ref. 4 and papers to appear). Observe that the latter function already appeared in the integer g case: One has

$$R(a_+,a_-,(N+1)a_+;x,p)=R_N(x,p), \tag{3.75}$$

where $R_N(x,p)$ is given by (2.71). [To check this, use (3.65) with $\alpha=+$, $N_+=N$ and $N_-=0$.]

To conclude this section, let us add one more observation on the auxiliary AΔOs B_δ (1.21). Since they are only defined for b of the form (1.20), we may specify their b-dependence by writing $B_\delta(N_+,N_-)$. Comparing (1.21) and (1.1), we now deduce

$$B_\delta(N_+,N_-)=A_\delta(-N_+a_+-N_-a_-), \quad \delta=+,-. \tag{3.76}$$

This coincidence agrees with (3.66). Indeed, the latter formula says that the joint $A_\delta(b)$-eigenfunction M for $b=-N_+a_+-N_-a_-$ is proportional to the joint $B_\delta(N_+,N_-)$-eigenfunction K_{N_+,N_-}. (See also the remarks at the end of Sec. IV in I, specialized to the hyperbolic case.)

IV. THE TRIGONOMETRIC SPECIALIZATION

At the end of the Introduction we have already delineated how various objects from the elliptic regime studied in I give rise to trigonometric counterparts. We will use the corresponding formulas (1.23)–(1.28) without further comment.

Until further notice, we restrict attention to the special choice $k=N+1\in\mathbb{N}^*$ in (1.27). Then the results in Sec. II of I can be readily specialized, giving rise to functions $\Psi(\pm x,y)$ that are joint eigenfunctions of

$$A=\frac{\sin r(x-i(N+1)\beta)}{\sin rx}T_{i\beta}+(i\rightarrow -i) \tag{4.1}$$

[the AΔO (1.24) for $b=(N+1)\beta$)] and Q (1.28). [Indeed, the relevant trigonometric function $E(x)$ I(2.8) has finite and equal limits for $\mathrm{Im}\,x\rightarrow\pm\infty$, entailing constancy.]

Comparing the trigonometric AΔO A (4.1) to its hyperbolic counterpart $A_-((N+1)a_+)$ (1.1), one sees they are related via the substitutions

$$a_+\rightarrow\beta, \quad a_-\rightarrow\pi/ir. \tag{4.2}$$

Moreover, these substitutions turn the second hyperbolic AΔO $A_+((N+1)a_+)$ (1.1) into $(-)^{N+1}Q$. Therefore, the joint $A_\delta((N+1)a_+)$-eigenfunctions from Sec. II can be exploited to obtain (A,Q)-eigenfunctions. (The latter will be shown to be essentially equal to those arising from the trigonometric specialization of Sec. II in I.)

Once more, we find it expedient to study first the pertinent eigenfunctions of the similarity transformed AΔO

$$B \equiv \prod_{j=-N}^{N} \sin r(x+ij\beta) \cdot A \cdot \prod_{j=-N}^{N} \sin r(x+ij\beta)^{-1} = \frac{\sin r(x+iN\beta)}{\sin rx} T_{i\beta} + (i \to -i). \quad (4.3)$$

While translating our results from Sec. II to trigonometric analogs, we retain the spectral variable y from Sec. II in I. As will soon become clear, this can be achieved by combining the substitutions (4.2) with

$$\pi p/a_- \to \beta(y+(N+1)r). \quad (4.4)$$

Equivalently, we can anticipate the relation to *loc. cit.* by taking

$$p \to \beta y_N/ir, \quad (4.5)$$

where we have set

$$y_N \equiv y+(N+1)r. \quad (4.6)$$

With the above substitutions in the hyperbolic (B_+, B_-)-eigenfunctions $K_N(a_+, a_-; x, p)$ (1.15), we obtain the trigonometric counterparts

$$L_N(r,\beta;x,y) \equiv K_N(\beta, \pi/ir; x, \beta y_N/ir). \quad (4.7)$$

More specifically, this yields

$$L_N(x,y) = \exp(ixy_N) \sum_{k,l=0}^{N} c_{kl}^{(N)}(q_t) \exp[i(N-2k)rx + (N-2l)\beta y_N]. \quad (4.8)$$

Here, we are using

$$q_t \equiv \exp(-\beta r), \quad (4.9)$$

to avoid confusion with the phase factor q (1.16), and the coefficients are defined by (2.2)–(2.5). Notice that in the present case all of the coefficients are real numbers, so that (4.8) entails

$$\overline{L_N}(x,y) = L_N(-x,y), \quad x,y \in \mathbb{R}. \quad (4.10)$$

In view of our hyperbolic result (2.1) (proved in Theorem II.1), we have

$$BL_N(x,y) = 2\cosh(\beta[y+(N+1)r])L_N(x,y). \quad (4.11)$$

Also, (1.19) translates into

$$QL_N(x,y) = -2\cos(\pi y/r)L_N(x,y). \quad (4.12)$$

[Just as (1.19), this is immediate from (4.8), of course.] Likewise, the dual eigenfunction properties (2.48)–(2.51) become

$$\tilde{B}L_N(x,y) = 2\cos(rx)L_N(x,y), \quad (4.13)$$

$$\tilde{Q}L_N(x,y)=(-)^N 2\cosh(\pi x/\beta)L_N(x,y), \tag{4.14}$$

where

$$\tilde{B}\equiv\frac{\sinh(\beta[y+(2N+1)r])}{\sinh(\beta[y+(N+1)r])}\tilde{T}_r+\frac{\sinh(\beta[y+r])}{\sinh(\beta[y+(N+1)r])}\tilde{T}_{-r}, \tag{4.15}$$

$$\tilde{Q}\equiv\tilde{T}_{i\pi/\beta}+\tilde{T}_{-i\pi/\beta}, \tag{4.16}$$

with

$$(\tilde{T}_\alpha G)(y)\equiv G(y-\alpha), \quad \alpha\in\mathbb{C}. \tag{4.17}$$

[Again, (4.14) is plain from (4.8).]

We can deduce a few more salient features from Theorem II.1. First, combining (4.7) with (2.14) and (4.10), we obtain

$$L_N(x,-y-2(N+1)r)=\overline{L_N}(x,y), \quad x,y\in\mathbb{R}. \tag{4.18}$$

Second, from (4.7) and (2.16) we infer

$$L_N(x,-r)=L_N(x,-(2N+1)r)=(-2)^N\prod_{k=N+1}^{2N}\sinh(k\beta r). \tag{4.19}$$

Finally, (4.7) and (2.18) entail

$$L_N(x,-(l+1)r)=L_N(x,-(2N+1-l)r)=C_l^{(N)}(\cos(rx)), \quad l=0,\ldots,N. \tag{4.20}$$

Here, $C_l^{(N)}(u)$ is a polynomial of degree l and parity $(-)^l$. Moreover, this polynomial has real coefficients in view of (4.18), and one has

$$BC_l^{(N)}(\cos(rx))=2\cosh((N-l)\beta r)C_l^{(N)}(\cos(rx)), \quad l=0,\ldots,N, \tag{4.21}$$

due to (4.11).

We proceed by obtaining the relation between $L_N(x,y)$ and the function $\mathcal{H}_N(x,y)$ I(2.34), specialized to the trigonometric context. To this end we exploit the arguments leading from (2.52) to (2.60). Specifically, (2.52) and (2.53) remain true when q is replaced by q_t. Assuming $t\neq q_t^{-2j}$, $j=1,\ldots,N$, one obtains (2.54)–(2.56). Hence the roots $\rho_j(q_t,t)$ are nonzero for $t\neq q_t^{2j}$, $j=1,\ldots,N$, and ρ_j can be chosen equal to q_t^{2j} for $t=0$.

In the case at hand, we need $t=\exp(-2\beta y_N)$, which entails

$$\prod_{j=1}^{N}\rho_j=\prod_{j=1}^{N}\frac{\sinh(\beta[y+jr])}{\sinh(\beta[y+(N+1+j)r])}. \tag{4.22}$$

Therefore, restricting attention to $\{\mathrm{Re}\, y>-r\}$, we may set

$$\rho_j=\exp(2irz_j), \quad z_j(y)\to ij\beta, \quad y\to\infty, \quad j=1,\ldots,N. \tag{4.23}$$

Now I(2.25) yields no restriction on $\beta=-iv$, since $a=\infty$ in the trigonometric regime. Hence it follows that the functions $z_j(y)$ thus defined may be identified with the zero functions $z_j(y)$ from Sec. II of I and that the desired relation reads

$$L_N(x,y)=(4ir)^N\prod_{j=1}^{N}[\sinh(\beta[y+jr])\sinh(\beta[y+(N+1+j)r])]^{1/2}\cdot\mathcal{H}_N(x,y),\quad y\in(K,\infty).\tag{4.24}$$

From the trigonometric specialization of I(2.45)–I(2.46) we also deduce that all of the zeros $z_j(y)$ are nonconstant.

It follows from Sec. II of I that all of the zeros $z_1(y),\ldots,z_N(y)$ belong to $i(0,\infty)$ for $y\in(R,\infty)$ and R large. For $N=1$ this is easily seen to be true for all $y\in(-r,\infty)$; cf. (4.22). But just as in the hyperbolic case, already for $N=2$ and a suitable choice of βr, the zeros move off the imaginary axis as y decreases from ∞ to 0, showing once more that the parameter K is necessary.

To see this phenomenon happen, we use (2.9) to calculate (note $y=0$ corresponds to $t=q_t^6$ for $N=2$)

$$S_2(q;s,q^6)=(q^9-q^7-q^5+q^3)P_2(q;s),\tag{4.25}$$

$$P_2(q;s)\equiv 1-(q^4+q^2+1+q^{-2}+q^{-4})s+(q^6+q^4+2q^2+2+2q^{-2}+q^{-4}+q^{-6})s^2.\tag{4.26}$$

Taking $q\to q_t$ and letting $q_t\uparrow 1$, one gets $P_2\to 1-5s+10s^2$. Since the limit polynomial has nonreal roots, it follows that the numbers $z_1(0),z_2(0)$ are not purely imaginary for βr small enough. [Recall that we need $s=\exp(-2irx)$ in the present case.]

Next, we calculate the Casorati determinant

$$D_N(x)\equiv L_N(x+i\beta/2,y)L_N(-x+i\beta/2,y)-(\beta\to-\beta).\tag{4.27}$$

The argument in Sec. II of I leading to I(2.43) is easily adapted, yielding

$$D_N(x)=\gamma_N(y)\prod_{n=-N+1/2}^{N-1/2}\sin r(x+in\beta).\tag{4.28}$$

Using (4.8) with $\operatorname{Im}x\to\infty$, we now obtain

$$\gamma_N(y)=(-2i)^{2N}[\exp(-\beta y_N)-\exp(\beta y_N)]\cdot\left(\sum_{l=0}^{N}c_{Nl}^{(N)}(q_t)\exp[(N-2l)\beta y_N]\right)\left(\sum_{l=0}^{N}c_{0l}^{(N)}(q_t)\exp[(N-2l)\beta y_N]\right).\tag{4.29}$$

From symmetry of the coefficients and (2.4)–(2.6) we then infer

$$\begin{aligned}\gamma_N(y)&=(-)^{N+1}2^{2N+1}\sinh(\beta y_N)\exp(2N\beta y_N)(-)^N Q_N(e^{-2\beta y_N})Q_0(e^{-2\beta y_N})\\&=-2^{4N+1}\prod_{j=1}^{2N+1}\sinh(\beta y+j\beta r).\end{aligned}\tag{4.30}$$

In view of the relation (4.24), it follows that in the trigonometric case the quantity α_N in I(2.43) becomes

$$\alpha_N=(-)^{N+1}2\sinh(\beta y+(N+1)\beta r).\tag{4.31}$$

From (4.30) we read off that $L_N(x,y)/L_N(-x,y)$ is not $i\beta$-periodic in x, unless y equals $y_{jk}\equiv -jr+i\pi k/\beta$ with $j=1,\ldots,2N+1$ and $k\in\mathbb{Z}$; in the latter case we readily obtain

$$L_N(x,y_{jk})/L_N(-x,y_{jk})=\exp(-2\pi kx/\beta),\quad j=1,\ldots,2N+1,\quad k\in\mathbb{Z}.\tag{4.32}$$

[The formulas (4.27)–(4.32) should be compared to their hyperbolic counterparts (2.63)–(2.68).]

Restricting attention to meromorphic B-eigenfunctions, it follows that the eigenspace corresponding to the eigenvalue $2\cosh(\beta[y+(N+1)r])$ is two-dimensional over the field of $i\beta$-periodic meromorphic functions, provided $y\neq y_{jk}$; cf. Appendix B in I. (It is not hard to see that for $y=y_{jk}$ this is still true; note in this connection that one need only handle the case $k=0$.) When we insist on joint (B,Q)-eigenfunctions with eigenvalues $(2\cosh(\beta[y+(N+1)r]),-2\cos(\pi y/r))$, we still obtain an infinite-dimensional eigenspace, since we can allow multipliers from the field of elliptic functions with periods $(\pi/r,i\beta)$.

Next, we turn to quantum-mechanical/functional-analytic properties of the operator B and its eigenfunctions. We begin by observing that the relation (4.24) can be used to define $\mathcal{H}_N(x,y)$ for complex y with $\mathrm{Re}\, y>-r$ (say), and, in particular, for $y=nr$, $n\in\mathbb{N}$. (We have already seen that $K>0$, in general, so this is a genuine extension.) We now study the functions

$$\psi_n(x)\equiv\mathcal{H}(x,nr)-\mathcal{H}(-x,nr),\quad n\in\mathbb{N}, \tag{4.33}$$

in relation to the Hilbert space

$$\mathcal{H}_{\hat{w}}\equiv L^2((0,\pi/r),\hat{w}(x)dx), \tag{4.34}$$

where

$$\hat{w}(x)\equiv\prod_{\pm n=1}^{N}\frac{r}{\sin r(x+in\beta)}. \tag{4.35}$$

First, let us note that all of the functions $\psi_n(x)$ belong to the dense subspace $\mathcal{O}_1$ I(4.11) (with $N_+=N,N_-=0$, of course). Indeed, from Sec. II of I we have

$$\mathcal{H}(ik\beta,y)=\mathcal{H}(-ik\beta,y),\quad |k|\leqslant N. \tag{4.36}$$

[See the paragraph containing I(2.39).] Moreover, $\psi_n(x)$ is π/r-periodic (π/r-antiperiodic) for n odd (even). Hence $\psi_n\in\mathcal{O}_1$, as asserted.

Second, it is easily checked that the operator B (4.3) is symmetric on $\mathcal{O}_1$. (One need only adapt the proof of Theorem IV.1 in I, which simplifies considerably in this case.) Now from (4.11) and (4.24) one gets

$$B\psi_n=2\cosh([n+N+1]\beta r)\psi_n,\quad n\in\mathbb{N}, \tag{4.37}$$

so the functions ψ_n are pairwise orthogonal.

Third, we combine (4.13), (4.15), and (4.24) to deduce that $\psi_n(x)$ satisfies the recurrence relation

$$C_n\psi_{n-1}(x)+C_{n+1}\psi_{n+1}(x)=2\cos(rx)\psi_n(x),\quad n\in\mathbb{N}, \tag{4.38}$$

where

$$C_n\equiv\left(\frac{\sinh([n+2N+1]\beta r)}{\sinh([n+N+1]\beta r)}\cdot\frac{\sinh(n\beta r)}{\sinh([n+N]\beta r)}\right)^{1/2},\quad n\in\mathbb{N}. \tag{4.39}$$

Now for $N=0$ we have

$$\mathcal{H}_0(x,y)=\exp ix(y+r),\quad \psi_n(x)=2i\sin(n+1)rx,\quad C_n=1,\quad n\in\mathbb{N}. \tag{4.40}$$

For $N>0$ we have $C_0=0$, so we deduce from (4.38) that $\psi_0(x)$ cannot vanish identically. [Indeed, $\psi_0=0$ would entail successively $\psi_1=0,\psi_2=0,\ldots$, contradicting the $y\to\infty$ asymptotics of $\mathcal{H}_N(x,y)$; cf. the specialization (4.46) of I(3.41).] In fact, using the hyperbolic result (2.78), the function $\psi_0(x)$ will be explicitly determined below.

Fourth, we use (4.38) with $N>0$ to infer

$$\psi_n(x)/\psi_0(x)=G_n(\cos rx),\quad n\in\mathbb{N},\tag{4.41}$$

where the functions $G_n(u)$ are polynomials of degree n and parity $(-)^n$ with real coefficients. As a consequence, the functions ψ_n, $n\in\mathbb{N}$, are an orthogonal base for the Hilbert space $\mathcal{H}_{\hat{w}}$ (4.34). Of course, this entails that the operator B is essentially self-adjoint on the linear span of $\psi_0,\psi_1,\ldots,$ and hence on $\mathcal{O}_1$, too. [For $N=0$ the analogous conclusions are immediate from (4.40).]

In the following theorem we summarize some of the above findings and add some new ones. In particular, we reinterpret the three-term recurrence (4.38) in terms of the *discrete* difference operator

$$D=\left(\frac{\sinh([n+2N+1]\beta r)}{\sinh([n+N+1]\beta r)}\right)^{1/2}S\left(\frac{\sinh([n+1]\beta r)}{\sinh([n+N+1]\beta r)}\right)^{1/2}+\text{h.c.},\tag{4.42}$$

on the Hilbert space $l^2(\mathbb{N})$. Here, S is the right shift,

$$(Sf)_n\equiv\begin{cases}0, & n=0,\\ f_{n-1}, & n>0,\end{cases}\tag{4.43}$$

with $f=(f_0,f_1,\ldots)\in l^2(\mathbb{N})$, and h.c. stands for hermitean conjugate. Clearly, D is a bounded self-adjoint operator on $l^2(\mathbb{N})$.

Theorem IV.1: *The B-eigenfunctions* $\{(r/2\pi)^{1/2}\psi_n(x)\}_{n=0}^{\infty}$ *are an orthonormal base for* $\mathcal{H}_{\hat{w}}$ *(4.34). The self-adjoint operator* D *on* $l^2(\mathbb{N})$ *has a purely absolutely continuous spectrum* $[-2,2]$ *with multiplicity one.*

Proof: Setting

$$N_n\equiv(\psi_n,\psi_n)^{1/2},\quad n\in\mathbb{N},\tag{4.44}$$

it follows from the above that the functions $\psi_0/N_0,\psi_1/N_1,\ldots,$ give rise to an isometric linear map U from $\mathcal{H}_{\hat{w}}$ onto $l^2(\mathbb{N})$. To prove that the normalization constants equal $(2\pi/r)^{1/2}$, we first show that they do not depend on n. Indeed, consider the inner product of the recurrence relation (4.38) with ψ_{n+1}. By virtue of orthogonality, this yields

$$C_{n+1}N_{n+1}^2=(\psi_{n+1}(x),\quad 2\cos(rx)\psi_n(x)).\tag{4.45}$$

Now when we rewrite 2 $\cos(rx)\psi_{n+1}(x)$ by using (4.38) with $n\to n+1$, then we deduce that the rhs of (4.45) equals $C_{n+1}N_n^2$. Hence we get $N_{n+1}=N_n$, and so our assertion follows. (This argument is probably not new, but we do not know a reference.)

Next, specializing I(3.41) to the trigonometric case, we obtain

$$\mathcal{H}(x,y)=\mathcal{H}^{(\infty)}(x,y)+O(e^{-2\beta y}),\quad y\to\infty,\tag{4.46}$$

uniformly on x-compacts, with

$$\mathcal{H}^{(\infty)}(x,y)=\prod_{j=1}^{N}\frac{\sin r(x+ij\beta)}{r}\cdot e^{irx(N+1)}e^{ixy}.\tag{4.47}$$

[This asymptotics can also be derived directly from (4.24) and (4.8).] In view of (4.33), this entails

$$\psi_n(x)=\prod_{j=1}^{N}\frac{\sin r(x+ij\beta)}{r}\cdot e^{irx(N+1+n)}-(x\to -x)+O(e^{-2n\beta r}),\quad n\to\infty,\tag{4.48}$$

the bound being uniform for $x\in[0,\pi/r]$. Since the inner product (ψ_n,ψ_n) does not depend on n, we can now calculate it by using (4.48) and taking $n\to\infty$. Thus we obtain the norm formula [recall (4.35)]

$$(\psi_n,\psi_n)=2\pi/r,\quad n\in\mathbb{N}, \tag{4.49}$$

and so the first assertion of the theorem follows.

Now (4.38) says that $(\psi_0(x),\psi_1(x),\ldots)$ is an improper D-eigenfunction with eigenvalue $2\cos rx$. More precisely, the unitary operator $U^{-1}:l^2(\mathbb{N})\to\mathcal{H}_{\hat{w}}$ sets up a spectral representation for D as multiplication by $2\cos rx$ on $\mathcal{H}_{\hat{w}}$:

$$(U^{-1}DUf)(x)=2\cos(rx)f(x),\quad f\in\mathcal{H}_{\hat{w}}. \tag{4.50}$$

Thus the second assertion is plain. □

We continue by determining $\psi_0(x)$ explicitly. To this end we note that (4.24) and (4.7) entail

$$\psi_0(x)=(4ir)^{-N}\Delta_N(x)\prod_{j=1}^{N}[\sinh(j\beta r)\sinh([N+1+j]\beta r)]^{-1/2}, \tag{4.51}$$

where

$$\Delta_N(x)\equiv K_N(\beta,\pi/ir;x,-i(N+1)\beta)-K_N(\beta,\pi/ir;-x,-i(N+1)\beta). \tag{4.52}$$

Now we use (2.14), (2.71), and (2.78) to calculate $\Delta_N(x)$. This yields

$$\Delta_N(x)=(-)^N2^{3N+1}i\prod_{n=-N}^{N}\sin r(x+in\beta)\cdot\prod_{j=1}^{N}\sinh(j\beta r), \tag{4.53}$$

so that

$$\psi_0(x)=r(2i/r)^{N+1}\prod_{j=1}^{N}\left(\frac{\sinh(j\beta r)}{\sinh([N+1+j]\beta r)}\right)^{1/2}\cdot\prod_{n=-N}^{N}\sin r(x+in\beta). \tag{4.54}$$

Comparing this explicit formula to (4.3), we deduce

$$B=\psi_0(x)^{-1}A\psi_0(x). \tag{4.55}$$

In view of (4.41), this entails that A (4.1) can be viewed as a self-adjoint operator on the Hilbert space

$$\mathcal{H}_A\equiv L^2((0,\pi/r),|\psi_0(x)|^2\hat{w}(x)dx), \tag{4.56}$$

yielding an orthonormal base of polynomials $(r/2\pi)^{1/2}G_n(\cos rx)$; the A-eigenvalues read $2\cosh([n+N+1]\beta r)$ [cf. (4.37)], and the polynomials are uniquely determined by the recurrence (4.38) and $G_0(u)=1$ (save for $N=0$; cf. below).

As already pointed out in Ref. 5 [cf. the paragraph in Ref. 5 containing Eq. (3.84)], the orthogonal polynomials thus obtained are not new: They are q-Gegenbauer polynomials generalizing the integer g Gegenbauer polynomials arising from the trigonometric specialization of the Lamé operator I(1.1). These q-Gegenbauer polynomials were studied in considerable detail by Askey and Ismail;[6] their parameters are related to ours via

$$q_{\mathrm{AI}}=q_t^2=\exp(-2\beta r),\quad \beta_{\mathrm{AI}}=\exp(-2g\beta r),\quad \lambda_{\mathrm{AI}}=g. \tag{4.57}$$

As far as we know, the two representations we have exploited to derive some important features of the integer g polynomials *are* new. In this connection we should also point out that the pertinent weight function integral is immediate from the above. Indeed, from (4.49) we have, in particular, $(\psi_0,\psi_0)=2\pi/r$. Hence (4.54) and (4.35) yield the integral

$$\frac{1}{2\pi}\int_0^\pi dy\,\sin y\prod_{n=-N}^{N}\sin(y+in\alpha)=\frac{1}{4^{N+1}}\prod_{j=1}^{N}\frac{\sinh([N+1+j]\alpha)}{\sinh(j\alpha)},\quad N\in\mathbb{N},\quad \alpha>0, \tag{4.58}$$

as a corollary.

In the remainder of this section we study the case $b=k\beta$ with $-k\in\mathbb{N}$. Since we intend to compare the insights obtained for these parameters to the state of affairs at the elliptic level, we follow the relevant part of Sec. IV in I [starting with the paragraph containing I(4.42)] to a large extent. We first need some preparations, however.

First, in keeping with the notation adopted for the elliptic case, we denote the $A((N+1)\beta)$-eigenfunctions corresponding to $\mathcal{H}_N(x,y)$ by $\Psi(x,y)$. Thus, we have

$$\Psi(x,y)=\mathcal{N}\prod_{j=-N}^{N}\sin r(x+ij\beta)^{-1}\cdot\mathcal{H}_N(x,y),\quad b=(N+1)\beta, \tag{4.59}$$

so that we get the asymptotics

$$\Psi(x,y)\sim c(x)\exp(ixy),\quad y\to\infty, \tag{4.60}$$

with

$$c(r,\beta,(N+1)\beta;x)=\mathcal{N}\prod_{j=0}^{N}\sin r(x-ij\beta)^{-1}\cdot\exp i(N+1)rx. \tag{4.61}$$

(As before, the symbol $\mathcal{N}$ is used to denote normalization constants.)

Second, the trigonometric c-function occurring here equals, more generally,

$$c(r,\beta,b;x)=\frac{G(r,\beta;-x-i\beta/2)}{G(r,\beta;-x+ib-i\beta/2)}, \tag{4.62}$$

where $G(r,\beta;z)$ is the generalized trigonometric gamma function from Ref. 2; just as in the elliptic case, the weight and scattering functions are then given by I(1.27) and I(1.28). [Compare also I(1.24)–I(1.26) with (4.60)–(4.62).]

We are now prepared to follow our elliptic reasoning. Denoting B (4.3) by $B(N)$, and comparing this AΔO to $A(b)$ (1.24), we obtain

$$A(-N\beta)=B(N),\quad N\in\mathbb{N}. \tag{4.63}$$

This equality is the trigonometric counterpart of the elliptic relation I(4.42) with $N_-=0$ and $\delta=-$; the prefactor r_- is absent here, since we omitted the prefactor $\exp(-br)$ occurring in I(1.15); cf. (1.24). (As will be explained at the end of this section, in the trigonometric case this omission is not motivated solely by a desire to avoid clutter from constants.)

Due to (4.63), (4.11), and (4.24) we now have

$$A(-N\beta)\mathcal{H}_N(\pm x,y)=2\cosh(\beta[y+(N+1)r])\mathcal{H}_N(\pm x,y), \tag{4.64}$$

which is the analog of I(4.46) with $N_-=0,\delta=-$. Similarly, I(4.48) and I(4.50) specialize to

$$\Psi(x,y)=\mathcal{N}\mathcal{H}_N(x,y-(2N+1)r),\quad b=-N\beta,\quad \mathrm{Re}\,y>2Nr. \tag{4.65}$$

[Note that (4.24) and (4.20) entail that $\Psi(x,y)$ becomes *singular* for $y=(N+j)r$ with $\pm j=1,\dots,N$.] This guarantees that (4.60) holds true, with

$$c(-N\beta;x)=\mathcal{N}\prod_{j=1}^{N}\sin r(x+ij\beta)\cdot\exp -iNrx. \tag{4.66}$$

[This formula agrees with the pertinent specializations of I(4.52) and (4.62); cf. also Proposition III.14 in Ref. 2.]

The upshot is that we have now obtained eigenfunctions $\Psi(b;\pm x,y)$ of $A(b)$ [and of Q (1.28), of course] for $b=k\beta$, $k\in\mathbb{Z}$. Denoting the eigenvalues by $E(y)$, they have asymptotics

$$E(y)\sim e^{br}e^{\beta y},\quad y\to\infty,\quad b=k\beta,\quad k\in\mathbb{Z}, \tag{4.67}$$

in accord with I(1.19) for $\delta=-$. [Recall that we omitted the factor exp($-br$) in I(1.15) for the trigonometric AΔO (1.24).]

In contrast to the elliptic and hyperbolic cases, the subset $\mathcal{D}_{\text{trig}}$ (1.27) is not dense in the trigonometric parameter domain $\mathcal{T}$(1.23). Therefore, our results have no direct implications for the existence and properties of joint $(A(b),Q)$-eigenfunctions $\Psi(x,y)$ for arbitrary $b\in\mathbb{R}$. But we can shed more light on the orthogonality question for the elliptic case by specializing to the b-values $-N\beta$, $N\in\mathbb{N}$, and studying the analogous trigonometric question. Then we are dealing with the Hilbert space of square-integrable functions on $(0,\pi/r)$ w.r.t. the measure $w(r,\beta,-N\beta;x)dx$; due to (4.66) and the relation $w(x)=1/c(x)c(-x)$, the weight function is proportional to $\hat{w}(x)$ (4.35), so we may as well use $\mathcal{H}_{\hat{w}}$ to study orthogonality.

First, we observe that due to (4.65) the functions

$$\zeta_k(x)\equiv\Psi(x,kr)-\Psi(-x,kr),\quad k\geqslant 2N+1,\quad b=-N\beta, \tag{4.68}$$

are proportional to the functions ψ_{k-2N-1} given by (4.33). Thus they yield an orthogonal base of eigenvectors of $A(-N\beta)$. But in this case we can actually rule out that for $N>0$ the even combinations

$$\chi_k(x)\equiv\Psi(x,kr)+\Psi(-x,kr),\quad k\geqslant 2N+1,\quad b=-N\beta, \tag{4.69}$$

are orthogonal in $\mathcal{H}_{\hat{w}}$. This entails that the elliptic generalizations for $b=-N_\delta a_\delta, N_\delta>0$, are not orthogonal as well [as announced below I(4.57)]. In the following theorem we prove not only nonorthogonality, but also an unexpected completeness property.

Theorem IV.2: *For all $N\in\mathbb{N}$ the functions $\chi_k(x)$ (4.69) satisfy*

$$A(-N\beta)\chi_k(x)=2\cosh([k-N]\beta r)\chi_k(x),\quad k\geqslant 2N+1. \tag{4.70}$$

Now let $N\in\mathbb{N}^*$. Then the functions $\{\chi_k(x)\}_{k=2N+1}^{\infty}$ are total, but not pairwise orthogonal in the Hilbert space $\mathcal{H}_{\hat{w}}$ (4.34).

Proof: The first statement follows from (4.64) and (4.65). In order to prove the second one, we begin by noting that the functions $\chi_k(x)$ satisfy the recurrence relation

$$C_n\chi_{2N+n}(x)+C_{n+1}\chi_{2N+n+2}(x)=2\cos(rx)\chi_{2N+n+1}(x),\quad n\in\mathbb{N}, \tag{4.71}$$

with C_n given by (4.39). [Indeed, this follows in the same way as (4.38).] Since $C_0=0$ and $C_n>0$ for all $n\in\mathbb{N}^*$, this entails as before that $\chi_{2N+1}(x)$ is not identically 0 and that

$$\chi_{2N+n+1}(x)/\chi_{2N+1}(x)=G_n(\cos rx),\quad n\in\mathbb{N}; \tag{4.72}$$

cf. (4.41).

Since $G_n(\cos rx)$ is a polynomial of degree n in $\cos rx$, it now follows from (4.72) that the linear span of the functions $\chi_{2N+1},\chi_{2N+2},\ldots,$ is dense in $\mathcal{H}_{\hat{w}}$. Thus, it remains to show that they are not pairwise orthogonal. We now prove this by deriving a contradiction from the assumption of pairwise orthogonality.

Indeed, this assumption entails (by virtue of the reasoning in the proof of Theorem IV.1) that the polynomials $(r/2\pi)^{1/2}G_n(\cos rx)$ are an orthonormal base for the Hilbert space

$$\tilde{\mathcal{H}}_A \equiv L^2((0,\pi/r),|\eta_0(x)|^2\hat{w}(x)dx), \tag{4.73}$$

where

$$\eta_0(x)\equiv \mathcal{H}_N(x,0)+\mathcal{H}_N(-x,0). \tag{4.74}$$

Since we have already proved that these polynomials have this property w.r.t. the Hilbert space $\mathcal{H}_A$ (4.56), it easily follows that $|\psi_0(x)|^2$ and $|\eta_0(x)|^2$ are equal for $x\in[0,\pi/r]$. Now this amounts to the real part of $\overline{\mathcal{H}_N}(x,0)\mathcal{H}_N(-x,0)$ being 0 for $x\in[0,\pi/r]$, so using (4.24) we infer $\mathrm{Re}\,\overline{L_N}(x,0)L_N(-x,0)$ vanishes for $x\in[0,\pi/r]$. Recalling (4.10), this entails $\mathrm{Re}(L_N(x,0)^2)=0$ for real x. But an inspection of (4.8) reveals that the function $L_N(x,0)^2$ is of the form $\Sigma_{m=2}^{4N+2}a_m \exp imrx$, with $a_m\in\mathbb{R}$. Thus, we infer $L_N(x,0)$ vanishes identically. Since this entails $\psi_0(x)=0$, we finally arrive at the desired contradiction. □

The alert reader will have noted that we excluded the choice $N=0$ from consideration. Indeed, from (4.65) and the $N=0$ formula (4.40), we have $\Psi(x,y)=\mathcal{N}\exp ixy$. Thus, the functions $\zeta_k(x)$ (4.68) and $\chi_k(x)$ (4.69) are proportional to $\sin krx$ and $\cos krx$, respectively, with $k=1,2,\ldots$. Moreover, $\mathcal{H}_{\hat{w}}$ reduces to $L^2((0,\pi/r),dx)$. Now, as before, the functions $\zeta_1,\zeta_2,\ldots,$ are an orthogonal base. But clearly the functions $\chi_1,\chi_2,\ldots,$ are *also* pairwise orthogonal, and they are *not* complete in $\mathcal{H}_{\hat{w}}$, since they are all orthogonal to the constant functions!

At first sight, this seems to contradict our previous reasoning. In fact, however, there is a subtle, but decisive difference with the case $N>0$: The pertinent recurrence coefficients C_n, $n\in\mathbb{N}$, are equal to 1 *including* C_0, whereas C_0 vanishes for $N>0$ [cf. (4.39)]. Hence it does not follow that (4.72) yields polynomials, and indeed the functions $\cos(n+1)rx/\cos rx$ are not polynomials in $\cos rx$.

We also observe that the $N=0$ recurrence is obeyed both by the second-kind Tchebichev polynomials $\sin(n+1)rx/\sin rx$ and by the first-kind ones $\cos nrx$, $n\in\mathbb{N}$. The latter *can* be used to define $A(0)$ as a self-adjoint operator on $L^2((0,\pi/r),dx)$, whereas the former are equal to the above q_t^2-Gegenbauer polynomials for $b=\beta$ [and as such were used to turn $A(\beta)$ into a self-adjoint operator on $L^2((0,\pi/r),\sin^2(rx)dx)$].

To conclude this section, we present some more observations on the relation between the cases $b=(N+1)\beta$ and $b=-N\beta$. When we transform the $A(b)$-eigenfunctions $\Psi(x,y)$ to functions $\mathcal{F}(x,y)$ by using I(1.29) [with $w(x)$ the trigonometric w-function $w(r,\beta,b;x)$], then the functions $\tilde{\mathcal{F}}(x,y)$ for the b-values $-N\beta$ are related to the functions $\mathcal{F}(x,y)$ for the b-values $(N+1)\beta$ via

$$\tilde{\mathcal{F}}(x,y)=\chi\mathcal{F}(x,y-(2N+1)r), \quad \mathrm{Re}\, y>2Nr, \tag{4.75}$$

with χ a phase. [This follows by combining (4.59) and (4.65) with the c-function formulas (4.61) and (4.66).]

This relation is in agreement with the identity

$$H(-b)=H(b+\beta), \tag{4.76}$$

satisfied by the Hamiltonian $H(b)$ (1.25). We have omitted the factor $\exp(-br)$ present in the elliptic counterpart H_- I(1.12), since the symmetry property I(1.13) does not admit a trigonometric specialization. Because we have done so, the symmetry property (4.76) appears instead. Note that this invariance property at the relativistic level turns into $g \to 1-g$ invariance at the nonrelativistic level; cf. I(1.1). By contrast, the invariance property I(1.13) has no nonrelativistic counterpart.

Let us observe finally that—again in contrast to the elliptic case—the AΔOs $H(b_1)$ and $H(b_2)$ are proportional (in fact, equal) only when $b_2 = b_1$ and $b_2 = -b_1 + \beta$. Of course, this easily verified assertion assumes that we restrict attention to $\mathcal{T}$ (1.23), as we have done throughout this section.

ACKNOWLEDGMENT

This work was partly supported by the Netherlands Organisation for the Advancement of Research (NWO).

[1] S. N. M. Ruijsenaars, ''Generalized Lamé functions. I. The elliptic case,'' J. Math. Phys. **40**, 1595–1626 (1999).
[2] S. N. M. Ruijsenaars, ''First order analytic difference equations and integrable quantum systems,'' J. Math. Phys. **38**, 1069–1146 (1997).
[3] S. N. M. Ruijsenaars, ''Action-angle maps and scattering theory for some finite-dimensional integrable systems. I. The pure soliton case,'' Commun. Math. Phys. **115**, 127–165 (1988).
[4] S. N. M. Ruijsenaars, ''Systems of Calogero–Moser type,'' to be published in *Proceedings of the 1994 Banff summer school 'Particles and fields'*, edited by G. Semenoff and L. Vinet.
[5] S. N. M. Ruijsenaars, ''Finite-dimensional soliton systems,'' in *Integrable and Superintegrable Systems*, edited by B. Kupershmidt (World Scientific, Singapore, 1990), pp. 165–206.
[6] R. Askey and M. E.-H. Ismail, ''A generalization of ultraspherical polynomials,'' in *Studies in Pure Mathematics*, edited by P. Erdös (Birkhäuser, Boston, 1983), pp. 55–79.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

On the reproducing kernel of the Segal-Bargmann space

Stephen Bruce Sontz[a)]
Universidad Autónoma Metropolitana-Iztapalapa, Col. Vicentina, México DF 09340, Mexico

(Received 22 July 1998; accepted for publication 29 October 1998)

This article revolves around the properties on the L^p scale of spaces of the integral kernel operator K whose kernel function is the reproducing kernel of the Segal-Bargmann space. We find sufficient conditions on p and q for K to be a Hille-Tamarkin (and hence compact) operator from L^p to L^q with respect to the standard Gaussian measure as well as with respect to a weighted measure on the codomain space. We also find sufficient conditions for K to be unbounded with respect to the standard Gaussian measure. Finally we give sufficent conditions for a Toeplitz operator to be Hille-Tamarkin on the L^p scale of spaces with respect to both the standard Gaussian measure and a weighted measure on the codomain space. © *1999 American Institute of Physics.* [S0022-2488(99)02702-4]

I. INTRODUCTION

The Segal-Bargmann space has enjoyed a long history in mathematics and mathematical physics. Originally introduced by Fischer,[1] it found its first use in physics in the early days of quantum mechanics by Fock[2] and has since come to be recognized as a quantum mechanical version of the phase space of classical mechanics (see Ref. 3). The aspects of this structure relevant to mathematical physics were put on a rigorous basis by Segal (see Refs. 4 and 5) in the infinite dimensional case, and by Bargmann[6] in the finite dimensional case. This last article describes a transform, often call the Segal-Bargmann transform, associated with the Segal-Bargmann space and with another space, sometimes called the configuration space or the Schrödinger space. The reader should be warned that unfortunately the Segal-Bargmann space has been given a wide variety of appellations, including all combinations and permutations of the names Bargmann, Fock, and Segal as well as the ''complex wave representation.'' Of course, the Segal-Bargmann transform has suffered a similar fate, all of which makes literature searches problematical even in an age of computerized data bases. Moreover, in the field of white noise analysis, the S-transform is closely related to the Segal-Bargmann transform (see Ref. 7, p. 337).

In this article, the role of the Segal-Bargmann transform is replaced completely by another structure, the reproducing kernel function, or more precisely, its associated integral kernel transform. (See Ref. 8 for a related analysis of the Segal-Bargmann transform itself.) Consequently, the theory here is developed in the Segal-Bargmann space and other related Banach spaces, without reference to the configuration space. Moreover, only the finite dimensional case is discussed here. The reproducing kernel integral transform is a principal ingredient in the quantization scheme which assigns to a function ϕ, which can be thought of as a classical observable (if it is real valued), a corresponding operator $T(\phi)$ (called a Toeplitz operator), which can be thought of as the corresponding quantum observable.

We now give a summary of this article. Precise definitions will be given in the next section. We work with the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$ (see Refs. 4–6) of holomorphic functions on $\mathbf{C}^n$ that are square integrable with respect to a Gaussian measure μ_n. This space has a reproducing kernel function $K(w,z)$, which in turn defines an integral kernel operator

$$Kf(w) := \int_{\mathbf{C}^n} d\mu_n(z) K(w,z) f(z),$$

[a)]Electronic mail: sontz@xanum.uam.mx

where $f:\mathbf{C}^n \rightarrow \mathbf{C}$ and $w \in \mathbf{C}^n$. This integral operator is well defined for all f in $L^p(\mathbf{C}^n,\mu_n)$ with $1<p\leqslant\infty$. It is proved in Theorem 3.1 that K is a compact (actually, Hille-Tamarkin) operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\mu_n)$ for $p'q<4$ and that in this case

$$||K||_{p\rightarrow q}\leqslant\min((1-p'q/4)^{-n/q},(1-p'q/4)^{-n/p'}),$$

where p' is the dual index of p. Note $p'q<4$ implies $p>q$. Next, calculations with test functions demonstrate that $||K||_{p\rightarrow q}=\infty$ if $p'q>4$. Note $p<q$ implies $p'q>4$. Also, $p=q\neq 2$ implies $p'q>4$. In Theorem 3.2 it is shown that K is a compact (actually, again Hille-Tamarkin) operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\nu_{ab})$ given only that $1<p\leqslant\infty$ and $1\leqslant q<\infty$ where ν_{ab} is the probability measure given by $d\nu_{ab}(z)=c_{ab}\exp(-ax^2-by^2)d\mu_n(z)$ and a and b are constants which depend only on p and q. Here c_{ab} is a normalization constant. Some of the results of Theorem 3.1 can be derived as special cases of Theorem 3.2, but it is hoped that the present exposition shows better how the ideas of Theorem 3.2 were formed. As is usual in L^p analysis, most of the results of this article depend in one way or other on Hölder's inequality, though interpolation theorems are used as well.

The results presented in this article are applied in two ways. First, they are used in Theorem 3.3 to prove a sufficient condition for a Toeplitz operator to be Hille-Tamarkin from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\mu_n)$ or to $L^q(\mathbf{C}^n,\nu_{ab})$. A second application is a demonstration of a reverse log-Sobolev inequality in the Segal-Bargmann space, which is proved in a subsequent article[9] by the author.

The organization of the paper is as follows. The next section introduces the requisite definitions and notation. Section III gives the statements of the theorems, while their proofs are given in Sec. IV. Section V concludes with some comments and remaining open questions.

II. NOTATION AND DEFINITIONS

This section contains a review of notation and definitions. A reference for most of the notation used here for the Segal-Bargmann space is Bargmann's article.[6] For each integer $n\geqslant 1$, the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$ (see Refs. 4–6) is the Hilbert space of holomorphic functions $f:\mathbf{C}^n\rightarrow\mathbf{C}$ which are square integrable with respect to a Gaussian measure given by

$$d\mu_n(z):=\pi^{-n}\exp(-z^*\cdot z)d^nxd^ny,$$

where for each $z=(z_1,\ldots,z_n)\in\mathbf{C}^n$ and $w=(w_1,\ldots,w_n)\in\mathbf{C}^n$ we define $z^*:=(z_1^*,\ldots,z_n^*)$ and $z\cdot w:=z_1w_1+\cdots+z_nw_n$. Here d^nxd^ny is Lebesgue measure on $\mathbf{C}^n$ and λ^* is the complex conjugate of the complex number λ. Also, we write $z^2:=z\cdot z$ for $z\in\mathbf{C}^n$; this specializes to $x^2:=x\cdot x$ for $x\in\mathbf{R}^n$. Notice that

$$HL^2(\mathbf{C}^n,\mu_n)=L^2(\mathbf{C}^n,\mu_n)\cap\mathcal{H}_n,$$

where $L^2(\mathbf{C}^n,\mu_n)$ is the space of all functions $f:\mathbf{C}^n\rightarrow\mathbf{C}$ that are square integrable with respect to μ_n and where $\mathcal{H}_n$ is the space of all holomorphic functions from $\mathbf{C}^n$ to $\mathbf{C}$. The reproducing kernel K for the Segal-Bargmann space is given by

$$K(w,z)=\exp(w\cdot z^*),$$

where $w,z\in\mathbf{C}^n$. It is then well known (see Ref. 6) that K does indeed satisfy the reproducing kernel property, namely, that

$$f(w)=\int_{\mathbf{C}^n}d\mu_n(z)K(w,z)f(z) \tag{2.1}$$

holds for all w in $\mathbf{C}^n$ and for all f in the Segal-Bargmann space. We can also define an integral kernel operator, also denoted by K, associated with the reproducing kernel. Explicitly, we define

$$Kf(w) := \int_{\mathbf{C}^n} d\mu_n(z) K(w,z) f(z), \tag{2.2}$$

where $f:\mathbf{C}^n \to \mathbf{C}$ is measurable and $w \in \mathbf{C}^n$, provided that the integral exists. An application of Hölder's inequality shows that $Kf(w)$ is well defined for all w in $\mathbf{C}^n$ and for all f in $L^p(\mathbf{C}^n, \mu_n)$ when $1 < p \leqslant \infty$; Kf is also a holomorphic function in this case. Then formula (2.1) says that $Kf = f$ for all f in the Segal-Bargmann space $HL^2(\mathbf{C}^n, \mu_n)$. Moreover, it is well known that K as an operator defined in all of $L^2(\mathbf{C}^n, \mu_n)$ is the orthogonal projection operator whose range is precisely the Segal-Bargmann space $HL^2(\mathbf{C}^n, \mu_n)$, which is a closed subspace of $L^2(\mathbf{C}^n, \mu_n)$.

Some of the usual notation for Lebesgue spaces has already been used in the previous paragraph. Here we review this. Whenever (X, μ) is a measure space and $1 \leqslant p \leqslant \infty$, we denote the usual L^p norm of a measurable function $f:X \to \mathbf{C}$ by $\|f\|_p$ or sometimes by $\|f\|_{L^p(\mu)}$ if we need to be more explicit. Also $L^p(X, \mu)$, or simply L^p, denotes the corresponding Banach space. If B is a linear map defined on a dense subspace D of $L^p(X, \mu)$ for $1 \leqslant p \leqslant \infty$ such that $Bf:Y \to \mathbf{C}$ is measurable for every f in D, where (Y, ν) is also a measure space, then we define the operator norm from $L^p(X, \mu)$ to $L^q(Y, \nu)$ where $1 \leqslant q \leqslant \infty$ by

$$\|B\|_{p \to q} := \sup\{\|Bf\|_{L^q(\nu)} : f \in D \text{ and } \|f\|_{L^p(\mu)} = 1\}.$$

It should be noted that this defines an element in $[0, \infty]$. Moreover, if this operator norm is a finite number, then B can be extended uniquely to a continuous linear map defined on all of $L^p(X, \mu)$ and with codomain $L^q(Y, \nu)$. Even though it is standard, the notation $\|\cdot\|_{p \to q}$ is ambiguous because it suppresses reference to the measure spaces as well as to their measures. As usual, context should resolve this ambiguity, though sometimes clarifying comments will be added in the text.

Suppose that (X, μ) and (Y, ν) are σ-finite measure spaces. We say that any measurable function $B:X \times Y \to \mathbf{C}$ is a kernel function. Given such a kernel function, we define its associated integral operator

$$Bf(x) := \int_Y d\nu(y) B(x,y) f(y),$$

where $x \in X$ and $f:Y \to \mathbf{C}$ is measurable, provided that the integral exists. Here we follow the common convention, already used above, of using the same symbol B for the kernel function and for its associated operator. We define the transpose kernel function $B^T:Y \times X \to \mathbf{C}$ by $B^T(y,x) := B(x,y)$ for all $(y,x) \in Y \times X$. For given indices p and q in $[1, \infty]$, we define the Hille-Tamarkin norm (see Refs. 10 and 11) of a kernel function B by

$$|||B|||_{p,q} := \|B_p\|_{L^q(\mu)}, \tag{2.3}$$

where $B_p:X \to \mathbf{C}$ is defined by

$$B_p(x) := \|B(x, \cdot)\|_{L^{p'}(\nu)},$$

where $x \in X$ and p' is the usual dual index of p. As in the case of the operator norm, the notation $|||\cdot|||_{p,q}$ is ambiguous and precisely for the same reason, namely that the notation suppresses reference to the measure spaces and to their measures. Context, sometimes including explicit comments, should again clarify the precise meaning. In the special case $1 < p \leqslant \infty$ and $1 \leqslant q < \infty$, the Hille-Tamarkin norm is given by the formula

$$|||B|||_{p,q} = \left\{ \int_X d\mu(x) \left(\int_Y d\nu(y) |B(x,y)|^{p'} \right)^{q/p'} \right\}^{1/q}.$$

Furthermore, when $p=q=2$, this norm becomes the Hilbert-Schmidt norm. Also one says that the associated operator is a Hille-Tamarkin operator with respect to p,q if its kernel function has a finite Hille-Tamarkin norm for p,q. Here are some facts about the Hille-Tamarkin norm. (See Ref. 11 for proofs of these and other properties.) The notation and hypotheses continue to be those already used in this section.

Proposition 2.1: $||B||_{p\to q}\leqslant|||B|||_{p,q}$ and $||B||_{p\to q}\leqslant|||B^T|||_{q',p'}$.

Proposition 2.2: If $|||B|||_{p,q}$ is finite and $1<p\leqslant\infty$ and $1\leqslant q<\infty$, then the associated integral operator B is a compact operator from $L^p(Y,\nu)$ to $L^q(X,\mu)$.

Proposition 2.3: If $p\geqslant q'$, then $|||B|||_{p,q}\leqslant|||B^T|||_{q',p'}$. If $p\leqslant q'$, then $|||B|||_{p,q}\geqslant|||B^T|||_{q',p'}$.

The proof of the first proposition is an application of Hölder's inequality, while the second is proved by a limiting argument similar to that which shows that the Hilbert-Schmidt operators are compact. Since the set of all Hille-Tamarkin operators with respect to p,q is a Banach space with the Hille-Tamarkin norm, the first part of the first proposition implies that this Banach space is contractively contained in the space of all bounded linear operators, while the second proposition says that this Banach space is a subspace of the space of all compact operators when $p\neq 1$ and $q\neq\infty$. The third proposition is proved by Minkowski's inequality for integrals. It tells us which of the two estimates in the first proposition is better. While this is an easy matter to decide in the applications given here, this proposition shows what is happening in its full generality.

Finally, we define the Toeplitz operator $T(\phi)$ associated to a measurable function $\phi:\mathbf{C}^n\to\mathbf{C}$ by

$$(T(\phi)f)(w):=K(\phi f)(w)=\int_{\mathbf{C}^n}d\mu_n(z)K(w,z)\phi(z)f(z), \tag{2.4}$$

where $w\in\mathbf{C}^n$ and $f:\mathbf{C}^n\to\mathbf{C}$ is measurable provided that the integral exists. This differs from the usual definition in that there are no requirements that f belong to a space of holomorphic functions nor that ϕ be bounded. When ϕ is real valued, it can be interpreted as a physical observable defined on the classical phase space $\mathbf{C}^n$, or equivalently, $\mathbf{R}^{2n}$. More details can be found in Refs. 3 and 12, and references therein.

III. STATEMENT OF THEOREMS

Now we can state the results of this article. The first theorem describes some of the L^p mapping properties of the integral kernel operator K.

Theorem 3.1: Let K denote the integral kernel operator associated to the reproducing kernel function of the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$, as defined in (2.2). Let p, p_1, q, and q_1 be indices in $[1,\infty]$.

(i) If $p\geqslant p_1$ and $q\leqslant q_1$, then $||K||_{p\to q}\leqslant||K||_{p_1\to q_1}$.

(ii) $||K||_{p\to q}\geqslant 1$.

(iii) $||K||_{2\to 2}=1$.

(iv) If $p\geqslant 2$ and $q\leqslant 2$, then $||K||_{p\to q}=1$.

(v) K is a Hille-Tamarkin operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\mu_n)$ if and only if $p'q<4$ where p' is the dual index of p. In this case,

$$1<|||K|||_{p,q}=(1-p'q/4)^{-n/q}.$$

In particular, $|||K|||_{2,2}=\infty$. (Notice that $p'q<4$ implies $p>q$.)

(vi) If $p'q<4$, then K is a compact operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\mu_n)$ with $||K||_{p\to q}\leqslant A_{pq}$, where $A_{pq}:=\min((1-p'q/4)^{-n/q},(1-p'q/4)^{-n/p'})$.

(vii) If $p'q<4$, then $||K||_{p_t\to q_t}\leq A^t_{pq}$ for all $0\leq t\leq 1$, where p_t and q_t are defined implicitly by

$$p_t^{-1}=(1-t)2^{-1}+tp^{-1} \quad \text{and} \quad q_t^{-1}=(1-t)2^{-1}+tq^{-1}. \tag{3.1}$$

(viii) If $p'q>4$, then $||K||_{p\to q}=\infty$. (Notice that $p<q$ implies $p'q>4$. Also $p=q\neq 2$ implies $p'q>4$.)
(ix) If $p'q\neq 4$, then $||K||_{p\to q}=||K||_{q'\to p'}$.

It seems reasonable to conjecture that $||K||_{p\to q}=||K||_{q'\to p'}$ for the remaining case $p'q=4$ as well. The next theorem also concerns L^p mapping properties of the integral kernel operator K, except that now a different measure (which implies a different norm) is used in the codomain space.

Theorem 3.2: Let K denote the integral kernel operator associated to the reproducing kernel function of the Segal-Bargmann space, as defined in (2.2).
(i) If $1<p\leq\infty$ and $1\leq q<\infty$, then there exist real numbers a and b such that K is a Hille-Tamarkin and compact operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\nu_{ab})$ where ν_{ab} is the probability measure $d\nu_{ab}(z):=c_{ab}\exp(-ax^2-by^2)d\mu_n(z)$. Here $z=x+iy\in\mathbf{C}^n$ where $x,y\in\mathbf{R}^n$ and where the normalization constant is $c_{ab}=(1+a)^{n/2}(1+b)^{n/2}$. Explicitly, we have $||K||_{p\to q}\leq|||K|||_{p,q}<\infty$ and $||K||_{p\to q}\leq|||K^T|||_{q',p'}<\infty$, where

$$1<|||K|||_{p,q}=((1+a)(1+b)(a+1-p'q/4)^{-1}(b+1-p'q/4)^{-1})^{n/2q}$$

and

$$1<|||K^T|||_{q',p'}=((1+a)(1+b)(a+1-p'q/4)^{-1}(b+1-p'q/4)^{-1})^{n/2p'},$$

and where a and b satisfy

$$a>p'q/4-1 \quad \text{and} \quad b>p'q/4-1. \tag{3.2}$$

Notice that the operator norm as well as the Hille-Tamarkin norms are with respect to the measure μ_n in the domain space and the measure ν_{ab} in the codomain space. Also, notice that the inequalities (3.2) imply $a>-1$ and $b>-1$ so that c_{ab} and hence ν_{ab} are well-defined.
(ii) If $1<p\leq\infty$ and $1\leq q<\infty$, then K is a bounded linear operator from $L^{p_t}(\mathbf{C}^n,\mu_n)$ to $L^{q_t}(\mathbf{C}^n,\nu^t_{ab})$, where $0\leq t\leq 1$ and

$$d\nu^t_{ab}(z):=c_{ab}^{tq_t/q}\exp\left(-\frac{tq_t}{q}(ax^2+by^2)\right)d\mu_n(z)$$

and p_t and q_t are defined above in (3.1) and a and b satisfy the inequalities (3.2) in part (i). Explicitly, we have the estimate

$$||K||_{p_t\to q_t}\leq(\min(|||K|||_{p,q},|||K^T|||_{q',p'}))^t,$$

where $|||K|||_{p,q}$ and $|||K^T|||_{q',p'}$ are given in part (i). Here the operator norm is with respect to the measure μ_n in the domain space and ν^t_{ab} in the codomain space. However, the Hille-Tamarkin norms are with respect to μ_n in the domain space and ν_{ab} in the codomain space.

As a particular case, we have the estimate

$$||f||_{L^{q_t}(\nu^t_{ab})}\leq M^t||f||_{L^{p_t}(\mu_n)} \tag{3.3}$$

for all f in the Segal-Bargmann space and $M=\min(|||K|||_{p,q},|||K^T|||_{q',p'})$.

Moreover, K is actually a Hille-Tamarkin operator with respect to μ_n in the domain space and ν^t_{ab} in the codomain space for $0<t\leq 1$.

The division of $\mathbf{C}^n$ into two copies of $\mathbf{R}^n$ (via the representation $z=x+iy$) and the corresponding introduction of the two constants a and b is somewhat arbitrary. One could have divided $\mathbf{C}^n$ into $2n$ summands with the introduction of $2n$ constants and have an analogous result.

The final theorem gives some sufficient conditions for a Toeplitz operator to be Hille-Tamarkin with respect to the L^p scale of spaces. This is done with both the standard measure and the weighted measure in the codomain space.

Theorem 3.3: (i) Suppose that $\phi\in L^r(\mathbf{C}^n,\mu_n)$ for some $r>4/3$. Suppose p satisfies $p^{-1}+r^{-1}<3/4$. Define s by $s^{-1}=p^{-1}+r^{-1}$. Suppose q satisfies $s'q<4$. Then the Toeplitz operator $T(\phi)$ is a Hille-Tamarkin and hence compact operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\mu_n)$ with

$$||T(\phi)||_{p\to q}\leq|||T(\phi)|||_{p,q}\leq|||K|||_{s,q}||\phi||_r=(1-s'q/4)^{-n/q}||\phi||_r<\infty.$$

(ii) Suppose $\phi\in L^r(\mathbf{C}^n,\mu_n)$ for some $r>1$. Suppose p satisfies $p^{-1}+r^{-1}<1$. Define s by $s^{-1}=p^{-1}+r^{-1}$. Suppose q satisfies $1\leq q<\infty$. Then the Toeplitz operator $T(\phi)$ is a Hille-Tamarkin and hence compact operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\nu_{ab})$, where ν_{ab} is the probability measure given in part (i) of Theorem 3.2 (where s replaces p). Moreover,

$$\begin{aligned}||T(\phi)||_{p\to q}&\leq|||T(\phi)|||_{p,q}\leq|||K|||_{s,q}||\phi||_r\\&=(1+a)^{n/2q}(1+b)^{n/2q}(a+1-s'q/4)^{-n/2q}(b+1-s'q/4)^{-n/2q}||\phi||_r<\infty,\end{aligned}$$

where a and b satisfy inequality (3.2), again with s replacing p.

It is an immediate consequence of this theorem that one obtains a compact operator with the same operator norm estimate when one restricts $T(\phi)$ to the holomorphic (or any other) subspace of L^p.

IV. PROOFS OF THEOREMS

In this section the theorems are proved and some additional commentary on them is given.

Proof of Theorem 3.1:

(i) If $||K||_{p_1\to q_1}=\infty$, the result is trivial. Otherwise, one uses Hölder's inequality to demonstrate that we have contractive inclusions

$$i_1:L^p(\mathbf{C}^n,\mu_n)\to L^{p_1}(\mathbf{C}^n,\mu_n)$$

and

$$i_2:L^{q_1}(\mathbf{C}^n,\mu_n)\to L^q(\mathbf{C}^n,\mu_n)$$

for $p\geq p_1$ and $q_1\geq q$, where $i_1(f):=f$ and $i_2(f):=f$ for f in the respective space. Since we can factor K, we obtain

$$||K||_{p\to q}=||i_2Ki_1||_{p\to q}\leq||i_2||_{q_1\to q}||K||_{p_1\to q_1}||i_1||_{p\to p_1}=||K||_{p_1\to q_1}.$$

Remark: I call this property of $||K||_{p\to q}$ the *northwest property* since the ordered pair (p^{-1},q^{-1}) is located to the upper left of the pair (p_1^{-1},q_1^{-1}) in the unit square $[0,1]\times[0,1]$. Here one sets $\infty^{-1}=0$. The pair (p^{-1},q^{-1}) is more convenient to use than the corresponding pair (p,q) because of interpolation theory, as we will see below. Obviously, this property also holds in general for integral kernel operators defined between probability (and even finite) measure spaces. In fact, only after completing a preliminary version of this article, did I become aware that this property is also known in the literature as *extrapolation*. (See Ref. 13, page 14.) It is also convenient to name the four quadrants of the unit square by their compass directions (northwest, southwest, northeast, and southeast). Each quadrant is understood to include the appropriate part of the boundary of the unit square, but to exclude the lines $p^{-1}=2^{-1}$ and $q^{-1}=2^{-1}$.

(ii) $K1=1$ where 1 denotes the function that is the constant 1. Since $\|1\|_p=1$ for all $1\leqslant p\leqslant\infty$, the result follows.
(iii) With the domain $L^2(\mathbf{C}^n,\mu_n)$, K is the orthogonal projection operator with range equal to the Segal-Bargmann space. Consequently, $\|K\|_{2\to 2}=1$.
(iv) This follows immediately from parts (i), (ii), and (iii).
(v) We have $|||K|||_{p,q}=\|K_p\|_q$, where $K_p(w)=\|K(w,\cdot)\|_{p'}$ for all $w\in\mathbf{C}^n$. If $p=1$, then

$$K_p(w)=K_1(w)=\|K(w,\cdot)\|_\infty=\sup_{z\in\mathbf{C}^n}|\exp(w\cdot z^*)|=\infty$$

if $w\neq 0$. So, $|||K|||_{1,q}=\|K_1\|_q=\infty$ for all $1\leqslant q\leqslant\infty$. For $1<p\leqslant\infty$ and $w=u+iv$ with u,v in $\mathbf{R}^n$, we have

$$K_p(w)=\left\{\int_{\mathbf{C}^n} d^nx\,d^ny\,\pi^{-n}e^{-x^2-y^2}|e^{w\cdot z^*}|^{p'}\right\}^{1/p'}=\exp(p'(u^2+v^2)/4) \tag{4.1}$$

by doing the Gaussian integral explicitly. If $q=\infty$, then

$$|||K|||_{p,q}=\|K_p\|_\infty=\sup_{w\in\mathbf{C}^n}\exp(p'(u^2+v^2)/4)=\infty.$$

If $1\leqslant q<\infty$, then by doing the Gaussian integral we have

$$|||K|||_{p,q}=\|K_p\|_q=\left\{\int_{\mathbf{C}^n} d^nu\,d^nv\,\pi^{-n}e^{-u^2-v^2}e^{p'q(u^2+v^2)/4}\right\}^{1/q}=(1-p'q/4)^{-n/q}$$

if $1-p'q/4>0$. Moreover, $|||K|||_{p,q}=\|K_p\|_q=\infty$ if $1-p'q/4\leqslant 0$. Finally, since $1-p'q/4<1$, we have that $(1-p'q/4)^{-n/q}>1$. This shows (v).

Remark: Notice that $|||K|||_{2,2}=\infty$ follows as a special case of (v). This says that K is not a Hilbert-Schmidt operator in $L^2(\mathbf{C}^n,\mu_n)$ even though it is bounded. Of course, we already know that K is an orthogonal projection operator in $L^2(\mathbf{C}^n,\mu_n)$ with infinite dimensional range, and so it is not a compact operator, and hence not Hilbert-Schmidt.
(vi) This follows immediately from part (v) and Propositions 2.1 and 2.2 and from a calculation, similar to that given in part (v), which shows that for $p'q<4$ we have

$$\|K\|_{p\to q}\leqslant|||K^T|||_{q',p'}=(1-p'q/4)^{-n/p'}.$$

Remarks: Part (iv) says that the situation for the norm $\|K\|_{p\to q}$ is trivial for all (p^{-1},q^{-1}) that are northwest of $(2^{-1},2^{-1})$. The condition $p'q<4$ includes all of those cases, but also allows other values of (p^{-1},q^{-1}). For instance, $\|K\|_{4\to q}$ is finite for all $2<q<3$ and $(4^{-1},q^{-1})$ lies southwest of $(2^{-1},2^{-1})$. Also, $\|K\|_{8/5\to q}$ is finite for $1\leqslant q<3/2$ and $(5/8,q^{-1})$ lies northeast of $(2^{-1},2^{-1})$. We can find pairs (p^{-1},q^{-1}) as near as we like to $(2^{-1},2^{-1})$ in the southwest or in the northeast and such that $p'q<4$. For example, $p^{-1}=2.01^{-1}$ and $q^{-1}=2.001^{-1}$ is one such pair in the southwest. An example in the northeast is $p^{-1}=1.9905^{-1}$ and $q^{-1}=1.99^{-1}$. Notice that in all of these cases $p>q$ and that the operator is compact. If we define the set

$$G:=\{(p^{-1},q^{-1})\in[0,1]\times[0,1]\ :\ p'q<4\}, \tag{4.2}$$

then the reflection S in the diagonal line $p^{-1}+q^{-1}=1$ leaves G invariant as the following shows. Explicitly, $S(x,y)=(1-y,1-x)$ for any (x,y) in $\mathbf{R}^2$ is the reflection in the line $x+y=1$. The transformation of Lebesgue indices corresponding to S sends (p,q) to (q',p') and so G is clearly invariant under S. However, the Hille-Tamarkin norm is *not* invariant under this transformation. Nonetheless, the estimate of the operator norm that we get in part (vi) from *two* Hille-Tamarkin norms indeed is invariant under S. This result should be compared with that of part (ix).
(vii) This follows immediately from parts (iii) and (vi) and the Riesz-Thorin interpolation theorem (see Ref. 14) applied to the pairs (p,q) and (2,2). Notice that the estimate in (vi) behaves

badly as (p^{-1},q^{-1}) approaches $(2^{-1},2^{-1})$, namely it approaches infinity. However, the estimate in part (vii) behaves well as t approaches zero (which implies that (p_t^{-1},q_t^{-1}) approaches $(2^{-1},2^{-1})$ along a line segment), namely it approaches 1, which is the value of $||K||_{2\to 2}$.

(viii) The idea is to use a family of trial functions to estimate $||K||_{p\to q}$ from below. Define

$$g(z)=\exp((a+bi)\cdot x+(c+di)\cdot y), \tag{4.3}$$

where x, y, a, b, c, and d are in $\mathbf{R}^n$ and $z=x+iy$ is in $\mathbf{C}^n$. The following three computations then consist of nothing more than evaluations of Gaussian integrals. We do these computations in the case $1\leqslant p<\infty$ and $1\leqslant q<\infty$. First, we have

$$||g||_p=\exp(p(a^2+c^2)/4).$$

Next, for w in $\mathbf{C}^n$, we obtain

$$Kg(w)=\exp\{\tfrac{1}{4}[2w\cdot(a+bi-ci+d)+(a+bi)^2+(c+di)^2]\}.$$

And finally, we compute

$$||Kg||_q=\exp\left\{\frac{1}{4}\left(a^2-b^2+c^2-d^2+q\left[\left(\frac{(a+d)}{2}\right)^2+\left(\frac{(b-c)}{2}\right)^2\right]\right)\right\}.$$

Then a lower bound on the norm $||K||_{p\to q}$ is given by

$$\frac{||Kg||_q}{||g||_p}=\exp\left\{\frac{1}{16}[\alpha a^2+\beta b^2+\alpha c^2+\beta d^2+2q(ad-bc)]\right\},$$

where $\alpha=4+q-4p$ and $\beta=q-4$. The quadratic form in the square brackets on the right hand side here can be diagonalized in the pair of variables a,d and separately in the pair of variables b,c. The result is that the quadratic form has two distinct eigenvalues, each of multiplicity two. They are denoted λ_+ and λ_- where

$$\lambda_\pm=(q-2p)\pm(q^2+4(p-2)^2)^{1/2}.$$

Since $\lambda_+>\lambda_-$, the quadratic form has at least one positive eigenvalue if and only if $\lambda_+>0$. Moreover, $\lambda_+>0$ implies

$$\sup_g\frac{||Kg||_q}{||g||_p}=\infty,$$

where g in the supremum ranges over all functions of the above form (4.3). It is straightforward to show that $p'q>4$ implies $\lambda_+>0$. Hence, $p'q>4$ implies $||K||_{p\to q}=\infty$ if both $1\leqslant p<\infty$ and $1\leqslant q<\infty$. The case $q=\infty$ and $1\leqslant p<\infty$ is proved using the northwest property and the previous result. The case $p=\infty$ and $1\leqslant q<\infty$ is handled by taking $a=c=0$ in the definition of g so that

$$\frac{||Kg||_q}{||g||_\infty}=\exp\left\{\frac{1}{16}((q-4)b^2+(q-4)d^2)\right\}.$$

Then $p'q=q>4$ implies $||K||_{\infty\to q}=\infty$. Finally, the case $p=q=\infty$ is demonstrated with the northwest property and the previous case. And so all cases of (viii) have been shown.

(ix) By part (viii), this result is trivial if $p'q>4$. So, we assume that $p'q<4$. In this case $|||K|||_{p,q}$ is finite by part (v). We will use the characterization of the operator norm given by

$$||K||_{p\to q}=\sup\{|(g,Kf)|\ :\ ||f||_p=1 \quad\text{and}\quad ||g||_{q'}=1\}, \tag{4.4}$$

where (g,Kf) denotes the usual duality pairing (without complex conjugation). So, for $||f||_p=1$ and $||g||_{q'}=1$, we have

$$(g,Kf)=\int_{\mathbf{C}^n} d\mu_n(w)g(w)Kf(w)=\int_{\mathbf{C}^n} d\mu_n(w)g(w)\int_{\mathbf{C}^n} d\mu_n(z)K(w,z)f(z)$$

$$=\int_{\mathbf{C}^n} d\mu_n(z)f(z)\int_{\mathbf{C}^n} d\mu_n(w)e^{w\cdot z^*}g(w). \tag{4.5}$$

The interchange of integrals is justified by Fubini's theorem for functions that are integrable in the product space, since by two applications of Hölder's inequality and by Fubini's theorem for non-negative functions we have

$$\int_{\mathbf{C}^n\times\mathbf{C}^n} d\mu_n(z)d\mu_n(w)|g(w)K(w,z)f(z)|\leqslant||f||_p||g||_{q'}|||K|||_{p,q}=|||K|||_{p,q}<\infty.$$

(As an aside, the reader should note that a calculation of the sort in the last line is the way one proves Proposition 2.1.) But similarly

$$||K||_{q'\to p'}=\sup\{|(f,Kg)|\ :\ ||g||_{q'}=1 \quad \text{and} \quad ||f||_p=1\} \tag{4.6}$$

and

$$(f,Kg)=\int_{\mathbf{C}^n} d\mu_n(w)f(w)Kg(w)=\int_{\mathbf{C}^n} d\mu_n(w)f(w)\int_{\mathbf{C}^n} d\mu_n(z)e^{w\cdot z^*}g(z)$$

$$=\int_{\mathbf{C}^n} d\mu_n(z^*)f(z^*)\int_{\mathbf{C}^n} d\mu_n(w^*)e^{w\cdot z^*}g(w^*) \tag{4.7}$$

by a simple change of variables. But this change of variables leaves invariant the measure μ_n and the classes of functions f (respectively, g) such that $||f||_p=1$ (respectively, $||g||_{q'}=1$). So the numbers which appear in Eq. (4.7) as f and g vary are the same as appear in (4.5). So the result follows from Eqs. (4.4) and (4.6). This completes the proof of part (ix) and of Theorem 3.1. □

We proceed next to the proof of the second theorem.

Proof of Theorem 3.2:

(i) We need to compute the Hille-Tamarkin norm of K much as in part (v) of Theorem 3.1. Only now we use a different measure in the codomain space, namely ν_{ab}. However, the definition and hence computation of $K_p(w)$ in formula (4.1) does not change. So, for $1<p\leqslant\infty$ and $1\leqslant q<\infty$, we have

$$|||K|||_{p,q}=||K_p||_q=\left\{\int_{\mathbf{C}^n} d^nu\,d^nv\,\pi^{-n}c_{ab}e^{-au^2-bv^2}e^{-u^2-v^2}e^{p'q(u^2+v^2)/4}\right\}^{1/q}$$

$$=((1+a)(1+b)(a+1-p'q/4)^{-1}(b+1-p'q/4)^{-1})^{n/2q} \tag{4.8}$$

provided that $a>p'q/4-1$ and $b>p'q/4-1$. Also, since $p'q>0$, it follows that $|||K|||_{p,q}>1$. Together with Propositions 2.1 and 2.2, this shows that the operator is compact and that the first formula in part (i) holds. The second formula follows by a similar calculation.

Remark: Notice that when $p'q/4-1<0$ we can take $a=b=0$ and so we recover one implication in part (v) of Theorem 3.1 as a special case. Notice also that we can always select a and b so that $a=b$. In this case, we can think of the codomain space as being a Banach space associated with a Segal-Bargmann space endowed with a Planck's constant distinct from that of the domain space. See Ref. 15, pp. 138 and 168, for a presentation of the Segal-Bargmann space with explicit use of Planck's constant. Actually, in the case $a\neq b$, we can think of the codomain as arising from a generalized type of Segal-Bargmann space with *two* distinct values for Planck's ''constant.'' Also, in the case $p=1$ or $q=\infty$ it is straightforward to show that there is no Gaussian weight factor for the codomain space such that K is Hille-Tamarkin.

To prove part (ii) of Theorem 3.2, we will again use interpolation theory, except that now the relevant result is a theorem due to Stein.[16] Since it is not as well known as the Riesz-Thorin theorem, here is a statement of Stein's theorem taken from Ref. 16 and transcribed into the notation of this article. First, recall that a simple function is a measurable function f with a finite range I such that $f^{-1}(i)$ has finite measure for every nonzero $i \in I$.

Theorem 4.1 (Stein): Let (M,ν) and (N,μ) be σ-finite measure spaces. Let T be a linear transformation defined on simple functions of M to measurable functions on N. Suppose that p_0, p_1, q_0, and q_1 are in $[1,\infty]$ and define p_t and q_t implicitly by

$$p_t^{-1}=(1-t)p_0^{-1}+tp_1^{-1} \quad \text{and} \quad q_t^{-1}=(1-t)q_0^{-1}+tq_1^{-1}$$

for $0 \leqslant t \leqslant 1$. Suppose that for some measurable functions $k_0,k_1:N \to [0,\infty)$ and $u_0,u_1:M \to [0,\infty)$ and for all simple $f:M \to \mathbf{C}$ one has

$$||(Tf)k_0||_{L^{q_0}(\mu)} \leqslant M_0||fu_0||_{L^{p_0}(\nu)} \tag{4.9}$$

and

$$||(Tf)k_1||_{L^{q_1}(\mu)} \leqslant M_1||fu_1||_{L^{p_1}(\nu)} \tag{4.10}$$

for some M_0 and M_1 in $\mathbf{R}$. (The right hand side of either of these inequalities may be infinite for some f.) Define new functions $k_t := k_0^{1-t}k_1^t$ on N and $u_t := u_0^{1-t}u_1^t$ on M for $0 \leqslant t \leqslant 1$. Then T can be extended uniquely to a linear transformation on all functions f satisfying $||fu_t||_{L^{p_t}(\nu)} < \infty$ so that

$$||(Tf)k_t||_{L^{q_t}(\mu)} \leqslant M_0^{1-t}M_1^t||fu_t||_{L^{p_t}(\nu)}. \tag{4.11}$$

To apply Stein's theorem, we take the measure spaces (M,ν) and (N,μ) to be $(\mathbf{C}^n,\mu_n)$. Also, we take the Lebesgue indices to be $p_1=p$, $q_1=q$ and $p_0=q_0=2$. Finally, we take the weight functions to be $k_0=u_0=u_1=1$ and

$$k_1(z)=c_{ab}^{1/q}\exp\left(-\frac{1}{q}(ax^2+by^2)\right), \tag{4.12}$$

where a and b satisfy (3.2). Then it is easy to show that

$$||(Kf)k_1||_{L^q(\mu_n)}=||Kf||_{L^q(\nu_{ab})}.$$

So, it follows from part (i) that

$$||(Kf)k_1||_{L^q(\mu_n)} \leqslant M_1||fu_1||_{L^p(\mu_n)},$$

where $M_1=\min(|||K|||_{p,q},|||K^T|||_{q',p'})$. This shows that inequality (4.10) of Stein's theorem holds. But we have

$$||(Kf)k_0||_{L^2(\mu_n)} \leqslant ||fu_0||_{L^2(\mu_n)}$$

since K is an orthogonal projection. This shows that inequality (4.9) of Stein's theorem holds with $M_0=1$. So we conclude that inequality (4.11) of Stein's theorem holds, that is to say, that for $0 \leqslant t \leqslant 1$

$$||(Kf)k_t||_{L^{q_t}(\mu_n)} \leqslant M_1^t||fu_t||_{L^{p_t}(\mu_n)}.$$

But $u_t = 1$ and

$$k_t(z) = k_0^{1-t}(z) k_1^t(z) = k_1^t(z) = c_{ab}^{t/q} \exp\left(-\frac{t}{q}(ax^2 + by^2) \right) \tag{4.13}$$

for all $0 \leqslant t \leqslant 1$. Since it is easy to see that

$$||(Kf)k_t||_{L^{q_t}(\mu_n)} = ||Kf||_{L^{q_t}(\nu_{ab}^t)}$$

we have that

$$||Kf||_{L^{q_t}(\nu_{ab}^t)} \leqslant M_1^t ||f||_{L^{p_t}(\mu_n)} \tag{4.14}$$

or in the notation of operator norms,

$$||K||_{p_t \to q_t} \leqslant M_1^t$$

which is the estimate we had to show.

The particular case (3.3) follows immediately from (4.14) since $Kf = f$ for f in the Segal-Bargmann space.

To show that K is Hille-Tamarkin with respect to μ_n and ν_{ab}^t we have to show, according to part (i), that

$$\frac{tq_t}{q} a > p_t' q_t /4 - 1 \text{ and } \frac{tq_t}{q} b > p_t' q_t /4 - 1$$

hold for $0 < t \leqslant 1$ given that a and b satisfy (3.2). (Notice that ν_{ab}^t is in general a finite measure, but not a probability measure. Nonetheless, the criterion of part (i) applies.) To do this, it suffices to show that

$$\frac{q}{tq_t}(p_t' q_t/4 - 1) \leqslant p' q/4 - 1, \tag{4.15}$$

since the right hand side of (4.15) is $< \min(a,b)$ by (3.2) and $t > 0$. But a computation for the case $1 < p < \infty$ shows that

$$\frac{q}{tq_t}(p_t' q_t/4 - 1) = \frac{q}{4} \frac{2(2 + pt - 2t)}{(p + pt - 2t)} - 1$$

and so it suffices to show that for $0 < t \leqslant 1$ we have

$$\frac{2(2 + pt - 2t)}{p + pt - 2t} \leqslant p' = \frac{p}{p-1}.$$

But this last inequality is trivial to verify. Since $p > 1$ by hypothesis, the remaining case is $p = \infty$, which is also straightforward. This concludes the proof of Theorem 3.2. □

Remark: Notice that in part (ii), the measure in the codomain space depends on the interpolating parameter t. Again, in the special case $a = b$, we can think of these interpolating measures as being given by a continuous deformation of Planck's constant between two extreme values. The behavior of the estimate of the operator norm $||K||_{p \to q}$ in part (i) is highly erratic as (p,q) approaches (2,2) because the parameters a and b depend on p and q, but only via the inequalities (3.2) rather than in a functional manner. Worse yet, at $(p,q) = (2,2)$, part (i) tells us that K is a Hille-Tamarkin operator but only if we pick $a > 0$ and $b > 0$. Of course, we know that K is actually a bounded operator at $(p,q) = (2,2)$ with $a = b = 0$, though *not* a Hille-Tamarkin operator with

$a=b=0$. So the interpolation in part (ii) gives us a better behaved estimate of $||K||_{p_t\to q_t}$ as $t\to 0$ (which implies that (p_t^{-1},q_t^{-1}) approaches $(2^{-1},2^{-1})$ along a line segment) since only one choice of parameters a and b is used, namely that at the endpoint corresponding to $t=1$.

Proof of Theorem 3.3: The two parts will be proved simultaneously. First notice that the hypotheses on r, p and q ensure that the rest of the assertions are non-vacuous and that $|||K|||_{s,q}$ is finite. Next notice that according to its definition in (2.4), $T(\phi)$ is an integral kernel operator with kernel function $K(w,z)\phi(z)$ for w,z in $\mathbf{C}^n$. Since the relation $s^{-1}=p^{-1}+r^{-1}$ implies $p'^{-1}=s'^{-1}+r^{-1}$, we have for all w in $\mathbf{C}^n$ that

$$||K(w,\cdot)\phi(\cdot)||_{p'}\leq||K(w,\cdot)||_{s'}||\phi||_r,$$

where the norms are with respect to $d\mu_n(z)$. Next, we take the L^q norm of both sides of the previous inequality. We do this with respect to $d\mu_n(w)$ for the first part of the theorem and correspondingly with respect to $d\nu_{a,b}(w)$ for the second part. Using the definition (2.3) of the Hille-Tamarkin norm, we obtain $|||T(\phi)|||_{p,q}\leq|||K|||_{s,q}||\phi||_r$. Together with Propositions 2.1 and 2.2 and the formulas for the Hille-Tamarkin norms in Theorems 3.1 and 3.2, this gives all the results. This concludes the proof of Theorem 3.3. □

V. CONCLUDING REMARKS

The results of this article depend critically on the estimates that come from the Hille-Tamarkin norm. However, this norm is usually much stronger than the operator norm, which is to say the estimates of Proposition 2.1 are often far from optimal. Despite this, the general situation of the transform K with respect to the L^p scale of spaces is revealed by this method. Specifically, if we define

$$E:=\{(p^{-1},q^{-1})\ :\ p'q=4\},$$

then for every (p^{-1},q^{-1}) that is not in E, we have either that K is compact (actually, Hille-Tamarkin) from L^p to L^q or that $||K||_{p\to q}=\infty$. For pairs (p^{-1},q^{-1}) in E, it remains an open problem to determine $||K||_{p\to q}$, except in the case $p=q=2$. It follows that K, as a transform from $L^2(\mathbf{C}^n,\mu_n)$ to itself is not the generic case on the L^p scale of spaces (i.e., is neither Hille-Tamarkin nor unbounded). The Segal-Bargmann transform is also generically either Hille-Tamarkin or unbounded (see Ref. 8) and, in fact, this is a general property of a Gaussian integral kernel operator G defined between two Euclidean spaces, each endowed with a Gaussian measure.

Of course, the lack of sharpness in the operator norm estimate is a question of finding best constants and this is an open problem for all of the estimates given of operator norms by Hille-Tamarkin norms in this article. It is perhaps interesting, in its own right, to know the value of the operator norm of K restricted to the subspace of holomorphic functions of $L^p(\mathbf{C}^n,\mu_n)$ with codomain $L^q(\mathbf{C}^n,\mu_n)$ or $L^q(\mathbf{C}^n,\nu_{ab})$. Moreover, the operator norm of K from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\mu_n)$ may obey a ''one-infinity'' law. This means that for every pair of indices p and q in $[1,\infty]$ one would have either $||K||_{p\to q}=1$ or $||K||_{p\to q}=\infty$. All of the results of this paper are consistent with this possibility. Also, various calculations with Gaussian trial functions failed to contradict this possibility. But it remains an open question.

Some comments are in order about the rather nice paper by Lieb[17] on Gaussian kernel integral transforms, especially since his results have not been used here. At first reading, it might appear that there is little overlap with the cases presented here and those found in Ref. 17, but the situation is more complicated. The fact that Lieb's results are presented with respect to Lebesgue measure (instead of Gaussian measure as is done in this article) is no obstacle; it is well known that there are Banach space isomorphisms (i.e., one-to-one, onto isometries) between $L^p(\mathbf{C}^n,\mu_n)$ and $L^p(\mathbf{C}^n,d^nxd^ny)$ for all p in $[1,\infty]$. Using these isomorphisms, one can easily translate the results of Lieb to the case of Gaussian measures. For example, using the lemma on p. 185 of Ref. 17, one can show that K is a compact operator from $L^p(\mathbf{C}^n,\mu_n)$ to $L^q(\mathbf{C}^n,\mu_n)$ provided that $p'q<4$. And one can even directly read off the estimate $||K||_{p\to q}\leq(1-p'q/4)^{-n/q}$ from the proof of this lemma, since Lieb uses implicitly the estimate $||B||_{p\to q}\le|||B|||_{p,q}$ of Proposition 2.1. (It should be remarked that the above Banach space isomorphisms preserve Hille-Tamarkin norm.) But the most interesting part of Ref. 17 is the determination of the maximizers, which in the context of Theorem 3.1 are those nonzero f in $L^p(\mathbf{C}^n,\mu_n)$ such that $||Kf||_q=||K||_{p\to q}||f||_p$, if indeed such f exist. However, Lieb's results for maximizers apply only under certain hypotheses. Either the situations described here do not fall directly under one of Lieb's conditions or, when they do as in Theorem 3.2, the resulting algebraic calculations are quite complicated.

It should be commented that the technique of the Hille-Tamarkin norm works here because these norms become Gaussian integrals, which can therefore be evaluated explicitly. This happens because the measures and the reproducing kernel function are Gaussian; consequently this analysis works as well in a more general Gaussian situation. See, for example, Ref. 8 where a similar analysis of the Segal-Bargmann transform is given. It seems that the Hille-Tamarkin norm has not seen much use in classical analysis because for many commonly occurring integral operators, such as convolutions with respect to Lebesgue measure or oscillatory integrals (for example, the Fourier transform $\mathcal{F}$), the Hille-Tamarkin norm is typically infinite. However, in this regard, note that $|||\mathcal{F}|||_{1,\infty}$ is finite and that the standard argument that $||\mathcal{F}||_{1\to\infty}$ is finite can be expressed in this terminology as $||\mathcal{F}||_{1\to\infty}\le|||\mathcal{F}|||_{1,\infty}<\infty$.

The results of Theorem 3.3 depend on the fact that the Toeplitz operator $T(\phi)$ has a factorization as a Hille-Tamarkin operator, namely K, multiplied by a bounded operator, namely multiplication by ϕ which we denote M_ϕ. It would be interesting to know what happens in contexts where M_ϕ is more singular since it may be possible that $T(\phi)$ is bounded or compact, though not Hille-Tamarkin, in such a case. Also it would be interesting to know whether the mapping properties of $T(\phi)$ are stronger when its domain of definition is taken to be the subspace of holomorphic functions of $L^p(\mathbf{C}^n,\mu_n)$.

Finally, the results of this article are exclusively for the case when n, the dimension, is finite and so it would be interesting to know what happens in the case of infinitely many dimensions.

ACKNOWLEDGMENTS

I wish to thank Lawrence Thomas for stimulating my interest in the Segal-Bargmann space as well as for much encouragement and support. I also thank Brian Hall for conversations about Segal-Bargmann analysis and for providing me with a preprint of his article (Ref. 3).

[1] E. Fischer, ''Über die differentiatensprozesse der algebra,'' J. Math. **148**, 1–78 (1917).
[2] V. Fock, ''Verallgemeinerung and lösung der Diracshen statistischen gleichung,'' Z. Phys. **49**, 339–357 (1928).
[3] B. C. Hall, *Quantum Mechanics in Phase Space,* in *Perspectives on Quantization,* edited by L. A. Coburn and M. A. Rieffel (American Mathematical Society, Providence, 1998), pp. 47–62, Contemporary Mathematics, Vol. 214.
[4] I. E. Segal, *Mathematical Problems of Relativistic Physics* (American Mathematical Society, Providence, 1963).
[5] I. E. Segal, *The Complex-Wave Representation of the Free Boson Field,* in *Topics in Functional Analysis: Essays Dedicated to M. G. Krein on the Occasion of His 70th Birthday,* edited by I. Gohberg and M. Kac (Academic, New York, 1978), pp. 321–343, Advances in Mathematics, Supplementary Studies, Vol. 3.
[6] V. Bargmann, ''On a Hilbert space of analytic functions and an associated integral transform, Part I,'' Commun. Pure Appl. Math. **14**, 187–214 (1961).
[7] H.-H Kuo, *White Noise Distribution Theory* (CRC Press, Boca Raton, 1996).
[8] S. B. Sontz, ''Entropy and the Segal-Bargmann transform,'' J. Math. Phys. **39**, 2402–2417 (1998).
[9] S. B. Sontz, ''A reverse log-Sobolev inequality in the Segal-Bargmann space,'' J. Math. Phys. **40**, 1677–1695 (1999).
[10] E. Hille and J. D. Tamarkin, ''On the theory of linear equations, II,'' Ann. Math. **35**, 445–455 (1934).
[11] K. Jörgens, *Linear Integral Operators* (Pitman, Boston, 1982).
[12] C. A. Berger and L. A. Coburn, ''Heat flow and Berezin-Toeplitz estimates,'' Am. J. Math. **116**, 563–590 (1994).
[13] M. A. Krasnoselskii *et al.*, *Integral Operators in Spaces of Summable Functions* (Noordhoff International, Leyden, 1976).
[14] E. Stein and G. Weiss, *Introduction to Fourier Analysis in Euclidean Space* (Princeton University Press, Princeton, 1971).
[15] N. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1980).
[16] E. Stein, ''Interpolation of linear operators,'' Trans. Am. Math. Soc. **83**, 482–492 (1956).
[17] E. M. Lieb, ''Gaussian kernels have only Gaussian maximizers,'' Invent. Math. **102**, 179–208 (1990).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 3 MARCH 1999

A reverse log-Sobolev inequality in the Segal-Bargmann space

Stephen Bruce Sontz[a)]
Universidad Autónoma Metropolitana-Iztapalapa, Col. Vicentina, México DF 09340, México

(Received 5 October 1998; accepted for publication 2 November 1998)

This article is based on recent results of the author on the properties of the reproducing kernel of the Segal-Bargmann space. Those results are used here to demonstrate a family of energy-entropy inequalities in the Segal-Bargmann space. In some cases this is a log-Sobolev inequality while in other cases this is actually a *reverse* log-Sobolev inequality, which means that the energy term is bounded above by the entropy term, plus a norm term. This implies that in the Segal-Bargmann space the entropy is finite if and only if the energy is finite. Applications of this result to the Segal-Bargmann transform are given as well as a discussion of its possible relation with reverse hypercontractivity. It should be noted that all of the results of this article are proved without using hypercontractivity estimates.
 [S0022-2488(99)03702-0]

I. INTRODUCTION

Suppose that (X,μ) is a probability measure space and that we have a distinguished non-negative quadratic form (say, a Dirichlet form) $Q(f)$ defined for all $f\in\mathcal{H}$ where $\mathcal{H}$ is a closed subspace of $L^2(X,\mu)$. One calls $Q(f)$ the energy of f. Then a log-Sobolev inequality in $\mathcal{H}$ is an inequality of the form

$$S(f)\leqslant a_1Q(f)+a_2||f||_2^2$$

for constants $a_1>0$ and $a_2\geqslant 0$ for all f in $\mathcal{H}$. Here $S(f)$ is the entropy of f, to be defined later. Such an inequality says that finite energy implies finite entropy. Such inequalities and their variants have played an important role in various fields of mathematics. The article[1] by Gross is the first systematic discussion of these inequalities. A later review article[2] of his has references to the literature before the date of that publication. One of the many recent references is Ref. 3. This gives a far from complete guide to the extensive literature on log-Sobolev inequalities.

On the other hand, we have a related situation in the following definition which says that finite entropy implies finite energy. We continue to use the hypotheses and notation introduced above.

Definition 1.1: An inequality of the form

$$Q(f)\leqslant b_1S(f)+b_2||f||_2^2$$

with $b_1>0$ and $b_2\geqslant 0$ and f in $\mathcal{H}$ is called a reverse log-Sobolev inequality.

In this article, I prove a parameterized family of inequalities involving energy and entropy for f in the Segal-Bargmann space of the form

$$(p^{-1}-q^{-1})S(f)\leqslant \frac{n}{q}(a-\log(a+1-p'q/4))||f||_2^2+\frac{a}{q}\langle f,Nf\rangle,$$

[a)]Electronic mail: sontz@xanum.uam.mx

where $1<p\leqslant\infty$, $1\leqslant q<\infty$, $a>p'q/4-1$, $\langle f,Nf\rangle$ is the Dirichlet form associated to the energy operator N in the Segal-Bargmann space, and p' is the dual index of p. (The parameter n is a dimension and will be specified later.) For some choices of the parameters, one gets a (regular) log-Sobolev inequality. This is a result that has been established already various times in the literature. (It was originally shown in Ref. 1, but also follows from a complex hypercontractivity inequality shown in Refs. 4–8.) For other choices of the parameters, one gets a trivial inequality that says that a non-positive quantity is bounded above by a positive quantity. But for other choices still, one gets a reverse log-Sobolev inequality in the Segal-Bargmann space. The best form (given by the method of this article) of the reverse log-Sobolev inequality is

$$\langle f,Nf\rangle\leqslant cS(f)+n\left(-1+c\log\left(\frac{4c^2}{(2c-1)(c-1)}\right)\right)||f||_2^2$$

for any $c>1$. This result is based on an earlier work of the author in Ref. 9 which in turn depends on an analysis of the reproducing kernel in the Segal-Bargmann space. Surprisingly, we obtain that finite entropy $S(f)$ implies finite energy $\langle f,Nf\rangle$, for all f in the Segal-Bargmann space. That finite energy implies finite entropy in the Segal-Bargmann space was already known by the above cited works.

The article is organized as follows. The next section reviews definitions and notation. Section III gives the statement of the theorem and its proof. In Secs. IV and V there are discussions of the question of best possible constants and of the relation to reverse hypercontractivity. In Sec. VI some applications of the reverse log-Sobolev inequality to the Segal-Bargmann transform are presented. In Sec. VII we conclude with comments and open questions.

II. NOTATION AND DEFINITIONS

We continue with the notation introduced in Ref. 9. We define the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$ (see Refs. 10–12) for each integer $n\geqslant 1$ as the Hilbert space of holomorphic functions $f:\mathbf{C}^n\to\mathbf{C}$ which are square integrable with respect to the Gaussian probability measure

$$d\mu_n(z):=\pi^{-n}\exp(-z^*\cdot z)d^nxd^ny.$$

Here for every $z=(z_1,\ \dots\ ,z_n)\in\mathbf{C}^n$ and $w=(w_1,\ \dots\ ,w_n)\in\mathbf{C}^n$ we define $z^*:=(z_1^*,\ \dots\ ,z_n^*)$ and $z\cdot w:=z_1w_1+\ \dots\ +z_nw_n$, where ω^* is the complex conjugate of the complex number ω. Also, d^nxd^ny denotes Lebesgue measure on $\mathbf{C}^n$. We write $z^2:=z\cdot z$ for $z\in\mathbf{C}^n$ as well. A special case of this notation is $x^2=x\cdot x$ for $x\in\mathbf{R}^n$. Moreover, we put $|z|^2:=z^*\cdot z$. We have

$$HL^2(\mathbf{C}^n,\mu_n)=L^2(\mathbf{C}^n,\mu_n)\cap\mathcal{H}_n$$

where $L^2(\mathbf{C}^n,\mu_n)$ is the space of all functions $f:\mathbf{C}^n\to\mathbf{C}$ that are square integrable with respect to μ_n and where $\mathcal{H}_n$ is the space of all holomorphic functions from $\mathbf{C}^n$ to $\mathbf{C}$.

Suppose (X,μ) is a measure space and that $1\leqslant p\leqslant\infty$. For a measurable function $f:X\to\mathbf{C}$ we denote the usual L^p norm of f by $||f||_{L^p(\mu)}$ or sometimes by $||f||_p$ if we do not need to be so explicit. Also $L^p(X,\mu)$ or simply L^p denotes the corresponding Banach space. We also write

$$HL^p(\mathbf{C}^n,\mu_n)=L^p(\mathbf{C}^n,\mu_n)\cap\mathcal{H}_n$$

for the space of holomorphic L^p functions. Moreover p' is the dual index of p, namely, $p^{-1}+p'^{-1}=1$.

Let (X,μ) be a probability measure space. Then for any function f in $L^2(X,\mu)$ define its entropy by

$$S(f):=\int_X d\mu|f|^2\log|f|^2-||f||_2^2\log||f||_2^2, \tag{2.1}$$

where $\log r$ denotes the natural logarithm (base e) of the real number $r>0$. We shall always use the usual convention that $0\log 0=0$. Then by Jensen's inequality, one has that $S(f)\geqslant 0$ though $S(f)=\infty$ is a possibility. The entropy, like the standard deviation, is a measure of the degree of concentration of a function. It was first introduced by Shannon[13] in information theory.

Next, $\langle f,Nf\rangle$ is the Dirichlet form in the Segal-Bargmann space associated to the non-negative operator N, known as the energy or number operator, which is given by

$$N:=z_1\frac{\partial}{\partial z_1}+\cdots+z_n\frac{\partial}{\partial z_n}.$$

The spectrum of N is the set of all non-negative integers. The Dirichlet form $\langle f,Nf\rangle$ is explicitly defined for *all* functions f in the Segal-Bargmann space by

$$\langle f,Nf\rangle:=\sum_{k=1}^{n}\left\|\frac{\partial f}{\partial z_k}\right\|_2^2, \tag{2.2}$$

where, as usual,

$$\frac{\partial}{\partial z_k}=\frac{1}{2}\left(\frac{\partial}{\partial x_k}-i\frac{\partial}{\partial y_k}\right).$$

Here we write the coordinate $z_k=x_k+iy_k$ where $x_k,y_k\in\mathbf{R}$ and i is the imaginary unit.

III. THE MAIN RESULT

Now we can state the result of this article which gives a parameterized family of energy-entropy inequalities in the Segal-Bargmann space. Under some conditions these are log-Sobolev inequalities but under others they are reverse log-Sobolev inequalities.

Theorem 3.1: For all f in the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$ satisfying $f\in HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$ we have the inequality

$$(p^{-1}-q^{-1})S(f)\leqslant\frac{n}{q}(a-\log(a+1-p'q/4))\|f\|_2^2+\frac{a}{q}\langle f,Nf\rangle \tag{3.1}$$

for $1<p\leqslant\infty$, $1\leqslant q<\infty$ and $a>p'q/4-1$ where $S(f)$ is the entropy defined in Eq. (2.1) and $\langle f,Nf\rangle$ is the Dirichlet form defined in Eq. (2.2). Moreover, this energy-entropy inequality (3.1) holds for all $f\in HL^2(\mathbf{C}^n,\mu_n)$ provided that $S(f)$ is finite and $1<p\leqslant 2$ and $1\leqslant q\leqslant 2$ and $a>p'q/4-1$.

In particular, $S(f)<\infty$ if and only if $\langle f,Nf\rangle<\infty$ for all f in the Segal-Bargmann space.

The coefficient of $\|f\|_2^2$ in (3.1) is always strictly positive. There are three qualitatively different cases for the other two coefficients in (3.1).

Case 1: $p^{-1}>q^{-1}$. It follows that $p'q/4-1>0$ and so $a>0$. The coefficients of $S(f)$ and $\langle f,Nf\rangle$ are positive, and so (3.1) is a (regular) log-Sobolev inequality.

Case 2: $p^{-1}\leqslant q^{-1}$ and $p'q/4-1\geqslant 0$. It follows that $a>0$ and so the coefficient of $S(f)$ is non-positive while the coefficient of $\langle f,Nf\rangle$ is positive. So, (3.1) is trivially true.

Case 3: $p'q/4-1<0$. It follows that $p^{-1}<q^{-1}$ so that the coefficient of $S(f)$ is negative. In this case we can take a with $0>a>p'q/4-1$ so that the coefficient of $\langle f,Nf\rangle$ is negative and so (3.1) is a reverse log-Sobolev inequality. (Taking $a\geqslant 0$, while valid, only yields a trivial inequality.)

Proof of Theorem 3.1: The proof of this theorem will be given in two passes. In the first pass some formulas will be derived in a nonrigorous way, but all the main ideas will be presented. Then in the second pass, the gaps in the first pass will be filled in.

To start the first pass, we first recall some results proved in Ref. 9. There it is shown that for every f in the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$ one has the inequality

$$||f||_{L^{q_t}(\nu_a^t)} \leqslant M^t ||f||_{L^{p_t}(\mu_n)} \tag{3.2}$$

for $0 \leqslant t \leqslant 1$, $1 < p \leqslant \infty$ and $1 \leqslant q < \infty$ with the following notation. First, p_t and q_t are defined implicitly by

$$p_t^{-1} = (1-t)2^{-1} + tp^{-1} \quad \text{and} \quad q_t^{-1} = (1-t)2^{-1} + tq^{-1} \tag{3.3}$$

for each $0 \leqslant t \leqslant 1$. The measure ν_a^t on $\mathbf{C}^n$ is defined by

$$d\nu_a^t(z) := c_a^{tq_t/q} \exp\left(-\frac{taq_t}{q} |z|^2 \right) d\mu_n(z) \tag{3.4}$$

for $z \in \mathbf{C}^n$ and $0 \leqslant t \leqslant 1$ where $c_a := (1+a)^n$ and a satisfies

$$a > p'q/4 - 1. \tag{3.5}$$

Finally, M can be taken to be the smaller of $D^{n/q}$ and $D^{n/p'}$ where

$$D := (a+1)(a+1-p'q/4)^{-1}. \tag{3.6}$$

The results in this paragraph are obtained from part (ii) of Theorem 3.2 of Ref. 9 by taking $a = b$ in the notation of that article. Notice that none of the results of Ref. 9 (nor of the present article) use hypercontractivity results. Rather the result (3.2) is based on interpolation theory.

Notice that for $t=0$, the above inequality (3.2) becomes an *equality* with both sides being $||f||_{L^2(\mu_n)}$. This allows us to use a technique of Hirschman.[14] This starts by fixing an element f in the Segal-Bargmann space as well as fixing the indices p and q such that $1 < p \leqslant \infty$ and $1 \leqslant q < \infty$. Then one takes the derivative of both sides of the inequality (3.2) at $t=0$ in order to generate the new inequality

$$\left.\frac{d}{dt}\right|_{t=0^+} ||f||_{q_t} \leqslant \left.\frac{d}{dt}\right|_{t=0^+} (M^t ||f||_{p_t})$$

which is to say

$$\left.\frac{d}{dt}\right|_{t=0^+} ||f||_{q_t} \leqslant (\log M) ||f||_2 + \left.\frac{d}{dt}\right|_{t=0^+} ||f||_{p_t}, \tag{3.7}$$

where we are only using the right hand derivative, since each side of the inequality (3.2) is a real valued function of t where $0 \leqslant t \leqslant 1$. Of course, in general the derivative is not an order preserving operator; however, in this situation the right hand derivative at $t=0$ is order preserving because of the *equality* in (3.2) at $t=0$.

Naturally, this discussion is still at a formal level, since one must show that the relevant derivatives indeed exist. However, it is also clear at a formal level that the derivative of

$$\psi(r) := \int_{\mathbf{C}^n} d\mu_n(z) |f(z)|^r \tag{3.8}$$

at $r=2$ is given by

$$\psi'(2) = \int_{\mathbf{C}^n} d\mu_n(z) |f(z)|^2 \log|f(z)| \tag{3.9}$$

since this is just differentiation under the integral sign. Using this formal result and the rules of elementary calculus, we get

$$\left.\frac{d}{dt}\right|_{t=0^+}||f||_{p_t}=\frac{p'(0)S(f)}{4||f||_2} \tag{3.10}$$

where $S(f)$ is the entropy of f, given in Eq. (2.1), and $p'(0)=4(2^{-1}-p^{-1})$ is the derivative at $t=0$ of $p(t):=p_t$. We do not intend to make (3.10) rigorous if $||f||_2=0$. However, the inequality (3.1) which we wish to prove is trivially true if $f=0$. So, hereafter, we consider only the case $f\neq 0$.

The derivative of $||f||_{q_t}$ is more complicated because the measure used to compute this norm also depends on t. Using the definition

$$k_t(z):=c_a^{t/q}\exp\left(-\frac{ta}{q}|z|^2\right) \tag{3.11}$$

for $z\in\mathbf{C}^n$ and $0\leqslant t\leqslant 1$ we have more specifically that

$$||f||_{q_t}=||f||_{L^{q_t}(\nu_a^t)}=||fk_t||_{L^{q_t}(\mu_n)}=\left\{\int_{\mathbf{C}^n}d\mu_n(z)(k_t(z)|f(z)|)^{q_t}\right\}^{1/q_t}.$$

It follows that

$$\left.\frac{d}{dt}\right|_{t=0^+}||f||_{q_t}=\frac{q'(0)S(f)}{4||f||_2}+\frac{1}{||f||_2}\int_{\mathbf{C}^n}d\mu_n(z)\left.\frac{dk_t}{dt}\right|_{t=0^+}|f(z)|^2 \tag{3.12}$$

by the rules of calculus again, a differentiation under the integral sign, and the seemingly innocuous, but sometimes false, statement that an integral of a sum of functions is the sum of the integrals. Here $q'(0)=4(2^{-1}-q^{-1})$ is the derivative at $t=0$ of $q(t):=q_t$. (The notation $q'(0)$ introduced here, as well as $p'(0)$ used earlier, has nothing to do with the dual index.) Substituting in the derivative of k_t,

$$\left.\frac{d}{dt}\right|_{t=0^+}k_t(z)=\left.\frac{d}{dt}\right|_{t=0^+}k_1(z)^t=\log k_1(z)=\frac{1}{q}\log c_a-\frac{1}{q}a|z|^2$$

which comes from Eq. (3.11), one gets

$$\left.\frac{d}{dt}\right|_{t=0^+}||f||_{q_t}=\frac{q'(0)S(f)}{4||f||_2}+\frac{n}{q}||f||_2\log(1+a)-\frac{a}{q||f||_2}\int_{\mathbf{C}^n}d\mu_n(z)|f(z)|^2|z|^2 \tag{3.13}$$

since $c_a=(1+a)^n$. Now we can use the formula

$$\int_{\mathbf{C}^n}d\mu_n(z)|f(z)|^2|z|^2=n||f||_2^2+\langle f,Nf\rangle \tag{3.14}$$

which comes from formula (3.17) of Bargmann's article,[14] and is proved there by using the Taylor series expansion of f. I will refer to (3.14) as Bargmann's identity. Here $\langle f,Nf\rangle$ is the Dirichlet form defined in Eq. (2.2). The formula (3.14) is valid for all f in the Segal-Bargmann space. In particular, one side of Eq. (3.14) is infinite if and only if the other side is infinite. One can think of (3.14) as a consequence of the canonical commutation relations. In this regard, see part (vii) of Theorem 13.7 in Ref. 15. Using Bargmann's identity (3.14) in Eq. (3.13), we get

$$\left.\frac{d}{dt}\right|_{t=0^+}||f||_{q_t}=\frac{q'(0)S(f)}{4||f||_2}+\frac{n}{q}||f||_2\log(1+a)-\frac{an}{q}||f||_2-\frac{a}{q||f||_2}\langle f,Nf\rangle. \tag{3.15}$$

Substituting the value $M=D^{n/q}$ and the Eqs. (3.6), (3.10) and (3.15) into inequality (3.7) and rearranging terms, some of which may possibly be infinite, we have

$$\left(\frac{q'(0)-p'(0)}{4}\right)S(f)\leqslant\frac{n}{q}(a-\log(a+1-p'q/4))||f||_2^2+\frac{a}{q}\langle f,Nf\rangle \tag{3.16}$$

for any $1<p\leqslant\infty$ and $1\leqslant q<\infty$ and $a>p'q/4-1$. Substituting the formulas for the derivatives $p'(0)$ and $q'(0)$ into (3.16) gives us precisely (3.1). Of course, we could have taken M to be $D^{n/p'}$ or simply equal to the optimal constant in the inequality (3.2). The value $M=D^{n/q}$ was used just to get a specific value for the coefficient of the norm term in (3.16). The proof of the assertion that the coefficient of the norm term is strictly positive is an exercise in elementary calculus. Using the value $M=D^{n/p'}$ gives a better bound if and only if $p^{-1}+q^{-1}>1$. But the resulting coefficient of the norm term is still strictly positive by another elementary exercise.

Before starting the second pass of the proof, here are some results which will be useful.

Lemma 3.2: For the left side derivative at $r=2$ of $\psi(r)$, as defined in Eq. (3.8), we have

$$\psi'(2^-):=\lim_{h<0,\,h\to 0}\frac{\psi(2+h)-\psi(2)}{h}=\int_{\mathbf{C}^n}d\mu_n(z)|f(z)|^2\log|f(z)| \tag{3.17}$$

for all $f\in L^2(\mathbf{C}^n,\mu_n)$, where the equality includes the case where both sides are $+\infty$. For the right side derivative we have

$$\psi'(2^+):=\lim_{h>0,\,h\to 0}\frac{\psi(2+h)-\psi(2)}{h}=\int_{\mathbf{C}^n}d\mu_n(z)|f(z)|^2\log|f(z)| \tag{3.18}$$

for all $f\in L^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$ where the equality is between finite real numbers.

The proof of this lemma is an exercise using convexity arguments and the monotone convergence theorem, and so will be left to the reader. Let us notice that this lemma implies that $S(f)$ is finite throughout the argument of the theorem. This is because we either have the hypothesis that $f\in L^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$ (in which case we have (3.18) which implies that $S(f)$ is finite) or we have $S(f)$ finite by explicit hypothesis. This justifies the interchange of integral and sum used to derive (3.12), since the first term on the right hand side of (3.12) is finite. Moreover, this implies that (3.10) is an equality between finite real numbers, provided that it is a valid equation, of course. But Lemma 3.2 also justifies (3.10). To show this we note that $||f||_{p_t}=[\psi(p_t)]^{1/p_t}$, where ψ is defined in (3.8) and that p_t approaches 2 monotonically as t decreases to zero by the definition (3.3) of p_t. If $1<p<2$, then p_t increases monotonically to 2 and we use (3.17). If $2<p\leqslant\infty$, then p_t decreases monotonically to 2, and we apply (3.18), which is applicable since we assume that f is in $HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$ when $2<p\leqslant\infty$. In either case we have (3.10) as a simple consequence. If $p=2$, then $p_t=2$ for all $0\leqslant t\leqslant 1$ implying $p'(0)=0$ so that (3.10) follows since each side is zero. This completes the justification of (3.10). Notice that the assumptions on f are what one expects from an examination of (3.2). Specifically, (3.2) is trivial for any f in the Segal-Bargmann space for which $||f||_{p_t}$ is infinite for all $0<t\leqslant 1$, and so one can not calculate a useful (i.e., finite) derivative in such a case. To have a finite derivative, it is necessary that $||f||_{p_t}<\infty$ for all $t\in(0,\delta]$ for some $0<\delta\leqslant 1$. If $1<p\leqslant 2$, then $||f||_{p_t}\leqslant||f||_2<\infty$ since $1<p_t\leqslant 2$. So we can take $\delta=1$ in this case and f arbitrary in the Segal-Bargmann space. If $2<p\leqslant\infty$, then we use the hypothesis that $f\in L^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$ which implies that $||f||_{p_t}\leqslant||f||_{2+\epsilon}<\infty$ whenever $2\leqslant p_t\leqslant 2+\epsilon$, which holds for all $t\in(0,\delta]$ for some $\delta>0$ sufficiently small. So the above argument based on Lemma 3.2 shows that this necessary condition on f for $||f||_{p_t}$ to have a finite derivative is also sufficient if $2<p\leqslant\infty$. If $1<p\leqslant 2$, we simply assume that the derivative is finite, or equivalently that $S(f)$ is finite. In the next lemma we find another consequence of the condition that f is in $HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$.

Lemma 3.3: The Dirichlet form $\langle f,Nf\rangle$ is finite if f is in $HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$.

To show this we note that

$$n\|f\|_2^2+\langle f,Nf\rangle=\int_{\mathbf{C}^n}d\mu_n(z)|f(z)|^2|z|^2\leq\|f\|_{2s}^2\|z\|_{2s'}^2, \tag{3.19}$$

for any $1\leq s\leq\infty$, by an application of Bargmann's identity (3.14) and Hölder's inequality. Here by $\|z\|_{2s'}$ we mean the $L^{2s'}$ norm of the function that maps z in $\mathbf{C}^n$ to $|z|$. By taking $s=1+\epsilon/2$ where f is in $L^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$, we have $s>1$ and so $s'<\infty$. But $\|z\|_{2s'}$ is then finite and so is $\|f\|_{2s}=\|f\|_{2+\epsilon}$. So the right hand side of (3.19) is finite, implying that $\langle f,Nf\rangle$ is finite. The inequality (3.19) shows the importance of the analyticity of the elements of the Segal-Bargmann space; it says that if a function in the Segal-Bargmann space has a bit more integrability than just square integrability, then the function has finite energy. So (3.19) is a *reverse* Sobolev inequality. This result appears to be basically new, though the author has used it previously in Ref. 16. This concludes the proof of Lemma 3.3.

Another basic result relating the quantities that we use in this theorem is the following lemma.

Lemma 3.4: Finite energy $\langle f,Nf\rangle$ implies finite entropy $S(f)$ for all f in the Segal-Bargmann space.

This can be derived from the log-Sobolev inequality $S(f)\leq\langle f,Nf\rangle$ which follows from a hypercontractivity result for the semigroup $\{e^{-tN}\}_{t\geq 0}$ (see Refs. 4–8). (A general reference for log-Sobolev inequalities and hypercontractivity is the article[1] by Gross.) Also, this lemma can be derived directly from the log-Sobolev inequality proved in Ref. 1 for Euclidean space. One then only gets $S(f)\leq 2\langle f,Nf\rangle$, but this is still sufficient to prove this lemma. However, hypercontractivity will not be used to derive any of the results of this article. So here is an elementary proof of the lemma.

First, for any f in the Segal-Bargmann space, one has the well known point-wise estimate $|f(z)|\leq\|f\|_2\exp(|z|^2/2)$ for all $z\in\mathbf{C}^n$. This follows from the fundamental property of the reproducing kernel function and the Cauchy-Schwarz inequality (see Ref. 10). Using this and Bargmann's identity (3.14), we then have for any f in the Segal-Bargmann space that

$$S(f)=\int_{\mathbf{C}^n}d\mu_n(z)|f(z)|^2\log\left(\frac{|f(z)|^2}{\|f\|_2^2}\right)\leq\int_{\mathbf{C}^n}d\mu_n(z)|f(z)|^2|z|^2=n\|f\|_2^2+\langle f,Nf\rangle \tag{3.20}$$

which shows that $\langle f,Nf\rangle<\infty$ implies that $S(f)<\infty$. Notice that the log-Sobolev inequality (3.20) is not as sharp as the log-Sobolev inequality given by the theory of hypercontractivity, but that it serves our purpose here. This ends the proof of Lemma 3.4.

We now start the second pass of the proof of the theorem. The formula (3.12) now has to be justified rigorously. We still must establish the interchange of integral and derivative

$$\left.\frac{d}{dt}\right|_{t=0^+}\int_{\mathbf{C}^n}d\mu_n(z)(k_t(z)|f(z)|)^{q_t}=\int_{\mathbf{C}^n}d\mu_n(z)\left(2|f(z)|^2\left.\frac{dk_t}{dt}\right|_{t=0^+}+q'(0)|f(z)|^2\log|f(z)|\right), \tag{3.21}$$

which is the only remaining step needed to justify (3.12). Inverting the defining relation given in Eq. (3.3) for q_t, we have that

$$t=\frac{q_t^{-1}-2^{-1}}{q^{-1}-2^{-1}}$$

provided that $q\neq 2$. (The case $q=2$ will be considered later.) Using this, and the relation $k_t(z)=(k_1(z))^t$, one can write the integrand on the left hand side of Eq. (3.21) as

$$(k_t(z)|f(z)|)^{q_t}=(k_1(z)^t|f(z)|)^{q_t}=(k_1(z)^{((q_t^{-1}-2^{-1})/(q^{-1}-2^{-1}))}|f(z)|)^{q_t}$$

which leads us to define the function

$$\eta(r) := \int_{\mathbf{C}^n} d\mu_n(z) (k_1(z)^{((r^{-1}-2^{-1})/(q^{-1}-2^{-1}))} |f(z)|)^r \tag{3.22}$$

for $r \in [1,\infty)$. Notice that $\eta(r) \in [0,\infty]$. So the left hand side of Eq. (3.21) is equal to

$$\left.\frac{d}{dt}\right|_{t=0^+} \eta(q(t)) = \eta'(2^{\pm}) q'(0), \tag{3.23}$$

where the right side (respectively, left side) derivative of $\eta(r)$ at $r=2$ is used if $q(t)$ is decreasing (respectively, increasing) to 2 as t decreases to 0. (Because of its definition in (3.3), the convergence of $q(t)$ to 2 is monotone as t decreases to 0.) Notice that the integrand of the expression (3.22) is a convex function of r, so that we may use standard measure theory arguments as before. However, the details are a bit more complicated and so will be presented in full. But before entering into the details of computing $\eta'(2^{\pm})$, notice that η itself depends on q, a and f. So the following argument will be by cases which depend on the hypotheses satisfied by q, a and f.

Also, notice that we will prove the two sides of (3.21) are equal without yet knowing that they are finite. On the right hand side there is no problem with the second term; it is integrable since $S(f)$ is finite. However, the first term, which includes an energy term when it is expanded out, may well be $+\infty$ or $-\infty$ as far as we know in this part of the argument. Similarly, the left hand side of (3.21), which is given again in (3.23), may be $+\infty$ or $-\infty$. Thus as far as we know $\eta'(2^{\pm})$ may be $+\infty$ or $-\infty$ as well. Only after completing this part of the argument will we be able to establish that in fact (3.21) is an equality between finite real numbers.

Before starting the cases, notice that since f is always in the Segal-Bargmann space, $\eta(2) = ||f||_2^2$ is finite. To compute the one-sided derivatives of $\eta(r)$ at $r=2$, we start with the difference quotient

$$\frac{\eta(2+h)-\eta(2)}{h} = \int_{\mathbf{C}^n} d\mu_n(z) k_1(z)^{\alpha} \left[\frac{(k_1(z)^{\beta}|f(z)|)^{2+h} - (k_1(z)^{\beta}|f(z)|)^2}{h} \right] \tag{3.24}$$

for $h \neq 0$, where $\alpha = 1/(q^{-1}-2^{-1})$ and $\beta = -2^{-1}/(q^{-1}-2^{-1})$. The pointwise limit of the integrand on the right hand side of (3.24), as $h \to 0$, is

$$\gamma(z) := |f(z)|^2 \log|f(z)| + \beta |f(z)|^2 \log k_1(z). \tag{3.25}$$

The goal is to prove that $\eta'(2^{\pm}) = \int_{\mathbf{C}^n} d\mu_n(z) \gamma(z)$, and then use (3.23) to show (3.21), since $q'(0)\gamma(z)$ is the integrand on the right hand side of (3.21). Since $k_1(z) > 0$, the convergence of the integrand of (3.24) to $\gamma(z)$ is monotone increasing as $h \nearrow 0$ (i.e., $h<0$, $h \to 0$), and it is monotone decreasing as $h \searrow 0$ (i.e., $h>0$, $h \to 0$). The difference quotient in (3.24) is a finite number if and only if the integrand on the right hand side of (3.24) is integrable if and only if $\eta(2+h)$ is a finite number. But by Hölder's inequality, we have

$$\begin{aligned} \eta(2+h) &= \int_{\mathbf{C}^n} d\mu_n(z) k_1(z)^{hq/(q-2)} |f(z)|^{2+h} \\ &\leq ||k_1^{hq/(q-2)}||_{s'} \, || \, |f|^{2+h} ||_s \\ &= (||k_1||_{hqs'/(q-2)})^{hq/(q-2)} (||f||_{(2+h)s})^{2+h} \end{aligned} \tag{3.26}$$

for any $1 \leq s \leq \infty$.

Now we are ready to argue by cases. The reader should be aware that these cases are not the same as the three cases in the statement of the theorem. First, let us consider the case of the left side derivative of $\eta(r)$ at $r=2$. Notice that this arises in the argument when we have $q_t \nearrow 2$, which itself occurs exactly when $q<2$. The idea is to use (3.26) to show $\eta(2+h)<\infty$ for some $-1 \leqslant h<0$ and apply the monotone convergence theorem, since the monotone increasing integrands in (3.24) will then be uniformly bounded below by an integrable function for all $h_0 \in [h,0)$. For each case, the condition imposed on a must be combined with the overall condition given in (3.5).

Case A: $q<2$, $a \geqslant 0$, $a>p'q/4-1$, $f \in HL^2(\mathbf{C}^n,\mu_n)$.

We take $s=1$. Then the first factor in (3.26) is finite since $a \geqslant 0$ while the second factor is finite for all $-1 \leqslant h<0$.

Case B: $q<2$, $a>p'q/4-1$, $f \in HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$.

Now we take $s>1$ and we use (3.26) for some h satisfying

$$2<(2+h)s \leqslant 2+\epsilon. \tag{3.27}$$

It follows that $\|f\|_{(2+h)s} \leqslant \|f\|_{2+\epsilon}$ and this makes the second factor on right hand side of (3.26) finite. Therefore the right hand side of (3.26) will be finite if and only if $\|k_1\|_{hqs'/(q-2)}$ is finite. So it remains to show that

$$\int_{\mathbf{C}^n} d\mu_n(z) k_1(z)^{hqs'/(q-2)}<\infty \tag{3.28}$$

for some $h<0$ which also satisfies (3.27). But by the definition (3.11) of $k_1(z)$, this last integral is finite if and only if

$$1+hs'a/(q-2)>0 \tag{3.29}$$

which is equivalent to

$$a>(2-q)/hs' \tag{3.30}$$

since $h<0$ and $q-2<0$. The lower bound here, namely $(2-q)/hs'$, is negative and decreases to $-\infty$ as $h \nearrow 0$ for *any* fixed $s>1$. So for $|h|$ sufficiently small, this new lower bound on a is less restrictive than (3.5) for any fixed $s>1$. Moreover, if s is fixed so that $1<s \leqslant 1+\epsilon/2$ (say $s=1+\epsilon/2$), then the inequality (3.27) is also satisfied for all $h<0$ with $|h|$ sufficiently small.

Case C: $q<2$, $a>q/2-1$, $a>p'q/4-1$, $f \in HL^2(\mathbf{C}^n,\mu_n)$.

Notice that the condition (3.27) can be weakened to $(2+h)s=2$. This still makes the second factor on the right hand side of (3.26) finite. The difference here is that now s is not fixed as h approaches zero. However, the argument works for any $f \in HL^2(\mathbf{C}^n,\mu_n)$. Moreover, it turns out that since $s'=-2/h$ the condition on a to make the first factor finite is then

$$a>q/2-1$$

which is independent of h and is more restrictive than (3.5) if $p>2$. Under this new condition on a, we then have Eq. (3.31), which is now an equality between finite real numbers, or both sides are $+\infty$. Weakening (3.27) further to $1 \leqslant (2+h)s<2$ leads to nothing new since the smaller value of s implies a larger value of s', thereby making larger the first factor on the right hand side of (3.26).

So it follows under Case A, B or C that

$$\eta'(2^-)=\int_{\mathbf{C}^n} d\mu_n(z)\gamma(z) \tag{3.31}$$

which is *a priori* either an equality of real numbers, or both sides are $+\infty$. [Recall that $\gamma(z)$ is defined in (3.25).]

Next, let us consider the right side derivative of $\eta(r)$ at $r=2$. Notice that this case occurs in the argument when $q_t \searrow 2$, which happens precisely when $q>2$. Now we want $\eta(2+h)$ to be finite for some $h>0$ so that we can use the monotone convergence theorem for the monotone decreasing integrands in (3.24), which will then be uniformly bounded above by an integrable function for all $h_0 \in (0,h]$. Of course, we can not solve $(2+h)s=2$ for $s \geqslant 1$ and $h>0$, so we need a bit more integrebility of f to make the right hand side of (3.26) finite. Now we want to satisfy $(2+h)s \leqslant 2+\epsilon$ for some $h>0$ and $s \geqslant 1$, so that the second factor on the right hand side of (3.26) is finite.

Case D: $q>2$, $a \geqslant 0$, $a>p'q/4-1$, $f \in HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$.

We take $s=1$ so that the first factor on the right hand side of (3.26) is finite since $a \geqslant 0$, which may be more restrictive than (3.5).

Case E: $q>2$, $a>p'q/4-1$, $f \in HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$.

Now we choose $s>1$ so the first factor is finite on the right hand side of (3.26) if and only if (3.29) holds as before, and this is again equivalent to (3.30), except now because $h>0$ and $q-2>0$. Again, the lower bound $(2-q)/hs'$ in (3.30) is negative and decreases to $-\infty$ as $h \searrow 0$, and the resulting condition on a is weaker than (3.5) for all $h>0$ with h sufficiently small. Also, by fixing s with $1<s<1+\epsilon/2$, we can satisfy (3.27) for all $h>0$ with h sufficiently small.

So it follows for Cases D and E that

$$\eta'(2^+) = \int_{\mathbf{C}^n} d\mu(z)\,\gamma(z)$$

which is an equality between real numbers or both sides are $-\infty$.

The formula (3.21) must still be justified in the case $q=2$. But then $q_t=2$ for all $0 \leqslant t \leqslant 1$, and so the integrand of the left hand side of (3.21) is

$$(k_t(z)|f(z)|)^{q_t} = k_1(z)^{2t}|f(z)|^2.$$

So the difference quotient used to compute the derivative on the left hand side of (3.21) is

$$\int_{\mathbf{C}^n} d\mu_n(z)\left(\frac{k_1(z)^{2h}-1}{h}\right)|f(z)|^2 \tag{3.32}$$

for $h>0$. The pointwise limit of the integrand here as $h \searrow 0$ is

$$(2\log k_1(z))|f(z)|^2 = 2|f(z)|^2 \left.\frac{dk_t}{dt}\right|_{t=0^+}$$

and the convergence is monotone decreasing. This is the integrand on the right hand side of (3.21) since $q'(0)=0$ in this case. So to interchange the limit and the integral using the monotone convergence theorem, we need the integral in (3.32) to be finite for some $h>0$. This is equivalent to the integrability of $k_1(z)^{2h}|f(z)|^2$. However, by Hölder's inequality

$$\int_{\mathbf{C}^n} d\mu_n(z) k_1(z)^{2h}|f(z)|^2 \leqslant \|k_1^{2h}\|_{s'} \| \, |f|^2 \|_s = (\|k_1\|_{2hs'})^{2h}(\|f\|_{2s})^2 \tag{3.33}$$

for any $1 \leqslant s \leqslant \infty$.

Case F: $q=2$, $a \geqslant 0$, $a>p'q/4-1$, $f \in HL^2(\mathbf{C}^n,\mu_n)$.

Since $a \geqslant 0$, we take $s=1$ in (3.33) since k_1 is then bounded and f is in $L^2(\mathbf{C}^n,\mu_n)$.

Case G: $q=2$, $a>p'q/4-1$, $f \in HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$.

Now we take $s>1$ satisfying $2s \leqslant 2+\epsilon$. So $\|f\|_{2s} \leqslant \|f\|_{2+\epsilon} < \infty$. To show the other factor in (3.33) is finite is equivalent to showing

$$\int_{\mathbf{C}^n} d\mu_n(z) k_1(z)^{2hs'} < \infty$$

and this is equivalent to $a>-q/2hs'$. But for s fixed (and satisfying $1<s<1+\epsilon/2$), the limit of the lower bound $-q/2hs'$ as $h\searrow 0$ is $-\infty$. So for all $h>0$ sufficiently small, this new restriction on a is weaker than (3.5) and so we have $\|k_1\|_{2hs'}<\infty$ in this case, and it follows that (3.21) holds in the case $q=2$, provided a satisfies (3.5).

So it follows for Cases F and G that

$$\eta'(2^+)=\int_{\mathbf{C}^n} d\mu(z)\,\gamma(z)$$

which is an equality between real numbers or both sides are $-\infty$.

At this point of the argument, we have established (3.21) in the seven cases A through G described above, but with the possibility that the expression in (3.21) could be infinite. And (3.21) in turn implies (3.12), but again with the understanding that both sides of (3.12) could be $+\infty$ or $-\infty$.

Now we will show that $S(f)$ finite implies that $\langle f,Nf\rangle$ is finite for any f in the Segal-Bargmann space. We take any p and q satisfying $1<p\leqslant 2$, $1\leqslant q<2$, and $p'q<4$. (These conditions are consistent and imply $q^{-1}>p^{-1}$.) Now we can apply Case C above, which is valid for all f in the Segal-Bargmann space, provided that a satisfies the two conditions specified in that case. But $p'q/4-1\geqslant q/2-1$ since $p'\geqslant 2$ so that a must satisfy merely the one condition $a>p'q/4-1$. This means we can (and do) pick a with $a<0$, this being possible since $p'q<4$. Then the inequality (3.7) says

$$\frac{q'(0)S(f)}{4\|f\|_2}+\frac{1}{\|f\|_2}\int_{\mathbf{C}^n} d\mu_n(z)|f(z)|^2\log k_1(z)\leqslant(\log M)\|f\|_2+\frac{p'(0)S(f)}{4\|f\|_2} \tag{3.34}$$

using (3.10) and (3.12). Since we are assuming that $S(f)$ is finite this can be rewritten as

$$\int_{\mathbf{C}^n} d\mu_n(z)|f(z)|^2\log k_1(z)\leqslant(\log M)\|f\|_2^2+(q^{-1}-p^{-1})S(f) \tag{3.35}$$

using the formulas for the derivatives $p'(0)$ and $q'(0)$. Now,

$$\log k_1(z)=\frac{n}{q}\log(1+a)-\frac{a}{q}|z|^2$$

so that (3.35) becomes

$$\left(\frac{n}{q}\log(1+a)-\frac{na}{q}\right)\|f\|_2^2-\frac{a}{q}\langle f,Nf\rangle\leqslant(\log M)\|f\|_2^2+(q^{-1}-p^{-1})S(f)$$

by an application of Bargmann's identity (3.14), and this is true even if $\langle f,Nf\rangle$ is not assumed to be finite, that is, even if (3.12) happened to be an equality between infinite quantities. In the last inequality the coefficient of $\langle f,Nf\rangle$ is positive since $a<0$ while the coefficient of $S(f)$ is positive since $q^{-1}>p^{-1}$. So $S(f)$ finite does imply $\langle f,Nf\rangle$ finite for all f in the Segal-Bargmann space, as claimed.

The result of the last paragraph together with Lemma 3.4 establishes that the entropy $S(f)$ is finite if and only if the energy $\langle f,Nf\rangle$ is finite, this being the third assertion of the theorem.

We are now ready to prove the first assertion in the theorem. So we have by hypothesis that $f\in HL^{2+\epsilon}(\mathbf{C}^n,\mu_n)$ for some $\epsilon>0$ and $1<p\leqslant\infty$ and $1\leqslant q<\infty$ and $a>p'q/4-1$. The assumption on f implies that $\langle f,Nf\rangle$ is finite as proved in Lemma 3.3. And this in turn implies that $S(f)$ is finite. We now apply Case B, Case E or Case G depending on the value of q. This gives us (3.21) and hence (3.12). To arrive at (3.16) from (3.12) now comes from simple substitution and the rearrangement of terms, all of which are now known to be finite. This concludes the demonstration of the first assertion of the theorem.

We proceed to the second assertion in the theorem. Now the hypotheses are that f is in the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$, $S(f)$ is finite, $a>p'q/4-1$, $1<p\leqslant 2$ and $1\leqslant q\leqslant 2$. If $1\leqslant q<2$, we use Case C. If $q=2$, we use Case F, which requires that $a\geqslant 0$. But $1<p\leqslant 2$ and $q=2$ imply $p'q/4-1=p'/2-1\geqslant 0$. So the hypothesis $a>p'q/4-1$ implies $a\geqslant 0$ and so Case F does apply. This gives us (3.12). Since $S(f)$ is finite by hypothesis, we have that $\langle f,Nf\rangle$ is finite by the above argument. So we can rearrange terms to arrive at (3.16).

Finally, the three cases in the statement of the theorem follow immediately from an elementary analysis of the coefficients of the energy and entropy terms. And this concludes the proof of the theorem. □

Cases A and D of the proof of this theorem were never used subsequently in that argument. They are included only for the sake of giving a complete analysis of all the possible cases. They do not imply the inequality (3.1) in any case not already established.

The inequality (3.1) has two limiting cases, one when the coefficient of the entropy term is zero and the other when the coefficient of the energy term is zero. The coefficient of the entropy being zero means that $p^{-1}=q^{-1}$ and (3.1) becomes trivial, as was already noted in Case 2 of the statement of the theorem. The coefficient of the energy being zero means $a=0$, and this implies that $p'q/4-1<a=0$ which in turn implies $p^{-1}<q^{-1}$. So again (3.1) becomes trivial since the left hand side is a non-positive quantity while the right hand side is non-negative. However, this leads one to ask if an inequality of the form $BS(f)\leqslant ||f||_2^2$ holds for all f in the Segal-Bargmann space with $B>0$. The negative answer to this question is provided by considering the holomorphic functions f_λ defined by

$$f_\lambda(z):=\exp\left(\frac{\sqrt{2}\lambda\cdot z}{2}-\frac{\lambda^2}{4}\right)$$

where $\lambda\in\mathbf{R}^n$ and $z\in\mathbf{C}^n$. These functions are in the Segal-Bargmann space. They also satisfy $||f_\lambda||_2=1$ and $S(f_\lambda)=|\lambda|^2/2$ by direct evaluation of the relevant Gaussian integrals. These functions also show that no inequality of the form $C||f||_2^2\leqslant S(f)$ with $C>0$ holds for all f in the Segal-Bargmann space. Essentially, this says that $S(f)$ and $||f||_2^2$ are inequivalent ways to gauge the size of f. This is to be expected; $S(f)$ measures the concentration of a state (that is, f satisfying $||f||_2=1$), and this concentration should be independent of the normalization of f. The analogous result in $L^2(\mathbf{C}^n,\mu_n)$ can be shown with characteristic functions, and so holds in quite great generality. The above example shows that the result still holds in the rather ''small'' subspace $HL^2(\mathbf{C}^n,\mu_n)$. The exponential functions f_λ also saturate the log-Sobolev inequality $S(f)\leqslant\langle f,Nf\rangle$ in the Segal-Bargmann space, and I conjecture that modulo a multiplicative constant these are the only functions to do so.

IV. BEST CONSTANTS

While the best constants for the reverse log-Sobolev inequality are not yet known, some partial results are given in this section. These have the effect of showing what the best constants *cannot* be, as well as what are the best constants given by the method of this article. First, we define

$$g_\lambda(z):=(1-\lambda^2)^{n/4}\exp(\lambda z^2/2) \tag{4.1}$$

for $-1<\lambda<1$ and $z\in\mathbf{C}^n$. Then g_λ is holomorphic with $||g_\lambda||_2=1$. So g_λ is in the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$. Moreover, we can calculate

$$\langle g_\lambda,Ng_\lambda\rangle=\frac{n\lambda^2}{1-\lambda^2}$$

and

$$S(g_\lambda)=\frac{n}{2}\log(1-\lambda^2)+\frac{n\lambda^2}{1-\lambda^2}.$$

But we can also write this in the form

$$\langle g_\lambda, Ng_\lambda\rangle = S(g_\lambda)-\frac{n}{2}(\log(1-\lambda^2))||g_\lambda||_2^2. \tag{4.2}$$

However, $-\log(1-\lambda^2)\to+\infty$ as $|\lambda|$ increases to 1 so that there is no reverse log-Sobolev inequality of the form

$$\langle f,Nf\rangle \leqslant cS(f)+n\beta||f||_2^2 \tag{4.3}$$

with $c\leqslant 1$. Also, let us note in passing that the functions g_λ show that the relation $S(f)=\langle f,Nf\rangle$ does not hold for all f in the Segal-Bargmann space.

Moreover, the reverse log-Sobolev inequality proved here can be brought into the form of (4.3) for any $c>1$ as the following argument shows. In this discussion we assume that $p'q<4$ and $0>a>p'q/4-1$, which is precisely the case of the reverse log-Sobolev inequality for (3.1). However, instead of analyzing (3.1), this analysis will start from

$$(p^{-1}-q^{-1})S(f)\leqslant \frac{a}{q}\langle f,Nf\rangle + n\left(\frac{a}{q}+\left(\frac{1}{p'}-\frac{1}{q}\right)\log(a+1)-\frac{1}{p'}\log(a+1-p'q/4)\right)||f||_2^2$$

which follows from (3.7) by using the value $M=D^{n/p'}$ instead of $M=D^{n/q}$ which was used to derive (3.1). This amounts to nothing more than a change in the coefficient of the norm term, but this will give a better constant than (3.1). First, write this in the form

$$\langle f,Nf\rangle \leqslant \left(\frac{q}{a}\right)(p^{-1}-q^{-1})S(f)+n\left(-1+\frac{q}{a}\left(\frac{1}{q}-\frac{1}{p'}\right)\log(a+1)+\frac{q}{ap'}\log(a+1-p'q/4)\right)||f||_2^2$$

by rearranging terms and multiplying by $-q/a$, which is positive. For fixed p and q, the coefficient of $S(f)$ is bounded below by

$$\inf_{0>a>p'q/4-1}\left(\frac{q}{a}\right)(p^{-1}-q^{-1})=\frac{4p^2-4p-4pq+4q}{4p^2-4p-p^2q} \tag{4.4}$$

though the infimum is not achieved as a runs over its allowed interval. Next, it is easy to show that the right hand side of (4.4) being $\geqslant 1$ is equivalent to $(p-2)^2\geqslant 0$, with the right hand side being equal to 1 if and only if $p=2$. So the coefficient of $S(f)$ is bounded below by 1, and it can be made to approach 1 precisely in the case $p=2$. So we now take $p=2$ and keep q fixed so that the standing assumptions become $q<2$ and $0>a>q/2-1$. We then get

$$\langle f,Nf\rangle \leqslant a^{-1}(q/2-1)S(f)+n\left(-1+\left(\frac{2-q}{2a}\right)\log(a+1)+\frac{q}{2a}\log(a+1-q/2)\right)||f||_2^2,$$

where the coefficient $a^{-1}(q/2-1)$ is greater than 1 and approaches 1 as a decreases to $q/2-1$ and approaches $+\infty$ as a increases to 0. This means that we can realize (4.3) for any $c>1$ as claimed. In fact, in (4.3) we can take

$$\beta=-1+\left(\frac{2-q}{2a}\right)\log(a+1)+\frac{q}{2a}\log(a+1-q/2).$$

Notice that then $\beta \to +\infty$ as a decreases to $q/2-1$. Now one would like to know the dependence of the coefficient $n\beta$ of the norm term as a function of the coefficient $c := a^{-1}(q/2-1)$ of the entropy term. Substituting $a = c^{-1}(q/2-1)$ we find that

$$\beta(q) = -1 - c \log(q/2 - 1 + c) + c \log c - \left(\frac{cq}{2-q}\right) \log\left(\frac{(2-q)(c-1)}{2c}\right)$$

for $c>1$ and $1 \leqslant q < 2$. By calculus, one verifies that $\beta(q)$ is a strictly increasing function for each fixed $c>1$ and for $1 \leqslant q < 2$. [Hint: Show $\beta''(q)>0$ and $\beta'(1)>0$. This implies that $\beta'(q)>0$ for $q \geqslant 1$.] So $q=1$ minimizes $\beta(q)$ and the minimum value is

$$\beta(1) = -1 + c \log\left(\frac{4c^2}{(2c-1)(c-1)}\right).$$

The appearance of the non-zero norm term in (4.3) is a bit troublesome. For example, the optimal form of the regular log-Sobolev inequality in the Segal-Bargmann space does not have such a term. Anyway, the best form obtainable by the method of this article is

$$\langle f, Nf \rangle \leqslant cS(f) + n\left(-1 + c \log\left(\frac{4c^2}{(2c-1)(c-1)}\right)\right) ||f||_2^2$$

for any $c>1$.

There is no point in optimizing (3.1) in the case that it is a log-Sobolev inequality, since the coefficient of the norm term in (3.1) is positive, while the optimal log-Sobolev inequality in the Segal-Bargmann space is $S(f) \leqslant \langle f, Nf \rangle$.

V. REVERSE HYPERCONTRACTIVITY

In Ref. 4, Carlen proves the hypercontractivity for the semigroup $\{e^{-tN}\}$ from $HL^p(\mathbf{C}^n, \mu_n)$ to $HL^q(\mathbf{C}^n, \mu_n)$ for $0<p<q$ and $t \geqslant 0$ provided that $t \geqslant \frac{1}{2}\log(q/p)$. Here one defines $e^{-tN}f(z) := f(e^{-t}z)$ for z in $\mathbf{C}^n$ and f in $HL^p(\mathbf{C}^n, \mu_n)$ using the motivation that this formula for the semigroup can be proved in the case $p=2$ where one uses spectral theory to define the semigroup. As remarked earlier, similar results are to be found in Refs. 5–8. However, Carlen also proves a result for e^{-tN} when t is negative, using the same defining formula for e^{-tN}. He shows that if $0<q<p$ and $t>\frac{1}{2}\log(q/p)$, then e^{-tN} is bounded from $HL^p(\mathbf{C}^n, \mu_n)$ to $HL^q(\mathbf{C}^n, \mu_n)$ with norm bounded above by $(1-e^{-2t}q/p)^{-n/q}$. This result is trivially true if $t \geqslant 0$, since it is easy to show that e^{-tN} has norm 1 in that case. But for $\frac{1}{2}\log(q/p)<t<0$, this is a new and interesting result. The question which we will address here is the relation of this reverse hypercontractivity result due to Carlen to the reverse log-Sobolev inequality of this article. Before proceeding, let us note that Carlen's results include Lebesgue indices between 0 and 1, but in this article we only have used Lebesgue indices between 1 and $+\infty$ due to the limitations of interpolation theory. The following discussion will be limited to the latter case.

Define $C(t,q,p,n)$ to be the supremum of $||e^{-tN}f||_q$ where $f:\mathbf{C}^n \to \mathbf{C}$ is holomorphic and $||f||_p = 1$. Then $C(t,q,p,n)$ could be infinite, though when it is finite it is just the above operator norm. Using this notation Carlen's result says that for $1<q \leqslant p$

$$||e^{-tN}f||_q \leqslant C(t,q,p,n) ||f||_p, \tag{5.1}$$

where $C(t,q,p,n)$ is finite, $t>\frac{1}{2}\log(q/p)$ and f is in $HL^p(\mathbf{C}^n, \mu_n)$. We use $C(t,q,p,n)$ here and not the estimate which Carlen derives, since his bound is not optimal. It remains an open problem to find a formula for $C(t,q,p,n)$ when $q<p$. Suppose that $p>1$ and $\epsilon>0$ are fixed. Suppose that s is a function mapping $(-\epsilon, 0]$ to $(1, +\infty)$. Suppose that s satisfies these conditions:

$$\text{(a)} \quad s(t) < pe^{2t} \quad \text{for} \quad -\epsilon < t < 0;$$

$$\text{(b)} \quad s(0)=p; \tag{5.2}$$

$$\text{(c)} \quad s'(0^-) \text{ exists.}$$

Here $s'(0^-)$ denotes the left side derivative of $s(t)$ at $t=0$. Then by condition (a) we can apply (5.1) to get

$$\|e^{-tN}f\|_{s(t)} \leqslant C(t,s(t),p,n)\|f\|_p \tag{5.3}$$

for all $t\in(-\epsilon,0)$. Moreover, (5.3) is valid also for $t=0$. In fact, it reduces to an equality for $t=0$, since $C(0,s(0),p,n)=1$ by condition (b). Notice that it is in this step that one uses the optimality of the constant $C(t,q,p,n)$. If we use Carlen's bound $M(t,q,p,n):=(1-e^{-2t}q/p)^{-n/q}$ instead, we find that $M(t,s(t),p,n)$ is undefined at $t=0$. Since the inequality (5.3) becomes an equality at the right end-point of its interval of validity (namely at $t=0$), the operation of taking the one-sided derivative at that end-point reverses the sense of the inequality, giving

$$\left.\frac{d}{dt}\right|_{t=0^-}\|e^{-tN}f\|_{s(t)} \geqslant \left.\frac{d}{dt}\right|_{t=0^-} C(t,s(t),p,n)\|f\|_p .$$

Proceeding here formally, we have

$$\left.\frac{d}{dt}\right|_{t=0^-}\|e^{-tN}f\|_{s(t)}=(p^{-2}s'(0^-)S_p(f)-\mathrm{Re}\langle f_p,Nf\rangle)\|f\|_p^{p-1},$$

where $f_p:=f|f|^{p-2}=\mathrm{sgn}(f)|f|^{p-1}$ and $S_p(f):=S(|f|^{p/2})$. One can find sufficient conditions for this interchange of derivative and integral in Refs. 1 and 2. Since $C(t,s(t),p,n)\geqslant 1$ (because $e^{-tN}1=1$ where 1 is the constant function) and $C(0,s(0),p,n)=1$, we see that $\kappa:=(d/dt)|_{t=0^-}C(t,s(t),p,n)\leqslant 0$. Moreover by conditions (a) and (b) of (5.2), $s'(0^-)\geqslant 2p$, which is positive. So we have shown

$$p^{-2}s'(0^-)S_p(f)-\mathrm{Re}\langle f_p,Nf\rangle \geqslant \kappa\|f\|_p^p,$$

which rearranges to

$$\mathrm{Re}\langle f_p,Nf\rangle \leqslant p^{-2}s'(0^-)S_p(f)-\kappa\|f\|_p^p. \tag{5.4}$$

For $p=2$, (5.4) becomes the following reverse log-Sobolev inequality

$$\langle f,Nf\rangle \leqslant \tfrac{1}{4}s'(0^-)S(f)-\kappa\|f\|_2^2, \tag{5.5}$$

which motivates calling (5.4) an index p reverse log-Sobolev inequality.

It seems now to be just a matter of arguing a bit more carefully to show that Carlen's reverse hypercontractivity result implies the reverse log-Sobolev inequality. But unfortunately this is not the case. First, note that there is an apparent contradiction between (5.5) and a result of Sec. IV. This is because the above formal argument should work for *any* function $s(t)$ satisfying the three conditions in (5.2). Let us consider the case $p=2$. We can take $s(t)$ so that it has first order contact at $t=0$ with the function $\phi(t):=2e^{2t}$ that majorizes $s(t)$. That is, we can pick $s(t)$ such that $s'(0^-)=\phi'(0^-)=4$. But then the coefficient of $S(f)$ in (5.5) is 1, and this seems to contradict what we have already shown in Sec. IV. This would indeed be a contradiction if the norm term in (5.5) were finite. So, we can conclude that the coefficient of the norm term is not finite or, equivalently, that $C(t,s(t),2,n)$ is not differentable from the left at $t=0$. This lack of smoothness must be inherited from a lack of smoothness in the operator norm $C(t,q,2,n)$ since $s(t)$ is by construction differentiable from the left at $t=0$. This means that at least one of the partial derivatives $\partial C/\partial t$ or $\partial C/\partial q$ does not exist when evaluated at the appropriate point. And this lack of smoothness in the operator norm then aborts any attempt to make rigorous the argument leading to (5.5), even in the ''good'' case when $s'(0^-)>4$. One can actually see directly from spectral theory that $C(t,2,2,n)=1$ for $t\geqslant 0$ and $C(t,2,2,n)=+\infty$ for $t<0$. And this shows that $(\partial C/\partial t)\times(t,2,2,n)$ evaluated at $t=0^-$ does not exist.

The above argument also can not be made rigorous for $p>2$. Using the functions g_λ from (4.1) we have in general for real λ satisfying $|\lambda|<\min(2/p,1)$

$$||g_\lambda||_p=(1-\lambda^2)^{n/4}(1-p^2\lambda^2/4)^{-n/2p},$$

$$\mathrm{Re}\langle (g_\lambda)_p, Ng_\lambda\rangle = np\lambda^2 2^{n+1}(1-\lambda^2)^{np/4}(4-p^2\lambda^2)^{-1-n/2},$$

$$S_p(g_\lambda)=np^2\lambda^2 2^n(1-\lambda^2)^{np/4}(4-p^2\lambda^2)^{-1-n/2} + (n/2)(1-\lambda^2)^{np/4}(1-p^2\lambda^2/4)^{-n/2}\log(1-p^2\lambda^2/4),$$

so that

$$\mathrm{Re}\langle (g_\lambda)_p, Ng_\lambda\rangle=(2/p)S_p(f)-(n/p)(\log(1-p^2\lambda^2/4))||g_\lambda||_p^p.$$

But $-(n/p)\log(1-p^2\lambda^2/4)\to+\infty$ as $|\lambda|$ increases to $2/p$, which is less than 1 for $p>2$. This shows that an index p reverse log-Sobolev inequality of the form

$$\mathrm{Re}\langle f, Nf\rangle \leqslant c_p S_p(f)+n\beta_p||f||_p^p \tag{5.6}$$

cannot hold for $c_p\leqslant 2/p$. But by picking $s(t)$ so that it has first order contact with $\psi(t):=pe^{2t}$ at $t=0$, one sees that the coefficient of the entropy term in (5.4) can be made equal to $2/p$. And so we observe the same sort of near contradiction as in the case $p=2$, and consequently the derivation of (5.4) fails.

However, it may be possible to make this argument rigorous in the case $p<2$. The above construction does not invalidate (5.4) in this case, since $|\lambda|$ can only increase to $1<2/p$. So it seems reasonable to conjecture that (5.4) holds for $p<2$ and that the above argument will establish this provided one can show that $C(t,q,p,n)$ is differentiable in this case. If this is so, it might be possible then to prove the index 2 reverse log-Sobolev inequality by taking the limit as p increases to 2 of the index p reverse log-Sobolev inequality.

The results of this section are a bit disappointing, since what at first looked like a straightforward way of relating reverse hypercontractivity to the reverse log-Sobolev inequality has not given us a positive result. Unfortunately, the usual method of smoothing out estimates by using interpolation theory is not available here in spaces of holomorphic functions. So, it still remains an open question whether an index p reverse log-Sobolev inequality (5.6) holds in the Segal-Bargmann space when $p\neq 2$.

VI. APPLICATIONS

The result of this article can be used to derive in a new way some results in Ref. 16. This makes the nature of those results more transparent. Let us recall in this paragraph some basic results from Refs. 10–12 expressed in the formulation of Ref. 16. One defines the configuration space (in the ground state representation) to be $L^2(\mathbf{R}^n,\nu_n)$ where ν_n is the Gaussian probability measure given by

$$d\nu_n(x):=\pi^{-n/2}\exp(-x^2)d^n x$$

for $x\in\mathbf{R}^n$ where $d^n x$ is Lebesgue measure on $\mathbf{R}^n$. Then the Segal-Bargmann transform A is defined by

$$Af(z):=\int_{\mathbf{R}^n} d\nu_n(x) A(z,x) f(x)$$

for $z\in\mathbf{C}^n$ and $f:\mathbf{R}^n\rightarrow\mathbf{C}^n$ measurable, provided that the integral converges absolutely. The kernel function is defined by $A(z,x):=\exp(-z^2/2+\sqrt{2}z\cdot x)$ for $z\in\mathbf{C}^n$ and $x\in\mathbf{R}^n$. Then by Hölder's inequality $Af(z)$ is well-defined for every z in $\mathbf{C}^n$ and for every element f of $L^p(\mathbf{R}^n,\nu_n)$ if $1<p\leqslant\infty$. Moreover, A maps $L^2(\mathbf{R}^n,\nu_n)$ unitarily onto the Segal-Bargmann space $HL^2(\mathbf{C}^n,\mu_n)$, and it furthermore intertwines the standard representations of the Heisenberg-Weyl group (exponentiated canonical commutation relations) on the spaces $L^2(\mathbf{R}^n,\nu_n)$ and $HL^2(\mathbf{C}^n,\mu_n)$. In particular, A preserves the energy, which can be expressed in terms of the Dirichlet forms as

$$\langle\psi,N_1\psi\rangle=\langle A\psi,NA\psi\rangle \tag{6.1}$$

for *all* $\psi\in L^2(\mathbf{R}^n,\nu_n)$ where

$$\langle\psi,N_1\psi\rangle:=\frac{1}{2}\sum_{j=1}^{n}\left|\left|\frac{\partial\psi}{\partial x_k}\right|\right|^2_{L^2(\nu_n)}$$

defines the Dirichlet form for *all* $\psi\in L^2(\mathbf{R}^n,\nu_n)$. (The factor of one-half in the previous definition corrects an error in Ref. 16.) Notice that one side of (6.1) is infinite if and only if the other side is infinite.

Entropy also enters in the analysis of the Segal-Bargmann space via the Segal-Bargmann transform, since that transform does not preserve entropy. One of the results from Ref. 16 which restricts how the Segal-Bargmann transform A can change entropy is a Hirschman type inequality (see Ref. 14) of the form

$$S(\psi)\leqslant c_1 S(A\psi)+c_2||\psi||_2^2 \tag{6.2}$$

with c_1 and c_2 being non-negative constants and ψ being in $L^2(\mathbf{R}^n,\nu_n)$. This inequality was derived by studying the L^p to L^q mapping properties of A for (p,q) close to $(2,2)$. Now it can be proved simply by noting

$$S(\psi)\leqslant 2\langle\psi,N_1\psi\rangle=2\langle A\psi,NA\psi\rangle\leqslant 2(a_1S(A\psi)+a_2||A\psi||_2^2)=2a_1S(A\psi)+2a_2||\psi||_2^2,$$

where the first step is the log-Sobolev inequality in $L^2(\mathbf{R}^n,\nu_n)$ originally proved in Ref. 1, the next step is just (6.1), the next step is the reverse log-Sobolev inequality in the Segal-Bargmann space and the last step is the unitarity of A. In this proof, one still uses mapping properties of A (preservation of norm and energy), but the energy-entropy properties pertain now to each space separately.

Another related fact about entropy in the Segal-Bargmann space is that there is no Hirschman type inequality of the form

$$S(A\psi)\leqslant b_1S(\psi)+b_2||\psi||_2^2 \tag{6.3}$$

for all $\psi\in L^2(\mathbf{R}^n,\nu_n)$ with b_1,b_2 non-negative constants. First, we can explicitly construct elements ψ in $L^2(\mathbf{R}^n,\nu_n)$ with finite entropy and infinite energy. I claim that any such ψ violates (6.3). This is because the right hand side of (6.3) is finite by construction. However, the left hand side of (6.3) must be infinite. Otherwise we would have $S(A\psi)<\infty$ implying $\langle A\psi,NA\psi\rangle<\infty$ by the reverse log-Sobolev inequality in the Segal-Bargmann space, and so $\langle\psi,N_1\psi\rangle<\infty$ because A preserves energy. But this contradicts the construction of ψ with infinite energy. So $S(A\psi)=\infty$ follows and (6.3) fails. Notice again the failure of (6.3) follows from energy-entropy properties of the individual spaces themselves plus the elementary properties that A preserves norm and energy. If one is inclined to think in categorical terms, one can formulate the result that (6.3) is false as saying that a certain commutative diagram cannot be constructed. This formulation is left to the interested reader. Also, using the functions f_λ introduced at the very end of Sec. III, one can show that none of the other possible non-trivial forms of a Hirschman type inequality in the Segal-Bargmann space can be valid.

VII. CONCLUSION

The result of this article shows a remarkable relation between entropy and energy in the Segal-Bargmann space. It seems reasonable to suppose that this comes about because the elements in the Segal-Bargmann space are holomorphic functions, rather than just merely square-integrable functions. For example, the proof here depends on the analysis of the reproducing kernel in Ref. 9, and the existence of a reproducing kernel is a hallmark of Hilbert spaces of holomorphic functions. It would be interesting therefore to investigate other Hilbert spaces of holomorphic functions to see if a natural reverse log-Sobolev inequality can be demonstrated even though a regular log-Sobolev inequality might not hold.

The presence of the norm term in the reverse log-Sobolev inequality proved here may not be optimal. It seems natural to conjecture that this term can be eliminated, and with it the only dependence on dimension in the coefficients of the inequality. This would then lead to an infinite dimensional version of the reverse log-Sobolev inequality. In any event, it would be interesting to know what is the exact relation of energy and entropy in the case of infinitely many dimensions.

Despite the results of Sec. V, there still might be another way of relating reverse hypercontractivity to the reverse log-Sobolev inequality. One might try to use the usual method of generating (regular) hypercontractivity inequalities from (regular) log-Sobolev inequalities. That method originated in Ref. 1 and has been modified in Ref. 5 for the holomorphic case. It is an open problem whether one can ''turn around'' the inequalities in that method to produce an argument for the reverse inequalities. As noted before, it remains an open problem if there is an index p reverse log-Sobolev inequality in the Segal-Bargmann space for $p \neq 2$. Another open problem from Sec. V is to determine the exact value of the operator norm $C(t,q,p,n)$ and hence whether it is differentiable at $t=0^-$ when $q<p$.

The failure of the Hirschman type inequality (6.3) is due to the fact that the Segal-Bargmann transform of a function with finite entropy can be a function with infinite entropy. It remains an open problem to identify the image of the finite entropy functions under the Segal-Bargmann transform.

Finally, the question of what are the best constants in the reverse log-Sobolev inequality remains an open problem. Let me just note that this question is independent of whether the norm term is really present or not.

ACKNOWLEDGMENTS

It is a great pleasure to thank Lawrence Thomas for initiating my interest in the Segal-Bargmann analysis. I thank Leonard Gross for bringing to my attention the reverse hypercontractivity result of Carlen and for various useful comments. I also thank Brian Hall for many stimulating discussions.

[1] L. Gross, ''Logarithmic Sobolev inequalities,'' Am. J. Math. **97**, 1061–1083 (1975).
[2] L. Gross, *Logarithmic Sobolev Inequalities and Contractivity Properties of Semigroups,* in *Lecture Notes in Mathematics* edited by G. Dell'Antonio and U. Mosco (Springer, Berlin, 1993), Vol. 1563, pp. 54–58.
[3] P. Diaconis and L. Saloff-Coste, ''Logarithmic Sobolev inequalities for Markov chains,'' Ann. Appl. Probab. **6**, 695–750 (1996).
[4] E. Carlen, ''Some integral identities and inequalities for entire functions and their application to the coherent state transform,'' J. Funct. Anal. **97**, 231–249 (1991).
[5] L. Gross, ''Hypercontractivity over complex manifolds,'' Cornell preprint, March 1998.
[6] S. Janson, ''On hypercontractivity for multipliers on orthogonal polynomials,'' Ark. Mat. **21**, 97–110 (1983).
[7] S. Janson, ''On complex hypercontractivity,'' J. Funct. Anal. **151**, 270–280 (1997).
[8] Z. Zhou, ''The contractivity of the free Hamiltonian semigroup in the L_p space of entire functions,'' J. Funct. Anal. **96**, 407–425 (1991).
[9] S.B. Sontz, ''On the reproducing kernel of the Segal-Bargmann space,'' J. Math. Phys. **40**, 1664–1676 (1999).

[10] V. Bargmann, ''On a Hilbert space of analytic functions and an associated integral transform, Part I,'' Commun. Pure Appl. Math. **14**, 187–214 (1961).
[11] I.E. Segal, *Mathematical Problems of Relativistic Physics,* Chap. VI in *Proceedings of the Summer Seminar*, Boulder, Colorado, 1960, edited by M. Kac, Lectures in Applied Mathematics (American Mathematical Society, Providence, RI, 1963), Vol 2.
[12] I.E. Segal, *The Complex Wave Representation of the Free Boson Field,* in *Topics in Functional Analysis: Essays Dedicated to M.G. Krein on the Occasion of His 70th Birthday,* edited by I. Gohberg and M. Kac, Advances in Mathematics Supplementary Studies (Academic, New York, 1978), Vol. 3, pp. 321–343.
[13] C. Shannon and W. Weaver, *A Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).
[14] I.I. Hirschman, Jr., ''A note on entropy,'' Am. J. Math. **79**, 152–156 (1957).
[15] S. Janson, *Gaussian Hilbert Spaces* (Cambridge University Press, Cambridge, 1997).
[16] S.B. Sontz, ''Entropy and the Segal-Bargmann transform,'' J. Math. Phys. **39**, 2402–2417 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

New approach to semiclassical analysis in mechanics

M. A. Alonso[a] and G. W. Forbes[b]
Department of Physics, Macquarie University, Sydney, New South Wales 2109, Australia

(Received 8 June 1998; accepted for publication 16 December 1998)

A new method is proposed for constructing approximate solutions to the Schrödinger equation. In place of the wave function, its Gaussian-windowed Fourier transform is used as the fundamental entity. This allows an intuitively attractive connection to be made with a family of classical trajectories and, at all times, the wave function is inferred from the present state of these trajectories. The fact that the connection between the wave function and the classical trajectories is consistently constructed in phase space allows this method to be free of the limitations of other methods. © *1999 American Institute of Physics.* [S0022-2488(99)02803-0]

I. INTRODUCTION

While classical mechanics models the motion of a particle in terms of a well-defined trajectory, quantum mechanics characterizes its evolution in a probabilistic fashion by means of a wave function $\Psi(\mathbf{x},t)$ that satisfies the Schrödinger equation:

$$i\hbar \frac{\partial\Psi}{\partial t}(\mathbf{x},t) = -\frac{\hbar^2}{2m}\nabla^2\Psi(\mathbf{x},t) + V(\mathbf{x},t)\Psi(\mathbf{x},t). \tag{1.1}$$

Here, m is the particle's mass, $V(\mathbf{x},t)$ is the potential, and $\hbar$ is Planck's constant. The spatial probability density for the location of the particle at time t is then $|\Psi(\mathbf{x},t)|^2$.

Except for some simple potentials, the determination of an exact solution of Eq. (1.1) given some initial condition, say $\Psi(\mathbf{x},t_0)$, is generally a formidable task. Schemes to construct approximate solutions are therefore important. The so-called *semiclassical* or *quasiclassical* methods offer one such option that is essentially expressed in terms of the classical description of the problem. These methods are direct analogs of ray optics, and allow intuition of the classical macroscopic world to be applied to quantum mechanics. An invaluable review of hundreds of key papers in this area was recently presented by Gutzwiller.[1]

In several of these methods, the wave function is associated with an n-parameter family of classical trajectories, where n is the number of spatial dimensions. Each of these trajectories carries an associated weight or amplitude. At any given time, the position and momentum of each trajectory identify a point in phase space, and the locus of all these points is known as the *Lagrange manifold*. For simplicity, the methods described in this work are presented for the one-dimensional case $(n=1)$. The Lagrange manifold or *phase space curve* is then composed of the points $(x,p)=[X(\xi,t),P(\xi,t)]$, where ξ is the parameter that labels the trajectories. Each trajectory, and hence the phase space curve itself, evolves in time according to the classical laws:

$$\frac{\partial X}{\partial t}(\xi,t) = \frac{P(\xi,t)}{m}, \tag{1.2a}$$

$$\frac{\partial P}{\partial t}(\xi,t) = -\frac{\partial V}{\partial x}[X(\xi,t),t]. \tag{1.2b}$$

[a]Electronic mail: alonso@physics.mq.edu.au
[b]Electronic mail: forbes@physics.mq.edu.au

There are several alternative prescriptions for estimating the wave function from the classical framework given by $[X(\xi,t),P(\xi,t)]$ once it is supplemented by the relative weight of each trajectory, say $D(\xi)$ (assumed for the moment to be independent of t). Some of the well-known methods are revisited in Sec. IX. The most direct of these estimates couples the spatial probability density to the spatial density of the trajectories, that is, $|\Psi[X(\xi,t),t]|^2 \propto D(\xi)/X'(\xi,t)$, where the prime denotes derivative with respect to ξ. This relation holds only while the phase space curve corresponds to a single-valued function of x [i.e., provided $X'(\xi,t)\neq 0$ for all ξ]. However, as X and P evolve according to Eqs. (1.2), this condition is eventually violated in most cases, and the estimate then diverges at the so-called *caustics* [where $X'(\xi,t)=0$].

Another possibility is first to estimate the *momentum distribution*, which is just the Fourier transform of $\Psi(x,t)$:

$$\tilde{\Psi}(p,t)=\frac{1}{\sqrt{2\pi\hbar}}\int \Psi(x,t)\exp\left(-i\frac{xp}{\hbar}\right)dx, \tag{1.3a}$$

$$\Psi(x,t)=\frac{1}{\sqrt{2\pi\hbar}}\int \tilde{\Psi}(p,t)\exp\left(i\frac{xp}{\hbar}\right)dp. \tag{1.3b}$$

Now, the probability density in momentum is coupled to the trajectory density in momentum, i.e., $|\tilde{\Psi}[P(\xi,t),t]|^2 \propto D(\xi)/P'(\xi,t)$. In this case, the phase space curve must be single valued in p, i.e., $P'(\xi,t)\neq 0$ for all ξ. Again, in most applications, this condition is ultimately violated.

These conditions on the form of the phase space curve are evidently tied to the representation that is used. Our aim is to state the connection between a wave function and an associated phase space curve in a way that is not inherently coupled to a particular representation. This is shown to lead to a more robust semiclassical method that is based upon the concept of representing the wave function as a phase space distribution by means of the *Gaussian-windowed Fourier transformation* (GWFT) defined in Sec. II. The evolution equation for this phase space distribution is derived in Sec. III. A form for the GWFT of a wave function is devised in Sec. IV so that the result is localized around a prescribed curve in phase space. The properties of this construction are studied in Sec. V and, in Sec. VI, it is shown that the result can readily be made to satisfy the evolution equation presented in Sec. III. The new estimate for the wave function is stated in Sec. VII, and some criteria for its use are given in Sec. VIII. Section IX gives a comparison with related methods. Finally, a summary of the new semiclassical method is given in Sec. X.

II. GAUSSIAN-WINDOWED FOURIER TRANSFORM

The GWFT of $\Psi(x,t)$ is defined here as

$$G(x,p;t)=\text{GWFT}\{\Psi(x,t)\}:=\frac{1}{\sqrt{2\pi\hbar}}l^{-1/4}\int \Psi(x',t)\exp\left[-\frac{(x'-x)^2}{2l\hbar}-\frac{ip}{\hbar}\left(x'-\frac{x}{2}\right)\right]dx', \tag{2.1a}$$

where ''$:=$'' means ''is defined to be equal to'' and l determines the width of the Gaussian window. The GWFT preserves its general form when expressed in terms of $\tilde{\Psi}(p,t)$, as can be seen upon substituting Eq. (1.3b) into (2.1a):

$$G(x,p;t)=\frac{1}{\sqrt{2\pi\hbar}}l^{1/4}\int \tilde{\Psi}(p',t)\exp\left[-\frac{l}{2\hbar}(p'-p)^2+\frac{ix}{\hbar}\left(p'-\frac{p}{2}\right)\right]dp'. \tag{2.1b}$$

The GWFT is linear in the wave function and, for the purposes to be considered here, this gives it several advantages over other phase space distributions. For example, it is straightforward to fully recover $\Psi(x,t)$ or $\tilde{\Psi}(p,t)$ from $G(x,p;t)$:

$$\Psi(x,t)=\frac{1}{\sqrt{2\pi\hbar}}l^{1/4}\int G(x,p;t)\exp\left(i\frac{px}{2\hbar}\right)dp, \tag{2.2a}$$

$$\tilde{\Psi}(p,t)=\frac{1}{\sqrt{2\pi\hbar}}l^{-1/4}\int G(x,p;t)\exp\left(-i\frac{px}{2\hbar}\right)dx. \tag{2.2b}$$

It is easily seen from either of Eqs. (2.1) that $G(x,p;t)$ satisfies

$$\left(l\frac{\partial}{\partial x}-i\frac{\partial}{\partial p}\right)G(x,p;t)=-\frac{1}{2\hbar}(x-ilp)G(x,p;t). \tag{2.3}$$

Notice that, given Eq. (2.3), the specification of G along a curve is sufficient in principle to infer its value over all of phase space. As a result, Ψ itself is fully determined once G is prescribed on a curve. While this option turns out to be unworkable, in general, Eq. (2.3) has implications in any work with the GWFT.

A useful insight into the structure of $|G|$ follows upon writing

$$G(x,p;t)=\exp[\gamma(x,p;t)+i\phi(x,p;t)], \tag{2.4}$$

where γ and ϕ are real functions. Two real-valued equations follow upon substituting Eq. (2.4) into Eq. (2.3):

$$\frac{\partial\gamma}{\partial x}+\frac{1}{l}\frac{\partial\phi}{\partial p}+\frac{x}{2\hbar l}=0, \tag{2.5a}$$

$$\frac{\partial\phi}{\partial x}-\frac{1}{l}\frac{\partial\gamma}{\partial p}-\frac{p}{2\hbar}=0. \tag{2.5b}$$

Equations (2.5) can now be decoupled by taking both the x and p derivatives of each of them, and then eliminating the cross-derivative terms to find

$$\left(\frac{\partial^2}{\partial x^2}+\frac{1}{l^2}\frac{\partial^2}{\partial p^2}\right)\gamma(x,p;t)=-\frac{1}{\hbar l}, \tag{2.6a}$$

$$\left(\frac{\partial^2}{\partial x^2}+\frac{1}{l^2}\frac{\partial^2}{\partial p^2}\right)\phi(x,p;t)=0. \tag{2.6b}$$

Notice that Eqs. (2.6) involve the Laplacian operator in the scaled phase space with coordinates (x,lp). From Eqs. (2.5) [especially through Eq. (2.6a)] it follows that if $|G|$ is localized along a curve—i.e., it has a ridge along which the first and second derivatives of γ are small—then the transverse profile of the ridge in this scaled phase space is approximately a Gaussian of width $\sqrt{\hbar l}$.

III. TIME EVOLUTION EQUATION FOR THE GWFT

The time evolution equation for G can be found by taking the GWFT of both sides of Eq. (1.1). From the p derivative of Eq. (2.1a), it is easy to show that

$$\mathrm{GWFT}\{x\Psi(x,t)\}=\hat{X}G(x,p;t):=\left(\frac{x}{2}+i\hbar\frac{\partial}{\partial p}\right)G(x,p;t), \tag{3.1}$$

where $\hat{X}$ is then the position operator for G and, more generally,

$$\mathrm{GWFT}\{x^n\Psi(x,t)\}=\hat{X}^n G(x,p;t). \tag{3.2}$$

By using integration by parts, it can be seen that

$$\mathrm{GWFT}\left\{\frac{\partial\Psi}{\partial x}(x,t)\right\}=\frac{-1}{\sqrt{2\pi\hbar}}l^{-1/4}\int \Psi(x',t)\frac{\partial}{\partial x'}\exp\left[-\frac{(x'-x)^2}{2l\hbar}-\frac{ip}{\hbar}\left(x'-\frac{x}{2}\right)\right]dx'$$
$$=\left(-\frac{x}{2l\hbar}+\frac{ip}{\hbar}+\frac{i}{l}\frac{\partial}{\partial p}\right)G(x,p;t), \tag{3.3}$$

where Eq. (3.1) has been used in the last step. This result can be written in a more suggestive form by using Eq. (2.3) to eliminate the p derivative and multiplying both sides by $-i\hbar$:

$$\mathrm{GWFT}\left\{-i\hbar\frac{\partial\Psi}{\partial x}(x,t)\right\}=\hat{P}G(x,p;t):=\left(\frac{p}{2}-i\hbar\frac{\partial}{\partial x}\right)G(x,p;t), \tag{3.4}$$

and, therefore,

$$\mathrm{GWFT}\left\{\left(-i\hbar\frac{\partial}{\partial x}\right)^n\Psi(x,t)\right\}=\hat{P}^nG(x,p;t). \tag{3.5}$$

By using Eqs. (3.2) and (3.5), the result of taking the GWFT of both sides of Eq. (1.1) can be written as

$$i\hbar\frac{\partial G}{\partial t}(x,p;t)=\frac{1}{2m}\hat{P}^2G(x,p;t)+V(\hat{X},t)G(x,p;t), \tag{3.6}$$

where V has been assumed to be analytic. This form of the Schrödinger equation turns out to be ideally suited for semiclassical analysis, and enables a robust connection to the classical domain.

IV. CONSTRUCTING *G* BY USING A PHASE SPACE CURVE

As mentioned in the Introduction, several semiclassical methods have been derived for cases where $\Psi(x,t)$ can be associated with a one-parameter family of classical trajectories, which, at any given time, is represented by a curve in phase space. However, this association is straightforward only when the curve corresponds to a single-valued function of x or p. This is a significant limitation, and the objective of this section is to give a form for a wave function that can be associated with a more general family of classical trajectories.

To begin, consider the GWFT of a Gaussian that has width $\sqrt{\hbar\Lambda}$ and a linear phase proportional to P:

$$\mathrm{GWFT}\left\{\exp\left[-\frac{(x-X)^2}{2\hbar\Lambda}+i\frac{P}{\hbar}\left(x-\frac{X}{2}\right)\right]\right\}=\frac{l^{1/4}\sqrt{\Lambda}}{\sqrt{l+\Lambda}}\exp\left[-\frac{(x-X)^2}{2\hbar(l+\Lambda)}-\frac{l\Lambda(p-P)^2}{2\hbar(l+\Lambda)}\right.$$
$$\left.+i\frac{2(\Lambda xP-lXp)+(l-\Lambda)(xp+XP)}{2\hbar(l+\Lambda)}\right]$$
$$=\frac{l^{1/4}\sqrt{\Lambda}}{\sqrt{l+\Lambda}}\exp\left[-\frac{z_{-l}(z_\Lambda-Z_\Lambda)-Z_\Lambda(z_{-l}-Z_{-l})}{2\hbar(l+\Lambda)}\right], \tag{4.1}$$

where the expression was simplified in the last step by using

$$z_\eta:=x+i\eta p, \tag{4.2a}$$

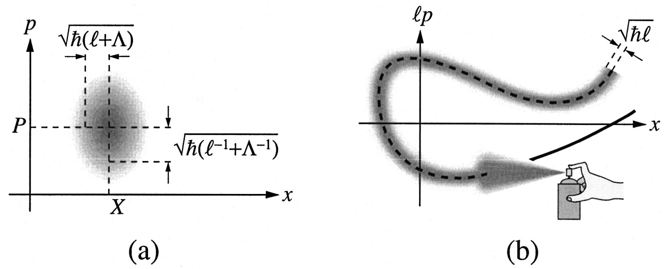

FIG. 1. (a) Dimensions in phase space of the Gaussian given in Eq. (4.1). (b) Construction of a wave function based on painting its GWFT over a curve in phase space. The grey level in both (a) and (b) represents $|G(x,p)|$. This process uses the Gaussian shown in (a) as the footprint of the notional spray can. Notice that the resulting ridge in $|G|$ has a transverse width that is essentially independent of the dimensions of the footprint.

$$Z_\eta := X + i\eta P, \tag{4.2b}$$

for η either $-l$ or Λ. The phase space distribution given in Eq. (4.1) is a Gaussian localized around (X,P), as shown in Fig. 1(a).

A form of G that is localized around a curve in phase space described by $[X(\xi),P(\xi)]$ can be built by superimposing a weighted sequence of Gaussians of the form given in Eq. (4.1):

$$G_\Lambda(x,p) = \frac{1}{\sqrt{2\pi\hbar}} \int w_\Lambda(\xi) g_\Lambda(x,p;\xi)\, d\xi, \tag{4.3}$$

where $w_\Lambda(\xi)$ is a weighting factor, and

$$g_\Lambda(x,p;\xi) = \frac{l^{1/4}\sqrt{\Lambda}}{\sqrt{l+\Lambda}} \exp\left\{ i\frac{\varphi(\xi)}{\hbar} - \frac{z_{-l}[z_\Lambda - Z_\Lambda(\xi)] - Z_\Lambda(\xi)[z_{-l} - Z_{-l}(\xi)]}{2\hbar(l+\Lambda)} \right\}. \tag{4.4}$$

The extra phase $\varphi(\xi)/\hbar$ in Eq. (4.4) could have been incorporated into $w_\Lambda(\xi)$, but it is included separately here for convenience.

The properties of $G_\Lambda(x,p)$ can be discussed more easily in terms of the picture presented in Fig. 1(b). With this, Eq. (4.3) associates a wave function to a curve in phase space by painting the GWFT of the wave function using something analogous to a spray can that is traced over the curve. The footprint of the spray can corresponds to $g_\Lambda(x,p;\xi)$. Increasing Λ gives a footprint that is wider in x and narrower in p. One might then think that the form of $G_\Lambda(x,p)$ would depend strongly on Λ. However, it is possible to choose $w_\Lambda(\xi)$ to ensure that G_Λ is insensitive to variations in Λ. This can be realized asymptotically when certain quantities turn out to be much larger than $\hbar$. The signature of the footprint then becomes effectively irrelevant in the continuous superposition. The weak dependence of G_Λ on Λ is of central importance, since it means that the corresponding wavefunction depends on just the essential elements of the classical framework, namely the phase space curve and its weighting factor, $w_\Lambda(\xi)$.

In the construction proposed in Eq. (4.3), $|G_\Lambda|$ is expected to take its largest values near the phase space curve. [Remember that g_Λ is largest at $z_{-l} = Z_{-l}$.] It is therefore required that this superposition of Gaussians is constructive in the neighborhood of the curve. From Eq. (4.4), the ξ derivative of g_Λ is seen to be given by

$$\begin{aligned}\frac{\partial g_\Lambda}{\partial \xi} &= \left[\frac{(z_{-l} - Z_{-l})}{\hbar(l+\Lambda)} \frac{\partial Z_\Lambda}{\partial \xi} + \frac{1}{2\hbar(l+\Lambda)} \left(\frac{\partial Z_\Lambda}{\partial \xi} Z_{-l} - Z_\Lambda \frac{\partial Z_{-l}}{\partial \xi} \right) + \frac{i}{\hbar} \frac{\partial \varphi}{\partial \xi} \right] g_\Lambda \\ &= \left[\frac{(z_{-l} - Z_{-l})}{\hbar(l+\Lambda)} \frac{\partial Z_\Lambda}{\partial \xi} + \frac{i}{\hbar} \left(\frac{\partial \varphi}{\partial \xi} - \frac{P}{2}\frac{\partial X}{\partial \xi} + \frac{X}{2}\frac{\partial P}{\partial \xi} \right) \right] g_\Lambda .\end{aligned} \tag{4.5}$$

Under the assumption that $w_\Lambda(\xi)$ does not vary significantly over the spread of g_Λ, constructive interference occurs along the curve if $\partial g_\Lambda/\partial\xi=0$ at $z_{-l}=Z_{-l}$. It then follows from Eq. (4.5) that $\varphi(\xi)$ must satisfy

$$\frac{\partial\varphi}{\partial\xi}(\xi)=\frac{1}{2}\left[P(\xi)\frac{\partial X}{\partial\xi}(\xi)-X(\xi)\frac{\partial P}{\partial\xi}(\xi)\right], \tag{4.6}$$

and Eq. (4.5) then reduces to

$$\frac{\partial g_\Lambda}{\partial\xi}=\frac{Z'_\Lambda\Delta}{\hbar}g_\Lambda, \tag{4.7}$$

where $\Delta:=(z_{-l}-Z_{-l})/(l+\Lambda)$, and the prime denotes a derivative with respect to ξ. Notice that the right-hand side of Eq. (4.6) is independent of Λ. Therefore, φ could depend on Λ only through an additive term that is independent of ξ. This term is now taken to be absorbed by w_Λ, so that φ is then independent of Λ.

It follows from Eqs. (4.3) and (4.4) that

$$\frac{\partial G_\Lambda}{\partial\Lambda}(x,p)=\frac{1}{\sqrt{2\pi\hbar}}\int\left\{\frac{\partial w_\Lambda}{\partial\Lambda}+w_\Lambda\left[\frac{l}{2\Lambda(l+\Lambda)}+\frac{\Delta^2}{2\hbar}\right]\right\}g_\Lambda(x,p;\xi)d\xi. \tag{4.8}$$

Notice that the second term inside the brackets on the right-hand side of Eq. (4.8) depends not only on ξ but also on x and p, due to the factor of Δ^2. This dependence can be eliminated by repeatedly using Eq. (4.7) and integrating by parts:

$$\frac{\partial G_\Lambda}{\partial\Lambda}=\frac{1}{\sqrt{2\pi\hbar}}\left\{\int\left[\frac{\partial w_\Lambda}{\partial\Lambda}+\frac{lw_\Lambda}{2\Lambda(l+\Lambda)}\right]g_\Lambda\,d\xi-\int\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda\Delta}{2Z'_\Lambda}\right)g_\Lambda\,d\xi\right\}$$

$$=\frac{1}{\sqrt{2\pi\hbar}}\int\left\{\frac{\partial w_\Lambda}{\partial\Lambda}+\frac{w_\Lambda}{2\Lambda}\frac{X'}{Z'_\Lambda}+\frac{\hbar}{2}\frac{\partial}{\partial\xi}\left[\frac{1}{Z'_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda}{Z'_\Lambda}\right)\right]\right\}g_\Lambda\,d\xi. \tag{4.9}$$

The only dependence on x and p in the final integrand of Eq. (4.9) is now within $g_\Lambda(x,p;\xi)$, so like Eq. (4.3), Eq. (4.9) is simply a superposition of Gaussians in phase space. More importantly, integration by parts has now clarified the asymptotic significance of each of the terms; this device is also used repeatedly below. Given the factor of $\hbar$ on the third term inside the braces of Eq. (4.9), w_Λ is now chosen to make the first two terms cancel in order to achieve the desired insensitivity. This condition fixes the dependence of $w_\Lambda(\xi)$ on Λ:

$$w_\Lambda(\xi)=\sqrt{D(\xi)Z'_\Lambda(\xi)/\Lambda}, \tag{4.10}$$

where D is an arbitrary function of ξ.

With this form for $w_\Lambda(\xi)$, G_Λ can be estimated at the phase space curve by using the saddle point method:[2]

$$G_\Lambda[X(\xi_0),P(\xi_0)]\sim l^{1/4}\sqrt{\frac{D(\xi_0)}{Z'_{-l}(\xi_0)}}\exp\left[i\frac{\varphi(\xi_0)}{\hbar}\right]. \tag{4.11}$$

As desired, this result is independent of Λ and does not vanish as long as $D(\xi_0)$ is nonzero. The asymptotic estimate of G_Λ can be found to decay as a Gaussian in any section that is transverse to the curve. (The saddle point then corresponds to a complex value of ξ.) Alternatively, it follows from the discussion at the end of Sec. II that, since $|G_\Lambda|$ evidently has a ridge along the phase space curve, it therefore has a transverse profile that is approximately Gaussian of width $\sqrt{\hbar l}$ in the scaled phase space with coordinates (x,lp).

By using Eqs. (4.10) and Eq. (4.9), it is found that

$$G_{\Lambda+\delta\Lambda}(x,p)\approx G_\Lambda(x,p)+\delta\Lambda\,\frac{\partial G_\Lambda}{\partial\Lambda}(x,p)=\frac{1}{\sqrt{2\pi\hbar}}\int\left\{w_\Lambda(\xi)+\delta\Lambda\,\frac{\hbar}{2}\,\delta w_\Lambda(\xi)\right\}g_\Lambda(x,p;\xi)d\xi, \tag{4.12}$$

where

$$\delta w_\Lambda(\xi):=\frac{1}{\sqrt{\Lambda}}\,\frac{\partial}{\partial\xi}\left(\frac{1}{Z'_\Lambda}\,\frac{\partial}{\partial\xi}\,\sqrt{\frac{D}{Z'_\Lambda}}\right). \tag{4.13}$$

Equation (4.12) states that, to first order, a change in Λ is equivalent to a change in the weighting. Given Eq. (4.11), it now follows that G_Λ is not significantly affected by a change $\delta\Lambda$ in Λ, provided $\delta\Lambda\ll|2w_\Lambda/(\hbar\,\delta w_\Lambda)|$ for all ξ. If we require G_Λ to be insensitive to changes in Λ that are of the order of Λ itself, this condition becomes

$$\left|\frac{\delta w_\Lambda(\xi)}{w_\Lambda(\xi)}\right|=\left|\frac{1}{\sqrt{DZ'_\Lambda}}\,\frac{\partial}{\partial\xi}\left(\frac{1}{Z'_\Lambda}\,\frac{\partial}{\partial\xi}\,\sqrt{\frac{D}{Z'_\Lambda}}\right)\right|\ll\frac{1}{\hbar\Lambda},\quad\text{for all}\quad\xi. \tag{4.14}$$

That is, as anticipated, G_Λ is insensitive to changes in Λ provided $\hbar^{-1}$ is sufficiently large compared to the value of a particular expression involving the curve and its weight function.

V. INTERPRETATION OF THE INSENSITIVITY CONDITION

Condition (4.14) can be used to delineate an acceptable interval for the value of Λ. Expanding the derivatives and multiplying both sides by $\hbar\Lambda$ leads to

$$\left|\left[5Z''^2_\Lambda-2Z'_\Lambda Z'''_\Lambda-4\,\frac{D'}{D}\,Z'_\Lambda Z''_\Lambda+\left(2\,\frac{D''}{D}-\frac{D'^2}{D^2}\right)Z'^2_\Lambda\right]\frac{\hbar\Lambda}{4Z'^4_\Lambda}\right|\ll 1,\quad\text{for all}\quad\xi. \tag{5.1}$$

[Remember that $Z_\Lambda(\xi)=X(\xi)+i\Lambda P(\xi)$.] Notice that, as $\Lambda\to 0$, condition (5.1) is satisfied for all ξ, except in the neighborhood of the values of ξ for which $X'(\xi)=0$. At these values, the left-hand side of condition (5.1) diverges—it is $O(\Lambda^{-3})$. Therefore, the points in the phase space curve where the local tangent is parallel to the p axis [that is, where $X'(\xi)=0$], fix a lower bound on Λ. Alternatively, for large Λ, condition (5.1) is satisfied everywhere except in the neighborhood of the points where the local tangent to the phase space curve is parallel to the x axis, that is, where $P'(\xi)=0$. At these points the left-hand side of condition (5.1) diverges—it is $O(\Lambda^3)$. This then fixes an upper bound for Λ. Since most phase space curves will have both vertical and horizontal segments, finite, nonzero values of Λ must be used and the allowed interval evidently follows from condition (5.1).

The functions $X(\xi)$, $P(\xi)$, $\varphi(\xi)$, and $D(\xi)$, can be reparametrized without affecting Eq. (4.3) or condition (4.14), provided the parametrization monotonically describes the phase space curve. If the relationship between the old and new parameters is $\xi=R(\check{\xi})$, then the reparametrized quantities are $\check{X}(\check{\xi}):=X[R(\check{\xi})]$, $\check{P}(\check{\xi}):=P[R(\check{\xi})]$, $\check{\varphi}(\check{\xi}):=\varphi[R(\check{\xi})]$, and $\check{D}(\check{\xi}):=R'(\check{\xi})D[R(\check{\xi})]$. A simple geometrical interpretation of condition (4.14) follows upon choosing the special parametrization where $|\check{Z}'_\Lambda(\check{\xi})|\equiv\sqrt{\Lambda}$ for all $\check{\xi}$. In this case, $\check{Z}'_\Lambda$ reduces to $\sqrt{\Lambda}\exp[i\alpha(\check{\xi})]$, where

$$\alpha(\check{\xi})=\arctan\left[\frac{\Lambda\check{P}'(\check{\xi})}{\check{X}'(\check{\xi})}\right]. \tag{5.2}$$

Now, consider the scaled phase space $(\Lambda^{-1/2}x,\Lambda^{1/2}p)$, where both axes again have the same dimensions. Here, $\alpha(\check{\xi})$ corresponds to the angle between the $\Lambda^{-1/2}x$ axis and the local tangent to the scaled phase space curve (see Fig. 2), and $\check{\xi}$ corresponds to the arclength along the curve. With this, condition (4.14) becomes

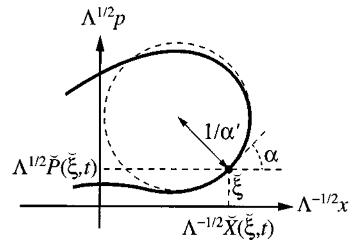

FIG. 2. At any point, labeled by $\check{\xi}$, along the scaled phase space curve, α is the angle between the $\Lambda^{-1/2}x$ axis and the local tangent to the curve, and α' is the local curvature.

$$\frac{1}{4}\left|2\frac{\check{D}''}{\check{D}}-\frac{\check{D}'^2}{\check{D}^2}-3\alpha'^2-2i\left(\alpha''+2\alpha'\frac{\check{D}'}{\check{D}}\right)\right|\ll\frac{1}{\hbar},\quad \text{for all } \check{\xi}, \tag{5.3}$$

where α' and α'' correspond, respectively, to the local curvature and the rate of change of the curvature of the scaled phase space curve.

Condition (5.3) is satisfied when any segment of length $\sqrt{\hbar}$ of the scaled phase space curve is sufficiently straight, and the relative variation in $\check{D}$ is small over the segment. That is, it is sufficient to require each term in condition (5.3) to be small, and this leads to

$$|\alpha'|\ll\frac{1}{\sqrt{\hbar}}, \tag{5.4a}$$

$$|\alpha''|\ll\frac{1}{\hbar}, \tag{5.4b}$$

$$\left|\frac{\check{D}'}{\check{D}}\right|\ll\frac{1}{\sqrt{\hbar}}, \tag{5.4c}$$

$$\left|\frac{\check{D}''}{\check{D}}\right|\ll\frac{1}{\hbar},\quad \text{for all } \check{\xi}. \tag{5.4d}$$

Notice that condition (5.4a) is consistent with what was said at the outset of this section: at points where $P'=0$, the curvature is given by $\Lambda^{3/2}P''/X'^2$, so Λ must satisfy $\Lambda\ll|X'^4/\hbar P''^2|^{1/3}$, while at points where $X'=0$, the curvature is $X''/(\Lambda^{3/2}P'^2)$ so Λ must be chosen such that $\Lambda\gg|\hbar X''^2/P'^4|^{1/3}$. It is clear that, when the curvature of the phase space curve is sufficiently tight, these two inequalities can be incompatible. Nevertheless, this limitation of the semiclassical method developed here turns out to be insignificant compared to the analogous limitations of other existing methods.

VI. TIME EVOLUTION

Each of X, P, φ, and D is now taken to be a function, not only of ξ, but also of time. The goal of this section is to show that G_Λ approximately satisfies the propagation equation given in Eq. (3.6) when X and P are, respectively, the parametric position and momentum for a classical trajectory, and φ is simply related to the classical action for that trajectory.

From Eqs. (4.4), (4.6), and (4.10), the time derivatives of g_Λ and w_Λ are

$$\frac{\partial g_\Lambda}{\partial t}=\frac{i}{\hbar}\left(\dot{\varphi}+\frac{X\dot{P}-\dot{X}P}{2}-i\dot{Z}_\Lambda\Delta\right)g_\Lambda, \tag{6.1a}$$

$$\frac{\partial w_\Lambda}{\partial t}=\left(\frac{\dot{D}}{D}+\frac{\dot{Z}'_\Lambda}{Z'_\Lambda}\right)\frac{w_\Lambda}{2}, \tag{6.1b}$$

where the overdot denotes a derivative with respect to t and Δ is defined after Eq. (4.7). It now follows from Eq. (4.3) that

$$i\hbar\frac{\partial G_\Lambda}{\partial t}=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda g_\Lambda\left(-\dot{\varphi}+\frac{\dot{X}P-X\dot{P}}{2}+i\dot{Z}_\Lambda\Delta+\frac{i\hbar}{2}\frac{\dot{D}}{D}+\frac{i\hbar}{2}\frac{\dot{Z}'_\Lambda}{Z'_\Lambda}\right)d\xi. \tag{6.2}$$

Notice that the third term inside the parentheses on the right-hand side of Eq. (6.2) contains a Δ, and therefore it depends on x and p. This dependence can be removed by using Eq. (4.7) and integrating by parts as in Eq. (4.9):

$$\int w_\Lambda\dot{Z}_\Lambda\,\Delta g_\Lambda\,d\xi=-\hbar\int g_\Lambda\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda\dot{Z}_\Lambda}{Z'_\Lambda}\right)d\xi=-\hbar\int w_\Lambda g_\Lambda\left[\frac{\dot{Z}_\Lambda}{w_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda}{Z'_\Lambda}\right)+\frac{\dot{Z}'_\Lambda}{Z'_\Lambda}\right]d\xi, \tag{6.3}$$

so Eq. (6.2) becomes

$$i\hbar\frac{\partial G_\Lambda}{\partial t}=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda g_\Lambda\left[-\dot{\varphi}+\frac{\dot{X}P-X\dot{P}}{2}-i\hbar\frac{\dot{Z}_\Lambda}{w_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda}{Z'_\Lambda}\right)-\frac{i\hbar}{2}\frac{\dot{Z}'_\Lambda}{Z'_\Lambda}+\frac{i\hbar}{2}\frac{\dot{D}}{D}\right]d\xi. \tag{6.4}$$

From the definition of $\hat{P}$ in Eq. (3.4), it follows that

$$\frac{\hat{P}^2}{2m}G_\Lambda=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda g_\Lambda\left(\frac{P^2}{2m}+\frac{\hbar}{2m(l+\Lambda)}+\frac{iP\Delta}{2}-\frac{\Delta^2}{2m}\right)d\xi. \tag{6.5}$$

Again, the Δ's are now eliminated by using Eq. (4.7) and integrating by parts, giving

$$\frac{\hat{P}^2}{2m}G_\Lambda=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda g_\Lambda\left\{\frac{P^2}{2m}-i\hbar\frac{P}{m}\frac{1}{w_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda}{Z'_\Lambda}\right)-\frac{i\hbar}{2}\frac{P'}{mZ'_\Lambda}-\frac{\hbar^2}{2m}\frac{1}{w_\Lambda}\frac{\partial}{\partial\xi}\left[\frac{1}{Z'_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda}{Z'_\Lambda}\right)\right]\right\}d\xi. \tag{6.6}$$

Following similar steps, it is shown in Appendix A that

$$\begin{aligned}V(\hat{X},t)G_\Lambda=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda g_\Lambda\Bigg\{&V_0-\hbar\Lambda\left[V_1\frac{1}{w_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda}{Z'_\Lambda}\right)+\frac{V_2}{2}\frac{X'}{Z'_\Lambda}\right]\\&+(\hbar\Lambda)^2\frac{Z'_\Lambda}{w_\Lambda}\left[\frac{V_2}{2}\hat{\partial}^2_{Z_\Lambda}+\frac{V_3}{6}(2\hat{\beta}\hat{\partial}_{Z_\Lambda}+\hat{\partial}_{Z_\Lambda}\hat{\beta})+\frac{V_4}{8}\hat{\beta}^2\right]\frac{w_\Lambda}{Z'_\Lambda}\\&+\sum_{n=3}^{\infty}(\hbar\Lambda)^n\sum_{j=0}^{n}V_{2n-j}\sum_{k=0}^{\binom{n}{j}}c_k^{n,j}\frac{Z'_\Lambda}{w_\Lambda}\Pi_k^{n,j}(\hat{\partial}_{Z_\Lambda},\hat{\beta})\frac{w_\Lambda}{Z'_\Lambda}\Bigg\}d\xi,\end{aligned} \tag{6.7}$$

where $V_n=\partial^nV/\partial x^n$ evaluated at $X(\xi,t)$, $\hat{\partial}_{Z_\Lambda}:=Z'^{-1}_\Lambda\partial/\partial\xi$, $\hat{\beta}:=X'/Z'_\Lambda$, $\Pi_k^{n,j}(\hat{\partial}_{Z_\Lambda},\hat{\beta})$ corresponds to the kth permutation of the product of j factors of $\hat{\partial}_{Z_\Lambda}$ and $n-j$ factors of $\hat{\beta}$, and $c_k^{n,j}$ is a numerical coefficient.

By substituting Eqs. (6.4), (6.6), and (6.7) into Eq. (3.6), the Schrödinger equation is found to take the form

$$\int w_\Lambda(\xi) g_\Lambda(x,p;\xi)\Bigg\{\left(\dot\varphi-\frac{\dot X P - X\dot P}{2}+\frac{P^2}{2m}+V_0\right)$$
$$-i\hbar\left[\frac{1}{w_\Lambda}\left(\frac{P}{m}-i\Lambda V_1-\dot Z_\Lambda\right)\frac{\partial}{\partial\xi}\left(\frac{w_\Lambda}{Z'_\Lambda}\right)+\frac{1}{2Z'_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{P}{m}-i\Lambda V_1-\dot Z_\Lambda\right)+\frac{1}{2}\frac{\dot D}{D}\right]$$
$$+(\hbar\Lambda)^2\frac{Z'_\Lambda}{w_\Lambda}\left[\frac{V_2-(m\Lambda^2)^{-1}}{2}\hat\partial^2_{Z_\Lambda}+\frac{V_3}{6}(2\hat\beta\hat\partial_{Z_\Lambda}+\hat\partial_{Z_\Lambda}\hat\beta)+\frac{V_4}{8}\hat\beta^2\right]\frac{w_\Lambda}{Z'_\Lambda}$$
$$+\sum_{n=3}^{\infty}(\hbar\Lambda)^n\sum_{j=0}^{n}V_{2n-j}\sum_{k=0}^{\binom{n}{j}}c_k^{n,j}\frac{Z'_\Lambda}{w_\Lambda}\Pi_k^{n,j}(\hat\partial_{Z_\Lambda},\hat\beta)\frac{w_\Lambda}{Z'_\Lambda}\Bigg\}d\xi=0. \tag{6.8}$$

The integral on the left-hand side of Eq. (6.8) must vanish for all x and p, and for any l. Notice that the only dependence on these parameters comes from g_Λ, and that, like Eqs. (4.3) and (4.9), Eq. (6.8) is just a weighted sum of the Gaussians. Equation (6.8) can evidently be satisfied by requiring the expression inside the braces to vanish. This is now realized by introducing three constraints:

$$-\dot\varphi+\frac{\dot X P-X\dot P}{2}=\frac{P^2}{2m}+V(X,t), \tag{6.9a}$$

$$\dot Z_\Lambda=\dot X+i\Lambda\dot P=\frac{P}{m}-i\Lambda\frac{\partial V}{\partial x}(X,t), \tag{6.9b}$$

$$\frac{\dot D}{D}=-i\hbar\Lambda^2\sqrt{\frac{Z'_\Lambda}{D}}\left\{[V_2-(m\Lambda^2)^{-1}]\hat\partial^2_{Z_\Lambda}+\frac{V_3}{3}(2\hat\beta\hat\partial_{Z_\Lambda}+\hat\partial_{Z_\Lambda}\hat\beta)+\frac{V_4}{4}\hat\beta^2\right\}\sqrt{\frac{D}{Z'_\Lambda}}$$
$$-2i\sum_{n=3}^{\infty}\hbar^{n-1}\Lambda^n\sum_{j=0}^{n}V_{2n-j}\sum_{k=0}^{\binom{n}{j}}c_k^{n,j}\sqrt{\frac{Z'_\Lambda}{D}}\Pi_k^{n,j}(\hat\partial_{Z_\Lambda},\hat\beta)\sqrt{\frac{D}{Z'_\Lambda}}, \tag{6.9c}$$

where Eq. (4.10) was used in Eq. (6.9c).

Notice that Eq. (6.9b) was chosen such that its real and imaginary parts correspond to the classical laws of motion [see Eqs. (1.2)], so $X(\xi,t)$ and $P(\xi,t)$ can be identified as the position and momentum of the classical trajectory labeled by ξ. Also, Eq. (6.9a) was chosen to be compatible with Eq. (4.6), since these two equations can be shown to lead to the same expression for $\partial^2\varphi/\partial\xi\,\partial t$. Further, if we write φ as

$$\varphi(\xi,t)=S(\xi,t)-\frac{X(\xi,t)P(\xi,t)}{2}, \tag{6.10}$$

then Eq. (6.9a) becomes

$$\dot S(\xi,t)=\frac{P^2(\xi,t)}{2m}-V[X(\xi,t),t], \tag{6.11}$$

where Eq. (1.2a) was used to eliminate $\dot X$. The right-hand side of Eq. (6.11) is equal to the Lagrangian for the trajectory, so S can be identified as the classical action. Upon substituting Eq. (6.10) into Eq. (4.6), it follows that S satisfies

$$\frac{\partial S}{\partial\xi}(\xi,t)=P(\xi,t)\frac{\partial X}{\partial\xi}(\xi,t), \tag{6.12}$$

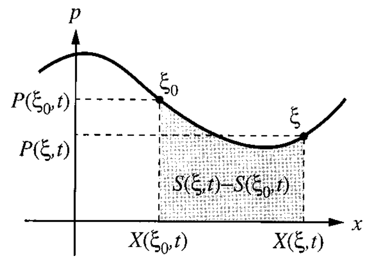

FIG. 3. $S(\xi,t)$ corresponds to an area under the phase space curve.

so S is also related to the area under the phase space curve, as shown in Fig. 3.

VII. SEMICLASSICAL ESTIMATE

Equation (6.9c) corresponds to the transport equation, which gives the time rate of change of the weighting D along a trajectory. While Eqs. (6.9a) and (6.9b) do not contain $\hbar$, Eq. (6.9c) has the form of a series in positive powers of $\hbar$. Unless V is a polynomial in x, this series is infinite. In conventional semiclassical methods, $\hbar$ is effectively regarded as a variable and $\hbar\to 0$ then underlies the asymptotic analysis. In this limit, Eq. (6.9c) states that D is conserved for each trajectory (i.e., $\dot{D}=0$), while Eqs. (6.9a) and (6.9b) remain unchanged. The form of the resulting semiclassical estimate follows from Eqs. (4.3) and (4.10):

$$G_\Lambda(x,p)=\frac{1}{\sqrt{2\pi\hbar\Lambda}}\int\sqrt{D(\xi,t_0)Z'_\Lambda(\xi,t)}\,g_\Lambda(x,p;\xi,t)\,d\xi, \tag{7.1}$$

where D has been frozen at the initial time t_0. The time interval over which this expression is valid can be estimated by using Eq. (6.9c): after a time given by the inverse of the absolute value of the right-hand side of Eq. (6.9c), the local estimate of $G(x,p)$ around the phase space point $[X(\xi,t),P(\xi,t)]$ is no longer expected to be accurate. In fact, Eq. (6.9c) gives a simple measure of the local deterioration of $G_\Lambda(x,p)$. It turns out that a suitable choice of Λ can help extend the validity of this estimate, and this is considered in Sec. VIII.

This approach allows a direct asymptotic estimate for any representation of the wave function. For example, by substituting Eq. (7.1) into Eq. (2.2a), and using Eqs. (4.2), (4.4), and (6.10), the associated estimate for $\Psi(x,t)$ is found to be

$$\Psi_\Lambda(x,t)=\frac{1}{\sqrt{2\pi\hbar\Lambda}}\int\sqrt{D(\xi,t_0)[X'(\xi,t)+i\Lambda P'(\xi,t)]}\times\exp\left\{-\frac{[x-X(\xi,t)]^2}{2\hbar\Lambda}\right\}\exp\left\{i\,\frac{S(\xi,t)+[x-X(\xi,t)]P(\xi,t)}{\hbar}\right\}d\xi. \tag{7.2}$$

Notice that the superfluous parameter l is entirely absent from Eq. (7.2). Also, provided condition (4.14) is satisfied, $\Psi_\Lambda(x,t)$ depends weakly on Λ. Alternatively, if Eq. (2.2b) is used in place of Eq. (2.2a) here, it is found that

$$\tilde{\Psi}_\Lambda(p,t)=\frac{1}{\sqrt{2\pi\hbar}}\int\sqrt{D(\xi,t_0)[X'(\xi,t)+i\Lambda P'(\xi,t)]}\times\exp\left\{-\Lambda\,\frac{[p-P(\xi,t)]^2}{2\hbar}\right\}\exp\left\{i\,\frac{S(\xi,t)-pX(\xi,t)}{\hbar}\right\}d\xi. \tag{7.3}$$

Again, this estimate does not depend on l, and has a weak dependence on Λ provided that condition (4.14) is satisfied. Further, $\tilde{\Psi}_\Lambda(p,t)$ is precisely the Fourier transform of $\Psi_\Lambda(x,t)$. As a result, it is easily shown that the mean square error is the same for these two estimates.

An important consideration for $\Psi_\Lambda(x,t)$ is that its norm is preserved in time. It is shown in Appendix B that $\int|\Psi_\Lambda(x,t)|^2\,dx \approx \int|D(\xi,t_0)|d\xi$. The weighting is therefore chosen to be normalized by

$$\int |D(\xi,t_0)|d\xi = 1. \tag{7.4}$$

VIII. CHOOSING Λ

It was mentioned in Sec. V that, for phase space curves that are multivalued in both x or p, a finite and nonzero value must be used for Λ. It was also anticipated in Sec. VII that a suitable choice for Λ may extend the validity of the estimate. To see this, consider rewriting Eq. (6.9c) in the form

$$\frac{\dot{D}}{D} = -i\hbar\left\{[\Lambda^2 V_2 - m^{-1}]\frac{1}{\sqrt{DZ'_\Lambda}}\frac{\partial}{\partial\xi}\left(\frac{1}{Z'_\Lambda}\frac{\partial}{\partial\xi}\sqrt{\frac{D}{Z'_\Lambda}}\right)\right.$$
$$\left.+\Lambda^2 V_3\left[\frac{X'}{Z'_\Lambda}\frac{1}{\sqrt{DZ'_\Lambda}}\frac{\partial}{\partial\xi}\sqrt{\frac{D}{Z'_\Lambda}} + \frac{1}{3Z'_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{X'}{Z'_\Lambda}\right)\right] + \frac{\Lambda^2 V_4}{4}\left(\frac{X'}{Z'_\Lambda}\right)^2\right\} + O(\hbar^2), \tag{8.1}$$

where we have concentrated on the terms that are first order in $\hbar$. The expected time of validity of the estimate is simply coupled to the modulus of the expression on the right-hand side of Eq. (8.1). Notice that the first term inside the braces in Eq. (8.1) has a factor that precisely corresponds to $\delta w_\Lambda / w_\Lambda$ [see Eqs. (4.10) and (4.13)], which was required to be small in order to achieve insensitivity to Λ. When this factor is expanded in the form shown in condition (5.1), one can see that the first term inside the braces in Eq. (8.1) is $O(\Lambda^{-4})$ for $\Lambda \to 0$ at points where $X'=0$, and it is $O(\Lambda^4)$ for $\Lambda \to \infty$ at points where $P'=0$. The rest of the expression in Eq. (8.1) is $O(\Lambda^0)$ and $O(\Lambda^3)$, respectively. Therefore, for both extreme cases, the first term dominates the expression inside the braces in Eq. (8.1). The insensitivity condition is then useful for enhancing the validity of the estimate. [In fact, the higher-order terms in Eq. (6.9c) are also kept from diverging by choosing Λ to make the scaled phase space curve satisfy conditions analogous to conditions (5.4) over all derivatives.]

So far, Λ has been treated as a constant. It is important, however, to see how the equations change when Λ is a function of time. [It is also natural to consider $\Lambda(\xi,t)$, but this is beyond the scope of this work.] In fact, by using Eq. (4.12) it follows that the only change is the appearance of an extra term of the form $i\hbar^2\,\delta w_\Lambda\,\dot{\Lambda}/(2w_\Lambda)$ inside the brackets in Eq. (6.4), and Eq. (8.1) then becomes

$$\frac{\dot{D}}{D} = -i\hbar\left\{[\Lambda^2 V_2 - m^{-1} - i\dot{\Lambda}]\frac{1}{\sqrt{DZ'_\Lambda}}\frac{\partial}{\partial\xi}\left(\frac{1}{Z'_\Lambda}\frac{\partial}{\partial\xi}\sqrt{\frac{D}{Z'_\Lambda}}\right)\right.$$
$$\left.+\Lambda^2 V_3\left[\frac{X'}{Z'_\Lambda}\frac{1}{\sqrt{DZ'_\Lambda}}\frac{\partial}{\partial\xi}\sqrt{\frac{D}{Z'_\Lambda}} + \frac{1}{3Z'_\Lambda}\frac{\partial}{\partial\xi}\left(\frac{X'}{Z'_\Lambda}\right)\right] + \frac{\Lambda^2 V_4}{4}\left(\frac{X'}{Z'_\Lambda}\right)^2\right\} + O(\hbar^2). \tag{8.2}$$

Since Eqs. (6.9a) and (6.9b) remain unchanged, the form of the estimates given in Eqs. (7.1)–(7.3) is preserved and they depend on Λ evaluated at the final time alone. That is, the history of Λ between the initial and final times turns out to be of no relevance for the estimates. The local deterioration of $\Psi_\Lambda(x,t)$ is now measured by Eq. (8.2). It follows that, for the purposes of arriving at a validity measure, the value of Λ should be chosen at all times by ensuring insensitivity for just the current form of the phase space curve.

Finally, also notice that Λ can be taken to be complex provided its real part is positive [so the Gaussian in Eq. (4.1) remains localized]. The imaginary part of Λ causes a shear in the footprint. Let $\Lambda^{-1}=\lambda^{-1}+i\nu^{-1}$, where both λ and ν are real, and $\lambda>0$. Equation (7.2) can now be rewritten as

$$\Psi_\Lambda(x,t)=\frac{1}{\sqrt{2\pi\hbar\lambda}}\exp\left(-i\frac{x^2}{2\hbar\nu}\right)\int\sqrt{D(\xi,t_0)[X'(\xi,t)+i\lambda\mathcal{P}'(\xi,t)]}$$
$$\times\exp\left\{-\frac{[x-X(\xi,t)]^2}{2\hbar\lambda}\right\}\exp\left\{i\frac{\mathcal{S}(\xi,t)+[x-X(\xi,t)]\mathcal{P}(\xi,t)}{\hbar}\right\}d\xi, \tag{8.3}$$

where

$$\mathcal{P}(\xi,t):=P(\xi,t)+\nu^{-1}X(\xi,t), \tag{8.4a}$$

$$\mathcal{S}(\xi,t):=S(\xi,t)+\nu^{-1}\frac{X^2(\xi,t)}{2}. \tag{8.4b}$$

Other than the chirp factor outside the integral, Eq. (8.3) has exactly the same form as Eq. (7.2), with Λ replaced by λ. Notice from Eq. (8.4a) that the parametric plot of $[X(\xi,t),\mathcal{P}(\xi,t)]$ corresponds to a sheared version of the phase space curve, where the degree of shearing is given by ν^{-1}. Also, it is easy to show from Eqs. (6.12) and (8.4b) that

$$\frac{\partial\mathcal{S}}{\partial\xi}(\xi,t)=\mathcal{P}(\xi,t)\frac{\partial X}{\partial\xi}(\xi,t), \tag{8.5}$$

so $\mathcal{S}$ corresponds to the area under the sheared phase space curve.

The condition for insensitivity on Λ can now be stated as follows: the real and imaginary parts of Λ must be chosen such that the scaled and sheared phase space curve described by $[\lambda^{-1/2}X(\xi,t),\lambda^{1/2}\mathcal{P}(\xi,t)]$ must have, for all ξ, a curvature much smaller than $\hbar^{-1/2}$, a rate of change of curvature much smaller than $\hbar^{-1}$, and first- and second-order rates of relative change in weighting that are negligible within any segment of the curve of length $\hbar^{1/2}$. Complex Λ therefore gives an extra degree of freedom that can be used to satisfy the validity condition for the semiclassical estimate given here.

IX. CONNECTION WITH OTHER SEMICLASSICAL ESTIMATES

As mentioned in Sec. V, when $X'(\xi,t)\neq 0$ for all ξ, condition (5.1) is satisfied for $\Lambda\to 0$. When this limit is taken, the spread in p of the footprint shown in Fig. 1(a) grows as $\Lambda^{-1/2}$, and the Gaussian in Eq. (7.2) together with the $\Lambda^{-1/2}$ factor approach a delta function. Therefore, the integral, evaluated at $x=X(\xi_0,t)$, gives

$$\Psi_{\Lambda\to 0}[X(\xi_0,t),t]=\sqrt{D(\xi_0,t_0)}\left|\frac{\partial X}{\partial\xi}(\xi_0,t)\right|^{-1/2}\exp\left[i\frac{S(\xi_0,t)}{\hbar}-i\frac{\pi}{2}M\right]. \tag{9.1}$$

Here, M is an integer known as the Maslov index, which accounts for the phase accumulated by the $\sqrt{Z'_\Lambda}$ factor in the integrand of Eq. (7.2) during propagation. [Notice that, since $\Lambda\to 0$, the phase of Z'_Λ is an integer multiple of π.] Equation (9.1) is precisely the conventional estimate mentioned in Sec. I, which is valid when the phase space curve is single valued in x. (Notice that the phase is now explicit.)

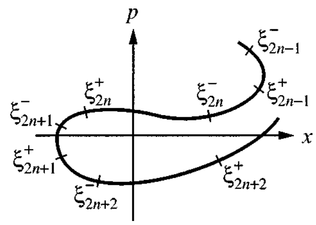

FIG. 4. Segmentation of the phase space curve for Maslov's method, where $X'(\xi,t)\neq 0$ for all $\xi\in[\xi_{2n}^-,\xi_{2n}^+]$, $P'(\xi,t)\neq 0$ for all $\xi\in[\xi_{2n+1}^-,\xi_{2n+1}^+]$, and $X'(\xi,t)P'(\xi,t)\neq 0$ for all $\xi\in[\xi_j^+,\xi_{j+1}^-]$.

Equivalently, when $P'(\xi,t)\neq 0$ for all ξ, condition (5.1) is satisfied as $\Lambda\to\infty$, which causes the footprint's spread in x to diverge as $\Lambda^{1/2}$. This limit leads to a delta function inside the integral in Eq. (7.3), which then gives

$$\tilde{\Psi}_{\Lambda\to\infty}[P(\xi_0,t),t]=\sqrt{D(\xi_0,t_0)}\left|\frac{\partial P}{\partial \xi}(\xi_0,t)\right|^{-1/2}\exp\left[i\,\frac{S(\xi_0,t)-X(\xi_0,t)P(\xi_0,t)}{\hbar}+i\,\frac{\pi}{4}-i\,\frac{\pi}{2}\tilde{M}\right],\tag{9.2}$$

where $\tilde{M}$ is the Maslov index for the momentum representation. Equation (9.2) corresponds to the estimate of the momentum distribution also mentioned in Sec. I, and is valid for phase space curves that are single valued in p.

The estimate given in Eq. (7.2) is significantly more general and robust than the conventional ones given in Eqs. (9.1) and (9.2). This follows from the fact that Λ can be chosen to reduce the deterioration of the estimate, as measured by $\dot{D}$. In particular, when the real part of Λ is finite and greater than zero, $\dot{D}$ remains finite since Z'_Λ [whose inverse appears repeatedly in Eq. (8.1)] never vanishes. For the two conventional estimates mentioned above, the corresponding measures of deterioration follow from the limits $\Lambda\to 0$ and $\Lambda\to\infty$ of Eq. (8.1), respectively. It is easily seen than that the time derivative of the weighting function then diverges at the caustics of the corresponding representation.

Another conventional semiclassical estimate that follows directly as a special case of Eq. (7.1) is Maslov's canonical operator method,[3,4] which is used at times when the phase space curve is multivalued in both x and p. [In such a case, neither Eq. (9.1) nor Eq. (9.2) can be used on their own.] To obtain this estimate, first define a series of switching functions $e_j(\xi)$, such that (i) $e_j(\xi)=1$ for ξ inside the interval $[\xi_j^-,\xi_j^+]$, which corresponds to a segment of the phase space curve that is single valued in either x (say, when j is even) or p (say, when j is odd), as shown in Fig. 4; (ii) $e_j(\xi)$ switches smoothly from 0 to 1 inside the interval (ξ_{j-1}^+,ξ_j^-) and from 1 to 0 inside the interval (ξ_j^+,ξ_{j+1}^-), where the phase space curve must be single valued in both x and p for both these intervals; (iii) $e_j(\xi)=0$ for all other ξ; and (iv) $\Sigma_j e_j(\xi)\equiv 1$ for all ξ. By inserting $\Sigma_j e_j(\xi)$ into Eq. (7.1), this expression can then be written as

$$G_\Lambda(x,p)=\sum_j \frac{1}{\sqrt{2\pi\hbar\Lambda}}\int \sqrt{D_j(\xi)Z'_\Lambda(\xi,t)}\,g_\Lambda(x,p;\xi,t)\,d\xi,\tag{9.3}$$

where $D_j(\xi):=e_j^2(\xi)D(\xi,t_0)$, and Λ is such that Eq. (7.1) satisfies the insensitivity condition (4.14).

Each of the terms in the sum in Eq. (9.3) is now effectively an independent estimate of the form given in Eq. (7.1). In fact, the value of Λ can now be varied independently for each term, provided the associated $D_j(\xi)$ satisfies condition (4.14). It follows from conditions (5.4c) and (5.4d) then, that the switching functions must themselves satisfy

$$\left|\frac{\check{e}_j'}{\check{e}_j}\right| \ll \frac{1}{\sqrt{\hbar}}, \tag{9.4a}$$

$$\left|\frac{\check{e}_j''}{\check{e}_j}\right| \ll \frac{1}{\hbar}, \quad \text{for } \check{\xi} \in (\check{\xi}_{j-1}^{+}, \check{\xi}_{j+1}^{-}), \tag{9.4b}$$

where $\check{e}_j(\check{\xi}) := e_j[R(\check{\xi})]$ and $R(\check{\xi}_j^{\pm}) = \xi_j^{\pm}$. The switching intervals should therefore be made as wide as possible given the restriction that they are necessarily constrained between two adjacent caustics—$X'(\xi,t)=0$ at one and $P'(\xi,t)=0$ at the other. These requirements impose added constraints on the validity of Maslov's method. This has been studied quantitatively (within the optics context) elsewhere,[5] where it is established that these are, in fact, the dominant limiting factors for this approach. [Of course, conditions (9.4) are necessarily violated within a neighborhood of ξ_{j-1}^{+} and ξ_{j+1}^{-}, where $e_j(\xi)$ vanishes, but since the weighting is small there, any associated errors are insignificant.]

To obtain an estimate for $\Psi(x,t)$, the even terms in Eq. (9.3) are substituted into Eq. (2.2a), and the odd ones are inserted into Eq. (2.2b) and inverse Fourier transformed according to Eq. (1.3b). Since the even terms correspond to segments of the phase space curve that are single valued in x, the associated integrals can be estimated, as in Eq. (9.1), by letting $\Lambda \to 0$. The resulting expressions (which are given in parametric form) are expressed explicitly as functions of x by evaluating them at $\xi = \Xi_j(x,t) \in (\xi_{j-1}^{+}, \xi_{j+1}^{-})$, which is the solution of $X[\Xi_j(x,t),t] = x$ for $x \in [X(\xi_{j-1}^{+},t), X(\xi_{j+1}^{-},t)]$. Similarly, the integral over ξ for the odd terms can be estimated by letting $\Lambda \to \infty$ as in Eq. (9.2), since they correspond to segments that are single valued in p. The result takes the form

$$\begin{aligned}\Psi_{\mathrm{M}}(x,t) = &\sum_n e_{2n}[\Xi_{2n}(x,t)]\sqrt{D[\Xi_{2n}(x,t),t_0]}\left|\frac{\partial \Xi_{2n}}{\partial x}(x,t)\right|^{1/2} \exp\left\{ i\,\frac{S[\Xi_{2n}(x,t),t]}{\hbar} - i\,\frac{\pi}{2} M_{2n}\right\}\\ &+ \frac{1}{\sqrt{2\pi\hbar}}\sum_n \int_{\xi_{2n}^{+}}^{\xi_{2n+2}^{-}} e_{2n+1}(\xi)\sqrt{D(\xi,t_0)}\left|\frac{\partial P}{\partial \xi}(\xi,t)\right|^{1/2}\\ &\times \exp\left\{ i\,\frac{S(\xi,t)+[x-X(\xi,t)]P(\xi,t)}{\hbar} + i\,\frac{\pi}{4} - i\,\frac{\pi}{2}\tilde{M}_{2n+1}\right\} d\xi,\end{aligned} \tag{9.5}$$

where M_{2n} and $\tilde{M}_{2n+1}$ are the Maslov indices for the corresponding segments. This is the standard result from Maslov's method and it is now seen to follow as a special case of Eq. (7.2). It is important to appreciate, however, that the dominant limitations on the validity of Ψ_{M} (i.e., the switching errors) are avoided altogether by the new method.

The form of Eq. (7.2) also suggests a clear link to propagation schemes based on summing Gaussian wave packets.[6] The basic idea is that the initial wave function is decomposed into a sum of Gaussians, and each of these is propagated independently. Similar ideas were considered in a different context by Daubechies,[7] who points out the clear link with the coherent states[8] of quantum electrodynamics. In these works, however, phase space acts only as a mathematical domain for field decomposition at a fixed time. This stands in contrast to the method developed here, where the field is explicitly propagated in phase space itself. Further, as shown in Sec. V, for Λ within a certain interval, the sum of Gaussians in Eq. (7.2) is effectively independent of Λ. It follows that the evolution of the individual Gaussians is irrelevant: Λ can simply be chosen such that it satisfies condition (4.14) at the final time alone.

This property of a sum of Gaussians was first noted by Heller,[9] who realized that a sum of propagated Gaussian beam elements may lead to a cancellation of the dominant errors within each element when their widths are frozen (i.e., there is then no beam waist). This is precisely what we have now derived directly from Eq. (1.1) and, as a result, arrived at explicit validity conditions. Further, we have shown in Sec. VIII that the beamwidths must actually be varied in time to determine the validity of the final results.

Herman and Kluk[10] justified Heller's heuristic approach quite differently and, just as in the Gabor method[11] that was further developed by Bastiaans,[12] they end up with a sum over all of phase space—not just over a curve. Kay recently introduced both an ansatz to generalize these integral approaches[13] and an empirical means to deal with the onset of chaos.[14] Walton and Manolopoulos[15] have recently combined some of the strengths of these ideas. However, the foundation of all this work is the conventional asymptotic propagator as developed by Van Vleck.[16] This propagator takes the form of Eq. (9.1), whose glaring limitations motivated this work. That is, unlike earlier semiclassical methods, the results developed here are derived directly from the Schrödinger equation and not from the fundamentally limited result given in Eq. (9.1). Nevertheless, these important contributions are clearly relevant to the method developed here.

X. SUMMARY OF THE METHOD

Our semiclassical method for estimating $\Psi(x,t)$ given the initial condition $\Psi(x,t_0)$ can be summarized in the following three steps.

(I) Find a phase space curve and an associated weighting—i.e., $X(\xi,t_0)$, $P(\xi,t_0)$, and $D(\xi,t_0)$, so that the right-hand side of Eq. (7.2) matches this prescribed initial condition. [Recall that $S(\xi,t_0)$ can be constructed from Eq. (6.12) evaluated at t_0.]

(II) Propagate the family of classical trajectories by using Eqs. (1.2), and construct $S(\xi,t)$ by integrating Eq. (6.11):

$$S(\xi,t)=S(\xi,t_0)+\int_{t_0}^{t}\left\{\frac{P^2(\xi,t')}{2m}-V[X(\xi,t'),t']\right\}dt'. \tag{10.1}$$

[Alternatively, $S(\xi,t)$ can be found, to within a constant, directly from X and P at t from the ξ integral of Eq. (6.12).]

(III) Find the estimate for $\Psi(x,t)$, namely, $\Psi_\Lambda(x,t)$, by using Eq. (7.2), where Λ is chosen to satisfy condition (4.14) at t.

For a significant set of wave functions, the solution to Step (I) is straightforward. For example, when the amplitude of $\Psi(x,t_0)$ is slowly varying, a prescription for determining $X(\xi,t_0)$, $P(\xi,t_0)$, and $D(\xi,t_0)$ is given by

$$X(\xi,t_0)=\xi, \tag{10.2a}$$

$$P(\xi,t_0)=\hbar\,\mathrm{Im}\left\{\frac{\Psi'(\xi,t_0)}{\Psi(\xi,t_0)}\right\}, \tag{10.2b}$$

$$D(\xi,t_0)=|\Psi(\xi,t_0)|^2. \tag{10.2c}$$

Alternatively, when the amplitude of $\tilde{\Psi}(p,t_0)$ varies slowly, one can use

$$X(\xi,t_0)=-\hbar\,\mathrm{Im}\left\{\frac{\tilde{\Psi}'(\xi,t_0)}{\tilde{\Psi}(\xi,t_0)}\right\}, \tag{10.3a}$$

$$P(\xi,t_0)=\xi, \tag{10.3b}$$

$$D(\xi,t_0)=|\tilde{\Psi}(\xi,t_0)|^2. \tag{10.3c}$$

Of course, $\Psi_\Lambda(x,t)$ may also accurately model wave functions for which the initial conditions do not fall into the categories considered above. A more general solution to the task outlined in Step (I) is clearly a natural next step.

Finally, we comment that the simplicity of Eq. (7.2) means that the generalization to higher dimensions is straightforward. The estimate for a wave function $\Psi(\mathbf{x},t)$, where $\mathbf{x}=(x_1,x_2,\ldots,x_n)$, is given in terms of an associated n-parameter family of trajectories by

$$\Psi_{\mathbb{L}^{-1}}(\mathbf{x},t)=(2\pi\hbar)^{-n/2}\int\left\{D(\boldsymbol{\xi},t_0)\,\frac{\partial[\mathbb{L}\cdot\mathbf{X}(\boldsymbol{\xi},t)+i\mathbf{P}(\boldsymbol{\xi},t)]}{\partial(\boldsymbol{\xi})}\right\}^{1/2} \times\exp\left\{-\frac{[\mathbf{x}-\mathbf{X}(\boldsymbol{\xi},t)]\cdot\mathbb{L}\cdot[\mathbf{x}-\mathbf{X}(\xi,t)]}{2\hbar}+i\,\frac{S(\boldsymbol{\xi},t)+[\mathbf{x}-\mathbf{X}(\boldsymbol{\xi},t)]\cdot\mathbf{P}(\boldsymbol{\xi},t)}{\hbar}\right\}d^n\xi, \tag{10.4}$$

where $\boldsymbol{\xi}=(\xi_1,\xi_2,\ldots,\xi_n)$, $\mathbb{L}$ is a real, symmetric $n\times n$ matrix with positive eigenvalues, and $\partial(\mathbf{Y})/\partial(\boldsymbol{\xi})$ is a Jacobian determinant. This estimate is insensitive to changes in $\mathbb{L}$ when any patch of radius $\hbar^{1/2}$ of the n-dimensional, scaled Lagrange manifold described by $[\mathbb{L}^{1/2}\cdot\mathbf{X}(\boldsymbol{\xi},t),\ \mathbb{L}^{-1/2}\cdot\mathbf{P}(\boldsymbol{\xi},t)]$ is sufficiently flat, and the relative variations of $D(\boldsymbol{\xi},t_0)$ over the patch are insignificant.

APPENDIX A: THE APPLICATION OF $V(\hat{X},t)$ TO G_Λ

From Eq. (4.3), it follows that

$$\begin{aligned}V(\hat{X},t)G_\Lambda&=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda(\xi,t)V(\hat{X},t)g_\Lambda(x,p;\xi,t)d\xi\\&=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda\sum_{n=0}^{\infty}\frac{V_n}{n!}(\hat{X}-X)^n g_\Lambda\,d\xi,\end{aligned}\tag{A1}$$

where $V_n=\partial^n V/\partial x^n$ evaluated at $X(\xi,t)$. Our goal in this appendix is to rewrite Eq. (A1) in a form where all of the dependence on x and p (as well as on l) is contained within g_Λ, and the remaining weighting expression takes the form of a series in powers of $\hbar$ that is consistent with an asymptotic treatment.

From Eqs. (3.1), (4.4), and (4.7), it follows that

$$(\hat{X}-X)g_\Lambda=\Lambda\,\Delta g_\Lambda=\hbar\Lambda Z_\Lambda'^{-1}\frac{\partial g_\Lambda}{\partial\xi}=\hbar\Lambda\,\hat{\partial}_{Z_\Lambda}g_\Lambda\,, \tag{A2}$$

where $\Delta:=(z_{-l}-Z_{-l})/(l+\Lambda)$ and $\hat{\partial}_{Z_\Lambda}:=Z_\Lambda'^{-1}\partial/\partial\xi$. Equation (A2) is now used to eliminate the powers of $(\hat{X}-X)$ progressively from Eq. (A1):

$$\sum_{n=0}^{\infty}\frac{V_n}{n!}(\hat{X}-X)^n g_\Lambda=V_0 g_\Lambda+\hbar\Lambda V_1\,\hat{\partial}_{Z_\Lambda}g_\Lambda+\hbar\Lambda\sum_{n=2}^{\infty}\frac{V_n}{n!}(\hat{X}-X)^{n-1}\,\hat{\partial}_{Z_\Lambda}g_\Lambda\,. \tag{A3}$$

The commutator between $(\hat{X}-X)$ and $\hat{\partial}_{Z_\Lambda}$ follows from

$$(\hat{X}-X)\hat{\partial}_{Z_\Lambda}=\frac{X'}{Z_\Lambda'}+\hat{\partial}_{Z_\Lambda}(\hat{X}-X)=\hat{\beta}+\hat{\partial}_{Z_\Lambda}(\hat{X}-X), \tag{A4}$$

where $\hat{\beta}:=X'/Z_\Lambda'$. Since $\hat{\beta}$ commutes with $(\hat{X}-X)$ (but not with $\hat{\partial}_{Z_\Lambda}$), it follows that

$$(\hat{X}-X)^{\mu}\hat{\partial}_{Z_\Lambda}=\mu\hat{\beta}(\hat{X}-X)^{\mu-1}+\hat{\partial}_{Z_\Lambda}(\hat{X}-X)^{\mu}. \tag{A5}$$

Equations (A2) and (A5) are used repeatedly in Eq. (A3) to find

$$\begin{aligned}\sum_{n=0}^{\infty}\frac{V_n}{n!}(\hat{X}-X)^n g_\Lambda &= V_0 g_\Lambda+\hbar\Lambda V_1\,\hat{\partial}_{Z_\Lambda}g_\Lambda \\ &\quad+\hbar\Lambda\sum_{n=2}^{\infty}\frac{V_n}{n!}[(n-1)\hat{\beta}(\hat{X}-X)^{n-2}+\hat{\partial}_{Z_\Lambda}(\hat{X}-X)^{n-1}]g_\Lambda \\ &= V_0 g_\Lambda+\hbar\Lambda V_1\,\hat{\partial}_{Z_\Lambda}g_\Lambda+\hbar\Lambda\frac{V_2}{2}\hat{\beta}g_\Lambda+(\hbar\Lambda)^2\frac{V_2}{2}\hat{\partial}^2_{Z_\Lambda}g_\Lambda \\ &\quad+(\hbar\Lambda)^2\sum_{n=3}^{\infty}\frac{V_n}{n!}[(n-1)\hat{\beta}(\hat{X}-X)^{n-3}+\hat{\partial}_{Z_\Lambda}(\hat{X}-X)^{n-2}]\hat{\partial}_{Z_\Lambda}g_\Lambda\,. \end{aligned}\tag{A6}$$

By repeating these steps, we ultimately obtain an expression of the form

$$\begin{aligned}\sum_{n=0}^{\infty}\frac{V_n}{n!}(\hat{X}-X)^n g_\Lambda &= V_0 g_\Lambda+\hbar\lambda\left[V_1\hat{\partial}_{Z_\Lambda}+\frac{V_2}{2}\hat{\beta}\right]g_\Lambda \\ &\quad+(\hbar\Lambda)^2\left[\frac{V_2}{2}\hat{\partial}^2_{Z_\Lambda}+\frac{V_3}{6}(2\hat{\beta}\hat{\partial}_{Z_\Lambda}+\hat{\partial}_{Z_\Lambda}\hat{\beta})+\frac{V_4}{8}\hat{\beta}^2\right]g_\Lambda \\ &\quad+\sum_{n=3}^{\infty}(\hbar\Lambda)^n\sum_{j=0}^{n}V_{2n-j}\sum_{k=0}^{\binom{n}{j}}a_k^{n,j}\Pi_k^{n,j}(\hat{\partial}_{Z_\Lambda},\hat{\beta})g_\Lambda\,, \end{aligned}\tag{A7}$$

where $\Pi_k^{n,j}(\hat{\partial}_{Z_\Lambda},\hat{\beta})$ corresponds to the kth permutation of the product of j factors of $\hat{\partial}_{Z_\Lambda}$ and $n-j$ factors of $\hat{\beta}$, and $a_k^{n,j}$ is a numerical coefficient. The terms for $n\leqslant 2$ have been written out explicitly here.

Remember that $\hat{\partial}_{Z_\Lambda}$ contains a ξ derivative that acts over the whole expression to its right. When Eq. (A7) is substituted into Eq. (A1), these derivatives can be removed from g_Λ by repeatedly integrating by parts in the form

$$\begin{aligned}\int BV_\mu\hat{\beta}^\eta\,\hat{\partial}_{Z_\Lambda}\hat{A}\,g_\Lambda\,d\xi &= \int\hat{\beta}^\eta\frac{B}{Z'_\Lambda}V_\mu\frac{\partial}{\partial\xi}(\hat{A}g_\Lambda)d\xi \\ &= -\int\frac{\partial}{\partial\xi}\left[\hat{\beta}^\eta\frac{B}{Z'_\Lambda}V_\mu\right]\hat{A}g_\Lambda\,d\xi \\ &= -\int\left[V_\mu\hat{\partial}_{Z_\Lambda}\left(\hat{\beta}^\eta\frac{B}{Z'_\Lambda}\right)+V_{\mu+1}\hat{\beta}^{\eta+1}\frac{B}{Z'_\Lambda}\right](Z'_\Lambda\hat{A}g_\Lambda)d\xi. \end{aligned}\tag{A8}$$

By substituting Eq. (A7) into Eq. (A1) and using Eq. (A8) repeatedly, the desired result is found to be

$$V(\hat{X},t)G_\Lambda=\frac{1}{\sqrt{2\pi\hbar}}\int w_\Lambda g_\Lambda\Bigg\{V_0-\hbar\Lambda\frac{Z'_\Lambda}{w_\Lambda}\left[V_1\hat{\partial}_{Z_\Lambda}+\frac{V_2}{2}\hat{\beta}\right]\frac{w_\Lambda}{Z'_\Lambda}$$
$$+(\hbar\Lambda)^2\frac{Z'_\Lambda}{w_\Lambda}\left[\frac{V_2}{2}\hat{\partial}^2_{Z_\Lambda}+\frac{V_3}{6}(2\hat{\beta}\hat{\partial}_{Z_\Lambda}+\hat{\partial}_{Z_\Lambda}\hat{\beta})+\frac{V_4}{8}\hat{\beta}^2\right]\frac{w_\Lambda}{Z'_\Lambda}$$
$$+\sum_{n=3}^{\infty}(\hbar\Lambda)^n\sum_{j=0}^{n}V_{2n-j}\sum_{k=0}^{\binom{n}{j}}c_k^{n,j}\frac{Z'_\Lambda}{w_\Lambda}\Pi_k^{n,j}(\hat{\partial}_{Z_\Lambda},\hat{\beta})\frac{w_\Lambda}{Z'_\Lambda}\Bigg\}d\xi. \tag{A9}$$

Notice that, as required, the term in braces is now a series in $\hbar$, with all terms given explicitly through $O(\hbar^2)$.

APPENDIX B: NORMALIZATION OF Ψ_Λ

By using Eq. (7.2), the integral of $|\Psi_\Lambda(x,t)|^2$ over all x can be carried out in closed form, leaving

$$\int |\Psi_\Lambda(x,t)|^2\,dx$$
$$=\frac{1}{2\sqrt{\pi\hbar\Lambda}}\int\int\sqrt{D(\xi,t_0)D^*(\tau,t_0)[X'(\xi,t)+i\Lambda P'(\xi,t)][X'(\tau,t)-i\Lambda P'(\tau,t)]}$$
$$\times\exp\left\{i\,\frac{S(\xi,t)-S(\tau,t)}{\hbar}+i\,\frac{[P(\xi,t)+P(\tau,t)][X(\xi,t)-X(\tau,t)]}{2\hbar}\right\}$$
$$\times\exp\left\{-\frac{[X(\xi,t)-X(\tau,t)]^2}{4\hbar\Lambda}-\Lambda\,\frac{[P(\xi,t)-P(\tau,t)]^2}{4\hbar}\right\}d\xi\,d\tau. \tag{B1}$$

Notice that the integrand of the right-hand side of Eq. (B1) is most significant when $\xi\approx\tau$. The integral over τ is now approximated by using the saddle point method: the exponent of Eq. (B1) is expanded in τ around ξ up to second order, and the local variation of the amplitude factor in the square root is neglected. This leads to

$$\int |\Psi_\Lambda(x,t)|^2\,dx\approx\frac{1}{2\sqrt{\pi\hbar\Lambda}}\int\int|D(\xi,t_0)Z'_\Lambda(\xi,t)|\exp\left[-\frac{|Z'_\Lambda(\xi,t)|^2(\tau-\xi)^2}{4\hbar\Lambda}\right]d\xi\,d\tau$$
$$=\int|D(\xi,t_0)|\,d\xi. \tag{B2}$$

ACKNOWLEDGMENTS

We are grateful to Dr. Misha Ivanov for bringing relevant publications to our attention and to the Australian Research Council for funding this project.

[1] M. C. Gutzwiller, ''Resource letter ICQM-1: The interplay between classical and quantum mechanics,'' Am. J. Phys. **66**, 304–324 (1998).
[2] E. T. Copson, *Asymptotic Expansions* (Cambridge University Press, Cambridge, 1965), pp. 91–94.
[3] V. P. Maslov and M. V. Fedoriuk, *Semiclassical Approximation in Quantum Mechanics* (Reidel, Boston, 1981).
[4] J. B. Delos, ''Semiclassical calculation of quantum mechanical wavefunctions,'' *Advances in Chemical Physics*, edited by I. Prigogine and S. A. Rice (Wiley Interscience, New York, 1986), Vol. 67, pp. 161–213.
[5] G. W. Forbes and M. A. Alonso, ''Asymptotic estimation of the optical wave propagator II: relative validity,'' J. Opt. Soc. Am. A **15**, 1341–1354 (1998).
[6] V. M. Babich and M. M. Popov, ''Gaussian summation method (review),'' Radiophys. Quantum Electron. **39**, 1063–1081 (1989).

[7] I. Daubechies, ''Time–frequency localization operators: a geometric phase space approach,'' IEEE Trans. Inf. Theory **34**, 605–612 (1988).
[8] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge University Press, Cambridge, UK, 1995), Chap. 11.
[9] E. J. Heller, ''Frozen Gaussians: a very simple semiclassical approximation,'' J. Chem. Phys. **75**, 2923–2931 (1981).
[10] M. F. Herman and E. Kluk, ''A semiclassical justification for the use of non-spreading wavepackets in dynamics calculations,'' Chem. Phys. **91**, 27–34 (1984).
[11] D. Gabor, ''Theory of communication,'' J. Inst. Electr. Eng. **93**, 429–457 (1946).
[12] M. J. Bastiaans, ''A sampling theorem for the complex spectrogram, and Gabor's expansion of a signal in Gaussian elementary signals,'' Opt. Eng. (Bellingham) **20**, 594–598 (1981).
[13] K. G. Kay, ''Integral expressions for the semiclassical time-dependent propagator,'' J. Chem. Phys. **100**, 4377–4392 (1994).
[14] K. G. Kay, ''Semiclassical propagation for multidimensional systems by an initial value method,'' J. Chem. Phys. **101**, 2250–2260 (1994).
[15] A. R. Walton and D. E. Manolopoulos, ''A new semiclassical initial value method for Frank–Condon spectra,'' Mol. Phys. **87**, 961–978 (1996).
[16] J. H. Van Vleck, ''The correspondence principle in the statistical interpretation of quantum mechanics,'' Proc. Natl. Acad. Sci. USA **14**, 178–188 (1928).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Bounded Bose fields in 1+1 dimensions commuting for space- and timelike distances

Klaus Baumann[a)]
Institut für Theoretische Physik, Universität Göttingen, Bunsenstraße 9, D-37073 Göttingen, Germany

Shortly after submission of this article, Klaus Baumann died at a much too young age. His scientific work was devoted to the foundations of relativistic quantum field theory and consists of masterly pieces of mathematical physics. He settled pertinent structural questions and helped to open our view for the subtle features of quantum field theory beyond perturbation theory.
Collegues and friends in Göttingen

(Received 29 August 1998; accepted for publication 29 October 1998)

We consider scalar Bose fields Φ in 1+1 dimensions which are bounded [i.e., $\Phi(F)$ is a bounded operator], commute for space- and timelike distances, and are dilation covariant with scaling dimension $d=1,3,5,\ldots$. We show that their truncated n-point-functions W_n^T are related to the truncated functions V_n^T of $\varphi(x)=\psi_{d/2}(x_+)\otimes\psi_{d/2}(x_-)$ via $W_n^T=c_nV_n^T$ with $c_n>0$. $\psi_{d/2}(x_\pm)$ is a free chiral real Fermi field of dimension $d/2$ depending on the light cone coordinates $x_\pm=t\pm x$. This comes close to the conjecture that under the above assumptions Φ is nothing but a weighted s-product of $\varphi=\psi_{d/2}\otimes\psi_{d/2}$. © *1999 American Institute of Physics.* [S0022-2488(99)00104-8]

INTRODUCTION

From ordinary quantum mechanics on $\mathbf{R}^d$ we are used to expecting that position and momentum operators are unbounded because of the Heisenberg commutation relations. The same is true for a relativistic quantum field theory in $d+1$-dimensional space–time fulfilling canonical commutation relations. But the Wightman axioms[1] for Bose fields do not ask explicitly for unbounded field operators and, furthermore, in two-dimensional space–time there exist examples of Bose fields, such that the smeared field operators $\Phi(F)$ are bounded operators (''bounded Bose fields'').[2] An especially simple class is constructed by tensor products of free chiral Fermi fields depending on the light cone coordinates $x_+=x_0+x_1$, respectively $x_-=x_0-x_1$, i.e., $\Phi(x)=\psi_+(x_+)\otimes\psi_-(x_-)$. Because of this construction they have the peculiar property to commute not only for spacelike but also for timelike distances. More complicated and therefore more interesting examples have been given by Rehren[3] based on tensor products of vertex operators. Starting with a scalar bounded Bose field Φ which commutes for space- and timelike distances and which is scale covariant, we try to prove the converse, namely we want to show that Φ looks like an s-product[4]—or a generalization thereof—of the above example given by Buchholz. In the case of chiral field theories we were able to show[5] that Bose fields are unbounded with the exception of c-number fields and that only free Fermi fields are bounded. The vertex operators considered by Rehren[3] fulfill more general commutation relations and he characterized which of them are bounded.

I. ASSUMPTIONS AND RESULTS

We want to do relativistic quantum field theory in the framework given by Wightman.[1] Our assumptions are as follows:

[a)]Deceased.

(A1) Φ is a scalar, neutral Bose field in $1+1$-dimensional space–time.

(A2) Φ commutes not only for spacelike, but also for timelike distances. Written in light cone coordinates $x_+ = t+x$ and $x_- = t-x$ we have

$$[\Phi(x),\Phi(y)]=0 \quad \text{if } (y_+-x_+)(y_--x_-)\neq 0. \tag{1.1}$$

(A3) Φ is scale covariant.

Remark: Lorentz invariance together with scale covariance imply the existence of unitary operators $U_{\lambda,\mu}$ such that

$$\Phi\begin{pmatrix}\lambda x_+\\ \mu x_-\end{pmatrix}=\frac{1}{(\lambda\mu)^{d/2}}U_{\lambda,\mu}\Phi\begin{pmatrix}x_+\\ x_-\end{pmatrix}U^\dagger_{\lambda,\mu} \quad \text{for all } \lambda,\mu>0, \tag{1.2}$$

and $U_{\lambda,\mu}\Omega=\Omega$ where Ω denotes the vacuum vector. Written with test functions this means

$$\Phi\begin{pmatrix}(f_+)_\lambda\\ (f_-)_\mu\end{pmatrix}=\frac{1}{(\lambda\mu)^{d/2-1}}U_{\lambda,\mu}\Phi\begin{pmatrix}f_+\\ f_-\end{pmatrix}U^\dagger_{\lambda,\mu} \quad \text{for all } \lambda,\mu>0, \tag{1.3}$$

if we define $(f)_\lambda(x):=f(x/\lambda)$ for $\lambda>0$. To fulfill space- and timelike commutativity, the allowed scaling dimensions d have to be integers from $\mathbf{N}_0$.

(A4) Φ is a bounded field, i.e., for every test function $F\in\mathcal{S}(\mathbf{R}^2)$ we have

$$\Phi(F) \text{ is a bounded operator with norm } \|\Phi(F)\|<\infty, \tag{1.4}$$

and because the operators $U_{\lambda,\mu}$ are unitary we get

$$\|\Phi(F_{\lambda,\mu})\|=\frac{1}{(\lambda\mu)^{d/2-1}}\|\Phi(F)\| \quad \text{for all } \lambda,\mu>0. \tag{1.5}$$

Remark: A nontrivial field which obviously fulfills the above assumptions **A** is given by $\varphi(x):=\psi(x_+)\otimes\psi(x_-)$ where ψ is a free neutral chiral Fermi field of (half-integer!) scaling dimension $d/2$. We shall discuss this example given by Buchholz in the next section.

Now we can formulate our main result:

Theorem: Let Φ fulfill the assumptions **A** with scaling dimension $d\in\mathbf{N}$, d odd. Then the truncated n-point functions $\mathcal{W}_n^T$ of the field Φ are proportional to the truncated n-point functions $\mathcal{V}_n^T$ of the just introduced field $\varphi=\psi\otimes\psi$, i.e., there exist positive numbers c_{2n} such that

$$\mathcal{W}_{2n}^T(F_1,\ldots,F_{2n})=c_{2n}\mathcal{V}_{2n}^T(F_1,\ldots,F_{2n}) \quad \text{and} \quad \mathcal{W}_{2n+1}^T\equiv 0. \tag{1.6}$$

Remark: If the sequence $(c_{2n}|n\in\mathbf{N})$ is of the form $c_{2n}=\Sigma_i\alpha_i^{2n}$ (a property which can be tested by an infinite set of inequalities such as $c_2^2\geqslant c_4$, and so on), we can write the field Φ as a weighted s-product of the field φ.

Remark: We can slightly generalize our theorem by relaxing Lorentz invariance and scale covariance in the following way:

$$\Phi\begin{pmatrix}\lambda x_+\\ \mu x_-\end{pmatrix}=\frac{1}{\lambda^{d_+/2}\mu^{d_-/2}}U_{\lambda,\mu}\Phi\begin{pmatrix}x_+\\ x_-\end{pmatrix}U^\dagger_{\lambda,\mu} \quad \text{for all } \lambda,\mu>0, \tag{1.7}$$

where d_+ and d_- are both half-integer numbers. If $d_+>d_-$, the field $\hat{\phi}:=\partial_-^{d_+-d_-}\phi$ fulfills the assumptions of our theorem. Therefore we have a candidate for all truncated functions and it is easy to show their uniqueness.

II. THE EXAMPLE GIVEN BY BUCHHOLZ

As remarked by Buchholz,[2] a simple method to construct a bounded Bose field in $1+1$-dimensional space–time is the following:

Take two bounded chiral Fermi fields in the light cone coordinates x_+, resp x_-. Their tensor product will define a bounded Bose field! But the class of bounded chiral Fermi fields coincides with free fields,[5] i.e., their anticommutator is a c-number. Therefore let ψ denote a neutral chiral Fermi field with scaling dimension $d_\psi = \frac{1}{2}$ fulfilling canonical anticommutation relations

$$\{\psi(u),\psi(v)\} = \delta(v-u)\mathbf{1} \quad \text{for all } u,v \in \mathbf{R}, \tag{2.1}$$

and bounded by

$$\|\psi(f)\| \leqslant \|f\|_2 \quad \text{for all } f \in \mathcal{S}(\mathbf{R}). \tag{2.2}$$

The field $\varphi(x) := \psi(x_+) \otimes \psi(x_-)$ defines a bounded scalar Bose field of scaling dimension $d=1$ and its commutation relations are

$$[\varphi(x),\, \varphi(y)] = \tfrac{1}{2}[\psi(x_+),\, \psi(y_+)] \otimes \{\psi(x_-),\psi(y_-)\} + \{\psi(x_+),\psi(y_+)\} \otimes \tfrac{1}{2}[\psi(x_-),\psi(y_-)] \tag{2.3}$$

$$= b(x_+,y_+)\,\delta(y_- - x_-) + b(x_-,y_-)\,\delta(y_+ - x_+) = 0$$

$$\text{if } (y_+ - x_+)(y_- - x_-) \neq 0 \tag{2.4}$$

i.e., φ commutes not only for spacelike but also for timelike distances.

As a pedagogical exercise let us calculate the double commutator.

$$\left[\left[\varphi\begin{pmatrix} x_+ \\ x_- \end{pmatrix},\varphi\begin{pmatrix} y_+ \\ y_- \end{pmatrix}\right]\varphi\begin{pmatrix} z_+ \\ z_- \end{pmatrix}\right] = \delta(y_- - x_-)\left\{\delta(z_+ - y_+)\,\varphi\begin{pmatrix} x_+ \\ z_- \end{pmatrix} - \delta(z_+ - x_+)\,\varphi\begin{pmatrix} y_+ \\ z_- \end{pmatrix}\right\}$$
$$+ \delta(y_+ - x_+)\left\{\delta(z_- - y_-)\,\varphi\begin{pmatrix} z_+ \\ x_- \end{pmatrix} - \delta(z_- - x_-)\,\varphi\begin{pmatrix} z_+ \\ y_- \end{pmatrix}\right\}. \tag{2.5}$$

To get a scalar bounded Bose field with scaling dimension d odd we start with a free neutral chiral Fermi field of scaling dimension $d/2$ given uniquely up to a constant by $\psi_{d/2}(u) = \partial^{(d-1)/2}\psi(u)$.

In the following we fix d odd and denote the corresponding bounded Bose field by φ, i.e., $\varphi = \psi_{d/2} \otimes \psi_{d/2}$. Many expressions can be written elegantly if we define

$$\langle fg \rangle := \int_{\mathbf{R}} f^{((d-1)/2)}(u)\, g^{((d-1)/2)}(u)\, du \quad \text{for all } f,g \in \mathcal{S}(\mathbf{R}), \tag{2.6}$$

for example, $\{\psi_{d/2}(f),\psi_{d/2}(g)\} = \langle fg \rangle \mathbf{1}$. The commutator and the double commutator of φ written with test functions are given by

$$\left[\varphi\begin{pmatrix} f_+ \\ f_- \end{pmatrix},\varphi\begin{pmatrix} g_+ \\ g_- \end{pmatrix}\right] = \langle f_- g_- \rangle b(f_+,g_+) + \langle f_+ g_+ \rangle b(f_-,g_-) \tag{2.7}$$

where

$$b(f,g) = \tfrac{1}{2}[\psi(f),\psi(g)], \tag{2.8}$$

respectively,

$$\left[\left[\varphi\begin{pmatrix}f_+\\f_-\end{pmatrix},\varphi\begin{pmatrix}g_+\\g_-\end{pmatrix}\right]\varphi\begin{pmatrix}h_+\\h_-\end{pmatrix}\right]=\langle f_-g_-\rangle\left\{\langle g_+h_+\rangle\varphi\begin{pmatrix}f_+\\h_-\end{pmatrix}-\langle f_+h_+\rangle\varphi\begin{pmatrix}g_+\\h_-\end{pmatrix}\right\}$$
$$+\langle f_+g_+\rangle\left\{\langle g_-h_-\rangle\varphi\begin{pmatrix}h_+\\f_-\end{pmatrix}-\langle f_-g_-\rangle\varphi\begin{pmatrix}h_+\\g_-\end{pmatrix}\right\}. \qquad (2.9)$$

Next we want to consider s-products of $\varphi=\psi\otimes\psi$, because with their help it is easy to outline and exemplify the proof of our theorem.

For fixed $K\in\mathbf{N}$ and given sequence $\underset{\sim}{\alpha}=(\alpha_1,\ldots,\alpha_K)$, let $\varphi_{\underset{\sim}{\alpha}}$ be the K-fold s-product of the fields $\varphi_i:=\alpha_i\varphi$, i.e., $\varphi_{\underset{\sim}{\alpha}}=(\alpha_1\varphi)s(\alpha_2\varphi)s\cdots s(\alpha_K\varphi)$ (see Ref. 4 for details).

From the very rules of s-products we get immediately for a commutator

$$[\varphi_{\underset{\sim}{\alpha}}(x),\varphi_{\underset{\sim}{\beta}}(y)]=(\alpha_1\beta_1[\varphi(x),\varphi(y)])s\cdots s(\alpha_K\beta_K[\varphi(x),\varphi(y)])$$
$$=[\varphi(x),\varphi(y)]_{\underline{\alpha\beta}}\quad\text{with}\quad\underline{\alpha\beta}=(\alpha_1\beta_1,\ldots,\alpha_K\beta_K), \qquad (2.10)$$

respectively for a double commutator

$$\left[\left[\varphi_{\underset{\sim}{\alpha}}\begin{pmatrix}f_+\\f_-\end{pmatrix},\varphi_{\underset{\sim}{\beta}}\begin{pmatrix}g_+\\g_-\end{pmatrix}\right]\varphi_{\underset{\sim}{\gamma}}\begin{pmatrix}h_+\\h_-\end{pmatrix}\right]=\langle f_-g_-\rangle\left\{\langle g_+h_+\rangle\varphi_{\underset{\sim}{\delta}}\begin{pmatrix}f_+\\h_-\end{pmatrix}-\langle f_+h_+\rangle\varphi_{\underset{\sim}{\delta}}\begin{pmatrix}g_+\\h_-\end{pmatrix}\right\}$$
$$+\langle f_+g_+\rangle\left\{\langle g_-h_-\rangle\varphi_{\underset{\sim}{\delta}}\begin{pmatrix}h_+\\f_-\end{pmatrix}-\langle f_-g_-\rangle\varphi_{\underset{\sim}{\delta}}\begin{pmatrix}h_+\\g_-\end{pmatrix}\right\}, \qquad (2.11)$$

with $\underset{\sim}{\delta}=\overline{\alpha\beta\gamma}$

Later on we shall show in a general setting the analoga to these two equations as a starting point of our proof!

Starting from $\Phi_1:=\Phi:=\varphi_{\underset{\sim}{\alpha}}$, respectively $\Delta_2(x,y):=[\varphi_{\underset{\sim}{\alpha}}(x),\varphi_{\underset{\sim}{\alpha}}(y)]=[\varphi(x),\varphi(y)]_{\overline{\alpha^2}}$, we define a sequence of field operators $\Phi_{2k-1}=\varphi_{\overline{\alpha^{2k-1}}}$, resp. commutator functions $\Delta_{2k}(x,y):=[\Phi_{2k-1}(x),\Phi_1(y)]$ for all $k\in\mathbf{N}$. Obviously we have

$$[\Phi_{2k-1}(x),\Phi_{2l-1}(y)]=\Delta_{2(k+l-1)}(x,y) \qquad (2.12)$$

and

$$\left[\left[\Phi_{2k-1}\begin{pmatrix}f_+\\f_-\end{pmatrix},\Phi_{2l-1}\begin{pmatrix}g_+\\g_-\end{pmatrix}\right]\Phi_{2m-1}\begin{pmatrix}h_+\\h_-\end{pmatrix}\right]=\langle f_-g_-\rangle\left\{\langle g_+h_+\rangle\Phi_n\begin{pmatrix}f_+\\h_-\end{pmatrix}-\langle f_+h_+\rangle\Phi_n\begin{pmatrix}g_+\\h_-\end{pmatrix}\right\}$$
$$+\langle f_+g_+\rangle\left\{\langle g_-h_-\rangle\Phi_n\begin{pmatrix}h_+\\f_-\end{pmatrix}-\langle f_-g_-\rangle\Phi_n\begin{pmatrix}h_+\\g_-\end{pmatrix}\right\} \qquad (2.13)$$

with $n=2(k+l+m)-3$.

Let V_n, resp. W_n, denote the n-point functions of φ, resp. of the s-product $\varphi_{\underset{\sim}{\alpha}}$. Because the underlying chiral Fermi field is a free field, all odd n-point functions have to vanish, i.e., $V_{2k-1}\equiv 0\equiv W_{2k-1}$ and by the governing rules of s-products we have for the even truncated n-point functions the relation $W_{2n}^T=c_{2n}V_{2n}^T$ with $c_{2n}=\Sigma_i\alpha_i^{2n}$. From the sequence $\{\Phi_{2k-1}|k\in\mathbf{N}\}$ we can reconstruct the c_n without referring to $\underset{\sim}{\alpha}$ as follows:

By using Eqs. (2.13) and (2.12) we get for any $(2n-1)$-fold commutator

$$[[\cdots[\Phi(x_1),\Phi(x_2)]\cdots]\Phi(x_{2n})]\sim\Delta_{2n}\quad(\text{sum of }2^{2n-2}\text{ terms!}). \qquad (2.14)$$

By scaling covariance we have for the vacuum expectation values

$$(\Omega,\Delta_{2n}(x,y)\Omega)=c_{2n}\langle[\varphi(x),\varphi(y)]\rangle. \qquad (2.15)$$

Because the algebraic relations (2.12), (2.13) and (2.7), (2.9) have an identical structure we get

$$(\Omega,[[\cdots[\Phi(x_1),\Phi(x_2)]\cdots]\Phi(x_{2n})]\Omega)=c_{2n}\langle[[\cdots[\varphi(x_1),\varphi(x_2)]\cdots]\varphi(x_{2n})]\rangle. \tag{2.16}$$

and this implies $W^T_{2n}=c_{2n}V^T_{2n}$! It is another task to reconstruct $\underset{\sim}{\alpha}$, but in Sec. VI we shall describe a very effective method for this job.

III. LEMMATA ON COMMUTATORS

In this section we want to exhibit the structure of commutators and double commutators. But in the course of our proof we shall encounter not only the field Φ, but also new fields which live in the same Hilbert space, are relatively local to Φ, and fulfill the assumptions **A**.

Lemma 1: Let A_1 and A_2 be two fields with scaling dimension d, relatively local to Φ, and fulfilling the assumptions **A**. For the commutator we have

$$\begin{aligned}[A_1(x),A_2(y)]&=(-1)^{(d-1)/2}\delta^{(d-1)}(y_- - x_-)B^{(+)}_{12}(x_+,y_+)\\&\quad+(-1)^{(d-1)/2}\delta^{(d-1)}(y_+ - x_+)B^{(-)}_{12}(x_-,y_-).\end{aligned} \tag{3.1}$$

Proof: (a) From space- and timelike commutativity $[A_1(x),A_2(y)]=0$ if $(y-x)^2\neq 0$ and scale covariance with dimension d we get

$$[A_1(x),A_2(y)]\Omega(y_+-x_+)^d(y_- - x_-)^d\equiv 0. \tag{3.2}$$

By the Reeh–Schlieder theorem this implies

$$\begin{aligned}[A_1(x),A_2(y)]&=\delta^{(d-1)}(y_- - x_-)B^{(+)}_{12}(x_+,y_+)+\delta^{(d-1)}(y_+-x_+)B^{(-)}_{12}(x_-,y_-)\\&\quad+\sum_{k=0}^{d-2}\left\{\delta^{(k)}(y_- - x_-)B^{(+)}_k\left(x_+,y_+,\frac{x_-+y_-}{2}\right)\right.\\&\quad\left.+\delta^{(k)}(y_+-x_+)B^{(-)}_k\left(x_-,y_-,\frac{x_++y_+}{2}\right)\right\}\end{aligned} \tag{3.3}$$

and because of scale covariance the coefficients $B^{(+)}_k$, resp. $B^{(-)}_k$, have to depend on $(x_- + y_-)/2$, resp. $(x_+ + y_+)/2$, too. The corresponding scaling dimension is $d-k-1$.

(b) To prove Lemma 1 we have to get rid of the terms $\Sigma_{k=0}^{d-2}\{\cdots\}$! These terms show up only if $d\geqslant 2$.

Let $F(x)=f_+(x_+)f_-(x_-)$ and $G(y)=g_+(y_+)g_-(y_-)$ with

(i) supp $f_+\cap$ supp $g_+=\varnothing$ [This eliminates all terms with an upper index $(-)$!]
(ii) supp $g_-\subseteq[-1,1]$.
(iii) $f_-\in\mathcal{S}(\mathbf{R})$ such that $f_-(x_-)\equiv 1$ on $[-1,1]$.

To evaluate integrals of the form $\int\int B((x_-+y_-)/2)\delta^{(k)}(y_- - x_-)f_-(x_-)g_-(y_-)dx_-dy_-$ we introduce new variables $u=(x_-+y_-)/2$ and $\xi=y_- - x_-$:

$$\begin{aligned}&\int\int B\left(\frac{x_-+y_-}{2}\right)\delta^{(k)}(y_- - x_-)f_-(x_-)g_-(y_-)dx_-dy_-\\&=\int B(u)\int\delta^{(k)}(\xi)f_-(u-\xi/2)g_-(u+\xi/2)d\xi\,du\\&=\int B(u)(-\partial_\xi)^k f_-(u-\xi/2)g_-(u+\xi/2)\Big|_{\xi=0}du\\&=(-\tfrac{1}{2})^k B(\partial^k g_-)\end{aligned} \tag{3.4}$$

because $f_- \equiv 1$ on the support of g_-. Therefore, we have

$$[A_1(F),A_2(G)] = \sum_{k=0}^{d-2} \left(-\frac{1}{2}\right)^k B_k^{(+)}(f_+,g_+,\partial^k g_-). \tag{3.5}$$

We remark that space- and timelike commutativity acts as a cutoff for f_-. Therefore we can replace f_- by $(f_-)_\mu(x_-) = f_-(x_-/\mu)$ with $\mu \geqslant 1$. This does not change the commutator because we still have $(f_-)_\mu(x_-) \equiv 1$ if $|x_-| \leqslant 1$, i.e.,

$$[A_1(F),A_2(G)] = [A_1(F_{(1,\mu)}),A_2(G)] \quad \text{if } \mu \geqslant 1, \tag{3.6}$$

but for $d>2$ we know by our assumption (**A4**)

$$\|A_i(F_{(1,\mu)})\| = \frac{1}{(\mu)^{d/2-1}} \|A_i(F)\| \to 0 \quad \text{as } \mu \to \infty \quad \text{for } i=1,2. \tag{3.7}$$

As a result of our assumptions we get

$$0 = \sum_{k=0}^{d-2} \left(-\frac{1}{2}\right)^k B_k^{(+)}(f_+,g_+,\partial^k g_-). \tag{3.8}$$

We can also interchange the roles of f_- and g_-, i.e., we put $\hat{F}(x) = f_+(x_+)g_-(x_-)$ and $\hat{G}(y) = g_+(y_+)f_-(y_-)$. Then as before we get for all $\mu \geqslant 1$

$$[A_1(\hat{F}),A_2(\hat{G})] = \sum_{k=0}^{d-2} \left(\frac{1}{2}\right)^k B_k^{(+)}(f_+,g_+,\partial^k g_-) = [A_1(\hat{F}),A_2(\hat{G}_{(1,\mu)})], \tag{3.9}$$

and therefore the additional equation

$$0 = \sum_{k=0}^{d-2} \left(\frac{1}{2}\right)^k B_k^{(+)}(f_+,g_+,\partial^k g_-). \tag{3.10}$$

Therefore, in the case $d=3$ we have

$$B_0^{(+)}(f_+,g_+,g_-) = 0 = B_1^{(+)}(f_+,g_+,\partial g_-), \tag{3.11}$$

which solves our task and for $d=5$ we get the two equations

$$B_0^{(+)}(f_+,g_+,g_-) + \tfrac{1}{4} B_2^{(+)}(f_+,g_+,\partial^2 g_-) = 0, \tag{3.12}$$

$$B_1^{(+)}(f_+,g_+,\partial g_-) + \tfrac{1}{4} B_3^{(+)}(f_+,g_+,\partial^3 g_-) = 0. \tag{3.13}$$

To get more (and, in fact, sufficiently many) equations, we consider the modified commutators $[A_1(x),A_2(y)](y_- - x_-)^l$.

Let $(\mathbf{X}^m f_-)(x_-) := x_-^m f_-(x_-)$. Then

$$\int\int \left[A_1\begin{pmatrix} f_+ \\ x_- \end{pmatrix}, A_2\begin{pmatrix} g_+ \\ y_- \end{pmatrix}\right](y_- - x_-)^l f_-(x_-) g_-(y_-) dx_- dy_-$$

$$= \sum_{m=0}^{l} \binom{l}{m} (-1)^m \left[A_1\begin{pmatrix} f_+ \\ \mathbf{X}^m f_- \end{pmatrix}, A_2\begin{pmatrix} g_+ \\ \mathbf{X}^{l-m} g_- \end{pmatrix}\right] \tag{3.14}$$

but

$$\left\| A_i\begin{pmatrix} f_+ \\ \mathbf{X}^m(f_-)_\mu \end{pmatrix}\right\| = \mu^m \left\| A_i\begin{pmatrix} f_+ \\ (\mathbf{X}^m f_-)_\mu \end{pmatrix}\right\| = \frac{1}{\mu^{d/2-1-m}}\left\| A_i\begin{pmatrix} f_+ \\ \mathbf{X}^m f_- \end{pmatrix}\right\| \xrightarrow{\mu\to\infty} 0 \quad \text{if } 0 \leqslant m < d/2-1, \tag{3.15}$$

and, therefore, as long as $l<d/2-1$, all the commutators (3.14) vanish. Evaluating the corresponding rhs we get

$$0=\sum_{k=l\geqslant 0}^{d-2}\left(-\frac{1}{2}\right)^k \frac{k!}{(k-l)!} B_k^{(+)}(f_+,g_+,\partial^{k-l}g_-), \tag{3.16}$$

where we have used

$$\partial_\xi^k \xi^l h(\xi)|_{\xi=0} = \begin{cases} 0, & \text{if } l>k, \\ \dfrac{k!}{(k-l)!}\partial_\xi^{k-l}h(0), & \text{if } l\leqslant k, \end{cases} \tag{3.17}$$

and h is of the form $h(\xi)=f_-(u-\xi/2)g_-(u+\xi/2)$. As before we can interchange f_- and g_- and end up with

$$0=\sum_{k=l\geqslant 0}^{d-2}\left(\frac{1}{2}\right)^k \frac{k!}{(k-l)!} B_k^{(+)}(f_+,g_+,\partial^{k-1}g_-), \tag{3.18}$$

i.e., we get separate equations for even and odd k's. For d odd we have $l=0,1,...,(d-3)/2$, and therefore $d-1$ independent equations for the coefficients $B_0^{(+)},\dots,B_{d-2}^{(+)}$. In the case $d=5$ the possible values of l are 0 and 1. So beside (3.12) and (3.13) we get the two further equations,

$$B_1^{(+)}(f_+,g_+,g_-)+\tfrac{3}{4}B_3^{(+)}(f_+,g_+,\partial^2 g_-)=0, \tag{3.19}$$

$$B_2^{(+)}(f_+,g_+,\partial g_-)=0, \tag{3.20}$$

which altogether imply $B_k^{(+)}(f_+,g_+,\partial^k g_-)=0,\ k=0,1,2,3,4.$

Remark: Because the operators $B_k^{(+)}(f_+,g_+,g_-),\ k=0,...,d-2$, have scaling dimension $d-k-1$ wrt u, an equation like $B_k^{(+)}(f_+,g_+,\partial^m g_-)=0$ implies $B_k^{(+)}(f_+,g_+,g_-)=0$ as can be seen from the two-point function.

Now we interchange the coordinates x_+,y_+ with x_-,y_-. By analogous reasoning we get $B_k^{(-)}(f_-,g_-,\partial^k g_+)=0$, $k=0,1,2,3,4$, as long as $\operatorname{supp} f_-\cap\ \operatorname{supp} g_-=\varnothing$.

(c) Up to now we have shown

$$[A_1(x),A_2(y)]=\delta^{(d-1)}(y_--x_-)B_{12}^{(+)}(x_+,y_+)+\delta^{(d-1)}(y_+-x_+)B_{12}^{(-)}(x_-,y_-) + \sum_{k,l=0}^{d-2}\underbrace{\delta^{(k)}(y_+-x_+)\delta^{(l)}(y_--x_-)}_{\text{contact terms!}}B_{kl}\left(\frac{x_++y_+}{2},\frac{x_-+y_-}{2}\right), \tag{3.21}$$

because part (b) of our proof relied heavily on the fact that either f_+ and g_+, or f_- and g_-, have disjoint supports!

In a first step we show for $m=0,...,(d-3)/2$

$$\lim_{\lambda\to\infty} B_{12}^{(\pm)}(\mathbf{X}^m(f)_\lambda,g)=0=\lim_{\lambda\to\infty} B_{12}^{(\pm)}(f,\mathbf{X}^m(g)_\lambda). \tag{3.22}$$

To prove this statement we start with

$$[A_1(x), A_2(y)](y_- - x_-)^{d-1} = (d-1)!\,\delta(y_- - x_-)B_{12}^{(+)}(x_+, y_+) + \delta^{(d-1)}(y_+ - x_+)B_{12}^{(-)}(x_-, y_-)(y_- - x_-)^{d-1} = \sum_{l=0}^{d-1}\binom{d-1}{l}(-1)^l[x_-^{d-1-l}A_1(x), y_-^l A_2(y)], \tag{3.23}$$

or, written with test functions,

$$\int\int\left[A_1\binom{f_+}{x_-}, A_2\binom{g_+}{y_-}\right](y_- - x_-)^{d-1} f_-(x_-)g_-(y_-)\,dx_-\,dy_- = \sum_{l=0}^{d-1}\binom{d-1}{l}(-1)^l\left[A_1\binom{f_+}{\mathbf{X}^{d-1-l}f_-}, A_2\binom{g_+}{\mathbf{X}^l g_-}\right] = (d-1)!\int f_-g_-(u)\,du\,B_{12}^{(+)}(f_+, g_+) + (-1)^{(d-1)/2}\int f_+^{((d-1)/2)}g_+^{((d-1)/2)}(u)\,du \times\int\int B_{12}^{(-)}(x,y)(y-x)^{d-1}f_-(x)g_-(y)\,dx\,dy. \tag{3.24}$$

Therefore we have the bound

$$\|B_{12}^{(+)}(f_+, g_+)\|(d-1)!\left|\int f_-g_-(u)du\right| \leq \sum_{l=0}^{d-1}\binom{d-1}{l}\left\|\left[A_1\binom{f_+}{\mathbf{X}^{d-1-l}f_-}, A_2\binom{g_+}{\mathbf{X}^l g_-}\right]\right\| + \left|\int f_+^{((d-1)/2)}g_+^{((d-1)/2)}(u)du\right| \times\left\|\int\int B_{12}^{(-)}(x,y)(y-x)^{d-1}f_-(x)g_-(y)\,dx\,dy\right\|. \tag{3.25}$$

Inserting $\mathbf{X}^m(f_+)_\lambda$ with $m = 1,...,(d-3)/2$ instead of f_+, using bounds similar to (3.15), and doing the limit $\lambda\to\infty$ gives

$$\lim_{\lambda\to\infty} B_{12}^{(+)}(\mathbf{X}^m(f_+)_\lambda, g_+) = 0.$$

The other equations of the statement (3.22) can be proven in an analogous way.

To show that all contact terms

$$B_{kl}\left(\frac{x_+ + y_+}{2}, \frac{x_- + y_-}{2}\right) \quad \text{for } k,l = 0,...,d-2$$

have to vanish, we start from

$$[A_1(x), A_2(y)](y_+ - x_+)^{m_+}(y_- - x_-)^{m_-}$$

for suitably chosen exponents m_+, resp. m_-. Let us demonstrate this for the simplest case $d = 3$:

$$\int\int \left[A_1\begin{pmatrix} f_+ \\ x\end{pmatrix}, A_2\begin{pmatrix} g_+ \\ y\end{pmatrix}\right](y-x)f_-(x)g_-(y)dxdy$$

$$=\left[A_1\begin{pmatrix} f_+ \\ f_-\end{pmatrix}, A_2\begin{pmatrix} g_+ \\ \mathbf{X}g_-\end{pmatrix}\right]-\left[A_1\begin{pmatrix} f_+ \\ \mathbf{X}f_-\end{pmatrix}, A_2\begin{pmatrix} g_+ \\ g_-\end{pmatrix}\right]$$

$$=B_{12}^{(+)}(f_+,g_+)\int (f_-g'_- - f'_-g_-)(u)du-\{B_{12}^{(-)}(f_-,\mathbf{X}g_-)-B_{12}^{(-)}(\mathbf{X}f_-,g_-)\}$$

$$\times\int f'_+g'_+(u)\,du \quad -B_{01}(f_+g_+,f_-g_-)+\frac{1}{2}B_{11}(f_+g'_+-f'_+g_+,f_-g_-).$$

By choosing $g_\pm\in\mathcal{D}([-1,1]$ and $f_\pm\in\mathcal{S}(\mathbf{R})$ with $f(x)\equiv 1$ on $[-1,1]$ we get

$$=-B_{01}(g_+,g_-)+\tfrac{1}{2}B_{11}(g'_+,g_-). \tag{3.26}$$

Now we replace f_+ by $(f_+)_\lambda$. For $\lambda\to\infty$, the lhs of (3.26) goes to zero and we have

$$0=-B_{01}(g_+,g_-)+\tfrac{1}{2}B_{11}(g'_+,g_-), \tag{3.27}$$

but by interchanging f_+ with g_+ we have also

$$0=-B_{01}(g_+,g_-)-\tfrac{1}{2}B_{11}(g'_+,g_-), \tag{3.28}$$

and therefore

$$B_{01}(g_+,g_-)=0=B_{11}(g'_+,g_-). \tag{3.29}$$

Interchanging the $+$ and $-$ coordinates we get

$$B_{10}(g_+,g_-)=0=B_{11}(g_+,g'_-). \tag{3.30}$$

Finally starting from

$$\left[A_1\begin{pmatrix} f_+ \\ f_-\end{pmatrix}, A_2\begin{pmatrix} g_+ \\ g_-\end{pmatrix}\right]=-B_{12}^{(+)}(f_+,g_+)\int f'_-g'_-(u)du-B_{12}^{(-)}(f_-,g_-)\int f'_+g'_+(u)\,du$$

$$+B_{00}(f_+g_+,f_-g_-)+\tfrac{1}{2}B_{01}(f_+g_+,f'_-g_- - f_-g'_-)$$

$$+\tfrac{1}{2}B_{10}(f'_+g_+-f_+g'_+,f_-g_-)+\tfrac{1}{4}B_{11}(f'_+g_+-f_+g'_+,f'_-g_- - f_-g'_-),$$

using the above methods, we end up with

$$B_{00}(g_+,g_-)=0. \tag{3.31}$$

For higher scaling dimensions d the calculations are a little bit cumbersome but not difficult. This finishes the proof of Lemma 1.

For the sequel it is convenient to write Lemma 1 in the form

$$[A_1(F),A_2(G)]=\langle f_-g_-\rangle B_{12}^{(+)}(f_+,g_+)+\langle f_+g_+\rangle B_{12}^{(-)}(f_-,g_-) \tag{3.32}$$

with $F=f_+\otimes f_-$, $G=g_+\otimes g_-$ and $\langle fg\rangle$ is given by definition (2.6).

Lemma 2: Let A_i, $i=1,2,3$, be three fields with scaling dimension d, relatively local to Φ, and fulfilling the assumptions **A**. For the double commutator we have

$$[[A_1(F),A_2(G)]A_3(H)]=\langle f_-g_-\rangle\left\{\langle g_+h_+\rangle C_1\begin{pmatrix} f_+\\ h_-\end{pmatrix}-\langle f_+h_+\rangle C_2\begin{pmatrix} g_+\\ h_-\end{pmatrix}\right\}$$
$$+\langle f_+g_+\rangle\left\{\langle g_-h_-\rangle C_3\begin{pmatrix} h_+\\ f_-\end{pmatrix}-\langle f_-h_-\rangle C_4\begin{pmatrix} h_+\\ g_-\end{pmatrix}\right\}, \qquad (3.33)$$

and the local fields C_j fulfill the same assumptions as the A_i.

Proof: (a) From Lemma 1 we know

$$[[A_1(F),A_2(G)]A_3(H)]=\langle f_-g_-\rangle[B_{12}^{(+)}(f_+,g_+),A_3(H)]+\langle f_+g_+\rangle[B_{12}^{(-)}(f_-,g_-),A_3(H)], \qquad (3.34)$$

and by the Jacobi identity

$$=[[A_1(F),A_3(H)]A_2(G)]+[[A_3(H),A_2(G)]A_1(F)]$$
$$=\langle f_-h_-\rangle[B_{13}^{(+)}(f_+,h_+),A_2(G)]+\langle f_+h_+\rangle[B_{13}^{(-)}(f_-,h_-),A_2(G)]+\langle h_-g_-\rangle$$
$$\times[B_{32}^{(+)}(h_+,g_+),A_1(F)]+\langle h_+g_+\rangle[B_{32}^{(-)}(h_-,g_-),A_1(F)].$$

Therefore each term on the rhs of (3.34) does contain a factor $\langle f_-g_-\rangle$, resp. $\langle f_+g_+\rangle$, combined with a factor out of the set $\{\langle f_-h_-\rangle,\langle f_+h_+\rangle,\langle g_-h_-\rangle,\langle g_+h_+\rangle\}$. On the other hand, $[B_{12}^{(+)}(f_+,g_+),A_3(H)]$ no longer depends on f_- and g_-! So only the combinations $\langle f_-g_-\rangle\langle f_+h_+\rangle$ and $\langle f_-g_-\rangle\langle g_+h_+\rangle$ can show up. Therefore

$$[[A_1(F),A_2(G)]A_3(H)]=\langle f_-g_-\rangle\left\{\langle g_+h_+\rangle C_1\begin{pmatrix} f_+\\ h_-\end{pmatrix}-\langle f_+h_+\rangle C_2\begin{pmatrix} g_+\\ h_-\end{pmatrix}\right\}$$
$$+\langle f_+g_+\rangle\left\{\langle g_-h_-\rangle C_3\begin{pmatrix} h_+\\ f_-\end{pmatrix}-\langle f_-h_-\rangle C_4\begin{pmatrix} h_+\\ g_-\end{pmatrix}\right\}. \qquad (3.35)$$

(b) We can prepare $C_1\binom{f_+}{h_-}$ in the following way:

We shift the test functions f_-, g_-, g_+, and h_+ by a and in the limit $a\to\infty$ we get

$$C_1\begin{pmatrix} f_+\\ h_-\end{pmatrix}=\lim_{a\to\infty}\left[\left[A_1\begin{pmatrix} f_+\\ \{h_-\}_a\end{pmatrix},\quad A_2\begin{pmatrix} \{g_+\}_a\\ \{g_-\}_a\end{pmatrix}\right]A_3\begin{pmatrix} \{h_+\}_a\\ h_-\end{pmatrix}\right] \qquad (3.36)$$

because for fixed f_+ and h_-

$$\lim_{a\to\infty}\langle f_+\{h_+\}_a\rangle=\lim_{a\to\infty}\int f_+^{((d-1)/2)}(u)h_+^{((d-1)/2)}(u-a)\,du=0,$$

and so on. The other operators C_i can be prepared in an analogous manner. From this explicit representation it is easy to see that all $C_i(x)$ are local fields and fulfill assumption **A** again. This proves Lemma 2.

Corollary 3: If in Lemma 2 all three fields A_i are equal, then the four fields C_j are equal, too.

Proof: By Lemma 2 and the Jacobi identity we have

$$0=[[A(F),A(G)]A(H)]+[[A(G),A(H)]A(F)]+[[A(H),A(F)]A(G)]$$
$$=\langle f_-g_-\rangle\langle g_+h_+\rangle\{C_1-C_4\}\begin{pmatrix} f_+\\ h_-\end{pmatrix}-\langle f_-g_-\rangle\langle f_+h_+\rangle\{C_2-C_3\}\begin{pmatrix} g_+\\ h_-\end{pmatrix}$$
$$+\langle f_+g_+\rangle\langle g_-h_-\rangle\{C_3-C_2\}\begin{pmatrix} h_+\\ f_-\end{pmatrix}-\langle f_+g_+\rangle\langle f_-h_-\rangle\{C_4-C_3\}\begin{pmatrix} h_+\\ g_-\end{pmatrix}$$

$$+\langle g_+h_+\rangle\langle f_-h_-\rangle\{C_3-C_2\}\binom{f_+}{g_-}+\langle g_-h_-\rangle\langle f_+h_+\rangle\{C_1-C_4\}\binom{g_+}{f_-}.$$

By a suitable limit procedure we can pick out every one of the six terms in curly brackets and therefore $C_1=C_2=C_3=C_4$.

IV. THE DEFINING SEQUENCE FOR MULTIPLE COMMUTATORS

Lemmas 1 and 2 together with the corollary enable us to calculate higher commutators of the field Φ we started with.

Lemma 4: Let Φ fulfill the assumptions (**A**). Then there exists a sequence of field operators

$$\Phi_1,\Phi_3,\Phi_5,\dots,\Phi_{2k-1},\dots,$$

and a sequence of commutator functions (bilocal operators)

$$\Delta_2,\Delta_4,\Delta_6,\dots,\Delta_{2k},\dots,$$

such that

(i) all $\Phi_{2k-1},k\in\mathbf{N}$, fulfill (**A**) again and are relatively local to Φ;

(ii) all $\Delta_{2k},k\in\mathbf{N}$, are of the form

$$\Delta_{2k}(F,G)=\langle f_-g_-\rangle B_{2k}^{(+)}(f_+,g_+)+\langle f_+g_+\rangle B_{2k}^{(-)}(f_-,g_-), \tag{4.1}$$

and the Δ_{2k} are antisymmetric, i.e., $\Delta_{2k}(F,G)=-\Delta_{2k}(G,F)$;

(iii)

$$[\Phi_{2k-1}(F),\Phi_{2l-1}(G)]=\Delta_{2(k+l-1)}(F,G); \tag{4.2}$$

(iv)

$$[\Delta_{2k}(F,G),\Phi_{2l-1}(H)]=\langle f_-g_-\rangle\langle g_+h_+\rangle\Phi_{2(k+l)-1}\binom{f_+}{h_-}-\langle f_-g_-\rangle\langle f_+h_+\rangle\Phi_{2(k+l)-1}\binom{g_+}{h_-}$$
$$+\langle f_+g_+\rangle\langle g_-h_-\rangle\Phi_{2(k+l)-1}\binom{h_+}{f_-}-\langle f_+g_+\rangle\langle f_-h_-\rangle\Phi_{2(k+l)-1}\binom{h_+}{g_-}. \tag{4.3}$$

Proof: We want to prove the lemma by an inductive definition of the two sequences.

(1) Let $\Phi_1:=\Phi$ and let $\Delta_2(F,G):=[\Phi(F),\Phi(G)]$. The antisymmetry of Δ_2 is obvious.

(2) Assume we have defined up to $n\in\mathbf{N}$ field operators $\Phi_1,\Phi_3,\dots,\Phi_{2n-1}$ relatively local to Φ and antisymmetric commutator functions $\Delta_2,\Delta_4,\dots,\Delta_{2n}$ with the following two properties:

(a)

$$[\Delta_{2(m-k)}(F,G),\Phi_{2k-1}(H)]=\langle f_-g_-\rangle\langle g_+h_+\rangle\Phi_{2m-1}\binom{f_+}{h_-}-\langle f_-g_-\rangle\langle f_+h_+\rangle\Phi_{2m-1}\binom{g_+}{h_-}$$
$$+\langle f_+g_+\rangle\langle g_-h_-\rangle\Phi_{2m-1}\binom{h_+}{f_-}-\langle f_+g_+\rangle\langle f_-h_-\rangle\Phi_{2m-1}\binom{h_+}{g_-}, \tag{4.4}$$

(b)

$$[\Phi_{2(m-k)-1}(F),\Phi_{2k-1}(G)]=\Delta_{2m}(F,G) \tag{4.5}$$

valid for all $m\leqslant n$ and $k=1,\dots,m$.

(3) To define Φ_{2n+1} we proceed as follows.

(α) By Lemma 2 we know

$$[\Delta_{2n}(F,G),\,\Phi_1(H)]=[[\Phi_{2(n-k)}(F),\Phi_{2k-1}(G)]\Phi_1(H)]$$
$$=\langle f_-g_-\rangle\langle g_+h_+\rangle C_1\binom{f_+}{h_-}-\langle f_-g_-\rangle\langle f_+h_+\rangle C_2\binom{g_+}{h_-}$$
$$+\langle f_+g_+\rangle\langle g_-h_-\rangle C_3\binom{h_+}{f_-}-\langle f_+g_+\rangle\langle f_-h_-\rangle C_4\binom{h_+}{g_-}, \tag{4.6}$$

but the antisymmetry $\Delta_{2n}(F,G)=-\Delta_{2n}(G,F)$ implies $C_1=C_2$ and $C_3=C_4$.

(β) Now we use the Jacobi identity

$$0=[[\Phi_{2(n-k)+1}(F),\,\Phi_{2k-1}(G)]\Phi_1(H)]+[[\Phi_{2k-1}(G),\,\Phi_1(H)]\Phi_{2(n-k)+1}(F)]$$
$$+[[\Phi_1(H),\,\Phi_{2(n-k)+1}(F)]\Phi_{2k-1}(G)]$$
$$=[\Delta_{2n}(F,G),\,\Phi_1(H)]+[\Delta_{2k}(G,H),\,\Phi_{2(n-k)+1}(F)]+[\Delta_{2(n-k+1)}(H,F),\,\Phi_{2k-1}(G)]$$

(Δ_{2n},Δ_{2k}, and $\Delta_{2(n-k+1)}$ are antisymmetric by assumption)

$$=\langle f_-g_-\rangle\langle g_+h_+\rangle\{C_1-D_3\}\binom{f_+}{h_-}-\langle f_-g_-\rangle\langle f_+h_+\rangle\{C_1-E_3\}\binom{g_+}{h_-}$$
$$+\langle f_+g_+\rangle\langle g_-h_-\rangle\{C_3-D_1\}\binom{h_+}{f_-}-\langle f_+g_+\rangle\langle f_-h_-\rangle\{C_3-E_1\}\binom{h_+}{g_-}$$
$$+\langle g_+h_+\rangle\langle f_-h_-\rangle\{D_3-E_1\}\binom{f_+}{g_-}+\langle g_-h_-\rangle\langle f_+h_+\rangle\{D_1-E_3\}\binom{g_+}{f_-},$$

where C_i, D_i, and E_i are the field operators corresponding to the three commutators. As in the proof of Corollary 3 we must have $C_1=C_3=D_1=D_3=E_1=E_3=:\Phi_{2n+1}$, and we take this as definition of Φ_{2n+1}. This proves (a) for the case $m=n+1$.

(4) As the definition of Δ_{2n+2} we take

$$\Delta_{2n+2}(F,G)=[\Phi_{2n+1}(F),\Phi_1(G)] \tag{4.7}$$

and, because of Lemma 1, this is meaningful and guarantees the right structure as given in (4.1). We still have to show the antisymmetry of Δ_{2n+2} and the relation

$$\Delta_{2n+2}(F,G)=[\Phi_{2(n-k)+3}(F),\,\Phi_{2k-1}(G)]\quad\text{for all}\quad k=1,2,\ldots,2n+1.$$

From part (3) we know already

$$[\Delta_{2k}(F,G),\,\Phi_{2(n-k)+1}(H)]=\langle f_-g_-\rangle\langle g_+h_+\rangle\Phi_{2n+1}\binom{f_+}{h_-}-\langle f_-g_-\rangle\langle f_+h_+\rangle\Phi_{2n+1}\binom{g_+}{h_-}$$
$$+\langle f_+g_+\rangle\langle g_-h_-\rangle\Phi_{2n+1}\binom{h_+}{f_-}-\langle f_+g_+\rangle\langle f_-h_-\rangle\Phi_{2n+1}\binom{h_+}{g_-},$$

and therefore

$$
\begin{aligned}
&[[\Delta_{2k}(F,G),\,\Phi_{2(n-k)+1}(U)]\Phi_1(V)]\\
&=\langle f_-g_-\rangle\langle g_+u_+\rangle\left[\Phi_{2n+1}\begin{pmatrix}f_+\\u_-\end{pmatrix},\,\Phi_1(V)\right]-\langle f_-g_-\rangle\langle f_+u_+\rangle\left[\Phi_{2n+1}\begin{pmatrix}g_+\\u_-\end{pmatrix},\,\Phi_1(V)\right]\\
&\quad+\langle f_+g_+\rangle\langle g_-u_-\rangle\left[\Phi_{2n+1}\begin{pmatrix}u_+\\f_-\end{pmatrix},\,\Phi_1(V)\right]-\langle f_+g_+\rangle\langle f_-u_-\rangle\left[\Phi_{2n+1}\begin{pmatrix}u_+\\g_-\end{pmatrix},\,\Phi_1(V)\right]\\
&=\langle f_-g_-\rangle\langle g_+u_+\rangle\Delta_{2n+2}\left(\begin{pmatrix}f_+\\u_-\end{pmatrix},V\right)-\langle f_-g_-\rangle\langle f_+u_+\rangle\Delta_{2n+2}\left(\begin{pmatrix}g_+\\u_-\end{pmatrix},V\right)\\
&\quad+\langle f_+g_+\rangle\langle g_-u_-\rangle\Delta_{2n+2}\left(\begin{pmatrix}u_+\\f_-\end{pmatrix},V\right)-\langle f_+g_+\rangle\langle f_-u_-\rangle\Delta_{2n+2}\left(\begin{pmatrix}u_+\\g_-\end{pmatrix},V\right)
\end{aligned}\tag{4.8}
$$

and by the Jacobi identity

$$
\begin{aligned}
&=[[\Delta_{2k}(F,G),\,\Phi_1(V)]\Phi_{2(n-k)+1}(U)]+\underbrace{[[\Phi_1(V),\,\Phi_{2(n-k)+1}(U)]}_{\Delta_{2(n-k+1)}(V,U)},\Delta_{2k}(F,G)]\\
&=\langle f_-g_-\rangle\langle g_+v_+\rangle\left[\Phi_{2k+1}\begin{pmatrix}f_+\\v_-\end{pmatrix},\,\Phi_{2(n-k)+1}(U)\right]-\langle f_-g_-\rangle\langle f_+v_+\rangle\left[\Phi_{2k+1}\begin{pmatrix}g_+\\v_-\end{pmatrix},\,\Phi_{2(n-k)+1}(U)\right]\\
&\quad+\langle f_+g_+\rangle\langle g_-v_-\rangle\left[\Phi_{2k+1}\begin{pmatrix}v_+\\f_-\end{pmatrix},\,\Phi_{2(n-k)+1}(U)\right]-\langle f_+g_+\rangle\langle f_-v_-\rangle\\
&\quad\times\left[\Phi_{2k+1}\begin{pmatrix}v_+\\g_-\end{pmatrix},\,\Phi_{2(n-k)+1}(U)\right]+[\Delta_{2(n-k+1)}(V,U),\Delta_{2k}(F,G)].
\end{aligned}
$$

Let us denote $[\Phi_{2k+1}(F),\,\Phi_{2(n-k)+1}(G)]=:\hat{\Delta}(F,G)$ for the moment. Then we can rewrite Eq. (4.8) as

$$
\begin{aligned}
[\Delta_{2(n-k+1)}(V,U),\Delta_{2k}(F,G)]=\langle f_-g_-\rangle\Bigg\{&\langle g_+u_+\rangle\Delta_{2n+2}\left(\begin{pmatrix}f_+\\u_-\end{pmatrix},V\right)-\langle g_+v_+\rangle\hat{\Delta}\left(\begin{pmatrix}f_+\\v_-\end{pmatrix},U\right)\\
&-\langle f_+u_+\rangle\Delta_{2n+2}\left(\begin{pmatrix}g_+\\u_-\end{pmatrix},V\right)+\langle f_+v_+\rangle\hat{\Delta}\left(\begin{pmatrix}g_+\\v_-\end{pmatrix},U\right)\Bigg\}\\
+\langle f_+g_+\rangle\Bigg\{&\langle g_-u_-\rangle\Delta_{2n+2}\left(\begin{pmatrix}u_+\\f_-\end{pmatrix},V\right)-\langle g_-v_-\rangle\hat{\Delta}\left(\begin{pmatrix}v_+\\f_-\end{pmatrix},U\right)\\
&-\langle f_-u_-\rangle\Delta_{2n+2}\left(\begin{pmatrix}u_+\\g_-\end{pmatrix},V\right)+\langle f_-v_-\rangle\hat{\Delta}\left(\begin{pmatrix}v_+\\g_-\end{pmatrix},U\right)\Bigg\}.
\end{aligned}\tag{4.9}
$$

Interchanging U and V on the lhs of Eq. (4.9) gives an additional $(-)$sign, whereas on the rhs we get

$$
\begin{aligned}
&\langle f_-g_-\rangle\Bigg\{\langle g_+v_+\rangle\Delta_{2n+2}\left(\begin{pmatrix}f_+\\v_-\end{pmatrix},U\right)-\langle g_+u_+\rangle\hat{\Delta}\left(\begin{pmatrix}f_+\\u_-\end{pmatrix},V\right)-\langle f_+v_+\rangle\Delta_{2n+2}\left(\begin{pmatrix}g_+\\v_-\end{pmatrix},U\right)\\
&\quad+\langle f_+u_+\rangle\hat{\Delta}\left(\begin{pmatrix}g_+\\u_-\end{pmatrix},V\right)\Bigg\}+\langle f_+g_+\rangle\Bigg\{\langle g_-v_-\rangle\Delta_{2n+2}\left(\begin{pmatrix}v_+\\f_-\end{pmatrix},U\right)-\langle g_-u_-\rangle\hat{\Delta}\left(\begin{pmatrix}u_+\\f_-\end{pmatrix},V\right)\\
&\quad-\langle f_-v_-\rangle\Delta_{2n+2}\left(\begin{pmatrix}v_+\\g_-\end{pmatrix},U\right)+\langle f_-u_-\rangle\hat{\Delta}\left(\begin{pmatrix}u_+\\g_-\end{pmatrix},V\right)\Bigg\}.
\end{aligned}
$$

This can only be true if

$$\hat{\Delta}(F,G)\equiv\Delta_{2n+2}(F,G). \tag{4.10}$$

This proves (b) for the case $m=n+1$. The antisymmetry of Δ_{2n+2} is a consequence of (b) and can be seen as follows:

$$\begin{aligned}\Delta_{2n+2}(F,G)&=[\Phi_{2(n-k)+1}(F),\,\Phi_{2k+1}(G)]\\ \text{but because of (b) we also have}&=[\Phi_{2k+1}(F),\Phi_{2(n-k)+1}(G)]\\ &=-[\Phi_{2(n-k)+1}(G),\,\Phi_{2k+1}(F)]\\ &=-\Delta_{2n+2}(G,F).\end{aligned} \tag{4.11}$$

This shows that the inductive definition of the two sequences Φ_{2k-1}, resp. Δ_{2k}, $k\in\mathbf{N}$, works and gives all the properties as stated in Lemma 4.

Remark: In proving Eq. (4.9) we have shown as a side result

$$\begin{aligned}[\Delta_{2k}(V,U),\Delta_{2l}(F,G)]=\langle f_-g_-\rangle\Big\{&\langle g_+u_+\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}f_+\\u_-\end{pmatrix},V\right)-\langle g_+v_+\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}f_+\\v_-\end{pmatrix},U\right)\\ &-\langle f_+u_+\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}g_+\\u_-\end{pmatrix},V\right)+\langle f_+v_+\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}g_+\\v_-\end{pmatrix},U\right)\Big\}\\ &+\langle f_+g_+\rangle\Big\{\langle g_-u_-\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}u_+\\f_-\end{pmatrix},V\right)-\langle g_-v_-\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}v_+\\f_-\end{pmatrix},U\right)\\ &-\langle f_-u_-\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}u_+\\g_-\end{pmatrix},V\right)+\langle f_-v_-\rangle\Delta_{2(k+l)}\left(\begin{pmatrix}v_+\\g_-\end{pmatrix},U\right)\Big\}.\end{aligned} \tag{4.12}$$

At first glance this looks quite asymmetric with respect to an interchange of (V,U) and (F,G). But this asymmetry disappears if we express everything in the bilocal operators $B_{2i}^{(\pm)}$

$$\begin{aligned}[B_{2k}^{(+)}(V,U),\,B_{2l}^{(+)}(F,G)]=&\langle g_+u_+\rangle B_{2(k+l)}^{(+)}(f_+,v_+)-\langle g_+v_+\rangle B_{2(k+l)}^{(+)}(f_+,u_+)\\ &-\langle f_+u_+\rangle B_{2(k+l)}^{(+)}(g_+,v_+)+\langle f_+v_+\rangle B_{2(k+l)}^{(+)}(g_+,u_+),\end{aligned} \tag{4.13}$$

and a similar relation for $B_{2i}^{(-)}$.

Remark: From the definitions of the bilocal operators $B_{2k}^{(\pm)}$ and the field operators ϕ_{2k+1} it is clear that all multiple commutators containing $2k$, resp. $2k+1$, $k\in\mathbf{N}$, fields Φ are linear in $B_{2k}^{(\pm)}$, resp. Φ_{2k+1}, e.g.,

$$[\Phi(F),\,\Phi(G)]=\langle f_-g_-\rangle B_2^{(+)}(f_+,g_+)+\langle f_+g_+\rangle B_2^{(-)}(f_-,g_-)$$

$$[[\Phi(F),\,\Phi(G)]\Phi(U)]=\langle f_-g_-\rangle\langle g_+u_+\rangle\Phi_3\begin{pmatrix}f_+\\u_-\end{pmatrix}+\cdots\quad\text{(another 3 terms)}$$

$$[[[\Phi(F),\,\Phi(G)]\Phi(U)]\Phi(V)]=\langle f_-g_-\rangle\langle g_+u_+\rangle\langle u_-v_-\rangle B_4^{(+)}(f_+,v_+)+\cdots\quad\text{(another 7 terms)}$$

$$\begin{aligned}[[\Phi(F),\,\Phi(G)]\cdots\Phi(W)]=&\langle f_-g_-\rangle\langle g_+u_+\rangle\langle u_-v_-\rangle\langle v_+w_+\rangle\Phi_5\begin{pmatrix}f_+\\w_-\end{pmatrix}\\ &+\cdots\quad\text{(another 15 terms)}.\end{aligned}$$

V. THE TRUNCATED *n*-POINT FUNCTIONS

In this section we shall show that all truncated *n*-point functions are fixed already by the sequences of field operators Φ_{2k-1} and of commutator functions Δ_{2k}. The idea is the following:

Let d be the scaling dimension of the field Φ under consideration. By assumption (**A**) d is an odd number. Let $\varphi(x)=\psi_{d/2}(x_+)\otimes\psi_{d/2}(x_-)$ as defined in Sec. II (Buchholz' example).

Up to a factor, Lorentz invariance and scaling covariance together determine the one- and two-point functions uniquely. But the scalar fields Φ_{2k-1}, $k\in\mathbf{R}$, and φ all have the scaling dimension d and therefore it is clear that for all $k\in\mathbf{N}$ we have

$$(\Omega,\Phi_{2k-1}(F)\Omega)=0 \quad \text{(one-point function)}, \tag{5.1}$$

$$(\Omega,\Delta_{2k}(F,G)\Omega)=c_{2k}(\Omega_0,[\varphi(F),\varphi(G)]\Omega_0) \quad \text{(two-point function)}, \tag{5.2}$$

and because of the special form of Δ_{2k} as given by (4.1), Eq. (5.2) is equivalent to

$$(\Omega,B_{2k}^{(\pm)}(f,g)\Omega)=c_{2k}(\Omega_0,\tfrac{1}{2}[\psi_{d/2}(f),\psi_{d/2}(g)]\Omega_0). \tag{5.3}$$

This fixes a sequence c_{2k}, $k\in\mathbf{N}$, of real numbers.

Let W_n, resp. V_n, denote the n-point functions of the fields Φ, resp. φ.

Lemma 5: For the truncated n-point functions we have

$$W_{2n-1}^T\equiv 0 \quad \text{and} \quad W_{2n}^T=c_{2n}V_{2n}^T, \tag{5.4}$$

and all c_{2n} are positive.

Proof: (a) The vacuum expectation value of a $2n$-fold commutator vanishes because, as shown in the last section, this multiple commutator is linear in the field Φ_{2n-1} and $(\Omega,\Phi_{2n-1}(F)\Omega)=0$.

On the other hand, the vacuum expectation value of a multiple commutator containing an even number of fields is given by

$$\begin{aligned}(\Omega,&\underbrace{[\ldots[}_{2n-1}\Phi(F^{(1)}),\Phi(F^{(2)})],\ldots,\Phi(F^{(2n)})]\Omega)\\ &=\langle f_-^{(1)}f_-^{(2)}\rangle\langle f_+^{(2)}f_+^{(3)}\rangle\cdots\langle f_-^{(2n-1)}f_-^{(2n)}\rangle(\Omega,B_{2n}^{(+)}(f_+^{(1)},f_+^{(2n)})\Omega)+\cdots(2^n-1)\ \text{further terms}\\ &=\langle f_-^{(1)}f_-^{(2)}\rangle\cdots\langle f_-^{(2n-1)}f_-^{(2n)}\rangle c_{2n}\left(\Omega,\frac{1}{2}[\psi_{d/2}(f_+^{(1)}),\psi_{d/2}(f_+^{(2n)})]\Omega\right)+\cdots(2^n-1)\ \text{further terms}\\ &=c_{2n}(\Omega_0,[\cdots[\varphi(F^{(1)}),\varphi(F^{(2)})],\cdots,\varphi(F^{(2n)})]\Omega_0),\end{aligned} \tag{5.5}$$

i.e., all the n-fold commutators of the fields Φ and $\varphi=\psi_{d/2}\otimes\psi_{d/2}$ are linearly related.

(b) By the very definition of the truncated n-point functions this implies

$$W_{2n}^T([\cdots[F^{(1)},F^{(2)}],\cdots,F^{(2n)}])=c_{2n}V_{2n}^T([\cdots[F^{(1)},F^{(2)}],\cdots,F^{(2n)}]). \tag{5.6}$$

As a consequence of the spectrum condition this linear relationship, valid for all $(2n-1)$-fold commutators, can be extended to $W_{2n}^T=c_{2n}V_{2n}^T$. To do this in a systematic way we write

$$W_{2n}^T(F_1,\ldots,F_{2n})=c_{2n}V_{2n}^T(F_1,\ldots,F_{2n})+R_{2n}^T(F_1,\ldots,F_{2n}) \tag{5.7}$$

and R_{2n}^T fulfills again the spectrum condition. From Eq. (5.6) we already know

$$R_{2n}^T([\ldots[F_1,F_2],\ldots,F_{2n}])=0 \tag{5.8}$$

and we have to show

$$R_{2n}^T(F_1,\ldots,F_{2n})\equiv 0. \tag{5.9}$$

We shall exemplify this for the case of the truncated two- and four-point functions:

For the test function $F\in\mathcal{S}(\mathbf{R}^2)$ we denote by F^+ the part of F which in momentum space is supported in $\overline{V^+}\backslash\{0\}$, i.e., ''closed forward cone with the origin exempted.'' F^- is defined analogously as the part of F which in momentum space is supported in $\overline{V^-}\backslash\{0\}$, i.e., ''closed backward cone with the origin exempted.'' By the spectrum condition we have $\Phi(F)\Omega=\Phi(F^+)\Omega$ and $\Phi(F^-)\Omega=0$.

We begin with the two-point function. We know $R_2^T([F,G])=0$:

$$R_2^T(F,G) \overset{sp.c.}{=} R_2^T(F,G^+)=R_2^T([F,G^+])+\underbrace{R_2^T(G^+,F)}_{=0}=0 \tag{5.10}$$

Next we consider the four-point function. We know

$$R_4^T([[[F_1,F_2]F_3]F_4])=0 \quad \text{and} \quad R_4^T([[F_1,F_2][F_3,F_4]])=0. \tag{5.11}$$

where the second equation follows from the Jacobi identity.

(i)

$$\begin{aligned} R_4^T([[F_1,F_2]F_3]F_4) &\overset{sp.c.}{=} R_4^T([[F_1,F_2]F_3]F_4^+) \\ &= R_4^T([[[F_1,F_2]F_3]F_4^+])+R_4^T(F_4^+[[F_1,F_2]F_3]) \\ &= 0 \text{ by Eq. (5.11) and spectrum condition,} \end{aligned} \tag{5.12}$$

and similarly also $R_4^T(F_1[[F_2,F_3]F_4])=0$.

(ii)

$$\begin{aligned} R_4^T([F_1,F_2][F_3,F_4]) &\overset{sp.c.}{=} R_4^T([F_1,F_2][F_3,F_4]^+) \\ &= R_4^T([[F_1,F_2],[F_3,F_4]^+])+R_4^T([F_3,F_4]^+[F_1,F_2]) \\ &= 0 \text{ by Eq. (5.11) and spectrum condition.} \end{aligned} \tag{5.13}$$

(iii)

$$\begin{aligned} R_4^T([F_1,F_2]F_3F_4) &\overset{sp.c.}{=} R_4^T([F_1,F_2]F_3F_4^+) \\ &= \underbrace{R_4^T([F_1,F_2][F_3,F_4^+])}_{=0 \text{ by Eq. (5.13)}}+R_4^T([F_1,F_2]F_4^+F_3) \\ &= R_4^T([[F_1,F_2]F_4^+]F_3)+R_4^T(F_4^+[F_1,F_2]F_3) \\ &= 0 \text{ by Eq. (5.12) and spectrum condition} \end{aligned} \tag{5.14}$$

and similarly also $R_4^T(F_1F_2[F_3,F_4])=0$.

(iv)

$$\begin{aligned} R_4^T(F_1[F_2,F_3]F_4) &\overset{sp.c.}{=} R_4^T(F_1[F_2,F_3]F_4^+) \\ &= R_4^T(F_1[[F_2,F_3]F_4^+])+R_4^T(F_1F_4^+[F_2,F_3]) \text{ by Eq. (5.12)} \\ &= R_4^T([F_1F_4^+][F_2,F_3])+R_4^T(F_4^+F_1[F_2,F_3]) \\ &= 0 \quad \text{by Eq. (5.13) and spectrum condition.} \end{aligned} \tag{5.15}$$

(v) From (iii) and (iv) it follows

$$R_4^T(F_1F_2F_3F_4) \overset{sp.c.}{=} R_4^T(F_1F_2F_3F_4^+)$$
$$=R_4^T(F_1F_2[F_3,F_4^+])+R_4^T(F_1[F_2F_4^+]F_3)+R_4^T([F_1,F_4^+]F_2F_3)$$
$$+R_4^T(F_4^+F_1F_2F_3)$$
$$=0 \text{ which proves our claim.} \tag{5.16}$$

The proof of $R_{2n}^T(F_1,\ldots,F_{2n})\equiv 0$ for $n>2$ is tedious but by no means more complicated!

From $W_{2n}^T=c_{2n}V_{2n}^T$ we get immediately $c_{2n}\geqslant 0$ because of positivity. If for some $n>1$ we have $c_{2n}=0$, then all $c_{2n}\equiv 0$ except possibly c_2 as well known. But then c_2 has to be 0, too, because otherwise Φ is a free field and this contradicts the assumption of boundedness! So we have shown Lemma 5.

VI. REDUCTION TO *s*-PRODUCTS?

Up to now we have shown that our field Φ with scaling dimension $d\in\mathbf{N}$, d odd is related to $\varphi=\psi_{d/2}\otimes\psi_{d/2}$ via

$$W_{2n}^T=c_{2n}V_{2n}^T, \quad W_{2n-1}^T\equiv 0\equiv V_{2n-1}^T,$$

where W_n, resp. V_n, denote the Wightman functions corresponding to Φ, resp. φ. This looks quite similar to an s-product of φ, but up to now we are not able to prove this conjecture. If Φ were an s-product of φ, then we would have

$$c_{2n}=\sum_{k=1}^{K}\alpha_k^{2n} \quad \text{if } \Phi=\varphi_{\underset{\sim}{\alpha}}=(\alpha_1\varphi)s\cdots s(\alpha_K\varphi). \tag{6.1}$$

We are not able to determine a sequence $\underset{\sim}{\alpha}$ such that $\Phi=\varphi_{\underset{\sim}{\alpha}}$ but at least we can find a function $\rho(\alpha)$ such that $c_{2n}=\int_{\mathbf{R}}\alpha^{2n-1}\rho(\alpha)\,d\alpha$.

The construction goes as follows:

For the special real test functions $F=f_+\otimes f_-$ and $G=g_+\otimes g_-$ s.t. $g_-\equiv f_-$, $\langle f_+g_+\rangle=0$ and $\langle f_-^2\rangle=1=\langle g_+^2\rangle$, we get

$$e^{-i\gamma\Phi(G)}\Phi(F)e^{i\gamma\Phi(G)}=\Phi(F)+\sum_{n=1}^{\infty}\frac{(i\gamma)^n}{n!}[\ldots[\Phi(F),\underbrace{\Phi(G)]\ldots,\Phi(G)}_{n}]$$
$$=\sum_{n=0}^{\infty}(-1)^n\frac{\gamma^{2n}}{(2n)!}\Phi_{2n+1}(F)$$
$$+i\sum_{n=0}^{\infty}(-1)^n\frac{\gamma^{2n+1}}{(2n+1)}B_{2n+2}^{(+)}(f_+,g_+). \tag{6.2}$$

Taking the vacuum expectation value

$$(\Omega, e^{-i\gamma\Phi(G)}\Phi(F)e^{i\gamma\Phi(G)}\Omega) = i\sum_{n=0}^{\infty}(-1)^n\frac{\gamma^{2n+1}}{(2n+1)!}(\Omega, B_{2n+2}^{(+)}(f_+,g_+)\Omega)$$

$$= \underbrace{i\sum_{n=0}^{\infty}(-1)^n\frac{\gamma^{2n+1}}{(2n+1)!}c_{2n+2}}_{\sigma(\gamma)}\left\langle\frac{1}{2}[\psi_{d/2}(f_+),\psi_{d/2}(g_+)]\right\rangle$$

$$= \sigma(\gamma)\left\langle\frac{1}{2}[\psi_{d/2}(f_+),\psi_{d/2}(g_+)]\right\rangle \tag{6.3}$$

defines the function $\sigma(\gamma)$.

Its Fourier transformed $\rho(\alpha) = \int_{\mathbf{R}} e^{-i\alpha\gamma}\sigma(\gamma)d\gamma/2\pi$ has the property

$$\int_{-\infty}^{\infty}\alpha^{2n-1}\rho(\alpha)\,d\alpha = \int_{-\infty}^{\infty}\alpha^{2n-1}\left(\int_{-\infty}^{\infty}e^{-i\alpha\gamma}\sigma(\gamma)\frac{d\gamma}{2\pi}\right)d\alpha$$

interchanging $d\alpha$ and $d\gamma$

$$= \int_{-\infty}^{\infty}\sigma(\gamma)\underbrace{\int_{-\infty}^{\infty}\alpha^{2n-1}e^{-i\alpha\gamma}\frac{d\alpha}{2\pi}}_{(i\partial_\gamma)^{2n-1}\delta(\gamma)}d\gamma$$

$$= i(-1)^n\sigma^{(2n-1)}(0) = c_{2n}. \tag{6.4}$$

For the s-product $\Phi = \varphi_{\underline{\alpha}}$ and the above special test functions F and G we get

$$e^{-i\gamma\Phi(G)}\Phi(F)e^{i\gamma\Phi(G)} = \varphi_{\underline{\alpha\cos(\gamma\alpha)}}(F) + i[\varphi(F),\varphi(G)]_{\underline{\alpha\sin(\gamma\alpha)}}, \tag{6.5}$$

the vacuum expectation value of which defines the function

$$\sigma(\gamma) = i\sum_{k=1}^{K}\alpha_k\sin(\gamma\alpha_k) \tag{6.6}$$

and obviously its Fourier transformed

$$\rho(\alpha) = \sum_{k=1}^{K}\frac{\alpha_k}{2}[\delta(\alpha-\alpha_k) - \delta(\alpha+\alpha_k)] \tag{6.7}$$

fulfills Eq. (6.1) for the c_{2n} and allows us to reconstruct the sequence $\underline{\alpha}$. Each α_k is determined up to a factor ± 1. This nonuniqueness corresponds to a unitary transformation. Positivity imposes strong restrictions on the weight functions $\rho(\alpha)$, e.g., there are convincing arguments that $\rho(\alpha)$ cannot have a continuous part.

VII. OUTLOOK

Let us make some comments on the work we have done:

(1) We can only conjecture, but we have not fully succeeded in proving it, that every field Φ fulfilling our assumptions **A** with scaling dimension $d \in \mathbf{N}$, d odd, is an s-product, eventually with infinitely many components, of the corresponding Buchholz field $\phi = \psi_{d/2} \otimes \psi_{d/2}$.

(2) What happens if the scaling dimension of Φ is an even natural number? We expect, in this case, that bounded fields do not exist.

For the simplest case, $d = 2$, the commutator looks like

$$[\Phi(x),\Phi(y)]=\delta'(y_- - x_-)B^{(+)}(x_+,y_+)+\delta'(y_+ - x_+)B^{(-)}(x_-,y_-)$$
$$+\delta(y_- - x_-)B_0^{(+)}\left(x_+,y_+,\frac{x_-+y_-}{2}\right)+\delta(y_+ - x_+)B_0^{(-)}\left(x_-,y_-,\frac{x_++y_+}{2}\right), \tag{7.1}$$

and we are not able to get rid of the second line, at least not with the methods used in Sec. III. The same difficulty appears for higher even d's. There are always two equations missing!

Nevertheless, let us explain why we think there is no bounded Bose field of dimension $d=2$ fulfilling our assumptions. Assume the second line in Eq. (7.1) to be absent. As in Sec. III we get

$$[[\Phi(x),\Phi(y)]\Phi(z)]=\delta'(y_- - x_-)\left\{\delta'(z_+ - y_+)\Phi_3\begin{pmatrix}x_+\\ z_-\end{pmatrix}+\delta'(z_+ - x_+)\Phi_3\begin{pmatrix}y_+\\ z_-\end{pmatrix}\right\}$$
$$+\delta'(y_+ - x_+)\left\{\delta'(z_- - y_-)\Phi_3\begin{pmatrix}z_+\\ x_-\end{pmatrix}+\delta'(z_- - x_-)\Phi_3\begin{pmatrix}z_+\\ y_-\end{pmatrix}\right\} \tag{7.2}$$

and so on. As a consequence, the Wightman functions have the following structure,

$$W^T_{2n-1}\equiv 0 \quad \text{and} \quad W^T_{2n}=c_{2n}V^T_{2n},$$

and all the c_{2n} are positive. The Wightman functions V_n correspond now to the field $\varphi(x)=\varphi_1(x_+)\otimes\varphi_1(x_-)$, where φ_1 is the free chiral Bose field of scaling dimension 1. But for a real factorized test function $f=f_+\otimes f_-$ we have

$$V^T_{2n}(f,\ldots,f)=(2n-1)!\,V_2(f,f)>0 \quad \text{and therefore} \quad W^T_{2n}(f,\ldots,f)>0,$$

too. This is incompatible with the boundedness of $\Phi(f)$ as can be easily seen from the characteristic functional

$$E(f)=(\Omega,e^{i\Phi(f)}\Omega)=\exp\sum_{n\geqslant 1}\frac{(-1)^n}{2n!}W^T_{2n}(f,\ldots,f), \tag{7.3}$$

because $|E(i\lambda f)|$ is not bounded by $\exp\{|\lambda|\cdot\|\Phi(f)\|\}$. This is no proof, but just a strong and plausible hint because we have neglected the second line in Eq. (7.1)! But perhaps these additional terms do not disturb the behavior too much.

(3) An important generalization in our opinion would be to give up strict covariance under dilations. But this step requires new ideas to prove an analogon to our theorem.

ACKNOWLEDGMENTS

Many thanks to Heide Narnhofer and Jakob Yngvason for inviting me to the Erwin Schrödinger Institut at Vienna and their hospitality. Most of the work has been done there. J. Yngvason and K.-H. Rehren contributed good ideas and I had interesting and helpful discussions with them. B. Lütge-Hampe gave many useful hints for typesetting the LATEX manuscript.

[1] R. F. Streater and A. S. Wightman, *PCT, SPIN AND STATISTICS, AND ALL THAT* (Benjamin/Cummings, Reading, MA, 1980).
[2] D. Buchholz, private communication, unpublished.
[3] K.-H. Rehren, ''Bounded Bose Fields,'' Lett. Math. Phys. **40**, 299–306 (1997).
[4] G. Hegerfeldt, ''Prime Field Decompositions and Infinitely Divisible States on Borchers' Tensor Algebra,'' Commun. Math. Phys. **45**, 137–151 (1975).
[5] K. Baumann, ''Unboundedness of Bose Fields in a One-Dimensional Model World,'' Lett. Math. Phys. **41**, 135–147 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

The Floquet analysis and noninteger higher harmonics generation

José M. Cerveró[a)]
Física Teórica, Facultad de Ciencias, Universidad de Salamanca, 37008 Salamanca, Spain

Juan D. Lejarreta
Escuela Técnica Superior de Ingeniería Industrial, Universidad de Salamanca, 37700 Béjar, Spain

(Received 7 July 1998; accepted for publication 10 December 1998)

We consider here the exact solution of a nonrelativistic quantum system composed of a two-level atom interacting with a laser with arbitrary large frequency and intensity. We use the analogy of this system and a $\hbar/2$-spin particle interacting with a time-dependent magnetic field. A systematic use of the dynamical symmetry underlying the physical system is made. Actually the Hamiltonian is a Hermitian element of the SU(2) Lie Algebra. The exact Temporal Evolution Operator in terms of a Generalized Displacement Operator of the group is constructed. It is possible to develop a nonperturbative method that allows us to solve exactly the model for any value of the relevant frequencies (Rabi, Laser, and Atom Frequencies) and in so doing the usual Rotating Wave and Small Interaction approximations are unnecessary. The properties of periodicity of this model and the phenomenon of harmonic generation are considered by using Floquet Analysis. We find that in addition to the so far well-known spectrum composed by odd harmonics, this model generates another type of noninteger harmonics whose frequencies and amplitudes are determined for any value of the relevant parameters of the system. © *1999 American Institute of Physics.* [S0022-2488(99)02704-8]

I. INTRODUCTION

The monochromatic electric field of a very intense laser is the source of a variety of nonlinear responses as it interacts with the matter fields, usually two- or three-level atoms. A particular area of interest in the last ten years has been the so-called higher harmonics generation: the appearance of optical harmonics of various frequencies close to that of the laser pump. These harmonics usually manifest themselves as radiation of the atoms with frequencies that are integer multiples of the incident field. This effect has been thought to be extremely useful to generate further laser beams of a shorter wavelength.[1] The spectrum of such outgoing radiation is given by the Fourier transform of the atom dipole moment, which encompasses the available dipolar oscillations among which the energy can be interchanged. The ultimate reason for the appearance of these harmonics is obviously the nonlinear response of the individual atoms to a strong pumping field. An adequate theoretical understanding requires *the description of the interaction of the atom with the strong external field*. This has been usually achieved by solving the Schrödinger Equation (2), with a time-dependent oscillating potential. When the pumping field is not very strong one can use safely conventional perturbation theory. However, as the intensity of the laser increases, perturbative methods are no longer reliable and a large body of literature has appeared to deal with the problem of interaction of the atom with very strong fields. The paper by Shirley[3] has always been considered the first step toward a reliable theory of atom–laser interaction for arbitrary large electric and magnetic fields. This approach was followed shortly after by Zel'dovich.[4] A recent paper by

[a)]Electronic mail: cervero@rs6000.usal.es

Compagno, Dietz, and Persico[5] contains a set of recent references in the subject and we address the interested reader to this paper for a complete account of the current literature on the field of strongly driven atoms. As a consequence of this interaction, a large variety of methods has been used to actually solve the equations leading to a more or less accurate prediction of higher harmonics generation. Numerical solutions of the Schrödinger Equation,[2] time-dependent (Hartree–Fock) iterative methods[1,6] and treatments involving the analysis of two-level atoms.[7,8] As we shall heavily rely in this paper on various aspects of Floquet Analysis it is necessary to mention that Floquet Theory has also been recently applied to high-order harmonic generation with the same aim than ours but with different perspective. (See Refs. 9–11.)

However, without leaving the paradigm of the two-level atom, the properties of the SU(2) dynamical symmetry of the system can be used to build the exact Temporal Evolution Operator (TEO),[12] that in our opinion becomes the core of the problem, as this operator determines with the appropriated initial conditions on the state vector the instantaneous state of the atom and all its relevant physical properties. Moreover, this method has an additional advantage, namely, that it does not require any kind of approximate treatment such as the Rotating Wave, Weak Field, or Adiabatical Approximations.

This paper is organized as follows: In Sec. II we describe in detail the main features of the model: the relevant mathematical properties of its dynamical symmetry and the specific application of the method to this physical problem in order to build the exact Temporal Evolution Operator. To identify this operator among the elements of the group one has to solve a first-order Riccati differential equation that can easily be linearized. Some transformations are considered in Sec. III, where an Invariant for the system is built and an exact solution in terms of a Taylor series is also found. In Sec. IV the relevant physical properties of the atom are described and compared to those of the Rotating Wave Approximation. We discuss in Secs. V and VI the properties of periodicity of this system by using Floquet Analysis. With these results in hand we can justify in Sec. VII the composition of the harmonic spectrum of the instantaneous electric dipole moment of the atom that is described completely, and both the resulting frequencies and amplitudes of the generated harmonics are discussed. Finally, Sec. VIII is one of conclusions and future prospects in pursuing this line of research.

II. THE PHYSICAL SYSTEM AND ITS DYNAMICAL SYMMETRY

Let us consider the physical system containing a *two-level atom* and its electric dipole interaction with the coherent field of a laser. The external monochromatic field of frequency ω will be treated classically and the possible amplification effects of the laser or spontaneous emission of the atom will be ignored. The system can be described by the Hamiltonian,[13]

$$H(t)=\hbar\omega_0|b\rangle\langle b|+\hbar\Omega_0 s(t)\cos(\omega t)[|a\rangle\langle b|+|b\rangle\langle a|], \tag{2.1}$$

where $|a\rangle$ and $|b\rangle$ are the atomic states whose energy is 0 and $\hbar\omega_0$. Also, $\hbar\Omega_0$ is the energy of the interaction and Ω_0 the *Rabi frequency*. Here $s(t)$ denotes a dimensionless function of time that describes a laser pulse. The introduction of the operators $J_0=J_0^+$ and $J_+^+=J_-$,

$$J_0=\tfrac{1}{2}\{|b\rangle\langle b|-|a\rangle\langle a|\},\quad J_-=|a\rangle\langle b|,\quad J_+=|b\rangle\langle a|, \tag{2.2}$$

with the commutation relations $[J_0,J_\pm]=\pm J_\pm$ and $[J_+,J_-]=2J_0$ allows the Hamiltonian to be expressed as

$$H(t)=\hbar\omega_0[\tfrac{1}{2}+J_0]+\hbar\Omega_0 s(t)\cos(\omega t)[J_+ + J_-], \tag{2.3}$$

and one can identify this system with a j-spin particle interacting with a time-dependent magnetic field. When the dynamical symmetry of this Hamiltonian is considered, $H(t)$ can be seen as a Hermitian element of the triparametric Lie-Algebra SU(2), whose properties[14] will be considered in the full solution of the problem. The physical states (rays) of the atom can be obtained by acting on the eigenstates of J_0 with the *Unitary Displacement Operator* $S(\beta)$ or $S(\eta)$ whose fragmentation properties are[15]

$$S(\beta)=\exp\{\beta J_+-\beta^* J_-\}=\exp\{\eta J_+\}\exp\{\delta J_0\}\exp\{-\eta^* J_-\}=S(\eta), \tag{2.4}$$

$$\beta=re^{i\phi}, \quad \eta=\tan re^{i\phi}, \quad \delta=\log\{1+|\eta|^2\}. \tag{2.5}$$

Every Displacement Operator generates a unitary transformation of the algebra,

$$S(\eta)J_+S^+(\eta)=\frac{1}{1+|\eta|^2}[J_++2\eta^* J_0-\eta^{*2}J_-], \tag{2.6}$$

$$S(\eta)J_0S^+(\eta)=\frac{1}{1+|\eta|^2}[-\eta J_++(1-|\eta|^2)J_0-\eta^* J_-], \tag{2.7}$$

$$S(\eta)J_-S^+(\eta)=\frac{1}{1+|\eta|^2}[-\eta^2 J_++2\eta J_0+J_-], \tag{2.8}$$

and a measure of the change with time of this operator can be given by the quantity

$$\dot{S}(\eta)S^+(\eta)=\frac{1}{1+|\eta|^2}[\dot{\eta}J_++(\dot{\eta}\eta^*-\eta\dot{\eta}^*)J_0-\dot{\eta}^* J_-]. \tag{2.9}$$

The explicit form of the $H(t)$ makes no reference to any specific representation of the algebra, and therefore all properties of the physical system determined by the dynamical symmetry can be extended to any representation. In particular, we can generalize the treatment of the problem to describe an *m-level-atom* interaction with either Equidistant Levels or with an Electric Dipole moment operator connecting only contiguous levels.

III. THE TEMPORAL EVOLUTION OPERATOR

The Temporal Evolution Operator that satisfies the Schrödinger equation,

$$i\hbar\dot{U}(t)U^+(t)=H(t), \quad U(0)=1, \tag{3.1}$$

can be obtained by mapping exponentially a Hermitian element of the algebra. We express this element by a Displacement Operator with arbitrary factorization as

$$U(t)=\exp\left\{-i\frac{\omega_0}{2}t\right\}S[\eta(t)]\exp\{ih(t)J_0\}, \tag{3.2}$$

and as a consequence of (3.1) and the properties of the algebra the following system of equations must hold:

$$-i\Omega_0 s(t)\cos(\omega t)(1+|\eta|^2)=\dot{\eta}-i\dot{h}\eta, \tag{3.3}$$

$$-i\omega_0(1+|\eta|^2)=\dot{\eta}\eta^*-\eta\dot{\eta}^*+i\dot{h}(1-|\eta|^2). \tag{3.4}$$

Finally, it is easy to see that the explicit identification of the TEO requires the knowledge of two functions: $h(t)$ real and $\eta(t)$ complex, whose evolution imposed by the Schrödinger equation is governed by the system of ordinary differential equations:

$$\dot{\eta}=-i\Omega_0 s(t)\cos(\omega t)(1-\eta^2)-i\omega_0\eta, \quad \eta(0)=0, \tag{3.5}$$

$$h(t) = -\omega_0 t + 2\Omega_0 \int_0^t s(u)\cos(\omega u)\mathrm{Re}[\eta(u)]du. \tag{3.6}$$

The nonlinear first-order Riccati equation is the heart of the problem. If we know the complex function $\eta(t)$, the solution of this Riccati equation, we can obtain the real function $h(t)$ by a quadrature, and in terms of these functions all the remaining relevant properties of the system. This equation is exactly solvable in the following particular cases: (i) $\eta = -i\tan\{(\Omega_0/\omega)\sin(\omega t)\}$ if the energy of the atom is negligible compared to the energy of the interaction ($\omega_0 \to 0$); (ii) $\eta = 0$ in the reciprocal case ($\Omega_0 \to 0$); and (iii) in the rotating wave approximation that will be considered later.
In the general case and with $\omega_0 \neq 0$ we consider the following transformations

(i) Conformal transformation:

$$\eta = \frac{1-\xi(t)}{1+\xi(t)}, \tag{3.7}$$

$$\dot{\xi} = 2i\Omega_0 s(t)\cos(\omega t)\xi + i\frac{\omega_0}{2}(1-\xi^2), \quad \xi(0) = 1. \tag{3.8}$$

The evolution of the new function $\xi(t)$ is governed by another Riccati equation with a constant coefficient for the nonlinear term. Therefore our system is equivalent to an atom with a time-dependent oscillating *energy*, interacting with an external stationary electric field.

(ii) Linearization of the Riccati equation: The new nonlinear Riccati equation can be linearized by introducing a new complex function $q(t)$:

$$\xi = -\frac{2i}{\omega_0}\frac{\dot{q}}{q}, \tag{3.9}$$

which must be a solution of the *second-order linear differential equation*,

$$\ddot{q} - 2i\Omega_0 s(t)\cos(\omega t)\dot{q} + \frac{\omega_0^2}{4}q = 0, \tag{3.10}$$

$$q(0) = 1, \quad \dot{q}(0) = i\frac{\omega_0}{2}. \tag{3.11}$$

This last differential equation shows the equivalence between our system and a harmonic oscillator subjected to viscous damping with an imaginary time-dependent coefficient. Introducing new variables given by $q = re^{i\phi}$ and $p = r^2\dot{\phi}$, we can construct an *invariant* of this system in the form

$$\dot{r}^2 + \frac{p^2}{r^2} + \frac{\omega_0^2}{4}r^2 = \frac{\omega_0^2}{2}, \tag{3.12}$$

$$\dot{p} = 2s(t)\frac{\Omega_0}{\omega_0}\cos(\omega t)r\dot{r}. \tag{3.13}$$

This invariant is reminiscent of the total energy of an isotropic two-dimensional oscillator. One should emphasize that the existence of the invariant (3.12) is independent of the function $s(t)$ that governs the shape of the incident pulse of the external field. The analogy cannot be pushed too far because in this system the ''angular momentum'' is not a conserved quantity.

If we introduce $\gamma = \Omega_0/\omega$, $\epsilon = \omega_0/\omega$, $x = \omega t$, and we use *primes* to denote d/dx, the new form of the linear equation (3.10) is

TABLE I. Taylor coefficients for $u(x)$.

Order	$u(x)$
0	1
1	0
2	$-\frac{\epsilon^2}{8}$
3	$-i\frac{\gamma\epsilon^2}{12}$
4	$-\frac{\epsilon^2}{384}[16\gamma^2+\epsilon^2]$
5	$i\frac{\gamma\epsilon^2}{480}[6+8\gamma^2+\epsilon^2]$
6	$-\frac{\epsilon^2}{46\,080}[576\gamma^2+256\gamma^4+48\epsilon^2\gamma^2+\epsilon^4]$
7	$-i\frac{\epsilon^2\gamma}{161\,280}[80+52\epsilon^2+3\epsilon^4+1216\gamma^2+64\epsilon^2\gamma^2+256\gamma^4]$
8	$\frac{\epsilon^2}{10\,321\,920}[\epsilon^6+16\,640\gamma^2+3328\epsilon^2\gamma^2+96\epsilon^4\gamma^2+34\,816\gamma^4+1280\epsilon^2\gamma^4+4096\gamma^6]$
9	$i\frac{\epsilon^2\gamma}{11\,612\,160}[112+160\epsilon^2+34\epsilon^4+\epsilon^6+18\,496\gamma^2+2144\epsilon^2\gamma^2+40\epsilon^4\gamma^2+14\,080\gamma^4+384\epsilon^2\gamma^4+1024\gamma^6]$
10	$-\frac{\epsilon^2}{3\,715\,891\,200}[\epsilon^8+394\,240\gamma^2+155\,904\epsilon^2\gamma^2+10\,880\epsilon^4\gamma^2+160\epsilon^6\gamma^2+3\,649\,536\gamma^4+286\,720\epsilon^2\gamma^4+3840\epsilon^4\gamma^4+1\,359\,872\gamma^6+28\,672\epsilon^2\gamma^6+65\,536\gamma^8]$

$$q''(x)-2i\gamma s(x)\cos x\, q'(x)+\frac{\epsilon^2}{4}q(x)=0, \tag{3.14}$$

$$q(0)=1, \quad q'(0)=i\frac{\epsilon}{2}, \tag{3.15}$$

which can be solved by applying the *Frobenius Theory* in terms of a *power series*,

$$u(x)=\sum_{k=0}^{\infty} A_k x^k, \quad v(x)=\sum_{k=0}^{\infty} B_k x^k, \quad A_0=B_1=1, \quad A_1=B_0=0, \tag{3.16}$$

whose coefficients A_k and B_k must satisfy the same following recurrence law:

$$A_{k+2}=\frac{1}{(k+2)(k+1)}\left[2i\gamma\sum_{r=0}^{k+1} a_{k+1-r}A_r r-\frac{\epsilon^2}{4}A_k\right], \tag{3.17a}$$

$$B_{k+2}=\frac{1}{(k+2)(k+1)}\left[2i\gamma\sum_{r=0}^{k+1} a_{k+1-r}B_r r-\frac{\epsilon^2}{4}B_k\right], \tag{3.17b}$$

where the a_j are the coefficients of the de Taylor series of the damping function $s(x)\cos x$ that in the case of a square pulse $s(x)=1$ are $a_{2k}=(-1)^k/(2k)!$, $a_{2k+1}=0$. Explicitly, the first terms of these series are collected in Tables I and II. An important consequence of this procedure is that all points of the real line are regular for the linear equation (3.14) that fully determines all physical properties of the model, and according to the Frobenius theorem the functions $u(x)$ and $v(x)$ satisfying $u(0)=\dot{v}(0)=1$ and $\dot{u}(0)=v(0)=0$ are *a fundamental system of solutions* of the linear differential equation. They *converge in the real line for all values of the parameters*,[16] although,

TABLE II. Taylor coefficients for $v(x)$.

Order	$v(x)$
0	0
1	1
2	$i\gamma$
3	$-\frac{1}{24}[16\gamma^2+\epsilon^2]$
4	$-i\frac{\gamma}{24}[2+8\gamma^2+\epsilon^2]$
5	$\frac{1}{1920}[256\gamma^4+49\gamma^2\epsilon^2+256\gamma^2+\epsilon^4]$
6	$i\frac{\gamma}{5760}[16+28\epsilon^2+3\epsilon^4+640\gamma^2+64\epsilon^2\gamma^2+256\gamma^4]$
7	$-\frac{1}{322\,560}[\epsilon^6+4096\gamma^2+1984\epsilon^2\gamma^2+96\epsilon^4\gamma^2+20\,480\gamma^4+1280\epsilon^2\gamma^4+4096\gamma^6]$
8	$-i\frac{\gamma}{322\,560}[16+64\epsilon^2+22\epsilon^4+\epsilon^6+5824\gamma^2+1376\epsilon^2\gamma^2+40\epsilon^4\gamma^2+8960\gamma^4+384\epsilon^2\gamma^4+1024\gamma^6]$
9	$\frac{1}{92\,897\,280}[\epsilon^8+65\,536\gamma^2+63\,744\epsilon^2\gamma^2+7424\epsilon^4\gamma^2+160\epsilon^6\gamma^2+1\,376\,256\gamma^4+194\,560\epsilon^2\gamma^4+3840\epsilon^4\gamma^4+917\,504\gamma^6+28\,672\epsilon^2\gamma^6+65\,536\gamma^8]$
10	$i\frac{\gamma}{464\,486\,400}[256+1856\epsilon^2+1376\epsilon^4+200\epsilon^6+5\epsilon^8+839\,680\gamma^2+363\,008\epsilon^2\gamma^2+23\,680\epsilon^4\gamma^2+320\epsilon^6\gamma^2+3\,956\,736\gamma^4+373\,760\epsilon^2\gamma^4+5376\epsilon^4\gamma^4+1\,376\,256\gamma^6+32\,768\epsilon^2\gamma^6+65\,536\gamma^8]$

in general, this convergence could be very slow. The method reveals itself as an intrinsically nonperturbative one, and it is valid regardless of the magnitude of the intensity of the coupling atom laser. In other words, the procedure is also applicable for arbitrarily intense strong fields. We can use these series to obtain $q(t)$ and $\eta(t)$ in the form

$$q(\omega t)=u(\omega t)+i\frac{\omega_0}{2}v(\omega t), \tag{3.18}$$

$$\eta(\omega t)=\frac{\omega_0 q(\omega t)+2i\dot{q}(\omega t)}{\omega_0 q(\omega t)-2i\dot{q}(\omega t)}, \tag{3.19}$$

which define the general solution.

IV. THE PHYSICAL OBSERVABLES AND THE ROTATING WAVE APPROXIMATION

Once both previous equations have been solved we immediately know $\eta(t)$, and we can obtain $h(t)$ through a quadrature as well as all other physical quantities of interest. For instance, the natural evolution of the eigenstates of the free atom is given by

$$U(t)|a\rangle=\frac{e^{-(i/2)[h+\omega_0 t]}}{\sqrt{1+|\eta|^2}}\{|a\rangle+\eta|b\rangle\}, \tag{4.1}$$

$$U(t)|b\rangle=\frac{e^{(i/2)[h-\omega_0 t]}}{\sqrt{1+|\eta|^2}}\{-\eta^*|a\rangle+|b\rangle\}. \tag{4.2}$$

The normalized transition probabilities between states is also given by

$$P_{ab}=P_{ba}=\frac{|\eta|^2}{1+|\eta|^2}=\sin^2|\beta|, \tag{4.3}$$

$$P_{aa}=P_{bb}=\frac{1}{1+|\eta|^2}=\cos^2|\beta|. \tag{4.4}$$

The instantaneous population inversion ($q=re^{i\phi}$) is

$$W(t)=P_{ab}-P_{aa}=-\frac{\xi+\xi^*}{1+|\xi|^2}=-\frac{2}{\omega_0}r^2\dot{\phi}. \tag{4.5}$$

The instantaneous electric dipole moment takes the form

$$d(t)=-\frac{\hbar\Omega_0}{E_0}\frac{\eta+\eta^*}{1+|\eta|^2}=-\frac{\hbar\Omega_0}{E_0}\{r^2-1\}, \tag{4.6}$$

and so we can identify the ''angular moment'' introduced above with the population inversion and conclude that the invariant describes the temporal evolution of the rescaled dipole moment $D(t)=r^2-1$,

$$\dot{D}^2+\omega_0^2D^2+\omega_0^2W^2=\omega_0^2. \tag{4.7}$$

In the tridimensional subspace of the phase space of the system labeled by the coordinates $\{\dot{D},D,W\}$, this equation of the invariant describes an ellipsoid with semiaxis $\{\omega_0,1,1\}$ and guarantees a bounded evolution of both the dipole moment and the population inversion for any value of the parameters. In particular, the dipole moment cannot be greater than $\hbar\Omega_0/E_0$ and the function $q(t)$ must be bound as it verifies $0\leqslant r=|q|\leqslant\sqrt{2}$.

Let us now consider the previous system in the traditional Rotating Wave Approximation just by ignoring fast oscillations of frequency $\{\omega+\omega_0\}$ that are much larger than $\{\omega-\omega_0\}$. The Hamiltonian of the system can now be written as

$$H(t)=\hbar\omega_0(\tfrac{1}{2}+J_0)+\tfrac{1}{2}\hbar\Omega_0(e^{-i\omega t}J_++e^{i\omega t}J_-), \tag{4.8}$$

and the Riccati equation is now simplified to read as

$$\dot{\eta}=-i\frac{\Omega_0}{2}(e^{-i\omega t}-\eta^2e^{i\omega t})-i\omega_0\eta, \quad \eta(0)=0, \tag{4.9}$$

whose exact solution is

$$\eta(t)=-i\frac{e^{-i\omega t}\sin(\Delta T)}{\sqrt{1+\kappa^2}\cos(\Delta t)+i\kappa\sin(\Delta t)}, \tag{4.10}$$

$$h(t)=-\omega t-2\arctan\left\{\frac{\kappa}{\sqrt{1+\kappa^2}}\tan(\Delta t)\right\}, \tag{4.11}$$

$$\kappa=\frac{\omega_0-\omega}{\Omega_0}, \quad \Delta=\frac{\Omega_0}{2}\sqrt{1+\kappa^2}. \tag{4.12}$$

When the frequencies ω and Δ are commensurable, (i.e., $\omega/\Delta = p/q$, a rational number), all the physical states of the atom are cyclic. They return to the $t=0$ situation after an evolution time equal to $T=(2\Pi/\Delta)q=(2\Pi/\omega)p$ plus a phase $\exp\{-ip\pi(1+\omega_0/\omega)\}$. In this approximation, the transition probabilities are expressed by the well-known formula:

$$P_{ab}=P_{ba}=\frac{\sin^2(\Delta t)}{1+\kappa^2}, \quad P_{\max}=\frac{1}{1+\kappa^2}, \quad t=\frac{l\pi}{2\Delta}, \tag{4.13}$$

with a maximum value just in resonance. The population inversion can be written as

$$W(t)=-\frac{\kappa^2+\cos(2\Delta t)}{1+\kappa^2}. \tag{4.14}$$

The quantity $W(t)$ oscillates between its maximum and minimum values with frequency 2Δ. The dipole moment is (ignoring spontaneous emission)

$$d(t)=-2\frac{\hbar\Omega_0}{E_0}\frac{\sin(\Delta t)}{1+\kappa^2}\{\sqrt{1+\kappa^2}\cos(\Delta t)\sin(\omega t)+\kappa\sin(\Delta t)\cos(\omega t)\}, \tag{4.15}$$

and thus the Fourier transform of the dipole moment contains just oscillations of frequencies $\pm\omega$ and $\pm(2\Delta\pm\omega)$. Only these frequencies can, in fact, be generated.

V. THE FLOQUET THEORY AND THE ATOM–LASER SYSTEM

The equivalent linear equation (3.14) with $s(x)=1$ has periodic coefficients of period 2π and *Floquet analysis (Bloch Theorem)*[16,17] can be applied. Note, however, that this analysis can also be applied to the case of nonconstant but periodic $s(x)$. Hereafter we shall concentrate just in the $s(x)=1$ case. Suppose that a pair of independent solutions $u(x)$ and $v(x)$ of the equation have been already built. The periodicity property of the coefficients leads to the fact that the translated functions $u(x+2\pi)$ and $v(x+2\pi)$ must be new solutions of the equation themselves and must be expressed as a linear combination of the previous ones. Explicitly,

$$u(x+2\pi)=u(2\pi)u(x)+u'(2\pi)v(x), \tag{5.1}$$

$$v(x+2\pi)=v(2\pi)u(x)+v'(2\pi)v(x). \tag{5.2}$$

There exist solutions—the *Floquet–Bloch functions* $q(x)$—verifying the *Floquet condition*

$$q(x+2\pi)=\lambda q(x), \tag{5.3}$$

where $\lambda(\gamma,\epsilon)$ is a complex number that depends on the specific values of the parameters of the system and must be a solution of the algebraic *characteristic equation*:

$$\lambda^2-[u(2\pi)+v'(2\pi)]\lambda+1=0. \tag{5.4}$$

The allowed values of $\lambda(\gamma,\epsilon)$ must be constrained by the the invariariant (3.12), which must be a C number of unit modulus. Therefore $\lambda=e^{2i\pi\nu}$, and its argument is not single-valued, since all physical situations can be described as we restrict their values to the interval $0\leqslant\nu\leqslant\frac{1}{2}$.

The most general form of the Floquet function $q(x)$ is

$$q(x)=e^{i\nu x}F(x), \quad F(x+2\pi)=F(x) \tag{5.5}$$

being $F(x)$ a 2π-*periodic* function. The Floquet solution $q(x)$ must be bound for all values of the parameters and for all x according to the condition obtained in the previous analysis. We can consider now two different cases.

(a) *Nondegenerate characteristic equation*. When the characteristic equation has two different roots, two independent Floquet functions $q_1(x)$ and $q_2(x)$ exist:

$$q_1(x)=e^{i\nu x}F^+(x), \quad q_2(x)=e^{-i\nu x}F^-(x), \tag{5.6}$$

where $F^\pm(x+2\pi)=F^\pm(x)$ are two linearly independent 2π-periodic functions, and the general solution can be expressed as

$$q(x)=Ae^{i\nu x}F^+(x)+Be^{-i\nu x}F^-(x), \tag{5.7}$$

with constants A and B depending on the initial conditions. It must pointed out that the general solution will not be itself a Floquet function if $AB\neq 0$. Thus, we can conclude that if $\nu\neq 0$ there are no 2π-periodic solutions. The general solution is not a Floquet function, but we can find particular Floquet solutions (quasiperiodic with period 2π) and those systems whose parameter is of the form $\nu=m_1/m_2$ (a rational number) are periodic. Its general solution is $2m_2\pi$ periodic.

(b) *Degenerate characteristic equation*. When $[u(2\pi)+v'(2\pi)]=\pm 2$, the characteristic equation has only one double solution $\lambda=\pm 1$, and there exists only one Floquet solution. However, a pair of functions $q_1(x)$ and $q_2(x)$ can be found, verifying Refs. 16 and 17:

$$q_1(x+2\pi)=\pm q_1(x), \quad q_2(x+2\pi)=\pm[q_1(x)+q_2(x)], \tag{5.8}$$

whose general form is

$$q_1(x)=e^{i\nu x}F_1(x), \quad q_2(x)=e^{i\nu x}\left[\frac{x}{2\pi}F_1(x)+F_2(x)\right], \tag{5.9}$$

with $\nu=0$ or $\frac{1}{2}$ and $F_1(x)$ and $F_2(x)$ 2π periodic. The general solution must now be of the form

$$q(x)=e^{i\nu x}\left\{AF_1(x)+B\left[\frac{x}{2\pi}F_1(x)+F_2(x)\right]\right\}, \tag{5.10}$$

and we can conclude in this case that there will be *particular solutions* of periodic character ($\nu=0$) and of antiperiodic ($\nu=\frac{1}{2}$) character, but the general solution is *neither bound nor periodic*.

VI. THE CONSTRUCTION OF THE FLOQUET FUNCTIONS

The explicit construction of the Floquet functions requires the previous knowledge of the characteristic exponent ν. In order to find ν we have to solve the *characteristic equation* plus the values of $u(2\pi)$ and $v'(2\pi)$. Nevertheless, the symmetry of Eq. (3.14) allows us to write the following set of conditions:

$$u(x+\pi)=u(\pi)u^*(x)+u'(\pi)v^*(x), \tag{6.1}$$

$$v(x+\pi)=v(\pi)u^*(x)+v'(\pi)v^*(x), \tag{6.2}$$

since $q(x+\pi)$ is a solution of the conjugate equation of (3.14), provided that $q(x)$ is a solution of (3.14) as well. One can reproduce the complete solution with the help of these functions $u(x)$ and $v'(x)$ defined in the closed neighbor $[0,\pi]$ and then to obtain the Floquet exponent as

$$2\cos(2\pi\nu)=u(\pi)u^*(\pi)+u'(\pi)v^*(\pi)+v(\pi)u'^*(\pi)+v'(\pi)v'^*(\pi). \tag{6.3}$$

These values, represented in Fig. 1, can be found either by numerical simulation of the equation (3.14) or using the Taylor series (3.16) for these functions. In order to obtain the correspondent 2π-*periodic function* we have to solve the Ordinary Differential Equation:

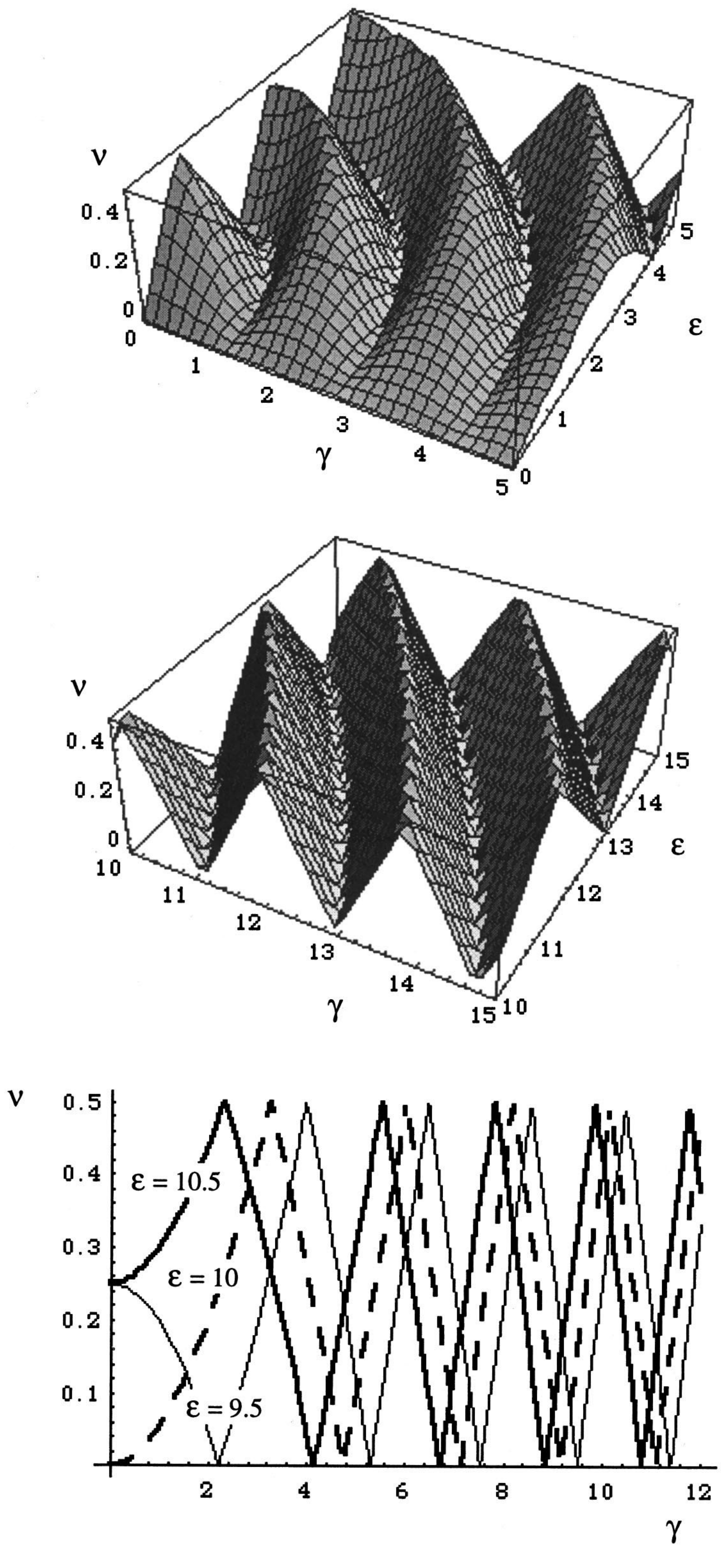

FIG. 1. Floquet exponents as a function of the γ,ϵ parameters.

$$F''+2i(\nu-\gamma\cos x)F'+\left(\frac{\epsilon^2}{4}-\nu^2+2\gamma\nu\cos x\right)F=0. \tag{6.4}$$

Just for the nondegenerate case, we have to find the solutions $F^{\pm}(x)$ corresponding to the two possible values of the Floquet exponents $\pm\nu$. As the periodicity is actually guaranteed, these functions can, in fact, be expressed as a usual Fourier superposition of harmonic components with a frequency multiple of that laser:

$$F^{\pm}(x)=\sum_{p=-\infty}^{\infty} F_p^{\pm}e^{ipx}, \quad F_0^{\pm}=1. \tag{6.5}$$

When we introduce this superposition in the ODE, we find the following relationship between three consecutive coefficients:

$$-\left\{\frac{\epsilon^2}{4\gamma}-\frac{(p\pm\nu)^2}{\gamma}\right\}F_p^{\pm}=(p\pm\nu-1)F_{p-1}^{\pm}+(p\pm\nu+1)F_{p+1}^{\pm}, \tag{6.6}$$

whose resolution requires the introduction of the relative amplitudes between contiguous harmonics,

$$G_p^{\pm}=\frac{F_p^{\pm}}{F_{p-1}^{\pm}}, \quad H_{-p}^{\pm}=\frac{F_{-p}^{\pm}}{F_{-p+1}^{\pm}}, \quad p\geqslant 1, \tag{6.7}$$

that must verify

$$G_p^{\pm}=-\frac{p\pm\nu-1}{V_p^{\pm}+(p\pm\nu+1)G_{p+1}^{\pm}}, \quad p\geqslant 1, \tag{6.8}$$

$$H_{-p}^{\pm}=\frac{p\mp\nu-1}{V_{-p}^{\pm}-(p\mp\nu+1)H_{-p-1}^{\pm}}, \quad p\geqslant 1, \tag{6.9}$$

$$V_l^{\pm}=\frac{\epsilon^2}{4\gamma}-\frac{(l\pm\nu)^2}{\gamma}. \tag{6.10}$$

These relationships determine every coefficient by a *continued fraction*.[7] The way to proceed is the following. If the series G_p and H_{-p} have a finite limit, this must be zero. Therefore, if we consider a value of $p=p_{\max}$ sufficiently high, we can also assume that the correspondent terms G_{p+1} and H_{-p-1} vanish. More precisely, what would vanish is the quantity $\pm\gamma/(p+1)$. Thus, all the previous terms can be calculated hereafter by using the recurrence relationship (6.5)–(6.7). Hence, all the coefficients are given by the quantities

$$F_p^{\pm}=G_p^{\pm}G_{p-1}^{\pm}C_{p-2}^{\pm},\ldots,G_3^{\pm}G_2^{\pm}G_1^{\pm}, \tag{6.11}$$

$$F_{-q}^{\pm}=H_{-q}^{\pm}H_{-q+1}^{\pm}H_{-q+2}^{\pm},\ldots,H_{-3}^{\pm}H_{-2}^{\pm}H_{-1}^{\pm}, \tag{6.12}$$

except the first one that remains undetermined and can be made equal to one $(F_0^{\pm}=1)$.

There exists an additional condition that so far has not been used, namely, the characteristic equation among the coefficients for the value $p=0$:

$$-\left\{\frac{\epsilon^2}{4\gamma}-\frac{\nu^2}{\gamma}\right\}=(\pm\nu-1)H_{-1}^{\pm}+(\pm\nu+1)G_1^{\pm}, \tag{6.13}$$

that must be considered a *compatibility condition*,[18] that connects all the relevant parameters of the system:

$$F(\gamma,\epsilon,\nu)=0, \tag{6.14}$$

and can be used to find through an iterative process the characteristic exponent if it is still unknown[18] or to classify all systems with a fixed exponent. The results coincide with those of Fig. 1. In particular, when $\nu=0$ or $\frac{1}{2}$, the equation (6.14) selects the parameters of those particular systems that have a general periodic or antiperiodic solution and forces the physical system described by our Hamiltonian to a truly periodical behavior.

All these properties allow us to establish an important relationship between the functions $F^{+}(x)$ and $F^{-}(x)$. It is easy to see that the following relations hold:

$$G_p^{+}=-H_{-p}^{-},\quad G_p^{-}=-H_{-p}^{+}. \tag{6.15}$$

As a consequence of these symmetries, one has to find the amplitudes of only one of the two functions because these coefficients verify

$$F_p^{-}=(-1)^p F_{-p}^{+}, \tag{6.16}$$

and $F^{\pm}(x)$ can be expressed by series of *even harmonics* and *odd harmonics* of the laser according to the expression

$$F^{+}(x)=E(x)+O(x),\quad F^{-}(x)=E^{*}(x)-O^{*}(x), \tag{6.17}$$

where

$$E(x)=\sum_{p=-\infty}^{\infty} F_{2p}e^{i2px},\quad O(x)=\sum_{p=-\infty}^{\infty} F_{2p+1}e^{i(2p+1)x}, \tag{6.18}$$

and we have redefined $F_j^{+}=F_j$ in order to simplify the notation. As soon as the *Floquet functions* have been constructed, the function $q(x)$ is totally determined by using the physical initial conditions of the problem. The result is

$$q(x)=Ae^{i\nu x}[E(x)+O(x)]+Be^{-i\nu x}[E^{*}(x)-O^{*}(x)], \tag{6.19}$$

$$A=\frac{1}{\Sigma}\left\{(c-d)\left(\nu+\frac{\epsilon}{2}\right)+e-f\right\}, \tag{6.20}$$

$$B=\frac{1}{\Sigma}\left\{(c+d)\left(\nu-\frac{\epsilon}{2}\right)+e+f\right\}, \tag{6.21}$$

$$\Sigma=2(ec-df)+2\nu(c^2-d^2), \tag{6.22}$$

where *c*, *d*, *e*, and *f* are constants depending on the initial values of the Floquet functions and their derivatives,

$$c=\sum_{p=-\infty}^{\infty} F_{2p},\quad d=\sum_{p=-\infty}^{\infty} F_{2p+1}, \tag{6.23}$$

$$e=2\sum_{p=-\infty}^{\infty} pF_{2p},\quad f=\sum_{p=-\infty}^{\infty} (2p+1)F_{2p+1}. \tag{6.24}$$

We would like to emphasize at this point that these series are always convergent as they correspond to Fourier coefficients of regular periodic functions.

VII. FLOQUET ANALYSIS AND THE HARMONIC GENERATION

The interaction between the atom and the coherent field forces the atomic dipole to oscillate with frequencies that can be different of the external field. Thus, one can analyze the phenomenon of higher harmonic generation, starting with the Fourier Transform of the instantaneous dipole moment of the atom. One has just to remember the relationship between this physical observable and the complex function $q(x)$. It is quite clear that we can obtain its harmonic spectrum with the help of $q(x)$. Alternatively, we can consider the equivalent system of equations, obtained from (3.12)–(3.13) and (4.5)–(4.7):

$$D''+\epsilon^2 D=2\gamma\epsilon\cos x W,\quad D(0)=0,\quad D'(0)=0, \tag{7.1}$$

$$W'+2\frac{\gamma}{\epsilon}\cos x D'=0,\quad W(0)=-1, \tag{7.2}$$

which describes the dipole moment forced by the population inversion. This is the treatment used in Ref. 7, where a stationary approximation of the problem is made and the instantaneous dipole moment of the atom is found in terms of a Fourier series built with a superposition of *only odd harmonics* of the frequency of the coherent field. The amplitude of the different harmonics is obtained using a *continued fraction*[7] totally equivalent to the one we have described above. However, the stationary approximation becomes unnecessary if we can describe in an analytical form the exact instantaneous dipole moment of the atom represented by its Fourier Transform whose frequencies and amplitudes are totally known, as we have just done. An exact expression would be

$$\begin{aligned} D(x)=(A^2+B^2)\sum_{s=0}^{\infty} P_s\cos(2sx)-1+2(A^2-B^2)\sum_{s=0}^{\infty} I_s\cos[(2s+1)x] \\ +2AB\sum_{s=0}^{\infty}\{T_s^+\cos[2(s+\nu)x]+T_s^-\cos[2(s-\nu)x]\}, \end{aligned} \tag{7.3}$$

whose coefficients are expressed in terms of the constants that we have just found:

$$P_0=\sum_{p=-\infty}^{\infty} F_pF_p,\quad T_0^{\pm}=\frac{1}{2}\sum_{p=-\infty}^{\infty}(F_{2p}F_{-2p}-F_{2p+1}F_{-2p-1}), \tag{7.4}$$

$$P_s=2\sum_{p=-\infty}^{\infty} F_pF_{p+2s},\quad s\geqslant 1, \tag{7.5}$$

$$I_s=\sum_{p=-\infty}^{\infty} F_{2p}(F_{2p+2s+1}+F_{2p-2s-1}),\quad s\geqslant 0, \tag{7.6}$$

$$T_s^{\pm}=\sum_{p=-\infty}^{\infty}(F_{2p}F_{\pm 2s-2p}-F_{2p+1}F_{\pm 2s-2p-1}),\quad s\geqslant 1. \tag{7.7}$$

One could, in principle, work directly with the system (7.1)–(7.2), basically because the convergence of the series appearing in $D(x)$ is faster than the series of $q(x)$. However, the periodic character of the functions $D(x)$ and $W(x)$ is still unknown as *there is no reason to assume that the solution of a linear system of first-order differential equations with periodic* *time-dependent coefficients is actually periodic*. In fact, only an adequate Floquet analysis (Bloch) of the type that has been performed in Ref. 3 can give us a clear cut answer to this question. According to our results the dipole moment must be described by a real function that must contain a purely 2π-periodic real function and another periodic function modulated by an exponential factor that contains the contribution due to the Floquet exponent. Let the primed coefficients be the Fourier coefficients of the Dipole Moment $D(x)$ that we have found using Eqs. (7.1) and (7.2). This function $D(x)$ can be expressed in this notation as $(T_0'^{-}=0)$

$$D(x)=\sum_{s=0}^{\infty}\{I_s'\cos((2s+1)x)+T_s'^{+}\cos(2(s+\nu)x)+T_s'^{-}\cos(2(s-\nu)x)\}. \tag{7.8}$$

This ansatz already contains the initial condition $D'(0)=0$ and the mathematical property concerning the fact that the system of equations itself cancels the even harmonics in $D(x)$. Also, the ansatz must yield the complete solution when the initial state has a good energy quantum number of the noninteracting atom. This is due to the property $D(t)=|q|^2-1$ that shifts just the even harmonics in an amount 2ν (instead of ν, as one might have initially expected). The same effect could be interpreted as a split of the odd harmonics followed by a shift of one of these in an amount $\pm(1-2\nu)$, giving rise to a sort of *triplet associated to each odd harmonic*, as can be seen in Fig. 2.

The relationships between contiguous harmonics $\Lambda_s=I_s'/I_{s-1}'$, $\Lambda_s^{\pm}=T_s'^{\pm}/T_{s-1}'^{\pm}$ must now verify

$$\Lambda_s=-\frac{2s-1}{2s}\left\{\frac{\epsilon^2-(2s+1)^2}{\gamma^2}+\frac{(2s+1)^2}{2s(s+1)}+\frac{2s+3}{2(s+1)}\Lambda_{s+1}\right\}^{-1},\quad s\geqslant 1, \tag{7.9}$$

$$\Lambda_s^{\pm}=-\frac{s-1\pm\nu}{2s-1\pm2\nu}\left\{\frac{\epsilon^2-4(s\pm\nu)^2}{2\gamma^2}+\frac{4(s\pm\nu)^2}{4(s\pm\nu)^2-1}+\frac{s+1\pm\nu}{2s+1\pm2\nu}\Lambda_{s+1}^{\pm}\right\}^{-1}, \tag{7.10}$$

and all these coefficients go over $\gamma^2/4s^2$ for large values of s and verifying the compatibility condition,

$$\frac{\epsilon^2-4\nu^2}{2\gamma^2}+\frac{4\nu^2}{4\nu^2-1}+\frac{1+\nu}{1+2\nu}\Lambda_1^{+}+\frac{1-\nu}{1-2\nu}\Lambda_1^{-}=0, \tag{7.11}$$

that can be used to adjust the value of ν. Therefore, starting with a sufficiently high value of s we can calculate the ratios $\Lambda_s=\Lambda_s^{\pm}=\gamma^2/4s^2$, and using (7.9)–(7.10) we can go backward and calculate all the others. With all these ratios we find the coefficients:

$$I_s'=\Lambda_s\Lambda_{s-1}\Lambda_{s-2}\Lambda_{s-3},\ldots,\Lambda_3\Lambda_2\Lambda_1I_0'=\Delta_sI_0',\quad s\geqslant 1, \tag{7.12}$$

$$\Lambda_s^{\pm}=\Lambda_s^{\pm}\Lambda_{s-1}^{\pm}\Lambda_{s-2}^{\pm}\Lambda_{s-3}^{\pm},\ldots,\Lambda_3^{\pm}\Lambda_2^{\pm}\Lambda_1^{\pm}T_0'^{+}=\Delta_s^{\pm}T_0'^{+}, \tag{7.13}$$

in terms of two of these I_0' and $T_0'^{+}$ that can be considered as integration constants. These integration constants are fixed by using the initial conditions $D(0)=0$ and $W(0)=-1$. This leads to the following set of algebraic equations:

$$C'I_0'+D'T_0'^{+}=0, \tag{7.14}$$

$$\left(\frac{\epsilon^2-1}{\gamma^2}-\frac{1}{2}-E'+\frac{3}{2}\Lambda_1\right)I_0'-4F'T_0'^{+}=-\frac{2\epsilon}{\gamma}, \tag{7.15}$$

where

TABLE III. The Fourier spectrum of the electric dipole moment. The symbol* means amplitudes less than 10^{-3}. The oscillations with a negative amplitude have an extra phase of $\exp\{i\pi\}$.

Harmonic	Ampl. 10^3 $\gamma=10$ $\epsilon=10$ $\nu=0.408\,239$	Harmonic	Ampl. 10^3 $\gamma=6.63$ $\epsilon=8.75$ $\nu=-0.305\,808$	Harmonic	Ampl. 10^3 $\gamma=13.84$ $\epsilon=8.75$ $\nu=-0.177\,815$
0.816 477	46.0882	0.611 616	16.3369	0.355 608	118.827
1	−533.334	1	−537.363	1	−490.609
1.183 523	26.6766	1.388 384	4.589 54	1.644 392	21.119
2.816 477	−21.5989	2.611 616	−15.5481	2.355 608	−38.502
3	95.009	3	77.6123	3	118.355
3.183 523	−5.318 95	3.388 384	*	3.644 392	−6.957
4.816 477	38.5086	4.611 616	43.0009	4.355 608	59.346
5	−29.3072	5	−20.1898	5	−50.194
5.183 523	1.713 39	5.388 384	*	5.644 392	3.467
6.816 477	−80.3143	6.611 616	−112.965	6.3556	92.023
7	9.472 38	7	6.4594	7	25.127
8.816 477	155.391	8.611 616	247.75	8.355 608	150.416
9	*	9	*	9	−11.635
9.183 523	*	9.388 384	*	9.644 392	1.522
10.816 477	−253.567	10.611 616	−378.329	10.355 608	−204.139
11	−7.98646	11	−4.495 26	*	*
12.816 477	303.189	12.611 616	212.905	12.355 608	225.252
13	13.1326	13	4.478 63	13	7.713
14.816 477	−170.788	14.611 616	295.105	14.355 608	−165.436
15	−9.369 89	15	4.226 36	15	−12.023
16.816 477	133.787	16.611 616	121.305	16.355 608	11.485
17	−4.784 48	17	1.5273	17	7.223
18.816 477	174.182	18.611 616	28.8898	18.355 608	135.262
19	8.852 01	19	*	19	4.534
20.816 477	208.911	20.611 616	4.734 98	20.355 608	−93.463
21	9.399 26	21	*	21	−8.745
22.816 477	107.257	22.611 616	*	22.355 608	−100.988
23	4.592 09	23	*	23	−3.268
24.816 477	35.3672	24.611 616	*	24.355 608	65.533
25	1.465 41	25	*	25	6.890
26.816 477	8.5114	26.611 616	*	26.355 608	136.739
27	*	27	*	27	8.4032
28.816 477	1.594 98	28.611 616	*	28.355 608	100.711
29	*	29	*	29	5.421
30.816 477	*	30.611 616	*	30.355 608	47.464

$$C'=\sum_{s=0}^{\infty}\Delta_s,\quad F'=\sum_{s=0}^{\infty}\left\{\frac{4(s+\nu)^2}{4(s+\nu)^2-1}\Delta_s^{+}+\frac{4(s-\nu)^2}{4(s-\nu)^2-1}\Delta_s^{-}\right\},\tag{7.16}$$

$$E'=\sum_{s=1}^{\infty}\frac{(2s+1)^2}{s(s+1)}\Delta_s,\quad D'=\sum_{s=0}^{\infty}\{\Delta_s^{+}+\Delta_s^{-}\}.\tag{7.17}$$

Finally, we identify

$$P_s=0,\quad I_s=I'_s,\quad T_s^{+}=T_s'^{+},\quad T_s^{-}=T_s'^{-},\tag{7.18}$$

as can easily be seen for any particular set of values of the parameters. The amplitude of every harmonic can be calculated either by (7.4)–(7.7) or (7.8)–(7.17).

The *Fourier Spectrum* so far described in the literature includes only *odd* harmonics. According to the results developed in this paper, the following composition of the *Fourier Spectrum* both in frequencies and amplitudes emerges. Indeed, *odd integer harmonics* must necessarily be an

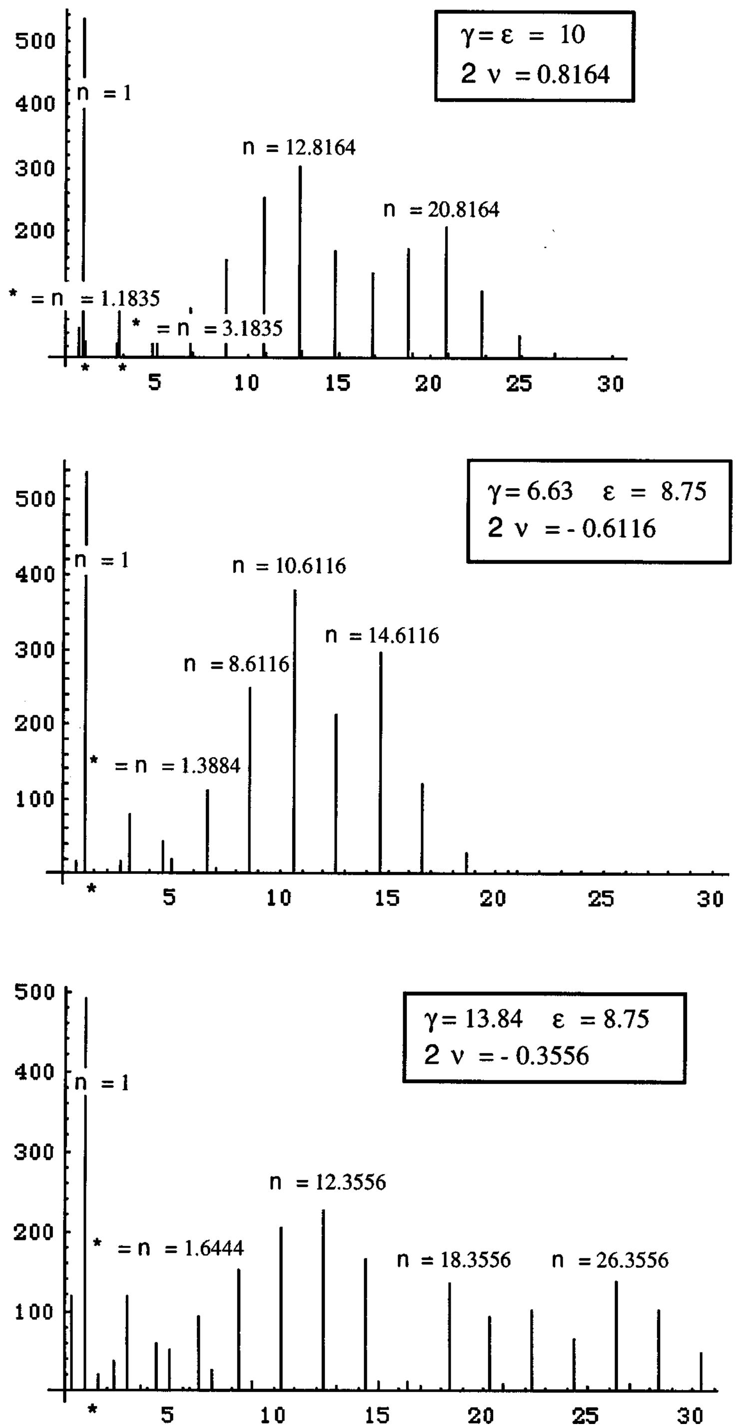

FIG. 2. Fourier spectra for different values of the ϵ,γ parameters.

unavoidable ingredient of the spectrum, but it must also contain *noninteger harmonics* that can be interpreted as the *odd harmonics displaced an amount* $\pm(1-2\nu)$ fixed by the *Floquet exponents* of the system. In both cases (integer and noninteger) the amplitude can be calculated by the expressions (7.4)–(7.7) or alternatively (7.8)–(7.17). We would like to emphasize that this result has arisen directly from the systematic use of the Floquet Analysis. An interesting question for further research is whether a shift in the physical initial conditions may change this Fourier Spectrum that we have just described. Some preliminary calculations show us that if the initial state of the atom is not an eigenstate of the energy, the formalism can be equally applied to the prediction of the Spectrum, but the results show nontrivial changes, both in the amplitudes and in the role played by the phases of the harmonics. Note that the formalism and hence the spectrum, depends only upon the full knowledge of $q(x)$.

A remaining question deals with the behavior of the spectrum in the limiting cases $\omega_0 \to 0$ and $\omega \to 0$. In the first case, any of the equations (3.5) or (3.10) lead to the same result. If the initial state of the atom is an eigenstate of energy, the Dipole Moment vanishes and no outgoing emission would be present. In the latter case, using Eq. (3.10) we obtain in this limit,

$$q(t)=\exp\{i\Omega_0 t\}\left\{ \cos\left(\sqrt{\Omega_0^2+\frac{\omega_0^2}{4}}\,t\right) - i\,\frac{\Omega_0-\frac{\omega_0}{2}}{\sqrt{\Omega_0^2+\frac{\omega_0^2}{4}}}\sin\left(\sqrt{\Omega_0^2+\frac{\omega_0^2}{4}}\,t\right)\right\}, \quad (7.19)$$

and the Dipole Moment takes the form in this case

$$D(t)=|q|^2-1=-\frac{\omega_0\Omega_0}{2\left(\Omega_0^2+\frac{\omega_0^2}{4}\right)}\left\{1-\cos\left(2\sqrt{\Omega_0^2+\frac{\omega_0^2}{4}}\,t\right)\right\}, \quad (7.20)$$

that describes a spectrum given by a constant contribution and just one frequency of value,

$$2\sqrt{\Omega_0^2+\frac{\omega_0^2}{4}}, \quad (7.21)$$

which is the standard result.

In Table III and Fig. 2 we give the results obtained in a *calculation of the Fourier Spectrum* of the atomic dipole moment based on the equations (7.4)–(7.8) and (7.8)–(7.17) for some values of the parameters. Only the Fourier components with an amplitude larger than 10^{-3} have been included. In the case $\epsilon=\gamma=10$, the energy of the atom–laser interaction equals the energy of the atomic transition, and one needs ten photons of the laser to produce the transition. A first estimation of the *Floquet exponent* was made, starting from the values of $u(\pi)$, $u'(\pi)$, $v(\pi)$, $v'(\pi)$ arising from a simulation of the ODE (3.14). The value $\nu=0.408\,202$ was obtained. This value was improved in the above described iterative process, yielding finally a value of $\nu=0.408\,238\,570\,386\,171\,53$. For this last numerical value the *compatibility conditions* (6.13) and (7.11), are zero with numerical precision of one part in 10^{14}. The values of the parameters do not have to be integers, as it is also shown in Table III for the cases $\epsilon=8.75$, $\gamma=6.63$, and $\epsilon=8.75$, $\gamma=13.84$. The theoretically calculated Floquet exponents are $-0.305\,808\,224\,854\,750\,90$ and $-0.177\,814\,603\,653\,674\,38$, respectively, fulfilling the compatibility conditions (6.13) and (7.11) also with an accuracy of one part in 10^{14}.

VIII. CONCLUSIONS

The physical system composed by a two-level atom interacting with a coherent external electromagnetic field has an *exact solution for all ratios between the relevant frequencies of the problem: Transition, Laser, and Rabi frequencies. It is possible to* make use of the dynamical symmetry of the Hamiltonian to develop a *nonperturbative method that allows us to solve exactly the model*. No approximations at all such as the Rotating Wave or Weak Field Approximations have been made. The Schrödinger equation is equivalent to a Riccati differential equation whose solution yields all the physical observables of the system. This Riccati equation can be linearized and becomes a linear second-order differential equation with periodic coefficients. This periodic character allows us to apply Floquet Theory for differential equations with periodic coefficients in order to describe the periodicity properties of the interacting atom. In particular, when the Fourier transform of the electric dipole moment of the atom is calculated and the problem of the Higher Harmonics Generation is considered, qualitatively new and rather surprising conclusions are found. To the usual harmonic spectrum composed by harmonics labeled just by *odd* integers, one should add a series of noninteger harmonics that heavily depend on the Floquet exponent. This Fourier Spectrum was implicit in other contributions on the two-level model (see Ref. 5 and references therein), but is not yet fully understood. Finally, the present authors claim that the new harmonics need to be more carefully analyzed experimentally, as they seem to arise naturally from a model that describes quite accurately the important effect of Higher Harmonic Generation. At any rate the Exact Solution for Strong Field Laser–Atom Interaction hereby presented is relevant in its own right, since the Two-State Model can also be used for a large variety of Physical Applications.

Note added in proof: A more detailed account of the numerical results hereby presented, including a general study for a wide range of the physical parameters as well as a careful numerical analysis dealing with the limiting cases, will be the subject of a forthcoming publication.

ACKNOWLEDGMENTS

We acknowledge with thanks Professor Joseph H. Eberly for his advice and support as well as for a careful reading of a previous version of this manuscript. We would also like to thank both Professor L. Roso-Franco for bringing Ref. 7 to our attention and to an anonymous referee whose suggestions greatly improved various technical aspects of the paper as well as the set of references. This research has been supported in part by the DGICYT under Contract No. PB95-0947.

[1] K. C. Kulander and B. W. Shore, J. Opt. Soc. Am. B **7**, 502–508 (1990).
[2] J. H. Eberly, Q. Su, J. Javanainen, K. C. Kulander, B. W. Shore, and L. Roso-Franco, J. Mod. Opt. **36**, 829–855 (1989).
[3] J. H. Shirley, Phys. Rev. **138**, 979–987 (1965).
[4] Ya. B. Zel'dovich, Sov. Phys. Usp. **16**, 427–441 (1973).
[5] G. Compagno, K. Dietz, and F. Persico, J. Phys. B **27**, 4779–4815 (1994).
[6] K. C. Kulander, Phys. Rev. A **38**, 778–784 (1988).
[7] L. Plaja and L. Roso-Franco, J. Opt. Soc. Am. B **9**, 2210–2213 (1992).
[8] B. Sundaram and P. W. Milonni, Phys. Rev. A **41**, 6571–6573 (1990).
[9] R. M. Potvliege and R. Shakeshaft, Phys. Rev. A **40**, 3061–3079 (1989).
[10] J. H. Eberly, Q. Su, and J. Javanainen, Phys. Rev. Lett. **62**, 881–885 (1989).
[11] N. Ben-Tal, N. Moseyev, R. Kosloff, and C. Cerjan, J. Phys. B **26**, 1445–1461 (1993).
[12] J. M. Cerveró and J. D. Lejarreta, J. Phys. A **29**, 7545–7560 (1996).
[13] R. Loudon, *The Quantum Theory of Light* (Oxford University Press, Oxford, 1991).
[14] B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1973).
[15] A. M. Perelomov, *Generalized Coherent States and Their Applications* (Springer-Verlag, Berlin, 1986).
[16] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956).
[17] E. T. Whittaker and G. N. Watson, *Course of Modern Analysis* (Cambridge Mathematical Library, Cambridge, 1922).
[18] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).

Generalization of the Birman–Schwinger method for the number of bound states

Khosrow Chadan
Laboratoire de Physique Théorique et Hautes Energies,[a] *Université Paris XI, Bâtiment 210, F-91405 Orsay Cedex, France*

Reido Kobayashi
Department of Mathematics, Science University of Tokyo, Noda, Chiba 278, Japan

Monique Lassaut
Groupe de Physique Théorique, Institut de Physique Nucléaire,[a] *Université Paris XI, Bâtiment 100, 91406 Orsay Cedex, France*

(Received 22 October 1998; accepted for publication 11 December 1998)

We generalize the Birman–Schwinger method, and derive a general upper bound on the number of bound states in the *S* wave for a spherically symmetric potential. This general bound includes, of course, the Bargmann bound, but also leads, for increasing (negative) potentials, to a Calogero–Cohn-type bound. Finally, we show that for a large class among these potentials, one can obtain further improvements.
 [S0022-2488(99)00804-X]

I. INTRODUCTION

The Birman–Schwinger method[1,2] was invented in order to put the Lippmann–Schwinger equation of scattering theory[3,4] on a firm basis within the framework of the Fredholm theory of integral equation in L^2 spaces. Among many important results, it led to a general proof of the Bargmann bound on the number of bound states,[5] without using the nodal theorem, which asserts that the number of bound states in each angular momentum state of a spherically symmetric potential is equal to the number of nodes (zeros) of the regular solution of the reduced radial Schrödinger equation at zero energy. The nodal theorem was used by Bargmann in his original proof.[5]

The reduced radial Schrödinger equation at zero energy for the lth angular momentum state reads

$$\varphi_l''(r)=\left[\frac{l(l+1)}{r^2}+V(r)\right]\varphi_l(r), \quad \varphi_l(0)=0. \tag{1}$$

The potential is assumed here to satisfy the integrability condition

$$\int_0^\infty r|V(r)|dr<\infty, \tag{2}$$

which entails the self-adjointness of the Hamiltonian $-(d^2/dr^2)+l(l+1)/r^2+V(r)$ in $L^2(0,\infty)$, and other good properties for a well-defined scattering theory.[3,4] Combining the Schrödinger equation and the boundary condition at the origin, one can write (1) as the Fredholm integral equation[3]

$$\varphi_l(r)=r^{l+1}-\frac{1}{(2l+1)}\int_0^\infty r_<^{l+1}r_>^{-l}V(r')\varphi_l(r')dr', \tag{3}$$

[a]Laboratoires associés au Centre National de la Recherche Scientifique.

where the symbols $r_<$ and $r_>$ mean, respectively, minimum (r,r') and maximum (r,r'). Here, the Birman–Schwinger method is to multiply both sides of (3) by $|V(r)|^{1/2}$, and write $\psi_l=|V(r)|^{1/2}\varphi_l$. The equation for ψ_l then has the kernel given by

$$K_l(r,r')=-\frac{r_<^{l+1}r_>^{-l}}{(2l+1)}|V(r)|^{1/2}|V(r')|^{1/2}\frac{V(r')}{|V(r')|}, \tag{4}$$

and it is easily seen that this kernel is indeed of Hilbert–Schmidt type, and has the trace

$$\text{Trace}\,K_l(r,r')=-\frac{1}{2l+1}\int_0^\infty r|V(r)|\text{sign}\,V(r)dr, \tag{5}$$

which, by (2), is finite.

Now, if we write $V=V_++V_-$, where V_+ is the positive part of $V:V_+=V(r)\theta[V(r)]$, and V_- its negative part defined similarly, we have, for the number of bound states n[1–4]

$$n(V)\leqslant n(V_-).$$

An upper bound for $n(V)$ is therefore given by $n(V_-)$. For V_-, the corresponding kernel (4) and its trace (5) are now positive. K_l becomes a symmetric positive kernel, and its characteristic values $\lambda_j^{(-)}$ (inverse of its eigenvalues) are all positive, and accumulate at infinity as $j\rightarrow\infty$ (remember that, for Hilbert–Schmidt kernels, the eigenvalues accumulate at the origin as $j\rightarrow\infty$). With V_-, these characteristic values are defined by

$$\psi_{l,j}(r)=\lambda_j^{(-)}\int_0^\infty\frac{r_<^{l+1}r_>^{-l}}{(2l+1)}|V_-(r)|^{1/2}|V_-(r')|^{1/2}\psi_{l,j}(r')dr'. \tag{6}$$

That there are an infinity of eigenvalues comes from the fact that the kernel is not of finite rank, unless V_- is a finite sum of δ potentials. Looking differently, if we multiply V_- by λ, and increase λ from $\lambda=0$, the characteristic values $\lambda_j^{(-)}$ are the values for which new bound states appear one by one at zero energy, the bottom of the continuum, and then move to the left to become real bound states with negative energy as we increase λ. That this is so follows from the Feynman–Hellmann theorem for bound states (V is negative here!)[4,6]

$$\frac{\partial E_j^{(-)}(\lambda)}{\partial\lambda}\leqslant 0,\quad \lambda\geqslant\lambda_j,\quad E_j^{(-)}(\lambda_j)=0. \tag{7}$$

In all rigor, one must first consider the bound states below some negative energy, $E\leqslant-\epsilon^2$, $\epsilon>0$, consider the number $n_\epsilon(V)$ of these bound states, and then make $\epsilon\downarrow 0$ in order to find the total number of bound states. As shown in Refs. 1, 4, and 6, one uses here the continuity theorem, which shows that one can take this limit for a large class of potentials. In the radial case, this class includes potentials satisfying (2).

We are interested now in those characteristic values $\lambda_j^{(-)}$ for V_- which are in the interval [0, 1] because they give the total number of bound states of V_-. Obviously

$$n(V_-)\leqslant\sum_{\lambda_j^{(-)}\leqslant 1}\frac{1}{\lambda_j^{(-)}}. \tag{8}$$

Therefore

$$n_l(V)\leqslant n_l(V_-)\leqslant\sum_{\lambda_j^{(-)}\leqslant 1}\frac{1}{\lambda_j^{(-)}}\leqslant\sum_{j=1}^{\infty}\frac{1}{\lambda_j^{(-)}}=\text{Trace}\,K_l(V_-)=\frac{1}{(2l+1)}\int_0^\infty r|V_-|dr, \tag{9}$$

which is the Bargmann bound. One can, of course, replace V_- by V inside the integral in (9), but this makes the upper bound bigger, and is unnecessary. Notice that, in (9), the finite sum is less than the infinite sum if all $\lambda_j^{(-)}$ are positive, which is the case here. The potential V_- is negative, and therefore λV_- cannot have bound states if λ also is negative.

The upper bound (9), while it can be saturated by sums of delta-function potentials[1–3]

$$V(r) = -\sum_{1}^{n} g_j \delta(r - r_j), \quad g_j > 0, \tag{10}$$

by adjusting g_j and r_j in order to get equality between the two sides of (9), has several weak points. The first is its bad behavior for strong attractive potentials. Indeed, if we replace V_- by λV_-, and make $\lambda \to \infty$, (9) would suggest that $n(\lambda V_-)$ should grow like λ, whereas we know that it actually grows like $\lambda^{1/2}$, as was shown some years ago.[7,8] This weak point was improved by a new bound by Calogero[9] and Cohn,[10] who showed that, at least for the class of purely attractive and increasing potentials, one has the bound, assuming of course the integral to be finite,

$$n_0(V) \leqslant \frac{2}{\pi} \int_0^\infty |V(r)|^{1/2} dr, \tag{11a}$$

$$V < 0, \quad V' \geqslant 0, \quad V(\infty) = 0. \tag{11b}$$

Here also, one can show that one can make the left-hand side of (11a) as close to the right-hand side as one wishes, with appropriate square-well potentials. The above bound, valid for the S wave $(l=0)$, and therefore for all higher angular momenta, has itself the defect of having, contrary to (9), no l dependence, and also contains the condition $V' \geqslant 0$. It has been shown recently that the condition $V' \geqslant 0$ can be relaxed to a large extent,[11] and also that (11a) can be generalized to $l > 0$ with appropriate l dependence.[12] Let us emphasize here that the proofs of Calogero and Cohn are both based on the nodal theorem mentioned earlier.

The second weak point of (9) is that it is not good for potentials having a strong repulsive (positive) part for, in such cases, the left-hand side of (9) may be much smaller than the right-hand side.

In the present paper, generalizing the Birman–Schwinger method, we are going to show that, first, for potentials satisfying (11b), one can get a bound of the Calogero–Cohn type by the trace method given above, though with a slightly larger constant than $2/\pi$ as in (11a), without using the nodal theorem; and second, generalize this same bound, with an extra term, for a large class of negative potentials which may oscillate (no condition on V'). Finally, we shall show that, for a large class of potentials, one can also improve these new bounds further.

II. CALOGERO–COHN-TYPE BOUND

We consider here purely negative potentials satisfying (2). As mentioned before, we are going to generalize the Birman–Schwinger method to obtain a more general bound than (9), which would contain, as a particular case, a Calogero–Cohn-type bound, of the form (11a), but with the constant 1 instead of $2/\pi$ in front of the integral. For the simplicity of algebra, we restrict ourselves to the case $l=0$ (S wave), and leave the case $l>0$ for a forthcoming paper. In essence, the Birman–Schwinger method relies on the Lippmann–Schwinger integral equation (3) in which the inhomogeneous term is the regular solution for the free case, i.e., without the potential V. However, this is not the only way of writing an integral equation for the full solution. One can write a more general integral equation by writing the Hamiltonian as (remember that we consider $l=0$)

$$H \equiv H_0 + V = -\frac{d^2}{dr^2} + V = \left(\frac{-d^2}{dr^2} + V_0 \right) + (V - V_0) \equiv \overline{H}_0 + \overline{V}, \tag{12}$$

where we assume that V_0 satisfies also (2). If we denote by φ_0 and χ_0 the regular and irregular solutions of $\overline{H}_0\psi=0$, such that $\varphi_0(0)=0$, $\varphi_0'(0)=1$, $\chi_0(0)=1$, $\lim_{r\to 0} r\chi_0'(r) = \lim_{r\to 0}\varphi_0(r)\chi_0'(r)=0$, and Wronskian $(\varphi_0,\chi_0)\equiv\varphi_0'\chi_0-\varphi_0\chi_0'=1$, we have the integral equation[3]

$$\varphi(r)=\varphi_0(r)-\int_0^\infty \varphi_0(r_<)\chi_0(r_>)\overline{V}(r')\varphi(r')dr', \tag{13}$$

as can be verified easily. Also, since we assume V_0 to satisfy (2), the above equation is, as for (3), a good equation of the Fredholm type, with well-defined Fredholm determinants (numerator as well as denominator).[3] One can again use the Birman–Schwinger method by multiplying both sides of (13) by $|\overline{V}|^{1/2}$ in order to obtain an L^2 kernel of Hilbert–Schmidt type with finite trace. Indeed, since V_0 is supposed to satisfy (2), we have the general bound[3]

$$|\varphi_0\chi_0|\leqslant Ar \tag{14}$$

for all values of r, with an appropriate finite constant $A(>0)$. When $V_0=0$, we get back, of course, (3) for $l=0$. *We shall assume henceforth that* $V_0(r)\geqslant 0$. Let us now look at the homogeneous integral equation which defines the characteristic values $\overline{\lambda}_j$ which correspond to the thresholds in $\overline{\lambda}$ for having new bound states appearing at zero energy ($\overline{V}=V-V_0$, $V\leqslant 0$, $V_0\geqslant 0$, $\overline{V}\leqslant 0$):

$$\overline{\varphi}(r)=-\overline{\lambda}\int_0^\infty \varphi_0(r_<)\chi_0(r_>)\overline{V}(r')\overline{\varphi}(r')dr' \tag{15}$$

and its corresponding differential equation

$$\overline{\varphi}''(r)=V_0\overline{\varphi}(r)+\overline{\lambda V}\overline{\varphi}(r). \tag{16}$$

We have to compare this equation with the original Schrödinger equation

$$\varphi''(r)=\lambda V(r)\varphi(r). \tag{17}$$

Since $\overline{V}=V-V_0$, we see that (16) and (17) are identical for $\overline{\lambda}=\lambda=1$, as it should be. Now, if $V(r)$ has n bound states, this means that there are n characteristic values λ_j in the interval (0, 1]. Therefore, there should exist also n values of $\overline{\lambda}_j$ in (0, 1] since for $\overline{\lambda}=1$, $V_0+\overline{\lambda V}$ has n bound states. There is a one-to-one correspondence between λ_j and $\overline{\lambda}_j$ for $j=1,2,...,n$. Also, since V_0 in (16) was assumed to be positive, it is obvious, by comparing $\varphi_j''=\lambda_j V\varphi_j$ with $\overline{\varphi}_j''=V_0\overline{\varphi}_j+\overline{\lambda}_j\overline{V}\overline{\varphi}_j=\overline{\lambda}_j V\overline{\varphi}_j+(1-\overline{\lambda}_j)V_0\overline{\varphi}_j$, that we should have $\overline{\lambda}_j>\lambda_j$ for $j=1,2,...,n$. This is simply because, for $j\leqslant n$, the potential $(1-\overline{\lambda}_j)V_0$ in the last equation is repulsive. For $j>n$, the reverse is true: $\lambda_j>\overline{\lambda}_j$. These are all simple consequences of Sturm-type comparison theorems.[15] Also, it is obvious that we cannot have any negative characteristic values $\overline{\lambda}$ for (16) since, for $\overline{\lambda}<0$, the total potential there is positive. In any case, for getting an upper bound for the number of bound states of $V(r)$, we can, as well, look at (15), and use

$$n(V)\leqslant\sum_{\overline{\lambda}_j\leqslant 1}\frac{1}{\overline{\lambda}_j}. \tag{18}$$

Proceeding as before, and remembering that, by hypotheses on V and V_0, we are dealing with positive definite kernels, we get the general bound

$$n(V) \leqslant -\int_0^\infty \varphi_0(r)\chi_0(r)[V-V_0]dr, \quad V\leqslant 0, \quad V_0 \geqslant 0, \tag{19}$$

which we are going now to exploit by making appropriate choices for V_0. Note that, if we take $V_0=0$, we have $\varphi_0=r$, $\chi_0=1$, and we get back the Bargmann bound.

Calogero–Cohn-type bound: Since V is assumed to be purely attractive, and we wish to obtain a Calogero–Cohn-type bound, we have also to assume, as they did, that $V'\geqslant 0$. The appropriate choice of V_0 would here be

$$V_0 = -V - \frac{d}{dr}\sqrt{-V(r)}, \quad \bar{V} = 2V + \frac{d}{dr}\sqrt{-V}, \tag{20}$$

in V_0 both parts are positive, and in $\bar{V}$ both parts are negative. The Schrödinger equation for this V_0 reads now

$$\left(-\frac{d}{dr}+\sqrt{-V(r)}\right)\left(\frac{d}{dr}+\sqrt{-V(r)}\right)\psi_0 = 0. \tag{21}$$

It is easily seen that its solutions φ_0 and χ_0 are given by

$$\chi_0(r) = \exp\left(-\int_0^r \sqrt{-V(t)}dt\right), \tag{22a}$$

$$\varphi_0(r) = \exp\left(-\int_0^r \sqrt{-V(u)}du\right)\int_0^r \exp\left(+2\int_0^t \sqrt{-V(u)}du\right)dt, \tag{22b}$$

so that

$$\varphi_0(r)\chi_0(r) = \int_0^r \exp\left[-2\int_t^r \sqrt{-V(u)}du\right]dt. \tag{23}$$

Using this in (19), and integrating by parts the term containing $\varphi_0\chi_0 d\sqrt{-V(r)}$, we get

$$n \leqslant \int_0^\infty (-V(r))^{1/2}dr, \quad V\leqslant 0, \quad V'\geqslant 0, \tag{24}$$

which is the desired result. Although (24) is not as good as (11a), it contains its main feature, which is that, for the Calogero–Cohn type potentials, the upper bound of $n(\lambda V)$ grows as $\lambda^{1/2}$ in the limit of λ becoming very large. But we did not use the nodal theorem.

More general bound: We can take, more generally,

$$V_0 = W_0^2 - W_0', \tag{25}$$

where W_0 is assumed to be positive and decreasing (non-increasing). Proceeding as before, we would then get

$$n \leqslant \int_0^\infty [W_0 - (V+W_0^2)\varphi_0\chi_0]dr, \tag{26}$$

where φ_0, χ_0, and $\varphi_0\chi_0$ are given by formulas identical to (22a), (22b), and (23) in which $\sqrt{-V}$ is replaced by W_0. Here, we have a formula in which V is given, and W_0 is arbitrary. One may think of setting up a variational principle for finding W_0 such that the integral in (26) is minimum. Unfortunately, the Euler equation one gets has no solution.

Oscillating negative potentials: Let us now first assume that the potential is of bounded variation. It is then known that such a function can be written as the difference of two positive and monotone functions.[13,14] Saying it differently,

$$V(r)=\tilde{V}_{+}(r)+\tilde{V}_{-}(r), \tag{27}$$

where $\tilde{V}_{+}$ is positive, $\tilde{V}_{-}$ is negative, and we have $\tilde{V}'_{+}\leqslant 0$ and $\tilde{V}'_{-}\geqslant 0$. Since $V(\infty)=0$, we also have $\tilde{V}_{\pm}(\infty)=0$. We can now choose our previous V_0 in (25) to be

$$\tilde{V}_0=-\tilde{V}_{-}-\frac{d}{dr}\sqrt{-\tilde{V}_{-}}(\geqslant 0). \tag{28}$$

We would then get

$$n(V)\leqslant\int_0^\infty[|\tilde{V}_{-}|^{1/2}-(\tilde{V}_{+})\tilde{\varphi}_{-}\tilde{\chi}_{-}]dr, \tag{29}$$

where $\tilde{\varphi}_{-}$ and $\tilde{\chi}_{-}$ are given by formulas similar to (22a) and (22b). *Remember that here* $\tilde{V}_{+}$ *and* $\tilde{V}_{-}$ *are not just simply the positive and negative parts of the potential*.

The bound (29) is somehow a generalization of (24) to the case of oscillating potentials. See also Ref. 11 for the generalization of (11a) for another class of oscillating potentials, as we mentioned earlier.

Remark: When the potential is not of bounded variation (too many oscillations around some points, going to $-\infty$ at $r=0$, etc.), it can often be the limit of such functions. If this is the case, and the limits $\tilde{V}_{-}$ and $\tilde{V}_{+}$ are such that the integrals on the right-hand side of (29) are finite, then (29) is valid. Note also that $\tilde{\varphi}_{-}$ and $\tilde{\chi}_{-}$ are explicitly given by (22a) and (22b). In conclusion, we have

Theorem: For negative potentials of total bounded variation on $[0, \infty)$, the number of bound states in the S wave satisfies the inequality (29). For negative potentials which satisfy (2), and are appropriate limits of potentials of bounded variation such that the two integrals on the right-hand side of (29) are each finite, the bound also holds.

III. IMPROVING THE BOUND (24)

So far, we have been writing

$$n\leqslant\sum_{\Lambda_j\geqslant 1}\bar{\Lambda}_j\leqslant\sum_{\text{all } j}\bar{\Lambda}_j, \tag{30}$$

where $\Lambda_j=1/\lambda_j$ are the eigenvalues of the positive kernels of our various integral equations, and then have calculated the last sum by the trace of the kernels. One way to improve on this is to stop at the middle sum. More specifically, to write

$$n\leqslant\sum_{\text{all } j}\bar{\Lambda}_j-\sum_{\Lambda_j<1}\bar{\Lambda}_j \tag{31}$$

and then try to find a lower bound for the second sum in (31), and use it there. If this can be done, one would then get an improvement of (30).

We are going to show that this can be done, at least for a class of potentials defined as follows. First, we have to remember that we had two sets of characteristic values, $\{\lambda_j\}$ for the original Schrödinger equation, and $\{\bar{\lambda}_j\}$ for the modified equation. And we had $\bar{\lambda}_j<\lambda_j$ for $j>n$, that is,

$$\Lambda_j<\bar{\Lambda}_j<1,\quad j>n. \tag{32}$$

Therefore, we can replace the second sum on the right-hand side of (31) by $\Sigma_{n+1}^{\infty}\Lambda_j$, and get still an upper bound for n. Now, assume that the potential V, which is purely attractive, satisfies $V'\geqslant 0$. We make now the Liouville transformation[15]

$$r\to Z=Z(r)=\int_0^r\sqrt{-V(t)}dt,$$

$$\overline{\varphi}(r)\to\psi(Z)=[-V(r)]^{1/4}\overline{\varphi}(r)|_{r=r(Z)}, \tag{33a}$$

in the Schrödinger equation (16) with (20):

$$\overline{\varphi}''=\left[(2\overline{\lambda}-1)V+(\overline{\lambda}-1)\frac{d}{dr}\sqrt{-V}\right]\overline{\varphi}. \tag{33b}$$

The change of variable $r\to Z$ is a good one-to-one and C^2 mapping from $[0,\infty)$ to $[0,I=\int_0^\infty\sqrt{-V(r)}dr<\infty]$. It is easily seen that the Schrödinger equation (33b), with $\overline{\varphi}(0)=0$, is transformed to[15]

$$\ddot{\psi}(Z)+(2\overline{\lambda}-1)\psi(Z)=\tilde{V}(Z)\psi(Z)+(\overline{\lambda}-1)\left[\frac{(\sqrt{-V})'}{-V}\right]\psi(Z),$$

$$\tilde{V}(Z)=\frac{\ddot{F}(Z)}{F(Z)},\quad F(Z)=[-V(r)]^{1/4}|_{r=r(Z)}, \tag{34}$$

$$Z\in[0,I],\quad \psi(0)=0,\quad I=\int_0^\infty\sqrt{-V(r)}dr.$$

The Liouville transformation does not change, of course, anything essential: self-adjointness of the new Hamiltonian is secured, and the characteristic values $\overline{\lambda}_j$, corresponding now to the eigenvalues $2\overline{\lambda}_j-1$, remain the same. As for the boundary condition at $Z=I$, we have to remember that for bound states in the r variable, the boundary condition at $r=\infty$ is $\overline{\varphi}'+\gamma\overline{\varphi}=0$, where $-\gamma^2$ is the energy of the bound state.[3] At zero energy, that is, at the thresholds $\overline{\lambda}_j$, the condition becomes simply[3]

$$\overline{\varphi}_j'(\infty)=0,\quad \overline{\varphi}_j(\infty)=\text{finite constant}. \tag{35}$$

Since $V(\infty)=0$, the boundary condition for $\psi(Z)$ is then

$$\psi(I)=0. \tag{36}$$

We have therefore the Dirichlet boundary condition at both ends of the interval for $\psi(Z)$.

For $\overline{\lambda}>1$, the second potential on the right-hand side of the differential equation (34) is negative. Now, assume that $\tilde{V}(Z)$ in (34) is also negative (attractive) everywhere in $[0,I]$. In the presence of both potentials, we have the eigenvalues of (34), $0<(2\overline{\lambda}_1-1)<(2\overline{\lambda}_2-1)<\cdots$. If we neglect these negative potentials, we get the ''free'' eigenvalues

$$2\overline{\lambda}_j^{(0)}-1=\frac{j^2\pi^2}{I^2},\quad j=1,2,... \tag{37}$$

and we know that $2\overline{\lambda}_j^{(0)}-1>2\overline{\lambda}_j-1$ for $j=n+1,...$. Therefore, we have the bound

$$\sum_{j=n+1}^{\infty} \frac{1}{\overline{\lambda_j^{(0)}}} \leqslant \sum_{j=n+1}^{\infty} \frac{1}{\overline{\lambda_j}}. \tag{38}$$

Using (37) on the left-hand side here, we get the sum (and its minorant integral)

$$2I^2 \sum_{n+1}^{\infty} \frac{1}{I^2 + j^2\pi^2} \geqslant 2I^2 \int_{n+1}^{\infty} \frac{dx}{I^2 + \pi^2 x^2}. \tag{39}$$

Using this to minorize the left-hand side of (38), we get ($\overline{\Lambda}_j = \overline{\lambda}_j^{-1}$)

$$\sum_{j=n+1}^{\infty} \overline{\Lambda}_j \geqslant \frac{2I}{\pi}\left[\frac{\pi}{2} - \mathrm{Arctg}\,\frac{\pi}{I}(n+1)\right]. \tag{40}$$

Using this now in (31), and remembering the trace (24), we get finally

$$n \leqslant \frac{2I}{\pi}\,\mathrm{Arctg}\left[\frac{\pi}{I}(n+1)\right], \tag{41}$$

which we have to solve to get a bound for n. Starting with $n \leqslant I$, and iterating (41), we end up, in fact, with an upper bound n_0 given by the solution of the equation

$$n \leqslant n_0 = \frac{2I}{\pi}\,\mathrm{Arctg}\left[\frac{\pi}{I}(n_0+1)\right], \tag{42}$$

which we have to solve for n_0. It can easily be seen that this bound is better than (24), but yet not as good as the Calogero–Cohn bound (11a).

It is also easily seen that a sufficient condition for $\tilde{V}(Z)$ to be purely attractive ($\leqslant 0$) is that $[-V(r)]^{-1/4}$ be a concave function of r (concave upward!). Of course, for such potentials, we can also improve the Bargmann bound. In any case, we have

Theorem: For any negative potential which satisfies (2), and is such that $G(r) = [-V(r)]^{-1/4}$ is a concave function (upward) of r, we have the upper bound $n \leqslant n_0$, where n_0 is given by (42). Note that, if $(-V)^{-1/4}$ is concave, we have $V' \geqslant 0$.

ACKNOWLEDGMENTS

One of the authors (K.C.) wishes to thank André Martin for useful correspondence, and another (R.K.) the LPTHE and IPN at Orsay for hospitality and financial support when this work was completed.

[1] S. Birman, Mat. Sb. **16**, 163 (1961).
[2] J. Schwinger, Proc. Natl. Acad. Sci. USA **47**, 122 (1961).
[3] R. G. Newton, *Scattering Theory of Waves and Particles*, 2nd ed. (Springer, Berlin, 1982).
[4] B. Simon, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms* (Princeton University Press, Princeton, 1971).
[5] V. Bargmann, Proc. Natl. Acad. Sci. USA **38**, 961 (1952).
[6] W. Thirring, *Quantum Mechanics of Atoms and Molecules* (Springer, Berlin, 1981).
[7] K. Chadan, Nuovo Cimento A **58**, 191 (1968).
[8] W. Frank, J. Math. Phys. **8**, 466 (1967).
[9] F. Calogero, Commun. Math. Phys. **1**, 80 (1965).
[10] J. H. E. Cohn, J. London Math. Soc. **40**, 523 (1965).
[11] K. Chadan, R. Kobayashi, A. Martin, and J. Stubbe, J. Math. Phys. **37**, 1106 (1996).
[12] K. Chadan, A. Martin, and J. Stubbe, J. Math. Phys. **36**, 1616 (1995).
[13] G. Valiron, *Théorie des Fonctions*, 2ème éd. (Masson et Cie, Paris, 1948).
[14] H. L. Royden, *Real Analysis*, 2nd ed. (Macmillan, New York, 1968).
[15] E. Hille, *Lectures on Ordinary Differential Equations* (Addison–Wesley, Reading, MA, 1969).

Energy levels of a quantal particle enclosed in *N* identical, mirror symmetric wells of a periodic potential. Numerical test of phase-integral formulas

Per Olof Fröman, Kilian Larsson, and Anders Hökback
Department of Theoretical Physics, University of Uppsala, Box 803, S-751 08 Uppsala, Sweden

(Received 24 March 1998; accepted for publication 2 December 1998)

An interesting structure prevails for the energy levels of a quantal particle in a periodic potential with N ($\geqslant 2$) mirror symmetric wells separated by $N-1$ mirror symmetric barriers, when the logarithmic derivative of the wave function is given at corresponding (periodically and mirror symmetrically situated) points in the barrier to the left of the first well and in the barrier to the right of the Nth well. It is shown that the quantization conditions that one obtains for these energy levels by means of a careful and rigorous phase-integral treatment are capable of giving extremely accurate results. The accuracy obtainable is demonstrated for $N=3$ by comparison with numerically exact results, which were obtained by means of the extended version of the phase-amplitude method presented in an Appendix. In the concluding section we summarize the results and point out unexpected features of the energy spectrum and the wave functions. Two different boundary conditions, commonly used in the theory of crystals, and closely related to the present investigation, are also discussed there. © *1999 American Institute of Physics.* [S0022-2488(99)02303-8]

I. INTRODUCTION

During several decades the solution of the one-dimensional Schrödinger equation has been the subject of numerous studies, both analytical and numerical. In spite of this fact, there are still problems of that kind, the solutions of which can give rise to new physical insight. One such problem concerns the energy levels of a quantal particle in a periodic potential subject to varying boundary conditions. Phase-integral formulas for the solution of this problem, derived by Fröman and Fröman (unpublished) on the basis of results in Chaps. 1 and 5 of Ref. 1, reveal an interesting structure of the energy spectrum and unexpected properties of the wave functions. Since it is the question of subtle effects, it seemed desirable to confirm the correctness of the phase-integral formulas and to test their accuracy by numerical calculations. It thereby turned out that the numerical phase-amplitude method,[2–6] though very powerful, had to be further extended as described in an Appendix. It is the purpose of this paper to present the phase-integral formulas along with results of numerical calculations for checking their accuracy, and to draw attention to the unexpected, and to our knowledge hitherto unknown, features of energy spectrum and wave functions that are revealed by the formulas.

II. PHASE-INTEGRAL QUANTIZATION CONDITIONS

Consider the one-dimensional Schrödinger equation

$$\frac{d^2\psi}{dx^2}+R(x)\psi=0, \tag{2.1}$$

where, with obvious notations,

0022-2488/99/40(4)/1764/16/$15.00

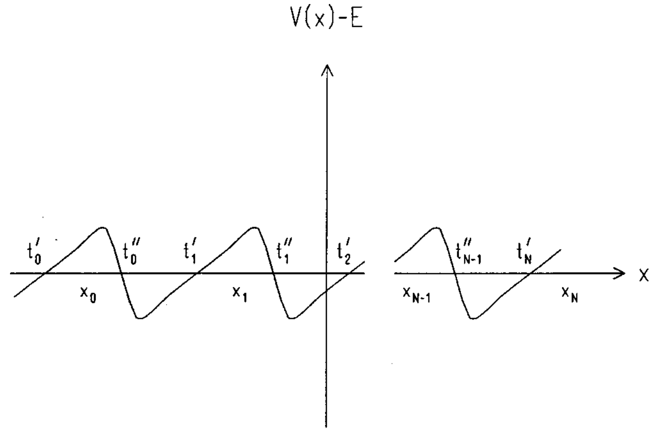

FIG. 1. The figure shows $V(x)-E$ as a function of x, where E is the energy of a quantal particle in a periodic potential $V(x)$ with N wells separated by $N-1$ barriers. The points x_0 and x_N are the points where the boundary conditions (2.3a) and (2.3b) are imposed on the wave function, and the points $x_1,\ldots,x_{N-1}$ correspond to the auxiliary points $z_1,\ldots,z_{N-1}$ that are used in the numerical phase-amplitude method described in Appendix B of the present paper. In the figure the wells and barriers are not mirror symmetric, although the treatment in this paper (except in the appendices) is restricted to the mirror symmetric case.

$$R(x)=\frac{2m}{\hbar^2}[E-V(x)]. \tag{2.2}$$

The potential $V(x)$ is assumed to be periodic and to consist of broad wells separated by barriers that are assumed to be mirror symmetric. Let x_0 be a point in a barrier, and let x_N ($N\geqslant 2$, $x_0<x_N$) be a point in another barrier such that between x_0 and x_N there are N wells separated by $N-1$ barriers; see Fig. 1. The points x_0 and x_N are assumed to be situated at the barrier maxima. Considering the wave function $\psi(x)$ in the interval $x_0\leqslant x\leqslant x_N$, we impose the boundary conditions

$$\psi'(x_0)/\psi(x_0)=k_0, \tag{2.3a}$$

$$\psi'(x_N)/\psi(x_N)=k_N. \tag{2.3b}$$

With the aid of the phase-integral approximation generated from an unspecified base function, described by Fröman and Fröman in Chap. 1 of Ref. 1, and also presented briefly in our Appendix A, Fröman and Fröman (unpublished) have derived quantization conditions for the problem described above.

For the case that $k_0=k_N=k$ the energy levels are obtained from the two quantization conditions

$$L=(s+1/2)\pi-\tilde{\phi}-\arcsin\frac{\cos(\nu\pi/N)}{[\exp(2K)+1]^{1/2}},$$

$$s=0,1,2,\ldots;\quad \nu=1,\ldots,N-1, \tag{2.4a}$$

$$L=(s+1/2)\pi-\tilde{\phi}+\arctan\left(\frac{k^2[1+\exp(-2K)]+|q(x_0)|^2}{k^2[1+\exp(-2K)]-|q(x_0)|^2}\exp(-K)\right),$$

$$s=0,1,2,\ldots, \tag{2.4b}$$

where (see Fig. 1)

$$L=\left|\int_{(t''_{n-1})}^{(t'_n)} q(z)dz\right|, \quad 1\leqslant n\leqslant N, \tag{2.5}$$

$$K=\left|\int_{(t'_n)}^{(t''_n)} q(z)dz\right|, \quad 1\leqslant n\leqslant N-1, \tag{2.6}$$

and $\tilde{\phi}$ is the function given by eq. (5.5.30), with ϕ replaced by $\tilde{\phi}$, $\lambda=1$ and $\ln(\overline{K_0}/\lambda)$ replaced by $\ln|\overline{K_0}|$, along with eqs. (5.5.25a–g), (5.4.21), and (5.4.23) in Chap. 5 of Ref. 1. In (2.5), as well as in (2.6), the integral, with the limits of integration within parentheses, is a simplified notation for half of the corresponding contour integral along a closed contour encircling the two limits of integration; see pp. 90–91 in Ref. 1. For the first order of the phase-integral approximation it is equal to an ordinary integral between the two limits of integration. The quantization condition (2.4a) has been derived under the assumption that the barriers are superdense, while the quantization condition (2.4b) has been derived under the assumption that the barriers are not only superdense but also thick. The accuracy of (2.4a) is expected to be greater than that of (2.4b) except when k approaches 0 or ∞. When $K(>0)$ is sufficiently large compared to unity, one can neglect $\exp(-2K)$ in (2.4b), getting

$$L=(s+1/2)\pi-\tilde{\phi}+\arctan\left(\frac{k^2+|q(x_0)|^2}{k^2-|q(x_0)|^2}\exp(-K)\right). \tag{2.4b$'$}$$

Numerical checks show that the accuracy of (2.4b′) is almost the same as that of (2.4b) even when K is as small as 2.7. When k is equal to zero or infinity, one can unify the two quantization conditions (2.4a) and (2.4b) into the single quantization condition

$$L=(s+1/2)\pi-\tilde{\phi}-\arcsin\frac{\cos(\nu\pi/N)}{[\exp(2K)+1]^{1/2}},$$
$$s=0,1,2,\ldots,\nu=0,\ldots,N-1 \ \text{ when } \ k=0, \quad \nu=1,\ldots,N \ \text{ when } \ k=\infty, \tag{2.7}$$

and for the validity of this quantization condition it is sufficient that the barriers are superdense. Since the potential is mirror symmetric, and the points x_0 and x_N are situated in a mirror symmetric way, the exact energy levels must be independent of the sign of k in the case when $k_0=k_N=k$. The approximate energy levels obtained from the quantization conditions (2.4a) and (2.4b), as well as (2.4a) and (2.4b′), also possess this property. These quantization conditions reveal an unexpected property of the energy spectrum and the corresponding eigenfunctions. For a given value of the quantum number s, the condition (2.4a) yields $N-1$ fine structure levels that depend on N but not on k, while the condition (2.4b), as well as (2.4b′), yields one *particular* level that depends on k but not on N. For each one of the $N-1$ *nonparticular* energy levels any solution of the differential equation (2.1) is an eigenfunction corresponding to a certain value of k. For the *particular* energy level there is a uniquely determined eigenfunction (except for a constant factor); it has the periodicity of the potential. When the absolute value of k increases from zero to infinity, one sees from (2.4b) or (2.4b′) that $L-(s+1/2)\pi+\tilde{\phi}$ decreases monotonically, and assumes for $|k|=0$ the approximate value $\pi-\exp(-K)$, then decreases slightly until $|k|$ approaches $|q(x_0)|[1+\exp(-2K)]^{-1/2}$, then decreases rapidly and assumes the value $\pi/2$ when $|k|$ is equal to $|q(x_0)|[1+\exp(-2K)]^{-1/2}$, then continues to decrease rapidly, and finally decreases slowly until $L-(s+1/2)\pi+\tilde{\phi}$ assumes the approximate value $\exp(-K)$ for $|k|=\infty$. The decrease of $L-(s+1/2)\pi+\tilde{\phi}$ from values slightly below π to values slightly above 0 takes place in an interval of $|k|$ values around $|q(x_0)|[1+\exp(-2K)]^{-1/2}$ that is small compared to $|q(x_0)|$.

For the case that $k_0=-k_N=k(>0)$ and $N\geqslant 3$, and the barriers are superdense and thick, the energy levels are obtained from the quantization condition

$$L=(s+1/2)\pi-\tilde{\phi}-\arcsin\Delta,\quad N\geqslant 3, \tag{2.8}$$

where Δ (small to its absolute value) is obtained from the equation (Fröman and Fröman, unpublished)

$$\left(\Delta+\frac{\kappa}{[\exp(2K)+1]^{1/2}}\right)\prod_{\mu=1}^{N-1}\left(\Delta-(-1)^s\frac{\cos(\mu\pi/N)}{[\exp(2K)+1]^{1/2}}\right)$$
$$-\frac{1-\kappa^2}{4[\exp(2K)+1]}\prod_{\mu=1}^{N-2}\left(\Delta-(-1)^s\frac{\cos[\mu\pi/(N-1)]}{[\exp(2K)+1]^{1/2}}\right)=0,\quad N\geqslant 3, \tag{2.9}$$

where

$$\kappa=\frac{k-|q(x_0)|}{k+|q(x_0)|}. \tag{2.10}$$

In the limit when $k\to 0$ or $k\to\infty$ one obtains the quantization condition (2.7) from (2.8), (2.9), and (2.10) by replacing in (2.9) μ by ν when s is even but by $N-\nu$ when s is odd. In these limits (2.9) is expected to give more accurate values of Δ than for other values of k. In the particular case when $N=3$ the quantum number s disappears from (2.9), and this equation has the three roots:

$$\Delta_1=\frac{(\kappa^2+8)^{1/2}-\kappa}{4[\exp(2K)+1]^{1/2}}\quad(>0), \tag{2.11a}$$

$$\Delta_2=-\frac{\kappa}{2[\exp(2K)+1]^{1/2}}, \tag{2.11b}$$

$$\Delta_3=-\frac{(\kappa^2+8)^{1/2}+\kappa}{4[\exp(2K)+1]^{1/2}}\quad(<0). \tag{2.11c}$$

It is easily seen that

$$\Delta_1>\Delta_2>\Delta_3. \tag{2.12}$$

When $k=0$ the root Δ_1, when inserted into (2.8), gives the particular state that one obtains from (2.7) for $\nu=0$. When $k=\infty$ the root Δ_3, when inserted into (2.8), gives the particular state that one obtains from (2.7) for $\nu=N=3$.

III. CHOICE OF POTENTIAL

To get information on the capability of the phase-integral quantization conditions presented in Sec. II to give accurate results it is convenient to choose the analytical expression for the potential $V(x)$ to be physically reasonable and to contain some parameters, so that the shape of the potential can be changed. By solving then the eigenvalue problem numerically as well as by means of the quantization conditions in Sec. II, and comparing the results thus obtained, one can get some information on the accuracy of the phase-integral eigenvalues associated with potentials of various shapes.

Our choice of potential is

$$V(x)=E_0\sum_{n=-\infty}^{+\infty}\exp\left[-\left(\frac{x-na}{b}\right)^2\right]. \tag{3.1}$$

When we introduce the dimensionless variable

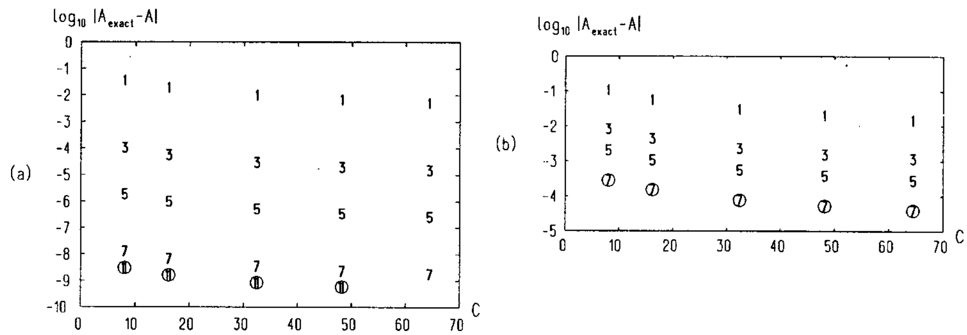

FIG. 2. In this figure $N=3$, $k_0=k_3=\infty$, $B=50$, $\nu=1$, and $K\approx 2.7$. In (a), where $\tilde{\phi}$ in (2.7) is retained, one obtains in the ninth-order approximation errors that are almost the same as those of the seventh-order approximation but of opposite sign. In (b), where $\tilde{\phi}$ in (2.7) is replaced by zero, one obtains in the ninth-order approximation errors that are almost the same as those of the fifth-order approximation but of opposite sign. In both (a) and (b) the quantum number s in (2.7) is $s=12$, 28, 60, 91, and 123 (corresponding to five different values of C).

$$z=\frac{x}{b} \tag{3.2}$$

and the dimensionless parameters

$$A=\frac{2mE_0b^2}{\hbar^2}, \tag{3.3a}$$

$$B=\frac{2mE_0b^2}{\hbar^2}, \tag{3.3b}$$

$$C=\frac{a}{b}, \tag{3.3c}$$

the Schrödinger equation (2.1) along with (2.2) and (3.1) becomes

$$\frac{d^2\psi}{dz^2}+\left\{A-B\sum_{n=-\infty}^{+\infty}\exp[-(z-nC)^2]\right\}\psi=0, \tag{3.4}$$

where we shall assume that

$$0<A<B, \tag{3.5a}$$

$$0<C. \tag{3.5b}$$

IV. NUMERICAL RESULTS

For a fixed value of B and various values of C we have calculated the eigenvalues A of the Schrödinger equation (3.4) by means of the quantization condition (2.7) with $N=3$ and $k=\infty$ for such values of s that K, defined by (2.6), remains approximately constant. In the present section the variable x in Sec. II is, of course, replaced by z; cf. (3.2). We have made the calculations with $\tilde{\phi}$ in (2.7) retained as well as replaced by zero. We have also calculated the corresponding exact eigenvalues A_{exact} by means of the phase-amplitude method described in Appendix B of the present paper. The values of $\log_{10}|A_{\text{exact}}-A|$ are shown for a fixed value of B and for various values of C in each one of Figs. 2 and 3. The digits in these figures indicate the orders of the phase-integral approximation. When there is a circle around such a digit, the quantity $A_{\text{exact}}-A$ is

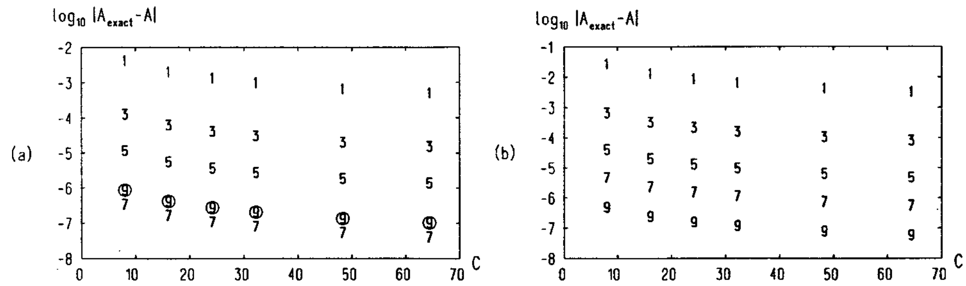

FIG. 3. In this figure $N=3$, $k_0=k_3=\infty$, $B=50$, $\nu=1$, and $K\approx 7.1$. In (a), where $\tilde{\phi}$ in (2.7) is retained, one obtains in the eleventh- and thirteenth-orders approximation errors that are almost the same as those of the ninth-order approximation and of the same sign. In (b), where $\tilde{\phi}$ in (2.7) is replaced by zero, one obtains in the 11th-order approximation errors that are only slightly smaller than those of the ninth-order approximation but of opposite sign, and in the 13th-order approximation one obtains errors that are almost the same as those of the ninth-order approximation but of opposite sign. In both (a) and (b) the quantum number s in (2.7) is $s=8$, 20, 32, 44, 68, and 92 (corresponding to six different values of C).

positive; otherwise $A_{\text{exact}}-A$ is negative. Each one of Figs. 2(a) and 3(a) shows the error when, in (2.7), $\tilde{\phi}$ is retained, while each one of Figs. 2(b) and 3(b) shows the error when, in (2.7), $\tilde{\phi}$ is replaced by zero. Since it is instructive to know how the wave function behaves for different values of ν, we have in Figs. 4 and 5 plotted the wave function (without caring about its normalization) for two different sets of values of B, C, s and the different possible values of ν. Note that, although $s=0$, the wave function in Fig. 4(b), as well as in Fig. 5(b), has a node in the middle potential well and is to its absolute value much smaller in this well than in the two adjacent wells.

For fixed values of B and C and values of $k_0=k_3=k$ ranging from -10 to $+10$ we have calculated energy eigenvalues A of the Schrödinger equation (3.4) by means of the numerically exact phase-amplitude method presented in Appendix B. The results are presented in Fig. 6.

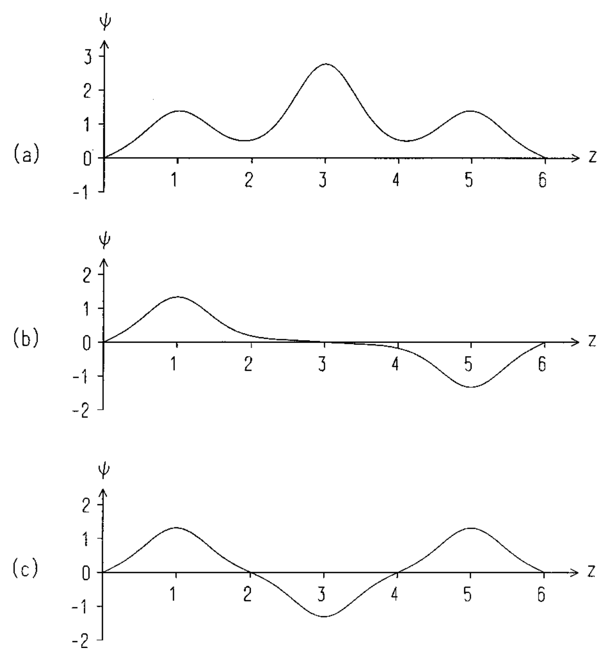

FIG. 4. In this figure $N=3$, $k_0=k_3=\infty$, $B=50$, $C=2.0034$, $s=0$, and $K=2.697$ in the fifth order of the phase-integral approximation. The classically allowed regions are $0.6<z<1.4$, $2.6<z<3.4$, and $4.6<z<5.4$. In (a) $\nu=1$ and $A_{\text{exact}}=41.939\,237\,827\,890$. In (b) $\nu=2$ and $A_{\text{exact}}=42.178\,416\,065\,134$. In (c) $\nu=3$ and $A_{\text{exact}}=42.312\,438\,121\,360$.

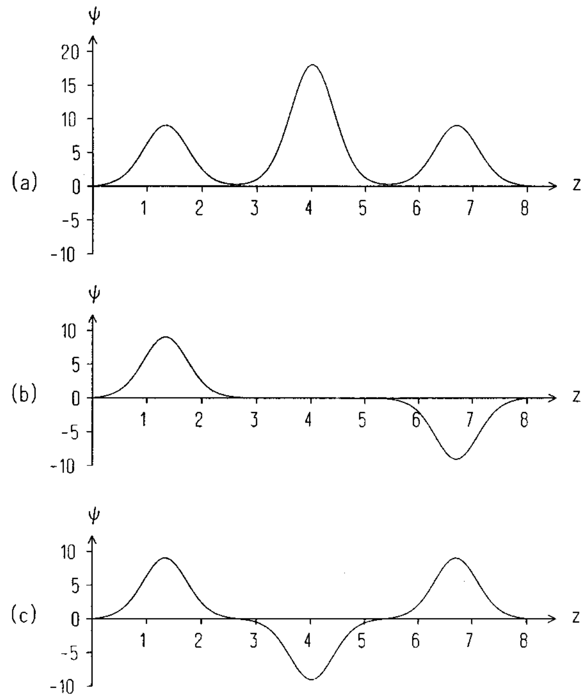

FIG. 5. In this figure $N=3$, $k_0=k_3=\infty$, $B=50$, $C=2.68055$, $s=0$, and $K=7.099$ in the fifth order of the phase-integral approximation. The classically allowed regions are $0.9<z<1.7$, $3.6<z<4.4$, and $6.3<z<7.1$. In (a) $\nu=1$ and $A_{\text{exact}}=22.845\,113\,870\,188$. In (b) $\nu=2$ and $A_{\text{exact}}=22.848\,428\,923\,516$. In (c) $\nu=3$ and $A_{\text{exact}}=22.850\,090\,112\,763$.

When $N=3$ and $k_0=-k_3=k>0$ we have for a fixed value of B and various values of C and k calculated the eigenvalues A of the Schrödinger equation (3.4) by means of the quantization condition (2.8) with $\tilde{\phi}$ retained and with Δ given by (2.11a) and (2.10) for such values of s that K, defined by (2.6), remains approximately constant. The phase-integral results thus obtained for the lowest fine structure level associated with the quantum number s under consideration have been compared with the corresponding exact energy levels obtained by means of the numerically exact phase-amplitude method presented in Appendix B. The absolute errors of the phase-integral values for the energy levels are presented graphically in Figs. 7 and 8. The digits in these figures indicate the orders of the phase-integral approximation. When there is a circle around such a digit, the quantity $A_{\text{exact}}-A$ is positive; otherwise $A_{\text{exact}}-A$ is negative. By recalling that the quantization conditions pertaining to $k=k_1=k_3$ and to $k=k_1=-k_3$ are the same in the limit $k\to\infty$, and by comparing Fig. 2(a) with Fig. 7 and Fig. 3(a) with Fig. 8, and by comparing also a number of figures analogous to these figures, but not presented in the present paper, we have found that even for very large values of k the figures may look quite different from those obtained in the limit $k\to\infty$.

V. CONCLUSIONS

For every quantum number s there are, in the case of $N(\geq 2)$ wells and $k_0=k_N=k$, in general N fine structure levels. Among these levels there is a *particular* one for which the energy is affected by the value of the logarithmic derivative k, but not by the number N of wells. The eigenfunction of this level is uniquely determined (except for a constant factor), and this eigenfunction has the same form in every well; only the amplitude is in general changed from one well to another. For the other $N-1$ *nonparticular* levels the energies are not affected by the value of the logarithmic derivative k, but by the number N of wells. For each one of these levels every solution of the Schrödinger equation is periodic (except sometimes for a change of sign) over the N wells. This unexpected behavior of the fine structure levels and the wave functions, which we have verified for $N=3$ (see Fig. 6), is to the best of our knowledge hitherto unknown. In certain

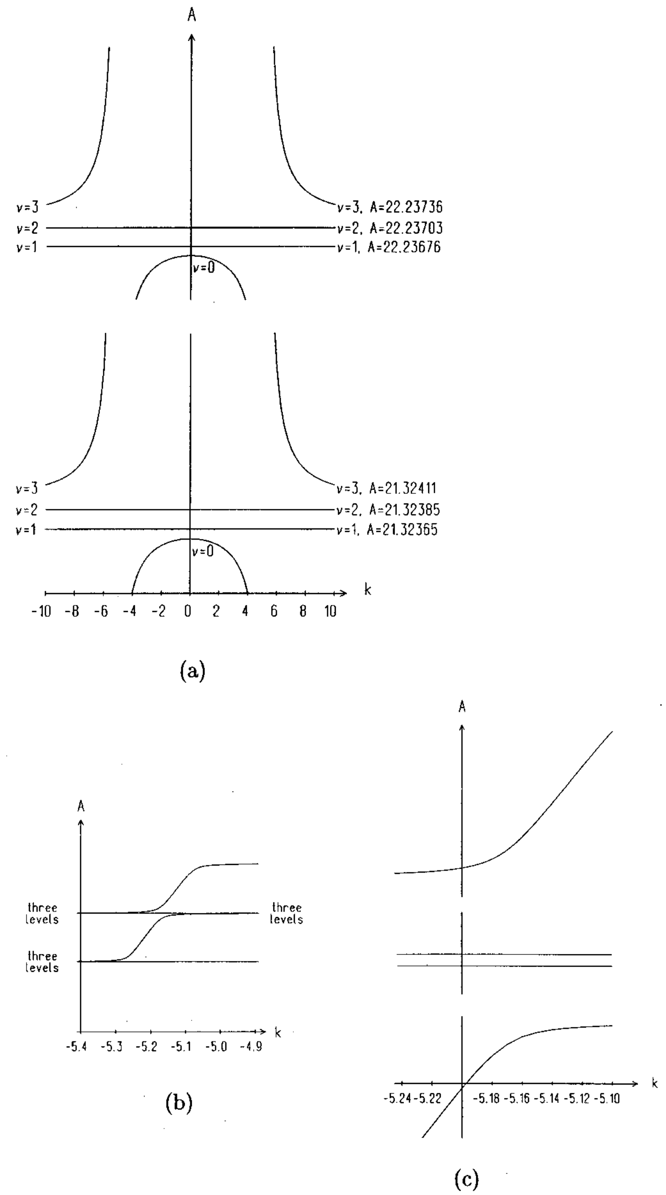

FIG. 6. In this figure $N=3$, $k=k_0=k_3$, $B=50$, and $C=32.6206$. The figure shows the dependence on k of the energy levels in two groups of fine structure levels which for $k=\infty$ have the quantum numbers $s=44$ and $s=45$ according to the phase-integral quantization condition (2.7) with $\nu=1$, 2, or 3. One more energy level is shown at the bottom of (a). The calculations have been performed by means of the numerically exact phase-amplitude method presented in Appendix B, but for $k=0$ and $k=\pm\infty$ the energy levels can be characterized by the quantum number ν in (2.7). In (a) the logarithmic derivative k assumes values in a large interval (in principle from $-\infty$ to $+\infty$). As mentioned above, the values of ν refer to $k=\pm\infty$ or $k=0$, but the values of A in the figure refer to $k=\pm 10$. Note that the vertical axis in (a) is interrupted. In (b) it is shown that A changes rapidly in a comparatively small interval of k-values. Note that the same eigenvalues A for two groups of fine structure levels are shown in (a) and (b) but that the scales of both horizontal and vertical axes differ strongly in (a) and (b). In (c), which is a magnification of a certain part of (b), we show the four highest-lying energy levels of (a) in a small interval of k values around $k=-5.17$. These energy levels form in this interval of k values a group of *four* close-lying levels (two of which are extremely close lying), which for $k=-5.17\,175$ have the energies $A=22.35\,084$, 22.23 703, 22.23 676, and 22.12 295. Note that the vertical axis in (c) is interrupted twice. The other energy levels of (a) form for $k=-5.17\,175$ a group of *three* still more close-lying levels (two of which are extremely close lying) with the energies $A=21.32\,385$, 21.32 365, and 21.31 252. The two groups of energy levels are clearly separated from each other.

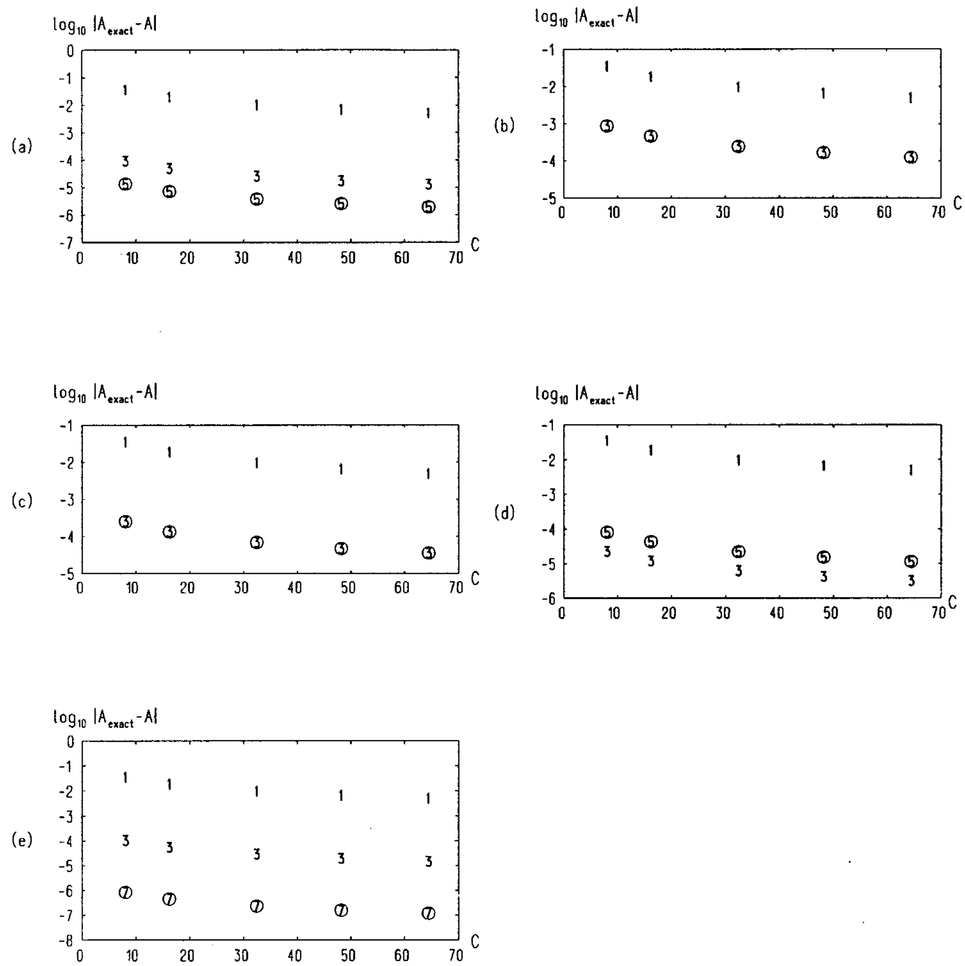

FIG. 7. This figure has reference to the lowest fine structure level associated with the quantum number s under consideration and the case when $N=3$, $k_0=-k_3$, $B=50$, and $K\approx 2.7$. The eigenvalues A have been obtained from (2.8) with $\tilde{\phi}$ retained and $s=12$, 28, 60, 91, and 123 (corresponding to five different values of C) along with (2.11a) and (2.10). In (a) $k_0=-k_3=10^{-2}$. In (b) $k_0=-k_3=2$. The errors of the fifth-order approximation are almost the same and of the same sign as those of the third-order approximation. In (c) $k_0=-k_3=20$. The third order of the phase-integral approximation is optimal, the fifth order being only slightly less accurate than the third order. In (d) $k_0=-k_3=100$. The errors of the seventh-order approximation are almost the same and of the same sign as those of the fifth-order approximation. In (e) $k_0=-k_3=10^4$. The seventh order of the phase-integral approximation is optimal. The errors of the fifth order are almost the same, but of opposite sign as those of the seventh-order approximation.

small intervals for the value of the logarithmic derivative in the boundary conditions there may be $N+1$ instead of, in general, N close-lying energy levels. We have also verified this, at a cursory glance unexpected, feature of the energy spectrum; see Fig. 6. In a single-well potential the wave function has s nodes in the well. It is, however, not common knowledge that in a multi-well potential there may exceptionally be a well in which there are $s+1$ nodes, if the absolute value of the wave function is much smaller in that well than in the two adjacent wells. We have verified this conclusion numerically for $N=3$ and $s=0$; see Figs. 4 and 5.

In the theory of crystals one uses boundary conditions implying either that the crystal is enclosed in a potential box with infinitely high walls or (according to Born and von Kármán) that the wave function (defined in the whole infinitely large space) has the periodicity of the whole crystal. For a one-dimensional crystal with N potential wells both these boundary conditions imply that the quantization condition is [cf. (2.7)]

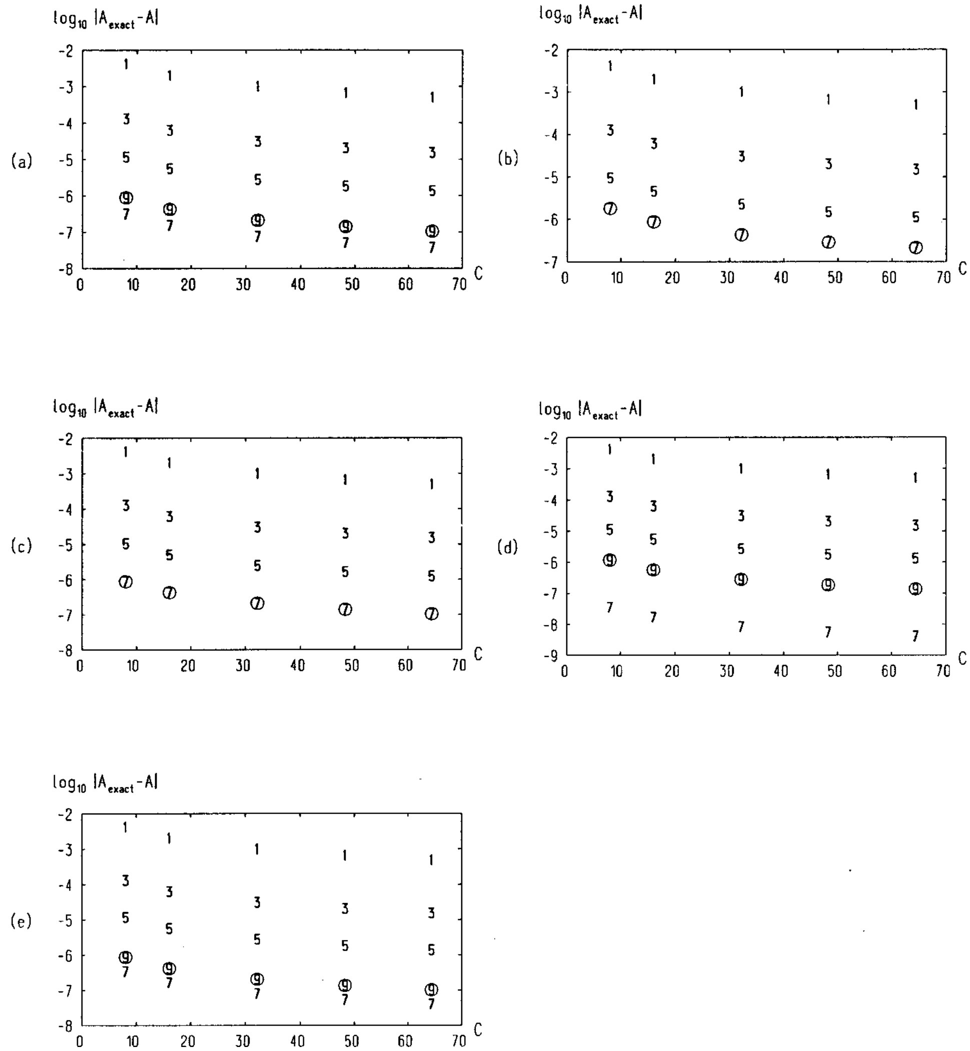

FIG. 8. This figure has reference to the lowest fine structure level associated with the quantum number s under consideration and the case when $N=3$, $k_0=-k_3$, $B=50$, and $K\approx 7.1$. The eigenvalues A have been obtained from (2.8) with $\tilde{\phi}$ retained and $s=8$, 20, 44, 68, and 92 (corresponding to five different values of C) along with (2.11a) and (2.10). In (a) $k_0=-k_3=10^{-2}$. In (b) $k_0=-k_3=2$. The seventh order of the phase-integral approximation is optimal. The errors of the ninth order are only slightly larger than and of the same sign as those of the seventh order. In (c) $k_0=-k_3=20$. The seventh order of the phase-integral approximation is optimal, and the errors of the ninth order are slightly larger than and of the same sign as those of the seventh order. In (d) $k_0=-k_3=100$. In (e) $k_0=-k_3=10^4$. The seventh order of the phase-integral approximation is optimal, and the errors of the 11th-order approximation are almost the same and of the same sign as those of the ninth-order approximation. As one should expect, this figure looks rather similar to Fig. 3(a); see, however, the comment at the end of Sec. IV.

$$L=(s+1/2)\pi-\tilde{\phi}-\arcsin\frac{\cos(\nu\pi/N)}{[\exp(2K)+1]^{1/2}}, \tag{5.1}$$

where for the boundary conditions associated with the potential box with infinitely high walls ν is *any* integer such that $1\leqslant\nu\leqslant N$, while for the boundary conditions of Born and von Kármán $0\leqslant\nu\leqslant N$ with ν even when s is even and $N-\nu$ even when s is odd. This shows the difference between the energy spectra associated with the two different boundary conditions in question. When the crystal is enclosed in a potential box with infinitely high walls, the wave function is uniquely determined (except for a constant normalization factor). When the Born–von Kármán boundary condition is imposed, the wave function is *uniquely* determined (except for a constant normalization factor) only for $\nu=0$ and $\nu=N$, while *every* solution of the Schrödinger equation has the periodicity of the whole crystal when $1\leqslant\nu\leqslant N-1$. For both boundary conditions now under consideration one finds, by comparing (2.4b) with (5.1), that the logarithmic derivative of the wave function at the center of each barrier is equal to zero when $\nu=0$ but equal to infinity when $\nu=N$. The discussion in the beginning of the present section of the particular and nonparticular energy levels applies, of course, also when one imposes the two boundary conditions now considered.

ACKNOWLEDGMENT

We are indebted to Professor Nanny Fröman who has read the manuscript and suggested improvements in the presentation.

APPENDIX A: PHASE-INTEGRAL APPROXIMATION GENERATED FROM AN UNSPECIFIED BASE FUNCTION

For a detailed description of the phase-integral approximation generated from an unspecified base function we refer to Chap. 1 in Ref. 1. A brief description is given below.

Approximate, but often very accurate, solutions of the differential equation

$$\frac{d^2\psi}{dz^2}+R(z)\psi=0, \tag{A1}$$

where $R(z)$ is an analytic function, can be obtained by means of the arbitrary-order phase-integral approximation generated from an unspecified base function. In this approximation there appears an unspecified function $Q(z)$ called the *base function*. This function is often chosen to be equal to $R^{1/2}(z)$, but in many physical problems it is important to use the possibility of choosing $Q(z)$ differently in order to achieve the result that the phase-integral approximation be valid close to certain exceptional points (e.g., the origin in connection with the radial Schrödinger equation), where the approximation would fail, if $Q(z)$ were chosen to be equal to $R^{1/2}(z)$. The function $Q(z)$ is in general chosen such that it is approximately equal to $R^{1/2}(z)$ except possibly in the neighborhood of the exceptional points.

To be able to write the phase-integral approximation in question in condensed form one introduces the new independent variable

$$\zeta=\int^{z}Q(z)\,dz \tag{A2}$$

and the function

$$\epsilon_0=Q^{-3/2}(z)\,\frac{d^2}{dz^2}\,Q^{-1/2}(z)+\frac{R(z)-Q^2(z)}{Q^2(z)}. \tag{A3}$$

It can be shown that in a local region of the complex z plane where the absolute value of ϵ_0 is small, the differential equation (A1) has the approximate solutions

$$\psi(z)=q^{-1/2}(z)\exp[\pm iw(z)], \tag{A4a}$$

$$w(z)=\int_{Z}^{z}q(z)\,dz, \tag{A4b}$$

where the lower limit of integration Z is an unspecified constant, and the function $q(z)$, pertaining to the phase-integral approximation of the order $2N+1$, is

$$q(z)=\sum_{n=0}^{N} q^{(2n+1)}(z), \tag{A5a}$$

$$q^{(2n+1)}(z)=Q(z)Y_{2n}, \tag{A5b}$$

with the first few functions Y_{2n} given by

$$Y_0=1, \tag{A6a}$$

$$Y_2=\tfrac{1}{2}\epsilon_0, \tag{A6b}$$

$$Y_4=-\frac{1}{8}\epsilon_0^2-\frac{1}{8}\frac{d^2\epsilon_0}{d\zeta^2}. \tag{A6c}$$

The choice of the function $Q(z)$ does not affect the expressions for Y_{2n} in terms of ϵ_0 and derivatives of ϵ_0 with respect to ζ, but only the expressions for ϵ_0 and ζ as functions of z.

It is an essential advantage of the phase-integral approximation, briefly described above, versus the Carlini (JWKB) approximation in higher order that the former approximation contains the unspecified base function $Q(z)$, which one can take advantage of in several ways. A criterion for the determination of the base function is that the function ϵ_0 be in some sense small in the region of the complex z plane relevant for the problem under consideration. However, this criterion does not determine the base function $Q(z)$ uniquely; it turns out that, within certain limits, the results are not very sensitive to the choice of $Q(z)$, when the approximation is used in higher orders. An inconvenient, but possible, choice of $Q(z)$ introduces in the first-order approximation an unnecessarily large error that is, however, in general corrected already in the third-order approximation. In many important cases the function $Q^2(z)$ can be chosen to be identical to $R(z)$. In other important cases, for instance when one wants to include the immediate neighborhood of a first- or second-order pole of $R(z)$ in the region of validity of the phase-integral approximation, the function $Q^2(z)$ is in general chosen to be similar to $R(z)$ except in the neighborhood of the pole. As regards a rather detailed discussion of the freedom that one has in the choice of the base function $Q(z)$, we refer to Chap. 1 in Ref. 1.

APPENDIX B: EXTENSION OF THE PHASE-AMPLITUDE METHOD FOR ACCURATE NUMERICAL SOLUTION OF THE ONE-DIMENSIONAL SCHRÖDINGER EQUATION

In this Appendix we present an extended version of the numerically exact phase-amplitude method, initiated by Milne,[2] Wilson,[3] Young,[4,5] and Wheeler,[6] to apply to the accurate numerical solution of the one-dimensional Schrödinger equation for a quantal particle in a potential consisting of N wells separated by $N-1$ superdense barriers.

Consider the differential equation

$$\frac{d^2\psi}{dz^2}+R(z)\psi=0, \tag{B1}$$

where, for real z values, $R(z)$ is positive in N regions that are separated by $N-1$ regions in which $R(z)$ is negative. The boundary conditions are assumed to be

$$\psi'(z_0)/\psi(z_0)=k_0, \tag{B2a}$$

$$\psi'(z_N)/\psi(z_N)=k_N, \tag{B2b}$$

where z_0 and $z_N(>z_0)$ lie in regions where $R(z)$ is negative and are separated by N regions where $R(z)$ is positive and $N-1$ regions where $R(z)$ is negative. We introduce in the regions where

$R(z)$ is negative further points $z_1,\ldots,z_{N-1}$ such that $z_0<z_1<\cdots<z_{N-1}<z_N$. These points correspond to the points $x_0,x_1,\ldots,x_{N-1}$, x_N in Fig. 1. For the eigenvalue problem considered in the present paper, the potential is periodic in the region to the right of x_0 (or z_0) and to the left of x_N (or z_N), but for the numerical phase-amplitude method to be described here this assumption need not be introduced, nor need we here introduce any assumption concerning symmetry.

To solve the differential equation (B1) numerically we put

$$\psi=F_n\hat{q}_n^{-1/2}(z)\sin\left(\int_{z_{n-1}}^{z}\hat{q}_n(z)dz\right)+G_n\hat{q}_n^{-1/2}(z)\cos\left(\int_{z_{n-1}}^{z}\hat{q}_n(z)dz\right),$$
$$z_{n-1}\leqslant z\leqslant z_n,\quad n=1,\ldots,N, \tag{B3}$$

where F_n and G_n are constants, and $\hat{q}_n(z)$ is to be determined numerically as a nonoscillating solution of the q equation

$$\frac{d^2}{dz^2}\hat{q}_n^{-1/2}+R(z)\hat{q}_n^{-1/2}=\hat{q}_n^{+3/2}, \tag{B4}$$

which one obtains by inserting (B3) into (B1) and requiring that the resulting equation be satisfied for all constant values of F_n and G_n. In passing we remark that the q equation appears not only in the phase-amplitude method[2–6] but also in connection with the Ermakov–Lewis invariant,[7,8] which is discussed in Chap. 1 of Ref. 1; see also Ref. 9. The continuity of $\psi(z)$ and $\psi'(z)$ at $z=z_n$, where $1\leqslant n\leqslant N-1$, gives the conditions

$$F_n\left[\hat{q}_n^{-1/2}(z)\sin\left(\int_{z_{n-1}}^{z}\hat{q}_n(z)dz\right)\right]_{z=z_n}+G_n\left[\hat{q}_n^{-1/2}(z)\cos\left(\int_{z_{n-1}}^{z}\hat{q}_n(z)dz\right)\right]_{z=z_n}$$
$$=F_{n+1}\left[\hat{q}_{n+1}^{-1/2}(z)\sin\left(\int_{z_n}^{z}\hat{q}_{n+1}(z)dz\right)\right]_{z=z_n}+G_{n+1}\left[\hat{q}_{n+1}^{-1/2}(z)\cos\left(\int_{z_n}^{z}\hat{q}_{n+1}(z)dz\right)\right]_{z=z_n} \tag{B5a}$$

and

$$F_n\left\{\frac{d}{dz}\left[\hat{q}_n^{-1/2}(z)\sin\left(\int_{z_{n-1}}^{z}\hat{q}_n(z)dz\right)\right]\right\}_{z=z_n}+G_n\left\{\frac{d}{dz}\left[\hat{q}_n^{-1/2}(z)\cos\left(\int_{z_{n-1}}^{z}\hat{q}_n(z)dz\right)\right]\right\}_{z=z_n}$$
$$=F_{n+1}\left\{\frac{d}{dz}\left[\hat{q}_{n+1}^{-1/2}(z)\sin\left(\int_{z_n}^{z}\hat{q}_{n+1}(z)dz\right)\right]\right\}_{z=z_n}$$
$$+G_{n+1}\left\{\frac{d}{dz}\left[\hat{q}_{n+1}^{-1/2}(z)\cos\left(\int_{z_n}^{z}\hat{q}_{n+1}(z)dz\right)\right]\right\}_{z=z_n}, \tag{B5b}$$

from which we obtain

$$F_{n+1}=\frac{\hat{q}'_{n+1}(z_n)/\hat{q}^2_{n+1}(z_n)-\hat{q}'_n(z_n)/\hat{q}^2_n(z_n)}{2\hat{q}^{1/2}_{n+1}(z_n)\hat{q}^{1/2}_n(z_n)}\left[F_n\sin\left(\int_{z_{n-1}}^{z_n}\hat{q}_n(z)dz\right)+G_n\cos\left(\int_{z_{n-1}}^{z_n}\hat{q}_n(z)dz\right)\right]$$
$$+\frac{\hat{q}^{1/2}_n(z_n)}{\hat{q}^{1/2}_{n+1}(z_n)}\left[F_n\cos\left(\int_{z_{n-1}}^{z_n}\hat{q}_n(z)dz\right)-G_n\sin\left(\int_{z_{n-1}}^{z_n}\hat{q}_n(z)dz\right)\right],\quad n=1,\ldots,N-1, \tag{B6a}$$

and

$$G_{n+1}=\frac{\hat{q}_{n+1}^{1/2}(z_n)}{\hat{q}_n^{1/2}(z_n)}\left[F_n \sin\left(\int_{z_{n-1}}^{z_n}\hat{q}_n(z)dz\right)+G_n \cos\left(\int_{z_{n-1}}^{z_n}\hat{q}_n(z)dz\right)\right],\quad n=1,...,N-1. \tag{B6b}$$

The boundary conditions (B2a) and (B2b) give

$$F_1\left\{\frac{d}{dz}\left[\hat{q}_1^{-1/2}(z)\sin\left(\int_{z_0}^{z}\hat{q}_1(z)dz\right)\right]\right\}_{z=z_0}+G_1\left\{\frac{d}{dz}\left[\hat{q}_1^{-1/2}(z)\cos\left(\int_{z_0}^{z}\hat{q}_1(z)dz\right)\right]\right\}_{z=z_0}$$
$$=k_0\left\{F_1\left[\hat{q}_1^{-1/2}(z)\sin\left(\int_{z_0}^{z}\hat{q}_1(z)dz\right)\right]_{z=z_0}+G_1\left[\hat{q}_1^{-1/2}(z)\cos\left(\int_{z_0}^{z}\hat{q}_1(z)dz\right)\right]_{z=z_0}\right\} \tag{B7a}$$

and

$$F_N\left\{\frac{d}{dz}\left[\hat{q}_N^{-1/2}(z)\sin\left(\int_{z_{N-1}}^{z}\hat{q}_N(z)dz\right)\right]\right\}_{z=z_N}+G_N\left\{\frac{d}{dz}\left[\hat{q}_N^{-1/2}(z)\cos\left(\int_{z_{N-1}}^{z}\hat{q}_N(z)dz\right)\right]\right\}_{z=z_N}$$
$$=k_N\left\{F_N\left[\hat{q}_N^{-1/2}(z)\sin\left(\int_{z_{N-1}}^{z}\hat{q}_N(z)dz\right)\right]_{z=z_N}+G_N\left[\hat{q}_N^{-1/2}(z)\cos\left(\int_{z_{N-1}}^{z}\hat{q}_N(z)dz\right)\right]_{z=z_N}\right\}. \tag{B7b}$$

From (B7a) and (B7b) we obtain

$$F_1=\left(\frac{k_0}{\hat{q}_1(z_0)}+\frac{\hat{q}_1'(z_0)}{2[\hat{q}_1(z_0)]^2}\right)G_1 \tag{B8a}$$

and

$$\int_{z_{N-1}}^{z_N}\hat{q}_N(z)dz=\hat{s}\pi+\arctan\frac{F_N-G_N(k_N/\hat{q}_N(z_N)+\hat{q}_N'(z_N)/2[\hat{q}_N(z_N)]^2)}{F_N(k_N/\hat{q}_N(z_N)+\hat{q}_N'(z_N)/2[\hat{q}_N(z_N)]^2)+G_N}, \tag{B8b}$$

where $\hat{s}$ is an integer.

To determine the functions $\hat{q}_n(z)$ one starts the integration of the differential equation (B4) in the middle of the actual potential well by using for $\hat{q}_n(z)$ a phase-integral expression of convenient order. When the functions $\hat{q}_n(z)$, $n=1,...,N$, have been calculated, we obtain from (B8a) F_1 expressed in terms of G_1 or G_1 expressed in terms of F_1, and from (B6a) and (B6b) we successively obtain $F_2,G_2;\ldots;F_N,G_N$. By inserting the values of F_N and G_N thus obtained into (B8b), we obtain a quantization condition from which the energy eigenvalues can be calculated by iteration.

There occur numerical difficulties when the potential is periodic and one tries to treat the *particular* fine structure level (corresponding to a given value of $\hat{s}$) by means of the general formulas that we have so far presented in this Appendix. One can overcome these difficulties by noting that the wave function of the *particular* fine structure level has the period of the potential and thus can be obtained if one puts $N=1$. Therefore we shall now give formulas that apply to $N=1$.

Particularization to the case when $N=1$

When $N=1$ we obtain from (B8a) and (B8b)

$$\int_{z_0}^{z_1}\hat{q}_1(z)dz=\hat{s}\pi+\arctan\frac{(k_0/\hat{q}_1(z_0)+\hat{q}_1'(z_0)/2[\hat{q}_1(z_0)]^2)-(k_1/\hat{q}_1(z_1)+\hat{q}_1'(z_1)/2[\hat{q}_1(z_1)]^2)}{(k_0/\hat{q}_1(z_0)+\hat{q}_1'(z_0)/2[\hat{q}_1(z_0)]^2)(k_1/\hat{q}_1(z_1)+\hat{q}_1'(z_1)/2[\hat{q}_1(z_1)]^2)+1}. \tag{B9}$$

For the further particularization that the function $R(z)$ is mirror symmetric, that the points z_0 and z_1 lie mirror symmetrically, and that the function $\hat{q}_1(z)$ is mirror symmetric [and hence $\hat{q}_1(z_0)=\hat{q}_1(z_1)$ and $\hat{q}_1'(z_0)=-\hat{q}_1'(z_1)$], the quantization condition (B9) simplifies to

$$\int_{z_0}^{z_1}\hat{q}_1(z)dz=\hat{s}\pi+\arctan\frac{(k_0-k_1)\hat{q}_1(z_0)+\hat{q}_1'(z_0)}{(k_0+\hat{q}_1'(z_0)/2\hat{q}_1(z_0))(k_1-\hat{q}_1'(z_0)/2\hat{q}_1(z_0))+[\hat{q}_1(z_0)]^2}. \tag{B10}$$

From (B10) we obtain when $k_1=k_0$

$$\int_{z_0}^{z_1}\hat{q}_1(z)dz=\hat{s}\pi+\arctan\frac{\hat{q}_1'(z_0)}{k_0^2-(\hat{q}_1'(z_0)/2\hat{q}_1(z_0))^2+[\hat{q}_1(z_0)]^2}, \tag{B11a}$$

and when $k_1=-k_0$

$$\int_{z_0}^{z_1}\hat{q}_1(z)dz=\hat{s}\pi+\arctan\frac{2k_0\hat{q}_1(z_0)+\hat{q}_1'(z_0)}{[\hat{q}_1(z_0)]^2-(k_0+\hat{q}_1'(z_0)/2\hat{q}_1(z_0))^2}. \tag{B11b}$$

We shall now show how approximate phase-integral quantization conditions can be obtained from the numerically exact phase-amplitude quantization conditions (B11a) and (B11b). When the barriers are thick, and the phase of the base function is chosen conveniently, we have approximately

$$\hat{q}_1(z_0)=-iq_1(z_0)\exp[-2iw_1(z_0)], \tag{B12}$$

where $q_1(z_0)$ is the asymptotic phase integrand given by (A5a) and (A5b), and $w_1(z_0)$ is the corresponding asymptotic phase integral (A4b) with the constant lower limit of integration for the first-order approximation chosen to lie at the right-hand turning point t_0'' of the barrier in which z_0 is situated. When a higher-order approximation is used, $w_1(z_0)$ is a related contour integral; see Eq. (4.3.3) in Ref. 1. Since the barrier in question is assumed to be mirror symmetric, with z_0 lying at its center, (B12) can be written as

$$\hat{q}_1(z_0)=|q_1(z_0)|\exp(-K), \tag{B12$'$}$$

with K given by (2.6). From (B12) we obtain approximately

$$\begin{aligned}\hat{q}_1'(z_0)&=-iq_1'(z_0)\exp[-2iw_1(z_0)]+2[-iq_1(z_0)]^2\exp[-2iw_1(z_0)]\\&=2[-iq_1(z_0)]^2\exp[-2iw_1(z_0)]\left[1+\frac{iq_1'(z_0)}{2[q_1(z_0)]^2}\right]\\&\approx 2[-iq_1(z_0)]^2\exp[-2iw_1(z_0)]\\&=2|q_1(z_0)|^2\exp(-K)\end{aligned} \tag{B13}$$

and hence, with the aid of (B12$'$),

$$\frac{\hat{q}_1'(z_0)}{2\hat{q}_1(z_0)}\approx|q_1(z_0)|. \tag{B14}$$

Inserting (B12′), (B13), and (B14) into (B11a) and (B11b), respectively, and noting that the barriers are assumed to be thick $[\exp(-2K)\ll 1]$, we obtain the *approximate* quantization conditions

$$\int_{z_0}^{z_1}\hat{q}_1(z)dz=\hat{s}\pi+\arctan\frac{2\exp(-K)}{|k_0/q_1(z_0)|^2-1+\exp(-2K)}$$

$$\approx\hat{s}\pi+\arctan\frac{2\exp(-K)}{|k_0/q_1(z_0)|^2-1},\quad k_1=k_0, \tag{B15a}$$

and

$$\int_{z_0}^{z_1}\hat{q}_1(z)dz=\hat{s}\pi-\arctan\frac{2\exp(-K)[k_0/|q_1(z_0)|+1]}{[k_0/|q_1(z_0)|+1]^2-\exp(-2K)}$$

$$\approx\hat{s}\pi-\arctan\frac{2\exp(-K)}{|k_0/q_1(z_0)|+1},\quad k_1=-k_0, \tag{B15b}$$

respectively. The numerically exact function $\hat{q}_1(z)$ is approximately the same as the asymptotic function $q_1(z)$ in the interior of the classically allowed region, but $\hat{q}_1(z)$ differs essentially from $q_1(z)$ in the classically forbidden regions; see (B12′). Putting

$$\int_{z_0}^{z_1}\hat{q}_1(z)dz=\int_{(t_0'')}^{(t_1')}q_1(z)dz+\pi/2+\tilde{\phi}-\exp(-K), \tag{B16a}$$

$$\hat{s}=s+1, \tag{B16b}$$

one can from (B15a) approximately obtain the phase-integral quantization condition (2.4b′), which is valid for $N=1$ when $k_1=k_0=k$, and one can from (B15b) approximately obtain

$$\int_{(t_0'')}^{(t_1')}q_1(z)dz=\left(s+\frac{1}{2}\right)\pi-\tilde{\phi}+\frac{k_0/|q_1(z_0)|-1}{k_0/|q_1(z_0)|+1}\exp(-K), \tag{B17}$$

which is the correct approximate phase-integral quantization condition when the barriers are thick and $N=1$ and $k_1=-k_0$. Because of the one-directional nature of the connection formulas, (B17) is expected to be valid only when $k_0\geqslant 0$.

[1] N. Fröman and P. O. Fröman, *Phase-Integral Method Allowing Nearlying Transition Points*, with adjoined papers by A. Dzieciol, N. Fröman, P. O. Fröman, A. Hökback, S. Linnaeus, B. Lundborg, and E. Walles, Springer Tracts in Natural Philosophy, edited by C. Truesdell (Springer-Verlag, New York, 1996).
[2] W. E. Milne, Phys. Rev. **35**, 863 (1930).
[3] H. A. Wilson, Phys. Rev. **35**, 948 (1930).
[4] L. A. Young, Phys. Rev. **38**, 1612 (1931).
[5] L. A. Young, Phys. Rev. **39**, 455 (1932).
[6] J. A. Wheeler, Phys. Rev. **52**, 1123 (1937).
[7] V. Ermakov, Univ. Izv. Kiev **20**(9), 1 (1880) (in Russian).
[8] H. R. Lewis, Jr., Phys. Rev. Lett. **18**, 510 (1967); **18**, 636 (1967).
[9] E. Pinney, Proc. Am. Math. Soc. **1**, 681 (1950).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Instability of a pseudo-relativistic model of matter with self-generated magnetic field

Marcel Griesemer[a),b)] and Christian Tix
Mathematik, Universität Regensburg, D-93040 Regensburg, Germany

(Received 23 November 1998; accepted for publication 23 November 1998)

For a pseudo-relativistic model of matter, based on the no-pair Hamiltonian, we prove that the inclusion of the interaction with the self-generated magnetic field leads to instability for all positive values of the fine structure constant. This is true no matter whether this interaction is accounted for by the Breit potential, by an external magnetic field which is chosen to minimize the energy, or by the quantized radiation field. © *1999 American Institute of Physics.* [S0022-2488(99)01203-7]

I. INTRODUCTION

The stability of matter problem concerns the questions whether the minimal energy of a system of particles is bounded from below (stability of the first kind), and whether it is bounded from below by a constant times the number of particles (stability of the second kind). Stability of the second kind for nonrelativistic quantum-mechanical electrons and nuclei was first proved in 1967 by Dyson and Lenard.[1,2] Since the new proofs of Lieb and Thirring, and Federbush in 1975 stability of matter is a subject of ongoing interest dealing with more and more realistic models of matter such as systems with a classical or quantized magnetic field included or with relativistic electrons (see Ref. 3 and the references therein). Stability with relativistic electrons is more subtle because of the uniform 1/length scaling behavior of the energy, which holds for massless particles (high particle–energy limit). Then the minimal energy is either non-negative or equal $-\infty$, so that stability of the second kind becomes equivalent to the statement that stability of the first kind holds for any given number of particles. We simply call this stability henceforth.

This paper is about a pseudo-relativistic model of matter which is stable, but which becomes unstable when the electrons are allowed to interact with the self-generated magnetic field. The self-generated magnetic field may be described using either an effective potential (the Breit-potential), an external magnetic field over which the energy is minimized, or the quantized radiation field. In all these cases we find instability for all positive values of the fine-structure constant. In contrast to most other models, where the collapse of the system, if it occurs, is due to the attraction of electrons and nuclei[4–7] (there would be no collapse without this interaction), the instability here is due to the attraction of parallel currents.

The model we study is based on a pseudo-relativistic Hamiltonian sometimes called the no-pair or Brown–Ravenhall Hamiltonian describing N relativistic electrons and K fixed nuclei interacting via Coulomb potentials. The electrons are vectors in the positive energy subspace of the free Dirac operator and their kinetic energy is described by this operator. For a physical justification of this model see the papers of Sucher,[8,9] for applications of the model in computational atomic physics and quantum chemistry; see Ishikawa and Koc[10,11] and Pyykkö.[12] The Brown–Ravenhall Hamiltonian yields stability for sufficiently small values of the fine-structure constant and the charge of the nuclei,[13–16,3] there are further rigorous results concerning the virial theorem[17] and eigenvalue estimates.[18]

We are interested in the minimal energy of this model when it is corrected to account for the interaction of the electrons with the self-generated magnetic field. This correction may be done,

---

[a)]Present address: Department of Mathematics, University of Alabama at Birmingham, Birmingham, Alabama 35294.
[b)]Electronic mail: marcel@uab.math.edu

for instance, by introducing a external magnetic field $\nabla\times\mathbf{A}$ to which the electrons are then minimally coupled and whose field energy is added to the energy of the system. The field **A** is now considered part of the system and hence the energy is to be minimized w.r.t. **A** as well. The minimizing **A** for a given electronic state is the self-generated one (to avoid instability for trivial reasons the gauge of **A** has to be fixed). The energy of this system is unbounded from below if $N\alpha^{3/2}$ is large, α being the fine structure constant, even if the vector potential is restricted to lie in a two parameter class $\{\gamma\mathbf{A}_0(\delta x):\gamma,\delta\in\mathbb{R}_+\}$ where $\mathbf{A}_0$ is fixed and obeys a weak condition requiring not much more than $\mathbf{A}_0\not\equiv 0$. This is our first main result. It extends a previous result of Lieb *et al.*[3] and is reminiscent of the fact that a static nonvanishing classical magnetic field in QED is not regular, in the sense that the dressed electron–positron emission and absorption operators do not realize a representation of the CAR on the Fock space of the free field.[19]

Alternatively the energy-shift due to the self-generated magnetic field may approximately be taken into account by including the Breit-potential in the energy. The resulting model is unstable as well. That is, the energy is unbounded from below if $N\alpha^{3/2}$ is large, no matter how small α is. This is our second main result. It concerns a Hamiltonian that is closely related the Dirac–Coulomb–Breit or Dirac–Breit Hamiltonian, which is the bases for most calculations of relativistic effects in many electron atoms.[8,12] We mention that for $\alpha=1/137$ the energy is bounded below if $N\leqslant 39$ and unbounded below if $N\geqslant 3.4\times 10^7$ (Theorem 4 and Theorem 3).

A third way of accounting for the self-generated field is to couple the electrons to the quantized radiation field. From a simple argument using coherent states (Lemma 5) it follows that the instability of this model is rather worse than the instability of the model first discussed.

As mentioned above the instability with the external magnetic field was previously found by Lieb *et al.*[3] Our result extends their result and our proof is simpler. The model with the Breit interaction corresponds to the classical system described by the Darwin Hamiltonian, which has been studied in the plasma physics literature (see Ref. 20 and the references therein). This classical model is thermodynamically unstable as well.[21]

In Secs. II, III, and IV we introduce the models with an external magnetic field, with the Breit-potential, and with quantized radiation field, and prove their instability (Theorem 1, Theorem 3, and Lemma 5). In Sec. III we also discuss dynamic nuclei for the model with Breit-potential. There is an appendix where numerical values for stability bounds on $N\alpha^{3/2}$ given in the main text are computed.

II. INSTABILITY WITH CLASSICAL MAGNETIC FIELD

We begin with the model of matter with an external magnetic field. For simplicity, the electrons are assumed to be noninteracting, and no nuclei are present. We could just as well treat a system of interacting electrons and static nuclei and would obtain essentially the same result (see Remark 4 below).

Consider a system of N noninteracting electrons in the external magnetic field $\nabla\times\mathbf{A}$. The energy of this system is

$$\mathcal{E}_N(\psi,\mathbf{A})=\left\langle \psi,\sum_{\mu=1}^{N} D_\mu(\mathbf{A})\psi\right\rangle+\frac{1}{8\pi}\int|\nabla\times\mathbf{A}(\mathbf{x})|^2d\mathbf{x},$$

where $D_\mu(\mathbf{A})$ is the Dirac operator $D(\mathbf{A})=\boldsymbol{\alpha}\cdot(-i\nabla+\alpha^{1/2}\mathbf{A}(\mathbf{x}))+\beta m$ acting on the μ-th particle, and the vector ψ, describing the state of the electrons, belongs to the Hilbert space,

$$\mathcal{H}_N=\bigwedge_{\mu=1}^{N}\Lambda_+L^2(\mathbb{R}^3,\mathbb{C}^4),$$

$$\Lambda_+=\chi_{(0,\infty)}(D(\mathbf{A}\equiv 0)),$$

or rather the dense subspace $\mathfrak{D}_N=\mathcal{H}_N\cap H^1[(\mathbb{R}^3\times\{1,...,4\})^N]$. That is, an electron is, by definition, a vector in the positive energy subspace of the free Dirac operator. We will always assume that the vector potential $\mathbf{A}$ belongs to the class $\mathcal{A}$ defined by the properties

$$(i)\quad \nabla\cdot\mathbf{A}=0,$$

$$(ii)\quad \mathbf{A}(\mathbf{x})\to 0,\quad \text{as } |\mathbf{x}|\to\infty,$$

$$(iii)\quad \int_{\mathbb{R}^3}|\nabla\times\mathbf{A}|^2<\infty.$$

Notice that $\mathcal{H}_N$ is not invariant under multiplication with smooth functions, in particular, it is not invariant under gauge transformations of the states. It follows that the minimal energy for fixed $\mathbf{A}$ is gauge-dependent. It can actually be driven to $-\infty$ by a pure gauge transformation (see Remark 3 below). To avoid this trivial instability we fixed the gauge of $\mathbf{A}$ by imposing conditions (i) and (ii).

The constants $\alpha>0$ and $m\geqslant 0$ in the definition of $D(\mathbf{A})$ are the fine-structure constant and the mass of the electron, respectively. In our units $\hbar=1=c$, so that $\alpha=e^2$ which is about 1/137 experimentally. We denote the Fourier transform of a function f by $\hat{f}$ or $\mathfrak{F}(f)$ and use $\mathbf{p}$ or $\mathbf{k}$ for its argument rather than $\mathbf{x}$ or $\mathbf{y}$. Our first result is the following.

Theorem 1: *Suppose $\mathbf{A}\in\mathcal{A}$ is such that $Re[\mathbf{e}\cdot\hat{\mathbf{A}}(\mathbf{p})]<0$ in $B(0,\epsilon)$ for some $\mathbf{e}\in\mathbb{R}^3$ and $\epsilon>0$. Then there exist a constant $C_\mathbf{A}$ such that for all $\alpha>0$, $m\geqslant 0$ and $N\geqslant C_\mathbf{A}\alpha^{-3/2}$,*

$$\inf_{\psi\in\mathfrak{D}_N,\|\psi\|=1,\gamma,\delta\in\mathbb{R}_+}\mathcal{E}_N(\psi,\gamma\mathbf{A}(\delta\mathbf{x}))=-\infty.$$

Remarks:

1. It is sufficient that $\mathbf{A}\in\mathcal{A}\cap L^1$ and $\int_{R^3}\mathbf{A}(\mathbf{x})d\mathbf{x}\neq 0$, since $\hat{\mathbf{A}}$ is then continuous and $\hat{\mathbf{A}}(0)\neq 0$. Thus, we have instability for virtually all nonvanishing $\mathbf{A}\in\mathcal{A}$.
2. The smallness of $N\alpha^{3/2}$ is not only necessary but also sufficient for stability (see Ref. 3, Sec. 4).
3. If the condition (ii) that $\mathbf{A}$ vanishes at infinity (and thus the gauge fixing) is dropped there is instability even for $N=1$ and the theorem becomes trivial. In fact, for $N=1$ and $\mathbf{A}(\mathbf{x})\equiv\mathbf{a}\neq 0$, $\mathcal{E}_{N=1}(\psi,\gamma\mathbf{A})=\langle\psi,D(0)\psi\rangle+\gamma\alpha^{1/2}\mathbf{a}\int\psi^+(\mathbf{x})\boldsymbol{\alpha}\psi(x)d\mathbf{x}$ which, as a function of γ, is unbounded from below for suitable $\psi\in\Lambda_+L^2(\mathbb{R}^3,\mathbb{C}^4)$.
4. The statement of the theorem also holds for the system of electrons and static nuclei with energy $\mathcal{E}_N(\psi,\mathbf{A})+\alpha\langle\psi,V_c\psi\rangle$ where
$$V_c:=-\sum_{\mu=1}^{N}\sum_{\kappa=1}^{K}\frac{Z_\kappa}{|\mathbf{x}_\mu-\mathbf{R}_\kappa|}+\sum_{\mu<\nu}^{N}\frac{1}{|\mathbf{x}_\mu-\mathbf{x}_\nu|}+\sum_{\kappa<\sigma}^{K}\frac{Z_\kappa Z_\sigma}{|\mathbf{R}_\kappa-\mathbf{R}_\sigma|},\tag{1}$$
if both N and $\Sigma_{\kappa=1}^K Z_\kappa$ are bigger than $C_\mathbf{A}\alpha^{-3/2}$ and if the energy is, in addition, minimized with respect to the pairwise distinct nuclear positions $\mathbf{R}_\kappa$. (See the proof of Theorem 3.)
5. Quantizing the radiation field does not improve the stability of the system (see Sec. IV).

The only way to restore stability we know is to replace $\mathcal{H}_N$ by the $\mathbf{A}$-dependent Hilbert space,

$$\mathcal{H}_{N,\mathbf{A}}=\bigwedge_{\mu=1}^{N}\chi_{(0,\infty)}(D(\mathbf{A}))L^2(\mathbb{R}^3,\mathbb{C}^4).$$

Obviously $\mathcal{E}_N(\psi,\mathbf{A})\geqslant 0$ for $\psi\in\mathcal{H}_{N,\mathbf{A}}$. In fact, even $\mathcal{E}_N(\psi,\mathbf{A})+\alpha\langle\psi,V_c\psi\rangle$ is non-negative for Z_κ and α small enough.[3]

Proof of Theorem 1: We will only work with Slater determinants and the following representation of one-particle orbitals. If $u\in L^2(\mathbb{R}^3;\mathbb{C}^2)$ then

$$\hat{\psi}(\mathbf{p})=\left(\frac{E(\mathbf{p})+m}{2E(\mathbf{p})}\right)^{1/2}\begin{pmatrix} u(\mathbf{p}) \\ \dfrac{\boldsymbol{\sigma}\cdot\mathbf{p}}{E(\mathbf{p})+m}u(\mathbf{p})\end{pmatrix}, \tag{2}$$

with $E(\mathbf{p})=\sqrt{\mathbf{p}^2+m^2}$, is the Fourier transform of a vector $\psi\in\Lambda_+L^2$, and the map $u\mapsto\psi$, $L^2(\mathbb{R}^3;\mathbb{C}^2)\to\Lambda_+L^2(\mathbb{R}^3;\mathbb{C}^4)$ is unitary.

It suffices to consider the case $m=0$ and find a Slater determinant $\psi=\psi_1\wedge\cdots\wedge\psi_N$ and γ, $\delta\in\mathbb{R}_+$ such that $\mathcal{E}_N(\psi,\gamma\mathbf{A}(\delta\mathbf{x}))<0$. In fact, by the scaling $\psi\mapsto\psi_\delta$, $\mathbf{A}\mapsto\mathbf{A}_\delta$ defined by $u_{\mu,\delta}=\delta^{-3/2}u_\mu(\delta^{-1}\mathbf{p})$ and $\mathbf{A}_\delta(\mathbf{x})=\delta\mathbf{A}(\delta\mathbf{x})$ we can then drive the energy with $m>0$ to $-\infty$ because $\mathcal{E}(\psi_\delta,\mathbf{A}_\delta,m)=\delta\mathcal{E}(\psi,\mathbf{A},m/\delta)$ and $\mathcal{E}(\psi,\mathbf{A},m/\delta)\to\mathcal{E}(\psi,\mathbf{A},m=0)$ for $\delta\to\infty$.

Choice of ψ: Let Q be the unit cube $\{\mathbf{p}\in\mathbb{R}^3|0\leq p_i\leq 1\}$, $u(\mathbf{p})=(\chi_Q(\mathbf{p}),0)^T$, and $\mathbf{e}\in\mathbb{R}^3$ an arbitrary unit vector. Set

$$u_\mu(\mathbf{p})=u(\mathbf{p}-\lambda N^{1/3}\mathbf{e}-\mathbf{n}_\mu),\quad \mu=1,...,N, \tag{3}$$

where λ is a positive constant to be chosen sufficiently large later on, and $(\mathbf{n}_\mu)_{\mu=1,\ldots,N}\subset\mathbb{Z}^3$ are the N lattice sites nearest to the origin, i.e., $\max_{\mu=1,\ldots,N}|\mathbf{n}_\mu|$ is minimal. We define $\psi=\psi_1\wedge\cdots\wedge\psi_N$ by

$$\hat{\psi}_\mu(\mathbf{p})=\frac{1}{\sqrt{2}}\begin{pmatrix} u_\mu(\mathbf{p}) \\ \boldsymbol{\sigma}\cdot\boldsymbol{\omega}_{\mathbf{p}}u_\mu(\mathbf{p})\end{pmatrix},\quad \boldsymbol{\omega}_{\mathbf{p}}=\frac{\mathbf{p}}{|\mathbf{p}|}, \tag{4}$$

which is (2) for $m=0$. Then $\psi\in\mathcal{H}_N$ and $\langle\psi_\mu,\psi_\nu\rangle=\langle u_\mu,u_\nu\rangle=\delta_{\mu\nu}$. Notice that

$$|\mathbf{p}-\lambda N^{1/3}\mathbf{e}|\leq N^{1/3},\quad \text{for all}\ \ \mathbf{p}\in\operatorname{supp}(u_\mu), \tag{5}$$

at least for large N (see the Appendix), i.e., in Fourier space all electrons are localized in a ball with radius $N^{1/3}$ and a distance from the origin which is large compared to the radius (since λ will be large).

Since $\psi=\psi_1\wedge\cdots\wedge\psi_N$ and $m=0$ we have

$$\mathcal{E}_N(\psi,\mathbf{A})=\sum_{\mu=1}^N\langle\psi_\mu,|\boldsymbol{\nabla}|\psi_\mu\rangle+\alpha^{1/2}\sum_{\mu=1}^N\int\mathbf{J}_\mu(x)\mathbf{A}(\mathbf{x})d\mathbf{x}+\frac{1}{8\pi}\int|\boldsymbol{\nabla}\times\mathbf{A}(\mathbf{x})|^2d\mathbf{x}, \tag{6}$$

where $\mathbf{J}_\mu(\mathbf{x})=\psi_\mu^*(x)\boldsymbol{\alpha}\psi_\mu(x)$. By definition of ψ_μ,

$$\hat{\mathbf{J}}_\mu(\mathbf{p})=\frac{1}{2}(2\pi)^{-3/2}\int u_\mu^*(\mathbf{k}-\mathbf{p})[\boldsymbol{\sigma}(\boldsymbol{\omega}_{\mathbf{k}}\cdot\boldsymbol{\sigma})+(\boldsymbol{\omega}_{\mathbf{k}-\mathbf{p}}\cdot\boldsymbol{\sigma})\boldsymbol{\sigma}]u_\mu(\mathbf{k})d\mathbf{k}. \tag{7}$$

Replace here u_μ by its defining expression and substitute $(\mathbf{k}-\lambda N^{1/3}\mathbf{e}-\mathbf{n}_\mu)\mapsto\mathbf{k}$. Since $\boldsymbol{\omega}_{\mathbf{k}+\lambda N^{1/3}\mathbf{e}+\mathbf{n}_\mu}\to\mathbf{e}$ as $\lambda\to\infty$, and since u has compact support, it follows that $\hat{\mathbf{J}}_\mu(\mathbf{p})$ converges to the current

$$\hat{\mathbf{J}}_0(\mathbf{p})=\mathbf{e}(2\pi)^{-2/3}\int u^*(\mathbf{k}-\mathbf{p})u(\mathbf{k})d\mathbf{k}, \tag{8}$$

as $\lambda\to\infty$. More precisely, $|\hat{\mathbf{J}}_\mu(\mathbf{p})-\hat{\mathbf{J}}_0(\mathbf{p})|\leq C\lambda^{-1}|\hat{\mathbf{J}}_0(\mathbf{p})|$ for $\lambda\geq\lambda_0$ where λ_0 and C are independent of μ and N. From $\hat{\mathbf{J}}_0(\mathbf{p})|\mathbf{p}|^{-1}$, $\hat{\mathbf{A}}(\mathbf{p})|\mathbf{p}|\in L^2$, it follows that

$$\int\hat{\mathbf{J}}_\mu^*(\mathbf{p})\hat{\mathbf{A}}(\mathbf{p})d\mathbf{p}=\int\hat{\mathbf{J}}_0^*(\mathbf{p})\hat{\mathbf{A}}(\mathbf{p})d\mathbf{p}+O(\lambda^{-1}),\quad \lambda\to\infty. \tag{9}$$

After a scaling $\mathbf{A}\mapsto\mathbf{A}_\delta$ we may assume that $\mathrm{Re}[\mathbf{e}\cdot\hat{\mathbf{A}}(\mathbf{p})]<0$ in the support of $\hat{\mathbf{J}}_0$ rather than in $B(0,\epsilon)$, so that (9) is bounded from above by some $-c_1<0$ for $\lambda\geqslant\lambda_0$ where c_1 and λ_0 are independent of μ and N. Observing finally that

$$\langle\psi_\mu,|\nabla|\psi_\mu\rangle=\int|\hat{\psi}_\mu(\mathbf{p})|^2|\mathbf{p}|d\mathbf{p}\leqslant(\lambda+1)N^{1/3}, \tag{10}$$

for all μ, we conclude that

$$\mathcal{E}_N(\psi,\gamma\mathbf{A})\leqslant(\lambda_0+1)N^{4/3}-\alpha^{1/2}\gamma Nc_1+\gamma^2c_2=(\lambda_0+1)N^{4/3}-\alpha\frac{c_1^2}{4c_2}N^2,$$

which is negative for $N\alpha^{3/2}$ large enough. At the end we inserted the optimal γ. □

The theorem has the obvious corollary.

Corollary 2: There is a constant C such that for all $\alpha>0$, $m\geqslant0$ and $N\geqslant C\alpha^{-3/2}$,

$$\inf_{\psi\in\mathfrak{D}_N,\|\psi\|=1;\mathbf{A}\in\mathcal{A}}\mathcal{E}_N(\psi,\mathbf{A})=-\infty.$$

This result is due to Lieb, Siedentop, and Solovej.[3]

Remark: It is sufficient that $C=1.4\times10^5$ or that $N\geqslant3.4\times10^7$ for $\alpha^{-1}=137$; see the Appendix.

To conclude this section we compute $\min_{\mathbf{A}\in\mathcal{A}}\mathcal{E}_N(\psi,\mathbf{A})$. This will provide a link to the instability with Breit-potential discussed in the next section. To exhibit the $\mathbf{A}$-dependence, we write the energy as

$$\mathcal{E}_N(\psi,\mathbf{A})=\mathcal{E}_N(\psi,\mathbf{A}\equiv0)+\alpha^{1/2}\int\mathbf{J}(\mathbf{x})\mathbf{A}(\mathbf{x})+\frac{1}{8\pi}\int|\nabla\times\mathbf{A}(\mathbf{x})|^2d\mathbf{x},$$

where $\mathbf{J}(\mathbf{x})$ is the probability current density associated with ψ. Its functional dependence on ψ is not crucial here. A straightforward computation shows that the Euler–Lagrange equation for $\mathbf{A}$ is $-\Delta\mathbf{A}=4\pi\alpha^{1/2}\mathbf{J}_T$ where $\mathbf{J}_T$ is the divergence-free—or transversal—part of $\mathbf{J}$. Comparing this equation with the Maxwell-equation for $\mathbf{A}$ in Coulomb gauge, which is $\Box\mathbf{A}=4\pi\alpha^{1/2}\mathbf{J}_T$, we find that the minimizing magnetic field is the self-generated one up to effects of retardation. Solving the Euler–Lagrange equation gives

$$\min_{\mathbf{A}\in\mathcal{A}}\mathcal{E}_N(\psi,\mathbf{A})=\mathcal{E}_N(\psi,\mathbf{A}\equiv0)-\frac{\alpha}{2}\int\frac{\mathbf{J}_T(\mathbf{x})\mathbf{J}_T(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{x}\,d\mathbf{y}. \tag{11}$$

III. INSTABILITY WITH BREIT-POTENTIAL

A. Static nuclei

We now consider a system of N (interacting) electrons in the external electric field of K static nuclei. There is no external magnetic field but a self-generated one which is approximately accounted for by the Breit-potential. The energy is now

$$\mathcal{E}_N(\psi,\mathbf{R})=\left\langle\psi,\left(\sum_{\mu=1}^{N}D_\mu+\alpha(V_c-B)\right)\psi\right\rangle, \tag{12}$$

where

$$B=\sum_{\mu<\nu}^{N}\frac{1}{2|\mathbf{x}_\mu-\mathbf{x}_\nu|}\left(\sum_i\alpha_{i,\mu}\otimes\alpha_{i,\nu}+\frac{\boldsymbol{\alpha}_\mu\cdot(\mathbf{x}_\mu-\mathbf{x}_\nu)\otimes\boldsymbol{\alpha}_\nu\cdot(\mathbf{x}_\mu-\mathbf{x}_\nu)}{|\mathbf{x}_\mu-\mathbf{x}_\nu|^2}\right), \tag{13}$$

and V_c is the Coulomb potential defined in (1). $\mathbf{R}$ denotes the K-tuple $(\mathbf{R}_1,\ldots,\mathbf{R}_K)$ of pairwise different nuclear positions, and $D_\mu = D_\mu(\mathbf{A}\equiv 0)$. As before, ψ belongs to $\mathfrak{D}_N \subset \mathcal{H}_N$. The interaction $-\alpha B$ is usually derived from the corresponding interaction in the Darwin Hamiltonian by the quantization $\mathbf{p}/m \mapsto \boldsymbol{\alpha}$[22] or from QED: treating the interactions of the electrons with the quantized radiation field in second order perturbation theory leads to a shift of the bound state energy levels approximately given by $-\alpha\langle\psi,B\psi\rangle$.[23] Important for our purpose is that

$$\langle\psi,B\psi\rangle + \begin{pmatrix}\text{self-energy \&}\\ \text{exchange terms}\end{pmatrix} = \frac{1}{2}\int \frac{\mathbf{J}_T(\mathbf{x})\mathbf{J}_T(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\, d\mathbf{x}\, d\mathbf{y}, \tag{14}$$

for any Slater determinant $\psi = \psi_1 \wedge \cdots \wedge \psi_N$ of orthonormal functions ψ_μ (see the proof of Theorem 3).

We are interested in the lowest possible energy,

$$E_{N,K} = \inf \mathcal{E}_N(\psi,\mathbf{R}),$$

where the infimum is taken over all $\psi \in \mathfrak{D}_N$ with $\|\psi\| = 1$ and all K-tuples $(\mathbf{R}_1,\ldots,\mathbf{R}_K)$ with $\mathbf{R}_j \neq \mathbf{R}_k$ for $j \neq k$. Our second main result is the following.

Theorem 3: There exists a constant C such that for all $\alpha > 0$, $m \geqslant 0$, $K \in \mathbb{N}$ and $Z_1,\ldots,Z_K \in \mathbb{R}_+$,

$$E_{N,K} = -\infty,$$

whenever N, $\sum_{\kappa=1}^K Z_\kappa \geqslant C \max(\alpha^{-3/2},1)$. *If* $\sum Z_\kappa^2 \geqslant 1$ it suffices that $C = 5\times 10^4$ or—when $\alpha^{-1} = 137$—that $N = \sum_{\kappa=1}^K Z_\kappa \geqslant 3.4\times 10^7$.

Remarks:

1. Similar to Section I, V_c and hence the condition on $\sum_{\kappa=1}^K Z_\kappa$ may be dropped. Then there is instability for $N \geqslant C \max(\alpha^{-3/2},1)$. It is for completeness of the model that we keep V_c in this section.
2. Without B the energy is proven to be non-negative if $\alpha Z_\kappa \leqslant 2/\pi$ for all κ and if $\alpha \leqslant 1/94$[6] (see also Ref. 3). One expects, however, stability, even for $\alpha Z_\kappa \leqslant 2(2/\pi + \pi/2)^{-1}, \alpha \leqslant 0.12$,[13,16] which would cover the atomic numbers of all known elements.

At least partly, this theorem can be understood from Corollary 2, Eq. (11) and Eq. (14).

Proof of Theorem 3: To begin with, we prove (14). Let $\psi = \psi_1 \wedge \cdots \wedge \psi_N$ with $\langle\psi_\mu,\psi_\nu\rangle = \delta_{\mu\nu}$ and let $\mathbf{J}(\mathbf{x}) = \sum_{\mu=1}^N \psi_\mu^+(x)\boldsymbol{\alpha}\psi_\mu(x)$ be the current density of ψ. Note that $\hat{J}_{T,i}(\mathbf{p}) = \sum_{j=1}^3 (\delta_{ij} - p_i p_j/p^2)\hat{J}_j(\mathbf{p})$ and that

$$\mathfrak{F}\frac{4\pi}{p^2}\left(\delta_{ij} - \frac{p_i p_j}{p^2}\right) = \frac{1}{2|x|}\left(\delta_{ij} + \frac{x_i x_j}{x^2}\right).$$

With $B(\mathbf{x})$ defined by

$$B(x) = \frac{1}{2|x|}\sum_{i,j} \alpha_i \left(\delta_{ij} + \frac{x_i x_j}{x^2}\right)\alpha_j = \frac{1}{2|x|}\left(\sum_i \alpha_i \otimes \alpha_i + \frac{\boldsymbol{\alpha}\cdot\mathbf{x} \otimes \boldsymbol{\alpha}\cdot\mathbf{x}}{|x|^2}\right),$$

it follows that

$$\frac{1}{2}\int \frac{\mathbf{J}_T(x)\mathbf{J}_T(y)}{|\mathbf{x}-\mathbf{y}|}\, d\mathbf{x}\, d\mathbf{y} = \frac{1}{2}\sum_{\mu,\nu} \langle\psi_\mu\otimes\psi_\nu, B(x-y)\psi_\mu\otimes\psi_\nu\rangle = \langle\psi,B\psi\rangle$$

$$+ \frac{1}{2}\sum_{\mu,\nu} \langle\psi_\mu\otimes\psi_\nu, B(x-y)\psi_\nu\otimes\psi_\mu\rangle, \tag{15}$$

which is Eq. (14). Similar as in the proof of Theorem 1 it suffices to consider the case $m=0$ and to find a Slater determinant $\psi=\psi_1\wedge\cdots\wedge\psi_N$ and nuclear positions $\mathbf{R}_1,\ldots,\mathbf{R}_K$ such that $\mathcal{E}_N(\psi,\mathbf{R})<0$.

Choice of the nuclear positions: A beautiful argument given in Ref. 3 shows that, after moving some electrons or nuclei far away from all others,

$$\langle\psi,V_c\psi\rangle\leqslant\epsilon+\frac{1}{2N^2}\sum_{\mu,\nu}\int\frac{|\psi_\mu(\mathbf{x})|^2|\psi_\nu(\mathbf{y})|^2}{|\mathbf{x}-\mathbf{y}|}\,d\mathbf{x}\,d\mathbf{y},$$

for suitably chosen nuclear positions. Here $\varepsilon>0$ is the (arbitrary small) contribution of the particles moved away. The second term can be dropped if $\Sigma_{\kappa=1}^K Z_\kappa^2\geqslant 1$. We use the inequality obtained in Ref. 14 to estimate it from above, and find

$$\langle\psi,V_c\psi\rangle\leqslant\epsilon+\mathrm{const}\frac{1}{N}\sum_{\mu=1}^N\langle\psi_\mu,D\psi_\mu\rangle. \tag{16}$$

The number N of remaining electrons obeys $N<\Sigma Z_\kappa+1$ which is the reason for the assumption on ΣZ_κ. Of course, the choice of the nuclear positions depends on ψ, which has not been specified yet.

Define one-particle orbitals ψ_μ and currents $\mathbf{J}_\mu$ and $\mathbf{J}_0$ exactly as in the proof of Theorem 1 with $\mathbf{e}$ being an arbitrary unit vector in $\mathbb{R}^3$. The convergence $\hat{\mathbf{J}}_\mu(\mathbf{p})\to\hat{\mathbf{J}}_0(\mathbf{p})$ as $\lambda\to\infty$ now implies that

$$\frac{1}{2}\int\frac{\mathbf{J}_T(\mathbf{x})\mathbf{J}_T(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\,d\mathbf{x}\,d\mathbf{y}=N^2\left[\frac{1}{2}\int\frac{\mathbf{J}_{0,T}(\mathbf{x})\mathbf{J}_{0,T}(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\,d\mathbf{x}\,d\mathbf{y}+O(\lambda^{-1})\right]\geqslant c_2N^2, \tag{17}$$

for $\lambda\geqslant\lambda_0$, where λ_0 and $c_2>0$ are independent of N.

To estimate the sum of exchange- and self-energy terms in (15), notice that

$$\langle\psi_\mu\otimes\psi_\nu,B(x-y)\psi_\nu\otimes\psi_\mu\rangle=\int\frac{4\pi}{p^2}|\hat{\mathbf{J}}_{\mu\nu,T}(\mathbf{p})|^2\,d\mathbf{p}, \tag{18}$$

where $\mathbf{J}_{\mu\nu}(\mathbf{x})=\psi_\mu^*(x)\boldsymbol{\alpha}\psi_\nu(x)$. After writing $\hat{\mathbf{J}}_{\mu\nu}(\mathbf{p})$ as an integral in Fourier space in terms of u_μ and u_ν similar as in (7), it is easily seen, using the support properties of u_μ and u_ν, that

$$|\hat{\mathbf{J}}_{\mu\nu,T}(\mathbf{p})|^2\leqslant|\hat{\mathbf{J}}_{\mu\nu}(\mathbf{p})|^2\leqslant 3(2\pi)^{-3}\chi(|\mathbf{p}+\mathbf{n}_\mu-\mathbf{n}_\nu|\leqslant\sqrt{3}). \tag{19}$$

The N balls $B(n_\nu,\sqrt{3})$, $\nu=1,\ldots,N$ all lie in the ball $B(0,N^{1/3})$ and cover a given point, at most, say, $4^3=64$ times (replace the balls by cubes with side $2\sqrt{3}$). Therefore (19) implies that

$$\sum_{\nu=1}^N|\hat{\mathbf{J}}_{\mu\nu}(\mathbf{p})|^2\leqslant 192(2\pi)^{-3}\chi(|\mathbf{p}+\mathbf{n}_\mu|<N^{1/3})\leqslant\frac{24}{\pi^3}\chi(|\mathbf{p}|<2N^{1/3}),$$

which, in conjunction with (18), gives

$$\frac{1}{2}\sum_{\mu,\nu}\langle\psi_\mu\otimes\psi_\nu,B(x-y)\psi_\nu\otimes\psi_\mu\rangle\leqslant\frac{384}{\pi}N^{4/3}. \tag{20}$$

Rewriting the energy using (15) and inserting the estimates (16), (10), (17), and (20) we arrive at

$$\mathcal{E}_N(\psi,\mathbf{R})\leqslant c_1(1+\alpha)N^{4/3}-c_2\alpha N^2,\quad c_2>0,$$

which is negative for $N>\mathrm{const}\max(\alpha^{-3/2},1)$. This proves the theorem.

For small N and small α there is stability. A similar result for the energy in Sec. I was proved in Ref. 3.

Theorem 4: Suppose $\tilde{\alpha}\leqslant 1/94$, $\max_\kappa Z_\kappa \leqslant 2/\pi\tilde{\alpha}^{-1}$ and $N-1\leqslant 2(2/\pi+\pi/2)(\alpha^{-1}-\tilde{\alpha}^{-1})$. Then $E_{N,K}\geqslant 0$. Inserting $\tilde{\alpha}=1/94$ and $\alpha=1/137$ we find stability for $N\leqslant 39$ and $\max Z_\kappa \leqslant 59$.

Proof: Since $B(x)\leqslant 2/|x|$ on $\mathbb{C}^4\otimes\mathbb{C}^4$ and $1/|x|\leqslant\delta^{-1}D$ on $\Lambda_+L^2(\mathbb{R}^3;\mathbb{C}^4)$ where $\delta=2(2/\pi+\pi/2)$,[14] one has, by the symmetry property of the states in $\mathcal{H}_N$,

$$B\leqslant\frac{N-1}{\delta}\sum_{\mu=1}^{N} D_\mu, \quad \text{on } \mathcal{H}_N. \tag{21}$$

Furthermore,

$$V_c\geqslant-\frac{1}{\tilde{\alpha}}\sum_{\mu=1}^{N} D_\mu, \quad \text{on } \mathcal{H}_N, \tag{22}$$

for all $\tilde{\alpha}>0$ with $\tilde{\alpha}\max Z_\kappa\leqslant 2/\pi$ and $\tilde{\alpha}q\leqslant 1/47$ by Ref. 6, where the number q of spin states may be set equal to 2.[3] Inserting (21) and (22) in the energy proves the theorem. □

B. Dynamic nuclei

Making the nuclei dynamical would improve stability if their kinetic energy were the only term we added to (12). However, if the nuclei are relativistic spin 1/2 particles like the electrons and if the Breit-potential couples all pairs of particles, taking their charges into account, then the instability will actually become worse.

Let us illustrate this for a system of N electrons and K identical nuclei of spin 1/2 and atomic number $Z>0$. These nuclei are described by vectors in the positive energy subspace of the free Dirac operator with the mass $M>0$ of the nuclei. To prove instability we adopt the strategy of the proof of Theorem 3 and thus assume $M=0$ and $m=0$. As a trial-wave function we take

$$\psi=(\psi_1\wedge\cdots\wedge\psi_N)\otimes(\phi_1\wedge\cdots\wedge\phi_K),$$

where ψ_μ is defined by Eqs. (3) and (4) and ϕ_κ is defined like ψ_κ, except that $\mathbf{e}$ and N are replaced by $-\mathbf{e}$ and K, respectively. It follows that in the limit $\lambda\to\infty$ we get $N+K$ (charge-) currents, the nuclear ones being larger than the electronic ones by a factor of Z but otherwise identical. The Breit interaction thus gives a negative contribution to the energy of order $\alpha(N+ZK)^2$. While the parallel currents of the $N+K$ particles add up, the opposite charges of the electrons and nuclei cancel themselves. In fact, for ψ defined as above,

$$\begin{aligned}\langle\psi,V_c\psi\rangle&\leqslant\sum_{\mu<\nu}^{N}\int d\mathbf{x}\,d\mathbf{y}\frac{|\psi_\mu(\mathbf{x})|^2|\psi_\nu(\mathbf{y})|^2}{|\mathbf{x}-\mathbf{y}|}+Z^2\sum_{\kappa<\sigma}^{K}\int d\mathbf{R}_1\,d\mathbf{R}_2\frac{|\phi_\kappa(\mathbf{R}_1)|^2|\phi_\sigma(\mathbf{R}_2)|^2}{|\mathbf{R}_1-\mathbf{R}_2|}\\&\quad-Z\sum_{\kappa=1}^{K}\sum_{\mu=1}^{N}\int d\mathbf{x}\,d\mathbf{R}\frac{|\psi_\mu(\mathbf{x})|^2|\phi_\kappa(\mathbf{R})|^2}{|\mathbf{x}-\mathbf{R}|}\\&=\left[\frac{N(N-1)}{2}+Z^2\frac{K(K-1)}{2}-NKZ\right](I+O(\lambda^{-1}))\\&=[(KZ-N)^2-KZ^2-N](I/2+O(\lambda^{-1})),\end{aligned}\tag{23}$$

where I is the limit of the above double integrals as $\lambda\to\infty$. Hence, $\langle\psi,V_c\psi\rangle$ is negative, e.g., if $KZ=N$ and λ is large. To achieve this in the static case we had to choose the nuclear positions properly. It is instructive to recall how this was done. The total energy is bounded from above by $c_1(N^{4/3}+K^{4/3})-c_2\alpha(N+KZ)^2$, $c_2>0$, for $N=KZ$ and λ large, and is therefore negative for $N=KZ$ large enough.

IV. STABILITY AND INSTABILITY WITH QUANTIZED RADIATION FIELD

Instability for the model with a classical magnetic field implies instability for the model with a quantized radiation field without UV-cutoff. In fact, for each classical magnetic field there is a coherent state of photons which reproduces the classical field as far as the energy is concerned. If an UV-cutoff is introduced the relativistic scale invariance of the energy is broken and stability of the first kind is restored. The lower bound depends on the cutoff and goes to $-\infty$ as the cutoff is removed.

The state of the system is now described by a vector $\Psi \in \mathcal{H}_N \otimes \mathcal{F}$ where $\mathcal{F}$ denotes the bosonic Fock-space over $L^2(\mathbb{R}^3)\otimes\mathbb{C}^2$, the factor $\mathbb{C}^2$ accounting for the two possible polarizations of the transversal photons, and the total energy of Ψ is

$$\mathcal{E}_N^{\mathrm{qed}}(\Psi) = \left\langle \Psi, \sum_{\mu=1}^{N} [\boldsymbol{\alpha}_\mu \cdot (-i\boldsymbol{\nabla}_\mu + \alpha^{1/2}\mathbf{A}(\mathbf{x}_\mu)) + \beta_\mu m]\Psi \right\rangle + \langle \Psi, (1\otimes H_f)\Psi\rangle,$$

$$H_f = \sum_{\lambda=1}^{2} \int d\mathbf{k}|\mathbf{k}| a_\lambda^\dagger(\mathbf{k}) a_\lambda(\mathbf{k}),$$

where

$$\mathbf{A}(\mathbf{x}) := \sum_{\lambda=1}^{2} \int dk [e_\lambda(\mathbf{k}) e^{i\mathbf{kx}} \otimes a_\lambda(\mathbf{k}) + \mathbf{e}_\lambda(\mathbf{k}) e^{-i\mathbf{kx}} \otimes a_\lambda^\dagger(\mathbf{k})] =: \mathbf{A}^+(\mathbf{x}) + \mathbf{A}^+(\mathbf{x})^*$$

is the quantized vector potential in the Coulomb gauge. The operators $a_\lambda(\mathbf{k})$ and $a_\lambda^\dagger(\mathbf{k})$ are creation and annihilation operators acting on $\mathcal{F}$ and obeying the CCR,

$$[a_\lambda(\mathbf{k}_1), a_\mu^\dagger(\mathbf{k}_2)] = \delta_{\lambda\mu}\delta(\mathbf{k}_1 - \mathbf{k}_2), \quad [a_\lambda^\#(\mathbf{k}_1), a_\mu^\#(\mathbf{k}_2)] = 0,$$

where $a_\lambda^\sharp = a_\lambda$ or $a_\lambda^\dagger$, and the two polarization vectors $\mathbf{e}_\lambda(\mathbf{k})$ are orthonormal and perpendicular to $\mathbf{k}$ for each $\mathbf{k}\in\mathbb{R}^3$. We use dk as a short hand for $(2\pi)^{-3/2}(2|\mathbf{k}|)^{-1/2} d\mathbf{k}$, and the subindex of $\boldsymbol{\alpha}_\mu$, $\boldsymbol{\nabla}_\mu$ and β_μ indicates that these one-particle operators act on the μ-th particle. While we used Gaussian units in Secs. II and III we now work with Heaviside Lorenz units.

Lemma 5: For each $\mathbf{A}_{cl} \in \mathcal{A}\cap L^2(\mathbb{R}^3)$ there exists a vector $\theta\in\mathcal{F}$ (coherent state) such that

$$\mathcal{E}_N^{\mathrm{qed}}(\psi\otimes\theta) = \mathcal{E}_N(\psi, \mathbf{A}_{cl}),$$

for all $\psi \in \mathfrak{D}_N$.

Proof: Pick $\mathbf{A}_{cl}\in\mathcal{A}\cap L^2(\mathbb{R}^3)$ and define $\eta_\lambda(\mathbf{k}) = (|\mathbf{k}|/2)^{1/2}\mathbf{e}_\lambda(\mathbf{k})\cdot\hat{\mathbf{A}}_{cl}(\mathbf{k})$ so that $\mathbf{A}_{cl}(\mathbf{x}) = \mathbf{A}_{cl}^+(\mathbf{x}) + \mathbf{A}_{cl}^+(\mathbf{x})^*$ with

$$\mathbf{A}_{cl}^+(\mathbf{x}) = \sum_{\lambda=1}^{2} \int dk\ \eta_\lambda(\mathbf{k})\mathbf{e}_\lambda(\mathbf{k}) e^{i\mathbf{kx}}. \tag{24}$$

Next, set

$$\Pi(\eta) := i\sum_{\lambda=1}^{2} \int d\mathbf{k}[\eta_\lambda(\mathbf{k}) a_\lambda^\dagger(\mathbf{k}) - \overline{\eta_\lambda(\mathbf{k})} a_\lambda(\mathbf{k})],$$

and $\Theta = e^{-i\Pi(\eta)}\Omega \in \mathcal{F}$. Θ is called a coherent state, it is normalized, and, most importantly, it is an eigenvector of all annihilation operators,

$$a_\lambda(\mathbf{k})\Theta = \eta_\lambda(\mathbf{k})\Theta. \tag{25}$$

From (24), (25), and the definition of $\eta_\lambda(\mathbf{k})$ it follows that

$$\boldsymbol{\alpha}_\mu \mathbf{A}^+(\mathbf{x}_\mu)\psi\otimes\Theta=(\boldsymbol{\alpha}_\mu \mathbf{A}^+_{cl}(\mathbf{x}_\mu)\otimes\mathbf{1})\psi\otimes\Theta$$

and

$$\langle\Theta,H_f\Theta\rangle=\int d\mathbf{k}|\mathbf{k}|\sum_\lambda |\eta_\lambda(\mathbf{k})|^2=\frac{1}{2}\int d\mathbf{k}\,k^2|\hat{\mathbf{A}}_{cl}(\mathbf{k})|^2.$$

Inserting this in the energy proves the theorem. □

If an ultraviolet cutoff is introduced in the field operator $\mathbf{A}(\mathbf{x})$ then stability of the first kind is restored for all N and a certain range of values for α and Z_κ. This follows from Ref. 24, Lemma I.5 and Ref. 3, Theorem 1.

ACKNOWLEDGMENTS

It is a pleasure to thank Heinz Siedentop for many discussions, and Arne Jensen, Jan Philip Solovej, and Erik Skibsted for the hospitality at Aarhus University in August 97, where this work was begun. M. G. also thanks Michael Loss for clarifying discussions. This work was partially supported by the European Union under Grant No. ERB4001GT950214 and under the TMR-network Grant No. FMRX-CT 96-0001.

APPENDIX: NUMERICAL ESTIMATES

To obtain the numerical values for the constants in Corollary 2 and Theorem 3 we follow the proof of Theorem 3 up to a few modifications, and explicitly evaluate the constants in this proof.

The main modifications are that the two-spinor u is now defined in terms of the (normalized) characteristic function of the ball with radius 1/2 contained in the unit cube $\{\mathbf{p}|0\leqslant p_i\leqslant 1\}$ and that the 4-spinors $\psi_{2\mu-1}$ are defined in terms of the $\psi_{2\mu}$'s by interchanging the components of u, while $\mathbf{n}_{2\mu-1}$ runs over the $N/2$ or—if N is odd—the $(N+1)/2$ lattice sites of $\mathbb{Z}^3$ closest to the origin. The balls simplify the computation of $\hat{\mathbf{J}}_0(\mathbf{p})$ and the double occupation $\mathbf{n}_{2\mu-1}=\mathbf{n}_{2\mu}$ reduces the kinetic energy. To begin with, we note that the n unit cubes of the lattice $\mathbb{Z}^3$ which are closest to the origin, all fit in a ball of radius

$$n^{1/3}\left(\frac{3}{4\pi}\right)^{1/3}+\sqrt{3}.$$

In particular, the $N/2$ or $(N+1)/2$ unit cubes containing the supports of the spinors ψ_μ, $\mu=1,...,N$ all lie in the ball of radius $bN^{1/3}$ centered at $\lambda N^{1/3}\mathbf{e}$ where $b=1/2$ if $N\geqslant 1.2\times 10^7$, $b=3/5$ if $N\geqslant 5\times 10^3$ and $b=\sqrt{3}$ if $N\geqslant 1$ (the ball of radius $\sqrt{3}n^{1/3}$ contains never less than n lattice cubes). This replaces Eq. (4) and implies, together with Eqs. (7) and (8), that

$$|\hat{\mathbf{J}}_\mu(\mathbf{p})-\hat{\mathbf{J}}_0(\mathbf{p})|\leqslant\frac{6b}{\lambda-b}|\hat{\mathbf{J}}_0(\mathbf{p})|,\quad \lambda>b.$$

Using this and $|\hat{\mathbf{J}}_0(\mathbf{p})|=1/2(2\pi)^{-3/2}(1-p)^2(2+p)$ one finds

$$\begin{aligned}\int\frac{\mathbf{J}_T(\mathbf{x})\mathbf{J}_T(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{x}\,d\mathbf{y}&=\sum_{\mu,\nu=1}^N\int d\mathbf{p}\frac{4\pi}{\mathbf{p}^2}\hat{\mathbf{J}}^*_\mu(\mathbf{p})T\hat{\mathbf{J}}_\nu(\mathbf{p})\\&\geqslant N^2\left[\int d\mathbf{p}\frac{4\pi}{p^2}\hat{\mathbf{J}}^*_0(\mathbf{p})T\hat{\mathbf{J}}_0(\mathbf{p})-\frac{12b}{\lambda-b}\int d\mathbf{p}\frac{4\pi}{p^2}|\hat{\mathbf{J}}_0(\mathbf{p})|^2\right]\\&=N^2\left[1-\frac{18b}{\lambda-b}\right]\frac{11}{35\pi},\end{aligned}\tag{A1}$$

where T is the 3×3 matrix with the components $T_{ij}=\delta_{ij}-p_ip_j/p^2$. This replaces (17). We proceed to bound the self-energy and exchange terms. Inequality (19) becomes

$$|\hat{\mathbf{J}}_{\mu\nu}(\mathbf{p})|^2\leqslant 3(2\pi)^{-3}\chi(|\mathbf{p}+\mathbf{n}_\mu-\mathbf{n}_\nu|\leqslant 1), \tag{A2}$$

because the support of u now has diameter 1 not $\sqrt{3}$. Since the N balls $B(\mathbf{n}_\nu,1)$ cover a given point, at most, 8 times (recall that now $\mathbf{n}_{2\nu-1}=\mathbf{n}_{2\nu}$) inequality (A2) leads to the bound

$$\frac{1}{2}\sum_{\mu,\nu}\langle\psi_\mu\otimes\psi_\nu,B(x-y)\psi_\nu\otimes\psi_\mu\rangle\leqslant\frac{48}{\pi}bN^{4/3}, \tag{A3}$$

improving (20). The kinetic energy is bounded by $(\lambda+b)N^{4/3}$ and $\langle\psi,V_c\psi\rangle\leqslant 0$ since $\Sigma Z_\kappa=N$ and $\Sigma Z_\kappa^2\geqslant 1$. In conjunction with (A1) and (A3) this gives

$$\mathcal{E}_N(\psi)\leqslant N^{4/3}\left[\lambda+b+\frac{48}{\pi}b\alpha-\alpha N^{2/3}\frac{11}{70\pi}\left(1-\frac{18b}{\lambda-b}\right)\right]. \tag{A4}$$

For $b=1/2$, $\alpha^{-1}=137$ and the optimal λ this is negative for $N\geqslant 3.4\times 10^7$. For $b=3/5$ and $\alpha>0$ arbitrary this is negative for $N\geqslant C\max(\alpha^{-3/2},1)$ with $C=43859$ where λ was chosen to minimize C. This explains the numbers in Theorem 3.

Now drop the term $(48/\pi)b\alpha$ in Eq. (A4) which was due to the exchange- and self-energy terms. By Eq. (11) what we are left with is a upper bound for $\inf_{\psi,\mathbf{A}\in\mathcal{A}}\mathcal{E}_N(\psi,\mathbf{A})$. It is negative for $b=\sqrt{3}$, the optimal λ and $\alpha^{3/2}N\geqslant 134'863$, or for $b=1/2$ the optimal λ, $\alpha^{-1}=137$, and $N\geqslant 3.4\times 10^7$. This explains the numbers in the remark after Corollary 2.

[1] F. J. Dyson and A. Lenard, ''Stability of matter I,'' J. Math. Phys. **8**, 423–434 (1967).
[2] F. J. Dyson and A. Lenard, ''Stability of matter II,'' J. Math. Phys. **9**, 698–711 (1967).
[3] E. H. Lieb, H. Siedentop, and J. P. Solovej, ''Stability and instability of relativistic electrons in classical electromagnetic fields,'' J. Stat. Phys. **89**, 37–59 (1997).
[4] J. Fröhlich, E. H. Lieb, and M. Loss, ''Stability of Coulomb systems with magnetic fields I. The one-electron atom,'' Commun. Math. Phys. **104**, 251–270 (1986).
[5] E. H. Lieb and M. Loss, ''Stability of Coulomb systems with magnetic fields. II: The many-electron atom and the one-electron molecule,'' Commun. Math. Phys. **104**, 271–282 (1986).
[6] E. H. Lieb and H.-T. Yau, ''The stability and instability of relativistic matter,'' Commun. Math. Phys. **118**, 177–213 (1988).
[7] M. Loss, ''The stability of matter interacting with fields,'' Not. Am. Math. Soc. **44**, 1288–1293 (1997).
[8] J. Sucher, ''Foundations of the relativistic theory of many-electron atoms,'' Phys. Rev. A **22**, 348–362 (1980).
[9] J. Sucher, ''Relativistic many-electron Hamiltonians,'' Phys. Scr. **36**, 271–281 (1987).
[10] Y. Ishikawa and K. Koc, ''Relativistic many-body perturbation theory based on the no-pair Dirac–Coulomb–Breit Hamiltonian: Relativistic correlation energies for the noble-gas sequence through Rn ($Z=86$), the group-IIB atoms through Hg, and the ions of Ne isoelectronic sequence,'' Phys. Rev. A **50**, 4733–4742 (1994).
[11] Y. Ishikawa and K. Koc, ''Relativistic many-body perturbation calculations for open-shell systems,'' Phys. Rev. A **56**, 1295–1304 (1997).
[12] P. Pyykkö, ''Relativistic effects in structural chemistry,'' Chem. Rev. **88**, 563–594 (1988).
[13] W. D. Evans, P. Perry, and H. Siedentop, ''The spectrum of relativistic one-electron atoms according to Bethe and Salpeter,'' Commun. Math. Phys. **178**, 733 (1996).
[14] C. Tix, ''Strict positivity of a relativistic Hamiltonian due to Brown and Ravenhall,'' Bull. London Math. Soc. **30**, 283–290 (1998).
[15] C. Tix, ''Lower bound for the ground state energy of the no-pair Hamiltonian,'' Phys. Lett. B **405**, 293–296 (1997).
[16] A. A. Balinsky and W. E. Evans, ''Stability of one-electron molecules in the Brown–Ravenhall model,'' preprint, 1998.
[17] A. A. Balinsky and W. E. Evans, ''On the virial theorem for the relativistic operator of Brown and Ravenhall, and the absence of embedded eigenvalues,'' Lett. Math. Phys. **44**, 233–248 (1998).
[18] M. Griesemer and H. Siedentop, ''A minimax principle for eigenvalue in spectral gaps,'' J. London Math. Soc. **30**, 283–290 (1998).
[19] G. Nenciu and G. Scharf, ''On regular external fields in quantum electrodynamics,'' Helv. Phys. Acta **51**, 412–424 (1978).
[20] W. Appel and A. Alastuey, ''Relativistic corrections for a classical one-component plasma,'' Physica A **252**, 238–268 (1998).
[21] M. Kiessling (private communication, 1998).

[22] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, New York, 1971).
[23] H. A. Bethe and E. E. Salpeter, ''Quantum mechanics of one and two-electron atoms,'' in *Handbuch der Physik*, 1st ed., edited by S. Flügge (Springer-Verlag, Berlin, 1957), Vol. XXXV, pp. 88–436.
[24] V. Bach, J. Fröhlich, and I. M. Sigal, ''Quantum electrodynamics of confined nonrelativistic particles,'' Adv. Math. **137**, 299–395 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Gauge symmetries of the master action in the Batalin–Vilkovisky formalism

M. A. Grigoriev, A. M. Semikhatov, and I. Yu. Tipunin
Lebedev Physics Institute, Russian Academy of Sciences, 53 Leninski prosp., Moscow 117924, Russia

(Received 29 June 1998; accepted for publication 10 December 1998)

We study the geometry of the Lagrangian Batalin–Vilkovisky (BV) theory on an antisymplectic manifold. We show that gauge symmetries of the BV theory are essentially the symmetries of an even symplectic structure on the stationary surface of the master action.
[S0022-2488(99)01803-4]

I. INTRODUCTION

In this paper, we investigate gauge symmetries in the Lagrangian Batalin–Vilkovisky (BV) formalism,[1,2] which is the most universal approach to the quantization of general gauge theories. The version of the BV quantization in which the coordinates are not explicitly separated into fields and antifields is known as the covariant approach.[3–9] The partition function is then given by a path integral of the exponential of the master action over a gauge-fixing surface $\mathcal{L}$ that is a Lagrangian submanifold of an odd-symplectic manifold $\mathcal{M}$. The gauge independence is realized as the independence from the choice of $\mathcal{L}$ and is ensured by the master equation imposed on the master action.

While the gauge symmetries of the original action are no longer symmetries of the formalism, the covariant formulation itself has its own ''gauge'' transformations. Each observable [a Becchi–Rouet–Stora–Tyutin- (BRST-) closed function] determines a gauge symmetry. Studied in Ref. 10 were the gauge symmetries corresponding to the trivial observables (BRST-exact functions). It was shown there that the space of functions modulo the BRST-exact ones, called *the space of gauge parameters*,[10] is endowed with a Lie algebra structure induced by the Lie algebra structure on the space of BRST-trivial gauge symmetries.

In this paper, we study symmetries of the BV formalism using the geometrical setting where the BV data are viewed as a QP-manifold.[7,11] This is a supermanifold equipped with an antisymplectic structure (the P-structure) and an odd nilpotent Hamiltonian vector field (the Q-structure); in the BV setting, the vector field is given by the antibracket with the master action.

An important characteristic of the Q-structure is the *zero locus* $\mathcal{Z}_{\mathbf{Q}}$ of the odd vector field. The most interesting case in applications is where $\mathcal{Z}_{\mathbf{Q}}$ is a smooth $(n|N-n)$-dimensional submanifold (assuming the antisymplectic supermanifold $\mathcal{M}$ to be $(N|N)$-dimensional). We call such QP-manifolds the proper QP-manifolds; then the QP-structure induces a symplectic structure on the zero locus of $\mathbf{Q}$ [see (II.8)]. [The existence of a symplectic structure on the zero locus can also be inferred from Ref. 11; the Poisson bracket on (,)-Lagrangian submanifolds was described in Ref. 12; see also Ref. 13.]

It turns out that symmetries of the BV ''master system'' are to a considerable degree determined by Hamiltonian vector fields on the stationary surface of the master action (the vector fields being Hamiltonian with respect to the *Poisson bracket*). We will explicitly define a nondegenerate Poisson bracket on the quotient algebra of all smooth functions modulo the functions vanishing on $\mathcal{Z}_{\mathbf{Q}}$; this generalizes the bilinear operation of Ref. 10, which is not a Poisson bracket since it fails to satisfy the Leibniz rule (in fact, it defined on the space that is not an algebra under the associative multiplication).

We show that each symmetry of a proper QP-manifold, i.e., a vector field preserving the QP-structure, can be restricted to $\mathcal{Z}_{\mathbf{Q}}$ and, moreover, this restriction is a locally Hamiltonian vector field with respect to the Poisson bracket on $\mathcal{Z}_{\mathbf{Q}}$. Conversely, locally Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$ can be lifted to vector fields on $\mathcal{M}$, into symmetries of the proper QP-manifold. At the same time, the *globally* Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$ lift to BRST-trivial symmetries. In this way, we obtain a ''translation table'' between the objects pertaining to the antisymplectic geometry on $\mathcal{M}$ and those of the symplectic geometry on $\mathcal{Z}_{\mathbf{Q}}$.

We further select the *on-shell symmetries*, i.e., symmetries modulo those vanishing on the stationary surface. We show that the Lie algebra $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ of the on-shell gauge symmetries is isomorphic to the Lie algebra of locally Hamiltonian vector fields on the stationary surface $\mathcal{Z}_{\mathbf{Q}}$. The Lie algebras of the on-shell gauge symmetries ($\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$), of gauge parameters[10] ($\tilde{\mathcal{O}}^c_{\text{triv}}$), and of gauge symmetries of the master action ($\tilde{\mathcal{O}}^c$), as well as their quantum counterparts, are related to each other as shown in (III.11) and to the cohomology, as shown in (III.16), (III.18), and (III.19).

We consider two examples of the general construction. We explicitly calculate the Lie algebras $\mathcal{O}^c$ and $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ in the Abelianized[14] gauge theory. It is not difficult to see then that the algebra of gauge symmetries of the original classical theory is embedded into the Lie algebra $\mathcal{O}^c$ as a subalgebra in such a way that the algebra of the on-shell gauge symmetries of the original theory is embedded into the Lie algebra $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$. As another example, we consider the theory with the vanishing action on a Lie group. Not surprisingly, the Poisson structure on the ''stationary surface'' is then related to the Kirillov bracket[15] on the coalgebra.

In Sec. II, we study the geometry of QP-manifolds. In Sec. III, we give a short reminder on the BV-quantization and then study the quantum and classical gauge symmetries. In Sec. IV, we demonstrate the main points of our construction in two characteristic examples.

II. GEOMETRY OF PROPER QP-MANIFOLDS

In this section, we study geometry of the QP-manifolds and define proper QP-manifolds by simply reformulating the condition from Ref. 1 for a solution of the master equation to be proper. We then show that the zero locus $\mathcal{Z}_{\mathbf{Q}}$ of the vector field $\mathbf{Q}$ on a proper QP-manifold is a symplectic manifold. Moreover, the vector fields that are *symmetries* of a proper QP-manifold correspond to locally Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$; the *BRST-trivial* symmetries then correspond to globally Hamiltonian vector fields.

A. A poisson structure

Let $\mathcal{M}$ be an $(N|N)$-dimensional supermanifold and let $\mathbb{C}_{\mathcal{M}}$ denote the algebra of smooth functions on $\mathcal{M}$. Let $(\cdot,\cdot)$: $\mathbb{C}_{\mathcal{M}}\times\mathbb{C}_{\mathcal{M}}\to\mathbb{C}_{\mathcal{M}}$ be an antibracket on $\mathcal{M}$. It satisfies

$$\epsilon((F,G))=\epsilon(F)+\epsilon(G)+1, \tag{II.1}$$

$$(F,G)=-(-1)^{(\epsilon(F)+1)(\epsilon(G)+1)}(G,F), \tag{II.2}$$

$$(F,GH)=(F,G)H+(-1)^{\epsilon(G)(\epsilon(F)+1)}G(F,H), \tag{II.3}$$

$$0=\underset{F,G,H}{\text{cycle}}(-1)^{(\epsilon(F)+1)(\epsilon(H)+1)}(F,(G,H)). \tag{II.4}$$

In a local coordinate system Γ^A, $A=1,...,2N$, we have the matrix $E^{AB}=(\Gamma^A,\Gamma^B)$ that defines a bivector E such that $(F,G)=E(dF,dG)$. We assume the antibracket to be nondegenerate.

Consider an odd vector field $\mathbf{Q}:\mathbb{C}_{\mathcal{M}}\to\mathbb{C}_{\mathcal{M}}$ on $\mathcal{M}$ satisfying the following conditions:

(1) $\mathbf{Q}$ preserves the antibracket, i.e., $\pounds_{\mathbf{Q}}E=0$, where $\pounds_{\mathbf{Q}}$ is the Lie derivative along $\mathbf{Q}$; equivalently, $\mathbf{Q}$ differentiates the antibracket:

$$\mathbf{Q}(F,G)-(\mathbf{Q}F,G)-(-1)^{\epsilon(F)+1}(F,\mathbf{Q}G)=0,\quad F,G\in\mathbb{C}_M; \tag{II.5}$$

(2) $\mathbf{Q}$ is nilpotent: $\mathbf{Q}(\mathbf{Q}F)=0$, $F\in\mathbb{C}_{\mathcal{M}}$, $\Leftrightarrow[\mathbf{Q},\mathbf{Q}]=0$.

Definition II.1[7,11]*: A supermanifold $\mathcal{M}$ equipped with a nondegenerate antibracket $(\cdot,\cdot)$ and an odd nilpotent vector field* **Q** *that satisfies condition (II.5) is called the* QP*-manifold.*

The main object of our analysis is the set $\mathcal{Z}_{\mathbf{Q}}$ of zeroes of **Q**, i.e., the set defined by the equations $Q^A=0$, where $\mathbf{Q}=Q^A\partial_A$ in a local coordinate system Γ^A. We assume $\mathcal{Z}_{\mathbf{Q}}$ to be a submanifold in $\mathcal{M}$. Denote by $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}\subset\mathbb{C}_{\mathcal{M}}$ the ideal of all smooth functions vanishing on $\mathcal{Z}_{\mathbf{Q}}$. We also assume the ''regularity condition,'' i.e., that **Q**-exact functions *generate* $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$, which means that any function $f\in\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ admits a representation $f=(\mathbf{Q}h)g$ with some $g,h\in\mathbb{C}_{\mathcal{M}}$. The quotient $\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}=\mathbb{C}_{\mathcal{M}}/\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ is the algebra of smooth functions on $\mathcal{Z}_{\mathbf{Q}}$. Obviously, $\mathbf{Q}\mathbb{C}_{\mathcal{M}}\subset\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$.

An additional requirement imposed on **Q** is that its local cohomology (that is, the cohomology evaluated in a sufficiently small neighborhood of a point) is trivial (constants only) at every point $p\in\mathcal{Z}_{\mathbf{Q}}$. In local coordinates, then, the nilpotent operator $\partial Q^A/\partial\Gamma^B|_{Q^A=0}$ has the vanishing cohomology on the tangent space to every $p\in\mathcal{Z}_{\mathbf{Q}}$.[7,11] This in turn implies the condition from Ref. 1:

$$\mathrm{rank}\left|\left(\frac{\partial Q^A}{\partial\Gamma^B}\right)\right|_{Q^A=0}=N. \tag{II.6}$$

In particular, it follows from (II.6) that $\mathcal{Z}_{\mathbf{Q}}$ is an $(n|N-n)$-dimensional submanifold.

Definition II.2: A proper QP*-manifold is a* QP*-manifold on which the local cohomology of* **Q** *is trivial at every point of* $\mathcal{Z}_{\mathbf{Q}}$.

Lemma II.3: The submanifold $\mathcal{Z}_{\mathbf{Q}}$ of a proper QP*-manifold is Lagrangian with respect to the antibracket. In particular, the ideal $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ is closed under the antibracket:*

$$(\mathbb{I}_{\mathcal{Z}_Q},\mathbb{I}_{Z_{\mathbf{Q}}})\subset\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}. \tag{II.7}$$

Proof: Since any function from $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ can be represented as the product $(\mathbf{Q}f)h$ with some $h\in\mathbb{C}_{\mathcal{M}}$, it suffices to check that the antibracket of **Q**-exact functions is **Q**-exact, which is obvious in view of $(\mathbf{Q}g,\mathbf{Q}h)=\mathbf{Q}(g,\mathbf{Q}h)$.

In fact, the submanifold $\mathcal{Z}_{\mathbf{Q}}$ is endowed with a natural Poisson structure. This is given by a construction of the type of those used, with some variations, in Refs. 10, 12, 13, and 16, namely,

$$\{F,G\}=(F,\mathbf{Q}G),\quad F,G\in\mathbb{C}_{\mathcal{M}}. \tag{II.8}$$

We interpret this structure as a bilinear mapping on the quotient algebra $\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}$. Functions from $\mathbb{C}_{\mathcal{M}}$ considered modulo $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ represent functions on $\mathcal{Z}_{\mathbf{Q}}$. We then have the following.

Theorem II.4: *For any proper* QP*-manifold,*

(1) *Equation (II.8) defines a Poisson bracket $\{\cdot,\cdot\}$: $\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}\times\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}\to\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}$ on the submanifold $\mathcal{Z}_{\mathbf{Q}}$.*
(2) *Moreover, the Poisson bracket $\{\cdot,\cdot\}$ is nondegenerate (thus, $\mathcal{Z}_{\mathbf{Q}}$ is symplectic).*

Proof: First of all, we must prove that definition (II.8) does not depend on the choice of representatives of the equivalence classes, i.e., $\{F+f,G+g\}-\{F,G\}\in\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ whenever $f,g\in\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$. Since **Q** differentiates the antibracket, we can check that

$$\{F+f,G+g\}-\{F,G\}=(-1)^{\epsilon(F)}(\mathbf{Q}F,g)+(-1)^{\epsilon(F)+1}\mathbf{Q}(F,g)+(f,\mathbf{Q}(G+g)).$$

The first and the third terms belong to $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ by Lemma (II.3) and the second term is in $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ because it is **Q**-exact. Thus (II.8) defines a mapping $\{\cdot,\cdot\}$: $\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}\times\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}\to\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}$. It is antisymmetric because

$$\{F,G\}+(-1)^{\epsilon(F)\epsilon(G)}\{G,F\}=(F,\mathbf{Q}G)+(-1)^{\epsilon(F)\epsilon(G)}(G,\mathbf{Q}F)=(-1)^{\epsilon(F)+1}\mathbf{Q}(F,G)\in\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}},$$

where we used (II.5) again. Next, to prove the Leibnitz rule, we evaluate

$$\{F,GH\}-\{F,G\}H-(-1)^{\epsilon(F)\epsilon(G)}G\{F,H\}$$
$$=(-1)^{(\epsilon(F)+1)(\epsilon(G)+1)}(\mathbf{Q}G)(F,H)+(-1)^{\epsilon(G)}(F,G)(\mathbf{Q}H), \quad \text{(II.9)}$$

which evidently vanishes modulo terms from $\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}$. Finally, to prove the Jacobi identity we have to show that

$$\underset{F,G,H}{\text{cycle}}(-1)^{\epsilon(F)\epsilon(H)}\{F,\{G,H\}\}=\underset{F,G,H}{\text{cycle}}(-1)^{\epsilon(F)\epsilon(H)}(F,\mathbf{Q}(G,\mathbf{Q}H))\equiv 0 \bmod \mathbb{I}_\mathbf{Q}.$$

In view of the Leibniz rule and the nilpotency condition this rewrites, modulo terms from $\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}$ as

$$\underset{\mathbf{Q}F,G,\mathbf{Q}H}{\text{cycle}}(-1)^{(\epsilon(\mathbf{Q}F)+1)(\epsilon(\mathbf{Q}H)+1)}(\mathbf{Q}F,(G,\mathbf{Q}H)),$$

which vanishes by virtue of the Jacobi identity for the antibracket.

To prove that the Poisson bracket is nondegenerate on $\mathcal{Z}_\mathbf{Q}$, we recall a standard fact from symplectic geometry, namely that in some neighborhood of $\mathcal{Z}_\mathbf{Q}$ there exists a coordinate system x^i, ξ_i such that the antibracket takes the canonical form $(x^i,\xi_j)=\delta^i_j$ and $\mathcal{Z}_\mathbf{Q}$ is determined by $\xi_i=0$. Locally, the vector field $\mathbf{Q}$ can be written in the form $\mathbf{Q}=(S,\cdot)$ with some function $S\in\mathbb{C}_\mathcal{M}$ [since a vector field preserving a nondegenerate (anti)bracket is locally Hamiltonian]. Expanding S as

$$S=S_0(x)+\xi_iS^i(x)+\xi_iS^{ij}(x)\xi_j+\xi_i\xi_j\xi_kS^{ijk}(x)+\cdots,$$

we see that $S_0(x)=\text{const}$ and $S^i(x)=0$, because $\mathbf{Q}=(S,\cdot)$ vanishes as $\xi=0$. Then condition (II.6) means that the matrix $S^{ij}(x)$ is nondegenerate at each point of $\mathcal{Z}_\mathbf{Q}$. On the other hand, $S^{ij}(x)|_{\xi=0}=-\frac{1}{2}\{x^i,x^j\}|_{\xi=0}$, which shows the theorem. ■

Note that the symbol $\{\cdot,\cdot\}$ is used for the formal operation (II.8) on the manifold $\mathcal{M}$ and also for the Poisson bracket on the submanifold $\mathcal{Z}_\mathbf{Q}$. We do not introduce two different symbols and hope that this will not lead to confusion.

B. Symmetries of QP-manifolds

For applications to the BV quantization in the subsequent sections, we will need some facts about symmetries of QP-manifolds.

Definition II.5: A vector field $\mathbf{X}$ *on a* QP-*manifold* $\mathcal{M}$ *is called a symmetry of* $\mathcal{M}$ *if*

(1) $\mathbf{X}$ *preserves the antibracket* $(\cdot,\cdot)$:

$$\mathbf{X}(F,G)-(\mathbf{X}F,G)-(-1)^{(\epsilon(F)+1)\epsilon(\mathbf{X})}(F,\mathbf{X}G)=0, \quad F,G\in\mathbb{C}_\mathcal{M}, \quad \text{(II.10)}$$

and

(2) $\mathbf{X}$ *preserves the odd vector field* $\mathbf{Q}$: $[\mathbf{Q},\mathbf{X}]=0$.

Our aim is to demonstrate that symmetries of a proper QP-manifold restrict to the zero locus of $\mathbf{Q}$, and to study the properties of such restrictions. We begin with characterizing, in the standard way, those vector fields on $\mathcal{M}$ that restrict to $\mathcal{Z}_\mathbf{Q}$:

Lemma II.6: A vector field $\mathbf{X}$ *on a proper* QP-*manifold* $\mathcal{M}$ *restricts to* $\mathcal{Z}_\mathbf{Q}$ *if and only if*

$$[\mathbf{X},\mathbf{Q}]|_{\mathcal{Z}_\mathbf{Q}}=0. \quad \text{(II.11)}$$

Proof: A vector field restricts to $\mathcal{Z}_\mathbf{Q}$ if and only if it preserves the ideal of functions vanishing on $\mathcal{Z}_\mathbf{Q}$: $\mathbf{X}\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}\subset\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}$. Now, $[\mathbf{X},\mathbf{Q}]|_{\mathcal{Z}_\mathbf{Q}}=0\Leftrightarrow[\mathbf{X},\mathbf{Q}]f\in\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}\forall f\in\mathbb{C}_\mathcal{M}$, which rewrites as $\mathbf{XQ}f\in\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}$. Since $\mathbf{Q}$-exact functions generate the ideal, we conclude that $\mathbf{X}\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}\subset\mathbb{I}_{\mathcal{Z}_\mathbf{Q}}$. The converse is now obvious. ■

It follows from this Lemma that any vector field $\mathbf{X}$ that is a symmetry of a proper QP-manifold can be restricted to $\mathcal{Z}_{\mathbf{Q}}$. We now recall that the zero locus of $\mathbf{Q}$ is endowed with a Poisson structure.

Theorem II.7: *If a vector field* $\mathbf{X}$ *is a symmetry of a proper* QP-*manifold* $\mathcal{M}$, *its restriction* $\mathbf{X}|_{\mathcal{Z}_{\mathbf{Q}}}$ *to the zero locus of* $\mathbf{Q}$ *preserves the Poisson bracket from Theorem II.4:*

$$\mathbf{X}|_{\mathcal{Z}_{\mathbf{Q}}}\{F,G\}-\{\mathbf{X}|_{\mathcal{Z}_{\mathbf{Q}}}F,G\}-(-1)^{\epsilon(F)\epsilon(\mathbf{X})}\{F,\mathbf{X}|_{\mathcal{Z}_{\mathbf{Q}}}G\}=0,\quad F,G\in\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}. \tag{II.12}$$

Proof: Let us choose two representatives $F,G\in\mathbb{C}_{\mathcal{M}}$ of the equivalence classes of functions on $\mathcal{Z}_{\mathbf{Q}}$. Using the properties stated in Definition (II.5), we have

$$\mathbf{X}\{F,G\}=\mathbf{X}(F,\mathbf{Q}G)=(\mathbf{X}F,\mathbf{Q}G)+(-1)^{(\epsilon(F)+1)\epsilon(\mathbf{X})}(F,\mathbf{X}\mathbf{Q}G)=\{\mathbf{X}F,G\}+(-1)^{\epsilon(F)\epsilon(\mathbf{X})}\{F,\mathbf{X}G\}.$$

■

It follows from the nondegeneracy of the Poisson bracket (II.8) that any vector field $\mathbf{x}$ on $\mathcal{Z}_{\mathbf{Q}}$ that preserves the Poisson bracket can be written as $\mathbf{x}=\{H,\cdot\}$ with some (locally) defined function H. We will refer to this as a *locally Hamiltonian* vector field. Every $\mathbf{x}=\{H,\cdot\}$ with a globally defined H will be called *globally Hamiltonian*. It is well known that globally Hamiltonian vector fields form an ideal in the Lie algebra of locally Hamiltonian vector fields.

C. Lifts and restrictions of vector fields

In this section, we are interested in relations between Hamiltonian vector fields on the symplectic manifold $\mathcal{Z}_{\mathbf{Q}}$ and the Hamiltonian vector fields on $\mathcal{M}$, in particular, the symmetries of the proper QP-manifold $\mathcal{M}$. (Thus, whenever we speak about Hamiltonian vector fields on $\mathcal{M}$, these are Hamiltonian with respect to the *anti*bracket, while the Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$ are Hamiltonian with respect to the Poisson bracket.)

To explain why there exists a correspondence between symmetries of a proper QP-manifold $\mathcal{M}$ and symmetries of its symplectic submanifold $\mathcal{Z}_{\mathbf{Q}}$, we begin with an example.[16] Consider an N-dimensional symplectic manifold $\mathcal{K}$ with the symplectic form $\hat{\omega}$ which defines the nondegenerate Poisson bracket $\{\cdot,\cdot\}$ (in general, $\mathcal{K}$ can be a supermanifold, but we assume for simplicity that it is an even manifold). In a local coordinate system x^i, we have the invertible matrix $\omega^{ij}=\{x^i,x^j\}$. Let $\Pi T^*\mathcal{K}$ be the cotangent bundle with the flipped parity. In the canonical coordinates x^i, ξ_i [with the antibracket $(x^i,\xi_j)=\delta^i_j$], the manifold $\mathcal{K}$ can be identified with the zero section $\xi_i=0$ of $\Pi T^*\mathcal{K}$. The function $S=\frac{1}{2}\xi_i\omega^{ij}\xi_j$ satisfies $(S,S)=0$ because ω^{ij} is the matrix of a Poisson bracket. Then the submanifold $\mathcal{K}$ is the zero locus of the vector field

$$\mathbf{Q}=(S,\cdot)=\frac{1}{2}\xi_i\left(\frac{\vec{\partial}}{\partial x^k}\omega^{ij}\right)\xi_j\frac{\vec{\partial}}{\partial\xi_k}-\xi_i\omega^{ij}\frac{\vec{\partial}}{\partial x^j}. \tag{II.13}$$

It is easy to see that $\mathbf{Q}$ satisfies the conditions of Definition II.1. In addition, condition (II.6) is satisfied because the dimension of $\mathcal{K}$ is N. Thus $\Pi T^*\mathcal{K}$ is a QP-manifold and, in fact, a proper QP-manifold (because ω^{ij} is nondegenerate). With the help of the symplectic form, we can identify $\Pi T^*\mathcal{K}$ with the tangent bundle $\Pi T\mathcal{K}$ (with $\xi^i=\omega^{ij}\xi_j$ being the coordinates on the fibers). Then we can rewrite $\mathbf{Q}$ as

$$\mathbf{Q}=\xi^i\frac{\vec{\partial}}{\partial x^i}. \tag{II.14}$$

Upon the identification of functions on $\Pi T\mathcal{K}$ with differential forms on $\mathcal{K}$, $\mathbf{Q}$ becomes the De Rham differential on $\mathcal{K}$.[11]

Further, every locally Hamiltonian vector field $\mathbf{x}$ on $\mathcal{K}$ can be lifted to a globally Hamiltonian vector field $\mathbf{X}$ on $\Pi T^*\mathcal{K}$. Namely, if $\mathbf{x}=\{H,\cdot\}$ (where we allow H to be multivalued), we take

$F=(-1)^{\epsilon(H)}\mathbf{Q}(\pi^*H)$, where π^* is the pullback with respect to the canonical projection $\pi:\Pi T^*\mathcal{K}\rightarrow\mathcal{K}$. Then the vector field $\mathbf{X}=(F,\cdot)$ *is well-defined* (independent of the multivaluedness of H) and satisfies

$$\mathbf{X}|_{\mathcal{K}}=\mathbf{x} \tag{II.15}$$

and is also a symmetry of $\Pi T^*\mathcal{K}$ in the sense of Definition II.5. Conversely, any symmetry of $\Pi T^*\mathcal{K}$ determines a locally Hamiltonian vector field on $\mathcal{K}$ (see Theorem II.7), which is obviously a symmetry of this symplectic manifold.

The above is a particular case of a more general construction, where $\mathcal{M}$ can be an arbitrary proper QP-manifold. There still exists a correspondence between symmetries of $\mathcal{M}$ and symmetries of its symplectic submanifold $\mathcal{Z}_{\mathbf{Q}}$.

We have seen in Theorem II.7 that every symmetry of a proper QP-manifold restricts to $\mathcal{Z}_{\mathbf{Q}}$ as a locally Hamiltonian vector field. Consider now the converse problem. We say that a symmetry $\mathbf{X}$ of a proper QP-manifold $\mathcal{M}$ is a lift from $\mathcal{Z}_{\mathbf{Q}}$ of a locally Hamiltonian vector field $\mathbf{x}$ if $\mathbf{X}$ restricts to $\mathcal{Z}_{\mathbf{Q}}$ and $\mathbf{X}|_{\mathcal{Z}_{\mathbf{Q}}}=\mathbf{x}$.

Theorem II.8: *Every locally Hamiltonian vector field* $\mathbf{x}$ *on* $\mathcal{Z}_{\mathbf{Q}}$ *admits a lift to a symmetry of* $\mathcal{M}$ *that is a globally Hamiltonian vector field on* $\mathcal{M}$ *with a* $\mathbf{Q}$*-closed Hamiltonian. If* $\mathbf{x}$ *is globally Hamiltonian on* $\mathcal{Z}_{\mathbf{Q}}$*, it is lifted to a globally Hamiltonian vector field with a* $\mathbf{Q}$*-exact Hamiltonian.*

In what follows, symmetries of $\mathcal{M}$ of the form $\mathbf{X}=(F,\cdot)$ with a $\mathbf{Q}$-exact Hamiltonian F are called the *BRST-trivial symmetries*.

Proof: For a (locally) Hamiltonian vector field $\mathbf{x}$ on $\mathcal{Z}_{\mathbf{Q}}$, the equation $\mathbf{x}=\{H,\cdot\}$ can be solved for H in a sufficiently small neighborhood of every point $p\in\mathcal{Z}_{\mathbf{Q}}$. Different such solutions can be considered as a multivalued Hamiltonian H. This can be extended to a multivalued function $\tilde{H}$ on $\mathcal{M}$ that restricts to H on $\mathcal{Z}_{\mathbf{Q}}$ (for example, consider a neighborhood U of $\mathcal{Z}_{\mathbf{Q}}$ in $\mathcal{M}$ and identify it with a neighborhood of the zero section of a vector bundle over $\mathcal{Z}_{\mathbf{Q}}$; if $\tilde{h}$ is the pullback of the multivalued function H to the bundle, we can choose a function $\alpha\in\mathbb{C}_{\mathcal{M}}$ such that $\alpha|_{\mathcal{Z}_{\mathbf{Q}}}=1$ and $\alpha=0$ outside U, which yields the lifting $\tilde{H}=\alpha\tilde{h}$ of the multivalued Hamiltonian H).

Then consider the function $F=(-1)^{\epsilon(H)}\mathbf{Q}\tilde{H}$ on $\mathcal{M}$; F is *single-valued,* is $\mathbf{Q}$-closed by construction, but in general is not $\mathbf{Q}$-exact (because its $\mathbf{Q}$-primitive is not necessarily single-valued). Now, let $\mathbf{X}=(F,\cdot)$. For any function $\tilde{G}\in\mathbb{C}_{\mathcal{M}}$, we have

$$(F,\tilde{G})|_{\mathcal{Z}_{\mathbf{Q}}}=(\tilde{H},\mathbf{Q}\tilde{G})|_{\mathcal{Z}_{\mathbf{Q}}},$$

which coincides with $\{\tilde{H}|_{\mathcal{Z}_{\mathbf{Q}}},\tilde{G}|_{\mathcal{Z}_{\mathbf{Q}}}\}=\mathbf{x}\tilde{G}|_{\mathcal{Z}_{\mathbf{Q}}}$ [see (II.8)]. Thus $\mathbf{X}$ is a lift of $\mathbf{x}$ to a symmetry of $\mathcal{M}$.

Whenever the Hamiltonian H of $\mathbf{x}$ on $\mathcal{Z}_{\mathbf{Q}}$ is globally defined on $\mathcal{Z}_{\mathbf{Q}}$, the function F is obviously $\mathbf{Q}$-exact and thus $\mathbf{X}=(F,\cdot)$ is a BRST-trivial symmetry of $\mathcal{M}$. ■

This theorem can also be seen by noticing that the cohomology of $\mathbf{Q}$ evaluated on an appropriately chosen neighborhood U of $\mathcal{Z}_{\mathbf{Q}}$ in $\mathcal{M}$ coincides with the De Rham cohomology of $\mathcal{Z}_{\mathbf{Q}}$.[11] Namely, one can identify the neighborhood U of $\mathcal{Z}_{\mathbf{Q}}$ with some neighborhood of the zero section of $\Pi T\mathcal{Z}_{\mathbf{Q}}$. Then we can write $\mathbf{Q}=\xi^i\vec{\partial}/\partial x^i$, where x^i and ξ^i are coordinates on $\mathcal{Z}_{\mathbf{Q}}$ and on the fibers, respectively. Thus $\mathbf{Q}$ coincides with the De Rham differential of $\mathcal{Z}_{\mathbf{Q}}$ if one identifies functions on U that are homogeneous in ξ with the differential forms on $\mathcal{Z}_{\mathbf{Q}}$. In particular, every closed but not exact one-form $f=dx^if_i$ gives rise to the function $F=\xi^if_i$ on U that is obviously in the cohomology of $\mathbf{Q}$. At the same time, the one-form $f=dx^if_i$ gives rise to the locally Hamiltonian vector field $\mathbf{x}=(-1)^{\epsilon(x^i)\epsilon(f_i)}f_i\omega^{ij}\vec{\partial}/\partial x^j$ on $\mathcal{Z}_{\mathbf{Q}}$ (where $\omega^{ij}=\{x^i,x^j\}$). Therefore, $\mathbf{x}$ lifts to the vector field $(-1)^{\epsilon(F)+1}(F,\cdot)$ on U whose Hamiltonian is the same function $F=\xi^if_i$.

We see, in particular, that a locally Hamiltonian vector field representing the first cohomology of $\mathcal{Z}_{\mathbf{Q}}$ corresponds to an element of the $\mathbf{Q}$-cohomology on $\mathcal{M}$. We now single out those $(\cdot,\cdot)$-Hamiltonian vector fields on $\mathcal{M}$ that restrict to $\{\cdot,\cdot\}$-Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$, and describe the full arbitrariness of the lifts of Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$ to Hamiltonian vector fields on $\mathcal{M}$. The following is proved by directly generalizing the proof of Theorem II.8.

Theorem II.9:

(1) *Let* $\mathbf{X}=(F,\cdot)$ *be a globally Hamiltonian vector field on a proper* QP*-manifold* $\mathcal{M}$ *with the Hamiltonian satisfying* $\mathbf{Q}F\in\mathbb{I}^3_{\mathcal{Z}_{\mathbf{Q}}}$ *(i.e.,* $\mathbf{Q}F=Q^AQ^BQ^CY_{ABC}$ *with some* $Y_{ABC}\in\mathbb{C}_{\mathcal{M}}$*). Then* $\mathbf{X}$ *restricts to a locally Hamiltonian vector field on* $\mathcal{Z}_{\mathbf{Q}}$.

(2) *Every locally Hamiltonian vector field* $\mathbf{x}$ *on* $\mathcal{Z}_{\mathbf{Q}}$ *admits a lift to a globally Hamiltonian vector field on* $\mathcal{M}$ *with the Hamiltonian F satisfying* $\mathbf{Q}F=0$. *If* $\mathbf{x}$ *is globally Hamiltonian on* $\mathcal{Z}_{\mathbf{Q}}$ *with the Hamiltonian H, the Hamiltonians of all its lifts to* $\mathcal{M}$ *are of the form*

$$F=(-1)^{\epsilon(\tilde{H})}\mathbf{Q}\tilde{H}+K+\text{const},\quad K\in\mathbb{I}^2_{\mathcal{Z}_{\mathbf{Q}}},\tag{II.16}$$

where $\tilde{H}$ *is any function on* $\mathcal{M}$ *such that* $\tilde{H}|_{\mathcal{Z}_{\mathbf{Q}}}=H$.

III. GAUGE SYMMETRIES OF THE MASTER ACTION

We now interpret the classical gauge symmetries in the covariant BV formalism as symmetries of the corresponding proper QP-manifold. Using the results of the previous section, we then show that the Lie algebra of locally Hamiltonian vector fields on the stationary surface of the master action coincides with the algebra of *on-shell gauge symmetries*. Section III A contains a brief reminder on the BV formalism, so the reader may wish to go directly to Sec. III B.

A. Batalin–Vilkovisky quantization

The geometrical background of the covariant formulation of the BV quantization is an $(N|N)$-dimensional supermanifold $\mathcal{M}$ equipped with a nondegenerate antibracket $(\cdot,\cdot)$ and a volume form $d\mu=\rho d\Gamma$, where $\rho=\rho(\Gamma)$ is a density (and Γ^A, $A=1,...2N$, are some local coordinates). The density should be compatible with the antibracket in such a way that the BV Δ operator

$$\Delta_\rho H=\tfrac{1}{2}\,\mathrm{div}_\rho(\mathbf{V}_H)\tag{III.1}$$

be nilpotent, $\Delta_\rho^2=\frac{1}{2}[\Delta_\rho,\Delta_\rho]=0$. Here, div_ρ denotes the divergence of a vector field with respect to the density ρ and $\mathbf{V}_H=(H,\cdot)$ is the globally Hamiltonian vector field with the Hamiltonian H.

The physics is determined by the quantum master action $W\in\mathbb{C}_{\mathcal{M}}[[\hbar]]$ (a formal power series in $\hbar$ with coefficients in $\mathbb{C}_{\mathcal{M}}$) that satisfies the quantum master equation

$$\Delta_\rho e^{(i/\hbar)W}=0\Leftrightarrow\tfrac{1}{2}(W,W)=i\hbar\Delta_\rho W.\tag{III.2}$$

Writing $W=S+i\hbar W_1+(i\hbar)^2W_2+\cdots$, we rewrite (III.2) as

$$(S,S)=0,\tag{III.3}$$

$$(S,W_1)=\Delta_\rho S,\tag{III.4}$$

and so on. Equation (III.3) is the classical master equation and the function $S=W|_{\hbar=0}$ is called the classical master action.

In addition to the master equation, one should impose boundary conditions on W. This requires fixing a Lagrangian submanifold $\mathcal{L}_0$ in $\mathcal{M}$ (in the canonical coordinates, the manifold of fields, with the antifields set to zero) and a function $\mathcal{S}$ on $\mathcal{L}_0$, which is the *original (''bare'') action* of the classical theory that is being quantized. Then one requires $W(\Gamma,\hbar)$ to be such that $W(\cdot,0)|_{\mathcal{L}_0}=\mathcal{S}$.

By definition, a quantum observable is a function $A\in\mathbb{C}_{\mathcal{M}}[[\hbar]]$ satisfying $\delta_WA=0$, where

$$\delta_WA=(W,A)-i\hbar\Delta_\rho A.\tag{III.5}$$

It follows from (III.2) that $\delta_W^2=0$, and, therefore, any function A of the form $A=\delta_W B$ is an observable; these are called *trivial observables*. Expanding $A=A_0+i\hbar A_1+(i\hbar)^2A_2+\cdots$, we rewrite the equation $\delta_W A=0$ as

$$(S,A_0)=0, \tag{III.6}$$

$$(S,A_1)+(W_1,A_0)=\Delta_\rho A_0, \tag{III.7}$$

and so on. An $\hbar$-independent function A_0 satisfying (III.6) is called a *classical observable*. It is easy to see that if $A=\delta_W B$, then $A_0=(S,B_0)$, where $B_0=B|_{\hbar=0}$. Any classical observable A_0 of the form $A_0=(S,B_0)$ with some $\hbar$-independent function B_0 is called a *trivial classical observable*.

The quantum expectation of an observable is defined via the path-integral over a Lagrangian submanifold $\mathcal{L}$,

$$\langle A\rangle=\int_{\mathcal{L}} d\lambda_\rho\, A e^{(i/\hbar)W}, \tag{III.8}$$

where $d\lambda_\rho$ is the volume form on $\mathcal{L}$ determined by the volume form $d\mu=\rho d\Gamma$ on $\mathcal{M}$ and by the antisymplectic structure as follows:[4,3,8]

$$d\lambda_\rho(e^1,\ldots,e^N)=(d\mu(e^1,\ldots,e^N,f_1,\ldots,f_N))^{1/2}, \tag{III.9}$$

where $e^i\in T\mathcal{L}$ and $f_j\in T\mathcal{M}$ are any vectors that satisfy $\hat{E}(e^i,f_j)=\delta^i_j$ and $\hat{E}$ is the antisymplectic two-form on $\mathcal{M}$. It follows from (III.9) that the volume form $d\lambda_{\rho'}$ corresponding to the density function $\rho'=\rho e^H$ is related to $d\lambda_\rho$ as $d\lambda_{\rho'}=d\lambda_\rho e^{(1/2)H}$ (this is the origin of the exponent in Definition III.1). If the submanifold $\mathcal{L}$ is determined by the equations $G_\alpha=0$, $\alpha=1,...,N$, it is Lagrangian whenever $(G_\alpha,G_\beta)=U^\gamma_{\alpha\beta}G_\gamma$.

An important part of the BV axioms is the nondegeneracy conditions. The submanifold $\mathcal{L}$ in (III.8) must be such that the restriction of $S=W|_{\hbar=0}$ to $\mathcal{L}$ be nondegenerate. In terms of the equations $G_\alpha=0$, the matrix $\partial_A G_\alpha$ and the Hessian matrix $\partial_A\partial_B S$ should have no common null vectors at the points where $\partial_A S=0$ and $G_\alpha=0$.[1–3] Whenever the set $\mathcal{Z}_{(S,\cdot)}$ defined by equations $\partial_A S=0$ is a submanifold, this requirement means that $\mathcal{Z}_{(S,\cdot)}$ intersects $\mathcal{L}$ transversely. It also follows that the rank of the Hessian matrix $H_{AB}=(\partial_A\partial_B S)|_{\partial_A S=0}$ satisfies $\mathrm{rank}(H_{AB})\geqslant N$. At the same time, the classical master equation (III.3) implies that $\mathrm{rank}(H_{AB})\leqslant N$, whence we have[1–3]

$$\mathrm{rank}\left(\frac{\partial^2 S}{\partial\Gamma^A\partial\Gamma^B}\right)\Bigg|_{\partial_A S=0}=N. \tag{III.10}$$

The solution of classical master equation (III.3) that satisfies (III.10) is called a *proper solution*.

The key statement of the BV formalism is that the path integral constructed as in (III.8) is invariant under infinitesimal deformations of the Lagrangian submanifold $\mathcal{L}$[1–4] for every quantum observable A. In the case where $A=1$, this is often called the gauge independence of the partition function.

B. Lie algebras of gauge symmetries

We now study Lie algebras of gauge symmetries in the BV quantization scheme. These are Lie algebras $\mathcal{O}^q$ and $\mathcal{O}^c$ of *quantum and classical gauge symmetries*, respectively. In addition to these two basic algebras, it is useful to consider several more Lie algebras, which we define in what follows and which can be arranged into the following commutative diagram of Lie algebra homomorphisms:

$$\begin{array}{ccccc} \tilde{\mathcal{O}}^q_{\mathrm{triv}} & \longrightarrow & \mathcal{O}^q_{\mathrm{triv}} & \longrightarrow & \mathcal{O}^q \\ \downarrow{\scriptstyle \hbar=0} & & \downarrow{\scriptstyle \hbar=0} & & \downarrow{\scriptstyle \hbar=0} \\ \tilde{\mathcal{O}}^c_{\mathrm{triv}} & \longrightarrow & \mathcal{O}^c_{\mathrm{triv}} & \longrightarrow & \mathcal{O}^c \\ & \searrow & \downarrow & \swarrow & \\ & & \mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}} & & \end{array} \tag{III.11}$$

In addition, we have homomorphisms (III.16), (III.18), and (III.19), whose constructions will also be explained in what follows. Here $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ is the Lie algebra of the *on-shell gauge symmetries* (which, as we show, is the algebra of locally Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$), $\mathcal{O}^q_{\mathrm{triv}}$ is the Lie algebra *of BRST-trivial quantum gauge symmetries*, and $\tilde{\mathcal{O}}^q_{\mathrm{triv}}$ is the Lie algebra *of quantum gauge parameters*.[10,6] The Lie algebras $\mathcal{O}^c$, $\mathcal{O}^c_{\mathrm{triv}}$, and $\tilde{\mathcal{O}}^c_{\mathrm{triv}}$ are the classical counterparts of $\mathcal{O}^q$, $\mathcal{O}^q_{\mathrm{triv}}$, and $\tilde{\mathcal{O}}^q_{\mathrm{triv}}$, respectively. We now proceed to the exact definitions.

Definition III.1:[10,17] *A vector field* $\mathbf{X}(\hbar)$ *is called a quantum gauge symmetry if it preserves the antibracket and the measure* $\rho e^{(2i/\hbar)W} d\Gamma$ *(viewed as formal power series in* $\hbar$*). The Lie algebra* $\mathcal{O}^q$ *of these vector fields is called the Lie algebra of quantum gauge symmetries.*

It follows from this definition that a quantum gauge symmetry $\mathbf{X}(\hbar)$ satisfies

$$\mathrm{div}_\rho(\mathbf{X}(\hbar)) + \frac{2i}{\hbar}\mathbf{X}(\hbar)W = 0, \tag{III.12}$$

$$\mathbf{X}(F,G) - (\mathbf{X}F,G) - (-1)^{(\epsilon(F)+1)\epsilon(\mathbf{X})}(F,\mathbf{X}G) = 0, \quad F,G \in \mathbb{C}_{\mathcal{M}}. \tag{III.13}$$

Equation (III.13) implies that there exists, at least locally, a function $A(\hbar)$ such that $\mathbf{X}(\hbar) = (A(\hbar),\cdot)$. Then Eq. (III.12) implies that $(W,A(\hbar)) - i\hbar\Delta_\rho A(\hbar) = 0$. Whenever $A(\hbar)$ is globally defined, it is a quantum observable. We explicitly indicate the $\hbar$ dependence of $\mathbf{X}(\hbar)$ because $\rho e^{(2i/\hbar)W}$ should be preserved for any value of $\hbar$; we assume $\mathbf{X}(\hbar)$ to be a formal power series in $\hbar$ with coefficients in the vector fields on $\mathcal{M}$.

Although $\mathbf{X}(\hbar)$ is not a symmetry of any classical system (in particular, it preserves neither the quantum master action nor the measure $d\mu$), we call it a quantum gauge symmetry because its classical counterpart, obtained by taking the $\hbar \to 0$ limit, does preserve the classical master action S. (This action can be considered as the action of some classical system defined on $\mathcal{M}$. Then the classical master equation can be viewed as an additional constraint imposed on the system with the action S. One can naturally identify gauge symmetries of this system with the transformations preserving both S and the master equation imposed on S.)

To make contact with the literature, we consider the Lie algebra $\mathcal{O}^q_{\mathrm{triv}}$ of quantum *BRST-trivial gauge symmetries*.[10] These are quantum gauge symmetries $\mathbf{X}_B(\hbar) = (\delta_W B(\hbar),\cdot)$ whose Hamiltonians are trivial observables [see (III.5)], which span an ideal in $\mathcal{O}^q$.

Now, the (Hamiltonian) mapping $\mathbb{C}_{\mathcal{M}}[[\hbar]] \to \mathcal{O}^q_{\mathrm{triv}}$ allows us to pull back the Lie bracket from $\mathcal{O}^q_{\mathrm{triv}}$ to the space of $\hbar$-dependent functions. Namely,

$$[B^1(\hbar), B^2(\hbar)]^q = (B^1(\hbar), \delta_W B^2(\hbar)), \tag{III.14}$$

which implies

$$[\mathbf{X}_{B^1}(\hbar), \mathbf{X}_{B^2}(\hbar)] = (\delta_W(B^1(\hbar), \delta_W B^2(\hbar)),\cdot) = \mathbf{X}_{[B^1(\hbar), B^2(\hbar)]^q}. \tag{III.15}$$

The bracket (III.14) was shown in Refs. 10 and 6 to determine a Lie algebra structure on the quotient space

$$\tilde{O}^q_{\mathrm{triv}} = \mathbb{C}_{\mathcal{M}}[[\hbar]]/\delta_W \mathbb{C}_{\mathcal{M}}[[\hbar]]$$

of all $\hbar$-dependent functions modulo the δ_W-exact ones. $\mathcal{O}^q_{\text{triv}}$ was called the *Lie algebra of quantum gauge parameters* in Refs. 10 and 6. [This is not an algebra under the *associative* multiplication because the multiplication does not preserve the equivalence classes $B(\hbar)\sim B(\hbar)+\delta_W C(\hbar)$; in particular, (III.14) is not a Poisson bracket.]

There is a nice way to ''measure'' how much $\mathcal{O}^q_{\text{triv}}$ differs from $\tilde{\mathcal{O}}^q_{\text{triv}}$. The (Hamiltonian) mapping $\mathbb{C}_{\mathcal{M}}[[\hbar]]\to\mathcal{O}^q_{\text{triv}}$ induces a homomorphism $\tilde{\mathcal{O}}^q_{\text{triv}}\to\mathcal{O}^q_{\text{triv}}$ [see diagram (III.11)], whose kernel consists of functions (modulo δ_W-exact ones) satisfying $\delta_W B(\hbar)=\text{const}(\hbar)$. However, the fact that a function F satisfies $\delta_W F(\hbar)=\text{const}(\hbar)$ implies $\delta_W F(\hbar)=0$. [In order to see this, consider first a function F_0 satisfying $\mathbf{Q}F_0=(S,F_0)=\text{const}$. Since (as we see in the next subsection) $\mathcal{M}$ is a proper QP-manifold, the function $\mathbf{Q}F_0$ vanishes on the zero locus of $\mathbf{Q}$, and therefore, $\mathbf{Q}F_0=0$. Now, to see that equation $\delta_W F(\hbar)=\text{const}(\hbar)$ leads to $\delta_W F(\hbar)=0$, we rewrite δ_W and $F(\hbar)$ as a power series in $\hbar$: $\delta_W=\delta^0_W+i\hbar\,\delta^1_W+(i\hbar)^2\delta^2_W+\cdots$, where in particular $\delta^0_W=\mathbf{Q}$, and $F=F_0+i\hbar F_1+(i\hbar)^2F_2+\cdots$. Since $\mathcal{M}$ is a proper QP-manifold, the equation $\mathbf{Q}F_0=(S,F_0)=0$ implies that in some neighborhood, we have $F_0=\mathbf{Q}\phi_0+\text{const}$ with some function ϕ_0. Then in the first order in $\hbar$, the equation $\delta^1_W F_0+\delta^0_W F_1=\text{const}$ implies that $\delta^1_W F_0+\delta^0_W F_1=0$ because $\delta^1_W F_0=-\delta^0_W\delta^1_W\phi_0$. A similar argument applies to higher orders in $\hbar$. Thus, the kernel of the homomorphism $\tilde{\mathcal{O}}^q_{\text{triv}}\to\mathcal{O}^q_{\text{triv}}$ coincides with the cohomology of δ_W evaluated on the space of formal power series in $\hbar$ with the coefficients in smooth function on $\mathcal{M}$.] We thus conclude that the homomorphism $\tilde{\mathcal{O}}^q_{\text{triv}}\to\mathcal{O}^q_{\text{triv}}$ is included into the exact sequence that involves the cohomology of δ_W:

$$0\to\mathcal{H}^q\to\tilde{\mathcal{O}}^q_{\text{triv}}\to\mathcal{O}^q_{\text{triv}}\to 0,\quad \mathcal{H}^q=\text{Ker}\,\delta_W/\text{Im}\,\delta_W. \tag{III.16}$$

The classical versions of these constructions are as follows.

Definition III.2: A vector field $\mathbf{X}_0$ *is called a classical gauge symmetry if* $\mathbf{X}_0 S=0$ *and* $\mathbf{X}_0$ *preserves the antibracket. The Lie algebra* $\mathcal{O}^c$ *of these vector fields is called the Lie algebra of classical gauge symmetries*.

The classical *BRST-trivial gauge symmetries* are the vector fields $\mathbf{X}_0=((S,B_0),\cdot)$ whose Hamiltonians are trivial classical observables [see (III.6)]. These vector fields span the ideal $\mathcal{O}^c_{\text{triv}}\subset\mathcal{O}^c$, which is called the *classical BRST-trivial gauge symmetries*.

We have the obvious homomorphism $\mathcal{O}^q\xrightarrow{\hbar\to 0}\mathcal{O}^c$. This induces a homomorphism from the ideal $\mathcal{O}^q_{\text{triv}}\subset\mathcal{O}^q$ into the ideal $\mathcal{O}^c_{\text{triv}}\subset\mathcal{O}^c$ [which are shown in (III.11)].

The classical counterpart of $\tilde{\mathcal{O}}^q_{\text{triv}}$ is the space $\tilde{\mathcal{O}}^c_{\text{triv}}$ of all functions on $\mathcal{M}$ modulo the functions of the form (S,C), where S is the classical master action satisfying $(S,S)=0$. One can see that the space $\tilde{\mathcal{O}}^c_{\text{triv}}$ is endowed with a Lie algebra structure with respect to the ''classical'' bracket

$$[B^1_0,B^2_0]^c=(B^1_0,(S,B^2_0)). \tag{III.17}$$

Thus, we have the Lie algebra homomorphism $\tilde{\mathcal{O}}^c_{\text{triv}}\to\mathcal{O}^c_{\text{triv}}$ shown in (III.11). The kernel of the homomorphism coincides with the cohomology of $\mathbf{Q}$; therefore, we have the following exact sequence involving the cohomology of $\mathbf{Q}$:

$$0\to\mathcal{H}^c\to\tilde{\mathcal{O}}^c_{\text{triv}}\to\mathcal{O}^c_{\text{triv}}\to 0,\quad \mathcal{H}^c=\text{Ker}\,\mathbf{Q}/\text{Im}\,\mathbf{Q}. \tag{III.18}$$

We also observe that the relation between the quantum and the classical bracket is given by $[B^1_0,B^2_0]^c=[B^1(\hbar),B^2(\hbar)]^q|_{\hbar=0}$, where $B^i_0=B^i|_{\hbar=0}$. Therefore, there exists a Lie algebra homomorphism $\tilde{\mathcal{O}}^q_{\text{triv}}\xrightarrow{\hbar\to 0}\tilde{\mathcal{O}}^c_{\text{triv}}$. Following Refs. 10 and 6 we call $\tilde{\mathcal{O}}^c_{\text{triv}}$ the *Lie algebra of classical gauge parameters*. We thus see how it is related to the other algebras in (III.11).

Of the algebras entering (III.11), it only remains to construct $\mathbb{H}_{\mathcal{Z}_\mathbf{Q}}$, which we now do in the BV setting.

C. The Hamiltonian algebra of on-shell gauge symmetries

The BV field–antifield manifold $\mathcal{M}$ and the classical master action S satisfying the BV quantization axioms are such that $\mathcal{M}$ is a proper QP-manifold. Indeed, the odd vector field $\mathbf{Q} = (S,\cdot)$ on the $(N|N)$-dimensional antisymplectic manifold $\mathcal{M}$ preserves the antibracket, and therefore, satisfies condition (II.5); the master equation imposed on S implies that $\mathbf{Q}$ is nilpotent; finally, the fact that S is a proper solution of the master equation implies the rank condition (II.6).

The zero locus $\mathcal{Z}_{\mathbf{Q}}$ of $\mathbf{Q}$ determined by the equations $\partial_A S = 0$ will be referred to as the *stationary surface* of the action S. As before, we assume $\mathcal{Z}_{\mathbf{Q}}$ to be a smooth submanifold. (Although in realistic examples the structure of the zero locus of $\mathbf{Q}$ can be very involved, we treat $\mathcal{Z}_{\mathbf{Q}}$ as a submanifold. Note in passing that the finite-dimensional models of gauge systems should be considered with some caution also in view of the results of Ref. 18). Then, according to Theorem II.4, $\mathcal{Z}_{\mathbf{Q}}$ has a natural symplectic structure. Further, the classical gauge symmetries (see Definition III.2) are in fact symmetries of the proper QP-manifold $\mathcal{M}$.

Theorem III.3: *Every classical gauge symmetry $\mathbf{X}_0$ determines a vector field $\mathbf{x} = \mathbf{X}_{0|_{\mathcal{Z}_{\mathbf{Q}}}}$ on $\mathcal{Z}_{\mathbf{Q}}$ that preserves the Poisson bracket on $\mathcal{Z}_{\mathbf{Q}}$.*

Proof: Indeed, any vector field $\mathbf{X}_0$ preserving the master action S and the antibracket commutes with $\mathbf{Q} = (S,\cdot)$ and is therefore a symmetry of $\mathcal{M}$ (Definition II.5). As we saw in Sec. II B, any vector field $\mathbf{X}_0$ that is a symmetry of $\mathcal{M}$ restricts to $\mathcal{Z}_{\mathbf{Q}}$, and $\mathbf{X}_{0|_{\mathcal{Z}_{\mathbf{Q}}}}$ is locally Hamiltonian on $\mathcal{Z}_{\mathbf{Q}}$. ■

We denote the Lie algebra of locally Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$ by $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$. As we are going to see, this is the algebra of the on-shell gauge symmetries.

Definition III.4: A classical gauge symmetry $\mathbf{X}_0$ is called on-shell trivial if it vanishes on the stationary surface $\mathcal{Z}_{\mathbf{Q}}$.

We now show that the Lie algebra of locally Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$ is isomorphic to the algebra of the on-shell gauge symmetries.

Theorem III.5: *The algebra $\mathcal{I}_0$ of the on-shell trivial symmetries is an ideal in the Lie algebra $\mathcal{O}^c$ of gauge symmetries and the quotient algebra $\mathcal{O}^c/\mathcal{I}_0$ is isomorphic to the Lie algebra $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ of locally Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$.*

Proof: Let $\mathbf{Y}_0 \in \mathcal{I}_0$ and $\mathbf{X}_0 \in \mathcal{O}^c$. For any function $F \in \mathbb{C}_{\mathcal{M}}$, we have $\mathbf{Y}_0 F \in \mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$. Since $\mathbf{X}_0 \mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}} \subset \mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ and $\mathbf{Y}_0 F \in \mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$, we have $[\mathbf{X}_0,\mathbf{Y}_0]F = \mathbf{X}_0\mathbf{Y}_0 F - (-1)^{\epsilon(\mathbf{X}_0)\epsilon(\mathbf{Y}_0)}\mathbf{Y}_0\mathbf{X}_0 F \in \mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$. Therefore, $\mathcal{I}_0$ is an ideal in the Lie algebra $\mathcal{O}^c$. Further, we have seen in Theorem II.8 that any locally Hamiltonian vector field $\mathbf{x}$ on $\mathcal{Z}_{\mathbf{Q}}$ is a restriction of some vector field $\mathbf{X} \in \mathcal{O}^c$. Thus we can identify the quotient algebra $\mathcal{O}^c/\mathcal{I}_0$ with the Lie algebra of locally Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$. ■

It follows from the theorem that the homomorphism $\mathcal{O}^c \to \mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ is included into the exact sequence

$$0 \to \mathcal{I}_0 \to \mathcal{O}^c \to \mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}} \to 0. \tag{III.19}$$

Note that we cannot replace $\mathcal{O}^c$ with $\mathcal{O}^c_{\mathrm{triv}}$ here, because the homomorphism $\mathcal{O}^c_{\mathrm{triv}} \to \mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ is *not* surjective whenever there exists the first cohomology of $\mathcal{Z}_{\mathbf{Q}}$. Indeed, a nonvanishing first cohomology implies that there exist locally Hamiltonian vector fields that are not globally Hamiltonian on $\mathcal{Z}_{\mathbf{Q}}$, which we have seen in Theorem II.8 to correspond to *BRST-nontrivial* gauge symmetries. Due to the existence of the latter, the mapping $\mathcal{O}^c_{\mathrm{triv}} \to \mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ is not surjective in general.

Looking at diagram (III.11), it is natural to ask the following question: What is the analog of $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ for the upper line of the diagram, i.e. what is the *quantum* analog of the on-shell gauge symmetries? We propose one possible answer to this question.

Note that Poisson bracket (II.8) on $\mathcal{Z}_{\mathbf{Q}}$ and Lie bracket (III.17) on the space of gauge parameters are defined by the same bilinear operation $(\cdot, \mathbf{Q}\ \cdot)$ on $\mathbb{C}_{\mathcal{M}}$. The difference between these two brackets is that (III.17) is defined on the quotient space $\mathbb{C}_{\mathcal{M}}/\mathrm{Im}\,\mathbf{Q}$, while the Poisson bracket is defined on $\mathbb{C}_{\mathcal{M}}/\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$. For a proper QP-manifold, $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}}$ is the ideal generated by $\mathrm{Im}\,\mathbf{Q}$, i.e., $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}} = \mathbb{C}_{\mathcal{M}} \cdot \mathrm{Im}\,\mathbf{Q}$. At the same time, (III.17) is the limit as $\hbar \to 0$ of the quantum construction (III.14) defined on $\mathbb{C}_{\mathcal{M}}[[\hbar]]$. Therefore, in the quantum case one can construct a Poisson bracket as a direct generalization of (II.8), as $\{\cdot,\cdot\}^q = (\cdot, \delta_W \cdot)$. The bracket $\{\cdot,\cdot\}^q$ would be well defined only on the quotient algebra of $\mathbb{C}_{\mathcal{M}}[[\hbar]]$ modulo the ideal $\mathbb{I}_{\delta_W}$ generated by $\mathrm{Im}\,\delta_W$. Obviously, $\mathrm{Im}\,\delta_W$ is generated by all series of the form $\delta_W f(\hbar)$, where $f(\hbar) = f_0 + f_1\hbar + f_2\hbar^2 + \cdots$ with the coefficients f_i taking independently each value 1, Γ^A, and $\Gamma^A\Gamma^B$. Since the matrix $E^{AB} = (\Gamma^A, \Gamma^B)$ is invertible, we thus see that $\mathbb{I}_{\delta_W}$ consists of the series of the form $\mathbb{I}_{\mathcal{Z}_{\mathbf{Q}}} + \mathbb{C}_{\mathcal{M}}\hbar + \mathbb{C}_{\mathcal{M}}\hbar^2 + \cdots$. Thus, the quotient algebra $\mathbb{C}_{\mathcal{M}}[[\hbar]]/\mathbb{I}_{\delta_W}$ coincides with the algebra $\mathbb{C}_{\mathcal{Z}_{\mathbf{Q}}}$ of functions on the zero locus of $\mathbf{Q}$. This means that the algebra $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ is in a certain sense the most general algebra of the on-shell symmetries not only of the classical but also of the quantum, master action.

IV. EXAMPLES

A. Abelianized gauge theory

We now consider the field–antifield space and the master action corresponding to the simplest gauge theory, the Abelianized gauge theory, which we choose as an instructive example that is free of additional complications because gauge symmetries are explicitly separated from the physical ones. We then explicitly construct the Poisson bracket and the Lie algebras $\mathcal{O}^c$ and $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$. Moreover, this example shows that the classical gauge symmetries $\mathcal{O}^c$ contain the Lie algebra of gauge symmetries of the *original* theory as a subalgebra and that, similarly, the on-shell gauge symmetries $\mathbb{H}_{\mathcal{Z}_{\mathbf{Q}}}$ contain the on-shell symmetries of the Abelianized gauge theory.

Let $S_0(X,x)$ be a polynomial action such that

$$\partial_\alpha S_0 = 0, \quad \det(\partial_i \partial_j S_0)\big|_{\partial_i S_0 = 0} \neq 0, \tag{IV.1}$$

where we denote $\partial_\alpha = \partial/\partial x^\alpha$ and $\partial_i = \partial/\partial X^i$ and assume X^i and x^α to be bosonic for simplicity. Due to rank condition (IV.1), the equations $\partial_i S_0 = 0$ admit only a finite set of solutions $\mathfrak{M}$. Thus, the stationary surface of this theory is the direct product of $\mathfrak{M}$ with the space parametrized by x^α. The gauge transformations preserving the action S_0 are of the form

$$\mathbf{Y}_0 = Y_0^\alpha(X,x)\partial_\alpha + \mu^{ij}(X,x)\partial_i S_0 \partial_j, \tag{IV.2}$$

where $\mu^{ij}(X,x)$ is an antisymmetric matrix. These vector fields span a Lie algebra $\mathcal{A}$ with respect to the commutator of vector fields; those vanishing on the stationary surface span the ideal $\mathcal{A}_{\mathrm{triv}}$ in $\mathcal{A}$. Then $\tilde{\mathcal{A}} = \mathcal{A}/\mathcal{A}_{\mathrm{triv}}$ is the algebra of the on-shell gauge symmetries, which can be identified with the Lie algebra of vector fields on the stationary surface.

To implement the BV scheme, we choose the gauge generators in the form $R^\alpha_\beta = \delta^\alpha_\beta$ and introduce the ghosts c^α and the antifields x^*_α, X^*_i, and c^*_α. The canonical antibracket is $(\phi^A, \phi^*_B) = \delta^A_B$, where $\phi^A = (X^i, x^\alpha, c^\alpha)$ and $\phi^*_A = (X^*_i, x^*_\alpha, c^*_\alpha)$. Then the master action

$$S = S_0 + x^*_\alpha c^\alpha \tag{IV.3}$$

is a proper solution of the master equation $(S,S) = 0$. This action gives rise to the vector field

$$\mathbf{Q} = (S,\cdot) = \partial_i S_0 \frac{\vec{\partial}}{\partial X^*_i} + c^\alpha \frac{\vec{\partial}}{\partial x^\alpha} + x^*_\alpha \frac{\vec{\partial}}{\partial c^*_\alpha}, \tag{IV.4}$$

whose stationary surface $\mathcal{Z}_{\mathbf{Q}}$ is determined by $\partial_i S_0 = 0$, $x^*_\alpha = 0$, $c^\alpha = 0$. Thus $\mathcal{Z}_{\mathbf{Q}}$ is the direct product of $\mathfrak{M}$ (the set of solutions to the system of equations $\partial_i S_0 = 0$) with the space parametrized by $Y^A = \{X^*_i, x^\alpha, c^*_\alpha\}$. The Poisson bracket (II.8) on $\mathcal{Z}_{\mathbf{Q}}$ is then represented by the matrix

$$\Omega^{AB}=\{Y^A,\, Y^B\}=\begin{pmatrix} \partial_i\partial_j S_0 & 0 & 0 \\ 0 & 0 & \delta_\alpha^\beta \\ 0 & -\delta_\gamma^\nu & 0 \end{pmatrix}. \tag{IV.5}$$

We now want to show that the Lie algebras $\mathcal{A}$ and $\tilde{\mathcal{A}}$ of the original theory are subalgebras in the Lie algebras $\mathcal{O}^c$ and $\mathbb{H}_{\mathcal{Z}_\mathbf{Q}}$, respectively. To do so, we calculate $\mathcal{O}^c$ in the master theory with the master action (IV.3). Note that for the Abelianized gauge theory, the first cohomology group of the field–antifield space vanishes, and, therefore, each Hamiltonian vector field has a globally defined Hamiltonian. Thus, in order to find $\mathcal{O}^c$, it suffices to find the kernel of $\mathbf{Q}$ evaluated on the space of globally defined functions. Any element $A\in \mathrm{Ker}\,\mathbf{Q}$ can be written in the form $A=\mathbf{Q}F+G$, where F is an arbitrary smooth function on the field–antifield space and G is a representative of the cohomology class of $\mathbf{Q}$. To calculate the cohomology of $\mathbf{Q}$, we write $\mathbf{Q}=\mathbf{Q}_1+\mathbf{Q}_2$, where

$$\mathbf{Q}_1=\partial_i S_0\frac{\vec{\partial}}{\partial X_i^*},\quad \mathbf{Q}_2=c^\alpha\frac{\vec{\partial}}{\partial x^\alpha}+x_\alpha^*\frac{\vec{\partial}}{\partial c_\alpha^*}\quad\text{and}\quad \mathbf{Q}_1^2=\mathbf{Q}_2^2=[\mathbf{Q}_1,\mathbf{Q}_2]=0. \tag{IV.6}$$

(Note that in the case where $S_0=\frac{1}{2}\delta_{ij}X^iX^j$, the vector field $\mathbf{Q}$ is nothing but the De Rham differential of $\mathcal{Z}_\mathbf{Q}$. In this case the cohomology of $\mathbf{Q}$ consists of constants only and $\mathcal{O}^c$ coincides with $\mathcal{O}^c_{\mathrm{triv}}$.)

By the Poincaré lemma, the cohomology of $\mathbf{Q}_2$ consists of constants only. Thus the cohomology of $\mathbf{Q}$ is determined by the cohomology of $\mathbf{Q}_1$ on the space of functions $F(X,X^*)$. A function $F(X,X^*)$ belongs to the image of $\mathbf{Q}_1$ whenever $F(X,X^*)=\partial_i S_0 f^i(X,X^*)$, i.e., $F(X,X^*)$ vanishes at each point where $\partial_i S_0=0$. Thus, any element A from $\mathrm{Ker}\,\mathbf{Q}$ is of the form

$$A(X,X^*)=\mathbf{Q}F(X,X^*)+G(X), \tag{IV.7}$$

where F is an arbitrary function and $G(X)$ is a function that does not vanish at least at one point of $\mathfrak{M}$. Whenever $\mathfrak{M}$ is an n-point set, the cohomology of $\mathbf{Q}$ is an n-dimensional vector space. (In this example, the group of ''physical'' symmetries is the group of permutations of these n points. This group obviously acts on the cohomology of $\mathbf{Q}$.)

We thus see that $\mathcal{O}^c$ and $\mathcal{O}^c_{\mathrm{triv}}$ are spanned by the vector fields of the form $(\mathbf{Q}F+G(X),\cdot)$ and $(\mathbf{Q}F,\cdot)$ respectively. The algebra $\mathbb{H}_{\mathcal{Z}_\mathbf{Q}}$ of the on-shell symmetries consists of Hamiltonian vector fields $\{H(\mathfrak{m},X^*,x,c^*),\cdot\}$ on $\mathcal{Z}_\mathbf{Q}$ (where we label the Hamiltonian by $\mathfrak{m}\in\mathfrak{M}$ enumerating the different components of $\mathfrak{M}$).

To see that the algebra $\mathcal{A}$ of gauge transformations of the original theory is embedded into the Lie algebra $\mathcal{O}^c$ of the classical gauge symmetries, we note that vector fields of the form

$$\mathbf{Y}=(Y_0^\alpha(X,x)x_\alpha^*+\mu^{ij}(X,x)\partial_i S_0 X_j^*,\cdot) \tag{IV.8}$$

form a subalgebra in $\mathcal{O}^c$. Moreover, these fields restrict to the subspace $c^\alpha=c_\alpha^*=x_\alpha^*=X_i^*=0$ as elements of $\mathcal{A}$ [see (IV.2)]. Thus, we have an embedding of $\mathcal{A}$ into $\mathcal{O}^c$ (obviously, the embedding is not unique).

As regards the on-shell gauge symmetries, observe that the vector fields on $\mathcal{Z}_\mathbf{Q}$ of the form

$$\mathbf{y}=\{y^\alpha(\mathfrak{m},x)c_\alpha^*,\cdot\} \tag{IV.9}$$

(which define a subalgebra in $\mathbb{H}_{\mathcal{Z}_\mathbf{Q}}$) restrict to the stationary surface of the original theory (which is a submanifold of $\mathcal{Z}_\mathbf{Q}$ determined by the equations $c_\alpha^*=0$ and $X_i^*=0$) and span $\tilde{\mathcal{A}}$. Thus the algebra of the on-shell gauge symmetries $\tilde{\mathcal{A}}$ of the original theory is embedded into the Lie algebra $\mathbb{H}_{\mathcal{Z}_\mathbf{Q}}$ of the on-shell gauge symmetries.

B. A ''topological'' field theory

We now apply Theorem II.4 to the ''topological'' theory with the vanishing action on a Lie group $\mathcal{G}$. We show that in this case, bracket (II.8) is related to the Kirillov bracket on the coalgebra. Denote by x^i a coordinate system in the neighborhood of $1 \in \mathcal{G}$. Let $\mathcal{R}_\alpha = \mathcal{R}^i_\alpha \partial_i$ (where the Greek indices have the same cardinality as the Latin ones) be the basis of the left invariant vector fields on $\mathcal{G}$. We have $[\mathcal{R}_\alpha, \mathcal{R}_\beta] = \mathcal{F}^\gamma_{\alpha\beta}\mathcal{R}_\gamma$, where $\mathcal{F}^\gamma_{\alpha\beta}$ are the structure constants.

In accordance with the BV prescription we introduce the ghosts c^α and the antifields x^*_i and c^*_α such that $(x^i, x^*_j) = \delta^i_j$ and $(c^\alpha, c^*_\beta) = \delta^\alpha_\beta$. The master action $S = x^*_i R^i_\alpha c^\alpha - \frac{1}{2} c^*_\gamma \mathcal{F}^\gamma_{\alpha\beta} c^\beta c^\alpha$ is a proper solution of $(S,S)=0$. Then

$$\mathbf{Q} = (S,\cdot) = c^\alpha R^i_\alpha \frac{\vec{\partial}}{\partial x^i} + \frac{1}{2}\mathcal{F}^\gamma_{\alpha\beta} c^\beta c^\alpha \frac{\vec{\partial}}{\partial c^\gamma} + (x^*_i R^i_\alpha - c^*_\gamma \mathcal{F}^\gamma_{\alpha\beta} c^\beta)\frac{\vec{\partial}}{\partial c^*_\alpha}, \tag{IV.10}$$

and therefore the zero locus $\mathcal{Z}_{\mathbf{Q}}$ of $\mathbf{Q}$ is determined by the equations $c^\alpha = 0$ and $x^*_i = 0$ and is coordinatized by $Y^A = \{x^i, c^*_\alpha\}$. In this case $\mathcal{Z}_{\mathbf{Q}} = T^*\mathcal{G} = \mathcal{G} \times \mathfrak{g}^*$ is the cotangent bundle to $\mathcal{G}$, where $\mathfrak{g}^*$ is the coalgebra. The matrix of Poisson bracket (II.8) takes the form

$$\Omega^{AB} = \{Y^A, Y^B\} = \begin{pmatrix} \{x^i, x^j\} & \{x^i, c^*_\beta\} \\ \{c^*_\alpha, x^j\} & \{c^*_\alpha, c^*_\beta\} \end{pmatrix} = \begin{pmatrix} 0 & R^i_\beta \\ -R^j_\alpha & -c^*_\gamma \mathcal{F}^\gamma_{\alpha\beta} \end{pmatrix}. \tag{IV.11}$$

It is nondegenerate because R^i_α are nondegenerate everywhere on $\mathcal{Z}_{\mathbf{Q}}$, since R^i_α are the coefficients of the left invariant vector fields on a Lie group. The cotangent bundle to a Lie group is trivial, therefore we have the embedding $\mathfrak{g}^* \to T^*\mathcal{G}$, which induces a Poisson bracket on $\mathfrak{g}^*$. This gives us the Kirillov bracket[15] on the coalgebra $\mathfrak{g}^*$ parametrized by the coordinates c^*.

V. CONCLUSIONS

We have seen that a number of objects of the antisymplectic BV geometry are essentially determined by objects of the symplectic geometry on the stationary surface of the master action, where the nondegenerate Poisson bracket is given by (II.8). In particular, every observable determines a symmetry of the master action, which in turn restricts to a locally Hamiltonian vector field on $\mathcal{Z}_{\mathbf{Q}}$; at the same time, every trivial observable determines a symmetry of the master action such that the corresponding vector field on $\mathcal{Z}_{\mathbf{Q}}$ is globally Hamiltonian. Those Hamiltonian vector fields on $\mathcal{Z}_{\mathbf{Q}}$ that are not globally Hamiltonian correspond then to the BRST-nontrivial observables.

Recalling how the master theory is constructed in terms of the bare classical action $\mathcal{S}$, we were able to explicitly see, in the Abelianized setting, that the gauge symmetries of $\mathcal{S}$ are dressed into master theory symmetries, i.e., into (,)-Hamiltonian vector fields; at the same time, the on-shell gauge symmetries of $\mathcal{S}$ are dressed into { , }-Hamiltonian vector fields on the symplectic manifold $\mathcal{Z}_{\mathbf{Q}}$. This would be interesting to extend to the general gauge theory setting.

Our analysis was performed in the framework of a finite-dimensional model; such models should be viewed with caution precisely for the reasons related to the existence of the BRST cohomology.[18] It would be interesting to see how our results can be reformulated in local field theory, where the gauge symmetries have been discussed in Ref. 19, and, possibly, also in application to string field theory,[5,6] which has been one of the motivations behind the geometrically covariant reformulation of the BV quantization.

ACKNOWLEDGMENTS

We are grateful to I. Batalin and I. Tyutin for many illuminating discussions on various aspects of the field–antifield quantization and the related problems. A.M.S. wishes to thank P. H. Damgaard for helpful discussions and for kind hospitality at the Niels Bohr Institute where a part of this paper was written. We also appreciated discussions with O. Khudaverdyan, A. Nersessian, and B. Voronov.

This work was supported in part by the RFBR Grant No. 96-01-00482 and by the INTAS-RFBR Grant No. 95-0829.

[1] I. A. Batalin and G. A. Vilkovisky, Phys. Lett. **B102**, 27 (1981).
[2] I. A. Batalin and G. A. Vilkovisky, Phys. Rev. D **28**, 2567 (1983).
[3] I. A. Batalin and I. V. Tyutin, Int. J. Mod. Phys. A **8**, 2333 (1993); Mod. Phys. Lett. A **8**, 3673 (1993); **9**, 1707 (1994).
[4] A. Schwarz, Commun. Math. Phys. **155**, 249 (1993).
[5] H. Hata and B. Zwiebach, Ann. Phys. **229**, 177 (1994).
[6] A. Sen and B. Zwiebach, Commun. Math. Phys. **177**, 305 (1996).
[7] A. Schwarz, Commun. Math. Phys. **158**, 373 (1993).
[8] O. M. Khudaverdian, J. Math. Phys. **32**, 1934 (1991); O. M. Khudaverdian and A. Nersessian, Mod. Phys. Lett. A **8**, 2377 (1993).
[9] K. Bering, ''Almost Parity Structure, Connections and Vielbeins in BV Geometry,'' MIT-CTP-2682, physics/9711010.
[10] A. Sen and B. Zwiebach, Phys. Lett. B **320**, 29 (1994).
[11] M. Alexandrov, M. Kontsevich, A. Schwarz, and O. Zaboronsky, Int. J. Mod. Phys. A **12**, 1405 (1997).
[12] A. Nersessian and P. H. Damgaard, Phys. Lett. B **355**, 150 (1995).
[13] I. A. Batalin, R. Marnelius, and A. M. Semikhatov, Nucl. Phys. B **446**, 249 (1995).
[14] I. A. Batalin and G. A. Vilkovisky, Nucl. Phys. B **234**, 106 (1984); I. A. Batalin and E. S. Fradkin, J. Math. Phys. **25**, 2426 (1984); B. L. Voronov and I. V. Tyutin, Teor. Mat. Fiz. **50**, 333 (1982).
[15] A. A. Kirillov, *Elements of the Theory of Representations* (Springer, Berlin, 1970).
[16] O. M. Khudaverdian and A. P. Nersessian, J. Math. Phys. **37**, 3713 (1996); A. P. Nersessian, Preprint JINR E2-93-225; Preprint JINR E2-93-358.
[17] A. Schwarz, Lett. Math. Phys. **31**, 299 (1994).
[18] I. A. Batalin and I. V. Tyutin, Theor. Math. Phys. **114**, 198 (1998).
[19] G. Barnich and M. Henneaux, J. Math. Phys. **37**, 5273 (1996).

On path integral localization and the Laplacian

Topi Kärki[a)]
Institut Mittag-Leffler, Auravagen 17, S-18262 Djursholm, Sweden

(Received 10 August 1998; accepted for publication 15 January 1999)

We introduce a new localization principle, which is a generalized canonical transformation. It unifies BRST localization, the non-Abelian localization principle and a special case of the conformal Duistermaat–Heckman integration formula of Paniak, Semenoff, and Szabo. The heat kernel on compact Lie groups is localized in two ways. First, using a non-Abelian generalization of the derivative expansion localization of Palo and Niemi and second, using the BRST localization principle and a configuration space path integral. In addition, we present some new formulas on homogeneous spaces, which might be useful in a possible localization of Selberg's trace formula on locally homogeneous spaces. © *1999 American Institute of Physics.* [S0022-2488(99)01804-6]

I. INTRODUCTION

Integral localization is a method to calculate path integrals. It conventionally involves a BRST symmetry and a one-parameter localization deformation of the action, which in the limit that the localization parameter is put to infinity evaluates the integral.[1–3] The result is usually a sum over the critical points of the action or an integral over a finite-dimensional subspace of the original infinite-dimensional integration domain. The method has been effective in topological field theories.[4]

A special case of it is the Duistermaat–Heckman integration formula[5] and its loop space version,[6] where the BRST symmetry is another way of writing the definition of the Hamiltonian vector field. A more technical introduction to the Hamiltonian BRST symmetry is given in Sec. II.

Recently there has emerged localizations that cannot be understood from the conventional BRST point of view: the non-Abelian localization principle[7] and the conformal Duistermaat–Heckman formula of Paniak, Semenoff, and Szabo.[8] In Secs. III, IV, and V they are unified in a new localization principle.

The two main questions behind this article are ''why is the heat kernel of the Laplacian on compact Lie groups semiclassically exact?''[9,10] and ''why are there Selberg's trace formulae?''[11–14] The new localization principle was also discovered in studying these questions. We review shortly the facts that motivated them. The heat kernel on Lie groups has been proved to be semiclassically exact[9,10] by direct comparison. On the other hand, the loop space Duistermaat–Heckman theorem[6] explains semiclassical exactness, in the case that the Hamiltonian vector field is also a Killing vector field, using a path integral localization proof. It is not, however, the case (in the obvious way) on Lie groups. In addition, there is Selberg's trace formula on constant negative curvature Riemann surfaces,

$$\operatorname{tr} e^{-\beta\Delta_0}=F+(4\pi\beta)^{-1/2}e^{-\beta/4}\sum_{n=1}^{\infty}\sum_{p}\frac{l(p)/2}{\sin l(p^n)/2}e^{-l^2(p^n)/4\beta}, \tag{1}$$

where F is the fixed point contribution.[13] p means a primitive geodesic and $l(p)$ the length of it. The formula is similar to the semiclassical approximation because it holds that $S=\frac{1}{4}\int_0^\beta g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu=l^2(p^n)/4\beta$ at the critical points of the action S. The negative constant curvature Riemann sur-

[a)]Electronic mail: karki@ml.kva.se

faces can be obtained as quotient spaces $\Gamma\backslash[\mathrm{SU}(1,1)/\mathrm{U}(1)]$, where Γ is a discrete subgroup.[12] This, together with the semiclassical exactness on Lie groups and the fact that there are also generalizations of Selberg's trace formula on some symmetric spaces,[14] leads to the speculation of a localization of the Laplacian on locally homogeneous manifolds $\Gamma\backslash(G/H)$.

The answers that we provide to the questions above are organized as follows.

A localization deformation of the Laplacian on homogeneous spaces is given in Sec. VI. The formulas and notations are presented in Appendices A and B because most of them are not needed in the rest of the discussion. They involve some new formulas: the use of a degenerate basis for vector fields is new as well as the Maurer–Cartan connection on the tangent bundle and the phase space metric. In Appendix D the scalar curvature is calculated. In addition, a formula is presented that might appear in the hypothetical localization on locally homogeneous manifolds.

The question about the heat kernel on Lie groups is answered partially in Sec. VII using a non-Abelian derivative expansion localization. The localization principle is even more interesting because it formally seems to apply to homogeneous manifolds as well, and because it may even apply to integrable models (Sec. IX). The final answer to the question is given in Sec. VIII using a configuration space path integral and the symplectic form pointed out by Picken in Ref. 10.

The question about Selberg's trace formulas is left open, despite the negative result, that it is not the deformation in Sec. VI, and the speculations in Sec. IX.

In addition, a new bound for the geodesic action is presented in Appendix C.

II. BRST LOCALIZATION PRINCIPLE

We review the (Hamiltonian) BRST localization principle, introducing some notations.

We assume that Γ is a $2D$-dimensional phase space, ω a symplectic form, and H a Hamiltonian. The classical partition function of the Hamiltonian system is

$$Z=\int_\Gamma \omega^D e^{-\beta H}, \tag{2}$$

and it is calculated using the BRST localization principle. The definition of the Hamiltonian vector field χ can be written as

$$(d+i\chi)(H+\omega)=0, \tag{3}$$

where d is the exterior derivative and i_χ is a contraction operator. The equivariant derivative,

$$d+i_\chi=d_\chi=Q,$$

is a BRST symmetry with the exception that it is closed only in the invariant subspace,

$$Q^2\alpha=L_\chi\alpha=0,$$

where L_χ is the Lie derivative and α is a differential form on the phase space. The BRST localization principle can be formulated as follows: Analogously to the Fradkin–Vilkovisky theorem, the partition function (see the explanation of notations below),

$$Z_\lambda=\int d\phi^\mu\, d\psi^\mu\, e^{-\beta(H+\omega)+\lambda\, d_\chi\psi}, \tag{4}$$

is independent of λ, provided that the arbitrary one-form ψ satisfies the Lie derivative condition

$$L_\chi\psi=0. \tag{5}$$

In (4), ϕ^μ are the coordinates on the phase space Γ and the Grassmann variables ψ^μ are associated with the one-forms $d\phi^\mu$ on the phase space. In particular, the symplectic form has been replaced by a bilinear in the fermionic variables $\omega_{\mu\nu}\, d\phi^\mu\wedge d\phi^\nu\to\omega_{\mu\nu}\psi^\mu\psi^\nu$. When λ vanishes the integral reduces to the partition function (2) modulo an irrelevant factor of $(-\beta)^D$. (We use throughout the paper the convention that such normalization factors are neglected, which is usual with path integrals.) If one makes a clever choice of the one-form ψ and takes the localization parameter λ to infinity, one is able to evaluate the integral.

One can also give a loop space version of the localization principle. The path integral describing the quantum partition function is

$$Z_\lambda(\beta)=\int [d\phi^\mu\, d\psi^\mu]\exp\int_0^\beta \theta_\mu\dot{\phi}^\mu - H+\omega+\lambda\, d_S\psi, \tag{6}$$

where ω is the loop space symplectic form, θ is the symplectic potential, and d_S is the loop space equivariant derivative,

$$d_S=d+i_{\dot{\phi}-\chi}=d+i_{\chi_S}. \tag{7}$$

(We use the convention that the imaginary unit i is missing.) In addition, χ_S is called the loop space Hamiltonian vector field and $\dot{\phi}=\dot{\phi}^\mu(\partial/\partial\phi^\mu)\equiv\int_0^\beta dt\, \dot{\phi}^\mu(t)[\delta/\delta\phi^\mu(t)]$ is a vector field on the loop space. We use the notation that the integral sign is not written and the functional derivatives are written as ordinary derivatives.[6] The loop space one-form ψ must also satisfy the Lie derivative condition,

$$L_{\chi_S}\psi=0.$$

Various applications of the principle can be found in Refs. 6, 15, 16 and 17, which, however, is not an exhaustive list.

III. NEW LOCALIZATION PRINCIPLE

We discuss the new localization principle and give the first example of a localization deformation that is obtained using it.

We define that

$$[S(\lambda),\chi(\lambda),\psi] \tag{8}$$

is a triple of the new localization principle (or triple) if the following occurs.

(1) $S(\lambda)=H(\lambda)+\omega(\lambda)$ is a one-parameter family of Hamiltonian structures, with the exception that the symplectic form can be degenerate except when λ is zero,

(2) $\chi(\lambda)$ is a one-parameter family of vector fields satisfying

$$d_{\chi(\lambda)}S(\lambda)=0. \tag{9}$$

(3) ψ is such a one-form that

$$\frac{dS}{d\lambda}=d_{\chi(\lambda)}\psi. \tag{10}$$

Theorem 1: *If* $[S(\lambda),\chi(\lambda),\psi]$ *is a triple, then the partition function,*

$$Z_\lambda=\int d\phi^\mu\, d\psi^\mu\, e^{S(\lambda)}, \tag{11}$$

is independent of λ.

Proof: If one makes an infinitesimal change of variables in the direction of the supervector field $V=\psi d_{\chi(\lambda)}$, the integrand is invariant, and the Jacobian is $1+\epsilon S\text{div}\, V=1+\epsilon\, d_{\chi(\lambda)}\psi$ which proves that $Z(\lambda+\epsilon)=Z(\lambda)$. □

We make few remarks on the theorem that is the new localization principle.

Remark 1: The definition of the principle is complicated due to the fact that one cannot assume that $\omega(\lambda)$ is nondegenerate for all λ. If one could neglect the complication, taking any point $H+\omega$ in the space of Hamiltonian structures and any one-form ψ on the phase space would give a one-parameter flow of Hamiltonian structures, because then $\chi(\lambda)$ is uniquely determined as

$$\chi(\lambda)=\omega(\lambda)^{-1}\, dH(\lambda). \tag{12}$$

Remark 2: There are some interesting special cases of the localization principle. If the one-form is exact, $\psi=dF$, it reduces to a canonical transformation generated by the function F. On the other hand, if $L_\chi\psi=0$ and $\chi(\lambda)=\chi$, it is the BRST localization principle.

Remark 3: The one-form ψ is in many examples of the principle associated with a phase space metric g that is contracted with some vector field, for example, $\psi=i_{\chi(0)}g$.

Remark 4: The loop space version of the principle is obtained by thinking that the phase space is the infinite-dimensional loop space.

Remark 5: If $\omega(\lambda)=\omega(0)+\lambda\, d\psi$ is nondegenerate for all λ, the localization principle further simplifies. We define another vector field,

$$u(\lambda)=\omega(\lambda)^{-1}\psi, \tag{13}$$

and write the flow equation (10) in the alternative form

$$\frac{dS}{d\lambda}=-L_{u(\lambda)}S, \tag{14}$$

which is a diffeomorphism of the phase space. It is then possible to transport any tensor T along the flow (14) according to the equation

$$\frac{dT}{d\lambda}=-L_{u(\lambda)}T. \tag{15}$$

One should not think that the invertibility of $\psi(\lambda)$ is generic, for example, the Duistermaat–Heckman theorem on a compact phase space[5] gives a counterexample, which can be deduced as follows: On a compact phase space the Hamiltonian function has a maximum and minimum, which cannot change under any diffeomorphism flow. (We thank Losev for this observation.) But in the Duistermaat–Heckman case the maximum of the Hamiltonian function goes to infinity under the localization flow because $H(\lambda)=H(0)+\lambda g(\chi(0),\chi(0))$, where g is a phase space metric. Thus, the flow cannot be a diffeomorphism and therefore $\omega(\lambda)$ must be degenerate for some λ.

Finally, we give the first example of the localization principle. Suppose that the phase space admits a metric g such that

$$\nabla_\chi\chi=0. \tag{16}$$

Then $[\frac{1}{2}g((1+\lambda)\chi,(1+\lambda)\chi)+d(i_{(1+\lambda)\chi}g),(1+\lambda)\chi,i_\chi g]$ is a triple of the new localization principle, provided that the two-form $d(i_\chi g)$ is nondegenerate. For example, the geodesic motion on homogeneous manifolds has this structure; there is actually a one-parameter family of metrics (B13) satisfying the condition (16). As a triple, it is a very special case because λ can be actually any function that satisfies $L_\chi\lambda=0$ and the partition function is still independent of λ. The conditions of a triple can be proved using the identity[5]

$$\nabla_\chi\chi=0\Leftrightarrow d_\chi[\tfrac{1}{2}g(\chi,\chi)+d(i_\chi g)]=0.$$

IV. NON-ABELIAN LOCALIZATION PRINCIPLE

We review the non-Abelian localization principle, making some additions to the original discussion in Ref. 7. In the end of the section it is interpreted using the new localization principle.

The partition function for two-dimensional gauge theories is

$$Z=\int [dA_\mu^B]e^{-(1/e)\int \mathrm{tr}\, F_{\mu\nu}F^{\mu\nu}}. \tag{17}$$

The configuration space $\mathcal{A}$ of gauge potentials that is integrated over has a symplectic form,

$$\omega=\int dx^\mu\wedge dx^\nu\ \delta A_\mu^B\wedge\delta A_\nu^B=\frac{1}{2}\int \sqrt{g}d^2x\sqrt{g}\epsilon^{\mu\nu}\ \delta A_\mu^B\wedge\delta A_\nu^B, \tag{18}$$

a metric,

$$g=\int \sqrt{g}d^2x\, g^{\mu\nu}\ \delta A_\mu^B\otimes\delta A_\nu^B, \tag{19}$$

and an almost complex structure,

$$J=\omega^{-1}g,\quad J^2=-\mathbf{1}. \tag{20}$$

The partition function (17) is actually an integration over the Liouville measure of the symplectic form (18). In addition, the ''Hamiltonian'' in the partition function (17) is the quadratic Casimir of the group of gauge transformations: The gauge transformations act symplectically and the momentum map is μ: $\mathcal{A}\to g^*$ (g is the Lie algebra of gauge transformations),

$$\mu(A)=F=dA+A\wedge A. \tag{21}$$

In other words,

$$\mu_\epsilon=\int F_{\mu\nu}^B\epsilon^B\, dx^\mu\wedge dx^\nu=\int \sqrt{g}d^2x(\sqrt{g}\epsilon^{\mu\nu}F_{\mu\nu}^B)\epsilon^B \tag{22}$$

generates the gauge transformation

$$\{\mu_\epsilon,A_\mu^B\}=D_\mu\epsilon^B. \tag{23}$$

The associated Hamiltonian vector fields are

$$v_\epsilon=\omega^{-1}\, d\mu_\epsilon=\int \sqrt{g}d^2x\, D_\mu^B\epsilon\frac{\delta}{\delta A_\mu^B}, \tag{24}$$

which generate isometries of the metric,

$$L_{v_\epsilon}g=0. \tag{25}$$

Using the Hamiltonian generators of the Lie algebra of gauge transformations,

$$\mu^{Bx}=\mu_{\epsilon=\delta^{BC}\delta(x-y)}=\sqrt{g}\epsilon^{\mu\nu}F_{\mu\nu}^B, \tag{26}$$

one can write the Hamiltonian in (17) as the quadratic Casimir,

$$S=\int \sqrt{g}d^2x\, \mu^{Bx}\mu^{Bx}=\int \sqrt{g}d^2x\, F_{\mu\nu}^BF_B^{\mu\nu}. \tag{27}$$

The discussion above is analogous to the following finite-dimensional situation: We assume that $C=\Sigma_i\mu_i^2$ is the quadratic Casimir of a group that acts symplectically on the phase space and that there is a metric g that is invariant. The partition function is

$$Z=\int \omega^n e^{-\Sigma\mu_i^2}=\int dx^\mu\, dc^\mu\, e^{-\Sigma\mu_i^2+\omega} \tag{28}$$

$$=\int d\phi_i e^{(1/4)\Sigma\phi_i^2}\int dx^\mu\, dc^\mu\, e^{\Sigma\phi_i\mu_i+\omega}, \tag{29}$$

where $\Sigma\phi^i\mu_i+\omega$ is closed by the equivariant derivative

$$d+i_{\phi^i v_i}, \tag{30}$$

and v_i is the Hamiltonian vector field of μ_i. (See Ref. 7 for the correct reinsertion of the imaginary unit.) We choose a one-form ψ on the phase space that is invariant under the action of the group,

$$L_{v_i}\psi=0, \tag{31}$$

for all v_i. One can use the BRST localization principle to prove that

$$Z=\int d\phi_i e^{(1/4)\Sigma\phi_i^2}\int dx^\mu\, dc^\mu\, e^{\Sigma\phi_i\mu_i+\omega+\lambda(d+i_{\phi^i v_i})\psi}, \tag{32}$$

$$=\int dx^\mu\, dc^\mu\, e^{-\Sigma_i(\mu_i+\lambda i_{v_i}\psi)^2+\omega+\lambda\, d\psi}, \tag{33}$$

is independent of λ. The limit $\lambda\to\infty$ yields a localization on

$$i_{v_i}\psi=0. \tag{34}$$

We choose more specifically $\psi=i_\chi g$ $(\chi=\{\Sigma\mu_i^2,\}=2\Sigma\mu_i v_i)$, which is equivalent to the one-form in Ref. 7, and get a localization on the critical points of $\Sigma\mu_i^2$, which is the non-Abelian localization principle.

Equation (33) can be interpreted as a consequence of the new localization principle because $[-\Sigma_i(\mu_i+\lambda i_{v_i}\psi)^2+\omega+\lambda\, d\psi, 2\Sigma(\mu_i+\lambda i_{v_i}\psi)v_i, i_\chi g]$ is a triple. The conditions of a triple can be proved using the identity $(d+i_{v_i})(\mu_i+\omega+\lambda\, d_{v_i}\psi)=0$. We are reminded that, neglecting possible complications due to the degeneracy of $\omega(\lambda)$, one deforms the action $S=-\Sigma_i\mu_i^2+\omega$ using the flow (10), where the one-form is $\psi=i_\chi g$.

V. CONFORMAL DUISTERMAAT–HECKMAN FORMULA

We derive the conformal Duistermaat–Heckman formula using the new localization principle. We have to impose, however, a restriction that is not present in the original approach in Ref. 8.

In the following it is assumed that the phase space admits a metric g such that the Hamiltonian vector field χ is a conformal Killing vector,

$$L_\chi g=\Lambda g. \tag{35}$$

Λ is a function on the phase space. In addition, the following is assumed.

(1) There is a one-parameter family of vector fields $\chi(\lambda)$, such that

$$d_{\chi(\lambda)}[H+\omega+\lambda d_\chi(i_\chi g)]=0. \tag{36}$$

(2) Here

$$g(\chi(\lambda),\chi)=g(\chi,\chi) \text{ at the points } p, \text{ where } \Lambda \text{ vanishes,} \tag{37}$$

which follows if $\chi(\lambda)=\chi$ at the points p or if $\omega(\lambda)=\omega+\lambda d(i_\chi g)$ is nondegenerate for all λ.

Lemma 1: $[H+\omega+\lambda d_\chi(i_\chi g),\chi(\lambda),i_\chi g]$ *is a triple of the new localization principle.*

Proof: We prove only the condition (10) of a triple because the other two conditions are evident. If λ is zero, (10) is trivially true, and we can assume that it is nonzero. A conformal Killing vector satisfies the equation

$$dg(\chi,\chi)=\Lambda i_\chi g-i_\chi d(i_\chi g),$$

and using the condition (36) we get the identity

$$i_{\chi(\lambda)}\omega(\lambda)=-d[H+\lambda g(\chi,\chi)]=i_\chi\omega(\lambda)-\lambda\Lambda i_\chi g.$$

If $\Lambda\neq 0$, we get that

$$i_{\chi(\lambda)}i_\chi g=\frac{1}{\lambda\Lambda}i_{\chi(\lambda)}i_\chi\omega(\lambda)=\frac{1}{\lambda\Lambda}i_\chi\, dH(\lambda)=g(\chi,\chi),$$

which is enough to prove (10). If Λ vanishes then (10) follows directly from the restriction (37).□

Because of the new localization principle, the partition function,

$$Z_\lambda=\int d\phi^\mu\, d\psi^\mu\, e^{S(\lambda)}, \tag{38}$$

is independent of λ, and in the limit $\lambda\to-\infty$ the action produces a delta function $\delta(\chi)$, which localizes the integral to the critical points of the Hamiltonian H. Calculation of the integral gives the Duistermaat–Heckman formula

$$Z=\sum_{dH=0}\frac{\sqrt{\det\omega_{\mu\nu}}}{\sqrt{\det\dfrac{\partial^2 H}{\partial\phi^\mu\partial\phi^\nu}}}e^{-\beta H}. \tag{39}$$

We have used the fact that $\Lambda=0$ at the critical points that follows from the formula[8]

$$\Lambda=\frac{1}{2D}\nabla_\lambda\chi^\lambda=\frac{1}{2D}(\nabla_\lambda\omega^{\lambda\rho})\partial_\rho H.$$

The details of the calculation can be found in Ref. 5.

VI. LOCALIZATION DEFORMATION OF THE LAPLACIAN ON HOMOGENEOUS SPACES

We derive a localization deformation (51) of the Laplacian (40), (43) on homogeneous spaces using the new localization principle. It is demonstrated that the deformation does not give the desired localization to the geodesics. However, it is worth presenting as a nontrivial solvable example of the triple, or more accurately a combination of the two of them. We mention also another flow that might localize, but it seems to be nonsolvable in closed form, and it is not clear if the conditions of a triple are satisfied (we conjecture that they are satisfied). See Appendices A and B for notations.

We consider the partition function

$$Z=\operatorname{tr} e^{-\beta[(1/2)\Delta_0-J^i v_i]}, \tag{40}$$

where β, J^i are constants and Δ_0 is the zero-form part of the Laplacian $\Delta = dd^* + d^*d$ on homogeneous manifolds. $e^{\beta J^i v_i}$ is the translation operator in the direction of the isometry $J^i v_i$. The partition function (40) coincides with the heat kernel that is integrated over the manifold:

$$Z = \int \sqrt{g} dx \langle x | e^{-\beta[(1/2)\Delta_0 - J^i v_i]} | x \rangle = \int \sqrt{g} dx \langle x | e^{-(1/2)\beta\Delta_0} | e^{\beta J^i v_i} x \rangle, \tag{41}$$

$$= \int \sqrt{g} dx\, k_\beta(x, e^{\beta J^i v_i} x). \tag{42}$$

The path integral presentation of (40) is

$$Z = \int [dx^\mu\, dp_\mu\, d\psi^\mu\, d\overline{\psi}_\mu] \exp \int_0^\beta I_i \omega^i(\dot{x}) - \frac{1}{2} K^{ij} I_i I_j + J^i I_i + d(I_i \omega^i), \tag{43}$$

where $\frac{1}{2}K^{ij}I_iI_j = \frac{1}{2}g^{\mu\nu}p_\mu p_\nu$ is the Hamiltonian of the geodesic motion and the Grassmann variables are associated with the one-forms, $\psi^\mu \sim dx^\mu$, $\overline{\psi}_\mu \sim dp_\mu$. In addition, there is DeWitt's term;[18] we assume that it is proportional to the scalar curvature. (There has been some controversy about this term; see Ref. 18 for a recent discussion and the speculation at the end of Sec. VIII.) We have neglected it because homogeneous manifolds are of constant scalar curvature and it yields only a shift in the energy levels.

The straightforward way to derive a localization deformation of the action in Eq. (43) would be to choose the loop space one-form,

$$\psi = i_{\chi_S} g,$$

where g is the phase space metric (B13) (lifted[6] on the loop space) and χ_S is the loop space Hamiltonian vector field,[6]

$$\chi_S = \dot{\phi} - \chi + J^i v_i^H, \quad \chi = \tfrac{1}{2} K^{ij} I_i v_j^H,$$

and apply the flow equation (10), neglecting the degeneracy problem. One can think (hopefully) that the inverse of the symplectic form becomes singular at some points of the phase space but $\chi_S(\lambda) = \omega(\lambda)^{-1} dH(\lambda)$ stays finite, we conjecture that this is the case. However, we have not been able to solve the flow equation. A power series solution experiment seems to give an infinity of different terms, which probably indicates that the flow is not solvable in closed form. However, this flow might localize to the geodesics if one could extract its asymptotic behavior somehow.

The following refined approach gives a localization deformation that is solvable and does not suffer from the degeneracy problem of the symplectic form; the latter is argued afterward for reasons of pedagogy. We use two one-forms [g_1 and g_2 are the two components of the metric (B13)],

$$\psi_i = i_{\chi_S} g_i, \quad i = 1,2,$$

both of which give a flow that is exactly solvable and localizes half of the degrees of freedom. In addition, ψ_2 satisfies

$$L_{\chi_S} \psi_2 = 0, \tag{44}$$

so that the flow in the direction of it is just a BRST flow,

$$S \to S + \alpha\, d_{\chi_S} \psi_2.$$

We combine the flows as follows:

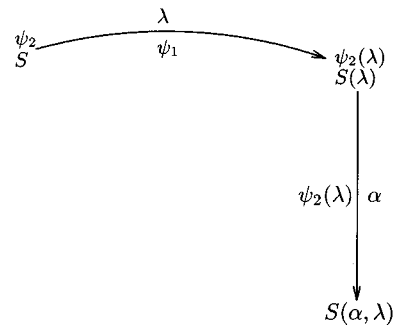

FIG. 1. Localization principle.

First, the action and the one-form ψ_2 are evolved by the flow of the one-form ψ_1, the localization parameter of the flow is λ. Second, the resulting action is evolved by the flow of the transported one-form $\psi_2(\lambda)$ and the localization parameter is α. The principle is described graphically in Fig. 1; by the arrows there is the localization parameter and the one-form that generates the flow. It is important to note that the first flow turns out to be such that $\omega(\lambda)$ is nondegenerate for all λ. It can be seen using the formulas in Appendices A and B. Thus, one can transport the one-form ψ_2 along it according to Eq. (15). In addition, the flow preserves the Lie derivative condition (44) in the form

$$L_{\chi_S(\lambda)}\psi_2(\lambda)=0.$$

Consequently, the flow along the transported one-form $\psi_2(\lambda)$ is again a BRST flow. The conditions of a triple are trivially satisfied because both the flows are familiar special cases: a diffeomorphism and a BRST flow. In the former the vector field $\chi_S(\lambda)$ is obtained from $\chi_S(0)$ by letting the diffeomorphism flow it according to Eq. (15). In the latter $\chi_S(\alpha,\lambda)=\chi_S(\lambda)$.

We get the total two-parameter localization deformation,

$$S(\alpha,\lambda)=S-S_B(\alpha,\lambda)-S_F(\alpha,\lambda), \tag{45}$$

$$S_B=(\lambda+\tfrac{1}{2}\lambda^2)K_{ij}\tilde{\chi}_x^i\tilde{\chi}_x^j+\alpha K^{ij}\,\partial_t^J I_i(\lambda)\partial_t^J I_j(\lambda), \tag{46}$$

$$S_F=\lambda d(K_{ij}\tilde{\chi}_x^i\omega^j)+\alpha K^{ij}d[\partial_t^J I_i(\lambda)dI_j(\lambda)], \tag{47}$$

where

$$\tilde{\chi}_x^i=\omega^i(\dot{x})-K^{ij}I_j+g_j^iJ^j, \tag{48}$$

$$I_i(\lambda)=I_i-\lambda K_{ij}\tilde{\chi}_x^j, \tag{49}$$

$$\partial_t^J=\delta_i^j\partial_t+J^kC_{ki}^j. \tag{50}$$

The partition function,

$$Z(\alpha,\lambda)=\int[dx\,dp\,d\psi\,d\bar{\psi}]e^{S(\alpha,\lambda)}, \tag{51}$$

is independent of the parameters α and λ and coincides with (43) when they vanish. The limit that the localization parameters are put to infinity localizes only half of the degrees of freedom, namely,

$$\tilde{\chi}_x^i=0, \tag{52}$$

leaving an infinite-dimensional integral. The full localization on the equations of motion (the geodesics) would be

$$\partial_t^J I_i = 0, \quad \tilde{\chi}_x^i = 0.$$

The fact that the localization deformation localizes only half of the degrees of freedom can be proved as follows: As λ and α are large there is exponential damping in the path integral due to the bosonic part (46), unless

$$\tilde{\chi}_x^i \sim \frac{1}{\lambda} f^i, \quad \partial_t^J I_i \sim \partial_t^J K_{ij} f^j + \frac{1}{\sqrt{\alpha}} g_i, \tag{53}$$

where f^i and g^i are finite. The fermionic part does not change the situation because the zeta function scaling of the fermionic determinant shows that it can only give a polynomial dependence. We see from the equations (53) that we get localization only on (52). However, it might be that taking first λ to infinity and then α would localize, or although there is no exponential damping, there might be a rational delta function $\delta(x) \sim \alpha/[1+(\alpha x)^2]$. Neither occurs, as can be proved in the case $J=0$ by integrating the fermions and expanding the bosonic action around $\tilde{\chi}_x^i = 0$ using the coordinate system x^μ, $\phi_S^\mu = v_i^\mu \tilde{\chi}_x^i$. One sees that the residual α and λ dependence cannot localize further. In addition, if there would be localization in the case that $J \neq 0$, there would also be localization in the limit that J vanishes.

VII. NON-ABELIAN DERIVATIVE EXPANSION LOCALIZATION

We localize the heat kernel on compact Lie groups using a derivative expansion localization. In the end of the section it is commented how it might be possible to obtain new localization formulas using the principle. Many of the equations in this section are formal; they involve distributions or possess singularities that cancel. We do not address the difficult mathematical problem of how to treat them rigorously.

We begin by studying the shifted heat kernel,

$$\langle x | e^{-\beta[(1/2)K^{ij}v_i v_j - J^i v_i]} | y \rangle, \tag{54}$$

on homogeneous manifolds $M = G/H$. If one puts the points $x,y \in M$ equal and integrates over x one gets the partition function (40).

Theorem 2: *Provided that the series* (57) *converges in the sense of distribution theory,*[19] *the following formal identity holds:*

$$\langle x | e^{-\beta[(1/2)K^{ij}v_i v_j - J^i v_i]} | y \rangle = e^{-(1/2)\beta K^{ij} u_i u_j} \langle x | e^{\varphi^i v_i} | y \rangle |_{\varphi = \beta J}, \tag{55}$$

where

$$u_i = \left(\frac{-\varphi^k C_k}{e^{-\varphi^k C_k} - 1} \right)_i^j \frac{\partial}{\partial \varphi^j}. \tag{56}$$

$[(C_i)_k^j = C_{ik}^j$ are the structure constants; see Appendix A.$]$

Proof: We expand the exponential function on the left-hand side as

$$\langle x | e^{-\beta[(1/2)K^{ij}v_i v_j - J^i v_i]} | y \rangle = \sum_{n=0}^{\infty} \langle x | e^{\varphi^i v_i} \left(-\frac{1}{2} \beta K^{ij} v_i v_j \right)^n | y \rangle \Big/ n!, \tag{57}$$

and by definition the right-hand side is

$$e^{-(1/2)\beta K^{ij}u_iu_j}\langle x|e^{\varphi^i v_i}|y\rangle=\sum_{n=0}^{\infty}\frac{1}{n!}\left(-\frac{1}{2}\beta K^{ij}u_iu_j\right)^n\langle x|e^{\varphi^i v_i}|y\rangle. \tag{58}$$

The terms in the series on the right-hand side are actually distributions: the $n=0$ term is

$$\langle x|e^{\varphi^i v_i}|y\rangle=\delta(x-e^{\varphi^i v_i}y), \tag{59}$$

and higher terms are obtained by differentiating it (it may occur that the terms diverge because of the singularities of u_i but then the series does not converge). Moreover, the terms on both sides are equal in each order (which suffices to prove the theorem because the series converge) because of the formal identity

$$u_i e^{\varphi^k v_k}=e^{\varphi^k v_k}v_i\,, \tag{60}$$

which can be proved as follows: Using Duhamel's formula[20]

$$\frac{\partial}{\partial\varphi^i}e^{\varphi^k v_k}=e^{\varphi^k v_k}\omega(\varphi)_i^j v_j\,, \tag{61}$$

where

$$\omega(\varphi)_i^j\,d\varphi^i=\left(\frac{e^{-\varphi^k C_k}-1}{-\varphi^k C_k}\right)_i^j d\varphi^i \tag{62}$$

can be understood as the left-invariant one-form of the Lie group G in exponential coordinates. It is, however, continued on the whole Lie algebra, including the points where the coordinate system is singular. The equation (60) follows, inverting it, which is possible, except at the singular points of the exponential coordinate system. However, formally the equation (60) holds also at these points, then u_i is singular, but the singularity cancels in the whole expression (60). □

One can turn the derivative expansion in Theorem 2 formally into an integral using the kernel

$$K_\beta(\varphi',\varphi)=\langle\varphi'|e^{-(1/2)\beta K^{ij}u_iu_j}|\varphi\rangle \tag{63}$$

as follows:

$$e^{-(1/2)\beta K^{ij}u_iu_j}\langle x|e^{\varphi^i v_i}|y\rangle=\int\sqrt{g'(\varphi')}d\varphi'\,K_\beta(\varphi,\varphi')\langle x|e^{\varphi'^i v_i}|y\rangle, \tag{64}$$

where g' is the degenerate metric,

$$g'(\varphi)_{kl}=K_{ij}\omega(\varphi)_k^i\omega(\varphi)_l^j\,, \tag{65}$$

on the Lie algebra.

In addition, one can formally calculate the kernel as follows:

$$K_\beta(\varphi',\varphi)=\langle 0|e^{-(1/2)\beta K^{ij}u_iu_j}|\xi(\varphi',\varphi)\rangle=K_\beta(0,\xi),$$

where ξ is defined as

$$e^{-\varphi'^k v_k}e^{\varphi^k v_k}=e^{\xi(\varphi',\varphi)^k v_k}. \tag{66}$$

An explicit expression for ξ can be obtained using the Campbell–Baker–Hausdorff theorem[21] in a neighborhood that φ' and φ are close to zero. The kernel $K_\beta(0,\xi)$ has been calculated in Ref. 9,

$$K_\beta(0,\varphi) = \frac{M}{(2\pi\beta)^{D/2}} \hat{A}(\varphi^i C_i) e^{-(1/2\beta)K_{ij}\varphi^i\varphi^j + (D/48)\beta}, \tag{67}$$

where $\hat{A}(X) = \Pi_{x_k>0}(x_k/2/\sin x_k/2)$ (ix_k are the eigenvalues of the antisymmetric real matrix X) and M is a normalization constant.[9,10]

We localize the Laplacian on Lie groups, assuming that the condition of Theorem 2 holds and one can transform the derivative operator into an integral as in Eq. (64), which is plausible because the singularities cancel. We calculate

$$k_\beta(\mathbf{1},g) = \langle \mathbf{1} | e^{-(1/2)\beta\Delta_0} | g \rangle, \tag{68}$$

$$= \int \sqrt{g'} d\varphi \, K_\beta(0,\varphi) \, \delta(\mathbf{1} - g e^{\varphi^i T_i}) = \sum_{\varphi \in L} K_\beta(\sigma), \tag{69}$$

where $\mathbf{1}$, $g \in G$ and we use the notation of matrix groups. T_i are the generators of the Lie algebra. One should notice that the delta function is normalized with respect to the volume on the Lie group whereas the integral is over φ, however, because one can associate φ as the exponential coordinate we get agreement. The lattice L is

$$L = \{\varphi | g = e^{\varphi^i T_i}\}. \tag{70}$$

[We have used the symmetry $K_\beta(\varphi) = K_\beta(-\varphi)$ to fix the sign convention for the lattice L.] It is in one to one correspondence with the geodesics starting at $\mathbf{1}$ and ending at g because one can associate with $\varphi \in L$ a geodesic $e^{t/\beta \varphi^i T_i}$, $t \in [0,\beta]$. The expression (69) coincides with the semi-classical approximation and it is studied in more detail in the next section.

Finally, we comment on how it might be possible to obtain new localization formulas using Theorem 2. Putting $y = x$ in Theorem 2 and integrating over x gives

$$\operatorname{tr} e^{-\beta[(1/2)K^{ij}v_i v_j - J^i v_i]} = e^{-(1/2)\beta K^{ij} u_i u_j} \operatorname{tr} e^{\varphi^i v_i}\big|_{\varphi = \beta J}. \tag{71}$$

There is, however, a minor subtlety: the condition of Theorem 2 does not necessarily hold. In the Lie group case the terms in the series (57) are not in the space of distributions because the delta functions restrict the vector field u_i to its singular points, hence, the series does not converge in the sense of distributions. The problem can be circumvented by replacing y instead by $e^{\varphi_0^i v_i} x$ and continuing analytically φ_0 to zero (or modifying the formulas by φ_0).

The linear partition function,

$$Z = \operatorname{tr} e^{\varphi^k v_k}, \tag{72}$$

is in the path integral form ($\varphi = \beta J$),

$$Z = \int [dx \, dp \, d\psi \, d\bar{\psi}] \exp \int_0^\beta I_i \omega^i(\dot{x}) + J^i I_i + d(I_i \omega^i). \tag{73}$$

It can be localized using the BRST localization principle because the Hamiltonian vector field of the Hamiltonian $J^i I_i$ generates an isometry of the phase space metric (B13). Consequently, we should get a localization for the partition function (71) that is a derivative expansion of ordinary localization formulas. (This principle appeared first in the Abelian form in Ref. 16. There is also a path integral derivation of it which lacks for the non-Abelian case; see also Ref. 22.) However, one has to be careful because the linear partition function (72) is a distribution, therefore the path integral (73) must diverge (for example, the divergent integral $\int_{-\infty}^{\infty} e^{ipx} \, dx$ can be associated with the delta function).

One can formally add a BRST exact term to the action in the path integral (73) with any of the following gauge fermions or their linear combinations, multiplied by the localization parameter λ:

$$\psi_1 = i_{J^i v_i^H} g, \tag{74}$$

$$\psi_2 = i_{\dot\phi} g, \tag{75}$$

$$\psi_3 = i_{\dot\phi - K^{ij} I_i v_j^H} g, \tag{76}$$

where g is the phase space metric (B13).

For example, the gauge fermion (75) should give a localization formula that is an integral over the phase space[15] of an equivariant characteristic class. (The Dirac genus should possibly be replaced by the Todd class[23] in Ref. 15.) The integral must be divergent, but if one would be able to differentiate the derivative expansion in the full formula (71) one would perhaps get a convergent integral over the phase space and a novel localization formula.

Moreover the gauge fermion $\psi_2 - \psi_1$ should give the loop space Duistermaat–Heckman formula[6] or its degenerate version, but it turns out to be singular because there are no critical points, except at fixed values of β. The localization by the gauge fermion (75) seems to be even more singular; it should localize to the nonexisting critical points of the Hamiltonian.[22] The gauge fermion (76) can possibly give a localization only if one first sums the derivative expansion and then takes the limit, which we have not been able to do.

VIII. CONFIGURATION SPACE LOCALIZATION

We localize the Laplacian on a Lie group using a configuration space path integral.

On the space ΩG of based loops (loops that start and end at the unit element **1** of the Lie group), there is a natural symplectic form that is right invariant,

$$\omega = \tfrac{1}{2} K_{ij} \omega_R^i \wedge \partial_t \omega_R^j, \quad d\omega = 0. \tag{77}$$

(The left- and right-invariant forms have changed places compared to Ref. 24.) The notation is that v_i, v_i^R are the left- and right-invariant vector fields and the dual one-forms are

$$\omega^i = K^{ij} g(v_j), \quad \omega_R^i = K^{ij} g(v_j^R),$$

where g is the bi-invariant metric and the vector fields coincide at **1**, $v_i|_{\mathbf{1}} = v_i^R|_{\mathbf{1}}$.

The configuration space path integral is

$$Z = \int [dx^\mu \, d\psi^\mu]_{\text{vbc}} \exp \int_0^\beta \frac{1}{2} K_{ij} [\omega^i(\dot x) + J^i][\omega^j(\dot x) + J^j] + \frac{1}{2} K_{ij} \omega_R^i \wedge \partial_t \omega_R^j, \tag{78}$$

where vbc means vanishing boundary conditions for both bosons and fermions, $x^\mu(\beta) = x^\mu(0) = 0$, $\psi^\mu(\beta) = \psi^\mu(0) = 0$, and the coordinates x^μ are chosen so that $x^\mu = 0$ corresponds to the unit element **1** of the Lie group.

Integration of the fermionic part of the integral gives a Pfaffian,

$$\sqrt{\det(\omega_{R\mu}^i K_{ij} \, \partial_t \omega_{R\nu}^j)} = \sqrt{\det g_{\mu\nu}} \sqrt{\det(\delta_j^i \partial_t)_{\text{vbc}}}, \tag{79}$$

where we have used the product rule for the determinant and the determinants obey vanishing boundary conditions. The equation (79) can be justified using the change of variables,

$$\psi^\mu = v_i^\mu M_{ij}^{-1} \theta^j, \quad K_{ij} = (M^T M)_{ij}, \tag{80}$$

writing the measure as $[d\theta^j]_{\text{vbc}} = [d\theta^j]_{\text{pbc}} \delta[\theta^j(0)]$ and expanding θ^j in Fourier modes (pbc means periodic boundary conditions). The procedure also evaluates the determinant,

$$\det(\delta^i_j\partial_t)_{\rm vbc}=\det{}'(\beta\partial_t)^D=1,$$

in terms of the standard determinant

$$\det{}'(\partial_t+\mu)_{\rm pbc}=\prod_{n\neq 0}\left(\mu+\frac{2\pi n i}{\beta}\right)=\beta\frac{\sinh\beta\mu/2}{\beta\mu/2}. \tag{81}$$

(We have chosen a regularization that preserves the ''charge conjugation'' symmetry[25,23] $\mu\leftrightarrow-\mu$.)

The result, after the integration of fermions, is

$$\int [\sqrt{g}dx^\mu]_{\rm vbc}\exp\int\frac{1}{2}K_{ij}[\omega^i(\dot x)+J^i][\omega^j(\dot x)+J^j],$$

which, using the change of variables,

$$x^\mu\to e^{-tJ^iv_i}x^\mu, \tag{82}$$

gives the equivalent path integral

$$\int [\sqrt{g}dx]_{\rm bc}\exp\frac{1}{2}\int_0^\beta K_{ij}\omega^i(\dot x)\omega^j(\dot x),$$

with the boundary conditions (bc)

$$x^\mu(\beta)=(e^{\beta J^iv_i}x^\mu)|_{x=0}\sim e^{\beta J^iT_i}=g,\quad x^\mu(0)=0\sim\mathbf{1}.$$

Thus, the configuration space path integral (78) coincides with the heat kernel $k_\beta(\mathbf{1},g)$.

We proceed to localize the integral (78) using the Hamiltonian BRST localization principle. The Hamiltonian vector field for the action,[24]

$$S=\int_0^\beta\frac{1}{2}K_{ij}[\omega^i(\dot x)+J^i][\omega^j(\dot x)+J^j], \tag{83}$$

is

$$\chi_S=\chi+J^iu_i, \tag{84}$$

where

$$\chi=\dot x-\omega^i(\dot x)|_{t=0}v_i^R, \tag{85}$$

$$u_i=v_i-v_i^R, \tag{86}$$

and one can use the Hamiltonian BRST symmetry with the equivariant derivative d_{χ_S}.

We construct the gauge fermion ψ starting with an invariant tensor,

$$g'=\int_0^\beta K_{ij}\,d\omega^i(\dot x)\otimes d\omega^j(\dot x),\quad L_{\chi_S}g'=0, \tag{87}$$

and contracting it with the loop space Hamiltonian vector field,

$$\psi=i_{\chi_S}g'. \tag{88}$$

This is analogous to the localization in Ref. 6. Adding the gauge fermion gives the action

$$S(\lambda)=S+\omega+\lambda d_{\chi_S}(i_{\chi_S}g'), \tag{89}$$

$$=S+\omega+\lambda K_{ij}\,\partial_t^J\omega^i(\dot{x})\partial_t^J\omega^j(\dot{x})+\lambda K_{ij}d[\partial_t^J\omega^i(\dot{x})d\omega^j(\dot{x})], \tag{90}$$

where $\partial_t^J=\delta_j^i\partial_t-J^kC_{kj}^i$.

The path integral,

$$Z=\int [dx^\mu\, d\psi^\mu]_{\rm vbc}e^{S(\lambda)}, \tag{91}$$

coincides with (78) when λ vanishes and localizes in the limit $\lambda\to-\infty$ on $\partial_t^J\omega^i(\dot{x})=0$, which are the classical geodesics $\partial_t\omega^i(\dot{x})=0$ before the change of variables in Eq. (82).

One can write the geodesics starting at **1** and ending at $g=e^{\beta J^iT_i}$ (in matrix group notation) as[9,10]

$$\gamma_\nu(t)=e^{(t/\beta)(\varphi^i+2\pi\nu^i)T_i},\quad \nu^i\in L,$$

where $\varphi^i=\beta J^i$ and

$$L=\{\nu^i|[\nu^iT_i,\varphi^jT_j]=0,\quad e^{2\pi\nu^iT_i}=\mathbf{1}\},$$

and we assume that φ^iT_i is in the generic position that its annihilator is a conjugated Cartan subalgebra. One can also write the lattice L in terms of the roots of the algebra; see Refs. 9 and 10. The solutions of the equation $\partial_t^J\omega^i(\dot{x})=0$ are obtained from these by inverting the change of variables in Eq. (82) [which is translated into matrix group notation as $x^\mu(t)\sim g(t)\to e^{-tJ^iv_i}x^\mu\sim g(t)e^{-tJ^iT_i}$], giving

$$\tilde{\gamma}(t)=e^{(t/\beta)2\pi\nu^iT_i}. \tag{92}$$

Then we calculate the limit $\lambda\to-\infty$ in the path integral (91).

First we make the change of variables,

$$[dx^\mu\, d\psi^\mu]_{\rm vbc}\to[dX^i\, d\eta^i],$$

where $X^i=\omega^i(\dot{x})$ and $\eta^i=dX^i=[\delta\omega^i(\dot{x})/\delta x^\mu]\psi^\mu$. It changes the vanishing boundary conditions into something more complicated. Fortunately it is enough to study what occurs near the critical points because of the localization and the fact that the first variation of the bosonic part in $S(\lambda)$ vanishes. In first order the bosonic coordinates are related by the identity

$$x^\mu=x^\mu_{\rm cr}+\delta x^\mu,\quad \delta\omega^i(\dot{x})=\left.\frac{\delta\omega^i(\dot{x})}{\delta x^\mu}\right|_{x_{\rm cr}}\delta x^\mu=(\tilde{D}_t^{M-C})^i_j(\omega^j_\mu\,\delta x^\mu),$$

where we have the Maurer–Cartan differential operator (A20). The critical points are given in equation (92) so that the Maurer–Cartan operator is

$$\tilde{D}_t^{M-C}=e^{-(t/\beta)2\pi\nu^iC_i}\,\partial_te^{(t/\beta)2\pi\nu^iC_i}.$$

We get that the vanishing boundary condition results in the condition

$$\int_0^\beta e^{(t/\beta)2\pi\nu^iC_i}\,\delta X^i=0, \tag{93}$$

for the coordinate $X^i = X^i_{\rm cr} + \delta X^i$ near the critical point. The fermionic coordinates η^i also obey the same condition (93).

Integrating the fermions and putting the localization parameter to infinity yields

$$Z = \sum_\nu \frac{1}{\mathrm{Pf}(\partial_t^J)} e^{-(1/2\beta)K_{ij}(\varphi^i + 2\pi\nu^i)(\varphi^j + 2\pi\nu_j)},$$

where the Pfaffian obeys the boundary condition (93). It is calculated conjugating the Hilbert space by the operator $e^{-(t/\beta)2\pi\nu^i v_i}$, which turns the boundary condition (93) into $\int_0^\beta \delta X^i = 0$ and the operator ∂_t^J into $\partial_t^{J+2\pi\nu}$. Furthermore, the diagonalization of $(\varphi^i + 2\pi\nu^i)C_i$ gives the eigenvalues μ_j,

$$\mu_j = \begin{cases} 0; & j = 1,...,r, \\ i\langle \varphi + 2\pi\nu, \alpha_j \rangle; & j = r+1,...,D, \end{cases} \tag{94}$$

where r is the rank of the group, α_i are the roots (they are in the annihilator of $\varphi^i T_i$, which is isomorphic to the Cartan subalgebra), and $\langle , \rangle$ is the contraction by the Killing tensor K_{ij}. The Pfaffian reduces into a product of usual determinants (81) and the final result is

$$k_\beta(\mathbf{1}, e^{\varphi^i T_i}) = \frac{M}{(2\pi\beta)^{D/2}} \sum_\nu \prod_{\alpha>0} \frac{\langle \varphi + 2\pi\nu, \alpha \rangle /2}{\sin\langle \varphi + 2\pi\nu, \alpha \rangle /2} e^{-(1/2\beta)\langle \varphi + 2\pi\nu, \varphi + 2\pi\nu \rangle}, \tag{95}$$

which, however, has to be corrected with DeWitt's term, which gives the extra factor exp $[-\beta(D/48)]$. It is consistent with DeWitt's original[26] proportionality constant $\frac{1}{12}$ in front of the scalar curvature $D/4$. The result coincides with the expression (69) and has been calculated using the different methods in Refs. 9 and 10. This is, however, disturbing because the natural value[18] of DeWitt's constant should be $\frac{1}{8}$. We hope that further research solves this puzzle. We speculate that both values may be correct, they just correspond to different path integral measures, and that this trivial looking problem may reveal deep insights in how to make the path integral rigorous and what the quantization is all about.

IX. CONCLUSIONS, SPECULATIONS, AND FUTURE PROSPECTS

We have developed two new localization methods: the new localization principle and the non-Abelian derivative expansion localization. We emphasize that both the principles are universal and probably have many other applications that are not yet known. For example, the non-Abelian derivative expansion localization may apply to integrable models: Many integrable models can be embedded in Poisson–Lie algebras.[27] The integrable hierarchy is the sequence of Casimirs with respect to the r-deformed Poisson bracket. The linear generators generate the coadjoint action and are therefore isometries of the Killing metric on the algebra. However, the algebra in many cases is noncompact and the metric that is pulled back on the phase space may be indefinite. There may also be anomalous problems associated with the quantization and the derivative expansion in Theorem 2 may be singular, so that before concrete examples are worked out we cannot say if the principle eventually works or not.

Another direction of research is to generalize the configuration space localization on Lie groups in Sec. VIII to homogeneous or locally homogeneous manifolds and try to understand Selberg's trace formulas. However, it may be that the existence of Selberg's trace formulas is a separate phenomenon; see the remarks below. Yet another direction of research would be to study if there would be an analog of the configuration space localization for two-dimensional sigma models, for example, the Laplacian on loop groups might be such. In addition, it and the derivative expansion localization in Sec. VII may have supersymmetric versions; the starting point would be the grand canonical partition function that was introduced in Ref. 18.

Finally, we collect few remarks on Selberg's trace formula on constant negative curvature Riemann surfaces, which might be helpful in a possible path integral localization.

Remark 6: Selberg's trace formula is the exact Gutzwiller's trace formula,[28] which suggests that the path integral should perhaps be enlarged by one degree of freedom,

$$Z(\beta)=\mathrm{tr}\, e^{-\beta\Delta_0}\rightarrow\int_0^\infty e^{-\beta E}Z(\beta). \tag{96}$$

Remark 7: The Riemannian surfaces have a natural symplectic form, the volume form $\sqrt{g}\epsilon_{\mu\nu}\,dx^\mu\wedge dx^\nu$, which makes it possible to write the path integral as follows:

$$Z(\beta)=\int [\sqrt{g}dx]\exp\int g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu=\int [dx^\mu\, d\psi^\mu]\exp\int g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu+\sqrt{g}\epsilon_{\mu\nu}\psi^\mu\psi^\nu. \tag{97}$$

Perhaps it should be taken to be the starting point for localization.

Remark 8: We have assumed in *Remark 7* that the path integral takes implicit care of the fact that the Riemann surfaces are not simply connected. It can be made explicit (it may even be that it is necessary) by integrating on the Poincaré upper half-plane and splitting the integral into a sum of path integrals with different boundary conditions. It is possible to transform the boundary conditions into periodic using an analogous change of variables that was used in Sec. VIII in Eq. (82), which allows one to use localization deformations. Furthermore, in order to replace the sum as an integral, one can perhaps make an analogous trick,[29] as in the case of the material particle on U(1):

$$\mathrm{tr}\, e^{\beta\partial_x^2}=\sum_{n=-\infty}^{\infty}\int_0^L dx(0)\int [dx]_{x(\beta)=x(0)+nL}e^{\int\dot{x}^2}=\int [dx]_{\mathrm{pbc}}\exp\int\left[\dot{x}+\left(\frac{x(0)}{L}\right)L\right]^2,$$

where $x(t)\in\mathbf{R}^1$ and $[x]$ is the greatest integer smaller than x. U(1) is associated with $\mathbf{R}^1 \bmod L$.

ACKNOWLEDGMENTS

We thank A. Niemi for pointing out the Ref. 8 and the connection with Selberg's trace formula. We thank him also for the problem of finding a non-Abelian derivative expansion localization in connection with integrable models. We thank O. Tirkkonen for pointing out the possibility of a configuration space localization. We thank K. Palo for giving his beautiful calculations in the degenerate Duistermaat–Heckman localization. In addition, we thank A. Alekseev, Cheeqer, B. S. DeWitt, L. D. Faddeev, J. Kalkkinen, A. Losev, M. Miettinen, T. Nassar, A. Niemi, K. Palo, P. Pasanen, V. Rubtsov, R. J. Szabo, and O. Tirkkonen for discussions.

APPENDIX A: FORMULAS ON HOMOGENEOUS SPACES

We introduce formulas on homogeneous spaces that use a degenerate basis for vector fields. In addition, we define a Maurer–Cartan connection on the tangent bundle.

A homogeneous space is, in this article, the quotient space $M=G/H$ (G and H are Lie groups) with the metric that is inherited from the unique bi-invariant metric on the Lie group G. Technically, one inverts the metric on G and pushes it on M using the canonical projection.[30] We require, in addition, that G is compact, although the formulas are valid also in the noncompact case, with the exception that the metric tensors are not necessarily positive definite.

As the quotient is taken by the right action we get the standard left action of G on M that is generated by the vector fields,

$$v_i,\quad i=1,\ldots,N, \tag{A1}$$

on the D-dimensional manifold M. They are isometries of the metric and satisfy the Lie algebra of G,

$$[v_i,v_j]=C_{ij}^k v_k. \tag{A2}$$

We define the Killing tensor,

$$K_{ij} = -\mathrm{tr}\, C_i C_j = -C_{il}^k C_{jk}^l, \tag{A3}$$

which is positive definite because G is compact. The inverse of K is K^{ij} and it is used to raise indices $i,j,k\cdots \in \{1,...,N\}$ as Greek indices $\alpha,\beta\cdots \in \{1,...,D\}$ are raised and lowered by the metric $g_{\mu\nu}$. For example, the tensor

$$C_{ijk} = K_{im} C_{jk}^m$$

is antisymmetric in all the three indices when the upper index is lowered.

We define the one-forms,

$$\omega^i = K^{ij} g(v_j), \tag{A4}$$

that are dual to the vector fields (A1). If x^μ are coordinates on M one can write the vector fields and their dual one-forms in the coordinate basis as $v_i = v_i^\mu \partial_\mu$ and $\omega^i = \omega_\mu^i dx^\mu$ that satisfy the relation

$$\omega_\mu^i v_i^\nu = \delta_\mu^\nu. \tag{A5}$$

The metric can be written as

$$g = K_{ij} \omega^i \otimes \omega^j, \quad g^{-1} = K^{ij} v_i \otimes v_j. \tag{A6}$$

We define a tensor g_j^i,

$$g_j^i = \omega^i \cdot v_j, \tag{A7}$$

which has the properties

$$g_{ij} = g_{ji}, \quad g_j^i g_k^j = g_k^i, \quad g_j^i \omega^j = \omega^i, \quad g_j^i v_i = v_j. \tag{A8}$$

Using it, we can associate the Lie algebra of the local isotropy group

$$H_p = \{g \in G | gp = p\}, \quad p \in M,$$

with

$$h_p = \{X^i v_i | X^i \in \mathbf{R}, g_j^i(p) X^j = 0\},$$

and the orthogonal complement of it, which is isomorphic to the tangent space at p, with

$$m_p = \{X^i v_i | X^i \in \mathbf{R}, [\delta_j^i - g_j^i(p)] X^j = 0\}.$$

The key observation is that $X^i v_i$ vanishes at p if and only if $g_j^i(p) X^j = 0$. Using the association the standard commutation relations on homogeneous spaces,[30]

$$[h_p, m_p] \subset m_p, \quad [h_p, h_p] \subset h_p,$$

give the identity

$$(\delta_n^m - g_n^m) C_{nk}^i g_j^k = (\delta_n^m - g_n^m) g_k^i C_{mj}^k. \tag{A9}$$

The dual of the algebra (A2) is

$$d\omega^i = -\tfrac{1}{2}(2\delta_m^i - g_m^i) C_{kl}^m \omega^k \wedge \omega^l, \tag{A10}$$

and we obtain the formulas for the connection,

$$\nabla_{v_i} v_j = \Gamma^k_{ij} v_k , \tag{A11}$$

$$\Gamma^k_{ij} = \tfrac{1}{2} K^{kl} (C^n_{ij} g_{ln} - C^n_{li} g_{jn} + C^n_{jl} g_{il}), \tag{A12}$$

and the curvature

$$R(v_i , v_j) v_k = R^l_{kij} v_l , \tag{A13}$$

$$R^l_{kij} = C^m_{ij} \Gamma^l_{mk} + C^m_{ik} \Gamma^l_{jm} - C^m_{jk} \Gamma^l_{im} + \Gamma^m_{jk} \Gamma^l_{mi} - \Gamma^m_{ik} \Gamma^l_{mj} . \tag{A14}$$

The formula for the connection can be checked using the formula[20]

$$2(\nabla_X Y \cdot Z) = X(Y \cdot Z) - Z(X \cdot Y) + Y(Z \cdot X) + [X,Y] \cdot Z - [Y,Z] \cdot X - [Z,X] \cdot Y,$$

where $\cdot$ means contraction with respect to the metric g. The Laplacian on zero-forms is actually the Casimir,

$$\Delta_0 = K^{ij} v_i v_j . \tag{A15}$$

One can also check that these manifolds are of constant scalar curvature by calculating

$$R = R^i_{jkl} g^j_i g^{kl},$$

and checking that

$$L_{v_i} R = 0$$

(Appendix D).

We make few comments on the connection and the curvature. First, we have on TM the usual Levi-Civita connection (A12), which is denoted as

$$D = D_{\mathrm{L-C}} = \delta^\alpha_\beta d + dx^\gamma \Gamma^\alpha_{\gamma\beta} . \tag{A16}$$

Since

$$\Gamma^\gamma_{\alpha\beta} = (\partial_\alpha \omega^j_\beta) v^\gamma_j + \omega^i_\alpha \omega^j_\beta \Gamma^k_{ij} v^\gamma_k = (\partial_\alpha \omega^j_\beta) v^\gamma_j + \tfrac{1}{2} \omega^i_\alpha \omega^j_\beta C^k_{ij} v^\gamma_k ,$$

we can write

$$D^\mu_\nu = v^\mu_i (\delta^i_j d + \tfrac{1}{2} \omega^k C^i_{kj}) \omega^j_\nu , \tag{A17}$$

$$= v^\mu_i \tilde{D}^i_j \omega^j_\nu , \tag{A18}$$

where $\tilde{D}$ is a differential operator. There is also another interesting connection on TM, which we call the Maurer–Cartan connection:

$$(D_{\mathrm{M-C}})^\mu_\nu = v^\mu_i (\tilde{D}_{\mathrm{M-C}})^i_j \omega^j_\nu , \tag{A19}$$

where

$$(\tilde{D}_{\mathrm{M-C}})^i_j = \delta^i_j d + \omega^k C^i_{kj} . \tag{A20}$$

The differential operator associated with it has the property

$$(\tilde{D}_{\mathrm{M-C}})^i_j g^j_k = g^i_j (\tilde{D}_{\mathrm{M-C}})^j_k . \quad \text{(A21)}$$

The fact that it really is a connection is checked by calculating the difference of it and the Levi-Civita connection that yields the tensor,

$$C_{\kappa\mu\nu} = \omega^i_\mu \omega^j_\nu C^k_{ij} K_{kl} \omega^l_\kappa . \quad \text{(A22)}$$

Then one can check easily that the axioms of a connection are satisfied, except the torsion-free axiom.[31] The definition of the tensor (A22) is invariant under rotations that take v_i into their linear combinations but may depend on the particular way that M is quotiented, $M = G/H$. In addition, it is antisymmetric in all the three indices. Consequently, on any two-dimensional homogeneous manifold D and $D_{\mathrm{M-C}}$ coincide.

Then we calculate the curvature of the Maurer–Cartan tensor, but first the Levi-Civita curvature:

$$D^2 = \tfrac{1}{2} R^\alpha_{\beta\gamma\delta}\, dx^\gamma \wedge dx^\delta ,$$

where the curvature tensor is easily expressed in terms of the tensor (A14) as $R^\alpha_{\beta\gamma\delta} = v_i^\alpha R^i_{jkl} \omega^j_\beta \omega^k_\gamma \omega^l_\delta$. One can calculate the Maurer–Cartan curvature,

$$F^\alpha_{\beta\gamma\delta} = v_i^\alpha F^i_{jkl} \omega^j_\beta \omega^k_\gamma \omega^l_\delta ; \quad \text{(A23)}$$

similarly,

$$D^2_{\mathrm{M-C}} = v_i^\mu \tilde{D}_{\mathrm{M-C}}\, g^i_j \tilde{D}_{\mathrm{M-C}} \omega^j_\nu = v_i^\mu \tilde{D}^2_{\mathrm{M-C}} \omega^j_\nu = \tfrac{1}{2} F^i_{jkl} v_i^\mu \omega^j_\nu \omega^k_\kappa \omega^l_\lambda\, dx^\kappa \wedge dx^\lambda ,$$

where

$$F^i_{jkl} = C^i_{jn} (\delta^n_m - g^n_m) C^m_{kl} . \quad \text{(A24)}$$

In the Lie group case $g^n_m = \delta^n_m$ and $F = 0$, which is very natural because on a Lie group,

$$D_{\mathrm{M-C}} = g^{-1}\, dg = d + \Omega ,$$

where $\Omega = g^{-1}\, dg$ is the Maurer–Cartan form (g is now exceptionally a group element $g \in G$, $g = e^{\theta^i T_i}$, where θ are the exponential coordinates on the Lie group and T_i are the generators of the Lie algebra), and the connection is zero-curvature by the Maurer–Cartan equation,[21]

$$D^2_{\mathrm{M-C}} = d\Omega + \Omega \wedge \Omega = g^{-1}\, d^2 g = 0.$$

APPENDIX B: FORMULAS ON THE COTANGENT BUNDLE

We introduce some formulas and a one-parameter family of invariant metrics on the cotangent bundle of homogeneous manifolds.

The phase space of the geodesic motion on a manifold is the cotangent bundle. If x^μ, p_μ are the standard coordinates on it,

$$p_\mu\, dx^\mu|_p \in T^*M, \quad p \in M,$$

the standard symplectic potential is

$$\theta = p_\mu\, dx^\mu ,$$

and the Hamiltonian of the geodesic motion is

$$H = \tfrac{1}{2}\, g^{\mu\nu} p_\mu p_\nu .$$

The G action on M has a Hamiltonian lift;[7] the Hamiltonians that generate it are

$$I_i = v_i^{\mu} p_{\mu}, \tag{B1}$$

and they satisfy the Poisson brackets

$$\{I_i, I_j\} = C_{ij}^k I_k. \tag{B2}$$

The associated Hamiltonian vector fields are

$$v_i^H = \{I_i, \ \} = v_i - \frac{\partial I_i}{\partial x^{\mu}} \frac{\partial}{\partial p_{\mu}}. \tag{B3}$$

The symplectic form[7] can be written as

$$\omega = d\theta = d(I_i \omega^i) = dI_i \wedge \omega^i - \tfrac{1}{2} I_i C_{jk}^i \omega^j \wedge \omega^k, \tag{B4}$$

where ω^i is actually the pull-back of the familiar form on M onto the cotangent bundle by the canonical projection.

The dual vector fields to the forms ω^i, dI_i are

$$u_i = v_i^H - C_{ij}^k I_k w^j, \quad w^i = \omega_{\mu}^i \frac{\partial}{\partial p_{\mu}},$$

and the word ''dual'' means that the identity tensor in the fiber of the tangent bundle can be written as

$$\mathbf{1}_{2D} = u_i \otimes \omega^i + w^i \otimes dI_i = v_i^H \otimes \omega^i + w^i \otimes (dI_i + C_{ij}^k I_k \omega^j). \tag{B5}$$

However, we prefer the vector fields v_i^H, w^i, whose contractions are

$$v_i^H \cdot dI_j = C_{ij}^k I_k, \tag{B6}$$

$$v_i^H \cdot \omega^j = g_i^j, \tag{B7}$$

$$w^i \cdot dI_j = g_j^i, \tag{B8}$$

$$w^i \cdot \omega^j = 0, \tag{B9}$$

and commutators,

$$[w^i, w^j] = 0, \quad [w^i, v_i^H] = C_{jk}^i w^k, \quad [v_i^H, v_j^H] = C_{jk}^i v_i^H.$$

The inverse of the symplectic form can be written using these vector fields as

$$\omega^{-1} = v_i^H \wedge w^i + \tfrac{1}{2} I_i C_{jk}^i w^j \wedge w^k. \tag{B10}$$

Finally, we describe a natural invariant metric on the cotangent bundle. Observing that

$$L_{v_i^H} \omega_j = C_{ij}^k \omega_k, \quad L_{v_i^H} dI_j = C_{ij}^k dI_k,$$

the following tensors are invariant under the action of the group G because of the Casimir structure:

$$g_1 = K_{ij} \omega^i \otimes \omega^j, \tag{B11}$$

$$g_2 = K^{ij}\, dI_i \otimes dI_j\,. \tag{B12}$$

From these two one can combine a metric,

$$g = \alpha g_1 + \beta g_2\,, \tag{B13}$$

provided that α and β are positive.

Proof: The combination is positive semidefinite because K_{ij} is positive definite. One needs to check that it is nondegenerate, which can be done by calculating the determinant of it,

$$\det g = \beta \det g_1^{\mu\nu} \det\,[\alpha g^1_{\mu\nu} + \beta (K^{ij} - g^{ij})\partial_\mu I_i\, \partial_\nu I_j],$$

using the block matrix formula, and proving that it is nonzero, which is not difficult because the last term in the latter determinant is positive semidefinite. □

The invariance property of the metric can be written explicitly as

$$L_{v_i^H} g = 0;$$

see also Ref. 7 for a different metric.

APPENDIX C: BOUND FOR THE GEODESIC ACTION

We derive a bound for the geodesic action,

$$S = \int_0^\beta g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu, \tag{C1}$$

starting from the expression[32]

$$S(t) = \int g_{\mu\nu}(\dot{x}^\mu + t g^{\mu\kappa}\theta_\kappa)(\dot{x}^\nu + t g^{\nu\lambda}\theta_\lambda) \geqslant 0, \tag{C2}$$

where θ is an arbitrary one-form. The inequality holds for all t, especially at the minimum,

$$t = -\frac{\int \theta_\mu \dot{x}^\mu}{\int g^{\mu\nu}\theta_\mu \theta_\nu}, \tag{C3}$$

where it gives the bound

$$\int g_{\mu\nu}\, \dot{x}^\mu\, \dot{x}^\nu \geqslant \frac{(\int \theta_\mu \dot{x}^\mu)^2}{\int g^{\mu\nu}\theta_\mu \theta_\nu}. \tag{C4}$$

For BRST quantization of the winding number action that appears in the bound, see Ref. 33.

APPENDIX D: CALCULATION OF THE SCALAR CURVATURE AND A FORMULA

We calculate the scalar curvature on homogeneous manifolds, but first we calculate

$$g^a_n g^k_b g^i_c g^j_d R^n_{kij} \tag{D1}$$

using the missing identity,

$$\Gamma^k_{ij} g^l_k g^i_a = \Gamma^k_{aj} g^l_k\,; \tag{D2}$$

the torsion,

$$\Gamma^k_{ij}-\Gamma^k_{ji}=g^k_l C^l_{ij}; \tag{D3}$$

and the following property of the connection coefficient:

$$g^i_a g^j_b \Gamma^k_{ij} g^c_k=\tfrac{1}{2} g^i_a g^j_b C^k_{ij} g^c_k. \tag{D4}$$

The point is to use (D2) to eliminate the Γ's using (D4) when one substitutes the expression (A14) in (D1). Another way is just hard work, but it requires perhaps the introduction of Feynman rules for the terms. The result is

$$=g^a_n g^k_b g^i_c g^j_d[-C^m_{ij}C^n_{mk}+\tfrac{1}{2}C^m_{ij}g^p_m C^n_{pk} \tag{D5}$$

$$+\tfrac{1}{4}(C^m_{ik}g^p_m C^n_{pj}-C^m_{jk}g^p_m C^n_{pi})], \tag{D6}$$

which is also useful in calculating the equation (D14). The scalar curvature is

$$R=R^i_{jkl}g^i_j g^{kl}, \tag{D7}$$

$$=K^{ij}C^m_{ik}g^k_l C^l_{jm}-\tfrac{3}{4}K^{ij}C^m_{ik}g^k_l C^l_{jn}g^n_m, \tag{D8}$$

$$=K^{ij}\,\mathrm{tr}(C_i g C_j)-\tfrac{3}{4}K^{ij}\,\mathrm{tr}(C_i g C_j g). \tag{D9}$$

Using the equation

$$L_{v_a}g^k_l=[g,C_a]^k_l=g^k_p C^p_{al}-C^k_{ap}g^p_l, \tag{D10}$$

it is easy to see that $L_{v_a}R=0$. If one puts $g^i_j=\delta^i_j$ one recovers the Lie group formula

$$=\frac{1}{4}K^{ij}\,\mathrm{tr}\,C_i C_j=\frac{D}{4}, \tag{D11}$$

where D is the dimension of the Lie group.

Finally we mention a formula. We define first

$$D_-=D_t-\tfrac{1}{2}\omega^i(\dot{x})C^l_{ik}v^\lambda_l\omega^k_\kappa=v^\lambda_i\,\partial_t\omega^i_\kappa, \tag{D12}$$

$$D_+=D_t+\tfrac{1}{2}\omega^i(\dot{x})C^l_{ik}v^\lambda_l\omega^k_\kappa=D^{\mathrm{M-C}}_t, \tag{D13}$$

where the operators are considered on classical trajectories of the geodesic motion which means that $\omega^i(\dot{x})$ is constant. The formula reads as

$$D_+D_-=D_-D_+=D^2_t+R^\alpha_{\beta\gamma\delta}\,\dot{x}^\beta\dot{x}^\gamma+F^\alpha_{\beta\gamma\delta}\,\dot{x}^\beta\dot{x}^\gamma, \tag{D14}$$

where F is the curvature of the Maurer–Cartan connection. In the Lie group case it reduces to the factorization of the geodesic deviation operator,

$$g^{\mu\kappa}\frac{\delta S}{\delta x^\kappa\,\delta x^\nu}\bigg|_{\delta S=0}=D_+D_-=D_-D_+=D^2_t+R^\alpha_{\beta\gamma\delta}\,\dot{x}^\beta\dot{x}^\gamma,$$

where $S=\tfrac{1}{2}\int g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$.

[1] K. Palo, Ph.D. thesis, Uppsala University, 1993.
[2] O. Tirkkonen, Ph.D. thesis, University of Helsinki, 1994.
[3] M. Blau and G. Thompson, J. Math. Phys. **36**, 2192 (1995).
[4] D. Birmingham, M. Blau, M. Rakowski, and G. Thompson, Phys. Rep. **209**, 129 (1991).
[5] T. Kärki and A. Niemi, Ahrenshoop Symposium, 1993, p. 175.

[6] M. Blau, E. Keski-Vakkuri, and A. Niemi, Phys. Lett. B **246**, 92 (1990).
[7] E. Witten, J. Geom. Phys. **9**, 303 (1992).
[8] L. D. Paniak, G. W. Semenoff, and R. J. Szabo, Phys. Lett. B **372**, 236 (1996).
[9] M. Marinov and M. Terentyev, Fortschr. Phys. **27**, 511 (1979).
[10] R. Picken, J. Phys. A **22**, 2285 (1989).
[11] M. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer-Verlag, Berlin, 1990).
[12] D. Hejhal, *The Selberg Trace Formula for PSL(2,R), Volume I* (Springer-Verlag, Berlin, 1976–1983).
[13] P. Buser, *Geometry and Spectra of Compact Riemann Surfaces* (Birkhäuser, Basel, 1992).
[14] C. Grosche, *Path Integrals, Hyperbolic Spaces, and Selberg's Trace Formulae* (World Scientific, Singapore, 1995).
[15] A. Niemi and O. Tirkkonen, Phys. Lett. B **293**, 339 (1992).
[16] A. Niemi and K. Palo, Mod. Phys. Lett. A **8**, 2311 (1993).
[17] A. Niemi and O. Tirkkonen, Ann. Phys. **235**, 318 (1994).
[18] T. Kärki and A. Niemi, Phys. Rev. D **56**, 2080 (1997).
[19] W. Rudin, *Functional Analysis* (McGraw-Hill, New York, 1989).
[20] N. Berline, E. Getzler, and M. Vergne, *Heat Kernels and Dirac Operators* (Springer-Verlag, Berlin, 1991).
[21] D. Sattinger and O. Weaver, *Lie Groups and Algebras with Applications to Physics, Geometry and Mechanics* (Springer-Verlag, Berlin, 1986).
[22] A. Niemi and O. Tirkkonen, Ann. Phys. **235**, 318 (1994).
[23] M. Miettinen, J. Math. Phys. **37**, 3141 (1996).
[24] A. Pressley and G. Segal, *Loop Groups* (Oxford University Press, Oxford, 1988).
[25] S. Elitzur, E. Rabinovici, Y. Frishman, and A. Schwimmer, Nucl. Phys. B **273**, 93 (1986).
[26] B. S. DeWitt, Rev. Mod. Phys. **29**, 377 (1957).
[27] M. Semenov-Tyan-Shanskii, Funct. Anal. Appl. **17**, 259 (1984).
[28] See Ref. 11, p. 290.
[29] W. Dittrich and M. Reuter, *Classical and Quantum Dynamics* (Springer-Verlag, Berlin, 1992).
[30] R. Coquereaux and A. Jadczyk, *Riemannian Geometry Fiber Bundles Kaluza–Klein Theories and All That* (World Scientific, Singapore, 1988).
[31] R. Wald, *General Relativity* (The University of Chicago Press, Chicago, 1984).
[32] E. Witten, J. Diff. Geom. **17**, 661 (1982).
[33] L. Baulieu and E. Rabinovici, Phys. Lett. B **316**, 93 (1993).

Two-particle scattering theory for anyons

C. Korff[a)]
Institut für Theoretische Physik, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany

G. Lang[b)]
Fachbereich Mathematik, Technische Universität Berlin, Strasse des 17. Juni 136, 10623 Berlin, Germany

R. Schrader[c)]
Institut für Theoretische Physik, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany

(Received 22 September 1998; accepted for publication 23 October 1998)

We consider potential scattering theory of a nonrelativistic quantum mechanical 2-particle system in $\mathbb{R}^2$ with anyon statistics. Sufficient conditions are given which guarantee the existence of Møller operators and the unitarity of the *S*-matrix. As examples the rotationally invariant potential well and the δ-function potential are discussed in detail. In case of a general rotationally invariant potential the angular momentum decomposition leads to a theory of Jost functions. The anyon statistics parameter gives rise to an interpolation for angular momenta analogous to the Regge trajectories for complex angular momenta. Levinson's theorem is adapted to the present context. In particular we find that in case of a zero energy resonance the statistics parameter can be determined from the scattering phase. [S0022-2488(99)01504-2]

I. INTRODUCTION

In recent years the theory of identical quantum mechanical particles with braid group statistics has received increasing attention (for a recent review and references on anyon physics (see, e.g., Refs. 1, 2). The original observation that in two-dimensional configuration space $\mathbb{R}^2$ identical particles may obey statistics different from Bose or Fermi statistics is due to Leinaas and Myrheim.[3] They based the notion of quantum statistics on the topological structure of the classical configuration space for identical particles. The relevant symmetry group is then shown to be the braid group which replaces the permutation group. Models for particles with a one-dimensional representation of the braid group were first discussed by Wilczek, who coined the name anyons for particles with these new statistics[4] (see also Refs. 5–7). Wilczek suggested the following physical picture of anyons: Magnetic flux tubes (vortices) are attached to either charged bosons or fermions. The latter then give rise to arbitrary Aharonov–Bohm phases when transported along paths exchanging the particle positions. The magnetic flux tubes are described by long range gauge potentials whose curvature vanishes and can be related to Chern–Simons theory.[4,8] This point of view was taken up by several authors (e.g., Refs. 9, 10) leading to a second quantized version of anyons obtained by coupling a Chern–Simons U(1)-gauge potential to a matter field.[11]

Also, the general case, where the finite-dimensional representation of the braid group is not one-dimensional, has been considered. The corresponding particles are called ''plektons.''[12,13] The relevance of braid group statistics in conformal quantum field theory was realized by Tsuchiya and Kanie[14] and in algebraic quantum field theory by Fröhlich[15] and Fredenhagen, Rehren and Schroer.[12,13]

[a)]Electronic mail: christian.korff@physik.fu-berlin.de
[b)]Electronic mail: lang@math.tu-berlin.de
[c)]Electronic mail: robert.schrader@physik.fu-berlin.de

However, in the discussion of particles with braid group statistics the main focus is on anyons (Abelian statistics), since in many particle theory they might provide an explanation for the fractional quantum Hall effect and for high T_c superconductivity (for a review see, e.g., Refs. 16, 17 and 18, 19).

In this article we want to discuss nonrelativistic two-particle potential scattering theory for anyons. This is done in the framework of ordinary quantum mechanics, i.e., we insert a gauge potential of the Aharonov–Bohm type in the center of mass Hamiltonian. Emphasis is then put on showing how well known techniques of scattering theory extend to the case of anyon statistics. The motivation is twofold. On the one hand scattering theory is a powerful tool in spectral analysis and thus might be helpful for a better understanding of fractional statistics. On the other hand, scattering data can be used to compute the virial coefficients of an *interacting* anyon gas.[20] Hence, it is possible to infer bulk properties from scattering theory. The latter are of main interest in the investigation of the above mentioned phenomena.

The first calculation of the second virial coefficient for the two-particle anyon system was done by Arovas *et al.*[9] They considered the case when only the statistics is present corresponding to the case of Aharonov–Bohm scattering.[21,22] See Ref. 23 and references therein for recent articles on the second virial coefficient of interacting anyon systems. In a forthcoming publication we intend to apply the results of this article to the calculation of the second virial coefficient.

The case of non-Abelian vortex–vortex scattering has been investigated in Refs. 24–26. By looking at the irreducible components (which are then one-dimensional Abelian) our discussion then also applies.

The article is organized as follows. In Sec. II we review the quantum mechanics of two ''free'' anyons in order to introduce our notation and to keep the paper self-contained. In particular, we give the energy eigenfunctions, the resolvent (Green's function) and the propagator. We also recapitulate the relation of the free anyon system and Aharonov–Bohm scattering which will be used in our discussion of the differential cross-section in Sec. IV. In Appendix A we recall the equivalent differential geometric formulation in terms of vector bundles, which in the case of anyons are line bundles. In particular, we recall that there exists a canonical Hermitian connection encoding the statistics such that the ''free'' Hamiltonian is the canonically associated Bochner Laplacian.

In Sec. III we consider scattering theory for the interacting anyon system obtained by adding a potential to the center of mass Hamiltonian. We give sufficient conditions for the existence of the Møller operators, which also cover the nonspherical symmetric case. Applying the Kuroda–Birman theorem[27] we derive the unitarity of the resulting S-matrix.

In Sec. IV we discuss the differential cross-section with the modifications necessary to accommodate anyon statistics. We show that the scattering amplitude splits into two parts: one describing the effects of the statistics the other the interaction represented by the potential.

In Sec. V, Jost functions are introduced which depend on the statistics parameter for anyons. The latter enters in the form of continuous angular momentum. This establishes a connection with Regge trajectories in the theory of complex angular momenta. We conclude Sec. V by showing how Levinson's theorem, which relates the scattering phase shift to the number of bound states, carries over to the present situation. In case the Jost function vanishes at zero energy, we derive an explicit formula giving the statistics parameter in terms of the scattering phase shift.

Sections VI and VII are devoted to explicitly solvable examples. In Sec. VI we examine the δ-potential. This case also figures under the name of anyons without a hard-core condition. The corresponding resolvent is calculated in closed form in Appendix D and the bound state problem is then considered. We also remark on the modification of Levinson's theorem and find an additional formula relating the statistics parameter to the scattering phase. In Sec. VII we discuss the square well potential. The Jost function is calculated and Regge trajectories are plotted which show the dependence of the point spectrum on angular momentum and the statistics parameter. In Secs. VI and VII we provide numerical examples for the differential cross-section which display the interpolation between Bose and Fermi statistics when the statistics parameter for anyons is varied.

The results presented here are based in part on the diploma thesis of two of the authors (Korff[28] and Lang[29]).

Throughout the article we will work in atomic units, $\hbar = e = m = 1$, where m denotes the mass of the particles. In particular, this sets the reduced mass of a two-particle system equal to 1/2. In estimates we make the convention that C, $C(\epsilon)$, etc. denote generic constants depending on ϵ, etc.

II. QUANTUM MECHANICS OF TWO ANYONS

The theory of identical particles with statistics differing from Bose or Fermi statistics may show up when the configuration (or momentum) space of one particle is the two-dimensional Euclidean space. Henceforth we will often use the complex plane $\mathbb{C}$ to describe such a space. For the configuration space $\mathbb{C}\times\mathbb{C}$ of two nonidentical particles with points labeled by (z_1,z_2), the relative coordinate $z=z_1-z_2$ changes into $-z$ if the coordinates z_1 and z_2 of the two particles are interchanged. The basic observation of Leinaas and Myrheim was, that by leaving out the case where the two particles are at the same point, i.e., where $z=0$, the configuration space in the center of mass frame of two identical particles should be the space obtained from $\mathbb{C}^\star=\mathbb{C}\setminus\{0\}$ by identifying the points z and $-z$. This space is therefore the orbit space $\mathbb{C}^\star/\mathbb{Z}_2$, where $\mathbb{Z}_2=\{+1,-1\}$ acts in the obvious way as a transformation group on $\mathbb{C}^\star$. This space is also obtained from the closed upper half-plane $\mathbb{H}$ in $\mathbb{C}$ minus the origin, i.e., the set $\mathbb{H}\setminus\{0\}$, by identifying the points x and $-x$ on the real axis. Geometrically this leads to a cone with removed apex as configuration space. The obvious choice of polar coordinates $(r,\theta)\in\mathbb{R}^+\times[0,\pi)$ on $\mathbb{H}$ carries over to the cone $\mathbb{C}^\star/\mathbb{Z}_2$. We take the state vectors of the system to be the square integrable functions on the cone $\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)$ or equivalently the π-periodic functions on the punctured plane $\mathbb{C}^\star$. The fact that the cone's apex or, respectively, the plane's origin is removed, allows for the particles to carry flux-tubes. This corresponds to the physical picture introduced by Wilczek (see, e.g., Refs. 4, 18), who named such particles *anyons*. The flux-tubes are taken into account by inserting a gauge potential of the Aharonov–Bohm type,

$$A_\alpha=\frac{\alpha}{r}e_\theta, \quad \alpha\in[0,1], \tag{II.1}$$

into the center of mass Hamiltonian. Here α is the so called statistics parameter and e_θ denotes the unit vector corresponding to the polar angle. Choosing units in such a way that $\hbar=1$ and setting the mass of the particles to one, the resulting Hamiltonian has the form

$$H_0(\alpha)=-(\nabla+iA_\alpha)^2=-\frac{\partial^2}{\partial r^2}-\frac{1}{r}\frac{\partial}{\partial r}-\frac{1}{r^2}\left(\frac{\partial}{\partial\theta}+i\alpha\right)^2,$$

in polar coordinates. This operator will be considered to be the free Hamiltonian in the center of mass frame for a two-particle anyon system with statistics parameter $\alpha\in[0,1]$. If $\alpha=0$ the particles actually behave like bosons, while for $\alpha=1$ they behave like fermions. In Appendix A we give a short review of a mathematically precise formulation of this model in terms of vector bundles, following Refs. 30, 31 (see also Refs. 32, 33). There we also argue why $\alpha=0,1$ corresponds to bosons and fermions, respectively, and why we may restrict the parameter α to the interval [0, 1].

We now determine the spectrum and the eigenfunctions of $H_0(\alpha)$. We start with the decomposition

$$\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)\cong\mathcal{L}^2(\mathbb{R}^+,r\,dr)\otimes\mathcal{L}^2(S^1,d\theta),$$

where the points in S^1, the unit circle in $\mathbb{C}$, are parameterized as $\exp(2i\theta)$ with $0\leq\theta<\pi$. This leads to the decomposition

$$\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)\cong \bigoplus_{m\in\mathbb{Z}} \mathfrak{h}_{2m}, \quad \mathfrak{h}_{2m}=\mathcal{L}^2(\mathbb{R}^+,r\,dr)\otimes\left\{\frac{e^{2im\theta}}{\sqrt{\pi}}\right\}. \tag{II.2}$$

This decomposition is of course related to the following fact. The rotation group SO/(2)≅U(1) maps $\mathbb{C}^\star$ into itself and commutes with the action of $\mathbb{Z}_2$, thus acts on $\mathbb{C}^\star/\mathbb{Z}_2$ and defines a unitary action on $L^2(\mathbb{C}^\star/\mathbb{Z}_2)$. Also $H_0(\alpha)$ commutes with this action of U(1); it is diagonal w.r.t. the decomposition, i.e.,

$$H_0(\alpha)=\bigoplus_{m\in\mathbb{Z}} H_{0,2m}(\alpha),$$

with

$$H_{0,2m}(\alpha)=-\left\{\frac{d^2}{dr^2}+\frac{1}{r}\frac{d}{dr}-\frac{(2m+\alpha)^2}{r^2}\right\},$$

on $\mathcal{L}^2(\mathbb{R}^+,r\,dr)\cong\mathfrak{h}_{2m}$. To find all solutions of the stationary Schrödinger equation for $H_0(\alpha)$ it therefore suffices to find the solutions of the Bessel equation for each $m\in\mathbb{Z}$,

$$\left\{\frac{d^2}{dr}+\frac{1}{r}\frac{d}{dr}-\frac{(2m+\alpha)^2}{r^2}+E\right\}R(r)=0. \tag{II.3}$$

For definiteness we choose those solutions of Eqs. (II.3) which are regular near $r=0$.

This gives the following improper eigenfunctions:

$$\phi_{\alpha;m,E}(r,\theta)=\frac{e^{2im\theta}}{\sqrt{2\pi}}\cdot J_{|2m+\alpha|}(\sqrt{E}r).$$

In what follows we will use the notations

$$k:=\sqrt{E}, \quad \mu:=|2m+\alpha|.$$

The identity

$$\int_0^\infty J_\mu(at)J_\mu(bt)t\,dt=\frac{1}{\sqrt{ab}}\delta(a-b), \quad \text{for } a,b>0,$$

gives the following orthogonality and completeness relations:

$$\langle\phi_{\alpha;m,E}|\phi_{\alpha;m',E'}\rangle=\delta_{m,m'}\delta(E-E'), \tag{II.4}$$

$$\sum_{m=-\infty}^{+\infty}\int_0^\infty dE\,\phi_{\alpha;m,E}(r,\theta)\overline{\phi_{\alpha;m,E}(r',E')}=\frac{1}{r}\delta(r-r')\delta(\theta-\theta'). \tag{II.5}$$

We will also need the integral kernel for the resolvent $R_{0,\alpha}(z):=(H_0(\alpha)-z)^{-1}$, also called the Green's function, given by

$$\langle r,\theta|R_{0,\alpha}(k\pm i\epsilon)|r',\theta'\rangle=\pm\frac{i}{2}\sum_{m=-\infty}^{+\infty}e^{2im(\theta-\theta')}J_\mu(kr_<)H_\mu^\pm(kr_>). \tag{II.6}$$

Here we have used the standard convention $r_>:=\max(r,r')$ and $r_<:=\min(r,r')$. To abbreviate the notation we will sometimes make the convention that $H_\mu^\pm$ denotes the first and second Hankel function $H_\mu^{(1)}$ and $H_\mu^{(2)}$, respectively. I_μ is the modified Bessel function and K_μ is the MacDonald function (see, e.g., Ref. 34). To establish (II.6) one uses the well-known formula (see, e.g., Ref. 35).

$$\int_0^\infty kdk\,\frac{J_\mu(kr)J_\mu(kr')}{k^2+c^2}=I_\mu(cr_<)K_\mu(cr_>)[\mathrm{Re}\,\mu>-1,\mathrm{Re}\,c>0].$$

Finally we give the kernel of the unitary time evolution $\exp(-itH_0(\alpha))$:

$$\mathfrak{k}_\alpha(r,\theta;r',\theta';t):=\langle r,\theta|e^{-itH_0(\alpha)}|r',\theta'\rangle. \tag{II.7}$$

We can make this explicit by adapting a calculation, which is quite standard in the context of the Aharonov–Bohm effect.[21] Along the lines of Refs. 36, 37 we obtain the following result:

$$\mathfrak{k}_\alpha(r,\theta;r',\theta';t)=\mathfrak{k}_{\alpha,0}(r,\theta;r',\theta';t)+\hat{\mathfrak{k}}_\alpha(r,\theta;r',\theta';t). \tag{II.8}$$

Here the two subexpressions $\mathfrak{k}_{\alpha,0}$ and $\hat{\mathfrak{k}}_\alpha$ are given as

$$\mathfrak{k}_{\alpha,0}(r,\theta;r',\theta';t)=\frac{1}{2\pi it}e^{-(1/4it)(r^2+r'^2)}e^{i\alpha(\theta-\theta'-(\pi/2)\mathrm{sgn}(\theta-\theta'))}\times\cos\left(-\frac{\alpha\pi}{2}\mathrm{sgn}(\theta-\theta')+\frac{rr'}{2t}\cos(\theta-\theta')\right), \tag{II.9}$$

$$\hat{\mathfrak{k}}_\alpha(r,\theta;r',\theta';t)=\frac{i}{2\pi t}\cdot\frac{\sin(\pi\alpha)}{\pi}e^{-(1/4it)(r^2+r'^2)}I_\alpha(rr'/2t,\theta-\theta'), \tag{II.10}$$

where we have introduced

$$I_\alpha(\rho,\chi)=\int_{-\infty}^{+\infty}dy\,e^{i\rho\cosh y}\frac{e^{-y\alpha}}{1-e^{-2y-2i\chi}}. \tag{II.11}$$

The formal relation to the Aharonov–Bohm effect used in the derivation of (II.9) and (II.10) does not come by accident. In fact, the Hamiltonian of the two-anyon system coincides with the Hamiltonian of the Aharonov–Bohm effect when restricted to the subspace of symmetric wave functions. As has been realized before (see, e.g., Ref. 18) one can exploit this by describing the anyonic dynamics with the help of Aharonov–Bohm scattering and thus demonstrate the nontrivial character of the statistical interaction. The description of the Aharonov–Bohm effect in terms of scattering theory was first given by Aharonov and Bohm themselves and later taken up by several authors, among others, Refs. 22, 38, 39. We will mostly follow the discussion presented in Ref. 22 because there the time-dependent as well as the time-independent scattering formalism are considered. The wave operators and the scattering operator of the Aharonov–Bohm effect are formally defined by

$$\Omega_{AB}^{\pm}:=\underset{t\to\pm\infty}{\text{s-lim}}\,e^{itH_0(\alpha)}e^{-itH_0}\quad\text{and}\quad S_{AB}:=(\Omega_{AB}^{+})^*\Omega_{AB}^{-}, \tag{II.12}$$

respectively. Here and henceforth H_0 denotes the bosonic Hamiltonian $H_0(\alpha=0)$ and the symbol s-lim stands for the strong operator limit. Furthermore, by writing (II.12) we have implied the restriction of the Aharonov–Bohm scattering to the subspace of symmetric functions or equivalently to the space $\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)$. It has been shown[22] that the wave operators exist and are complete, whence the scattering operator is unitary. For later use we give the explicit form of the integral kernels of $\Omega_{AB}^{\pm}$, i.e., the stationary scattering states, which can be obtained by symmetrizing the results given in Refs. 22, 39,

$$\langle r,\theta|\Omega^{\pm}_{AB}|k,\theta'\rangle=\frac{1}{\pi}\sum_{m\in\mathbb{Z}}(\pm i)^{|2m+\alpha|}J_{|2m+\alpha|}(kr)e^{i2m(\theta-\theta')},\tag{II.13}$$

where $|k,\theta'\rangle$ denotes the symmetric plane wave, i.e.,

$$\langle r,\theta|k,\theta'\rangle=\frac{1}{\pi}\cos(kr\cos(\theta-\theta')).$$

The scattering phase shift corresponding to S_{AB} was first derived by Henneberger.[38] Using his result we can rewrite the scattering operator in the compact form

$$S_{AB}=e^{i\pi\alpha}P_{\geqslant}+e^{-i\pi\alpha}P_{<},\tag{II.14}$$

with $P_{\geqslant}$ and $P_{<}$ denoting the spectral projections onto the subspaces of positive, respectively, negative angular momentum. Equation (II.14) serves physical intuition by clarifying the effect of the gauge potential A_α defined in (II.1).

However, of practical importance is the integral kernel of the *S*-matrix in momentum space which was first given in Ref. 22. After symmetrization we obtain for $\alpha\in[0,1]$,

$$\langle k,\theta|S_{AB}|k',\theta'\rangle=4\,\delta(k^2-k'^2)\left[\cos\pi\alpha\,\delta(\Theta)+\frac{\sin\pi\alpha}{2\pi i}PV(1+i\cot\Theta)\right].\tag{II.15}$$

Here, $\Theta=\theta-\theta'$ and *PV* stands for the principal value prescription. From (II.15) one immediately derives the scattering amplitude,

$$f_{AB}(k,\Theta)=\left(\frac{\pi}{ik}\right)^{1/2}\langle\theta|T_{AB}(E=k^2)|\theta'\rangle=\left(\frac{\pi}{ik}\right)^{1/2}2\left[(\cos\pi\alpha-1)\,\delta(\Theta)+\frac{\sin\pi\alpha}{2\pi i}PV(1+i\cot\Theta)\right],\tag{II.16}$$

where $\langle\theta|T_{AB}(E)|\theta'\rangle$ denotes the on-shell matrix element of $T_{AB}:=S_{AB}-\mathbf{1}$ defined by the equation

$$\langle k,\theta|T_{AB}|k',\theta'\rangle=2\,\delta(E-E')\langle\theta|T_{AB}(E)|\theta'\rangle,$$

with $E=k^2$ and $E'=k'^2$. Note that Eq. (II.16) differs from what one would get when symmetrizing the amplitude in the way calculated by Aharonov and Bohm.[21] In their result the contribution of the δ-function was left out. This violates the unitarity of the *S*-matrix, as was pointed out by Ruijsenaars.[22] Thus, away from the forward direction we end up with the following expression for the differential cross-section:

$$\frac{d\sigma_{AB}}{d\theta}=|f_{AB}(k,\theta)|^2=\frac{\sin^2\pi\alpha}{\pi k}(1+\cot^2\theta),\quad\theta\neq 0,\tag{II.17}$$

which has a nontrivial angular dependence. Note that the total cross-section is infinite if $\alpha\notin\mathbb{Z}$ because of the singular contribution in the forward direction, as displayed in (II.16). The latter has been interpreted to be characteristic of anyon statistics reflecting the long range nature of the statistical interaction (Ref. 18, p. 23). For $\alpha=0,1$, i.e., for bosons and fermions, the cross-section vanishes. This is obvious from Eq. (II.14) which shows that bosons are not scattered at all while fermions pick up a factor minus one.

III. POTENTIAL SCATTERING

In this section we discuss two-particle potential scattering for anyons. Time-dependent scattering theory involves the comparison between two dynamics, one of which is considered to be ''free'' in an appropriate sense. Here, we shall consider the ''free'' dynamics to be the one defined by $H_0(\alpha)$. This appears to be a natural choice, if one views statistics to be an inherent property of the particles. It is a generalization of the fermionic picture, where usually the free time evolution acts on the space of antisymmetric functions. Note, however, that this ''free'' dynamics contains an interaction given by the long range gauge forces encoding the statistics. In particular, noninteger values of α lead to a highly nontrivial ''free'' time evolution similar to the Aharonov–Bohm effect (see our discussion at the end of Sec. II). As a consequence not all methods used in potential scattering theory may be applied. For example, we do not know how to adopt Enss'[40] method, which makes use of the Fourier transformation and is therefore not suitable in the present context. Thus, in order to prove the existence of the wave operators and the unitarity of the *S*-matrix, we will rely on Cook's method (see, e.g., Ref. 41) and the Kuroda–Birman theorem,[27] respectively.

Following our discussion in Sec. II we will formulate the two-particle scattering theory in the center of mass system using the Hilbert space $\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)$ with $H_0(\alpha)=-\Delta_\alpha$ being the free Hamiltonian. Let V, the potential, be a measurable function on $\mathbb{C}^\star/\mathbb{Z}_2$. For example, V may result from a function, also denoted by $V(z)$, on $\mathbb{C}$ with $V(-z)=V(z)$ and which acts as a multiplication operator. We set

$$H(\alpha):=H_0(\alpha)+V=-\Delta_\alpha+V$$

as an operator on $\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)$. At this point we are not concerned with giving a criterium for self-adjointness of $H_0(\alpha)$; below we specify a certain class of spherically symmetric potentials for which this operator is self-adjoint. The wave operators $\Omega_\alpha^\pm$ for the pair $(H(\alpha),H_0(\alpha))$ are defined as follows:

$$\Omega_\alpha^\pm:=\operatorname*{s-lim}_{t\to\pm\infty} e^{itH(\alpha)}e^{-itH_0(\alpha)},$$

provided the strong operator limit, denoted by s-lim, exists. Note that in the present context the absolute continuous spectrum of $H_0(\alpha)$ is the positive real axis including the origin and the associated space is all of $\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)$. The *S*-matrix is then defined as

$$S_\alpha:=(\Omega_\alpha^+)^*\Omega_\alpha^-.$$

Now we have the following theorem.

Theorem III.1: *Let $(1+r)V$ be in $\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2)$. Then the wave operators $\Omega_\alpha^\pm$ for the pair $(H(\alpha),H_0(\alpha))$ exist for all α.*

According to Cook's theorem (see, e.g., Ref. 41) it suffices to prove the following lemma.

Lemma III.1: Under the conditions on V stated in the Theorem one has

$$\int_t^\infty \|Ve^{-isH_0(\alpha)}\phi\|ds<\infty,$$

for a dense set of ϕ's and for some $t=t(\phi)<\infty$.

We will provide two proofs of this lemma. The first one will use the explicit form of the integral kernel of $e^{-itH_0(\alpha)}$ given in expressions (II.9), (II.10) and (II.11). The second proof will make use of the asymptotic form of Bessel functions near the origin and at infinity.

1. Proof of lemma III.1: It suffices to consider the case $\alpha\in(0,1)$ since the cases $\alpha=0$ and $\alpha=1$ are the already well known bosonic and fermionic situation, respectively. We will choose ϕ to be of the form $r\phi(r,\theta)=(\partial/\partial r)\Phi(r,\theta)$ with $\Phi\in C_0^\infty(\mathbb{C}^\star/\mathbb{Z}_2)$. It is easy to see that such ϕ's form a dense set.

Using (II.8), we write

$$|(e^{-itH(\alpha)}\phi)(r,\theta)|\leq F_{\alpha,0}(r,\theta;t)+\hat{F}_\alpha(r,\theta;t),$$

with

$$F_{\alpha,0}(r,\theta;t)=\left|\int_0^\infty\int_0^\pi \mathfrak{k}_{\alpha,0}(r,\theta;r',\theta';t)\phi(r',\theta')r'\,dr'\,d\theta'\right|,$$

$$\hat{F}_{\alpha}(r,\theta;t)=\left|\int_0^\infty\int_0^\pi \hat{\mathfrak{k}}_{\alpha}(r,\theta;r',\theta';t)\phi(r',\theta')r'\,dr'\,d\theta'\right|.$$

We start with an estimate of $F_{\alpha,0}$. Partial integration w.r.t. r' gives

$$F_{\alpha,0}(r,\theta;t)\leqslant\frac{1}{t^2}\int_0^\infty\int_0^\pi G_0(r,\theta;r',\theta';t)|\Phi(r',\theta')|dr'\,d\theta',$$

where G_0 satisfies an estimate of the form

$$0\leqslant G_0(r,\theta;r',\theta';t)\leqslant(1+r)C,$$

uniformly in $r,\theta,t,\alpha\in(0,1)$ and (r',θ') in the support of Φ. This gives

$$F_{\alpha,0}(r,\theta;t)\leqslant C\frac{(1+r)}{t^2},$$

and hence

$$\int_0^\infty\int_0^\pi|V(r,\theta)F_{\alpha,0}(r,\theta;t)|^2r\,dr\,d\theta\leqslant\frac{C}{t^4}\int_0^\infty\int_0^\pi|(1+r)V(r,\theta)|^2r\,dr\,d\theta, \tag{III.1}$$

uniformly in $\alpha\in(0,1)$. We turn to an estimate of $\hat{F}_\alpha$. By (II.10),

$$0\leqslant\hat{F}_\alpha(r,\theta;t)\leqslant\frac{C}{t}\cdot\left|\int_0^\infty\int_0^\pi e^{-(1/4it)(r^2+r'^2)}I_\alpha(\rho,\chi)\phi(r',\theta')r\,dr'\,d\theta'\right|,$$

with the notation

$$\rho=\frac{rr'}{2t},\quad \chi=\theta-\theta'.$$

Adding and subtracting $I_\alpha(0,\chi)$ gives

$$0\leqslant\hat{F}_\alpha(r,\theta;t), \tag{III.2}$$

$$\leqslant\frac{C}{t}\int_0^\infty\int_0^\pi|I_\alpha(\rho,\chi)-I_\alpha(0,\chi)||\phi(r',\theta')|r'\,dr'\,d\theta'$$

$$+\frac{C}{t}\left|\int_0^\infty\int_0^\pi e^{-(1/4it)(r^2+r'^2)}I_\alpha(0,\chi)\frac{\partial}{\partial r'}\Phi(r',\theta')dr'\,d\theta'\right|. \tag{III.3}$$

We need the following lemma which will be proved in Appendix B.

Lemma III.2: The quantity $I_\alpha(\rho,\chi)$ satisfies the estimates

$$|I_\alpha(\rho,\chi)|\leqslant C,$$

$$|I_\alpha(\rho,\chi)-I_\alpha(0,\chi)|\leqslant C(\epsilon)\rho^\epsilon,$$

uniformly in ρ and χ for all $0\leqslant\epsilon<\min(\alpha,1-\alpha)$.

We use this lemma combined with a partial integration w.r.t. r' in the second term of the r.h.s of (III.2) to obtain

$$|\hat{F}_\alpha(r,\theta;t)| \leq C(\epsilon)\left(\frac{r^\epsilon}{t^{1+\epsilon}}+\frac{1}{t^2}\right).$$

Therefore, since $r^\epsilon \leq 1+r$ we finally arrive at

$$\int_0^\infty \int_0^\pi |V(r,\theta)|^2 |\hat{F}_\alpha(r,\theta;t)|^2 r\,dr\,d\theta \leq C(\epsilon)\left(\frac{1}{t^2}+\frac{1}{t^{1+\epsilon}}\right)^2 \int_0^\infty \int_0^\pi (1+r)^2 |V(r,\theta)|^2 r\,dr\,d\theta. \tag{III.4}$$

Combining (III.1) and (III.4) shows that $\|Ve^{-itH_0(\alpha)}\|$ is integrable in t on the interval $[1,\infty)$, say, concluding the first Proof of lemma III.1. □

2. Proof of lemma III.1: For fixed and given $\alpha \in (0,1)$ we use the spectral decomposition of $H_0(\alpha)$ to obtain an α-dependent unitary equivalence,

$$\hat{V}_\alpha : \mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2) \to \hat{\mathcal{L}} = \bigoplus_{m\in\mathbb{Z}} \mathcal{L}^2(\mathbb{R}^+,dE), \tag{III.5}$$

defined by

$$\psi \mapsto \hat{\psi} \equiv \{\hat{\psi}_m\}_{m\in\mathbb{Z}}, \quad \hat{\psi}_m(E) = \langle \phi_{\alpha;m,E} | \psi \rangle,$$

such that $H_0(\alpha)$ just turns into a multiplication operator $\widehat{H_0(\alpha)}$:

$$(\widehat{H_0(\alpha)\psi})_m(E) = (\widehat{H_0(\alpha)}\hat{\psi})_m(E) = E\hat{\psi}_m(E).$$

Via this isomorphism we have

$$\|Ve^{-itH_0(\alpha)}\|^2 = \sum_{m,m'\in\mathbb{Z}} \int_0^\infty dE \int_0^\infty dE'\, e^{-it(E'-E)} \overline{\hat{\psi}_m(E)} v_\alpha^{m,m'}(E,E') \hat{\psi}_{m'}(E'), \tag{III.6}$$

where

$$v_\alpha^{m,m'}(E,E') = \langle \phi_{\alpha;m,E} | V^2 \phi_{\alpha;m',E'} \rangle.$$

We now choose the following dense subspace $\hat{\mathcal{D}} \subset \mathcal{D}(\widehat{H_0(\alpha)})$ in $\hat{\mathcal{L}}$:

$$\hat{\mathcal{D}} = \{\hat{\psi} \in \hat{\mathcal{L}} \,|\, \hat{\psi}_m \in C_0^\infty(\mathbb{R}^+) \text{ and } \hat{\psi}_m \equiv 0 \text{ for almost every } m\}.$$

Correspondingly, we set $\mathcal{D} = (\hat{V}_\alpha)^{-1}\hat{\mathcal{D}}$. We need the following lemma, which will be proved in Appendix B.

Lemma III.3: Let V satisfy the conditions of the Theorem. For arbitrary m and m' the functions $v_\alpha^{m,m'}(E,E')$ have measurable partial derivatives up to order 3 in E and E' on $(0,\infty)$, which are essentially bounded on compact sets.

With this lemma at hand we now proceed as follows. We use the identity

$$e^{itE} = \left(\frac{1}{it}\frac{\partial}{\partial E}\right)^n e^{itE}$$

to perform 3 partial integrations w.r.t. E in (III.6). This gives for $\psi \in \mathcal{D}$,

$$\|Ve^{-itH_0(\alpha)}\psi\|^2 \leq \frac{1}{t^3}\sum_{m,m'}\int dE\int dE'|\partial_E^3\{\overline{\hat{\psi}_m(E)}v_\alpha^{m,m'}(E,E')\}||\hat{\psi}_{m'}(E')| \leq C\frac{1}{t^3},$$

which again shows that $\|Ve^{-itH_0(\alpha)}\psi\|$ is integrable in t in the interval $[1,\infty)$ say. This concludes the second proof of lemma III.1. □

Now we turn to the situation, where the potential is centrally symmetric, i.e., where $V = V(r)$. According to Ref. 42 it is possible to define $H(\alpha)=H_0(\alpha)+V$ as a self-adjoint operator, if the potential is of the form $\gamma/r+\beta/r^b+W(r)$ say, where γ and β are arbitrary reals, $b \in [0,2)$ and W is a bounded function on $\mathbb{R}^+$. In general, $H(\alpha)$ is obviously diagonal w.r.t. the decomposition (II.2) such that

$$H(\alpha)=\bigoplus_{m\in\mathbb{Z}} H_{2m}(\alpha) \quad \text{with} \quad H_{2m}(\alpha)=H_{0,2m}(\alpha)+V,$$

acting on $\mathcal{L}^2(\mathbb{R}^+,r\,dr)$. This leads to a corresponding decomposition for the wave operators and the S-matrix,

$$\Omega_\alpha^\pm=\bigoplus_{m\in\mathbb{Z}}\Omega_{\alpha,2m}^\pm, \quad \text{where} \quad \Omega_{\alpha,2m}^\pm=\operatorname*{s-lim}_{t\to\pm\infty} e^{itH_{2m}(\alpha)}e^{-itH_{0,2m}(\alpha)} \tag{III.7}$$

and

$$S_\alpha=\bigoplus_{m\in\mathbb{Z}} S_{\alpha,2m}, \quad \text{with} \quad S_{\alpha,2m}=(\Omega_{\alpha,2m}^+)^*\Omega_{\alpha,2m}^-.$$

The Kuroda–Birman theorem[27] now leads to the following result.

Theorem III.2: *Let the centrally symmetric potential V satisfy*

$$\int_0^1|V(r)|r\,dr+\int_1^\infty|V(r)|dr<\infty.$$

Then the wave operators $\Omega_{\alpha,2m}^\pm$ exist and are complete. In particular, the S-matrices $S_{\alpha,2m}$ are unitary and so are the S_α.

We recall that the wave operators for an arbitrary pair (H,H_0) are called complete if they are unitary when considered as operators from the absolutely continuous subspace of H_0 to the absolutely continuous subspace of H. Note that these conditions on V in the central symmetric case are weaker than the conditions used in theorem III.1.

Proof: We have to show that the operator

$$|V|^{1/2}(H_{0,2m}(\alpha)+k^2)^{-1}V^{1/2}$$

in $\mathcal{L}^2(\mathbb{R}^+,r\,dr)$ has finite trace class norm $\|\cdot\|_1$, where $V^{1/2}(r)=\text{sgn}(V(r))\cdot|V(r)|^{1/2}$. Here it suffices to choose any $k^2>0$. By (II.6) we have

$$\langle r|(H_{0,2m}(\alpha)+k^2)^{-1}|r'\rangle=I_\mu(kr_<)K_\mu(kr_>),$$

with $\mu=|2m+\alpha|$. Therefore we obtain the following *a priori* estimate:

$$\||V|^{1/2}(H_{0,2m}(\alpha)+k^2)^{-1}V^{1/2}\|_1\leq\int_0^\infty|V(r)||I_\mu(kr)||K_\mu(kr)|r\,dr. \tag{III.8}$$

The following estimates for I_μ,K_μ can be derived from the asymptotic behavior near the origin and at infinity (see, e.g., Refs. 34, 35)

$$\text{for } kr\leq 1: \quad |I_\mu(kr)|\leq C(\mu)(kr)^\mu, \quad |K_\mu(kr)|\leq C(\mu)(kr)^{-\mu};$$

$$\text{for } kr \geqslant 1: \quad |I_\mu(kr)| \leqslant C(\mu)\frac{e^{kr}}{\sqrt{kr}}, \quad |K_\mu(kr)| \leqslant C(\mu)\frac{e^{-kr}}{\sqrt{kr}}.$$

Now we choose $k=1$. For this choice of k the right hand side of (III.8) is bounded by

$$C(\mu)\left\{\int_0^1 |V(r)| r dr + \int_1^\infty |V(r)| dr\right\},$$

concluding the proof of the theorem. □

IV. THE DIFFERENTIAL CROSS-SECTION

We now turn to the discussion of the two-particle scattering cross-section. The conventional approach when dealing with identical particles is to compute first the cross-section for *distinguishable* particles and in a second step to symmetrize or anti-symmetrize it in order to describe boson or fermion scattering. This is allowed because all observables commute with the projection operators onto the subspaces of symmetric and antisymmetric wave functions, respectively (see, e.g., Ref. 43 for a detailed account on this issue). In the context of anyon statistics this procedure is not applicable since no corresponding projection operators on the anyonic Hilbert spaces are available. Hence, we will investigate the two-particle cross-section in the center of mass system in a way similar to the single particle case. Doing this one encounters an additional difficulty, namely, the presence of the gauge potential A_α encoding the statistics. The latter gives rise to a nontrivial differential cross-section, even if no *dynamical* potential is present, i.e., $V\equiv 0$. This should not come as a surprise since the ''free'' dynamics generated by $H_0(\alpha)$ is described by Aharonov–Bohm scattering in the infinite time limit (see our discussion in Sec. II). Thus, the cross-section will display both the statistical as well as the dynamical interaction represented by A_α and V, respectively.

Let us first consider the simplest case where $\alpha=0$, i.e., the particles are bosons. Then the natural choice of a basis for describing the scattering is given by the *symmetric* plane waves $|k,\theta\rangle$, because the incoming asymptote is usually taken to have a sharply peaked momentum distribution. Moreover, we recall that the localization of some state ψ at large times can be determined by means of its Fourier transform,

$$\lim_{t\to\pm\infty}\int_{\mathcal{C}} |e^{-itH_0}\psi|^2 = \int_{\mathcal{C}} d\theta k dk |\langle k,\theta|\psi\rangle|^2, \tag{IV.1}$$

where $\mathcal{C} := \{(k,\theta): k\in\mathbb{R}_+, \theta\in(a,b)\subset[0,\pi]\}$ denotes a cone in real space on the left hand side and in momentum space on the right hand side. The above identity can be found in several text books on scattering theory; see, e.g., Ref. 44.

If $\alpha\notin\mathbb{Z}$ the flux-tube destroys translation invariance, whence the plane waves are not eigenstates of $H_0(\alpha)$. Thus, we have to look for an appropriate replacement such that the new basis diagonalizes $H_0(\alpha)$ and we still can form an incident wave packet with a sharply peaked momentum distribution. It turns out that this can be achieved by making use of Aharonov–Bohm scattering theory. We assign to each symmetric plane wave denoted by $|k,\theta\rangle$ the corresponding stationary Aharonov–Bohm scattering state, that is,

$$|k,\theta\rangle^{in} := \Omega_{AB}^{-}|k,\theta\rangle \quad \text{and} \quad |k,\theta\rangle^{out} := \Omega_{AB}^{+}|k,\theta\rangle. \tag{IV.2}$$

The symbols $\Omega_{AB}^{\pm}$ stand for the Aharonov–Bohm wave operators, as defined in (II.12). Note that each of the sets $\{|k,\theta\rangle^{in}\}$ and $\{|k,\theta\rangle^{out}\}$ forms a complete orthonormal system since the Aharonov–Bohm wave operators are unitary.[22] Moreover, the basis elements are (improper) eigenstates of $H_0(\alpha)$ due to the intertwining relation

$$H_0(\alpha)\Omega^{\pm}_{AB}=\Omega^{\pm}_{AB}H_0.$$

Their explicit form was given in Sec. II, Eq. (II.13). Now, by construction every wave packet in the $|k,\theta\rangle^{in}_{out}$ basis under the ''free'' anyonic time evolution $\exp(-itH_0(\alpha))$ approaches the corresponding wave packet in the plane wave basis $|k,\theta\rangle$ as $t\to\mp\infty$. Therefore, we shall refer to $|k,\theta\rangle^{in}_{out}$ as the anyonic state with incoming, respectively, outgoing (relative) momentum (k,θ). Note that according to their definition the ''in'' states are transformed into the ''out'' states by the Aharonov–Bohm S-matrix,

$$^{out}\langle k,\theta|k',\theta'\rangle^{in}=\langle k,\theta|S_{AB}|k',\theta'\rangle.$$

An additional argument for the usefulness of the two bases $|k,\theta\rangle^{in}_{out}$ is the following generalization of Dollard's theorem to anyonic dynamics:

$$\lim_{t\to\pm\infty}\int_{\mathcal{C}}|e^{-itH_0(\alpha)}\psi|^2=\int_{\mathcal{C}}d\theta k dk|^{in}_{out}\langle k,\theta|\psi\rangle|^2. \tag{IV.3}$$

The proof is immediate. By construction, the state $e^{-itH_0}(\Omega^{\pm}_{AB})^*\psi$ approaches the anyonic state $e^{-itH_0(\alpha)}\psi$ in the norm as $t\to\pm\infty$. Applying (IV.1) yields the desired relation.

We are now prepared to consider a single scattering event of two anyons when a potential V is present. Denote now by ψ the incoming wave function. According to (IV.3) the probability that the particles are scattered into $\mathcal{C}$ is given by

$$\lim_{t\to\infty}\int_{\mathcal{C}}|e^{-itH_0(\alpha)}S_\alpha\psi|^2=\int_{\mathcal{C}}d\theta\, k dk|^{out}\langle k,\theta|S_\alpha\psi\rangle|^2, \tag{IV.4}$$

where S_α is the scattering operator introduced in Sec. III. On physical grounds ψ is assumed to have a sharply peaked momentum distribution at $t=-\infty$, whence it will be given in the basis $|k,\theta\rangle^{in}$. Therefore, we perform the following transformation:

$$\begin{aligned} ^{out}\langle k,\theta|S_\alpha\psi\rangle &= \int d\theta'\, k'\, dk'\ ^{out}\langle k,\theta|S_\alpha|k',\theta'\rangle^{in\ in}\langle k,',\theta'|\psi\rangle \\ &= \int d\theta'\, k'\, dk' \langle k,\theta|(\Omega^{+}_{AB})^* S_\alpha \Omega^{-}_{AB}|k',\theta'\rangle\ ^{in}\langle k',\theta'|\psi\rangle, \end{aligned} \tag{IV.5}$$

where in the second step we have used the defining relation (IV.2). Thus, we are lead to consider the operator

$$S^{tot}_\alpha := (\Omega^{+}_{AB})^* S_\alpha \Omega^{-}_{AB},$$

which can be identified with the S-matrix resulting from the wave operators

$$\Omega^{\pm}(H(\alpha),H_0)=\operatorname*{s-lim}_{t\to\pm\infty} e^{itH(\alpha)}e^{-itH_0}.$$

This can be easily verified by using the well known chain rule for wave operators (see, e.g., Ref. 41)

$$\Omega^{\pm}(H(\alpha),H_0)=\Omega^{\pm}(H(\alpha),H_0(\alpha))\Omega^{\pm}(H_0(\alpha),H_0)=\Omega^{\pm}_\alpha\Omega^{\pm}_{AB}. \tag{IV.6}$$

Notice that the scattering operator S^{tot}_α incorporates the statistical as well as the dynamical interaction while S_α only takes care of the latter. This is best displayed by setting the potential equal to zero,

$$V \equiv 0 \Rightarrow S_\alpha = \mathbf{1}, \quad S_\alpha^{tot} = S_{AB}.$$

Therefore, in order to accommodate the anyon statistics of the particles we shall define the scattering amplitude f_α, referring to the dynamical interaction V as follows. Denote by T_α^{tot} the operator

$$T_\alpha^{tot} := (\Omega_{AB}^+)^* (S_\alpha - \mathbf{1}) \Omega_{AB}^- = S_\alpha^{tot} - S_{AB}.$$

Then the scattering amplitude is given by

$$f_\alpha(k,\theta-\theta') := \left(\frac{\pi}{ik}\right)^{1/2} \langle \theta | T_\alpha^{tot}(E=k^2) | \theta' \rangle, \tag{IV.7}$$

where $\langle \theta | T_\alpha^{tot}(E) | \theta' \rangle$ denotes the on-shell matrix element defined by the relation

$$\langle k,\theta | T_\alpha^{tot} | k',\theta' \rangle = 2\,\delta(E-E') \langle \theta | T_\alpha^{tot}(E) | \theta' \rangle,$$

with $E=k^2$ and $E'=k'^2$. From (IV.4) and (IV.5) one can now derive the differential cross-section analogously to the single-particle case, e.g., Refs. 43, 44. Away from the forward direction one ends up with the expression

$$\begin{aligned} \frac{d\sigma}{d\theta}(E_0,\theta_0) &= \pi E_0^{-1/2} |\langle \theta | T_\alpha^{tot}(E_0) + T_{AB}(E_0) | \theta_0 \rangle|^2 \\ &= |f_\alpha(k_0,\theta-\theta_0) + f_{AB}(k_0,\theta-\theta_0)|^2, \end{aligned} \tag{IV.8}$$

with $\theta \neq \theta_0$. Here, $E_0 = k_0^2, \theta_0$ determine the energy and direction of the incident particle beam and $\langle \theta | T_{AB}(E) | \theta' \rangle$ denotes the on-shell matrix element of the operator $T_{AB} = S_{AB} - \mathbf{1}$ [see Sec. II, Eq. (II.17) for the explicit expression of the Aharonov–Bohm scattering amplitude f_{AB}]. Note that we have a normalization factor π instead of 2π in (IV.8) because the polar angle is restricted to the interval $[0,\pi]$. As mentioned in Sec. II the total Aharonov–Bohm cross-section is infinite for $\alpha \notin \mathbb{Z}$, whence the total cross-section corresponding to (IV.8) is infinite as well. However, the above differential cross-section coincides with the usual one for bosons or fermions if we set $\alpha=0$ and $\alpha=1$, respectively. This can be most easily seen by use of the defining relation (IV.2) for the ''in'' and ''out'' states which for the values $\alpha=0,1$ become symmetric and antisymmetric plane waves, respectively [compare (II.13)]. In case $\alpha=1$, however, this is only true up to an angular-dependent phase factor which comes in by representing fermions as bosons with attached flux-tubes. This does not influence the outcome since the differential cross-section is given by the square modulus of the scattering amplitude. Thus, as special cases we obtain the familiar result that the differential cross-section for bosons and fermions is given by the square modulus of the symmetrized and anti-symmetrized scattering amplitude, respectively. In particular, f_{AB} vanishes and the total cross-section becomes finite for suitable short range interactions V.

V. JOST FUNCTIONS AND LEVINSON'S THEOREM

In this section we will introduce Jost functions[45] indexed by a continuous angular momentum and discuss their properties. This continuous parameter leads to an alternative formulation of the generalized Levinson theorem. We start by adapting the standard theory and results of Jost functions to the present situation (see e.g., Refs. 41, 43, 46–48). Conditions on the spherically symmetric potential will be presented at the appropriate places.

Consider the unitary map $U:\mathcal{L}^2(\mathbb{R}^+, r\,dr) \to \mathcal{L}^2(\mathbb{R}^+, dr)$, given by

$$\psi(r) \mapsto \sqrt{r}\,\psi(r). \tag{V.1}$$

Under this map the Hamiltonian $H_{2m}(\alpha) = H_{0,2m}(\alpha) + V$ turns into the operator

$$h_\mu = h_{0,\mu} + V, \tag{V.2}$$

with $\mu = |2m+\alpha| > 0$, where

$$h_{0,\mu} = -\left(\frac{d^2}{dr^2} - \frac{\mu^2 - \frac{1}{4}}{r^2}\right). \tag{V.3}$$

In this section μ will be allowed to be any positive number. Also in this section the notation $\phi_0(r;k,\mu)$ will be reserved for the regular solutions (so called because of their behavior near $r=0$) of the free Schrödinger equation $(h_{0,\mu}-k^2)\psi=0$ given as

$$\phi_0(r;k,\mu) = \sqrt{\frac{\pi k r}{2}}\, J_\mu(kr). \tag{V.4}$$

With the help of these solutions the orthogonality and completeness relations (II.4) and (II.5) are now reformulated as

$$\int_0^\infty \phi_0(r;k,\mu)\phi_0(r';k,\mu)\,dk = \frac{\pi}{2}\,\delta(r-r'), \tag{V.5a}$$

$$\int_0^\infty \phi_0(r;k,\mu)\phi_0(r;k',\mu)\,dr = \frac{\pi}{2}\,\delta(k-k'). \tag{V.5b}$$

We will also need the irregular, free solutions $\chi_0^\pm$ given as

$$\chi_0^\pm(r;k,\mu) = \pm i\sqrt{\frac{\pi k r}{2}}\, H_\mu^\pm(kr).$$

With help of these solutions the free Green's functions read [compare (II.6)] as

$$G_0^\pm(r,r';k,\mu) = \langle r|(h_{0,\mu}-k^2 \mp i\epsilon)^{-1}|r'\rangle = \frac{1}{k}\,\phi_0(r_<;k,\mu)\chi_0^\pm(r_>;k,\mu). \tag{V.6}$$

Definition V.1: For given V the function $\phi=\phi(r;k,\mu)$ is the regular solution of the equation $(h_\mu - k^2)\psi=0$ which for $r\to 0$ approximates the free, regular solution ϕ_0:

$$\lim_{r\to 0}\left(\frac{2}{kr}\right)^{\mu+1/2}\frac{\Gamma(\mu+1)}{\sqrt{\pi}}\,\phi(r;k,\mu) = 1.$$

As in the standard theory (see, e.g., Refs. 46, 47) one establishes the following facts.

(A0) ϕ is the solution to the integral equation

$$\psi = \phi_0 - G_0^< V\psi, \quad \text{with} \quad G_0^<(r,r';k,\mu) = \Theta(r-r')g_0(r,r';k,\mu). \tag{V.7}$$

Here Θ denotes the Heaviside step function and g_0 is given as

$$g_0(r,r';k,\mu) = \frac{1}{k}[\phi_0(r;k,\mu)\chi_0^-(r';k,\mu) - \chi_0^-(r;k,\mu)\phi_0(r';k,\mu)]. \tag{V.8}$$

(B0) If the potential V satisfies the condition

$$\int_0^\infty r\,dr|V(r)|<\infty,$$

then the regular solution ϕ to the eigenvalue problem $(h_{0,\mu}+\lambda V-k^2)\psi=0$ has a convergent power series expansion in λ of the form

$$\phi=\sum_{n=0}^\infty \lambda^n\phi_n, \quad \text{with} \quad \phi_{n=0}=\phi_0 \quad \text{and} \quad \phi_n=-G_0^< V\phi_{n-1}, \quad n\geqslant 1. \tag{V.9}$$

(C0) For fixed $\mu>0$, $r>0$ and real V the solution ϕ has an analytic continuation in k into the complex plane with a cut along the negative imaginary axis. There one has the relation $(k>0)$

$$\phi(r;-ik-0,\mu)=e^{i\pi(\mu+1/2)}\phi(r;ik,\mu)=e^{2i\pi(\mu+1/2)}\phi(r;-ik+0,\mu). \tag{V.10}$$

The proofs of (A0), (B0) and (C0) are as in ordinary 3-dimensional Schrödinger theory (see, e.g., Ref. 47). As a byproduct of the proof one also has the two estimates:

$$|\phi_n(r;k,\mu)|<e^{|\mathrm{Im}\,k|r}\left(\frac{|k|r}{1+|k|r}\right)^{\mu+1/2}\frac{C(\mu)^{n+1}}{n!}\left[\int_0^r dr'\frac{r'|V(r')|}{1+|k|r'}\right]^n, \tag{V.11a}$$

$$|\phi(r;k,\mu)|<C(\mu)e^{|\mathrm{Im}\,k|r}\left(\frac{|k|r}{1+|k|r}\right)^{\mu+1/2}. \tag{V.11b}$$

We will prove these estimates in Appendix C. Relation (V.10) is a consequence of the relation

$$J_\mu(e^{in\pi})=e^{in\pi\mu}J_\mu(z).$$

We now introduce the irregular solutions of the eigenvalue equation with potential as those solutions $\chi^\pm$ which approximate the free irregular solutions $\chi_0^\pm$ at $r=\infty$.

Definition V.2: The functions $\chi^\pm=\chi^\pm(r;k,\mu)$ are the solutions of the eigenvalue problem $(h_\mu-k^2)\psi=0$ satisfying the boundary conditions at infinity of the form

$$\lim_{r\to\infty} e^{\mp i(kr-(\pi/2)\mu+\pi/4)}\chi^\pm(r;k,\mu)=1.$$

These solutions will be called Jost solutions. With the help of these solutions ϕ and $\chi^\pm$ the Green's function for the full Hamiltonian h_μ now takes a form analogous to the free case [see (V.6)],

$$G^\pm(r,r';k,\mu)=\langle r|(h_\mu-k^2\mp i\epsilon)^{-1}|r'\rangle=\frac{\phi(r_<;k,\mu)\chi^\pm(r_>;k,\mu)}{W(\chi^\pm,\phi)}, \tag{V.12}$$

where W is the Wronskian, i.e.,

$$W(\psi_1,\psi_2)(r)=\psi_1(r)\partial_r\psi_2(r)-\psi_2(r)\partial_r\psi_1(r).$$

We recall that since ϕ and $\chi^\pm$ are solutions of the same eigenvalue equation, the Wronskian in (V.12) is actually independent of r, whence the notation.

Again one may establish the following facts (see, e.g., Ref. 47).

(A$\pm$) The functions $\chi^\pm$ are the solutions to the integral equations,

$$\psi=\chi_0^\pm-G_0^> V\psi, \quad G_0^>(r,r';k,\mu):=\Theta(r'-r)g_0(r,r';k,\mu). \tag{V.13}$$

Here, g_0 denotes the function defined in (V.8) which can be rewritten as

$$g_0(r,r';k,\mu)=\frac{i}{2k}[\chi_0^+(r;k,\mu)\chi_0^-(r';k,\mu)-\chi_0^-(r;k,\mu)\chi_0^+(r';k,\mu)].$$

(B±) If the potential V satisfies the condition

$$\int_0^\infty r\,dr(1+r)|V(r)|<\infty, \tag{V.14}$$

the irregular solutions $\chi^\pm$ to the eigenvalue problem $(h_{0,\mu}+\lambda V-k^2)\psi=0$ with $\pm\operatorname{Im} k\geqslant 0$, $k\neq 0$ have a convergent power series expansion in λ of the form

$$\chi^\pm=\sum_{n=0}^\infty \lambda^n\chi_n^\pm, \quad \text{with} \quad \chi_{n=0}^\pm=\chi_0^\pm \quad \text{and} \quad \chi_n^\pm=-G_0^> V\chi_{n-1}^\pm, \quad n\geqslant 1.$$

(C±) For V real and $\mu>0$, $r>0$ the functions $\chi^\pm$ are analytic in k for $\pm\operatorname{Im} k>0$ with continuous extensions to the real axis, except possibly for a singularity of order $\mu-1/2$ at $k=0$. In case V satisfies

$$\int_0^\infty r\,dr|V(r)|e^{\eta r}<\infty, \quad \eta>0, \tag{V.15}$$

then the functions $\chi^\pm$ have an analytic continuation into the domains given by $\pm\operatorname{Im} k>-\eta/2$ for all $\mu>0$, $r>0$, with the exception of a branch cut from $k=0$ to $k=\mp i(\eta/2)$. There one has the relation $(k>0)$

$$\chi^\pm(r;\mp(ik+0),\mu)=\chi^\pm(r;\mp(ik-0),\mu)+2i\cos\pi\mu\;\chi^\pm(r;\pm ik,\mu).$$

The proof of these statements is similar to the one in the regular case and involves a proof of the bounds,

$$|\chi_n^\pm(r;k,\mu)|<e^{\mp\operatorname{Im}(k)r}\left(\frac{|k|r}{1+|k|r}\right)^{-\mu+1/2}\frac{C(\mu)^{n+1}}{n!}\left[\int_r^\infty dr'\,e^{|\operatorname{Im} k|r'\mp\operatorname{Im}(k)r}\frac{r'|V(r')|}{1+|k|r'}\right]^n \tag{V.16}$$

and

$$|\chi^\pm(r;k,\mu)|<C(\mu)e^{\mp\operatorname{Im}(k)r}\left(\frac{|k|r}{1+|k|r}\right)^{-\mu+1/2}, \tag{V.17}$$

for all $\mu>0$ and $r>0$. We will establish these bounds in Appendix C. Due to the uniqueness of the solutions $\chi^\pm$ and the multivaluedness of the Hankel functions (see, e.g., Ref. 35) one has the relation

$$\overline{\chi^\pm(r;k,\mu)}=\chi^\mp(r;\bar{k},\mu)=\mp i e^{\pm i\pi\mu}\chi^\pm(r,-\bar{k},\mu).$$

We are now in the position to introduce the Jost functions in the present context and establish their properties. Since χ^+ and χ^- form a basis of solutions, ϕ is a linear combination of these two solutions. As in the standard theory (see, e.g., Ref. 47) it turns out that this linear combination is of the form

$$\phi(r;k,\mu)=\frac{i}{2}[F(k,\mu)\chi^-(r;k,\mu)-F(-k,\mu)\chi^+(r;k,\mu)],$$

where the coefficient function $F=F(k,\mu)$ is called the Jost function given by the Wronskian

$$F=\frac{1}{k}W(\chi^{+},\phi). \tag{V.18}$$

The Jost function was first introduced by and named after R. Jost. Note, however, that our definition of F follows the convention introduced by Newton,[47] which differs from the original one given in Ref. 49. We have the first main result concerning these Jost functions.

Theorem V.1: *Let the potential V satisfy the condition (V.14). Then for fixed $\mu>0$ $F(k,\mu)$ extends to an analytic function in k in the open upper half plane, which is continuous up to the real axis with the exception of the origin.*

In case the potential satisfies the stronger condition (V.15), $F(k,\mu)$ is analytic in k in $\{k|\mathrm{Im}\,k>-\eta/2\}$, except for a cut on the negative imaginary axis. Then F also satisfies the relation $(k>0)$

$$F(-ik-0,\mu)=-e^{2i\pi\mu}F(-ik+0,\mu)+2\cos\pi\mu F(ik,\mu).$$

The proof follows from the representation for F:

$$F(k,\mu)=1+\frac{1}{k}\langle\chi_0^{-}(\cdot;k,\mu)|V\phi(\cdot;k,\mu)\rangle, \tag{V.19}$$

the estimate (V.11b) for ϕ and the estimate (V.17), which is also valid for χ_0^{+}. To establish this relation we use the integral equation (V.7) for ϕ, which for large r gives

$$\phi(r;k,\mu)\xrightarrow{r\to\infty}\frac{i}{2}\left[\chi_0^{-}(r;k,\mu)\left\{1+\frac{1}{k}\langle\chi_0^{-}(\cdot;k,\mu)|V\phi(\cdot;k,\mu)\rangle\right\}\right.$$
$$\left.-\chi_0^{+}(r;k,\mu)\left\{1+\frac{1}{k}\langle\chi_0^{+}(\cdot;k,\mu)|V\phi(\cdot;k,\mu)\rangle\right\}\right]. \tag{V.20}$$

Since $\chi^{\pm}(r;k,\mu)$ approaches $\chi_0^{\pm}(r;k,\mu)$ for $r\to\infty$, relation (V.19) follows from this behavior.

Note that making use of the analytic properties just proven there is an equivalent definition of the Jost function in terms of Fredholm theory, see, e.g., Refs. 47, 50. Consider the operator $G_0^{+}(k,\mu)V$ with G_0^{+} the free Green's operator and V satisfying relation (V.14). Then for k positive imaginary $(k\in i\mathbb{R}^{+})$ the operator $G_0^{+}(k,\mu)V$ is trace class, whence its Fredholm determinant $\det(\mathbf{1}+G_0^{+}(k,\mu)V)$ is well defined. It turns out that the Jost function is just the analytic continuation of this determinant to the upper half complex plane.

We note that the series expansion (V.9) for the regular solution ϕ gives the series expansion

$$F=1+\frac{1}{k}\sum_0^{\infty}F^{n},\quad\text{with}\quad F^{n}(k,\mu)=\langle\chi_0^{-}(\cdot;k,\mu)|V\phi^{n}(\cdot;k,\mu)\rangle.$$

The bound (V.11a) for ϕ_n combined with the bound (V.17) for χ_0^{+} gives the bound

$$|F^{n}(k,\mu)|<C(\mu)\frac{C^{n}}{n!}\int_0^{\infty}dr\ e^{(|\mathrm{Im}\,k|-\mathrm{Im}\,k)r}\frac{|k|r}{1+|k|r}|V(r)|,$$

while the bound (V.11b) gives

$$|F(k,\mu)-1|\leqslant C(\mu)\int_0^{\infty}dr\ e^{(|\mathrm{Im}\,k|-\mathrm{Im}\,k)r}\frac{r|V(r)|}{1+|k|r}.$$

The importance of the Jost function stems from its role in the discussion of the S-matrix. Let $\Omega_{\mu}^{\pm}$ be the wave operators for the pair $(h_{\mu},h_{0,\mu})$ such that one has [see (III.7) and (V.1)]

$$\Omega_\mu^\pm = U\Omega_{\alpha,2m}^\pm U^{-1} \quad \text{and} \quad S_\mu = (\Omega_\mu^+)^* \Omega_\mu^- = US_{\alpha,2m}U^{-1}.$$

Since S_μ commutes with $h_{0,\mu}$ there is a decomposition of S_μ in the form

$$S_\mu = \int_\oplus dk S(k,\mu),$$

with respect to the spectral decomposition of $h_{0,\mu}$. Since the spectrum of $h_{0,\mu}$ is not degenerate, $S(k,\mu)$ acts in a one-dimensional Hilbert space and therefore is a complex number of absolute value one, i.e.,

$$S(k,\mu) = e^{2i\delta(k,\mu)}, \quad k \geqslant 0, \quad \mu \geqslant 0,$$

with $\delta(k,\mu)$ being the phase shift. We define by

$$f(k,\mu) := \frac{e^{2i\delta(k,\mu)} - 1}{\sqrt{\pi i k}}$$

the partial wave amplitude for the angular momentum $\pm\mu$, such that we obtain the following partial wave decomposition of the scattering amplitude given in (IV.7):

$$\begin{aligned} f_\alpha(k,\theta-\theta') &= \left(\frac{\pi}{ik}\right)^{1/2} \langle \theta | T_\alpha^{tot}(k^2) | \theta' \rangle \\ &= \left(\frac{\pi}{ik}\right)^{1/2} \frac{1}{\pi} \sum_{m\in\mathbb{Z}} e^{-i\pi|2m+\alpha|}(e^{2i\delta(k,|2m+\alpha|)} - 1) e^{i2m(\theta-\theta')} \\ &= \sum_{m\in\mathbb{Z}} e^{-i\pi|2m+\alpha|} f(k,|2m+\alpha|) e^{i2m(\theta-\theta')}. \end{aligned} \tag{V.21}$$

Here we have used the relation $\langle k,\theta | T_\alpha^{tot} | k',\theta' \rangle = {}^{out}\langle k,\theta | S_\alpha - \mathbf{1} | k',\theta' \rangle^{in}$ as well as the identities

$$\langle r,\theta | k,\theta' \rangle_{out}^{in} = \sqrt{2\pi}\, \frac{1}{\pi} \sum_{m\in\mathbb{Z}} (\mp i)^{|2m+\alpha|} \phi_{\alpha;m,E}(r,\theta) e^{-i2m\theta'},$$

$$\langle \phi_{\alpha;m,E} | S_\alpha \phi_{\alpha;m',E'} \rangle = \delta_{mm'} \delta(E-E') e^{2i\delta(k,|2m+\alpha|)}.$$

Now set

$$\phi^\pm(r;k,\mu) = (\Omega_\mu^\pm \phi_0)(r;k,\mu).$$

By construction one has a solution of the eigenvalue equation $(h_\mu - k^2)\psi = 0$ and $\phi^\pm$ satisfies the Lippmann–Schwinger equation,

$$\phi^\pm(\cdot;k,\mu) = \phi_0(\cdot;k,\mu) - (h_\mu - k^2 \pm i0)^{-1} V \phi^\pm(\cdot;k,\mu).$$

With the help of $\phi^\pm$ one may express the partial wave amplitude as

$$f(k,\mu) = -\frac{2}{k^2}\left(\frac{ik}{\pi}\right)^{1/2} \langle \phi_0(\cdot;k,\mu) | V\phi^+(\cdot;k,\mu) \rangle.$$

The solution ϕ^+ of the Lippmann–Schwinger equation has the following asymptotic behavior for $r \to \infty$:

$$\phi^{+}(r;k,\mu) \xrightarrow[r\to\infty]{} \phi_0(r;k,\mu)+\tfrac{1}{2}\sqrt{\pi k}f(k,\mu)e^{i(kr-(\pi/2)\mu)},$$

where we have used $\chi_0^{+}(r;k,\mu) \xrightarrow[r\to\infty]{} e^{i(kr-(\pi/2)\mu+\pi/4)}$. Alternatively, using the relation $\phi_0 = i/2(\chi_0^{-}-\chi_0^{+})$ one gets

$$\phi^{+}(r;k,\mu) \xrightarrow[r\to\infty]{} \frac{i}{2}(\chi_0^{-}(r;k,\mu)-e^{2i\delta(k,\mu)}\chi_0^{+}(r;k,\mu)).$$

With the same arguments as in the standard theory (see, e.g., Refs. 41, 46, 47) one proves the following.

Lemma V.1: Let $\mu>0$ be fixed and assume that V satisfies the condition (V.14). Then one has the following.

(i) For k on the real axis ($k\in\mathbb{R}^{+}$), the following identities are valid:

$$\phi(r;k,\mu)=F(k,\mu)\phi^{+}(r;k,\mu), \tag{V.22}$$

$$S(k,\mu)=e^{2i\delta(k\mu)}=\frac{\overline{F}(k,\mu)}{F(k,\mu)}=\frac{F(-k,\mu)}{F(k,\mu)}. \tag{V.23}$$

Furthermore, $F(k,\mu)$ can only vanish at the origin.

(ii) The zeros of $F(k,\mu)$ in the upper half plane $\mathrm{Im}\,k>0$ all lie on the positive imaginary axis and are simple. In particular, $k\in i\mathbb{R}^{+}$ is a zero of $F(k,\mu)$ *iff* k^2 is the energy of a bound state of h_μ with angular momentum $\pm\mu$.

(iii) The following limit relation is valid in the closed upper half plane:

$$\lim_{k\to\infty} F(k,\mu)=1.$$

Since the proof is analogous to the one given in three-dimensional Schrödinger theory[41,46,47] we omit it here. Now, the famous Levinson's theorem follows as a corollary by use of the residue theorem. Only the so called resonance case, i.e., $F(k=0,\mu)=0$, requires some more work.

Theorem V.2: *Let $\mu>0$ and V satisfy* (V.14). *If $F(k=0,\mu)\neq 0$ one has the following relation between the phase shift $\delta(k,\mu)$ and the number n_μ of bound states with angular momentum $\pm\mu$,*

$$\delta(k=0,\mu)-\delta(k=\infty,\mu)=\pi n_\mu. \tag{V.24}$$

Proof: The proof of Eq. (V.24) is standard and can be found in, e.g., Refs. 41, 43, 46, 47. For the sake of completeness we recall the main arguments. Integrate the logarithmic derivative of F with respect to k over a closed semi-circle in the complex upper half plane centered at the origin. According to the preceding lemma the integrand has simple poles on the positive imaginary axis, each of which corresponds to a bound state of h_μ. While the contribution of the semi-circle vanishes in the limit of infinite radius [see (III.4)] the integration over the real axis yields the phase difference by Eq. (V.23). Applying the residue theorem completes the proof. □

We comment on the results of the above theorem. Equation (III.5) generalizes Levinson's theorem to continuous angular momentum μ. One might wonder what happens when μ varies. On physical grounds one expects that for higher angular momentum μ there are fewer bound states since then the centrifugal barrier is stronger. Thus, when μ is steadily increased the zeros of the Jost function F move along the positive imaginary axis towards the origin. Every time a zero moves out of the upper half plane we get a discontinuous jump in the phase shift indicating the disappearance of a bound state.

Theorem V.3: *Let $\mu\notin\mathbb{Z}$, $\mu>0$ and assume V satisfies* (V.15). *In case $F(k=0,\mu)=0$ the above form of Levinson's theorem remains valid when $\mu>1$ and is modified as*

$$\delta(k=0,\mu)-\delta(k=\infty,\mu)=\pi(n_\mu+\mu), \tag{V.25}$$

when $\mu<1$.

So far we have not been able to settle the case $\mu \in \mathbb{Z}$ corresponding to boson and fermion statistics. For real anyons, however, note that in case of positive angular momentum we have $\mu=\alpha$ when $\mu<1$. Thus, Eq. (V.25) then specializes to

$$\delta(k=0,\alpha)-\delta(k=\infty,\alpha)=\pi(n_\mu+\alpha), \tag{V.26}$$

showing that one can determine the statistics parameter from the scattering phase independently of the detailed form of the short range potential V. Below we will find a similar formula for the delta potential. Setting $\alpha=1/2$ (i.e., for semions) in (V.25) we obtain the well known resonance case of three-dimensional Schrödinger theory,[46] since then the radial Schrödinger equation is formally identical to the one in three dimensions [compare (V.2) and (V.3)].

Proof: The strategy for proving (V.25) is similar to the one used for (V.24). The techniques used are analogous to those in three-dimensional Schrödinger theory (see Refs. 43, 46) but the line of argument is slightly changed.

Because $F(k=0,\mu)=0$ we change the integration contour as follows. We cut the semi-circle at the origin and insert a small semi-circle C_ρ of radius $\rho \ll 1$ centered at the origin. Then there is an additional contribution to Eq. (V.24), namely,

$$\pi n'_\mu=\delta(k=0,\mu)-\delta(k=\infty,\mu)+\lim_{\rho\to 0}\frac{1}{2i}\oint_{C_\rho} d\ln F, \tag{V.27}$$

where n'_μ is the number of bound states with the exception of a possible zero energy eigenstate. In order to compute the second term of the right hand side in the above equation we only need to know the small energy behavior of F. Since we have assumed that the potential is decaying exponentially we can continue F in the lower half plane and expand it into a convergent power series around the origin for suitably small k. We start by rewriting expression (V.19),

$$\begin{aligned} F(k,\mu)&=1+\frac{1}{k}\langle \chi_0^-(\cdot;k,\mu)|V\phi(\cdot;k,\mu)\rangle \\ &=1+\frac{1}{k\sin\pi\mu}[\langle \phi_0(\cdot;k,-\mu)|V\phi(\cdot;k,\mu)\rangle - e^{-i\pi\mu}\langle\phi_0(\cdot;k,\mu)|V\phi(\cdot;k,\mu)\rangle]. \end{aligned} \tag{V.28}$$

Here we have used the identity $H_\mu^{(1)}(z)=(1/i\sin\pi\mu)[J_{-\mu}(z)-e^{-i\pi\mu}J_\mu(z)]$ which can be found in, e.g., Refs. 34, 35. Note that in case of integer μ the Hankel function is defined by the limit of the r.h.s. of the last equation. Then the above decomposition (V.28) is no longer valid, whence we restrict ourselves to the noninteger case. Both integrals in (V.28) converge uniformly as can be shown by use of the estimates

$$|\phi_0(r;k,\mu)\phi(r;k,\mu)|\le C(\mu)e^{2|\mathrm{Im}\,k|r}\left(\frac{|k|r}{1+|k|r}\right)^{2\mu+1},$$

$$|\phi_0(r;k,-\mu)\phi(r;k,\mu)|\le C(\mu)e^{2|\mathrm{Im}\,k|r}\left(\frac{|k|r}{1+|k|r}\right),$$

which follow from (V.11a) and (V.11b). The functions $\phi_0(r;k,-\mu)$ and $\phi_0(r;k,\mu)$ can be written as a convergent power series in k^2 times a factor $k^{-\mu+1/2}$ and $k^{\mu+1/2}$, respectively. This follows from their definition (V.4) and the series expansions of the Bessel functions $J_{-\mu}(kr),J_\mu(kr)$; see, e.g., Refs. 34, 35. From (V.9) we see that also $\phi(r;k,\mu)$ can be written as a convergent power series in k^2 times $k^{\mu+1/2}$, since the Green's function is an even function in k. Therefore, in a suitably small chosen neighborhood of $k=0$ the Jost function can be expanded in a convergent series of the form

$$F(k,\mu)=\sum_{n=0}^{\infty} a_n(\mu)k^{2n}+k^{2\mu}\sum_{n=0}^{\infty} b_n(\mu)k^{2n}, \quad |k|\ll 1. \tag{V.29}$$

Recall our convention to place the branch cut of F along the negative imaginary axis. The assumption $F(0,\mu)=0$ gives $a_0=0$. In order to deduce from this expansion the small energy behavior of F, we need to know the first nonvanishing coefficient in (V.29). We claim $a_1(\mu)\neq 0$ for $\mu>1$ and $b_0(\mu)\neq 0$ for $\mu<1$. But then we have

$$F(k,\mu)=\begin{cases}\mathcal{O}(k^{2\mu}), & \text{if } \mu<1,\\ \mathcal{O}(k^2), & \text{if } \mu>1.\end{cases}$$

Thus, evaluating the integral over the semi-circle C_ρ yields

$$\lim_{\rho\to 0}\frac{1}{2i}\oint_{C_\rho} d\ln F=\begin{cases}-\pi\mu, & \text{if } \mu<1,\\ -\pi, & \text{if } \mu>1.\end{cases} \tag{V.30}$$

In order to prove the claim we consider the limit $\lim_{k\to 0}k^{-1}(\partial/\partial k)F$. If we can show that

$$\lim_{k\to 0}k^{-1}\frac{\partial}{\partial k}F(k,\mu)=\begin{cases}\infty, & \text{if } \mu<1,\\ C(\mu)\neq 0, & \text{if } \mu>1,\end{cases}$$

then the claim follows, as can be inferred directly from (V.29). We calculate the expression $(\partial/\partial k)F$ by making use of the identity (V.18). Since the solutions χ^+, ϕ are ill defined for $k\to 0$ we consider instead the modified solutions

$$\Psi(r;k,\mu):=k^{\mu-1/2}\chi^+(r;k,\mu) \quad \text{and} \quad \Phi(r;k,\mu):=k^{-\mu-1/2}\phi(r;k,\mu),$$

which are both finite and nonvanishing at $k=0$. This can be seen by taking the limit $k\to 0$ in the defining integral equations (V.7) and (V.13), respectively. In particular, one then derives that $\Phi(r;0,\mu)$ behaves like $r^{\mu+1/2}$ for $r\to 0$ and $\Psi(r;0,\mu)$ like $r^{-\mu+1/2}$ for $r\to\infty$. In terms of Φ, Ψ the Jost function and its derivative w.r.t. k (denoted by a dot) then read as

$$F=W(\Psi,\Phi) \quad \text{and} \quad \dot F=W(\dot\Psi,\Phi)+W(\Psi,\dot\Phi).$$

Since Ψ is a solution of the radial equation one easily verifies the identity

$$\frac{\partial}{\partial r}W[\Psi(r;k,\mu),\Psi(r;k',\mu)]=(k^2-k'^2)\Psi(r;k,\mu)\Psi(r;k',\mu).$$

Differentiating w.r.t. k and setting subsequently $k=k'$ one obtains from this the relation

$$W[\dot\Psi(r;k,\mu),\Psi(r;k,\mu)]=-2k\int_r^\infty dr'\,\Psi(r';k,\mu)^2, \tag{V.31}$$

where the r.h.s. is finite for $\operatorname{Im}k>0$. The assumption $F(0,\mu)=0$ implies that the regular and the irregular solution are proportional when $k=0$, i.e., $\Psi(r;0,\mu)=\kappa(\mu)\Phi(r;0,\mu)$ holds. Note that the proportionality factor $\kappa(\mu)$ is nonzero and finite because both solutions are nonvanishing. Thus, in the limit $k\to 0$ we have

$$\dot F(0,\mu)=\kappa(\mu)^{-1}W(\dot\Psi(r;0,\mu),\Psi(r;0,\mu))+\kappa(\mu)W(\Phi(r;0,\mu),\dot\Phi(r;0,\mu)).$$

However, $\dot{F}(0,\mu)$ does not depend on r whence we can set $r=0$ in the last equation. Since the regular solution Φ vanishes at $r=0$ we obtain from (V.31),

$$\lim_{k\to\infty} k^{-1}\frac{\partial}{\partial k}F(k,\mu)=\lim_{k\to 0} k^{-1}\kappa(\mu)^{-1}W[\dot{\Psi}(0;k,\mu),\Psi(0;k,\mu)]$$

$$=-2\kappa(\mu)^{-1}\int_0^\infty dr\Psi(r;0,\mu)^2$$

$$=-2\kappa(\mu)\int_0^\infty dr\Phi(r;0,\mu)^2. \tag{V.32}$$

The irregular solution Ψ falls off like $r^{-\mu+1/2}$ for large r and is proportional to the regular solution Φ at $k=0$, therefore the r.h.s. of (V.32) is finite for $\mu>1$. This means that the regular solution is square integrable and hence a proper eigensolution of the full Hamiltonian h_μ. The number of bound states is thus $n_\mu=n'_\mu+1$, whence we read off from (V.27) and (V.30) that Levinson's theorem also holds in case $F(0,\mu)=0$ when $\mu>1$.

For $\mu<1$, however, the regular solution is no longer square integrable and the expression (V.32) is divergent. Hence the number of bound states is given by $n_\mu=n'_\mu$ and from (V.27) and (V.30) we infer the modified equation (V.25). This completes the proof. □

VI. THE δ-POTENTIAL

In this section we will introduce and discuss the δ-potential, often also called the contact potential or zero-range interaction potential because it describes an interaction which is ''nonvanishing'' only if the two particles are at the same place, i.e., if $z_{rel}=0$. In the literature this figures also under the notion of anyons without a hard-core condition.[23] See, e.g., Refs. 51–53 for previous articles on this subject also in connection with field theoretic considerations.

To introduce the δ-potential we follow a standard strategy (see, e.g., Ref. 54), i.e., we consider the free Hamiltonian (in the center of mass system) restricted to a definition domain of wave functions which vanish at $z_{rel}=0$. The resulting operator has deficiency indices (1,1) and so there is a one-parameter family of self-adjoint extensions, one of which is of course the ordinary free Hamiltonian given as the Friedrich's extension. We discuss the bound state problem and the resulting scattering theory. We also check the validity of Levinson's theorem.

Note that Aharonov–Bohm scattering on $\mathbb{R}^2$ with δ-type interactions have been discussed in Refs. 55, 56. On $\mathbb{R}^2$ deficiency indices are (2, 2), leading to a four-parameter family of self-adjoint extensions. By specializing to symmetric functions one can obtain some of the results below.

For a start, consider the operator

$$h_{0,\mu}=-\partial_r^2+\frac{\mu^2-4^{-1}}{r^2},$$

on $\mathcal{L}^2(\mathbb{R}^+,dr)$ [recall (V.1)] with domain of definition $\mathcal{D}(h_{0,\mu})$ consisting of wave functions $\psi(r)$ having compact support away from the origin such that ψ and $\partial_r\psi$ are locally absolutely continuous and such that $h_{0,\mu}\psi$ (defined in the sense of distributions) is an element of $\mathcal{L}^2(\mathbb{R}^+,dr)$. Recall that locally absolutely continuous functions are such that their derivatives in the sense of distributions are locally integrable functions (see, e.g., Ref. 57).

Let us now investigate the most general solution of the equation $h_{0,\mu}\psi=0$, given by

$$c_1\cdot r^{1/2+\mu}+c_2\cdot r^{1/2-\mu}.$$

Using Weyl's terminology (consult, e.g., Ref. 58), one has the limit point case at $r=\infty$. At $r=0$ one also has the limit point case if and only if $\mu\geqslant 1$, otherwise one has the limit circle case. If now $0\leqslant\mu<1$, by Weyl's alternative there is exactly one solution of $h_{0,\mu}\psi=0$, which is $\mathcal{L}^2$ near infinity. By construction this solution must be $\mathcal{L}^2$ at both infinity and at zero, and thus it follows that $\dim(\ker(h_{0,\mu}\pm i))=1$ by Weyl's alternative. Therefore the deficiency indices are (1,1) and there is a one-parameter family of s.a. extensions (see e.g. Ref. 50).

In terms of the statistics parameter α we have the following situation: If one chooses $\alpha=1$, then $\mu=|2m+\alpha|\geqslant 1$ for all $m\in\mathbb{Z}$ and all the $\{h_{0,\mu}\}$ are essentially self-adjoint. Now, if $\alpha\in[0,1)$ we have $\mu<1$ if and only if $m=0$. In this case, just as in the three-dimensional case, the δ-potential affects only one of the angular momentum channels, namely, the one associated with the smallest eigenvalue in the sense of the modulus (in three dimensions called the *s*-channel).

Let α be in $[0,1)$. To construct the s.a. extensions of the operators $h_{0,\alpha}$ with domain $\mathcal{D}(h_{0,\alpha})$ given above, following Refs. 42, 54 we define the regular and irregular solutions of $h_{0,\alpha}\psi=0$:

$$F_\alpha(r):=r^{1/2+\alpha}\quad\text{(regular)},$$

$$G_\alpha(r):=F_\alpha(r)\int_r^{r_0}dr'(F_\alpha(r'))^{-2}\quad\text{(irregular)},$$

where $r_0\in\mathbb{R}^+$ may be chosen arbitrary. The choice $r_0=1$ gives

$$G_\alpha(r)=\begin{cases}\dfrac{1}{2\alpha}(r^{1/2-\alpha}-r^{1/2+\alpha}), & \alpha>0,\\ -\sqrt{r}\ln r, & \alpha=0.\end{cases}$$

In order to explicitly construct the s.a. extensions of the $\{h_{0,\alpha}\}$ we take recourse to the following theorem, to be found in, e.g., Ref. 58.

Theorem VI.1: *For $\alpha<1$ the operator $h_{0,\alpha}$ with domain $\mathcal{D}(h_{0,\alpha})$ has a one-parameter family of s.a. extensions $\{h_{0,\alpha}(s)\}_{s\in\mathbf{R}}$ which are essentially s.a. on the domains*

$$\mathcal{D}(h_{0,\alpha}(s))=\{\psi\in\mathcal{D}(h_{0,\alpha}^*)|\lim_{r\downarrow 0}W(\psi_{\alpha,s},\psi)(r)=0\}.\tag{VI.1}$$

Here again W is the Wronskian and

$$\psi_{\alpha,s}:=G_\alpha+sF_\alpha.$$

The choice $s=\infty$ gives the Friedrich's extension.

In Appendix D we will prove the following.

Theorem VI.2: *The integral kernel of the resolvent of $h_{0,\alpha}(s)$ for $\alpha<1$ is given as*

$$\frac{1}{h_{0,\alpha}(s)-k^2}(r,r')=\frac{i\pi}{2}\sqrt{rr'}J_\alpha(kr_<)H_\alpha^{(1)}(kr_>)-A(k,\alpha;s)\cdot\sqrt{rr'}H_\alpha^{(1)}(kr)H_\alpha^{(1)}(kr'),\tag{VI.2}$$

where k satisfies $0<\arg(k)<\pi$. $A(k,\alpha;s)$ is given as

$$A(k,\alpha;s)=\frac{\pi}{2}\left[\left(\frac{k}{2}\right)^{-2\alpha}(2s\alpha-1)\frac{1}{\sin\pi\alpha}\frac{\Gamma(1+\alpha)}{\Gamma(1-\alpha)}+\cot\pi\alpha-i\right]^{-1}.\tag{VI.3}$$

Note that $\lim_{s\to\infty}A(k,\alpha;s)=0$ such that with (VI.2) in the limit $s\to\infty$ we indeed recover the kernel of the resolvent of the Friedrich's extension (V.6). Taking the ''bosonic limit'' $\alpha\to 0$ one arrives at the expression

$$A(k,0;s)=\frac{\pi^2}{4}\left[\ln\frac{k}{2i}+\gamma+s\right]^{-1}\quad(\gamma:\ \text{Euler's constant}).$$

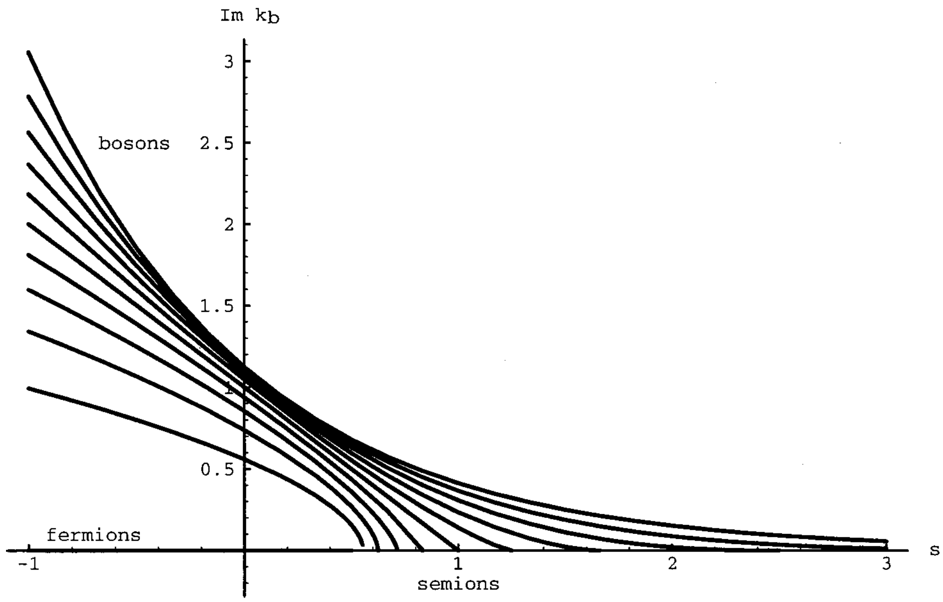

FIG. 1. Bound state energies.

Combined with (VI.2) we have arrived at an expression for the resolvent of the δ-potential in the bosonic case, which agrees with Ref. 54.

We turn to the discussion of the bound states. It can be shown that the essential spectrum as a set is equal to $\mathbb{R}_0^+$ and that $\mathbb{R}^+$ is the purely absolutely continuous spectrum. We therefore have to look for poles of the resolvent at energies $E_b=k_b^2<0$. By (VI.2) we have to search for the poles of $A(k,\alpha;s)$ as a function of k, and by (VI.3) this gives the condition

$$k_b^{2\alpha}=2^{2\alpha}e^{i\pi\alpha \bmod 2\pi i}\frac{\Gamma(1+\alpha)}{\Gamma(1-\alpha)}(1-2\alpha s). \tag{VI.4}$$

The restriction $0<\arg(k)<\pi$ now implies $0<\arg(k^{2\alpha})<2\pi\alpha$. Applying this condition to (VI.4) leads to three different cases.

(i) $1-2\alpha s>0$: Then (VI.4) may be solved to give

$$k_b=k_b(\alpha,s)=2i\left(\frac{\Gamma(1+\alpha)}{\Gamma(1-\alpha)}(1-2\alpha s)\right)^{1/2\alpha},$$

i.e., k_b is purely imaginary. The case $k_b=0$ occurs when $\alpha\uparrow 1$. Note that this requires $0\leqslant s<\frac{1}{2}$. Hence there is exactly one bound state.

(ii) $1-2\alpha s<0$: Since by definition $0\leqslant\alpha<1$, it is easy to see that in this case there is no solution to Eq. (VI.4), only in the limit $\alpha\uparrow 1$ one obtains $k_b=0$. This means that there is no bound state.

(iii) $1-2\alpha s=0$: This immediately gives $k_b=0$, which implies that there is no bound state.

For comparison we recall that in the three-dimensional case the attractive δ-potential supports, at most, one bound state. A bound state exists if the particles obey bosonic statistics. For fermions there is none. In the present context a bound state always exists for $\alpha=0$ (bosons) and disappears at the latest when $\alpha=1$ (fermions). The fact that for $\alpha=0$ there is always a bound state suggests that in two dimensions there is no such thing as a repulsive δ-potential.

Below we argue that case (iii) corresponds to a zero energy resonance in three-dimensional Schrödinger theory. Figure 1 gives a plot of $|k_b(\alpha,s)|$ as a function of s for various values of α in case (i), i.e., when $1-2\alpha s>0$. To calculate the bound state wave function ψ_b when $1-2\alpha s<0$, we note that

$$\frac{1}{h_{0,\alpha}(s)-k^2}\approx\frac{1}{k_b^2-k^2}P_b,$$

for $k\approx k_b$ where P_b is the orthogonal projector onto the eigenspace of energy $E_b=k_b^2<0$. Since this pole of the resolvent is isolated, P_b may be calculated as a residue. The calculation is easy and shows that P_b is indeed one-dimensional and that the associated (normalized) eigenfunction ψ_b is given as

$$\psi_b(r;\alpha)=\left(\frac{2\sin\pi\alpha}{\pi\alpha}\right)^{1/2}\sqrt{r|E_b|}K_\alpha(|k_b|r),$$

for $\alpha<1$ and hence $|k_b(\alpha,s)|>0$. For any fixed $\alpha<1$ the wave function decays exponentially for large r, which guarantees square-integrability. In the limit $\alpha\uparrow 1$ the bound state wave function approaches $r^{-1/2}$ times a normalization constant, which is not square integrable. The case $\alpha=1$ should therefore be considered as a resonance at $E=0$.

We turn to a discussion of the resulting scattering theory. First we note that due to relation (VI.2) the difference of the resolvents for $h_{0,\alpha}(s)$ and the free Hamiltonian $h_{0,\alpha}(s=\infty)$ is a rank-one operator and therefore in particular trace class. Therefore the S-matrix exists and is unitary by the Kuroda–Birman theorem (see, e.g., Ref. 27).

Theorem VI.3: *For given α and $E=k^2>0$ the outgoing states for the Hamiltonian $h_{0,\alpha}(s)$ are given as*

$$\phi^+(r;k,\alpha)=\phi_0(r;k,\alpha)+\frac{2i}{\pi}A(k,\alpha;s)\cdot\chi_0^+(r;k,\alpha).$$

In particular, the resulting partial wave amplitude and the phase shift take the form

$$f(k,\alpha;s)=\frac{4i}{\pi}\frac{A(k,\alpha;s)}{\sqrt{\pi ik}},\tag{VI.5a}$$

$$e^{2i\delta(k,\alpha;s)}=1+\frac{4i}{\pi}A(k,\alpha;s).\tag{VI.5b}$$

Proof: We start by rewriting the resolvent (VI.2) as

$$\frac{1}{h_{0,\alpha}(s)-k^2}(r,r')=\left(\phi_0(r_<;k,\alpha)+\frac{2i}{\pi}A(k,\alpha;s)\chi_0^+(r_<;k,\alpha)\right)\frac{i}{k}\chi_0^+(r_>;k,\alpha).\tag{VI.6}$$

Now we observe that (see, e.g., Ref. 54, p. 37)

$$\lim_{\epsilon\downarrow 0}\lim_{r'\to\infty}e^{-i(k+i\epsilon)r'}\frac{1}{h_{0,\alpha}(s)-(k+i\epsilon)^2}(r,r')\tag{VI.7}$$

must be proportional to the outgoing solution $\phi^+(r;k,\alpha)$. The claim now follows by applying (VI.7) to (VI.6). □

In view of Levinson's theorem we would like to make the following comments on this last result: The limiting values of $\delta(k,\alpha;s)$ at $k\to 0^+$ and $k\to\infty$ are *a priori* determined up to an additive constant in $\pi\cdot\mathbb{Z}$, only. We shall choose these constants in such a way, that $\delta(k,\alpha;s)$ as a function of $k \in \mathbb{R}^+$ can be defined as a continuous function of k. Now, $\exp(i2\delta(k,\alpha;s))$ is of the form $(a(k;\alpha;s)+i)(a(k,\alpha;s)-i)^{-1}$. Here $a(k,\alpha;s)$ is a real-valued, continuous function of k. Explicitly,

$$a(k,\alpha;s)=\frac{\pi}{2}A(k,\alpha;s)^{-1}+i.$$

[This observation also shows that the right hand side of (VI.5b) is indeed unimodular.] Hence

$$\delta(k,\alpha;s)=\arg(a(k,\alpha;s)+i).$$

Here the branch of $\arg(\cdot)$ is chosen such that $\delta(k,\alpha;s)$ is a continuous function of all its arguments. In particular, we can achieve $\delta(k,\alpha;s)\in(0,\pi)$, for any k and fixed $\alpha<1$ and s. For the cases, where there is a bound state, this is consistent with the attractive nature of the interaction. By the explicit form of $a(k,\alpha;s)$, it is easily seen that

$$\delta(k,\alpha;s)\xrightarrow{k\to 0^+}\pi\frac{1-\mathrm{sgn}(2\alpha s-1)}{2},$$

$$\delta(k,\alpha;s)\xrightarrow{k\to\infty}\pi\alpha.$$

The fraction $n:=[1-\mathrm{sgn}(2\alpha s-1)]/2$ takes values in $\{0, 1\}$ and is, in fact, nothing but the number of bound states given by (VI.4). We therefore have the following relation:

$$\delta(0^+,\alpha;s)-\delta(\infty,\alpha;s)=\pi(n-\alpha), \tag{VI.8}$$

which is *not* of the form one would naively expect in view of Levinson's theorem V.2. In this context we note that the statement of Levinson's theorem also fails in the nonanyonic three-dimensional case, if a δ-interaction is considered. However, we obtain a formula similar to (V.26) relating the phase shift to the statistics parameter α.

By relations (IV.8) and (V.21), we obtain the following expression for the differential cross-section:

$$\frac{d\sigma}{d\theta}=|f(k,\alpha;s)\cdot e^{-i\pi\alpha}+f_{AB}(k,\theta)|^2,$$

where f_{AB} and f are given by expressions (II.16) and (VI.5a), respectively. Inserting these explicit expressions, one obtains—away from the forward direction—the following relation:

$$\frac{d\sigma}{d\theta}=\frac{1}{\pi k}\left|\sin\pi\alpha(\cot\theta-i)+e^{-i\pi\alpha}\frac{4i}{\pi}A(k,\alpha;s)\right|^2.$$

So far, we have been unable to analyze this in a non-numerical way. Figures 2 and 3 show some numerical results of the differential cross-section as a function of the scattering angle θ. Displayed is the deviation from pure Aharonov–Bohm scattering, which is asymptotically reached when $s\to\pm\infty$. Note the simultaneous cross over at $\theta=\pi\alpha$ for varying s in each of the plots with $s\neq\pm\infty$. In particular, the point corresponding to the cross over always lies on the graph ($s=\pm\infty$) associated with the Aharonov–Bohm cross-section. We like to point out that in both Figs. 2 and 3 only for $s=0,2$ and $\alpha=0.5$ (semions) the differential cross-section actually vanishes for some scattering angle θ. In the other cases the cross-section becomes very small but is nonvanishing for all θ.

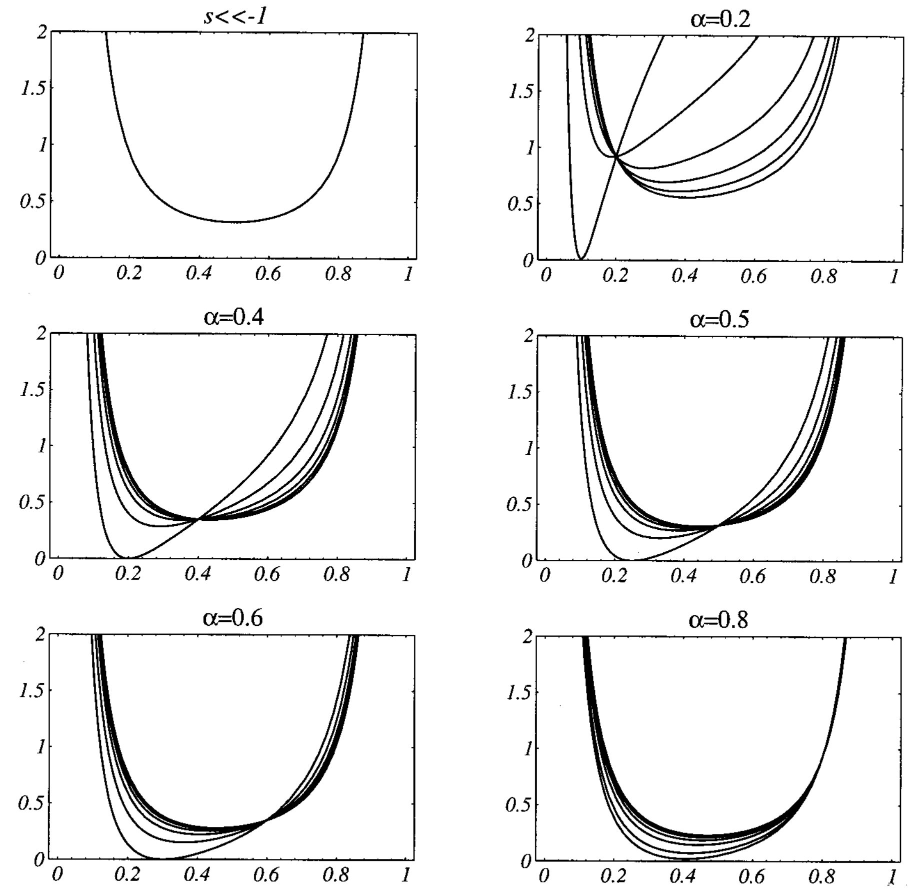

FIG. 2. Cross-sections for contact interaction 1.

VII. THE SQUARE WELL POTENTIAL

Another explicitly solvable problem in scattering theory is given by the square well,

$$V(r)=\begin{cases}-V_0, & r\leqslant d,\\ 0, & r>d,\end{cases} \quad \text{with} \quad V_0>0.$$

The regular solution of the eigenvalue problem $(h_\mu-k^2)\psi=0$ in the interval $0<r\leqslant d$ is given by

$$\phi(r;k,\mu)=\left(\frac{k}{q}\right)^{\mu+1/2}\phi_0(r;q,\mu), \quad \text{where} \quad q=\sqrt{k^2+V_0}.$$

For $r>d$ the eigensolutions of h_μ obey the free radial equation and hence the Jost solutions in this interval are just given by the free solutions,

$$\chi^{\pm}(r;k,\mu)=\chi_0^{\pm}(r;k,\mu), \quad r>d.$$

Making the usual requirement that all solutions are continuously differentiable we can evaluate the Wronskian $W(\chi^{\pm},\phi)$ at $r=d$ and obtain the following expression for the Jost function:

$$F(k,\mu)=\left(\frac{k}{q}\right)^{\mu-1/2}\left[\chi_0^{+}(d;k,\mu)\phi_0(d;q,\mu-1)-\frac{k}{q}\chi_0^{+}(d;k,\mu-1)\phi_0(d;q,\mu)\right].$$

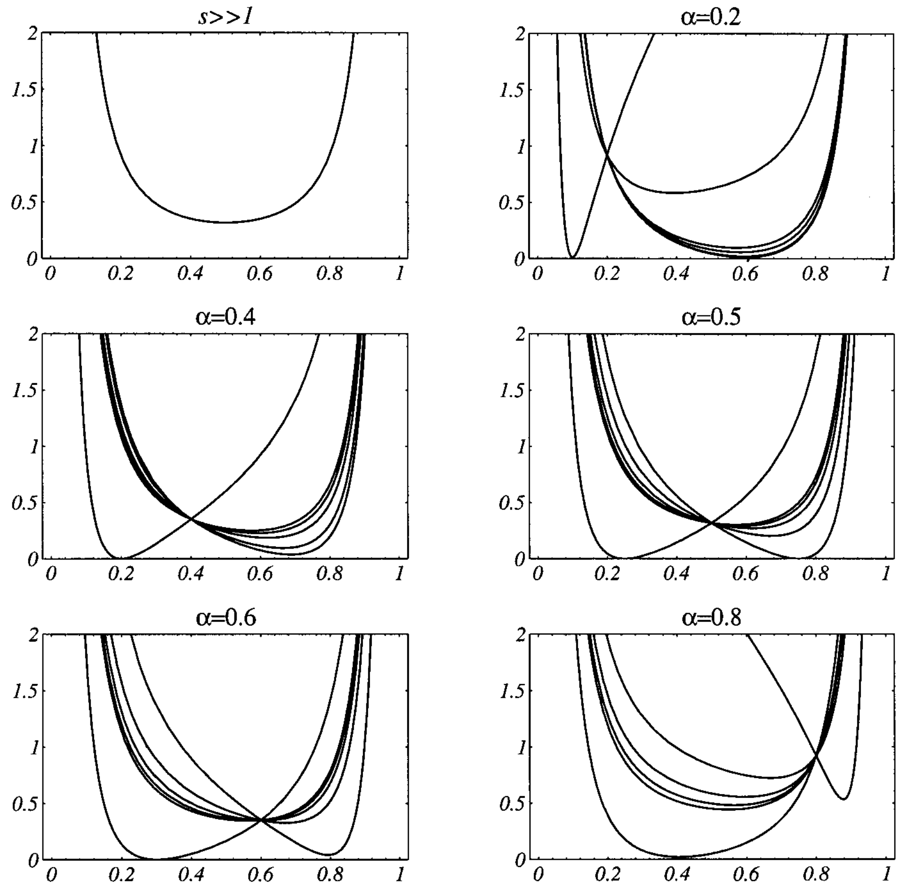

FIG. 3. Cross-sections for contact interaction 2.

Here, we have used the common differentiation rules for Bessel and Hankel functions (see, e.g., Refs. 34, 35). With the help of the Jost function we can now calculate the phase shift and the partial wave amplitude. We find

$$\tan \delta(k,\mu) = i\,\frac{F-\overline{F}}{F+\overline{F}} = \frac{qJ_\mu(kd)J_{\mu-1}(qd) - kJ_{\mu-1}(kd)J_\mu(qd)}{qY_\mu(kd)J_{\mu-1}(qd) - kY_{\mu-1}(kd)J_\mu(qd)},$$

$$f(k,\mu) = \frac{\overline{F}-F}{\sqrt{\pi ik}F} = \frac{-2}{\sqrt{\pi ik}}\,\frac{qJ_\mu(kd)J_{\mu-1}(qd) - kJ_{\mu-1}(kd)J_\mu(qd)}{qH_\mu^{(1)}(kd)J_{\mu-1}(pd) - kH_{\mu-1}^{(1)}(kd)J_\mu(qd)}.$$

Inserting the last expression into the formula (V.21) for the partial wave decomposition of the scattering amplitude we can calculate the differential cross-sections for various parameters α. Note that for small momenta we can make use of the small energy behavior of the partial wave amplitude, which is given by

$$k \to 0: \quad f(k,\mu) = \mathcal{O}(k^{2\mu-1}).$$

Thus, when calculating the scattering amplitude in the partial wave decomposition the partial wave amplitudes for higher angular momentum hardly contribute to the infinite sum in (V.21). According to our discussion in Sec. IV we have to add the Aharonov–Bohm scattering amplitude f_{AB} in order to obtain the differential cross-section [see (IV.8)]. A numerical example is displayed in Fig.

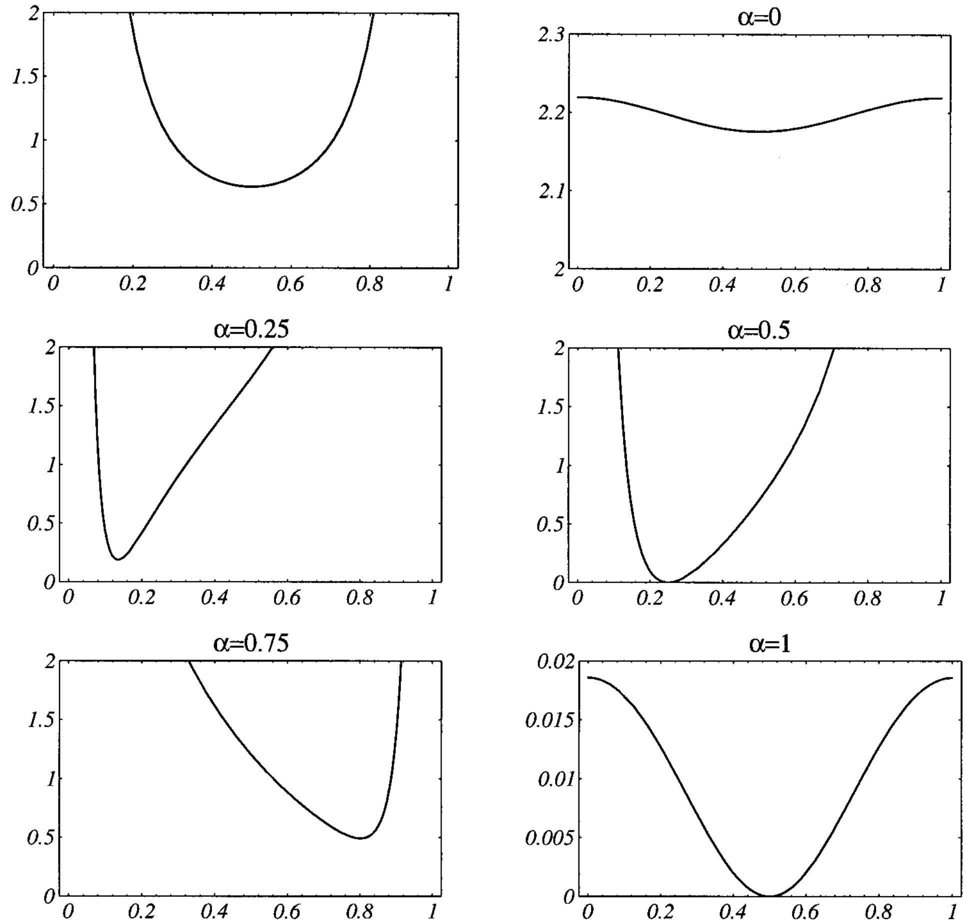

FIG. 4. Square well cross-sections.

4. For different values of the statistics parameter α one obtains very different angular distributions of the scattered particles. One observes the interpolation between the bosonic ($\alpha=0$) and the fermionic ($\alpha=1$) cross-section at energy $E=k^2=0.25$. Note that the singularities at the scattering angles $\theta=0$ and $\theta=\pi$ for $\alpha\notin\mathbb{Z}$ enter due to the long range nature of the statistics. In fact, according to (IV.8) they result from adding the Aharonov–Bohm amplitude given in (II.16). For fermions ($\alpha=1$) there is no scattering under an angle of $\theta=\pi/2$, as is well known. The figure for $\alpha=0.5$ shows that a similar effect can take place for fractional statistics.

According to our discussion in Sec. V the zeros of the Jost function F lie on the positive imaginary axis and determine the bound states. Thus, plotting the solutions of

$$F(ik,\mu)=0, \quad k>0,$$

we get so called (real) Regge trajectories (see, e.g., Ref. 47) displaying the μ dependence of the discrete spectrum of h_μ. Figure 5 provides a numerical example. In contrast to three-dimensional scattering theory here every point on the Regge trajectories has a specific physical meaning because the statistics parameter α interpolates between the integer-valued angular momentum channels.

VIII. CONCLUSION

We have demonstrated that it is possible to formulate nonrelativistic quantum scattering theories for two particles obeying anyon statistics. In particular, we have proven the existence of

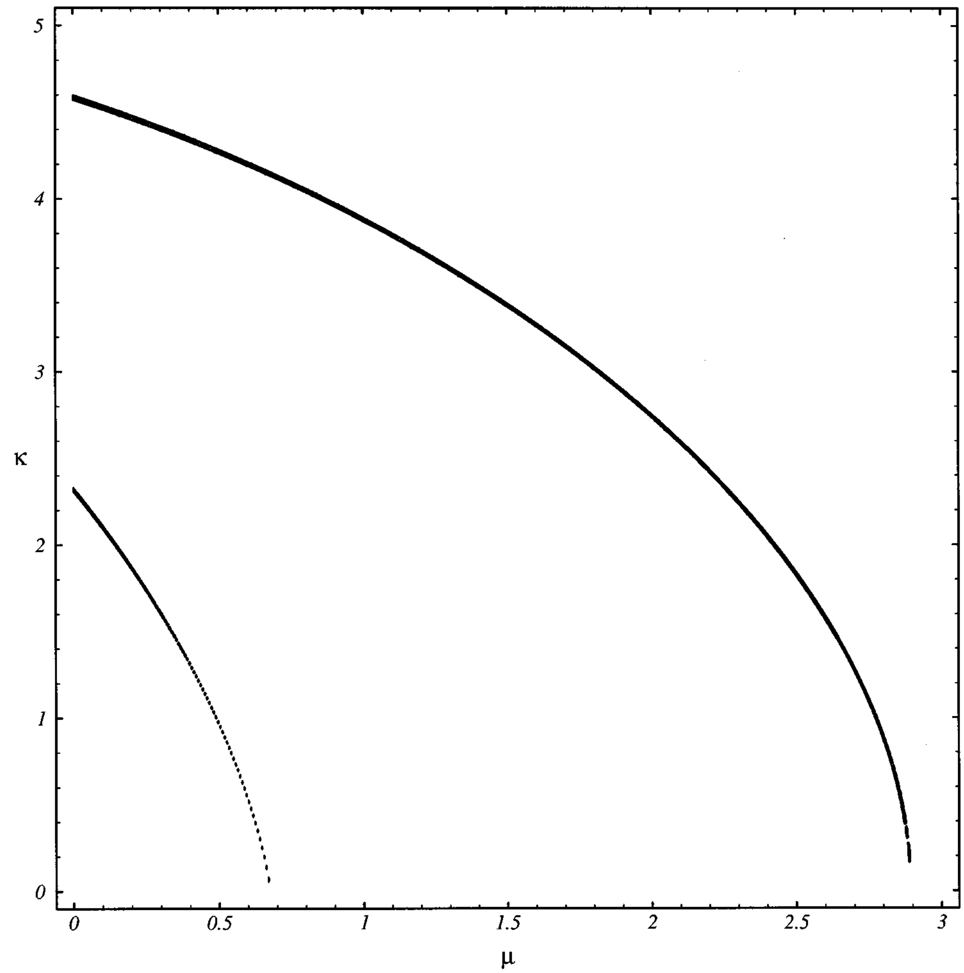

FIG. 5. Regge trajectories for the square well.

the wave operators under certain assumptions concerning the interaction. For spherically symmetric potentials we gave the criteria for completeness, i.e., unitarity of the *S*-matrix. These criteria turned out to be the same as in three-dimensional Schrödinger theory. It remains an open problem to show completeness in the general case in a way comparable to Enss' method. To generalize the latter, we believe it is crucial to obtain a better analytic control of the propagator. (There is an extensive literature on a similar problem, namely, to obtain a closed form of the Aharonov–Bohm propagator, see, e.g., Ref. 37.) We extended the notion of a differential cross-section to two dimensions and showed that in case of anyon statistics, the corresponding scattering amplitude consists of two parts. The first encodes the anyon statistics and resembles the Aharonov–Bohm amplitude, while the second is relevant to spectral analysis of the perturbed Hamiltonian. For spherically symmetric potentials we carried out this analysis by introducing Jost functions and showed how fractional statistics is equivalent to fractional angular momentum. We also showed that Levinson's theorem holds in the conventional case and gave the modifications necessary in the presence of a zero energy resonance. In the latter case we found that for positive angular momentum $\mu<1$ the statistics parameter—independently of the detailed form of the short range potential—can be determined from the scattering phase, namely [compare (V.26)]

$$\alpha \equiv \frac{1}{\pi}(\delta(k=0,\alpha)-\delta(k=\infty,\alpha)) \bmod \mathbb{Z}.$$

We then applied our results to the following two examples. First we considered the δ-potential, where we extracted information regarding the point spectrum of the perturbed Hamiltonian from its resolvent and showed that the above relation between the statistics parameter and scattering phase is also valid in this context [compare (VI.8)]. Second, we considered the square well potential, where we obtained similar information from the corresponding Jost function. In both situations we numerically evaluated the differential cross-section for noninteger values of the statistics parameter. It then became evident that fractional statistics is fundamentally different from Bose and Fermi statistics: The statistical gauge forces produce singularities of the differential cross-section in the forward direction, showing the long range nature of these forces. Due to this fact the angular dependence of the statistical interaction will invariably show up in the scattering data. In particular, this might cause a differential cross-section which is constant for bosons or fermions, but which becomes angular-dependent for intermediate statistics. Our discussion of the δ-potential provides an example.

The results presented in this paper may also be relevant to the investigation of bulk properties of anyon matter, via the relation between the virial coefficients and the scattering data.[20] This we intend to discuss in a forthcoming publication.

ACKNOWLEDGMENTS

The authors would like to thank Jens Mund for numerous discussions concerning the differential geometric formulation of fractional statistics and valuable comments on the results of this paper. This work was supported in part by DFG Sfb 288 ''Differentialgeometrie und Quantenphysik.''

APPENDIX A: THE BUNDLE THEORETIC FORMULATION

In Sec. II we used the physical picture of particles carrying flux-tubes in order to describe the anyon model. Anyons, however, are but a special case in a more general class of particles, called plektons, obeying neither Bose nor Fermi statistics. In this appendix we give a short review of plektonic systems, i.e., systems of n particles whose statistics are determined by a finite-dimensional unitary representation of the braid group B_n. The appropriate mathematical tool to describe such particles is the concept of vector bundles. This will, in particular, provide the Hilbert space and a canonical ''free'' dynamics described by a certain Laplace operator. Our presentation will closely follow the one given in Refs. 30, 31, where, however, emphasis was mainly given to relativistic formulations using momentum space considerations. For other references on the bundle theoretic formulation see, e.g., Refs. 32, 33. We continue to use $\mathbb{C}$ to describe the one-particle configuration space. Let $\mathbb{C}^{\times n}$ denote the n-fold product, viewed as the configuration space for n distinguishable particles. The set D_n in $\mathbb{C}^{\times n}$ is the set of all points $z=(z_1,z_2,\ldots,z_n)$ with $z_i=z_j$ for at least one pair of different indices. Any element σ of the permutation group S_n acts in an obvious way on $\mathbb{C}^{\times n}$ via $(z\sigma)_i=z_{\sigma(i)}$. We define the configuration space of n identical particles in two dimensions to be

$$^{n}\mathbb{C}:=(\mathbb{C}^{\times n}\setminus D_n)/S_n,$$

with points in it denoted by $\underline{z}$. Let $^{n}\tilde{\mathbb{C}}$ denote the universal covering space of $^{n}\mathbb{C}$ with points there being written as $\tilde{z}$ and let π: $^{n}\tilde{\mathbb{C}}\to {}^{n}\mathbb{C},\tilde{z}\mapsto\pi(\tilde{z})=\underline{z}$ be the associated projection mapping.

It is well-known that the fundamental group of $^{n}\mathbb{C}$, denoted by $\pi_1(^{n}\mathbb{C})$, is isomorphic to the braid group B_n and that any element $b\in B_n$ acts in a standard way from the right $\tilde{z}\mapsto\tilde{z}b$ on the manifold $^{n}\tilde{\mathbb{C}}$. Furthermore, the universal covering $^{n}\tilde{\mathbb{C}}\mapsto{}^{n}\mathbb{C}$ can be viewed as a principal fiber bundle over $^{n}\mathbb{C}$ with structure group $\pi_1(^{n}\mathbb{C})$ and fiber $\pi^{-1}(\underline{z})$ for any $\underline{z}\in{}^{n}\mathbb{C}$. Given any finite-dimensional unitary representation $b\mapsto\rho(b)$ of B_n in a Hilbert space F with scalar product $\langle\cdot|\cdot\rangle$ there is an associated Hermitian vector bundle. This vector bundle $\mathcal{F}$ with base space $^{n}\mathbb{C}$ is given by $^{n}\tilde{\mathbb{C}}\times_{\rho,B_n}F$, which by definition is the set of orbits in $^{n}\tilde{\mathbb{C}}\times F$ under the following action of B_n on this space:

$$b:(\tilde{z},f)\mapsto(\tilde{z}b,\rho(b^{-1})f),\quad \tilde{z}\in{}^{n}\tilde{\mathbb{C}},\ \ f\in F. \tag{A1}$$

$\mathcal{F}$ is a smooth fibered space with base ${}^{n}\mathbb{C}$ and fibers isomorphic to F. On the fiber over $\underline{z}$ there is a natural scalar product denoted by $\langle\cdot|\cdot\rangle_{\underline{z}}$. Furthermore, there is a canonical measure $d\mu(\underline{z})$ on ${}^{n}\mathbb{C}$ inherited from the Lebesgue measure on $\mathbb{C}^{\times n}$. This defines a scalar product on the space $\Gamma_c(\mathcal{F})$ of smooth sections in $\mathcal{F}$ with compact support via

$$\langle\psi|\phi\rangle:=\int_{{}^{n}\mathbb{C}}\langle\psi(\underline{z})|\phi(\underline{z})\rangle_{\underline{z}}d\mu(\underline{z}). \tag{A2}$$

By $\mathcal{L}^2(\mathcal{F})$ we denote the resulting Hilbert space completion. There is another Hilbert space equally well suited and canonically isomorphic to this space. Consider the set of maps Ψ: ${}^{n}\tilde{\mathbb{C}}\mapsto F$ which are smooth with $\pi(\mathrm{supp}\Psi)\subset{}^{n}\mathbb{C}$ being compact and which satisfy the equivariance property

$$\Psi(\tilde{z}b)=\rho(b^{-1})\Psi(\tilde{z}),\quad \forall\tilde{z}\in{}^{n}\tilde{\mathbb{C}},\ \ b\in B_n. \tag{A3}$$

For any two such functions Φ and Ψ we therefore have

$$\langle\Psi(\tilde{z}b)|\Phi(\tilde{z}b)\rangle=\langle\Psi(\tilde{z})|\Phi(\tilde{z})\rangle,\quad \forall\tilde{z}\in{}^{n}\tilde{\mathbb{C}},\ \ b\in B_n. \tag{A4}$$

Hence this expression depends on $\underline{z}=\pi(\tilde{z})$ only and so the integral over the base space makes sense and we may define a scalar product by

$$\langle\Psi|\Phi\rangle:=\int_{{}^{n}\mathbb{C}}\langle\Psi(\tilde{z})|\Phi(\tilde{z})\rangle d\mu(\underline{z}). \tag{A5}$$

The resulting Hilbert space obtained again by completion is denoted by $\mathcal{L}^2_{eq}({}^{n}\tilde{\mathbb{C}},F)$ and there is a canonical isomorphism between $\mathcal{L}^2(\mathcal{F})$ and $\mathcal{L}^2_{eq}({}^{n}\tilde{\mathbb{C}},F)$ (see, e.g., Refs. 30, 31). Furthermore the canonical (flat) Levi-Civita connection on $\mathbb{C}^{\times n}$ induces a (flat) connection on ${}^{n}\mathbb{C}$ which in turn defines a Hermitian connection $\boldsymbol{\nabla}$ on $\mathcal{F}$. The associated Bochner or generalized Laplacian (see, e.g., Ref. 59) $\Delta=\boldsymbol{\nabla}\circ\boldsymbol{\nabla}$ on $\mathcal{L}^2(\mathcal{F})$ is what defines a free Hamiltonian $H_0=-\Delta/2m$ for a system of n plektons, if $m>0$ is taken to be the mass of one particle.

When the unitary representation ρ of B_n is one-dimensional one speaks of anyons. Any such representation is obviously Abelian and can be shown to be of the form $b_k\mapsto\exp(i\alpha\pi)$ in terms of the generators $b_1,\dots,b_{n-1}\in B_n$ for a fixed $\alpha\in[0,2)$. This follows easily from the observation that all the b_k's are conjugate to each other. We denote the resulting line bundle by $\mathcal{F}_\alpha$. Another consequence of F being one-dimensional is that all anyonic line bundles are actually trivial. This observation was first made by Dowker,[60] based on Arnold's result that $H^2({}^{n}\mathbb{C},\mathbb{Z})=0$[61] and the classification theorem of Cartan, Kostant, Souriau and Isham. It was rediscovered by Gaberdiel[62] and is implicitly contained in Refs. 63 and 64. Another proof is given in Refs. 30, 31.

Let us now consider in more detail the Hilbert spaces constructed above when $n=2$. We first introduce relative coordinates by considering the following transformation of $\mathbb{C}^{\times 2}$:

$$(z_1,z_2)\mapsto(2^{-1}(z_1+z_1),z_1-z_2)=(z_{cen},z_{rel}). \tag{A6}$$

The transposition in $S_2\cong\mathbb{Z}_2$ obviously maps (z_{cen},z_{rel}) into $(z_{cen},-z_{rel})$. Therefore ${}^{2}\mathbb{C}$ is diffeomorphic to $\mathbb{C}\times\mathbb{C}^\star/\mathbb{Z}_2$. The first factor is the configuration space for the center of mass motion. The corresponding quantum mechanical discussion is analogous to ordinary multi-particle systems, since it is not affected by the statistics. Therefore we will concentrate on the second factor and the associated quantum mechanical description. In the same way as $\mathbb{R}$ is the universal covering space of S^1; the universal covering space of $\mathbb{C}^\star/\mathbb{Z}_2\cong\mathbb{R}^+\times S^1$ is given by $\mathbb{R}^+\times\mathbb{R}$ with the projection mapping

$$\pi:\mathbb{R}^+\times\mathbb{R}\to\mathbb{R}^+\times S^1,\quad \text{with}\quad (r,\theta)\mapsto(r,e^{2i\theta}). \tag{A7}$$

The action of B_2 on the universal covering is given in terms of its only generator b_1 as

$$(r,\theta)\mapsto(r,\theta)b_1:=(r,\theta+\pi). \tag{A8}$$

Choosing the representation $b_1\mapsto\exp(-i\pi\alpha)$ for a fixed $\alpha\in[0,1]$ we are prepared to construct the line bundle $\mathcal{F}_\alpha$ and the Hilbert space $\mathcal{L}^2(\mathcal{F}_\alpha)$, or equivalently $\mathcal{L}^2_{eq}({}^n\tilde{\mathbb{C}},F)$. The fact that the bundle $\mathcal{F}_\alpha$ is trivial is now reflected in the possibility to ''pull-down'' the theory from either of those $\mathcal{L}^2$ spaces onto the space of square integrable functions on ${}^n\mathbb{C}$ itself, denoted by $\mathcal{L}^2({}^n\mathbb{C})$. Explicitly there is the following unitary map from $\mathcal{L}^2({}^n\mathbb{C})$ onto $\mathcal{L}^2_{eq}({}^n\tilde{\mathbb{C}})$ given as

$$\psi\in\mathcal{L}^2({}^n\mathbb{C})\mapsto\Psi\in\mathcal{L}^2_{eq}({}^n\tilde{\mathbb{C}}),\quad\text{with}\quad\Psi(\tilde z)=e^{i\pi\alpha\cdot\theta(\tilde z)}\cdot\psi(\pi(\tilde z)),$$

where $\theta(\tilde z)$ is a real-valued, continuous function of $\tilde z$ induced by the polar angle. With the help of this unitary mapping, we can pull-down the Bochner Laplacean from $\mathcal{L}^2(\mathcal{F}_\alpha)$ to obtain an expression for the free Hamiltonian on $\mathcal{L}^2({}^n\mathbb{C})$. The result of this procedure is precisely the Hamiltonian we introduced in Sec. II on physical grounds.

Let it be noted that α was introduced as a parameter fixing the representation $B_2\ni b_1\mapsto\exp(-i\pi\alpha)$ and is hence determined up to an additive even integer, only. Therefore we lose no generality if we restrict α to the interval $[0, 2)$. In fact, we have $H_0(\alpha+2)=\exp(-i2\theta)H_0(\alpha)\exp(i2\theta)$, which is a gauge transformation reflecting the arbitrariness we have. On the other hand, time inversion causes α to change sign, i.e., if T denotes the anti-unitary operator of time inversion, we have $H_0(-\alpha)=TH_0(\alpha)T$. However, $H_0(-\alpha)$ is, in turn, equivalent to $H_0(2-\alpha)$ and consequently the anyon models with $\alpha\in[0,1]$ are connected to those with $\alpha\in[1,2]$ by time reversal. This justifies the restriction of α to the interval $[0, 1]$.

Furthermore, we remark that $\mathcal{L}^2(\mathbb{C}^\star/\mathbb{Z}_2,r\,dr\,d\theta)$ is unitary equivalent to the square integrable symmetric functions on the punctured plane $P_+\mathcal{L}^2(\mathbb{C}^\star,1/2r\,dr\,d\theta)$ and therefore the case $\alpha=0$ corresponds to the bosonic case [since $H_0(\alpha=0)=-\Delta$]. If $\alpha=1$ our model is equivalent to $H_0(\alpha=1)$ acting on $P_+\mathcal{L}^2(\mathbb{C}^\star)$, which is, in turn, unitary equivalent to $-\Delta$ on $P_-\mathcal{L}^2(\mathbb{C}^\star,1/2r\,dr\,d\theta)$, the Hilbert space of square integrable, anti-symmetric functions on the punctured plane. Hence we call anyons obeying statistics corresponding to $\alpha=1$ fermions.

APPENDIX B: SOME ESTIMATES

This appendix is devoted to a proof of the lemmas III.2 and III.3.

Proof of lemma III.2: For the moment let $0<\epsilon<1$ be arbitrary. Since $|e^{ix}-1|\le\min(2,|x|)$, we have $|e^{ix}-1|<2|x|^\epsilon$ for all $x\in\mathbb{R}$. This gives

$$|I_\alpha(\rho,\chi)-I_\alpha(0,\chi)|\le\left|\int_{-\infty}^{+\infty}dy(e^{i\rho\cosh y}-1)\,\frac{e^{-\alpha y}}{1-e^{-2y-2i\chi}}\right|$$

$$\le2\rho^\epsilon\int_0^\infty dy\,\cosh^\epsilon y\left|\frac{e^{-\alpha y}}{1-e^{-2y-2i\chi}}+\frac{e^{\alpha y}}{1-e^{2y-2i\chi}}\right|.$$

To proceed further, we split the integral into a part from 0 to 1 and the rest. This gives

$$|I_\alpha(\rho,\chi)-I_\alpha(0,\chi)|\le2\rho^\epsilon(\cosh(1)\cdot I_1(\chi)+I_2(\chi)).$$

For I_1 we have an expression involving the sum of three integrals $I_{1,1}$, $I_{1,2}$ and $I_{1,3}$, given by

$$I_1(\chi)=I_{1,1}(\chi)+I_{1,2}(\chi)+I_{1,3}(\chi)=\int_0^1dy\left|\frac{e^{-\alpha y}-1}{1-e^{-2y-2i\chi}}\right|+\int_0^1dy\left|\frac{e^{\alpha y}-1}{1-e^{2y-2i\chi}}\right|$$

$$+\int_0^1dy\left|\frac{1}{1-e^{-2y-2i\chi}}+\frac{1}{1-e^{2y-2i\chi}}\right|.$$

For I_2 we have the expression

$$I_2(\chi)=\int_1^\infty dy\,\cosh^\epsilon y\left(\left|\frac{e^{-\alpha y}}{1-e^{-2y-2i\chi}}\right|+\left|\frac{e^{\alpha y}}{1-e^{2y-2i\chi}}\right|\right).$$

The aim is to estimate these quantities for $\chi\in[-\pi,+\pi]$. To estimate $I_{1,1}(\chi)$ we use the following estimates:

$$2|1-e^{-\alpha y}|\leqslant 2y,\quad \text{for } y>0 \text{ and } \alpha\in(0,1); \tag{B1}$$

$$|1-e^{-2y-2i\chi}|\geqslant|1-e^{-2y}|,\quad \text{for } y\in\mathbb{R}. \tag{B2}$$

This gives

$$I_{1,1}(\chi)\leqslant\int_0^1 dy\,\frac{\alpha y}{1-e^{-2y}}<C.$$

$I_{1,2}(\chi)$ is estimated similarly, if one replaces the estimate (B1) by

$$0\leqslant e^{\alpha y}-1\leqslant 2e^2 y,\quad \text{for } 0\leqslant y\leqslant 1 \text{ and } \alpha\in(0,1).$$

To estimate $I_{1,3}(\chi)$ for $\chi\in[-\pi,+\pi]$ we first observe that for $\min(|\chi|,\pi-\chi,\pi+\chi)\geqslant\pi/4$ it is obviously bounded. Hence, it suffices to estimate the three remaining cases $|\chi|\leqslant\pi/4$, $3/4\pi\leqslant\chi\leqslant\pi$ or $-\pi\leqslant\chi\leqslant-3/4\pi$. We only consider the first case, since the other two cases may be discussed with similar arguments. So let $|\chi|\leqslant 1/4\pi$ and add and subtract $(\pm 2y+2i\chi)^{-1}$ in the integrand of $I_{1,3}(\chi)$. This way we can estimate $I_{1,3}$ by a sum of three integrals, which we will denote by $\hat{I}_i$,

$$I_{1,3}(\chi)\leqslant\hat{I}_1(\chi)+\hat{I}_2(\chi)+\hat{I}_3(\chi)=\int_0^1 dy\left|\frac{1}{1-e^{-2y-2i\chi}}-\frac{1}{2y+2i\chi}\right|+\int_0^1 dy\left|\frac{1}{1-e^{2y-2i\chi}}-\frac{1}{-2y+2i\chi}\right|+\int_0^1 dy\left|\frac{1}{2y+2i\chi}+\frac{1}{-2y+2i\chi}\right|.$$

To estimate $\hat{I}_1(\chi)$ and $\hat{I}_2(\chi)$ we note that the function

$$G(\zeta)=\frac{1}{1-e^{-\zeta}}-\frac{1}{\zeta},$$

is obviously analytic in $\{\zeta\in\mathbb{C}\,|\,|\mathrm{Im}(\zeta)|\leqslant\pi/2\}$, except possibly at the origin $\zeta=0$. However, $G(\zeta=0)=1$ and hence by Riemann's theorem, $G(\zeta)$ is an analytic function in that domain. In particular, $G(\zeta)$ is bounded on every compact subset, and this gives the boundedness of $\hat{I}_1(\chi)$ and $\hat{I}_2(\chi)$ for $\chi\in[-1/4\pi,1/4\pi]$. $\hat{I}_3(\chi)$ can be estimated as follows:

$$\hat{I}_3(\chi)\leqslant\int_0^1 dy\,\frac{|\chi|}{\chi^2+y^2}\leqslant\int_0^\infty d\eta\,\frac{1}{1+\eta^2}<\infty.$$

This concludes the estimate for $I_1(\chi)$ and it remains to estimate $I_2(\chi)$. By (B2) we have

$$I_2(\chi)\leqslant\int_1^\infty dy\,2e^{\epsilon y}\left(\frac{e^{-\alpha y}}{1-e^{-2y}}+\frac{e^{\alpha y}}{e^{2y}-1}\right).$$

Now this integral is finite whenever $\epsilon<\min(\alpha,1-\alpha)$. □

Proof of lemma III.3: We will consider the third partial derivatives w.r.t. E only, since this is actually the case we will need and since the other cases may be discussed similarly. By interchanging integration and differentiation we formally have

$$\frac{\partial^n}{\partial E^n} v_\alpha^{m,m'}(E,E') = \int_{\mathbb{H}\setminus\{0\}} \frac{\partial^n}{\partial E^n} [\overline{\phi_{\alpha;m,E}(z)} V(z)^2 \phi_{\alpha;m',E'}(z)] d\mu(z), \tag{B3}$$

for $n=1,2,3$. This is permitted provided the integrand is a measurable function in z, E and E', bounded in the sense of the modulus by a function which is in $\mathcal{L}^1$ w.r.t. z and with uniform bounds in E and E' in compact sets in $(0, \infty)$. To prove this claim it obviously suffices to replace $\partial/\partial E$ by $\partial/\partial k$ with $E=k^2$ in (B3). By the explicit form of $\phi_{\alpha;m,E}$ given in Sec. II, for fixed α, m and m' we therefore have to estimate the product

$$|\partial_k^n J_\mu(kr)| \cdot |J_\mu(k'r)| \cdot |V(r,\theta)|^2, \quad \text{for } n=1,2,3.$$

Using the formula to be found in Refs. 34 or 35,

$$\frac{d}{dk} J_\sigma(kr) = \frac{r}{2}(J_{\sigma-1}(kr) - J_{\sigma+1}(kr)), \tag{B4}$$

it follows by iteration that it suffices to estimate

$$r^n |J_{\mu+l}(kr)| \cdot |J_{\mu'}(k'r)| \cdot |V(r,\theta)|^2, \tag{B5}$$

for $-n \leqslant l \leqslant n$ with $0 \leqslant n \leqslant 3$. Let the compact set in question be given by $E_1 \leqslant E$, $E' \leqslant E_2$ such that $k_1 \leqslant k$, $k' \leqslant k_2$. Using the well known estimates

$$(a) \quad |J_\sigma(kr)| \leqslant C(\sigma), \quad \text{for } |kr| \leqslant 1,$$

$$(b) \quad |J_\sigma(kr)| \leqslant C(\sigma)(kr)^{-1/2}, \quad \text{for } |kr| \geqslant 1,$$

we see that (B5) with $k_1 \leqslant k$, $k' \leqslant k_2$ is bounded by

$$C(\mu,\mu') \cdot r^n |V(r,\theta)|^2 \leqslant C(\mu,\mu') \cdot \max(1,k_2^{-1})^3 |V(r,\theta)|^2,$$

for $r \leqslant k_2^{-1}$ and by

$$C(\mu,\mu') \cdot r^{n-1} k_1^{-1} |V(r,\theta)|^2 \leqslant C(\mu,\mu') \cdot \max(k_1^2, k_1^{-1}) \cdot r^2 |V(r,\theta)|^2,$$

for $r \geqslant k_1^{-1}$. Obviously (B5) is bounded for r in the interval (k_2^{-1}, k_1^{-1}) uniformly for $k_1 \leqslant k$, $k' \leqslant k_2$. This proves the claim and by our previous remark this concludes the proof. □

APPENDIX C: BOUNDS FOR JOST FUNCTIONS

In this section we will prove the estimates (V.11b), (V.17), (V.11a) and (V.16) for the functions ϕ and $\chi^\pm$ and their power series coefficients ϕ^n and $\chi^{\pm,n}$. Also, we will show that the analyticity properties for ϕ and $\chi^\pm$ also extend to their derivatives with respect to r. We start with two preparations. Observe first that (V.11b) and (V.17) hold when $V=0$, i.e., for $\phi=\phi_0$ and $\chi^\pm=\chi_0^\pm$, respectively. Indeed, for ϕ_0 (V.11b) follows from the relations (see, e.g., Refs. 34, 35)

$$J_\mu(z) \xrightarrow[z\to 0]{} C(\mu) z^\mu \quad \text{and} \quad J_\mu(z) \xrightarrow[|z|\to\infty]{} \frac{\cos\left(z - \frac{\pi}{2}(2\mu+1)\right)}{\sqrt{\frac{\pi}{2} z}}.$$

For $\chi_0^{\pm}$ the equation (V.17) follows from the following asymptotic behavior:

$$H_\mu^{\pm}(z) \xrightarrow[z\to 0]{} C(\mu) z^\mu \quad \text{and} \quad H_\mu^{\pm}(z) \xrightarrow[|z|\to\infty]{} \frac{e^{\pm i(z-(\pi/2)\mu-1/4)}}{\sqrt{\frac{\pi}{2}z}}.$$

Second, for g_0 [see (V.7)] one establishes analogously

$$0 \leqslant r' \leqslant r: \ |g_0(r,r';k,\mu)| \leqslant C(\mu) e^{|\mathrm{Im}(k)|(r-r')} \left(\frac{r}{1+|k|r}\right)^{\mu+1/2} \left(\frac{r'}{1+|k|r'}\right)^{-\mu+1/2}, \tag{C1}$$

$$0 \leqslant r \leqslant r': \ |g_0(r,r';k,\mu)| \leqslant C(\mu) e^{|\mathrm{Im}(k)|r-\mathrm{Im}(kr')} \left(\frac{r}{1+|k|r}\right)^{\mu+1/2} \left(\frac{r'}{1+|k|r'}\right)^{-\mu+1/2}. \tag{C2}$$

Now (V.11a) is proved by induction on n as follows. Define A_n by

$$|\phi_n(r;k,\mu)| = e^{|\mathrm{Im}(k)|r} \left(\frac{r}{1+|k|r}\right)^{\mu+1/2} A_n(r;k,\mu),$$

and it suffices to show that

$$A_n(r;k,\mu) \leqslant |k|^{\mu+1/2} \frac{C(\mu)^{n+1}}{n!} \left[\int_0^r dr' \frac{r'|V(r')|}{1+|k|r'}\right]^n,$$

where $C(\mu)$ is the maximum of the $C(\mu)$ in (V.11b) and the one in (C1). By the preparatory remarks (V.11a) holds for $n=0$, i.e., for $\phi_0 = \phi_{n=0}$. To perform the induction step, by construction we have

$$\phi_n(r;k,\mu) = \int_0^r dr' g_0(r,r';k,\mu) V(r') \phi_{n-1}(r';k,\mu),$$

which gives the inequality

$$A_n(r;k,\mu) \leqslant C(\mu) \int_0^r dr' \frac{r'|V(r')|}{1+|k|r'} A_{n-1}(r';k,\mu),$$

from which the induction step follows easily. This completes the proof of (V.11a) and (V.11b) follows from it by summation.

The proof of the bounds (V.16) and (V.17) is similar and we will consider the case χ^- only. Now write

$$|\chi_n^-(r;k,\mu)| = e^{\mathrm{Im}(kr)} \left(\frac{r}{1+|k|r}\right)^{-\mu+1/2} B_n(r;k,\mu),$$

such that it suffices to prove

$$B_n(r;k,\mu) \leqslant \frac{C(\mu)^{n+1}}{n!} |k|^{-\mu+1/2} \left(\int_r^\infty dr' \frac{r'|V(r')|}{1+|k|r'}\right)^n.$$

By the preparatory remarks, this inequality is valid for $n=0$. To make the induction step, we note that by construction

$$\chi_n^-(r;k,\mu)=\int_r^\infty dr' g_0(r,r';k,\mu)\chi_{n-1}^-(r';k,\mu).$$

Hence one has the estimate

$$B_n(r;k,\mu)\leqslant C(\mu)\int_r^\infty dr' \frac{r'|V(r')|}{1+|k|r'} B_{n-1}(r';k,\mu),$$

from which the induction step easily follows. This completes the proof of (V.16) and (V.17) follows by summation.

APPENDIX D: THE RESOLVENT OF THE δ-POTENTIAL

Proof of theorem VI.2: For any $\eta\in\mathcal{L}^2(\mathbb{R}^+)$ define

$$\psi:=\frac{1}{h_{0,\alpha}(s)-k^2}\,\eta,$$

using the integral kernel given by (VI.2). It is easily seen that ψ is well-defined. In order to prove the theorem we shall show that (VI.2) formally defines a resolvent, i.e., we have

$$\left(-\frac{\partial^2}{\partial r}-\frac{\alpha^2-\frac{1}{4}}{r^2}\right)\psi-k^2\psi=\eta. \tag{D1}$$

That this ''resolvent property'' is formally satisfied can easily be verified with the aid of the well-known formulas for the first derivatives of Bessel functions. In particular, one uses the relations

$$C'_\alpha=\frac{\alpha}{z}C_\alpha-C_{\alpha+1},\quad C'_{\alpha+1}=C_\alpha-\frac{\alpha+1}{z}C_{\alpha+1},\quad \text{for } C_\alpha=J_\alpha \text{ or } H_\alpha.$$

Furthermore, we shall show that ψ satisfies the boundary condition in (VI.1), i.e.,

$$\lim_{r\downarrow 0} W(\psi_{\alpha,s},\psi)(r)=0. \tag{D2}$$

For convenience we introduce the following notation:

$$I_<(r):=\int_0^r dr'\,\sqrt{r'}J_\alpha(kr')\,\eta(r'),\quad I:=\int_0^\infty dr'\,\sqrt{r'}H_\alpha^{(1)}(kr')\,\eta(r'),$$

$$I_>(r):=\int_r^\infty dr'\,\sqrt{r'}H_\alpha^{(1)}(kr')\,\eta(r').$$

ψ can now be cast into the form

$$\psi(r)=\sqrt{r}\left[\frac{i\pi}{2}H_\alpha^{(1)}(kr)I_<(r)+\frac{i\pi}{2}J_\alpha(kr)I_>(r)-A(k,\alpha;s)H^{(1)}(kr)I\right]. \tag{D3}$$

To verify the boundary condition (D2), we note that

$$\psi_{\alpha,s}(r)=\frac{1}{2\alpha}\cdot r^{1/2-\alpha}+\tilde{s}\cdot r^{1/2+\alpha},\quad \text{with}\quad \tilde{s}=s-\frac{1}{2\alpha},$$

giving

$$\psi'_{\alpha,s}(r)=\left(\frac{1}{4\alpha}-\frac{1}{2}\right)\cdot r^{1/2-\alpha}+\left(\frac{\tilde{s}}{2}+\tilde{s}\alpha\right)r^{-1/2+\alpha},$$

and hence

$$W(\psi_{\alpha,s},\psi)(r)=B'(r)\left(\frac{1}{2\alpha}\cdot r^{1-\alpha}+\tilde{s}r^{1+\alpha}\right)+B(r)\left(\frac{1}{2}\cdot r^{-\alpha}-\alpha\tilde{s}r^{\alpha}\right),$$

where $B(r)$ denotes the quantity in the square brackets in (D3). In order to take the limit, we remark that $z^{-\alpha}J_\alpha(z)$ is an analytic function for any value of $\alpha\in\mathbb{R}$. From its power series expansion at $z=0$ (see, e.g., Ref. 35) one obtains the following asymptotic relations when $\alpha>0$:

$$\lim_{r\downarrow 0} r^{-\alpha}J_\alpha(kr)=\iota_\alpha(k):=\left(\frac{k}{2}\right)^{\alpha}\frac{1}{\Gamma(1+\alpha)},$$

$$\lim_{r\downarrow 0} r^{\alpha}Y_\alpha(kr)=\gamma_\alpha(k):=-\left(\frac{2}{k}\right)^{\alpha}\frac{1}{\sin\pi\alpha}\,\frac{1}{\Gamma(1-\alpha)},$$

$$\lim_{r\downarrow 0} r^{\alpha}Y_{-\alpha}(kr)=\gamma_{-\alpha}(k):=-\left(\frac{k}{2}\right)^{-\alpha}\frac{1}{\tan\pi\alpha}\,\frac{1}{\Gamma(1-\alpha)}.$$

For the definition of the integral kernel (VI.2) we restricted the Bessel functions to the sheet given by $0<\arg(k)<\pi$. Since the irrational powers of k in the above expressions stem from the asymptotic behavior of the Bessel functions, they have to be evaluated on the same sheet. Bearing this in mind, we are prepared to take the limit $r\to 0^+$, which will make most of the terms in (D3) disappear. Since $|I_<(r)|\sim C\cdot r^{\alpha+1}$ and $I_>\sim I$ for $r\ll 1$, the remaining terms are

$$\lim_{r\downarrow 0} W(\psi_{\alpha,s},\psi)(r)=I\cdot\lim_{r\downarrow 0}\left[\frac{i\pi}{2}r^{-\alpha}J_\alpha(kr)+\tilde{s}\cdot k\cdot A(k,\alpha;s)r^{1+\alpha}H^{(1)}_{\alpha+1}(kr)+\frac{k}{2\alpha}A(k,\alpha;s)r^{1-\alpha}\left(H^{(1)}_{\alpha+1}(kr)-\frac{2\alpha}{kr}H^{(1)}_{\alpha}(kr)\right)\right].$$

Now, by a standard theorem for Bessel functions, we have $H^{(1)}_{\alpha+1}(kr)-(2\alpha/kr)H^{(1)}_\alpha(kr)=-H^{(1)}_{\alpha-1}(kr)$. The boundary condition (D2) therefore implies

$$A(k,\alpha;s)=\frac{i\pi}{2}\iota_\alpha(k)\left[\iota_\alpha(k)-2i\tilde{s}\alpha\cdot\gamma_\alpha+\frac{ik}{2\alpha}\gamma_{\alpha-1}\right]^{-1}.$$

Substituting $\tilde{s}=s-1/2\alpha$ and the expressions for ι_α, γ_α and $\gamma_{\alpha-1}$ we see that $A(k,\alpha;s)$ is given by (VI.3), thus completing the proof of theorem (VI.2). □

[1] A. Lerda, *Anyons*, Vol. m14 of Lecture Notes in Physics (Springer-Verlag, Berlin, 1992).
[2] C. Montonen, ''The many anyon problem,'' in hep th/9502071, Lectures Presented at the VI Mexican School of Particles and Fields, Mexican School, October, 1994, p. 225.
[3] J. M. Leinaas and J. Myrheim, Nuovo Cimento B **37**, 1 (1977).
[4] F. Wilczek, Phys. Rev. Lett. **49**, 957 (1982).
[5] G. Goldin, R. Menikoff, and D. H. Sharp, J. Math. Phys. **22**, 1664 (1981).
[6] G. Goldin and D. H. Sharp, Phys. Rev. D **28**, 830 (1983).
[7] G. Goldin, R. Menikoff, and D. H. Sharp, Phys. Rev. Lett. **54**, 603 (1985).
[8] F. Wilczek, Phys. Rev. Lett. **48**, 1144 (1982).
[9] D. P. Arovas, J. R. Schrieffer, F. Wilczek, and A. Zee, Nucl. Phys. B **251**, 117 (1985).
[10] T. H. Hansson, M. Rocek, I. Zahed, and S. C. Zhang, Phys. Lett. B **214**, 475 (1988).

[11] R. Jackiw and S.-Y. Pi, Phys. Rev. D **42**, 3500 (1990).
[12] K. Fredenhagen, K. H. Rehren, and B. Schroer, Commun. Math. Phys. **125**, 201 (1989).
[13] K. Fredenhagen, K. H. Rehren, and B. Schroer, Rev. Math. Phys. **4**, 113 (1992).
[14] A. Tsuchiya and Y. Kanie, Adv. Stud. Pure Math. **16**, 297 (1988).
[15] J. Fröhlich and P. A. Marchetti, Commun. Math. Phys. **121**, 177 (1989).
[16] R. E. Prange and S. M. Girvin, in *The Quantum Hall Effect* (Springer-Verlag, New York, 1987).
[17] M. Stone, *The Quantum Hall Effect* (World Scientific, Singapore, 1992).
[18] F. Wilczek, *Fractional Statistics and Anyon Superconductivity* (World Scientific, Singapore, 1990).
[19] J. D. Lykken, J. Sonnenschein, and N. Weiss, Int. J. Mod. Phys. A **6**, 5155 (1991).
[20] E. Beth and G. E. Uhlenbeck, Physica (The Hague) **4**, 915 (1937).
[21] Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).
[22] S. N. M. Ruijsenaars, Ann. Phys. (N.Y.) **146**, 1 (1983).
[23] C. Kim, Phys. Lett. B **431**, 374 (1998).
[24] F. Wilczek and Y. S. Wu, Phys. Rev. Lett. **65**, 13 (1990).
[25] M. Bucher, Nucl. Phys. B **350**, 163 (1991).
[26] H. K. Lo and J. Preskill, Phys. Rev. D **48**, 4821 (1993).
[27] S. Kuroda, J. Math. Phys. **3**, 933 (1962).
[28] C. Korff, ''Nichtrelativistische Streutheorie für das Zweiteilchensystem mit Zopfgruppenstatistik,'' Diploma thesis, Freie Universität Berlin, 1997.
[29] G. Lang, ''Anyonische Quantenmechanik: Streutheorie und gebundene Zustände im Zweiteilchenfall,'' Diploma thesis, Freie Universität Berlin, 1997.
[30] J. Mund, ''Quantenmechanik von nichtrelativistischen Teilchen mit Zopfgruppenstatistik,'' Diploma thesis, Freie Universität Berlin, 1992.
[31] J. Mund and R. Schrader, ''Hilbert spaces for nonrelativistic and relativistic 'free' plektons (particles with braid group statistics),'' in Advances in Dynamical Systems and Quantum Systems, edited by S. Albeverio, R. Figari, E. Orlandi, and A. Teta, (World Scientific, Singapore, 1995), p. 235.
[32] P. A. Horvathy, G. Morandi, and E. C. G. Surdarshan, Nuovo Cimento D **11**, 201 (1989).
[33] T. D. Imbo, C. Shah Imbo, and E. C. G. Surdarshan, Phys. Lett. B **234**, 103 (1990).
[34] G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge University Press, Cambridge, 1922).
[35] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Springer-Verlag, New York, 1966).
[36] M. Alvarez, Phys. Rev. A **54**, 1128 (1996).
[37] S. Gerbert and R. Jackiw, Commun. Math. Phys. **124**, 229 (1989).
[38] W. C. Henneberger, Phys. Rev. A **22**, 1383 (1980).
[39] S. Sakoda and M. Omote, J. Math. Phys. **38**, 716 (1997).
[40] V. Enss, Commun. Math. Phys. **61**, 285 (1978).
[41] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1979), Vol. 3.
[42] W. Bulla and F. Gesztesy, J. Math. Phys. **26**, 2520 (1985).
[43] R. Taylor, *The Quantum Theory of Nonrelativistic Collisions* (Krieger, Malabar, FL, 1983).
[44] D. B. Pearson, *Quantum Scattering and Spectral Theory, Techniques of Physics* (Academic, New York, 1988).
[45] R. Jost, Helv. Phys. Acta **20**, 256 (1947).
[46] R. G. Newton, J. Math. Phys. **1**, 319 (1960).
[47] R. G. Newton, *Scattering Theory of Waves and Particles, Texts and Monographs in Physics* (Springer-Verlag, New York, 1986).
[48] V. De Alfaro and T. Regge, *Potential Scattering* (North-Holland, Amsterdam, 1965).
[49] R. Jost and A. Pais, Phys. Rev. **82**, 840 (1951).
[50] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1979), Vol. 2.
[51] C. Manuel and R. Tarrach, Phys. Lett. B **268**, 222 (1991).
[52] M. Bourdeau and R. D. Sorkin, Phys. Rev. D **45**, 687 (1992).
[53] G. Amelino-Camelia, Phys. Lett. B **326**, 282 (1994).
[54] S. Albeverio, F. Gesztesy, R. Hoegh-Krohn, and H. Holden, *Solvable Models in Quantum Mechanics* (Springer-Verlag, Berlin, 1988).
[55] L. Dabrowski and P. Stovicek, J. Math. Phys. **39**, 47 (1998).
[56] R. Adami and A. Teta, Lett. Math. Phys. **43**, 43 (1998).
[57] W. Thirring, *Lehrbuch der Mathematischen Physik, Band 3* (Springer-Verlag Wien, New York, 1979).
[58] J. Weidmann, *Lineare Operatoren in Hilberträumen* (Teubner, Stuttgart, 1976).
[59] N. Berline, E. Getzler, and M. Vergne, *Heat Kernels and Dirac Operators, Grundlehren der mathematischen Wissenschaften 298* (Springer-Verlag, Berlin, 1992).
[60] J. S. Dowker, J. Phys. A **18**, 3521 (1985).
[61] V. I. Arnol'd, Trudi Moskovskogo Mathematicheskogo Obshchestva **21**, 27 (1970).
[62] M. Gaberdiel (private communication, 1993).
[63] Y. S. Wu, Phys. Rev. Lett. **53**, 111 (1984).
[64] J. Fröhlich and P. A. Marchetti, Nucl. Phys. B **356**, 533 (1991).

Poisson brackets of normal-ordered Wilson loops

C.-W. H. Lee and S. G. Rajeev
Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627

(Received 2 November 1998; accepted for publication 2 December 1998)

We formulate Yang–Mills theory in terms of the large-N limit, viewed as a classical limit, of gauge-invariant dynamical variables, which are closely related to Wilson loops, via deformation quantization. We obtain a Poisson algebra of these dynamical variables corresponding to normal-ordered quantum (at a finite value of $\hbar$) operators. Comparing with a Poisson algebra one of us introduced in the past for Weyl-ordered quantum operators, we find, using ideas closely related to topological graph theory, that these two Poisson algebras are, roughly speaking, the same. More precisely speaking, there exists an invertible Poisson morphism between them. © *1999 American Institute of Physics.* [S0022-2488(99)02204-5]

I. INTRODUCTION

Among the many different approaches to Yang–Mills theory, one of the most widely studied is the large-N limit.[1] The large-N limit can be formulated as a classical limit,[2] with a well-defined phase space and a Poisson bracket between dynamical variables which are functions of the phase space. The hallmark of Yang–Mills theory is the gauge invariance of physical observables, and it is natural for us to think that the dynamical variables should also be gauge-invariant functions. Next comes naturally this question: Is there a sensible Poisson bracket between these gauge-invariant functions? If so, this will be a major step towards the classical formulation of Yang–Mills theory in the large-N limit.

One of us, together with Turgut, introduced in a previous article[3] such a Poisson bracket. Consider a Yang–Mills theory with matter fields z^i, where different matter fields are distinguished by different values of the index i, in the adjoint representation. Such a theory can be obtained, for example, by dimensionally reducing a D-dimensional Yang–Mills theory to a two-dimensional one. The two color indices carried by the adjoint matter field can be regarded as matrix entries. In this sense, the adjoint matter fields are Hermitian matrices. Consider the trace of a product of these matrices. Under a gauge transformation characterized by a unitary matrix g, the adjoint matter fields are changed in the following manner:

$$z^i \to g z^i g^\dagger. \tag{1}$$

Hence the trace remains unchanged, and is thus a gauge-invariant function, and a dynamical variable of the theory. We call this gauge-invariant function a loop variable, as this was originally motivated from the study of Wilson loops.[4]

A convenient way to quantize such loop variables is via deformation quantization. (Deformation quantization was proposed by Flato, Lichnerowicz and Sternheimer.[5] See also Ref. 6. Reference 7 gives a pedagogical introduction. A more comprehensive list of references can be found in Ref. 8.) The essential idea is that the commutative product of these loop variables is deformed in such a way that when we multiply two loop variables, it is as if we are multiplying the two operators they represent. (We say that the loop variables are the symbols of these operators.) As there are different ways to order a product of operators, there are also different schemes of deformation quantization. In Ref. 3, the operators are Weyl-ordered. Then the Poisson bracket of two loop variables can be defined as the large-N limit of the commutator of them. We will review the precise definition of this Poisson bracket at the beginning of Sec. III. In a sense, we have obtained a classical limit not by setting $\hbar$ to 0 but by letting N go to infinity. This Poisson bracket dictates the classical dynamics of a system in which the dynamical variables are expressed in terms of these loop variables.

However, as most finite-$\hbar$ quantum theory are formulated in terms of normal-ordered operators, it should be interesting to find another Poisson bracket which corresponds to normal-ordered operators, i.e., the loop variables should be multiplied in such a way that it is as if we are multiplying normal-ordered operators. This is the goal in Sec. II.

This Poisson algebra is closely related to the Lie algebras we presented in previous papers,[9–11] though we derived those Lie algebras in a manner thoroughly independent of this Poisson algebra. We believe that the loop variables have a meaning in noncommutative geometry, and, in some sense, the Lie algebras are linear approximations of this Poisson algebra. We have not yet precisely identified the nature of this approximation, and this is a subject worthy of being pursued in the future. Nevertheless, at the end of Sec. II, we will indicate in a crude manner how the Poisson algebra can be truncated to obtain these Lie algebras.

The next interesting question which comes to mind is: what is the relationship between these two seemingly different Poisson algebras? It turns out that when there are only a finite number of distinct Hermitian matrices z^i, i.e., when i can take on a finite number of distinct values only, these two Poisson brackets are, roughly speaking, the same. More precisely speaking, there exists an invertible Poisson morphism between the two Poisson algebras. We are going to show the existence of this Poisson morphism in Sec. III. The astute reader will notice that many of the lemmas in the proof have simple interpretations in terms of topological graph theory. (For an introductory account on topological graph theory, see, e.g., Ref. 12. However, we will not use any of the results there because of the difference between the underlying topologies discussed in that reference and the topologies here.) Indeed, we will derive from first principles some properties of topological graphs which, we hope, will be of interest to topological graph theorists.

II. DEFORMATION QUANTIZATION

We are going to derive a Poisson algebra pertinent to gauge theory via deformation quantization in this section. Deformation quantization refers to the procedure of defining an algebra of smooth functions in such a way that when the functions are multiplied, it is as if we are multiplying suitably ordered operators these smooth functions represent. To be more specific, consider the set of all smooth functions on a one-dimensional complex Euclidean space. Let z be a coordinate of this one-dimensional space. Then we can associate a smooth function $f(z,\bar{z})$ on it with a Weyl-ordered operator in the way described by Chari and Pressley.[7] The way to associate $f(z,\bar{z})$ with a normal-ordered operator is similar. Indeed, the first step is to obtain the Fourier transform $\hat{f}(\xi,\eta)$ of $f(z,\bar{z})$ first,

$$\hat{f}(\xi,\eta)=\frac{1}{(2\pi)^2}\int dz\,d\bar{z}\,f(z,\bar{z})e^{-(i/\hbar)(\xi z+\eta\bar{z})}. \tag{2}$$

Here ξ and η are still complex variables and $\hbar$ is a quantization parameter. Then the associated normal-ordered operator $\Phi(f)$ is defined as

$$\Phi(f)=\int d\xi\,d\eta\,\hat{f}(\xi,\eta)e^{i\hbar\xi a^\dagger}e^{i\hbar\eta a}, \tag{3}$$

where a and $a^\dagger$ are the annihilation and creation operators satisfying $[a,a^\dagger]=1$, respectively. We then define a noncommutative associative product $*_\hbar$ such that

$$\Phi(f_1 *_\hbar f_2)=\Phi(f_1)\Phi(f_2). \tag{4}$$

Equation (4) is satisfied if this product is defined as follows:

$$f_1 *_\hbar f_2 \equiv e^{\hbar(\partial/\partial\bar{z})(\partial/\partial z')} f_1(z,\bar{z}) f_2(z',\bar{z'})|_{z=z',\bar{z}=\bar{z}'}. \tag{5}$$

Then this operation $*_\hbar$ is a deformation of the algebra of functions on a one-dimensional complex Euclidean space.

In the physical systems we are interested, the dynamical variables are represented by loop variables. Mathematically these loop variables are traces of $N\times N$ matrices. Thus we would like to generalize the above formulation of deformation quantization from ordinary complex variables to $N\times N$ matrices. Furthermore, physically each matrix corresponds to a state with a particular set of quantum numbers other than color (e.g., momentum). There are, of course, more than one possible set of quantum numbers and so we would also generalize the quantization scheme from a one-dimensional space to a multidimensional space. For the sake of simplicity, this dimension is still finite though in the actual physical context it should be infinite.

Having said this, let us generalize the formulation of deformation quantization to a system of bosons. Consider a complex Euclidean space of dimension $2\Lambda N^2$, where Λ is an arbitrary positive integer. Let z^i, where $i=-\Lambda, -\Lambda+1, \ldots, -1, 1, 2, \ldots$, or Λ, be a Hermitian $N\times N$ matrix. An entry of z^i is denoted by z_b^{ia}, a and b being the row and column indices, respectively. Denote z^{-i} by $\overline{z^i}$. A *normal-ordered loop variable* is a function of the form

$$\tilde{\phi}^I(z,\bar{z}) = \operatorname{Tr} z^{i_1}\cdots z^{i_m}. \tag{6}$$

Here I represents the sequence of nonzero integers $i_1, \ldots, i_m$ between $-\Lambda$ and Λ inclusive. $\tilde{\phi}^I(z,\bar{z})$ is gauge-invariant since it remains unchanged under the gauge transformation given by Eq. (1). Linear combinations of products of normal-ordered loop variables form a function space $\mathcal{N}$. Equation (5) can be generalized to

$$\tilde{\phi}^I * \tilde{\phi}^J(z,\bar{z}) = e^{\hbar\gamma^{\mu\nu}(\partial/\partial z_b^{\mu a})(\partial/\partial z'^{\nu b}_a)} \tilde{\phi}^I(z,\bar{z})\tilde{\phi}^J(z',\bar{z'})|_{z=z',z^\dagger=z'^\dagger} \tag{7}$$

with $\gamma^{\mu\nu}=0$ unless $\mu<0$ and $\nu>0$. In the limit $\hbar\to\infty$, Eq. (7) produces the ordinary Poisson bracket.

Let us derive from Eq. (7) a Poisson bracket for a finite value of $\hbar$. This is done by expanding Eq. (7) as a power series of $\hbar$. Indeed, we obtain

$$\tilde{\phi}^I *_\hbar \tilde{\phi}^J = \tilde{\phi}^I\tilde{\phi}^J + \sum_{r=1}^{\infty} \frac{\hbar^r}{r!} \gamma^{i_{\mu_1} j_{\nu_1}} \cdots \gamma^{i_{\mu_r} j_{\nu_r}} \frac{\partial^r \tilde{\phi}^I}{\partial z_{b_1}^{i_{\mu_1} a_1} \cdots z_{b_r}^{i_{\mu_r} a_r}} \frac{\partial^r \tilde{\phi}^J}{\partial z_{a_1}^{j_{\nu_1} b_1} \cdots z_{a_r}^{j_{\nu_r} b_r}}, \tag{8}$$

where $i_{\mu_1}, i_{\mu_2}, \ldots, i_{\mu_r} < 0$ and $j_{\nu_1}, j_{\nu_2}, \ldots, j_{\nu_r} > 0$. We can always bring the first set of indices to the order $\mu_1 < \mu_2 < \cdots < \mu_r$ by relabeling the indices. Then the set of indices $\nu_1, \nu_2, \ldots, \nu_r$ will be rearranged in one of all $r!$ possible permutations. We note that

$$\frac{\partial \tilde{\phi}^I}{\partial z_b^{ka}} = 0 \tag{9}$$

unless k is equal to one of the elements of the loop $(i_1, \ldots, i_m)$. If $k=i_\mu$ for some $\mu = 1, \ldots, m$, then

$$\frac{\partial \tilde{\phi}^I}{\partial z_b^{ka}} = [z^{i_{\mu+1}} z^{i_{\mu+2}} \cdots z^{i_m} z^{i_1} \cdots z^{i_{\mu-1}}]_a^b. \tag{10}$$

More generally, when $\mu_1 < \mu_2 < \cdots < \mu_r$,

$$\frac{\partial^r \tilde{\phi}^I}{\partial z^{i_{\mu_1} a_1}_{b_1} \partial z^{i_{\mu_2} a_2}_{b_2} \cdots \partial z^{i_{\mu_r} a_r}_{b_r}} = P^{b_1}_{a_2}(I(\mu_1,\mu_2)) P^{b_2}_{a_3}(I(\mu_2,\mu 3)) \cdots P^{b_r}_{a_1}(I(\mu_r,\mu_1)), \tag{11}$$

where

$$P^{b_1}_{a_2}(I(\mu_1,\mu_2)) = \begin{cases} [z^{i_{\mu_1+1}} z^{i_{\mu_1+2}} \cdots z^{i_{\mu_2-1}}]^{b_1}_{a_2} & \text{if } \mu_2 > \mu_1 \\ [z^{i_{\mu_1+1}} z^{i_{\mu_1+2}} \cdots z^{i_m} z^{i_1} \ldots z^{i_{\mu_2-1}}]^{b_1}_{a_2} & \text{if } \mu_2 < \mu_1 \end{cases} \tag{12}$$

and so on for the other P's. Hence, we can substitute Eq. (11) into Eq. (8) to get

$$\begin{aligned} \sum_{\{\sigma\}} \frac{\hbar^r}{r!} \gamma^{i_{\mu_1} j_{\nu_{\sigma(1)}}} \cdots \gamma^{i_{\mu_r} j_{\nu_{\sigma(r)}}} \\ \cdot P^{b_1}_{a_2}(I(\mu_1,\mu_2)) P^{b_2}_{a_3}(I(\mu_2,\mu_3)) \cdots \\ \cdot P^{b_r}_{a_1}(I(\mu_r,\mu_1)) P^{a_{\sigma(1)}}_{b_{\sigma(2)}}(J(\nu_{\sigma(1)},\nu_{\sigma(2)})) P^{a_{\sigma(2)}}_{b_{\sigma(3)}}(J(\nu_{\sigma(2)},\nu_{\sigma(3)})) \cdots \\ \cdot P^{a_{\sigma(r)}}_{b_{\sigma(1)}}(J(\nu_{\sigma(r)},\nu_{\sigma(1)})) \end{aligned} \tag{13}$$

for the rth order term. Here σ is any possible permutation of $\nu_1, \ldots, \nu_r$.

In the large-N limit, the term with the largest number of traces will dominate. This occurs if the ν indices are in decreasing order up to a cyclic permutation, e.g., $\nu_2 > \nu_3 > \cdots > \nu_r > \nu_1$, etc. Then to the first two orders in the large-N limit,

$$\begin{aligned} \tilde{\phi}^I *_\hbar \tilde{\phi}^J \simeq \tilde{\phi}^I \tilde{\phi}^J \\ + \sum_{r=1}^{\infty} \sum_{\substack{\mu_1 < \mu_2 < \cdots < \mu_r \\ (\nu_1 > \nu_2 > \cdots > \nu_r)}} \hbar^r \gamma^{i_{\mu_1} j_{\nu_1}} \cdots \gamma^{i_{\mu_r} j_{\nu_r}} \\ \cdot \tilde{\phi}^{I(\mu_1,\mu_2) J(\nu_2,\nu_1)} \tilde{\phi}^{I(\mu_2,\mu_3) J(\nu_3,\nu_2)} \ldots \tilde{\phi}^{I(\mu_r,\mu_1) J(\nu_1,\nu_r)}, \end{aligned} \tag{14}$$

where

$$\tilde{\phi}^{I(\mu_1,\mu_2) J(\nu_2,\nu_1)} = P^b_a(I(\mu_1,\mu_2)) P^a_b(J(\nu_2,\nu_1)). \tag{15}$$

To ensure that the large-N limit is well defined, we need to normalize the functions ϕ^I by some N-dependent factor. The normalization is such that the vacuum expectation value of ϕ^I remains finite as $N \to \infty$. Consider the vacuum state of the Hamiltonian $g_{ij} \operatorname{Tr} z^i \overline{z^j}$, where $i,j = 1, \ldots, \Lambda$. Then the vacuum expectation value of $z^{ia}_b z^{jc}_d$ is $\langle z^{ia}_b z^{jc}_d \rangle = \gamma^{ij} \delta^a_d \delta^c_b$. Thus the vacuum expectation value of the product of an odd number of z's will vanish whereas that of an even number of z's will be given by Wick's theorem. A short calculation reveals that the $\langle \tilde{\phi}^I \rangle$ for the ϕ^I defined in Eq. (6) with m even is of order $N^{(m/2)+1}$. This can further be shown to be independent of the particular form of the Hamiltonian. Consequently, we define the normalized functions

$$\phi^I = \frac{1}{N^{(m/2)+1}} \tilde{\phi}^I. \tag{16}$$

Combining Eqs. (14) and (15), we get

$$\phi^I *_\hbar \phi^J = \phi^I \phi^J + \frac{1}{N_c^2} \sum_{r=1}^{\infty} \sum_{\substack{\mu_1<\mu_2<\cdots<\mu_r \\ (\nu_1>\nu_2>\cdots>\nu_r)}} \hbar^r \gamma^{i_{\mu_1} j_{\nu_1}} \cdots \gamma^{i_{\mu_r} j_{\nu_r}} \cdot \phi^{I(\mu_1,\mu_2)J(\nu_2,\nu_1)} \phi^{I(\mu_2,\mu_3)J(\nu_3,\nu_2)} \cdots \phi^{I(\mu_r,\mu_1)J(\nu_1,\nu_r)} + O\left(\frac{1}{N^3}\right). \tag{17}$$

Let us define the Poisson bracket by

$$\{\phi^I, \phi^J\}_N \equiv \lim_{N\to\infty} N^2 (\phi^I * \phi^J - \phi^J * \phi^I). \tag{18}$$

We then finally obtain

$$\{\phi^I, \phi^J\}_N = a \sum_{r=1}^{\infty} \sum_{\substack{\mu_1<\mu_2<\cdots<\mu_r \\ (\nu_1>\nu_2>\cdots>\nu_r)}} \hbar^r \gamma^{i_{\mu_1} j_{\nu_1}} \cdots \gamma^{i_{\mu_r} j_{\nu_r}} \cdot \phi^{I(\mu_1,\mu_2)J(\nu_2,\nu_1)} \phi^{I(\mu_2,\mu_3)J(\nu_3,\nu_2)} \cdots \phi^{I(\mu_r,\mu_1)J(\nu_1,\nu_r)} - (I \leftrightarrow J). \tag{19}$$

We can visualize Eq. (19) by the diagrammatic representations in Fig. 1.

Equation (19) characterizes the Poisson algebra of loop variables corresponding to normal-ordered operators, and we call this the *normal-ordered Poisson algebra*. In comparison with the Poisson algebra found in a previous paper,[3] where the loop variables correspond to Weyl-ordered operators, we notice that the antisymmetric tensors ω^{ij} in the last equation of Ref. 3 are here replaced by γ^{ij}, which are nonzero only if $i<0$ and $j>0$. In addition, terms of order $\hbar^1, \hbar^3, \hbar^5, \ldots$, etc. vanish in the last equation of Ref. 3 but they are nonzero here in general.

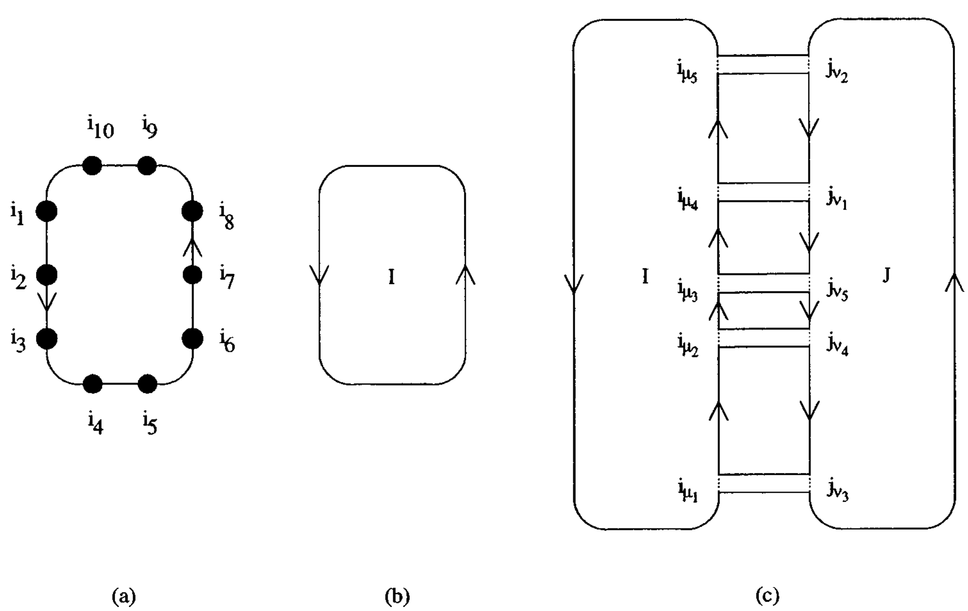

FIG. 1. (a) A typical loop variable ϕ^I. Each solid circle represents a z^i. Notice the cyclic symmetry of the figure. (b) A simplified diagrammatic representation of ϕ^I. We use the capital letter I to denote the whole sequence $i_1, i_2, \ldots, i_m$. (c) A typical term in $\{\phi^I, \phi^J\}_N$. This is a product of the loop variables $\phi^{I(\mu_1,\mu_2)J(\nu_4,\nu_3)}$, $\phi^{I(\mu_2,\mu_3)J(\nu_5,\nu_4)}$, $\phi^{I(\mu_3,\mu_4)J(\nu_1,\nu_5)}$, $\phi^{I(\mu_4,\mu_5)J(\nu_2,\nu_1)}$, and $\phi^{I(\mu_5,\mu_1)J(\nu_3,\nu_2)}$.

Nevertheless, these two Poisson algebras have a deep relationship — there is an invertible Poisson morphism between the Poisson algebra of Weyl-ordered operators and that of normal-ordered operators, whose proof will be given in the next section.

In previous papers, we defined and discussed a number of Lie algebras like the cyclix algebra,[9] the centrix algebra,[10] and the heterix algebra.[11] These Lie algebras arise from taking the planar large-N limit of gauge theory. We can actually think of them as various approximations of the Poisson algebra given by Eq. (19). For example, to get the centrix algebra from this Poisson algebra, we choose $\hbar=1$ and $\gamma^{ij}=\delta^{-i,j}$ and restrict ourselves to loop variables of the form $\sigma_J^I \equiv \phi^{IJ^*}=\mathrm{Tr}\, z^{i_1}z^{i_2}\cdots z^{i_{\#(I)}}\overline{z^{j_{\#(J)}}}\,\overline{z^{j_{\#(J)-1}}}\cdots\overline{z^{j_1}}$, where $\#(I)$ and $\#(J)$ are the numbers of integers in I and J, respectively, and all the indices $i_1, i_2, \ldots, i_{\#(I)}, j_1, j_2, \ldots,$ and $j_{\#(J)}$ are positive integers between 1 and Λ inclusive. (J^* is defined as the reverse sequence of J.) If we now compute the Poisson bracket between two loop variables of this form using Eq. (19), we should obtain

$$\{\sigma_J^I,\sigma_L^K\}_N=\sum_{r=1}^{\infty}\sum_{\substack{\mu_1>\mu_2>\cdots>\mu_r\\(\nu_1>\nu_2>\cdots>\nu_r)}}\phi^{J^*(\mu_1,\mu_2)K(\nu_2,\nu_1)}\cdot\phi^{J^*(\mu_2,\mu_3)K(\nu_3,\nu_2)}\cdots\phi^{J^*(\mu_{r-1},\mu_r)K(\nu_r,\nu_{r-1})}$$
$$\cdot\,\phi^{J^*(\mu_r,0)I(0,\#(I)+1)J^*(\#(J)+1,\mu_1)K(\nu_r,\#(K)+1)L^*(\#(L)+1,0)K(0,\nu_1)}-(I\leftrightarrow K,J\leftrightarrow L). \quad (20)$$

If we now retain only those terms in which $\mu_1, \mu_2, \ldots, \mu_r$ are consecutive integers in the reverse order, i.e., $\mu_2=\mu_1-1, \mu_3=\mu_2-1, \ldots,$ and $\mu_r=\mu_{r-1}-1$, and in which $\nu_1, \nu_2, \ldots, \nu_r$ are also consecutive integers in the reverse order, we will get precisely the Lie bracket of the centrix algebra. If we retain some more terms, we will obtain the heterix algebra. The cyclix algebra is obtained from the heterix algebra by identifying certain products of loop variables as a linear combination of single loop variables. We believe that the loop variables have a geometrical meaning in a noncommutative space, and thus there should be a geometrical meaning of these truncating approximations. We hope to understand the geometry better in the future.

III. A POISSON MORPHISM

We are going to show that there exists an invertible Poisson morphism between the Poisson algebra of Weyl-ordered loop variables described in Ref. 3 and the Poisson algebra of normal-ordered loop variables given in Eq. (19). It should be interesting for the reader to discern, with the help of the accompanying diagrams, the meanings of many of the following lemmas in topological graph theory.

Let us remind ourselves the definition of the Weyl-ordered Poisson algebra here. Consider 2Λ distinct $N\times N$ Hermitian matrices $\eta^{-\Lambda}, \eta^{-\Lambda+1}, \ldots, \eta^{-1}, \eta^1, \eta^2, \ldots,$ and η^Λ. A *Weyl-ordered loop variable* is a trace of an arbitrary sequence of these matrices $f^I=\mathrm{Tr}\,\eta^{i_1}\eta^{i_2}\cdots\eta^{i_m}$, where m is a positive integer called the *degree of* f^I, $i_k\in\{-M,-M+1,\ldots,-1,1,2,\ldots,M\}\forall k=1,2,\ldots,m,$ and I denotes the integer sequence $i_1,i_2,\ldots,i_m$. Linear combinations of products of Weyl loops form a function space $\mathcal{W}$. The Poisson bracket between two Weyl-ordered loop variables f^I and $f^J=\mathrm{Tr}\,\eta^{j_1}\eta^{j_2}\cdots\eta^{j_n}$, where n, j_k, and J have analogous definitions as m, i_k, and I, is given by the following formula:

$$\{f^I,f^J\}_W=2i\sum_{r=1,\mathrm{odd}}^{\infty}\sum_{\substack{\mu_1<\mu_2<\cdots<\mu_r\\(\nu_1>\nu_2>\cdots>\nu_r)}}\left(-\frac{i\hbar}{2}\right)^r\tilde{\omega}^{i_{\mu_1}j_{\nu_1}}\cdots\tilde{\omega}^{i_{\mu_r}j_{\nu_r}}$$
$$\cdot\, f^{I(\mu_1,\mu_2)J(\nu_2,\nu_1)}f^{I(\mu_2,\mu_3)J(\nu_3,\nu_2)}\cdots f^{I(\mu_r,\mu_1)J(\nu_1,\nu_r)}. \quad (21)$$

In this equation, for every value of r, we sum over all possible sets of integers $\mu_1,\mu_2,\ldots,\mu_r \in\{1,2,\ldots,m\}$ such that $\mu_1<\mu_2<\ldots<\mu_r$, and all sets of integers $\nu_1,\nu_2,\ldots,\nu_r$

$\in\{1,2,\ldots,m\}$ such that $\nu_1,\nu_2,\ldots,\nu_r$ form a decreasing sequence up to a cyclic permutation. $\tilde{\omega}^{ij}$ is an antisymmetric tensor. Equation (21) defines the *Weyl-ordered Poisson algebra* for the space $\mathcal{W}$.

Now we are going to define a linear transformation $F:\mathcal{W}\to\mathcal{N}$. Nevertheless, we need a number of lemmas first in order to show that F is well defined. Introduce two matrices S and J as follows:

$$S=2\begin{pmatrix} I & 0 \\ 0 & -iI \end{pmatrix}; \tag{22}$$

and

$$J=\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \tag{23}$$

where I is the $\Lambda\times\Lambda$ unit matrix. The index of each row and column of S and J runs from 1 to Λ, then from -1 to $-\Lambda$. From Eq. (22), we see that

$$S^{-1}=\frac{1}{2}\begin{pmatrix} I & 0 \\ 0 & iI \end{pmatrix}. \tag{24}$$

Let η'^{i}, where $i\in\{-\Lambda,-\Lambda+1,\ldots,-1,1,2,\ldots,\Lambda\}$, be defined as

$$\eta'^{i}=(S^{-1})^{i}(z^{i}+J^{i\bar{i}}z^{\bar{i}}), \tag{25}$$

where

$$(S^{-1})^{i}=(S^{-1})^{ii} \tag{26}$$

and

$$\bar{i}=-i. \tag{27}$$

Moreover, let

$$C^{ij}=\gamma^{ij}+J^{i\bar{i}}\gamma^{\bar{i}j}+J^{j\bar{j}}\gamma^{i\bar{j}}+J^{i\bar{i}}J^{j\bar{j}}\gamma^{\bar{i}\bar{j}} \tag{28}$$

and

$$T^{ij}=\frac{\hbar}{2}(S^{-1})^{i}(S^{-1})^{j}(C^{ij}+C^{ji}). \tag{29}$$

We will need these formulas in the definition of $F(f^{I})$.

Next we want to introduce the concepts of an allowable set of contracted indices, a forbidden set of contracted indices and leftover indices. Choose an ordered sequence I_a of $2k$ integers, where k is a non-negative integer with $2k\leqslant m$, from $i_1,\ldots,i_m$ with distinct subscripts. Let us call the integers $i_{a(-1)},i_{a(1)},i_{a(-2)},i_{a(2)},\ldots,i_{a(-k)}$ and $i_{a(k)}$, respectively, where $a(-1),a(1)$, $a(-2),a(2),\ldots,a(-k),a(k)\in\{1,2,\ldots,m\}$ and $a(r)\neq a(s)$ if $r\neq s$ for integers r and s such that $1\leqslant|r|\leqslant k$ and $1\leqslant|s|\leqslant k$. Then $I_a=(i_{a(-1)},i_{a(1)},i_{a(-2)},i_{a(2)},i_{a(-k)},i_{a(k)})$ will be called *an allowable set of contracted indices* (or in short I_a is *allowable*) if any arbitrary integers r and s such that $1\leqslant r<s\leqslant k$,

Condition 1: either $i_{a(\pm s)}\in I(a(-r),a(r))$ *or* $i_{a(\pm s)}\in I(a(r),a(-r))$.

Otherwise, I_a will be a *forbidden set of contracted indices* (or in short I_a is *forbidden*). We illustrate in Fig. 2 examples of an allowable set of contracted indices, and one of a forbidden set of contracted indices.

We have the following lemmas characterizing an allowable set of contracted indices.

Lemma 1: Let I_a be an allowable set of $2k$ contracted indices, and r any integer between 1 and k inclusive. Let $b(-1),b(1),b(-2),b(2),\ldots,b(-k),b(k)$ be an ordered sequence of integers such that for each r, either $b(\pm r)=a(\pm r)$ or $b(\pm r)=a(\mp r)$. Then $I_b=(i_{b(-1)},i_{b(1)},i_{b(-2)},i_{b(2)},\ldots,i_{b(-k)},i_{b(k)})$ is also an allowable set of contracted indices. If I_a is forbidden, then I_b is also forbidden.

Proof: Trivial.

Lemma 2: Let $I_a=(i_{a(-1)},i_{a(1)},i_{a(-2)},i_{a(2)},\ldots,i_{a(-k)},i_{a(k)})$ be an allowable set of contracted indices, p an arbitrary integer between 1 and $k-1$ inclusive, r any integer between 1 and k inclusive, and $(b(-1),b(1),b(-2),b(2),\ldots,b(-k),b(k))$ a sequence of integers such that

$$\begin{cases} b(\pm r)=a(\pm r) \text{ if } r\neq p \text{ and } r\neq p+1; \\ b(\pm p)=a(\pm(p+1)); \text{ and} \\ b(\pm(p+1))=a(\pm p). \end{cases} \tag{30}$$

Then $I_b=(i_{b(-1)},i_{b(1)},i_{b(-2)},i_{b(2)},\ldots,i_{b(-k)},i_{b(k)})$ is also an allowable set of contracted indices. If I_a is forbidden, then I_b is also forbidden.

Proof: Assume that I_a is allowable. From Lemma 1, we can assume without loss of generality that $a(-p)<a(p)$. It is clear that the set of integers $(i_{b(-1)},i_{b(1)},i_{b(-2)},i_{b(2)},\ldots,i_{b(-p)},i_{b(p)})$ satisfies Condition 1. Consider $i_{b(-p-1)}$ and $i_{(p+1)}$. Since I_a is allowable, we have from Condition 1 that either $i_{b(\pm(p+1))}=i_{a(\pm p)}\in I(a(-s),a(s))=I(b(-s),b(s))$ or $i_{b(\pm(p+1))}\in I(b(s),b(-s))\forall s=1,2,\ldots,p-1$. Moreover, we have either case (1) that $i_{b(\pm p)}=i_{a(\pm(p+1))}\in I(a(-p),a(p))=I(b(-p-1),b(p+1))$; case (2) that $i_{b(\pm p)}<i_{b(-p-1)}$; case (3) that $i_{b(\pm p)}>i_{b(p+1)}$; or case (4) that $i_{b(-p)}<i_{b(-p-1)}$ and $i_{b(p)}>i_{b(p+1)}$. If one of the first 3 cases holds, then $i_{b(\pm(p+1))}\in I(b(p),b(-p))$. If case (4) holds, then $i_{b(\pm(p+1))}\in I(b(-p),b(p))$. Hence in all cases, the set of integers $(i_{b(-1)},i_{b(1)},i_{b(-2)},i_{b(2)},\ldots,i_{b(-p-1)},i_{b(p+1)})$ satisfies Condition 1. It is now easy to deduce that the whole set $(i_{b(-1)},i_{b(1)},i_{b(-2)},i_{b(2)},\ldots,i_{b(-k)},i_{b(k)})$ satisfies Condition 1. Hence I_b is also allowable. The proof that I_b is forbidden if I_a is forbidden is similar. Q.E.D.

Lemma 3: Consider an allowable set of contracted indices

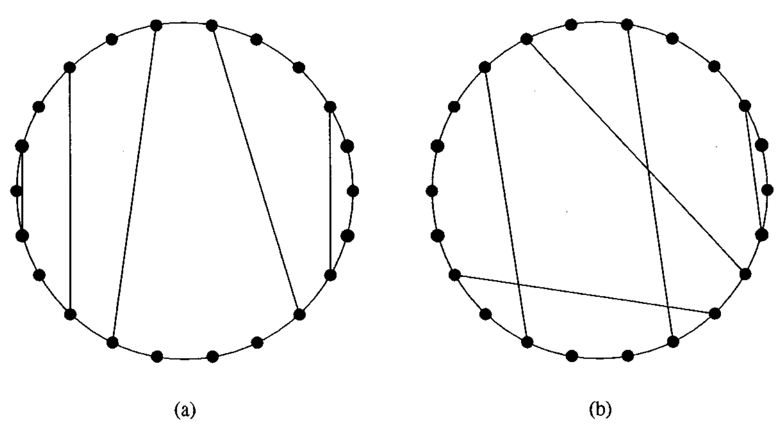

FIG. 2. (a) An allowable set of contracted indices in a loop variable (Weyl-ordered or normal-ordered). Each straight line joins $i_{a(-r)}$ and $i_{a(r)}$ together. Note that no two straight lines cross each other. (b) A forbidden set of contracted indices. Note that some straight lines cross one other.

$$I_a=(i_{a(-1)},i_{a(1)},i_{a(-2)},i_{a(2)},\ldots,i_{a(-k)},i_{a(k)}).$$

Let $\sigma:\{1,2,\ldots,k\}\rightarrow\{1,2,\ldots,k\}$ *be a permutation of the set of integers 1, 2, ..., and k. Then*

$$I_{\sigma(a)}=(i_{a(-\sigma(1))},i_{a(\sigma(1))},i_{a(-\sigma(2))},i_{a(\sigma(2))},\ldots,i_{a(-\sigma(k))},i_{a(\sigma(k))})$$

is also allowable. If I_a *is forbidden, then* $I_{\sigma(a)}$ *is also forbidden.*

Proof: This can be easily deduced from Lemma 2. Q.E.D.

In short, we see from Lemma 3 that whether a set of contracted indices I_a is allowable or not is independent of the order of the pairs of indices $i_{a(s)},i_{a(-s)}$'s. Each of these pairs will be called a *contraction pair*.

Let us concentrate on an allowable set of contracted indices I_a. For the i_l's such that $l\in\{1,2,\ldots,m\}$ but that $l\neq a(s)\in I_a\forall s=\pm1,\ldots,\pm k$ (these i_l's are called the *leftover indices*), form *subloops* by defining an integer-valued auxiliary function L of some positive integers as below. Let $L(1)=l$. If $L(v)$ is defined for an integer v, then we define $L^{(i)}(v+1)$ for some integers i by the following:

Algorithm 1: (cf. Fig. 3 below) In the following, $L(v)+_m1$ *means precisely* $L(v)+1$ *if* $L(v)\neq m$, *and it means 1 if* $L(v)=m$. *Similarly,* $L(v-_m1)$ *means* $L(v)-1$ if $L(v)\neq1$, *and it means m if* $L(v)=1$.

Step 1: Set $i=1$.

Step 2: $L^{(i)}(v+1)=L(v)+_m1$.

Step 3: If $L^{(i)}(v+1)\neq a(s)\forall s=\pm1,\ldots,\pm k$, *then end this algorithm.*

Step 4: Let s_i *be such that* $a(s_i)=L^{(i)}(v+1)$.

Step 5: Increment the value of i by 1.

Step 6: Set $L^{(i)}(v+1)=a(-s_i)+_m1$.

Step 7: Go back to Step 3.

If $L^{(i)}(v+1)\neq L(1)$, where i is the maximum integer such that $L^{(i)}(v+1)$ is defined, then define $L(v+1)=L^{(i)}(v+1)$; otherwise, $L(v+1)$ and thus $L(v+2)$, $L(v+3)$, ..., etc. are all left undefined.

Before proceeding on using the auxiliary function L to define a subloop, we need to show that the above algorithm is well defined by

Lemma 4: In the notations of Algorithm 1, if $L(v)$ *is well defined, then there is an i such that* $L^{(i)}(v+1)\neq a(s)\forall s=\pm1,\ldots,\pm k$.

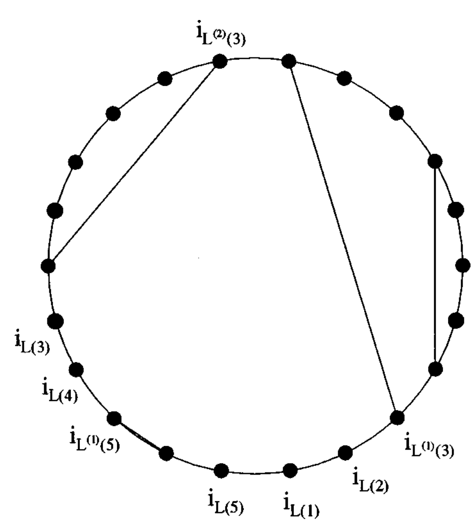

FIG. 3. A typical subloop with 5 numbers in the integer sequence L.

Proof: Assume on the contrary that such an i does not exist. Then we have an infinite sequence $L^{(1)}(v+1), L^{(2)}(v+1), \ldots$ and so on. Since there are a finite number of $a(s)$'s for $s = \pm 1, \ldots, \pm k$ only, there is an integer i_2 such that $L^{(i_2)}(v+1) = L^{(i_1)}(v+1) \exists i_1 \in \{1,2,\ldots,i_2 - 1\}$. Consider the case $i_1 \neq 1$. Let $L^{(i_1-1)}(v+1) = a(s_1)$ and $L^{(i_2-1)} = a(s_2)$ for some integers s_1 and s_2. Then $L^{(i_1)}(v+1) = a(-s_1) +_m 1$ and $L^{(i_2)}(v+1) = a(-s_2) +_m 1$. Hence $a(-s_1) +_m 1 = a(-s_2) +_m 1$ and thus $a(s_1) = a(s_2)$, i.e., $L^{(i_1-1)}(v+1) = L^{(i_2-1)}(v+1)$, contradicting the assumption that $L^{(i_2)}(v+1)$ is the first integer that repeats one of the previous numbers in the sequence. Now consider the case $i_1 = 1$. Then $L^{(i_1)}(v+1) = L(v) +_m 1$ and $L^{(i_2)}(v+1) = a(-s_2) +_m 1$. Hence $L(v) = a(-s_2)$. However, $L(v)$ does not belong to I_a and this equation is impossible. Consequently, there is an i such that $L^{(i)}(v+1) \neq a(s)$ for all $s = \pm 1, \ldots, \pm k$. Q.E.D.

Let u be the maximum integer such that $L(u)$ is defined. Then the *subloop of f^I with respect to I_a* including i_l is given by $\phi^L = \mathrm{Tr}\, \eta'^{i_{L(1)}} \eta'^{i_{L(2)}} \cdots \eta'^{i_{L(u)}}$, where η'^{i} is defined in Eq. (25). Obviously any one of the leftover indices belongs to at least one of these subloops. Moreover, no two distinct subloops ϕ^{L_1} and ϕ^{L_2} of f^I with respect to I_a share even one common η'^{i_l} for an arbitrary leftover index i_l because of the following two lemmas.

Lemma 5: Consider a subloop $\phi^L = \mathrm{Tr}\, \eta'^{i_{L(1)}} \cdots \eta'^{i_{L(u)}}$ of f^I with respect to I_a. Let r and s be integers between 1 and u inclusive. Then $L(r) \neq L(s)$ if $r \neq s$.

Proof: Assume on the contrary that there exist some integers $\tilde{r}$ and $\tilde{s}$ such that $\tilde{r} \neq \tilde{s}$ but $L(\tilde{r}) = L(\tilde{s})$. Choose the smallest integer r out of these $\tilde{r}$ and $\tilde{s}$'s. Then $r > 1$ from the statement immediately after Algorithm 1. Let s be the smallest integer distinct from r such that $L(s) = L(r)$. Then $s > r > 1$. Again from the statement immediately after Algorithm 1, there exists an unknown integer x such that $L(r) = L^{(x)}(r)$. Consider the following reverse of Algorithm 1:

Algorithm 2: Here is the procedure of this algorithm.

Step 1: Set $i = 0$.

Step 2: Let an integer y be such that it satisfies the equation $L^{(x-i)}(r) = y +_m 1$.

Step 3: If y does not belong to I_a, then $y = L(r-1)$ from Step 2 of Algorithm 1 (or else $L^{(x-i)}(r) = y +_m 1$, where $y \in I_a$ because of Step 6 of Algorithm 1, which is impossible). Hence $x - i = 1 \Rightarrow x = i + 1$. End the algorithm.

Step 4: Since $y \in I_a$, $L^{(x-i)}(r) = a(-s_{i+1}) +_m 1 \exists$ integer s_{i+1}.

Step 5: Increment the value of i by 1.

Step 6: From Steps 6 and 4 of Algorithm 1, $L^{(x-i)}(r) = a(s_i)$.

Step 7: Go back to Step 2.

Hence $L(r-1)$ can be uniquely determined just from the value of $L(r)$ by Algorithm 2. Moreover, $L(s-1)$ can be uniquely determined just from the value of $L(s)$ by the same algorithm. Since $L(r) = L(s)$, we must have $L(r-1) = L(s-1)$, contradicting the assumption that r is the smallest number such that $L(r) = L(s)$ for a number $s > r$. Q.E.D.

Corollary 1: The degree of a subloop is a finite positive integer.

Proof: Since the degree of a loop is a finite number only, a subloop of it also has a finite degree by Lemma 5.

Lemma 6: For each distinct $l \in \{1,2,\ldots,m\}$ such that i_l is a leftover index, η'^{i_l} is contained in at most one of the distinct subloops produced from all the leftover indices.

Proof: Let $\eta'^{i_l} \in \phi^L = \mathrm{Tr}\, \eta'^{i_{L(1)}} \cdots \eta'^{i_{L(u)}}$, where $L(1) = l$. Consider another subloop $\phi^{\tilde{L}} = \mathrm{Tr}\, \eta'^{i_{\tilde{L}(1)}} \cdots \eta'^{i_{\tilde{L}(v)}} \cdots \eta'^{i_{\tilde{L}(\tilde{u})}}$, where $\tilde{L}(v) = l$. From Algorithm 1, it is clear that $\tilde{L}(v+1) = L(2)$, $\tilde{L}(v+2) = L(3)$, $\ldots$, and $\tilde{L}(\tilde{u}) = L(\tilde{u} - v + 1)$. Then $\tilde{L}^{(i)}(\tilde{u}+1) = \tilde{L}(1)$ for the maximum integer i such that $\tilde{L}^{(i)}(\tilde{u}+1)$ is defined. On the other hand, $\tilde{L}^{(i)}(\tilde{u}+1) = L^{(i)}(\tilde{u} - v + 2) = L(\tilde{u} - v + 2)$. Hence $\tilde{L}(1) = L(\tilde{u} - v + 2)$. Then $\tilde{L}(2) = L(\tilde{u} - v + 3)$, $\ldots$, and $\tilde{L}(u - \tilde{u} + v - 1) = L(u)$. $L^{(j)}(u+1) = L(1)$ for the maximum integer j such that $L^{(j)}(u+1)$ is defined. However, $L^{(j)}(u+1) = \tilde{L}^{(j)}(u - \tilde{u} + v) = \tilde{L}(u - \tilde{u} + v)$. As a result, $\tilde{L}(u - \tilde{u} + v) = L(1) = \tilde{L}(v)$. By Lemma 5, $\tilde{L}(u - \tilde{u} + v) = \tilde{L}(v)$ only if $u = \tilde{u}$. Now it is clear that $\phi^L = \phi^{\tilde{L}}$. Q.E.D.

Figure 3 shows a typical subloop.

The following lemmas and corollary pertaining to subloops will be found useful later.

Lemma 7: Let us consider a particular pair of indices $a(-s_0)$ and $a(s_0)$, where $1 \leqslant |s_0|$

$\leqslant k$, and the sequence L^{ext} of the numbers $L(1), L^{(1)}(2), L^{(2)}(2), \ldots, L^{(\iota_2)}(2) = L(2), L^{(1)} \times(3), L^{(2)}(3), \ldots, L^{(\iota_3)}(3) = L(3), \ldots, L^{(1)}(u), L^{(2)}(u), \ldots, L^{(\iota_u)}(u) = L(u), L^{(1)}(u+1), L^{(2)}(u+1), \ldots$, and $L^{(\iota_{u+1}-1)}(u+1)$, where $\iota_2, \iota_3, \ldots, \iota_{u+1}$ are the maximum integers such that $L^{(\iota)}(x)$ is defined for $2 \leqslant x \leqslant u+1$ and $\iota \leqslant \iota_x$. Moreover, $L^{(\iota_{u+1})}(u+1) = L(1)$ by the definition of u. Then either all numbers in $L^{ext} \in I(a(-s_0), a(s_0)) \cup \{a(s_0)\}$ or all numbers in $L^{ext} \in I(a(s_0), a(-s_0)) \cup \{a(-s_0)\}$.

Proof: Let $L(1) \in I(a(s_0), a(-s_0)) \cup \{a(-s_0)\}$. Note that in the sequence L^{ext}, the immediately succeeding number $\lambda^{(s)}$ of a preceding number $\lambda^{(p)}$ is always obtained either by (1) $\lambda^{(s)} = \lambda^{(p)} +_m 1$ if $\lambda^{(p)}$ does not belong to I_a, or by (2) $\lambda^{(s)} = a(-s^{(p)}) +_m 1$ for an integer $s^{(p)}$ such that $\lambda^{(p)} = a(s^{(p)})$ if $\lambda^{(p)} \in I_a$. Assume that $\lambda^{(p)} \in I(a(s_0), a(-s_0)) \cup \{a(-s_0)\}$. In Case (1), $\lambda^{(p)} \in I(a(s_0), a(-s_0))$ and so $\lambda^{(s)} \in I(a(s_0), a(-s_0))$ or $\lambda^{(s)} = a(-s_0)$. Hence $\lambda^{(s)} \in I(a(s_0), a(-s_0)) \cup \{a(-s_0)\}$. In Case (2), if $a(s^{(p)}) \neq a(-s_0)$, then $a(s^{(p)}) \in I(a(s_0), a(-s_0))$ and so $a(-s^{(p)}) \in I(a(s_0), a(-s_0))$ (because I_a is allowable and because of Lemma 3). This implies $a(-s^{(p)}) +_m 1 \in I(a(s_0), a(-(s_0)) \cup \{a(-s_0)\}$. If, on the other hand, $a(s^{(p)}) = a(-s_0)$, then $a(-s^{(p)}) = a(s_0)$ and so $a(-s^{(p)}) +_m 1 \in I(a(s), a(-s_0)) \cup \{a(-s_0)\}$. By induction, all numbers in $L^{ext} \in I(a(s_0), a(-s_0)) \cup \{a(-s_0)\}$. The case for $L(1) \cup I(a(-s_0), a(s_0)) \cup \{a(s_0)\}$ is similar. Q.E.D.

Lemma 8: Consider a subloop $\phi^L = \mathrm{Tr}\, \eta'^{i_{L(1)}} \cdots \eta'^{i_{L(u)}}$ of f^I with respect to I_a. Without loss of generality, $L(1)$ can be chosen to be the smallest among $L(1), L(2), \ldots, L(u)$ by a cyclic permutation of the η' matrices. Then using the notations of Lemma 5, we have $L(1) < L^{(1)}(2) < L^{(2)}(2) < \cdots < L^{(\iota_2)}(2) < L^{(1)}(3) < L^{(2)}(3) < \cdots < L^{(\iota_3)}(3) < \cdots < L^{(1)}(u) < L^{(2)}(u) < \cdots < L^{(\iota_u)}(u)$.

Proof: Consider the numbers $\lambda^{(p)}$ and $\lambda^{(s)}$ defined in the proof of Lemma 7. Assume that $L(1) < L^{(1)}(2) < \cdots < L^{(\iota_2)}(2) < \cdots < \lambda^{(p)}$. If $\lambda^{(p)} = L^{(\iota_u)}(u)$, the lemma is proved. If $\lambda^{(p)}$ is a number before $L^{(\iota_u)}(u)$ in the sequence L^{ext}, then $\lambda^{(s)}$ is obtained by the two alternatives described in the proof of Lemma 7. For the case $\lambda^{(p)}$ does not belong to I_a, we have $\lambda^{(s)} = \lambda^{(p)} +_m 1$. Thus either $\lambda^{(s)} = 1 \leqslant L(1)$ or $\lambda^{(s)} > \lambda^{(p)}$. For the case $\lambda^{(p)} \in I_a$, we have $\lambda^{(s)} = a(-s^{(p)}) +_m 1$ where $s^{(p)}$ is defined in the proof of the Lemma 7. If $a(s^{(p)}) < a(-s^{(p)})$, then either $\lambda^{(s)} \leqslant L(1)$ or $\lambda^{(s)} > \lambda^{(p)}$. If $a(s^{(p)}) > a(-s^{(p)})$, then we deduce from $\lambda^{(p)} \in I(a(-s^{(p)}), a(s^{(p)})) \cup \{a(s^{(p)})\}$ and Lemma 7 that $L(1) \in I(a(-s^{(p)}), a(s^{(p)}))$. Then $\lambda^{(s)} = a(-s^{(p)}) +_m 1 \leqslant L(1)$. Hence this lemma is proved if we can show that $\lambda^{(s)} \leqslant L(1)$ is impossible.

Clearly, $\lambda^{(s)} = L(1)$ is impossible by Lemma 5. Let $\lambda^{(s)} < L(1)$, and let $\lambda^{(s)} = L^{(\iota)}(x)$ for some numbers ι and x. Consider the numbers $L^{(\iota)}(x), L^{(\iota+1)}(x), \ldots$, and $L^{(\iota_x)}(x) = L(x)$. Since $L(x) > L(1)$, there is a smallest integer ι_c such that $L^{(\iota_c)}(x) < L(1)$ but $L^{(\iota_c+1)}(x) > L(1)$. Let $L^{(\iota_c)}(x) = a(s_c)$. Then $L^{(\iota_c+1)}(x) = a(-s_c) +_m 1 > L(1)$ and so $a(-s_c) > L(1)$. This, together with $a(s_c) < L(1)$, implies $L(1) \in I(a(s_c), a(-s_c)) \cup \{a(-s_c)\}$. By Lemma 7, $a(s_c) = L^{(\iota_c)}(x) \in I(a(s_c), a(-s_c)) \cup \{a(-s_c)\}$, and this is clearly impossible. Q.E.D.

Corollary 2: Consider a subloop $\phi^L = \mathrm{Tr}\, \eta'^{i_{L(1)}} \cdots \eta'^{i_{L(u)}}$ of f^I with respect to I_a. $L(1)$ can be chosen to be the smallest among $L(1), L(2), \ldots, L(u)$ without loss of generality. Then $L(1) < L(2) < \cdots < L(u)$.

Proof: This follows directly from Lemmas 6 and 8. Q.E.D.

We are now ready to define $F: \mathcal{W} \to \mathcal{N}$. Let $f^I = \mathrm{Tr}\, \eta^{i_1} \cdots \eta^{i_m} \in \mathcal{W}$. Then,

$$F(f^I) = \sum_{\substack{\text{all distinct allowable sets} \\ \text{of contracted indices } I_a}} T^{i_{a(-1)} i_{a(1)}} T^{i_{a(-2)} i_{a(2)}} \ldots T^{i_{a(-k)} i_{a(k)}} \cdot \prod_{\substack{\text{all distinct subloops } L \\ \text{of } f^I \text{ w.r.t. } I_a.}} \mathrm{Tr}\, \eta'^{i_{L(1)}} \eta'^{i_{L(2)}} \cdots \eta'^{i_{L(u)}}. \tag{31}$$

For instance,

$$F(\text{Tr } \eta^{i_1}\eta^{i_2}\eta^{i_3}\eta^{i_4}) = \text{Tr } \eta'^{i_1}\eta'^{i_2}\eta'^{i_3}\eta'^{i_4} + T^{i_1i_2}\text{ Tr } \eta'^{i_3}\eta'^{i_4} + T^{i_1i_3}\text{ Tr } \eta'^{i_2}\text{ Tr } \eta'^{i_4} + T^{i_1i_4}\text{ Tr } \eta'^{i_2}\eta'^{i_3} + T^{i_2i_3}\text{ Tr } \eta'^{i_1}\eta'^{i_4} + T^{i_2i_4}\text{ Tr } \eta'^{i_1}\text{ Tr } \eta'^{i_3} + T^{i_3i_4}\text{ Tr } \eta'^{i_1}\eta'^{i_2} + T^{i_1i_2}T^{i_3i_4} + T^{i_1i_4}T^{i_2i_3}. \tag{32}$$

Let us give another example. In $F(\text{Tr } \eta^{i_1}\eta^{i_2}\eta^{i_3}\eta^{i_4}\eta^{i_5}\eta^{i_6})$, there are terms like $T^{i_2i_3}T^{i_5i_6}\text{ Tr } \eta'^{i_1}\eta'^{i_4}$, $T^{i_1i_6}T^{i_3i_4}\text{ Tr } \eta'^{i_2}\eta'^{i_5}$ and $T^{i_1i_2}T^{i_4i_5}\text{ Tr } \eta'^{i_3}\eta'^{i_6}$. Moreover, we define $F(f^I f^J) = F(f^I)F(f^J)$ for some $f^I, f^J \in \mathcal{W}$. The following lemma shows that F is invertible.

Lemma 9: Consider the mapping $F:\mathcal{W}\rightarrow\mathcal{N}$ defined in Eq. (32). Then F is invertible.

Proof: Let $P(n)$ be the proposition that for every normal-ordered loop variable ϕ^I in $\mathcal{N}$ of degree n, there is a unique element in $\mathcal{W}$ such that F maps this element to ϕ^I. From Eq. (25), we see that

$$\begin{cases} z^j = \eta'^j + i\,\eta'^{-j} \\ z^{-j} = \eta'^j - i\,\eta'^{-j} \end{cases} \tag{33}$$

for $j \in \{1, \ldots, M\}$. Hence

$$\begin{cases} \text{Tr } z^j = \text{Tr } \eta'^j + i\text{ Tr } \eta'^{-j} \\ \text{Tr } z^{-j} = \text{Tr } \eta'^j - i\text{ Tr } \eta'^{-j} \end{cases} \tag{34}$$

for each j. Hence $P(1)$ is true. Assume that $P(k)$ is true. Consider $\phi^J = \text{Tr } z^{j_1}\cdots z^{j_{k+1}}$. From Eq. (33), ϕ^J is a linear combination of $\text{Tr } \eta'^{j'_1}\eta'^{j'_2}\cdots\eta'^{j'_{k+1}}$, where $j'_1 = j_1$ or $-j_1$, $j'_2 = j_2$ or $-j_2$, ..., and $j'_{k+1} = j_{k+1}$ or $-j_{k+1}$. Each of these in turn differs from $F(\text{Tr } \eta^{j'_1}\eta^{j'_2}\cdots\eta^{j'_{k+1}})$ by normal-ordered loop variables of degrees less than $k+1$. By the induction hypothesis, there is an element f' in W which is mapped by F to the sum of these normal-ordered loop variables of lower degree. Hence $\text{Tr } \eta'^{j'_1}\eta'^{j'_2}\cdots\eta'^{j'_{k+1}} = F(\text{Tr } \eta^{j'_1}\eta^{j'_2}\cdots\eta^{j'_{k+1}} + f')$. Therefore, $P(k+1)$ is true and there is an element $f^I \in \mathcal{W}$ such that $F(f^I) = \phi^J$. If there is another element $f^{I'} \neq f^I$ such that $F(f^{I'}) = \phi^J$, then $F(f^{I'} - f^I) = 0$ for $f^{I'} - f^I \neq 0$. However, this is impossible from Eq. (31). Q.E.D.

Having defined a mapping $F:\mathcal{W}\rightarrow\mathcal{N}$, we are going to prove that this is a Poisson morphism. Let f^I and $f^J \in \mathcal{W}$. F is a Poisson morphism if (1) every term in $\{F(f^I), F(f^J)\}_N$ is also a term in $F(\{f^I, f^J\}_W)$, which is a product of normal-ordered loop variables, and (2) every term in $F(\{f^I, f^J\}_W)$ is also a term in $\{F(f^I), F(f^J)\}_N$.

Let us derive expressions for $\{F(f^I), F(f^J)\}_N$ and $F(\{f^I, f^J\}_W)$ first before proving these two statements. From Eq. (31) and the Leibniz property of a Poisson bracket,

$$\begin{aligned} \{F(f^I), F(f^J)\}_N = & \sum_{\substack{\text{distinct} \\ \text{allowable } I_a}} T^{i_{a(-1)}i_{a(1)}}T^{i_{a(-2)}i_{a(2)}}\cdots T^{i_{a(-k)}i_{a(k)}} \\ & \cdot \sum_{\substack{\text{distinct} \\ \text{allowable } J_b}} T^{j_{b(-1)}j_{b(1)}}T^{j_{b(-2)}j_{b(2)}}\cdots T^{j_{b(-l)}j_{b(l)}} \\ & \cdot \prod_{\substack{\text{distinct subloops } L \text{ of } f^I \\ \text{w.r.t. } I_a \text{ except subloop } A}} \text{Tr } \eta'^{i_{L(1)}}\eta'^{i_{L(2)}}\cdots\eta'^{i_{L(u)}} \\ & \cdot \prod_{\substack{\text{distinct subloops } M \text{ of } f^J \\ \text{w.r.t. } J_b \text{ except subloop } B}} \text{Tr } \eta'^{j_{M(1)}}\eta'^{j_{M(2)}}\cdots\eta'^{j_{M(v)}} \\ & \cdot \{\text{Tr } \eta'^{i_{A(1)}}\cdots\eta'^{i_{A(\alpha)}},\ \text{Tr } \eta'^{j_{B(1)}}\cdots\eta'^{j_{B(\beta)}}\}_N, \end{aligned} \tag{35}$$

where $J_b=(j_{b(-1)},j_{b(1)},j_{b(-2)},j_{b(2)},\ldots,j_{b(-l)},j_{b(l)})$ is an allowable set of contracted indices in J for a positive integer l, u, and v are the degrees of the subloops L and M, respectively, and α and β are the degrees of the subloops A and B. Furthermore, from Eqs. (25) and (19),

$$\begin{aligned}&\{\mathrm{Tr}\,\eta'^{i_{A(1)}}\cdots\eta'^{i_{A(\alpha)}},\ \mathrm{Tr}\,\eta'^{j_{B(1)}}\cdots\eta'^{j_{B(\beta)}}\}_N\\ &=\sum_{r'=1}^{\infty}\sum_{\substack{\mu_1<\mu_2<\cdots<\mu_{r'}\\(\nu_1>\nu_2>\cdots>\nu_{r'})}}\hbar^{r'}(S^{-1})^{i_{A(\mu_1)}}(S^{-1})^{i_{A(\mu_2)}}\cdots(S^{-1})^{i_{A(\mu_{r'})}}\\ &\quad\cdot(S^{-1})^{j_{B(\nu_1)}}(S^{-1})^{j_{B(\nu_2)}}\cdots(S^{-1})^{j_{B(\nu_{r'})}}\\ &\quad\cdot(C^{i_{A(\mu_1)}j_{B(\nu_1)}}C^{i_{A(\mu_2)}j_{B(\nu_2)}}\cdots C^{i_{A(\mu_{r'})}j_{B(\nu_{r'})}}-C^{j_{B(\nu_1)}i_{A(\mu_1)}}C^{j_{B(\nu_2)}i_{A(\mu_2)}}\cdots C^{j_{B(\nu_{r'})}i_{A(\mu_{r'})}})\\ &\quad\cdot H^{I_A(\mu_1,\mu_2)J_B(\nu_2,\nu_1)}H^{I_A(\mu_2,\mu_3)J_B(\nu_3,\nu_2)}\cdots H^{I_A(\mu_{r'},\mu_1)J_B(\nu_1,\nu_{r'})}.\end{aligned}\tag{36}$$

In this equation, if $\mu_1<\mu_2$, then

$$I_A(\mu_1,\mu_2)=i_{A(\mu_1+1)},i_{A(\mu_1+2)},\ldots,i_{A(\nu_2-1)}.\tag{37}$$

If, instead, $\mu_1\geqslant\mu_2$, then

$$I_A(\mu_1,\mu_2)=i_{A(\mu_1+1)},i_{A(\mu_1+2)},\ldots,i_{A(\alpha)},i_{A(1)},i_{A(2)},\ldots,i_{A(\nu_2-1)}.\tag{38}$$

We have a similar definition for $J_B(\nu_2,\nu_1)$. In addition,

$$H^{I_A(\mu_1,\mu_2)J_B(\nu_2,\nu_1)}=\mathrm{Tr}\,\eta'^{i_{A(\mu_1+1)}}\cdots\eta'^{i_{A(\mu_2-1)}}\eta'^{j_{B(\nu_2+1)}}\cdots\eta'^{j_{B(\nu_1-1)}}\tag{39}$$

if $\mu_1<\mu_2$ and $\nu_1>\nu_2$, and so on for $H^{I_A(\mu_2,\mu_3)J_B(\nu_3,\nu_2)}$, …, etc.

Let us define,

$$\omega^{ij}=\mathrm{i}(S^{-1})^i(S^{-1})^j(C^{ij}-C^{ji}).\tag{40}$$

Then Eq. (36) can be simplified by the following lemma:

Lemma 10: (Within the statements and proofs of Lemmas 10 and 11, $i_{A(\mu_k)}$ and $j_{B(\nu_k)}$ will be abbreviated as i_k and j_k, respectively.) The following identity holds true:

$$\begin{aligned}&(S^{-1})^{i_1}(S^{-1})^{i_2}\cdots(S^{-1})^{i_{r'}}(S^{-1})^{j_1}(S^{-1})^{j_2}\cdots(S^{-1})^{j_{r'}}\\ &\quad\cdot(C^{i_1j_1}C^{i_2j_2}\cdots C^{i_{r'}j_{r'}}-C^{j_1i_1}C^{j_2i_2}\cdots C^{j_{r'}i_{r'}})\\ &=2\sum_{\substack{\text{distinct sets of choices for}\\ \Delta\text{ with an odd number of }\omega'\text{s}}}(\Delta)^{i_1j_1}(\Delta)^{i_2j_2}\cdots(\Delta)^{i_{r'}j_{r'}}\end{aligned}\tag{41}$$

where each $(\Delta)^{ij}$ can be chosen as either $-(i/2)\omega^{ij}$ or $(1/\hbar)T^{ij}$.

In order to prove Lemma 10, we need to state and prove Lemma 11 simultaneously.

Lemma 11: The following identity holds true:

$$\begin{aligned}&(S^{-1})^{i_1}(S^{-1})^{i_2}\cdots(S^{-1})^{i_{r'}}(S^{-1})^{j_1}(S^{-1})^{j_2}\cdots(S^{-1})^{j_{r'}}\\ &\quad\cdot(C^{i_1j_1}C^{i_2j_2}\cdots C^{i_{r'}j_{r'}}+C^{j_1i_1}C^{j_2i_2}\cdots C^{j_{r'}i_{r'}})\\ &=2\sum_{\substack{\text{distinct sets of choices of}\\ \Delta\text{ with an even number of }\omega'\text{s}}}(\Delta)^{i_1j_1}(\Delta)^{i_2j_2}\cdots(\Delta)^{i_{r'}j_{r'}}\end{aligned}\tag{42}$$

Proof of Lemmas 10 and 11: Let us calculate the coefficient of the term

$$(S^{-1})^{i_1}(S^{-1})^{i_2}\cdots(S^{-1})^{i_{r'}}(S^{-1})^{j_1}(S^{-1})^{j_2}\cdots(S^{-1})^{j_{r'}}C^{i_1j_1}C^{i_2j_2}\cdots C^{i_{r'}j_{r'}} \tag{43}$$

on the right-hand sides of Eqs. (41) and (42) first. This is $(1/2^{r-1})$ the number of summands on the right-hand sides of these two equations because each summand contributes to Formula (43) whatever set of choices of Δ's we choose. Since the number of distinct choices is $C_1^{r'}+C_3^{r'}+\cdots+C_{r'-1}^{r'}$ or $C_1^{r'}+C_3^{r'}+\cdots+C_{r'}^{r'}=2^{r'-1}$ for Eq. (41) and $C_0^{r'}+C_2^{r'}+\cdots+C_{r'-1}^{r'}$ or $C_0^{r'}+C_2^{r'}+\cdots+C_{r'}^{r'}=2^{r'-1}$ for Eq. (42), this coefficient is 1. Similarly, the numerical coefficient of the expression

$$(S^{-1})^{i_1}(S^{-1})^{i_2}\cdots(S^{-1})^{i_{r'}}(S^{-1})^{j_1}(S^{-1})^{j_2}\cdots(S^{-1})^{j_{r'}}C^{j_1i_1}C^{j_2i_2}\cdots C^{j_{r'}i_{r'}} \tag{44}$$

is -1 on the R.H.S. of Eq. (41) and 1 on that of Eq. (42), the negative sign in Eq. (41) being due to the fact that we choose an odd number of the Δ's to be ω's, whereas in Eq. (42) we choose an even number. Hence, every term on the left-hand sides of Eq. (41) and Eq. (42) are contained in the right-hand sides of the same equations with the same coefficient. We are going to show that there are no other terms on the R.H.S.'s besides the terms present on the left-hand sides.

Indeed, let $C^{\langle ij\rangle}=C^{ij}$ or C^{ji}, and let $P(r',k,-)$ be the proposition that the coefficient of $C^{\langle i_1j_1\rangle}C^{\langle i_2j_2\rangle}\cdots C^{\langle i_{r'}j_{r'}\rangle}$, where k of the $C^{\langle ij\rangle}$'s are C^{ji}'s and $r'-k$ of them are C^{ij}'s vanishes on the R.H.S. of Eq. (41) for $1\leqslant k\leqslant r'-1$. Similarly, let $P(r',k,+)$ be the proposition that this coefficient vanishes on the R.H.S. of Eq. (42) for $1\leqslant k\leqslant r'-1$. Consider $P(2,1,-)$ and $P(2,1,+)$. The R.H.S. of Eq. (41) is

$$-i\frac{T^{i_1j_1}\omega^{i_2j_2}}{\hbar}-i\frac{\omega^{i_1j_1}T^{i_2j_2}}{\hbar}=(S^{-1})^{i_1}(S^{-1})^{i_2}(S^{-1})^{j_1}(S^{-1})^{j_2}(C^{i_1j_1}C^{i_2j_2}-C^{j_1i_1}C^{j_2i_2}). \tag{45}$$

Therefore, $P(2,1,-)$ is true. Similarly, the R.H.S. of Eq. (42) is

$$2\frac{T^{i_1j_1}T^{i_2j_2}}{\hbar^2}-\frac{1}{2}\omega^{i_1j_1}\omega^{i_2j_2}=(S^{-1})^{i_1}(S^{-1})^{i_2}(S^{-1})^{j_1}(S^{-1})^{j_2}(C^{i_1j_1}C^{i_2j_2}+C^{j_1i_1}C^{j_2i_2}). \tag{46}$$

Hence $P(2,1,+)$ is also true.

Now assume that $P(r'',k,-)$ and $P(r'',k,+)$ are true for a positive integer $r''\geqslant 2$ and $1\leqslant k\leqslant r''-1$. Consider $P(r''+1,k,-)$. There are 2 types of summands on the R.H.S. of Eq. (41) which contributes to $t=C^{\langle i_1j_1\rangle}\cdots C^{\langle i_{r''}j_{r''}\rangle}C^{\langle i_{r''+1}j_{r''+1}\rangle}$. One type (*type 1 summands*) is of the general form $(\Delta)^{i_1j_1}\cdots(\Delta)^{i_{r''}j_{r''}}T^{i_{r''+1}j_{r''+1}}$. Here an odd number of the Δ's are ω's, and the rest are T's. The other type (*type 2 summands*) is of the general form $(\Delta)^{i_1j_1}\cdots(\Delta)^{i_{r''}j_{r''}}\omega^{i_{r''+1}j_{r''+1}}$. Here an even number of the Δ's are ω's. There are several different cases.

- *Case 1*: $2\leqslant k\leqslant r''-1$.

Subcase a: $t=C^{\langle i_1j_1\rangle}\cdots C^{\langle i_{r''}j_{r''}\rangle}C^{i_{r''+1}j_{r''+1}}$.

Since $P(r'',k,-)$ is true, the coefficient of t derived from type 1 summands, where the first r'' $C^{\langle ij\rangle}$'s come from $(\Delta)^{ij}$'s and $C^{i_{r''+1}j_{r''+1}}$ comes from $T^{i_{r''+1}j_{r''+1}}$, is 0. Since $P(r'',k,+)$ is also true, the coefficient of t derived from type 2 summands, where $C^{i_{r''+1}j_{r''+1}}$ comes from $\omega^{i_{r''+1}j_{r''+1}}$ instead, is also 0. As a result, $P(r''+1,k,-)$ is true in this subcase.

Subcase b: $t=C^{\langle i_1j_1\rangle}\cdots C^{\langle i_{r''}j_{r''}\rangle}C^{j_{r''+1}i_{r''+1}}$.

Since $P(r'',k-1,-)$ is true, the coefficient of t derived from type 1 summands is 0. Since $P(r'',k-1,+)$ is also true, the coefficient of t from type 2 summands is also 0. Hence $P(r''+1,k,-)$ is also true in this subcase.

- *Case 2:* $k=1$.

Subcase a: $t=C^{\langle i_1j_1\rangle}\cdots C^{\langle i_{r''}j_{r''}\rangle}C^{i_{r''+1}j_{r''+1}}$.

This is exactly the same as Subcase 1a.

Subcase b: $t=C^{i_1j_1}\cdots C^{i_{r''}j_{r''}}C^{j_{r''+1}i_{r''+1}}$.

The coefficient of t derived from type 1 summands is $\frac{1}{2}$ (this $\frac{1}{2}$ comes from the term $\frac{1}{2}C^{j_{r''+1}i_{r''+1}}$ in $T^{i_{r''+1}j_{r''+1}}$), and the coefficient of t derived from type 2 summands is $-\frac{1}{2}$ (because of the term $-\frac{1}{2}C^{j_{r''+1}i_{r''+1}}$ in $\omega^{j_{r''+1}i_{r''+1}}$). Hence the total coefficient is $\frac{1}{2}-\frac{1}{2}=0$, i.e., $P(r''+1,1,-)$ is true in this case.

• *Case 3:* $k=r''$

Subcase a: $t=C^{j_1 i_1}\cdots C^{j_{r''}i_{r''}}C^{i_{r''+1}j_{r''+1}}$.

The coefficient of t derived from type 1 summands is $-\frac{1}{2}$, whereas that derived from type 2 summands is $\frac{1}{2}$. Hence the total coefficient vanishes and $P(r''+1,r'',-)$ is true in this case.

Subcase b: $t=C^{\langle i_1 j_1\rangle}\cdots C^{\langle i_{r''}j_{r''}\rangle}C^{j_{r''+1}i_{r''+1}}$.

This is the same as Subcase 1b.

Thus $P(r''+1,k,-)$ is true for all cases for $1\leqslant k\leqslant r''$. With the same induction hypothesis, $P(r''+1,k,+)$ is also true by a similar analysis. By induction, $P(r',k,-)$ and $P(r',k,+)$ are always true for $r'\geqslant 2$ and $1\leqslant k\leqslant r'-1$. Q.E.D.

With the help of Lemma 10, we can derive from Eqs. (35) and (36) that $\{F(f^I),F(f^J)\}_N$ is a linear combination of all terms of the form

$$
\begin{aligned}
&T^{i_{a(-1)}i_{a(1)}}T^{i_{a(-2)}i_{a(2)}}\cdots T^{i_{a(-k)}i_{a(k)}}T^{j_{b(-1)}j_{b(1)}}T^{j_{b(-2)}j_{b(2)}}\cdots T^{j_{b(-l)}j_{b(l)}}\\
&\cdot\prod_{\substack{\text{distinct subloops } L \text{ of } f^I\\ \text{w.r.t. } I_a \text{ except subloop } A}}\operatorname{Tr}\eta'^{i_{L(1)}}\eta'^{i_{L(2)}}\cdots\eta'^{i_{L(u)}}\\
&\cdot\prod_{\substack{\text{distinct subloops } M \text{ of } f^J\\ \text{w.r.t. } J_b \text{ except subloop } B}}\operatorname{Tr}\eta'^{j_{M(1)}}\eta'^{j_{M(2)}}\cdots\eta'^{j_{M(v)}}\\
&\cdot 2\hbar^{r'}(\Delta)^{i_{A(\mu_1)}j_{B(\nu_1)}}(\Delta)^{i_{A(\mu_2)}j_{B(\nu_2)}}\cdots(\Delta)^{i_{A(\mu_{r'})}j_{B(\nu_{r'})}}\\
&\cdot H^{I_{A(\mu_1,\mu_2)}J_{B(\nu_2,\nu_1)}}H^{I_{A(\mu_2,\mu_3)}J_{B(\nu_3,\nu_2)}}\cdots H^{I_{A(\mu_{r'},\mu_1)}J_{B(\nu_1,\nu_{r'})}}
\end{aligned}
\tag{47}
$$

with an arbitrary allowable I_a, an arbitrary allowable J_b, an arbitrary positive integer r', arbitrary sets of μ's and ν's such that $\mu_1<\mu_2<\cdots<\mu_{r'}$ and $(\nu_1>\nu_2>\cdots>\nu_{r'})$, and an arbitrary set $\mathcal{C}$ of choices of Δ's with an odd number of ω's. On the other hand, from Eqs. (21) and (31), $F(\{f^I,f^J\}_W)$ is a linear combination of all terms of the form

$$
\begin{aligned}
&2\mathrm{i}\left(-\frac{\mathrm{i}\hbar}{2}\right)^r\omega^{i_{\rho_1}j_{\sigma_1}}\cdots\omega^{i_{\rho_r}j_{\sigma_r}}\cdot\prod_{p=1}^{r}T^{\kappa^{(p)}_{a(-1)}\kappa^{(p)}_{a(1)}}T^{\kappa^{(p)}_{a(-2)}\kappa^{(p)}_{a(2)}}\cdots T^{\kappa^{(p)}_{a(-k)}\kappa^{(p)}_{a(k)}}\\
&\cdot\prod_{\substack{\text{distinct subloops } L\\ \text{of } f^K \text{ w.r.t. } K^{(p)}_a}}\operatorname{Tr}\eta'^{\kappa^{(p)}_{L(1)}}\eta'^{\kappa^{(p)}_{L(2)}}\cdots\eta'^{\kappa^{(p)}_{L(u)}}
\end{aligned}
\tag{48}
$$

with an arbitrary positive odd integer r, arbitrary sets of ρ's and σ's such that $\rho_1<\rho_2<\cdots<\rho_r$ and $(\sigma_1>\sigma_2>\cdots>\sigma_r)$, and an arbitrary allowable set of contracted indices $K^{(p)}_a$ in

$$
K^{(p)}=\begin{cases} I(\rho_p,\rho_{p+1})J(\sigma_{p+1},\sigma_p) & \text{for } 1\leqslant p\leqslant r-1\\ I(\rho_r,\rho_1)J(\sigma_1,\sigma_r) & \text{for } p=r.\end{cases}
\tag{49}
$$

Moreover, the indices of $K^{(p)}$ are $\kappa^{(p)}_1,\kappa^{(p)}_2,\ldots,$ etc.

It is possible to rewrite Eqs. (47) and (48) in the same form. Because of the emergence of a large number of new parameters, before writing out the new expression (Eq. (52) below), we would like to introduce these parameters first with the help of Fig. 4.

In Fig. 4, the oval object on the left, which is delineated by a thick closed line with an arrow, is the Weyl-ordered loop variable f^I. The oval object on the right is f^J. If we map them to

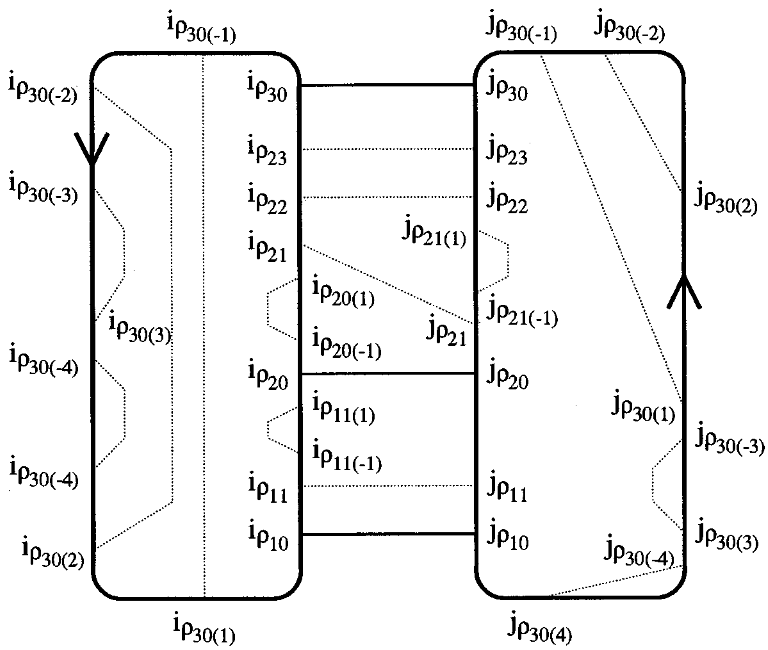

FIG. 4. An illustration for a typical term in Eq. (52). See the text for the legend of this figure.

normal-ordered loop variables and then take the Poisson bracket, we will obtain Eq. (47); if we take the Poisson bracket first and map the resultant expression to the space of normal-ordered loop variables later, we will obtain Eq. (48) instead. There are a number of ω^{ij}'s in both Eqs. (47) and (48). They will be labeled as $\omega^{i_{\rho_{10}}j_{\sigma_{10}}}$, $\omega^{i_{\rho_{20}}j_{\sigma_{20}}}$, $\ldots$, and $\omega^{i_{\rho_{r0}}j_{\sigma_{r0}}}$, where r is a positive *odd* integer. Moreover, it is always possible to arrange the indices in such a way that $i_{\rho_{10}}<i_{\rho_{20}}<\cdots<i_{\rho_{r0}}$ and $(j_{\sigma_{10}}>j_{\sigma_{20}}>\cdots>j_{\sigma_{r0}})$. We represent these ω's as solid lines joining the two oval objects in Fig. 4.

There are also a number of T^{ij}'s, where the index i comes from I and j comes from J, in both Eqs. (47) and (48). We will show in the following two lemmas that if they are generically labeled as $T^{i_{\rho_{x_1x_2}}j_{\sigma_{x_1x_2}}}$, where $1\leq x_1\leq r$ and x_2 is an integer between 1 and a certain positive integer s_{x_1}, in such a way that

$$(\rho_{10}<\rho_{11}<\cdots<\rho_{1s_1}<\rho_{20}<\rho_{21}<\cdots<\rho_{2s_2}<\cdots<\rho_{r0}<\rho_{r1}<\cdots<\rho_{rs_r}), \tag{50}$$

then

$$(\sigma_{10}>\sigma_{11}>\cdots>\sigma_{1s_1}>\sigma_{20}>\sigma_{21}>\cdots>\sigma_{2s_2}>\cdots>\sigma_{r0}>\sigma_{r1}>\cdots>\sigma_{rs_r}). \tag{51}$$

These T's are depicted as broken lines joining the two oval objects in Fig. 4.

There are other T^{ij} in both Eqs. (47) and (48) such that both i and j come from I. We will also show in the following two lemmas that they can be generically labeled as $T^{i_{a_{x_1x_2(-x_3)}}i_{a_{x_1x_2(x_3)}}}$, where x_3 is an integer between 1 and a certain positive integer $k_{x_1x_2}$ in such a way that $i_{a_{x_1x_2}(\pm x_3)}\in I(\rho_{x_1x_2},\rho_{x_1x_2+1})$ for $x_2<s_{x_1}$ or $i_{a_{x_1x_2}(\pm x_3)}\in I(\rho_{x_1x_2},\rho_{x_1+_m1,0})$ for $x_2=s_{x_1}$. These T's are depicted as broken lines within the left oval object in Fig. 4. There are still other T^{ij} in both equations such that both i and j come from J. Likewise, we will show that they can be generically labeled $T^{j_{b_{x_1x_2(-x_3)}}j_{b_{x_1x_2(x_3)}}}$, where x_3 is an integer between 1 and a certain positive integer $l_{x_1x_2}$ in such a way that $j_{b_{x_1x_2}(\pm y_3)}\in I(\sigma_{x_1x_2},\sigma_{x_1x_2+1})$ for $x_2<s_{x_1}$ or $j_{b_{x_1x_2}(\pm y_3)}\in I(\sigma_{x_1x_2},\sigma_{x_1+_m1,0})$ for $x_2=s_{x_1}$. These T's are depicted as broken lines within the right oval object in Fig. 4.

We are now ready to introduce the following lemmas.

Lemma 12: $i\{F(f^I),F(f^J)\}_N$ *is equal to a linear combination of all terms of the form*

$$2i\prod_{x_1=1}^{r}\left(-\frac{i\hbar}{2}\right)\omega^{i_{\rho_{x_1 0}}j_{\sigma_{x_1 0}}}\prod_{x_2=1}^{s_{x_1}}T^{i_{\rho_{x_1x_2}}j_{\sigma_{x_1x_2}}}\cdot\prod_{x_3=1}^{k_{x_1x_2}}T^{i_{a_{x_1x_2}(-x_3)}i_{a_{x_1x_2}(x_3)}}\prod_{y_3=1}^{l_{x_1x_2}}T^{\sigma_{b_{x_1x_2}(-y_3)}j_{b_{x_1x_2}(y_3)}}$$
$$\cdot\prod_{\substack{\text{distinct subloops } L \text{ within } I\\ \text{w.r.t. all contracted indices in } I}}\operatorname{Tr}\eta'^{i_{L(1)}}\eta'^{i_{L(2)}}\cdots\eta'^{i_{L(u)}}$$
$$\cdot\prod_{\substack{\text{distinct subloops } M \text{ within } J\\ \text{w.r.t. all contracted indices in } J}}\operatorname{Tr}\eta'^{j_{M(1)}}\eta'^{j_{M(2)}}\cdots\eta'^{j_{M(v)}}$$
$$\cdot\prod_{\substack{\text{distinct subloops } QR \text{ between } I \text{ and } J\\ \text{w.r.t. all contracted indices in } I \text{ and } J}}\operatorname{Tr}\eta'^{i_{Q(1)}}\eta'^{i_{Q(2)}}\cdots\eta'^{i_{Q(w_1)}}\cdot\eta'^{j_{R(1)}}\eta'^{j_{R(2)}}\cdots\eta'^{j_{R(w_2)}},\tag{52}$$

where

(1) *r is an arbitrary odd positive integer;*
(2) *the set of indices $\rho_{x_1x_2}$ is arbitrary except that these indices have to satisfy Eq. (50). Similarly, the set of indices $\sigma_{x_1x_2}$ is arbitrary except that these indices have to satisfy Eq. (51).*
(3) *the set of all $i_{a_{x_1x_2}(\pm x_3)}$'s is an arbitrary allowable set of contracted indices in I satisfying the stipulations in a paragraph on p. 42, and with $i_{a_{x_1x_2}(-x_3)}$ and $i_{a_{x_1x_2}(x_3)}$ forming a contraction pair. Similarly, the set of all $j_{b_{x_1x_2}(\pm y_3)}$'s is an arbitrary allowable set of contracted indices in J also satisfying the stipulations in the same paragraph, and with $j_{b_{x_1x_2}(-y_3)}$ and $j_{b_{x_1x_2}(y_3)}$ forming a contraction pair;*
(4) *u and v are the degrees of the subloops L and M, respectively. w_1 and w_2 are integers such that w_1+w_2 is the degree of the subloop QR; and*
(5) *the set of all indices in I belonging to any subloop QR comes from one subloop in I with respect to the set of all $i_{a_{x_1x_2}(\pm x_3)}$'s. Similarly, the set of all indices in J belonging to any subloop QR comes from one subloop in J with respect to the set of all $J_{b_{x_1x_2}(\pm y_3)}$'s.*

Proof: First of all, let us prove that any expression of the form shown in Eq. (47) can be rewritten in the manner shown in Eq. (52). Indeed, let r be the number of Δ's in Eq. (47) which are ω's. r is then a positive odd number (Statement (1)). By Corollary 2, $\mu_1<\mu_2<\cdots<\mu_{r'}$ implies $A(\mu_1)<A(\mu_2)<\cdots<A(\mu_{r'})$ and $(\nu_1<\nu_2<\cdots<\nu_{r'})$ implies $(B(\nu_1)<B(\nu_2)<\cdots<B(\nu_{r'}))$. Let us rename the indices $A(\mu_1),A(\mu_2),\ldots,A(\mu_{r'})$ and $B(\nu_1)<B(\nu_2),\ldots,B(\nu_{r'})$ by the following:

Algorithm 3: Here is the procedure of this algorithm.

Step 1: Set $x_1=0$ and $x_2=0$.

Step 2: Set $y=1$.

Step 3: If $(\Delta)^{i_{A(\mu_y)}j_{B(\nu_y)}}=T^{i_{A(\mu_y)}j_{B(\nu_y)}}$, then increment the value of x_2 by 1. Put $\rho_{x_1x_2}=A(\mu_y)$ and $\sigma_{x_1x_2}=B(\nu_y)$.

Step 4: If $(\Delta)^{i_{A(\mu_y)}j_{B(\nu_y)}}=\omega^{i_{A(\mu_y)}j_{B(\nu_y)}}$, then put $s_{x_1}=x_2$. Increment the value of x_1 by 1 and set the value of x_2 to 0. Put $\rho_{x_10}=A(\mu_y)$ and $\sigma_{x_10}=B(\nu_y)$.

Step 5: If $y\neq r'$, then increment the value of y by 1. Go back to Step 3.

Step 6: Put $s_r=x_2+s_0$.

Step 7: Set $x_2'=1$.

Step 8: If $x_2'>s_0$, then end the algorithm.

Step 9: Put $\rho_{r,x_2+x_2'}=\rho_{0,x_2'}$ *and* $\sigma_{r,x_2+x_2'}=\sigma_{0,x_2'}$.
Step 10: Increment the value of x_2' *by 1.*
Step 11: Go back to Step 8.

It is now clear that Eqs. (50) and (51) are true (Statement (2)). Consider $i_{a(-s)}$ and $i_{a(s)}$ in Eq. (47), where $1\le|s|\le k$. If $i_{a(-s)}\in I(\rho_{x_1x_2},\rho_{x_1,x_2+1})$ where $x_2<s_{x_1}$ but $i_{a(s)}\in I(\rho_{x_1,x_2+1},\rho_{x_1x_2})$, then $i_{\rho_{x_1x_2}}\in I(a(s),a(-s))$ and $i_{\rho_{x_1,x_2+1}}\in I(a(-s),a(s))$. Thus two subloops of f^I are involved to produce the γ's by Lemma 7 and then the Δ's in the Poisson bracket with f^J. However, only one subloop of f^I, namely A in Eq. (47), should be involved and this leads to a contradiction. Hence, if $i_{a(-s)}\in I(\rho_{x_1x_2},\rho_{x_1,x_2+1})$, then $i_{a(s)}\in I(\rho_{x_1x_2},\rho_{x_1,x_2+1})$ also. Similarly, if $i_{a(-s)}\in I(\rho_{x_1,x_2+1},\rho_{x_1x_2})$, then $i_{a(s)}\in I(\rho_{x_1,x_2+1},\rho_{x_1x_2})$ also. In addition, $i_{a(-s)}$ and $i_{a(s)}\in I(\rho_{x_1s_{x_1}},\rho_{x_1+_r1,0})$ or $i_{a(-s)}$ and $i_{a(s)}\in I(\rho_{x_1+_r1,0},\rho_{x_1s_{x_1}})$. Let $a_{x_1x_2}(\pm 1),a_{x_1x_2}(\pm 2),\ldots,a_{x_1x_2}(\pm k_{x_1x_2})$ be those $a(\pm s)$'s such that $a(\pm s)\in I(\rho_{x_1x_2},\rho_{x_1,x_2+1})$ for $x_2<s_{x_1}$ or $a(\pm s)\in I(\rho_{x_1s_{x_1}},\rho_{x_1+_r1,0})$ for $x_2=s_{x_1}$, and let the $b_{x_1x_2}(\pm x_3)$'s have analogous definitions. These definitions of $a_{x_1x_2}(x_3)$'s and $b_{x_1x_2}(y_3)$'s are consistent with the ones given in the paragraph preceding this Lemma. Furthermore, Statement (3) should be clear. Statements (4) and (5) are direct consequences of Eq. (47).

The subloops L and M are still defined by using Algorithm 1. From Eqs. (47) and (52), every QR lies within $I(\rho_{x_1x_2},\rho_{x_1x_2+1})$ and $J(\sigma_{x_1x_2+1},\sigma_{x_1x_2})$ for $x_2<s_{x_1}$ or within $I(\rho_{x_1s_{x_1}},\rho_{x_1+_r1,0})$ and $J(\sigma_{x_1s_{x_1}},\sigma_{x_1+_r1,0})$. Let $Q(0)=\rho_{x_1x_2}$. If $Q(v)$ is defined for an integer v, then we define $Q^{(i)}(v+1)$ for some integers i as follows:

Algorithm 4: Here is the procedure of this algorithm.
Step 1: Set $i=1$.
Step 2: $Q^{(i)}(v+1)=Q(v)+_m1$.
Step 3: If $Q^{(i)}(v+1)=\rho_{x_1x_2}\exists x_1$ *and* x_2, *then jump to Step 9.*
Step 4: If $Q^{(i)}(v+1)\neq a_{x_1x_2}(x_3)\forall x_1,x_2,x_3$ *(where* x_3 *can be positive or negative), then end the algorithm.*
Step 5: Let $\{x_1,x_2,x_3\}$ *be a set of numbers such that* $a_{x_1x_2}(x_3)=Q^{(i)}(v+1)$.
Step 6: Increment the value of i by 1.
Step 7: Set $Q^{(i)}(v+1)=a_{x_1x_2}(-x_3)+_m1$.
Step 8: Go back to Step 3.
Step 9: Let $\{x_1,x_2\}$ *be a set of numbers such that* $\rho_{x_1x_2}=Q^{(i)}(v+1)$.
Step 10: Increment the value of i by 1.
Step 11: Set $Q^{(i)}(v+1)=\sigma_{x_1x_2}$.
Step 12: End the algorithm.

If the algorithm was ended in Step 4, then define $Q(v+1)=Q^{(i)}(v+1)$; if the algorithm was ended in Step 12, then define $R(0)=Q^{(i)}(v+1)$. If $R(v)$ is defined for an integer v, then we define $R^{(i)}(v+1)$ for some integers i by *Algorithm 5*, which is the same as Algorithm 4 except that $Q^{(i)}(v+1)$ is changed to $R^{(i)}(v+1)$, $+_m$ to $+_n$, $\rho_{x_1x_2}$ to $\sigma_{x_1x_2}$, $a_{x_1x_2}(x_3)$ to $b_{x_1x_2}(y_3)$, and $\sigma_{x_1x_2}$ to $\rho_{x_1x_2}$. Then, if Algorithm 5 was ended in Step 4, we will define $R(v+1)=R^{(i)}(v+1)$; if it was ended in Step 12, then this $R^{(i)}(v+1)$ should be exactly $Q(0)$.

Conversely, now let us prove that any expression in the form shown in Eq. (52) and satisfying the five ensuing statements can be rewritten in the way shown in Eq. (47). It should be clear by a reversal of the procedure described earlier in this proof that we can rewrite the $-\mathrm{i}\omega^{i_{\rho_{x_10}}j_{\sigma_{x_10}}}/2$'s and $T^{i_{\rho_{x_1x_2}}j_{\sigma_{x_1x_2}}}/\hbar$'s as $(\Delta)^{i_{A(\mu_x)}j_{B(\nu_x)}}$'s with $\mu_1<\mu_2<\cdots<\mu_{r'}$ and $(\nu_1>\nu_2>\cdots>\nu_{r'})$. Moreover, the $T^{i_{a_{x_1x_2}(-x_3)}i_{a_{x_1x_2}(x_3)}}$'s can be rewritten as $T^{i_{a(-x)}i_{a(x)}}$'s, and the $T^{j_{b_{x_1x_2}(-y_3)}j_{b_{x_1x_2}(y_3)}}$'s can be rewritten as $T^{j_{b(-y)}j_{b(y)}}$'s. If we can show that the set of all $i_{\rho_{x_1x_2}}$'s come from one subloop with respect to the set of $i_{a_{x_1x_2}(x_3)}$'s in I, and the set of all $j_{\sigma_{x_1x_2}}$'s come from one subloop with respect to the set of $j_{b_{x_1x_2}(y_3)}$'s in J, then the subloops QR can be rewritten as those H's in Eq. (47).

To show that all $i_{\rho_{x_1x_2}}$'s come from one subloop, we need to show that if $L(v')=\rho_{x_1x_2}$ for some values of x_1 and x_2, then there exists an integer v such $L(v)=\rho_{x_1x_2+1}$ for $x_2<s_{x_1}$ or $L(v)=\rho_{x_1+_r1,0}$ for $x_2=s_{x_1}$. We will write $L(v)=\rho_{x_1x_2+1}$ generically. We can set $v'=1$ without loss of generality by Lemma 6. Obviously $(L(1)<\rho_{x_1x_2+1})$. Assume that $(L(1)<L(2)<\cdots L(v-1)<\rho_{x_1x_2+1})$ for an integer $v>1$, and assume that $L(v)$ does not exist. Then there should be an $L^{\iota'}(v)$ for an integer $\iota'\geqslant 1$ such that $L^{(\iota')}(v)=a_{\tilde{x}_1\tilde{x}_2}(-\tilde{x}_3)$ for some integers $\tilde{x}_1,\tilde{x}_2,\tilde{x}_3$ and $a_{\tilde{x}_1\tilde{x}_2}(\tilde{x}_3)+_m1=L(1)$. Thus $(a_{\tilde{x}_1\tilde{x}_2}(\tilde{x}_3)<L(1)<L^{(1)}(v)<\rho_{x_1x_2+1})$. Hence there exists a smallest integer ι such that $(a_{x_1x_2}(x_3)<L(1)<L^{(\iota)}(v)<\rho_{x_1x_2+1})$, where $a_{x_1x_2}(-x_3)=L^{(\iota)}(v)$. However, this is impossible and so $L(v)$ exists. Assume $(\rho_{x_1x_2+1}<L(v)<L(1))$. Then there should be an $L^{(\iota')}(v)$ for an integer $\iota'\geqslant 1$ such that $L^{(\iota')}(v)=a_{\tilde{x}_1\tilde{x}_2}(-\tilde{x}_3)$ for some integers $\tilde{x}_1,\tilde{x}_2,\tilde{x}_3$ and $a_{\tilde{x}_1\tilde{x}_2}(\tilde{x}_3)+_m1=L(v)$. Hence $(L^{(1)}(v)<\rho_{x_1x_2+1}<a_{\tilde{x}_1\tilde{x}_2}(x_3)<L(1))$. This implies the existence of a smallest integer ι such that $L^{(\iota)}(v)=a_{x_1x_2}(-x_3)$ for an integer x_3 and $(L^{(\iota)}\times(v)<\rho_{x_1x_2+1}<a_{x_1x_2}(x_3)<L(1))$. Again this is impossible. Hence $(L(1)<L(v)\leqslant\rho_{x_1x_2+1})$. By Corollary 2, $(L(1)<L(2)<\cdots<L(v)\leqslant\rho_{x_1x_2+1})$. Since there are only a finite number of indices between $L(1)$ and $\rho_{x_1x_2+1}$, there exists a number $\tilde{v}$ such that $L(\tilde{v})=\rho_{x_1x_2+1}$. Hence $i_{\rho_{x_1x_2}}$ and $i_{\rho_{x_1x_2+1}}$ belong to the same subloop of I. Consequently, the set of all $i_{\rho_{x_1x_2}}$'s for $1\leqslant x_1\leqslant r$ and $0\leqslant x_2\leqslant s_{x_1}$ belongs to one subloop of I. Similarly, the set of all $j_{\sigma_{x_1x_2}}$'s belongs to one subloop of J. Q.E.D.

Before we prove Lemma 13, we remark that in the following, by $I(\rho_{x_10},\rho_{x_1+_r1,0})J(\sigma_{x_1+_r1,0},\sigma_{x_10})(\rho_{x_1x_2},\sigma_{x_1x_2})$ we mean the sequence $i_{\rho_{x_1x_2}+1},i_{\rho_{x_1x_2}+2},\ \ldots,i_{\rho_{x_1+_r1,0}},j_{\sigma_{x_1+_r1,0}},\ldots,j_{\sigma_{x_1x_2}}$.

Lemma 13: $F(\{f^I,f^j\}_W)$ can also be written as a linear combination of all terms of the form shown in Eq. (52) with the five accompanying statements.

Proof: First of all, let us show that any expression of the form given in Eq. (48) can be rewritten as shown in Eq. (52) with the accompanying five statements being satisfied. Indeed, Statement (1) is obvious. Let us rename the indices $\rho_1,\rho_2,\ \ldots,\rho_r,\sigma_1,\sigma_2,\ \ldots,\sigma_r$ in Eq. (48) as $\rho_{10},\rho_{20},\ \ldots,\rho_{r0},\sigma_{10},\sigma_{20},\ \ldots,\sigma_{r0}$. Moreover, within the loop $f^{I(\rho_{x_10},\rho_{x_1+_r1,0})J(\sigma_{x_1+_r1,0},\sigma_{x_10})}$, for those contraction pairs with one index coming from I and the other one from J, call these indices $i_{\rho_{x_1x_2}}$ and $j_{\sigma_{x_1x_2}}$ in such a way that $\rho_{x_10}<\rho_{x_11}<\cdots<\rho_{x_1s_{x_1}}<\rho_{x_1+1,0}$ if there are s_{x_1} such pairs for $x_1<r$, or $(\rho_{r0}<\rho_{r1}<\cdots<\rho_{rs_r}<\rho_{10})$ if $x_1=r$. Thus Eq. (50) is true. Obviously $(\sigma_{x_1+_r1,0}>\sigma_{x_11}>\sigma_{x_10})$. Assume that $(\sigma_{x_1+_r1,0}>\sigma_{x_1x_2}>\sigma_{x_1x_2-1}>\cdots>\sigma_{x_10})$. If $x_2\neq s_{x_1}$, consider σ_{x_1,x_2+1}. Since $i_{\rho_{x_1x_2+1}}\in I(\rho_{x_10},\rho_{x_1+_r1,0})J(\sigma_{x_1+_r1,0},\sigma_{x_10})(\rho_{x_1x_2},\sigma_{x_1x_2})$, we get $j_{\sigma_{x_1,x_2+1}}\in I(\rho_{x_10},\rho_{x_1+_r1,0})J(\sigma_{x_1+_r1,0},\sigma_{x_10})(\rho_{x_1x_2},\sigma_{x_1x_2})$ also. Hence $(\sigma_{x_1+_r1,0}>\sigma_{x_1,x_2+1}>\sigma_{x_1x_2}>\sigma_{x_1,x_2-1}>\cdots>\sigma_{x_10})$. Thus $\sigma_{x_1+_r1,0}>\sigma_{x_1s_{x_1}}>\sigma_{x_1,s_{x_1}-1}>\cdots>\sigma_{x_10}$. Therefore Statement (2) is true. Statements (3) and (4) are obvious.

Now let us fix the values of x_1 and x_2, and consider those contraction pairs with both indices coming from I, and one of them, say i_c, belonging to $I(\rho_{x_1x_2},\rho_{x_1,x_2+1})$ for $x_2<s_{x_1}$ or $I(\rho_{x_1x_2},\rho_{x_1+_r1,0})$ for $x_2=s_{x_1}$. Let $i_{c'}$ be the other index of this contraction pair. Clearly $i_{c'}\in I(\rho_{x_10},\rho_{x_1+_r1,0})$. If i'_c does not belong to $I(\rho_{x_1x_2},\rho_{x_1,x_2+1})$ for $x_2<s_{x_1}$ or $I(\rho_{x_1x_2},\rho_{x_1+_r1,0})$ for $x_2=s_{x_1}$, then $i_c\in I(\rho_{x_10},\rho_{x_1+_r1,0})J(\sigma_{x_1+_r1,0},\sigma_{x_10})(\rho_{x_1x_2},\sigma_{x_1x_2})$ but $i_{c'}\in J(\sigma_{x_1+_r1,0},\sigma_{x_10})I(\rho_{x_10},\rho_{x_1+_r1,0})(\sigma_{x_1x_2},\rho_{x_1x_2})$, which is impossible. Thus any contraction pairs coming only from I can be written as $i_{a_{x_1x_2}(\pm x_3)}$ because both of them must belong to a sequence $I(\rho_{x_1x_2},\rho_{x_1x_2+1})$ with $x_2<s_{x_1}$, or belong to a sequence $I(\rho_{x_1x_2},\rho_{x_1+_m1,0})$ with $x_2=s_{x_1}$.

We can now say that for a fixed value of x_1, the set of all $i_{\rho_{x_1x_2}}$, $j_{\sigma_{x_1x_2}}$, $i_{a_{x_1x_2}(\pm x_3)}$ and $j_{b_{x_1x_2}(\pm y_3)}$, where $1\leqslant x_2\leqslant s_{x_1}$, $1\leqslant x_3\leqslant k_{x_1x_2}$ and $1\leqslant y_3\leqslant l_{x_1x_2}$, together form an allowable set of contracted indices in the loop $I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})$ with $i_{\rho_{x_1 x_2}}$ and $j_{\sigma_{x_1 x_2}}$ being contraction pairs, $i_{a_{x_1 x_2}(x_3)}$ and $i_{a_{x_1 x_2}(-x_3)}$ being contraction pairs, and $j_{b_{x_1 x_2}(y_3)}$ and $j_{b_{x_1 x_2}(-y_3)}$ being contraction pairs.

The remaining thing to do to prove that Statement (5) is satisfied is to show that the set of all $i_{a_{x_1 x_2}(\pm x_3)}$'s for $1 \leqslant x_1 \leqslant r$, $0 \leqslant x_2 \leqslant s_{x_1}$ and $1 \leqslant x_3 \leqslant k_{x_1 x_2}$ is allowable in I, and the set of all $j_{b_{x_1 x_2}(\pm y_3)}$'s for $1 \leqslant x_1 \leqslant r$, $0 \leqslant x_2 \leqslant s_{x_1}$ and $1 \leqslant y_3 \leqslant l_{x_1 x_2}$ is allowable in J. Indeed, since for each fixed x_1, the set of all $i_{a_{x_1 x_2}(\pm x_3)}$'s is allowable in $I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})$, this set alone is allowable in I also. Now let us choose a fixed set of values for x_1, x_2 and x_3. Consider two integers $\tilde{x}_1$ and $\tilde{x}_2$ such that $1 \leqslant \tilde{x}_1 \leqslant r$ and $1 \leqslant \tilde{x}_2 \leqslant s_{\tilde{x}_1}$. In addition, either $\tilde{x}_1 \neq x_1$ or $\tilde{x}_2 \neq x_2$, or both. Then $i_{a_{x_1 x_2}(\pm x_3)} \in I(a_{\tilde{x}_1 \tilde{x}_2}(\tilde{x}_3), a_{\tilde{x}_1 \tilde{x}_2}(-\tilde{x}_3))$ for $1 \leqslant \tilde{x}_3 \leqslant k_{\tilde{x}_1 \tilde{x}_2}$ if $a_{\tilde{x}_1 \tilde{x}_2}(\tilde{x}_3) \geqslant a_{\tilde{x}_1 \tilde{x}_2}(-\tilde{x}_3)$ and $\rho_{\tilde{x}_1 0} < \rho_{\tilde{x}_1+_r 1,0}$, or if $a_{\tilde{x}_1 \tilde{x}_2}(\tilde{x}_3) > a_{\tilde{x}_1 \tilde{x}_2}(-\tilde{x}_3) > \rho_{\tilde{x}_1 0} > \rho_{\tilde{x}_1+_r 1,0}$, or if $\rho_{\tilde{x}_1 0} > \rho_{\tilde{x}_1+_r 1,0} > a_{\tilde{x}_1 \tilde{x}_2}(\tilde{x}_3) > a_{\tilde{x}_1 \tilde{x}_2}(-\tilde{x}_3)$. Similarly, $i_{a x_1 x_2(\pm x_3)} \in I(a_{\tilde{x}_1 \tilde{x}_2}(-\tilde{x}_3), a_{\tilde{x}_1 \tilde{x}_2}(\tilde{x}_3))$ if $a_{\tilde{x}_1 \tilde{x}_2}(\tilde{x}_3) > \rho_{\tilde{x}_1 0} > \rho_{\tilde{x}_1+_r 1,0} > a_{\tilde{x}_1 \tilde{x}_2}(-\tilde{x}_3)$. As a result, the set of all $i_{a_{x_1 x_2}(\pm x_3)}$'s is allowable in I. A similar argument holds for all $j_{b_{x_1 x_2}(\pm y_3)}$'s in J.

The subloops L and M are found by using Algorithm 1, and the subloops QR can again be determined by Algorithms 4 and 5.

Let us consider the converse, i.e., whether any expression of the form shown in Eq. (52) and satisfying the five accompanying statements is a term in Eq. (48). The only thing we need to do to substantiate this converse statement is to prove that if the set of all $i_{a_{x_1 x_2}(\pm x_3)}$'s is allowable in I, the set of all $j_{b_{x_1 x_2}(\pm y_3)}$'s is allowable in J, the set of all $i_{\rho_{x_1 x_2}}$'s satisfies Eq. (50) and the set of all $j_{\sigma_{x_1 x_2}}$'s satisfies Eq. (51), then for each fixed x_1, the set of all $i_{a_{x_1 x_2}(\pm x_3)}$'s and $j_{b_{x_1 x_2}(\pm y_3)}$'s together with $i_{\rho_{x_1 x_2}}$ and $j_{\sigma_{x_1 x_2}}$ is allowable in $f^{I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})}$.

Indeed, let all the five statements accompanying Eq. (52) be satisfied. Obviously the pair $i_{\rho_{x_1 1}}$ and $j_{\sigma_{x_1 1}}$ forms an allowable set of contraction pairs in the loop $I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})$. Assume that $(i_{\rho_{x_1 1}}, j_{\sigma_{x_1 1}}, i_{\rho_{x_1 2}}, j_{\sigma_{x_1 2}}, \ldots, i_{\rho_{x_1 x_2}}, j_{\sigma_{x_1 x_2}})$ is allowable. If $x_2 < s_{x_1}$, consider the set $(i_{\rho_{x_1 1}}, j_{\sigma_{x_1 1}}, i_{\rho_{x_1 2}}, j_{\sigma_{x_1 2}}, \ldots, i_{\rho_{x_1 x_2+1}}, j_{\sigma_{x_1 x_2+1}})$. It is clear that $i_{\rho_{x_1 x_2+1}} \in I(\rho_{x_1 x_2},\rho_{x_1+_r 1,0})$ and $j_{\sigma_{x_1 x_2+1}} \in J(\sigma_{x_1+_r 1,0},\sigma_{x_1 x_2})$ from Eqs. (50) and (51). Hence both $i_{\rho_{x_1 x_2+1}}$ and $j_{\sigma_{x_1 x_2+1}} \in I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})(\rho_{x_1 \tilde{x}_2},\sigma_{x_1 \tilde{x}_2}) \forall \tilde{x}_2 = 1, 2, \ldots,$ and x_2. As a result, $(i_{\rho_{x_1 1}}, j_{\sigma_{x_1 1}}, i_{\rho_{x_1 2}}, j_{\sigma_{x_1 2}}, \ldots, i_{\rho_{x_1 s_{x_1}}}, j_{\sigma_{x_1 s_{x_1}}})$ is allowable in $I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})$.

Let us turn to $i_{a_{x_1 x_2}(\pm x_3)}$'s for the fixed x_1 we are considering. Since $i_{a_{x_1 x_2}(\pm x_3)} \in I(\rho_{x_1 x_2},\rho_{x_1, x_2+1})$ for $x_2 < s_{x_1}$ or $I(\rho_{x_1 x_2},\rho_{x_1+_r 1,0})$ for $x_2 = s_{x_1}$, we have $i_{a_{x_1 x_2}(\pm x_3)} \in I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})(\rho_{x_1 \tilde{x}_2},\sigma_{x_1 \tilde{x}_2})$ for $\tilde{x}_2 \leqslant x_2$, or $i_{a_{x_1 x_2}(\pm x_3)} \in I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})(\sigma_{x_1 \tilde{x}_2},\rho_{x_1 \tilde{x}_2})$ for $\tilde{x}_2 > x_2$. Moreover, the set of all $i_{a_{x_1 x_2}(\pm x_3)}$'s for the fixed x_1 is allowable in I, so this set of $i_{a_{x_1 x_2}(\pm x_3)}$'s alone is also allowable in $I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})$. Furthermore, $i_{a_{x_1 x_2}(\pm x_3)} \in I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0}) \times (b_{x_1 \tilde{x}_2(-y_3)}, b_{x_1 \tilde{x}_2(y_3)})$ for $0 \leqslant \tilde{x}_2 \leqslant s_{x_r}$ and $(\sigma_{x_1+_r 1,0} < b_{x_1 \tilde{x}_2(-y_3)} < b_{x_1 \tilde{x}_2(y_3)} < \sigma_{x_1 0})$. A similar argument applies to the set of all $j_{b_{x_1 x_2}(\pm y_3)}$'s for the fixed x_1. Consequently, for each fixed x_1, the set of all $i_{a_{x_1 x_2}(\pm x_3)}, j_{b_{x_1 x_2}(\pm y_3)}, i_{\rho_{x_1 x_2}}, j_{\sigma_{x_1 x_2}}$'s is allowable in $I(\rho_{x_1 0},\rho_{x_1+_r 1,0})J(\sigma_{x_1+_r 1,0},\sigma_{x_1 0})$. Q.E.D.

Theorem 1: *There exists an invertible Poisson morphism between the Poisson algebra of Weyl-ordered loop variables and the Poisson algebra of normal-ordered loop variables.*

Proof: This is a direct consequence of Lemmas 9, 12 and 13. Q.E.D.

ACKNOWLEDGMENTS

We would like to thank O.T. Turgut for useful discussions. Indeed, C.W.H.L. is indebted to him for explaining the content of Ref. 3 in detail. C.W.H.L. would also like to thank C. Macesanu for his technical help in the use of LaTeX at an early stage of this work. This work was supported in part by the U.S. Department of Energy under Grant No. DE-FG02-91ER40685.

[1] G. 't Hooft, Nucl. Phys. B **75**, 461 (1974); E. Witten, *ibid.* **160**, 57 (1979).
[2] L. Yaffe, Rev. Mod. Phys. **54**, 407 (1982).
[3] S. G. Rajeev and O. T. Turgut, J. Math. Phys. **37**, 637 (1996).
[4] S. G. Rajeev and O. T. Turgut, Int. J. Mod. Phys. A **10**, 2479 (1995).
[5] M. Flato, A. Lichnerowicz, and D. Sternheimer, C. R. Acad. Sci. Paris Sér. A **279**, 877 (1974); Compositio Mathematica **31**, 47 (1975); J. Math. Pures Appl. **54**, 445 (1975).
[6] L. Hormander, Commun. Pure Appl. Math. **32**, 359 (1979); G. B. Folland, *Harmonic Analysis in Phase Space* (Princeton University Press, Princeton, 1989).
[7] V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge University Press, Cambridge, 1994).
[8] M. Flato, hep-th/9809073, in *Particles, Fields and Gravitation*, edited by J. Rembieliński (AIP, New York, to be published); M. Flato and D. Sternheimer, *The Deformation Programme in Physical Theories*, Lectures at the Mittag–Leffler Institute in the program Topology and Geometry of Quantum Fields, October 22, 1998 (unpublished).
[9] C.-W. H. Lee and S. G. Rajeev, Phys. Rev. Lett. **80**, 2285 (1998).
[10] C.-W. H. Lee and S. G. Rajeev, Nucl. Phys. B **529**, 656 (1998).
[11] C.-W. H. Lee and S. G. Rajeev, J. Math. Phys. **39**, 5199 (1998).
[12] J. L. Gross and T. W. Tucker, *Topological Graph Theory* (Wiley–Interscience, New York, 1987).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Differential equations for scaling relation in $N=2$ supersymmetric SU(2) Yang–Mills theory coupled with massive hypermultiplet

Yűji Ohta
Research Institute for Mathematical Sciences, Kyoto University, Sakyoku, Kyoto 606, Japan

(Received 7 October 1998; accepted for publication 12 January 1999)

Differential equations for the scaling relation of prepotential in $N=2$ supersymmetric SU(2) Yang–Mills theory coupled with massive matter hypermultiplet are proposed and are explicitly demonstrated in one flavor ($N_f=1$) theory. By applying Whitham dynamics, the first-order derivative of the prepotential over the T_0 variable corresponding to the mass of the hypermultiplet, which has a line integral representation, is found to satisfy a differential equation. As a result, the closed form of this derivative can be obtained by solving this equation. In this way, the scaling relation of massive prepotential is established. Furthermore, as an application of another differential equation for the massive scaling relation, the massive prepotential in a strong coupling region is derived. © *1999 American Institute of Physics.* [S0022-2488(99)01704-1]

I. INTRODUCTION

It is well-known that the low energy effective action of $N=2$ supersymmetric Yang–Mills theory is described in terms of the holomorphic prepotential $\mathcal{F}$.[1] In this case, the perturbative part of the prepotential is not modified beyond one-loop order according to the nonrenormalization theorem,[2–5] but is known to be affected by instantons. In the case of the SU(2) gauge group, Seiberg and Witten[6,7] proposed a general prescription to determine the nonperturbative prepotential with the aid of Riemann surface of genus one. Based on their observation, Klemm *et al.*[8] determined the instanton corrected prepotential by adopting a method of Picard–Fuchs equation, which was often used in the mirror symmetry of Calabi–Yau manifold.[9–11] On the other hand, Matone[12] derived a recurrence formula of the instanton expansion coefficients of the prepotential by noticing a modularity. As a bonus, he obtained a quite simple relation between prepotential and moduli often referred as a scaling relation. After this discovery, the existence of such relation in $N=2$ supersymmetric Yang–Mills theory coupled with or without massive quark hypermultiplets was pointed out independently by two groups.[13,14] Sonnenschein *et al.*[13] proved it by noticing the homogeneity of the prepotential, while Eguchi and Yang[14] established the same result (2.17) in the language of Whitham dynamics.[15,16] However, in the framework of Whitham dynamics, we always encounter the problem of evaluating the derivative of $\mathcal{F}$ over the T_0 variable in order to establish a scaling relation when the mass of the hypermultiplets is not ignored. In the Whitham theory, this quantity is represented by a line integral interpolating two coverings of Riemann surface and therefore its evaluation is quite complicated, especially, in a massive case. In fact, explicit calculation of this integral is not found in the literature.

One of the aims of the paper is to give a solution to this problem. However, since the calculation in the case of a theory with N_f massive flavors is very complicated even for SU(2), we treat only SU(2) $N_f=1$ case. In addition, extension to $N_f>1$ is straightforward, so it is recommended that the reader try to proceed to other cases. Our construction starts from comparing the Wronskian of a massive Picard–Fuchs equation with the modular invariant of Matone.[12] These two are shown to be related by a Fuchsian differential equation. Applying the Whitham theory of soliton to this equation, we can obtain a differential equation for $\partial\mathcal{F}/\partial T_0$. In this way, it is determined as a solution to this equation. The details are discussed in Sec. II. On the other hand, in the course of the calculus used in the derivation of this differential equation, we can find another simple differential equation, which indicates a relation between prepotential and moduli like that in five dimensions.[17] This equation is also a consequence of scaling relation in the massive theory. As an application, the prepotential in the strong coupling region (dual prepotential) is derived in Sec. III. At first sight, the dual prepotential looks very complicated, but the massless limit coincides with that in the massless theory obtained by Ito and Yang.[18] Section IV is a brief summary.

II. SCALING RELATION AND THE WHITHAM HIERARCHY

A. The Picard–Fuchs equation

First of all, let us recall the basics of the massive $N_f=1$ theory in the Seiberg–Witten approach.[7] In the case of SU(2), we can take two kinds of curves, one of which is elliptic type[7] and the other is hyperelliptic type.[19] Though an elliptic curve is used in the next section, here we take the hyperelliptic curve. In this case, it is given by

$$y^2=(x^2-u)^2-\Lambda^3(x+m), \tag{2.1}$$

where $u=\langle \mathrm{tr}\,\phi^2\rangle$ is the moduli (ϕ is the complex scalar field of $N=2$ chiral superfield), m is the mass of the hypermultiplet, and Λ is the mass scale parameter of this $N_f=1$ theory. For this curve, the Seiberg–Witten differential 1-form λ is determined from the basic relation $d\lambda/du \propto dx/y$. Taking into account the numerical normalization factor, we find[19,20]

$$\lambda=\frac{\sqrt{2}}{4\pi i}\frac{x\,dx}{y}\left[\frac{x^2-u}{2(x+m)}-2x\right]. \tag{2.2}$$

In general, the Seiberg–Witten 1-form in a theory involving massive matter hypermultiplets in the fundamental representation of the gauge group is endowed with a pole structure whose residue is linearly proportional to the masses of the hypermultiplets.[7,19]

The Seiberg–Witten ansatz requires that the vacuum expectation value of ϕ and its dual are quantum mechanically given by the two periods

$$a=\oint_\alpha \lambda, \quad a_D=\oint_\beta \lambda, \tag{2.3}$$

respectively, along the canonical basis ($\alpha\cap\beta=+1$) of 1-cycles on (2.1). Then the periods satisfy the Picard–Fuchs equation

$$\frac{d^3\Pi}{du^3}+X\frac{d^2\Pi}{du^2}+Y\frac{d\Pi}{du}=0, \tag{2.4}$$

where

$$X=\frac{d}{du}\ln\frac{\Delta}{4m^2-3u},\qquad Y=-\frac{8}{\Delta}\left[4(2m^2-3u)+3\,\frac{3m\Lambda^3-4u^2}{4m^2-3u}\right]. \tag{2.5}$$

Here,

$$\Delta(u)=256u^3-256m^2u^2-288m\Lambda^3u+256m^3\Lambda^3+27\Lambda^6 \tag{2.6}$$

is the discriminant of the curve (2.1).

B. Differential equations for scaling relation

Differential equations for the Wronskian can be used to make a relation to prepotential. The reader might know the example in the context of mirror symmetry presented by Candelas *et al.*,[10,11] who used the ''Wronskian'' in order to make contact with Yukawa coupling in a complex structure moduli space. A similar presentation is possible also in $N=2$ supersymmetric Yang–Mills theory coupled with or without massless matter hypermultiplets. In the case of SU(2), the integral of Wronskian of Picard–Fuchs equation yields the scaling relation of the prepotential.[14] However, when hypermultiplets are massive, it is not easy to see such a simple relation connecting prepotential and moduli. We encountered a similar problem in the five-dimensional gauge theory.[17] Since we did not know a five-dimensional analog of the scaling relation in four-dimensional gauge theory, we proposed a differential relation between prepotential and the Wronskian. The method used in the course of this calculation reveals further aspects of the scaling relation, provided it is applied to four-dimensional gauge theory.

For this, let us prepare the following two quantities:

$$W=a'a_D''-a''a_D', \quad w=aa_D'-a'a_D, \tag{2.7}$$

where $'=d/du$. The first equation is precisely the Wronskian for the third-order Picard–Fuchs equation (2.4), while the integration of the second one produces the modular invariant of Matone[12]

$$\int_0 w\,du=aa_D-2\mathcal{F}. \tag{2.8}$$

Remark: The Modular invariance mentioned here is the sense of pure SU(2) theory.

Here, the integral symbol indicates that it is an integration constant free integral, namely, the integration constant is set to zero. What we would like to clarify is a relation between (2.8) and (integration of) W, so the first task is to try to connect W and w.

Fortunately, this is simply done by differentiating w over u repeatedly modulo the Picard–Fuchs equation (2.4), and we can find

$$w''+Xw'+Yw=W. \tag{2.9}$$

It is interesting to notice that w satisfies a Picard–Fuchs equation with W as a source term. On the other hand, for W we can easily obtain

$$W'+XW=0, \tag{2.10}$$

which implies

$$W=c\,\frac{4m^2-3u}{\Delta}, \tag{2.11}$$

where c is an integration constant. Note that the discriminant of the curve appears in the denominator and a similar relation was noticed in five dimensions.[17] In order to fix c, we may use a massless or double scaling limit as a boundary condition. For example, in the massless limit, it is well-known that $w=i3/(4\pi)$,[21] therefore, from (2.9)

$$c=-i\,\frac{16}{\pi}. \tag{2.12}$$

Of course, c must be uniquely fixed, and the same value is obtained from double scaling limit. (2.11) with (2.12) plays an important role in the next section.

In this way, we arrive at

$$w''+Xw'+Yw=-i\frac{16}{\pi}\frac{4m^2-3u}{\Delta}. \tag{2.13}$$

Since the solution to this equation gives w and must be related to (2.8), w obtained from (2.13) is expected to give a scaling relation in this massive theory[13]

$$a\frac{\partial\mathcal{F}}{\partial a}-2\mathcal{F}=-m\frac{\partial\mathcal{F}}{\partial m}-\Lambda\frac{\partial\mathcal{F}}{\partial\Lambda}, \tag{2.14}$$

where $\mathcal{F}$ is regarded here as a homogeneous function $\mathcal{F}=\mathcal{F}(a,\Lambda,m)$ in variables. Specifically, the solution of (2.13) will include information on the right-hand side of (2.14) and what we would like to do next is extract such information from (2.13). The best way to accomplish this is to introduce the Whitham theory.[15,16]

C. Relation to Whitham dynamics

Let us consider a consequence of (2.13) in view of Whitham dynamics in $N=2$ Yang–Mills theory. For details of Whitham dynamics in the context of Yang–Mills theory, see Refs. 14–16.

Let T_n $(n\in\mathbf{N}\cup\{0\})$ be time variables coupled to $(n+1)$th order pole of the Seiberg–Witten differential 1-form. In the SU(2) $N_f=1$ theory, the prepotential is available from the relation

$$\frac{\partial\mathcal{F}}{\partial a}=\oint_\beta\lambda,\quad \frac{\partial\mathcal{F}}{\partial T_n}=-2\pi i\,\mathrm{res}(z^{-n}\lambda),\quad \frac{\partial\mathcal{F}}{\partial T_0}=-2\pi i\int_{z_*=-m}^{z=-m}\lambda, \tag{2.15}$$

where residue is evaluated at $x=1/z=\infty$ and $z_*(=1/x)$ is the coordinate on the other sheet of the curve. At $x=\infty$, λ is expanded as

$$\lambda=\left[-\sum_{n>0}^{\infty}nT_nz^{-n-1}+T_0z^{-1}-\frac{1}{2\pi i}\sum_{n>0}^{\infty}\frac{\partial\mathcal{F}}{\partial T_n}z^{n-1}\right]dz. \tag{2.16}$$

Then, Whitham dynamics in $N=2$ Yang–Mills theory implies the homogeneity relation of the prepotential[14]

$$a\frac{\partial\mathcal{F}}{\partial a}-2\mathcal{F}=-T_0\frac{\partial\mathcal{F}}{\partial T_0}-T_1\frac{\partial\mathcal{F}}{\partial T_1}. \tag{2.17}$$

Note that the right-hand side of (2.17) should be identified with that in (2.14).

In the framework of Whitham hierarchy, T_0, T_1, and $\partial\mathcal{F}/\partial T_1$ are read from (2.16) as

$$T_0=-i\frac{m}{4\sqrt{2}\pi},\quad T_1=i\frac{3}{4\sqrt{2}\pi},\quad \frac{\partial\mathcal{F}}{\partial T_1}=\sqrt{2}\left(\frac{m^2}{4}-u\right). \tag{2.18}$$

Due to our normalization, the numerical factors are different from those used by Eguchi and Yang.[14] Note that T_0 vanishes when $m=0$. This simplifies (2.17) and because of this, the scaling relation in the massless theory is easily determined by using the Whitham hierarchy. In this case, it is not necessary to know $\partial\mathcal{F}/\partial T_0$, but in a massive case it must be known. However, in a massive theory the calculation of $\partial\mathcal{F}/\partial T_0$ requires care because the third equation in (2.15) is a line integral from one covering of the Riemann surface to the other. In the case at hand, this integral consists of two pieces of integral from $-m$ to ∞ on one sheet and a copy of it on the other sheet (when the mass vanishes, it reduces to a familiar integral[14]). Due to the pole of λ at $x=-m$, the evaluation of it is not easy and it is therefore still at a challenging stage. However, we can develop another method and show that $\partial\mathcal{F}/\partial T_0$ satisfies a differential equation with the aid of (2.13). Then $\partial\mathcal{F}/\partial T_0$ is obtained as a solution to this equation.

Differentiation of (the left-hand side of) (2.17) gives w, so a differential equation for $\partial\mathcal{F}/\partial T_0$ can be obtained by substituting w calculated from (2.17) with (2.18) into (2.13)

$$\Theta''' + X\Theta'' + Y\Theta' = \frac{4\sqrt{2}}{m}\left[i\frac{\pi}{2}W + \frac{6}{\Delta}\left(-8m^2 + 3\frac{3m\Lambda^3 - 4u^2}{4m^2 - 3u}\right)\right], \tag{2.19}$$

where $\Theta = \partial\mathcal{F}/\partial T_0$. It is easy to get a solution to this equation

$$\Theta' = c_1 a' + c_2 a_D' + i\frac{2\sqrt{2}\pi}{m}w + i\frac{3\sqrt{2}\pi}{2m}\left[a_D'\int_0 \frac{a'Z}{4m^2-3u}du - a'\int_0 \frac{a_D'Z}{4m^2-3u}du\right], \tag{2.20}$$

where c_i are integration constants and

$$Z = -8m^2 + 3\frac{3m\Lambda^3 - 4u^2}{4m^2-3u}. \tag{2.21}$$

In the derivation of (2.20), we have used the fact that a second-order differential equation

$$\frac{d^2y}{dx^2} + P(x)\frac{dy}{dx} + Q(x)y = R(x) \tag{2.22}$$

with any function $P(x)$, $Q(x)$ and $R(x)$ has a general solution in the form

$$y = c_1 y_1 + c_2 y_2 - y_1\int_0 \frac{y_2 R(x)}{W(y_1,y_2)}dx + y_2\int_0 \frac{y_1 R(x)}{W(y_1,y_2)}dx, \tag{2.23}$$

where y_i are two independent solutions in the case of $R(x)=0$ and $W(y_1,y_2)$ is its Wronskian. Since (2.19) without the right-hand side is nothing but the massive Picard–Fuchs equation, the above y_i may be chosen as a' and a_D'. Then $W(y_1,y_2)$ is identified with W defined in (2.7). Furthermore, (2.11) is used to arrive at the final expression (2.20).

Let us see the massless limit of (2.20). In this limit, naively, the factor $1/m$ diverges, but we leave it for the moment. When m vanishes, the integrals can be easily evaluated and the resulting terms cancel out the third term in (2.20). Therefore, it follows that

$$\left.\frac{\partial\mathcal{F}}{\partial T_0}\right|_{m\to 0} = c_1 a + c_2 a_D + \text{const.} \tag{2.24}$$

This result reflects the fact that (2.19) reduces to the total differentiation of the massless Picard–Fuchs equation because of vanishing of the right-hand side of (2.19) for $m\to 0$. Actually, $\partial\mathcal{F}/\partial T_0$ corresponds to $\partial\mathcal{F}/\partial m$, so all constants in (2.24) should be zero for $m\to 0$.

On the other hand, integrating (2.20) for $m\neq 0$, we obtain

$$aa_D - 2\mathcal{F} = i\frac{m}{6\sqrt{2}\pi}(c_1 a + c_2 a_D) + c_3 - \frac{1}{4}\int_0\left[a_D'\int_0 \frac{a'Z}{4m^2-3u}du - a'\int_0 \frac{a_D'Z}{4m^2-3u}du\right]du, \tag{2.25}$$

where c_3 is an integration constant, with the aid of (2.17) and (2.18). This is the general form of the massive scaling relation.

Remark: If the prepotential is known,[20] c_i *are easily determined as*

$$c_1 = -3\pi i n', \quad c_2 = 3\pi i n, \quad c_3 = -\frac{m^2}{16}(i + 4i\ln 2 - 2\pi n n'), \tag{2.26}$$

where $n,n' \in \mathbf{Z}$ are the winding numbers of 1-cycles around the pole corresponding to $x=-m$ of λ. These constants support the massless limit behavior of $\partial\mathcal{F}/\partial T_0$. However, precisely, these integration constants must be determined by comparing them with the result of lower order expansion of the third equation in (2.15).

III. MASSIVE PREPOTENTIAL IN STRONG COUPLING REGIME

Next, let us study the prepotential in the strong coupling region following the technology recently developed in the five-dimensional gauge theory.[17] The basic tool in our case is (2.11).

For later convenience, we take the elliptic curve[7]

$$y^2=x^2(x-u)+\frac{1}{4}m\Lambda^3x-\frac{\Lambda^6}{64}, \tag{3.1}$$

whose discriminant coincides with (2.6). The Seiberg–Witten differential 1-form is given by

$$\lambda=\frac{\sqrt{2}}{8\pi}\frac{dx}{y}\left[2u-3x-\frac{m\Lambda^3}{4x}\right]. \tag{3.2}$$

For this curve, we choose the 1-cycles on the surface (3.1) as Ito-Yang cycles,[18] which reduce for large u with vanishing mass

$$\alpha:\quad -i\frac{\Lambda^3}{8\sqrt{u}}\rightarrow +i\frac{\Lambda^3}{8\sqrt{u}},\qquad \beta:\quad u\rightarrow -i\frac{\Lambda^3}{8\sqrt{u}}. \tag{3.3}$$

Also in this case, the periods are defined by (2.3) and satisfy (2.4).

To find strong coupling regime, let us decompose the discriminant as

$$\Delta(u)=256\prod_{i=1}^{3}(u-e_i), \tag{3.4}$$

where e_i are given by the vanishing points of $\Delta(u)=0$, which correspond to the strong coupling regime. e_i can be easily obtained by solving $\Delta(u)=0$ for u, but we should take care of the derivation. In general, any cubic equation in x,

$$x^3+ax^2+bx+c=0, \tag{3.5}$$

where a, b, and c are some constants independent of x, has three independent solutions. One of them is

$$x=-\frac{2^{1/3}(-a^2+3b)}{3[-2a^3+9ab-27c+\sqrt{4(-a^2+3b)^3+(-2a^3+9ab-27c)^2}]^{1/3}}+\cdots, \tag{3.6}$$

where the ellipsis $(\cdots)$ means omission of the remaining terms. However, if the denominator vanishes

$$-2a^3+9ab-27c+\sqrt{4(-a^2+3b)^3+(-2a^3+9ab-27c)^2}=0, \tag{3.7}$$

i.e.,

$$(a^2-3b)^3=0, \tag{3.8}$$

formula (3.6) is not valid any more. In this case, one must start again from (3.5) under condition (3.8).

In the case at hand, the condition (3.8) for (2.6) corresponds to

$$m^3(2m+3\Lambda)^3(4m^2-6m\Lambda+9\Lambda^2)^3=0, \tag{3.9}$$

which implies that the directly obtained location under $m\neq 0$ [cf. (3.6)], does not reflect the precise massless limit of the massive discriminant. To understand more illustratively, let us set $u=x+iy$, where $x,y\in\mathbf{R}$. Then the equation $\Delta(u)=0$ produces $3(x-m^2/3)^2-y^2=m(8m^3+27)/24$. For a nonzero m, this is nothing but hyperbolic curves, but for $m=0$ it becomes crossing lines. Clearly, the zeros of the discriminant at $m=0$ behave singularly because transition from the hyperbolic curve to lines is always suppressed.

In fact, though the correct massless location in the strong coupling regime must be

$$e_1^{(0)}=-\frac{3\Lambda^2}{2^{8/3}},\quad e_2^{(0)}=\frac{3(1-i\sqrt{3})\Lambda^2}{2^{11/3}},\quad e_3^{(0)}=\frac{3(1+i\sqrt{3})\Lambda^2}{2^{11/3}}, \tag{3.10}$$

these cannot be obtained from (3.6) with $m=0$. Therefore, (3.9) distinguishes regions in quantum moduli space at $m=0$ and $m\neq 0$. To see a connection with massless theory, for example, we realize the zero locus of the massive discriminant as a small mass perturbation in the form $e_i=e_i^{(0)}+(\text{series in } m)$, where this series converges for $|m|<1$. Substituting this into the equation $\Delta=0$ and equating coefficient of powers in m to zero, one finds

$$\begin{aligned}
e_1&=e_1^{(0)}-\frac{m}{2^{1/3}}\Lambda+\frac{m^2}{3}-\frac{4}{27}\frac{2^{1/3}m^3}{\Lambda}+\frac{4}{81}\frac{2^{2/3}m^4}{\Lambda^2}+\cdots,\\
e_2&=e_2^{(0)}-\frac{(1+i\sqrt{3})}{2^{4/3}}m\Lambda+\frac{m^2}{3}+\frac{2^{4/3}(1-i\sqrt{3})m^3}{27\Lambda}+\cdots,\\
e_3&=e_3^{(0)}+\frac{(1-i\sqrt{3})}{2^{4/3}}m\Lambda+\frac{m^2}{3}+\frac{2^{4/3}(1+i\sqrt{3})m^3}{27\Lambda}+\cdots.
\end{aligned} \tag{3.11}$$

As a final check, (3.4) must be satisfied, but this can be easily confirmed by order in m greater than m^3. Note that (3.11) coincides with the massless strong coupling points for $m=0$ and thus the above e_i are the expected ones for small but finite mass. Below, to make a contact with the result of Ito and Yang,[18] e_1 is chosen as a representative of the strong coupling regime. On the other hand, for m greater than or equal to 1, (3.6) is available to derive zeros of the discriminant.

Remark: In the case of the SU(2) *gauge group, this situation is characteristic of the* $N_f=1$ *theory. For* $N_f=2$ *and* 3 *theories' discriminants, we do not encounter such a sensitive problem in the determination of the strong coupling region.*

Performing differential calculation between periods and dual prepotential

$$\begin{aligned}
\frac{da}{du}&=\frac{\mathcal{F}_D''}{u'},\\
\frac{d^2a}{du^2}&=\frac{1}{u'^3}(\mathcal{F}_D'''u'-\mathcal{F}_D''u''),\\
\frac{da^3}{du^3}&=\frac{1}{u'^5}[(\mathcal{F}_D^{(4)}u'-\mathcal{F}_D'''u''')u'-3u''(\mathcal{F}_D'''u'-\mathcal{F}_D''u'')],
\end{aligned} \tag{3.12}$$

where $'=d/da_D$ and $\mathcal{F}_D$ is the prepotential defined by

$$a=\frac{d\mathcal{F}_D}{da_D}, \tag{3.13}$$

and using (2.4) for $\Pi = a$ with inverse relation du/da_D (see also Ref. 17), we can arrive at the differential equation for $\mathcal{F}_D$,

$$\mathcal{F}_D^{(4)} + u'\left(X - \frac{3u''}{u'^2}\right)\mathcal{F}'''_D = 0. \tag{3.14}$$

This equation is integrated to give

$$\mathcal{F}'''_D = c\,\frac{4m^2 - 3u}{\Delta}\,u'^3, \tag{3.15}$$

where c is an integration constant to be determined later. The factor $(4m^2 - 3u)/\Delta$ corresponds to the Wronskian W defined in (2.7).

Since $\mathcal{F}'''_D$ manifestly vanishes at $u = 4m^2/3$, $\mathcal{F}_D$ is represented by a linear combination of a_D^2, a_D, and 1, although $O(a_D)$ terms can be neglected, thus it follows immediately from (3.13) that $a \propto a_D$. This induces the trivial monodromy

$$\begin{pmatrix} a \\ a_D \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a \\ a_D \end{pmatrix}, \tag{3.16}$$

which indicates that the indicial indices of the Picard–Fuchs equation at $u = 4m^2/3$ are integers and the BPS spectrum is unchanged at this point.

Remark: The same monodromy can be seen from a version of (3.15) in the weak coupling region.

As is easy to find, all we need to determine $\mathcal{F}_D$ is a function $u = u(a_D)$, which is obtainable from inverting the solution of the Picard–Fuchs equation. Therefore, substituting these data into (3.15) and triply integrating it, we will be able to determine $\mathcal{F}_D$. On the other hand, it would be sufficient to once determine c at a representative point in the moduli space. Since c has mass dimension zero, it can be regarded as a pure number. In fact, c can be determined at the massless point and the result

$$c = \frac{i}{16\pi} \tag{3.17}$$

follows from comparing (3.15) with the massless prepotential.[21]

Next we calculate u', but this is an easy task. Since $u' = (da_D/du)^{-1}$, it is sufficient to obtain the solution to the Picard–Fuchs equation at $u = e_1$. With the help of (3.4), it is found by a linear combination

$$\frac{da_D}{du} = \rho_1\varphi_1 + \rho_2\varphi_2, \tag{3.18}$$

where

$$\begin{aligned} \varphi_1 &= 1 - \frac{6u}{4m^2 - 3e_1} - \frac{u^2}{e_1(2e_1 - e_2)(2e_1 - e_3)(3e_1 - 4m^2)^2}[9e_1^2(36e_1 - e_2 - e_3) - 12(112e_1^2 \\ &\quad - 3e_1e_2 - 3e_1e_3 + e_2e_3)m^2 + 512(3e_1 - m^2)m^4 + 36m(3e_1 - 4m^2)\Lambda^3] - \cdots, \\ \varphi_2 &= u - \frac{3e_1^2(4e_1 - e_2 - e_3) - 4(8e_1^2 - 3e_1e_2 - 3e_1e_3 + e_2e_3)m^2}{2e_1(2e_1 - e_2)(2e_1 - e_3)(3e_1 - 4m^2)}u^2 - \cdots. \end{aligned} \tag{3.19}$$

Since these φ_i are extremely complicated functions in the language of explicit e_i, i.e., the right-hand sides of e_i in (3.11), we do not try to express φ_i by using them. The constants ρ_i are determined from an asymptotic expansion of the period integral of a_D, i.e.,

$$\rho_1 = -\frac{1}{2^{1/6}\sqrt{3}\Lambda}, \quad \rho_2 = -\frac{25}{9\Lambda^3}\sqrt{\frac{2}{3}}. \tag{3.20}$$

Integration produces a_D itself with an integration constant, but from dimensional analysis, it is found that it has a unit mass dimension. Since all dependence of the scale parameter (instanton correction) should be entered on the right-hand side of (3.18), the integration constant must be proportional to m. In addition, since the massless theory does not have this term, it can be regarded as a characteristic feature in massive theory. Actually, it is a residue contribution from the massive meromorphic 1-form, and this is the case. Thus,

$$a_D = 2\pi i n \operatorname{res}(\lambda) + \int_0 (\rho_1\varphi_1 + \rho_2\varphi_2)du, \tag{3.21}$$

where $n \in \mathbf{Z}$ is the winding number of the β-cycle which loops around the pole of the massive meromorphic 1-form and the residue is evaluated at $x=0$. Denoting $\tilde{a}_D = a_D - 2\pi i n \operatorname{res}(\lambda)$ and repeatedly solving (3.21), one can arrive at the inverse relation $u = u(a_D)$.

In this way, the result

$$\mathcal{F}_D = \frac{1}{2}c_1\tilde{a}_D^2 + c_2\tilde{a}_D + c_3 - i\frac{\Lambda^2}{\pi}\frac{\tilde{a}_D^2}{(e_1-e_2)(e_1-e_3)}\Bigg[\frac{3}{2^{17/3}}(3e_1-4m^2)(\ln\tilde{a}_D^2 - 3)$$
$$+ \frac{1}{2^{29/6}3^{2/3}(e_1-e_2)(e_1-e_3)}[125(e_1-e_2)(e_1-e_3)(3e_1-4m^2) + 9\cdot 2^{7/3}(e_2+e_3-2e_1)$$
$$\times m^2\Lambda^2 + 27\cdot 2^{1/3}(6e_1^2 + 4e_2e_3 - 5e_1(e_2+e_3))\Lambda^2]\frac{\tilde{a}_D}{\Lambda} + \cdots\Bigg], \tag{3.22}$$

where c_i are integration constants, follows from expanding (3.15). $c_2\tilde{a}_D + c_3$ may be neglected because this term does not change the effective coupling constant, but c_1 is nontrivial. Dimensional analysis shows that c_1 is a pure number. Therefore, it is sufficient to choose it as a constant such that $\mathcal{F}_D$ for $m \to 0$ coincides with that in the massless theory. As a check, one can see that this $\mathcal{F}_D$ reduces to that of massless theory for $m \to 0$, provided

$$c_1 = -\frac{i}{4\pi}\ln\Lambda^2. \tag{3.23}$$

The reader might already have noticed that actually this prepotential was obtained without regard to the details of e_1, although it was calculated under the assumption of small mass to see a connection with the massless theory. For general m in the strong coupling regime, only the differences of the final form of the prepotential are explicit values of e_i, c_i and c. For instance, for a large mass case, it is enough to simply expand (3.6) near $m = \infty$ to get the zeros of Δ and replace e_i by, e.g.,

$$e_1 = m^2 + \frac{\Lambda^3}{8m} + \cdots, \quad e_2 = \Lambda^{3/2}\sqrt{m} - \frac{\Lambda^3}{16m} + \cdots, \quad e_3 = -\Lambda^{3/2}\sqrt{m} - \frac{\Lambda^3}{16m} + \cdots. \tag{3.24}$$

The period a_D can be obtained by simply substituting (3.24) into (3.18). In this way, the dual prepotential for a large mass is calculated, but its form is not so attractive for us because it is again written by a complicated function in e_i. For this reason, it would not be necessary to write down the dual prepotential for $|m| \geqq 1$, and for other N_f cases.

IV. SUMMARY

In this paper, it has been shown that the Wronskian of the Picard–Fuchs equation in massive theory is related to (differentiation of) Matone's modular invariant through a differential equation. By comparing the solution to this equation with the Whitham dynamics, we have found a closed form of $\partial\mathcal{F}/\partial T_0$, which has not been evaluated so far. In this way, we have found a general form of the massive scaling relation. In the pure SU(2) theory, the scaling relation of the prepotential is known to also be obtained as an anomalous superconformal Ward identity,[22] therefore, it would be a natural question to ask whether (2.25) can be obtained as ''an anomalous superconformal Ward identity'' in this $N_f=1$ massive theory.

On the other hand, we have also found that the differential equation for the Wronskian gives a differential relation between prepotential and moduli, which is quite reminiscent of the one presented in the five-dimensional gauge theory[17] and the dual prepotential is calculated from this equation.

A more detailed study of the relation between the Wronskian and modular invariant of Matone will provide us with more useful information on the massive prepotentials in the future.

ACKNOWLEDGMENT

The author acknowledges Professor K. Ito (YITP) for discussions on miscellaneous aspects of $N=2$ Seiberg–Witten models.

[1] N. Seiberg, Phys. Lett. B **206**, 75 (1988).
[2] M. T. Grisaru and W. Siegel, Nucl. Phys. B **201**, 292 (1982).
[3] P. S. Howe, K. S. Stelle, and P. K. Townsend, Nucl. Phys. B **214**, 519 (1983).
[4] P. S. Howe, K. S. Stelle, and P. K. Townsend, Nucl. Phys. B **236**, 125 (1984).
[5] P. S. Howe, K. S. Stelle, and P. C. West, Phys. Lett. **124B**, 55 (1983).
[6] N. Seiberg and E. Witten, Nucl. Phys. B **431**, 484 (1994).
[7] N. Seiberg and E. Witten, Nucl. Phys. B **435**, 129 (1994).
[8] A. Klemm, W. Lerche, and S. Theisen, Int. J. Mod. Phys. A **11**, 1929 (1996).
[9] *Essays on Mirror Manifolds*, edited by S. T. Yau (International, Hong Kong, 1992).
[10] P. Candelas, X. C. de la Ossa, P. S. Green, and L. Parkes, Nucl. Phys. B **359**, 21 (1991).
[11] P. Candelas, X. C. de la Ossa, P. S. Green, and L. Parkes, Phys. Lett. B **258**, 118 (1991).
[12] M. Matone, Phys. Lett. B **357**, 342 (1995).
[13] J. Sonnenschein, S. Theisen, and S. Yankielowicz, Phys. Lett. B **367**, 145 (1996).
[14] T. Eguchi and S.-K. Yang, Mod. Phys. Lett. A **11**, 131 (1996).
[15] A. Gorsky, I. Krichever, A. Marshakov, A. Mironov, and A. Morozov, Phys. Lett. B **355**, 466 (1995).
[16] T. Nakatsu and K. Takasaki, Mod. Phys. Lett. A **11**, 157 (1996).
[17] H. Kanno and Y. Ohta, Nucl. Phys. B **530**, 73 (1998).
[18] K. Ito and S.-K. Yang, Phys. Lett. B **366**, 165 (1996).
[19] A. Hanany and Y. Oz, Nucl. Phys. B **452**, 283 (1995).
[20] Y. Ohta, J. Math. Phys. **37**, 6074 (1996).
[21] K. Ito and S.-K. Yang, ''Picard-Fuchs equations and prepotentials in $N=2$ supersymmetric QCD,'' hep-th/9603073.
[22] P. S. Howe and P. C. West, Nucl. Phys. B **486**, 425 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

The flux-across-surfaces theorem for short range potentials and wave functions without energy cutoffs

S. Teufel[a] and D. Dürr
Mathematisches Institut der Universität München, Theresienstraße 39, 80333 München, Germany

K. Münch-Berndl[b]
Institut für Angewandte Mathematik, Universität Zürich-Irchel, Winterthurer Strasse 190, 8057 Zürich, Switzerland

(Received 24 June 1998; accepted for publication 5 January 1999)

The quantum probability flux of a particle integrated over time and a distant surface gives the probability for the particle crossing that surface at some time. The relation between these crossing probabilities and the usual formula for the scattering cross section is provided by the flux-across-surfaces theorem, which was conjectured by Combes, Newton, and Shtokhamer [Phys. Rev. D **11**, 366–372 (1975)]. We prove the flux-across-surfaces theorem for short range potentials and wave functions without energy cutoffs. The proof is based on the free flux-across-surfaces theorem (Daumer *et al.*) [Lett. Math. Phys. **38**, 103–116 (1996)], and on smoothness properties of generalized eigenfunctions: It is shown that if the potential $V(x)$ decays like $|x|^{-\gamma}$ at infinity with $\gamma > n \in \mathbb{N}$ then the generalized eigenfunctions of the corresponding Hamiltonian $-1/2\Delta + V$ are $n-2$ times continuously differentiable with respect to the momentum variable. © *1999 American Institute of Physics.* [S0022-2488(99)00604-0]

I. INTRODUCTION

Potential scattering theory is concerned with the long-time behavior of wave functions Ψ_t. Its relation to experiment, i.e., to the definition of the scattering cross section, is, however, only rarely discussed. One such relation is provided by Dollard's scattering-into-cones theorem.[1] It asserts that, assuming asymptotic completeness of the wave operators, the probability of finding a particle with a wave function $\Psi_t = e^{-iHt}\Psi_0 \in \mathcal{H}_{ac}(H)$, the absolutely continuous subspace for the Hamiltonian H, in the far future in a given cone $C \subset \mathbb{R}^3$ (with a vertex at the origin) equals the probability that the quantum mechanical momentum of the asymptotic outgoing wave $W_+^{-1}\Psi_0$ lies in the same cone,

$$\lim_{t\to\infty} \int_C |\Psi_t(x)|^2\, dx = \int_C |\widehat{W_+^{-1}\Psi_0}(k)|^2\, dk, \tag{1}$$

where $\widehat{\ }$ denotes the Fourier transform, $W_+ := s-\lim_{t-\infty} e^{iHt}e^{-iH_0 t}$ is the wave operator, and $H = H_0 + V$ with the free Hamiltonian $H_0 = -\frac{1}{2}\Delta$ (we choose units such that $\hbar = m = 1$) and the potential V. The scattering-into-cones theorem is regarded as fundamental for quantum mechanical scattering theory. The expression for the differential cross section $d\sigma/d\Omega = |f(\theta,\phi)|^2$ from the time-independent scattering theory can be derived from the right-hand side of (1).

Combes, Newton, and Shtokhamer[3] observed, however, that what is relevant for scattering theory is a formula for the probability that the particle crosses some distant surface at some time during the scattering process. Heuristically, this probability should be given by integrating the

[a]Electronic mail: teufel@rz.mathematik.uni-muenchen.de
[b]Electronic mail: kmb@amath.unizh.ch

quantum mechanical probability flux $j^{\Psi_t} := \mathrm{Im}(\Psi_t^* \nabla \Psi_t)$ over the relevant time interval and this surface.[4–6] Combes, Newton, and Shtokhamer hence conjectured the flux-across-surfaces theorem,

$$\lim_{R\to\infty} \int_0^\infty dt \int_{R\Sigma} j^{\Psi_t} \cdot n \, d\sigma = \int_{C_\Sigma} |\widehat{W_+^{-1}\Psi_0}(k)|^2 \, dk, \tag{2}$$

where Σ is a measurable subset of S_1, the sphere with radius 1, $R\Sigma := \{Rx \in \mathbb{R}^3 : x \in \Sigma\}$ and $C_\Sigma := \{\lambda x \in \mathbb{R}^3 : x \in \Sigma, \lambda \geq 0\}$ is the cone spanned by Σ. The free flux-across-surfaces theorem (2) was proven by Daumer, Dürr, Goldstein, and Zanghi[4] for the case $V=0$. Recently (2) has been established for a large class of short range potentials by Amrein and Zuleta.[7] For long range potentials Amrein and Pearson[8] showed that the left-hand side of (2) equals the left-hand side of (1). [In this case, modified wave operators have to be introduced to define the right-hand side of (1) and (2).] However, since the proofs in Refs. 7 and 8 apply the usual time-dependent methods, they have to assume that $\widehat{W_+^{-1}\Psi_0}$ has compact support not containing the origin. Although this condition is a natural idealization of the experimental situation often encountered in scattering theory, and these wave functions form a dense set in L^2, there are no physical or mathematical reasons that (2) should hold only for this restricted class of wave functions. Furthermore, there are situations, i.e., the decay of an unstable system, where the physically interesting wave functions do have momentum support at zero. But the set of wave functions for which (2) holds cannot be enlarged by a simple limiting procedure in L^2, since the expression

$$\int_0^\infty dt \int_{R\Sigma} j^{\Psi_t} \cdot n \, d\sigma$$

is an unbounded sesquilinear form. Therefore the essential propagation estimates have to be proven directly for wave functions without energy cutoffs. Some results in this direction, so-called L^p estimates, have been established under rather restrictive conditions on the potential.[9,10] However, these estimates alone are not sufficient to prove (2).

In this paper we will give an elementary proof of (2) for a class of wave functions without energy cutoffs. We must assume, however, that the potential is short range with decay of order $|x|^{-4-\epsilon}$, $\epsilon>0$, at infinity, and that it does not have a zero energy resonance or eigenvalue. Our proof as well as the proof in Ref. 7 are based on the results of the free case $(V=0)$ established in Ref. 4. We employ stationary phase methods and the so-called generalized eigenfunctions $\Phi(x,k)$, which are certain solutions of the stationary Schrödinger equation $(-\frac{1}{2}\Delta + V(x))\Phi(x,k) = k^2\Phi(x,k)$, $k \in \mathbb{R}^3$, not belonging to $L^2(\mathbb{R}^3)$. This strategy of proof has been put forward in Ref. 5. We need $\Phi(x,k)$ to be differentiable with respect to k as well as to be uniformly bounded in both variables. Furthermore, we need that $\sup_{k\in\mathbb{R}\setminus\{0\}} |\partial_{k_l}\Phi(x,k)| \leq c(1+|x|)$ for some constant c and $l=1,2,3$.

In Sec. II, the flux-across-surfaces theorem will be established under suitable conditions on the generalized eigenfunctions. In Sec. III we will prove a theorem on the regularity of the generalized eigenfunctions, which will, among other things, justify the assumptions made in Sec. II: $\Phi(x,k)$ is n times partially differentiable with respect to k if $V(x) = O(|x|^{-n-2-\epsilon})$ for $|x| \to \infty$ and some $\epsilon>0$. Moreover, consider a family of Hamiltonians $H_c := H_0 + cV$, $c \in \mathbb{R}$. Then, if $V(x) = O(|x|^{-3-\epsilon})$,[11] the eigenfunctions corresponding to H_c are uniformly bounded and their partial derivatives of order n with respect to k grow not faster than $(1+|x|)^n$, except for a discrete set of constants, $c \in \mathbb{R}$.

II. THE FLUX-ACROSS-SURFACES THEOREM

We start with notation. Points in position space will be denoted by $x \in \mathbb{R}^3$; points in momentum space by $k \in \mathbb{R}^3$. By dx and dk integration with respect to the Lebesgue measure on $\mathbb{R}^3$ is understood. For $n \geqslant 2$, the following conditions on the potential V will be denoted by $(\mathbf{V})_n$:$(\mathbf{V})_n$ $V:\mathbb{R}^3 \to \mathbb{R}$ and (i) V is locally Hölder continuous except at a finite number of singularities;[12] (ii) $V \in L^2(\mathbb{R}^3)$; and (iii) $|V(x)| = O(|x|^{-n-\epsilon})$ for $|x| \to \infty$ and some $\epsilon > 0$.

For $n=2$ these are the conditions of Ikebe.[13] Under these conditions H is self-adjoint on the domain of H_0. The absolutely continuous part of the spectrum is $[0, \infty)$. Furthermore, H has neither positive eigenvalues nor a singular continuous spectrum. The wave operators $W_\pm = s-\lim_{t\to\pm\infty} e^{-iHt}e^{iH_0t}$ exist and are complete, i.e., Ran $W_\pm = \mathcal{H}_{\mathrm{ac}}(H)$.

The time-dependent wave function will be denoted by $\Psi_t := e^{-iHt}\Psi_0$, $\Psi_0 \in L^2(\mathbb{R}^3)$. To simplify notation we will abbreviate $\Psi_{\mathrm{out}} := W_+^{-1}\Psi_0$ for the outgoing asymptotic wave. By S we denote the set of Schwartz functions.

Zero is said to be a resonance of H if there exists a solution f of $-1/2\Delta f(x) + V(x)f(x) = 0$ such that $(1+|x|)^{-\gamma} f(x) \in L^2(\mathbb{R}^3)$ for any $\gamma > \frac{1}{2}$ and not for $\gamma = 0$.[14] The appearance of zero-energy resonances or eigenvalues is an exceptional event: $H_c = H_0 + cV$ can have a zero-energy resonance or eigenvalue only for c in discrete subset of $\mathbb{R}$.[14]

Theorem 2.1: Let the potential satisfy the condition $(\mathbf{V})_4$ and let zero be neither a resonance nor an eigenvalue of H. Let $\Psi_{\mathrm{out}} \in S$. Then $\Psi_t = e^{-iHt}W_+\Psi_{\mathrm{out}}$ is continuously differentiable, except at the singularities of V and for any measurable $\Sigma \subset S_1$ and any $T \in \mathbb{R}$,

$$\lim_{R\to\infty} \int_T^\infty dt \int_{R\Sigma} j^{\Psi_t}(x)\cdot n \, d\sigma = \lim_{R\to\infty} \int_T^\infty dt \int_{R\Sigma} |j^{\Psi_t}(x)\cdot n| d\sigma = \int_{C_\Sigma} |\hat{\Psi}_{\mathrm{out}}(k)|^2 \, dk. \tag{3}$$

Remark 2.2: The first equality in (3) shows that far away from the scattering center the flux is essentially outgoing, i.e., that there the particles cross spherical surfaces only once and do not return. Thus (3) yields the crossing probability of interest.

Remark 2.3: It would be, of course, more satisfactory if we could prove (3) under a suitably general condition on Ψ_0, not on Ψ_{out}. However, the set of wave functions $\Psi_0 = W_+\Psi_{\mathrm{out}}$ for which Theorem 2.1 holds is dense in $\mathcal{H}_{\mathrm{ac}}(H)$, since S is dense in L^2 and $W_+ : L^2 \to \mathcal{H}_{\mathrm{ac}}(H)$ is unitary. For an explicit characterization of the domain W_+S one would need suitable mapping properties of the wave operators. Some mapping properties for wave functions without energy cutoffs have been established by Yajima,[10] however, they are not sufficient for our purpose.

Remark 2.4: Due to the so-called intertwining property of the wave operators, $W_\pm e^{-iH_0t} = e^{-iHt}W_\pm$, and the fact that S is left invariant under the free time evolution, the condition imposed on Ψ in Theorem 2.1 is invariant under the full time evolution: $e^{-iHt}W_+S = W_+e^{-iH_0t}S = W_+S$.

As already mentioned, we will make use of the generalized eigenfunctions $\Phi(x,k)$ that diagonalize H in the same sense as the ordinary plane waves $\{e^{ik\cdot x}, k \in \mathbb{R}^3\}$ diagonalize H_0. We define $\Phi(x,k)$ and state the properties that we will use in the proof of Theorem 2.1 in a proposition.

Proposition 2.5: Let V satisfy $(\mathbf{V})_2$. Then for any $k \in \mathbb{R}^3 \backslash \{0\}$ there are unique continuous solutions $\Phi_\pm(\cdot,k):\mathbb{R}^3 \to \mathbb{C}$ of the Lippmann–Schwinger equations,

$$\Phi_\pm(x,k) = e^{ik\cdot x} - \frac{1}{2\pi} \int \frac{e^{\mp i|k||x-y|}}{|x-y|} V(y)\Phi_\pm(y,k)dy, \tag{4}$$

with the boundary conditions $\lim_{|x|\to\infty}[\Phi_\pm(x,k) - e^{ik\cdot x}] = 0$, which are also classical solutions of the stationary Schrödinger equation,

$$\left[-\frac{1}{2}\Delta + V(x)\right]\Phi_\pm(x,k) = \frac{k^2}{2}\Phi_\pm(x,k), \tag{5}$$

such that the following holds.

(i) For any compact $D \subset \mathbb{R}^3 \backslash \{0\}$ the functions $\Phi_\pm(\cdot,\cdot):\mathbb{R}^3 \times D \to \mathbb{C}$ are uniformly continuous.

(ii) For any $f \in L^2(\mathbb{R}^3)$ the generalized Fourier transforms,

$$(\mathcal{F}_{\pm} f)(k) = \frac{1}{(2\pi)^{3/2}} \text{l.i.m.} \int \Phi^{*}_{\pm}(x,k) f(x) dx,$$

exist in $L^2(\mathbb{R}^3)$.[15]

(iii) Ran $\mathcal{F}_{\pm} = L^2(\mathbb{R}^3)$ and $\mathcal{F}_{\pm} : \mathcal{H}_{\text{ac}}(H) \to L^2(\mathbb{R}^3)$ are unitary and the inverse of $\mathcal{F}_{\pm}$ is given by

$$(\mathcal{F}^{-1}_{\pm} f)(x) = \frac{1}{(2\pi)^{3/2}} \text{l.i.m.} \int \Phi_{\pm}(x,k) f(k) dk.$$

(iv) For any $f \in D(H) \cap \mathcal{H}_{\text{ac}}(H)$, we have

$$Hf(x) = \left(\mathcal{F}^{-1}_{\pm} \frac{k^2}{2} \mathcal{F}_{\pm} f \right)(x) = \frac{1}{(2\pi)^{3/2}} \text{l.i.m.} \int \frac{k^2}{2} \Phi_{\pm}(x,k) (\mathcal{F}_{\pm} f)(k) dk, \tag{6}$$

and therefore for any $f \in \mathcal{H}_{\text{ac}}(H)$,

$$e^{-iHt} f(x) = (\mathcal{F}^{-1}_{\pm} e^{-i(k^2/2)t} \mathcal{F}_{\pm} f)(x) = \frac{1}{(2\pi)^{3/2}} \text{l.i.m.} \int e^{-i(k^2 t/2)} \Phi_{\pm}(x,k) (\mathcal{F}_{\pm} f)(k) dk. \tag{7}$$

(v) For any $f \in \mathcal{H}_{\text{ac}}(H)$ the relations $W_{\pm} f = \mathcal{F}^{-1}_{\pm} \mathcal{F} f$ hold, where $\mathcal{F}$ denotes the ordinary Fourier transform.

(vi) If V satisfies $(\mathbf{V})_3$, then $\Phi_{\pm}(x,k)$ are continuously differentiable with respect to k for all $x \in \mathbb{R}^3$, except at $k=0$. The partial derivatives $\partial_{k_l} \Phi_{\pm}(x,k)$ are continuous in x and k. If, in addition, zero is not an eigenvalue or resonance of H, then $\Phi_{\pm}(x,k)$ is well defined and continuous for $x,k \in \mathbb{R}^3$,

$$\sup_{x \in \mathbb{R}^3, k \in \mathbb{R}^3} |\Phi_{\pm}(x,k)| < \infty,$$

and there is a $c < \infty$ such that

$$\sup_{k \in \mathbb{R}^3 \setminus \{0\}} \left| \frac{\partial}{\partial k_l} \Phi_{\pm}(x,k) \right| < c(1 + |x|),$$

for $l = 1,2,3$.

The proof of (i)–(v) of Proposition 2.5 is due to Ikebe.[13] (vi) is a special case of Theorem 3.1 on the regularity of generalized eigenfunctions that we shall state and prove in Sec. III.

Remark 2.6: Similar eigenfunction expansions can be obtained also for potentials with slower decay, but then, in general, the continuity in k will not hold anymore.[16]

Proof (of Theorem 2.1): Let $\Psi_t = e^{-iHt} W_+ \Psi_{\text{out}}$, $\Psi_{\text{out}} \in S$. Using Proposition 2.5.(iv), (v) and $\eta(x,k) := \Phi_+(x,k) - e^{ik \cdot x}$, we have that

$$\begin{aligned}
\Psi_t(x) &= \frac{1}{(2\pi)^{3/2}} \int e^{-i(k^2 t/2)} \hat{\Psi}_{\text{out}}(k) \Phi_+(x,k) dk \\
&= \frac{1}{(2\pi)^{3/2}} \int e^{-i(k^2 t/2)} \hat{\Psi}_{\text{out}}(k) e^{ik \cdot x} dk + \frac{1}{(2\pi)^{3/2}} \int e^{-i(k^2 t/2)} \hat{\Psi}_{\text{out}}(k) \eta(x,k) dk \\
&=: \alpha(x,t) + \beta(x,t).
\end{aligned} \tag{8}$$

The flux generated by this wave function is

$$j^{\Psi_t}(x) = \text{Im}(\alpha^* \nabla \alpha + \alpha^* \nabla \beta + \beta^* \nabla \alpha + \beta^* \nabla \beta), \tag{9}$$

where the differentiability of α is obvious and that of β will be established later.

The first part $j_0 = \mathrm{Im}(\alpha^* \nabla \alpha)$ is the flux generated by the free time evolution of Ψ_{out} and according to the free flux-across-surfaces theorem,[2]

$$\lim_{R\to\infty} \int_T^\infty dt \int_{R\Sigma} j_0(x,t)\cdot n\, d\sigma = \lim_{R\to\infty} \int_T^\infty dt \int_{R\Sigma} |j_0(x,t)\cdot n|\, d\sigma = \int_{C_\Sigma} |\hat{\Psi}_{\mathrm{out}}(k)|^2\, dk.$$

Therefore to prove (3) we need only show that the last three terms in (9) do not contribute to the flux across distant surfaces, i.e., that for $j_1 := \mathrm{Im}(\alpha^* \nabla \beta + \beta^* \nabla \alpha + \beta^* \nabla \beta)$,

$$\lim_{R\to\infty} \int_T^\infty dt \int_{S_R} |j_1(x,t)\cdot n|\, d\sigma = 0. \tag{10}$$

For some fixed $T>0$ this will follow from the estimates (which we shall prove below)

$$\sup_{x\in S_R} |\alpha(x,t)| \leqslant t^{-3/2} f_1(R,t), \quad \forall t \geqslant T \tag{11}$$

$$\sup_{x\in S_R} |\nabla \alpha(x,t)| \leqslant t^{-3/2} f_2(R,t), \quad \forall t \geqslant T, \tag{12}$$

where there exists a $c<\infty$ such that $f_i(R,t)$ satisfy

$$\lim_{R\to\infty} f_i(R,t) = 0, \quad \forall t \geqslant T \tag{13}$$

and

$$\sup_{R\in[0,\infty),t\geqslant T} f_i(R,t) < c, \tag{14}$$

for $i=1,2$, and there is $R_0 \geqslant 0$ such that

$$\sup_{x\in S_R} |\beta(x,t)| \leqslant c\, \frac{1}{R(t+R)}, \quad \forall R>0, \tag{15}$$

$$\sup_{x\in S_R} |\nabla \beta(x,t)| \leqslant c\, \frac{1}{R(t+R)}, \quad \forall R>R_0, \tag{16}$$

for $t \geqslant T$. Note that the constants in these estimates depend on T.

Using (11) and (16) we obtain by dominated convergence

$$\begin{aligned} \lim_{R\to\infty} \int_T^\infty \int_{S_R} |\mathrm{Im}(\alpha^* \nabla \beta)\cdot n|\, d\sigma\, dt &\leqslant \lim_{R\to\infty} 4\pi \int_T^\infty \sup_{x\in S_R} R^2 |\alpha||\nabla \beta|\, dt \\ &\leqslant c \lim_{R\to\infty} \int_T^\infty \frac{R^2 f_1(R,t)}{t^{3/2} R(t+R)}\, dt = c \int_T^\infty \lim_{R\to\infty} \frac{R f_1(R,t)}{t^{3/2}(t+R)}\, dt = 0, \end{aligned} \tag{17}$$

for $T>0$, where we observed that the integrand in (17) is bounded by an integrable function uniformly in R,

$$\frac{R f_1(R,t)}{t^{3/2}(t+R)} \leqslant c t^{-3/2}.$$

The terms $|\beta^* \nabla \alpha|$ and $|\beta^* \nabla \beta|$ can be treated analogously and thus (10) holds for positive times T.

According to Remark 2.4, the set of wave functions for which (3) holds as well as the right-hand side of (3) are invariant under finite time shifts:

$$\lim_{R\to\infty} \int_T^\infty dt \int_{R\Sigma} j^{\Psi_t}(x)\cdot n\, d\sigma = \lim_{R\to\infty} \int_{\tilde{T}}^\infty dt \int_{R\Sigma} j^{\Psi_{t+T-\tilde{T}}}(x)\cdot n\, d\sigma$$

$$= \int_{C_\Sigma} |e^{-i(k^2/2)(T-\tilde{T})}\hat{\Psi}_{\text{out}}(k)|^2\, dk = \int_{C_\Sigma} |\hat{\Psi}_{\text{out}}(k)|^2\, dk.$$

Therefore if (3) holds (for all $\Psi_{\text{out}} \in S$) for some fixed T, then (3) will hold for all T; hence (3) is proved for all T.

We turn now to the proof of the estimates (11)–(16). Recalling that $\alpha(x,t) = (e^{-iH_0 t}\Psi_{\text{out}})(x)$ and, since ∇ commutes with the free time evolution, $\nabla\alpha(x,t) = (e^{-iH_0 t}\nabla\Psi_{\text{out}})(x)$, we can write

$$\alpha(x,t) = \frac{1}{(2\pi i t)^{3/2}} \int e^{i(|x-y|^2/2t)}\Psi_{\text{out}}(y)\,dy \tag{18}$$

and

$$\nabla\alpha(x,t) = \frac{1}{(2\pi i t)^{3/2}} \int e^{i(|x-y|^2/2t)}\nabla\Psi_{\text{out}}(y)\,dy. \tag{19}$$

Now (11)–(14) are immediate consequences of (18) and (19) and the fact that, for every fixed $t \geqslant T$, $\alpha(x,t)$, and $\nabla\alpha(x,t)$ are Schwartz functions.

By (4) $\eta(x,k) = \Phi_+(x,k) - e^{ik\cdot x} = -1/2\pi \int (e^{i|k||x-y|}/|x-y|)V(y)\Phi_+(y,k)\,dy$, and therefore

$$\beta(x,t) = \frac{1}{(2\pi)^{3/2}} \int e^{-i(k^2 t/2)}\hat{\Psi}_{\text{out}}(k)\,\eta(x,k)\,dk$$

$$= -\frac{1}{(2\pi)^{5/2}} \int e^{-i(k^2 t/2)}\hat{\Psi}_{\text{out}}(k)\left[\int \frac{e^{-i|k||x-y|}}{|x-y|} V(y)\Phi_+(y,k)\,dy\right]dk$$

$$= -\frac{1}{(2\pi)^{5/2}} \int \frac{V(y)}{|x-y|} \int e^{-i(k^2 t/2+|k||x-y|)}\hat{\Psi}_{\text{out}}(k)\Phi_+(y,k)\,dk\,dy$$

$$=: -\frac{1}{(2\pi)^{5/2}} \int \frac{V(y)}{|x-y|} f(x,y,t)\,dy, \tag{20}$$

where

$$f(x,y,t) := \int e^{-i(k^2 t/2+|k||x-y|)}\hat{\Psi}_{\text{out}}(k)\Phi_+(y,k)\,dk. \tag{21}$$

The change of order of integration in (20) is justified by Fubini's theorem. We shall now apply ''stationary phase'' methods to estimate (21). We set

$$\chi := \frac{\frac{k^2 t}{2} + |k||x-y|}{\frac{t}{2} + |x-y|},$$

$$\chi' = \frac{d}{d|k|}\chi = \frac{|k|t + |x-y|}{\frac{t}{2} + |x-y|} = \frac{|k| + |x-y|t^{-1}}{\frac{1}{2} + |x-y|t^{-1}} \geqslant \min(1, 2|k|),$$

$$\omega := \frac{t}{2} + |x-y|.$$

In the following, $'$ will denote differentiation with respect to $|k|$. Introducing spherical coordinates, with $d\Omega$ denoting the Lebesgue measure on the unit sphere, we estimate (21):

$$\begin{aligned} |f(x,y,t)| &= \left| \int \frac{1}{\omega \chi'} \left[\frac{d}{d|k|} e^{-i\omega\chi} \right] \hat{\Psi}_{\text{out}}(k) \Phi_+(y,k) |k|^2 d|k| d\Omega(k) \right| \\ &= \left| \frac{1}{\omega} \int e^{-i\omega\chi} \frac{d}{d|k|} \left[\frac{1}{\chi'} \hat{\Psi}_{\text{out}}(k) \Phi_+(y,k) |k|^2 \right] d|k| d\Omega(k) \right| \\ &\leqslant \frac{1}{\omega} \int \left| \frac{d}{d|k|} \left[\frac{1}{\chi'} \hat{\Psi}_{\text{out}}(k) \Phi_+(y,k) |k|^2 \right] \right| d|k| d\Omega(k). \end{aligned} \tag{22}$$

For the second equality in (22), the boundary term from the partial integration at $|k| = \infty$ vanishes since $\chi'^{-1} \leqslant \max(1, 1/2|k|)$, $\lim_{|k|\to\infty} |k|^2 \hat{\Psi}_{\text{out}}(k) = 0$, and Ψ_+ is bounded according to Proposition 2.5 (vi). The boundary term at $|k| = 0$ vanishes since $\hat{\Psi}_{\text{out}}$ and Φ_+ are bounded and $\chi'^{-1}|k|^2 \leqslant \max(|k|^2, |k|/2)$. Note that the differentiability of Φ_+ is ensured by Proposition 2.5 (vi). Next, observe that

$$\begin{aligned} \left| \frac{d}{d|k|} \left[\frac{1}{\chi'} \hat{\Psi}_{\text{out}} \Phi_+ |k|^2 \right] \right| \leqslant & \left| \frac{1}{\chi'^2} \chi'' \hat{\Psi}_{\text{out}} \Phi_+ |k|^2 \right| + \left| \frac{1}{\chi'} \hat{\Psi}'_{\text{out}} \Phi_+ |k|^2 \right| + \left| \frac{1}{\chi'} \hat{\Psi}_{\text{out}} \Phi'_+ |k|^2 \right| \\ & + \left| \frac{1}{\chi'} \hat{\Psi}_{\text{out}} \Phi_+ 2|k| \right|. \end{aligned} \tag{23}$$

Since $\chi'' = (\frac{1}{2} + |x-y|t^{-1})^{-1} \leqslant 2$, we obtain for the first term

$$\int \left| \frac{1}{\chi'^2} \chi'' \hat{\Psi}_{\text{out}} \Phi_+ \right| dk \leqslant \sup_{y,k \in \mathbb{R}^3} |\Phi_+(y,k)| \left(\int_{|k|<1/2} \frac{|\hat{\Psi}_{\text{out}}(k)|}{2|k|^2} dk + \int_{|k|\geqslant 1/2} 2|\hat{\Psi}_{\text{out}}(k)| dk \right) \leqslant c_1. \tag{24}$$

Analogously we get for the second and fourth term in (23),

$$\int \left| \frac{1}{\chi'} \hat{\Psi}'_{\text{out}} \Phi_+ \right| dk \leqslant c_2 \quad \text{and} \quad \int \left| \frac{2}{\chi'|k|} \hat{\Psi}_{\text{out}} \Phi_+ \right| dk \leqslant c_4.$$

By Proposition 2.5 (vi), the third term satisfies a bound linear in $|y|$:

$$\int \left| \frac{1}{\chi'} \hat{\Psi}_{\text{out}} \Phi'_+ \right| dk \leqslant \tilde{c}_3 \sup_{k \in \mathbb{R}^3 \setminus \{0\}} |\Phi'_+(y,k)| \leqslant c_3(1+|y|).$$

Combining the four estimates, we arrive at

$$|f(x,y,t)| \leqslant c\,\frac{1}{\omega}(1+|y|) = c\,\frac{1+|y|}{\frac{t}{2}+|x-y|}, \tag{25}$$

which inserted into (20) yields

$$\sup_{x\in S_R}|\beta(x,t)| = \sup_{x\in S_R}\left|\frac{1}{(2\pi)^{5/2}}\int \frac{V(y)}{|x-y|} f(x,y,t)\,dy\right| \leqslant c\,\sup_{x\in S_R}\int \frac{|V(y)|(1+|y|)}{|x-y|\left(\frac{t}{2}+|x-y|\right)}\,dy. \tag{26}$$

Now, substituting $z=x-y$,

$$\begin{aligned}\int \frac{|V(y)|(1+|y|)}{|x-y|\left(\frac{t}{2}+|x-y|\right)}\,dy &= \int_{|z|<|x|/2} \frac{|V(x-z)|(1+|x-z|)}{|z|\left(\frac{t}{2}+|z|\right)}\,dz \\ &\quad + \int_{|x-y|\geqslant|x|/2} \frac{|V(y)|(1+|y|)}{|x-y|\left(\frac{t}{2}+|x-y|\right)}\,dy \\ &\leqslant \sup_{z\in B_{|x|/2}} |V(x-z)|(1+|x-z|)\int_0^{|x|/2} \frac{4\pi|z|^2}{|z|\left(\frac{t}{2}+|z|\right)}\,d|z| \\ &\quad + \frac{1}{\frac{|x|}{2}\left(\frac{t}{2}+\frac{|x|}{2}\right)}\int |V(y)|(1+|y|)\,dy,\end{aligned}$$

where B_r denotes the ball with radius r in $\mathbb{R}^3$ centered at the origin. Since $V(x)=O(|x|^{-4-\epsilon})$ for some $\epsilon>0$,

$$\sup_{z\in B_{|x|/2}} |V(x-z)|(1+|x-z|) \leqslant c|x|^{-3},$$

for $|x|$ sufficiently large. Using

$$\int_0^{\delta} \frac{z}{t+z}\,dz = \delta + t\ln\left(\frac{t}{t+\delta}\right) \leqslant \delta + t\left(\frac{t}{t+\delta}-1\right) = \frac{\delta^2}{t+\delta},$$

we compute

$$\int_0^{|x|/2} \frac{|z|^2}{|z|\left(\frac{t}{2}+|z|\right)}\,d|z| \leqslant \frac{1}{2}\,\frac{|x|^2}{t+|x|}.$$

Finally, $\int |V(y)|(1+|y|)\,dy<\infty$, so that altogether

$$\sup_{x\in S_R}|\beta(x,t)| \leqslant c\,\sup_{x\in S_R}\left(|x|^{-3}\frac{|x|^2}{t+|x|} + \frac{1}{|x|(t+|x|)}\right) = \frac{c}{R(t+R)}. \tag{27}$$

We now show that the same bound holds for $\sup_{x\in S_R}|\nabla\beta(x,t)|$ and $R>R_0$, where R_0 is chosen such that all singularities of V lie in the ball with radius R_0. Then

$$|\nabla\beta(x,t)|\leqslant\frac{1}{(2\pi)^{5/2}}\left|\int\frac{V(y)}{|x-y|^2}\int e^{-i(k^2t/2+|k||x-y|)}\hat{\Psi}_{\text{out}}(k)\Phi_+(y,k)dk\,dy\right|$$
$$+\frac{1}{(2\pi)^{5/2}}\left|\int\frac{V(y)}{|x-y|}\int e^{-i(k^2/2+|k||x-y|)}|k|\hat{\Psi}_{\text{out}}(k)\Phi_+(y,k)dk\,dy\right|,\tag{28}$$

where the exchange of differentiation and integration will be justified below. The second term can be treated analogously to $|\beta(x,t)|$, since also $|k|\hat{\Psi}_{\text{out}}(k)\in S$. The first term can as well be estimated along the same lines: in Eq. (26) $|x-y|^2$ will appear in the denominator instead of $|x-y|$, which leads to a stronger bound than (27).

To get (28) from (20) we note that, according to Proposition 2.5, $\Phi(\cdot,k)$ is a classical solution of the stationary Schrödinger equation. Thus, $\Phi(\cdot,k)$ as well as $\eta(\cdot,k)$ are differentiable with respect to *x*, except at the singularities of V. We will show that

$$\nabla_x\eta(x,k)=\nabla_x\left(-\frac{1}{2\pi}\int\frac{e^{-i|k||x-y|}}{|x-y|}V(y)\Phi_+(y,k)dy\right)$$
$$=-\frac{1}{2\pi}\int\nabla_x\left(\frac{e^{-i|k||x-y|}}{|x-y|}\right)V(y)\Phi_+(y,k)dy,\tag{29}$$

and that therefore

$$|\nabla_x\eta(x,k)|\leqslant c_1+c_2|k|,\tag{30}$$

for some $c_1,c_2<\infty$. Then, changing the order of differentiation and integration in the first line of (20) is justified by dominated convergence and (28) follows for all *x* which are not singularities of V.

To get (29) for some $x_0\in\mathbb{R}^3$ that is no singularity of V, we split the domain of integration into $B_{2R}(x_0):=\{y\in\mathbb{R}^3:|x_0-y|\leqslant 2R\}$ and its complement $B^c_{2R}(x_0)$, where R is chosen such that $B_{2R}(x_0)$ contains no singularity of V. Then one can change the order of integration and differentiation in the $B^c_{2R}(x_0)$ term, since there the integrand is bounded by an integrable function uniformly in x for $x\in B_R(x_0)$. To see that the $B_{2R}(x_0)$ term can be made arbitrarily small by appropriately choosing R, we write down the difference quotient for this term. Using that $\sup_{y\in B_{2R}(x_0)}(V(y)\Phi+(y,k))\leqslant c_k<\infty$ and that $|re^{i\theta}-(r+\Delta r)e^{i(\theta+\Delta\theta)}|\leqslant|r\Delta\theta|+|\Delta r|$, we compute

$$\lim_{|\epsilon|\to 0}\frac{1}{|\epsilon|}\left|\int_{B_{2R}(x_0)}\left(\frac{e^{-i|k||x+\epsilon-y|}}{|x+\epsilon-y|}-\frac{e^{-i|k||x-y|}}{|x-y|}\right)V(y)\Phi_+(y,k)dy\right|$$
$$\leqslant\lim_{|\epsilon|\to 0}\frac{1}{|\epsilon|}c_k\int_{B_{2R}(x_0)}\left|\frac{e^{i|k||x+\epsilon-y|}|x-y|-e^{-i|k||x-y|}|x+\epsilon-y|}{|x+\epsilon-y||x-y|}\right|dy$$
$$\leqslant\lim_{|\epsilon|\to 0}\frac{1}{|\epsilon|}c_k\int_{B_{2R}(x_0)}\frac{|x-y||k||\epsilon|+|\epsilon|}{|x+\epsilon-y||x-y|}dy$$
$$\leqslant\lim_{|\epsilon|\to 0}c_k\int_{B_{2R}(x_0)}\left(\frac{|k|}{|x+\epsilon-y|}+\frac{1}{|x+\epsilon-y||x-y|}\right)dy\leqslant c_k12\pi R,$$

where the last inequality is justified by elementary integrations. The bound (30) can be obtained by a simple calculation.

From that we also conclude that $\Psi_t(x)$ is differentiable outside the singularities of $V(x)$ since α and β are. □

Remark 2.7: We used the domain $\Psi_{\text{out}} \in S$ to simplify the proof and avoid tedious estimates. A more detailed analysis of the proof shows that Theorem 2.1 also holds for $\Psi_{\text{out}} \in L^2$ such that $(1+k^2)^{(q+p)/2}(1+\Delta)^{p/2}\hat{\Psi}_{\text{out}} \in L^2$ for some $p > \frac{7}{2}$ and $q > \frac{9}{2}$. Then $(1+k^2)^{q/2}(1+\Delta)^{p/2}e^{-i(k^2/2)t}\hat{\Psi}_{\text{out}} \in L^2$ for all $t \in \mathbb{R}$, and this is enough regularity to prove the free theorem as well as our estimates.

III. REGULARITY OF THE GENERALIZED EIGENFUNCTIONS

In this section we will prove a theorem about the generalized eigenfunctions that connects the differentiability of $\Phi(x,k)$ with respect to k with the behavior of the potential at infinity and gives uniform bounds on $\Phi(x,k)$ and its partial derivatives. At the end of the section we state two simple corollaries that show other applications of our results.

Theorem 3.1: Let the potential satisfy the condition $(\mathbf{V})_{\mathbf{n}}$ for some $n \geqslant 3$, $n \in \mathbb{N}$. Then we have the following.

(i) $\Phi_{\pm}(x,\cdot) \in C^{n-2}(\mathbb{R}^3 \backslash \{0\})$ for all $x \in \mathbb{R}^3$ and the partial derivatives $\partial_k^{\alpha}\Phi_{\pm}(x,k)$, $|\alpha| \leqslant n-2$, are continuous with respect to x and k.[17]

(ii) If, in addition, zero is not an eigenvalue or a resonance of H, then

$$\sup_{x \in \mathbb{R}^3, k \in \mathbb{R}^3} |\Phi_{\pm}(x,k)| < \infty,$$

and for any α with $|\alpha| \leqslant n-2$ there is a $c_{\alpha} < \infty$ such that

$$\sup_{k \in \mathbb{R}^3 \backslash \{0\}} |\partial_k^{\alpha}\Phi_{\pm}(x,k)| < c_{\alpha}(1+|x|)^{|\alpha|}.$$

Remark 3.2: Proposition 2.5 (vi) follows from Theorem 3.1 by taking $n=3$.

Proof (of Theorem 3.1): To simplify notation we will give the proof for $\Phi_{+}(x,k) =: \Phi(x,k)$ since the proof for $\Phi_{-}(x,k)$ is exactly the same apart from the change of some signs.

The structure of the proof will be as follows: First we introduce some notation and results from Ikebe and Povzner that we will use frequently. Then part (i) of Theorem 3.1 is shown for $|\alpha|=1$ involving several lemmas and results proven in the appendix. The generalization to $|\alpha| \geqslant 1$ will be sketched afterward.

In the proof of part (ii) of Theorem 3.1 we will establish boundedness of $\Phi(x,k)$ for k near zero and for $|k| \to \infty$ separately in two propositions.

We start with an investigation of Eq. (4). If $\Phi(x,k) = e^{ik\cdot x} + \eta(x,k)$ is a continuous solution of the Lippmann–Schwinger equation (4) with $\lim_{|x|\to\infty} \eta(x,k) = 0$ for $k \in \mathbb{R}^3$, then $\eta(x,k)$ is a solution of the integral equation

$$\eta(x,k) = -\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|} V(y)[e^{ik\cdot y} + \eta(y,k)]dy, \tag{31}$$

and $\eta(\cdot,k) \in C_{\infty}(\mathbb{R}^3)$. Therefore Eq. (31) is examined on the Banach space $B = C_{\infty}(\mathbb{R}^3)$; the set of continuous functions vanishing at infinity, equipped with the norm $\|f\|_B = \sup_{x \in \mathbb{R}^3}|f(x)|$. $\mathcal{L}(B)$ denotes the space of bounded linear operators mapping B into itself, equipped with the operator norm. Following Ikebe,[13] we define the linear operators $T_k \in \mathcal{L}(B)$, $k \in \mathbb{R}^3$, by

$$(T_k f)(x) = -\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|} V(y) f(y) dy. \tag{32}$$

Since we will make use of some results of Ikebe and Povzner, we state them as a lemma.

Lemma 3.3: Let the potential satisfy the condition $(\mathbf{V})_{\mathbf{3}}$. Then the following holds.

(i) The operator $T_k \in \mathcal{L}(B)$ defined in (32) is compact for all $k \in \mathbb{R}^3$.

(ii) Let $f(x)$ be a bounded continuous function on $\mathbb{R}^3$; then

$$h(x) := -\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|} V(y) f(y) dy$$

is an element of B for all $k \in \mathbb{R}^3$ and $h(x) = O(|x|^{-1})$ for $|x| \to \infty$.

(iii) Let

$$g(x,k) := -\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|} V(y) e^{ik\cdot y}\, dy = (T_k e^{ik\cdot\cdot})(x);$$

then $g(\cdot,k) \in B$ for all $k \in \mathbb{R}^3$ and $g(\cdot,k)$ is continuous with respect to k.

(iv) Let $f(\cdot,k) \in B$ be a solution of the homogeneous equation $f(\cdot,k) = T_k f(\cdot,k)$ for $k \in \mathbb{R}^3$. If $|k|>0$, then $f=0$, and if $k=0$ then $(-\frac{1}{2}\Delta + V(x)) f(x,0) = 0$.

(v) The map $T:\mathbb{R}^3 \to \mathcal{L}(B)$, $k \mapsto T_k$ is continuous.

For the proofs of (i), (ii), (iii), and (iv) see Ikebe;[13] for the proof of (v) see Povzner.[18]

Since we will use similar reasoning, we will briefly repeat Ikebe's proof of the existence of continuous solutions of Eq. (31), starting from Lemma 3.3. Equation (31) now reads as

$$\eta(\cdot,k) = g(\cdot,k) + T_k \eta(\cdot,k). \tag{33}$$

According to Lemma 3.3 (iv) the homogeneous equation $\eta(\cdot,k) = T_k\eta(\cdot,k)$ has only the trivial solution $\eta(x,k)=0$ if $k\neq 0$. Thus 1 is not an eigenvalue of T_k and therefore 1 is in the resolvent set since T_k is compact,[19] i.e., $(1-T_k)^{-1} \in \mathcal{L}(B)$ exists. The unique solution of (31) for $|k|>0$ is then given by

$$\eta(\cdot,k) = (1-T_k)^{-1} g(\cdot,k). \tag{34}$$

Since $\mathcal{L}(B)$ is a Banach algebra in which the map $A \mapsto A^{-1}$ is continuous,[19] from Lemma 3.3 (v) it follows that $(1-T_k)^{-1}$ is continuous in k. Thus, since according to Lemma 3.3 (iii) $g(\cdot,k)$ is continuous with respect to k, we have that $\eta(x,k)$ is continuous with respect to k.

We will now prove part (i) of Theorem 3.1 for $|\alpha|=1$ and assume $(\mathbf{V})_3$. The generalization to $|\alpha|>1$ will then be immediate. Consider arbitrary $l \in \{1,2,3\}$ and $k^0 \in \mathbb{R}^3 \setminus \{0\}$. We use the following notation: k_l denotes the lth Cartesian coordinate of a vector $k \in \mathbb{R}^3$ and $k_{\bar{l}}$ the tuple of the other coordinates. Symbolically we will write $k = (k_l, k_{\bar{l}})$.

By (formally) differentiating (31) we obtain

$$\frac{\partial}{\partial k_l}\eta(x,k) = \frac{\partial}{\partial k_l} g(x,k) + \frac{i}{2\pi}\frac{k_l}{|k|}\int e^{-i|k||x-y|} V(y)\,\eta(y,k) dy - \frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|} V(y) \frac{\partial}{\partial k_l}\eta(y,k) dy. \tag{35}$$

Assume that for $k_l \in I_l := [k_l^0 - \delta_l, k_l^0 + \delta_l]$ and $k_{\bar{l}} \in I_{\bar{l}} := [k_{\bar{l}}^0 - \delta_{\bar{l}}, k_{\bar{l}}^0 + \delta_{\bar{l}}]$, where δ_l and $\delta_{\bar{l}}$ are chosen such that, in particular, $0 \notin I := I_l \times I_{\bar{l}}$, the equation

$$\xi(x,k) = \frac{\partial}{\partial k_l} g(x,k) + \frac{i}{2\pi}\frac{k_l}{|k|}\int e^{-i|k||x-y|} V(y) \left[\int_{k_l^0}^{k_l} \xi(y,(k_l', k_{\bar{l}})) dk_l' + \eta(y,(k_l^0, k_{\bar{l}}))\right] dy - \frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|} V(y)\xi(y,k) dy, \tag{36}$$

which arises from (35) by substituting $\eta(x,k) = \int_{k_l^0}^{k_l} \xi(x,(k_l', k_{\bar{l}})) dk_l' + \eta(x,(k_l^0, k_{\bar{l}}))$, has a continuous solution $\xi(x,k)$. Integrating (36) with respect to k_l and using Fubini's theorem we get

$$
\begin{aligned}
\int_{k_l^0}^{k_l}\xi(x,(k_l',k_{\bar{l}}))dk_l' &= g(x,k)-g(x,(k_l^0,k_{\bar{l}}))\\
&\quad -\frac{1}{2\pi}\int V(y)\left[\frac{e^{-i|k||x-y|}}{|x-y|}\left(\int_{k_l^0}^{k_l'}\xi(y,(k_l'',k_{\bar{l}}))dk_l''+\eta(y,(k_l^0,k_{\bar{l}}))\right)\right]_{k_l^0}^{k_l}dy\\
&= g(x,k)-g(x,(k_l^0,k_{\bar{l}}))+\frac{1}{2\pi}\int\frac{e^{-i|(k_l^0,k_{\bar{l}})||x-y|}}{|x-y|}V(y)\,\eta(y,(k_l^0,k_{\bar{l}}))dy\\
&\quad -\frac{1}{2\pi}\int\frac{e^{-i|k||x-y|}}{|x-y|}V(y)\left(\int_{k_l^0}^{k_l}\xi(y,(k_l',k_{\bar{l}}))dk_l'+\eta(y,(k_l^0,k_{\bar{l}}))\right)dy. \qquad (37)
\end{aligned}
$$

Since $\eta(x,(k_l^0,k_{\bar{l}}))$ is a solution of (31), the second and third term of the right-hand side of (37) combine to $-\eta(x,(k_l^0,k_{\bar{l}}))$. Therefore (37) simply reads as

$$
\begin{aligned}
&\eta(x,(k_l^0,k_{\bar{l}}))+\int_{k_l^0}^{k_l}\xi(x,(k_l',k_{\bar{l}}))dk_l'\\
&\quad = g(x,k)-\frac{1}{2\pi}\int\frac{e^{-i|k||x-y|}}{|x-y|}V(y)\left(\eta(y,(k_l^0,k_{\bar{l}}))+\int_{k_l^0}^{k_l}\xi(y,(k_l',k_{\bar{l}}))dk_l'\right)dy.
\end{aligned}
$$

In other words, if $\xi(x,k)$ is a continuous solution of (36), then the function $f(x,k) = \eta(x,(k_l^0,k_{\bar{l}}))+\int_{k_l^0}^{k_l}\xi(x,(k_l',k_{\bar{l}}))dk_l'$ is a solution of Eq. (31). Since (31) has a unique solution in B, we may conclude that $f(x,k)=\eta(x,k)$, i.e., that $\partial_{k_l}\eta(x,k)=\xi(x,k)$ for $x\in\mathbb{R}^3$ and $k\in I$ once we have shown that $f(\cdot,k)\in B$.

We show now that Eq. (36) has a solution $\xi(x,k)$ that is continuous with respect to $x\in\mathbb{R}^3$ and $k\in I$, such that $\eta(\cdot,(k_l^0,k_{\bar{l}}))+\int_{k_l^0}^{k_l}\xi(\cdot,(k_l',k_{\bar{l}}))dk_l'\in B$. From the physical argument that $\Phi(x,k) \sim e^{ik\cdot x}+e^{i|k||x|}/|x|$ for $|x|\to\infty$, we expect $|\nabla_k\eta(x,k)|\sim e^{i|k||x|}$, $|x|\to\infty$, to be a uniformly bounded function, but we will only show that $|\nabla_k\eta(x,k)|\leq c(1+|x|)^s$ for any $s>0$. We start by multiplying Eq. (36) by $\langle x\rangle^{-s}:=(1+|x|)^{-s}$, $s>0$:

$$
\begin{aligned}
\tilde{\xi}(x,k) &= \langle x\rangle^{-s}\,\partial_{k_l}g(x,k)-\frac{1}{2\pi}\int\frac{e^{-i|k||x-y|}}{\langle x\rangle^s|x-y|}\langle y\rangle^s V(y)\tilde{\xi}(y,k)dy\\
&\quad +\frac{i}{2\pi}\frac{k_l}{|k|}\int\frac{e^{-i|k||x-y|}}{\langle x\rangle^s}\langle y\rangle^s V(y)\left[\int_{k_l^0}^{k_l}\tilde{\xi}(y,(k_l',k_{\bar{l}}))dk_l'+\langle y\rangle^{-s}\eta(y,(k_l^0,k_{\bar{l}}))\right]dy.
\end{aligned}
\qquad (38)
$$

To see that $\xi(x,k)=\langle x\rangle^s\tilde{\xi}(x,k)\leq c(1+|x|)^s$, we show that (38) has a unique solution in

$$\tilde{B}:=\{f(x,k)\in C(\mathbb{R}^3\times I):\ \lim_{|x|\to\infty}\sup_{k\in I}|f(x,k)|=0\}.$$

In the appendix we prove that $\tilde{B}$ equipped with the norm $\|f\|_{\tilde{B}}=\sup_{x\in\mathbb{R}^3,k\in I}|f(x,k)|$ is a Banach space (see Lemma A.1); and that for $f\in\tilde{B}$ the operators,

$$(\tilde{T}f)(x,k) := -\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s|x-y|}\langle y\rangle^s V(y)f(y,k)dy,$$

$$(\tilde{T}'f)(x,k) := \frac{i}{2\pi}\frac{k_l}{|k|}\int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s}\langle y\rangle^s V(y)f(y,k)dy,$$

and

$$(\tilde{K}f)(x,k) := \int_{k_l^0}^{k_l} f(x,(k_l',k_{\bar{l}}))dk_l',$$

belong to $\mathcal{L}(\tilde{B})$ if $s>0$ is chosen, such that $\langle x\rangle^s V(x)$ still satisfies $(\mathbf{V})_3$ (see Lemma A.2).

Noting that $\langle x\rangle^{-s}\,\partial_{k_l}g(x,k)$ and $\langle x\rangle^{-s}\,\eta(x,(k_l^0,k_{\bar{l}}))$ belong to $\tilde{B}$ (see Lemma A.2), Eq. (38) can be written as

$$\tilde{\xi} = \langle\cdot\rangle^{-s}\,\partial_{k_l}g + \tilde{T}'\tilde{K}\tilde{\xi} + \tilde{T}'\langle\cdot\rangle^{-s}\,\eta(\cdot,(k_l^0,k_{\bar{l}})) + \tilde{T}\tilde{\xi},$$

where $\langle\cdot\rangle^{-s}$ denotes the operator of multiplication with $\langle x\rangle^{-s}$ in $\tilde{B}$. To prove that this equation has a unique solution in $\tilde{B}$, we show that $(1-\tilde{T})^{-1}\in\mathcal{L}(\tilde{B})$ exists, and that

$$\tilde{\xi} = (1-\tilde{T})^{-1}(\langle\cdot\rangle^{-s}\partial_{k_l}g + \tilde{T}'\langle\cdot\rangle^{-s}\,\eta(\cdot,(k_l^0,k_{\bar{l}}))) + (1-\tilde{T})^{-1}\tilde{T}'\tilde{K}\tilde{\xi} \tag{39}$$

has a unique solution. The former will be the content of Lemma 3.4, and to see the latter note that, according to Lemma A.2.(i),

$$\|(1-\tilde{T})^{-1}\tilde{T}'\tilde{K}\|_{\mathcal{L}(\tilde{B})} \le \|(1-\tilde{T})^{-1}\tilde{T}'\|_{\mathcal{L}(\tilde{B})} 2\delta_l .$$

Also $\|(1-\tilde{T})^{-1}\|_{\mathcal{L}(\tilde{B})}$ and $\|\tilde{T}\|_{\mathcal{L}(\tilde{B})}$ depend on δ_l since be space $\tilde{B}$ itself depends on δ_l. But the norm of these operators decreases as δ_l decreases, since according to Lemma A.1.(ii) and the constructions in the proofs of Lemma A.2 and Lemma 3.4, $\|\tilde{T}'\|_{\mathcal{L}(\tilde{B})} \le \sup_{k\in I}\|T_k'^s\|_{\mathcal{L}(B)}$ and $\|(1-\tilde{T})^{-1}\|_{\mathcal{L}(\tilde{B})} \le \sup_{k\in I}\|(1-T_k^s)^{-1}\|_{\mathcal{L}(B)}$. Thus, one can choose δ_l such that

$$\|(1-\tilde{T})^{-1}\tilde{T}'\tilde{K}\|_{\mathcal{L}(\tilde{B})} < 1.$$

Then (39) has a unique solution $\tilde{\xi}\in\tilde{B}$ since $(1-\tilde{T})^{-1}\tilde{T}'\tilde{K}$ is a contraction in a complete metric space.

Now $\xi(x,k) = \langle x\rangle^s\tilde{\xi}(x,k)$ is a solution of (36) and $f(x,k) = \eta(x,(k_l^0,k_{\bar{l}})) + \int_{k_l^0}^{k_l}\xi(x,(k_l',k_{\bar{l}}))dk_l'$ is a solution of (31). Recall that to conclude $f(x,k) = \eta(x,k)$, i.e., that ξ is the partial derivative of η with respect to k_l, we need to show $f(\cdot,k)\in B$. By construction, $\sup_{k\in I}|\xi(x,k)| = O(|x|^s)$ for $|x|\to\infty$ and therefore also $|f(x,k)| = O(|x|^s)$ for any $k\in I$. Thus, writing $V(x)f(x,k) = \langle x\rangle^s V(x)\langle x\rangle^{-s}f(x,k)$ and observing that $\langle x\rangle^s V(x)$ satisfies $(\mathbf{V})_3$ and $\langle x\rangle^{-s}f(x,k)$ is bounded we use Lemma 3.3 (ii) to conclude that $f(\cdot,k)\in B$.

To complete the proof of part (i) for $|\alpha|=1$ we need to show the following lemma.

Lemma 3.4: $(1-\tilde{T})^{-1}\in\mathcal{L}(\tilde{B})$ exists.

Proof: First we show that $(1-T_k^s)^{-1}\in\mathcal{L}(B)$ exists, where $T_k^s := \langle\cdot\rangle^{-s}T_k\langle\cdot\rangle^s\in\mathcal{L}(B)$. Since $\langle x\rangle^s V(x)$ meets the requirements of Lemma 3.3 and multiplication by $\langle x\rangle^{-s}$ is a bounded operation on B,T_k^s is compact. Therefore $(1-T_k^s)^{-1}$ exists if the homogeneous equation $(1-T_k^s)f_s = 0$ has only the trivial solution $f_s=0$. Now let $f_s\in B$ be a solution of the homogeneous equation, which explicitly reads as

$$f_s(x)=(T_k^s f_s)(x)=-\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s|x-y|}V(y)\langle y\rangle^s f_s(y)dy.$$

Then $f(x):=\langle x\rangle^s f_s(x)$ is a solution of

$$f(x)=-\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|}V(y)f(y)dy,$$

and Lemma 3.3 (ii) implies $f\in B$ since $\langle x\rangle^s V(x)$ satisfies $(\mathbf{V})_\mathbf{3}$ and $\langle x\rangle^{-s}f(x)$ is bounded. Using Lemma 3.3 (iv), we conclude that $f(x)=f_s(x)=0$ for $k\neq 0$. Therefore $(1-T_k^s)^{-1}$ exists for any $k\in I$. The continuity of $(1-T_k^s)^{-1}$ with respect to k follows again from the continuity of $(1-T_k^s)$.

Using Lemma A.1 (ii), we define the operator $(1-\tilde{T})^{-1}\in\mathcal{L}(\tilde{B})$. Since $(1-T_k^s)(1-T_k^s)^{-1}=(1-T_k^s)^{-1}(1-T_k^s)=1$ holds for all $k\in I$, also $(1-\tilde{T})(1-\tilde{T})^{-1}=(1-\tilde{T})^{-1}(1-\tilde{T})=1$ holds. □

It is now easy to prove the existence of higher-order derivatives by induction. From the proof for $|\alpha|=1$ we conclude that if $\eta(x,k)\in B$ is a solution of (31) then $\langle x\rangle^{-s}\,\partial_{k_l}\eta(x,k)$ is given by the unique solution $\xi(x,k)$ in B of

$$\xi(x,k)=\langle x\rangle^{-s}\,\partial_{k_l}g(x,k)+\frac{i}{2\pi}\frac{k_l}{|k|}\int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s}V(y)\,\eta(y,k)dy$$
$$-\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s|x-y|}\langle y\rangle^s V(y)\,\xi(y,k)dy,$$

for any $k\in\mathbb{R}^3\backslash\{0\}$. In general, assume that $\eta(x,\cdot)\in C^p(\mathbb{R}^3\backslash\{0\})$ for some $p<n-2$ and that $\langle x\rangle^{-s-p+1}\partial_k^\alpha\eta(x,k)$, $|\alpha|=p$, is given by the unique solution $\xi(x,k)$ of

$$\xi(x,k)=\langle x\rangle^{-s-p+1}[\partial_k^\alpha g(x,k)+\partial_k^\alpha(T_k\eta)(x,k)-(T_k\partial_k^\alpha\eta)(x,k)]+(T_k^{s+p-1}\xi)(x,k), \quad (40)$$

in B, where T_k^{s+p-1} is given by

$$(T_k^{s+p-1}f)(x):=-\frac{1}{2\pi}\int \frac{e^{-i|k||x-y|}}{\langle x\rangle^{s+p-1}|x-y|}V(y)\langle y\rangle^{s+p-1}f(y)dy.$$

Then one can prove by exactly the same method as in the case $|\alpha|=1$ that $\partial_{k_l}\xi(x,k)$ exists: Equation (40) is analogous to (33), where g is replaced by $\langle x\rangle^{-s-p+1}[\partial_k^\alpha g(x,k)+\partial_k^\alpha(T_k\eta)(x,k)-(T_k\,\partial_k^\alpha\eta)(x,k)]\in B$ and T_k by T_k^{s+p-1}. As long as $\langle y\rangle^{s+p-1}V(y)$ satisfies $(\mathbf{V})_\mathbf{3}$, the proof of differentiability of the solution of (40) can be done along the same lines as for $|\alpha|=1$.

Proof [of part (ii) of Theorem 3.1]: From the continuity of Φ and the fact that $\lim_{|x|\to\infty}(\Phi(x,k)-e^{ik\cdot x})=0$ for all $k\neq 0$, Ikebe already concluded that for compact $D\subset\mathbb{R}^3\backslash\{0\}$,

$$\sup_{x\in\mathbb{R}^3,k\in D}|\Phi(x,k)|<\infty,$$

holds. It remains to examine the cases $k\to 0$ and $|k|\to\infty$. If H has a zero-energy resonance or eigenvalue, according to Jensen and Kato,[14] the spectral density is singular at $E=0$. Since the spectral density and the generalized eigenfunctions are closely related,[20] we expect that in this case also the generalized eigenfunctions become singular at $k=0$.

But assuming that H has neither a resonance nor an eigenvalue at $E=0$, the eigenfunctions are uniformly bounded near $k=0$:

Proposition 3.5: Let the potential satisfy $(\mathbf{V})_\mathbf{n}$ for some $n\geqslant 3$. If H has no zero-energy resonance or eigenvalue, then for any compact $D\subset\mathbb{R}^3$,

$$\sup_{x\in\mathbb{R}^3,k\in D}|\Phi(x,k)|<\infty,$$

and, for any α with $|\alpha|<n-2$, there is a $c_\alpha<\infty$ such that

$$\sup_{k\in D\setminus\{0\}}|\partial_k^\alpha\Phi(x,k)|<c_\alpha(1+|x|)^{|\alpha|}.$$

Proof: If H has no zero-energy resonance or eigenvalue, the homogeneous equation $f=T_0f$ has no solution in B since, according to Lemma 3.3, under the conditions $(\mathbf{V})_\mathbf{3}$ any solution of $f=T_0f$ is a solution of $(-\frac{1}{2}\Delta+V(x))f(x)=0$ with $f(x)=O(|x|^{-1})$. And a solution $f\in B$ of $Hf=0$ with $f(x)=O(|x|^{-1})$ is, in particular, a resonance. Thus, either $f\in L^2$, i.e., zero is an eigenvalue, or $f\notin L^2$, but then zero is a resonance.

Thus $(1-T_k)^{-1}$ exists for all $k\in D$ and, recalling $\eta(x,k)=((1-T_k)^{-1}g)(x,k)$,

$$\sup_{x\in\mathbb{R}^3,k\in D}|\eta(x,k)|=\sup_{k\in D}\|\eta_k\|_B\leqslant\sup_{k\in D}\|(1-T_k)^{-1}\|_{\mathcal{L}(B)}\|g_k\|_B<\infty,$$

since $\|(1-T_k)^{-1}\|_{\mathcal{L}(B)}$ is a continuous function on a compact set and therefore bounded. Recalling $\Phi(x,k)=e^{ik\cdot x}+\eta(x,k)$, the proof of the first statement is complete.

The bounds for the partial derivatives near zero also follow from the fact that $(1-T_0)^{-1}$ exists if zero is neither a resonance nor an eigenvalue of H. To see this we introduce spherical coordinates $(|k|,\omega)$, $|k|\in(0,\infty)$ and $\omega\in S^2$ for k. If we replace $\partial_{k_i}|k|=k_i/|k|=\omega\cdot e_i$ in Eq. (40), it has a unique solution $\xi(\cdot,|k|,\omega)\in B$ also for $|k|=0$. Thus, $\lim_{|k|\to 0}\partial_k^\alpha\eta(x,|k|,\omega)$ exists for all $\omega\in S^2$. As in the first part of this proof,

$$\sup_{x\in\mathbb{R}^3,k\in D\setminus\{0\}}\frac{\partial_k^\alpha\eta(x,k)}{\langle x\rangle^{s+|\alpha|-1}}\leqslant\sup_{x\in\mathbb{R}^3,|k|\in[0,R],\omega\in S^2}|\xi(x,|k|,\omega)|<\infty,$$

for some R such that $D\subset K_R$, follows from the fact that $\xi(\cdot,|k|,\omega)\in B$ depends continuously on k. Noting $|\partial_k^\alpha e^{ik\cdot x}|=|x_1^{\alpha_1}x_2^{\alpha_2}x_3^{\alpha_3}e^{ik\cdot x}|<\langle x\rangle_{|\alpha|}$ completes the proof. □

To prove the uniform bound on Φ and its derivatives, it remains to examine their behavior for large k. This can be done using the Born series. As expected on physical grounds, the generalized eigenfunctions for large momentum are essentially plane waves.

Proposition 3.6: Let the potential V satisfy $(\mathbf{V})_\mathbf{n}$ for some $n\geqslant 3$. Then

$$\lim_{|k|\to\infty}\sup_{x\in\mathbb{R}^3}\langle x\rangle^{-|\alpha|}|\partial_k^\alpha\Phi(x,k)-\partial_k^\alpha e^{ik\cdot x}|=0,$$

for every $|\alpha|\leqslant n-2$.

Proof: First we will show that the function $\eta(x,k)=\Phi(x,k)-e^{ik\cdot x}$ converges uniformly to zero for $|k|\to\infty$. Recall (34):

$$\eta(\cdot,k)=(1-T_k)^{-1}g(\cdot,k).$$

We shall show that $\lim_{|k|\to\infty}\|g(\cdot,k)\|_B=0$, but we have no simple control of the norm of $(1-T_k)^{-1}$, for example, in terms of the Born series, since

$$\|T_k\|_{\mathcal{L}(B)}=\sup_{x\in\mathbb{R}^3}\int\frac{|V(y)|}{|x-y|}dy=\text{const},$$

does not depend on $|k|$. Following Zemach and Klein[21] we iterate Eq. (33) once and obtain

$$\eta(\cdot,k)=g(\cdot,k)+T_kg(\cdot,k)+T_k^2\eta(\cdot,k), \tag{41}$$

with the formal solution

$$\eta(\cdot,k)=(1-T_k^2)^{-1}(g(\cdot,k)+T_k g(\cdot,k)). \tag{42}$$

If Eq. (41) has a unique solution, it must equal the unique solution of (33), since any solution of (33) is clearly also a solution of (41). We shall now first establish that (a) $(1-T_k^2)^{-1}\to 1$ for $|k|\to\infty$ and then that (b) $\|g(\cdot,k)+T_k g(\cdot,k)\|_B\to 0$ for $|k|\to\infty$, since then

$$\lim_{|k|\to\infty}\|\eta(\cdot,k)\|_B \leqslant \lim_{|k|\to\infty}\|(1-T_k^2)^{-1}\|_{\mathcal{L}(B)}\|(g(\cdot,k)+T_k g(\cdot,k)\|_B=0.$$

(a) follows from the following.

Lemma 3.7: Let $V\in L^1\cap L^2$; then

$$\lim_{|k|\to\infty}\|T_k^2\|_{\mathcal{L}(B)}=0. \tag{43}$$

Here (43) also holds, if T_k^2 is understood as an operator on bounded continuous functions.

Now $|k|$ can be chosen such that $\|T_k^2\|_{\mathcal{L}(B)}<1$, and then $(1-T_k^2)^{-1}$ is given as the norm convergent Born series:

$$(1-T_k^2)^{-1}=\sum_{n=0}^{\infty}(T_k^2)^n.$$

Thus, $\lim_{|k|\to\infty}\|(1-T_k^2)^{-1}-1\|_{\mathcal{L}(B)}=0.$

Proof (of Lemma 3.7): We compute for $f\in B$,

$$\begin{aligned}(T_k^2 f)(x)&=\frac{1}{4\pi^2}\int\frac{e^{-i|k||x-y|}}{|x-y|}V(y)\int\frac{e^{-i|k||y-z|}}{|y-z|}V(z)f(z)dz\,dy\\&=\frac{1}{4\pi^2}\int V(z)f(z)\int\frac{e^{-i|k|(|x-y|+|y-z|)}}{|x-y||y-z|}V(y)dy\,dz\\&=\frac{1}{4\pi^2}\int\frac{V(z)}{|x-z|}f(z)\left[|x-z|\int\frac{e^{-i|k|(|x-y|+|y-z|)}}{|x-y||y-z|}V(y)dy\right]dz\\&=:\frac{1}{4\pi^2}\int\frac{V(z)}{|x-z|}f(z)I^V(x,z,|k|)dz.\end{aligned}$$

Zemach and Klein[21] showed that, for $V\in C_0^1(\mathbb{R}^3)$,

$$\lim_{|k|\to\infty}\sup_{x,z\in\mathbb{R}^3}|I^V(x,z,|k|)|=0,$$

i.e., that $\lim_{|k|\to\infty}\|T_k^2\|_{\mathcal{L}(B)}=0$ holds for $V\in C_0^1(\mathbb{R}^3)$.

Potentials $V\in L^1\cap L^2$ will be approximated in the following norm:

$$|||V|||=\sup_{x\in\mathbb{R}^3}\int\frac{|V(y)|}{|x-y|}dy.$$

Observing that

$$|||V|||\leqslant\sup_{x\in\mathbb{R}^3}\int_{|x-y|<1}\frac{|V(y)|}{|x-y|}dy+\sup_{x\in\mathbb{R}^3}\int_{|x-y|\geqslant 1}\frac{|V(y)|}{|x-y|}dy\leqslant c(\|V\|_{L^2}+\|V\|_{L^1})<\infty, \tag{44}$$

where we used Schwarz's inequality for the $\|V\|_{L^2}$ term, we conclude that, since any $V \in L^1 \cap L^2$ can be approximated by a function $U \in C_0^1$ simultaneously in the L^1 and L^2 norm, this is also true for the $\||\cdot\||$ norm. Thus we get the following bound for the norm of T_k^2:

$$\|T_k^2 f\|_B = \sup_{x\in\mathbb{R}^3}\left|\frac{1}{(2\pi)^2}\int \frac{e^{-i|k||x-y|}}{|x-y|}(V(y)-U(y)+U(y))\int \frac{e^{-i|k||y-z|}}{|y-z|}V(z)f(z)dz\,dy\right|$$
$$\leqslant \|f\|_B \||V\||(\||V-U\||+\sup_{x,z\in\mathbb{R}^3}|I^U(x,z,|k|)|).$$

The first term in the brackets becomes small for appropriately chosen $U \in C_0^1$ while the second one converges to zero for $|k|\to\infty$. □

We now proceed to (b), namely, that $\lim_{|k|\to\infty}\|g(\cdot,k)+T_k g(\cdot,k)\|_B=0$. By $T_k g(x,k) = T_k^2 e^{ik\cdot x}$ and Lemma 3.7, $\lim_{|k|\to\infty}\|T_k g(\cdot,k)\|_B=0$ follows immediately. To get the analogous statement for $g(\cdot,k)$ we assume first again $V\in C_0^1(\mathbb{R}^3)$. Then

$$|g(x,k)| = \frac{1}{2\pi}\left|\int \frac{e^{-i|k||x-y|}}{|x-y|}V(y)e^{-ik\cdot(x-y)}\,dy\right|$$
$$\leqslant \frac{1}{2\pi}\left|\int_{\theta<\theta_0}\frac{e^{-i|k||z|(1+\cos\theta)}}{|z|}V(x-z)dz\right| + \frac{1}{2\pi}\int_{\theta\geqslant\theta_0}\frac{|V(x-z)|}{|z|}dz \tag{45}$$

holds. Herein θ denotes the angle between $z=(x-y)$ and k. Stationary phase methods on the first term yields

$$\frac{1}{2\pi}\left|\int_{\theta<\theta_0}\frac{1}{i|k|(1+\cos\theta)}\left(\frac{d}{d|z|}e^{-i|k||z|(1+\cos\theta)}\right)V(x-z)|z|d|z|d\Omega(z)\right|$$
$$=\frac{1}{2\pi}\left|\int_{\theta<\theta_0}\frac{1}{i|k|(1+\cos\theta)}e^{-i|k||z|(1+\cos\theta)}\left(\frac{d}{d|z|}V(x-z)|z|\right)d|z|d\Omega(z)\right|$$
$$\leqslant\frac{1}{2\pi}\frac{1}{|k|(1+\cos\theta_0)}\int\left|\frac{d}{d|z|}V(x-z)|z|\right|d|z|d\Omega(z)\leqslant\frac{c}{|k|(1+\cos\theta_0)}\xrightarrow{|k|\to\infty}0,$$

where

$$\sup_{x\in\mathbb{R}^3}\int\left|\frac{\frac{d}{d|z|}V(x-z)|z|}{|z|^2}\right|dz=c<\infty$$

was used. This follows directly from $V\in C_0^1$.

The second term in (45) is an integration over a cone with opening angle θ_0, where the potential in the integrand has compact support, is bounded and displaced by $-x$. Thus

$$\lim_{\theta_0\to\pi}\sup_{x\in\mathbb{R}^3}\int_{\theta\geqslant\theta_0}\frac{|V(x-z)|}{|z|}dz\leqslant c\lim_{\theta_0\to\pi}\int_{\theta_0}^{\pi}|\sin\theta|d\theta=0,$$

and $\lim_{|k|\to\infty}\sup_{x\in\mathbb{R}^3}|g(x,k)|=0$ follows. To get this for $V\in L^1\cap L^2$, we approximate V by $U \in C_0^1$, as in Lemma 3.7.

Analogously we show that $\langle x\rangle^{-|\alpha|}|\partial_k^\alpha\Phi(x,k)-\partial_k^\alpha e^{ik\cdot x}|$ vanish uniformly for $|k|\to\infty$. According to the proof of part (i) of Theorem 3.1, $\partial_k^\alpha\eta(x,k)$ is obtained as the unique solution of (40) in B. For large $|k|$ the operator $(1-(T_k^{s+p-1})^2)^{-1}$ with $p=|\alpha|$ can be expanded in terms of the Born series, and the solution of the modified equation is given by

$$\xi=(1-(T_k^{s+p-1})^2)^{-1}(\langle\cdot\rangle^{-s-p+1}[\partial_k^\alpha g+\partial_k^\alpha T_k\,\eta-T_k\,\partial_k^\alpha\eta]$$
$$+T_k^{s+p-1}\langle\cdot\rangle^{-s-p+1}[\partial_k^\alpha g+\partial_k^\alpha T_k\,\eta-T_k\,\partial_k^\alpha\eta]).$$

It can be shown by the same methods as in the case of $g(x,k)$ that the term, on which $(1-(T_k^{s+p-1})^2)^{-1}$ acts, uniformly approaches zero for $|k|\to\infty$, which completes the proof of the proposition. □

The uniform boundedness of Φ as well as the bounds on its partial derivatives with respect to k now follow from Proposition 3.5 and Proposition 3.6. □

We will conclude this section with two corollaries to Theorem 3.1 and Proposition 3.6: The first one states that the Riemann–Lebesgue lemma holds also for the generalized Fourier transformation and its inverse. Furthermore, the differentiability of the generalized Fourier transform of a function is connected to its decay as in the case of the ordinary Fourier transform. Related results can be found in a work by Isozaki.[22]

Corollary 3.8: Let V satisfy $(\mathbf{V})_{\mathbf{n}}$ with some $n\geqslant 3$ and let zero not be an eigenvalue or resonance of H. Then, for any $N\leqslant n-2$ and any f such that $\langle x\rangle^N f\in L^1(\mathbb{R}^3)\mathcal{F}_\pm f$ and $\mathcal{F}_\pm^{-1}f$ are in $C^N(\mathbb{R}^3)$ and $\partial_k^\alpha\mathcal{F}_\pm f\in C_\infty(\mathbb{R}^3)$ and $\partial_k^\alpha\mathcal{F}_\pm^{-1}f\in C_\infty(\mathbb{R}^3)$ for all α with $|\alpha|\leqslant N$.

Proof: Let $\langle x\rangle^N f\in L^1$, $0\leqslant|\alpha|\leqslant N$, then, e.g.,

$$\partial_k^\alpha(\mathcal{F}_+f)(k)=\partial_k^\alpha\frac{1}{(2\pi)^{3/2}}\int\Phi_+^*(x,k)f(x)dx$$
$$=\frac{1}{(2\pi)^{3/2}}\int\partial_k^\alpha\Phi_+^*(x,k)f(x)dx$$
$$=\frac{1}{(2\pi)^{3/2}}\int\partial_k^\alpha e^{-ik\cdot x}f(x)dx+\frac{1}{(2\pi)^{3/2}}\int\partial_k^\alpha\eta_+^*(x,k)f(x)dx \tag{46}$$

is bounded and continuous since $|\partial_k^\alpha\Phi_+^*(x,k)|$ is bounded by $c_\alpha\langle x\rangle^{|\alpha|}$ according to Theorem 3.1 and $\langle x\rangle^{|\alpha|}f\in L^1$. Furthermore, the first term in the second line belongs to C_∞ by the ordinary Riemann–Lebesgue lemma and the second term belongs to C_∞ since $\langle x\rangle^{-|\alpha|}|\partial_k^\alpha\eta_+^*(x,k)|$ tends uniformly to zero for $|k|\to\infty$ according to Proposition 3.6. □

The second corollary concerns the so-called T matrix, an object widely discussed in quantum mechanical scattering theory. Let V satisfy $(\mathbf{V})_{\mathbf{3}}$; then the T matrix $\mathbf{T}(\cdot,\cdot)$ is defined by

$$\mathbf{T}(k,k')=(2\pi)^{-3}\int e^{-ik\cdot x}V(x)\Phi_-(x,k')dx. \tag{47}$$

There are several results about the analyticity of the T matrix for potentials with exponential decay.[23] The following corollary gives sufficient conditions for $\mathbf{T}(k,k')$ to be continuously differentiable.

Corollary 3.9: Let V satisfy $(\mathbf{V})_{\mathbf{n}}$ for some $n\geqslant 3$ and let zero be neither a resonance nor an eigenvalue of H. Then (i) $\mathbf{T}(\cdot,\cdot)\in C^{n-3}(\mathbb{R}^3\times(\mathbb{R}^3\setminus\{0\}))$; (ii) for every multi-index α with $|\alpha|\leqslant n-3$,

$$\sup_{k\in\mathbb{R}^3,k'\in\mathbb{R}^3\setminus\{0\}}|\partial_{k'}^\alpha\mathbf{T}(k,k')|<\infty.$$

ACKNOWLEDGMENTS

This work was supported by the DFG. We thank W. Amrein for very helpful correspondence, for Reference 14 and for sending us his preprints 7 and 8 prior to publication. Very helpful discussions with S. Goldstein, A. Soffer and N. Zanghi are gratefully acknowledged.

APPENDIX

In this appendix we prove two lemmas used in Sec. III.

Lemma A.1: (i) The space $\tilde{B}$ equipped with the norm

$$\|f\|_{\tilde{B}} = \sup_{x\in\mathbb{R}^3, k\in I} |f(x,k)|$$

is a Banach space.

(ii) Let $\{A_k\}_{k\in I}\subset\mathcal{L}(B)$ be a family of bounded operators on B such that A_k depends continuously on k with respect to the operator norm. Then

$$(Af)(x,k) := (A_k f(\cdot,k))(x)$$

defines an operator $A\in\mathcal{L}(\tilde{B})$ and $\|A\|_{\mathcal{L}(\tilde{B})}\leqslant \sup_{k\in I}\|A_k\|_{\mathcal{L}(B)}$.

Proof [of part (i)]: Let $\{f_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in $\tilde{B}\subset C(\mathbb{R}^3\times I)$. Then there exists $f\in C(\mathbb{R}^3\times I)$ such that $\lim_{n\to\infty}\|f-f_n\|_{\tilde{B}}=0$. It remains to show that $f\in\tilde{B}$, i.e., that $\lim_{|x|\to\infty}\sup_{k\in I}|f(x,k)|=0$. But

$$\sup_{k\in I}|f(x,k)|\leqslant \sup_{k\in I}|f(x,k)-f_n(x,k)| + \sup_{k\in I}|f_n(x,k)| \leqslant \|f-f_n\|_{\tilde{B}} + \sup_{k\in I}|f_n(x,k)|.$$

The first term can be made arbitrarily small by appropriately choosing n, and the second term vanishes for $|x|\to\infty$.

Proof [of part (ii)]: Let $f\in\tilde{B}$. Then for any fixed $k\in I$, $f(\cdot,k)\in B$ and therefore $A_k f(\cdot,k)\in B$. First we show that $Af(\cdot,\cdot)\in C(\mathbb{R}^3\times I)$:

$$\begin{aligned}|(Af)(x,k)-(Af)(x',k')| &= |(A_k f(\cdot,k))(x)-(A_{k'}f(\cdot,k'))(x')|\\ &\leqslant |(A_k f(\cdot,k))(x)-(A_k f(\cdot,k))(x')| + |((A_k-A_{k'})f(\cdot,k))(x')|\\ &\quad + |(A_{k'}(f(\cdot,k)-f(\cdot,k')))(x')|. \end{aligned} \tag{A1}$$

Since $A_k f(\cdot,k)\in B$, the first term can be made arbitrarily small by choosing $|x-x'|$ small enough. The second term becomes small uniformly in x' by choosing $|k-k'|$ small enough, since

$$\sup_{x'\in\mathbb{R}^3}|((A_k-A_{k'})f(\cdot,k))(x')| = \|(A_k-A_{k'})f(\cdot,k)\|_B \leqslant \|A_k-A_{k'}\|_{\mathcal{L}(B)}\|f\|_{\tilde{B}},$$

and A_k depends continuously on k. The third term in (A1) yields

$$\begin{aligned}\sup_{x'\in\mathbb{R}^3}|(A_{k'}(f(\cdot,k)-f(\cdot,k')))(x')| &\leqslant \|A_{k'}\|_{\mathcal{L}(B)}\|f(\cdot,k)-f(\cdot,k')\|_B\\ &\leqslant c\max\left(\sup_{|x'|>R}|f(x',k)-f(x',k')|,\ \sup_{|x'|\leqslant R}|f(x',k)-f(x',k')|\right),\end{aligned}$$

where we used $\sup_{k\in I}\|A_k\|_{\mathcal{L}(B)}\leqslant c<\infty$. This holds because $\|A_k\|_{\mathcal{L}(B)}$ is a continuous function of k on a compact set. The first term in $\max(\cdots)$ can be made arbitrarily small by choosing R large since $f\in\tilde{B}$. The second term vanishes for $|k-k'|\to 0$ since a continuous function on a compact domain is uniformly continuous.

We now show that

$$\lim_{|x|\to\infty}\sup_{k\in I}(Af)(x,k) = \lim_{|x|\to\infty}\sup_{k\in I}(A_k f(\cdot,k))(x)=0.$$

Suppose that this is wrong; then there exists an $\epsilon > 0$ and a sequence $\{x_n, k_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^3 \times I$ with $\lim_{n\to\infty} x_n = \infty$, such that $|(A_{k_n} f(\cdot, k_n))(x_n)| > \epsilon, \forall n \in \mathbb{N}$. Since I is compact, $\{k_n\}$ contains a convergent subsequence (for simplicity also denoted by $\{k_n\}$) with $\lim_{n\to\infty} k_n = k \in I$. Now

$$|A_{k_n} f(x_n, k_n)| \leqslant |A_{k_n}(f(x_n, k_n) - f(x_n, k))| + |(A_{k_n} - A_k) f(x_n, k)| + |A_k f(x_n, k)|,$$

where the first two terms get arbitrarily small as $n \to \infty$, as has just been shown, and the third term gets arbitrarily small as $n \to \infty$ since $A_k f(\cdot, k) \in B$. Thus, we have a contradiction and $Af \in \tilde{B}$ follows.

The estimate for the norm follows directly from

$$\|Af\|_{\tilde{B}} = \sup_{x\in\mathbb{R}^3, k\in I} |(A_k f(\cdot, k))(x)| \leqslant \sup_{k\in I} \|A_k\|_{\mathcal{L}(B)} \|f(\cdot, k)\|_B \leqslant \sup_{k\in I} \|A_k\|_{\mathcal{L}(B)} \|f\|_{\tilde{B}}.$$

□

Lemma A.2: Let V satisfy $\mathbf{(V)_3}$ and let $s > 0$ such that $\langle x\rangle^s V(x)$ still satisfies $\mathbf{(V)_3}$. For $f \in \tilde{B}$ let

$$(\tilde{T} f)(x,k) := -\frac{1}{2\pi} \int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s |x-y|} \langle y\rangle^s V(y) f(y,k) dy,$$

$$(\tilde{T}' f)(x,k) := \frac{i}{2\pi} \frac{k_l}{|k|} \int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s} \langle y\rangle^s V(y) f(y,k) dy,$$

and

$$(\tilde{K} f)(x,k) := \int_{k_l^0}^{k_l} f(x, (k_l', k_{\bar{l}})) dk_l';$$

then (i) the operators $\tilde{T}$, $\tilde{T}'$, and $\tilde{K}$ belong to $\mathcal{L}(\tilde{B})$ and $\|\tilde{K}\|_{\mathcal{L}(\tilde{B})} \leqslant 2\delta_l$, where $2\delta_l$ is the length of the interval I_l; (ii) the functions $\langle x\rangle^{-s} \partial_{k_l} g(x,k)$ and $\langle x\rangle^{-s} \eta(x, (k_l^0, k_{\bar{l}}))$ belong to $\tilde{B}$.

Proof [of part (i)]: Let $f \in B$ and define

$$(T_k^s f)(x) = -\frac{1}{2\pi} \int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s |x-y|} \langle y\rangle^s V(y) f(y) dy,$$

$$(T_k'^s f)(x) = \frac{i}{2\pi} \frac{k_l}{|k|} \int \frac{e^{-i|k||x-y|}}{\langle x\rangle^s} \langle y\rangle^s V(y) f(y) dy;$$

then for $\tilde{f} \in B$ we have $(\tilde{T}\tilde{f})(x,k) = (T_k^s \tilde{f}(\cdot, k))(x)$ and analogously for $\tilde{T}'$. We shall use Lemma A.1 (ii) to prove that $\tilde{T}$ and $\tilde{T}'$ are in $\mathcal{L}(\tilde{B})$. We have to show that $(T_k^s)_{k\in I}$ and $(T_k'^s)_{k\in I}$ are families of operators in $\mathcal{L}(B)$ continuously depending on k.

Now $\langle y\rangle^s V(y)$ still satisfies the conditions $\mathbf{(V)_3}$. According to Lemma 3.3 (i) and (v) $\langle x\rangle^s T_k^s$ satisfies the conditions of Lemma A.1 (ii). Multiplication by $\langle x\rangle^{-s}$ is a bounded operation in B, and thus also $(T_k^s)_{k\in I}$ satisfies the conditions of Lemma A.1 (ii). Hence $\tilde{T} \in \mathcal{L}(\tilde{B})$.

Next consider $T_k'^s$. From $\langle y\rangle^s V(y) \in L^1$ and

$$|(T_k'^s f)(x)| \leqslant \langle x\rangle^{-s} \frac{1}{2\pi} \int |\langle y\rangle^s V(y)| dy \|f\|_B,$$

$\lim_{|x|\to\infty} |(T_k'^s f)(x)| = 0$ follows. With $|e^{i\alpha} - e^{i\alpha'}| \leqslant |\alpha - \alpha'|$ for $\alpha, \alpha' \in \mathbb{R}$ we estimate

$$|\langle x\rangle^s(T_k'^s f)(x)-\langle x'\rangle^s(T_k'^s f)(x')| \leq \frac{1}{2\pi}\int |e^{-i|k||x-y|}-e^{-i|k||x'-y|}||\langle y\rangle^s V(y)||f(y)|dy$$

$$\leq \frac{1}{2\pi}\int |k|||x-y|-|x'-y|||\langle y\rangle^s V(y)|dy\|f\|_B$$

$$\leq |x-x'||k|\frac{1}{2\pi}\int |\langle y\rangle^s V(y)|dy\|f\|_B,$$

which proves the continuity of $\langle x\rangle^s(T_k'^s f)(x)$ in x and thus that of $(T_k'^s f)(x)$ itself. Therefore $T_k'^s \in \mathcal{L}(B)$. It remains to show that $T_k'^s$ is norm continuous with respect to k:

$$\|(T_k'^s - T_{k'}'^s)f\|_B = \sup_{x\in\mathbb{R}^3}\left|\frac{1}{2\pi}\int \frac{\langle y\rangle^s V(y)f(y)}{\langle x\rangle^s}\left(\frac{k_l}{|k|}e^{-i|k||x-y|}-\frac{k_l'}{|k'|}e^{-i|k'||x-y|}\right)\right|$$

$$\leq \sup_{x\in\mathbb{R}^3}\left|\frac{1}{2\pi}\int \frac{\langle y\rangle^s V(y)f(y)}{\langle x\rangle^s}e^{-i|k||x-y|}\left(\frac{k_l}{|k|}-\frac{k_l'}{|k'|}\right)\right|$$

$$+\sup_{x\in\mathbb{R}^3}\left|\frac{1}{2\pi}\int \frac{\langle y\rangle^s V(y)f(y)}{\langle x\rangle^s}\frac{k_l'}{|k'|}(e^{-i|k||x-y|}-e^{-i|k'||x-y|})\right|$$

$$\leq c\left|\frac{k_l}{|k|}-\frac{k_l'}{|k'|}\right|\|f\|_B+\sup_{x\in\mathbb{R}^3}\frac{1}{2\pi}\int \frac{|\langle y\rangle^s V(y)|}{\langle x\rangle^s}|e^{-i|k||x-y|}-e^{-i|k'||x-y|}|dy\|f\|_B.$$

Since we can achieve $c|k_l/|k|-k_l'/|k'||<\epsilon/2$ for any $\epsilon>0$ by choosing $|k-k'|$ small, it remains to show that also

$$\sup_{x\in\mathbb{R}^3}\frac{1}{2\pi}\int \frac{|\langle y\rangle^s V(y)|}{\langle x\rangle^s}|e^{-i|k||x-y|}-e^{-i|k'||x-y|}|dy<\frac{\epsilon}{2},$$

for $|k-k'|$ small. Since $\langle x\rangle^s V(x)\in L^1$ and $|e^{-i|k||x-y|}-e^{-i|k'||x-y|}|\leq 2$, there exists R_1 such that

$$\sup_{|x|>R_1}\frac{1}{2\pi}\int \frac{|\langle y\rangle^s V(y)|}{\langle x\rangle^s}|e^{-i|k||x-y|}-e^{-i|k'||x-y|}|dy \leq \sup_{|x|>R_1}\frac{1}{\langle x\rangle^s}\frac{1}{2\pi}\int 2|\langle y\rangle^s V(y)|dy<\frac{\epsilon}{2}.$$

Similarly, there is R_2 such that

$$\sup_{x\in\mathbb{R}^3}\frac{1}{2\pi}\int_{y>R_2}\frac{|\langle y\rangle^s V(y)|}{\langle x\rangle^s}|e^{-i|k||x-y|}-e^{-i|k'||x-y|}|dy \leq \sup_{x\in\mathbb{R}^3}\frac{1}{\langle x\rangle^s}\frac{1}{2\pi}\int_{y>R_2}2|\langle y\rangle^s V(y)|dy<\frac{\epsilon}{4},$$

holds. Observing that from $|x|<R_1$ and $|y|<R_2|x-y|<R_1+R_2$ follows, we obtain for the remaining part,

$$\sup_{|x|<R_1}\frac{1}{2\pi}\int_{|y|<R_2}\frac{|\langle y\rangle^s V(y)|}{\langle x\rangle^s}|e^{-i|k||x-y|}-e^{-i|k'||x-y|}|dy$$

$$\leq ||k|-|k'||\frac{1}{2\pi}(R_1+R_2)\int |\langle y\rangle^s V(y)|dy \leq C|k-k'|<\frac{\epsilon}{4},$$

for $|k-k'|$ sufficiently small. Combining these results, we get that, for any $\epsilon>0$,

$$\|(T_k'^s - T_{k'}'^s)f\|_B<\epsilon\|f\|_B,$$

for $|k-k'|$ small enough, which proves the norm continuity of $T_k'^s$. Therefore $T_k'^s$ meets the requirements of Lemma A.1 (ii), and we conclude that $\tilde{T}' \in \mathcal{L}(\tilde{B})$.

Finally, consider $\tilde{K}$. For $f \in \tilde{B}$ the continuity of $(\tilde{K}f)(x,k) = \int_{k_l^0}^{k_l} f(x,(k_l',k_{\bar{l}}))dk_l'$ in x and k is clear. Furthermore,

$$\lim_{|x|\to\infty} \sup_{k\in I} |(\tilde{K}f)(x,k)| \leq \lim_{|x|\to\infty} 2\delta_l \sup_{k\in I} |f(x,k)| = 0,$$

so that $\tilde{K} \in \mathcal{L}(\tilde{B})$ and

$$\|\tilde{K}f\|_{\tilde{B}} = \sup_{x\in\mathbb{R}^3, k\in I} \left| \int_{k_l^0}^{k_l} f(x,(k_l',k_{\bar{l}}))dk_l' \right| \leq 2\delta_l \|f\|_{\tilde{B}}.$$

Proof [of part (ii)]: Since $\eta(x,k) \in B$ for all $k \neq 0$ and I is compact, $\langle x\rangle^{-s}\eta(x,(k_l^0,k_{\bar{l}}))$ $\in \tilde{B}$ is obvious. Observing

$$\partial_{k_l} g(x,k) = \frac{i}{2\pi}\frac{k_l}{|k|}\int e^{-i|k||x-y|}V(y)e^{ik\cdot y}\,dy - \frac{i}{2\pi}\int \frac{e^{-i|k||x-y|}}{|x-y|}V(y)y_l e^{ik\cdot y}\,dy,$$

$\langle x\rangle^{-s}\partial_{k_l} g(x,k) \in \tilde{B}$ can be shown using the same types of estimates as in the proof of part (i) of this lemma. □

[1] J. D. Dollard, ''Scattering into cones I, Potential scattering,'' Commun. Math. Phys. **12**, 193–203 (1969).
[2] See, e.g., p. 278 in W. O. Amrein, J. M. Jauch, and K. B. Sinha, ''*Scattering theory in quantum mechanics*,'' (Benjamin, Reading, MA, 1977).
[3] J.-M. Combes, R. G. Newton, and R. Shtokhamer, ''Scattering into cones and flux across surfaces,'' Phys. Rev. D **11**, 366–372 (1975).
[4] M. Daumer, D. Dürr, S. Goldstein, and N. Zanghì, ''On the flux-across-surfaces theorem,'' Lett. Math. Phys. **38**, 103–116 (1996).
[5] M. Daumer, ''Streutheorie aus der Sicht Bohmscher Mechanik,'' Ph.D. thesis, Ludwig-Maximilians-Universität, München, 1995.
[6] M. Daumer, D. Dürr, S. Goldstein, and N. Zanghì, ''On the quantum probability flux through surfaces,'' J. Stat. Phys. **88**, 967–977 (1997).
[7] W. O. Amrein, and J. L. Zuleta, ''Flux and scattering into cones in potential scattering,'' Helv. Phys. Acta **70**, 1 (1997).
[8] W. O. Amrein and D. B. Pearson, ''Flux and scattering into cones for long range and singular potentials,'' J. Phys. A **30**, 5361–5379 (1997).
[9] J.-L. Journé, A. Soffer, and C. D. Sogge, ''Decay estimates for Schrödinger operators,'' Commun. Pure Appl. Math. **44**, 573–604 (1991).
[10] K. Yajima, ''The $W^{k,p}$-continuity of wave operators for Schrödinger operators,'' J. Math. Soc. Jpn. **47**, 551–581 (1995).
[11] Throughout this paper ''$O(|x|^\alpha)$'' will always mean ''$O(|x|^\alpha)$ as $|x|\to\infty$.''
[12] $V:D\to\mathbb{R}$ is locally Hölder continuous if for every $x\in D$ there is an open neighborhood $U_x \subset D$ and an $\alpha_x > 0$, $c_x > 0$ such that $|V(x)-V(y)| \leq c_x|x-y|^{\alpha_x}$ for all $y\in U_x$.
[13] T. Ikebe, ''Eigenfunction expansion associated with the Schrödinger operators and their applications to scattering theory,'' Arch. Ration. Mech. Anal. **5**, 1–34 (1960).
[14] A. Jensen and T. Kato, ''Spectral properties of Schrödinger operators and time-decay of the wave function,'' Duke Math. J. **46**, 583–611 (1979).
[15] l.i.m. $\int$ is a short way of writing $s-\lim_{R\to\infty}\int_{B_R}$ and $s-\lim$ denotes the strong limit.
[16] S. Agmon, ''Spectral properties of Schrödinger operators and scattering theory,'' Ann. Scuola Norm. Sup. Pisa Ser. IV **2**, 151–217 (1975).
[17] We use the usual multi-index notation: $\alpha=(\alpha_1,\alpha_2,\alpha_3)$, $\alpha_i\in\mathbb{N}_0$, $\partial_k^\alpha f(k) := \partial_{k_1}^{\alpha_1}\partial_{k_2}^{\alpha_2}\partial_{k_3}^{\alpha_3} f(k)$ and $|\alpha| := \alpha_1+\alpha_2+\alpha_3$.
[18] A. J. Povzner, ''The expansion of arbitrary functions in terms of eigenfunctions of the operator $-\Delta u = cu$,'' Math. Sbo. **32**, 109–156 (1953); Am. Math. Soc. Trans. **60**, 1–49 (1967).
[19] See, e.g., K. Yoshida, *Functional Analysis* (Springer-Verlag, New York, 1974).
[20] See, e.g., Theorem XI.41 (d) in M. Reed, and B. Simon, *Methods of Modern Mathematical Physics III* (Academic, London, 1979).
[21] C. Zemach and A. Klein, ''The Born expansion in non-relativistic quantum theory,'' Nuovo Cimento **X**, 1078–1087 (1958).
[22] H. Isozaki, ''Differentiability of generalized Fourier transforms associated with Schrödinger operators,'' J. Math. Kyoto Univ. **25**, 789–806 (1985).
[23] See, e.g., Theorems XI.47 and XI.48 in Ref. 20.

On conserved quantities at spatial infinity

Shyan-Ming Perng[a]
Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637

(Received 31 July 1998; accepted for publication 20 November 1998)

There is a well-known short list of asymptotic conserved quantities for a physical system at spatial infinity. We search for new ones. This is carried out within the asymptotic framework of Ashtekar and Romano, in which spatial infinity is represented as a smooth boundary of space–time. We first introduce, for physical fields on space–time, a characterization of their asymptotic behavior as certain fields on this boundary. Conserved quantities at spatial infinity, in turn, are constructed from these fields. We find, in Minkowski space–time, that each of a Klein–Gordon field, a Maxwell field, and a linearized gravitational field yields an entire hierarchy of conserved quantities. Only certain quantities in this hierarchy survive into curved space–time. © *1999 American Institute of Physics.* [S0022-2488(99)00504-6]

I. INTRODUCTION

In the description of isolated systems in flat space–time, conserved quantities have often been found to be useful. Examples of such conserved quantities include electric charge, energy–momentum, angular momentum, and, in certain circumstances, various multipole moments. These conserved quantities are usually expressed as surface integrals in the limit as the surface approaches infinity. In general relativity, by contrast, the construction of such conserved quantities is more complicated. Not the least of these complications is that ''infinity'' is so much more difficult to pin down in the presence of curvature.

The study of isolated systems in general relativity was pioneered by Arnowitt, Deser, and Misner.[1] They defined asymptotic flatness of a space–time in terms of the existence of an initial data set, which, expressed in suitable coordinates, has the initial data approach the flat values at suitable rates. Conserved quantities, such as energy–momentum and angular momentum, were then expressed as limits of certain surface integrals.

One unfortunate aspect of the approach of Arnowitt, Deser, and Misner is that their asymptotic conditions are tied so closely to coordinates. Their approach was subsequently geometrized and extended by Geroch[2] via a conformal completion by a single point ''at spatial infinity.'' Multipole moments for certain fields in flat space–time were generalized to static asymptotically flat space–times within this framework.[3] An alternative geometrical framework, which unifies spatial and null infinity and is thus adapted to the relation between these two asymptotic regimes, was introduced by Ashtekar and Hansen.[4] This framework involves a conformal completion of the entire space–time, null infinity becoming a null cone with spatial infinity its vertex. This framework is used, for example, both to formulate and to prove the assertion that the ADM mass is the past limit of the future Bondi mass.[5]

In both of the geometrical frameworks outlined above spatial infinity is squeezed into a point, and there smoothness of the completed manifold fails. So, inevitably, one is forced to deal with complicated differentiable structures there. This circumstance is less satisfactory than that of null infinity, which is formulated as a smooth boundary of space–time. Early attempts to restore smoothness to spatial infinity include those of Sommers[6] and Persides.[7] Beig and Schmidt,[8,9] using a coordinate-dependent treatment similar to that of Bondi *et al.*,[10] obtained fields on the surface at spatial infinity order by order, and noticed that these fields there satisfy hyperbolic equations. This

[a]Electronic mail: sperng@rainbow.uchicago.edu

work culminates in that of Ashtekar and Romano,[11] who introduced a new geometrical framework for asymptotic flatness in which spatial infinity was indeed expressed as a smooth boundary of space–time. Their definition also provided a natural geometrical setting for the results of Beig and Schmidt. The Ashtekar and Romano framework is somewhat of a hybrid, in that it involves both the conformal and projective structure. By their definition, a space–time is asymptotically flat at spatial infinity, provided one can attach to it a smooth boundary $\mathcal{H}$ and introduce a smooth function Ω vanishing at $\mathcal{H}$ such that the induced metric on and the normal to the constant-Ω surfaces are, after rescaling by suitable powers of Ω, smoothly extendible to $\mathcal{H}$. This new definition has proven to be useful in the study of asymptotic properties of space–time at spatial infinity, since various physical fields turn out to be smooth there.

We return now to conserved quantities. It is natural to ask the following: Do, in some sense, the well-known conserved quantities—energy–momentum, angular momentum, electric charge—at spatial infinity exhaust all conserved quantities that could possibly be defined there? To settle this question would clearly provide insights into the asymptotic properties of the physical fields and of the space–time. The framework introduced by Ashtekar and Romano is perfectly suited to addressing this question. One has a simple, universal smooth structure at spatial infinity, enabling one to investigate fields at spatial infinity order by order. The notion of a conserved quantity had already been formulated by Ashtekar and Romano: Each conserved quantity is to be expressed as an integral over a two-sphere section of spatial infinity, where the value of the integral is independent of section. In particular, the well-known conserved quantities are so expressed. We seek others.

This paper is organized as follows. Section II contains the basic framework, which underlies the rest of the paper. We first review briefly (a slight modification of) the Ashtekar–Romano definition of asymptotic flatness. We then formulate within this framework the asymptotic structure of the physical fields. In particular, we introduce the notion of a conserved quantity, and give some familiar examples. In Sec. III, we consider the special but important case of fields in Minkowski space–time. We construct all linear conserved quantities associated with a Klein–Gordon field, with a Maxwell field, and with a linearized gravitational field and having a certain ''polynomial dependence'' on asymptotic translations. We then study the symmetry properties and the ''gauge behavior'' (the dependence on a certain freedom in the formulation of an asymptotic structure) of these quantities. In Sec. IV, we consider fields in a curved, asymptotically flat space–time. We first derive the equations, at spatial infinity, satisfied by the asymptotic fields. We then show that—at least in the Klein–Gordon and Maxwell cases—certain of the conserved quantities found in Sec. III for Minkowski space–time can be generalized to these curved space–times. In Sec. V, we discuss various related issues. In particular, we formulate two conjectures. One asserts that a certain conserved quantity for linearized gravity in Minkowski space–time can be generalized to curved space–time. The other asserts that we have here found all conserved quantities in curved space–time for Klein–Gordon, Maxwell, and gravitational fields.

II. PRELIMINARIES

A. Asymptotic flatness

Fix a space–time $(\tilde{M},\tilde{g}_{ab})$.

Definition 1: By a completion of $(\tilde{M},\tilde{g}_{ab})$, *we mean (cf. Ashtekar and Romano in Ref. 11): A manifold M with boundary* $\mathcal{H}$, *a smooth function* Ω *defined on M vanishing on* $\mathcal{H}$, *and a diffeomorphism from* $\tilde{M}$ *to* $M-\mathcal{H}$ *(by means of which we identify* $\tilde{M}$ *with its image in M) satisfying the following three classes of conditions.*

(1) The combinations (i) $\nabla_a\Omega$, *(ii)* $\Omega^{-4}\tilde{g}^{ab}\nabla_b\Omega(\equiv n^a)$, *and (iii)* $\Omega^2[\tilde{g}_{ab}-(\tilde{g}^{cd}\nabla_c\Omega\,\nabla_d\Omega)^{-1}\nabla_a\Omega\,\nabla_b\Omega](\equiv q_{ab})$ *admit smooth, nowhere-vanishing extensions to* $\mathcal{H}$ *such that (iv)* $n^a\nabla_a\Omega(\equiv\lambda^{-2})|_{\mathcal{H}}=1$ *and (v)* $\mathcal{L}_n[(n^m\nabla_m\Omega)^{-1}q_{ab}]|_{\mathcal{H}}=0$.

(2) $(\mathcal{H},q_{ab}|_{\mathcal{H}})$ *is a standard timelike hyperboloid, i.e.,* $\mathcal{H}$ *has topology* $S^2\times\mathbb{R}$, $q_{ab}|_{\mathcal{H}}\equiv\overset{0}{q}_{ab}$ *is of constant positive curvature and is geodesically complete.*

(3) The combinations (i) $n^k n^l \tilde{G}_{kl}$, *(ii)* $\Omega^{-1} q_a{}^k n^l \tilde{G}_{kl}$, *and (iii)* $\Omega^{-2} q_a{}^k q_b^l \tilde{G}_{kl}$, *are smoothly extendible to* $\mathcal{H}$, *where* $\tilde{G}_{ab}$ *is the Einstein tensor of* $\tilde{g}_{ab}$.

The boundary $\mathcal{H}$ represents spatial infinity. Conditions (1) describe the falloff behavior of the metric $\tilde{g}_{ab}$ and conditions (3) that of its second derivative. Conditions (2) ensure, among other things, that we are dealing with (all of) spatial infinity. There is some redundancy in the above conditions. Specifically, the constancy both of the left side of (1) (iv) and of the curvature of q_{ab} already follow from the other conditions. In light of this, the choice of the constant ''1'' in condition (i) (iv) (which is equivalent to the demand that $\overset{0}{q}_{ab}$ be the metric of a unit hyperboloid) serves only to restrict the freedom of multiplying Ω by a constant factor. Condition (1)(v) is essentially the condition, $\overset{0}{B}_{ab}=0$ [cf. Eq. (3)], introduced by Ashtekar and Hansen[4] in order to define angular momentum. More precisely, when $\overset{0}{B}_{ab}=0$, condition (1)(v) can always be achieved without affecting the other conditions by choosing a suitable Ω.

Definition 1 is essentially the same as the definition, given by Ashtekar and Romano,[11] of what they call an asymptotically Minkowskian space–time. However, there are three differences. First, our conditions on the Einstein tensor are weaker than the corresponding condition, namely, $\lim_{\Omega\to 0} \Omega^{-1}\tilde{G}_{ab}=0$, in their definition. Their condition, expressed in the present language, is equivalent to the smooth extendibility to $\mathcal{H}$ of $\Omega^{-2} n^k n^l \tilde{G}_{kl}$, $\Omega^{-2} q_a{}^k n^l \tilde{G}_{kl}$, and $\Omega^{-2} q_a{}^k q_b^l \tilde{G}_{kl}$. Indeed, our condition holds while theirs fails (for $n^k n^l \tilde{G}_{kl}|_{\mathcal{H}} \neq 0$) in the Reissner–Nordstrom solution. Second, we impose condition (1)(iv), which, as mentioned above, is effectively a gauge restriction on the conformal factor Ω, a restriction that is absent in the definition of Ashtekar and Romano. Finally, we impose condition (1)(v), which Ashtekar and Romano omit from the general definition of asymptotic Minkowskian space–times, but subsequently impose for their discussion of angular momentum.

We give a few simple examples of completion. As a first example, let $(\tilde{M},\tilde{g}_{ab})$ be Minkowski space–time, and let (t,r,θ,ϕ) be ordinary spherical polar coordinates. Set $\Omega=(r^2-t^2)^{-1/2}$ and $\tanh \chi = t/r$. Let M be $\tilde{M}$ together with the boundary $\mathcal{H}$ consisting of the points labeled by $\Omega=0$ in the (hyperbolic coordinate) chart $(\Omega,\chi,\theta,\phi)$, with a differentiable structure given by that chart. Then this (M,Ω) is a completion of Minkowski space–time. As a second example, let $(\tilde{M},\tilde{g}_{ab})$ be Reissner–Nordstrom solution, and let (t,r,θ,ϕ) be the usual Schwarzschild-like coordinates therein. Repeat the same construction as in Minkowski space–time to obtain a manifold with boundary (M,Ω). Then this choice of Ω satisfies all conditions in definition 1, except condition (1)(v). Condition (1)(v), in turn, can be achieved, without violating other conditions, by choosing a new conformal factor Ω' of the form $\Omega'=\Omega(1+\omega\Omega)$ with a suitable smooth function ω. In general, all stationary vacuum space–times asymptotically flat by the usual definition[3] admit completions in the present sense.[12]

Two completions, (M,Ω), (M',Ω') of $(\tilde{M},\tilde{g}_{ab})$, are said to be equivalent if the identity map of $\tilde{M}$ extends to a diffeomorphism from M to M'. It turns out that a space–time may admit inequivalent completions.[13] Minkowski space–time, for instance, has a four-parameter family of inequivalent completions related to each other by what are called ''logarithmic translations.''[14] Indeed, let x^μ be a usual Minkowskian coordinate system in Minkowski space–time $\tilde{M}$, and c^μ any constant vector. Then the hyperbolic coordinates associated with x'^μ given by $x^\mu = x'^\mu - c^\mu \log \Omega'$ yield a new completion of $\tilde{M}$ inequivalent to that arising from x^μ. In this case we can single out the usual completion to be the preferred one among this four-parameter family since it is the only one in which all Killing fields are smoothly extendible to the boundary at spatial infinity. Similarly, any stationary asymptotically flat space–time admits at least a one-parameter family of inequivalent completions, arising from logarithmic time translations. There is also a sort of converse to this: the existence of two inequivalent completions related by such a logarithmic translation implies that the space–time admits an asymptotic translational Killing field—a vector field $\tilde{\xi}^a$ with the properties that $\Omega^{-1}\tilde{\xi}^a$ is smoothly extendible to, and vanishes nowhere on, $\mathcal{H}$; and that $\tilde{\nabla}_{(a}\tilde{\xi}_{b)}$ and all its derivatives vanishes on $\mathcal{H}$. In the spatial-infinity framework of Geroch and Ashtekar–Hansen, it has been shown by Chrusciel[15] that these logarithmic translations are the only kind of inequivalent completions that may arise. We conjecture the following in the present framework: Any two inequivalent completions are related by such a logarithmic translation. If this conjecture is true, then our work will not be affected by the possible existence of inequivalent completions. In what follows, we will always fix a specific completion and only consider completions smoothly related (i.e., equivalent) to the fixed one.

B. Physical fields and their remnants

We now set up the framework for dealing with the asymptotic structure of physical fields. Let $(\tilde{M},\tilde{g}_{ab})$ be a space–time, with (M,Ω) a completion. Let $\tilde{v}_{a_1\cdots a_m}$ be a smooth, covariant, mth rank tensor field on $\tilde{M}$, and consider the 2^m tensor fields that result from contracting each index of $\tilde{v}$ with either $\Omega^2 n^a$ or $\Omega q^a{}_b$. We say $\tilde{v}$ is *asymptotically regular of order s*, provided each of these 2^m tensor fields, multiplied by Ω^{-s}, is smoothly extendible to $\mathcal{H}$. Asymptotic regularity of a general tensor field is defined by lowering any contravariant indices with $\tilde{g}_{ab}$ and applying the definition above to the resulting covariant field. Note that conditions (1)(*ii*), (*iii*) above are precisely the statement that $\tilde{g}_{ab}$ is asymptotically regular of order 0; and conditions (3)(*i*)–(*iii*) are precisely the statement that $\tilde{G}_{ab}$ is asymptotically regular of order 4. The outer product of two asymptotically regular fields, of respective orders s and s', is asymptotically regular, of order $s+s'$. Contractions using $\tilde{g}^{ab}$ preserve asymptotic regularity, and order.

Thus, an asymptotically regular physical field gives rise, on M, to 2^m smooth fields, with ranks ranging from m down to zero, whose behavior near $\mathcal{H}$ reflects the asymptotic behavior of the physical field. Let $u_{a_1\cdots a_m}$ denote any one of these fields. Then set, for k any non-negative integer,

$$\overset{k}{u}_{a_1\cdots a_m} \equiv \underset{\leftarrow}{\psi}[(\mathcal{L}_{(n\cdot\nabla\Omega)^{-1}n})^k u_{a_1\cdots a_m}], \tag{1}$$

where $\underset{\leftarrow}{\psi}$ stands for the pull-back to $\mathcal{H}$ via the natural embedding map $\mathcal{H}\overset{\psi}{\rightarrow}M$. Note the right side of Eq. (1) exists since $u_{a_1\cdots a_m}$ (and therefore each of its derivatives) is smoothly extendible to $\mathcal{H}$. The $\overset{k}{u}_{a_1\ldots a_m}$ so defined will be called the kth-*order remnant* of $u_{a_1\cdots a_m}$. These remnants, ($k=0,1,\ldots$), clearly carry, order by order, the asymptotic information contained in $u_{a_1\cdots a_m}$, and, therefore, the asymptotic information in the original physical field $\tilde{v}$. Suppose, next, that the physical field $\tilde{v}$ satisfies various field equations. Then these field equations yield partial differential equations on M on the u's that arise via asymptotic regularity from $\tilde{v}$, and so partial differential equations on $\mathcal{H}$ on the remnants $\overset{k}{u}$ that arise via Eq. (1) from the u's. We will refer to these as the *remnant field equations*.

We give some examples of asymptotically regular fields and their associated remnants. Fix a space–time $(\tilde{M},\tilde{g}_{ab})$ with a completion (M,Ω). For the first example, consider the space–time metric $\tilde{g}_{ab}$. Then, as we mentioned above, this field is asymptotically regular of order 0. The corresponding u's are $q_{ab}(\equiv\Omega^2 q_a{}^k q_b{}^l \tilde{g}_{kl})$, $0(\equiv\Omega^3 q_a{}^k n^l \tilde{g}_{kl})$, and $\lambda^{-2}(\equiv[\Omega^4 n^a n^b \tilde{g}_{ab}])$. Their corresponding remnants, $\overset{k}{q}_{ab}$ and $\overset{k}{\lambda}$, carry the asymptotic information contained in the space–time geometry. We note that conditions $(1)(iv)$ and $(1)(v)$ in the definition of a completion are actually conditions on these remnants: namely, $\overset{0}{\lambda}=1$, and $\overset{1}{q}_{ab}=-2\overset{1}{\lambda}\overset{0}{q}_{ab}$, respectively. For the second example, consider the Einstein tensor $\tilde{G}_{ab}$. Then, as we mentioned above, this field is asymptotically regular of order 4. The corresponding u's, written in terms of the stress–energy tensor $\tilde{T}_{ab}$ $(=\tilde{G}_{ab}/\kappa$, with $\kappa=8\pi G/c^4)$ are

$$T\equiv\lambda^2 n^a n^b(\kappa\tilde{T}_{ab}),\quad T_a\equiv\lambda\Omega^{-1}q_a{}^k n^l(\kappa\tilde{T}_{kl}),\quad \text{and}\quad T_{ab}\equiv\Omega^{-2}q_a{}^k q_b{}^l\kappa(\tilde{T}_{kl}-\tfrac{1}{2}\tilde{T}\tilde{g}_{kl}),\tag{2}$$

where we have introduced certain powers of λ and have used the trace-reversed version of $\tilde{T}_{ab}$ in defining T_{ab} for later convenience. We denote by $\overset{k}{T},\overset{k}{T}_a,\overset{k}{T}_{ab}$, the remnants of T, T_a, and T_{ab}, respectively. For the third example, consider the Weyl tensor, $\tilde{C}_{abcd}$, of this space–time. It is shown in Appendix C [cf. the discussion around Eq. (C14)] that this field is asymptotically regular of order 3. The u's in this case may be taken to be

$$E_{ab}\equiv\Omega^3\lambda^2 q_a{}^j q_b{}^l n^k n^m\tilde{C}_{jklm}\quad\text{and}\quad B_{ab}\equiv\Omega^3\lambda^2 q_a{}^j q_b{}^l n^k n^m{}^*\tilde{C}_{jklm}.\tag{3}$$

Denote their remnants $\overset{k}{E}_{ab}$ and $\overset{k}{B}_{ab}$. Note that condition $(1)(v)$ in the definition of a completion is actually a condition on one of these remnants, namely, $\overset{0}{B}_{ab}=0$ [cf. Eq. (C17)]. For the final example, consider a Maxwell field $\tilde{F}_{ab}$ in this space–time. We demand that it be asymptotically regular of order 2,[16] i.e., that each of

$$E_a\equiv\Omega\lambda n^b\tilde{F}_{ab}\quad\text{and}\quad B_a\equiv\Omega\lambda n^b{}^*\tilde{F}_{ab}\tag{4}$$

be smoothly extendible to $\mathcal{H}$. These are effectively the u's. This demand reflects the idea that a physically reasonable Maxwell field must fall off like $1/r^2$ near spatial infinity. We denote by $\overset{k}{E}_a,\overset{k}{B}_a$ the remnants of E_a,B_a, respectively. Note that it follows that the stress–energy tensor of this Maxwell field has the fall-off rate consistent with that of Eq. (2). Indeed, from $\tilde{T}_{ab}=\tfrac{1}{2}(\tilde{F}_{am}\tilde{F}_b{}^m-\tfrac{1}{4}\tilde{F}^2\tilde{g}_{ab})$ we have

$$T=\tfrac{1}{2}\kappa(E^2+B^2),\quad T_a=-\kappa\epsilon_{amn}E^m B^n,\quad T_{ab}=\kappa[E_aE_b+B_aB_b-\tfrac{1}{2}(E^2+B^2)q_{ab}].\tag{5}$$

There remains, as it turns out, some gauge freedom in the present framework. Fix a space–time $(\tilde{M},\tilde{g}_{ab})$, and let (M,Ω) and (M,Ω'), be two completions of $(\tilde{M},\tilde{g}_{ab})$. It then follows that $\Omega'=\Omega(1+\omega\Omega)$, for some smooth function ω on M such that $\overset{0}{\omega}\equiv\omega|_{\mathcal{H}}$ satisfies Eq. (7) below (i.e., $\overset{0}{\omega}$ is an ''asymptotic translation''); and, conversely, for (M,Ω) any completion and ω and Ω' as above, then (M,Ω') is also a completion. Thus, the gauge freedom consists precisely of such ω fields. The asymptotic gauge freedom, then, is described by the remnants, $\overset{k}{\omega}$, of ω. It turns out[8] that one can, utilizing this gauge freedom, always achieve

$$\overset{k}{\lambda}=0,\quad k\geqslant 2,\tag{6}$$

and that this exhausts the gauge freedom associated with the remnants $\overset{k}{\omega}$, for $k\geqslant 1$. Thus, making this gauge choice, the remaining gauge freedom is represented by a single $\overset{0}{\omega}$ satisfying Eq. (7).

C. Asymptotic translations

In order to construct conserved quantities, it will be convenient to have on hand some facts about asymptotic translations. Denote by $\mathcal{T}$ the set of functions v on $\mathcal{H}$ satisfying the differential equation,

$$D_a D_b v + v \overset{0}{q}_{ab} = 0, \tag{7}$$

where D_a denotes the derivative operator of $\overset{0}{q}_{ab}$. This $\mathcal{T}$ is a four-dimensional vector space [since, by virtue of the fact that the curl of Eq. (7) is an identity, v is completely determined by its value and derivative at any one point] equipped with a Lorentz metric $\langle v,w\rangle \equiv \overset{0}{q}{}^{ab} D_a v D_b w + vw$ [since, by virtue of Eq. (7), the right side is a constant]. Elements of $\mathcal{T}$ can be interpreted[11] as asymptotic translations on M in the following sense: For $\tilde{\xi}^a$ a vector field on $\tilde{M}$ asymptotically regular of order 0 such that $\mathcal{L}_{\tilde{\xi}} \tilde{g}_{ab}$ is asymptotically regular of order 2, then $\Omega^{-2} \mathcal{L}_{\tilde{\xi}} \Omega|_{\mathcal{H}} \in \mathcal{T}$.

It is convenient to introduce an index notation for tensors over $\mathcal{T}$: Greek superscripts and subscripts denote, respectively, elements of $\mathcal{T}$ and its dual $\mathcal{T}^*$. Thus, a solution v of Eq. (7) might be denoted v^μ, while a linear map $\mathcal{T} \overset{w}{\mapsto} \mathbb{R}$ might be denoted w_μ. The action of w on v would be expressed by contraction: $w(v) = w_\mu v^\mu$. We denote by $\eta_{\mu\nu}$ the above Lorentz metric on $\mathcal{T}$, i.e., we set $\eta_{\mu\nu} v^\mu w^\nu = \langle v,w \rangle$. We shall use $\eta_{\mu\nu}$ (and its inverse) to lower and raise indices of tensors over $\mathcal{T}$. The objects with which we shall be concerned are fields on $\mathcal{H}$ that may have Latin indices (indicating tensor character over the manifold $\mathcal{H}$) and Greek indices (indicating tensor character over the vector space $\mathcal{T}$). Thus, for example, ζ_α would denote a $\mathcal{T}^*$-valued function on $\mathcal{H}$, ζ^a would denote an ordinary tangent vector field on $\mathcal{H}$, and $\zeta^a{}_\alpha$ would denote a $\mathcal{T}^*$-valued tangent vector field on $\mathcal{H}$. In particular, an element v^μ in $\mathcal{T}$ is now viewed as a $\mathcal{T}$-valued constant function on $\mathcal{H}$. We lower and raise Greek indices of such fields with $\eta_{\alpha\beta}$ and its inverse, and lower and raise Latin indices with $\overset{0}{q}_{ab}$ and its inverse. There is a natural field, α_μ, defined by the property that, for any $v^\mu \in \mathcal{T}$, $\alpha_\mu v^\mu$ is the corresponding solution of Eq. (7).[17] Then, e.g., $\alpha_\mu \alpha^\mu = 1$. The derivative operator D_a on $\mathcal{H}$ associated with $\overset{0}{q}_{ab}$ extends to a derivative operator on our indexed fields by demanding that $D_a v^\alpha = 0$, for v^α any constant field. There now follows $D_a \overset{0}{q}_{bc} = 0$, $D_a \eta_{\alpha\beta} = 0$,

$$D_a D_b \alpha_\mu + \alpha_\mu \overset{0}{q}_{ab} = 0 \tag{8}$$

[from Eq. (7)],

$$D_a \alpha_\mu D^a \alpha_\nu + \alpha_\mu \alpha_\nu = \eta_{\mu\nu} \tag{9}$$

(from the definition of $\eta_{\mu\nu}$), and $\eta^{\mu\nu} D_a \alpha_\mu D_b \alpha_\nu = \overset{0}{q}_{ab}$.[18]

D. Conserved quantities

Now imagine that we were somehow able to find a divergence-free vector field, constructed from (the remnants of) some physical fields and the background geometry of $\mathcal{H}$. Integrating (the dual of) this vector field over a two-sphere cut (i.e., a noncontractible two-sphere submanifold) of $\mathcal{H}$, we obtain a number—one clearly independent of the choice of cut. Think of such an integral as being the limit of an integral over a spacelike 2-sphere in space–time, as the 2-sphere approaches the cut at spatial infinity. These integrals we call conserved quantities. In each of the examples we shall consider, the divergence-free vector field is multilinear in α_μ,[19] and so the conserved quantities may be viewed as a tensor over $\mathcal{T}$.

We now give three well-known[1,2,4,11] examples of conserved quantities. Some of the computations are relegated to Sec. IV and Appendix C. Fix a space–time $(\tilde{M},\tilde{g}_{ab})$, a completion (M,Ω) thereof, and a cut C of $\mathcal{H}$.

For the first example, let $\tilde{F}_{ab}$ be a Maxwell field on $\tilde{M}$, regular of order 2. Consider the right side of

$$Q=\frac{1}{4\pi}\int_C \overset{0}{E}{}^a\, dS_a, \tag{10}$$

where $\overset{0}{E}{}^a$ is the (zeroth-order) remnant of E_a given by Eq. (4). Maxwell's equations imply the integrand above is divergence-free [cf. Eq. (51)]. Thus, Eq. (10) defines a conserved quantity. This Q is precisely the electric charge, for the right side of Eq. (10) is the limit of the integral of $^*F_{ab}$ over a large spacelike 2-sphere in the space–time as that 2-sphere approaches the cut C. For the second example, consider the right side of

$$\mathcal{P}^\mu=\frac{1}{8\pi}\int_C \overset{0}{E}{}^{ab} D_b \alpha^\mu\, dS_a, \tag{11}$$

where $\overset{0}{E}_{ab}$ is the remnant of E_{ab}, a portion of the Weyl tensor, given in Eq. (3). The remnant field equation [Eq. (C20)] together with Eq. (8) on α_μ, imply that the integrand above is divergence-free. Thus, Eq. (11) defines a conserved quantity, which is a vector over $\mathcal{T}$. This $\mathcal{P}^\mu$ is precisely[1,2] the total mass momentum of the space–time. For the third example, consider the right side of

$$\mathcal{M}_{\mu\nu}=-\frac{1}{16\pi}\epsilon_{\mu\nu}{}^{\tau\sigma}\int_C \overset{1}{B}{}^{ab}\alpha_\tau D_b \alpha_\sigma\, dS_a, \tag{12}$$

where $\epsilon_{\mu\nu\tau\sigma}$ denotes the η-alternating tensor on $\mathcal{T}$. In order for the integrand above to be divergence-free, we must impose on the space–time the additional condition[20] that

$$D_{[a}\overset{0}{T}_{b]}=0. \tag{13}$$

Under this additional condition, Eq. (12) defines a conserved quantity, which is a two-form over $\mathcal{T}$. This $\mathcal{M}_{\mu\nu}$ is precisely[4] the total angular momentum of the space–time.

Finally, we revisit the issue of gauge. Fix a space–time $(\tilde{M},\tilde{g}_{ab})$, and a completion (M,Ω) thereof. Demand further that the completion satisfy the gauge condition (6), so the remaining gauge freedom is represented by the choice of some $\overset{0}{\omega}\in\mathcal{T}$. Applying such a gauge transformation, the remnants of any physical field, and thus also of any conserved quantities associated with that field, will, in general, change. Specifically, let Q_A be any conserved quantity or any remnant field, where the subscript A is an abbreviation for all the indices of Q. Then, for each translation $\overset{0}{\omega}\in\mathcal{T}$, there corresponds a ''gauge-transformed'' quantity—$Q_A[\overset{0}{\omega}]$. Thus, we may regard our quantity Q_A as a tensor field on the 4-manifold $\mathcal{T}$ so defined that its value at $\overset{0}{\omega}\in\mathcal{T}$ is $Q_A[\overset{0}{\omega}]$. In short, the gauge behavior of our original quantity Q_A is coded in the position dependence of this tensor field on $\mathcal{T}$. The derivative of this tensor field reflects the behavior of the quantity under ‘‘infinitesimal gauge transformation.’’ Indeed, from

$$Q_A[\overset{0}{\omega}+\delta\overset{0}{\omega}]=Q_A[\overset{0}{\omega}]+(\delta\overset{0}{\omega})^{\mu}Q^{(1)}{}_{\mu A}[\overset{0}{\omega}]+O((\delta\overset{0}{\omega})^2), \tag{14}$$

we have

$$\nabla_{\mu}Q_A=Q^{(1)}{}_{\mu A}, \tag{15}$$

where ∇_{μ} denote the natural derivative operator on the 4-manifold $\mathcal{T}$. As examples, consider the conserved quantities (10)–(12). Under a gauge transformation, $\Omega'=\Omega(1+\omega\Omega)$ with $\overset{0}{\omega}\in\mathcal{T}$ the remnants $\overset{0}{E}_a,\overset{0}{E}_{ab}$ remain unchanged, while $\overset{1}{B}_{ab}$ changes to $\overset{1}{B'}_{ab}=\overset{1}{B}_{ab}-2\epsilon_{(a}{}^{kl}\overset{0}{E}_{b)k}D_l\overset{0}{\omega}$. In terms of the corresponding tensor fields on the 4-manifold $\mathcal{T}$, these become $\nabla_{\mu}\overset{0}{E}_a=0$, $\nabla_{\mu}\overset{0}{E}_{ab}=0$, and $\nabla_{\mu}\overset{1}{B}_{ab}=-2\epsilon_{(a}{}^{kl}\overset{0}{E}_{b)k}D_l\alpha_{\mu}$. It follows that the total electric charge Q (10) and the 4-momentum $\mathcal{P}^{\mu}$ (11) are gauge invariant, and that[11] the angular momentum $\mathcal{M}_{\mu\nu}$ (12) changes via[21]

$$\mathcal{M}'_{\mu\nu}=\mathcal{M}_{\mu\nu}-\overset{0}{\omega}_{[\mu}\mathcal{P}_{\nu]}. \tag{16}$$

In terms of the corresponding tensor fields on the 4-manifold $\mathcal{T}$, these become, respectively, $\nabla_{\lambda}Q=0$, $\nabla_{\lambda}\mathcal{P}^{\mu}=0$, and

$$\nabla_{\lambda}\mathcal{M}_{\mu\nu}=-\eta_{\lambda[\mu}\mathcal{P}_{\nu]}. \tag{17}$$

III. MINKOWSKI SPACE–TIME

We now apply the framework developed in the previous section to the study of conserved quantities associated with physical fields in Minkowski space–time. Minkowski space–time is a good starting point: It is simple and suggestive of what might happen in the presence of curvature. We shall take as the physical field successively a Klein–Gordon field, a Maxwell field, and a linearized gravitational field. We will write down, for each of these cases, all conserved quantities linear in the physical fields and multilinear in asymptotic translations.

Let $(\tilde{M},\tilde{\eta}_{ab})$ be Minkowski space–time. Fix a point $p\in\tilde{M}$; let Ω be the inverse geodesic distance from p. Then this Ω yields a completion (M,Ω) of Minkowski space–time that we call the *standard completion*. In this completion, we have $\overset{n}{\lambda}=0$ and $\overset{n}{q}_{ab}=0$, for $n\geqslant 1$.

A. Remnant field equations

Here we derive the remnant field equations for Klein–Gordon, Maxwell, and linearized gravitational fields for later use in constructing conserved quantities. For what follows we fix a standard completion of Minkowski space–time $\tilde{M}$ and denote by D_a the derivative operator associated with q_{ab} on constant-Ω surfaces.

Let $\tilde{\phi}$ be a Klein–Gordon field on $\tilde{M}$ asymptotically regular of order 1. Setting $\phi=\Omega^{-1}\tilde{\phi}$, we have

$$0=\tilde{\nabla}^2\tilde{\phi}=\Omega^3[(D^2-1)\phi+\Omega\mathcal{L}_n\phi+\Omega^2(\mathcal{L}_n)^2\phi]. \tag{18}$$

Taking the remnants of the above equation, we obtain

$$D^2 \overset{n}{\phi} = (-n^2+1)\overset{n}{\phi}, \tag{19}$$

for $n=0,1,2,\ldots$.

Let $\tilde{F}_{ab}$ be a Maxwell field on $\tilde{M}$ asymptotically regular of order 2, with remnants $\overset{n}{E}_a$ and $\overset{n}{B}_a$. Using Eq. (4), Maxwell's equation can be written as

$$0 = \tilde{\nabla}^m \tilde{F}_{ma} = \Omega D^m E_m \nabla_a \Omega - \Omega^2 (\Omega \mathcal{L}_n E_a - \epsilon_{akl} D^k B^l), \tag{20}$$

$$0 = \tilde{\nabla}^m {}^*\tilde{F}_{ma} = \Omega D^m B_m \nabla_a \Omega - \Omega^2 (\Omega \mathcal{L}_n B_a + \epsilon_{akl} D^k E^l). \tag{21}$$

Taking the remnants of the above equations, we obtain

$$D_a \overset{0}{E}{}^a = 0, \quad D_a \overset{0}{B}{}^a = 0, \tag{22}$$

$$\epsilon^{abc} D_b \overset{n}{B}_c = n \overset{n}{E}{}^a, \quad -\epsilon^{abc} D_b \overset{n}{E}_c = n \overset{n}{B}{}^a, \tag{23}$$

for $n=0,1,2,\ldots$. Note that Eqs. (23) imply

$$D^2 \overset{n}{E}_a = (-n^2+2)\overset{n}{E}_a, \quad D^2 \overset{n}{B}_a = (-n^2+2)\overset{n}{B}_a, \tag{24}$$

for $n=0,1,2,\ldots$.

Let $\tilde{K}_{abcd}$ be a linearized gravitational field on $\tilde{M}$, i.e., a tensor field on $\tilde{M}$ having the same symmetry and contractions as the Weyl tensor and satisfying the linearized Bianchi identity:

$$\tilde{\nabla}_{[a} \tilde{K}_{bc]de} = 0. \tag{25}$$

Let $\tilde{K}_{abcd}$ be asymptotically regular of order 3, so $E_{ab} \equiv \Omega^3 \tilde{K}_{akbl} n^k n^l$ and $B_{ab} \equiv \Omega^3 {}^*\tilde{K}_{akbl} n^k n^l$ are smoothly extendible to $\mathcal{H}$. Their remnants, denoted $\overset{n}{E}_{ab}$ and $\overset{n}{B}_{ab}$, are symmetric and trace-free. The linearized Bianchi identity can be written as

$$\begin{aligned} 0 &= \tilde{\nabla}^m \tilde{K}^*_{mabc} \\ &= [\Omega^{-1} \nabla_a \Omega \, D^m B_{m[b} - (\Omega \mathcal{L}_n B_{a[b} + \epsilon_{akl} D^k E^l{}_{[b]})] \nabla_{c]} \Omega \\ &\quad + \tfrac{1}{2} [\nabla_a \Omega \, D^k E_{km} - \Omega (\Omega \mathcal{L}_n E_{ma} - \epsilon_{mkl} D^k B^l{}_a)] \epsilon^m{}_{bc}. \end{aligned} \tag{26}$$

Taking the remnants of the above equation, we obtain

$$\epsilon^{lma} D_l \overset{n}{B}_m{}^b = n \overset{n}{E}{}^{ab}, \quad -\epsilon^{lma} D_l \overset{n}{E}_m{}^b = n \overset{n}{B}{}^{ab}, \tag{27}$$

for $n=0,1,2,\ldots$. Note that Eqs. (27) imply

$$D^2 \overset{n}{E}_{ab} = (-n^2+3)\overset{n}{E}_{ab}, \quad D^2 \overset{n}{B}_{ab} = (-n^2+3)\overset{n}{B}_{ab}, \tag{28}$$

for $n=0,1,2,\ldots$.

Recall that the present framework is subject to a class of restricted gauge transformations [namely, replacements of Ω by $\Omega' = \Omega(1+\omega\Omega)$], which preserve the gauge conditions $\overset{n}{\lambda} = 0$, n

$\geqslant 2$, and that each such gauge transformation is completely characterized by an $\overset{0}{\omega}\in\mathcal{T}$. For completeness, we summarize the behavior of the remnants above under such a gauge transformation:

$$\nabla_\mu \overset{n}{\phi} = n(\mathcal{L}_{D\alpha_\mu} \overset{n-1}{\phi} - n\alpha_\mu \overset{n-1}{\phi}), \tag{29}$$

$$\nabla_\mu \overset{n}{E}_a = n(\mathcal{L}_{D\alpha_\mu} \overset{n-1}{E}_a - n\alpha_\mu \overset{n-1}{E}_a + \epsilon_a{}^{kl} \overset{n-1}{B}_k D_l\alpha_\mu), \tag{30}$$

$$\nabla_\mu \overset{n}{B}_a = n(\mathcal{L}_{D\alpha_\mu} \overset{n-1}{B}_a - n\alpha_\mu \overset{n-1}{B}_a - \epsilon_a{}^{kl} \overset{n-1}{E}_k D_l\alpha_\mu), \tag{31}$$

$$\nabla_\mu \overset{n}{E}_{ab} = n(\mathcal{L}_{D\alpha_\mu} \overset{n-1}{E}_{ab} - n\alpha_\mu \overset{n-1}{E}_{ab} + 2\epsilon^{kl}{}_{(a} \overset{n-1}{B}_{b)k} D_l\alpha_\mu), \tag{32}$$

$$\nabla_\mu \overset{n}{B}_{ab} = n(\mathcal{L}_{D\alpha_\mu} \overset{n-1}{B}_{ab} - n\alpha_\mu \overset{n-1}{B}_{ab} - 2\epsilon^{kl}{}_{(a} \overset{n-1}{E}_{b)k} D_l\alpha_\mu), \tag{33}$$

for $n=0,1,2,\ldots$. Note that $\overset{0}{\phi}$, $\overset{0}{E}_a$, $\overset{0}{B}_a$, $\overset{0}{E}_{ab}$, and $\overset{0}{B}_{ab}$ are gauge invariant. As a consistency check, we note also that the ∇_μ curl of the right side of each of the above equations vanishes, by virtue of $\nabla_\mu \alpha_\nu = 0$, as it must. Of course, these gauge-transformed fields satisfy the same equations as the original fields.

B. Remnant radiation multipoles

It is perhaps most natural to seek conserved quantities that are linear in the remnants, since, e.g., this category includes all well-known conserved quantities.[22] In this section we shall find all such conserved quantities for a Klein–Gordon field, a Maxwell field, and a linearized gravitational field in Minkowski space–time $\tilde{M}$. Again, we fix the standard completion of $\tilde{M}$.

We begin with the Klein–Gordon field. Let $\tilde{\phi}$ be a Klein–Gordon field asymptotically regular of order 1, with remnants $\overset{n}{\phi}$.

Theorem 1: *(i) The conserved quantities linear in this Klein–Gordon field consist precisely of the family*

$$\mathcal{K}_{\mu_1\cdots\mu_{n-1}}[\overset{n}{\phi}] \equiv \int_C [\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}}) D^a \overset{n}{\phi} - \overset{n}{\phi} D^a \mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})] dS_a, \quad n \geqslant 1, \tag{34}$$

where $\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$ denotes the symmetric, trace-free part of $\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}}$.[23]

(ii) The $\mathcal{K}_{\mu_1\cdots\mu_{n-1}}$ are totally symmetric and trace-free.

(iii) The behavior of $\mathcal{K}_{\mu_1\cdots\mu_{n-1}}$ under restricted gauge transformations is given by

$$\nabla_\mu \mathcal{K}_{\mu_1\cdots\mu_{n-1}} = \tfrac{1}{2} n(n-2)\, \eta_{(\mu_1\mu_2} \mathcal{K}_{\mu_3\cdots\mu_{n-1})\mu} - n(n-1)\, \eta_{\mu(\mu_1} \mathcal{K}_{\mu_2\cdots\mu_{n-1})}\,. \tag{35}$$

We will refer to these $\mathcal{K}$'s as the remnant radiation multipoles of a Klein–Gordon field.

To see that Eq. (34) indeed defines a conserved quantity, take the divergence of the integrand, and use that both $\overset{n}{\phi}$ and $\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$ satisfy Eq. (19). To prove *(iii)*, use Eq. (29), the definition of $\mathcal{K}$, and a certain identity on $\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$. See Appendix B for details. Note the right side of Eq. (35) is, up to an overall factor, the only $(n-1)$th rank, symmetric, trace-free tensor linear in $\overset{n-2}{\mathcal{K}}$. Equation (35) states that the dependence of the $\mathcal{K}$'s on position in $\mathcal{T}$ is exactly that of ordinary multipole moments. The proof that the family given by Eq. (34) exhausts the linear conserved quantities in the Klein–Gordon case is outlined in Appendix B. As an example of these remnant radiation multipoles, let $\tilde{\phi}=(f(t+r)-f(t-r))/r$. Then, provided $k_{\pm}(x)\equiv f(\pm 1/x)$, $x>0$, are both smoothly extendible to zero, this $\tilde{\phi}$ will be asymptotically regular of order 1. Then the remnants of $\tilde{\phi}$ are given by $\overset{n}{\phi}=(1+\zeta^2)^{-1/2}\{k_+^{(n)}(0)[(1+\zeta^2)^{1/2}-\zeta]^n-k_-^{(n)}(0)[(1+\zeta^2)^{1/2}+\zeta]^n\}$, where we have set $\zeta=-\Omega^{-2}(\partial\Omega/\partial t)|_{\mathcal{H}}\in\mathcal{T}$. The $\mathcal{K}$'s in this example involve various derivatives of $k_{\pm}$ at zero. Explicitly, the first two are given by $\mathcal{K}=4\pi[k_+'(0)+k_-'(0)]$, $\mathcal{K}_\mu=4\pi[k_-''(0)+k_-''(0)]\langle\alpha_\mu,\zeta\rangle$. Thus, the $\mathcal{K}$'s in this example describe radiation emanating from future and past timelike infinity.

We turn next to the Maxwell case. Let $\tilde{F}_{ab}$ be a Maxwell field asymptotically regular of order 2, with remnants $\overset{n}{E}_a$ and $\overset{n}{B}_a$.

Theorem 2: *(i) The conserved quantities linear in this Maxwell field consist precisely of the electric charge [given by Eq. (10)], the magnetic charge [obtained by replacing $\overset{0}{E}_a$ by $\overset{0}{B}_a$ in Eq. (10)], and the family*

$$\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}\equiv\mathcal{K}_{\mu_1\cdots\mu_{n-1}}[\overset{n}{E}{}^m D_m\alpha_\mu]$$
$$=\int[D^a(\overset{n}{E}{}^m D_m\alpha_\mu)\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})-\overset{n}{E}{}^m D_m\alpha_\mu D^a\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})]dS_a, \quad (36)$$

for $n=1,2,3,\ldots$.

(ii) The $\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}$ are trace-free in all indices, totally symmetric in the indices $\mu_1\cdots\mu_{n-1}$, and satisfy

$$\mathcal{E}_{(\mu\mu_1\cdots\mu_{n-1})}=0. \quad (37)$$

(iii) The gauge behavior of $\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}$ is given by

$$\nabla^\mu\mathcal{E}^\nu{}_{\mu_1\cdots\mu_{n-1}}=\tfrac{1}{2}n(n-2)\,\eta_{(\mu_1\mu_2}\mathcal{E}^{\nu\mu}{}_{\mu_3\cdots\mu_{n-1})}-n(n-1)\,\delta^\mu{}_{(\mu_1}\mathcal{E}^\nu{}_{\mu_2\cdots\mu_{n-1})}$$
$$+\frac{n(n-2)}{n-1}\,\eta_{(\mu_1\mu_2}\mathcal{E}^{[\nu\mu]}{}_{\mu_3\cdots\mu_{n-1})}-2n\,\delta^{[\mu}{}_{(\mu_1}\mathcal{E}^{\nu]}{}_{\mu_2\cdots\mu_{n-1})}. \quad (38)$$

We will refer to the $\mathcal{E}$'s as the remnant radiation multipoles of a Maxwell field.

To see that Eq. (36) indeed defines a conserved quantity, take the divergence of the integrand and use that $\overset{n}{E}{}^m D_m\alpha_\mu$ (and $\overset{n}{B}{}^m D_m\alpha_\mu$) satisfy Eq. (19). To prove Eq. (37), we note that its integrand is the divergence of an antisymmetric tensor.[24] Equation (37) implies, in particular, that $\mathcal{E}_\mu$ is zero, and that $\mathcal{E}_{\mu\nu}$ is antisymmetric. While a second family of conserved quantities, $*\mathcal{E}$, associated similarly with $\overset{n}{B}_a$ could be defined, they yield nothing new, for we have[25]

$$*\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}=\frac{n-1}{n}\,\epsilon_{\mu(\mu_1}{}^{\nu\nu_1}\mathcal{E}_{|\nu\nu_1|\mu_2\cdots\mu_{n-1})}. \quad (39)$$

Note that $*\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}$ has the symmetries *(ii)* in Theorem 2 above and that $**\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}=-\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}$. The gauge behavior, Eq. (38), is proved in Appendix B. Note that Eq. (38) yields, in particular, that $\mathcal{E}_{\mu\nu}$ is gauge invariant. The proof that the quantities given by Eq. (36) exhaust the linear, Maxwell conserved quantities is outlined in Appendix B. Here is an example of these electromagnetic conserved quantities. Let $\tilde{\phi}$ be a Klein–Gordon field asymptotically regular of order 1, $\tilde{w}^{ab}$ a constant antisymmetric tensor field on $\tilde{M}$, and set $\tilde{F}_{ab}=\tilde{\boldsymbol{\nabla}}_{[a}(\tilde{w}_{b]m}\tilde{\boldsymbol{\nabla}}^m\tilde{\phi})$. Then this $\tilde{F}_{ab}$ is a solution of Maxwell's equations, asymptotically regular of order 2. Its remnants are given in terms of those of $\tilde{\phi}$ by

$$\overset{n}{E}_a=D_a D_b \overset{n-1}{\phi}\ \xi^b-(n+1)D_m\xi_a D^m \overset{n-1}{\phi}+n^2\overset{n-1}{\phi}\ \xi_a, \tag{40}$$

where we have set $\xi^a=\tilde{w}^{ab}x_b$. Then, the remnant radiation multipoles of $\tilde{F}_{ab}$ can be expressed in terms of those of $\tilde{\phi}$. For instance, we have $\mathcal{E}_{\mu\nu}=\mathcal{K}w_{\mu\nu}$, where we have set $w_{\mu\nu}\equiv 2\xi_a\alpha_{[\mu}D^a\alpha_{\nu]}+D_a\xi_b D^a\alpha_\mu D^b\alpha_\nu$.

We turn finally to linearized gravity. Let $\tilde{K}_{abcd}$ be a linearized gravitational field asymptotically regular of order 3, with remnants $\overset{n}{E}_{ab}$ and $\overset{n}{B}_{ab}$.

Theorem 3: *(i) The conserved quantities linear in this linearized gravitational field consist precisely of the mass–momentum [given by Eq. (11)], the angular momentum [given by Eq. (12)], and*

$$\begin{aligned}\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}&\equiv\mathcal{K}_{\mu_1\cdots\mu_{n-1}}[\overset{n}{E}{}^{kl}D_k\alpha_\mu D_l\alpha_\nu]\\&=\mathcal{E}_{\nu\mu_1\cdots\mu_{n-1}}[\overset{n}{E}_{ab}D^b\alpha_\mu]\\&=\int[D^a(\overset{n}{E}{}^{kl}D_k\alpha_\mu D_l\alpha_\mu)\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})\\&\quad-\overset{n}{E}{}^{kl}D_k\alpha_\mu D_l\alpha_\mu D^a\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})]dS_a,\end{aligned} \tag{41}$$

for $n=1,2,3,\ldots$.

(ii) The $\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}$ are trace-free in all indices, totally symmetric in the indices $\mu_1\cdots\mu_{n-1}$, symmetric in indices μ, ν, and satisfy

$$\mathcal{G}_{\mu(\nu\mu_1\cdots\mu_{n-1})}=0. \tag{42}$$

(iii) The gauge behavior of $\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}$ is given by

$$\begin{aligned}\boldsymbol{\nabla}^\sigma\mathcal{G}^{\mu\nu}{}_{\mu_1\cdots\mu_{n-1}}&=\tfrac{1}{2}n(n-2)\,\eta_{(\mu_1\mu_2}\mathcal{G}^{\mu\nu\sigma}{}_{\mu_3\cdots\mu_{n-1})}-n(n-1)\,\delta^\sigma{}_{(\mu_1}\mathcal{G}^{\mu\nu}{}_{\mu_2\cdots\mu_{n-1})}\\&\quad+\frac{n(n-2)}{n-1}\,\eta_{(\mu_1\mu_2}(\mathcal{G}^{\mu[\nu\sigma]}{}_{\mu_3\cdots\mu_{n-1})}+\mathcal{G}^{\nu[\mu\sigma]}{}_{\mu_3\cdots\mu_{n-1})})\\&\quad-2n\,\delta^{[\sigma}{}_{(\mu_1}(\mathcal{G}^{\mu]\nu}{}_{\mu_2\cdots\mu_{n-1})}+\mathcal{G}^{\nu]\mu}{}_{\mu_2\cdots\mu_{n-1})}).\end{aligned} \tag{43}$$

We will refer to the $\mathcal{G}$'s as the remnant radiation multipoles of a linearized gravitational field.

To see that Eq. (41) indeed defines a conserved quantity, take the divergence of the integrand, and use that $\overset{n}{E}{}^{kl}D_k\alpha_\mu D_l\alpha_\nu$ (and $\overset{n}{B}{}^{kl}D_k\alpha_\mu D_l\alpha_\nu$) satisfy Eq. (19). Equation (42), which is actually equivalent to $\mathcal{E}_{(\nu\mu_1\cdots\mu_{n-1})}=0$, implies, in particular, that $\mathcal{G}_{\mu\nu}=\mathcal{G}_{\mu\nu\sigma}=0$. While a second family of conserved quantities, $*\mathcal{G}$, associated similarly with $\overset{n}{B}_{ab}$ could be defined, they yield nothing new, for we have[26]

$$*\mathcal{G}^{\mu\nu}{}_{\mu_1\cdots\mu_{n-1}}=\frac{n-1}{n}\,\epsilon^{\sigma(\mu}{}_{\lambda(\mu_1}\mathcal{G}^{\nu)\lambda}{}_{|\sigma|\mu_2\cdots\mu_{n-1})}\,. \tag{44}$$

Note that $*\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}$ also satisfies *(ii)* in Theorem 3 and that $**\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}=-\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}$. The proof that the quantities given by Eq. (41) exhaust the linear, gravitational conserved quantities is outlined in Appendix B. We omit the proof of the gauge behavior [Eq. (43)], which is similar to the Maxwell case. Examples of linearized gravitational fields, their remnants, and their remnant radiation multipoles can be constructed in a manner similar to that of the Maxwell case.[27]

One might expect, on physical grounds, that a static field would be characterized completely by its static multipole moments and that its remnant radiation multipoles would all vanish identically. This indeed turns out to be the case. See Appendix A.

IV. CURVED SPACE–TIME

It is natural to ask whether the remnant radiation multipoles constructed above for various fields in Minkowski space–time can be generalized to curved space–time. To address this issue, we first obtain the remnant equations. Let $\tilde{\phi}$ be a Klein–Gordon field asymptotically regular of order 1, so, $\phi(\equiv\Omega^{-1}\tilde{\phi})$ is smoothly extendible to $\mathcal{H}$. Then the Klein–Gordon equation on $\tilde{\phi}$ yields

$$\begin{aligned}0=\tilde{\nabla}^2\tilde{\phi}=\Omega^3[D^2\phi+\lambda^{-1}D^a\lambda\,D_a\phi+\Omega\lambda^{-2}(2\overset{1}{\phi}+\Omega\overset{2}{\phi})\\ +(\phi+\Omega\overset{1}{\phi})(-\lambda^{-2}-\Omega\lambda^{-3}\overset{1}{\lambda}+\tfrac{1}{2}\Omega\lambda^{-2}q^{ab}\overset{1}{q}_{ab})],\end{aligned} \tag{45}$$

where D_a is, as before, the derivative operator on constant-Ω surfaces induced from $\tilde{\boldsymbol{\nabla}}_a$. Evaluating (45) and its first two normal derivatives on $\mathcal{H}$, we obtain, respectively,

$$0=(D^2-1)\overset{0}{\phi}, \tag{46}$$

$$0=D^2\overset{1}{\phi}, \tag{47}$$

$$\begin{aligned}0=D^2\overset{2}{\phi}+3\overset{2}{\phi}-\overset{2}{q}{}^{ab}D_aD_b\overset{0}{\phi}-16\overset{1}{\lambda}\overset{1}{\phi}-14\overset{1}{\lambda}\,D^a\overset{1}{\lambda}\,D_a\overset{0}{\phi}\\ -2(D_aD_b\overset{1}{\lambda})D^a\overset{1}{\lambda}\,D^b\overset{0}{\phi}+32\overset{1}{\lambda}{}^2\overset{0}{\phi}+2(D\overset{1}{\lambda})^2\overset{0}{\phi}.\end{aligned} \tag{48}$$

For the nth derivative, the equation that results has the form

$$0=(D^2+n^2-1)\overset{n}{\phi}-4n^2(n-1)\overset{1}{\lambda}\,\overset{n-1}{\phi}+\cdots, \tag{49}$$

where $\cdots$ involves only remnants of ϕ of order $\leqslant n-2$.

Next, let $\tilde{F}_{ab}$ be a Maxwell field asymptotically regular of order 2. Then Maxwell's equations yield

$$0=\tilde{\boldsymbol{\nabla}}^m\tilde{F}_{ma}=-\lambda\Omega\, D^m E_m\, \boldsymbol{\nabla}_a\Omega-\lambda^{-1}\Omega^2\epsilon_a{}^{bc}[D_b(\lambda B_c)+\tfrac{1}{2}\Omega\mathcal{L}_{\lambda^2 n}(E^m\epsilon_{mbc})]. \tag{50}$$

Evaluating (50) and its first two normal derivatives on $\mathcal{H}$, we obtain, respectively,

$$D_a\overset{0}{E}{}^a=0, \quad D_a\overset{0}{B}{}^a=0, \tag{51}$$

$$D_{[a}\overset{0}{E}_{b]}=0, \quad D_{[a}\overset{0}{B}_{b]}=0, \tag{52}$$

$$D_{[a}\overset{1}{\mathcal{B}}_{b]}=-\tfrac{1}{2}\epsilon_{ab}{}^c(\overset{1}{\mathcal{E}}_c-2\overset{1}{\lambda}\overset{0}{E}_c), \tag{53}$$

$$D_{[a}\overset{1}{\mathcal{E}}_{b]}=\tfrac{1}{2}\epsilon_{ab}{}^c(\overset{1}{\mathcal{B}}_c-2\overset{1}{\lambda}\overset{0}{B}_c), \tag{54}$$

$$D_{[a}\overset{2}{\mathcal{B}}_{b]}=-\epsilon_{ab}{}^c[\overset{2}{\mathcal{E}}_c-4\overset{1}{\lambda}\overset{1}{\mathcal{E}}_c+w_{cd}\overset{0}{E}{}^d], \tag{55}$$

$$D_{[a}\overset{2}{\mathcal{E}}_{b]}=\epsilon_{ab}{}^c[\overset{2}{\mathcal{B}}_c-4\overset{1}{\lambda}\overset{1}{\mathcal{B}}_c+w_{cd}\overset{0}{B}{}^d], \tag{56}$$

where we have set $\overset{1}{\mathcal{E}}_a=\overset{1}{E}_a+\overset{1}{\lambda}\overset{0}{E}_a$, $\overset{2}{\mathcal{E}}_a=\overset{2}{E}_a+2\overset{1}{\lambda}\overset{1}{E}_a+\overset{2}{\lambda}\overset{0}{E}_a$, $\overset{1}{\mathcal{B}}_a=\overset{1}{B}_a+\overset{1}{\lambda}\overset{0}{B}_a$, $\overset{2}{\mathcal{B}}_a=\overset{2}{B}_a+2\overset{1}{\lambda}\overset{1}{B}_a+\overset{2}{\lambda}\overset{0}{B}_a$, and $w_{ab}=-\overset{2}{q}_{ab}+(\tfrac{1}{2}q^2+3\overset{1}{\lambda}{}^2-\overset{2}{\lambda})\overset{0}{q}_{ab}$. For the nth derivative, the equations that result have the form

$$D_{[a}(\overset{n}{B}_{b]}+n\overset{1}{\lambda}\overset{n-1}{B}_{b]})=-\frac{n}{2}\epsilon_{ab}{}^c(\overset{n}{E}_c-n\overset{1}{\lambda}\overset{n-1}{E}_c+\cdots), \tag{57}$$

$$D_{[a}(\overset{n}{E}_{b]}+n\overset{1}{\lambda}\overset{n-1}{E}_{b]})=\frac{n}{2}\epsilon_{ab}{}^c(\overset{n}{B}_c-n\overset{1}{\lambda}\overset{n-1}{B}_c+\cdots), \tag{58}$$

where $\cdots$ involves only remnants of E_a, B_a of order $\leqslant n-2$.

We turn finally to the gravitational field. The remnant field equations of order $\leqslant 2$ were obtained by Beig and Schmidt[8,9] in the vacuum case under the gauge conditions (6). We here drop the assumption of the vacuum and the gauge condition. See Appendix C for an outline of our derivation. The zeroth-order equations are satisfied identically. The first-order equations are

$$\overset{1}{q}_{ab}=-2\overset{1}{\lambda}\overset{0}{q}_{ab}, \tag{59}$$

$$(D^2+3)\overset{1}{\lambda}=0. \tag{60}$$

The second-order equations are

$$\overset{2}{q}=2(D\overset{1}{\lambda})^2+24\overset{1}{\lambda}{}^2-D^2\overset{2}{\lambda}-6\overset{2}{\lambda}-D^m\overset{0}{T}_m-2\overset{0}{T}, \tag{61}$$

$$D_b\overset{2}{q}_a{}^b=32\overset{1}{\lambda}\,D_a\overset{1}{\lambda}+4\,D^b\overset{1}{\lambda}\,D_a\,D_b\overset{1}{\lambda}-4\,D_a\overset{2}{\lambda}-6D_a(D^2\overset{2}{\lambda})+D_a(-D^m\overset{0}{T}_m-2\overset{0}{T})+2\overset{0}{T}_a, \tag{62}$$

$$(D^2-2)\overset{2}{q}_{ab}=8(D\overset{1}{\lambda})^2\overset{0}{q}_{ab}+20\,D_a\overset{1}{\lambda}\,D_b\overset{1}{\lambda}+28\overset{1}{\lambda}D_a\,D_b\overset{1}{\lambda}-36\overset{1}{\lambda}{}^2\overset{0}{q}_{ab}+4D_a\,D^c\overset{1}{\lambda}\,D_b\,D_c\overset{1}{\lambda}$$
$$+4D^c\overset{1}{\lambda}D_aD_b\,D_c\overset{1}{\lambda}-4D_a\,D_b\overset{2}{\lambda}+4\overset{2}{\lambda}\overset{0}{q}_{ab}-D_aD_b(D^2\overset{2}{\lambda})$$
$$-4\overset{0}{T}_{ab}+D_aD_b(-D^m\overset{0}{T}_m-2\overset{0}{T})+4D_{(a}\overset{0}{T}_{b)}\,. \tag{63}$$

The third- and fourth-order equations are not used in what follows and are collected in Appendix C, where we also rewrite the second-order equations in terms of the remnants of the Weyl tensor.

We turn now to the issue of whether or not the various conserved quantities that we defined in flat space–time can be generalized to curved space–time. Recall that a conserved quantity is given by the integral over a cut of $\mathcal{H}$ of a vector field $v^a{}_\Gamma$ on $\mathcal{H}$, where that field is expressed as an algebraic function of the preferred field α_μ of the universal background geometry of $\mathcal{H}$, the remnants of the physical field, the remnants of the geometry, and their derivatives. The divergence of this $v^a{}_\Gamma$ must, for independence of cut, vanish by virtue of the equations satisfied by α_μ and the various remnants. In the special case of flat space–time, we have (or, at least, achieved via gauge) $\overset{k}{\lambda}=0$, $\overset{k}{q}_{ab}=0$, for $k\geqslant 1$, i.e., we have effectively no "remnants of the geometry." Clearly, every conserved quantity, in general, remains a conserved quantity in the special case of flat space–time. But the converse need not hold. Given a conserved quantity in flat space–time—i.e., given a divergence-free field $v^a{}_\Gamma$ constructed from the preferred field α_μ, the remnants of the physical field, and their derivatives—then it may or may not be the case that it is the specialization to flat space–time of some conserved quantity in curved space–time. When it is, we say we have produced a *generalization* of our given flat-space conserved quantity.

Consider first the Klein–Gordon case. We have immediately from Eq. (47), the following.

Theorem 4: *The conserved quantity $\mathcal{K}$ [$n=1$ in Eq. (34)] for the Klein–Gordon field in flat space–time admits a generalization, in the sense described above, to a conserved quantity in curved space–time, namely, that given by*

$$\mathcal{K}=\int D^a\overset{1}{\phi}\,dS_a\,. \tag{64}$$

For the higher-order Klein–Gordon remnant radiation multipoles, we have the following.

Theorem 5: *Fix $n\geqslant 2$. Then the conserved quantity $\mathcal{K}_{\mu_1\cdots\mu_{n-1}}$ for the Klein–Gordon field in flat space–time, given by Eq. (34), does not admit generalization to curved space–time.*

Proof: Let, for contradiction, $v^a{}_{\mu_1\cdots\mu_{n-1}}$ be a generalization to curved space–time. By a simple scaling argument (using, respectively, linearity of the Klein–Gordon remnant field equations in the $\overset{k}{\phi}$ and homogeneity of all remnant field equations in order), we may assume that $v^a{}_{\mu_1\cdots\mu_{n-1}}$ is linear in the $\overset{k}{\phi}$, and of total order n in all remnants. Thus, $v^a{}_{\mu_1\cdots\mu_{n-1}}$ contains no $\overset{k}{\phi}$, for $k>n$, and the term involving $\overset{n}{\phi}$ is, because $v^a{}_{\mu_1\cdots\mu_{n-1}}$ must reduce to the integrand of $\mathcal{K}_{\mu_1\cdots\mu_{n-1}}$ in flat space–time, precisely $\psi_{\mu_1\cdots\mu_{n-1}}D^a\overset{n}{\phi}-\overset{n}{\phi}D^a\psi_{\mu_1\cdots\mu_{n-1}}$, where we have set $\psi_{\mu_1\cdots\mu_{n-1}}=\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$. Denote by $u^a{}_{\mu_1\cdots\mu_{n-1}}$ the term of $v^a{}_{\mu_1\cdots\mu_{n-1}}$ involving $\overset{n-1}{\phi}$. Then, the vanishing of divergence of $v^a{}_{\mu_1\cdots\mu_{n-1}}$ implies

$$D_au^a{}_{\mu_1\cdots\mu_{n-1}}\hat{=}-4n^2(n-1)\overset{1}{\lambda}\psi_{\mu_1\cdots\mu_{n-1}}\overset{n-1}{\phi}\,, \tag{65}$$

where $\hat{=}$ stands for equality modulo Klein–Gordon remnants of order $\leqslant n-2$. But there exists no such $u^a{}_{\mu_1\cdots\mu_{n-1}}$, as one sees by the following steps. First, add to $v^a{}_{\mu_1\cdots\mu_{n-1}}$ a divergence of an antisymmetric tensor field to achieve the form

$$u^a{}_{\mu_1\cdots\mu_{n-1}}=w_{\mu_1\cdots\mu_{n-1}}D^a\overset{n-1}{\phi}-\overset{n-1}{\phi}D^a w_{\mu_1\cdots\mu_{n-1}}, \tag{66}$$

with $w_{\mu_1\cdots\mu_{n-1}}$ linear in $\overset{1}{\lambda}$ and $\psi_{\mu_1\cdots\mu_{n-1}}$, and from Eqs. (65) and (49), satisfying

$$(D^2+n^2-2n)w_{\mu_1\cdots\mu_{n-1}}=4n^2(n-1)\overset{1}{\lambda}\psi_{\mu_1\cdots\mu_{n-1}}. \tag{67}$$

Second, replace every occurrence of $\overset{1}{\lambda}$ in Eq. (67) by α_μ. Then, under this substitution, $w_{\mu_1\cdots\mu_{n-1}}$ reduces to the form $w_{\mu_1\cdots\mu_{n-1}}=c\,D_a\alpha_\mu D^a\psi_{\mu_1\cdots\mu_{n-1}}+c'\alpha_\mu\psi_{\mu_1\cdots\mu_{n-1}}$, for some constants c,c'. Since α_μ satisfies Eq. (60), which is the only property of $\overset{1}{\lambda}$ that may be used in establishing (67), it follows that Eq. (67) must continue to hold after replacing $\overset{1}{\lambda}$ therein by α_μ. However, under this substitution, Eq. (67) becomes

$$2[(n+1)c+c'][D_a\alpha_\mu D^a\psi_{\mu_1\cdots\mu_{n-1}}+(n-1)\alpha_\mu\psi_{\mu_1\cdots\mu_{n-1}}]=4n^2(n-1)\alpha_\mu\psi_{\mu_1\cdots\mu_{n-1}}, \tag{68}$$

which can never hold. □

We turn next to Maxwell fields. We have the following.

Theorem 6: *Let* $\overset{0}{B}_a=0$, *and let the stress–energy tensor* T_{ab} *satisfy*

$$\overset{0}{T}_a=0,\quad \overset{0}{T}_{ab}\overset{0}{E}{}^b-\overset{0}{T}\overset{0}{E}_a=0. \tag{69}$$

Then the conserved quantity $\mathcal{E}_{\mu\nu}$ *[of Eq. (36)] for the Maxwell field in flat space–time admits a generalization to a conserved quantity in curved space–time, namely, that given by*

$$\mathcal{E}_{\mu\nu}=\int s^{ab}\alpha_{[\nu}D_{|b|}\alpha_{\mu]}\,dS_a, \tag{70}$$

where we have set

$$\begin{aligned}s_{ab}={}&2D_{(a}\overset{2}{\mathcal{E}}_{b)}-2D^c\overset{2}{\mathcal{E}}_c\overset{0}{q}_{ab}+16\overset{1}{\mathcal{E}}_{(a}D_{b)}\overset{1}{\lambda}-8\overset{1}{\mathcal{E}}{}^cD_c\overset{1}{\lambda}\overset{0}{q}_{ab}+16\psi\overset{1}{\lambda}D_aD_b\overset{1}{\lambda}+8\overset{1}{\lambda}\overset{0}{E}_{(a}D_{b)}\overset{1}{\lambda}\\&+8\overset{1}{\lambda}{}^2D_a\overset{0}{E}_b+[12\psi\overset{1}{\lambda}{}^2-20\overset{1}{\lambda}\overset{0}{E}{}^cD_c\overset{1}{\lambda}-4\psi(D\overset{1}{\lambda})^2]\overset{0}{q}_{ab}+4\psi w_{ab}+4\psi\overset{0}{T}_{ab}-4\psi\overset{0}{T}\overset{0}{q}_{ab}\\&-11(\psi D_aD_b\overset{2}{\lambda}+\overset{2}{\lambda}D_aD_b\psi-2D_{(a}\psi D_{b)}\overset{2}{\lambda})-\psi(11D^2\overset{2}{\lambda}+22\overset{2}{\lambda})\overset{0}{q}_{ab},\end{aligned} \tag{71}$$

where ψ *is so chosen to satisfy* $D_a\psi=\overset{0}{E}_a$.[28]

The integrand reduces, in flat space–time, to the integrand of $\mathcal{E}_{\mu\nu}$ therein plus a divergence, namely, $D_b(\overset{2}{E}{}^{[a}\alpha_{[\mu}D^{b]}\alpha_{\nu]})$, of an antisymmetric tensor field. The demonstration of $D_bs^{ab}=0$ is given in Appendix B. The above conditions, (69), on T_{ab} are satisfied when the space–time is vacuum, and also when the source is the Maxwell field itself. But the condition (69) need not be satisfied in the presence of other matter sources. It is readily verified that this generalized $\mathcal{E}_{\mu\nu}$ is again gauge invariant.

For higher-order Maxwell remnant radiation multipoles, we have the following theorem.

Theorem 7: *Fix* $n \geqslant 3$. *Then the conserved quantity* $\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}$ *for the Maxwell field in flat space–time, given by Eq. (36), does not admit generalization to curved space–time.*

The proof of Theorem 7 is similar to that of Theorem 5, and is therefore omitted.

V. CONCLUSION

We have constructed, for each of a Klein–Gordon field, a Maxwell field, and a linearized gravitational field in Minkowski space–time, a hierarchy of conserved quantities that we call the remnant radiation multipoles. In the cases of Klein–Gordon and Maxwell, we have generalized the remnant radiation monopoles to curved space–time. There follows a discussion of some outstanding issues.

Does the remnant gravitational monopole admit generalization to curved space–time? We conjecture that the answer is yes. In Appendix C we give the remnant field equations necessary for addressing this question. There we also display a candidate for a curved-space gravitational remnant monopole. This candidate has the attractive feature that its divergence, which could, in principle, have contained remnants of order as high as 3, contains only remnants of order $\leqslant 2$. Although the existence of this candidate lends some support to the conjecture, it is, of course, far from a proof of it. Work is in progress to settle this conjecture. We further conjecture that none of the higher-order remnant radiation multipoles for linearized gravitational fields admit generalization to curved space–time.

The way we introduced the Klein–Gordon and Maxwell remnant radiation monopoles in a curved space–time involves a quite strong falloff condition, namely, $\overset{0}{B}_{ab}=0$, on the gravitation remnants. In the Klein–Gordon case, this restriction is, in fact, unnecessary. Indeed, in the absence of this condition, the first-order remnant field equation on $\overset{1}{\phi}$ becomes $D_a v^a = 0$, with $v^a = D^a \overset{1}{\phi} - \overset{1}{q}{}^{ab} D_b \overset{0}{\phi} + \frac{1}{2} \overset{2}{q} D^a \overset{0}{\phi} + \overset{1}{\lambda} D^a \overset{0}{\phi}$. Thus, $\mathcal{K} \equiv \int_C v^a \, dS_a$ remain conserved in asymptotic conditions weaker than the ones presently imposed. Can other conserved quantities be defined with such weaker asymptotic conditions?

Do there exist conserved quantities analogous to our remnant radiation multipoles, but defined at null rather than spatial infinity? And if so, are there any simple relations between the values of corresponding quantities at spatial and null infinity? In Minkowski space–time, it should not be too difficult to answer these two questions. A relevant observation[29] is that, in the case of Minkowski space–time corresponding conserved quantities, in general, take different values at spatial and null infinity. This result suggests that ''remnant radiation'' is capable of escaping between spatial infinity and null infinity. Recall that Newman, Penrose, and Exton[30,31] have introduced certain conserved quantities at null infinity in curved space–time. Are these quantities analogs, in any sense, of the remnant radiation monopoles?

It is unfortunate that the present treatment of asymptotic quantities involves such complicated algebra. It is not entirely clear whether these complications are inherent in the subject itself, or merely a reflection of the present techniques. One case in which we know that these techniques are the culprit is that of stationary space–times. It is not hard to convince oneself that the present framework, in the case of stationary space–times, is essentially equivalent to the usual formalism involving a three-dimensional manifold of trajectories. Since the stationary gravitational multipole moments of all order can be defined within this three-dimensional formalism, it should also be possible, in principle, to define these very same moments within the present framework. However, it already seems difficult to define even the first few stationary multipoles in the present framework. Unlike the three-dimensional formalism, the present framework is not well adapted to the presence of Killing fields. For example, to treat Killing's equation order by order yields a complicated set of remnant equations. Finding a more natural way of dealing with stationary space–times within the present framework may give some clue as to how to tame its algebraic complexity. Indeed, it may further lead to a generalization of the stationary multipoles to more general asymptotically flat space–times.

Here we have restricted our consideration to conserved quantities that are linear in (the highest-order part of the remnants of) the physical fields. More generally, one might allow polynomial dependence on the remnants. A candidate for a conserved quantity quadratic in the remnants has been given by Beig:[9] $\int \epsilon^{kl(a} D_k \overset{1}{\lambda} \overset{0}{E}_l{}^{b)} \alpha_{[\mu} D_{|b|} \alpha_{\nu]} \, dS_a$, where ξ^a is any Killing field in $\mathcal{H}$. However, as shown in Ref. 9, this quantity vanishes identically by virtue of the second-order gravitational remnant equations. It would be of interest to carry out a systematic search for polynomial conserved quantities. One might even search for conserved quantities with nonpolynomial dependence on the remnant fields, but the fact that these remnant fields have complicated gauge behavior rather suggests that no such quantities will exist.

ACKNOWLEDGMENTS

I am very grateful to R. Geroch, my thesis advisor, for his guidance, encouragement, and many helpful suggestions. I wish to thank R. M. Wald for several valuable suggestions. This work was supported in part by National Science Foundation Grant No. PHY 95-14726 to the University of Chicago.

APPENDIX A: STATIONARY FIELDS IN MINKOWSKI SPACE–TIME

Consider in Minkowski space–time a physical field that is static, i.e., that is invariant under a time translation in the space–time. In this appendix, we do two things: Express within the present framework the ordinary static multipole moments of such a field; and show that, in this static case, all the remnant radiation multipoles vanish. We will only discuss linearized gravity here since the treatment of Klein–Gordon and Maxwell fields is similar and simpler.

Let $\tilde{K}_{abcd}$ be a linearized gravitational field in Minkowski space–time $\tilde{M}$, asymptotically regular of order 3. Further, let $\tilde{K}_{abcd}$ be static, i.e., let

$$\mathcal{L}_{\tilde{t}} \tilde{K}_{abcd} = 0, \tag{A1}$$

where $\tilde{t}^a$ is a (unit) timelike Killing field in $\tilde{M}$. Denote by $\zeta\ (=\Omega^{-2}\mathcal{L}_{\tilde{t}}\Omega)$ the corresponding unit time translation on $\mathcal{H}$. Taking the remnants of Eq. (A1), we obtain the following equations on the remnant fields:[32]

$$\mathcal{L}_{D\zeta} \overset{n}{E}_{ab} - (n+1)\zeta \overset{n}{E}_{ab} + 2\epsilon^{lm}{}_{(a} \overset{n}{B}_{b)l} D_m \zeta = 0, \tag{A2}$$

for $n=0,1,2,\ldots$. Set $\overset{n}{\phi}_E = \overset{n}{E}_{ab} D^a\zeta D^b\zeta$. Then this $\overset{n}{\phi}_E$ satisfies Eq. (19), and, from Eq. (A2), also

$$\mathcal{L}_{D\zeta} \overset{n}{\phi}_E - (n+1)\zeta \overset{n}{\phi}_E = 0. \tag{A3}$$

Under a gauge transformation $\overset{n}{\phi}_E$ changes according to Eq. (29). The same equations hold, similarly, on $\overset{n}{\phi}_B = \overset{n}{B}_{ab} D^a\zeta D^b\zeta$, $n=0,1,\ldots$. We note that the $\overset{n}{E}_{ab}$ and $\overset{n}{B}_{ab}$ for this static linearized gravitational field can be expressed in terms of $\overset{n}{\phi}_E$ and $\overset{n}{\phi}_B$. Indeed, we have, from Eq. (A2) and Eq. (27), that

$$\overset{n}{E}_{ab}=\frac{2\zeta^2+1}{(n+1)(n+2)}D_a D_b \overset{n}{\phi}_E+\frac{3D_{(a}\zeta D_{b)} D_m \overset{n}{\phi}_E D^m\zeta}{2(n+1)(n+2)}-\frac{D^k D^l \overset{n}{\phi}_E D_k\zeta D_l\zeta \overset{0}{q}_{ab}}{2(n+1)(n+2)}$$
$$+\frac{5\zeta D_{(a}\zeta D_{b)}\overset{n}{\phi}_E}{2(n+1)}+\frac{4n+5}{2(n+2)}\overset{n}{\phi}_E D_a\zeta D_b\zeta+\frac{(n-3)\zeta^2+2n+1}{2(n+2)}\overset{n}{\phi}_E \overset{0}{q}_{ab}$$
$$+\frac{2\zeta\epsilon^{kl}{}_{(a}D_{b)} D_k\overset{n}{\phi}_B D_l\zeta}{(n+1)(n+2)}+\frac{2\epsilon^{kl}{}_{(a}D_{b)}\zeta D_k\overset{n}{\phi}_B D_l\zeta}{n+1}, \tag{A4}$$

and similarly for $\overset{n}{B}_{ab}$.

Now consider, for $n=0,1,2,...$,

$$M_{\mu_1\cdots\mu_n}[\overset{n}{\phi}_E]\equiv\frac{(2n+1)!}{2^{n+1}(n!)^3}\int_C[\overset{n}{\phi}_E(\alpha_{\mu_1}+\langle\alpha_{\mu_1},\zeta\rangle\zeta)\cdots(\alpha_{\mu_n}+\langle\alpha_{\mu_n},\zeta\rangle\zeta)(1+\zeta^2)^{-1}D^a\zeta]dS_a. \tag{A5}$$

The integrand on the right above is divergence-free, by Eq. (A3), and so Eq. (A5) defines, for each n, a conserved quantity, $M_{\mu_1\cdots\mu_n}[\overset{n}{\phi}_E]$. These are precisely the ordinary static electric multipole moments of the linearized gravitational field.[33] They are totally symmetric, and satisfy

$$0=\zeta^{\mu_1}M_{\mu_1\mu_2\mu_3\cdots\mu_n}, \tag{A6}$$

$$0=\eta^{\mu_1\mu_2}M_{\mu_1\mu_2\mu_3\cdots\mu_n}, \tag{A7}$$

$$\boldsymbol{\nabla}_\mu M_{\mu_1\cdots\mu_n}=-(2n-1)h_{\mu(\mu_1}M_{\mu_2\cdots\mu_n)}+(n-1)h_{(\mu_1\mu_2}M_{\mu_3\cdots\mu_n)\mu}, \tag{A8}$$

where we have set $h_{\mu\nu}=\eta_{\mu\nu}+\zeta_\mu\zeta_\nu$. To prove Eq. (A7), use that $\overset{n}{\phi}_E$ satisfies Eq. (19); to prove Eq. (A8), use the gauge behavior (29) of $\overset{n}{\phi}_E$. Similarly, we obtain the magnetic multipole moments, $M_{\mu_1\cdots\mu_n}[\overset{n}{\phi}_B]$. These two sets of moments are the linearized versions of Hansen's mass and angular momentum multipole moments, respectively.

Finally, we show that all of the gravitational remnant radiation multipoles [the $\mathcal{G}$'s introduced in Eq. (41), Sec. III], vanish for a static linearized gravitational field in Minkowski space–time. To see this, substitute Eq. (A4) into the integrand of $\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}$, to obtain

$$\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}=\mathcal{K}[c_1\eta_{0\mu}\eta_{0\nu}\overset{n}{\phi}_E+c_2\eta_{\mu\nu}\overset{n}{\phi}_E+c_3\eta_{0(\mu}\mathcal{L}_{\xi^*_{\nu)}}\overset{n}{\phi}_B+c_4\mathcal{L}_{\xi^*_{(\mu}}\mathcal{L}_{\xi^*_{\nu)}}\overset{n}{\phi}_E$$
$$+c_5\eta_{0(\mu}\mathcal{L}_{\xi_{\nu)}}\overset{n}{\phi}_E+c_6\mathcal{L}_{\xi_{(\mu}}\mathcal{L}_{\xi^*_{\nu)}}\overset{n}{\phi}_B+c_7\mathcal{L}_{\xi_{(\mu}}\mathcal{L}_{\xi_{\nu)}}\overset{n}{\phi}_E]_{\mu_1\cdots\mu_{n-1}}, \tag{A9}$$

where $c_1,...,c_7$ are certain constants, and $\xi^a{}_\mu$ and $\xi^{*a}{}_\mu$ are the Killing fields given by $\xi^a{}_\mu=\zeta D^a\alpha_\mu-\alpha_\mu D^a\zeta$ and $\xi^{*a}{}_\mu=\epsilon^{abc}D_b\zeta D_c\alpha_\mu$. Let C denote the $\zeta=0$ 2-sphere section of $\mathcal{H}$. We show that each term on the right in eqn. (A9) contributes zero by evaluating the integral over C. The first four terms contribute zero by virtue of the fact that each of the terms satisfies the same equations as a static nth-order Klein–Gordon remnant field $\overset{n}{\phi}$, and that, for any such $\overset{n}{\phi}$,

$\int_C \overset{n}{\phi}\alpha_{\mu_1}\cdots\alpha_{\mu_k}D^a\zeta\, dS_a=0$, for $0\leqslant k<n$. The fifth and sixth terms contribute zero because for any $\overset{n}{\phi}$ as above, $\mathcal{L}_{\xi_\mu}\overset{n}{\phi}$ vanishes on C, and $\mathcal{L}_{D\zeta}(\mathcal{L}_{\xi_\mu}\overset{n}{\phi})$ is a sum of two terms, one of which [namely $(n+1)\zeta\mathcal{L}_{\xi_\mu}\overset{n}{\phi}$] vanishes on C and the other [namely $-(\overset{n}{\phi})_\mu\equiv -D^a\alpha_\mu D_a\overset{n}{\phi}+(n+1)\alpha_\mu\overset{n}{\phi}$] satisfies the $(n+1)$th remnant field equation and is static. Finally, the last term contributes zero because $(\mathcal{L}_{\xi_\mu}\mathcal{L}_{\xi_\nu}\overset{n}{\phi}_E)$ is a translation times a term that satisfies the $(n+1)$th remnant field equation and is static, and because $\mathcal{L}_{D\zeta}(\mathcal{L}_{\xi_\mu}\mathcal{L}_{\xi_\nu}\overset{n}{\phi}_E)$ is a sum of two terms, one of which [namely, $(n+1)\zeta\mathcal{L}_{\xi_\mu}\mathcal{L}_{\xi_\nu}\overset{n}{\phi}_E-\mathcal{L}_{\xi_{(\mu}}(\overset{n}{\phi}_E)_{\nu)}$] vanishes on C and the other (namely, $-[\mathcal{L}_{\xi_{(\mu}}\overset{n}{\phi}_E]_{\nu)}$) satisfies the $(n+1)$th remnant field equation and is equal to a static field $\overset{n+1}{\phi}$ on C. □

APPENDIX B: MISCELLANEOUS RESULTS

Appendix B 1 contains the proofs of item *(i)* of each of Theorems 1–3. Appendix B 2 contains the proofs of item *(iii)* of each of Theorems 1–2. Appendix B 3 completes the proof that the $\mathcal{E}_{\mu\nu}$ we introduced in Theorem 6 is indeed conserved.

1. The remnant radiation multipoles exhaust the conserved quantities in Minkowski space–time

We first show that, in the Klein–Gordon case, the $\mathcal{K}$'s of Eq. (34) exhaust all conserved quantities in Minkowski space–time linear in $\overset{n}{\phi}$ and multilinear in $\mathcal{T}$. Sketch of proof: Let $v^a{}_\Gamma$ be a divergence-free vector field on $\mathcal{H}$, constructed linearly in the $\overset{n}{\phi}$, and multilinearly in $\mathcal{T}$. (We introduce the subscript Γ to stand for any Greek indices that may be attached to v^a.) Since the various $\overset{n}{\phi}$ are uncoupled in (19), we may take $v^a{}_\Gamma$ to depend on just a single remnant field, say $\overset{n}{\phi}$. Then $v^a{}_\Gamma$ takes the form

$$v^a{}_\Gamma=\sum_{k=0}^{s} w^{aa_1\cdots a_k}{}_\Gamma D_{a_1}\cdots D_{a_k}\overset{n}{\phi}, \tag{B1}$$

where s is the order of the highest derivative in $v^a{}_\Gamma$. We may assume $w^{aa_1\cdots a_s}{}_\Gamma=w^{(aa_1\cdots a_s)}{}_\Gamma$, since any parts of $w^{aa_1\cdots a_s}{}_\Gamma$ antisymmetric between "a" and another index can be eliminated by adding to $v^a{}_\Gamma$ the divergence of an antisymmetric second-rank tensor field, and any parts antisymmetric between two indices neither "a" can be eliminated using the definition of the Riemann tensor. It now follows, from $D_a v^a{}_\Gamma=0$, that $w^{aa_1\cdots a_s}{}_\Gamma=q^{(aa_1}u^{a_2\cdots a_s)}{}_\Gamma$, for some tensor field $u^{a_2\cdots a_4}{}_\Gamma$. Were $s\geqslant 2$, then this term could now be eliminated in its entirety by adding to $v^a{}_\Gamma$ a divergence, namely $D_{a_2}[2(D^{[a}D_{a_3}\cdots D_{a_s}\overset{n}{\phi})u^{a_2]a_3\cdots a_4}{}_\Gamma]$, of an antisymmetric tensor. So we may assume $s=1$ in Eq. (B1), i.e., we may set $v^a=w_\Gamma D^a\overset{n}{\phi}-\overset{n}{\phi}w^a{}_\Gamma$. It now follows, again from $D_a v^a{}_\Gamma=0$, that $w^a{}_\Gamma=D^a w_\Gamma$, where w_Γ is some solution of Eq. (19). But this w_Γ must be multilinear in $\mathcal{T}$, and the only[34] such solution of Eq. (19) is $\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$.[35]

We next show that, in the Maxwell case, the $\mathcal{E}$'s of Eq. (36), together with the electric and magnetic charges, exhaust all conserved quantities in Minkowski space–time linear in the remnants of the Maxwell field and multilinear in $\mathcal{T}$. Sketch of proof: Let $v^a{}_\Gamma$ be a divergence-free vector field on $\mathcal{H}$, constructed linearly in the $\overset{n}{E}_a$ and multilinearly in $\mathcal{T}$. As before, we may take $v^a{}_\Gamma$ to depend on a single remnant field, $\overset{n}{E}_a$. If $n=0$, the result, that $v^a = \overset{0}{E}{}^a$ follows by setting $\overset{0}{E}_a = D_a\psi$ with $D^2\psi=0$, and using the Klein–Gordon result. So, let $n\geqslant 1$. Then $v^a{}_\Gamma$ takes the form

$$v^a{}_\Gamma = \sum_{k=0}^{s} w^{aa_1\cdots a_{k+1}}{}_\Gamma D_{a_1}\cdots D_{a_k}\overset{n}{E}_{a_{k+1}}, \quad n\geqslant 1, \tag{B2}$$

where s is the order of the highest derivative in $v^a{}_\Gamma$. An argument similar to the Klein–Gordon case shows that $v^a{}_\Gamma$ can be brought to the form

$$v^a{}_\Gamma = w_{b\Gamma}\, D^a\overset{n}{E}{}^b - \overset{n}{E}_b\, D^a w^b{}_\Gamma + \mu_\Gamma \overset{n}{E}{}^a, \tag{B3}$$

where $(D^2+n^2-2)w_{a\Gamma} = D_a\mu_\Gamma$. We may achieve $\mu_\Gamma=0$ in (B2) by adding to $v^a{}_\Gamma$ a divergence of an antisymmetric tensor field, namely, $D_b(2\overset{n}{E}{}^{[a}\, D^{b]}w_\Gamma + 2w_\Gamma\, D^{[a}\overset{n}{E}{}^{b]} + (2/n^2)c_\Gamma\, D^{[a}\overset{n}{E}{}^{b]})$, where c_Γ is a certain constant and where we have set $w_\Gamma = (1/n^2)(-\mu_\Gamma + D_a w^a{}_\Gamma)$. Now $w^a{}_\Gamma$ satisfies precisely the same equations as $\overset{n}{E}{}^a$. The conserved quantity thus arises from the ''symplectic product'' between $\overset{n}{E}_a$ and $w_{a\Gamma}$. But this $w^a{}_\Gamma$ must be multilinear in $\mathcal{T}$, and the only such solution of Eq. (23) is $w^a{}_{\mu\mu_1\cdots\mu_{n-1}} = \mathcal{C}(\alpha_\mu\cdots\alpha_{\mu_{n-1}})D^a\alpha_\mu + (1/n^2)D^a(D^b\alpha_\mu\, D_b\mathcal{C}(\alpha_\mu\cdots\alpha_{\mu_{n-1}})) - \alpha_\mu\mathcal{C}(\alpha_\mu\cdots\alpha_{\mu_{n-1}})$.

Finally, we show that, in the case of linearized gravity, the $\mathcal{G}$'s [of Eq. (41)], together with the mass–momentum and angular momentum exhaust all conserved quantities in Minkowski space–time linear in remnants of the linearized gravitational field and multilinear in $\mathcal{T}$. Sketch of proof: Let $v^a{}_\Gamma$ be a divergence-free vector field on $\mathcal{H}$, constructed linearly in the $\overset{n}{E}_{ab}$ and multilinearly in $\mathcal{T}$. As before, we may take $v^a{}_\Gamma$ to depend on a single remnant field, $\overset{n}{E}_{ab}$. If $n=0$, the result, that $v^a{}_\mu = \overset{0}{E}{}^{ab}D_b\alpha_\mu$, follows by setting $\overset{0}{E}_{ab} = D_a\, D_b\psi + \psi \overset{0}{q}_{ab}$ with $D^2\psi = -3\psi$, and using the Klein–Gordon result. If $n=1$, the result, that $v^a{}_{\mu\nu} = \overset{1}{E}{}^{ab}\alpha_{[\mu}\, D_b\alpha_{\nu]}$, follows by setting $\overset{1}{E}_{ab} = D_{(a}u_{b)}$ with $D^2u_a = -2u_a$, $D_a u^a = 0$, and using the Maxwell result. So, let $n\geqslant 2$. Then $v^a{}_\Gamma$ takes the form

$$v^a{}_\Gamma = \sum_{k=0}^{s} w^{aa_1\cdots a_{k+2}}{}_\Gamma D_{a_1}\cdots D_{a_k}\overset{n}{E}_{a_{k+1}a_{k+2}}, \quad n\geqslant 2, \tag{B4}$$

where s is the order of the highest derivative in $v^a{}_\Gamma$. An argument similar to the Klein–Gordon case shows that $v^a{}_\Gamma$ can be brought to the form

$$v^a{}_\Gamma = w_{bc\Gamma}D^a\overset{n}{E}{}^{bc} - \overset{n}{E}_{bc}D^a w^{bc}{}_\Gamma + \overset{n}{E}{}^{ab}u_{b\Gamma}, \tag{B5}$$

where $w_{ab\Gamma}$ is symmetric and trace-free, and satisfies $(D^2+n^2-3)w_{ab\Gamma} = D_{(a}u_{b)\Gamma} - \frac{1}{3}\overset{0}{q}_{ab}D_m u^m{}_\Gamma$. We may achieve $u_{a\Gamma}=0$ in (B5) by adding to $v^a{}_\Gamma$ a divergence of an antisymmetric tensor field. The result now follows from an argument similar to the Maxwell case.

2. Gauge behavior of remnant radiation multipoles

Here we prove that the gauge behavior of the Klein–Gordon and Maxwell remnant radiation multipoles is that given, respectively, by Eqs. (35) and (38).

For the Klein–Gordon case, denote by $k^a{}_{\mu_1\cdots\mu_{n-1}}$ the integrand of Eq. (34). Then we have

$$\begin{aligned}\nabla_\mu k^a{}_{\mu_1\cdots\mu_{n-1}} &= nD^a(\mathcal{L}_{D\alpha_\mu}\overset{n-1}{\phi} - n\alpha_\mu \overset{n-1}{\phi})\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})\\ &\quad - n(\mathcal{L}_{D\alpha_\mu}\overset{n-1}{\phi} - n\alpha_\mu\overset{n-1}{\phi})D^a\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}}).\\ &= \tfrac{1}{2}n(n-2)\,\eta_{(\mu_1\mu_2}k^a_{\mu_3\cdots\mu_{n-1})\mu} - n(n-1)\,\eta_{\mu(\mu_1}k^a_{\mu_2\cdots\mu_{n-1})}\\ &\quad + D_m[2n\,D^{[a}\overset{n-1}{\phi}\,D^{m]}\alpha_\mu\,\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})\\ &\quad + 2n\overset{n-1}{\phi}\,D^{[a}\alpha_\mu D^{m]}\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})],\end{aligned} \tag{B6}$$

where we used, in the first step, Eqs. (34) and (29), and, in the second, the identity[36]

$$\begin{aligned}D^m\alpha_\mu D_m\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}}) &= -(n-1)\,\alpha_\mu\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}}) + (n-1)\,\eta_{\mu(\mu_1}\mathcal{C}(\alpha_{\mu_2}\cdots\alpha_{\mu_{n-1})})\\ &\quad - \frac{n-2}{2}\,\eta_{(\mu_1\mu_2}\mathcal{C}(\alpha_{\mu_3}\cdots\alpha_{\mu_{n-1})}\alpha_\mu).\end{aligned} \tag{B7}$$

Integrate over a cut of $\mathcal{H}$.

For the Maxwell case, denote by $(e^E)^a_{\nu\mu_1\cdots\mu_{n-2}}$ the integrand of Eq. (36), and set $\xi^a{}_{\mu\nu} = 2\alpha_{[\mu}D^a\alpha_{\nu]}$. Then we have

$$\begin{aligned}\nabla^\mu[(e^E)^a]^\nu{}_{\mu_1\cdots\mu_{n-1}} &= nD^a[(\mathcal{L}_{D\alpha_\mu}\overset{n-1}{E}{}_m - n\alpha_\mu\overset{n-1}{E}{}_m + \epsilon_{mkl}\overset{n-1}{B}{}^k D^l\alpha_\mu)D^m\alpha_\nu]\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})\\ &\quad - n[(\mathcal{L}_{D\alpha_\mu}\overset{n-1}{E}{}_m - n\alpha_\mu\overset{n-1}{E}{}_m + \epsilon_{mkl}\overset{n-1}{B}{}^k D^l\alpha_\mu)D^m\alpha_\nu]D^a\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})\\ &= \nabla_\mu k^a_{\mu_1\cdots\mu_{n-1}}[\overset{n-1}{E}{}_m D^m\alpha_\nu] + \frac{n(n-2)}{n-1}\,\eta_{(\mu_1\mu_2}[(e^E)^a]^{[\nu\mu]}{}_{\mu_3\cdots\mu_{n-1})}\\ &\quad - 2n\,\delta^{[\mu}{}_{(\mu_1}[(e^E)^a]^{\nu]}{}_{\mu_2\cdots\mu_{n-1})} - \frac{n}{n-1}D_m[2(D^{[a}\overset{n-1}{E}{}_k D^{m]}\xi^k{}_{\mu\nu}\\ &\quad + \overset{n-1}{E}{}^{[a}\xi^{m]}{}_{\mu\nu})\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}}) + 2\overset{n-1}{E}{}^k D^{[a}\xi_{k\mu\nu}D^{m]}\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})],\end{aligned} \tag{B8}$$

where we used, in the first step, Eqs. (30) and (36), and, in the second, the identity, Eq. (B7), again. Integrate over a cut of $\mathcal{H}$.

3. Completion of proof of Theorem 6

In our proof of Theorem 6, we omitted one step: The demonstration the the s_{ab} of Eq. (71) is indeed divergence-free. Here we supply that step. We have

$$D^b s_{ab} = -2\epsilon_{abc} D^b(w^{cd}\overset{0}{B}_d) + 16\overset{1}{\lambda}\epsilon_{abc} D^b\overset{1}{\lambda}\overset{0}{B}{}^c - 8\psi\overset{0}{T}_a + 4(\overset{0}{T}_{ab}\overset{0}{E}{}^b - \overset{0}{T}\overset{0}{E}_a) = 0, \tag{B9}$$

where, in the first step, we used the following six equations:

$$D^b(2D_{(a}\overset{2}{\mathcal{E}}_{b)} - 2D^c\overset{2}{\mathcal{E}}_c\overset{0}{q}_{ab}) = 16(D_{[a}\overset{1}{\mathcal{E}}_{b]})\, D^b\overset{1}{\lambda} + 24\overset{1}{\lambda}\overset{1}{\mathcal{E}}_a - 16\overset{1}{\lambda}{}^2\overset{0}{E}_a - 4w_{ab}\overset{0}{E}{}^b - 2\epsilon_{abc}D^b(w^{cd}\overset{0}{B}_d)$$
$$+16\overset{1}{\lambda}\epsilon_{abc}\, D^b\overset{1}{\lambda}\overset{0}{B}{}^c, \tag{B10}$$

$$D^b(16\overset{1}{\mathcal{E}}_{(a}\, D_{b)}\overset{1}{\lambda} - 8\overset{1}{\mathcal{E}}{}^c\, D_c\overset{1}{\lambda}\,\overset{0}{q}_{ab}) = -16(D_{[a}\overset{1}{\mathcal{E}}_{b]})D^b\overset{1}{\lambda} - 24\overset{1}{\lambda}\overset{1}{\mathcal{E}}_a + 16(\overset{0}{E}{}^c\, D_c\overset{1}{\lambda})D_\alpha\overset{1}{\lambda}, \tag{B11}$$

$$D^b(4\psi w_{ab}) = 4w_{ab}\overset{0}{E}{}^b - 8\psi\overset{1}{\lambda}\, D_a\overset{1}{\lambda} - 8\psi D_a\, D_b\overset{1}{\lambda}\, D^b\overset{1}{\lambda} + 22\psi D_a(D^2\overset{2}{\lambda})$$
$$+\psi D_a(2\, D_b T^b + 4T) - 8\psi T_a, \tag{B12}$$

$$D^b[16\psi\overset{1}{\lambda}D_a\, D_b\overset{1}{\lambda} + 8\overset{1}{\lambda}\overset{0}{E}_{(a}\, D_{b)}\overset{1}{\lambda} + 8\overset{1}{\lambda}{}^2\, D_a\overset{0}{E}_b + (12\psi\overset{1}{\lambda}{}^2 - 20\overset{1}{\lambda}\overset{0}{E}{}^c\, D_c\overset{1}{\lambda} - 4\psi(D\overset{1}{\lambda})^2)\overset{0}{q}_{ab}]$$
$$= 16\overset{1}{\lambda}\overset{0}{E}_a - 16(\overset{0}{E}{}^b D_b\overset{1}{\lambda})D_a\overset{1}{\lambda} + 8\psi\overset{1}{\lambda}\, D_a\overset{1}{\lambda} + 8\psi D_a\, D_b\overset{1}{\lambda}\, D^b\overset{1}{\lambda}, \tag{B13}$$

$$D^b[-11(\psi D_a\, D_b\overset{2}{\lambda} + \overset{2}{\lambda}D_a\, D_b\psi - 2D_{(a}\psi\, D_{b)}\overset{2}{\lambda}) + \psi(-11\, D^2\overset{2}{\lambda} + 22\overset{2}{\lambda})\overset{0}{q}_{ab}\,,] = -22\psi D_a(D^2\overset{2}{\lambda}), \tag{B14}$$

$$D^b(4\psi\overset{0}{T}_{ab} - 4\psi\overset{0}{T}\overset{0}{q}_{ab}) = 4(\overset{0}{T}_{ab}\overset{0}{E}{}^b - \overset{0}{T}\overset{0}{E}_a) + 4\psi(D^b\overset{0}{T}_{ab} - D_a\overset{0}{T}) \tag{B15}$$

(themselves consequences of the remnant field equations [(50)–(56), (60)–(63), (C12), (C13)], and, in the second step, $\overset{0}{B}_a = 0$ and Eq. (69) of the theorem.

APPENDIX C: GRAVITATIONAL REMNANT EQUATIONS

In Appendix C 1 we discuss the issue of generalizing the remnant radiation multipoles from linearized to full gravitation. In Appendix C 2, we outline the derivation of gravitational remnant field equations. In Appendix C 3 we present an alternative version of the second-order gravitational remnant field equations, involving the remnants of the Weyl tensor.

1. Generalization of gravitational remnant radiation monopole

In Sec. IV, we generalized the flat-space Klein–Gordon and Maxwell remnant radiation monopoles to curved space–time. However, we have been unable to determine whether there exists a similar generalization for linearized gravity. Here is how far we have gotten.

The remnant equations for gravitation were given, up to second order, in (60)–(63). We shall need the next two orders. The third-order remnant equations are

$$\overset{3}{q} = \overset{2}{q}{}^{ab}D_a\, D_b\overset{1}{\lambda} + 2D^a\, D^b\overset{1}{\lambda}\, D_a\overset{1}{\lambda}\, D_b\overset{1}{\lambda} + 12\overset{1}{\lambda}(D\overset{1}{\lambda})^2 - 24\overset{1}{\lambda}{}^3, \tag{C1}$$

$$D^b\overset{3}{q}_{ab} = D_a\overset{3}{q} + 2\overset{2}{q}_{ab}\, D^b\overset{1}{\lambda} + 4(D\overset{1}{\lambda})^2\, D_a\overset{1}{\lambda} + 64\overset{1}{\lambda}{}^2\, D_a\overset{1}{\lambda}, \tag{C2}$$

$$
\begin{aligned}
D^2\overset{3}{q}_{ab}={}&-\overset{3}{q}_{ab}+D_a D_b\overset{3}{q}-12\,D\overset{1}{\lambda}\cdot D\overset{2}{q}_{ab}+24\overset{1}{\lambda}\overset{2}{q}_{ab}+16D_{(a}(\overset{2}{q}_{b)m}D^m\overset{1}{\lambda})+16D_{(a}\overset{1}{\lambda}\,D_{b)}D_c\overset{1}{\lambda}\,D^c\overset{1}{\lambda}\\
&-4(D\overset{1}{\lambda})^2D_aD_b\overset{1}{\lambda}+160\overset{1}{\lambda}\,D_a\overset{1}{\lambda}\,D_b\overset{1}{\lambda}-52\overset{1}{\lambda}^2D_aD_b\overset{1}{\lambda}-[5\overset{3}{q}-12\overset{1}{\lambda}(D\overset{1}{\lambda})^2+372\overset{1}{\lambda}^3]\overset{0}{q}_{ab}.
\end{aligned}
\tag{C3}
$$

The fourth-order remnant equations are

$$
\begin{aligned}
\overset{4}{q}={}&\tfrac{2}{3}\overset{3}{q}{}^{ab}D_aD_b\overset{1}{\lambda}+\tfrac{1}{3}D\overset{1}{\lambda}\cdot D\overset{3}{q}+\tfrac{10}{3}\overset{2}{q}{}^{ab}D_a\overset{1}{\lambda}\,D_b\overset{1}{\lambda}+5\overset{1}{\lambda}\overset{2}{q}{}^{ab}D_aD_b\overset{1}{\lambda}+2\overset{2}{q}{}^{ab}\overset{2}{q}_{ab}\\
&-\tfrac{4}{3}(D\overset{1}{\lambda})^2(D\overset{1}{\lambda})^2+\tfrac{92}{3}\overset{1}{\lambda}^2(D\overset{1}{\lambda})^2+10\overset{1}{\lambda}D^aD^b\overset{1}{\lambda}\,D_a\overset{1}{\lambda}\,D_b\overset{1}{\lambda}-192\overset{1}{\lambda}^4,
\end{aligned}
\tag{C4}
$$

$$
\begin{aligned}
D^b\overset{4}{q}_{ab}={}&D_a\overset{4}{q}+4\overset{3}{q}_{ab}D^b\overset{1}{\lambda}+2\overset{3}{q}\,D_a\overset{1}{\lambda}+3D^b(\overset{2}{q}_{ac}\overset{2}{q}{}^c{}_b)-\tfrac{9}{4}D_a(\overset{2}{q}_{bc}\overset{2}{q}{}^{bc})\\
&-\tfrac{3}{2}\overset{2}{q}_{ab}D^b\overset{2}{q}-10\overset{1}{\lambda}\overset{2}{q}_{ab}D^b\overset{1}{\lambda}+76\overset{1}{\lambda}(D\overset{1}{\lambda})^2D_a\overset{1}{\lambda}+832\overset{1}{\lambda}^3D_a\overset{1}{\lambda},
\end{aligned}
\tag{C5}
$$

$$
\begin{aligned}
D^2\overset{4}{q}_{ab}={}&-6\overset{4}{q}_{ab}+2D_{(a}D^m\overset{4}{q}_{b)m}+2\overset{4}{q}\overset{0}{q}_{ab}+16\overset{1}{\lambda}D_{(a}D^m\overset{3}{q}_{b)m}+16D_{(a}\overset{3}{q}_{b)m}D^m\overset{1}{\lambda}+72\overset{1}{\lambda}\overset{3}{q}_{ab}-20\overset{1}{\lambda}\overset{3}{q}\overset{0}{q}_{ab}\\
&-8\lambda\,D^2\overset{3}{q}_{ab}-16D\overset{1}{\lambda}\cdot D\overset{3}{q}_{ab}+96\overset{1}{\lambda}^2D_{(a}D^m\overset{2}{q}_{b)m}+96\overset{1}{\lambda}^2\overset{2}{q}_{ab}-96\overset{1}{\lambda}^2\overset{2}{q}\overset{0}{q}_{ab}\\
&+120\overset{1}{\lambda}\,D_{(a}\overset{2}{q}_{b)m}D^m\overset{1}{\lambda}-48\overset{1}{\lambda}^2D^2\overset{2}{q}_{ab}-156\overset{1}{\lambda}\,D\overset{1}{\lambda}\cdot D\overset{2}{q}_{ab}+16D^cD_{(a}\overset{1}{\lambda}\overset{3}{q}_{b)c}+16\,D^c\overset{3}{q}_{c(a}D_{b)}\overset{1}{\lambda}\\
&-8\overset{3}{q}{}^{cd}D_cD_d\overset{1}{\lambda}\,\overset{0}{q}_{ab}-8\,D^c\overset{3}{q}_{cd}D^d\overset{1}{\lambda}\,\overset{0}{q}_{ab}-12\,D_c\overset{2}{q}{}^{cd}D_{(a}\overset{2}{q}_{b)d}-12\overset{2}{q}{}^{cd}D_cD_{(a}\overset{2}{q}_{b)d}\\
&+6\,D_c\overset{2}{q}{}^{cd}D_d\overset{2}{q}_{ab}+6\overset{2}{q}{}^{cd}D_cD_d\overset{2}{q}_{ab}+192\overset{1}{\lambda}\,D_{(a}\overset{1}{\lambda}\,D^c\overset{2}{q}_{b)c}+192\overset{1}{\lambda}D^cD_{(a}\overset{1}{\lambda}\overset{2}{q}_{b)c}\\
&+192\,D_{(a}\overset{1}{\lambda}\,\overset{2}{q}_{b)c}D^c\overset{1}{\lambda}-96\overset{1}{\lambda}\,D_c\overset{2}{q}{}^{cd}D_d\overset{1}{\lambda}\,\overset{0}{q}_{ab}-96\overset{1}{\lambda}\overset{2}{q}{}^{cd}D_cD_d\overset{1}{\lambda}\,\overset{0}{q}_{ab}-96\overset{2}{q}{}^{cd}D_c\overset{1}{\lambda}\,D_d\overset{1}{\lambda}\,\overset{0}{q}_{ab}\\
&+432\overset{1}{\lambda}^3D_aD_b\overset{1}{\lambda}+1440\overset{1}{\lambda}^2D_a\overset{1}{\lambda}\,D_b\overset{1}{\lambda}+1488\overset{1}{\lambda}^2(D\overset{1}{\lambda})^2\overset{0}{q}_{ab}+1296\overset{1}{\lambda}^4\overset{0}{q}_{ab}+48\overset{1}{\lambda}\,D\overset{1}{\lambda}\cdot D\overset{2}{q}\,\overset{0}{q}_{ab}\\
&-3\,D\overset{2}{q}\cdot D\overset{2}{q}_{ab}+6\,D_{(a}\overset{2}{q}_{b)c}D^c\overset{2}{q}+4\,D\overset{1}{\lambda}\cdot D\overset{3}{q}\,\overset{0}{q}_{ab}+48(D\overset{1}{\lambda})^2(D\overset{1}{\lambda})^2\overset{0}{q}_{ab}+24\overset{2}{q}_{ac}\overset{2}{q}{}^c{}_b\\
&-12\overset{2}{q}_{cd}\overset{2}{q}{}^{cd}\overset{0}{q}_{ab}-48(D\overset{1}{\lambda})^2\overset{2}{q}_{ab}-6\,D_c\overset{2}{q}_{d(a}D^d\overset{2}{q}_{b)}{}^c-3\,D_a\overset{2}{q}_{cd}D_b\overset{2}{q}{}^{cd}+6\,D_c\overset{2}{q}_{da}D^c\overset{2}{q}_b{}^d\\
&-D_cD_b\overset{4}{q}-8D_aD_b(\overset{1}{\lambda}\overset{3}{q})+3D_aD_b(\overset{2}{q}{}^{cd}\overset{2}{q}_{cd})-48D_aD_b(\overset{1}{\lambda}^2\overset{2}{q}).
\end{aligned}
\tag{C6}
$$

We begin by noting that, introducing a potential h_{ab} for the linearized gravitational field, the integrand of $\mathcal{G}_{\mu\nu\lambda\sigma}$ in flat space–time in Eq. (41) is a multiple of $D_{(a}\overset{4}{h}_{bc)}\xi^b{}_{(\mu|(\lambda}\xi^c{}_{\sigma)|\nu)}$, where we have set $\xi^a{}_{\mu\nu}=\alpha_{[\mu}D^a\alpha_{\nu]}$. Note also that $D_{(a}\overset{4}{h}_{bc)}$ is divergence-free, by virtue of the remnant field equations on $\overset{k}{h}_{ab}$. Thus, the problem of generalizing to curved space–time the $\mathcal{G}_{\mu\nu\lambda\sigma}$ of flat space–time is equivalent to that of finding a third-rank, totally symmetric, divergence-free tensor field s_{abc} on $\mathcal{H}$, constructed from the gravitational remnants, such that s_{abc} reduces, in the case of linearized gravity, to $D_{(a}\overset{4}{h}_{bc)}$. Consider, in this connection, the candidate $\hat{s}_{abc}$ given by

$$\begin{aligned}\hat{s}_{abc}=&D_{(a}\overset{4}{q}_{bc)}+(\tfrac{82}{3}+4c)\lambda_{(a}\overset{3}{q}_{bc)}-(24+4c)\overset{0}{q}_{(ab}\overset{3}{q}_{c)d}\lambda^{d}+(\tfrac{20}{3}+2c)\overset{1}{\lambda}\,D_{(a}\overset{3}{q}_{bc)}\\&-\left(\tfrac{10}{3}+c\right)\lambda^{d}{}_{(ab}\overset{3}{q}_{c)d}+\left(\tfrac{2}{3}+\tfrac{c}{2}\right)\overset{0}{q}_{(ab}\lambda_{c)de}\overset{3}{q}^{de}+\tfrac{4}{3}\lambda^{d}{}_{(a}D_{b}\overset{3}{q}_{c)d}-\left(\tfrac{10}{3}+\tfrac{c}{2}\right)\overset{0}{q}_{(ab}D_{c)}\overset{3}{q}_{de}\lambda^{de}\\&+\tfrac{8}{3}\overset{0}{q}_{(ab}D^{d}\overset{3}{q}^{e}{}_{c)}\lambda_{de}+(2+c)\lambda^{d}D_{(a}D_{b}\overset{3}{q}_{c)d}-(\tfrac{8}{3}+c)\lambda^{d}D_{d}D_{(a}\overset{3}{q}_{bc)}+c\lambda_{d(a}D^{d}\overset{3}{q}_{bc)},\end{aligned}\tag{C7}$$

where c is any constant, and where we have set $\lambda_a\equiv D_a\overset{1}{\lambda}$, $\lambda_{ab}\equiv D_a D_b\overset{1}{\lambda}+\overset{1}{\lambda}\overset{0}{q}_{ab}$, and $\lambda_{abc}\equiv D_a\lambda_{bc}$. This $\hat{s}_{abc}$ has all the required properties, except that its divergence, instead of vanishing, includes remnants of order not exceeding 2. The issue, then, is whether one can add to this $\hat{s}_{abc}$ terms of order not exceeding two to achieve vanishing divergence. In any case, the mere existence of this field $\hat{s}_{abc}$ lends support to the conjecture that $\mathcal{G}_{\mu\nu\lambda\sigma}$ admits a generalization to curved space–time. Work is in progress to settle this conjecture.

2. Derivation of gravitational remnant field equations

The Einstein equation gives rise to certain differential equations on the gravitational remnants, $\overset{k}{\lambda}_{ab}, \overset{k}{q}_{ab}$. These equations for a vacuum space–time were first systematically studied by Beig and Schmidt.[8,9] We have here utilized the nonvacuum equations, of order one (59)–(60), two (61)–(63), and vacuum equations of order three (C1)–(C3), and four (C4)–(C6). We summarize how these were derived. First write Einstein's equation in a $3+1$ form, adapted to the surfaces $\Omega=\text{const}$:

$$\Omega^2 T=-\tfrac{1}{2}[\mathcal{R}+p^{mn}p_{mn}-p^2],\tag{C8}$$

$$\Omega^2 T_a=D^m(p_{am}-pq_{am}),\tag{C9}$$

$$\Omega^2 T_{ab}=\mathcal{R}_{ab}+2p_a{}^{m}p_{mb}-pp_{ab}-\lambda^{-1}D_a D_b\lambda+\lambda^{-1}p_{ab}-\Omega\mathcal{L}_{\lambda n}p_{ab},\tag{C10}$$

where D_a denotes the derivative operator of the metric q_{ab} of these surfaces, $\mathcal{R}_{ab}$ its Ricci curvature, and p_{ab} the rescaled extrinsic curvature of these surfaces, defined by

$$p_{ab}\equiv\Omega q_a{}^{k}q_b{}^{l}\tilde{\nabla}_k(\Omega^{-2}\lambda\,\tilde{\nabla}_l\Omega)=-\lambda^{-1}q_{ab}+\tfrac{1}{2}\Omega\mathcal{L}_{\lambda n}q_{ab}.\tag{C11}$$

Taking the remnants of Eqs. (C8)–(C10) through fourth order, we obtain Eqs. (59)–(60), (60)–(63), (C1)–(C3), and (C4)–(C6).

We remark, finally, that the conservation equation of the stress–energy tensor, $\tilde{\nabla}^a\tilde{T}_{ab}=0$, yields, for the zeroth-order remnants of $\tilde{T}_{ab}$, the following equations:

$$0=D^a\overset{0}{T}_a+2\overset{0}{T}+2\overset{0}{T}{}^{m}{}_{m},\tag{C12}$$

$$0=D^b\overset{0}{T}_{ab}-D_a(\overset{0}{T}+\overset{0}{T}{}^{m}{}_{m}).\tag{C13}$$

3. Second-order equations in Weyl remnants

We first remark that, for any space–time with completion, the Weyl tensor is asymptotically regular of order 3. To see this, rewrite $2\tilde{\nabla}_{[a}\tilde{\nabla}_{b]}n_c=\tilde{R}_{abc}{}^d n_d$ as

$$E_{ab}=\Omega^{-1}(-\overset{q}{\mathcal{R}}_{ab}-p_a{}^m p_{bm}+pp_{ab})+\Omega[\tfrac{1}{2}(T_{ab}-\tfrac{1}{3}q_{ab}T^m{}_m)-\tfrac{2}{3}q_{ab}T], \tag{C14}$$

$$B_{ab}=\Omega^{-1}\epsilon_{mn(a}\overset{q}{D}{}^m p^n{}_{b)}, \tag{C15}$$

with E_{ab} and B_{ab} given by Eq. (3), p_{ab} given by Eq. (C11), and T_{ab} given by Eq. (2). But, by the conditions in Definition 1, the right sides are smooth on M.

We next remark that the gravitational remnant equations, (60)–(63), can be written in terms of the Weyl remnants, $\overset{k}{E}_{ab}, \overset{k}{B}_{ab}$. To see this, first take the zeroth-order remnants of Eqs. (C14), (C15) above, to obtain

$$\overset{0}{E}_{ab}=-(D_a D_b\overset{1}{\lambda}+\overset{1}{\lambda}\overset{0}{q}_{ab}), \tag{C16}$$

$$\overset{0}{B}_{ab}=\epsilon_{kl(a}D^k(\overset{1}{q}{}^l{}_{b)}+2\overset{1}{\lambda}\overset{0}{q}{}^l{}_{b)})=0, \tag{C17}$$

and the first-order remnants, to obtain

$$\overset{1}{E}_{ab}=-\tfrac{1}{2}\overset{2}{q}_{ab}+[(D\overset{1}{\lambda})^2+5\overset{1}{\lambda}{}^2]\overset{0}{q}_{ab}+\overset{1}{\lambda}D_a D_b\overset{1}{\lambda}-2D_a\overset{1}{\lambda}\, D_b\overset{1}{\lambda}-\tfrac{1}{2}\overset{0}{T}_{ab}-(\tfrac{2}{3}\overset{0}{T}+\tfrac{1}{6}\overset{0}{T}{}^m{}_m)\overset{0}{q}_{ab}, \tag{C18}$$

$$\overset{1}{B}_{ab}=\tfrac{1}{2}\epsilon_{mn(a}D^m\overset{2}{q}{}^n{}_{b)}. \tag{C19}$$

These Weyl remnants satisfy, by virtue of Eqs. (60)–(63), the equations

$$D_{[a}\overset{0}{E}_{b]c}=0, \tag{C20}$$

$$D_{[a}\overset{1}{\mathcal{E}}_{b]c}=\tfrac{1}{2}\epsilon_{ab}{}^m[\overset{1}{B}_{mc}+4\epsilon_{kl(m}(D^k\overset{1}{\lambda})\overset{0}{E}{}^l{}_{c)}+\tfrac{1}{2}\epsilon_{mc}{}^n\overset{0}{T}_n], \tag{C21}$$

$$D_{[a}\overset{1}{B}_{b]c}=-\tfrac{1}{2}\epsilon_{ab}{}^m[\overset{1}{\mathcal{E}}_{mc}-2\overset{0}{T}_{mc}-\overset{0}{T}\overset{0}{q}_{mc}+D_c\overset{0}{T}_m], \tag{C22}$$

where we have set

$$\overset{1}{\mathcal{E}}_{ab}=\overset{1}{E}_{ab}-\overset{1}{\lambda}\overset{0}{E}_{ab}+\tfrac{1}{2}\overset{0}{T}_{ab}+(\tfrac{1}{6}\overset{0}{T}{}^m{}_m-\tfrac{2}{3}\overset{0}{T})\overset{0}{q}_{ab}. \tag{C23}$$

Now fix a space–time with completion, and define $\overset{0}{E}_{ab}$, $\overset{1}{E}_{ab}$, and $\overset{1}{B}_{ab}$ by Eqs. (C16)–(C19). Then Eqs. (60)–(63) are equivalent to the statements that the $\overset{0}{E}_{ab}$, $\overset{1}{E}_{ab}$, $\overset{1}{B}_{ab}$, so defined are trace-free and satisfy (C20)–(C22).

[1] R. Arnowitt, S. Deser, and C. W. Misner, Phys. Rev. **117**, 1695 (1960); *ibid.* **121**, 1566 (1961); *ibid.* **122**, 997 (1961); *Gravitation, an Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).

[2] R. Geroch, J. Math. Phys. **13**, 956 (1972).
[3] R. Geroch, J. Math. Phys. **11**, 2580 (1970).
[4] A. Ashtekar and R. O. Hansen, J. Math. Phys. **19**, 1542 (1978).
[5] A. Ashtekar and A. Magnon-Ashtekar, J. Math. Phys. **20**, 793 (1979).
[6] P. Sommers, J. Math. Phys. **19**, 549 (1978).
[7] S. Persides, J. Math. Phys. **20**, 1731 (1979); *ibid.* **21**, 135 (1980); *ibid.* **21**, 142 (1980).
[8] R. Beig and B. G. Schmidt, Commun. Math. Phys. **87**, 65 (1982).
[9] R. Beig, Proc. R. Soc. London, Ser. A **391**, 295 (1984).
[10] H. Bondi, A. W. K. Metzner, and M. J. G. Van Der Berg, Proc. R. Soc. London, Ser. A **269**, 21 (1962).
[11] A. Ashtekar and J. D. Romano, Class. Quantum Grav. **9**, 1069 (1992).
[12] Sketch of proof: Let $(\tilde{M},\tilde{g}_{ab})$ be a stationary asymptotically flat vacuum space–time with Killing field ξ^a. Denote by μ and ω the norm and twist of the Killing field, respectively. Let $(\tilde{V},\tilde{h}_{ab})$ denote the (Riemannian) manifold of orbits of the Killing field, (V,Λ) its completion, Ω_G a conformal factor, and $h_{ab}=\Omega_G^2\tilde{h}_{ab}$ (see Ref. 3). It follows[37] that each of $\Omega_G^{-1/2}\mu^{1/4}(\mu-\mu^{-1}+\mu^{-1}\omega^2)$, $\Omega_G^{-1/2}\mu^{1/4}(\mu^{-1}\omega)$, and $\mu+\mu^{-1}+\mu^{-1}\omega^2$ is smooth on V. Fix any smooth coordinates x^i on V near Λ such that $h_{ij}|_\Lambda=\delta_{ij}$, and Lie drag them into $\tilde{M}$ by ξ^a. Perform an inversion on these coordinates to obtain $\tilde{x}^i$ on $\tilde{M}$. Pick a τ'_a on V satisfying $D_{[a}\tau'_{b]}=-\frac{1}{2}\mu^{-3/2}\epsilon_{abc}D^c\omega$ and such that τ'_a is smooth in y' and vanishes at Λ (See, e.g., the appendix of Ref. 37 for motivation). Let τ_a be the pull-back of τ'_a to $\tilde{M}$. Define $\tilde{x}^0$ on $\tilde{M}$ such that $\boldsymbol{\nabla}_a\tilde{x}^0=\mu^{-1}\xi_a-\tau_a$ (note the right side is curl-free and yields 1 when contracted with $\tilde{\xi}^a$). Then the hyperbolic coordinates associated with the $\tilde{x}^\mu$ coordinates yield a completion of $\tilde{M}$ in the sense above.
[13] Such freedom in choices of (inequivalent) completion are known to exist also for other frameworks such as that of Geroch and that of Ashtekar-Hansen.
[14] P. G. Bergmann, Phys. Rev. **124**, 274 (1961).
[15] P. T. Chrusciel, J. Math. Phys. **30**, 2094 (1989).
[16] More generally, for a spin-s field, we would demand asymptotic regularity of order $s+1$.
[17] In essence, one may view α_μ as a choice of an orthonormal basis of $\mathcal{T}$, and $v=v^\mu\alpha_\mu$ as the expansion of an element $v\in\mathcal{T}$ in the basis α_μ with constant coefficients v^μ. Our presentation serves the same purpose without committing ourselves to a particular choice of basis. Thus, the α_μ might be better viewed as an ''abstract orthonormal basis.''
[18] To see this, evaluate $D_{[a}D_{b]}(\eta^{\mu\nu}D_c\alpha_\mu D_d\alpha_\nu)$ using Eq. (8) and equate the result to $2R_{ab(c}{}^m(\eta^{\mu\nu}D_{d)}\alpha_\mu D_m\alpha_\nu)$, to obtain $\eta^{\mu\nu}D_a\alpha_\mu D_b\alpha_\nu=\eta^{\mu\nu}\alpha_\mu\alpha_\nu \overset{0}{q}_{ab}$. Now contract with $\overset{0}{q}{}^{ab}$ using Eq. (9).
[19] One might be tempted to consider, in addition, those divergence-free vector fields that are multilinear in the Killing fields on $\mathcal{H}$. However, this adds nothing new, since every antisymmetric second rank tensor $F^{\mu\nu}$ in $\mathcal{T}$ yields a Killing field in $\mathcal{H}$ when contracted with $\alpha_{[\mu}D^a\alpha_{\nu]}$ and, conversely, for every Killing field ξ^a in $\mathcal{H}$, there exists an antisymmetric second rank tensor over $\mathcal{T}$ (namely, $F_{\mu\nu}\equiv 2\xi_a\alpha_{[\mu}D^a\alpha_{\nu]}+D_a\xi_b D^a\alpha_\mu D^b\alpha_\nu$) that gives rise to it. Similarly, multilinearity in conformal Killing fields yields nothing new, for every vector v^μ in $\mathcal{T}$ yields a curl-free conformal Killing field in $\mathcal{H}$ when contracted with $D^a\alpha_\mu$ and conversely, for every curl-free conformal Killing field ζ^a in $\mathcal{H}$, there exists a vector over $\mathcal{T}$ [namely, $v^\mu=\zeta^a D_a\alpha^\mu-\frac{1}{3}(D_a\zeta^a)\alpha^\mu$] that gives rise to it.
[20] In Ref. 11, Ashtekar and Romano used instead the condition $\lim_{\Omega\to 0}\Omega^{-1}\tilde{G}_{ab}=0$ to show that the angular momentum is conserved. As we have noted earlier, their condition is too strong. The condition we are imposing is the necessary and sufficient condition for $\overset{1}{B}_{ab}$ to be divergence-free on $\mathcal{H}$ [cf. Eq. (C22)]. An example of a space–time satisfying our additional condition is the Kerr–Newman solution. In fact, the Kerr–Newman solution satisfies a stronger condition: $\overset{0}{T}_a=0$. In general, it is not clear how restrictive is the condition given by Eq. (13). However, the condition is presumably satisfied for all stationary asymptotically flat space–times since in that case one expects the angular momentum is well defined and equal to Hansen's angular-momentum dipole moment.[38]
[21] To see this, note that $\mathcal{M}'_{\mu\nu}-\mathcal{M}_{\mu\nu}=-(1/16\pi)\epsilon_{\mu\nu}{}^{\tau\sigma}\int_C\{(-2\epsilon^{mn(a}\overset{0}{E}{}^{b)}{}_m D_n\overset{0}{\omega})\alpha_\tau D_b\alpha_\sigma\}dS_a=-(1/8\pi)\overset{0}{\omega}_{[\mu}\int_C\overset{0}{E}{}^{ab}D_{|b|}\alpha_{\nu]}\,dS_a=-\overset{0}{\omega}_{[\mu}\mathcal{P}_{\nu]}$.
[22] The linearity is clear for electric charge. For total energy–momentum and angular momentum, we have in mind the linearized gravity in Minkowski space–time in which these quantities are linear in the gravitational field and are expressible as surface integrals.[39]
[23] That is, $\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})\equiv\Sigma_{m=0}^{[n/2]}(-1/4)^m\binom{n-m-1}{m}\eta_{(\mu_1\mu_2}\cdots\eta_{\mu_{2m-1}\mu_{2m}}\alpha_{\mu_{2m+1}}\cdots\alpha_{\mu_{n-1})}$, with $[n/2]$ denoting the largest integer not exceeding $n/2$.
[24] Indeed, we have $D^a(\overset{n}{E}{}^m D_m\alpha_{(\mu})\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1})})-\overset{n}{E}{}^m D_m\alpha_{(\mu}D^a\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1})})=D_b\{2\overset{n}{E}{}^{[a}D^{b]}\alpha_{(\mu}\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1})})+\epsilon^{abc}\overset{n}{B}_c[\alpha_{(\mu}\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1})})-(1/2)\eta_{(\mu\mu_1}\mathcal{C}(\alpha_{\mu_2}\cdots\alpha_{\mu_{n-1})})]\}$, which can be seen by using Eq. (B7) and the identity $\alpha_{(\mu}D^a\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1})})=(n-1)\mathcal{C}(\alpha_{(\mu_1}\cdots\alpha_{\mu_{n-1}})D^a\alpha_{\mu)}+(1/2)\eta_{(\mu\mu_1}D^a\mathcal{C}(\alpha_{\mu_2}\cdots\alpha_{\mu_{n-1})})$.
[25] To see this, note that the difference between the integrands of $*\mathcal{E}_{\mu\mu_1\cdots\mu_{n-1}}$ and that of the right of Eq. (39) is given by $D_k\{(2/n)D_m\alpha_\mu\,\epsilon^{lm[a}[(D^{k]}\overset{n}{E}_l)\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})-\overset{n}{E}_l D^{k]}\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})]\}$. Integrate over a cut of $\mathcal{H}$.

[26]To see this, note that the difference between the integrands of $*\mathcal{G}_{\mu\nu\mu_1\cdots\mu_{n-1}}$ and the right side of Eq. (44) is given by $D_k\{(1/n)\epsilon^{akl}\overset{n}{E}_{lm}\alpha_{(\mu}D^m\alpha_{\nu)}\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})+(2/n)D_m\alpha_{(\mu}D^s\alpha_{\nu)}\epsilon^{lm[a}[(D^{k]}\overset{n}{E}_{ls})\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})-\overset{n}{E}_{ls}D^{k]}\mathcal{C}$ $\times(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})]\}$. Integrate over a cut of $\mathcal{H}$.

[27]Indeed, we have the following. Let $\overset{n-2}{\phi}$ satisfy the $n-2$th remnant equation for a Klein–Gordon field, ξ^a any Killing field. Denote by $\overset{n-2}{\psi}_{a_1\cdots a_s}$ the symmetric and trace-free part of $D_{a_1}\cdots D_{a_s}\overset{n-2}{\phi}$. Then $\overset{n}{E}_{ab}\equiv\overset{n-2}{\psi}_{abcd}\xi^c\xi^{*d}+\frac{12}{7}(n+1)$ $\times(n+2)[\xi^{*m}\overset{n-2}{\psi}_{m(a}\xi_{b)}+\xi^m\overset{n-2}{\psi}_{m(a}\xi^*_{b)}-\frac{2}{3}\overset{0}{q}_{ab}\overset{n-2}{\psi}_{cd}\xi^c\xi^{*d}]-\frac{4}{5}n(n+1)(n+2)\overset{n-2}{\psi}{}^m D_m(\xi_{(a}\xi^*_{b)})+\frac{4}{5}n(n-1)(n+1)(n+2)$ $\times\overset{n-2}{\psi}\xi_{(a}\xi^*_{b)}+(n+2)\overset{n-2}{\psi}_{cd(a}D_{b)}(\xi^{(c}\xi^{*d)})$, satisfies the nth remnant equations for a linearized gravitational field.

[28]Adding to ψ a constant changes the integrand of $\mathcal{E}_{\mu\nu}$ in Eq. (70) by a divergence of an antisymmetric tensor.

[29]Let $\tilde{\phi}$ be a Klein–Gordon field in Minkowski space–time asymptotically regular of order 1. Consider $I=\int_{S_\infty}$ $-\tilde{\epsilon}_{abcd}x^c\tilde{\nabla}^d[(x^e\tilde{\nabla}_e+1)\tilde{\phi}]$, where S_∞ denotes a two-sphere at infinity, and x^a a position vector field. When S_∞ is any two-sphere cut at spatial infinity, the above integral reproduces the remnant radiation monopole $\mathcal{K}$ associated with $\tilde{\phi}$. However, in general, the integral evaluates to a different value when the two-sphere S_∞ is at null infinity. For example, let $\tilde{\phi}=(f(t+r)-f(t-r))/r$, with $k_\pm(x)\equiv f(\pm 1/x)$, $x>0$, both smoothly extendible to zero. Then, when S_∞ is at spatial infinity, we have $I=\mathcal{K}=4\pi[k'_+(0)+k'_-(0)]$, while for S_∞ any cut at future null infinity, $I=4\pi k'_+(0)$, and for S_∞ any cut at past null infinity, $I=4\pi k'_-(0)$.

[30] Newman and Penrose, Proc. R. Soc. London Ser. A **305**, 175 (1968).

[31]A. R. Exton, E. T. Newman, and R. Penrose, J. Math. Phys. **10**, 1566 (1969).

[32]The analogous equation for $\overset{n}{B}_{ab}$,

$$\mathcal{L}_{D\zeta}\overset{n}{B}_{ab}-(n+1)\zeta\overset{n}{B}_{ab}-2\epsilon^{lm}{}_{(a}\overset{n}{E}_{b)}D_m\zeta=0,$$

follows from Eq. (A2) and the remnant field equations. For a Maxwell field the corresponding equation is $\mathcal{L}_{D\zeta}\overset{n}{E}_a-(n+1)\zeta\overset{n}{E}_a+\epsilon_a{}^{kl}\overset{n}{B}_k D_l\zeta=0$, and for a Klein–Gordon field, Eq. (A3).

[33]We remark that the multipole moment M defined here is related to the Q defined by Geroch in Ref. 40 by a normalization factor: $Q_{i_1\cdots i_n}=(-\frac{1}{3})^n n!M_{i_1\cdots i_n}$.

[34]We are concerned only with ''irreducible'' solutions. Thus, for example, $\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$ and $\eta_{\mu\nu}\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$ are viewed as equivalent solutions to Eq. (19).

[35]To see this, embed $\mathcal{H}$ as the unit hyperboloid in Minkowski space–time M'. Let x^a denote the position vector field from some origin. Then $\nabla_a x^b=\delta_a{}^b$ and $\mathcal{H}$ is specified by $x^a x_a=1$. Let k^a be a constant vector field in M'. Then $k^a x_a$ is a translation on $\mathcal{H}$. Thus, the most general function multilinear in translations is a sum of terms of the form $w(s)$ $\equiv w^{a_1\cdots a_s}x_{a_1}\cdots x_{a_s}$, with $w_{a_1\cdots a_s}$ some symmetric, trace-free constant tensor. This $w(s)$ satisfies the Klein–Gordon equation in M'. Using $\nabla^2 w=[D^2+(x\cdot x)^{-1}((x\cdot\nabla)^2+2x\cdot\nabla)]w$, we see that $w(n-1)$ satisfies Eq. (19) on $\mathcal{H}$. Such $w(n-1)$'s clearly exhaust all solutions of Eq. (19), which are multilinear in translations. But each such $w(n-1)$'s on $\mathcal{H}$ is the contraction of $\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$ with some tensor over $\mathcal{T}$.

[36]One way to prove the identity is to note the following: $D^a\alpha_\mu D_a\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})+(n-1)\alpha_\mu\mathcal{C}(\alpha_{\mu_1}\cdots\alpha_{\mu_{n-1}})$ satisfies Eq. (19) with n replaced by $n-1$ and is trace-free in $\mu_1\cdots\mu_{n-1}$. The overall normalization factor can be fixed by comparing, say, the coefficients of the term $\eta_{(\mu\mu_1}\mathcal{C}(\alpha_{\mu_2}\cdots\alpha_{\mu_{n-1})})$ on both sides.

[37]R. Geroch, J. Math. Phys. **12**, 918 (1970).

[38]R. O. Hansen, J. Math. Phys. **15**, 46 (1974).

[39]R. Penrose, Proc. R. Soc. London, Ser. A **381**, 53 (1982).

[40]R. Geroch, J. Math. Phys. **11**, 1955 (1970).

Group invariant solutions for the $N=2$ super Korteweg–de Vries equation

M. A. Ayari,[a] V. Hussin,[b] and P. Winternitz[c]
Centre de Recherches Mathématiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montreal, Quebec H3C 3J7, Canada

(Received 3 August 1998; accepted for publication 9 November 1998)

The method of symmetry reduction is used to solve Grassmann-valued differential equations. The $(N=2)$ supersymmetric Korteweg–de Vries equation is considered. It admits a Lie superalgebra of symmetries of dimension 5. A two-dimensional subsuperalgebra is chosen to reduce the number of independent variables in this equation. We are then able to give different types of exact solutions, in particular soliton solutions.
[S0022-2488(99)02903-5]

I. INTRODUCTION

The study of integrability and conservation laws for systems of Grassmann-valued differential equations (SGVDEs) has known a wide expansion in the area of supersymmetry.[1–13] Many of these equations have been constructed[2–6] in order to combine bosonic and fermionic degrees of freedom in such a way that these equations are invariant under supersymmetry transformations. This means that in these cases, there exists a symmetry relating bosonic and fermionic fields.

Many authors have proven integrability by finding Lax pairs and conservation laws. For example, Yung[12] has studied the supersymmetric Boussinesq hierarchies and his results were confirmed later by Bellucci *et al.*[13] Mathieu[3,4] has investigated the supersymmetric (SUSY) Korteweg–de Vries (KdV) equation for $N=1,2$ odd independent variables and has found that integrability occurs for special values of a parameter figuring in the superequations.

These SUSY KdV equations are the starting point of our approach and more particularly the case[4] $N=2$. These equations contain both the KdV and modified KdV equations in the limit where odd Grassmann dependent variables are set equal to zero. Moreover, the supersymmetry group is richer in the $N=2$ case than for $N=1$ and it contains significant subgroups.[14] We will use the technique of symmetry reduction adapted to the super case to give some solutions of this $(N=2)$ SUSY KdV equation. This technique does not depend on the integrability of the equation and consists of a systematic application of group theory to reduce the SGVDE to a system of ordinary differential equations (ODEs).

The problem of computing solutions for SGVDEs has recently received a large amount of attention.[15–19] In these approaches, the SGVDEs are decomposed and give rise to systems of classical partial differential equations (PDEs) which can be solved. Our approach is based on symmetries and supersymmetries and does not require that we start with such a decomposition.

The Lie-point symmetries of a SGVDE have been obtained using an extension to Grassmann variables of the procedure described, e.g., by Olver.[20] The $N=2$ SUSY KdV admits a symmetry superalgebra with three even and two odd generators.[14] The equation, its invariance superalgebra, and the corresponding group will be given in Sec. II. The technique of symmetry reduction allows us to consider solutions which are invariant under subsuperalgebras with at least one supersymmetric (or odd) generator. In Sec. III, a subsuperalgebra of dimension 2 will be used to derive the

[a]Electronic mail: ayaria@dms.umontreal.ca
[b]Electronic mail: hussin@dms.umontreal.ca
[c]Electronic mail: wintern@crm.umontreal.ca

reduced equations and the solutions of the bosonic and fermionic part of the superfield. Some conclusions are drawn in Sec. IV.

II. THE SUPERSYMMETRY GROUPS OF THE $N=2$ SUSY KdV EQUATION

Let us recall that a SGVDE is a system of s partial differential equations of order $k=(k_1;k_2)$ of the form

$$\Delta_\nu(\mathbf{X},\Theta;\mathbf{A}^{(k_1)},\mathbf{\Gamma}^{(k_2)})=0, \quad \nu=1,...,s \tag{2.1}$$

with m independent even variables $\mathbf{X}=\{x_1,\ldots,x_m\}$, n independent odd variables $\mathbf{\Theta}=\{\theta_1,\ldots,\theta_n\}$, q dependent even variables $\mathbf{A}=\{A^1,A^2,\ldots,A^q\}$, and p dependent odd variables $\mathbf{\Gamma}=\{\Gamma^1,\Gamma^2,\ldots,\Gamma^p\}$. Note that odd variables η_i must satisfy

$$\eta_i\eta_j=-\eta_j\eta_i, \quad \eta_i^2=0, \quad 1\leqslant i,j\leqslant r.$$

The equation we are interested in is precisely of this type. It takes the form:[4]

$$A_t=-A_{xxx}+3(AD_1D_2A)_x+\frac{(a-1)}{2}(D_1D_2A^2)_x+3aA^2A_x, \tag{2.2}$$

where D_1,D_2 are the covariant superderivatives

$$D_i=\theta_i\partial_x+\partial_{\theta_i}, \quad i=1,2. \tag{2.3}$$

Equation (2.2) thus represents a one-parameter ($a\in\mathbb{R}$) family of Grassmann-valued partial differential equations having four (two even and two odd) independent variables $(x,t;\theta_1,\theta_2)$ and one dependent variable A which is supposed to be a bosonic (or even) superfield. It has been constructed by Mathieu *et al.*[4] as a nontrivial SUSY equation which contains both the KdV and the modified KdV as nonSUSY limits.

A more suitable form of Eq. (2.2) is given using partial derivatives, i.e.,

$$\begin{aligned}
&A_t+A_{xxx}-3a\,\theta_1\theta_2A_xA_{xx}-(a+2)\,\theta_1AA_{xx\theta_2}-(a+2)\{\theta_1\theta_2AA_{xxx}-\theta_2AA_{xx\theta_1}\}\\
&\quad+(2a+1)\,\theta_2A_xA_{x\theta_1}+(a+2)\{A_xA_{\theta_1\theta_2}+AA_{x\theta_1\theta_2}\}-(2a+1)\,\theta_1A_xA_{x\theta_2}\\
&\quad-(a-1)\{\theta_1A_{\theta_2}A_{xx}-\theta_2A_{\theta_1}A_{xx}+A_{\theta_1}A_{x\theta_2}-A_{\theta_2}A_{x\theta_1}\}-3aA^2A_x=0.
\end{aligned} \tag{2.4}$$

Since the superfield A is bosonic, it can be decomposed into

$$A(x,t;\theta_1,\theta_2)=u(x,t)+\theta_1\rho^1(x,t)+\theta_2\rho^2(x,t)+\theta_1\theta_2v(x,t), \tag{2.5}$$

where u and v are even functions of (x,t) while ρ^1 and ρ^2 are odd functions of the same independent even variables. It is then easy to see that the superequation (2.4) reduces to the following system of two bosonic and two fermionic equations:

$$\begin{gathered}
u_t+u_{xxx}-3au^2u_x+(a+2)(uv)_x-(a-1)(\rho^1\rho^2)_x=0,\\
v_t+v_{xxx}+6vv_x-3au_xu_{xx}-(a+2)uu_{xxx}-3\rho^2\rho^2_{xx}-(a+2)\rho^1\rho^1_{xx}\\
-3a(u^2v_x+2uu_xv-2u\rho^1\rho^2_x+2u\rho^2\rho^1_x-2u_x\rho^1\rho^2)=0,
\end{gathered} \tag{2.6}$$

and

TABLE I. The supercommutator table of the SUSY KdV equation ($N=2$).

	$\mathcal{P}_1$	$\mathcal{P}_0$	$\mathcal{D}$	$\mathcal{Q}_1$	$\mathcal{Q}_2$
$\mathcal{P}_1$	0	0	$\mathcal{P}_1$	0	0
$\mathcal{P}_0$	0	0	$3\mathcal{P}_0$	0	0
$\mathcal{D}$	$-\mathcal{P}_1$	$-3\mathcal{P}_0$	0	$(-1/2)\mathcal{Q}_1$	$(-1/2)\mathcal{Q}_2$
$\mathcal{Q}_1$	0	0	$(1/2)\mathcal{Q}_1$	$-2\mathcal{P}_1$	0
$\mathcal{Q}_2$	0	0	$(1/2)\mathcal{Q}_2$	0	$-2\mathcal{P}_1$

$$\begin{aligned}
&\rho_t^1+\rho_{xxx}^1+(a+2)(\rho^1 v)_x-(a-1)(\rho^1 v)_x-(a+2)u\rho_{xx}^2-(2a+1)u_x\rho_x^2\\
&\quad-(a-1)\rho^2 u_{xx}-3a(u^2\rho_x^1+2uu_x\rho^1)=0,\\
&\rho_t^2+\rho_{xxx}^2+(a+2)(\rho^2 v)_x+(a+2)u\rho_{xx}^1+(2a+1)u_x\rho_x^1+(a-1)\rho^1 u_{xx}\\
&\quad-(a-1)(\rho^2 v)_x-3a(u^2\rho_x^2+2uu_x\rho^2)=0.
\end{aligned}\tag{2.7}$$

We see that the first bosonic equation reduces to the mKdV equation when $v=0$ and $\rho^1=\rho^2=0$. The second one reduces to the KdV equation when $u=0$ and $\rho^1=\rho^2=0$. So we expect to find, for example, supersolitonic solutions for the complete supersymmetric system. Concerning the fermionic equations, they form a system of coupled linear equations in ρ^1 and ρ^2 once u and v are known. We will see later that the method of symmetry reduction will decouple these equations and will help in its resolution.

The Lie superalgebra of symmetries for Eq. (2.4) has been computed making use of a MAPLE program GLIE.[21] It is a $(3|2)$-dimensional superalgebra with basis

$$\begin{gathered}
\mathcal{P}_1=\partial_x,\quad \mathcal{P}_0=\partial_t,\\
\mathcal{D}=x\partial_x+3t\partial_t+\tfrac{1}{2}\theta_1\partial_{\theta_1}+\tfrac{1}{2}\theta_2\partial_{\theta_2}-A\partial_A,\\
\mathcal{Q}_1=\theta_1\partial_x-\partial_{\theta_1},\quad \mathcal{Q}_2=\theta_2\partial_x-\partial_{\theta_2},
\end{gathered}\tag{2.8}$$

where $\mathcal{P}_0,\mathcal{P}_1,\mathcal{D}$ are the three even generators and $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are the two odd ones. This result is true independent of the value of the parameter a entering in the superequation. The supercommutator table of the Lie superalgebra is given in Table I, where, as usual, commutation relations are satisfied for even–even and even–odd products while anticommutation relations are satisfied for odd–odd products.

As usual in super Lie group theory, starting with a Lie superalgebra, we obtain the corresponding Lie group by exponentiation. The group G of Lie-point symmetries for Eq. (2.4) is generated by the elements $g=(x_0,t_0,d;\eta_1,\eta_2)$ where x_0,t_0,d are even Grassmann numbers and η_1,η_2 odd ones. They satisfy the composition law

$$g\equiv(x_0,t_0,d;\eta_1,\eta_2)=(x_0^2,t_0^2,d_2;\eta_1^2,\eta_2^2)(x_0^1,t_0^1,d_1;\eta_1^1,\eta_2^1)=g_2g_1\tag{2.9}$$

with

$$\begin{gathered}
d=d_2+d_1,\quad t_0=t_0^2+e^{3d_2}t_0^1,\\
x_0=x_0^2+e^{d_2}x_0^1+\exp\left[\left(d_2+\frac{d_1}{2}\right)\right](\eta_1^2\eta_1^1+\eta_2^2\eta_2^1),\\
\eta_1=\eta_1^1+e^{-d_1/2}\eta_1^2,\quad \eta_2=\eta_2^1+e^{-d_1/2}\eta_2^2.
\end{gathered}\tag{2.10}$$

The identity element is $e=(0,0,0;0,0)$ and the inverse g^{-1} of g is

$$g^{-1}=(-e^{-3d}t_0,-e^{-d}x_0,-d;-e^{-d/2}\eta_1,-e^{-d/2}\eta_2). \tag{2.11}$$

This group acts on the superspace $(x,t;\theta_1,\theta_2)$ and on the superfield A as

$$g(x,t;\theta_1,\theta_2)=(e^d(x+\eta_1\theta_1+\eta_2\theta_2)-x_0,e^{3d}t-t_0;e^{d/2}(\theta_1-\eta_1),e^{d/2}(\theta_2-\eta_2)) \tag{2.12}$$

and

$$gA(x,t;\theta_1,\theta_2)=e^{-d}A(g^{-1}(x,t;\theta_1,\theta_2)). \tag{2.13}$$

The even parameters x_0 and t_0 correspond to space and time translations, respectively, while d corresponds to a dilatation. The two odd parameters η_1 and η_2 correspond to supersymmetric transformations mixing even and odd variables.

III. INVARIANT SOLUTIONS

Different subgroups may be chosen to get invariant solutions of the SUSY KdV equation (2.4). The more interesting ones contain at least one SUSY transformation. In fact, since both $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are like square roots of the translation generator $\mathcal{P}_1$, they come together in a subgroup structure and lead to invariant solutions constant in the variable x. We are not interested in such trivial solutions. Nevertheless, a combination like $\mathcal{Q}_+=\mathcal{Q}_1+i\mathcal{Q}_2$ is allowed giving $(\mathcal{Q}_+)^2=0$. The subgroup $G_1=\{g_0=(cb,b,0;\eta,i\eta)\}$ will be considered since it has the superalgebra $\mathcal{G}_1=\{\mathcal{P}_0-c\mathcal{P}_1,\mathcal{Q}_+\}$ and will give rise to traveling-wave solutions of Eq. (2.4). In order to perform the symmetry reduction using such a subgroup of the symmetry group G, we have first to find the invariants of the action of this subgroup on the independent and dependent variables and then rewrite the equations in terms of them. This will reduce the number of independent variables and we will get superequations with one even and one odd variables.

A. The reduced superequations

From Eqs. (2.12) and (2.13), we see that the subgroup G_1 acts on the independent and dependent variables as follows:

$$g_0(x,t;\theta_1,\theta_2)=(x+\eta(\theta_1+i\theta_2)-cb,t-b;\theta_1-\eta,\theta_2-i\eta),$$

$$g_0A(x,t;\theta_1,\theta_2)=A(g_0^{-1}(x,t;\theta_1,\theta_2)).$$

It is then easy to compute the invariants of this action. They are

$$y=x+ct+i\theta_1\theta_2,\quad \theta=\theta_1+i\theta_2. \tag{3.1}$$

Now, if we take $A=A(y,\theta)$, we get, from Eq. (2.4) and the fact that $D_1=\partial_\theta+\theta\partial_y$, $D_2=i(\partial_\theta-\theta\partial_y)$, the reduced superequation

$$\begin{aligned}cA_y={}&-A_{yyy}+3aA^2A_y-i(a+2)(A_y)^2-i(a+2)AA_{yy}\\&+2i(2a+1)\theta A_yA_{y\theta}+2i(a+2)\theta AA_{yy\theta}+2i(a-1)\theta A_\theta A_{yy}.\end{aligned} \tag{3.2}$$

Such an equation may be integrated once with respect to y and gives

$$A_{yy}-aA^3+i(a+2)AA_y-2i(a+2)\theta AA_{y\theta}-2i(a-1)\theta A_\theta A_y+cA+c_1-\theta k=0, \tag{3.3}$$

where c_1 and k are even and odd integration constants, respectively. In fact, it is again a PDE but much simpler than the original one. Indeed, expanding now the bosonic superfield A as

$$A(y,\theta)=u(y)+\theta\rho(y),$$

we finally get the system of ODEs:

$$u'' - au^3 + i(a+2)uu' + cu + c_1 = 0, \tag{3.4}$$

$$\rho'' - i(a+2)u\rho' + (c - 3au^2 + i(4-a)u')\rho = k. \tag{3.5}$$

The prime ($'$) means that we differentiate with respect to y. We notice that Eq. (3.4) is a second-order nonlinear differential equation in u which does not depend on ρ. For $a=-2$, we recover a reduction of the mKdV equation. Since u is an even function in y, the method of resolution of Eq. (3.4) is a standard one and we will use the classification given in Ince's book[22] to get explicit solutions. We will see that this will give a selection of admissible values for the parameter a. Equation (3.5) is linear in ρ once u is known and it can be solved by the usual techniques for linear equations, once we take $\rho(y) = \psi f(y)$, where ψ is an odd constant parameter and $f(y)$ an even function.

B. Solution of the equation for *u*

It has been noted that Eq. (3.4) reduces to the mKdV equation for the special value $a=-2$ and then has well-known solutions such as the soliton solution $u = 1/\cosh y$. Requiring that all solutions of Eq. (3.4) be free of movable critical points, we will see that this value of a appears in a list of only five admissible values, namely $a=-4$, -2, -1, 1, and 4. Here let us mention that the complete integrability of Eq. (2.2), in terms of the existence of Lax pairs, has been proven by Mathieu *et al.*[4] only for $a=-2$ and 4. Conservation laws have been found for $a=1$ but not a Lax pair.

To make use of the classification of second-order ODEs without movable critical points given in Ince's book,[22] we first make the following change of variables:

$$u(y) = \alpha(y)w(z(y)) + \beta(y), \quad \alpha(y) \neq 0, \quad z'(y) \neq 0, \tag{3.6}$$

in order to take Eq. (3.4) to one of the canonical forms of the equation

$$w_{zz} = (A(z)w + B(z))w_z + C(z)w^3 + D(z)w^2 + E(z)w + F(z). \tag{3.7}$$

This corresponds to the case 14.31 i) listed by Ince. The functions *A,B,C,D,E,F* are written as follows:

$$A(z) = -i(a+2)\frac{\alpha}{z'}, \quad B(z) = -\frac{1}{\alpha(z')^2}(2\alpha'z' + \alpha z'' + i(a+2)\beta\alpha z'),$$

$$C(z) = \frac{a\alpha^2}{(z')^2}, \quad D(z) = \frac{1}{(z')^2}(3a\alpha\beta - i(a+2)\alpha'),$$

$$E(z) = \frac{1}{\alpha(z')^2}(3a\alpha\beta^2 - i(a+2)(\alpha'\beta + \alpha\beta') - \alpha'' - \alpha c),$$

$$F(z) = \frac{1}{\alpha(z')^2}(a\beta^3 - c\beta - c_1 - i(a+2)\beta'\beta - \beta'').$$

In order to obtain an equation with solutions free from movable critical points, the pair of functions A and C must belong to the following list (after a suitable choice of the functions α, β and z):

$$\text{(i.a)} \quad A=0, \quad C=0; \quad \text{(i.b)} \quad A=-2, \quad C=0;$$

$$\text{(i.c)} \quad A=-3, \quad C=-1; \quad \text{(i.d)} \quad A=-1, \quad C=1;$$

$$\text{(i.e)}\quad A=0,\quad C=2.$$

The first two cases cannot occur in our context. Indeed, no value of a satisfies (i.a). For the case (i.b), a must be zero and $\alpha=-iz'$. Continuing with the classification,[23] we have to satisfy at least one of the following conditions:

$$\text{(i)}\quad B=D,\quad E=0;\quad \text{(ii)}\quad D=F=0,\quad E=B',$$

to get an integrable equation. This is impossible for $c\neq 0$.

The three other cases effectively occur and give rise to a selection of the admissible values for a. This leads to $a=1$ or $a=4$ for the case (i.c), $a=-4$ or $a=-1$ for the case (i.d) and finally $a=-2$ for the case (i.e). Let us now specify the classification process and exhibit the solutions for all these cases.

1. Case (i.c): a=1, or a=4

We have to take $\alpha=-3iz'/(a+2)$ and the functions B, D, E, and F must verify $B=D$ and $E=F=0$. The first equality is trivially satisfied when $a=1$ or $a=4$ and the other constraints lead to the following equations on β and z':

$$\beta''-a\beta^3+i(a+2)\beta\beta'+c\beta+c_1=0, \tag{3.8}$$

$$z'''+i(a+2)(z'\beta)'-3a\beta^2z'+cz'=0. \tag{3.9}$$

Equation (3.8) is the same as the original Eq. (3.4) but now we only need one particular solution. It is obtained as the constant solution

$$\beta=\left(\frac{-c_1}{2a}\right)^{1/3}=\left(\frac{c}{3a}\right)^{1/2}, \tag{3.10}$$

which gives a relation between c and the integration constant c_1. With such a solution β, the function z satisfying Eq. (3.9) can be simply taken as $z(y)=y$.

Thus, Eq. (3.7) reduces to the canonical form

$$w_{yy}=-3ww_y-w^3+q(w_y+w^2), \tag{3.11}$$

where the constant q is given by

$$q=-i(a+2)\left(\frac{c}{3a}\right)^{1/2}=-i(3c)^{1/2}, \tag{3.12}$$

which is the same for $a=1$ and $a=4$. The solution of Eq. (3.11) is easily computed using the substitution $w=v_y/v$, where v satisfies the linear equation $v'''=qv''$, i.e.,

$$v(y)=c_2e^{qy}+c_3y+c_4. \tag{3.13}$$

Finally, since $u(y)=(-3i/(a+2))w(y)+\beta$, we get

$$u(y)=\frac{i}{a+2}\left[q-3\,\frac{v_y}{v}\right]. \tag{3.14}$$

We see that v given by Eq. (3.13) depends on three integration constants but u depends only on two independent constants. Indeed for $c_2\neq 0$, we get

$$u(y)=\frac{i}{a+2}\left[q-3\,\frac{qe^{qy}+c_3}{e^{qy}+c_3y+c_4}\right]. \tag{3.15}$$

For $c_2=0, c_3\neq 0$, the solution is

$$u(y)=\frac{i}{a+2}\left[q-\frac{3}{y+c_4}\right], \tag{3.16}$$

and finally for $c_2=c_3=0$, we get the constant solution $u(y)=\beta$. Thus we have obtained the general solution of Eq. (3.4) with $a=1$, or $a=4$ subject to the constraint (3.10) on c and c_1.

2. Case (i.d): a=−4, or a=−1

We choose $\alpha=-iz'/(a+2)$ and all solutions of Eq. (3.7) will be free from movable critical points if $B=D=0$ and $F=-E'$. These conditions are satisfied for $\beta=0$, $z''=0$ and this implies that the integration constant c_1 must be zero. Hence Eq. (3.7) reduces to the canonical form

$$w_{zz}=-ww_z+w^3-pw, \tag{3.17}$$

where $p=-c/(z')^2$ is a constant. The standard way of solving Eq. (3.17) is to introduce a new function $v(z)$ that satisfies

$$v_z^2=p_3v^3+p_2v^2+p_1v+p_0, \tag{3.18}$$

for some constants p_i $(i=0,1,2,3)$ and to write $w=v_z/(v-1)$. From Eq. (3.18), we immediately get

$$v_{zz}=\tfrac{1}{2}(3p_3v^2+2p_2v+p_1), \quad v_{zzz}=(3p_3v+p_2)v_z. \tag{3.19}$$

Inserting now the new expression of w in Eq. (3.17) and using the expressions (3.19), we get the admissible values for the constants p_1, p_2, and p_3. A canonical choice gives

$$p_3=\frac{p}{3}, \quad p_2=0, \quad p_1=-p.$$

The solution of Eq. (3.18) is expressed in terms of the P-Weierstrass elliptic function, $\mathcal{P}(z,g_2,g_3)$, (see, e.g., Byrd and Friedman[23]) and is written as follows:

$$v(z)=\left(\frac{12}{p}\right)^{1/3}\mathcal{P}(z), \tag{3.20}$$

where

$$g_2=(12p^2)^{1/3}, \quad g_3=-p_0.$$

Finally, for the simplest choice $z(y)=y$, we obtain the solution of Eq. (2.4) as

$$u(y)=-\frac{i}{(a+2)}\left(\frac{v'(y)}{v(y)-1}\right)=-\frac{i}{(a+2)}\left(\frac{\mathcal{P}'(y)}{\mathcal{P}(y)-(p/12)^{1/3}}\right), \tag{3.21}$$

where $\mathcal{P}'(y)=-\sqrt{4\mathcal{P}^3(y)-g_2\mathcal{P}(y)+p_0}$.

3. Case (i.e): a=−2

This corresponds to the mKdV equation and all solutions are free from movable critical points. We set $\alpha^2=-(z')^2$ and have $B=D=0$. We choose the auxiliary quantities to be $\alpha=i$, $\beta=0$, and $z=y$ so that from Eq. (3.6), we see that $u(y)=iw(y)$. Hence, we can solve directly Eq. (3.4), which reduces to

$$u''+2u^3+cu+c_1=0. \tag{3.22}$$

Multiplying it by u' and integrating once, we get

$$(u')^2 = -(u^4 + cu^2 + 2c_1 u - c_2) = H(u), \tag{3.23}$$

where c_2 is an integration constant. This is a well-known equation solvable in terms of elliptic functions and their degenerate cases. Let us give some particular solutions.

We first consider the solitonic solution

$$u_S(y) = u_2 + \frac{\omega^2}{2u_2 + (u_1 + u_2)\cosh(\omega y)}, \tag{3.24}$$

where $\omega = \sqrt{(u_1 - u_2)(u_1 + 3u_2)}$, u_1, u_2 being roots of the polynomial $H(u)$ in Eq. (3.23). Indeed, if we denote by u_1, u_2, u_3 and u_4 the four roots of $H(u)$, the solution u_S corresponds to the case where $u_1 \geqslant u_S > u_2 = u_3 > u_4$ (note that $u_4 = -u_1 - u_2 - u_3$). The constants c and c_1 (written now as c_S and c_{1S}) are then given by

$$c_S = -(u_1^2 + 2u_1u_2 + 3u_2^2), \quad c_{1S} = -2u_2(u_1^2 + u_1u_2 + u_2^2).$$

The second solution we are interested in, is the rational one

$$u_R(y) = u_1 - \frac{4u_1}{4u_1^2 y^2 + 1}. \tag{3.25}$$

It corresponds to the case where three roots of $H(u)$ are equal and such that $u_1 = u_2 = u_3 > u_R > u_4$. Again the constants c and c_1 in Eq. (3.23) may be written as

$$c_R = -6u_1^2, \quad c_{1R} = -8u_1^3.$$

C. Solutions of the equation for ρ

Now we will use the solutions u of the PDE (3.4) to solve Eq. (3.5) that is linear in ρ. We can take $\rho(y) = \psi f(y)$ and $k = \psi k_1$, where ψ is an odd constant. We get a linear equation satisfied by the even function f:

$$f''(y) + p_1 f'(y) + p_0 f(y) = k_1, \tag{3.26}$$

where

$$p_1(y) = -i(a+2)u, \quad p_0(y) = c - 3au^2 + i(4-a)u' \tag{3.27}$$

and u is a solution of Eq. (3.4). For general solutions $u(y)$, Eq. (3.26) cannot be solved in terms of elementary functions, nor the standard special functions. Some particular solutions for u lead to simple forms of Eq. (3.26) that can be solved explicitly.

For the case $a=1$, or $a=4$, the general solution $u(y)$ was given by Eq. (3.14) with v satisfying the expression (3.13). It is easy to see that p_1 and p_0 in Eq. (3.27) are now given by

$$p_1(y) = \left(q - 3\frac{v_y}{v}\right), \quad p_0(y) = -2q\frac{v_y}{v} + (a-1)\left(\frac{v_y}{v}\right)^2. \tag{3.28}$$

We see that $p_1(y)$ does not depend on a and that $p_0(y)$ has been simplified taking into account the admissible values of a. We now distinguish the two values of a. For $a=1$, we use the particular solution (3.16) which leads to the linear equation

$$f'' + \left(q - \frac{3}{y}\right)f' - \frac{2q}{y}f = k_1, \tag{3.29}$$

where without loss of generality we put the constant c_4 equal to zero. The general solution is

$$f(y)=C_1(6-4qy+q^2y^2)+C_2(qy+3)e^{-qy}+\frac{k_1}{2q^2}(3-2qy), \tag{3.30}$$

C_1 and C_2 being arbitrary integration constants. For $a=4$, the same solution $u(y)$ inserted in Eq. (3.26) gives

$$f''+\left(q-\frac{3}{y}\right)f'+\left(\frac{-2q}{y}+\frac{3}{y^2}\right)f=k_1, \tag{3.31}$$

which admits the general solution

$$f(y)=qy[C_1(1-qy)+C_2e^{-qy}]-\frac{k_1y}{q}, \tag{3.32}$$

where again C_1 and C_2 are arbitrary constants.

We now turn to the case $a=-2$ which leads to the solitonic solution $u_S(y)$ of Eq. (3.24). Equation (3.26) takes the form

$$f''(y)+p_0(y)f(y)=k_1 \tag{3.33}$$

with

$$\begin{aligned} p_0(y)=\frac{1}{[(u_1+u_2)\cosh(\omega y)+2u_2]^2}\{&-(u_1^2+2u_1u_2+3u_2^2)((u_1+u_2)^2\cosh(\omega y)+2u_2)^2\\ &+6(u_2(u_1+u_2)\cosh(\omega y)+u_1^2+2u_1u_2-u_2^2)^2\\ &-6i(u_1-u_2)(u_1+u_2)(u_1+3u_2)\omega\sinh(\omega y)\}. \end{aligned} \tag{3.34}$$

Suitable changes of variables lead to a hypergeometric equation that can be solved exactly. Indeed, we first consider the homogeneous equation and use the change of variable $s=e^{\omega y}$ to get the new equation

$$f_{ss}+P(s)f_s+Q(s)f=0, \tag{3.35}$$

where

$$P(s)=\frac{1}{s},\quad Q(s)=\frac{-(s^2+10\kappa s+(\kappa)^2)}{s^2(s-\kappa)^2},$$

with $\kappa=-(2u_2-i\omega)/(u_1+u_2)$. A second change of variable will bring us to the hypergeometric equation

$$\varphi_{zz}+\frac{3(2z-1)}{z(z-1)}\varphi_z-\frac{6}{z(z-1)}\varphi=0, \tag{3.36}$$

where $z=s/(s-\kappa)$ and $\varphi=[1/z(z-1)]f$. One solution of Eq. (3.36) is easily found as $\varphi_1(z)=2z-1$, which leads us to a solution for the original homogeneous equation

$$f_1(y)=\frac{\kappa e^{\omega y}(e^{\omega y}+\kappa)}{(e^{\omega y}-\kappa)^3}. \tag{3.37}$$

The second linearly independent solution $\varphi_2(z)$ is obtained from $\varphi_1(z)$ using the formula

$$\varphi_2(z)=\varphi_1(z)\int \frac{1}{\varphi_1^2(z)}e^{-\int P(z)dz}dz, \tag{3.38}$$

or, explicitly,

$$\varphi_2(z)=32+\frac{7(2z-1)^2}{z(z-1)}-\frac{(2z-1)^2}{2z^2(z-1)^2}+30(2z-1)\ln\left(\frac{z-1}{z}\right). \tag{3.39}$$

We then get the other solution of the original homogeneous equation by replacing z in terms of y, i.e.,

$$f_2(y)=\frac{32\kappa e^{\omega y}+7(e^{\omega y}+\kappa)^2}{(e^{\omega y}-\kappa)^2}-\frac{(e^{\omega y}+\kappa)^2}{2\kappa e^{\omega y}}+30f_1(y)(-\omega y+\ln(\kappa)). \tag{3.40}$$

The corresponding solution of the inhomogeneous equation is obtained by the method of variation of constants and it is the linear combination of f_1 and f_2 and a particular solution given by

$$f_p(z(y))=\frac{k}{\omega^2}\left(-1-3z+33z^2-30z^3-6z(1-3z+2z^2)\ln\left(z-\frac{1}{z}\right)\right). \tag{3.41}$$

Finally let us take the rational solution u_R given by Eq. (3.25) and solve

$$f''+\left(\frac{48u_1^2}{(1-2iu_1y)^2}\right)f=k_1. \tag{3.42}$$

Once we put $r=1-2iu_1y$, we get an Euler type equation of the form

$$r^2f_{rr}-12f=\frac{k_1r^2}{-4u_1^2}. \tag{3.43}$$

We easily see that the general solution of Eq. (3.42) is

$$f(y)=C_1(1-2iu_1y)^4+C_2(1-2iu_1y)^{-3}+\frac{k_1}{40u_1^2}(1-2iu_1y)^2, \tag{3.44}$$

with C_1 and C_2 arbitrary integration constants.

D. Super traveling-wave solution

Returning to the original super KdV equation (2.4), let us give the expression for the solutions A which are invariant under G_1. Let us recall that we have solved Eq. (2.4) with $A=A(y,\theta)=u(y)+\theta\rho(y)$ where y is an even Grassmann variable given by Eq. (3.1). Developing u and ρ, we get (in view of the nilpotent character of θ_1 and θ_2)

$$A=u(x+ct)+i\theta_1\theta_2\frac{du}{d\xi}\bigg|_{\xi=x+ct}+(\theta_1+i\theta_2)\rho(x+ct). \tag{3.45}$$

This means that the components of A in Eq. (2.5) are not independent. Indeed we have

$$v=\frac{du}{d\xi}\bigg|_{\xi=x+ct},\quad \rho^2=i\rho^1=i\rho(\xi). \tag{3.46}$$

So we can view the solutions u and ρ as functions of $\xi=x+ct$ instead of y and $A=u(\xi)+i\theta_1\theta_2(du/d\xi)+(\theta_1+i\theta_2)\rho(\xi)$. For example, we have for $a=1$, the solutions

$$u(\xi)=\frac{i}{3}\left[q-\frac{3}{\xi}\right], \tag{3.47}$$

$$\rho(\xi)=\psi\left[C_1(6-4q\xi+q^2\xi^2)+C_2(q\xi+3)e^{-q\xi}+\frac{k_1}{2q^2}(3-2q\xi)\right]. \tag{3.48}$$

For $a=4$, we get

$$u(\xi)=\frac{i}{6}\left[q-\frac{3}{\xi}\right], \tag{3.49}$$

$$\rho(\xi)=\psi\left[q\xi[C_1(1-q\xi)+C_2e^{-q\xi}]-\frac{k_1\xi}{q}\right]. \tag{3.50}$$

Let us now develop the supersolitonic solution occurring for $a=-2$. First, we take some particular values for the constants appearing in the solution (3.24). Thus, with $u_1=1$, $u_2=0$, and $c=-1$, we get

$$u(x-t)=\frac{1}{\cosh(x-t)}, \tag{3.51}$$

the well-known solitonic solution of the mKdV equation. The first solution for the homogeneous equation in ρ becomes very simple. It takes the form $\rho_1(x-t)=\psi f_1(x-t)$ where

$$f_1(x-t)=\frac{1}{\cosh^3(x-t)}\left[\sinh(x-t)+\frac{i}{2}(1-\sinh^2(x-t))\right]. \tag{3.52}$$

In Fig. 1, we see the behavior of u, $\mathrm{Re}(f_1)$ and $\mathrm{Im}(f_1)$ as functions of $\xi=x-t$. As function of x and t, we have the graphs of $\mathrm{Re}(f_1)$ and $\mathrm{Im}(f_1)$ in Figs. 2 and 3. The solution (3.52) has a very interesting behavior, discussed earlier by Ibort *et al.*[16] for a different supersymmetric equation. They had a real solution which they called a *solitino* and its graph was similar to that of $\mathrm{Re}(f_1)$ in Fig. 2. Here we get a complex solution for which the norm of the corresponding even function $f_1(x-t)$ is $|f_1(x-t)|=\frac{1}{2}u(x-t)$. We also see that the real part of f_1 is the derivative of the usual soliton solution (up to a multiplication by a constant factor) of the KdV equation, while the imaginary part is related to u by $\mathrm{Im}\, f_1(\xi)=d^2u(\xi)/d\xi^2$.

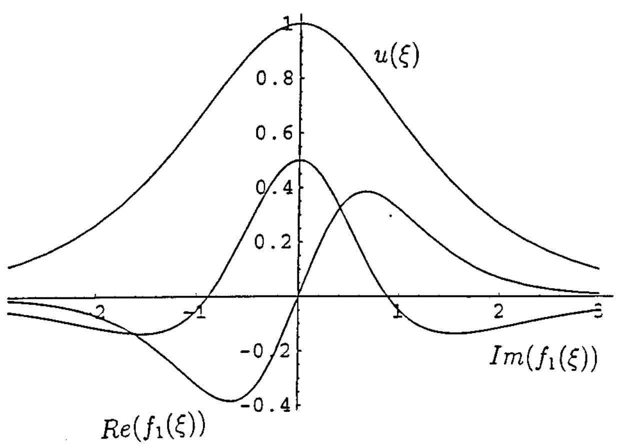

FIG. 1. The functions $u(\xi)$, $\mathrm{Re}(f_1(\xi))$, and $\mathrm{Im}(f_1(\xi))$.

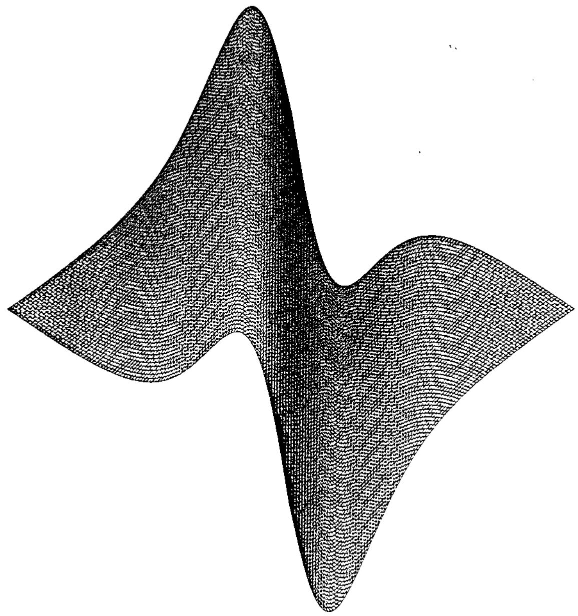

FIG. 2. The function $\mathrm{Re}(f_1(x-t))$.

Let us emphasize that our soliton type solution of the SUSY mKdV equation provides a complete solution of Eq. (2.4), since we have obtained all components of A [see Eq. (3.45)]. The behavior of the other independent solutions $\rho_2(\xi)$ and $\rho_p(\xi)$ is more complex, but we can compute it in the particular case where u is given by Eq. (3.51). They take the form

$$f_2(\xi)=\mathrm{Re}(f_2(\xi))+i\ \mathrm{Im}(f_2(\xi)) \tag{3.53}$$

with

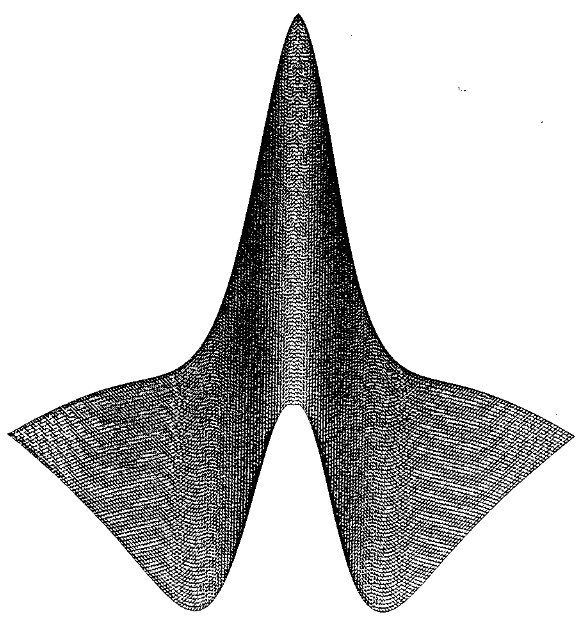

FIG. 3. The function $\mathrm{Im}(f_1(x-t))$.

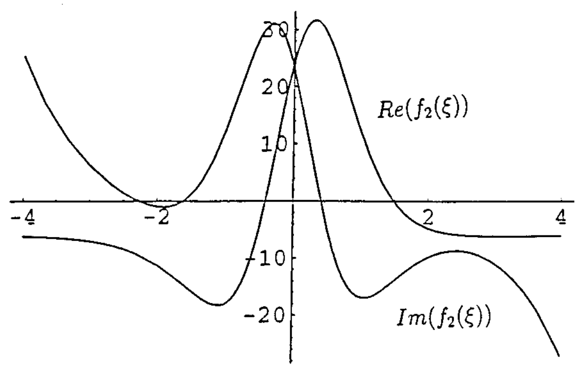

FIG. 4. The functions $\mathrm{Re}(f_2(\xi))$ and $\mathrm{Im}(f_2(\xi))$.

$$\mathrm{Re}(f_2(\xi)) = -6 + 15\pi \frac{\sinh(\xi)}{(\cosh(\xi))^3} + 30 \frac{1}{(\cosh(\xi))^2} - 30\xi \frac{\sinh(\xi)}{(\cosh(\xi))^3}, \tag{3.54}$$

$$\mathrm{Im}(f_2(\xi)) = -\sinh(\xi) + 15 \frac{\xi}{\cosh(\xi)} - 30 \frac{\xi}{(\cosh(\xi))^3} - \frac{15}{2}\pi \frac{1}{\cosh(\xi)} + 15\pi \frac{1}{(\cosh(\xi))^3} - 30 \frac{\sinh(\xi)}{(\cosh(\xi))^2}. \tag{3.55}$$

Finally, the particular solution f_p is

$$f_p(\xi) = \mathrm{Re}(f_p(\xi)) + i\,\mathrm{Im}(f_p(\xi)) \tag{3.56}$$

with

$$\mathrm{Re}(f_p(\xi)) = k_1\left(-1 - \frac{3}{2}\frac{e^{\xi}}{\cosh(\xi)} + \frac{33}{4}\frac{(-1+e^{2\xi})}{4(\cosh(\xi))^2} - \frac{15}{4}\frac{(-3e^{\xi}+e^{3\xi})}{(\cosh(\xi))^3} - 6\left(\frac{\pi}{2} - \xi\right)\left[\frac{e^{\xi}}{2\cosh(\xi)} - \frac{3}{4}\frac{(-1+e^{2\xi})}{(\cosh(\xi))^2} + \frac{(-3e^{\xi}+e^{3\xi})}{4(\cosh(\xi))^3}\right]\right), \tag{3.57}$$

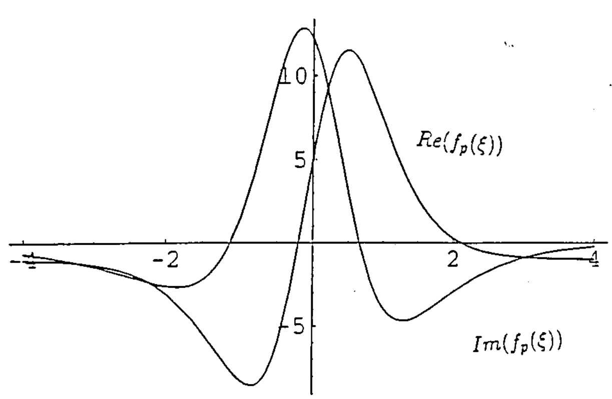

FIG. 5. The functions $\mathrm{Re}(f_p(\xi))$ and $\mathrm{Im}(f_p(\xi))$.

$$\mathrm{Im}(f_p(\xi))=k_1\left(-\frac{3}{2\cosh(\xi)}+\frac{33}{2}\frac{e^{\xi}}{(\cosh(\xi))^2}-15\frac{(-1+3e^{2\xi})}{4(\cosh(\xi))^3}\right.$$
$$\left.-6\left(\frac{\pi}{2}-\xi\right)\left[\frac{1}{2\cosh(\xi)}-\frac{3e^{\xi}}{2(\cosh(\xi))^2}+\frac{(-1+3e^{2\xi})}{4(\cosh(\xi))^3}\right]\right). \tag{3.58}$$

All these functions are represented in Figs. 4 and 5 as functions of ξ and for $k_1=1$.

IV. CONCLUSION

Starting from the $N=2$ SUSY KdV equation (2.4), we have used a subgroup of the supersymmetry group to obtain Grassmann-valued solutions depending on one even and one odd Grassmann independent variable. Such an equation contains a lot of information, since if we decompose the bosonic superfield in terms of its components, As in Eq. (2.5), it produces a set of four partial differential equations. Our way of determining symmetries has the advantage of avoiding that decomposition, of working with the concise equation (2.4), and of producing a superalgebra (2.8) of symmetries. In the search for solutions using the method of symmetry reduction, we work with Grassmann variables until we get a PDE with one even and one odd Grassmann variable. At this stage, the nilpotency of the odd variable leads, by expansion of the dependent variable, to a set of two ODEs.

The invariant solutions that we have obtained are based on the choice of G_1 as a subgroup of the SUSY group. Let us recall that it contains a SUSY transformation and a combination of both spatial and temporal translations. It gives rise to interesting solutions such as the supersolitonic ones. Another group G_2 containing the same SUSY transformation and the dilation generated by $\mathcal{D}$ may also be considered. Here the difficulty is to solve explicitly the linear equation for the odd field ρ. Indeed, if we take the group $G_2=\{g_0'=(0,0,d;\eta,i\eta)\}$, it acts on the independent and dependent variables as

$$g_0'(x,t;\theta_1,\theta_2)=(e^d(x+\eta(\theta_1+i\theta_2)),e^{3d}t;e^{d/2}(\theta_1-\eta),e^{d/2}(\theta_2-i\eta)),$$

$$g_0'A(x,t;\theta_1,\theta_2)=e^{-d}A((g_0')^{-1}(x,t;\theta_1,\theta_2)),$$

and the invariants of this action are

$$y=t^{-1/3}(x+i\theta_1\theta_2),\quad \theta=\theta_1+i\theta_2,\quad W=t^{1/3}A. \tag{4.1}$$

It is easy to show that the reduced equation in terms of $W=W(y,\theta)$ may be developed as a set similar to Eqs. (3.4) and (3.5) where the constant c is replaced by $-y/3$. This means that the solution for u essentially follows the lines described in Sec. III. In particular, this implies the selection of the same values for a, important information for the integrability of the SUSY KdV equation. We plan to return to the study of self-similar SUSY solutions in the near future.

ACKNOWLEDGMENTS

This work is partly supported by research grants from NSERC of Canada and FCAR du gouvernement du Québec.

[1] B. A. Kupershmidt, J. Phys. A **17**, 869 (1984).
[2] Yu I. Manin and A. O. Radul, Commun. Math. Phys. **98**, 65 (1985).
[3] P. Mathieu, J. Math. Phys. **29**, 2499 (1988).
[4] P. Labelle and P. Mathieu, J. Math. Phys. **32**, 923 (1991).
[5] G. H. M. Roelofs and P. H. M. Kersten, J. Math. Phys. **33**, 2185 (1992).
[6] J. C. Brunelli and A. Das, Phys. Lett. B **337**, 303 (1994).
[7] B. W. Fatyga, V. A. Kostelecký, and D. R. Truax, J. Math. Phys. **30**, 1464 (1989).
[8] P. M. Kersten, Phys. Lett. A **34**, 25 (1988).
[9] M. H. Roelofs, Ph.D. thesis, University of Twente, Enschede, The Netherlands, 1993.
[10] I. S. Krasil'shchik and P. H. M. Kersten, Acta Applicandae Mathematicae **41**, 167 (1995).

[11] C. Elphick, J. Math. Phys. **28**, 1243 (1987).
[12] C. M. Yung, Phys. Lett. B **309**, 75 (1993).
[13] S. Belluci, E. Ivanov, S. Krivonos, and A. Pichugin, Phys. Lett. B **312**, 463 (1993).
[14] M. A. Ayari and V. Hussin, ''Computations of Lie Supersymmetries for Grassmann-Valued Differential Equations,'' *Group 21, Physical Applications and Mathematical Aspects of Geometry, Groups and Algebras*, Vol. 2, edited by H. D. Doebner, W. Scherer and C. Schulte (World Scientific, Singapore, 1997), p. 1032.
[15] R. Cianci, J. Math. Phys. **29**, 2156 (1988).
[16] A. Ibort, A. L. Martínez, and R. Medina, J. Math. Phys. **37**, 6157 (1996).
[17] C. Z. Qu, Commun. Theor. Phys. **24**, 177 (1995).
[18] M. A. Ayari, M. I. Ayari, and V. Hussin, Comput. Phys. Commun., Special Issue: Computer Algebra in Physics Research **115**, 1 (1998).
[19] M. A. Ayari and V. Hussin, ''Supersolitonic solutions of the $N=2$ super-KdV equation,'' in *Proceedings of the Fifth International Wigner Symposium*, edited by P. Kasperkovitz and D. Grau (World Scientific, Singapore, 1998), p. 248.
[20] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer, New York, 1986).
[21] M. A. Ayari and V. Hussin, Comput. Phys. Commun. **100**, 157 (1997).
[22] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956).
[23] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals For Engineers and Physicists* (Springer, New York, 1971).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

On the integrability of nonlinear partial differential equations

H. J. S. Dorren[a)]
Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

(Received 31 August 1998; accepted for publication 15 January 1999)

We investigate the integrability of Nonlinear Partial Differential Equations (NPDEs). The concepts are developed by first discussing the integrability of the KdV equation. We proceed by generalizing the ideas introduced for the KdV equation to other NPDEs. The method is based upon a linearization principle that can be applied on nonlinearities that have a polynomial form. The method is further illustrated by finding solutions of the nonlinear Schrödinger equation and the vector nonlinear Schrödinger equation, which play an important role in optical fiber communication. Finally, it is shown that the method can also be generalized to higher dimensions. © *1999 American Institute of Physics.* [S0022-2488(99)01904-0]

I. INTRODUCTION

The conditions under which Nonlinear Partial Differential Equations (NPDEs) can be solved are even in one dimension not well understood.[1] Roughly speaking, the majority of the integrable systems can be classified in three main groups. In the first of these groups are those equations that can be reduced to a quadrature through the existence of an adequate number of integrals of motion. In the second class are those equations that can be mapped into a linear system by applying a number of transformations (hereafter to be called C integrable[2]). The last group consists of differential equations that can be solved by Inverse Scattering Transformations (IST). In the following, we will call equations that can be solved by inverse scattering methods ''*S* integrable.'' The discovery of the IST has led to considerable progress in understanding the topic of integrability, since this technique made it possible to investigate the integrability of large classes of NPDEs systematically.[3]

Another important consideration is that most of the work on the integrability of NPDEs has been carried out in one space dimension only. Although the inverse problem of the Schrödinger equation can be generalized to three dimensions, the method is far too complicated to solve higher-dimensional NPDEs. An alternative is the $\bar{\partial}$ approach, which is also successfully generalized to N dimensions (see, for instance, the book by Ablowitz and Clarkson[3]). Nevertheless, for both these methods the existence of the obtained solutions is difficult to prove. The concept of C integrability, however, has the potential to be generalized to dimensions higher than one. In this paper, we will demonstrate a simple method based upon linearization principles that enables us to compute solutions of large classes of NPDEs by solving a linear algebraic recursion relationship. The result suggests that the method can be generalized to higher-dimensional NPDEs.

In this paper we aim to find integrable differential equations that can be solved by linearization. The basic idea of the method goes back to Stokes,[4] and is used several times to obtain solutions of nonlinear evolution equations.[5–9] We will apply the method in a slightly different form to find conditions on the integrability of nonlinear evolution equations. Since it is not clear what integrability exactly means, we use in this paper the heuristic definition that a NPDE is integrable if given a sufficiently general initial condition, we can find analytic expressions the time evolution of the solution. For NPDEs that can be solved by inverse scattering techniques, this

[a)]Electronic mail: H.J.S.Dorren@ele.tue.nl notion is equivalent with the existence of N-soliton solutions, since it is implicitly assumed that the obtained solution can be expanded on a Fourier basis.[8] It is shown that the condition of expansion in a Fourier can be replaced by an arbitrary other infinite set of basis functions.

We present the following results. First, we derive a simple method to test NPDEs with a polynomial type of nonlinearity in the presence of N-soliton solutions. Necessary conditions that indicate whether a NPDE has N-soliton solutions includes that the nonlinearity can be expanded in the same basis functions as the linear part, and second that the dispersion relationship associated with the linearized problem can be solved. The method is demonstrated by first discussing the integrability of the KdV equation in Sec. II. In Sec. III, the concepts derived for the KdV equation are generalized to discuss the integrability of more general NPDEs. Finally, in Sec. IV, the results are applied to investigate the integrability of the coupled nonlinear Schrödinger equation. Moreover, it is indicated that the method can also be used to obtain solutions to higher-dimensional NPDEs. The paper is concluded with a discussion.

II. THE INTEGRABILITY OF THE KdV EQUATION

In order to illustrate the machinery developed throughout this paper, we first discuss the integrability of the KdV equation as an example. The integrability of the KdV equation is a well-studied problem.[3] This makes the KdV equation an ideal object to test the validity of newly developed ideas with respect to the integrability of NPDEs. We will introduce our methods on the integrability of NPDEs by discussing the existence of N-soliton solutions for the KdV equation, which is given by

$$u_t+u_{xxx}=6u_xu. \tag{1}$$

We try to find solutions of Eq. (1) by substitution of the following Fourier series:

$$u(x,t)=\sum_{n=1}^{\infty} A_n e^{in(kx-\omega t)}. \tag{2}$$

If we substitute the solution $u(x,t)$ into Eq. (2), we obtain

$$\sum_{n=1}^{\infty} (n\omega+k^3n^3)A_n e^{in(kx-\omega t)}=-6k\sum_{n=1}^{\infty}\sum_{l=1}^{n-1} lA_lA_{n-l}e^{in(kx-\omega t)}. \tag{3}$$

We can now determine the coefficients A_n by deriving a recursion relationship. This can be achieved by comparing the exponential functions in Eq. (3). If we compare all the terms for which $n=1$, we find

$$(\omega+k^3)A_1e^{i(kx-\omega t)}=0. \tag{4}$$

For a nonzero A_1, we find that Eq. (4) is satisfied if

$$\omega=-k^3. \tag{5}$$

If we put $n=2$ in Eq. (3), we can determine A_2 by solving the following relationship:

$$(2\omega+8k^3)A_2e^{2i(kx-\omega t)}=-6kA_1A_1e^{2i(kx-\omega t)}. \tag{6}$$

If we use the dispersion relationship (5), we find that A_2 is given by

$$A_2=-\frac{A_1^2}{k^2}. \tag{7}$$

By repeating this procedure, we can compute all the expansion coefficients A_n of the solutions $u(x,t)$. In general, all the coefficients A_n can be computed by solving the following linear algebraic problem:

$$L^{(n)}(k)A_n=R^{(n)}(k). \tag{8}$$

The operators $L^{(n)}(k)$ and $R^n(k)$ in Eq. (8) are given by

$$L^{(n)}(k)=n[n^2-1]k^3, \quad R^{(n)}(k)=-6k\sum_{l=1}^{n-1} lA_lA_{n-l}. \tag{9}$$

If we compute all the coefficients A_n by using Eq. (9), we then obtain the Fourier expansion of $u(x,t)$, for which the first terms are given by

$$u(x,t)=A_1e^{i(kx-\omega t)}-\frac{A_1^2}{k^2}e^{2i(kx-\omega t)}+\frac{3A_1^3}{4k^4}e^{-3i(kx-\omega t)}+\cdots. \tag{10}$$

If we substitute $k=2i\beta$ and $A_1=4\,d\beta$ into Eq. (10), we find

$$u(x,t)=4\,d\beta\,e^{-2(\beta x-4\beta^3t)}+16\,d^2e^{-4(\beta x-4\beta^3t)}+\frac{24d^3}{\beta}e^{-6(\beta x-4\beta^3t)}+\cdots. \tag{11}$$

By carrying out the summation in Eq. (11), we can formulate this equation more compactly:

$$u(x,t)=\frac{8\,d\beta\,e^{-2(\beta x-4\beta^3t)}}{\left(1+\frac{d}{\beta}e^{-2(\beta x-4\beta^3t)}\right)^2}. \tag{12}$$

Hence, if we put

$$\beta=\frac{1}{2}\sqrt{c}, \quad x_0=-\frac{1}{\sqrt{c}}\log\left(-\frac{d}{\beta}\right), \quad d<0, \tag{13}$$

we can simplify Eq. (12) one step further to

$$u(x,t)=-\frac{c}{2}\operatorname{sech}^2\left\{\frac{1}{2}\sqrt{c}(x-ct+x_0)\right\}. \tag{14}$$

Equation (14) describes the well-known KdV soliton.

What did we learn from this simple exercise? At first, the KdV equation has solutions because of the special structure of the nonlinearity. If we substitute the special solution (2) in the nonlinear part of the KdV equation, we find that we can expand the nonlinearity in the same basis functions as the linear part:

$$6u_xu=\sum_{n=1}^{\infty} D_ne^{in(kx-\omega t)}; \quad D_n=-6k\sum_{l=1}^{n-1} lA_lA_{n-l}. \tag{15}$$

This guarantees that we can find an iteration relationship for the expansion coefficients A_n. As we will see later, we do not have to restrict to a Fourier expansion of the solution only. In principle, this method works for any set of basis functions as long as we can expand the nonlinearity in the same basis functions as the linear part. In the following, we will show that the structure of the nonlinearity of the KdV equation enables us to construct the Fourier expansion of the N soliton of the KdV equation. In order to systematically solve these solutions it is illustrative to discuss also the two-soliton solutions, which are assumed to have the following series expansion:

$$u(x,t)=\sum_{\mu_1,\mu_2=1}^{\infty} C(\mu_1,\mu_2)e^{i(\mu_1 k_1 z_1+\mu_2 k_2 z_2)} \begin{cases} z_1=x-\dfrac{\omega(k_1)}{k_1}, \\ z_2=x-\dfrac{\omega(k_2)}{k_2}. \end{cases} \tag{16}$$

If we substitute Eq. (16) into the KdV equation (1), we obtain the following result:

$$\sum_{\mu_1,\mu_2=1}^{\infty} L^{(\mu_1,\mu_2)}(k_1,k_2)C(\mu_1,\mu_2)e^{i(\mu_1 k_1 z_1+\mu_2 k_2 z_2)}$$
$$=-6\sum_{\mu_1,\mu_2=1}^{\infty}\sum_{\eta_1,\eta_2=1}^{\mu_1-1,\mu_2-1} M^{(\eta_1,\eta_2)}(k_1,k_2)C(\mu_1-\eta_1,\mu_2-\eta_2)C(\eta_1,\eta_2)e^{i(\mu_1 k_1 z_1+\mu_2 k_2 z_2)}, \tag{17}$$

where

$$L^{(n_1,n_2)}(k_1,k_2)=\sum_{i=1}^{2} n_i[n_i^2-1]k_i^3, \quad M^{(n_1,n_2)}(k_1,k_2)=\sum_{i=1}^{2} n_i k_i\,. \tag{18}$$

We solve Eq. (17) by comparing equal exponential powers on both sides. This can be done by defining a parameter $\Gamma=\mu_1+\mu_2$ and subsequently comparing the powers for $\Gamma=1,2,3,\ldots$. We first discuss the case in which $\Gamma=1$ in which only the coefficients $C(1,0)$ and $C(0,1)$ contribute:

$$[\omega_1+k_1^3]C(1,0)e^{ik_1z_1}+[\omega_2+k_2^3]C(0,1)e^{ik_2z_2}=0. \tag{19}$$

If we put $C(1,0)=A_1$ and $C(0,1)=A_2$, we find that the following linear dispersion relationships must be valid:

$$\omega(k_1)=-k_1^3 \quad \text{and} \quad \omega(k_2)=-k_2^3. \tag{20}$$

Once the linear dispersion relationships are determined and if the coefficients $C(1,0)$ and $C(0,1)$ have taken their values A_1 and A_2, we can compute all the other coefficients $C(\mu,\eta)$ by applying the following linear recursion relation:

$$L^{(\mu_1,\mu_2)}(k_1,k_2)C(\mu_1,\mu_2)=R^{(\mu_1,\mu_2)}(k_1,k_2), \tag{21}$$

where

$$R^{(\mu_1,\mu_2)}(k_1,k_2)=-6\sum_{\eta_1,\eta_2=1}^{\mu_1-1,\mu_2-1} M^{(\eta_1,\eta_2)}(k_1,k_2)C(\mu_1-\eta_1,\mu_2-\eta_2)C(\eta_1,\eta_2). \tag{22}$$

Equation (21) has a similar structure as Eq. (8). In principle, Eq. (21) provides an efficient tool to compute all the coefficients $C(\mu,\eta)$. We can generalize this result to the N-soliton case by assuming that the solution $u(x,t)$ takes the following form:

$$u(x,t)=\sum_{\mu_1\cdots\mu_N=1}^{\infty} C(\mu_1\cdots\mu_N)^{i(\mu_1 k_1 z_1+\cdots+\mu_N k_N z_N)} \begin{cases} z_1=x-\dfrac{\omega(k_1)}{k_1} \\ \vdots \\ z_N=x-\dfrac{\omega(k_N)}{k_N}. \end{cases} \tag{23}$$

We can determine the nonzero coefficients $C(\mu_1\cdots\mu_N)$ by substituting Eq. (23) into the KdV equation (1):

$$\sum_{\mu_1\cdots\mu_N=1}^{\infty} L^{(\mu_1\cdots\mu_N)}(k_1\cdots k_N)C(\mu_1\cdots\mu_N)e^{i(\mu_1k_1z_1+\cdots+\mu_Nk_Nz_N)}$$
$$=-6\sum_{\mu_1\cdots\mu_N=1}^{\infty}\sum_{\eta_1\cdots\eta_N=1}^{\mu_1-1\cdots\mu_N-1} M^{(\eta_1\cdots\eta_N)}(k_1\cdots k_N)C(\mu_1-\eta_1\cdots\mu_N-\eta_N)C(\eta_1\cdots\eta_N)$$
$$\times e^{i(\mu_1k_1z_1+\cdots+\mu_Nk_Nz_N)}, \tag{24}$$

where

$$L^{(n_1\cdots n_N)}(k_1\cdots k_N)=\sum_{i=1}^{N} n_i[n_i^2-1]k_i^3;\quad M^{(n_1\cdots n_N)}(k_1\cdots k_N)=\sum_{i=1}^{N} n_ik_i. \tag{25}$$

If we use that $\omega(k_i)=-k_i^3$, $(i\in 1\cdots N)$ and $A_1=C(1,0,0,...,0), A_2=C(0,1,0,...,0),\ldots,A_N=C(0,...,0,1)$, we find that the expansion coefficients of the N-soliton solution for the KdV equation can be computed by solving the following linear relationship:

$$L^{(\mu_1\cdots\mu_N)}(k_1\cdots k_N)C(\mu_1\cdots\mu_N)=R^{(\mu_1\cdots\mu_N)}(k_1\cdots k_N), \tag{26}$$

where

$$R^{(\mu_1\cdots\mu_N)}(k_1\cdots k_N)=-6\sum_{\eta_1\cdots\eta_N=1}^{\mu_1-1\cdots\mu_N-1} M^{(\eta_1\cdots\eta_N)}(k_1\cdots k_N)C(\mu_1-\eta_1\cdots\mu_N-\eta_N)C(\eta_1\cdots\eta_N). \tag{27}$$

From the exercise performed in this section, we can conclude that general solutions of the KdV equation can be obtained by solving Eqs. (26). This implies that the KdV equation can be transformed into a simple linear algebraic equation in the coefficient space. We can conclude that the KdV equation has N-soliton solutions because the following two conditions are satisfied.

(i) The structure of the nonlinearity of the kdV equation guarantees that the equation has solutions of the form (23). This result implies that the coefficients $R^{(\mu_1\cdots\mu_N)}(k_1\cdots k_N)$ exist.

(ii) $L^{(n_1\cdots n_N)}(k_1\cdots k_N)$ is not equal to zero if $k_1\cdots k_N\neq 0$ and $n_1\cdots n_N\neq 0$. This implies that $L^{(n_1\cdots n_N)}(k_1\cdots k_N)$ has an inverse.

In the following section we will show that a similar condition must hold for other NPDEs. In the following section it is shown that the concepts derived for the KdV equation can be generalized to large classes of NPDEs. The results obtained in this section are derived by assuming that the solution of the KdV equation can be expanded in Fourier basis functions. In the following section, it will be shown that similar principles apply for other basis functions.

III. GENERALIZATIONS

In this section we will present more general results with respect to the integrability of nonlinear evolution equations. This will be done by generalizing the results obtained for the KdV equation. In this section, we focus on NPDEs of the following type:

$$\mathcal{L}[\mathbf{u}(x,t)]=Q[\mathbf{u}(x,t)]. \tag{28}$$

In Eq. (28), the function $\mathbf{u}(x,t)$ is an M-component vector function having entries $u_i(x,t)$. The operator $\mathcal{L}[\cdot]$ is assumed to take the following form:

$$\mathcal{L}[\mathbf{u}(x,t)]=\left[i\mathbf{I}\frac{\partial}{\partial t}+\sum_{n=1}^{K}\mathbf{A}^{(n)}\frac{\partial^n}{\partial x^n}\right]\mathbf{u}(x,t). \tag{29}$$

The matrices $\mathbf{A}^{(n)}$ in Eq. (29) are $M\times M$ matrices and $\mathbf{I}$ is the identity matrix. As concluded from the previous section, integrability puts strong constraints on the nonlinearity represented by the operator Q. As a necessary condition for the integrability we require that if a solution of Eq. (28) has the following form:

$$\mathbf{u}(x,t)=\sum_{\mu_1\cdots\mu_N=1}^{\infty}\mathbf{C}(\mu_1\cdots\mu_N)\exp\left[i\sum_{r=1}^{N}\sum_{s=1}^{M}\mu_r k_{rs}z_{rs}\right];\quad z_{rs}=x-\frac{\omega(k_{rs})}{t}, \tag{30}$$

then the operator Q must satisfy the following property:

$$Q[\mathbf{u}(x,t)]=\sum_{\mu_1\cdots\mu_N=1}^{\infty}\mathbf{R}(\mu_1\cdots\mu_N)\exp\left[i\sum_{r=1}^{N}\sum_{s=1}^{M}\mu_r k_{rs}z_{rs}\right], \tag{31}$$

where $\mathbf{C}(\mu_1\cdots\mu_N)$ and $\mathbf{R}(\mu_1\cdots\mu_N)$ are M-dimensional vector functions. Similarly as for the KdV equation, the vector function $\mathbf{R}(\mu_1\cdots\mu_N)$ is specified by the nonlinearity. In other words, we require that given a solution of the form (30), the nonlinear operator $Q[\mathbf{u}(x,t)]$ can be expanded in the same set of basis functions as $\mathcal{L}[\mathbf{u}(x,t)]$. In the previous section, we have shown that the nonlinearity of the KdV equation satisfies this condition. In general, large classes of nonlinear operators will have the property (31) and among them we are especially interested in the subclass $\hat{P}$, which plays an important role in nonlinear optics:

$$\hat{P}[\mathbf{u}(x,t)]=P_N\left(\mathbf{u},\frac{\partial\mathbf{u}}{\partial x},\frac{\partial\mathbf{u}}{\partial t},\ldots,\frac{\partial^p\mathbf{u}}{\partial x^q\partial t^{p-q}}\right), \tag{32}$$

where P_N are polynomials of order N. If we let act the linear operator $\mathcal{L}$ onto the solution (30), we obtain the following relationship:

$$\mathcal{L}[\mathbf{u}(x,t)]=\left(\mathbf{I}\sum_{r=1}^{N}\sum_{s=1}^{M}\mu_r\omega(k_{rs})+\sum_{n=1}^{K}\mathbf{A}^{(n)}\left[i\sum_{r=1}^{N}\sum_{s=1}^{M}\mu_r k_{rs}\right]^n\right)\mathbf{u}(x,t). \tag{33}$$

From this result, we can identify a matrix $\mathbf{L}^{(\mu_1\cdots\mu_N)}(k_{ij})$, which is given by

$$\mathbf{L}^{(\mu_1\cdots\mu_N)}(k_{ij})=\mathbf{I}\sum_{r=1}^{N}\sum_{s=1}^{M}\mu_r\omega(k_{rs})+\sum_{n=1}^{K}\mathbf{A}^{(n)}\left[i\sum_{r=1}^{N}\sum_{s=1}^{M}\mu_r k_{rs}\right]^n. \tag{34}$$

This result implies that the coefficients $\mathbf{C}(\mu_1\cdots\mu_N)$ that determine the solution (30) can be determined by solving

$$\mathbf{L}^{(\mu_1\cdots\mu_N)}(k_{ij})\mathbf{C}(\mu_1\cdots\mu_N)=\mathbf{R}(\mu_1\cdots\mu_N). \tag{35}$$

The coefficients $\mathbf{C}(1,0,0,...,0),\mathbf{C}(0,1,0,...,0),...,\mathbf{C}(0,...,0,1)$ are determined by the initial condition.

In principle, we expand the solution $u(x,t)$ in an arbitrary set of basis functions. Suppose as an example a function $\hat{u}(x,t)$ that can be expanded in the set of basis functions $f^{(n)}(x,t|k,\omega)$:

$$\hat{u}(x,t)=\sum_{n=1}^{\infty}\alpha_n f^{(n)}(x,t|k,\omega). \tag{36}$$

We define the set $\mathcal{S}$ as the basis function:

$$\mathcal{S}=\{f^{(1)}(x,t|k,\omega),f^{(2)}(x,t|k,\omega),f^{(3)}(x,t|k,\omega),\ldots\}, \tag{37}$$

which have the following properties:

I: if $f^{(n)}(x,t|k,\omega)\in\mathcal{S}\Rightarrow\frac{\partial}{\partial t}f^{(n)}(x,t|k,\omega)=\hat{\alpha}_n(k,\omega)f^{(m)}(x,t|k,\omega)\quad(f^{(m)}(x,t)\in\mathcal{S})$,

II: if $f^{(n)}(x,t|k,\omega)\in\mathcal{S}\Rightarrow\frac{\partial}{\partial x}f^{(n)}(x,t|k,\omega)=\hat{\beta}_n(k,\omega)f^{(m)}(x,t|k,\omega)\quad(f^{(m)}(x,t)\in\mathcal{S})$,

III: if $f^{(n)}(x,t)\in\mathcal{S}$ and $f^{(m)}(x,t)\in\mathcal{S}\Rightarrow f^{(n)}(x,t)\cdot f^{(m)}(x,t)\in\mathcal{S}$. (38)

The properties I and II guarantee that $\mathcal{L}[\hat{u}(x,t)]$ can be expanded in basis functions $f^{(n)}(x,t|k,\omega)$:

$$\mathcal{L}[\hat{u}(x,t)]=\sum_{n=1}^{\infty}\hat{L}^{(n)}\alpha_n f^{(n)}(x,t|k,\omega),\tag{39}$$

where the precise structure of the operator $\hat{L}^{(n)}$ is determined by the linear differential operator $\mathcal{L}$. Property III in Eq. (38) guarantees nonlinearities of the type $\hat{P}$ can be expanded in the same basis functions $f^{(n)}(x,t|k,\omega)$. If the nonlinearity represented by the operator $\hat{P}$ can also be expanded in the same basis functions $f^{(n)}(x,t|k,\omega)$:

$$Q[\hat{u}(x,t)]=\sum_{n=1}^{\infty}\hat{R}_n f^{(n)}(x,t),\tag{40}$$

then, we can compute the expansion coefficients α_n by solving the relationship

$$\alpha_n=[\hat{L}^{(n)}]^{-1}\hat{R}_n,\tag{41}$$

where α_1 is determined by the initial condition. Of course, we can generalize this result further by replacing Eq. (30) by

$$\mathbf{u}(x,t)=\sum_{\mu_1\cdots\mu_N=1}^{\infty}\hat{\mathbf{C}}(\mu_1\cdots\mu_N)\prod_{i=1}^{N}\prod_{j=1}^{M}f^{(i)}(x,t|\hat{k}_{ij},\hat{\omega}_{ij}).\tag{42}$$

The structure of the solutions proposed in Eq. (42) is, in fact, a generalization of Eq. (30). If we replace $f^{(i)}(x,t|\hat{k}_{ij},\hat{\omega}_{ij})$ by $\exp[i\mu_i k_{ij}z_{ij}]$, the form (30) is retained. Following a similar approach as in the case of Fourier basis functions, we find that if the conditions (38) hold for the solution (42), the linear part of the differential equation acts on the solution (42) like

$$\mathcal{L}[\mathbf{u}(x,t)]=\left(i\mathbf{I}\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\omega}_{ij}+\sum_{n=1}^{K}\mathbf{A}^{(n)}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{k}_{ij}\right]^n\right)\mathbf{u}(x,t),\tag{43}$$

where it is assumed that $\partial_t f^{(i)}(x,t|\hat{k}_{ij},\hat{\omega}_{ij})=\hat{\omega}_{ij}f^{(i)}(x,t|\hat{k}_{ij},\hat{\omega}_{ij})$ and $\partial_x f^{(i)}(x,t)=\hat{k}_{ij}f^{(i)}(x,t|\hat{k}_{ij},\hat{\omega}_{ij})$. This relationship enables us to identify an operator $\hat{\mathbf{L}}^{(ij)}\times(\hat{\omega}_{ij},\hat{k}_{ij})$ according to

$$\hat{\mathbf{L}}^{(ij)}(\hat{\omega}_{ij},\hat{k}_{ij})=\left(i\mathbf{I}\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\omega}_{ij}+\sum_{n=1}^{K}\mathbf{A}^{(n)}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{k}_{ij}\right]^n\right).\tag{44}$$

If we, moreover, assume that the operator Q is of the class $\hat{P}$ so that

$$Q[\mathbf{u}(x,t)]=\sum_{\mu_1\cdots\mu_N=1}^{\infty}\hat{\mathbf{R}}(\mu_1\cdots\mu_N)\prod_{i=1}^{N}\prod_{j=1}^{M}f^{(i)}(x,t|\hat{k}_{ij},\hat{\omega}_{ij}),\tag{45}$$

then the expansion coefficients are determined by the following linear iteration series:

$$\hat{\mathbf{L}}^{(ij)}(\hat{\omega}_{ij},\hat{k}_{ij})\hat{\mathbf{C}}(\mu_1\cdots\mu_N)=\hat{\mathbf{R}}(\mu_1\cdots\mu_N). \tag{46}$$

From this result, we can conclude that we can transform Eq. (28) into Eq. (46). We can conclude that a NPDE of the form of Eq. (28) is integrable if the following two conditions are satisfied.

(i) The nonlinearity must have such a structure that it can be expanded in the same basis functions as the linear part. In other words, the nonlinearity must guarantee that Eq. (45) is satisfied.

(ii) The inverse matrix $\hat{\mathbf{L}}^{(ij)}(\hat{\omega}_{ij},\hat{k}_{ij})$ must exist.

From this result we can conclude that provided a solution (30) exists, the integrability of the NPDE is completely determined by the linear part of the evolution equation. These are also the conditions that guarantee the integrability of Eq. (28). In the following section, we apply these concepts to examine the integrability of some NPDEs.

IV. EXAMPLES

In this section, we will apply the machinery developed in the previous sections to investigate the integrability of various NPDEs. As a first example, we consider the nonlinear Schrödinger equation:

$$i\,\partial_t u=\partial_{xx}u+2uu^*u. \tag{47}$$

If we substitute

$$u(x,t)=e^{iax}e^{i(a^2-b^2)t}e^{i\phi}\sum_{n=1}^{\infty} A_n e^{-n(bx-2abt)}, \tag{48}$$

into Eq. (47), we obtain

$$\sum_{n=1}^{\infty}[(1-n^2)b^2]A_n e^{-n(bx-2abt)}=2\sum_{n=1}^{\infty}\sum_{l=1}^{n-2}\sum_{m=1}^{n-l-1} A_l A_m A_{n-m-l}e^{-n(bx-2abt)}. \tag{49}$$

It can be verified that for $n=1$ the linear dispersion relationship $\omega=-k^2$ $(k=a+bi)$ is satisfied. Since both the left-hand side and the right-hand side can be expanded in the same Fourier basis functions, we can determine the expansion coefficients by the following recursion relationship:

$$L^{(n)}(k)A_n=R^{(n)};\quad k=a+bi, \tag{50}$$

where

$$L^{(n)}=[1-n^2]b^2;\qquad R^{(n)}=2\sum_{l=1}^{n-2}\sum_{m=1}^{n-l-1} A_l A_m A_{n-m-l}. \tag{51}$$

If we assume that $A_1=A$, then by computing all the coefficients A_n, and carrying out the summation, similarly as in Eq. (11), we obtain the NLS soliton:

$$u(x,t)=Ae^{iax}e^{i(a^2-b^2)t}e^{i\phi}e^{\xi_0}\,\mathrm{sech}(bx-2abt+\xi_0),\qquad \xi_0=-\frac{1}{2}\log\left(\frac{A^2}{4b^2}\right). \tag{52}$$

Similarly as for the KdV equation, the two-soliton solution of the nonlinear Schrödinger equation can be computed by considering solutions:

$$u(x,t)=e^{i(a_1+a_2)x}e^{i(a_1^2+a_2^2-b_1^2-b_2^2)t}e^{i\phi}\sum_{n,m=1}^{\infty} C(n,m)e^{-n([b_1+b_2]x-2[a_1b_1+a_2b_2]t)}. \tag{53}$$

By generalizing this procedure, as presented in Sec. III, the N-soliton solution of the nonlinear Schrödinger equation can be computed.

As a second example we consider the coupled nonlinear Schrödinger equation:

$$\begin{aligned} iu_{1t}&=u_{1xx}+(|u_1|^2+|u_2|^2)u_1=0,\\ iu_{2t}&=u_{2xx}+(|u_2|^2+|u_1|^2)u_2=0. \end{aligned} \tag{54}$$

If we make the following substitution for the solution $\mathbf{u}(x,t)=[u_1(x,t),u_2(x,t)]^T$:

$$\mathbf{u}(x,t)=e^{iax}e^{i(a^2-b^2)t}\sum_{n=1}^{\infty} \mathbf{A}_n e^{-n(bx-2abt)}, \quad \mathbf{A}^{(n)}=(A_1^{(n)},A_2^{(n)})^T, \tag{55}$$

into Eq. (54), it can be verified that both the left-hand side and the right-hand side of Eq. (54) can be expanded in the same basis functions. This is due to the fact that both $u_1(x,t)$ and $u_2(x,t)$ have the same dispersion relation $\omega(k)=-k^2$. As a result, we can determine the expansion coefficients $\mathbf{A}^{(n)}$ by solving the following recursion relation:

$$\mathbf{L}^{(n)}(k)\mathbf{A}^{(n)}=\mathbf{R}^{(n)}, \quad k=a+bi, \tag{56}$$

where

$$\mathbf{L}^{(n)}=\mathbf{I}[1-n^2]b^2; \quad \mathbf{R}^{(n)}=\sum_{l=1}^{n-2}\sum_{m=1}^{n-l-1}\begin{pmatrix} A_1^{(l)}A_1^{(m)}A_1^{(n-m-l)}+A_2^{(l)}A_2^{(m)}A_1^{(n-m-l)}\\ A_1^{(l)}A_1^{(m)}A_2^{(n-m-l)}+A_2^{(l)}A_2^{(m)}A_2^{(n-m-l)}\end{pmatrix}. \tag{57}$$

As a last example, we consider the three-dimensional nonlinear Schrödinger equation:

$$i\,\partial_t u=\sum_{n=1}^{3}\partial_{x_n}^2 u+2uu^*u. \tag{58}$$

If we substitute

$$u(\mathbf{x},t)=e^{i\mathbf{a}\cdot\mathbf{x}}e^{i(\mathbf{a}\cdot\mathbf{a}-\mathbf{b}\cdot\mathbf{b})t}e^{i\phi}\sum_{n=1}^{\infty} A_n e^{-n(\mathbf{b}\cdot\mathbf{x}-2\mathbf{a}\cdot\mathbf{b}t)}, \tag{59}$$

into Eq. (47), we obtain

$$\sum_{n=1}^{\infty}[(1-n^2)\mathbf{b}\cdot\mathbf{b}]A_n e^{-n(\mathbf{b}\cdot\mathbf{x}-2\mathbf{a}\cdot\mathbf{b}t)}=2\sum_{n=1}^{\infty}\sum_{l=1}^{n-2}\sum_{m=1}^{n-l-1} A_lA_mA_{n-m-l}e^{-n(\mathbf{b}\cdot\mathbf{x}-2\mathbf{a}\cdot\mathbf{b}t)}. \tag{60}$$

In Eq. (59) and Eq. (60), it is used that $\mathbf{x}=(x_1,x_2,x_3)^T$, $\mathbf{a}=(a_1,a_2,a_3)^T$, and $\mathbf{b}=(b_1,b_2,b_3)^T$. It can be verified that for $n=1$ the linear dispersion relationship $\omega^2=-\mathbf{k}\cdot\mathbf{k}(\mathbf{k}=\mathbf{a}+\mathbf{b}i)$ is satisfied. Since both the left-hand side and the right-hand side can be expanded in the same Fourier basis functions, we can determine the expansion coefficients by the following recursion relationship:

$$L^{(n)}(\mathbf{k})A_n=R^{(n)}; \quad \mathbf{k}=\mathbf{a}+\mathbf{b}i, \tag{61}$$

where

$$L^{(n)}(\mathbf{k})=[1-n^2]\mathbf{b}\cdot\mathbf{b};\quad R^{(n)}=2\sum_{l=1}^{n-2}\sum_{m=1}^{n-l-1}A_lA_mA_{n-m-l}. \tag{62}$$

Similarly as in the one-dimensional case, explicit solutions of the three-dimensional nonlinear Schrödinger equation can be obtained by carrying out the summation of the expansion coefficients. The discussion can be made more general by using other expansion functions, similarly as in Eq. (53).

V. DISCUSSION AND CONCLUSIONS

We have presented a method to investigate the integrability for NPDEs having a polynomial type of nonlinearity. It has to be remarked that we have assumed throughout this paper that integrability is equivalent with the existence of N solitons. It is shown that two conditions play an important role. The first condition is that the nonlinearity can be expanded in the same basis functions as the linear part. The second condition is that the linearized part of the NPDE has nontrivial solutions. The method is presented by investigating the integrability of the KdV equation as an example. In Sec. III the method is first generalized for NPDEs having solutions that can be expanded in an infinite set of Fourier basis functions. Later on, it is shown that we do not have to restrict ourselves to Fourier basis functions only. Moreover, it is likely that the method also works for nonpolynomial types of nonlinearity, at least if the nonlinearity can be expanded in polynomial form. The paper is concluded by applying the method on the nonlinear Schrödinger equation, the coupled nonlinear Schrödinger equation, and a three-dimensional example. It is shown that we can derive special solutions of the three-dimensional nonlinear Schrödinger equation.

There is an interesting link between the work carried out in this paper and Hirota's method[5,9] in which it is shown for the KdV equation that by applying the transformation

$$u=2(\log F)_{xx}, \tag{63}$$

the solution F can be written as

$$F(x,t)=\det|M|, \tag{64}$$

where the $N\times N$ matrix M has the entries

$$M_{ij}(x,t)=\delta_{ij}+\frac{2(P_iP_j)^{1/2}}{P_i+P_j}e^{(1/2)(\xi_i+\xi_j)};\quad \xi_i=P_ix-P_i^3t-\xi_i^0, \tag{65}$$

and P_i and ξ_i^0 are arbitrary constants. The result presented above was obtained by assuming that the solution $F(x,t)$ can be expanded in a similar series that formed the starting point in this paper:

$$F(x,t)=1+F^{(1)}(x,t)+F^{(2)}(x,t)+\cdots. \tag{66}$$

The major difference between the method presented in this paper and Hirota's method is that the latter succeeded to formulate solutions of the KdV equation by using a finite number of functions $F^{(N)}(x,t)$, whereas our method needs an infinite number of basis functions. It is also interesting to mention that the solutions obtained by Hirota have a similar structure as inverse scattering solutions for rational reflection coefficients as obtained by Sabatier.[10] Moreover, in Ref. 11 it is shown for the KdV equation that Fourier expansion of the inverse scattering solutions as derived by Sabatier is equal to the series (11). The solutions derived in this paper can therefore be regarded as a Fourier expansion of Hirota's solution.

ACKNOWLEDGMENTS

An anonymous referee is gratefully thanked for the comments made on the manuscript. This research was supported by the Netherlands Organization for Scientific Research (N.W.O.) through the ''N.R.C. Photonics'' grant.

[1] V. E. Zakharov, in *What is Integrability* (Springer-Verlag, New York, 1990).
[2] F. Calogero and W. Eckhaus, ''Nonlinear evolution equations, rescaling model PDEs and their integrability. I & II,'' Inverse Probl. **3**, 229–262 (1987); **4**, 11–33 (1988).
[3] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[4] G. G. Stokes, ''On the theory of oscillatory waves,'' Cambridge Trans. **8**, 441–473 (1847).
[5] G. B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974).
[6] R. R. Rosales, ''Exact solutions of some nonlinear evolution equations,'' Stud. Appl. Math. **59**, 117–151 (1978).
[7] M. Wadati and K. Sawada, ''New representations of the soliton solution of the Korteweg-de Vries equation,'' J. Phys. Soc. Jpn. **48**, 312–318 (1980).
[8] H. J. S. Dorren, ''A linearizing transformation for the Korteweg-de Vries equations; generalizations to higher-dimensional nonlinear partial differential equations,'' J. Math. Phys. **39**, 3711–3729 (1998).
[9] R. Hirota, ''Exact solution of the Korteweg-de Vries equation for multiple collisions of solitons,'' Phys. Rev. Lett. **27**, 1192 (1971).
[10] P. C. Sabatier, ''Rational reflection coefficients and inverse scattering on the line,'' Nuovo Cimento B **78**, 235–248 (1983).
[11] H. J. S. Dorren and R. K. Snieder, ''On the stability of inverse problems,'' Inverse Probl. **11**, 889–911 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Isospectral problem for Schrödinger operator: Evolutional viewpoint

V. M. Eleonsky and V. G. Korolev[a)]
Lukin Research Institute of Physical Problems, Zelenograd, Moscow, Russia, 103460

(Received 15 May 1998; accepted for publication 7 December 1998)

An isospectral transform of the Schrödinger operator is considered as an evolutional problem. For a transform defined by the McKean–Trubowitz flows associated evolutional equations are derived. It is shown that for one-level and two-level flows these equations can be split into integrable Liouville equations. A relationship between the Liouville equations and the Darboux transforms is discussed; this analysis suggests that the evolutional equations can be split into the Liouville equations in the general case. A Hamiltonian formulation of the isospectral transform defined by the McKean–Trubowitz flows is presented. It is shown that this transform is performed by a canonical change of variables, which is related to the Darboux transform. © *1999 American Institute of Physics.*
[S0022-2488(99)02403-2]

I. INTRODUCTION

Any progress in a study of isospectral transforms for the Schrödinger operator always gives rise to particular interest. On the one hand, such transforms provide a tool to extend a class of models for which the Schrödinger equation can be solved exactly; this explains the special interest of isospectral transforms among physicists. For mathematicians, to develop a theory of the Sturm–Liouville operator (and for them, the Schrödinger operator is nothing but its particular case) always was a generous classical challenge. In some branches of mathematics (i.e., in the theory of the Korteweg–de Vries equation) the Schrödinger operator plays an auxiliary role; its isospectral properties are related with an existence of soliton solutions to evolutional equations.[1]

The Darboux transform[2] is known as the most popular isospectral transform of the Schrödinger operator. Since Darboux's earliest papers,[3] this method was extensively developed;[4,5] it is mostly used in building new operators whose spectrum differs from the spectrum of the initial Schrödinger operator by one eigenvalue.

The Darboux transform of the Schrödinger operator is closely related with a number of modern approaches, such as the factorization method[6–8] (which, in its turn, is a basis for the theory of supersymmetry in physics[9,10]); the method of dressing chains;[11] the method of nonlinear spectrum shift operators,[12] etc.

There exists another approach to the isospectral problem; it involves an investigation of isospectral phase flows. This branch of the problem goes back to the theory of integrable ''soliton'' equations and is related with a Hamiltonian description of the evolution that is governed by such equations. Comparatively recently, an important step was made in this direction: there appeared the paper by McKean and Trubowitz;[13] the authors studied phase flows of the form

$$U_t=\sum_n \tau_n[\psi_n^2]_x; \quad \tau_n\in\mathcal{R} \tag{1}$$

(ψ_n are eigenfunctions of discrete states); it was shown that in this case the deformation of the potential as well as of the eigenfunctions is isospectral; the flows (1) were analyzed from the

[a)]Electronic mail: korolev@nonlin.msk.ru viewpoint of symplectic geometry in associated function spaces. Later, the results exposed in this paper were made more strict, generalized and extended by Levitan in Ref. 14.

Note that in both the Darboux transform and the McKean and Trubowitz approach *closed evolutional equations* (i.e., partial differential equations where one of the partial derivatives is in the deformation parameter t) for eigenfunctions as well as for a potential are not derived. So the solutions for the isospectral deformation of eigenfunctions and a potential were not looked for as a result of the integration of such evolutional equations. This (quite natural) problem was formulated and solved in our paper.[15] There we derived coupled evolutional equations of the Liouville type that govern the isospectral deformation defined by the flows (1). In the simplest case, when the sum in (1) contains the only term, those equations degenerate into an integrable Liouville equation for the evolutions both of the eigenfunction that ''specifies'' the flow and of the potential itself; and into d'Alembert equations for the other eigenfunctions. The simple asymptotical analysis performed in Ref. 15 showed that such an evolution is accompanied by splitting off a local potential well, which asymptotically gains the form of the soliton reflectionless potential $\gamma^2/\cosh^2[\gamma\xi]$; and this scenario holds for a wide class of potentials, including scattering potentials as well as the potential of the harmonic (and anharmonic) oscillator.

In the present paper we develop the ''evolutional approach'' to the isospectral problem using the flow (1) in the following directions.

(i) We consider a generalization of this flow in the form

$$U_t=\sum_k \tau_k[\psi_k\tilde{\psi}_k]_x, \tag{2}$$

where $\psi_k(x,t)$ and $\tilde{\psi}_k(x,t)$ are a pair of solutions to the Schrödinger equation with the parameter E_k, but only ψ_k must be an eigenfunction. For this case we derive proper PDEs for an evolution of the eigenfunctions of the Schrödinger operator.

(ii) Returning to the original McKean–Trubowitz flows we study a ''two-level'' flow [in the expression (1) two constants τ_n are nonzero] in detail. It is found out that an associated pair of evolutional equations splits into *two integrable Liouville equations* for certain combinations of the eigenfunctions for those states and of their derivatives in x. The integration of the evolutional equation for the potential provides a solution, which can be rewritten as the well-known formula for the isospectral deformation of the potential that arises in the Gelfand–Levitan–Marchenko theory. It seems likely that the conclusion about the splitting of the coupled evolutional equation into independent integrable Liouville equations can be extended to the general case of arbitrary multi-level flows (1).

(iii) We show that the Darboux transform is also directly related with the integrable Liouville equations and the McKean and Trubowitz ''flow'' approach.

(iv) Finally, in the case of two-level flow, we present a Hamiltonian formulation of the isospectral evolution that obeys the Liouville equations; we show that splitting of the initial coupled equations into the Liouville equations is performed by a canonical change of variables.

The paper is arranged as follows: in Sec. II we discuss the role of the McKean–Trubowitz flows in the isospectral problem; in Sec. III we derive evolutional equations for the generalization of the McKean–Trubowitz flows. In Sec. IV we present a brief review of the results related to the simplest flow (''individual,'' in the terminology of Ref. 13). In Sec. V we study evolutional equations for the two-level flow; we discuss relations between the Darboux transform and the Liouville equations and between them both and the McKean–Trubowitz flows. Finally, in Sec. VI we develop a Hamiltonian description for the two-level case.

II. ON A NOTION OF ISOSPECTRAL FLOWS. THE McKEAN–TRUBOWITZ FLOWS

To begin with, we would like to say a few common words concerning the ''flow'' approach to the Schrödinger isospectral problem; this will help us in clearing up a role of the flows studied by McKean and Trubowitz in the general context of the isospectral problem.

Consider the Sturm–Liouville problem for the 1D Schrödinger operator in $\mathcal{R}^1$:

$$-\tfrac{1}{2}(\psi_n)_{xx}+U\psi_n=E_n\psi_n, \quad n\geqslant 0, \quad \lim_{x\to\pm\infty}\psi_n=0. \tag{3}$$

Let U and ψ_n be functions of parameter t (it is a deformation parameter and should not be confused with the physical time: we deal with a *stationary* Schrödinger equation). A dependence of the eigenelements of the problem (3) on the parameter t will be specified by the following relation:

$$U_t=F(U,U_x,\ldots;\{\psi_k\},\{(\psi_k)_x\},\ldots). \tag{4}$$

If the right-hand side of this relation can be represented as a gradient (in x) of a variational derivative for a certain functional of the potential U, then such a relation has a meaning of *a phase flow*. Since we will mostly deal with relations (4) that admit of this representation, we will say that relation (4) specifies a phase flow.

Demand that the eigenvalues E_n do not vary as the potential deforms. As follows from (3), the conditions,

$$(E_n)_t=0, \quad \forall t; \quad n\geqslant 0, \tag{5}$$

lead to the system of equations,

$$\int_{-\infty}^{+\infty} U_t\psi_n^2=0, \quad n\geqslant 0. \tag{6}$$

Hence, we get a *necessary* condition for a flow of type (4) to be *isospectral* (to preserve the spectrum of the Schrödinger operator):

$$\int_{-\infty}^{+\infty} F\psi_n^2=0, \quad n\geqslant 0. \tag{7}$$

In the other words, for any isospectral flow the functional F is to belong to the orthogonal complement to the set $\{\psi_n^2\}$.

Return to the Schrödinger equation (3). Differentiating it in x gives us the following equations:

$$\tfrac{1}{8}(\psi_n)^2_{xxx}-\tfrac{1}{2}U_x\psi_n^2-(U-E_n)(\psi_n^2)_x=0, \quad n\geqslant 0, \tag{8}$$

which can be rewritten in the form

$$\mathbf{L}[\psi_n^2]_x=E_n[\psi_n^2]_x, \quad n\geqslant 0, \tag{9}$$

if one introduces the integro-differential operator,

$$\mathbf{L}\equiv -\tfrac{1}{8}D^2-\tfrac{1}{2}U_xD^{-1}+U, \tag{10}$$

where $Dg(x)\equiv dg/dx$, $D^{-1}g(x)\equiv\int_x^{+\infty}g(x')dx'$ (this remark is usually assigned to Hermite; see, for instance, the monograph[16]).

The operator $\mathbf{L}$ has two important properties (quite a large number of them is listed in the book[17]).

(1) The functions $[\psi_n^2]_x$ are its eigenfunctions, which correspond to the eigenvalues E_n; this fact is expressed by formula (9). So, if the set of pairs $\{(E_m,\psi_m)\}$ is a set of eigenelements for the operator $H=-(1/2)D^2+U$, then the set $\{(E_m,[\psi_m^2]_x)\}$ is a set of eigenelements for the operator $\mathbf{L}$. Note that $\mathbf{L}$ is not a Hermitian operator, so the set $[\psi_n^2]_x$ is not to be complete.

(2)

$$\int_{-\infty}^{+\infty} \psi_n^2 \mathbf{L} f(x)\, dx = E_n \int_{-\infty}^{+\infty} \psi_n^2 f(x)\, dx, \quad n \geqslant 0, \tag{11}$$

for quite a wide class of functions $f(x)$ (this class includes functions that grow at infinity slower than the exponential). This property implies that if some relation $U_t = F$ determines an isospectral flow, then all the flows $U_t = \mathbf{L}^k F$, $k > 0$ are also isospectral: each of them satisfies the condition (6). The last feature of the operator $\mathbf{L}$ is widely used when one builds hierarchies of integrable evolutional equations, and is the reason to call it *a recursion operator*.

Note that the operator $\mathbf{L}$ can be written in the following symmetric form:

$$\mathbf{L} = -\tfrac{1}{8} D^2 - \sqrt{U} D \sqrt{U} D^{-1} \tag{12}$$

(here U is assumed to be non-negative; but, obviously, the formula can be generalized).

Now let us return to the isospectral problem and compare two simple isospectral flows. The first of them is a so-called ''shift flow:''

$$U_t = c U_x = \frac{c}{2} \frac{\partial}{\partial x} \frac{\delta \Lambda}{\delta U}; \quad \Lambda = \int dx U^2. \tag{13}$$

Really,

$$\int_{-\infty}^{+\infty} U_x \psi_n^2\, dx = 0, \tag{14}$$

in the same class of potentials that grow at infinity slower than the exponential; this simple result immediately follows from Eq. (3). Usually, however, to avoid a singularity in t, one demands a ''good behavior'' of U_x at infinity (in particular, for scattering potentials $U_x \to 0$ as $x \to \pm\infty$).

Sequentially applying the recursion operator $\mathbf{L}$ to the shift flow (13), one builds a hierarchy of KdV flows:

$$\begin{gathered} k=0: \quad U_t = c_0 U_x, \\ k=1: \quad U_t = c_1 (U_{xxx} + 6(U^2)_x), \\ \cdots \\ k=m: \quad U_t = c_m \mathbf{L}^m U_x. \\ \cdots \end{gathered}$$

Let us stress again that all these flows lie in the orthogonal complement to the set $\{\psi_n^2\}$.

The second elementary type of isospectral flows is given by the expression

$$U_t = \sum_n \tau_n [\psi_n^2]_x; \quad \tau_n \in \mathcal{R}; \tag{15}$$

actually, it can be shown that

$$\int_{-\infty}^{+\infty} \psi_n^2 [\psi_m^2]_x\, dx = 0, \quad \forall n,m \geqslant 0, \tag{16}$$

by virtue of Eq. (3); in the other words, the sets $\{\psi_n^2\}$ and $\{[\psi_n^2]_x\}$ lie in orthogonal complements. The flows (15) and their isospectral property were studied in detail in Ref. 13; in what follows they will be referred to as ''the McKean–Trubowitz flows'' or simply ''MKT flows.''

What sets off the MKT flows among other isospectral flows? Note two their important properties.

(a) The functional F that determines the MKT flow does not depend on the potential (only on the eigenfunctions, which always behave well). So these flows exist both for scattering potentials and for potentials *growing at infinity* (e.g., for the potential of the harmonic oscillator).

(b) As it follows from the first property of the recursion operator $\mathbf{L}$, the functions $[\psi_n^2]_x$ are its eigenfunctions. It means that it is impossible to build an $\mathbf{L}$-hierarchy starting with the MKT flow:

$$\mathbf{L}\sum_n \tau_n[\psi_n^2]_x=\sum_n \tau_n\mathbf{L}[\psi_n^2]_x=\sum_n \tau_n'[\psi_n^2]_x. \tag{17}$$

Thus, the operator $\mathbf{L}$ simply changes the weights τ_n, but the flow is related with the linear envelope of the set $\{[\psi_n^2]_x\}$, as before. On the contrary, the elements of a functional space that are associated with different flows of the KdV hierarchy, are sequentially mapped onto each other by the operator $\mathbf{L}$. One can say loosely that the MKT flows are ''eigenflows'' of the recursion operator $\mathbf{L}$; this, in particular, explains an interest in a deeper study of their features.

Note that these two elementary isospectral flows can coincide: as is well known (see, for example, Ref. 17), the reflectionless soliton potentials can be represented in the form

$$U=2\sqrt{2}\sum_{n=0}^{N} \sqrt{-E_n}\psi_n^2, \tag{18}$$

where the sum is taken over all the discrete states of the Schrödinger operator. So, if we specify the weights τ_n by the relations $\tau_n=c\sqrt{-E_n}$, then the MKT flow with these weights reduces to the simplest shift flow (the first flow of the KdV hierarchy) for those potentials; hence, each of the higher k-th flows of the KdV hierarchy coincides with the MKT flow with properly changed weights:

$$\tau_n^{(k)}=c\sqrt{-E_n}E_n^k. \tag{19}$$

It is clear that the expressions for the KdV flows are evolutional equations for the potential U themselves. The situation with the MKT flows is more complex. In the paper[15] we showed that in the case of ''individual'' MKT flow (all $\tau_n=0$ but one) the deformation both of the potential and of the eigenfunctions is related to the integrable Liouville equation $S_{x\tau}=\exp 2S$. In the present paper we show that an analogous statement holds for more complex MKT flows; moreover, everything indicates that this is true for any McKean–Trubowitz flow.

III. EQUATIONS OF ISOSPECTRAL DEFORMATION FOR SCHRÖDINGER OPERATORS: GENERAL CASE

Let us supplement the Schrödinger problem (3) by the relation

$$U_t=\sum_k \tau_k(\psi_k\tilde{\psi}_k)_x, \tag{20}$$

where $\psi_k(x,t)$ and $\tilde{\psi}_k(x,t)$ are a pair of solutions to the Schrödinger equation with the parameter E_k and $\{\psi_k,E_k\}$ is an eigenelement:

$$\int_{-\infty}^{+\infty}\psi_k^2(x',t)dx'<\infty, \quad \forall t. \tag{21}$$

Evidently, relation (20) is an analog and a generalization of Eq. (15), which determines MKT flows.

Linear independence of the functions ψ_k and $\tilde{\psi}_k$ is characterized by the Wronskian,

$$w_k(t)=[\psi_k,\tilde{\psi}_k]\equiv\psi_k(\tilde{\psi}_k)_x-(\psi_k)_x\tilde{\psi}_k. \tag{22}$$

In the case $w_k(t)\neq 0$ we do not know a representation of the relation (20) in the form of a phase flow (with a gradient of a variational derivative on the right-hand side).

Let us derive equations that determine an evolution of the eigenelements $\{\psi_m,E_m\}$, $m=0,1,...$, of the Schrödinger operator under the condition (20). Differentiating Eq. (3) in t and substituting the right-hand side of Eq. (20) for U_t, we obtain the system of equations:

$$\psi_m^2(E_m)_t+\frac{1}{2}[\psi_m(\psi_m)_{xt}-(\psi_m)_x(\psi_m)_t]_x=\sum_k \tau_k(\psi_k\tilde{\psi}_k)_x, \quad m=0,1,\dots . \tag{23}$$

Now we take into account the following identities:

$$\psi_m^2(\psi_k\tilde{\psi}_k)_x\equiv\tfrac{1}{2}(\psi_m^2\psi_k\tilde{\psi}_k)_x+\tfrac{1}{2}\{\psi_m^2(\psi_k\tilde{\psi}_k)_x-(\psi_m^2)_x\psi_k\tilde{\psi}_k\}, \tag{24}$$

and

$$\psi_m^2(\psi_m\tilde{\psi}_m)_x-(\psi_m^2)_x\psi_m\tilde{\psi}_m\equiv\psi_m^2 w_m(t),$$

$$\psi_m^2(\psi_k\tilde{\psi}_k)_x-(\psi_m^2)_x\psi_k\tilde{\psi}_k\equiv\frac{\{[\psi_m,\psi_k]\cdot[\psi_m,\tilde{\psi}_k]\}_x}{2(E_m-E_k)}, \quad k\neq m.$$

Then the system (23) can be written as follows:

$$\left\{(E_m)_t-\frac{1}{2}\tau_m w_m(t)\right\}\psi_m^2=-\frac{1}{2}[\psi_m(\psi_m)_{xt}-(\psi_m)_x(\psi_m)_t]_x+\frac{1}{2}\sum_k \tau_k[\psi_m\psi_k\tilde{\psi}_k]_x$$
$$+\frac{1}{2}\sum_{k\neq m}\frac{\tau_k}{2(E_m-E_k)}[\psi_m,\psi_k][\psi_m,\tilde{\psi}_k]. \tag{25}$$

Integrating Eqs. (25) in $x\in\mathcal{R}$, we eventually find that the evolution of the eigenvalues E_m and the eigenfunctions ψ_m obeys the following equations:

$$(E_m)_t=\tfrac{1}{2}\tau_m w_m(t), \quad m=0,1,..., \tag{26}$$

$$\left[\frac{(\psi_m)_x}{\psi_m}\right]_t=\sum_k \tau_k\psi_k\tilde{\psi}_k+\sum_{k\neq m}\frac{\tau_k}{2\psi_m^2(E_m-E_k)}[\psi_m,\psi_k][\psi_m,\tilde{\psi}_k]. \tag{27}$$

Thus, for all the states with ''weights'' $\tau_m\neq 0$ the evolution of the eigenvalues $E_m(t)$ is described by the simple formula:

$$E_m(t)=E_m^0+\frac{\tau_m}{2}\int_0^t w_m(t')dt'. \tag{28}$$

In the other words, only those eigenvalues move, for which $w_m(t)\neq 0$ [i.e., the relation (20) contains a corresponding pair of linearly *independent* functions]. Eigenvalues are constant if $w_m(t)=0$. All the other eigenvalues (for which $\tau_m=0$) are also invariant [regardless of $w_m(t)$, naturally].

IV. ONE-LEVEL ISOSPECTRAL DEFORMATION

Let us present a brief review of results for the simplest case: the relation (20) contains only one product of the functions ψ_m, $\tilde{\psi}_m$; the detailed analysis, including a study of asymptotical behavior both of the eigenfunctions and of the potential, together with a description of the one-level isospectral deformation in terms of explicit unitary operators, was presented in our paper.[15]

Let all $\tau_m=0$ except for the only $\tau_n=1$. In this case the general equations (26), (27) split into separate equations for the evolution of the eigenelement (ψ_n, E_n) and for the evolution of the other eigenelements $\{(\psi_m, E_m), m\neq n\}$:

$$(E_n)_t=\tfrac{1}{2}w_n(t),\quad (E_m)_t=0,\quad m\neq n;\quad (\mathrm{D}_{\log}\psi_n)_t=\psi_n\tilde{\psi}_n,$$

$$(\mathrm{D}_{\log}\psi_m)_t=\psi_n\tilde{\psi}_n+\frac{[\psi_m,\psi_n][\psi_m,\tilde{\psi}_n]}{2\psi_m^2(E_m-E_n)},\quad m\neq n$$

[for brevity, hereafter we use the operator of ''logarithmic derivative'' $\mathrm{D}_{\log}$: $\mathrm{D}_{\log}G(x)\equiv G_x/G \equiv(\ln|G|)_x$].

In the case $\tilde{\psi}_n=\psi_n$ (''individual'' MKT flow) we have $w_n(t)=0$, and the above equations take the form

$$(E_n)_t=(E_m)_t=0,\quad (\mathrm{D}_{\log}\psi_n)_t=\psi_n^2, \tag{29}$$

$$(\mathrm{D}_{\log}\psi_m)_t=\psi_n^2+\frac{[\psi_m,\psi_n]^2}{2\psi_m^2(E_m-E_n)},\quad m\neq n. \tag{30}$$

Equation (29), which determines evolution (in the parameter t) of the ''specifying'' eigenfunction ψ_n can be rewritten in the classical form of the *integrable Liouville equation* after the substitution $\psi_n\rightarrow\exp S_n$. It has the following solution:

$$\psi_n(x,t)=\frac{\psi_n^0}{1+(e^t-1)\int_x^{+\infty}[\psi_n^0]^2dx'}. \tag{31}$$

Here the t-function, which is arbitrary in the general case, is chosen exponential to keep the norm of the eigenfunction.

After a few transforms, the equations (30) for the other eigenfunctions can be written in the simple form

$$\frac{\partial}{\partial t}\mathrm{D}_{\log}\frac{\psi_m}{[\psi_n,\psi_m]}=0. \tag{32}$$

They are integrated in an obvious way; this allows us to write down the evolution of any function $\psi_m(x,t)$ if we know a solution for the ''main'' eigenfunction $\psi_n(x,t)$.

Note that by virtue of the Schrödinger operator the Wronskian of these functions can always be expressed through the integral of their product, and *vice versa*:

$$[\psi_n,\psi_m]=(E_n-E_m)\int_x^{\infty}\psi_n\psi_m dx'; \tag{33}$$

this relation can be used if convenient.

Of course, formula (31) as well as the solutions for the other eigenfunctions ψ_m, $m\neq n$ are known in the literature for quite a long time; they were also presented by McKean and Trubowitz in Ref. 13; we have obtained these results in the framework of the evolutional approach, considering the isospectral deformation as an evolution in the parameter t; deriving associated PDEs and integrating them.

Return to the generalization of the individual flow. Let $\tilde{\psi}_n \neq \psi_n$ $[w_n(t) \neq 0]$. Expressing $\tilde{\psi}_n$ from Eq. (20) and substituting it into the equation for the evolution of ψ_n, we get

$$(\mathrm{D}\log\psi_n)_t = \psi_n^2\left(1 + w_n(t)\int^x \frac{dx'}{\psi_n^2(x',t)}\right). \tag{34}$$

On the other hand, expressing $\tilde{\psi}_n$ from the equation for the evolution of ψ_n and substituting it into the formula for the flow (20), we obtain an equation, which is convenient to rewrite introducing the function $\phi_n \equiv 1/\psi_n$:

$$\left[\frac{(\phi_n)_{xx}}{\phi_n}\right]_t = -w_n(t); \tag{35}$$

it can be integrated in t; as a result we find

$$\frac{(\phi_n)_{xx}}{\phi_n} - \frac{(\phi_n^0)_{xx}}{\phi_n^0} = -2(E_n - E_n^0). \tag{36}$$

If ψ_n^0 is given, then this is the Schrödinger equation for ϕ_n. Taking into account that $(\psi_n^0)_{xx}/\psi_n^0 = 2(U^0 - E_n^0)$, we can rewrite it in the form

$$-\frac{1}{2}(\phi_n)_{xx} + \left\{U^0 - \frac{d^2}{dx^2}\ln|\psi_n^0|\right\}\phi_n = E_n\phi_n. \tag{37}$$

This formula clearly witnesses that in the case of individual flow defined by a pair of linearly independent functions, the integration of the associated evolutional equations leads to the same results as does the Darboux transform.

Note also that the functions $\psi_n(x,t)$ can be expressed through solutions of the Schrödinger equations at $t=0$ (this is especially important if those solutions are known explicitly):

$$\psi_n(x,t) = \frac{\psi_n^0}{[\psi_n^0, \chi^0]}; \tag{38}$$

here χ^0 is a solution of the following Schrödinger equation:

$$-\tfrac{1}{2}(\chi^0)_{xx} + U^0\chi^0 = E_n(t)\chi^0 \tag{39}$$

[with the initial potential U^0, but for the value $E_n(t)$ already shifted].

We stress that all the formulas related to the case $w_n(t) \neq 0$ directly transform to the corresponding formulas for the case $w_n(t) = 0$; except for expressions (38), (39): when they are derived, the assumption $E_n(t) - E_n^0 \neq 0$ is used. Note also that the above formulas for the generalized MKT flow were also obtained by McKean and Trubowitz in their approach.[13]

Up to this point we dealt with equations that describe a deformation of *eigenfunctions* of the Schrödinger operator. Using them, one can derive an equation that governs a deformation of *the potential* $U(x,t)$ for the generalization of the individual flow:

$$\frac{\partial}{\partial t}\left[\mathrm{D}_{\log} Y - \frac{w_n(t)}{Y}\right] = 2Y, \tag{40}$$

where

$$Y=\frac{\partial}{\partial t}\int^{x}U(x',t)dx'. \tag{41}$$

If $w_n(t)=0$ it also degenerates into the Liouville equation. In this case the analysis of the asymptotical behavior for $U(x,t)$ shows,[15] that for a wide class of potentials, such a deformation involves the formation of moving local wells in the potential relief; their form is asymptotically close to the reflectionless soliton potential:

$$\Delta U(x\sim x_0(t);t\to\infty)\approx-\frac{2(U^0(x_0)-E_n)}{\cosh^2\sqrt{2(U^0(x_0)-E_n)}(x-x_0(t))}, \tag{42}$$

where the ''center'' $x_0(t)$ of the moving soliton well is determined by the equation

$$\int^{x_0}\sqrt{2(U^0(x')-E_n)}dx'=\frac{t}{2}. \tag{43}$$

The eigenfunction that defines the flow (for which $\tau_n\neq 0$) is localized in this well; as $t\to\infty$ the domains of localization move to infinity, which is related with disappearance of the corresponding state from the spectrum. This process is accompanied by a deformation of all the other functions in such a way that the number of nodes and zeros for the other eigenfunctions corresponds to the new spectrum after the ''governing'' state is removed.

It is important that such a scenario (a splitting off of local potential wells of soliton form) is realized not only for the scattering potentials but also for the potential of *the harmonic oscillator*.

V. ANALYSIS OF ISOSPECTRAL EVOLUTION FOR TWO-LEVEL FLOW

Now consider a case of the ''two-level'' MKT flow. Assume that in the relation (20) τ_n, $\tau_l\neq 0$ for two states, and for all the other states $\tau_m=0$, $m\neq n,l$; let also $w_m=0$, $m=0,1,\ldots$. Since $w_n=w_l=0$, the evolution of the system is not accompanied by a shift of the eigenvalues E_n, E_l (as we know, for the other states eigenvalues do not change in any case). Then Eqs. (26), (27), which describe a deformation of the eigenfunctions of the Schrödinger problem, take the form

$$(\mathrm{D}_{\log}\psi_l)_t=\tau_n(\psi_n)^2+\tau_l(\psi_l)^2-\frac{\tau_n}{2\Delta_{nl}}\left\{\frac{[\psi_l,\psi_n]}{\psi_l^2}\right\}^2, \tag{44}$$

$$(\mathrm{D}_{\log}\psi_n)_t=\tau_n(\psi_n)^2+\tau_l(\psi_l)^2+\frac{\tau_l}{2\Delta_{nl}}\left\{\frac{[\psi_l,\psi_n]}{\psi_n^2}\right\}^2, \tag{45}$$

where $\Delta_{nl}=E_n-E_l$.

For the convenience of the further analysis, rewrite this pair of coupled evolutional equations as follows:

$$(\mathrm{D}_{\log}\psi_l)_t=\Sigma-\frac{\tau_n}{2\Delta_{nl}}\left(\psi_n\,\mathrm{D}_{\log}\left[\frac{\psi_n}{\psi_l}\right]\right)^2, \tag{46}$$

$$(\mathrm{D}_{\log}\psi_n)_t=\Sigma+\frac{\tau_l}{2\Delta_{nl}}\left(\psi_l\,\mathrm{D}_{\log}\left[\frac{\psi_n}{\psi_l}\right]\right)^2; \tag{47}$$

here we denoted $\Sigma\equiv\tau_n\psi_n^2+\tau_l\psi_l^2$.

Subtracting these equations, we get

$$\left(\mathrm{D}_{\log}\left[\frac{\psi_n}{\psi_l}\right]\right)_t=\frac{\Sigma}{2\Delta_{nl}}\left(\mathrm{D}_{\log}\left[\frac{\psi_n}{\psi_l}\right]\right)^2, \tag{48}$$

so

$$\Sigma=-\frac{\partial}{\partial t}\left\{\frac{2\Delta_{nl}}{\mathrm{D}_{\log}\left[\frac{\psi_n}{\psi_l}\right]}\right\}. \tag{49}$$

Furthermore, from a pair of Schrödinger equations for the functions ψ_n, ψ_l for some potential U follows:

$$\frac{1}{2}\frac{(\psi_n)_{xx}}{\psi_n}+E_n=\frac{1}{2}\frac{(\psi_l)_{xx}}{\psi_l}+E_l, \tag{50}$$

which we rewrite in the form

$$[\psi_n,\psi_l]_x=2\Delta_{nl}\psi_n\psi_l. \tag{51}$$

Note that this relation must be satisfied by any pair of eigenfunctions of the Schrödinger operator for which the difference between eigenvalues equals Δ_{nl}. In the other words, for the equations that determine the evolution in t the expression (51) is an additional condition, which is to hold at any t; *a priori* it is not known whether this condition is substantial or holds always by virtue of the equations themselves (i.e., in some sense, it is a *a local integral* of those equations).

Dividing both sides of Eq. (51) by $[\psi_n,\psi_l]$, and comparing the result with Eq. (49), we find that

$$\Sigma=\frac{\partial}{\partial t}\mathrm{D}_{\log}[\psi_n,\psi_l]\equiv a(x,t). \tag{52}$$

If we return to a definition of the flow ($U_t=\Sigma_x$), we notice that the function $a(x,t)$ that denotes the right-hand side of the relation (52) coincides with the antiderivative of the potential $V(x,t)$: $V_x=U$ up to an arbitrary term depending on x only.

Substituting the expression for Σ from (52) into the coupled equations (44)–(45), we obtain two *separated* equations for the functions $\{[\psi_n,\psi_l]/\psi_l\}^2$ and $\{[\psi_n,\psi_l]/\psi_n\}^2$:

$$\frac{\partial}{\partial t}\mathrm{D}_{\log}\left\{\frac{[\psi_n,\psi_l]}{\psi_l}\right\}^2=+\frac{\tau_n}{\Delta_{nl}}\left\{\frac{[\psi_n,\psi_l]}{\psi_l}\right\}^2, \tag{53}$$

$$\frac{\partial}{\partial t}\mathrm{D}_{\log}\left\{\frac{[\psi_n,\psi_l]}{\psi_n}\right\}^2=-\frac{\tau_l}{\Delta_{nl}}\left\{\frac{[\psi_n,\psi_l]}{\psi_n}\right\}^2. \tag{54}$$

Thus, for the McKean–Trubowitz flow defined at two eigenfunctions of the Schrödinger operator, the evolution of these combinations obeys two *independent* integrable equations that reduce to classical Liouville equations after substitutions of the kind $[\psi_n,\psi_l]/\psi_{n,l}\rightarrow\exp S_{n,l}$.

It can be proved that expression (51) is *a local integral* for that pair of independent equations (for any choice of two arbitrary functions of t in their solutions). In terms of Dirac's description of Hamiltonian systems with constraints (see the next section) this expression is classified as ''a complementary condition that is not substantial by virtue of a dynamics of the system'' (this means that such a condition does not violate an integrability of the dynamical system).

Consider the ''diagonalizing'' transform

$$\{\psi_n,\psi_l\}\Rightarrow\left\{\Phi_n=+\frac{\psi_l}{[\psi_n,\psi_l]},\quad \Phi_l=-\frac{\psi_n}{[\psi_n,\psi_l]}\right\} \tag{55}$$

in more detail. This transform is antisymmetric:

$$\Phi_n=+\frac{\psi_l}{[\psi_n,\psi_l]},\quad \psi_n=-\frac{\Phi_l}{[\Phi_n,\Phi_l]},$$
$$\Phi_l=-\frac{\psi_n}{[\psi_n,\psi_l]},\quad \psi_l=+\frac{\Phi_n}{[\Phi_n,\Phi_l]}; \tag{56}$$

since

$$[\Phi_n,\Phi_l]\cdot[\psi_n,\psi_l]=-1. \tag{57}$$

The condition (50) or (51) (which was written as a condition for the functions ψ_n and ψ_l to be solutions of the Schrödinger equation for the same potential but for different values E_n and E_l) can be rewritten in the form

$$\frac{1}{2}\frac{(\Phi_n)_{xx}}{\Phi_n}+E_n=\frac{1}{2}\frac{(\Phi_l)_{xx}}{\Phi_l}+E_l. \tag{58}$$

It means that the functions $\psi_l/[\psi_n,\psi_l]$, $\psi_n/[\psi_n,\psi_l]$ are also solutions of the Schrödinger equation for a potential, which differs from the potential for the functions ψ_n,ψ_l, but for the same values of the parameters E_n,E_l. Indeed, it is not difficult to prove that the transform (55) can be represented as a product of two Darboux transforms defined at the states with eigenvalues E_n and E_l, respectively.

The explicit formula for the change of the potential by the transform (55) is as follows:

$$U'=U+\frac{d}{dx}\mathrm{D}_{\log}[\psi_n,\psi_l]. \tag{59}$$

Let us try to understand the following: why it is after the transform to the functions Φ_n^{-2},Φ_l^{-2} that the initial coupled evolutional equations for the functions ψ_n,ψ_l are diagonalized in our problem? The answer is related to the following simple observation: if some function $F(x,t)$ is a solution to the equation

$$(\mathrm{D}_{\log}F)_t=F \tag{60}$$

(which reduces to the Liouville equation $S_{xt}=\exp S$), then the function $\Phi=F^{-1/2}$ is a solution to the equation

$$\left\{\frac{\Phi_{xx}}{\Phi}\right\}_t=0, \tag{61}$$

and *vice versa* (generically, in the latter case the ''times'' t in these equations are not the same, but they are related by a simple scaling).

It is easy to see that Eq. (61) is satisfied by a family of solutions to the Schrödinger equation with a fixed potential and a fixed parameter E, if we identify t with a substantial parameter of that family (the parameter that cannot be removed by scaling). In the other words, if we continuously change the parameter t and thus move over solutions $\Phi^{(t)}$ of the Schrödinger equation with a fixed potential and fixed parameter E, then the functions $F^{(t)}=(\Phi^{(t)})^{-2}$ ''evolve'' in accordance with the Liouville equations (60); it is important that E is constant.

Thus, the Liouville equations for the functions Φ^{-2} arise in the problem on the isospectral deformation quite naturally. A switch from the set of eigenfunctions ψ_k to the set of functions Φ_k is performed by transforms of the Darboux type: in the case of ''individual'' MKT flows it is a simple Darboux transform; in the case of ''two-level'' MKT flow, which we study, it is the transform (55) that can be represented as an analogous double (consecutive) transform.

Now return to the separated equations for the functions $\Phi_n^{-2}=([\psi_n,\psi_l]/\psi$ and $\Phi_l^{-2}=([\psi_n,\psi_l]/\psi_n)^2$ (53), (54). They can be written in the form

$$\frac{\partial}{\partial t_1}\mathrm{D}_{\log}\frac{1}{\Phi_n^2}=+\frac{1}{\Phi_n^2}, \tag{62}$$

$$\frac{\partial}{\partial t_2}\mathrm{D}_{\log}\frac{1}{\Phi_l^2}=-\frac{1}{\Phi_l^2}, \tag{63}$$

where $t_1=\tau_n/\Delta_{nl}t,\ t_2=\tau_l/\Delta_{nl}t$. Solutions of these equations are as follows:

$$\frac{1}{\Phi_n^2}=+2\frac{\partial}{\partial t_1}\mathrm{D}_{\log}\left[1-\frac{1}{2}\int_0^{t_1}\exp g_1(t')dt'\int_x^\infty \frac{1}{\Phi_n^2}\bigg|_{t=0}dx'\right], \tag{64}$$

$$\frac{1}{\Phi_l^2}=-2\frac{\partial}{\partial t_2}\mathrm{D}_{\log}\left[1+\frac{1}{2}\int_0^{t_2}\exp g_2(t')dt'\int_x^\infty \frac{1}{\Phi_l^2}\bigg|_{t=0}dx'\right]. \tag{65}$$

Denote the right-hand sides of these solutions by F_1 and F_2, respectively. Then final formulas for the isospectral deformation of the eigenfunctions in the case of two-level flow will be given in the form

$$\psi_l=1\bigg/\sqrt{F_2}\int_x^\infty\frac{dx'}{\sqrt{F_1F_2}},\quad \psi_n=1\bigg/\sqrt{F_1}\int_x^\infty\frac{dx'}{\sqrt{F_1F_2}}. \tag{66}$$

Equations (53), (54) can also be rewritten as a pair of separated equations for certain functions $Z^\pm$, which are only expressed through the function $a(x,t)$ [see Eq. (52)] and its derivative; i.e., in fact, we can obtain equations that describe the isospectral deformation of the *potential* in closed form for the case of two-level McKean–Trubowitz flow. Define the functions $Z^\pm$ as follows:

$$Z^+=\frac{a_x}{a}-a+\frac{2\Delta_{nl}}{a}, \tag{67}$$

$$Z^-=\frac{a_x}{a}-a-\frac{2\Delta_{nl}}{a}; \tag{68}$$

and use, together with Eq. (3), the identity

$$\frac{a_x}{a}-a=\mathrm{D}_{\log}\left[\frac{a^2}{\psi_n\psi_l}\right]. \tag{69}$$

It can be shown that Eqs. (53), (54) and relations (3), (69) lead to the separated equations for the functions $Z^\pm(x,t)$:

$$\frac{\partial}{\partial t}\mathrm{D}_{\log}Z_t^\pm=-Z_t^\pm. \tag{70}$$

Any of them can be used to determine the evolution of the potential, since the compatibility condition for their solutions (i.e., a condition under which they give the same potential),

$$2(Z^+)_x+(Z^+)^2+4\Delta_{nl}=2(Z^-)_x+(Z^-)^2-4\Delta_{nl}, \tag{71}$$

is nothing but the condition (51) rewritten in the new variables.

The solutions of Eqs. (70) have the form

$$Z^{\pm}=Z^{\pm}+2\mathrm{D}_{\log}\,\theta^{\pm}(x,t),$$

$$\theta\equiv 1-\frac{1}{2}\int_0^t \exp g^{\pm}(t')dt'\int_x^{\infty}|Z_t^{\pm}|_{t=0}dx'. \tag{72}$$

Using these expressions, we can write down a solution for the function $a(x,t)$ (an antiderivative of the potential); the same solution can be obtained from formulas (64), (65):

$$\frac{1}{a(x,t)}=\frac{1}{a^0(x)}+\frac{1}{2\Delta_{nl}}\mathrm{D}_{\log}\frac{1-2\tau_l\Delta_{nl}\alpha(t)\int_x^{\infty}(\psi_l/a)^2|_{t=0}dx'}{1+2\tau_n\Delta_{nl}\beta(t)\int_x^{\infty}(\psi_n/a)^2|_{t=0}dx'}, \tag{73}$$

where $\alpha(t)=\int_0^t \exp g_1(t')dt'$, $\beta(t)=\int_0^t \exp g_2(t')dt'$.

It can be proved that this expression can be transformed to a particular case of a well-known formula for the isospectral deformation of the potential in the Schrödinger operator,

$$a(x,t)=a^0-\mathrm{D}_{\log}\begin{vmatrix} 1+\alpha(t)\int_x^{\infty}(\psi_n^0)^2dx' & \alpha(t)\int_x^{\infty}\psi_n^0\psi_l^0dx' \\ \beta(t)\int_x^{\infty}\psi_n^0\psi_l^0dx' & 1+\beta(t)\int_x^{\infty}(\psi_l^0)^2dx' \end{vmatrix}, \tag{74}$$

which is derived both in the method related to the integral equations of the Gelfand–Levitan and Marchenko types, and in the McKean and Trubowitz approach. The transformation from formula (73) to (74) is not obvious; but it is not difficult to check, as an intermediate step, that Eqs. (73) and (74) can be rewritten as follows:

$$\frac{1}{a(x,t)}=\frac{1}{a^0}-\frac{1}{(a^0)^2}\frac{-R_x}{R-R_x[a^0]^{-1}} \tag{75}$$

and

$$a(x,t)=a^0-\mathrm{D}_{\log}R, \tag{76}$$

respectively.

Thus, we have shown that the known formulas for the isospectral deformation of the potential and of the eigenfunctions can also be obtained as solutions of PDEs (reducible to the Liouville equations), which we derived using the ‘‘evolutional’’ approach to the description of the isospectral deformation of the Schrödinger operator together with the formulas for the McKean–Trubowitz flow.

VI. ON HAMILTONIAN DESCRIPTION OF TWO-LEVEL ISOSPECTRAL DEFORMATION

Performing the scaling $x/\sqrt{\Delta_{nl}}\Rightarrow\xi$, $\sqrt{\Delta_{nl}}t\Rightarrow\tau$, $\sqrt{\tau_n}\psi_n\Rightarrow\Psi_n$, $\sqrt{\tau_l}\psi_l\Rightarrow\Psi_l$, and defining the variables Γ_n, Γ_l, Γ_{nl} by the relations

$$\Gamma_n\equiv\frac{(\Psi_n)_\xi}{\Psi_n},\quad \Gamma_l\equiv\frac{(\Psi_l)_\xi}{\Psi_l},\quad \Gamma_{nl}\equiv\Gamma_n-\Gamma_l, \tag{77}$$

rewrite the initial system of two coupled evolutional equations in the form

$$(\Gamma_n)_\tau-(\Psi_n^2+\Psi_l^2)=+\tfrac{1}{2}(\Gamma_{nl}\Psi_l)^2, \tag{78}$$

$$(\Gamma_l)_\tau-(\Psi_n^2+\Psi_l^2)=-\tfrac{1}{2}(\Gamma_{nl}\Psi_n)^2. \tag{79}$$

In the new variables the relation (50) (the condition for the functions ψ_n, ψ_l to be solutions of the Schrödinger equation with one and the same potential for E_n and E_l) takes the form of a relation between the logarithmic derivatives Γ_n and Γ_l:

$$\tfrac{1}{2}(\Gamma_{nl})_\xi = -1 - \tfrac{1}{2}\Gamma_n^2 + \tfrac{1}{2}\Gamma_l^2. \tag{80}$$

Note that Eqs. (78), (79) and the relation (80) do not contain any parameters.

The evolutional equations (78), (79) can be written as Hamiltonian equations with respect to the dynamical variables Γ_n, Γ_l:

$$(\Gamma_n)_\tau = -\frac{\partial}{\partial\xi}\frac{\delta\mathbf{H}}{\delta\Gamma_n}, \quad (\Gamma_l)_\tau = +\frac{\partial}{\partial\xi}\frac{\delta\mathbf{H}}{\delta\Gamma_l}; \tag{81}$$

these equations are generated by the functional

$$\mathbf{H} = \frac{1}{4}\int_{-\infty}^{+\infty} d\xi\,\Gamma_{nl}^2(\xi,\tau)\left[\exp\left(2\int_{-\infty}^{\xi} d\xi'\,\Gamma_n(\xi',\tau)\right) + \exp\left(2\int_{-\infty}^{\xi} d\xi'\,\Gamma_l(\xi',\tau)\right)\right]. \tag{82}$$

In fact, one can prove that Eqs. (81) together with the condition (80) and the definitions (77) lead to the initial system of evolutional equations (78), (79).

Thus, the evolutional formulation of the problem on isospectral deformations of the Schrödinger operator is related with dynamical systems of the Hamiltonian type with constraints. The Hamiltonian formalism for such a class of dynamical systems was first studied by Dirac.[18]

Now pass from the variables $\{\Psi_{n,l}, \Gamma_{n,l} = \mathrm{D}_{\log}\Psi_{n,l}\}$ to the variables $\{\zeta_{n,l}, \gamma_{n,l} = \mathrm{D}_{\log}\zeta_{n,l}\}$ by the relations

$$\gamma_n = \Gamma_n + \frac{(\Gamma_{nl})_\xi}{\Gamma_{nl}}, \quad \gamma_l = \Gamma_l + \frac{(\Gamma_{nl})_\xi}{\Gamma_{nl}},$$

$$\zeta_n = \Psi_n\Gamma_{nl}, \quad \zeta_l = \Psi_l\Gamma_{nl}.$$

In these variables the evolutional equations (78), (79) and the constraint (80) take the form

$$(\gamma_n)_\tau = -\tfrac{1}{2}\zeta_n^2, \quad (\gamma_l)_\tau = +\tfrac{1}{2}\zeta_l^2, \tag{83}$$

$$-\tfrac{1}{2}(\gamma_{nl})_\xi + \tfrac{1}{2}(\gamma_n^2 - \gamma_l^2) = -1. \tag{84}$$

Thus, the transform $(\Gamma_i, \Psi_i) \Rightarrow (\gamma_i, \zeta_i)$ leads to a system of two *separated* equations. Moreover, this transform *preserves a Hamiltonian structure* of the dynamical system: the equations

$$(\gamma_n)_\tau = -\frac{\partial}{\partial\xi}\frac{\delta\mathbf{h}}{\delta\gamma_n}, \quad (\gamma_l)_\tau = +\frac{\partial}{\partial\xi}\frac{\delta\mathbf{h}}{\delta\gamma_l}, \tag{85}$$

generated by the functional

$$\mathbf{h} = \frac{1}{4}\int_{-\infty}^{+\infty} d\xi\left[\exp\left(2\int_{-\infty}^{\xi} d\xi'\,\gamma_n(\xi',\tau)\right) + \exp\left(2\int_{-\infty}^{\xi} d\xi'\,\gamma_l(\xi',\tau)\right)\right] \tag{86}$$

provide the evolutional equations (83); note that $\mathbf{h} = \mathbf{H}$.

Furthermore, the transform $(\Gamma_i, \Psi_i) \Rightarrow (\gamma_i, \zeta_i)$ preserves ‘‘naive’’ (in terms of Ref. 16) Poisson brackets. For example, if they are defined as follows:

$$\{a,b\}_{P(\Gamma)} \equiv \int_{-\infty}^{+\infty} d\xi\,\frac{\delta a}{\delta\Gamma(\xi)}\frac{\partial}{\partial\xi}\frac{\delta b}{\delta\Gamma(\xi)}, \tag{87}$$

then we find that

$$\{\gamma_m,\gamma_{m'}\}_{P(\gamma)}=\{\gamma_m,\gamma_{m'}\}_{P(\Gamma)}, \tag{88}$$

since the variational derivative of a logarithmic derivative always vanishes.

Thus, the transform $(\Gamma_i,\Psi_i)\Rightarrow(\gamma_i,\zeta_i)$ is *a canonical transform* for the Hamiltonian dynamical system with the Hamiltonian function $\mathbf{h}=\mathbf{H}$.

Note also that the constraint (84) is *a local integral* of the system (83). This means that the expression on the right-hand side of Eq. (84) (denote it by J_{nl}) satisfies the equation $(J_{nl})_\tau=0$ by virtue of Eqs. (83). Hence, the equations give us $J_{nl}=C(\xi)$; and only choosing the ''level'' $C(\xi)=-1$ we can satisfy the condition (84).

Now let us formulate principal results of this section.

(i) The above-considered system of two coupled evolutional equations can be written as Hamiltonian equations generated by the functional (82) (where ''logarithmic derivatives'' of the eigenfunctions are canonical variables).

(ii) The switch from the system of two coupled evolutional equations to the pair of split evolutional equations of Liouville form is performed by a canonical change of variables.

APPENDIX: SEQUENCE OF CANONICAL TRANSFORMS

Consider the following sequence of transforms, each of which preserves a Hamiltonian structure of the equations (78), (79).

(i): $(\Gamma_i,\Psi_i)\Rightarrow(\gamma_i,\zeta_i)$. The initial system of evolutional equations (78), (79) and the constraint (80) are transformed to (83), (84).

(ii): $(\gamma_i,\zeta_i)\Rightarrow(\tilde{\gamma}_i,\tilde{\zeta}_i)$: $\tilde{\gamma}_n=-\gamma_n$, $\tilde{\gamma}_l=-\gamma_l$, $\tilde{\zeta}_n=\zeta_n^{-1}$, $\tilde{\zeta}_l=\zeta_l^{-1}$. Now the system of equations (83), (84) is transformed to the form

$$(\tilde{\gamma}_n)_\tau=-\tfrac{1}{2}\tilde{\zeta}_n^{-2},\quad (\tilde{\gamma}_l)_\tau=+\tfrac{1}{2}\tilde{\zeta}_l^{-2}, \tag{A1}$$

$$+\tfrac{1}{2}(\tilde{\gamma}_{nl})_\xi-\tfrac{1}{2}\tilde{\gamma}_{nl}(\tilde{\gamma}_n+\tilde{\gamma}_l)=-1. \tag{A2}$$

(iii): $(\tilde{\gamma}_i,\tilde{\zeta}_i)\Rightarrow(\tilde{\Gamma}_i,\tilde{\Psi}_i)$:

$$\tilde{\Gamma}_n=\tilde{\gamma}_n+\frac{(\tilde{\gamma}_{nl})_\xi}{\tilde{\gamma}_{nl}},\quad \tilde{\Gamma}_l=\tilde{\gamma}_l+\frac{(\tilde{\gamma}_{nl})_\xi}{\tilde{\gamma}_{nl}},$$

$$\tilde{\Psi}_n=\tilde{\Gamma}_{nl}\tilde{\zeta}_n,\quad \tilde{\Psi}_l=\tilde{\Gamma}_{nl}\tilde{\zeta}_l.$$

Here the system (A1), (A2) takes the form

$$\left(\tilde{\Gamma}_n+\frac{2}{\tilde{\Gamma}_{nl}}\right)_\tau=+\frac{\tilde{\Gamma}_{nl}^2}{2\tilde{\Psi}_n^2},\quad \left(\tilde{\Gamma}_l+\frac{2}{\tilde{\Gamma}_{nl}}\right)_\tau=-\frac{\tilde{\Gamma}_{nl}^2}{2\tilde{\Psi}_l^2}, \tag{A3}$$

$$-\tfrac{1}{2}(\tilde{\Gamma}_{nl})_\xi+\tfrac{1}{2}\tilde{\Gamma}_{nl}(\tilde{\Gamma}_n+\tilde{\Gamma}_l)=-1. \tag{A4}$$

(iv): $(\tilde{\Gamma}_i,\tilde{\Psi}_i)\Rightarrow(\bar{\Gamma}_i,\bar{\Psi}_i)$: $\bar{\Gamma}_n=-\tilde{\Gamma}_n$, $\bar{\Gamma}_l=-\tilde{\Gamma}_l$, $\bar{\Psi}_n=\tilde{\Psi}_n^{-1}$, $\bar{\Psi}_l=\tilde{\Psi}_l^{-1}$.
After this (last) step the system (A3), (A4) is written as follows:

$$\left(\bar{\Gamma}_n+\frac{2}{\bar{\Gamma}_{nl}}\right)_\tau=-\frac{1}{2}(\bar{\Gamma}_{nl}\bar{\Psi}_n)^2, \tag{A5}$$

$$\left(\overline{\Gamma}_l+\frac{2}{\overline{\Gamma}_{nl}}\right)_\tau=-\tfrac{1}{2}(\overline{\Gamma}_{nl}\overline{\Psi}_l)^2, \tag{A6}$$

$$+\tfrac{1}{2}(\overline{\Gamma}_{nl})_\xi+\tfrac{1}{2}\overline{\Gamma}_{nl}(\overline{\Gamma}_n+\overline{\Gamma}_l)=-1. \tag{A7}$$

It is not difficult to prove that

$$\overline{\Gamma}_n=\Gamma_n,\quad \overline{\Gamma}_l=\Gamma_l,\quad \overline{\Psi}_n=-\Psi_n,\quad \overline{\Psi}_l=-\Psi_l, \tag{A8}$$

whereas Eqs. (A5), (A7) coincide with the initial system of evolutional equations (78), (79) and the constraint (80).

Thus, the chain of transforms $(\Gamma_i,\Psi_i)\Rightarrow(\gamma_i,\zeta_i)\Rightarrow(\tilde{\gamma}_i,\tilde{\zeta}_i)\Rightarrow(\tilde{\Gamma}_i,\tilde{\Psi}_i)\Rightarrow(\overline{\Gamma}_i,\overline{\Psi}_i)$ maps the initial Hamiltonian system with constraints into itself (up to change of signs of the solutions Ψ_i).

Note that links of this chain coincide, principally, with links of the double Darboux transform. This points to a possibility to define a correspondence between the Darboux approach in the theory of integrable evolutional equations of the isospectral deformation in the Schrödinger problem and canonical transforms of an associated Hamiltonian system with constraints. This statement should be made more formal by a strict analysis of the multi-flow case and a more correct definition of Poisson–Dirac brackets.[16]

[1] A. C. Newell, *Solitons in Mathematics and Physics* (SIAM, Philadelphia, PA 1985).
[2] J. Poschel and E. Trubowitz, *Inverse Spectral Theory* (Academic, New York, 1987).
[3] G. Darboux, C. R. Acad. Sci. (Paris) **94**, 1456 (1882).
[4] M. M. Crum, Q. J. Math. **6**, 121 (1955).
[5] M. G. Krein, Dokl. Akad. Nauk SSSR **113**, 970 (1957).
[6] E. Schrödinger, Proc. R. Ir. Acad. A, Math. Phys. Sci. **46**, 9; **46**, 183 (1940); **47**, 53 (1941).
[7] L. Infeld and E. T. Hull, Rev. Mod. Phys. **23**, 21 (1951).
[8] B. Mielnik, J. Math. Phys. **25**, 3387 (1984).
[9] E. Witten, Nucl. Phys. B **188**, 513 (1981).
[10] C. V. Sukumar, J. Phys. A **18**, L57 (1985); **18**, 2917 (1985); **21**, L455 (1988).
[11] A. P. Veselov and A. B. Shabat, Funkt. Anal. Pril. **27**, 1 (1993) (in Russian).
[12] V. M. Eleonsky and V. G. Korolev, J. Phys. A **28**, 4973 (1995); **29**, L241 (1996); JETP **83**, 1084 (1996).
[13] H. P. McKean and E. Trubowitz, Commun. Math. Phys. **82**, 471 (1982).
[14] B. M. Levitan, Mat. Sb. **132**, 73 (1987) (in Russian).
[15] V. M. Eleonsky and V. G. Korolev, Phys. Rev. A **55**, 2580 (1997).
[16] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Approach in Soliton Theory* (Springer-Verlag, Berlin, 1985).
[17] F. Calogero and A. Degasperis, *Spectral Transform and Solitons* (North-Holland, New York, 1982), Vol. I.
[18] P. Dirac, *Lectures on Quantum Mechanics*, Belfer Graduate School of Science, Yeshiva University, NY, 1964.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Soliton solutions, Liouville integrability and gauge equivalence of Sasa Satsuma equation

Sasanka Ghosh[a)]
Physics Department, Indian Institute of Technology, Guwahati Panbazar, Guwahati 781 001, India

Anjan Kundu[b)]
TNP Division, Saha Institute of Nuclear Physics 1/AF Bidhan Nagar, Calcutta 700 069, India

Sudipta Nandy[c)]
Physics Department, Indian Institute of Technology, Guwahati Panbazar, Guwahati 781 001, India

(Received 5 October 1998; accepted for publication 14 December 1998)

Exact integrability of the Sasa Satsuma equation (SSE) in the Liouville sense is established by showing the existence of an infinite set of conservation laws. The explicit form of the conserved quantities in terms of the fields are obtained by solving the Riccati equation for the associated 3×3 Lax operator. The soliton solutions, in particular, one and two soliton solutions, are constructed by the Hirota's bilinear method. The one soliton solution is also compared with that found through the inverse scattering method. The gauge equivalence of the SSE with a generalized Landau Lifshitz equation is established with the explicit construction of the new equivalent Lax pair.
[S0022-2488(99)00204-2]

I. INTRODUCTION

Nonlinear Schrödinger equation and its various generalized versions (higher order nonlinear Schrödinger equation) is well known in describing various physical phenomenas.[1] A common property in all these physical systems is the appearance of solitons, as a result of a balance between the nonlinear and dispersive terms of the wave equations. With the advancement of experimental accuracy, solitons having more complicated dynamics can also be detected and observed now.[2] The Sasa Satsuma equation[3]

$$iQ_T+\frac{1}{2}Q_{XX}+|Q|^2Q+\frac{i}{6\epsilon}(Q_{XXX}+6|Q|^2Q_X+3|Q|^2_XQ)=0 \tag{1}$$

describing the evolution of a complex scalar field, is an example of such a system, whose soliton solutions have been obtained through inverse scattering method (ISM) in Ref. 3.

A limited class of soliton bearing equations exhibits further interesting properties and belongs to the exclusive club of integrable systems. The most prominent definition of integrability is the integrability in the Liouville sense, i.e., the existence of a set of infinite numbers of conserved quantities in involution,[4] which can be considered as the action variables. This criterion of integrability is extendable also to the quantum case. The Lax pair associated with the model is usually a sign of such integrability, while the Painlevé singularity analysis[5] is supposed to be a direct test of integrability for the given equation.

[a)]Electronic mail: sasanka@iitg.ernet.in
[b)]Electronic mail: anjan@tnp.saha.ernet.in
[c)]Electronic mail: sudipta@iitg.ernet.in

It should be mentioned that the Lax pair for the SSE as well as its one soliton through ISM were found in Ref. 3, while the Painleve analysis for the equation was carried out in Ref. 6. However, extraction of the higher conserved quantities for the Sasa Satsuma system and thus establishing the integrability of the whole hierarchy in the Liouville sense remained unexplored. A possible reason of this may be the difficulty involved due to the unusual 3×3 matrix form of the Lax operator, associated with the SSE.

Our objective is, therefore, to find the Riccati equation for the 3×3 Lax operator, associated with the SSE and consequently, to obtain the whole hierarchy of conserved charges in a systematic way. This construction will be somewhat involved due to the extended form of the Lax operator.

For investigating SSE from a different view point, we further find the explicit soliton solutions of the equation through Hirota's bilinear method. This is a direct and much more effective method compared to ISM for obtaining the soliton solutions, since it does not require the knowledge of the Lax pair. Moreover, the construction of the τ function becomes straightforward in this method.

One should recall in this context another interesting fact about the NLS equation that it is gauge related to the well known Landau Lifshitz equation (LLE).[7] This equivalence can also be established through the space curve method.[8] It is, therefore, natural to ask what is the gauge equivalent equation to the SSE. Though such an equivalent system has been discovered by the space curve method in Ref. 6, we complete the investigation for SSE by showing the equivalence of it through an explicit gauge transformation, which not only reproduces the generalized LLE (GLLE), but also constructs the associated Lax pair for the GLLE.

The organization of this paper is as follows. In Sec. II we study the soliton solutions of SSE by Hirota's bilinear method. We compute explicitly the one and two soliton solutions and compare our result of one soliton solution with the known one.[3] In Sec. III, we construct the related Riccati equation using the 3×3 Lax operator of the SSE and subsequently find the infinite number of conserved quantities through the recursion relation. The time invariance of the conserved quantities is checked directly by using the evolution equation. This proves the Liouville integrability of the SSE. Section IV provides the gauge equivalent generalized LLE and gives the associated new Lax operators in the explicit form. Section V is the concluding one.

II. SOLITON SOLUTIONS THROUGH HIROTA'S METHOD

Let us begin with the SSE (1), which through a change of variable and a Galelian transformation:

$$Q(X,T)=u(x,t)\exp\left\{i\epsilon\left(x+\frac{\epsilon t}{6}\right)\right\},$$

$$T=t,$$

$$X=x+\frac{\epsilon}{2}t, \tag{2}$$

may be simplified to the form

$$u_t+\frac{1}{6\epsilon}(u_{xxx}+6|u|^2u_x+3(|u|^2)_xu)=0. \tag{3}$$

This is an example of a complex modified KdV type equation and goes to mKdV for the real valued field. The associated spectral problem can be studied through the pair of linear equations

$$\Psi_x=\mathbf{U}(x,t,\lambda)\Psi, \tag{4a}$$

$$\Psi_t=\mathbf{V}(x,t,\lambda)\Psi, \tag{4b}$$

where $\mathbf{U}(x,t,\lambda)$ and $\mathbf{V}(x,t,\lambda)$ are 3×3 matrices and λ is the spectral parameter. The explicit form of $\mathbf{U}$ and $\mathbf{V}$ may be given using the result of Ref. 3 as

$$\mathbf{U}=-i\lambda\Sigma+\mathbf{A}, \tag{5a}$$

$$\mathbf{V}=-i4\epsilon\lambda^3\Sigma+4\epsilon(\lambda^2-|u|^2)\mathbf{A}-i2\lambda\epsilon\Sigma(\mathbf{A}^2-\mathbf{A}_x)-\epsilon\mathbf{A}_{xx}+\epsilon(\mathbf{A}_x\mathbf{A}-\mathbf{A}\mathbf{A}_x), \tag{5b}$$

with

$$\Sigma=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

$$A=\begin{pmatrix} 0 & 0 & u \\ 0 & 0 & u^* \\ -u^* & -u & 0 \end{pmatrix}.$$

Compatibility of (4a) and (4b) leads to SSE (3), which can be shown easily by using the following properties of Σ and $\mathbf{A}$ matrices:

$$\{\Sigma,\mathbf{A}\}=0,$$

$$\Sigma^2=1,$$

$$\mathbf{A}^3+2|u|^2\mathbf{A}=0.$$

Note that SSE in the form (3), though suitable for studying inverse scattering technique, is not convenient for casting it into Hirota's bilinear form. On the other hand, the higher order nonlinear Schrödinger equation form (1) for SSE is more suitable for this purpose. Now, in order to write (1) in the bilinear form, we make the transformation

$$Q(X,T)=G(X,T)/F(X,T), \tag{6}$$

where, G is complex and F is real. Consequently, in these new variables, we have the following set of equations:

$$\left(iD_T+\frac{1}{2}D_X^2+\frac{i}{6\epsilon}D_X^3\right)G\cdot F=0, \tag{7a}$$

$$D_X^2F\cdot F=4G^*\cdot G, \tag{7b}$$

$$\left(1-\frac{2i}{\epsilon}D_X\right)G^*\cdot G=0, \tag{7c}$$

which follow from (1). D_T, D_X, D_{XX} etc. in (7) are Hirota derivatives.[9] Equation (7) belongs to a new class of bilinear equations, whose general form would be of the type

$$\mathcal{B}(D_X,D_T,\ldots)G\cdot F=0, \tag{8a}$$

$$\mathcal{A}(D_X,D_T,\ldots)F\cdot F=\mathcal{C}(D_X,D_T,\ldots)G^*\cdot G, \tag{8b}$$

$$\mathcal{E}(1-D_X,D_T,\ldots)G^*\cdot G=0. \tag{8c}$$

The additional bilinear equation (7c) involving G^*G imposes one further condition on the complex parameter P (shown below), which is absent in other examples of higher order nonlinear Schrödinger equations.[10]

For obtaining one soliton solution of SSE (1), we choose G and F in the following form

$$G = L\exp(\eta), \tag{9a}$$

$$F = 1 + K\exp(\eta+\eta^*), \tag{9b}$$

where, L is a complex c-number parameter and

$$\eta = PX + \Omega T + \cdots, \tag{10}$$

with P, Ω are in general complex parameters. Substituting the expressions (9) for G and F in (7), we see that G and F are the solutions of (7) provided the following relations hold:

$$i\Omega + \frac{1}{2}P^2 + i\frac{P^3}{6\epsilon} = 0, \tag{11a}$$

$$K = \frac{LL^*}{2\mu^2}, \tag{11b}$$

$$P - P^* = 2i\epsilon, \tag{11c}$$

where the complex parameter, P is of the form

$$P = \mu + i\epsilon. \tag{12a}$$

Equation (11a) is nothing but the dispersion relation and (11b) determines K. In the above solution, so far, L is an arbitrary complex parameter. We will see shortly that in order to compare our result with the one obtained through the ISM,[3] the parameter L is to be chosen in a specific form. It follows from the dispersion relation (11a) and the expression of P (12a) that Ω should be of the form

$$\Omega = -\mu\left(\frac{\mu^2}{6\epsilon} + \frac{\epsilon}{2}\right) - i\frac{\epsilon^2}{3}. \tag{12b}$$

Substituting (12a) and (12b) in (10) and using (11b) the one soliton solution in the explicit form becomes

$$Q(X,T) = \frac{L\exp\eta}{1+K\exp(\eta+\eta^*)} = \frac{L\exp\left\{(\mu+i\epsilon)X - \left(\frac{\mu^3}{6\epsilon} + \frac{\mu\epsilon}{2} + i\frac{\epsilon^2}{3}\right)T\right\}}{1 + \frac{|L|^2}{2\mu^2}\exp\left\{2\mu X - \left(\frac{\mu^3}{3\epsilon} + \mu\epsilon\right)T\right\}}. \tag{13}$$

To compare (13) with that of ISM, we choose L as

$$L = 2\mu\exp(-i\epsilon X^{(1)} - \mu X^{(0)}), \tag{14}$$

which reduces (13) to the form

$$Q(X,T)=\frac{\frac{\mu}{\sqrt{2}}\exp\left\{i\epsilon\left(X-\frac{\epsilon}{3}T-X^{(1)}\right)\right\}}{\cosh\left(\mu X-\mu\left(\frac{\mu^2}{6\epsilon}+\frac{\epsilon}{2}\right)T-\mu X^{(0)}+\log\sqrt{2}\right)} \tag{15}$$

and this is in agreement with the ISM result of Ref. 3.

Two soliton solutions of the Sasa Satsuma equation may be obtained, following Ref. 10, by choosing G and F of the form

$$G=L(\exp\eta_1+M\exp(\eta_1+\eta_A)), \tag{16a}$$

$$F=1+k\exp(\eta_1+\eta_1^*)+\exp(\eta_A)+kMM^*\exp(\eta_1+\eta_1^*+\eta_A), \tag{16b}$$

where, η_1 is complex as in the case of one solition, but η_A is chosen to be real. Substituting F and G (16) in the bilinear forms (7), we observe that η_1 soliton satisfies the same dispersion relation as (11a), whereas the dispersion relation for η_A becomes $P_A^2=0$. The degree 2 term in (7a) determines the parameter M to be unity. Once again following Ref. 10, we define the degree of a term by the number of η's present in the exponent. Now the degree 2 and 4 terms in (7b), yield the value of k as in (11b). However, because of the additional bilinear form (7c) in our case, we obtain one more relation, $\epsilon(P_1^*-P_1)=2i$. Note that, in general for complex bosonic systems having complex parameters, the two soliton solutions may have some relation, which is analogous to the three soliton conditions.[10] But, the choice of the real parameter η_A, makes the three soliton condition trivial in our case. A more general choice of F and G for two soliton solutions will give such a nontrivial condition.

III. RICCATI EQUATION AND THE CONSERVED QUANTITIES

To show the Liouville integrability, i.e., the existence of an infinite number of conserved quantities related to SSE (3), we first write the associated Riccati equation. Since the Lax operators in this case are 3×3 matrices, the Riccati equation becomes more complicated, though tactable. Let us write the auxiliary field Ψ in the component form as

$$\Psi=\begin{pmatrix}\chi_1\\ \chi_2\\ \chi_3\end{pmatrix}. \tag{17}$$

Substituting (17) in (4a), we get a set of three coupled equations:

$$\chi_{1x}=-i\lambda\chi_1+u\chi_3, \tag{18a}$$

$$\chi_{2x}=-i\lambda\chi_2+u^*\chi_3, \tag{18b}$$

$$\chi_{3x}=i\lambda\chi_3-u^*\chi_1-u\chi_2. \tag{18c}$$

Now expressing (18) in terms of $\Gamma_1=(\chi_1/\chi_3)$ and $\Gamma_2=(\chi_2/\chi_3)$, and eliminating χ_1, χ_2, and χ_3, one obtains

$$\Gamma_{1x}=u-2i\lambda_1+u^*\Gamma_1^2+u\Gamma_1\Gamma_2, \tag{19a}$$

$$\Gamma_{2x}=u^*-2i\lambda_2+u\Gamma_2^2+u^*\Gamma_1\Gamma_2. \tag{19b}$$

The first order nonlinear coupled equations (19) for Γ_1 and Γ_2 are the Riccati equations in our case. Notice that neither the integral of Γ_1 nor of Γ_2, plays the role of generating functions for conserved quantities, but a suitable combination of them does. The infinite number of conserved quantities (Hamiltonians), H_{2n+1}; $n=0,1,2,\ldots$ can be obtained from (19) by identifying

$$a(\lambda)=\exp(-i\lambda x)|_{x\to\infty}\Psi_3(\infty,\lambda)=\exp\left\{-\int_{-\infty}^{\infty}(u^*\Gamma_1+u\Gamma_2)dx\right\}, \tag{20a}$$

where H_{2n+1} are related to $a(\lambda)$ as

$$\ln a(\lambda)=-2\sum_{n=0}^{\infty}(2i)^{-2n-1}H_{2n+1}\lambda^{-2n-1}. \tag{20b}$$

We will see that Hamiltonians with odd indices only survive, while the terms with even indices become trivial. This property is similar to that of the real KdV or the modified KdV equation.

We may look for series solutions of (19) by assuming Γ_1 and Γ_2 in the form

$$\Gamma_1=\sum_{n=0}^{\infty}C_n^1\lambda^{-n}, \tag{21a}$$

$$\Gamma_2=\sum_{n=0}^{\infty}C_n^2\lambda^{-n}, \tag{21b}$$

which yield the following recursion relations from (19a) and (21):

$$C_0^1=0,\quad C_1^1=\frac{u}{2i},$$

$$2iC_{n+2}^1=-(C_{n+1}^1)_x+\sum_{m=0}^{n+1}(u^*C_m^1C_{n-m+1}^1+uC_m^1C_{n-m+1}^2). \tag{22}$$

Similarly (19b) and (21) determine another, though quite similar, set of recursion relations given as

$$C_0^2=0,\quad C_1^2=\frac{u^*}{2i},$$

$$2iC_{n+2}^2=-(C_{n+1}^2)_x+\sum_{m=0}^{n+1}(uC_m^2C_{n-m+1}^2+u^*C_m^2C_{n-m+1}^1). \tag{23}$$

Inserting the expressions of C_n^1 and C_n^2, thus obtained through the recursion relations (22), (23) in (21), we get from (20a) and (20b) the explicit form of all conserved quantities, H_{2n+1}. These expressions are the integrals taken over the functions of the fields u and u^* and their derivatives. The first few conserved quantities of the infinite set of the SSE hierarchies are given by

$$H_1=\int u^*u\,dx, \tag{24}$$

$$H_3=\int(-u_x^*u_x+2|u|^4)dx, \tag{25}$$

$$H_5=\int(u_{xx}^*u_{xx}-8|u|^2u_x^*u_x-3((|u|^2)_x)^2+8|u|^6)dx. \tag{26}$$

We have checked explicitly by using equations of motion (3) that H_1, H_3, and H_5 are, indeed, the constants of motion.

IV. GENERALIZED LANDAU LIFSHITZ TYPE EQUATION AS THE GAUGE EQUIVALENT SYSTEM

We now show an interesting connection between the SSE in the form (3) and the generalized Landau Lifshitz type equation, by exploiting the gauge equivalence of the Lax pairs of these two dynamical systems. The procedure is similar to that between the NLS and the standard Landau Lifshitz equation.[7] Under a local gauge transformation, the Jost function, $\Psi(x,t,\lambda)$ changes as

$$\tilde{\Psi}(x,t,\lambda)=g^{-1}(x,t)\Psi(x,t,\lambda), \tag{27}$$

where $g(x,t)=\Psi(x,t,\lambda)|_{\lambda=0}$, may be taken as an element of the gauge group. Consequently, the Lax equations (4) under this gauge transformation (27) become

$$\tilde{\Psi}_x=\tilde{\mathbf{U}}(x,t,\lambda)\tilde{\Psi}, \tag{28a}$$

$$\tilde{\Psi}_t=\tilde{\mathbf{V}}(x,t,\lambda)\tilde{\Psi}, \tag{28b}$$

where $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are the new gauge transformed Lax pair, given by

$$\tilde{\mathbf{U}}(x,t,\lambda)=g^{-1}(\mathbf{U}-\mathbf{U}_0)g, \tag{29a}$$

$$\tilde{\mathbf{V}}(x,t,\lambda)=g^{-1}(\mathbf{V}-\mathbf{V}_0)g, \tag{29b}$$

with $\mathbf{U}_0=\mathbf{U}|_{\lambda=0}=g^{-1}(x,t)g_x(x,t)$ and $\mathbf{V}_0=\mathbf{V}|_{\lambda=0}=g^{-1}(x,t)g_t(x,t)$.

We may identify the spin field of the Landau Lifshitz type equation as

$$S=g(x,t)^{-1}\Sigma g(x,t), \quad S^2=1. \tag{30}$$

With this identification, the gauge transformed Lax pair (29) can be expressed in terms of the spin field S (30) and its derivatives only, yielding

$$\tilde{\mathbf{U}}=-i\lambda S, \tag{31a}$$

$$\tilde{\mathbf{V}}=-4i\epsilon\lambda^3S+2\epsilon\lambda^2SS_x+i\epsilon\lambda(S_{xx}+\tfrac{3}{2}SS_x^2). \tag{31b}$$

In deriving (31) one has to use the following important identities:

$$SS_x=2g^{-1}\mathbf{A}g, \quad SS_x^2=-4g^{-1}\Sigma\mathbf{A}^2g \quad \text{and} \quad S_{xx}+SS_x^2=2g^{-1}\Sigma\mathbf{A}_xg.$$

The zero curvature condition of (31);

$$\tilde{\mathbf{U}}_t-\tilde{\mathbf{V}}_x+[\tilde{\mathbf{U}},\tilde{\mathbf{V}}]=0 \tag{32}$$

leads to the generalized Landau Lifshitz type equation

$$S_t+\epsilon S_{xxx}+\tfrac{3}{2}\epsilon(S_x^3+SS_{xx}S_x+SS_xS_{xx})=0 \tag{33}$$

with $S\in SU(3)/SU(2)$.

V. CONCLUSION

In this paper, we have bilinearized the higher order nonlinear Schrödinger equation via the SSE following Hirota's method. Hirota's method is an effective and important method to obtain multisoliton solutions. We have found explicitly one and two soliton solutions and recovered the ISM result of one soliton solution from that of Hirota's method after a specific choice of the parameter involved. The result related to higher soliton solutions are more complicated and will be given elsewhere. It is found that the SSE falls under a new class of bilinear forms.

The linear problem of SSE is a nontrivial one in the sense that the Lax operator corresponding to this dynamical equation is a 3×3 matrix, which makes the related Riccati equation more involved. However, by solving such coupled Riccati equations we are able to compute explicitly the infinite number of conserved quantities through the recursion relations and to show that terms with odd indices only contribute to the conserved charges like KdV and mKdV systems. This result establishes explicitly the integrability of the SSE in the Liouville sense. Finding out the Poisson bracket structures among the dynamical fields and consequently revealing the explicit form of the hierarchy of SSE from the conserved charges obtained here will be an interesting future problem.

The gauge equivalence of the GLLE with the SSE has been established here. This equivalence not only gives a direct relationship between the fields, which would help to find the soliton solutions of the GLLE using those of the SSE, but also yields explicit Lax operators for GLLE from which one would be able to extract the related higher conserved quantities in the similar way following the present results of SSE.

[1] R. S. Johnson, Proc. R. Soc. London, Ser. A **357**, 131 (1977); M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981); F. Calogero and W. Eckhaus, Inverse Probl. **3**, 229 (1987); F. Calogero and S. de Lillo, *ibid.* **3**, 633 (1987).
[2] Y. Kodama, J. Stat. Phys. **39**, 597 (1985); Y. Kodama and A. Hasegawa, IEEE J. Quantum Electron. **23**, 510 (1987).
[3] N. Sasa and J. Satsuma, J. Phys. Soc. Jpn. **60**, 409 (1991).
[4] L. A. Takhtajan and L. D. Faddeev, *Hamiltonian Approach in Soliton Theory* (Nauka, Moscow, 1986).
[5] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[6] K. Porsezein, M. Daniel, and M. Lakshmanan, *Proceedings of the International Conference on Nonlinear Evolution Equations and Dynamical Systems*, edited by V. G. Makhankov, I. Puzynin, and O. Pashaev (World Scientific, Singapore, 1993) p. 436.
[7] V. I. Zakharov and L. A. Takhtajan, Teor. Mat. Fiz. **38**, 26 (1979); A. Kundu, J. Math. Phys. **25**, 3433 (1984).
[8] M. Lakshmanan, Phys. Lett. A **61**, 53 (1977).
[9] R. Hirota, in *Solitons*, edited by R. K. Bullough and P. J. Courdrey (Springer, New York, 1980), p. 157; R. Hirota, Prog. Theor. Phys. **52**, 1498 (1974).
[10] J. Hietarinta, J. Math. Phys. **29**, 628 (1988).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Hirota bilinear approach to a new integrable differential-difference system

Xing-Biao Hu
CCAST, World Laboratory, P.O. Box 8730, Beijing 100080, China and State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific Engineering Computing, Academia Sinica, P.O. Box 2719, Beijing 100080, China

Yong-Tang Wu
Department of Computing Studies, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

Xian-Guo Geng
CCAST, World Laboratory, P.O. Box 8730, Beijing 100080, China and Department of Mathematics, Zhengzhou University, Henan 450052, China

(Received 19 May 1998; accepted for publication 9 November 1998)

A new integrable differential-difference system is proposed. By the dependent variable transformation, the system is transformed into multilinear form. By introducing an auxiliary variable, we further transform it into the bilinear form. A corresponding Bäcklund transformation for it is obtained. Furthermore a nonlinear superposition formula is presented. As an application of the obtained results, soliton solutions to the system are derived. © *1999 American Institute of Physics.* [S0022-2488(99)03202-8]

I. INTRODUCTION

While continuous soliton equations have been under active investigation, only a relatively small portion of the literature was devoted to the subject of discrete systems. However, this situation has been improved by now. Recently there has been a renewal of interest in discrete soliton systems. As a result, various approaches are being further developed to discrete soliton equations, among which Hirota's method is one of the most important ones. As is known, Hirota's bilinear method has played an important role in solitons and integrable systems since its inception.[1] In a series of papers,[2] Hirota discretized several soliton equations and obtained soliton solutions to the discretized equations.

The purpose of this paper is to propose a new differential-difference system and then study it by using Hirota's method. It is noted that recently the so-called Belov–Chaltikian lattice[3]

$$b_t(n)=b(n)(b(n+1)-b(n-1))-c(n)+c(n-1), \tag{1}$$

$$c_t(n)=c(n)(b(n+2)-b(n-1)), \tag{2}$$

and the Blaszak–Marciniak lattice[4]

$$a_t(n)=c(n+1)-c(n-1), \tag{3}$$

$$b_t(n)=a(n-1)c(n-1)-a(n)c(n), \tag{4}$$

$$c_t(n)=c(n)(b(n)-b(n+1)), \tag{5}$$

are transformed into the following bilinear forms:[5,6]

$$(D_t^2 e^{(1/2)D_n}-D_z e^{(1/2)D_n})f(n)\cdot f(n)=0, \tag{6}$$

$$(D_z e^{D_n} - D_t^2 e^{D_n} + 2e^{2D_n} - 2e^{D_n}) f(n)\cdot f(n) = 0, \tag{7}$$

and

$$(D_t^2 - 2D_z e^{D_n}) f(n)\cdot f(n) = 0, \tag{8}$$

$$(D_z D_t - 4\sinh^2(\tfrac{1}{2}D_n)) f(n)\cdot f(n) = 0, \tag{9}$$

respectively, where z is an auxiliary variable and the bilinear operators are defined as follows:[7–9]

$$D_z^m D_t^n a\cdot b \equiv \left(\frac{\partial}{\partial z} - \frac{\partial}{\partial z'}\right)^m \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial t'}\right)^n a(z,t) b(z',t')\Big|_{z'=z, t'=t},$$

$$\exp(\delta D_n) a(n)\cdot b(n) \equiv \exp\left[\delta\left(\frac{\partial}{\partial n} - \frac{\partial}{\partial n'}\right)\right] a(n) b(n')\Big|_{n'=n} = a(n+\delta) b(n-\delta).$$

On the other hand, in Ref. 10 a new, and relatively simple, procedure for finding new integrable differential-difference equation was reported. By combining these two thinkings, it is natural to search for new integrable systems such that the systems have the bilinear forms of the type (6) and (7) and (8) and (9) and the corresponding bilinear Bäcklund transformations could be found. With such a motivation in mind and after some tests and guesses, we now propose the following new system:

$$v_t(n) = v(n)(2u(n) - u(n+1) - u(n-1)), \tag{10}$$

$$u_t(n+1) + u_t(n-1) + v(n) u_t(n) + \tfrac{3}{4}(u(n+1) - u(n-1))^2 + \tfrac{1}{4}(u(n+1) + u(n-1) - 2u(n))^2 + \tfrac{1}{4}(v(n) - 1) = 0. \tag{11}$$

By the dependent variable transformation

$$u(n) = (\ln f(n))_t, \quad v(n) = \frac{f^2(n)}{f(n+1) f(n-1)}$$

and by use of (A1), (A2), (A3), (10) and (11) can be transformed into the following form:

$$[(3D_t^2 e^{D_n} + 3D_t^2 - e^{D_n} + 1) f(n)\cdot f(n)] f(n)^2 + [(1 - e^{D_n}) f(n)\cdot f(n)](D_t^2 f(n)\cdot f(n)) + D_t^2[(e^{D_n} - 1) f(n)\cdot f(n)]\cdot f^2(n) = 0. \tag{12}$$

Further, by introducing an auxiliary variable z and using (A4), we can decouple (12) into the following bilinear form:

$$(D_z D_t - 2e^{D_n} + 2) f(n)\cdot f(n) = 0, \tag{13}$$

$$(D_t^3 D_z + 6D_t^2 e^{D_n} + 6D_t^2 - 2e^{D_n} + 2) f(n)\cdot f(n) = 0. \tag{14}$$

Equation (11) may be rewritten as

$$u_t(n) + (T_+ + T_- + v(n))^{-1}[\tfrac{3}{4}(u(n+1) - u(n-1))^2 + \tfrac{1}{4}(u(n+1) + u(n-1) - 2u(n))^2 + \tfrac{1}{4}(v(n) - 1)] = 0, \tag{15}$$

where $T_\pm u(n) \equiv u(n\pm 1)$. Obviously (15) is nonlocal. However, to our knowledge, most integrable differential-difference systems appeared in literature such as the Toda lattice, Volterra equation are local. Therefore it is far from obvious to find such a transformation if there exists some transformation (point transformation, Miura-like transformation, etc.) which relates the system (10) and (11) to some other one already in the literature. On the other hand, from the viewpoint of bilinear formalism, to our knowledge it is the first time to consider the bilinear equations (13) and (14) simultaneously with z being an auxiliary variable, although (13) is just the bilinear form of the two-dimensional Toda lattice. Based on these explanations, it would be reasonable to view (13) and (14) as a new system.

The paper is organized as follows. In Sec. II we give a Bäcklund transformation for Eqs. (13) and (14). Then in Sec. III, we give a brief proof of a nonlinear superposition formula. Some particular solutions of Eqs. (13) and (14) are then found through this formula. Finally, Sec. IV summarizes the obtained results. An Appendix lists some bilinear operator identities made use of in this paper.

II. A BÄCKLUND TRANSFORMATION

In this section, we shall derive a Bäcklund transformation for Eqs. (13) and (14). The result obtained is:

Proposition 1: A Bäcklund transformation (BT) for (13) and (14) is

$$(D_z+\lambda^{-1}e^{-D_n}+\mu)f(n)\cdot g(n)=0, \tag{16}$$

$$(D_te^{-(1/2)D_n}-\lambda e^{(1/2)D_n}+\gamma e^{-(1/2)D_n})f(n)\cdot g(n)=0, \tag{17}$$

$$\begin{aligned}&[\lambda^{-1}D_t^3e^{-D_n/2}+3D_t^2e^{D_n/2}+3\lambda^{-1}\gamma D_t^2e^{-(1/2)D_n}+6\gamma D_te^{(1/2)D_n}+(6\gamma^2-1)e^{(1/2)D_n}\\&\quad+\omega e^{-(1/2)D_n}]f(n)\cdot g(n)=0,\end{aligned} \tag{18}$$

where λ, μ, γ, and ω are arbitrary constants.

Proof: Let $f(n)$ be a solution of Eqs. (13) and (14). If we can show that Eqs. (16)–(18) guarantee that the following two relations

$$P_1\equiv(D_zD_t-2e^{D_n}+2)g(n)\cdot g(n)=0,$$

$$P_2\equiv(D_t^3D_z+6D_t^2e^{D_n}+6D_t^2-2e^{D_n}+2)g(n)\cdot g(n)=0,$$

hold, then Eqs. (16)–(18) form a Bäcklund transformation.

In fact, same as the proof in Ref. 11, we know that $P_1=0$ can be proved using Eqs. (16) and (17). Thus it suffices to show that $P_2=0$. For this, by making use of (A5)–(A11), (16)–(18), we have

$$\begin{aligned}-f(n)^2P_2&=[(D_t^3D_z+6D_t^2e^{D_n}+6D_t^2-2e^{D_n}+2)f(n)\cdot f(n)]g(n)^2\\&\quad-f(n)^2[(D_t^3D_z+6D_t^2e^{D_n}+6D_t^2-2e^{D_n}+2)g(n)\cdot g(n)]\\&\quad+3[(D_zD_t-2e^{D_n}+2)f(n)\cdot f(n)][D_t^2g(n)\cdot g(n)]\\&\quad-3[D_t^2f(n)\cdot f(n)][(D_zD_t-2e^{D_n}+2)g(n)\cdot g(n)]\\&=2D_t^3(D_zf(n)\cdot g(n))\cdot f(n)g(n)\\&\quad+12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\&\quad-(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(D_te^{-(1/2)D_n}f(n)\cdot g(n))]\\&\quad-4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\end{aligned}$$

$$
\begin{aligned}
&= -2\lambda^{-1}D_t^3(e^{-D_n}f(n)\cdot g(n))\cdot f(n)g(n)\\
&\quad +12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad -(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(D_te^{-(1/2)D_n}f(n)\cdot g(n))]\\
&\quad -4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)[(D_t^3e^{-(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +3(D_te^{-(1/2)D_n}f(n)\cdot g(n))\cdot(D_t^2e^{-(1/2)D_n}f(n)\cdot g(n))]\\
&\quad +12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad -(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(D_te^{-(1/2)D_n}f(n)\cdot g(n))]\\
&\quad -4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)(D_t^3e^{-(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +12\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(D_t^2e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad -12\lambda^{-1}\gamma\sinh(\tfrac{1}{2}D_n)(e^{-(1/2)D_n}f(n)\cdot g(n))\cdot(D_t^2e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad -(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(D_te^{-(1/2)D_n}f(n)\cdot g(n))]\\
&\quad -4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)(D_t^3e^{-(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +12\sinh(\tfrac{1}{2}D_n)(D_t^2e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad -12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(D_te^{-(1/2)D_n}f(n)\cdot g(n))]+12\lambda^{-1}\gamma\sinh(\tfrac{1}{2}D_n)\\
&\quad \times(D_t^2e^{-(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad -(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(D_te^{-(1/2)D_n}f(n)\cdot g(n))]\\
&\quad -4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)(D_t^3e^{-(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +12\sinh(\tfrac{1}{2}D_n)(D_t^2e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +24D_t\cosh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(\gamma e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad +12\lambda^{-1}\gamma\sinh(\tfrac{1}{2}D_n)(D_t^2e^{-(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))\\
&\quad -4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f(n)\cdot g(n))\cdot(e^{-(1/2)D_n}f(n)\cdot g(n))
\end{aligned}
$$

$$
\begin{aligned}
&= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)(D_t^3 e^{-(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad +12\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad +24\gamma\sinh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad -(e^{(1/2)D_n} f(n)\cdot g(n))\cdot(D_t e^{-(1/2)D_n} f(n)\cdot g(n))]\\
&\quad +12\lambda^{-1}\gamma\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{-(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad -4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)(D_t^3 e^{-(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad +12\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad +24\gamma\sinh(\tfrac{1}{2}D_n)(D_t e^{(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad +24\gamma^2\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad +12\lambda^{-1}\gamma\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{-(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))\\
&\quad -4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n} f(n)\cdot g(n))\cdot(e^{-(1/2)D_n} f(n)\cdot g(n))=0.
\end{aligned}
$$

Thus we have completed the proof of proposition 1.

By using (16)–(18), we can easily obtain the following solutions from the trivial solution $f(n)=1$:

$$g(n)=1+\exp(\eta)$$

with

$$\lambda=\pm\sqrt{\frac{e^{(1/2)p}-e^{-(1/2)p}}{4e^{(1/2)p}-4e^{-(5/2)p}}},\quad \mu=-\lambda^{-1},\quad \gamma=\lambda,\quad \omega=-6\lambda^2+1,$$

where $\eta=pn+qz+rt+\eta^0$, $q=\lambda^{-1}(e^p-1)$, $r=\lambda(1-e^{-p})$.

III. A NONLINEAR SUPERPOSITION FORMULA

In the following, we shall simply denote, without confusion, $f(n,t)=f(n)$ or f. The result reached is

Proposition 2: Let f_0 be a solution of Eqs. (13) and (14) and suppose that f_i $(i=1,2)$ are solutions of (13) and (14) which are related to f_0 under the BT Eqs. (16)–(18) with parameters $(\lambda_i,\mu_i,\gamma_i,\omega_i)$, i.e.,

$$f_0 \xrightarrow{(\lambda_i,\mu_i,\gamma_i,\omega_i)} f_i \quad (i=1,2),\quad \lambda_1\lambda_2\neq 0,\ f_j\neq 0\ \ (j=0,1,2).$$

Then f_{12} defined by

$$\exp(-\tfrac{1}{2}D_n) f_0\cdot f_{12}=k[\lambda_1\exp(-\tfrac{1}{2}D_n)-\lambda_2\exp(\tfrac{1}{2}D_n)] f_1\cdot f_2 \quad (k \text{ is a nonzero constant}) \tag{19}$$

is a new solution which is related to f_1 and f_2 under the BT (16)–(18) with parameters $(\lambda_2,\mu_2,\gamma_2,\omega_2)$, $(\lambda_1,\mu_1,\gamma_1,\omega_1)$, respectively.

Proof: Same as the deduction in Refs. 11 and 6, we can show that

$$(D_z+\lambda_2^{-1}e^{-D_n}+\mu_2)f_1\cdot f_{12}=0, \tag{20}$$

$$(D_z+\lambda_1^{-1}e^{-D_n}+\mu_1)f_2\cdot f_{12}=0, \tag{21}$$

$$(D_te^{-(1/2)D_n}-\lambda_2e^{(1/2)D_n}+\gamma_2e^{-(1/2)D_n})f_1\cdot f_{12}=0, \tag{22}$$

$$(D_te^{-(1/2)D_n}-\lambda_1e^{(1/2)D_n}+\gamma_1e^{-(1/2)D_n})f_2\cdot f_{12}=0, \tag{23}$$

$$-D_zf_1\cdot f_2+(\mu_1-\mu_2)f_1f_2-\frac{1}{k\lambda_1\lambda_2}e^{-D_n}f_0\cdot f_{12}=0, \tag{24}$$

$$(\lambda_2D_te^{(1/2)D_n}+\lambda_1D_te^{-(1/2)D_n}-2\lambda_2\gamma_1e^{(1/2)D_n}+2\lambda_1\gamma_2e^{-(1/2)D_n})f_1\cdot f_2+\frac{1}{k}D_te^{-(1/2)D_n}f_0\cdot f_{12}=0. \tag{25}$$

Therefore in order to prove proposition 2, it suffices to show that

$$[\lambda_2^{-1}D_t^3e^{-D_n/2}+3D_t^2e^{D_n/2}+3\lambda_2^{-1}\gamma_2D_t^2e^{-(1/2)D_n}+6\gamma_2D_te^{(1/2)D_n}+(6\gamma_2^2-1)e^{(1/2)D_n} +\omega_2e^{-(1/2)D_n}]f_1\cdot f_{12}=0, \tag{26}$$

$$[\lambda_1^{-1}D_t^3e^{-(D_n/2)}+3D_t^2e^{(D_n/2)}+3\lambda_1^{-1}\gamma_1D_t^2e^{-(1/2)D_n}+6\gamma_1D_te^{(1/2)D_n}+(6\gamma_1^2-1)e^{(1/2)D_n} +\omega_1e^{-(1/2)D_n}]f_2\cdot f_{12}=0. \tag{27}$$

Since f_1 and f_2 are two solutions of Eqs. (13) and (14), we have, by use of (A5)–(A15), (19), (20), (22), (24), (25) and

$$f_0\xrightarrow{(\lambda_2,\mu_2,\gamma_2,\omega_2)}f_2,$$

that

$$\begin{aligned}
0=&[(D_t^3D_z+6D_t^2e^{D_n}+6D_t^2-2e^{D_n}+2)f_1\cdot f_1]f_2^2\\
&-[(D_t^3D_z+6D_t^2e^{D_n}+6D_t^2-2e^{D_n}+2)f_2\cdot f_2]f_1^2\\
&+3[(D_zD_t-2e^{D_n}+2)f_1\cdot f_1][D_t^2f_2\cdot f_2]\\
&-3[(D_zD_t-2e^{D_n}+2)f_2\cdot f_2][D_t^2f_1\cdot f_1]\\
=&2D_t^3(D_zf_1\cdot f_2)\cdot f_1f_2+12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f_1\cdot f_2)\cdot(e^{-(1/2)D_n}f_1\cdot f_2)\\
&-(e^{(1/2)D_n}f_1\cdot f_2)\cdot(D_te^{-(1/2)D_n}f_1\cdot f_2)]-4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f_1\cdot f_2)\cdot(e^{-(1/2)D_n}f_1\cdot f_2)\\
=&-\frac{2}{k\lambda_1\lambda_2}D_t^3(e^{-D_n}f_0\cdot f_{12})\cdot f_1f_2\\
&+12D_t\cosh(\tfrac{1}{2}D_n)[(D_te^{(1/2)D_n}f_1\cdot f_2)\cdot(e^{-(1/2)D_n}f_1\cdot f_2)-(e^{(1/2)D_n}f_1\cdot f_2)\cdot(D_te^{-(1/2)D_n}f_1\cdot f_2)]\\
&-4\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n}f_1\cdot f_2)\cdot(e^{-(1/2)D_n}f_1\cdot f_2)\\
=&-\frac{2}{k\lambda_1\lambda_2}e^{-(1/2)D_n}[(D_t^3e^{-(1/2)D_n}f_0\cdot f_2)\cdot(e^{-(1/2)D_n}f_1\cdot f_{12})-3(D_t^2e^{-(1/2)D_n}f_0\cdot f_2)\\
&\cdot(D_te^{-(1/2)D_n}f_1\cdot f_{12})
\end{aligned}$$

$$+3(D_t e^{-(1/2)D_n} f_0 \cdot f_2)\cdot(D_t^2 e^{-(1/2)D_n} f_1 \cdot f_{12}) - (e^{-(1/2)D_n} f_0 \cdot f_2)\cdot(D_t^3 e^{-(1/2)D_n} f_1 \cdot f_{12})]$$

$$+12 D_t \cosh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} f_1 \cdot f_2)\cdot(e^{-(1/2)D_n} f_1 \cdot f_2) - (e^{(1/2)D_n} f_1 \cdot f_2)\cdot(D_t e^{-(1/2)D_n} f_1 \cdot f_2)]$$

$$+\frac{4}{\lambda_1}\sinh\left(\frac{1}{2}D_n\right)[(\lambda_1 e^{-(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_1 \cdot f_2]\cdot(e^{(1/2)D_n} f_1 \cdot f_2)$$

$$= -\frac{2}{k\lambda_1\lambda_2} e^{-(1/2)D_n}\{[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n}) f_0 \cdot f_2]\cdot(e^{-(1/2)D_n} f_1 \cdot f_{12})$$

$$-(e^{-(1/2)D_n} f_0 \cdot f_2)\cdot[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n}) f_1 \cdot f_{12}]$$

$$-3\lambda_2(D_t^2 e^{-(1/2)D_n} f_0 \cdot f_2)\cdot(e^{(1/2)D_n} f_1 \cdot f_{12}) + 3\lambda_2(e^{(1/2)D_n} f_0 \cdot f_2)\cdot(D_t^2 e^{-(1/2)D_n} f_1 \cdot f_{12})\}$$

$$+12 D_t \cosh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} f_1 \cdot f_2)\cdot(e^{-(1/2)D_n} f_1 \cdot f_2) - (e^{(1/2)D_n} f_1 \cdot f_2)\cdot(D_t e^{-(1/2)D_n} f_1 \cdot f_2)]$$

$$+\frac{4}{k\lambda_1}\sinh\left(\frac{1}{2}D_n\right)(e^{-(1/2)D_n} f_0 \cdot f_{12})\cdot(e^{(1/2)D_n} f_1 \cdot f_2)$$

$$= -\frac{2}{k\lambda_1\lambda_2} e^{-(1/2)D_n}\{[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n}) f_0 \cdot f_2]\cdot(e^{-(1/2)D_n} f_1 \cdot f_{12})$$

$$-(e^{-(1/2)D_n} f_0 \cdot f_2)\cdot[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n}) f_1 \cdot f_{12}]\}$$

$$+\frac{12}{k\lambda_1} D_t \cosh\left(\frac{1}{2}D_n\right)[(D_t e^{-(1/2)D_n} f_0 \cdot f_{12})\cdot(e^{(1/2)D_n} f_1 \cdot f_2) + (e^{-(1/2)D_n} f_0 \cdot f_{12})$$

$$\cdot(D_t e^{(1/2)D_n} f_1 \cdot f_2)]$$

$$+12 D_t \cosh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} f_1 \cdot f_2)\cdot(e^{-(1/2)D_n} f_1 \cdot f_2) - (e^{(1/2)D_n} f_1 \cdot f_2)\cdot(D_t e^{-(1/2)D_n} f_1 \cdot f_2)]$$

$$+\frac{2}{k\lambda_1} e^{-(1/2)D_n}[(e^{(1/2)D_n} f_0 \cdot f_2)\cdot(e^{-(1/2)D_n} f_1 \cdot f_{12}) - (e^{-(1/2)D_n} f_0 \cdot f_2)\cdot(e^{(1/2)D_n} f_1 \cdot f_{12})]$$

$$= -\frac{2}{k\lambda_1\lambda_2} e^{-(1/2)D_n}\{[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_0$$

$$\cdot f_2]\cdot(e^{-(1/2)D_n} f_1 \cdot f_{12})$$

$$-(e^{-(1/2)D_n} f_0 \cdot f_2)\cdot[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_1 \cdot f_{12}]\}$$

$$-24\gamma_2 D_t \cosh(\tfrac{1}{2}D_n)(e^{-(1/2)D_n} f_1 \cdot f_2)\cdot(e^{(1/2)D_n} f_1 \cdot f_2)$$

$$= -\frac{2}{k\lambda_1\lambda_2} e^{-(1/2)D_n}\{[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_0 \cdot f_2]$$

$$\cdot(e^{-(1/2)D_n} f_1 \cdot f_{12})$$

$$-(e^{-(1/2)D_n} f_0 \cdot f_2)\cdot[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_1 \cdot f_{12}]\}$$

$$-\frac{24\gamma_2}{k\lambda_1} D_t \cosh\left(\frac{1}{2}D_n\right)(e^{-(1/2)D_n} f_0 \cdot f_{12})\cdot(e^{(1/2)D_n} f_1 \cdot f_2)$$

$$= -\frac{2}{k\lambda_1\lambda_2} e^{-(1/2)D_n}\{[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_0$$

$$\cdot f_2]\cdot(e^{-(1/2)D_n} f_1 \cdot f_{12})$$

$$-(e^{-(1/2)D_n} f_0 \cdot f_2)\cdot[(D_t^3 e^{-(1/2)D_n} + 3\gamma_2 D_t^2 e^{-(1/2)D_n} + 3\lambda_2 D_t^2 e^{(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_1 \cdot f_{12}]\}$$

$$-\frac{12\gamma_2}{k\lambda_1}e^{-(1/2)D_n}[(D_t e^{(1/2)D_n}f_0\cdot f_2)\cdot(e^{-(1/2)D_n}f_1\cdot f_{12})-(e^{(1/2)D_n}f_0\cdot f_2)\cdot(D_t e^{-(1/2)D_n}f_1\cdot f_{12})$$
$$+(D_t e^{-(1/2)D_n}f_0\cdot f_2)\cdot(e^{(1/2)D_n}f_1\cdot f_{12})-(e^{-(1/2)D_n}f_0\cdot f_2)\cdot(D_t e^{(1/2)D_n}f_1\cdot f_{12})]$$
$$=\frac{2}{k\lambda_1\lambda_2}e^{-(1/2)D_n}(e^{-(1/2)D_n}f_0\cdot f_2)\cdot[(D_t^3 e^{-(1/2)D_n}+3\gamma_2 D_t^2 e^{-(1/2)D_n}+3\lambda_2 D_t^2 e^{(1/2)D_n}$$
$$-\lambda_2 e^{(1/2)D_n}+6\lambda_2\gamma_2 D_t e^{-(1/2)D_n}+6\lambda_2\gamma_2^2 e^{(1/2)D_n}+\omega_2 e^{-(1/2)D_n})f_1\cdot f_{12}],$$

which implies that (26) holds. Similarly we can prove (27) also holds. Therefore we have completed the proof of the proposition 2.

As an application of the result, we can construct soliton solutions of Eqs. (13) and (14). Choose for example $f_0=1$, $c=1/(\lambda_1-\lambda_2)$. It is easily verified that

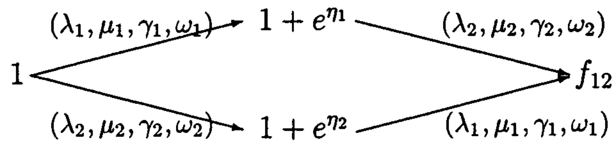

where

$$f_{12}=1+\frac{\lambda_1 e^{-p_1}-\lambda_2}{\lambda_1-\lambda_2}e^{\eta_1}+\frac{\lambda_1-\lambda_2 e^{-p_2}}{\lambda_1-\lambda_2}e^{\eta_2}+\frac{\lambda_1 e^{-p_1}-\lambda_2 e^{-p_2}}{\lambda_1-\lambda_2}e^{\eta_1+\eta_2}$$

with

$$\eta_i=p_i n+q_i z+r_i t+\eta_i^0,\quad q_i=\lambda_i^{-1}(e^{p_i}-1),\quad r_i=\lambda_i(1-e^{-p_i})$$

and

$$\lambda_i=\pm\sqrt{\frac{e^{(1/2)p_i}-e^{-(1/2)p_i}}{4e^{(1/2)p_i}-4e^{-(5/2)p_i}}},\quad \mu_i=-\lambda_i^{-1},\quad \gamma_i=\lambda_i,\quad \omega_i=-6\lambda_i^2+1.$$

In general, along this line, we can obtain multisoliton solutions for Eqs. (13) and (14) step by step.

IV. CONCLUSION AND DISCUSSIONS

A new integrable differential-difference system is proposed. By the dependent variable transformation, the system is transformed into multilinear form. By introducing an auxiliary variable, we further transform it into the bilinear form. A corresponding Bäcklund transformation for it is obtained. Furthermore a nonlinear superposition formula is presented. As an application of the obtained results, soliton solutions to the system are derived. As is known, it is of both theoretical and practical value to find as many new integrable systems as possible and to elucidate in depth their algebraic and geometric properties. In theory, it will greatly help to formulate a criterion for integrability, which is a long-standing open problem; and in practice it will provide us with dozens of nonlinear systems, which are of potential value in physical applications. Therefore we hope our proposed system (10) and (11) will model the behavior of some physical, or biological system. Moreover, it also would be of interest to study other algebraic and geometric properties for (10) and (11), e.g., to see if it has Hamiltonian structure and belongs to hierarchies associated to some linear problem. Besides, we can further consider the following extended form of (13) and (14):

$$(D_z D_t-2e^{D_n}+2)f(n)\cdot f(n)=0, \tag{28}$$

$$(D_t^3 D_z + 6D_t^2 e^{D_n} + 6D_t^2 + AD_t e^{D_n} - 2e^{D_n} + 2) f(n) \cdot f(n) = 0, \tag{29}$$

where A is an arbitrary constant.

ACKNOWLEDGMENTS

This work was partially supported by Hong Kong RGC/97-98/21, the National Natural Science Foundation of China (Project No. 19871082), and the Chinese Academy of Sciences.

APPENDIX: HIROTA BILINEAR OPERATOR IDENTITIES

The following bilinear operator identities hold for arbitrary functions *a*, *b*, *c*, and *d*:

$$(D_t^2 a \cdot a) b^2 + a^2 D_t^2 b \cdot b = 2(D_t^2 a \cdot b) ab - 2(D_t a \cdot b)^2, \tag{A1}$$

$$D_t^2 ca \cdot a = c_{tt} a^2 + c D_t^2 a \cdot a, \tag{A2}$$

$$D_t^2 a^2 \cdot a^2 = 2a^2 D_t^2 a \cdot a, \tag{A3}$$

$$(D_t^3 D_z a \cdot a) a^2 + (D_z D_t a \cdot a)(D_t^2 a \cdot a) = D_t^2 (D_z D_t a \cdot a) \cdot a^2, \tag{A4}$$

$$\begin{aligned}(D_t^3 D_z a \cdot a) b^2 - a^2 (D_t^3 D_z b \cdot b) + 3(D_z D_t a \cdot a)(D_t^2 b \cdot b) - 3(D_t^2 a \cdot a)(D_t D_z b \cdot b)\\ = 2D_t^3 (D_z a \cdot b) \cdot ab,\end{aligned} \tag{A5}$$

$$\begin{aligned}(D_t^2 e^{D_n} a \cdot a) b^2 - a^2 D_t^2 e^{D_n} b \cdot b + (D_t^2 a \cdot a)(e^{D_n} b \cdot b) - (e^{D_n} a \cdot a)(D_t^2 b \cdot b)\\ = 2D_t \cosh(\tfrac{1}{2} D_n)[(D_t e^{(1/2)D_n} a \cdot b) \cdot (e^{-(12)D_n} a \cdot b) - (e^{(1/2)D_n} a \cdot b) \cdot (D_t e^{-(1/2)D_n} a \cdot b)],\end{aligned} \tag{A6}$$

$$\begin{aligned}D_t^3 (e^{-D_n} a \cdot b) \cdot ab = -2 \sinh(\tfrac{1}{2} D_n)[(D_t^3 e^{-(1/2)D_n} a \cdot b) \cdot (e^{-(1/2)D_n} a \cdot b)\\ + 3(D_t e^{-(1/2)D_n} a \cdot b) \cdot (D_t^2 e^{-(1/2)D_n} a \cdot b)],\end{aligned} \tag{A7}$$

$$\begin{aligned}\sinh(\tfrac{1}{2} D_n)[(e^{(1/2)D_n} a \cdot b) \cdot (D_t^2 e^{-(1/2)D_n} a \cdot b) - (D_t^2 e^{(1/2)D_n} a \cdot b) \cdot (e^{-(1/2)D_n} a \cdot b)]\\ = -D_t \cosh(\tfrac{1}{2} D_n)[(D_t e^{(1/2)D_n} a \cdot b) \cdot (e^{-(1/2)D_n} a \cdot b) + (e^{(1/2)D_n} a \cdot b) \cdot (D_t e^{-(1/2)D_n} a \cdot b)],\end{aligned} \tag{A8}$$

$$\begin{aligned}D_t \cosh(\tfrac{1}{2} D_n)(e^{(1/2)D_n} a \cdot b) \cdot (e^{-(1/2)D_n} a \cdot b) = \sinh(\tfrac{1}{2} D_n)[(D_t e^{(1/2)D_n} a \cdot b) \cdot (e^{-(1/2)D_n} a \cdot b)\\ - (e^{(1/2)D_n} a \cdot b) \cdot (D_t e^{-(1/2)D_n} a \cdot b)],\end{aligned} \tag{A9}$$

$$D_t \cosh(\tfrac{1}{2} D_n) a \cdot a = 0, \tag{A10}$$

$$\sinh(\tfrac{1}{2} D_n) a \cdot a = 0, \tag{A11}$$

$$\begin{aligned}D_t^3 (e^{-D_n} a \cdot b) \cdot cd = e^{-(1/2)D_n}[(D_t^3 e^{-(1/2)D_n} a \cdot d) \cdot (e^{-(1/2)D_n} c \cdot b) - 3(D_t^2 e^{-(1/2)D_n} a \cdot d)\\ \cdot (D_t e^{-(1/2)D_n} c \cdot b) + 3(D_t e^{-(1/2)D_n} a \cdot d) \cdot (D_t^2 e^{-(1/2)D_n} c \cdot b) - (e^{-(1/2)D_n} a \cdot d)\\ \cdot (D_t^3 e^{-(1/2)D_n} c \cdot b)],\end{aligned} \tag{A12}$$

$$
\begin{aligned}
&e^{-(1/2)D_n}[(D_t^2 e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)-(e^{(1/2)D_n}a\cdot b)\cdot(D_t^2 e^{-(1/2)D_n}c\cdot d)]\\
&\quad = e^{-(1/2)D_n}[(e^{-(1/2)D_n}a\cdot b)\cdot(D_t^2 e^{(1/2)D_n}c\cdot d)-(D_t^2 e^{(1/2)D_n}a\cdot b)\cdot(e^{-(1/2)D_n}c\cdot d)]\\
&\qquad +2D_t\cosh(\tfrac{1}{2}D_n)[(D_t e^{-(1/2)D_n}a\cdot d)\cdot(e^{(1/2)D_n}c\cdot b)+(e^{-(1/2)D_n}a\cdot d)\cdot(D_t e^{(1/2)D_n}c\cdot b)],
\end{aligned}
\tag{A13}
$$

$$
\begin{aligned}
&2\sinh(\tfrac{1}{2}D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\
&\quad = e^{-(1/2)D_n}[(e^{(1/2)D_n}a\cdot d)\cdot(e^{-(1/2)D_n}c\cdot b)-(e^{-(1/2)D_n}a\cdot d)\cdot(e^{(1/2)D_n}c\cdot b)],
\end{aligned}
\tag{A14}
$$

$$
\begin{aligned}
&2D_t\cosh(\tfrac{1}{2}D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\
&\quad = e^{-(1/2)D_n}[(D_t e^{(1/2)D_n}a\cdot d)\cdot(e^{-(1/2)D_n}c\cdot b)-(e^{(1/2)D_n}a\cdot d)\cdot(D_t e^{-(1/2)D_n}c\cdot b)\\
&\quad +(D_t e^{-(1/2)D_n}a\cdot d)\cdot(e^{(1/2)D_n}c\cdot b)-(e^{-(1/2)D_n}a\cdot d)\cdot(D_t e^{(1/2)D_n}c\cdot b)].
\end{aligned}
\tag{A15}
$$

[1] R. Hirota, Phys. Rev. Lett. **27**, 1192 (1971).
[2] R. Hirota, J. Phys. Soc. Jpn. **43**, 1423 (1977); **43**, 2074 (1977); **43**, 2079 (1977); **45**, 321 (1978); **46**, 312 (1978); **50**, 3785 (1981).
[3] A. A. Belov and K. D. Chaltikian, Phys. Lett. B **309**, 268 (1993).
[4] M. Blaszak and K. Marciniak, J. Math. Phys. **35**, 4661 (1994).
[5] X. B. Hu and Z. N. Zhu, J. Phys. A **31**, 4755 (1998).
[6] X. B. Hu and Z. N. Zhu, J. Math. Phys. **39**, 4766 (1998).
[7] R. Hirota, *Direct Methods in Soliton Theory*, in Solitons, edited by R. K. Bullough and P. J. Caudrey (Springer, Berlin, 1980).
[8] R. Hirota and J. Satsuma, Prog. Theor. Phys. Suppl. **59**, 64 (1976).
[9] Y. Matsuno, *Bilinear Transformation Method* (Academic, New York, 1984).
[10] X. B. Hu, P. A. Clarkson, and R. Bullough, J. Phys. A L669 (1997).
[11] X. B. Hu, J. Phys. A **27**, 201 (1994).

High-frequency soliton-like waves in a relaxing medium

V. O. Vakhnenko[a)]
Division of Geodynamics of Explosion, Institute for Geophysics, 252054 Kyiv, Ukraine

(Received 23 March 1998; accepted for publication 6 October 1998)

A nonlinear evolution equation is suggested to describe the propagation of waves in a relaxing medium. It is shown that for low-frequency approach this equation is reduced to the KdVB equation. The high-frequency perturbations are described by a new nonlinear equation. This equation has ambiguous looplike solutions. It is established that a dissipative term, with a dissipation parameter less than some limit value, does not destroy these looplike solutions. © *1999 American Institute of Physics.* [S0022-2488(99)00503-4]

I. INTRODUCTION

As a rule the behavior of media under the action of high-frequency wave perturbations is not described in the framework of equilibrium models of continuum mechanics. So, to develop physical models for wave propagation through media with complicated inner kinetics, the notions based on the relaxational nature of a phenomenon are regarded to be promising and fruitful.

The description of nonlinear processes arising in different areas of research can often be reduced to the well-known Korteweg–de Vries (KdV) equation.[1,2] It turns out that low-frequency perturbations in a relaxing medium satisfy the KdV equation, too. The high-frequency perturbations are described by a new nonlinear evolution equation which has been investigated in Refs. 3 and 4. This equation has an ambiguous solution in the form of a solitary wave. This work deals with the looplike solutions of the model evolution equation. It is proved that the dissipative term, with a dissipation parameter less than some limit value, does not destroy the looplike solutions.

II. LOW-FREQUENCY AND HIGH-FREQUENCY PERTURBATIONS IN RELAXING MEDIUM

Thermodynamic equilibrium is disturbed owing to the propagation of fast perturbations in a medium. There are processes of the interaction that tend to return the equilibrium. The parameters characterizing this interaction are referred to as the inner variables unlike the macroparameters such as the pressure p, mass velocity u, and density ρ. In essence, the change of macroparameters caused by the changes of inner parameters is a relaxation process. From the nonequilibrium thermodynamics standpoint, the models of a relaxing medium are more general than the equilibrium models for describing the evolution of the wave perturbations.

We restrict our attention to barothropic media. An equilibrium state equation of a barothropic medium is a one-parameter equation. As a result of relaxation, an additional variable ξ (inner parameter) appears in the state equation. It defines the completeness of the relaxation process

$$p=p(\rho,\xi). \tag{2.1}$$

There are two limiting cases:

(*i*) the lack of the relaxation (inner interaction processes are frozen) $\xi=1$.

$$p=p(\rho,1)\equiv p_f(\rho), \tag{2.2}$$

(*ii*) the relaxation complete (there is the local thermodynamic equilibrium) $\xi=0$:

[a)]Electronic mail: vakhnenko@bitp.kiev.ua

$$p = p(\rho,0) \equiv p_e(\rho). \tag{2.3}$$

These relationships enable us to introduce the sound velocities for fast processes

$$c_f^2 = dp_f/d\rho \tag{2.4}$$

and for slow processes

$$c_e^2 = dp_e/d\rho. \tag{2.5}$$

Slow and fast processes are compared by means of the relaxation time τ_p. The dynamic state equation is written down in the form of the differential first-order equation

$$\tau_p\left(\frac{dp}{dt} - c_f^2\frac{d\rho}{dt}\right) + (p - p_e) = 0. \tag{2.6}$$

Clearly, for the fast processes ($\omega\tau_p \gg 1$) we have the relation (2.2), and for the slow ones ($\omega\tau_p \ll 1$) we obtain (2.3).

The substantiation of this equation within the framework of the thermodynamics of irreversible processes has been given in Refs. 5–8. The mechanism of the exchange (inner) processes is not defined concretely when the equation (2.6) is derived, and the thermodynamic and kinetic parameters appear in this equation only. These characteristics can be found by experiment. The dynamic state equation (2.6) enables us to take into account the exchange processes completely. We note that the phenomenological approach for describing the relaxation processes in hydrodynamics is developed in many works.[7–10] The dynamic state equation was used to describe the propagation of sound in a relaxing medium,[7] to take into account the exchange processes within media (gas–solid particles),[8] and to study wave fields in gas-liquid media[9] and in soils.[10] In most works the state equation has been derived from the concept of some concrete mechanism for the inner process.

To analyze the wave motion, we shall use the hydrodynamic equations: the law of the conservation of mass

$$\frac{\partial V}{\partial t} - \frac{\partial u}{\rho_0 \partial x} = 0 \tag{2.7}$$

and the law of the conservation of momentum

$$\frac{\partial u}{\partial t} + \frac{\partial p}{\rho_0 \partial x} = 0. \tag{2.8}$$

Here $V \equiv \rho^{-1}$ is specific volume and x is Lagrangian space coordinate.

The closed system of equations consists of two motion equations (2.7) and (2.8) and the dynamic state equation (2.6). The motion equations (2.7) and (2.8) are written in Lagrangian coordinates, since the state equation (2.6) is related to the element of the mass of medium.

Let us consider a small perturbation $p' < p_0$. The state equations for fast [(2.2)] and slow [(2.3)] processes are considered to be known. They can be expanded as the power series with accuracy $O(p'^2)$

$$V_f(p_0 + p') = V_0 - c_f^{-2} V_0^2 p' + \frac{1}{2}\left.\frac{d^2 V_f}{dp^2}\right|_{p=p_0} p'^2 + \ldots,$$

$$V_e(p_0+p')=V_0-c_e^{-2}V_0^2p'+\frac{1}{2}\frac{d^2V_e}{dp^2}\bigg|_{p=p_0}p'^2+\ \ldots .$$

Hereafter, the velocities c_e, c_f are related to initial pressure p_0. Combining these two relationships with the motion equations (2.7) and (2.8), we obtain the equation in one unknown (the dash in p' is omitted):

$$\tau_p\frac{\partial}{\partial t}\left(\frac{\partial^2 p}{\partial x^2}-c_f^{-2}\frac{\partial^2 p}{\partial t^2}+\frac{1}{2V_0^2}\frac{d^2V_f}{dp^2}\bigg|_{p=p_0}\frac{\partial^2 p^2}{\partial t^2}\right)+\left(\frac{\partial^2 p}{\partial x^2}-c_e^{-2}\frac{\partial^2 p}{\partial t^2}+\frac{1}{2V_0^2}\frac{d^2V_e}{dp^2}\bigg|_{p=p_0}\frac{\partial^2 p^2}{\partial t^2}\right)=0. \tag{2.9}$$

A similar equation has been obtained in Ref. 5, though without nonlinear terms.

Now we shall show that for low-frequency perturbations the equation (2.9) is reduced to the Korteweg–de Vries–Burgers (KdVB) equation, while for high-frequency waves we shall obtain the equation with hydrodynamic nonlinearity and term that appeared in the Klein–Gordon equation.

To analyze the equation (2.9), let us apply the multiscale method.[11,12] The value $\varepsilon\equiv\tau_p\omega$ is chosen to be small (large) parameter where the quantity ω is the characteristic frequency of wave perturbation. For the sake of convenience we rewrite the equation (2.9) as follows:

$$\tau_p\omega\frac{\partial}{\partial t\omega}\left(\frac{\partial^2 p}{\partial (x\omega)^2}-c_f^{-2}\frac{\partial^2 p}{\partial (t\omega)^2}+\alpha_f\frac{\partial^2 p^2}{\partial (t\omega)^2}\right)+\left(\frac{\partial^2 p}{\partial (x\omega)^2}-c_e^{-2}\frac{\partial^2 p}{\partial (t\omega)^2}+\alpha_e\frac{\partial^2 p^2}{\partial (t\omega)^2}\right)=0, \tag{2.10}$$

$$\alpha_f=\frac{1}{2V_0^2}\frac{d^2V_f}{dp^2}\bigg|_{p=p_0},\qquad \alpha_e=\frac{1}{2V_0^2}\frac{d^2V_e}{dp^2}\bigg|_{p=p_0},$$

and introduce new independent variables

$$T_0=t\omega,\quad X_0=x\omega,\quad T_{-2}=t\omega/\varepsilon^2,\quad X_{-2}=x\omega/\varepsilon^2.$$

It is precisely these variables that cause the equations, obtained within the framework of multiscale method[11,12]

$$O(\varepsilon^{+1}):\frac{\partial}{\partial T_0}\left(\frac{\partial^2 p}{\partial X_0^2}-c_f^{-2}\frac{\partial^2 p}{\partial T_0^2}+\alpha_f\frac{\partial^2 p^2}{\partial T_0^2}\right)=0,$$

$$O(\varepsilon^{0}):\frac{\partial^2 p}{\partial X_0^2}-c_e^{-2}\frac{\partial^2 p}{\partial T_0^2}+\alpha_e\frac{\partial^2 p^2}{\partial T_0^2}=0,$$

$$O(\varepsilon^{-1}):\left(\frac{\partial^3}{\partial X_0^2\partial T_{-2}}+2\frac{\partial^3}{\partial T_0\partial X_0\partial X_{-2}}\right)p-3c_f^{-2}\frac{\partial^3 p}{\partial T_0^2\partial T_{-2}}+3\alpha_f\frac{\partial^3 p^2}{\partial T_0^2\partial T_{-2}}=0,$$

$$O(\varepsilon^{-2}):\frac{\partial^2 p}{\partial X_0\partial X_{-2}}-c_e^{-2}\frac{\partial^2 p}{\partial T_0\partial T_{-2}}+\alpha_e\frac{\partial^2 p^2}{\partial T_0\partial T_{-2}}=0, \tag{2.11}$$

$$O(\varepsilon^{-3}):\left(\frac{\partial^3}{\partial T_0\partial X_{-2}^2}+2\frac{\partial^3}{\partial X_0\partial X_{-2}\partial T_{-2}}\right)p-3c_f^{-2}\frac{\partial^3 p}{\partial T_0\partial T_{-2}^2}+3\alpha_f\frac{\partial^3 p^2}{\partial T_0\partial T_{-2}^2}=0,$$

$$O(\varepsilon^{-4}):\frac{\partial^2 p}{\partial X_{-2}^2}-c_e^{-2}\frac{\partial^2 p}{\partial T_{-2}^2}+\alpha_e\frac{\partial^2 p^2}{\partial T_{-2}^2}=0,$$

$$O(\varepsilon^{-5}):\frac{\partial}{\partial T_{-2}}\left(\frac{\partial^2 p}{\partial X_{-2}^2}-c_f^{-2}\frac{\partial^2 p}{\partial T_{-2}^2}+\alpha_f\frac{\partial^2 p^2}{\partial T_{-2}^2}\right)=0,$$

to be partially uncoupled. The two leading equations depend on T_0 and X_0 only, while the last two equations include the independent variables T_{-2} and X_{-2} only. Thus, the low-frequency perturbations are described by the two leading equations, and the high-frequency perturbations are described by the last two equations. An interaction between these perturbations is described by the three center equations.

Let us write out the motion equations for low-frequency and high-frequency perturbations in the initial variables x and t. For low-frequency perturbations the main terms $\partial^2 p/\partial x^2$ and $c_e^{-2}\partial^2 p/\partial t^2$ appear in the second equation of the system (2.11), while for high-frequency perturbations the main terms $\partial^2 p/\partial x^2$ and $c_f^{-2}\partial^2 p/\partial t^2$ appear in the seventh equation of (2.11).

For low-frequency perturbations $(\tau_p\omega\ll 1)$ propagating in one direction, we obtain an evolution equation

$$\frac{\partial p}{\partial t}+c_e\frac{\partial p}{\partial x}+\alpha_e c_e^3 p\frac{\partial p}{\partial x}-\beta_e\frac{\partial^2 p}{\partial x^2}+\gamma_e\frac{\partial^3 p}{\partial x^3}=0,$$

$$\beta_e=\frac{c_e^2\tau_p}{2c_f^2}(c_f^2-c_e^2),\qquad \gamma_e=\frac{c_e^3\tau_p^2}{8c_f^4}(c_f^2-c_e^2)(c_f^2-5c_e^2). \tag{2.12}$$

This equation can be obtained in the following way. A dispersion relation for the linearized equation (2.10) can be written down with an accuracy $O(k^3)$ in the form $\omega=c_e k+i\beta_e k^2-\gamma_e k^3$, if the terms $c_e^{-1}\partial p/\partial t$ and $\partial p/\partial x$ are the main ones. For this dispersion relation we write a linear equation in which a nonlinear term is reconstructed in agreement with the initial equation.

The equation (2.12) is the well-known KdVB equation. It is encountered in many chapters of physics to describe nonlinear wave processes.[1] In Ref. 2 it was shown how hydrodynamic equations reduce to either the KdV or Burgers equation according to the choices for the state equation and the generalized force when analyzing the gasdynamical waves, waves in shallow water,[2] hydrodynamic waves in cold plasma,[13] or ion-acoustic waves in cold plasma.[14] The KdV equation $(\beta_e=0)$ has stationary solutions (solitons). In the case of $\beta_e\neq 0$ the stationary solutions of the equation (2.12) are known also.[15]

For high-frequency perturbations $(\tau_p\omega\gg 1)$, using the last two equations of the system (2.11), we get the following evolution equation:

$$\frac{\partial^2 p}{\partial x^2}-c_f^{-2}\frac{\partial^2 p}{\partial t^2}+\alpha_f c_f^2\frac{\partial^2 p^2}{\partial x^2}+\beta_f\frac{\partial p}{\partial x}+\gamma_f p=0,$$

$$\beta_f=\frac{c_f^2-c_e^2}{\tau_p c_e^2 c_f},\qquad \gamma_f=\frac{c_f^4-c_e^4}{2\tau_p^2 c_e^4 c_f^2}. \tag{2.13}$$

In addition to the nonlinear term with coefficient α_f, the equation has dissipative $\beta_f\,\partial p/\partial x$ and dispersive $\gamma_f p$ terms. If $\alpha_f=\beta_f=0$, this is a linear Klein–Gordon equation. There is a Green's function for this equation[16,17] that enables us to find the solution in quadrature, at least. Numerical solutions of the Klein–Gordon equation modeling the propagation of high-frequency perturbations in gas–liquid media have been presented in Ref. 17. Whitham's monograph[18] has also described a similar evolution equation for high-frequency perturbations, but its form coincides with that of Eq. (2.13) only when $\alpha_f=0$ and $\gamma_f=0$.

Landau and Lifshitz have shown that for high frequencies the dissipative term under high transport of heat agrees with corresponding term in the equation (2.13) (see section 79 and 81 in Ref. 7). Thus, the dynamic state equation (2.6) enable us to take into account the dissipative processes completely. But the form of the dissipative terms describing the inner exchange processes (transport of heat and momentum) are different for the high and low frequencies.

We call attention to the fact that the dispersion relations $\omega=\omega(k)$ for the linearized equations (2.12) and (2.13) have been restricted by the finite power series in k and in k^{-1}, respectively:

$$\omega=c_e k+i\beta_e k^2-\gamma_e k^3, \quad \tau_p\omega\ll 1,$$

$$c_f^{-2}\omega^2=k^2+i\beta_f k-\gamma_f, \quad \tau_p\omega\gg 1.$$

In the general case the equation (2.13) has been investigated insufficiently. It is likely that this is connected with the fact noted by Whitham[18] that the high-frequency perturbations attenuate very fast. However, in Ref. 18 the evolution equation without nonlinear and dispersive terms was considered. Certainly, the lack of such terms restricts the class of solutions. At least, there is no solution in the form of a solitary wave which is caused by nonlinearity and dispersion.

The studies of the equation (2.13) have some scientific interest both from the viewpoint of the investigation of the propagation of high-frequency perturbations and from the viewpoint of the existence of stable wave formations.

III. EVOLUTION EQUATION FOR HIGH-FREQUENCY PERTURBATIONS

The equation (2.13), which we are interested in, is written down in dimensionless form. Let us restrict our consideration to the propagation of high-frequency waves in positive direction x, then with necessary accuracy we can write the operator

$$\frac{\partial^2}{\partial x^2}-c_f^{-2}\frac{\partial^2}{\partial t^2}=\left(\frac{\partial}{\partial x}-c_f^{-1}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x}+c_f^{-1}\frac{\partial}{\partial t}\right)\rightarrow 2\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}+c_f^{-1}\frac{\partial}{\partial t}\right)$$

(for example, see section 93 in Ref. 7). In the moving coordinates system with velocity c_f, the equation has the form in dimensionless variables $\tilde{x}=\sqrt{\gamma_f/2}(x-c_f t)$, $\tilde{t}=\sqrt{\gamma_f/2}c_f t$, $\tilde{u}=\alpha_f c_f^2 p$ (tilde over variables $\tilde{x},\tilde{t}$, $\tilde{u}$ is omitted):

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}\right)u+\alpha\frac{\partial u}{\partial x}+u=0. \tag{3.1}$$

The constant $\alpha=\beta_f/\sqrt{2\gamma_f}$ is always positive.

The equation (3.1) without the dissipative term has the form of the nonlinear equation[3,4]

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}\right)u+u=0. \tag{3.2}$$

These equations are related to that of Whitham[18] with the kernels $K(x)=\frac{1}{2}[\alpha(2\Theta(x)-1)+|x|]$ and $K(x)=\frac{1}{2}|x|$ (see Eq. (2) in Ref. 3) and are written as

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+\alpha u+\frac{1}{2}\int_{-\infty}^{\infty}|x-s|\frac{\partial u}{\partial s}ds=0, \tag{3.3}$$

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+\frac{1}{2}\int_{-\infty}^{\infty}|x-s|\frac{\partial u}{\partial s}ds=0, \tag{3.4}$$

where $\Theta(x)$ is a Heaviside function. There is no derivative in the dissipative term αu of Eq. (3.3).

Papers[3,4] are devoted to the analysis of the equation (3.2). In Ref. 4 it was named Vakhnenko's equation. The equation (3.2) has two families of traveling wave solutions.[3,4] In one case the solutions have looplike form (see Fig. 1 in Ref. 3). Only in this case is there a solitary wave solution. In Ref. 4 it is predicted that both families of solutions are stable to long wavelength perturbations. The existence of singular points, at which the derivatives tend to infinity, required the application of a nonstandard method.[4] The ambiguous structure is similar to the loop soliton solution to an equation that models a stretched rope.[19] The looplike solitons on a vortex filament were investigated by Hasimoto[20] and Lamb, Jr.[21]

The material described below deals with the ambiguous looplike solutions of the equation (3.1). From the mathematical point of view the ambiguous solution does not present difficulties while the physical interpretation of ambiguity always has some difficulties. In this connection the problem of ambiguous solutions is regarded to be important. The problem consists in whether the ambiguity has a physical nature or is related to the incompleteness of mathematical model, in particular to the lack of dissipation.

We will consider the problem related to the singular points when the dissipation takes place. At these points the dissipative term $\alpha\,\partial u/\partial x$ tends to infinity. The question arises: are there solutions of the equation (3.1) in a looplike form? That the dissipation is likely to destroy the looplike solutions can be associated with the following well-known fact.[1] For a simplest nonlinear equation without dispersion and dissipation,

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=0, \tag{3.5}$$

any initial smooth solution with boundary conditions

$$u|_{x\to+\infty}=0, \quad u|_{x\to-\infty}=u_0=\text{const}>0$$

becomes ambiguous in the final analysis. When the dissipation is considered, we have a Burgers equation[22]

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+\mu\frac{\partial^2 u}{\partial x^2}=0.$$

The dissipative terms of this equation and Eq. (2.13) for low frequency are coincided. The inclusion of the dissipative term transforms the solutions so that they cannot be ambiguous as a result of evolution. The wave parameters are always unambiguous. What happens in our case for the high frequency when the dissipative term has the form αu [Eq. (3.3)]? Will the inclusion of dissipation give rise to unambiguous solutions? It turns out that, and here this has been proved, the dissipative term, with a dissipation parameter less than some limit value, does not destroy the looplike solutions. A physical interpretation is given to ambiguous solutions.

IV. AMBIGUOUS SOLUTIONS

Let us pass to the coordinates in which the equation (3.2) has stationary periodic solutions [see Eq. (7) in Ref. 3]

$$\eta=x-vt, \quad \tau=t, \tag{4.1}$$

where v is a nonzero constant. Eq. (3.2) has looplike solutions when $v>0$. Parkes noted[4] that there are no stationary periodic solution of (3.2) when $v=0$. After substitution of

$$z = u - v$$

into Eq. (3.1) we get the evolution equation

$$z_{\tau\eta} + (z z_\eta)_\eta + (z + v) + \alpha z_\eta = 0. \tag{4.2}$$

We investigate the solution behavior within the neighborhood of singular points $z=0$ where $z_\eta \to \pm\infty$ and $z_\tau \ll z_\eta$. Therefore in the investigated equation (4.2) we neglect the term z in comparison with v, and also discard the term z_τ, to obtain

$$(z z_\eta)_\eta + v + \alpha z_\eta = 0. \tag{4.3}$$

It is convenient to use the inverse function $\eta = \eta(z)$. Taking into account $z_\eta = 1/\eta_z$ and $z_{\eta\eta} = -\eta_{zz}/\eta_z^3$, Eq. (4.3) is rewritten as

$$-z\,\eta_{zz} + v\,\eta_z^3 + \alpha\,\eta_z^2 + \eta_z = 0.$$

Introducing the definition $q \equiv \eta_z$, this equation can be integrated to obtain

$$\int \frac{dq}{q(vq^2 + \alpha q + 1)} = \int \frac{dz}{z}.$$

Depending on the sign of the quantity $1 - \alpha^2/4v$, the latter expression has two different forms. We have introduced the critical value α^* of the parameter α defined by

$$\alpha^* = 2\sqrt{v}. \tag{4.4}$$

For $\alpha < \alpha^*$ (i.e., $1 - \alpha^2/4v > 0$), we get

$$\ln\left[\frac{z^2}{q^2}(vq^2 + \alpha q + 1)\right] = -\frac{2\alpha}{\sqrt{4v - \alpha^2}} \mathrm{tg}^{-1} \frac{2vq + \alpha}{\sqrt{4v - \alpha^2}} + \ln c_1, \tag{4.5}$$

and for $\alpha > \alpha^*$ (i.e., $1 - \alpha^2/4v < 0$), we have

$$\ln\left[\frac{z^2}{q^2}(vq^2 + \alpha q + 1)\right] = \frac{\alpha}{\sqrt{\alpha^2 - 4v}} \ln\left|\frac{2vq + \alpha + \sqrt{\alpha^2 - 4v}}{2vq + \alpha - \sqrt{\alpha^2 - 4v}}\right| + \ln c_2. \tag{4.6}$$

We analyze the expression (4.5). First let us verify the special case $\alpha = 0$. We have

$$\frac{z^2}{q^2}(vq^2 + 1) = c_1,$$

or

$$vz^2 + \frac{1}{4}(z^2)_\eta^2 = c_1.$$

Hence in the vicinity of $z = 0$,

$$\eta + \eta_0 = \pm\frac{1}{2}\int \frac{dz^2}{\sqrt{c_1 - vz^2}} = \mp\frac{\sqrt{c_1 - vz^2}}{v}.$$

We arrive at the result given in Ref. 3, namely that with the lack of dissipation ($\alpha=0$) the integral curves pass over an ellipse at $z=0$.

Now we investigate the case $0<\alpha<\alpha^*$. It is easy to show that the r.h.s. of (4.5) is always bounded for any value $q\equiv z_\eta^{-1}$. In the neighborhood of $z=0$ the r.h.s. of (4.5) is close to value

$$-\frac{2\alpha}{\sqrt{4v-\alpha^2}}\,\mathrm{tg}^{-1}\frac{\alpha}{\sqrt{4v-\alpha^2}}+\ln c_1\equiv\ln c_3.$$

Consequently, we arrive at the equation

$$\frac{z^2}{q^2}(vq^2+\alpha q+1)=c_3.$$

Even not integrating this equation, it is easy to show that at $z=0$ we must have $q=0$ since in general $c_3\neq 0$. This means that at $z=0$ the derivatives have the values

$$\eta_z=0,\ z_\eta=\pm\infty.$$

At $z=0$ the solution becomes ambiguous.

In the case $\alpha>\alpha^*$ there is the solution

$$z=0,\quad q=\eta_z\neq 0,\qquad z_\eta\neq\pm\infty.$$

In fact, at $z=0$ we obtain from the r.h.s. of (4.6)

$$q=\eta_z=-\frac{\alpha}{2v}-\frac{\sqrt{\alpha^2-4v}}{2v}\neq 0. \tag{4.7}$$

Thus, the derivative z_η at $z=0$ is bounded by a finite value. The solution is always unambiguous.

Let us consider the solution behavior in the neighborhood of $z=0$ as $\alpha\to\alpha^*$. We first consider the case $\alpha\to\alpha^*-0$. According to (4.5) the r.h.s. of this equation tends to minus infinity, i.e., at $z\approx 0$ we have $q=\eta_z\neq 0$. Consequently, there is no looplike solution.

When $\alpha\to\alpha^*+0$ there is also a solution with $q=\eta_z\neq 0$ at $z=0$. The root $q=0$ at $z=0$ seems possible in this case since (4.6) transforms to

$$\ln\left[\frac{z^2}{q^2}(vq^2+\alpha q+1)\right]=\frac{2\alpha}{2vq+\alpha}+\ln c_2. \tag{4.8}$$

However, as appears from (4.7), the r.h.s. of the equation (4.8) tends to minus infinity so that $q\neq 0$ at $z=0$. Therefore, in the case $\alpha\to\alpha^*$ the dissipation destroys the looplike solutions.

We have proved the following statement. For values of $\alpha<\alpha^*$ the inclusion of the dissipative term does not change the looplike solutions of equation (3.1), while for $\alpha\geqslant\alpha^*$ there is no solution with an infinite gradient.

The common form of the dissipative term for high-frequency perturbations αu (which does not depend on the nature of the exchange processes) cannot preclude the possibility of a formation of a multi-valued solution from an initial single-valued profile. In this case there are the infinite gradients in contrast to the profiles of a wave for the low frequencies when the dissipative term has the form $\beta\,\partial^2u/\partial x^2$.

The problem of a multi-valued solution can be forestalled in a following way. The equation (3.2) can be rewritten into new independent variables $y=y(x,t)$ and $t_1=t$ so that the dependent variable $u=u(y,t_1)$ will be a single-valued function of y. These variables have been defined by the relationships

$$\varphi dy = dx - u dt, \quad t_1 = t$$

in which the equation (3.2) has been reduced to a system of the equations in the unknowns u and φ:

$$\frac{\partial \varphi}{\partial t_1} = \frac{\partial u}{\partial y}, \quad \frac{\partial^2 \varphi}{\partial t_1^2} + u\varphi = 0.$$

For example, for a one-soliton solution we have $\varphi = 1 - u/v$ [see Eqs. (12) and (14) in Ref. 3]. In the space of new variables y and t_1 a solution is a single-valued function [Eq. (13) from Ref. 3]. Each state has been uniquely defined by the variable y at any time t.

Considering the dependent variable u and the coordinate x as the functions of new variable y we solve the problem of the ambiguous solution. A number of the states with their thermodynamic parameters can occupy one microvolume, but these states are distinguished by the coordinate y. It is assumed that the interaction between the separated states occupying one microvolume can be neglected in comparison with the interaction between the particles of one thermodynamic state. Even if we shall take into account the interaction between the separated states in accord with the dynamic state equation (2.6), then for high frequencies the dissipative term arises which is similar to the corresponding term in Eq. (2.13), but with the other relaxation time. In this sense the separated terms are distributed in space, but describing the wave process we consider them as interpenetratable. The similar situation, when several components with different hydrodynamic parameters occupied one microvolume, has been assumed in the mixture theory (see, for instance Refs. 23 and 24). Such a fundamental assumption in the theory of mixtures is physically impossible (see Ref. 23, p. 7), but it is appropriate in the sense that separated components are multi-velocity interpenetratable continua.

Thus in the frameworks of this model approach, the high-frequency perturbation can be described by the multi-valued functions.

V. CONCLUSIONS

The KdV and KdVB equations are employed to describe a number of evolution processes when the low-frequency approach turns out to be adequate. In these cases thermodynamic parameters of a medium are close to the equilibrium values, the microvolume state is defined by one set of thermodynamic values, and the disturbance from the equilibrium is taken into account by means of expansion in gradients.[25] If the low-order expansions within the framework of such an approach give rise to an inadequate description, we could take into account the terms of higher order and as a result consider higher frequencies. For example, if Eq. (3.5) has an ambiguous solution (or discontinuous solution), the improvement of models by means of adding higher degree derivatives excludes the ambiguous solutions. So, in the low-frequency approach, an ambiguity is connected with the incompleteness of the mathematical model.

In contrast to this, in models for the propagation of high-frequency perturbations the disturbance from the ''frozen'' state is taken into account by means of expansion in integral terms [see Eq. (3.3) and (3.4)]. The integral terms contain the prehistory of the process. We have just established that a higher order of expansion (in particular, the dissipative term) for the high-frequency evolution equation still allows ambiguous solutions. Consequently, the ambiguity of solution does not relate to the incompleteness of the mathematical model, in particular to the lack of dissipation. In addition there is the space of new independent variables where the solution is the single-valued function.

The following three circumstances show that in the framework of the approach considered here there are the multi-valued solutions when we model the high-frequency wave processes: (1) All parts of looplike solutions are stable to perturbations; this was proved by Parkes.[4] (2) The dissipation does not destroy the looplike solutions (the result of this work). (3) The investigation regarding the interaction of the solitons has shown that it is necessary to take into account the whole ambiguous solution, and not just the separate parts.

It is necessary to note that the substantiation of the nonlinear evolution equation (3.1) within the framework of statistic physics remains an important problem. At present this problem is too difficult since it is connected with the description of high nonequilibrium systems.

ACKNOWLEDGMENTS

The author would like to express his thanks to Professor V. A. Danylenko for fruitful discussions of the results.

[1] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. C. Morris, *Solitons and Nonlinear Wave Equations* (Academic, London, 1984).
[2] C. H. Su and C. S. Gardner, J. Math. Phys. **10**, 536 (1969).
[3] V. A. Vakhnenko, J. Phys. A **25**, 4181 (1992).
[4] E. J. Parkes, J. Phys. A **26**, 6469 (1993).
[5] J. F. Clarke, Ann. Phys. Fr. **9**, 211 (1984).
[6] V. A. Danylenko, V. V. Sorokina, and V. A. Vladimirov, J. Phys. A **26**, 7125 (1993).
[7] L. D. Landau and E. M. Lifshitz, *Fluids Mechanics* (Pergamon, New York, 1988).
[8] G. P. Yasnikov and V. S. Belousov, J. Eng. Phys. **35**, 872 (1978) (in Russian).
[9] V. E. Nakoryakov, B. G. Pokusaev, and I. R. Shreber, *The Waves Propagation in Gas- and Steam Liquid Media* (Institute Heatphysics Press, Novosibirsk, 1983) (in Russian).
[10] G. M. Lyakhov, *Waves in Soils and Porous Multicomponent Media* (Nauka, Moscow, 1982) (in Russian).
[11] A. H. Nayfeh, *Perturbation Methods* (Wiley, New York, 1973).
[12] Yu. A. Mitropolsky, A. M. Samoilenko, and D. I. Martinyuk, *Systems of Evolution Equations with Periodic and Quasiperiodic Coefficients* (Kluwer Academic, Dordrecht, 1993).
[13] C. S. Gardner and G. K. Morikawa, Courant Institute of Math. Sciences Rep. NYU 9082 (1960).
[14] H. Washimi and T. Taniuti, Phys. Rev. Lett. **17**, 996 (1966).
[15] N. A. Kudryashov, Sov. Phys. Dokl. **34**, 798 (1989) (English translation).
[16] V.S. Vladimirov, *Equations of Mathematical Physics* (Dekker, New York, 1971).
[17] K. V. Malykh and I. L. Ogorodnikov, in *Dynamics of Continuous Media* (Hydrodynamics Institute, Novosibirsk, 1977), pp. 140–3 (in Russian).
[18] G.B. Whitham, *Linear and Nonlinear Waves* (Wiley-Interscience, New York, 1974).
[19] K. Konno, Y. H. Ichikawa, and M. Wadati, J. Phys. Soc. Jpn. **50**, 1025 (1981).
[20] H. Hasimoto, J. Fluid Mech. **51**, 477 (1972).
[21] G. L. Lamb, Jr, J. Math. Phys. **18**, 1654 (1977).
[22] J. M. Burgers, Adv. Mech. **1**, xxx (1948).
[23] K.R. Rajagopal and L. Tao, *Mechanics of Mixtures* (World Scientific, Singapore, 1995).
[24] V. V. Struminskii, *Mechanics and Technical Progress* (Nauka, Moscow, 1980) (in Russian).
[25] V. N. Rudyak, *The Statistical Theory of Dissipative Processes in Gases and Liquids* (Nauka, Moscow, 1987) (in Russian).

Asymptotically Schwarzschild space–times

Uchida Gen
Department of Earth and Space Science, Graduate School of Science, Osaka University, Toyonaka 560-0043, Japan
and Department of Physics, The University of Tokyo, Tokyo 113-0033, Japan

Tetsuya Shiromizu[a)]
DAMTP, University of Cambridge, Silver Street, Cambridge CB3 9EW, United Kingdom;
Department of Physics, The University of Tokyo, Tokyo 113-0033, Japan;
and Research Center for the Early Universe (RESCEU), The University of Tokyo, Tokyo 113-0033, Japan

(Received 16 October 1998; accepted for publication 30 November 1998)

It is shown that if an asymptotically flat space–time is asymptotically stationary, in the sense that $\mathcal{L}_\xi g_{ab}$ vanishes at the rate $\sim t^{-3}$ for asymptotically timelike vector field ξ^a, and the energy–momentum tensor vanishes at the rate $\sim t^{-4}$, then the space–time is an asymptotically Schwarzschild space–time. This gives a new aspect of the uniqueness theorem of a black hole. © *1999 American Institute of Physics.* [S0022-2488(99)00704-5]

I. INTRODUCTION

There are many astrophysical phenomena that are best explained by black holes, e.g., active galactic nuclei and x-ray binaries. The analyses on the phenomena are done assuming that those black holes are described by the Kerr space–times. This is because the uniqueness theorem of a black hole guarantees that a space–time which is stationary, vacuum, and asymptotically flat is uniquely the Kerr space–time, and we believe that when gravitational collapse takes place and a black hole is formed, the space–time around it becomes vacuum and accordingly stationary. (For details on the uniqueness theorem of a black hole, see Ref. 1.) However, one may argue that such a space–time does not become *exactly* vacuum or *exactly* stationary: the space–time becomes *asymptotically* vacuum, and accordingly *asymptotically* stationary at a certain rate of the time. In this context, a more adequate ''uniqueness theorem'' is the one that states an asymptotically stationary, vacuum and flat space–time is uniquely an asymptotically Schwarzschild space–time. This is what we show in this paper. (One may find it peculiar that an asymptotically *stationary* space–time is an asymptotically Schwarzschild space–time. However, if we define *asymptotically* Schwarzchild space–times as a class of space–times that asymptotically approach the exact Schwarzschild space–time, the space–times comprise a wide class of space–times, including the Kerr space–time, which is stationary. See the end of Sec. IV for details.)

To show the theorem, we use the notion of the asymptotic flat space–time, first introduced by Ashtekar and Romano[2] at spacelike infinity and succeedingly developed in our previous study[3] at timelike infinity. We investigate the asymptotic behavior of the gravitational field at the future timelike infinity, because we would like to know whether the gravitational field approaches asymptotically that of the Schwarzschild space–time at the *late time* when the space–time becomes asymptotically stationary. The standard definition of the asymptotic flatness[4] is based on the conformal completion method, which is used to obtain the Penrose diagram. This method is useful in that spacelike and timelike infinities can be simultaneously treated with null infinity, and thus that investigation concerning the global causal structure can be done. However, the method compresses the spacelike and timelike infinities, which possess rich 3-manifold information, down to points, and this compression results in a complicated differential structure at these points and makes it difficult to obtain the comprehensive picture of the behavior of gravitational fields in

[a)]JSPS Postdoctoral Fellowship for Research Abroad.

general asymptotically flat space–times. In contrast, the completion method introduced by Ref. 2 for definining the asymptotic flatness at spacelike infinity leaves the infinity as a 3-manifold. As a result, the complicated differential structure in the former treatment can be avoided. Subsequently, in our previous study,[3] we applied the method to timelike infinity and clarified that the method leads to a definite picture of hierarchy in the asymptotic behavior of the gravitational field and the symmetry, and that it is suitable to discuss such a notion as an ''asymptotically Schwarzschild space–time.'' (Perng also investigates hierarchy in the asymptotic gravitational field at spacelike infinity,[5] although it does not directly correspond to the hierarchy discussed in Ref. 3.) In this paper, we further investigate the hierarchy and prove the theorem stating that an asymptotically stationary, vacuum and flat space–time is uniquely an asymptotically Schwarzschild space–time.

The plan of this paper is the following. Section II is devoted to preliminaries. The definition of the asymptotic flatness at timelike infinity is recalled, and some of its useful consequences are summarized for the subsequent discussion. In Sec. III, the first-order asymptotic structure is explored, that is, the one order higher structure than the basic asymptotic structure that all the asymptotic flat space–times possess. Then, in Sec. IV, we introduce the notion of asymptotic stationarity and investigate how the condition that an asymptotic flat space–time be asymptotically stationary, constrain the asymptotic structure. The investigation leads to the proof of the main theorem.

Throughout the paper, we follow the notation of Wald.[6]

II. PRELIMINARIES

In this section, we recall the definition of asymptotic flatness at timelike infinity of Ref. 3 that will be used in the main proof and fix the notation.

Definition: A physical space–time $(\hat{\mathcal{M}},\hat{g}_{ab})$ is said to possess an *asymptote at future timelike infinity* $\breve{\imath}^+$ to *order* n(ATI-n) for a non-negative integer n, if there exists a manifold $\mathcal{M}$ with boundary $\mathcal{H}$, a smooth function Ω defined on $\mathcal{M}$, and an imbedding Ψ of an open subset $\hat{\mathcal{F}}$ in $\hat{\mathcal{M}}$ to $\mathcal{M}-\mathcal{H}$ satisfying the following conditions:

(1) $\breve{\imath}^+ := \partial\mathcal{F}\cap(\mathcal{M}-\Psi(\hat{\mathcal{M}}))$ is not empty and $\breve{\imath}^+\subset I^+(\mathcal{F})$ where $\mathcal{F}:=\Psi(\hat{\mathcal{F}})$;
(2) $\Omega \breve{=} 0$ and $\nabla_a\Omega \breve{\neq} 0$, where $\breve{=}$ denotes the equality evaluated on $\breve{\imath}^+$;
(3) $n^a := \Omega^{-4}\Psi^*\hat{g}^{ab}\nabla_b\Omega$ and $q_{ab}:=\Omega^2(\Psi^*\hat{g}_{ab}+\Omega^{-4}F^{-1}\nabla_a\Omega\nabla_b\Omega)$ admit smooth limits to $\breve{\imath}^+$ with q_{ab} having signature$(+++)$ on $\breve{\imath}^+$, where $F:=-\mathcal{L}_n\Omega$; and
(4) $\lim_{\to\breve{\imath}^+}\Omega^{-(2+n)}T_{\hat{\mu}\hat{\nu}}\breve{=}0$ where $\hat{T}_{\hat{\mu}\hat{\nu}}:=\Psi^*[(\hat{e}_\mu)^a(\hat{e}_\nu)^b\hat{T}_{ab}]$ in which $\{(\hat{e}_\mu)^a\}$ and $\hat{T}_{ab}$ are a tetrad and the physical energy–momentum tensor of $(\hat{\mathcal{M}},\hat{g}_{ab})$, respectively.

Henceforth, we use a tetrad consisting of a unit vector field $\hat{n}^a$ that is normal to the Ω-const surfaces and a triad $\{(\hat{e}_{\mathrm{I}})^a\}_{\mathrm{I}=1,2,3}$ of the metric $\hat{q}_{ab}:=\hat{g}_{ab}+\hat{n}_a\hat{n}_b$ on the Ω-const surface. We denote the timelike components with the subscript 0 and the spacelike component with capital-Roman-letter subscript, e.g., $\hat{A}^{\hat{0}}:=\hat{n}_a\hat{A}^a$ and $\hat{A}^{\hat{\mathrm{I}}}:=(\hat{e}_{\mathrm{I}})_a\hat{A}^a$. If a tensor $A^{a\cdots b}{}_{c\cdots d}$ admits a smooth limit to $\breve{\imath}^+$, it is useful to define the nth order term of $A^{a\cdots b}{}_{c\cdots d}$ as

$$^{(0)}A^{a\cdots b}{}_{c\cdots d}:=\lim_{\to\breve{\imath}^+}A^{a\cdots b}{}_{c\cdots d},$$

$$^{(n)}A^{a\cdots b}{}_{c\cdots d}:=\lim_{\to\breve{\imath}^+}\Omega^{-n}\left(A^{a\cdots b}{}_{c\cdots d}-\sum_{\ell=0}^{n-1}{}^{(\ell)}A^{a\cdots b}{}_{c\cdots d}\Omega^{\ell}\right)\quad\text{for } n\geqslant 1. \tag{1}$$

This definition of the nth order terms of a tensor implies that in the vicinity of $\breve{\imath}^+$, $A^{a\cdots b}{}_{c\cdots d}$ can be expanded as

$$A^{a\cdots b}{}_{c\cdots d}=\sum_{n=0}^{\infty}{}^{(n)}A^{a\cdots b}{}_{c\cdots d}\Omega^n. \tag{2}$$

Since all the equations appearing in the following discussion are those on $\mathcal{M}$, unless it may cause ambiguity, we omit hereafter Ψ^* in front of the tensors defined on $\hat{\mathcal{M}}$ for brevity.

Before we examine the properties of ATI-n space–times, we introduce some valuable tensors. First, the projection operator with respect to the Ω-const surface can be introduced as

$$q^a{}_b := \sum_{I=1,2,3} (\hat{e}_I)^a (\hat{e}_I)_b = \delta^a{}_b + F^{-1} n^a \nabla_b \Omega .$$

This operator admits a smooth limit to $\breve{\imath}^+$ by virtue of the definition of an ATI-n space–time. Second, note that the above definition of an ATI-n space–time implies that $\breve{\imath}^+$ is a 3-submanifold of $\mathcal{M}$ with an imbedding, say Π. Hence, if a tensor field $A^{a\cdots b}{}_{c\cdots d}$ is tangential to $\breve{\imath}^+$, or $A^{a\cdots b}{}_{c\cdots d} := q^a{}_e \cdots q^b{}_f q^g{}_c \cdots q^h{}_d A^{e\cdots f}{}_{g\cdots h}$, it is useful to consider the tensor field $\Pi^* A^{a\cdots b}{}_{c\cdots d}$ defined on $\breve{\imath}^+$. Hereafter, we denote such a tensor field in boldface, i.e., $\mathbf{A}^{a\cdots b}{}_{c\cdots d} := \Pi^* A^{a\cdots b}{}_{c\cdots d}$, and say that $A^{a\cdots b}{}_{c\cdots d}$ induces $\mathbf{A}^{a\cdots b}{}_{c\cdots d}$ on $\breve{\imath}^+$.

Now we explore the consequence of the definition of an ATI-n space–time for $n=0$. Solving the Einstein equation under the falloff condition on the energy–momentum tensor, $\lim_{\to \breve{\imath}^+} \Omega^{-2} \hat{T}_{\hat{\mu}\hat{\nu}} = 0$, it is found that

$$\mathbf{F} \stackrel{\smile}{=} 1, \quad \mathbf{q}_{ab} \stackrel{\smile}{=} \mathbf{h}_{ab} \tag{3}$$

in an ATI-0 space–time, where $h_{ab} = (d\chi)_a (d\chi)_b + \sinh^2 \chi [(d\theta)_a (d\theta)_b + \sin^2 \theta (d\phi)_a (d\phi)_b]$ is the 3-metric of the unit timelike 3-hyperboloid. (For the details of the derivation, see Ref. 3.) Because Eq. (3) is a gravitational structure common to all the ATI-0 space–times, we call it the *zeroth order asymptotic structure*. Using conformal time $\eta := \ln \Omega$, these results imply

$$\hat{g}_{ab} = {}^{(0)}\hat{g}_{ab} + \Omega\, {}^{(1)}\hat{g}_{ab} + O(\Omega^2)$$

where

$${}^{(0)}\hat{g}_{ab} = (e^{-\eta})^2 [-(d\eta)_a (d\eta)_b + h_{ab}] \tag{4}$$

in which ${}^{(n)}\hat{g}_{ab}$ is defined by

$${}^{(n)}\hat{g}_{ab} := \sum_{\mu,\nu} (\hat{e}_\mu)_a (\hat{e}_\nu)_b\, {}^{(n)}g_{\hat{\mu}\hat{\nu}} \tag{5}$$

with a function $g_{\hat{\mu}\hat{\nu}} := (\hat{e}_\mu)^a (\hat{e}_\nu)^b \hat{g}_{ab}$ that admits a smooth limit to $\breve{\imath}^+$ and thus is expanded in the manner described in Eq. (1). ${}^{(0)}\hat{g}_{ab}$ is a metric of the Milne universe and is equivalent to the metric of a Minkowski space–time, $\hat{g}^{\mathrm{Min}}_{ab}$. In other words, Eq. (4) tell us that an ATI-0 space–time is an asymptotically Minkowski space–time:

$$\hat{g}_{ab} = \hat{g}^{\mathrm{Min}}_{ab} + O(\Omega). \tag{6}$$

Hence, it is no surprise that the Riemann tensor asymptotically vanishes in such a space–time. The trace part of the Riemann tensor asymptotically vanishes by virtue of the falloff condition on the energy–momentum tensor. The traceless part, or the Weyl tensor $\hat{C}_{ambn}$, can be best investigated by decomposing the tensor into the electric part $\hat{E}_{ab} := \hat{C}_{ambn} \hat{n}^m \hat{n}^n$ and the magnetic part $\hat{B}_{ab} := {}^*\hat{C}_{ambn} \hat{n}^m \hat{n}^n$ where ${}^*\hat{C}_{ambn}$ denotes the dual of the 2-form $\hat{C}_{[am]bn}$. In terms of $\hat{q}_{ab}$ and $\hat{n}^a$, $\hat{E}_{ab}$ and $\hat{B}_{ab}$ are given by

$$\begin{aligned} \hat{E}_{ab} &= \hat{K}_{ar} \hat{K}^r{}_b - \mathcal{L}_{\hat{n}} \hat{K}_{ab} + \hat{D}_{(a} \hat{a}_{b)} + \hat{a}_a \hat{a}_b + \tfrac{1}{2} (\hat{q}^r_a \hat{q}^s_b - \hat{q}_{ab} \hat{n}^r \hat{n}^s) \hat{L}_{rs}, \\ \hat{B}_{ab} &= \hat{\epsilon}_{ra}{}^s \hat{D}^r \hat{K}_{bs} + \tfrac{1}{2} \hat{\epsilon}_{ab}{}^r \hat{n}^s \hat{L}_{rs}, \end{aligned} \tag{7}$$

where $\hat{K}_{ab} := \frac{1}{2}\mathcal{L}_{\hat{n}}\hat{q}_{ab}$, $\hat{L}_{ab} := \hat{R}_{ab} - \frac{1}{6}\hat{R}\hat{g}_{ab}$, $\hat{a}_a := \hat{q}_{ar}\hat{n}^s\nabla_s\hat{n}^r$, and $\hat{D}_a$ and $\hat{\epsilon}_{abc}$ are the derivative operator and the volume element associated with $\hat{q}_{ab}$, respectively. Using the definitions of n^a and q_{ab}, and imposing the falloff condition on the energy–momentum tensor, it can be found that both admit a smooth limit to $\breve{\imath}^+$ and thus can be expanded by the manner described in Eq. (1). Their leading terms are

$$ {}^{(0)}\mathbf{E}_{ab} \doteq 0, \quad {}^{(0)}\mathbf{B}_{ab} \doteq 0. \tag{8}$$

A simple calculation shows that the behaviors of ${}^{(n)}\mathbf{E}_{ab}$ and ${}^{(n)}\mathbf{B}_{ab}$ for $n \geqslant 1$ depend on that of the higher order energy–momentum tensor, which is arbitrary in an ATI-0 space–time. (See Ref. 3 for the derivations.)

III. THE FIRST-ORDER ASYMPTOTIC STRUCTURE

In this section, we derive the first order asymptotic structure, that is, a structure which is possessed by all the ATI-1 space–times but not by ATI-0 space–times. The part of the structure, i.e., the $O(\Omega)$ term of F, has been already derived in Ref. 3. Here, we treat the whole structure in a systematic way. The point is that the first-order asymptotic structure may be considered as the perturbation to the Milne universe, Eq. (4), where Ω plays the role of a small parameter of the perturbation. Hence, we can apply the technique of the cosmological perturbation[7,8] to derive the first-order asymptotic structure.

The energy–momentum tensor of an ATI-1 space–time satisfies a stronger falloff condition, $\lim_{\to \breve{\imath}^+}\Omega^{-3}\hat{T}_{\hat{\mu}\hat{\nu}} = 0$, than that of an ATI-0 and thus the behavior of asymptotic gravitational fields is constrained stronger. In other words, ATI-1 space–times possess more asymptotic gravitational structure in common. The structure can be derived by solving the Einstein equation under the condition $\lim_{\to \breve{\imath}^+}\Omega^{-3}\hat{T}_{\hat{\mu}\hat{\nu}} = 0$. To obtain the equation, we first decompose the first-order metric ${}^{(1)}\hat{g}_{ab}$ as

$$ {}^{(1)}\hat{g}_{ab} = (e^{-\eta})^2[{}^{(1)}F(d\eta)_a(d\eta)_b - 2\,{}^{(1)}\beta_{(a}(d\eta)_{b)} - 2\,{}^{(1)}\psi h_{ab} + 2\,{}^{(1)}\chi_{ab}], \tag{9}$$

where ${}^{(1)}\beta_a$ and ${}^{(1)}\chi_{ab}$ are tangential to the Ω-const surfaces, i.e., $q_a{}^{b}\,{}^{(1)}\beta_b = {}^{(1)}\beta_a$ and $q_a{}^c q_b{}^d\,{}^{(1)}\chi_{cd} = {}^{(1)}\chi_{ab}$; and ${}^{(1)}\chi_{ab}$ is traceless, i.e., $q^{ab}\,{}^{(1)}\chi_{ab} = 0$. With these quantities, the Einstein equation induces on $\breve{\imath}^+$ the following set of differential equations in an ATI-1 space–time satisfying $\lim_{\to \breve{\imath}^+}\Omega^{-3}\hat{T}_{\hat{\mu}\hat{\nu}} = 0$:

$$\begin{aligned} &3\,{}^{(1)}\mathbf{F} + 2\Delta\,{}^{(1)}\boldsymbol{\psi} - 2\mathbf{D}_m\,{}^{(1)}\boldsymbol{\beta}^m + \mathbf{D}^m\mathbf{D}_n\,{}^{(1)}\boldsymbol{\chi}_m{}^n \doteq 0,\\ &\mathbf{D}_a({}^{(1)}\mathbf{F} + 2\,{}^{(1)}\boldsymbol{\psi}) + \tfrac{1}{2}(\Delta - 2){}^{(1)}\boldsymbol{\beta}_a - \tfrac{1}{2}\mathbf{D}_a(\mathbf{D}_m\,{}^{(1)}\boldsymbol{\beta}^m) + \mathbf{D}_m\,{}^{(1)}\boldsymbol{\chi}_a{}^m \doteq 0,\\ &(\mathbf{h}_{ab}\Delta - \mathbf{D}_a\mathbf{D}_b)({}^{(1)}\mathbf{F} + 2\,{}^{(1)}\boldsymbol{\psi}) + 2\mathbf{D}_{(a}\,{}^{(1)}\boldsymbol{\beta}_{b)} + 2(\Delta + 3){}^{(1)}\boldsymbol{\chi}_{ab}\\ &\quad - 2\mathbf{h}_{ab}\mathbf{D}_m\,{}^{(1)}\boldsymbol{\beta}^m - 4\mathbf{D}_{(a}\mathbf{D}^m\,{}^{(1)}\boldsymbol{\chi}_{b)m} + 2\mathbf{h}_{ab}\mathbf{D}^m\mathbf{D}_n\,{}^{(1)}\boldsymbol{\chi}_m{}^n \doteq 0, \end{aligned} \tag{10}$$

where $\mathbf{D}_a$ is the derivative operator associated with the metric $\mathbf{h}_{ab}$ of $\breve{\imath}^+$, and $\Delta := \mathbf{D}^a\mathbf{D}_a$. To simplify the equation above, we consider the Poisson gauge in which ${}^{(1)}\boldsymbol{\beta}_a$ is transverse and ${}^{(1)}\boldsymbol{\chi}_{ab}$ is transverse traceless. Noting that the quantities are transformed as

$$\begin{aligned} &{}^{(1)}\overline{\mathbf{F}} \doteq {}^{(1)}\mathbf{F}, \quad {}^{(1)}\overline{\boldsymbol{\beta}}_a \doteq {}^{(1)}\boldsymbol{\beta}_a + \mathbf{D}_a\mathbf{T} - \mathbf{L}_a,\\ &{}^{(1)}\overline{\boldsymbol{\psi}} \doteq {}^{(1)}\boldsymbol{\psi} + \mathbf{T} - \tfrac{1}{3}\mathbf{D}^a\mathbf{L}_a, \quad {}^{(1)}\overline{\boldsymbol{\chi}}_{ab} \doteq {}^{(1)}\boldsymbol{\chi}_{ab} + \mathbf{D}_{(a}\mathbf{L}_{b)} - \tfrac{1}{3}\mathbf{h}_{ab}\mathbf{D}_m\mathbf{L}^m \end{aligned} \tag{11}$$

under a gauge transformation generated by $\xi^a = \Omega T(\partial_\eta)^a + \Omega L^a$, we find that the Poisson gauge can be always chosen if we set

$$\mathbf{T} :\doteq -\tfrac{1}{2}\Delta^{-1}[3(\Delta-3)^{-1}\mathbf{D}^m\mathbf{D}^{n\,(1)}\boldsymbol{\chi}_{mn}+2\mathbf{D}^{a\,(1)}\boldsymbol{\beta}_a],$$
$$L_a :\doteq \tfrac{1}{2}(\Delta-2)^{-1}\{\mathbf{D}_a[(\Delta-3)^{-1}\mathbf{D}^m\mathbf{D}^{n\,(1)}\boldsymbol{\chi}_{mn}]-4\mathbf{D}^{m\,(1)}\boldsymbol{\chi}_{ma}\}, \tag{12}$$

for the generator $\xi^a=\Omega T(\partial_\eta)^a+\Omega L^a$, which satisfy $\mathbf{D}^{a\,(1)}\boldsymbol{\beta}_{ab}+\mathbf{D}^a\mathbf{D}_{(a}\mathbf{L}_{b)}-\frac{1}{3}\mathbf{D}_b(\mathbf{D}_m\mathbf{L}^m)\doteq 0$ and $\mathbf{D}^{a\,(1)}\boldsymbol{\beta}_a+\Delta\mathbf{T}-\mathbf{D}^a\mathbf{L}_a\doteq 0$. Note that this Poisson gauge is preserved under the transformation generated by $\xi^a=\Omega T(\partial_\eta)^a+\Omega L^a$ where T and L_a satisfies

$$\Delta\mathbf{T}-\mathbf{D}^m\mathbf{L}_m\doteq 0,\quad (\Delta-2)\mathbf{L}_b+\tfrac{1}{3}\mathbf{D}_b\mathbf{D}_m\mathbf{L}^m\doteq 0. \tag{13}$$

In this gauge, the Einstein equation (10) simplifies to

$$3\,{}^{(1)}\mathbf{F}+2\Delta\,{}^{(1)}\boldsymbol{\psi}\doteq 0, \tag{14a}$$

$$\mathbf{D}_a({}^{(1)}\mathbf{F}+2\,{}^{(1)}\boldsymbol{\psi})+\tfrac{1}{2}(\Delta-2)^{(1)}\boldsymbol{\beta}_a^T\doteq 0, \tag{14b}$$

$$(\mathbf{h}_{ab}\Delta-\mathbf{D}_a\mathbf{D}_b)({}^{(1)}\mathbf{F}+2\,{}^{(1)}\boldsymbol{\psi})+2\mathbf{D}_{(a}\,{}^{(1)}\boldsymbol{\beta}_{b)}^T+2(\Delta+3)^{(1)}\boldsymbol{\chi}_{ab}^{\mathrm{TT}}\doteq 0, \tag{14c}$$

where the overbar is omitted which shows that the quantity is transformed and the subscripts T and TT denote that the quantities are transverse and transverse traceless, respectively.

First, we derive the scalar ${}^{(1)}\mathbf{F}$. Subtracting Eq. (14a) from the divergence of Eq. (14b), we obtain

$$(\Delta-3)^{(1)}\mathbf{F}\doteq 0. \tag{15}$$

The general solution of the above equation can be derived by first expanding the function ${}^{(1)}\mathbf{F}$ as

$${}^{(1)}\mathbf{F}(\boldsymbol{\chi},\boldsymbol{\theta},\boldsymbol{\phi})\doteq\sum\ \mathbf{a}_{\ell m}\,{}^{(1)}\mathbf{F}^{\ell m}(\boldsymbol{\chi},\boldsymbol{\theta},\boldsymbol{\phi})$$

where

$${}^{(1)}\mathbf{F}^{\ell m}(\boldsymbol{\chi},\boldsymbol{\theta},\boldsymbol{\phi}):=\mathbf{T}_0^{\ell}(\boldsymbol{\chi})\mathbf{Y}^{\ell m}(\boldsymbol{\theta},\boldsymbol{\phi}) \tag{16}$$

in which the summation is taken over $\ell\in\mathbb{Z}$ and $|m|\leqslant|\ell|$, and $\mathbf{Y}^{\ell m}(\theta,\phi)$ are the two-dimensional spherical harmonics. Substituting Eq. (16) into Eq. (15), it is found that $\mathbf{T}_0^{\ell}$ is given by

$$\mathbf{T}_0^{\ell}(\chi)=\boldsymbol{\mathcal{P}}_2^{\ell}(\chi), \tag{17}$$

where the functions $\boldsymbol{\mathcal{P}}_n^{\ell}(\chi)$ are defined by

$$\boldsymbol{\mathcal{P}}_n^{\ell}(\chi)\doteq\frac{1}{\sqrt{\sinh\boldsymbol{\chi}}}\mathbf{P}_{n-(1/2)}^{\ell+(1/2)}(\cosh\boldsymbol{\chi}), \tag{18}$$

and satisfies

$$\boldsymbol{\mathcal{P}}_n^{\ell\,\prime\prime}(\boldsymbol{\chi})+\frac{2}{\tanh\boldsymbol{\chi}}\boldsymbol{\mathcal{P}}_n^{\ell\,\prime}(\boldsymbol{\chi})-\left(n^2-1+\frac{\ell(\ell+1)}{\sinh^2\boldsymbol{\chi}}\right)\boldsymbol{\mathcal{P}}_n^{\ell}(\boldsymbol{\chi})\doteq 0, \tag{19}$$

where the prime ($'$) denotes the derivative with respect to $\boldsymbol{\chi}$; and $P_\nu^\mu(z)$ is an associated Legendre function normalized as

$$P_\nu^\mu(z):=\frac{1}{\Gamma(1-\mu)}\left(\frac{z+1}{z-1}\right)^{\mu/2}F\left(-\nu,\nu+1,1-\mu;\frac{1-z}{2}\right)\quad\text{for } z>1. \tag{20}$$

[Note here that the unconventional choice of ℓ is taken. It is related to the conventional choice ℓ_c by $\ell = -\ell_c - 1$, and thus it is $\ell \leq -1$ terms that are regular. The reason for the choice is to have the coefficient a_{00} in Eq. (16) correspond to the mass of a space–time. See Eq. (51).]

Second, we derive the scalar ${}^{(1)}\psi$. Adding Eq. (15) to Eq. (14a), we obtain

$$\Delta({}^{(1)}\mathbf{F} + 2\,{}^{(1)}\boldsymbol{\psi}) \doteq 0. \tag{21}$$

Let $\mathbf{A}$ be the solution of $\Delta \mathbf{A} \doteq 0$. Then, ${}^{(1)}\psi$ may be given by

$${}^{(1)}\boldsymbol{\psi} \doteq -\frac{1}{2}\,{}^{(1)}\mathbf{F} + \frac{\mathbf{A}}{2}. \tag{22}$$

Now, perform the gauge transformation generated by $\xi^a = \Omega T(\partial_\eta)^a$ where $\mathbf{T} \doteq -\mathbf{A}/2$. This transformation preserves the Poisson gauge, which can be seen from Eq. (13), and simplifies ${}^{(1)}\psi$:

$${}^{(1)}\boldsymbol{\psi} \doteq -\tfrac{1}{2}\,{}^{(1)}\mathbf{F}. \tag{23}$$

Next, we consider the vector field ${}^{(1)}\boldsymbol{\beta}^T_a$. Substitution of Eq. (23) into Eq. (14b) yields

$$(\Delta - 2)\,{}^{(1)}\boldsymbol{\beta}^T_a \doteq 0. \tag{24}$$

Hence, together with the fact ${}^{(1)}\boldsymbol{\beta}^T_a$ is transverse, we see from Eq. (13) that the gauge transformation generated by $\xi^a = \Omega L^a$ where $\mathbf{L}_a \doteq {}^{(1)}\boldsymbol{\beta}^T_a$ preserves the Poisson gauge, and results in

$${}^{(1)}\boldsymbol{\beta}^T_a \doteq 0. \tag{25}$$

Finally, we derive the tensor field ${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}}_{ab}$. Substituting Eqs. (23)–(25) into Eq. (14c), we obtain

$$(\Delta + 3)\,{}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}}_{ab} \doteq 0. \tag{26}$$

To solve this equation, we consider the spherical harmonic expansion again. As generally done, we decompose ${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}}_{ab}$ into the even (or electric-type) parity part ${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(+)}_{ab}$ and the odd (or magnetic-type) parity part ${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(-)}_{ab}$: ${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}}_{ab} \doteq {}^{(1)}\check{\boldsymbol{\chi}}^{\mathrm{TT}(+)}_{ab} + {}^{(1)}\check{\boldsymbol{\chi}}^{\mathrm{TT}(-)}_{ab}$. In the present case, the even parity part that satisfies Eq. (26) is found to be[9]

$${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(+)}_{ab} \doteq (\mathbf{D}_a \mathbf{D}_b - \mathbf{h}_{ab})\mathbf{X}, \tag{27}$$

where $\mathbf{X}$ is a function that satisfies $(\Delta - 3)\mathbf{X} \doteq 0$. On the other hand, the odd parity part is found to be[10]

$${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(-)}_{ab} \doteq \sum_{\ell \neq 0} \mathbf{b}^{(-)}_{\ell m}\,{}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(-)\ell m}_{ab}, \tag{28}$$

where the summation is taken over $\ell \in \mathbb{Z}$ for $\ell \neq 0$, $|m| \leq |\ell|$; and

$${}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(-)\ell m}_{\chi\chi} :\doteq 0, \quad {}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(-)\ell m}_{\chi A} :\doteq \mathbf{T}^\ell_1(\boldsymbol{\chi})\boldsymbol{\epsilon}_A{}^B \boldsymbol{\mathcal{D}}_B \mathbf{Y}^{\ell m}, \quad {}^{(1)}\boldsymbol{\chi}^{\mathrm{TT}(-)\ell m}_{AB} :\doteq \mathbf{T}^\ell_2(\boldsymbol{\chi})\boldsymbol{\epsilon}^C{}_{(A}\boldsymbol{\mathcal{D}}_{B)}\boldsymbol{\mathcal{D}}_C \mathbf{Y}^{\ell m} \tag{29}$$

in which functions $\mathbf{T}^\ell_1(\boldsymbol{\chi})$ and $\mathbf{T}^\ell_2(\boldsymbol{\chi})$ are given by

$$\mathbf{T}^\ell_1(\boldsymbol{\chi}) \doteq \boldsymbol{\mathcal{P}}^\ell_0(\boldsymbol{\chi}), \quad \mathbf{T}^\ell_2(\boldsymbol{\chi}) \doteq \frac{\sinh^2 \boldsymbol{\chi}}{(\ell - 1)(\ell + 2)}\left[\partial_\chi + 2\,\frac{\cosh \boldsymbol{\chi}}{\sinh \boldsymbol{\chi}}\right]\boldsymbol{\mathcal{P}}^\ell_0(\boldsymbol{\chi}) \quad \text{for } \ell \neq 1, -2. \tag{30}$$

(Here, there is no need to derive $\mathbf{T}_2^{\ell}(\boldsymbol{\chi})$ for $\ell=1$ or $\ell=-2$ since $\boldsymbol{\epsilon}^{c}{}_{(A}\boldsymbol{\mathcal{D}}_{B)}\boldsymbol{\mathcal{D}}_C\mathbf{Y}^{\ell m}$ vanishes for these values of ℓ.) Next, consider the gauge transformation generated by $\xi^a=\Omega T(\partial_\eta)^a+\Omega L^a$ where $\mathbf{T}:\hat{=}-\mathbf{X}$ and $\mathbf{L}_a:\hat{=}-\mathbf{D}_a\mathbf{X}$. This transformation leaves ${}^{(1)}\boldsymbol{\beta}_a$ and ${}^{(1)}\boldsymbol{\psi}$ unchanged and kills the even parity part of ${}^{(1)}\boldsymbol{\chi}_{ab}^{\mathrm{TT}}$:

$$ {}^{(1)}\boldsymbol{\chi}_{ab}^{\mathrm{TT}}\hat{=}\sum_{\ell\neq 0}\mathbf{b}_{\ell m}^{(-)}\,{}^{(1)}\boldsymbol{\chi}_{ab}^{\mathrm{TT}(-)\ell m}. \tag{31} $$

To summarize, the solutions of the Einstein equation (10) is given by

$$ \begin{aligned} {}^{(1)}\overline{\mathbf{F}}&\hat{=}\sum\,\mathbf{a}_{\ell m}\,{}^{(1)}\mathbf{F}^{\ell m}, \quad {}^{(1)}\overline{\boldsymbol{\beta}}_a\hat{=}0,\\ {}^{(1)}\overline{\boldsymbol{\psi}}&\hat{=}-\frac{1}{2}\sum\,\mathbf{a}_{\ell m}\,{}^{(1)}\mathbf{F}^{\ell m}, \quad {}^{(1)}\overline{\boldsymbol{\chi}}_{ab}\hat{=}\sum_{\ell\neq 0}\mathbf{b}_{\ell m}^{(-)}\,{}^{(1)}\boldsymbol{\chi}_{ab}^{\mathrm{TT}(-)\ell m} \end{aligned} \tag{32} $$

in the suitably chosen Poisson gauge. This is the structure of the gravitational field common to all the ATI-1 space–times, and thus we define the *first-order asymptotic structure* as follows.

Definition: ${}^{(1)}\hat{g}_{ab}$ given by Eq. (9) is called the *first-order asymptotic structure* of an AFTI-1 space–time, where ${}^{(1)}\mathbf{F}$, ${}^{(1)}\boldsymbol{\beta}_a$, ${}^{(1)}\boldsymbol{\psi}$, and ${}^{(1)}\boldsymbol{\chi}_{ab}$ takes the form Eq. (32) on $\breve{\imath}^+$, in the Poisson gauge.

In such a space–time, it can be calculated that

$$ {}^{(1)}\mathbf{E}_{ab}\hat{=}\tfrac{1}{2}(\mathbf{D}_a\mathbf{D}_b-\mathbf{h}_{ab})^{(1)}\mathbf{F}, \quad {}^{(1)}\mathbf{B}_{ab}\hat{=}\tfrac{1}{2}\boldsymbol{\epsilon}_{ra}{}^{s}\mathbf{D}^{r(1)}\boldsymbol{\chi}_{bs} \tag{33} $$

and that ${}^{(n)}\mathbf{E}_{ab}$ and ${}^{(n)}\mathbf{B}_{ab}$ for $n\geqslant 2$ depend on the behavior of the higher order energy–momentum tensor, which is arbitrary in an ATI-1 space–time. (See Ref. 3 for the details of the calculation.) Equation (33) clearly shows that ${}^{(1)}\mathbf{F}$ of the first-order asymptotic structure forms the first-order term of the electric part of the Weyl tensor and that ${}^{(1)}\boldsymbol{\chi}_{ab}$ forms the magnetic part. In other words, if we specify the sets of coefficients $\{\mathbf{a}_{\ell m}\}$ and $\{\mathbf{b}_{\ell m}\}$, the first-order terms of the electric part and the magnetic part are determined, respectively.

IV. ASYMPTOTIC STATIONARITY

In this section, we introduce the notion of *asymptotic stationarity* of ATI-n space–times, and prove that an ATI-1 space–time that is asymptotically stationary to order 2 must be an asymptotically Schwarzschild space–time. The plan of the proof is: (1) we derive, in the Poisson gauge, the reduced first asymptotic structure of an ATI-1 space–time that is asymptotically stationary to order 2 in the lemma; (2) we perform a suitable gauge transformation as to show explicitly that such an asymptotic structure approaches asymptotically the Schwarzschild metric in the theorem.

A Killing vector field $\hat{\xi}^a$ is a vector field with respect to which the Lie derivative of the metric vanishes, $\mathcal{L}_{\hat{\xi}}\hat{g}_{ab}=0$. This fact motivates us to define an asymptotic Killing vector field and its order as follows.

Definition: An ATI-m spacetime $(\hat{\mathcal{M}},\hat{g}_{ab})$ is said to admit an *asymptotic Killing field* $\hat{\xi}^a$ to order n if

$$ \lim_{\to\breve{\imath}^+}\Omega^{-n}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}=0, \tag{34} $$

where $(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}:=\Psi^*((\hat{e}_\mu)^a(\hat{e}_\nu)^b\mathcal{L}_{\hat{\xi}}\hat{g}_{ab})$.

Next, we consider how to define asymptotic stationarity, using this notion of an asymptotic Killing vector field. A vector field $\hat{\xi}^a$ is said to be a stationary Killing vector field in an asymptotically flat space–time, if $\hat{\xi}^a$ is a timelike Killing field and satisfies $\hat{g}_{ab}\hat{\xi}^a\hat{\xi}^b=-1$ at infinity. Hence, we define its asymptotic correspondence as follows.

Definition: A vector field ξ^a is said to be an asymptotic stationary Killing vector field to order n of an ATI-m space–time if $\hat{\xi}^a$ is admitted as an asymptotic Killing vector field to order n in the ATI-m space–time and satisfies

$$\hat{g}_{ab}\hat{\xi}^a\hat{\xi}^b \doteq -1 \tag{35}$$

on $\breve{\imath}^+$.

It is important to note that the definition of the asymptotic stationary Killing vector field implies that the leading term of $\hat{\xi}^{\hat{\mu}}=\hat{\xi}^a(\hat{e}_\mu)_a$ is of the order Ω^0. Hence, $\xi^{\hat{\mu}}:=\hat{\xi}^{\hat{\mu}}$ admits a smooth limit to $\breve{\imath}^+$ and can be expanded as

$$\xi^{\hat{\mu}}=\sum_{n=1}^{\infty}{}^{(n)}\xi^{\hat{\mu}}\Omega^n. \tag{36}$$

(Here, ${}^{(n)}\xi^{\hat{0}}$ and ${}^{(n)}\xi^{\hat{\mathrm{I}}}$ correspond to ${}^{(n+2)}\xi^{\hat{0}}$ and ${}^{(n+1)}\xi^{\hat{\mathrm{I}}}$ in Ref. 3. We do not use the notation of Ref. 3 for $\hat{\xi}^a$, because the above notation reflects the nature of the completion more intrinsically.) Simple calculation shows that the function $(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}$ admits a smooth limit to $\breve{\imath}^+$, and thus can be expanded in the manner described by Eq. (1).

Now we are ready to prove the following lemma.

Lemma: An ATI-1 space–time is asymptotically stationary to order 2 if and only if $\mathbf{a}_{\ell m}\doteq 0$ for $\ell\neq 0$ and $\mathbf{b}_{\ell m}\doteq 0$.

Proof of only if: If an ATI-1 space–time is asymptotically stationary to order 2,

$${}^{(n)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}\doteq 0 \quad \text{for } n\leqslant 2, \tag{37}$$

$$\hat{g}_{ab}\hat{\xi}^a\hat{\xi}^b\doteq -1 \tag{38}$$

hold. First, we note that Eq. (38) is equivalent to

$$-({}^{(0)}\boldsymbol{\xi}^{\hat{0}})^2+{}^{(0)}\boldsymbol{\xi}^{a\,(0)}\boldsymbol{\xi}_a\doteq -1, \tag{39}$$

where ${}^{(\ell)}\boldsymbol{\xi}^a:={}^{(\ell)}\boldsymbol{\xi}^{\hat{K}}(\mathbf{e}_K)^a$ and $\{(\mathbf{e}_{\mathrm{I}})_a\}_{\mathrm{I}=1,2,3}$ is a triad of $\mathbf{h}_{ab}$. Second, we simplify the $n\leqslant 1$ part of Eq. (37). In an ATI-1 space–time, the $O(\Omega^0)$ and $O(\Omega^1)$ terms of $(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}$ are given by

$${}^{(0)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{0}\hat{0}}\doteq 0,\quad {}^{(0)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{0}\hat{\mathrm{I}}}\doteq 0,\quad {}^{(0)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mathrm{I}}\hat{\mathrm{J}}}\doteq 0,$$

$${}^{(1)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{0}\hat{0}}\doteq 0,\quad {}^{(1)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{0}\hat{\mathrm{I}}}\doteq[{}^{(0)}\boldsymbol{\xi}_a-\mathbf{D}_a\,{}^{(0)}\boldsymbol{\xi}^{\hat{0}}](\mathbf{e}_{\mathrm{I}})^a,$$

$${}^{(1)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mathrm{I}}\hat{\mathrm{J}}}\doteq[\mathbf{D}_{(a}\,{}^{(0)}\boldsymbol{\xi}_{b)}-{}^{(0)}\boldsymbol{\xi}^{\hat{0}}\mathbf{h}_{ab}](\mathbf{e}_{\mathrm{I}})^a(\mathbf{e}_{\mathrm{J}})^b.$$

Hence, the $n\leqslant 1$ part of Eq. (37) is equivalent to

$${}^{(0)}\boldsymbol{\xi}_a-\mathbf{D}_a\,{}^{(0)}\boldsymbol{\xi}^{\hat{0}}\doteq 0,\quad \mathbf{D}_{(a}\,{}^{(0)}\boldsymbol{\xi}_{b)}-{}^{(0)}\boldsymbol{\xi}^{\hat{0}}\mathbf{h}_{ab}\doteq 0. \tag{40}$$

Solving Eqs. (39) and (40) simultaneously, we find that

$${}^{(0)}\boldsymbol{\xi}^{\hat{0}}\doteq\cosh\boldsymbol{\chi},\quad {}^{(0)}\boldsymbol{\xi}^{\hat{\mathrm{K}}}(\mathbf{e}_{\mathrm{K}})^a\doteq\sinh\boldsymbol{\chi}(\partial_\chi)^a. \tag{41}$$

Next, we consider the $n=2$ part of Eq. (37). Let us choose the Poisson gauge in which the first-order asymptotic structure takes the form Eq. (32) in an ATI-1 space–time. Because

$${}^{(2)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{0}\hat{0}}\doteq -{}^{(1)}\boldsymbol{\xi}^{\hat{0}}+\tfrac{1}{2}({}^{(0)}\boldsymbol{\xi}^{\hat{0}}+{}^{(0)}\boldsymbol{\xi}^m\mathbf{D}_m){}^{(1)}\mathbf{F}, \tag{42}$$

$$^{(2)}(\mathcal{L}_{\hat\xi}\hat g)_{\hat 0\hat{\rm I}} \doteq [2\,^{(1)}\boldsymbol{\xi}_a - \mathbf{D}_a\,^{(1)}\boldsymbol{\xi}^{\hat 0} + (\mathbf{D}_a\,^{(0)}\boldsymbol{\xi}^{\hat 0} - {}^{(0)}\boldsymbol{\xi}_a)\,^{(1)}\mathbf{F} - 2\,^{(0)}\boldsymbol{\xi}_m\,^{(1)}\boldsymbol{\chi}_a{}^m](\mathbf{e}_{\rm I})^a,$$

the $n=2$ part of Eq. (37) for $\hat\mu=\hat 0$ is equivalent to

$$^{(1)}\boldsymbol{\xi}^{\hat 0} \doteq \tfrac{1}{2}(^{(0)}\boldsymbol{\xi}^{\hat 0} + {}^{(0)}\boldsymbol{\xi}^m\mathbf{D}_m)\,^{(1)}\mathbf{F}, \quad ^{(1)}\boldsymbol{\xi}^{\hat{\rm K}}(e_{\rm K})^a \doteq \tfrac{1}{2}[\mathbf{D}_a\,^{(1)}\boldsymbol{\xi}^{\hat 0} - (\mathbf{D}_a\,^{(0)}\boldsymbol{\xi}^{\hat 0} - {}^{(0)}\boldsymbol{\xi}_a)\,^{(1)}\mathbf{F} + 2\,^{(0)}\boldsymbol{\xi}_m\,^{(1)}\boldsymbol{\chi}_a{}^m]. \tag{43}$$

Using Eqs. (41)–(43), we find that $^{(2)}(\mathcal{L}_{\hat\xi}\hat g)_{\hat{\rm I}\hat{\rm J}}$ are given by

$$^{(2)}(\mathcal{L}_{\hat\xi}\hat g)_{\hat{\rm I}\hat{\rm J}} \doteq 2\sinh\boldsymbol{\chi}[\,^1\boldsymbol{\mathcal{L}}_{ab}\,^{(1)}\mathbf{F} + {}^2\boldsymbol{\mathcal{L}}_{(a}{}^{m}\,^{(1)}\boldsymbol{\chi}_{b)m}](\mathbf{e}_{\rm I})^a(\mathbf{e}_{\rm J})^b \tag{44}$$

where the derivative operators $^1\boldsymbol{\mathcal{L}}_{ab}$ and $^2\boldsymbol{\mathcal{L}}_{ab}$ are given by

$$^1\boldsymbol{\mathcal{L}}_{ab} :\doteq \frac{3}{\tanh\boldsymbol{\chi}}(\mathbf{D}_a\mathbf{D}_b - \mathbf{h}_{ab}) + (\partial_\chi)^m\mathbf{D}_{(a}\mathbf{D}_{b)}\mathbf{D}_m - (d\boldsymbol{\chi})_{(a}\mathbf{D}_{b)}, \quad ^2\boldsymbol{\mathcal{L}}_{ab} :\doteq 2(d\boldsymbol{\chi})_p\mathbf{h}_a{}^{[p}\mathbf{h}_b{}^{m]}\mathbf{D}_m, \tag{45}$$

respectively. By the definition of $^{(1)}\mathbf{F}$ and $^{(1)}\boldsymbol{\chi}_{ab}$, $^{(2)}(\mathcal{L}_{\hat\xi}\hat g)_{\hat{\rm I}\hat{\rm J}}$ vanishes if and only if $^1\boldsymbol{\mathcal{L}}_{ab}\,^{(1)}\mathbf{F}$ and $^2\boldsymbol{\mathcal{L}}_{(a}{}^{s}\,^{(1)}\boldsymbol{\chi}_{b)s}$ vanish independently. With the help of eqs. (16)–(32), $^1\boldsymbol{\mathcal{L}}_{ab}\,^{(1)}\mathbf{F}\doteq 0$ can be rewritten as

$$\sum a_{\ell m}\Bigg[\left(\mathbf{T}_0^{\ell\,'''} + \frac{3\mathbf{T}_0^{\ell\,''}}{\tanh\boldsymbol{\chi}} - \mathbf{T}_0^{\ell\,'} - \frac{3\mathbf{T}_0^{\ell}}{\tanh\boldsymbol{\chi}}\right)(d\boldsymbol{\chi})_a(d\boldsymbol{\chi})_b + 2\left(\mathbf{T}_0^{\ell\,''} - \frac{\mathbf{T}_0^{\ell\,'}}{\tanh\boldsymbol{\chi}} - \left(5 + \frac{4}{\sinh^2\boldsymbol{\chi}}\right)\mathbf{T}_0^{\ell}\right)$$
$$\times(d\boldsymbol{\chi})_{(a}\boldsymbol{\mathcal{D}}_{b)} + \sinh\boldsymbol{\chi}\cosh\boldsymbol{\chi}\left(\mathbf{T}_0^{\ell\,''} + \frac{2\mathbf{T}_0^{\ell}}{\tanh\boldsymbol{\chi}} - 3\mathbf{T}_0^{\ell} - \frac{\ell(\ell+1)}{2\sinh\boldsymbol{\chi}\cosh\boldsymbol{\chi}}\left(\mathbf{T}_0^{\ell\,'} + \frac{\mathbf{T}_0^{\ell}}{\tanh\boldsymbol{\chi}}\right)\right)(d\boldsymbol{\sigma})_{ab}$$
$$+\left(\mathbf{T}_0^{\ell\,'} + \frac{\mathbf{T}_0^{\ell}}{\tanh\boldsymbol{\chi}}\right)(\boldsymbol{\mathcal{D}}_a\boldsymbol{\mathcal{D}}_b - \tfrac{1}{2}(d\boldsymbol{\sigma})_{ab}\boldsymbol{\mathcal{D}}^c\boldsymbol{\mathcal{D}}_c)\Bigg]\mathbf{Y}^{\ell m} \doteq 0. \tag{46}$$

We first consider the $\ell\neq 0$ terms of the above equation. Noting that all the components are independent and using Eqs. (17)–(19), it is found that if $\mathbf{a}_{\ell m}\not\doteq 0$ the above equation (46) is equivalent to

$$\frac{\ell(\ell+1)}{\sinh^2\boldsymbol{\chi}}\left(\boldsymbol{\mathcal{P}}_2^{\ell\,'} + \frac{\boldsymbol{\mathcal{P}}_2^{\ell}}{\tanh\boldsymbol{\chi}}\right) \doteq 0, \quad \left(\frac{\ell^2+\ell-4}{\sinh^2\boldsymbol{\chi}} - 2\right)\boldsymbol{\mathcal{P}}_2^{\ell} - 3\boldsymbol{\mathcal{P}}_2^{\ell\,'} \doteq 0, \quad \frac{\ell(\ell+1)}{\sinh^2\boldsymbol{\chi}}\boldsymbol{\mathcal{P}}_2^{\ell} \doteq 0,$$
$$\boldsymbol{\mathcal{P}}_2^{\ell\,'} + \frac{\boldsymbol{\mathcal{P}}_2^{\ell}}{\tanh\boldsymbol{\chi}} \doteq 0. \tag{47}$$

Apparently, there is no integer ℓ that is not equal to 0 and that satisfies Eq. (47), simultaneously. Therefore, $\mathbf{a}_{\ell m}\doteq 0$ for $\ell\neq 0$. Noting that $\boldsymbol{\mathcal{D}}_a\mathbf{Y}^{\ell m}$ vanishes for $\ell=0$, we find that all the components of the $\ell=0$ terms of the right-hand side of Eq. (46) vanish. Therefore, $\mathbf{a}_{00}$ can take arbitrary value. Next consider $^2\boldsymbol{\mathcal{L}}_{(a}{}^{s}\,^{(1)}\boldsymbol{\chi}_{b)s}\doteq 0$. This equation is satisfied if and only if

$$\sum_{\ell\neq 0}\mathbf{b}_{\ell m}\left[\left(\mathbf{T}_1^{\ell\,'} + \frac{\mathbf{T}_1^{\ell}}{\tanh\boldsymbol{\chi}}\right)(d\boldsymbol{\chi})_{(a}\boldsymbol{\epsilon}_{b)}{}^{r}\boldsymbol{\mathcal{D}}_r + \left(\mathbf{T}_2^{\ell\,'} - \frac{\mathbf{T}_2^{\ell}}{\tanh\boldsymbol{\chi}} - \mathbf{T}_1^{\ell}\right)\boldsymbol{\epsilon}^{r}{}_{(a}\boldsymbol{\mathcal{D}}_{b)}\boldsymbol{\mathcal{D}}_r\right]\mathbf{Y}^{\ell m} \doteq 0. \tag{48}$$

With the same reasoning, we find that if $\mathbf{b}_{\ell m}\not\doteq 0$ Eq. (48) is equivalent to

$$\frac{\ell(\ell+1)}{\sinh^2\boldsymbol{\chi}}\boldsymbol{\mathcal{P}}_0^{\ell} \doteq 0, \quad \frac{(\ell+2)(\ell-1)}{\sinh^2\boldsymbol{\chi}}\boldsymbol{\mathcal{P}}_0^{\ell} \doteq 0 \tag{49}$$

and that there is no ℓ that satisfies the above equations simultaneously. Hence, we conclude $\mathbf{b}_{\ell m} \cong 0$. □

Proof of if: If $\mathbf{a}_{\ell m} \cong 0$ for $\ell \neq 0$ and $\mathbf{b}_{\ell m} = 0$ in an ATI-1 space–time, the vector field $\hat{\xi}^a$ whose $O(\Omega^0)$ and $O(\Omega^1)$ terms are given by Eqs. (41)–(43) satisfy ${}^{(n)}(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}} \cong 0$ for $n \leq 2$ and $\hat{g}_{ab}\hat{\xi}^a\hat{\xi}^b \cong -1$. Hence, $\hat{\xi}^a$ is an asymptotically stationary Killing vector field to order 2. □

The fact that $\mathbf{a}_{\ell m} \cong 0$ for $\ell \neq 0$ and $\mathbf{b}_{\ell m} \cong 0$ means that in such an ATI-1 space–time, the first asymptotic structure takes the simple form

$$ {}^{(1)}\mathbf{F} \cong \mathbf{a}_{00}\,{}^{(1)}\mathbf{F}^{00}, \quad {}^{(1)}\boldsymbol{\psi} \cong -\tfrac{1}{2}\mathbf{a}_{00}\,{}^{(1)}\mathbf{F}^{00}, \quad {}^{(1)}\boldsymbol{\beta} \cong 0, \quad {}^{(1)}\boldsymbol{\chi}_{ab} \cong 0. \tag{50} $$

Before we show that such an ATI-1 space–time is an asymptotically Schwarzschild space–time, we remark an important fact relating to the definition of the angular-momentum of an asymptotically flat space–time. To define angular-momentum, one must impose the condition that the $O(\Omega^1)$ term of the magnetic part of the Weyl tensor vanish.[2–4] However, the physical meaning of the condition was left unclear. The lemma and Eqs. (33)–(50) tells us that the meaning is that the angular-momentum of an asymptotically spacetime can be defined if the spacetime is asymptotically stationary to order 2.

Theorem: An ATI-1 space–time which is asymptotically stationary to order 2 is an asymptotically Schwarzschild space–time with mass a_{00} in the sense that the metric takes the form

$$ \hat{g}_{ab} = \hat{g}^{\mathrm{Sch}}_{ab} + O(r^{-2}) + O(t^{-2}) $$

where

$$ \hat{g}^{\mathrm{Sch}}_{ab} := -\left(1 - \frac{2a_{00}}{r}\right)(dt)_a(dt)_b + \left(1 - \frac{2a_{00}}{r}\right)^{-1}(dr)_a(dr)_b + r^2(d\sigma)_{ab} \tag{51} $$

in which $t > r$.

Proof: From the lemma, in an ATl-1 space–time which is asymptotically stationary to order 2, $\mathbf{a}_{\ell m} = 0$ for $\ell \neq 0$ and $\mathbf{b}_{\ell m} = 0$ hold. Thus, the first-order term of the metric ${}^{(1)}\hat{g}_{ab}$ takes the form

$$ {}^{(1)}\hat{g}_{ab} = (e^{-\eta})^2[a_{00}\,{}^{(1)}F^{00}(d\eta)_a(d\eta)_a + a_{00}\,{}^{(1)}F^{00}h_{ab}]. \tag{52} $$

Under a gauge transformation generated by $\xi^a = \Omega T(\partial_\eta)^a + \Omega D^a L$ where

$$ \mathbf{T} :\cong \mathbf{a}_{00}(2\boldsymbol{\chi}\cosh\boldsymbol{\chi} + \sinh\boldsymbol{\chi}), \quad \mathbf{L} :\cong -4\mathbf{a}_{00}\sinh\boldsymbol{\chi} + \mathbf{T}, \tag{53} $$

${}^{(1)}\hat{g}_{ab}$ transforms as

$$ {}^{(1)}\hat{g}_{ab} \mapsto (e^{-\eta})^2[a_{00}\,{}^{(1)}F^{00}(d\eta)_a(d\eta)_a - 8a_{00}\cosh\chi(d\eta)_{(a}(d\chi)_{b)} + a_{00}\,{}^{(1)}F^{00}(d\chi)_a(d\chi)_b]. \tag{54} $$

Then, the change of variables, $t = \Omega^{-1}\cosh\chi$ and $r = \Omega^{-1}\sinh\chi$, leads us to

$$ O(\Omega^2) = \sinh^{-2}\chi O(\Omega^2) + \cosh^{-2}\chi O(\Omega^2) = O(r^{-2}) + O(t^{-2}), \quad \frac{r}{t} = \tanh\chi < 1. \tag{55} $$

and

$$ {}^{(0)}\hat{g}_{ab} + \Omega\,{}^{(1)}\hat{g}_{ab} = \hat{g}^{\mathrm{Sch}}_{ab} + O(\Omega^2). \tag{56} $$

Hence, Eq. (51) holds. □

It is important to note here that asymptotically Schwarzschild space–times, that are defined in Eq. (51), comprise the Kerr space–time also. This can be understood by writing the Kerr metric $\hat{g}^{\mathrm{Ker}}_{ab}$ with the coordinates (Ω,χ) where $t = \cosh\chi$ and $r = \sinh\chi$:

$$\hat{g}_{ab}^{\mathrm{Ker}}=\hat{g}_{ab}^{\mathrm{Sch}}+\Omega^{2}\,{}^{(2)}\hat{g}_{ab}+O(\Omega^3), \tag{57}$$

where

$$\begin{aligned}{}^{(2)}\hat{g}_{ab}=&(4m^2-a^2\sin^2\theta)\left((d\eta)_{(a}(d\eta)_{b)}-2\,\frac{(d\chi)_{(a}(d\eta)_{b)}}{\tanh\chi}\right)+4ma\,\frac{\sin^2\theta}{\tanh\chi}(d\phi)_{(a}(d\eta)_{b)}\\&+\frac{4m^2-a^2\sin^2\theta}{\tanh\chi}(d\chi)_a(d\chi)_b-4am\sin^2\theta^2(d\phi)_{(a}(d\chi)_{b)}+a^2\cos^2\theta(d\theta)_a(d\theta)_b\\&+a^2\sin^2\theta(d\phi)_a(d\phi)_b\,.\end{aligned} \tag{58}$$

In other words, the Kerr space–time is a special space–time of asymptotically Schwarzschild space–times, which possesses a particular second-order asymptotic structure, i.e., ${}^{(2)}\hat{g}_{ab}$ given by Eq. (58).

V. SUMMARY AND REMARKS

In this paper, we have proved that an asymptotically flat space–time as defined in definition 1 of Sec. II is an asymptotically Schwarzschild spacetime in the sense of Eq. (51), if the energy-momentum of the space–time falls off at the rate faster than $O(\Omega^3)$ and the space–time is asymptotically stationary to order 2 in the sense that $(\mathcal{L}_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}$ falls off at the rate faster than $O(\Omega^2)$ for the asymptotically timelike vector $\hat{\xi}^a, \hat{g}_{ab}\hat{\xi}^a\hat{\xi}^b\doteq-1$.

Finally, we give a remark. Although we have solved the Einstein equation (10) and obtained the first-order asymptotic structure, we did not impose any physically suitable boundary conditions on the solutions. Hence, the obtained first-order asymptotic structure may include those that are *unphysical*. For example, a solution that describes an incoming gravitational wave from future null infinity or an outcoming wave from the event horizon. In other words, *physically* acceptable gravitational fields around a black hole may be obtained only after a suitable boundary condition are imposed on the solutions. The derivation of the physical first-order asymptotic structure may be profitable because it may be possible to show that an ATI-1 space–time with such a physical structure is intrinsically asymptotically stationary to order 1, and thus is an asymptotically Schwarzschild space–time. This result is anticipated because we expect that a space–time with a black hole that becomes vacuum also becomes stationary, due to the nature of the black hole. This is an important point that should be clarified.

ACKNOWLEDGMENTS

We would like to thank K. Sato and Y. Suto for their encouragement. The discussion with T. T. Nakamura was very fruitful. U.G. thanks M. Sasaki, M. Shibata, T. Tanaka, and Y. Mino for the valuable discussions. T.S. thanks Gary Gibbons and the Relativity Group at Cambridge for their hospitality. Finally, we wish to thank Misao Sasaki especially for his comments on the draft of this paper. This work is partially supported by the JSPS fellowship.

[1] M. Heusler, *Black Hole Uniqueness Theorem* (Cambridge University Press, Cambridge, 1996).
[2] A. Ashtekar and J. D. Romano, Class. Quantum Grav. **9**, 1069 (1992).
[3] U. Gen and T. Shiromizu, gr-qc/9709009ver3.
[4] A. Ashtekar and R. O. Hansen, J. Math. Phys. **19**, 1542 (1978); a comprehensive introduction of the ideas can also be found in Ref. 6.
[5] S. Perng, gr-qc/9807082
[6] R. M. Wald, *General Relativity* (The University of Chicago Press, Chicago, 1984).
[7] H. Kodama and M. Sasaki, Prog. Theor. Phys. Suppl. **78**, 1 (1984).
[8] V. F. Mukhanov, H. A. Feldman, and R. H. Brandenberger, Phys. Rep. **2**, 203 (1992).
[9] T. Tanaka and M. Sasaki, Prog. Theor. Phys. **9**, 243 (1997).
[10] K. Tomita, Prog. Theor. Phys. **6**, 310 (1982).

Weyl-type fields with geodesic lines of force

Brendan S. Guilfoyle[a)]
Department of Mathematics, Institute of Technology
Tralee, Tralee, County Kerry, Ireland

(Received 25 September 1998; accepted for publication 12 January 1999)

The static electrogravitational equations are studied and it is shown that an aligned type D metric that has a Weyl-type relationship between the gravitational and electric potential has shear-free geodesic lines of force. All such fields are then found and turn out to be the fields of a charged sphere, charged infinite rod and charged infinite plate. A further solution is also found with shearing geodesic lines of force. This new solution can have $m>|e|$ or $m<|e|$, but cannot be in the Majumdar–Papapetrou class (in which $m=|e|$). It is algebraically general and has flat equipotential surfaces.
[S0022-2488(99)01604-7]

I. INTRODUCTION

Static fields have been a very fruitful area of study for general relativity. On the one hand, the simplifications in the field equations for static fields are substantial, and on the other hand, it is this area that lends itself most easily to comparison with Newtonian gravitation and classical potential theory. For example, the 44-component of the metric tensor, g_{44}, plays the role of a gravitational potential. This allows one to talk of gravitational equipotential surfaces (surfaces of constant g_{44}) and lines of force (the integral curves of the gradient of g_{44}).

Das[1] and Kota and Perjes[2] investigated vacuum metrics in which the lines of force are geodesic. The Newtonian analog for such a field is one with straight lines of force, which is generated by a sphere or an infinite rod or an infinite plate. Das found that among all vacuum fields with geodesic lines of force is included the exterior field of a sphere and an infinite rod and an infinite plate. These constitute all of the vacuum fields in which the lines of force are, in addition, shear-free. Das also discovered a shearing field with no Newtonian analog.

When the source of the field is an electric field, one also has the electric potential ϕ, and the corresponding electric equipotential surfaces and lines of force. Fields in which there is a functional relationship between the two potentials has been investigated by numerous authors (e.g., Weyl,[3] Majumdar,[4] Papapetrou,[5] Gatreau and Hoffman,[6] Guilfoyle[7]). We will refer to such fields as being of the *Weyl-type*.

In this paper we show that among all Weyl-type fields the algebraically special (type D) ones have shear-free geodesic lines of force. We then go on to find all Weyl-type fields with geodesic lines of force. The shear-free metrics turn out to describe the fields of a charged sphere, an infinite charged rod and an infinite charged plate. The only asymptotically flat member of this class is the Reissner–Nordström metric. Again, there is a further shearing field which is algebraically general and has no Newtonian analog. We find, however, that this is not in the Majumdar–Papapetrou class of electrovac solutions. In fact, the only Majumdar–Papapetrou field with geodesic lines of force is found to be the extreme Reissner–Nordström solution.

II. BACKGROUND AND GENERAL RESULTS

A space–time $(M,g_{\mu\nu})$ is *static* if there exists a timelike hypersurface orthogonal Killing vector ξ^μ on M. With a suitable choice of co-ordinates (x^i,t) the metric takes the form

[a)]Electronic mail: brendan.guilfoyle@ittralee.ie

$$ds^2 = g_{ij}\, dx^i\, dx^j - e^{\omega}\, dt^2,$$

where both the spatial metric g_{ij} and e^{ω} depend only on x^i. Here, and throughout, Greek letters will take values 1 to 4, while Latin letters will take values 1, 2, 3. We will raise and lower all indices using g.

The 3+1 split above induces the following decomposition of the Riemann tensor of $(M, g_{\mu\nu})$ in terms of that of the spatial slices (V^3, g_{ij}) (see Synge[8]):

$$R_{ijkl} = {}^3R_{ijkl}, \quad R_{i44l} = -e^{(1/2)\,\omega}(e^{(1/2)\,\omega})_{|il}, \quad R_{ij4l} = 0. \tag{1}$$

A stroke represents covariant differentiation with respect to the Levi-Çivita connection of the spatial metric g_{ij}. The Ricci tensors of $(M, g_{\mu\nu})$ and (V^3, g_{ij}) are then found to be related by

$$R_{ij} = {}^3R_{ij} + e^{-(1/2)\,\omega}(e^{(1/2)\,\omega})_{|ij}, \tag{2}$$

$$R_{44} = -\tfrac{1}{2} e^{\omega}(\triangle\omega + \tfrac{1}{2}|d\omega|^2), \tag{3}$$

where $\triangle$ is the covariant Laplacian $\triangle\omega = g^{ij}\omega_{|ij}$ and $|d\omega|^2 = g^{ij}\omega_{,i}\omega_{,j}$.

We note that the 3-dimensional Riemann tensor of (V^3, g_{ij}) is entirely determined by the 3-dimensional Ricci tensor through

$${}^3R_{ijkl} = g_{i[l}{}^3R_{k]j} + g_{j[k}{}^3R_{l]i} + \tfrac{1}{2}{}^3R g_{i[k}g_{l]j}, \tag{4}$$

where ${}^3R$ is the Ricci scalar ${}^3R \equiv {}^3R^i_i$. Skew-symmetrization (symmetrization) is denoted by square (round) brackets on pairs of indices.

It is often useful to turn to the conformally related background space $(\overline{V}^3, \overline{g_{ij}})$ given by

$$g_{ij} = e^{-\omega}\overline{g_{ij}}.$$

The relationship between the Ricci curvatures of (V^3, g_{ij}) and $(\overline{V}^3, \overline{g_{ij}})$ is

$${}^3R_{ij} = {}^3\overline{R_{ij}} - e^{-(1/2)\,\omega}(e^{(1/2)\,\omega})_{|ij} + \tfrac{1}{2}\omega_{,i}\omega_{,j} - \tfrac{1}{2}g_{ij}(\triangle\omega + \tfrac{1}{2}|d\omega|^2). \tag{5}$$

Einstein's field equations for the electric field ϕ are

$$R_{ij} = 2e^{-\omega}\phi_{,i}\phi_{,j} - e^{-\omega}|d\phi|^2 g_{ij}, \tag{6}$$

$$\Delta\omega = 2e^{-\omega}|d\phi|^2 - \tfrac{1}{2}|d\omega|^2, \tag{7}$$

$$R = 0, \tag{8}$$

while Maxwell's equations reduce to

$$\Delta\phi = \tfrac{1}{2}\omega^{,i}\phi_{,i}. \tag{9}$$

III. ALGEBRAIC STRUCTURE

We now turn to the algebraic structure of the electrostatic field. The most widely used classification scheme for the gravitational field is due to Petrov[9] (see also Kramer *et al.*,[10] Papapetrou[11] and Synge[12]). In this, a metric is classified by the degeneracy of certain null directions associated with the Weyl conformal curvature tensor. This tensor is given in terms of the Riemann and Ricci curvature tensors by

$$C_{\alpha\beta\gamma\delta} = R_{\alpha\beta\gamma\delta} + \tfrac{1}{2}(g_{\alpha[\gamma}R_{\delta]\beta} + g_{\beta[\delta}R_{\gamma]\alpha}) + \tfrac{1}{6}R g_{\alpha[\delta}g_{\gamma]\beta}.$$

A *principal null direction* (*p.n.d.* for short) of the Weyl tensor is a null vector, k^μ, satisfying

$$k_{[\mu}C_{\alpha]\beta\gamma[\delta}k_{\nu]}k^\beta k^\gamma=0.$$

As the name suggests, a p.n.d. is defined only up to multiplication by a function. A repeated p.n.d. for the Weyl tensor is a null vector, k^μ, satisfying

$$C_{\alpha\beta\gamma[s}k_{\nu]}k^\beta k^\gamma=0.$$

Using Eqs. (1) to (8), this can be shown to be equivalent to

$$e^{-(1/2)\omega}(e^{(1/2)\omega})_{|ij}=e^{-\omega}\phi_{,i}\phi_{,j}-\chi\left(\frac{k_ik_j}{|k|^2}-\frac{1}{3}g_{ij}\right), \tag{10}$$

where

$$\chi\equiv\frac{3}{2|k|^2}[e^{-\omega}(k^i\phi_{,i})^2-e^{-(1/2)\omega}(e^{(1/2)\omega})_{|ij}k^ik^j], \tag{11}$$

and the p.n.d. is

$$k^\mu=(k^i,e^{-(1/2)\omega}|k|). \tag{12}$$

We note as a check on these calculations that contracting (10) gives

$$e^{-(1/2)\omega}\Delta(e^{(1/2)\omega})=e^{-\omega}|d\phi|^2,$$

which is precisely the field equation (7).

Thus, a necessary and sufficient condition that a static electrovac space–time is algebraically special (and therefore, type D) is that there exists a 3-vector k^i satisfying (10) and (11). Then the repeated p.n.d. of the Weyl tensor is given by (12) and the Ricci tensor of $(V^3,g_{ij})$is [put (10) in (2)]

$${}^3R_{ij}=\chi\left(g_{ij}-\frac{3k_ik_j}{|k|^2}\right)+e^{-\omega}(\phi_{,i}\phi_{,j}-|d\phi|^2g_{ij}). \tag{13}$$

In the vacuum case $(\phi=0)$ this tells us that a repeated p.n.d. of the Weyl tensor is also an eigenvector of the spatial Ricci tensor ${}^3R_{ij}$. Furthermore, any vector on (V^3,g_{ij}) that is orthogonal to the spatial component, k^i, of the repeated p.n.d., is also an eigenvector of the spatial Ricci tensor. This decomposition of ${}^3R_{ij}$ allows one to integrate the field equations fully, and determine explicitly all type D static vacuum fields, as was done almost 75 years ago by Levi-Çivita.[13] Although the Schwarzchild solution is of this type, it was found that all the other 6 solutions in this class are unphysical (Kramer *et al.*,[10] Ehlers and Kundt[14]).

If we now include the electromagnetic field, the spatial Ricci tensor differs only by the projection operator onto directions orthogonal to $\phi_{,i}$. Thus, we have the following theorem.

Theorem 1: *A repeated p.n.d. of the Weyl conformal tensor of an algebraically special static electrovac field is also an eigenvector of the spatial Ricci tensor if one of the following holds: (i) The electric field vanishes* $(\phi=0)$*; (ii) The electric field is aligned* $(\phi_{,[i}k_{j]}=0)$*; (iii) The electric field is perpendicular* $(\phi_{,i}k^i=0)$.

Since the field equations are tractable in case *(i)* and the algebraic structure of ${}^3R_{ij}$ is the same in (*i*), (*ii*), and (*iii*), we are led to consider the possibility that all aligned or perpendicular type D static electrovac fields may be explicitly determined using Eq. (13) to integrate the field equations.

We shall look at the case where the field is aligned and $\omega_{,i}$ is also coincident with $\phi_{,i}$. In this case, there is a functional relationship between e^ω and ϕ. It is well known that this must be of the form (see Majumdar[4])

$$e^\omega=A+B\phi+\phi^2,$$

from which we find that

$$\phi,_i=\frac{e^\omega}{2(e^\omega+\lambda)^{1/2}}\omega,_i\,,$$

and

$$|d\phi|^2=\frac{e^{2\omega}|d\omega|^2}{4(e^\omega+\lambda)},$$

where

$$\lambda\equiv\frac{B^2}{4}-A,\quad A,B \textit{ constants.}$$

From these we find that (10) becomes

$$\omega_{|ij}=-\frac{\lambda}{2(e^\omega+\lambda)}\omega,_i\omega,_j+2\chi\left(\frac{1}{3}g_{ij}-\frac{\omega,_i\omega,_j}{|d\omega|^2}\right), \tag{14}$$

or transvecting with $\omega^{,j}$,

$$\omega_{|ij}\omega^{,j}=-\left[\frac{4}{3}\chi+\frac{\lambda|d\omega|^2}{2(e^\omega+\lambda)}\right]\omega,_i\,.$$

This is just the geodesic equation,

$$\omega_{|ij}\omega^{,j}=\frac{\omega_{lk}\omega^{,l}\omega^{,k}}{|d\omega|^2}\omega_{,i}\,.$$

We have therefore established the following theorem

Theorem 2: *If a static Weyl-type electrovac space is aligned and type D, then the lines of force are geodesic in* (V^3,g_{ij}).

In fact, we can go further by looking at these equations in the conformal space $(\overline{V}^3,\overline{g_{ij}})$. There, Eq. (14) becomes

$$\omega_{\|ij}=\left[-\frac{e^\omega}{4(e^\omega+\lambda)}+\frac{3}{2}\frac{\omega^{\|kl}\omega,_k\omega,_l}{\|d\omega\|^4}\right]\omega,_i\omega,_j+\left[\frac{\|d\omega\|^2e^\omega}{4(e^\omega+\lambda)}-\frac{1}{2}\frac{\omega^{\|kl}\omega,_k\omega,_l}{\|d\omega\|^2}\right]\overline{g}_{ij}\,,$$

where we have denoted covariant differentiation with respect to the metric $\overline{g_{ij}}$ by a double stroke subscript and $\|d\omega\|^2=\overline{g^{ij}}\omega_i\omega_j$. Thus, the lines of force are also geodesic in $(\overline{V}^3,\overline{g_{ij}})$. Now, if we let

$$k_i=\frac{\omega,_i}{\|d\omega\|},$$

then k^i is a geodesic in $(\overline{V}^3,\overline{g_{ij}})$ and k^μ [given by (12)] is also a null geodesic in the full space–time $(M,g_{\mu\nu})$. In addition, k^i is shear-free in $(\overline{V}^3,\overline{g_{ij}})$:

$$k_{(i\|j)}k^{i\|j}-(\overline{g^{ij}}k_{i\|j})^2=0.$$

This also means, as one would expect from the Goldberg–Sachs theorem,[15] that the p.n.d. k^μ is shear-free in $(M,g_{\mu\nu})$. Thus we have the following:

Theorem 3: *If a static Weyl-type electrovac space is aligned and type D, then the lines of force are geodesic and shear-free in* $(\overline{V}^3,\overline{g}_{ij})$.

In the next section we determine all such metrics explicitly.

IV. TYPE D FIELDS

It is well known that the Reissner–Nordström solution is the unique static asymptotically flat electrovac field with geodesic shear-free lines of force (or *eigenrays*) Kota and Perjés.[2] In this section we present all static electrovac fields with $g_{44}=g_{44}(\phi)$, which possess shear-free geodesic lines of force.

In the vacuum case, all static fields with geodesic lines of force were found by Das[1] and Kota and Perjés.[2] Here we shall adopt the formalism of Das, generalizing it to include the electrostatic field. This formalism entails the setting up of a 3-dimensional complex triad field in $(\overline{V}^3,\overline{g}_{ij})$ and then expressing the field equations in terms of the associated 3-dimensional Ricci rotation co-efficients.

Let $\lambda^i_{(A)}$ be an orthonormal frame in $(\overline{V}^3,\overline{g}_{ij})$ and

$$\Gamma_{(ABC)}\equiv\lambda_{(A)j\|k}\lambda^j_{(B)}\lambda^k_{(C)},$$

be the associated Ricci rotation co-efficients. Here and throughout we use bracketed capital Latin letters $A,B,...=1,2,3$ to represent frame components and summation is implied over any repeated indices. These satisfy the commutation relations

$$\lambda^i_{(A)},_j\lambda^j_{(B)}-\lambda^i_{(B)},_j\lambda^j_{(A)}+\Gamma_{(C[AB])}\lambda^i_{(C)}=0, \tag{15}$$

and the Riemann curvature of $(\overline{V}^3,\overline{g}_{ij})$ is given in terms of them by

$${}^3\overline{R}_{(ABCD)}=\Gamma_{(AB)[(C),(D)]}+\Gamma_{(ABM)}\Gamma_{(M[CD])}+\Gamma_{(MAD)}\Gamma_{(MBC)}-\Gamma_{(MAC)}\Gamma_{(MBD)}. \tag{16}$$

We choose the frame so that the congruence of $\lambda^i_{(1)}$ is normal to the surface $\omega=\text{const}$. Thus

$$\lambda_{(1)i}=U\omega,_i, \quad \Gamma_{(123)}=\Gamma_{(132)}. \tag{17}$$

We choose co-ordinates so that

$$\omega=x^1, \quad \overline{g}_{12}=\overline{g}_{13}=0.$$

Hence (17) can be raised to give

$$\lambda^i_{(1)}=U^{-1}\delta^i_1.$$

Since the lines of force are geodesic, we have that

$$\Gamma_{(131)}=\Gamma_{(121)}=0.$$

Parallelly propagate $\lambda^i_{(2)}$, $\lambda^i_{(3)}$ along $\lambda^i_{(1)}$, so that

$$\Gamma_{(231)}=0.$$

Finally, since we have vanishing shear,

$$\Gamma_{(122)}=\Gamma_{(133)}, \quad \Gamma_{(123)}=0.$$

Label the remaining nonvanishing rotation co-efficients by

$$H\equiv\tfrac{1}{2}(\Gamma_{(122)}+\Gamma_{(133)}),$$

$$\sigma \equiv \frac{1}{\sqrt{2}}(\Gamma_{(233)} - i\Gamma_{(232)}).$$

Here and throughout capital Latin letters will indicate real valued functions and Greek letters will denote complex valued functions.

Furthermore, we make a formal transformation to complex conjugate co-ordinates defined by

$$z^2 \equiv x^2 + ix^3, \quad z^{\bar{2}} \equiv x^2 - ix^3.$$

It should be noted that a co-ordinate transformation of the form

$$x^1 \rightarrow x'^1 = x^1, \quad z^2 \rightarrow z'^2 = f(x^1, z^2), \quad z^{\bar{2}} \rightarrow z'^{\bar{2}} = \overline{f(x^1, z^2)},$$

where $f(x^1,z^2)$ is an analytic function of x^1,z^2 and the bar denotes complex conjugation, does not alter the static form of the metric. This corresponds to a co-ordinate transformation of (x^2,x^3) on the equipotential surfaces $\omega = x^1 = \text{const}$.

The field equations written in terms of frame components in $(\overline{V}^3, \overline{g}_{ij})$ are

$$^3\overline{R}_{(AB)} = 2e^{-\omega}\phi,_{(A)}\phi,_{(B)} - \tfrac{1}{2}\omega,_{(A)}\omega,_{(B)}, \tag{18}$$

$$\omega,_{(AA)} + \Gamma_{(CAA)}\omega,_{(C)} = 2e^{-\omega}\phi,_{(A)}\phi,_{(A)}, \tag{19}$$

with Maxwell's equation

$$\phi,_{(AA)} + \Gamma_{(CAA)}\phi,_{(C)} = \omega,_{(A)}\phi,_{(A)}. \tag{20}$$

Finally, introduce the complex triad field $\Lambda^j_{(A)}$ in complex co-ordinates by

$$\Lambda^j_{(1)} \equiv \lambda^j_{(1)}, \quad \Lambda^j_{(2)} \equiv \frac{1}{\sqrt{2}}(\lambda^j_{(2)} + i\lambda^j_{(3)}),$$

$$\Lambda^j_{(\bar{2})} \equiv \frac{1}{\sqrt{2}}(\lambda^j_{(2)} - i\lambda^j_{(3)}),$$

where complex frame indices take the values 1, 2, $\bar{2}$. We choose complex co-ordinates on the equipotential surfaces so that

$$\Lambda^j_{(2)} = \Sigma\,\delta^j_2, \quad \Lambda^j_{(\bar{2})} = \overline{\Sigma}\,\delta^j_{\bar{2}}$$

for some complex function Σ.

For a Weyl-type electric field the gravitational and electric potentials are related by

$$e^\omega = B + C\phi + \phi^2, \quad B, C \textit{ constants}.$$

Assembling the field equations (18)–(20) [using (4) and (16) together with the commutation relations (15)] gives

$$(\ln\sigma),_1 = -HU, \tag{21}$$

$$UH,_1 + (HU)^2 + \frac{1}{4}\left(1 + \frac{e^{x^1}}{\lambda}\right)^{-1} = 0, \tag{22}$$

$$H,_2 = 0, \tag{23}$$

$$\Sigma\,\overline{\sigma},_2 + \overline{\Sigma}\,\sigma,_{\bar{2}} + 2|\sigma|^2 + H^2 - \frac{1}{4}U^{-2}\left(1 + \frac{e^{x^1}}{\lambda}\right)^{-1} = 0, \tag{24}$$

$$(\ln U)_{,1} - 2HU + \frac{e^{x^1}}{2\lambda}\left(1 + \frac{e^{x^1}}{\lambda}\right)^{-1} = 0, \tag{25}$$

$$U_{,2} = 0, \tag{26}$$

$$(\ln \Sigma)_{,1} = -HU, \tag{27}$$

$$\bar{\Sigma}\Sigma_{,\bar{2}} + \bar{\sigma}\Sigma = 0, \tag{28}$$

where

$$\lambda \equiv \frac{C^2}{4} - B.$$

By (23), (26) we have that $U = U(x^1)$ and $H = H(x^1)$, and so (22) and (25) are ordinary differential equations. In order to integrate this system of equations, we must consider separately the cases $\lambda = 0$, $\lambda > 0$ and $\lambda < 0$. In the asymptotically flat case these correspond to $m = |e|$, $m > |e|$ and $m < |e|$, respectively.

A. $\lambda = 0$

Here, (22), (24) and (25) reduce to

$$UH_{,1} + (HU)^2 = 0 \tag{29}$$

$$HU_{,1} - 2(HU)^2 + \tfrac{1}{2}HU = 0, \tag{30}$$

$$\bar{\Sigma}\sigma_{,\bar{2}} + \Sigma\bar{\sigma}_{,2} + 2|\sigma|^2 + H^2 = 0., \tag{31}$$

Adding the first two of these, we get the Bernoulli equation

$$\frac{d(HU)}{dx^1} + \frac{1}{2}HU = (HU)^2,$$

with the solution

$$HU = \frac{e^{-(1/2)x^1}}{a + 2e^{-(1/2)x^1}}, \tag{32}$$

where a is an arbitrary constant of integration.

Now, (32) in (30) reads as

$$(\ln U)_{,1} - \frac{2e^{-(1/2)x^1}}{a + 2e^{-(1/2)x^1}} + \frac{1}{2} = 0,$$

or, integrating this up,

$$U = \frac{be^{-(1/2)x^1}}{(a + 2e^{-(1/2)x^1})^2}, \tag{33}$$

with b another arbitrary (nonzero) constant of integration. When this is put in (32), we find that

$$H = \frac{1}{b}(a + 2e^{-(1/2)x^1}). \tag{34}$$

Now, having determined the $\overline{g_{11}}$ component of the metric, we return to Eq. (27), which, using (32), integrates up to

$$\Sigma = (a + 2e^{-(1/2)x^1})e^{S+is},$$

where S and s are otherwise arbitrary real-valued functions of $(z^2, \overline{z^2})$. However, s does not contribute to the metric form and so we can, without loss of generality, put $s=0$. Thus

$$\Sigma = (a + 2e^{-(1/2)x^1})e^{S}, \tag{35}$$

and we need only determine S to completely solve the problem.

We do this by first putting (35) in (28), yielding

$$\sigma = -(a + 2e^{-(1/2)x^1})S_{,2}e^{S}, \tag{36}$$

and then Eqs. (34)–(36) in (31) boils down to the simple equation

$$S_{,2\overline{2}} = \frac{e^{-2S}}{2b^2}. \tag{37}$$

In real co-ordinates this is the 2-dimensional Poisson-type equation,

$$\frac{\partial^2 S}{\partial x^{22}} + \frac{\partial^2 S}{\partial x^{32}} = \frac{e^{-2S}}{2b^2}.$$

The general solution of (37) is (see Nehari[16])

$$e^{S} = \frac{1 + |\psi|^2}{\sqrt{2}|b||\psi_{,2}|}, \tag{38}$$

where ψ is an otherwise arbitrary function of z^2. Finally, (21) is identically satisfied by (32) and (36).

We can now assemble the metric with (33), (35), and (38),

$$ds^2 = b^2 e^{-x^1}\left[\frac{e^{-x^1}}{(a+2e^{-(1/2)x^1})^4}(dx^1)^2 + \frac{4|\psi_{,2}|^2|dz^2|^2}{(a+2e^{-(1/2)x^1})^2(1+|\psi|^2)^2}\right] - e^{x^1}dt^2,$$

or making the co-ordinate transformation,

$$x^1 \to x'^1 \equiv x^1, \quad z^2 \to z'^2 \equiv 2\psi, \quad \overline{z^2} \to \overline{z'^2} \equiv 2\overline{\psi}, \tag{39}$$

and subsequently dropping the primes and returning to real co-ordinates,

$$ds^2 = b^2 e^{-x^1}\left[\frac{e^{-x^1}}{(a+2e^{-(1/2)x^1})^4}(dx^1)^2 + \frac{1}{(a+2e^{-(1/2)x^1})^2}\,\frac{(dx^2)^2 + (dx^3)^2}{(1 + (1/4)[(x^2)^2 + (x^3)^2])^2}\right] - e^{x^1}\,dt^2. \tag{40}$$

By writing this metric on a null tetrad and analyzing it with the REDUCE computer algebra system we find that this metric generates a Petrov type D field, as expected. Furthermore, the surfaces

$x^1=$constant in (40) are 2-spaces of constant positive curvature, i.e., spheres and so we expect, by Birkhoff's Theorem, that (40) is just the extreme Reissner–Nordström solution in an unusual co-ordinate system. To see that this is indeed the case we introduce new co-ordinates (θ,φ) on these 2-spaces, defined by

$$\theta\equiv\tan^{-1}\left[\frac{1}{2}((x^2)^2+(x^3)^2)^{1/2}\right],\quad \varphi\equiv\tan^{-1}\left(\frac{x^3}{x^2}\right), \tag{41}$$

so that (40) reduces to

$$ds^2=b^2e^{-x^1}\left[\frac{e^{-x^1}}{(a+2e^{-(1/2)x^1})^4}(dx^1)^2+\frac{d\theta^2+\sin^2\theta d\varphi^2}{(a+2e^{-(1/2)x^1})^2}\right]-e^{x^1}dt^2.$$

Now, if we relabel the equipotential surfaces with ϱ defined by

$$e^{-(1/2)x^1}=-\frac{a}{2}\left(1+\frac{b}{2\varrho}\right),$$

and rescale the time co-ordinate by

$$t\rightarrow t'\equiv\frac{2t}{a},$$

we find, dropping the primes, that (40) becomes

$$ds^2=\left(1+\frac{b}{2\varrho}\right)^2[d\varrho^2+\varrho^2(d\theta^2+\sin^2\theta\,d\varphi^2)]-\left(1+\frac{b}{2\varrho}\right)^{-2}dt^2,$$

which is precisely the extreme Reissner–Nordström solution with

$$b=2m=2|e|.$$

B. $\lambda>0$

In this case we must solve (21)–(28). We still have, by (23) and (26) $U=U(x^1)$ and $H=H(x^1)$, but when we add (22) and (25) we get the Ricatti equation

$$\frac{d(HU)}{dx^1}-(HU)^2+\frac{1}{2}HU\frac{e^{x^1}}{\lambda}\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1}+\frac{1}{4}\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1}=0, \tag{42}$$

with the first integral

$$HU=\frac{1}{2}\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1/2}.$$

So, returning to (42) with a substitution of the form

$$HU=\frac{1}{2}\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1/2}+F(x^1), \tag{43}$$

we get a Bernoulli equation for F with the solution

$$F=\frac{1}{ae^{-I}-1}\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1/2}, \tag{44}$$

where a is an arbitrary constant of integration and I is defined,

$$I\equiv\int\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1/2}dx^1.$$

Plugging (44) in (43) gives the general solution to (42) as

$$HU=\frac{1}{2}\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1/2}+\frac{e^I}{a-e^I}\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1/2}=\frac{1}{2}\left(\frac{ae^{-I}+1}{ae^{-I}-1}\right)\left(1+\frac{e^{x^1}}{\lambda}\right)^{-1/2}. \tag{45}$$

Putting this in (25) and integrating,

$$U=\frac{be^I}{(1+(e^{x^1}/\lambda))^{1/2}(a-e^I)^2}, \tag{46}$$

where b is a nonzero constant of integration. Then (46) in (45) gives

$$H=\frac{1}{2b}e^{-I}(a^2-e^{2I}). \tag{47}$$

Again, having determined $\overline{g}_{11}$ we return to (27), which, with the help of (46) and (47), integrates to

$$\Sigma=(a-e^I)e^{-(1/2)I}e^{S+is}, \tag{48}$$

where S and s are arbitrary functions of $(z^2,\bar{z}^2)$. Without loss of generality, we can again put $s=0$.

Then, (48) in (28) gives

$$\sigma=-(a-e^I)e^{-(1/2)I}S_{,2}e^S, \tag{49}$$

and (21) is identically satisfied by (48) and (49).

Our remaining equation (24) boils down, from (46)–(49) to

$$S_{,2\bar{2}}=\frac{a}{2b^2}e^{-2S}. \tag{50}$$

Here we must consider the solutions to (50) for $a>0$, $a=0$ and $a<0$ separately. These correspond to the cases where the equipotential surfaces are of constant positive, zero and negative curvature, respectively.

$a>0$: In this case the general solution to (50) is

$$e^S=\sqrt{\frac{a}{2b^2}}\left[\frac{1+|\psi|^2}{|\psi_{,2}|}\right],$$

where ψ is an otherwise arbitrary function of z^2.

Assembling our metric, we find that

$$ds^2=b^2e^{-x^1}\left[\frac{e^{2I}(dx^1)^2}{(1+e^{x^1}/\lambda)(a-e^I)^4}+\frac{4e^I|\psi,_2|^2|dz^2|^2}{a(a-e^I)^2(1+|\psi|^2)^2}\right]-e^{x^1}\,dt^2,$$

and then transforming co-ordinates by (39) and dropping the primes, we find

$$ds^2=b^2e^{-x^1}\left[\frac{e^{2I}(dx^1)^2}{(1+e^{x^1}/\lambda)(a-e^I)^4}+\frac{e^I((dx^2)^2+(dx^3)^2)}{a(a-e^I)^2[1+(1/4)\,((x^2)^2+(x^3)^2)]^2}\right]-e^{x^1}\,dt^2.$$

Again, the field is Petrov type D, the equipotential surfaces are 2-spaces of constant positive curvature, i.e., spheres, and we thus expect it to represent the under-charged ($m>|e|$) or over-charged ($m<|e|$) Reissner–Nordström solution in an unfamiliar co-ordinate system. This is indeed the case, as can be seen by utilizing the co-ordinate transformation (41) and relabeling the equipotential surfaces with ϱ defined by

$$e^I\equiv a\left(\frac{b-\varrho}{b+\varrho}\right),$$

or, equivalently,

$$e^{x^1}\equiv\frac{4\lambda a}{(1-a)^2}\,\frac{(1-b^2/\varrho^2)^2}{[1+2((1+a)/(1-a))(b/\varrho)+(b^2/\varrho^2)]^2}.$$

The metric then becomes

$$ds^2=\frac{(1-a)^2}{64a^3}\left[1+2\left(\frac{1+a}{1-a}\right)\frac{b}{\varrho}+\frac{b^2}{\varrho^2}\right]^2[d\varrho^2+\varrho^2(d\theta^2+\sin^2\theta\,d\varphi^2)]$$
$$-\frac{4\lambda a(1-b^2/\varrho^2)^2}{(1-a)^2}\left[1+2\left(\frac{1+a}{1-a}\right)\frac{b}{\varrho}+\frac{b^2}{\varrho^2}\right]^{-2}dt^2,$$

which, with a final rescaling of

$$t\rightarrow\left(\frac{1-a}{2\sqrt{\lambda a}}\right)t,\qquad \varrho\rightarrow\left(\frac{(4a)^{3/2}}{1-a}\right)\varrho,$$

takes the standard isotropic form

$$ds^2=\left[1+2\left(\frac{1+a}{1-a}\right)\frac{b}{\varrho}+\frac{b^2}{\varrho^2}\right]^2[d\varrho^2+\varrho^2(d\theta^2+\sin^2\theta\,d\varphi^2)]$$
$$-\left(1-\frac{b^2}{\varrho^2}\right)^2\left[1+2\left(\frac{1+a}{1-a}\right)\frac{b}{\varrho}+\frac{b^2}{\varrho^2}\right]^{-2}dt^2,$$

with mass and charge parameters given by

$$m=\frac{2(1+a)b}{1-a},\qquad e^2=\frac{16ab^2}{(1-a)^2}.$$

$a=0$: In this case the general solution to (50) is

$$e^{-S}=b|\psi,_2|,\qquad \psi=\psi(z^2),$$

and using (39) our metric takes the simple form

$$ds^2 = b^2 e^{-x^1}\left[\frac{e^{-2I}}{(1+e^{x^1}/\lambda)}(dx^1)^2 + e^{-I}[(dx^2)^2+(dx^3)^2]\right] - e^{x^1}\,dt^2.$$

From this we can see that the equipotential surfaces are flat, and so our metric represents the field of a charged infinite plate, which also turns out to be of Petrov type D. The fact that it is not an asymptotically flat field can be most easily seen by transforming into isotropic co-ordinates. However, rather than do this here, we shall change to the canonical form of the plane symmetric static Einstein–Maxwell field, as given by McVitie.[17] This is achieved by relabeling the equipotential surfaces by z defined by

$$e^{x^1} \equiv \frac{4(\lambda^2+\lambda)^{1/2}}{bz} + \frac{4(1+\lambda)}{b^2 z^2},$$

and rescaling,

$$t \to t' \equiv \left(\frac{\lambda^{1/2} b^2}{(\lambda+1)^{1/2}}\right) t$$

$$x^2 \to x'^2 \equiv \left(\frac{b^2}{(\lambda+1)^{1/2}}\right) x^2, \quad x^3 \to x'^3 \equiv \left(\frac{b^2}{(\lambda+1)^{1/2}}\right) x^3,$$

we get (dropping the primes)

$$ds^2 = \left(\frac{m}{z} + \frac{e^2}{z^2}\right)^{-1} dz^2 + z^2[(dx^2)^2 + (dx^3)^2] - \left(\frac{m}{z} + \frac{e^2}{z^2}\right) dt^2,$$

where

$$m \equiv \frac{4(\lambda+1)^{3/2}}{\lambda^{1/2} b^5}, \quad e^2 \equiv \frac{4(\lambda+1)^2}{\lambda b^6}.$$

$a<0$: In this case (50) has the general solution

$$e^S = \sqrt{\frac{-a}{2b^2}}\left[\frac{1-|\psi|^2}{|\psi,_2|}\right],$$

and so, assembling our metric we find that, after transforming by (39) that

$$d\overline{s^2} = d\left(\frac{1}{e^I - a}\right)^2 - \frac{e^I}{a(e^I-a)^2}\,\frac{(dx^2)^2+(dx^3)^2}{[1-\frac{1}{4}[(x^2)^2+(x^3)^2]]^2}.$$

If we write the metric for the equipotential 2-surfaces of constant negative curvature in the form Stephani[18]

$$d\overline{\sigma^2} = \frac{du^2}{1+u^2} + u^2\,d\varphi^2,$$

and make the co-ordinate transformations

$$x^1 \to \varrho \equiv \left[\frac{-e^I}{a(e^I-a)^2}\right]^{1/2} u, \quad u \to h \equiv (1+u^2)^{1/2}\left[\frac{1}{4a^2} - \frac{\varrho^2}{u^2}\right]^{1/2},$$

we find that the metric is in the Weyl canonical form for a static axially symmetric field. We furthermore find that the field is singular along the axis of symmetry and has only two principal null directions associated with the Weyl curvature tensor (i.e., it is Petrov type D). We therefore interpret the field as being generated by an infinite rod situated along the axis of symmetry $\varrho = 0$.

C. $\lambda<0$

By a series of calculations similar to the $\lambda>0$ case, one obtains the metric generated by an overcharged ($m<|e|$) sphere, infinite rod and infinite plate.

V. SHEARING FIELDS

By a procedure similar to that of the last section, all Weyl-type fields with shearing geodesic lines of force can be found. The metric $\bar{g}$ turns out to be (for details, see Guilfoyle[19])

$$\bar{g}_{11} = \frac{F^{-2}}{\lambda^2(\lambda+e^{x^1})}[1+2k\lambda F[\tfrac{1}{4}((x^2)^2+(x^3)^2)^2+[(x^2)^2-(x^3)^2]\Omega+\Omega^2]],$$

$$\bar{g}_{12} = -\frac{F^{-1}x^2}{\lambda(\lambda+e^{x^1})^{1/2}}[\tfrac{1}{2}((x^2)^2-3(x^3)^2)+\Omega]$$

$$\bar{g}_{13} = -\frac{F^{-1}x^3}{\lambda(\lambda+e^{x^1})^{1/2}}[\tfrac{1}{2}((x^3)^2-3(x^2)^2)-\Omega]$$

$$\bar{g}_{22} = \bar{g}_{33} = \frac{1}{k\lambda}((x^2)^2+(x^3)^2)F^{-1},$$

where k is a constant, Ω is an arbitrary function of x^1 and

$$F(x^1) = \exp\left[(\lambda+4k^2)^{1/2}\int(\lambda+e^{x^1})^{-1/2}\,dx^1\right] = \begin{cases} \left[\dfrac{(\lambda+e^{x^1})^{1/2}-\lambda^{1/2}}{(\lambda+e^{x^1})^{1/2}+\lambda^{1/2}}\right]^n, & \text{for } \lambda>0 \\ \exp\left[-2n\tan^{-1}\left[-\left(1+\dfrac{e^{x^1}}{\lambda}\right)\right]^{1/2}\right], & \text{for } \lambda<0, \end{cases}$$

for $n \equiv [1+(2k)^2/\lambda]^{1/2}$. As expected, this metric is algebraically general, with shear given by

$$|\beta|^2 = k^2\lambda^2F^2 \neq 0.$$

Therefore this cannot be of the Majumdar–Papapetrou type. That is, $\lambda>0$ or $\lambda<0$, but $\lambda=0$ is not possible.

The equipotential surfaces $x^1=\text{const}$ are easily seen to be flat.

VI. CONCLUSION

We have established that all algebraically special Weyl-type electric fields have shear-free geodesic lines of force. In the over-charged ($m<|e|$) and under-charged ($m>|e|$) cases these are the fields generated by a charged sphere, infinite charged rod and infinite charged plate. These correspond exactly to the fields in Newtonian theory that have straight lines of force. Of these, only the sphere is asymptotically flat. In the Majumdar–Papapetrou case ($m=|e|$) the only solution is the extreme Reissner–Nordström solution. All metrics with shearing geodesic lines of force giving rise to a Weyl-type electric field were also found. This class of solutions depends on an arbitrary function of the gravitational potential and the equipotential surfaces are flat. It is algebraically general and cannot be in the Majumdar–Papapetrou class.

ACKNOWLEDGMENT

The author would like to thank Petros Florides, under whose supervision most of the above work was carried out.

[1] A. Das, J. Math. Phys. **14**, 1099 (1973).
[2] J. Kota and Z. Perjes, J. Math. Phys. **13**, 1695 (1972).
[3] H. Weyl, Ann. Phys. (Leipzig) **54**, 117 (1917).
[4] S. D. Majumdar, Phys. Rev. **72**, 390 (1947).
[5] A. Papapetrou, Proc. R. I. Acad. Sect. A, Math. Astron. Phys. Sci. **51**, 191 (1947).
[6] R. Gatreau and R. B. Hoffman, Nuovo Cimento B **16**, 162 (1973).
[7] B. S. Guilfoyle, to appear in Gen. Relativ. Gravit.
[8] J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1966).
[9] A. Z. Petrov, *Einstein Spaces* (Pergamon, New York, 1969).
[10] D. Kramer, H. Stephani, E. Herlt, M. MacCallum and E. Schmutzer, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, England, 1980).
[11] A. Papapetrou, *Lectures on General Relativity* (Reidel, Dordrecht, 1974).
[12] J. L. Synge, Commun. Dublin Inst. Adv. Stud. A **15**, (1964).
[13] T. Levi-Çivita, Rend. Acc. Lincei **26**, 307; **27**, 3, 183, 220, 240, 283, 343; **28**, 3, 101 (1917–1919).
[14] J. Ehlers and W. Kundt, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[15] J. N. Goldberg and R. K. Sachs, Acta Phys. Pol. **22**, 13 (1962).
[16] Z. Nehari, *Introduction to Complex Analysis* (Allyn and Bar, New York, 1968).
[17] G. C. Mcvitie, Proc. R. Soc. London **124**, 366 (1929).
[18] H. Stephani, *General Relativity* (Cambridge University Press, England, 1982).
[19] B. S. Guilfoyle, ''Weyl-type electrostatic fields: New exterior and interior solutions,'' Master's thesis, Trinity College, Dublin, 1996.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

On relativistic spin network vertices

Michael P. Reisenberger[a)]
Instituto de Física, Universidad de la República Igua, 4225 esq. Mataojo, C.P. 11400 Montevideo, Uruguay

(Received 30 September 1998; accepted for publication 15 October 1998)

Barrett and Crane have proposed a model of simplicial Euclidean quantum gravity in which a central role is played by a class of Spin(4) spin networks called ''relativistic spin networks'' which satisfy a series of physically motivated constraints. Here a proof is presented that demonstrates that the intertwiner of a vertex of such a spin network is uniquely determined, up to normalization, by the representations on the incident edges and the constraints. Moreover, the constraints, which were formulated for four valent spin networks only, are extended to networks of arbitrary valence, and the generalized relativistic spin networks proposed by Yetter are shown to form the entire solution set (mod normalization) of the extended constraints. Finally, using the extended constraints, the Barrett–Crane model is generalized to arbitrary polyhedral complexes (instead of just simplicial complexes) representing space-time. It is explained how this model, like the Barret–Crane model can be derived from BF theory, a simple topological field theory [G. Horowitz, Commun. Math. Phys. **125**, 417 (1989)], by restricting the sum over histories to ones in which the left-handed and right-handed areas of any 2-surface are equal. It is known that the solutions of classical Euclidean general relativity form a branch of the stationary points of the BF action with respect to variations preserving this condition. [S0022-2488(99)00904-4]

I. INTRODUCTION

The ''Relativistic spin networks'' defined by Barrett and Crane (BC) are a fundamental ingredient of their proposal for a simplicial model of quantum general relativity in four dimensions. In Ref. 1 the space of intertwiners that relativistic spin networks are allowed to carry is defined implicitly by a set of physically motivated constraints. They also exhibit a single solution to these constraints. Soon after Barbieri[2] gave a partial proof of the uniqueness up to normalization of this solution, which relies on some unproven hypothesies. BC's constraints apply only to 4-valent spin networks, but their solution to their constraints has been generalized in a natural way to arbitrary valence by Yetter,[3] and Barrett[4] has given a very transparent characterization of this extension in the non-q-deformed case, $q=1$.

Here a proof (without auxiliary assumptions) will be given showing that the BC solution is the only one, up to normalization, and similarly Yetter's generalization of the solution (for $q=1$) is the unique solution up to normalization of a natural generalization of the BC constraints. In addition a physically motivated extension of the BC model to polyhedral complexes is outlined. The BC model can be obtained from a sum over histories quantization of Spin(4) BF theory by restricting the histories to ones that assign equal (suitably defined) left-handed and right-handed areas to any 2-surface. At the classical continuum level such a constrained BF theory does reproduce general relativity (GR). The solutions of GR form a branch of the stationary points of the Spin(4) BF action with respect to variations that preserve the constraint that the left- and right-handed areas be equal for all 2-surfaces.[5] This procedure for obtaining the BC model generalizes straightforwardly to complexes of convex 4-polyhedra.

For background information on spin networks see Refs. 6–10.

[a)]Electronic mail: miguel@fisica.edu.uy

II. DEFINITION OF RELATIVISTIC SPIN NETWORKS

I will adopt the following definition of relativistic spin networks, which generalizes BC's definition from 4-valent vertices to vertices of arbitrary valence.

Definition 1: A relativistic spin network is a Spin(4) spin network such that

(1) on each edge the left-handed spin, j_L, and the right-handed spin, j_R, are equal.

(2) in any expansion of an n-valent vertex into a sum of trivalent trees only trivalent trees with $j_L = j_R$ on each of the internal (virtual) edges appear.

In general the edges of a spin network carry nontrivial irreducible representations (irreps) of the gauge group, and the vertices carry intertwiners. The intertwiner for a vertex can be any invariant tensor of the product representation R formed by the product of the irreps carried by the incoming edges and the duals of the irreps on the outgoing edges. (The dual of a representation D of a group is the representation D^{-1T} formed by the transposes of the inverses of the representation matrices of D. If D is unitary its dual is the complex conjugate representation D^* formed by complex conjugating each matrix element in the representation matrices of D.) The space of invariant tensors of R will be denoted $\mathrm{Inv}(R)$. The *evaluation* of a spin network is a complex number calculated by contracting the intertwiners of the vertices along the edges. An intertwiner carries one index for each incident edge. In the evaluation of a spin network the pair of indices associated with each edge (one index lives at each end) is contracted, leaving ultimately a single $\mathbf{C}$ number. In the BC model histories of the gravitational field determine relativistic spin networks on the boundaries of the 4-simplices that form the simplicial space–time, and the probability amplitude of each history is the product of the evaluations of these spin networks (times some simple further factors associated with the lower dimensional simplices).

Relativistic spin networks are spin networks of Spin(4), the covering group of SO(4). Spin(4) is the product of two SU(2)s: Spin(4)=SU(2)×SU(2), where the first SU(2) factor will be called $\mathrm{SU}(2)_L$, the ''left-handed'' subgroup, and the second $\mathrm{SU}(2)_R$, the ''right-handed'' subgroup. This factorization extends to the irreps of Spin(4). These are tensor products of an irrep of $\mathrm{SU}(2)_L$ and an irrep of $\mathrm{SU}(2)_R$, and their carrying spaces, i.e., the vector spaces on which they act, are the tensor products of the carrying spaces of the $\mathrm{SU}(2)_L$ and $\mathrm{SU}(2)_R$ irreps [M. Tinkham, *Group Theory and Quantum Mechanics* (McGraw-Hill, New York, 1964), p. 43].

The factorization of the irreps implies that the product R of irreps incident on a spin network vertex also factorizes into a left-handed factor, R_L, and a right handed factor, R_R, and hence that the intertwiner space factorizes according to

$$\mathrm{Inv}_{\mathrm{Spin}(4)}(R) = \mathrm{Inv}_{\mathrm{Spin}(4)}(R_L \otimes R_R) = \mathrm{Inv}_{\mathrm{SU}(2)}(R_L) \otimes \mathrm{Inv}_{\mathrm{SU}(2)}(R_R) \tag{1}$$

into a tensor product of two SU(2) intertwiner spaces.

SU(2) irreps are determined by their spin, modulo the freedom to change basis in the carrying space. Spin(4) irreps are therefore characterized in the same sense by the spins (j_L, j_R) of their left- and right-handed factors. To keep the mathematics as concrete as possible it is convenient to fix the bases in the carrying spaces so that the irreps take a standard form completely determined by their spins (j_L, j_R). (Note that the evaluation of a spin network is invariant under changes of basis in the carrying spaces of the irreps on the edges. The contractions of the indices of the intertwiners are all between vector indices of some irrep and corresponding covector indices, i.e., vector indices of the dual of the irrep.) We shall adopt as our standard (j_L, j_R) irrep

$$U^{(j_L)} \otimes U^{(j_R)*}, \tag{2}$$

where for each spin j, $U^{(j)}$ is a particular, *unitary*, spin j SU(2) irrep fixed once and for all by some conventions, and $U^{(j)*}$ is its dual irrep. [For instance we may choose

$$U^{(j)m}{}_n(g) = \binom{2j}{j+m}^{-1/2} \binom{2j}{j+n}^{-1/2} \sum_{\Sigma M_i = m, \Sigma N_i = n} g^{M_1}{}_{N_1} \cdots g^{M_{2j}}{}_{N_{2j}} \quad \forall g \in \mathrm{SU}(2). \tag{3}$$

Here the indices M_i and N_i range over $\{-\frac{1}{2},\frac{1}{2}\}$.] This choice can be made because, first, the compactness of SU(2) implies that its irreps preserve a Hermitean inner product, and are thus unitary in orthonormal bases with respect to this inner product, and second because the dual of a spin j SU(2) irrep is also a spin j irrep, so that the Spin(4) irrep chosen really has right-handed spin j_R.

Now let us consider a spin network vertex. First we will define some notation. Once the bases are fixed the product representation R formed by the incident irreps is completely determined by the incident spins and whether edges are incoming or outgoing. This information can be gathered into two vectors, $\mathbf{j}_L$ and $\mathbf{j}_R$, which we shall, in a slight abuse of language, refer to as the vectors of incident left-handed and right-handed spins. Each entry of $\mathbf{j}_L$ corresponds to an incident edge and consists of the left-handed spin, j_L, if the edge is incoming and $-j_L$ if the edge is outgoing. The left-handed factor of R is then $R_L = R(\mathbf{j}_L)$ where

$$R(\mathbf{j}) = \bigotimes_n U^{(j_n)} \tag{4}$$

(n numbers the edges and we have defined $U^{(-j)} \equiv U^{(j)*}$). $\mathbf{j}_R$ is defined in complete analogy to $\mathbf{j}_L$, so our conventions for the Spin(4) irreps on the edges imply that $R = R(\mathbf{j}_L) \otimes R(\mathbf{j}_R)^*$. If $\mathcal{H}(\mathbf{j})$ is the carrying space of the SU(2) representation $R(\mathbf{j})$, and $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}) \equiv \mathrm{Inv}_{\mathrm{SU(2)}}(R(\mathbf{j}))$ is its invariant subspace then the carrying space of $R = R(\mathbf{j}_L) \otimes R(\mathbf{j}_R)^*$ is $\mathcal{H}(\mathbf{j}_L) \otimes \mathcal{H}(\mathbf{j}_R)^*$ and the Spin(4) intertwiner space is $\mathrm{Inv}_{\mathrm{Spin(4)}}(\mathbf{j}_L, \mathbf{j}_R) = \mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_L) \otimes \mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_R)^*$.

The inner product preserved by the unitary representation $R(\mathbf{j})$ establishes a one to one correspondence between vectors of $\mathcal{H}(\mathbf{j})^*$ and linear functions $\mathcal{H}(\mathbf{j}) \to \mathbf{C}$. A tensor $\Psi \in \mathcal{H}(\mathbf{j}_L) \otimes \mathcal{H}(\mathbf{j}_R)^*$ can thus be viewed as a linear mapping $\Psi : \mathcal{H}(\mathbf{j}_R) \to \mathcal{H}(\mathbf{j}_L)$. If the tensor Ψ is an intertwiner it maps $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_R)$ into $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_L)$ and the orthogonal complement of $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_R)$ in $\mathcal{H}(\mathbf{j}_R)$ to zero.

Condition 1 in the definition of relativistic spin networks implies that $\mathbf{j}_L = \mathbf{j}_R$ at their vertices. Thus a relativistic intertwiner Φ may be viewed as a mapping of $\mathcal{H}(\mathbf{j}_R)$ into *itself*, that furthermore maps $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_R)$ into itself and the orthogonal complement of $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_R)$ to zero. It is the composition of the orthogonal projector P onto $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_R)$ and a linear mapping X of $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j}_R)$ to itself: $\Phi = XP$.

Condition 2 in the definition of relativistic spin networks refers to trivalent tree expansions of spin network vertices. It is well known (see Refs. 6 and 11) that for SU(2) spin networks each trivalent tree graph having the same external edges as a given vertex defines a basis of the intertwiner space $\mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j})$ of the vertex. (The trivalent tree has oriented edges and a cyclic ordering of the edges incident at each vertex.) Each element of such a ''trivalent tree basis'' is associated with an assignment $\mathbf{J}$ of (possibly zero) spins to the internal edges of the tree (also known as ''virtual'' edges because they are not present in the actual spin network). The basis element is evaluated by contracting the intertwiners of the trivalent vertices along the internal edges as in a spin network evaluation. This leaves just the intertwiner indices associated with the external edges free. To complete the definition one needs to specify the trivalent intertwiners. The trivalent intertwiner spaces $\mathrm{Inv}_{\mathrm{SU(2)}}(j_1, j_2, j_3)$ of SU(2) are all one dimensional, so it is sufficient to fix the freedom to multiply the intertwiners by scalar factors. We will choose the trivalent intertwiners to be normalized. Then, if a normalizing factor $\sqrt{2J+1}$ is included for each internal edge, the trivalent tree intertwiners will be normalized. There remains a phase which must be chosen by convention. We will suppose that such a convention has been adopted, so that the trivalent tree T and the vector $\mathbf{J}$ of spins on the internal edges determine a unique intertwiner $|T,\mathbf{j},\mathbf{J}\rangle \in \mathrm{Inv}_{\mathrm{SU(2)}}(\mathbf{j})$. [The Wigner 3-$jm$ symbols, $\begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix}$, with the standard convention[6] that $(-1)^{j_1-j_2+j_3}\begin{pmatrix} j_1 & j_2 & j_3 \\ j_1 & j_3-j_1 & -j_3 \end{pmatrix}$ is real and non-negative for all j_1, j_2, and j_3, form a basis of trivalent intertwiners consistent with the convention for $U^{(j)}$ of (3). Aside from being real this basis also has the attractive feature that the intertwiners have simple symmetry properties under permutations of the incident edges: If $j_1+j_2+j_3$ is even they are symmetric, if it is odd they are antisymmetric.] Spin(4) trivalent tree bases can then be constructed from the SU(2) trivalent tree bases: If the trivalent tree T spans a Spin(4) vertex with incident spins $(\mathbf{j}_L, \mathbf{j}_R)$ then the multiplet

$$\{|T,\mathbf{j}_L,\mathbf{J}_L\rangle \otimes \langle T,\mathbf{j}_R,\mathbf{J}_R|\}_{\mathbf{J}_D,\mathbf{J}_R} \tag{5}$$

spans the intertwiner space $\mathrm{Inv}_{\mathrm{Spin}(4)}(\mathbf{j}_L,\mathbf{j}_R)$. Here $\langle T,\mathbf{j},\mathbf{J}|$ denotes the complex conjugate of the tensor $|T,\mathbf{j},\mathbf{J}\rangle$, which, as has been explained, defines a linear function $\mathcal{H}(\mathbf{j})\to\mathbf{C}$ via the Hermitean inner product, thus justifying the Dirac bra notation. The definition of a relativistic spin network implies that the expansion of an intertwiner Φ of such a spin network on a trivalent tree basis has the form

$$\Phi = \sum_{\mathbf{J}} a^T_{\mathbf{jJ}} |T,\mathbf{j},\mathbf{J}\rangle \otimes \langle T,\mathbf{j},\mathbf{J}|, \tag{6}$$

i.e., the left- and right-handed spins are equal on both external and internal edges, for *any* tree T spanning the vertex.

III. THE SOLUTION TO THE CONSTRAINTS DEFINING RELATIVISTIC VERTICES

Now that relativistic spin networks have been defined and explained we are ready to state and prove our result:

Theorem: *The intertwiner of any vertex in a relativistic spin network is uniquely determined, up to a numerical factor, by the irreps on the incident edges. Let* $\mathbf{j}=\mathbf{j}_L=\mathbf{j}_R$ *be the common value of the vectors of left- and right-handed incident spins (as defined in Sec. II) at a vertex, and let* $\mathcal{H}(\mathbf{j})$ *be the space of tensors transforming under the product of the* SU(2) *irreps with these spins, equipped with the Hermitean inner product preserved by* SU(2). *When the bases in the carrying spaces of the* $\mathrm{SU}(2)_L$ *and* $\mathrm{SU}(2)_R$ *irreps on the edges of the relativistic spin network are chosen so that the* $\mathrm{SU}(2)_L$ *irrep is the dual of the* $\mathrm{SU}(2)_R$ *irrep on each edge incident on the vertex then*

(1) tensors, like the intertwiner, that transform under the product of the incident Spin(4) *irreps live in the tensor product of* $\mathcal{H}(\mathbf{j})$ *and its dual, and can thus be viewed as linear mappings of* $\mathcal{H}(\mathbf{j})$ *to itself, and*

(2) the intertwiner is proportional to the orthogonal projector from $\mathcal{H}(\mathbf{j})$ *to the subspace of invariant tensors* $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})\subset\mathcal{H}(\mathbf{j})$.

Proof: Part (1) has already been established in Sec. II, so only part (2) remains to be proven. (6) shows that an intertwiner, Φ, of a relativistic spin network is the composition of the projector P from the space $\mathcal{H}(\mathbf{j})$ of tensors with spins $\mathbf{j}$ onto the subspace of intertwiners $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$ and a linear map X of $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$ to itself. (6) furthermore requires that X is *diagonal* in each of the trivalent tree bases of $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$. Obviously $X=c\mathbf{1}$ with $c\in\mathbf{C}$ satisfies this condition, so cP is a relativistic intertwiner. To establish the theorem it remains to be shown that the set of all trivalent tree bases is rich enough so that this is the only solution to the condition.

Let us first consider four valent vertices. There are three (unoriented) trivalent trees matching the four incident edges,

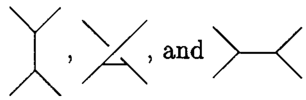

each of which has one internal, or virtual, edge. Once the orientations of the external edges are fixed to match those of the four valent vertex being expanded there remains the freedom to choose the orientation of the internal edge and the cyclic ordering of the incident edges at the two trivalent vertices. However these choices only affect the *signs* of the corresponding trivalent tree basis, modulo sign it is determined by the unoriented tree graph.

Let us number the incident edges 1, 2, 3, 4 clockwise from the top left, and let G_{ni} for $n \in \{1,2,3,4\}$ be the generators, acting in $\mathcal{H}(\mathbf{j})$, of the SU(2) irreps on the incident edges. Let us also define $G_{mni} = (G_m + G_n)_i$. (If the edge n is incoming then $[G_{ni}, G_{nj}] = i\epsilon_{ij}^{k} G_{nk}$. If the edge is outgoing the negatives $-G_{ni}$ satisfy these communication relations. The generators belonging to distinct edges of course commute.) The first of the trivalent trees drawn above corresponds to a pairing of edges 1 and 2, which join at a trivalent vertex. The corresponding basis intertwiners

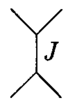

are contractions of a trivalent intertwiner at this vertex and one at the other vertex of the graph. The invariance of the intertwiner at the first vertex implies that

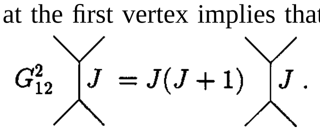

(7)

The intertwiner basis

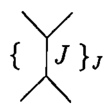

is thus the eigenbasis of G_{12}^2 in $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$. Similarly the trivalent tree bases associated with the second and third trees diagonalize G_{13}^2 and G_{14}^2, respectively.

Since X, the restriction of Φ to $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$, is diagonal in all the trivalent tree bases it commutes with G_{12}^2, G_{13}^2, and G_{14}^2. Further, since the spectra of these operators $(\{J(J+1)\})$ are nondegenerate, X can be expressed as a function of any one of them. Choosing G_{12}^2 we write

$$X = \sum_{q=0}^{d-1} b_q [G_{12}^2]^q, \tag{8}$$

where d is the dimension of $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$. Since the spin J defined by the eigenvalues $J(J+1)$ of G_{12}^2 can take only d values a polynomial of degree $d-1$ can reproduce any dependence of X on G_{12}^2.

Because X also commutes with G_{13}^2

$$0 = [G_{13}^2, X] = \sum_{q=0}^{d-1} b_q [G_{13}^2, (G_{12}^2)^q]. \tag{9}$$

This condition implies that $b_q = 0\ \forall q \neq 0$, so that $X = b_0 \mathbf{1}$. To prove this it is sufficient to consider the matrix elements of (9) between the basis intertwiners

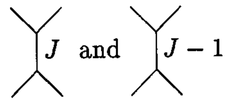

(which I will denote $|J\rangle$ and $|J-1\rangle$ in the following) for all allowed values of J. $\langle J|[G_{13}^2,(G_{12}^2)^q]|J-1\rangle$ is obviously zero when $q=0$. When $q\geqslant 1$ it equals

$$\sum_{r=0}^{q-1} \langle J|(G_{12}^2)^r[G_{13}^2,G_{12}^2](G_{12}^2)^{q-1-r}|J-1\rangle=\beta_J P_q(J), \tag{10}$$

where $\beta_J=\langle J|[G_{13}^2,G_{12}^2]|J-1\rangle$, and

$$P_q(J)=\sum_{r=0}^{q-1} [(J)(J+1)]^r[(J-1)J]^{q-1-r}. \tag{11}$$

(9) therefore implies

$$0=\beta_J\sum_{q=1}^{d-1} b_q P_q(J). \tag{12}$$

The matrix elements of $[G_{13}^2,G_{12}^2]$ have been worked out explicitly by Levý-Leblond and Levý-Nahas.[12] They find

$$\beta_J=\frac{2j_4+1}{\sqrt{4J^2-1}}\{[(j_1+j_2+1)^2-J^2][J^2-(j_1-j_2)^2]\}^{1/2}\{[(j_3+j_4+1)^2-J^2][J^2-(j_3-j_4)^2]\}^{1/2}. \tag{13}$$

[(13) corresponds to Eq. (2.17) of Ref. 12. Their J is called j_4 in our notation and their l is our J. Our formula has an extra factor $(2j_4+1)$ relative to theirs because, while they are evaluating the matrix element between states of *three* spins having definite values (j_4 and m_4) of the magnitude and 3 axis component of the total spin, we are calculating the matrix elements between states of *four* spins which have total spin zero. In this latter calculation one must sum over the $2j_4+1$ possible values of m_4. It is also interesting to note that the operator $[G_{13}^2,\ G_{12}^2]=-4i\epsilon^{ijk}G_{1i}G_{2j}G_{3k}$ has a physical interpretation in loop quantum gravity. It is $-4i$ times the squared 3-volume associated with the four valent vertex.[13–15]]

The important feature of (13) for us is that it is nonzero for all values of J such that both J and $J-1$ satisfy the triangle inequalities

$$|j_1-j_2|\leqslant \text{spin}\leqslant j_1+j_2, \tag{14}$$

$$|j_3-j_4|\leqslant \text{spin}\leqslant j_3+j_4 \tag{15}$$

for the spin on the internal edge of

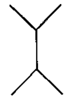

.

It is thus nonzero for all but the smallest of the values of J corresponding to the intertwiners $|J\rangle$ spanning $\mathrm{Inv}_{SU(2)}(\mathbf{j})$. The condition (12) therefore implies

$$0=\sum_{q=1}^{d-1} b_q P_q(J) \tag{16}$$

for all but one of the d allowed values of J.

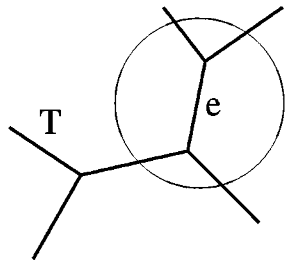

FIG. 1. The circle contains a four valent fragment of the trivalent tree graph T.

Now note that, first, the highest order term in $P_q(J)$ is $J^{2(q-1)}$, and, second, that $P_q(-J) = P_q(J)$. It follows that the P_q have the form $P_q(J) = \Sigma_{p=0}^{d-2} A_{qp} J^{2p}$, where A is a $(d-1)\times(d-1)$ matrix. Moreover, since their leading powers of J are all distinct, the P_q are linearly independent polynomials, implying that A is invertible. From (16) it follows that $\Sigma_{q=1}^{d-1} b_q A_{qp} = 0 \forall p \in \{0,...,d-2\}$. The invertibility of A then shows that $b_q = 0 \forall_q \in \{1,...,d-1\}$. As claimed $X = b_0 \mathbf{1}$.

The result is thus established for four valent vertices. Let us now consider a vertex of arbitrary valence. Fix a particular trivalent tree graph T expanding the vertex and consider a four valent fragment of the graph consisting of an internal edge e, and the four (internal or external) edges attached to it (see Fig. 1). The preceeding arguments relating to four valent vertices can be applied directly to this fragment and show that the eigenvalue $a_{\mathbf{jJ}}^T$ of X on $|T,\mathbf{j},\mathbf{J}\rangle$ is independent of the spin on e. Since this holds for any internal edge of T the eigenvalue is independent of all the internal spins, i.e., it has a common value $c \in \mathbf{C}$ on each element of the basis $\{|T,\mathbf{j},\mathbf{J}\rangle\}_{\mathbf{J}}$. Hence $X = c\mathbf{1}$ and the relativistic intertwiner is cP. ■

I close with a few observations.

(1) The BC model has a simple, physically motivated extension to arbitrary polyhedral complexes (as opposed to simplicial complexes) representing space–time. The generalized BC constraints of definition (1) are equivalent to the requirement that relativistic spin networks define equal left- and right-handed areas for any surface, including ones cutting through vertices. Here the left- and right-handed areas are determined from the left- and right-handed spins on the spin network edges and on the virtual edges of trivalent tree expansions of the vertices using the area operator of loop quantum gravity.[13,16,17]

In Ref. 5 it has been shown that GR is a branch of the theory obtained by restricting SO(4) (or Spin(4)) BF theory to histories in which left-handed and right-handed areas are equal. That is, the solutions of GR form a branch of the stationary points of the SO(4) BF action with respect to variations that respect the constraint that left-handed areas equal right-handed areas. This provides a motivation of the BC model which may be extended to polyhedral complexes: Spin(4) BF theory is just two noninteracting SU(2) BF theories, corresponding to $SU(2)_L$ and $SU(2)_R$, respectively. Thus Ooguri's[18] simplicial lattice sum over histories quantization of SU(2) BF theory [this model is also known as the Crane–Kauffman–Yetter model[19]] immediately provides a simplicial quantization of Spin(4) BF theory. The BC model is then obtained from simplicial Spin(4) BF theory by restricting the histories to ones in which left- and right-handed areas are equal. A history in the Spin(4) BF theory defines a Spin(4) spin network on the boundary of each 4-simplex (or more precisely, on the 1-skeleton of the dual of the boundary seen as a three-dimensional simplicial complex), which plays the role of boundary data in the sense that the spins and intertwiners of the spin networks on two neighboring 4-simplices must match in their mutual boundary. The requirement that left- and right-handed areas be equal then reduces the allowed Spin(4) spin networks to just relativistic spin networks.

Ooguri's quantization of BF theory is most easily generalized to arbitrary polyhedral complexes in the connection formulation (see Ref. 20 for a detailed discussion). In this formulation the boundary data on each 4-cell is a lattice connection [of Spin(4) in our case] defined by the parallel transport matrices across the 2-cells separating the 3-cells of the boundary of the 4-cell. (Equivalently, it is a lattice connection on the 1-skeleton of the dual of the boundary). The amplitude of this connection is a delta distribution with support on flat connections. Clearly the sum over histories yields the same states on the boundary of the space–time (once infinities stemming from redundancies in the delta functions are factored out) whether simplices or arbitrary convex polyhedral 4-cells form the space–time complex.

If one transforms the sum over connection boundary data to a sum over spin network boundary data (see Ref. 20) one finds, for polyhedral complexes as for simplicial complexes, that the amplitudes of each history is just the product of the evaluations of the spin networks on the 4-cells, times a factor $(2j+1)^2$ for each 2-cell, with $j=j_L=j_R$ the common value of the spins carried by the spin network edges crossing that 2-cell. Applying the constraint that left areas equal right areas for all surfaces, even ones crossing the vertices of the spin networks, restricts the spin networks on the boundaries of the 4-cells to be relativistic spin networks in the extended sense of our definition (1). This defines the generalization of the BC model to polyhedral complexes.

(2) To completely determine the BC sum over histories a normalization has to be chosen for the relativistic intertwiners. The four valent relativistic intertwiner given by BC in Ref. 1 seems to be just P. On the other hand, if the sum over histories is to be truly a restriction of the sum over histories for BF theory then the relativistic intertwiner must be normalized in the sense that its contraction on all indices with its complex conjugate must be 1. Thus it must be $P/\sqrt{d}$ up to a phase, where d is the dimensionality of $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$.

(3) Barrett[4] has shown that Yetter's extension to arbitrary valence of the four valent relativistic intertwiner found by BC is equal to

$$\int_{\mathrm{SU}(2)} dg \prod_{j\in\mathbf{j}} U^{(j)}(g) \tag{17}$$

(when the conventions fixing the bases in the irrep carrying spaces are adopted). Here $U^{(j)}(g)$ is the spin j representation matrix of $g\in\mathrm{SU}(2)$ corresponding to the basis convention, and the normalized Haar measure is used to integrate over the group. (17) is precisely the orthogonal projector P on $\mathrm{Inv}_{\mathrm{SU}(2)}(\mathbf{j})$. Thus the theorem shows that the unique solution (mod normalization) of our generalization of BC's constraints is Yetter's extension of their solution.

(4) Since j_R is everywhere equal to j_L the BC state sum can be viewed as a sum over histories of left-handed ''fields,'' i.e., the j_L only. Thus the BC model can be viewed, like the model of Ref. 20 as a (proposal for) a formulation of quantum GR in terms of ''self-dual'' variables. (Note that this does *not* mean that it is a model of only the self-dual sector of GR, in which the anti-self-dual curvature vanishes. It only means that exclusively self-dual variables are used to express the configuration of the gravitational field.)

ACKNOWLEDGMENTS

I would like to thank John Barrett, Rodolfo Gambini, and Carlo Rovelli for encouragement and discussions regarding the subject of this note and closely related issues. This work was funded by the Comision Sectorial de Investigación Cientifica (CSIC) of the Universidad de la República.

[1] J. W. Barrett and L. Crane, ''Relativistic spin networks and quantum gravity,'' J. Math. Phys. **39**, 3296 (1998); also available as gr-qc 9709028, 1997.
[2] A. Barbieri, ''Space of the vertices of relativistic spin networks,'' gr-qc/9709076, 1997.
[3] D. Yetter, ''Generalized Barrett–Crane vertices and invariants of embedded graphs,'' gr-qc/9801131, 1998.
[4] J. W. Barrett, ''The classical evaluation of relativistic spin networks,'' gr-qc/9803063, 1998.
[5] M. Reisenberger, ''Classical Euclidean general relativity from left-handed area=right-handed area,'' gr-qc/9804061, 1998.
[6] A. P. Yutsis, I. B. Levinson, and V. V. Vanagas, ''Mathematical apparatus of the theory of angular momentum,'' Israel program for scientific translations, Jerusalem, 1962.
[7] R. Penrose, ''Angular momentum: An approach to combinatorial spacetime,'' in *Quantum Theory and Beyond*, edited by T. Bastin (Cambridge University Press, Cambridge, 1969).
[8] C. Rovelli and L. Smolin, ''Spin networks and quantum gravity,'' Phys. Rev. D **52**, 5743 (1995).
[9] J. Baez, ''Spin network states in gauge theory,'' Adv. Math. **117**, 253 (1996).

[10] M. Reisenberger, ''Worldsheet formulations of gauge theories and gravity,'' gr-qc/9412035, 1994.
[11] R. DePietri and C. Rovelli, ''Geometry eigenvalues and scalar product from recoupling theory in loop quantum gravity,'' Phys. Rev. D **54**, 2664 (1996).
[12] J-M. Lévy-Leblond and M. Lévy-Nahas, ''Symmetrical coupling of three angular momenta,'' J. Math. Phys. **6**, 1372 (1965).
[13] C. Rovelli and L. Smolin, ''Discreteness of volume and area in quantum gravity,'' Nucl. Phys. B **442**, 593 (1995). Erratum, **456**, 734 (1995).
[14] A. Ashtekar and J. Lewandowski, ''Quantum theory of geometry. II. Volume operators,'' gr-qc/9711031, 1997.
[15] A. Barbieri, ''Quantum tetrahedra and simplicial spin networks,'' Nucl. Phys. B **518**, 714 (1998); also available as gr-qc 9707010, 1997.
[16] A. Ashtekar and J. Lewandowski ''Quantum theory of geometry I: Area operators,'' Class. Quantum Grav. **14**, A55 (1997).
[17] S. Fritelli, L. Lehner, and C. Rovelli, ''The complete spectrum of the area from recoupling theory in loop quantum gravity,'' Class. Quantum Grav. **13**, 2921 (1996).
[18] H. Ooguri, ''Topological lattice models in four dimensions,'' Mod. Phys. Lett. A **7**, 2799 (1992).
[19] L. Crane, L. Kauffman, and D. Yetter, ''State sum invariants of 4-manifolds,'' J. Knot Theory Ram. **6**, 177 (1997).
[20] M. Reisenberger, ''A lattice worldsheet sum for 4-d Euclidean general relativity,'' gr-qc/9711052, 1997.

Generalized Kronecker and permanent deltas, their spinor and tensor equivalents and applications

R. L. Agacy[a)]
42 Brighton Street, Gulliver, Townsville, Qld 4812, Australia

(Received 14 September 1998; accepted for publication 8 December 1998)

The aim of this paper is fourfold: (i) to introduce a *generalized permanent delta* on an equal footing with the generalized Kronecker delta, to use for the symmetries of any tensor or spinor, (ii) to cite an ancillary reference source of comprehensive tensorial and spinorial combinatorial formulas for both, (iii) to table spinor equivalents of these individual tensors and give examples of their usage, and (iv) to tabulate the tensor equivalents of various useful combinations of their spinor forms.
 [S0022-2488(99)03303-4]

I. INTRODUCTION

The generalized Kronecker delta (gKd) is well known as an alternating function or *antisymmetrizer*, e.g., $\delta^{abc}_{def}X_{abc}=3!X_{[def]}$. In contrast, although symmetric tensors are seen constantly in the mathematical language for general relativity (GR), there does not appear to be employment of any *symmetrizer*, for e.g., $X_{(abc)}$, in analogy with the antisymmetrizer. Historically, nomenclature for both date back to Cauchy 1812 in terms of ''fonctions symétriques alternées'' and ''fonctions symétriques permanentes.''[1] A permanent symmetrized tensor was probably first introduced by Cramlet,[2,3] but seems to have been neglected, never acquiring the same prominence as the gKd. However, it is exactly the combination of *both* types of symmetrizers, treated equally, that give us the flexibility to describe any type of tensor symmetry, being exactly the symmetrizers needed to describe Young tableaux tensors.[4] We define such a ''permanent'' symmetrizer as a generalized permanent delta (gpd) below. Our purpose is to restore an imbalance between the gKd and gpd and cite a reference of combinatorial formulas for them, most of which is not in the literature. The contents of this Ref. 5, referred to as PAPS, which contains a *comprehensive* tabulation of gKd and gpd formulas is amplified at the end of this section. In Sec. II we define the gpd and gKd and state some very basic relations. In Sec. III we illustrate the usage of the generalized deltas in the symmetries of tensors. In Sec. IV examples are given of finding spinor equivalents of simple tensors almost instantly, using gKd and gpd spinors. A table of spinor equivalents of individual tensor generalized deltas are tabulated in Appendix A. Converse to finding spinor equivalents, we next tabulate the tensor equivalents of spinor generalized deltas in Appendix B. They can be applied to obtain a variety of spinor formulas.

General indices range from $1,...,n$. All index sets are permutations of each other. Two-component spinor indices are in capital lower case roman.

The PAPS reference[5] contains the following parts and sections: Part I; formulas for *n*-dimensions, (1) Introduction, (2) Definitions and illustrations of the gKd and gpd, (3) Generalized Kronecker δ, (4) Generalized permanent δ, (5) Combined gKd and gpd, Part II; formulas for tensors and spinors in the mathematical language of GR, (6) Spinor equivalents of gKd and gpd tensors, (7) Tensor equivalents of gKd and gpd spinors. Appendixes A–F provide tabulations of all formulas. The *derivation* of each and every formula is provided in the above sections.

II. DEFINITIONS AND BASICS OF THE gKd AND gpd

Complete symmetry of (a tensor's) indices, as opposed to total antisymmetry, is manifested by all positive signs in any *p*-linear expression. Whereas the gKd's antisymmetry comes about

[a)]Electronic mail: ragacy@ultra.net.au

through a *determinant* (interchanges of rows/columns or indices changes the sign), total positive or pure or *permanent* symmetry, as we term it, comes about through the use of a *permanent*. Then in complete analogy to the gKd we introduce the *generalized permanent delta* or *gpd*. This is defined, like the gKd determinant of a matrix, *except that we take all positive signs*. We use the kernel letter π to denote a gpd and *double* vertical lines for the *permanent* of the defining matrix. The gKd and gpd are completely complementary to each other and are defined, for $p(\leqslant n)$ distinct indices, respectively by,

$$\delta^{i_1\cdots i_p}_{j_1\cdots j_p}=\begin{vmatrix} \delta^{i_1}_{j_1} & . & . & \delta^{i_1}_{j_p} \\ . & . & . & . \\ . & . & . & . \\ \delta^{i_p}_{j_1} & . & . & \delta^{i_p}_{j_p} \end{vmatrix}, \qquad \pi^{i_1\cdots i_p}_{j_1\cdots j_p}=\left\Vert \begin{matrix} \delta^{i_1}_{j_1} & . & . & \delta^{i_1}_{j_p} \\ . & . & . & . \\ . & . & . & . \\ \delta^{i_p}_{j_1} & . & . & \delta^{i_p}_{j_p} \end{matrix} \right\Vert ,$$

where the first is a determinant and the second a permanent. The gKd has value $+1$ (-1) depending on whether $(j_1,\ldots,j_p)$ is an even (odd) permutation of $(i_1,\ldots,i_p)$. The gpd has the *permanent* value $+1$ for any permutation of the index sets. Note that $\pi^a_b=\delta^a_b$. The gpd is a complete *symmetrizer* (*permanent symmetrizer*) and now one easily sees that $\pi^{abc}_{def}X_{abc}=3!X_{(def)}$.

A simple but important interaction between the gKd and gpd is that in any expression containing them as a product, where there is a summation on a pair of indices between them, such expression vanishes. For

$$\delta^{\cdots a\ldots b\cdots}_{\cdots\cdots\cdots\cdots}\pi^{\cdots\cdots\cdots\cdots}_{\cdots a\ldots b\cdots}=\delta^{\cdots a\ldots b\cdots}_{\cdots\cdots\cdots\cdots}\pi^{\cdots\cdots\cdots\cdots}_{\cdots b\ldots a\cdots}=-\delta^{\cdots b\ldots a\cdots}_{\cdots\cdots\cdots\cdots}\pi^{\cdots\cdots\cdots\cdots}_{\cdots b\ldots a\cdots}=-\delta^{\cdots a\ldots b\cdots}_{\cdots\cdots\cdots\cdots}\pi^{\cdots\cdots\cdots\cdots}_{\cdots a\ldots b\cdots}=0.$$

From this it also obviously follows that the product of any gKd and any gpd with *two or more common contracted indices* vanishes.

If $A=[a^i_j]$ is an $n\times n$ matrix its *determinant* and *permanent* are

$$\det A=\frac{1}{n!}a^{i_1}_{j_1}\cdots a^{i_n}_{j_n}\delta^{j_1\cdots j_n}_{i_1\cdots i_n}, \qquad \operatorname{per} A=\frac{1}{n!}a^{i_1}_{j_1}\cdots a^{i_n}_{j_n}\pi^{j_1\cdots j_n}_{i_1\cdots i_n}.$$

III. gKd's AND gpd's IN SYMMETRIES OF TENSORS

The use of symmetrizing parentheses and antisymmetrizing brackets for specification of tensor symmetries can become both convoluted and ambiguous. For example $4T_{[(i|j|]k)}$ can be confusing at first sight; taken to mean performing symmetry in *i, k* and skew symmetry in *i, j*, it is still ambiguous, depending on which operation is performed first. It can be appreciated that symmetry specifications in this way for tensors with a large number of indices and intertwining brackets can be quite horrendous. Specification of tensor symmetries with *generalized deltas* (gd's—collectively gKd's and gpd's) gives clean, unambiguous expressions. Performing symmetry then antisymmetry and vice versa on $4T_{[(i|j|]k)}$ gives two quite different expressions:

$$2(T_{[ij]k}+T_{[kj]i})=(T_{ijk}+T_{kji}-T_{jik}-T_{jki})=\pi^{pn}_{ik}(T_{pjn}-T_{jpn})=\delta^{lm}_{pj}\pi^{pn}_{ik}T_{lmn},$$

$$2(T_{(i|j|k)}-T_{(j|i|k)})=(T_{ijk}-T_{jik}+T_{kji}-T_{kij})=\delta^{pm}_{ij}(T_{pmk}+T_{kmp})=\pi^{ln}_{pk}\delta^{pm}_{ij}T_{lmn}.$$

Specification of what one means by the gd's is unequivocal. They are also exactly what is needed for tensors obeying the symmetry of Young tableaux illustrated next.

The Riemann tensor R_{abcd} is a *Young tableau* (YT) *tensor* in its *algebraic* symmetries, expressible as[4,6] $R_{abcd}=\frac{1}{12}R_{\{ac,bd\}}$, obeying the partial symmetries and antisymmetries as determined by its tableau $\{ac,bd\}$. Fully expanded, it is[4]

$$\begin{aligned}R_{abcd}&=\tfrac{1}{12}R_{\{ac,bd\}}\\&=\tfrac{1}{12}[R_{abcd}+R_{adcb}+R_{cbad}+R_{cdab}-R_{abdc}-R_{acdb}-R_{dbac}-R_{dcab}\\&\quad-R_{bacd}-R_{bdca}-R_{cabd}-R_{cdba}+R_{badc}+R_{bcda}+R_{dabc}+R_{dcba}].\end{aligned}$$

Its symmetries are inbuilt. It is easy to see the skewsymmetry in (a,b) and in (c,d). Then too the interchange $(a,b)\Leftrightarrow(c,d)$ can also be seen, while the cyclic symmetry $R_{abcd}+R_{acdb}+R_{adbc}=0$, though tedious, is easy enough to verify. Conversely, using these symmetries in the rhs does indeed produce the lhs.

The above expression can be put into a convenient form involving gKd's and gpd's. Let E_{abcd} be the expression on the rhs within brackets:

$$\begin{aligned}E_{abcd}&=R_{abcd}+R_{cbad}+R_{adcb}+R_{cdab}-R_{abdc}-R_{dbac}-R_{acdb}-R_{dcab}\\&\quad-R_{bacd}-R_{cabd}-R_{bdca}-R_{cdba}+R_{badc}+R_{dabc}+R_{bcda}+R_{dcba}\\&=\delta^{pq}_{ab}[R_{pqcd}+R_{cqpd}+R_{pdcq}+R_{cdpq}-R_{pqdc}-R_{dqpc}-R_{pcdq}-R_{dcpq}]\\&=\delta^{pq}_{ab}\delta^{rs}_{cd}[R_{pqrs}+R_{rqps}+R_{psrq}+R_{rspq}]\\&=\delta^{pq}_{ab}\delta^{rs}_{cd}\pi^{eg}_{pr}[R_{eqgs}+R_{esgq}]\\&=\delta^{pq}_{ab}\delta^{rs}_{cd}\pi^{eg}_{pr}\pi^{fh}_{qs}R_{efgh},\end{aligned}$$

so that

$$R_{abcd}=\tfrac{1}{12}\delta^{pq}_{ab}\delta^{rs}_{cd}\pi^{eg}_{pr}\pi^{fh}_{qs}R_{efgh}$$

and there are no intertwined, or indeed any parentheses or brackets.

As for recognition of symmetries from this, it is evident that because of the gKd's the expression is skew in (a,b) and in (c,d). With $(a,b)\Leftrightarrow(c,d)$ and slight index manipulation it is also visible that the expression is symmetric. The cyclic identity is not obvious from either the 16 term expression or its compacted gd equivalent. But it is here that the PAPS reference tables come into play.

Observing the gd factor $\delta^{pq}_{ab}\delta^{rs}_{cd}\pi^{eg}_{pr}\pi^{fh}_{gs}$ above and formula **m3** from the table in PAPS Appendix C 7.3.2, i.e., $\delta^{bcd}_{\dots}\delta^{pq}_{.b}\delta^{rs}_{cd}\pi^{..}_{pr}\pi^{..}_{qs}=0$ (dots allow any free indices), we can immediately write down the symmetry/condition as $\delta^{bcd}_{ijk}R_{abcd}=0$. But let us show it in reverse (in fact deriving the **m3** identity).

Multiply the above relation for R_{abcd} by δ^{bcd}_{ijk} to get

$$\begin{aligned}\delta^{bcd}_{ijk}R_{abcd}&=\tfrac{1}{12}\delta^{bcd}_{ijk}\delta^{pq}_{ab}\delta^{rs}_{cd}\pi^{eg}_{pr}\pi^{fh}_{qs}R_{efgh}\\&=\tfrac{1}{6}\delta^{brs}_{ijk}\delta^{pq}_{ab}\pi^{eg}_{pr}\pi^{fh}_{qs}R_{efgh}\quad(\text{or see PAPS-formula }\mathbf{k8}\text{ or, better, Eq. }(4))\\&=\tfrac{1}{6}[\delta^{qrs}_{ijk}\delta^{p}_{a}-\delta^{prs}_{ijk}\delta^{q}_{a}]\pi^{eg}_{pr}\pi^{fh}_{qs}R_{efgh}\\&=0\qquad(\text{two repeated indices for a gKd and a gpd}).\end{aligned}$$

Expanding the lhs (six terms), but using the antisymmetry in the last two indices, produces the cyclic identity. This reverse derivation relied on ''already knowing'' the cyclic identity of the Riemann tensor (by applying δ^{bcd}_{ijk}).

If instead we have the tensor

$$\begin{aligned} U_{abcd} &= T_{abcd} - T_{cbad} + T_{dbac} - T_{dbca} + T_{cbda} - T_{abdc} \\ &\quad + T_{bacd} - T_{bcad} + T_{bdac} - T_{bdca} + T_{bcda} - T_{badc} \\ &= \delta^{pgh}_{acd}(T_{pbgh} + T_{bpgh}) \\ &= \delta^{pgh}_{acd}\pi^{ef}_{pb}T_{efgh} \end{aligned}$$

one may, with some scrutiny, discern the antisymmetry in indices (a,c) and in (a,d) (and hence total antisymmetry in all three of these indices) from the expanded expression. The antisymmetry in (a,c,d) is immediately obvious from the compacted gd expression. However, there is another ''hidden'' symmetry, unobvious from either expression for U_{abcd}. It is

$$U_{abcd} - U_{bcda} + U_{cdab} - U_{dabc} = 0.$$

But how can this be discovered? Here, again, inspection of the identity **o9** in the same table in PAPS, i.e., $\delta^{efgh}_{....}\delta^{p..}_{efg}\pi^{..}_{hp} = 0$, provides the answer. Adapted to our indices by $(e,f,g,h) \rightarrow (a,c,d,b)$ we can write (sign changes do not matter here)

$$\delta^{abcd}_{ijkl}\delta^{pgh}_{acd}\pi^{ef}_{pb}T_{efgh} = 0.$$

Thus the ''hidden'' identity is $\delta^{abcd}_{ijkl}U_{abcd} = 0$. One may wish to leave it like this; however, expansion of it, and using antisymmetry in first, third, and fourth indices produces the 4-term alternating cyclic identity above (with subscripts i,j,k,l). Further, we also remark that while the antisymmetries can be encompassed within the bracketed notation, $U_{abcd} = U_{[a|b|cd]}$, what can be suggested to accommodate the hidden symmetry? Clearly the gd notation to express and ''discover'' symmetries of tensors seems superior.

The simpler Lanczos tensor[7,8] satisfies algebraic (nondifferential, or not involving covariant differentiation) relations $L_{ijk} = -L_{jik}$, $L_{ijk} + L_{jki} + L_{kij} = 0$, and $L_{ij}{}^{j} = 0$ (optional algebraic gauge condition, often accepted). In any case, it is the symmetries of the (free) 3-index object that interests us. These first two symmetries qualify it as a YT tensor, expressible as[4] $L_{ijk} = \frac{1}{3}L_{\{ik,j\}}$. This again is a one-line expression for the tensor, encompassing the first two symmetries. Fully expanded, we have

$$\begin{aligned} L_{ijk} = \tfrac{1}{3}L_{\{ik,j\}} &= \tfrac{1}{3}[L_{ijk} - L_{jik} + L_{kji} - L_{kij}] \\ &= \tfrac{1}{3}\delta^{pm}_{ij}[L_{pmk} + L_{kmp}] \\ &= \tfrac{1}{3}\delta^{pm}_{ij}\pi^{ln}_{kp}L_{lmn}. \end{aligned}$$

It is then easy enough to verify the skewsymmetry and the cyclic symmetry (multiply by δ^{ijk}_{abc}) from this equation—and in reverse; that with these symmetries, employed on the rhs, we do get the lhs. The cyclic relation for the Lanczos tensor gives rise to the identity $\delta^{ijk}_{...}\delta^{p.}_{ij}\pi^{..}_{kp} = 0$, recorded as identity **o2** in the PAPS reference document.

Other than determining symmetries of tensors, the use of gd's may possibly help in ''seeing'' the number of independent components of a tensor with various symmetries.

Some computer algebra packages allow for symmetrization and antisymmetrization of indices. It is suggested that perhaps *specific* gpd and gKd objects (over and above a single Kronecker delta) be constructed in these packages allowing easy expansions of products of them over summed indices. This would be most useful for checking identities and formulas of various sorts.

IV. SPINOR AND TENSOR EQUIVALENTS OF gKd's AND gpd's AND EXAMPLES

A. Spinor equivalents of tensors

It is quite easy to construct spinor equivalents of the gKd's and gpd's from the basic spinor equivalent of the Kronecker δ, i.e., $\delta^{a}_{b} \Leftrightarrow \delta^{A}_{B}\delta^{A'}_{B'}$, and the use of determinants and permanents. This leads to a tabulation of spinor equivalents of individual gKd's and gpd's in Appendix A. The derivation of such formulas can be done by sight and by simple manipulations.

Apart from this table it is also our purpose to demonstrate in a couple of examples that use of the gd's can facilitate extremely quick derivation of spinor equivalents.

In general, if one expresses the symmetries of a tensor using gd's then (using the table of Appendix A) we can write down their spinor equivalents, replacing all tensor indices by corresponding spinor equivalent ones. Once this is done the spinor equivalent is technically found. However, it may involve several relations and consequent algebraic manipulations in order to get decompositions, such as the Weyl and Ricci parts for the Riemann tensor equivalent.

The two illustrations of obtaining the spinor equivalents of 2-index permanent symmetric and skewsymmetric tensors follow.

First, we determine the spinor equivalent of the skewsymmetric tensor $F_{ab}=-F_{ba}=F_{[ab]}$. This can be written as $F_{ab}=\frac{1}{2}F_{cd}\delta^{cd}_{ab}$ in terms of a gKd. Then taking the spinor equivalent (see Δ_2 in Appendix A) gives

$$\begin{aligned} F_{ABA'B'} &= \tfrac{1}{4}F_{CDC'D'}[\delta^{CD}_{AB}\pi^{C'D'}_{A'B'}+\pi^{CD}_{AB}\delta^{C'D'}_{A'B'}] \\ &= -F_{BAB'A'}=F_{[AB](A'B')}+F_{(AB)[A'B']}=\epsilon_{AB}\psi_{A'B'}+\epsilon_{A'B'}\phi_{AB}, \end{aligned}$$

where $\psi_{A'B'}=\frac{1}{2}F_{X}{}^{X}{}_{(A'B')}$ and $\phi_{AB}=\frac{1}{2}F_{(AB)X'}{}^{X'}$. If F_{ab} is *real*, $F_{ab}=\overline{F}_{ab}$, then $\psi_{A'B'}=\overline{\phi}_{A'B'}=\overline{\phi_{AB}}$, so that the spinor equivalent of the real skew-tensor (Maxwell tensor) F_{ab} is

$$F_{ab}=F_{[ab]}\Leftrightarrow\epsilon_{AB}\overline{\phi}_{A'B'}+\epsilon_{A'B'}\phi_{AB}.$$

Second, we determine the spinor equivalent of a permanent symmetric (or just symmetric by common usage if the context is clear) tensor $T_{ab}=T_{ba}=T_{(ab)}$. This can be written $T_{ab}=\frac{1}{2}T_{cd}\pi^{cd}_{ab}$ in terms of a gpd. Then taking the spinor equivalent (see Π_2 in Appendix A) gives

$$\begin{aligned} T_{ABA'B'} &= \tfrac{1}{4}T_{CDC'D'}[\delta^{CD}_{AB}\delta^{C'D'}_{A'B'}+\pi^{CD}_{AB}\pi^{C'D'}_{A'B'}] \\ &= T_{BAB'A'} \\ &= T_{[AB][A'B']}+T_{(AB)(A'B')} \\ &= \tfrac{1}{4}\epsilon_{AB}\epsilon_{A'B'}T+T_{(AB)(A'B')}, \end{aligned}$$

where $T=T_{X}{}^{X}{}_{X'}{}^{X'}=T_{XX'}{}^{XX'}$, which is also the *trace* of the tensor $T_{a}{}^{a}$. Thus

$$T_{ab}=T_{(ab)}\Leftrightarrow\tfrac{1}{4}\epsilon_{AB}\epsilon_{A'B'}T+T_{(AB)(A'B')}.$$

If one writes R in place of T, then the spinor equivalent of the trace-free Ricci tensor is

$$R_{ab}-\tfrac{1}{4}g_{ab}R\Leftrightarrow R_{(AB)(A'B')}.$$

B. Tensor equivalents of spinors

The tensor equivalents of gd spinors in combination are elegantly derived in PAPS. The results, along with related formulas are accumulated in the table in Appendix B.

We feel it worth mentioning at least one formula, one that enables us to easily remember the usual rather complicated formula for the spinor equivalent of the alternating tensor—by using gKd/gpd spinors. Abbreviate by putting

$$\Delta=\delta^{AB}_{CD}\qquad\Pi=\pi^{AB}_{CD}$$
$$\Delta'=\delta^{A'B'}_{C'D'}\quad\Pi'=\pi^{A'B'}_{C'D'}.$$

The formula, with the understanding that the first relation is a symbolic equivalence, is (see full derivation in PAPS)

$$\frac{i}{2}\,\delta^{\Delta\Pi}_{\Delta'\Pi'}\equiv\frac{i}{2}(\Delta\Pi'-\Pi\Delta')=\frac{i}{2}(\delta^{AB}_{CD}\pi^{A'B'}_{C'D'}-\pi^{AB}_{CD}\delta^{A'B'}_{C'D'})\Leftrightarrow\epsilon^{ab}{}_{cd}\,.$$

To give an example from the table, consider the tensor equivalent of the spinor (last but one table entry in Appendix B)

$$\delta^A_C\delta^B_D\delta^{A'}_{D'}\delta^{B'}_{C'}\Leftrightarrow\tfrac{1}{2}(\pi^{ab}_{cd}-g^{ab}g_{cd}-i\epsilon^{ab}{}_{cd}).$$

Multiply the lhs by $H_{ABA'B'}$ and the rhs by its tensor equivalent H_{ab} and one immediately obtains (see Ref. 6, p. 153)

$$H_{CDD'C'}\Leftrightarrow\tfrac{1}{2}(\pi^{ab}_{cd}-g^{ab}g_{cd}-i\epsilon^{ab}{}_{cd})H_{ab}=\tfrac{1}{2}(H_{cd}+H_{dc}-g_{cd}H_a{}^a-i\epsilon_{abcd}H^{ab}).$$

Now take the conjugate of the spinor/tensor equivalent relation, to get $\delta^A_D\delta^B_C\delta^{A'}_{C'}\delta^{B'}_{D'}\Leftrightarrow\tfrac{1}{2}(\pi^{ab}_{cd}-g^{ab}g_{cd}+i\epsilon^{ab}{}_{cd}$, multiply as before by $H_{ABA'B'}$ and its equivalent H_{ab}, and arrive at the *different* tensor equivalent (with unprimed indices interchanged)

$$H_{DCC'D'}\Leftrightarrow\tfrac{1}{2}(H_{cd}+H_{dc}-g_{cd}H_a{}^a+i\epsilon_{abcd}H^{ab}).$$

APPENDIX A: SPINOR EQUIVALENTS FOR THE gKd AND gpd

In the reference formulas below it should be mentioned that there is a good deal of interplay between the gKd and gpd in such specifications, there being a variety of ways to express expansions of some gKd's, gpd's (and also their spinor equivalents).

1. Spinor equivalents of the gKd's

$$\Delta_1=\delta^a_b\Leftrightarrow\delta^A_B\delta^{A'}_{B'}\,,$$

$$\Delta_2=\delta^{ab}_{cd}=\begin{vmatrix}\delta^a_c & \delta^a_d\\ \delta^b_c & \delta^b_d\end{vmatrix}\Leftrightarrow\begin{vmatrix}\delta^A_C\delta^{A'}_{C'} & \delta^A_D\delta^{A'}_{D'}\\ \delta^B_C\delta^{B'}_{C'} & \delta^B_D\delta^{B'}_{D'}\end{vmatrix}=\tfrac{1}{2}[\delta^{AB}_{CD}\pi^{A'B'}_{C'D'}+\pi^{AB}_{CD}\delta^{A'B'}_{C'D'}],$$

$$\Delta_3=\delta^{abc}_{def}=\begin{vmatrix}\delta^a_d & \delta^a_e & \delta^a_f\\ \delta^b_d & \delta^b_e & \delta^b_f\\ \delta^c_d & \delta^c_e & \delta^c_f\end{vmatrix}$$

$$=\delta^a_d\delta^{bc}_{ef}-\delta^a_e\delta^{bc}_{df}+\delta^a_f\delta^{bc}_{de}\Leftrightarrow\tfrac{1}{2}[\delta^A_D\delta^{A'}_{D'}(\delta^{BC}_{EF}\pi^{B'C'}_{E'F'}+\pi^{BC}_{EF}\delta^{B'C'}_{E'F'})$$

$$-\delta^A_E\delta^{A'}_{E'}(\delta^{BC}_{DF}\pi^{B'C'}_{D'F'}+\pi^{BC}_{DF}\delta^{B'C'}_{D'F'})+\delta^A_F\delta^{A'}_{F'}(\delta^{BC}_{DE}\pi^{B'C'}_{D'E'}+\pi^{BC}_{DE}\delta^{B'C'}_{D'E'})].$$

We can expand the gKd with 4 (upper/lower) indices by a Laplace expansion of its first two rows and complementary minors. The result is

$$\Delta_4=\delta^{abcd}_{efgh}=\begin{vmatrix}\delta^a_e & \delta^a_f & \delta^a_g & \delta^a_h\\ \delta^b_e & \delta^b_f & \delta^b_g & \delta^b_h\\ \delta^c_e & \delta^c_f & \delta^c_g & \delta^c_h\\ \delta^d_e & \delta^d_f & \delta^d_g & \delta^d_h\end{vmatrix}$$

$$=\delta^{ab}_{ef}\delta^{cd}_{gh}-\delta^{ab}_{eg}\delta^{cd}_{fh}+\delta^{ab}_{eh}\delta^{cd}_{fg}+\delta^{ab}_{fg}\delta^{cd}_{eh}-\delta^{ab}_{fh}\delta^{cd}_{eg}+gd^{ab}_{gh}\delta^{cd}_{ef}.$$

The rhs can be written as the sum/difference of *permanents* of gKd's, if desired

$$\left\|\begin{matrix}\delta^{ab}_{ef} & \delta^{ab}_{gh}\\ \delta^{cd}_{ef} & \delta^{cd}_{gh}\end{matrix}\right\|-\left\|\begin{matrix}\delta^{ab}_{eg} & \delta^{ab}_{fh}\\ \delta^{cd}_{eg} & \delta^{cd}_{fh}\end{matrix}\right\|+\left\|\begin{matrix}\delta^{ab}_{eh} & \delta^{ab}_{fg}\\ \delta^{cd}_{eh} & \delta^{cd}_{fg}\end{matrix}\right\|.$$

The spinor equivalent of δ^{abcd}_{efgh} is

$$\Delta_4\Leftrightarrow\tfrac{1}{4}[(\delta^{AB}_{EF}\pi^{A'B'}_{E'F'}+\pi^{AB}_{EF}\delta^{A'B'}_{E'F'})(\delta^{CD}_{GH}\pi^{C'D'}_{G'H'}+\pi^{CD}_{GH}\delta^{C'D'}_{G'H'})-(\delta^{AB}_{EG}\pi^{A'B'}_{E'G'}+\pi^{AB}_{EG}\delta^{A'B'}_{E'G'})$$
$$\times(\delta^{CD}_{FH}\pi^{C'D'}_{F'H'}+\pi^{CD}_{FH}\delta^{C'D'}_{F'H'})+(\delta^{AB}_{EH}\pi^{A'B'}_{E'H'}+\pi^{AB}_{EH}\delta^{A'B'}_{E'H'})(\delta^{CD}_{FG}\pi^{C'D'}_{F'G'}+\pi^{CD}_{FG}\delta^{C'D'}_{F'G'})$$
$$+(\delta^{AB}_{FG}\pi^{A'B'}_{F'G'}+\pi^{AB}_{FG}\delta^{A'B'}_{F'G'})(\delta^{CD}_{EH}\pi^{C'D'}_{E'H'}+\pi^{CD}_{EH}\delta^{C'D'}_{E'H'})-(\delta^{AB}_{FH}\pi^{A'B'}_{F'H'}+\pi^{AB}_{FH}\delta^{A'B'}_{F'H'})$$
$$\times(\delta^{CD}_{EG}\pi^{C'D'}_{E'G'}+\pi^{CD}_{EG}\delta^{C'D'}_{E'G'})+(\delta^{AB}_{GH}\pi^{A'B'}_{G'H'}+\pi^{AB}_{GH}\delta^{A'B'}_{G'H'})(\delta^{CD}_{EF}\pi^{C'D'}_{E'F'}+\pi^{CD}_{EF}\delta^{C'D'}_{E'F'})].$$

2. Spinor equivalents of the gpd's

$$\Pi_1=\pi^a_b=\delta^a_b\Leftrightarrow\delta^A_B\delta^{A'}_{B'},$$

$$\Pi_2=\pi^{ab}_{cd}=\left\|\begin{matrix}\delta^a_c & \delta^a_d\\ \delta^b_c & \delta^b_d\end{matrix}\right\|\Leftrightarrow\left\|\begin{matrix}\delta^A_C\delta^{A'}_{C'} & \delta^A_D\delta^{A'}_{D'}\\ \delta^B_C\delta^{B'}_{C'} & \delta^B_D\delta^{B'}_{D'}\end{matrix}\right\|=\tfrac{1}{2}[\delta^{AB}_{CD}\delta^{A'B'}_{C'D'}+\pi^{AB}_{CD}\pi^{A'B'}_{C'D'}],$$

$$\Pi_3=\pi^{abc}_{def}=\left\|\begin{matrix}\delta^a_d & \delta^a_e & \delta^a_f\\ \delta^b_d & \delta^b_e & \delta^b_f\\ \delta^c_d & \delta^c_e & \delta^c_f\end{matrix}\right\|$$

$$=\delta^a_d\pi^{bc}_{ef}+\delta^a_e\pi^{bc}_{df}+\delta^a_f\pi^{bc}_{de}\Leftrightarrow\tfrac{1}{2}[\delta^A_D\delta^{A'}_{D'}(\delta^{BC}_{EF}\delta^{B'C'}_{E'F'}+\pi^{BC}_{EF}\pi^{B'C'}_{E'F'})$$
$$+\delta^A_E\delta^{A'}_{E'}(\delta^{BC}_{DF}\delta^{B'C'}_{D'F'}+\pi^{BC}_{DF}\pi^{B'C'}_{D'F'})+\delta^A_F\delta^{A'}_{F'}(\delta^{BC}_{DE}\delta^{B'C'}_{D'E'}+\pi^{BC}_{DE}\pi^{B'C'}_{D'E'})].$$

Since permanents only involve positive signs, the following Laplace expansion is also clear:

$$\Pi_4=\pi^{abcd}_{efgh}=\left\|\begin{matrix}\delta^a_e & \delta^a_f & \delta^a_g & \delta^a_h\\ \delta^b_e & \delta^b_f & \delta^b_g & \delta^b_h\\ \delta^c_e & \delta^c_f & \delta^c_g & \delta^c_h\\ \delta^d_e & \delta^d_f & \delta^d_g & \delta^d_h\end{matrix}\right\|$$

$$=\pi^{ab}_{ef}\pi^{cd}_{gh}+\pi^{ab}_{eg}\pi^{cd}_{fh}+\pi^{ab}_{eh}\pi^{cd}_{fg}+\pi^{ab}_{fg}\pi^{cd}_{eh}+\pi^{ab}_{fh}\pi^{cd}_{eg}+\pi^{ab}_{gh}\pi^{cd}_{ef}.$$

The rhs can be written as the sum of permanents of permanents, if desired

$$\left\|\begin{matrix}\pi^{ab}_{ef} & \pi^{ab}_{gh}\\ \pi^{cd}_{ef} & \pi^{cd}_{gh}\end{matrix}\right\| + \left\|\begin{matrix}\pi^{ab}_{eg} & \pi^{ab}_{fh}\\ \pi^{cd}_{eg} & \pi^{cd}_{fh}\end{matrix}\right\| + \left\|\begin{matrix}\pi^{ab}_{eh} & \pi^{ab}_{fg}\\ \pi^{cd}_{eh} & \pi^{cd}_{fg}\end{matrix}\right\|.$$

The spinor equivalent of π^{abcd}_{efgh} is

$$\begin{aligned}\Pi_4 \Leftrightarrow \tfrac{1}{4}[&(\delta^{AB}_{EF}\delta^{A'B'}_{E'F'}+\pi^{AB}_{EF}\pi^{A'B'}_{E'F'})(\delta^{CD}_{GH}\delta^{C'D'}_{G'H'}+\pi^{CD}_{GH}\pi^{C'D'}_{G'H'})+(\delta^{AB}_{EG}\delta^{A'B'}_{E'G'}+\pi^{AB}_{EG}\pi^{A'B'}_{E'G'})\\ &\times(\delta^{CD}_{FH}\delta^{C'D'}_{F'H'}+\pi^{CD}_{FH}\pi^{C'D'}_{F'H'})+(\delta^{AB}_{EH}\delta^{A'B'}_{E'H'}+\pi^{AB}_{EH}\pi^{A'B'}_{E'H'})(\delta^{CD}_{FG}\delta^{C'D'}_{F'G'}+\pi^{CD}_{FG}\pi^{C'D'}_{F'G'})\\ &+(\delta^{AB}_{FG}\delta^{A'B'}_{F'G'}+\pi^{AB}_{FG}\pi^{A'B'}_{F'G'})(\delta^{CD}_{EH}\delta^{C'D'}_{E'H'}+\pi^{CD}_{EH}\pi^{C'D'}_{E'H'})+(\delta^{AB}_{FH}\delta^{A'B'}_{F'H'}+\pi^{AB}_{FH}\pi^{A'B'}_{F'H'})\\ &\times(\delta^{CD}_{EG}\delta^{C'D'}_{E'G'}+\pi^{CD}_{EG}\pi^{C'D'}_{E'G'})+(\delta^{AB}_{GH}\delta^{A'B'}_{G'H'}+\pi^{AB}_{GH}\pi^{A'B'}_{G'H'})(\delta^{CD}_{EF}\delta^{C'D'}_{E'F'}+\pi^{CD}_{EF}\pi^{C'D'}_{E'F'})].\end{aligned}$$

APPENDIX B: SPINOR⇔TENSOR EQUIVALENTS

The mixed mode spinor appears with its covariant form below it.

Spinor	Tensor
δ^A_B ϵ_{AB}	
$\delta^A_B\delta^{A'}_{B'}$ $\epsilon_{AB}\epsilon_{A'B'}$	δ^a_b g_{ab}
$\delta^{AB}_{CD}=\Delta,\ \delta^{A'B'}_{C'D'}$ $\epsilon_{AB}\epsilon_{CD}=\epsilon_{AC}\epsilon_{BD}-\epsilon_{AD}\epsilon_{BC}$	
$\pi^{AB}_{CD}=\Pi,\ \pi^{A'B'}_{C'D'}=\Pi'$ $\epsilon_{AC}\epsilon_{BD}+\epsilon_{AD}\epsilon_{BC}$	
$\frac{1}{2}(\delta^{AB}_{CD}\pi^{A'B'}_{C'D'}+\pi^{AB}_{CD}\delta^{A'B'}_{C'D'})=\frac{1}{2}(\Delta\Pi'+\Pi\Delta')$ $\epsilon_{AC}\epsilon_{BD}\epsilon_{A'C'}\epsilon_{B'D'}-\epsilon_{AD}\epsilon_{BC}\epsilon_{A'D'}\epsilon_{B'C'}$	δ^{ab}_{cd} $g_{ac}g_{bd}-g_{ad}g_{bc}$
$\frac{1}{2}(\delta^{AB}_{CD}\delta^{A'B'}_{C'D'}+\pi^{AB}_{CD}\pi^{A'B'}_{C'D'})=\frac{1}{2}(\Delta\Delta'+\Pi\Pi')$ $\epsilon_{AC}\epsilon_{BD}\epsilon_{A'C'}\epsilon_{B'D'}+\epsilon_{AD}\epsilon_{BC}\epsilon_{A'D'}\epsilon_{B'C'}$	π^{ab}_{cd} $g_{ac}g_{bd}+g_{ad}g_{bc}$
$i(\delta^A_C\delta^B_D\delta^{A'}_{D'}\delta^{B'}_{C'}-\delta^A_D\delta^B_C\delta^{A'}_{C'}\delta^{B'}_{D'})=\frac{i}{2}\delta^{\Delta\Pi}_{\Delta'\Pi'}$ $i(\epsilon_{AC}\epsilon_{BD}\epsilon_{A'D'}\epsilon_{B'C'}-\epsilon_{AD}\epsilon_{BC}\epsilon_{A'C'}\epsilon_{B'D'})$	$\epsilon_{cd}{}^{ab}=\epsilon^{ab}{}_{cd}$ ϵ_{abcd}
$\delta^{AB}_{CD}\delta^{A'B'}_{C'D'}=\Delta\Delta'$ $\epsilon_{AB}\epsilon_{CD}\epsilon_{A'B'}\epsilon_{C'D'}$	$g^{ab}g_{cd}$ $g_{ab}g_{cd}$
$\delta^{AB}_{CD}\pi^{A'B'}_{C'D'}=\Delta\Pi'$ $\epsilon_{AC}\epsilon_{BD}\epsilon_{A'C'}\epsilon_{B'D'}-\epsilon_{AD}\epsilon_{BC}\epsilon_{A'D'}\epsilon_{B'C'}$ $+\epsilon_{AC}\epsilon_{BD}\epsilon_{A'D'}\epsilon_{B'C'}-\epsilon_{AD}\epsilon_{BC}\epsilon_{A'C'}\epsilon_{B'D'}$ $=\epsilon_{AB}\epsilon_{CD}(\epsilon_{A'C'}\epsilon_{B'D'}+\epsilon_{A'D'}\epsilon_{B'C'})$	$\delta^{ab}_{cd}-i\epsilon^{ab}{}_{cd}$ $g_{ac}g_{bd}-g_{ad}g_{bc}-i\epsilon_{abcd}$
$\pi^{AB}_{CD}\delta^{A'B'}_{C'D'}=\Pi\Delta'$ $\epsilon_{AC}\epsilon_{BD}\epsilon_{A'C'}\epsilon_{B'D'}-\epsilon_{AD}\epsilon_{BC}\epsilon_{A'D'}\epsilon_{B'C'}$ $-\epsilon_{AC}\epsilon_{BD}\epsilon_{A'D'}\epsilon_{B'C'}+\epsilon_{AD}\epsilon_{BC}\epsilon_{A'C'}\epsilon_{B'D'}$ $=(\epsilon_{AC}\epsilon_{BD}+\epsilon_{AD}\epsilon_{BC})\epsilon_{A'B'}\epsilon_{C'D'}$	$\delta^{ab}_{cd}+i\epsilon^{ab}{}_{cd}$ $g_{ac}g_{bd}-g_{ad}g_{bc}+i\epsilon_{abcd}$
$\pi^{AB}_{CD}\pi^{A'B'}_{C'D'}=\Pi\Pi'$ $(\epsilon_{AC}\epsilon_{BD}+\epsilon_{AD}\epsilon_{BC})\times(\epsilon_{A'C'}\epsilon_{B'D'}+\epsilon_{A'D'}\epsilon_{B'C'})$	$2\pi^{ab}_{cd}-g^{ab}g_{cd}$ $2g_{ac}g_{bd}+2g_{ad}g_{bc}-g_{ab}g_{cd}$
$\delta^A_C\delta^B_D\delta^{A'}_{D'}\delta^{B'}_{C'}$ $\epsilon_{AC}\epsilon_{BD}\epsilon_{A'D'}\epsilon_{B'C'}$	$\frac{1}{2}(\pi^{ab}_{cd}-g^{ab}g_{cd}-i\epsilon^{ab}{}_{cd})$ $\frac{1}{2}(g_{ac}g_{bd}+g_{ad}g_{bc}-g_{ab}g_{cd}-i\epsilon_{abcd})$

[1] T. Muir, *The Theory of Determinants in the Historical Order of Development* (Macmillan and Co., Limited, London, 1906), Vol. 1, p. 94.

[2] C. M. Cramlet, *Invariant tensors and their application to the study of determinants and allied tensor functions*, Ph.D. thesis, University of Washington (1926).

[3] C. M. Cramlet, ''Applications of the determinant and permanent tensors to determinants of general class and allied tensor functions,'' Am. J. Math. **49**, 87–96 (1927).

[4] R. L. Agacy, *Generalized Kronecker, permanent delta and Young tableaux applications to tensors and spinors; Lanczos–Zund spinor classification and general spinor factorizations*, Ph.D. thesis, London University (1997). The expression $L_{ijk}=\frac{1}{3}L_{\{ik,j\}}$ is correctly stated on p. 28 of this reference, but incorrectly stated on p. 26 with $L_{\{ij,k\}}$ on the rhs.

[5] See AIP document No. PAPS JMAPAQ-Vol. 40-033903 for 33 pages of the document entitled ''Generalized Kronecker and Permanent deltas, their spinor and tensor equivalents—Reference Formulae.'' Order by PAPS number and journal reference from the American Institute of Physics, Physics Auxiliary Publications Service, 500 Sunnyside Boulevard, Woodbury, NY 11797-2999. Fax: 516-576-2223, email: paps@aip.org. The price is $1.50 for each microfiche or $5.00 for photocopies of up to 30 pages, and $0.15 for each additional page over 30 pages. Airmail additional. Make checks payable to the American Institute of Physics.

[6] In this connection we remark that our definition for the algebraic symmetries of the Riemann tensor $R_{abcd}=\frac{1}{12}R_{\{ac,bd\}}$, agrees with the result in R. Penrose and W. Rindler, *Spinors and Space-time* (Cambridge University Press, England, 1984), Vol. 1, p. 144, $\frac{3}{4}R_{abcd}=R_{\overline{(ac)}\atop(\underline{bd})}$. This is because the symmetrization on two letters introduces a factor of 1/2, which together with 2 rows gives a factor of 1/4. Antisymmetrization of columns produces another factor of 1/4. Hence the rhs of the latter is 1/16 of our result, i.e., $\frac{3}{4}R_{abcd}=\frac{1}{16}R_{\{ac,bd\}}$, agreeing precisely with our definition.

[7] S. B. Edgar and A. Höglund, ''The Lanczos potential for the Weyl curvature tensor: existence, wave equation, and algorithms,'' Proc. R. Soc. London, Ser. A **453**, 835–851 (1997).

[8] P. Dolan and C. Kim, ''The wave equation for the Lanczos potential,'' Proc. R. Soc. London, Ser. A **447**, 557–575 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

The algebra of two symmetric matrices: Proving completeness and deriving syzygies for a set of invariants of the Riemann tensor

S. Bonanos[a)]
Institute of Nuclear Physics, N.C.S.R. DEMOKRITOS, Aghia Paraskevi, 15310 Greece

(Received 8 October 1998; accepted for publication 9 December 1998)

A large number of Riemann tensor invariants can be written as traces of products of two 3×3 matrices, representing the Weyl tensor and the Weyl-like square of the Ricci tensor. It is pointed out that finding a complete set, $\mathcal{I}$, for all of these invariants is a simple consequence of the more general problem of finding a complete set of symmetric *matrices*, $\mathcal{M}$, in terms of which all symmetric matrix polynomials in these two matrices can be expressed. Such a set is constructed and a formal proof of its completeness is given. Several matrix identities and a scalar syzygy, obtained recently by Sneddon, are rederived and their interrelationships clarified. They are shown to be, ultimately, consequences of the Cayley–Hamilton theorem. A ''minimal set'' of invariants, that must be contained in the complete set of invariants of the general problem, is identified and it is concluded that no set proposed so far is complete. © *1999 American Institute of Physics.* [S0022-2488(99)02503-7]

I. INTRODUCTION

The study of the invariants of the Riemann tensor was originally motivated by their obvious usefulness in the invariant characterization of space–time properties[1–4] and, thereby, in the invariant classification of gravitational fields.[5–8] Gradually, however, interest has focused on more mathematical questions regarding the ''complete set of polynomial curvature invariants.'' *Polynomial curvature invariants* are those that are expressible as contractions of arbitrary products of the curvature tensor R_{abcd}, the metric tensor g_{ab} and the totally antisymmetric tensor η_{abcd}, and are, therefore, polynomials in the curvature components. A *complete set* $\mathcal{I}=\{I_1,I_2,\ldots,I_n\}$ of invariants is one having the property that any polynomial invariant can be expressed as a polynomial in $I_1,\ldots,I_n$ and no invariant in the set can be so expressed in terms of the remaining I_i.[9,10] It is known that such a set will be redundant, i.e., it will contain more than the maximum number $(=14)$ of gauge-independent components of the Riemann tensor.[10] There will then exist $n-14$ polynomial identities (syzygies) relating these invariants.

The mathematical questions alluded to above are:

(i) *How does one decide which invariants to include in a complete set?*
(ii) *How does one prove that a certain set of invariants is complete?*
(iii) *How does one derive the syzygies that relate the invariants of a complete set?*

The answers to these questions are anything but obvious, as is indicated by the number of different sets of invariants that have been proposed,[5,6,11–15] some of which have later been shown to be deficient in some sense,[14,16] and by the lack of any attempt to prove completeness (in the sense used here) for any of them.

In this paper we shall answer these questions for a restricted set of Riemann tensor invariants (to be precisely defined in Sec. II). The essential idea is to consider these invariants as traces of products of a pair of symmetric matrices and obtain the desired result as a corollary of the solution of the more general problem of finding a complete set of *matrices*, $\mathcal{M}$, in terms of which all such

[a)]Electronic mail: sbonano@mail.ariadne-t.gr

symmetric matrix products can be expressed. (Sneddon[17] has also considered the invariants built out of these two matrices.) This more general problem is actually easier to handle because of the Cayley–Hamilton theorem which implies polynomial identities between matrices. In fact, all matrix identities obtained here are consequences of this theorem. In a recent paper,[18] Sneddon gives a reference to work by Spencer and Rivlin[19,20] who apparently solved this problem (in even greater generality) 40 years ago! The proof of completeness given here, however, is different and conceptually simpler. It also leads to a more efficient computer algebra algorithm for simplification of products of matrix polynomials.

In Sec. II we define the set of scalars we will be considering and formulate the problem in terms of traces of products of two symmetric matrices. In Sec. III we systematically construct a set of matrices, $\mathcal{M}$, (and a corresponding set of scalars, $\mathcal{I}$) and prove that it is complete. In Sec. IV we derive a matrix identity satisfied by the matrices of the complete set which leads to a scalar identity (syzygy) between the invariants of the complete set $\mathcal{I}$. This syzygy was first obtained by Sneddon[17] in a rather tortuous manner. We also show that the nine symmetric matrix identities obtained recently by Sneddon[18] are direct consequences of the first one. In Sec. V we show that a wider class of scalars than originally considered can actually be expressed in terms of the complete set obtained. Finally, in Sec. VI we extend our set to a ''minimal set'' that must be contained in the complete set of invariants of the Riemann tensor, and conclude that none of the sets proposed so far can be complete. In particular neither the set of Carminati and McLenaghan[14] nor the set of Zakhary and McIntosh[15] are complete, even though the latter has been shown to yield the required number of invariants for every combination of Petrov and Segre types.

II. MOTIVATION AND FORMULATION OF THE PROBLEM

In the Newman–Penrose formalism the Riemann tensor is described by the (complex) Weyl spinor Ψ_{ABCD}, the traceless Ricci spinor $\Phi_{ABC'D'}$ and the scalar curvature R. The polynomial curvature invariants defined in the introduction can, therefore, be equivalently defined as those that are expressible as complete contractions (with ε^{AB}, $\varepsilon^{A'B'}$) of arbitrary products of the spinors Ψ_{ABCD} and $\Phi_{ABC'D'}$ together with the scalar R.

The 4-index spinors $\Psi_{AB}{}^{CD}$ and $\Phi_{AB}{}^{C'D'}$ can be represented as 3×3 matrices (Ψ symmetric traceless and Φ Hermitian) with respect to an orthonormal basis[21,22] for 2-index spinors as follows:

$$\Psi=\frac{1}{2}\begin{pmatrix} 2\Psi_2-\Psi_0-\Psi_4 & -i(\Psi_0-\Psi_4) & 2(\Psi_1-\Psi_3)\\ -i(\Psi_0-\Psi_4) & 2\Psi_2+\Psi_0+\Psi_4 & 2i(\Psi_1+\Psi_3)\\ 2(\Psi_1-\Psi_3) & 2i(\Psi_1+\Psi_3) & -4\Psi_2\end{pmatrix}, \tag{2.1}$$

$$\Phi=\frac{1}{2}\begin{pmatrix} \Phi_{00}+\Phi_{22}-\Phi_{02}-\Phi_{20} & -i(\Phi_{00}-\Phi_{22}+\Phi_{02}-\Phi_{20}) & 2(\Phi_{21}-\Phi_{01})\\ i(\Phi_{00}-\Phi_{22}-\Phi_{02}+\Phi_{20}) & \Phi_{00}+\Phi_{22}+\Phi_{02}+\Phi_{20} & -2i(\Phi_{21}+\Phi_{01})\\ 2(\Phi_{12}-\Phi_{10}) & 2i(\Phi_{12}+\Phi_{10}) & 4\Phi_{11}\end{pmatrix}. \tag{2.2}$$

Then contraction of two 4-index spinors over a pair of indices (which gives again a 4-index spinor!) corresponds to multiplication of the corresponding matrices, i.e., the matrix products $\Psi\Psi$ and $\Psi\Phi$ have components $\Psi_{AB}{}^{MN}\Psi_{MN}{}^{CD}$ and $\Psi_{AB}{}^{CD}\Phi_{CD}{}^{C'D'}$, respectively, arranged as in (2.1) (the traceless part of $\Psi\Psi$) and (2.2). [The form (2.2) for $\Psi\Phi$ refers to the arrangement of its spinor components only. Reality relations between these components, like those that make the Φ matrix Hermitian, are not implied. The Hermitian conjugate of $\Psi\Phi$ is the matrix $\Phi\bar{\Psi}$, and one must take the combinations $\Psi\Phi+\Phi\bar{\Psi}$ and $i(\Psi\Phi-\Phi\bar{\Psi})$ to obtain Hermitian matrices.]

To avoid the complication of having to distinguish primed from unprimed indices, we will consider in this paper only those scalars that can be formed by contraction of arbitrary products of the following two spinors: Ψ_{ABCD} and $\chi_{ABCD}=\Phi_{(AB}{}^{C'D'}\Phi_{CD)C'D'}$, the Weyl-like square of the Ricci spinor. In addition we will require that contractions between these spinors will always involve *pairs* of indices. (In Sec. V we show that this restriction can be weakened considerably.) Under these conditions, the set of scalars we will be considering can be defined as the set of traces of arbitrary products of the traceless 3×3 matrices Ψ and χ corresponding to ${\Psi_{AB}}^{CD}$ and ${\chi_{AB}}^{CD}$.

As stated in the Introduction, the question of finding a complete set for these traces follows immediately if we have solved the problem of finding a complete set of symmetric matrices in terms of which all symmetric products of two given symmetric matrices can be expressed. The required complete set will consist of the traces of the matrices in the complete set of matrices, together with any other scalars that were used in expressing higher order matrices in terms of those in the complete set. Thus the mathematical problem to be solved becomes:

Given two symmetric, traceless 3×3 matrices A and B, find a set of $\mathcal{M}$ of symmetric matrices such that every symmetric product of A's and B's can be written as a linear combination of the matrices in the set, with coefficients that are scalar combinations of the components of A and B.

Although the problem of finding a complete set of matrices seems, at first, more difficult than the corresponding one for scalars, it is actually easier because of the Cayley–Hamilton theorem, which gives polynomial identities between matrices. For the traceless matrices A and B the Cayley–Hamilton theorem reads

$$A^3 = i_a A + 2 j_a \mathbf{1}, \quad B^3 = i_b B + 2 j_b \mathbf{1}, \tag{2.3}$$

where $i_a = 1/2\,\mathrm{tr}(A^2)$, $2j_a = 1/3\,\mathrm{tr}(A^3) = \det A$ and similarly for i_b and j_b.

Now, if A, B, C are three 3×3 matrices, then evaluating the expression $(A+B+C)^3 - (-A+B+C)^3 - (A-B+C)^3 - (A+B-C)^3$ (i) by expanding each term and (ii) by applying the Cayley–Hamilton theorem to each term we obtain a very useful identity. When A, B, C are traceless, this identity takes the simple form (square brackets denote the trace of the enclosed matrix)

$$ABC + ACB + BCA + BAC + CAB + CBA = A[BC] + B[AC] + C[AB] + ([ABC] + [ACB])\mathbf{1}. \tag{2.4}$$

This identity can also be obtained[17] by antisymmetrizing the outer product of the 3 matrices and the identity over 4 indices. Our derivation shows that it is a simple consequence of the Cayley–Hamilton theorem.

III. SOLUTION OF THE PROBLEM

To build the complete set of matrices $\mathcal{M}$, we will first systematically construct all symmetric matrices of low order (=number of factors) in A and B and collect those that are independent. The symmetric matrices for each order $\leqslant 4$ (after using Eq. (2.3) to eliminate powers of A and B greater than 2) are:

order 1: A, B,
order 2: A^2, $AB+BA$, B^2,
order 3: A^2B+BA^2, ABA, BAB, B^2A+AB^2,
order 4: $A^2BA+ABA^2$, AB^2A, $A^2B^2+B^2A^2$, $ABAB+BABA$, BA^2B, $B^2AB+BAB^2$.

The matrices of order 1 and 2 are clearly independent. Now, putting $C=A$ and $C=B$ in Eq. (2.4) we obtain the identities:

$$A^2B + BA^2 + ABA = i_a B + sA + p\mathbf{1}, \tag{3.1}$$

$$B^2A + AB^2 + BAB = i_b A + sB + q\mathbf{1}, \tag{3.2}$$

where $s=[AB]$, $p=[A^2B]$, and $q=[AB^2]$. Thus, of the third order matrices only A^2B+BA^2 and B^2A+AB^2 need to be included in the complete set, Eqs. (3.1) and (3.2) giving expressions for ABA and BAB in terms of matrices in $\mathcal{M}$. Next, multiplying each of (3.1), (3.2) by A on the left and on the right and adding, and replacing B in (3.1) by the traceless matrix $B^2-(2/3)i_b\mathbf{1}$, we find that all symmetric 4th order matrices can be expressed in terms of $A^2B^2+B^2A^2$ and lower order matrices. Thus, the set[23] $\mathcal{M}$ (so far) includes the matrices

$$\mathcal{M}=\{\mathbf{1},\ A,\ B,\ A^2,\ AB+BA,\ B^2,\ A^2B+BA^2,\ B^2A+AB^2,\ A^2B^2+B^2A^2\}, \tag{3.3}$$

while the set of scalars $\mathcal{I}$ contains the traces of these matrices and of $A^3,\ B^3$,

$$\mathcal{I}=\{i_a,\ i_b,\ j_a,\ j_b,\ s,\ p,\ q,\ t\}, \tag{3.4}$$

where $i_a=1/2[A^2],\ i_b=1/2[B^2],\ j_a=1/6[A^3],\ j_b=1/6[B^3],\ s=[AB],\ p=[A^2B],\ q=[AB^2],\ t=[A^2B^2]$.

Proceeding in a similar manner we can show that all 10 symmetric matrices of fifth order in A, B can be written as linear combinations of the matrices in $\mathcal{M}$ with coefficients that are polynomials in $\mathcal{I}$. In fact, the fifth order matrices are expressible in terms of third or lower order matrices, as there are no first order scalars in $\mathcal{I}$. It thus appears that the set $\mathcal{M}$ may be the required complete set. To prove this we will also need the expressions for the 15 sixth order symmetric matrices. They can be obtained by appropriate multiplication by A and/or B of the lower order equations, or substitutions $A\rightarrow A^2-(2/3)i_a\mathbf{1},\ B\rightarrow B^2-(2/3)i_b\mathbf{1}$ in (3.1), (3.2). The resulting expressions for all 4th, 5th and 6th order symmetric matrices in terms of matrices in $\mathcal{M}$ with coefficients in $\mathcal{I}$ are given in the Appendix.

We are now in a position to formally prove completeness of the set $\mathcal{M}$ (and, hence, of $\mathcal{I}$). We will first make the following definition.

Definition: A symmetric matrix will be called *reducible* if it can be expressed as a linear combination of matrices in $\mathcal{M}$ with coefficients that are polynomials in $\mathcal{I}$.

Now let $P(n)$ stand for the proposition *every symmetric polynomial matrix of order n in A and B is reducible*. We shall prove the following theorem.

Theorem: $P(n)$ holds provided that $P(m)$ holds for all $m\leqslant n-2$.

Having already verified that $P(m)$ holds for all $m\leqslant 6$, this theorem guarantees that $P(n)$ holds for all n. To prove the theorem we shall need the following lemma.

Lemma: For any $M\in\mathcal{M}$, the symmetric matrices (i) *AMA*, (ii) BMB, (iii) $A^\alpha MB^\beta+B^\beta MA^\alpha$ $(\alpha,\beta=1$ or $2)$ are reducible.

Proof of the lemma: As the maximum order of $M\in\mathcal{M}$ is 4, cases (i), (ii), and (iii) for $\alpha=\beta=1$ are trivial because the resulting symmetric matrices are of order $\leqslant 6$, and all such matrices have been shown to be reducible. When either α or β equals 2, we need to consider only those $M\in\mathcal{M}$ for which the resulting matrix in (iii) has order greater than 6. In every such case we can isolate factors that are symmetric matrices of order 5 or 6 and are therefore reducible to matrices of order 3 or 4, respectively. For example,

$$A^2(A^2B+BA^2)B^\beta+B^\beta(A^2B+BA^2)A^2=A^4B^{1+\beta}+A^2BA^2B^\beta+B^\beta A^2BA^2+B^{1+\beta}A^4, \tag{3.5}$$

Substituting for the 5th order symmetric matrix A^2BA^2 in the middle two terms (and for A^4 in the other two) we find that the sum reduces to a symmetric matrix of order $3+\beta<6$. Similarly

$$A^2(A^2B^2+B^2A^2)B^\beta+B^\beta(A^2B^2+B^2A^2)A^2=A^4B^{2+\beta}+A^2B^2A^2B^\beta+B^\beta A^2B^2A^2+B^{2+\beta}A^4 \tag{3.6}$$

and reducibility is effected by substituting for $A^2B^2A^2$ in the middle two terms (and for A^4 in the other two), giving a $4+\beta\leqslant 6$ symmetric matrix, which is again reducible.

Proof of the theorem: An arbitrary symmetric polynomial matrix of order n in A and B will have highest order terms of the form

$$K=A^{\alpha_1}B^{\beta_1}A^{\alpha_2}B^{\beta_2}\cdots A^{\alpha_m}B^{\beta_m}+B^{\beta_m}A^{\alpha_m}\cdots B^{\beta_2}A^{\alpha_2}B^{\beta_1}A^{\alpha_1}, \tag{3.7}$$

where the sum of all exponents α_i and β_i is n. We implicitly define the matrices V and W by writing K in the form

$$K=A^{\alpha_1}VB^{\beta_m}+B^{\beta_m}V^TA^{\alpha_1}=A^{\alpha_1}B^{\beta_1}WA^{\alpha_m}B^{\beta_m}+B^{\beta_m}A^{\alpha_m}W^TB^{\beta_1}A^{\alpha_1}. \tag{3.8}$$

The proof is different depending on whether $\alpha_1\beta_m=0$ or not.

Case (i): If $\alpha_1\beta_m=0$, i.e., the first and last matrix in each term of K is the same (A or B), then we can write K (if, say, $\beta_m=0$) as

$$K=AUA+AU^TA=A(U+U^T)A,$$

where U stands for the remaining factors in the product and is, therefore, of order $n-2$. By using the lemma, K is reducible provided that $U+U^T$ is reducible, proving the theorem for this case.

Case (ii): If neither α_1 nor β_m vanishes, we write K using (3.8) as

$$K=A^{\alpha_1}(V+V^T)B^{\beta_m}+B^{\beta_m}(V+V^T)A^{\alpha_1}-A^{\alpha_1}V^TB^{\beta_m}-B^{\beta_m}VA^{\alpha_1}. \tag{3.7a}$$

Now $V+V^T$ is a symmetric matrix of order $n-\alpha_1-\beta_m\leqslant n-2$ and, by the lemma, the first two terms are reducible if $V+V^T$ is. The last two terms can be written $-A^{\alpha_1+\alpha_m}W^TB^{\beta_1+\beta_m}-B^{\beta_1+\beta_m}WA^{\alpha_1+\alpha_m}$ which is of the same form as K except that the first and last exponents are larger. If either $\alpha_1+\alpha_m$ or $\beta_1+\beta_m$ is greater than 2, the Cayley–Hamilton theorem reduces these two terms to a symmetric matrix of order $n-2$ in A and B, proving the theorem. If not, we can apply again the steps of case (ii) to these two terms, resulting in exponents which are now necessarily greater than 2 (since all α_i and β_i are $\geqslant 1$) and thus the Cayley–Hamilton theorem will be applicable to lower the order of these terms, completing the proof of the theorem.

Corollary: The set of scalars $\mathcal{I}$ is a complete set for the traces of arbitrary matrix products of A and B.

The proof follows trivially from the theorem and the fact that an arbitrary matrix can be written as the sum of a symmetric and an antisymmetric matrix—the latter having vanishing trace!

The proof of this theorem suggests a recursive procedure for fully reducing a symmetric matrix polynomial of arbitrary order in A, B: (i) for each term of order >6 apply the steps of the proof to obtain terms of lower order; (ii) for each term of order $\leqslant 6$ use the expressions in the Appendix. This procedure has been implemented in a set of MATHEMATICA routines which have been used in carrying out the calculations in the next section. (They are available by e-mail from the author.)

IV. A MATRIX IDENTITY AND ITS CONCOMITANTS

Let Q be an *antisymmetric* 3×3 matrix. Then the antisymmetric part of the Cayley–Hamilton theorem for Q reads

$$Q^3=1/2[Q^2]Q \tag{4.1}$$

(the symmetric part is just $\det Q=0$). Multiplying (4.1) by Q, we conclude that the *symmetric* matrix $X=Q^2$ satisfies the identity

$$Z\equiv X^2-1/2[X]X=O. \tag{4.2}$$

Thus, if Q is any antisymmetric matrix polynomial in A and B, then, by the results of the previous section, (4.2) will be an identity relating the matrices in $\mathcal{M}$! Taking $Q=AB-BA$ and evaluating (4.2) we find

$$\begin{aligned}Z=&2(2t^2-s^2t-4i_ai_bt+2ia_ai_bs^2+pqs+12j_aj_bs-8i_aj_bp-8i_bj_aq)\mathbf{1}\\&+2(qt+i_bps-2i_aj_bs-2i_ai_bq-8i_b^2j_a)A+2(pt+i_aqs-2i_bj_as-2i_ai_bp-8i_a^2j_b)B\\&+2(2i_bt-i_bs^2-q^2+6j_bp)A^2+2(2i_at-i_as^2-p^2+6j_aq)B^2+(st+pq-2i_ai_bs-36j_aj_b)\\&\times(AB+BA)+2(i_bp-qs+6i_aj_b)(A^2B+BA^2)+2(i_aq-ps+6i_bj_a)(AB^2+B^2A)\\&+2(s^2-3t+2i_ai_b)(A^2B^2+B^2A^2),\end{aligned} \tag{4.3}$$

while taking $Q=A^2B-BA^2$ we obtain a lengthy expression, Z_1, which can be shown to equal (after proper reduction of all matrix products) to

$$Z_1=\cdots=-2j_a(AZ+ZA). \tag{4.4}$$

Similarly, $Q=A^2B^2-B^2A^2$ gives

$$Z_2=\cdots=4j_aj_b(AZB+BZA+ABZ+ZBA). \tag{4.5}$$

It is clear that there is essentially only one independent matrix identity $Z=0$. One can obtain eight others by multiplying Z, as given by Eq. (4.3), symmetrically by the elements of $\mathcal{M}$ and reducing the resulting expressions. These are precisely the nine symmetric matrix identities relating the elements of $\mathcal{M}$ whose existence was inferred by Sneddon.[18] In fact it can be checked that the matrices S_1, S_2, and S_4 given in his Appendix A are equal, in our notation, to $6S_1=Z$, $24S_2=BZ+ZB$ and $6S_4=AZB+BZA+ABZ+ZBA$.

Now, multiplying Eq. (4.3) by each matrix in $\mathcal{M}$ in turn and taking traces, we find that all but one are identically zero (as polynomials in $\mathcal{I}$). The only nonvanishing trace, written as a polynomial in t, is

$$\begin{aligned}[Z(A^2B^2+B^2A^2)]=&-4t^3+(s^2+20i_ai_b)t^2-2(pqs-36sj_aj_b+2i_ai_bs^2-2i_aq^2-2i_bp^2\\&+12i_aj_bp+12i_bj_aq+16i_a^2i_b^2)t+p^2q^2+4i_a^2i_b^2s^2-16j_aj_bs^3+8i_aj_bps^2\\&+8i_bj_aqs^2-8i_a^2j_bqs-8i_b^2j_aps-96i_ai_bj_aj_bs-4i_a^2i_bq^2-4i_b^2i_ap^2-8j_aq^3\\&-8j_bp^3+72j_aj_bpq+32i_a^2i_bj_bp+32i_b^2i_aj_aq-432j_a^2j_b^2+16i_a^3j_b^2+16i_b^3j_a^2\\&+16i_a^3i_b^3.\end{aligned} \tag{4.6}$$

The vanishing of this expression gives the single syzygy connecting the invariants of the complete set $\mathcal{I}$. This syzygy was first obtained by Sneddon[17] (his Eq. (19)) in a very laborious and rather inelegant way. (Note that in Sneddon's expression the term $+4i_a^2i_b^2s^2$ is missing and the terms $-8i_a^2j_bqs-8i_b^2j_aps$ appear with a plus sign.) Our derivation shows that it is, ultimately, a consequence of the Cayley–Hamilton theorem.

V. DIFFERENT SPINOR CONTRACTIONS

The traceless parts of the eight symmetric matrices in $\mathcal{M}$ define eight symmetric 4-index spinors by the correspondence (2.1). The discussion in Sec. III proved completeness of the set $\mathcal{I}$ for contractions over *pairs* of indices of arbitrary products of these spinors. It is natural to ask what happens when different spinor contractions are performed.

Consider first triple index contractions. If U_{ABCD} and V_{ABCD} are any two of the eight basic spinors, then it is easy to show that the 2-index spinor $Q_{AB}=U_{(A}{}^{CDE}V_{B)CDE}$ corresponds[22] to the antisymmetric matrix $Q=VU-UV$. Thus the scalar $u=U_{(A}{}^{CDE}V_{B)CDE}X^{(A}{}_{PQR}Y^{B)PQR}$ is proportional to the trace of the symmetric matrix $(UV-VU)(XY-YX)+(XY-YX)(UV-VU)$, which, when expanded, is of the form considered in III (symmetric products of elements of $\mathcal{M}$). It follows that scalars like u can be expressed as polynomials in $\mathcal{I}$.

Single index contractions lead to a greater variety of spinors and scalars. Consider next the 4-index spinor (symmetrized over *ABCD*)

$$K_{ABCD}=U_A{}^{KLM}V_{BK}{}^{PQ}X_{CLP}{}^{R}Y_{DMQR}, \tag{5.1}$$

where a single index is contracted between every pair of spinors U, V, X, Y. Is the traceless 3×3 matrix corresponding to K_{ABCD} reducible? The answer can be found by using the fundamental ε identity[24]

$$\varepsilon_{AB}\varepsilon_{CD}=\varepsilon_{AC}\varepsilon_{BD}-\varepsilon_{BC}\varepsilon_{AD}. \tag{5.2}$$

We first write all contractions in terms of ε's

$$K_{ABCD}=U_A{}^{KLM}V_B{}^{K_1PQ}X_C{}^{L_1P_1R}Y_D{}^{M_1Q_1R_1}\varepsilon_{KK_1}\varepsilon_{LL_1}\varepsilon_{MM_1}\varepsilon_{PP_1}\varepsilon_{QQ_1}\varepsilon_{RR_1}$$

and use the identity (5.2) to replace $\varepsilon_{MM_1}\varepsilon_{PP_1}$. Regrouping terms we find that K_{ABCD} can be written as

$$\begin{aligned}&(U_A{}^{KLM}V_B{}^{K_1PQ}\varepsilon_{KK_1}\varepsilon_{MP})(X_C{}^{L_1P_1R}Y_D{}^{M_1Q_1R_1}\varepsilon_{M_1P_1}\varepsilon_{RR_1})\varepsilon_{LL_1}\varepsilon_{QQ_1}\\&\quad-(U_A{}^{KLM}X_C{}^{L_1P_1R}\varepsilon_{LL_1}\varepsilon_{MP_1})(V_B{}^{K_1PQ}Y_D{}^{M_1Q_1R_1}\varepsilon_{PM_1}\varepsilon_{QQ_1})\varepsilon_{KK_1}\varepsilon_{RR_1}\\&\quad=(U_A{}^{KLM}V_{BKM}{}^{Q})(X_C{}^{L_1P_1R}Y_{DR\,P_1}{}^{Q_1})\varepsilon_{LL_1}\varepsilon_{QQ_1}\\&\quad-(U_A{}^{KLM}X_{CLM}{}^{R})(V_B{}^{K_1PQ}Y_{DPQ}{}^{R_1})\varepsilon_{KK_1}\varepsilon_{RR_1}.\end{aligned}\tag{5.3}$$

The last expression involves contractions of *pairs* of indices only and, hence, is reducible.

Consider finally the 6-index spinor $\Pi^{(UV)}{}_{ABCDEF}=U_{(ABC}{}^{S}V_{DEF)S}$. Such a spinor can be formed from any two of the eight 4-index spinors, but cannot be represented in terms of 3×3 matrices (it corresponds to a third-rank tensor). We can form scalars by contracting any number of different spinors $\Pi^{(K)}$ among themselves. For example,

$$s_4=\Pi^{(1)ABCDEF}\Pi^{(2)}{}_{AB}{}^{KLMN}\Pi^{(3)}{}_{CDKL}{}^{RS}\Pi^{(4)}{}_{EFMNRS},$$

$$s_7=\Pi^{(1)ABCDEF}\Pi^{(2)}{}_{A}{}^{GHKLM}\Pi^{(3)}{}_{BG}{}^{PQRS}\Pi^{(4)}{}_{CHP}{}^{TUV}\Pi^{(5)}{}_{DKQT}{}^{XY}\Pi^{(6)}{}_{ELRUX}{}^{Z}\Pi^{(7)}{}_{FMSVYZ}.$$

In s_4 two indices are contracted between any two $\Pi^{(K)}$'s, while in s_7 only one. Clearly, there are many other possibilities. It is very likely that, using the identity (5.2), all such scalars can be reduced to successive contractions over pairs of indices of the constituent 4-spinors. However, lacking a general theorem to this effect, we will not attempt to examine the reducibility of these scalars. In any case, it is safe to say that the technical requirement imposed at the beginning, that only contractions of products of Ψ_{ABCD} and χ_{ABCD} over *pairs* of indices will be considered (which was needed in order to translate the problem into the language of 3×3 matrices), can be weakened considerably if not dropped altogether.

VI. A MINIMAL SET OF RIEMANN TENSOR INVARIANTS

When A is identified with Ψ and B with χ the invariants found in the complete set $\mathcal{I}$ can be identified with known Riemann tensor invariants. Thus i_a and j_a are the two (complex) Weyl tensor invariants I and J. The invariants i_b and j_b, being of 4th and 6th order in $\Phi_{ABC'D'}$, respectively, can be expressed as polynomials in the 2nd, 3rd and 4th order invariants of $\Phi_{ABC'D'}$. Specifically, let

$$\begin{aligned}r_2&=\Phi_{AA'}{}^{BB'}\Phi_{BB'}{}^{AA'},\\r_3&=\Phi_{AA'}{}^{BB'}\Phi_{BB'}{}^{CC'}\Phi_{CC'}{}^{AA'},\\r_4&=\Phi_{AA'}{}^{BB'}\Phi_{BB'}{}^{CC'}\Phi_{CC'}{}^{DD'}\Phi_{DD'}{}^{AA'}.\end{aligned}\tag{6.1}$$

Then it can be shown[8] that

$$i_b=(7/12)r_2^2-r_4,\tag{6.2}$$

$$6j_b=(1/3)r_3^2+(17/36)r_2^3-r_2r_4.\tag{6.3}$$

The introduction of these lower order invariants (r_2,r_3,r_4), however, may make some of the higher order invariants (p,q,t) become dependent on the lower order ones. Counting powers of Ψ and Φ, we find that only the following relations are possible:

$$p=c_1 i_a r_2,$$

$$q=c_2 s r_2, \tag{6.4}$$

$$t=c_3 p r_2+c_4 s^2+c_5 i_a r_4+c_6 i_a r_2^2,$$

where the coefficients c_i are constants. However, it is easy to check that there are no c_i that can make any one of (6.4) an identity in the components of Ψ and Φ. Hence the invariants p, q and t remain independent even when r_2, r_3, and r_4 are added to the set. Including the scalar curvature R, we conclude that the set $\{R,r_2,r_3,r_4,I,J,s,p,q,t\}$ of 4 real and 6 complex invariants (subject to the complex syzygy (4.8)) is a minimal set that must be contained in the complete set of invariants of the Riemann tensor. The set of Carminati and McLenaghan does not contain the invariants q and t, while the set of Zakhary and McIntosh does not contain t. (The additional invariants in these sets, containing both Ψ and $\overline{\Psi}$, cannot be used to write q and t.) Thus neither of these sets can be complete according to the definition used here. Moreover, as the number of invariants minus the number of syzygies for the minimal set equals 14, for every additional invariant that is added to the set an additional syzygy must be found to maintain the number of algebraically independent invariants equal to 14, the number of gauge-independent components of the Riemann tensor.

VII. SUMMARY AND CONCLUSIONS

The main results of this paper are the proof of completeness of the sets $\mathcal{M}$ and $\mathcal{I}$ given in Sec. III and the derivation of the matrix identities, together with the scalar syzygy (4.8), given in Sec. IV. They have all been obtained using entirely elementary means: beginning with Eq. (2.4), everything is a consequence of the Cayley–Hamilton theorem!

Considering the general problem, these results reinforce a conclusion reached recently in a related publication;[25] a definitive study of the Riemann tensor invariants can only be done after the algebra of the matrices Ψ, Φ (and their complex conjugates) has been fully worked out.

Note added in proof: Sneddon (private communication) has pointed out that the conjecture at the end of Sec. V (scalars obtained by different spinor index contractions can be converted to contractions over *pairs* of indices) is correct and can be proved using the rotor (bivector) notation. Particularly useful in the proof is the equation following Eq. (A5) in Appendix A of Ref. 17.

APPENDIX: ALL 4th, 5th AND 6th ORDER SYMMETRIC MATRICES

4th Order:

$$A^2BA+ABA^2=sA^2+pA-2j_aB,$$

$$AB^2A=-(A^2B^2+B^2A^2)+2i_bA^2+i_aB^2+qA+(t-2i_ai_b)\mathbf{1},$$

$$ABAB+BABA=(A^2B^2+B^2A^2)+s(AB+BA)-2(i_bA^2+i_aB^2)+2(2i_ai_b-t)\mathbf{1}.$$

There are two more equations, that can be obtained from the first two through the interchanges, $A\leftrightarrow B$, $i_a\leftrightarrow i_b$, $j_a\leftrightarrow j_b$, $p\leftrightarrow q$.

5th Order:

$$A^2B^2A+AB^2A^2=-2j_aB^2+qA^2+tA+4i_bj_a\mathbf{1},$$

$$A^2BAB+BABA^2=-i_a(AB^2+B^2A)+s(A^2B+BA^2)-2j_aB^2+qA^2-(t-2i_ai_b)A,$$

$$ABA^2B+BA^2BA=i_a(AB^2+B^2A)+p(AB+BA)-2j_aB^2-qA^2+(t-2i_ai_b)A,$$

$$ABABA=-s(A^2B+BA^2)+2j_aB^2+(s^2-t+i_ai_b)A+si_aB+(ps-2i_bj_a)\mathbf{1},$$

$$A^2BA^2=i_a(A^2B+BA^2)-2j_a(AB+BA)+pA^2-i_a^2B+(2sj_a-pi_a)\mathbf{1}.$$

There are five more equations, that can be obtained from these through the interchanges, $A\leftrightarrow B$, $i_a\leftrightarrow i_b$, $j_a\leftrightarrow j_b$, $p\leftrightarrow q$.

6th Order:

$$AB^2A^2B+BA^2B^2A=-s(A^2B^2+B^2A^2)+p(AB^2+B^2A)+q(A^2B+BA^2)+i_ai_b(AB+BA)$$
$$+s(i_bA^2+i_aB^2)+(2i_aj_b-i_bp)A+(2i_bj_a-i_aq)B$$
$$+(st-pq-2si_ai_b-4j_aj_b)\mathbf{1},$$

$$A^2BAB^2+B^2ABA^2=p(AB^2+B^2A)+q(A^2B+BA^2)+(i_ai_b-t)(AB+BA)+s(i_bA^2+i_aB^2)$$
$$-(2i_aj_b+i_bp)A-(2i_bj_a+i_aq)B+(st-pq-2si_ai_b-4j_aj_b)\mathbf{1},$$

$$ABABAB+BABABA=s(A^2B^2+B^2A^2)+(s^2+i_ai_b-t)(AB+BA)-2s(i_bA^2+i_aB^2)$$
$$-2(st-2si_ai_b-4j_aj_b)\mathbf{1},$$

$$AB^2ABA+ABAB^2A=-s(A^2B^2+B^2A^2)-q(A^2B+BA^2)+2si_bA^2+si_aB^2-2(i_aj_b-qs)A$$
$$+i_aqB+(st+pq-2si_ai_b-4j_aj_b)\mathbf{1},$$

$$A^2B^2AB+BAB^2A^2=s(A^2B^2+B^2A^2)-p(AB^2+B^2A)+(t-i_ai_b)(AB+BA)-s(i_bA^2+i_aB^2)$$
$$-(2i_aj_b-i_bp)A+(2i_bj_a+i_aq)B-(st-pq-2si_ai_b+4j_aj_b)\mathbf{1},$$

$$A^2B^2A^2=i_a(A^2B^2+B^2A^2)-2j_a(AB^2+B^2A)+(t-2i_ai_b)A^2-i_a^2B^2+4i_bj_aA$$
$$+(2i_a^2i_b+2qj_a-i_at)\mathbf{1},$$

$$ABA^2BA=-p(A^2B+BA^2)-2j_a(AB^2+B^2A)+(t-i_ai_b)A^2+(ps+2i_bj_a)A+i_apB+p^2\mathbf{1},$$

$$A^2BABA+ABABA^2=2j_a(AB^2+B^2A)+(s^2+2i_ai_b-2t)A^2+(ps-4i_bj_a)A-2sj_aB,$$

$$A^2BA^2B+BA^2BA^2=-i_a(A^2B^2+B^2A^2)+2j_a(AB^2+B^2A)+p(A^2B+BA^2)+2i_a(i_bA^2+i_aB^2)$$
$$-4i_bj_aA-2(2i_a^2i_b+2qj_a-i_at)\mathbf{1}.$$

There are six more equations, that can be obtained from the last six through the interchanges, $A\leftrightarrow B$, $i_a\leftrightarrow i_b$, $j_a\leftrightarrow j_b$, $p\leftrightarrow q$.

[1] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973), Sec. 8.1, p. 260.
[2] G. F. R. Ellis and B. G. Schmidt, Gen. Relativ. Gravit. **8**, 915 (1977).
[3] G. F. R. Ellis and B. G. Schmidt, Gen. Relativ. Gravit. **10**, 989 (1979).
[4] S. T. C. Siklos, Gen. Relativ. Gravit. **13**, 433 (1981).
[5] L. Witten, Phys. Rev. **113**, 357 (1959).
[6] A. Z. Petrov, *Einstein Spaces* (Pergamon, Oxford, 1969), Sec. 22.
[7] A. Karlhede, Gen. Relativ. Gravit. **12**, 693 (1980).
[8] G. C. Jolly and M. A. H. MacCallum, Class. Quantum Grav. **7**, 541 (1990).
[9] G. B. Gurevich, *Foundations of the Theory of Algebraic Invariants* (Noordhoff, Groningen, 1964).
[10] R. Penrose and W. Rindler, *Spinors and Space–time* (Cambridge University Press, Cambridge, 1986), Sec. 8.6, Vol. II, p. 263.

[11] V. V. Narlikar and K. R. Karmakar, Proc.-Indian Acad. Sci., Sect. A **29**, 91 (1948); A. Harvey, Class. Quantum Grav. **7**, 715 (1990).
[12] J. Géhéniau and R. Debever, Bull. Cl. Sci., Acad. R. Belg. **42**, 114 (1956).
[13] P. J. Greenberg, Stud. Appl. Math. **51**, 277 (1972).
[14] J. Carminati and R. G. McLenaghan, J. Math. Phys. **32**, 3135 (1991).
[15] E. Zakhary and C. B. G. McIntosh, Gen. Relativ. Gravit. **29**, 539 (1997).
[16] G. E. Sneddon, Class. Quantum Grav. **3**, 1031 (1986).
[17] G. E. Sneddon, J. Math. Phys. **37**, 1059 (1996).
[18] G. E. Sneddon, J. Math. Phys. **39**, 1659 (1998).
[19] R. S. Rivlin, J. Ration. Mech. Anal. **4**, 681 (1955).
[20] A. J. M. Spencer and R. S. Rivlin, Arch. Rat. Mech. Anal. **2**, 309 (1958).
[21] In Ref. 10, Sec. 8.3, p. 234.
[22] S. Bonanos, J. Math. Phys. **32**, 188 (1991).
[23] This set was first obtained in Refs. 19 and 20. The work reported here was completed without knowledge of Rivlin's and Spencer's results.
[24] In Ref. 10, Eq. (2.5.22).
[25] S. Bonanos, Gen. Relativ. Gravit. **30**, 653 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Darboux method and search of invariants for the Lotka–Volterra and complex quadratic systems

Laurent Cairó[a)]
Département de Mathématiques, MAPMO, UMR 6628, Université d'Orléans, B.P. 6759, 45067 Orléans, Cédex 2, France

Marc R. Feix
Département de Mathématiques, MAPMO, UMR 6628, Université d'Orléans, B.P. 6759, 45067 Orléans, Cédex 2, France and SUBATECH, Ecole des Mines, 4 rue Alfred Kastler, B.P. 20722, 44307 Nantes Cedex 3, France

Jaume Llibre[b)]
Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193–Bellaterra, Barcelona, Spain

(Received 10 June 1998; accepted for publication 9 December 1998)

The Darboux method introduces algebraic solutions quite useful to obtain invariant and first integrals of polynomial differential systems. Here we study the 2D Lotka–Volterra (LVS) and the complex quadratic system (QS) using straight lines for both and conics for the LVS. The conditions needed to obtain these invariants are given and a study of the phase space portrait is done. © *1999 American Institute of Physics.* [S0022-2488(99)02604-3]

I. INTRODUCTION

We consider the search of invariants for the two-dimensional differential system

$$\frac{\mathrm{d}x}{\mathrm{d}t}=P(x,y), \quad \frac{\mathrm{d}y}{\mathrm{d}t}=Q(x,y), \tag{1}$$

where P and Q are polynomials with coefficients in $\mathbf{F}$, where $\mathbf{F}$ is either the real field $\mathbf{R}$ or the complex field $\mathbf{C}$. We say that $m=\max\{\deg P,\deg Q\}$ is the *degree* of the polynomial differential system. This type of equations appear in the modelization of natural phenomena described in different branches of the science such as biology, chemistry, astrophysics, fluid mechanics, electronics, etc. Of particular interest are the systems such that $m=2$. The polynomial differential systems of degree 2 will be called *quadratic systems* (QS). One particularly well known quadratic system is the *Lotka-Volterra system* (LVS) which has been used to model the time evolutions of conflicting species in biology and of chemical reactions.[1,2] Among other applications, we find a QS in the equations of continuity describing the interactions of ions, electrons and neutral species in plasma physics (with the assumption of quasineutrality to eliminate either the ion or electron equation).[3] Moreover a reduced QS is obtained from a generalized Blasius equation for fluid flow around a wedge-shaped obstacle in boundary layer theory.[4] In the context of plasma physics, all the nonlinear terms represent binary interactions or model certain transport across the boundary of the system. There is a long history of research on finding sufficient conditions for which periodic solutions (center problem) exist for systems equivalent to the QS, and numerous results were obtained which we are not able to fully survey.[5] However, most of the previous works assumed that the origin is a linear center (i.e., having eigenvalues $\pm i$) which we do not assume here as starting point.

[a)]Electronic mail: lcairo@labomath.univ-orleans.fr
[b)]Electronic mail: jllibre@mat.uab.es

Although our main interest concerns the integrability, we will also consider cases for which one can exhibit a constant of the motion which is not a *first integral*. As it is well known in 2D a polynomial system is integrable if it possess a first integral. However, it can be interesting sometimes to know if it can have an *invariant*. Roughly speaking, with a first integral we can describe completely the phase portrait of the polynomial system, while with an invariant (a time-dependent first integral) we only can describe its asymptotic behavior. In fact, an invariant is a function $I(x,y,t)$ such that when $(x,y)=(x(t),y(t))$ is a solution of (1), I is a constant. Consequently, I must satisfy the equation

$$\frac{\partial I}{\partial t}+\frac{\partial I}{\partial x}P+\frac{\partial I}{\partial y}Q=0. \tag{2}$$

When I does not depend explicitly of t, the invariant is called a first integral.

In spite of the fact that they differ in the details of computation, all the methods to obtain invariants (Carleman linearization,[6,2] Hamiltonian method[7,8]) are based on an *a priori* hypothesis on the form of the invariant I. So, in the Carleman method this form is introduced in system (1) and the parameters entering in it are selected in order that (2) becomes an identity or, in the Hamiltonian method, through rescaling, the form of the equations leading to this invariant is determined and again we select the different parameters and unknowns which are at our disposal to identify these equations with the one we want to study.

Here we review a method proposed long time ago by Darboux.[9] A survey of many works triggered by the Darboux theorem together with a study of the integrability of real quadratic systems having an invariant conic is given in Ref. 10.

The Darboux method is based on the possibility of writing the invariant [or at least an integrating factor for system (1)] as the product of different functions $f_i(x,y)$ raised at a given power λ_i. It is on the form of these functions f_i that we introduce an ansatz and the subsequent need to identify the parameters. In that sense the Darboux method is not so different from the others cited above. Nevertheless the experience shows that we have somehow divided the difficulties of the unavoidable identification leading to algebraic but nonlinear equations with sometimes a number of equations greater than the number of unknowns with, consequently, conditions imposed on the parameters describing the dynamical system. From a physicist point of view the rule of the game is to find invariants with as few conditions as possible on the parameters of the given system.

The paper is organized as follows. In Sec. II we present the Darboux method, and apply it in Sec. III to the Lotka–Volterra system. In Sec. IV we apply it to the quadratic systems after reduction to a canonical form already used by two of the authors in the Painlevé analysis of such systems.[11] In Sec. V we give our conclusion.

II. THE DARBOUX METHOD

Although the method can be extended to more than 2 dimensions we present it for the planar system (1) where P and Q are polynomials in x and y.

As has been said in the Introduction, the Darboux method is based on the existence of functions $f_i(x,y)$ which are in fact *algebraic solutions* of the differential system. Suppose that we can determine two polynomials $f_i(x,y)$ and $K_i(x,y)$ such that

$$\frac{\partial f_i}{\partial x}P+\frac{\partial f_i}{\partial y}Q=K_i f_i . \tag{3}$$

Then equation $f_i(x,y)=0$ describes an algebraic curve which is formed by trajectories.

These f_i are going to be the ''bricks'' with which we will build the invariants. Suppose that we have obtained q functions f_i. The polynomial K_i is called the *cofactor* of f_i. We look for an invariant of the form

$$I=\prod_{i=1}^{q} f_i^{\mu_i}(x,y)\exp(st). \tag{4}$$

We obtain

$$\frac{\mathrm{d}I}{\mathrm{d}t}=\frac{\partial I}{\partial t}+\frac{\partial I}{\partial x}P+\frac{\partial I}{\partial y}Q=I\left[s+\sum_{i=1}^{q}\frac{\mu_i}{f_i}\left(\frac{\partial f_i}{\partial x}P+\frac{\partial f_i}{\partial y}Q\right)\right]. \tag{5}$$

Taking into account (3) and imposing that I is an invariant, we obtain

$$s+\sum_{i=1}^{q}\mu_i K_i=0. \tag{6}$$

The equations for μ_i are now linear equations. How many f_i do we need? Equation (3) shows that if the system is of degree m, K_i is at most of degree $m-1$ independently of the degree of f_i. The left-hand side of (6) is consequently a polynomial in x,y of degree at most $m-1$ with a total of $m(m+1)/2$ terms. Consequently (6) produces $m(m+1)/2$ equations where the unknowns are the μ_i. We see from (6) that if we allow s to be different from zero the system of equations for the μ_i is inhomogeneous, and we need $q=m(m+1)/2$. If we want a first integral then s must be zero and the system becomes homogeneous and we need a priori either a new f_i, or a new condition by imposing that the determinant of the system of equations for the μ_i is zero. We will call λ_i the solution of this system.

In fact it is not impossible to find cases where q is smaller than $m(m+1)/2$ with μ_i such that $\Sigma\mu_i K_i+s\equiv 0$. We see that the possibility of solving for μ_i in Eq. (6) depends on the conditions we put on the coefficients of (1). We should not forget that some conditions come from Eq. (6) and other ones come from Eq. (3).

Another possibility indicated also by Darboux is to obtain an integrating factor as follows. We seek an integrating factor of the form

$$R=\prod_{i=1}^{q} f_i^{\lambda_i}. \tag{7}$$

For R to be an integrating factor we must have

$$\operatorname{div}(RP,RQ)=\frac{\partial}{\partial x}(RP)+\frac{\partial}{\partial y}(RQ)=0. \tag{8}$$

We have

$$\frac{\partial R}{\partial x}=R\sum_i \frac{\lambda_i}{f_i}\frac{\partial f_i}{\partial x},\qquad \frac{\partial R}{\partial y}=R\sum_i \frac{\lambda_i}{f_i}\frac{\partial f_i}{\partial y}, \tag{9}$$

and

$$\operatorname{div}(RP,RQ)=R\operatorname{div}(P,Q)+R\sum_{i=1}^{q}\frac{\lambda_i}{f_i}\left(P\frac{\partial f_i}{\partial x}+Q\frac{\partial f_i}{\partial y}\right). \tag{10}$$

Taking (3) into account we obtain

$$\operatorname{div}(RP,RQ)=R\left[\operatorname{div}(P,Q)+\sum_{i=1}^{q}\lambda_i K_i\right]=0. \tag{11}$$

Therefore R will be an integrating factor if we can solve for λ_i the following system:

$$\sum_i^q \lambda_i K_i + \mathrm{div}(P,Q) = 0, \tag{12}$$

which is of the same type as Eq. (6). They differ only through the inhomogeneous terms. (12) gives $m(m+1)/2$ equations and, *a priori*, we need the same number of pairs f_i, K_i. However, for some usually nongeneric cases a smaller number can be enough.

A comment is in order about the number of conditions and its variation with m and the order of the algebraic solutions tested. For the existence of algebraic solutions of higher order we require larger number of conditions than for curves of smaller order. Since the number of algebraic solutions needed increases quickly with m, we see that the most interesting cases (i.e., with a not too high number of conditions) will be obtained using straight lines and conics as algebraic solutions for Lotka–Volterra and quadratic systems. This conclusion was already found using the Carleman or Hamiltonian method.

III. LOTKA–VOLTERRA SYSTEM

The bidimensional LVS writes

$$\dot{x} = x(a_1 + b_{11}x + b_{12}y), \quad \dot{y} = y(a_2 + b_{21}x + b_{22}y), \tag{13}$$

We begin with straight lines as algebraic solutions, i.e.,

$$f(x,y) = f_{00} + f_{10}x + f_{01}y, \quad K(x,y) = K_{00} + K_{10}x + K_{01}y. \tag{14}$$

Two first straight lines are obvious,

$$f(x,y) = x, \quad K(x,y) = a_1 + b_{11}x + b_{12}y, \tag{15}$$

$$f(x,y) = y, \quad K(x,y) = a_2 + b_{21}x + b_{22}y. \tag{16}$$

We remark that these two straight lines do not impose any condition. The interest of using these straight lines was noted by Cairó and Feix who were seeking invariants of the form

$$x^\alpha y^\beta P(x,y)\exp(st),$$

and have noted that the two terms $x^\alpha y^\beta$ were giving extra freedom. Now the fulfillment of (3) imposes either $f_{00}=0$ or $K_{00}=0$. We begin with the solution $K_{00}=0$. We obtain

$$K_{10} = b_{11}, \quad K_{01} = b_{22}, \quad f(x,y) = f_{00}\left(1 + \frac{b_{11}}{a_1}x + \frac{b_{22}}{a_2}y\right), \tag{17}$$

but we need one condition, namely,

$$r_{12} = \frac{b_{11}}{a_1}(b_{12} - b_{22}) + \frac{b_{22}}{a_2}(b_{21} - b_{11}) = 0, \tag{18}$$

found in Ref. 2 to obtain the invariant of type III (in fact a first integral).

Note that the straight line $f(x,y)=0$ with f given by (17) joins the two equilibrium points $-a_1/b_{11},0$ and $0,-a_2/b_{22}$ while the cofactor straight line $b_{11}x + b_{22}y = 0$ joins the origin to the equilibrium point outside the axes.

Another invariant straight line is obtained by taking $f_{00}=0$. It is easily proved that we need one condition namely $a_1 = a_2 = a$, and that $K_{10} = b_{11}$, $K_{01} = b_{22}$ and $K_{00} = a$. This condition is needed to obtain the invariant of type II.

TABLE I. Algebraic solutions and conditions for Theorem 1.

$f(x,y)$	$K(x,y)$	Condition	No.
x	$a_1+b_{11}x+b_{12}y$	—	1
y	$a_2+b_{21}x+b_{22}y$	—	2
$a_1+b_{11}x$	$b_{11}x$	$b_{12}=0$	3
$a_2+b_{22}y$	$b_{22}y$	$b_{21}=0$	4
$(b_{21}-b_{11})x+(b_{22}-b_{12})y$	$a_1+b_{11}x+b_{22}y$	$a_1=a_2$	5
$a_1a_2+b_{11}x+b_{22}y$	$b_{11}x+b_{22}y$	$r_{12}=0$	6

Two other possible linear algebraic solutions are $f(x,y)=x+a_1/b_{11}$ with $b_{12}=0$ as condition, and $f(x,y)=y+a_2/b_{22}$ with $b_{21}=0$ as condition. Table I recapitulates the 6 possible invariant straight lines.

We see that to build an invariant with straight lines we must include first the straight lines 1 and 2 which do not bring any condition. If we add the invariant straight line 5 we recover the invariant II of Cairó and Feix, and if we add the straight line 6 we recover invariant III which is in fact a first integral. Straight lines 3 and 4 do not bring interesting invariants since the associated invariant concerns the species fully decoupled from the other (we have either b_{12} or b_{21} equal to zero). Finally we try to build an invariant with no condition, and for that we select the two straight lines $f=x$ and $f=y$. If μ_1 and μ_2 are the respective exponents, we obtain, taking into account the expression of the cofactors,

$$\mu_1a_1+\mu_2a_2+s=0,\quad b_{11}\mu_1+b_{21}\mu_2=0,\quad b_{12}\mu_1+b_{22}\mu_2=0. \tag{19}$$

Since s is at our disposal we can ignore the first equation, but the two others being homogeneous will have a nontrivial solution only if $\Delta=b_{11}b_{22}-b_{12}b_{21}=0$. This is the invariant of type I in Ref. 2. The LVS need at least one condition to posses an invariant. So the following theorem can be established.

Theorem 1: *Let $f_1=x=0$ and $f_2=y=0$ be the two trivial algebraic solutions of a LVS. Then the following statement holds*:

(a) *If $r_{12}=0$ the expression $f_1^{\lambda_1}f_2^{\lambda_2}$ is an integrating factor of the LVS with $\lambda_1=[(a_2-a_1)b_{22}+a_2b_{12}]/(a_1b_{22}-a_2b_{12})$ and $\lambda_2=(2a_1b_{22}-a_2b_{12})/(a_1b_{22}-a_2b_{12})$. If the condition is $\Delta=0$, then the expression $f_1^{\mu_1}f_2^{\mu_2}e^{st}$ is a Darboux invariant of the LVS provided that μ_1 and μ_2 be a solution of the system (19).*
If a LVS has a third algebraic solution $f_3=0$ of degree 1, then modulus the symmetry $(x,y,b_{11},b_{12},a_1,b_{21},b_{22},a_2)\rightarrow(y,x,b_{22},b_{21},a_2,b_{12},b_{11},a_1)$ and provided that the LVS satisfies one of the conditions indicated in Table I, there exist a Darboux invariant $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}e^{st}$ and an integrating factor $f_1^{\lambda_1}f_2^{\lambda_2}f_3^{\lambda_3}$ having the following values for f_3 and the exponents λ_i,μ_i:

(b) *If $a_1b_{11}b_{22}\neq 0$, and condition (3) is satisfied, then*

$$f_3=b_{11}x+a_1,$$

$$\mu_1=-s/a_1,\quad \mu_2=0,\quad \mu_3=s/a_1,$$

$$\lambda_1=(a_2-a_1)b_{11},\quad \lambda_2=-2a_1b_{11},\quad \lambda_3=a_1(b_{21}-b_{11})-a_2b_{11}.$$

(c) *If $a_1(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$, and condition (5) is satisfied, then*

$$f_3=(b_{21}-b_{11})x+(b_{22}-b_{12})y,$$

$$\mu_1 = s\frac{b_{22}}{a_1(b_{12}-b_{22})}, \quad \mu_2 = s\frac{b_{11}}{a_1(b_{21}-b_{11})}, \quad \mu_3 = s\frac{b_{11}b_{22}-b_{12}b_{21}}{a_1(b_{12}-b_{22})(b_{21}-b_{11})},$$

$$\lambda_1 = b_{12}(b_{21}-b_{11}), \quad \lambda_2 = b_{21}(b_{12}-b_{22}), \quad \lambda_3 = -[b_{11}(b_{12}-b_{22})+b_{22}(b_{21}-b_{11})].$$

(d) *If* $a_2^2b_{11}^2+a_1^2b_{22}^2\neq 0$, *and condition (6) is satisfied, then there exist a first integral* $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}$ *with*

$$f_3 = 1+\frac{b_{11}}{a_1}x+\frac{b_{22}}{a_2}y,$$

$$\mu_1 = b_{22}(b_{21}-b_{11}), \quad \mu_2 = b_{11}(b_{12}-b_{22}), \quad \mu_3 = b_{11}b_{22}-b_{12}b_{21}.$$

Next we turn to quadratic algebraic solutions $f(x,y)=0$, which we write

$$f(x,y) = f_{00}+f_{10}x+f_{01}y+f_{20}x^2+f_{11}xy+f_{02}y^2.$$

The cofactor is of the form $K(x,y)=K_{00}+K_{10}x+K_{01}y$ but K_{00} must be zero if $f_{00}\neq 0$. The identification of the terms in x^3 and y^3 give

$$f_{20}K_{10} = 2f_{20}b_{11}, \quad f_{02}K_{01} = 2f_{02}b_{22},$$

and if $f_{20}f_{02}\neq 0$ (we will come back to this possibility later on), we get

$$K_{10} = 2b_{11}, \quad K_{01} = 2b_{22}. \tag{20}$$

Taking $f_{00}=1$, the identification of terms in x and y gives

$$f_{10} = \frac{2b_{11}}{a_1}, \quad f_{01} = \frac{2b_{22}}{a_2}. \tag{21}$$

Identification of terms in x^2 and y^2 gives, taking (20) and (21) into account

$$f_{20} = \frac{b_{11}^2}{a_1^2}, \quad f_{02} = \frac{b_{22}^2}{a_2^2}. \tag{22}$$

All quantities except f_{11} are known and we are left with three equations from the identification of terms in xy, x^2y, and xy^2. Consequently we will need two conditions. After some calculations we get

$$f_{11} = -2\frac{b_{11}b_{22}}{a_1a_2},$$

with the two conditions

$$b_{12} = b_{22}\left(2+\frac{a_1}{a_2}\right), \quad b_{21} = b_{11}\left(2+\frac{a_2}{a_1}\right). \tag{23}$$

The conic is

$$f(x,y) = \left(1+\frac{b_{11}}{a_1}x+\frac{b_{22}}{a_2}y\right)^2 - 4\frac{b_{11}b_{22}}{a_1a_2}xy,$$

and the associated cofactor is

$$K(x,y) = 2b_{11}x+2b_{22}y.$$

TABLE II. Consequences of the identification of the different terms of (3).

x	y	x^3	y^3	
$f_{10}=2b_{11}/a_1$	$f_{01}=K_{01}/a_2$	$K_{10}=2b_{11}$	$f_{02}=0$	
x^2	y^2	xy^2	x^2y	xy
$f_{20}=b_{11}^2/a_1^2$	$K_{01}=b_{22}$	$f_{11}=0$	$b_{22}=2b_{12}$	$b_{21}=b_{11}(2+a_2/a_1)$

It must be realized that the presence of two conditions as given by (23) leaves a one parameter LVS equation. Indeed it is easily proved that (13) can be written as a three parameter system. Introducing $X=b_{11}x/a_1$, $Y=b_{22}y/a_2$, $C_1=b_{12}a_2/b_{22}a_1$, $C_2=b_{21}a_1/b_{11}a_2$, (13) becomes

$$\frac{dX}{a_1 dt}=X(1+X+C_1Y), \quad \frac{dY}{a_2 dt}=Y(1+C_2X+Y), \tag{24}$$

where

$$C_1=1+2\frac{a_2}{a_1}, \quad C_2=1+2\frac{a_1}{a_2},$$

and, indeed, writing the system (24) in the form dY/dX, we are left with one relevant parameter, (a_1/a_2).

We briefly discuss the other possibilities. The identification of the coefficients of the terms in x^3 and y^3 can also be obtained by taking f_{20} (for x^3) and f_{02} (for y^3) equal to zero. Identifying both leads to an uninteresting case: either $f_{11}=0$ or $b_{12}=b_{21}=0$. Breaking the symmetry we take $f_{02}=0$. Table II gives the result of the identification of the different terms. We recover again two conditions,

$$b_{22}=2b_{12}, \quad b_{21}=b_{11}\left(2+\frac{a_2}{a_1}\right).$$

One is identical to the previous case, the other is different. However, there is still another case for invariant conics of type (19). The identification of the y^2 term gives $f_{01}b_{22}=K_{01}f_{01}$ and assuming $f_{01}\neq 0$ we deduce $K_{01}=b_{22}$. Finally another solution is possible with $f_{01}=0$. Since (4th column in Table II) $f_{01}=K_{01}/a_2$, we must take $K_{01}=0$ and the table is modified. The identification of the term in xy^2 gives

$$b_{12}+b_{22}=0,$$

which is the first condition. The next term x^2y gives f_{11} and the term in xy gives the second condition

$$b_{21}=b_{11}\left(2+\frac{a_2}{a_1}\right),$$

the same as in the previous two cases.

We must now consider a nonzero constant term in the cofactor with, consequently, no constant term in f with

$$K(x,y)=K_{00}+K_{10}x+K_{01}y,$$

$$f(x,y)=f_{10}x+f_{01}y+f_{20}x^2+f_{11}xy+f_{02}y^2.$$

The terms in x^3 and y^3 provide the usual results $K_{10}=2b_{11}$, $K_{01}=2b_{22}$ with f_{20} and f_{02} different from zero. Then the terms in x and y give $f_{10}K_{00}=f_{10}\,a_1$ and $f_{01}K_{00}=f_{01}a_2$ with a first condition $a_1=a_2$ (assuming $f_{10}\neq 0$ and $f_{01}\neq 0$). However, we have found that this condition is sufficient to provide the invariant II. The existence of a conic as an invariant curve implies a second condition. Obtaining an invariant with this conic and the two straight lines $x=0$, $y=0$ does not present any interest since we can get an invariant with only the first condition. However, could it be that we recover a first integral? Pushing the computation we obtain for the conic the following relations (in addition to $a_1=a_2=a$):

$$f_{10}=(a/b_{11})f_{20}, \quad f_{01}=(a/b_{22})f_{02},$$

$$f_{20}=(b_{21}-b_{11})/[2(b_{22}-b_{12})]f_{11},$$

$$f_{02}=(b_{12}-b_{22})/[2(b_{11}-b_{21})]f_{11}.$$

The second condition comes from the identification of terms in x,y. We get in addition to $r_{12}=0$, the relation

$$3b_{11}b_{22}+b_{12}b_{21}-2b_{11}b_{12}-2b_{22}b_{21}=0, \tag{25}$$

which is the interesting one. Now the three cofactors are

$$a+b_{11}x+b_{12}y, \quad a+b_{21}x+b_{22}y, \quad a+2b_{11}x+2b_{22}y,$$

and we obtain a first integral if the determinant of the matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ b_{11} & b_{21} & 2b_{11} \\ b_{12} & b_{22} & 2b_{22} \end{pmatrix}$$

cancels, but this gives again (25) which together with $a_1=a_2$ provide the two conditions needed to obtain a first integral. If we assume that one of the coefficients, say f_{01}, is zero, then $K_{00}=a_1$ and the cancellation of the terms in x^2 and y^2 gives, respectively,

$$a_1 f_{20}=b_{11}f_{10} \ \text{ and } \ a_1=2a_2. \tag{26}$$

Finally looking at the terms in xy^2, x^2y and xy we obtain

$$f_{11}=2f_{02}(b_{11}-b_{21})/(b_{12}-b_{22}), \tag{27}$$

$$f_{11}=2f_{20}(b_{22}-b_{12})/(b_{21}-b_{11}), \tag{28}$$

$$f_{11}=(2b_{22}-b_{12})f_{10}/a_2. \tag{29}$$

Equations (28) and (29) and the first of (26) give the second condition, the first being $a_1=2a_2$,

$$b_{21}(2b_{22}-b_{12})-b_{11}(3b_{22}-2b_{12})=0.$$

The following theorem recapitulates the five cases obtained with a conic as algebraic solution in addition to the two straight lines $f_1=x, f_2=y$. We note that a trivial interchange between x and y, a_1 and a_2, b_{11} and b_{22}, b_{12} and b_{21}, give three other invariants in cases (a), (b), and (d) while cases (c) and (e) are invariant under this interchange.

Theorem 2: *Let $f_1=x=0$ and $f_2=y=0$. If a LVS has a third algebraic solution $f_3=0$ of degree 2, then modulus the symmetry* $(x,y,b_{11},b_{12},a_1,b_{21},b_{22},a_2)$

TABLE III. Invariant conditions of Theorem 2.

Statement	Condition 1	Condition 2
a	$b_{22}-2b_{12}=0$	$(2a_1+a_2)b_{11}-a_1b_{21}=0$
b	$b_{12}+b_{22}=0$	$(2a_1+a_2)b_{11}-a_1b_{21}=0$
c	$(a_1+2a_2)b_{22}-b_{12}a_2=0$	$(2a_1+a_2)b_{11}-a_1b_{21}=0$
d	$a_1=2a_2$	$b_{21}(2b_{22}-b_{12})-b_{11}(3b_{22}-2b_{12})=0$
e	$a_1=a_2$	$b_{21}(2b_{22}-b_{12})-b_{11}(3b_{22}-2b_{12})=0$

$\rightarrow(y,x,b_{22},b_{21},a_2,b_{12},b_{11},a_1)$, *and under the conditions given in Table III, a Darboux invariant* $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}e^{st}$ *and an integrating factor* $f_1^{\lambda_1}f_2^{\lambda_2}f_3^{\lambda_3}$ *can be found having the following values for* f_3 *and the exponents* μ_i,λ_i:

(a) If $a_1a_2b_{11}\neq 0$, *then*

$$f_3=\left(1+\frac{b_{11}}{a_1}x\right)^2+\frac{b_{22}}{a_2}y,$$

$$\mu_1=-s/a_1,\quad \mu_2=0,\quad \mu_3=s/2a_1,$$

$$\lambda_1=-2,\quad \lambda_2=(a_1-a_2)/a_2,\quad \lambda_3=-(2a_1+a_2)/(2a_2).$$

(b) If $a_1(a_1+a_2)b_{11}\neq 0$, *then*

$$f_3=\left(1+\frac{b_{11}}{a_1}x\right)^2+2\frac{b_{11}b_{22}}{a_1(a_1+a_2)}xy,$$

$$\mu_1=\mu_2=-s/(a_1+a_2),\quad \mu_3=(3a_1+a_2)s/2a_1(a_1+a_2),$$

$$\lambda_1=-a_1/(a_1+a_2),\quad \lambda_2=-(2a_1+a_2)/(a_1+a_2),\quad \lambda_3=(a_1-a_2)/[2(a_1+a_2)].$$

(c) If $a_1a_2b_{11}\neq 0$, *then*

$$f_3=\left(1+\frac{b_{11}}{a_1}x+\frac{b_{22}}{a_2}y\right)^2-4\frac{b_{11}b_{22}}{a_1a_2}xy,$$

$$\mu_1=-s/(2a_1),\quad \mu_2=-s/(2a_2),\quad \mu_3=(a_1+a_2)s/(2a_1a_2),$$

$$\lambda_1=-1,\quad \lambda_2=-1,\quad \lambda_3=-1/2.$$

(d) If $a_1b_{11}(2b_{22}-b_{12})\neq 0$, *then*

$$f_3=a_1b_{11}x+[b_{11}x+(2b_{22}-b_{12})y]^2,$$

$$\mu_1=2b_{22}s/a_1(b_{12}-2b_{22}),\quad \mu_2=2s/a_1,\quad \mu_3=2(b_{22}-b_{12})s/a_1(b_{12}-2b_{22}),$$

$$\lambda_1=(b_{22}-b_{12})/(b_{12}-2b_{22}),\quad \lambda_2=-2,\quad \lambda_3=b_{12}/[2(b_{12}-2b_{22})].$$

(e) Finally, a first integral $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}$ *exists if* $a_1b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$. *Then*

$$f_3=a_1[b_{22}(b_{21}-b_{11})^2x+b_{11}(b_{12}-b_{22})^2y]+b_{11}b_{22}[(b_{21}-b_{11})x-(b_{12}-b_{22})y]^2,$$

$$\mu_1=b_{21}-2b_{11},\quad \mu_2=b_{11},\quad \mu_3=b_{11}-b_{21}.$$

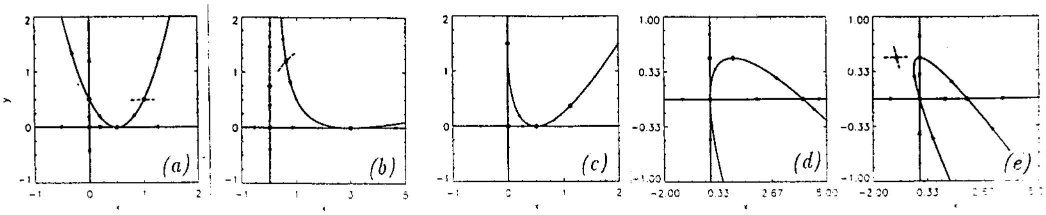

FIG. 1. Typical phase portraits for the LVS having invariants formed with a conic.

The qualitative picture of the phase portrait at finite distance from the origin for each statement of Theorem 2 is given in Fig. 1.

IV. THE COMPLEX QUADRATIC SYSTEM

The LVS are a particular case of the *quadratic systems*. Its most general form is

$$\dot{x}=\frac{dx}{dt}=k_1+a_{11}x+a_{12}y+b_{11}x^2+b_{12}xy+c_1y^2,$$
$$\dot{y}=\frac{dy}{dt}=k_2+a_{21}x+a_{22}y+b_{21}xy+b_{22}y^2+c_2x^2, \tag{30}$$

where the coefficients and the variables x,y are complex, but the independent variable t is real. Here we are going to consider a canonical version of (30), which is the result first, of a translation of the origin eliminating the constant terms k_1 and k_2,

$$x=\overline{x}+m_1, \quad y=\overline{y}+m_2, \tag{31}$$

and second, of a linear transformation

$$x=\alpha_{11}\overline{x}+\alpha_{12}\overline{y}, \quad y=\alpha_{21}\overline{x}+\alpha_{22}\overline{y}, \tag{32}$$

to eliminate the terms c_1y^2 and c_2x^2 in (30). Since the linear transformations (31) and (32) do not introduce higher order terms, we arrive at a new system with the same form as (30) (with of course $k_1=k_2=0$) but with all the terms in overbar variables. Of interest are

$$\overline{c_1}=-\alpha_{12}^3c_2+\alpha_{12}^2\alpha_{22}(b_{11}-b_{21})+\alpha_{12}\alpha_{22}^2(b_{12}-b_{22})+\alpha_{22}^3c_1,$$
$$\overline{c_2}=\alpha_{11}^3c_2-\alpha_{11}^2\alpha_{21}(b_{11}-b_{21})-\alpha_{11}\alpha_{21}^2(b_{12}-b_{22})-\alpha_{21}^3c_1.$$

Introducing

$$l=\frac{\alpha_{12}}{\alpha_{22}}, \quad m=\frac{\alpha_{11}}{\alpha_{21}}, \tag{33}$$

it is easily checked that the two equations $\overline{c_1}=0$ and $\overline{c_2}=0$ coincide and write

$$c_2X^3+(b_{21}-b_{11})X^2+(b_{22}-b_{12})X-c_1=0. \tag{34}$$

In (34) we must take for l and m two different roots, otherwise $\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21}=0$ and the linear transformation becomes degenerate. The transformation is nondegenerate if we take different roots among the three roots of (34) for l and m. Note at this point that when two roots are complex conjugate, the resulting complex quadratic system has complex coefficients, without modifying the generality of our subsequent study. Two coefficients (for example α_{11} and α_{22}) are still arbitrary and one could hope to be chosen in order to diagonalize the linear terms. If that were possible, we could reduce the quadratic system to the LVS. To see the impossibility we use (32) to diagonalize the linear terms. We obtain a similar result, namely, that l and m, as given by (33), are the two different roots of the second degree equation

$$a_{21}X^2+(a_{22}-a_{11})X-a_{12}=0. \tag{35}$$

Of course the same l and m cannot be simultaneously the roots of (34) and (35). Excluding the case of a double root we can say that the quadratic system becomes, without loss of generality

$$\begin{aligned}\dot{x}&=a_{11}x+a_{12}y+b_{11}x^2+b_{12}xy,\\ \dot{y}&=a_{21}x+a_{22}y+b_{21}xy+b_{22}y^2,\end{aligned} \tag{36}$$

where we have, once again, dropped the over-bar on the variables. The only—but essential—difference with the LVS is the presence of the off-diagonal terms $a_{12}y$ and $a_{21}x$. It is clear that the above transformation, can in general, generate complex coefficients. For this reason system (36) is called *complex quadratic system*, which in short will be called here QS.

The search for invariant straight lines is quite similar to the one given above for the LVS. To begin with we look for a (complex) straight line going through the origin, i.e.,

$$f(x,y)=f_{10}x+f_{01}y, \quad K(x,y)=K_{00}+K_{10}x+K_{01}y,$$

where we suppose first that f_{10} and f_{01} are different from zero. The terms in x^2 and y^2 in (3) gives $K_{10}=b_{11}$ and $K_{01}=b_{22}$. The terms in xy, x, and y give, respectively,

$$f_{10}(b_{22}-b_{12})+f_{01}(b_{11}-b_{21})=0, \tag{37}$$

$$f_{10}(a_{11}-K_{00})+f_{01}a_{21}=0, \quad f_{10}a_{12}+f_{01}(a_{22}-K_{00})=0. \tag{38}$$

If we suppose $b_{21}\neq b_{11}$ and $b_{12}\neq b_{22}$, we deduce f_{10}/f_{01} from (37) and introducing it in (38) we obtain K_{00} and a condition, the values of which are given below in **(5)** of Proposition 3. A particularly interesting case is obtained when $b_{12}=b_{22}$ and $b_{21}=b_{11}$. Then (37) is automatically fulfilled and the two Eqs. (38) have a nontrivial solution only in the case of K_{00} being the solution of the equation

$$K_{00}^2-(a_{11}+a_{22})K_{00}+\Delta_a=0, \tag{39}$$

where $\Delta_a=a_{11}a_{22}-a_{12}a_{21}$ and the two roots of K_{00} are the eigenvalues of the matrix (a)

$$(a)=\begin{pmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}. \tag{40}$$

The precise values are given below in **(5′)** and **(5″)** of Proposition 3.

The other general case is obtained with

$$f(x,y)=f_{00}+f_{10}x+f_{01}y, \quad K(x,y)=K_{10}x+K_{01}y.$$

Again the terms in x^2 and y^2 give $K_{10}=b_{11}$ and $K_{01}=b_{22}$ (assuming $f_{00},f_{10},f_{01}\neq 0$). The terms in x and y give

$$f_{10}a_{11}+f_{01}a_{21}=f_{00}b_{11}, \quad f_{10}a_{12}+f_{01}a_{22}=f_{00}b_{22}. \tag{41}$$

The solution of (41) give f_{10} and f_{01} which introduced in (37) provides a condition (case **(6)** of Proposition 3).

As in the LVS we must also consider cases where one of f_{10} or f_{01} is equal to zero. In fact we recuperate the two axes and the two parallel to the axes given in Table I (under the numbers 1, 2, 3, and 4) with conditions differently written. Now these straight lines and especially 3 and 4, are more interesting. In the LVS two conditions were needed ($b_{12}=b_{21}=0$), resulting in a decoupling of the two equations, while now, it is a relation between four coefficients (a_{12}, a_{11}, b_{12}, and b_{11}) or (a_{21}, a_{22}, b_{21}, and b_{22}) which are needed. Proposition 3 recapitulates the eight different invariant straight lines, their cofactors, and the conditions for their existence.

Proposition 3: *A QS has a linear algebraic solution* $f=0$ with cofactor K in the following cases:

(1) *If* $a_{12}=0$, *then* $f=x$ *and* $K=a_{11}+b_{11}x+b_{12}y$.
(2) *If* $a_{21}=0$, *then* $f=y$ *and* $K=a_{22}+b_{21}x+b_{22}y$.
(3) *If* $a_{11}b_{11}\neq 0$ *and* $a_{12}=a_{11}b_{12}/b_{11}$, *then* $f=a_{11}+b_{11}x$ *and* $K=b_{11}x+b_{12}y$.
(4) *If* $a_{22}b_{22}\neq 0$ *and* $a_{21}=a_{22}b_{21}/b_{22}$, *then* $f=a_{22}+b_{22}y$ *and* $K=b_{21}x+b_{22}y$.
(5) *If* $(b_{21}-b_{11})(b_{12}-b_{22})\neq 0$ *and* $r_1\equiv(b_{12}-b_{22})(b_{21}-b_{11})(a_{11}-a_{22})+a_{12}(b_{21}-b_{11})^2-a_{21}(b_{12}-b_{22})^2=0$, *then* $f=(b_{21}-b_{11})x+(b_{22}-b_{12})y$ *and* $K=a_{11}-a_{21}(b_{12}-b_{22})/(b_{21}-b_{11})+b_{11}x+b_{22}y$.
(5′) *If* $b_{12}=b_{22}$ *and* $b_{21}=b_{11}$, *then* $f=a_{21}x+(K_{00}^{[1]}-a_{11})y$ *and* $K=K_{00}^{[1]}+b_{11}x+b_{22}y$, *where* $K_{00}^{[1]}$ *is the first root of (39), which writes* $(a_{11}+a_{22})/2+[(a_{11}-a_{22})^2+4a_{12}a_{21}]^{1/2}/2$.
(5″) *If* $b_{12}=b_{22}$ *and* $b_{21}=b_{11}$, *then* $f=a_{12}y+(K_{00}^{[2]}-a_{22})x$ *and* $K=K_{00}^{[2]}+b_{11}x+b_{22}y$, *where* $K_{00}^{[2]}$ *is the second root of (39), which writes* $(a_{11}+a_{22})/2-[(a_{11}-a_{22})^2+4a_{12}a_{21}]^{1/2}/2$.
(6) *If* $b_{11}b_{22}(b_{21}-b_{11})(b_{12}-b_{22})(a_{11}a_{22}-a_{12}a_{21})\neq 0$ *and* $r_2\equiv(a_{12}b_{11}-a_{11}b_{22})(b_{21}-b_{11})+(a_{21}b_{22}-a_{22}b_{11})(b_{12}-b_{22})=0$, *then* $f=a_{11}a_{22}-a_{12}a_{21}+(a_{22}b_{11}-a_{21}b_{22})x+(a_{11}b_{22}-a_{12}b_{11})y$ *and* $K=b_{11}x+b_{22}y$.

A comment is in order about the two invariant straight lines **(5’)** and **(5’’)**. They do not contain any b_{ij} and are the support of the two eigenvectors of the matrix (a). Moreover it is easily shown that in this case ($b_{12}=b_{22}, b_{21}=b_{11}$) we have only three equilibrium points at finite distance from the origin: the origin, one on **(5′)** and the last one on **(5″)**. Moreover the invariant straight line **(6)** goes through these two last points.

Now with these eight ‘‘bricks’’ we are going to build the invariants, first integrals and integrating factors. Let us begin with the obtention of Darboux invariants. Getting q invariant straight lines labeled with i going from 1 to q, we obtain the identity

$$\sum_{i=1}^{q}\mu_i(K_{00i}+K_{10i}x+K_{01i}y)+s=0,$$

i.e., two homogeneous linear equations in μ_i,

$$\sum_{i=1}^{q}K_{10i}\mu_i=0,\quad \sum_{i=1}^{q}K_{01i}\mu_i=0, \tag{42}$$

and the inhomogeneous equation

$$\sum_{i=1}^{q}K_{00i}\mu_i+s=0. \tag{43}$$

Since s is at our disposal, let us leave for a moment (43) which will simply fix s. For the two Eqs. (42) to be fulfilled we need either three straight lines (with 3 μ_i) or two straight lines, but then the determinant of the system must be zero. Since one straight line implies one condition we need in principle three conditions. However, one can find exceptional cases with two straight lines having a determinant automatically equal to zero. For example, the cofactors of straight lines **(5)** and **(6)** have the same K_{10i} and K_{01i} (respectively, b_{11} and b_{22}) and we get an invariant with only two straight lines and two conditions.

Concerning the integrating factor, the things are somewhat similar except that now s is replaced by $\mathrm{div}(P,Q)$, with $\mathrm{div}(P,Q)=A_0+A_1x+A_2y$. We have now three inhomogeneous equations with three conditions.

A priori the most expensive in conditions should be the first integral. Since we get (42) and (43) and this time with $s=0$, we have a system of three homogeneous linear equations. The identification of the determinant plus the obtention of the three straight lines gives four conditions, but again special cases can be found. For example straight lines **(5′)**, **(5″)** and **(6)** altogether, ask for only two conditions ($b_{12}=b_{22}$ and $b_{21}=b_{11}$). Moreover they all possess the same $K_{10i}=b_{11}$ and $K_{01i}=b_{22}$ in their three cofactors and we get a first integral with only two conditions. Consequently each case must be separately examined.

In the rest of the paper we use the following notation: If a first integral, an integrating factor or a Darboux invariant of a complex QS has been obtained using for instance the algebraic solutions of statements **(5)** and **(6)** of Proposition 3, we only say that the integrable case is **(56)**, etc. We present three types of results.

A. Invariant (56)

It is the invariant built with straight lines **(5)** and **(6)**. The coexistence of straight lines **(5)** and **(6)** implies $r_1=r_2=0$ and a first solution is

$$a_{12}=a_{22}(b_{22}-b_{12})/(b_{11}-b_{21}),\qquad a_{21}=a_{11}(b_{11}-b_{21})/(b_{22}-b_{12}),\tag{44}$$

but it is easily checked that under conditions (44), the straight lines **(5)** and **(6)** are identical (the straight line **(6)** passing also through the origin) with, as a consequence, not enough straight lines to build an invariant. Fortunately another solution is possible with

$$b_{11}(b_{22}-b_{12})+b_{22}(b_{11}-b_{21})=0,\tag{45}$$

$$(a_{11}-a_{22})b_{11}b_{22}-a_{12}b_{11}^2+a_{21}b_{22}^2=0.\tag{46}$$

The cofactors being, respectively, $b_{11}x+b_{22}y+K_{00}=0$ for **(5)** and $b_{11}x+b_{22}y=0$ for **(6)**, the equations write

$$b_{11}(\mu_1+\mu_2)=0,\quad b_{22}(\mu_1+\mu_2)=0,\quad K_{00}\mu_1+s=0.$$

Taking $\mu_2=-\mu_1$, we get an invariant with only two conditions, (45) and (46). Now this result must be compared with the two conditions given for that system in Ref. 11. For the passing of the Painlevé test as given in Ref. 11, (45) is the index condition (4.31) and (46) is the compatibility condition (4.33). Indeed with, for example, $\mu_1=1$ and $\mu_2=-1$ we obtain an analytical invariant in agreement with the conclusion obtained in Ref. 11.

B. Invariant (5′5″) and (5′5″6)

In the case where the invariant is built with **(5′)** and **(5″)** or **(5′)**, **(5″)** and **(6)** we have the two conditions $b_{21}=b_{11}$, $b_{12}=b_{22}$ and the corresponding cofactors are $b_{11}x+b_{22}y+K_{00}^{[1]}$, $b_{11}x+b_{22}y+K_{00}^{[2]}$ and $b_{11}x+b_{22}y$, where $K_{00}^{[1]}$ and $K_{00}^{[2]}$ are the eigenvalues of matrix (a) given by (39). First we search an invariant built only with **(5′)** and **(5″)**. We get $\mu_1+\mu_2=0$, $\mu_1K_{00}^{[1]}+\mu_2K_{00}^{[2]}+s=0$ with, for example, $\mu_1=1$, $\mu_2=-1$ and $s=K_{00}^{[2]}-K_{00}^{[1]}$. The invariant writes

$$I=\frac{a_{21}x+(K_{00}^{[1]}-a_{11})y}{(K_{00}^{[2]}-a_{22})x+a_{12}y}\exp(-\sqrt{(a_{11}-a_{22})^2+4a_{12}a_{21}})t.\tag{47}$$

However, we can have a first integral in addition to (47) without introducing new conditions. For that we build with **(5′)**, **(5″)**, and **(6)** obtaining $\mu_1+\mu_2+\mu_3=0$ and $\mu_1 K_{00}^{[1]}+\mu_2 K_{00}^{[2]}+s=0$. The interesting point is that we get twice the equation $\mu_1+\mu_2+\mu_3=0$. Now taking $\mu_1=K_{00}^{[2]}$ and $\mu_2=-K_{00}^{[1]}$ we obtain $s=0$ (and consequently a first integral). Having a first integral and an invariant the problem is now fully algebraic and, consequently, this system can be completely solved and the integration is as complete as possible. Note, however that this case does not pass the Painlevé test, a clear indication that passing the Painlevé test and getting explicit invariant, first integral or explicit solutions are two different approaches not always connected (although correlated in many cases).

C. Other cases

We exclude cases treated above built with both straight lines **(5)** and **(6)** and cases where $b_{12}=b_{22}$, $b_{21}=b_{11}$. Also we omit those invariants involving only one variable obtained when one equation is decoupled from the other one. As a result, 13 invariants are obtained: 8 build with 3 straight lines, 4 with 2 straight lines and 1 with 1 line. We denote the invariants with the names of the straight lines upon which they have been built. With three straight lines we have

(125), **(236)**, **(345)**, **(146)**, which are the generalization of invariant II of the LVS. **(125)** is LVS ($a_{12}=0$, $a_{21}=0$) with the three straight lines passing through one of the equilibrium point.

(126), **(235)**, **(346)**, **(145)**, which are the generalization of invariant III of the LVS. **(126)** is LVS with three straight lines forming a triangle.

With two straight lines, we have

(12), **(14)**, **(23)**, **(34)**, which are the generalization of the invariant I of the LVS (**(12)** is LVS). These four cases require the third condition $\Delta=0$.

A last case can be found with only one line. Obviously the cofactor must be a constant and the only interesting case is given by the straight line **(5)** with $b_{11}=b_{22}=0$. Taking $f=f_{10}x+f_{01}y$ and $K=K_{00}$ we obtain

$$f_{10}b_{12}+f_{01}b_{21}=0, \tag{48}$$

$$f_{10}(a_{11}-K_{00})+f_{01}a_{21}, \quad f_{10}a_{12}+f_{01}(a_{22}-K_{00})=0, \tag{49}$$

and the invariant is

$$I=(b_{21}x-b_{12}y)\exp[(a_{12}b_{21}-a_{22}b_{12})t/b_{12}], \tag{50}$$

with three conditions

$$b_{11}=b_{22}=0, \quad a_{11}-a_{22}=a_{21}\frac{b_{12}}{b_{21}}-a_{12}\frac{b_{21}}{b_{12}}. \tag{51}$$

which is invariant I in Ref. 11.

It is worth to note that in all cases (including the last one) each invariant straight line goes through two equilibrium points. We have seen also that many cases are generalization of the types I, II, and III of the LVS. One of the most interesting invariants (**(56)** since we need only 2 conditions) splits for the LVS in both invariants II and III. The introduction of the canonical form (36) through transformations (31) and (32) has greatly simplified the problem of expliciting invariants and conditions and has shown the very fundamental connection between LVS and QS. We can establish the following theorem.

Theorem 4: *A complex QS has a first integral, an integrating factor or a Darboux invariant formed by invariant straight lines in the following cases:*

(a) If $b_{11}(2b_{22}-b_{12})-b_{21}b_{22}=0$, $r_2=0$ *and* $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$, *then the straight lines* $f_1=b_{11}x+b_{22}y=0$ *and* $f_2=a_{11}a_{22}-a_{12}a_{21}+(a_{22}b_{11}-a_{21}b_{22})x+(a_{11}b_{22}-a_{12}b_{11})y=0$ *are algebraic solutions of the QS with cofactors* $K_1=b_{11}x+b_{22}y+a_{11}+a_{21}b_{22}/b_{11}$ *and* $K_2=b_{11}x+b_{22}y$, *respectively, and* $f_1^{\mu_1}f_2^{\mu_2}e^{st}$ *is a Darboux invariant for*

$\mu_1=-\mu_2$ and $\mu_2=sb_{11}/(a_{21}b_{22}+a_{11}b_{11})$. This corresponds to the integrable case **(56)**.

(b) *If $b_{21}=b_{11}$, $b_{12}=b_{22}$, $a_{11}a_{22}b_{11}b_{22}\neq 0$, and let us call $K_{00}^{[1]}$, $K_{00}^{[2]}$ the eigenvalues of matrix (a), then the straight lines $f_1=a_{21}x+(K_{00}^{[1]}-a_{11})y=0$ and $f_2=(K_{00}^{[2]}-a_{22})x+a_{12}y=0$ are algebraic solutions of the QS with cofactors $K_1=b_{11}x+b_{22}y+K_{00}^{[1]}$ and $K_2=b_{11}x+b_{22}y+K_{00}^{[2]}$, respectively, such that $f_1^{\lambda_1}f_2^{\lambda_2}$ is an integrating factor for $\lambda_1=(a_{11}+a_{22})/(2K_{00}^{[1]}-2K_{00}^{[2]})-3/2$ and $\lambda_2=-(a_{11}+a_{22})/(2K_{00}^{[1]}-2K_{00}^{[2]})-3/2$. Moreover $f_1^{\mu_1}f_2^{\mu_2}e^{st}$ is a Darboux invariant for $\mu_1=-s/(K_{00}^{[1]}-K_{00}^{[2]})$ and $\mu_2=-\mu_1$. This corresponds to the integrable case* **(5′5″)**.

(c) *If $b_{21}=b_{11}$, $b_{12}=b_{22}$, $a_{11}a_{22}b_{11}b_{22}\neq 0$, and let us call $K_{00}^{[1]}$, $K_{00}^{[2]}$ the eigenvalues of matrix (a), then the straight lines $f_1=a_{21}x+(K_{00}^{[1]}-a_{11})y=0$, $f_2=(K_{00}^{[2]}-a_{22})x+a_{12}y=0$ and $f_3=a_{11}a_{22}-a_{12}a_{21}+(a_{22}b_{11}-a_{21}b_{22})x+(a_{11}b_{22}-a_{12}b_{11})y=0$ are algebraic solutions of the QS with cofactors $K_1=b_{11}x+b_{22}y+K_{00}^{[1]}$, $K_2=b_{11}x+b_{22}y+K_{00}^{[2]}$ and $K_3=b_{11}x+b_{22}y$, respectively, and $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}$ is a first integral for $\mu_1=K_{00}^{[2]}$, $\mu_2=-K_{00}^{[1]}$ and $\mu_3=K_{00}^{[1]}-K_{00}^{[2]}$. This corresponds to the integrable case* **(5′5″6)**.

(d) *If $a_{12}=0$, $a_{21}=a_{22}b_{21}/b_{22}$, $r_2=0$* and *$a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$, then the straight lines $f_1=y=0$, $f_2=a_{11}+b_{11}x=0$ and $f_3=b_{11}x+(b_{22}-b_{12})y+a_{11}a_{22}=0$ are algebraic solutions of the QS with cofactors $K_1=b_{21}x+b_{22}y+a_{22}$, $K_2=b_{11}x+b_{12}y$ and $K_3=b_{11}x+b_{22}y$, respectively, such that $f_1^{\lambda_1}f_2^{\lambda_2}f_3^{\lambda_3}$ is an integrating factor for $\lambda_1=-(a_{11}+a_{22})/a_{22}$, $\lambda_2=b_{12}/(b_{22}-b_{12})$ and $\lambda_3=a_{11}/a_{22}+b_{22}/(b_{12}-b_{22})$. Moreover, $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}e^{st}$ is a Darboux invariant for $\mu_1=-s/a_{11}$, $\mu_2=sb_{11}/[a_{22}(b_{21}-b_{11})]$ and $\mu_3=s/a_{22}+sb_{12}/[a_{11}(b_{12}-b_{22})]$. This corresponds to the integrable case* **(236)**.

(e) *If $a_{12}=a_{11}b_{12}/b_{11}$, $a_{21}=a_{22}b_{21}/b_{22}$, $r_1=0$, and $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$ then there exist the straight lines $f_1=a_{11}+b_{11}x=0$, $f_2=a_{22}+b_{22}y=0$ and $f_3=a_{22}b_{11}x-a_{11}b_{22}y=0$ which are algebraic solutions of the QS with cofactors $K_1=b_{11}x+b_{12}y$, $K_2=b_{21}x+b_{22}y$, $K_3=b_{11}x+b_{22}y-(b_{12}-b_{22})a_{22}/b_{22}$, respectively, such that $f_1^{\lambda_1}f_2^{\lambda_2}f_3^{\lambda_3}$ is an integrating factor for $\lambda_1=a_{22}b_{11}/[a_{11}(b_{11}-b_{21})]-1$, $\lambda_2=b_{21}/(b_{11}-b_{21})$ and $\lambda_3=b_{11}(a_{11}+a_{22})/[a_{11}(b_{21}-b_{11})]$. Moreover, $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}e^{st}$ is a Darboux invariant of the QS for $\mu_1=-s/a_{22}$, $\mu_2=sb_{22}/[a_{11}(b_{22}-b_{12})]$ and $\mu_3=s/a_{22}+sb_{12}/[a_{11}(b_{12}-b_{22})]$. This corresponds to the integrable case* **(345)**.

(f) *If $a_{12}=0$,$a_{21}=a_{22}b_{21}/b_{22}$, $r_2=0$ and $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$, then the straight lines $f_1=x=0$, $f_2=a_{22}+b_{22}y=0$ and $f_3=(b_{11}-b_{21})x+b_{22}y+a_{11}a_{22}=0$ are algebraic solutions of the QS with cofactors $K_1=b_{11}x+b_{12}y+a_{11}$, $K_2=b_{21}x+b_{22}y+a_{22}$ and $K_3=b_{11}x+b_{22}y+a_{11}$, respectively, such that $f_1^{\lambda_1}f_2^{\lambda_2}f_3^{\lambda_3}$ is an integrating factor for $\lambda_1=-(a_{11}+a_{22})/a_{11}$, $\lambda_2=b_{21}/(b_{11}-b_{21})$ and $\lambda_3=b_{11}/(b_{21}-b_{11})+a_{22}/a_{11}$. Moreover, $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}e^{st}$ is a Darboux invariant for $\mu_1=-s/a_{11}$,$\mu_2=sb_{11}/[a_{22}(b_{21}-b_{11})]$ and $\mu_3=sb_{21}/[a_{22}(b_{21}-b_{11})]+s/a_{11}$. This corresponds to the integrable case* **(146)**.

(g) *If $a_{12}=a_{11}b_{12}/b_{11}$, $a_{21}=0$, $r_1=0$ and $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$, then the straight lines $f_1=y=0$, $f_2=a_{11}+b_{11}x=0$ and $f_3=(b_{21}-b_{11})x-(b_{12}-b_{22})y=0$ are algebraic solutions of the QS with cofactors $K_1=a_{22}+b_{21}x+b_{22}y$, $K_2=b_{11}x+b_{12}y$ and $K_3=a_{11}+b_{11}x+b_{22}y$, respectively, and $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}$ is a first integral for $\mu_1=b_{22}(a_{11}-a_{22})$,$\mu_2=-a_{11}b_{12}$ and $\mu_3=a_{22}b_{12}$. This corresponds to the integrable case* **(235)**.

(h) *If $a_{11}+a_{22}=0$, $a_{12}=a_{11}b_{12}/b_{11}$, $a_{21}=a_{22}b_{21}/b_{22}$ and $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\Delta\neq 0$, then there exist the straight lines $f_1=a_{11}+b_{11}x=0$, $f_2=a_{22}+b_{22}y=0$ and $f_3=a_{11}\Delta-b_{11}b_{22}(b_{21}-b_{11})x+b_{22}b_{11}(b_{12}-b_{22})y=0$ which are algebraic solutions of the QS with the cofactors $K_1=b_{11}x+b_{12}y$, $K_2=b_{21}x+b_{22}y$, $K_3=b_{11}x+b_{22}y$, such that $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}$ is a first integral for $\mu_1=b_{22}(b_{21}-b_{11})$,$\mu_2=b_{11}(b_{12}-b_{22})$ and $\mu_3=\Delta$. This corresponds to the integrable case* **(346)**.

(i) *If $a_{12}=0$, $a_{21}=a_{22}b_{21}/b_{22}$, $r_1=0$ and $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$, then the straight lines $f_1=x=0$, $f_2=a_{22}+b_{22}y=0$ and $f_3=(b_{21}-b_{11})x-(b_{12}-b_{22})y=0$ are alge-*

braic solutions of the QS with cofactors $K_1=a_{11}+b_{11}x+b_{12}y$, $K_2=b_{21}x+b_{22}y$ *and* $K_3=a_{22}+b_{11}x+b_{22}y$, *respectively, and* $f_1^{\mu_1}f_2^{\mu_2}f_3^{\mu_3}$ *is a first integral for* $\mu_1=-a_{22}b_{21}$, $\mu_2=b_{11}(a_{22}-a_{11})$ *and* $\mu_3=a_{11}b_{21}$. *This corresponds to the integrable case* **(145)**.

(j) *If* $a_{12}=0$, $a_{21}=a_{22}b_{21}/b_{22}$ *and* $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$, *then the straight lines* $f_1=x=0$ *and* $f_2=a_{22}+b_{22}y=0$ *are algebraic solutions of the QS with cofactors* $K_1=b_{11}x+b_{12}y+a_{11}$ *and* $K_2=b_{21}x+b_{22}y$, *respectively, such that with the additional condition* $r_1=0$, $f_1^{\lambda_1}f_2^{\lambda_2}$ *is an integrating factor for* $\lambda_1=-(a_{11}+a_{22})/a_{11}$ *and* $\lambda_2=(a_{22}b_{12}-2a_{11}b_{22})/(a_{11}b_{22})$. *On the other hand if the additional condition is* $\Delta=0$ *then* $f_1^{\mu_1}f_2^{\mu_2}e^{st}$ *is a Darboux invariant for* $\mu_1=-s/a_{11}$ *and* $\mu_2=sb_{12}/(a_{11}b_{22})$. *This corresponds to the integrable case* **(14)**.

(k) *If* $a_{12}=0$, $a_{21}=a_{22}b_{21}/b_{22}$, $\Delta=0$ *and* $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$ *then the straight lines* $f_1=y=0$ *and* $f_2=a_{11}+b_{11}x=0$ *are algebraic solutions of the QS with cofactors* $K_1=b_{21}x+b_{22}y+a_{22}$ *and* $K_2=b_{11}x+b_{12}y$, *respectively, such that with the additional condition* $r_1=0$, $f_1^{\lambda_1}f_2^{\lambda_2}$ *is an integrating factor for* $\lambda_2=-(a_{11}+a_{22})/a_{22}$ *and* $\lambda_1=(a_{11}b_{21}-2a_{22}b_{11})/(a_{22}b_{11})$. *On the other hand if the additional condition is* $\Delta=0$ *then* $f_1^{\mu_1}f_2^{\mu_2}e^{st}$ *is a Darboux invariant for* $\mu_1=-s/a_{22}$ *and* $\mu_2=sb_{22}/(a_{22}b_{12})$. *This corresponds to the integrable case* **(23)**.

(l) *If* $a_{12}=a_{11}b_{12}/b_{11}$, $a_{21}=a_{22}b_{21}/b_{22}$, $\Delta=0$ *and* $a_{11}a_{22}b_{11}b_{22}(b_{12}-b_{22})(b_{21}-b_{11})\neq 0$ *then the straight lines* $f_1=a_{11}+b_{11}x=0$ *and* $f_2=a_{22}+b_{22}y=0$ *are algebraic solutions of the QS with cofactors* $K_1=b_{11}x+b_{12}y$ *and* $K_2=b_{21}x+b_{22}y$, *respectively, such that* $f_1^{\mu_1}f_2^{\mu_2}$ *is a first integral for* $\mu_1=b_{22}$ *and* $\mu_2=-b_{12}$. *This corresponds to the integrable case* **(34)**.

(m) *If* $b_{11}=b_{22}=0$, $r_1=0$ *and* $b_{12}b_{21}\neq 0$, *then the straight line* $f=b_{21}x-b_{12}y=0$, *is an algebraic solution of the QS with cofactor* $K=a_{11}-a_{21}b_{12}/b_{21}$ *such that* $f_1^{\mu}e^{st}$ *is a Darboux invariant for* $\mu=(a_{12}b_{21}-a_{22}b_{12})/b_{12}$. *This corresponds to the integrable case* **(5)**.

A qualitative picture of the phase space portrait at a finite distance from the origin for each of the statements of Theorem 4 appears in Fig. 2. The figures concerning the invariants **(126)**, **(145)**, **(235)**, and **(346)** deserve a special comment. They suggest that the equilibrium point, which is not a summit of the triangle formed by the three straight lines is a center. Indeed the four cases are first integrals and if we show that this equilibrium point is a linear center, the possession of this first integral will confer a center status.

To check this linear center property we will have to discuss the real or complex nature of the roots of the second degree equation obtained by linearization of the system (36). While the relation between the coefficients of the QS, needed for the existence of the invariants or first integral remains equally valid for complex or real coefficients of (36), we will suppose real in this paragraph all the coefficients of (36).

Let x_0 and y_0 be the coordinates of an equilibrium point. The linearized equation for δx and δy with $x=x_0+\delta x$ and $y=y_0+\delta y$ writes

$$\begin{aligned}\dot{\delta x}&=(a_{11}+2b_{11}x_0+b_{12}y_0)\,\delta x+(a_{12}+b_{12}x_0)\,\delta y,\\ \dot{\delta y}&=(a_{21}+b_{21}y_0)\,\delta x+(a_{22}+2b_{22}y_0+b_{21}x_0)\,\delta y,\end{aligned}\tag{52}$$

which will describe a linear center if the two eigenvalues of the matrix are purely imaginary. As a consequence, two relations are necessary. The first concerns the trace of the matrix of the system. We must have

$$a_{11}+a_{22}+b_{12}y_0+b_{21}x_0+2(b_{11}x_0+b_{22}y_0)=0,\tag{53}$$

relation which can be written, taking into account the fact that x_0,y_0 are the coordinates of an equilibrium point

$$a_{12}(y_0/x_0)+a_{21}(x_0/y_0)=b_{11}x_0+b_{22}y_0.\tag{54}$$

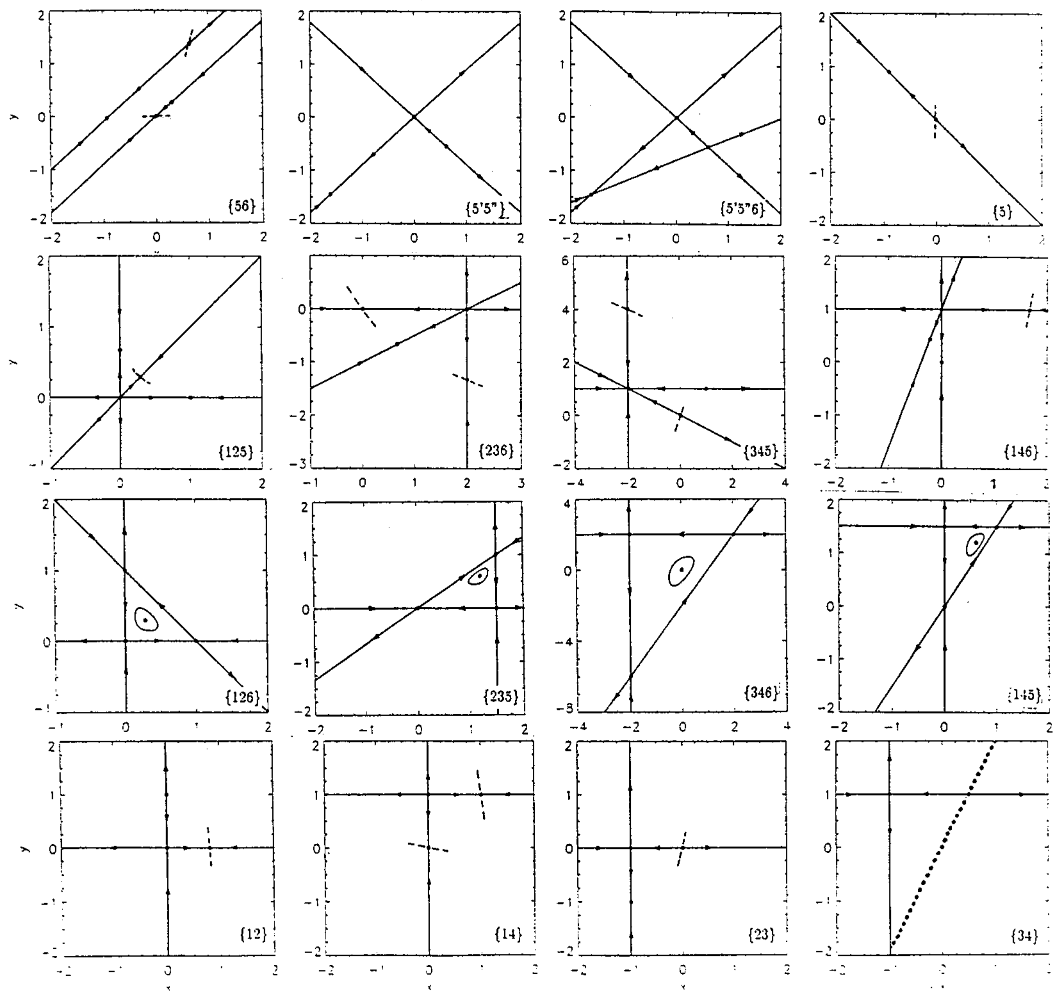

FIG. 2. Typical phase portraits for the QS having invariants or first integrals formed with straight lines.

The second relation is an inequality concerning the sign of the term under the square root to obtain a purely imaginary root. Taking into account (53), the eigenvalues are

$$\lambda_{1,2}=\sqrt{-D}, \tag{55}$$

where

$$D=\begin{vmatrix} a_{11}+2b_{11}x_0+b_{12}y_0, & a_{12}+b_{12}x_0 \\ a_{21}+b_{21}y_0, & a_{22}+2b_{22}y_0+b_{21}x_0 \end{vmatrix}$$

and the inequality to obtain a linear center is $D>0$. Now we show that the four cases mentioned above fulfill (54). We begin with **(346)** where the origin is the candidate to the status of a center. With $x_0=y_0=0$, (53) becomes $a_{11}+a_{22}=0$, which is precisely one of the conditions needed to obtain a first integral. If $a_{11}a_{22}-a_{12}a_{21}>0$ the origin is consequently a center. For **(126)** we are in the LVS case ($a_{12}=a_{21}=0$). Computing x_0 and y_0 and introducing it in (54) we check that this last relation is fulfilled. For values of the parameters such that $D>0$ this equilibrium point will be a center. Finally we consider **(145)** (**(235)** being symmetric). Taking into account the three conditions and doing a little algebra, we obtain for x_0 and y_0,

$$x_0=-a_{11}b_{22}/\Delta, \quad y_0=a_{11}b_{21}/\Delta. \tag{56}$$

Introducing (56) into (53), we see that this last relation is fulfilled and we have just to check that $D>0$.

Consequently the four equilibrium points of figures **(126)**, **(235)**, **(346)**, and **(145)** fulfill the relations needed to have the status of a center and are, indeed, centers if $D>0$.

V. CONCLUSION

Although some of the results presented here have been obtained before, we see that the Darboux method is more systematic and simplifies the obtention and resolution of the resulting algebraic—but nonlinear—equations. These equations are originated from the identification of the different terms after the introduction of ansatzes on the expression of the algebraic solutions (also called invariant curves). Each of these algebraic solutions brings its own conditions to exist. For bidimensional quadratic systems, three algebraic solutions are needed. The simplest are straight lines, each bringing one condition in the QS case. We need altogether three conditions except essentially in two cases. The LVS is the special quadratic system with $a_{12}=a_{21}=0$ and, consequently only one more condition is needed. These algebraic solutions are invariant straight lines connecting two equilibrium points. For the LVS the two axes provide the two first algebraic solutions and if the third one is a conic we need two conditions. The invariant conics also connect two equilibrium points. Both an increase of the dimensionality and the degree of the system increase quickly the number of needed conditions.

[1] H. T. Davis, *Introduction to Nonlinear Differential and Integral Equations* (Dover, New York, 1962), Chap. 6.
[2] L. Cairó and M. R. Feix, ''Families of invariants of the motion for the Lotka–Volterra equations: The linear polynomials family,'' J. Math. Phys. **33**, 2440–2455 (1992).
[3] J. R. Roth, ''Periodic small-amplitude solutions to Volterra's problem of two conflicting populations and their applications to the plasma continuity equations,'' J. Math. Phys. **10**, 1412–1414 (1969).
[4] W. A. Coppel, ''On a differential equation of boundary layer theory,'' Philos. Trans. R. Soc. London, Ser. A **253**, 101–136 (1960).
[5] J. W. Reyn, ''A bibliography of the qualitative theory of quadratic systems of differential equations in the plane,'' 3rd ed., Delft University of Technology, Report No. 94-02, 1994.
[6] T. Carleman, ''Application de la théorie des équations intégrales linéaires aux systémes d'équations différentielles nonlinéaires,'' Acta Math. **59**, 63–87 (1932).
[7] L. Cairó, M. R. Feix, D. D. Hua, S. Bouquet, and A. Dewisme, ''Hamiltonian method and invariant search for 2D quadratic systems,'' J. Phys. A **26**, 4371–4386 (1993).
[8] D. D. Hua, L. Cairó, and M. R. Feix, ''Time-independent invariants of motion for the quadratic system,'' J. Phys. A **26**, 7097–7114 (1993).
[9] G. Darboux, ''Mémoire sur les équations différentielles algébriques du premier ordre et du primier degre'' (Mélanges), Bull. Sci. Math. **60–96**, 123–144 (1978); **60–96**, 151–200 (1978).
[10] L. Cairó, J. Llibre, and M. R. Feix ''Integrability and algebraic solutions for planar polynomial differential systems with emphasis on the quadratic systems,'' Université d' Orleans preprint 97-03, 1997.
[11] D. D. Hua, L. Cairó, M. R. Feix, K. S. Govinder, and P. Leach, ''Connection between the existence of first integrals and the Painlevé property in two-dimensional Lotka–Volterra and quadratic systems,'' Proc. R. Soc. London **452**, 859–880 (1996).

Bäcklund transformation of partial differential equations from the Painlevé-Gambier classification. II. Tzitzéica equation

Robert Conte
Service de physique de l'état condensé, CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France

Micheline Musette
Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, B-1050 Brussel, België

A. Michel Grundland
Centre de recherches mathématiques, Université de Montréal, Case postale 6128, Succursale Centre ville, Montréal, Québec H3C 3J7, Canada

(Received 11 August 1998; accepted for publication 19 November 1998)

From the existing methods of singularity analysis only, we derive the two equations which define the Bäcklund transformation of the Tzitzéica equation. This is achieved by defining a truncation in the spirit of the approach of Weiss *et al.*, so as to preserve the Lorentz invariance of the Tzitzéica equation. If one assumes a third-order scattering problem, this truncation admits a unique solution, thus leading to a matrix Lax pair and a Darboux transformation. In order to obtain the Bäcklund transformation (BT), which is the main new result of this paper, one represents the Lax pair by an equivalent two-component Riccati pseudopotential. This yields two different BTs; the first one is a BT for the Hirota–Satsuma equation, while the second one is a BT for the Tzitzéica equation. One of the two equations defining the BT is the fifth ordinary differential equation of Gambier.
 [S0022-2488(99)01503-0]

I. INTRODUCTION

This paper is a follow-up of a previous article,[1] hereafter referred to as paper I, in which was defined a scheme, based on singularity analysis only, in order to find the auto-Bäcklund transformation (BT) of a given partial differential equation (PDE) when one exists. This general scheme was designed so as to also succeed when the *singular manifold method* of Weiss *et al.*[2,3] and its later improvements, see references in paper I, fail to provide this auto-BT.

After the Kaup–Kupershmidt equation, we consider the Tzitzéica equation

$$E(u) \equiv u_{xt} + \alpha e^{u} + a_2 e^{-2u} = 0, \quad \alpha a_2 \neq 0, \quad (\alpha, a_2)\,\text{const.} \tag{1}$$

Its invariance under the permutation (Lorentz transformation)

$$\mathcal{P}: (\partial_x, \partial_t) \rightarrow (\partial_t, \partial_x), \tag{2}$$

which does not exist for evolution equations like the one considered in paper I, must be present in all the expected results, and this will require some extra care. For the definitions, notation and procedure, we refer to paper I. This equation appears in several domains of mathematics and physics, which we summarize below.

(1) It was originally found in the field of geometry in 1907 by Georges Tzitzéica[4] who looked for surfaces on which the total curvature at each point is proportional to the fourth power of the distance from a fixed point to the tangent plane. One of the beautiful properties of this equation is the periodicity of its Laplace sequence,[5] with a period three. In the opinion of Tzitzéica however (Ref. 6, p. 255), ‘‘...l'intégration de l'équation (1) ...est très difficile.’’

(2) In one-dimensional ideal gas dynamics, when the equation of state is

$$P=\frac{\rho^3}{M^4}, \tag{3}$$

in which P, ρ, M denote the pressure, density and Lagrangian mass, the system of Euler's equations is reducible[7,8] to the Tzitzéica equation. This equation of state is identical, *mutatis mutandis*, to the geometrical definition of Tzitzéica. The geometrical link with the equation for affine spheres[6] has been recently rederived in the context of gas dynamics.[9]

(3) In magnetohydrodynamics, there exits a $2+1$-dimensional system which, after elimination of the magnetic field, becomes

$$r_x+s_y=0, \quad r_t-r_{xxx}+(rs)_x=0, \tag{4}$$

in which (r,s) are the components of the velocity on the x and y axes. The stationary reduction (independence on t) is equivalent to the so-called Hirota–Satsuma PDE,[10] which we define as

$$\mathrm{HS}(w)\equiv[w_{xxt}+(6/a)w_x w_t]_x=0, \quad a\neq 0 \tag{5}$$

[notation is $r=(6/a)w_y$, $s=-(6/a)w_x$, with a constant and y renamed t]. Its potential form

$$\mathrm{pHS}(w)\equiv w_{xxt}+(6/a)w_x w_t+F(t)=0, \quad a\neq 0, \tag{6}$$

in the particular case $F(t)=0$ (this restriction is important, see Sec. II), has a one-to-one correspondence[11] with the Tzitzéica equation, obtained by elimination of a_2 in Eq. (1)

$$\left(F(t)=0, e^u=\frac{2}{a\alpha}w_t\right)\Rightarrow\left(e^{-2u}(e^{2u}E(u))_x=\left(\frac{\mathrm{pHS}(w)}{w_t}\right)_t\right). \tag{7}$$

(4) The search for all equations of the type

$$u_{xt}+f(u)=0 \tag{8}$$

which admit an infinite number of integrals of motion led Dodd and Bullough[12,13] to the following finite list of admissible functions f (up to linear transformations on u): 0 (d'Alembert), u, e^u (Liouville), $\sin u$ or $\sinh u$ (sine- or shine-Gordon, not different in the complex plane), and $e^u - e^{-2u}$ (Tzitzéica). This infinity of integrals of motion for (1) was later found by Zhiber and Shabat.[14]

(5) The search of separating solutions of (8), i.e., of functions φ, X,T of one variable such that

$$u=\varphi(X(x)T(t)) \tag{9}$$

satisfies (8), gives rise[15] to a finite list of admissible f including the above mentioned list.

(6) In the nonlinear SL(3,R) σ-model

$$\mathbf{r}_{xt}+\rho^{-1}(\mathbf{r}_{xt},\mathbf{r}_x,\mathbf{r}_t)^{1/2}\mathbf{r}=0, \quad \rho \ \text{ constant}, \tag{10}$$

$$(\mathbf{r}_{xx},\mathbf{r}_x,\mathbf{r}_t)=(\mathbf{r}_{tt},\mathbf{r}_x,\mathbf{r}_t)=0, \tag{11}$$

in which boldface characters denote vectors in R^3, the field $u=(1/2)\mathrm{Log}(\mathbf{r}_{xt},\mathbf{r}_x,\mathbf{r}_t)$ satisfies the Tzitzéica equation.[16]

(7) Equation (1) is a reduction of the two-dimensional Toda lattice.[17]

The purpose of this paper is to derive all the integrability results of (1) (Lax pair, Darboux transformation and Bäcklund transformation) from singularity analysis *only*.

We handle at the same time all the equations of the type considered by Zhiber and Shabat:[14]

$$E(u) \equiv u_{xt} + \alpha e^{u} + a_1 e^{-u} + a_2 e^{-2u} = 0, \quad \alpha \neq 0, \tag{12}$$

i.e., the Liouville equation $a_1 = a_2 = 0$, the sine-Gordon equation $(a_1 \neq 0,\ a_2 = 0)$ and the Tzitzéica equation $(a_1 = 0,\ a_2 \neq 0)$. The results to be found are the auto-BT of each equation, and for Liouville the hetero-BT to the d'Alembert equation. We use the notation of Ref. 18.

The paper is organized as follows. In Sec. II, we recall the existing results, whether obtained from singularity analysis or from other methods, and we point out what is missing. In Sec. III, we recall the singularity structure of the Tzitzéica PDE and argue that, just like the Kaup–Kupershmidt equation in paper I, it possesses only one relevant family of movable singularities. In the next two sections, we successively choose the order two, then three, for the order of the unknown scattering problem. In Sec. IV, we thus give a very short singularity-based derivation of the hetero-BT between Liouville and d'Alembert equations; we also give a unified derivation of the auto-BT of the Liouville and sine-Gordon equations. In Sec. V, we assume an unknown Lax pair invariant under the Lorentz transformation (2) and define it in its equivalent projective Riccati representation; the truncation which it defines admits a unique solution, thus providing a matrix Lax pair and a Darboux transformation. The last section is divided into two subsections, each corresponding to one of the two possible eliminations leading to a Bäcklund transformation. In Sec. VI A, the first elimination provides the BT of the Hirota–Satsuma PDE (5). In Sec. VI B, the second elimination provides the expected result, namely the BT of the Tzitzéica equation (1). We write this BT as one second-order nonlinear ODE and one first-order PDE, which are both linearizable. The ODE part

$$y'' + 3yy' + y^3 + ry + q = 0, \tag{13}$$

in which q and r depend on a solution of (1) and on the Bäcklund parameter λ, belongs to the fifth equivalence class of Gambier.[19]

II. PREVIOUS RESULTS

Two Lax pairs of (1) are known. The first one has been given by Tzitzéica,[20]

$$-\tau_{xx} + U_x \tau_x + \sqrt{a_2}\lambda e^{-U} \tau_t = 0, \tag{14}$$

$$-\tau_{tt} + U_t \tau_t + \sqrt{a_2}\lambda^{-1} e^{-U} \tau_x = 0, \tag{15}$$

$$-\tau_{xt} - \alpha e^{U} \tau = 0, \tag{16}$$

in which the arbitrary complex constant λ is the spectral parameter [this ''pair'' is in fact a set of three scalar equations in the entire function τ for which the three commutators vanish whenever these three equations hold and U satisfies (1)]. The third-order traceless matrix Lax pair

$$L = \begin{pmatrix} -U_x/2 & 0 & -\sqrt{\alpha}\lambda^{-1} e^{U/2} \\ -\sqrt{a_2} e^{-U} & U_x/2 & 0 \\ 0 & \sqrt{\alpha}\lambda e^{U/2} & 0 \end{pmatrix}, \tag{17}$$

$$M = \begin{pmatrix} U_t/2 & -\sqrt{a_2} e^{-U} & 0 \\ 0 & -U_t/2 & -\sqrt{\alpha}\lambda^{-1} e^{U/2} \\ \sqrt{\alpha}\lambda e^{U/2} & 0 & 0 \end{pmatrix}, \tag{18}$$

$$L_t - M_x + [L, M] = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} E(U) \tag{19}$$

has been established by Mikhailov.[17]

Note that the system (14)–(16) is invariant under the involution[21,8]

$$(\tau,e^U,\lambda)\to\left(\frac{1}{\tau},-e^U-\frac{2}{\alpha}\frac{\tau_x\tau_t}{\tau^2},-\lambda\right), \tag{20}$$

which defines the Darboux transformation

$$e^u=-e^U-\frac{2}{\alpha}\frac{\tau_x\tau_t}{\tau^2} \tag{21}$$

between two different solutions u and U of (1).

Both the prolongation method[22] and the conditional symmetries method[23] allow us to retrieve the Lax pair of Mikhailov in its equivalent projective Riccati representation.

The Bäcklund transformation, according to the classical definition recalled in paper I (see also the fifth step of Sec. II in that paper), is a set of two PDEs in two solutions u and U of (1) such that the PDE in U resulting from the elimination of u is equivalent to $E(U)=0$.

It is worthwhile to underline the fact that the BT is *not* a couple made of a Lax pair and a Darboux transformation; it is the explicit result of the elimination of the solution of the Lax pair between them.

Such a BT has never been obtained for the Tzitzéica equation. All the classical authors (Tzitzéica,[21,24] A. Demoulin and H. Jonas[25]) considered the Lax pair and the DT and none of them gave the explicit form of the Bäcklund transformation. McLaughlin and Scott[26] proved that its type is certainly different from the well-known one of the sine-Gordon equation.

Gaffet was the first modern author to analyze the difficulty of finding the BT (section 5.1 of Ref. 8). His main conclusion is that obtaining the BT would be quite easy if one knew a *scalar* Lax pair whose commutator is equivalent to the PDE (1) and generates the one-soliton solution. Such a pair is given below, Eqs. (114) and (115).

By elimination of ∂_x (resp. ∂_t) between the three scalar linear PDEs (14)–(16), Gaffet obtained[8] the two linear third-order ODEs,

$$\tau_{xxx}-(U_{xx}+U_x^2)\tau_x+\alpha\sqrt{a_2}\lambda\tau=0, \tag{22}$$

$$\tau_{ttt}-(U_{tt}+U_t^2)\tau_t+\alpha\sqrt{a_2}\lambda^{-1}\tau=0, \tag{23}$$

but these two ODEs, taken alone, do not define a scalar Lax pair since their commutator is equivalent to (1) only *modulo* the triplet (14)–(16). Therefore, the elimination of τ between these two ODEs and the DT cannot yield the BT.

Sharipov and Yamilov[27] (see also subsequent papers, Refs. 28 and 29) eliminated τ between the DT (21) and three differential consequences of (14)–(16), to obtain three complicated PDEs, instead of two, without the property that two of them constitute a BT obeying the above definition.

For completion, let us also mention the existence of a nonlinear superposition formula between four different solutions of (1).[25,30,31]

From the correspondence (7) with the Hirota–Satsuma equation, one might think of transposing the well-established results for that equation (Lax pair,[32] DT and BT,[33] all easy to find by singularity analysis[34]) to get the BT of the Tzitzéica equation. Unfortunately, this is impossible because the correspondence (7) restricts the potential form of the Hirota–Satsuma equation to $F(t)=0$. This impossibility is illustrated by the solution to the Hirota–Satsuma equation[35]

$$w_x=-a(P(x-f(t))+Q(x+f(t))), \tag{24}$$

$$w_t=a(P(x-f(t))-Q(x+f(t))f'(t), \quad f'\not\equiv 0, \tag{25}$$

$$P''(z)=6P(z)^2+kz+A, \quad Q''(z)=6Q(z)^2+kz+B, \tag{26}$$

$$F(t)=af'(t)(2kf(t)-A+B), \tag{27}$$

in which k,A,B are arbitrary constants and f is an arbitrary function. Both P and Q are either a first Painlevé function $(k\neq 0)$ or a Weierstrass elliptic function $(k=0)$. This also defines a solution to the Tzitzéica equation only under the two additional restrictions $F(t)=0$, $f''(t)=0$. This implies $k=0$, $A=B$ and thus forbids the first Painlevé function, only leaving the elliptic solution [formula (104) in Ref. 18].

One must therefore turn to the study of the singularity structure of the Tzitzéica equation.

If the underlying scattering problem is not of second order, then the usual truncation of Weiss (as well as its two-singular manifold extension[18,36]) is not adapted and it should generically fail. Indeed, the determining equations generated by the Weiss truncation [condition of identical vanishing of a polynomial in $(\varphi-\varphi_0)^{-1}$] are equivalent[37] to those generated by the assumption that the field ψ defined by

$$\frac{\varphi_x}{\varphi-\varphi_0}-\frac{\varphi_{xx}}{2\varphi_x}=\frac{1}{\chi}=\frac{\psi_x}{\psi} \tag{28}$$

satisfies the *second-order* linear system

$$\psi_{xx}+\frac{S}{2}\psi=0, \tag{29}$$

$$\psi_t+C\psi_x-\tfrac{1}{2}C_x\psi=0, \tag{30}$$

in which S and C are invariant under a homography on φ and linked by the cross-derivative condition

$$X\equiv S_t+CS_x+2C_xS+C_{xxx}=0. \tag{31}$$

As briefly explained in paper I formulas (26) and (27), when the scattering problem is of third order, it is inconsistent to perform the Weiss truncation, then to insert in each of the generated determining equations an assumption for a third-order Lax pair. For the same reason, it is also inconsistent to introduce the so-called *singular manifold equation*. Although such an inconsistency may still provide the full result for some ''robust'' equations (Boussinesq,[38] Sawada–Kotera,[39] Hirota–Satsuma[34]), there do exist equations for which it leads to a failure, and the Kaup–Kupershmidt equation studied in paper I is one of them.

The correct procedure[40] when one assumes a scattering problem with an order $n\geqslant 2$ is to generate determining equations [(27) in paper I] by expansion on a $(n-1)$-component basis.

The Tzitzéica equation is one of the equations which, up to now, has escaped the Weiss truncation, as found by three previous attempts in his formalism:

(1) the truncation of e^{-u} has no solution at all,[41,42] not even the one-soliton;
(2) the truncation of e^{u} (Ref. 18) does provide some particular solutions (in particular a two-wave solution) but, when the involved computations are finished, it should probably fail to provide a Lax pair;
(3) the truncation of e^{u} contains a particular solution[43] linking φ to the pair (τ,τ'), in which τ satisfies the Tzitzéica triplet (14)–(16) and τ' its adjoint.[20,25]

To summarize, the results of singularity analysis which are missing are the following.

(1) Definition of a truncation whose general solution is unique and defines a single Lax pair and a Darboux transformation.
(2) The BT as exactly two DEs, namely one nonlinear ODE (necessarily linearizable) plus one first-order PDE, with a commutator equivalent to the Tzitzéica equation.

III. SINGULARITY STRUCTURE OF THE ZHIBER–SHABAT EQUATIONS

Equation (12) always possesses the family

$$e^{u} \sim -(2/\alpha)\varphi_x\varphi_t(\varphi-\varphi_0)^{-2}, \quad \text{indices } (-1,2), \quad \mathcal{D}=(2/\alpha)\partial_x\partial_t. \tag{32}$$

For the Liouville equation, this is the only family. For Sine-Gordon (SG), there also exists the family

$$e^{-u} \sim (2/a_1)\varphi_x\varphi_t(\varphi-\varphi_0)^{-2}, \quad \text{indices } (-1,2), \quad \mathcal{D}=-(2/a_1)\partial_x\partial_t, \tag{33}$$

which reflects the invariance of SG under $u \to -u + \text{Log}(-a_1/\alpha)$.

For the Tzitzéica equation, there also exists the family

$$e^{-u} \sim \sqrt{(1/a_2)\varphi_x\varphi_t}(\varphi-\varphi_0)^{-1}, \quad \text{indices } (-1,2). \tag{34}$$

However, the existence of the single family for the Hirota–Satsuma equation (6)

$$w \sim a\varphi_x(\varphi-\varphi_0)^{-1}, \quad \text{indices } (-1,1,6), \quad \mathcal{D}=a\partial_x, \tag{35}$$

together with the link (7), proves that this second family (34) is irrelevant and must not be considered. This conclusion is confirmed by the negative result of Weiss[41,42] obtained when performing a truncation on e^{-u}.

All the truncations will accordingly take the form

$$e^{u} = \mathcal{D}\,\text{Log}\,\tau + e^{U}, \quad E(u)=0, \quad \mathcal{D}=(2/\alpha)\partial_x\partial_t, \tag{36}$$

without assuming that U should satisfy $E(U)=0$.

Remark: In paper I, the similar assumption $E(U)=0$ is, also, unnecessary for the equation of Kaup–Kupershmidt, since the condition $E(U)=0$ comes out of the truncation.

IV. ASSUMPTION OF A SECOND-ORDER SCATTERING PROBLEM

We are going to derive, by singularity analysis only, first the general solution of the Liouville equation and its hetero-BT to the d'Alembert equation and second the second-order Lax pair and the auto-BT of both the Liouville and sine-Gordon equations.

A. Hetero-BT between Liouville and d'Alembert equations

The special form of the Liouville equation allows the assumption (36) to be integrated twice to yield

$$u = -2\log\tau + V, \quad E(u) = \sum_{j=0}^{2} E_j \tau^{j-2} = 0, \tag{37}$$

in which nothing is assumed on V. The three determining equations are then quite simple,

$$E_0 \equiv 2\tau_x\tau_t + \alpha e^{V} = 0, \tag{38}$$

$$E_1 \equiv \tau_{xt} = 0, \tag{39}$$

$$E_2 \equiv V_{xt} = 0, \tag{40}$$

and their general solution depends on two arbitrary functions of one variable,

$$\tau = f(x) + g(t), \tag{41}$$

$$e^{V} = -\frac{2}{\alpha}\tau_x\tau_t = -\frac{2}{\alpha}f'(x)g'(t), \tag{42}$$

$$e^u = -\frac{2}{\alpha}\frac{\tau_x\tau_t}{\tau^2} = -\frac{2}{\alpha}\frac{f'(x)g'(t)}{(f(x)+g(t))^2}, \tag{43}$$

$$e^U = \tau^{-2}e^V + \frac{2}{\alpha}\frac{\tau_x\tau_t}{\tau^2} = 0. \tag{44}$$

Thus, the two fields u and V are the general solution of, respectively, the Liouville and d'Alembert equations. The hetero-BT between these two equations is provided by the elimination of f and g between (42), (43) and the x- and t-derivatives of (37):

$$(u-v)_x = \alpha\lambda e^{(u+v)/2}, \tag{45}$$

$$(u+v)_t = -2\lambda^{-1}e^{(u-v)/2}, \tag{46}$$

in which v is another solution of d'Alembert equation defined by

$$e^v = (\lambda\tau_t)^{-2}e^V = -\frac{2}{\alpha}\lambda^{-2}\frac{f'(x)}{g'(t)}. \tag{47}$$

Remark: When performing the truncation (36), Tamizhmani and Lakshmanan[44] already found $e^U=0$, $\tau_{xt}=0$ as a *particular* solution, while the above truncation (37) proves it to be the *general* solution. Another difference between the two truncations is the presence of a field V in (37), which allows one to find in addition the hetero-BT between the Liouville and d'Alembert equations.

B. Auto-BT of Liouville and sine-Gordon equations

One takes here $a_2=0$ and either $a_1=0$ (Liouville) or, by a linear change on variable u, $a_1=-\alpha$ (sine-Gordon). Let us define a unique truncation for both equations.

For sine-Gordon, the two equivalent families provide the two pieces of information

$$e^u = \mathcal{D}\,\mathrm{Log}\,\tau_1 + e^{U_1}, \qquad e^{-u} = \mathcal{D}\,\mathrm{Log}\,\tau_2 + e^{-U_2}, \tag{48}$$

in which each τ_j satisfies some second-order Lax pair and U_1, U_2 are two *a priori* unknown quantities. Consequently, the field

$$Y = \frac{\tau_1}{\tau_2} \tag{49}$$

satisfies a Riccati system. The condition that Y be a homographic transform of χ and vanish as χ when $\varphi-\varphi_0$ goes to zero imposes the existence of two nonzero quantities (A,B) such that[34,36]

$$Y^{-1} = B(\chi^{-1}+A), \quad AB\neq 0; \tag{50}$$

in which the Riccati system satisfied by the expansion variable Y is easily deduced from the canonical one satisfied by χ:

$$\chi_x = 1 + \frac{S}{2}\chi^2, \tag{51}$$

$$\chi_t = -C + C_x\chi - \frac{CS+C_{xx}}{2}\chi^2, \tag{52}$$

with S and C linked by (31). The elimination of (τ_1,τ_2) between (48) and (49) leads to

$$e^u - e^{-u} = \mathcal{D} \operatorname{Log} Y + e^{U_1} - e^{U_2}, \quad E(u) = 0, \tag{53}$$

which can be integrated twice to yield[18,36]

$$u = -2 \operatorname{Log} Y + W, \quad E(u) = 0. \tag{54}$$

For Liouville, if we assume the most general form for the unknown second-order Lax pair, we arrive at an assumption identical to (54).

Accordingly, in both cases, the truncation is

$$u = -2 \operatorname{Log} Y + W, \quad Y^{-1} = B(\chi^{-1} + A), \quad E(u) = \sum_{j=0}^{4} E_j(W,A,B,S,C) Y^{j-2} = 0, \tag{55}$$

in which nothing is imposed on W [it is to avoid any confusion with (37) that we use different symbols V and W]. This generates six determining equations: $E_j = 0$ and (31), in which B and W only contribute by the product $B^2 e^W$. They are solved as usual by ascending values of j:

$$E_0 : B^2 e^W = \frac{2}{\alpha} C, \tag{56}$$

$$E_1 : A = -\tfrac{1}{2} (\operatorname{Log} C)_x, \tag{57}$$

$$E_2 \equiv 0, \tag{58}$$

$$E_3 : S = -F(x) + \frac{C_x^2}{2C^2} - \frac{C_{xx}}{C}, \tag{59}$$

$$E_4 : C C_{xt} - C_x C_t + F(x) C^3 + a_1 \alpha F(x)^{-1} C = 0, \tag{60}$$

$$X : a_1 F'(x) = 0, \tag{61}$$

in which F is a function of integration. For sine-Gordon, $F(x)$ must be a constant

$$F(x) = 2\lambda^2. \tag{62}$$

In the Liouville case, for which the truncation imposes no restriction on $F(x)$, let us also require that $F(x)$ be a constant. Then, for both equations, Log C is proportional to a second solution $\tilde{U}$ of the PDE

$$C = \frac{\alpha}{2} \lambda^{-2} e^{\tilde{U}}, \quad E(\tilde{U}) = 0, \tag{63}$$

and one has obtained the Darboux transformation

$$u = -2 \operatorname{Log} y + \tilde{U}, \quad y = \lambda B Y, \tag{64}$$

in which y satisfies the Riccati system

$$y_x = \lambda + \tilde{U}_x y - \lambda y^2, \tag{65}$$

$$y_t = -\frac{\alpha}{2} \lambda^{-1} e^{\tilde{U}} - \frac{a_1}{2} \lambda^{-1} e^{-\tilde{U}} y^2. \tag{66}$$

The second-order matrix Lax pair (Refs. 45 and 46 in the Liouville case) results from the linearization $y=\psi_1/\psi_2$:

$$(\partial_x - L)\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=0,\quad L=\begin{pmatrix}\tilde{U}_x/2 & \lambda\\ \lambda & -\tilde{U}_x/2\end{pmatrix}, \tag{67}$$

$$(\partial_t - M)\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=0,\quad M=\begin{pmatrix}0 & -(\alpha/2)\lambda^{-1}e^{\tilde{U}}\\ (a_1/2)\lambda^{-1}e^{-\tilde{U}} & 0\end{pmatrix}. \tag{68}$$

The auto-BT (classical for sine-Gordon, Ref. 46 for Liouville) results from the substitution $y=e^{-(u-\tilde{U})/2}$ into (65) and (66):

$$(u+\tilde{U})_x=-4\lambda \sinh\frac{u-\tilde{U}}{2}, \tag{69}$$

$$(u-\tilde{U})_t=\lambda^{-1}(\alpha e^{(u+\tilde{U})/2}+a_1 e^{-(u+\tilde{U})/2}). \tag{70}$$

The ODE part of the BT is a Riccati equation, and the link between the entire function τ and the solution ψ of a *scalar* Lax pair is $\tau=\psi$ [see (120) and (121)].

For sine-Gordon, the Riccati system (65) and (66) is invariant under the involutions (denoting $y=\lambda\tau/\tau_x$)

$$(Y,U,\lambda)\to(Y^{-1},U-2\,\mathrm{Log}\,Y,-\lambda),(Y^{-1},-U,\lambda), \tag{71}$$

$$(Y,U,\lambda,u,\partial_x,\partial_t)\to(e^{-U}Y,-U,-\alpha/(2\lambda),u,\partial_t,\partial_x), \tag{72}$$

$$(Y,e^U,\lambda)\to(Y^{-1},e^{-U}+\mathcal{D}\,\mathrm{Log}\,\tau,-\lambda). \tag{73}$$

This defines the two usual forms of the DT of sine-Gordon:

$$e^u=e^{-sU}+\mathcal{D}\,\mathrm{Log}\,\tau, \tag{74}$$

$$u=sU-2\,\mathrm{Log}\,Y, \tag{75}$$

in which s can take either sign ±1.

C. The case of the Tzitzéica equation

We have already summarized in Sec. II the consequences of the assumption of a second-order scattering problem. Although, in our opinion, this question is not settled yet, we will not comment any more on it.

An additional computational complication arises from the existence of a single family. This indeed forbids the integration of the operator $\mathcal{D}$ twice in order to transform the assumption (36) into one with a form similar to either (37) or (54).

V. ASSUMPTION OF A THIRD-ORDER SCATTERING PROBLEM

There only remains the Tzitzéica equation.

The only Lax pair invariant under (2) is the matrix pair in the basis (τ_x,τ_t,τ),

$$(\partial_x - L)\begin{pmatrix}\tau_x\\ \tau_t\\ \tau\end{pmatrix}=0,\quad L=\begin{pmatrix}f_1 & f_2 & f_3\\ g_1 & g_2 & g_3\\ 1 & 0 & 0\end{pmatrix}, \tag{76}$$

$$(\partial_t - M)\begin{pmatrix}\tau_x\\ \tau_t\\ \tau\end{pmatrix} = 0, \quad M = \begin{pmatrix} g_1 & g_2 & g_3\\ h_1 & h_2 & h_3\\ 0 & 1 & 0\end{pmatrix}, \tag{77}$$

in which the nine coefficients f_j, g_j, h_j, $j=1,2,3$, are functions to be determined. We define it in the equivalent projective Riccati components (Y_1, Y_2) with the property $Y_{1,t} = Y_{2,x}$:

$$Y_1 = \frac{\tau_x}{\tau}, \quad Y_2 = \frac{\tau_t}{\tau}, \tag{78}$$

$$Y_{1,x} = -Y_1^2 + f_1 Y_1 + f_2 Y_2 + f_3, \tag{79}$$

$$Y_{2,x} = -Y_1 Y_2 + g_1 Y_1 + g_2 Y_2 + g_3, \tag{80}$$

$$Y_{1,t} = -Y_1 Y_2 + g_1 Y_1 + g_2 Y_2 + g_3, \tag{81}$$

$$Y_{2,t} = -Y_2^2 + h_1 Y_1 + h_2 Y_2 + h_3. \tag{82}$$

The truncation defined by

$$e^u = \mathcal{D} \operatorname{Log} \tau + e^U, \tag{83}$$

$$E(u) \equiv \sum_{k=0}^{3} \sum_{l=0}^{3-k} E_{kl}(f_j, g_j, h_j, U) Y_1^k Y_2^l = 0 \tag{84}$$

generates ten determining equations,

$$\forall k,l : E_{kl}(f_j, g_j, h_j, U) = 0, \tag{85}$$

in U and the nine unknown coefficients. One must add to them the six equations $X_j = 0$ defined by the cross-derivative conditions

$$(Y_{1,x})_t - (Y_{1,t})_x = X_0 + X_1 Y_1 + X_2 Y_2 = 0, \tag{86}$$

$$(Y_{2,x})_t - (Y_{2,t})_x = X_3 + X_4 Y_1 + X_5 Y_2 = 0. \tag{87}$$

During their resolution, one first proves that the product $f_2 h_1$ cannot vanish (otherwise a_2 would be zero). This makes the 16 equations algebraically independent and equivalent to the 15 differential relations

$$f_{j,t}, g_{j,x}, g_{j,t}, h_{j,x}, g_{j,xt} = P(\{f_k, g_k, h_k\}, \quad k=1,2,3), \quad j=1,2,3, \tag{88}$$

with P polynomials whose coefficients depend on U, U_x, U_t, U_{xt}, plus the single algebraic relation

$$E_{00} \equiv a_2 - \frac{4}{\alpha^2}(g_3 + g_1 g_2 + (\alpha/2) e^U)^2 = 0. \tag{89}$$

They are solved successively as [equations are numbered as in (88) and (89)]

$g_{3,xt} - (g_{3,x})_t$, : $E(U) = 0$,
$g_{1,x} - g_{2,t}$: $\exists g_0(x,t) : g_1 = g_{0,t}$, $g_2 = g_{0,x}$
$g_{2,t}$: $g_3 = -\alpha e^U - g_{0,x} g_{0,t} - g_{0,xt}$
E_{00} : $\exists f_0(x,t) \neq 0 : f_2 = \sqrt{a_2} W^{-1} f_0$, $h_1 = \sqrt{a_2} W^{-1} f_0^{-1}$,
notation $W = e^U + (2/\alpha) g_{0,xt}$
$g_{2,x}$: $f_3 = -\sqrt{a_2} W^{-1} f_0 g_{0,t} - f_1 g_{0,x} - g_{0,x}^2 + g_{0,xx}$,

$$
\begin{aligned}
g_{3,x} &: \quad f_1 = W_x/W + 2g_{0,x},\\
f_{2,t} &: \quad h_2 = W_t/W + 2g_{0,t} - f_{0,t}/f_0,\\
h_{1,x} &: \quad f_{0,x} = 0,\\
g_{1,t} &: \quad h_3 = g_{0,t}(-W_t/W + f_{0,t}/f_0 - g_{0,t}) + g_{0,tt} - \sqrt{a_2}W^{-1}g_{0,x}/f_0,\\
g_{3,t} &: \quad f_{0,t} = 0,\\
h_{2,x} &: \quad g_{0,xt} = 0.
\end{aligned}
\tag{90}
$$

The irrelevant arbitrary function g_0 reflects the freedom in the definition (83) of τ and can be absorbed by redefining τ as τe^{-g_0}. Thus the solution is unique: the field U must satisfy the Tzitzéica PDE, and f_0 is an arbitrary nonzero complex constant λ. Accordingly, one has obtained a Lax pair and a Darboux transformation. This Lax pair is the rewriting in matrix form of the scalar triplet; it reads in its equivalent projective Riccati representation

$$Y_{1,x} = -Y_1^2 + U_x Y_1 + \sqrt{a_2}\lambda e^{-U} Y_2, \tag{91}$$

$$Y_{2,x} = -Y_1 Y_2 - \alpha e^U, \tag{92}$$

$$Y_{1,t} = -Y_1 Y_2 - \alpha e^U, \tag{93}$$

$$Y_{2,t} = -Y_2^2 + U_t Y_2 + \sqrt{a_2}\lambda^{-1} e^{-U} Y_1, \tag{94}$$

with cross-derivative conditions proportional to the Tzitzéica equation

$$(Y_{1,x})_t - (Y_{1,t})_x = Y_1 E(U), \quad (Y_{2,x})_t - (Y_{2,t})_x = Y_2 E(U). \tag{95}$$

It admits by construction the involution (20), equivalent to

$$(\tau, e^U, \lambda) \to (1/\tau, e^U + \mathcal{D}\,\mathrm{Log}\,\tau, -\lambda). \tag{96}$$

This defines two equivalent writings for the DT, (21) and (36).

Remark: Knowing these results, one can also write this DT[17,22,23] as a difference of the two fields $u-U$ in terms of the two components of a projective Riccati pseudopotential

$$u = U + \mathrm{Log}(-2\lambda^2 y_1 y_2 - 1), \quad y_j = \alpha^{-1/2}\lambda^{-1} Y_j e^{-U/2}, \tag{97}$$

in a quite similar manner to the DT of Liouville and sine-Gordon (64).

In order to find the BT, one must now eliminate one of the two equivalent projective components, and this defines two possible, different, eliminations.

VI. BÄCKLUND TRANSFORMATION

A. First elimination: BT of the Hirota–Satsuma equation

One takes Y_2 from (91) and substitutes it into the three remaining equations, which results in

$$Y_2 = (Y_{1,x} + Y_1^2 - U_x Y_1) e^U / (\sqrt{\alpha_2}\lambda), \tag{98}$$

$$\mathrm{ODE} \equiv Y_{1,xx} + 3Y_1 Y_{1,x} + Y_1^3 - e^{-U}(e^U)_{xx} Y_1 + \alpha\sqrt{a_2}\lambda = 0, \tag{99}$$

$$\mathrm{PDE} \equiv Y_{1,t} + e^U((Y_1 Y_{1,x} + Y_1^3) - Y_1^2 U_x)/(\sqrt{a_2}\lambda) + \alpha e^U = 0, \tag{100}$$

$$(94) \equiv -Y_1 E(U) - \frac{e^U Y_1}{\sqrt{a_2}\lambda}\mathrm{ODE} + (2Y_1 - U_x + \partial_x)\mathrm{PDE} = 0, \tag{101}$$

$$[\mathrm{ODE}, \mathrm{PDE}] = (Y_{1,xx})_t - (Y_{1,t})_{xx} = Y_1 (e^{2U} E(U))_x. \tag{102}$$

Only two of them are functionally independent, as shown by relation (101), but the commutator (102) of Eqs. (99) and (100) shows that this elimination fails to generate the auto-BT of the Tzitzéica equation.

However, it does provide another result, which we now derive. The ODE (99) belongs to the classification of Gambier—this is the number 5, see Eq. (13). It is linearizable by the transformation $Y_1 = \partial_x \operatorname{Log} \psi$ into the third-order linear ODE (22) considered by Gaffet, with the relation $\tau = \psi$ between the two entire functions. This transformation also linearizes the PDE (100) (the linearization is best seen on the equivalent combination $Y_{1,t} - Y_{2,x}$ of the projective Riccati system), and the resulting linear system.

$$\tau_{xxx} - (U_{xx} + U_x^2)\tau_x + \sqrt{a_2}\alpha\lambda\tau = 0, \tag{103}$$

$$-\sqrt{a_2}\lambda\tau_t + e^U\tau_{xx} - U_x e^U \tau_x = 0, \tag{104}$$

which cannot be a scalar Lax pair of the Tzitzéica equation, is, in fact, the scalar Lax pair of the Hirota–Satsuma equation (5):

$$\tau_{xxx} - (6/a)w_x\tau_x + \Lambda\tau = 0, \tag{105}$$

$$\Lambda\tau_t - (2/a)w_t\tau_{xx} + (2/a)w_{xt}\tau_x = 0, \tag{106}$$

under the change of variables (7).

Thus, the underlying Gambier equation for the Hirota–Satsuma equation is the fifth one (13). Let us recall that the Painlevé analysis of the Hirota–Satsuma equation[34] directly provides this result.

B. Second elimination: BT of the Tzitzéica equation

The second elimination is to take Y_1 from (92) and to substitute it into the three remaining equations:

$$Y_1 = -(Y_{2,x} + \alpha e^U)/Y_2, \tag{107}$$

$$\mathrm{ODE} \equiv Y_2 Y_{2,xx} - 2Y_{2,x}^2 - (U_x Y_2 + 3\alpha e^U)Y_{2,x} + \sqrt{a_2}\lambda e^{-U} Y_2^3 - \alpha^2 e^{2U} = 0, \tag{108}$$

$$\mathrm{PDE} \equiv Y_2 Y_{2,t} + Y_2^3 - U_t Y_2^2 + \sqrt{a_2}\alpha\lambda^{-1}(1 + e^{-U}Y_{2,x}) = 0, \tag{109}$$

$$(93) \equiv E(U) - \sqrt{a_2}\lambda^{-1} e^{-U} Y_2^{-2}\mathrm{ODE} + (-\alpha e^U Y_2^{-1} + \partial_x)\mathrm{PDE} = 0, \tag{110}$$

$$[\mathrm{ODE},\mathrm{PDE}] = (Y_{2,xx})_t - (Y_{2,t})_{xx} = (3\alpha e^U + U_x Y_2 + 3Y_{2,x} - Y_2\partial_x)E(U). \tag{111}$$

Only two of them are functionally independent, as shown by the relation (110), and the vanishing of the commutator (111) of Eqs. (108) and (109) is equivalent to the vanishing of the Tzitzéica equation for U. This elimination therefore generates the auto-BT of the Tzitzéica equation by the substitution

$$Y_2 = (\alpha/2)\int (e^u - e^U)\mathrm{d}x \tag{112}$$

into (108) and (109).

This BT, contrary to all the previously proposed ones, obeys the definition given in Sec. II.

The nonlinear ODE (108) again belongs to the equivalence class of the fifth Gambier equation (13), and its linearization

$$Y_2^{-1} = -\alpha^{-1} e^{-U}\partial_x \operatorname{Log}(e^U\psi) \tag{113}$$

transforms the two equations (108) and (109) into the third-order scalar Lax pair of the Gelfand and Dikii types:

$$\mathcal{L}\psi \equiv \psi_{xxx} + (2U_{xx} - U_x^2)\psi_x + ((2U_{xx} - U_x^2)_x/2 + \sqrt{a_2}\alpha\lambda)\psi = 0, \tag{114}$$

$$\mathcal{M}\psi \equiv \psi_t + \sqrt{a_2}(\alpha\lambda)^{-1}e^{-2U}(\psi_{xx} + U_x\psi_x + U_{xx}\psi) + \left(U_t + \int (\alpha e^U + a_2 e^{-2U})\mathrm{d}x\right)\psi = 0, \tag{115}$$

$$[\mathcal{L},\mathcal{M}] = 3E\partial_x^2 + (2(e^U)_x E + E_x)\partial_x + (e^U)_x E_x + (3U_{xx} - U_x^2)E, \quad E = E(U). \tag{116}$$

Remarks:

(1) The results of these two different eliminations show that the scalar Lax pair deduced from a matrix one crucially depends on the choice of the scalar component.

(2) In the third step of the method presented in paper I, which we have not strictly followed here, one had to assume a link $\partial_x\partial_t$ Log $\tau = f(\psi)$ between the two entire functions τ and ψ. This link is here *a posteriori* provided by the linearizing formula (113) and the Riccati equation (93); this is the invertible transformation

$$\tau = e^{-U}(e^U\psi)_x, \quad \psi = -(1/\alpha)e^{-U}\tau_t. \tag{117}$$

(3) The number three, rather than two, of linear scalar equations (14)–(16) in τ has been the source of all the difficulties to obtain the correct BT. Its reduction to two, i.e., to a true scalar Lax pair, is achieved by taking the parametric representation of the third one (16),

$$\exists\Psi: \tau = e^{-U}\Psi_x, \quad \tau_t = -\alpha\Psi, \tag{118}$$

and by inserting it in the two other equations. After the rescaling $\Psi = e^U\psi$, Eq.(14) becomes exactly (114), but Eq. (15) becomes (115) without its last term $(\int E(U)dx)\psi$; this one is recovered by adding a suitable term $e\psi$ with $e_t = 0$ so as to match the cross-derivative condition $(\psi_{xxx})_t = (\psi_t)_{xxx}$ with the condition $E(U) = 0$. Note that such a matching term also arises in the sine-Gordon case[36] when converting the Riccati form (65) and (66) into a one-component scalar Lax pair:

$$y^{-1} = \lambda^{-1}\partial_x \operatorname{Log} \tau, \tag{119}$$

$$L_1\tau \equiv \tau_{xx} + U_x\tau_x - \lambda\tau = 0, \tag{120}$$

$$L_2\tau \equiv \tau_t + \frac{\alpha}{2\lambda}e^U\tau_x - \left(\int (\alpha e^U + a_1 e^{-U})\mathrm{d}x\right)\tau = 0, \tag{121}$$

$$[L_1, L_2] = -E(U)\partial_x. \tag{122}$$

(4) Denoting (τ_1, τ_2, τ_3) the wave vectors of the Lax pair of Mikhailov (17) and (18), the correspondence is

$$\tau_1 = (\sqrt{\alpha}\lambda)^{-1}e^{-U/2}\tau_t, \quad \tau_2 = (\sqrt{\alpha}\lambda)^{-1}e^{-U/2}\tau_x, \quad \tau_3 = \tau. \tag{123}$$

VII. CONCLUSION

The results of this paper are twofold: the explicit analytic expression for the Bäcklund transformation, and the proof that there indeed exists an application of the truncation method able to yield all the integrability results. This reinforces the power of the singularity approach to PDEs initiated by Weiss *et al.*

TABLE I. The two writings of the Darboux transformation (on e^u or on u) of the Liouville, sine-Gordon and Tzitzéica equations $E(u) \equiv u_{xt} + f(u) = 0$. The fields u and U satisfy the same PDE and are linked by the auto-BT. The symbol $\mathcal{D}$ is $(2/\alpha)\partial_x\partial_t$. The Liouville case has two entries, one for its hetero-BT with the d'Alembert equation for v, and one for its auto-BT. For sine-Gordon, $s = \pm 1$ and the coordinate z is either x or t.

$f(u)$	DT on e^u	DT on u	Lax pair
αe^u	$e^u = \mathcal{D}\,\text{Log}\,\tau$	$u = -2\,\text{Log}((\lambda\tau_t)^{-1}\tau) + v$	$\tau_{xt} = 0,\ v_{xt} = 0$
αe^u	$e^u = \mathcal{D}\,\text{Log}\,y + e^U$	$u = -2\,\text{Log}\,y + U$	Riccati(y), $y = \dfrac{\psi_1}{\psi_2}$
$\alpha(e^u - e^{-u})$	$e^u = \mathcal{D}\,\text{Log}\,\tau + e^{-sU}$	$u = -2\,\text{Log}\,Y + sU$	Riccati(y), $\dfrac{\lambda}{y} = \dfrac{\tau_z}{\tau}$
$\alpha e^u + a_2 e^{-2u}$	$e^u = \mathcal{D}\,\text{Log}\,\tau + e^U$	$u = \text{Log}(-2\lambda^2 y_1 y_2 - 1) + U$	Proj. Riccati $\left(\dfrac{\tau_x}{\tau}, \dfrac{\tau_t}{\tau}\right)$

Table I represents a unified picture of the Darboux transformation of each of the three equations of the Zhiber–Shabat group.

ACKNOWLEDGMENTS

It is a pleasure to thank Bernard Gaffet and Wolfgang Schief for numerous discussions. RC and MM acknowledge the financial support of the Tournesol Grant No. T95/004. MM acknowledges the financial support extended within the framework of the IUAP Contract No. P4/08 funded by the Belgian government and the support of CEA. AMG acknowledges the financial support of CEA and of research grants from NSERC of Canada and the Fonds FCAR du Gouvernement du Québec.

[1] M. Musette and R. Conte, ''Bäcklund transformation of partial differential equations from the Painlevé-Gambier classification, I. Kaup-Kupershmidt equation,'' J. Math. Phys. **39**, 5617–5630 (1998).
[2] J. Weiss, M. Tabor, and G. Carnevale, ''The Painlevé property for partial differential equations,'' J. Math. Phys. **24**, 522–526 (1983).
[3] J. Weiss, ''The Painlevé property for partial differential equations. II: Bäcklund transformation, Lax pairs, and the Schwarzian derivative,'' J. Math. Phys. **24**, 1405–1413 (1983).
[4] G. Tzitzéica, ''Sur une nouvelle classe de surfaces,'' C. R. Acad. Sci. Paris **144**, 1257–1259 (1907).
[5] G. Darboux, *Leçons sur la théorie générale des surfaces et les applications géométriques du calcul infinitésimal*, vol. III (Gauthier-Villars, Paris, 1894). Reprinted, *Théorie générale des surfaces* (Chelsea, New York, 1972). Reprinted (Gabay, Paris, 1993).
[6] G. Tzitzéica, *Géométrie différentielle projective des réseaux* (Gauthier-Villars, Paris, 1924).
[7] B. Gaffet, ''SU(3) symmetry of the equations of unidimensional gas flow, with arbitrary entropy distribution,'' J. Math. Phys. **25**, 245–255 (1984).
[8] B. Gaffet, ''A class of 1-d gas flows soluble by the inverse scattering transform,'' Physica D **26**, 123–139 (1987).
[9] W. Schief and C. Rogers, ''The affinsphären equation. Moutard and Bäcklund transformations,'' Inverse Probl. **10**, 711–731 (1994).
[10] R. Hirota and J. Satsuma, ''*N*-soliton solutions of model equations for shallow water waves,'' J. Phys. Soc. Jpn Lett. **40**, 611–612 (1976).
[11] R. Hirota and A. Ramani, ''The Miura transformations of Kaup's equation and of Mikhailov's equation,'' Phys. Lett. **76A**, 95–96 (1980).
[12] R. K. Dodd and R. K. Bullough, ''Bäcklund transformations for the sine-Gordon equations,'' Proc. R. Soc. London, Ser. A **351**, 499–523 (1976).
[13] R. K. Dodd and R. K. Bullough, ''Polynomial conserved densities for the sine-Gordon equations,'' Proc. R. Soc. London, Ser. A **352**, 481–503 (1977).
[14] A. V. Zhiber and A. B. Shabat, ''Klein-Gordon equations with a nontrivial group,'' Dokl. Akad. Nauk SSSR **247**, 1103–1107 (1979) [English: Sov. Phys. Dokl. **24**, 607–609 (1979)].
[15] A. M. Grundland and E. Infeld, ''A family of nonlinear Klein-Gordon equations and their solutions,'' J. Math. Phys. **33**, 2498–2503 (1992).
[16] V. V. Nesterenko, ''Nonlinear σ model for the Dodd-Bullough equation,'' Teoreticheskaya i Matematicheskaya Fizika **58** 192–199 (1984) [English: Theor. Math. Phys. **58**, 126–131 (1984)]
[17] A. V. Mikhailov, ''Integrability of a two-dimensional generalization of the Toda chain,'' Pis'ma Zh. Eksp. Teor. Fiz. **30**, 443–448 (1979). [English: JETP Lett. **30**, 414–418 (1979)]; ''The reduction problem and the inverse scattering method,'' Physica D **3**, 73-117 (1981).

[18] M. Musette and R. Conte, ''The two-singular manifold method, I. Modified KdV and sine-Gordon equations,'' J. Phys. A **27**, 3895–3913 (1994).
[19] B. Gambier, ''Sur les équations différentielles du second ordre et du premier degré dont l'intégrale générale est à points critiques fixes,'' Thèse, Paris (1909); Acta Math. **33**, 1–55 (1910).
[20] G. Tzitzéica, ''Sur une nouvelle classe de surfaces,'' Rendiconti del Circolo Matematico di Palermo **25**, 180–187 (1908).
[21] G. Tzitzéica, ''Sur une nouvelle classe de surfaces,'' C. R. Acad. Sci. Paris **150**, 955–956 (1910)
[22] Yang Huan-xiong and Li You-quan, ''Prolongation approach to Bäcklund transformation of Zhiber–Mikhailov–Shabat equation,'' J. Math. Phys. **37**, 3491–3497 (1996).
[23] I. Ayari and A. M. Grundland, ''Conditional symmetries for *k*th order partial differential equations,'' preprint CRM-2488 (1997).
[24] G. Tzitzéica, ''Sur une nouvelle classe de surfaces,'' C. R. Acad. Sci. Paris **150**, 1227–1229 (1910)
[25] H. Jonas, ''Sopra una classe di trasformazioni asintotiche, applicabili in particolare alle superficie la cui curvatura è proporzionale alla quarta potenza della distanza del piano tangente da uno punto fisso,'' Annali di Mat. pura ed applicata **30**, 223–255 (1921).
[26] D. W. McLaughlin and A. C. Scott, ''A restricted Bäcklund transformation,'' J. Math. Phys. **14**, 1817–1828 (1973).
[27] R. A. Sharipov and R. I. Yamilov, ''Bäcklund transformation and the construction of the integrable boundary-value problem for the equation $u_{xx}-u_{tt}=e^{u}-e^{-2u}$,'' Trans. Inst. Math. BNC, URO AN SSSR, Ufa **66**, 1–9 (1991).
[28] A. Yu. Boldin, S. S. Safin, and R. A. Sharipov, ''On an old article of Tzitzéica and the inverse scattering method,'' J. Math. Phys. **34**, 5801–5809 (1993).
[29] S. S. Safin and R. A. Sharipov, ''Bäcklund autotransformation for the equation $u_{xt}=e^{u}-e^{-2u}$,'' Teoreticheskaya i Matematicheskaya Fizika **95**, 146–159 (1993) [English: Theor. Math. Phys. **95**, 462–470 (1993)].
[30] Y. V. Brezhnev, ''Darboux transformation and some multi-phase solutions of the Dodd–Bullough–Tzitzéica equation: $U_{xt}=e^{U}-e^{-2U}$,'' Phys. Lett. A **211**, 94–100 (1996).
[31] W. Schief, ''Self-dual Einstein spaces via a permutability theorem for the Tzitzéica equation,'' Phys. Lett. A **223**, 55–62 (1996).
[32] M. Ito, ''An extension of nonlinear evolution equations of the K-dV (mK-dV) type to higher orders,'' J. Phys. Soc. Jpn. **49**, 771–778 (1980).
[33] R. Hirota, ''Direct methods in soliton theory,'' in Solitons, edited by R. K. Bullough and P. J. Caudrey, Springer Topics in Current Physics (Springer Verlag, Berlin, 1980), pp. 157–176.
[34] M. Musette and R. Conte, ''Algorithmic method for deriving Lax pairs from the invariant Painlevé analysis of nonlinear partial differential equations,'' J. Math. Phys. **32**, 1450–1457 (1991).
[35] P. A. Clarkson and E. L. Mansfield, ''On a shallow water wave equation,'' Nonlinearity **7**, 975–1000 (1994).
[36] A. Pickering, ''The singular manifold method revisited,'' J. Math. Phys. **37**, 1894–1927 (1996).
[37] R. Conte, ''Universal invariance properties of Painlevé analysis and Bäcklund transformation in nonlinear partial differential equations,'' Phys. Lett. A **134**, 100–104 (1988).
[38] J. Weiss, ''The Painlevé property and Bäcklund transformations for the sequence of Boussinesq equations,'' J. Math. Phys. **26**, 258–269 (1985).
[39] J. Weiss, ''On classes of integrable systems and the Painlevé property,'' J. Math. Phys. **25**, 13–24 (1984).
[40] M. Musette and R. Conte, ''Solitary waves and Lax pairs from polynomial expansions of nonlinear differential equations,'' in *Nonlinear Evolution Equations and Dynamical Systems*, edited by M. Boiti, L. Martina, and F. Pempinelli (World Scientific, Singapore, 1992), pp. 161–170.
[41] J. Weiss, ''Bäcklund transformation and the Painlevé property,'' J. Math. Phys. **27**, 1293–1305 (1986).
[42] R. Conte, ''Invariant Painlevé analysis of partial differential equations,'' Phys. Lett. A **140**, 383–390 (1989).
[43] W. Schief, ''The Tzitzéica equation: a Bäcklund transformation interpreted as truncated Painlevé expansion,'' J. Phys. A **29**, 5153–5155 (1996).
[44] K. M. Tamizhmani and M. Lakshmanan, ''Linearization and Painlevé property of Liouville and Cheng equations,'' J. Math. Phys. **27**, 2257–2258 (1986).
[45] V. A. Andreev, ''Application of the inverse scattering method to the equation $\sigma_{xt}=e^{\sigma}$,'' Teoreticheskaya i Matematicheskaya Fizika **29**, 213–220 (1976) [English: Theor. Math. Phys. **29**, 1027–1032 (1976)].
[46] Z. Popowicz, ''The gauge field copies and Bäcklund transformation,'' J. Phys. A **20**, 2685–2691 (1987).

On the Groenewold–Van Hove problem for $\mathbf{R}^{2n}$

Mark J. Gotay[a)]
Department of Mathematics, University of Hawaii,
2565 The Mall, Honolulu, Hawaii 96822

(Received 28 September 1998; accepted for publication 16 December 1998)

We discuss the Groenewold–Van Hove problem for $\mathbf{R}^{2n}$, and completely solve it when $n=1$. We rigorously show that there exists an obstruction to quantizing the Poisson algebra of polynomials on $\mathbf{R}^{2n}$, thereby filling a gap in Groenewold's original proof. Moreover, when $n=1$ we determine the largest Lie subalgebras of polynomials which can be consistently quantized, and explicitly construct all their possible quantizations. © *1999 American Institute of Physics.* [S0022-2488(99)00304-7]

I. INTRODUCTION

In 1946 Groenewold[1] presented a remarkable result which states that one cannot consistently quantize the Poisson algebra of all polynomials in the positions q^i and momenta p_i on $\mathbf{R}^{2n}$ as symmetric operators on some Hilbert space $\mathcal{H}$, subject to the requirement that the q^i and p_i be irreducibly represented. Van Hove[2] subsequently refined Groenewold's result. Thus it is *in principle* impossible to quantize—by *any* means—every classical observable on $\mathbf{R}^{2n}$, or even every polynomial observable, in a way consistent with Schrödinger quantization (which, according to the Stone–Von Neumann theorem, is the import of the irreducibility requirement on the q^i and p_i). At most one can consistently quantize certain Lie subalgebras of observables, for instance polynomials which are at most quadratic, or observables which are affine functions of the momenta.

This is not quite the end of the story, however; there are two loose ends which need to be tied up. The first is that there is a technical gap in Groenewold's proof.[3] This gap has been filled in Ref. 2 (see also Ref. 4) by means of a certain functional analytic assumption. Although ''small,'' this gap is nevertheless vexing, and its elimination in this manner is not entirely satisfactory. Second, in the absence of such a polynomial quantization, it is important to determine the largest Lie subalgebras of polynomials that can be consistently quantized along with their quantizations. While some results are known along these lines, this program has not yet been fully carried out.

In this paper we consider the Groenewold–Van Hove problem for $\mathbf{R}^{2n}$. We present two variants of Groenewold's theorem (''strong'' and ''weak''); the weak one is the version that Groenewold actually proved, while the strong one is the result referred to above. We then show that the strong version follows from the weak one *without* introducing extra hypotheses. Thus we fill the gap in Groenewold's proof. Moreover, when $n=1$ we determine the largest quantizable Lie subalgebras of polynomials and explicitly construct all their possible quantizations.

To make the presentation self-contained, we include a detailed discussion of previous work on the Groenewold–Van Hove problem.

II. BACKGROUND

Let $P(2n)$ denote the Poisson algebra of polynomials on $\mathbf{R}^{2n}$ with Poisson bracket

$$\{f,g\}=\sum_{k=1}^{n}\left[\frac{\partial f}{\partial p_k}\frac{\partial g}{\partial q^k}-\frac{\partial g}{\partial p_k}\frac{\partial f}{\partial q^k}\right].$$

a)Electronic mail: gotay@math.hawaii.edu

$P(2n)$ contains several distinguished Lie subalgebras: The *Heisenberg algebra*

$$\mathrm{h}(2n)=\mathrm{span}\{1,q^i,p_i|i=1,\ldots,n\};$$

the *symplectic algebra*

$$\mathrm{sp}(2n,\mathbf{R})=\mathrm{span}\{q^iq^j,q^ip_j,p_ip_j|i,j=1,\ldots,n\};$$

the *extended* (or *inhomogeneous*) *symplectic algebra*

$$\mathrm{hsp}(2n,\mathbf{R})=\mathrm{span}\{1,q^i,p_j,q^iq^j,q^ip_j,p_ip_j|i,j=1,\ldots,n\},$$

which is the semidirect product of $\mathrm{h}(2n)$ with $\mathrm{sp}(2n,\mathbf{R})$; and the *coordinate* (or *position*) *algebra*

$$C(2n)=\left\{\sum_{i=1}^{n} f^i(q)p_i+g(q)\right\},$$

where f^i and g are polynomials. All of these will play an important role in our development.

Let $P^k(2n)$ denote the subspace of polynomials of degree at most *k*, and $P_k(2n)$ the subspace of homogeneous polynomials of degree *k*. Then $P^1(2n)=\mathrm{h}(2n)$, $P_2(2n)=\mathrm{sp}(2n,\mathbf{R})$, and $P^2(2n)=\mathrm{hsp}(2n,\mathbf{R})$. When *n* is fixed, we simply write $P=P(2n)$, etc.

We now state what it means to ''quantize'' a Lie algebra of polynomials on $\mathbf{R}^{2n}$. Throughout, the Heisenberg algebra $\mathrm{h}(2n)$ is regarded as a ''basic algebra of observables.''[5]

Definition 1: Let $\mathcal{O}$ be a Lie subalgebra of $P(2n)$ containing the Heisenberg algebra $\mathrm{h}(2n)$. A *quantization* of $\mathcal{O}$ is a linear map $\mathcal{Q}$ from $\mathcal{O}$ to the linear space $\mathrm{Op}(D)$ of symmetric operators which preserve a fixed dense domain *D* in some separable Hilbert space $\mathcal{H}$, such that for all $f,g\in\mathcal{O}$,

(Q1) $\mathcal{Q}(\{f,g\})=i/\hbar[\mathcal{Q}(f),\mathcal{Q}(g)]$,

(Q2) $\mathcal{Q}(1)=I$,

(Q3) If the Hamiltonian vector field X_f of *f* is complete, then $\mathcal{Q}(f)$ is essentially self-adjoint on *D*,

(Q4) $\mathcal{Q}$ represents $\mathrm{h}(2n)$ irreducibly, and

(Q5) *D* contains a dense set of separately analytic vectors for the standard basis of $\mathcal{Q}(\mathrm{h}(2n))$.

We briefly comment on these conditions; a full exposition along with detailed motivation is given in Ref. 5.

Condition (Q1) is Dirac's famous ''Poisson bracket→commutator'' rule; here $\hbar$ is Planck's reduced constant. The second condition reflects the fact that if an observable *f* is a constant *c*, then the probability of measuring $f=c$ is one regardless of which quantum state the system is in. Regarding (Q3), we remark that in contradistinction with Van Hove,[2] we do not confine our considerations to only those classical observables whose Hamiltonian vector fields are complete. Rather than taking the point of view that ''incomplete'' classical observables cannot be quantized, we simply do not demand that the corresponding quantum operators be essentially self-adjoint (''e.s.a.'').

(Q4) and (Q5) emphasize the fundamental role of the Heisenberg algebra. The technical condition (Q5) guarantees the integrability of the Lie algebra representation $\mathcal{Q}(\mathrm{h}(2n))$ on *D*.[6] [There do exist nonintegrable representations of the Heisenberg algebra;[7] however, none of them seem to have physical significance. (Q5) thus serves to eliminate these ''spurious'' representations.] By virtue of the Stone–Von Neumann theorem, (Q5) along with the irreducibility criterion (Q4) imply that $\mathcal{Q}(\mathrm{h}(2n))$ is unitarily equivalent to a restriction of the *Schrödinger quantization* $d\Pi$*:*

$$q^i\mapsto q^i,\quad p_j\mapsto -i\hbar\,\partial/\partial q^j,\quad \text{and}\quad 1\mapsto I \tag{1}$$

on the Schwartz space $\mathcal{S}(\mathbf{R}^n)\subset L^2(\mathbf{R}^n)$. Indeed, by (Q5) $\mathcal{Q}(\mathrm{h}(2n))$ can be integrated to a representation τ of the Heisenberg group $\mathrm{H}(2n)$ which, according to (Q4), is irreducible. The Stone–Von Neumann theorem then states that this representation of $\mathrm{H}(2n)$ is unitarily equivalent to the Schrödinger representation Π, and hence $\tau=U\Pi U^{-1}$ for some unitary map $U:L^2(\mathbf{R}^n)\rightarrow\mathcal{H}$. Consequently,[8] $\mathcal{Q}(f)=U\overline{d\Pi(f)}U^{-1}\restriction D$ for all $f\in\mathrm{h}(2n)$, where the bar denotes closure. It now follows from (1), the invariance of the domain D, and Sobolev's lemma that $U^{-1}D\subseteq\mathcal{S}(\mathbf{R}^n)$, so that $U^{-1}\mathcal{Q}U$ is the restriction of $d\Pi$ to $U^{-1}D$.

Finally, there is a sixth criterion that a quantization must satisfy in general, viz., that $\mathcal{Q}$ be faithful when restricted to the given basic algebra of observables.[5] In the case of the Heisenberg algebra, however, this follows automatically by virtue of (Q1) and (Q2).

III. THE WEAK NO-GO THEOREM

In the next two sections we argue that there are no quantizations of $P(2n)$. Extensive discussions can be found in Refs. 1–4 and 9–13. We shall state the main results for $\mathbf{R}^{2n}$ but, for convenience, usually prove them only for $n=1$. The proofs for higher dimensions are immediate generalizations of these.

We begin by observing that there *does* exist a quantization $d\varpi$ of $\mathrm{hsp}(2n,\mathbf{R})$. It is given by the familiar formulas

$$d\varpi(q^i)=q^i,\quad d\varpi(1)=I,\quad d\varpi(p_j)=-i\hbar\frac{\partial}{\partial q^j},$$

$$d\varpi(q^iq^j)=q^iq^j,\quad d\varpi(p_ip_j)=-\hbar^2\frac{\partial^2}{\partial q^i\partial q^j},\tag{2}$$

$$d\varpi(q^ip_j)=-i\hbar\left(q^i\frac{\partial}{\partial q^j}+\frac{1}{2}\delta^i_j\right),\tag{3}$$

on the domain $\mathcal{S}(\mathbf{R}^n)\subset L^2(\mathbf{R}^n)$. Properties (Q1)–(Q3) are readily verified. (Q4) follows automatically since the restriction of $d\varpi$ to P^1 is just the Schrödinger representation. For (Q5) we recall that the Hermite functions $h_{k_1\cdots k_n}(q^1,\ldots,q^n)=h_{k_1}(q^1)\cdots h_{k_n}(q^n)$, where

$$h_k(q)=e^{q^2/2}\frac{d^k}{dq^k}e^{-q^2}$$

for $k=0,1,2,\ldots$, form a dense set of separately analytic vectors for $d\varpi(P^1)$. As these functions are also separately analytic vectors for $d\varpi(P_2)$,[14] the operator algebra $d\varpi(P^2)$ is integrable to a unique representation ϖ of the universal cover $\widetilde{\mathrm{HSp}}(2n,\mathbf{R})$ of the extended (or inhomogeneous) symplectic group $\mathrm{HSp}(2n,\mathbf{R})$ (thereby justifying our notation "$d\varpi$").[15] ϖ is known as the "extended metaplectic representation;" detailed discussions of it may be found in Refs. 10 and 13.

We call $d\varpi$ the "extended metaplectic quantization." It has the following crucial property.

Proposition 1: The extended metaplectic quantization is the *unique* quantization of $\mathrm{hsp}(2n,\mathbf{R})$ which exponentiates to a unitary representation of $\widetilde{\mathrm{HSp}}(2n,\mathbf{R})$.

By "unique," we mean up to unitary equivalence and restriction of representations.

Proof: Suppose $\mathcal{Q}$ were another such quantization of $\mathrm{hsp}(2n,\mathbf{R})$ on some domain D in a Hilbert space $\mathcal{H}$. Then $\mathcal{Q}(\mathrm{hsp}(2n,\mathbf{R}))$ can be integrated to a representation τ of $\widetilde{\mathrm{HSp}}(2n,\mathbf{R})$, and (Q4) implies that τ, when restricted to $\mathrm{H}(2n)\subset\widetilde{\mathrm{HSp}}(2n,\mathbf{R})$, is irreducible. The Stone–Von Neumann theorem then states that this representation of $\mathrm{H}(2n)$ is unitarily equivalent to the Schrödinger representation, and hence $\tau=U\varpi U^{-1}$ for some unitary map $U:L^2(\mathbf{R}^n)\rightarrow\mathcal{H}$.[16] Consequently, $\mathcal{Q}(f)=U\overline{d\varpi(f)}U^{-1}\restriction D$ for all $f\in \mathrm{hsp}(2n,\mathbf{R})$. Arguing as in the discussion following Definition 1, we see that $U^{-1}\mathcal{Q}U$ is in fact the *restriction* of $d\varpi$ to $U^{-1}D\subseteq \mathcal{S}(\mathbf{R}^n)$. □

The existence of an obstruction to quantization now follows from Theorem 2.

Theorem 2 (Weak No-Go Theorem): The extended metaplectic quantization of P^2 cannot be extended beyond P^2 in P.

Since P^2 is a maximal Lie subalgebra of P,[17] (Q1) implies that any quantization which extends $d\varpi$ must be defined on all of P. Thus we may restate this as: *There exists no quantization of P which reduces to the extended metaplectic quantization on* P^2.

Proof: Let $\mathcal{Q}$ be a quantization of P which extends the metaplectic quantization of P^2. As noted previously, we may assume that the domain $D\subseteq \mathcal{S}(\mathbf{R}^n)$. We will show that a contradiction arises when cubic polynomials are considered.

Take $n=1$. By inspection of (1)–(3) we see that the ''Von Neumann rules''

$$\mathcal{Q}(q^2)=\mathcal{Q}(q)^2,\quad \mathcal{Q}(p^2)=\mathcal{Q}(p)^2, \tag{4}$$

$$\mathcal{Q}(qp)=\tfrac{1}{2}(\mathcal{Q}(q)\mathcal{Q}(p)+\mathcal{Q}(p)\mathcal{Q}(q)), \tag{5}$$

hold. These in turn lead to higher degree Von Neumann rules.[9,10]

Lemma 1: For all real-valued polynomials r,

$$\mathcal{Q}(r(q))=r(\mathcal{Q}(q)),\quad \mathcal{Q}(r(p))=r(\mathcal{Q}(p)),$$

$$\mathcal{Q}(r(q)p)=\tfrac{1}{2}[r(\mathcal{Q}(q))\mathcal{Q}(p)+\mathcal{Q}(p)r(\mathcal{Q}(q))],$$

$$\mathcal{Q}(qr(p))=\tfrac{1}{2}[\mathcal{Q}(q)r(\mathcal{Q}(p))+r(\mathcal{Q}(p))\mathcal{Q}(q)].$$

Proof: We illustrate this for $r(q)=q^3$. The other rules follow similarly using induction. Now $\{q^3,q\}=0$ whence by (Q1) we have $[\mathcal{Q}(q^3),\mathcal{Q}(q)]=0$. Since also $[\mathcal{Q}(q)^3,\mathcal{Q}(q)]=0$, we may write $\mathcal{Q}(q^3)=\mathcal{Q}(q)^3+T$ for some operator T which (weakly) commutes with $\mathcal{Q}(q)$. We likewise have using (4)

$$[\mathcal{Q}(q^3),\mathcal{Q}(p)]=-i\hbar\,\mathcal{Q}(\{q^3,p\})=3i\hbar\,\mathcal{Q}(q^2)=3i\hbar\,\mathcal{Q}(q^2)=[\mathcal{Q}(q)^3,\mathcal{Q}(p)]$$

from which we see that T commutes with $\mathcal{Q}(p)$ as well. Consequently, T also commutes with $\mathcal{Q}(q)\mathcal{Q}(p)+\mathcal{Q}(p)\mathcal{Q}(q)$. But then from (5),

$$\begin{aligned}\mathcal{Q}(q^3)=\frac{1}{3}\mathcal{Q}(\{pq,q^3\})&=\frac{i}{3\hbar}[\mathcal{Q}(pq),\mathcal{Q}(q^3)]\\ &=\frac{i}{3\hbar}\left[\frac{1}{2}(\mathcal{Q}(q)\mathcal{Q}(p)+\mathcal{Q}(p)\mathcal{Q}(q)),\mathcal{Q}(q)^3+T\right]\\ &=\frac{i}{6\hbar}[\mathcal{Q}(q)\mathcal{Q}(p)+\mathcal{Q}(p)\mathcal{Q}(q),\mathcal{Q}(q)^3]=\mathcal{Q}(q)^3.\end{aligned}$$ ▽

With this lemma in hand, it is now a simple matter to prove the no-go theorem. Consider the classical equality

$$\tfrac{1}{9}\{q^3,p^3\}=\tfrac{1}{3}\{q^2p,p^2q\}.$$

Quantizing and then simplifying this, the formulas in Lemma 1 give

$$\mathcal{Q}(q)^2\mathcal{Q}(p)^2-2i\hbar\,\mathcal{Q}(q)\mathcal{Q}(p)-\tfrac{2}{3}\hbar^2I$$

for the lhs, and

$$\mathcal{Q}(q)^2\mathcal{Q}(p)^2-2i\hbar\,\mathcal{Q}(q)\mathcal{Q}(p)-\tfrac{1}{3}\hbar^2 I$$

for the rhs, which is a contradiction. □

IV. THE STRONG NO-GO THEOREM

In Groenewold's paper[1] a stronger result was claimed: His assertion was that there is no quantization of P, period. This is *not* what Theorem 2 states. For if $\mathcal{Q}$ is a quantization of P, then while of course $\mathcal{Q}(P^1)$ must coincide with Schrödinger quantization, it is not obvious that $\mathcal{Q}$ need be the extended metaplectic quantization when restricted to P^2. Referring to Proposition 1, the problem is that $\mathcal{Q}(P^2)$ is not *a priori* integrable; (Q5) only guarantees that $\mathcal{Q}(P^1)$ can be integrated. From a different point of view, the problem lies in *deducing* the relations (2) and (3) or, equivalently, the Von Neumann rules ([4]) and (5) from the quantization axioms (Q1)–(Q5) and the properties of the extended symplectic algebra alone.

Van Hove[2] supplied an extra assumption guaranteeing the integrability of $\mathcal{Q}(P^2)$, which in particular implies: If the Hamiltonian vector fields of f,g are complete and $\{f,g\}=0$, then $\mathcal{Q}(f)$ and $\mathcal{Q}(g)$ *strongly* commute.[18] This assumption is used to derive relations (2) and (3) in Refs. 4 and 9. It is also possible to enforce the integrability of $\mathcal{Q}(P^2)$ in a more direct manner.[3]

We now show that Van Hove's assumption is unnecessary; in fact, we may establish the integrability of $\mathcal{Q}(P^2)$ directly, via the following generalization of Proposition 1.

Proposition 3: Let $\mathcal{Q}$ be a quantization of P^2 on a dense invariant domain D in a Hilbert space $\mathcal{H}$. Then there is a unitary transformation U: $L^2(\mathbf{R}^n)\to\mathcal{H}$ such that $\mathcal{Q}(f)=U d\varpi(f)U^{-1}\restriction D$ for all $f\in P^2$.

Thus, up to unitary equivalence, $\mathcal{Q}$ must be either $d\varpi$ or a restriction thereof. As such, $\mathcal{Q}(P^2)$ must be integrable and, consequently, Van Hove's strong commutativity assumption holds for elements of P^2.

Before giving the proof, we establish two technical lemmas.

Lemma 2: Let A be an e.s.a. operator on a Hilbert space, and B a closable operator, both of which have a common dense invariant domain D. Suppose that D consists of analytic vectors for A, and that A (weakly) commutes with B. Then $\exp(i\overline{A})$ (weakly) commutes with $\overline{B}$ on D.

Proof: Recall that as $\psi\in D$ is analytic for A,

$$e^{i\overline{A}}\psi=\sum_{k=0}^{\infty}\frac{1}{k!}(iA)^k\psi=:\phi.$$

Define the partial sums

$$\phi_K=\sum_{k=0}^{K}\frac{1}{k!}(iA)^k\psi\in D;$$

then using the (weak) commutativity of A and B,

$$B\phi_K=\sum_{k=0}^{K}\frac{1}{k!}(iA)^k B\psi\in D.$$

Since $B\psi\in D$ is analytic for A, the sequence $B\phi_K$ converges:

$$\chi:=\lim_{K\to\infty}B\phi_K=e^{i\overline{A}}B\psi=e^{i\overline{A}}\overline{B}\psi.$$

But $\overline{B}$ is closed, hence $\phi=\lim_{K\to\infty}\phi_K$ is in the domain of $\overline{B}$ and $\chi=\overline{B}\phi$, i.e.,

$$e^{i\overline{A}}\overline{B}\psi=\overline{B}e^{i\overline{A}}\psi$$

for all $\psi\in D$. ▽

Lemma 3: Let B be a closable operator. If a bounded operator T (weakly) commutes with B on $D(B)$, then they also commute on $D(\bar{B})$.

Proof: If $\psi\in D(\bar{B})$, then from the definition of closure there exists a sequence $\{\psi_k\}$ in $D(B)$ with $\psi_k\to\psi$ such that $B\psi_k\to\bar{B}\psi$. Because the operator T is continuous,

$$T\bar{B}\psi = T\lim_{k\to\infty} B\psi_k = \lim_{k\to\infty} TB\psi_k = \lim_{k\to\infty} \bar{B}T\psi_k$$

as T commutes with $\bar{B}$ on $D(B)$. Again applying the definition of closure to the sequence $\{T\psi_k\}$ in $D(\bar{B})$, we get that $\lim_{k\to\infty}T\psi_k = T\psi\in D(\bar{B})$ and

$$\bar{B}T\psi = \lim_{k\to\infty}\bar{B}T\psi_k = T\bar{B}\psi$$

for every $\psi\in D(\bar{B})$. ▽

Proof of Proposition 3: Let $\mathcal{Q}$ be a quantization of P^2. As discussed earlier, we may assume that $\mathcal{Q}(P^1)$ is the Schrödinger representation (1) on $L^2(\mathbf{R}^n)$, and that the domain $D\subseteq\mathcal{S}(\mathbf{R}^n)$. Again taking $n=1$, we will prove by brute force that the Von Neumann rules (4) and (5) hold.

We begin by determining $\mathcal{Q}(q^2)$. Set $\Delta=\mathcal{Q}(q^2)-\mathcal{Q}(q)^2$. We readily verify that $[\Delta,\mathcal{Q}(q)]=0$ and $[\Delta,\mathcal{Q}(p)]=0$ on D. Now let $D_\omega\subseteq D$ be the space of separately analytic vectors for $\mathcal{Q}(q)$ and $\mathcal{Q}(p)$; by (Q5) we have that D_ω is dense. According to Proposition 1 of Ref. 6, Δ leaves D_ω invariant. By Corollary 2 in Sec. X.6 of Ref. 7, $\mathcal{Q}(q)\restriction D_\omega$ is e.s.a.; moreover, $\Delta_\omega := \Delta\restriction D_\omega$ is symmetric and hence closable. Upon taking $A=\mathcal{Q}(q)\restriction D_\omega$ and $B=\Delta_\omega$ in Lemma 2, it follows that $\exp(i\overline{\mathcal{Q}(q)\restriction D_\omega})=\exp(i\overline{\mathcal{Q}(q)})$ and Δ_ω commute on D_ω. Lemma 3 than shows that $\exp(i\overline{\mathcal{Q}(q)})$ and $\overline{\Delta_\omega}$ commute on $D(\overline{\Delta_\omega})$. Likewise $\exp(i\overline{\mathcal{Q}(p)})$ and $\overline{\Delta_\omega}$ commute on $D(\overline{\Delta_\omega})$. But now the unbounded version of Schur's lemma[19] implies that $\overline{\Delta_\omega}=EI$ for some real constant E on $D(\overline{\Delta_\omega})=L^2(\mathbf{R})$. Since $\overline{\Delta_\omega}$ is the smallest closed extension of Δ_ω and $\Delta_\omega\subset\Delta\subset\bar{\Delta}$, it follows that $\bar{\Delta}=EI$, whence Δ itself is a constant multiple of the identity on D. Thus $\mathcal{Q}(q^2)=\mathcal{Q}(q)^2+EI$ on D.

An identical argument yields $\mathcal{Q}(p^2)=\mathcal{Q}(p)^2+FI$ on D. Quantizing the relation $4pq=\{p^2,q^2\}$ and using these formulas then gives

$$\mathcal{Q}(pq)=\tfrac{1}{2}(\mathcal{Q}(p)\mathcal{Q}(q)+\mathcal{Q}(q)\mathcal{Q}(p))$$

on D. But upon quantizing $2q^2=\{pq,q^2\}$ we find that $E=0$. Similarly $F=0$. It follows from (1)–(3) that $\mathcal{Q}=d\varpi\restriction D$. □

Thus, up to unitary equivalence and restriction of representations, we may as well suppose that $D=\mathcal{S}(\mathbf{R}^n)$. If we were to take this as our starting point, then we could reverse our constructions and derive (4) and (5) in a simpler fashion, cf. Sec. 5.1 of Ref. 5.

If $\mathcal{Q}$ were a quantization of P, $\mathcal{Q}(P^2)$ must therefore be unitarily equivalent to (a restriction of) the extended metaplectic quantization, and this contradicts Theorem 2. Thus we have proven our main result.

Theorem 4 (Strong No-Go Theorem): There is no quantization of P.

Van Hove[2] gave a slightly different analysis using only those observables $f\in C^\infty(\mathbf{R}^{2n})$ with complete Hamiltonian vector fields, and still obtained an obstruction [but now to quantizing all of $C^\infty(\mathbf{R}^{2n})$]. Yet other variants of Groenewold's theorem are presented in Refs. 12 and 20. Related results can be found in Refs. 21 and 22.

V. QUANTIZABLE LIE SUBALGEBRAS OF POLYNOMIALS

We hasten to add that there are Lie subalgebras of $P(2n)$ other than $P^2(2n)$ which can be quantized. For example, consider the coordinate subalgebra $C(2n)$. It is straightforward to verify that for each $\eta\in\mathbf{R}$, the map $\sigma_\eta : C(2n)\to\mathrm{Op}(\mathcal{S}(\mathbf{R}^n))$ given by

$$\sigma_\eta\left(\sum_{i=1}^n f^i(q)p_i+g(q)\right)=-i\hbar\sum_{i=1}^n\left(f^i(q)\frac{\partial}{\partial q^i}+\left[\frac{1}{2}+i\eta\right]\frac{\partial f^i(q)}{\partial q^i}\right)+g(q) \tag{6}$$

is a quantization of $C(2n)$. σ_0 is the familiar ''position'' or ''coordinate representation.'' The significance of the parameter η is explained in Ref. 23 and 24 (see also Ref. 25). There it is shown that while the quantizations σ_η and $\sigma_{\eta'}$ are unitarily inequivalent if $\eta\neq\eta'$, they are related by a *nonlinear* norm-preserving isomorphism.

Proposition 5: C is a maximal Lie subalgebra of *P*.

Proof: We take $n=1$. Suppose that *V* were a Lie subalgebra of *P* strictly containing *C*. *V* must contain a polynomial *h* of the form

$$h(q,p)=f(q)p^k+\text{terms of degree at most } k-1 \text{ in } p$$

for some $k>1$ and some polynomial $f\neq 0$ of degree *l*. Now both $q,p\in V$, and so by bracketing *h* with q $(k-2)$-times, we get

$$\frac{k!}{2}f(q)p^2+\text{terms of degree at most degree 1 in } p\in V.$$

Since $C\subset V$ this implies that $f(q)p^2\in V$. By bracketing this expression with *p* *l*-times, we conclude that $p^2\in V$. Now both $q^2,qp\in V$, so $P^2\subset V$. The maximality of P^2 implies that $V=P$, whence *C* is maximal. □

As a consequence, any quantization which extends σ_η must be defined on all of *P*. Thus Theorem 4 yields

Corollary 6: The quantizations σ_η of *C* cannot be extended beyond *C* in *P*.

Furthermore a variant of Proposition 3 (see also Theorem 8 in Ref. 25) yields ''uniqueness:''

Proposition 7: Let $\mathcal{Q}$ be a quantization of *C* on a dense invariant domain *D* in a Hilbert space $\mathcal{H}$. Then there is an $\eta\in\mathbf{R}$ and a unitary transformation $U:L^2(\mathbf{R}^n)\to\mathcal{H}$ such that $\mathcal{Q}(f)=U\sigma_\eta(f)U^{-1}\restriction D$ for all $f\in C$.

Proof: Again set $n=1$. As in the proof of Proposition 3, we may assume that $\mathcal{Q}(P^1)$ is given by (1) on $L^2(\mathbf{R})$ and that $D\subseteq\mathcal{S}(\mathbf{R})$.

Just as before, we first compute that $\mathcal{Q}(q^2)=\mathcal{Q}(q)^2+EI$ on *D* for some real constant *E*.

Now consider $\mathcal{Q}(q,p)$. Set

$$\Delta=\mathcal{Q}(qp)-\tfrac{1}{2}(\mathcal{Q}(q)\mathcal{Q}(p)+\mathcal{Q}(p)\mathcal{Q}(q)).$$

It is straightforward to verify that Δ commutes with both $\mathcal{Q}(q)$ and $\mathcal{Q}(p)$. The same argument based on Lemmas 2 and 3 and the unbounded Schur's lemma that was used in the proof of Propositin 3 can be applied *mutatis mutandis* to give $\Delta=GI$ on *D* for some real constant *G*. Thus

$$\mathcal{Q}(qp)=\tfrac{1}{2}(\mathcal{Q}(q)\mathcal{Q}(p)+\mathcal{Q}(p)\mathcal{Q}(q))+GI \tag{7}$$

on *D*. By quantizing the Poisson bracket relation $\{qp,q^2\}=2q^2$ we find that $E=0$. Arguing as in the proof of Lemma 1, we then find that on *D*

$$\mathcal{Q}(q^k)=\mathcal{Q}(q)^k. \tag{8}$$

Next, quantizing the Poisson bracket relations $\{q^kp,q\}=q^k$ and $\{q^kp,p\}=-kq^{k-1}p$ yields

$$[\mathcal{Q}(q^kp),\mathcal{Q}(q)]=-i\hbar\mathcal{Q}(q^k)\quad\text{and}\quad[\mathcal{Q}(q^kp),\mathcal{Q}(p)]=i\hbar k\mathcal{Q}(q^{k-1}p), \tag{9}$$

respectively. Now consider the classical relation $(1-k)q^kp=\{q^kp,qp\}$. Quantizing this and simplifying by means of (7), (9), and (8) produces the recursion relation

$$\mathcal{Q}(q^k p) = \frac{1}{1-k}(\mathcal{Q}(q^k)\mathcal{Q}(p) - k\mathcal{Q}(q)\mathcal{Q}(q^{k-1}p)).$$

Iterating this computation $(k-1)$-times gives

$$\mathcal{Q}(q^k p) = (1-k)\mathcal{Q}(q^k)\mathcal{Q}(p) + k\mathcal{Q}(q)^{k-1}\mathcal{Q}(qp).$$

Again using (7) and simplifying, we finally get

$$\mathcal{Q}(q^k p) = \mathcal{Q}(q^k)\mathcal{Q}(p) + k\left(G - \frac{i\hbar}{2}\right)\mathcal{Q}(q)^{k-1}.$$

Recalling (1) and (8), this can be rewritten

$$\mathcal{Q}(q^k p) = -i\hbar\left[q^k\frac{d}{dq} + \left(\frac{1}{2} + \frac{iG}{\hbar}\right)\frac{dq^k}{dq}\right]$$

on D. Consolidating this with (8), we obtain (6) where $\eta = G/\hbar$. We thus have $\mathcal{Q} = \sigma_\eta \restriction D$, as claimed. □

Notice that unlike in the proof of Proposition 3, we cannot quantize the Poisson bracket relation $\{q^2, p^2\} = -4qp$ to obtain $G=0$ since $p^2 \notin C$. The fact that G remains arbitrary is mirrored by the presence of the parameter η in (6).

Thus far we have encountered two maximal Lie subalgebras of P containing P^1: P^2 and C. When $n=1$, it turns out that these are essentially the *only* such subalgebras.

Theorem 8: $(n=1)$ Up to isomorphism, P^2 and C are the only maximal Lie subalgebras of P which contain P^1.

Proof: Suppose that W were a maximal Lie subalgebra of P containing P^1, distinct from P^2. We will show that W must be isomorphic to C. Denote $W^k = W \cap P^k$, etc.

Since $W \neq P^2$ there must exist a polynomial of degree $k, k>2$, in W. By bracketing this polynomial $(k-2)$ times with an appropriate number of $p, q \in W$, we obtain a nonzero polynomial $h \in W^2$. Since $P^1 \subset W$, we may subtract off terms of degree one or less, so we may assume that h is homogeneous quadratic. By means of a rotation we may diagonalize h; thus we may further suppose that canonical coordinates have been chosen so that $h(q,p) = ap^2 + cq^2$. Now $\dim W_2 \neq 3$, for otherwise $P^2 \subset W$, and then the maximality of P^2 implies that $W = P$. We break the argument into parts, depending on whether $\dim W_2 = 1$ or 2.

(i) $\dim W_2 = 1$: Then W_2 is spanned by h. We first claim that either $W^3 = W^2$ or $W^3 \subset C^3$. Indeed, if $f \in W^3$, then the quadratic terms of both $\{p,f\},\{f,q\} \in W^2$ must be proportional to h: $\{p,f\} = rh + \text{l.d.t.}$ and $\{f,q\} = sh + \text{l.d.t.}$, where "l.d.t." means lower degree terms. The particular form of h then implies that

$$f(q,p) = \tfrac{1}{3}(sap^3 + rcq^3) + \text{l.d.t.},$$

along with $sc=0$ and $ra=0$. Since $h \neq 0$, both a, c cannot vanish. If both $r, s = 0$, then $f \in W^2$ and so $W^3 = W^2$. If both $s, a = 0$, then h is proportional to q^2 and f must be of the form

$$\tfrac{1}{3}rcq^3 + \alpha q^2 + \beta qp + \gamma p^2 + \text{l.d.t.}$$

But then $\{f,h\} \propto 2\beta q^2 + 4\gamma qp \in W^2$, which forces $\gamma = 0$. Thus $W^3 \subset C^3$. The canonical transformation $q \mapsto p$, $p \mapsto -q$ reduces the subcase with $r, c = 0$ to the previous one.

If $W^3 = W^2$ then $W = W^2 \subset P^2$, which contradicts the assumed maximality of W.

If $W^3 \subset C^3$, then a similar argument shows that $W^4 \subset C^4$, and so on. Thus $W \subset C$, which again contradicts the maximality of W.

(ii) $\dim W_2 = 2$: Now we may suppose that h, g form a basis for W_2, where h is as above and

$$g(q,p) = rp^2 + spq + tq^2.$$

If $s=0$ then, as h,g are linearly independent, both $p^2,q^2\in W_2$. But then $\{p^2,q^2\}=4pq\in W$, so that $\dim W_2=3$. Without loss of generality, we may thus assume that $s=1$.

Now $\{h,g\}\in W_2$, and a computation shows that $h,g,\{h,g\}$ are linearly dependent iff

$$ac+(at-cr)^2=0. \tag{10}$$

Again we consider various subcases. If $a=0$ then (10) gives $r=0$, and it follows from the above expressions for h,g that $W_2=C_2$. As in case (i), the subcase $c=0$ can be reduced to that of $a=0$ by means of a linear canonical transformation. It remains to consider the subcase $ac \neq 0$. We may suppose that $a=1$; (10) then implies that $c<0$. Setting $\beta=t-rc$, we may thus take

$$h=p^2-\beta^2q^2 \quad \text{and} \quad \{g,h\}=2(p^2+2\beta pq+\beta^2q^2).$$

as a basis for W_2. But now the canonical transformation

$$p\mapsto\frac{1}{\sqrt{2}\beta}(p-\beta q), \quad q\mapsto\frac{1}{\sqrt{2}}(p+\beta q)$$

reduces this subcase to that of $a=0$. Thus up to isomorphism we have $W_2=C_2$.

Similarly we have $W_k\subseteq C_k$, and so $W\subseteq C$. The maximality of W now implies that $W=C$.□

In particular, the subalgebras $\{f(\mu p+q)(p-\mu q)+g(\mu p+q)\}$, where f,g are polynomials and $\mu\in\mathbf{R}$, are all maximal Lie subalgebras of $P(2)$ containing $P^1(2)$ isomorphic to $C(2)$. [These are the normalizers in $P(2)$ of the polarizations $\{g(\mu p+q)\}$.] So is the ''momentum algebra'' consisting of polynomials which are at most affine in the position.

As both $P^2(2)$ and $C(2)$ are quantizable, it follows from Theorem 2 and Corollary 6 that the following corollary is true.

Corollary 9: Up to isomorphism, $P^2(2)$ and $C(2)$ are the largest quantizable subalgebras of $P(2)$ containing $P^1(2)$.

Unfortunately, neither Theorem 8 nor Corollary 9 hold in higher dimensions. To see this, take $n=2$ and consider the Lie algebra

$$\{f(q^1)p_1+g(q^1,q^2,p_2)\},$$

where f,g are polynomials. This subalgebra is maximal, but not isomorphic to either $C(4)$ or $P^2(4)$. It is also not quantizable—if it were, we would obtain a quantization of the polynomial algebra in q^2,p_2, contrary to the Strong No-Go Theorem. Furthermore, the subalgebra thereof for which g is at most quadratic in q^2,p_2 is maximal quantizable, but also not isomorphic to either $C(4)$ or $P^2(4)$.

VI. DISCUSSION

We have thus completely solved the Groenewold–Van Hove problem for $\mathbf{R}^2$ in that we have identified (the isomorphism classes of) the largest quantizable Lie subalgebras of $P(2)$ [viz. $P^2(2)$ and $C(2)$] and explicitly constructed all their possible quantizations [given by (1)–(3) and (6), respectively]. It remains to carry out this program in higher dimensions; the key missing ingredient is a classification of the maximal Lie subalgebras of $P(2n)$ containing $P^1(2n)$. Unfortunately, this appears to be a difficult problem. We emphasize, however, that all the results of this paper other than Theorem 8 and Corollary 9 hold for arbitrary n.

Of course, Groenewold's classical result is valid only for $\mathbf{R}^{2n}$. Similar obstructions appear when trying to quantize certain other phase spaces, e.g., S^2 and T^*S^1. Complete solutions of the corresponding Groenewold–Van Hove problems in these two examples are given in Refs. 26 and 25, respectively. On the other hand, in some instances there are *no* obstructions to quantization, such as T^2 and $T^*\mathbf{R}_+$, cf. Refs. 27 and 12, respectively. (Although probably not of physical interest, it is amusing to wonder what happens for $\mathbf{R}^{2n}$, $n>1$, with an exotic symplectic structure.)

It is important, therefore, to understand the mechanisms which are responsible for these divergent outcomes. Already some results have been established along these lines, to the effect that under certain circumstances there are obstructions to quantizing both compact and noncompact phase spaces.[12,28–30] We refer the reader to Ref. 5 for an up-to-date summary.

ACKNOWLEDGMENTS

We wish to thank H. Grundling for valuable discussions, including his help in pointing out and correcting an error in a previous version of this manuscript. This research was supported in part by NSF Grant No. DMS 96-23083.

[1] H. J. Groenewold, ‘‘On the principles of elementary quantum mechanics,’’ Physica (Amsterdam) **12**, 405–460 (1946).
[2] L. van Hove, ‘‘Sur certaines représentations unitaires d'un groupe infini de transformations,’’ Proc. R. Acad. Sci. Belgium **26**, 1–102 (1951).
[3] M. J. Gotay, H. Grundling, and G. M. Tuynman, ‘‘Obstruction results in quantization theory,’’ J. Nonlinear Sci. **6**, 469–498 (1996).
[4] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin–Cummings, Reading, MA, 1978), 2nd ed.
[5] M. J. Gotay, ‘‘Obstructions to quantization,’’ in *The Juan Simo Memorial Volume*, edited by J. E. Marsden and S. Wiggins (Springer, New York, 1999) (in press).
[6] M. Flato and J. Simon, ‘‘Separate and joint analyticity in Lie groups representations,’’ J. Funct. Anal. **13**, 268–276 (1973).
[7] M. Reed and B. Simon, *Functional Analysis I, II* (Academic, New York, 1972, 1975), Sec. VIII.5.
[8] See Ref. 6, Cor. 1.
[9] P. R. Chernoff, ‘‘Mathematical obstructions to quantization,’’ Hadronic J. **4**, 879–898 (1981).
[10] G. B. Folland, *Harmonic Analysis in Phase Space*, Ann. Math. Ser. Vol. 122 (Princeton University Press, Princeton, NJ, 1989).
[11] M. J. Gotay, ‘‘Functorial geometric quantization and Van Hove's theorem,’’ Int. J. Theor. Phys. **19**, 139–161 (1980).
[12] M. J. Gotay and J. Grabowski, ‘‘On quantizing nilpotent and solvable basic algebras,’’ preprint math-ph/9902012 (1999).
[13] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics* (Cambridge University Press, Cambridge, 1984).
[14] See Ref. 10, Prop. 4.49.
[15] This representation actually drops to the double cover of $\mathrm{HSp}(2n,\mathbf{R})$, but we do not need this fact here.
[16] See Ref. 10, Prop. 4.58.
[17] See Ref. 13, Sec. 16.
[18] Recall that two e.s.a. operators *strongly commute* iff their spectral resolutions commute, cf. Sec. VIII.5 of Ref. 7. Two operators *A*, *B* *weakly commute* on a domain *D* if they commute in the ordinary sense, i.e., $[A,B]$ is defined on *D* and vanishes.
[19] A. Robert, *Introduction to the Representation Theory of Compact and Locally Compact Groups*, London Math. Soc. Lect. Note Ser. Vol. 80 (Cambridge University Press, Cambridge, 1983).
[20] M. Dubois-Violette (private communication, 1998).
[21] R. Arens and D. Babbit, ‘‘Algebraic difficulties of preserving dynamical relations when forming quantum-mechanical operators,’’ J. Math. Phys. **6**, 1071–1075 (1965).
[22] A. Joseph, ‘‘Derivations of Lie brackets and canonical quantization,’’ Commun. Math. Phys. **17**, 210–232 (1970).
[23] B. Angermann, H.-D. Doebner, and J. Tolar, ‘‘Quantum kinematics on smooth manifolds,’’ in *Nonlinear Partial Differential Operators and Quantization Procedures*, edited by S. I. Andersson and H.-D. Doebner, Lect. Notes Math. **1087**, 171–208 (1983).
[24] B. Hennig, ‘‘Nonlinear Schrödinger equations and Hamiltonian dynamics,’’ in *Nonlinear, Deformed, and Irreversible Quantum Systems*, edited by H.-D. Doebner, V. K. Dobrev, and P. Nattermann (World Scientific, Singapore, 1995), pp. 155–165.
[25] M. J. Gotay and H. B. Grundling, ‘‘On quantizing T^*S^1,’’ Rep. Math. Phys. **40**, 107–123 (1997).
[26] M. J. Gotay, H. Grundling, and C. A. Hurst, ‘‘A Groenewold–Van Hove theorem for S^2,’’ Trans. Am. Math. Soc. **348**, 1579–1597 (1996).
[27] M. J. Gotay, ‘‘On a full quantization of the torus,’’ in *Quantization, Coherent States and Complex Structures*, edited by J.-P. Antoine *et al.* (Plenum, New York, 1995), pp. 55–62.
[28] V. L. Ginzburg and R. Montgomery, ‘‘Geometric quantization and no-go theorems, ’’ preprint dg-ga/9703010 (1997).
[29] M. J. Gotay, J. Grabowski, and H. B. Grundling, ‘‘An obstruction to quantizing compact symplectic manifolds,’’ Proc. Am. Math. Soc. (in press).
[30] M. J. Gotay and H. Grundling, ‘‘Nonexistence of finite-dimensional quantizations of a noncompact symplectic manifold,’’ in *Differential Geometry and Its Applications 1998*, edited by J. Slovák and O. Kowalski (Masaryk University, Brno, Czech Republic, 1999) (in press).

The Jacobi polynomial ensemble and the Painlevé VI equation

Luc Haine[a)] and Jean-Pierre Semengue
Department of Mathematics, Université Catholique de Louvain, Chemin du Cyclotron 2, 1348 Louvain-la-Neuve, Belgium

(Received 11 September 1998; accepted for publication 9 December 1998)

We consider the Jacobi polynomial ensemble of $n\times n$ random matrices. We show that the probability of finding no eigenvalues in the interval $[-1,z]$ for a random matrix chosen from the ensemble, viewed as a function of z, satisfies a second-order differential equation. After a simple change of variable, this equation can be reduced to the Okamoto–Jimbo–Miwa form of the Painlevé VI equation. The result is achieved by a comparison of the Tracy–Widom and the Virasoro approaches to the problem, which both lead to different third-order differential equations. The Virasoro constraints satisfied by the tau functions are obtained by a systematic use of the moments, which drastically simplifies the computations. © *1999 American Institute of Physics.* [S0022-2488(99)02203-3]

I. INTRODUCTION

Consider a family of orthonormal polynomials $P_n(x)$, $n=0,1,2,...$, on an interval $[a,b]$, for a weight function $w(x)=\exp(-V(x))$. The associated orthogonal polynomial ensemble assigns a probability measure on the space of $n\times n$ Hermitian matrices proportional to

$$\exp(-\mathrm{Tr}\,V(M))dM, \tag{1.1}$$

with dM the product of Lebesgue measures over the independent elements of the Hermitian matrix M. As in the case of the unitary group, where the probability measure is the Haar measure, there is an exact analog of the Weyl integral formula, giving the induced density distribution on the eigenvalues of the matrices:

$$\prod_{1\leq k<l\leq n}(x_k-x_l)^2\prod_{j=1}^{n}w(x_j)dx_j\,. \tag{1.2}$$

We refer the reader to the book of Mehta,[1] as well as to the introductory lectures by Tracy and Widom,[2] for a discussion of this material.

In this paper, we shall be concerned with the (finite Fredholm) determinant,

$$\det(\mathbf{I}-K(z)), \tag{1.3}$$

viewed as a function of z, where $K(z)=(K_{ij}(z))_{0\leq i,j\leq n-1}$, denotes the $n\times n$ matrix with entries

$$K_{ij}(z)=\int_a^z P_i(x)P_j(x)w(x)dx. \tag{1.4}$$

The determinant (1.3) gives the probability of finding no eigenvalues in the interval $[a,z]$ for a random matrix chosen from the ensemble.[1,2] Tracy and Widom[3] have shown that, in the case of the Hermite and the Laguerre polynomials, as a function of z, the logarithmic derivatives of the

[a)]Electronic mail: haine@agel.ucl.ac.be

determinants (1.3) (multiplied by an appropriate factor related to the interval of orthogonality) solve, respectively, the Painlevé IV and the Painlevé V equations. In the case of the Jacobi polynomials, they have obtained[3] a third-order differential equation that they could not recognize as being connected with some Painlevé transcendents.

The methodology of Tracy and Widom goes well beyond the special examples mentioned above. It extends very successfully the pioneering work of Jimbo, Miwa, Môri, and Sato,[4] who showed that the Fredholm determinant constructed from the sine kernel solves the Painlevé V equation. Adler, Shiota, and van Moerbeke[5] have proposed another approach to the problem, based on the use of ''Virasoro constraints decoupling into a time-part and a boundary-part.'' As examples of orthogonal polynomial ensembles, the authors discuss the cases of the Hermite and the Laguerre polynomials, which lead to the same equations as those obtained by Tracy and Widom.[3]

We note that, in the case of the Hermite and the Laguerre polynomials, the ''Virasoro constraints without boundary-part'' were first obtained by Haine and Horozov.[6] The latter paper was motivated by the explicit construction of all the highest weight representations of the Virasoro algebra (with central charge $c=1$) in terms of tau functions. The Jacobi polynomials did not play any role in that construction. Grünbaum and Haine[7] came back to the Jacobi case, in relation with a discrete-continuous version of the bispectral problem, but more about this below.

We are now ready to state the results of our paper. We show that, in the case of the Jacobi polynomials, the Tracy–Widom and the Virasoro approaches lead to *different third-order* differential equations for the determinants (1.3). Comparing the results gives then a *second-order* differential equation, which, after some manipulations, we recognize as the Okamoto–Jimbo–Miwa ''σ representation'' of the Painlevé VI equation (Theorem IV.3). The derivation of our result depends on a new construction of the so-called master symmetries of the (semi-infinite) Toda lattice hierarchy. It is based on the classical correspondence (valid under appropriate hypothesis),

$$\{\mu_0,\mu_1,\mu_2,\ldots\}\to L, \tag{1.5}$$

between sequences of moments and semi-infinite tridiagonal matrices, as explained, for instance, in the treatises of Akhiezer[8] or Chihara.[9] Thus, we do not assume any knowledge of Sato's Grassmannian from the reader. Our main tool, Theorem III.1, allows us to establish the Virasoro constraints by checking some ''trivial'' relations satisfied by the moments [see Lemma IV.1 and formulas (A2) and (A8) in the Appendix, for illustrations of this principle].

There is, in fact, an intimate relation between some subalgebras of the algebra of master symmetries of the Toda lattice hierarchy, which appear naturally in the context of the *bispectral problem*,[7,10] and the *special* solutions of the Painlevé equations that we consider here. To explain this, let us introduce the moments

$$\mu_k(z)=\int_z^b x^k w(x)\,dx, \quad z\in[a,b]. \tag{1.6}$$

Elementary row and column manipulations show that

$$\det(I-K(z))=\frac{\tau_n(z)}{\tau_n(a)}, \tag{1.7}$$

with

$$\tau_n(z)=\det(\mu_{i+j}(z))_{0\leq i,j\leq n-1}. \tag{1.8}$$

The bispectral problem, in the form that is relevant to the considerations of this paper, asks for all families of orthogonal polynomials that are eigenfunctions of a second-order differential equation. In this (special) form, the problem was already solved by Bochner:[11] all the solutions are given by the classical orthogonal polynomials. Although the above mentioned work[7] was concerned with a bi-infinite version of the problem, it contains, in particular, the result that all the (semi-infinite) tridiagonal matrices L, which solve the problem, are *fixed points of some subalgebras of the algebra of master symmetries* of the Toda lattice hierarchy. Under the identification (1.5), the corresponding sequences of moments are obtained by putting $z=a$ and $w(x)$ to be the weight function of any family of classical orthogonal polynomials in (1.6). In a nutshell, our proof that (1.8) is a tau function for the appropriate Painlevé equations, amounts to checking that precisely the *same subalgebras of symmetries* become *tangent* along the curve (1.6), obtained by letting z move away from a.

II. ORTHOGONAL POLYNOMIALS AND SATO THEORY

In this section we summarize some of the results of Haine and Horozov,[6] where a precise connection was established between Sato's theory[12] and the theory of orthogonal polynomials. We have, however, removed all ''Grassmannian considerations,'' hoping to make the material accessible to a larger audience. Our starting point is the classical correspondence,[8,9]

$$\{\mu_0,\mu_1,\mu_2,\ldots\}\to L, \tag{2.1}$$

between sequences of moments and (semi-infinite) Jacobi matrices. We shall deal with monic orthogonal polynomials and consequently write L in the form

$$L=\begin{pmatrix} b_1 & 1 & & & & \\ a_1 & b_2 & 1 & & & \\ & a_2 & b_3 & 1 & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot \end{pmatrix}. \tag{2.2}$$

The correspondence is valid with the assumption that all entries a_n are nonzero and that all the $n\times n(n=1,2,3,...)$ determinants,

$$\tau_n=\det(\mu_{i+j})_{0\leqslant i,j\leqslant n-1}, \tag{2.3}$$

do not vanish. In one direction, it is given by the formulas

$$a_n=\frac{\tau_{n-1}\tau_{n+1}}{\tau_n^2},\quad b_n=\frac{\dot{\tau}_n}{\tau_n}-\frac{\dot{\tau}_{n-1}}{\tau_{n-1}}, \tag{2.4}$$

where, by convention, $\tau_0=1$, $\dot{\tau}_0=0$, and $\dot{\tau}_n$ denotes the determinant

$$\dot{\tau}_n=\det(\mu_{ij}^{(n)})_{0\leqslant i,j\leqslant n-1}, \tag{2.5}$$

with

$$\mu_{ij}^{(n)}=\mu_{i+j}\quad\text{for}\quad j\leqslant n-2\quad\text{and}\quad\mu_{in-1}^{(n)}=\mu_{i+n}.$$

Conversely, given the Jacobi matrix L, one constructs inductively the sequence of polynomials,

$$p_n(x)=(x-b_n)p_{n-1}(x)-a_{n-1}p_{n-2}(x),\quad n\geqslant 1, \tag{2.6}$$

with

$$p_{-1}(x)=0 \quad \text{and} \quad p_0(x)=1. \tag{2.7}$$

By Favard's theorem,[9] as long as the entries a_n are nonzero, there exists a unique (up to a multiplicative constant) sequence $\{\mu_n\}_{n\geq 0}$, such that the sequence $\{p_n(x)\}_{n\geq 0}$ is an orthogonal polynomial sequence for the associated moment functional M, that is

$$M(p_n p_m)=0, \quad \text{for} \quad m\neq n \quad \text{and} \quad M(p_n^2)\neq 0.$$

We remind the reader that the moment functional M corresponding to the sequence $\{\mu_n\}_{n\geq 0}$, is defined by $M(x^n)=\mu_n$ and then extended to the vector space of all polynomials by linearity.

It is a beautiful and elementary fact that, under the correspondence (2.1), the vector field,

$$\dot{\mu}_k=\mu_{k+1}, \tag{2.8}$$

translates on the Jacobi matrix L into the celebrated Toda lattice equation. More generally, the sequence of vector fields,

$$T_i: \frac{\partial \mu_k}{\partial t_{i+1}}=\mu_{k+i+1}, \quad i\geq -1, \tag{2.9}$$

with T_0, giving (2.8), form a family of commuting vector fields,

$$[T_i,T_j]=0, \tag{2.10}$$

and we have the following.

Theorem II.1:[13,14] *Under the correspondence* (2.1), *the vector fields* (2.9) *translate on the Jacobi matrix L into the Toda lattice hierarchy,*

$$T_i: \frac{\partial L}{\partial t_{i+1}}=[L_+^{i+1},L], \tag{2.11}$$

where L_+^i *denotes the upper part (including the diagonal) of* L^i.

Remark: Although the proof of the above theorem is elementary, to the best of our knowledge, in the context of semi-infinite matrices, it was first formulated by Witten.[14] In that paper, Witten considers polynomials orthogonal with respect to some measure $d\sigma_0$ and shows that the time deformation of the measure

$$d\sigma_t=\exp\left(\sum_n t_n x^n\right) d\sigma_0, \tag{2.12}$$

defines orthogonal polynomials for which the corresponding Jacobi matrices satisfy the Toda lattice hierarchy. Starting with the Hermite polynomials leads then to the partition function of two-dimensional (2-D)-lattice quantum gravity. This corresponds to the situation when the moment functional M is defined by a measure (i.e., all $a_n>0$). It is straightforward to check that the deformation (2.12) induces the flows (2.9) on the moments. When the starting measure $d\sigma_0$ is purely discrete, one is led to the finite Toda lattice, and Eq. (2.12) can be traced back to the integration by Moser[13] of the finite Toda lattice, via the inverse Stieltjes transform.

The next theorem[6] establishes the precise link between Sato's theory[12] and the theory of orthogonal polynomials. For the convenience of the reader we present here an elementary proof, which avoids the use of Grassmannians. Let $t=(t_1,t_2,t_3,\ldots)$ and let us denote by

$$\tau_n(t)=\det(\mu_{i+j}(t))_{0\leq i,j\leq n-1}, \tag{2.13}$$

the determinants obtained by flowing from (2.3) under the flows of the commuting vector fields T_i defined in (2.9). With a similar definition for $L(t)$, we have the following.

Theorem II.2: *The functions* $\tau_n(t)$, $t=(t_1,t_2,t_3,\ldots)$, *are tau functions in the precise sense of Sato theory, that is,*

$$p_n(x,t)=x^n\frac{\tau_n(t-[x^{-1}])}{\tau_n(t)}, \tag{2.14}$$

where $p_n(x,t)$ *denote the (monic) orthogonal polynomials defined by the Jacobi matrix* $L(t)$ *and*

$$[x]=(x,\tfrac{1}{2}x^2,\tfrac{1}{3}x^3,\ldots). \tag{2.15}$$

Proof: Introduce the elementary Schur polynomials via the generating function

$$\exp\left(\sum_{k=1}^{\infty} t_k x^k\right)=\sum_{n\in\mathbf{Z}} S_n(t)x^n. \tag{2.16}$$

The (formal) solution of the family of commuting flows T_i in (2.9) is given by

$$\mu_k(t)=\sum_{n\in\mathbf{Z}} S_n(t)\mu_{n+k}(0). \tag{2.17}$$

From (2.16), we deduce easily that

$$S_n(t-[x^{-1}])=S_n(t)-x^{-1}S_{n-1}(t), \tag{2.18}$$

which, using (2.17), gives

$$\mu_k(t-[x^{-1}])=\mu_k(t)-x^{-1}\mu_{k+1}(t), \tag{2.19}$$

and therefore

$$x^n\tau_n(t-[x^{-1}])=\begin{vmatrix} \mu_0(t) & \mu_1(t) & \cdots & \mu_n(t)\\ \vdots & \vdots & & \vdots\\ \mu_{n-1}(t) & \mu_n(t) & \cdots & \mu_{2n-1}(t)\\ 1 & x & \cdots & x^n \end{vmatrix}, \tag{2.20}$$

which can be checked by expanding the determinant along the last line. The ratio of the right-hand side of (2.20) by $\tau_n(t)$ is the classical formula[9] that expresses the (monic) orthogonal polynomials $p_n(x,t)$ in terms of the moments $\mu_k(t)$. This establishes the theorem.

It follows from Theorem II.2 that each of the functions $\tau_n(t)$, $n=1,2,3,4,\ldots$, must be a solution of the system of Hirota's bilinear equations,[12,15] which characterizes tau functions. In what follows we shall just need to know that they satisfy the first equation of this hierarchy, which is the classical Kadomtzev–Petviashvili (KP in short) equation:

$$\left(\frac{\partial^4}{\partial t_1^4}+3\frac{\partial^2}{\partial t_2^2}-4\frac{\partial^2}{\partial t_1\partial t_3}\right)\log\tau_n(t)+6\left(\frac{\partial^2}{\partial t_1^2}\log\tau_n(t)\right)^2=0. \tag{2.21}$$

III. OSCILLATOR REPRESENTATIONS OF THE VIRASORO ALGEBRA AND MASTER SYMMETRIES OF THE TODA LATTICE HIERARCHY

We introduce the following vector fields on the space of moments:

$$V_i:\frac{\partial\mu_k}{\partial s_i}=(i+k+1)\mu_{i+k},\quad i\geqslant -1. \tag{3.1}$$

These vector fields satisfy the commutation relations,

$$[V_i,V_j]=(j-i)V_{i+j}. \tag{3.2}$$

They can be used to generate, via commutators, the vector fields T_i of the Toda lattice hierarchy (2.9),

$$[V_i,T_j]=(j+1)T_{i+j}, \tag{3.3}$$

and, for this reason, they are often called master symmetries.

In order to formulate the result of this section, we shall need the so-called oscillator representations of the Virasoro algebra:[15]

$$\begin{aligned} L_0^{(n)}&=n^2+\sum_{k=1}^{\infty} kt_k\frac{\partial}{\partial t_k},\\ L_j^{(n)}&=\sum_{k=1}^{j-1}\frac{\partial^2}{\partial t_k\partial t_{j-k}}+2n\frac{\partial}{\partial t_j}+\sum_{k=1}^{\infty} kt_k\frac{\partial}{\partial t_{k+j}},\quad j\geqslant 1,\\ L_{-j}^{(n)}&=\frac{1}{4}\sum_{k=1}^{j-1}k(j-k)t_kt_{j-k}+njt_j+\sum_{k=j+1}^{\infty}kt_k\frac{\partial}{\partial t_{k-j}},\quad j\geqslant 1. \end{aligned} \tag{3.4}$$

The operators $L_j^{(n)}$ satisfy the commutation relations of the Virasoro algebra, with central charge $c=1$:

$$[L_i^{(n)},L_j^{(n)}]=(i-j)L_{i+j}^{(n)}+\delta_{i,-j}\frac{i^3-i}{12}. \tag{3.5}$$

Starting with a sequence $\{\tau_n\}$ as in (2.3), let us denote by $\{\tau_n^{s_j}(t)\}$ the sequence that is obtained by first flowing along the vector field V_j during a time s_j, and then flowing along the (commuting) family of Toda flows. Notice that, since the master symmetries and the Toda flows *do not* commute between themselves, it is very important to specify *in which order* the flows are followed. With this convention, we can formulate our main result.

Theorem III.1: For $j\geqslant -1$,

$$\frac{\partial}{\partial s_j}\tau_n^{s_j}(t)|_{s_j=0}=L_j^{(n)}\tau_n(t). \tag{3.6}$$

This theorem is *the key* to the derivation of the various ''Virasoro-type constraints'' that are satisfied by *special* (Toda-type) tau functions. Indeed, it tells us, for example, that if the vector field V_j vanishes at $\{\tau_n\equiv\tau_n(0)\}$ (that is, at $\{\mu_n\equiv\mu_n(0)\}$), then, by definition, $\tau_n^{s_j}(t)$ is independent of s_j, and thus $L_j^{(n)}\tau_n(t)=0$, for all n. The proof of Theorem III.1 will depend on an expansion of the tau function in terms of Schur polynomials.

For any partition $j_1\geqslant j_2\geqslant\cdots\geqslant j_k>0$, with k parts, the Schur polynomials $S_{j_1,\ldots,j_k}(t)$ are defined in terms of the elementary Schur polynomials (2.16) by the determinants

$$S_{j_1,\ldots,j_k}(t)=\det(S_{j_r+s-r}(t))_{1\leqslant r,s\leqslant k}. \tag{3.7}$$

Substituting (2.17) into (2.13), one computes that the tau functions $\tau_n(t)$ admit the following expansion (which is a particular case of a formula of Sato[12] for general tau functions):

$$\tau_n(t)=\sum_{0\leqslant i_1<i_2<\cdots<i_n} p_{i_1,i_2,\ldots,i_n}S_{i_n-n+1,i_{n-1}-n+2,\ldots,i_1}(t), \tag{3.8}$$

where $p_{i_1,i_2,\ldots,i_n}$ denote the Plücker coordinates:

$$p_{i_1,i_2,\ldots,i_n}=\begin{vmatrix} \mu_{i_1} & \mu_{i_2} & \cdots & \mu_{i_n} \\ \mu_{i_1+1} & \mu_{i_2+1} & \cdots & \mu_{i_n+1} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{i_1+n-1} & \mu_{i_2+n-1} & \cdots & \mu_{i_n+n-1} \end{vmatrix}. \tag{3.9}$$

We shall need three lemmas. In order to formulate them, we introduce some notations. Given n vectors $x_1,\ldots,x_n\in\mathbf{R}^n$, we shall denote by $|x_1x_2\cdots x_n|$ the determinant of the $n\times n$ matrix formed with the columns x_i. Also, given two vectors x and y, $x\wedge y$ denotes the usual wedge product, with components $(x\wedge y)_{rs}=x_ry_s-x_sy_r$. Finally, for an $n\times n$ matrix A, A_r will denote the rth column of A and $\mathrm{tr}(A)$ will mean the trace of A. With these conventions, we have the following.

Lemma III.2: Let A and B be $n\times n$ matrices, with A invertible, and let D be an $n\times n$ diagonal matrix $D=\mathrm{diag}(d_1,d_2,\ldots,d_n)$; then

$$(i)\ \ \Sigma_{r=1}^n|A_1\cdots A_{r-1}(DB)_rA_{r+1}\cdots A_n|=(\det A)\mathrm{tr}(DBA^{-1}),$$

$$(ii)\ \ \Sigma_{1\leqslant r<s\leqslant n}|A_1\cdots A_{r-1}(DB)_rA_{r+1}\cdots A_{s-1}(DB)_sA_{s+1}\cdots A_n|$$
$$=(\det A)\sum_{1\leqslant r<s\leqslant n}((DBA^{-1})_r\wedge(DBA^{-1})_s)_{rs}.$$

Proof: Both assertions are easy consequences of Cramer's formula. For simplicity of notations, we just establish (i). One has

$$\text{the left-hand side of } (i)=(\det A)\sum_{r=1}^n(A^{-1}DB)_{rr}, \text{ by Cramer's formula,}$$
$$=(\det A)\mathrm{tr}(A^{-1}DB)=(\det A)\mathrm{tr}(DBA^{-1}). \tag{3.10}$$

This proves Lemma III.2.

The next lemma translates the master symmetries on Plücker coordinates.

Lemma III.3: Let $V_jp_{i_1,\ldots,i_n}$ denote the Lie derivative of the Plücker coordinates (3.9) *in the direction of the vector fields V_j introduced in* (3.1); *then*

$$(i)\ \ V_{-1}\ \ p_{i_1,\ldots,i_n}=\sum_{r=1}^n i_rp_{i_1,\ldots,i_r-1,\ldots,i_n},$$

$$(ii)\ \ V_0\ \ p_{i_1,\ldots,i_n}=\left(\frac{n(n+1)}{2}+\sum_{r=1}^n i_r\right)p_{i_1,\ldots,i_n},$$

$$(iii)\ \ V_1\ \ p_{i_1,\ldots,i_n}=\sum_{r=1}^n(n+1+i_r)p_{i_1,\ldots,i_r+1,\ldots,i_n},$$

$$(iv)\ \ V_2\ \ p_{i_1,\ldots,i_n}=\sum_{r=1}^n(n+2+i_r)p_{i_1,\ldots,i_r+2,\ldots,i_n}+\sum_{1\leqslant r<s\leqslant n}p_{i_1,\ldots,i_r+1,\ldots,i_s+1,\ldots,i_n}.$$

Proof: Let us fix $0\leqslant i_1<i_2<\cdots<i_n$ and introduce the $n\times n$ matrices,

$$A=(\mu_{i_{s+1}+r})_{0\leqslant r,s\leqslant n-1},\quad B_j=(\mu_{i_{s+1}+r+j})_{0\leqslant r,s\leqslant n-1}, \tag{3.11}$$

as well as the $n\times n$ diagonal matrix,

$$D=\mathrm{diag}(0,1,2,...,n-1). \tag{3.12}$$

Notice that, by the definition (3.9) of the Plücker coordinates, we have

$$\det A=p_{i_1,\dots,i_n}. \tag{3.13}$$

From the definition of V_j and of the Plücker coordinates, respectively, in (3.1) and (3.9), using first Leibniz rule and then Lemma III.2 (i), we find

$$\begin{aligned} V_j p_{i_1,\dots,i_n} &= \sum_{r=1}^{n} (j+i_r+1) p_{i_1,\dots,i_r+j,\dots,i_n} + \sum_{r=1}^{n} |A_1\cdots A_{r-1}(DB_j)_r A_{r+1}\cdots A_n|, \\ &= \sum_{r=1}^{n} (j+i_r+1) p_{i_1,\dots,i_r+j,\dots,i_n} + (\det A)\mathrm{tr}(DB_j A^{-1}). \end{aligned} \tag{3.14}$$

When $j=-1$, one checks easily that $\mathrm{tr}(DB_{-1}A^{-1})=0$, using (3.11) and (3.12), which gives (i). When $j=0$, $B_0=A$, and thus $\mathrm{tr}(DB_0A^{-1})=\mathrm{tr}(D)=n(n-1)/2$ which, using (3.13), gives (ii).

When $j=1$,

$$\begin{aligned} (\det A)\mathrm{tr}(DB_1A^{-1}) &= (n-1)(\det A)\mathrm{tr}(B_1A^{-1}) \\ &\qquad [\text{using the definition of } B_1 \text{ and } D \text{ in (3.11) and (3.12)}], \\ &= (n-1)\sum_{r=1}^{n} p_{i_1,\dots,i_r+1,\dots,i_n} \quad (\text{using Lemma III.2 (i) with } D=\text{identity}), \end{aligned} \tag{3.15}$$

and thus (3.14) reduces to (iii).

It remains to discuss the case $j=2$ of (3.14). One computes that

$$\begin{aligned} (\det A)\mathrm{tr}(DB_2A^{-1}) &= (n-1)(\det A)\mathrm{tr}(B_2A^{-1}) \\ &\quad + (\det A)\sum_{1\leqslant r<s\leqslant n} ((B_1A^{-1})_r \wedge (B_1A^{-1})_s)_{rs}, \end{aligned} \tag{3.16}$$

noticing that, in the last sum, only the term $r=n-1$, $s=n$ is nonzero. Using Lemma III.2 (i) and (ii), both with $D=$identity, it follows that

$$\begin{aligned} (\det A)\mathrm{tr}(DB_2A^{-1}) &= (n-1)\sum_{r=1}^{n} p_{i_1,\dots,i_r+2,\dots,i_n} \\ &\quad + \sum_{1\leqslant r<s\leqslant n} p_{i_1,\dots,i_r+1,\dots,i_s+1,\dots,i_n}, \end{aligned} \tag{3.17}$$

which shows that (3.14), with $j=2$, reduces to (iv). This concludes the proof of Lemma III.3.

The last lemma computes the action of the Virasoro operators, defined in (3.4), on the Schur polynomials.

Lemma III.4: Let $0\leqslant i_1<\cdots<i_k$; *then*

$$(i)\quad L_{-1}^{(n)}S_{i_k-k+1,\ldots,i_1}(t)=\sum_{r=1}^{k}(i_r+1)S_{i_k-k+1,\ldots,i_r-r+2,\ldots,i_1}(t)+(n-k)t_1S_{i_k-k+1,\ldots,i_1}(t),$$

$$(ii)\quad L_0^{(n)}S_{i_k-k+1,\ldots,i_1}(t)=\left(n^2-\frac{k(k-1)}{2}+\sum_{r=1}^{k}i_r\right)S_{i_k-k+1,\ldots,i_1}(t),$$

$$(iii)\quad L_1^{(n)}S_{i_k-k+1,\ldots,i_1}(t)=\sum_{r=1}^{k}(2n-k+i_r)S_{i_k-k+1,\ldots,i_r-r,\ldots,i_1}(t),$$

$$(iv)\quad L_2^{(n)}S_{i_k-k+1,\ldots,i_1}(t)=\sum_{r=1}^{k}(2n-k+i_r)S_{i_k-k+1,\ldots,i_r-r-1,\ldots,i_1}(t)+\sum_{1\leqslant r<s\leqslant k}S_{i_k-k+1,\ldots,i_r-r,\ldots,i_s-s,\ldots,i_1}(t).$$

Proof: The proof is very similar in spirit to the proof of Lemma III.3, so we just sketch the main steps. We introduce the operators

$$M_{-1}=\sum_{k=2}^{\infty}kt_k\frac{\partial}{\partial t_{k-1}}\quad\text{and}\quad M_j=\sum_{k=1}^{\infty}kt_k\frac{\partial}{\partial t_{k+j}},\quad j\geqslant 0. \tag{3.18}$$

From the definition of the elementary Schur polynomials in (2.16), it follows easily that

$$M_{-1}S_i(t)=(i+1)S_{i+1}(t)-t_1S_i(t)\quad\text{and}\quad M_jS_i(t)=(i-j)S_{i-j}(t),\quad j\geqslant 0. \tag{3.19}$$

We introduce the $n\times n$ matrices,

$$A=(S_{i_{s+1}-r}(t))_{0\leqslant r,s\leqslant k-1},\quad B_j=(S_{i_{s+1}-r-j}(t))_{0\leqslant r,s\leqslant k-1}, \tag{3.20}$$

as well as the diagonal matrix,

$$D=\mathrm{diag}(0,-1,-2,\ldots,-(k-1)). \tag{3.21}$$

Notice that

$$\det A=S_{i_k-k+1,\ldots,i_1}(t). \tag{3.22}$$

Since the operators M_j, $j=-1,0,1,2$, act as derivations on the space of polynomials in $t_1,t_2,t_3,\ldots,$ using (3.19), (3.20), (3.21), and (3.22), one derives explicit formulas for $M_jS_{i_k-k+1,\ldots,i_1}(t)$ in terms of Schur polynomials, using exactly the same arguments as in the proof of Lemma III.3. From these formulas, using that

$$\frac{\partial}{\partial t_j}S_i(t)=S_{i-j}(t), \tag{3.23}$$

one deduces easily formulas (i), (ii), (iii), and (iv), giving the action of the Virasoro operators on the Schur polynomials. This finishes the proof of Lemma III.4.

Given the previous lemmas, the proof of Theorem III.1 is now straightforward.

Proof of Theorem III.1: We shall use the shorthand notation

$$\tau_n(t)=\Sigma_I p_I S_I(t), \tag{3.24}$$

for the Plücker expansion (3.8). For $j=-1,0,1,2$, we have

$$\begin{aligned}\frac{\partial}{\partial s_j}\tau_n^{s_j}(t)|_{s_j=0}&=\Sigma_I(V_jp_I)S_I(t),\\&=\Sigma_I p_I(L_j^{(n)}S_I(t)),\quad \text{using Lemmas III.3 and III.4, with}\\&\quad k=n,\ \text{and some relabeling of the indices},\\&=L_j^{(n)}\tau_n(t).\end{aligned}\tag{3.25}$$

The rest of the argument follows from the commutation relations (3.2) and (3.5), that are satisfied by the master symmetries and the Virasoro operators. Notice that for $j\geqslant -1$, they differ by a sign! Assuming the theorem for some $j\geqslant 2$, we establish it for $j+1$. Indeed,

$$\begin{aligned}(j-1)\frac{\partial}{\partial s_{j+1}}\tau_n^{s_{j+1}}(t)|_{s_{j+1}=0}&=\Sigma_I([V_1,V_j]p_I)S_I(t),\ \text{using (3.2)},\\&=\Sigma_I p_I([L_j^{(n)},L_1^{(n)}]S_I(t)),\ \text{by induction hypothesis},\\&=(j-1)L_{j+1}^{(n)}\tau_n(t),\ \text{using (3.5)}.\end{aligned}\tag{3.26}$$

This establishes Theorem III.1.

Using Theorem III.1 and formula (2.4), which gives the entries a_n and b_n of the tridiagonal matrix L in (2.2) in terms of $\{\tau_n\}$, it is possible to obtain an expression for the master symmetries V_j, $j\geqslant -1$, acting on L, in terms of a Lax pair. Let us define P to be the strictly lower matrix, which represents d/dx in the basis formed by the *monic* orthogonal polynomials $p_n(x)$:

$$\frac{d}{dx}p=Pp,\quad \text{with } p=(p_0(x)=1,p_1(x),p_2(x),\ldots)^T.\tag{3.27}$$

For a semi-infinite matrix $A=(a_{ij})_{i,j\geqslant 0}$, we denote by $\operatorname{tr}A=(a_{00},a_{00}+a_{11},a_{00}+a_{11}+a_{22},\ldots)$, the diagonal matrix formed with the partial traces of A. Also, A_- means the strictly lower part of A, and A_{++}, the strictly upper part of A. With these notations, we have

$$\frac{\partial L}{\partial s_j}=\left[-(PL^{j+1})_- - \sum_{k=0}^{j-1}(\operatorname{tr}L^k)L_-^{j-k}-\sum_{k=1}^{j-2}(L_{++}^kL_-^{j-k})_-,L\right],\quad j\geqslant -1.\tag{3.28}$$

This Lax pair is different from the one obtained by Adler and van Moerbeke,[16] which requires L to be in symmetric form. We shall not pursue this aspect here, since it will not be needed in what follows. We refer the interested reader to the Ph.D. thesis of the second author,[17] for a detailed proof. To the best of our knowledge, the theory of master symmetries of Toda-type systems has not yet been extended beyond tridiagonal matrices. We feel that (3.28) may be useful in this regard since, in general, a finite-band matrix cannot be assumed to be in symmetric form. As a final comment, we mention that, specializing our theory to the case of discrete orthogonal polynomials, we find back the master symmetries of the finite Toda lattice, which were studied by Damianou[18] and Fernandes.[19] However, a general Lax pair formulation was not given in these works.

IV. THE JACOBI POLYNOMIALS AND THE PAINLEVÉ VI EQUATION

In this section, we illustrate the use of Theorem III.1. The choice of the Jacobi polynomials is motivated by the fact that, as explained in the Introduction, the study of this example allows us to connect the Jacobi polynomial ensemble to the Painlevé VI equation. To the best of our knowledge this result is new.

Let α and β denote real numbers with $\alpha>-1$ and $\beta>-1$. We consider the following curve in the space of moments:

$$\mu_k(z)=\int_z^1 x^k(1-x)^\alpha(1+x)^\beta\,dx,\quad z\in[-1,1]. \tag{4.1}$$

When $z=-1$, the μ_k's are the moments of the orthogonality measure for the Jacobi polynomials. Thinking of z as a parameter, we shall denote by $\tau_n(z,t)$ the tau functions associated with the sequence $\{\mu_k(z)\}$, i.e.,

$$\tau_n(z,t)=\det[(\mu_{i+j}(z))(t)]_{0\leqslant i,j\leqslant n-1}. \tag{4.2}$$

The key to our story is the following.

Lemma IV.1: For all $k\geqslant 0$ and for all $j\geqslant 0$, we have

$$z^j(z^2-1)\frac{d}{dz}\mu_k(z)=[V_{j+1}-V_{j-1}+(\alpha+\beta)T_j+(\alpha-\beta)T_{j-1}]\mu_k(z), \tag{4.3}$$

where the notation on the right-hand side of (4.3) means the Lie derivative in the direction of the vector field between the brackets.

Proof: By definition of the vector fields T_j and V_j in (2.9) and (3.1), the right-hand side of (4.3) is equal to

$$\begin{aligned}
&(k+j+2+\alpha+\beta)\mu_{k+j+1}(z)+(\alpha-\beta)\mu_{k+j}(z)-(k+j)\mu_{k+j-1}(z),\\
&=\int_z^1[(i+2+\alpha+\beta)x^{i+1}+(\alpha-\beta)x^i-ix^{i-1}](1-x)^\alpha(1+x)^\beta\,dx,\quad \text{with}\ \ i=j+k,\\
&=\int_z^1-\frac{d}{dx}[x^i(1-x)^{\alpha+1}(1+x)^{\beta+1}]dx,\\
&=\text{the left-hand side of (4.3)},
\end{aligned} \tag{4.4}$$

which proves the lemma.

It was observed in Grünbaum and Haine[7] that the vector fields,

$$\mathcal{V}_j=V_{j+1}-V_{j-1}+(\alpha+\beta)T_j+(\alpha-\beta)T_{j-1},\quad j=0,1,2,..., \tag{4.5}$$

form a subalgebra of the algebra of master symmetries,

$$[\mathcal{V}_i,\mathcal{V}_j]=(j-i)(\mathcal{V}_{i+j+1}-\mathcal{V}_{i+j-1}), \tag{4.6}$$

and that the *vanishing* of *all* these vector fields characterize the Jacobi polynomials, which correspond to $z=-1$. In a similar spirit, we can interpret the lemma above as saying that *all* vector fields $\mathcal{V}_j$, $j\geqslant 0$, become *tangent* along the curve (4.1).

Combining Lemma IV.1 with Theorem III.1, we obtain the following.

Corollary IV.2: The tau functions $\tau_n(z,t)$, as defined in (4.2), satisfy the following ''Virasoro type'' constraints,

$$z^j(z^2-1)\frac{\partial}{\partial z}\tau_n(z,t)=\left[L_{j+1}^{(n)}-L_{j-1}^{(n)}+(\alpha+\beta)\frac{\partial}{\partial t_{j+1}}+(\alpha-\beta)\frac{\partial}{\partial t_j}\right]\tau_n(z,t), \tag{4.7}$$

for $j=0,1,2,...$, with $L_j^{(n)}$ defined as in (3.4), *where $\partial/\partial t_0\ \tau_n(z,t)$ is interpreted as $n\tau_n(z,t)$.*

Proof: Expanding (4.2) in Plücker coordinates as in (3.8), and using Lemma IV.1, we have that

$$z^j(z^2-1)\frac{\partial}{\partial z}\tau_n(z,t)=\sum_{0\leq i_1<\cdots<i_n}\mathcal{V}_j p_{i_1,\ldots,i_n}(z)S_{i_n-n+1,\ldots,i_1}(t), \tag{4.8}$$

where $p_{i_1,\ldots,i_n}(z)$ are the Plücker coordinates (3.9) formed with the moments $\mu_k(z)$, and $\mathcal{V}_j p_{i_1,\ldots,i_n}(z)$ denotes the Lie derivative in the direction of the vector field $\mathcal{V}_j$, defined in (4.5). Equation (4.8) can be rewritten as

$$z^j(z^2-1)\frac{\partial}{\partial z}\tau_n(z,t)=\frac{\partial}{\partial s_{j+1}}\tau_n^{s_{j+1}}(z,t)\bigg|_{s_{j+1}=0}-\frac{\partial}{\partial s_{j-1}}\tau_n^{s_{j-1}}(z,t)\bigg|_{s_{j-1}=0}$$
$$+\left[(\alpha+\beta)\frac{\partial}{\partial t_{j+1}}+(\alpha-\beta)\frac{\partial}{\partial t_j}\right]\tau_n(z,t). \tag{4.9}$$

By Theorem III.1, the right-hand side of (4.9) is identical with the right-hand side of (4.7), which establishes the corollary.

We shall now show that the Virasoro constraints (4.7) combined with the KP equation (2.21) imply that the functions

$$r(z)=(z^2-1)\frac{d}{dz}\log\tau_n(z),\quad n=1,2,3,\ldots, \tag{4.10}$$

with $\tau_n(z)\equiv\tau_n(z,0)$, all satisfy (after some manipulations) the Painlevé VI equation. Put

$$f(z,t)=\log\tau_n(z,t). \tag{4.11}$$

From now on we fix the integer n and, for this reason, we drop n from our notations.

Remembering the definition of $L_j^{(n)}$ in (3.4), we can rewrite the first three Virasoro constraints in (4.7), for $j=0,1,2$, as follows:

$$(z^2-1)\frac{\partial f}{\partial z}=n(\alpha-\beta-t_1)+(2n+\alpha+\beta)\frac{\partial f}{\partial t_1}+t_1\frac{\partial f}{\partial t_2}$$
$$+\sum_{k=2}^{\infty}kt_k\left(\frac{\partial}{\partial t_{k+1}}-\frac{\partial}{\partial t_{k-1}}\right)f, \tag{4.12}$$

$$z(z^2-1)\frac{\partial f}{\partial z}=-n^2+(\alpha-\beta)\frac{\partial f}{\partial t_1}+(2n+\alpha+\beta)\frac{\partial f}{\partial t_2}+\frac{\partial^2 f}{\partial t_1^2}+\left(\frac{\partial f}{\partial t_1}\right)^2$$
$$+\sum_{k=1}^{\infty}kt_k\left(\frac{\partial}{\partial t_{k+2}}-\frac{\partial}{\partial t_k}\right)f, \tag{4.13}$$

$$z^2(z^2-1)\frac{\partial f}{\partial z}=-2n\frac{\partial f}{\partial t_1}+(\alpha-\beta)\frac{\partial f}{\partial t_2}+(2n+\alpha+\beta)\frac{\partial f}{\partial t_3}$$
$$+2\frac{\partial^2 f}{\partial t_1\partial t_2}+2\frac{\partial f}{\partial t_1}\frac{\partial f}{\partial t_2}+\sum_{k=1}^{\infty}kt_k\left(\frac{\partial}{\partial t_{k+3}}-\frac{\partial}{\partial t_{k+1}}\right)f. \tag{4.14}$$

These three equations can be exploited to express all the partial t derivatives of f, involved in the KP equation (2.21), at $t\equiv(t_1,t_2,t_3,\ldots)=(0,0,0,\ldots)$, in terms of z derivatives of the function $r(z)$ defined in (4.10). We note that a similar method[20] was already used in the case of the tau functions associated with the Laguerre polynomials, to show that the Virasoro constraints uniquely characterize these tau functions (up to a constant).

The constraint (4.12) evaluated at $t\equiv(t_1,t_2,t_3,\dots)=(0,0,0,...)$ gives, remembering the definition of $r(z)$ in (4.10),

$$\left.\frac{\partial f}{\partial t_1}\right|_{t=0}=\frac{r(z)-n(\alpha-\beta)}{2n+\alpha+\beta}. \tag{4.15}$$

We then proceed by induction. We call

$$\frac{\partial^n f}{\partial t_{j_1}\cdots\partial t_{j_n}}, \tag{4.16}$$

a t derivative of weighted degree $|j|=j_1+\cdots+j_n$. Then, for $k\geqslant 1$, we compute the system formed by

$$\begin{aligned}&\text{all } t-\text{derivatives of weighted degree } k \text{ of the constraint (4.12)},\\ &\text{all } t-\text{derivatives of weighted degree } k-1 \text{ of the constraint (4.13)},\\ &\text{all } t-\text{derivatives of weighted degree } k-2 \text{ of the constraint (4.14)},\\ &\qquad\qquad\qquad\qquad\text{evaluated at } t=0.\end{aligned} \tag{4.17}$$

For instance, for $k=1$, (4.17) reduces to

$$\begin{aligned}(z^2-1)\frac{d}{dz}\left(\left.\frac{\partial f}{\partial t_1}\right|_{t=0}\right)&=-n+(2n+\alpha+\beta)\left.\frac{\partial^2 f}{\partial t_1^2}\right|_{t=0}+\left.\frac{\partial f}{\partial t_2}\right|_{t=0},\\ zr(z)=-n^2+(\alpha-\beta)\left.\frac{\partial f}{\partial t_1}\right|_{t=0}&+(2n+\alpha+\beta)\left.\frac{\partial f}{\partial t_2}\right|_{t=0}+\left.\frac{\partial^2 f}{\partial t_1^2}\right|_{t=0}+\left(\left.\frac{\partial f}{\partial t_1}\right|_{t=0}\right)^2,\end{aligned} \tag{4.18}$$

which, after substitution of (4.15), can be solved for $\partial^2 f/\partial t_1^2|_{t=0}$ and $\partial f/\partial t_2|_{t=0}$ in terms of $r(z)$ and $r'(z)$.

In general, the system (4.17) is an (overdetermined) system that can be solved for all t derivatives of f of weighted degree $k+1$, evaluated at $t=0$, in terms of lower (weighted) degree t derivatives of f (at $t=0$). Assuming, by induction, that the latter have already been expressed in terms of $r(z)$ and its first $k-1$ derivatives, we conclude that all t derivatives of f of weighted degree $k+1$, at $t=0$, can be expressed purely in z, in terms of $r(z)$ and its first k derivatives.

Since the KP equation (2.21) contains t derivatives of f of weighted degree less or equal to 4, by performing the above scheme up to $k=3$, we can express all these derivatives, evaluated at $t=0$, in terms of $r(z)$ and its first three derivatives, which gives us *a third-order* differential equation for $r(z)$:

$$\begin{aligned}(1-z^2)^2r'''=&\,2z(1-z^2)r''+6(1-z^2)(r')^2+8zrr'\\ &-[4n^2+4(\alpha+\beta)n-(\alpha-\beta)^2+2(\beta^2-\alpha^2)z-(2n+\alpha+\beta)^2z^2]r'\\ &-2r^2-[\alpha^2-\beta^2+(2n+\alpha+\beta)^2z]r.\end{aligned} \tag{4.19}$$

Now Tracy and Widom,[3] using a different method, find that $r(z)$ satisfies *another third-order* differential equation,

$$\begin{aligned}(1-z^2)^2r'''=&\,(1-z^2)^2\frac{(r'')^2}{r'}+2z(1-z^2)r''+2(1-z^2)(r')^2\\ &+2\left(1-\frac{2\alpha_1^2}{r'}\right)r^2+4\alpha_1(\alpha_0+\alpha_1 z)r,\end{aligned} \tag{4.20}$$

with

$$\alpha_0=\frac{\beta^2-\alpha^2}{2(2n+\alpha+\beta)},\quad \alpha_1=-\left(n+\frac{\alpha+\beta}{2}\right). \tag{4.21}$$

Substracting Eq. (4.20) from Eq. (4.19), shows that, in fact, $r(z)$ does satisfy a *second-order* differential equation. We now show that this second-order equation is nothing but the celebrated Painlevé VI equation in disguised form.

Let us put

$$r(z)=2\rho\left(\frac{z+1}{2}\right), \tag{4.22}$$

which amounts to normalizing the Jacobi polynomials to be orthogonal on the interval [0,1], and to reparametrize our curve of moments (4.1) in terms of the variable

$$s=\frac{z+1}{2}. \tag{4.23}$$

Then, one computes that $\rho(s)$ satisfies the following second-order equation:

$$\begin{aligned}(4.19)-(4.20)\Leftrightarrow &(s(s-1)\rho'')^2+4s(s-1)(\rho')^3+4\rho'\rho^2-(2n+\alpha+\beta)^2\rho^2+4(1-2s)\rho(\rho')^2\\ &-2[2n^2+2(\alpha+\beta)n+\beta(\alpha+\beta)-(2n+\alpha+\beta)^2s]\rho\rho'\\ &-[\beta^2-2(2n^2+2(\alpha+\beta)n+\beta(\alpha+\beta))s+(2n+\alpha+\beta)^2s^2](\rho')^2=0.\end{aligned} \tag{4.24}$$

We can now state the following.

Theorem IV.3: *Let $r(z)=(z^2-1)(d/dz)\log\tau_n(z)$ be defined as in (4.10). Then, the function*

$$\sigma(s)=\frac{1}{2}r(2s-1)-\frac{(2n+\alpha+\beta)^2}{4}s+\frac{2n^2+2(\alpha+\beta)n+\alpha(\alpha+\beta)}{4}, \tag{4.25}$$

satisfies the Okamoto–Jimbo–Miwa form[21] *of the Painlevé VI equation:*

$$\begin{aligned}&\sigma'(s(s-1)\sigma'')^2+\{(2s-1)(\sigma')^2-2\sigma\sigma'-\nu_1\nu_2\nu_3\nu_4\}^2\\ &=(\sigma'+\nu_1^2)(\sigma'+\nu_2^2)(\sigma'+\nu_3^2)(\sigma'+\nu_4^2),\end{aligned} \tag{4.26}$$

with ν_i, $1\leqslant i\leqslant 4$, specified in terms of n, α and β as in the equations (4.28), (4.31), (4.32), (4.33) and (4.34) below.

Proof: Observe that, in fact, σ' factors out of Eq. (4.26), so that it can be written in the less elegant form

$$\begin{aligned}&(s(s-1)\sigma'')^2+4s(s-1)(\sigma')^3+4\sigma'\sigma^2+4(1-2s)\sigma(\sigma')^2\\ &-c_1(\sigma')^2+[2(1-2s)c_4-c_2]\sigma'+4c_4\sigma-c_3=0,\end{aligned} \tag{4.27}$$

with

$$c_1=\sum_i \nu_i^2,\quad c_2=\sum_{i<j}\nu_i^2\nu_j^2,$$

$$c_3=\sum_{i<j<k}\nu_i^2\nu_j^2\nu_k^2,\quad c_4=\nu_1\nu_2\nu_3\nu_4. \tag{4.28}$$

We try to match (4.24) with (4.27) by putting

$$\rho(s)=\sigma(s)+xs+y, \tag{4.29}$$

and determining the constants x and y to eliminate the coefficients of ρ^2 and $\rho\rho'$ in (4.24). We find that

$$x=\frac{(2n+\alpha+\beta)^2}{4} \quad \text{and} \quad y=-\frac{2n^2+2(\alpha+\beta)n+\alpha(\alpha+\beta)}{4}. \tag{4.30}$$

With this choice of x and y, the new function $\sigma(s)$ satisfies Eq. (4.27) if we take

$$c_1=2n^2+2(\alpha+\beta)n+\alpha^2+\alpha\beta+\beta^2, \tag{4.31}$$

$$\begin{aligned} c_2=\tfrac{1}{8}[&3(\alpha+\beta)^2(\alpha^2+\beta^2)+4(\alpha+\beta)(3\alpha^2+2\alpha\beta+3\beta^2)n \\ &+4(5\alpha^2+6\alpha\beta+5\beta^2)n^2+16(\alpha+\beta)n^3+8n^4], \end{aligned} \tag{4.32}$$

$$\begin{aligned} c_3=&\frac{(2n+\alpha+\beta)^2}{16} \\ &\times[(\alpha+\beta)^2(\alpha^2-\alpha\beta+\beta^2)+2(\alpha+\beta)(\alpha^2+\beta^2)n+2(\alpha^2+\beta^2)n^2], \end{aligned} \tag{4.33}$$

$$c_4=\frac{(\alpha^2-\beta^2)(2n+\alpha+\beta)^2}{16}. \tag{4.34}$$

Remembering the relations between the functions $r(z)$, $\rho(s)$, and $\sigma(s)$ in (4.22), (4.29), and (4.30), Theorem IV.3 is established.

Remarks: (1) One can check that

$$\begin{aligned} &2r''(z)\times(\text{left-hand side–right-hand side) of (4.19)} \\ &=\frac{d}{dz}r'(z)\times[\text{right-hand side of (4.20)–right-hand side of (4.19)}]. \end{aligned} \tag{4.35}$$

This shows that, in fact, the third-order equation (4.19), which is obtained by the ''Virasoro method,'' can indeed be integrated into a second-order equation, which we know to be equivalent [via (4.25)] to the Painlevé VI equation. Thus, in principle, we could have derived our result without using the equation (4.20) found by Tracy and Widom. But (4.35) is not so easy to guess *a priori*!

(2) From our point of view, it is no more difficult to obtain systems of partial differential equations satisfied by the probability of finding no eigenvalues in a disjoint union of intervals $\cup_{i=1}^{r}[z_{2i-1},z_{2i}]\subset[-1,1]$, as a function of the boundary points. One just needs to consider the ''manifolds'' in the space of moments,

$$\mu_k(z_1,\ldots,z_{2r})=\sum_{i=0}^{r}\int_{z_{2i}}^{z_{2i+1}}x^k(1-x)^\alpha(1+x)^\beta\,dx, \tag{4.36}$$

with $z_0=-1$ and $z_{2r+1}=1$, the crucial point being that the quantity under the integral sign in (4.4) is an *exact derivative*. One then replaces the ordinary differential operator in z on the left-hand side of (4.3) by the partial differential operator

$$\sum_{i=1}^{2r} z_i^j(z_i^2-1)\frac{\partial}{\partial z_i}. \tag{4.37}$$

Using a recipe similar to the one described in (4.17), one can express all the time derivatives of the logarithmic derivative of the tau function, evaluated at time $t=0$, in terms of partial derivatives with respect to the variables z_i, $1\leqslant i\leqslant 2r$. Substituting the result into Hirota's hierarchy of bilinear equations satisfied by the tau function, produces the desired system of partial differential equations.

APPENDIX: THE HERMITE AND LAGUERRE ENSEMBLES

For the sake of completeness, we write the analogues of Eqs. (4.3) and (4.7), in the cases of the Laguerre and the Hermite polynomials, though it does not lead to new results on the corresponding orthogonal polynomial ensembles. Only the methodology is different.

In the case of the Laguerre polynomials the curve of moments,

$$\mu_k(z)=\int_z^\infty x^{k+\alpha}e^{-x}\,dx, \quad z\in[0,\infty[,\alpha>-1, \tag{A1}$$

satisfies

$$z^{j+1}\frac{d}{dz}\mu_k(z)=(V_j-T_j+\alpha T_{j-1})\mu_k(z), \quad j=0,1,2,..., \tag{A2}$$

which, using Theorem III.1, immediately leads to the Virasoro constraints on $\tau_n(z,t)$:

$$z^{j+1}\frac{\partial}{\partial z}\tau_n(z,t)=\left(L_j^{(n)}-\frac{\partial}{\partial t_{j+1}}+\alpha\frac{\partial}{\partial t_j}\right)\tau_n(z,t), \quad j=0,1,2,\dots . \tag{A3}$$

The functions

$$\sigma(z)=z\frac{d}{dz}\log\tau_n(z), \tag{A4}$$

satisfy the Okamoto–Jimbo–Miwa form[21] of the Painlevé V equation:

$$\begin{aligned}(z\sigma'')^2=(\sigma-z\sigma'+2(\sigma')^2+(\nu_0+\nu_1+\nu_2+\nu_3)\sigma')^2\\ -4(\sigma'+\nu_0)(\sigma'+\nu_1)(\sigma'+\nu_2)(\sigma'+\nu_3),\end{aligned} \tag{A5}$$

with

$$\nu_0=n, \quad \nu_1=n+\alpha, \quad \nu_2=\nu_3=0. \tag{A6}$$

For the Hermite polynomials, the curve of moments,

$$\mu_k(z)=\int_z^\infty x^k e^{-x^2}\,dx, \quad z\in]-\infty,\infty[, \tag{A7}$$

satisfies

$$z^{j+1}\frac{d}{dz}\mu_k(z)=(V_j-2T_{j+1})\mu_k(z), \quad j=-1,0,1,2,..., \tag{A8}$$

leading to the Virasoro constraints

$$z^{j+1}\frac{\partial}{\partial z}\tau_n(z,t)=\left(L_j^{(n)}-2\frac{\partial}{\partial t_{j+2}}\right)\tau_n(z,t), \quad j=-1,0,1,2,\dots . \tag{A9}$$

The functions

$$\sigma(z)=\frac{d}{dz}\log\tau_n(z), \tag{A10}$$

satisfy the Okamoto–Jimbo–Miwa form[21] of the Painlevé IV equation:

$$(\sigma'')^2=4(z\sigma'-\sigma)^2-4(\sigma'+\nu_0)(\sigma'+\nu_1)(\sigma'+\nu_2), \tag{A11}$$

with

$$\nu_0=2n, \quad \nu_1=\nu_2=0. \tag{A12}$$

We notice that, both in (A2) and (A8), the subalgebra $\{\mathcal{V}_i\}$ of the algebra of master symmetries, which becomes tangent along the curves of moments (A1) and (A7), satisfies the commutation relations

$$[\mathcal{V}_i,\mathcal{V}_j]=(j-i)\mathcal{V}_{i+j}. \tag{A13}$$

Equations (A5) and (A11) were first obtained by Tracy and Widom.[3] When $z\to 0$ (respectively, $z\to-\infty$), the left-hand side of (A2) [respectively, (A8)] vanishes; the corresponding Virasoro constraints for the tau functions (A3) [respectively (A9)] were obtained by Haine and Horozov.[6] The introduction of a ''boundary-part'' into (A3) and (A9) appears in Adler, Shiota, and van Moerbeke.[5] We hope that the methodology developed in this paper will bring further progress in the field.

ACKNOWLEDGMENTS

We thank the referee for suggesting a clarification in the exposition. The first author (L.H.) is a Research Associate of the Belgian National Fund for Scientific Research and is grateful for its support.

[1] M. L. Mehta, *Random Matrices*, 2nd ed. (Academic, San Diego, 1991).
[2] C. A. Tracy and H. Widom, ''Introduction to random matrices,'' in *Geometric and Quantum Aspects of Integrable Systems*, Lecture Notes in Physics, edited by G. F. Helminck (Springer-Verlag, Berlin, 1993), Vol. 424, pp. 103–130.
[3] C. A. Tracy and H. Widom, ''Fredholm determinants, differential equations and matrix models,'' Commun. Math. Phys. **163**, 33–72 (1994).
[4] M. Jimbo, T. Miwa, Y. Môri, and M. Sato, ''Density matrix of an impenetrable Bose gas and the fifth Painlevé transcendent,'' Physica D **1**, 80–158 (1980).
[5] M. Adler, T. Shiota, and P. van Moerbeke, ''Random matrices, vertex operators and the Virasoro algebra,'' Phys. Lett. A **208**, 67–78 (1995).
[6] L. Haine and E. Horozov, ''Toda orbits of Laguerre polynomials and representations of the Virasoro algebra,'' Bull. Sci. Math. **117**, 485–518 (1993).
[7] F. A. Grünbaum and L. Haine, ''A theorem of Bochner, revisited,'' in *Algebraic Aspects of Integrable Systems: In Memory of Irene Dorfman*, Progress in Nonlinear Differential Equations, edited by A. S. Fokas and I. M. Gelfand (Birkhäuser, Boston, 1997), Vol. 26, pp. 143–172.
[8] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis* (Hafner, New York, 1965).
[9] T. S. Chihara, *An Introduction to Orthogonal Polynomials, Mathematics and its Applications* (Gordon and Breach, New York, 1978), Vol. 13.
[10] L. Haine, ''Beyond the classical orthogonal polynomials,'' in *The Bispectral Problem*, CRM Proc. Lecture Notes, edited by J. Harnad and A. Kasman (American Mathematical Society, Providence, RI, 1998), Vol. 14, pp. 47–65.
[11] S. Bochner, ''Über Sturm–Liouvillesche polynomsysteme,'' Math. Z. **29**, 730–736 (1929).
[12] M. Sato and Y. Sato, ''Soliton equations as dynamical systems on infinite dimensional Grassmann manifolds,'' in *Nonlinear Partial Differential Equations in Applied Science*, North-Holland Math. Stud. 81, edited by H. Fujita, P. D. Lax, and G. Strang (Kinokuniya/North-Holland, Tokyo, 1983), pp. 259–271.
[13] J. Moser, ''Finitely many mass points on the line under the influence of an exponential potential—An integrable system,'' in *Dynamical Systems, Theory and Applications*, Lecture Notes in Physics, edited by J. Moser (Springer-Verlag, Berlin, 1975), Vol. 38, pp. 467–497.
[14] E. Witten, ''Two-dimensional gravity and intersection theory on moduli space,'' Surveys in Differential Geometry 1, A Supplement to the Journal of Differential Geometry, edited by C. C. Hsiung and S. T. Yau (1991), pp. 243–310.

[15] V. G. Kac and A. K. Raina, *Bombay Lectures on Highest Weight Representations of Infinite Dimensional Lie Algebras*, Adv. Series in Math. Phys. (World Scientific, Singapore, 1987), Vol. 2.
[16] M. Adler and P. van Moerbeke, ''Matrix integrals, Toda symmetries, Virasoro constraints, and orthogonal polynomials,'' Duke Math. J. **80**, 863–911 (1995).
[17] J. P. Semengue, Ph.D. thesis, Université catholique de Louvain, 1998.
[18] P. Damianou, ''Master symmetries and R-matrices for the Toda lattice,'' Lett. Math. Phys. **20**, 101–112 (1990).
[19] R. L. Fernandes, ''On the master symmetries and bi-Hamiltonian structure of the Toda lattice,'' J. Phys. A **26**, 3797–3803 (1993).
[20] L. Haine and E. Horozov, ''Tau-functions and modules over the Virasoro algebra,'' in *Abelian Varieties*, Proceedings of the International Conference held in Egloffstein, Germany, 3–8 October 1993, edited by W. Barth, K. Hulek, and H. Lange (Walter de Gruyter, Berlin, 1995), pp. 85–104.
[21] M. Jimbo and T. Miwa, ''Monodromy preserving deformation of linear ordinary differential equations with rational coefficients. II,'' Physica D **2**, 407–448 (1981).

Representations of the cyclically symmetric *q*-deformed algebra $so_q(3)$

M. Havlíček
Department of Mathematics, FNSPE, Czech Technical University, CZ-120 00, Prague 2, Czech Republic

A. U. Klimyk
Institute for Theoretical Physics, Kiev 252143, Ukraine

S. Pošta
Department of Mathematics, FNSPE, Czech Technical University, CZ-120 00, Prague 2, Czech Republic

(Received 18 March 1998; accepted for publication 30 September 1998)

An algebra homomorphism ψ from the nonstandard q-deformed (cyclically symmetric) algebra $U_q(so_3)$ to the extension $\hat{U}_q(sl_2)$ of the Hopf algebra $U_q(sl_2)$ is constructed. Not all irreducible representations of $U_q(sl_2)$ can be extended to representations of $\hat{U}_q(sl_2)$. Composing the homomorphism ψ with irreducible representations of $\hat{U}_q(sl_2)$ we obtain representations of $U_q(so_3)$. Not all of these representations of $U_q(so_3)$ are irreducible. Reducible representations of $U_q(so_3)$ are decomposed into irreducible components. In this way we obtain all irreducible representations of $U_q(so_3)$ when q is not a root of unity. A part of these representations turns into irreducible representations of the Lie algebra so_3 when $q\to 1$. Representations of the other part have no classical analog. Using the homomorphism ψ it is shown how to construct tensor products of finite-dimensional representations of $U_q(so_3)$. Irreducible representations of $U_q(so_3)$ when q is a root of unity are constructed. Some of them are obtained from irreducible representations of $\hat{U}_q(sl_2)$ by means of the homomorphism ψ. *© 1999 American Institute of Physics.* [S0022-2488(99)01003-8]

I. INTRODUCTION

It is well known that the Lie algebras sl_2 and so_3 of the Lie groups $SL(2,C)$ and $SO(3)$, respectively, are isomorphic. But these algebras differ from each other if we consider their embedding to the wider Lie algebra sl_3. There is no automorphism of sl_3 which transfers the embedding $sl_2\subset sl_3$ to the embedding $so_3\subset sl_3$. Note that the embedding $so_3\subset sl_3$ is of great importance for nuclear physics: it is used in spectroscopy.

The definition of the q-analog of the universal enveloping algebra $U(sl_2)$ is well known. It is the quantum algebra $U_q(sl_2)$ which is a Hopf algebra. If we wish to have a q-analog of the universal enveloping algebra so_3 such that at $q\to 1$ we obtain the classical embedding $so_3\subset sl_3$, then the algebra sl_2 is not appropriate for this role. By other words, an algebra $U_q(so_3)$ must differ from $U_q(sl_2)$. This algebra $U_q(so_3)$ is well known. It is the associative algebra generated by three elements I_1, I_2, and I_3 satisfying the relations

$$q^{1/2}I_1I_2-q^{-1/2}I_2I_1=I_3, \tag{1}$$

$$q^{1/2}I_2I_3-q^{-1/2}I_3I_2=I_1, \tag{2}$$

$$q^{1/2}I_3I_1-q^{-1/2}I_1I_3=I_2. \tag{3}$$

Such (and more general) deformation of the commutator $[I_i, I_j]=I_iI_j-I_jI_i$ was defined in 1967 by R. Santilli in Ref. 1 (see also Refs. 2 and 3) under studying a generalization of the Lie theory. Afterwards (in 1990), the algebra $U_q(so_3)$ with commutation relations (1)–(3) was determined by D. Fairlie (see Ref. 4). An algebra which can be reduced to $U_q(so_3)$ was defined in 1986 by M. Odesski in Ref. 5.

Fairlie gave finite-dimensional irreducible representations of the algebra $U_q(so_3)$ which at $q\to 1$ give the well-known finite-dimensional irreducible representations of the Lie algebra so_3 (see Ref. 4). These representations are given by integral or half-integral non-negative numbers. Odesski also gave some classes of irreducible representations in Ref. 5.

It was shown in Refs. 5–7 that the algebra $U_q(so_3)$ has irreducible finite-dimensional representations which have no classical analog (that is, which do not admit the limit $q\to 1$). It was not clear why such strange representations of the algebra $U_q(so_3)$ appear. What is their nature? The answer to this question is one of the aims of this paper.

We construct a homomorphism from $U_q(so_3)$ to the algebra $\hat{U}_q(sl_2)$ which is an extension of the well-known quantum algebra $U_q(sl_2)$ [note that there is no homomorphism from $U_q(so_3)$ to $U_q(sl_2)$]. Irreducible finite-dimensional representations of $U_q(sl_2)$ (but not all) can be extended to finite-dimensional representations of the algebra $\hat{U}_q(sl_2)$. Composing a homomorphism $U_q(so_3)\to\hat{U}_q(sl_2)$ with these representations of $\hat{U}_q(sl_2)$, we obtain representations of the algebra $U_q(so_3)$. But some irreducible representations of $\hat{U}_q(sl_2)$ lead to reducible representations of the algebra $U_q(so_3)$. Decomposing these reducible representations of $U_q(so_3)$ we obtain irreducible representations of this algebra which have no analog for the Lie algebra so_3. If q is not a root of unity, then in this way we obtain all finite-dimensional irreducible representations of $U_q(so_3)$. But there are infinite-dimensional irreducible representations of $U_q(so_3)$ which cannot be obtained in this way.

Existence of the homomorphism $U_q(so_3)\to\hat{U}_q(sl_2)$ allows us to define tensor products of representations of the algebra $U_q(so_3)$ which is not a Hopf algebra.

Using the homomorphism $U_q(so_3)\to\hat{U}_q(sl_2)$ and irreducible representations of $\hat{U}_q(sl_2)$ we obtain representations of $U_q(so_3)$ when q is a root of unity. Taking irreducible representations of $U_q(so_3)$ obtained in this way and decomposing reducible representations, we obtain several series of irreducible representations of $U_q(so_3)$. In addition, we construct irreducible representations of $U_q(so_3)$ which cannot be derived from $\hat{U}_q(sl_2)$.

When q is not a root of unity, then each irreducible (finite or infinite dimensional) representation of $U_q(so_3)$ is equivalent to one of the representations constructed below. (We do not give a proof of this assertion in this paper because it would take much space; this proof will be given in a separate paper.) We think that in this paper we constructed also all irreducible representations of $U_q(so_3)$ when q is a root of unity. But in this case we have no proof of this assertion. The reason of this is that in this case there are many classes of irreducible representations and a proof of completeness of irreducible representations becomes very tedious.

Let us remark that in Ref. 5 there were constructed irreducible finite-dimensional representations of $U_q(so_3)$ when q is not a root of unity and a part of irreducible infinite-dimensional representations. In Refs. 6 and 7, there were constructed irreducible representations of $U_q(so_3)$ which satisfy the conditions of $*$-representations [that is, such that $T(I_j^*)=-T(I_j)$, $j=1,2$]. These $*$-representations are a part of irreducible representations of $U_q(so_3)$ constructed in this paper. We started to study irreducible representations of $U_q(so_3)$ for q a root of unity in Ref. 8, where a part of irreducible representations for this case were constructed. Note that in Refs. 5–8 there are no relations of representations of $U_q(so_3)$ to representations of $\hat{U}_q(sl_2)$. This relation makes representations of $U_q(so_3)$ clear and understandable.

We suppose that in Secs. II and III q is any complex number different from -1. In Secs. IV–VII, q is not a root of unity. In Secs. VIII–X, q is a root of unity.

II. THE ALGEBRAS $\mathrm{U}_q(\mathrm{so}_3)$ AND $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$

The algebra $\mathrm{U}_q(\mathrm{so}_3)$ is obtained by a q-deformation of the standard commutation relations

$$[I_1,I_2]=I_3,\quad [I_2,I_3]=I_1,\quad [I_3,I_1]=I_2$$

of the Lie algebra so_3. So, $\mathrm{U}_q(\mathrm{so}_3)$ is defined as the complex associative algebra with unit element generated by the elements I_1,I_2,I_3 satisfying the defining relations

$$[I_1,I_2]_q:=q^{1/2}I_1I_2-q^{-1/2}I_2I_1=I_3, \tag{4}$$

$$[I_2,I_3]_q:=q^{1/2}I_2I_3-q^{-1/2}I_3I_2=I_1, \tag{5}$$

$$[I_3,I_1]_q:=q^{1/2}I_3I_1-q^{-1/2}I_1I_3=I_2. \tag{6}$$

Unfortunately, a Hopf algebra structure is not known on $\mathrm{U}_q(\mathrm{so}_3)$. However, it can be embedded into the Hopf algebra $\mathrm{U}_q(\mathrm{sl}_3)$ as a Hopf coideal (see Ref. 9). This embedding is very important for the possible application in spectroscopy.

It follows from the relations (4)–(6) that for the algebra $\mathrm{U}_q(\mathrm{so}_3)$ the Poincaré–Birkhoff–Witt theorem is true and this theorem can be formulated as: *The elements* $I_1^kI_2^mI_3^n$, $k,m,n=0,1,2,...$, *form a basis of the linear space* $\mathrm{U}_q(\mathrm{so}_3)$. Indeed, by using the relations (4)–(6) any product $I_{j_1}I_{j_2}\cdots I_{j_s}$, $j_1,j_2,\cdots,j_s=1,2,3$, can be reduced to a sum of the elements $I_1^kI_2^mI_3^n$ with complex coefficients.

Note that by (4) the element I_3 is not independent: it is determined by the elements I_1 and I_2. Thus, the algebra $\mathrm{U}_q(\mathrm{so}_3)$ is generated by I_1 and I_2, but now instead of quadratic relations (4)–(6) we must take the relations

$$I_1I_2^2-(q+q^{-1})I_2I_1I_2+I_2^2I_1=-I_1, \tag{7}$$

$$I_2I_1^2-(q+q^{-1})I_1I_2I_1+I_1^2I_2=-I_2, \tag{8}$$

which are obtained if we substitute the expression (4) for I_3 into (5) and (6). The equation $I_3=q^{1/2}I_1I_2-q^{-1/2}I_2I_1$ and the relations (7) and (8) restore the relations (4)–(6).

Remark that the definition of $\mathrm{U}_q(\mathrm{so}_3)$ by means of relations (7) and (8) was used in Ref. 9 for the embedding of $\mathrm{U}_q(\mathrm{so}_3)$ to $\mathrm{U}_q(\mathrm{sl}_3)$. The relations (7) and (8) differ from Serre's relations in the definition of quantum algebras by V. Drinfeld and M. Jimbo by the appearance of nonvanishing right-hand sides.

The algebra $\mathrm{U}_q(\mathrm{so}_3)$ is closely related to (but does not coincide with) the quantum algebra $\mathrm{U}_q(\mathrm{sl}_2)$. The last algebra is generated by the elements q^H, q^{-H}, E, and F satisfying the relations

$$q^Hq^{-H}=q^{-H}q^H=1,\quad q^HEq^{-H}=qE,\quad q^HFq^{-H}=q^{-1}F, \tag{9}$$

$$[E,F]:=EF-FE=\frac{q^{2H}-q^{-2H}}{q-q^{-1}}. \tag{10}$$

Note that $\mathrm{U}_q(\mathrm{sl}_2)$ is the associative algebra equipped with a Hopf algebra structure (a comultiplication, a counit, and an antipode). In particular, the comultiplication Δ is determined by the formulas

$$\Delta(q^{\pm H})=q^{\pm H}\otimes q^{\pm H},\quad \Delta(E)=E\otimes q^H+q^{-H}\otimes E,\quad \Delta(F)=F\otimes q^H+q^{-H}\otimes F.$$

In order to relate the algebras $\mathrm{U}_q(\mathrm{so}_3)$ and $\mathrm{U}_q(\mathrm{sl}_2)$ we need to extend $\mathrm{U}_q(\mathrm{sl}_2)$ by the elements $(q^kq^H+q^{-k}q^{-H})^{-1}$ in the sense of Ref. 10. We denote by $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ the associative algebra with unit element generated by the elements

$$q^H,\quad q^{-H},\quad E,\quad F,\quad (q^k q^H+q^{-k}q^{-H})^{-1},\quad k\in\mathbf{Z},$$

satisfying the defining relations (9) and (10) of the algebra $\mathrm{U}_q(\mathrm{sl}_2)$ and the following natural relations:

$$(q^kq^H+q^{-k}q^{-H})^{-1}(q^kq^H+q^{-k}q^{-H})=(q^kq^H+q^{-k}q^{-H})(q^kq^H+q^{-k}q^{-H})^{-1}=1, \tag{11}$$

$$q^{\pm H}(q^kq^H+q^{-k}q^{-H})^{-1}=(q^kq^H+q^{-k}q^{-H})^{-1}q^{\pm H}, \tag{12}$$

$$(q^kq^H+q^{-k}q^{-H})^{-1}E=E(q^{k+1}q^H+q^{-k-1}q^{-H})^{-1}, \tag{13}$$

$$(q^kq^H+q^{-k}q^{-H})^{-1}F=F(q^{k-1}q^H+q^{-k+1}q^{-H})^{-1}. \tag{14}$$

Note that the algebra $\mathrm{U}_q(\mathrm{sl}_2)$ has finite-dimensional irreducible representations $T_l\equiv T_l^{(1)}$, $T_l^{(-1)}$, $T_l^{(i)}$, $T_l^{(-i)}$, $l=0,\frac{1}{2},1,\frac{3}{2},\ldots,$ acting on the vector spaces $\mathcal{H}_l$ with bases $|m\rangle$, $m=-l,-l+1,...,l$. These representations are given by the formulas

$$T_l^{(1)}(q^H)|m\rangle=q^m|m\rangle,\quad T_l^{(1)}(E)|m\rangle=[l-m]|m+1\rangle, \tag{15}$$

$$T_l^{(1)}(F)|m\rangle=[l+m]|m-1\rangle, \tag{16}$$

where a number in square brackets means a q-number, defined by the formula

$$[a]=\frac{q^a-q^{-a}}{q-q^{-1}},$$

and by the formulas

$$T_l^{(-1)}(q^H)|m\rangle=-q^m|m\rangle,\quad T_l^{(-1)}(E)=T_l^{(1)}(E),\quad T_l^{(-1)}(F)=T_l^{(1)}(F), \tag{17}$$

$$T_l^{(i)}(q^H)|m\rangle=iq^m|m\rangle,\quad T_l^{(i)}(E)=T_l^{(1)}(E),\quad T_l^{(i)}(F)=-T_l^{(1)}(F), \tag{18}$$

$$T_l^{(-i)}(q^H)|m\rangle=-iq^m|m\rangle,\quad T_l^{(-i)}(E)=T_l^{(1)}(E),\quad T_l^{(-i)}(F)=-T_l^{(1)}(F). \tag{19}$$

The representations $T_l^{(1)}$, $T_l^{(-1)}$, $T_l^{(i)}$, $T_l^{(-i)}$, $l=0,\frac{1}{2},1,\frac{3}{2},\ldots,$ are pairwise nonequivalent, and any finite-dimensional irreducible representation of $\mathrm{U}_q(\mathrm{sl}_2)$ is equivalent to one of these representations (see, for example, Ref. 11, Chap. 3).

Now we wish to extend these representations of $\mathrm{U}_q(\mathrm{sl}_2)$ to the representations of $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ by using the relation

$$T((q^kq^H+q^{-k}q^{-H})^{-1}):=(q^kT(q^H)+q^{-k}T(q^{-H}))^{-1}.$$

Clearly, only those irreducible representations T of $\mathrm{U}_q(\mathrm{sl}_2)$ can be extended to $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ for which the operators $q^kT(q^H)+q^{-k}T(q^{-H})$ are invertible. From formulas (15)–(19) it is clear that these operators are always invertible for the representations $T_l^{(1)}$, $T_l^{(-1)}$, $l=0,\frac{1}{2},1,\frac{3}{2},\ldots,$ and for the representations $T_l^{(i)}$, $T_l^{(-i)}$, $l=\frac{1}{2},\frac{3}{2},\frac{5}{2},\ldots$. For the representations $T_l^{(i)}$, $T_l^{(-i)}$, $l=0,1,2,...$, some of these operators are not invertible since they have zero eigenvalue. Denoting the extended representations by the same symbols, we can formulate the following statement:

Proposition 1: The algebra $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ *has the irreducible finite-dimensional representations* $T_l^{(1)}$, $T_l^{(-1)}$, $l=0,\frac{1}{2},1,\frac{3}{2},\ldots,$ *and* $T_l^{(i)}$, $T_l^{(-i)}$, $l=\frac{1}{2},\frac{3}{2},\frac{5}{2},\ldots$. *Any irreducible finite-dimensional representation of* $\mathrm{U}_q(\mathrm{sl}_2)$ *is equivalent to one of these representations.*

III. THE ALGEBRA HOMOMORPHISM $\mathrm{U}_q(\mathrm{so}_3) \to \hat{\mathrm{U}}_q(\mathrm{sl}_2)$

The aim of this section is to give (in an explicit form) the homomorphism of the algebra $\mathrm{U}_q(\mathrm{so}_3)$ to $\hat{U}_q(\mathrm{sl}_2)$. This homomorphism is described by the following proposition:

Proposition 2: There exists a unique algebra homomorphism $\psi:\mathrm{U}_q(\mathrm{so}_3)\to\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ *such that*

$$\psi(I_1)=\frac{i}{q-q^{-1}}(q^H-q^{-H}), \tag{20}$$

$$\psi(I_2)=(E-F)(q^H+q^{-H})^{-1}, \tag{21}$$

$$\psi(I_3)=(iq^{H-1/2}E+iq^{-H-1/2}F)(q^H+q^{-H})^{-1}, \tag{22}$$

where $q^{H+a}:=q^Hq^a$ *for* $a\in\mathbf{C}$.

Proof: In order to prove this proposition we have to show that

$$\begin{aligned} &q^{1/2}\psi(I_1)\psi(I_2)-q^{-1/2}\psi(I_2)\psi(I_1)=\psi(I_3),\\ &q^{1/2}\psi(I_2)\psi(I_3)-q^{-1/2}\psi(I_3)\psi(I_2)=\psi(I_1),\\ &q^{1/2}\psi(I_3)\psi(I_1)-q^{-1/2}\psi(I_1)\psi(I_3)=\psi(I_2).\end{aligned} \tag{23}$$

Let us prove the relation (23). (Other relations are proved similarly.) Substituting the expressions (20)–(22) for $\psi(I_i)$, $i=1,2,3$, into (23) we have [after multiplying both sides of the equality by (q^H+q^{-H}) on the right], the relation

$$\begin{aligned} &q(E-F)Eq^H(qq^H+q^{-1}q^{-H})^{-1}+q(E-F)Fq^{-H}(q^{-1}q^H+qq^{-H})^{-1}\\ &\quad -qE^2q^H(qq^H+q^{-1}q^{-H})^{-1}-q^{-1}FEq^{-H}(qq^H+q^{-1}q^{-H})^{-1}\\ &\quad +q^{-1}EFq^H(q^{-1}q^H+qq^{-H})^{-1}+qF^2q^{-H}(q^{-1}q^H+qq^{-H})^{-1}=i\,\frac{q^{2H}-q^{-2H}}{q-q^{-1}}.\end{aligned}$$

The formula (23) is true if and only if this relation is correct. We multiply both its sides by $(qq^H+q^{-1}q^{-H})(q^{-1}q^H+qq^{-H})$ on the right and obtain the relation in the algebra $\mathrm{U}_q(\mathrm{sl}_2)$ [that is, without the expressions $(q^kq^H+q^{-k}q^{-H})^{-1}$]. This relation is easily verified by using the defining relations (9) and (10) of the algebra $\mathrm{U}_q(\mathrm{sl}_2)$. Proposition is proved.

IV. FINITE-DIMENSIONAL REPRESENTATIONS OF $\mathrm{U}_q(\mathrm{so}_3)$: q IS NOT A ROOT OF UNITY

We assume in Secs. IV–VII that q is not a root of unity.

If T is a representation of the algebra $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ on a linear space $\mathcal{V}$, then the mapping $R:\mathrm{U}_q(\mathrm{so}_3)\to\mathcal{V}$ defined as the composition $R=T\circ\psi$, where ψ is the homomorphism from Proposition 2, is a representation of $\mathrm{U}_q(\mathrm{so}_3)$. Let us consider the representations

$$R_l^{(1)}=T_l^{(1)}\circ\psi,\quad R_l^{(-1)}=T_l^{(-1)}\circ\psi,\quad R_l^{(i)}=T_l^{(i)}\circ\psi,\quad R_l^{(-i)}=T_l^{(-i)}\circ\psi$$

of $\mathrm{U}_q(\mathrm{so}_3)$, where $T_l^{(1)}$, $T_l^{(-1)}$, $T_l^{(i)}$, and $T_l^{(-i)}$ are the irreducible representations of $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ from Proposition 1.

Using formulas for the representations $T_l^{(\pm1)}$ of $\mathrm{U}_q(\mathrm{sl}_2)$ and the expressions (20)–(22) for $\psi(I_j)$, $j=1,2,3$, we find that

$$R_l^{(1)}(I_1)|m\rangle=i[m]|m\rangle,$$

$$R_l^{(1)}(I_2)|m\rangle = \frac{1}{q^m+q^{-m}}\{[l-m]|m+1\rangle - [l+m]|m-1\rangle\},$$

$$R_l^{(1)}(I_3)|m\rangle = \frac{iq^{1/2}}{q^m+q^{-m}}\{q^m[l-m]|m+1\rangle + q^{-m}[l+m]|m-1\rangle\}$$

for the representation $R_l^{(1)}$ and

$$R_l^{(-1)}(I_1)|m\rangle = -i[m]|m\rangle,\quad R_l^{(-1)}(I_2) = -R_l^{(1)}(I_2),\quad R_l^{(-1)}(I_3) = R_l^{(1)}(I_3).$$

Denoting the vectors $|m\rangle$ by $|-m\rangle$ for the representation $R_l^{(-1)}$ we easily find that the matrices of the representation $R_l^{(-1)}$ in the basis $|-m\rangle$, $m=-l,-l+1,...,l$, coincide with the corresponding matrices of the representation $R_l^{(1)}$. Thus, the nonequivalent representations $T_l^{(1)}$ and $T_l^{(-1)}$ of the algebra $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ lead to equivalent representations of $\mathrm{U}_q(\mathrm{so}_3)$.

For the representations $R_l^{(i)}$ and $R_l^{(-i)}$ we have

$$R_l^{(i)}(I_1)|m\rangle = -\frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle,$$

$$R_l^{(i)}(I_2)|m\rangle = i\frac{[l-m]}{q^m-q^{-m}}|m+1\rangle + i\frac{[l+m]}{q^m-q^{-m}}|m-1\rangle,$$

$$R_l^{(i)}(I_3)|m\rangle = -\frac{iq^{m+1/2}[l-m]}{q^m-q^{-m}}|m+1\rangle - \frac{iq^{-m+1/2}[l+m]}{q^m-q^{-m}}|m-1\rangle,$$

and

$$R_l^{(-i)}(I_1)|m\rangle = \frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle,$$

$$R_l^{(-i)}(I_2)|m\rangle = -i\frac{[l-m]}{q^m-q^{-m}}|m+1\rangle - i\frac{[l+m]}{q^m-q^{-m}}|m-1\rangle,$$

$$R_l^{(-i)}(I_3)|m\rangle = -\frac{iq^{m+1/2}[l-m]}{q^m-q^{-m}}|m+1\rangle - \frac{iq^{-m+1/2}[l+m]}{q^m-q^{-m}}|m-1\rangle.$$

Proposition 3: The representations $R_l^{(1)}$ of $\mathrm{U}_q(\mathrm{so}_3)$ *are irreducible. The representations $R_l^{(i)}$ and $R_l^{(-i)}$ are reducible.*

Proof: To prove the first part of the proposition we first note that since q is not a root of unity, the eigenvalues $i[m]$, $m=-l,-l+1,...,l$, of the operator $R_l^{(1)}(I_1)$ are pairwise different.

Let V be an invariant subspace of the space $\mathcal{H}_l$ of the representation $R_l^{(1)}$, and let $\mathbf{v} \equiv \Sigma_{m_i}\alpha_i|m_i\rangle \in V$, where $|m_i\rangle$ are eigenvectors of $R_l^{(1)}(I_1)$. Then $|m_i\rangle \in V$. We prove this for the case when $\mathbf{v}=\alpha_1|m_1\rangle+\alpha_2|m_2\rangle$. (The case with a larger number of summands is proved similarly.) We have $R_l^{(1)}(I_1)\mathbf{v}=i\alpha_1[m_1]|m_1\rangle+i\alpha_2[m_2]|m_2\rangle$. Since

$$\mathbf{v}=\alpha_1|m_1\rangle+\alpha_2|m_2\rangle \in V,\quad \mathbf{v}'\equiv i\alpha_1[m_1]|m_1\rangle+i\alpha_2[m_2]|m_2\rangle \in V,$$

one derives that

$$i[m_1]\mathbf{v}-\mathbf{v}' = i\alpha_2([m_1]-[m_2])|m_2\rangle \in V.$$

Since $[m_1]\neq[m_2]$, then $|m_2\rangle \in V$ and hence $|m_1\rangle \in V$.

In order to prove that $V=\mathcal{H}_l$ we obtain from the above formulas for $R_l^{(1)}(I_2)|m\rangle$ and $R_l^{(1)}(I_3)|m\rangle$ that

$$\{R_l^{(1)}(I_3)-iq^{m+1/2}R_l^{(1)}(I_2)\}|m\rangle=iq^{1/2}|m-1\rangle,$$

$$\{R_l^{(1)}(I_3)+iq^{-m+1/2}R_l^{(1)}(I_2)\}|m\rangle=iq^{1/2}|m+1\rangle.$$

Since V contains at least one basis vector $|m\rangle$, it follows from these relations that V contains the vectors $|m-1\rangle,|m-2\rangle,\dots,|-l\rangle$ and the vectors $|m+1\rangle,|m+2\rangle,\dots,|l\rangle$. This means that $V=\mathcal{H}_l$ and the representation $R_l^{(1)}$ is irreducible.

Let us show that the representations $R_l^{(i)}$ are reducible. The eigenvalues of the operator $R_l^{(i)}(I_1)$ are

$$-\frac{q^m+q^{-m}}{q-q^{-1}},\quad m=-l,-l+1,\dots,l,$$

that is, every spectral point has multiplicity 2. Namely, the pairs of vectors $|m\rangle$ and $|-m\rangle$ are of the same eigenvalue. Let V_1 be the subspace of the representation space $\mathcal{H}_l$ spanned by the vectors

$$|\tfrac{1}{2}\rangle+i|-\tfrac{1}{2}\rangle,\quad |\tfrac{3}{2}\rangle-i|-\tfrac{3}{2}\rangle,\quad |\tfrac{5}{2}\rangle+i|-\tfrac{5}{2}\rangle,\quad |\tfrac{7}{2}\rangle-i|-\tfrac{7}{2}\rangle,\quad \dots, \tag{24}$$

and let V_2 be the subspace spanned by the vectors

$$|\tfrac{1}{2}\rangle-i|-\tfrac{1}{2}\rangle,\quad |\tfrac{3}{2}\rangle+i|-\tfrac{3}{2}\rangle,\quad |\tfrac{5}{2}\rangle-i|-\tfrac{5}{2}\rangle,\quad |\tfrac{7}{2}\rangle+i|-\tfrac{7}{2}\rangle,\quad \dots . \tag{25}$$

We denote the vectors (24) by

$$|\tfrac{1}{2}\rangle',\quad |\tfrac{3}{2}\rangle',\quad |\tfrac{5}{2}\rangle',\quad |\tfrac{7}{2}\rangle',\dots \tag{26}$$

and the vectors (25) by

$$|\tfrac{1}{2}\rangle'',\quad |\tfrac{3}{2}\rangle'',\quad |\tfrac{5}{2}\rangle'',\quad |\tfrac{7}{2}\rangle'',\quad \dots . \tag{27}$$

Then

$$R_l^{(i)}(I_1)|m\rangle'=-\frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle',\quad R_l^{(i)}(I_1)|m\rangle''=-\frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle''.$$

We also have

$$R_l^{(i)}(I_2)|\tfrac{1}{2}\rangle'=i\frac{[l-\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|\tfrac{3}{2}\rangle+i\frac{[l+\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|-\tfrac{1}{2}\rangle+\frac{[l+\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|\tfrac{1}{2}\rangle+\frac{[l-\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|-\tfrac{3}{2}\rangle$$
$$=\frac{[l+\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|\tfrac{1}{2}\rangle'+i\frac{[l-\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|\tfrac{3}{2}\rangle'.$$

We derive similarly that

$$R_l^{(i)}(I_2)|\tfrac{1}{2}\rangle''=-\frac{[l+\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|\tfrac{1}{2}\rangle''+i\frac{[l-\frac{1}{2}]}{q^{1/2}-q^{-1/2}}|\tfrac{3}{2}\rangle''$$

and that

$$R_l^{(i)}(I_2)|m\rangle' = i\,\frac{[l-m]}{q^m-q^{-m}}|m+1\rangle' + i\,\frac{[l+m]}{q^m-q^{-m}}|m-1\rangle', \quad m>\tfrac{1}{2},$$

$$R_l^{(i)}(I_2)|m\rangle'' = i\,\frac{[l-m]}{q^m-q^{-m}}|m+1\rangle'' + i\,\frac{[l+m]}{q^m-q^{-m}}|m-1\rangle'', \quad m>\tfrac{1}{2}.$$

Thus, the subspaces V_1 and V_2 are invariant with respect to the operators $R_l^{(i)}(I_1)$ and $R_l^{(i)}(I_2)$. This means that they are invariant with respect to the representation $R_l^{(i)}$.

It is proved similarly that the subspace V_1 of the space $\mathcal{H}_l$ of the representation $R_l^{(-i)}$ spanned by the vectors (24) and the subspace V_2 of $\mathcal{H}_l$ spanned by the vectors (25) are invariant with respect to the operators $R_l^{(-i)}(I_1)$ and $R_l^{(-i)}(I_2)$. That is, the representation $R_l^{(-i)}$ is also reducible. Proposition is proved.

Let $R_n^{(i,+)}$ and $R_n^{(i,-)}$, $n=l+\frac{1}{2}=\dim V_1=\dim V_2$, be the representations of $\mathrm{U}_q(\mathrm{so}_3)$ which are restrictions of $R_l^{(i)}$ to the subspaces V_1 and V_2, respectively. Denoting the vectors (26) of the subspace V_1 by

$$|1\rangle,|2\rangle,|3\rangle,|4\rangle,\ldots,|n\rangle \equiv |l+\tfrac{1}{2}\rangle, \tag{28}$$

respectively, we have

$$R_n^{(i,+)}(I_1)|k\rangle = -\frac{q^{k-1/2}+q^{-k+1/2}}{q-q^{-1}}|k\rangle,$$

$$R_n^{(i,+)}(I_2)|1\rangle = \frac{[n]}{q^{1/2}-q^{-1/2}}|1\rangle + i\,\frac{[n-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_n^{(i,+)}(I_2)|k\rangle = i\,\frac{[n-k]}{q^{k-1/2}-q^{-k+1/2}}|k+1\rangle + i\,\frac{[n+k-1]}{q^{k-1/2}-q^{-k+1/2}}|k-1\rangle, \quad k\neq 1.$$

For the operator $R_n^{(i,+)}(I_3)$ we have

$$R_n^{(i,+)}(I_3)|1\rangle = -\frac{[n]}{q^{1/2}-q^{-1/2}}|1\rangle - i\,\frac{q[n-1]}{q^{1/2}-q^{1/2}}|2\rangle,$$

$$R_n^{(i,+)}(I_3)|k\rangle = -i\,\frac{q^k[n-k]}{q^{k-1/2}-q^{-k+1/2}}|k+1\rangle - i\,\frac{q^{-k+1}[n+k-1]}{q^{k-1/2}-q^{-k+1/2}}|k-1\rangle, \quad k\neq 1.$$

Denoting the vectors (27) of the subspace V_2 by the symbols (28), respectively, we obtain

$$R_n^{(i,-)}(I_1)|k\rangle = -\frac{q^{k-1/2}+q^{-k+1/2}}{q-q^{-1}}|k\rangle,$$

$$R_n^{(i,-)}(I_2)|1\rangle = -\frac{[n]}{q^{1/2}-q^{-1/2}}|1\rangle + i\,\frac{[n-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_n^{(i,-)}(I_2)|k\rangle = R_n^{(i,+)}(I_2)|k\rangle, \quad k\neq 1.$$

For the operator $R_l^{(i,-)}(I_3)$ we find that

$$R_n^{(i,-)}(I_3)|1\rangle = \frac{[n]}{q^{1/2}-q^{-1/2}}|1\rangle - i\,\frac{q[n-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_n^{(i,-)}(I_3)|k\rangle = R_n^{(i,+)}(I_3)|k\rangle, \quad k\neq 1.$$

Let now $R_n^{(-i,+)}$ and $R_n^{(-i,-)}$, $n=l+\frac{1}{2}$, be the representations of $U_q(so_3)$ which are restrictions of the representation $R_l^{(-i)}$ to the subspaces V_1 and V_2, respectively. Introducing the vectors similar to the vectors (28), for the representation $R_n^{(-i,+)}$ we have

$$R_n^{(-i,+)}(I_1)|k\rangle = \frac{q^{k-1/2}+q^{-k+1/2}}{q-q^{-1}}|k\rangle, \quad R_n^{(-i,+)}(I_2) = -R_n^{(i,+)}(I_2),$$

$$R_n^{(-i,+)}(I_3)|1\rangle = \frac{[n]}{q^{1/2}-q^{-1/2}}|1\rangle + i\frac{q[n-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_n^{(-i,+)}(I_3)|k\rangle = i\frac{q^k[n-k]}{q^{k-1/2}-q^{-k+1/2}}|k+1\rangle + i\frac{q^{-k+1}[n+k-1]}{q^{k-1/2}-q^{-k+1/2}}|k-1\rangle, \quad k\neq 1.$$

For the representation $R_l^{(-i,-)}(I_3)$ we obtain

$$R_n^{(-i,-)}(I_1)|k\rangle = \frac{q^{k-1/2}+q^{-k+1/2}}{q-q^{-1}}|k\rangle, \quad R_n^{(-i,-)}(I_2) = -R_n^{(i,-)}(I_2),$$

$$R_n^{(-i,-)}(I_3)|1\rangle = -\frac{[n]}{q^{1/2}-q^{-1/2}}|1\rangle + i\frac{q[n-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_n^{(-i,-)}(I_3)|k\rangle = R_n^{(-i,+)}(I_3)|k\rangle.$$

Thus, we constructed the representations $R_n^{(i,+)}$, $R_n^{(i,-)}$, $R_n^{(-i,+)}$, and $R_n^{(-i,-)}$ of the algebra $U_q(so_3)$. The following theorem characterizes them.

Theorem 1: *The representations $R_n^{(i,+)}$, $R_n^{(i,-)}$, $R_n^{(-i,+)}$, and $R_n^{(-i,-)}$ are irreducible and pairwise nonequivalent. For any l the representation $R_l^{(1)}$ is not equivalent to any of these representations.*

Proof: The irreducibility is proved exactly in the same way as in Proposition 3. Equivalence relations may exist only for irreducible representations of the same dimension. That is, we have to show that under fixed n no pair of the representations $R_n^{(i,+)}$, $R_n^{(i,-)}$, $R_n^{(-i,+)}$, and $R_n^{(-i,-)}$ is equivalent. It follows from the above formulas that the operators $R_n^{(i,+)}(I_1)$ and $R_n^{(i,-)}(I_1)$, as well as the operators $R_n^{(-i,+)}(I_1)$ and $R_n^{(-i,-)}(I_1)$, have the same set of eigenvalues. Moreover, the spectrum of the first pair of operators differs from that of the second pair. Hence, neither of the representations $R_n^{(i,+)}$ and $R_n^{(i,-)}$ is equivalent to $R_n^{(-i,+)}$ or $R_n^{(-i,-)}$. The representations $R_n^{(i,+)}$ and $R_n^{(i,-)}$ are not equivalent since the operators $R_n^{(i,+)}(I_2)$ and $R_n^{(i,-)}(I_2)$ have different traces (for equivalent representations these operators must have the same trace). For the same reason, the representations $R_n^{(-i,+)}$ and $R_n^{(-i,-)}$ are not equivalent. The last assertion of the theorem follows from the fact that the spectrum of the operator $R_l^{(1)}(I_3)$ differs from the spectra of the operators $R_n^{(i,+)}(I_1)$, $R_n^{(i,-)}(I_1)$, $R_n^{(-i,+)}(I_1)$, and $R_n^{(-i,-)}(I_1)$. Theorem is proved.

Clearly, the reducible representations $R_n^{(i)}$ and $R_n^{(-i)}$ decompose into irreducible components as

$$R_n^{(i)} = R_n^{(i,+)} \oplus R_n^{(i,-)}, \quad R_n^{(-i)} = R_n^{(-i,+)} \oplus R_n^{(-i,-)}. \tag{29}$$

It can be proved that *every irreducible finite-dimensional representation of* $U_q(so_3)$ *is equivalent to one of the representations* $R_l^{(1)}$, $R_n^{(i,+)}$, $R_n^{(i,-)}$, $R_n^{(-i,+)}$, $R_n^{(-i,-)}$. That is, these representations exhaust, up to equivalence, all irreducible finite-dimensional representations of $U_q(so_3)$. A proof of this statement will be given in a separate paper.

V. TENSOR PRODUCTS OF REPRESENTATIONS OF $U_q(so_3)$

As mentioned above, no Hopf algebra structure is known for the algebra $U_q(so_3)$. Therefore, we cannot construct a tensor product of finite-dimensional representations of $U_q(so_3)$ by using a comultiplication as we do in the case of the quantum algebra $U_q(sl_2)$. However, we may construct some tensor product representations by using the algebra homomorphism of Proposition 2.

First we determine which tensor products of irreducible representations of $U_q(sl_2)$ can be extended to representations of the algebra $\hat{U}_q(sl_2)$. Verifying for which tensor products $T=T'\otimes T''$ of irreducible representations of $U_q(sl_2)$ the operators

$$q^k T(q^H)+q^{-k}T(q^{-H}), \quad k\in \mathbf{Z},$$

are invertible, we conclude that only the tensor products

$$T_l^{(\pm 1)}\otimes T_{l'}^{(\pm 1)}, \quad T_l^{(\pm 1)}\otimes T_{l'}^{(\mp 1)}, \quad l,l'=0,\tfrac{1}{2},1,\tfrac{3}{2},\dots,$$

$$T_l^{(\pm 1)}\otimes T_{l'}^{(\pm i)}, \quad T_l^{(\pm 1)}\otimes T_{l'}^{(\mp i)}, \quad l=0,1,2,..., \quad l'=\tfrac{1}{2},\tfrac{3}{2},\tfrac{5}{2},\dots,$$

$$T_l^{(\pm i)}\otimes T_{l'}^{(\pm 1)}, \quad T_l^{(\pm i)}\otimes T_{l'}^{(\mp 1)}, \quad l=\tfrac{1}{2},\tfrac{3}{2},\tfrac{5}{2},\dots, \quad l'=0,1,2,...,$$

$$T_l^{(\pm i)}\otimes T_{l'}^{(\pm i)}, \quad T_l^{(\pm i)}\otimes T_{l'}^{(\mp i)}, \quad l,l'=\tfrac{1}{2},\tfrac{3}{2},\tfrac{5}{2},\dots,$$

can be extended to the algebra $\hat{U}_q(sl_2)$. Taking into account the decompositions of tensor products of irreducible representations of $U_q(sl_2)$ (see, for example, the end of Subsection 3.2.1 and Proposition 3.22 in Ref. 11) we find that

$$T_l^{(\omega)}\otimes T_{l'}^{(\omega')}\simeq T_{l+l'}^{(\omega\omega')}\oplus T_{l+l'-1}^{(\omega\omega')}\oplus\cdots\oplus T_{|l+l'|}^{(\omega\omega')}, \tag{30}$$

$$T_l^{(\omega)}\otimes t_{l'}^{(\pm i)}\simeq T_{l+l'}^{(\pm\omega i)}\oplus T_{l+l'-1}^{(\pm\omega i)}\oplus\cdots\oplus T_{|l+l'|}^{(\pm\omega i)}, \tag{31}$$

$$T_l^{(\pm i)}\otimes T_{l'}^{(\omega)}\simeq T_{l+l'}^{(\pm\omega i)}\oplus T_{l+l'-1}^{(\pm\omega i)}\oplus\cdots\oplus T_{|l+l'|}^{(\pm\omega i)}, \tag{32}$$

$$T_l^{(\omega i)}\otimes T_{l'}^{(\omega' i)}\simeq T_{l+l'}^{(-\omega\omega')}\oplus T_{l+l'-1}^{(-\omega\omega')}\oplus\cdots\oplus T_{|l+l'|}^{(-\omega\omega')}, \tag{33}$$

where $\omega,\omega'=\pm 1$.

Now we define tensor products of representations of $U_q(so_3)$ corresponding to the above tensor product representations of $\hat{U}_q(sl_2)$ as

$$R\otimes R'=(T\otimes T')\circ\psi,$$

where $R=T\circ\psi$ and $R'=T'\circ\psi$. Taking into account the definitions of tensor products of representations of $U_q(sl_2)$ by means of the comultiplication and the definition of the mapping ψ we have

$$\begin{aligned}(R\otimes R')(I_1)&=(T\otimes T')\circ\psi(I_1)\\ &=\frac{i}{q-q^{-1}}(T(q^H)\otimes T'(q^H)-T(q^{-H})\otimes T'(q^{-H})).\end{aligned}$$

Similarly,

$$\begin{aligned}(R\otimes R')(I_2)=&(T(E)\otimes T'(q^H)+T(q^{-H})\otimes T'(E)-T(F)\otimes T'(q^H)\\ &-(T(q^{-H})\otimes T'(F))(T(q^H)\otimes T'(q^H)+T(q^{-H})\otimes T'(q^{-H}))^{-1}.\end{aligned}$$

Composing both sides of the relations (30)–(33) with the mapping ψ of Proposition 2, we find the decomposition into representations of $U_q(so_3)$ for the tensor products

$$R_l^{(1)}\otimes R_{l'}^{(1)},\quad R_l^{(1)}\otimes R_{l'}^{(\pm i)},\quad R_{l'}^{(\pm i)}\otimes R_l^{(1)},\quad R_l^{(\pm i)}\otimes R_{l'}^{(\pm i)},\quad R_l^{(\pm i)}\otimes R_{l'}^{(\mp i)},$$

where the second and the third tensor products are defined only for $l=0,1,2,\ldots$. (Note that the representations $R_l^{(\pm i)}$ are defined only for $l=\frac{1}{2},\frac{3}{2},\frac{5}{2},\ldots$.) We have

$$R_l^{(1)}\otimes R_{l'}^{(1)}\simeq R_{l+l'}^{(1)}\oplus R_{l+l'-1}^{(1)}\oplus\cdots\oplus R_{|l-l'|}^{(1)},$$

$$R_l^{(1)}\otimes R_{l'}^{(\pm i)}\simeq R_{l+l'}^{(\pm i)}\oplus R_{l+l'-1}^{(\pm i)}\oplus\cdots\oplus R_{|l-l'|}^{(\pm i)},$$

$$R_l^{(\pm i)}\otimes R_{l'}^{(1)}\simeq R_{l+l'}^{(\pm i)}\oplus R_{l+l'-1}^{(\pm i)}\oplus\cdots\oplus R_{|l-l'|}^{(\pm i)},$$

$$R_l^{(\omega i)}\otimes R_{l'}^{(\omega' i)}\simeq R_{l+l'}^{(1)}\oplus R_{l+l'-1}^{(1)}\oplus\cdots\oplus R_{|l-l'|}^{(1)}.$$

In these formulas the representations $R_l^{(\pm i)}$ are reducible. Unfortunately, our definition of tensor products of representations of $U_q(so_3)$ does not allow us to determine the tensor products containing the irreducible representations $R_n^{(\pm i,\pm)}$ and $R_n^{(\pm i,\mp)}$.

VI. INFINITE-DIMENSIONAL REPRESENTATIONS OF $U_q(so_3)$ OBTAINED FROM REPRESENTATIONS OF $U_q(sl_2)$

By using the homomorphism $\psi:U_q(so_3)\to\hat{U}_q(sl_2)$ from Proposition 2 and infinite-dimensional irreducible representations of the algebra $\hat{U}_q(sl_2)$ we can construct infinite-dimensional irreducible representations of the algebra $U_q(so_3)$.

Let us first describe irreducible infinite-dimensional representations of the algebra $U_q(sl_2)$. Note that by an infinite-dimensional representation T of $U_q(sl_2)$ we mean a homomorphism of $U_q(sl_2)$ into the algebra of linear operators (bounded or unbounded) on a Hilbert space, defined on an everywhere dense invariant subspace D, such that the operator $T(q^H)$ can be diagonalized, has a discrete spectrum, and its eigenvectors belong to D. Infinite-dimensional representations T of $U_q(so_3)$ are described in the same way replacing the operator $T(q^H)$ by $T(I_1)$.

Two representations T and T' of $U_q(sl_2)$ on spaces $\mathcal{H}$ and $\mathcal{H}'$, respectively, are called (algebraically) equivalent if there exist everywhere dence invariant subspaces $V\subset\mathcal{H}$ and $V'\subset\mathcal{H}'$ and a one-to-one linear operator $A:V\to V'$ such that $AT(a)\mathbf{v}=T'(a)A\mathbf{v}$ for all $a\in U_q(sl_2)$ and $\mathbf{v}\in V$. Equivalence of infinite-dimensional representations of $U_q(so_3)$ is defined in the same way.

Let ϵ be a fixed complex number such that $0\leqslant \mathrm{Re}\,\epsilon<1$, and let $\mathcal{H}_\epsilon$ be a complex Hilbert space with the orthonormal basis

$$|m\rangle,\quad m=n+\epsilon,\quad n=0,\pm1,\pm2,\ldots . \tag{34}$$

For every complex number a we construct the representation $T_{a\epsilon}$ on the Hilbert space $\mathcal{H}_\epsilon$ defined by

$$T_{a\epsilon}(q^H)|m\rangle=q^m|m\rangle,\quad T_{a\epsilon}(E)|m\rangle=[a-m]|m+1\rangle,\quad T_{a\epsilon}(F)|m\rangle=[a+m]|m-1\rangle,$$

where $[a\pm m]$ is the q-number (see, for example, Ref. 12). The equivalence relations in the set of the representations $T_{a\epsilon}$ can be extracted from Ref. 12.

Note that the representation $T_{a\epsilon}$ is irreducible if and only if $a\neq\pm\epsilon(\mathrm{mod}\,\mathbf{Z})$.

All the representations $T_{a\epsilon}$ can be extended to representations of the algebra $\hat{U}_q(sl_2)$ except for the case when $\epsilon=\pm i\pi/2\tau$, where $q=e^\tau$. (We suppose below that $\epsilon\neq\pm i\pi/2\tau$.) We denote these extended representations by the same symbols $T_{a\epsilon}$.

The formula $R_{a\epsilon}=T_{a\epsilon}\circ\psi$ associates with every irreducible representation $T_{a\epsilon}$, $\epsilon\neq\pm i\pi/2\tau$, of $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ a representation of the algebra $\mathrm{U}_q(\mathrm{so}_3)$.

Let $\epsilon\neq\pm i\pi/2\tau$ and $\epsilon\neq\pm i\pi/2\tau+\frac{1}{2}$. Then for the representations $R_{a\epsilon}$ of $\mathrm{U}_q(\mathrm{so}_3)$ we have

$$R_{a\epsilon}(I_1)|m\rangle=i[m]|m\rangle, \tag{35a}$$

$$R_{a\epsilon}(I_2)|m\rangle=\frac{1}{q^m+q^{-m}}\{[a-m]|m+1\rangle-[a+m]|m-1\rangle\}, \tag{35b}$$

$$R_{a\epsilon}(I_3)|m\rangle=\frac{iq^{1/2}}{q^m+q^{-m}}\{q^m[a-m]|m+1\rangle+q^{-m}[a+m]|m-1\rangle\}. \tag{35c}$$

If $\epsilon=i\pi/2\tau+\frac{1}{2}$, then denoting the basis elements $|m\rangle$, $m=n+\epsilon$, $n\in\mathbf{Z}$, by $|n+\frac{1}{2}\rangle$, $n\in\mathbf{Z}$, respectively, we obtain

$$R_{a\epsilon}(I_1)|k\rangle=-\frac{q^k+q^{-k}}{q-q^{-1}}|k\rangle,$$

$$R_{a\epsilon}(I_2)|k\rangle=i\,\frac{[a'-k]}{q^k-q^{-k}}|k+1\rangle+i\,\frac{[a'+k]}{q^k-q^{-k}}|k-1\rangle,$$

$$R_{a\epsilon}(I_3)|k\rangle=-\frac{iq^{k+1/2}[a'-k]}{q^k-q^{-k}}|k+1\rangle-\frac{iq^{-k+1/2}[a'+k]}{q^k-q^{-k}}|k-1\rangle,$$

where $a'=a+i\pi/2\tau$ and $k=n+\frac{1}{2}$. If $\epsilon=-i\pi/2\tau+\frac{1}{2}$, then using the same notations for basis elements we obtain

$$R'_{a\epsilon}(I_1)|k\rangle=\frac{q^k+q^{-k}}{q-q^{-1}}|k\rangle,$$

$$R'_{a\epsilon}(I_2)|k\rangle=-i\,\frac{[a'-k]}{q^k-q^{-k}}|k+1\rangle-i\,\frac{[a'+k]}{q^k-q^{-k}}|k-1\rangle,$$

$$R'_{a\epsilon}(I_3)|k\rangle=-\frac{iq^{k+1/2}[a'-k]}{q^k-q^{-k}}|k+1\rangle-\frac{iq^{-k+1/2}[a'+k]}{q^k-q^{-k}}|k-1\rangle$$

(to distinguish these representations from the previous ones we supplied $R_{a\epsilon}$ with a prime).

Proposition 4: The representations $R_{a\epsilon}$ *of* $\mathrm{U}_q(\mathrm{so}_3)$ *are irreducible for irreducible representations* $T_{a\epsilon}$, $\epsilon\neq\pm i\pi/2\tau+\frac{1}{2}$ of $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$. *The representations* $R_{a\epsilon}$, $\epsilon=i\pi/2\tau+\frac{1}{2}$, *and* $R'_{a\epsilon}$, $\epsilon=-i\pi/2\tau+\frac{1}{2}$, *are reducible.*

Proof is given in the same way as in the case of Proposition 3.

As in the case of finite-dimensional representations in Sec. IV, decomposing the representations $R_{a\epsilon}$, $\epsilon=i\pi/2\tau+\frac{1}{2}$, and $R'_{a\epsilon}$, $\epsilon=-i\pi/2\tau+\frac{1}{2}$, we obtain irreducible infinite-dimensional representations of $\mathrm{U}_q(\mathrm{so}_3)$ which will be denoted by $R_{a'}^{(i,\pm)}$ and $R_{a'}^{(-i,\pm)}$, $a'=a+i\pi/2\tau$. In the basis

$$|n\rangle,\quad n=1,2,3,...,$$

they are given by the formulas

$$R_{a'}^{(i,\pm)}(I_1)|k\rangle=-\frac{q^{k-1/2}+q^{-k+1/2}}{q-q^{-1}}|k\rangle,$$

$$R_{a'}^{(i,\pm)}(I_2)|1\rangle = \pm\frac{[a']}{q^{1/2}-q^{-1/2}}|1\rangle + i\,\frac{[a'-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_{a'}^{(i,\pm)}(I_2)|k\rangle = i\,\frac{[a'-k]}{q^{k-1/2}-q^{-k+1/2}}|k+1\rangle + i\,\frac{[a'+k-1]}{q^{k-1/2}-q^{-k+1/2}}|k-1\rangle, \quad k\neq 1.$$

$$R_{a'}^{(i,\pm)}(I_3)|1\rangle = \mp\frac{[a']}{q^{1/2}-q^{-1/2}}|1\rangle - i\,\frac{q[a'-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_{a'}^{(i,\pm)}(I_3)|k\rangle = -i\,\frac{q^k[a'-k]}{q^{k-1/2}-q^{-k+1/2}}|k+1\rangle - i\,\frac{q^{-k+1}[a'+k-1]}{q^{k-1/2}-q^{-k+1/2}}|k-1\rangle, \quad k\neq 1$$

and by the formulas

$$R_{a'}^{(-i,\pm)}(I_1)|k\rangle = \frac{q^{k-1/2}+q^{-k+1/2}}{q-q^{-1}}|k\rangle, \quad R_{a'}^{(-i,\pm)}(I_2) = -R_{a'}^{(i,\pm)}(I_2),$$

$$R_{a'}^{(-i,\pm)}(I_3)|1\rangle = \pm\frac{[a']}{q^{1/2}-q^{-1/2}}|1\rangle + i\,\frac{q[a'-1]}{q^{1/2}-q^{-1/2}}|2\rangle,$$

$$R_{a'}^{(-i,\pm)}(I_3)|k\rangle = i\,\frac{q^k[a'-k]}{q^{k-1/2}-q^{-k+1/2}}|k+1\rangle + i\,\frac{q^{-k+1}[a'+k-1]}{q^{k-1/2}-q^{-k+1/2}}|k-1\rangle, \quad k\neq 1.$$

Theorem 2: *The representations* $R_{a'}^{(i,\pm)}$ and $R_{a'}^{(-i,\pm)}$ *are irreducible and pairwise nonequivalent. For any a the irreducible representation* $R_{a\epsilon}$ *is not equivalent to some of these representations*.

Proof is given in the same way as in the finite-dimensional case (see the proof of Theorem 1).

The algebra $U_q(sl_2)$ also has irreducible infinite-dimensional representations with highest weights or with lowest weights. They are classified in Ref. 12. All of these representations T can be extended to the algebra $\hat{U}_q(sl_2)$. Using the composition $R=T\circ\psi$ we obtain the corresponding representations R of $U_q(so_3)$. As above, it can be easily proved that to nonequivalent representations T of $\hat{U}_q(sl_2)$ with highest or lowest weight there correspond nonequivalent irreducible representations of $U_q(so_3)$. We give a list of these representations.

Let $l=\frac{1}{2},1,\frac{3}{2},2,\ldots$. We denote by R_l^+ the representation of $U_q(so_3)$ acting on the Hilbert space $\mathcal{H}_l$ with the orthonormal basis $|m\rangle$, $m=l,l+1,l+2,\ldots$, and given by formulas (35) with $a=-l$. By R_l^- we denote the representation of $U_q(so_3)$ acting on the Hilbert space $\hat{\mathcal{H}}_l$ with the orthonormal basis $|m\rangle$, $m=-l,-l-1,-l-2,\ldots$, and given by formulas (35) with $a=l$.

Now let $a\neq 0(\mathrm{mod}\,\mathbf{Z})$ and $a\neq\frac{1}{2}(\mathrm{mod}\,\mathbf{Z})$. We denote by $\mathcal{H}_a$ the Hilbert space with the orthonormal basis $|m\rangle$, $m=-a,-a+1,-a+2,\ldots$. On this space the representation R_a^+ acts as given by formulas (35). On the Hilbert space $\hat{\mathcal{H}}_a$ with the orthonormal basis $|m\rangle$, $m=a,a-1,a-2,\ldots$, the representation R_a^- acts as given by formulas (35).

Proposition 5: The above representations $R_l^{\pm}$ *and* $R_a^{\pm}$ *are irreducible and pairwise nonequivalent.*

Proof of this proposition is contained in Ref. 13.

VII. OTHER INFINITE DIMENSIONAL REPRESENTATIONS OF $U_q(so_3)$

The algebra $U_q(so_3)$ has also irreducible infinite-dimensional representations which cannot be obtained from representations of $\hat{U}_q(sl_2)$. We describe these representations in this section.

Let $\mathcal{H}$ be the infinite-dimensional vector space with the basis $|m\rangle$, $m=0,\pm 1,\pm 2,\ldots$, and let $\lambda=q^{\tau}$ be a nonzero complex number such that $0\leqslant \mathrm{Re}\,\tau<1$. Then a direct calculation shows that the operators $Q_\lambda^+(I_1)$ and $Q_\lambda^+(I_2)$ given by the formulas

$$Q_\lambda^+(I_1)|m\rangle = \frac{\lambda q^m + \lambda^{-1} q^{-m}}{q-q^{-1}}|m\rangle,$$

$$Q_\lambda^+(I_2)|m\rangle = \frac{1}{q-q^{-1}}|m+1\rangle + \frac{1}{q-q^{-1}}|m-1\rangle$$

satisfy the relations (7) and (8) and hence determine a representation of $U_q(so_3)$ which will be denoted by Q_λ^+. Similarly, the operators $Q_\lambda^-(I_1)$ and $Q_\lambda^-(I_2)$ given on the space $\mathcal{H}$ by

$$Q_\lambda^-(I_1)|m\rangle = -\frac{\lambda q^m + \lambda^1 q^{-m}}{q-q^1}|m\rangle, \quad Q_\lambda^-(I_2) := Q_\lambda^+(I_2)$$

determine a representation of $U_q(so_3)$ which is denoted by Q_λ^-. The operators $Q_\lambda^\pm(I_3)$ can be calculated by means of formula (4).

Proposition 6: If $\lambda \neq 1$ and $\lambda \neq q^{1/2}$, then the representations Q_λ^+ and Q_λ^- are irreducible. The representations $Q_1^\pm$ and $Q_{\sqrt{q}}^\pm$ are reducible.

Proof: The first part is proved in the same way as that of Proposition 3. Let us prove the second part. The representations $Q_1^\pm$ and $Q_{\sqrt{q}}^\pm$ are the only representations in the set $\{Q_\lambda^\pm\}$ for which the operator $Q_\lambda^\pm(I_1)$ has not a simple spectrum. The operators $Q_1^\pm(I_1)$ has the spectrum

$$\ldots, \quad q^{-2}+q^2, \quad q^{-1}+q, \quad 2, \quad q+q^{-1}, \quad q^2+q^{-2}, \quad \ldots .$$

Thus, only the spectral point 2 has multiplicity 1. All other points have multiplicity 2. Let V_1 and V_2 be the vector subspaces of $\mathcal{H}$ with the bases

$$|0\rangle, \quad |m\rangle' = |m\rangle - |-m\rangle, \quad m=1,2,\ldots,$$

and

$$|m\rangle'' = |m\rangle + |-m\rangle, \quad m=1,2,\ldots,$$

respectively. These basis vectors are eigenvectors of the operator $Q_1^\pm(I_1)$:

$$Q_1^\pm(I_1)|m\rangle' = \pm\frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle', \quad Q_1^\pm(I_1)|m\rangle'' = \pm\frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle'',$$

and

$$Q_1^\pm(I_2)|0\rangle = \frac{1}{q-q^{-1}}|1\rangle', \quad Q_1^\pm(I_2)|1\rangle'' = \frac{1}{q-q^{-1}}|2\rangle'',$$

$$Q_1^\pm(I_2)|m\rangle' = \frac{1}{q-q^{-1}}|m+1\rangle' + \frac{1}{q-q^{-1}}|m-1\rangle', \quad m>0,$$

$$Q_1^\pm(I_2)|m\rangle'' = \frac{1}{q-q^{-1}}|m+1\rangle'' + \frac{1}{q-q^{-1}}|m-1\rangle'', \quad m>1.$$

Thus, the subspaces V_1 and V_2 are invariant with respect to the representation Q_1^+ (and the representation Q_1^-). We denote the subrepresentations of $Q_1^\pm$ realized on V_1 and V_2 by $Q_1^{1,\pm}$ and $Q_1^{2,\pm}$, respectively.

The eigenvalues of the operators $Q_{\sqrt{q}}^\pm(I_1)$ are

$$\ldots, \quad q^{-3/2}+q^{3/2}, \quad q^{-1/2}+q^{1/2}, \quad q^{1/2}+q^{-1/2}, \quad q^{3/2}+q^{-3/2}, \quad \ldots .$$

Thus, every spectral point has multiplicity 2. We denote by W_1 and W_2 the vector subspaces of $\mathcal{H}$ spanned by the basis vectors

$$|\tfrac{1}{2}\rangle' = |0\rangle - |-1\rangle, |\tfrac{3}{2}\rangle' = |1\rangle - |-2\rangle, \ldots, |m+\tfrac{1}{2}\rangle' = |m\rangle - |-m-1\rangle, \ldots$$

and

$$|\tfrac{1}{2}\rangle'' = |0\rangle + |-1\rangle, |\tfrac{3}{2}\rangle'' = |1\rangle + |-2\rangle, \ldots, |m+\tfrac{1}{2}\rangle'' = |m\rangle + |-m-1\rangle, \ldots,$$

respectively. These basis vectors are eigenvectors of the operator $Q^{\pm}_{\sqrt{q}}(I_1)$:

$$Q^{\pm}_{\sqrt{q}}(I_1)|m+\tfrac{1}{2}\rangle' = \pm \frac{q^{m+1/2}+q^{-m-1/2}}{q-q^{-1}}|m+\tfrac{1}{2}\rangle',$$

$$Q^{\pm}_{\sqrt{q}}(I_1)|m+\tfrac{1}{2}\rangle'' = \pm \frac{q^{m+1/2}+q^{-m-1/2}}{q-q^{1}}|m+\tfrac{1}{2}\rangle'',$$

and

$$Q^{\pm}_{\sqrt{q}}(I_2)|\tfrac{1}{2}\rangle' = -\frac{1}{q-q^{-1}}|\tfrac{1}{2}\rangle' + \frac{1}{q-q^{-1}}|\tfrac{3}{2}\rangle',$$

$$Q^{\pm}_{\sqrt{q}}(I_2)|m+\tfrac{1}{2}\rangle' = \frac{1}{q-q^{-1}}|m+\tfrac{3}{2}\rangle' + \frac{1}{q-q^{-1}}|m-\tfrac{1}{2}\rangle', \quad m>0,$$

$$Q^{\pm}_{\sqrt{q}}(I_2)|\tfrac{1}{2}\rangle'' = \frac{1}{q-q^{-1}}|\tfrac{1}{2}\rangle'' + \frac{1}{q-q^{-1}}|\tfrac{3}{2}\rangle'',$$

$$Q^{\pm}_{\sqrt{q}}(I_2)|m+\tfrac{1}{2}\rangle'' = \frac{1}{q-q^{-1}}|m+\tfrac{3}{2}\rangle'' + \frac{1}{q-q^{-1}}|m-\tfrac{1}{2}\rangle'', \quad m>0.$$

Thus, the subspaces W_1 and W_2 are invariant with respect to the representations $Q^{\pm}_{\sqrt{q}}$. We denote the subrepresentations of $Q^{\pm}_{\sqrt{q}}$ realized on W_1 and W_2 by $Q^{1,\pm}_{\sqrt{q}}$ and $Q^{2,\pm}_{\sqrt{q}}$, respectively. Proposition is proved.

Theorem 3: *The representations $Q^{1,\pm}_{1}$, $Q^{2,\pm}_{1}$, $Q^{1,\pm}_{\sqrt{q}}$, and $Q^{2,\pm}_{\sqrt{q}}$ are irreducible and pairwise nonequivalent. For any admissible value of λ the representation Q^{+}_{λ} (as well as the representation Q^{-}_{λ}) is not equivalent to some of these representations.*

Proof: Proof is similar to that of Theorem 1 if to take into account spectra of the operators $Q^{1,\pm}_{1}(I_1)$, $Q^{2,\pm}_{1}(I_1)$, $Q^{1,\pm}_{\sqrt{q}}(I_1)$, $Q^{2,\pm}_{\sqrt{q}}(I_1)$, and $Q^{\pm}_{\lambda}(I_1)$ and traces of the operators $Q^{1,\pm}_{1}(I_2)$, $Q^{2,\pm}_{1}(I_2)$, $Q^{1,\pm}_{\sqrt{q}}(I_2)$, and $Q^{2,\pm}_{\sqrt{q}}(I_2)$.

It will be proved in a separate paper that every irreducible infinite-dimensional representation of $U_q(so_3)$ is equivalent to one of the representations described in this and previous sections.

VIII. FINITE-DIMENSIONAL REPRESENTATIONS OF $\hat{U}_q(sl_2)$: q IS A ROOT OF UNITY

Everywhere below q is a root of unity, that is, there is a smallest positive integer p such that $q^p=1$. We suppose that $p\neq 1,2$. We introduce the number p' setting $p'=p$ if p is odd and $p'=p/2$ if p is even.

As in the case of the algebra $U_q(sl_2)$ (see Ref. 11, Chap. 3), if q is a root of unity, then $U_q(so_3)$ is a finite-dimensional vector space over the center of $U_q(so_3)$. If q is a primitive root of unity, then this assertion is stated in Ref. 5. If q is any root of unity, then this assertion may be proved in the following way. If $q^p=1$, then the center $\mathcal{C}$ of $U_q(so_3)$ contains the elements

$$P_p = I_j^p + aI_j^{p-2} + bI_j^{p-4} + \cdots + dI_j^r, \quad j=1,2,3,$$

where $r=0$ if p is even and $r=1$ if p is odd and $a,b,\ldots,d$ are certain fixed complex numbers expressed in terms of q. [They are the polynomials P defined in Ref. 5 if q is a primitive root of unity. Unfortunately, we could not find the explicit expressions for the coefficients $a,b,\ldots,d$. But note that $P_3 = I_j^3 + I_j$, $P_4 = I_j^4 + I_j^2$, and $P_5 = I_j^5 + (1+(q+q^{-1})^{-1})I_j^3 + (q+q^{-1})^{-1}I_j$.] Therefore, I_j^s, $s>n$, can be reduced to the linear combination of I_j^i, $i<n$, with coefficients from the center $\mathcal{C}$. Now our assertion follows from this and from the Poincaré–Birkhoff–Witt theorem for $\mathrm{U}_q(\mathrm{so}_3)$.

Theorem 4: *If q is a root of unity, then any irreducible representation of* $\mathrm{U}_q(\mathrm{so}_3)$ *is finite dimensional.*

Proof: Let T be an irreducible representation of $\mathrm{U}_q(\mathrm{so}_3)$. Then T maps central elements into scalar operators. Since the linear space $\mathrm{U}_q(\mathrm{so}_3)$ is finite dimensional over the center $\mathcal{C}$ with the basis $I_1^k I_2^m I_3^n$, $k,m,n<p$ then for any $a \in \mathrm{U}_q(\mathrm{so}_3)$ we have $T(a) = \Sigma_{k,m,n<p} T(I_1^k I_2^m I_3^n)$. Hence, if $\mathbf{v}$ is a nonzero vector of the representation space $\mathcal{V}$, then $T(\mathrm{U}_q(\mathrm{so}_3))\mathbf{v} = \mathcal{V}$ and $\mathcal{V}$ is finite dimensional. Theorem is proved.

Taking into account Theorem 4, below we consider only finite-dimensional representations of $\mathrm{U}_q(\mathrm{so}_3)$.

In order to find irreducible representations of $\mathrm{U}_q(\mathrm{so}_3)$ for q a root of unity, we use the same method as before, that is, we apply the homomorphism ψ from Proposition 2 and irreducible representations of the algebra $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ for q a root of unity.

Let us find irreducible representations of $\hat{\mathrm{U}}_q(\mathrm{sl}_2)$ for q a root of unity. The quantum algebra $\mathrm{U}_q(\mathrm{sl}_2)$ for q a root of unity has the following irreducible representations (see Ref. 11, Subsection 3.3.2):

(a) The representations $T_l^{(1)}$, $T_l^{(-1)}$, $T_l^{(i)}$, $T_l^{(-i)}$, $2l<p'$, given by the formulas (15)–(19).

(b) The representations $T_{ab\lambda}$, a, b, $\lambda \in \mathbf{C}$, $\lambda \neq 0$, acting on a p'-dimensional vector space $\mathcal{H}$ with the basis $|j\rangle$, $j=0,1,2,\ldots,p'-1$, and given by the formulas

$$T_{ab\lambda}(q^H)|i\rangle = q^{-i}\lambda|i\rangle, \quad T_{ab\lambda}(F)|p'-1\rangle = b|0\rangle, \tag{36}$$

$$T_{ab\lambda}(F)|i\rangle = |i+1\rangle, \quad i<p'-1, \quad T_{ab\lambda}(E)|0\rangle = a|p'-1\rangle, \tag{37}$$

$$T_{ab\lambda}(E)|i\rangle = \left(ab + [i]\frac{\lambda^2 q^{1-i} - \lambda^{-2} q^{i-1}}{q-q^{-1}} \right)|i-1\rangle, \quad i>0. \tag{38}$$

The representations $T_{ab\lambda}$ with $(a,b)=(0,0)$ and $\lambda = \pm q^n$, $n=0,1,2,\ldots,p'-2$, are reducible and must be taken out from this set.

(c) The representations $T'_{0b\lambda}$, b, $\lambda \in \mathbf{C}$, $\lambda \neq 0$, acting on a p'-dimensional vector space $\mathcal{H}$ with the basis $|j\rangle$, $j=0,1,2,\ldots,p'-1$, and given by the formulas

$$T'_{0b\lambda}(q^H)|i\rangle = q^i \lambda^{-1}|i\rangle, \quad T'_{0b\lambda}(E)|p'-1\rangle = b|0\rangle, \tag{39}$$

$$T'_{0b\lambda}(E)|i\rangle = |i+1\rangle, \quad i<p'-1, \quad T_{0b\lambda}(F)|0\rangle = 0, \tag{40}$$

$$T'_{0b\lambda}(F)|i\rangle = [i]\frac{\lambda^2 q^{1-i} - \lambda^{-2} q^{i-1}}{q-q^{-1}}|i-1\rangle, \quad i>0. \tag{41}$$

The representations $T'_{00\lambda}$ with $\lambda = \pm q^n$, $n=0,1,2,\ldots,p'-2$, are reducible and must be taken out from this set.

Remark 1: In the set of representations (a)–(c) there exist equivalent representations (see, for example, Propositions 3.17 and 3.18 in Ref. 11).

Remark 2: In Ref. 11, Subsection 3.3.2, irreducible representations of the algebra generated by the elements $E,F,K := q^{2H}$, $K^{-1} := q^{-2H} \in \mathrm{U}_q(\mathrm{sl}_2)$ are given. Clearly, this algebra is a subalgebra in $U_q(sl_2)$. It is easy to generalize the results of Subsection 3.3.2 in Ref. 11 for $U_q(sl_2)$. Let us note that the algebra $U_q(sl_2)$ has a unique automorphism φ such that $\varphi(q^H)=iq^H$, $\varphi(E)=-E$, and $\varphi(F)=F$. (If q is not a root of unity, then this automorphism transforms the representations $T_l^{(1)}$ to the representations $T_l^{(i)}$, respectively.) Therefore, the mapping $\tilde{T}_{-a,b,\lambda}=T_{ab\lambda}\circ\varphi$ is also a representation of $U_q(sl_2)$. We have

$$\tilde{T}_{ab\lambda}(q^H)|i\rangle=iq^{-i}\lambda|i\rangle,\quad \tilde{T}_{ab\lambda}(F)|p'-1\rangle=b|0\rangle, \tag{42}$$

$$\tilde{T}_{ab\lambda}(F)|i\rangle=|i+1\rangle,\quad i<p'-1,\quad \tilde{T}_{ab\lambda}(E)|0\rangle=a|p'-1\rangle, \tag{43}$$

$$\tilde{T}_{ab\lambda}(E)|i\rangle=\left(ab-[i]\frac{\lambda^2q^{1-i}-\lambda^{-2}q^{i-1}}{q-q^{-1}}\right)|i-1\rangle,\quad i>0. \tag{44}$$

However, it is easy to see by comparing (36)–(38) with (42)–(44) that the representation $\tilde{T}_{ab\lambda}$ is equivalent to $T_{a,b,i\lambda}$. This means that for q a root of unity we do not obtain new representations of $U_q(sl_2)$ from $T_{ab\lambda}$ applying the automorphism φ as in the case of the representations $T_l^{(1)}$.

We have described irreducible representations of the algebra $U_q(sl_2)$. Now we wish to extend these representations to obtain representations of the algebra $\hat{U}_q(sl_2)$ by using the relation

$$T((q^kq^H+q^{-k}q^{-H})^{-1}):=(q^kT(q^H)+q^{-k}T(q^{-H}))^{-1}.$$

Clearly, only those irreducible representations T of $U_q(sl_2)$ can be extended to $\hat{U}_q(sl_2)$ for which the operators $q^kT(q^H)+q^{-k}T(q^{-H})$ are invertible. From formulas (15)–(19) it is clear that these operators are always invertible for the irreducible representations $T_l^{(1)}$, $T_l^{(-1)}$, $l=0,\frac{1}{2},1,\frac{3}{2},\ldots,(p'-1)/2$, and for the irreducible representations $T_l^{(i)}$, $T_l^{(-i)}$, $l=\frac{1}{2},\frac{3}{2},\frac{5}{2},\ldots,(p'-1)/2$ [or $(p'-2)/2$.] (For the representations $T_l^{(i)}$, $T_l^{(-i)}$, $l=0,1,2,\ldots$, some of these operators are not invertible since they have zero eigenvalue.) We denote the extended representations by the same symbols $T_l^{(1)}$, $T_l^{(-1)}$, $T_l^{(i)}$, and $T_l^{(-i)}$, respectively.

Similarly, the representation $T_{ab\lambda}$ (and the representation $T'_{0b\lambda}$) can be extended to a representation of the algebra $\hat{U}_q(sl_2)$ if and only if $\lambda\neq\pm iq^k$, $k\in\mathbf{Z}$.

Proposition 7: The algebra $\hat{U}_q(sl_2)$ *for q a root of unity has the irreducible representations* $T_l^{(1)}$, $T_l^{(-1)}$, $l=0,\frac{1}{2},1,\frac{3}{2},\ldots,(p'-1)/2$, *the irreducible representations* $T_l^{(i)}$, $T_l^{(-i)}$, $l=\frac{1}{2},\frac{3}{2},\frac{5}{2},\ldots,(p'-1)/2$ [*or* $(p'-2)/2$], *and the irreducible representations* $T_{ab\lambda}$, $T'_{0b\lambda}$, $\lambda\neq\pm iq^k$, $k\in\mathbf{Z}$. *Any irreducible representation of* $\hat{U}_q(sl_2)$ *for q a root of unity is equivalent to one of these representations.*

IX. REPRESENTATIONS OF $U_q(so_3)$ FOR q A ROOT OF UNITY OBTAINED FROM THOSE OF $\hat{U}_q(sl_2)$

As in Sec. IV, we shall obtain representations of $U_q(so_3)$ for q a root of unity by applying the homomorphism ψ from Proposition 2. Namely, if T is a representation of $\hat{U}_q(sl_2)$, then

$$R=T\circ\psi \tag{45}$$

is a representation of $U_q(so_3)$. As in Sec. IV, application of this method to the pair of the irreducible representations $T_l^{(1)}$ and $T_l^{(-1)}$ of $\hat{U}_q(sl_2)$ leads to the same representation of $U_q(so_3)$ which will be denoted by $R_l^{(1)}$. Applying the formula (45) to the irreducible representations $T_l^{(i)}$ and $T_l^{(-i)}$ of $\hat{U}_q(sl_2)$ gives the representations of $U_q(so_3)$ which will be denoted by $R_l^{(i)}$ and $R_l^{(-i)}$, respectively.

Proposition 8: The representations $R_l^{(1)}$ *of* $U_q(so_3)$ *are irreducible. The representations* $R_l^{(i)}$ *and* $R_l^{(-i)}$ *are reducible.*

Proof of this proposition is the same as that of Proposition 3.

Repeating word-by-word the reasoning of Sec. IV, we decompose the representations $R_l^{(i)}$ and $R_l^{(-i)}$ into the direct sums of representations of $U_q(\mathrm{so}_3)$ which are denoted by $R_n^{(\pm i,+)}$ and $R_n^{(\pm i,-)}$:

$$R_l^{(i)}=R_n^{(i,+)}\oplus R_n^{(i,-)},\quad R_l^{(-i)}=R_n^{(-i,+)}\oplus R_n^{(-i,-)},\quad n=l+\tfrac{1}{2}.$$

Moreover, the representations $R_n^{(\pm i,+)}$ and $R_n^{(\pm i,-)}$ are given in the appropriate bases $|1\rangle,|2\rangle,\ldots,|n\rangle$ by the corresponding formulas of Sec. IV.

Theorem 5: *The representations $R_n^{(i,+)}$, $R_n^{(i,-)}$, $R_n^{(-i,+)}$, and $R_n^{(-i,-)}$, $n=1,2,3,...,p'/2$ [or $(p'-1)/2$] are irreducible and pairwise nonequivalent. For any l, $l=0,\frac{1}{2},1,\frac{3}{2},\ldots,(p'-1)/2$, the representation $R_l^{(1)}$ is not equivalent to some of these representations.*

Proof is the same as that of Theorem 1.

Now we apply formula (45) to the representations $T_{ab\lambda}$ and $T'_{0b\lambda}$. As a result, we obtain the representations

$$R_{ab\lambda}=T_{-a,b,-i\lambda}\circ\psi,\quad R'_{0b\lambda}=T'_{0,b-i\lambda}$$

given in the bases $|j\rangle$, $j=0,1,2,...,p'-1$, by the formulas

$$R_{ab\lambda}(I_1)|i\rangle=\frac{-1}{q-q^{-1}}(q^{-i}\lambda+q^{i}\lambda^{-1})|i\rangle, \tag{46}$$

$$R_{ab\lambda}(I_2)|0\rangle=\frac{i}{\lambda-\lambda^{-1}}(a|p'-1\rangle+|1\rangle), \tag{47}$$

$$R_{ab\lambda}(I_2)|p'-1\rangle=\frac{i}{q^{-p'+1}\lambda-q^{p'-1}\lambda^{-1}}\times\left\{b|0\rangle+\left(ab+[p'-1]\frac{q^{-p'+2}\lambda^2-q^{p'-2}\lambda^{-2}}{q-q^{-1}}\right)|p'-2\rangle\right\}, \tag{48}$$

$$R_{ab\lambda}(I_2)|i\rangle=\frac{i}{q^{-i}\lambda-q^{i}\lambda^{-1}}\left\{\left(ab+[i]\frac{q^{-i+1}\lambda^2-q^{i-1}\lambda^{-2}}{q-q^{-1}}\right)|i-1\rangle+|i+1\rangle\right\},\quad 0<i<p'-1, \tag{49}$$

and by the formulas

$$R'_{0b\lambda}(I_1)|i\rangle=\frac{1}{q-q^{-1}}(q^{-i}\lambda+q^{i}\lambda^{-1})|i\rangle,\quad R'_{0b\lambda}(I_2)|0\rangle=\frac{-i}{\lambda-\lambda^{-1}}|1\rangle,$$

$$R'_{0b\lambda}(I_2)|p'-1\rangle=\frac{-i}{q^{-p'+1}\lambda-q^{p'-1}\lambda^{-1}}\left(b|0\rangle+[p'-1]\frac{q^{-p'+2}\lambda^2-q^{p'-2}\lambda^{-2}}{q-q^{-1}}|p'-2\rangle\right),$$

$$R'_{0b\lambda}(I_2)|i\rangle=\frac{-i}{q^{-i}\lambda-q^{i}\lambda^{-1}}\left(|i+1\rangle+[i]\frac{q^{-i+1}\lambda^2-q^{i-1}\lambda^{-2}}{q-q^{-1}}|i-1\rangle\right),\quad 0<i<p'-1.$$

The operators $R_{ab\lambda}(I_3)$ and $R'_{0b\lambda}(I_3)$ can be calculated by means of the relation

$$R(I_3)=q^{1/2}R(I_1)R(I_2)-q^{-1/2}R(I_2)R(I_1).$$

Recall that the representations $R_{ab\lambda}$ and $R'_{0b\lambda}$ are determined for $\lambda\neq 0$ and $\lambda\neq\pm q^k$, $k\in\mathbf{Z}$.

It is seen from the above formulas that

$$R'_{0b\lambda}(I_1)=R_{0,b,-\lambda}(I_1),\quad R'_{0b\lambda}(I_2)=R_{0,b,-\lambda}(I_2),$$

that is, the representations $R_{0,b,-\lambda}$ and $R'_{0b\lambda}$ are equivalent. For this reason, we consider below only the representations $R_{ab\lambda}$.

In order to study the representations $R_{ab\lambda}$ of $\mathrm{U}_q(\mathrm{so}_3)$ we consider the spectrum of the operator $R_{ab\lambda}(I_1)$. It coincides with the set of points

$$-\frac{\lambda+\lambda^{-1}}{q-q^{-1}},\quad -\frac{q^{-1}\lambda+q\lambda^{-1}}{q-q^{-1}},\quad -\frac{q^{-2}\lambda+q^2\lambda^{-1}}{q-q^{-1}},\quad \ldots,\quad -\frac{q^{1-p'}\lambda+q^{p'-1}\lambda^{-1}}{q-q^{-1}}. \tag{50}$$

It is easy to see that there exist coinciding points in this set if and only if λ is equal to one of the numbers

$$\pm q^{1/2},\pm q^{3/2},\pm q^{5/2},\ldots,\pm q^{(p'-1)/2}\quad (\text{or}\ \ \pm q^{(p'-2)/2}).$$

(Here we have to take $\pm q^{(p'-1)/2}$ if p' is even and $\pm q^{(p'-2)/2}$ if p' is odd.) Moreover, the set (50) splits into pairs of coinciding points if and only if $\lambda=\pm q^{(p'-1)/2}$. In all other cases there exists at least one spectral point which coincides with no other point. In particular, if $\lambda=\pm q^{(p'-2)/2}$, then in this set there exists only one eigenvalue with multiplicity 1. In all other cases there are more than one eigenvalue with multiplicity 1.

Proposition 9: If $\lambda\neq\pm q^{(p'-1)/2}$ for even p' and $\lambda\neq\pm q^{(p'-2)/2}$ for odd p', then the representation $R_{ab\lambda}$ is irreducible.

Proof: Let $\lambda\neq\pm q^{(p'-1)/2}$ for even p' and $\lambda\neq\pm q^{(p'-2)/2}$ for odd p'. We distinguish two cases: when the spectrum of the operator $R_{ab\lambda}(I_1)$ is simple and when there exists at least one spectral point of this operator having multiplicity 2. In the first case the proof is the same as the first part of the proof of Proposition 3. For the second case, we give a proof only for $\lambda=q^{1/2}$. (Proofs for other values of q are similar.) Then in the set (50) there are only two coinciding points $-(\lambda+\lambda^{-1})/(q-q^{-1})$ and $-(q^{-1}\lambda+q\lambda^{-1})/(q-q^{-1})$ corresponding to the eigenvectors $|0\rangle$ and $|1\rangle$. Let V be an invariant subspace of the representation space $\mathcal{H}$. As in the proof of Proposition 3, it is shown that V is a linear span of eigenvectors of the operator $R_{ab\lambda}(I_1)$, that is, a certain part of the vectors $|i\rangle$, $i\neq 0$, 1, $\alpha_0|0\rangle+\alpha_1|1\rangle$, $\beta_0|0\rangle+\beta_1|1\rangle$ constitutes a basis of V. Let V contain some basis vector $|j\rangle$. Then as in the proof of Proposition 3, acting successively upon $|j\rangle$ by certain linear combinations of the operators $R_{ab\lambda}(I_2)$ and $R_{ab\lambda}(I_3)$ we generate all the vectors $|i\rangle$, $i=0,1,\ldots,\frac{1}{2}(p'-1)$. This means that $V=\mathcal{H}$ and the representation $R_{ab\lambda}$ is irreducible. If V contains no vector $|j\rangle$, $j\neq 0,1$, then some linear combination $\alpha_0|0\rangle+\alpha_1|1\rangle$ belongs to V. Then the vector $\mathbf{v}=R_{ab\lambda}(I_2)(\alpha_0|0\rangle+\alpha_1|1\rangle)$ belongs to V. Since $\mathbf{v}$ contains the summand $\alpha|2\rangle$ with nonzero coefficient α, then $|2\rangle\in V$. This is a contradiction. Hence, the representation $R_{ab\lambda}$ is irreducible. Proposition is proved.

Let p' be even. Let us study the representations $R_{ab\lambda}$ for $\lambda=\pm q^{(p'-1)/2}$. For $\lambda=q^{(p'-1)/2}$ we have

$$R_{ab\lambda}(I_1)|i\rangle=\frac{-1}{q-q^{-1}}(q^{-i+(p'-1)/2}+q^{i-(p'-1)/2})|i\rangle, \tag{51}$$

$$R_{ab\lambda}(I_2)|0\rangle=c_{(p'-1)/2}(a|p'-1\rangle+|1\rangle), \tag{52}$$

$$R_{ab\lambda}(I_2)|p'-1\rangle=-c_{(p'-1)/2}((ab+[p'-1]^2)|p'-2\rangle+b|0\rangle), \tag{53}$$

$$R_{ab\lambda}(I_2)|i\rangle=c_{-i+(p'-1)/2}((ab+[i]^2)|i-1\rangle+|i+1\rangle), \tag{54}$$

where

$$c_j=\frac{i}{q^j-q^{-j}}.$$

The operator $R_{a,b,(p'-1)/2}(I_1)$ has the spectrum

$$\frac{-1}{q-q^{-1}}(q^{-i+(p'-1)2}+q^{i-(p'-1)/2}),\quad i=0,1,2,...,p'-1,$$

that is, if p' is even, then all spectral points are of multiplicity 2.

We assume that $ab\neq-[j]^2$, $j=0,1,...,p'-1$, and go over from the basis $\{|i\rangle\}$ to the basis $\{|i\rangle^\circ\}$, where

$$|i\rangle^\circ=\prod_{j=0}^{i}(ab+[j]^2)^{-1/2}|i\rangle,\quad i=0,1,2,...,p'-1.$$

Then the formula (51) does not change and the formulas (52)–(54) turn into

$$R_{ab\lambda}(I_2)|0\rangle^\circ=c_{(p'-1)/2}\left(a\prod_{j=1}^{p'-1}(ab+[j]^2)^{1/2}|p'-1\rangle^\circ+(ab+1)^{1/2}|1\rangle^\circ\right),$$

$$R_{ab\lambda}(I_2)|p'-1\rangle^\circ=-c_{(p'-1)/2}\left((ab+1)^{1/2}|p'-2\rangle^\circ+\frac{b}{\Pi_{j=1}^{p'-1}(ab+[j]^2)^{1/2}}|0\rangle^\circ\right),$$

$$R_{ab\lambda}(I_2)|i\rangle^\circ=c_{-i+(p'-1)/2}((ab+[i]^2)^{1/2}|i-1\rangle^\circ+(ab+[i+1]^2)^{1/2}|i+1\rangle^\circ).$$

We split the representation space $\mathcal{H}$ into the direct sum of two linear subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$ spanned by the basis vectors $|j\rangle'$, $j=0,1,2,...,\frac{1}{2}(p'-2)$, and $|j\rangle''$, $j=0,1,2,...,\frac{1}{2}(p'-2)$, where

$$|j\rangle'=|j\rangle^\circ+i(-1)^{-j-1+p'/2}|p'-j-1\rangle^\circ,\quad |j\rangle''=|j\rangle^\circ+i(-1)^{-j+p'/2}|p'-j-1\rangle^\circ.$$

Then as in Sec. IV, we derive

$$R_{a,b,(p'-1)/2}(I_1)|j\rangle'=\frac{-1}{q-q^{-1}}(q^{-j+(p'-1)/2}+q^{j-(p'-1)/2})|j\rangle',$$

$$R_{a,b,(p'-1)/2}(I_1)|j\rangle''=\frac{-1}{q-q^{-1}}(q^{-j+(p'+1)/2}+q^{j-(p'-1)/2})|j\rangle'',$$

for the operator $R_{a,b,(p'-1)/2}(I_1)$ and

$$R_{a,b,(p'-1)/2}(I_2)|j\rangle'=c_{-j+(p'-1)/2}((ab+[j+1]^2)^{1/2}|j+1\rangle'+(ab+[j]^2)^{1/2}|j-1\rangle'),$$

$$R_{a,b,(p'-1)/2}(I_2)|j\rangle''=c_{-j+(p'-1)/2}((ab+[j+1]^2)^{1/2}|j+1\rangle''+(ab+[j]^2)^{1/2}|j-1\rangle''),$$

where $j\neq0$, $p'/2-1$,

$$R_{a,b,(p'-1)/2}(I_2)\left|\frac{p'}{2}-1\right\rangle'=\frac{1}{q^{1/2}-q^{1/2}}\left(ab+\left[\frac{p'}{2}\right]^2\right)^{1/2}\left|\frac{p'}{2}-1\right\rangle'$$
$$+\frac{i}{q^{1/2}-q^{-1/2}}\left(ab+\left[\frac{p'}{2}-1\right]^2\right)^{1/2}\left|\frac{p'}{2}-2\right\rangle',$$

$$R_{a,b,(p'-1)/2}(I_2)\left|\frac{p'}{2}-1\right\rangle'' = -\frac{1}{q^{1/2}-q^{-1/2}}\left(ab+\left[\frac{p'}{2}\right]^2\right)^{1/2}\left|\frac{p'}{2}-1\right\rangle'' + \frac{i}{q^{1/2}-q^{-1/2}}\left(ab+\left[\frac{p'}{2}-1\right]^2\right)^{1/2}\left|\frac{p'}{2}-2\right\rangle'',$$

$$R_{a,b,(p'-1)/2}(I_2)|0\rangle' = c_{(p'-1)/2}\left(a\prod_{j=1}^{p'-1}(ab+[j]^2)^{1/2}|p'-1\rangle^\circ + (ab+1)^{1/2}|1\rangle^\circ\right) - i(-1)^{(p'-2)/2}c_{(p'-1)/2}\left((ab+1)^{1/2}|p'-2\rangle^\circ + \frac{b}{\prod_{j=1}^{p'-1}(ab+[j]^2)^{1/2}}|0\rangle^\circ\right).$$

When

$$a\prod_{j=1}^{p'-1}(ab+[j]^2)^{1/2} = \frac{b}{\prod_{j=1}^{p'-1}(ab+[j]^2)^{1/2}}, \tag{55}$$

then the last relation reduces to

$$R_{a,b,(p'-1)/2}(I_2)|0\rangle' = \frac{(-1)^{(p'-2)/2}}{q^{(p'-1)/2}-q^{-(p'-1)/2}}a\prod_{j=1}^{p'-1}(ab+[j]^2)^{1/2}|0\rangle' + c_{(p'-1)/2}(ab+1)^{1/2}|1\rangle'.$$

Similarly, if the condition (55) is fulfilled, then

$$R_{a,b,(p'-1)/2}(I_2)|0\rangle'' = \frac{(-1)^{p'/2}}{q^{(p'-1)/2}-q^{-(p'-1)/2}}a\prod_{j=1}^{p'-1}(ab+[j]^2)^{1/2}|0\rangle'' + c_{(p'-1)/2}(ab+1)^{1/2}|1\rangle''.$$

Thus, the subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are invariant with respect to the representation $R_{a,b,(p'-1)/2}$ if the condition (55) is fulfilled. We denote the corresponding subrepresentations by $R^{1,+}_{a,b,(p'-1)/2}$ and $R^{2,+}_{a,b,(p'-1)/2}$, respectively.

Similarly, if $\lambda = -q^{(p'-1)/2}$, then

$$R_{a,b,-(p'-1)/2}(I_1) = -R_{a,b,(p'-1)2}(I_1), \quad R_{a,b,-(p'-1)/2}(I_2) = -R_{a,b,(p'-1)/2}(I_2)$$

and the subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are invariant with respect to the representation $R_{a,b,-(p'-1)/2}$ if the condition (55) is fulfilled. We denote the corresponding subrepresentations by $R^{1,-}_{a,b,-(p'-1)/2}$ and $R^{2,-}_{a,b,-(p'-1)/2}$, respectively.

Proposition 10: Let the condition (55) be satisfied. Then the representations $R^{i,+}_{a,b,(p'-1)/2}$ *and* $R^{i,-}_{a,b,-(p'-1)/2}$, $i=1,2$, *of the algebra* $\mathrm{U}_q(\mathrm{so}_3)$ *are irreducible and pairwise nonequivalent. If the condition (55) is not satisfied, then the representations* $R_{a,b,(p'-1)/2}$ *and* $R_{a,b,-(p'-1)/2}$ *are irreducible.*

Proof is similar to that of the previous propositions and we omit it.

Remark that the representations $R^{i,+}_{a,b,(p'-1)/2}$ and $R^{i,-}_{a,b,-(p'-1)/2}$, $i=1,2$, have two nonzero diagonal matrix elements $\langle p'/2-1|R|p'/2-1\rangle$ and $\langle 0|R|0\rangle$.

Let now p' be odd and $\lambda = q^{(p'-2)/2}$. For this value of λ we have

$$R_{ab\lambda}(I_1)|i\rangle = \frac{-1}{q-q^{-1}}(q^{-i+(p'-2)/2}+q^{i-(p'-2)/2})|i\rangle,$$

$$R_{ab\lambda}(I_2)|0\rangle = c_{(p'-2)/2}(a|p'-1\rangle + |1\rangle),$$

$$R_{ab\lambda}(I_2)|p'-1\rangle = -c_{p'/2}((ab+\epsilon[p'-1][p'])|p'-2\rangle + b|0\rangle),$$

$$R_{ab\lambda}(I_2)|i\rangle = c_{-i+(p'-2)/2}((ab+\epsilon[i][i+1])|i-1\rangle + |i+1\rangle),$$

where $\epsilon=1$ for $p'=p/2$, $\epsilon=-1$ for $p'=p$, and c_j is such as in (51)–(54). The operator $R_{a,b,(p'-2)/2}(I_1)$ has the spectrum

$$\frac{-1}{q-q^{-1}}(q^{-i+(p'-2)/2}+q^{i-(p'-2)/2}), \quad i=0,1,2,...,p'-1,$$

that is, all spectral points are of multiplicity 2 except for the point $-(q^{p'/2}+q^{-p'/2})/(q-q^{-1})$ which is of multiplicity 1.

We assume that $ab \neq -\epsilon[j][j+1]$, $j=0,1,...,p'-1$, and go over from the basis $\{|i\rangle\}$ to the basis $\{|i\rangle^\circ\}$, where

$$|i\rangle^\circ = \prod_{j=0}^{i}(ab+\epsilon[j][j+1])^{-1/2}|i\rangle, \quad i=0,1,2,...,p'-1.$$

Then

$$R_{ab\lambda}(I_1)|i\rangle^\circ = \frac{-1}{q-q^{-1}}(q^{-i+(p'-2)/2}+q^{i-(p'-2)/2})|i\rangle^\circ,$$

$$R_{ab\lambda}(I_2)|0\rangle^\circ = c_{(p'-2)/2}\left(a\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2}|p'-1\rangle^\circ + (ab+\epsilon[2])^{1/2}|1\rangle^\circ\right),$$

$$R_{ab\lambda}(I_2)|p'-1\rangle^\circ = c_{p'/2}\left((ab+\epsilon[p'-1][p'])^{1/2}|p'-2\rangle^\circ + b\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{-1/2}|0\rangle^\circ\right),$$

$$R_{ab\lambda}(I_2)|i\rangle^\circ = c_{-i+(p'-2)/2}((ab+\epsilon[i][i+1])^{1/2}|i-1\rangle^\circ + (ab+\epsilon[i+1][i+2])^{1/2}|i+1\rangle^\circ),$$

where $\lambda = q^{(p'-2)/2}$. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two linear subspaces of the representation space $\mathcal{H}$ spanned by the basis vectors

$$|j\rangle' = |j\rangle^\circ + i(-1)^j|p'-j-2\rangle^\circ, \quad j=0,1,2,...,\frac{p'-3}{2},$$

and the basis vectors

$$|j\rangle'' = |j\rangle^\circ + i(-1)^{j+1}|p'-j-2\rangle^\circ, \quad j=0,1,2,...,\frac{p'-3}{2},$$

respectively. Then the operator $R_{a,b,(p'-2)/2}(I_1)$ acts on the basis elements $|j\rangle'$ and $|j\rangle''$ as on the vectors $|j\rangle$ and

$$R_{a,b,(p'-1)/2}(I_2)|j\rangle' = c_{-j+(p'-2)/2}((ab+\epsilon[j+1][j+2])^{1/2}|j+1\rangle' + (ab+\epsilon[j][j+1])^{1/2}|j-1\rangle'),$$

$$R_{a,b,(p'-2)/2}(I_2)|j\rangle'' = c_{-j+(p'-2)/2}((ab+\epsilon[j+1][j+2])^{1/2}|j+1\rangle'' + (ab+\epsilon[j][j+1])^{1/2}|j-1\rangle''),$$

where $j\neq 0,(p'-3)/2$,

$$R_{a,b,(p'-2)/2}(I_2)\left|\frac{p'-3}{2}\right\rangle' = \frac{(-1)^{(p'-3)/2}}{q^{1/2}-q^{-1/2}}\left(ab+\epsilon\left[\frac{p'-1}{2}\right]\left[\frac{p'+1}{2}\right]\right)^{1/2}\left|\frac{p'-3}{2}\right\rangle' + \frac{i}{q^{1/2}-q^{-1/2}}\left(ab+\epsilon\left[\frac{p'-1}{2}\right]\left[\frac{p'-3}{2}\right]\right)^{1/2}\left|\frac{p'-5}{2}\right\rangle',$$

$$R_{a,b,(p'-2)/2}(I_2)\left|\frac{p'-3}{2}\right\rangle'' = -\frac{(-1)^{(p'-3)/2}}{q^{1/2}-q^{-1/2}}\left(ab+\epsilon\left[\frac{p'-1}{2}\right]\left[\frac{p'+1}{2}\right]\right)^{12}\left|\frac{p'-3}{2}\right\rangle'' + \frac{i}{q^{1/2}-q^{-1/2}}\left(ab+\epsilon\left[\frac{p'-1}{2}\right]\left[\frac{p'-3}{2}\right]\right)^{1/2}\left|\frac{p'-5}{2}\right\rangle'',$$

$$R_{a,b,(p'-2)/2}(I_2)|0\rangle' = c_{(p'-2)/2}\left(a\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2}|p'-1\rangle^\circ + (ab+\epsilon[2])^{1/2}|1\rangle^\circ - i(ab+\epsilon[2])^{1/2}|p'-1\rangle^\circ - i(ab+\epsilon[p'-2][p'-1])^{1/2}|p'-3\rangle^\circ\right),$$

$$R_{a,b,(p'-2)/2}(I_2)|0\rangle'' = c_{(p'-2)/2}\left(a\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2}|p'-1\rangle^\circ + (ab+\epsilon[2])^{1/2}|1\rangle^\circ + i(ab+\epsilon[2])^{1/2}|p'-1\rangle^\circ + i(ab+\epsilon[p'-2][p'-1])^{1/2}|p'-3\rangle^\circ\right).$$

If

$$a\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2} + i(ab+\epsilon[2])^{1/2} = 0, \tag{56}$$

$$(ab+\epsilon[2])^{1/2}\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2} = ib, \tag{57}$$

then

$$R_{a,b,(p'-2)/2}(I_2)|p'-1\rangle = \frac{-bc_{p'/2}}{\Pi_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2}}|0\rangle',$$

$$R_{a,b,(p'-2)/2}(I_2)|0\rangle' = \frac{i(ab+\epsilon[2])^{1/2}}{q^{(p'-2)/2}-q^{-(p'-2)/2}}|1\rangle' + c|p'-1\rangle',$$

$$R_{a,b,(p'-2)/2}(I_2)|0\rangle'' = \frac{i(ab+\epsilon[2])^{1/2}}{q^{(p'-2)/2}-q^{-(p'-2)/2}}|1\rangle'',$$

where c is a nonzero coefficient easily determined from the above formulas. Hence, the subspaces $\mathcal{H}_1+\mathbf{C}|p'-1\rangle$ and $\mathcal{H}_2$ of the representation space are invariant with respect to the representation $R_{a,b,(p'-2)/2}$ (we denote these subrepresentations by $R^1_{a,b,(p'-2)/2}$ and $R^2_{a,b,(p'-2)/2}$, respectively). Remark that

$$\dim \mathcal{H}_1+\mathbf{C}|p'-1\rangle=\tfrac{1}{2}(p'+1),\quad \dim \mathcal{H}_2=\tfrac{1}{2}(p'-1).$$

If

$$a\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2}-i(ab+\epsilon[2])^{1/2}=0, \tag{58}$$

$$(ab+\epsilon[2])^{1/2}\prod_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2}=-ib, \tag{59}$$

then

$$R_{a,b,(p'-2)/2}(I_2)|p'-1\rangle=\frac{-bc_{p'/2}}{\Pi_{j=1}^{p'-1}(ab+\epsilon[j][j+1])^{1/2}}|0\rangle'',$$

$$R_{a,b,(p'-2)/2}(I_2)|0\rangle'=\frac{i(ab+\epsilon[2])^{1/2}}{q^{(p'-2)/2}-q^{-(p'-2)/2}}|1\rangle',$$

$$R_{a,b,(p'-2)/2}(I_2)|0\rangle''=\frac{i(ab+\epsilon[2])^{1/2}}{q^{(p'-2)/2}-q^{-(p'-2)/2}}|1\rangle''+c|p'-1\rangle,$$

where c is a nonzero coefficient. Hence, now the subspaces $\mathcal{H}_1$ and $\mathcal{H}_2+\mathbf{C}|p'-1\rangle$ of the representation space are invariant. We denote the subrepresentations on these subspaces by $\hat{R}^1_{a,b,(p'-2)/2}$ and $\hat{R}^2_{a,b,(p'-2)/2}$, respectively). Note that the representation $\hat{R}^1_{a,b,(p'-2)/2}$ is not equivalent to $R^2_{a,b,(p'-2)/2}$ (and the representation $\hat{R}^2_{a,b,(p'-2)/2}$ is not equivalent to $R^1_{a,b,(p'-2)/2}$) since the parameters a and b determining these representations satisfy different equations.

If a and b do not satisfy the relations (56) and (57) or the relations (58) and (59), then the representation $R_{a,b,(p'-2)/2}$ is irreducible.

Let now p' be odd and $\lambda=-q^{(p'-2)/2}$. In this case, the representation $R_{a,b,-(p'-2)/2}$ is irreducible if a and b do not satisfy the relations (56) and (57) or the relations (58) and (59). If a and b satisfy the relations (56) and (57), then $R_{a,b,-(p'-2)/2}$ is a reducible representation and decomposes into the direct sum of two subrepresentations acting on the subspaces $\mathcal{H}_1+\mathbf{C}|p'-1\rangle$ and $\mathcal{H}_2$. These subrepresentations are denoted by $\hat{R}^1_{a,b,-(p'-2)/2}$ and $R^2_{a,b,-(p'-2)/2}$, respectively, and are determined as

$$R^i_{a,b,-(p'-2)/2}(I_1)=-R^i_{a,b,(p'-2)/2}(I_1),\quad R^i_{a,b,-(p'-2)/2}(I_2)=-R^i_{a,b,(p'-2)/2}(I_2),\quad i=1,2.$$

Similarly, if a and b satisfy the relations (58) and (59), then $R_{a,b,-(p'-2)/2}$ is a reducible representation and decomposes into the direct sum of two subrepresentations acting on the subspaces $\mathcal{H}_1$ and $\mathcal{H}_2+\mathbf{C}|p'-1\rangle$. These subrepresentations are denoted by $\hat{R}^1_{a,b,-(p'-2)/2}$ and $\hat{R}^2_{a,b,-(p'-2)/2}$, respectively, and are determined as

$$\hat{R}^i_{a,b,-(p'-2)/2}(I_1)=-\hat{R}^i_{a,b,(p'-2)/2}(I_1),\quad \hat{R}^i_{a,b,-(p'-2)/2}(I_2)=-\hat{R}^i_{a,b,(p'-2)/2}(I_2),\quad i=1,2.$$

Proposition 11: Let the conditions (56) and (57) be satisfied. Then the representations $R^1_{a,b,(p'-2)/2}$, $R^2_{a,b,(p'-2)/2}$, $R^1_{a,b,-(p'-2)/2}$, *and* $R^2_{a,b,-(p'-2)/2}$ *are irreducible and pairwise nonequivalent. If the conditions (58) and (59) are satisfied, the representations* $\hat{R}^1_{a,b,(p'-2)/2}$, $\hat{R}^2_{a,b,(p'-2)/2}$, $\hat{R}^1_{a,b,-(p'-2)/2}$, *and* $\hat{R}^2_{a,b,-(p'-2)/2}$ *are irreducible and pairwise nonequivalent.*

Proof is similar to that of the previous propositions and we omit it.

X. OTHER REPRESENTATIONS OF $U_q(so_3)$ FOR q A ROOT OF UNITY

In the previous section we described irreducible representations of $U_q(so_3)$ obtained from irreducible representations of the algebra $\hat{U}_q(sl_2)$ for q a root of unity. However, at q a root of unity the algebra $U_q(so_3)$ has irreducible representations which cannot be derived from those of $\hat{U}_q(sl_2)$. They are obtained as irreducible components of the representations Q_λ from Sec. VII when one put q equal to a root of unity. We describe these representations of $U_q(so_3)$ in this section.

Let $\lambda = q^\tau$ be a nonzero complex number such that $0 \leq \mathrm{Re}\,\tau < 1$ and let $\mathcal{H}$ be the p'-dimensional complex vector space with basis

$$|m\rangle, \quad m=0,1,2,...,p'-1.$$

We define on this space the operators $Q'_\lambda(I_1)$ and $Q'_\lambda(I_2)$ determined by the formulas

$$Q'_\lambda(I_1)|m\rangle = \frac{\lambda q^m + \lambda^{-1} q^{-m}}{q-q^{-1}}|m\rangle,$$

$$Q'_\lambda(I_2)|0\rangle = \frac{1}{q-q^{-1}}|1\rangle + \frac{1}{q-q^{-1}}|p'-1\rangle,$$

$$Q'_\lambda(I_2)|p'-1\rangle = \frac{1}{q-q^{-1}}|p'-2\rangle + \frac{1}{q-q^{-1}}|0\rangle,$$

$$Q'_\lambda(I_2)|m\rangle = \frac{1}{q-q^{-1}}|m-1\rangle + \frac{1}{q-q^{-1}}|m+1\rangle, \quad m \neq 0, p'-1.$$

A direct computation shows that these operators satisfy the relations (7) and (8) and hence determine a representation of $U_q(so_3)$ which will be denoted by Q'_λ.

Theorem 6: *If* $\lambda \neq 1$ *and* $\lambda \neq q^{1/2}$, *then the representation* Q'_λ *is irreducible.*

Proof of this proposition is the same as that of the first part of Proposition 3.

The representations Q'_1 and $Q'_{\sqrt{q}}$ are studied in the same way as the representations Q_1 and $Q_{\sqrt{q}}$ in Sec. VII. This study leads to the irreducible representations of $U_q(so_3)$ which are described below. (Note that the description of these representations for p' even and for p' odd is different.)

Let p' be odd. We denote by $\mathcal{H}_r$ and $\mathcal{H}_s$, $r=\frac{1}{2}(p'+1)$ and $s=\frac{1}{2}(p'-1)$, the complex vector spaces with the bases

$$|0\rangle, |1\rangle, |2\rangle, \ldots, |\tfrac{1}{2}(p'-1)\rangle, \quad \text{and} \quad |1\rangle, |2\rangle, \ldots, |\tfrac{1}{2}(p'-1)\rangle,$$

respectively. Four representations $Q_1^{\pm,\pm}$ act on the space $\mathcal{H}_r$ and are given by the formulas

$$Q_1^{+,\pm}(I_1)|m\rangle = \frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle, \quad m=0,1,2,...,\tfrac{1}{2}(p'-1), \tag{60}$$

$$Q_1^{+,\pm}(I_2)|\tfrac{1}{2}(p'-1)\rangle = \pm\frac{1}{q-q^{-1}}|\tfrac{1}{2}(p'-1)\rangle + \frac{1}{q-q^{-1}}|\tfrac{1}{2}(p'-3)\rangle, \tag{61}$$

$$Q_1^{+,\pm}(I_2)|m\rangle = \frac{1}{q-q^{-1}}|m+1\rangle + \frac{1}{q-q^{-1}}|m-1\rangle, \quad m<\tfrac{1}{2}(p'-1), \tag{62}$$

and by the formulas

$$Q_1^{-,\pm}(I_1)|m\rangle = -\frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle, \quad m=0,1,2,...,\tfrac{1}{2}(p'-1), \tag{63}$$

$$Q_1^{-,\pm}(I_2) := Q_1^{+,\pm}(I_2). \tag{64}$$

Note that the upper sign corresponds to the representations $Q_1^{+,+}$ and $Q_1^{-,+}$ and the lower sign to the representations $Q_1^{+,-}$ and $Q_1^{-,-}$.

On the space $\mathcal{H}_s$, four representations $\hat{Q}_1^{\pm,\pm}$ act by the corresponding formulas (60)–(64), but now m runs over the values $1,2,3,...,\frac{1}{2}(p'-1)$.

Let now $\mathcal{H}_r'$ and $\mathcal{H}_s'$, $r=\frac{1}{2}(p'+1)$ and $s=\frac{1}{2}(p'-1)$, be the complex vector spaces with the bases

$$|m+\tfrac{1}{2}\rangle, \quad m=0,1,2,...,\tfrac{1}{2}(p'-1), \quad \text{and} \quad |m+\tfrac{1}{2}\rangle, \quad m=0,1,2,...,\tfrac{1}{2}(p'-3),$$

respectively. The four representations $Q_{\sqrt{q}}^{\pm,\pm}$ act on the space $\mathcal{H}_r'$ and are given by the formulas

$$Q_{\sqrt{q}}^{+,\pm}(I_1)|m+\tfrac{1}{2}\rangle = \frac{q^{m+1/2}+q^{-m-1/2}}{q-q^{-1}}|m+\tfrac{1}{2}\rangle, \quad m=0,1,2,...,\tfrac{1}{2}(p'-1), \tag{65}$$

$$Q_{\sqrt{q}}^{+,\pm}(I_2)|\tfrac{1}{2}\rangle = \pm\frac{1}{q-q^{-1}}|\tfrac{1}{2}\rangle + \frac{1}{q-q^{-1}}|\tfrac{3}{2}\rangle, \tag{66}$$

$$Q_{\sqrt{q}}^{+,\pm}(I_2)|m+\tfrac{1}{2}\rangle = \frac{1}{q-q^{-1}}|m+\tfrac{3}{2}\rangle + \frac{1}{q-q^{-1}}|m-\tfrac{1}{2}\rangle, \quad m\neq 0, \tag{67}$$

where $|m+\frac{3}{2}\rangle \equiv 0$ if $m=\frac{1}{2}(p'-1)$, and by the formulas

$$Q_{\sqrt{q}}^{-,\pm}(I_1)|m+\tfrac{1}{2}\rangle = -\frac{q^{m+1/2}+q^{-m-1/2}}{q-q^{-1}}|m+\tfrac{1}{2}\rangle, \quad m=0,1,2,...,\tfrac{1}{2}(p'-1), \tag{68}$$

$$Q_{\sqrt{q}}^{-,\pm}(I_2) := Q_{\sqrt{q}}^{+,\pm}(I_2). \tag{69}$$

On the space $\mathcal{H}_s'$, four representations $\check{Q}_{\sqrt{q}}^{\pm,\pm}$ act by the corresponding formulas (65)–(69), but now m runs through the values $0,1,2,...,\frac{1}{2}(p'-3)$.

Let now p' be even. We denote by $\mathcal{H}_r$ and $\mathcal{H}_s$, $r=\frac{1}{2}(p'+2)$ and $s=\frac{1}{2}(p'-2)$, the complex vector spaces with the bases

$$|0\rangle,|1\rangle,|2\rangle,\ldots,|\tfrac{1}{2}p'\rangle, \quad \text{and} \quad |1\rangle,|2\rangle,\ldots,|\tfrac{1}{2}(p'-2)\rangle,$$

respectively. The representations $Q_1^{1,\pm}$ and $Q_1^{2,\pm}$ act on $\mathcal{H}_r$ and $\mathcal{H}_s$, respectively, which are given by the formulas

$$Q_1^{i,\pm}(I_1)|m\rangle = \pm\frac{q^m+q^{-m}}{q-q^{-1}}|m\rangle, \quad i=1,2,$$

$$Q_1^{i,\pm}(I_2)|m\rangle = \frac{1}{q-q^{-1}}|m+1\rangle + \frac{1}{q-q^{-1}}|m-1\rangle, \quad i=1,2,$$

where $|m+1\rangle$ or $|m-1\rangle$ must be put equal to 0 if the corresponding vector does not exist.

Let $\mathcal{H}_{p'/2}$ be the complex vector space with the basis

$$|m+\tfrac{1}{2}\rangle, \quad m=0,1,2,...,\tfrac{1}{2}(p'-2).$$

Four representations $\hat{Q}^{\pm,\pm}_{\sqrt{q}}$ act on this space which are given by the formulas

$$\hat{Q}^{+,\pm}_{\sqrt{q}}(I_1)|m+\tfrac{1}{2}\rangle=\frac{q^{m+1/2}+q^{-m-1/2}}{q-q^{-1}}|m+\tfrac{1}{2}\rangle,$$

$$\hat{Q}^{+,\pm}_{\sqrt{q}}(I_2)|\tfrac{1}{2}\rangle=\pm\frac{1}{q-q^{-1}}|\tfrac{1}{2}\rangle+\frac{1}{q-q^{-1}}|\tfrac{3}{2}\rangle,$$

$$\hat{Q}^{+,\pm}_{\sqrt{q}}(I_2)|\tfrac{1}{2}(p'-2)\rangle=\pm\frac{1}{q-q^{-1}}|\tfrac{3}{2}(p'-2)\rangle+\frac{1}{q-q^{-1}}|\tfrac{1}{2}(p'-4)\rangle,$$

$$\hat{Q}^{+,\pm}_{\sqrt{q}}(I_2)|m+\tfrac{1}{2}\rangle=\frac{1}{q-q^{-1}}|m-\tfrac{1}{2}\rangle+\frac{1}{q-q^{-1}}|m+\tfrac{3}{2}\rangle, \quad m\neq\tfrac{1}{2},\tfrac{1}{2}(p'-2),$$

and by the formulas

$$\hat{Q}^{-,\pm}_{\sqrt{q}}(I_1)|m+\tfrac{1}{2}\rangle=-\frac{q^{m+1/2}+q^{-m-1/2}}{q-q^{-1}}|m+\tfrac{1}{2}\rangle,$$

$$\hat{Q}^{-,\pm}_{\sqrt{q}}(I_2)=\hat{Q}^{+,\pm}_{\sqrt{q}}(I_2).$$

Let us mention peculiarities of the representations described above. The operators $Q_1^{\pm,\pm}(I_2)$, $Q^{\pm,\pm}_{\sqrt{q}}(I_2)$, $\check{Q}^{\pm,\pm}_{\sqrt{q}}(I_2)$, and $\hat{Q}^{\pm,\pm}_{\sqrt{q}}(I_2)$ have nonzero diagonal matrix elements and nonzero traces. Moreover, the operators $\hat{Q}^{\pm,\pm}_{\sqrt{q}}(I_2)$ have two such diagonal elements. Spectra of the operators $Q_1^{\pm,\pm}(I_1)$, $Q^{\pm,\pm}_{\sqrt{q}}(I_1)$, $Q_1^{1,\pm}(I_1)$, $Q_1^{2,\pm}(I_1)$, and $\hat{Q}^{\pm,\pm}_{\sqrt{q}}(I_1)$ are not symmetric with respect to the zero point.

Proposition 12: The representations $Q_1^{\pm,\pm}$, $Q^{\pm,\pm}_{\sqrt{q}}$, $\check{Q}^{\pm,\pm}_{\sqrt{q}}$, $Q_1^{1,\pm}$, $Q_1^{2,\pm}$, and $\hat{Q}^{\pm,\pm}_{\sqrt{q}}$ *are irreducible and pairwise nonequivalent. No representation* Q'_λ *is equivalent to any of them.*

Proof is the same as that of Proposition 3.

ACKNOWLEDGMENTS

The second coauthor (AUK) was supported in part by Civilian Research and Development Foundation Grant No. UP1-309. This coauthor would like to thank the Department of Mathematics of FNSPE, Czech Technical University, for the hospitality.

[1] R. M. Santilli, Nuovo Cimento **51**, 570 (1967).
[2] R. M. Santilli, Hadronic J. **1**, 574 (1978).
[3] R. M. Santilli, Hadronic J. **4**, 1166 (1981).
[4] D. B. Fairlie, J. Phys. A **23**, L183 (1990).
[5] M. Odesski, Funct. Anal. Appl. **20(2)**, 78 (1986).
[6] Yu. S. Samoilenko and L. B. Turovska, in *Quantum Groups and Quantum Spaces*, Banach Center Publications, Vol. 40 (Banach Center, Warsaw, 1997), p. 21.
[7] O. V. Bagro and S. A. Kruglyak, ''Representations of Fairlie algebra,'' Kiev University, preprint, 1996.
[8] M. Havlíček, A. Klimyk, and E. Pelantová, Czech. J. Phys. **47**, 13 (1997).
[9] M. Noumi, Adv. Math. **123**, 11 (1996).
[10] J. Dixmier, *Algébres Enveloppantes* (Gauthier-Villars, Paris, 1974).
[11] A. Klimyk and K. Schmüdgen, *Quantum Groups and Their Representations* (Springer, Berlin, 1997).
[12] I. M. Burban and A. U. Klimyk, J. Phys. A **26**, 2139 (1993).
[13] A. M. Gavrilik and A. U. Klimyk, J. Math. Phys. **35**, 3670 (1994).

Alternative sets of hyperspherical harmonics: Satisfying cusp conditions through frame transformations

Thomas A. Heim[a)] and Dmitry Green[b)]
James Franck Institute, University of Chicago, Chicago, Illinois 60637

(Received 4 February 1998; accepted for publication 30 November 1998)

By extending the concept of Euler-angle rotations to more than three dimensions, we develop the systematics under rotations in higher-dimensional space for a novel set of hyperspherical harmonics. Applying this formalism, we determine all pairwise Coulomb interactions in a few-body system without recourse to multipole expansions. Our approach combines the advantages of relative coordinates with those of the hyperspherical description. In the present method, each Coulomb matrix element reduces to the ''$1/r$'' form familiar from the two-body problem. Consequently, our calculation accounts for all the cusps in the wave function whenever an interparticle separation vanishes. Unlike a truncated multipole expansion, the calculation presented here is exact. Following the systematic development of the procedure for an arbitrary number of particles, we demonstrate it explicitly with the simplest nontrivial example, the three-body system. © *1999 American Institute of Physics.* [S0022-2488(99)02804-2]

I. INTRODUCTION

A system consisting of N charged particles gives rise to $N(N-1)/2$ pairwise Coulomb interaction terms in its Hamiltonian. Since only the two-body problem ($N=2$) can actually be solved exactly, conventional atomic physics methods view the complete system of particles at the outset as a conglomerate of independent two-body systems, adding the interactions between these independent particles in a second step. This approach amounts to selecting a suitable subset of the $N(N-1)/2$ Coulomb terms for which a solution in terms of ''simultaneous two-body wave functions'' can be given. For instance, the independent-particle model for atomic systems treats each electron as interacting primarily with the nucleus (or with the ionic core in the case of valence electrons). Each (valence) electron's position introduces an independent spherical direction. Just as in the familiar solution of the hydrogen atom, each electron contributes to the angular part of the total system's wave function a spherical harmonic Y_{lm} of the angles specifying its direction in space. In the next step, all the electron–electron interactions are calculated by expanding the corresponding separations into Legendre polynomials of the interelectronic angles, thus yielding the familiar multipole expansion. In more general terms, the independent-particle model first selects a specific particle 1 (on physical grounds, typically the nucleus or the ionic core) and solves for each of the remaining $N-1$ particles the two-body problem $\{1,j\}$, $j=2,\ldots,N$, thereby providing a basis for expanding the total wave function. In the next step, the interaction between particles i and j is calculated by adapting the reciprocal of their separation, $1/r_{ij}$, to the coordinate system pertaining to the ''two-body'' basis functions for i and j ($i,j=2,\ldots,N$).

The Coulomb interaction is singular whenever an interparticle separation vanishes. Since these singularities are isolated from one another, they do not pose fundamental difficulties in calculating the Hamiltonian matrix. However, they give rise to *cusps* (discontinuous derivatives) in the wave function,[1] thus slowing down the convergence of partial-wave expansions of the wave function in their vicinity.[2,3] The independent-particle wave functions can only account (through

[a)]Present address: Department of Physics and Astronomy, University of Basel, CH-4056 Basel, Switzerland.
[b)]Present address: Department of Physics, Yale University, P.O. Box 208120, New Haven, CT 06520.

the corresponding *s*-wave components) for the cusps arising from vanishing separations r_{1j}, $j=2,\ldots,N$. Cusps due to vanishing r_{ij} with $i\neq 1$ are not reproduced. Possible remedies to these shortcomings include the explicit use of coordinates like the interelectronic angle θ_{ij} or even r_{ij} besides r_i and r_j; various approximations for $1/r_{ij}$;[4] or replacing r_i,r_j by $r_<=\min(r_i,r_j)$,$r_>=\max(r_i,r_j)$ in the Hamiltonian.[5] All these approaches amount to *adapting the interaction operator* $1/r_{ij}$ to a single coordinate system.

The present investigation explores the alternative approach of *adapting the basis functions* to match the relevant interparticle separations r_{ij} by utilizing several coordinate systems simultaneously. We calculate a specific interaction term $\sim 1/r_{ij}$ in the coordinate system best suited for this particular purpose. Thus, the particle separations r_{ij} dictate the choice of coordinate systems, the wave functions being transformed between the relevant reference frames to evaluate the different terms of the Hamiltonian. The success of this approach hinges on our ability to perform the numerous transformations between reference frames with high efficiency. Section II describes hyperspherical Jacobi coordinates appropriate for this task. Section III provides the main results by implementing the relevant transformations for a system with an arbitrary number of particles and by constructing basis functions (harmonics) suitable for extensive transformation between reference frames. The resulting set of hyperspherical harmonics, derived here in the context of calculating the Hamiltonian matrix for a system of *N* charged particles, has in fact much wider applicability.

Hyperspherical coordinates and corresponding hyperspherical harmonics have been applied in various areas of physics since the 1950s, for instance, in three-body scattering,[6–8] nuclear[9,10] and atomic[11–19] physics, as well as in quantum chemistry.[20,21] However, the sets of functions introduced in the present investigation are equivalent to, but much more flexible than, the hyperspherical harmonics discussed in Refs. 6–21. Beyond constituting a complete orthogonal set of functions appropriate for expansions, their frame independence affords greater flexibility in analyzing selection rules and other relations between harmonics. These aspects are conveniently investigated through ladder operators; they are determined entirely by the symmetry properties of the *N*-particle system, independent of any coordinate representation. Only calculating the nonvanishing matrix elements requires an appropriate coordinate representation of the generic basis functions. Section IV illustrates the relevant procedures with the simplest nontrivial example, the three-body problem. The concluding Sec. V discusses advantages and limitations of the new technique, as compared to the conventional multipole expansion.

II. HYPERSPHERICAL JACOBI COORDINATES

The inverse proportionality between pairwise interaction and interparticle separation suggests replacing individual particles' positions with *relative* (Jacobi) coordinates. Starting from one pair of particles, the construction of Jacobi coordinates proceeds hierarchically by joining (the centers of mass of) increasingly complex groups of particles. This hierarchical structure is commonly referred to as a ''Jacobi tree.''[10] Alternative choices for the initial particle pair, as well as the ordering of successive particle groups, correspond to different Jacobi trees. In the next step, mass scaling of the Jacobi coordinates fully exposes the symmetry of the kinetic energy operator for the multiparticle system. Let p,q denote two (groups of) particles with masses M_p,M_q and center of mass positions $\boldsymbol{r}_p$ and $\boldsymbol{r}_q$, respectively. Appropriate mass scaling of the relative coordinate in the form[8] $\boldsymbol{\xi}_{p,q}=\{M_pM_q/(M_p+M_q)\}^{1/2}(\boldsymbol{r}_p-\boldsymbol{r}_q)$ removes the individual mass dependence from the expression for the kinetic energy:

$$-\sum_{i=1}^{N}\frac{\hbar^2}{2M_i}\Delta_{\boldsymbol{r}_i}=-\frac{\hbar^2}{2}\sum_{k=1}^{N-1}\Delta_{\boldsymbol{\xi}_k}-\frac{\hbar^2}{2M_{\text{tot}}}\Delta_{\boldsymbol{r}_{\text{c.m.}}}. \tag{2.1}$$

Setting the origin at the center of mass of the whole system allows discarding the c.m.'s position and motion. With the individual mass factors removed, the kinetic energy (generalized Laplacian) displays complete symmetry under rotations in $(3N-3)$-dimensional space.

Hyperspherical coordinates exploit this symmetry by separating the ''shape'' of the system (described by $3N-4$ angular coordinates) from its overall ''size'' $R^2=\Sigma_k \boldsymbol{\xi}_k^2$ (with the dimension of a moment of inertia). Eigenfunctions of the Laplacian's angular part, termed ''hyperspherical harmonics'' by generalizing the two- and three-dimensional cases, constitute a basis for expanding the complete wave function. In higher-dimensional spaces, the angular Laplacian's eigenvalues $\Lambda(\Lambda+3N-5)$ are highly degenerate. In addition to the ''grand angular momentum'' Λ,[8] the hyperspherical harmonics thus require numerous labels—analogs of the single ''magnetic'' quantum number m for three-dimensional spherical harmonics—for their unique specification. Different sets of hyperspherical harmonics, resulting from solving the Laplacian eigenvalue problem by separation of variables in alternative sets of coordinates, have been investigated extensively (see, e.g., Refs. 10,17, and 21 for systematic studies). However, their construction through separation of variables inevitably ties these harmonics to a specific coordinate system. They are thus not well suited for extensive transformation between different reference frames. The definition of harmonics based only on their behavior under the relevant transformations is the main result of this paper, to be derived in Sec. III.

The construction of each Jacobi tree starts with a pair of particles. Thus, each Jacobi tree contains at least one Jacobi vector joining two particles only (rather than centers of mass of particle *groups*). We denote such a vector as a ''primary'' coordinate. In any one Jacobi tree, up to $\lfloor N/2 \rfloor$ Jacobi vectors are directly proportional to actual interparticle separations, $\boldsymbol{r}_{ij}$, the remaining relative coordinates necessarily involving larger complexes of particles. (Here and in the following, $\lfloor x \rfloor$ denotes the largest integer not exceeding x.) We will calculate each of the $N(N-1)/2$ Coulomb terms of the Hamiltonian in a Jacobi tree where it occurs as a ''primary'' coordinate. Obviously, all these terms then take the same form as the simple one-electron integral over $1/r$ in hydrogen, but the variables now refer to a multitude of different reference frames. Having thus eliminated all nested integrations, our next task consists in determining the transformations between different Jacobi trees.

The most general transformation from one Jacobi tree to another (for the same system of particles) resolves into a sequence of elementary operations. Each elementary step consists in ''transplanting'' a subcomplex q from some particle complex $\{pq\}$ to a complex $\{qr\}$;[10] it is achieved by a two-dimensional kinematic rotation[8] through the angle

$$\phi=\tan^{-1}\sqrt{M_q(M_p+M_q+M_r)/M_pM_r} \tag{2.2}$$

in the $(\boldsymbol{\xi}_{p,q},\boldsymbol{\xi}_{pq,r})$-plane of the $(N-1)$-dimensional space of mass-scaled Jacobi vectors:

$$\boldsymbol{\xi}'_{q,r}=\cos\phi\,\boldsymbol{\xi}_{p,q}-\sin\phi\,\boldsymbol{\xi}_{pq,r}, \tag{2.3a}$$

$$\boldsymbol{\xi}'_{p,qr}=\sin\phi\,\boldsymbol{\xi}_{p,q}+\cos\phi\,\boldsymbol{\xi}_{pq,r}. \tag{2.3b}$$

As each Jacobi vector is a vector in three-dimensional physical space, the basic rotation in the $(\boldsymbol{\xi}_j,\boldsymbol{\xi}_k)$ plane implies three rotations through the same angle ϕ in the (x_j,x_k), the (y_j,y_k), and the (z_j,z_k) planes, where (x,y,z) denote the Cartesian components of $\boldsymbol{\xi}$. An elementary kinematic rotation through a finite angle ϕ in the $(\boldsymbol{\xi}_j,\boldsymbol{\xi}_k)$ plane then reads[22]

$$T_{\boldsymbol{\xi}_j,\boldsymbol{\xi}_k}(\phi)=\exp(i\phi J_{x_jx_k})\exp(i\phi J_{y_jy_k})\exp(i\phi J_{z_jz_k}), \tag{2.4}$$

with infinitesimal rotation operators

$$J_{uv}\equiv -i\left(u\frac{\partial}{\partial v}-v\frac{\partial}{\partial u}\right),\quad (u,v)=(x_j,x_k),(y_j,y_k),(z_j,z_k). \tag{2.5}$$

[For a heuristic explanation of (2.4), recall the Taylor expansion of $f(x+\Delta x)$ which may formally be written as $\exp(i\Delta xp)f(x)$ with the infinitesimal ''translation operator'' $p\equiv -i(\partial/\partial x)$ corresponding to the quantum mechanical linear momentum (with $\hbar\equiv 1$). In the present context, the translation from x to $x+\Delta x$ is replaced by a rotation through an angle ϕ from some direction in multidimensional space to a new direction.] In order to determine T_{ξ_j,ξ_k}'s effect on the basis functions, we need to (i) express the generic rotation operators J_{uv} in terms of ''ladder operators'' whose action on the basis functions is straightforward to calculate, and (ii) construct basis functions suitable for such transformations.

This section concludes with the outline of a strategy for determining the most efficient sequence of two-dimensional rotations to achieve the transformation between arbitrary Jacobi trees. Note first that the individual terms in the Hamiltonian are referred to by means of *particle labels*, implying the need for appropriate antisymmetrization of the total wave function under interchange of identical particles. This consideration dominates the construction of the first Jacobi tree to which all subsequent transformations are applied. In an atomic or molecular system, antisymmetrization concerns primarily the electronic part of the wave function. Thus the basic Jacobi tree starts out with a pair of electrons. It grows by adding one electron at a time, finishing with the addition of nuclei or ionic core(s). This procedure results in a ''canonical'' Jacobi tree,[10] represented by the sequence of particle labels

$$(\cdots((((12)3)4)5)\ \ldots), \tag{2.6}$$

whose parentheses separate subcomplexes. Alternatively, one could start by first forming as many *pairs* of particles as possible, i.e., $\lfloor N/2 \rfloor$ pairs, before joining these pairs into larger complexes. Although the antisymmetrization of an electron pair is particularly compact (namely, their relative angular momentum and their spin must add to an *even* value), antisymmetrization among different pairs requires breaking up all these pairs, in essence going back to a canonical tree. Since breaking up subcomplexes involves about as many elementary rotations as forming the new complexes, this procedure is not efficient. Nevertheless, a set of N (for odd particle number) or $N-1$ (even particle number) trees with the maximum number of $\lfloor N/2 \rfloor$ pairs suffices to isolate all the $N(N-1)/2$ interparticle separations. A successful strategy therefore aims at building these particular Jacobi trees from the canonical tree. The simplest of these ''pair trees,'' $(\cdots((12)(34))\times(56)\ \ldots)$, is obtained from the canonical tree with only $\lfloor N/2 \rfloor-1$ rotations. In general, the pair (jk) with $j<k$ is $(k-j-\delta_{j,1})$ rotations away from the canonical tree, but many other ''useful'' pairs are formed in the course of these ''transplantations.''

III. TRANSFORMING THE BASIS FUNCTIONS

This section provides the tools required to rotate harmonics from one coordinate system to another. The reader familiar with Lie algebra will of course recognize the relevant aspects of the algebra of rotations in $(3N-3)$ dimensional space, so $(3N-3)$. However, such familiarity is not presumed, and we hardly use the terminology of group theory. The presentation reflects a more pragmatic point of view, emphasizing both the technical implementation as well as its relation to the physical application at hand, rather than full mathematical generality, let alone mathematical rigor. For the latter aspects, the reader should turn to the mathematical literature.[23]

A. Rotations in *d*-dimensional space

The most intuitive description of a rotation in three dimensions requires two elements: the (invariant) *axis*, and the *angle* of rotation. The rotation itself occurs in a *plane* perpendicular to the axis of rotation. In three dimensions, specifying the direction of the axis of rotation is equivalent to, but more economical than, describing the actual plane of rotation. For a higher dimensional space this is no longer the case, because there are several invariant directions perpendicular to a given plane. Thus, in *d*-dimensional space, a basic rotation is more appropriately characterized as occurring in the *plane* spanned by two coordinates rather than by an invariant axis orthogonal to it. The generic infinitesimal rotation operators then take the form (2.5). The number of different planes, $\frac{1}{2}d(d-1)$ in d dimensions, equals the number of basic rotations. Because rotations occurring in nonintersecting planes affect different pairs of coordinates, they are evidently independent of each other; the corresponding rotation operators commute with one another. As there are $\lfloor d/2 \rfloor$

nonintersecting planes in a d-dimensional space, the largest set of simultaneously commuting operators J_{uv} contains $\ell \equiv \lfloor d/2 \rfloor$ elements. These commuting operators are commonly denoted as H_j, $j=1,\ldots,\ell$.[23] For our application, the power of the Lie algebraic method derives from the structure and the properties of rotations being completely independent of any particular coordinate representation of the operators H_j. In fact, the following manipulations are most efficiently carried out in generic Cartesian coordinates without any reference to a specific Jacobi tree.

Simultaneous eigenfunctions of the H_j with (integer) eigenvalues m_j, $j=1,\ldots,\ell$, provide a basis for building hyperspherical harmonics suitable for extensive rotations between reference frames. If the infinitesimal rotation corresponding to H_j occurs in the (u_j,v_j) plane, the simultaneous eigenfunction of the ℓH_j with eigenvalues m_j, respectively, reads

$$F(\ldots)\times\prod_{j=1}^{\ell}(u_j+iv_j)^{m_j} \tag{3.1}$$

with a function F that bears further specification in Sec. III C. At this point, we only require that $H_j F\equiv 0$ for all j. The set of eigenvalues $\boldsymbol{\mu}=\{m_1,m_2,\ldots,m_\ell\}$ serves as a label identifying different harmonics with the same value of $\Lambda(\Lambda+d-2)$ in the angular Laplacian's eigenvalue equation. Additional labels required for a unique specification will be introduced in Sec. III C.

Appropriate linear combinations of the remaining infinitesimal rotation operators act as *ladder operators* $E_{\boldsymbol{\alpha}}$ satisfying

$$[H_j,E_\alpha]=\alpha_j E_\alpha, \quad j=1,\ldots,\ell, \tag{3.2}$$

with an ℓ-dimensional vector index $\boldsymbol{\alpha}$ with components $\alpha_j=0,\pm 1$. Here $\alpha_j=0$ indicates that $E_{\boldsymbol{\alpha}}$ does not change the part of the eigenfunction pertaining to H_j, whereas $\alpha_j=\pm 1$ means that it maps this part to the eigenfunction with eigenvalue $m_j\pm 1$. The ladder operators interrelate eigenfunctions with different $\boldsymbol{\mu}$ but degenerate Λ. Raising and lowering operators form Hermitian conjugate pairs $E_{-\boldsymbol{\alpha}}=E_{\boldsymbol{\alpha}}^\dagger$.

In general (for $\ell>1$), each ladder operator affects two of the m_j simultaneously; for odd-dimensional spaces a subset of ℓ pairs of raising and lowering operators change one of the m_j only.[23] Abbreviating the vector label $\boldsymbol{\alpha}$ by giving only its two nonvanishing components α_j and α_k, the ladder operator that raises m_j and simultaneously lowers m_k takes the form

$$E_{jk}^{+-}=-\frac{i}{2}\left((u_j+i\,v_j)\frac{\partial}{\partial(u_k+iv_k)}-(u_k-iv_k)\frac{\partial}{\partial(u_j-iv_j)}\right), \tag{3.3a}$$

whereas the operator raising both m_j and m_k is represented by

$$E_{jk}^{++}=-\frac{i}{2}\left((u_j+iv_j)\frac{\partial}{\partial(u_k-iv_k)}-(u_k+iv_k)\frac{\partial}{\partial(u_j-iv_j)}\right). \tag{3.3b}$$

[Straightforward application of these operators to eigenfunctions (3.1) verifies their behaving as ladder operators: they contribute to, or remove from, (3.1) factors $(u_j\pm iv_j)$ and $(u_k\pm iv_k)$ as appropriate for the intended ''ladder operator action.''] For odd-dimensional spaces, a residual coordinate, denoted here by w, does not occur in any of the H_j. The ladder operators changing only m_j (rather than a pair m_j,m_k) read then

$$E_j^{\pm}=-i\left((u_j\pm iv_j)\frac{\partial}{\partial w}-w\frac{\partial}{\partial(u_j\mp iv_j)}\right), \quad j=1,\ldots,\ell. \tag{3.3c}$$

[For $d=3$, setting $(u,v,w)=(x,y,z)$ and transforming the derivatives reveals the familiar pair of ladder operators $l_x\pm i\,l_y$, the single H_j occurring in this case coinciding with l_z.] This symbolic representation allows for efficient implementation on the computer.

A subset of ℓ ladder operators $E_{\boldsymbol{\epsilon}_j}$ (and their Hermitian conjugates) suffices to interrelate all harmonics with the same eigenvalue Λ. A convenient choice[23] for these $E_{\boldsymbol{\epsilon}_j}$ is given by the ladder operators that raise m_j and simultaneously lower m_{j+1} for $j=1,\ldots,\ell-1$. $E_{\boldsymbol{\epsilon}_\ell}$'s form depends on whether d is even or odd. For d even, $E_{\boldsymbol{\epsilon}_\ell}$ raises both $m_{\ell-1}$ and m_ℓ, and for odd d, it raises m_ℓ only, without changing any of the other m_j. The set $\{\boldsymbol{\epsilon}_j\}$, $j=1,\ldots,\ell$, then provides a basis for the ℓ-dimensional space of the vector labels $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$. In an ℓ-component vector notation, this basis reads (we assume $\ell \geqslant 5$ in order to expose the generic structure clearly)

$$\begin{aligned}
\boldsymbol{\epsilon}_1&=(1,-1,0,0,\ldots,0),\\
\boldsymbol{\epsilon}_2&=(0,1,-1,0,\ldots,0),\\
\boldsymbol{\epsilon}_3&=(0,0,1,-1,\ldots,0),\\
&\vdots\\
\boldsymbol{\epsilon}_{\ell-1}&=(0,0,\ldots,0,1,-1),\\
\boldsymbol{\epsilon}_\ell&=\begin{cases}(0,0,\ldots,0,1,1) & \text{for } d=2\ell\\ (0,0,\ldots,0,0,1) & \text{for } d=2\ell+1.\end{cases}
\end{aligned} \tag{3.4}$$

In general these basis vectors are *not* orthogonal in ℓ-dimensional space. Note, however, the following special cases: (i) In four-dimensional space (with $\ell=2$), the two basis vectors $\boldsymbol{\epsilon}_1=(1,-1)$ and $\boldsymbol{\epsilon}_2=(1,1)$ *are orthogonal*, indicating that the two basic ladder operator pairs, $E_{\pm\boldsymbol{\epsilon}_1}$ and $E_{\pm\boldsymbol{\epsilon}_2}$, commute with each other. In group theoretical language, this feature reflects the direct product structure $SO(4)=SO(3)\otimes SO(3)$. However, the two $SO(3)$ components do not refer to m_1 and m_2 directly, but rather to $(m_1-m_2)/2$ and $(m_1+m_2)/2$. (ii) For rotations in three-dimensional space, we have $\ell=1$, thus only one quantum number m which is being changed by one pair of ladder operators $E_{\pm\boldsymbol{\epsilon}_1}\equiv l_\pm$. (iii) In $d=2$ dimensions, *there are no ladder operators*. Since all rotations occur in the same plane (the ''only'' plane of two-dimensional space), the order in which rotations through different angles are performed does not matter; they are all independent of each other. In the present formulation, the (single) rotation operator H_1 generates all rotations. Each rotation is associated with its own harmonic function, $\exp(i\phi)$, the phase functions for different rotations (i.e., for different rotation angles ϕ) not being related to one another through linear operators.

B. Rotation of hyperspherical harmonics

We now turn to the analysis of the transformation described by (2.4). Note first that the three rotations occur in three nonintersecting planes, affording the more suitable representation $\exp(i\phi[J_{x_jx_k}+J_{y_jy_k}])\exp(i\phi J_{z_jz_k})$. Arranging the Cartesian components of the $N-1$ mass-scaled Jacobi vectors $\boldsymbol{\xi}_k$ in the form

$$\{x_1,y_1,x_2,y_2,\ldots,x_{N-1},y_{N-1},z_1,z_2,\ldots,z_{N-1}\}, \tag{3.5}$$

we choose the H_j by selecting pairs of coordinates from this list, starting from the left:

$$H_j=J_{x_jy_j},\quad j=1,\ldots,N-1, \tag{3.6a}$$

$$H_{N-1+k}=J_{z_{2k-1}z_{2k}},\quad k=1,\ldots,\left\lfloor\frac{N-1}{2}\right\rfloor, \tag{3.6b}$$

in the notation of (2.5). We have stressed repeatedly our aim of defining hyperspherical harmonics entirely through their behavior when acted upon by the operators H_j and E_{ϵ_j}, because these operators embody the structure and symmetry of the whole system in a frame-independent way. The explicit coordinate representation (3.6a)–(3.6b) of the relevant operators appears at variance with our intended frame independence. However, the spatial coordinates, (x_j,y_j,z_j), $j=1,\ldots,N-1$, in (3.6a)–(3.6b) should be viewed as *generic Cartesian coordinates*; they are *not* tied to any particular Jacobi tree. Arranging the generic Cartesian coordinates as we did in (3.5), on the other hand, does reflect physical considerations beyond the purely mathematical structure of rotations in higher-dimensional space: The latter would instead label the coordinates most appropriately as $(X_1,X_{-1},X_2,X_{-2},\ldots,X_\ell,X_{-\ell})$, supplemented possibly with an X_0 for odd-dimensional spaces. Our arrangement takes into account that the dimension $(3N-3)$ arises from a *product structure* of an $(N-1)$-dimensional particle space with the three-dimensional physical space of each Jacobi vector. In particular, the arrangement (3.5) affords attributing *relevant physical meaning* to the first $(N-1)$ eigenvalues m_j, $(j=1,\ldots,N-1)$: They represent the z projections of physical angular momenta; their sum constitutes the z projection L_z of the total orbital angular momentum, an invariant of the system.

As is evident from (3.6a), the coordinates (x_j,x_k,y_j,y_k)occur in H_j and H_k. The x and y parts of the rotation (2.4) thus involve the ladder operators that change m_j and m_k only. A more detailed analysis shows that (3.3a–c) can be inverted to read:

$$J_{x_jx_k}+J_{y_jy_k}=E_{jk}^{+-}+E_{jk}^{-+}\equiv E_{\boldsymbol{\alpha}(jk)}+E_{-\boldsymbol{\alpha}(jk)}, \tag{3.7}$$

for $1\leq j<k\leq N-1$, with $\boldsymbol{\alpha}(jk)=\sum_{s=j}^{k-1}\boldsymbol{\epsilon}_s$, i.e., the ℓ-component vector with $+1$ as its jth component and -1 as its kth component, all other entries being 0.

The following consideration is central to our development, providing the crucial link between the ladder operator representation (3.7) for $(J_{x_jx_k}+J_{y_jy_k})$ and their transformation matrix elements, by means of Euler-angle rotations. In analogy to the relation $l_x=\frac{1}{2}(l_++l_-)$ familiar from rotations in three dimensions, we view the sum of a ladder operator $E_{\boldsymbol{\alpha}(jk)}$ and its inverse (Hermitian conjugate) $E_{-\boldsymbol{\alpha}(jk)}$ as describing a rotation about an analog of the x axis. In three dimensions, an arbitrary rotation conventionally resolves into a sequence of three rotations: about the z axis, about the resulting y' axis, and about the new z' axis, through the Euler angles (φ,θ,ψ), respectively.[22] In terms of these three Euler-angle rotations, a rotation about the x axis through an angle ϕ results from the following sequence of operations: The first Euler-angle rotation about the z axis through the angle $\varphi=-\pi/2$ rotates the y axis onto the original x axis; in the second step one rotates about this y' axis (which is the original x axis) through the angle $\theta=\phi$; the third Euler-angle rotation finally moves the x' axis back to the original x direction by rotating about the z' axis (lying in the original (yz) plane) through the angle $\psi=\pi/2$. Wigner's d-symbol is the matrix element for the rotation of a spherical harmonic about the y axis, $d_{m'm}^{(l)}(\phi)=\langle Y_{lm'}|\exp(i\phi l_y)|Y_{lm}\rangle$, whereas the two z-type rotations only contribute phase factors $\exp(im'\pi/2-im\pi/2)$,[22] giving for the rotation of a spherical harmonic about the x axis

$$\langle Y_{lm'}|\exp(i\phi l_x)|Y_{lm}\rangle=e^{im'(\pi/2)d_{m'm}^{(l)}(\phi)e^{-im(\pi/2)}}. \tag{3.8}$$

Generalization of the concept of Euler-angle rotations from three to more dimensions hinges on the following key observations: m and m' serve to distinguish between degenerate harmonics with the same l. In higher-dimensional spaces, the vector indices $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ play the same role. However, the parameter l in the d-symbol indicates not only the angular Laplacian's eigenvalue $l(l+1)$, but also—more importantly—the range of possible m values, $-l\leq(m,m')\leq l$. More precisely, it sets the upper limit of $2l$ for the number of times either one of the two ladder operators l_+,l_- can act in direct succession before necessarily mapping any spherical harmonic Y_{lm} to zero. Viewed in this way, the harmonic's parameter m indicates its ''position'' along the string (of length $2l+1$) of degenerate harmonics interrelated by a ladder operator. We now extend these concepts to rotations in more than three dimensions.

The single m (l_z's eigenvalue) in three dimensions belongs to a string ranging from $-l$ to $+l$, accessed by the single pair of ladder operators l_+, l_-. In more than three dimensions, we replace it with an ℓ-component vector $\boldsymbol{\mu}$ whose components are changed (typically in pairs) along different strings labeled by corresponding ladder operator pairs $E_{\boldsymbol{\alpha}}$ and $E_{-\boldsymbol{\alpha}}$. (Recall, however, that the ℓ vectors $\boldsymbol{\epsilon}_j$ provide a basis for the vectors $\boldsymbol{\alpha}$; thus any ladder operator $E_{\boldsymbol{\alpha}}$ can be expressed as products of $E_{\boldsymbol{\epsilon}_j}$.) The role of m as an indicator of the harmonic's position along the single string in three dimensions extends therefore to higher dimensions if we project the ''indicator'' $\boldsymbol{\mu}$ onto the ''direction'' of any particular string $\boldsymbol{\alpha}$ of interest, i.e., if we define

$$m_{\boldsymbol{\alpha}}=\frac{\boldsymbol{\mu}\cdot\boldsymbol{\alpha}}{\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}}, \quad m'_{\boldsymbol{\alpha}}=\frac{\boldsymbol{\mu}'\cdot\boldsymbol{\alpha}}{\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}}, \tag{3.9}$$

where the scalar product in the denominators accounts for $\boldsymbol{\alpha}$'s *two* nonzero components in the case of a pairwise change of m_j's. In analogy to l in the three-dimensional case, a parameter $\lambda_{\boldsymbol{\alpha}}$ determines how many times the ladder operators $E_{\pm\boldsymbol{\alpha}}$ can be applied in direct succession. The modulus of $m_{\boldsymbol{\alpha}}$ in (3.9) provides a lower limit for the relevant ''string length'' $\lambda_{\boldsymbol{\alpha}}$. Since the $\boldsymbol{\epsilon}_j$ form a basis for the $\boldsymbol{\alpha}$, we anticipate that $\lambda_{\boldsymbol{\alpha}}$ emerges from an ℓ-component vector $\boldsymbol{\lambda}=(\lambda_{\boldsymbol{\epsilon}_1},\ldots,\lambda_{\boldsymbol{\epsilon}_\ell})$. The latter provides the additional parameters required for a unique specification of harmonics (besides $\boldsymbol{\mu}$ and the eigenvalue Λ) . Note that in the general case $\lambda_{\boldsymbol{\alpha}}$ is not identical with the eigenvalue Λ, at variance with the situation in three dimensions.

To summarize the procedure so far, we note that the x and y parts of the finite rotation (2.4) affect only the two components m_j and m_k of the vector $\boldsymbol{\mu}$, shifting m_j by an integer amount n while simultaneously changing m_k by the same amount in the opposite direction. The projections (3.9) evaluate to

$$m_{\boldsymbol{\alpha}(jk)}=\tfrac{1}{2}(m_j-m_k), \tag{3.10a}$$

$$m'_{\boldsymbol{\alpha}(jk)}=\tfrac{1}{2}(m'_j-m'_k)=m_{\boldsymbol{\alpha}(jk)}+n, \tag{3.10b}$$

with nonvanishing matrix elements for the rotation of a hyperspherical harmonic $Y_{\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda}}$ into another harmonic $Y_{\Lambda',\boldsymbol{\mu}',\boldsymbol{\lambda}'}$ occurring only if $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ lie along the *same* $\boldsymbol{\alpha}$-string of harmonics with degenerate Λ, i.e., if an integral number n of ladder operator steps $E_{\pm\boldsymbol{\alpha}(jk)}$ separates $\boldsymbol{\mu}'$ from $\boldsymbol{\mu}$. With our previous considerations on Euler-angle rotations leading to (3.8), this matrix element reads

$$\langle Y_{\Lambda',\boldsymbol{\mu}',\boldsymbol{\lambda}'}|\exp(i\phi[J_{x_jx_k}+J_{y_jy_k}])|Y_{\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda}}\rangle=e^{i(m'_{\boldsymbol{\alpha}}-m_{\boldsymbol{\alpha}})\pi/2}\,d^{(\lambda_{\boldsymbol{\alpha}})}_{m'_{\boldsymbol{\alpha}}m_{\boldsymbol{\alpha}}}(2\phi)\,\delta_{\Lambda',\Lambda}\,\delta_{\boldsymbol{\lambda}',\boldsymbol{\lambda}}\,\delta_{\boldsymbol{\mu}',\boldsymbol{\mu}+n\boldsymbol{\alpha}}, \tag{3.11}$$

where we omitted $\boldsymbol{\alpha}$'s parameters j,k for brevity. The Kronecker symbols δ ensure that only harmonics with the same eigenvalue Λ, lying along the same $\boldsymbol{\alpha}$-string, are connected. The factor 2 multiplying the rotation angle ϕ reflects the absence of the factor 1/2 on the right-hand side of (3.7), as compared to the expression for $l_x=\frac{1}{2}(l_++l_-)$, thereby effectively doubling the rotation angle in (3.8).

It remains to apply the same concepts to the z part of the rotation (2.4). This part affects only $\boldsymbol{\mu}$'s components that correspond to the H_κ involving z_j and z_k. Because we chose to gather all the z coordinates after all (xy) pairs in (3.5), the Jacobi vector indices j,k are shifted relative to $\boldsymbol{\mu}$'s components for the z coordinates. To simplify the notation, we use

$$\iota\equiv N-1+\left\lfloor\frac{j+1}{2}\right\rfloor;\quad \kappa\equiv N-1+\left\lfloor\frac{k+1}{2}\right\rfloor, \tag{3.12}$$

so that H_ι and m_ι now refer to the operator and quantum number pertaining to the coordinate z_j, with the same connection between H_κ, m_κ, and z_k. Depending on the indices j and k, i.e., depending on the positions of the coordinates z_j and z_k in the sequence (3.5), $J_{z_j z_k}$ takes alternative forms in terms of ladder operators or H_κ. The possible cases are:

(1) k even and $j=k-1$. According to (3.6b), $J_{z_j z_k}$ coincides with $H_{N-1+k/2}\equiv H_\kappa$. The rotation matrix element is simply

$$\langle Y_{\Lambda',\boldsymbol{\mu}',\boldsymbol{\lambda}'}|\exp(i\phi J_{z_j z_k})|Y_{\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda}}\rangle=\exp(im_\kappa\phi)\delta_{\Lambda',\Lambda}\delta_{\boldsymbol{\lambda}',\boldsymbol{\lambda}}\delta_{\boldsymbol{\mu}',\boldsymbol{\mu}}. \tag{3.13}$$

(2) k even and $j<k-1$ even. We find

$$J_{z_j z_k}=\tfrac{1}{4}(E_{\iota\kappa}^{+-}+E_{\iota\kappa}^{-+}+E_{\iota\kappa}^{++}+E_{\iota\kappa}^{--}). \tag{3.14}$$

Viewing again the sum of a ladder operator and its inverse (Hermitian conjugate) as proportional to an analog of a rotation about the x axis, we readily recognize the last expression as the sum of *two* x-type rotations [compare with (3.7)'s central expression]. With the explicit formulas (3.3a)–(3.3b), it is straightforward to verify that the ladder operator $E_{\iota\kappa}^{+-}$ raising m_ι and lowering m_κ commutes with the ''raising–raising'' operator $E_{\iota\kappa}^{++}$. The two x-type rotations are therefore independent of each other. As in the previous discussion of the x and y parts, an integer number of steps $E_{\iota\kappa}^{+-}$ must connect (m_ι,m_κ) to (m'_ι,m'_κ), i.e., $m'_\iota=m_\iota+n_1$, $m'_\kappa=m_\kappa-n_1$. However, the second rotation involves the same components of $\boldsymbol{\mu}$, only this time changing both in the *same* direction: $m'_\iota=m_\iota+n_2$, $m'_\kappa=m_\kappa+n_2$. Compatibility of the two conditions requires $n_1=n_2=0$, thus

$$m_{\boldsymbol{\beta}_-(\iota\kappa)}=\tfrac{1}{2}(m_\iota-m_\kappa)=m'_{\boldsymbol{\beta}_-(\iota\kappa)}, \tag{3.15a}$$

$$m_{\boldsymbol{\beta}_+(\iota\kappa)}=\tfrac{1}{2}(m_\iota+m_\kappa)=m'_{\boldsymbol{\beta}_+(\iota\kappa)}, \tag{3.15b}$$

and the extra phase factors occurring in (3.11) drop out in this case:

$$\langle Y_{\Lambda',\boldsymbol{\mu}',\boldsymbol{\lambda}'}|\exp(i\phi J_{z_j z_k})|Y_{\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda}}\rangle=d^{(\lambda_{\boldsymbol{\beta}_-})}_{m_{\boldsymbol{\beta}_-},m_{\boldsymbol{\beta}_-}}(\phi/2)\,d^{(\lambda_{\boldsymbol{\beta}_+})}_{m_{\boldsymbol{\beta}_+},m_{\boldsymbol{\beta}_+}}(\phi/2)\;\delta_{\Lambda',\Lambda}\delta_{\boldsymbol{\lambda}',\boldsymbol{\lambda}}\delta_{\boldsymbol{\mu}',\boldsymbol{\mu}}. \tag{3.16}$$

The factor 1/2 multiplying the rotation angle stems of course from the factor 1/4 in (3.14).

(3) k even and $j<k-1$ odd. In this case,

$$J_{z_j z_k}=\frac{1}{4i}(E_{\iota\kappa}^{+-}-E_{\iota\kappa}^{-+}+E_{\iota\kappa}^{++}-E_{\iota\kappa}^{--}), \tag{3.17}$$

obviously the sum of two y-type rotations. The ''selection rules'' for m_ι and m_κ are the same as in the previous case, and since the extra phase factors distinguishing an x from a y-type rotation cancel in (3.16), the *matrix element is identical to the one in case (2).*

(4) $k<N-1$ odd and j even. $J_{z_j z_k}$ turns into

$$J_{z_j z_k}=\frac{1}{4i}(E_{\iota\kappa}^{+-}-E_{\iota\kappa}^{-+}-E_{\iota\kappa}^{++}+E_{\iota\kappa}^{--}). \tag{3.18}$$

Exactly the same considerations apply again, except that the *difference* of two y-type rotations corresponds to a rotation through the angle $-\phi/2$ for the $E_{\iota\kappa}^{++}$ part.

(5) $k<N-1$ odd and j even, yielding

$$J_{z_j z_k}=\tfrac{1}{4}(E_{\iota\kappa}^{+-}+E_{\iota\kappa}^{-+}-E_{\iota\kappa}^{++}-E_{\iota\kappa}^{--}), \tag{3.19}$$

i.e., the difference of two x-type rotations. The matrix element coincides with the one of case (4).

(6) $k=N-1$ odd and j even. z_k does not occur in any of the H_κ, and the relevant ladder operators act on m_ι only:

$$J_{z_j z_k}=\tfrac{1}{2}(E_\iota^{+}+E_\iota^{-}). \tag{3.20}$$

This is an x-type rotation, exactly as in three dimensions. Projecting $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ onto the appropriate direction $\boldsymbol{\gamma}(\iota)$: $m_{\boldsymbol{\gamma}(\iota)}=m_\iota$, $m'_{\boldsymbol{\gamma}(\iota)}=m'_\iota=m_{\boldsymbol{\gamma}(\iota)}+\nu$, we obtain for the matrix element

$$\langle Y_{\Lambda',\boldsymbol{\mu}',\boldsymbol{\lambda}'}|\exp(i\phi J_{z_jz_k})|Y_{\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda}}\rangle = d^{(\lambda_\gamma)}_{m'_\gamma,m_\gamma}(\phi)e^{i(m'_\gamma - m_\gamma)\pi/2}\delta_{\Lambda',\Lambda}\delta_{\boldsymbol{\lambda}',\boldsymbol{\lambda}}\delta_{\boldsymbol{\mu}',\boldsymbol{\mu}+\nu\boldsymbol{\gamma}}. \quad (3.21)$$

(7) $k=N-1$ odd and j odd.

$$J_{z_jz_k} = \frac{1}{2i}(E^+_\iota - E^-_\iota), \quad (3.22)$$

a y-type rotation. Its matrix element differs from the one in case (6) only by the absence of the phase factor $\exp(i\nu\pi/2)$.

Thus, the z part of the rotation (2.4) takes essentially three different forms: (i) a phase factor $\exp(im_\kappa\phi)$, diagonal in the $\boldsymbol{\mu}$, if $J_{z_jz_k}$ coincides with H_κ; (ii) an x or y-type rotation through the angle ϕ, changing m_ι into $m'_\iota = m_\iota + \nu$, as in three dimensions, if z_k is the unpaired coordinate z_{N-1} of an odd-dimensional space; (iii) the product of two rotations through the angles $\phi/2$ and $(-1)^k\phi/2$, respectively, but diagonal in $\boldsymbol{\mu}$, if z_j and z_k belong to different H_ι and H_κ.

C. Harmonics suitable for rotation

In Sec. III B, we have completely determined the matrix elements for the rotations of our interest when applied to hyperspherical harmonics, before actually specifying these functions explicitly. This was possible because we expressed the matrix elements in terms of the labels (''quantum numbers'') $(\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda})$ identifying the harmonics, rather than through integrals in coordinate space. So far, the harmonics are functions of the generic Cartesian coordinates (x_j,y_j,z_j). Upon rotation from one Jacobi tree to another, these coordinates transform into another set (x'_j,y'_j,z'_j), $j=1,\ldots,N-1$, but the harmonics, when expressed in terms of the new coordinates, retain their functional form. In this sense, these hyperspherical harmonics are frame independent.

Direct solution of the angular Laplacian's eigenvalue problem by separation of the $(3N-4)$ coordinates in the second-order differential equation leads to hyperspherical harmonics represented by standard spherical harmonics and Jacobi or Gegenbauer polynomials, depending on the choice of hyperspherical coordinates.[10,17,21] However, these harmonics are not simultaneous eigenfunctions of all the H_j. Consequently, they are not suitable for our purpose, because rotating the harmonics requires knowledge of the ladder operators' effects on these functions, at least for the base set of ladder operators $E_{\pm\boldsymbol{\epsilon}_j}$, $j=1,\ldots,\ell$, which in turn requires uniquely specifying the harmonics with labels $(\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda})$. We now construct complete sets of functions defined exclusively by their behavior under the action of the first-order differential operators H_j and $E_{\pm\boldsymbol{\epsilon}_j}$, $j=1,\ldots,\ell$. The resulting functions are ''hyperspherical harmonics,'' too, because they satisfy the generalized Laplacian's symmetry under rotations.

The hyperspherical description separates the ''hyperradial'' momentum from the generalized angular momentum. Each of the Cartesian coordinates x_j,y_j,z_j, $(j=1,\ldots,N-1)$, is proportional to the ''hyperradius'' R. The angular Laplacian's eigenvalue parameter Λ determines the harmonic's *degree* by setting the radial scale as R^Λ. Neither the H_j nor the ladder operators $E_{\pm\boldsymbol{\epsilon}_j}$ affect this radial factor, as is to be expected of angular momentum-like operators. With $(3N-4)$ angular coordinates, a complete specification of the harmonics requires $(3N-5)$ labels in addition to Λ. The vector $\boldsymbol{\mu}$ provides $\ell=\lfloor(3N-3)/2\rfloor$ of them in the form of the eigenvalues m_j of all the H_j. The remaining labels are taken from the ℓ-component vector $\boldsymbol{\lambda}$ consisting of the ''string lengths'' $\lambda_{\boldsymbol{\epsilon}_j}$ for the ladder operators $E_{\pm\boldsymbol{\epsilon}_j}$. The first $(\ell-2)$ components of $\boldsymbol{\lambda}$ suffice to reach a total of $(3N-4)$ labels if N is odd, as do the first $(\ell-1)$ components for N even.

It remains to determine the function F introduced in (3.1). The requirement $H_jF\equiv 0$ for all j implies that F depends only on $(u_j^2+v_j^2)$ [and possibly on the single unpaired coordinate w in case $(3N-3)$ is odd]. Furthermore, if $\Sigma_j|m_j|=\Lambda$ the product $\Pi_j(u_j+i\,v_j)^{m_j}$ already accounts for the radial factor R^Λ, i.e., $F=\text{const}$ in this case. This circumstance suggests constructing complete sets of degenerate harmonics with the same Λ as follows: We set m_1 to its maximum value and all other $m_j=0$, i.e., $\boldsymbol{\mu}=(\Lambda,0,\ldots,0)$.[24] For $j=1,\ldots,\ell$, the modulus of the projection $m_{\boldsymbol{\epsilon}_j}=\boldsymbol{\mu}\cdot\boldsymbol{\epsilon}_j/(\boldsymbol{\epsilon}_j\cdot\boldsymbol{\epsilon}_j)$ sets the lower limit for the corresponding string length $\lambda_{\boldsymbol{\epsilon}_j}$. However, $\lambda_{\boldsymbol{\epsilon}_j}$ cannot be larger than $m_{\boldsymbol{\epsilon}_j}$ either, for applying any of the *raising* operators $E_{+\boldsymbol{\epsilon}_j}$ necessarily results in a vector $\boldsymbol{\mu}'=\boldsymbol{\mu}+\boldsymbol{\epsilon}_j$ having $\Sigma_j|m_j'|>\Lambda$, i.e., in a harmonic not belonging to the same set of degenerate functions. With $\boldsymbol{\lambda}$ completely specified, the first harmonic reads

$$Y_{\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda}}=c_\Lambda(x_1+iy_1)^\Lambda,\quad \boldsymbol{\lambda}=(m_{\boldsymbol{\epsilon}_1},0,...,0),\quad \boldsymbol{\mu}=(\Lambda,0,...,0),\tag{3.23}$$

where c_Λ denotes a normalization constant. Starting from this first function, we generate harmonics by applying all the lowering operators $E_{-\boldsymbol{\epsilon}_j}$ first to (3.23), then applying the $E_{-\boldsymbol{\epsilon}_j}$ to the harmonics so obtained, and so on in a *recursive procedure*:

$$Y_{\Lambda,\boldsymbol{\mu}-\boldsymbol{\epsilon}_j,\boldsymbol{\lambda}'}=\frac{1}{\sqrt{(\lambda_{\boldsymbol{\epsilon}_j}+m_{\boldsymbol{\epsilon}_j})(\lambda_{\boldsymbol{\epsilon}_j}-m_{\boldsymbol{\epsilon}_j}+1)}}E_{-\boldsymbol{\epsilon}_j}Y_{\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda}}\tag{3.24}$$

for all $\boldsymbol{\epsilon}_j$ strings that have not yet terminated, i.e., the strings having $m_{\boldsymbol{\epsilon}_j}>-\lambda_{\boldsymbol{\epsilon}_j}$. The recursive nature of the process suggests gathering together the harmonics in ''levels,'' with the level of a harmonic indicating the number of lowering operator steps separating it from the first harmonic (3.23). At level 0 with the harmonic (3.23), the $\boldsymbol{\epsilon}_1$ string is the only string with nonzero length, and we can only generate one harmonic of level 1. But for $\ell>1$, this level-1 harmonic will already have several nonvanishing $m'_{\boldsymbol{\epsilon}_j}$ and corresponding $\lambda'_{\boldsymbol{\epsilon}_j}$, giving rise to additional $\boldsymbol{\epsilon}_j$ strings starting from this level. Repeating this procedure level by level, we work our way down until finally reaching the last harmonic with $\boldsymbol{\mu}=(-\Lambda,0,\ldots,0)$. The procedure stops automatically, because applying any of the lowering operators to this last harmonic maps it to zero (exactly as $l_-Y_{l,-l}\equiv 0$ in three dimensions).

In d dimensions, the total number of different hyperspherical harmonics with the same ''grand angular momentum'' Λ is (Ref. 22, p. 265).

$$\dim(\Lambda;d)=\frac{2\Lambda+d-2}{\Lambda+d-2}\binom{\Lambda+d-2}{d-2}.\tag{3.25}$$

For a given Λ, the procedure outlined above generates exactly $\dim(\Lambda,d)$ independent functions, i.e., a complete set of hyperspherical harmonics. However, some of these harmonics still require modifying for our purpose, as the following observations illustrate.

Note first that with our choice for $E_{\boldsymbol{\epsilon}_j}$, the first $(\ell-1)$ ladder operators always change two m_j in opposite directions, thus leaving $\Sigma_j|m_j|=\Lambda$ invariant when starting from the first harmonic. For all these harmonics, the function F reduces to a constant. At some stage during the above construction, however, an $\boldsymbol{\epsilon}_\ell$ string will appear along which the sum $\Sigma_j|m_j|$ no longer remains constant. Since the ladder operators do not change the overall radial factor R^Λ, the function F accounts for any ''missing powers'' in $\Pi_j(u_j+iv_j)^{m_j}$ for harmonics having $\boldsymbol{\mu}$ with $\Sigma_j|m_j|<\Lambda$. For instance, in three dimensions the lowering operator l_- [E_1^- in (3.3c)'s notation] removes one power of $(x+iy)$ from the ''highest'' harmonic $r^lY_{ll}\sim(x+iy)^l=r^l\sin^l\theta\exp(il\phi)$ while simultaneously adding a factor $F=z=r\cos\theta$ to the next harmonic $r^lY_{l,l-1}\sim z\,(x+iy)^{l-1}=r^l\cos\theta\sin^{l-1}\theta\exp(i[l-1]\phi)$. In the same way, repeated application of l_- generates the higher-degree Legendre polynomials in $\cos\theta$ making up F in this case.

So far, Λ and the vector $\boldsymbol{\mu}$ provide enough information to uniquely specify the hyperspherical harmonics; we need their additional label $\boldsymbol{\lambda}$ only for the purpose of determining transformation matrix elements, not to distinguish the harmonics from one another. For these harmonics, the construction described above provides the appropriate function F.

In a higher-dimensional setting, however, the $E_{\boldsymbol{\epsilon}_j}$ do not all commute with each other. It is thus possible to arrive at *different* functions with the *same* $\boldsymbol{\mu}$ along different ladder operator sequences, starting from the first harmonic (3.23). This is exactly the situation of ''degenerate eigenvalues:'' in this case, $\boldsymbol{\mu}$, the set of eigenvalues of the H_j, has multiplicity higher than one. Although our construction generates the appropriate number of independent functions, thus providing a basis system for the higher-dimensional eigenspace associated with $\boldsymbol{\mu}$, we now need additional labels—to be taken from $\boldsymbol{\lambda}$—to distinguish the different eigenfunctions with degenerate eigenvalue $\boldsymbol{\mu}$. Suppose, therefore, the vector $\boldsymbol{\mu}$ occurs with multiplicity $\nu>1$. Our procedure generates ν independent functions Φ_ρ, $\rho=1,\ldots,\nu$, all having the same $\Pi_j(u_j+iv_j)^{m_j}$. These functions differ only in their F. Harmonics suitable for the calculation of rotation matrix elements are expressed as linear combinations

$$\tilde{\Phi}_\sigma=\sum_{\rho=1}^{\nu} a_{\sigma\rho}\Phi_\rho, \tag{3.26}$$

and the requirement

$$(E_{+\boldsymbol{\epsilon}_j})^{\lambda_{\boldsymbol{\epsilon}_j}-m_{\boldsymbol{\epsilon}_j}+1}\tilde{\Phi}_\sigma=0, \qquad \text{for } j=1,\ldots,j_{\max} \tag{3.27}$$

determines the coefficients $a_{\sigma\rho}$ for $\sigma=1,\ldots,\nu$ (and thus the functions F). In this way, appropriate sets of parameters $(\lambda_{\boldsymbol{\epsilon}_1},\ldots,\lambda_{\boldsymbol{\epsilon}_{j_{\max}}})$ serve to distinguish the harmonics by enforcing specific lengths for the different $\boldsymbol{\epsilon}_j$ strings. As noted previously, the number of additional labels required is $j_{\max}=\ell-2$ (or $\ell-1$) for even (or odd)-dimensional spaces, respectively. With this modification, even the harmonics corresponding to degenerate vectors $\boldsymbol{\mu}$ show the desired behavior under rotations in d dimensions, and their transformation matrix elements can be deduced from their ''quantum numbers'' $(\Lambda,\boldsymbol{\mu},\boldsymbol{\lambda})$ directly.

Finally, a remark concerning the degeneracy of $\boldsymbol{\mu}$ seems in order. In three dimensions, it is impossible to arrive at the same $\boldsymbol{\mu}\equiv m$ along different strings, because there is *only one pair of ladder operators*. Nevertheless, even in this case we note that the m components arising for a given l $(\equiv\Lambda)$ occur again for $l'>l$. Due to the mostly pairwise change of m_j in higher-dimensional spaces, we expect that a vector $\boldsymbol{\mu}$ occurring for a given Λ will appear again as a label for harmonics with $\Lambda'=\Lambda+2,\Lambda+4,\ldots$. For $d>4$ there are noncommuting ladder operators. Because the number of different pathways leading from the highest harmonic to a specific $\boldsymbol{\mu}$ increases with the length of this path, $\boldsymbol{\mu}$'s multiplicity increases with Λ. Furthermore, since different functions F imply contributions from different $(u_j^2+v_j^2)$ terms, higher multiplicity—even in spaces with more than three dimensions—can only arise for $\Lambda-\Sigma_j|m_j|\geqslant 2$. A detailed analysis of the recursion in Λ with $\boldsymbol{\mu}$ held fixed confirms these expectations, yielding for the multiplicity of $\boldsymbol{\mu}$

$$\mathrm{mult}(\boldsymbol{\mu};\Lambda,d)=\binom{p+q}{q} \tag{3.28}$$

$$\text{with } p=\left\lfloor\frac{\Lambda-\Sigma_j|m_j|}{2}\right\rfloor, \quad q=\left\lfloor\frac{d-3}{2}\right\rfloor,$$

a useful result to test the implementation of the (recursive) procedure.

IV. EXAMPLE: TWO-ELECTRON SYSTEM

For the purpose of illustration, we apply the method outlined in the preceding sections to a Coulombic three-body system. Specifically, we consider a two-electron atom or ion with atomic number Z, i.e., a system with only one heavy particle. The case of diatomic molecules (two-center Coulomb system) requires additional modifications of the hyperspherical approach.[25]

A. Coulomb interactions among three particles

After elimination of the c.m. motion, a three-particle system requires two Jacobi vectors for its description. We use three different Jacobi trees, $\boldsymbol{\xi}_i$, $\boldsymbol{\eta}_i$, $\boldsymbol{\zeta}_i$, $(i=1,2)$. With $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_N$ denoting the positions of the two electrons and the nucleus, respectively, the relevant mass-weighted Jacobi vectors read

$$\boldsymbol{\xi}_1=\sqrt{\frac{M_e}{2}}(\boldsymbol{r}_1-\boldsymbol{r}_2), \tag{4.1a}$$

$$\boldsymbol{\xi}_2=\sqrt{\frac{2M_eM_N}{M_N+2M_e}}\left(\frac{\boldsymbol{r}_1+\boldsymbol{r}_2}{2}-\boldsymbol{r}_N\right), \tag{4.1b}$$

$$\boldsymbol{\eta}_1=\sqrt{\frac{M_NM_e}{M_N+M_e}}\,(\boldsymbol{r}_N-\boldsymbol{r}_1)=\cos\beta\boldsymbol{\xi}_1+\sin\beta\boldsymbol{\xi}_2, \tag{4.2}$$

$$\boldsymbol{\zeta}_1=\sqrt{\frac{M_NM_e}{M_N+M_e}}\,(\boldsymbol{r}_2-\boldsymbol{r}_N)=\cos\gamma\boldsymbol{\xi}_1+\sin\gamma\boldsymbol{\xi}_2, \tag{4.3}$$

where

$$\cos\beta=-\sqrt{\frac{M_N}{2(M_N+M_e)}}, \tag{4.4a}$$

$$\sin\beta=-\sqrt{\frac{M_N+2M_e}{2(M_N+M_e)}}, \tag{4.4b}$$

and $\gamma=-\beta$. M_N and M_e denote the nuclear and electron mass, respectively. The vectors $\boldsymbol{\eta}_2$ and $\boldsymbol{\zeta}_2$ will not be needed, since the Coulomb interaction among the three particles takes the form

$$V_C=\sqrt{\frac{M_e}{2}}\frac{1}{|\boldsymbol{\xi}_1|}-\sqrt{\frac{M_NM_e}{M_N+M_e}}\left(\frac{Z}{|\boldsymbol{\eta}_1|}+\frac{Z}{|\boldsymbol{\zeta}_1|}\right). \tag{4.5}$$

At this point, the familiar approach using the transformation of the *coordinates* would exploit (4.2)–(4.3) in a multipole expansion of the electron–nucleus interactions in terms of the coordinates $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$. In our method, however, we apply the (kinematic) rotations (4.2)–(4.3) to the *wave functions* instead. This amounts to calculating the three interaction matrix elements in three different coordinate systems. The immediately obvious advantage of this method is that each of the three terms takes exactly the same form. Each of the integrals reduces to the same radial integral as in the textbook example of hydrogen. Higher-order multipoles and nested integrations over powers of the coordinates do not occur.

B. Choice of H_j and ladder operators

The kinetic energy has rotational symmetry in six dimensions. Partitioning the six coordinates into pairs defines three nonintersecting planes, and thus $\ell=3$ mutually commuting rotations. We choose the corresponding first-order operators H_j as $J_{x_1y_1}$, $J_{x_2y_2}$, and $J_{z_1z_2}$. Here $\{x_1,y_1,z_1,x_2,y_2,z_2\}$ denote the Cartesian components of the mass-weighted Jacobi vectors. Accordingly, the harmonics—simultaneous eigenfunctions of the three H_j with eigenvalues $\boldsymbol{\mu}=(m_1,m_2,m_3)$—take the form

$$F(x_1^2+y_1^2,\,x_2^2+y_2^2,z_1^2+z_2^2)\prod_{j=1}^{2}\,(x_j+i\,\mathrm{sign}(m_j)\,y_j)^{|m_j|}(z_1+i\,\mathrm{sign}(m_3)\,z_2)^{|m_3|}, \tag{4.6}$$

where we have rewritten (3.1) so as to yield non-negative powers of the hyperradius for either sign of the m_j. The appropriate base set of three ladder operator pairs $E_{\pm\boldsymbol{\epsilon}_j}$ is then specified by the vector labels $\boldsymbol{\epsilon}_1=(1,-1,0)$, $\boldsymbol{\epsilon}_2=(0,1,-1)$, and $\boldsymbol{\epsilon}_3=(0,1,1)$, corresponding to $E_{12}^{\pm\mp}$, $E_{23}^{\pm\mp}$, and $E_{23}^{\pm\pm}$ in (3.3a)–(3.3b).

C. Labeling the basis functions

Five angles specify each point on a sphere of fixed hyperradius R in six-dimensional space. Consequently, our hyperspherical harmonics require five labels—related to the numbers of nodes in the five angular variables—for their identification. Besides Λ and $\boldsymbol{\mu}=(m_1,m_2,m_3)$ we need one more label, $\lambda_{\boldsymbol{\epsilon}_1}$, the string length along the ladder operator sequences spanned by $E_{\pm\boldsymbol{\epsilon}_1}$.

The set of harmonics so defined differs from the more familiar set labeled by quantum numbers $(l_1,m_1,l_2,m_2,n_\alpha)$.[17,22] However, the latter set of harmonics does not make use of the third eigenvalue m_3. Since rotating the harmonics is accomplished by acting on them with ladder operators which in turn act on *all* the m_j, the conventional harmonics specified by $(l_1,m_1,l_2,m_2,n_\alpha)$ are not suited for our purpose. They may, of course, still serve as a basis set in an application, being then expanded into the new set labeled by $(\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1})$ prior to the actual rotation. A straightforward transformation links the two basis sets.

D. Explicit construction of harmonics with degenerate $\boldsymbol{\mu}$

As an example, we derive the expressions for some harmonics that are not completely characterized by Λ and $\boldsymbol{\mu}$. Consider for instance harmonics with $\Lambda=4$, $\boldsymbol{\mu}=(1,0,1)$. According to (3.28), there are two harmonics with these labels, to be distinguished by the additional parameter $\lambda_{\boldsymbol{\epsilon}_1}$. With the particular vector $\boldsymbol{\mu}$ of this example, we obtain $m_{\boldsymbol{\epsilon}_1}=\frac{1}{2}$, thus setting the lower limit for $\lambda_{\boldsymbol{\epsilon}_1}$. The two essentially different ways of arriving at the set of labels $(1,0,1)$ from harmonics of the next-lower level are given by an $E_{-\boldsymbol{\epsilon}_1}$-step from $\boldsymbol{\mu}'=(2,-1,1)$, and by an $E_{-\boldsymbol{\epsilon}_3}$-step from $\boldsymbol{\mu}''=(1,1,2)$. Both of these labels have multiplicity 1 since $\Sigma_j|m_j|=\Lambda$; the function $F(\ldots)$ in the corresponding harmonics reduces to a normalization constant. Up to these normalizing factors, c_1 and c_2, (4.6) gives these harmonics as

$$Y_{4,(2,-1,1),\,3/2}=c_1(x_1+iy_1)^2(x_2-iy_2)(z_1+iz_2), \tag{4.7a}$$

$$Y_{4,(1,1,2),1}=c_2(x_1+iy_1)(x_2+iy_2)(z_1+iz_2)^2. \tag{4.7b}$$

Applying the appropriate ladder operators to these harmonics provides a basis for the two-dimensional eigenspace of degenerate $\boldsymbol{\mu}=(1,0,1)$:

$$\Phi_1=E_{-\boldsymbol{\epsilon}_1}Y_{4,(2,-1,1),\,3/2}=c_3\,(x_1+iy_1)(z_1+iz_2)([x_1^2+y_1^2]-2[x_2^2+y_2^2]), \tag{4.8a}$$

$$\Phi_2=E_{-\boldsymbol{\epsilon}_3}Y_{4,(1,1,2),1}=c_4(x_1+iy_1)(z_1+iz_2)(2[x_2^2+y_2^2]-[z_1^2+z_1^2]), \tag{4.8b}$$

with constant factors c_3, c_4 to be ultimately absorbed into the normalization. Because Φ_1 is obtained from $Y_{4,(2,-1,1),3/2}$ by applying $E_{-\boldsymbol{\epsilon}_1}$, it behaves under the relevant rotations exactly as is required for $Y_{4,(1,0,1),3/2}$. Apart from the normalization constant c_3, we thus find

$$Y_{4,(1,0,1),\,3/2}=c_3(x_1+iy_1)(z_1+iz_2)([x_1^2+y_1^2]-2[x_2^2+y_2^2]). \tag{4.9}$$

However, the second harmonic with the same $\boldsymbol{\mu}=(1,0,1)$ does not simply coincide with Φ_2, because the latter has the same string-length $\lambda_{\boldsymbol{\epsilon}_1}=\frac{3}{2}$ as Φ_1: One easily verifies that

$$E_{\boldsymbol{\epsilon}_1}\Phi_2\neq 0. \tag{4.10}$$

We can, however, construct a harmonic $Y_{4,(1,0,1),1/2}$ as a linear combination of Φ_1 and Φ_2 by requiring

$$E_{\boldsymbol{\epsilon}_1}(a\Phi_1+b\Phi_2)\equiv 0, \tag{4.11}$$

yielding the condition $3ac_3-2bc_4=0$ and thus

$$Y_{4,(1,0,1),\,1/2}=c_5(x_1+iy_1)(z_1+iz_2)(2[x_1^2+y_1^2]+2[x_2^2+y_2^2]-3[z_1^2+z_2^2]). \tag{4.12}$$

The two harmonics with ''degenerate'' $\boldsymbol{\mu}=(1,0,1)$ are now distinguished by their respective $\lambda_{\boldsymbol{\epsilon}_1}$.

E. Rotating the harmonics

Equation (4.2) describes the *coordinate* transformation between the Jacobi trees $(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ and $(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)$. Accordingly, the hyperspherical harmonics transform as

$$|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle_{\boldsymbol{\xi}}=\sum_{\boldsymbol{\mu}'} D^{(\lambda_{\boldsymbol{\epsilon}_1})}_{\boldsymbol{\mu}',\boldsymbol{\mu}}(\beta)|\Lambda,\boldsymbol{\mu}',\lambda_{\boldsymbol{\epsilon}_1}\rangle_{\boldsymbol{\eta}} \tag{4.13}$$

where the subscripts on the ket vectors indicate the respective Jacobi tree. The transformation matrix elements are given by

$$\begin{aligned} D^{(\lambda_{\boldsymbol{\epsilon}_1})}_{\boldsymbol{\mu}',\boldsymbol{\mu}}(\beta)&=\langle\Lambda,\boldsymbol{\mu}',\lambda_{\boldsymbol{\epsilon}_1}|\exp(i\beta[J_{x_1x_2}+J_{y_1y_2}+J_{z_1z_2}])|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle\\ &=\langle\Lambda,\boldsymbol{\mu}',\lambda_{\boldsymbol{\epsilon}_1}|\exp(i\beta[E_{\boldsymbol{\epsilon}_1}+E_{-\boldsymbol{\epsilon}_1}+H_3])|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle, \end{aligned} \tag{4.14}$$

according to Sec. III B. Following the analysis given there, we find

$$D^{(\lambda_{\boldsymbol{\epsilon}_1})}_{\boldsymbol{\mu}',\boldsymbol{\mu}}(\beta)=e^{i[m'-m]\pi/2}d^{(\lambda_{\boldsymbol{\epsilon}_1})}_{m',m}(2\beta)e^{im_3\beta}\delta_{\boldsymbol{\mu}',\boldsymbol{\mu}+n\boldsymbol{\epsilon}_1}, \tag{4.15}$$

with

$$m=\tfrac{1}{2}(m_1-m_2), \tag{4.16a}$$

$$m'=\tfrac{1}{2}(m_1'-m_2'), \tag{4.16b}$$

$$n=m_1'-m_1=-m_2'+m_2 \tag{4.16c}$$

in terms of $\boldsymbol{\mu}$ components. Replacing the angle β with $\gamma=-\beta$ yields the expressions for the transformation to Jacobi tree $\boldsymbol{\zeta}$.

F. Symmetry properties of basis functions

Due to the very high degeneracy of harmonics with the same grand angular momentum Λ as expressed in (3.25), it is important to exploit various symmetries of the functions in order to reduce the size of the hyperspherical basis. Reflection through the origin of the coordinate system transforms all six coordinates into their negatives. According to (4.6), the harmonic $|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle$ picks up a factor $(-1)^{|m_1|+|m_2|+|m_3|}$ under this operation. Since all the ladder operators change two of the m_j at a time, the sum in the exponent has the same parity as Λ. This first observation thus restricts the basis set by allowing only even Λ for even-parity states and odd Λ for odd-parity states.

Consider next the harmonics' symmetry under interchange of identical particles (i.e., the two electrons). This interchange is most easily described in the Jacobi tree $\{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\}$:

$$\boldsymbol{\xi}_1 \leftrightarrow -\boldsymbol{\xi}_1, \quad \boldsymbol{\xi}_2 \leftrightarrow \boldsymbol{\xi}_2. \tag{4.17}$$

Accordingly, the harmonic $|\Lambda,m_1,m_2,m_3,\lambda_{\boldsymbol{\epsilon}_1}\rangle$ turns into the harmonic $|\Lambda,m_1,m_2,-m_3,\lambda_{\boldsymbol{\epsilon}_1}\rangle$ under interchange of the electrons while acquiring a factor $(-1)^{m_1+m_3+S}$ (where S denotes the total spin of the two electrons). Antisymmetrized basis functions may thus be labeled by non-negative m_3 only, and $m_3=0$ is possible only for harmonics with (m_1+S) *even*.

Furthermore, when choosing the first-order operators H_j we have arranged the six coordinates in a way that allows us to identify m_1 and m_2 with the z projections of the three-dimensional relative angular momenta $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$. Therefore, the sum $m_1+m_2=m_{\text{tot}}$ is the z component of the coupled (total) orbital angular momentum L. The fact that the Coulomb interaction does not couple states with different m_{tot} reduces the size of a basis consisting of antisymmetrized harmonics accordingly. This point is particularly interesting because the system's invariant m_{tot} restricts our basis, even though the corresponding total orbital angular momentum L is *not defined* in this basis. The reason for this seemingly surprising fact lies in our use of first-order operators only; hence L_z can be identified, but not the second-order operators $\boldsymbol{l}_j^2$ or L^2. The absence of the invariant three-dimensional L is the main trade-off we have to accept when treating all transformations as rotations in a genuinely six-dimensional space (Ref. 22, especially Sec. 10.2., p. 267ff).

G. Calculation of matrix elements

While Cartesian coordinates prove most appropriate for manipulating the hyperspherical harmonics using ladder operators, hyperspherical coordinates lend themselves for the calculation of matrix elements. Specifically, the familiar representation of Cartesian coordinates

$$\begin{aligned} x_1 &= R\cos\alpha\sin\theta_1\cos\varphi_1, \quad x_2 = R\sin\alpha\sin\theta_2\cos\varphi_2, \\ y_1 &= R\cos\alpha\sin\theta_1\sin\varphi_1, \quad y_2 = R\sin\alpha\sin\theta_2\sin\varphi_2 \\ z_1 &= R\cos\alpha\cos\theta_1, \quad z_2 = R\sin\alpha\cos\theta_2, \end{aligned} \tag{4.18}$$

transforms (4.6) into

$$R^{\Lambda}F\sin^{|m_1|}\theta_1 e^{im_1\varphi_1}\sin^{|m_2|}\theta_2 e^{im_2\varphi_2}(\cos\alpha\cos\theta_1 + i\,\text{sign}(m_3)\sin\alpha\cos\theta_2)^{|m_3|}, \tag{4.19}$$

where F is a function of $(\cos^2\alpha\sin^2\theta_1)$, $(\sin^2\alpha\sin^2\theta_2)$, and $(\cos^2\alpha\cos^2\theta_1+\sin^2\alpha\cos^2\theta_2)$. Note that this form remains the same, regardless of whether the Cartesian components (x,y,z) refer to the Jacobi vectors in tree $\boldsymbol{\xi}$, in $\boldsymbol{\eta}$, or in $\boldsymbol{\zeta}$. The relevant interaction operator entering into the matrix element is always $1/(R\cos\alpha)$, for each of the three pairwise Coulomb interactions, with the angle α referring to a different coordinate system in each case. The θ and α integrals arising in the calculation are related to the Euler beta function,[26] namely,

$$\int_0^{2\pi} d\phi e^{i(m'-m)\varphi} = 2\pi\delta_{m'm}, \tag{4.20a}$$

$$\int_0^{\pi} d\theta\sin^p\theta\cos^q\theta = (1+(-1)^q)\,\frac{(p-1)!!(q-1)!!}{(p+q)!!}c_{pq}, \tag{4.20b}$$

$$\int_0^{\pi/2} d\alpha\sin^p\alpha\cos^q\alpha = \frac{(p-1)!!(q-1)!!}{(p+q)!!}c_{pq}, \tag{4.20c}$$

where

$$c_{pq}=\begin{cases} \frac{\pi}{2} & p,q \text{ both even} \\ 1 & \text{otherwise.} \end{cases} \tag{4.21}$$

A multipole expansion would have forced us to split the last integral into two parts, with different integrands depending on whether $|\cos\beta\boldsymbol{\xi}_1|$ is greater or smaller than $|\sin\beta\boldsymbol{\xi}_2|$. The resulting integral could only be expressed as a sum of factorial quotients, rather than *a single term*, as in (4.20c).

Gathering together all the pieces developed in this section, we find for the angular part of the matrix elements

$$\langle\Lambda',\boldsymbol{\mu}',\lambda'_{\boldsymbol{\epsilon}_1}|V_C|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle=\langle\Lambda',\boldsymbol{\mu}',\lambda'_{\boldsymbol{\epsilon}_1}|\frac{1}{r_{12}}-\frac{Z}{r_{1N}}-\frac{Z}{r_{2N}}|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle \tag{4.22}$$

$$=\frac{1}{R}\sqrt{\frac{M_e}{2}}\langle\Lambda',\boldsymbol{\mu}',\lambda'_{\boldsymbol{\epsilon}_1}|\frac{1}{\cos\alpha}|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle-\frac{Z}{R}\sqrt{\frac{M_NM_e}{M_N+M_e}}$$

$$\times\sum_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}\{(D^{(\lambda'_{\boldsymbol{\epsilon}_1})}_{\boldsymbol{\mu}_1,\boldsymbol{\mu}'}(\beta))^\dagger D^{(\lambda_{\boldsymbol{\epsilon}_1})}_{\boldsymbol{\mu}_2,\boldsymbol{\mu}}(\beta)+(D^{(\lambda'_{\boldsymbol{\epsilon}_1})}_{\boldsymbol{\mu}_1,\boldsymbol{\mu}'}(\gamma))^\dagger D^{(\lambda_{\boldsymbol{\epsilon}_1})}_{\boldsymbol{\mu}_2,\boldsymbol{\mu}}(\gamma)\}$$

$$\times\langle\Lambda',\boldsymbol{\mu}_1,\lambda'_{\boldsymbol{\epsilon}_1}|\frac{1}{\cos\alpha}|\Lambda,\boldsymbol{\mu}_2,\lambda_{\boldsymbol{\epsilon}_1}\rangle, \tag{4.23}$$

with the further simplification $\gamma=-\beta$. The double summation (transforming the bra and ket vectors between Jacobi trees) seems to spoil the present approach's advantage over a multipole expansion. After all, the latter also leads to two summations, namely, a sum over the multipole order and another summation stemming from the analytical evaluation of the nested integral over α. Note, however, that the present method achieves significantly more with two summations: it accounts for *all the cusps* in the wave functions whenever an interparticle separation vanishes.

Finally, the particular case of a three-body system involves two relevant Jacobi vectors only. All possible Jacobi trees are thus related to one another by rotations in the *same plane* $(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$. This particularity of the three-body problem might suggest arranging the Cartesian coordinates of the Jacobi vectors in the following way:

$$\{x_1,x_2,y_1,y_2,z_1,z_2\}, \tag{4.24}$$

rather than our choice (3.5). The H_j resulting from the above arrangement of coordinates coincide with $J_{x_1x_2}$, $J_{y_1y_2}$, and $J_{z_1z_2}$, thereby simplifying the rotation of harmonics: *All* rotations reduce to the first case of (z_1z_2)-type rotations discussed in Sec. III B, with matrix elements

$$\langle\Lambda',\boldsymbol{\mu}',\lambda'_{\boldsymbol{\epsilon}_1}|\exp(i\beta[J_{x_1x_2}+J_{y_1y_2}+J_{z_1z_2}])|\Lambda,\boldsymbol{\mu},\lambda_{\boldsymbol{\epsilon}_1}\rangle=e^{i(m_1+m_2+m_3)\beta}\delta_{\Lambda'\Lambda}\delta_{\boldsymbol{\mu}'\boldsymbol{\mu}}\delta_{\lambda'_{\boldsymbol{\epsilon}_1}\lambda_{\boldsymbol{\epsilon}_1}}. \tag{4.25}$$

This orthogonality relation virtually eliminates the double summation over $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2$ in (4.23). However, the corresponding harmonics are now functions of $(x_1\pm ix_2)$, $(y_1\pm iy_2)$, and $(z_1\pm iz_2)$. Transforming to hyperspherical coordinates using

$$x_1=R\sin\alpha_1\cos\alpha_2\cos\varphi_1,\quad x_2=R\sin\alpha_1\cos\alpha_2\sin\varphi_1,$$

$$y_1=R\sin\alpha_1\sin\alpha_2\cos\varphi_2,\quad y_2=R\sin\alpha_1\sin\alpha_2\sin\varphi_2, \tag{4.26}$$

$$z_1=R\cos\alpha_1\cos\varphi_3,\quad z_2=R\cos\alpha_1\sin\varphi_3,$$

yields $H_j = -i(\partial/\partial\varphi_j)$, with $0 \leqslant \varphi_j \leqslant 2\pi$ and $0 \leqslant \alpha_k \leqslant \pi/2$. While the H_j, as well as the harmonics, now obviously attain their simplest form, the choice (4.24) has two serious drawbacks: (i) None of the three eigenvalues m_j of the H_j have physical significance, and (ii) $1/|\boldsymbol{\xi}_1| = (x_1^2 + y_1^2 + z_1^2)^{-1/2}$ is a (complicated) function of all five angles $(\alpha_1, \alpha_2, \varphi_1, \varphi_2, \varphi_3)$. The latter problem is solved by using hyperspherical coordinates (4.18) instead of (4.26), yielding for each of the combinations $(x_1 \pm i x_2)$, $(y_1 \pm i y_2)$, and $(z_1 \pm i z_2)$ a sum of two terms [as opposed to the single terms obtained for $(x_j \pm i y_j)$ in the previous subsections]. Expanding powers of these binomials leads to two additional summations, leaving us with no net gain.

V. CONCLUSION

Recognizing the independent-particle model's failure to account for cusps due to variables on which the wave function does not depend explicitly, we have developed a method that satisfies Kato's cusp condition through reference frame transformations. By transforming the *wave function* to the appropriate reference frame, we expose the wave function's cusp arising from vanishing of any given particle separation r_{ij}. In addition to satisfying the cusp condition on the wave function, this technique also simplifies the calculation of the pairwise Coulomb interaction $\sim 1/r_{ij}$, as compared to the conventional multipole expansion.

Implementation of the approach outlined above resolves into three major tasks, all addressed in the present investigation: (i) the systematic study of the relevant transformations between reference frames; (ii) the definition of functions suitable for such transformations; (iii) the determination of the transformation matrix. While the literature on Lie algebra provides ready-made solutions to problem (i), it usually fails to do so for (ii) and (iii). Furthermore, the mathematical literature does not exploit the particularity of an atomic or molecular N-body system.

More specifically, problem (i) is solved by using mass-scaled Jacobi coordinates, since the transformations between reference frames reduce then to generalized rotations in $(3N-3)$ dimensions. The Lie algebra so$(3N-3)$ describes these transformations completely, embodied in the sets of commuting rotation operators $\{H_j\}$ and ladder operators $\{E_{\pm\boldsymbol{\epsilon}_j}\}$, $j = 1, \ldots, \lfloor (3N-3)/2 \rfloor$. To solve problem (ii) mentioned above, we have introduced basis functions defined entirely through their behavior under infinitesimal rotations, i.e., when acted upon by the operators H_j and $E_{\pm\boldsymbol{\epsilon}_j}$. Simultaneous eigenfunctions of all the H_j constitute appropriate basis functions, classified further according to their matrix elements for the transformation between reference frames. The latter step removes possible ambiguities whenever the H_j have degenerate eigenvalues. Finally, by extending the concept of Euler-angle rotations from three to higher dimensions, we have provided the solution to task (iii), the determination of the transformation matrix elements. For an arbitrary N-body system, each change of reference frames considered here resolves into a sequence of basic transformations described by these matrix elements.

In an N-body system, the high dimensionality arises from the *product structure* of the $(N-1)$-dimensional particle space and the three-dimensional physical space. The basic step in the transformation between reference frames reduces to a two-dimensional rotation in a plane of particle space. Upon expansion of the particle space variables into their physical-space components, the basic rotation induces three two-dimensional rotations in $(3N-3)$-dimensional space that are analyzed using the Lie algebra so$(3N-3)$. To take advantage of the efficiency offered by the Lie algebraic method, we treat the transformations as rotations in a genuine $(3N-3)$-dimensional space. Thus, we work with the *first-order* differential operators H_j and $E_{\pm\boldsymbol{\epsilon}_j}$ throughout. Arranging the generic Cartesian components (x_j, y_j, z_j) of the particle-space variables $\boldsymbol{\xi}_j$ appropriately, we can *partially* recover the product structure characteristic for the physical application at hand: The eigenvalues of $(N-1)$ among the H_j carry physical significance; and we interpret them as z projections of orbital angular momenta. Their sum represents the total angular momentum's z projection, an invariant of the system. However, neither the individual angular momenta, nor the coupled (total) angular momentum can appear in our treatment, as they are represented by second-order operators. Other methods exploit the product structure of the $(3N-3)$-dimensional coordinate space to a larger extent by solving the Laplacian's (second-order) eigenvalue problem through separation of variables. However, the resulting hyperspherical harmonics incorporate these features of the three-dimensional physical space at considerable cost: Transforming these functions between reference frames proves very inefficient, requiring essentially their expansion into the equivalent sets of harmonics introduced in the present investigation.

In conclusion, we have demonstrated by explicit construction the possibility of describing a system of N charged particles without recourse to multipole expansions. Kato's cusp condition is satisfied implicitly through reference frame transformations of the *wave function*. We introduced appropriate basis functions and discussed their symmetries under particle interchange and reflection through the origin. We derived the matrix elements for the transformation between reference frames for arbitrary numbers of particles, and we showed that the interaction matrix elements attain a simpler form as compared to the conventional multipole expansion. Because the formalism underlying our approach relies exclusively on first-order differential operators, it does not incorporate angular momenta, thereby limiting its usefulness for bound state problems. However, the technique introduced here provides significant simplifications in scattering problems where partial angular momenta are not resolved.

Note added in proof. Our transformation matrices are analogous to the so-called Raynal–Revai coefficients. See, e.g., A. Novoselsky and J. Katriel, Phys. Rev. A **49**, 833 (1994), especially Sec. 5.

ACKNOWLEDGMENT

We are grateful to Professor Ugo Fano whose critical questions and comments on earlier drafts of this article helped us clarify the conceptual implications of the present approach.

[1] T. Kato, Commun. Pure Appl. Math. **10**, 151 (1957).
[2] D. P. Carroll, H. J. Silverstone, and R. M. Metzger, J. Chem. Phys. **71**, 4142 (1979).
[3] R. N. Jill, J. Chem. Phys. **83**, 1173 (1985).
[4] E. Ley-Koo and C. F. Bunge, Phys. Rev. A **40**, 477 (1989); **40**, 1215 (1989).
[5] S. P. Goldman, Phys. Rev. Lett. **78**, 2325 (1997).
[6] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw–Hill, New York, 1953), p. 1730.
[7] G. Wannier, Phys. Rev. **90**, 817 (1953).
[8] F. T. Smith, Phys. Rev. **120**, 1058 (1960); J. Math. Phys. **3**, 735 (1962).
[9] L. M. Delves, Nucl. Phys. **9**, 391 (1958–59).
[10] See also: Yu. F. Smirnov and K. V. Shitikova, Sov. J. Part. Nucl. **8**, 344 (1978), and references therein.
[11] J. H. Macek, J. Phys. B **1**, 831 (1968).
[12] See: C. D. Lin, Adv. At. Mol. Phys. **22**, 77 (1986) for references to Lin's decisive contributions in the 1970s and 1980s.
[13] Z. Chen and C. D. Lin, Phys. Rev. A **42**, 18 (1990).
[14] J.-Z. Tang, S. Watanabe, and M. Matsuzawa, Phys. Rev. A **46**, 2437 (1992).
[15] T. A. Heim, G. B. Armen, and A. R. P. Rau, Phys. Rev. A **55**, 2674 (1997).
[16] X. Yang, C. G. Bao, and C. D. Lin, Phys. Rev. A **53**, 3934 (1996); applications to three-electron atomic systems in full detail.
[17] See, e.g., M. Cavagnero, Phys. Rev. A **30**, 1169 (1984); **33**, 2877 (1986); **36**, 523 (1987) for the systematic extension to more than two electrons.
[18] See also U. Fano, Rep. Prog. Phys. **46**, 97 (1983).
[19] U. Fano and A. R. P. Rau, *Atomic Spectra and Collisions* (Academic, Orlando, 1986), Chap. 10.
[20] See, e.g., V. Aquilanti, L. Bonnet, and S. Cavalli, Mol. Phys. **89**, 1 (1996); V. Aquilanti and J. Avery, Chem. Phys. Lett. **267**, 1 (1997); V. Aquilanti, S. Cavalli, C. Coletti, D. De Fazio, and G. Grossi, in *New Methods in Quantum Theory*, edited by C. A. Tsipis, V. S. Popov, D. R. Herschbach, and J. S. Avery (Kluwer, Dordrecht, 1996), p. 233ff; several contributions by other authors in the same volume prove relevant to the subject as well.
[21] J. S. Avery, *Hyperspherical Harmonics* (Kluver, Dordrecht, 1989), and references therein.
[22] U. Fano and A.R.P. Rau, *Symmetries in Quantum Physics* (Academic, San Diego, 1996), Chaps. 2, 4, and 10.
[23] See, e.g., J. F. Cornwell, *Group Theory in Physics* (Academic, San Diego, 1984), Vol. II, Chaps. 13–16; Appendices F, G.
[24] The procedure outlined in the text generates only weights of real, symmetric representations. As these representations account for all sets of harmonics needed in our applications, irreducible representations of so(d) with highest weight other than $(\Lambda,0,\ldots,0)$ are not considered here.
[25] O. Tolstikhin, S. Watanabe, and M. Matsuzawa, Phys. Rev. Lett. **74**, 3573 (1995).
[26] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972), Sec. 6.2.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

Operator integrals and phase space observables

Pekka Lahti[a] and Juha-Pekka Pellonpää[b]
Department of Physics, University of Turku, 20014 Turku, Finland

Kari Ylinen[c]
Department of Mathematics, University of Turku, 20014 Turku, Finland

(Received 20 October 1998; accepted for publication 14 January 1999)

The operator integral $\int f\,dE$ of a complex valued measurable function f with respect to a positive operator measure E is considered. If F is a Neumark dilation of E into a projection measure, then the ''projected'' operator integral $\mathrm{pr}(\int f dF)$ is a restriction of the operator $\int f\,dE$. Necessary and sufficient conditions for the equality $\mathrm{pr}(\int f\,dF) = \int f\,dE$ are obtained. The results are applied to determine the moment operators of the phase space observables generated by the number states. © *1999 American Institute of Physics.* [S0022-2488(99)01104-4]

I. INTRODUCTION

The operator integral of a complex valued measurable function f with respect to an operator measure E is defined as the unique linear operator $L^E(f)$ for which the number $\langle \psi | L^E(f)\varphi\rangle$ equals the integral of f with respect to the complex measure $E_{\psi,\varphi}$ defined by E and the vectors ψ and φ, with ψ ranging over the whole of the Hilbert space $\mathcal{H}$ and φ the natural domain of $L^E(f)$. For positive operator measures E this domain contains, as a subspace, the set of those vectors φ for which f is square integrable with respect to the positive measure $E_{\varphi,\varphi}$. For a positive operator measure E one may also use a Neumark dilation of E into a projection measure F to obtain a ''projected'' operator integral $\mathrm{pr}(L^F(f))$. In general, the thus obtained operator $\mathrm{pr}(L^F(f))$ is a restriction of the operator $L^E(f)$. It will be shown that the two operators are the same exactly when the domain of $L^E(f)$ consists only of those vectors φ for which f is square integrable with respect to $E_{\varphi,\varphi}$. The theory of operator integrals, developed here and in a previous article,[1] will be applied to the so-called phase space observables generated by the number eigenstates.

The structure of the paper is as follows. Section II reviews the basic notions and results of the theory of operator integrals. Section III relates this theory in the case of positive operator measures to the operator integrals obtained via the Neumark dilation theory. In Sec. IV the general results will be applied to determine the moment operators of the number state generated phase space observables and their Cartesian margins.

II. INTEGRATION WITHOUT DILATION

A. The operator integral

Let $\mathcal{H}$ be a complex Hilbert space, with the inner product $\langle\cdot|\cdot\rangle$, and let $\mathcal{L}(\mathcal{H})$ denote the set of bounded operators on $\mathcal{H}$. Let Ω be a nonempty set and $\mathcal{A}$ a σ algebra of subsets of Ω. Let $E:\mathcal{A}\to\mathcal{L}(\mathcal{H})$ be an operator measure, i.e., a set function which is σ additive with respect to the weak, or, equivalently, the strong operator topology of $\mathcal{L}(\mathcal{H})$. For any $\varphi,\psi\in\mathcal{H}$ we denote by $E_{\psi,\varphi}$ the complex measure defined by $E_{\psi,\varphi}(X) = \langle\psi|E(X)\varphi\rangle$. Let $f:\Omega\to\mathbf{C}$ be an $\mathcal{A}$-measurable function and let $\mathcal{D}(L^E(f))$ denote the set of those vectors $\varphi\in\mathcal{H}$ for which f is $E_{\psi,\varphi}$ integrable for each $\psi\in\mathcal{H}$. $\mathcal{D}(L^E(f))$ is a vector subspace of $\mathcal{H}$ and there is a unique linear operator $L^E(f)$, with the domain $\mathcal{D}(L^E(f))$, such that

---

[a]Electronic mail: pekka.lahti@utu.fi
[b]Electronic mail: juhpello@utu.fi
[c]Electronic mail: ylinen@utu.fi

$$\langle \psi | L^E(f)\varphi \rangle = \int_\Omega f \, dE_{\psi,\varphi}$$

for all $\psi \in \mathcal{H}$. We call $L^E(f)$ the *integral* of f with respect to E and we also write $L^E(f) = \int f \, dE$. For further details, see Ref. 1.

Let $\mathcal{D}_f^E$ denote the set of those vectors $\varphi \in \mathcal{H}$ for which $|f|^2$ is $E_{\varphi,\varphi}$ integrable. The following result, proved in Ref. 1, will be important.

Theorem: (a) If the operator measure E is positive, that is, $E(X) \geqslant O$ for each $X \in \mathcal{A}$, then $\mathcal{D}_f^E$ is a subspace of $\mathcal{H}$ and

$$\mathcal{D}_f^E \subseteq \mathcal{D}(L^E(f)).$$

(b) If the operator measure E is projection valued, that is, $E(X)^2 = E(X) = E(X)^*$ for all $X \in \mathcal{A}$, then $\mathcal{D}_f^E = \mathcal{D}(L^E(f))$.

We note that in the above case (a) the inclusion may be proper. Indeed, if $\mu:\mathcal{A} \to [0,1]$ is a probability measure and $E:X \mapsto E(X) := \mu(X)I$, then for any f which is μ integrable, $L^E(f) = (\int f \, d\mu)I$, with $\mathcal{D}(L^E(f)) = \mathcal{H}$, but, if $|f|^2$ is not μ integrable, then $\mathcal{D}_f^E = \{0\}$. It is equally obvious that the equality $\mathcal{D}_f^E = \mathcal{D}(L^E(f))$ does not require E to be projection valued. For instance, if E is a projection valued measure and $g:\Omega \to [a,b]$, with $0 < a < b < \infty$, an $\mathcal{A}$-measurable function, then $F(X) := \int_X g \, dE$ defines a positive operator measure. Then, for any f, and for all $\psi, \varphi \in \mathcal{H}$, the integral $\int f \, dF_{\psi,\varphi}$ exists exactly when the integral $\int fg \, dE_{\psi,\varphi}$ exists, which, in turn, exists if and only if $\int f \, dE_{\psi,\varphi}$ exists. This shows that $\mathcal{D}_f^F = \mathcal{D}_f^E = \mathcal{D}(L^E(f)) = \mathcal{D}(L^F(f))$. Clearly, if f is bounded, then $\mathcal{D}_f^E = \mathcal{D}(L^E(f)) = \mathcal{H}$ for any positive operator measure E. Finally, we note that the equality $\|L^E(f)\varphi\|^2 = \int |f|^2 dE_{\varphi,\varphi}$, $\varphi \in \mathcal{D}_f^E$, does not hold, in general. Section IV D provides examples of this phenomenon.

B. Remarks on the literature

In mathematical literature the operator integrals $L^E(f)$ are mostly discussed only for positive operator measures and the domain of the operator $L^E(f)$ is taken to be $\mathcal{D}_f^E$ instead of the wider domain $\mathcal{D}(L^E(f))$. See, e.g., Refs. 2–5. Riesz and Sz.-Nagy,[2] however, point out on page 460 that the integral $\int f \, dE_{\psi,\varphi}$, $\psi \in \mathcal{H}$, can also converge for certain φ which do not belong to $\mathcal{D}_f^E$. On the other hand, Akhiezer and Glazman[4] report on page 132 of Vol. II an example for which the set $\mathcal{D}_f^E$, as well as the set $\mathcal{D}(L^E(f))$, consists only of the null vector. Schroeck[6] develops a theory of operator integrals under the additional assumption that the positive operator measures involved are absolutely continuous; the question of the domain is not explicitly addressed there. As an example of recent physics literature we mention Ozawa,[7] who also takes the set $\mathcal{D}_f^E$ as the domain of the operator defined by the integral of a measurable function f with respect to a positive operator measure E. The natural domain of the linear operator $L^E(f)$ is, however, the wider set $\mathcal{D}(L^E(f))$. In Sec. III we shall see that the restriction of the operator $L^E(f)$ to $\mathcal{D}_f^E$ equals the projected operator integral $\mathrm{pr}(L^F(f))$ obtained via a Neumark dilation of the (positive) operator measure E into a projection measure F.

C. Densely defined operator integrals

From now on we assume that the operator measure E is positive and so normalized that $E(\Omega) = I$.

Let $\overline{f}$ be the complex conjugate of f. The operator $L^E(\overline{f})$ has the domain $\mathcal{D}(L^E(\overline{f})) = \mathcal{D}(L^E(f))$. If $L^E(f)$ is densely defined, then the adjoint of $L^E(f)$ exists and it extends the operator $L^E(\overline{f})$, that is, $L^E(\overline{f}) \subseteq L^E(f)^*$ (Ref. 1, Lemma A4). Furthermore, if $\mathcal{D}(L^E(f))$ is dense in $\mathcal{H}$, then also $L^E(f)^*$ is densely defined, implying that $L^E(f)$ is closable, with $L^E(f)^{**}$ being the smallest closed linear extension of $L^E(f)$ (Ref. 2, Theorem on p. 305). Thus, if $L^E(f)$ is densely defined, we have

$$L^E(\overline{f}) \subseteq L^E(f)^*,$$

$$L^E(f)\subseteq L^E(f)^{**}.$$

It is well known that $L^E(\overline{f})=L^E(f)^*$ whenever E is projection valued. But it is obvious that this is not a necessary condition for the equality $L^E(\overline{f})=L^E(f)^*$. The operator measure $E(X)=\mu(X)I$ defined by a probability measure μ also demonstrates this statement. Indeed, if f is μ integrable, then $L^E(f)^*=L^E(\overline{f})$, though E is not projection valued. Finally, we recall that $L^E(f)=L^E(f)^{**}$ holds if and only if $L^E(f)$ is densely defined and closed.

D. Symmetric operator integrals

Assume that the function f is real valued. Then the operator $L^E(f)$ is symmetric, that is, for any $\varphi,\psi\in\mathcal{D}(L^E(f))$,

$$\langle\psi|L^E(f)\varphi\rangle=\langle L^E(f)\psi|\varphi\rangle,$$

see Ref. 1 for a proof. Therefore, if, in addition, $L^E(f)$ is densely defined one has the extensions

$$L^E(f)\subseteq L^E(f)^{**}\subseteq L^E(f)^*.$$

E. The real and imaginary parts

Let $f=f_1+if_2$ be the decomposition of f into its real and imaginary parts. Then f is $E_{\psi,\varphi}$ integrable if and only if f_1 and f_2 are $E_{\psi,\varphi}$ integrable, in which case $\int f\,dE_{\psi,\varphi}=\int f_1\,dE_{\psi,\varphi}+i\int f_2\,dE_{\psi,\varphi}$. We thus have the operator equalities

$$L^E(f)=L^E(f_1)+iL^E(f_2),$$

$$L^E(\overline{f})=L^E(f_1)-iL^E(f_2),$$

with $\mathcal{D}(L^E(f))=\mathcal{D}(L^E(\overline{f}))=\mathcal{D}(L^E(f_1))\cap\mathcal{D}(L^E(f_2))$. Assume now that $\overline{\mathcal{D}(L^E(f))}=\mathcal{H}$. Then $L^E(\overline{f})\subset L^E(f)^*$ so that the operators $\alpha L^E(f)+\beta L^E(\overline{f})$ and $\alpha L^E(f)+\beta L^E(f)^*$ are equal for all nonzero α, $\beta\in\mathbf{C}$, and, in particular,

$$\mathrm{Re}\,L^E(f):=\tfrac{1}{2}[L^E(f)+L^E(f)^*]=\tfrac{1}{2}[L^E(f)+L^E(\overline{f})],$$

$$\mathrm{Im}\,L^E(f):=\frac{1}{2i}[L^E(f)-L^E(f)^*]=\frac{1}{2i}[L^E(f)-L^E(\overline{f})].$$

By definition, $\mathcal{D}(\mathrm{Re}(L^E(f)))=\mathcal{D}(L^E(f))\subseteq\mathcal{D}(L^E(\mathrm{Re}\,f))$, and for any $\psi\in\mathcal{H}$, $\varphi\in\mathcal{D}(L^E(f))$,

$$\langle\psi|\mathrm{Re}(L^E(f))\varphi\rangle=\langle\psi|\tfrac{1}{2}(L^E(f)+L^E(\overline{f}))\varphi\rangle=\langle\psi|L^E(\tfrac{1}{2}(f+\overline{f}))\varphi\rangle=\langle\psi|L^E(\mathrm{Re}\,f)\varphi\rangle,$$

which shows that $\mathrm{Re}(L^E(f))\subseteq L^E(\mathrm{Re}\,f)$. Since the functions $\mathrm{Re}\,f$ and $\mathrm{Im}\,f$ are real the operators $L^E(\mathrm{Re}\,f)$ and $L^E(\mathrm{Im}\,f)$, as well as their restrictions $\mathrm{Re}\,L^E(f)$ and $\mathrm{Im}\,L^E(f)$, are symmetric.

III. INTEGRATION VIA DILATION

A. Neumark dilation

Consider the (positive normalized) operator measure $E:\mathcal{A}\to\mathcal{L}(\mathcal{H})$ and let $F:\mathcal{A}\to\mathcal{L}(\mathcal{K})$ be a (Neumark) dilation of it into a projection measure acting on a Hilbert space $\mathcal{K}$. Let $V:\mathcal{H}\to\mathcal{K}$ be the isometry such that

$$E(X)=V^*F(X)V$$

for all $X\in\mathcal{A}$. For any $\mathcal{A}$-measurable $f:\Omega\to\mathbf{C}$ the operator $L^F(f)=\int f\,dF$ is densely defined with the domain $\mathcal{D}(L^F(f))=\mathcal{D}_f^F$. The question arises under which conditions the ''projected'' operator $\mathrm{pr}(L^F(f)):=V^*L^F(f)V$ equals the operator $L^E(f)$. This will be answered next.

Theorem: With the above notations, the operator $V^*L^F(f)V$ is the restriction of the operator $L^E(f)$ to $\mathcal{D}_f^E$, so that $V^*L^F(f)V=L^E(f)$ if and only if $\mathcal{D}_f^E=\mathcal{D}(L^E(f))$.

Proof: For any $\varphi,\psi\in\mathcal{H}$ the complex measures $E_{\psi,\varphi}$ and $F_{V\psi,V\varphi}$ are the same. Hence, for any $\varphi\in\mathcal{D}_f^E$,

$$\int |f|^2 dF_{V\varphi,V\varphi}=\int |f|^2 dE_{\varphi,\varphi}<\infty,$$

showing that $V(\mathcal{D}_f^E)\subseteq\mathcal{D}_f^F=\mathcal{D}(L^F(f))$. But then for any $\varphi\in\mathcal{D}_f^E$ and for each $\psi\in\mathcal{H}$,

$$\langle\psi|L^E(f)\varphi\rangle=\int f\,dE_{\psi,\varphi}=\int f\,dF_{V\psi,V\varphi}=\langle V\psi|L^F(f)V\varphi\rangle=\langle\psi|V^*L^F(f)V\varphi\rangle.$$

This means that

$$L^E(f)|_{\mathcal{D}_f^E}=V^*L^F(f)V|_{\mathcal{D}_f^E}.$$

But we have

$$\begin{aligned}\mathcal{D}(V^*L^F(f)V)&=\mathcal{D}(L^F(f)V)=\{\varphi\in\mathcal{H}|V\varphi\in\mathcal{D}(L^F(f))\}\\&=\left\{\varphi\in\mathcal{H}\,\middle|\,\int |f|^2 dF_{V\varphi,V\varphi}<\infty\right\}=\left\{\varphi\in\mathcal{H}\,\middle|\,\int |f|^2 dE_{\varphi,\varphi}<\infty\right\}\\&=\mathcal{D}_f^E\subseteq\mathcal{D}(L^E(f)).\end{aligned}$$

Therefore, $V^*L^F(f)V$ is the restriction of $L^E(f)$ to $\mathcal{D}_f^E$, and $V^*L^F(f)V=L^E(f)$ if and only if $\mathcal{D}_f^E=\mathcal{D}(L^E(f))$.

Remark: For compactly supported operator measures E and for functions f which are bounded on the support of the measure, the equality $V^*L^F(f)V=L^E(f)$ is already proved in Ref. 8. The authors are indebted to an anonymous referee for bringing this reference to their attention. In the general case, however, the inclusion $\mathcal{D}_f^E\subset\mathcal{D}(L^E(f))$ may be proper, as pointed out in Sec. II A.

B. A theorem of Foiaş

As an application of the above result we obtain a slight generalization of a theorem of Foiaş[3] which characterizes subnormal operators in terms of operator measures. To that end, assume now that E is defined on the σ-algebra $\mathcal{B}(\mathbf{C})$ of Borel subsets of the complex plane $\mathbf{C}$. Let F and V be as in Sec. III A and choose f to be the identity function id on $\mathbf{C}$. Let A stand for the operator $\int z\,dE(z)$. The operator $N:=\int z\,dF(z)$ is now normal and by the proof of the above theorem we have $V^*NV=A|_{\mathcal{D}_{\mathrm{id}}^E}$, where $\mathcal{D}_{\mathrm{id}}^E=\{\varphi\in\mathcal{H}|\int|z|^2 dE_{\varphi,\varphi}<\infty\}$. We also recall[9] that a (not necessarily densely defined) operator $B:\mathcal{D}(B)\to\mathcal{H}$ is said to be subnormal if there is a Hilbert space $\mathcal{K}$, a linear isometry $V:\mathcal{H}\to\mathcal{K}$, and a normal operator $N:\mathcal{D}(N)\to\mathcal{K}$ such that $\mathcal{D}(B)\subseteq V^*(V(\mathcal{H})\cap\mathcal{D}(N))$, $N(V(\mathcal{D}(B)))\subseteq V(\mathcal{H})$, and $B\varphi=V^*NV\varphi$ for all $\varphi\in\mathcal{D}(B)$.

Theorem: An operator $B:\mathcal{D}(B)\to\mathcal{H}$ is subnormal if and only if there is a positive normalized operator measure $E:\mathcal{B}(\mathbf{C})\to\mathcal{L}(\mathcal{H})$ such that for each $\psi\in\mathcal{H}$, $\varphi\in\mathcal{D}(B)$, the following two conditions are satisfied:

$$\text{(a)}\qquad \langle\psi|B\varphi\rangle=\int z\,dE_{\psi,\varphi},$$

$$\text{(b)}\qquad \|B\varphi\|^2=\int |z|^2 dE_{\varphi,\varphi}.$$

Proof: Assume that B is subnormal, and let $N:\mathcal{D}(N)\to\mathcal{K}$ be a normal extension of it on the Hilbert space $\mathcal{K}$, with $V:\mathcal{H}\to\mathcal{K}$ being the accompanying isometry as described above. Let $F:\mathcal{B}(\mathbf{C})\to\mathcal{L}(\mathcal{K})$ be the spectral measure of N. Then $E(Z):=V^*F(Z)V$ defines a positive normalized operator measure $E:\mathcal{B}(\mathbf{C})\to\mathcal{L}(\mathcal{H})$ which satisfies for all $\psi\in\mathcal{H}$, $\psi\in\mathcal{D}(B)$, the conditions:

$$\begin{aligned}\langle\psi|B\varphi\rangle&=\langle\psi|V^*NV\varphi\rangle=\langle V\psi|NV\varphi\rangle\\&=\int z\,dF_{V\psi,V\varphi}=\int z\,dE_{\psi,\varphi},\end{aligned}$$

$$\begin{aligned}\|B\varphi\|^2=\|V^*NV\varphi\|^2&=\langle V^*NV\varphi|V^*NV\varphi\rangle\\&=\langle NV\varphi|VV^*NV\varphi\rangle\\&=\langle NV\varphi|NV\varphi\rangle\\&=\int |z|^2\,dF_{V\varphi,V\varphi}=\int |z|^2\,dE_{\varphi,\varphi},\end{aligned}$$

where we have used the fact that VV^* is the projection onto $V(\mathcal{H})$ and that $N(V(\mathcal{D}(B)))\subseteq V(\mathcal{H})$.

Conversely, assume that there is a normalized positive operator measure E associated with B via the formulas (a) and (b). We observe first that these conditions imply that B is a restriction of $A|_{\mathcal{D}^E_{\rm id}}$, that is, $B\subseteq A|_{\mathcal{D}^E_{\rm id}}$, where A denotes $\int z\,dE$. Let $F:\mathcal{B}(\mathbf{C})\to\mathcal{L}(\mathcal{K})$ be a dilation of E into a projection measure, $N=\int z\,dF$, and let $V:\mathcal{H}\to\mathcal{K}$ be the associated isometry. Then, by the theorem of Sec. II A, for any $\varphi\in\mathcal{D}(B)$,

$$B\varphi=A\varphi=V^*NV\varphi,$$

and

$$\|B\varphi\|^2=\int |z|^2\,dE_{\varphi,\varphi}=\int |z|^2\,dF_{V\varphi,V\varphi}=\|NV\varphi\|^2.$$

Since $\|NV\varphi\|=\|B\varphi\|=\|V^*NV\varphi\|=\|VV^*NV\varphi\|$ for all $\varphi\in\mathcal{D}(B)$, and VV^* is the projection onto $V(\mathcal{H})$, this gives $VV^*(NV\varphi)=NV\varphi$ for all $\varphi\in\mathcal{D}(B)$, that is, $N(V(\mathcal{D}(B)))\subseteq V(\mathcal{H})$, which concludes the proof that B is subnormal.

Let E,F,V,f be as in Sec. III A. We then have $V^*L^F(\mathrm{Re}\,f)V\subseteq L^E(\mathrm{Re}\,f)$, with the equality sign if and only if $\mathcal{D}^E_{\mathrm{Re}\,f}=\mathcal{D}(L^E(\mathrm{Re}\,f))$. Moreover,

$$\begin{aligned}\mathrm{Re}(V^*L^F(f)V)&=\tfrac{1}{2}[V^*L^F(f)V+(V^*L^F(f)V)^*]\\&=\tfrac{1}{2}[V^*L^F(f)V+V^*L^F(f)^*V)]\\&=V^*\,\mathrm{Re}(L^F(f))V\subseteq V^*L^F(\mathrm{Re}\,f)V,\end{aligned}$$

where the second equality is due to the fact that $(V^*L^F(f)V)^* \supseteq V^*L^F(f)^*V$ and $\mathcal{D}(V^*L^F(f)V) = \mathcal{D}(V^*L^F(f)^*V) \subseteq \mathcal{D}((V^*L^F(f)V)^*)$.

IV. EXAMPLES

In this section we apply the above results to some phase space observables and their Cartesian margins. From now on the Hilbert space $\mathcal{H}$ will be separable.

A. The phase space observable $A_{|s\rangle}$

For any $\varphi,\psi\in\mathcal{H}$ we let $|\varphi\rangle\langle\psi|$ denote the rank one operator $|\varphi\rangle\langle\psi|(\eta) = \langle\psi|\eta\rangle\varphi$. Let $\{|n\rangle\}_{n\geq 0}$ be a fixed orthonormal basis of $\mathcal{H}$. Let $N=\Sigma_{n\geq 0} n|n\rangle\langle n|$, $a=\Sigma_{n\geq 0}\sqrt{n+1}|n\rangle\langle n+1|$, and $a^* = \Sigma_{n\geq 0}\sqrt{n+1}|n+1\rangle\langle n|$ be the associated number, lowering, and raising operators, respectively. Their domains are $\mathcal{D}(N)=\{\varphi\in\mathcal{H}|\Sigma_{n\geq 0}n^2|\langle n|\varphi\rangle|^2<\infty\}$, and $\mathcal{D}(a)=\mathcal{D}(a^*)=\{\varphi\in\mathcal{H}|\Sigma_{n\geq 0}n|\langle n|\varphi\rangle|^2<\infty\}$. Clearly, $a^*a=N$ and $aa^*=N+I$. Moreover, $\mathrm{Re}(a)$ and $\mathrm{Im}(a)$ are (densely defined) symmetric operators whose closures are, apart from a scaling factor, the canonical position and momentum operators Q and P, respectively, that is, $\sqrt{2}\,\mathrm{Re}(a)^{**}=Q$, $\sqrt{2}\,\mathrm{Im}(a)^{**}=P$.

Let $\lambda:\mathcal{B}(\mathbf{C})\to[0,\infty]$ be the two-dimensional Lebesgue measure. Let $D_z=e^{za^*-\bar{z}a}$, $z\in\mathbf{C}$, be the (unitary) shift operator, and let $|s\rangle$ be a fixed number state. The map $\mathbf{C}\ni z\mapsto D_z|s\rangle\in\mathcal{H}$ is norm-continuous and $1/\pi\int_{\mathbf{C}}D_z|s\rangle\langle s|D_z^*d\lambda(z)=I$, where the integral converges in the weak operator topology. The formula

$$A_{|s\rangle}(Z) := \frac{1}{\pi}\int_Z D_z|s\rangle\langle s|D_z^* d\lambda(z)\in\mathcal{L}(\mathcal{H}),\quad Z\in\mathcal{B}(\mathbf{C}),$$

thus defines a normalized positive operator measure, the phase space observable $A_{|s\rangle}$ associated with the number state $|s\rangle$. For further details on the phase space observables and their dilations (Sec. IV B) the reader may wish to consult Refs. 10 and 11.

B. A dilation of $A_{|s\rangle}$

Let $\mathcal{K}$ denote the Hilbert space $L^2(\mathbf{C},(1/\pi)d\lambda)$ and let $F:\mathcal{B}(\mathbf{C})\to\mathcal{L}(\mathcal{K})$ be the canonical spectral measure $(F(Z)\phi=\chi_Z\phi)$. Consider the mapping $V_{|s\rangle}:\mathcal{H}\to\mathcal{K}$, with

$$(V_{|s\rangle}\varphi)(z) := \langle s|D_z^*\varphi\rangle_{\mathcal{H}},\quad z\in\mathbf{C}.$$

$V_{|s\rangle}$ is linear and preserves the norm:

$$\|V_{|s\rangle}\varphi\|_{\mathcal{K}}^2=\frac{1}{\pi}\int_{\mathbf{C}}\overline{(V_{|s\rangle}\varphi)(z)}(V_{|s\rangle}\varphi)(z)d\lambda=\frac{1}{\pi}\int_{\mathbf{C}}\langle\varphi|D_z|s\rangle\langle s|D_z^*\varphi\rangle d\lambda=\|\varphi\|_{\mathcal{H}}^2$$

for any $\varphi\in\mathcal{H}$. The final projection of this isometry $V_{|s\rangle}$, that is, the projection onto the closed subspace $V_{|s\rangle}(\mathcal{H})=\{V_{|s\rangle}\varphi|\varphi\in\mathcal{H}\}\subset\mathcal{K}$, is easily seen to be $P_{|s\rangle}=\Sigma_{n=0}^{\infty}P[\phi_n^s]$, where $\phi_n^s=V_{|s\rangle}|n\rangle$, $n\geq 0$.

For any $\varphi,\psi\in\mathcal{H}$, and for each $Z\in\mathcal{B}(\mathbf{C})$,

$$\langle\varphi|A_{|s\rangle}(Z)\psi\rangle=\langle V_{|s\rangle}\varphi|F(Z)V_{|s\rangle}\psi\rangle,$$

which shows that the projection measure F is a dilation of the operator measure $A_{|s\rangle}$.

Let $L^F(f)=\int_{\mathbf{C}}f(z)dF(z)$ be the normal operator defined by the canonical projection measure $F(Z)\phi=\chi_Z\phi$ and the Borel function $f:\mathbf{C}\to\mathbf{C}$. Then $L(f)$ is a multiplication operator, $(L^F(f)\phi)(z)=f(z)\phi(z)$, $z\in\mathbf{C}$, $\phi\in\mathcal{D}(L^F(f))$. By the theorem of Sec. III A we now have that $V_{|s\rangle}^*L^F(f)V_{|s\rangle}\subseteq L^{A_{|s\rangle}}(f)$ and the domain of $V_{|s\rangle}^*L^F(f)V_{|s\rangle}$ is $D_f^{A_{|s\rangle}}$.

C. A completeness relation

For each $s=0,1,2,...$ the functions ϕ_n^s, $n=0,1,2,...$, are mutually orthogonal unit vectors. We shall proceed to show that the set $\{\phi_n^s|s,n=0,1,2,...\}$ is a complete orthonormal system. The orthogonality of the functions ϕ_n^s can readily be computed. Indeed, these functions are related to the associated Laguerre polynomials as follows:

$$\begin{aligned}\phi_n^s(z)&=\langle s|D_z^*|n\rangle\\&=\frac{1}{\sqrt{s!}}e^{-|z|^2/2}\sum_{l=0}^{\min\{s,n\}}(-1)^{s-l}\binom{s}{l}\frac{\sqrt{n!}}{(n-l)!}z^{s-l}\overline{z}^{n-l}\\&=e^{i(s-n)\theta}(-1)^{\max\{0,s-n\}}e^{-|z|^2/2}\sqrt{\frac{(\min\{n,s\})!}{(\max\{n,s\})!}}|z|^{|s-n|}L_{\min\{n,s\}}^{|s-n|}(|z|^2),\end{aligned}$$

where $z=|z|e^{i\theta}$ and L_m^k, $k\in\mathbf{Z}$, $m\in\mathbf{N}$, is the associated Laguerre polynomial. Using the Vandermonde convolution formula[12] or the properties of the Laguerre polynomials it is a straightforward computation to confirm that for any $s,r,n,m\geqslant 0$

$$\langle\phi_n^s|\phi_m^r\rangle=\frac{1}{\pi}\int_{\mathbf{C}}\overline{\phi_n^s(z)}\phi_m^r(z)d\lambda=\frac{1}{\pi}\int_{\mathbf{C}}\langle n|D_z|s\rangle\langle r|D_z^*|m\rangle d\lambda=\delta_{sr}\delta_{nm}.$$

This shows that for $s\neq r$ the operator measures $A_{|s\rangle}$ and $A_{|r\rangle}$ are Neumark projections of F into mutually orthogonal subspaces $V_{|s\rangle}(\mathcal{H})$ and $V_{|r\rangle}(\mathcal{H})$ of $\mathcal{K}$. To show that the functions $\phi_n^s,s,n=0,1,2,...$, constitute a complete system we write

$$L^2\left(\mathbf{C},\frac{1}{\pi}d\lambda(z)\right)\simeq L^2\left([0,2\pi),\frac{d\theta}{2\pi}\right)\otimes L^2(\mathbf{R}^+,d(r^2)),$$

and recall that the functions $e^{ik\theta}$, $k\in\mathbf{Z}$, and

$$\sqrt{\frac{m!}{(l+m)!}}e^{-r^2/2}r^lL_m^l(r^2),$$

$l,m\in\mathbf{N}$, form orthonormal bases for the respective component spaces. We may thus conclude that

$$\sum_{s=0}^{\infty}P_{|s\rangle}=\sum_{s=0}^{\infty}\sum_{n=0}^{\infty}P[\phi_n^s]=I_{\mathcal{K}}.$$

In other words, we have the orthogonal decomposition of the bigger space

$$\mathcal{K}=L^2\left(\mathbf{C},\frac{1}{\pi}d\lambda(z)\right)=V_{|0\rangle}(\mathcal{H})\oplus V_{|1\rangle}(\mathcal{H})\oplus V_{|2\rangle}(\mathcal{H})\oplus\cdots.$$

D. The moment operators of $A_{|s\rangle}$

In Ref. 1 all the operator integrals $L^{A_{|s\rangle}}(f)$ for the functions $z\mapsto z^k$ and $z\mapsto\overline{z}^k$, with $k=0,1,2,...$, were determined. They are:

$$\text{(a)}\qquad L^{A_{|s\rangle}}(z^k)=\int_{\mathbf{C}}z^k dA_{|s\rangle}=a^k,$$

$$\text{(b)}\qquad L^{A_{|s\rangle}}(\overline{z}^k)=\int_{C}\overline{z}^k dA_{|s\rangle}=(a^*)^k,$$

with the domains

$$\mathcal{D}^{A_{|s\rangle}}_{\mathrm{id}^k}=\mathcal{D}(L^{A_{|s\rangle}}(z^k))=\mathcal{D}(a^k)=\mathcal{D}(a^{*k}).$$

Accordingly, we also have:

$$a^k=V^*_{|s\rangle}L^F(z^k)V_{|s\rangle}=V^*_{|s\rangle}L^F(z)^kV_{|s\rangle}=(V^*_{|s\rangle}L^F(z)V_{|s\rangle})^k,$$

$$(a^*)^k=V^*_{|s\rangle}L^F(\overline{z}^k)V_{|s\rangle}=V^*_{|s\rangle}(L^F(z)^*)^kV_{|s\rangle}=(V^*_{|s\rangle}L^F(z)^*V_{|s\rangle})^k,$$

where $(L^F(z)\phi)(z)=z\phi(z)$, $z\in\mathbf{C}$, for all $\phi\in\mathcal{D}(L^F(z))$.

In Ref. 1 it was shown, in addition, that for any $s=0,1,2,...$,

$$\text{(c)} \qquad \int_{\mathbf{C}}|z|^2dA_{|s\rangle}=N+(s+1)I.$$

Since $aa^*=N+I$ we observe that the operator integral (c) equals aa^* exactly when $s=0$. Indeed, the phase space measure $A_{|0\rangle}$, associated with the ground state $|0\rangle$, satisfies the condition

$$\text{(d)} \qquad a^k(a^*)^l=\int_{\mathbf{C}}z^k\overline{z}^l\,dA_{|0\rangle}$$

for $k,l=0,1$. This shows the well-known fact that the operator a^* is subnormal. Holevo[13] referred to the subnormality of a^* as an explanation of the validity of the formulas (a)–(d) for the operator measure $A_{|0\rangle}$. It is worth mentioning that the operator a is not subnormal[14] so that there is no positive operator measure $E:\mathcal{B}(\mathbf{C})\to\mathcal{L}(\mathcal{H})$ for which $a=\int_{\mathbf{C}}z\,dE$, $a^*=\int_{\mathbf{C}}\overline{z}\,dE$, and $a^*a=\int_{\mathbf{C}}|z|^2\,dE$. This is in accordance with formula (c) which gives $a^*a=\int|z|^2\,dA_{|s\rangle}-(s+1)I$.

E. The real and imaginary parts of $L^{A_{|s\rangle}}(z^k)$

The above results (a) and (b) give immediately

$$\mathrm{Re}\,L^{A_{|s\rangle}}(z^k)=\tfrac{1}{2}(L^{A_{|s\rangle}}(z^k)+L^{A_{|s\rangle}}(\overline{z}^k))=\tfrac{1}{2}(a^k+a^{*k}),$$

$$\mathrm{Im}\,L^{A_{|s\rangle}}(z^k)=\frac{1}{2i}(L^{A_{|s\rangle}}(z^k)-L^{A_{|s\rangle}}(\overline{z}^k))=\frac{1}{2i}(a^k-a^{*k}),$$

with the domain $\mathcal{D}(a^k)$. Writing $z=x+iy$ we also have that the operators $L^{A_{|s\rangle}}(x)$ and $L^{A_{|s\rangle}}(y)$ extend the operators $\frac{1}{2}(a+a^*)$ and $(1/2i)(a-a^*)$, respectively. Since $L^{A_{|s\rangle}}(x)$ and $L^{A_{|s\rangle}}(y)$ are densely defined symmetric operators, and since $\mathrm{Re}(a)^{**}=1/\sqrt{2}Q$ and $\mathrm{Im}(a)^{**}=1/\sqrt{2}P$, we conclude that

$$\frac{1}{2}(a+a^*)\subset L^{A_{|s\rangle}}(x)\subset L^{A_{|s\rangle}}(x)^{**}=\frac{1}{\sqrt{2}}Q,$$

$$\frac{1}{2i}(a-a^*)\subset L^{A_{|s\rangle}}(y)\subset L^{A_{|s\rangle}}(y)^{**}=\frac{1}{\sqrt{2}}P.$$

Similarly, one gets:

$$\tfrac{1}{2}[N+(s+1)I+\tfrac{1}{2}(a^2+a^{*2})]=\tfrac{1}{2}[L^{A_{|s\rangle}}(|z|^2)+\tfrac{1}{2}(a^2+a^{*2})]$$
$$\subset L^{A_{|s\rangle}}(x^2)\subset L^{A_{|s\rangle}}(x^2)^{**}$$
$$=\tfrac{1}{2}Q^2+\tfrac{1}{2}(s+\tfrac{1}{2})I,$$
$$\tfrac{1}{2}[N+(s+1)I-\tfrac{1}{2}(a^2+a^{*2})]=\tfrac{1}{2}[L^{A_{|s\rangle}}(|z|^2)-\tfrac{1}{2}(a^2+a^{*2})]$$
$$\subset L^{A_{|s\rangle}}(y^2)\subset L^{A_{|s\rangle}}(y^2)^{**}$$
$$=\tfrac{1}{2}P^2+\tfrac{1}{2}(s+\tfrac{1}{2})I,$$
$$\frac{1}{4i}(a^2-a^{*2})\subset L^{A_{|s\rangle}}(xy)\subset L^{A_{|s\rangle}}(xy)^{**}=\frac{1}{4}(QP+PQ).$$

Finally, we note that the operator measure $A_{|s\rangle}$ and its Cartesian marginal measures $X\mapsto A_{|s\rangle}(X\times\mathbf{R})=:A^x_{|s\rangle}(X)$ and $Y\mapsto A_{|s\rangle}(\mathbf{R}\times Y)=:A^y_{|s\rangle}(Y)$ carry the following respective noises:

$$R(A_{|s\rangle}):=L^{A_{|s\rangle}}(z^2)-L^{A_{|s\rangle}}(z)^2=O,$$
$$R(A^x_{|s\rangle}):=L^{A_{|s\rangle}}(x^2)-L^{A_{|s\rangle}}(x)^2=\tfrac{1}{2}(s+\tfrac{1}{2})I,$$
$$R(A^y_{|s\rangle}):=L^{A_{|s\rangle}}(y^2)-L^{A_{|s\rangle}}(y)^2=\tfrac{1}{2}(s+\tfrac{1}{2})I.$$

The fact that the positive operators $R(A^x_{|s\rangle})$ and $R(A^y_{|s\rangle})$ are nonzero reflects the fact that the margins $A^x_{|s\rangle}$ and $A^y_{|s\rangle}$ are not projection valued.[2] Recalling that the marginal measures $A^x_{|s\rangle}$ and $A^y_{|s\rangle}$ are—apart from the scaling factor $\sqrt{2}$—unsharp position and momentum observables, the above noise equations lead to the following s-dependent uncertainty product for these observables (with $q=\sqrt{2}x$ and $p=\sqrt{2}y$):

$$\begin{aligned}\mathrm{Var}(A^q_{|s\rangle},\varphi)\mathrm{Var}(A^p_{|s\rangle},\varphi)&=(\mathrm{Var}(Q,\varphi)+\langle\varphi|R(A^q_{|s\rangle})\varphi\rangle)(\mathrm{Var}(P,\varphi)+\langle\varphi|R(A^p_{|s\rangle})\varphi\rangle)\\&=(\mathrm{Var}(Q,\varphi)+(s+\tfrac{1}{2}))(\mathrm{Var}(P,\varphi)+(s+\tfrac{1}{2}))\\&=(\mathrm{Var}(Q,\varphi)\mathrm{Var}(P,\varphi)+(s+\tfrac{1}{2})(\mathrm{Var}(Q,\varphi)+\mathrm{Var}(P,\varphi))+(s+\tfrac{1}{2})^2\\&\geqslant(s+1)^2.\end{aligned}$$

For $s=0$, this result is well known, see, e.g., Ref. 11, p. 55.

[1] P. Lahti, M. Maczynski, and K. Ylinen, Rep. Math. Phys. **41**, 319–331 (1998).
[2] F. Riesz and B. Sz.-Nagy, *Functional Analysis* (Dover, New York, 1990).
[3] C. Foiaş, Rev. Roum. Math. Pures Appl. **7**, 241–282 (1962).
[4] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space* (Dover, New York, 1993).
[5] W. Mlak, *Hilbert Space and Operator Theory, PWN* (Polish Scientific Publishers, Warszawa, 1991).
[6] F. E. Schroeck, Jr., *Quantum Mechanics on Phase Space* (Kluwer, Dordrecht, 1996).
[7] M. Ozawa, Ann. Phys. (Leipzig) **257**, 65–83 (1997).
[8] M. Schreiber, Trans. Am. Math. Soc. **87**, 108–118 (1958).
[9] P. Halmos, *Hilbert Space Problem Book* (Springer, Berlin, 1985).
[10] E. B. Davies, *Quantum Theory of Open Systems* (Academic, London, 1976).
[11] W. Stulpe, *Classical Representations of Quantum Mechanics* (Wissenschaft & Technik, Berlin, 1997).
[12] J. Riordan, *Combinatorial Identities* (Wiley, New York, 1968).
[13] A. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).
[14] J. Stochel and F. H. Szafraniec, J. Operator Theory **14**, 31–55 (1985).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 4 APRIL 1999

On Verma bases for representations of $sl(n,\mathbb{C})$

K. N. Raghavan[a] and P. Sankaran[b]
SPIC Mathematical Institute, 92 G. N. Chetty Road, T. Nagar, Chennai 600 017, India

(Received 11 February 1998; accepted for publication 17 December 1998)

Conjectural bases for irreducible representation spaces of simple Lie algebras, due to Verma, have been described by Li, Moody, Nicolescu, and Patera. In this paper the basis conjecture for the algebras of type A_n is established. © *1999 American Institute of Physics.* [S0022-2488(99)02504-9]

I. INTRODUCTION

Conjectural bases for irreducible representation spaces of simple Lie algebras, due to Verma, are described in Ref. 1. The purpose of the present paper is to establish the basis conjecture for the algebras of type A_n.

The proof consists of two steps. Recall that isomorphism classes of irreducible representation spaces are in one-to-one correspondence with shapes of Young tableaus. The first step in the proof is to exhibit a ''natural'' bijective correspondence between the conjectural basis vectors of an irreducible representation space on the one hand and the standard tableaus of the corresponding shape on the other. From the well-known result that the number of such standard tableaus equals the dimension of the representation, we conclude that the basis vectors are right in number. It then remains only to prove that they are linearly independent, and this is step two of the proof.

The set up for step two is borrowed from Ref. 2. In fact, the linear independence can be deduced as a special case of the main Proposition 7.1 of Ref. 2. In the present paper, we deduce the linear independence as an immediate consequence of Proposition 2 below. Although the two propositions appear at first sight to be identical, they are not so, for the partial orders are different. The proof of linear independence in the present paper is much simpler than the proof in Ref. 2 even after specializing the latter to the case of type A_n. Referring to Ref. 2 for the proof of linear independence would greatly reduce the readability of the present paper.

Another strategy for a proof of the basis conjecture, quite different from the one of our proof, is outlined in Ref. 1 itself (see Sec. V of that paper). We believe the conjectures on which this strategy is based to be true, but our attempts at obtaining a proof along these lines have not met with success.

There naturally arises the question whether our method of proof can be generalized to cover the other types of algebras. A ''natural'' bijective correspondence between the conjectural basis vectors and standard tableaus presumably exists for the other types too, but we do not see it. Of course, even after getting hold of such a correspondence, one still has to establish the linear independence of conjectural basis vectors.

The organization of this paper is as follows. In Sec. II we recall the basis conjecture for the algebras of type A_n. The bijective correspondence between the basis vectors and the standard tableaus is the subject of Sec. III. The proof of linear independence is straight forward for the fundamental representations as pointed out in Sec. IV. The linear independence in the case of a non-fundamental representation is established in Sec. V.

II. THE VERMA CONJECTURE

Let $\mathfrak{g}=sl(n;\mathbb{C})$, the Lie algebra of $n\times n$ matrices of trace 0 over the complex numbers. When we speak of roots, weights, positive roots, and such other terms, it is to be understood that these

---

[a]Electronic mail: knr@smi.ernet.in
[b]Electronic mail: sankaran@smi.ernet.in

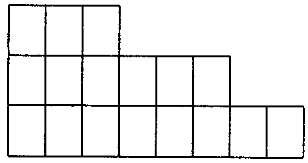

FIG. 1.

are with respect to the standard Cartan subalgebra (the subalgebra of traceless diagonal matrices) and the standard Borel subalgebra (the subalgebra of traceless upper triangular matrices).

Let $l:=n-1$, the *rank* of $\mathfrak{g}$. Let $\varpi_1,\ldots,\varpi_l$ be the fundamental weights. A weight λ is *dominant integral* if, in the expression $\lambda=m_1\varpi_1+\cdots+m_l\varpi_l$, the m_i are all non-negative integers. Recall that isomorphism classes of irreducible finite dimensional $\mathfrak{g}$-modules are in bijective correspondence with dominant integral weights. Given an irreducible $\mathfrak{g}$-module, the corresponding dominant integral weight is just the highest weight of that $\mathfrak{g}$-module.

Let $\lambda=m_1\varpi_1+\cdots+m_l\varpi_l$ be a dominant integral weight, V_λ the corresponding irreducible finite dimensional $\mathfrak{g}$-module, and v_λ a highest weight vector in V_λ. (The vector v_λ is uniquely determined up to a scalar factor.) Note that $s_1(s_2s_1)\cdots(s_i\cdots s_1)\cdots(s_l\cdots s_1)$ is a reduced expression for the longest element of the Weyl group of $\mathfrak{g}$; here s_i denotes the simple reflection corresponding to the simple root α_i which corresponds to the fundamental weight ϖ_i. Thus the vectors obtained by the action on v_λ of the monomials

$$Y_1^{(a_1^1)}(Y_2^{(a_2^2)}Y_1^{(a_1^2)})\cdots(Y_i^{(a_i^i)}Y_{i-1}^{(a_{i-1}^i)}\cdots Y_2^{(a_2^i)}Y_1^{(a_1^i)})\cdots(Y_l^{(a_l^l)}\cdots Y_1^{(a_1^l)}) \tag{1}$$

span V_λ; here Y_i is the lower triangular matrix with all but one entry being 0, the sole non-zero entry being 1 on row $i+1$ and column i, a_k^c are non-negative integers, and $Y_i^{(a)}$ denotes $Y_i^a/a!$. It was conjectured by Verma[1] that the subset of these vectors defined by the following constraint forms a basis for V_λ: for each pair (c,k) such that $1\leqslant k\leqslant c\leqslant l$,

$$0\leqslant a_k^c\leqslant \min\{a_{k-1}^c+m_{l-c+k},a_{k+1}^{c+1}\}, \tag{2}$$

where

$$a_k^{l+1}:=\infty \quad \text{and} \quad a_0^c:=0.$$

Let us call those monomials (1) that are subject to the constraint (2) the *Verma monomials*. Let us call *Verma vectors* those vectors obtained by the action of the Verma monomials on v_λ. The theorem of the present paper can now be phrased thus:

Theorem 1 (Verma's Conjecture): *The Verma vectors form a basis for* V_λ. The proof is given in Sec. V.

III. STANDARD TABLEAUS AND VERMA MONOMIALS

The purpose of this section is to establish a bijective correspondence between the set of standard tableaus of a fixed shape λ on the one hand and the Verma monomials associated to λ on the other. We first recall the notion of tableaus and standard tableaus.

Fix a positive integer l. Let $\lambda=(m_l,\ldots,m_1)$ be an l-tuple of non-negative integers. To λ we associate a *shape* as follows. The shape consists of boxes $lm_l+(l-1)m_{l-1}+\cdots+1\cdot m_1$ in number, left-justified and bottom-justified, with l boxes in the first m_l columns, $l-1$ boxes in the next m_{l-1} columns, and so on. For example, if $l=3$, the shape corresponding to $\lambda=(3,3,2)$ is shown in Fig. 1.

A *tableau* of shape λ is a filling up of the boxes in the shape associated to λ by integers $1,\ldots,n:=l+1$, such that the entries in each column are strictly increasing downwards. A tableau is

2	1	1					
3	2	3	3	2	3		
4	3	4	4	3	4	4	2

FIG. 2.

standard if the numbers in each row are non-increasing rightwards. For example, if $l=3$ and $\lambda=(3,3,2)$, the tableau in Fig. 2 is not standard while the one in Fig. 3 is standard.

In what follows, l and λ are fixed, and (c,k) denotes an ordered pair of integers satisfying $1\leqslant c\leqslant l$ and $1\leqslant k\leqslant c$. Given a standard tableau σ, define for each (c,k) the integer b_k^c to be the number of entries on row c of σ that are (strictly) greater than k, where the topmost row is row 1, the one just below it is row 2, and so on until the bottom-most row is row l. It is clear, from the definition of b_k^c, that $b_k^c\leqslant b_{k-1}^c$ and further, since σ is a tableau, that $b_k^c\leqslant b_{k+1}^{c+1}$. In other words, the b_k^c satisfy

$$b_k^c\leqslant \min\{b_{k-1}^c,b_{k+1}^{c+1}\},$$

$$b_k^c\leqslant m_l+\cdots+m_{l-c+1}. \tag{3}$$

Conversely, given integers b_k^c subject to the constraint (3), we can construct a standard tableau as follows. Fill up the shape corresponding to λ with integers so that in row c there are precisely b_k^c entries greater than k and the entries in each row are nonincreasing rightwards. There is clearly a unique way of doing this. The result is a tableau since $b_k^c\leqslant b_{k+1}^{c+1}$, and it is standard by construction. Thus the standard tableaus are in bijective correspondence with sets of integers $\{b_k^c\}$ subject to (3).

Setting $a_k^c:=b_k^c-(m_l+\cdots+m_{l-c+k+1})$, it is easily checked that (3) transforms to (2). Thus the Verma monomials are in bijective correspondence with $\{b_k^c\}$ subject to (3) and so also with standard tableaus. The Verma monomial corresponding to the standard tableau of Fig. 3 is

$$Y_1^{(1)}(Y_2^{(2)}Y_1^{(2)})(Y_3^{(2)}Y_2^{(3)}Y_1^{(1)}).$$

IV. THE CASE OF A FUNDAMENTAL REPRESENTATION

Let $V_1,\ldots,V_l$ denote the representations that correspond respectively to the fundamental weights $\varpi_1,\ldots,\varpi_l$. These representations admit the following simple description. The representation V_1 is the *standard representation* of $\mathfrak{g}=sl(n;\mathbb{C})$: the elements of V_1 are complex matrices of size $n\times 1$ and $\mathfrak{g}$ acts on V_1 by left multiplication. For $1\leqslant j\leqslant l$, the representation V_j is the exterior product $\wedge^j V_1$. Let e_i denote the $n\times 1$ matrix with the entry on row i being 1 and all other entries being 0. The vector $v_j:=e_1\wedge\cdots\wedge e_j$ is a highest weight vector in V_j.

A standard tableau σ of shape ϖ_j consists of a column of j integers $1\leqslant i_1<\cdots<i_j\leqslant n$ arranged in increasing order downwards. The Verma monomial corresponding to σ is

$$(Y_{i_1-1}\cdots Y_2Y_1)\cdots(Y_{i_{j-k+1}-1}\cdots Y_{j-k+2}Y_{j-k+1})\cdots(Y_{i_{j-1}-1}\cdots Y_jY_{j-1})(Y_{i_j-1}\cdots Y_{j+2}Y_{j+1}Y_j).$$

As can be readily checked, the action of this on v_j results in $e_{i_1}\wedge\cdots\wedge e_{i_j}$, which we denote $u(\sigma)$. Since the $u(\sigma)$ form a basis for V_j as σ varies over (standard) tableaus, Theorem 1 holds for a fundamental representation.

2	1	1					
3	3	2	2	2	1		
4	4	3	3	3	3	2	1

FIG. 3.

V. THE CASE OF A GENERAL REPRESENTATION

Let $\lambda = m_l \varpi_l + \cdots + m_1 \varpi_1$ be an arbitrary dominant integral weight. Let V_λ be the corresponding finite dimensional irreducible representation. We can realize V_λ as follows. Consider the tensor product representation

$$V := \underbrace{V_l \otimes \cdots \otimes V_l}_{m_l \text{ times}} \otimes \cdots \otimes \underbrace{V_1 \otimes \cdots \otimes V_1}_{m_1 \text{ times}}.$$

The element

$$v_\lambda := \underbrace{v_l \otimes \cdots \otimes v_l}_{m_l \text{ times}} \otimes \cdots \otimes \underbrace{v_1 \otimes \cdots \otimes v_1}_{m_1 \text{ times}}$$

of V then is a highest weight vector and the $\mathfrak{g}$-submodule of V generated by v_λ is a model for V_λ.

We think of a tableau ζ of shape λ as a concatenation $\zeta = (\zeta_1, \ldots, \zeta_m)$, where $m := m_l + \cdots + m_1$; $\zeta_1, \ldots, \zeta_{m_l}$ are tableaus of shape ϖ_l; $\zeta_{m_l+1}, \ldots, \zeta_{m_l+m_{l-1}}$ are tableaus of shape ϖ_{l-1}; and so on. We associate to ζ the vector $u(\zeta) := u(\zeta_1) \otimes \cdots \otimes u(\zeta_m)$ of V, where $u(\zeta_j)$ is the basis vector of $V_{r(j)}$ associated as in Sec. IV to ζ_j, where $r(j)$ is the number of rows in ζ_j. It is clear that as ζ varies over all tableaus of shape λ, standard or otherwise, the $u(\zeta)$ form a basis for V.

Proof of Theorem 1: For a fundamental weight λ, the proof was given in Sec. IV. Now let λ be a nonfundamental weight, σ be a standard tableau of shape λ, and $v(\sigma)$ be the corresponding Verma vector in V_λ. Thinking of $v(\sigma)$ as a vector in V, we can express it as a sum of the basis vectors $u(\zeta)$, ζ varying over all tableaus. Proposition 2 below is an assertion about this expression. It follows immediately from the proposition that the Verma vectors $v(\sigma)$ are linearly independent and therefore also that they form a basis (since the number of standard tableaus is well-known to equal the dimension of V_λ). □

To state Proposition 2, we introduce the following partial order on the set of all tableaus of shape λ. Given a tableau ζ of shape λ, let $\zeta(i,j)$ denote the entry of ζ on row i and column j, where the rows and columns are indexed as for a matrix—of course, in the case of a tableau as opposed to that of a matrix, the range of possible values for j depends on the value of i and vice versa. We call those pairs (i,j) *admissible* for which $\zeta(i,j)$ makes sense. For admissible pairs (i,j) and (i',j'), we say $(i,j) \leqslant (i',j')$ if either $i < i'$ or $i = i'$ and $j \leqslant j'$. For tableaus ζ, ζ' of shape λ, we say $\zeta \leqslant \zeta'$ if $\zeta(i,j) \leqslant \zeta'(i,j)$ for the smallest such pair (i,j) that $\zeta(i,j) \neq \zeta'(i,j)$.

There clearly exists a smallest tableau of a given shape. The smallest tableau of shape (3,3,2) is shown in Fig. 4. The exponents a_k^c in the Verma monomial of the smallest tableau are all 0. For

1	1	1					
2	2	2	1	1	1		
3	3	3	2	2	2	1	1

FIG. 4.

a standard tableau σ, the relation between the numbers b_k^c and a_k^c defined in Sec. III may be rephrased thus

$$a_k^c(\sigma)=b_k^c(\sigma)-b_k^c(\tau), \tag{4}$$

where τ is the smallest tableau of the same shape as σ.

Proposition 2: With notation as above, one has

$$v(\sigma)=u(\sigma)+\sum_{\zeta\lessdot\sigma} q(\zeta)u(\zeta),$$

where $q(\zeta)$ are some scalars (in fact, integers).

Proof: We proceed by an induction on the order $\leqslant$ defined above. If σ is the smallest tableau of shape λ with respect to $\leqslant$, then $v(\sigma)=u(\sigma)=v_\lambda$ and the statement is clear.

So suppose that σ is not the smallest tableau. Let a_k^c be the exponents in the Verma monomial corresponding to σ. Let C be the smallest such value of c that $a_k^c\neq 0$ for some k. Let K be the largest such value that $a_K^C\neq 0$. Changing a_K^C to 0 but leaving the remaining a_k^c unchanged, we get another Verma monomial, the standard tableau corresponding to which we denote by μ. It is readily checked that μ and σ satisfy the following:

(1) In positions $(i,j)\leqslant(C,m_l+\cdots+m_{l-C+K})$, the entries of μ are equal to the corresponding entries of the smallest tableau [see Eq. (4)]. In particular, for $(i,j)\leqslant(C,m_l+\cdots+m_{l-C+K+1})$, if $\mu(i,j)=K$, then $i<C$ and $\mu(i+1,j)=K+1$.

(2) To go from μ to σ the only change that needs to be made is to replace K by $K+1$ in certain boxes, a_K^C in number, on row C. The boxes figuring in these changes occur in consecutive columns starting with the column $m_l+\cdots+m_{l-C+K+1}+1$. In each of these columns of μ, K occurs but not $K+1$.

Note that $\mu\lessdot\sigma$. The induction hypothesis applied to μ says

$$v(\mu)=u(\mu)+\sum_{\eta\lessdot\mu} q(\eta)u(\eta)$$

and so

$$v(\sigma)=Y_K^{(p)}u(\mu)+\sum_{\eta\lessdot\mu} q(\eta)Y_K^{(p)}u(\eta),$$

where we have written p for a_K^C for simplicity of notation.

We now investigate $Y_K^{(p)}u(\eta)$ for a general tableau η. An entry of η is called *marked* if it equals K and is either on the last row or the entry immediately below it is not $K+1$. Let $S(\eta,K)$ be the set of marked positions of η, and let $Z(\eta):=Z(\eta,K,p)$ be the subsets of cardinality p of $S(\eta,K)$. For t an element of $Z(\eta)$, let η_t be the tableau obtained by changing the entries of η from K to $K+1$ in those marked positions that belong to t. We have

$$Y_K^{(p)}u(\eta)=\sum_{t\in Z(\eta)} u(\eta_t)$$

and so

$$v(\sigma)=\sum_{t\in Z(\mu)} u(\mu_t)+\sum_{\eta\lessdot\mu}\ \sum_{t\in Z(\eta)} q(\eta)u(\eta_t). \tag{5}$$

Let t and t' be sets of cardinality p of positions (i,j) admissible for the shape λ. We say $t \leqslant t'$ if $t_r \lessdot t'_r$ for the least such r that $t_r \neq t'_r$, where $t=\{t_1 \lessdot \cdots \lessdot t_p\}$ and $t'=\{t'_1 \lessdot \cdots \lessdot t'_p\}$ are arranged in increasing order.

Let η,ζ be tableaus of shape λ. Let $s \in Z(\eta)$ and $t \in Z(\zeta)$. The following statements are easily seen to be true:

(a) If $\eta \leqslant \zeta$ and $s \geqslant t$, then $\eta_s \leqslant \zeta_t$.
(b) If $\eta \leqslant \zeta$, $s \geqslant t$, and either $\eta \lessdot \zeta$ or $s \gtrdot t$, then $\eta_s \lessdot \zeta_t$.

Let $\eta \leqslant \mu$. It is clear that statement (1) above holds for η because it holds for μ. Thus every element s of $Z(\eta)$ satisfies

$$s \geqslant s_0 := \{(C, m_l + \cdots + m_{l-C+K+1} + 1), \ldots, (C, m_l + \cdots + m_{l-C+K+1} + p)\}.$$

By statement (2) above, we see that s_0 belongs to $Z(\mu)$ and that $\sigma = \mu_{s_0}$. Now by the statements (a) and (b) above, we obtain $\eta_s \leqslant \mu_{s_0} = \sigma$ and that equality holds if and only if $\eta = \mu$ and $s = s_0$. The proposition now follows from Equation (5). $\square$

ACKNOWLEDGMENTS

We thank Professor D.-N. Verma for bringing to our notice the existence of Ref. 1 and Professor R. V. Moody for providing us with a reprint of Ref. 1. This research was funded in part by a grant from the Department of Science and Technology, Government of India.

[1] Sai-Ping Li, R. V. Moody, M. Nicolescu, and J. Patera, ''Verma bases for representations of classical simple Lie algebras,'' J. Math. Phys. **27**, 668–677 (1986).
[2] K. N. Raghavan and P. Sankaran, ''A new approach to standard monomial theory for classical groups,'' Transformation Groups **3**, 59–75 (1998).

Erratum: "Information gain within generalized thermostatistics" [J. Math. Phys. 39, 6490 (1998)]

L. Borland, A. R. Plastino, and C. Tsallis
Centro Brasileiro de Pesquisas Fisicas, Rua Dr. Xavier Sigaud 150, 22290-180 Rio de Janeiro, RJ Brazil

(Received 31 December 1998; accepted for publication 20 January 1999)

[S0022-2488(99)02304-X]

(1) Equation (5) should read

$$K_q(p,p')=-\sum_j p_j \ln_q \frac{p'_j}{p_j}.$$

(2) Equation (7) should read

$$K_q(p,p')=\int dx \frac{p(x)^q}{1-q}(p(x)^{1-q}-p'(x)^{1-q})=-\int dx p(x) \ln_q\left(\frac{p'(x)}{p(x)}\right).$$

(3) Above Eq. (34) the sentence should read as follows:
...consequently the conditional entropy given y

$$\Sigma_q(x|y)\equiv-\sum_x p^q(x|y) \ln_q p(x|y) \tag{34}$$

(which is related to the conditional entropy averaged over the variable y through $S_q(x|y)=\Sigma_y p^q(y)\Sigma_q(x|y)$) must vanish. This case... .

(4) In Eqs. (35) and (36), $S_q(x|y)$ should be replaced by $\Sigma_q(x|y)$.

A. K. Rajagopal is sincerely thanked for discussions and in particular for directing our attention to the above points.

Erratum: ‘‘Toward a theory of the integer quantum Hall transition: Continuum limit of the Chalker–Coddington model’’ [J. Math. Phys. 38, 2007 (1997)]

Martin R. Zirnbauer
Universitat zu Koln, Institut fur Theoretische Physik, Zulpicher Str. 77, 50937 Koln, Germany

(Received 20 April 1998; accepted for publication 21 April 1998)

[S0022-2488(99)02404-4]

The value of the coupling constant $\sigma_{xx}^{(0)}$ computed in Sec. VII A is not correct. Its calculation was based on the reasonable but false assumption that the field configurations which dominate at large wavelengths are smooth. As a matter of fact, the relevant configurations contain a small high-momentum component due to the alternating structure of the Chalker–Coddington network. When this is taken into account, the value of the longitudinal conductivity changes to $\sigma_{xx}^{(0)}=1/2$ for the single-channel model and $\sigma_{xx}^{(0)}=N/2$ for the N-channel model. The rest of the paper is unaffected by this change.

$\mathcal{PT}$-symmetric quantum mechanics

Carl M. Bender
Department of Physics, Washington University, St. Louis, Missouri 63130

Stefan Boettcher
Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Maxico 87545 and CTSPS, Clark Atlanta University, Atlanta, Georgia 30314

Peter N. Meisinger
Department of Physics, Washington University, St. Louis, Missouri 63130

(Received 7 December 1998; accepted for publication 26 January 1999)

This paper proposes to broaden the canonical formulation of quantum mechanics. Ordinarily, one imposes the condition $H^{\dagger}=H$ on the Hamiltonian, where † represents the mathematical operation of complex conjugation and matrix transposition. This conventional Hermiticity condition is sufficient to ensure that the Hamiltonian H has a real spectrum. However, replacing this mathematical condition by the weaker and more physical requirement $H^{\ddagger}=H$, where ‡ represents combined parity reflection and time reversal $\mathcal{PT}$, one obtains new classes of complex Hamiltonians whose spectra are still real and positive. This generalization of Hermiticity is investigated using a complex deformation $H=p^2+x^2(ix)^{\epsilon}$ of the harmonic oscillator Hamiltonian, where ϵ is a real parameter. The system exhibits two phases: When $\epsilon\geqslant 0$, the energy spectrum of H is real and positive as a consequence of $\mathcal{PT}$ symmetry. However, when $-1<\epsilon<0$, the spectrum contains an infinite number of complex eigenvalues and a finite number of real, positive eigenvalues because $\mathcal{PT}$ symmetry is spontaneously broken. The phase transition that occurs at $\epsilon=0$ manifests itself in both the quantum-mechanical system and the underlying classical system. Similar qualitative features are exhibited by complex deformations of other standard real Hamiltonians $H=p^2+x^{2N}(ix)^{\epsilon}$ with N integer and $\epsilon>-N$; each of these complex Hamiltonians exhibits a phase transition at $\epsilon=0$. These $\mathcal{PT}$-symmetric theories may be viewed as analytic continuations of conventional theories from real to complex phase space.
[S0022-2488(99)00105-X]

I. INTRODUCTION

In a recent letter[1] a class of complex quantum-mechanical Hamiltonians of the form

$$H=p^2+x^2(ix)^{\epsilon}\quad(\epsilon\ \text{real})\tag{1.1}$$

was investigated. Despite the lack of conventional Hermiticity the spectrum of H is real and positive for all $\epsilon\geqslant 0$. As shown in Fig. 11 in this paper and Fig. 1 of Ref. 1, the spectrum is discrete and each of the energy levels increases as a function of increasing ϵ. We will argue below that the reality of the spectrum is a consequence of $\mathcal{PT}$ invariance.

The operator $\mathcal{P}$ represents parity reflection and the operator $\mathcal{T}$ represents time reversal. These operators are defined by their action on the position and momentum operators x and p:

$$\begin{aligned}\mathcal{P}&:x\to -x,\quad p\to -p,\\ \mathcal{T}&:x\to x,\quad p\to -p,\quad i\to -i.\end{aligned}\tag{1.2}$$

When the operators x and p are real, the canonical commutation relation $[x,p]=i$ is invariant under both parity reflection and time reversal. We emphasize that this commutation relation remains invariant under $\mathcal{P}$ and $\mathcal{T}$ even if x and p are complex provided that the above transformations hold. In terms of the real and imaginary parts of x and p, $x=\mathrm{Re}\,x+i\,\mathrm{Im}\,x$ and $p=\mathrm{Re}\,p+i\,\mathrm{Im}\,p$, we have

$$\begin{aligned}\mathcal{P}&:\mathrm{Re}\,x\to-\mathrm{Re}\,x,\quad \mathrm{Im}\,x\to-\mathrm{Im}\,x,\\ &\mathrm{Re}\,p\to-\mathrm{Re}\,p,\quad \mathrm{Im}\,p\to-\mathrm{Im}\,p,\\ \mathcal{T}&:\mathrm{Re}\,x\to\mathrm{Re}\,x,\quad \mathrm{Im}\,x\to-\mathrm{Im}\,x,\\ &\mathrm{Re}\,p\to-\mathrm{Re}\,p,\quad \mathrm{Im}\,p\to\mathrm{Im}\,p.\end{aligned}\tag{1.3}$$

While there is as yet no proof that the spectrum of H in Eq. (1.1) is real,[2] we can gain some insight regarding the reality of the spectrum of a $\mathcal{PT}$-invariant Hamiltonian H as follows: Note that eigenvalues of the operator $\mathcal{PT}$ have the form $e^{i\theta}$. To see this, let Ψ be an eigenfunction of $\mathcal{PT}$ with eigenvalue λ: $\mathcal{PT}\Psi=\lambda\Psi$. Recalling that $(\mathcal{PT})^2=1$, we multiply this eigenvalue equation by $\mathcal{PT}$ and obtain $\lambda^*\lambda=1$, where we have used the fact that $i\to-i$ under $\mathcal{PT}$. Thus, $\lambda=e^{i\theta}$. We know that if two linear operators commute, they can be simultaneously diagonalized. By assumption, the operator $\mathcal{PT}$ commutes with H. Of course, the situation here is complicated by the nonlinearity of the $\mathcal{PT}$ operator ($\mathcal{T}$ involves complex conjugation). However, let us suppose for now that the eigenfunctions ψ of H are simultaneously eigenfunctions of the operator $\mathcal{PT}$ with eigenvalue $e^{i\theta}$. Then applying $\mathcal{PT}$ to the eigenvalue equation $H\psi=E\psi$, we find that the energy E is real: $E=E^*$.

We have numerically verified the supposition that the eigenfunctions of H in Eq. (1.1) are also eigenfunctions of the operator $\mathcal{PT}$ when $\epsilon\geqslant0$. However, when $\epsilon<0$, the $\mathcal{PT}$ symmetry of the Hamiltonian is spontaneously broken; even though $\mathcal{PT}$ commutes with H, the eigenfunctions of H are *not* all simultaneously eigenfunctions of $\mathcal{PT}$. For these eigenfunctions of H the energies are complex. Thus, a transition occurs at $\epsilon=0$. As ϵ goes below 0, the eigenvalues as functions of ϵ pair off and become complex, starting with the highest-energy eigenvalues. As ϵ decreases, there are fewer and fewer real eigenvalues and below approximately $\epsilon=-0.57\,793$ only one real energy remains. This energy then begins to increase with decreasing ϵ and becomes infinite as ϵ approaches -1. In summary, the theory defined by Eq. (1.1) exhibits two phases, an unbroken-symmetry phase with a purely real energy spectrum when $\epsilon\geqslant0$ and a spontaneously-broken-symmetry phase with a partly real and partly complex spectrum when $\epsilon<0$.

A primary objective of this paper is to analyze the phase transition at $\epsilon=0$. We will demonstrate that this transition occurs in the classical as well as in the quantum theory. As a classical theory, the Hamiltonian H describes a particle subject to complex forces, and therefore the trajectory of the particle lies in the complex-x plane. The position and momentum coordinates of the particle are complex functions of t, a real time parameter. We are interested only in solutions to the classical equations of motion for which the energy of the particle is real. We will see that in the $\mathcal{PT}$-symmetric phase of the theory, the classical motion is periodic and is thus a complex generalization of a pendulum. We actually observe two kinds of closed classical orbits, one in which the particle oscillates between two complex turning points and another in which the particle follows a closed orbit. In many cases these closed orbits lie on an elaborate multisheeted Riemann surface. On such Riemann surfaces the closed periodic orbits exhibit remarkable knotlike topological structures. All of these orbits exhibit $\mathcal{PT}$ symmetry; they are left–right symmetric with respect to reflections about the imaginary-x axis in accordance with Eq. (1.3). In the broken-symmetry phase classical trajectories are no longer closed. Instead, the classical path spirals out to infinity. These spirals lack $\mathcal{PT}$ symmetry.

There have been many previous instances of non-Hermitian $\mathcal{PT}$-invariant Hamiltonians in physics. Energies of solitons on a *complex* Toda lattice have been found to be real.[3] Hamiltonians rendered non-Hermitian by an imaginary external field have been used to study population biology[4] and to study delocalization transitions such as vortex flux-line depinning in type II superconductors.[5] In these cases, initially real eigenvalues bifurcate into the complex plane due to the increasing external field, indicating the growth of populations or the unbinding of vortices.

The $\mathcal{PT}$-symmetric Hamiltonian considered in this paper has many generalizations: (i) Introducing a mass term of the form m^2x^2 yields a theory that exhibits several phase transitions; transitions occur at $\epsilon=-1$ and $\epsilon=-2$ as well as at $\epsilon=0$.[1] (ii) Replacing the condition of Hermiticity by the weaker constraint of $\mathcal{PT}$ symmetry also allows one to construct new classes of quasi-exactly solvable quantum theories.[6] (iii) In this paper we consider complex deformations of real Hamiltonians other than the harmonic oscillator. We show that Hamiltonians of the form

$$H=p^2+x^{2K}(ix)^{\epsilon} \tag{1.4}$$

have the same qualitative properties as H in Eq. (1.1). As ϵ decreases below 0, all of these theories exhibit a phase transition from an unbroken $\mathcal{PT}$-symmetric regime to a regime in which $\mathcal{PT}$ symmetry is spontaneously broken.

The Hamiltonian H in (1.1) is especially interesting because it can be generalized to quantum field theory. A number of such generalizations have recently been examined. The $\mathcal{PT}$-symmetric scalar field theory described by the Lagrangian[7]

$$\mathcal{L}=\tfrac{1}{2}(\partial\phi)^2+\tfrac{1}{2}m^2\phi^2+g\phi^2(i\phi)^{\epsilon}\quad(\epsilon\geqslant 0) \tag{1.5}$$

is intriguing because it is not invariant under parity reflection. This is manifested by a nonzero value of $\langle\phi\rangle$. It is interesting that this broken symmetry persists even when $\epsilon>0$ is an even integer.[7] The Hamiltonian for this theory is not Hermitian and, therefore, the theory is not unitary in the conventional sense. However, there is strong evidence that the spectrum for this theory is real and bounded below. For $\epsilon=1$ one can understand the positivity of the spectrum in terms of summability. The weak-coupling expansion for a conventional $g\phi^3$ theory is real, and apart from a possible overall factor of g, the Green's functions are formal power series in g^2. These series are not Borel summable because they do not alternate in sign. Nonsummability reflects the fact that the spectrum of the underlying theory is not bounded below. However, when we replace g by ig, the perturbation series remains real but now alternates in sign. Thus, the perturbation series becomes summable, and this suggests that the underlying theory has a real positive spectrum.

Replacing conventional $g\phi^4$ or $g\phi^3$ theories by $\mathcal{PT}$-symmetric $-g\phi^4$ or $ig\phi^3$ theories has the effect of reversing signs in the beta function. Thus, theories that are not asymptotically free become asymptotically free and theories that lack stable critical points develop such points. There is evidence that $-g\phi^4$ in four dimensions is nontrivial.[8]

Supersymmetric quantum field theory that is $\mathcal{PT}$ invariant has also been studied.[9] When we construct a two-dimensional supersymmetric quantum field theory by using a superpotential of the form $S(\phi)=-ig(i\phi)^{1+\epsilon}$, the supersymmetric Lagrangian resulting from this superpotential is

$$\begin{aligned}\mathcal{L}&=\tfrac{1}{2}(\partial\phi)^2+\tfrac{1}{2}i\bar{\psi}\partial\!\!\!/\psi+\tfrac{1}{2}S'(\phi)\bar{\psi}\psi+\tfrac{1}{2}[S(\phi)]^2\\&=\tfrac{1}{2}(\partial\phi)^2+\tfrac{1}{2}i\bar{\psi}\partial\!\!\!/\psi+\tfrac{1}{2}g(1+\epsilon)(i\phi)^{\epsilon}\bar{\psi}\psi-\tfrac{1}{2}g^2(i\phi)^{2+2\epsilon},\end{aligned} \tag{1.6}$$

where ψ is a Majorana spinor. The Lagrangian (1.3) has a broken parity symmetry. This poses the question, Does the parity violation induce a breaking of supersymmetry? To answer this question, both the ground-state energy E_0 and the fermion–boson mass ratio R were calculated as series in powers of the parameter ϵ. Through second order in ϵ, $E_0=0$ and $R=1$, which strongly suggests that supersymmetry remains unbroken. We believe that these results are valid to all orders in powers of ϵ. This work and our unpublished numerical studies of SUSY quantum mechanics show that complex deformations do not break supersymmetry.

Quantum field theories having the property of $\mathcal{PT}$ invariance exhibit other interesting features. For example, the Ising limit of a $\mathcal{PT}$-invariant scalar quantum field theory is intriguing because it is dominated by solitons rather than by instantons as in a conventional quantum field theory.[10] In addition, a model of $\mathcal{PT}$-invariant quantum electrodynamics has been studied.[11] The massless theory exhibits a stable, nontrivial fixed point at which the renormalized theory is finite. Moreover, such a theory allows one to revive successfully the original electron model of Casimir.

Since $\langle\phi\rangle \neq 0$ in $\mathcal{PT}$-symmetric theories, one can in principle calculate directly (using the Schwinger–Dyson equations, for example) the real positive Higgs mass in a renormalizable $\mathcal{PT}$-symmetric theory in which symmetry breaking occurs naturally. No symmetry-breaking parameter needs to be introduced. This most intriguing idea could lead to an experimental vindication of our proposed generalization of the notion of Hermiticity to $\mathcal{PT}$ symmetry.

This paper is organized as follows: In Sec. II we study the classical version of the Hamiltonian in Eq. (1.1). The behavior of classical orbits reveals the nature of the phase transition at $\epsilon = 0$. Next, in Sec. III we analyze the quantum version of this Hamiltonian. We derive several asymptotic results regarding the behavior of the energy levels near the phase transition. In Sec. IV we discuss the classical and quantum properties of the broad class of $\mathcal{PT}$-symmetric Hamiltonians in Eq. (1.4) of which H in Eq. (1.1) is a special case.

II. CLASSICAL THEORY

The classical equation of motion for a particle described by H in (1.1) is obtained from Hamilton's equations:

$$\frac{dx}{dt} = \frac{\partial H}{\partial p} = 2p,$$
$$\frac{dp}{dt} = -\frac{\partial H}{\partial x} = i(2+\epsilon)(ix)^{1+\epsilon}. \tag{2.1}$$

Combining these two equations gives

$$\frac{d^2x}{dt^2} = 2i(2+\epsilon)(ix)^{1+\epsilon}, \tag{2.2}$$

which is the complex version of Newton's second law, $F = ma$.

Equation (2.2) can be integrated once to give[12]

$$\frac{1}{2}\frac{dx}{dt} = \pm\sqrt{E + (ix)^{2+\epsilon}}, \tag{2.3}$$

where E is the energy of the classical particle (the time-independent value of H). We treat time t as a real variable that parametrizes the complex path $x(t)$ of this particle.

This section is devoted to studying and classifying the solutions to Eq. (2.3). By virtue of the $\mathcal{PT}$ invariance of the Hamiltonian H, it seems reasonable to restrict our attention to real values of E. Given this restriction, we can always rescale x and t by real numbers so that without loss of generality Eq. (2.3) reduces to

$$\frac{dx}{dt} = \pm\sqrt{1 + (ix)^{2+\epsilon}}. \tag{2.4}$$

The trajectories satisfying Eq. (2.4) lie on a multi-sheeted Riemann surface. On this surface the function $\sqrt{1+(ix)^{2+\epsilon}}$ is single valued. There are two sets of branch cuts. The cuts in the first set radiate outward from the roots of

$$1 + (ix)^{2+\epsilon} = 0. \tag{2.5}$$

These roots are the classical turning points of the motion. There are many turning points, all lying at a distance of unity from the origin. The angular separation between consecutive turning points

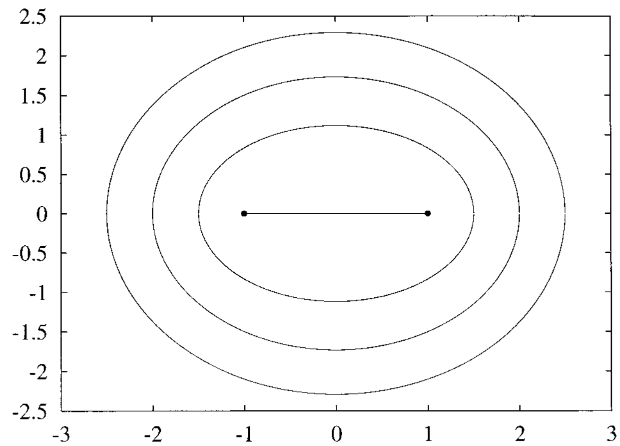

FIG. 1. Classical trajectories in the complex-x plane for the harmonic oscillator whose Hamiltonian is $H=p^2+x^2$. These trajectories represent the possible paths of a particle whose energy is $E=1$. The trajectories are nested ellipses with foci located at the turning points at $x=\pm 1$. The real line segment (degenerate ellipse) connecting the turning points is the usual periodic classical solution to the harmonic oscillator. All closed paths [see Eq. (2.6)] have the same period 2π.

is $2\pi/(2+\epsilon)$. The second set of branch cuts is present only when ϵ is noninteger. In order to maintain explicit $\mathcal{PT}$ symmetry (left–right symmetry in the complex-x plane), we choose these branch cuts to run from the origin to infinity along the positive imaginary axis.

A. Case $\epsilon=0$

Because the classical solutions to Eq. (2.4) have a very elaborate structure, we begin by considering some special values of ϵ. The simplest case is $\epsilon=0$. For this case there are only two turning points and these lie on the real axis at ± 1.

In order to solve Eq. (2.4) we need to specify an initial condition $x(0)$. The simplest choice for $x(0)$ is a turning point. If the path begins at ± 1, there is a unique direction in the complex-x plane along which the phases of the left side and the right side of Eq. (2.4) agree. This gives rise to a trajectory on the real axis that oscillates between the two turning points. This is the well-known sinusoidal motion of the harmonic oscillator.

Note that once the turning points have been fixed, the energy is determined. Thus, choosing the initial position of the particle determines the initial velocity (up to a plus or minus sign) as well. So, if the path of the particle begins anywhere on the real axis between the turning points, the initial velocity is fixed up to a sign and the trajectory of the particle still oscillates between the turning points.

Ordinarily, in conventional classical mechanics the only possible initial positions for the particle lie on the real-x axis between the turning points because the velocity is real; all other points on the real axis lie in the classically forbidden region. However, because we are analytically continuing classical mechanics into the complex plane, we can choose any point $x(0)$ in the complex plane as an initial position. For all complex initial positions outside of the conventional classically allowed region the classical trajectory is an ellipse whose foci are the turning points. The ellipses are nested because no trajectories may cross. (See Fig. 1.) The exact solution to Eq. (2.4) is

$$x(t)=\cos[\arccos x(0)\pm t], \tag{2.6}$$

where the sign of t determines the direction (clockwise or anticlockwise) in which the particle traces the ellipse. For *any* ellipse the period of the motion is 2π. The period is the same for all trajectories because we can join the square-root branch cuts emanating from the turning points,

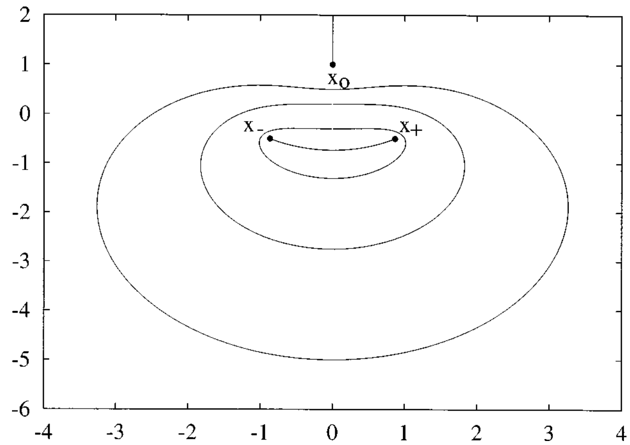

FIG. 2. Classical trajectories in the complex-x plane for a particle described by the Hamiltonian $H=p^2+ix^3$ and having energy $E=1$. An oscillatory trajectory connects the turning points $x_\pm$. This trajectory is enclosed by a set of closed, nested paths that fill the finite complex-x plane except for points on the imaginary axis at or above the turning point $x_0=i$. Trajectories originating at one of these exceptional points go off to $i\infty$ or else they approach x_0, stop, turn around, and then move up the imaginary axis to $i\infty$.

creating a single finite branch cut lying along the real axis from $x=-1$ to $x=1$. The complex path integral that determines the period can then be shrunk (by Cauchy's theorem) to the usual real integral joining the turning points.

Finally, we remark that all of the classical paths (elliptical orbits) are symmetric with respect to parity $\mathcal{P}$ (reflections through the origin) and time reversal $\mathcal{T}$ (reflections about the real axis), as well as $\mathcal{PT}$ (reflections about the imaginary axis). Furthermore, $\mathcal{P}$ and $\mathcal{T}$ individually preserve the directions in which the ellipses are traversed.

B. Case $\epsilon=1$

The case $\epsilon=1$ is significantly more complicated. Now there are three turning points. Two are located below the real axis and these are symmetric with respect to the imaginary axis: $x_-=e^{-5i\pi/6}$ and $x_+=e^{-i\pi/6}$. That is, under $\mathcal{PT}$ reflection x_- and x_+ are interchanged. The third turning point lies on the imaginary axis at $x_0=i$.

As in the case $\epsilon=0$, the trajectory of a particle that begins at the turning point x_- follows a unique path in the complex-x plane to the turning point at x_+. Then, the particle retraces its path back to the turning point at x_-, and it continues to oscillate between these two turning points. This path is shown on Fig. 2. The period of this motion is $2\sqrt{3\pi}\Gamma(\frac{4}{3})/\Gamma(\frac{5}{6})$. The periodic motion between $x_\pm$ is clearly time-reversal symmetric.

A particle beginning at the third turning point x_0 exhibits a completely distinct motion: It travels up the imaginary axis and reaches $i\infty$ in a finite time $\sqrt{\pi}\Gamma(\frac{4}{3})/\Gamma(\frac{5}{6})$. This motion is not periodic and is not symmetric under time reversal.

Paths originating from all other points in the finite complex-x plane follow closed periodic orbits. No two orbits may intersect; rather they are all nested, like the ellipses for the case $\epsilon=0$. All of these orbits encircle the turning points $x_\pm$ and, by virtue of Cauchy's theorem, have the same period $2\sqrt{3\pi}\Gamma(\frac{4}{3})/\Gamma(\frac{5}{6})$ as the oscillatory path connecting $x_\pm$. Because these orbits must avoid crossing the trajectory that runs up the positive imaginary axis from the turning point $x_0=i$, they are pinched in the region just below x_0, as shown on Fig. 2.

As these orbits become larger they develop sharper indentations in the vicinity of x_0. We observe that the characteristic radius of a large orbit approaches the reciprocal of the distance d between x_0 and the point where the orbit intersects the positive imaginary axis. Thus, it is appropriate to study these orbits from the point of view of the renormalization group: We scale the distance d down by a factor L and then plot the resulting orbit on a graph whose axis are scaled

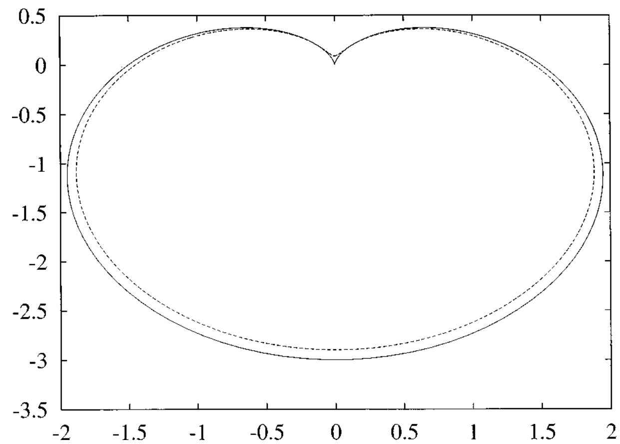

FIG. 3. Approach to the limiting cardioid in Eq. (2.8). As the orbits shown in Fig. 2 approach the turning point x_0, they get larger. Using a renormalization-group approach, we plot successively larger orbits (one such orbit is shown as a dashed line) scaled down by the characteristic size of the orbit. The limiting cardioid is indicated by a solid line. The indentation in the limiting cardioid develops because classical trajectories may not intersect and thus must avoid crossing the trajectory (shown in Fig. 2) on the imaginary axis above x_0.

down by the same factor L. Repeated scaling gives a limiting orbit whose shape resembles a cardioid (see Fig. 3). The equation of this limiting orbit is obtained in the asymptotic regime where we neglect the dimensionless energy 1 in Eq. (2.4):

$$\frac{dx}{dt}=\pm(ix)^{3/2}. \tag{2.7}$$

The solution to this differential equation, scaled so that it crosses the negative imaginary axis at $-3i$, is

$$x(t)=\frac{4i}{(t+2i/\sqrt{3})^2} \quad (-\infty<t<\infty). \tag{2.8}$$

This curve is shown as the solid line in Fig. 3. (Strictly speaking, this curve is not a true cardioid, but its shape so closely resembles a true cardioid that we shall refer to it in this paper as the *limiting* cardioid.)

In the infinite scaling limit all periodic orbits [all these orbits have period $2\sqrt{3\pi}\Gamma(\frac{4}{3})/\Gamma(\frac{5}{6})$], which originally filled the entire finite complex-x plane, have been squeezed into the region inside the limiting cardioid (2.8). The nonperiodic orbit still runs up the positive imaginary axis. The obvious question is, What complex classical dynamics is associated with all of the other points in the scaled complex-x plane that lie outside of the limiting cardioid? We emphasize that all of these points were originally at infinity in the unscaled complex-x plane.

We do not know the exact answer to this question, but we can draw a striking and suggestive analogy with some previously published work. It is generally true that the region of convergence in the complex-x plane for an infinitely iterated function is a cardioid-shaped region. For example, consider the continued exponential function

$$f(x)=e^{xe^{xe^x}}. \tag{2.9}$$

The sequence $e^x, e^{xe^x}, \ldots$ is known to converge in a cardioid-shaped region of the complex-x plane (see Figs. 2–4 in Ref. 13). It diverges on the straight line that emerges from the indentation of the cardioid. The remaining part of the complex-x plane is divided into an extremely elaborate mosaic of regions in which this sequence converges to limit cycles of period 2, period 3, period 4, and so

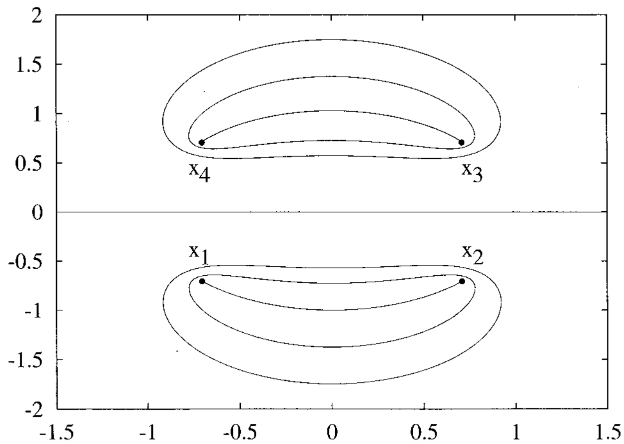

FIG. 4. Classical trajectories in the complex-x plane for a particle described by the Hamiltonian $H=p^2-x^4$ and having energy $E=1$. There are two oscillatory trajectories connecting the pairs of turning points x_1 and x_2 in the lower-half x-plane and x_3 and x_4 in the upper-half x-plane. [A trajectory joining any other pair of turning points is forbidden because it would violate $\mathcal{PT}$ (left-right) symmetry.] The oscillatory trajectories are surrounded by closed orbits of the same period. In contrast to these periodic orbits there is a class of trajectories having unbounded path length and running along the real-x axis. These are the only paths that violate time-reversal symmetry.

on. These regions have fractal structure. It would be interesting if unbounded complex classical motion exhibits this remarkable fractal structure. In other words, does the breaking of $\mathcal{P}$ and $\mathcal{T}$ symmetry allow for unbounded chaotic solutions?

C. Case $\epsilon=2$

When $\epsilon=2$ there are four turning points, two located below the real axis and symmetric with respect to the imaginary axis, $x_1=e^{-3i\pi/4}$ and $x_2=e^{-i\pi/4}$, and two more located above the real axis and symmetric with respect to the imaginary axis, $x_3=e^{i\pi/4}$ and $x_4=e^{3i\pi/4}$. Classical trajectories that oscillate between the pair x_1 and x_2 and the pair x_3 and x_4 are shown on Fig. 4. The period of these oscillations is $2\sqrt{2\pi}\Gamma(\frac{5}{4})/\Gamma(\frac{3}{4})$. Trajectories that begin elsewhere in the complex-x plane are also shown on Fig. 4. Note that by virtue of Cauchy's theorem all these nested nonintersecting trajectories have the same period. All motion is periodic except for trajectories that begin on the real axis; a particle that begins on the real-x axis runs off to $\pm\infty$, depending on the sign of the initial velocity. These are the only trajectories that are nonperiodic.

The rescaling argument that gives the cardioid for the case $\epsilon=1$ yields a doubly indented cardioid for the case $\epsilon=2$ (see Fig. 5). This cardioid is similar to that in Fig. 5 of Ref. 13. However, for the case $\epsilon=2$ the limiting double cardioid consists of two perfect circles, which are tangent to one another at the origin $x=0$. Circles appear because at $\epsilon=2$ in the scaling limit the equation corresponding to (2.7) is $dx/dt=\pm x^2$. The solutions to this equation are the inversions $x(t)=\pm[1/(t+i)]$, which map the real-t axis into circles in the complex-x plane.

D. Case $\epsilon=5$

When $\epsilon=5$ there are seven turning points, one located at i and three pairs, each pair symmetric with respect to reflection about the imaginary axis ($\mathcal{PT}$ symmetric). We find that each of these pairs of turning points is joined by an oscillatory classical trajectory. (A trajectory joining any other two turning points would violate $\mathcal{PT}$ symmetry.) Surrounding each of the oscillatory trajectories are nested closed loops, each loop having the same period as the oscillatory trajectory it encloses. These classical trajectories are shown on Fig. 6. The periods for these three families of trajectories are

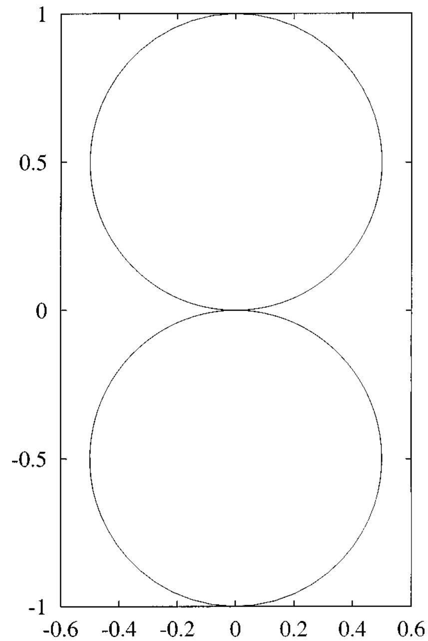

FIG. 5. Limiting double cardioid for the case $\epsilon=2$. As the orbits in Fig. 4 approach the real axis, they get larger. If we scale successively larger orbits down by their characteristic size, then in the limiting case the orbits approach two circles tangent at the origin. In this limit the four turning points in Fig. 4 coalesce at the point of tangency.

$$4\sqrt{\pi}\frac{\Gamma(8/7)}{\Gamma(9/14)}\cos\theta,$$

where $\theta=5\pi/14$ for the lowest pair of turning points, $\theta=\pi/14$ for the middle pair, and $\theta=3\pi/14$ for the pair above the real axis.

One other class of trajectory is possible. If the initial position of the classical particle lies on the imaginary axis at or above the turning point at i, then depending on the sign of the initial

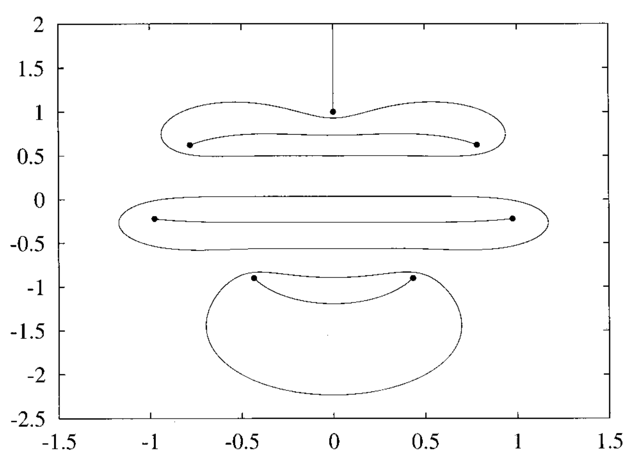

FIG. 6. Classical trajectories in the complex-x plane for a particle described by the Hamiltonian $H=p^2+ix^7$ and having energy $E=1$. Shown are oscillatory trajectories surrounded by periodic trajectories. Unbounded trajectories run along the positive-imaginary axis above $x=i$.

velocity, the particle either runs off to $i\infty$ or it approaches the turning point, reverses its direction, and then goes off to $i\infty$. These purely imaginary paths are the only possible nonperiodic trajectories. They are also shown on Fig. 6.

E. General case: Noninteger values of $\epsilon>0$

Because Eq. (2.4) contains a square root function, the turning points, which are solutions to Eq. (2.5), are square root branch points for all values of ϵ. Thus, in principle, the complex trajectories $x(t)$ lie on a multi-sheeted Riemann surface. However, when ϵ is a non-negative integer, we can define the branch cuts so that the classical trajectories satisfying Eq. (2.4) never leave the principal sheet of this Riemann surface. We do this as follows: We choose to join the $\mathcal{PT}$-symmetric (left–right-symmetric) pairs of turning points by branch cuts that follow exactly the oscillatory solutions connecting these pairs. (There are three such pairs in Fig. 6, two in Fig. 4, and one in Figs. 2 and 1.) If ϵ is odd, there is one extra turning point that lies on the positive imaginary axis (see Figs. 2 and 6); the branch cut emanating from this turning point runs up the imaginary-x axis to $i\infty$. Since classical paths never cross, there are no trajectories that leave the principal sheet of the Riemann surface.

When ϵ is noninteger, we can see from the argument of the square root function in Eq. (2.4) that there is an entirely new branch cut, which emerges from the origin in the complex-x plane. To preserve $\mathcal{PT}$ symmetry we choose this branch cut to run off to ∞ along the positive-imaginary x-axis. If ϵ is rational, the Riemann surface has a finite number of sheets, but if ϵ is irrational, then there are an infinite number of sheets.

If a classical trajectory crosses the branch cut emanating from the origin, then this trajectory leaves the principal sheet of the Riemann surface. In Fig. 7 we illustrate some of the possible classical trajectories for the case $\epsilon=\pi-2$. The top plot shows some trajectories that do not cross the positive-imaginary-x axis and thus do not leave the principal sheet of the Riemann surface. The trajectories shown are qualitatively similar to those in Fig. 2; all trajectories have the same period.

In the middle plot of Fig. 7 is a trajectory that crosses the positive-imaginary-x axis and visits *three* sheets of the Riemann surface. The solid line and the dotted line outside of the solid line lie on the principal sheet, while the remaining two portions of the dotted line lie on two other sheets. Note that this trajectory does *not* cross itself; we have plotted the projection of the trajectory onto the principal sheet. The trajectory continues to exhibit $\mathcal{PT}$ symmetry. The period of the trajectory is greater than that of the period of the trajectories shown in the top plot. This is because the trajectory encloses turning points that are not on the principal sheet. In general, as the size of the trajectory increases, it encloses more and more complex turning points; each time a new pair of turning points is surrounded by the trajectory the period jumps by a discrete quantity.

Although the trajectory in the bottom plot in Fig. 7 has the same topology as that in the middle plot, it is larger. As the trajectory continues to grow, we observe a phenomenon that seems to be universal; namely, the appearance of a limiting cardioid shape (solid line) on the principal surface. The remaining portion of the trajectory (dotted line) shrinks relative to the cardioid and becomes compact and knotlike.

In Fig. 8 we examine the case $\epsilon=0.5$. In this figure we observe behavior that is qualitatively similar to that seen in Fig. 7; namely, as the trajectory on the principal sheet of the Riemann surface becomes larger and approaches a limiting cardioid, the remaining portion of the trajectory becomes relatively small and knotlike.

To summarize, for any $\epsilon>0$ the classical paths are always $\mathcal{PT}$ symmetric. The simplest such path describes oscillatory motion between the pair of turning points that lie just below the real axis on the principal sheet. In general, the period of this motion as a function of ϵ is given by

$$T=4\sqrt{\pi}E^{-\epsilon/(4+2\epsilon)}\frac{\Gamma((3+\epsilon)/(2+\epsilon))}{\Gamma((4+\epsilon)/(4+2\epsilon))}\cos\left(\frac{\epsilon\pi}{4+2\epsilon}\right). \tag{2.10}$$

Other closed paths having more complicated topologies (and longer periods) also exist, as shown in Figs. 7 and 8.

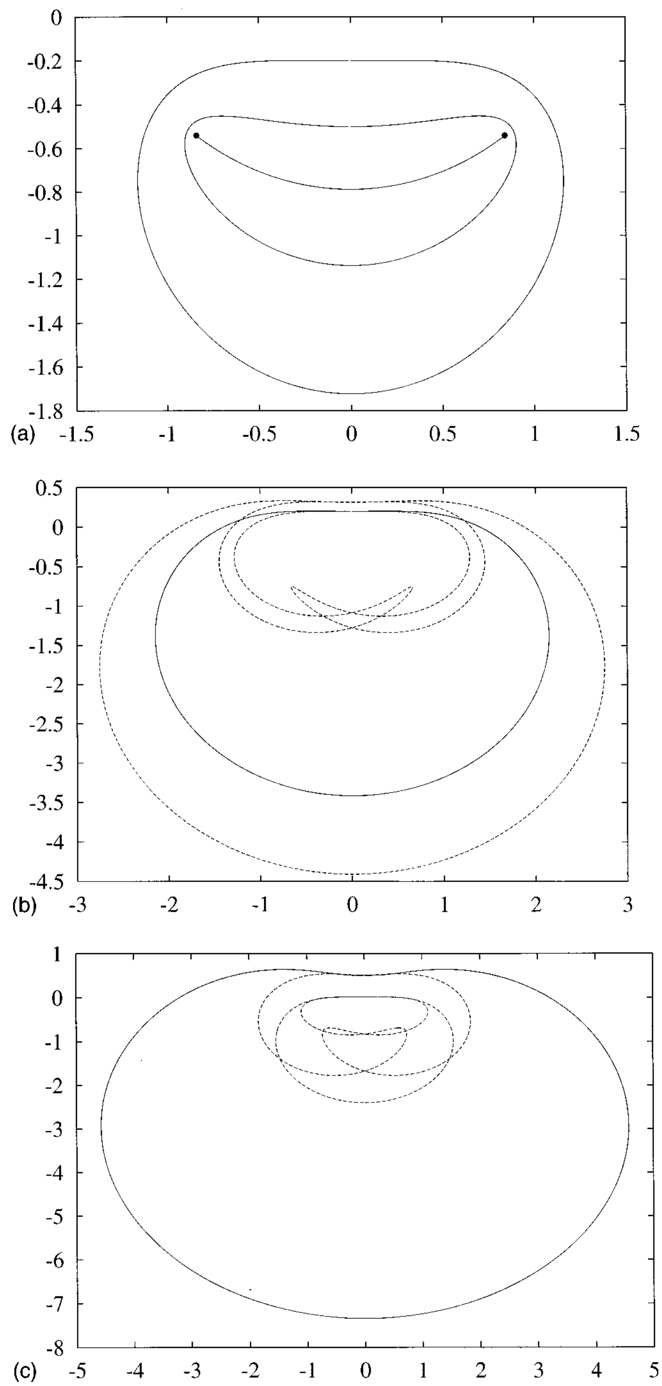

FIG. 7. Classical trajectories for $H=p^2-(ix)^\pi$ corresponding to the case $\epsilon=\pi-2$. Observe that as the classical trajectory increases in size, a limiting cardioid appears on the principal sheet of the Riemann surface. On the other sheets the trajectory becomes relatively small and knotlike.

Whenever the classical motion is periodic, we expect, the quantized version of the theory to exhibit real eigenvalues. Although we have not yet done so, we intend to investigate the consequences of quantizing a theory whose underlying classical paths have complicated topological structures traversing several sheets of a Riemann surface. The properties of such a theory of quantum knots might well be novel.

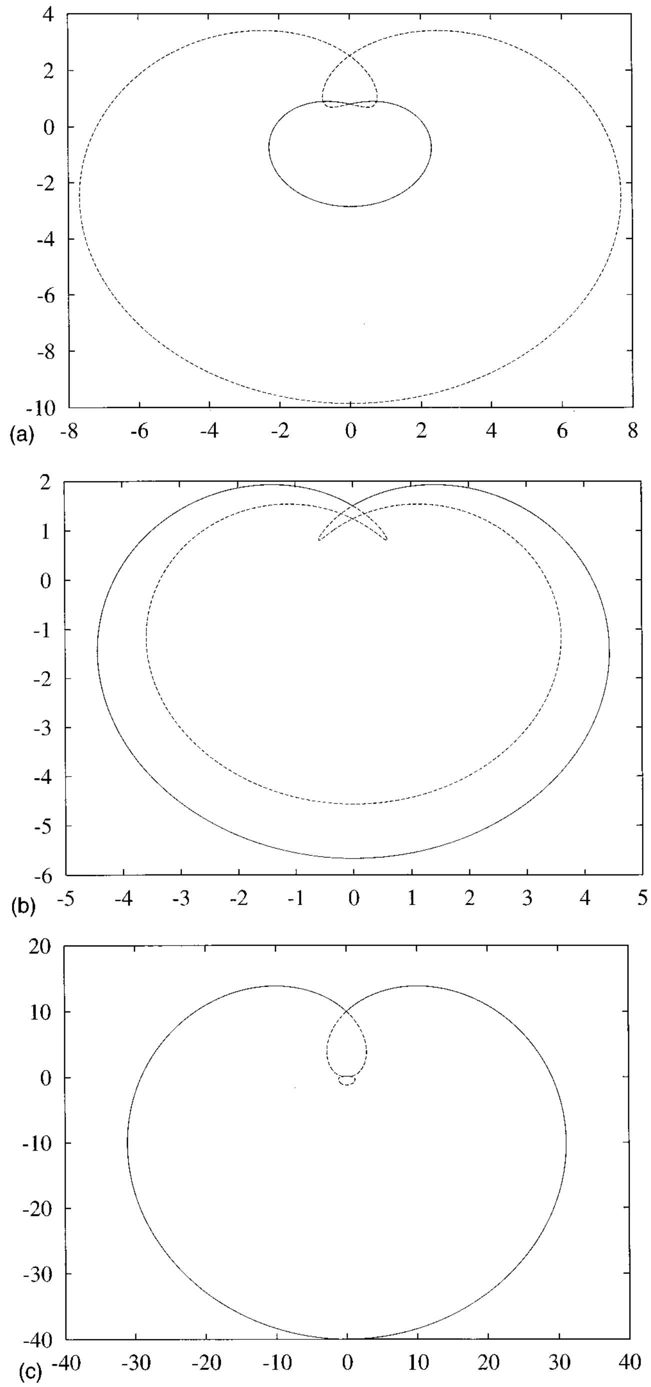

FIG. 8. Classical trajectories for the case $\epsilon=0.5$. As the classical path on the principal sheet of the Riemann surface increases in size it approaches a limiting cardioid, just as in Fig. 7. The remaining portion of the path becomes relatively small and knotlike.

F. Case $-1<\epsilon<0$

Classical paths for negative values of ϵ are fundamentally different from those corresponding to non-negative values of ϵ; such paths no longer exhibit $\mathcal{PT}$ symmetry. Furthermore, we no longer see paths that are periodic; all paths eventually spiral outwards to infinity. In general, the time that it takes for a particle to reach infinity is infinite.

We interpret the abrupt change in the global nature of the classical behavior that occurs as ϵ passes through 0 as a change in phase. For all values of ϵ the Hamiltonian in Eq. (1.1) is $\mathcal{PT}$ (left–right) symmetric. However, for $\epsilon<0$ the solutions cease to exhibit $\mathcal{PT}$ symmetry. Thus, we say that $\epsilon\geqslant 0$ is a $\mathcal{PT}$-symmetric phase and that $\epsilon<0$ is a spontaneously broken $\mathcal{PT}$-symmetric phase.

To illustrate the loss of $\mathcal{PT}$ (left–right) symmetry, we plot in Fig. 9 the classical trajectory for a particle that starts at a turning point $x_-=-\pi(4+\epsilon)/(4+2\epsilon)$ in the second quadrant of the complex-x plane (Re $x<0$, Im $x>0$) for three values of ϵ: -0.2, -0.15, and -0.1. We observe that a path starting at this turning point moves toward but *misses* the $\mathcal{PT}$-symmetric turning point $x_+=-\pi(\epsilon/(4+2\epsilon))$ because it crosses the branch cut on the positive-imaginary-x axis. This path spirals outward, crossing from sheet to sheet on the Riemann surface, and eventually veers off to infinity asymptotic to the angle θ_∞, where

$$\theta_\infty=-\frac{2+\epsilon}{2\epsilon}\pi. \tag{2.11}$$

This formula shows that the total angular rotation of the spiral is finite for all $\epsilon\neq 0$ but becomes infinite as $\epsilon\rightarrow 0^-$. In the top figure ($\epsilon=-0.2$) the spiral makes $2\frac{1}{4}$ turns before moving off to infinity; in the middle figure ($\epsilon=-0.15$) the spiral makes $3\frac{1}{12}$ turns; in the bottom figure ($\epsilon=-0.1$) the spiral makes $4\frac{3}{4}$ turns.

Note that the spirals in Fig. 9 pass many classical turning points as they spiral clockwise from x_-. {From Eq. (2.5) we see that the nth turning point lies at the angle $[(4-\epsilon-4n)/(4+2\epsilon)]\pi$ (x_- corresponds to $n=0$).} As ϵ approaches 0 from below, when the classical trajectory passes a new turning point, there is a corresponding merging of the quantum energy levels (as shown in Fig. 11). As pointed out in Ref. 1, this correspondence becomes exact in the limit $\epsilon\rightarrow 0^-$ and is a manifestation of Ehrenfest's theorem.

G. Case $\epsilon=-1$

For this special case we can solve the equation (2.4) exactly. The result,

$$x(t)=(1-b^2+\tfrac{1}{4}t^2)i+bt \quad (b \text{ real}), \tag{2.12}$$

represents a family of parabolas that are symmetric with respect to the imaginary axis (see Fig. 10). Note that there is one degenerate parabola corresponding to $b=0$ that lies on the positive imaginary axis above i.

III. QUANTUM THEORY

In this section we discuss the quantum properties of the Hamiltonian H in Eq. (1.1). The spectrum of this Hamiltonian is obtained by solving the corresponding Schrödinger equation

$$-\psi''(x)+[x^2(ix)^\epsilon-E]\psi(x)=0 \tag{3.1}$$

subject to appropriate boundary conditions imposed in the complex-x plane. These boundary conditions are described in Ref. 1. A plot of the spectrum of H is shown in Fig. 11.

There are several ways to obtain the spectrum that is displayed in Fig. 11. The simplest and most direct technique is to integrate the differential equation using Runge–Kutta. To do so, we convert the complex differential equation (3.1) to a system of coupled, real, second-order equations. We find that the convergence is most rapid when we integrate along anti-Stokes lines and then patch the two solutions together at the origin. This procedure, which is described in Ref. 1, gives highly accurate numerical results.

To verify the Runge–Kutta approach, we have solved the differential equation (3.1) using an independent and alternative procedure. We construct a matrix representation of the Hamiltonian in Eq. (1.1) in harmonic oscillator basis functions $e^{-x^2/2}H_n(x)\pi^{-1/4}/\sqrt{2^n n!}$:

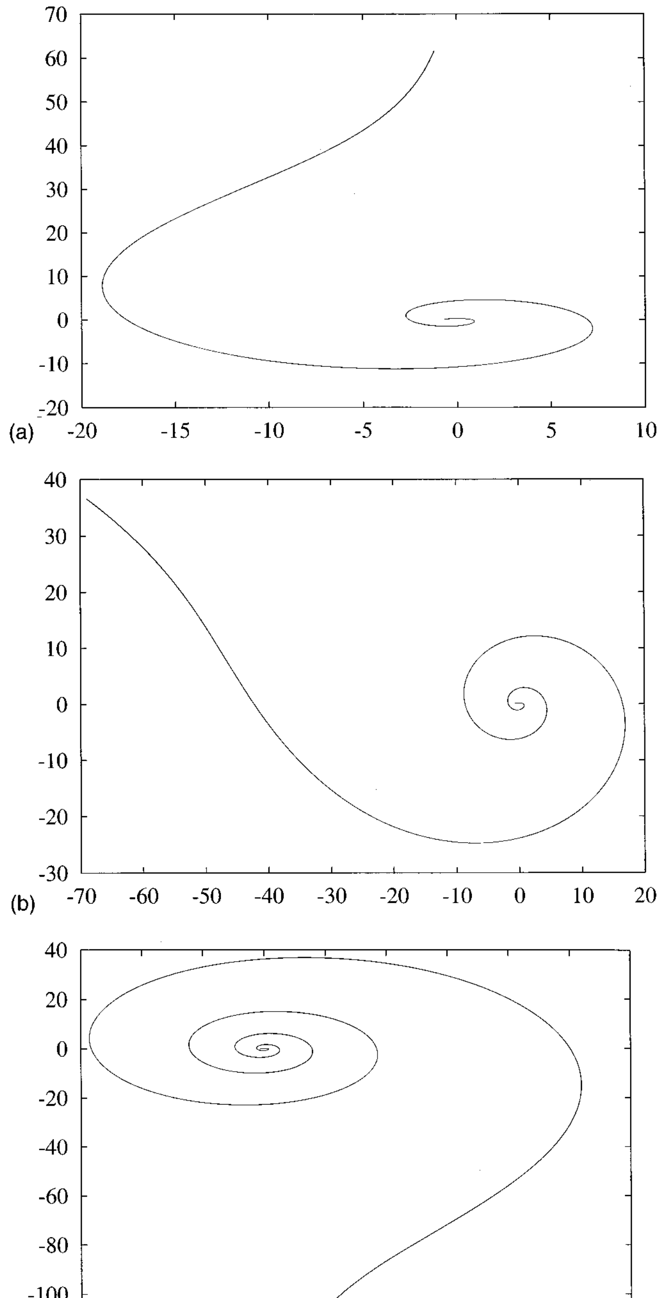

FIG. 9. Classical trajectories that violate $\mathcal{PT}$ symmetry. The top plot corresponds to the case $\epsilon=-0.2$, the middle plot to $\epsilon=-0.15$, and the bottom plot to $\epsilon=-0.1$. The paths in each plot begin at a turning point and spiral outwards to infinity in an infinite amount of time.

$$M_{m,n}=-\int_{-\infty}^{\infty}dx\,\frac{1}{\sqrt{\pi 2^{m+n}m!n!}}e^{-x^2/2}H_m(x)\left\{\frac{d^2}{dx^2}-i^{m+n}\right.$$
$$\left.\times\cos\left[\frac{\pi}{2}(\epsilon-m-n)\right]|x|^{2+\epsilon}\right\}e^{-x^2/2}H_n(x). \tag{3.2}$$

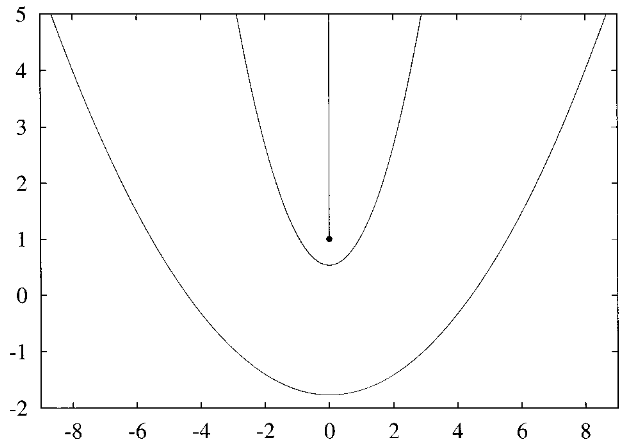

FIG. 10. Classical trajectories in the complex-x plane for a particle described by the Hamiltonian $H=p^2-ix$ and having energy $E=1$. Shown are parabolic trajectories and a turning point at i. All trajectories are unbounded.

The Kth approximant to the spectrum comes from diagonalizing a truncated version of this matrix $M_{m,n}$ $(0\leqslant m,\ n\leqslant K)$. One drawback of this method is that the eigenvalues of $M_{m,n}$ approximate those of the Hamiltonian H in (1.1) only if $-1<\epsilon<2$. Another drawback is that the convergence to the exact eigenvalues is slow and not monotone because the Hamiltonian H is not Hermitian in a conventional sense. We illustrate the convergence of this truncation and diagonalization procedure for $\epsilon=-\frac{1}{2}$ in Fig. 12.

A third method for finding the eigenvalues in Fig. 11 is to use WKB (Wentzel–Kramers–Brillouin). Complex WKB theory (see Ref. 1) gives an excellent analytical approximation to the spectrum.

In the next two subsections we examine two aspects of the spectrum in Fig. 11. First, we study

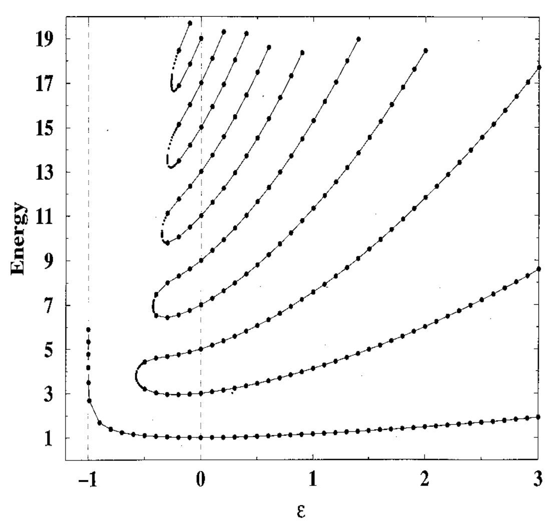

FIG. 11. Energy levels of the Hamiltonian $H=p^2+x^2(ix)^\epsilon$ as a function of the parameter ϵ. There are three regions.: When $\epsilon\geqslant 0$, the spectrum is real and positive and the energy levels rise with increasing ϵ. The lower bound of this region, $\epsilon=0$, corresponds to the harmonic oscillator, whose energy levels are $E_n=2n+1$. When $-1<\epsilon<0$, there are a finite number of real positive eigenvalues and an infinite number of complex conjugate pairs of eigenvalues. As ϵ decreases from 0 to -1, the number of real eigenvalues decreases; when $\epsilon\leqslant -0.57793$, the only real eigenvalue is the ground-state energy. As ϵ approaches -1^+, the ground-state energy diverges. For $\epsilon\leqslant -1$ there are no real eigenvalues.

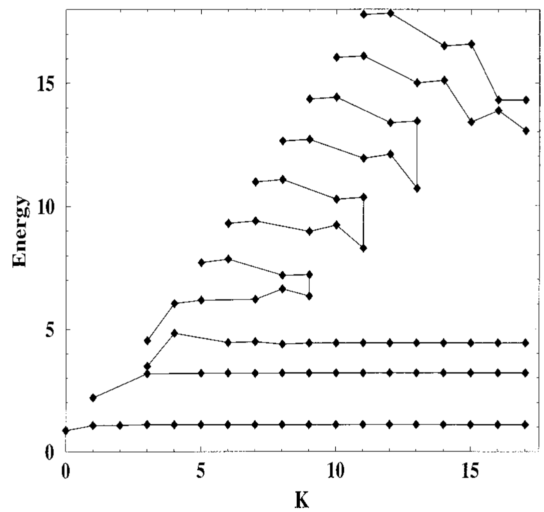

FIG. 12. Real eigenvalues of the $(K+1)\times(K+1)$ truncated matrix $M_{m,n}$ in Eq. (3.2) ($K=0,1,...,17$) for $\epsilon=-\frac{1}{2}$. As K increases, the three lowest eigenvalues converge to the three real energy levels in Fig. 11 at $\epsilon=-\frac{1}{2}$. The other real eigenvalues do not stabilize, and instead disappear in pairs.

the asymptotic behavior of the ground-state energy as $\epsilon\to-1$. Second, we examine the phase transition in the vicinity of $\epsilon=0$.

A. Behavior of the ground-state energy near $\epsilon=-1$

In this subsection we give an analytic derivation of the behavior of the lowest real energy level in Fig. 11 as $\epsilon\to-1$. We show that in this limit the eigenvalue grows logarithmically.

When $\epsilon=-1$, the differential equation (3.1) reduces to

$$-\psi''(x)-ix\psi(x)=E\psi(x), \tag{3.3}$$

which can be solved exactly in terms of Airy functions.[14] The anti-Stokes lines at $\epsilon=-1$ lie at 30° and at $-210°$ in the complex-x plane. We find the solution that vanishes exponentially along each of these rays and then rotate back to the real-x axis to obtain

$$\psi_{L,R}(x)=C_{L,R}\ \ \mathrm{Ai}(\mp xe^{\pm i\pi/6}+Ee^{\pm 2i\pi/3}). \tag{3.4}$$

We must patch these solutions together at $x=0$ according to the patching condition

$$\frac{d}{dx}|\psi(x)|^2\bigg|_{x=0}=0. \tag{3.5}$$

But for real E, the Wronskian identity for the Airy function[14] is

$$\frac{d}{dx}|\mathrm{Ai}(xe^{-i\pi/6}+Ee^{-2i\pi/3})|^2\bigg|_{x=0}=-\frac{1}{2\pi} \tag{3.6}$$

instead of 0. Hence, there is no real eigenvalue.

Next, we perform an asymptotic analysis for $\epsilon=-1+\delta$ where δ is small and positive:

$$\begin{gathered}-\psi''(x)-(ix)^{1+\delta}\psi(x)=E\psi(x),\\ \psi(x)\sim y_0(x)+\delta y_1(x)+O(\delta^2)\quad(\delta\to0+).\end{gathered} \tag{3.7}$$

We assume that $E\to\infty$ as $\delta\to 0+$ and obtain

$$\begin{aligned} y_0''(x)+ixy_0(x)+Ey_0(x)&=0,\\ y_1''(x)+ixy_1(x)+Ey_1(x)&=-ix\ln(ix)y_0(x), \end{aligned} \tag{3.8}$$

and so on.

To leading order we again obtain the Airy equation (3.3) for $y_0(x)$. The solution for $y_0(x)$ $(x\geqslant 0)$ is given by $\psi_R(x)$ in Eq. (3.4) and we are free to choose $C_R=1$. We can expand the Airy function in $y_0(x)$ for large argument in the limit $E\to\infty$:

$$y_0(x)=\mathrm{Ai}(xe^{-i\pi/6}+Ee^{-2i\pi/3})\sim(xe^{-i\pi/6}+Ee^{-2i\pi/3})^{-1/4}\exp[\tfrac{2}{3}(xe^{-i\pi/6}+Ee^{-2i\pi/3})^{3/2}]. \tag{3.9}$$

At $x=0$ we get

$$y_0(0)=\mathrm{Ai}(Ee^{-2i\pi/3})\sim e^{i\pi/6}E^{-1/4}e^{(2/3)E^{3/2}}/(2\sqrt{\pi}). \tag{3.10}$$

To next order in ϵ we simplify the differential equation for $y_1(x)$ in (3.8) by substituting

$$y_1(x)=Q(x)y_0(x). \tag{3.11}$$

Using the differential equation for $y_0(x)$ in (3.8), we get

$$y_0(x)Q''(x)+2y_0'(x)Q'(x)=-ix\ln(ix)y_0(x). \tag{3.12}$$

Multiplying this equation by the integrating factor $y_0(x)$, we obtain

$$[y_0^2(x)Q'(x)]'=-ix\ln(ix)y_0^2(x), \tag{3.13}$$

which integrates to

$$Q'(x)=\frac{i}{y_0^2(x)}\int_x^\infty dt\, t\ln(it)y_0^2(t), \tag{3.14}$$

where the upper limit of the integral ensures that $Q'(x)$ is bounded for $x\to\infty$. Thus, we obtain

$$Q'(0)=\frac{i}{y_0^2(0)}\int_0^\infty dx\, x\ln(ix)y_0^2(x). \tag{3.15}$$

To determine the asymptotic behavior of the ground-state eigenvalue as $\delta\to 0$, we insert

$$\psi(x)\sim y_0(x)+\delta y_1(x)+O(\delta^2)=y_0(x)[1+\delta Q(x)]+O(\delta^2) \tag{3.16}$$

into the quantization condition:

$$\begin{aligned} 0=\frac{d}{dx}[\psi^*(x)\psi(x)]\Big|_{x=0}&\sim\frac{d}{dx}[|y_0(x)|^2(1+\delta Q^*(x))(1+\delta Q(x))]\Big|_{x=0}\\ &\sim\frac{d}{dx}[|y_0(x)|^2]\Big|_{x=0}+2\delta|y_0(0)|^2\,\mathrm{Re}[Q'(0)]\\ &\quad+2\delta\frac{d}{dx}[|y_0(x)|^2]\Big|_{x=0}\mathrm{Re}[Q(0)]. \end{aligned} \tag{3.17}$$

TABLE I. Comparison of the exact ground-state energy E near $\epsilon=-1$ and the asymptotic results in Eq. (3.21). The explicit dependence of E on $\epsilon=-1+\delta$ is roughly $E\propto(-\ln\delta)^{2/3}$ as $\delta-0+$.

δ	E_{exact}	Eq. (3.21)
0.01	1.6837	2.0955
0.01	2.6797	3.9624
0.001	3.4947	3.6723
0.0001	4.1753	4.3013
0.00 001	4.7798	4.8776
0.000 001	5.3383	5.4158
0.0 000 001	5.8943	5.9244

We are free to choose $Q(0)=0$, and doing so eliminates the last term on the right side. The leading-order result for the quantization condition in Eq. (3.6) then gives

$$\frac{1}{2\pi}\sim 2\delta|y_0(0)|^2\,\mathrm{Re}[Q'(0)]. \tag{3.18}$$

Next, we substitute the asymptotic form for y_0 in Eq. (3.10) and the result for $Q'(0)$ in Eq. (3.15) and obtain

$$\sqrt{E}e^{-(4/3)E^{3/2}}\sim 2\delta\,\mathrm{Re}\int_0^\infty dx\,ix\ln(ix)\left[\frac{y_0(x)}{y_0(0)}\right]^2. \tag{3.19}$$

Because the ratio of the unperturbed wave functions in the integrand in Eq. (3.19) is bounded and vanishes exponentially for large x, we know that the integral can grow at most as a power of E. Thus,

$$\delta\sim CE^\alpha e^{-(4/3)E^{3/2}} \tag{3.20}$$

for some power α and constant C and the controling behavior of the ground-state energy as $\delta\to0$ is given by

$$E\sim[-\tfrac{3}{4}\ln\delta]^{2/3}, \tag{3.21}$$

where we have neglected terms that vary at most like $\ln(\ln\delta)$. Equation (3.21) gives the asymptotic behavior of the lowest energy level and is the result that we have sought. This asymptotic behavior is verified numerically in Table I.

B. Behavior of energy levels near $\epsilon=0$

In this subsection we examine analytically the phase transition that occurs at $\epsilon=0$. In particular, we study high-lying eigenvalues for small negative values of ϵ and verify that adjacent pairs of eigenvalues pinch off and become complex.

For small ϵ we approximate H in Eq. (1.1) to first order in ϵ:

$$H=p^2+x^2+\epsilon x^2\ln(ix)+O(\epsilon^2). \tag{3.22}$$

Using the identity $\ln(ix)=\ln(|x|)+\frac{1}{2}i\pi\,\mathrm{sgn}(x)$, we then have

$$H=p^2+x^2+\epsilon x^2\left[\ln(|x|)+\frac{i\pi}{2}\mathrm{sgn}(x)\right]+O(\epsilon^2). \tag{3.23}$$

The simplest way to continue is to truncate this approximate Hamiltonian to a 2×2 matrix. We introduce a harmonic oscillator basis as follows: The nth eigenvalue of the harmonic oscillator Hamiltonian p^2+x^2 is $E_n=2n+1$ and the corresponding x-space normalized eigenstate $|n\rangle$ is

$$\psi_n(x)=\frac{\pi^{-1/4}}{\sqrt{2^n n!}}e^{-x^2/2}H_n(x), \tag{3.24}$$

where $H_n(x)$ is the nth Hermite polynomial $[H_0(x)=1,\ H_1(x)=2x,\ H_2(x)=4x^2-2,\ H_3(x)=8x^3-12x$, and so on]. We then have the following diagonal matrix elements:

$$\langle n|p^2+x^2|n\rangle=2n+1, \tag{3.25}$$

$$\langle n|x^2\ln(|x|)|n\rangle=a_n-\left(\frac{\gamma}{2}+\ln 2\right)\left(n+\frac{1}{2}\right), \tag{3.26}$$

where γ is Euler's constant and

$$a_n=n+1+\left[\frac{n}{2}\right]+\left(n+\frac{1}{2}\right)\sum_{0}^{[n+1/2]}\frac{1}{2k-1}. \tag{3.27}$$

We also have the off-diagonal matrix element

$$\langle 2n-1|\frac{1}{2}i\pi x^2\,\mathrm{sgn}(x)|2n\rangle=\frac{1}{3}i(8n+1)\left[\frac{\Gamma^2(n+1/2)}{n!(n-1)!}\right]^{1/2}. \tag{3.28}$$

In the $(2n-1)-(2n)$ subspace, the matrix $H-E$ then reduces to the following 2×2 matrix:

$$\begin{pmatrix} A-E & iB \\ iB & C-E \end{pmatrix}, \tag{3.29}$$

where for large n and small ϵ we have

$$\begin{aligned} A&\sim 4n-1+\epsilon(n-1/2)\ln(2n),\\ B&\sim \tfrac{8}{3}\epsilon n,\\ C&\sim 4n+1+\epsilon n\ln(2n). \end{aligned} \tag{3.30}$$

The determinant of the matrix in Eq. (3.\29) gives the following roots for E:

$$E=\tfrac{1}{2}\left(A+C\pm\sqrt{(A-C)^2-4B^2}\right). \tag{3.31}$$

We observe that the roots E are degenerate when the discriminant (the square root) in Eq. (3.31) vanishes. This happens when the condition

$$\epsilon=\frac{3}{8n} \tag{3.32}$$

is met. Hence, the sequence of points in Fig. 11 where the eigenvalues pinch off approaches $\epsilon=0$ as $n\to\infty$. For example, Eq. (3.32) predicts (using $n=4$) that E_7 and E_8 become degenerate and move off into the complex plane at $\epsilon\approx-0.1$. In Fig. 13 we compare our prediction for the behavior of E in Eq. (3.31) with a blow-up of of a small portion of Fig. 11. We find that while our prediction is qualitatively good, the numerical accuracy is not particularly good. The lack of accuracy is not associated with truncating the expansion in powers of ϵ but rather with truncating

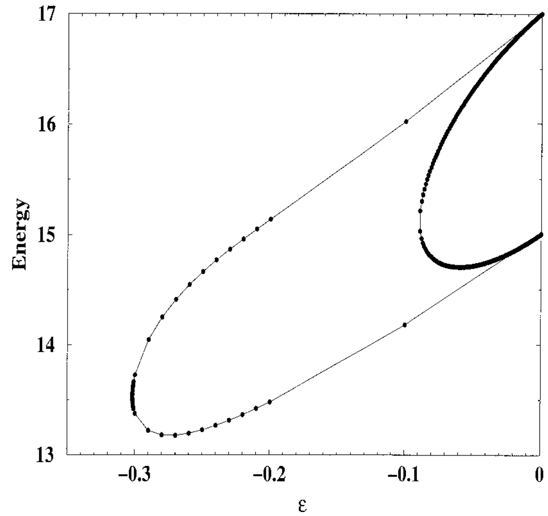

FIG. 13. A comparison of the prediction in Eq. (3.31) and a magnification of Fig. 11. Our prediction for the point at which E_7 and E_8 become degenerate is not very accurate numerically but is qualitatively quite good.

the Hamiltonian H to a 2×2 matrix. Our numerical studies indicate that as the size of the matrix truncation increases, we obtain more accurate approximations to the behavior of the energy levels E in Fig. 11.

IV. MORE GENERAL CLASSES OF THEORIES

In this section we generalize the results of Secs. II and III to a much wider class of theories. In particular, we consider a complex deformation of the x^{2K} anharmonic oscillator, where $K=1,2,3,\ldots$ [see Eq. (1.4)]. The Schrödinger equation for this oscillator has the form

$$-\psi''(x)+[x^{2K}(ix)^{\epsilon}-E]\psi(x)=0. \tag{4.1}$$

To determine the energy levels E as functions of the deformation parameter ϵ, we must impose appropriate boundary conditions on Eq. (4.1). We require that the wave function vanish as $|x|\to\infty$ inside of two wedges symmetrically placed about the imaginary-x axis. The right wedge is centered about the angle $\theta_{\rm right}$, where

$$\theta_{\rm right}=-\frac{\epsilon\pi}{4K+2\epsilon+4}, \tag{4.2}$$

and the left wedge is centered about the angle $\theta_{\rm left}$, where

$$\theta_{\rm left}=-\pi+\frac{\epsilon\pi}{4K+2\epsilon+4}. \tag{4.3}$$

The opening angle of each of these wedges is

$$\frac{2\pi}{2K+\epsilon+2}. \tag{4.4}$$

This pair of wedges is $\mathcal{PT}$ (left–right) symmetric.

The orientation of these wedges is determined by analytically continuing the differential equation eigenvalue problem (4.1) and associated boundary conditions in the variable ϵ using the techniques explained in Ref. 15. The rotation of the boundary conditions is obtained from the asymptotic behavior of the solution $\psi(x)$ for large $|x|$:

$$\psi(x)\sim\exp\left(\pm\frac{i^{\epsilon/2}x^{K+1+\epsilon/2}}{K+1+\epsilon/2}\right). \tag{4.5}$$

(In this formula we give the *controlling factor* of the asymptotic behavior of the wave function; we neglect algebraic contributions.) Note that at the center of the wedges the behavior of the wave function is most strongly exponential; the centerline of each wedge is an anti-Stokes line. At the edges of the wedges the asymptotic behavior is oscillatory. The lines marking the edges of the wedges are Stokes lines.

For all positive integer values of K the results are qualitatively similar. At $\epsilon=0$ the two wedges are centered about the positive and negative real axes. As ϵ increases from 0 the wedges rotate downward and become thinner. In the region $\epsilon\geqslant 0$ the eigenvalues are all real and positive and they rise with increasing ϵ. As $\epsilon\to\infty$, the two wedges become infinitely thin and lie along the negative imaginary axis. There is no eigenvalue problem in this limit because the solution contour for the Schrödinger equation (4.1) can be pushed off to infinity. Indeed, we find that in this limit the eigenvalues all become infinite.

When ϵ is negative, the wedges rotate upward and become thicker. The eigenvalues gradually pair off and become complex starting with the highest eigenvalues. Thus, $\mathcal{PT}$ symmetry is spontaneously broken for $\epsilon<0$. Eventually, as ϵ approaches $-K$, only the lowest eigenvalue remains real. At $\epsilon=-K$ the two wedges join at the positive imaginary axis. Thus, again there is no eigenvalue problem and there are no eigenvalues at all. In the limit $\epsilon\to-K$ the one remaining real eigenvalue diverges logarithmically.

The spectrum for the case of arbitrary positive integer K is quite similar to that for $K=1$. However, in general, when $K>1$, a novel feature emerges: A new transition appears for all negative integer values of ϵ between 0 and $-K$. At these isolated points the spectrum is entirely real. Just above each of these negative-integer values of ϵ the energy levels reemerge in pairs from the complex plane and just below these special values of ϵ the energy levels once again pinch off and become complex.

A. Quantum $x^4(ix)^\epsilon$ theory

The spectrum for the case $K=2$ is displayed in Fig. 14. This figure resembles Fig. 11 for the case $K=1$. However, at $\epsilon=-1$ there is a new transition. This transition is examined in detail in Fig. 15.

An important feature of the spectrum in Fig. 14 is the disappearance of the eigenvalues and divergence of the lowest eigenvalue as ϵ decreases to -2. Following the approach of Sec. III A, we now derive the asymptotic behavior of the ground-state energy as $\epsilon\to-2^+$. To do so we let $\epsilon=-2+\delta$ and obtain from Eq. (4.1) the Schrödinger equation

$$-\psi''(x)-x^2(ix)^\delta\psi(x)=E\psi(x). \tag{4.6}$$

We study this differential equation for small positive δ.

When $\delta=0$ this differential equation (4.6) reduces to

$$-\psi''(x)-x^2\psi(x)=E\psi(x). \tag{4.7}$$

The anti-Stokes lines for this equation lie at 45° and at $-225°$. Thus, we rotate the integration contour from the real axis to the anti-Stokes lines and substitute

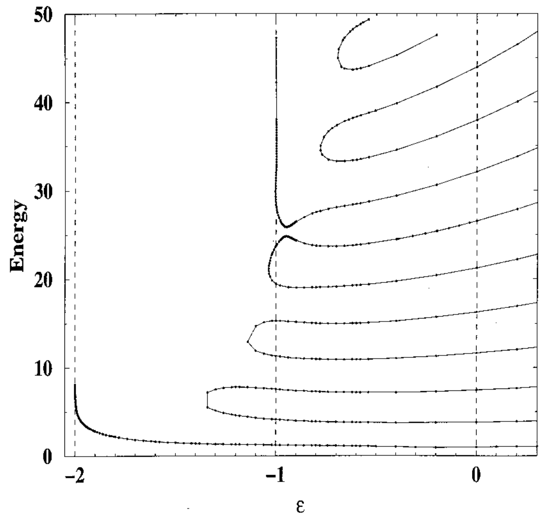

FIG. 14. Energy levels of the Hamiltonian $H=p^2+x^4(ix)^\epsilon$ as a function of the parameter ϵ. This figure is similar to Fig. 11, but now there are four regions: When $\epsilon\geqslant 0$, the spectrum is real and positive and it rises monotonically with increasing ϵ. The lower bound $\epsilon=0$ of this $\mathcal{PT}$-symmetric region corresponds to the pure quartic anharmonic oscillator, whose Hamiltonian is given by $H=p^2+x^4$. When $-1<\epsilon<0$, $\mathcal{PT}$ symmetry is spontaneously broken. There are a finite number of real positive eigenvalues and an infinite number of complex conjugate pairs of eigenvalues; as a function of ϵ the eigenvalues pinch off in pairs and move off into the complex plane. By the time $\epsilon=-1$ only eight real eigenvalues remain; these eigenvalues are continuous at $\epsilon=1$. Just as ϵ approaches -1 the entire spectrum reemerges from the complex plane and becomes real. (Note that at $\epsilon=-1$ the entire spectrum agrees with the entire spectrum in Fig. 11 at $\epsilon=1$.) This reemergence is difficult to see in this figure but is much clearer in Fig. 15 in which the vicinity of $\epsilon=-1$ is blown up. Just below $\epsilon=-1$, the eigenvalues once again begin to pinch off and disappear in pairs into the complex plane. However, this pairing is different from the pairing in the region $-1<\epsilon<0$. Above $\epsilon=-1$ the lower member of a pinching pair is even and the upper member is odd (that is, E_8 and E_9 combine, E_{10} and E_{11} combine, and so on); below $\epsilon=-1$ this pattern reverses (that is, E_7 combines with E_8, E_9 combines with E_{10}, and so on). As ϵ decreases from -1 to -2, the number of real eigenvalues continues to decrease until the only real eigenvalue is the ground-state energy. Then, as ϵ approaches -2^+, the ground-state energy diverges logarithmically. For $\epsilon\leqslant -2$ there are no real eigenvalues.

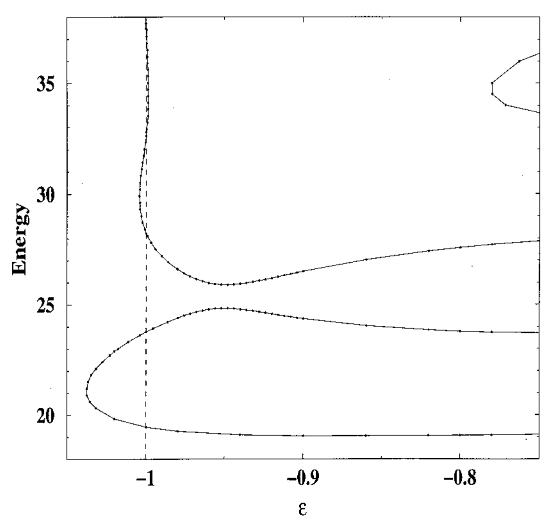

FIG. 15. A magnification of Fig. 14 in the vicinity of the transition at $\epsilon=-1$. Just above $\epsilon=-1$ the entire spectrum reemerges from the complex plane, and just below $\epsilon=-1$ it continues to disappear into the complex plane. The spectrum is entirely real at $\epsilon=-1$.

$$x=\begin{cases}\dfrac{s}{\sqrt{2}}e^{-5i\pi/4} & (\mathrm{Re}\,x<0),\\[2ex] \dfrac{r}{\sqrt{2}}e^{i\pi/4} & (\mathrm{Re}\,x>0),\end{cases} \tag{4.8}$$

for x in the left-half and in the right-half complex plane, respectively. Note as s and r increase, x moves towards complex infinity in both the left- and right-half plane.

The wave function in the left-half plane, $\psi_L(s)$, and the wave function in the right-half plane, $\psi_R(r)$, satisfy the differential equations

$$\begin{aligned}-\frac{d^2}{ds^2}\psi_L(s)+\left(\frac{s^2}{4}-\frac{1}{2}\right)\psi_L(s)=\nu\psi_L(s),\\ -\frac{d^2}{dr^2}\psi_R(r)+\left(\frac{r^2}{4}-\frac{1}{2}\right)\psi_R(r)=(-\nu-1)\psi_R(r),\end{aligned} \tag{4.9}$$

where we have set $\nu=-(i/2)E-\frac{1}{2}$. For each of these equations the solution that vanishes at infinity is a parabolic cylinder function:[16]

$$\begin{aligned}\psi_L(s)=C_L D_\nu(s)=C_L D_\nu(x\sqrt{2}e^{5i\pi/4}),\\ \psi_R(r)=C_R D_{-\nu-1}(r)=C_R D_{-\nu-1}(x\sqrt{2}e^{-i\pi/4}),\end{aligned} \tag{4.10}$$

where C_L and C_R are arbitrary constants.

We impose the quantization condition by patching these solutions together at $x=0$ on the real-x axis according to the patching conditions

$$\begin{aligned}\psi_L(x)\big|_{x=0}=\psi_R(x)\big|_{x=0},\\ \frac{d}{dx}\psi_L(x)\bigg|_{x=0}=\frac{d}{dx}\psi_R(x)\bigg|_{x=0}\end{aligned} \tag{4.11}$$

To eliminate the constants C_L and C_R we take the ratio of these two equations and simplify the result by cross multiplying:

$$\left[\psi_R(x)\frac{d}{dx}\psi_L(x)-\psi_L(x)\frac{d}{dx}\psi_R(x)\right]\bigg|_{x=0}=0. \tag{4.12}$$

We now show that this condition cannot be satisfied by the $\delta=0$ wave function in Eq. (4.10). For this case, the quantization condition (4.12) states that

$$D_\nu(s)\frac{d}{ds}D_{-\nu-1}(is)-D_{-\nu-1}(is)\frac{d}{ds}D_\nu(s) \tag{4.13}$$

vanishes at $s=0$. (We have simplified the argument by setting $s=x\sqrt{2}e^{5i\pi/4}$). But Eq. (4.13) for any value of s is just the Wronskian for parabolic cylinder functions[16] and this Wronskian equals $-ie^{-i\nu\pi/2}$. This is a *nonzero* result. Thus, when $\delta=0$, there cannot be any eigenvalue E, real or complex, and the spectrum is empty.

The quantization condition (4.12) can be satisfied when $\delta>0$. We investigate this region for the case when δ is small and positive by performing an asymptotic analysis. We assume that $E\to\infty$ as $\delta\to 0+$, but slower than any power of δ, and that the wave function $\psi(x)$ has a formal power series expansion in δ:

$$\psi(x) \sim y_0(x) + \delta y_1(x) + O(\delta^2) \quad (\delta \to 0+). \tag{4.14}$$

Next, we expand the Schrödinger equation (4.6) in powers of δ:

$$\begin{aligned} y_0''(x) + x^2 y_0(x) + E y_0(x) &= 0, \\ y_1''(x) + x^2 y_1(x) + E y_1(x) &= -x^2 \ln(ix) y_0(x), \end{aligned} \tag{4.15}$$

and so on.

Of course, to zeroth order in δ we obtain Eq. (4.7) for $y_0(x)$. Thus, in the left- and right-half complex x-plane we get

$$\begin{aligned} y_0^L(x) &= C_L D_\nu(x\sqrt{2} e^{5i\pi/4}), \\ y_0^R(x) &= C_R D_{-\nu-1}(x\sqrt{2} e^{-i\pi/4}). \end{aligned} \tag{4.16}$$

To first order in δ, we simplify the differential equation for $y_1(x)$ in (4.15) by substituting

$$y_1(x) = Q(x) y_0(x). \tag{4.17}$$

Using the differential equation for $y_0(x)$ in (4.15), we get

$$y_0(x) Q''(x) + 2 y_0'(x) Q'(x) = -x^2 \ln(ix) y_0(x). \tag{4.18}$$

Multiplying this equation by the integrating factor $y_0(x)$, we obtain

$$[y_0^2(x) Q'(x)]' = -x^2 \ln(ix) y_0^2(x). \tag{4.19}$$

The integral of this equation gives

$$\begin{aligned} Q_L'(x) &= \int_x^{\infty e^{-5i\pi/4}} dt\, t^2 \ln(it) \left[\frac{y_0^L(t)}{y_0^L(x)} \right]^2, \\ Q_R'(x) &= \int_x^{\infty e^{i\pi/4}} dt\, t^2 \ln(it) \left[\frac{y_0^R(t)}{y_0^R(x)} \right]^2, \end{aligned} \tag{4.20}$$

where the limit of the integral at infinity ensures that $Q'(x)$ is bounded for $|x| \to \infty$.

To determine the asymptotic behavior of the ground-state eigenvalue as $\delta \to 0^+$, we insert

$$\psi_{L,R}(x) \sim y_0^{L,R}(x) + \delta_{y_1}^{L,R}(x) + O(\delta^2) = y_0^{L,R}(x)[1 + \delta Q^{L,R}(x)] \tag{4.21}$$

into the quantization condition (4.12):

$$\begin{aligned} 0 &= \left[\psi_R(x) \frac{d}{dx} \psi_L(x) - \psi_L(x) \frac{d}{dx} \psi_R(x) \right] \Bigg|_{x=0} \\ &= \left[y_0^R(x) \frac{d}{dx} y_0^L(x) - y_0^L(x) \frac{d}{dx} y_0^R(x) \right] \Bigg|_{x=0} [1 + \delta(Q_R(0) + Q_L(0))] \\ &\quad + \delta y_0^R(0) y_0^L(0) [Q_L'(0) - Q_R'(0)]. \end{aligned} \tag{4.22}$$

We are free to choose $Q_R(0) + Q_L(0) = 0$ to simplify this result.

Substituting the Wronskian for the parabolic cylinder function and the result for $y_0(0)$ in Eq. (4.16), we obtain

$$\sqrt{2}e^{-\pi E/4} \sim \delta D_\nu(0) D_{-\nu-1}(0)[Q_R'(0) - Q_L'(0)]. \tag{4.23}$$

We can simplify this result using the identity

$$D_\nu(0) = \frac{\sqrt{\pi}2^{\nu/2}}{\Gamma(\frac{1}{2} - \nu/2)} \tag{4.24}$$

and $\nu = -(i/2)E - \frac{1}{2}$ to obtain

$$D_\nu(0) D_{-\nu-1}(0) = \frac{\Gamma(\frac{1}{2} + \nu/2)\cos(\pi\nu/2)}{\Gamma(1+\nu/2)\sqrt{2}} \sim \frac{\epsilon^{(\pi/4)E}}{\sqrt{E}}, \tag{4.25}$$

where we have used the reflection formula, $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$, and the asymptotic behavior $\Gamma(x+1/2)/\Gamma(x+1) \sim x^{-1/2}$ for large x. Thus, Eq. (4.23) reduces to

$$\frac{\sqrt{2E}}{\delta} e^{-\pi E/2} \sim Q_R'(0) - Q_L'(0). \tag{4.26}$$

We can further show that

$$\begin{aligned} Q_R'(0) - Q_L'(0) &= \int_0^{\infty e^{i\pi/4}} dt\, t^2 \ln(it) \left[\frac{D_{-\nu-1}(\sqrt{2}te^{-i\pi/4})}{D_{-\nu-1}(0)} \right]^2 - \int_0^{\infty e^{-5i\pi/4}} dt\, t^2 \ln(it) \left[\frac{D_\nu(\sqrt{2}te^{5i\pi/4})}{D_\nu(0)} \right]^2 \\ &= -\int_0^\infty \frac{t^2 dt}{2^{3/2}} e^{-i\pi/4} \ln\left(\frac{s}{\sqrt{2}} e^{3i\pi/4} \right) \left[\frac{D_{-\nu-1}(t)}{D_{-\nu-1}(0)} \right]^2 \\ &\quad - \int_{0\infty} \frac{t^2 dt}{2^{3/2}} e^{i\pi/4} \ln\left(\frac{s}{\sqrt{2}} e^{-3i\pi/4} \right) \left[\frac{D_\nu(t)}{D_\nu(0)} \right]^2. \end{aligned} \tag{4.27}$$

We observe that the previous expression is real because $\nu^* = -\nu - 1$ implies that $D_\nu(t)^* = D_{-\nu-1}(t)$ and thus the two integrals are complex conjugates. Thus, Eq. (4.27) is real, and E is a real function of δ. Furthermore, because the ratio $D_\nu(t)/D_\nu(0)$ appears in both integrals, the expression can at most vary as a power of E. Hence, the contribution of $Q_R'(0) - Q_L'(0)$ to the balance in Eq. (4.26) is subdominant and can be neglected. Our final result for the small-δ behavior of the lowest eigenvalue is that

$$E \sim -\frac{2}{\pi} \ln \delta + O[\ln(\ln \delta)] \quad (\delta \to 0^+). \tag{4.28}$$

In Fig. 16 we show that Eq. (4.28) compares well with the numerical data for the lowest eigenvalue in the limit as $\delta \to 0$.

B. Classical $x^4(ix)^\epsilon$ theory

It is instructive to compare the quantum mechanical and classical mechanical theories for the case $K=2$. Our objective in doing so is to understand more deeply the breaking of $\mathcal{PT}$ symmetry that occurs at $\epsilon = 0$. For the case $K=1$ we found that $\mathcal{PT}$ symmetry is broken at the classical level in a rather obvious way: Left–right symmetric classical trajectories become spirals as ϵ becomes negative (see Fig. 9). However, we find that when $K=2$ spirals do not occur until $\epsilon < -2$. The classical manifestation of $\mathcal{PT}$ symmetry breaking for $-2 \leq \epsilon < 0$ and the transition that occurs at $\epsilon = 0$ is actually quite subtle.

For purposes of comparison we begin by examining the classical trajectories for the positive value $\epsilon = 0.7$. In Fig. 17 we plot three classical trajectories in the complex-x plane. The first is an

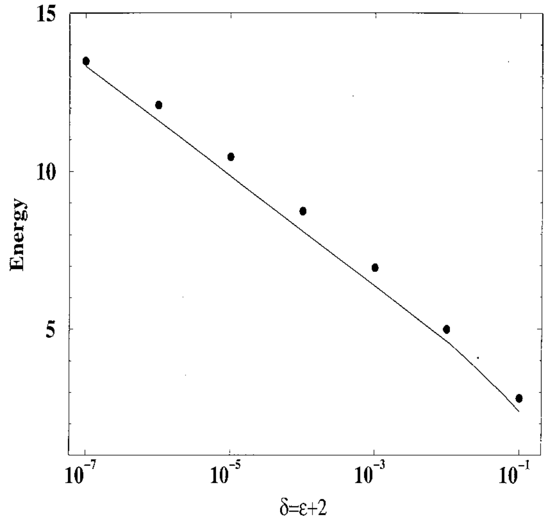

FIG. 16. A comparison of the lowest eigenvalue of the Hamiltonian $H=p^2+x^4(ix)^\epsilon$ (solid circles) with the asymptotic prediction in (4.28) (solid line) near $\epsilon=-2$. The solid line includes a one parameter fit of terms that grow like ln(ln δ) as $\delta\to 0^+$.

arc that joins the classical turning points in the lower-half plane. The other two are closed orbits that surround this arc. The smaller closed orbit remains on the principal sheet and has a period $(T\approx 4.9)$, which is equal to that of the arc. The more complicated trajectory is left–right symmetric but extends to three sheets of the Riemann surface. The period $(T\approx 26.1)$ of this third orbit is significantly different from and larger than the period of the other two.

Next, we consider the negative value $\epsilon=-0.7$. In Fig. 18 we plot two classical trajectories for this value. The first (solid line) is an arc joining the classical turning points in the upper-half plane. This arc extends to three sheets of the Riemann surface. The other trajectory (dashed line) is a closed orbit that surrounds this arc. Both have the period $T\approx 22.3$. This figure illustrates the first

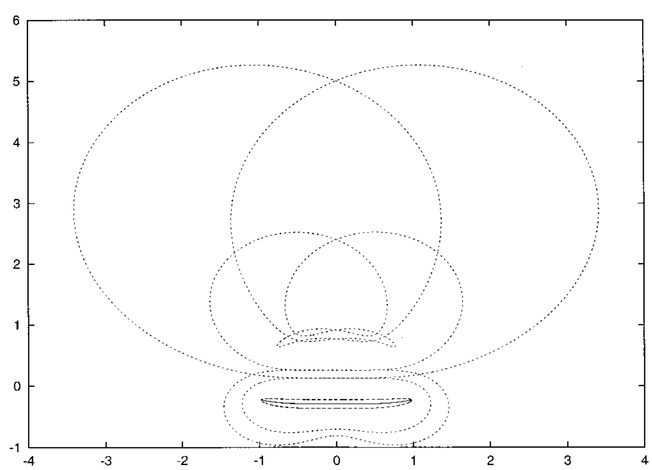

FIG. 17. Three classical trajectories in the complex-x plane for a particle described by the Hamiltonian $H=p^2+x^4(ix)^\epsilon$ with $\epsilon=0.7$. The solid line represents oscillatory motion between the classical turning points. The long-dashed line is a nearby trajectory that encloses and has the same period as the solid-line trajectory. The short dashed line has a different topology (it enters three sheets of the Riemann surface) from the long-dashed line, even though these trajectories are very near one another in the vicinity of the turning points. The period of this motion is much longer than that of the solid and long-dashed trajectories.

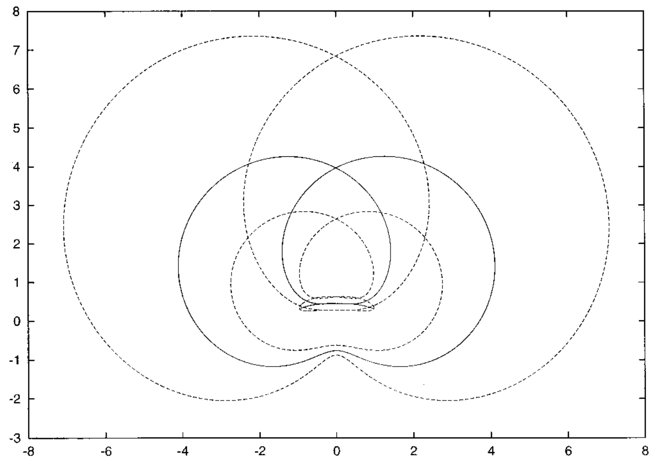

FIG. 18. Two classical trajectories in the complex-x plane for a particle described by the Hamiltonian $H=p^2+x^4(ix)^\epsilon$ with $\epsilon=-0.7$. The solid line represents oscillatory motion between the classical turning points. This trajectory enters three sheets of the Riemann surface. The dashed line is a nearby trajectory that encloses and has the same period as the solid-line trajectory.

of two important changes that occur as ϵ goes below zero. The trajectory that joins the two turning points no longer lies on the principal sheet of the Riemann surface; it exhibits a multisheeted structure.

Figure 19 illustrates the second important change that occurs as ϵ goes below zero. On this figure we again plot two classical trajectories for the negative value $\epsilon=-0.7$. The first (solid line) is the arc joining the classical turning points in the upper-half plane. This arc is also shown on Fig. 18. The second trajectory (dashed line) is a closed orbit that passes near the turning points. The two trajectories do not cross; the apparent points of intersection are on different sheets of the Riemann surface. The period of the dashed trajectory is $T\approx 13.7$, which is considerably *smaller* than that of the solid line. Indeed, on the basis of extensive numerical studies, it appears that all trajectories for $-2<\epsilon<0$, while they are $\mathcal{PT}$ (left–right) symmetric, have periods that are less than or equal to that of the solid line. When $\epsilon>0$, the periods of trajectories increase as the trajectories move away from the oscillatory trajectory connecting the turning points.

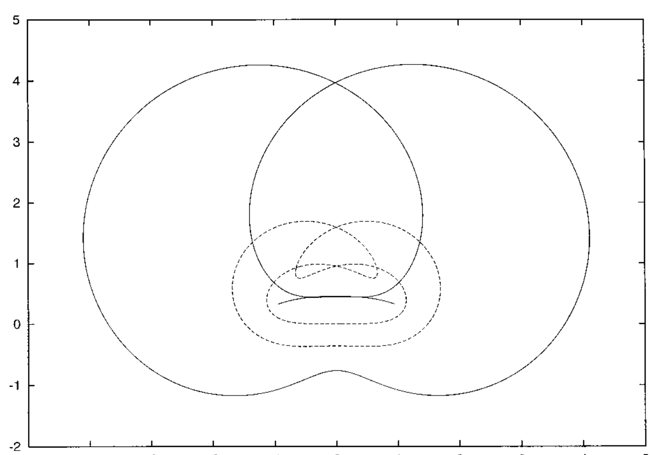

FIG. 19. Two classical trajectories in the complex-x plane for a particle described by the Hamiltonian $H=p^2+x^4(ix)^\epsilon$ with $\epsilon=-0.7$. The solid line represents oscillatory motion between the classical turning points and is the same as that in Fig. 18. The dashed line is a nearby trajectory whose period is smaller than the period of the solid-line trajectory.

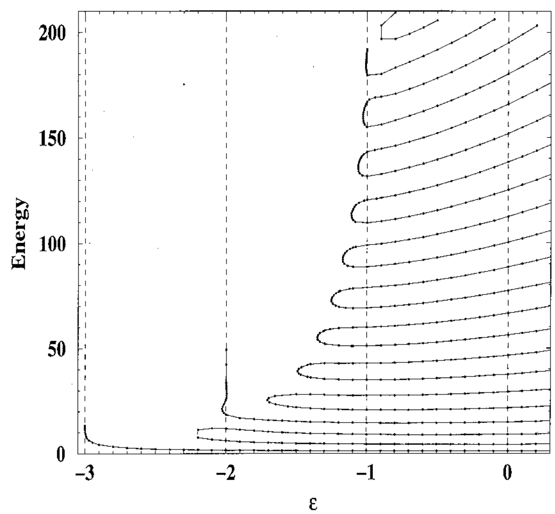

FIG. 20. Energy levels of the Hamiltonian $H=p^2+x^6(ix)^\epsilon$ as a function of the parameter ϵ. This figure is similar to Fig. 14, but now there are five regions: When $\epsilon \geqslant 0$, the spectrum is real and positive and it rises monotonically with increasing ϵ. The lower bound $\epsilon=0$ of this $\mathcal{PT}$-symmetric region corresponds to the pure sextic anharmonic oscillator, whose Hamiltonian is given by $H=p^2+x^6$. The other four regions are $-1<\epsilon<0$, $-2<\epsilon<-1$, $-3<\epsilon<-2$, and $\epsilon<-3$. The $\mathcal{PT}$ symmetry is spontaneously broken when ϵ is negative, and the number of real eigenvalues decreases as ϵ becomes more negative. However, at the boundaries $\epsilon=-1,-2$ there is a complete real positive spectrum. When $\epsilon=-1$, the eigenspectrum is identical to the eigenspectrum in Fig. 14 at $\epsilon=1$. For $\epsilon \leqslant -3$ there are no real eigenvalues.

We speculate that for negative values of ϵ the appearance of complex eigenvalues in the quantum theory (see Fig. 14) is associated with an instability. The path integral for a quantum theory is ordinarily dominated by paths in the vicinity of the classical trajectory connecting the turning points. However, when ϵ is negative, we believe that these trajectories no longer dominate the path integral because there are more remote trajectories whose classical periods are *smaller*. Thus, the action is no longer dominated by a stationary point in the form of a classical path having $\mathcal{PT}$ symmetry. Hence, the spectrum can contain complex eigenvalues.

The appearance of a purely real spectrum for the special value $\epsilon=-1$ is consistent with this conjecture. For integer values of $\epsilon>-2$ we find that all classical trajectories lie on the principal sheet of the Riemann surface and have the *same* period.

C. Quantum $x^6(ix)^\epsilon$ theory

The spectrum for the case $K=3$ is displayed in Fig. 20. This figure resembles Fig. 14 for the case $K=2$. However, now there are transitions at both $\epsilon=-1$ and $\epsilon=-2$.

ACKNOWLEDGMENTS

We thank D. Bessis, M. Flato, J. Wess, A. Wightman, and Y. Zarmi for helpful conversations. We are grateful to the U.S. Department of Energy for financial support.

[1] C. M. Bender and S. Boettcher, Phys. Rev. Lett. **80**, 5243 (1998). Note that in this reference the notation N represents the quantity $2+\epsilon$.
[2] Several years ago, D. Bessis and J. Zinn-Justin conjectured that the spectrum of H in Eq. (1.1) for the special case $\epsilon=1$ is real (private communication). A partial proof for the reality of the spectrum in this special case has been given by M. P. Blencowe, H. Jones, and A. P. Korte, Phys. Rev. D **57**, 5092 (1998) using the linear delta expansion and by E. Delabaere and F. Pham, Ann. Phys. **261**, 180 (1997) using WKB methods.
[3] T. J. Hollowood, Nucl. Phys. B **386**, 166 (1992).
[4] D. R. Nelson and N. M. Shnerb, Phys. Rev. E **58**, 1383 (1998).
[5] N. Hatano and D. R. Nelson, Phys. Rev. Lett. **77**, 570 (1996); Phys. Rev. B **56**, 8651 (1997).
[6] C. M. Bender and S. Boettcher, J. Phys. A **31**, L273 (1998).

[7] C. M. Bender and K. A. Milton, Phys. Rev. D **55**, R3255 (1997).
[8] Nontriviality of $-g\phi^4$ field theory has been argued using analytic continuation techniques [see K. Gawadzki and A. Kupiainen, Nucl. Phys. B **257**, 474 (1985)] and using Gaussian approximation [see B. Rosenstein and A. Kovner, Phys. Rev. D **40**, 504 (1989)].
[9] C. M. Bender and K. A. Milton, Phys. Rev. D **57**, 3595 (1998).
[10] C. M. Bender, S. Boettcher, H. F. Jones, and P. N. Meisinger, in preparation.
[11] C. M. Bender and K. A. Milton, J. Phys. A **32**, L87 (1999).
[12] Equation (2.3) is a complex version of the statement that the velocity is the time derivative of the position ($v = dx/dt$). Here, the time is real but the velocity and position are complex.
[13] C. M. Bender and J. P. Vinson, J. Math. Phys. **37**, 4103 (1996).
[14] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1964).
[15] C. M. Bender and A. Turbiner, Phys. Lett. A **173**, 442 (1993).
[16] A. Erdelyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. 2.

Hidden supersymmetry and Berezin quantization of $N=2$, $D=3$ spinning superparticles

I. V. Gorbunov[a)]
Department of Physics, Tomsk State University, Lenin Avenue 36, Tomsk 634050, Russia

S. L. Lyakhovich[b)]
Department of Physics, Tomsk State University, Lenin Avenue 36, Tomsk 634050, Russia and International Centre for Theoretical Physics, P.O. Box 586, 3410 Trieste, Italy

(Received 23 September 1998; accepted for publication 1 December 1998)

The first quantized theory of $N=2$, $D=3$ massive superparticles with arbitrary fixed central charge and (half) integer or fractional superspin is constructed. The quantum states are realized on the fields carrying a finite-dimensional, or a unitary infinite-dimensional, representation of the supergroups OSp $(2|2)$ or SU $(1,1|2)$. The construction originates from quantization of a classical model of the superparticle we suggest. The physical phase space of the classical superparticle is embedded in a symplectic superspace $T^*(R^{1,2})\times\mathcal{L}^{1|2}$, where the inner Kähler supermanifold $\mathcal{L}^{1|2}\cong\mathrm{OSp}(2|2)/[\mathrm{U}(1)\times\mathrm{U}(1)]\cong\mathrm{SU}(1,1|2)/[\mathrm{U}(2|2)\times\mathrm{U}(1)]$ provides the particle with superspin degrees of freedom. We find the relationship between Hamiltonian generators of the global Poincaré supersymmetry and the ''internal'' SU$(1,1|2)$ one. Quantization of the superparticle combines the Berezin quantization on $\mathcal{L}^{1|2}$ and the conventional Dirac quantization with respect to space–time degrees of freedom. Surprisingly, to retain the supersymmetry, quantum corrections are required for the classical $N=2$ supercharges as compared to the conventional Berezin method. These corrections are derived and the Berezin correspondence principle for $\mathcal{L}^{1|2}$ underlying their origin is verified.
[S0022-2488(99)01303-1]

I. INTRODUCTION

In this paper we construct an $N=2$, $D=1+2$ massive spinning superparticle model and study the symplectic supergeometry behind it. This supergeometry is compatible with the Berezin quantization method which is applied to construct the one-particle quantum theory. The main part of our consideration is based on the observation that the $N=2$ superextension of the $D=3$ spinning particle results in the classical model which possesses simultaneously Poincaré supersymmetry (SUSY) and Lorentz supersymmetry of the superspin degrees of freedom. This ''double'' supersymmetry can be lifted to the quantum level and we obtain the realization of the $N=2$, $D=3$ Poincaré supermultiplet on the fields carrying an irreducible representation of the supergroup SU$(1,1|2)$ [''Lorentz supergroup'' whose even part is $\mathrm{SO}^{\uparrow}(1,2)\times\mathrm{U}(2)\times$central charge]. A nonlinear mutual involvement of the Hamiltonian generators of two supersymmetries requires the careful geometric quantization of the superparticle. At first, we try to explain the most important motivations of the problem.

In the hierarchy of all known entities, the particles living in three-dimensional space–time stand out mostly due to a possibility of fractional spin and statistics (anyons). Anyon excitations are actually presented in some planar physics phenomena,[1,2] and the relevant theoretical concept has both topological[3–5] and group-theoretical[6–8] grounds. It is well known that in the field theory

[a)]Electronic mail: ivan@phys.tsu.ru
[b)]Electronic mail: sll@phys.tsu.ru

fractional statistics originates usually from a coupling of the matter fields to the gauge field with the Chern–Simons mass term.[9,10] The supersymmetric extension of this approach[11] implies a direct interaction between anyon excitations.

The group-theoretical methods may give an alternative way to understand the anyon concept. One can start from the mechanical model of the $D=3$ spinning particle, whose quantization leads to the one-particle quantum mechanics for the fractional spin state.[12,13,8,14–17] It is established that $D=3$ spinning particles possess the following remarkable features: (i) the spinning particle carries as many physical degrees of freedom as a spinless one; (ii) there is the so-called *canonical model*[12] of the spinning particle, which implies a deformation of the canonical symplectic structure of the spinless particle by the use of the Dirac monopole two-form, without extension of the phase space introducing any ''spinning'' variables; (iii) it is a promising feature of the canonical model to be adapted for the construction of consistent couplings of the particle to external fields[18,19,15,20,16] and self-interaction of anyons.[21,22] In higher dimensions, the interaction problem for spinning particles becomes more involved, although some progress has recently been achieved there as well;[23,24] (iv) the anyon wave equations may be formulated in analogy with the ones for bosons and fermions. An essential difference is that the fractional spin, in contrast to the (half) integer, is naturally described in terms of infinite component fields carrying infinite-dimensional representations of the universal covering group $\overline{SO^{\uparrow}(1,2)} \cong \overline{SU(1,1)}$; (v) representations of fractional spin are multivalued.

There is no consistent quantum field theory of anyons up to now; nevertheless, the Chern–Simons and group-theoretical constructions are deemed to lead to a unified consistent theory. In this regard, it would be interesting to understand how the supersymmetry may be included into a group-theoretical description of anyons in terms of the infinite component fields.

Another reason to investigate the $D=3$ superparticle is the exceptional fact that not only the Poincaré supersymmetry is possible in $1+2$ dimensions, but the Lorentz one is, too. The Lorentz group $SO^{\uparrow}(1,2)$ coincides with the $D=2$ anti-de Sitter group; the latter admits the superextension regardless of specific space–time dimension. Although the Lorentz and the Poincaré supersymmetries are not compatible with each other, surprisingly, we will show that the Lorentz supersymmetry of the $D=3$ spinning superparticle (which is invariant by construction with respect to the global Poincaré SUSY transformations) manifests itself as a hidden supersymmetry of internal degrees of freedom associated to the particle superspin and to the underlying superextended monopolelike symplectic structure.

The hidden $OSp(2|2)$ supersymmetry of $N=1$ superanyons has been found in Ref. 25 where the respective model is constructed. The presence of the $OSp(2|2)$ supersymmetry already in the classical mechanics appears to be crucial for a consistent first quantization of the $N=1$, $D=3$ superanyon. As a result, one obtains in quantum theory the realization of the $N=1$ Poincaré supermultiplet on the fields carrying an atypical unitary infinite-dimensional representation of the $OSp(2|2)$.[25] It is a direct $N=1$ superextension of description in terms of infinite-dimensional unitary representation of the $D=3$ Lorentz group[6–8] or the ones of the deformed Heisenberg algebra.[26] We argued in this manner the relevance of the group-theoretical approach for $N=1$ supersymmetric anyons. In this paper we suggest a nontrivial generalization of this construction to the case of the $N=2$, $D=3$ massive spinning superparticle with arbitrary fixed central charge.

We construct a superparticle model, which gives $N=2$ superextension of the canonical description of the $D=3$ spinning particle mentioned above. It is essential for our consideration that the Hamiltonian formalism of the canonical model may be built either in terms of the minimal phase space, or in an extended phase space restricted by constraints.[14,16,17,25] In both cases the reduced phase space could be thought of as a space of motion of a Souriau's ''elementary system.''[27]

A general concept of elementary physical systems, including spinning particles and superparticles, is based on the so-called Kostant–Souriau–Kirillov (KSK) construction.[28,27,29] The idea of the KSK construction is to identify the *physical phase space (=space of motion)* of any elementary system with a *coadjoint orbit* $\mathcal{O}$ of the symmetry group G. The symplectic action G on $\mathcal{O}$ (classical mechanics) lifts to a representation of the group in a space of functions $\mathcal{H}$ on the classical manifold (prequantization). Then the quantization problem reduces to an appropriate choice of polarization, that is, a global Lagrangian section in $T(\mathcal{O})$ being invariant under the action of the symmetry group.

In the special case of Kähler homogeneous spaces, perfect results can be achieved in the framework of the Berezin quantization method,[30,31] which implies one-to-one correspondence between the phase-space functions (covariant Berezin symbols) and linear operators in a Hilbert space. The latter is realized by holomorphic sections, because the Kähler homogeneous manifold admits a natural complex polarization.[32] Moreover, the multiplication of the operators in the Hilbert space induces a noncommutative binary $*$-operation for the covariant Berezin symbols and a correspondence principle can be proved.[30,31]

Physically speaking, it would not always be satisfactory to describe elementary systems in terms of the coadjoint orbits. In particular, the dynamics of relativistic particles and superparticles is usually supposed to evolve in a fiber bundle $\mathcal{M}$ over a *space–time manifold* that is crucial for the interaction problem. Thus, the coadjoint orbit of the spinning (super)particle arises from embedding into evolution (super)space. The projection π: $\mathcal{M}\rightarrow\mathcal{O}_G$, where G is a Poincaré (super)group, generates G-invariant constraints and gauge symmetries in $\mathcal{M}$. The construction of interactions, being consistent with the gauge symmetries, and the quantization problem for π provide a subject of current interest in the problems of spinning particle and superparticle models.[12,33–35]

Concerning the $D=3$ spinning particle, it is established[13,14,16,25] that the quantization problem for the canonical model is naturally solved by means of an embedding of the maximal (four-dimensional) coadjoint orbit of the group $\mathrm{ISO}^{\uparrow}(1,2)$ into eight-dimensional phase space (that is, *extended phase space*) $\mathcal{M}^8\cong T^*(R^{1,2})\times\mathcal{L}$. Here $\mathcal{L}\cong\mathrm{SU}(1,1)/\mathrm{U}(1)$ is a Lobachevsky plane and the character $\cong$ denotes a symplectomorphism. The projection π: $\mathcal{M}^8\rightarrow\mathcal{O}_{m,s}$ onto coorbit $\mathcal{O}_{m,s}$ of the particle of mass m and spin s is provided by the constraints. The auxiliary variables parametrizing $\mathcal{L}$ are used to describe the particle spin. One can interpret the (holomorphic) automorphisms of the Lobachevskian Kähler metric as a hidden symmetry of the internal particle's structure, which is related to the spin. In this approach the quantization of the anyon could be achieved as a compromise of the conventional Dirac quantization on $T^*(R^{1,2})$ and of the geometric quantization in the Lobachevsky plane. Constraints of the classical mechanics are converted into wave equations of the anyon according to the Dirac prescriptions.

The starting point of this paper is a mechanical model of $N=2$ superparticles with arbitrary fixed mass $m>0$, superspin $s\neq 0$, and central charge $\mathcal{Z}\equiv mb$, $|b|\leqslant 1$, briefly announced before.[36] For this elementary system the maximal coadjoint orbit $\mathcal{O}_{m,s,b}$ of real dimension 4/4 is related to the case $|b|<1$. In our model, this orbit appears embedded into 8/4-dimensional extended phase superspace $\mathcal{M}^{8|4}$ of a special geometry: $\mathcal{M}^{8|4}\cong T^*(R^{1,2})\times\mathcal{L}^{1|2}$, where $\mathcal{L}^{1|2}=\mathrm{SU}(1,1|2)/[\mathrm{U}(2|2)\times\mathrm{U}(1)]\cong\mathrm{OSp}(2|2)/[\mathrm{U}(1)\times\mathrm{U}(1)]$ is an atypical Kähler coadjoint orbit of the supergroup SU (1,1|2) and the typical one of OSp (2|2). The inner supermanifold $\mathcal{L}^{1|2}$, providing the particle model with a nonzero superspin, was studied originally in Refs. 38 and 39 in relation to OSp (2|2) supercoherent states and called the $N=2$ *superunit disc*. The projection of $\mathcal{M}^{8|4}$ onto physical subspace follows similarly to the nonsupersymmetric model. In fact, introducing the supersymmetry for the $D=3$ particle, we need to superextend only the inner submanifold $\mathcal{L}$ of the extended phase space. The extended phase superspace $\mathcal{M}^{8|4}\cong T^*(R^{1,2})\times\mathcal{L}^{1|2}$ carries ''double supersymmetry:'' one is related to the Poincaré supergroup and acts on the associated coorbit $\mathcal{O}_{m,s,b}\subset\mathcal{M}^{8|4}$, another one lives in the inner subsupermanifold $\mathcal{L}^{1|2}$. Moreover, the model allows an extended hidden $N=4$ supersymmetry with special values of the central charges saturating the Bogomol'ny–Prassad–Sommerfield (BPS) bound.

We will quantize the theory similarly to the quantization of the canonical model of the particle on $\mathcal{M}^8$.[16,25] Specifically, we combine the geometric quantization in the inner subsupermanifold $\mathcal{L}^{1|2}$ for the internal SU (1,1|2) supersymmetry and the canonical Dirac quantization in $T^*(R^{1,2})$.

This quantization scheme implies from the outset that the mentioned ''double supersymmetry'' must survive in the quantum theory. The crucial point is to express the Hamiltonian generators of the Poincaré supersymmetry in $\mathcal{M}^{8|4}$ in terms of the ones of internal SU (1,1|2) supersymmetry (as well as of space–time coordinates and momenta). These expressions appear to be *nonlinear*. As a consequence, some renormalization of the Poincaré supergenerators should be required for the closure of the Poincaré supersymmetry algebra. Roughly speaking, the corrections to generators could be treated as a manifestation of the ordering ambiguity for operators in quantum theory. We will see that the origin of the corrections may also be clarified from the viewpoint of the Berezin quantization in $\mathcal{L}^{1|2}$ and the underlying correspondence principle. However, the Berezin method itself does not provide a regular technique of deriving the closing corrections which have to recover the representation of the Poincaré superalgebra in quantum theory. Moreover, it is unclear *a priori* whether the consistent corrections exist at all. Surprisingly, the problem is solved if a simple ansatz is taken for the renormalized Poincaré generators. Then the closing corrections, which appear in the order of $\mathcal{O}(s^{-2})$, can be *exactly* calculated.

We arrive eventually to the realization of the unitary representation of the $N=2$, $D=3$ supermultiplet on the fields carrying atypical irreps of the supergroup SU(1,1|2) and the typical ones of the subsupergroup OSp(2|2). These irreps are certainly infinite dimensional for the case of fractional superspin, but for the habitual case of (half) integer superspin they may be chosen to be finite dimensional.

The model of the $N=2$ superparticle reduces to the one of the $N=1$ superparticle in the Bogomol'ny–Prassad–Sommerfield (BPS) limit for central charge, when $|b|=1$. One can trace the BPS limit both at the classical and quantum levels. Classically, it corresponds to the degenerate coadjoint orbit of the $D=3$, $N=2$ superparticle of dimension 4/2. When $|b|=1$, the extended phase superspace becomes degenerate and reduces to $\mathcal{M}^{8|2}\cong T^*(R^{1,2})\times\mathcal{L}^{1|1}$ with inner supermanifold $\mathcal{L}^{1|1}\cong \mathrm{OSp}(2|2)/\mathrm{U}(1|1)\cong \mathrm{OSp}(1|2)/\mathrm{U}(1)$. $\mathcal{M}^{8|2}$ is exactly the extended phase superspace of the $N=1$ superanyon.[25] In this exceptional case, the generators of the $N=1$ Poincaré supersymmetry and the internal OSp(2|2) one are *linearly* expressible to one another. Thus, the geometric quantization immediately gives the quantum theory of the $N=1$ superparticle, without extra constructions and corrections. In this paper we touched on the BPS limit briefly; the detailed theory is considered in Refs. 25 and 37.

The geometric quantization in the OSp(2|2) coadjoint orbits was constructed in Refs. 38 and 39 and we follow these results. At the same time we have to clarify two important points, which have seemingly been unknown. First, we found out that the Kähler geometry of the regular coorbit $\mathcal{L}^{1|2}$ admits the symplectic holomorphic action of the supergroup SU(1,1|2), which is larger than the supergroup OSp(2|2) in itself. We construct the geometric quantization on $\mathcal{L}^{1|2}$ provided for this extended supersymmetry supergroup. Second, we perform Berezin quantization for $\mathcal{L}^{1|2}$ to establish a correspondence principle and to explain the origin of quantum corrections to the $N=2$ Poincaré supercharges in $\mathcal{M}^{8|4}$.

The paper is organized as follows. In Sec. II we recall briefly the canonical model of the $D=3$ spinning particle in terms of the minimal and extended phase spaces. Specifically, we focus on symplectic structure and symmetries of the minimal and extended spaces.

Then we are going to construct the superextension of the canonical model. The classical mechanics of the $N=2$, $D=3$ massive spinning superparticle with arbitrary central charge is considered in Sec. III. Starting from a first-order Lagrangian we study the supergeometry of the phase superspace and identify it with $\mathcal{M}^{8|4}=T^*(R^{1,2})\times\mathcal{L}^{1|2}$. We construct explicitly the embeddings of the $N=2$ Poincaré and Lorentz supergroup's coadjoint orbits into $\mathcal{M}^{8|4}$ and find out the Hamiltonian generators of corresponding supersymmetries. The relation, being crucial for quantization, is established between the $N=2$ Poincaré and SU(1,1|2) Hamiltonian generators. We also reveal a degenerate $N=4$ supersymmetry in the model and a special case of degenerate coorbits, which appear in the BPS limit.

In Sec. IV we suggest a quantization procedure for the classical mechanics constructed in Sec. III. At first the Berezin quantization is considered on the regular OSp(2|2) coadjoint orbit. In particular, we construct the correspondence between symbols and operators on $\mathcal{L}^{1|2}$ and prove the underlying correspondence principle. Then these results are applied to the consistent quantization of the $D=3$, $N=2$ superparticle, which is the final object of construction.

II. MINIMAL AND EXTENDED PHASE SPACES OF A CANONICAL MODEL OF SPINNING PARTICLE

First consider the nonsupersymmetric canonical model of the particle (various formulations, see Refs. 12, 6, 7, 14, and 17), which serves as an initial subject for further generalizations. The particle lives originally on six-dimensional phase space $\mathcal{M}^6$ with a symplectic two-form

$$\Omega_s = -dx^a \wedge dp_a + \Omega_m, \quad \Omega_m = \frac{s}{2}\frac{\epsilon^{abc} p_a dp_b \wedge dp_c}{(-p^2)^{3/2}}, \quad p^2 < 0, \tag{2.1}$$

where Ω_m is known as the Dirac monopole form. [We use Latin letters to denote the $D=3$ Lorentz vectors and Greek letters for the SU(1,1) spinors; the Minkowski metric is chosen to be $\eta_{ab} = \mathrm{diag}(-1,1,1)$, the totally antisymmetric tensor is normalized by the condition $\epsilon_{012} = -\epsilon^{012} = 1$; the spinor indices are raised and lowered with the use of the spinor metric $\epsilon^{\alpha\beta} = -\epsilon^{\beta\alpha} = -\epsilon_{\alpha\beta}$ $(\alpha,\beta = 0,1)$, $\epsilon^{01} = -1$ by the rule $\psi_\alpha = \epsilon_{\alpha\beta}\psi^\beta$, $\psi^\alpha = \epsilon^{\alpha\beta}\psi_\beta$.] The Poincaré transformations are generated by the following functions:

$$\mathcal{P}_a = p_a, \quad \mathcal{J}_a = \epsilon_{abc} x^b p^c - s\frac{p_a}{(-p^2)^{1/2}}, \tag{2.2}$$

which constitute the $D=3$ Poincaré algebra with respect to Poisson brackets (PBs)

$$\{\mathcal{P}_a, \mathcal{P}_b\} = 0, \quad \{\mathcal{J}_a, \mathcal{P}_b\} = \epsilon_{abc}\mathcal{P}^c, \quad \{\mathcal{J}_a, \mathcal{J}_b\} = \epsilon_{abc}\mathcal{J}^c, \tag{2.3}$$

The fundamental PBs read

$$\{x^a, x^b\} = s\frac{\epsilon^{abc} p_c}{(-p^2)^{3/2}}, \quad \{x^a, p_b\} = \delta^a{}_b, \quad \{p_a, p_b\} = 0. \tag{2.4}$$

The last two PBs mean that x^a and p_a transform as coordinates and momenta by Poincaré translations. Moreover, they are Lorentz vectors because of $\{\mathcal{J}_a, x_b\} = \epsilon_{abc} x^c$ and $\{\mathcal{J}_a, p_b\} = \epsilon_{abc} p^c$.

Let us assume that the particle dynamics on $\mathcal{M}^6$ is governed by the mass shell *constraint*

$$p^2 + m^2 = 0, \tag{2.5}$$

whereas the canonical Hamilton function is identically zero. On the mass shell, the Casimir functions of the enveloping Poincaré algebra are identically conserved: $\mathcal{P}^2 = -m^2$, $(\mathcal{P},\mathcal{J}) = ms$. We conclude that the $D=3$ particle of mass *m*, spin *s*, and energy sign $p^0/|p^0|$ lives on mass shell. From now on, we take a further restriction $p^0 > 0$, bearing in mind the supersymmetric theory, when the energy is positive essentially. The mass shell constraint generates the reparametrization (gauge) invariance for every world line of the particle. The set of world lines, being considered modulo to the gauge equivalence, is named the particle history space, the latter is isomorphic to the physical state space $\mathcal{O}_{m,s}$ of the spinning particle. The reduced symplectic manifold $\mathcal{O}_{m,s}$ is symplectomorphic to the maximal *coadjoint orbit*[27,29,32] of the $D=3$ Poincaré group.

There is a standard way to extend the canonical model to the Poincaré supersymmetry. One may substitute $dx^a \to dx^a - i(\gamma^a)_{\alpha\beta}\theta^{\alpha I} d\theta^{\beta I}$ in Eq. (2.1), introducing real Grassmann variables $\theta^{\alpha I}$, $I = 1,...,N$. The resulting symplectic superform appears to be invariant under the *N*-extended Poincaré supergroup without central charges. One may further generate central charges introducing some Wess–Zumino-type terms[40,34] in Eq. (2.1). Then, imposing the mass shell constraint (2.5), one may build the classical model of the $D=3$ superparticle of mass *m*, superspin *s*, and arbitrary fixed central charges in the 6/2*N*-dimensional phase superspace. However, it is hardly possible to conceive satisfactory quantization of this model.

Even for the canonical model without supersymmetry the realization of the coordinate operators $\hat{x}^a$ is a nontrivial problem accounting for the complicated form of the first Poisson bracket in Eqs. (2.4). A detailed analysis of Ref. 14 shows that the manifest covariance of the canonical model, being formulated in terms of the ''minimal'' phase space $\mathcal{M}^6$, is inevitably lost in quantum theory. The superextension of the canonical model makes the Poisson brackets, being quantized, much more complicated. In fact, the quantization problem in the reduced nonlinear phase superspace is not solved even for the spinless $D=3$ superparticle. Thus we will reformulate from the outset the canonical model in an ''extended'' phase space, where a hidden symmetry of the spinning particle becomes transparent and gives an efficient method for quantization making use of this symmetry. Moreover, the construction will be appropriate for intriguing superextension.

An adapted reformulation of the canonical model is suggested in Refs. 16 and 25. We observe, that the monopole two-form Ω_m in Eq. (2.1) is nothing else but the Kähler two-form on the mass hyperboloid (2.5), which gives the realization of the Lobachevsky plane $\mathcal{L}$. It will be convenient to make use of another realization of $\mathcal{L}\cong\{z\in C^1,|z|<1\}$ by an open unit disc of complex plane C^1. We rewrite the symplectic two-form (2.1) as follows:

$$\Omega_s=-dx^a\wedge dp_a+\Omega_{\mathcal{L}},\quad \Omega_{\mathcal{L}}=-2is\frac{dz\wedge d\overline{z}}{(1-z\overline{z})^2}, \tag{2.6}$$

where (recall that $p^2<0$ and we have taken $p^0>0$)

$$p^a=\sqrt{-p^2}n^a,\quad n^a=\left(\frac{1+z\overline{z}}{1-z\overline{z}},-\frac{z+\overline{z}}{1-z\overline{z}},i\frac{\overline{z}-z}{1-z\overline{z}}\right),\quad n^2\equiv-1. \tag{2.7}$$

The unit timelike Lorentz vector n^a parametrizes the points of the Lobachevsky plane.

Let us look at Eq. (2.6) from a different viewpoint. Consider a new phase space $\mathcal{M}^8\cong T^*(R^{1,2})\times\mathcal{L}$ with a symplectic two-form (2.6) and an elementary system on $\mathcal{M}^8$, whose dynamics is subjected by three constraints

$$p^a=mn^a. \tag{2.8}$$

Apparently these constraints project the extended phase space $\mathcal{M}^8$ into the same coadjoint orbit as the mass shell constraint (2.5) does for $\mathcal{M}^6$. Alternatively, one can solve explicitly only two constraints $p^a=\sqrt{-p^2}n^a$ providing the reduction $\pi_1:\mathcal{M}^8\to\mathcal{M}^6$ of extended phase space to the minimal one. In other words, we have constructed the sequence of embeddings $\mathcal{O}_{m,s}\subset\mathcal{M}^6\subset\mathcal{M}^8$. Hence, we get an equivalent description of the $D=3$ spinning particle in terms of the extended phase space $T^*(R^{1,2})\times\mathcal{L}$. The Hamiltonian generators of the canonical Poincaré transformations in $\mathcal{M}^8$ read

$$\mathcal{P}_a=p_a,\quad \mathcal{J}_a=\epsilon_{abc}x^bp^c+J_a, \tag{2.9}$$

where the spin vector J_a is expressed in terms of the ''inner'' space $\mathcal{L}$:

$$J_a=-sn_a. \tag{2.10}$$

The Hamiltonians (2.9) generate the Poincaré algebra with respect to PBs in $\mathcal{M}^8$, whereas the spin generators (2.10) span the internal Lorentz algebra related to the (holomorphic) automorphism group of the Lobachevsky plane. The latter group can be recognized as a hidden symmetry of the internal structure of spinning particle. Although this concept may seem artificial at the moment, below we will observe essentially nontrivial superextension of the hidden symmetry.

The Poincaré Casimir functions are identically conserved owing to constraints (2.8):

$$p^2+m^2=0,\quad (p,J)-ms=0. \tag{2.11}$$

A crucial detail is that the equations (2.11) define the same surface in the extended phase space as the constraints (2.8) do.

The quantization of the model in $\mathcal{M}^8$ is almost transparent. We can combine the canonical Dirac quantization in $T^*(R^{1,2})$ and the Berezin quantization in the Lobachevsky plane.[25] Constraints (2.11) will be imposed in Hilbert space to separate the one-particle states.

Finally, write down the Lagrangian of the theory. One may choose the action functional as an integrand of the one-form Θ, where $d\Theta=\Omega_s+V$ and V vanishes on shell. Let us take

$$S=\int\Theta,\quad \Theta=p_a dx^a+is\frac{\bar{z}dz-zd\bar{z}}{1-z\bar{z}}\equiv p_a dx^a+\Sigma_{\mathcal{L}},\quad d\Sigma_{\mathcal{L}}=\Omega_{\mathcal{L}}. \tag{2.12}$$

It is implied here that the virtual paths lay in the constraint surface (2.8). Excluding the momenta accounting for constraints (2.8) and making pull back of Θ, one obtains the action functional

$$S=\int_{\tau_1}^{\tau_2}L\,d\tau,\quad L=m(\dot{x},n)+is\frac{\bar{z}\dot{z}-z\dot{\bar{z}}}{1-z\bar{z}}, \tag{2.13}$$

with the first-order Lagrangian being invariant under reparametrizations. Notice that the Lagrangian is also strongly invariant under translations and spatial rotations, whereas the Lorentz boosts change it by a total derivative.

III. CLASSICAL MODEL OF $D=3$ SPINNING SUPERPARTICLE

A. A first-order Lagrangian

Introduce an $N=2$ superextension of the Lagrangian (2.13) providing both the super-Poincaré invariance of the theory and other hidden supersymmetry as well. Introducing a pair of Majorana anticommuting spinors $\theta^{\alpha I}=(\theta^\alpha,\chi^\alpha)$, $I=1,2$, we suggest

$$L=m(\Pi,n)+mb(\theta_\alpha\dot{\chi}^\alpha-\chi_\alpha\dot{\theta}^\alpha)-mb\,\theta^\alpha n_{\alpha\gamma}\dot{n}^\gamma{}_\beta\chi^\beta+is\frac{\bar{z}\dot{z}-z\dot{\bar{z}}}{1-z\bar{z}}, \tag{3.1}$$

where m,b,s are real parameters, n^a is a unit Lorentz vector in the Lobachevsky plane, being defined by Eq. (2.7), and

$$n_{\alpha\beta}\equiv n^a\gamma_{a\alpha\beta},\quad \Pi^a=\dot{x}^a-i\gamma^a_{\alpha\beta}(\theta^\alpha\dot{\theta}^\beta+\chi^\alpha\dot{\chi}^\beta).$$

The three-dimensional Dirac matrices γ_a are chosen in the form

$$(\gamma_0)_{\alpha\beta}=\begin{pmatrix}0&1\\1&0\end{pmatrix},\quad(\gamma_1)_{\alpha\beta}=\begin{pmatrix}1&0\\0&1\end{pmatrix},\quad(\gamma_2)_{\alpha\beta}=\begin{pmatrix}-i&0\\0&i\end{pmatrix},$$

$$(\gamma_a)_{\alpha\gamma}(\gamma_b)^\gamma{}_\beta=i\epsilon_{abc}(\gamma^c)_{\alpha\beta}-\eta_{ab}\epsilon_{\alpha\beta}.$$

[One may wonder why the γ-matrices are not Hermitian. It is instructive to note that the reality condition for SU(1,1) spinor formalism is not trivial, as for isomorphic SL(2,$\mathbb{R}$) ones. For any $g\in$ SU(1,1) the complex conjugation reads $\bar{g}=cgc$, where $c=c^{-1}=\mathrm{antidiag}(-1,-1)$. The matrices $c\gamma^a$ are truly Hermitian. The covariant Majorana (reality) condition looks like

$$c\bar{\psi}=\psi \tag{3.2}$$

for two-component SU(1,1)-spinor ψ.] The first term in the Lagrangian (3.1) is a conventional superextension of the respective expression in Eq. (2.13) and the second addend represents the Wess–Zumino-type term generating the central charge for the supersymmetry.[40] At last, the third term accounts for the specific of the $D=3$ spinning superparticle model. Owing to this addend, the supertranslations, underlying Poincaré supersymmetry of the Lagrangian, read rather unusually:

$$\delta_\epsilon x^a = i\gamma^a_{\alpha\beta}\epsilon^\alpha\theta^\beta + ib\epsilon^{abc}n_b\gamma_{c\alpha\beta}\epsilon^\alpha\chi^\beta + bn^a\epsilon^\alpha\chi_\alpha, \quad \delta_\epsilon\theta^\alpha = \epsilon^\alpha, \quad \delta_\epsilon\chi^\alpha = 0, \quad \delta_\epsilon z = 0$$
$$\delta_\eta x^a = i\gamma^a_{\alpha\beta}\eta^\alpha\chi^\beta - ib\epsilon^{abc}n_b\gamma_{c\alpha\beta}\eta^\alpha\theta^\beta - bn^a\eta^\alpha\theta_\alpha, \quad \delta_\eta\theta^\alpha = 0, \quad \delta_\eta\chi^\alpha = \eta^\alpha, \quad \delta_\eta z = 0. \tag{3.3}$$

Here ϵ^α and η^α are odd real parameters. For completeness, expose also the even infinitesimal Poincaré transformations and U(1) transformations as well:

$$\delta_\omega x^a = \epsilon^{abc}\omega_b x_c, \quad \delta_\omega\theta^\alpha = -\frac{i}{2}\omega^a\gamma_a{}^\alpha{}_\beta\theta^\beta \quad \delta_\omega\chi^\alpha = -\frac{i}{2}\omega^a\gamma_a{}^\alpha{}_\beta\chi^\beta, \quad \delta_\omega z = i\omega^a\xi_a,$$
$$\delta_f x^a = f^a, \quad \delta_f\theta^\alpha = \delta_f\chi^\alpha = 0, \quad \delta_f z = 0, \tag{3.4}$$
$$\delta_\mu x^a = 0, \quad \delta_\mu\theta^\alpha = -\mu\theta^\alpha, \quad \delta_\mu\chi^\alpha = \mu\theta^\alpha, \quad \delta_\mu z = 0,$$

with the even real parameters ω^a, f^a, and μ and the holomorphic object $\xi_a = -1/2(2z, 1+z^2, i(1-z^2))$. The infinitesimal transformations (3.4) and (3.3) generate the $N=2$ Poincaré superalgebra, which is discussed in Sec. III C.

B. Extended phase superspace

We show in this subsection that the superparticle being described by the Lagrangian (3.1) lives in a supersymplectic phase space $\mathcal{M}^{8|4}$ of very special supergeometry: $\mathcal{M}^{8|4} \cong T^*(R^{1,2}) \times \mathcal{L}^{1|2}$. Then we identify $\mathcal{L}^{1|2}$ with regular (when $|b|<1$) or degenerate (when $|b|=1$) coadjoint orbit of the OSp(2|2) supergroup. Having the goal to quantize the theory in $\mathcal{M}^{8|4}$, we will need detailed information about SUSYs and quantization in $\mathcal{L}^{1|2}$. The supersymplectic geometry of $\mathcal{L}^{1|2}$ is considered in Sec. III D, while the Berezin quantization will be constructed in IV A.

The model (3.1) fits naturally into the formulation in symplectic language. The theory originates from the action functional

$$S = \int \Theta^{\mathrm{SUSY}}, \quad \Theta^{\mathrm{SUSY}} = p_a dx^a + \Sigma_{\mathcal{L}^{1|2}}, \tag{3.5}$$

$$\Sigma_{\mathcal{L}^{1|2}} = -imn_{\alpha\beta}\theta^\alpha d\theta^\beta - imn_{\alpha\beta}\chi^\alpha d\chi^\beta + mb\theta_\alpha d\chi^\alpha - mb\chi_\alpha d\theta^\alpha$$
$$-2mb\frac{z^\alpha z^\beta\theta_\alpha\chi_\beta d\bar{z} - \bar{z}^\alpha\bar{z}^\beta\theta_\alpha\chi_\beta dz}{(1-z\bar{z})^2} + is\frac{\bar{z}dz - zd\bar{z}}{1-z\bar{z}}, \tag{3.6}$$

where the virtual paths belong to the surface

$$p^a = mn^a, \tag{3.7}$$

as follows from the definition $p_a = \partial L/\partial\dot{x}^a$. Introduce the objects

$$z^\alpha \equiv (1,z), \quad \bar{z}^\alpha \equiv (\bar{z},1), \quad \alpha = 0,1, \tag{3.8}$$

which simplifies Eq. (3.6) and many of the forthcoming formulas.

Relation (3.5) shows that the particle dynamics is embedded in phase superspace $\mathcal{M}^{8|4} \cong T^*(R^{1,2}) \times \mathcal{L}^{1|2}$ with some inner superspace of a real dimension 2/4 denoted by $\mathcal{L}^{1|2}$. The symplectic two-superform in $\mathcal{M}^{8|4}$ reads

$$\Omega_s^{\mathrm{SUSY}} = d\Theta^{\mathrm{SUSY}} = -dx^a \wedge dp_a + \Omega_{\mathcal{L}^{1|2}}, \quad \Omega_{\mathcal{L}^{1|2}} = d\Sigma_{\mathcal{L}^{1|2}}. \tag{3.9}$$

The inner superspace is an $N=2$ superextension of the Lobachevsky plane. We show at first that $\mathcal{L}^{1|2}$ coincides with a coadjoint orbit of the OSp(2|2) supergroup. Let us introduce new complex Grassmann variables ($m\neq 0$, $s\neq 0$)

$$\theta=\sqrt{\frac{m}{s}}(iz^{\alpha}\chi_{\alpha}-z^{\alpha}\theta_{\alpha})\left[1+m\frac{1-b}{4s}(\theta^{\alpha}\theta_{\alpha}+\chi^{\alpha}\chi_{\alpha})\right],\quad \bar{\theta}=\overline{(\theta)},$$
$$\chi=\sqrt{\frac{m}{s}}(iz^{\alpha}\theta_{\alpha}-z^{\alpha}\chi_{\alpha})\left[1+m\frac{1+b}{4s}(\theta^{\alpha}\theta_{\alpha}+\chi^{\alpha}\chi_{\alpha})\right],\quad \bar{\chi}=\overline{(\chi)}, \tag{3.10}$$

which are in one-to-one correspondence with the Majorana spinors θ^{α} and χ^{α} used before.

It is easy to check that the symplectic two-superform

$$\Omega_{\mathcal{L}^{1|2}}=d\Sigma_{\mathcal{L}^{1|2}} \tag{3.11}$$

in the new holomorphic variables (3.10) exactly coincides with the one deduced by Gradechi and Nieto[39] in the supercoherent state's approach for the OSp(2|2) coadjoint orbits. $\Omega_{\mathcal{L}^{1|2}}$ is nondegenerate iff $|b|\neq 1$. In the case $|b|<1$, the supermanifold $\mathcal{L}^{1|2}$ is the regular OSp(2|2) coadjoint orbit $\mathcal{L}^{1|2}\cong$OSp(2|2)/[U(1)×U(1)] and is called the $N=2$ superunit disc. The degenerate orbit OSp(2|2)/U(1|1), which is denoted usually by $\mathcal{L}^{1|1}$ and called the $N=1$ superunit disc, appears when $|b|=1$. The other possibility, $|b|>1$, has no physical significance: neither the Poincaré supersymmetry nor the internal OSp(2|2) one admit unitary representations. It is seen from further consideration that the inequality $|b|>1$ contradicts the BPS bound.

C. Observables and the physical subspace

Consider in detail the realization of the Poincaré supersymmetry in the extended phase superspace $\mathcal{M}^{8|4}\cong T^{*}(R^{1,2})\times\mathcal{L}^{1|2}$. The Poincaré supergroup is realized by a symplectic action leaving the coadjoint orbit (3.7) invariant. The vector superfields generating the transformations (3.3) and (3.4) are related to the corresponding canonical Hamiltonian generators by

$$X_{H}\lrcorner\Omega_{s}^{\mathrm{SUSY}}=-(-1)^{\epsilon_{H}}dH, \tag{3.12}$$

where ϵ_{H} is the Grassmann parity of the Hamiltonian H. Solving these equations one gets the following Hamiltonian generators [we denote the generator of isotopic U(1) rotations by P_{3}]:

$$\mathcal{P}_{a}=p_{a},\quad \mathcal{J}_{a}=\epsilon_{abc}x^{b}p^{c}-sn_{a}+\tfrac{1}{2}mn_{a}(\theta^{\alpha}\theta_{\alpha}+\chi^{\alpha}\chi_{\alpha}-2ibn_{\alpha\beta}\theta^{\alpha}\chi^{\beta}),$$
$$\mathcal{Q}_{\alpha}^{1}=ip_{\alpha\beta}(\theta^{\beta}-ibn^{\beta}{}_{\gamma}\chi^{\gamma})+m(in_{\alpha\beta}\theta^{\beta}+b\chi_{\alpha}),\quad p_{\alpha\beta}\equiv p_{a}\gamma^{a}_{\alpha\beta},$$
$$\mathcal{Q}_{\alpha}^{2}=ip_{\alpha\beta}(\chi^{\beta}+ibn^{\beta}{}_{\gamma}\theta^{\gamma})+m(in_{\alpha\beta}\chi^{\beta}-b\theta_{\alpha}), \tag{3.13}$$
$$P_{3}=imn_{\alpha\beta}\theta^{\alpha}\chi^{\beta}-\frac{mb}{2}(\theta^{\alpha}\theta_{\alpha}+\chi^{\alpha}\chi_{\alpha}).$$

With respect to Poisson superbrackets on $\mathcal{M}^{8|4}$ they generate the following superalgebra:

$$\{\mathcal{J}_{a},\mathcal{J}_{b}\}=\epsilon_{abc}\mathcal{J}^{c},\quad \{\mathcal{J}_{a},\mathcal{P}_{b}\}=\epsilon_{abc}\mathcal{P}^{c},\quad \{\mathcal{J}_{a},\mathcal{Q}_{\alpha}^{I}\}=-\frac{i}{2}(\gamma_{a})_{\alpha}{}^{\beta}\mathcal{Q}_{\beta}^{I},$$
$$\{\mathcal{Q}_{\alpha}^{I},P_{3}\}=-\tfrac{1}{2}\epsilon^{IJ}\mathcal{Q}_{\alpha}^{J},\quad \{\mathcal{Q}_{\alpha}^{I},\mathcal{Q}_{\beta}^{J}\}\approx-2i\delta^{IJ}p_{\alpha\beta}-2\epsilon^{IJ}\epsilon_{\alpha\beta}\mathcal{Z},\quad \mathcal{Z}=mb, \tag{3.14}$$

the other brackets being equal to zero and $I,J=1,2$, $\epsilon^{IJ}=-\epsilon^{JI}$, $\epsilon^{01}=1$. We stress that the latter bracket $\{\mathcal{Q}_{\alpha}^{I},\mathcal{Q}_{\beta}^{J}\}$ is closed only in a weak sense, that is, modulo to constraints (3.7). What we have obtained is the $N=2$, $D=3$ Poincaré superalgebra with central charge $\mathcal{Z}=mb$ and isotopic charge P_{3} acting on the internal indices of supercharges $\mathcal{Q}_{\alpha}^{I}$.

One can easily examine that the mass and the spin Casimir functions of the superalgebra (3.14) read $C_1 \equiv \mathcal{P}^a \mathcal{P}_a = p^2$ and $C_2 \equiv \mathcal{P}^a \mathcal{J}_a + \frac{1}{8} \mathcal{Q}^{I\,\alpha} \mathcal{Q}^I_\alpha - \mathcal{Z} P_3 = -s(p,n)$. On the constraint surface (3.7),

$$p^2 + m^2 = 0, \quad (p,n) + m = 0, \tag{3.15}$$

the Casimirs are conserved identically. Equations (3.15) and (3.7) are completely equivalent to each other, in other words, they define one and the same surface in the phase superspace $\mathcal{M}^{8|4}$. We conclude that the mechanical model describes the $N=2$, $D=3$ superparticle of mass m, superspin s, and central charge mb.

Regular and degenerate cases are essentially distinguished for the coadjoint orbit, being associated for the superparticle. Since the massless and spinless particles are not covered in our model, the Bogomol'nyi–Prassad–Sommerfield bound of central charge (see, for instance, Ref. 41) assumes the only possibility for the degeneracy. The BPS bound $m \geqslant |\mathcal{Z}|$ provides, as is known, consistency of the quantum theory; the opposite inequality breaks the unitarity. As we have the goal to construct the quantum theory, we may restrict the consideration to the case of $|b| \leqslant 1$. Furthermore, the limiting point $|b| = 1$ corresponds to the multiplet-shortening.[41] It is the case $m = |\mathcal{Z}|$ when the massive multiplet contains the same number of particles as a massless one. These massive multiplets are called hypermultiplets. In the case of the $N=2$, $D=3$ Poincaré superalgebra, a massive supermultiplet of superspin s describes a quartet of particles with spins $s, s+\frac{1}{2}, s+\frac{1}{2}, s+1$ for $m > |\mathcal{Z}|$ and a doublet $s, s+\frac{1}{2}$ for $m = |\mathcal{Z}|$. The shortening of the superparticle multiplet has the respective origin in the classical mechanics: the number of odd physical degrees of freedom of the superparticle halved in the BPS limit. Let us show that it is the case which is described by our model.

Reducing to the constraints (3.15) [or, equivalently, (3.7)] we come to the smaller 5/4-dimensional phase space $\mathcal{M}^{5|4} \subset \mathcal{M}^{8|4}$ with a degenerate symplectic two-superform

$$\Omega_s^{\mathrm{SUSY}}\big|_{p_a = m n_a} \equiv \Omega_s^{\mathrm{red}} = -m\, dx^a \wedge dn_a + \Omega_{\mathcal{L}^{1|2}}, \quad dn_a \equiv \frac{2\xi_a d\bar{z} + 2\bar{\xi}_a dz}{(1 - z\bar{z})^2}, \tag{3.16}$$

where $\Omega_{\mathcal{L}^{1|2}}$ is defined by Eq. (3.11) and

$$\xi_a = -\tfrac{1}{2}(\gamma_a)_{\alpha\beta} z^\alpha z^\beta = -\tfrac{1}{2}(2z, 1+z^2, i(z^2-1)), \quad \bar{\xi}_a = \overline{(\xi_a)}. \tag{3.17}$$

The kernel of the two-superform (3.16) contains obviously the even one-dimensional null space $\mathrm{Ker}_0 \Omega_s^{\mathrm{red}}$, related to the reparametrization invariance of the world lines. In the coset superspace $\mathcal{O}_{m,s,b} = \mathcal{M}^{5|4} / \mathrm{Ker}_0\, \Omega_s^{\mathrm{red}}$ the induced symplectic two-superform is nondegenerate when $|b| < 1$; the same is true in $\mathcal{L}^{1|2}$ for the respective superform $\Omega_{\mathcal{L}^{1|2}}$. Therefore, $\mathcal{O}_{m,s,b}$, $\dim \mathcal{O}_{m,s,b} = 4/4$, $|b| < 1$ is isomorphic to a regular coadjoint orbit of the $N=2$, $D=3$ Poincaré supergroup. We have established both the embedding of the regular orbit into the original phase superspace and the underlying projection $\pi: \mathcal{M}^{8|4} \to \mathcal{O}_{m,s,b}$, provided by constraints (3.15).

In the BPS limit $|b| = 1$, the inner two-superform $\Omega_{\mathcal{L}^{1|2}}$ generates a 0/2-dimensional null-vector superspace. To make the degeneracy more evident we introduce for a while new odd variables

$$\tilde{\theta}^\alpha = \theta^\alpha - i n^\alpha{}_\beta \chi^\beta, \quad \tilde{\chi}^\alpha = \chi^\alpha - i n^\alpha{}_\beta \theta^\beta$$

instead of $\theta^\alpha, \chi^\alpha$. This change of the odd variables is one-to-one, and the original Lagrangian (3.1) reads in new variables as

$$L = m(\dot{x}, n) - im \frac{1+b}{2} n_{\alpha\beta} \tilde{\theta}^\alpha \dot{\tilde{\theta}}^\beta - im \frac{1-b}{2} n_{\alpha\beta} \tilde{\chi}^\alpha \dot{\tilde{\chi}}^\beta + is \frac{\bar{z}\dot{z} - z\dot{\bar{z}}}{1 - z\bar{z}}. \tag{3.1$'$}$$

It is seen immediately that half of the odd degrees of freedom of the superparticle drops out from the theory in the case of $|b|=1$. The full kernel Ker Ω_s^{red} of the symplectic two-superform on $\mathcal{M}^{5|4}$ becomes 1/2-dimensional if $|b|=1$. The 4/2-dimensional coset superspace $\mathcal{O}_{m,s}=\mathcal{M}^{5|4}/\mathrm{Ker}\,\Omega_s^{\mathrm{red}}$ corresponds to a degenerate orbit of the $N=2$ Poincaré supergroup. Hence, the number of odd physical degrees of freedom of the $N=2$, $D=3$ superparticle halved actually in the BPS limit and we observe an evident classical analog of the multiplet-shortening. Moreover, in the BPS limit, expression (3.1′) reduces to the Lagrangian of the $N=1$, $D=3$ superparticle[25] and does describe after quantization not a superquartet, but a supersymmetric doublet of particles of equal mass m and spins s and $s+\frac{1}{2}$ only.

D. Hidden su(1,1|2) supersymmetry of the superspin degrees of freedom

We have shown that the superparticle dynamics is embedded in the phase superspace $\mathcal{M}^{8|4}\cong T^*(R^{1,2})\times\mathcal{L}^{1|2}$. One can imply that the inner supermanifold $\mathcal{L}^{1|2}$ carries internal (both even and odd) degrees of freedom of the $D=3$ particle. Then the symplectomorphisms of $\mathcal{L}^{1|2}$ should be treated as the hidden supersymmetry of the particle internal structure. Consider this supersymmetry in more detail. To be specific, let us assume that $|b|<1$. The degenerate case is already discussed in Ref. 25.

We have already mentioned that $\mathcal{L}^{1|2}$ is a homogeneous OSp(2|2) superspace. Introducing new odd complex variables (3.10), we established that the symplectic two-superform (3.11) reduces to the superform on the regular OSp(2|2) coadjoint orbit obtained earlier in Refs. 38 and 39 in the framework of the supercoherent state technique. A crucial point is that $\mathcal{L}^{1|2}$ reveals a *Kähler* supermanifold structure with the superpotential

$$\Phi=-2s\ln(1-z\bar{z})-s(1+b)\frac{\theta\bar{\theta}}{1-z\bar{z}}-s(1-b)\frac{\chi\bar{\chi}}{1-z\bar{z}}+\frac{s(1-b^2)}{2}\frac{\theta\bar{\theta}\chi\bar{\chi}}{(1-z\bar{z})^2},\tag{3.18}$$

so that

$$\Omega_{\mathcal{L}^{1|2}}=i\left(d\bar{z}\frac{\partial}{\partial\bar{z}}+d\bar{\theta}\frac{\vec{\partial}}{\partial\bar{\theta}}+d\bar{\chi}\frac{\vec{\partial}}{\partial\bar{\chi}}\right)\wedge\left(dz\frac{\partial}{\partial z}+d\theta\frac{\vec{\partial}}{\partial\theta}+d\chi\frac{\vec{\partial}}{\partial\chi}\right)\Phi,$$

and OSp(2|2) acts on the $N=2$ superunit disc by the *superholomorphic* transformations. Moreover, the supergroup of the superholomorphic symplectomorphisms of $\mathcal{L}^{1|2}$ is, in fact, essentially larger than OSp(2|2) and it contains at least the supergroup SU(1,1|2). The corresponding infinitesimal transformations read

$$\delta z=i\omega^a\xi_a-\frac{\sqrt{1+b}}{2}\epsilon_\alpha z^\alpha\theta-\frac{\sqrt{1-b}}{2}\eta_\alpha z^\alpha\chi,$$

$$\delta\theta=\frac{i}{2}\omega^a\partial\xi_a\theta+\frac{i}{2}\sqrt{\frac{1-b}{1+b}}\mu_1\chi-\frac{i}{2}(\mu_2+\mu_3)\theta-\frac{1}{\sqrt{1+b}}\bar{\epsilon}_\alpha z^\alpha-\frac{\sqrt{1-b}}{2}\eta_\alpha\partial z^\alpha\theta\chi,\tag{3.19}$$

$$\delta\chi=\frac{i}{2}\omega^a\partial\xi_a\chi-\frac{i}{2}\sqrt{\frac{1+b}{1-b}}\bar{\mu}_1\theta-\frac{i}{2}(\mu_2-\mu_3)\chi+\frac{\sqrt{1+b}}{2}\epsilon_\alpha\partial z^\alpha\theta\chi-\frac{1}{\sqrt{1-b}}\bar{\eta}_\alpha z^\alpha,$$

where $\partial\equiv\partial/\partial z$, even parameters ω^a, μ_2, and μ_3 are real, even parameter μ_1 is complex, and the odd ones ϵ_α and η_α are complex. Transformations (3.19) are generated by the following Hamiltonians, which may be obtained straightforwardly solving Eqs. (3.12). There are seven (real) even Hamiltonians,

$$J_a=-sn_a\left(1-\frac{1+b}{2}\frac{\theta\overline{\theta}}{1-z\overline{z}}-\frac{1-b}{2}\frac{\chi\overline{\chi}}{1-z\overline{z}}+\frac{1-b^2}{2}\frac{\theta\overline{\theta}\chi\overline{\chi}}{(1-z\overline{z})^2}\right),$$

$$P_1=s\frac{\sqrt{1-b^2}}{2}\frac{\theta\overline{\chi}-\overline{\theta}\chi}{1-z\overline{z}},\quad P_3=-s\left(\frac{1+b}{2}\frac{\theta\overline{\theta}}{1-z\overline{z}}-\frac{1-b}{2}\frac{\chi\overline{\chi}}{1-z\overline{z}}\right), \tag{3.20a}$$

$$P_2=is\frac{\sqrt{1-b^2}}{2}\frac{\theta\overline{\chi}+\overline{\theta}\chi}{1-z\overline{z}},\quad P_4=-s\left(\frac{1+b}{2}\frac{\theta\overline{\theta}}{1-z\overline{z}}+\frac{1-b}{2}\frac{\chi\overline{\chi}}{1-z\overline{z}}-\frac{1-b^2}{2}\frac{\theta\overline{\theta}\chi\overline{\chi}}{(1-z\overline{z})^2}\right),$$

and eight odd ones,

$$E^\alpha=s\sqrt{1+b}\left(\frac{z^\alpha\overline{\theta}-\overline{z}^\alpha\theta}{1-z\overline{z}}\right)\left(1-\frac{1-b}{2}\frac{\chi\overline{\chi}}{1-z\overline{z}}\right),\quad F^\alpha=in^\alpha{}_\beta E^\beta$$

$$G^\alpha=s\sqrt{1-b}\left(\frac{z^\alpha\overline{\chi}-\overline{z}^\alpha\chi}{1-z\overline{z}}\right)\left(1-\frac{1+b}{2}\frac{\theta\overline{\theta}}{1-z\overline{z}}\right),\quad H^\alpha=in^\alpha{}_\beta G^\beta. \tag{3.20b}$$

These Hamiltonians, together with one more even element $Z\equiv s$, generate a closed superalgebra with respect to Poisson superbrackets on $\mathcal{L}^{1|2}$ (the explicit form of these superbrackets may be found in Ref. 36). This is the so-called su(1,1|2) superalgebra,[42] whose even part is $\mathrm{su}(1,1|2)_0=\mathrm{su}(1,1)\oplus\mathrm{u}(2)\oplus R=\{J_a,P_I,Z\}$ and the odd part constitutes an eight-dimensional module of the even part; Z presents a central charge. The osp(2|2) subsuperalgebra found in Refs. 38 and 39 is spanned by J_a, B, $\sqrt{ms}V^\alpha$, and $\sqrt{ms}W^\alpha$, where V^α and W^α are defined below by Eqs. (3.26) and, $B=P_3-bZ$. We reveal that the $N=2$ superunit disc is not only a typical coadjoint orbit of the OSp(2|2) supergroup, $\mathcal{L}^{1|2}\cong\mathrm{OSp}(2|2)/[\mathrm{U}(1)\times\mathrm{U}(1)]$, but it can be treated simultaneously as an atypical Kähler orbit of the supergroup $\mathrm{SU}(1,1|2):\mathcal{L}^{1|2}\cong\mathrm{SU}(1,1|2)/[\mathrm{U}(2|2)\times\mathrm{U}(1)]$.

E. Hidden $N=4$ Poincaré supersymmetry

Subalgebra u(2) of the internal su(1,1|2) superalgebra acts on the odd variables, as is seen from Eqs. (3.19). It is exactly the subalgebra of the isotopic symmetry. However, the isotopic U(2) symmetry may now be involved in the Poincaré supersymmetry. The isotopic rotations together with the $N=2$ Poincaré transformations (3.3) and (3.4) generate (when $|b|\neq 1$) the wider $D=3$, $N=4$ Poincaré superalgebra. In addition to (3.3) there are the following supersymmetry transformations:

$$\delta_{\tilde{\epsilon}}x^a=ib\gamma^a_{\alpha\beta}\tilde{\epsilon}^\alpha\chi^\beta-i\epsilon^{abc}n_b\gamma_{c\alpha\beta}\tilde{\epsilon}^\alpha\theta^\beta+n^a\tilde{\epsilon}^\alpha\theta_\alpha,\quad \delta_{\tilde{\epsilon}}\theta^\alpha=-in^\alpha{}_\beta\tilde{\epsilon}^\beta,\quad \delta_{\tilde{\epsilon}}\chi^\alpha=\delta_{\tilde{\epsilon}}z=0,$$

$$\delta_{\tilde{\eta}}x^a=-ib\gamma^a_{\alpha\beta}\tilde{\eta}^\alpha\theta^\beta-i\epsilon^{abc}n_b\gamma_{c\alpha\beta}\tilde{\eta}^\alpha\chi^\beta+n^a\tilde{\eta}^\alpha\chi_\alpha,\quad \delta_{\tilde{\eta}}\chi^\alpha=-in^\alpha{}_\beta\tilde{\eta}^\beta,\quad \delta_{\tilde{\eta}}\theta^\alpha=\delta_{\tilde{\eta}}z=0, \tag{3.21}$$

where $\overline{\epsilon}^\alpha,\overline{\eta}^\alpha$ are odd infinitesimal parameters. The respective Hamiltonians on $\mathcal{M}^{8|4}$ read

$$\tilde{\mathcal{Q}}^1_\alpha=ip_{\alpha\beta}(in^\beta{}_\gamma\theta^\gamma+b\chi^\beta)-m(\theta_\alpha-ibn_{\alpha\beta}\chi^\beta)\approx-\frac{i}{m}p_\alpha{}^\beta\mathcal{Q}^1_\beta,$$

$$\tilde{\mathcal{Q}}^2_\alpha=ip_{\alpha\beta}(in^\beta{}_\gamma\chi^\gamma-b\theta^\beta)-m(\chi_\alpha+ibn_{\alpha\beta}\theta^\beta)\approx-\frac{i}{m}p_\alpha{}^\beta\mathcal{Q}^2_\beta. \tag{3.22}$$

New supercharges together with $N=2$ superpoincaré Hamiltonians (3.13) and isotopic U(2) Hamiltonians P^I, $I=1,2,3,4$ generate the closed $N=4$ Poincaré superalgebra with one central charge. It can be seen by introducing a new basis for supercharges $2\mathcal{R}^I_\alpha=(\mathcal{Q}^1_\alpha+\tilde{\mathcal{Q}}^2_\alpha,\mathcal{Q}^2_\alpha-\tilde{\mathcal{Q}}^1_\alpha)$, $2\tilde{\mathcal{R}}^I_\alpha=(\tilde{\mathcal{Q}}^1_\alpha+\mathcal{Q}^2_\alpha,\tilde{\mathcal{Q}}^2_\alpha-\mathcal{Q}^1_\alpha)$, $I=1,2$. On shell (3.7) we have

$$\{\mathcal{R}^I_\alpha,\mathcal{R}^J_\beta\}\approx(1-b)(-i\delta^{IJ}p_{\alpha\beta}+m\epsilon^{IJ}\epsilon_{\alpha\beta}),$$
$$\{\tilde{\mathcal{R}}^I_\alpha,\tilde{\mathcal{R}}^J_\beta\}\approx(1+b)(-i\delta^{IJ}p_{\alpha\beta}+m\epsilon^{IJ}\epsilon_{\alpha\beta}),\quad \{\mathcal{R}^I_\alpha,\tilde{\mathcal{R}}^J_\beta\}\approx 0. \tag{3.23}$$

The invariance of the original Lagrangian (3.1) under the transformations (3.21) can be examined straightforwardly. Thus, the model, being $N=2$ super-Poincaré invariant by construction, allows the hidden $N=4$ supersymmetry. The appearance of the enhanced supersymmetry is hardly surprising in the model. This $N=4$ supersymmetry is degenerate in a sense that the corresponding central charges equals to m and, so, they saturate the BPS bound for the $N=4$ Poincaré superalgebra. It reflects the degeneracy of the $N=4$ supersymmetry and the shortening of the $N=4$ superparticle multiplet to the $N=2$ supermultiplet in quantum theory. Moreover, it is a general property of extended supersymmetry that some of the degenerate multiplets of a larger SUSY (those which saturate the BPS bound) have the same particle content, as is observed in the respective multiplets of a smaller SUSY. This fact provides a simple reason why some of supersymmetric theories may have the extended supersymmetries. The precedents are known both for $D=4,6,10$ superparticle models[43] and supersymmetric field theories (for example, the theories with nontrivial topological charge[44]). The $D=3$, $N=1$ superparticle allows the hidden $N=2$ SUSY.[25]

F. Relationship between Hamiltonian generators of the Poincaré and internal supersymmetries

We have observed that the model contains both the global Poincaré SUSY and the hidden SU(1,1|2). The latter is closely related to the superspin intrinsic structure. Thus, the relevant quantization procedure should make a provision for either symmetry to survive in quantum theory. This quantization can be based on the simple fact that the Hamiltonian generators (3.13) and (3.22) of the Poincaré supersymmetries, being the functions on $\mathcal{M}^{8|4}\cong T^*(R^{1,2})\times\mathcal{L}^{1|2}$, can be expressed in terms of the Minkowski-space coordinates and momenta (x^a,p_a) and of the su(1,1|2) Hamiltonians J_a, P_I, E^α, F^α, G^α, and H^α (3.20), which parametrize the coadjoint orbit $\mathcal{L}^{1|2}$. We give here the explicit form of these expressions:

$$\mathcal{J}_a=\epsilon_{abc}x^bp^c+J_a,\quad \mathcal{P}_a=p_a,\quad \mathcal{Z}=mb,$$
$$\mathcal{Q}^1_\alpha=(ip_{\alpha\beta}W^\beta+m\tilde{W}_\alpha)[1+q^{cl}(bP_3-\sqrt{1-b^2}P_2-P_4)], \tag{3.24}$$
$$\mathcal{Q}^2_\alpha=(ip_{\alpha\beta}V^\beta+m\tilde{V}_\alpha)[1+q^{cl}(bP_3+\sqrt{1-b^2}P_2-P_4)],$$

$$\tilde{\mathcal{Q}}^1_\alpha=(ip_{\alpha\beta}\tilde{W}^\beta-mW_\alpha)[1+q^{cl}(bP_3-\sqrt{1-b^2}P_2-P_4)],$$
$$\tilde{\mathcal{Q}}^2_\alpha=(ip_{\alpha\beta}\tilde{V}^\beta-mV_\alpha)[1+q^{cl}(bP_3+\sqrt{1-b^2}P_2-P_4)], \tag{3.25}$$

where

$$W^\alpha=\frac{1}{2\sqrt{ms}}(\sqrt{1+b}E^\alpha+\sqrt{1-b}H^\alpha),\quad \tilde{W}^\alpha=\frac{1}{2\sqrt{ms}}(\sqrt{1+b}F^\alpha-\sqrt{1-b}G^\alpha),$$
$$V^\alpha=\frac{1}{2\sqrt{ms}}(\sqrt{1+b}F^\alpha+\sqrt{1-b}G^\alpha),\quad \tilde{V}^\alpha=\frac{1}{2\sqrt{ms}}(\sqrt{1-b}H^\alpha-\sqrt{1+b}E^\alpha), \tag{3.26}$$

and constant q^{cl} reads as

$$q^{cl}=\frac{1}{4s}. \tag{3.27}$$

To construct an appropriate operator realization of these expressions we shall quantize (x^a, p_a) canonically and extend simultaneously su(1,1|2) Hamiltonian vector fields to a representation by Hermitian operators in Hilbert space.

Notice at once some important details in relation to the quantization which should be compatible to the full symmetry of the superparticle. First, expressions (3.24) and (3.25) are essentially *nonlinear* in the generators (3.20) of the inner su(1,1|2) superalgebra. Thus, even though the operator realization of the Poisson su(1,1|2) superalgebra is found and the corresponding operators are substituted in Eq. (3.24), we may not be sure that the representation of the Poincaré superalgebra (neither $N=2$ nor $N=4$) is reproduced for certain in quantum theory. Because of the nonlinearity, the superalgebra of operators, corresponding to (3.24), might be disclosed, and it is the parameter q that controls the possible disclosure of the Poincaré superalgebra. We will see that the parameter q^{cl} should be renormalized in quantum theory to reproduce a representation of the Poincaré supersymmetry.

Second, it is a matter of direct verification that the Hamiltonians W^α and $\tilde{W}^\alpha$ have vanishing Poisson superbrackets with $bP_3 - \sqrt{1-b^2}P_2 - P_4$, whereas V^α and $\tilde{V}^\alpha$ commute to $bP_3 + \sqrt{1-b^2}P_2 - P_4$. This point is important for Hermitian properties of supercharge operators in quantum mechanics.

IV. FIRST QUANTIZATION OF THE SUPERPARTICLE

It is a primary objective of previous consideration to present the classical model of the $N=2$, $D=3$ superparticle in the form, well adapted for a quantizing procedure. We have obtained an embedding of the (maximal) coadjoint orbit of the $N=2$ Poincaré supergroup in the extended phase superspace $\mathcal{M}^{8|4} \cong T^*(R^{1,2}) \times \mathcal{L}^{1|2}$. Going to quantum theory we will combine the canonical Dirac quantization on $T^*(R^{1,2})$ and the geometric quantization methods on the SU(1,1|2) coorbit $\mathcal{L}^{1|2}$. In particular, a combination of the standard real polarization in $T^*(R^{1,2})$ and the Kähler one in $\mathcal{L}^{1|2}$ will be used to construct the superparticle's Hilbert space.

The quantization scheme implies from the outset that the internal SU(1,1|2) supersymmetry must survive at the quantum level. Mutual relation between Hamiltonians of SU(1,1|2) and Poincaré supersymmetries, being expressed by Eq. (3.24) and (3.25), is crucial in our approach. At first, we construct the operator realization for the Hamiltonians of the su(1,1|2) superalgebra in the framework of Berezin quantization. Then the expressions (3.24) [possibly together with (3.25)] are used to obtain the realization of a unitary irreducible representation (UIR) for the $N=2$ (respectively, enhanced $N=4$) Poincaré superalgebra. We find that the classical meaning of the parameter q in (3.27) in the relations (3.24) and (3.25) should be accompanied by certain quantum corrections, referred to as a renormalization, for consistency of the quantum theory.

Eventually we obtain the straightforward $N=2$ supergeneralization of the conventional realization of the unitary irreducible representations (UIRs) of the $D=3$ Poincaré group on the fields carrying representations of $\overline{\mathrm{SO}^{\uparrow}(1,2)}$.[6,13] Two cases should be distinguished among these representations. The fields describing fractional superspin (superanyons) carry a atypical unitary *infinite-dimensional* UIRs of SU(1,1|2), whereas the UIRs of (half) integer superspin can be realized on the spin-tensor fields carrying atypical *finite dimensional* nonunitary representations of SU(1,1|2). The realization of the superparticle Hilbert space is slightly different in these two cases.

A. Berezin quantization on $\mathcal{L}^{1|2}$

The Berezin technique[30,31,45,46] provides the perfect quantization method for the Kähler homogeneous spaces. We consider here briefly the application of this method to the supermanifold $\mathcal{L}^{1|2} \cong \mathrm{OSp}(2|2)/[\mathrm{U}(1)\times\mathrm{U}(1)] \cong \mathrm{SU}(1,1|2)/[\mathrm{U}(2|2)\times\mathrm{U}(1)]$ with the nondegenerate symplectic structure when $|b|<1$. The geometric quantization on $\mathcal{L}^{1|2}$, being considered as a regular coadjoint orbit, is studied in Refs. 38 and 39 in detail. However, as we know, $\mathcal{L}^{1|2}$ has not been considered as an irregular SU(1,1|2) coorbit nor as a detailed Berezin quantization, and the underlying correspondence principle is not explicitly established.

In the following subsection we apply the obtained results for quantization of the $D=3$ superparticle.

1. Antiholomorphic sections and an inner product

Let us consider the space $\mathcal{O}_{s,b}$ of superantiholomorphic sections of the superholomorphic line bundle over $\mathcal{L}^{1|2}$, whose elements are represented by functions

$$f(\overline{\Gamma})\equiv f(\overline{z},\overline{\theta},\overline{\chi})=f_0(\overline{z})+\sqrt{s(1+b)}\,\overline{\theta}f_1(\overline{z})+\sqrt{s(1-b)}\,\overline{\chi}f_2(\overline{z})+\sqrt{s(s+1/2)(1-b^2)}\,\overline{\theta}\overline{\chi}f_3(\overline{z}), \tag{4.1}$$

where $f_i(\overline{z})$, $i=0,1,2,3$, are ordinary antiholomorphic functions on the unit disc of the complex plane. We denote by $\Gamma\equiv\{\Gamma^A\}=\{z,\theta,\chi\}$ and $\overline{\Gamma}\equiv\{\overline{\Gamma}^A\}=\{\overline{z},\overline{\theta},\overline{\chi}\}$ the sets of the superholomorphic and superantiholomorphic variables, respectively. The space $\mathcal{O}_{s,b}$ is equipped naturally by an inner product

$$\langle f|g\rangle_{\mathcal{L}^{1|2}}=\int_{\mathcal{L}^{1|2}}\overline{f(\overline{\Gamma})}g(\overline{\Gamma})e^{-\Phi(\Gamma,\overline{\Gamma})}\,d\mu(\Gamma,\overline{\Gamma}). \tag{4.2a}$$

Here $\Phi(\Gamma,\overline{\Gamma})$ is the Kähler superpotential (3.18) and $d\mu(\Gamma,\overline{\Gamma})$ is an SU(1,1|2) invariant Liouville supermeasure on $\mathcal{L}^{1|2}$. Taking into account the definition of the symplectic two-superform (3.11) $\Omega_{\mathcal{L}^{1|2}}\equiv d\Gamma^A\Omega_{A\overline{B}}d\overline{\Gamma}^B$, one can derive the supermeasure explicitly:[38,39]

$$d\mu(\Gamma,\overline{\Gamma})=-\frac{1}{4\pi}\,\mathrm{sdet}\|\Omega_{A\overline{B}}\|\,d\Gamma d\overline{\Gamma}=\frac{d\Gamma d\overline{\Gamma}}{i\pi s(1-b^2)},\quad d\Gamma d\overline{\Gamma}\equiv dz\,d\overline{z}\,d\theta\,d\overline{\theta}\,d\chi\,d\overline{\chi}. \tag{4.3}$$

Using Eqs. (3.18), (4.1), and (4.3) we can integrate out the Grassmann variables in Eq. (4.2a), which reduces the inner product to the following form:

$$\langle f|g\rangle_{\mathcal{L}^{1|2}}=\langle f_0|g_0\rangle_{\mathcal{L}}^{s}+\langle f_1|g_1\rangle_{\mathcal{L}}^{s+1/2}+\langle f_2|g_2\rangle_{\mathcal{L}}^{s+1/2}+\langle f_3|g_3\rangle_{\mathcal{L}}^{s+1}, \tag{4.2b}$$

where

$$\langle\varphi|\chi\rangle_{\mathcal{L}}^{l}=(2l-1)\int_{|z|<1}\frac{dz\,d\overline{z}}{2\pi i}(1-z\overline{z})^{2l-2}\overline{\varphi(\overline{z})}\chi(\overline{z}) \tag{4.4}$$

is an inner product in the representation space D_+^l of $\overline{\mathrm{SO}^{\uparrow}(1,2)}$ discrete series bounded below, being realized by antiholomorphic functions in the unit disc $|z|<1$. {The monomials $\phi_n^l=[\Gamma(2l+n)/\Gamma(n-1)\Gamma(2l)]^{1/2}\overline{z}^n$, n is a non-negative integer, serve as a standard orthonormal basis in D_+^l.} The inner product (4.4) is well defined and positive if $l>\frac{1}{2}$. Moreover, for values $0<l<1/2$ one can still use Eq. (4.4) if suitable analytic continuations are made. The case of $l=\frac{1}{2}$ should be understood in the sense of the limit. We conclude that the inner product (4.2) in $\mathcal{O}_{s,b}$ is well defined if $s>0$ (and, of course, if $|b|<1$.).

In view of the transformation law for Kähler superpotential $\Phi(\Gamma,\overline{\Gamma})$ under the action of SU(1,1|2) supergroup, the inner product (4.2) holds to be SU(1,1|2) invariant, if an appropriate transformation law for $f(\Gamma)\in\mathcal{O}_{s,b}$ is implemented. In other terms, the Hamiltonian action of SU(1,1|2) on $\mathcal{L}^{1|2}$ can be lifted to a unitary representation in $\mathcal{O}_{s,b}$. We give below an infinitesimal form of this representation only, that is, explicit representation of corresponding superalgebra su(1,1|2). To obtain it, we first consider a conventional correspondence between linear operators in $\mathcal{O}_{s,b}$ and Berezin's symbols.

2. Classical observables and operators

Let $A(\underline{\Gamma},\overline{\Gamma})$ be a "classical observable," that means it is a real function on $\mathcal{L}^{1|2}$ to be continuously differentiable in $z,\overline{z}$ that the integrals considered below do exist. We associate a linear operator $\hat{A}$ in $\mathcal{O}_{s,b}$ to the classical observable $A(\underline{\Gamma},\overline{\Gamma})$ by the rule

$$(\hat{A}f)(\overline{\Gamma})=\int_{\mathcal{L}^{1|2}} A(\underline{\Gamma}_1,\overline{\Gamma})f(\overline{\Gamma}_1)L_{s,b}(\underline{\Gamma}_1,\overline{\Gamma})e^{-\Phi(\underline{\Gamma}_1\overline{\Gamma}_1)}\,d\mu(\underline{\Gamma}_1,\overline{\Gamma}_1), \tag{4.5}$$

where $A(\underline{\Gamma}_1,\overline{\Gamma})$ serves only as an analytic continuation in $\mathcal{L}^{1|2}\times\mathcal{L}^{1|2}$ for classical observable $A(\underline{\Gamma},\overline{\Gamma})$. The generating kernel $L_{s,b}(\underline{\Gamma}_1,\overline{\Gamma})$ can be constructed by the use of an arbitrary complete orthonormal basis $f_k(\overline{\Gamma})$ in $\mathcal{O}_{s,b}$, and appears to be related immediately to the analytic continuation in $\mathcal{L}^{1|2}\times\mathcal{L}^{1|2}$ of the Kähler superpotential:

$$L_{s,b}(\underline{\Gamma}_1,\overline{\Gamma})=\sum_{k=1}^{\infty} f_k(\overline{\Gamma})\overline{f_k(\overline{\Gamma}_1)}=\exp[\Phi(\underline{\Gamma}_1,\overline{\Gamma})]. \tag{4.6}$$

The state, being presented by the function $\Phi_{\underline{\Gamma}}(\overline{\Gamma}_1)=L_{s,b}(\underline{\Gamma},\overline{\Gamma}_1)$ with fixed $\underline{\Gamma}\equiv\{\overline{z},\overline{\theta},\overline{\chi}\}$ is denoted by $|\overline{z},\overline{\theta},\overline{\chi}\rangle$, is called as an SU(1,1|2) [or OSp(2|2)] supercoherent state. The analytic continuation in $\mathcal{L}^{1|2}\times\mathcal{L}^{1|2}$ for any classical observable could be expressed in terms of the supercoherent states as follows:

$$A(\underline{\Gamma}_1,\overline{\Gamma}_2)=\frac{\langle\Phi_{\overline{\Gamma}_2}|\hat{A}|\Phi_{\overline{\Gamma}_1}\rangle_{\mathcal{L}^{1|2}}}{\langle\Phi_{\overline{\Gamma}_2}|\Phi_{\overline{\Gamma}_1}\rangle_{\mathcal{L}^{1|2}}}. \tag{4.7}$$

So, the symbol of the unit operator $\hat{I}$ is just 1. Hence, the one-to-one correspondence between classical observables on $\mathcal{L}^{1|2}$ and linear operators in $\mathcal{O}_{s,b}$ is established. In view of Eq. (4.7), classical observables are also referred to as (covariant) Berezin symbols.

3. Atypical unitary and finite-dimensional representations of the su(1,1|2) superalgebra

Using Eq. (4.5), one can now obtain the operators which correspond to the Hamiltonian generators (3.20) of holomorphic transformations of the $N=2$ superunit disc. One gets

$$\hat{J}_a=-\overline{\xi_a}\overline{\partial}-(\overline{\partial\xi_a})\left(s+\frac{1}{2}\overline{\theta}\frac{\partial}{\partial\overline{\theta}}+\frac{1}{2}\overline{\chi}\frac{\partial}{\partial\overline{\chi}}\right),\quad \hat{Z}=s\hat{I},$$

$$\hat{P}_1=-\frac{1}{\sqrt{1-b^2}}\left(\frac{1-b}{2}\overline{\chi}\frac{\partial}{\partial\overline{\theta}}+\frac{1+b}{2}\overline{\theta}\frac{\partial}{\partial\overline{\chi}}\right),\quad \hat{P}_3=\frac{1}{2}\overline{\theta}\frac{\partial}{\partial\overline{\theta}}-\frac{1}{2}\overline{\chi}\frac{\partial}{\partial\overline{\chi}},$$

$$\hat{P}_2=\frac{i}{\sqrt{1-b^2}}\left(\frac{1-b}{2}\overline{\chi}\frac{\partial}{\partial\overline{\theta}}-\frac{1+b}{2}\overline{\theta}\frac{\partial}{\partial\overline{\chi}}\right),\quad \hat{P}_4=\frac{1}{2}\overline{\theta}\frac{\partial}{\partial\overline{\theta}}+\frac{1}{2}\overline{\chi}\frac{\partial}{\partial\overline{\chi}},$$

$$\hat{E}^{\alpha}=\frac{\sqrt{1+b}}{2}\overline{\theta}\left[\overline{z^{\alpha}}\overline{\partial}+(\overline{\partial}\overline{z^{\alpha}})\left(2s+\overline{\chi}\frac{\partial}{\partial\overline{\chi}}\right)\right]-\frac{1}{\sqrt{1+b}}\overline{z^{\alpha}}\frac{\partial}{\partial\overline{\theta}}, \tag{4.8}$$

$$\hat{F}^{\alpha}=-i\frac{\sqrt{1+b}}{2}\overline{\theta}\left[\overline{z^{\alpha}}\overline{\partial}+(\overline{\partial}\overline{z^{\alpha}})\left(2s+\overline{\chi}\frac{\partial}{\partial\overline{\chi}}\right)\right]-i\frac{1}{\sqrt{1+b}}\overline{z^{\alpha}}\frac{\partial}{\partial\overline{\theta}},$$

$$\hat{G}^{\alpha}=\frac{\sqrt{1-b}}{2}\overline{\chi}\left[\overline{z^{\alpha}}\overline{\partial}+(\overline{\partial}\overline{z}^{\alpha})\left(2s+\overline{\theta}\frac{\partial}{\partial\overline{\theta}}\right)\right]-\frac{1}{\sqrt{1-b}}\overline{z}^{\alpha}\frac{\partial}{\partial\overline{\chi}},$$

$$\hat{H}^{\alpha}=-i\,\frac{\sqrt{1-b}}{2}\overline{\chi}\left[\overline{z^{\alpha}}\overline{\partial}+(\overline{\partial}\overline{z}^{\alpha})\left(2s+\overline{\theta}\frac{\partial}{\partial\overline{\theta}}\right)\right]-i\,\frac{1}{\sqrt{1-b}}\overline{z}^{\alpha}\frac{\partial}{\partial\overline{\chi}},$$

where $\overline{\partial}\equiv\partial/\partial\overline{z},$ all the derivatives are left, and $\overline{\xi_a},\overline{z^{\alpha}}$ are defined by Eqs. (3.17) and (3.8), respectively. It is readily verified that the derived operators generate an irreducible representation of the su(1,1|2) superalgebra, and it is the same case for any values s and b, not only for $s>0$ and $|b|<1$. The anticommutation relations for operators (4.8) completely correspond to the Poisson superbrackets of the classical observables (3.20), and it is sufficient to apply the correspondence rules, that is, to replace $\{,\}\rightarrow 1/i[\,,]_{\mp}$ (anticommutator for two odd operators and commutator in the rest cases). By reduction to the orthosymplectic subsuperalgebra we reproduce just the typical UIRs of the osp(2|2) obtained in Refs. 38 and 39.

The constructed representation is infinite dimensional for $s>0$ and $|b|<1$, and unitary in the sense that the operators (4.8) are Hermitian with respect to inner product (4.2). It means, in particular, that $\langle f|\hat{J}_a|g\rangle_{\mathcal{L}^{1|2}}=\langle g|\hat{J}_a|f\rangle_{\mathcal{L}^{1|2}}$ and $\langle f|\hat{P}_I|g\rangle_{\mathcal{L}^{1|2}}=\langle g|\hat{P}_I|f\rangle_{\mathcal{L}^{1|2}}$ for any $f,g\in\mathcal{O}_{s,b}$. The Hermitian self-conjugation conditions for the odd operators may reveal some subtlety. Any odd classical observable among (3.20) is the Majorana spinor and we have, for example, $\overline{E^0}=-E^1$, $\overline{E^1}=-E^0$ with respect to the reality condition (3.2). $\hat{E}^{\alpha}$ (and any odd operator with the spinor index) is Hermitian in the sense that $\langle f|\hat{E}^0|g\rangle_{\mathcal{L}^{1|2}}=-\langle g|\hat{E}^1|f\rangle_{\mathcal{L}^{1|2}}$.

We denote the UIR obtained by $\mathbf{D}_{+}^{s,b}$. With respect to the su(1,1) subalgebra, it is decomposed into the direct sum $D_{+}^{s}\oplus D_{+}^{s+1/2}\oplus D_{+}^{s+1/2}\oplus D_{+}^{s+1}$ of the unitary representations of discrete series, and the components $f_0, f_{1,2}, f_3$ of the state (4.1) transform by the representations of higher weights s, $s+1/2$, and $s+1$, respectively.

The representations being obtained for $s\leqslant 0$ or $|b|>1$ are nonunitary. The case of $s+1=-j$, j is non-negative integer or half integer, is special. Then the operators (4.8) generate a finite-dimensional representation $\mathbf{D}^{j}$ of dimension $8j+8$. It is a superquartet of finite-dimensional representations of su(1,1), $\mathbf{D}^{j}=D^{j+1}\oplus D^{j+1/2}\oplus D^{j+1/2}\oplus D^{j}$, and the state's components $f_0, f_{1,2}, f_3$ transform by the $2j+3$-, $2j+2$-, and $2j+1$-dimensional representations, respectively.

It should be mentioned that the representations of the su(1,1|2) being considered here correspond to an *irregular* coadjoint orbit $\mathcal{L}^{1|2}$ of the supergroup SU(1,1|2) and, hence, they are *atypical* representations. By reduction to the orthosymplectic subsuperalgebra we get just the *typical* representations of the osp(2|2).

Keeping in mind the spinning superparticle, we remember that the representations D_{+}^{s} of the universal covering $\overline{\mathrm{SO}^{\uparrow}(1,2)}$ are commonly used for conventional realizations of the UIRs of the $D=3$ Poincaré symmetry of fractional spin,[6,7,14] whereas the finite-dimensional irreps D^{j} serve the ones of integer or of half-integer spin. It will be natural to extend these realizations to the $N=2$ Poincaré supersymmetry by means of representations $\mathbf{D}_{+}^{s,b}$ and $\mathbf{D}^{j}$ of the inner su(1,1|2) superalgebra.

4. The correspondence principle

To complete the quantization procedure on $\mathcal{L}^{1|2}$ let us return to the relation between observables and linear operators. We have examined for the supersymmetry generators of the su(1,1|2) superalgebra that there is an exact correspondence between supercommutators of the operators in $\mathcal{O}_{s,b}$ and the Poisson superbrackets of respective classical observables. In this sense we do have ''the quantization'' of a classical mechanics on the $N=2$ superunit disc. Consider now the correspondence between the algebras of *arbitrary* linear operators in the Hilbert space and their symbols.

The problem we are concerned with is thoroughly studied for Kähler homogeneous manifolds. Berezin proved the general ''correspondence principle,''[30,31] which roughly consists of the following. The multiplication of operators induces a binary $*$-operation for corresponding symbols; $*$-multiplication is noncommutative. Furthermore, the theory contains a ''Planck constant'' h related to one of the quantum numbers, and in the limit when $h \to 0$ $*$-algebra transforms to the ordinary commutative algebra of functions on the manifold. Finally, the first-order reset with respect to h of the commutator of symbols coincides with their Poisson bracket. The Lobachevsky plane has originally served as a test example for the Berezin technique.[31] The parameter s^{-1} plays the role of the Planck constant.

Similar principles hold true for the $N=2$ superunit disc being a natural $N=2$ superextension of the Lobachevsky plane.

Let $\hat{A}_1$ and $\hat{A}_2$ be two linear operators in $\mathcal{O}_{s,b}$ and $A_1(\underline{\Gamma},\overline{\Gamma})$ and $A_2(\underline{\Gamma},\overline{\Gamma})$ be the respective Berezin covariant symbols. It follows from Eq. (4.5) that the symbol being corresponded to the product $\hat{A}_2 \cdot \hat{A}_1$ (and denoted by $A_2 * A_1$) reads

$$A_2 * A_1(\underline{\Gamma},\overline{\Gamma}) = \int_{\mathcal{L}_{1|2}} A_2(\underline{\Gamma}_1,\overline{\Gamma}) A_1(\underline{\Gamma},\overline{\Gamma}_1) \frac{L_{s,b}(\underline{\Gamma},\overline{\Gamma}_1) L_{s,b}(\underline{\Gamma}_1,\overline{\Gamma})}{L_{s,b}(\underline{\Gamma}_1,\overline{\Gamma}_1) L_{s,b}(\underline{\Gamma},\overline{\Gamma})} d\mu(\underline{\Gamma}_1,\overline{\Gamma}_1). \tag{4.9}$$

Hence the multiplication of the operators induces the $*$-multiplications of the symbols.

Theorem (the correspondence principle): The following estimations take place:

$$(1) \quad \lim_{s\to\infty} A_2 * A_1(\underline{\Gamma},\overline{\Gamma}) = A_2(\underline{\Gamma},\overline{\Gamma}) \cdot A_1(\underline{\Gamma},\overline{\Gamma})$$

$$(2) \quad \lim_{s\to\infty} s\big(A_2 * A_1(\underline{\Gamma},\overline{\Gamma}) - A_1 * A_2(\underline{\Gamma},\overline{\Gamma})\big) = is\{A_2,A_1\},$$

where $\{,\}$ is the Poisson superbracket on $\mathcal{L}^{1|2}$.

Proof: It is based on the asymptotic estimation

$$A_2 * A_1(\underline{\Gamma},\overline{\Gamma}) = A_2(\underline{\Gamma},\overline{\Gamma}) \cdot A_1(\underline{\Gamma},\overline{\Gamma}) + i A_2(\underline{\Gamma},\overline{\Gamma}) \frac{\overleftarrow{\partial}}{\partial \underline{\Gamma}^A} \Omega^{A\overline{B}} \frac{\overrightarrow{\partial}}{\partial \overline{\underline{\Gamma}^B}} A_1(\underline{\Gamma},\overline{\Gamma}) + \mathcal{O}(s^{-2}), \tag{4.10}$$

from which both propositions of the theorem are easily obtained. Here $\Omega^{A\overline{B}} = \{\underline{\Gamma}^A, \overline{\underline{\Gamma}^B}\}$ are fundamental Poisson superbrachets on $\mathcal{L}^{1|2}$, whose explicit form is written down in the extended version of this paper.[37] The validity of the relation (4.10) is sufficient to prove when $z=0$. If it is the case, Eqs. (4.10) hold true at any z in consequence of the SU(1,1) invariance of the symplectic structure. Taking this fact into account, the verification of Eq. (4.10) is made by means of an ordinary expansion of the symbols in (finite) series in the odd variables and the comparison of the lhs and rhs of Eqs. (4.10) for the respective components. It is a trivial but cumbersome exercise, which may be successfully performed using the known estimation[31]

$$\hat{T}_l[\varphi] \equiv \frac{2l-1}{2\pi i} \int_{|z|<1} \varphi(z,\overline{z})(1-z\overline{z})^{2l-2}\, dz d\overline{z} = \varphi(0,0) + \frac{1}{2l} \Delta \varphi(z,\overline{z}) \Big|_{z=\overline{z}=0} + \mathcal{O}(l^{-2}) \tag{4.11}$$

and its consequence

$$l(\hat{T}_1[\varphi] - \hat{T}_{l+1/2}[\varphi]) = \frac{1}{4l} \Delta \varphi(z,\overline{z})|_{z=\overline{z}=0} + \mathcal{O}(l^{-2}).$$

Here $l>\frac{1}{2}$, $\varphi(z,\overline{z})$ is an arbitrary function to be continuously differentiable into the unit disc in a complex plane, and $\Delta=(1-z\overline{z})^2\partial\overline{\partial}$ is an invariant Laplace–Beltrami operator in $\mathcal{L}$. It is exactly the estimation (4.11) that was originally applied by Berezin for the proof of the correspondence principle in the Lobachevsky plane.[31] In this sense, we reduce the correspondence principle in $\mathcal{L}^{1|2}$ to the one in $\mathcal{L}$ by means of the expansion in the odd variables. ■

B. Operator realization of the Poincaré superalgebra. Renormalization of the supercharges

Now we are in a position to proceed directly to the quantization of the $D=3$ spinning superparticle. Consider the space $\mathcal{H}$ of functions of the form

$$F(p,\overline{\Gamma})\equiv F(p,\overline{z},\overline{\theta},\overline{\chi})=F_0(p,\overline{z})+\sqrt{s(1+b)}\,\overline{\theta}F_1(p,\overline{z})+\sqrt{s(1-b)}\,\overline{\chi}F_2(p,\overline{z})$$
$$+\sqrt{s(s+1/2)(1-b^2)}\,\overline{\theta}\overline{\chi}F_3(p,\overline{z}),\qquad(4.12)$$

where $p\equiv p^a\in R^{1,2}$, and $F_p(\overline{\Gamma})\equiv F(p,\overline{\Gamma})\in\mathcal{O}_{s,b}$ at each fixed p. We would like to suppose that the Hamiltonians (3.24) [which are the same as in relation (3.13)] present "the classical symbols" of respective operators of the $N=2$ Poincaré superalgebra acting in $\mathcal{H}$. We take the following ansatz for these operators:

$$\hat{\mathcal{J}}_a=-i\epsilon_{abc}p^b\frac{\partial}{\partial p_c}+\overline{J}_a,\quad \overline{\mathcal{P}}_a=p_a,\quad \overline{\mathcal{Z}}=mb,$$

$$\hat{\mathcal{Q}}^1_\alpha=(ip_{\alpha\beta}\hat{W}^\beta+m\hat{\hat{W}}_\alpha)[1+q(b\hat{P}_3-\sqrt{1-b^2}\hat{P}_2-\hat{P}_4)],\qquad(4.13)$$

$$\hat{\mathcal{Q}}^2_\alpha=(ip_{\alpha\beta}\hat{V}^\beta+m\hat{\hat{V}}_\alpha)[1+q(b\hat{P}_3+\sqrt{1-b^2}\hat{P}_2-\hat{P}_4)].$$

Here the operators $\hat{W}^\alpha$, $\hat{\hat{W}}^\alpha$, $\hat{V}^\alpha$, and $\hat{\hat{V}}^\alpha$ are expressed as linear combinations of $\hat{E}^\alpha$, $\hat{F}^\alpha$, $\hat{G}^\alpha$, and $\hat{H}^\alpha$ according to relations (3.26), whereas the latter, together with the operators $\hat{J}_a$ and $\hat{P}_I$, are defined by the expressions (4.8).

Recall that the classical observables (3.24) or (3.13) generate the Poincaré superalgebra on shell only, that is, modulo to the constraints (3.15). The operator counterparts of the constraints are now imposed to annihilate the physical states according to Dirac quantization prescriptions. The linear operator in $\mathcal{O}_{s,b}$, which corresponds to the Berezin covariant symbol $-sn_a(s>0)$, reads

$$\widehat{-sn}_a=\hat{J}_a\left(1-\frac{2}{2s+1}\hat{P}_4+\frac{2}{(2s+1)(2s+2)}\overline{\theta}\frac{\vec{\partial}}{\partial\overline{\theta}}\overline{\chi}\frac{\vec{\partial}}{\partial\overline{\chi}}\right)=\hat{J}_a\left(1+\frac{1}{s}\hat{P}_4\right)^{-1}.$$

Thus, the wave equations for the superparticle are easily brought to the form

$$(p^2+m^2)F^{\mathrm{phys}}(p,\overline{z},\overline{\theta},\overline{\chi})=0,\quad[(p,\hat{J})-m\hat{P}_4-ms]F^{\mathrm{phys}}(p,\overline{z},\overline{\theta},\overline{\chi})=0.\qquad(4.14)$$

[It is worth noting that the second constraint equation (3.15) should be written in an equivalent form $(p,J)-mP_4-ms=0$.] Solutions of the wave equations generate a subspace $\mathcal{H}_{m,s,b}$ in $\mathcal{H}$. Furthermore, if $F\in\mathcal{H}_{m,s,b}$ and $\hat{\mathcal{S}}$ is any one of the operators (4.13), then $\hat{\mathcal{S}}F$ is a physical state again, regardless of the particular value of the parameter q. In this sense, the wave equations are super-Poincaré invariant.

It is crucial now to examine explicitly whether the operators (4.13) actually generate the $N=2$ Poincaré superalgebra. One gets in $\mathcal{H}$ [compare with Eq. (3.13)]

$$[\hat{\mathcal{J}}_a,\hat{\mathcal{J}}_b]_-=i\epsilon_{abc}\hat{\mathcal{J}}^c,\quad[\hat{\mathcal{J}}_a,\hat{\mathcal{P}}_b]_-=i\epsilon_{abc}\hat{\mathcal{P}}^c,\quad[\hat{\mathcal{P}}_a,\hat{\mathcal{P}}_b]_-=0,\qquad(4.15)$$

$$[\hat{\mathcal{J}}_a,\hat{\mathcal{Q}}^I_\alpha]_- = \frac{1}{2}(\gamma_a)_\alpha{}^\beta\hat{\mathcal{Q}}^I_\beta,\quad [\hat{\mathcal{P}}_a,\hat{\mathcal{Q}}^I_\alpha]_- = 0,\quad [\hat{\mathcal{Q}}^I_\alpha,\hat{P}_3]_- = -\frac{i}{2}\epsilon^{IJ}\hat{\mathcal{Q}}^J_\alpha .$$

These relations hold true with an arbitrary number taken for q. However, the anticommutator of the supercharges is strongly dependent on the particular value of q:

$$[\hat{\mathcal{Q}}^I_\alpha,\hat{\mathcal{Q}}^J_\beta]_+ = 2\delta^{IJ}p_{\alpha\beta} - 2imb\epsilon^{IJ}\epsilon_{\alpha\beta} + c^{1IJ}_{\alpha\beta}(p^2+m^2) + c^{2IJ}_{\alpha\beta}((p,\hat{J}) - m\hat{P}_4 - ms) + \mathcal{O}(s^{-2}), \tag{4.16}$$

where $c^{\cdot IJ}_{\alpha\beta}$ are some functions, and $\mathcal{O}(s^{-2})$ are the corrections of higher orders in s^{-1}, which depend on q and the other parameters like m and b. These corrections do not vanish if $q=q^{cl}=1/4s$.

Notice that the quantum value of q is not uniquely determined. The value $q^{cl}=1/4s$ is derived from the expressions (3.24), when the relationship between the super-Poincaré and su(1,1|2) generators (3.20) is taken into account. However, one can start immediately from the symbols (3.13) and restore the operators applying the correspondence rule (4.5). What is remarkable is that one obtains the same operators (4.13), but the parameter q changes, and q^{cl} appears to be $q^{cl}_1=1/(4s+2)$. But the Poincaré superalgebra is disclosed by $q=1/(4s+2)$; the same is true for $q=1/4s$ too.

Both the appearance of corrections $\mathcal{O}(s^{-2})$ on the rhs of Eq. (4.16) and the ambiguity in the definition of q have the same origin, that is, a *nonlinearity* of the Poincaré supercharge operators (4.13) in the generators (4.8) of the inner su(1,1|2) superalgebra. In consequence of the nonlinearity, different operator factor orderings may lead to the different forms for $\hat{\mathcal{Q}}^I_\alpha$, and the corrections appear in response to the correspondence principle in $\mathcal{L}^{1|2}$.

We show that the disclosure of the Poincaré superalgebra at the quantum level has transparent mathematical ground in view of the Berezin correspondence principle. However, this disclosure is quite unsatisfactory from the physical viewpoint for the quantization of the elementary system. The latter is completely characterized by its inherent symmetries (in the present case it is the $D=3$ Poincaré SUSY). *It is the representation of these symmetries in Hilbert space that allows us to identify the obtained quantum theory with the quantized elementary system*. According to these reasons, to quantize the $D=3$ superparticle we now have to provide an *exact* realization of the representation of the Poincaré superalgebra in the physical Hilbert space, without any corrections in the parameters of the model. To find the true quantum realization for the representation, we can try, starting from Eqs. (4.13)–(4.16), to introduce some renormalized terms in the observables (4.13), which should be sufficient for the closure of the anticommutators (4.16).

Certainly, we do not have an *a priori* reason, which may ensure the consistency of the renormalization procedure; a structure of the possible higher-order corrections to (4.13) is unclear in general. Surprisingly, the exact corrections may be obtained from the simplest ansatz (4.13) for the quantum observables. In other words, a true ordering exists for the su(1,1|2) superalgebra operators, entering in $\hat{\mathcal{Q}}^I_\alpha$ in Eq. (4.13), that *allows us to restore a representation of the Poincaré superalgebra by the renormalization of the only parameter q*.

It is examined by a direct calculation that the corrections $\mathcal{O}(s^{-2})$ on the rhs of Eq. (4.16) vanish and the operators (4.13) generate the closed Poincaré superalgebra if, and only if,

$$q = q^{\text{quant}}_\mp = 1 \mp \sqrt{1-\frac{1}{2s+1}}. \tag{4.17}$$

Some details of calculations of the anticommutators (4.16) are given in the Appendix of Ref. 37. The renormalized value $q^{\text{quant}}_- = q^{cl} + \mathcal{O}(s^{-2})$ can be treated as a perturbative correction to the classical symbols of the supercharges. The other possible value q^{quant}_+ emerges from the hidden $N=4$ supersymmetry and could be understood from the following reasons.

Let $q=q_-$. The operators of supercharges corresponding to the classical observables (3.25) [see also Eqs. (3.22)] and providing the hidden $N=4$ supersymmetry in $\mathcal{H}_{m,s,b}$ are presented by

$$\hat{\bar{\mathcal{Q}}}^I_\alpha = -\frac{i}{m} p_\alpha{}^\beta \hat{\mathcal{Q}}^I_\beta = i\hat{K}\cdot\hat{\mathcal{Q}}^I_\alpha\Big|_{q\to 2-q}, \tag{4.18}$$

where the *parity operator* $\hat{K}$ is introduced. It acts on the components of the wave function (4.12) by the rule

$$\hat{K}: F=(F_0,F_1,F_2,F_3)\to\hat{K}F=(F_0,-F_1,-F_2,F_3), \quad \hat{K}^2=\hat{I}, \tag{4.19}$$

and $\hat{\mathcal{Q}}^I_\alpha|_{q\to 2-q}$ denote the supercharges (4.13) being considered when the constant q is substituted for the expression $(2-q)$.

The same critical values (4.17) evidently provide the closure of the $N=4$ Poincaré superalgebra. Moreover, the parity operator (4.19) possesses remarkable features: it commutes with the even generators of the $N=4$ Poincaré superalgebra and anticommutes with the supercharges:

$$[\hat{\mathcal{J}}_a,\hat{K}]_-=[\hat{\mathcal{P}}_a,\hat{K}]_-=0, \quad [\hat{P}_I,\hat{K}]_-=0, \quad I=1,2,3,4,$$
$$[\hat{\mathcal{Q}}^I_\alpha,\hat{K}]_+=[\hat{\bar{\mathcal{Q}}}^I_\alpha,\hat{K}]_+=0, \quad I=1,2. \tag{4.20}$$

Therefore, the operators $\hat{\mathcal{Q}}'^I_\alpha=-i\hat{K}\hat{\mathcal{Q}}^I_\alpha=\hat{\bar{\mathcal{Q}}}^I_\alpha|_{q\to 2-q}$ and $\hat{\bar{\mathcal{Q}}}'^I_\alpha=-i\hat{K}\hat{\bar{\mathcal{Q}}}^I_\alpha=\hat{\mathcal{Q}}^I_\alpha|_{q\to 2-q}$ satisfy the (anti)commutation relations being identical with Eqs. (4.15) and (4.16) for the supercharges $\hat{\mathcal{Q}}^I_\alpha$ and $\hat{\bar{\mathcal{Q}}}^I_\alpha$ themselves. This observation clarifies to some extent the origin of the nonperturbative value q_+^{quant} for the parameter q. Notice that two representations of the Poincaré superalgebra corresponding to either possible value q are equivalent to each other. It is seen straightforwardly from relations

$$\hat{\mathcal{Q}}'^I_\alpha=\hat{U}\hat{\bar{\mathcal{Q}}}^I_\alpha\hat{U}, \quad \hat{\bar{\mathcal{Q}}}'^I_\alpha=\hat{U}\hat{\mathcal{Q}}^I_\alpha\hat{U}, \quad [\hat{\mathcal{J}}_a,\hat{U}]_-=[\hat{\mathcal{P}}_a,\hat{U}]_-=[\hat{P}_I,\hat{U}]_-=0,$$

where the operator $\hat{U}$ reads

$$\hat{U}=1-2\bar{\theta}\frac{\vec{\partial}}{\partial\bar{\theta}}\bar{\chi}\frac{\vec{\partial}}{\partial\bar{\chi}}, \quad \hat{U}^2=\hat{I}.$$

We do not observe, however, either any classical counterpart for the supercharges $\hat{\mathcal{Q}}'^I_\alpha, \hat{\bar{\mathcal{Q}}}'^I_\alpha$ or any algebraic construction (for instance, superalgebra) involving both sets of the $N=4$ supercharges $\hat{\mathcal{Q}}^I_\alpha, \hat{\bar{\mathcal{Q}}}^I_\alpha$ and $\hat{\mathcal{Q}}'^I_\alpha, \hat{\bar{\mathcal{Q}}}'^I_\alpha$ on equal footing.

To summarize briefly, the ''double'' SUSY of the classical mechanics of the $D=3$ spinning superparticle can be lifted to the operator representation in the quantum theory. The key step of the construction is the renormalization (4.17) for the Poincaré supercharges (4.13). Equation (4.17) displays two exceptional values of the parameter q providing the closure for the anticommutator of supercharges (4.16) and recovering the consistent representation of the Poincaré superalgebra. In accordance with the analysis of Sec. IV A 3 we conclude that the off-shell wave function (4.12) carries the representation of the superalgebra su(1, 1|2) which is unitary infinite dimensional when $s>0$, and nonunitary finite dimensional when $s+1=j,j$ being non-negative integer or half-integer. The first possibility corresponds to the $N=2$ superanyon, whereas in the second case we obtain conventional realization of the representations of a habitual half-integer superspin. One can easily establish that the theory describes the massive quartet of $N=2$, $D=3$ superparticles when $|b|<1$, and this quartet contracts to the $N=1$ superdoublet[25] in the BPS limit $|b|=1$.

The last remarkable step is that the representations of the Poincaré superalgebra obtained are *unitary*. For example, in the case of fractional superspin $(s>0)$, the physical space $\mathcal{H}_{m,s,b}$ is naturally endowed by an inner product

$$(F|G)=\mathcal{N}\int \frac{d\mathbf{p}}{p^0}\langle F|G\rangle_{\mathcal{L}^{1|2}}, \quad p^0=\sqrt{\mathbf{p}^2+m^2}<0, \tag{4.21}$$

where $\langle F|G\rangle_{\mathcal{L}^{1|2}}$ denotes the inner product (4.2) in $\mathcal{O}_{s,b}$, $p^a=(p^0,\mathbf{p})$, and $\mathcal{N}$ is an arbitrary normalization constant. The operators (4.13) of the $N=2$ Poincaré superalgebra are Hermitian with respect to the inner product. However, the integrand in Eq. (4.2) [and, thus, (4.21)] has the inherent singularity at $|z|=1$ in the case of $s<0$, and it is ill defined in the BPS limit $|b|=1$. The correct realization of the inner product for the case of half-integer superspin and in the BPS limit has been found in Refs. 37 and 25, respectively.

V. SUMMARY AND DISCUSSION

In the present paper we have constructed the consistent first quantized theory of an $N=2$, $D=3$ superanyon as well as the one of massive superparticle of the habitual (half) integer superspin. The starting point for the quantization is the classical model of the superparticle in the nonlinear phase superspace $\mathcal{M}^{8|4}=T^*(R^{1,2})\times\mathcal{L}^{1|2}$, which is different from the standard approach.

A traditional viewpoint in the construction of the spinning particle models[34] is to describe the spinning degrees of freedom by some variables being simultaneously translation invariant and Lorentz covariant (as usual, those are Lorentz vectors or spinors). Such variables parametrize some linear space L and then the extended phase space is chosen to be $\mathcal{M}=T^*(R^{1,D-1})\times L$ or $\mathcal{M}=T^*(R^{1,D-1}\times L)$. The only difference for superparticles is to replace D-dimensional Minkowski space $R^{1,D-1}$ by the respective superspace. The advantage of the covariant (super) space $\mathcal{M}$ is in the linear (''covariant'') action of the Poincaré supergroup. In this approach, however, an embedding of the (super)particle physical space $\mathcal{O}$ (that is, the underlying coadjoint orbit) in the covariant phase (super)space may be ambiguous. Moreover, it is a common usage in this approach to give little attention to the geometry underlying the embedding $\mathcal{O}\rightarrow\mathcal{M}$.

We have demonstrated that the nonlinear phase superspace $\mathcal{M}^{8|4}=T^*(R^{1,2})\times\mathcal{L}^{1|2}$ of the $D=3$ spinning superparticle has the following remarkable features:

(i) The embedding of an appropriate coadjoint $\mathcal{O}_{m,s,b}$ orbit, being associated to the $N=2$, $D=3$ superparticle of arbitrary fixed mass $m>0$, superspin $s\neq 0$, and central charge mb ($|b|<1$ in $\mathcal{M}^{8|4}$), is realized by two constraints, which provide the identical conservation of any Casimir function of the Hamiltonian Poincaré superalgebra. These constraints have transparent geometric origin and, after quantization, they are converted into wave equations of the superparticle in a natural way.

(ii) The ''inner'' subsupermanifold $\mathcal{L}^{1|2}$ of $\mathcal{M}^{8|4}$ appears to be in itself the coadjoint orbit for some supergroups. $\mathcal{L}^{1|2}$ is shown to be symplectomorphic to the Kähler homogeneous superspace of the supergroup SU(1, 1|2) or its subsupergroup OSp(2|2). In this sense the model admits the second supersymmetry [SU(1, 1|2) SUSY] along with the original Poincaré one.

To describe the superparticle in a standard way, it is convenient, starting from an ordinary particle living in $R^{1,D-1}$, to extend the geometry of the Minkowski space to the supergeometry of the respective Minkowski superspace. We have found an alternative way, at least for dimension $D=3$. The intrinsic structure of $D=3$ spinning particle may be described in terms of the Lobachevsky geometry. To introduce the supersymmetry we may extend the inner manifold, going to the Lobachevsky supergeometry.

The following interpretation is admissible. The $D=3$ particle lives in an ordinary Minkowski space $R^{1,2}$. In addition the *super*spin degrees of freedom are associated to its internal structure and generate the internal phase superspace $\mathcal{L}^{1|2}$ with an inherent SU(1, 1|2) supersymmetry, which is different from the Poincaré (super)symmetry.

(iii) We suggest nontrivial quantization of the superparticle in the extended phase superspace $\mathcal{M}^{8|4}$, which combines the canonical quantization in $T^*(R^{1,2})$ and the Berezin quantization in the inner phase superspace $\mathcal{L}^{1|2}$. This quantization scheme leads naturally to the fields carrying infinite-dimensional or finite-dimensional representation of the supergroup SU(1, 1|2) depending on the fractional or habitual (half)integer value of spin. The results are completely consistent with the previous known description of $D=3$ nonsupersymmetric particles as mechanical systems.[6,13,14]

Surprisingly, there are two, unitary equivalent to one another, series of $N=2$ supercharges in quantum theory, which correspond to different possibilities for the parameter q in Eq. (4.17). Only one of them, namely q_-, is related directly to a conventional classical limit. Another possible value q_+ is shown to be related to the special properties (4.20) of the parity operator (4.19). However, the classical counterpart of this parity structure remains unclear, and the origin of the second possible value for q may seem enigmatic. Notice that the parity operator generates the structure of the deformed Heisenberg algebra in Hilbert space of anyon[26] or $N=1$ superanyon.[25] It would be interesting to understand what is a geometry behind the parity operator for the $N=2$ superanyon.

The significance of this one-particle theory may vary, in particular, depending on the possibility of an efficient second quantization of the model. One of the problems here is to construct a Lagrangian of the theory, which leads to the one-particle wave equation we have deduced from the classical mechanical action. The first step of construction may be to present two independent wave equations of superanyon [like Eqs. (4.14)] in the form of one spinor equation, when the mass and spin shell fixing conditions may emerge as integrability conditions. It is known that the similar construction for anyons gives a simple action functional,[8,26] which may be relevant for the second quantization of fractional spin particles. An adequate superextension (at least for $N=1$) may be constructed probably using the representations of the su(1, 1|2) superalgebra in the same way, as the spinor set of the anyon wave equations was constructed in Ref. 26 using the atypical UIRs of osp(2|2). In this connection it should be noted that the exploitation of the atypical UIRs of the osp(2|2) superalgebra and of the deformed Heisenberg algebra produces the linear set of spinor wave equations of the $N=1$, $D=3$ superparticle only for special (half) integer $j=\frac{1}{2}$ and $j=1$ values of the superspin.[26]

And, of course, the consistent interaction of (super)anyons remains an intriguing problem. Even in the first quantized theory the suggested approach to the description of anyon, being attempted for the extension to an interaction with an external field, implies (in the framework of minimal phase space) a perturbative representation for nonlinear commutation relations in terms of a series in powers of the field strengths.[18] In particular, it is unclear whether any consistent generalization exists for the wave equations of (super)anyons obtained in this paper in the presence of arbitrary external fields.

ACKNOWLEDGMENTS

SLL is thankful to Professor S. Randjbar-Daemi for the hospitality at the High Energy Section, ICTP, where the manuscript of the paper has been completed.

This work is partially supported by Grants of the International Soros Science Education Program and by the RBRF Grant No. 98-02-16261. SLL was supported in part by the Joint INTAS-RBRF Grant No. 95-829. IVG was supported partially by the INTAS Grant No. 96-0457 within the research program of the International Center for Fundamental Physics in Moscow.

[1] R. B. Laughlin, Phys. Rev. Lett. **50**, 1395 (1983).
[2] F. Wilczek, in *Fractional Statistics and Anyon Superconductivity*, edited by F. Wilczek (World Scientific, Singapore, 1990).
[3] J. M. Leinaas and J. Myrheim, Nuovo Cimento B **37**, 1 (1977).
[4] F. Wilczek, Phys. Rev. Lett. **48**, 1144 (1982); **49**, 957 (1982).
[5] G. A. Goldin, R. Menikoff, and D. H. Sharp, J. Math. Phys. **21**, 650 (1980); **22**, 1664 (1981).
[6] R. Jackiw and V. P. Nair, Phys. Rev. D **43**, 1933 (1991).
[7] S. Forte and T. Jolicoeur, Nucl. Phys. B **350**, 589 (1991).
[8] D. P. Sorokin and D. V. Volkov, Nucl. Phys. B **409**, 547 (1993).

[9] S. Deser, R. Jackiw, and S. Templeton, Ann. Phys. (N.Y.) **140**, 372 (1982).
[10] G. W. Semenoff, Phys. Rev. Lett. **61**, 517 (1988); J. Fröhlich and P. A. Marchetti, Lett. Math. Phys. **16**, 347 (1988); Commun. Math. Phys. **121**, 177 (1988); Nucl. Phys. B **356**, 533 (1991), and references therein.
[11] Z. Hlousek and D. Spector, Nucl. Phys. B **344**, 763 (1990).
[12] A. Balachandran, G. Marmo, B. Skagerstam, and A. Stern, *Gauge Symmetries and Fibre Bundles* (Springer, Berlin, 1983).
[13] M. S. Plyushchay, Phys. Lett. B **248**, 107 (1990); Int. J. Mod. Phys. A **7**, 7045 (1992).
[14] J. L. Cortés and M. S. Plyushchay, Mod. Phys. Lett. A **10**, 409 (1995); Int. J. Mod. Phys. A **11**, 3331 (1996).
[15] S. Ghosh, Phys. Rev. D **51**, 5827 (1995).
[16] I. V. Gorbunov, S. M. Kuzenko, and S. L. Lyakhovich, Int. J. Mod. Phys. A **12**, 4199 (1997).
[17] A. Nersessian, Mod. Phys. Lett. A **12**, 1783 (1997).
[18] C. Chou, V. P. Nair, and A. P. Polychronakos, Phys. Lett. B **304**, 105 (1993).
[19] C. Chou, Phys. Lett. B **323**, 147 (1994).
[20] S. Ghosh and S. Mukhopadhyay, Phys. Rev. D **51**, 6843 (1995).
[21] N. Banerjee and S. Ghosh, Phys. Rev. D **52**, 6130 (1995).
[22] S. Ghosh, J. Phys. A **30**, L821 (1997); Phys. Rev. D **57**, 6342 (1998).
[23] S. L. Lyakhovich, A. Yu. Segal, and A. A. Sharapov, Phys. Rev. D **54**, 5223 (1996).
[24] S. L. Lyakhovich, A. A. Sharapov, and K. M. Shekhter, Nucl. Phys. B **537**, 640 (1999).
[25] I. V. Gorbunov, S. M. Kuzenko, and S. L. Lyakhovich, Phys. Rev. D **56**, 3744 (1997).
[26] M. S. Plyushchay, Phys. Lett. B **320**, 91 (1994); Ann. Phys. (N.Y.) **245**, 339 (1996); Mod. Phys. Lett. A **11**, 2953 (1996); **12**, 1153 (1997).
[27] J. M. Souriau, *Structure des Systémes Dynamique* (Dunod, Paris, 1970).
[28] B. Kostant, *Lecture Notes in Mathematics, Vol. 170* (Springer-Verlag, New York, 1970), p. 87.
[29] A. A. Kirillov, *Elements of the Representation Theory* (Mir, Moscow, 1974).
[30] F. A. Berezin, Math. USSR-Izvestija **38**, 1116 (1974); **39**, 363 (1975).
[31] F. A. Berezin, Commun. Math. Phys. **40**, 153 (1975).
[32] N. M. J. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1980).
[33] R. Marnelius and U. Martensson, Nucl. Phys. B **335**, 395 (1991); Int. J. Mod. Phys. A **6**, 807 (1991).
[34] A. Frydryszak, in *From Field Theory to Quantum Mechanics* edited by B. Jancewicz and J. Sobczyk (World Scientific, Singapore, 1996).
[35] S. M. Kuzenko, S. L. Lyakhovich, and A. Yu. Segal, Int. J. Mod. Phys. A **10**, 1529 (1995); Phys. Lett. B **348**, 421 (1995).
[36] I. V. Gorbunov and S. L. Lyakhovich, Phys. Lett. B **422**, 293 (1998).
[37] I. V. Gorbunov and S. L. Lyakhovich, ''Hidden supersymmetries and Berezin quantization of $N=2$, $D=3$ spinning superparticles,'' preprint ICTP IC/98/138, hep-th/9809104.
[38] A. B. Balantekin, H. A. Schmitt, and P. Halse, J. Math. Phys. **30**, 274 (1989).
[39] E. A. M. Gradechi and L. M. Nieto, Commun. Math. Phys. **175**, 521 (1996).
[40] J. A. de Azcárraga and J. Lukierski, Phys. Lett. B **113**, 170 (1982); Phys. Rev. D **28**, 1337 (1983).
[41] P. Fayet, Nucl. Phys. B **51**, 135 (1976); M. F. Sohnius, *ibid.* **138**, 109 (1978); Phys. Rep. **128**, 39 (1985).
[42] L. Frappat, A. Sciarrino, and P. Sorba, ''Dictionary on Lie Superalgebras,'' preprint ENSLAPP-AL-600/96, DSF-T-30/96, hep-th/9607161.
[43] L. Brink and M. B. Green, Phys. Lett. B **106**, 393 (1981).
[44] Z. Hlousek and D. Spector, Nucl. Phys. B **370**, 143 (1992); **397**, 173 (1993).
[45] M. A. Perelomov, *Generalized Coherent States* (Nauka, Moscow, 1987);
[46] D. Bar-Moshe and M. S. Marinov, in *Topics in Statistical and Theoretical Physics*, edited by R. L. Dobrushin *et al.*; J. Phys. A **27**, 6287 (1994).

An inversion inequality for potentials in quantum mechanics

Richard L. Hall[a)]
Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Boulevard West, Montréal, Québec H3G 1M8, Canada

(Received 19 November 1998; accepted for publication 2 February 1999)

We suppose: (1) that the ground-state eigenvalue $E=F(v)$ of the Schrödinger Hamiltonian $H=-\Delta+vf(x)$ in one dimension is known for all values of the coupling $v>0$; and (2) that the potential shape can be expressed in the form $f(x)=g(x^2)$, where g is monotone increasing and convex. The inversion inequality $f(x)\leqslant \bar{f}(1/4x^2)$ is established, in which the ''kinetic potential'' $\bar{f}(s)$ is related to the energy function $F(v)$ by the transformation $\{\bar{f}(s)=F'(v),\ s=F(v)-vF'(v)\}$. As an example, f is approximately reconstructed from the energy function F for the potential $f(x)=ax^2+b/(c+x^2)$.
[S0022-2488(99)01705-3]

I. INTRODUCTION

We suppose that a discrete eigenvalue $E=F(v)$ of the Schrödinger Hamiltonian,

$$H=-\Delta+vf(x), \tag{1.1}$$

is known for all sufficiently large values of the coupling parameter $v>v_c$, and we try to use this data to reconstruct the potential shape f. Here v_c is the critical value of the coupling parameter for the eigenvalue considered. For the present discussion we shall assume that the potential shape $f(x)$ is nonconstant, symmetric, and monotone increasing for $x>0$. The usual ''forward'' problem would be the following: given the potential (shape) f, find the energy trajectory F; ''geometric spectral inversion'' is the inverse of this, that is to say, $F\rightarrow f$.

This problem should be distinguished from the ''inverse problem in the coupling constant'' discussed, for example, by Chadan and Sabatier.[1] In this latter problem, the discrete part of the ''input data'' is a set $\{v_n\}$ of values of the coupling constant that all yield the identical energy eigenvalue E. The index n might typically represent the number of nodes in the corresponding eigenfunction. In contrast, for the problem discussed in the present paper, n is kept fixed and the input data is the graph $(F_n(v),v)$, where the coupling parameter has any value $v>v_c(n)$, and $v_c(n)$ is the critical value of v for the support of a discrete eigenvalue with n nodes; for the ground state in one dimension we have, in general, $v_c(0)=0$. For the excited states, the critical coupling depends on the potential shape; for example, in the case of the square well, $v_c(n)=(n\pi/2)^2$; for the sech-squared potential with shape $f(x)=-\mathrm{sech}^2(x)$, we have[2] $v_c(n)=n(n+1)$, $n=0,1,2,\ldots$. There are strong indications on the basis of studies involving the inversion of the WKB approximation[3] that inversion from a single fixed energy trajectory F_n becomes more efficient as n increases (and the problem becomes more classical). However, the present paper will be concerned only with inversion from the ground-state energy function $F_0(v)=F(v)$ for the problem in one spatial dimension.

By making suitable assumptions concerning the class of potential shapes, theoretical progress has already been made with this inversion problem.[4–7] In Ref. 5 a ''concentration lemma'' is proved. If we suppose that $H\psi=E\psi$ and $\|\psi\|=1$, this lemma quantifies the monotone increase in

[a)]Electronic mail: rhall@cicma.concordia.ca

concentration toward $x=0$ of the probability density $\psi^2(x,v)$ with increasing v. In Ref. 6 this lemma is used to establish the uniqueness of the potential shape f corresponding to a given energy function F. The class of potentials for which this uniqueness proof applies are those nonconstant potential shapes f that are symmetric, continuous at $x=0$, piecewise analytic,[8] and monotone increasing for $x>0$. The ''envelope inversion'' discussed in Ref. 6 involved a class of potentials that could be expressed as a smooth monotone transformation $f(x)=g(h(x))$ of a soluble potential $h(x)$. The approximation obtained was *ad hoc*, in the sense that nothing was known *a priori* concerning the relationship between the approximation and the (unknown) exact potential corresponding to the given energy function $F(v)$. In the present paper we establish an inversion *inequality* for a special case of envelope inversion, namely the case in which the ''envelope basis'' is the harmonic-oscillator shape $h(x)=x^2$. Thus, we assume that the potential shape $f(x)$ has the representation

$$f(x)=g(x^2), \tag{1.2}$$

where g is monotone increasing and convex $(g''>0)$. This is a strong assumption, but, as we prove in Sec. II, it yields a corresponding strong result, that is to say,

$$f(x)\leqslant \bar{f}\left(\frac{1}{4x^2}\right), \tag{1.3}$$

where $\bar{f}(s)$ is the ''kinetic potential'' corresponding to the potential $f(x)$. The parameter s is equal to the mean kinetic energy $\langle -\Delta\rangle$ and, in terms of s, the eigenvalue $F(v)$ may be represented[9] *exactly* by the semiclassical expression

$$E=F(v)=\min_{s>0}\{s+v\bar{f}(s)\}. \tag{1.4}$$

The transformations $F\leftrightarrow\bar{f}$ are essentially Legendre transformations.[10] This is so because we know[5] that F and $\bar{f}$ have definite and opposite convexity; more particularly, we know

$$\bar{f}''(s)F''(v)=-\frac{1}{v^3}<0. \tag{1.5}$$

The transformation in the direction needed here $F\rightarrow\bar{f}$ will be given explicitly in Sec. II, below where we also prove the inequality (1.3), the main result of this paper. In Section (III) we discuss an example for which we compare the upper approximation given by (1.3) with the corresponding exact result.

II. PROOF OF THE INVERSION INEQUALITY

We suppose that the exact normalized wave function corresponding to the potential $vf(x)$ is given by $\psi(x,v)$, where the coupling parameter $v>0$. Thus, $(\psi,H\psi)=F(v)$. We know how this total expectation value is divided between kinetic and potential energies for, in more detail, we have

$$\begin{aligned}\langle -\Delta\rangle=(\psi,-\Delta\psi)=F(v)-vF'(v)=s,\\ \langle f\rangle=(\psi,f\psi)=F'(v)=\bar{f}(s).\end{aligned} \tag{2.1}$$

These equations also define the kinetic potential $\bar{f}(s)$ parametrically in terms of the parameter $v>0$. We first use Heisenberg's uncertainty inequality, which gives us

$$\langle -\Delta\rangle\langle x^2\rangle=s\langle x^2\rangle\geqslant\tfrac{1}{4}. \tag{2.2}$$

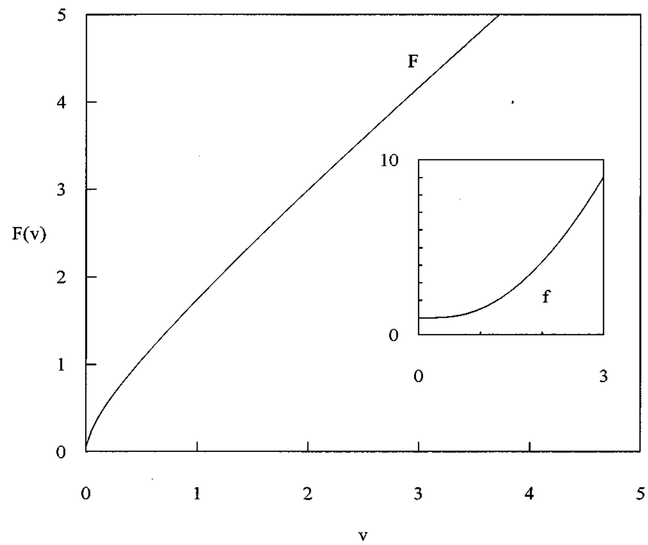

FIG. 1. The potential $f(x)=x^2+1/(1+x^2)$ is shown in the inset graph, along with the corresponding ground-state energy function $E=F(v)$. The aim of geometric spectral inversion is to reconstruct f from F.

We now consider

$$\overline{f}(s)=\langle f(x)\rangle=\langle g(x^2)\rangle\geqslant g(\langle x^2\rangle). \tag{2.3}$$

This inequality follows from Jensen's inequality[11] and the fact that g is convex. By applying (2.2) in (2.3), and using the monotony of g, we find

$$\overline{f}(s)\geqslant g\left(\frac{1}{4s}\right)=f\left(\frac{1}{2\sqrt{s}}\right). \tag{2.4}$$

Finally, by letting $x=1/2\sqrt{s}$ we establish the inversion inequality,

$$f(x)\leqslant\overline{f}\left(\frac{1}{4x^2}\right). \tag{2.5}$$

□

Since the transformation in the direction $F\rightarrow\overline{f}$ is already expressed by (2.1), the upper approximation provided by the inversion inequality is now completely determined.

III. AN EXAMPLE

We consider the potential shape given by

$$f(x)=ax^2+b/(c+x^2), \quad a,b,c>0. \tag{3.1}$$

The case $a=b=c=1$ is illustrated in Fig. 1, which shows the potential shape $f(x)$, in the inset graph, and also the ground-state energy function $F(v)$ generated from it. In Fig. 2 the upper approximation A obtained by the inversion inequality is shown along with the exact potential shape f itself. The set of corresponding ''exact'' wave functions $\psi(x,v)$ are also shown for $3\times10^{-4}\leqslant v\leqslant 10$. The wave function normalization is arbitrarily taken here to be $\psi(0,v)=20$, so that the graphs fit on the same figure as the potentials. As the coupling v increases, the wave functions become monotonically more concentrated near zero, in agreement with the ''concentration lemma'' mentioned in Sec. I.

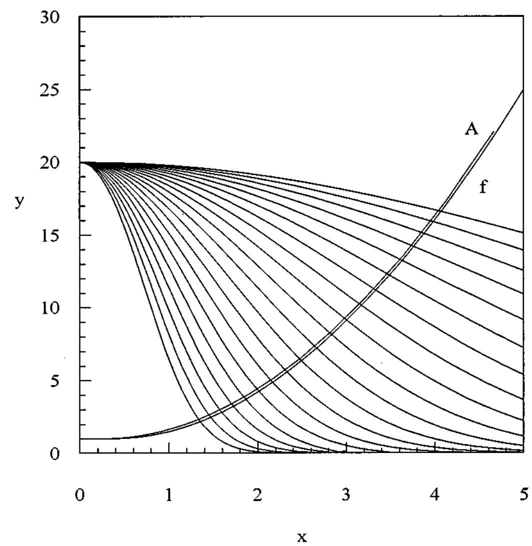

FIG. 2. The approximation *A* obtained from the inversion inequality (2.5) is compared to the exact potential *f*. The family of corresponding ''exact'' wave functions $\psi(x,v)$ satisfying $\psi(0,v)=20$ is also shown: the wave functions become monotonically concentrated toward zero as v is increased from $v=3\times10^{-4}$ to 10.

IV. CONCLUSION

Although the assumption behind the inversion inequality is strong, the fact that such an inequality exists may be important, especially if it can eventually be generalized. The expression of this result in terms of kinetic potentials could be avoided, in principle. However, the representation of the energy functions $F(v)$ in terms of $\bar{f}(s)$ has already yielded some very effective bounds in the forward direction, and it is natural to explore this same apparatus for the more difficult inversion problem. For example, in the forward direction the envelope method[9] may be expressed succinctly as

$$f(x)=g(h(x))\Rightarrow\bar{f}(s)\approx g(\bar{h}(s)), \tag{4.1}$$

where $\approx=\geqslant$ if g is convex and $\approx=\leqslant$ if g is concave. Once one has such an approximation for $\bar{f}(s)$, it can immediately be inserted in the expression (1.4) to yield an approximation for the corresponding eigenvalue $E=F(v)$. In the present paper we have found one case $h(x)=x^2$ for which an inequality is retained for the inverse problem. In Ref. 6 we also explored the idea of inverting the Rayleigh–Ritz variational method, and we obtained an inversion approximation with respect to a chosen family of ''trial'' functions. However, unlike the situation in the forward direction, the inversion approximation obtained was again *not* an inequality. Our experience with this problem so far suggests that it is difficult to generate potential inequalities for geometric spectral inversion.

ACKNOWLEDGMENTS

Partial financial support of this work under Grant No. GP3438 from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged. The author is indebted to Y. V. Kurylev for discussions concerning the uniqueness proof of Ref. 6.

[1] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer-Verlag, New York, 1989); the ''inverse problem in the coupling constant'' is discussed on p. 406.
[2] S. Flügge, *Practical Quantum Mechanics* (Springer-Verlag, New York, 1974); the sech-squared potential is discussed on p. 94.
[3] R. L. Hall, Phys. Rev. A **51**, 1787 (1995).
[4] R. L. Hall, J. Phys. A **25**, 4459 (1992).

[5] R. L. Hall, Phys. Rev. A **50**, 2876 (1995).
[6] R. L. Hall, J. Phys. A **28**, 1771 (1995).
[7] R. L. Hall, J. Math. Phys. **40**, 699 (1999).
[8] We are indebted to Y. V. Kurylev for pointing out that the uniqueness proof given in Ref. 6 requires an additional condition on $f(x)$ to control rapid oscillations; it is sufficient to assume that $f(x)$ is continuous at $x=0$ and piecewise analytic.
[9] R. L. Hall, J. Math. Phys. **25**, 2708 (1984).
[10] I. M. Gelfand and S. V. Fomin, *Calculus of Variations* (Prentice–Hall, Englewood Cliffs, NJ, 1963); Legendre transformations are discussed on p. 72.
[11] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1971); Jensen's inequality is discussed on p. 153.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Weyl anomaly in higher dimensions and Feynman rules in coordinate space

Shoichi Ichinose[a)]
Department of Physics, Brookhaven National Laboratory, Upton, New York 11973

Noriaki Ikeda[b)]
Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-01, Japan

(Received 26 October 1998; accepted for publication 22 December 1998)

An algorithm to obtain the Weyl anomaly in higher dimensions is presented. It is based on the heat-kernel method. Feynman rules, such as the vertex rule and the propagator rule, are given in (regularized) coordinate space. A graphical calculation is introduced. The six-dimensional scalar-gravity theory is taken as an example, and its explicit result is obtained. © *1999 American Institute of Physics.* [S0022-2488(99)01805-8]

I. INTRODUCTION

The anomaly phenomenon is, as well as the renormalization, one of the important aspects of the quantum field theories. It is a local quantum effect originating from regularization of the continuous space–time. We may say it is the problem of how to define the space–time as the continuum. Generally, the symmetry, imposed at the classical level does not hold at the quantum level due to the anomaly. As for the chiral anomaly, which is the anomaly concerned with the γ_5 symmetry (chiral symmetry), the general form is well established in arbitrary even dimensions.[1] It is fixed, except for an overall coefficient, by its cohomology structure, and is given in a beautiful form in terms of the differential form: $\mathrm{tr}(R\wedge R\wedge\cdots\wedge R)$, $\mathrm{tr}(F\wedge F\wedge\cdots\wedge F)$. The topological nature due to the γ_5 projection is the origin of its simplicity. On the other hand, as for the Weyl anomaly, the general form in higher dimensions is not so simple. This is the anomaly concerned with local scale transformations. Cohomology analysis restricts the Weyl anomaly to some extent, but there generally remains numerous candidate terms in higher dimensions.[2–5] The undetermined coefficients of those terms must be fixed in some way. The situation originates from the ''rich'' structure of the functional space on which the scale transformation acts. In the development of the string theory or *M* theory, it becomes more and more necessary to investigate the Weyl anomaly in higher-dimensional field theories.[6] At this circumstance we present a new formalism for it.

The Weyl anomaly itself is not harmful to the construction of physical theories because we know that realistic theories have some dimensional parameters, such as the mass or the cosmological constant, and they break the Weyl symmetry. However, when we try to understand the origin of those massive parameters, the Weyl anomaly is so important because it triggers the symmetry breaking of the conformal invariant vacuum. In this sense the Weyl anomaly can be regarded as a ''softer'' anomaly than the chiral one. The latter threatens the consistency of the theory whereas the former does not. The nonrenormalization theorem is valid for the latter, whereas generally not for the former.

In supersymmetric theories, however, it is also known that both anomalies make a supermultiplet together with the super-current.[7] Under the supersymmetric treatment, both anomalies are intimately related. In this connection there has been a lot of work. Because our interest here is to establish a general algorithm to treat the Weyl anomaly of the higher-dimensional theories, we

[a)]On leave of absence from the Department of Physics, University of Shizuoka, Yada 52-1, Shizuoka 422, Japan. Electronic mail: ichinose@u-shizuoka-ken.ac.jp
[b)]Electronic mail: nori@kurims.kyoto-u.ac.jp

do not consider supersymmetry here. We suppose, using the present result, that we can investigate how the Weyl anomaly is constrained by the supersymmetry.

We analyze the Weyl anomaly in the heat-kernel regularization. The content is essentially the same as, or overlaps with, topics with other names such as the effective action analysis, the background field method, the Schwinger–De Witt technique, Seeley series, etc. They all treat some background functional that represents (one-loop) quantum effects. It has a long history since Schwinger,[8] DeWitt,[9] and Seeley.[10] Especially the first author, in 1951, treated quantum matter theories in external gauge fields, keeping the gauge invariance of the effective action. He introduced an additional ''coordinate'' into the ordinary space(–time) coordinates, called ''proper time'' or ''temperature'' in a natural way. It was exploited by Alvarez-Gaumé and Witten[1] in the modern problem of the chiral anomaly of the gravitational theories. Later it was generalized to the Weyl anomaly by a SUNY group.[11,12]

The effective action has been analyzed in various ways. Its divergent part was systematically and algorithmically analyzed by t'Hooft[13] in 1973 and with Veltman[14] in 1974 using dimensional regularization. As will be shown later, the Weyl anomaly is obtained from the ''heat kernel,'' which is the one-loop effective action, with the proper time as the ultraviolet regularization. The evaluation of the heat kernel has been well developed by many people.[10,15–21] Each of them has their own characteristic formalisms. Generally they start with expressing a differential operator by some general background quantities. Most of them put an emphasis on the covariance in the formulas in the intermediate stages. As originally done by tHooft,[13] the ''intermediate covariance'' is powerful as far as the lower-order calculations, such as the heat kernel on two and four dimensions, are concerned. However, at higher order (this means higher dimensions too), it hinders one from performing explicit calculations because many more ''intermediate invariants'' appear than those appearing in the final explicit answer. The important thing is not the formulas at intermediate stages, but the final explicit result. Aiming at the treatment of the Weyl anomaly in higher dimensions, we present a new algorithm that is based on a simple weak-field expansion. Of course, we assume the invariance of the general coordinate and gauge symmetries at the final result, because it is guaranteed by the background (effective action) formalism.[9,13,22]

The following are new points of the present approach.

(1) We take a formalism in the regularized coordinate space, that is, the space with the ordinary space(–time) coordinates and the proper time.[23,24] The Feynman rules in the coordinate space are given for the anomaly (effective action) calculation.[25] The ultraviolet regularization is done with the Schwinger's parameter (proper time). The regularization lets us properly evaluate both chiral and Weyl anomalies.[26,27] The Feynman rules help us to investigate the Weyl anomaly in higher-dimensional theories.

(2) The graphical representation is taken at all stages. The familiar procedures in the ordinary perturbative treatment, such as the propagator rule and the vertex rule, are all graphically represented in coordinate space. The expanded terms themselves are also graphically represented. Especially, we introduce a ''graphical calculation.'' This makes the higher-order calculation transparent and the Weyl anomaly calculation in the higher dimensions tractable.

(3) We have not introduced any covariant quantities in intermediate stages. It is based on the weak-field perturbation. Although losing manifest general covariance in intermediate stages, the expanded elements have the simplest symmetry (w.r.t. the suffix permutation) and are preferable for the computer algorithm.

(4) We focus only on a special type of term: $(\partial\partial h)^n$, in the n-dimensional space. These are the lowest order of the product of n Riemann tensors. We demonstrate that they can determine the Weyl anomaly up to ''trivial terms.''

In Sec. II we briefly give the present formalism of the Weyl anomaly. The anomaly is given by the trace of the heat kernel. In Sec. III we present the Feynman rules to compute the heat kernel *to any higher order*. They are naturally expressed in the *regularized* coordinate space, that is, the space of n-dim ordinary space coordinates plus the the proper time. As always in perturbative approaches, a graphical representation helps to systematically compute the heat kernel at the higher order. The Taylor expansion, with respect to the ''regularization'' parameter t, is explained in Sec. IV. The more the number of space dimensions becomes, the more expansion is necessary. We apply the general algorithm to the 6-dim scalar-gravity theory in Sec. V. We focus on four special graphs and introduce a new graphical calculation. The final result of the Weyl anomaly for the model is obtained in Sec. VI. We conclude in Sec. VII. Four appendices are provided to supplement the text. The propagator rules are given in Appendix A. Some supplementary calculation, relegated by Sec. V of the text, is done in Appendix B. Weak field expansion of the 6-dim scalar gravity is graphically given in Appendix C. The table of coefficients, when some general invariants are weak-field expanded, is given in Appendix D.

II. FORMALISM

Let us explain the present formulation of anomalies taking a simple example: n-dim Euclidean gravity-scalar coupled system (see Ref. 26),

$$\mathcal{L}[g_{\mu\nu},\phi]=\sqrt{g}\,\phi\left(-\tfrac{1}{2}\nabla^2+\tfrac{1}{2}qR\right)\phi\equiv\tfrac{1}{2}\tilde{\phi}\,\mathbf{D}\tilde{\phi},$$
$$q=-\frac{n-2}{4(n-1)},\quad \mathbf{D}_x\equiv\sqrt[4]{g}\left(-\nabla_x^2-\frac{n-2}{4(n-1)}R(x)\right)\frac{1}{\sqrt[4]{g}},\tag{1}$$

where $g_{\mu\nu}$ and ϕ are the metric field and the scalar field, respectively. We have introduced $\tilde{\phi}\equiv\sqrt[4]{g}\,\phi$ for the measure $\mathcal{D}\tilde{\phi}$ to be general coordinate (BRS) invariant.[28] This Lagrangian is invariant under the local Weyl transformation:

$$g^{\mu\nu}(x)'=e^{2\alpha(x)}g^{\mu\nu}(x),\quad \tilde{\phi}(x)'=e^{-\alpha(x)}\tilde{\phi}(x)\quad [\phi(x)'=e^{(n-2)\alpha(x)/2}\phi(x)],\tag{2}$$

where $\alpha(x)$ is the parameter of the local Weyl transformation.

Even if the Lagrangian is Weyl invariant, the quantum theory is generally not (Weyl anomaly). The effective action $\Gamma[g_{\mu\nu}]$, defined by

$$e^{-\Gamma[g_{\mu\nu}]}=\int\mathcal{D}\tilde{\phi}\exp\left\{-\int d^nx\,\mathcal{L}[g_{\mu\nu},\phi]\right\},\tag{3}$$

changes as

$$e^{-\Gamma[g'_{\mu\nu}]}=\int\mathcal{D}\tilde{\phi}'\exp\left\{-\int d^nx\,\mathcal{L}[g'_{\mu\nu},\phi']\right\}$$
$$=\int\mathcal{D}\tilde{\phi}(x)\det\frac{\partial\tilde{\phi}'(y)}{\partial\tilde{\phi}(x)}\exp\left\{-\int d^nx\,\mathcal{L}[g_{\mu\nu},\phi]\right\}.\tag{4}$$

The Weyl anomaly is given by the Jacobian,[29,30]

$$J\equiv\det\left[\frac{\partial\tilde{\phi}'(y)}{\partial\tilde{\phi}(x)}\right]=\exp(-\mathrm{Tr}\,\alpha(x)\delta^n(x-y)+O(\alpha^2)).\tag{5}$$

Taking the heat-kernel regularization, we obtain the expression for the Weyl anomaly as

$$J=\exp\Big(-\lim_{t\to+0}\mathrm{Tr}[\alpha(x)G(x,y;t)]+O(\alpha^2)\Big),$$
$$\text{Weyl anomaly}=\left.\frac{\delta\Gamma}{\delta\alpha(x)}\right|_{\alpha=0}=-\left.\frac{\delta\ln J}{\delta\alpha(x)}\right|_{\alpha=0}=-2g^{\mu\nu}\langle T_{\mu\nu}\rangle=\lim_{t\to+0}\mathrm{Tr}\,G(x,y;t),\tag{6}$$

where $G(x,y;t)$ is the heat kernel defined by

$$\left(\frac{\partial}{\partial t}+\mathbf{D}_x\right)G(x,y;t)=0,\quad G(x,y;t)\left(\frac{\overleftarrow{\partial}}{\partial t}+\overleftarrow{D}_y^{\dagger}\right)=0,$$
$$\lim_{t\to+0} G(x,y;t)=\delta^n(x-y), \tag{7}$$

where $\overleftarrow{D}_x$ means it operates on the left. The parameter t is regarded here as a regularization parameter and is called Schwinger's *proper time*.[8] The last equation expresses the regularization of the delta function $\delta^n(x-y)$.[31,32] $G(x,y;t)$ can be symbolically written as

$$G(x,y;t)\equiv\langle x|e^{-t\mathbf{D}}|y\rangle,\quad t>0. \tag{8}$$

We note here the physical dimension of the space coordinate x^μ and the proper time t are

$$[x^\mu]=L,\quad [t]=L^2, \tag{9}$$

where L is some length. For other (conformally invariant) theories, the Weyl anomaly is obtained by replacing the operator $\mathbf{D}$ above. See Ref. 26 for details.

III. FEYNMAN RULES IN COORDINATE SPACE

A. Heat kernel and Feynman graph

Let us solve Eq. (7) in weak-field perturbation theory. For a general theory with the derivative couplings up to the second order, the operator $\mathbf{D}_x^{ij}$ can be always divided into the field-independent (free) part and the field-dependent (perturbation) one:

$$\mathbf{D}_x^{ij}=-\delta_{\mu\nu}\delta^{ij}\,\partial_\mu\partial_\nu-\mathbf{V}^{ij}(x),$$
$$\mathbf{V}^{ij}(x)\equiv W_{\mu\nu}^{ij}(x)\partial_\mu\partial_\nu+N_\mu^{ij}(x)\partial_\mu+M^{ij}(x), \tag{10}$$
$$i,j=1,2,...,N,$$

where $W_{\mu\nu}^{ij}$, N_μ^{ij}, and M^{ij} are external fields (background coefficient fields) and the field suffixes (such as a fermion suffix and a vector suffix) i,j are introduced for general use. In the present example (1), the above quantities are explicitly written as $(N=1)$

$$\mathbf{V}(x)=\sqrt[4]{g}(\boldsymbol{\nabla}^\mu\boldsymbol{\nabla}_\mu-qR)\frac{1}{\sqrt[4]{g}}-\delta_{\mu\nu}\partial_\mu\partial_\nu,$$
$$W_{\mu\nu}=g^{\mu\nu}-\delta_{\mu\nu}=-h_{\mu\nu}+h_{\mu\lambda}h_{\lambda\nu}+O(h^3),$$
$$N_\lambda=-g^{\mu\nu}\Gamma^\lambda_{\mu\nu}-g^{\lambda\mu}\Gamma^\nu_{\mu\nu}=-\partial_\mu h_{\lambda\mu}+O(h^2), \tag{11}$$
$$M=-qR+\tfrac{1}{4}g^{\mu\nu}\{\Gamma^\lambda_{\mu\lambda}\Gamma^\sigma_{\nu\sigma}+2\Gamma^\lambda_{\mu\nu}\Gamma^\sigma_{\lambda\sigma}-2\partial_\nu\Gamma^\lambda_{\mu\lambda}\}=-q(\partial^2h-\partial_\alpha\partial_\beta h_{\alpha\beta})-\tfrac{1}{4}\partial^2h+O(h^2),$$

where $g_{\mu\nu}=\delta_{\mu\nu}+h_{\mu\nu}$, $h\equiv h_{\mu\mu}$. The graphical representation of $W_{\mu\nu}$, N_λ, and M above is given, up to h^3 order, for the $n=6$-dim case in Appendix C: Fig. 16 for $W_{\mu\nu}$ and N_λ, Figs. 17–21 for M. The usage of general coefficients $W_{\mu\nu}^{ij}$, N_μ^{ij}, and M^{ij}, instead of their concrete contents, makes it possible to perform the general treatment valid for many theories.[26,33,34] All equations below are valid for all theories insofar as they have no higher-derivative interactions. Let us solve the differential equation (7) perturbatively for the case of weak external fields $(W_{\mu\nu}^{ij},N_\mu^{ij},M^{ij})$. (In the present example, this corresponds to perturbation around *flat* space.) The differential equation (7) becomes

$$\left(\frac{\partial}{\partial t}-\partial^2\right)G^{ij}(x,y;t)=\mathbf{V}^{ik}(x)G^{kj}(x,y;t),$$
$$\partial^2\equiv\delta_{\mu\nu}\partial_\mu\partial_\nu=\sum_{\mu=1}^{n}\left(\frac{\partial}{\partial x^\mu}\right)^2. \tag{12}$$

In the following we suppress the field suffixes $i,j,\ldots,$ and take the matrix notation. This equation (12) is the n-dim heat equation with the small perturbation $\mathbf{V}$. We give two important quantities in order to obtain the solution. (This approach is popular in the perturbative quantum field theory under the name *propagator approach*.[35] In Ref. 35, the momentum representation is taken, which is to be compared with the coordinate one of the present approach. The Weyl anomaly in the string theory was analyzed in this approach by Alvarez.[36])

1. Heat equation

The heat equation,

$$\left(\frac{\partial}{\partial t}-\partial^2\right)G_0(x,y;t)=0,\quad t>0, \tag{13}$$

has the solution

$$G_0(x,y;t)=G_0(x-y;t)=\int\frac{d^nk}{(2\pi)^n}\exp\{-k^2t+ik^\mu(x-y)^\mu\}I_N=\frac{1}{(4\pi t)^{n/2}}c^{-(x-y)^2/4t}I_N,$$
$$k^2\equiv\sum_{\mu=1}^{n}(k^\mu)^2, \tag{14}$$

where I_N is the identity matrix of the size $N\times N$. G_0 satisfies the initial condition: $\lim_{t\to+0}G_0(x-y;t)=\delta^n(x-y)I_N$. We define

$$G_0(x,y;t)=0,\quad \text{for } t\leqslant 0. \tag{15}$$

2. Heat propagator

The heat equation with the delta-function source defines the heat propagator,

$$\left(\frac{\partial}{\partial t}-\partial^2\right)S(x,y;t-s)=\delta(t-s)\delta^n(x-y)I_N,$$
$$S(x,y;t)=S(x-y;t)=\int\frac{d^nk}{(2\pi)^n}\frac{dk^0}{2\pi}\frac{\exp\{-ik^0t+ik\cdot(x-y)\}}{-ik^0+k^2}I_N=\theta(t)G_0(x-y;t), \tag{16}$$
$$k^2\equiv\sum_{\mu=1}^{n}k^\mu k^\mu,\quad k\cdot x\equiv\sum_{\mu=1}^{n}k^\mu x^\mu.$$

$\theta(t)$ is the *step function* defined by $\theta(t)=1$ for $t>0$; $\theta(t)=0$ for $t<0$. $S(x-y;t)$ satisfies the initial condition $\lim_{t\to+0}S(x-y;t)=\delta^n(x-y)I_N$ and $S(x,y;t)=0$ for $t\leqslant 0$.

Now the formal solution of (12) with the initial condition (7) is given by

$$G(x,y;t)=G_0(x-y;t)+\int d^nz\int_{-\infty}^{\infty}ds\,S(x-z;t-s)\mathbf{V}(z)G(z,y;s). \tag{17}$$

$G(x,y;t)$ appears in both sides above. We can iteratively solve (17) as[8]

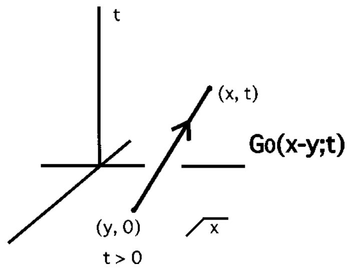

FIG. 1. Graph of $G_0(x-y;t)$.

$$
\begin{aligned}
G(x,y;t)&=G_0(x-y;t)+\int S\mathbf{V}G_0+\int S\mathbf{V}\int S\mathbf{V}G_0+\cdots,\\
G_1(x,y;t)&\equiv\int S\mathbf{V}G_0=\int d^nz\,dsS(x-z;t-s)\mathbf{V}(z)G_0(z-y;s)\\
&=\int d^nz\int_0^t ds\,G_0(x-z;t-s)\mathbf{V}(z)G_0(z-y;s),\\
G_2(x,y;t)&\equiv\int S\mathbf{V}\int S\mathbf{V}G_0=\int d^nz'\,ds'\,S(x-z';t-s')\mathbf{V}(z')\\
&\quad\times\int d^nz\,ds\,S(z'-z;s'-s)\mathbf{V}(z)G_0(z-y;s)\\
&=\int d^nz'\int_0^t ds'\,G_0(x-z';t-s')\mathbf{V}(z')\\
&\quad\times\int d^nz\int_0^{s'} ds\,G_0(z'-z;s'-s)\mathbf{V}(z)G_0(z-y;s),
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
G_3(x,y;t)&=\int d^nz''\int_0^t ds''\,G_0(x-z'';t-s'')\mathbf{V}(z'')\int d^nz'\int_0^{s''} ds'\,G_0(z''-z';s''-s')\mathbf{V}(z')\\
&\quad\times\int d^nz\int_0^{s'} ds\,G_0(z'-z;s'-s)\mathbf{V}(z)G_0(z-y;s).
\end{aligned}
$$

Expressions for higher-order terms are similarly obtained.

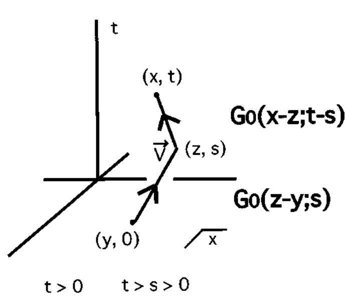

FIG. 2. Graph of $G_1(x,y;t)$.

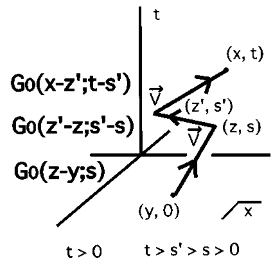

FIG. 3. Graph of $G_2(x,y;t)$.

As in the ordinary field theory, it is illuminating to represent this perturbative solution in graphs. We do it in the $(n+1)$-dimensional space, which is composed of the space of the ordinary n-dim (Euclidean) space plus 1-dim positive proper time. In the above solution, $G_0(x-y;t)$ plays the role of ''propagator.'' It is represented by a directed[37] straight line as in Fig. 1. $\mathbf{V}(z)$ plays the role of ''vertex,'' which acts on the system, during the process of ''propagation of particles,'' sometimes according to the perturbation order. The situation up to the third order is represented in Figs. 2–4. We can easily write down the expression of any higher order, $G_k(x,y;t)$, with the help of this graphical representation.

Generally, in n-dim, the terms up to $G_{n/2}$ are sufficient for the anomaly calculation.[26] In this sense, the present expansion looks like that with respect to the space(–time) dimension.

B. Factoring out the scale parameter *t*

Because of the presence of the positive regularization parameter t, we can safely (without singularity) take the trace part, $x=y$, in the above equations:

$$G(x,x;t)=G_0(0;t)+G_1(x,x;t)+G_2(x,x;t)+G_3(x,x;t)+\cdots,$$
$$G_0(0;t)=\frac{1}{(4\pi t)^{n/2}}I_N. \tag{19}$$

As shown in (6), the t^0 part of $G(x,x;t)$ is the Weyl anomaly. In order to see their t dependence, it is best to replace the dimensional coordinate variables such as (z,s) in (18) by the dimensionless ones (w,r),

$$G_1(x,x;t)\equiv\int S\mathbf{V}G_0\Big|_{x=y}=\frac{1}{t^{(n/2)-1}}\int d^nw\int_0^1 dr\, G_0(w;1-r)\mathbf{V}(x+\sqrt{t}w)G_0(w;r), \tag{20}$$

$$1>r=\frac{s}{t}>0,\quad w=(z-x)/\sqrt{t}, \tag{21}$$

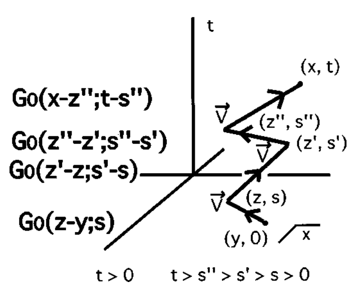

FIG. 4. Graph of $G_3(x,y;t)$.

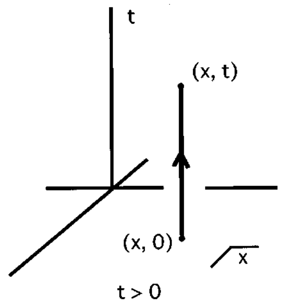

FIG. 5. Graph of $G_0(0;t)$.

where the relation (A2) is used. In the same way, we obtain

$$G_2(x,x;t) \equiv \int S\mathbf{V} \int S\mathbf{V} G_0 \bigg|_{x=y} = \frac{1}{t^{(n/2)-2}} \int d^n w' \int_0^1 dr' \, G_0(w';1-r')\mathbf{V}(x+\sqrt{t}w')$$

$$\times \int d^n w \int_0^{r'} dr \, G_0(w'-w;r'-r)\mathbf{V}(x+\sqrt{t}w) G_0(w;r), \tag{22}$$

where

$$1 > r' = s'/t > r = s/t > 0, \quad w' = (z'-x)/\sqrt{t}, \quad w = (z-x)/\sqrt{t},$$

$$G_3(x,x;t) \equiv \int S\mathbf{V} \int S\mathbf{V} \int S\mathbf{V} G_0 \bigg|_{x=y} = \frac{1}{t^{(n/2)-3}} \int d^n w'' \int_0^1 dr'' \, G_0(w'';1-r'')\mathbf{V}(x+\sqrt{t}w'')$$

$$\times \int d^n w' \int_0^{r''} dr' \, G_0(w''-w';r''-r')\mathbf{V}(x+\sqrt{t}w')$$

$$\times \int d^n w \int_0^{r'} dr \, G_0(w'-w;r'-r)\mathbf{V}(x+\sqrt{t}w) G_0(w;r), \tag{23}$$

where

$$1 > r'' = s''/t > r' = s'/t > r = s/t > 0,$$

$$w'' = (z''-x)/\sqrt{t}, \quad w' = (z'-x)/\sqrt{t}, \quad w = (z-x)/\sqrt{t}.$$

The above quantities of $G_k(x,x;t)$, $(k=0,...,3)$ are depicted in Figs. 5–8, with the dimensionless

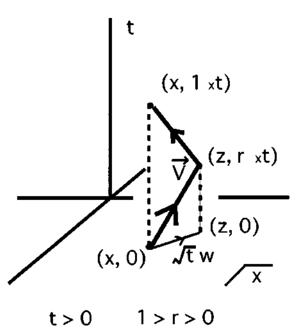

FIG. 6. Graph of $G_1(x,x;t)$.

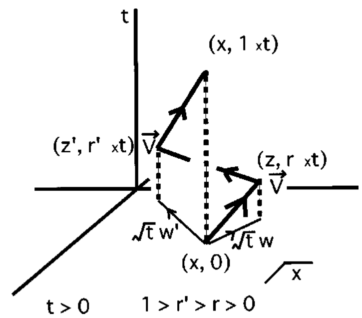

FIG. 7. Graph of $G_2(x,x;t)$.

quantities. We can easily write down higher-order terms with the help of graphs.

C. Vertex and propagator rules

The vertex operator $\mathbf{V}(x+\sqrt{t}w)$ in the expression $G_k(x,x;t)$ has differentials,

$$\mathbf{V}(x+\sqrt{t}w)=\frac{1}{t}W_{\mu\nu}(x+\sqrt{t}w)\frac{\partial}{\partial w^\mu}\frac{\partial}{\partial w^\nu}+\frac{1}{\sqrt{t}}N_\mu(x+\sqrt{t}w)\frac{\partial}{\partial w^\mu}+M(x+\sqrt{t}w), \tag{24}$$

and acts only on the ''adjacently right'' G_0 in Eqs. (20)–(23). Here we have the following *vertex rule*: Vertex Rule 1,

$$\begin{aligned}\mathbf{V}(x+\sqrt{t}w')G_0(w'-w;r'-r)&=\frac{1}{\{4\pi(r'-r)\}^{n/2}}\mathbf{V}(x+\sqrt{t}w')e^{-(w'-w)^2/4(r'-r)}\\&=\left\{\frac{1}{t}W_{\mu\nu}(x+\sqrt{t}w')\left(-\frac{\delta_{\mu\nu}}{2(r'-r)}+\frac{(w'-w)_\mu(w'-w)_\nu}{4(r'-r)^2}\right)\right.\\&\quad\left.+\frac{1}{\sqrt{t}}N_\mu(x+\sqrt{t}w')\left(-\frac{(w'-w)_\mu}{2(r'-r)}\right)+M(x+\sqrt{t}w')\right\}\\&\quad\times G_0(w'-w;r'-r)\\&=V(x+\sqrt{t}w';w'-w,r'-r;t)G_0(w'-w;r'-r),\end{aligned}$$

where

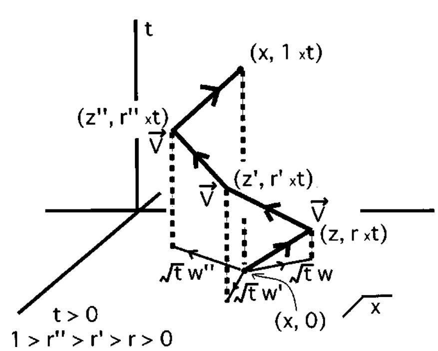

FIG. 8. Graph of $G_3(x,x;t)$.

$$V(x;w,r;t)\equiv\frac{1}{t}W_{\mu\nu}(x)\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\frac{1}{\sqrt{t}}N_\mu(x)\left(-\frac{w_\mu}{2r}\right)+M(x), \tag{25}$$

where V does *not* have differentials. Especially taking $r=0$, $w=0$ in the above and dropping primes ($'$), we obtain the following relation: Vertex Rule 2,

$$\begin{aligned}\mathbf{V}(x+\sqrt{t}w)G_0(w;r)&=\left\{\frac{1}{t}W_{\mu\nu}(x+\sqrt{t}w)\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)\right.\\&\quad\left.+\frac{1}{\sqrt{t}}N_\mu(x+\sqrt{t}w)\left(-\frac{w_\mu}{2r}\right)+M(x+\sqrt{t}w)\right\}G_0(w;r)\\&=V(x+\sqrt{t}w;w,r;t)G_0(w;r).\end{aligned}\tag{26}$$

Making use of the above rule, we can evaluate $G_1(x,x;t)$ of Eq. (20) as

$$\begin{aligned}G_1(x,x;t)&=\frac{1}{t^{(n/2)-1}}\int d^nw\int_0^1 dr\,G_0(w;1-r)G_0(w;r)V(x+\sqrt{t}w;w,r;t)\\&=\frac{1}{(4\pi)^{n/2}t^{(n/2)-1}}\int d^nw\int_0^1 dr\,G_0(w;(1-r)r)V(x+\sqrt{t}w;w,r;t),\end{aligned}\tag{27}$$

where the relation $G_0(w;1-k)G_0(w;r)=[1/(4\pi)^{n/2}]G_0(w;(1-r)r)$, $1>r>0$, which is derived from a *propagator rule* 4 (A8), is used in order to reduce the number of G_0's by one. In Appendix A, we list the propagator rules (A5)–(A8), and their graphical (''geometrical'') expressions in the coordinate space: Figs. 10–13. $G_2(x,x;t)$ of Eq. (22) is written, in terms of V, as

$$\begin{aligned}G_2(x,x;t)=&\frac{1}{t^{(n/2)-2}}\int d^nw'\int_0^1 dr'\,G_0(w';1-r')\int d^nw\int_0^{r'} dr\,V(x+\sqrt{t}w';w'-w,r'-r;t)\\&\times G_0(w'-w;r'-r)V(x+\sqrt{t}w;w,r;t)G_0(w;r).\end{aligned}\tag{28}$$

By changing the integration variable w' to a new one $\overline{w}$ defined below, we can make all G_0's have space variables independent of each other. Using the following relations:

$$G_0(w';1-r')G_0(w'-w;r'-r)=G_0\left(\overline{w};\frac{(1-r')(r'-r)}{1-r}\right)G_0(w;1-r),$$

$$1>r'>r,\quad \overline{w}=w'-\frac{1-r'}{1-r}w, \tag{29}$$

$$G_0(w;1-r)G_0(w;r)=\frac{1}{(4\pi)^{n/2}}G_0(w;(1-r)r),\quad 1>r>0,$$

which are obtained from Rule 2, (A6), and Rule 3, (A8), respectively, we can finally evaluate Eq. (28) as

$$\begin{aligned}G_2(x,x;t)=&\frac{1}{(4\pi)^{n/2}t^{(n/2)-2}}\int d^n\overline{w}\int_0^1 dr'\int d^nw\int_0^{r'} dr G_0\left(\overline{w};\frac{(1-r')(r'-r)}{1-r}\right)\\&\times G_0(w;(1-r)r)V(x+\sqrt{t}w';w'-w,r'-r;t)V(x+\sqrt{t}w;w,r;t),\\&w'=\overline{w}+\frac{1-r'}{1-r}w,\end{aligned}\tag{30}$$

where the integration variable w' is changed to $\overline{w}$ ($\det[\partial(w',w)/\partial(\overline{w},w)]=1$). We note the following things in the above evaluation.

(1) The number of propagators, G_0's, decreases by one and becomes the same as that of the space integration variables.
(2) All ''Gaussian'' coordinates in G_0's ($\overline{w}$ and w in the present case) can be taken to be independent.
(3) In the above reduction, we have applied the propagator rules on G_0's from ''left'' to ''right.'' There are some other choices (for example, from ''right'' to ''left'') which give expressions different from (30) in appearance.

These properties are valid for any order of $G_k(x,x;t)$.

Similarly, we can evaluate G_3 as

$$G_3(x,x;t)=\frac{1}{(4\pi)^{n/2}t^{(n/2)-3}}\int d^n\bar{\bar{w}}\int d^n\overline{w}\int d^nw\int_0^1 dr''\int_0^{r''}dr'\int_0^{r'}dr$$
$$\times G_0\left(\bar{\bar{w}};\frac{(1-r'')(r''-r')}{1-r'}\right)G_0\left(\overline{w};\frac{(1-r')(r'-r)}{1-r}\right)G_0(w;(1-r)r)$$
$$\times V(x+\sqrt{t}w'';w''-w',r''-r';t)V(x+\sqrt{t}w';w'-w,r'-r;t)V(x+\sqrt{t}w;w,r;t), \tag{31}$$
$$w''=\bar{\bar{w}}+\frac{1-r''}{1-r'}\overline{w}+\frac{1-r''}{1-r}w,\quad w'=\overline{w}+\frac{1-r'}{1-r}w,$$

where the integration variables w'' and w' are changed to $\bar{\bar{w}}$ and $\overline{w}$ ($\det[\partial(w'',w',w)/\partial(\bar{\bar{w}},\overline{w},w)]=1$). As in G_2, there are some different expressions depending on how we apply the propagator rules in G_0's.

Because of the integration formulas (A3) of Appendix A, the space integrations in (27), (30) and (31) give a product of Kronecker's deltas times some rational function of dimensionless parameters. From the above results we can write down $G_k(x,x;t)$ for any higher k. $G_k(x,x;t)$ has k-fold parameter integrals.

IV. EVALUATION OF t^0 PART OF $G(x,x;t)$—TAYLOR EXPANSION FOR TAKING THE LIMIT $t\to+0$—

Let us evaluate $G(x,x;t)|_{t^0}$. We consider $n=6$-dim space. We focus on the $(\partial\partial h)^3$-type terms, because that part has the sufficient information to determine the Weyl anomaly in 6 dim. Taking the Taylor expansion of $W_{\mu\nu}(x+\sqrt{t}v)$, $N_\mu(x+\sqrt{t}v)$, and $M(x+\sqrt{t}v)$ around $\sqrt{t}v=0$, $V(x+\sqrt{t}v;w,r;t)$ can be expanded as

$$V(x+\sqrt{t}v;w,r;t)=\frac{1}{t}W_{\mu\nu}(x+\sqrt{t}v)\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\frac{1}{\sqrt{t}}N_\mu(x+\sqrt{t}v)\left(-\frac{w_\mu}{2r}\right)+M(x+\sqrt{t}v)$$
$$=\frac{1}{t}V_{-1}(x,v;w,r)+V_0('')+tV_1('')+\cdots+\frac{1}{\sqrt{t}}V_{-1/2}('')+\sqrt{t}V_{1/2}('')+\cdots, \tag{32}$$

where $V_{-1/2}$, $V_{1/2},\ldots,$ are irrelevant parts because they all have odd-time derivatives of h's: ∂h, $\partial\partial\partial h,\ldots$. $V_{-1}('')=W_{\mu\nu}(x)(-\delta_{\mu\nu}/2r+w_\mu w_\nu/4r^2)$ is also irrelevant because it has no derivatives of h's.[38] V_0, V_1, V_2 are given by

$$V_0(x,v;w,r)=\frac{1}{2!}\partial_{\alpha_1}\partial_{\alpha_2}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\partial_{\alpha_1}N_\mu\cdot v^{\alpha_1}\left(-\frac{w_\mu}{2r}\right)+M,$$

$$V_1(x,v;w,r)=\frac{1}{4!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_4}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_4}\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\frac{1}{3!}\partial_{\alpha_1}\partial_{\alpha_2}\partial_{\alpha_3}N_\mu \cdot v^{\alpha_1}v^{\alpha_2}v^{\alpha_3}\left(-\frac{w_\mu}{2r}\right)+\frac{1}{2!}\partial_{\alpha_1}\partial_{\alpha_2}M\cdot v^{\alpha_1}v^{\alpha_2}, \tag{33}$$

$$V_2(x,v;w,r)=\frac{1}{6!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_6}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_6}\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\frac{1}{5!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_5}N_\mu \cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_5}\left(-\frac{w_\mu}{2r}\right)+\frac{1}{4!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_4}M\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_4}.$$

The Taylor expansion of $W_{\mu\nu}$, N_μ, M, for the case of the 6-dim scalar-gravity theory is graphically shown in Appendix C: (C3) for V_0, (C5) for V_1, and (C7) for V_2.

Now we pick up the t^0 part[39] and focus on the $(\partial\partial h)^3$ terms in the final form. For $G_1(x,x;t)$,

$$G_1(x,x;t)=\frac{1}{(4\pi)^3}\int_0^1 dr\int d^6w G_0(w;(1-r)r)\left[\frac{1}{t^2}V(x+\sqrt{t}w;w,r;t)\right],$$
$$\left.\left[\frac{1}{t^2}V('')\right]\right|_{t^0}=V_2(x,w;w,r). \tag{34}$$

We notice the $w^{\alpha_1}w^{\alpha_2}\cdots$ parts give, after the w integration, different products of Kronecker's deltas times powers of $(1-r)r$. See (A3).

As for $G_2(x,x;t)$, from (30),

$$G_2(x,x;t)=\frac{1}{(4\pi)^3}\int_0^1 dr'\int_0^{r'} dr\int d^6\overline{w}\int d^6w G_0\left(\overline{w};\frac{(1-r')(r'-r)}{1-r}\right)G_0(w;(1-r)r)\times\left[\frac{1}{t}V(x+\sqrt{t}w';w'-w,r'-r;t)V(x+\sqrt{t}w;w,r;t)\right],$$

$$\left.\left[\frac{1}{t}V('')V(''')\right]\right|_{t^0}=V_1(x,w';w'-w,r'-r)V_0(x,w;w,r)+V_0(x,w';w'-w,r'-r)V_1(x,w;w,r)+\text{irrel. terms},$$
$$w'=\overline{w}+R_1w,\quad w'-w=\overline{w}-S_1w, \tag{35}$$

$$R_1\equiv\frac{1-r'}{1-r}\equiv R(r,r'),\quad S_1\equiv\frac{r'-r}{1-r}\equiv S(r,r'),\quad R_1>0,\quad S_1>0,\quad R_1+S_1=1.$$

The functions $R(r,r')$ and $S(r,r')$ will be used in (37). The irrelevant terms (''irrel. terms'') are such ones as V_2V_{-1}, $V_{1/2}V_{1/2}$, etc.

As for $G_3(x,x;t)$, from (31),

$$G_3(x,x;t)=\frac{1}{(4\pi)^3}\int_0^1 dr''\int_0^{r''}dr'\int_0^{r'}dr\int d^6\bar{\bar{w}}\int d^6\overline{w}\int d^6w\,G_0\left(\bar{\bar{w}};\frac{(1-r'')(r''-r')}{1-r'}\right)$$
$$\times G_0\left(\overline{w};\frac{(1-r')(r'-r)}{1-r}\right)G_0(w;(1-r)r)[V(x+\sqrt{t}w'';w''-w',r''-r';t)$$
$$\times V(x+\sqrt{t}w';w'-w,r'-r;t)V(x+\sqrt{t}w;w,r;t)],$$
$$[V('')V(''')V('''')]|_{t^0}=V_0(x,w'';w''-w',r''-r')V_0(x,w';w'-w,r'-r)V_0(x,w;w,r)$$
$$+\text{irrel. terms},\tag{36}$$

where ''irrel. terms'' above are such ones as $V_1V_{-1}V_0$, $V_{1/2}V_{1/2}V_0$, etc. In the above, the coordinate variables in the integrand are written by the integration variables $\bar{\bar{w}}$, $\overline{w}$, and w as

$$w''=\bar{\bar{w}}+R_2\overline{w}+R_3w,\quad w''-w'=\bar{\bar{w}}-S_2\overline{w}-T_1w,$$
$$w'=\overline{w}+R_1w,\quad w'-w=\overline{w}-S_1w,$$

where

$$R_1=R(r,r')=\frac{1-r'}{1-r},\quad R_2=R(r',r'')=\frac{1-r''}{1-r'},\quad R_3=R(r,r'')=\frac{1-r''}{1-r},$$
$$S_1=S(r,r')=\frac{r'-r}{1-r},\quad S_2=S(r',r'')=\frac{r''-r'}{1-r'},\quad T_1=\frac{r''-r'}{1-r}\equiv T(r,r',r''),\tag{37}$$

where $R(r,r')$ and $S(r,r')$ was introduced in the previous order (35), and $T(r,r',r'')$ is a new function. R_i, S_i, and T_1 have the following relations:

$$R_1+S_1=1,\quad R_2+S_2=1,\quad R_3+T_1=R_1,\quad \frac{R_3}{R_2}=R_1,\quad \frac{T_1}{S_2}=R_1,$$
$$R_1,R_2,R_3,S_1,S_2,T_1>0.\tag{38}$$

Here we finish treating the Weyl anomaly $G(x,x;t)|_{t^0}$ in terms of the general ones: $(W_{\mu\nu},N_\lambda,M)$ and their derivatives. Their explicit forms depend on each model. The content of Appendix C comes from the 6-dim scalar-gravity theory (11). In the following two sections, we explain how we obtain the explicit form of the Weyl anomaly using the obtained formulas.

V. FOUR SPECIAL GRAPHS

In the 6-dim space, there are four ''important general invariants'' as the Weyl anomaly terms, which will be explained in the next section. In order to obtain the four coefficients of those terms, let us determine the coefficients of the following four $(\partial\partial h)^3$ terms that appear in $G(x,x;t)|_{t^0}$,

$$\text{Graph }3=\partial_\sigma\partial_\tau h_{\mu\nu}\cdot\partial_\nu\partial_\lambda h_{\tau\omega}\cdot\partial_\omega\partial_\mu h_{\lambda\sigma},$$
$$\text{Graph }67=\partial_\tau\partial_\omega h_{\mu\nu}\cdot\partial_\mu\partial_\nu h_{\lambda\sigma}\cdot\partial_\lambda\partial_\sigma h_{\tau\omega},$$
$$\text{Graph }1=\partial_\mu\partial_\nu h_{\nu\lambda}\cdot\partial_\lambda\partial_\sigma h_{\sigma\tau}\cdot\partial_\tau\partial_\omega h_{\omega\mu},$$
$$\text{Graph }2=\partial_\mu\partial_\nu h_{\tau\sigma}\cdot\partial_\sigma\partial_\lambda h_{\lambda\nu}\cdot\partial_\tau\partial_\omega h_{\omega\mu}.\tag{39}$$

Here we introduce a useful graphical representation to express SO(n) invariants like those given above. (See Ref. 40 for details.) We graphically express the basic ingredient $\partial_\mu\partial_\nu h_{\lambda\sigma}$ as

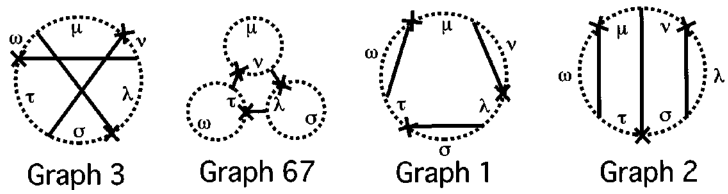

FIG. 9. Four Graphs of 3, 67, 1, and 2. See the (40) for the definition of the graph.

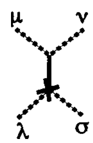 $\equiv \partial_\mu \partial_\nu h_{\lambda\sigma}.$ (40)

The contraction of suffixes is expressed by gluing the corresponding suffix lines. Then the above four terms (39) are graphically shown in Fig. 9. The advantage of this representation is that the way suffixes are contracted can be read by the topology of the graph. We need not bother about the dummy contracted suffixes. For later use, we introduce the following usage too:

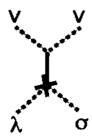 $\equiv \partial_\mu \partial_\nu h_{\lambda\sigma} \cdot v^\mu v^\nu,$ 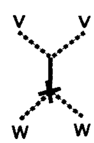 $\equiv \partial_\mu \partial_\nu h_{\lambda\sigma} \cdot v^\mu v^\nu w^\lambda w^\sigma.$ (41)

The choice of those graphs (39) is an important step of this algorithm. As will be explained in Sec. VI, it is done by looking at the table of Appendix D and the content of the ''important invariants.'' In the evaluation, we exploit the topology of graphs in order to efficiently select relevant terms. All four graphs above come only from $G_3(x,x;t)|_{t^0}$. It is graphically seen by the following common features:

- There is no ,
- No tadpoles (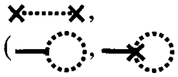),

the structure of G_1 and G_2:

$$G_1(x,x;t)|_{t^0} \sim \int dr \int d^6w\, G_0(w;(1-r)r) V_2(x,w;w,r),$$
$$G_2(x,x;t)|_{t^0} \sim \int dr' \int dr \int d^6\overline{w} \int d^6w\, G_0\left(\overline{w}; \frac{(1-r')(r'-r)}{1-r}\right) G_0(w;(1-r)r)$$
$$\{V_1(x,w';w'-w,r'-r)V_0(x,w;w,r) + V_0(x,w';w'-w,r'-r)V_1(x,w;w,r)\}, \quad (42)$$

where V_2, V_1, and V_0 are graphically shown in Appendix C 2. Both $G_1|_{t=0}$ and $G_2|_{t=0}$ contain, at least, one of two graph ingredients itemized above. For simplicity we focus, in this section, only on Graphs 3 and 67. Graph 1 and 2 are evaluated in Appendix B. Because of the following common features of Graph 3 and 67:

- No tadpoles,

- There is no  ,

we can reduce the expression of $G_3(x,x;t)|_{t^0}$, (36) with V_0 given by (C2) and (C3), to the following form as the relevant part:

$$G_3(x,x;t)|_{t^0}=\frac{1}{(4\pi)^3}\int_0^1 dr''\int_0^{r''}dr'\int_0^{r'}dr\int d^6\bar{\bar{w}}\int d^6\bar{w}\int d^6w$$

$$\times G_0\left(\bar{\bar{w}};\frac{(1-r'')(r''-r')}{1-r'}\right)G_0\left(\bar{w};\frac{(1-r')(r'-r)}{1-r}\right)G_0(w;(1-r)r)$$

$$\times\left(-\frac{1}{2}\right)^3\times\frac{1}{4(r''-r')^2}\times\frac{1}{4(r'-r)^2}\times\frac{1}{4r^2}\times(R_2)^2(S_2)^2$$

$\times$ 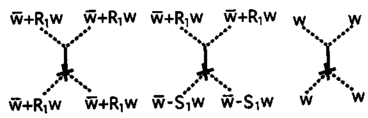

+irrelevant terms, (43)

where the relations $R_3/R_2=R_1$, $T_1/S_2=R_1$ are used [see (38)]. Furthermore, again by the common features of the Graphs 3 and 67, we see no contribution comes from $\overline{w}^8$ and $\overline{w}^6$ terms.[41]

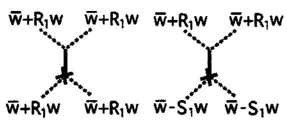

= 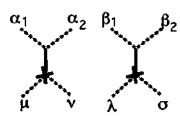

$$\times\{(R_1)^6(S_1)^2w^{\alpha_1}w^{\alpha_2}w^{\mu}w^{\nu}w^{\beta_1}w^{\beta_2}w^{\lambda}w^{\sigma}$$
$$+(R_1)^6w^{\alpha_1}w^{\alpha_2}w^{\mu}w^{\nu}w^{\beta_1}w^{\beta_2}\bar{w}^{\lambda}\bar{w}^{\sigma}$$
$$-(R_1)^5S_1(w^{\alpha_1}w^{\alpha_2}w^{\mu}w^{\nu}w^{\beta_1}\bar{w}^{\beta_2}+{}_6C_1)\bar{w}^{\lambda}w^{\sigma}\times 2$$
$$+(R_1)^4(S_1)^2(w^{\alpha_1}w^{\alpha_2}w^{\mu}w^{\nu}\bar{w}^{\beta_1}\bar{w}^{\beta_2}+{}_6C_2)w^{\lambda}w^{\sigma}$$
$$+(R_1)^2(S_1)^2(\bar{w}^{\alpha_1}\bar{w}^{\alpha_2}\bar{w}^{\mu}\bar{w}^{\nu}w^{\beta_1}w^{\beta_2}+{}_6C_4)w^{\lambda}w^{\sigma}$$
$$-(R_1)^3(S_1)(\bar{w}^{\alpha_1}\bar{w}^{\alpha_2}\bar{w}^{\mu}w^{\nu}w^{\beta_1}w^{\beta_2}+{}_6C_3)\bar{w}^{\lambda}w^{\sigma}\times 2$$
$$+(R_1)^4(\bar{w}^{\alpha_1}\bar{w}^{\alpha_2}w^{\mu}w^{\nu}w^{\beta_1}w^{\beta_2}+{}_6C_2)\bar{w}^{\lambda}\bar{w}^{\sigma}\}+O(\bar{w}^6)+O(\bar{w}^8),$$

(44)

where ${}_mC_r$ means ''other choices of $r\overline{w}$'s among m suffixes appearing in the first term within the brackets.'' The $\bar{\bar{w}}$, $\overline{w}$, w integrations are three independent Gaussian integrations, and they give the product of two parts for each term. One part is a rational function of the parameters and the other is a symmetrized product of the Kronecker's deltas. In terms of a ''components'' notation:

$$\langle c_1,c_2\rangle\equiv c_1(\text{Graph } 3)+c_2(\text{Graph } 67), \tag{45}$$

the final result is evaluated as

$$G_3(x,x;t)|_{t^0}=\frac{1}{(4\pi)^3}\int_0^1 dr''\int_0^{r''}dr'\int_0^{r'}dr\times\left(-\frac{1}{2}\right)^3\times\frac{1}{4(r''-r')^2}\times\frac{1}{4(r'-r)^2}\times\frac{1}{4r^2}$$
$$\times(R_2)^2(S_2)^2\times\Bigg[(R_1)^6(S_1)^2\{2(1-r)r\}^6\times\langle 64,16\rangle$$
$$+2\,\frac{(1-r')(r'-r)}{1-r}\{2(1-r)r\}^5\times\{(R_1)^6\langle 0,0\rangle-2(R_1)^5S_1\langle 32,8\rangle$$
$$+(R_1)^4(S_1)^2\langle 64,16\rangle\}+\left\{2\,\frac{(1-r')(r'-r)}{1-r}\right\}^2\{2(1-r)r\}^4\times\{(R_1)^2(S_1)^2\langle 0,8\rangle$$
$$-2(R_1)^3S_1\langle 32,0\rangle+(R_1)^4\langle 0,8\rangle\}\Bigg]+\text{other graph terms}$$
$$=\frac{1}{(4\pi)^3}\left\langle\frac{4}{27\times 105},-\frac{1}{36\times 63}\right\rangle+\text{other graph terms.}\tag{46}$$

The above numbers $\langle c_1,c_2\rangle$ are obtained by a computer calculation.[42] The numbers c_1 and c_2 show the ‘‘weights’’ when a symmetrized product of Kronecker’s deltas are multiplied to a $(\partial\partial h)^3$ tensor. For example the first one of (46), $\langle 64,16\rangle$, says

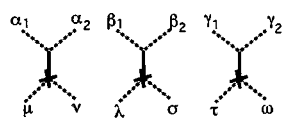

$$\times[\alpha_1\alpha_2\mu\nu\beta_1\beta_2\lambda\sigma\gamma_1\gamma_2\tau\omega]$$
$$=64(\text{Graph }3)+16(\text{Graph }67)+\text{other graph terms,}\tag{47}$$

where the notation $[\alpha_1\alpha_2\cdots]$ is the symmetrized product of Kronecker’s deltas and is defined in (A4) of Appendix A. Note that in the expression just before the parameter integral, poles at $r=1$, $r'=1$ cancel out and there remain no divergences of the parameter integrals. This occurs also in the calculation in Appendix B. Adding the result for Graph 1 and 2 explained in Appendix B, we obtain finally the total contribution of $G(x,x;t)|_{t^0}$ to the four graphs as

$$G(x,x;t)|_{t^0}=\frac{1}{(4\pi)^3}\left\{\frac{4}{27\times 105}(\text{Graph }3)-\frac{1}{36\times 63}(\text{Graph }67)+\frac{1}{36\times 630}(\text{Graph }1)\right.$$
$$\left.+\frac{1}{36\times 35}(\text{Graph }2)\right\}+\text{other terms.}\tag{48}$$

VI. FINAL INVARIANT FORM OF 6-DIM WEYL ANOMALY

The general structure of the Weyl anomaly has been analyzed by Refs. 43–45. They claim that the Weyl anomaly is composed of three types of terms: (a) the Euler term; (b) conformal invariants; (c) trivial terms. Trivial terms are those that can be absorbed by local counterterms made of the metric. This statement is checked by the cohomology analysis up to six dimensions.[2] We called, in Sec. V, the (a) and (b) terms ‘‘important invariants.’’ Therefore we may write, for the present 6-dim case,

$$G(x,x;t)|_{t^0}=\frac{1}{(4\pi)^3}\sqrt{g}\{xC_1+yC_2+zC_3+wE+\text{trivial terms}\}$$

$$C_1=C_{\mu\nu\lambda\sigma}C^{\sigma\lambda\alpha\beta}C_{\beta\alpha}{}^{\nu\mu},\quad C_2=C_{\mu\nu\alpha\sigma}C^{\nu\lambda\beta\alpha}C_\lambda{}^{\mu\sigma}{}_\beta,$$

$$C_3\sim C_{\mu\nu\alpha\beta}\nabla^2C^{\mu\nu\alpha\beta}+\cdots,\quad E\sim R_{\mu\nu\alpha\beta}R_{\lambda\sigma\gamma\delta}R_{\tau\omega\epsilon\theta}\epsilon^{\mu\nu\lambda\sigma\tau\omega}\epsilon^{\alpha\beta\gamma\delta\epsilon\theta}, \tag{49}$$

$$C^\lambda_{\mu\nu\sigma}=R^\lambda{}_{\mu\nu\sigma}+\frac{1}{4}(\delta^\lambda_\nu R_{\mu\sigma}+g_{\mu\sigma}R^\lambda_\nu-\nu\leftrightarrow\sigma)+\frac{1}{20}(\delta^\lambda_\sigma g_{\mu\nu}-\nu\leftrightarrow\sigma)R,$$

where $C^\lambda{}_{\mu\nu\sigma}$ is the Weyl tensor, C_1, C_2, and C_3 are three independent conformal invariants, and E is the Euler term. x,y,z, and w are some constants to be determined. It is well established that all possible independent invariants (parity even) with mass dimension 6 are, in the 6 and higher space dimensions, given by the following 17 terms:[2,46,47]

$$P_1=RRR,\quad P_2=RR_{\mu\nu}R^{\mu\nu},\quad P_3=RR_{\mu\nu\lambda\sigma}R^{\mu\nu\lambda\sigma},$$

$$P_4=R_{\mu\nu}R^{\nu\lambda}R_\lambda{}^\mu,\quad P_5=-R_{\mu\nu\lambda\sigma}R^{\mu\lambda}R^{\nu\sigma},\quad P_6=R_{\mu\nu\lambda\sigma}R_\tau{}^{\nu\lambda\sigma}R^{\mu\tau},$$

$$A_1=R_{\mu\nu\lambda\sigma}R^{\sigma\lambda}{}_{\tau\omega}R^{\omega\tau\nu\mu},\quad B_1=R_{\mu\nu\tau\sigma}R^\nu{}_{\lambda\omega}{}^\tau R^{\lambda\mu\sigma\omega},$$

$$O_1=\nabla^\mu R\cdot\nabla_\mu R,\quad O_2=\nabla^\mu R_{\lambda\sigma}\cdot\nabla_\mu R^{\lambda\sigma},$$

$$O_3=\nabla^\mu R^{\lambda\rho\sigma\tau}\cdot\nabla_\mu R_{\lambda\rho\sigma\tau},\quad O_4=\nabla^\mu R_{\lambda\nu}\cdot\nabla^\nu R^\lambda{}_\mu, \tag{50}$$

$$T_1=\nabla^2R\cdot R,\quad T_2=\nabla^2R_{\lambda\sigma}\cdot R^{\lambda\sigma},\quad T_3=\nabla^2R_{\lambda\rho\sigma\tau}\cdot R^{\lambda\rho\sigma\tau},$$

$$T_4=\nabla^\mu\nabla^\nu R\cdot R_{\mu\nu},$$

$$S=\nabla^2\nabla^2R.$$

In terms of the above 17 ''basis,'' we can rewrite (49) as follows:

$$C_1=\tfrac{9}{200}P_1-\tfrac{27}{40}P_2+\tfrac{3}{10}P_3+\tfrac{5}{4}P_4+\tfrac{3}{2}P_5-3P_6+A_1,$$

$$C_2=-\tfrac{19}{800}P_1+\tfrac{57}{160}P_2-\tfrac{3}{40}P_3-\tfrac{7}{16}P_4-\tfrac{9}{8}P_5+\tfrac{3}{4}P_6+B_1,$$

$$C_3=P_1-8P_2-2P_3+10P_4+10P_5-\tfrac{1}{2}T_1+5T_2-5T_3, \tag{51}$$

$$E=P_1-12P_2+3P_3+16P_4+24P_5-24P_6+4A_1-8B_1,$$

$$G(x,x,t)|_{t^0}=\frac{1}{(4\pi)^3}\sqrt{g}\left\{(x+4w)A_1+(y-8w)B_1+\left(\frac{5}{4}x-\frac{7}{16}y+10z+16w\right)P_4\right.$$
$$\left.+\left(\frac{3}{2}x-\frac{9}{8}y+10z+24w\right)P_5+\text{other invariants}\right\},$$

where ''other invariants'' means other than A_1, B_1, P_4, and P_5. Now we see how nicely we have chosen the four graphs in Sec. V.

(i) We note that the trivial terms are written in the total derivative form, therefore they do not contain any one of A_1, B_1, P_4, and P_5.

(ii) Now we know, from Appendix D,

$$P_4|_{(\partial\partial h)^3} = -\tfrac{1}{4}(\text{Graph1}) + \text{other } (\partial\partial h)^3 \text{ terms},$$

$$P_5|_{(\partial\partial h)^3} = \tfrac{1}{4}(\text{Graph2}) + \text{other } (\partial\partial h)^3 \text{ terms},$$

$$A_1|_{(\partial\partial h)^3} = -(\text{Graph3}) + \text{other } (\partial\partial h)^3 \text{ terms},$$

$$B_1|_{(\partial\partial h)^3} = -\tfrac{1}{4}(\text{Graph3}) + \tfrac{1}{4}(\text{Graph67}) + \text{other } (\partial\partial h)^3 \text{ terms}, \tag{52}$$

where ''other $(\partial\partial h)^3$ terms'' means ''other terms than Graph1,2,3 and 67''.[48]

(iii) We also note, again from Appendix D, Graphs 1, 2, 3, and 67 do not come from other invariants than (A_1, B_1, P_4, P_5).

These properties are due to the chosen four graphs. Then we can rewrite (51) as

$$\begin{aligned} G(x,x;t)|_{t^0,(\partial\partial h)^3} = \frac{1}{(4\pi)^3}\Bigg\{ & \left(-x - \frac{1}{4}y - 2w\right)(\text{Graph3}) + \frac{1}{4}(y - 8w)(\text{Graph67}) \\ & + \frac{1}{4}\left(\frac{3}{2}x - \frac{9}{8}y + 10z + 24w\right)(\text{Graph2}) \\ & - \frac{1}{4}\left(\frac{5}{4}x - \frac{7}{16}y + 10z + 16w\right)(\text{Graph1}) + \text{other } (\partial\partial h)^3 \text{ terms}\Bigg\}. \end{aligned} \tag{53}$$

Comparing the result of Sec. V, we finally obtain

$$\begin{gathered} x = -\frac{83}{189\times 60}, \quad y = \frac{31}{189\times 15}, \quad z = -\frac{11}{189\times 50}, \quad w = \frac{3}{189\times 10}, \\ G(x,x;t)|_{t^0} = -2g^{\mu\nu}\langle T_{\mu\nu}\rangle = \frac{1}{(4\pi)^3}\frac{1}{189}\sqrt{g}\left\{-\frac{83}{60}C_1 + \frac{31}{15}C_2 - \frac{11}{50}C_3 + \frac{3}{10}E + \text{trivial terms}\right\}. \end{gathered} \tag{54}$$

The trivial terms can be similarly obtained, but they do change by introducing the local counterterms in the action and seem to be unimportant. This result (54) is similar to that of Ref. 17.

VII. DISCUSSION AND CONCLUSION

We have presented a new algorithm to obtain the Weyl anomaly in the higher dimensions. The Feynman rules in coordinate space are presented. The graphical representation is exploited at all stages. Especially the graphical calculation is taken to efficiently compute the coefficients of the four graphs given in Sec. V. An explicit result of the Weyl anomaly for the 6-dim scalar gravity is obtained.

As the space–time dimension increases, the number of suffixes to deal with increases because we must treat higher products of Riemann tensors. The aid of the computer is indispensable in this algebraic calculation. As mentioned at some places in the text and the Appendices, the weight

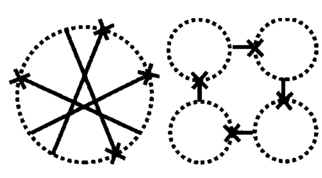

FIG. 10. Two important $(\partial\partial h)^4$ graphs in 8 dim.

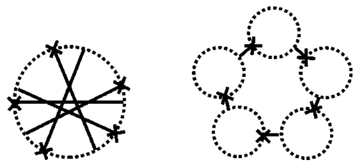

FIG. 11. Two important $(\partial\partial h)^5$ graphs in 10 dim.

calculation in the contraction of multisuffixed quantities relies on a computer (a C program[42]). Although we do not touch on the computer algorithm in the present paper, it helps to obtain the concrete results.

In the recent rapid development of string theory, the low-energy effective actions play an important role. They are field theories (supergravities) on various (higher) dimensions. The chiral anomalies of gravitational and gauge symmetries were well analyzed for those low-energy effective actions. The famous one is the pioneering work by Green and Schwarz,[49] where they found the SO(32) and $E_8\times E_8$ gauge group from the analysis of the chiral anomaly structure of 10-dim $N=1$ supergravity. Clearly the analysis from the Weyl anomaly side is lacking. From the present standpoint, we stress the importance of examining the string from the structure of the Weyl anomaly in the low-energy effective theories.[50] One reason is we know some important models often have vanishing Weyl anomaly due to conformal symmetry. Another is that the supersymmetry surely relates the Weyl anomaly to the chiral one. We can examine the role of supersymmetry from this point. Reference 51 recently analyzed the Weyl anomaly of conformal field theories using the adS/CFT correspondence conjectured by Ref. 52. Weyl anomalies are explicitly obtained for some theories in 2, 4, and 6 dimensions. We hope the present result will become useful when we go over the quantum field theory through the string theory.

We have most experience with quantization only in 0–4-dim field theories. Higher-dimensional quantum field theories have not been so thoroughly examined (at least systematically) so far, because they are unrenormalizable except for free theories. In the text we consider free theories on curved space. It is meaningful if we ignore the trivial terms and the divergent massive terms, which should be explained by the quantum gravitational mode. It might be possible to find a clue to make meaningful the higher-dimensional field theories through the present analysis.

Although the nonperturbative aspect is recently stressed in string theory, the perturbative analysis is still important because it is one of few reliable approaches to analyze dynamical aspects systematically. We suppose the analysis will become important to understand how the quantum field theory is generalized in the coming new era of Planck physics. Finally, we point out that the graphs shown in Figs. 10 and 11 are expected to become some of the ''important graphs'' in the Weyl anomaly calculation in 8-dim and 10-dim, respectively.

ACKNOWLEDGMENTS

The authors thank Professor M. Creutz for reading the manuscript carefully. One of the authors (S.I.) is thankful for the hospitality at the Department of Physics, Brookhaven National Laboratory, where this work was finished.

APPENDIX A: PROPAGATOR RULES

$G_0(x;t)$ is the solution of the heat equation, Eq. (13), and is given by

$$G_0(x;t)=\frac{1}{(4\pi t)^{n/2}}e^{-x^2/4t}I_N,\quad t>0,\quad x^2=(x_1)^2+(x_2)^2+\cdots+(x_n)^2, \tag{A1}$$

where I_N is the $N\times N$ unit matrix and $x=(x_1,x_2,\ldots,x_n)$ is the n-dim Euclidean space coordinates. It has the basic properties

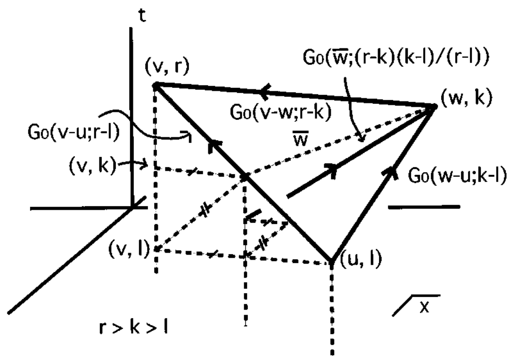

FIG. 12. Propagator Rule 1, Eq. (A5).

$$G_0(x;t)=G_0(-x;t)\text{:even function of } x,$$
$$G_0(\sqrt{a}x;at)=a^{-n/2}G_0(x;t), \quad a>0. \tag{A2}$$

The integral formula is given by

$$\int d^nx\, x^{\mu_1}\cdots x^{\mu_{2s}}G_0(x;t)=[\mu_1\cdots\mu_{2s}](2t)^sI_N, \tag{A3}$$

where $[\mu_1\cdots\mu_{2s}]$ is the totally symmetric sum of products of Kronecker's deltas and is concretely defined by

$$[\mu\nu]\equiv\delta^{\mu\nu},$$
$$[\mu\nu\lambda\sigma]\equiv\delta^{\mu\nu}\delta^{\lambda\sigma}+\delta^{\mu\sigma}\delta^{\nu\lambda}+\delta^{\mu\lambda}\delta^{\nu\sigma},$$
$$[\mu_1\cdots\mu_6]\equiv\delta^{\mu_1\mu_2}[\mu_3\cdots\mu_6]+\delta^{\mu_1\mu_3}[\cdots]+\cdots+\delta^{\mu_1\mu_6}[\cdots],$$
$$\cdots \tag{A4}$$
$$\cdots$$
$$[\mu_1\cdots\mu_{2s}]\equiv\delta^{\mu_1\mu_2}[\mu_3\cdots\mu_{2s}]+\delta^{\mu_1\mu_3}[\cdots]+\cdots+\delta^{\mu_1\mu_{2s}}[\cdots].$$

$G_0(x,t)$ has the ''convolution'' property. Propagator Rule 1:

$$G_0(v-w;r-k)G_0(w-u;k-l)=G_0\left(\bar{w};\frac{(r-k)(k-l)}{r-l}\right)G_0(v-u;r-l),$$
$$r>k>l, \quad \bar{w}=w-\frac{(k-l)v+(r-k)u}{r-l}. \tag{A5}$$

This relation is geometrically expressed in the coordinate space as in Fig. 12. Some special cases of above are given by Propagator Rule 2: $v=0$, $r=1$ in Eq. (A5),

$$G_0(w;1-k)G_0(w-u;k-l)=G_0\left(\bar{w};\frac{(1-k)(k-1)}{1-l}\right)G_0(u;1-l),$$
$$1>k>l, \quad \bar{w}=w-\frac{1-k}{1-l}u; \tag{A6}$$

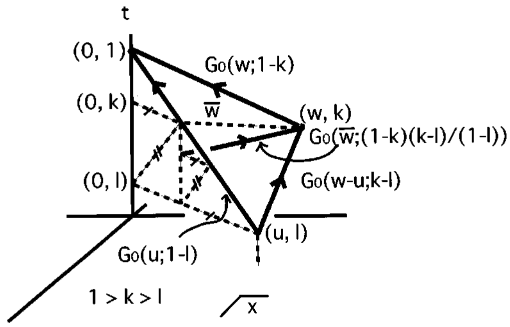

FIG. 13. Propagator Rule 2, Eq. (A6).

Propagator Rule 3: $u=0$, $l=0$ in Eq. (A5),

$$G_0(v-w;r-k)G_0(w;k)=G_0\left(\overline{w};\frac{(r-k)k}{r}\right)G_0(v;r),$$
$$r>k>0,\quad \overline{w}=w-\frac{k}{r}v; \tag{A7}$$

Propagator Rule 4: $u=0$, $v=0$, $l=0$, in Eq. (A5),

$$G_0(w;r-k)G_0(w;k)=\frac{1}{(4\pi r)^{n/2}}G_0\left(w;\frac{(r-k)k}{r}\right),\quad r>k>0. \tag{A8}$$

The above three relations are geometrically represented as in Figs. 13–15.

APPENDIX B: SUPPLEMENTARY CALCULATION OF SEC. V

In Sec. V of the text, we have evaluated $G(x,x;t)|_{t^0}$, focusing on four special $(\partial\partial h)^3$ terms (graphs) in order to fix four coefficients of the main Weyl anomaly terms given in (49). Among the four graphs, Graph3 and 67 only are explained for simplicity. We evaluate the remaining graphs, Graph1 and 2 to supplement Sec. V.

From (36),

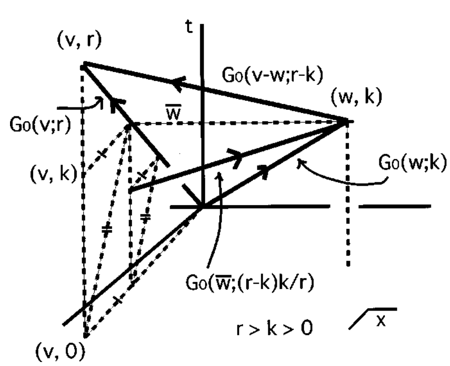

FIG. 14. Propagator Rule 3, Eq. (A7).

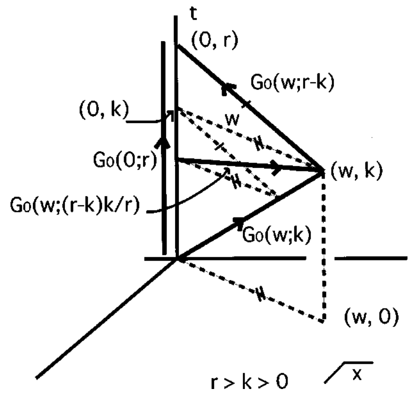

FIG. 15. Propagator Rule 4, Eq. (A8).

$$
G_3(x,x;t)|_{t^0} = \frac{1}{(4\pi)^3} \int d^6\bar{\bar{w}} \int d^6\overline{w} \int d^6 w \int_0^1 dr'' \int_0^{r''} dr' \int_0^{r'} dr
$$
$$
\times G_0\left(\bar{\bar{w}}; \frac{(1-r'')(r''-r')}{1-r'} \right) G_0\left(\overline{w}; \frac{(1-r')(r'-r)}{1-r} \right) G_0(w;(1-r)r)
$$
$$
\times V_0(x,w'';w''-w',r''-r') V_0(x,w';w'-w,r'-r) V_0(x,w;w,r) + \text{irrel.},
$$

$$
w'' = \bar{\bar{w}} + R_2 \overline{w} + R_3 w, \quad w'' - w' = \bar{\bar{w}} - S_2 \overline{w} - T_1 w, \tag{B1}
$$

$$
w' = \overline{w} + R_1 w, \quad w' - w = \overline{w} - S_1 w,
$$

where R_i, S_i, and T_1 are defined in (37). Because Graph1 and 2 do not have tadpoles and  subgraph, we can reduce V_0 of (C2) with (C3) to the following relevant form:

$$
V_0(x,v;w,r) = -\frac{1}{2} \times \frac{1}{4r^2} \times
$$
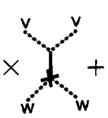

$$
+ \frac{1}{2r} \times
$$
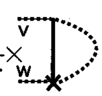

$$
+ \text{irrel. terms.} \tag{B2}
$$

The first V_0 in (B1) is evaluated as

$$
V_0(x,w'';w''-w',r''-r') = -\frac{1}{8(r''-r')^2} \times
$$
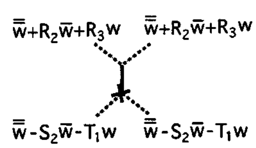

$$
+ \frac{1}{2(r''-r')} \times
$$
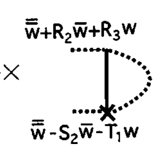

$$
+ \text{irrel. terms,} \tag{B3}
$$

$$\text{Second term of (B3)}=\frac{1\times(-1)}{2(r''-r')}\times$$
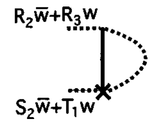

$$+\text{irrel. terms,}$$

$$\text{First term of (B3)}=-\frac{1}{8(r''-r')^2}\Bigg(2\times2\times(-1)\times$$
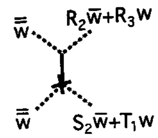

$$+$$
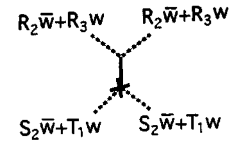

$$\Bigg)+\text{irrel. terms.} \qquad \text{(B4)}$$

The above V_0, with the $\bar{\bar{w}}$ integration, gives

$$\int d^6\bar{\bar{w}}\, V_0(x,w'';w''-w',r''-r')$$

$$=-\frac{1}{8(r''-r')^2}\times$$
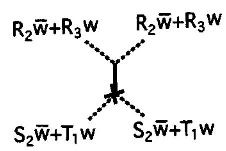

$$+\left\{\frac{-1}{2(r''-r')}+\frac{2\times2}{8(r''-r')^2}\times2\frac{(1-r'')(r''-r')}{(1-r')}\right\}\times$$
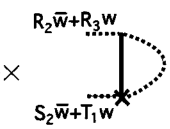

$$+\text{irrel. terms}$$

$$=-\frac{(R_2)^2(S_2)^2}{8(r''-r')^2}\times$$
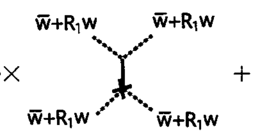

$$+\frac{1+r'-2r''}{2(r''-r')(1-r')}R_2S_2$$
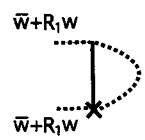

$$+\text{irrel. terms,} \qquad \text{(B5)}$$

where $\bar{\bar{w}}$ is integrated out and the relations $R_3/R_2=T_1/S_2=R_1$ are used. Therefore the $V_0V_0V_0$ part of (B1), with the $\bar{\bar{w}}$ integration, can be written as

$$\int d^6\bar{\bar{w}} V_0(x,w'';w''-w',r''-r')V_0(x,w';w'-w,r'-r)V_0(x,w;w,r)$$

$$=\left\{-\frac{(1-r'')^2}{8(1-r')^4}\times\right.$$
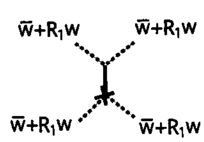

$$+\frac{(1+r'-2r'')(1-r'')}{2(1-r')^3}\times$$
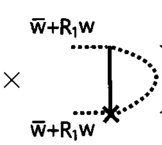

$$\left.\right\}$$

$$\times\left\{-\frac{1}{8(r'-r)^2}\times\right.$$
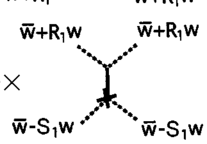

$$+\frac{1}{2(r'-r)}\times$$
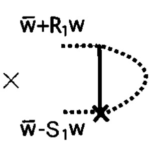

$$\left.\right\}$$

$$\times\left\{-\frac{1}{8r^2}\times\right.$$
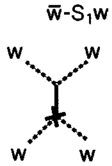

$$+\frac{1}{2r}\times$$
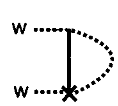

$$\left.\right\}+\text{irrel.} \qquad \text{(B6)}$$

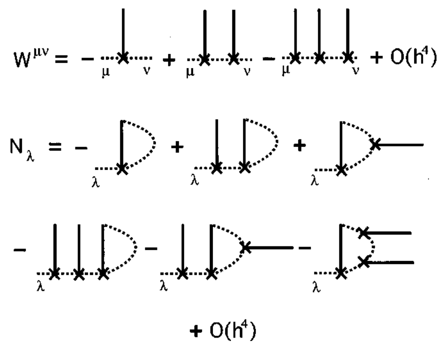

FIG. 16. $W^{\mu\nu}$ and N_λ.

Now we can further evaluate (B1) by picking up relevant powers of $\overline{w}$ in (B6) in the same way as Sec. V. After integration of the remaining coordinates $(\overline{w},w)$ and the parameters (r,r',r''), the final result is

$$G_3(x,x;t)|_{t^0}=\frac{1}{(4\pi)^3}\left\{\frac{1}{36\times 630}(\text{Graph1})+\frac{1}{36\times 35}(\text{Graph2})\right\}+\text{other terms.} \tag{B7}$$

APPENDIX C: WEAK FIELD EXPANSION OF 6-DIM SCALAR-GRAVITY THEORY

In the text, we have taken 6-dim scalar-gravity theory (1) as a higher-dimensional model. The expressions for the general theories are expressed in terms of $W_{\mu\nu}$, N_λ, and M (10). In this appendix we graphically express those general expressions for the case of the explicit model.

1. Graphs of *W*, *N*, and *M*

For the 6-dim scalar-gravity theory, $W_{\mu\nu}$, N_λ, and M are given by (11). Their weak-field expansions ($g_{\mu\nu}=\delta_{\mu\nu}+h_{\mu\nu}$) up to $O(h^3)$ are given in Fig. 16 for $W_{\mu\nu}$ and N_λ, and in Figs. 17–21 for M. Especially they are classified by the number of closed suffix loops.

2. Taylor Expansion of *W*, *N*, and *M*

In Sec. IV of the text, we have Taylor expanded $W_{\mu\nu}$, N_μ, and M, which are the background functional appearing in the differential operator **D**

$$\begin{aligned}W_{\mu\nu}(x+v)&=W_{\mu\nu}(x)+\partial_{\alpha_1}W_{\mu\nu}\cdot v^{\alpha_1}+\tfrac{1}{2}\partial_{\alpha_1}\partial_{\alpha_2}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}+\cdots,\\ N_\mu(x+v)&=N_\mu(x)+\partial_{\alpha_1}N_\mu\cdot v^{\alpha_1}+\tfrac{1}{2}\partial_{\alpha_1}\partial_{\alpha_2}N_\mu\cdot v^{\alpha_1}v^{\alpha_2}+\cdots,\\ M(x+v)&=M(x)+\partial_{\alpha_1}M\cdot v^{\alpha_1}+\tfrac{1}{2}\partial_{\alpha_1}\partial_{\alpha_2}M\cdot v^{\alpha_1}v^{\alpha_2}+\cdots.\end{aligned} \tag{C1}$$

We focus on those terms that have only $\partial\partial h$-type ones. They are sufficient to determine the Weyl anomaly, and are used in the text.

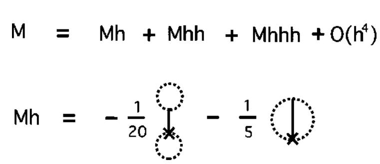

FIG. 17. M, order of h.

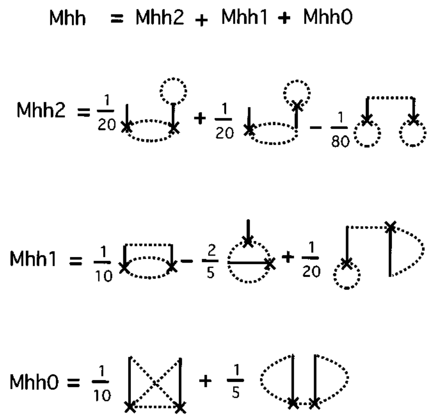

FIG. 18. *M*, order of h^2.

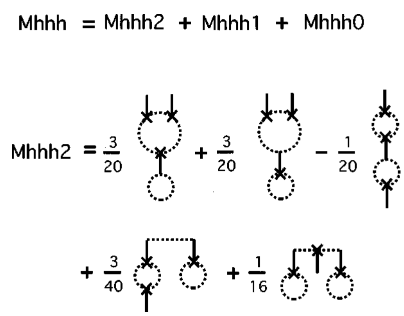

FIG. 19. *M*, order of h^3 and loop No.=2.

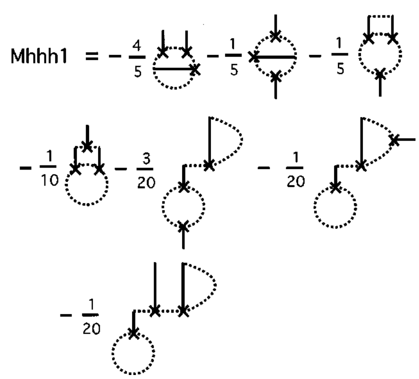

FIG. 20. *M*, order of h^3 and loop No.=1.

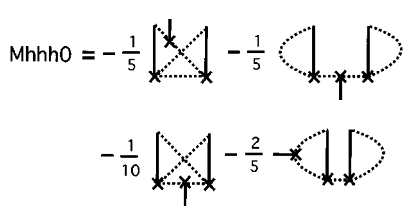

FIG. 21. *M*, order of h^3 and loop No.=0.

(i) Expansion terms appearing in V_0 of (33),

$$V_0(x,v;w,r)=\frac{1}{2!}\partial_{\alpha_1}\partial_{\alpha_2}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\partial_{\alpha_1}N_\mu\cdot v^{\alpha_1}\left(-\frac{w_\mu}{2r}\right)+M, \quad \text{(C2)}$$

$\frac{1}{2!}\partial_{\alpha_1}\partial_{\alpha_2}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}= -\frac{1}{2}$ 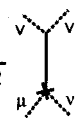 + irrel. terms,

$\partial_{\alpha_1}N_\lambda\cdot v^{\alpha_1}= -$ 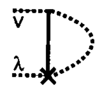 + irrel. terms,

$M= -\frac{1}{20}$  $-\frac{1}{5}$  + irrel. terms. (C3)

(ii) Expansion terms appearing in V_1 of (33),

$$V_1(x,v;w,r)=\frac{1}{4!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_4}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_4}\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\frac{1}{3!}\partial_{\alpha_1}\partial_{\alpha_2}\partial_{\alpha_3}N_\mu$$

$$\cdot v^{\alpha_1}v^{\alpha_2}v^{\alpha_3}\left(-\frac{w_\mu}{2r}\right)+\frac{1}{2!}\partial_{\alpha_1}\partial_{\alpha_2}M\cdot v^{\alpha_1}v^{\alpha_2}, \quad \text{(C4)}$$

$\frac{1}{4}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_4}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_4}= \frac{1}{2}$ 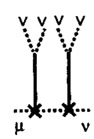 + irrel. terms,

$\frac{1}{3!}\partial_{\alpha_1}\partial_{\alpha_2}\partial_{\alpha_3}N_\lambda\cdot v^{\alpha_1}v^{\alpha_2}v^{\alpha_3}= \frac{1}{2}\Bigg[$ 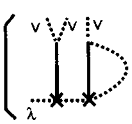 + 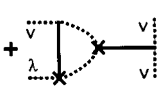 $\Bigg]$ + irrel. terms,

$\frac{1}{2!}\partial_{\alpha_1}\partial_{\alpha_2}M\cdot v^{\alpha_1}v^{\alpha_2}=\frac{1}{2}\Bigg\{\frac{1}{20}$ 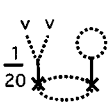 $+\frac{1}{20}$ 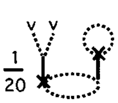 $-\frac{2}{80}$ 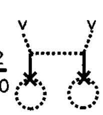

$+\frac{2}{10}$ 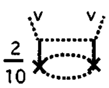 $-\frac{2}{5}$ 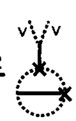 $+\frac{2}{20}$ 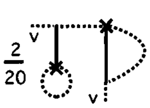

$+\frac{2}{10}$ 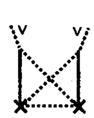 $+\frac{2}{5}$ 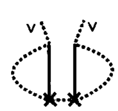 $\Bigg\}$ + irrel. terms. (C5)

(iii) Expansion terms appearing in V_2 of (33),

$$V_2(x,v;w,r)=\frac{1}{6!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_6}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_6}\left(-\frac{\delta_{\mu\nu}}{2r}+\frac{w_\mu w_\nu}{4r^2}\right)+\frac{1}{5!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_5}N_\mu \cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_5}\left(-\frac{w_\mu}{2r}\right)+\frac{1}{4!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_4}M\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_4}. \tag{C6}$$

$$\frac{1}{6!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_6}W_{\mu\nu}\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_6}=-\frac{3}{4}$$

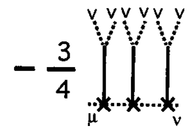

$+$ irrel. terms,

$$\frac{1}{5!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_5}N_\lambda\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_5}$$

$=$

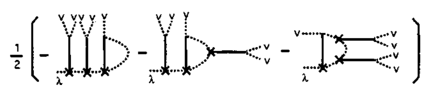

$+$ irrel. terms,

$$\frac{1}{4!}\partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_4}M\cdot v^{\alpha_1}v^{\alpha_2}\cdots v^{\alpha_4}$$

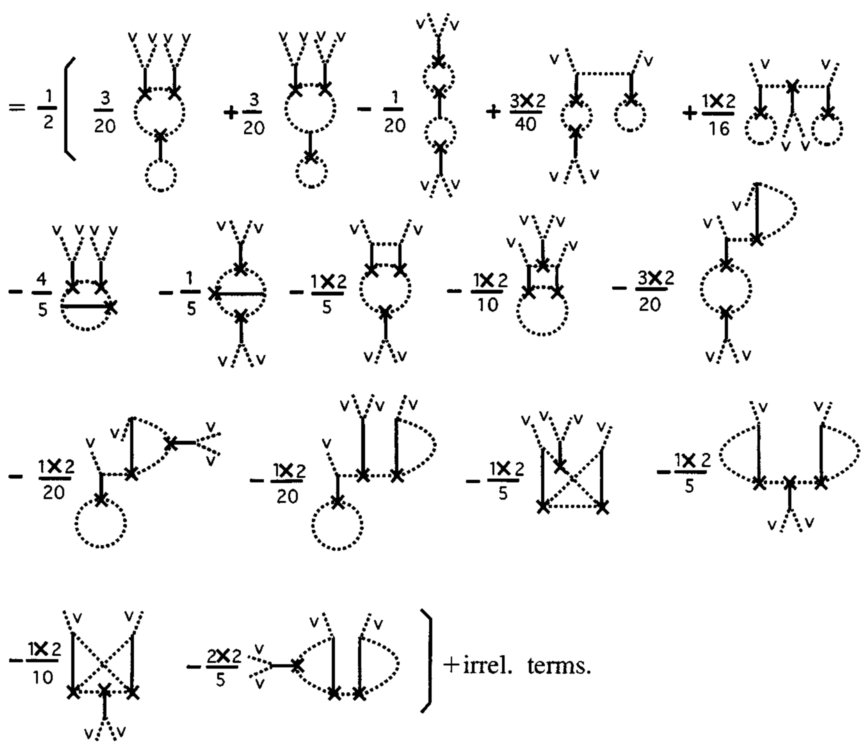

$+$ irrel. terms. (C7)

APPENDIX D: WEAK-EXPANSION OF INVARIANTS WITH (MASS)^6-DIM

In 6-dim space, there are totally 17 independent general invariants given in (50). In this appendix, we list the coefficients of $(\partial\partial h)^3$ terms when the invariants are weak-field expanded $(g_{\mu\nu}=\delta_{\mu\nu}+h_{\mu\nu})$. They are obtained by a computer using a C program.[42] Among 17 ones, $O_i(i=1-4)$ terms do not have $(\partial\partial h)^3$ terms. As shown in the first column of Table I, there are 90 independent $(\partial\partial h)^3$ terms. Some terms (G3,G67,G1,G2) are given, in the ordinary literal form, in (39) and graphically in Fig. 9. Their complete list is graphically given in Ref. 40. For example, the A_1 column says

TABLE I. Weak Expansion of Invariants with (Mass)^6-Dim: $(\partial\partial h)^3$. (a) G1–G13($\underline{l}$[No of suffix loops]$=1$); (b) G14-G42 ($\underline{l}=2$); (c) G43-G69 ($\underline{l}=3$); (d) G70-G85 ($\underline{l}=4$); (e) G86-G89 ($\underline{l}=5$); (f) G90 ($\underline{l}=6$).

Graph	P_1	P_2	P_3	P_4	P_5	P_6	A_1	B_1	T_1	T_2	T_3	T_4	S
						(a) $\underline{l}=1$							
G1	0	0	0	$-\frac{1}{4}$	0	0	0	0	0	0	0	0	0
G2	0	0	0	0	$-\frac{1}{4}$	0	0	0	0	0	0	0	0
G3	0	0	0	0	0	0	-1	$-\frac{1}{4}$	0	0	0	0	0
G4	0	0	0	$-\frac{3}{4}$	0	0	0	0	0	0	0	-2	-4
G5	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	0	-1	-2
G6	0	0	0	0	0	$\frac{1}{2}$	0	0	0	-1	0	0	0
G7	0	0	0	0	0	0	0	$\frac{3}{2}$	0	0	2	0	0
G8	0	0	0	0	$\frac{1}{2}$	0	0	0	0	-1	0	0	0
G9	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	0	0	0
G10	0	0	0	0	$\frac{1}{4}$	0	0	0	0	-1	0	0	-8
G11	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	0	0	-16
G12	0	0	0	0	0	0	0	$\frac{3}{2}$	0	0	6	0	-8
G13	0	0	0	0	0	0	-3	$-\frac{3}{4}$	0	0	-2	0	-4
						(b) $\underline{l}=2$							
G14	0	0	0	0	0	0	1	0	0	0	-1	0	12
G15	0	0	0	0	0	0	3	0	0	0	-1	0	0
G16	0	0	0	0	-1	0	0	0	0	1	0	-2	-4
G17	0	0	0	0	0	$-\frac{1}{2}$	0	0	0	1	0	0	-16
G18	0	0	0	0	0	0	0	$-\frac{3}{4}$	0	0	-1	0	-4
G19	0	0	0	0	0	$-\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0	$\frac{3}{2}$	7
G20	0	0	0	0	0	0	0	$-\frac{3}{4}$	0	0	-2	0	6
G21	0	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	0	0	0
G22	0	0	0	0	0	0	0	$-\frac{3}{2}$	0	0	-2	0	0
G23	0	$-\frac{1}{2}$	0	0	0	0	0	0	-2	0	0	0	0
G24	0	0	2	0	0	0	0	0	-1	0	0	0	0
G25	0	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	0	0	0
G26	0	$-\frac{1}{2}$	0	0	0	0	0	0	0	0	0	0	0
G27	0	0	0	$\frac{3}{8}$	0	0	0	0	0	0	0	1	2
G28	0	0	0	0	0	$-\frac{1}{4}$	0	0	0	0	0	$\frac{1}{2}$	1
G29	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	$\frac{5}{4}$	0	0	8
G30	0	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	0	0	8

TABLE I. (*Continued.*)

Graph	P_1	P_2	P_3	P_4	P_5	P_6	A_1	B_1	T_1	T_2	T_3	T_4	S
G31	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	$\frac{1}{4}$	0	0	0
G32	0	0	0	$\frac{3}{4}$	0	0	0	0	0	0	0	1	2
G33	0	0	0	$\frac{3}{8}$	0	0	0	0	0	0	0	1	2
G34	0	0	0	0	0	$-\frac{1}{4}$	0	0	0	$\frac{1}{2}$	0	0	0
G35	0	0	0	$\frac{3}{8}$	0	0	0	0	0	$-\frac{1}{4}$	0	0	-4
G36	0	0	0	0	0	$-\frac{1}{4}$	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	-2
G37	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	$\frac{1}{2}$	0	0	-8
G38	0	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	1	0	-4
G39	0	0	0	$\frac{3}{8}$	0	0	0	0	0	$-\frac{1}{4}$	0	1	-2
G40	0	0	0	0	0	$-\frac{1}{4}$	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	-1
G41	0	0	0	0	$-\frac{1}{2}$	0	0	0	0	$-\frac{1}{2}$	0	0	0
G42	0	0	0	$\frac{3}{4}$	0	0	0	0	0	$-\frac{1}{2}$	0	0	0
						(c) $\underline{l}=3$							
G43	0	0	0	0	$\frac{1}{4}$	0	0	0	0	$\frac{1}{2}$	0	0	-2
G44	0	0	0	$-\frac{3}{4}$	0	0	0	0	0	1	0	0	-8
G45	0	0	0	0	0	$\frac{1}{4}$	0	0	0	$-\frac{1}{2}$	$-\frac{1}{2}$	0	14
G46	0	0	0	0	0	$\frac{1}{4}$	0	0	0	$-\frac{1}{4}$	0	$-\frac{3}{4}$	$-\frac{7}{2}$
G47	0	0	0	0	0	$\frac{1}{4}$	0	0	0	$-\frac{1}{2}$	0	0	8
G48	0	0	0	0	$\frac{1}{2}$	0	0	0	0	$-\frac{1}{4}$	0	1	2
G49	0	0	0	0	0	0	0	$\frac{3}{4}$	0	0	1	0	4
G50	0	0	-1	0	0	0	0	0	$\frac{3}{2}$	0	0	0	0
G51	0	0	0	$-\frac{3}{4}$	0	0	0	0	0	$\frac{1}{2}$	0	0	4
G52	0	0	0	0	$\frac{1}{2}$	0	0	0	0	$\frac{1}{4}$	0	0	4
G53	0	$\frac{1}{2}$	0	0	0	0	0	0	2	0	0	0	0
G54	0	0	-2	0	0	0	0	0	1	0	0	0	0
G55	0	0	0	$-\frac{3}{4}$	0	0	0	0	0	$\frac{1}{2}$	0	-1	2
G56	0	$\frac{1}{2}$	0	0	0	0	0	0	0	0	0	0	0
G57	0	0	0	0	0	$\frac{1}{4}$	0	0	0	$-\frac{1}{4}$	$-\frac{1}{2}$	$-\frac{3}{4}$	$\frac{9}{2}$
G58	0	0	0	0	$\frac{1}{2}$	0	0	0	0	0	0	1	4
G59	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	-1	0	0
G60	0	0	0	0	$\frac{1}{2}$	0	0	0	0	$\frac{1}{2}$	0	1	-2

TABLE I. (*Continued.*)

Graph	P_1	P_2	P_3	P_4	P_5	P_6	A_1	B_1	T_1	T_2	T_3	T_4	S
G61	0	1	0	0	0	0	0	0	-2	0	0	0	0
G62	0	0	0	$-\frac{3}{4}$	0	0	0	0	0	0	0	$-\frac{3}{2}$	-3
G63	0	0	0	0	$\frac{1}{4}$	0	0	0	0	$-\frac{1}{4}$	0	0	0
G64	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	0	0	0
G65	0	0	0	0	$\frac{1}{2}$	0	0	0	0	$-\frac{1}{2}$	0	1	2
G66	0	1	0	0	0	0	0	0	2	0	0	0	0
G67	0	0	0	0	0	0	0	$\frac{1}{4}$	0	0	0	0	0
G68	0	0	-1	0	0	0	0	0	0	0	0	0	0
G69	-1	0	0	0	0	0	0	0	0	0	0	0	0
						(d) $\underline{l}=4$							
G70	0	0	0	$\frac{1}{8}$	0	0	0	0	0	$-\frac{1}{4}$	0	0	4
G71	0	0	0	0	$-\frac{1}{4}$	0	0	0	0	$-\frac{1}{8}$	0	$-\frac{1}{2}$	-2
G72	0	0	1	0	0	0	0	0	$-\frac{3}{2}$	0	0	0	0
G73	0	0	0	$\frac{3}{8}$	0	0	0	0	0	$-\frac{1}{2}$	0	0	4
G74	0	-1	0	0	0	0	0	0	2	0	0	0	0
G75	0	0	0	0	$-\frac{1}{4}$	0	0	0	0	$-\frac{1}{2}$	0	$-\frac{1}{2}$	3
G76	0	$-\frac{1}{4}$	0	0	0	0	0	0	1	0	0	0	0
G77	0	0	0	$\frac{1}{8}$	0	0	0	0	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$
G78	0	0	0	0	$-\frac{1}{4}$	0	0	0	0	$\frac{1}{8}$	0	$-\frac{1}{2}$	-1
G79	0	$-\frac{1}{4}$	0	0	0	0	0	0	$-\frac{1}{2}$	0	0	0	0
G80	0	0	0	$\frac{3}{8}$	0	0	0	0	0	$-\frac{1}{4}$	0	$\frac{1}{4}$	$-\frac{1}{2}$
G81	0	-1	0	0	0	0	0	0	-2	0	0	0	0
G82	0	0	0	0	$-\frac{1}{4}$	0	0	0	0	$-\frac{1}{4}$	0	$-\frac{1}{2}$	1
G83	0	$-\frac{1}{2}$	0	0	0	0	0	0	1	0	0	0	0
G84	0	0	1	0	0	0	0	0	0	0	0	0	0
G85	3	0	0	0	0	0	0	0	0	0	0	0	0
						(e) $\underline{l}=5$							
G86	0	$\frac{1}{4}$	0	0	0	0	0	0	-1	0	0	0	0
G87	0	$\frac{1}{4}$	0	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0
G88	-3	0	0	0	0	0	0	0	0	0	0	0	0
G89	0	$\frac{1}{2}$	0	0	0	0	0	0	-1	0	0	0	0
						(f) $\underline{l}=6$							
G90	1	0	0	0	0	0	0	0	0	0	0	0	0

$$A_1(R_{\mu\nu\lambda\sigma}R^{\sigma\lambda}{}_{\tau\omega}R^{\omega\tau\nu\mu}) = (-1)\times\text{Graph3}(\partial_\sigma\partial_\tau h_{\mu\nu}\cdot\partial_\nu\partial_\lambda h_{\tau\omega}\cdot\partial_\omega\partial_\mu h_{\lambda\sigma}) + (-3)$$
$$\times\text{Graph13}(\partial_\mu\partial_\nu h_{\tau\sigma}\cdot\partial_\mu\partial_\omega h_{\lambda\sigma}\cdot\partial_\tau\partial_\omega h_{\nu\lambda}) + (1)\times\text{Graph14}(\partial_\mu\partial_\nu h_{\omega\sigma}$$
$$\cdot\partial_\nu\partial_\lambda h_{\tau\omega}\cdot\partial_\lambda\partial_\mu h_{\sigma\tau}) + (3)\times\text{Graph15}(\partial_\mu\partial_\nu h_{\omega\sigma}\cdot\partial_\tau\partial_\omega h_{\nu\lambda}\cdot\partial_\lambda\partial_\mu h_{\sigma\tau})$$
$$+O(h^4). \tag{D1}$$

The content of the table is fully used in Secs. V and VI of the text. Especially the choice of the four graphs of Fig. 9 relies on the table.

[1] L. Alvarez-Gaumé and E. Witten, Nucl. Phys. B **234**, 269 (1983).
[2] L. Bonora, P. Pasti, and M. Bregola, Class. Quantum Grav. **3**, 635 (1986).
[3] T. Arakelyan, D. R. Karakhanyan, R. P. Manvelyan, and R. L. Mkrtchyan, Phys. Lett. B **353**, 52 (1995).
[4] D. R. Karakhanyan, R. P. Manvelyan, and R. L. Mkrtchyan, Mod. Phys. Lett. A **11**, 409 (1996).
[5] S. Ichinose and S. D. Odintsov, Nucl. Phys. B **539**, 643 (1999).
[6] S. Deser, Uniqueness of $D=11$ supergravity, *Lecture at 'Quantum Mechanics of Fundamental Systems VI'*, Santiago, Chile, August 1997, hep-th/9712064.
[7] S. Ferrara and B. Zumino, Nucl. Phys. B **87**, 207 (1975).
[8] J. Schwinger, Phys. Rev. **82**, 664 (1951).
[9] B. DeWitt, *Dynamical Theory of Groups and Fields* (Gordon and Breach, New York, 1965).
[10] R. T. Seeley, Proc. Symp. Pure Math. **10**, 288 (1967).
[11] F. Bastianelli, Nucl. Phys. B **376**, 113 (1992).
[12] F. Bastianelli and P. van Nieuwenhuizen, Nucl. Phys. B **389**, 53 (1993).
[13] G. t'Hooft, Nucl. Phys. B **62**, 444 (1973).
[14] G. t'Hooft and M. Veltman, Ann. Inst. Henri Poincaré, Sect. A **20**, 69 (1974).
[15] P. B. Gilkey, J. Diff. Geom. **10**, 601 (1975).
[16] I. G. Avramidi, ''The covariant methods for calculation of the effective action in quantum field theory and the investigation of higher derivative quantum gravity,'' Ph.D. thesis, Moscow State University, Moscow, 1986; hep-th/9510140.
[17] I. G. Avramidi, Nucl. Phys. B **55**, 712 (1991).
[18] A. O. Barvinsky and G. A. Vilkovisky, Phys. Rep. C **119**, 1 (1985).
[19] J. de Boer, B. Peeters, K. Skenderis, and P. van Nieuwenhuizen, Nucl. Phys. B **446**, 211 (1995).
[20] J. de Boer, B. Peeters, K. Skenderis, and P. van Nieuwenhuizen, Nucl. Phys. B **459**, 631 (1996).
[21] S. Yajima, Class. Quantum Grav. **14**, 2853 (1997).
[22] I. L. Buchbinder, S. D. Odintsov, and I. L. Shapiro, *Effective Action in Quantum Gravity* (IOP, Bristol, 1992).
[23] Other examples of the coordinate space approach are the operator product expansion method and the lattice field theory. The former was exploited in the 2 dim conformal field theory. Traditionally the coordinate space has been used for the analysis of conformally invariant or covariant Green's functions. For example, see a recent progress of Ref. 24.
[24] J. Pachos and R. Schiappa, ''Conformal symmetry and the three point function for the gravitational axial anomaly,'' MIT-CTP#2760, hep-th/9807128, to be published in Phys. Rev. D.
[25] The heat kernel is formulated in the coordinate space, using the quantum-mechanical nonlinear sigma model, and the Feynman rules are given in Refs. 11, 12, 19, 20.
[26] S. Ichinose and N. Ikeda, Phys. Rev. D **53**, 5932 (1996).
[27] It must be compared with the dimensional regularization.
[28] K. Fujikawa, Nucl. Phys. B **226**, 437 (1983).
[29] K. Fujikawa, Phys. Rev. Lett. **42**, 1195 (1979); **44**, 1733 (1980); Phys. Rev. D **21**, 2848 (1980); **22**, 1499(E) (1980).
[30] K. Fujikawa, *Lecture in Quantum Gravity and Cosmology*, Kyoto Summer *Institute 1985*, edited by Sato and Inami (World Scientific, Singapore, 1985), p. 106. This is the good review of the series of works by Fujikawa.
[31] E. Abdalla, M. C. B. Abdalla, and K. D. Rothe, *Non-Perturbative Methods in 2 Dimensional Quantum Field Theory* (World Scientific, Singapore, 1991).
[32] E. D'Hoker, ''Tasi Lectures on critical string theory,'' UCLA/92/TEP/30, 1992.
[33] S. Ichinose and N. Ikeda, ''Gauge symmetry of the heat kernel and anomaly formulae,'' DAMTP/96-87, hep-th/9610136.
[34] The case of special $W^{ij}_{\mu\nu}$ (Riemann geometry) was done by Gilkey,[15] where he analyzed using the normal coordinates in order to keep the general covariance manifestly.
[35] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, New York, 1964).
[36] O. Alvarez, Nucl. Phys. B **216**, 125 (1983).
[37] The condition $t>0$ in $G_0(x-y;t)$ requires the directedness.
[38] These properties come only from the physical dimensions of $W_{\mu\nu}$, N_μ, and M, not the speciality of the present model. The relevant-part selection in this section is valid for general theories.
[39] The divergent parts t^{-1}, t^{-2},..., are considered to be canceled by the divergent counterterms. They generally correspond to power divergences (not log divergence) and the counterterms are those operators that have massive couplings: for example, the cosmological term, the Einstein term.
[40] S. Ichinose and N. Ikeda, J. Math. Phys. **38**, 6475 (1997).

[41] The $\overline{w}^8$ term gives disconnected graphs, and $\overline{w}^6$ terms give graphs that conflict with the second one of the common features of Graph3 and 67.
[42] S. Ichinose, Int. J. Mod. Phys. C **9**, 243 (1998).
[43] M. J. Duff, Nucl. Phys. B **125**, 334 (1977).
[44] S. Deser and A. Schwimmer, Phys. Lett. B **309**, 279 (1993).
[45] S. Ichinose, US-97-08, ''General structure of conformal anomaly and 4 Dimensional photon–dilaton gravity,'' hep-th/9801056.
[46] S. A. Fulling, R. C. King, B. G. Wybourne, and C. J. Cummins, Class. Quantum Grav. **9**, 1151 (1992).
[47] S. Ichinose, Class. Quantum Grav. **12**, 1021 (1995).
[48] The total independent $(\partial\partial h)^3$ terms are given by 90 terms shown in the first column of the table in Appendix D.
[49] M. B. Green and J. H. Schwarz, Phys. Lett. B **149**, 117 (1984).
[50] M. J. Duff, Class. Quantum Grav. **11**, 1387 (1994).
[51] M. Henningson and K. Skenderis, J. High Energy Phys. **9807**, 23 (1998).
[52] J. M. Maldacena, Adv. Theor. Math. Phys. **2**, 231 (1998).

Superintegrability on the two-dimensional hyperboloid. II

E. G. Kalnins
Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand

W. Miller, Jr.
School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455

Ye. M. Hakobyan and G. S. Pogosyan
Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, Dubna, Moscow Region 141980, Russia

(Received 14 December 1998; accepted for publication 7 January 1999)

This work is devoted to the investigation of the quantum mechanical systems on the two-dimensional hyperboloid which admits separation of variables in at least two coordinate systems. Here we consider two potentials introduced in a paper of C. P. Boyer, E. G. Kalnins, and P. Winternitz [J. Math. Phys. **24**, 2022 (1983)], which have not yet been studied. We give an example of an interbasis expansion and work out the structure of the quadratic algebra generated by the integrals of motion. © *1999 American Institute of Physics.* [S0022-2488(99)00505-8]

I. INTRODUCTION

Superintegrable systems on the two-dimensional hyperboloid were introduced and developed in Refs. 1–3. In distinction to the cases of two-dimensional Euclidean space and the two-sphere, the classification of superintegrable systems on the hyperboloid is difficult. To date only the four potentials studied in Ref. 3 and two more listed in Ref. 1 are known. In the present paper two potentials are considered, which were constructed in Ref. 1 but have not previously been investigated. These potentials both have only a finite number of bound states. At this point we have treated all the potentials that arise by restriction from Hermitean hyperbolic space. We follow the approach of Ref. 3, which contains an introduction and motivation.

The two-dimensional hyperboloid is characterized via the Cartesian coordinates $\omega_0, \omega_1, \omega_2$ where $\omega_0^2-\omega_1^2-\omega_2^2=1$, $\omega_0>1$. The requirement $\omega_0>1$ means that we consider only the upper sheet of the double-sheet hyperboloid. Throughout this paper we will consider the Schrödinger equation on the hyperboloid in the form ($\hbar=m=1$)

$$H\Psi\equiv\left(-\tfrac{1}{2}\Delta_{\mathrm{LB}}+V\right)\Psi=E\Psi, \tag{1}$$

where V is a potential function and the Laplace–Beltrami operator Δ_{LB} is written as

$$\Delta_{\mathrm{LB}}=K_3^2+K_2^2-M_1^2. \tag{2}$$

Here K_3, K_2, M_1 generate the Lie algebra so(2,1) (Refs. 4 and 5):

$$K_3=\omega_0\partial_{\omega_1}+\omega_1\partial_{\omega_0}, \quad K_2=\omega_0\partial_{\omega_2}+\omega_2\partial_{\omega_0}, \quad M_1=\omega_1\partial_{\omega_2}-\omega_2\partial_{\omega_1}, \tag{3}$$

and

$$[K_3,K_2]=M_1, \quad [K_2,M_1]=-K_3, \quad [K_3,M_1]=K_2. \tag{4}$$

The Schrödinger equation (1) for $V=0$ separates in nine coordinate systems.[6] Introduction of a potential breaks the symmetry and, in general, reduces the number of coordinate systems permitting separability, usually to zero. We consider the following two potentials (see Table I), constructed in Ref. 1, for which (1) is superintegrable.

TABLE I. Superintegrable potentials.

Potential $V(\omega)$	Coordinate system
$V_1=\frac{\alpha^2}{\omega_2^2}-\frac{\gamma^2}{(\omega_0-\omega_1)^2}+\beta^2\frac{\omega_0+\omega_1}{(\omega_0-\omega_1)^3}$	Equidistant Elliptic-parabolic Hyperbolic-parabolic Horicyclic
$V_2=\frac{\alpha^2}{\omega_2^2}+\gamma^2\frac{\omega_0\omega_1}{(\omega_0^2+\omega_1^2)^2}+(\alpha^2-\beta^2)\frac{\omega_0^2-\omega_1^2}{(\omega_0^2+\omega_1^2)^2}$	Equidistant Semi-hyperbolic

Recall that (1) is *superintegrable* for a given potential V if it is separable simultaneously in at least two coordinate systems.

II. FIRST POTENTIAL

The first considered potential is

$$V_1=\frac{\alpha^2}{\omega_2^2}-\frac{\gamma^2}{(\omega_0-\omega_1)^2}+\beta^2\frac{\omega_0+\omega_1}{(\omega_0-\omega_1)^3}, \tag{5}$$

where α, β, γ are positive constants. The corresponding Schrödinger equation admits separable solutions in four coordinate systems: equidistant, elliptic–parabolic, hyperbolic–parabolic, and horicyclic.

A. Solutions of the Schrödinger equation

1. Equidistant coordinates

In this coordinate system

$$\omega_0=\cosh\tau_1\cosh\tau_2,\quad \omega_1=\cosh\tau_1\sinh\tau_2,\quad \omega_2=\sinh\tau_1$$

$[\tau_1,\tau_2\in(-\infty,\infty)]$, the potential V_1 has the form

$$V_1(\tau_1,\tau_2)=\frac{\alpha^2}{\sinh^2\tau_1}+\frac{1}{\cosh^2\tau_1}\frac{\beta^2-\gamma^2(\cosh\tau_2-\sinh\tau_2)^2}{(\cosh\tau_2-\sinh\tau_2)^4}. \tag{6}$$

After putting

$$\Psi(\tau_1,\tau_2)=(\cosh\tau_1)^{-1/2}S_1(\tau_1)S_1(\tau_2), \tag{7}$$

we come to the system of equations:

$$\frac{d^2S_2}{d\tau_2^2}+[-\mu^2-2\beta^2e^{4\tau_2}+2\gamma^2e^{2\tau_2}]S_2=0, \tag{8}$$

$$\frac{d^2S_1}{d\tau_1^2}+\left[\left(2E-\frac{1}{4}\right)+\frac{\mu^2-\frac{1}{4}}{\cosh^2\tau_1}-\frac{2\alpha^2}{\sinh^2\tau_1}\right]S_1=0, \tag{9}$$

where μ is the equidistant separation constant. The first equation (8) could be considered as a one-dimensional Schrödinger equation for the Morse potential[7] and the orthonormalized solution is given by the expression

$$S_2(\tau_2)\equiv S_m^{(\beta,\mu)}(z)=\sqrt{\frac{2\mu\Gamma(m+\mu+1)}{m!\Gamma^2(\mu+1)}}e^{-z/2}z^{\mu/2}{}_2F_1(-m,\mu+1;z)$$

$$=\sqrt{\frac{2\mu m!}{\Gamma(m+\mu+1)}}e^{-z/2}z^{\mu/2}L_m^{\mu}(z),\quad z=\sqrt{2}\beta e^{2\tau_2} \tag{10}$$

where $L_m^{\mu}(z)$ are the Laguerre polynomials.[8] The separation constant is quantized as

$$\mu=-2m-1+\frac{\gamma^2}{\sqrt{2}\beta},\quad 0\leqslant m\leqslant\left[\frac{1}{2}\left(\frac{\gamma^2}{\sqrt{2}\beta}-1\right)\right]. \tag{11}$$

The second equation (9) represents the modified Pöschl–Teller equation.[3,9] The orthonormalized wave function is given by

$$S_1(\tau_1)\equiv S_n^{(\alpha,\mu)}(\tau_1)=\sqrt{\frac{2(\mu-\sqrt{2\alpha^2+1/4}-2n-1)\Gamma(\mu-n)n!}{\Gamma(\mu-\sqrt{2\alpha^2+1/4}-n)\Gamma(1+n+\sqrt{2\alpha^2+1/4})}}$$

$$\times(\sinh\tau_1)^{1/2+\sqrt{2\alpha^2+1/4}}(\cosh\tau_1)^{1/2-\mu}P_n^{(\sqrt{2\alpha^2+1/4},-\mu)}(\cosh 2\tau_1), \tag{12}$$

with $n=0,1,...,[\frac{1}{2}(\mu-1-\sqrt{2\alpha^2+\frac{1}{4}})]$, where $P_n^{(\alpha,\beta)}(x)$ is the Jacobi polynomial.[8] The quantized energy is

$$E_N=-\frac{1}{2}\left(\mu-\sqrt{2\alpha^2+\frac{1}{4}}-2n-1\right)^2+\frac{1}{8}=-\frac{1}{2}\left(2N+2+\sqrt{2\alpha^2+\frac{1}{4}}-\frac{\gamma^2}{\sqrt{2}\beta}\right)^2+\frac{1}{8}, \tag{13}$$

where $N=m+n$ is the principal quantum number and the bound states occur for

$$0\leqslant N\leqslant\left[\frac{1}{2}\left(\frac{\gamma^2}{\sqrt{2}\beta}-\sqrt{2\alpha^2+1/4}-2\right)\right]. \tag{14}$$

The orthonormalized total wave function $\Psi_{nm}(\tau_1,\tau_2)$ is given by (7), (10), and (12).

The symmetry operator describing this coordinate system is

$$L_1\Psi_{nm}(\tau_1,\tau_2)\equiv\left[K_3^2-2\beta^2\left(\frac{\omega_0+\omega_1}{\omega_0-\omega_1}\right)^2+2\gamma^2\frac{\omega_0+\omega_1}{\omega_0-\omega_1}\right]\Psi_{nm}(\tau_1,\tau_2)$$

$$=\left(-2m-1+\frac{\gamma^2}{\sqrt{2}\beta}\right)^2\Psi_{nm}(\tau_1,\tau_2). \tag{15}$$

2. Horicyclic coordinates

In the horicyclic coordinates,

$$\omega_0=\frac{x^2+y^2+1}{2y},\quad \omega_1=\frac{x^2+y^2-1}{2y},\quad \omega_2=\frac{x}{y} \tag{16}$$

$[y>0, x\in(-\infty,\infty)]$, the potential V_1 is

$$V_1(x,y)=y^2\left[\frac{\alpha^2}{x^2}+\beta^2(x^2+y^2)-\gamma^2\right] \tag{17}$$

and the Schrödinger equation has the following form:

$$-\frac{1}{2}y^2\left[\frac{\partial^2}{\partial x^2}-\frac{2\alpha^2}{x^2}-2\beta^2x^2+\frac{\partial^2}{\partial y^2}-2\beta^2y^2+2\gamma^2\right]\Psi(x,y)=E\Psi(x,y). \tag{18}$$

Via putting

$$\Psi(x,y)=\psi_1(x)\psi_2(y), \tag{19}$$

it admits a separation

$$\frac{d^2\psi_1}{dx^2}+2\left[\gamma^2(\lambda_1+1)-\beta^2x^2-\frac{\alpha^2}{x^2}\right]\psi_1=0, \tag{20}$$

$$\frac{d^2\psi_2}{dy^2}+2\left[\gamma^2(\lambda_2-1)-\beta^2y^2+\frac{E}{y^2}\right]\psi_2=0, \tag{21}$$

where λ_1 and λ_2 are the horicyclic separation constants with the relation $\lambda_1+\lambda_2=1$.

The orthonormalized solutions of the equations (20) and (21) for $(-2E+\frac{1}{4})>0$ are

$$\psi_1(x)\equiv\psi_{n_1}^{(\alpha,\beta)}(x)=\sqrt{\frac{n_1!(\sqrt{2}\beta)^{1/2}}{\Gamma(n_1+\sqrt{2\alpha^2+\frac{1}{4}}+1)}}e^{-\beta x^2/\sqrt{2}}(\sqrt{\sqrt{2}\beta x^2})^{1/2+\sqrt{2\alpha^2+1/4}}L_{n_1}^{\sqrt{2\alpha^2+1/4}}(\sqrt{2}\beta x^2), \tag{22}$$

$$\psi_2(y)\equiv\psi_{n_2}^{(\gamma,\beta)}(y)=\sqrt{\frac{n_2!(\sqrt{2}\beta)^{1/2}}{\Gamma(n_2+\sqrt{-2E+\frac{1}{4}}+1)}}e^{-\beta y^2/\sqrt{2}}(\sqrt{\sqrt{2}\beta y^2})^{1/2+\sqrt{-2E+1/4}}L_{n_2}^{\sqrt{-2E+1/4}}(\sqrt{2}\beta y^2). \tag{23}$$

The separation constants λ_1,λ_2 are quantized as

$$\lambda_1=\frac{\sqrt{2}\beta}{\gamma^2}\left(2n_1+\sqrt{2\alpha^2+\frac{1}{4}}+1\right)-1;\quad \lambda_2=\frac{\sqrt{2}\beta}{\gamma^2}\left(2n_2+\sqrt{-2E+\frac{1}{4}}+1\right)+1, \tag{24}$$

and according to the relation $\lambda_1+\lambda_2=1$, we come to the energy spectrum as in (13). The operator characterizing the separation in horicyclic coordinates is

$$\begin{aligned}L_2\Psi_{n_1n_2}(x,y)&\equiv\left[(K_2-M_1)^2-\frac{2\beta^2\omega_2^2}{(\omega_0-\omega_1)^2}-\frac{2\alpha^2(\omega_0-\omega_1)^2}{\omega_2^2}+2\gamma^2\right]\Psi_{n_1n_2}(x,y)\\&=-[2\sqrt{2}\beta(2n_1+\sqrt{2\alpha^2+\tfrac{1}{4}}+1)+2\gamma^2]\Psi_{n_1n_2}(x,y).\end{aligned} \tag{25}$$

3. Elliptic–parabolic coordinates

In this coordinate system,

$$\omega_0=\frac{\cosh^2 a+\cos^2\theta}{2\cosh a\cos\theta},\quad \omega_1=\frac{\sinh^2 a-\sin^2\theta}{2\cosh a\cos\theta},\quad \omega_2=\tanh a\tan\theta \tag{26}$$

$[a>0,\ \theta\in(-\pi/2,\pi/2)]$, the potential V_1 has the form

$$V_1(a,\theta)=\frac{\cosh^2 a\cos^2\theta}{\cosh^2 a-\cos^2\theta}\Bigg[\beta^2(\cosh^2 a\sinh^2 a+\cos^2\theta\sin^2\theta)-\gamma^2(\cosh^2 a-\cos^2\theta)+\alpha^2\left(\frac{1}{\sinh^2 a}+\frac{1}{\sin^2\theta}\right)\Bigg]. \tag{27}$$

The Schrödinger equation is

$$-\frac{1}{2}\,\frac{\cosh^2 a\cos^2\theta}{\cosh^2 a-\cos^2\theta}\Bigg[\frac{\partial^2}{\partial a^2}-2\beta^2\cosh^2 a\sinh^2 a+2\gamma^2\cosh^2 a-\frac{2\alpha^2}{\sinh^2 a}+\frac{\partial^2}{\partial\theta^2}-2\beta^2\cos^2\theta\sin^2\theta-2\gamma^2\cos^2\theta-\frac{2\alpha^2}{\sin^2\theta}\Bigg]\Psi(a,\theta)=E\Psi(a,\theta). \tag{28}$$

Putting for the wave function $\Psi(a,\theta)=S(a)S(\theta)$, after separation of variables we get two identical equations:

$$\frac{d^2S(\rho)}{d\rho^2}+\left[\lambda-2\beta^2\cosh^2\rho\sinh^2\rho+2\gamma^2\cosh^2\rho-\frac{2\alpha^2}{\sinh^2\rho}-\frac{2E}{\cosh^2\rho}\right]S(\rho)=0, \tag{29}$$

where λ is the elliptic–parabolic separation constant and $\rho\equiv a,i\theta$. After changing the variables $x=\cosh^2\rho$ in Eq. (29), we obtain

$$4x(x-1)\frac{d^2S}{dx^2}+2(2x-1)\frac{dS}{dx}+\left[\lambda-2\beta^2x(x-1)+2\gamma^2x-\frac{2\alpha^2}{x-1}-\frac{2E}{x}\right]S=0. \tag{30}$$

Thus the region $x\in[1,\infty]$ in Eq. (30) belongs to the wave function $S(a)$ and $x\in[0,1]$ to the wave function $S(\theta)$. Putting

$$S(x)=(x-1)^s x^t e^{-\beta x/\sqrt{2}}G(x), \tag{31}$$

where

$$s=\frac{1}{4}+\frac{1}{\sqrt{2}}\sqrt{\alpha^2+\frac{1}{8}},\qquad t=\frac{1}{4}+\frac{1}{\sqrt{2}}\sqrt{-E+\frac{1}{8}}, \tag{32}$$

we get

$$\frac{d^2G}{dx^2}+\frac{1}{2}\left[\frac{1+4t}{x}+\frac{1+4s}{x-1}-\frac{4\beta}{\sqrt{2}}\right]\frac{dG}{dx}+\frac{1}{4}\left\{\frac{[2\gamma^2-4\beta(1+2(t+s))/\sqrt{2}]x+\nu+\sqrt{2}\beta(1+4t)+4(t+s)^2}{x(x-1)}\right\}G=0. \tag{33}$$

If we now substitute

$$G(x)=\prod_{i=1}^{N}(x-\theta_i) \tag{34}$$

and take into account (32), we find that θ_i satisfies the equation

$$2\theta_i(1-\theta_i)\left(\sum_{\substack{k=1\\k\neq i}}^{N}\frac{1}{\theta_k-\theta_i}+\frac{\beta}{\sqrt{2}}\right)+2(1-\theta_i)N+\frac{\sqrt{2}\gamma^2}{4\beta}\theta_i+\frac{\gamma^2}{\sqrt{2}\beta}+\sqrt{2\alpha^2+\frac{1}{4}}+1=0. \quad (35)$$

The quantization for the energy is given via

$$\sqrt{-2E+\frac{1}{4}}+\sqrt{2\alpha^2+\frac{1}{4}}+2N+2-\frac{\gamma^2}{\sqrt{2}\beta}=0, \quad (36)$$

and we obtain the expression (13). The separation constant λ is

$$\lambda=\frac{8\beta}{\sqrt{2}}\sum_{i=1}^{N}\theta_i-\left(\frac{\gamma^2}{\sqrt{2}\beta}-1\right)^2+\frac{4\beta}{\sqrt{2}}\left(1+\sqrt{2\alpha^2+\frac{1}{4}}\right)-2\gamma^2. \quad (37)$$

Thus the total solution $\Psi(a,\theta)$ is represented as

$$\begin{aligned}\Psi_{Npq}(a,\theta)&=S_{Np}(a)S_{Nq}(\theta)\\&=(\sinh a\sin\theta)^{1/2+\sqrt{2\alpha^2+1/4}}(\cosh a\cos\theta)^{\gamma^2/\sqrt{2}\beta-\sqrt{2\alpha^2+1/4}-2N-3/2}\\&\quad\times\exp\left\{-\frac{\beta}{\sqrt{2}}(\cosh^2 a+\cos^2\theta)\right\}\prod_{i=1}^{N}(\cosh^2 a-\theta_i)(\cos^2\theta-\theta_i),\end{aligned} \quad (38)$$

where p and q are the number of zeros for the wave functions $S(a)$ and $S(\theta)$ in the regions $[0,1]$, $[1,\infty]$ correspondingly, and the total number of zeros is $N=p+q$.

Eliminating the energy E from Eq. (30), we see that the additional integral of motion here is

$$\begin{aligned}L_3\Psi_{Npq}(a,\theta)&=\frac{1}{\cos^2\theta-\cosh^2 a}\left\{\cosh^2 a\frac{\partial^2}{\partial a^2}+\cos^2\theta\frac{\partial^2}{\partial\theta^2}-2\beta^2(\cosh^4 a\sinh^2 a+\cos^4\theta\sin^2\theta)\right.\\&\qquad\left.+2\gamma^2(\cosh^4 a-\cos^4\theta)-2\alpha^2(\coth^2 a-\cot^2\theta)\right\}\Psi_{Npq}(a,\theta)\\&=\left\{-(K_2-M_1)^2-K_3^2+2\beta^2\frac{(w_0+w_1)^2+w_2^2}{(w_0-w_1)^2}+2\alpha^2\left(\frac{w_0-w_1}{w_2}\right)^2\right.\\&\qquad\left.-4\gamma^2\frac{w_0}{w_0-w_1}\right\}\Psi_{Npq}(a,\theta)\\&=\lambda\Psi_{Npq}(a,\theta).\end{aligned} \quad (39)$$

4. Hyperbolic–parabolic coordinates

In this coordinate system,

$$\omega_0=\frac{\cosh^2 b+\cos^2\theta}{2\sinh b\sin\theta},\quad \omega_1=\frac{\sinh^2 b-\sin^2\theta}{2\sinh b\sin\theta},\quad \omega_2=\coth b\cot\theta \quad (40)$$

$[b>0,\theta\in(-\pi/2,\pi/2)]$, the potential V_1 has the form

$$\begin{aligned}V_1(b,\theta)=\frac{\sinh^2 b\sin^2\theta}{\sinh^2 b+\sin^2\theta}&\left[\beta^2(\sinh^2 b\cosh^2 b+\sin^2\theta\cos^2\theta)\right.\\&\left.-\gamma^2(\sinh^2 b+\sin^2\theta)+\alpha^2\left(\frac{1}{\cos^2\theta}-\frac{1}{\cosh^2 b}\right)\right].\end{aligned} \quad (41)$$

The Schrödinger equation is

$$-\frac{1}{2}\frac{\sinh^2 b\sin^2\theta}{\sinh^2 b+\sin^2\theta}\left[\frac{\partial^2}{\partial b^2}-2\beta^2\sinh^2 b\cosh^2 b+2\gamma^2\sinh^2 b+\frac{2\alpha^2}{\cosh^2 b}\right.$$
$$\left.+\frac{\partial^2}{\partial\theta^2}-2\beta^2\sin^2\theta\cos^2\theta+2\gamma^2\sin^2\theta-\frac{2\alpha^2}{\cos^2\theta}\right]\Psi(b,\theta)=E\Psi(b,\theta). \quad (42)$$

Putting for the wave function $\Psi(b,\theta)=S(b)S(\theta)$, after separation of variables we get two identical equations:

$$\frac{d^2S(\rho)}{d\rho^2}+2\left[\frac{\tau}{2}-\beta^2\sinh^2\rho\cosh^2\rho+\gamma^2\sinh^2\rho+\frac{\alpha^2}{\cosh^2\rho}+\frac{E}{\sinh^2\rho}\right]S(\rho)=0, \quad (43)$$

where τ is the hyperbolic–parabolic separation constant and $\rho\equiv b,i\theta$. After changing the variables $x=\sinh^2\rho$ in Eq. (43), we come to the equation

$$4x(x+1)\frac{d^2S}{dx^2}+2(2x+1)\frac{dS}{dx}+\left[\tau-2\beta^2x(x+1)+2\gamma^2x+\frac{2\alpha^2}{x+1}+\frac{2E}{x}\right]S=0. \quad (44)$$

Choosing

$$P(x)=(1+x)^s x^t e^{-\beta x/\sqrt{2}}\prod_{i=1}^{N}(x-\theta_i), \quad (45)$$

where t and s are given by the formulas (32), we obtain the energy spectrum (36). Here θ_i satisfies the equations

$$2\theta_i(1+\theta_i)\left(\sum_{\substack{k=1\\k\neq i}}^{N}\frac{1}{\theta_i-\theta_k}-\frac{\beta}{\sqrt{2}}\right)-2(1+\theta_i)N+\frac{\sqrt{2}\gamma^2}{4\beta}\theta_i+\frac{\gamma^2}{\sqrt{2}\beta}-\sqrt{2\alpha^2+\frac{1}{4}}-1=0. \quad (46)$$

The separation constant τ is

$$\tau=\frac{8\beta}{\sqrt{2}}\sum_{i=1}^{N}\theta_i-\left(\frac{\gamma^2}{\sqrt{2}\beta}-1\right)^2-\frac{4\beta}{\sqrt{2}}\left(1+\sqrt{2\alpha^2+\frac{1}{4}}\right)+2\gamma^2, \quad (47)$$

so the total solution $\Psi(b,\theta)$ is represented as

$$\Psi_{Nlk}(b,\theta)=S_{Nl}(b)S_{Nk}(\theta)$$
$$=(\cosh b\cos\theta)^{1/2+\sqrt{2\alpha^2+1/4}}(\sinh b\sin\theta)^{\gamma^2/\sqrt{2}\beta-\sqrt{2\alpha^2+1/4}-2N-3/2}$$
$$\cdot\exp\left\{-\frac{\beta}{\sqrt{2}}(\sinh^2 b-\sin^2\theta)\right\}\prod_{i=1}^{N}(\sinh^2 b-\theta_i)(\sin^2\theta+\theta_i). \quad (48)$$

The total number of zeros is N, and k of them are located in the interval $[-1,0]$ and l are in $[0,\infty]$.

Each solution $\Psi_{Nlk}(b,\theta)$ satisfies the eigenvalue equation

$$
\begin{aligned}
L_4\Psi_{Nlk}(b,\theta) = & -\frac{1}{\sin^2 b+\sin^2\theta}\left\{\sinh^2 b\,\frac{\partial^2}{\partial b^2}-\sin^2\theta\,\frac{\partial^2}{\partial\theta^2}-2\beta^2(\cosh^2 b\,\sinh^4 b-\cos^2\theta\,\sin^4\theta)\right.\\
&\left.+2\gamma^2(\sinh^4 b-\sin^4\theta)+2\alpha^2(\tanh^2 b+\tan^2\theta)\right\}\Psi_{Nlk}(b,\theta)\\
=&\left\{(K_2-M_1)^2-K_3^2+2\beta^2\frac{(w_0+w_1)^2-w_2^2}{(w_0-w_1)^2}-2\alpha^2\left(\frac{w_0-w_1}{w_2}\right)^2\right.\\
&\left.-4\gamma^2\frac{w_1}{w_0-w_1}\right\}\Psi_{Nlk}(b,\theta)=\tau\Psi_{Nlk}(b,\theta).
\end{aligned}
\tag{49}
$$

B. Algebra

Among the operators $\{L_1,L_2,L_3,L_4\}$, corresponding to the four separable coordinate systems, only two are independent, as

$$
L_3=-L_2-L_1,\quad L_4=L_2-L_1. \tag{50}
$$

Consider the operators N_1, N_2, and R where

$$
\begin{aligned}
N_1=&\tilde{L}_2=L_1,\quad N_2=\tilde{L}_1=L_2-2\gamma^2,\\
R\equiv[N_1,N_2]=&2\{K_3,\{K_2,M_1\}\}-2\{K_3,K_2^2\}-2\{K_3,M_1^2\}\\
&+8\left[\alpha^2\left(\frac{\omega_0-\omega_1}{\omega_2}\right)^2+\beta^2\left(\frac{\omega_2}{\omega_0-\omega_1}\right)^2\right]K_3+\frac{16\beta^2\omega_2}{(\omega_0-\omega_1)^2}(\omega_0K_2-\omega_1M_1)\\
&+\frac{8\gamma^2\omega_2}{\omega_0-\omega_1}(M_1-K_2)-4\left[\gamma^2+2\alpha^2\left(\frac{\omega_0-\omega_1}{\omega_2}\right)^2-2\beta^2\frac{1+2\omega_2^2}{(\omega_0-\omega_1)^2}\right].
\end{aligned}
\tag{51}
$$

We have

$$
[R,N_2]=-8N_2^2-64\beta^2H-16\gamma^2N_2-32\beta^2N_1+16\beta^2(4\alpha^2-1), \tag{52}
$$

$$
[R,N_1]=4\{N_1,N_2\}+32\gamma^2H-16N_2+16\gamma^2N_1+16\gamma^2(2\alpha-1), \tag{53}
$$

$$
\begin{aligned}
R^2=&\tfrac{8}{3}\{N_2,N_2,L_1\}-\tfrac{176}{3}N_2^2+32\beta^2N_1^2+128\beta^2H^2+64\gamma^2HN_2+128\beta^2HN_1\\
&+16\gamma^2\{N_1,N_2\}+(\tfrac{128}{3}+256\alpha^2\beta^2)H+(64\alpha^2\gamma^2-\tfrac{352}{3}\gamma^2)N_2+(\tfrac{352}{3}-128\alpha^2\beta^2)N_1\\
&+(128\alpha^4\beta^2+128\gamma^4\alpha^2-\tfrac{128}{3}\alpha^2\beta^2-\tfrac{64}{3}\beta^2-48\gamma^2),
\end{aligned}
$$

where $\{A,B\}=AB+BA$ and

$$
\{A,B,C\}=ABC+ACB+BCA+BAC+CAB+CBA.
$$

The integrals of motion N_1, N_2, and H generate a quadratic algebra.

C. Interbasis expansion

For a fixed value of energy, we can write the equidistant wave function (7) in terms of the horicyclic ones (19) as

$$
\Psi_{n_1n_2}(x,y)=\sum_{m=0}^{n_1+n_2}W_{n_1n_2}^{nm}(\alpha,\beta,\gamma)\Psi_{nm}(a,b), \tag{54}
$$

where $n_1+n_2=n+m$. The connection between the equidistant (a,b) and horicyclic (x,y) coordinates is

$$x=e^b \tanh a, \quad y=e^b\frac{1}{\cosh a}. \tag{55}$$

Going over to the horicyclic coordinates on the left side of expansion (54), then considering the limit $b\to\infty$ and using the asymptotic formula for Laguerre polynomials[8]

$$\lim_{x\to\infty} L_n^{\alpha}(x)\to(-1)^n\frac{x^n}{n!}, \tag{56}$$

we see that dependence on b cancels on both sides of (54). Now, using the orthogonality condition for the angular wave functions (12), we find the following expression for the interbasis coefficients $W_{n_1n_2}^{nm}$:

$$W_{n_1n_2}^{nm}=(-1)^n\sqrt{\frac{m!n!\sqrt{2}\beta(\mu-d-2n-1)\Gamma(\mu+m+1)\Gamma(\mu-n)}{n_1!n_2!\mu\Gamma(n_1+d+1)\Gamma(n_2+d+1)\Gamma(n+d+1)\Gamma(\mu-d-n)}}B_{n_1n_2}^{nm}, \tag{57}$$

where

$$B_{n_1n_2}^{nm}=\int_{-\infty}^{+\infty}(\sinh a)^{1+2d+2n_1}(\cosh a)^{1-2\mu-2m}P_n^{(d,-\mu)}(\cosh 2a)da \tag{58}$$

and $d=\sqrt{2\alpha^2+\frac{1}{4}}$. The integral B_{n_1,n_2}^{nm} can be evaluated by expressing the Jacobi polynomial through the hypergeometric function ${}_2F_1$:[8]

$$P_n^{(\alpha,\beta)}(x)=(-1)^n\frac{\Gamma(n+\beta+1)}{\Gamma(\beta+1)n!}\,{}_2F_1\left(\begin{matrix}-n,n+\alpha+\beta+1\\ \beta+1\end{matrix}\middle|\frac{1+x}{2}\right). \tag{59}$$

Representing the function ${}_2F_1$ as a series we come to a sum of integrals, each of which can be calculated by using the formula[8]

$$\int_0^{+\infty}(\sinh\tau)^{\alpha}(\cosh\tau)^{-\beta}\,d\tau=\frac{1}{2}B\left(\frac{1+\alpha}{2},\frac{\beta-\alpha}{2}\right) \quad [\mathrm{Re}\,\alpha>-1,\mathrm{Re}(\alpha-\beta)<0]. \tag{60}$$

We thus obtain

$$W_{n_1n_2}^{nm}=\frac{(-1)^n}{2}\sqrt{\frac{m!\sqrt{2}\beta(\mu-d-2n-1)(\mu+m)\Gamma(n_1+d+1)}{n!n_1!n_2!\mu\Gamma(n_2+d+1)\Gamma(n+d+1)\Gamma(\mu-n-d)}}$$
$$\times\frac{\Gamma(\mu)\Gamma(\mu+m-d-n_1-1)}{\sqrt{\Gamma(\mu-n)\Gamma(\mu+m)}}\,{}_3F_2\left(\begin{matrix}-n,n+d-\mu+1,1-\mu-m\\ 1-\mu,2+n_1+d-\mu-m\end{matrix}\middle|1\right). \tag{61}$$

Alternatively, by using the formula[10] for the Hahn polynomials $h_n^{(\alpha,\beta)}(x,N)$,

$$h_n^{(\alpha,\beta)}(x,N)=\frac{(-1)^n\Gamma(N)\Gamma(\beta+n+1)}{n!\Gamma(N-n)\Gamma(\beta+1)}\,{}_3F_2\left(\begin{matrix}-n;\alpha+\beta+n+1;-x\\ \beta+1;1-N\end{matrix}\middle|1\right), \tag{62}$$

we obtain the following expression for the expansion coefficients:

$$W_{n_1 n_2}^{nm}=\frac{(-1)^n}{2}\sqrt{\frac{m!n!\sqrt{2}\beta(\mu-d-2n-1)(\mu+m)}{n_1!n_2!\mu\Gamma(n+d+1)\Gamma(\mu-n-d)}}\cdot\sqrt{\frac{\Gamma(n_1+d+1)\Gamma(\mu-n)}{\Gamma(n_2+d+1)\Gamma(\mu+m)}}$$
$$\times\Gamma(\mu+m-d-n_1-n-1)\cdot h_n^{(d,-\mu)}(\mu+m+1,\mu+m-d-n_1-1), \tag{63}$$

in terms of Hahn polynomials.

III. SECOND POTENTIAL

The second considered potential is

$$V_2=\frac{\alpha^2}{\omega_2^2}+\gamma^2\frac{\omega_0\omega_1}{(\omega_0^2+\omega_1^2)^2}+(\alpha^2-\beta^2)\frac{\omega_0^2-\omega_1^2}{(\omega_0^2+\omega_1^2)^2}, \tag{64}$$

where α, β, and γ are positive constants. The corresponding Schrödinger equation admits separable solutions in two coordinate systems: equidistant and semi-hyperbolic.

A. Solutions of the Schrödinger equation

1. Equidistant coordinates

In this coordinate system,

$$\omega_0=\cosh\tau_1\cosh\tau_2,\quad \omega_1=\cosh\tau_1\sinh\tau_2,\quad \omega_2=\sinh\tau_1 \tag{65}$$

$[\tau_1,\tau_2\in(-\infty,\infty)]$, the potential V_2 has the form

$$V_2(\tau_1,\tau_2)=-\frac{\alpha^2}{\sinh^2\tau_1}+\frac{1}{\cosh^2\tau_1}\frac{\alpha^2-\beta^2+\gamma^2\cosh\tau_2\sinh\tau_2}{(\cosh^2\tau_2+\sinh^2\tau_2)^2}. \tag{66}$$

After putting

$$\Psi(\tau_1,\tau_2)=(\cosh\tau_1)^{-1/2}Z(\tau_1)S(\tau_2), \tag{67}$$

we arrive at two equations:

$$\frac{d^2S}{d\tau_2^2}+\left[-\mu^2-\frac{2(\alpha^2-\beta^2)+\gamma^2\sinh(2\tau_2)}{\cosh^2(2\tau_2)}\right]S=0, \tag{68}$$

$$\frac{d^2Z}{d\tau_1^2}+\left[2E-\frac{1}{4}+\frac{\mu^2-\frac{1}{4}}{\cosh^2\tau_1}-\frac{2\alpha^2}{\sinh^2\tau_1}\right]Z=0, \tag{69}$$

where μ is the equidistant separation constant.

Let us consider the first equation (68). The substitution $x=\sinh 2\tau_2$ transforms this equation to

$$4(1+x^2)\frac{d^2S}{dx^2}+4x\frac{dS}{dx}+\left[-\mu^2+\frac{2(\beta^2-\alpha^2)-\gamma^2x}{(1+x^2)}\right]S=0, \tag{70}$$

where the physical region is $x\in(-\infty,\infty)$. The equation (68) has three regular singularities in the points $x=-i,i,\infty$ and may be solved in term of hypergeometric functions. The solution of the equation (68) for a large $|x|$ can be written as

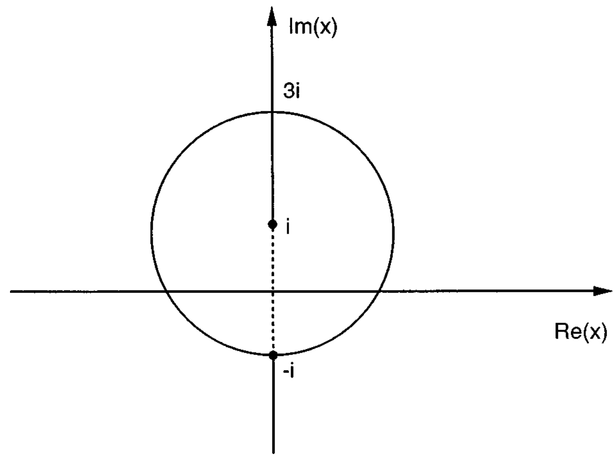

FIG. 1. Domain of convergence.

$$S(x)=A_1(x-i)^{-(b+\mu)/2-1/4}(x+i)^{b/2+1/4}\,{}_2F_1\left(\frac{a+b+1+\mu}{2},\frac{b-a+1+\mu}{2};1+\mu;\frac{2i}{i-x}\right)$$
$$+A_2(x-i)^{-(b-\mu)/2-1/4}(x+i)^{b/2+1/4}\,{}_2F_1\left(\frac{a+b+1-\mu}{2},\frac{b-a+1-\mu}{2};1-\mu;\frac{2i}{i-x}\right) \tag{71}$$

with

$$a^2=(b^2)^*=\frac{2\beta^2-2\alpha^2+1-i\gamma^2}{4}. \tag{72}$$

Let the separation constant μ be a positive number [the equation (71) is symmetric with respect to the replacement $\mu\to-\mu$]. Then the second term in formula (71) behaves like $|x|^{\mu/2}$ at ∞ and must be omitted. Thus for $S(x)$ we obtain

$$S(x)=A(x-i)^{-(b+\mu)/2-1/4}(x+i)^{b/2+1/4}\,{}_2F_1\left(\frac{a+b+1+\mu}{2},\frac{b-a+1+\mu}{2};\mu+1;\frac{2i}{i-x}\right). \tag{73}$$

The hypergeometric function in Eq. (73) converges if x lies out of the circle C on Fig. 1, defined by $|i-x|=2$, and converges on the circle C with the condition $\mathrm{Re}(b)<0$. The function $S(x)$ exists everywhere inside C except the interval $x\in[-i,i]$, since the hypergeometric function in (73) has a cut along the argument $2i/(i-x)\in[1,\infty)$. That means that the solution (73) along the real axes inside C in general is not a continuous function and may have a jump at the point $x=0$. Let us now consider the analytic continuation of (73) inside the circle C:

$$S(x)=A\Bigg\{(x-i)^{a/2+1/4}(x+i)^{b/2+1/4}\frac{\Gamma(\mu+1)\Gamma(-a)}{\Gamma((b-a+1+\mu)/2)\Gamma((-b-a+1+\mu)/2)(2i)^{(a+b+1+\mu)/2}}$$
$$\cdot{}_2F_1\left(\frac{a+b+1+\mu}{2},\frac{a+b+1-\mu}{2};a+1;\frac{i-x}{2i}\right)$$
$$+(x-i)^{-a/2+1/4}(x+i)^{b/2+1/4}\frac{\Gamma(\mu+1)\Gamma(a)}{\Gamma((b+a+1+\mu)/2)\Gamma((-b+a+1+\mu)/2)(2i)^{(-a+b+1+\mu)/2}}$$
$$\cdot{}_2F_1\left(\frac{-a+b+1+\mu}{2},\frac{-a+b+1-\mu}{2};-a+1;\frac{i-x}{2i}\right)\Bigg\}. \tag{74}$$

From Eq. (72) follow two possibilities

$$a=b^*,\quad a=-b^*. \tag{75}$$

Putting the $a=b^*$ [Re(a)=Re(b)<0] we find that the first term in (74) represents an analytic function, while the second term is discontinuous at $x=0$. [Note since the both terms in Eq. (74) transform to each other with replacement $a\to-a$, the choice $a=-b^*$ means that the first term in (74) is discontinuous while the second term is continuous at $x=0$.] Thus the *sufficient* condition for the existence of the continuous solution requires the relation

$$\mu+a+a^*+1=-2m,\quad m=0,1,2,...,\left[-\frac{a+a^*+1}{2}\right], \tag{76}$$

so from (72) we have

$$\mu=-2m-1+\frac{1}{\sqrt{2}}\sqrt{2\beta^2-2\alpha^2+1+\sqrt{(2\beta^2-2\alpha^2+1)^2+\gamma^4}}. \tag{77}$$

Finally, the orthonormalized eigenfunction of Eq. (68) may be written in the form

$$S(\tau_2)=(-1)^{3m/2}\Gamma(-a)\sqrt{\frac{(-2m-a-a^*-1)\Gamma(-m-a^*)}{\pi m!2^{a+a^*+1}\Gamma(-m-a)\Gamma(-m-a-a^*)}}$$
$$\cdot(1+i\sinh 2\tau_2)^{a/2+1/4}(1-i\sinh 2\tau_2)^{a^*/2+1/4}$$
$$\cdot{}_2F_1\left(-m,m+a+a^*+1;a+1;\frac{1+i\sinh 2\tau_2}{2}\right)$$
$$=(-1)^{m/2}\sqrt{\frac{(-2m-a-a^*-1)m!\Gamma(-m-a)\Gamma(-m-a^*)}{\pi 2^{a+a^*+1}\Gamma(-m-a-a^*)}}$$
$$\cdot(1+i\sinh 2\tau_2)^{a/2+1/4}(1-i\sinh 2\tau_2)^{a^*/2+1/4}P_m^{(a,a^*)}(-i\sinh 2\tau_2), \tag{78}$$

where

$$a=\frac{1}{2^{3/2}}\{-\sqrt{\sqrt{(2\beta^2-2\alpha^2+1)^2+\gamma^4}+2\beta^2-2\alpha^2+1}$$
$$+i\sqrt{\sqrt{(2\beta^2-2\alpha^2+1)^2+\gamma^4}-(2\beta^2-2\alpha^2+1)}\}.$$

The second equation (69) is quite like (9) and has a solution

$$Z(\tau_1)\equiv S_n^{(\alpha,\mu)}(\tau_1)=\sqrt{\frac{2(\mu-\sqrt{2\alpha^2+\frac{1}{4}}-2n-1)\Gamma(\mu-n)n!}{\Gamma(\mu-\sqrt{2\alpha^2+\frac{1}{4}}-n)\Gamma(1+n+\sqrt{2\alpha^2+\frac{1}{4}})}}$$
$$\times(\sinh\tau_1)^{1/2+\sqrt{2\alpha^2+1/4}}(\cosh\tau_1)^{1/2-\mu}P_n^{(\alpha,-\mu)}(\cosh 2\tau_1) \tag{79}$$

with $n=0,1,\ldots$.

The quantized energy is

$$E=-\frac{1}{2}(\mu-\sqrt{2\alpha^2+1/4}-2n-1)^2+\frac{1}{8}$$
$$=-\frac{1}{2}\left\{2N+2+\sqrt{2\alpha^2+\tfrac{1}{4}}-\frac{1}{\sqrt{2}}\sqrt{2\beta^2-2\alpha^2+1\sqrt{(2\beta^2-2\alpha^2+1)^2+\gamma^4}}\right\}^2+\frac{1}{8}, \tag{80}$$

where $N=n+m$ is the principal quantum number and the bound state occurs for

$$0\leqslant N\leqslant\left[\frac{1}{\sqrt{8}}\sqrt{2\beta^2-2\alpha^2+1+\sqrt{(2\beta^2-2\alpha^2+1)^2+\gamma^4}}-\frac{1}{2}\sqrt{2\alpha^2+\tfrac{1}{4}}-1\right]. \tag{81}$$

The additional operator describing this coordinate system is

$$L_1\Psi_{nm}(\tau_1,\tau_2)\equiv\left[K_3^2-2(\alpha^2-\beta^2)\left(\frac{\omega_0^2-\omega_1^2}{\omega_0^2+\omega_1^2}\right)^2-2\gamma^2\frac{\omega_0\omega_1(\omega_0^2-\omega_1^2)}{(\omega_0^2+\omega_1^2)^2}\right]\Psi_{nm}(\tau_1,\tau_2)$$
$$=\left\{2m+1-\frac{1}{\sqrt{2}}\sqrt{2\beta^2-2\alpha^2+1+\sqrt{(2\beta^2-2\alpha^2+1)^2+\gamma^4}}\right\}^2\Psi_{nm}(\tau_1,\tau_2). \tag{82}$$

2. Semi-hyperbolic coordinates

Here

$$\omega_0^2=-\frac{(\mu-e_3)(\nu-e_3)}{2[(e_3-a)^2+b^2]}+\frac{1}{2}-\frac{1}{2b}\left[\frac{[(\mu-a)^2+b^2][(\nu-a)^2+b^2]}{(e_3-a)^2+b^2}\right]^{1/2},$$
$$\omega_1^2=\frac{(\mu-e_3)(\nu-e_3)}{2[(e_3-a)^2+b^2]}-\frac{1}{2}-\frac{1}{2b}\left[\frac{[(\mu-a)^2+b^2][(\nu-a)^2+b^2]}{(e_3-a)^2+b^2}\right]^{1/2}, \tag{83}$$
$$\omega_2^2=-\frac{(\mu-e_3)(\nu-e_3)}{(e_3-a)^2+b^2}$$

$[\nu<e_3<\mu]$, where $\sinh 2f=(e_3-a)/b$ and $2f$ is the distance between the focii of the semi-hyperbolas and the bases of their equidistants.[6]

If we change variables according to

$$\omega_0=\frac{(s_1+s_2)}{\sqrt{2}},\quad \omega_1=\frac{-i(s_1-s_2)}{\sqrt{2}},\quad \omega_2=-is_3, \tag{84}$$

the Schrödinger equation becomes

$$\frac{1}{2}\left[\left(s_1\frac{\partial}{\partial s_2}-s_2\frac{\partial}{\partial s_1}\right)^2+\left(s_1\frac{\partial}{\partial s_3}-s_3\frac{\partial}{\partial s_1}\right)^2+\left(s_3\frac{\partial}{\partial s_2}-s_2\frac{\partial}{\partial s_3}\right)^2\right]\Psi$$
$$+\left[-E-\frac{1}{2}\left(\frac{k_1^2-\frac{1}{4}}{s_1^2}+\frac{k_2^2-\frac{1}{4}}{s_2^2}+\frac{k_3^2-\frac{1}{4}}{s_3^2}\right)\right]\Psi=0 \tag{85}$$

with

$$\frac{1}{2}\left(k_1^2-\frac{1}{4}\right)=\frac{1}{4}(\beta^2-\alpha^2)-\frac{i}{8}\gamma^2,\quad \frac{1}{2}\left(k_2^2-\frac{1}{4}\right)=\frac{1}{4}(\beta^2-\alpha^2)+\frac{i}{8}\gamma^2,\quad \frac{1}{2}\left(k_3^2-\frac{1}{4}\right)=\alpha^2.$$

Noting

$$\omega_0^2-\omega_1^2-\omega_2^2=s_1^2+s_2^2+s_3^2=1$$

and considering Eq. (85), we see that the problem we wish to solve using the real coordinates ω_0, ω_1, and ω_2 is a real case of the corresponding problem on the sphere with coordinates s_1,s_2,s_3 and energy $\varepsilon=-E$.

Inverting the relations (84) we have

$$s_1=\frac{(\omega_0+i\omega_1)}{\sqrt{2}},\quad s_2=\frac{(\omega_0-i\omega_1)}{\sqrt{2}},\quad s_3=i\omega_2.$$

Now choose elliptic coordinates on the complex sphere according to

$$s_1^2=\frac{(\mu-e_1)(\nu-e_1)}{(e_1-e_2)(e_1-e_3)},\quad s_2^2=\frac{(\mu-e_2)(\nu-e_2)}{(e_2-e_1)(e_2-e_3)},\quad s_3^2=\frac{(\mu-e_3)(\nu-e_3)}{(e_3-e_2)(e_3-e_1)}.$$

This choice of real coordinates μ,ν will work for the real coordinates ω_k, $k=0,1,2$, if we take $e_1=e_2^*=a+ib,a,b$ real and $\nu<e_3<\mu$.

In terms of the coordinates μ and ν the Schrödinger equation has the form

$$\frac{4}{(\mu-\nu)}\left\{(\mu-e_2^*)(\mu-e_2)(\mu-e_3)\left[\frac{\partial^2\Psi}{\partial\mu^2}+\frac{1}{2}\left(\frac{1}{\mu-e_2^*}+\frac{1}{\mu-e_2}+\frac{1}{\mu-e_3}\right)\frac{\partial\Psi}{\partial\mu}\right]\right.$$
$$\left.-(\nu-e_2^*)(\nu-e_2)(\nu-e_3)\left[\frac{\partial^2\Psi}{\partial\nu^2}-\frac{1}{2}\left(\frac{1}{\nu-e_2^*}+\frac{1}{\nu-e_2}+\frac{1}{\nu-e_3}\right)\frac{\partial\Psi}{\partial\nu}\right]\right\}$$
$$+\left[\left(k_1^2-\frac{1}{4}\right)\frac{(e_2^*-e_2)(e_2^*-e_3)}{(\mu-e_2^*)(\nu-e_2^*)}+\left(k_2^2-\frac{1}{4}\right)\frac{(e_2-e_2^*)(e_2-e_3)}{(\mu-e_2)(\nu-e_2)}\right.$$
$$\left.+\left(k_3^2-\frac{1}{4}\right)\frac{(e_3-e_2)(e_3-e_2^*)}{(\mu-e_3)(\nu-e_3)}\right]\Psi=-2E\Psi. \tag{86}$$

The separation equations are

$$(\rho-e_2^*)(\rho-e_2)(\rho-e_3)\left[\frac{d^2\Psi}{d\rho^2}+\frac{1}{2}\left(\frac{1}{\rho-e_2^*}+\frac{1}{\rho-e_2}+\frac{1}{\rho-e_3}\right)\frac{d\Psi}{d\rho}\right]$$
$$-\frac{1}{4}\left[\left(k_1^2-\frac{1}{4}\right)\frac{(e_2^*-e_2)(e_2^*-e_3)}{(\rho-e_2^*)}+\left(k_2^2-\frac{1}{4}\right)\frac{(e_2-e_2^*)(e_2-e_3)}{(\rho-e_2)}\right.$$
$$\left.+\left(k_3^2-\frac{1}{4}\right)\frac{(e_3-e_2)(e_3-e_2^*)}{(\rho-e_3)}-2E\rho+\lambda\right]\psi(\rho)=0, \tag{87}$$

where $\rho=\mu,\nu$. The operator L_2 with eigenvalue λ is

$$L_2\Psi=\frac{-4}{(\mu-\nu)}\left\{\nu(\mu-e_1)(\mu-e_2)(\mu-e_3)\left[\frac{\partial^2\Psi}{\partial\mu^2}+\frac{1}{2}\left(\frac{1}{\mu-e_2^*}+\frac{1}{\mu-e_2}+\frac{1}{\mu-e_3}\right)\frac{\partial\Psi}{\partial\mu}\right]\right.$$
$$\left.-\mu\left[(\nu-e_1)(\nu-e_2)(\nu-e_3)\left[\frac{\partial^2\Psi}{\partial\nu^2}+\frac{1}{2}\left(\frac{1}{\nu-e_2^*}+\frac{1}{\nu-e_2}+\frac{1}{\nu-e_3}\right)\frac{\partial\Psi}{\partial\nu}\right]\right]\right\}$$
$$-\left[\left(k_1^2-\frac{1}{4}\right)\frac{(e_2^*-e_2)(e_2^*-e_3)}{(\mu-e_2^*)(\nu-e_2^*)}(\mu+\nu-e_2^*)+\left(k_2^2-\frac{1}{4}\right)\frac{(e_2-e_2^*)(e_2-e_3)}{(\mu-e_2)(\nu-e_2)}(\mu+\nu-e_2)\right.$$
$$\left.+\left(k_3^2-\frac{1}{4}\right)\frac{(e_3-e_2)(e_3-e_2^*)}{(\mu-e_3)(\nu-e_3)}(\mu+\nu-e_3)\right]\Psi. \tag{88}$$

In order to find the bound state solutions of this system in semi-hyperbolic coordinates we first observe the identity

$$\frac{s_1^2}{\theta_j-e_2^*}+\frac{s_2^2}{\theta_j-e_2}+\frac{s_3^2}{\theta_j-e_3}=\frac{(\omega_0^2-\omega_1^2)(\theta_j-a)-2\omega_0\omega_1 b}{(\theta_j-a)^2+b^2}-\frac{\omega_2^2}{\theta_j-e_3}$$
$$=\frac{(\mu-\theta_j)(\nu-\theta_j)}{(\theta_j-e_2^*)(\theta_j-e_2)(\theta_j-e_3)}. \tag{89}$$

If we then look for solutions of the form

$$\Psi=\prod_{\ell=1}^{3}s_\ell^{k_\ell+1/2}\prod_{j=1}^{N}\left(\frac{s_1^2}{\theta_j-e_2^*}+\frac{s_2^2}{\theta_j-e_2}+\frac{s_3^2}{\theta_j-e_3}\right), \tag{90}$$

we see that the corresponding zeros satisfy the equations

$$\frac{k_1+1}{\theta_m-e_2^*}+\frac{k_2+1}{\theta_m-e_2}+\frac{k_3+1}{\theta_m-e_3}+\sum_{j\neq m}^{N}\frac{2}{(\theta_m-\theta_j)}=0. \tag{91}$$

For the energy E we have

$$E=-\tfrac{1}{2}(2N+2+k_1+k_2+k_3)^2+\tfrac{1}{8}, \tag{92}$$

which coincides with the formula (73) [note (86)]. For the separation constant λ we obtain

$$\lambda=-2[k_1(e_2+e_3)+k_2(e_2^*+e_3)+k_3(e_2+e_2^*)]-2[e_3k_1k_2+e_2k_1k_3+e_2^*k_2k_3]$$
$$-\frac{3}{2}(e_2^*+e_2+e_3)-4e_2e_3(k_1+1)\sum_{m=1}^{q}\frac{1}{(\theta_m-e_2^*)}-e_2^*e_3(k_2+1)\sum_{m=1}^{q}\frac{1}{(\theta_m-e_2)}$$
$$-4e_2e_2^*(k_3+1)\sum_{m=1}^{q}\frac{1}{(\theta_m-e_3)}. \tag{93}$$

In terms of variables w_i the total wave function is written

$$\Psi=\left(\frac{\omega_0+i\omega_1}{\sqrt{2}}\right)^{k_1+1/2}\left(\frac{\omega_0-i\omega_1}{\sqrt{2}}\right)^{k_2+1/2}(i\omega_2)^{k_3+1/2}\prod_{j=1}^{N}\left[\frac{(\omega_0^2-\omega_1^2)(\theta_j-a)-2\omega_0\omega_1 b}{(\theta_j-a)^2+b^2}-\frac{\omega_2^2}{\theta_j-e_3}\right].$$

The algebra of second-order symmetries for this potential is generated by the operators

$$L_{jk}=(s_j\partial_{s_k}-s_k\partial_{s_j})^2+\left(\frac{1}{4}-k_j^2\right)\frac{s_k^2}{s_j^2}+\left(\frac{1}{4}-k_k^2\right)\frac{s_j^2}{s_k^2} \tag{94}$$

for $j,k=1,2,3$ and $j\neq k$. The Hamiltonian of the system is expressed in terms of L_{jk} as

$$H=\frac{1}{2}(L_{12}+L_{13}+L_{23})-\frac{1}{2}\sum_{i=1}^{3} k_i^2+\frac{3}{4}. \tag{95}$$

The relevant generators in the real case we are considering are then

$$L_{12}=-K_3^2+\left(\frac{1}{4}-k_1^2\right)\left(\frac{\omega_0-i\omega_1}{\omega_0+i\omega_1}\right)^2+\left(\frac{1}{4}-k_2^2\right)\left(\frac{\omega_0+i\omega_1}{\omega_0-i\omega_1}\right)^2, \tag{96}$$

$$L_{13}=\frac{1}{2}(M_1-iK_2)^2+\left(\beta^2-\alpha^2-\frac{i}{2}\gamma^2\right)\frac{\omega_2^2}{(\omega_0+i\omega_1)^2}+\alpha^2\frac{(\omega_0+i\omega_1)^2}{\omega_2^2}, \tag{97}$$

$$L_{23}=\frac{1}{2}(M_1+iK_2)^2+\left(\beta^2-\alpha^2+\frac{i}{2}\gamma^2\right)\frac{\omega_2^2}{(\omega_0-i\omega_1)^2}+\alpha^2\frac{(\omega_0-i\omega_1)^2}{\omega_2^2}. \tag{98}$$

The commutation relations and resulting quadratic algebra can then be deduced from the relations for the complex forms in terms of the L_{ij}. It is easy to show that the additional integrals of motion, corresponding to the separation in equidistant and semi-hyperbolic coordinates, can be written as

$$L_1=-L_{12}+\beta^2-\alpha^2 \tag{99}$$

and

$$L_2=e_3L_{12}+e_2L_{13}+e_1L_{32}-k_1^2(e_2+e_3-e_1)-k_2^2(e_1+e_3-e_2)-k_3^2(e_1+e_2-e_3)+\tfrac{1}{4}(e_1+e_2+e_3). \tag{100}$$

The algebra for the operators (99) and (100) is found in Ref. 11.

ACKNOWLEDGMENTS

The authors thank Professor V. M. Ter-Antonyan and Professor P. Winternitz for interesting discussions.

[1] C. P. Boyer, E. G. Kalnins, and P. Winternitz, ''Completely integrable relativistic Hamiltonian systems and separation of variables in Hermitian hyperbolic spaces,'' J. Math. Phys. **24**, 2022 (1983).
[2] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, ''Path integral approach to superintegrable potentials. The two-dimensional hyperboloid,'' Phys. Part. Nuclei **27**, 244 (1996).
[3] E. G. Kalnins, W. Miller, Jr., and G. S. Pogosyan, ''Superintegrability on the two dimensional hyperboloid,'' J. Math. Phys. **38**, 5416 (1997).
[4] P. Winternitz, I. Lukac, and Ya. A. Smorodinskii, ''Quantum numbers in the little groups of the Pioncaré group,'' Sov. J. Nucl. Phys. **7**, 139 (1968).
[5] E. G. Kalnins and W. Miller, Jr., ''Lie theory and separation of variables, 4: The groups SO(2,1) and SO(3),'' J. Math. Phys. **15**, 1263 (1974).
[6] M. N. Olevskiĭ, ''Triorthogonal systems in spaces of constant curvature in which the equation $\Delta_2 u+\lambda u=0$ allows the complete separation of variables,'' Math. Sb. **27**, 379 (1950) (in Russian).
[7] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory* (Pergamon, Oxford, 1977).
[8] *Higher Transcendental Functions* edited by A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi (McGraw–Hill, New York, 1953), Vols. I and II.
[9] A. Frank and K. B. Wolf, ''Lie Algebras for systems with mixed spectra. I. The scattering Pöschl–Teller potential,'' J. Math. Phys. **26**, 973 (1985).
[10] A. F. Nikiforov, S. K. Suslov, and V. B. Uvarov, *Classical Orthogonal Polynomials of Discrete Variables* (Nauka, Leningrad, 1985).
[11] E. G. Kalnins, W. Miller, Jr., and G. S. Pogosyan, ''Superintegrability and associated polynomial solutions. Euclidean space and the sphere in two dimensions,'' J. Math. Phys. **37**, 6439 (1996).

On the Coulomb Sturmian matrix elements of relativistic Coulomb Green's operators

B. Kónya and Z. Papp[a)]
Institute of Nuclear Research of the Hungarian Academy of Sciences, P. O. Box 51, H–4001 Debrecen, Hungary

(Received 28 October 1998; accepted for publication 16 November 1998)

The Hamiltonian of the radial Coulomb Klein–Gordon and second order Dirac equations are shown to possess an infinite symmetric tridiagonal matrix structure on the relativistic Coulomb Sturmian basis. This allows us to give an analytic representation for the corresponding Coulomb Green's operators in terms of continued fractions. The poles of the Green's matrix reproduce the exact relativistic hydrogen spectrum. © *1999 American Institute of Physics.*
[S0022-2488(99)02104-0]

I. INTRODUCTION

In quantum mechanics the knowledge of the Green's operator is equivalent to the complete knowledge of the system. So, having an analytic basis representation for the Green's operator can tremendously simplify the actual calculations. If we know the Green's operator only of the asymptotic part of the Hamiltonian we can treat the remaining terms as perturbations and approximate them by finite matrices.

In a recent publication, Ref. 1, we have proposed a method for calculating matrix representation of Green's operators. If, in some basis representation, the Hamiltonian possesses an infinite symmetric tridiagonal (Jacobi) matrix structure the corresponding Green's operator can be given in terms of continued fractions. In Ref. 1, this theorem was exemplified with the Green's operator of the nonrelativistic Coulomb and harmonic oscillator Hamiltonian, and, in Ref. 2, an exactly solvable nonrelativistic potential problem was considered which provides a smooth transition between the Coulomb and the harmonic oscillator problems.

Our aim in this paper is to extend this result for relativistic Coulomb Green's operators, i.e., for the Coulomb Green's operator of the Klein–Gordon and of the second order Dirac equations. This later is equivalent to the conventional Dirac equation and seems to have several advantages. For details see Ref. 3 and references therein. The Coulomb Sturmian matrix elements of the second order Dirac equation has already been obtained by Hostler[3] via evaluating complicated contour integrals. Our derivation, however, is much simpler, it relies only on the Jacobi-matrix structure of the Hamiltonian, and the result obtained is also better suited for numerical calculations. In Ref. 3 the result appears in terms of Γ and hypergeometric functions, while our procedure results in an easily computable and analytically continuable continued fraction.

II. MATRIX ELEMENTS OF RELATIVISTIC COULOMB–GREEN'S OPERATORS

The radial Klein–Gordon and second order Dirac equations for Coulomb interaction are given by

$$H_u|\xi^u\rangle = 0, \tag{1}$$

where

[a)]Electronic mail: pz@indigo.atomki.hu

$$H_u = \left(\frac{E}{\hbar c}\right)^2 - \mu^2 + \frac{2\alpha Z}{\hbar c}\frac{E}{r} + \frac{d^2}{dr^2} - \frac{u(u+1)}{r^2}. \tag{2}$$

Here $\mu = mc/\hbar$, $\alpha = e^2/\hbar c$, m is the mass and Z denotes the charge. For the Klein–Gordon case u is given by

$$u = -\tfrac{1}{2} + \sqrt{\tfrac{1}{4} + l(l+1) - (Z\alpha)^2}, \tag{3}$$

and in the case of the second order Dirac equation for the different spin states we have

$$u_{\pm} = -\tfrac{1}{2} \mp \tfrac{1}{2} + \sqrt{(j+\tfrac{1}{2})^2 - (Z\alpha)^2}. \tag{4}$$

The relativistic Coulomb Green's operator is defined as the inverse of the Hamiltonian H_u:

$$H_u G_u = G_u H_u = \mathbf{1}_u, \tag{5}$$

where $\mathbf{1}_u$ denotes the unit operator of the radial Hilbert space $\mathcal{H}_u$.

In complete analogy with the nonrelativistic case we can define the relativistic Coulomb Sturmian functions as solutions of the Sturm–Liouville problem,

$$\left(-\frac{d^2}{dr^2} + \eta^2 + \frac{u(u+1)}{r^2} - \frac{2\eta(n+u+1)}{r}\right) S^u_{n;\eta}(r) = 0, \tag{6}$$

where η is a real parameter and $n = 0,1,2,\ldots,\infty$ is the radial quantum number. In coordinate space representation they take the form

$$\langle r|nu;\eta\rangle = \left[\frac{n!}{(n+2u+1)!}\right]^{1/2} (2\eta r)^{u+1} e^{-\eta r} L_n^{2u+1}(2\eta r), \tag{7}$$

where L is a Laguerre-polinom. The Coulomb Sturmian functions, together with their biorthogonal partner $\langle r|\widetilde{nu;\eta}\rangle = 1/r \cdot \langle r|nu;\eta\rangle$, form a basis: i.e., they are orthogonal,

$$\langle \widetilde{nu;\eta}|mu;\eta = \langle nu;\eta|\widetilde{mu;\eta}\rangle = \delta_{nm}, \tag{8}$$

and form a complete set in $\mathcal{H}_u$,

$$\sum_{n=0}^{\infty} |nu;\eta\rangle\langle\widetilde{nu;\eta}| = \sum_{n=0}^{\infty} |nu;\eta\rangle\langle\widetilde{nu;\eta}| = \mathbf{1}_u. \tag{9}$$

A straightforward calculation yields

$$\langle nu;\eta|mu;\eta\rangle = \frac{1}{2\eta}[\delta_{nm}(2u+2n+2) - \delta_{nm-1}\sqrt{(n+1)(n+2u+2)} - \delta_{nm+1}\sqrt{n(2u+n+1)}]. \tag{10}$$

Utilizing this relation and considering Eq. (6) we can easily calculate the Coulomb Sturmian matrix elements of H_u,

TABLE I. Energy levels of hydrogen-like atoms in atomic units. E_{cf} is the relativistic spectrum calculated via a continued fraction, E_D and E_S are textbook values of the relativistic Dirac and the nonrelativistic Schrödinger spectrum, respectively.

	Energy levels	E_{cf}	E_{D}	E_{S}
Hydrogen	$1S_{1/2}$	−0.5000066521	−0.5000066521	−0.5
$Z=1$	$2P_{1/2}$	−0.1250020801	−0.1250020801	−0.125
	$2P_{3/2}$	−0.1250004160	−0.1250004160	−0.125
	$50P_{1/2}$	−0.0002000002	−0.0002000002	−0.0002
	$50P_{3/2}$	−0.0002000001	−0.0002000001	−0.0002
Uranium	$1S_{1/2}$	−4861.1483347	−4861.1483347	−4232
$Z=92$	$100D_{3/2}$	−0.4241695002	−0.4241695002	−0.4232
	$100D_{5/2}$	−0.4238303306	−0.4238303306	−0.4232

$$\underline{H}_{nm}:=\langle nu;\eta|H_u|mu;\eta\rangle=+\delta_{nm}\left(\frac{2\alpha zE}{\hbar c}-2(u+n+1)\eta+2(u+n+1)\frac{(E/\hbar c)^2-\mu^2+\eta^2}{2\eta}\right)$$
$$-\delta_{nm-1}\left(\frac{(E/\hbar c)^2-\mu^2+\eta^2}{2\eta}\sqrt{(n+1)(n+2u+2)}\right)$$
$$-\delta_{nm+1}\left(\frac{(E/\hbar c)^2-\mu^2+\eta^2}{2\eta}\sqrt{n(n+2u+1)}\right), \tag{11}$$

which happens to possess a Jacobi-matrix structure. So, the theorem of Ref. 1 is readily applicable here.

Let us consider the $\infty\times\infty$ Green's matrix,

$$(\underline{G}_u)_{nm}\equiv\langle\widetilde{nu;\eta}|G_u|\widetilde{mu;\eta}\rangle, \tag{12}$$

and let us denote its rank-N leading principal submatrix by $\underline{G}_u^{(N)}$. Then, according to Ref. 1,

$$(\underline{G}_u^{(N)})_{ij}^{-1}=\underline{H}_{ij}+\delta_{jN}\delta_{iN}\underline{H}_{NN+1}\mathcal{F}, \tag{13}$$

where $\mathcal{F}$ is a continued fraction,

$$\mathcal{F}=-K_{i=N}^{\infty}\left(\frac{a_i}{b_i}\right)=-\frac{a_{1+N}}{b_{1+N}+}\frac{a_{2+N}}{b_{2+N}+}\cdots+\frac{a_{n+N}}{b_{n+N}+}\cdots, \tag{14}$$

whose coefficients are related to the Jacobi matrix

$$a_i=-\frac{\underline{H}_{ii-1}}{\underline{H}_{ii+1}},\quad b_i=-\frac{\underline{H}_{ii}}{\underline{H}_{ii+1}}. \tag{15}$$

This continued fraction convergent for bound-state energies, but, by using the method presented in Ref. 1, can be continued analytically to the whole complex energy plane. Simple matrix inversion now gives the desired Green's matrix.

In Table I we demonstrate the numerical accuracy of the method by evaluating the ground and some highly excited sates of relativistic hydrogen-like atoms, which, in fact, correspond to the poles of the Dirac Coulomb Green's matrix. In particular, the zeros of the determinant of (13) were located. It should be noted that irrespective of the rank N the zeros should provide the exact Dirac results. In Table I we have taken 2×2 matrices. Indeed, the results of this method, E_{cf}, agree with the exact one in all cases, practically up to the machine accuracy, allowing thus to study the fine structure splitting.

III. SUMMARY

In this paper we have presented a practical and easy-to-apply procedure for calculating the Coulomb Sturmian matrix elements of the Coulomb Green's operator of the Klein–Gordon and of the second order Dirac equations. The method is relied only on the Jacobi-matrix structure of the corresponding Hamiltonians and results in a continued fraction which can be continued analytically to the whole complex energy plane.

ACKNOWLEDGMENTS

This work has been supported by the OTKA Contracts No. T17298 and T026233.

[1] B. Kónya, G. Lévai, and Z. Papp, J. Math. Phys. **38**, 4832 (1997).
[2] G. Lévai, B. Kónya, and Z. Papp, J. Math. Phys. **39**, 5811 (1998).
[3] L. C. Hostler, J. Math. Phys. **28**, 2984 (1987).

Quantum geometry of field extensions

Shahn Majid[a)]
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 9EW, United Kingdom

(Received 9 October 1998; accepted for publication 20 October 1998)

We introduce a new kind of topological gauge configuration or ''soliton'' associated to the extension $\mathbb{C}\supseteq\mathbb{R}$ of the real numbers to the complex ones. These configurations describe zero-curvature gauge fields and nontrivial cohomology over the real line, but with a quantum choice of differential calculus. In general, the quantum differential 1-forms on the line with coordinate algebra $k[x]$ are in correspondence with field extensions of k, and the quantum cohomology detects the nontriviality of the extension.
[S0022-2488(99)04002-5]

I. INTRODUCTION

Usually quantum differential geometry aims to model or q-deform classical geometrical constructions, e.g., Ref. 1 as a step towards a more general geometry appropriate to quantum integrable systems and (in principle) to the Planck scale. In this paper we observe that in quantum differential geometry one also has the possibility of entirely new kinds of topological gauge configurations or ''solitons'' not even visible classically. We demonstrate the ideas explicitly for the real line, which we work with algebraically as the coordinate algebra of real polynomials $\mathbb{R}[x]$ in one variable. The existence of these topological configurations implies, among other things, a novel quantum geometrical approach to number theory that we have not seen considered elsewhere. From a physical point of view, the extension from real numbers to complex ones is particularly central to physics and the introduction of a geometrical way of looking at it is a unification that could certainly be significant at the Planck scale.

First, note that the differential calculus on a space is additional information to the space itself, but for Lie groups there is a unique translation-invariant calculus which is usually assumed. In quantum differential geometry we drop the assumption that the product of coordinate functions and differential 1-forms from the left and right coincide, i.e., in 1-variable we allow $x\,dx \neq (dx)x$. This is the natural situation, for example, on a lattice, where multiplication from the left and right differ due to the difference between the starting-point and end-point of a finite-difference differential. On the other hand, the choice of differential calculus is then no longer unique, even for covariant ones[2] on a classical (or quantum) group. A lot of attention in recent years has been given to classifying the possible covariant calculi on various groups and quantum groups, a project recently solved generically for finite groups and for the coordinate algebras of standard quantum groups $U_q(g)$.[3]

In the present paper we will concentrate on the simplest Hopf algebra of all, namely the coordinate algebra $k[x]$ of the line, but where k is a general field. In this case the classification of calculi itself is quite elementary, although we have not seen it treated explicitly elsewhere. We provide this explicit treatment in the preliminary Sec. II and see that the coirreducible calculi are of the form $k_\lambda[x]$, where $k_\lambda\supseteq k$ is a field extension of k. We then proceed in Sec. III to compute the quantum differentials associated to the extension $\mathbb{C}\supseteq\mathbb{R}$ (so the functions are $\mathbb{R}[x]$ and the 1-forms are $\mathbb{C}[x]$). In Sec. IV we consider quantum cohomology and gauge theory in this case, and compute the moduli of algebraic flat connections. In both cases we find nontrivial configurations

---

[a)]Royal Society University Research Fellow and Fellow of Pembroke College, Cambridge, England. Electronic mail: majid@damtp.cam.ac.uk

reflecting the nontriviality of the field extension from real to complex numbers. Section V concludes with some further quantum geometric considerations.

Note that the possible quantum differentials in 1-variable over the field $\mathbb{C}$ are already known[4] to be parametrized by $\lambda_0\in\mathbb{C}$ and have the form

$$\mathrm{d}f(x)=\mathrm{d}x\,\frac{f(x+\lambda_0)-f(x)}{\lambda_0},\quad x\,\mathrm{d}x-(\mathrm{d}x)x=\lambda_0.$$

Of these, $\lambda_0=0$ is the classical Newtonian calculus while the others are clearly the natural differential calculi for a version of lattice geometry based on finite differences. The calculi in these cases do not have nontrivial cohomology; the new phenomenon to be described in the present paper arises only when we consider quantum differentials over a nonalgebraically closed field such as $\mathbb{R}$.

II. Ω^1 AND FIELD EXTENSIONS

Let A be an algebra, which we consider as playing the role of ''co-ordinates'' in algebraic geometry, except that we do not require the algebra to be commutative. The appropriate notion of cotangent space or differential 1-forms in this case is the following:[2]

1. Ω^1 an A-bimodule.
2. $\mathrm{d}:A\to\Omega^1$ a linear map obeying the Leibniz rule $\mathrm{d}(ab)=a\,\mathrm{d}b+(\mathrm{d}a)b$ for all $a,b\in A$.
3. The map $A\otimes A\to\Omega^1$, $a\otimes b\mapsto a\,\mathrm{d}b$ is surjective.

When A has a Hopf algebra structure with coproduct $\Delta:A\to A\otimes A$ and counit $\epsilon:A\to k$ (k the ground field), we say that Ω^1 is bicovariant if the following occurs:

4. Ω^1 is a bicomodule with coactions $\Delta_L:\Omega^1\to A\otimes\Omega^1$, $\Delta_R:\Omega^1\to\Omega^1\otimes A$ bimodule maps (with the tensor product bimodule structure on the target spaces, where A is a bimodule by left and right multiplication).
5. d is a bicomodule map with the left and right regular coactions on A provided by Δ.

A morphism of calculi means a bimodule and bicomodule map forming a commuting triangle with the respective d maps. One says[3] that a calculus is coirreducible if it has no proper quotients.

Note that we do not demand that $a\,\mathrm{d}b=(\mathrm{d}b)a$, for in this case axiom 2 would imply that $\mathrm{d}(ab-ba)=0$, which we cannot naturally suppose when A is noncommutative. This possible noncommutativity of forms and ''functions'' is the main generalization featuring in the above axioms; we say that a differential calculus is noncommutative or ''quantum'' if the left and right multiplication of forms by functions do not coincide. Also, given Ω^1, there are natural prolongations to higher order differential forms, i.e., the entire exterior algebra $\Omega^\cdot$. We recall this when it is needed, in Sec. IV. Note that this approach is somewhat different from Ref. 5, where an entire $\Omega^\cdot$ on an algebra is effectively specified via a ''spectral triple.''

On the other hand, when Ω^1 is required to be bicovariant, there is a standard argument[2] that it must be of the form

$$\Omega^1=\Omega_0\otimes A,\quad \mathrm{d}a=(\pi\otimes\mathrm{id})(\Delta a-1\otimes a),$$

where $\Omega_0=\ker\epsilon/\mathcal{M}$ with canonical projection $\pi:\ker\epsilon\to\Omega_0$ and $\mathcal{M}$ is a left ideal contained in $\ker\epsilon$ and stable under the Hopf algebra adjoint coaction Ad. The right (co)module structures are those of A alone by (co)multiplication. The left (co)module structures are the tensor product of those on Ω_0 as inherited from $\ker\epsilon\subset A$ (where A acts by left multiplication and coacts by Ad) and those on A by (co)multiplication. We recall that modules and comodules of a Hopf algebra A have a tensor product induced by the coproduct and product of A, respectively. Then bicovariant Ω^1 are in 1–1 correspondence with the Ad-stable left ideals $\mathcal{M}\subseteq\ker\epsilon$. When A is cocommutative the adjoint coaction Ad is trivial.

Proposition 2.1: When $A=k[x]$, the coirreducible bicovariant Ω^1 are in 1–1 correspondence with irreducible monic polynomials $m\in k[\lambda]$, and take the form $\Omega^1=k_\lambda[x]$, where $k_\lambda=k[\lambda]/\langle m\rangle$ is the corresponding field extension. The bimodule structures and d differential are

$$f(x)\cdot P(\lambda,x)=f(x+\lambda)P(\lambda,x),\quad P(\lambda,x)\cdot f(x)=P(\lambda,x)f(x),\quad \mathrm{d}f(x)=\frac{f(x+\lambda)-f(x)}{\lambda}$$

for all $f\in k[x]$, $P\in k_\lambda[x]$.

Proof: We are interested in $A=k[x]$, the polynomials over a field k, forming a Hopf algebra with its additive coproduct counit and antipode,

$$\Delta x=x\otimes 1+1\otimes x,\quad \epsilon x=0,\quad Sx=-x.$$

According to the above, bicovariant differential calculi on $k[x]$ are in 1–1 correspondence with ideals $\mathcal{M}\subseteq\ker\epsilon$. Here $\ker\epsilon=\langle x\rangle$, the ideal generated by x in $k[x]$. Since $k[x]$ is a principal ideal domain (P.I.D.), the ideal $\mathcal{M}$ above is generated by a polynomial. Since $\mathcal{M}\subseteq\ker\epsilon$, this polynomial is divisible by x, i.e., $\mathcal{M}=\langle xm\rangle$. Coirreducible calculi correspond to m irreducible and monic.

We identify the corresponding $\Omega_0=\langle x\rangle/\langle xm\rangle\cong k[\lambda]/\langle m\rangle=k_\lambda$ by $xf(x)\mapsto f(\lambda)$. Under this identification, $\Omega^1=\Omega_0\otimes k[x]\cong k_\lambda[x]$. The action from the right is by the inclusion $k[x]\subset k_\lambda[x]$. The action from the left is by

$$f(x)\cdot x^m\otimes x^n=f(x\otimes 1+1\otimes x)x^m\otimes x^n,$$

as the tensor product action. Hence $f(x)\cdot\lambda^{m-1}x^n=f(\lambda+x)\lambda^{m-1}x^n$ under our identification. The quotient by $\langle xm(x)\rangle$ or $\langle m(\lambda)\rangle$ is understood in these expressions.

We compute $\mathrm{d}f=f(x\otimes 1+1\otimes x)-1\otimes f(x)$ modulo $\langle xm\rangle$ in the first tensor factor. Under our isomorphism this is $[f(\lambda+x)-f(x)]/\lambda$ modulo $\langle m(\lambda)\rangle$. Note that $\mathrm{d}x=x\otimes 1$ modulo $\langle xm\rangle$ becomes $\mathrm{d}x=1\in k_\lambda[x]$.

To see explicitly that the correspondence here is indeed 1–1, suppose that $k_{\lambda_1}[x]\cong k_{\lambda_2}[x]$ as quantum differential calculi associated to $m_1(\lambda_1)$ and $m_2(\lambda_2)$. Since the isomorphism is, in particular, a right module map under $k[x]$, it restricts to the identity on $k[x]$. And since the isomorphism forms a commutative triangle with the d maps, it identifies $[f(x+\lambda_1)-f(x)]/\lambda_1$ with $[f(x+\lambda_2)-f(x)]/\lambda_2$ for all f. Taking $f=x^2$ we have $2x+\lambda_1$ and $2x+\lambda_2$ identified, hence λ_1,λ_2 identified. Similarly by induction λ_1^n and λ_2^n are identified for all $n\geqslant 0$. One can also use the left module map property to conclude this. Hence $m_1(\lambda_2)=0$ in k_{λ_2}. Hence m_2 divides m_1. Since m_1 is monic and irreducible, we conclude that $m_1=m_2$ as required. The converse direction is clear. □

This generalizes the observation in Ref. 4 that coirreducible bicovariant quantum differential calculi over $\mathbb{C}[x]$ are parametrized by $\lambda_0\in\mathbb{C}$ (say). Here $m(\lambda)=\lambda-\lambda_0$ and $\pi(\lambda)=\lambda_0$. Hence, in this case,

$$\mathrm{d}f=\mathrm{d}x\,\frac{f(x+\lambda_0)-f(x)}{\lambda_0}. \tag{1}$$

The ratio on the right should be understood as the coefficient of λ_0 in $f(x+\lambda_0)-f(x)$, i.e., we include the usual differential calculus as the case $\lambda_0=0$. More generally, if the extension is Galois, the roots λ_i of m are as many as its degree and are primitive elements of k_λ, i.e., $k_\lambda\cong k[\lambda_i]$ by setting $\lambda=\lambda_i$, for each i. This gives us different ways of thinking of the differentials in Proposition 2.1 concretely as finite differences, all of them equivalent via the action of the Galois group of k_λ automorphisms that permute the λ_i.

It is easy to verify that Ω^1 in Proposition 2.1 is bicovariant under the left and right coactions,

$$\Delta_R P(\lambda,x)=P(\lambda,x+y)\in k_\lambda[x]\otimes k[y],\quad \Delta_L P(\lambda,x)=P(\lambda,y+x)\in k[y]\otimes k_\lambda[x], \tag{2}$$

induced by the coproduct Δ, as it must be by construction. Here the coacting copy of A is denoted by $k[y]$. The space Ω_0 is the subspace of Ω^1 invariant under the left coaction Δ_L, again by the general theory. Clearly, the dimension of Ω_0 over k, which is the dimension of the quantum differential calculus, is the degree of m, the degree of the associated field extension. The elements $\{1=\mathrm{d}x,\lambda,\cdots,\lambda^{\deg(m)-1}\}$ of $k_\lambda[x]$ are a basis of right-invariant 1-forms.

III. QUANTUM DIFFERENTIALS FOR THE COMPLEX EXTENSION OF THE REALS

In this section we consider in detail the case $k=\mathbb{R}$ and $m(\lambda)=\lambda^2+1$. Then $k_\lambda=\mathbb{C}$. The space of right-invariant 1-forms has the basis

$$\Omega_0=\{1=\mathrm{d}x,\quad \lambda=\mathrm{d}x^2-2(\mathrm{d}x)x\}.$$

We use the notations $\mathrm{d}x$ and $\omega\equiv\mathrm{d}x^2-2(\mathrm{d}x)x=x\,\mathrm{d}x-(\mathrm{d}x)x$ (by the Leibniz rule) for these two 1-forms in what follows.

Lemma 3.1: The left part of the bimodule structure on Ω^1 in this basis is given by

$$x\cdot\mathrm{d}x=(\mathrm{d}x)x+\omega,\quad x\cdot\omega=\omega x-\mathrm{d}x.$$

Proof: The first equality is the definition of ω (given the Leibniz rule). The second depends on the irreducible polynomial m according to $x\cdot\omega=(x+\lambda)\lambda=\lambda x-1=\omega x-\mathrm{d}x$. □

Proposition 3.2: The exterior differential is given by

$$\mathrm{d}f(x)=(\mathrm{d}x)\mathfrak{I}f(x+\iota)+\omega(f(x)-\mathfrak{R}f(x+\iota)),$$

where $f\in\mathbb{R}[x]$ is continued to $\mathbb{C}$ and $\mathfrak{I}$, $\mathfrak{R}$ denote imaginary and real parts. The left and right multiplication of forms by functions are related by

$$f(x)\cdot(\mathrm{d}x\quad\omega)=(\mathrm{d}x\quad\omega)\begin{pmatrix}\mathfrak{R} & -\mathfrak{I}\\ \mathfrak{I} & \mathfrak{R}\end{pmatrix}f(x+\iota).$$

Proof: This follows directly as an example of Proposition 2.1 on writing $\lambda=\iota$. Here we provide a more conventional direct proof based on the more conventional description in Lemma 3.1. First we write Lemma 3.1 in matrix form

$$x\cdot(\mathrm{d}x\quad\omega)=(\mathrm{d}x\quad\omega)\begin{pmatrix}x & -1\\ 1 & x\end{pmatrix}.$$

Then $f(x)\cdot\mathrm{d}x=(\mathrm{d}x)f(x+\Lambda)^1{}_1+\omega f(x+\Lambda)^2{}_1$ where $\Lambda=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}$ and the numerical indices denote the matrix element. We regard x for these purposes as multiplied by the identity matrix. Similarly for $f(x)\cdot\omega$.

Now, by induction on the Leibniz rule,

$$\begin{aligned}\mathrm{d}x^m&=x^{m-1}\,\mathrm{d}x+x^{m-2}(\mathrm{d}x)x+\cdots+(\mathrm{d}x)x^{m-1}\\ &=(\mathrm{d}x\quad\omega)(x^{m-1}+x^{m-2}(x+\Lambda)+\cdots+x(x+\Lambda)^{m-2}+(x+\Lambda)^{m-1})\begin{pmatrix}1\\0\end{pmatrix}\\ &=(\mathrm{d}x\quad\omega)\left(\frac{(x+\Lambda)^m-x^m}{\Lambda}\right)\begin{pmatrix}1\\0\end{pmatrix}\\ &=(\mathrm{d}x\quad\omega)((x+\Lambda)^m-x^m)\begin{pmatrix}0\\-1\end{pmatrix}.\end{aligned}$$

This provides the formula

$$\mathrm{d}f=(\mathrm{d}x\quad \omega)(f(x+\Lambda)-f(x))\begin{pmatrix}0\\-1\end{pmatrix}.$$

Since $\Lambda^2=-1$, we then identify the Λ^0 and Λ^1 parts of $f(x+\Lambda)-f(x)$ with the real and imaginary parts of $f(x+\iota)-f(x)$ as stated. Similarly for $f(x)\cdot \mathrm{d}x$ and $f(x)\cdot\omega$. □

To gain further insight into this differential calculus it is useful to embed it in the 1-parameter family corresponding to $m(\lambda)=\lambda^2+q^2$, $q\in\mathbb{R}$. This is isomorphic to the above $q=1$ case for all $q\neq 0$ and not irreducible for $q=0$. It does, however, have an interesting limit as $q\to 0$. Briefly, the relevant formulas are

$$x\cdot(\mathrm{d}x\quad \omega)=(\mathrm{d}x\quad \omega)\begin{pmatrix}x & -q^2\\ 1 & x\end{pmatrix}, \tag{3}$$

resulting in

$$\mathrm{d}f=q^{-2}\omega(f(x)-\Re f(x+\iota q))+q^{-1}(\mathrm{d}x)\Im f(x+\iota q). \tag{4}$$

This has a limit as $q\to 0$:

$$x\omega=\omega x,\quad x\,\mathrm{d}x=(\mathrm{d}x)x+\omega,\quad \mathrm{d}f=\omega\,\tfrac{1}{2}f''+(\mathrm{d}x)f', \tag{5}$$

in terms of the usual Newtonian derivative f'. This is the 2-jet calculus in Ref. 4 whereby up to second order derivatives are viewed as ''first order'' with respect to the new calculus and an appropriate ''braided derivation''[3] rule. We see that this calculus, although not coirreducible, arises naturally as a degenerate limit of coirreducibles corresponding to the extension $\mathbb{R}\subset\mathbb{C}$.

IV. QUANTUM COHOMOLOGY AND GAUGE THEORY OF FIELD EXTENSIONS

In this section, we consider two natural prolongations of the $\Omega^1(k[x])$ associated to a field extension to ''exterior algebras'' $\Omega^n(k[x])$ of degree $n>1$. We compute the first quantum cohomology for each prolongation in the case of the extension $\mathbb{R}\subseteq\mathbb{C}$, and the associated gauge theory.

We recall first that a differential graded algebra $\Omega^{\cdot}$ over a unital algebra A means a graded algebra with degree zero part A itself, and $\mathrm{d}:\Omega^{\cdot}\to\Omega^{\cdot}$ which increases the degree by 1 and obeys $\mathrm{d}^2=0$ and the graded Leibniz rule. In other words, $\Omega^{\cdot}$ has the algebraic properties of an ''exterior algebra'' in DeRahm theory and one may likewise compute its ''quantum de Rahm cohomology.'' Thus,

$$H^1=\{\omega\in\Omega^1|\mathrm{d}\omega=0\}/\{\mathrm{d}a|a\in A\}. \tag{6}$$

Given Ω^1, its maximal prolongation is defined as follows. First of all, we recall that in view of Axiom 3 above we can write Ω^1 as a quotient of the universal calculus $\Omega^1_U=\ker(\cdot:A\otimes A\to A)$ by a sub-bimodule $\mathcal{N}$. Here Ω^1_U has the obvious bimodule structure from $A\otimes A$ and $\mathrm{d}_U a=a\otimes 1-1\otimes a$. [Note that when A is a Hopf algebra then $A\otimes A\cong A\otimes A$ by $a\otimes b\mapsto(\Delta a)b$ restricts to $\Omega^1_U\cong\ker\epsilon\otimes A$ and $\mathcal{N}\cong\mathcal{M}\otimes A$, giving the description used in Sec. II.] Moreover, Ω^1_U is the degree 1 part of a canonical $\Omega^{\cdot}_U$ (albeit with trivial quantum cohomology). Here $\Omega^n_U\subset A^{\otimes n+1}$ as elements in the joint kernel of all product maps $\cdot_i$ multiplying the $i,i+1$'th copies of A. This can also be identified with $\Omega^n_U=\Omega^1_U\otimes_A\cdots\otimes_A\Omega^1_U$ in the obvious way; see Refs. 5 and 1. Here

$$\mathrm{d}_U(a_0\otimes a_1\cdots\otimes a_n)=\sum_{i=0}^{n+1}(-1)^{n+1-i}a_0\otimes\cdots\otimes a_{i-1}\otimes 1\otimes a_i\otimes\cdots\otimes a_n. \tag{7}$$

One may check that $\mathrm{d}_U\circ\mathrm{d}_U=0$. The product $\wedge$ of $\Omega^{\cdot}_U$ is given by multiplication between the two adjacent copies of A. A general $\Omega^{\cdot}$ over A is a quotient of $\Omega^{\cdot}_U$ by a differential graded ideal (i.e., an ideal stable under d_U). Without loss of generality we assume that the degree 0 part of the ideal is zero. The degree 1 part is some subbimodule $\mathcal{N}\subseteq\Omega^1_U$ and conversely, given $\mathcal{N}$ the maximal prolongation is provided by the differential ideal generated by $\mathcal{N}$. Its degree 2 part is $\mathcal{F}=\Omega^1_U\wedge\mathcal{N}+\mathcal{N}\wedge\Omega^1_U+d_U\mathcal{N}$ and $\Omega^2=\Omega^2_U/\mathcal{F}$.

Lemma 4.1: For the field extension $\mathbb{R}\subset\mathbb{C}$, the maximal Ω^2 is generated as an $\mathbb{R}[x]$-module by the two forms $\mathrm{d}x\wedge\mathrm{d}x$ and $\mathrm{d}x\wedge\omega$. Moreover,

$$d\omega=2\,\mathrm{d}x\wedge\mathrm{d}x=2\,\omega\wedge\omega,\quad \mathrm{d}x\wedge\omega=-\omega\wedge\mathrm{d}x.$$

Proof: The subbimodule $\mathcal{N}$ in our case is generated by $x\omega-\omega x+\mathrm{d}x$ where ω is defined as above. Now, $\mathrm{d}\omega=\mathrm{d}(x\,\mathrm{d}x-(\mathrm{d}x)x)=2\,\mathrm{d}x\wedge\mathrm{d}x$ from the definition of ω and the graded Leibniz rule and $\mathrm{d}^2=0$. Hence the subbimodule $\mathcal{F}$ is generated by $\Omega^1_U\wedge\mathcal{N}$, $\mathcal{N}\wedge\Omega^1_U$ and $\mathrm{d}x\wedge\omega+\omega\wedge\mathrm{d}x+2x\,\mathrm{d}x\wedge\mathrm{d}x-2(\mathrm{d}x\wedge\mathrm{d}x)x$. From Lemma 3.1 we have $x\,\mathrm{d}x\wedge\mathrm{d}x=(\mathrm{d}x)\wedge x\,\mathrm{d}x+\omega\wedge\mathrm{d}x=(\mathrm{d}x\wedge\mathrm{d}x)x+\mathrm{d}x\wedge\omega+\omega\wedge\mathrm{d}x$ up to terms in $\Omega^1_U\wedge\mathcal{N}$, $\mathcal{N}\wedge\Omega^1_U$. Therefore, $\mathcal{F}$ is generated by these and $\omega\wedge\mathrm{d}x+\mathrm{d}x\wedge\omega$.

Finally, from the definition of ω, and the relations in Ω^1 and $\mathcal{N}$, we have

$$\begin{aligned}\omega\wedge\omega&=(x\,\mathrm{d}x-(\mathrm{d}x)x)\wedge\omega\\&=-x\omega\wedge\mathrm{d}x-(\mathrm{d}x)x\wedge\omega\\&=\mathrm{d}x\wedge\mathrm{d}x-\omega\wedge x\,\mathrm{d}x-(\mathrm{d}x)x\wedge\omega\\&=\mathrm{d}x\wedge\mathrm{d}x-\omega\wedge\omega-\omega\wedge\mathrm{d}x\,x+\mathrm{d}x\wedge\mathrm{d}x-(\mathrm{d}x)\wedge\omega x\\&=2\,\mathrm{d}x\wedge\mathrm{d}x-\omega\wedge\omega,\end{aligned}$$

which gives the stated description of Ω^2 as a quotient of the tensor square over $\mathbb{R}[x]$ of Ω^1. □

Proposition 4.2: With the maximal Ω^2, the quantum de Rahm cohomology H^1 associated to $\mathbb{R}\subset\mathbb{C}$ vanishes.

Proof: Suppose $\mathrm{d}((\mathrm{d}x)f+\omega g)=0$, i.e., $-\mathrm{d}x\wedge\mathrm{d}f+2(\mathrm{d}x\wedge\mathrm{d}x)g-\omega\wedge\mathrm{d}g=0$. Putting in the form of $\mathrm{d}f$ and $\mathrm{d}g$ from Proposition 3.2, we see this is equivalent to

$$\mathfrak{I}g(x+\iota)=f(x)-\mathfrak{R}f(x+\iota),\quad \mathfrak{R}g(x+\iota)+g(x)=\mathfrak{I}f(x+\iota),$$

which can be combined into the single equation

$$f(x+\iota)-f(x)=\iota(g(x)+g(x+\iota)). \tag{8}$$

We now show that such f,g are necessarily of the form

$$f=\mathfrak{I}h(x+\iota),\quad g=h(x)-\mathfrak{R}h(x+\iota),$$

for some $h(x)$. Note first that if (f,g) obey (8) and without loss of generality $f=nx^{n-1}+\text{lower}$ degree, say, then

$$g=\frac{n(n-1)}{2}x^{n-2}+\text{lower degree}.$$

Indeed, writing $g=\mu x^p+\text{lower degree}$, the second half of (8) implies that $\mu x^p+\mu\mathfrak{R}(x+\iota)^p+\cdots=2\mu x^p+\cdots=n\mathfrak{I}(x+\iota)^{n-1}-nx^{n-1}+\cdots=n(n-1)x^{n-2}+\cdots$. Equating leading terms gives $p=n-2$ and $2\mu=n(n-1)$.

Now let

$$f_n=\mathfrak{I}(x+\iota)^n,\quad g_n=x^n-\mathfrak{R}(x+\iota)^n=x^n-(x+\iota)^n+\iota f_n,$$

for $n>0$. Note that the leading term of f_n is nx^{n-1} and the leading term of g_n is $[n(n-1)/2]x^{n-2}$. Hence

$$f=f_n+\overline{f},\quad g=g_n+\overline{g}$$

defines two polynomials $\overline{f},\overline{g}$ of lower degree. Now, since (f_n,g_n) are the components of the differential of x^n, and since $d^2=0$, we know that they obey (8). Hence $(\overline{f},\overline{g})$ obeys (8) and has a lower degree.

Therefore we have a proof by induction. The case where $n=2$ is easily seen to be true. That is, if $f=2x+\mu$ then (8) implies as above that $g=1$. Then, indeed, $f=\mathfrak{I}((x+\iota)^2+\mu(x+\iota))$ and $x^2+\mu x-\mathfrak{R}((x+\iota)^2+\mu(x+\iota))=1=g$, as required. In terms of differential forms, the assertion is that if $(\mathrm{d}x)(2x+\mu)+\omega h(x)$ is closed then $h(x)=1$ and the form is $\mathrm{d}(x^2+\mu x)$. This may also be verified directly from the relations in Lemma 4.1. □

Next we consider a natural quotient of the above prolongation which always exists when A is a Hopf algebra and Ω^1 is bicovariant. In this case $\Omega^1=\Omega_0\otimes A$ as explained in Sec. II, and $\Omega^\cdot$ is defined in such a way that the invariant differential forms ''braided-anticommute'' where the braiding is the one associated to the quantum double of A.[2] Fortunately, in our case where A is commutative and cocommutative, the quantum double braiding is the trivial flip map (the usual transposition). Hence, in this case we have simply $\Omega^n=\Lambda^n\Omega_0\otimes A$, where Λ^n denotes the usual exterior algebra of the vector space Ω_0. We call this the skew exterior algebra.

Proposition 4.3: The skew Ω^2 in the case $\mathbb{R}\subset\mathbb{C}$ is 1-dimensional with basis $\mathrm{d}x\wedge\omega$ [i.e., as in Lemma 3.1, with the additional relations $(\mathrm{d}x)^2=0=\omega^2$]. The first quantum cohomology in this case is $H^1=\mathbb{R}\omega$, i.e., 1-dimensional and spanned by ω.

Proof: This time $\mathrm{d}((\mathrm{d}x)f+\omega g)=0$ and $\mathrm{d}f,\mathrm{d}g$ from Proposition 3.2 implies only that

$$\mathfrak{I}g(x+\iota)=f(x)-\mathfrak{R}f(x+\iota), \tag{9}$$

as the coefficient of $\mathrm{d}x\wedge\omega$. [The first half of (8) does not apply since $\mathrm{d}x\wedge\mathrm{d}x=0$.] This equation still implies that if $f=nx^{n-1}+$lower degree and $n>2$ then $g=[n(n-1)/2]x^{n-2}+$lower degree, as before. Indeed, if $g=\mu x^p+\cdots$ then it says $\mu p x^{p-1}+\cdots=n(n-1)(n-2)/2+\cdots$. This is weaker than before because it does not fix μ when $n=2$. We proceed as before by writing $f=f_n+\overline{f}$, $g=g_n+\overline{g}$ so that $\overline{f},\overline{g}$ obey (9) and have a lower degree. In this way we obtain (without loss of generality by scaling f,g suitably) $f=F+2x+\mu$ and $g=G+\tau$ where $(\mathrm{d}x)F+\omega G=\mathrm{d}h$ for some h. Adding $f_2+\mu f_1$ and $g_2=1$ (here $g_1=0$) to F,G, we have $(\mathrm{d}x)f+\omega g=(1-\tau)\omega+\mathrm{d}h'$ for $h'=h+x^2+\mu x$. Hence $H^1=\mathbb{R}\omega$. Indeed $\mathrm{d}\omega=0$ for this choice of Ω^2 but ω is not exact. □

Finally, associated to any $\Omega^\cdot$ over a unital algebra A one has further ''quantum geometrical'' constructions, such as gauge theory. In its simplest form we consider a gauge field as any $\alpha\in\Omega^1$ and a gauge transform as an any invertible $\gamma\in A$. The group of gauge transforms acts on the set of α by

$$\alpha^\gamma=\gamma^{-1}\alpha\gamma+\gamma^{-1}\,\mathrm{d}\gamma. \tag{10}$$

The fundamental lemma of gauge theory is that the curvature

$$F(\alpha)=\mathrm{d}\alpha+\alpha\wedge\alpha\in\Omega^2 \tag{11}$$

is covariant in the sense that $F(\alpha^\gamma)=\gamma^{-1}F(\alpha)\gamma$. Moreover, one can consider sections $\psi\in A$ and a covariant derivative $\nabla\psi=\mathrm{d}\psi+\alpha\psi\in\Omega^1$. One has an action of the group of gauge transformations by $\psi^\gamma=\gamma^{-1}\psi$ and $\nabla^\gamma\psi^\gamma=(\nabla\psi)^\gamma$. These facts require only that Ω^1,Ω^2 obey the natural axioms as part of a differential graded algebra; see Ref. 4. Note that when Ω^1 is ''quantum,'' the nonlinearity in F does not necessarily collapse, even though the ''structure group'' here is trivial, i.e., one has many of the features of non-Abelian gauge theory. One may also consider α with values in some other algebra.

In our present setting where $A=k[x]$, only 1 will be invertible as a polynomial. One may enlarge A and our constructions above to handle this. Alternatively, instead of the ''finite'' gauge transformations γ one can consider only ''infinitesimal'' ones. Here an infinitesimal gauge transformation means $\theta\in k[x]$ acting by

$$\alpha^\theta=\alpha+\mathrm{d}\theta+\alpha\theta-\theta\alpha,\quad F(\alpha^\theta)=F(\alpha)+F(\alpha)\theta-\theta F(\alpha),\tag{12}$$

to lowest order in θ. This can be stated more formally as a vector field associated to each θ on the space of connections, etc., in the usual way. The covariant derivative $\nabla=\mathrm{d}+\alpha\wedge$ is covariant to lowest order under $\psi^\theta=\psi-\theta\psi$. By the same methods as in Ref. 4, one may check that any Ω^1,Ω^2 which are part of an exterior algebra will do for these features of gauge theory. Covariance of the curvature means that the vector fields associated to θ restrict to vector fields on the space of flat connections. They may not, however, restrict to only the algebraic (i.e., polynomial) part.

Proposition 4.4: For the extension $\mathbb{R}\subset\mathbb{C}$ and the maximal prolongation Ω^2 we write $\alpha=(\mathrm{d}x)a+\omega b$ and $F(\alpha)=(\mathrm{d}x)^2F_0+\mathrm{d}x\wedge\omega F_1$, say, then

$$F_0+\iota F_1=(a(x)+\iota(b(x)+1))(a(x+\iota)-\iota(b(x+\iota)+1))-1,$$

and the infinitesimal gauge transformations are

$$(a(x)+\iota(b(x)+1))\mapsto(a(x)+\iota(b(x)+1))(1+\theta(x)-\theta(x+\iota)).$$

The algebraic part of the space of flat connections is a circle,

$$\mathrm{Flat}=\{\mathrm{d}x\,s+\omega t|s,t\in\mathbb{R},\quad s^2+(t+1)^2=1\}=S^1\subset\mathbb{C}.$$

Here $s+\iota t\in\mathbb{C}\subset\mathbb{C}[x]\subset\Omega^1$ is a circle of unit radius centered at $-\iota$. The action of $\theta(x)=x$ is a unit vector field along the circle, so that the algebraic moduli space is the class of the zero connection.

Proof: We use Proposition 3.2 to compute $\mathrm{d}\alpha+\alpha\wedge\alpha$ in Ω^2. We then use Lemma 4.1 and collect the coefficients of $\mathrm{d}x\wedge\mathrm{d}x$ and $\mathrm{d}x\wedge\omega$ as

$$F_0=(-\Im a(x+\iota)+\Re b(x+\iota))(1+b(x))+b(x)+(\Re a(x+\iota)+\Im b(x+\iota))a(x),$$

$$F_1=(\Re a(x+\iota)+\Im b(x+\iota))(1+b(x))-a(x)+(\Im a(x+\iota)-\Re b(x+\iota))a(x).$$

Likewise from Proposition 3.2, the action of infinitesimal gauge transformation $\theta\in\mathbb{R}[x]$ is

$$a\mapsto a(x)(1+\theta(x))+\Im\theta(x+\iota)(1+b(x))-\Re\theta(x+\iota)a(x),$$

$$b\mapsto b(x)(1+\theta(x))+\theta-\Re\theta(x+\iota)(1+b(x))-\Im\theta(x+\iota)a(x).$$

We can then combine these expressions into the expressions shown for $F_0+\iota F_1$ and $a+\iota b$. Note that $\alpha=\mathrm{d}x\,a+\omega b=a+\iota b$ in the identification of Proposition 2.1, and similarly $F(\alpha)=\mathrm{d}x\wedge(F_0+\iota F_1)$ by an extension of this identification.

Next we compute the algebraic part of the space of flat connections. Suppose that

$$\alpha=a+\iota b=sx^n+\cdots+\iota(tx^m+\cdots),\quad s,t\neq0$$

are the leading terms for the real and imaginary parts. Here, $n,m\geqslant0$. Then

$$\begin{aligned}F(\alpha)=&-snx^{n-1}-\iota s\frac{n(n-1)}{2}x^{n-2}+2tx^m+\iota tmx^{m-1}+(sx^n+\iota tx^m)\\&\times\left(sx^n-\iota tx^m+\iota snx^{n-1}+tmx^{m-1}-s\frac{n(n-1)}{2}x^{n-2}+\iota t\frac{m(m-1)}{2}x^{m-2}\right)+\cdots\\=&-snx^{n-1}-\iota s\frac{n(n-1)}{2}x^{n-2}+2tx^m+\iota tmx^{m-1}+s^2x^{2n}+t^2x^{2m}+\iota s^2nx^{2n-1}+\iota t^2mx^{2m-1}\\&+st(m-n)x^{m+n-1}+\frac{\iota}{2}st(m(m-1)-n(n-1))x^{m+n-2}+\cdots.\end{aligned}$$

Now, since $n\geqslant 0$ the x^{n-1} and ιx^{n-2} terms can be dropped against the x^{2n} and ιx^{2n-1} terms, respectively.

Suppose that $m\geqslant 1$. Then the x^m and ιx^{m-1} terms can likewise be dropped. If $m=n$ then $(s^2+t^2)x^{2n}+\iota(s^2+t^2)nx^{2n-1}$ is dominant, in which case $F=0$ would imply $s=0$, $t=0$. So this case is excluded under our initial assumption. If $m>n$ then $t^2x^{2m}+\iota t^2mx^{2m-1}$ is dominant, in which case $t=0$. Likewise $m<n$ would imply $s=0$.

Hence $m=0$ for a flat connection under our assumption $s,t\neq 0$. In this case, if $n\geqslant 1$ then $s^2x^{2n}+\iota s^2nx^{2n-1}$ is dominant and $F=0$ would imply that $s=0$. Hence $n=0$ as well for a flat connection.

It remains to consider the simpler cases where $t=0$ or $s=0$ in our leading terms (i.e., real or imaginary α). If $t=0$ and $s\neq 0$ we similarly conclude that $n=0$ for a nonzero flat connection. And if $s=0$ and $t\neq 0$ then $m=0$ for a nonzero flat connection in the same way. Hence, for an algebraic connection of zero curvature, we are left with $\alpha=s+\iota t$ for $s,t\in\mathbb{R}$. Then

$$F(\mathrm{d}x\, s+\omega t)=\mathrm{d}(\mathrm{d}x\, s+\omega t)+(\mathrm{d}x\, s+\omega t)\wedge(\mathrm{d}x\, s+\omega t)=(t^2+2t+s^2)\mathrm{d}x\wedge\mathrm{d}x,$$

via Lemma 4.1, which tells us that $s^2+(t+1)^2=1$ for zero curvature.

For $\theta(x)=x\epsilon$, where $\epsilon\in\mathbb{R}$, we have the infinitesimal gauge transform $s+\iota(t+1)\mapsto(s+\iota(t+1))(1-\iota\epsilon)$ to lowest order in ϵ. This is an infinitesimal rotation of $s+\iota t$ about $-\iota$. □

Although we can consider only infinitesimal gauge transformations in our present algebraic setup, it is clear that the exponentiation of the infinitesimal gauge transformations associated to $\theta(x)=x\epsilon$ rotate us around the stated S^1. Since this S^1 passes through the origin, we see that all the algebraic zero curvature solutions stated are connected in this way to the zero connection by finite gauge transformations. Note also that infinitesimal gauge transformations by $\theta(x)=x^n\epsilon$, $n>1$ take us out of the space of algebraic zero curvature connections. This tells us that additional zero curvature connections beyond those in the proposition certainly exist in a suitable context, just not as polynomials. For example, the formal exponentiation of the gauge transform by $\theta(x)=x^2\epsilon$ of the $\alpha=-2\iota$ solution is

$$\alpha=-\iota(1+e^{\tau(1-2x\iota)}),\quad \tau\in\mathbb{R}.$$

It corresponds to the gauge transformation of $\alpha=-2\iota$ by $\gamma(x)=e^{\tau x^2}$, where (10) for the $\mathbb{R}\subset\mathbb{C}$ calculus comes out as

$$\alpha^\gamma+\iota=(\alpha+\iota)\frac{\gamma(x)}{\gamma(x+\iota)}. \tag{13}$$

Although we are not able to consider such finite gauge transformations and exponentials in our polynomial setting, we see that the infinitesimal gauge transforms do give us some information about the entire space of solutions.

Proposition 4.5: For the extension $\mathbb{R}\subset\mathbb{C}$ and the skew prolongation Ω^2 we write $\alpha=(\mathrm{d}x)a+\omega b$ as above and $F(\alpha)=\mathrm{d}x\wedge\omega F_1$, say. Then

$$F_1=(\Re a(x+\iota)+\Im b(x+\iota))(1+b(x))-a(x)+(\Im a(x+\iota)-\Re b(x+\iota))a(x),$$

and the infinitesimal gauge transformations by θ as in the preceding proposition. The algebraic part of the space of flat connections is the complex plane,

$$\mathrm{Flat}=\{\mathrm{d}s+\omega t|s,t\in\mathbb{R}\}=\mathbb{C},$$

where $s+\iota t\in\mathbb{C}\subset\mathbb{C}[x]=\Omega^1$. The algebraic moduli space of flat connections modulo gauge transformations is the half-line $\mathbb{R}_+$.

Proof: We take the same form for α with leading coefficients *s,t* as in the preceding proof. This time, however, the zero-curvature condition is only half of the preceding one. Indeed,

$$F_1=-s\frac{n(n-1)}{2}x^{n-2}+tmx^{m-1}+tx^m\left(-s\frac{n(n-1)}{2}x^{n-2}+tmx^{m-1}\right)$$
$$+sx^n\left(snx^{n-1}+t\frac{m(m-1)}{2}x^{m-2}\right)+\cdots,$$

for the leading terms after cancellations. We used the same expression for F_1 as the coefficient of $\mathrm{d}x\wedge\omega$ in the preceding proof. We drop x^{n-2} against x^{2n-1} and, assuming $m\geqslant 1$ we drop x^{m-1} as well. If $m=n$ we drop x^{m+n-2} and the dominant term is $(s^2+t^2)nx^{2n-1}$, which would imply $s=t=0$ for a flat connection. If $m>n$ the dominant term is t^2x^{2m-1} which would imply $t=0$. If $m<n$ the dominant term is s^2x^{2n-1} which would imply $s=0$. Hence $m=0$. Hence the dominant term is s^2nx^{2n-1} which would imply $s=0$ if $n\geqslant 1$. Hence $n=0$ as well. Finally, if we consider the similar form of α with $s=0$, the leading term for $m\geqslant 1$ would be t^2x^{2m-1} and imply $\alpha=0$, so $m=0$ in this case for a nonzero flat connection. If we consider α with $t=0$ then the leading term is s^2nx^{2n-1} as before, which would imply $n=0$. These are similar arguments to those in the preceding proof but relying now only on the imaginary part of the curvature. We deduce that an algebraic flat connection is of the form $\alpha=\mathrm{d}x\,s+\omega t$. This time, however, $F(\alpha)=0$ for all $s,t\in\mathbb{R}$ since $\mathrm{d}x\wedge\mathrm{d}x=0$ in the skew prolongation.

Infinitesimal gauge transformations are computed as before without change. Hence, the ones of the form $\theta(x)=x\epsilon$ rotate about $-\iota$ in the $s+\iota t$ plane. The orbits are circles of constant radius $s^2+(t+1)^2\in\mathbb{R}_+$. The different orbits are, however, inequivalent at least by such θ. (On the other hand, higher degree θ take us out of the class of polynomial connections.) Hence the algebraic part of the moduli space of flat connections is $\mathbb{R}_+$. □

Finally, the cohomology and moduli spaces in the maximal and skew prolongations are much more easily computed in the simpler 2-jet calculus resulting from the degenerate $q\to 0$ limit of the parametrized version of the $\mathbb{R}\subset\mathbb{C}$ extension. We first compute the maximal prolongation as having relations

$$\omega\wedge\omega=q^2\,\mathrm{d}x\wedge\mathrm{d}x,\quad \mathrm{d}\omega=2\,\mathrm{d}x\wedge\mathrm{d}x,\quad \mathrm{d}x\wedge\omega=-\omega\wedge\mathrm{d}x. \tag{14}$$

The proof is entirely similar to that of Lemma 4.1 (and equivalent to it after a rescaling), so we omit it. The degenerate limit is therefore

$$\omega\wedge\omega=0,\quad \mathrm{d}\omega=2\,\mathrm{d}x\wedge\mathrm{d}x,\quad \mathrm{d}x\wedge\omega=-\omega\wedge\mathrm{d}x. \tag{15}$$

The skew prolongation has the additional relation $\mathrm{d}x\wedge\mathrm{d}x=0$.

Proposition 4.6: The quantum cohomology for the 2-jet calculus is $H^1=0$ in the maximal prolongation and $H^1=\mathbb{R}\omega$ in the skew prolongation.

Proof: Here $\mathrm{d}((\mathrm{d}x)f+\omega g)=0$ implies

$$\tfrac{1}{2}f'=g,\quad \tfrac{1}{2}f''=g'.$$

Letting h be such that $h'=f$, we have $(\mathrm{d}x)f+\omega g=\mathrm{d}h$, so that H^1 is trivial. For the skew prolongation we have only $(1/2)f''=g'$, which implies $(1/2)f'=g-\mu$ where $\mu\in\mathbb{R}$. Choosing h such that $f=h'$, we have $(\mathrm{d}x)f+\omega g=\mathrm{d}h+\mu\omega$, so that $H^1=\mathbb{R}\omega$. □

Gauge theory in the $q\to 0$ limit is described in Ref. 4, and we now compute the moduli space of flat connections in this case.

Proposition 4.7: Writing $\alpha=(\mathrm{d}x)a+\omega b$, the curvature in the 2-jet calculus with the maximal prolongation is

$$F(\alpha)=\mathrm{d}x\wedge\mathrm{d}x(2b-a'+a^2)+\mathrm{d}x\wedge\omega(b'-\tfrac{1}{2}a''+a'a),$$

and is invariant under the gauge transformation,

$$a\mapsto a+\theta',\quad b\mapsto b-a\theta'+\tfrac{1}{2}(\theta''-(\theta')^2),$$

by $\theta\in\mathbb{R}[x]$. The moduli space of flat connections in the maximal prolongation is trivial and in the skew prolongation is $\mathbb{R}$, with flat connections gauge equivalent to $\alpha=\mu\omega$ for unique $\mu\in\mathbb{R}$.

Proof: We compute $F(\alpha)$ using the relations in Ω^2 and the commutation rules for Ω^1 at the end of Sec. III. The gauge transformation is likewise the infinitesimal gauge transformations as above but computed for this calculus, and corrected by the $-1/2(\theta')^2$ to make, in the present case, an exact gauge symmetry of the curvature (not only to lowest order in θ). These formulas are obtained by formally writing $\gamma=e^{\theta}$ in the finite gauge transformation formulas computed for the 2-jet calculus in Ref. 4; in our present case the result involves only polynomials in derivatives of θ, i.e., makes sense in terms of θ at our algebraic level. One then verifies directly at this level that $F(\alpha^{\theta})=F(\alpha)$.

The zero-curvature condition in the maximal prolongation is therefore

$$b=\tfrac{1}{2}(a'-a^2).$$

If this is the case then choose θ such that $\theta'=-a$. This gauge transforms $a\mapsto 0$. On the other hand, $b\mapsto b-a(-a)+1/2(-a'-a^2)=b+1/2(a^2-a')=0$ as well. Hence every flat connection is gauge equivalent to the zero one. By contrast, in the skew prolongation, the zero-curvature equation is

$$b'=\tfrac{1}{2}a''-a'a,$$

which means $b=1/2(a'-a^2)+\mu$ for some constant $\mu\in\mathbb{R}$. Making the same gauge transformation as before now sends $a\mapsto 0$ and $b\mapsto\mu$. Any further gauge transformation preserving $a=0$ would require $\theta'=0$, which would therefore not change the b component, i.e., the different μ cannot be related by any further gauge transformation. Hence the moduli space is $\mathbb{R}$ in the skew prolongation. □

V. CONCLUDING REMARKS

We conclude the paper with two miscellaneous pieces of general theory, demonstrated for our particular quantum exterior algebras. First, by Ref. 6, the Woronowicz $\Omega^{\cdot}$ (which in our case means the skew prolongation) is always a $\mathbb{Z}_2$-graded Hopf algebra with the coproduct extended by $\Delta=\Delta_L+\Delta_R$ on Ω^1. The same applies, in general, to the maximal prolongation, which again gives a $\mathbb{Z}_2$-graded Hopf algebra. From (2), we know (for any field extension) that $\Delta_L\lambda^n=1\otimes\lambda^n$ and $\Delta_R\lambda^n=\lambda^n\otimes 1$ (i.e., $\Omega_0=k_\lambda$ is left and right invariant). Hence the coproduct structure is with the basis of Ω_0 primitive, and the original coproduct of $k[x]$.

For example, for the extension $\mathbb{R}\subset\mathbb{C}$ we have the maximal prolongation $\Omega^{\cdot}$ as the $\mathbb{Z}_2$-graded Hopf algebra generated over $\mathbb{R}$ by x of degree zero and $\theta\equiv dx$, ω of degree 1, and the relations and coproduct

$$\begin{gathered} x\theta-\theta x=\omega,\quad x\omega-\omega x=-\theta,\quad \omega\theta=-\theta\omega,\quad \theta^2=\omega^2,\\ \Delta x=x\otimes 1+1\otimes x,\quad \Delta\theta=\theta\otimes 1+1\otimes\theta,\quad \Delta\omega=\omega\otimes 1+1\otimes\omega. \end{gathered}\tag{16}$$

The skew prolongation is the quotient of this by the additional relation $\theta^2=0$.

Finally, we consider what should be the notion of ''differentiable'' map $k[x]\to k[x]$ where the source and target are considered with differential calculi defined by m_1, m_2, respectively. A full analysis of the dependence of the above quantum geometric constructions on the choice of m will be developed elsewhere, but one may conjecture that at least some ''geometric'' invariants obtained from constructions of this type will be invariants of the field extension; i.e., if m_1, m_2 give isomorphic field extensions then some of the invariants should coincide. This is a long-term goal suggested by the above results, and would have applications in number theory (where the question of which monic polynomials give equivalent extensions is poorly understood for many fields k). The analysis of which maps $k[x]\to k[x]$ are indeed differentiable should be a first step in this geometric program.

We recall that any $\Omega^1(A)$ over a unital algebra A is a quotient $\Omega^1_U A/\mathcal{N}_A$ of the universal 1-forms $\Omega^1_U A\subset A\otimes A$. Any algebra map $\phi:A\to B$ (between unital algebras A,B) clearly induces a map $\phi\otimes\phi:\Omega^1_U A\to\Omega^1_U B$. Given this situation, we say that ϕ is differentiable if $\phi\otimes\phi$ descends to a map $\Omega^1(A)\to\Omega^1(B)$. If so, we denote the map by ϕ_* and note that it obeys the commutative diagram

$$\begin{array}{ccc} A & \xrightarrow{\phi} & B \\ \mathrm{d}\downarrow & & \downarrow\mathrm{d} \\ \Omega^1(A) & \xrightarrow{\phi_*} & \Omega^1(B) \end{array}, \tag{17}$$

since the universal d_U for A,B clearly obey this. The condition for differentiability is that $(\phi\otimes\phi)(\mathcal{N}_A)\subseteq\mathcal{N}_B$.

Proposition 5.1: In the setting of Proposition 2.1, an algebra map $\phi:k[x]\to k[x]$ defined by $\phi(x)=\Phi\in k[x]$ is differentiable with respect to calculi defined by $m_1(\lambda_1)$, $m_2(\lambda_2)$ on the source and target, respectively, iff

$$\mathrm{d}\Phi=0,\quad \text{or}\quad m_1(\Phi(\lambda_2+x)-\Phi(x))=0,$$

in $k_{\lambda_2}[x]$. Then $\phi_*(P(\lambda_1,x))=(\mathrm{d}\Phi)P(\Phi(\lambda_2+x)-\Phi(x),\Phi(x))$, where the product is in $k_{\lambda_2}[x]$.

Proof: We use the explicit isomorphism $\theta:\Omega^1_U A\cong\ker\epsilon\otimes A$ provided by $\theta(a\otimes b)=a_{(1)}\otimes a_{(2)}b$ and $\theta^{-1}(a\otimes b)=a_{(1)}\otimes(Sa_{(2)})b$ where S is the antipode and $\Delta a=a_{(1)}\otimes a_{(2)}$ (the summation is understood). In view of this, the map $\phi\otimes\phi$ becomes the map $\phi^U_*:\ker\epsilon\otimes A\to\ker\epsilon\otimes A$ as given by

$$\phi^U_*(a\otimes b)=\theta(\phi(a_{(1)})\otimes\phi(Sa_{(2)})\phi(b))=\phi(a_{(1)})_{(1)}\otimes\phi(a_{(1)})_{(2)}\phi(Sa_{(2)})\phi(b),$$

for all $a\in\ker\epsilon$ and $b\in A$. In the present setting, this becomes

$$\phi^U_*(yg(y)\otimes f(x))=(\Phi(y+x)-\Phi(x))g(\Phi(y+x)-\Phi(x))f(\Phi(x)),$$

for polynomials f,g (we write $A\otimes A=k[y,x]$). As in the proof of Proposition 2.1, we further identify the source $\ker\epsilon=k[\lambda_1]$ by $yg(y)\mapsto g(\lambda)$. We likewise identify the target $\ker\epsilon=k[\lambda_2]$ in the similar, say. With these identifications understood, we have

$$\phi^U_*(g(\lambda_1)\otimes f(x))=\frac{\Phi(\lambda_2+x)-\Phi(x)}{\lambda_2}g(\Phi(\lambda_2+x)-\Phi(x))f(\Phi(x)).$$

This map descends to the quotients $k_{\lambda_1}=k[\lambda_1]/\langle m_1\rangle$ and $k_{\lambda_2}=k[\lambda_2]/\langle m_2\rangle$ iff

$$\phi^U_*(m_1(\lambda_1)\otimes 1)=0,$$

in $k_{\lambda_2}[x]$, i.e., iff

$$(\mathrm{d}\Phi)m_1(\Phi(\lambda_2+x)-\Phi(x))=0,$$

in $k_{\lambda_2}[x]$, where we used the description of d in the target calculus from Proposition 2.1. This is the condition stated. Of the two possibilities, the second is more interesting in view of the form of ϕ_*. □

For example, for the differential calculus associated to $\mathbb{R}\subset\mathbb{C}$ in the source and target, the differentiability condition is

$$\Phi(x+\iota)-\Phi(x)=\begin{cases}0,\\ \pm\iota,\end{cases} \tag{18}$$

which at the algebraic level means $\Phi(x)=\pm x+\mu$ or $\Phi(x)=\mu$ for $\mu\in\mathbb{R}$. If we allow nonpolynomials then other possibilities, such as $\Phi(x)=e^{2\pi x}$, certainly open up. By contrast, for the degenerate 2-jet calculus in the source and target, the differentiability condition is automatically satisfied for all $\Phi\in\mathbb{R}[x]$. Here $m(\lambda)=\lambda^2$ is not irreducible but one can use the same formulas (the calculus is merely not coirreducible). Then $\Phi(\lambda+x)-\Phi(x)=\lambda\Phi'$ and $m(\lambda\Phi')=\lambda^2(\Phi')^2=0$ for all Φ.

ACKNOWLEDGMENT

I would like to thank D. Solomon for useful discussions at the start of the project.

[1] T. Brzeziński and S. Majid, ''Quantum group gauge theory on quantum spaces,'' Commun. Math. Phys. **157**, 591–638 (1993); **167**, 235 (1995) (erratum).
[2] S. L. Woronowicz, ''Differential calculus on compact matrix pseudogroups (quantum groups),'' Commun. Math. Phys. **122**, 125–170 (1989).
[3] S. Majid, ''Classification of bicovariant differential calculi,'' J. Geom. Phys. **25**, 119–140 (1998).
[4] S. Majid, ''Advances in quantum and braided geometry,'' in *Quantum Group Symposium at Group XXI*, edited by H.-D. Doebner and V. K. Dobrev (Heron Press, Sofia, 1997), pp. 11–26.
[5] A. Connes, *Noncommutative Geometry* (Academic, New York, 1994).
[6] T. Brzeziński, ''Remarks on bicovariant differential calculi and exterior Hopf algebras,'' Lett. Math. Phys. **27**, 287 (1993).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Nonunitary representations of the SU(2) algebra in the Dirac equation with a Coulomb potential

R. P. Martínez-y-Romero[a)] and A. L. Salas-Brito[b)]
Laboratorio de Sistemas Dinámicos, Departamento de Ciencias Básicas, UAM-Azcapotzalco, Apartado Postal 21-726, Coyoacán 04000 DF, Mexico

Jaime Saldaña-Vega
Facultad de Ciencias, Universidad Nacional Autónoma de México, Apartado Postal 50-542, 04510 DF, Mexico

(Received 13 November 1997; accepted for publication 7 January 1999)

A novel realization of the classical SU(2) algebra is introduced for the Dirac relativistic hydrogen atom defining a set of operators that allow the factorization of the problem. An extra phase is needed as a new variable in order to define the algebra. We take advantage of the operators to solve the Dirac equation using algebraic methods. A similar path to the one used in the angular momentum case is used; hence, the radial eigenfunctions so calculated comprise nonunitary representations of the algebra. One of the interesting properties of such nonunitary representations is that they are not labeled by integer nor by half-integer numbers, as occurs in the usual angular momentum representation. © *1999 American Institute of Physics.* [S0022-2488(99)02004-6]

I. INTRODUCTION

The unitary representations of groups and algebras are of a great interest in physics. We can mention the ubiquitous example of angular momentum or the SU(2) algebra. In this instance, as it is well known, the representations are labeled by two real numbers j and m, which may take only integer or half-integer values; the representations depend only on two parameters, θ and ϕ, say, which are defined in the compact sets $[0, \pi]$ and $[0, 2\pi]$, respectively.[1] However, there are many other physically interesting groups or algebras that are not unitary. In such a case, the representations are not restricted to parameters defined in compact sets nor are its generators necessarily Hermitian, but they nevertheless can play an important role in physics. The Lorentz group being a very important example of the class of physically relevant noncompact groups whose algebra is not necessarily unitary.[2]

It is our purpose in this article to introduce a realization of the cyclic SU(2) algebra in terms of non-Hermitian operators and then to use these operators to factorize and solve the relativistic Dirac hydrogen atom. The solution is obtained using algebraic methods, using the basic operators of the system following a route parallel to the unitary compact case, but introducing an extra variable (it is found below to play the role of a phase), which is required in our approach.[3]

The use of algebraic techniques has been common for these sort of problems. For example, in the relativistic hydrogen atom à la Dirac, one approach has been the use of shift operators;[4] whereas in the nonrelativistic case a successful approach requires the use of ladder operators and the factorization method.[5,6] Our approach is more akin to the introduction of ladder operators than to the shift operator method customarily used for this problem.[4]

The paper is organized as follows. In Sec. II we introduce the equations of the problem and define our notation. In Sec. III we construct the basic operators spanning the SU(2) algebra as a useful tool for the problem. In Sec. IV we define the inner product needed to investigate the

[a)]On sabbatical leave from Facultad de Ciencias UNAM; electronic mail: rodolfo@dirac.fciencias.unam.mx
[b)]Electronic mail: asb@hp9000a1.uam.mx

properties of our basic operators. In Sec. V some nonunitary representations of the symmetry algebra are constructed as the radial eigenfunctions of the problem. We find that they are noncompact representations that play an equivalent role to the spherical harmonics in the unitary case. In order to write such representations we find it convenient to define a family of polynomials. In Sec. VI, using algebraic methods and the operators defined in Sec. II, we find the energy spectrum of the hydrogen atom. In Sec. VII we give our conclusions. In the Appendix, we list and plot the explicit expressions of the first six mentioned polynomials; they are then explicitly related to the generalized associated Laguerre polynomials previously used by Davis for expressing the radial eigenfunctions of the hydrogen atom.[7]

II. THE DIRAC HYDROGEN ATOM

The symmetry algebra for the bound states of the nonrelativistic hydrogen atom, even in the classical case, is well known to be the SO(4),[8,9] but in the relativistic case the situation is different, as it is well known. The algebra associated with the symmetry of the radial part of the problem can be regarded as SU(2), as occurs with the angular momentum. The beautiful thing that follows is that we can then proceed to solve the relativistic hydrogen atom following a method that essentially parallels the calculation of eigenfunctions and eigenvalues of the angular momentum at the only expense of introducing an extra phase.

Let us begin with the Dirac Hamiltonian of the hydrogen atom,

$$H_{\mathrm{D}}=\boldsymbol{\alpha}\cdot\mathbf{p}+\beta m-\frac{Ze^2}{r}, \tag{1}$$

where m is the mass of an electron and $\boldsymbol{\alpha}$ and β are the standard Dirac matrices,[10]

$$\boldsymbol{\alpha}=\begin{pmatrix}0 & \boldsymbol{\sigma}\\ \boldsymbol{\sigma} & 0\end{pmatrix}, \quad \beta=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \tag{2}$$

where the 1's and 0's stand, respectively, for 2×2 unit and zero matrices and the $\boldsymbol{\sigma}$ is the standard vector composed by the three Pauli matrices $\boldsymbol{\sigma}=(\sigma_x,\sigma_y,\sigma_z)$. Since the Hamiltonian (1) is invariant under rotations, the solutions of the problem can be written in the form

$$\psi(r,\theta,\phi)=\frac{1}{r}\begin{pmatrix}F(r)\mathcal{Y}_{jm}(\theta,\phi)\\ iG(r)\mathcal{Y}'_{jm}(\theta,\phi)\end{pmatrix}, \tag{3}$$

where $\mathcal{Y}_{jm}$ and $\mathcal{Y}'_{jm}$ are spinor spherical harmonics of opposed parity. Parity is a good quantum number in the problem because the Coulomb potential is invariant under reflections; parity goes as $(-1)^l$ and, according to the triangle's rule of addition of momenta, the orbital angular momentum l is given by $l=j\pm\frac{1}{2}$. But, instead of working directly with parity, we prefer to introduce the quantum number ϵ, defined by

$$\epsilon=\begin{cases}1, & \text{if } l=j+\frac{1}{2},\\ -1, & \text{if } l=j-\frac{1}{2}.\end{cases} \tag{4}$$

Thus $l=j+\epsilon/2$ in all cases; we also define $l'=j-\epsilon/2$. Accordingly, the spherical spinor $\mathcal{Y}_{jm}$ depends on l, whereas the spherical spinor $\mathcal{Y}'_{jm}$, which has an opposite parity, depends on l'. Writing the solutions in the form (3) completely solves the angular part of the problem.

Let us now address the radial part of the problem; we are interested in its bound states, then of the quantity $k\equiv\sqrt{m^2-E^2}$ is positive definite; furthermore, let us define

$$\zeta\equiv Ze^2, \quad \tau_j\equiv\epsilon\left(j+\frac{1}{2}\right), \quad \nu\equiv\sqrt{\frac{m-E}{m+E}}; \tag{5}$$

then, we can write the differential equations for the radial part of the problem, in terms of the dimensionless variable $\rho = kr$, as

$$\left(-\frac{d}{d\rho} + \frac{\tau_j}{\rho} \right) G(\rho) = \left(-\nu + \frac{\zeta}{\rho} \right) F(\rho), \tag{6}$$

and

$$\left(+\frac{d}{d\rho} + \frac{\tau_j}{\rho} \right) F(\rho) = \left(\nu^{-1} + \frac{\zeta}{\rho} \right) G(\rho); \tag{7}$$

these equations are to be regarded as the initial formulation of our problem.

The first thing we want to do is to show that Eqs. (6) and (7) can be rewritten using a set of three operators whose commutation relations define a SU(2) algebra. To this end, let us first introduce the new variable x through the relation [but please notice that this change is not required for any of the conclusions that follow; see the second remark below after Eq. (15)]

$$\rho = e^x, \tag{8}$$

so x is defined in the open interval $(-\infty, \infty)$ and redefine the radial functions $F(\rho)$ and $G(\rho)$, introduced in Eqs. (6) and (7), in the form

$$F(\rho(x)) = \sqrt{m+E}\, [\psi_-(x) + \psi_+(x)], \tag{9}$$

$$G(\rho(x)) = \sqrt{m-E}\, [\psi_-(x) - \psi_+(x)]. \tag{10}$$

In terms of the new functions, $\psi_+(x)$ and $\psi_-(x)$, we thus arrive at the following set of equations for our problem:

$$\left[\frac{d}{dx} + e^x - \frac{\zeta E}{\sqrt{m^2 - E^2}} \right] \psi_+(x) = \left(\frac{\zeta m}{\sqrt{m^2 - E^2}} - \tau_j \right) \psi_-(x), \tag{11}$$

and

$$-\left[\frac{d}{dx} - e^x + \frac{\zeta E}{\sqrt{m^2 - E^2}} \right] \psi_-(x) = \left(\frac{\zeta m}{\sqrt{m^2 - E^2}} + \tau_j \right) \psi_+(x). \tag{12}$$

This first-order system can be uncoupled, multiplying by the left the first equation [Eq. (11)] times the operators that appear between square brackets in the second equation and, *vice versa*, by multiplying the second equation [Eq. (12)] times the operators that appear (between square brackets) in the first one. This procedure gives us the second-order system,

$$\left[\frac{d^2}{dx^2} + 2\mu e^x - e^{2x} - \frac{1}{4} \right] \psi_+(x) = \left(\tau_j^2 - \zeta^2 - \frac{1}{4} \right) \psi_+(x), \tag{13}$$

and

$$\left[\frac{d^2}{dx^2} + 2(\mu - 1) e^x - e^{2x} - \frac{1}{4} \right] \psi_-(x) = \left(\tau_j^2 - \zeta^2 - \frac{1}{4} \right) \psi_-(x), \tag{14}$$

where we have defined

$$\mu \equiv \frac{\zeta E}{\sqrt{m^2 - E^2}} + 1. \tag{15}$$

At this point there are several remarks that need be made. First, in the next section the seemingly odd term $-\frac{1}{4}$ in Eqs. (13) and (14) is shown to be necessary to construct the algebra. Second, as we said before, the change of variable from ρ to x is not really necessary for any of the calculations in the article, however, we prefer to work in the x rather than in the ρ variable because this choice simplifies the appearance of some of the equations and, mainly, because it makes the inner product introduced in Sec. III [in Eq. (33)] to look more familiar; but to make contact with the usual description, we sometimes, at our convenience, revert to the variable ρ. As a third remark, notice that we can regard Eqs. (13) and (14) as two eigenvalue equations, where the common eigenvalue ω is given by

$$\omega=\tau_j^2-\zeta^2-\tfrac{1}{4}=j(j+1)-\zeta^2; \tag{16}$$

as it is obvious, we do not need to calculate ω because it follows directly from the radial symmetry of the problem and from the intensity of the interaction that is needed to set the scale. The fourth remark we want to make is that, as the minimum value of j is $\frac{1}{2}$, then $\omega\geqslant 0$ for at least $Z=1,2,...,$ up to 118; for a discussion of the significance of this number see Ref. 4, p. 236.

III. AN OPERATOR ALGEBRA FOR THE DIRAC HYDROGEN ATOM

Our main purpose of this article is the construction of nonunitary representations of the SU(2) algebra for the Dirac hydrogen atom;[3] let us introduce the operator

$$\Omega_3\equiv -i\frac{\partial}{\partial\xi}, \tag{17}$$

depending exclusively on the new variable ξ, which is essentially an extra phase, as must be clear in what follows, and

$$\Omega_{\pm}\equiv e^{\pm i\xi}\left(\frac{\partial}{\partial x}\mp e^{x}\mp i\frac{\partial}{\partial\xi}+\frac{1}{2}\right), \tag{18}$$

which depend both on ξ and on the transformed ''radial'' variable x. These three operators satisfy the following algebraic relations:

$$[\Omega_3,\Omega_{\pm}]=\pm\Omega_{\pm}, \tag{19}$$

and

$$[\Omega_+,\Omega_-]=2\Omega_3. \tag{20}$$

We can alternatively define the two operators, Ω_1 and Ω_2, as

$$\Omega_1=\frac{1}{2}(\Omega_++\Omega_-),\quad \Omega_2=\frac{1}{2i}(\Omega_+-\Omega_-), \tag{21}$$

in terms of which the algebraic properties (19) and (20) read as

$$[\Omega_i,\Omega_j]=i\epsilon_{ijk}\Omega_k,\quad i,j,k=1,2,3. \tag{22}$$

It is now clear that the commutation relations (19) and (20)—or just (22)—correspond to an SU(2) algebra.[1,3,8] To complete the discussion, we also need to introduce the Casimir operator of the algebra; let us consider the operator

$$\Omega^2=\Omega_1^2+\Omega_2^2+\Omega_3^2; \tag{23}$$

we can easily show that (23) is indeed a Casimir for the algebra (19) and (20) [or (22)],

$$[\Omega^2,\Omega_i]=0, \quad \text{for} \quad i=1,2,3. \tag{24}$$

Here, as in the angular momentum case, we can regard Ω^2 as the square of $\mathbf{\Omega}=\Omega_1\hat{\mathbf{i}}+\Omega_2\hat{\mathbf{j}}+\Omega_3\hat{\mathbf{k}}$. To obtain the explicit expression of Ω^2, it is better to calculate first the product $\Omega_-\Omega_+$:

$$\Omega_-\Omega_+=\frac{\partial^2}{\partial x^2}-e^{2x}-2ie^x\frac{\partial}{\partial\xi}+i\frac{\partial}{\partial\xi}+\frac{\partial^2}{\partial\xi^2}-\frac{1}{4}, \tag{25}$$

and obtain the Casimir operator from the relationship $\Omega^2=\Omega_-\Omega_++\Omega_3(\Omega_3-1)$. We remark that we do not have a linear term in $\partial/\partial x$ in Eq. (22) because we choose the last constant in Eq. (18) precisely as $\frac{1}{2}$. We then easily conclude that the Casimir operator is given explicitly by

$$\Omega^2=\frac{\partial^2}{\partial x^2}-e^{2x}-2ie^x\frac{\partial}{\partial\xi}-\frac{1}{4}. \tag{26}$$

Although we are not restricted to a compact set of parameters anymore due to the presence of the variable x, in analogy with the conventions used for the *spherical harmonics* $Y_l^m(\theta,\phi)$, we label the simultaneous eigenfunctions of Ω^2 and Ω_3 as $V_\omega^\mu(x,\alpha)$, where the numbers ω are the eigenvalues of Ω^2 and the numbers μ are the eigenvalues of Ω_3. Therefore, we write

$$\Omega_3V_\omega^\mu(x,\xi)=\mu V_\omega^\mu(x,\xi) \tag{27}$$

and

$$\Omega^2V_\omega^\mu(x,\xi)=\omega V_\omega^\mu(x,\xi); \tag{28}$$

we thus have

$$V_\omega^\mu(x,\xi)=e^{i\mu\xi}\mathcal{P}_\omega^\mu(x), \tag{29}$$

where, again, we have used the angular momentum analogy to write the x functions as $\mathcal{P}_\omega^\mu(x)$, resembling the usual notation for the Legendre polynomials $P_l^m(\theta)$. In this equation it becomes clear the role of ξ as just an extra phase.

A comparison with Eqs. (13) and (14) tells us that

$$\psi_+(x)=\mathcal{P}_\omega^\mu(x), \tag{30}$$

$$\psi_-(x)=\mathcal{P}_\omega^{\mu-1}(x). \tag{31}$$

The operators $\Omega_\pm$ are thus ladder operators for the problem; they move along the set of eigenfunctions, changing the eigenvalue μ to the eigenvalue $\mu\pm1$, in a completely analogous way to the the case of the angular momentum algebra,

$$\Omega_\pm V_\omega^\mu(x,\xi)\propto V_\omega^{\mu\pm1}(x,\xi); \tag{32}$$

the required proportionality constants are evaluated in Sec. V [Eq. (52)].

IV. THE INNER PRODUCT AND RELATED PROPERTIES OF THE OPERATORS

To establish the properties of the Ω operators and to construct the representations of the SU(2) algebra they span, we need a properly defined inner product. Here we cannot longer rely on the angular momentum analogy, since two of the parameters of the algebra we purport to construct are defined over the noncompact interval $(-\infty,\infty)$, making it completely different from the angular momentum. This situation implies that not all the generators of the algebra are to be Hermitian (indeed, Ω_3 is found to be Hermitian but Ω_1 and Ω_2 are are found to be anti-Hermitian) and, as a consequence, that the corresponding group could not be unitary.

Let us consider the operation

$$(\phi,\psi)=\int_0^{2\pi}\frac{d\xi}{2\pi}\int_{-\infty}^{\infty}\phi^*(\xi,x)\psi(\xi,x)dx. \tag{33}$$

It is easy to show that Eq. (33) defines an inner product, since, as it is not difficult to prove, it satisfies the three basic properties: (i) If $\psi(x)=0$, then $(\psi,\psi)=0$; (ii) if $\psi(x)\neq 0$, then $(\psi,\psi)\geqslant 0$; (iii) if c is any complex number, then $(\psi,c\phi)=c(\psi,\phi)$ and $(c\psi,\phi)=c^*(\psi,\phi)$.

To study the behavior of the generators of the algebra in terms of the inner product (33), we consider a certain function $\psi(x,\xi)$ associated with a certain fixed value μ_0. We define the elements of the associated Hilbert space H_{μ_0} with functions of the form

$$\psi(x,\xi)=e^{i(\mu_0+m-n)\xi}\mathcal{F}(x), \tag{34}$$

where m and n are integer numbers and $\mathcal{F}(x)$ is a well-behaved function depending only on x. As we exhibit in Sec. V, this is actually the general form of the functions inhabiting the Hilbert space of our problem.

With the help of the inner product (33), we can establish the properties of the Ω_a $(a=1,2,3)$ operators. Let us consider first Ω_3; in this case the important product is

$$(\psi',\Omega_3\psi)=\delta_{m'-n',m-n}(\mu_0+m-n)\int_{-\infty}^{\infty}\mathcal{F}^*(x)\mathcal{F}(x)dx; \tag{35}$$

this last equation, due to the presence of the Kronecker delta, can be written as

$$\int_0^{2\pi}\frac{d\xi}{2\pi}\int_{-\infty}^{\infty}\left[i\frac{\partial}{\partial\xi}e^{-i(\mu_0+m'-n')\xi}\right]e^{-i(\mu_0+m-n)\xi}\mathcal{F}^*(x)\mathcal{F}(x)dx, \tag{36}$$

and this is precisely $(\Omega_3\psi',\psi)$. We have in this way proved that Ω_3 is a Hermitian operator:

$$\Omega_3^\dagger=\Omega_3. \tag{37}$$

To study the operators $\Omega_\pm$, let us first consider Ω_+ and evaluate the product

$$(\psi',\Omega_+\psi)=\int_0^{2\pi}\frac{d\xi}{2\pi}\int_{\infty}^{\infty}e^{-i(\mu_0+m'-n')\xi}\mathcal{F}^*(x)dx\,e^{i\xi}\left(\frac{\partial}{\partial x}-e^x-i\frac{\partial}{\partial\xi}+\frac{1}{2}\right)e^{i(\mu_0+m-n)\xi}\mathcal{F}(x)dx. \tag{38}$$

To analyze this integral, it is simpler to split it in two parts. Let us consider first the $e^{i\xi}(-\partial/\partial\xi+1/2)$ part; its contribution to the inner product (38) is

$$\left(\mu_0+m-n+\frac{1}{2}\right)\delta_{m'-n',m-n+1}\int_{-\infty}^{\infty}\mathcal{F}^*(x)\mathcal{F}(x)dx; \tag{39}$$

this expression can be written as

$$-\int_0^{2\pi}\frac{d\xi}{2\pi}\left[e^{-i\xi}\left(i\frac{\partial}{\partial\xi}+\frac{1}{2}\right)e^{i(\mu_0+m'-n')\xi}\mathcal{F}(x)\right]^\dagger e^{i(\mu_0+m-n)\xi}\mathcal{F}(x)dx. \tag{40}$$

We now take care of the contribution of the term $e^{i\xi}(\partial/\partial\xi-e^x)$. After a partial integration this contribution becomes

$$\delta_{m'-n'-1,m-n}\int_{-\infty}^{\infty}\left[-\left(\frac{\partial}{\partial x}+e^{x}\right)\mathcal{F}(x)\right]^{\dagger}\mathcal{F}(x)\,dx. \tag{41}$$

Taking together the two previous results, it is easy to see that the operator complies with $\Omega_{+}^{\dagger} = -\Omega_{-}$. A similar calculation establish the analogous property $\Omega_{-}^{\dagger} = -\Omega_{+}$. Therefore, we have established that

$$\Omega_{\pm}^{\dagger} = -\Omega_{\mp}. \tag{42}$$

We have proved that not all the operators are Hermitian, thence arriving to the anticipated results: We showed that the operator Ω_3 is indeed Hermitian; as a consequence, we expect the range of the only parameter it depends on, ξ, to be compact; this is certainly the case since $\xi \in [0,2\pi]$, a compact set. The other two operators Ω_1 and Ω_2 depend upon x, a variable defined over the noncompact set $(-\infty, \infty)$, but here such operators are anti-Hermitian,

$$\Omega_a^{\dagger} = -\Omega_a, \quad a=1,2, \tag{43}$$

as it is easy to show from Eqs. (21) and (42).

V. REPRESENTATIONS OF THE SU(2) ALGEBRA (RADIAL EIGENFUNCTIONS OF THE PROBLEM)

With the inner product defined in the previous section, we are in the position of introducing a complete orthogonal basis of simultaneous eigenfunctions for Ω^2 and Ω_3, which, accordingly, must carry a representation of the algebra—and, besides, they solve our radial eigenvalue problem [Eqs. (6) and (7)]. We have decided to choose the commuting operators in a similar fashion to the standard SU(2) case. Let us define then

$$V_{\omega}^{\mu}(x,\xi) \equiv |\omega\mu\rangle, \tag{44}$$

where the kets $|\omega\mu\rangle$ are both assumed orthogonal and normalized respect the inner product (33),

$$\langle \omega' \mu' | \omega\mu\rangle = \delta_{\omega,\omega'}\delta_{\mu,\mu'}. \tag{45}$$

Now, as Ω_1 and Ω_2 are not Hermitian, the Casimir operator Ω^2 defined in (23) is not necessarily positive definite, we can nevertheless introduce a positive definite operator as

$$\boldsymbol{\Omega}^{\dagger}\cdot\boldsymbol{\Omega} = -\Omega_1^2 - \Omega_2^2 + \Omega_3^2 = 2\Omega_3^2 - \Omega^2; \tag{46}$$

the positivity of this operator allows us to show that

$$2\mu^2 \geqslant \omega, \tag{47}$$

that is, $|\mu|$ is bounded by below. As a consequence, there must exist a minimum value for $|\mu|$, let us say $\lambda \equiv |\mu|_{\min}$. We also know the that the ket $|\omega\lambda\rangle$ is annihilated by Ω_- or, equivalently, that $\Omega_+\Omega_-|\omega\lambda\rangle = 0$; so

$$\omega - \lambda^2 - \lambda = 0 \quad \text{or} \quad \omega = \lambda(\lambda-1). \tag{48}$$

Given Eq. (47), let us introduce a slight change in notation, writing λ instead of ω in the eigenfunctions $\mathcal{P}_{\omega}^{\mu}(x)$ of Eq. (29), that is, $\mathcal{P}_{\omega}^{\mu}(x)$ is to be replaced by $\mathcal{P}_{\lambda}^{\mu}(x)$. Furthermore, as λ has to be positive and since $\omega = \tau_j^2 - \zeta^2 - \frac{1}{4}$, we find that the minimum $|\mu|$ value is

$$|\mu|_{\min} \equiv \lambda = s + \tfrac{1}{2}, \tag{49}$$

where we have defined

$$s \equiv +\sqrt{\tau_j^2-\zeta^2}=\sqrt{(j+\tfrac{1}{2})^2-\zeta^2}. \tag{50}$$

Notice that s is a real quantity for $Z=1,2,3,...,137$.

However, the most important conclusion we can draw from Eqs. (49) and (50), is that λ no longer has to be an integer or half-integer number, as necessarily occurs in the unitary compact angular momentum case. We may see that this result is a direct consequence of both Ω_1 and Ω_2 not being Hermitian operators. So the operators introduced for the SU(2) algebra associated with the Dirac hydrogen atom lead naturally to nonunitary representations labeled by real numbers, or, at least, not necessarily integer nor half-integer ones.

For constructing the functions comprising the representations, let us introduce the constants $C_\mu^\pm$ as follows [compare with Eq. (32)]:

$$\Omega_\pm|\omega\mu\rangle = C_\mu^\pm|\omega\mu\pm 1\rangle; \tag{51}$$

these constants can be explicitly evaluated from $\langle\omega\mu|\Omega_+\Omega_-|\omega\mu\rangle = C_\mu^- C_{\mu-1}^+$ and from $(C_{\mu-1}^+)^* = -C_\mu^-$; if we further assume these constants to be real, we easily get

$$C_\mu^\pm = \pm\sqrt{\mu(\mu\pm 1)-\lambda(\lambda-1)}, \tag{52}$$

a result with a slightly different form than in the analogous angular momentum case.[4,8]

The ground state of the hydrogen atom can be obtained from the equation $\Omega_-|\lambda\lambda\rangle=0$ for the positive set of eigenvalues. Such an equation becomes

$$e^{-i\xi}\left(\frac{\partial}{\partial x}+e^x-\lambda+\frac{1}{2}\right)e^{i\lambda\xi}\mathcal{P}_\lambda^\lambda(x)=0, \tag{53}$$

whose solution is

$$\mathcal{P}_\lambda^\lambda(x)=d_\lambda e^{sx}\exp(-e^x)=d_\lambda e^{(\lambda-1/2)x}\exp(-e^x), \tag{54}$$

where

$$d_\lambda \equiv \frac{2^{(\lambda-1/2)}}{\sqrt{\Gamma(2\lambda-1)}}, \tag{55}$$

is a normalization constant and $\Gamma(y)$ stands for the Euler-gamma function. Since λ is the lowest eigenvalue, in this instance we should have

$$\psi_+(x)=\mathcal{P}_\lambda^\lambda(x), \tag{56}$$

and

$$\psi_-(x)=0. \tag{57}$$

In terms of the radial variable ρ, we can see from Eqs. (9) and (10) that the ground state solutions are

$$F(\rho)=\sqrt{\frac{(m+E)}{2m(\lambda-1/2)}}\,\rho^s e^{-\rho}, \tag{58}$$

and

$$G(\rho)=-\sqrt{\frac{(m-E)}{2m(\lambda-1/2)}}\,\rho^s e^{-\rho}, \tag{59}$$

normalized in the sense

$$\int_0^\infty (|F(\rho)|^2+|G(\rho)|^2)d\rho=1. \tag{60}$$

We can also see that in the case of negative eigenvalues the solutions behave as $\sim\rho^s e^\rho$, giving divergent behavior as $\rho\to\infty$. This behavior makes the negative energy solutions not square integrable; therefore, we have to discard them if we want to describe physically realizable states.[2,3,10]

The excited states can be obtained, as we do in the Appendix, by applying successively Ω_+ to the ground state $|\lambda\lambda\rangle$, as it is customarily done for the spherical harmonics. The result involves the functions $\mathcal{P}_\lambda^\mu(x)$ introduced by Eqs. (29) and (34). The $\mathcal{P}_\lambda^\mu(x)$ are polynomials multiplied by the weight factor $W(\rho)=\rho^{\lambda-1/2}e^{-\rho}$. This weight factor assures that the behavior of the big and the small components of the spinor are regular both at the origin as well as $\rho\to\infty$. As an illustration of the solutions discussed here, we quote in the Appendix the first few cases of $\mathcal{P}_\lambda^{\lambda+p}(x)$, for $p=0,...,5$; there we also plot these polynomial parts of the functions [i.e., we plot $\mathcal{P}_\lambda^\mu(x)$ without the weight factors] for the first five excited energy levels. Notice that the polynomial part in the eigenfunction $\mathcal{P}_\lambda^{\lambda+p}(x)$ is always of order p.

The matrix representations of the Ω_a, $a=1,2,3$, for $\lambda=s+\frac{1}{2}+p$, $p=0,1,2,3,...$, are constructed from Eqs. (27), (28), (51), and (52). In contrast with the standard SU(2) case, here the representations are non-Hermitian and infinite dimensional for each λ, except for Ω_3. The matrix elements of the operator Ω_3, including the negative eigenvalues series, are given by

$$\langle\omega\mu|\Omega_3|\omega\mu'\rangle=\mu\delta_{\mu\mu'}, \tag{61}$$

where $\mu=\pm(\lambda+p)$, $p=0,1,2,...$; therefore, its trace vanishes and the determinant of any element of the corresponding group with the form $\exp(i\Omega_3)\xi$, is always 1.

For the other two operators the only nonvanishing matrix elements are

$$\langle\omega\mu|\Omega_1|\omega\mu\pm1\rangle=\mp\tfrac{1}{2}\sqrt{\mu(\mu\pm1)-\lambda(\lambda-1)}, \tag{62}$$

$$\langle\omega\mu|\Omega_2|\omega\mu\pm1\rangle=-\frac{i}{2}\sqrt{\mu(\mu\pm1)-\lambda(\lambda-1)}. \tag{63}$$

This means again that the trace of both Ω_1 and Ω_2 vanish. Notice also that the determinant of a group element generated by Ω_1 or Ω_2 is 1 only for a purely imaginary parameter.

VI. THE ENERGY SPECTRUM

We can now evaluate the bound energy spectrum for the problem. As we mentioned in Sec. V, the bound state energy spectrum comprises only the positive series eigenvalues.[2,10] Let us first express the energy in terms of the eigenvalue μ from Eq. (15),

$$E=m\left[1+\frac{\zeta^2}{(\mu-1/2)^2}\right]^{-1/2}; \tag{64}$$

then, for the case of positive eigenvalues—the only ones with physically appropriate eigenfunctions—we have that $\mu=\lambda+p$, where p is a non-negative integer $p=0,1,2,...$, or, equivalently, that $\mu-\frac{1}{2}=s+p$. This gives precisely the energy spectrum of the relativistic hydrogen atom. To rewrite our result in a more familiar form, we need only to define the principal quantum number n and the auxiliary quantity ϵ_j as follows:

$$n=j+\tfrac{1}{2}+p, \tag{65}$$

$$\epsilon_j=n-s-p=j+\tfrac{1}{2}-s; \tag{66}$$

we then finally conclude that $\mu-\frac{1}{2}=s+p=n-\epsilon_j$, which gives precisely the well-known energy spectrum.[10]

VII. CONCLUDING REMARKS

In summary, we have constructed a SU(2) algebra for the hydrogen atom in the Dirac formulation, introducing the Hermitian operator Ω_3 and the anti-Hermitian operators Ω_1 and Ω_2. The result of all of this is that the representations are labeled by numbers λ that are neither integers nor half-integers, as in the case of the more familiar unitary representations. Nevertheless, the algebra introduced predicts precisely the energy eigenvalues and eigenfunctions of the Dirac hydrogen atom. One of the most noteworthy features of the representations reported here is the mixing of a spinorial angular momentum character, implying an equally spaced spectrum, with the energy requirements of the problem—requiring a differently spaced spectrum; the interplay of these two spectral requirements is basically reflected in the fact that the eigenvalues associated with Eqs. (13) and (14) follows from both the generic radial symmetry and the specific features of the interaction in the Dirac equation (1). From Eq. (15), it also follows that in the limit of a vanishing interaction, i.e., $\zeta\to 0$, our representations collapse and, in this special case, $\mu=1$ always. Such behavior is precisely as expected because there is no longer any restriction over the eigenvalues and thus the spectrum becomes continuous, corresponding to a free Dirac particle. The operator algebra we introduced allows an essentially algebraic solution of the Dirac hydrogen atom, which may have various applications.[3,4,6]

It is to be noted also the possible connections that our formulation may have with systems with hidden supersymmetric properties,[11–14] as we will discuss in a forthcoming article. The energy spectrum of the problem has some peculiarities that also appear in the spectrum of a Dirac oscillator;[12,13,15] in particular, the equally spaced energy solutions for $\psi_+(x)$ and $\psi_-(x)$ resemble, respectively, the behavior of the big and the small components of the aforementioned system. As a result of this resemblance, we are studying the hidden supersymmetric properties and the superconformal algebra, in the sense of Refs. 13 and 16, associated with this problem. The similitude of Eqs. (13) and (14) with the corresponding ones for a Morse oscillator should be also noticed.[6]

It is worth pinpointing that we are forced to introduce the new variable ξ in order to define the algebra; in terms of the solutions of the Dirac equation, ξ just plays the role of a phase. To exemplify, when we perform a ''rotation'' using Ω_3, the big component changes from $F(x)\propto[\psi_-(x)+\psi_+(x)]$ to $F(x)\propto[e^{i(\mu-1)\xi}\psi_-(x)+e^{i\mu\xi}\psi_+(x)]=e^{i\mu\xi}[e^{-i\xi\psi_-(x)+\psi_+(x)}]$. The phase $e^{i\mu\xi}$ does not play any observable role, but the term $e^{-i\xi}$ changes the relative phase between the $\psi_+(x)$ and the $\psi_-(x)$ components of the eigenfunction and, in consequence, changes the radial function $F(\rho)$ itself although the energy spectrum is still invariant under such a transformation; this is a consequence of the fact that $|\omega\mu\rangle$ and $\Omega_3|\omega\mu\rangle$ both correspond to the same eigenvalue μ. In a way, this resembles what happens when there are superselection rules in a system.[17–19]

To finalize, let us comment that it is not widely known that the radial eigenfunctions of the Dirac hydrogen atom can be expressed in terms of generalized associated Laguerre polynomials, as was realized by Davis a long time ago.[7,20] These polynomials, which are a generalization to nonintegral indices of the usual associated Laguerre polynomials, are defined as[7]

$$\mathcal{L}_p^{\alpha}(x)=\frac{\Gamma(\alpha+p+1)}{n!\,\Gamma(\alpha+1)}\,{}_1F_1(-p;\alpha+1;x), \tag{67}$$

where $\Gamma(x)$ is again the Euler gamma function, p is a positive integer, and ${}_1F_1(-p,\alpha+1;x)$ stands for the confluent hypergeometric function—having one of their arguments negative, the hypergeometric function reduces to a polynomial.[21] The polynomial representation of the radial eigenfunctions introduced here is related to that used by Davis in Eqs. (A8) and (A9) of the Appendix.

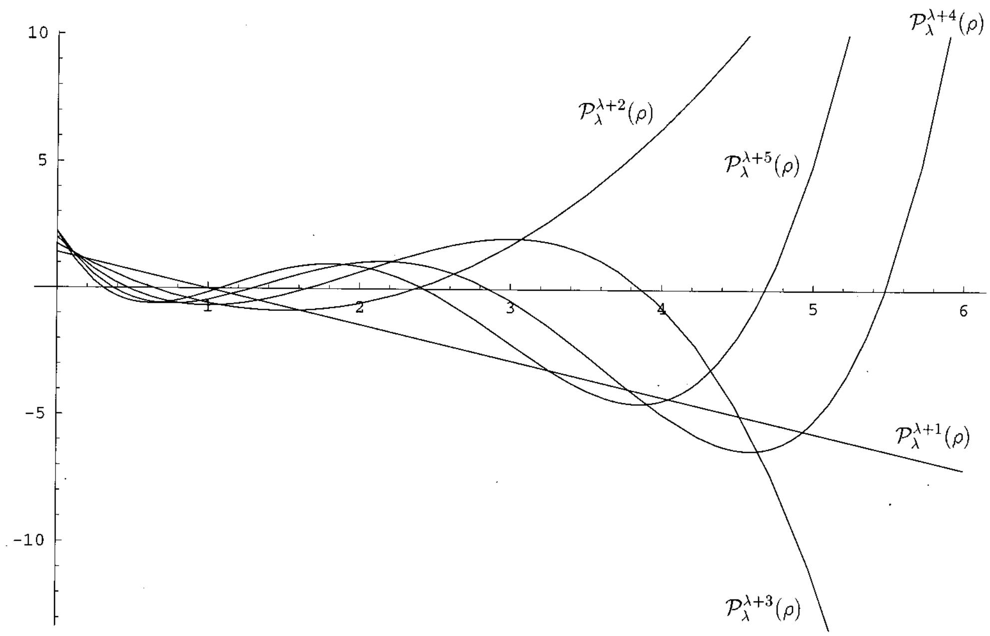

FIG. 1. We show the graph of the first polynomials $\mathcal{P}_\lambda^{\lambda+p}(\rho)/W(\rho)$, for $p=1,2,3,4,5$, $Z=1$, and $j=\frac{1}{2}$. Notice the similarity of the behavior of all polynomials near the origin.

ACKNOWLEDGMENTS

This work has benefitted from the insightful comments of L. F. Urrutia, R. Jauregui, and H. N. Núñez-Yépez and has been partially supported by CONACyT (grant No. 1343P-E9607). ALSB acknowledges the help of F. C. Minina, G. Tigga, U. Kim, Ch. Ujaya, B. Caro, M. X'Sac, M. Osita, Ch. Dochi, Ch. Mec, A. Abdul, and F. C. Bonito. Last but not least, this paper is dedicated to the memory of Q. Motita, N. Kuro, M. Mina, M. Tlahui, L. Tuga, and M. Miztli.

APPENDIX: THE POLYNOMIAL PART OF THE RADIAL EIGENFUNCTIONS

Our purpose in this appendix is to illustrate the behavior of the first few polynomials associated with the radial eigenfunction of the hydrogen atom in the base $|\lambda\mu\rangle$ and to relate this description to the old but little known results of Davis.[7,11] We explicitly calculate the first six functions of the positive eigenvalue series; such functions are always of the form of the weight factor $W(\rho)$ times a polynomial; the weight factor is $W(\rho)=\rho^s e^{-\rho}$. The polynomials are plotted in Fig. 1. The radial eigenfunctions are essentially generated from the basic relationship $\Omega_+|\omega\mu\rangle=C_\mu^+|\omega\mu+1\rangle$. In the ρ variable, the first equation in the series can be written as

$$e^{+i\xi}\left(\rho\frac{\partial}{\partial\rho}-\rho-i\frac{\partial}{\partial\xi}+\frac{1}{2}\right)e^{i\mu\xi}\mathcal{P}_\lambda^\lambda(\rho)=C_\mu^+e^{i(\mu+1)\xi}\mathcal{P}_\lambda^{\lambda+1}(\rho), \tag{A1}$$

which is just the first term in the whole ascending series $\mathcal{P}_\lambda^{\lambda+p}(\rho)=\Omega_+^p\mathcal{P}_\lambda^\lambda(\rho)$ used to recursively calculate (A2)–(A7). Note that the polynomial part of the function $\mathcal{P}_\lambda^{\lambda+p}(x)$ is always of the form $\Sigma_{i=0}^p C_i(\lambda)\rho^i$, where $C_i(\lambda)$ is also an order $(p-i)$ polynomial in λ.

Starting with $\mu=\lambda$, the first few functions in the positive series are then given by

$$\mathcal{P}_\lambda^\lambda(\rho)=W(\rho), \tag{A2}$$

$$\mathcal{P}_{\lambda}^{\lambda+\mathbf{1}}(\rho)=\sqrt{\frac{2}{\lambda}}(\lambda-\rho)W(\rho), \tag{A3}$$

$$\mathcal{P}_{\lambda}^{\lambda+\mathbf{2}}(\rho)=\frac{W(\rho)}{\sqrt{2\lambda(\lambda+1/2)}}[2\rho^2-(2\lambda+1)(2\rho-\lambda)], \tag{A4}$$

$$\mathcal{P}_{\lambda}^{\lambda+\mathbf{3}}(\rho)=\frac{W(\rho)}{\sqrt{3\lambda(\lambda+1/2)(\lambda+1)}}[-2\rho^3+6\rho^2(\lambda+1)-3\rho(2\lambda^2+3\lambda+1)+\lambda(2\lambda^2+3\lambda+1)], \tag{A5}$$

$$\mathcal{P}_{\lambda}^{\lambda+\mathbf{4}}(\rho)=\sqrt{\frac{2}{3}}\frac{W(\rho)}{\sqrt{\lambda(\lambda+1/2)(\lambda+1)(\lambda+3/2)}}\Bigg[\rho^4-2(3+2\lambda)\rho^3+3(3+5\lambda+2\lambda^2)\rho^2 -(3+11\lambda+12\lambda^2+4\lambda^3)\rho+\left(\lambda^3+3\lambda^2+\frac{11}{4}\lambda+\frac{3}{4}\right)\lambda\Bigg], \tag{A6}$$

$$\mathcal{P}_{\lambda}^{\lambda+\mathbf{5}}(\rho)=\frac{W(\rho)}{\sqrt{15\lambda(\lambda+1/2)(\lambda+1)(\lambda+3/2)(\lambda+2)}}\Bigg[-\rho^5+\left(5\lambda+\frac{29}{2}\right)\rho^4-(10\lambda^2+51\lambda+54)\rho^3 +\left(10\lambda^3+66\lambda^2+\frac{235}{2}\lambda+\frac{123}{2}\right)\rho^2-\left(5\lambda^4+37\lambda^3+\frac{319}{4}\lambda^2+\frac{257}{4}\lambda+\frac{33}{2}\right)\rho+\lambda^5 +\frac{15}{2}\lambda^4+\frac{65}{4}\lambda^3+\frac{105}{8}\lambda^2+\frac{27}{8}\lambda\Bigg]. \tag{A7}$$

The relationship of these polynomials to those used by Davis can be seen from Eqs. (13) and (14), putting $\rho=\exp x$ and introducing the function $v(\rho)$ according to $\psi_+\equiv\rho^s\exp(-\rho)v(\rho)$, and, finally, using $\mathcal{L}(\rho)\equiv v(\rho/2)$, to get

$$\rho\frac{d^2\mathcal{L}}{d\rho^2}+[(2s+1)-\rho]\frac{d\mathcal{L}}{d\rho}+\left[\frac{(s^2+\zeta^2-\tau_j^2)}{\rho}+(\mu-s-1/2)\right]\mathcal{L}=0, \tag{A8}$$

where $\mu=s+\frac{1}{2}+n'$. This equation can be regarded as a generalization to a noninteger index of the usual associated Laguerre differential equation. The equation corresponding to ψ_- can be obtained in an analogous fashion. This equation reduces to the one used by Davis if and only if $s^2=\tau_j^2-\zeta^2$, a result that just recovers the definition in Eq. (50). Now we can give the explicit relationship between our representation of the radial eigenfunctions to that used by Davis as

$$\psi_+(\rho)=\mathcal{P}_{s+1/2}^{s+1/2+n'}(\rho)=\rho^s\exp(-\rho)\mathcal{L}_{n'}^{2s}(2\rho),$$

$$\psi_-(\rho)=\mathcal{P}_{s+1/2}^{s-1/2+n'}(\rho)=\rho^s\exp(-\rho)\mathcal{L}_{n'-1}^{2s}(2\rho) \tag{A9}$$

save for normalization factors (unimportant for the point at hand), where $s=\lambda-\frac{1}{2}$, and where the generalized associated Laguerre polynomials used by Davis are defined in Eq. (67).

[1] C. Itzykson and M. Naunberger, Rev. Mod. Phys. **38**, 1 (1966).
[2] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1966).
[3] R. P. Martínez-y-Romero, J. Saldaña-Vega, and A. L. Salas-Brito, J. Phys. A **31**, L157 (1998).
[4] O. L. de Lange and R. E. Raab, *Operator Methods in Quantum Mechanics* (Clarendon, Oxford, 1991).
[5] L. Infeld and T. E. Hull, Rev. Mod. Phys. **23**, 21 (1951).
[6] H. N. Núñez-Yépez, J. L. López, D. Navarrete, and A. L. Salas-Brito, Int. J. Quantum Chem. **62**, 177 (1997).
[7] L. Davis, Jr., Phys. Rev. **56**, 186 (1939).
[8] A. Böhm, *Quantum Mechanics* (Springer-Verlag, New York, 1979), Chap. VI.

[9] A. L. Salas-Brito, R. P. Martínez-y-Romero, and H. N. Núñez-Yépez, Int. J. Mod. Phys. A **12**, 271 (1997).
[10] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (MacGraw-Hill, New York, 1964).
[11] J. Benítez, R. P. Martínez-y-Romero, H. N. Núñez-Yépez, and A. L. Salas-Brito, Phys. Rev. Lett. **64**, 1643 (1990); **65**, 2085 (1990) (Erratum).
[12] M. Moreno and A. Zentella, J. Phys. A **22**, L821 (1989).
[13] R. P. Martínez-y-Romero and A. L. Salas-Brito, J. Math. Phys. **33**, 1831 (1992).
[14] R. P. Martínez-y-Romero, M. Moreno, and A. Zentella, Phys. Rev. D **43**, 2036 (1991).
[15] M. Moshinsky and J. Szcepaniak, J. Phys. A **22**, L821 (1989).
[16] R. Haag, V. T. Lopuszanski, and M. Sohnius, Nucl. Phys. B **88**, 383 (1976).
[17] R. P. Martínez-y-Romero, H. N. Núñez-Yépez, and A. L. Salas-Brito, Rev. Mex. Fis. **35**, 617 (1989).
[18] R. P. Martínez-y-Romero, H. N. Núñez-Yépez, and A. L. Salas-Brito, Phys. Lett. A **142**, 318 (1989).
[19] C. Cisneros, R. P. Martínez-y-Romero, H. N. Núñez-Yépez, and A. L. Salas-Brito, Eur. J. Phys. **19**, 237 (1998).
[20] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms* (Springer-Verlag, Berlin, 1957).
[21] J. B. Seaborn, *Hypergeometric Functions and their Applications* (Springer-Verlag, New York, 1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Physical properties of quantum field theory measures

J. M. Mourão[a)]
Departamento de Física, Instituto Superior Técnico, Av. Rovisco País, 1096 Lisboa, Portugal

T. Thiemann[b)]
Max-Planck-Institut für Gravitationphysik, Albert-Einstein-Institut, Schlaatzweg 1, 14473 Postdam, Germany

J. M. Velhinho[c)]
Unidade de Ciências Exactas e Humanas, Universidade do Algarve, Campus de Gambelas, 8000 Faro, Portugal

(Received 27 April 1998; accepted for publication 17 December 1998)

Well known methods of measure theory on infinite dimensional spaces are used to study physical properties of measures relevant to quantum field theory. The difference of typical configurations of free massive scalar field theories with different masses is studied. We apply the same methods to study the Ashtekar–Lewandowski (AL) measure on spaces of connections. In particular we prove that the diffeomorphism group acts ergodically, with respect to the AL measure, on the Ashtekar–Isham space of quantum connections modulo gauge transformations. We also prove that a typical, with respect to the AL measure, quantum connection restricted to a (piecewise analytic) curve leads to a parallel transport discontinuous at every point of the curve.
[S0022-2488(99)00404-1]

I. INTRODUCTION

Path integrals play an important role in modern quantum field theory. The application in this context of methods of the mathematical theory of measures on infinite dimensional spaces is due to constructive quantum field theorists.[1–5] With the help of these methods important physical results have been obtained, especially concerning two and three dimensional theories. Recently analogous methods have been applied within the framework of Ashtekar nonperturbative quantum gravity to give a rigorous meaning to the connection representation,[6–10] solve the diffeomorphism constraint,[11] and define the Hamiltonian constraint.[12–18] These works all crucially depend on the use of (generalized) Wilson loop variables which have been considered for the first time, within a Hamiltonian formulation of gauge theories, by Gambini, Trias, and collaborators[19–21] and were rediscovered for canonical quantum gravity by Rovelli and Smolin.[22] In fact, instead of working in the connection representation for which Wilson loops are just convenient functions, one can use a so-called loop representation (see Ref. 23 and references therein) by means of which a rich arsenal of (formal) results was obtained which complement those obtained in the connection representation. For previous works on measures on spaces of connections see, e.g., Refs. 24 and 25.

The present paper has two main goals. The first consists of studying physical properties of the support of the path integral measure of free massive scalar fields. We use a system with a countable number of simple random variables which probe the typical scalar fields over cubes, with volume L^{d+1}, placed far away in (Euclidean) space–time. With these probes we are able to study the difference between the supports of two free scalar field theories with different masses.

[a)]Electronic mail: jmourao@galaxia.ist.utl.pt
[b)]Electronic mail: thiemann@aeipotsdam.mpg.de
[c)]Electronic mail: jvelhi@ualg.pt

The results that we obtain provide a characterization of the supports which is physically more transparent than those obtained previously.[26–30]

Our second goal consists of providing the proof of analogous results for the Ashtekar–Lewandowski (AL) measure on the space $\overline{\mathcal{A}/\mathcal{G}}$ of quantum connections modulo gauge transformations.[6,7] First we prove that the group of diffeomorphisms acts ergodically, with respect to the AL measure, on $\overline{\mathcal{A}/\mathcal{G}}$. Second we show that the AL measure is supported on connections which, restricted to a curve, lead to parallel transports discontinuous at every point of the curve.

Quantum scalar field theories have been intensively studied both from the mathematical and the physical point of view. Divergences in Schwinger functions (which, in measure theoretical terminology, are the moments of the measure) are directly related with the fact that the relevant measures are supported not on the space of nice smooth scalar field configurations, that enter the classical action, but rather on spaces of distributions. This gives a strong motivation for more detailed studies of the support of relevant measures. In physical terms this corresponds to finding ''typical (quantum) scalar field configurations,'' the set of which has measure one.

The present paper is organized as follows. In Sec. II we recall some results from the theory of measures on infinite dimensional spaces. Namely Bochner–Minlos theorems are stated and the concept of ergodicity of (semi-)group actions is introduced. In Sec. III we study properties of simple measures: The countable product of identical one-dimensional Gaussian measures and the white noise measure. In the first example we illustrate the disjointness of the support of different measures, which are invariant under a fixed ergodic action of the same group. For the white noise measure we choose random variables which probe the support and which will be also used in Sec. IV to study the support of massive scalar field theories. In Sec. V we obtain properties of the support of the AL measure that complement previously obtained results[8] and prove that the diffeomorphism group acts ergodically on the space of connections modulo gauge transformations. In Sec. VI we present our conclusions.

II. REVIEW OF RESULTS FROM MEASURE THEORY

A. Bochner–Minlos theorems

In the characterization of typical configurations of measures on functional spaces the so-called Bochner–Minlos theorems play a very important role. These theorems are infinite dimensional generalizations of the Bochner theorem for probability measures on $\mathbb{R}^N$. Let us, for the convenience of the reader, recall the latter result. Consider any (Borel) probability measure μ on $\mathbb{R}^N$, i.e., a finite measure, normalized so that $\mu(\mathbb{R}^N)=1$. The generating functional χ_μ of this measure is its Fourier transform, given by the following function on $\mathbb{R}^N$ [$\cong(\mathbb{R}^N)'$, the prime denotes the topological dual, see below]

$$\chi_\mu(\lambda)=\int_{\mathbb{R}^N} d\mu(x)e^{i(\lambda,x)}, \tag{1}$$

where $(\lambda,x)=\Sigma_{j=1}^N \lambda^j x_j$. Generating functionals of measures satisfy the following three basic conditions,

(i) Normalization: $\chi(0)=1$;
(ii) Continuity: χ is continuous on $\mathbb{R}^N$;
(iii) Positivity: $\Sigma_{k,l=1}^m c_k\overline{c_l}\chi(\lambda_k-\lambda_l)\geqslant 0$, for all $m\in\mathbb{N}$, $c_1,\ldots,c_m\in\mathbb{C}$ and $\lambda_1,\ldots,\lambda_m\in\mathbb{R}^N$.

The last condition comes from the fact that $\|f\|_\mu\geqslant 0$, for $f(x)=\Sigma_k^m c_k e^{i(\lambda_k,x)}$, where $\|\cdot\|_\mu$ denotes the $L^2(\mathbb{R}^N,d\mu)$ norm. The finite dimensional Bochner theorem states that the converse is also true. Namely, for any function χ on $\mathbb{R}^N$ satisfying (i), (ii), and (iii) there exists a unique probability measure on $\mathbb{R}^N$ such that χ is its generating functional.

Both in statistical mechanics and in quantum field theory one is interested in the so-called correlators, or in probabilistic terminology, the moments of the measure μ,

$$\langle (x_{i_1})^{p_1} \dots (x_{i_k})^{p_k} \rangle := \int_{\mathbb{R}^N} d\mu(x) (x_{i_1})^{p_1} \dots (x_{i_k})^{p_k}. \tag{2}$$

For the correlators of any order to exist the measure μ must have a rapid decay at x-infinity, in order to compensate the polynomial growth in (2) (examples are Gaussian measures and measures with compact support). In the λ-space the latter condition turns out to be equivalent to χ being infinitely differentiable (C^∞). The correlators are then just equal to partial derivatives of χ at the origin, multiplied by an appropriate power of $-i$.

Let us now turn to the infinite dimensional case. The role of the space of λ's will be played by $\mathcal{S}(\mathbb{R}^{d+1})$, the Schwarz space of C^∞-functions on (Euclideanized) space-time with fast decay at infinity. So we have the indices $\lambda(i) := \lambda^i$ replaced by $f(x)$. The space $\mathcal{S}(\mathbb{R}^{d+1})$ has a standard (nuclear) topology. Its elements are functions with regularity properties both for small and for large distances. The physically interesting measures will ''live'' on spaces dual to $\mathcal{S}(\mathbb{R}^{d+1})$. Consider the space $\mathcal{S}'(\mathbb{R}^{d+1})$ of all continuous linear functionals on $\mathcal{S}(\mathbb{R}^{d+1})$ [i.e., the topological dual of $\mathcal{S}(\mathbb{R}^{d+1})$]. This is the so-called space of tempered distributions, which includes delta functions and their derivatives, as well as functions which grow polynomially at infinity. We will consider also the even bigger space $\mathcal{S}^a(\mathbb{R}^{d+1})$ of all linear (not necessarily continuous) functionals on $\mathcal{S}(\mathbb{R}^{d+1})$. Then the simplest generalization of the Bochner theorem states that a function $\chi(f)$ on $\mathcal{S}(\mathbb{R}^{d+1})$ satisfies the following conditions,

(i′) Normalization: $\chi(0)=1$;
(ii′) Continuity: χ is continuous on any finite dimensional subspace of $\mathcal{S}(\mathbb{R}^{d+1})$;
(iii′) Positivity: $\Sigma_{k,l=1}^m c_k \overline{c_l} \chi(f_k - f_l) \geqslant 0$, for all $m \in \mathbb{N}$, $c_1,\dots,c_m \in \mathbb{C}$ and $f_1,\dots,f_m \in \mathcal{S}(\mathbb{R}^{d+1})$,

if and only if it is the Fourier transform of a probability measure μ on $\mathcal{S}^a(\mathbb{R}^{d+1})$, i.e.,

$$\chi(f) = \int_{\mathcal{S}^a(\mathbb{R}^{d+1})} d\mu(\phi) e^{i\phi(f)}. \tag{3}$$

The topology of convergence on finite dimensional subspaces of $\mathcal{S}(\mathbb{R}^{d+1})$ is unnaturally strong. Demanding in (ii′) continuity of χ with respect to the much weaker standard nuclear topology on $\mathcal{S}(\mathbb{R}^{d+1})$ yields a measure supported on the topological dual $\mathcal{S}'(\mathbb{R}^{d+1})$ of $\mathcal{S}(\mathbb{R}^{d+1})$.[31] This is the first version of the Bochner–Minlos theorem. Further refinement can be achieved if χ is continuous with respect to an even weaker topology induced by an inner product. We present a special version of this result, suitable for the purposes of the present work; for different, more general formulations see Refs. 3 and 32. Let P be a linear continuous operator from $\mathcal{S}(\mathbb{R}^{d+1})$ onto $\mathcal{S}(\mathbb{R}^{d+1})$, with continuous inverse. Suppose further that P is positive when viewed as an operator on $L^2(\mathbb{R}^{d+1}, d^{d+1}x)$ and that the bilinear form

$$\langle f_1, f_2 \rangle_{P^{1/2}} := (P^{1/2} f_1, P^{1/2} f_2), \quad f_1, f_2 \in \mathcal{S}(\mathbb{R}^{d+1}) \tag{4}$$

defines an inner product on $\mathcal{S}(\mathbb{R}^{d+1})$, where (,) denotes the $L^2(\mathbb{R}^{d+1}, d^{d+1}x)$ inner product. Let χ satisfy (i′), (iii′) and be continuous with respect to the norm associated with the inner product $\langle , \rangle_{P^{1/2}}$. Natural examples are provided by Gaussian measures μ_C with covariance C and Fourier transform

$$\chi_C(f) = e^{-(1/2)(f, Cf)}, \tag{5}$$

in which case one can take the positive operator P to be the covariance C itself. A particular case is the path integral measure for free massive scalar fields with mass m, which is the Gaussian measure with covariance

$$C_m = (-\Delta + m^2)^{-1}, \tag{6}$$

where Δ denotes the Laplacian on $\mathbb{R}^{d+1}$. In the general (not necessarily Gaussian) case let $\mathcal{H}_{P^{1/2}}(\mathcal{H}_{P^{-1/2}})$ denote the completion of $\mathcal{S}(\mathbb{R}^{d+1})$ with respect to the inner product $\langle,\rangle_{P^{1/2}}(\langle,\rangle_{P^{-1/2}})$. Then the measure on $\mathcal{S}'(\mathbb{R}^{d+1})$ corresponding to χ is actually supported on a proper subset of $\mathcal{S}'(\mathbb{R}^{d+1})$ given by an extension of $\mathcal{H}_{P^{-1/2}}$ defined by a Hilbert–Schmidt operator on $\mathcal{H}_{P^{1/2}}$. We see that in the scalar field case $\mathcal{H}_{C_m^{-1/2}}$ is the space of finite action configurations and therefore typical quantum configurations live in a bigger space. In order to define the extension mentioned above, recall that an operator H on a Hilbert space is said to be Hilbert–Schmidt if given an (arbitrary) orthonormal basis $\{e_k\}$ one has

$$\sum_{k=1}^{\infty} \langle He_k, He_k \rangle < \infty.$$

Given such a Hilbert–Schmidt operator H on $\mathcal{H}_{P^{1/2}}$, which we require to be invertible, self-adjoint and such that $H(\mathcal{S}(\mathbb{R}^{d+1})) \subset \mathcal{S}(\mathbb{R}^{d+1})$, define the new inner product $\langle,\rangle_{P^{-1/2}H}$ on $\mathcal{S}(\mathbb{R}^{d+1})$ by

$$\langle f_1, f_2 \rangle_{P^{-1/2}H} := (P^{-1/2}Hf_1, P^{-1/2}Hf_2), \quad f_1, f_2 \in \mathcal{S}(\mathbb{R}^{d+1}). \tag{7}$$

Consider $\mathcal{H}_{P^{-1/2}H}$, the completion of $\mathcal{S}(\mathbb{R}^{d+1})$ with respect to the inner product $\langle,\rangle_{P^{-1/2}H}$, and identify its elements with linear functionals on $\mathcal{S}(\mathbb{R}^{d+1})$ through the $L^2(\mathbb{R}^{d+1}, d^{d+1}x)$ inner product. Under the above conditions, the (second version of the) Bochner–Minlos Theorem states that

> **(Bochner–Minlos):** A generating functional χ, continuous with respect to the inner product $\langle,\rangle_{P^{1/2}}$, is the Fourier transform of a unique measure supported on $\mathcal{H}_{P^{-1/2}H}$, for every Hilbert–Schimdt operator H on $\mathcal{H}_{P^{1/2}}$ such that $\mathcal{S}(\mathbb{R}^{d+1}) \subset \mathrm{Ran}\, H$, $H^{-1}(\mathcal{S}(\mathbb{R}^{d+1}))$ is dense and H^{-1}: $\mathcal{S}(\mathbb{R}^{d+1}) \to \mathcal{H}_{\langle,\rangle_{P^{1/2}}}$ is continuous.

In Sec. III A we will also use an obvious adaptation of this result to the space of infinite sequences $\mathbb{R}^N$.

A common feature of the two versions of the Bochner–Minlos theorem is that they give the support as a linear subspace of the original measure space. Nonlinear properties of the support have to be obtained in a different way. In particular one can show (see Sec. III B) that the white noise measures with

$$\chi_{\sigma_1}(f) := e^{-(\sigma_1/2)(f,f)} \tag{8}$$

and

$$\chi_{\sigma_2}(f) := e^{-(\sigma_2/2)(f,f)} \tag{9}$$

have disjoint supports for $\sigma_1, \sigma_2 > 0$, and $\sigma_1 \neq \sigma_2$, while the Bochner–Minlos theorem would give the same results in both cases.

B. Ergodic actions

We review here some concepts and results from ergodic theory.[32,33] Let φ denote an action of the group G on the space M, endowed with a probability measure μ, by measure preserving transformations $\varphi_g : M \to M$, $g \in G$, i.e., $\varphi_{g_*}\mu = \mu$, $\forall g \in G$ or, equivalently, for every measurable set $A \subset M$ and for every $g \in G$, the measure of A equals the measure of the pre-image of A by φ_g. The action φ is said to be ergodic if all G-invariant sets have either measure zero or one. The fact that φ is measure preserving implies that the (right) linear representation U of G on $L^2(M, d\mu)$ induced by φ

$$(U_g\psi)(x) := \psi(\varphi_g x) \tag{10}$$

is unitary. The action φ on M is ergodic if and only if the only U_G-invariant vectors on $L^2(M, d\mu)$ are the (almost everywhere) constant functions. This follows easily from the fact that the linear space spanned by characteristic functions of measurable sets (equal to one on the set and zero outside) is dense in $L^2(M,d\mu)$. The above also applies for a discrete semi-group generated by a single (not necessarily invertible) measure preserving transformation *T*, in which case the role of the group *G* is played by the additive semi-group $\mathbb{N}$, *T* being identified with φ_1. Notice that in the latter case the linear representation on $L^2(M,d\mu)$ may fail to be unitary, although isometry still holds. For actions of $\mathbb{R}$ and $\mathbb{N}$, respectively, the following properties are equivalent to ergodicity[33]

$$\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}dt\,\psi(\varphi_t x_0)=\int_M d\mu(x)\psi(x), \tag{11a}$$

$$\lim_{N\to\infty}\frac{1}{N+1}\sum_{n=0}^{N}\psi(\varphi_n x_0)=\int_M d\mu(x)\psi(x), \tag{11b}$$

where $\varphi_n:=\varphi_1^n$ and the equalities hold for μ-almost every x_0 and for all $\psi\in L^1(M,d\mu)$. One important consequence of (11) is that if μ_1 and μ_2 are two different measures and a given action φ is ergodic with respect to both μ_1 and μ_2 then these measures must have disjoint supports [points x_0 for which (11) holds]. Recall also that the action of $\mathbb{R}$ and $\mathbb{N}$, respectively, is called mixing if for every $\psi_1,\psi_2\in L^2(M,d\mu)$ we have

$$\lim_{t\to\infty}\langle\psi_1,U_t\psi_2\rangle=\langle\psi_1,1\rangle\langle 1,\psi_2\rangle, \tag{12a}$$

$$\lim_{n\to\infty}\langle\psi_1,U_n\psi_2\rangle=\langle\psi_1,1\rangle\langle 1,\psi_2\rangle. \tag{12b}$$

It follows from (12) that every *U*-invariant $L^2(M,d\mu)$-function is constant almost everywhere and therefore every mixing action is ergodic (see Ref. 33 for details). If *M* is a linear space then (11) gives a nonlinear characterization of the support. Indeed if x_1 and x_2 are typical configurations, in the sense that (11) holds for them, then x_1+x_2 and λx_1 for $\lambda\neq 1$, are in general not typical configurations. The nonlinearity of supports is best illustrated by the action of $\mathbb{N}$ on the space of infinite sequences endowed with a Gaussian measure that we recall in the next subsection.

III. SUPPORT PROPERTIES OF SIMPLE MEASURES

A. Countable product of Gaussian measures

We will consider here the simplest case of a Gaussian measure in an infinite dimensional space. We will see however that many aspects of Gaussian measures on functional spaces can be rephrased in this simple context.

Let $M=\mathbb{R}^N$, the set of all real sequences (maps from $\mathbb{N}$ to $\mathbb{R}$)

$$x=\{x_1,x_2,\ldots\}$$

and consider on this space the measure given by the infinite product of identical Gaussian measures on $\mathbb{R}$, of mean zero and variance ρ

$$d\mu_\rho(x)=\prod_{n=1}^{\infty}e^{-x_n^2/2\rho}\frac{dx_n}{\sqrt{2\pi\rho}}. \tag{13}$$

As we saw above, an equivalent way of defining μ_ρ is by giving its Fourier transform. Let

$$\langle y,z\rangle_{\sqrt{\rho}}:=\rho(y,z),\quad y,z\in\mathcal{S}, \tag{14}$$

where $(y,z)=\Sigma_{n=1}^{\infty}y_n z_n$ and $\mathcal{S}$ is the space of rapidly decreasing sequences

$$\mathcal{S}:=\left\{y\in\mathbb{R}^{N}:\quad \sum_{n=1}^{\infty} n^k y_n^2\langle\infty,\forall k\rangle 0\right\}.$$

Then

$$\chi_\rho(y):=e^{-(1/2)\langle y,y\rangle_{\sqrt{\rho}}}=\int e^{i(y,x)}d\mu_\rho(x), \tag{15}$$

where $y\in\mathcal{S},\ x\in\mathbb{R}^{N}$.

Consider now the ergodic (in fact mixing) action φ of $\mathbb{N}$ generated by $T\equiv\varphi_1$,

$$(\varphi_1(x))_n:=x_{n+1},\quad x\in\mathbb{R}^{N} \tag{16}$$

which, as we will see, is a discrete analog of the action of $\mathbb{R}$ by translations on the quantum configuration space of a free scalar field theory. The transformation φ_1 is clearly measurable and measure preserving, since all the measures in the product (13) are equal. It is also not difficult to see that φ is mixing. In fact, invoking linearity and continuity, one only needs to show that (12b) is verified for a set of L^2-functions whose span is dense. The functions of the form $\exp(i(y,x))$, $y\in\mathcal{S}$ form such a set, and one has for them

$$\lim_{n\to\infty}\langle e^{i(y,x)},U_n e^{i(z,x)}\rangle=e^{-(1/2)(\langle z,z\rangle_{\sqrt{\rho}}+\langle y,y\rangle_{\sqrt{\rho}})}=\langle e^{i(y,x)},1\rangle\langle 1,e^{i(z,x)}\rangle,\quad \forall y,z\in\mathcal{S},$$

where, in the first and last terms, $\langle,\rangle$ denotes the $L^2(\mathbb{R}^N,d\mu_\rho)$ inner product and U denotes the isometric representation associated with φ. Thus, φ is an ergodic action with respect to μ_ρ, for any ρ. Of course, $\rho\neq\rho'$ implies $\mu_\rho\neq\mu_{\rho'}$ and one concludes that μ_ρ and $\mu_{\rho'}$ must be mutually singular (i.e., have disjoint supports). In fact, taking in (11b) $\psi(x)=\exp(i(y,x))$ for both μ_ρ and $\mu_{\rho'}$ leads to a contradiction, unless a x_0 satisfying (11b) for both μ_ρ and $\mu_{\rho'}$ cannot be found.

The aim now is to find properties of typical configurations which allow us to distinguish the supports of μ_ρ and $\mu_{\rho'}$. Unfortunately the Bochner–Minlos theorem cannot help, due to the fact that the inner products (14), that define the measures μ_ρ and $\mu_{\rho'}$ via (15), are proportional to each other and therefore the corresponding extensions in the Bochner–Minlos theorem are equal: $\mathcal{H}_{\rho^{-1/2}H}=\mathcal{H}_{\rho'^{-1/2}H}(=\mathcal{H}_H)$ for any ρ,ρ'.

Let us now find a better characterization of the support, for which the mutual singularity of μ_ρ and $\mu_{\rho'}$ becomes explicit. In order to achieve this we use a slight modification of an argument given in Ref. 1, that provides convenient sets, both of measure zero and one.

Proposition 1: Given a sequence $\{\Delta_j\}$, $\Delta_j>1$ the μ_ρ-measure of the set

$$Z_\rho(\{\Delta_j\}):=\{x:\exists N_x\in\mathbb{N}\ \text{s.t.}\ |x_n|<\sqrt{2\rho\ln\Delta n},\ \text{for}\ n\geq N_x\} \tag{17}$$

is one (zero) if $\Sigma 1/(\Delta_j\sqrt{\ln\Delta_j})$ converges (diverges).

This can be proven as follows. For fixed integer N and positive sequence $\{\Lambda_j\}$ define sets $Z_N(\{\Lambda_j\})$ by

$$Z_N(\{\Lambda_j\}):=\{x:|x_n|<\Lambda_n,\ \text{for}\ n\geq N\}. \tag{18}$$

The μ_ρ-measure of each of these sets is

$$\mu_\rho(Z_N(\{\Lambda_j\}))=\prod_{n=N}^{\infty}\mathrm{Erf}\left(\frac{\Lambda_n}{\sqrt{2\rho}}\right), \tag{19}$$

where $\mathrm{Erf}(x)=1/\sqrt{\pi}\int_{-x}^{x}e^{-\xi^2}d\xi$ is the error function. The sequence of sets $Z_N(\{\Lambda_j\})$ is an increasing sequence for fixed $\{\Lambda_j\}$, and the set $Z_\rho(\{\Delta_j\})$ defined in (17) is just their infinite union, for $\Delta_j=\exp(\Lambda_j^2/2\rho)$:

$$Z_\rho(\{\Delta_j\}) = \bigcup_{N\in\mathbb{N}} Z_N(\{\Lambda_j\}). \tag{20}$$

From σ-additivity one gets

$$\mu_\rho(Z_\rho(\{\Delta_j\})) = \lim_{N\to\infty} \mu_\rho(Z_N(\{\Lambda_j\})) \tag{21}$$

and therefore

$$\mu_\rho(Z_\rho(\{\Delta_j\})) = \exp\left(\lim_{N\to\infty} \sum_{n=N}^{\infty} \ln(\mathrm{Erf}(\sqrt{\ln \Delta_n}))\right). \tag{22}$$

Notice that the exponent is the limit of the remainder of order N of a series. Since only divergent sequences $\{\Delta_j\}$ may lead to a nonzero measure, one can use the asymptotic expression for $\mathrm{Erf}(x)$, which gives

$$\mu_\rho(Z_\rho(\{\Delta_j\})) = \exp\left(-\lim_{N\to\infty} \sum_{n=N}^{\infty} \frac{1}{\Delta_n \sqrt{\ln \Delta_n}}\right). \tag{23}$$

Depending on the sequence $\{\Delta_j\}$, only two cases are possible: either the series exists, or it diverges to plus infinity, since $\Delta_j > 1$, $\forall j$. In the first case the limit of the remainder is zero, and so the measure of $Z_\rho(\{\Delta_j\})$ will be one. If the sum diverges so does the remainder of any order and therefore the measure of $Z_\rho(\{\Delta_j\})$ is zero. □

Let us now discuss the meaning of this result. To begin with, it is easy to present disjoint sets A, A' s.t $\mu_\rho(A)=1$, $\mu_{\rho'}(A)=0$ and $\mu_\rho(A')=0$, $\mu_{\rho'}(A')=1$, for $\rho\neq\rho'$. Without loss of generality, take $\rho = a\rho'$, $a>1$. The set $Z_\rho(\{n\})$ has μ_ρ-measure zero, since $\Sigma 1/(n\sqrt{\ln n})$ diverges. But $Z_\rho(\{n\}) = Z_{\rho'}(\{n^a\})$ (see (17)) and $Z_{\rho'}(\{n^a\})$ has $\mu_{\rho'}$-measure one, since $\Sigma 1/(n^a \sqrt{\ln n^a})$ converges, for $a>1$. On the other hand the sets $Z_\rho(\{n^{1+\epsilon}\})$ are such that

$$\mu_\rho(Z_\rho(\{n^{1+\epsilon}\})) = \mu_{\rho'}(Z_\rho(\{n^{1+\epsilon}\})) = 1, \quad \forall \epsilon > 0$$

and since $Z_\rho(\{n\}) \subset Z_\rho(\{n^{1+\epsilon}\})$, $\forall \epsilon>0$, the difference sets

$$A_\rho^\epsilon := Z_\rho(\{n^{1+\epsilon}\}) \setminus Z_\rho(\{n\}) \tag{24}$$

are such that $\mu_\rho(A_\rho^\epsilon)=1$ and $\mu_{\rho'}(A_\rho^\epsilon)=0$, $\forall \epsilon>0$.

Notice that the ''square-root-of-logarithm'' nature of the support A_ρ^ϵ of μ_ρ does not mean that the typical sequence x approaches $\sqrt{2\rho \ln n}$, as $n\to\infty$. To clarify this point let us appeal to the Bochner–Minlos Theorem (see Sec. II A). Take $\mathcal{H}_{P^{1/2}}$ to be l^2, the completion of $\mathcal{S}$ with respect to the inner product (,):

$$l^2 = \left\{ y \in \mathbb{R}^{\mathbb{N}} : \sum_{n=1}^{\infty} y_n^2 < \infty \right\}.$$

Consider a vector $a = \{a_1, a_2, \ldots\} \in l^2$ and a Hilbert–Schmidt operator H_a defined by

$$(H_a(x))_n := a_n x_n, \quad x \in l^2. \tag{25}$$

Then the Bochner–Minlos Theorem leads to the conclusion that a typical sequence x in the support of the measure must satisfy

$$\sum_{n=1}^{\infty} a_n^2 x_n^2 < \infty, \tag{26}$$

which is certainly not true for a sequence behaving asymptotically like $\sqrt{2\rho\ln n}$, if we choose appropriately $a\in l^2$. However, the Bochner–Minlos theorem does not forbid the appearance of a subsequence behaving asymptotically even worse than $\sqrt{2\rho\ln n}$. Therefore proposition 1 means that in a typical sequence $\{x_n\}$ no subsequence $\{x_{n_k}\}$ can be found such that, $\forall n_k$, $|x_{n_k}| > \sqrt{2(1+\epsilon)\rho\ln n_k}$, for any arbitrary but fixed ϵ greater than zero, and that one is certainly found if ϵ is taken to be zero. But this subsequence is rather sparse, as demanded by the Bochner–Minlos Theorem; from a stochastic point of view the occurrence of values $|x_n|$ greater than $\sqrt{2\rho\ln n}$ is a rare event. The typical sequence in the support is one that is generated with a probability distribution given by the measure. The measure in this case is just a product of identical Gaussian measures in $\mathbb{R}$, so the typical sequence is one obtained by throwing a ''Gaussian dice'' an infinite number of times.

Notice that a typical $\mu_{\rho'}$ sequence can be obtained from a μ_ρ typical sequence simply by multiplying by $\sqrt{\rho'/\rho}$. This follows from the fact that the map $x\mapsto\sqrt{\rho'/\rho}x$ is an isomorphism of measure spaces $(\mathbb{R}^{\mathbb{N}},\mu_\rho)\to(\mathbb{R}^{\mathbb{N}},\mu_{\rho'})$.

B. The white noise measure

We consider now the so called ''white noise'' measure, which in some sense is the continuous analog of the previous case.[4] Again, we will look for convenient sets of measure one, in the sense given in Sec. III A. This will be achieved by a proper choice of random variables, i.e., measurable functions, which will reduce the present case to the previous discrete one.

As mentioned in Sec. II A the $d+1$-dimensional white noise is the Gaussian measure μ_σ with Fourier transform $\chi_\sigma(f)=\exp(-\sigma/2(f,f))$. Notice that here σ has dimensions of inverse mass squared.

The Euclidean group $\mathcal{E}$ acts on $\mathcal{S}(\mathbb{R}^{d+1})$, $\mathcal{S}'(\mathbb{R}^{d+1})$ and (unitarily) on $L^2(\mathcal{S}'(\mathbb{R}^{d+1}),d\mu_\sigma)$, respectively, by

$$
\begin{aligned}
(\tilde{\varphi}_g f)(x)&=f(g^{-1}x),\\
(\varphi_g\phi)(f)&=\phi(\tilde{\varphi}_{g^{-1}}f),\\
(U_g\psi)(\phi)&=\psi(\varphi_g\phi),
\end{aligned}
\tag{27}
$$

where $g\in\mathcal{E}$, gx denotes the standard action of $\mathcal{E}$ on $\mathbb{R}^{d+1}$ by translations, rotations and reflections, $f\in\mathcal{S}(\mathbb{R}^{d+1})$, $\phi\in\mathcal{S}'(\mathbb{R}^{d+1})$ and $\psi\in L^2(\mathcal{S}'(\mathbb{R}^{d+1}),d\mu_\sigma)$.

It is easy to see that a subgroup of translations in a fixed direction, say the time direction, is mixing. One just has to consider the set with dense span of L^2-functions of the form $\exp(i\phi(f))$, $f\in\mathcal{S}(\mathbb{R}^{d+1})$ and use the Riemann–Lebesgue Lemma to prove that

$$\lim_{t\to\infty}\int f(x_0+t,\ldots,x_d)g(x_0,\ldots,x_d)d^{d+1}x=0,\ \ \forall f,g\in\mathcal{S}(\mathbb{R}^{d+1}). \tag{28}$$

This implies that the measures μ_σ and $\mu_{\sigma'}$ for $\sigma\neq\sigma'$ have disjoint supports, even though the Bochner–Minlos theorem gives us for the support in both cases an extension of $L^2(\mathbb{R}^{d+1},d^{d+1}x)$ through a Hilbert–Schmidt operator (see Sec. III A). Since the choice of a complete $(\ ,\ )$-orthonormal system $\{f_n\}$ gives us an isomorphism of measure spaces

$$
\begin{aligned}
(\mathcal{S}'(\mathbb{R}^{d+1}),\mu_\sigma)&\to(\mathbb{R}^{\mathbb{N}},\mu_\rho)_{|_{\rho=\sigma}}\\
\phi&\mapsto\{\phi(f_n)\},
\end{aligned}
\tag{29}
$$

for every such basis one can find sets of the type of those found in the previous subsection and which put in evidence the mutual singularity of μ_σ and $\mu_{\sigma'}$. However, for the convenience of our analysis of free massive scalar fields in the next section, let us study the x-behavior of typical white noise configurations ϕ. Since

$$\delta_x : \phi \mapsto \phi(x) \tag{30}$$

is not a good random variable, we fix in $\mathbb{R}^{d+1}$ a family of nonintersecting cubic boxes, $\{B_j\}_{j=1}^{\infty}$, with sides of length L. Then the mean value of ϕ over B_j is a well defined random variable

$$F_{B_j} : \phi \mapsto F_{B_j}(\phi) \equiv \phi(f_j) = \frac{1}{L^{d+1}} \int_{B_j} \phi(x) d^{d+1}x, \tag{31}$$

where f_j denotes the characteristic function of the set B_j divided by the volume L^{d+1}, and the map

$$\mathcal{S}'(\mathbb{R}^{d+1}) \to \mathbb{R}^{\mathbb{N}}$$

$$\phi \mapsto \{\phi(f_j)\} \tag{32}$$

defines (by push-forward) a measure on $\mathbb{R}^{\mathbb{N}}$ of the form (13) with $\rho = \sigma/L^{d+1}$. This can be seen from the fact that

$$\int_{\mathcal{S}'(\mathbb{R}^{d+1})} d\mu_\sigma \exp\left(i \sum_{j=1}^{\infty} y_j \phi(f_j) \right) = \exp\left(-\frac{\sigma}{L^{d+1}} \frac{\Sigma_{j=1}^{\infty} y_j^2}{2} \right). \tag{33}$$

We then conclude from Sec. III A that the sets

$$W_\sigma^\epsilon := \{\phi \in \mathcal{S}'(\mathbb{R}^{d+1}) : \exists N_\phi \in \mathbb{N} \text{s.t.} |\phi(f_n)| < \sqrt{2(1+\epsilon)(\sigma/L^{d+1}) \ln n}, \quad \text{for } n \geqslant N_\phi\} \tag{34}$$

have μ_σ-measure one for every positive ϵ, and that for $\sigma' > \sigma$ an $\epsilon(\sigma') > 0$ can be found such that $\mu_{\sigma'}(W_\sigma^{\epsilon(\sigma')}) = 0$. This shows that the supremum of the mean value of ϕ over N boxes with volume L^{d+1} goes like $\sqrt{2(1+\epsilon)(\sigma/L^{d+1}) \ln N}$. A white noise with a bigger variance $\sigma' > \sigma$ has the latter behavior on larger boxes with volume $L'^{d+1} = (\sigma'/\sigma) L^{d+1}$.

IV. QUANTUM SCALAR FIELD THEORIES

A. Constructive quantum scalar field theories

Here we recall briefly some aspects of constructive quantum field theory that will be relevant for the next subsection.[2] A quantum scalar field theory on $d+1$-dimensional (flat Euclideanized) space-time is a measure μ on $\mathcal{S}'(\mathbb{R}^{d+1})$ with Fourier transform χ (generating functional or $\chi(f) = Z(-if)/Z(0)$ in theoretical physics terminology) satisfying the Osterwalder–Schrader (OS) axioms. We will be interested in the axioms which state the Euclidean invariance of the measure (OS2) and ergodicity of the action of the time translation subgroup (OS4), i.e., for $g = T_t$: $T_t(t',x) = (t'+t,x)$,

$$\lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} dt \, \psi(\varphi_{T_t} \phi_0) =_{\text{a.e.}} \int_{\mathcal{S}'(\mathbb{R}^{d+1})} \psi(\phi) d\mu(\phi). \tag{35}$$

The action of the Euclidean group on $\mathcal{S}(\mathbb{R}^{d+1})$, $\mathcal{S}'(\mathbb{R}^{d+1})$ and $L^2(\mathcal{S}'(\mathbb{R}^{d+1}), d\mu)$ is defined as in Sec. III B. So OS2 states that $\varphi_{g*}\mu = \mu$ for all elements g of the Euclidean group $\mathcal{E}$. Notice that OS2+OS4 imply that ergodicity under the subgroup of time translations is equivalent to ergodicity under the full Euclidean group. The vacua of the theory correspond to Euclidean invariant vectors on $L^2(\mathcal{S}'(\mathbb{R}^{d+1}), d\mu)$ and the axioms OS2 and OS4 imply that the vacuum is unique and given by the constant function. Examples of measures satisfying the OS axioms are the Gaussian measures μ_{C_m} corresponding to free massive quantum scalar field theories [see Eqs. (5) and (6)].

B. Support of free scalar field measures

Since the Euclidean group $\mathcal{E}$ acts ergodically on $(\mathcal{S}'(\mathbb{R}^{d+1}),\mu_{C_m})$ we conclude that the measures μ_{C_m} and $\mu_{C_{m'}}$ [see Eq. (6)] with $m\neq m'$ must have disjoint supports. Like in Sec. III B we will characterize the difference of supports in terms of the mean value of ϕ over a region with volume L^{d+1}. Before going into the details of the calculations notice that the map

$$(\mathcal{S}'(\mathbb{R}^{d+1}),\mu_\sigma)\to\mathcal{S}'(\mathbb{R}^{d+1}),\mu_{C_m})$$

$$\phi\mapsto[\sigma(-\Delta+m^2)]^{-1/2}\phi \tag{36}$$

is an isomorphism of measure spaces which maps typical white noise configurations to typical μ_{C_m}-configurations. Heuristically this means that for big distances $(\Delta x\gg 1/m)$ or small momenta the correlation imposed by the kinetic term in the action is lost and the typical configurations approach those of white noise with $\sigma=1/m^2$. Let us now obtain a formal derivation of this fact as far as the x-space behavior of typical configurations of free massive scalar fields is concerned.

Consider the same random variables F_{B_j} as in Sec. III B but, in order to eliminate the correlation, the cubic boxes B_j will be chosen centered in the points $x^j=(x_0^j,\ldots,x_d^j)=(j^2/m,0,...,0)$ and with sides parallel to the coordinate axes. The push-forward of μ_{C_m} with respect to the map

$$\mathcal{S}(\mathbb{R}^{d+1})\to\mathbb{R}^{\mathbb{N}}$$

$$\phi\mapsto\{\phi(f_j)\} \tag{37}$$

is a Gaussian measure μ_{M_m} in $\mathbb{R}^{\mathbb{N}}$ with covariance matrix M_m given by

$$(M_m)_{jl}=C_m(f_j,f_l)=\left(\frac{2}{\pi}\right)^{d+1}\frac{1}{L^{2(d+1)}}\int_{\mathbb{R}^{d+1}}d^{d+1}p\,e^{i(p_0/m)(j^2-l^2)}\frac{1}{p^2+m^2}\prod_{k=0}^{d}\frac{\sin^2(p_kL/2)}{p_k^2}. \tag{38}$$

Let us denote the (constant) value of the diagonal elements of M_m by C_m^L, i.e.,

$$C_m^L:=(M_m)_{ii}=\left(\frac{2}{\pi}\right)^{d+1}\frac{1}{L^{2(d+1)}}\int_{\mathbb{R}^{d+1}}d^{d+1}p\,\frac{1}{p^2+m^2}\prod_{k=0}^{d}\frac{\sin^2(p_kL/2)}{p_k^2}. \tag{39}$$

Proposition 2: The set

$$Y^{\epsilon}_{(m)}:=\{\phi\in\mathcal{S}'(\mathbb{R}^{d+1}):\exists N_\phi\in\mathbb{N}\ \text{ s.t.}|\phi(f_n)|<\sqrt{2(1+\epsilon)C_m^L\ln n},\ \ n\geq N_\phi\} \tag{40}$$

has μ_{C_m}-measure one for any $\epsilon>0$.

Like in Sec. III B, we will show that the μ_{M_m}-measure of the image of $Y^{\epsilon}_{(m)}$ in $\mathbb{R}^{\mathbb{N}}$ is one. To prove this we will relate the measure μ_{M_m} with a diagonal Gaussian measure of the form (13). Let $\mu_{C_m^L}$ be the Gaussian measure in $\mathbb{R}^{\mathbb{N}}$ with diagonal covariance matrix $C_m^L\delta_{ij}$.

Lemma 1: The measures μ_{M_m} and $\mu_{C_m^L}$ are mutually absolutely continuous, i.e., have the same zero measure sets.

To prove the lemma we will rely on Theorem I.23 on p. 41 of Ref. 1 (see also Theorem 10.1 on p. 160 of Ref. 32), which gives necessary and sufficient conditions for two covariances to give rise to mutually absolutely continuous Gaussian measures. In our case, since the covariance of $\mu_{C_m^L}$ is proportional to the identity, it is sufficient to show that (i) the operator $T:=M_m-C_m^L\mathbf{1}$ is

Hilbert–Schmidt and (ii) the operator M_m is bounded, positive with bounded inverse in l^2. Let us first prove that T is Hilbert–Schmidt. The matrix elements of T are $T_{ii}=0$ and $T_{jl}=(M_m)_{jl}$, for $j\neq l$. One can see from (38) that the off-diagonal elements of $(M_m)_{jl}$ are the values at the points j^2-l^2 of the Fourier transform of a real function f. Explicitly,

$$(M_m)_{jl}=\int_{\mathbb{R}} d\nu_0 e^{i\nu_0(j^2-l^2)} f(\nu_0), \tag{41}$$

where

$$f(\nu_0):=\left(\frac{2}{\pi}\right)^{d+1} \frac{1}{m^{d+3}L^{2(d+1)}} \frac{\sin^2(mL\nu_0/2)}{\nu_0^2} \cdot \int_{\mathbb{R}^d} d^d\nu \frac{1}{1+\nu_0^2+\Sigma_1^d \nu_k^2} \prod_1^d \frac{\sin^2(mL\nu_k/2)}{\nu_k^2}. \tag{42}$$

Since both f and its derivative f' and L^1, one gets for the Fourier transform $\tilde{f}$ of f:

$$|\tilde{f}(t)|=\frac{|\tilde{f}'(t)|}{|t|}, \tag{43}$$

with $\tilde{f}'$ continuous, bounded, and approaching zero at infinity. Therefore $A>0$ exists such that

$$|(M_m)_{jl}|^2 \leq \frac{A}{(j^2-l^2)^2}, \quad \text{for } j\neq l, \tag{44}$$

and therefore

$$\sum_{j,l} |T_{jl}|^2 \leq A \sum_{j\neq l} \frac{1}{(j^2-l^2)^2} < \infty, \tag{45}$$

thus proving that T is Hilbert–Schmidt. Let us now prove (ii). The operator M_m is a positive operator on l^2 since it is given by the restriction of the positive covariance C_m on $L^2(\mathbb{R}^{d+1})$ to the linearly independent system $\{f_j\}_{j\in\mathbb{N}}$. Positivity of M_m and the fact that $M_m=C_m^L\mathbf{1}+T$, T being compact (in fact Hilbert–Schmidt), implies that M_m is bounded, has a trivial kernel and therefore is invertible with bounded inverse (see, e.g., Theorem 4.25 and open mapping theorem in Ref. 34). Proposition 2 now follows, given the characterization of the support of $\mu_{C_m^L}$ one gets from Sec. III A. □

Using the fact that

$$\frac{1}{\alpha\pi} \frac{\sin^2(\alpha p)}{p^2} \tag{46}$$

tends to $\delta(p)$ when $\alpha\to\infty$, one sees from (39) that

$$\lim_{L\to\infty} L^{d+1} C_m^L = \frac{1}{m^2}. \tag{47}$$

Thus, comparing (40) with (34) we see that, in accordance with the discussion above, when averaged over widely separated large boxes ($L \gg 1/m$), the typical free field distribution approaches white noise with $\sigma=1/m^2$.

The explicitness of the mutual singularity of μ_{C_m} and $\mu_{C_{m'}}$, with $m \neq m'$ now follows easily from (24), (40), and the fact that C_m^L is a monotonous (decreasing) function of *m*.

V. PROPERTIES OF THE ASHTEKAR–LEWANDOWSKI MEASURE ON SPACES OF CONNECTIONS

A. Ergodic action of the group of diffeomorphisms

The diffeomorphism-invariant Ashtekar–Lewandowski measure μ_{AL} on the space of connections modulo gauge transformations $\mathcal{A}/\mathcal{G}$ over the manifold Σ plays an important role in rigorous attempts to find a quantization of canonical general relativity. In the present section we study properties of this measure. We show that $\mathrm{Diff}_0(\Sigma)$, the connected component of the group of diffeomorphisms of (the connected analytic manifold) Σ, acts ergodically on $\overline{\mathcal{A}/\mathcal{G}}$. In the next subsection we also obtain results concerning the properties of the support of μ_{AL}.

Let us recall the definition of μ_{AL}. We denote by a hoop $[\alpha]$ in Σ the equivalence class of (piecewise analytic) loops $\{\tilde{\alpha}\}$ based on $x_0 \in \Sigma$ such that

$$U_{\tilde{\alpha}}(A) = U_\alpha(A), \quad \forall A \in \mathcal{A}, \tag{48}$$

where $U_\alpha(A)$ denotes the holonomy corresponding to the loop α, the connection A and a chosen point in the fiber over x_0 of the (fixed) principal G-bundle over Σ and $\mathcal{A}$ is the space of all connections on this bundle. The set of all hoops forms a group $\mathcal{HG}$, called the hoop group.[7] We note that for all $G = \mathrm{SU}(N)$ with $N \geqslant 2$ the group $\mathcal{HG}$ does not depend on N nor on the principal bundle.[7] Throughout the present section we will assume G to be a compact connected Lie group. It is well known that a connection A defines through $U(A)$ a homomorphism from $\mathcal{HG}$ to the gauge group G: $[\alpha] \mapsto U_\alpha(A)$. In fact, the space $\mathcal{A}/\mathcal{G}$ is in a natural bijection with the space of all appropriately smooth homomorphisms of this type, modulo conjugation at the base point.[35–37] On the other hand, and as expected from the example of scalar fields, the measure μ_{AL} lives in a space bigger than the classical space $\mathcal{A}/\mathcal{G}$. This is the space of all, not necessarily smooth or even continuous, homomorphisms from $\mathcal{HG}$ to G modulo conjugation, denoted by $\overline{\mathcal{A}/\mathcal{G}}$.[6,7] The space $\mathcal{A}/\mathcal{G}$ of smooth classes $[A]$ was shown to be of zero measure in $\overline{\mathcal{A}/\mathcal{G}}$.[8] In the next subsection we will deepen this result.

The measure μ_{AL} is, as in the scalar field case, completely specified by giving the result of integrating the so-called cylindrical functions, which in this case are gauge-invariant functions of a finite number of parallel transports along (analytic embedded) edges

$$f(\bar{A}) = F(\bar{A}(e_1), \ldots, \bar{A}(e_n)), \tag{49}$$

where different edges may intersect only on the ends and $\bar{A} \in \bar{\mathcal{A}}$, the space of all connections realized as parallel transports in a natural way.[9,11] The measure μ_{AL} is then defined by

$$\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_{AL}\, f(\bar{A}) = \int_{G^n} dg_1 \cdots dg_n F(g_1, \ldots, g_n), \tag{50}$$

where dg is the normalized Haar measure on G.

The group $\mathrm{Diff}_0(\Sigma)$ has a natural action on $\bar{\mathcal{A}}$ which leaves μ_{AL} invariant

$$\varphi^* \bar{A}(e) := \bar{A}(\varphi \cdot e). \tag{51}$$

As we have seen in Sec. II B this action induces a unitary action of $\mathrm{Diff}_0(\Sigma)$ on $L^2(\overline{\mathcal{A}/\mathcal{G}}, d\mu_{AL})$

$$(U_\varphi f) = f(\varphi^* \bar{A}). \tag{52}$$

Consider now the so-called spin-network states[38–40] $\{T_s\}$, indexed by triples $s = (\gamma, \pi, c)$, where γ is a graph, $\pi := (\pi_1, \ldots, \pi_n)$ is a labeling of the edges of γ with nontrivial irreducible representations π_i of G and $c:=(c_1,\ldots,c_m)$ is a labeling of the vertices $v_1,\ldots,v_m$ of γ with contractors c_j, i.e., nonzero intertwining operators from the tensor product of the representations corresponding to the incoming edges at v_j to the tensor product of the representations associated with the outgoing edges. The unitary action of $\mathrm{Diff}_0(\Sigma)$ on the spin-network states is particularly simple, being given by

$$U_\varphi T_{\gamma,\pi,c}=T_{\varphi\gamma,\varphi\pi,\varphi c} \tag{53}$$

or, in short

$$U_\varphi T_s=T_{\varphi s}, \tag{54}$$

where $\varphi\gamma$ is the image of the graph γ under the diffeomorphism φ and $\varphi\pi$ and φc are the corresponding representations and contractors associated with $\varphi\gamma$. A crucial property is that the contractors can be chosen in such a way that the spin-network states form an orthonormal basis on $L^2(\overline{\mathcal{A}/\mathcal{G}},d\mu_{AL})$.[39–41] We will assume that this has been done to prove the following.

Theorem 1: The group $\mathrm{Diff}_0(\Sigma)$ acts ergodically on $\overline{\mathcal{A}/\mathcal{G}}$, with respect to the measure μ_{AL}.

To prove the theorem, we will show that the only $\mathrm{Diff}_0(\Sigma)$-invariant vector on $L^2(\overline{\mathcal{A}/\mathcal{G}},d\mu_{AL})$ is the constant function. Therefore there can be no measurable $\mathrm{Diff}_0(\Sigma)$-invariant subsets of $\overline{\mathcal{A}/\mathcal{G}}$ with measure different from zero or one.

Using the completeness of the spin-network states, every $\psi\in L^2(\overline{\mathcal{A}/\mathcal{G}},d\mu_{AL})$ can be represented in the form

$$\psi=\sum_s c_s T_s, \tag{55}$$

where no more than countably many coefficients c_s are nonzero. Since for any s and any diffeomorphism φ, $T_{\varphi s}$ belongs to the same orthonormal basis ($\varphi s\neq s \Rightarrow T_{\varphi s}\perp T_s$), we conclude that ψ in (55) is $\mathrm{Diff}_0(\Sigma)$-invariant if and only if

$$c_{\varphi s}=c_s \quad \forall \varphi\in \mathrm{Diff}_0(\Sigma). \tag{56}$$

An L^2-vector cannot have more than finitely many equal coefficients, and since for every nontrivial spin-network (with nontrivial graph and representations) there is an infinite (actually uncountable) number of (distinct) spin-networks in the orbit

$$\mathrm{Diff}_0(\Sigma)s:=\{\varphi s,\varphi\in \mathrm{Diff}_0(\Sigma)\}, \tag{57}$$

we conclude that an invariant ψ in (55) is necessarily constant almost everywhere. □

From the proof it follows that if H is a subgroup of $\mathrm{Diff}_0(\Sigma)$ s.t. the H-orbit Hs through every nontrivial spin-network s is infinite, then H acts ergodically on $\overline{\mathcal{A}/\mathcal{G}}$. So for example

Corollary 1: If $\Sigma=\mathbb{R}^N$ or $\Sigma=T^N$ the measure space $(\overline{\mathcal{A}/\mathcal{G}},\mu_{AL})$ admits the ergodic action of subgroups H of $\mathrm{Diff}_0(\Sigma)$ isomorphic to $\mathbb{Z}$.

For $\Sigma=\mathbb{R}^N$ we take the $\mathbb{Z}$-subgroup of $\mathrm{Diff}_0(\Sigma)$ generated by φ_0:

$$\varphi_0(x^1,\ldots,x^N)=(x^1+\omega^1,\ldots x^N+\omega^N), \tag{58}$$

where $(\omega^1,\ldots,\omega^N)$ is a fixed nonvanishing vector in $\mathbb{R}^N$ (recall that spin-networks are defined here only for finite graphs). If $\Sigma=T^N$ we take in (58) the vector $(\omega^1,\ldots,\omega^N)$ to have (at least) two irrational and incommensurable components, where $(x^1,\ldots,x^N)$ are now mod 1 coordinates of T^N. In both cases for every nontrivial spin network s, $\{\varphi_0^n s,n\in\mathbb{Z}\}$ contains an infinite number of distinct spin networks and so the group $\{\varphi_0^n,n\in\mathbb{Z}\}$ acts ergodically on $(\overline{\mathcal{A}/\mathcal{G}},\mu_{AL})$. □

B. Support properties

As we mentioned in the beginning of Sec. V A the space $\mathcal{A}/\mathcal{G}$ of smooth connections modulo gauge transformations is contained in a zero measure subset of $\overline{\mathcal{A}/\mathcal{G}}$, the space where the measure μ_{AL} is naturally defined.[6–8] The latter space is naturally identified with the space of all (not necessarily continuous) homomorphisms from the hoop group $\mathcal{HG}$ to the gauge group G modulo conjugation at the base point.[9] In the present subsection we deepen the result of Ref. 8 by showing that the parallel transport of a μ_{AL}-typical connection $\bar{A}\in\bar{\mathcal{A}}$ along an edge e leads to a nowhere continuous map

$$g(\cdot):[0,1]\to G. \tag{59}$$

Indeed let e: $[0,1]\to\Sigma$ be an arbitrary edge and consider for $s\in[0,1]$ the part of the edge e_s given by

$$e_s(t)=e(st),\quad t\in[0,1].$$

We then have a map

$$v:\bar{\mathcal{A}}\to G^{[0,1]}$$

$$\bar{A}\mapsto v_{\bar{A}},\quad v_{\bar{A}}(s):=\bar{A}(e_s), \tag{60}$$

where $G^{[0,1]}$ denotes the space of all maps from $[0,1]$ to G. By choosing in $G^{[0,1]}$ the standard product space topology and as algebra of measurable sets the Borel σ-algebra map v becomes measurable.

It is easy to see that, due to the properties of the Haar measure, the push-forward $v_*\mu_{AL}$ of μ_{AL}[42] to $G^{[0,1]}$ is a product of Haar measures, one for each point $s\in[0,1]$:

$$d\nu(g(\cdot))=v_* d\mu_{AL}=\prod_{s\in[0,1]} dg(s). \tag{61}$$

The main result of this subsection is the following.

Theorem 2: The measure μ_{AL} is supported on the set W of all connections $\bar{A}$ such that $v_{\bar{A}}$ is everywhere discontinuous as a map from $[0,1]$ to G.

Since $W=v^{-1}(W_1)$, where

$$W_1=\{g(\cdot)\in G^{[0,1]},\quad \text{s.t. } g(\cdot) \text{ is nowhere continuous}\}, \tag{62}$$

it is sufficient to prove that the complement of W_1,

$$W_1^c=\{g(\cdot)\in G^{[0,1]}:\exists s_0\in[0,1] \text{ s.t. } g(\cdot) \text{ is continuous at } s_0\}, \tag{63}$$

is contained in a zero ν-measure subset of $G^{[0,1]}$. Consider the sets

$$\Theta_U=\{g(\cdot)\in G^{[0,1]}:\exists I \text{ s.t. } g(I)\subset U\}, \tag{64}$$

where U is a (measurable) subset of G with $0<\mu_H(U)<1$ ($\mu_H(U)$ denoting the Haar measure of U) and I is an open subset of $[0,1]$. We need the following.

Lemma 2: For every $U\subset G$ with $0<\mu_H(U)<1$ the set Θ_U is contained in a zero measure subset of $G^{[0,1]}$.

To prove the lemma recall that the open balls

$$B(q,1/m)=\{s\in[0,1]:|s=q|<1/m\} \tag{65}$$

with rational q and integer m are a countable basis for the topology of $[0,1]$. Thus Θ_U is the countable union of the sets

$$\Theta_{U,q,m} := \{g(\cdot) \in G^{[0,1]} : g(B(q,1/m)) \subset U\}, \quad q \in \mathbb{Q}, m \in \mathbb{N}. \tag{66}$$

It is easy to construct zero measure subsets containing $\Theta_{U,q,m}$. For this fix an infinite sequence $\{s_i\}_{i=1}^{\infty} \subset B(q,1/m)$ of distinct points, $s_i \neq s_j$ for $i \neq j$. Then the set $Z = \{g(\cdot) \in G^{[0,1]} : g(s_i) \in U, i \in \mathbb{N}\}$ contains $\Theta_{U,q,m}$ and has zero measure:

$$\nu(Z) = \lim_{n \to \infty} (\mu_H(U))^n = 0.$$

From the σ-additivity of ν we conclude that, for every subset $U \subset G$ with Haar measure less than one, the set Θ_U is contained in a zero ν-measure subset. □

We can now conclude the proof of the theorem. Let us choose a number r, $0 < r < 1$ and a finite open covering $\{U_i\}_{i=1}^{k}$ of the compact group G with $\mu_H(U_i) = r$, $i = 1,...,k$. This is clearly possible since we can take a neighborhood of each point with measure r and then take a finite subcovering. Consider $g(\cdot) \in W_1^c$. Then there exists a $s_0 \in [0,1]$ such that $g(\cdot)$ is continuous at s_0. Let i_0 be such that $g(s_0) \in U_{i_0}$. Continuity implies that there exists a neighborhood I of s_0 such that $g(I) \subset U_{i_0}$ and therefore we have $W_1^c \subset \cup_{i=1}^{k} \Theta_{U_i}$. □

VI. CONCLUSION AND DISCUSSION

The knowledge of the support of measures on infinite dimensional spaces, used in quantum field theory, gives a grasp on the behavior of typical quantum field configurations associated with these measures. This may be important for a better understanding, both from the physical and mathematical points of view, of problems afflicting interacting theories like the problem of divergences.

The Bochner–Minlos theorem is very effective in capturing linear properties of the support of measures on the space $\mathcal{S}'(\mathbb{R}^{d+1})$ of quantum scalar field configurations [the support is a linear subspace of $\mathcal{S}'(\mathbb{R}^{d+1})$ spanned by configurations with a given norm finite]. It allows, e.g., to distinguish the support of the measure μC_m, corresponding to free scalar field theory with mass m, from that of the white noise measure. However, this theorem would predict the same support for the measures μ_{C_m} and $\mu_{C_{m'}}$ with $m \neq m'$ even though these measures must have disjoint supports. The latter is due to the fact that any one parameter subgroup of translations of $\mathbb{R}^{d+1}$ acts ergodically on $\mathcal{S}'(\mathbb{R}^{d+1})$ with respect to both measures. The above two claims do not contradict and rather complement each other since the subset of a ''support'' which is thick, i.e., such that any measurable subset of its complement has measure zero, is also a (finer) support for the same measure. So, to distinguish between μ_{C_m} and $\mu_{C_{m'}}$, one has to find properties of the supports which complement those given by the Bochner–Minlos theorem. Support properties of μ_{C_m} were studied in Refs. 26 and 27 (see also Refs. 28–30). Our goal in Sec. IV B consisted in showing that a simple system of random variables can be used to study the difference of supports of μ_{C_m} and $\mu_{C_{m'}}$ for large distances. This simplifies part of the results of Refs. 26 and 27 and makes them physically more transparent.

The use of the AL measure in attempts to construct a quantum theory of gravity[6–18,43–51] justifies the importance to study its properties. In Sec. V we improved the results of Ref. 8. We fixed an arbitrary edge e, or equivalently a piecewise analytic curve on Σ, and considered the random variables given by $\bar{A} \mapsto \bar{A}(e_s)$, where $\bar{A}$ is a quantum connection, e_s represents the edge e up to the value $s \in [0,1]$ of the parameter, and $\bar{A}(e_s)$ is the parallel transport corresponding to $\bar{A}$ and e_s. These random variables were suggested to the authors by Abhay Ashtekar and were motivated by studies of the volume operator in quantum gravity.[43–53] Varying s we obtain a (measurable) map v from the space of quantum connections $\bar{\mathcal{A}}$ to the space $G^{[0,1]}$ of maps,

, $g(\cdot):[0,1]\rightarrow G$. The *AL* measure becomes, by push-forward, an infinite product of Haar measures. Using this we have shown that for a typical μ_{AL}-connection $\bar{A}$ the parallel transport $\bar{A}(e_s)$ is everywhere discontinuous.

We also showed that $\mathrm{Diff}_0(\Sigma)$ acts ergodically on $\overline{\mathcal{A}/\mathcal{G}}$ with respect to the *AL* measure. The importance of this stems from the fact that in quantum gravity one has to solve the diffeomorphism constraint and therefore naively one would have to take functions on the quotient

$$\overline{\mathcal{A}/\mathcal{G}}/\mathrm{Diff}_0(\Sigma).$$

The ergodicity of the action of $\mathrm{Diff}_0(\Sigma)$ on $\overline{\mathcal{A}/\mathcal{G}}$ implies that the only solution to the diffeomorphism constraint in $L^2(\overline{\mathcal{A}/\mathcal{G}},d\mu_{AL})$ is the constant function. This explains why in Ref. 11, and even though $\overline{\mathcal{A}/\mathcal{G}}$ is compact, one had to use distributional elements to solve the diffeomorphism constraint. If the action were not ergodic one would have diffeomorphism invariant measurable subsets of $\overline{\mathcal{A}/\mathcal{G}}$ (pre-images of sets in $\overline{\mathcal{A}/\mathcal{G}}/\mathrm{Diff}_0(\Sigma)$) with μ_{AL}-measure different from zero or one. The characteristic functions of these sets would provide L^2 solutions to the diffeomorphism constraint.

ACKNOWLEDGMENTS

We would like to thank Abhay Ashtekar, Donald Marolf, and Jerzy Lewandowski for useful discussions. J.M.M. and T.T. would like to thank ESI, where a significative part of their contribution to this work was accomplished. J.M.M. and J.M.V. were in part supported by PCERN/P/FAE/1111/96, CENTRA/IST and CENTRA/Algarve. T.T. was supported in part by DOE-Grant DE-FG02-94ER25228 of Harvard University. J.M.V. was supported by grant PRAXIS XXI BD/3429/94.

[1] B. Simon, *The $P(\phi)_2$ Euclidean (Quantum) Field Theory* (Princeton University Press, Princeton, NJ, 1974).
[2] J. Glimm and A. Jaffe, *Quantum Physics* (Springer, New York, 1987).
[3] V. Rivasseau, *From Perturbative to Constructive Renormalization* (Princeton University Press, Princeton, NJ, 1991).
[4] T. Hida, H.-H. Kuo, J. Potthoff, and L. Streit, *White Noise* (Kluwer, Dordrecht, 1993).
[5] S. Albeverio, J. Jost, S. Paycha, and S. Scarlatti, *A Mathematical Introduction to String Theory* (Cambridge University Press, Cambridge, 1997).
[6] A. Ashtekar and C. J. Isham, Class. Quantum Grav. **9**, 1433 (1992).
[7] A. Ashtekar and J. Lewandowski, ''Representation Theory of Analytic Holonomy $C^\star$ Algebras,'' in *Knots and Quantum Gravity*, edited by J. Baez (Oxford University Press, Oxford, 1994).
[8] D. Marolf and J. Mourão, Commun. Math. Phys. **170**, 583 (1995).
[9] A. Ashtekar and J. Lewandowski, J. Math. Phys. **36**, 2170 (1995).
[10] J. Lewandowski, Int. J. Mod. Phys. D **3**, 207 (1994).
[11] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, J. Math. Phys. **36**, 6456 (1995).
[12] T. Thiemann, Phys. Lett. B **380**, 257 (1996).
[13] T. Thiemann, Class. Quantum Grav. **15**, 839 (1998).
[14] T. Thiemann, Class. Quantum Grav. **15**, 875 (1998).
[15] T. Thiemann, Class. Quantum Grav. **15**, 1207 (1998).
[16] T. Thiemann, Class. Quantum Grav. **15**, 1249 (1998).
[17] T. Thiemann, Class. Quantum Grav. **15**, 1281 (1998).
[18] T. Thiemann, Class. Quantum Grav. **15**, 1463 (1998).
[19] R. Gambini and A. Trias, Phys. Rev. D **22**, 1380 (1980).
[20] C. di Bartolo, F. Nori, R. Gambini, and A. Trias, Lett. Nuovo Cimento **38**, 497 (1983).
[21] R. Gambini and A. Trias, Nucl. Phys. B **278**, 436 (1986).
[22] C. Rovelli and L. Smolin, Nucl. Phys. B **331**, 80 (1990).
[23] R. Gambini and J. Pullin, *Loops, Knots, Gauge Theories, and Quantum Gravity* (Cambridge University Press, Cambridge, 1996).
[24] S. Albeverio, R. Høegh–Krohn, and H. Holden, J. Funct. Anal. **78**, 154 (1988).
[25] L. Gross, Ch. King, and A. Sengupta, Ann. Phys. (Paris) **194**, 65 (1989).
[26] Ph. Colella and O. E. Lanford, ''Sample field behavior for the free Markov random field,'' in *Constructive Quantum Field Theory, The 1973 Ettore Majorana School*, edited by G. Velo and A. Wightman (Springer-Verlag, Berlin, 1973).
[27] M. Reed and L. Rosen, Commun. Math. Phys. **36**, 123 (1974).
[28] M. Roeckner, J. Funct. Anal. **79**, 211 (1988).
[29] Z. Haba, Commun. Math. Phys. **69**, 247 (1979).
[30] S. Kusuoka, J. Fac. Sci., Univ. Tokyo, Sect. 1 **A29**, 387 (1982).

[31] The support of the measure μ on M is any (not necessarily measurable) subset $X \subset M$ such that any measurable subset Y of the complement $Y \subset X^c$ has measure zero. Then μ can be identified with the σ-additive measure μ' on X defined on sets of the form $X \cap Y$, with Y measurable on M, by $\mu'(X \cap Y) := \mu(Y)$.
[32] Y. Yamasaki, *Measures on Infinite Dimensional Spaces* (World Scientific, Singapore, 1985).
[33] Ya. G. Sinai, *Topics in Ergodic Theory* (Princeton University Press, NJ, 1994).
[34] W. Rudin, *Functional Analysis* (McGraw–Hill, New York, 1991).
[35] J. W. Barret, Int. J. Theor. Phys. **30**, 1171 (1991).
[36] J. Lewandowski, Class. Quantum Grav. **10**, 879 (1993).
[37] A. Caetano and R. F. Picken, Int. J. Math. **5**, 835 (1994).
[38] C. Rovelli and L. Smolin, Phys. Rev. D **52**, 5743 (1995).
[39] J. Baez, Adv. Math. **117**, 253 (1996).
[40] J. Baez, ''Spin-networks in Non-Perturbative Quantum Gravity,'' in *The Interface of Knots and Physics*, edited by L. Kauffman (AMS, Providence, RI, 1996).
[41] T. Thiemann, J. Math. Phys. **39**, 1236 (1998).
[42] Here μ_{AL} is considered on $\overline{\mathcal{A}}$. It is defined analogously to Eq. (50) with the difference that the functions of parallel transports are not necessarily gauge invariant (see Ref. 7). The push-forward of this measure to $\overline{\mathcal{A}/\mathcal{G}} = \overline{\mathcal{A}}/\overline{\mathcal{G}}$ gives the Ashtekar–Lewandowski measure on this space.
[43] C. Rovelli and L. Smolin, Nucl. Phys. B **442**, 593 (1995); **456**, 753 (1995).
[44] R. Loll, Phys. Rev. Lett. **75**, 3048 (1995).
[45] R. Loll, Nucl. Phys. B **460**, 143 (1996).
[46] R. Loll, Class. Quantum Grav. **14**, 1725 (1997).
[47] R. Loll, Nucl. Phys. B **500**, 405 (1997).
[48] A. Ashtekar and J. Lewandowski, ''Quantum Field Theory of Geometry,'' in *Proceedings of the March 1995 Boston Conference on the Foundations of Quantum Field Theory*.
[49] A. Ashtekar and J. Lewandowski, Class. Quantum Grav. **14**, A55 (1997).
[50] A. Ashtekar and J. Lewandowski, Adv. Theo. Math. Phys. **1**, 388 (1997).
[51] J. Lewandowski, Class. Quantum Grav. **14**, 71 (1997).
[52] R. De Pietri and C. Rovelli, Phys. Rev. D **54**, 2664 (1996).
[53] T. Thiemann, J. Math. Phys. **39**, 3347 (1998).

New generalized coherent states

K. A. Penson[a)]
Université Pierre et Marie Curie, Laboratoire de Physique Théorique des Liquides, Tour 16, 5ème étage, 4, place Jussieu, 75252 Paris Cedex 05, France

A. I. Solomon[b)]
Faculty of Mathematics and Computing, Open University, Milton Keynes MK7 6AA, United Kingdom

(Received 20 July 1998; accepted for publication 23 October 1998)

We construct a new family of boson coherent states using a specially designed function which is a solution of a functional equation $d\epsilon(q,x)/dx=\epsilon(q,qx)$ with $0\leq q\leq 1$ and $\epsilon(q,0)=1$. We use this function in place of the usual exponential to generate new coherent states $|q,z\rangle$ from the vacuum, which are normalized and continuous in their label z. These states allow the resolution of unity, and a corresponding weight function is furnished by the exact solution of the associated Stieltjes moment problem. They also permit exact evaluation of matrix elements of an arbitrary polynomial given as a normally-ordered function of boson operators. We exemplify this by showing that the photon number statistics for these states is sub-Poissonian. For any $q<1$ the states $|q,z\rangle$ are squeezed; we obtain and discuss their signal to quantum noise ratio. The function $\epsilon(q,x)$ allows a natural generation of multiboson coherent states of arbitrary multiplicity, which is impossible for the usual coherent states. For $q=1$ all the above results reduce to those for conventional coherent states. Finally, we establish a link with q-deformed bosons. © *1999 American Institute of Physics.* [S0022-2488(99)01404-8]

I. INTRODUCTION

The conventional boson coherent states are a family of collective states of the harmonic oscillator which are parameterized by a single complex number z.[1] They are sums of the eigenstates of the number operator $N=a^\dagger a$, with $[a,a^\dagger]=1$,

$$N|n\rangle=n|n\rangle,\quad (a^\dagger)^n|0\rangle=\sqrt{n!}|n\rangle,\quad \langle n|n'\rangle=\delta_{n,n'}, \tag{1}$$

with coefficients given by the exponential function. More precisely, the coherent state $|z\rangle$ is defined by

$$|z\rangle=\mathcal{N}^{-1/2}\exp(za^\dagger)|0\rangle, \tag{2}$$

$$=\mathcal{N}^{-1/2}\sum_{n=0}^{\infty}\frac{z^n}{\sqrt{n!}}|n\rangle, \tag{3}$$

with the normalization $\mathcal{N}(|z|^2)=\exp(|z|^2)$. Whereas the states $|z\rangle$ are normalized, $\langle z|z\rangle=1$, their overlap is nonvanishing,[2]

$$\langle z|z'\rangle=\exp(-\tfrac{1}{2}|z-z'|^2+i\,\mathrm{Im}(z^*z')). \tag{4}$$

[a)]Electronic mail: penson@lptl.jussieu.fr
[b)]Electronic mail: a.i.solomon@open.ac.uk

The eigenfunctions of N form a complete set and their resolution of unity is given by a sum of orthogonal projection operators, $\Sigma_{n=0}^{\infty}|n\rangle\langle n|=I$. The set $|z\rangle$ gives a resolution of unity with weight function $W(z)$ if

$$\int\int d^2z|z\rangle W(z)\langle z|=I. \tag{5}$$

Consequently, the requirement that $|z\rangle$ form a complete (in fact, an overcomplete) set is equivalent to the condition

$$\sum_{n=0}^{\infty}\left\{\frac{\pi}{n!}\int_0^{\infty}e^{-x}x^nW(x)dx\right\}|n\rangle\langle n|=I, \tag{6}$$

where $x\equiv|z|^2$, from which a (trivial) Stieltjes power-moment problem[3] results, namely, determine a positive weight function $\widetilde{W}(x)$ with $\widetilde{W}(x)\equiv e^{-x}W(x)$, such that

$$\int_0^{\infty}x^n\widetilde{W}(x)dx=\frac{n!}{\pi}\quad(n=0,1,2,\ldots,\infty), \tag{7}$$

which immediately yields a (unique) solution $W(x)=1/\pi$. The reason for the appearance of the weight function in

$$\int\int d^2z|z\rangle W(|z|^2)\langle z|=I$$

is that the states $|z\rangle$ are eigenstates of a non-Hermitian operator $a, a|z\rangle=z|z\rangle$; then in the resolution of unity for $|z\rangle$ we have a weighted sum of nonorthogonal projection operators $|z\rangle\langle z|$ as in Eq. (4). Following the prescription of Klauder,[4] a general coherent state is defined as one satisfying the following minimal set of conditions:[1]

The states $|\mu\rangle$ are coherent states if the following occurs

(1) $|\mu\rangle$ are normalizable, i.e., $\langle\mu|\mu\rangle=1$.
(2) $|\mu\rangle$ are continuous in the label μ, i.e.,

$$|\mu-\mu'|\rightarrow 0\Rightarrow\||\mu\rangle-|\mu'\rangle|^2\rightarrow 0.$$

(3) The set $|\mu\rangle$ allows a resolution of unity, i.e., there exists a weight function $W(|\mu|^2)>0$ such that

$$\int\int d^2\mu|\mu\rangle W(|\mu|^2)|\langle\mu|=I.$$

Whereas Condition 1 is evident for any allowable vector in the Hilbert space by definition, Condition 2 follows from the continuity of the overlapping factor through

$$||\mu\rangle|-|\mu'\rangle|^2=2(1-\mathrm{Re}\,\langle\mu|\mu'\rangle), \tag{8}$$

and can be realized rather easily in practice. However, Condition 3 imposes very severe restrictions on possible sets $|\mu\rangle$. In this paper we present results obtained by creating a set of states through a specially designed function $\epsilon(q,z)$. These states are coherent in the sense described above, in that the requirements 1, 2 and 3 are satisfied. One remark appears in order, concerning the weight function $W(|\mu|^2)$: if the coherent states are based on an underlying group structure [Heisenberg–Weyl, $SU(2)$ (Ref. 5) or $SU(1,1)$ (Ref. 6) coherent states, for example] then finding $W(|\mu|^2)$ is equivalent to considering the Kähler potential $F(\mu,\mu^*)$ defined by[7]

$$F(\mu,\mu^*)=\log(|\langle 0|\mu\rangle|^{-2}), \tag{9}$$

from which the weight function, up to a multiplicative constant, can be *uniquely* obtained through

$$W(|\mu|^2)=\frac{\partial^2}{\partial\mu\partial\mu^*}F(\mu,\mu^*). \tag{10}$$

The Kähler potential mirrors the geometry and the group structure of the underlying phase space. In the following we set aside any group structure and will use exclusively analytic methods.

II. NEW GENERALIZED EXPONENTIAL FUNCTION AND THE STIELTJES MOMENT PROBLEM

Consider the following functional equation for a function of the complex variable:

$$d\epsilon(q,z)/dz=\epsilon(q,qz), \quad (\epsilon(q,0)=1, \quad 0\leqslant q\leqslant 1). \tag{11}$$

When $q=1$ these are defining equations for exp(z). When $q\neq 1$, then $\epsilon(q,z)\neq\exp(z)$ and a solution analytic in some neighborhood of $z=0$ may be assumed to be given by $\epsilon(q,z)=\Sigma_{n=0}^{\infty}a_n(q)z^n$. Equation (11) produces the following recursion relation:

$$a_{n+1}(q)=a_n(q)\frac{q^n}{(n+1)}, \qquad n=1,2,\ldots,\infty, \quad a_0=1, \tag{12}$$

with solution $a_n(q)=q^{n(n-1)/2}/n!$ and

$$\epsilon(q,z)=\sum_{n=0}^{\infty} q^{n(n-1)/2}z^n/n!, \tag{13}$$

which is convergent for all z when $q\leqslant 1$. Note that $\epsilon(q,z)$ interpolates between $\epsilon(0,z)=1+z$ and $\epsilon(1,z)=\exp(z)$. The function $\epsilon(q,z)$ has an infinitely countable number of roots, of which none lies on the positive real axis.

We now define a new family of physical states $|q,z\rangle$ labeled by q and z,

$$|q,z\rangle=\{\mathcal{N}(q,|z|^2)\}^{-1/2}\epsilon(q,za^\dagger)|0\rangle, \tag{14}$$

and show that the states $|q,z\rangle$ are coherent in the general sense described above:

A. Normalization

From $\langle q,z|q,z\rangle=1$ we obtain the normalization

$$\mathcal{N}(q,|z|^2)=\sum_{n=0}^{\infty}\frac{q^{n(n-1)}}{n!}|z|^{2n}, \tag{15}$$

$$=\epsilon(q^2,|z|^2)>0. \tag{16}$$

Then the normalized state is

$$|q,z\rangle=[\epsilon(q^2,|z|^2)]^{-1/2}\sum_{n=0}^{\infty}\frac{q^{\frac{n(n-1)}{2}}}{\sqrt{n!}}z^n|n\rangle. \tag{17}$$

Note that $|q,z\rangle$ is not an eigenstate of a. For a given q we calculate the overlap,

$$\langle q,z|q,z'\rangle=[\epsilon(q^2,|z|^2)\epsilon(q^2,|z'|^2)]^{-1/2}\sum_{n=0}^{\infty} q^{n(n-1)}(z'z^*)^n/n!$$

$$=[\epsilon(q^2,|z|^2)\epsilon(q^2,|z'|^2)]^{-1/2}\epsilon(q^2,z^*z'). \tag{18}$$

For z fixed we can always choose z' such that $z^*z'=r_p(q)$, where $r_p(q)$ is the pth root of $\epsilon(q^2,z)$ $(p=1,2,\ldots)$. It follows that the state $|q,z\rangle$ can be made orthogonal to an infinity of states $|q,z'\rangle$. However, since the infinite set of roots $r_p(q)$ is still of measure zero with respect to the whole complex plane, we conclude that $\langle q,z|q,z'\rangle$ never vanishes except on a set of measure zero. [We recall that the $SU(2)$ coherent states $|\mu\rangle$ and $|\mu'\rangle$ are never orthogonal, except for $\mu'=-1/\mu^*$, since $\langle\mu|\mu'\rangle$ has only one root.[5]]

B. Continuity in z

We use Eq. (18) and substitute it into Eq. (8), which after a few standard steps leads to

$$||q,z'\rangle-|q,z\rangle|^2\leqslant|z-z'|D_q(z',z), \tag{19}$$

where

$$D_q(z',z)=\sum_{p=1}^{\infty}\frac{q^{p(p-1)}}{p!}\sum_{k=0}^{p-1}(z'^*)^k z^{(p-1-k)}, \tag{20}$$

with $D_q(z,z)\geqslant 0$. Therefore the states $|q,z\rangle$ are continuous in their label, as are the canonical coherent states.

C. Resolution of unity

In close analogy with Eq. (5) the condition for the resolution of unity for the set of states $|q,z\rangle$ is

$$\int\int d^2z|q,z\rangle W(q,|z|^2)\langle q,z|=I, \tag{21}$$

which transforms into ($x\equiv|z|^2$)

$$\sum_{n=0}^{\infty}\left[\frac{\pi q^{n(n-1)}}{n!}\int_0^{\infty}\frac{x^n W(q,x)}{\epsilon(q^2,x)}dx\right]|n\rangle\langle n|=I, \tag{22}$$

where the unknown positive weight function $W(q,x)$ is the solution of the following Stieltjes power-moment problem for $\widetilde{W(q,x)}\equiv W(q,x)/\epsilon(q^2,x)$:

$$\int_0^{\infty}x^n\widetilde{W(q,x)}dx=\frac{n!}{\pi q^{n(n-1)}}\equiv c_q(n),\quad n=1,2,\ldots,\infty. \tag{23}$$

The general theory of the Stieltjes power-moment problem[3] tells us that a condition for the solvability of Eq. (23) is the positivity of the two series $\{h_0^{(n)}\},\{h_1^{(n)}\}(n=1,2,\ldots,\infty)$ of so-called Hankel–Hadamard matrices (HH-matrices), defined by

$$h_0^{(n)}(i,j)=c_q(i+j-2)=\frac{(i+j-2)!}{q^{(i+j-2)(i+j-3)}}\quad(i,j=1,2,\ldots,n), \tag{24}$$

$$h_1^{(n)}(i,j)=c_q(i+j-1)=\frac{(i+j-1)!}{q^{(i+j-1)(i+j-2)}}\quad (i,j=1,2,\ldots,n). \tag{25}$$

It turns out that indeed all the left-upper corner determinants of HH-matrices are positive.[8] Therefore a solution of Eq. (23) exists. However, in contradistinction to the case of group-defined coherent states, the solution here is not unique. In fact, there exists an infinity of solutions, as the so-called Carleman criterion[3] indicates,

$$\sum_{n=0}^{\infty}[c_q(n)]^{-1/2n}<\infty. \tag{26}$$

We have generated a few families of solutions of Eq. (23) using the combined methods of Mellin and Laplace transforms.[9,10,11] We quote here the simplest solution,

$$\widetilde{W}(q,x)=[2q(\pi)^{3/2}\sqrt{\log(q^{-1})}]^{-1}\int_0^{\infty}\exp\left(-\frac{xy}{q}-\frac{(\log y)^2}{4\log(q^{-1})}\right)dy, \tag{27}$$

which is effectively the Laplace transform with respect to x/q of the lognormal distribution function. It satisfies $\widetilde{W}(q,0)=\pi q^{-2}$ and $\widetilde{W}(q,\infty)=0$. With the change of variable $y=q^{-2}\exp(2\sqrt{\log(q^{-1})}s)$, $\widetilde{W}(q,x)$ transforms into

$$\widetilde{W}(q,x)=[q^2(\pi)^{3/2}]^{-1}\int_{-\infty}^{\infty}\exp\left(-\frac{x}{q^3}e^{2\sqrt{\log(q^{-1})}s}-s^2\right)ds, \tag{28}$$

from which we observe that for $q\to 1$, $\widetilde{W}(q,x)\to\pi^{-1}\exp(-x)$. A detailed derivation of these results, together with a discussion of analyticity properties and a graphical representation of Eq. (27), as well as of other solutions of Eq. (23), will be given in a forthcoming publication.[11] The above results have two main consequences.

First, it follows from Eq. (21) that $\langle q,z|q,z'\rangle$ is a reproducing kernel, since

$$\int\int d^2z\langle q,z'|q,z\rangle W(q,|z|^2)\langle q,z|q,z''\rangle=\langle q,z'|q,z''\rangle. \tag{29}$$

Equivalently, any state $|q,z'\rangle$ can be expressed in terms of the others through

$$|q,z'\rangle=\int\int d^2z|q,z\rangle\langle q,z|q,z'\rangle|W(q,|z|^2). \tag{30}$$

Second, any arbitrary state $|\phi\rangle$ can be expressed in terms of states $|q,z\rangle$

$$|\phi\rangle=\int\int d^2z|q,z\rangle|W(q,|z|^2)\langle q,z|\phi\rangle, \tag{31}$$

where from Eq. (17),

$$\langle q,z|\phi\rangle=[\epsilon(q^2,|z|^2)]^{-1/2}\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}}{\sqrt{n!}}z^n\langle\phi|n\rangle. \tag{32}$$

We conclude from these considerations that the family of $|q,z\rangle$ states is a fully-fledged family of coherent states in the sense described above.

III. PHYSICAL APPLICATIONS

The utility of the general coherent states $|q,z\rangle$ in physical applications is enhanced by the observation that the following quite general matrix element can be calculated exactly for $p,r=0,1,2,\ldots,$

$$\langle q,z|(a^{\dagger})^r a^p|q,z\rangle=(z^*)^r z^p q^{(1/2)(r^2+p^2-r-p)}\frac{\epsilon(q^2,q^{(p+r)}|z|^2)}{\epsilon(q^2,|z|^2)}. \tag{33}$$

More generally, the nondiagonal matrix elements $\langle q,z|(a^{\dagger})^r a^p|q,z'\rangle$ are also readily evaluated, but will not be quoted here.

We now give three simple physical applications of the formula of Eq. (33).

A. Photon number distribution

First, the average number of photons in the state $|q,z\rangle$ is given by

$$\langle q,z|N|q,z\rangle=|z|^2\frac{\epsilon(q^2,q^2|z|^2)}{\epsilon(q^2,|z|^2)}\leqslant|z|^2 \quad (N\equiv a^{\dagger}a). \tag{34}$$

The variance $(\Delta Q)^2$ of an operator Q in the state $|\psi\rangle$ is defined by

$$(\Delta Q)^2=\langle\psi|Q^2|\psi\rangle-\langle\psi|Q|\psi\rangle^2. \tag{35}$$

We have the following expression for the number variance in the state $|q,z\rangle$:

$$(\Delta N)^2\equiv\langle q,z|N^2|q,z\rangle-\langle q,z|N|q,z\rangle^2 \tag{36}$$

$$=x\frac{\epsilon(q^2,x)\epsilon(q^2,q^2x)+x[q^2\epsilon(q^2,x)\epsilon(q^2,q^4x)-\epsilon^2(q^2,q^2x)]}{\epsilon^2(q^2,x)}, \tag{37}$$

where $x\equiv|z|^2$. Upon closer inspection one sees that the number variance is always smaller than $\langle N\rangle$. Consequently, the probability distribution generated by $\epsilon(q,x)$ is sub-Poissonian. (The ordinary coherent states are exactly Poissonian.)

B. Squeezing properties

Introduce the self-adjoint quadrature operators $X=(a+a^{\dagger})/\sqrt{2}$ and $P=(a-a^{\dagger})/i\sqrt{2}$ which satisfy $[X,P]=iI$. The variances of X and P in any state $|\psi\rangle$ satisfy the Heisenberg uncertainty relation,

$$(\Delta X)^2(\Delta P)^2\geqslant\tfrac{1}{4}. \tag{38}$$

A state $|\psi\rangle$ is called *squeezed* for the quadrature X if $(\Delta X)^2<\frac{1}{2}$ (the vacuum value). We now show that for $q<1$, there is a neighborhood of $z=0$ for which the state $|q,z\rangle$ is squeezed. To this end we observe that $(\Delta X)^2<\frac{1}{2}$ may be rewritten as

$$\langle a^2\rangle-\langle a\rangle^2+\langle a^{\dagger}a\rangle-\langle a\rangle\langle a^{\dagger}\rangle<0, \tag{39}$$

where all the averages in Eq. (39) are understood as taken in the state $|q,z\rangle$. Using Eq. (33) gives the following condition for squeezing ($x\equiv|z|^2$):

$$(1+q)\epsilon(q^2,q^2x)\epsilon(q^2,x)-2\epsilon^2(q^2,qx)<0. \tag{40}$$

We note that the inequality (40) is always satisfied for $x=0$, as it then reduces to $q<1$, which is true by hypothesis. From the continuity of $\epsilon(q,x)$ for fixed q there is therefore a neighborhood of $x=0$ for which the inequality (40) is satisfied; that is, there is a constant $\delta(q)>0$ such that squeezing in X occurs for $0<x<\delta(q)$.

C. Signal-to-quantum noise ratio

The last physical quantity calculated in the state $|q,z\rangle$ will be the signal-to-quantum noise ratio σ, which is defined by

$$\sigma=\frac{\langle X\rangle^2}{(\Delta X)^2}. \tag{41}$$

Repeated use of Eq. (33) yields the result

$$\sigma(q,x)=2x\epsilon^2(q^2,qx)/\{x[(1+q)\epsilon(q^2,x)\epsilon(q^2,q^2x)-2\epsilon^2(q^2,qx)]+\tfrac{1}{2}\epsilon^2(q^2,x)\}. \tag{42}$$

For the standard coherent states ($q=1$) the quantity $\sigma(1,x)$ attains the value $4N_s$, where N_s is the number of photons in the signal, $N_s=N_s(1,x)$, where $N_s(q,x)=\langle q,z|a^\dagger a|q,z\rangle$. In our case we may show that in a positive neighborhood of x,

$$\frac{\sigma(q,x)}{4N_s(q,x)}=1+(1-q^2)x\cdots, \tag{43}$$

which is greater than 1 for $q^2<1$, thus improving on the usual coherent state value.

The classical result of Yuen,[12,13] valid for any state whatsoever, implies that

$$\sigma(q,x)\leqslant 4N_s(q,x)(N_s(q,x)+1), \tag{44}$$

where the upper value is attained by a conventional squeezed state.[14] Calculations on our coherent state $|q,z\rangle$, which as we have seen above is a squeezed state, confirm this upper bound.

IV. TIME EVOLUTION

Since the state $|q,z\rangle$ is not an eigenstate of the Hamiltonian $H=\omega a^\dagger a$, it is natural to analyze its time dependence, obtained from the time-dependent Schrödinger equation

$$i\frac{\partial}{\partial t}|q,z;t\rangle=H|q,z;t\rangle \quad (\hbar=1). \tag{45}$$

The time-dependent state $|q,z;t\rangle$ is obtained by acting on $|q,z\rangle\equiv|q,z;0\rangle$ by the time-evolution operator $U(t)=\exp(-iHt)$. That is,

$$\begin{aligned}|q,z;t\rangle&=\exp(-iHt)|q,z;0\rangle=\exp(-iHt)[\epsilon(q^2,|z|^2)]^{-1/2}\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}}{\sqrt{n!}}z^n|n\rangle\\ &=[\epsilon(q^2,|z(t)|^2)]^{-1/2}\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}}{\sqrt{n!}}z(t)^n|n\rangle=|q,z(t)\rangle,\end{aligned} \tag{46}$$

where $z(t)\equiv z\exp(-i\omega t)$. The eigenproperties of the set $|n\rangle$ have been used, $\exp(-iHt)|n\rangle=\exp(-in\omega t)|n\rangle$. The corresponding evolution of the propagator,

$$\langle q,z|q,z';t\rangle=\langle q,z|q,\exp(-i\omega t)z'\rangle=\langle q,\exp(i\omega t)z|q,z'\rangle, \tag{47}$$

indicates that the coherent state $|q,z'\rangle$ goes over into another coherent state described by $z'(t)\equiv z'\exp(-i\omega t)$ under the action of the time-evolution operator; this latter state is simply the time-dependent state of a *classical* harmonic oscillator. Then, very much as for standard coherent states, the set of all states $|q,z\rangle$ is invariant under time evolution with the Hamiltonian H. The state $|q,z\rangle$ remains as close as possible to its classical analog.

V. RELATION TO DEFORMED BOSONS

The reader will have noticed by now that the states $|q,z\rangle$ share practically all the properties of the usual coherent states $|z\rangle \equiv |1,z\rangle$ except one: $|q,z\rangle$ is not an eigenstate of a, whereas $|z\rangle$ is, $a|z\rangle = z|z\rangle$. Although this did not impair our ability to apply the new states $|q,z\rangle$ in actual calculations, it is legitimate to ask the following question: is the state $|q,z\rangle$ an eigenstate of some non-Hermitian operator b ? The answer to that question is affirmative, with the operator b and its Hermitian conjugate playing the role of *deformed bosons*; in other words, one has to deform the basic commutation relations in order to introduce the so-called deformed bosons. It is natural in our context to parametrize such a deformation by the parameter q, i.e., $b \equiv b(q)$ and $b^\dagger \equiv b^\dagger(q)$ with $b(1) = a$. The study of deformed commutation relations and their representations (more specifically as *quantum groups*) has been the subject of much activity in recent years.[15,16] Of course, there exists an infinite number of possible deformations. All of them can be systematically investigated using the function $[x]_q$ (''box'' x),[17] which enters the deformed commutator as follows:

$$b(q)b^\dagger(q) - b^\dagger(q)b(q) = [N+1]_q - [N]_q, \tag{48}$$

where $N = a^\dagger a$, $N|n\rangle = n|n\rangle$. Furthermore, it is possible to assume that $b^\dagger(q)$ and $a^\dagger$ (along with their conjugates) act on the same Fock space with basis $\{|n\rangle\}$. From Eq. (48), it follows that

$$b(q)|n\rangle = \sqrt{[n]_q}|n-1\rangle,$$

so the operators $b(q)$ may be thought of as functions of a according to $b(q) = \sqrt{[N+1]_q/(N+1)}a$, etc.

For historical reasons some specific forms of the function $[N]_q$ gained a particular popularity.[18–20] The general coherent states $|q,z\rangle$ of this paper are related naturally to a specific ''box'' function, and a particular set of deformed commutation relations. Defining $[n]_q \equiv nq^{2(1-n)}$, then

$$[n]_q! = [n]_q[n-1]_q \cdots [1]_q = n!/q^{n(n-1)},$$

and the state $|q,z\rangle$ of Eq. (17) may be rewritten as

$$|q,z\rangle = [\epsilon(q^2,|z|^2)]^{-1/2} \sum_{n=0}^{\infty} \frac{z^n}{\sqrt{[n]_q!}}|n\rangle. \tag{49}$$

This shows that $|q,z\rangle$ is an eigenstate of a q-deformed boson $b(q)$ satisfying

$$b(q)|q,z\rangle = z|q,z\rangle, \tag{50}$$

where, from Eq. (48), the operators $b(q)$ satisfy the following commutation relations:

$$b(q)b^\dagger(q) - (1/q^2)b^\dagger(q)b(q) = q^{-2N}. \tag{51}$$

This gives a new q-deformation, different from the ones previously used. Note that the coherent state of Eq. (14) or Eq. (49) corresponds to a new q-exponential function $\epsilon(q,x) = \Sigma x^n/[n]!$, where $[n]$ is defined above, and $[n]! = [n][n-1]\cdots[1], [0]! = 1$. This is not one of the family of q-exponential functions considered, for example, by Exton,[21] but may be thought of as a member of a 3-parameter family,

$$E(p,q,\lambda;x)\equiv\sum\frac{x^n q^{\lambda n(n-1)}}{[n]_p!};$$

namely, $\epsilon(q,x)=E(1,q,1;x)$ with $[n]_p\equiv(1-p^n)/(1-p)$ here being the traditional definition.

VI. MULTIBOSON STATES

The use of the function $\epsilon(q,x)$ can be extended to investigate multiboson coherent states. First, recall that if we set out to create k excitations of the harmonic oscillator at the time, we are led to use, for fixed k,

$$(a^\dagger)^{kn}|0\rangle=\sqrt{(kn)!}|kn\rangle,\quad n=0,1,\ldots,\infty. \tag{52}$$

Using the conventional exponential function to create such a k-state leads us to consider, for complex z,

$$|k,z\rangle=\mathcal{N}^{-1/2}(k,|z|^2)\sum_{n=0}^{\infty}\frac{[z(a^\dagger)^k]^n}{n!}|0\rangle=\mathcal{N}^{-1/2}(k,|z|^2)\sum_{n=0}^{\infty}\frac{z^n\sqrt{(kn)!}}{n!}|kn\rangle, \tag{53}$$

where the normalization is

$$\mathcal{N}(k,|z|^2)=\sum_{n=0}^{\infty}\frac{|z|^{2n}(kn)!}{(n!)^2}. \tag{54}$$

(The functions $|k,z\rangle$ are not of only theoretical interest: $|2,z\rangle$ was the first example of an exponentially-generated state displaying squeezing.[14]) Unfortunately, the power series in Eq. (54) diverges for $k>2$.[22] Therefore no multiboson coherent states for $k>2$ exist when generated by the usual exponential, unless a number-dependent convergence factor is introduced.[23] We now turn to $\epsilon(q,z)$ and define

$$|q,k,z\rangle=\mathcal{N}^{-1/2}(q,k,|z|^2)\epsilon(q,z(a^\dagger)^k)|0\rangle=\mathcal{N}^{-1/2}(q,k,|z|^2)\sum_{n=0}^{\infty}\frac{z^n\sqrt{(kn)!}q^{n(n-1)/2}}{n!}|kn\rangle, \tag{55}$$

where the normalization is now

$$\mathcal{N}(q,k,|z|^2)=\sum_{n=0}^{\infty}\frac{|z|^{2n}(kn)!q^{n(n-1)}}{(n!)^2}, \tag{56}$$

and for $q<1$ is a convergent series for all $|z|^2$ and for $k=1,2,\ldots$. This allows us to define normalizable multiboson states of any multiplicity k. However, the $|q,k,z\rangle$ do not form a complete set as their Stieltjes problem does not possess a positive weight function. However, such states can be understood to be ''complete'' in a different sense: there is a way to extract every eigenstate, say $|k,m\rangle$, from a state of the type Eq. (55) either by taking derivatives at $z=0$, or by dividing by $m+1$ and performing a contour integral about $z=0$. We refer to Ref. 24 for a description of such an approach.

VII. CONCLUSIONS

We have shown how to define new coherent states by the introduction of a variant of the exponential function $\epsilon(q,x)$. We have shown these states to be coherent in that we have explicitly given the weight function which satisfies the associated Stieltjes power-moment problem and thus the resolution of unity property. The analytic behavior of this new exponential is such as to enable ready evaluation of matrix elements of operators in these new coherent states, and we have exemplified this by evaluating their squeezing and related physical properties, as well as their time dependence. Further, the stronger convergence behavior of $\epsilon(q,x)$ allows the definition of multi-boson coherent states, not permitted in the convention exponential case. Finally, we have shown that the new function may be considered as the q-exponential associated with a simple deformation of the boson commutation relations, and the related coherent states as the q-boson eigenstates. As a postscript, we mention another construction of coherent states which follows what might be termed an inverse path as compared to our construction of Eqs. (23) and (27). Any positive function having well-behaved moments can serve as a weight function for a family of coherent states. In this way one may define new and interesting coherent states with particularly interesting geometrical properties.[25,26]

The new coherent state defined in this paper possesses a combination of strong convergence properties and easy evaluability which make it a good candidate for the description of solvable model systems.

ACKNOWLEDGMENTS

The authors are grateful to L. Haddad for a significant contribution to this work, H. Widom for providing an ingenious proof of the positivity of the Hankel–Hadamard matrices of Eqs. (24) and (25), and A. Comtet, D. Zeilberger, and M. Ismail for interesting discussions.

[1] J. R. Klauder and B.-S. Skagerstam, *Coherent States* (World Scientific, Singapore, 1985).
[2] J. R. Klauder and E. C. G. Sudarshan, *Fundamentals of Quantum Optics* (Benjamin, New York, 1968).
[3] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis* (Oliver and Boyd, London, 1965).
[4] J. Klauder, J. Math. Phys. **4**, 1058 (1963).
[5] J. M. Radcliffe, J. Phys. A **4**, 313 (1971).
[6] A. I. Solomon, J. Math. Phys. **12**, 390 (1971).
[7] E. A. Kochetov, J. Phys. A **31**, 4478 (1998); Phys. Lett. A **217**, 65 (1996).
[8] H. Widom (private communication).
[9] I. N. Sneddon, *The Use of Integral Transforms* (McGraw-Hill, New York, 1974).
[10] D. J. Fernandez, V. Hussin, and L. M. Nieto, J. Phys. A **27**, 3547 (1994).
[11] L. Haddad, K. A. Penson, and A. I. Solomon (to be published).
[12] H. P. Yuen, Phys. Lett. A **56**, 105 (1976).
[13] A. I. Solomon, Phys. Lett. A **188**, 215 (1994).
[14] J. N. Hollenhorst, Phys. Rev. D **19**, 1669 (1979).
[15] M. Chaichian and A. Demichev, *Introduction to Quantum Groups* (World Scientific, Singapore, 1996).
[16] T. T. Truong, J. Phys. A **27**, 3829 (1994),
[17] A. I. Solomon, Phys. Lett. A **196**, 29 (1994).
[18] M. Arik and D. D. Coon, J. Math. Phys. **17**, 524 (1976).
[19] A. J. MacFarlane, J. Phys. A **22**, 4581 (1989).
[20] L. C. Biedenharn, J. Phys. A **22**, L873 (1989).
[21] H. Exton, *q-Hypergeometric Functions and Applications* (Chichester-Harwood, London, 1983).
[22] R. A. Fisher, M. M. Nicto, and V. D. Sandberg, Phys. Rev. D **29**, 1107 (1984).
[23] J. Katriel, A. I. Solomon, G. d'Ariano, and M. Rasetti, J. Opt. Soc. Am. B **4**, 1728 (1987).
[24] A. Vourdas, Phys. Rev. A **54**, 4544 (1996).
[25] J. Klauder, Mod. Phys. Lett. A **8**, 1735 (1993).
[26] J. Klauder, Ann. Phys. (N.Y.) **237**, 147 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Waves and particles in Kaluza–Klein theory

W. N. Sajko[a] and P. S. Wesson[b]
Department of Physics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

H. Liu
Department of Physics, Dalian University of Technology, Dalian, People's Republic of China

(Received 17 June 1998; accepted for publication 17 November 1998)

We examine three overlapping problems in the application of five-dimensional (5D) manifolds to physics. First, we linearize the 5D theory along the lines of the four-dimensional (4D) theory, using the harmonic gauge condition. The resulting wave equations have sources, and can in principle describe gravitons and scalar particles with finite masses, but the natural choice of gauge parameters makes both massless. Second, we generalize the 5D metric by including separate conformal factors on its 4D and extra parts. Then the 5D harmonic gauge gives back the 4D harmonic gauge of gravitational waves and the Lorentz gauge of electromagnetic waves, but both particles are massless. Third, we again take a conformally rescaled metric, but rewrite the field equations in a novel manner. This allows us to interpret the finite masses of ordinary particles in terms of the wavelengths associated with 4D spaces embedded in a 5D space.
[S0022-2488(99)01304-3]

I. INTRODUCTION

Five-dimensional (5D) Kaluza–Klein theory is a generalization of 4D Einstein theory and is commonly regarded as a unified theory of the gravitational, electromagnetic, and scalar field whose quantum analogs are the spin-2 graviton, spin-1 photon, and the spin-0 scalar (for a recent review of Kaluza–Klein theory see Ref. 1). As a possible bridge between classical field theory and quantum theory, there has in recent years been significant work on wavelike solutions on 5D manifolds which might be interpreted as four-dimensional (4D) particles.[2–5] In the present work we wish to give fairly generic accounts of three overlapping problems in this area.

In Sec. II we take a new look at the case where the 5D metric can be written as a flat piece plus a perturbation. The linearized 5D field equations can be made algebraically tractable if one chooses the 5D harmonic gauge. This gives a wave equation for the 4D components of the perturbation and a Klein–Gordon type equation for the extra part of the perturbation. These wave equations in general have sources, and can represent gravitons and scalar particles with finite masses. However, both masses go to zero for what should be considered the most natural choice of gauge parameters. In Sec. III, we generalize the metric by conformally rescaling its 4D and 5D parts separately. We consider the full (nonlinearized) field equations and again choose the 5D harmonic gauge. For a certain choice of constant parameters involved in the conformal rescaling, the general equations reduce to the 4D harmonic gauge and the Lorentz gauge. In Sec. IV we use the results derived in previous sections to explore the relationship between 5D field theory and 4D particle mass following lines suggested by previous work.[6–8] We find that particle masses may be interpreted as a wavelength in the 4D subspace of a 5D space.

[a]Electronic mail: wnsajko@astro.uwaterloo.ca
[b]Electronic mail: wesson@astro.uwaterloo.ca

II. WAVES IN LINEARIZED 5D THEORY

To linearize 5D gravity we assume as in the 4D problem that the metric can be written as

$$\hat{g}_{AB}=\hat{\eta}_{AB}+\hat{h}_{AB}, \tag{2.1}$$

where $\hat{h}_{AB}$ is viewed as a small perturbation from flat 5D Minkowski space. Here and in what follows we use careted 5D objects and uncareted 4D ones. Also, upper-case Latin letters run from 0 to 4, Greek letters run from 0 to 3, and the block diagonal form for the 5D Minkowski space is $\hat{\eta}_{AB}=(\hat{\eta}_{\alpha\beta},\epsilon)$ where ϵ can be spacelike (-1) or timelike $(+1)$. The inverse of the above metric to $O(\hat{h})$ is

$$\hat{g}^{AB}=\hat{\eta}^{AB}-\hat{h}^{AB}, \tag{2.2}$$

and the Christoffel symbols to $O(\hat{h})$ are

$$\hat{\Gamma}^{A}_{BC}=\tfrac{1}{2}(\partial_B\hat{h}^{A}{}_{C}+\partial_C\hat{h}^{A}{}_{B}-\partial^{A}\hat{h}_{BC}). \tag{2.3}$$

Here $\partial_A=\partial/\partial x^{A}$ and the flat-space metric is responsible for raising and lowering indices $(\hat{h}^{A}{}_{C}=\hat{\eta}^{AB}\hat{h}_{BC}=\hat{h}_{C}{}^{A})$. We take as our starting point the 15 field equations representing the 5D vacuum:

$$\hat{R}_{AB}=0. \tag{2.4}$$

These to $O(\hat{h})$ give

$$\hat{R}_{AB}=\tfrac{1}{2}(\partial_A\partial_C\hat{h}^{C}{}_{B}+\partial_B\partial_C\hat{h}^{C}{}_{A}-\hat{\Box}\hat{h}_{AB}-\partial_A\partial_B\hat{h}^{C}{}_{C}), \tag{2.5}$$

where the trace of $\hat{h}_{AB}$ is defined as

$$\hat{h}\equiv\hat{h}^{A}{}_{A}=\hat{\eta}^{AB}\hat{h}_{AB} \tag{2.6}$$

and the 5D box operator is defined as

$$\hat{\Box}\equiv\hat{\eta}^{AB}\partial_A\partial_B. \tag{2.7}$$

In order to reduce the algebraic complexity of the 5D Ricci tensor we now choose the harmonic gauge:

$$\hat{\Gamma}^{C}\equiv\hat{g}^{AB}\hat{\Gamma}^{C}_{AB}=0. \tag{2.8}$$

This to $O(\hat{h})$ gives

$$\partial_A\hat{h}^{AC}=\tfrac{1}{2}\partial^{C}\hat{h}, \tag{2.9}$$

and the 5D Ricci tensor reduces to

$$\hat{R}_{AB}=\tfrac{1}{2}\hat{\Box}\hat{h}_{AB}. \tag{2.10}$$

If we now impose (2.4), we obtain a wave equation for the $\hat{h}_{AB}$, namely

$$\hat{\Box}\hat{h}_{AB}=0. \tag{2.11}$$

The trace of this gives

$$\hat{R}=\tfrac{1}{2}\hat{\Box}\hat{h}=0. \tag{2.12}$$

However, it is sometimes more convenient to use the tensor

$$\hat{\psi}_{AB}\equiv\hat{h}_{AB}-\tfrac{1}{2}\hat{\eta}_{AB}\hat{h}, \tag{2.13}$$

and by using (2.10) and (2.12) we see that $\hat{\psi}_{AB}$ satisfies the wave equation

$$\hat{G}^{AB}=\tfrac{1}{2}\hat{\Box}\hat{\psi}^{AB}=0. \tag{2.14}$$

Now the harmonic gauge (2.8) applied to $\hat{\psi}_{AB}$ becomes

$$\partial_A\hat{\psi}^{AB}=0. \tag{2.15}$$

Since this equation is a manifestation of the (noncovariant) harmonic gauge, we cannot demand that it be invariant under arbitrary 5D coordinate transformations. Instead what we do demand (as in 4D theory) is that (2.15) be invariant to $O(\hat{\xi})$ for the transformation

$$x^A\rightarrow x'^A=x^A+\hat{\xi}^A, \tag{2.16}$$

where $\hat{\xi}^A$ is an infinitesimal vector. Under this transformation $\hat{h}_{AB}, \hat{\psi}_{AB}$ and the gauge condition transform as

$$\hat{h}_{AB}\rightarrow\hat{h}'_{AB}=\hat{h}_{AB}-\partial_B\hat{\xi}_A-\partial_A\hat{\xi}_B\,, \tag{2.17}$$

$$\hat{\psi}_{AB}\rightarrow\hat{\psi}'_{AB}=\hat{\psi}_{AB}-\partial_B\hat{\xi}_A-\partial_A\hat{\xi}_B+\hat{\eta}_{AB}\partial_C\hat{\xi}^C, \tag{2.18}$$

$$\partial_A\hat{\psi}^{AB}\rightarrow\partial'_A\hat{\psi}'^{AB}=\partial_A\hat{\psi}^{AB}-\hat{\Box}\hat{\xi}^B. \tag{2.19}$$

The invariance of (2.15) under the gauge transformation (2.16) then holds provided the following wave equation for $\hat{\xi}^B$ is satisfied:

$$\hat{\Box}\hat{\xi}^B=0. \tag{2.20}$$

The transformation (2.16) represents the only gauge freedom left in the theory and is important in deducing the actual degrees of freedom of the metric $\hat{h}_{AB}$. Now $\hat{h}_{AB}$ has fifteen independent components. But we have used five coordinate degrees of freedom in (2.16) and imposed five constraints through (2.8). We therefore conclude that $\hat{h}_{AB}$ has only $15-5-5=5$ degrees of freedom left. These correspond to two degrees of freedom for the gravitational field, two degrees for the electromagnetic field, and one degree for the scalar field.

We now turn to the field equations and look at their reduction to 4D quantities and interpret their meaning. The field equations for 5D linearized theory are

$$\hat{R}^{AB}=\tfrac{1}{2}\hat{\Box}\hat{h}^{AB}=0$$

or

$$\hat{G}^{AB}=\tfrac{1}{2}\hat{\Box}\hat{\psi}^{AB}=0, \tag{2.21}$$

$$\partial_A\hat{h}^{AB}=\tfrac{1}{2}\partial^B\hat{h}$$

or

$$\partial_A \hat{\psi}^{AB} = 0, \tag{2.22}$$

$$\hat{\Box} \hat{\xi}^A = 0. \tag{2.23}$$

Looking at the $(\alpha\beta)$, $(\alpha 4)$, and (44) components of the right part of (2.21) we obtain

$$G^{\alpha\beta} = \tfrac{1}{2}\Box_M \hat{\psi}^{\alpha\beta} = \tfrac{1}{2}(-2\,\hat{\eta}^{4\gamma}\partial_4\partial_\gamma\hat{\psi}^{\alpha\beta} - \hat{\eta}^{44}\partial_4^2\hat{\psi}^{\alpha\beta}), \tag{2.24}$$

$$\tfrac{1}{2}\Box_M \hat{\psi}^{4\alpha} = \tfrac{1}{2}(-2\,\hat{\eta}^{4\gamma}\partial_4\hat{\psi}^{4\alpha} - \hat{\eta}^{44}\partial_4^2\hat{\psi}^{4\alpha}), \tag{2.25}$$

$$\tfrac{1}{2}\Box_M \hat{\psi}^{44} = \tfrac{1}{2}(-2\,\hat{\eta}^{4\gamma}\partial_4\partial_\gamma\hat{\psi}^{44} - \hat{\eta}^{44}\partial_4^2\hat{\psi}^{44}). \tag{2.26}$$

The β and fourth components of the constraint equations yield

$$\partial_\alpha \hat{\psi}^{\alpha\beta} = -\partial_4 \hat{\psi}^{4\beta}, \tag{2.27}$$

$$\partial_\alpha \hat{\psi}^{4\alpha} = -\partial_4 \hat{\psi}^{44}. \tag{2.28}$$

And the wave equation for $\hat{\xi}^A$ gives

$$\Box_M \hat{\xi}^\alpha = -2\,\hat{\eta}^{4\beta}\partial_4\partial_\beta\hat{\xi}^\alpha - \hat{\eta}^{44}\partial_4^2\hat{\xi}^\alpha, \tag{2.29}$$

$$\Box_M \hat{\xi}^4 = -2\,\hat{\eta}^{4\beta}\partial_\beta\partial_4\hat{\xi}^4 - \hat{\eta}^{44}\partial_4^2\hat{\xi}^4. \tag{2.30}$$

In the above, the 5D box operator was expanded as

$$\hat{\Box} \equiv \hat{\eta}^{AB}\partial_A\partial_B = \hat{\eta}^{\alpha\beta}\partial_\alpha\partial_\beta + 2\,\hat{\eta}^{4\alpha}\partial_4\partial_\alpha + \hat{\eta}^{44}\partial_4^2, \tag{2.31}$$

and using the 4D Minkowski box operator $\Box_M \equiv \hat{\eta}^{\alpha\beta}\partial_\alpha\partial_\beta$ (2.31) reduces to

$$\hat{\Box} = \Box_M + 2\,\hat{\eta}^{4\alpha}\partial_4\partial_\alpha + \hat{\eta}^{44}\partial_4^2. \tag{2.32}$$

The most important of the sets of equations given above is (2.24) since it describes an induced energy–momentum tensor from the definition $G^{\alpha\beta} \equiv T^{\alpha\beta}$ (with $8\pi G = c = \hbar = 1$). Its explicit form and the wave equation it satisfies are

$$T^{\alpha\beta} \equiv \tfrac{1}{2}(-2\,\hat{\eta}^{4\gamma}\partial_4\partial_\gamma\hat{\psi}^{\alpha\beta} - \hat{\eta}^{44}\partial_4^2\hat{\psi}^{\alpha\beta}), \tag{2.33}$$

$$\Box_M \hat{\psi}^{\alpha\beta} = 2T^{\alpha\beta}. \tag{2.34}$$

These are the equations for 4D linearized gravity,[5] but with a source term. It should be noted that the procedure used above, wherein 5D terms are shifted in a vacuum relation to produce a 4D type of equation with a source, is now in common use (see Ref. 1 for a review). In our case, it would be necessary to have $\partial_4\hat{g}_{AB} = 0$, $\hat{g}_{4A} = 0$ and $\hat{g}_{44} = \text{const.}$ for a 5D vacuum to go to a 4D vacuum. These conditions will not in general be met by significant solutions of the 5D field equations (2.4). So, in general, Kaluza–Klein theory in 5D generates an energy–momentum tensor for Einstein theory in 4D.

The question of whether the induced energy–momentum tensor is conserved is now addressed. Since $T^{\alpha\beta}$ is of $O(\hat{\psi})$ we only need to verify that

$$\partial_\alpha T^{\alpha\beta} = 0, \tag{2.35}$$

since the product of $T^{\alpha\beta}$ and the Christoffel symbols is of $O(\hat{\psi}^2)$, which we neglect in our approximation. Taking the partial derivative of both sides of (2.34) and using (2.27) we obtain

$$\partial_\alpha T^{\alpha\beta} = \tfrac{1}{2}\Box_M \partial_\alpha \hat{\psi}^{\alpha\beta} = -\tfrac{1}{2}\Box_M \partial_4 \hat{\psi}^{4\beta}. \tag{2.36}$$

Therefore, for the induced stress–energy to be conserved in our approximation we must have $\Box_M \partial_4 \hat{\psi}^{4\beta}=0$. There are three cases in which this statement will hold, and these are worked out in detail in the Appendix. The first case is when $\Box_M \hat{\psi}^{4\alpha}=0$ which implies a massless spin-1 field, and leaves open the possibility for massive gravitons and scalar particles. The second case is when $\partial_4 \hat{\psi}^{4\beta}=0$ so that the spin-1 field has no x^4 dependence. The spin-1 field in this case is massless, but the graviton and the scalar field may be massive. The third and final case is when $\hat{\psi}^{4\beta}=0$ by a choice of the coordinate frame. This can be achieved by setting from the outset $\hat{\eta}_{4\beta}=\hat{h}_{4\beta}=0$, and this defines what we will refer to as the natural frame. This condition removes the spin-1 field (up to a coordinate transformation involving the extra coordinate), and allows for either massive or massless gravitons, but constrains the scalar field to be massless.

We now look at a simple example of plane waves in the natural frame. The flat-space metric and the perturbation are

$$\hat{\eta}_{AB} = \begin{pmatrix} \hat{\eta}_{\alpha\beta} & 0 \\ 0 & \epsilon \end{pmatrix}, \tag{2.37}$$

$$\hat{h}_{AB} = \begin{pmatrix} \hat{h}_{\alpha\beta} & 0 \\ 0 & \hat{h}_{44} \end{pmatrix}. \tag{2.38}$$

We can assume that $\hat{\psi}_{\alpha\beta}$ has the form of a 4D gravitational wave and a scalar wave:

$$\hat{\psi}^{AB} = \begin{pmatrix} \hat{\psi}^{\alpha\beta} & 0 \\ 0 & \hat{\psi}^{44} \end{pmatrix}, \tag{2.39}$$

$$\hat{\psi}^{AB} = \begin{pmatrix} A^{\alpha\beta} e^{i(k_\alpha x^\alpha + a x^4)} & 0 \\ 0 & A^{44} e^{i l_\alpha x^\alpha} \end{pmatrix}. \tag{2.40}$$

Here $A^{\alpha\beta}$ is the 4D polarization tensor, A^{44} is the amplitude of the scalar wave, and a is a constant with dimensions of inverse length which parametrizes the extra coordinate dependence. The field equations (2.24)–(2.26) simplify considerably and give

$$\Box_M \hat{\psi}^{\alpha\beta} = -\epsilon \partial_4^2 \hat{\psi}^{\alpha\beta} \Rightarrow k_\gamma k^\gamma = -\epsilon a^2, \tag{2.41}$$

$$\Box_M \hat{\psi}^{44} = -\epsilon \partial_4^2 \hat{\psi}^{44} \Rightarrow l_\gamma l^\gamma = 0. \tag{2.42}$$

We see that (2.42) can be interpreted as a massless scalar field, while (2.41) can be interpreted as a massive graviton when the parameter a is identified as

$$a = \pm i \sqrt{\epsilon} m. \tag{2.43}$$

The induced stress–energy is

$$T^{\alpha\beta} = -\frac{\epsilon}{2} a^2 A^{\alpha\beta} e^{i(k_\alpha x^\alpha + a x^4)}, \tag{2.44}$$

and its trace is given by

$$T^{\alpha}{}_{\alpha} = -\frac{\epsilon}{2} a^2 A^{\alpha}{}_{\alpha} e^{i(k_\alpha x^\alpha + a x^4)}. \tag{2.45}$$

This will be zero either if the 4D polarization tensor can be put into a trace-free form or if we choose $a=0$ (or both). This would imply a radiationlike equation of state induced on the 4D manifold representing gravitational radiation. The gauge condition (2.27) gives

$$k_\beta A^{\alpha\beta} = 0, \tag{2.46}$$

so the propagation of the 4D gravitational plane wave is transverse as it is in 4D theory.

So far there exists little difference in the qualitative features of 5D and 4D linearized theory, but the difference becomes apparent when the gauge freedom in $\hat{\xi}^A$ is used to obtain a transverse-traceless (TT) representation for $A^{\alpha\beta}$. The wave equations for the components of $\hat{\xi}^A$ which must be satisfied are

$$\Box_M \hat{\xi}^\alpha = -\epsilon \partial_4^2 \hat{\xi}^\alpha, \tag{2.47}$$

$$\Box_M \hat{\xi}^4 = -\epsilon \partial_4^2 \hat{\xi}^4. \tag{2.48}$$

The choice

$$\hat{\xi}^\alpha = (\hat{\xi}^\alpha, \hat{\xi}^4) = (-i e^\alpha e^{i(k_\gamma x^\gamma + a x^4)}, -i \epsilon e^4 e^{i(l_\gamma x^\gamma)}) \tag{2.49}$$

satisfies these wave equations, since (2.41) and (2.42) hold. In the natural frame, the transformations for the polarization components of $\hat{\psi}^{AB}$ under the gauge transformation are given by the following [see (2.18)]:

$$A'^{\alpha\beta} = A^{\alpha\beta} - k^\alpha e^\beta - k^\beta e^\alpha + \eta^{\alpha\beta} k_\gamma e^\gamma, \tag{2.50}$$

$$A'^{4\alpha} = -a \epsilon e^\alpha e^{i(k_\gamma x^\gamma + a x^4)} - \epsilon e^4 l_\alpha e^{i l_\gamma x^\gamma}, \tag{2.51}$$

$$A'^{44} = A^{44} + \epsilon k_\alpha e^\alpha e^{i[(k-l)_\gamma x^\gamma + a x^4]}. \tag{2.52}$$

We see that off-diagonal amplitudes have been generated and that the scalar amplitude has also been changed. What is interesting about the off-diagonal components is that they are a linear superposition of plane waves with different wave vectors, and functions of all five coordinates. The choice of setting $a = e^4 = 0$ is consistent with all the equations derived above but physically limiting since is sets the off-diagonal components in the transformed natural frame to zero, and therefore removes the electromagnetic effects which are usually associated with these components. This choice also sets the graviton mass to zero, and hence the 5D theory would give a conventional TT 4D gravitational wave and a scalar wave with an oscillating amplitude. Another simplification that takes place is if one chooses to remove the periodic behavior from the scalar field amplitude with the choice $k_\alpha = l_\alpha$. Since l_α is null, this forces k_α to be null as well, and hence a must equal zero, which implies a massless graviton again. In this case, the off-diagonal terms survive and give a simple expression:

$$A'^{4\alpha} = -\epsilon k^\alpha e^4 e^{i k_\gamma x^\gamma}. \tag{2.53}$$

This has the form of an electromagnetic plane wave propagating with the same null wave vector as the gravitational plane wave.

Let us now consider an example of a plane gravitational wave in the natural frame propagating in the z direction. The wave vector is

$$k_\gamma = (\omega, 0, 0, k), \tag{2.54}$$

and obeys the conditions (2.41), (2.43) and (2.46). These give

$$\omega^2-k^2=m^2,\qquad \omega A^{0\alpha}=kA^{3\alpha}. \tag{2.55}$$

Performing the gauge transformations and using the above we see that the only independent components of the $\hat{\psi}^{AB}$ are

$$\begin{aligned}
A'^{00}&=A^{00}-\omega e^0-ke^3,\\
A'^{01}&=A^{01}-\omega e^1,\\
A'^{02}&=A^{02}-\omega e^2,\\
A'^{11}&=A^{11}-\omega e^0+ke^3,\\
A'^{12}&=A^{12},\\
A'^{22}&=A^{22}-\omega e^0+ke^3,\\
A'^{44}&=A^{44}-\epsilon(\omega e^0-ke^3).
\end{aligned} \tag{2.56}$$

Then the choice

$$\begin{aligned}
e^0&=\frac{1}{2\omega}\left(A^{00}+\frac{1}{2}A^{11}+\frac{1}{2}A^{22}\right),\\
e^1&=A^{01},\\
e^2&=A^{02},\\
e^3&=\frac{1}{2k}\left(A^{00}-\frac{1}{2}A^{11}-\frac{1}{2}A^{22}\right)
\end{aligned} \tag{2.57}$$

will bring the 4D polarization tensor to a transverse-traceless form, but will generate off-diagonal terms and an oscillating scalar field amplitude. The choice $e^4=a=0$ corresponds to setting the off-diagonal terms to zero, and the $\hat{\psi}^{AB}$ represents a massless TT gravitational wave and a scalar wave. If we choose $l_\alpha=k_\alpha$ (which again forces the graviton mass to zero), the 4D gravitional wave is again TT, but the off-diagonal components with the choice of $e^4=\epsilon C/k$ (where $C\ll 1$ is a constant), are:

$$A'^{04}=-Ce^{ik_\gamma x^\gamma},\qquad A'^{34}=Ce^{ik_\gamma x^\gamma}. \tag{2.58}$$

These are simple plane waves with constant amplitude.

What we have shown in this section is that 5D gravity can be linearized in the same way as 4D gravity, but that the harmonic gauge (2.8) is very restrictive; and since it is noncovariant, conservation of the induced energy–momentum tensor on the 4D subspace holds only when certain restrictions are imposed on the extra off-diagonal components of the perturbation tensor. Also, while the equations allow in principle for massive gravitons, the most natural choice of parameters which removes electromagnetic effects also reduces the graviton mass to zero.

III. WAVES IN CONFORMAL 5D GRAVITY

In this section we investigate the consequences of imposing the 5D harmonic gauge condition for a metric with two conformal factors, one for the 4D part of the metric and another for the 5D off-diagonal components. We then look at the propagation of electromagnetic waves and their interaction with the scalar field. To end this section we give a solution to the field equations which has physical relevance.

Let us consider the metric

$$dS^2=\phi^{2a}g_{\alpha\beta}\,dx^\alpha\,dx^\beta-\phi^{2b}(dx^4+A_\alpha\,dx^\alpha)^2, \tag{3.1}$$

where a and b are constants. The demand that this metric satisfy the 5D harmonic gauge condition (2.8) gives two equations:

$$\hat{\Gamma}^\gamma\equiv\hat{g}^{AB}\hat{\Gamma}^\gamma_{AB}=0\Rightarrow g^{\alpha\beta}\Gamma^\gamma{}_{\alpha\beta}+(2a+b)\,\frac{\phi^\gamma}{\phi}=0, \tag{3.2}$$

$$\hat{\Gamma}^4\equiv\hat{g}^{AB}\hat{\Gamma}^4_{AB}=0\Rightarrow\nabla_\alpha A^\alpha+(2a+b)\,\frac{\phi_\alpha A^\alpha}{\phi}=0. \tag{3.3}$$

Here ∇_α is the 4D covariant derivative operator associated with the metric $g_{\alpha\beta}$. In order for the 5D harmonic gauge to induce the 4D harmonic gauge as well as the Lorentz gauge, we are forced to choose $2a+b=0$. With this constraint and changing the conformal factor to $\phi=e^{2\sigma/b\sqrt{3}}$, the metric takes the form

$$dS^2=e^{-2\sigma/\sqrt{3}}g_{\alpha\beta}\,dx^\alpha dx^\beta-e^{4\sigma/\sqrt{3}}(dx^4+A_\alpha\,dx^\alpha)^2, \tag{3.4}$$

which is the usual metric for Kaluza–Klein gravity in the Einstein frame.[9] The field equations are obtained by calculating the $(\alpha\beta)$, (4α), and (44) components of $\hat{R}_{AB}=0$. We obtain

$$R_{\alpha\beta}=-\tfrac{1}{2}e^{2\sqrt{3}\sigma}(F_{\alpha\gamma}F_\beta{}^\gamma-\tfrac{1}{4}g_{\alpha\beta}F^2)+2\sigma_\alpha\sigma_\beta, \tag{3.5}$$

$$\nabla_\beta(e^{2\sqrt{3}\sigma}F^{\alpha\beta})=0, \tag{3.6}$$

$$\Box\sigma+\frac{\sqrt{3}}{8}e^{2\sqrt{3}\sigma}F^2=0. \tag{3.7}$$

The Einstein tensor can be used to give the induced matter of an electromagnetic field coupled to a massless scalar field. Thus

$$G_{\alpha\beta}\equiv T^{\mathrm{EM}}_{\alpha\beta}+T^S_{\alpha\beta}, \tag{3.8}$$

where

$$T^{\mathrm{EM}}_{\alpha\beta}=-\tfrac{1}{2}e^{2\sqrt{3}\sigma}(F_{\alpha\gamma}F_\beta{}^\gamma-\tfrac{1}{4}g_{\alpha\beta}F^2), \tag{3.9}$$

$$T^S_{\alpha\beta}=2(\sigma_\alpha\sigma_\beta-\tfrac{1}{2}g_{\alpha\beta}\sigma^\gamma\sigma_\gamma). \tag{3.10}$$

The Ricci scalar is simply

$$R=2\sigma^\gamma\sigma_\gamma. \tag{3.11}$$

We now investigate the propagation of electromagnetic waves using the Lorentz gauge (3.3) with $2a+b=0$ and Maxwell's equations (3.6). We postpone the use of the 4D harmonic gauge so the following equations hold 4D covariantly for the induced matter from the Kaluza–Klein metric in the Einstein frame. The 4D harmonic gauge will however be imposed when we discuss an exact solution at the end of this section. The ansatz for the electromagnetic vector potential is

$$A_\alpha = a_\alpha e^{i\omega S}, \tag{3.12}$$

where $a_\alpha \in \mathrm{Im}$, $S \in \mathrm{Re}$, and this assumption represents a good approximation for large ω only. Substituting this ansatz into both the Lorentz gauge condition and Maxwell's equations and setting the coefficients of ω^2 and ω separately to zero, gives, respectively,

$$\nabla_\alpha A^\alpha = 0 \Rightarrow k_\alpha a^\alpha = 0, \tag{3.13}$$

$$O(\omega^2):\ \ k_\alpha k^\alpha = 0, \tag{3.14}$$

$$O(\omega):\ \ k^\alpha \nabla_\beta a^\beta - a^\alpha \nabla_\beta k^\beta - 2k^\beta \nabla_\beta a^\alpha = -2\sqrt{3}\sigma_\beta (k^\alpha a^\beta - k^\beta a^\alpha). \tag{3.15}$$

Here $k_\alpha \equiv \partial_\alpha S$ is the gradient to surfaces of constant phase, and as such the null condition implies that the k_α propagate on null geodesics:

$$k^\beta \nabla_\beta k^\alpha = 0. \tag{3.16}$$

The contraction of (3.15) with $\overline{a_\alpha}$ and the orthogonality condition (3.13) allows us to obtain

$$\nabla_\alpha (a^2 k^\alpha) = -\sqrt{3}\sigma_\alpha (a^2 k^\alpha), \tag{3.17}$$

where $a^2 \equiv a_\alpha \overline{a^\alpha}$. If we define a photon current $j^\alpha \equiv a^2 k^\alpha$, the right-hand side here can be interpreted as a violation of photon number due to the coupling of the gradient of the scalar field to the wave vector. The equation governing the propagation of the unit polarization vector can be obtained from (3.17) by the substitution of $f^\alpha = a^\alpha / a^2$, and is

$$k^\beta \nabla_\beta f^\alpha = \sqrt{3}\sigma_\beta f^\alpha + \frac{1}{2}\left(\nabla_\beta f^\beta + \frac{f^\beta \nabla_\beta a}{a} \right) k^\alpha. \tag{3.18}$$

This again differs from the usual 4D result due to the scalar-field coupling. Since the photons follow 4D null geodesics, their motion in the 4D part of the 5D space is the same as their motion in 4D space–time. However, the scalar field does modify the conservation equation for photon number, and also introduces a new term to the propagation of the unit polarization vector. Both things could in principle be tested.

So far we have not used the 4D harmonic gauge in deriving the above equations (3.5)–(3.18). We will illustrate the use of this gauge by deriving an exact, gravitational plane wave solution accompanied by a plane electromagnetic wave and a scalar wave. The choice of an electromagnetic plane wave simplifies the induced 4D field equations since $F^2 = 0$ for plane waves. This reduces the field equations (3.5)–(3.7) along with the 4D harmonic and Lorentz gauge to

$$R^H_{\alpha\beta} = -\tfrac{1}{2} e^{2\sqrt{3}\sigma} F_{\alpha\gamma} F_\beta{}^\gamma + 2\sigma_\alpha \sigma_\beta, \tag{3.19}$$

$$\partial_\alpha F^{\alpha\beta} + \partial_\alpha (\ln \sqrt{g} F^{\alpha\beta}) = -2\sqrt{3}\sigma_\alpha F^{\alpha\beta}, \tag{3.20}$$

$$\Box \sigma = g^{\alpha\beta} \partial_\alpha \partial_\beta \sigma = 0, \tag{3.21}$$

$$g^{\alpha\beta} \Gamma^\gamma_{\alpha\beta} = 0, \tag{3.22}$$

$$\nabla_\alpha A^\alpha = g^{\alpha\beta} \partial_\alpha A_\beta = 0. \tag{3.23}$$

Here $R^H_{\alpha\beta}$ is the Ricci tensor in the harmonic gauge. We would like a 4D gravitational wave metric that would simplify these equations even further. The candidate metric should have a determinant equal to a constant, which would remove the second term on the left-hand side of (3.20), while also satisfying the 4D harmonic gauge (3.22). A simple metric which satisfies these conditions is that of an exact gravitational plane wave with parallel rays (*pp*-wave) traveling along the z direction. Such a metric has the form:[10]

$$ds^2 = K(u,\xi,\overline{\xi})du^2 + 2\,du\,dv - d\xi\,d\overline{\xi}. \tag{3.24}$$

Here u and v are retarded and advanced coordinates, and $\xi = x + iy$ and $\overline{\xi} = x - iy$ are complex transverse coordinates. If we choose the electromagnetic vector potential and the scalar field to be independent of the complex transverse coordinates, we have

$$A_\alpha = (0,0,\tfrac{1}{2}A(u),\tfrac{1}{2}A(u)), \quad \sigma = \sigma(u). \tag{3.25}$$

The vector potential then corresponds to an electric field oscillating in the x direction and a magnetic field oscillating in the y direction, while the wave propagates in the z direction. One can check that the scalar wave equation (3.21) and Maxwell's equations (3.20) are satisfied by arbitrary functions $\sigma(u)$ and $A(u)$, and that the only surviving component of the Ricci tensor is R_{uu}, which gives for (3.19) the equation

$$\partial_\xi \partial_{\overline{\xi}} K(u,\xi,\overline{\xi}) = \tfrac{1}{4} e^{2\sqrt{3}\sigma}[\partial_u A(u)]^2 + [\partial_u \sigma(u)]^2. \tag{3.26}$$

This equation can be integrated immediately to give

$$K(u,\xi,\overline{\xi}) = (\tfrac{1}{4} e^{2\sqrt{3}\sigma}[\partial_u A(u)]^2 + [\partial_u \sigma(u)]^2)\xi\overline{\xi} + f(u)\xi^2 + \overline{f}(u)\overline{\xi}^2, \tag{3.27}$$

where $f(u)$ is an arbitrary complex function which represents the solution to the homogeneous equation. Since we are interested in electromagnetic and scalar waves we assume that they can be written as

$$A(u) = \mathrm{Re}\, A_0 e^{i\omega u}, \quad \sigma(u) = \mathrm{Re}\, \sigma_0 e^{i\lambda u}, \tag{3.28}$$

where A_0 and σ_0 are real constants. This gives on taking the real parts only,

$$K(u,\xi,\overline{\xi}) = (\tfrac{1}{4} e^{2\sqrt{3}\sigma_0 \cos(\lambda u)}[A_0^2 \omega^2 \sin^2(\omega u)] + \sigma_0^2 \lambda^2 \sin^2(\lambda u))\xi\overline{\xi}. \tag{3.29}$$

This particular solution is simple and realistic, and would probably repay future investigation. However, all of the work we have done in this section presumes the 5D harmonic gauge, in which the graviton and photon are massless. This is acceptable, but leaves open the question of how the theory should be interpreted in order to accommodate the many kinds of particles in nature which have finite masses. We turn our attention to this now.

IV. PARTICLE MASSES IN CONFORMAL 5D FIELD THEORY

It is now well known that because Einstein's equations with matter are a subset of the Kaluza–Klein equations in vacuum, phenomenological fluids in 4D can be generated by appropriate coordinate transformations in 5D.[1] The idea behind this, namely that 4D matter is a manifestation of 5D geometry, can in principle be extended from continuous fluids to particles.[6–8] For example, a connection between the mass of a particle as measured by its Compton wavelength and the dimension of the geometry as measured by its Ricci scalar was put forward as a realization of Mach's principle.[6] An extension of this approach was to envisage 4D space–time as having principle curvatures related via de Broglie waves to the 4-momenta of a particle, the mass again being determined by the scalar curvature.[7] The latter proposal involved a reinterpretation of the right-hand side of Einstein's field equations, and an alternative way of doing this has recently been proposed that introduces the Planck length as a fundamental parameter and derives particle energies dependent on wave modes in the extra dimension.[8] In this section, we wish to inquire into the possible relationship between the mass of a particle and a conformal factor in the geometry.

Following Sec. III and Ref. 7, let us consider metric (3.1) with $a=1$ and $A_\alpha=0$ (this does not restrict the 4D metric but removes electromagnetic effects). Thus the metric we consider has the form

$$\begin{aligned} dS^2 &= \phi^2\, ds^2 - \phi^{2b}(dx^4)^2, \\ ds^2 &= g_{\alpha\beta}\, dx^\alpha\, dx^\beta. \end{aligned} \tag{4.1}$$

Using as elsewhere a caret to distinguish 5D quantities from purely 4D quantities as conventionally defined, the Christoffel symbols of the second kind for (4.1) are

$$\begin{aligned} \hat\Gamma^\alpha_{\beta\gamma} &= \Gamma^\alpha_{\beta\gamma} + \phi^{-1}(\delta^\alpha_\beta\phi_\gamma + \delta^\alpha_\gamma\phi_\beta - g_{\alpha\beta}\phi^\alpha), \\ \hat\Gamma^\alpha_{\beta\alpha} &= \Gamma^\alpha_{\beta\alpha} + 4\,\frac{\phi_\alpha}{\phi}, \\ \hat\Gamma^\alpha_{44} &= b\,\phi^{2b-3}\phi^\alpha, \\ \hat\Gamma^\alpha_{4\beta} &= \hat\Gamma^4_{\alpha\beta} = \hat\Gamma^4_{44} = 0, \\ \hat\Gamma^4_{4\alpha} &= b\,\frac{\phi_\alpha}{\phi}, \\ \hat\Gamma^D_{\alpha D} &= \Gamma^\delta_{\alpha\delta} + (b+4)\,\frac{\phi_\alpha}{\phi}, \\ \hat\Gamma^D_{4D} &= 0. \end{aligned} \tag{4.2}$$

Here as before, $\phi_\alpha \equiv \partial_\alpha\phi$ and $\phi^\alpha \equiv g^{\alpha\beta}\phi_\beta$. The components of the 5D Ricci tensor are

$$\hat R_{\alpha\beta} = R_{\alpha\beta} - (b+2)\phi^{-1}\nabla_\alpha\phi_\beta - (b+1)(b-4))\phi^{-2}\phi_\alpha\phi_\beta - \phi^{-2}[\phi\nabla_\gamma\phi^\gamma + (b+1)\phi_\gamma\phi^\gamma]g_{\alpha\beta}, \tag{4.3}$$

$$\hat R_{4\alpha} = 0, \tag{4.4}$$

$$\hat R_{44} = b\,\phi^{2b-4}[\phi\nabla_\alpha\phi^\alpha + (b+1)\phi_\alpha\phi^\alpha]. \tag{4.5}$$

Clearly, the electromagnetic components of the field equations $\hat R_{AB}=0$ are trivially satisfied, while the remaining equations read

$$R_{\alpha\beta} = (b+2)\phi^{-1}\nabla_\alpha\phi_\beta + (b+1)(b-4)\phi^{-2}\phi_\alpha\phi_\beta, \tag{4.6}$$

$$\phi\nabla_\alpha\phi^\alpha + (b+1)\phi_\alpha\phi^\alpha = 0. \tag{4.7}$$

These are a set of ten tensor equations and one scalar equation, and are what we will be concerned with in what follows.

From (4.6) using (4.7) we find that the 4D Ricci scalar is

$$R = -6(b+1)\,\frac{\phi_\alpha\phi^\alpha}{\phi^2}. \tag{4.8}$$

This combined with (4.6) allows us to form the Einstein tensor $G_{\alpha\beta}=R_{\alpha\beta}-\frac{1}{2}Rg_{\alpha\beta}$. Then Einstein's equations $G_{\alpha\beta}\equiv T_{\alpha\beta}$ define the induced energy–momentum tensor:

$$T_{\alpha\beta}=(b+2)\phi^{-1}\nabla_{\alpha}\phi_{\beta}+(b+1)(b-4)\phi^{-2}\phi_{\alpha}\phi_{\beta}+3(b+1)\phi^{-2}\phi_{\gamma}\phi^{\gamma}g_{\alpha\beta}. \tag{4.9}$$

This obviously contains a term that is second order in the derivatives of the scalar field, and two terms that are first order as in ordinary quantum field theory. We can in principle have matter that depends solely on second derivatives by setting $b=-1$, or matter that depends solely on first derivatives by setting $b=-2$. However, we have not so far imposed the 5D harmonic gauge (3.2), and if we do so then it corresponds to setting $b=-2$, and we specialize to this case in what follows. Thus the 5D line element is given by

$$dS^2=\phi^2\,ds^2-\phi^{-4}(dx^4)^2, \tag{4.10}$$

where ds^2 is the 4D line element and the metric components are independent of the extra coordinate $(\partial_4 g_{\alpha\beta}=\partial_4\phi=0)$. The 5D field equations $\hat{R}^{AB}=0$ then yield the 4D ones

$$R_{\alpha\beta}=6\,\frac{\phi_{\alpha}\phi_{\beta}}{\phi^2}, \tag{4.11}$$

$$\nabla_{\alpha}\phi^{\alpha}=\frac{\phi_{\alpha}\phi^{\alpha}}{\phi}, \tag{4.12}$$

with an effective 4D energy–momentum tensor

$$T_{\alpha\beta}=6\phi^{-2}(\phi_{\alpha}\phi_{\beta}-\tfrac{1}{2}\phi^{\gamma}\phi_{\gamma}g_{\alpha\beta}). \tag{4.13}$$

It is our objective to evaluate this for a particle.

To this end, let us briefly review 4D particle dynamics. The Lagrangian is

$$L\equiv m\,\frac{ds}{d\lambda}=m\sqrt{g_{\alpha\beta}\,\frac{dx^{\alpha}}{d\lambda}\,\frac{dx^{\beta}}{d\lambda}}, \tag{4.14}$$

where λ is a parameter and m is the particle mass, which we assume to be constant. The covariant and contravariant components of the 4-momenta are

$$p_{\alpha}=\frac{\partial L}{\partial\left(\dfrac{dx^{\alpha}}{d\lambda}\right)}=mg_{\alpha\beta}\,\frac{dx^{\beta}}{ds}, \tag{4.15}$$

$$p^{\alpha}=g^{\alpha\beta}p_{\beta}=m\,\frac{dx^{\alpha}}{ds}, \tag{4.16}$$

and obey $p_{\alpha}p^{\alpha}=m^2$. The action corresponding to the Lagrangian is

$$I=\int L\,d\lambda=\int\left(m\,\frac{ds}{d\lambda}\right)d\lambda=\int m\,ds=\int m\left(g_{\alpha\beta}\,\frac{dx^{\alpha}}{ds}\,\frac{dx^{\beta}}{ds}\right)ds=\int p_{\alpha}\,dx^{\alpha}, \tag{4.17}$$

where we have used (4.14) and (4.15), respectively. However, in place of the action we can use a wave function

$$\Psi\equiv e^{-iI/\hbar} \tag{4.18}$$

involving Planck's constant. In terms of this, the momenta are given by

$$-i\hbar\nabla_\alpha\Psi=\frac{\partial}{\partial x^\alpha}\Psi\equiv p_\alpha\Psi. \tag{4.19}$$

If the p_α were constants as in flat space–time, they would be the eigenvalues and Ψ the eigenfunction of the 4D momentum operator. Now the p_α are not generally constants in curved space–time, but we can keep (4.19) as the defining relation for the momenta because the preceeding relations (4.14)–(4.18) are covariant. Then following the philosophy of induced-matter theory, the question is how to relate the particle formalism involving Ψ to the Kaluza–Klein equations (4.7) involving ϕ.

We answer this by adopting the ansatz

$$\phi\equiv\Psi^d, \tag{4.20}$$

where d is a dimensionless parameter. Then (4.11), (4.12) with (4.19), (4.20) give

$$R_{\alpha\beta}=-\frac{6d^2}{\hbar^2}p_\alpha p_\beta, \tag{4.21}$$

$$\Box\Psi+\frac{m^2}{\hbar^2}\Psi=0. \tag{4.22}$$

The corresponding induced energy–momentum tensor (4.13) is

$$T_{\alpha\beta}=-6\frac{d^2}{\hbar^2}\left(p_\alpha p_\beta-\frac{m^2}{2}g_{\alpha\beta}\right). \tag{4.23}$$

These relations make sense in the context of induced-matter theory and justify the assumption (4.20). Thus (4.22) is the curved-space–time Klein–Gordon equation. The implication is that for the harmonic gauge the extra potential of the 5D space is related to the wave function of the 4D space. That wave function, by (4.1), represents a conformal modification of space–time. It can be regarded, alternatively, as a particle with mass m and momenta p_α. Then (4.21) says that the de Broglie wavelengths associated with p_α describe by their product $p_\alpha p_\beta$ the curvature of the 4D space as represented by the Ricci tensor $R_{\alpha\beta}$. The corresponding energy–momentum tensor (4.13) is really a dynamical quantity, which by virtue of $\nabla_\alpha T^{\alpha\beta}=0$ can be shown to yield the standard equations of motion $p^\beta\nabla_\beta p^\alpha=0$. [See Ref. 6: It can be shown by some tedious algebra that the 4D Bianchi identities $\nabla_\alpha G^{\alpha\beta}=\nabla_\alpha T^{\alpha\beta}=0$ with (4.13) give relations which are the same as those derived from the 4D part of the 5D geodesic equation.] In other words, (4.13)–(4.20) are the field equations for what might be called Kaluza–Klein–Gordon field theory.

To go further, we need exact solutions. The most convenient way to show such is to write the 4D part of the metric as

$$ds^2=dt^2-e^{2\lambda}dx^2-e^{2\mu}dy^2-dz^2, \tag{4.24}$$

where $\lambda=\lambda(t,z)$ and $\mu=\mu(t,z)$. This metric can describe a plane wave moving in the z-direction.[10] We expect the associated particle to have momenta which are constants and given by (4.15) as

$$p_0=m\frac{dt}{ds}\equiv E,\quad p_1=0,\quad p_2=0,\quad p_3=-m\frac{dz}{ds}\equiv -p. \tag{4.25}$$

These particle properties are connected to the wave properties by the field equations (4.21) and (4.22). The components of (4.21) can be evaluated using (4.25) and read

$$
\begin{aligned}
R_{00} &= -(\lambda+\mu)_{00} - (\lambda_0^2+\mu_0^2) = -6d^2E^2/\hbar^2, \\
R_{03} &= -(\lambda+\mu)_{03} - (\lambda_0\lambda_3+\mu_0\mu_3) = 6d^2Ep/\hbar^2, \\
R_{33} &= -(\lambda+\mu)_{33} - (\lambda_3^2+\mu_3^2) = -6a^2p^2/\hbar^2, \\
R_{11} &= e^{2\lambda}[\lambda_{00}-\lambda_{33}+\lambda_0(\lambda+\mu)_0-\lambda_3(\lambda+\mu)_3] = 0, \\
R_{22} &= e^{2\mu}[\mu_{00}-\mu_{33}+\mu_0(\lambda+\mu)_0-\mu_3(\lambda+\mu)_3] = 0, \\
R_{\alpha\beta} &= 0 \quad \text{otherwise.}
\end{aligned}
\tag{4.26}
$$

It is obvious when written in this way that there is a solution given by

$$
\lambda = -\mu = \sqrt{3}a(Et-pz)/\hbar. \tag{4.27}
$$

This with (4.17) and (4.18) gives the action and the wave function, respectively,

$$
I = \int p_\alpha \, dx^\alpha = Et - pz, \tag{4.28}
$$

$$
\Psi = e^{-i(Et-pz)/\hbar}. \tag{4.29}
$$

The latter should satisfy the fifth or Klein–Gordon part of the field equations (4.22) and we find that it does so. The induced or effective 4D energy–momentum tensor given by (4.23) and (4.25) has the following nonzero components:

$$
\begin{aligned}
T^0{}_0 &= -6d^2(E^2-m^2/2)/\hbar^2, \\
T^0{}_3 &= -T^3{}_0 = 6d^2Ep/\hbar^2, \\
T^1{}_1 &= T^2{}_2 = 3d^2m^2/\hbar^2, \\
T^3{}_3 &= 6d^2(p^2+m^2/2)/\hbar^2.
\end{aligned}
\tag{4.30}
$$

These obey $T^\alpha{}_\alpha = 6d^2m^2/\hbar^2$, and we recall from (4.15) and (4.16) that E, p, and m are related by $p_\alpha p^\alpha = E^2 - p^2 = m^2$. This constraint means that we have a class of exact solutions that depends on two constants, E and p. [The dimensionless constant d which we introduced in (4.20) to connect the scalar field and the wave function could in principle be absorbed into E and p, or into $\hbar$.] We can sum up the 5D wave solutions by

$$
dS^2 = \Psi^{2d}\, ds^2 - \Psi^{-4d}\, dx^{4^2}, \tag{4.31}
$$

$$
ds^2 = dt^2 - e^{2\lambda}\, dx^2 - e^{-2\lambda}\, dy^2 - dz^2, \tag{4.32}
$$

$$
\Psi = e^{-i(Et-pz)/\hbar}, \quad \lambda = \sqrt{3}d(Et-pz)/\hbar. \tag{4.33}
$$

We have examined the algebraic properties of this class of solutions using GRTENSOR,[11] and apart from confirming $\hat{R}_{AB}=0$, we find $\hat{R}_{ABCD}\neq 0$ so that these solutions are not 4D curved and 5D flat like some others in the literature. [See Ref. 1; for (4.31)–(4.33) above we find for example that $\hat{R}_{1414} = -2d^2(3E^2+p^2)\Psi^{-4d}\neq 0$ and $K \equiv \hat{R}_{ABCD}\hat{R}^{ABCD}\neq 0$.] However, while the solutions (4.31)–(4.33) are in general curved, they have $\sqrt{-g}=1$ for the determinant of the 4D metric and so are algebraically special.

It is useful to review the argument of this section before discussing some implications. The nonelectromagnetic metric (4.1) with $\hat{R}_{AB}=0$ yields field equations (4.6), (4.7) with tensor and scalar parts. Together they give a Ricci scalar (4.8), and in the induced-matter picture[1] an energy–momentum tensor (4.9), both of which depend on a free parameter b. This is fixed as $b=-2$ by the harmonic condition (3.2), or alternatively by the requirement that matter depend on first-order derivatives of the scalar field. The resulting metric (4.10) has field equations (4.11), (4.12) and energy–momentum tensor (4.13). Conventional particle dynamics (4.14)–(4.19) can be connected to classical field theory via the ansatz (4.20), which effectively says that the wave function depends on the scalar field or fifth part of the metric, or equivalently on the conformal factor that modifies the 4D metric. The field equations are then (4.21) and (4.22), of which the latter is the Klein–Gordon equation with an effective particle mass m. The induced energy–momentum tensor (4.23) is actually a dynamical quantity, and the Bianchi identities of field theory yield the standard equations of motion of particle physics. The special solution (4.33) satisfies the field equations of classical Kaluza–Klein theory, but is written in terms of a wave function associated with quantum theory.

The implications of this are clear and can be summed up succinctly by the relation

$$m^2=\frac{k\hbar^2}{c^2}R. \tag{4.34}$$

Here we have conventional units for Planck's constant and the speed of light, and k is a coupling constant. But the essential thing is that a particle's rest mass can be related to the scalar curvature of the 4D subspace of a 5D manifold. This idea is not new (see Refs. 6–8), but the results of this section are significant in that they demonstrate its viability for a fairly wide class of x^4-independent metrics. However, two comments need to be made. First, it is not always possible to write the 4D Ricci tensor as a product of two (momentum) vectors as in (4.21). More work is needed on this, presumably using the Segré classification of the Ricci tensor. Second, we recall from differential geometry that the Ricci tensor $\tilde{R}$ in the frame where there is a 4D conformal factor ϕ^{2a} is related to the Ricci tensor R in the frame where there is no conformal factor by

$$\tilde{R}=\phi^{-2a}\left(R-6a\,\frac{\Box\phi}{\phi}+6a(1-a)\,\frac{\phi_\gamma\phi^\gamma}{\phi^2}\right). \tag{4.35}$$

This means that if the mass is known in one frame it is different in another frame by an amount that depends on the conformal factor. More work is needed on this in order to see how conformal factors relate to the observed hierarchy of particle masses.

V. CONCLUSION

The equations of 4D general relativity are succinct and in the form $R_{\alpha\beta}=0$ are known from the classical tests to be in agreement with observation. However, the dimensionality of the equations, as recognized by Einstein and Kaluza, is open. This has led to much work on Kaluza–Klein theory, supergravity, and superstrings. In the present article, we have gone back to basic issues, and looked at three overlapping problems in 5D theory with the field equations $\hat{R}_{AB}=0$.

Waves in 5D can be approached as in 4D, using a linearized metric and the harmonic gauge condition. However, the last is algebraically restrictive, and physically implies that the particle analogs of the gravitational and scalar waves are massless. Introducing conformal factors into the metric (a traditional step in Kaluza–Klein theory) generalizes the field equations, and imposing the harmonic gauge describes massless gravitons and photons. Now there is nothing heinous about this, but if 5D theory is to describe the real world then one needs to find a way to codify particles of finite mass. This can apparently be done if, again using a conformal factor, the field equations are rewritten so that the mass of a particle is defined by the scalar curvature of a 4D space embedded in a 5D space.

Our results follow because, as in the induced-matter approach to fluids,[1] we allow dependency on the extra coordinate. This algebraic latitude has not hitherto been much applied to elucidating the physical properties of waves and particles; and as noted at several places above, in this regard there are vistas to be explored.

ACKNOWLEDGMENTS

We thank NSERC and OGS for financial support.

APPENDIX: CONSERVATION OF THE LINEARIZED INDUCED STRESS–ENERGY TENSOR

As discussed in Sec. II, conservation of the induced stress–energy tensor in our linearized approximation will hold only if

$$\Box_M \partial_4 \hat{\psi}^{4\alpha} = 0. \tag{A1}$$

There are three cases in which this is satisfied and they are presented here.

Case 1: $\Box_M \hat{\psi}^{4\alpha} = 0$. The spin-1 field in this case is massless, and implies from the right-hand side of (2.25) that

$$-2\,\hat{\eta}^{4\gamma}\partial_\gamma \partial_4 \hat{\psi}^{4\alpha} = \hat{\eta}^{44}\partial_4^2 \hat{\psi}^{4\alpha}. \tag{A2}$$

This constraint does not have a direct physical interpretation but follows from the requirement that the stress–energy be conserved. The wave equations (2.24) and (2.26) allow for massive gravitons $(\partial_4 \hat{\psi}^{\alpha\beta} \neq 0)$ or massless gravitons $(\partial_4 \hat{\psi}^{\alpha\beta} = 0)$, and a massive $(\partial_4 \hat{\psi}^{44} \neq 0)$ or massless $(\partial_4 \hat{\psi}^{44} = 0)$ scalar field.

Case 2: $\partial_4 \hat{\psi}^{4\alpha} = 0$. This implies that the spin-1 field has no x^4 dependence and so $\hat{\psi}^{4\alpha} = \hat{\psi}^{4\alpha}(x^\Sigma)$. The gauge constraint equations (2.27) and (2.28) then reduce to

$$\partial_\alpha \hat{\psi}^{\alpha\beta} = 0, \tag{A3}$$

$$\partial_\alpha \hat{\psi}^{4\alpha} = -\partial_4 \hat{\psi}^{44}, \tag{A4}$$

and by operating on (A4) with ∂_4 and using the commutation of partial derivatives along with $\partial_4 \hat{\psi}^{4\alpha} = 0$, we see that

$$\partial_4^2 \hat{\psi}^{44} = 0. \tag{A5}$$

The wave equations (2.24)–(2.26) then reduce to

$$\Box_M \hat{\psi}^{\alpha\beta} = -2\,\hat{\eta}^{4\gamma}\partial_4 \partial_\gamma \hat{\psi}^{\alpha\beta} - \hat{\eta}^{44}\partial_4^2 \hat{\psi}^{\alpha\beta}, \tag{A6}$$

$$\Box_M \hat{\psi}^{4\alpha} = 0, \tag{A7}$$

$$\Box_M \hat{\psi}^{44} = -2\,\hat{\eta}^{4\gamma}\partial_4 \partial_\gamma \hat{\psi}^{44}. \tag{A8}$$

These are the wave equations for a massive graviton $(\partial_4 \hat{\psi}^{\alpha\beta} \neq 0)$ or massless graviton $(\partial_4 \hat{\psi}^{\alpha\beta} = 0)$, a massless spin-1 field and massive $(\partial_4 \hat{\psi}^{44} \neq 0)$ or massless $(\partial_4 \hat{\psi}^{44} = 0)$ scalar field.

Case 3: $\hat{\eta}_{4\alpha} = \hat{h}_{4\alpha} = 0 \Rightarrow \hat{g}_{4\alpha} = 0$. In this case the off-diagonal components of the flat-space metric and the perturbation tensor are set to zero from the outset, defining what we refer to as the natural frame. This automatically sets $\hat{\psi}_{4\alpha} = 0$ by the definition (2.13). Therefore the spin-1 field is removed from the theory in this frame, but can be reintroduced using a coordinate transformation involving the extra coordinate. The gauge constraints (2.27) and (2.28) reduce to

$$\partial_\alpha \hat{\psi}^{\alpha\beta} = 0, \tag{A9}$$

$$\partial_4 \hat{\psi}^{44} = 0. \tag{A10}$$

These simplify the wave equations (2.24) and (2.26) in the natural frame so [using (A10) in (2.26)] they become

$$\Box_M \hat{\psi}^{\alpha\beta} = -\hat{\eta}^{44} \partial_4^2 \hat{\psi}^{\alpha\beta}, \tag{A11}$$

$$\Box_M \hat{\psi}^{44} = 0. \tag{A12}$$

The scalar field in this case is massless, and the gravitons are massive if $\partial_4 \hat{\psi}^{\alpha\beta} \neq 0$ and massless if $\partial_4 \hat{\psi}^{\alpha\beta} = 0$. Since off-diagonal components are absent in the natural frame, we can use the coordinate transformation (2.16) to generate off-diagonal terms. In the natural frame, the coordinate transformation for the extra coordinate is

$$x^4 \rightarrow x'^4 = x^4 + \hat{\xi}^4 (x^\Sigma), \tag{A13}$$

where $\hat{\xi}^4$ obeys the wave equation

$$\Box_M \hat{\xi}^4 = 0. \tag{A14}$$

The off-diagonal components $\hat{\psi}_{4\alpha}$ transform [using $\partial_4 \hat{\xi}_\alpha = 0$ and (2.18)] as

$$\hat{\psi}_{4\alpha} \rightarrow \hat{\psi}'_{4\alpha} = \hat{\psi}_{4\alpha} - \partial_\alpha \hat{\xi}_4 . \tag{A15}$$

These equations can be compared to the coordinate transformation which shifts the electromagnetic vector potential in regular 5D Klauza–Klein theory

$$x^4 \rightarrow x'^4 = x^4 + \lambda (x^\Sigma), \tag{A16}$$

$$A_\alpha \rightarrow A'_\alpha = A_\alpha + \partial_\alpha \lambda (x^\Sigma).$$

This is the usual gauge freedom in electromagnetism, where the function λ satisfies

$$\Box_M \lambda = 0.$$

This implies that off-diagonal components $\hat{\psi}^{4\alpha}$ can be created in the same way as electromagnetic potentials A^α, and hence should resemble electromagnetic terms.

[1] J. M. Overduin and P. S. Wesson, Phys. Rep. **283**, 303 (1997).
[2] H. Liu and S. Wanzhong, Gen. Relativ. Gravit. **20**, 303 (1988).
[3] H. Liu and P. S. Wesson, Int. J. Mod. Phys. D **3**, 627 (1994).
[4] A. Billyard and P. S. Wesson, Gen. Relativ. Gravit. **28**, 129 (1996).
[5] W. N. Sajko, P. S. Wesson, and H. Liu, J. Math. Phys. **39**, 2193 (1998).
[6] P. S. Wesson and H. Liu, Int. J. Theor. Phys. **36**, 1865 (1997).
[7] H. Liu and P. S. Wesson, Int. J. Mod. Phys. D **5**, 737 (1998).
[8] E. A. Matute, Class. Quantum Grav. **14**, 2771 (1997).
[9] G. W. Gibbons and D. L. Wiltshire, Ann. Phys. (N.Y.) **167**, 201 (1986).
[10] D. Kramer, H. Stephani, E. Herlt, M. MacCallum, and E. Schmutzer, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).
[11] K. Lake, P. Musgrave, and D. Pollney, GRTENSOR II Version 1.19 (Department of Physics, Queen's University, Canada, 1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Classical and quantum mechanics of jointed rigid bodies with vanishing total angular momentum

Toshihiro Iwai
Department of Applied Mathematics and Physics, Kyoto University, Kyoto 606-8501, Japan

(Received 19 October 1998; accepted for publication 8 February 1999)

A gauge theoretical treatment proves to have been successful in the study of systems of point particles; the center-of-mass system is made into a principal fiber bundle, on which is defined a natural connection. The gauge theoretical approach may be generalized to be applicable to a system of rigid bodies. The present article deals with a system of two identical axially symmetric cylinders jointed together by a special type of joint. This system is the model made by Kane and Scher and reformulated later by Montgomery, in order to study the falling cats who can land on their legs when released upside down. With the no-twist condition, the system turns out to have the configuration space diffeomorphic with SO(3), which is made into a principal O(2) bundle over $\mathbf{R}P^2$, the real projective space of dimension two, and endowed with a natural connection. An optimal control problem for this system with the vanishing total angular momentum is satisfactorily treated in this bundle picture. Along with a certain performance index, the Maximum Principle gives rise to a Hamiltonian system on the cotangent bundle $T^*(\mathrm{SO}(3))$ of SO(3). This Hamiltonian system is shown to admit a symmetry group O(2), which is not the structure group, but comes from the material symmetry of the respective cylinders. Moreover, quantization of this ''classical'' system is carried out, giving rise to a quantum system with the constraints of the vanishing total angular momentum. Through the symmetry by the structure group O(2), the reduction procedure is performed for both the classical and the quantum systems. It then turns out that the respective reduced systems, classical and quantum, admit the material symmetry group O(2), in general.
[S0022-2488(99)01905-2]

I. INTRODUCTION

Systems of point particles have been successfully studied in the gauge theoretical manner; the center-of-mass system is made into a principal fiber bundle, on which is defined a natural connection. The bundle picture (or gauge theoretical method) for systems of point particles dates back to 1984 when Guichardet[1] first defined rotational and vibrational motions precisely and thereby showed that rotations cannot be separated from vibrations by the use of the connection theory or the gauge theory. As Guichardet[1] pointed out, the nonseparability of rotation and vibration has been known to cats who can land on their legs when launched in the air, that is, who can accomplish a rotation after a vibrational motion. The gauge theoretical method has been developed to study and analyze classical and quantum mechanics of many-particle systems.[2–10]

Systems of rigid bodies are more suitable than systems of point particles in the understanding of falling cats. In fact, before introduction of the bundle picture for many-particle systems, Kane and Scher[11] considered a system of rigid bodies as a model of the falling cat, and later Montgomery[12] gave it a bundle picture. In this article, the gauge theoretical approach is generalized to be applicable to systems of rigid bodies. In particular, a system of two identical axially symmetric cylinders jointed along their symmetry axes by a special kind of joint is studied in an explicit manner, after Kane and Scher[11] and Montgomery.[12] Under the no-twist condition, the present system has the configuration space diffeomorphic with SO(3), which is made into an O(2) principal fiber bundle over $\mathbf{R}P^2$, the real projective space of dimension two. This fact was first observed by Montgomery.[12] A natural connection is defined on this bundle, like the natural connection for systems of point particles. In this bundle picture, a control problem is well formulated for the jointed cylinders to move with the vanishing total angular momentum. Along with a performance index which describes the horizontal (or vibrational) kinetic energy, the Maximum Principle applied to the optimal control problem provides an optimal Hamiltonian system on the cotangent bundle $T^*(\mathrm{SO}(3))$ of SO(3), whose Hamiltonian flows, when projected on SO(3), bring about optimal paths for the control problem. The reduction of the optimal Hamiltonian system is performed through the structure group symmetry to show that, in general, the reduced phase space is not the cotangent bundle $T^*(\mathbf{R}P^2)$ of $\mathbf{R}P^2$, but a double cover of $T^*(\mathbf{R}P^2)$. Further, the reduced system is shown to admit the ''material'' symmetry group O(2).

With the optimal Hamiltonian system is associated a quantum system on SO(3), which may be looked upon as a quantum system with the constraint of the vanishing total angular momentum. In a simplified case, the energy eigenvalues and eigenfunctions can be calculated explicitly. Like the ''classical'' system, the quantum system on SO(3) is reduced, through the structure group symmetry, to a system defined on the vector bundle over $\mathbf{R}P^2$ with fiber $\mathbf{C}^2$, in general. It will be shown further that the ''material'' symmetry group O(2) has unitary representations in energy eigenspaces.

The organization of this article is as follows: Section II contains the setting up of a system of two identical axially symmetric cylinders jointed together by a special type of joint. The configuration space of this system is SO(3)$\times$SO(3), clearly made into an SO(3) principal fiber bundle with base space SO(3), a local section of which is obtained explicitly. In Sec. III, the no-twist condition is taken into account, with which the motion of the jointed cylinders is restricted to that without twist at the joint. It turns out that the no-twist configuration space is diffeomorphic with SO(3), which is made into a principal O(2) bundle over $\mathbf{R}P^2$. A close look is given at the structure group O(2) and at its action on SO(3) along with the explicit expression in terms of local coordinates. Section IV is concerned with the connection and the curvature defined on the principal bundle $\mathrm{SO}(3)\to\mathbf{R}P^2$. By using the connection, a control problem is well formulated for the jointed cylinders to move under the condition of the vanishing total angular momentum. In Sec. V, a metric is defined on SO(3) in association with the kinetic energy of the jointed cylinders. This metric is decomposed into the horizontal (or vibrational) and vertical (or rotational) components. By the use of the horizontal part of the metric, a performance index is defined to be the vibrational energy of the jointed cylinders, and thereby the above-mentioned control problem is set up as an optimal control problem. The Maximum Principle can then be applied to give rise to an optimal Hamiltonian system on the cotangent bundle $T^*(\mathrm{SO}(3))$ of SO(3), whose Hamiltonian function is equal to the vibrational energy of the jointed cylinders. In Sec. VI, the reduction of the optimal Hamiltonian system is performed through the symmetry by the structure group O(2) action. The reduced phase space proves to be not $T^*(\mathbf{R}P^2)$ but a double cover of it, in general. Section VII is concerned with the symmetry of the jointed cylinders, which results from the material symmetry of the cylinders. It is shown that both the optimal Hamiltonian system and its reduced system admits O(2) as a symmetry group. Section VIII is devoted to the quantization of the optimal Hamiltonian system obtained in Sec. V. The vibrational energy functional of wave functions gives rise to a Hamiltonian operator corresponding to the optimal Hamiltonian function. If the Hamiltonian operator is chosen in a simplified form, its eigenvalues can be obtained quite easily. Section IX contains the reduction and the symmetry of the quantum system set up in Sec. VIII. The reduced quantum system is defined on the vector bundle associated with the O(2) bundle $\mathrm{SO}(3)\to\mathbf{R}P^2$. The energy eigenspaces are obtained, in which the symmetry group O(2) is represented unitarily.

II. SETTING UP

Let a rigid body be laid in the space $\mathbf{R}^3$ in such a way that the principal axes of the inertia tensor relative to the center of mass of the rigid body are parallel to the standard basis $\mathbf{e}_j$, j

$=1,2,3$, of $\mathbf{R}^3$. Let $\mathbf{X}_\alpha$ denote the position vector of a point of the rigid body with respect to the frame of the principal axes placed at the center of mass, where α is a continuous parameter. Then a generic position vector $\mathbf{x}_\alpha$ of points of the rigid body is expressed as

$$\mathbf{x}_\alpha=\mathbf{r}+g\mathbf{X}_\alpha, \tag{2.1}$$

where $\mathbf{r}$ is the position vector of the center of mass of the rigid body and $g\in \mathrm{SO}(3)$. Equation (2.1) implies that one needs a pair $(\mathbf{r},g)\in\mathbf{R}^3\times\mathbf{SO}(3)$ to specify the position and the attitude of the rigid body.

In what follows, our interest will center on a system of two identical rigid bodies in the center-of-mass system. This means that we assume that the center of mass of the two identical rigid bodies are kept fixed at the origin of $\mathbf{R}^3$. Further, we assume that the rigid bodies are axially symmetric cylinders jointed together along their symmetry axes by a special type of joint. The joint is supposed to be ball-and-socket; that is, the joint will give no constraint on the relative motion of the cylinders other than that they are jointed. We assume that the symmetry axes of the respective cylinder are parallel to $g_1\mathbf{e}_3$ and $g_2\mathbf{e}_3$, where $g_1,g_2\in\mathrm{SO}(3)$ denote the attitude of each body half. Let ℓ be the distance of each body half's center of mass from the joint. Then the center of mass of each body half is pointed by $\mathbf{r}_i=\ell g_i\mathbf{e}_3-\ell/2(g_1\mathbf{e}_3+g_2\mathbf{e}_3)$, $i=1,2$. This means that $\mathbf{r}_1$ and $\mathbf{r}_2$ are determined by g_1 and g_2. Hence, the configuration of our jointed cylinders is described by a pair (g_1,g_2) of rotation matrices. Thus, the configuration space of our system is the product space $Q:=\mathrm{SO}(3)\times\mathrm{SO}(3)$, on which the rotation group SO(3) acts in a natural manner,

$$(g_1,g_2)\mapsto(kg_1,kg_2),\quad k\in\mathrm{SO}(3). \tag{2.2}$$

Since SO(3) is compact and acts freely on Q, the factor space $Q/\mathrm{SO}(3)$ becomes a manifold. The natural projection to the factor space is obviously given by

$$\pi:(g_1,g_2)\mapsto g_1^{-1}g_2, \tag{2.3}$$

which means that the factor space is diffeomorphic with SO(3). This factor space is called the shape space, which consists of shapes of our jointed cylinders.

Let $g_1^{-1}g_2\in\mathrm{SO}(3)$ be expressed in terms of the Euler angles such that

$$g_1^{-1}g_2=e^{-\theta_1\hat{\mathbf{e}}_3}e^{\psi\hat{\mathbf{e}}_1}e^{\theta_2\hat{\mathbf{e}}_3},\quad 0\leqslant\theta_1\leqslant 2\pi,\;\; 0\leqslant\psi\leqslant\pi,\;\; 0\leqslant\theta_2\leqslant 2\pi, \tag{2.4}$$

where $\hat{\mathbf{e}}_j$, $j=1,2,3$, are the 3×3 skew-symmetric matrices defined, in general, to be

$$\hat{\mathbf{a}}=\begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix}\quad \text{for}\;\; \mathbf{a}=\sum_{j=1}^{3} a_j\mathbf{e}_j\in\mathbf{R}^3. \tag{2.5}$$

Then we are allowed to describe (g_1,g_2) as

$$\begin{aligned} g_1&=he^{-\psi/2\hat{\mathbf{e}}_1}e^{\theta_1\hat{\mathbf{e}}_3},\\ g_2&=he^{\psi/2\hat{\mathbf{e}}_1}e^{\theta_2\hat{\mathbf{e}}_3}, \end{aligned}\quad h\in\mathrm{SO}(3). \tag{2.6}$$

This expression is capable of natural interpretation; the body half 1 (respectively, 2) is first laid in the space $\mathbf{R}^3$ in such a way that its symmetry axis is parallel to $\mathbf{e}_3$, and then rotated around the $\mathbf{e}_3$ axis by θ_1 (respectively, by θ_2) and further on around the $\mathbf{e}_1$ axis by $-\psi/2$ (respectively, by $\psi/2$). Thus ψ stands for the angle made by the symmetry axes of respective body halves, and θ_1 and θ_2 are the rotation angles of the respective body halves around their symmetry axes. The jointed rigid bodies laid tentatively along the $\mathbf{e}_2-\mathbf{e}_3$ plane are acted on by $h\in\mathrm{SO}(3)$ to take a generic attitude in the space $\mathbf{R}^3$.

We have to point out here that if $\psi=0$ or π a pair of rotation angles (θ_1,θ_2) are not determined uniquely from the shape of the jointed cylinders. In fact, if $\psi=0$ (respectively, $\psi=\pi$), the shapes of the jointed cylinders designated by the pairs of rotation angles (θ_1,θ_2) and (θ_1+c,θ_2+c) [respectively, (θ_1,θ_2) and (θ_1-c,θ_2+c)] with $c\in\mathbf{R}^3$ are not distinguished from each other. This singularity of the Euler angles at $\psi=0$ or π can be observed from (2.4) as well.

Equation (2.6) also describes the local triviality of our principal SO(3) bundle, $Q\to SO(3)$. In fact, for an open subset U of SO(3) designated by the Euler angles (θ_1,θ_2,ψ) with $\psi\neq 0$, π, a local section $\sigma=(\sigma_1,\sigma_2)$: U→SO(3)×SO(3) is defined to be

$$\begin{aligned}\sigma_1(\theta_1,\theta_2,\psi)&=e^{-\psi/2\hat{\mathbf{e}}_1}e^{\theta_1\hat{\mathbf{e}}_3},\\ \sigma_2(\theta_1,\theta_2,\psi)&=e^{\psi/2\hat{\mathbf{e}}_1}e^{\theta_2\hat{\mathbf{e}}_3},\end{aligned}\tag{2.7}$$

and thereby the local triviality $\pi^{-1}(\mathrm{U})\cong\mathrm{U}\times SO(3)$ is realized as $(g_1,g_2)=(h\sigma_1,h\sigma_2)$, as is seen from (2.6).

Another local coordinate system is also possible. If we start with

$$g_1^{-1}g_2=e^{-\theta_1\hat{\mathbf{e}}_2}e^{2\phi\hat{\mathbf{e}}_1}e^{\theta_2\hat{\mathbf{e}}_2},\quad 0\leqslant\theta_1\leqslant 2\pi,\ \ 0\leqslant 2\phi\leqslant\pi,\ \ 0\leqslant\theta_2\leqslant 2\pi,\tag{2.8}$$

in place of (2.4), we will obtain the following local section in place of (2.7);

$$\begin{aligned}\tau_1(\theta_1,\theta_2,\phi)&=e^{-\phi\hat{\mathbf{e}}_1}e^{\theta_1\hat{\mathbf{e}}_2},\\ \tau_2(\theta_1,\theta_2,\phi)&=e^{\phi\hat{\mathbf{e}}_1}e^{\theta_2\hat{\mathbf{e}}_2}.\end{aligned}\tag{2.9}$$

This local section allows of an interpretation; at first we assume that the symmetry axes of respective body halves are oriented in such a way that the positive direction is outward from the joint. With this assumption, the body half 1 (respectively, 2) is set in such a way that its symmetry axis is in parallel with $-\mathbf{e}_2$ (respectively, $\mathbf{e}_2$), then rotated around $\mathbf{e}_2$ by θ_1 (respectively, θ_2) and further on around $\mathbf{e}_1$ by $-\phi$ (respectively, ϕ). Thus, the jointed body halves are laid tentatively in the $\mathbf{e}_2-\mathbf{e}_3$ plane, which will be rotated, as a whole, by the action of SO(3) to take its generic position;

$$(g_1,g_2)=(k\tau_1,k\tau_2),\quad k\in SO(3).\tag{2.10}$$

III. NO-TWIST CONFIGURATION SPACE

Our system of jointed cylinders is a model of the cat that somersaults when launched in the air. Kane and Scher[11] imposed the no-twist condition on the jointed cylinders, which implies that two of body halves rotate around the respective symmetry axes without twist at the joint. According to Montgomery,[12] the no-twist condition is characterized as follows: Let ϵ be a rotation such that $\epsilon^2=I$. Then an involution i acting on SO(3)×SO(3) is defined as

$$i(g_1,g_2)=(\epsilon g_2\epsilon,\epsilon g_1\epsilon).\tag{3.1}$$

The no-twist configuration space is defined to be the subset of fixed points for i,

$$Q_0=\{(g_1,g_2)\in SO(3)\times SO(3);\ \ i(g_1,g_2)=(g_1,g_2)\}.\tag{3.2}$$

The fixed-point condition $\epsilon g_1\epsilon=g_2$ implies that Q_0 is diffeomorphic with SO(3),

$$Q_0=\{(g,\epsilon g\epsilon);\quad g\in SO(3)\}\cong SO(3).\tag{3.3}$$

Further, to see that Q_0 is indeed subject to the no-twist condition of Kane and Scher, we choose to express ϵ and an arbitrary point (g_1,g_2) of the total space, respectively, as

$$\epsilon = e^{\pi \hat{\mathbf{e}}_2}, \quad \begin{cases} g_1 = e^{\chi_1 \hat{\mathbf{e}}_2} e^{\phi_1 \hat{\mathbf{e}}_1} e^{\theta_1 \hat{\mathbf{e}}_2}, \\ g_2 = e^{\chi_2 \hat{\mathbf{e}}_2} e^{\phi_2 \hat{\mathbf{e}}_1} e^{\theta_2 \hat{\mathbf{e}}_2}. \end{cases} \tag{3.4}$$

Then, the condition $\epsilon g_2 \epsilon = g_1$ implies that

$$\chi_1 = \chi_2, \quad \phi_1 = -\phi_2, \quad \theta_1 = \theta_2, \tag{3.5}$$

where use has been made of $e^{\pi \hat{\mathbf{e}}_2} e^{t \hat{\mathbf{e}}_1} = e^{-t \hat{\mathbf{e}}_1} e^{\pi \hat{\mathbf{e}}_2}$, $t \in \mathbf{R}$. Setting $\chi_1 = \chi_2 = \chi$, $\phi_1 = -\phi_2 = -\phi$, $\theta_1 = \theta_2 = \theta$ in (3.4), one has

$$\begin{aligned} g_1 &= e^{\chi \hat{\mathbf{e}}_2} e^{-\phi \hat{\mathbf{e}}_1} e^{\theta \hat{\mathbf{e}}_2}, \\ g_2 &= e^{\chi \hat{\mathbf{e}}_2} e^{\phi \hat{\mathbf{e}}_1} e^{\theta \hat{\mathbf{e}}_2}, \end{aligned} \quad 0 \leqslant \chi \leqslant 2\pi, \ 0 \leqslant \phi \leqslant \pi, \ 0 \leqslant \theta \leqslant 2\pi, \tag{3.6}$$

which is a specialization of (2.9) and (2.10). We are then allowed to interpret that the body halves are rotated by the same angle θ around respective symmetry axes without twist at the joint. However, the structure group SO(3) has reduced to a subgroup.

Now we wish to study to what subgroup the structure group reduces. A first glance at (3.6) will tempt us to be convinced that the structure group SO(3) has reduced to a one-parameter subgroup SO(2) which consists of $e^{\chi \hat{\mathbf{e}}_2}$. However, the reduced structure group is not SO(2) but O(2).[12] To see this, one has to be aware of the fact that if ϕ is increased by $\pi (\phi \mapsto \phi + \pi)$, each of the body halves is rotated about $\mathbf{e}_1$ to be set in the direction opposite to the previous one, so that the shape of the jointed cylinders is not altered. In fact, one verifies from (3.6) that $\pi(g_1, g_2) = e^{-\theta \hat{\mathbf{e}}_2} e^{2\phi \hat{\mathbf{e}}_1} e^{\theta \hat{\mathbf{e}}_2}$ does not change under the transformation $\theta \mapsto \phi + \pi$. This means that the jointed cylinders do not change in shape. Under the transformation $\phi \mapsto \phi + \pi$, one has the factor $e^{\chi \hat{\mathbf{e}}_2} e^{\pi \hat{\mathbf{e}}_1}$ in the expression of g_1 and g_2 in (3.6). Written out, $e^{\chi \hat{\mathbf{e}}_2} e^{\pi \hat{\mathbf{e}}_1}$ is expressed as

$$e^{\chi \hat{\mathbf{e}}_2} e^{\pi \hat{\mathbf{e}}_1} = \begin{pmatrix} \cos\chi & 0 & -\sin\chi \\ 0 & -1 & 0 \\ -\sin\chi & 0 & -\cos\chi \end{pmatrix}, \tag{3.7}$$

which describes an action of O(2) in the $\mathbf{e}_3 - \mathbf{e}_1$ plane. Thus the reduced structure group turns out to be O(2);

$$\mathrm{O}(2) = \mathrm{SO}(2) \cup \mathrm{SO}(2) e^{\pi \hat{\mathbf{e}}_1} \quad \text{with } \mathrm{SO}(2) = \{ e^{\chi \hat{\mathbf{e}}_2} \}. \tag{3.8}$$

We notice, for confirmation, that this subgroup is indeed closed under multiplication. In fact, one can easily verify that

$$e^{\chi_1 \hat{\mathbf{e}}_2} e^{\pi \hat{\mathbf{e}}_1} e^{\chi_2 \hat{\mathbf{e}}_2} e^{\pi \hat{\mathbf{e}}_1} = e^{(\chi_1 - \chi_2) \hat{\mathbf{e}}_2} \tag{3.9}$$

which also implies that O(2) is not Abelian.

Accompanying the reduction of the structure group, the base space will reduce to $Q_0/\mathrm{O}(2) \cong \mathrm{SO}(3)/\mathrm{O}(2)$, which is diffeomorphic with $S^2/Z_2 \cong \mathbf{R}P^2$, the real projective space of dimension two. Restricting the projection given in (2.3) to the subset Q_0 ($\pi_0 := \pi|_{Q_0}$), one must have $\pi_0(Q_0) \cong \mathbf{R}P^2$. That is, $\pi_0(Q_0)$ must be an embedding of $\mathbf{R}P^2$ in SO(3), which can be verified as follows: From (3.6) it follows that

$$\pi_0(g_1, g_2) = g_1^{-1} g_2 = e^{-\theta \hat{\mathbf{e}}_2} e^{2\phi \hat{\mathbf{e}}_1} e^{\theta \hat{\mathbf{e}}_2}. \tag{3.10}$$

We then set

$$\begin{aligned} x_3 &= \sin\phi \cos\theta, \\ x_1 &= \sin\phi \sin\theta, \\ x_2 &= \cos\phi, \end{aligned} \quad 0 \leqslant \theta \leqslant 2\pi, \ 0 \leqslant \phi \leqslant \pi/2, \tag{3.11}$$

and form the matrix $X=(x_i x_j)$, $i,j=1,2,3$, which represents a point of $\mathbf{R}P^2$. Then we obtain, after calculation,

$$\mathrm{diag}(1,-1,-1)(I_3-2X)\mathrm{diag}(1,1,-1)=e^{-\theta\hat{\mathbf{e}}_2}e^{2\phi\hat{\mathbf{e}}_1}e^{\theta\hat{\mathbf{e}}_2}, \tag{3.12}$$

where I_3 is the 3×3 identity matrix. This provides the embedding of $\mathbf{R}P^2$ into SO(3). From (3.10) and (3.12), it follows that $\pi_0(Q_0)\cong\mathbf{R}P^2$. We note in addition that

$$e^{-\theta\hat{\mathbf{e}}_2}e^{2\phi\hat{\mathbf{e}}_1}e^{\theta\hat{\mathbf{e}}_2}=e^{2\phi(\hat{\mathbf{e}}_1\cos\theta+\hat{\mathbf{e}}_3\sin\theta)}, \tag{3.13}$$

which means that $(2\phi,\theta)$ serve as ''polar'' coordinates of the projective plane $\mathbf{R}P^2$. Thus we have the following:

Proposition 1: The configuration space Q_0 of the jointed cylinders with no-twist condition is diffeomorphic with SO(3), which is made into a principal O(2) bundle with base space $\mathbf{R}P^2$. Equation (3.6) provides a local coordinate system of this bundle.

In conclusion, we are to describe the structure group action in terms of local coordinates (χ,ϕ,θ). Letting $e^{t\hat{\mathbf{e}}_2}$ and $e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}$ act to the left on Q_0, one obtains, from (3.6),

$$\begin{aligned} e^{t\hat{\mathbf{e}}_2}&: \quad (\chi,\phi,\theta)\mapsto(\chi+t,\phi,\theta),\\ e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}&: \quad (\chi,\phi,\theta)\mapsto(-\chi+t,\phi+\pi,\theta), \end{aligned} \tag{3.14}$$

respectively, where use has been made of $e^{\pi\hat{\mathbf{e}}_1}e^{\chi\hat{\mathbf{e}}_2}=e^{-\chi\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}$.

IV. CONNECTION AND CURVATURE

We are interested in making our system move with the condition of the vanishing total angular momentum. To this end, we calculate both the total angular momentum and the total inertia tensor of our system, and thereby define a connection form.

From (2.1), the angular momentum of a rigid body is given by and calculated as

$$\sum_\alpha m_\alpha \mathbf{x}_\alpha\times d\mathbf{x}_\alpha=m\mathbf{r}\times d\mathbf{r}+gA_0\mathbf{\Theta}, \tag{4.1}$$

where Σ_α stands for the sum throughout the body, $m=\Sigma_\alpha m_\alpha$ is the total mass of the rigid body, A_0 is the inertia tensor of the rigid body around its center of mass, which is expressed, from the assumption we have made in the last paragraph of Sec. II, as

$$A_0=\mathrm{diag}(I_1,I_2,I_3), \quad I_3=I_1 \tag{4.2}$$

with respect to the standard basis $\mathbf{e}_j$, $j=1,2,3$, and $\mathbf{\Theta}$ is the vector of one-forms defined through

$$g^{-1}dg=\hat{\mathbf{\Theta}}. \tag{4.3}$$

The total angular momentum, $\mathbf{\Lambda}$, of our system is the sum of the angular momentum of the respective cylinders;

$$\mathbf{\Lambda}=\sum_{i=1}^{2} m\mathbf{r}_i\times d\mathbf{r}_i+\sum_{i=1}^{2} g_i A_0\mathbf{\Theta}_i, \tag{4.4}$$

where the subscripts $i=1,2$ indicate that the subscripted quantities ($\mathbf{r}_i$, etc.) are concerned with the respective body halves 1 and 2.

We are going to calculate the total angular momentum $\mathbf{\Lambda}$ in terms of the local coordinates given in (3.6). We assume that our system is a center-of-mass system, that is, the center of mass of the jointed cylinders is kept fixed at the origin. The position vectors, $\mathbf{r}_1$ and $\mathbf{r}_2$, of the center of mass of the respective body halves are then given by

$$\mathbf{r}_1 = -\ell (g_1 \mathbf{e}_2 \cdot \mathbf{e}_2)\mathbf{e}_2, \quad \mathbf{r}_2 = \ell (g_2 \mathbf{e}_2 \cdot \mathbf{e}_2)\mathbf{e}_2, \tag{4.5}$$

where ℓ is the distance of the center of mass of the respective body halves from the joint, and g_1, g_2 are the matrices given in (3.6). A straightforward calculation then provides

$$\mathbf{r}_1 \times d\mathbf{r}_1 + \mathbf{r}_2 \times d\mathbf{r}_2 = 0, \tag{4.6}$$

which implies that the first term of the right-hand side of (4.4) vanishes. As for the latter term of the right-hand side of (4.4), we obtain, after a straightforward calculation,

$$\sum_{i=1}^{2} g_i A_0 \mathbf{\Theta}_i = 2((I_1 \sin^2\phi + I_2 \cos^2\phi)d\chi + I_2 \cos\phi\, d\theta)\mathbf{e}_2. \tag{4.7}$$

From (4.4) along with (4.6), and (4.7), it follows that

$$\mathbf{\Lambda} = 2((I_1 \sin^2\phi + I_2 \cos^2\phi)d\chi + I_2 \cos\phi d\theta)\mathbf{e}_2. \tag{4.8}$$

We turn to the inertia tensor of the jointed cylinders. From (2.1), the inertia tensor of a rigid body as a map of $\mathbf{R}^3$ to $\mathbf{R}^3$ is defined to be and calculated as

$$\sum_{\alpha} m_\alpha \mathbf{x}_\alpha \times (\mathbf{v} \times \mathbf{x}_\alpha) = m\mathbf{r} \times (\mathbf{v} \times \mathbf{r}) + gA_0 g^{-1}\mathbf{v}, \tag{4.9}$$

where $\mathbf{v} \in \mathbf{R}^3$, and m and A_0 are the same as those in (4.1). The total inertia tensor, A_q: $\mathbf{R}^3 \to \mathbf{R}^3$, $q \in Q_0$, is the sum of those for respective body halves;

$$A_q \mathbf{v} = \sum_{i=1}^{2} m\mathbf{r}_i \times (\mathbf{v} \times \mathbf{r}_i) + \sum_{i=1}^{2} g_i A_0 g_i^{-1}\mathbf{v}, \quad q \in Q_0. \tag{4.10}$$

As for the first term on the right-hand side of (4.10), one obtains, after a calculation, the matrix,

$$\left(\mathbf{e}_j \cdot \sum_{i=1}^{2} (\mathbf{r}_i \times (m\mathbf{e}_k \times \mathbf{r}_i)) \right)_{j,k=1,2,3} = 2m\ell^2 \operatorname{diag}(\cos^2\phi, 0, \cos^2\phi). \tag{4.11}$$

Further calculation with g_1 and g_2 given in (3.6) provides

$$\sum_{i=1}^{2} g_i A_0 g_i^{-1} = 2\begin{pmatrix} A_1\cos^2\chi + A_3\sin^2\chi & 0 & -A_1\cos\chi\sin\chi + A_3\cos\chi\sin\chi \\ 0 & A_2 & 0 \\ -A_1\cos\chi\sin\chi + A_3\cos\chi\sin\chi & 0 & A_1\sin^2\chi + A_3\cos^2\chi \end{pmatrix}, \tag{4.12}$$

where

$$\begin{aligned} A_1 &= I_1, \\ A_2 &= I_1 \sin^2\phi + I_2 \cos^2\phi, \\ A_3 &= I_1 \cos^2\phi + I_2 \sin^2\phi. \end{aligned} \tag{4.13}$$

Hence the total inertia tensor A_q is put in the matrix form as the sum of the right-hand sides of (4.11) and (4.12).

In the same manner as in particle systems, the connection form is defined to be

$$\omega_q = \widehat{A_q^{-1}\Lambda_q}, \quad q \in Q_0, \tag{4.14}$$

and turns out, from (4.8), (4.11), and (4.12), to be expressed as

$$\omega = \left(d\chi + \frac{I_2 \cos\phi}{I_1 \sin^2\phi + I_2 \cos^2\phi} d\theta \right) \hat{\mathbf{e}}_2 . \tag{4.15}$$

We have to notice here that the ω satisfies the following equations:

$$\omega\left(\frac{\partial}{\partial\chi}\right) = \hat{\mathbf{e}}_2, \quad \frac{\partial}{\partial\chi} = \frac{d}{dt} e^{t\hat{\mathbf{e}}_2} q \bigg|_{t=0},$$

$$\omega_{gq} = \mathrm{Ad}_g\, \omega_q, \quad g \in \mathrm{O}(2), \quad q \in Q_0 . \tag{4.16}$$

In what follows, we introduce the parameter

$$\lambda = I_2 / I_1 . \tag{4.17}$$

Then the ω takes the form

$$\omega = \left(d\chi + \frac{\lambda \cos\phi\, d\theta}{\sin^2\phi + \lambda \cos^2\phi} \right) \hat{\mathbf{e}}_2 . \tag{4.18}$$

A vector field X on Q_0 is called horizontal (or vibrational) if it satisfies $\omega(X)=0$. One can easily observe, from (4.18), that the following vector fields are independent horizontal ones:

$$X_1 = \frac{\partial}{\partial\phi}, \quad X_2 = \frac{\partial}{\partial\theta} - \frac{\lambda\cos\phi}{\sin^2\phi + \lambda\cos^2\phi} \frac{\partial}{\partial\chi}. \tag{4.19}$$

A curve $q(t)$ in Q_0 is called horizontal, if its tangent vector $\dot{q}(t)$ is horizontal at any point $q(t)$. Hence, any smooth horizontal curve $q(t)$ is subject to a differential equation

$$\frac{dq}{dt} = u_1 X_1 + u_2 X_2, \tag{4.20}$$

where u_1, u_2 are functions of t. Since $\omega(\dot{q})=0$ is equivalent to $\Lambda_q(\dot{q})=0$, as is observed from (4.14), Eq. (4.20) means kinematically that a curve $q(t)$ satisfying (4.20) describes a motion of the no-twist jointed cylinders with the vanishing total angular momentum.

The curvature of the connection ω is defined, in general, to be

$$\Omega = d\omega - \omega \wedge \omega, \tag{4.21}$$

where the minus sign on the right-hand side of (4.21) is due to the left action of the structure group, in our case. We notice here that from (4.16) the Ω is also subject to the transformation

$$\Omega_{gq} = \mathrm{Ad}_g\, \Omega_q, \quad g \in \mathrm{O}(2), \quad q \in Q_0 . \tag{4.22}$$

Applying the formula (4.21) to (4.18), we obtain the connection form Ω on Q_0,

$$\Omega = -\lambda \frac{1 + (1-\lambda)\cos^2\phi}{(\sin^2\phi + \lambda\cos^2\phi)^2} \sin\phi\, d\phi \wedge d\theta\ \hat{\mathbf{e}}_2 . \tag{4.23}$$

A casual glance at (4.23) makes us tend to think of Ω as defined on the base space $\mathbf{R}P^2$, since (4.23) is independent of χ, the vertical coordinate. However, the curvature form Ω defined on Q_0 does not project to a two form on $\mathbf{R}P^2$. This is because Ω is not invariant under the action of the structure group O(2). In fact, for $g=e^{\pi\hat{\mathbf{e}}_1}\in \mathrm{O}(2)$, on account of (4.22) along with $e^{\pi\hat{\mathbf{e}}_1}\hat{\mathbf{e}}_2 e^{-\pi\hat{\mathbf{e}}_1}=-\hat{\mathbf{e}}_2$, the curvature form Ω is subject to $\Omega_{gq}=-\Omega_q$.

The curvature form (4.23) seems to vanish if $\phi=0$, i.e., if the body halves get in line. However, it does not vanish, as we show in the following. The factor two-form $\sin\phi\, d\phi\wedge d\theta$ on the right-hand side of (4.23) does not vanish even if $\phi=0$, since this form is a kind of an area form defined in an open subset containing the ''origin'' of the polar coordinates $(2\phi,\theta)$ [see (3.13)] in the projective plane $\mathbf{R}P^2$. In fact, on setting $\xi_1=\sin(\phi/2)\cos\theta$, $\xi_2=\sin(\phi/2)\sin\theta$, one obtains $d\xi_1\wedge d\xi_2=\frac{1}{4}\sin\phi\, d\phi\wedge d\theta$. On the other hand, if $\lambda>2$, one has $1+(1-\lambda)\cos^2\phi>0$ in (4.23). However, the condition $\lambda>2$ is satisfied for any axially symmetric cylinder. In the same manner, we see that Ω does not vanish if $\phi=\pi$. Thus we observe that the curvature form vanishes nowhere on Q_0.

V. AN OPTIMAL HAMILTONIAN SYSTEM

Equation (4.20) defines a control problem, if u_1 and u_2 are regarded as control variables, and if a performance index is given. To define a performance index, we consider the metric on Q_0 which is associated with the kinetic energy of the no-twist jointed cylinders.

By using (2.1), the metric associated with the kinetic energy for a rigid body is given by and calculated as

$$\sum_{\alpha} m_\alpha d\mathbf{x}_\alpha\cdot d\mathbf{x}_\alpha=m\, d\mathbf{r}\cdot d\mathbf{r}+\mathbf{\Theta}\cdot A_0\mathbf{\Theta}, \tag{5.1}$$

where m is the mass of the rigid body and $\mathbf{\Theta}$ is the vector of one-forms defined by (4.3). The metric of our system is then defined to be

$$ds^2=\sum_{i=1}^{2} m\, d\mathbf{r}_i\cdot d\mathbf{r}_i+\sum_{i=1}^{2}\mathbf{\Theta}_i\cdot A_0\mathbf{\Theta}_i. \tag{5.2}$$

From (4.5), it follows that

$$\sum_{i=1}^{2} m\, d\mathbf{r}_i\cdot d\mathbf{r}_i=2m\ell^2\sin^2\phi\, d\phi^2. \tag{5.3}$$

In a similar manner, one obtains

$$\sum_{i=1}^{2}\mathbf{\Theta}_i\cdot A_0\mathbf{\Theta}_i=2I_1(\sin^2\phi\, d\chi^2+d\phi^2)+2I_2(\cos\phi\, d\chi+d\theta)^2. \tag{5.4}$$

The sum of the right-hand sides of (5.3) and (5.4) provides the metric ds^2, as is desired. In view of the connection form (4.18), we put the metric in the form

$$\begin{aligned} ds^2=2I_1&\left[(1+\kappa\sin^2\phi)d\phi^2+\frac{\lambda\sin^2\phi}{\sin^2\phi+\lambda\cos^2\phi}d\theta^2\right]\\ &+2I_1\left[(\sin^2\phi+\lambda\cos^2\phi)\left(d\chi+\frac{\lambda\cos\phi}{\sin^2\phi+\lambda\cos^2\phi}d\theta\right)^2\right], \end{aligned}\tag{5.5}$$

where

$$\kappa=m\ell^2/I_1. \tag{5.6}$$

The last term of the right-hand side of (5.5) is the vertical (or rotational) component which vanishes for X_1, X_2 given in (4.19), and the first term is the horizontal (or vibrational) component which vanishes for the vertical vector field $\partial/\partial\chi$, the infinitesimal generator of the structure group action. Since ds^2 is O(2) invariant, as is observed from the transformation (3.14) applied to (5.5), the metric ds^2 projects to the base space $Q_0/\mathrm{O}(2)\cong \mathbf{R}P^2$, defining a metric

$$ds_0^2 := 2I_1\left[(1+\kappa\sin^2\phi)d\phi^2 + \frac{\lambda\sin^2\phi}{\sin^2\phi+\lambda\cos^2\phi}d\theta^2\right]. \tag{5.7}$$

To be strict, ds_0^2 is defined through

$$ds_0^2(\partial/\partial\phi,\partial/\partial\theta) := ds^2(X_1,X_2), \quad \text{etc.}, \tag{5.8}$$

where X_1, X_2 are the horizontal vector fields given in (4.19), which project to $\partial/\partial\phi$, $\partial/\partial\theta$, respectively; $\pi_{0*}X_1 = \partial/\partial\phi$, $\pi_{0*}X_2 = \partial/\partial\theta$.

We proceed to the cotangent bundle $T^*(Q_0)\cong T^*(\mathrm{SO}(3))$ which is endowed with the standard one-form expressed as

$$\Theta = p_\chi d\chi + p_\phi d\phi + p_\theta d\theta, \tag{5.9}$$

where $(p_\chi, p_\phi, p_\theta)$ are related to tangent vectors $(\dot\chi, \dot\phi, \dot\theta)$ through ds^2;

$$p_\chi d\chi + p_\phi d\phi + p_\theta d\theta = ds^2\left(\dot\chi\frac{\partial}{\partial\chi} + \dot\phi\frac{\partial}{\partial\phi} + \dot\theta\frac{\partial}{\partial\theta}, \cdot\right). \tag{5.10}$$

Let us denote the horizontal metric tensor by

$$h_{ij} := ds^2(X_i, X_j), \quad i,j=1,2, \tag{5.11}$$

which are actually given in (5.7), because of (5.8).

Now we are going to consider the control problem (4.20), wishing to determine controls u_1, u_2 in such a way that the performance index

$$\frac{1}{2}\int_0^T \sum_{i,j=1}^{2} h_{ij}(q(t))u_i(t)u_j(t)dt \tag{5.12}$$

is minimized among all controls which steer the state $q(t)\in Q_0$ from an initial state q_0 to a final state q_1 in time T. To apply the Maximum Principle to this problem, we define the conjugate variables by

$$P_1 := \Theta(X_1) = p_\phi, \quad P_2 := \Theta(X_2) = p_\theta - \frac{\lambda\cos\phi}{\sin^2\phi+\lambda\cos^2\phi}p_\chi, \tag{5.13}$$

where X_1, X_2 are given by (4.19). Then the Maximum Principle tells us that optimal controls for normal extremals are determined so that the function

$$\mathcal{H} = \sum_{i=1}^{2} u_i P_i - \frac{1}{2}\sum_{i,j=1}^{2} h_{ij}u_i u_j \tag{5.14}$$

may take its maximum value in u_i. Thus one obtains the optimal control $u_i = \Sigma_{j=1}^2 h^{ij}P_j$ with $(h^{ij}) = (h_{ij})^{-1}$, and thereby the optimal Hamiltonian

$$H=\frac{1}{2}\sum_{i,j=1}^{2} h^{ij}P_iP_j=\frac{1}{4I_1}\left(\frac{1}{1+\kappa\sin^2\phi}p_\phi^2+\frac{\sin^2\phi+\lambda\cos^2\phi}{\lambda\sin^2\phi}\left(p_\theta-\frac{\lambda\cos\phi}{\sin^2\phi+\lambda\cos^2\phi}p_\chi\right)^2\right). \tag{5.15}$$

Thus we have the following:

Proposition 2: The optimal control problem described by (4.20) together with the performance index (5.12) gives rise to the Hamiltonian system $(T^*(Q_0),d\Theta,H)$, where $d\Theta$ is the differential of (5.9) and H is the Hamiltonian given by (5.15).

According to Montgomery,[11] abnormal extremals for $\mathcal{H}$ do not occur when the curvature does not vanish, so that our optimal control problem can be solved through the normal extremals only. Hence, optimal paths are obtained by projecting trajectories for the optimal Hamiltonian system $(T^*(Q_0),d\Theta,H)$ to the base space Q_0.

We note in addition that the Hamiltonian (5.15) is just equal to the horizontal (or vibrational) part of the total kinetic energy T that can be obtained in the usual manner as follows:

$$T=\frac{1}{4I_1}\left(\frac{1}{1+\kappa\sin^2\phi}p_\phi^2+\frac{\sin^2\phi+\lambda\cos^2\phi}{\lambda\sin^2\phi}\left(p_\theta-\frac{\lambda\cos\phi}{\sin^2\phi+\lambda\cos^2\phi}p_\chi\right)^2\right)$$
$$+\frac{1}{4I_1}\frac{1}{(\sin^2\phi+\lambda\cos^2\phi)}p_\chi^2. \tag{5.16}$$

In fact, the Lagrangian defined through (5.5) is Legendre transformed into (5.16) by using (5.10). Note that the last term of the right-hand side of (5.16) stands for the rotational energy associated with the total angular momentum. Thus we understand that the optimal Hamiltonian system describes the motion with the vanishing total angular momentum.

In conclusion, we should mention that the choice of the performance index is not unique, of course, rather it is in our hands. Hence, we may choose a simpler one. If we set $\kappa=0$, i.e., $\ell=0$, then the metric (5.5) becomes the metric known as that for a symmetrical top. Further, if $\lambda=1$ in addition, the metric (5.12) takes the form

$$ds^2_{\mathrm{can}}=2I_1(d\chi^2+d\phi^2+d\theta^2+2\cos\phi\,d\chi\,d\theta), \tag{5.17}$$

which is known, within a constant multiple, as the canonical metric on SO(3) expressed in the Euler angles. If we start with ds^2_{can}, the optimal Hamiltonian takes a simpler form,

$$H'=\frac{1}{4I_1}\left(p_\phi^2+\frac{1}{\sin^2\phi}(p_\theta-\cos\phi p_\chi)^2\right). \tag{5.18}$$

VI. REDUCTION

The equations of motion of the optimal Hamiltonian system $(T^*(Q_0),d\Theta,H)$ are obtained, as usual, through the Hamiltonian vector field X_H determined by $\iota(X_H)d\Theta=-dH$; $dp/dt=X_H$, $p\in T^*(Q_0)$. One of the equations is expressed as $dp_\chi/dt=-\partial H/\partial\chi=0$, so that $p_\chi=\mu=\mathrm{const}$. This constant of motion must be a consequence of the O(2) symmetry of our system. However, since O(2) is neither connected nor Abelian, we must be careful in treating the O(2) symmetry. First, the O(2) action on Q_0 is lifted to that on $T^*(Q_0)$ so that the lifted action may preserve the standard one-form Θ. According to (3.14), the O(2) action on Q_0 gives rise to the transformation $(d\chi,d\phi,d\theta)\mapsto(-d\chi,d\phi,d\theta)$ for $e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}$, so that one has $(p_\chi,p_\phi,p_\theta)\mapsto(-p_\chi,p_\phi,p_\theta)$ from the invariance of Θ. Thus the lifted action is expressed, in terms of local coordinates of $T^*(Q_0)$, as

$$e^{t\hat{\mathbf{e}}_2}:\begin{cases}(\chi,\phi,\theta)\mapsto(\chi+t,\phi,\theta),\\(p_\chi,p_\phi,p_\theta)\mapsto(p_\chi,p_\phi,p_\theta),\end{cases}$$
$$e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}:\begin{cases}(\chi,\phi,\theta)\mapsto(-\chi+t,\phi+\pi,\theta),\\(p_\chi,p_\phi,p_\theta)\mapsto(-p_\chi,p_\phi,p_\theta).\end{cases}\tag{6.1}$$

The momentum map J associated with the O(2) symmetry, which takes its value in the Lie algebra of O(2), proves to be adjoint-equivariant on account of (6.1) and of $e^{\pi\hat{\mathbf{e}}_1}\hat{\mathbf{e}}_2e^{-\pi\hat{\mathbf{e}}_1}=-\hat{\mathbf{e}}_2$;

$$J:=\Theta(\partial/\partial\chi)\hat{\mathbf{e}}_2=p_\chi\hat{\mathbf{e}}_2,$$
$$J(gp)=\mathrm{Ad}_g\,J(p),\quad g\in\mathrm{O}(2),\quad p\in T^*(Q_0).\tag{6.2}$$

We should note here that J is not O(2)-invariant. However, the Hamiltonian (5.15) is O(2)-invariant, i.e., invariant under (6.1).

The submanifold determined by $J^{-1}(\mu)$ with $\mu\neq 0$ is not invariant under the action of O(2) but invariant under that of SO(2). Thus the reduced phase space should be

$$P_\mu:=J^{-1}(\mu)/\mathrm{SO}(2)=J^{-1}(\mu)\times_{\mathrm{O}(2)}(\mathrm{O}(2)/\mathrm{SO}(2))=J^{-1}(\mu)\times_{\mathrm{O}(2)}\mathbf{Z}_2,\tag{6.3}$$

so that P_μ is a fiber bundle with base space $T^*(\mathbf{R}P^2)$ and fiber $\mathbf{Z}_2$, that is, a double cover of $T^*(\mathbf{R}P^2)$. The reduced symplectic form $d\Theta_\mu$ and the reduced Hamiltonian H_μ are apparently expressed in terms of local coordinates as

$$d\Theta_\mu=dp_\phi\wedge d\phi+dp_\theta\wedge d\theta,\tag{6.4}$$

$$H_\mu=\frac{1}{4I_1}\left(\frac{1}{1+\kappa\sin^2\phi}p_\phi^2+\frac{\sin^2\phi+\lambda\cos^2\phi}{\lambda\sin^2\phi}\left(p_\theta-\frac{\lambda\cos\phi}{\sin^2\phi+\lambda\cos^2\phi}\mu\right)^2\right),\tag{6.5}$$

respectively. It is to be noted here that H_μ is a function not on $T^*(\mathbf{R}P^2)$ but on P_μ. In fact, it is not a single-valued function on $T^*(\mathbf{R}P^2)$, that is, it is not invariant when $(\phi,\theta,p_\phi,p_\theta)$ is replaced by $(\phi+\pi,\theta,p_\phi,p_\theta)$. If $\mu=0$, the phase space P_μ becomes $T^*(\mathbf{R}P^2)$. In this case, H_0 is a function on $T^*(\mathbf{R}P^2)$, of course.

Theorem 3: The optimal Hamiltonian system $(T^*(Q_0),d\Theta,H)$ is reduced to a Hamiltonian system $(P_\mu,d\Theta_\mu,H_\mu)$, where P_μ with $\mu\neq 0$ is a double cover of $T^*(\mathbf{R}P^2)$, and $d\Theta_\mu$ and H_μ are given, respectively, by (6.4) and (6.5). If $\mu=0$, the P_μ becomes $T^*(\mathbf{R}P^2)$.

In conclusion, we have to point out that the mechanical system $(T^*(Q_0),d\Theta,T)$ has also the constant of motion $p_\chi=\mu=\mathrm{const}$, and reduces to $(P_\mu,d\Theta_\mu,T_\mu)$, where T_μ is defined from (5.16) along with $p_\chi=\mu$. If $\mu=0$, then $T_0=H_0$, so that the two reduced systems coincide. This implies that if $p_\chi=0$ the mechanical and the controlled system have the same trajectories.

VII. SYMMETRY

We wish to discuss the symmetry of the jointed cylinders. Since each of the body halves is an axially symmetric cylinder, the system of jointed cylinders has to admit symmetry arising from the axial symmetry of the body halves.

We have assumed in Sec. II that the reference attitude of the respective body halves are such that their principal axes are set in parallel to the standard basis $\mathbf{e}_j$, $j=1,2,3$, of $\mathbf{R}^3$. However, the choice of the reference attitude of the body halves is not unique. We can choose to set the body halves in such a way that the principal axes are parallel to $h\mathbf{e}_j$, $h\in\mathrm{SO}(3)$. Then the inertia tensor of each body half becomes expressed as $h^{-1}A_0h$, where A_0 is the diagonal inertia tensor given in (4.2). For the rigid body in this attitude, the position vector of a generic point from its center of mass is described as $gh\mathbf{X}_\alpha$ in place of $g\mathbf{X}_\alpha$. This implies that SO(3) acts on the configuration space Q_0 to the right,

$$(g_1,g_2)\mapsto(g_1h,g_2h),\quad h\in \mathrm{SO}(3). \tag{7.1}$$

In what follows, we are to ask how the system of jointed cylinders is transformed under the right action of SO(3). We first observe that under the right action, the vector, $\mathbf{\Theta}$, of one-forms defined by (4.3) receives the transformation

$$(gh)^{-1}d(gh)=h^{-1}\hat{\mathbf{\Theta}}h=\widehat{h^{-1}\mathbf{\Theta}}. \tag{7.2}$$

In view of this and the transformation $A_0\rightarrow h^{-1}A_0h$, we conceive that the subgroup

$$G_0=\{h\in \mathrm{SO}(3);\ \ A_0=h^{-1}A_0h\} \tag{7.3}$$

will provide a symmetry group of the jointed cylinders. It turns out from (4.2) and (7.3) that

$$G_0\cong\begin{cases}\mathrm{O}(2) & \text{for } \lambda\neq 1\\ \mathrm{SO}(3) & \text{for } \lambda=1,\end{cases} \tag{7.4}$$

where O(2) is a subgroup of SO(3) expressed as

$$\left\{\begin{pmatrix} a & 0 & b\\ 0 & 1 & 0\\ c & 0 & d\end{pmatrix},\ \begin{pmatrix} a' & 0 & b'\\ 0 & -1 & 0\\ c' & 0 & d'\end{pmatrix};\quad \begin{matrix}\kappa=\begin{pmatrix} a & b\\ c & d\end{pmatrix},\ \ \nu=\begin{pmatrix} a' & b'\\ c' & d'\end{pmatrix}\in \mathrm{O}(2),\\ \det\kappa=1, \det\nu=-1\end{matrix}\right\}. \tag{7.5}$$

On the other hand, under the right action of h, the position vector of the center of mass of the body half 1 is subject to the transformation

$$\mathbf{r}_1=-\ell(g_1\mathbf{e}_2\cdot\mathbf{e}_2)\mathbf{e}_2\mapsto\mathbf{r}_1'=-\ell(g_1h\mathbf{e}_2\cdot h\mathbf{e}_2)h\mathbf{e}_2. \tag{7.6}$$

Since $h\mathbf{e}_2=\pm\mathbf{e}_2$ for $h\in G_0\cong\mathrm{O}(2)$ with $\lambda\neq 1$, one has $\mathbf{r}_1'=\pm\mathbf{r}_1$. However, for $h\in G_0\cong\mathrm{SO}(3)$ with $\lambda=1$, one has $\mathbf{r}_1'\neq\pm\mathbf{r}_1$ in general. For the body half 2, similar equations hold.

Thus, from the definitions (4.4), (4.10), (4.14), and (5.2), one has the following.

Proposition 4: Both the connection form ω and the metric ds^2 on the no-twist configuration space Q_0 are invariant under the right action of $G_0\cong\mathrm{O}(2)$. The metric ds_0^2 on the shape space $Q_0/\mathrm{O}(2)\cong\mathbf{R}P^2$ is invariant under G_0 as well, since the structure group O(2) and the symmetry group G_0 commute. In the case of $\lambda=1$ and $\kappa=0$ [see (4.17) and (5.6)], both ω and ds^2 become invariant under the symmetry group $G_0\cong\mathrm{SO}(3)$.

Now it is clear that the optimal Hamiltonian system $(T^*(Q_0),d\Theta,H)$ referred to in Proposition 5 will be also invariant under the $G_0\cong\mathrm{O}(2)$ action, where one should note that the action on Q_0 is naturally lifted to a symplectic action on the cotangent bundle $T^*(Q_0)$. To show this in an explicit manner, we describe the right action of O(2) in local coordinates. We notice first that the group $G_0\cong\mathrm{O}(2)$ given in (7.5) is put in the form

$$\mathrm{O}(2)=\mathrm{SO}(2)\cup e^{\pi\hat{\mathbf{e}}_1}\mathrm{SO}(2),\quad \mathrm{SO}(2)=\{e^{t\hat{\mathbf{e}}_2}\}. \tag{7.7}$$

Letting $e^{t\hat{\mathbf{e}}_2}$ and $e^{\pi\hat{\mathbf{e}}_1}e^{t\hat{\mathbf{e}}_2}$ act to the right on Q_0 with the local expression (3.6), one obtains

$$\begin{aligned} e^{t\hat{\mathbf{e}}_2}:&\quad (\chi,\phi,\theta)\mapsto(\chi,\phi,\theta+t),\\ e^{\pi\hat{\mathbf{e}}_1}e^{t\hat{\mathbf{e}}_2}:&\quad (\chi,\phi,\theta)\mapsto(\chi,\phi+\pi,-\theta+t),\end{aligned} \tag{7.8}$$

on account of $e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}=e^{\pi\hat{\mathbf{e}}_1}e^{-t\hat{\mathbf{e}}_2}$. The lifted symplectic action is then expressed as

$$e^{t\hat{\mathbf{e}}_2}: \quad \begin{cases} (\chi,\phi,\theta)\mapsto(\chi,\phi,\theta+t), \\ (p_\chi,p_\phi,p_\theta)\mapsto(p_\chi,p_\phi,p_\theta), \end{cases}$$
$$e^{\pi\hat{\mathbf{e}}_1}e^{t\hat{\mathbf{e}}_2}: \quad \begin{cases} (\chi,\phi,\theta)\mapsto(\chi,\phi+\pi,-\theta+t), \\ (p_\chi,p_\phi,p_\theta)\mapsto(p_\chi,p_\phi,-p_\theta). \end{cases} \tag{7.9}$$

The following Proposition is now easy to prove.

Proposition 5: The Hamiltonian system $(T^*(Q_0),d\Theta,H)$ admits a symmetry group $G_0 \cong \mathrm{O}(2)$ which acts on the phase space to the right. If $\lambda=1$ and $\kappa=0$ [see (4.17) and (5.6)], the symmetry group G_0 becomes SO(3).

Since the symmetry group and the structure group commute, the reduced system admits the same symmetry group G_0 as well.

Theorem 6: The reduced system $(P_\mu,d\Theta_\mu,H_\mu)$ admits the symmetry group O(2) as well. If $\lambda=1$ and $\kappa=0$, the symmetry group becomes SO(3).

VIII. QUANTIZATION

We have obtained in Sec. V the Hamiltonian system arising from the optimal control problem. We now wish to quantize this system, regarding the jointed cylinders as a model of a molecule.

For convenience' sake, we choose to treat the optimal Hamiltonian given by (5.18) along with $I_1=1/2$. The canonical metric ds^2_{can} is now expressed as

$$ds^2_{\mathrm{can}}=d\phi^2+\sin^2\phi\, d\theta^2+(d\chi+\cos\phi\, d\theta)^2. \tag{8.1}$$

Then the volume element is defined to be

$$d\phi\wedge\sin\phi\, d\theta\wedge(d\chi+\cos\phi\, d\theta)=\sin\phi\, d\phi\wedge d\theta\wedge d\chi, \tag{8.2}$$

and the horizontal metric tensor and horizontal vector fields become, respectively,

$$(h_{ij})=\mathrm{diag}(1,\sin^2\phi), \quad (h^{ij})=(h_{ij})^{-1}, \tag{8.3}$$

$$X_1=\frac{\partial}{\partial\phi}, \quad X_2=\frac{\partial}{\partial\theta}-\cos\phi\frac{\partial}{\partial\chi}. \tag{8.4}$$

By the use of these quantities, the horizontal (or vibrational) energy functional for wave functions f on Q_0 is expressed as

$$\frac{1}{2}\int_{Q_0}\sum_{i,j=1}^{2} h^{ij}\overline{X_i f}X_j f\sqrt{|h|}dq, \tag{8.5}$$

where the overbar denotes the complex conjugate, and

$$\sqrt{|h|}dq=\sin\phi\, d\phi\wedge d\theta\wedge d\chi, \quad |h|=\det(h_{ij})=\sin^2\phi. \tag{8.6}$$

Assuming that f is a function of compact support in the local coordinate system, we carry out the integration of (8.5) by part to obtain

$$-\frac{1}{2}\int_{Q_0}\bar{f}\frac{1}{\sqrt{|h|}}\sum_{i,j=1}^{2} X_i(h^{ij}\sqrt{|h|}X_j f)\sqrt{|h|}dq, \tag{8.7}$$

which provides the Hamiltonian operator $\hat{H}'$ in the form

$$\hat{H}' = -\frac{1}{2}\frac{1}{\sqrt{|h|}}\sum_{i,j=1}^{2} X_i(h^{ij}\sqrt{|h|}X_j) = -\frac{1}{2}\left(\frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial}{\partial\phi}\right) + \frac{1}{\sin^2\phi}\left(\frac{\partial}{\partial\theta} - \cos\phi\frac{\partial}{\partial\chi}\right)^2\right). \tag{8.8}$$

This operator is manifestly O(2) invariant, as is observed by using (3.14).

On the other hand, we denote by Δ_3 the Laplacian on $Q_0 \cong \mathrm{SO}(3)$ with respect to the canonical metric given in (8.1). By definition, Δ_3 is expressed as

$$\Delta_3 = -\frac{1}{\sqrt{|h|}}\sum_{k,\ell=1}^{3}\frac{\partial}{\partial q_k}\left(g^{k\ell}\sqrt{|h|}\frac{\partial}{\partial q_\ell}\right), \quad (q_k) = (\phi,\theta,\chi), \tag{8.9}$$

where $(g^{k\ell}) = (g_{k\ell})^{-1}$ with $ds^2_{\mathrm{can}} = \Sigma^3_{k,\ell=1} g_{k\ell}\, dq_k\, dq_\ell$, and we have used the fact that $\det(g_{k\ell}) = \det(h_{ij})$. The Δ_3 can also be obtained from the total energy functional which is broken up into the sum of the horizontal (or vibrational) and the vertical (or rotational) energy functionals;

$$\frac{1}{2}\int_{Q_0}\sum_{k,\ell=1}^{3} g^{k\ell}\overline{\frac{\partial f}{\partial q_k}}\frac{\partial f}{\partial q_\ell}\sqrt{|h|}dq = \frac{1}{2}\int_{Q_0}\left(\sum_{i,j=1}^{2} h^{ij}\overline{X_i f}X_j f + \overline{\frac{\partial f}{\partial\chi}}\frac{\partial f}{\partial\chi}\right)\sqrt{|h|}dq. \tag{8.10}$$

The integration of (8.10) by part yields

$$\frac{1}{2}\Delta_3 = \hat{H}' - \frac{1}{2}\frac{\partial^2}{\partial\chi^2}. \tag{8.11}$$

This means that $\hat{H}'$ is the horizontal part of the total energy operator $\frac{1}{2}\Delta_3$. This fact is the same as that H' is the horizontal part of the total energy T', where T' is T with $\kappa = 0$, $\lambda = 1$, and $I_1 = 1/2$. Thus, the operator $\hat{H}'$ should be regarded as a quantization of the classical Hamiltonian H' given by (5.18) with $I_1 = 1/2$. In other words, the $\hat{H}'$ is a quantization of the classical Hamiltonian satisfying the nonholonomic constraint of the vanishing total angular momentum.

Using the relation (8.11), we can find the eigenvalues of $\hat{H}'$. As for Δ_3, its eigenvalues and the associated eigenfunctions are known,[13] respectively, to be given by

$$J(J+1), \quad J = 0,1,2,\ldots, \tag{8.12}$$

$$Y_{JKM}(\phi,\theta,\chi) = \Phi_{JKM}(\phi)e^{iK\theta}e^{iM\chi}, \quad |K| \leqslant J, |M| \leqslant J, \tag{8.13}$$

where $\Phi_{JKM}(\phi)$ is defined, by the use of the Jacobi polynomial $P^{(\alpha,\beta)}_{\gamma}(x)$, to be

$$\Phi_{JKM}(\phi) = \left(\sin\frac{\phi}{2}\right)^{\alpha}\left(\cos\frac{\phi}{2}\right)^{\beta} P^{(\alpha,\beta)}_{\gamma}(\cos\phi),$$
$$\alpha = |K-M|, \quad \beta = |K+M|, \quad \gamma = J - \tfrac{1}{2}|K-M| - \tfrac{1}{2}|K+M|. \tag{8.14}$$

Applying $\hat{H}'$ to Y_{JKM} results in

$$\hat{H}' Y_{JKM} = \tfrac{1}{2}(J(J+1) - M^2)Y_{JKM}, \quad J = 0,1,2,\ldots, \quad |M| \leqslant J, \tag{8.15}$$

which provides the eigenvalues of $\hat{H}'$.

Theorem 7: The classical Hamiltonian system having the Hamiltonian function H' given by (5.18) with $I_1 = 1/2$ is quantized to be a quantum system having the Hamiltonian operator $\hat{H}'$ given in (8.8), whose eigenvalues and associated eigenfunctions are given in (8.15) and (8.13), respectively.

It is possible to define the Hamiltonian operator corresponding to the optimal Hamiltonian (5.15). To do this, we put the ds^2, given in (5.5), in the form

$$ds^2 = 2I_1 \sum_{\alpha,\beta=1}^{3} g_{\alpha\beta}\nu_\alpha\nu_\beta, \tag{8.16}$$

where ν_α are one-forms defined to be

$$\begin{aligned} \nu_1 &= d\phi, \\ \nu_2 &= d\theta, \\ \nu_3 &= d\chi + \frac{\lambda \cos\phi}{\sin^2\phi + \lambda\cos^2\phi} d\theta. \end{aligned} \tag{8.17}$$

Let $(g^{\alpha\beta}) = (g_{\alpha\beta})^{-1}$, and let X_α be the dual vector fields to (8.17), $\nu_\alpha(X_\beta) = \delta_{\alpha\beta}$, where X_1 and X_2 are equal to those given in (4.19) and $X_3 = \partial/\partial\chi$. The volume element is then defined, from (8.16), to be

$$\sqrt{|g|}dq = \sqrt{\lambda}\sqrt{1+\kappa\sin^2\phi}\sin\phi d\phi\wedge d\theta\wedge d\chi. \tag{8.18}$$

With this set up, the kinetic energy functional is put in the form

$$\begin{gathered} \frac{1}{4I_1}\int_{Q_0}\sum_{\alpha\beta=1}^{3} g^{\alpha\beta}\overline{X_\alpha f}X_\beta f\sqrt{|g|}dq = \frac{1}{4I_1}\int_{Q_0}\left[\sum_{i,j=1}^{2} h^{ij}\overline{X_i f}X_j f + g^{33}\overline{X_3 f}X_3 f\right]\sqrt{|g|}dq, \\ (h^{ij}) = (g^{ij}), \quad i,j=1,2. \end{gathered} \tag{8.19}$$

Thus one defines the Hamiltonian operator $\hat{H}$ corresponding to the optimal Hamiltonian (5.15) to be

$$\begin{aligned} \hat{H} &= -\frac{1}{4I_1}\frac{1}{\sqrt{|g|}}\sum_{i,j=1}^{2} X_i(h^{ij}\sqrt{|g|}X_j) \\ &= -\frac{1}{4I_1}\left[\frac{1}{\sqrt{1+\kappa\sin^2\phi}\sin\phi}\frac{\partial}{\partial\phi}\left(\frac{\sin\phi}{\sqrt{1+\kappa\sin^2\phi}}\frac{\partial}{\partial\phi}\right)\right. \\ &\quad \left. + \frac{\sin^2\phi+\lambda\cos^2\phi}{\lambda\sin^2\phi}\left(\frac{\partial}{\partial\theta} - \frac{\lambda\cos\phi}{\sin^2\phi+\lambda\cos^2\phi}\frac{\partial}{\partial\chi}\right)^2\right], \end{aligned} \tag{8.20}$$

which is the horizontal (or vibrational) part of the full energy operator $\hat{T}$,

$$\begin{aligned} \hat{T} &= -\frac{1}{4I_1}\frac{1}{\sqrt{|g|}}\sum_{\alpha,\beta=1}^{3} X_\alpha(g^{\alpha\beta}\sqrt{|g|}X_\beta) \\ &= -\frac{1}{4I_1}\left[\frac{1}{\sqrt{1+\kappa\sin^2\phi}\sin\phi}\frac{\partial}{\partial\phi}\left(\frac{\sin\phi}{\sqrt{1+\kappa\sin^2\phi}}\frac{\partial}{\partial\phi}\right)\right. \\ &\quad \left. + \left(\frac{\cos^2\phi}{\sin^2\phi} + \frac{1}{\lambda}\right)\frac{\partial^2}{\partial\theta^2} - \frac{2\cos\phi}{\sin^2\phi}\frac{\partial^2}{\partial\chi\partial\theta} + \frac{1}{\sin^2\phi}\frac{\partial^2}{\partial\chi^2}\right] \\ &= \hat{H} - \frac{1}{4I_1(\sin^2\phi+\lambda\cos^2\phi)}\frac{\partial^2}{\partial\chi^2}. \end{aligned} \tag{8.21}$$

If $\kappa=0$, the $\hat{T}$ becomes the operator known as the Hamiltonian operator for a symmetrical top,[13] the eigenfunctions and eigenvalues of which are known[13] to be given by Y_{JKM} and

$$E=\frac{1}{4I_1}\left(J(J+1)+K^2\left(\frac{1}{\lambda}-1\right)\right), \tag{8.22}$$

respectively. If $\kappa\neq 0$, the eigenvalue problem becomes quite difficult to solve.

IX. REDUCTION AND SYMMETRY OF THE QUANTUM SYSTEM

We wish to reduce the quantum system on $Q_0\cong \mathrm{SO}(3)$ with the Hamiltonian $\hat{H}'$ to a quantum system on $Q_0/\mathrm{O}(2)\cong \mathbf{R}P^2$. To do this, we have to set up vector bundles associated with the principal O(2) bundle, $Q_0\to Q_0/\mathrm{O}(2)$. Let ρ be a unitary irreducible representation of the structure group O(2) in a complex vector space V. For example, in the vector space $\mathbf{C}^2$, one has a unitary irreducible representation ρ_M such that

$$\rho_M(e^{t\hat{\mathbf{e}}_2})\mathbf{u}_1=e^{iMt}\mathbf{u}_1,\quad \rho_M(e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1})\mathbf{u}_1=e^{-iMt}\mathbf{u}_2,$$
$$\rho_M(e^{t\hat{\mathbf{e}}_2})\mathbf{u}_2=e^{-iMt}\mathbf{u}_2,\quad \rho_M(e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1})\mathbf{u}_2=e^{iMt}\mathbf{u}_1, \tag{9.1}$$

where $\mathbf{u}_i$, $i=1,2$, are the standard basis of $\mathbf{C}^2$ and M is an integer. It is an easy matter to observe that ρ_M is indeed a homomorphism. For example, one has

$$\rho_M(e^{t_1\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}e^{t_2\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1})=\rho_M(e^{t_1\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1})\rho_M(e^{t_2\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}). \tag{9.2}$$

We are to give the definition of the vector bundle associated with a representation ρ. First we define a left action of O(2) on $Q_0\times V$ by

$$(q,v)\mapsto(gq,\rho(g)v),\quad (q,v)\in Q_0\times V,\quad g\in \mathrm{O}(2). \tag{9.3}$$

This action gives an equivalence relation in $Q_0\times V$, and then yields the quotient manifold, denoted by $V_\rho := Q_0\times_\rho V$, which is made into a complex vector bundle with base space $Q_0/\mathrm{O}(2)$ and fiber V.

A V-valued function F on Q_0 is called equivariant with respect to ρ (or ρ-equivariant), if it satisfies

$$F(gq)=\rho(g)F(q),\quad g\in \mathrm{O}(2),\ q\in Q_0. \tag{9.4}$$

A map $s:B:=Q_0/\mathrm{O}(2)\to V_\rho$ is called a cross section in V_ρ, if $\pi\circ s=\mathrm{id}_B$, where id_B is the identity map of B. The ρ-equivariant functions are in one-to-one correspondence with the cross sections in V_ρ. In fact, one has the correspondence $s(\pi(q))=[(q,F(q))]$, where $[\cdot]$ denotes an equivalence class in $Q_0\times V$. We denote by $q_\rho^\#$ the correspondence; $q_\rho^\# s=F$.

We proceed to the linear connection associated with the connection ω defined on the principal O(2) bundle $Q_0\to B$. Let ξ be a vector field on the base space B and ξ^* its horizontal lift; $\omega(\xi^*)=0$, $\pi_*\xi_q^*=\xi_{\pi(q)}$, $q\in Q_0$. Then, for a cross section s in V_ρ, its covariant derivative with respect to ξ is defined to be

$$\nabla_\xi s=q_\rho^{\#^{-1}}\xi^*(q_\rho^\# s). \tag{9.5}$$

In the same manner, the reduced Hamiltonian operator $\hat{H}'_\rho$ is defined through

$$\hat{H}'_\rho s=q_\rho^{\#^{-1}}\hat{H}'(q_\rho^\# s). \tag{9.6}$$

This is because $\hat{H}'$ is O(2) invariant, so that $\hat{H}'(q_\rho^\# s)$ is ρ-equivariant. It then follows that if F is a ρ-equivariant eigenfunction of $\hat{H}'$ with an eigenvalue E then $s=q_\rho^{\#^{-1}}F$ becomes an eigen-cross section of $\hat{H}'_\rho$ with the same eigenvalue E.

We are going to work with the representation ρ_M given in (9.1). For the ρ_M, we choose to denote the vector bundle V_ρ and the correspondence $q_\rho^\#$ by V_M and by $q_M^\#$, respectively. Applying (9.6) to (8.8), one has the reduced Hamiltonian operator $\hat{H}'_M=q_M^{\#^{-1}}\hat{H}' q_M^\#$ expressed as

$$\hat{H}'_M=-\frac{1}{2}\frac{1}{\sqrt{|h|}}\sum_{i,j=1}^{2}\nabla_i(h^{ij}\sqrt{|h|}\nabla_j), \tag{9.7}$$

where ∇_1, ∇_2 are the covariant differential operators with respect to $\partial/\partial\phi$ and $\partial/\partial\theta$, respectively.

We now turn to the eigenfunctions Y_{JKM} for $\hat{H}'$, in order to obtain eigenvalues and eigen-cross sections of $\hat{H}'_M$. The first task for us to do is to find ρ_M-equivariant $\mathbf{C}^2$-valued functions in terms of $Y_{JKM}(q)$, $q\in Q_0$. A calculation along with (3.14), (8.13), and (8.14) provides

$$\begin{aligned} Y_{JKM}(e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}q)&=(-1)^{\beta+\gamma}e^{iMt}Y_{JK,-M}(q),\\ Y_{JK,-M}(e^{t\hat{\mathbf{e}}_2}e^{\pi\hat{\mathbf{e}}_1}q)&=(-1)^{\alpha+\gamma}e^{-iMt}Y_{JKM}(q), \end{aligned} \tag{9.8}$$

where we have used the following formula for the Jacobi polynomials;

$$P_\gamma^{(\alpha,\beta)}(-x)=(-1)^\gamma P_\gamma^{(\beta,\alpha)}(x). \tag{9.9}$$

It turns out from (9.8) that, for $M\neq 0$, the $\mathbf{C}^2$-valued function

$$F_{JKM}=Y_{JKM}\mathbf{u}_1+Y_{JK,-M}\mathbf{u}_2 \tag{9.10}$$

becomes ρ_M-equivariant if both $\beta+\gamma$ and $\alpha+\gamma$ are even numbers. Incidentally, this condition is satisfied, if all of *J, M, K* are either even or odd. Hence the F_{JKM} proves to be a ρ_M-equivariant function if all of *J, M, K* are either even or odd. We denote by $\gamma_{JKM}=q_M^{\#^{-1}}F_{JKM}$ the associated cross section in the complex vector bundle V_M. If $M=0$, one has

$$Y_{JK0}(\phi,\theta,\chi)=\frac{J!}{(J+|K|)!}P_J^{|K|}(\cos\phi)e^{iK\theta}, \tag{9.11}$$

which seem to be spherical harmonics on S^2. However, under the antipodal map $(\phi,\theta)\to(\phi+\pi,\theta+\pi)$, Y_{JK0} transforms to $(-1)^{J+K}Y_{JK0}$. In our case, since J and K are even or odd simultaneously, Y_{JK0} has the same value on a pair of antipodal points of S^2. Thus Y_{JK0} becomes a function on $\mathbf{R}P^2\cong Q_0/\mathrm{O}(2)$.

Theorem 8: The quantum system on $Q_0\cong\mathrm{SO}(3)$ with the Hamiltonian operator $\hat{H}'$ is reduced to a quantum system on $Q_0/\mathrm{O}(2)\cong\mathbf{R}P^2$, which consists of the complex vector bundle V_M $(M\neq 0)$ with base space $\mathbf{R}P^2$ and fiber $\mathbf{C}^2$, and the Hamiltonian operator $\hat{H}'_M$ acting on the space of cross sections in V_M. The local expressions of $\hat{H}'_M$ are of the form (9.7). The eigenvalues and the associated eigen-cross sections for $\hat{H}'_M$ are given by $\frac{1}{2}(J(J+1)-M^2)$, $J=|M|,|M|+2,|M|+4,\ldots$, and γ_{JKM}, $K=-J,-J+2,\ldots,J-2,J$, respectively, where γ_{JKM} are the cross sections corresponding to F_{JKM} given in (9.10). If $M=0$, the complex vector bundle V_M is replaced by a trivial complex line bundle over $\mathbf{R}P^2$, and the eigen-cross sections become the eigenfunctions on $\mathbf{R}P^2$.

Now it is an easy matter to find the symmetry of the quantum reduced system. Like the classical reduced system, the quantum reduced system is expected to admit the symmetry group O(2) which acts on eigenspaces of $\hat{H}'_M$. We wish to find how O(2) is represented in the eigenspace. In a similar manner to (9.8), a calculation along with (7.8) provides

$$Y_{JKM}(qe^{\pi\hat{\mathbf{e}}_1}e^{t\hat{\mathbf{e}}_2})=e^{iKt}Y_{J,-KM}(q),$$
$$Y_{J,-KM}(qe^{\pi\hat{\mathbf{e}}_1}e^{t\hat{\mathbf{e}}_2})=e^{-iKt}Y_{JKM}(q), \tag{9.12}$$

where we have used the assumption that all of J, K, M are either even or odd. Thus one finds that the ρ_M-equivariant functions F_{JKM} given in (9.10) are subject to the transformation

$$F_{JKM}(qe^{\pi\hat{\mathbf{e}}_1}e^{t\hat{\mathbf{e}}_2})=e^{iKt}F_{J,-KM}(q),$$
$$F_{J,-KM}(qe^{\pi\hat{\mathbf{e}}_1}e^{t\hat{\mathbf{e}}_2})=e^{-iKt}F_{JKM}(q). \tag{9.13}$$

This equation implies that the symmetry group O(2) is unitarily irreducibly represented in the two-dimensional subspace spanned by eigen-cross sections γ_{JKM} and $\gamma_{J,-KM}$ in the energy eigenspace, if $K\neq 0$. If $K=0$, the symmetry group O(2) is trivially represented in the one-dimensional subspace spanned by γ_{J0M}.

Theorem 9: If $M\neq 0$, the reduced quantum system $(V_M,\hat{H}'_M)$ admits the symmetry group O(2), which is unitarily represented in the energy eigenspace associated with the eigenvalue $\frac{1}{2}(J(J+1)-M^2)$. If both J and M are odd, the eigenspace is completely reducible to the direct sum of two-dimensional irreducible subspaces spanned by γ_{JKM} and $\gamma_{J,-KM}$ with J, M fixed and $K=-J,-J+2,...,J$. If both J and M are even, the one-dimensional subspace spanned by γ_{J0M} occurs in addition to the two-dimensional subspace spanned by γ_{JKM} and $\gamma_{J,-KM}$. If $M=0$, the symmetry group becomes SO(3), which acts on the energy eigenspace spanned by γ_{JK0} with J a non-negative even number and $K=-J,-J+2,...,J$.

In conclusion, we have to mention that the quantum system on $Q_0\cong \mathrm{SO}(3)$ with the Hamiltonian operator $\frac{1}{2}\Delta_3$ [see (8.9)] is reduced to a quantum system on $\mathbf{R}P^2$, which consists of the vector bundle $V_M(M\neq 0)$ and the Hamiltonian operator $\hat{T}_M:=\hat{H}'_M+M^2/2$. If $M=0$, one has $\hat{T}_0=\hat{H}_0$, and the eigenfunction are the same for both of $\hat{T}_0$ and $\hat{H}_0$. This implies that if $M=0$, the free and the controlled quantum systems have the same states.

[1] A. Guichardet, Ann. Inst. Henri Poincaré Phys. Theor. **40**, 329 (1984).
[2] A. Tachibana and T. Iwai, Phys. Rev. A **33**, 2262 (1986).
[3] T. Iwai, J. Math. Phys. **28**, 964 (1987).
[4] T. Iwai, J. Math. Phys. **28**, 1315 (1987).
[5] T. Iwai, J. Math. Phys. **29**, 1325 (1988).
[6] T. Iwai, Ann. Inst. Henri Poincaré Phys. Theor. **47**, 199 (1987).
[7] R. G. Littlejohn and M. Reinsch, Rev. Mod. Phys. **69**, 213 (1997), and the references therein.
[8] T. Iwai and A. Tachibana, Ann. Inst. H. Poincaré (to be published).
[9] T. Iwai, J. Phys. A **31**, 3849 (1998).
[10] M. Reinsch, J. Math. Phys. **39**, 1223 (1998).
[11] T. R. Kane and M. P. Scher, Int. J. Solids Struct. **5**, 663 (1969).
[12] R. Montgomery, *Fields Institute Communications* (Amer. Math. Soc., Providence, 1993), Vol. 1, p. 193.
[13] S. Flügge, *Practical Quantum Mechanics* (Springer, Berlin, 1971), Vol. I.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Algebraic structure of discrete zero curvature equations and master symmetries of discrete evolution equations

Wen-Xiu Ma[a)]
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, Peoples Republic of China

Benno Fuchssteiner[b)]
Department of Mathematics, University of Paderborn, D-33098 Paderborn, Germany

(Received 7 January 1998; accepted for publication 20 January 1999)

An algebraic structure related to discrete zero curvature equations is established. It is used to give an approach for generating master symmetries of the first degree for systems of discrete evolution equations and an answer to why there exist such master symmetries. The key of the theory is to generate nonisospectral flows ($\lambda_t = \lambda^l, l \geqslant 0$) from the discrete spectral problem associated with a given system of discrete evolution equations. Three examples are given. © *1999 American Institute of Physics.* [S0022-2488(99)01305-5]

I. INTRODUCTION

The theory of integrable systems has various aspects, although the term ''integrable'' is somewhat ambiguous, especially for systems of partial differential equations. Symmetries are one of those important aspects and have a deep mathematical and physical background. When any special character, for example the Lax pair, has not been found for a given system of continuous or discrete equations, among the most efficient ways is to consider its symmetries in order to obtain exact solutions. It is through symmetries that Russian scientists *et al.* developed some theories for testing the integrability of systems of evolution equations, both continuous and discrete, and classified many types of systems of nonlinear equations that possess higher differential or differential-difference degree symmetries (for example, see Refs. 1 and 2). Usually an integrable system of equations is referred to as a system possessing infinitely many symmetries.[3,4] Moreover, these symmetries form nice and interesting algebraic structures.[3,4]

For a given system of evolution equations $u_t = K(u)$, both continuous and discrete, a vector field $\sigma(u)$ is called its symmetry if $\sigma(u)$ satisfies its linearized system,

$$\frac{d\sigma(u)}{dt} = K'[\sigma], \quad \text{i.e.,} \quad \frac{\partial \sigma}{\partial t} = [K,\sigma] := K'[\sigma] - \sigma'[K], \tag{1}$$

where the prime means the Gateaux derivative. Starting from a Lie-point symmetry, we can often construct the corresponding explicit group-invariant solutions. A symmetry σ may, of course, depend explicitly on the evolution variable *t*. If a symmetry σ of the system $u_t = K(u)$ not depending explicitly on *t* is a polynomial in *t*, i.e.,

$$\sigma(t,u) = \sum_{i=0}^{n} \frac{t^i}{i!} \rho_i(u), \quad n \geqslant 1, \tag{2}$$

then we have

[a)]Electronic mail: mawx@cityu.edu.hk
[b)]Electronic mail: benno@uni-paderborn.de

$$\rho_i=[K,\rho_{i-1}],\quad 1\leqslant i\leqslant n, \tag{3}$$

and

$$(\mathrm{ad}_K)^{n+1}\rho_0=0,\quad \text{where } (\mathrm{ad}_K)\rho_0=[K,\rho_0]. \tag{4}$$

Therefore the symmetry (2) is totally determined by a vector field ρ_0 satisfying (4). This kind of vector field ρ_0 has been discussed in considerable detail and is called a master symmetry of degree n of $u_t=K(u)$ by one of the authors (BF) in Ref. 5.

The appearance of first degree master symmetries gives a common character for integrable systems of continuous evolution equations, both in $1+1$ dimensions and in $2+1$ dimensions, for example, the KdV equation and the KP equation. The resulting symmetries are sometimes called τ-symmetries (for more information, see Ref. 6, for example) and usually constitute centerless Virasoro algebras together with time-independent symmetries.[7–9] Moreover these τ-symmetries may be generated by use of zero curvature equations or Lax equations,[10] and the corresponding master symmetry flows may also be solved by the inverse scattering method.[11,12] In the case of systems of discrete evolution equations, there exist some similar results. For example, many systems of discrete evolution equations have τ-symmetries and centerless Virasoro symmetry algebras,[13–15] and the inverse scattering method may still be applied in solving themselves and their master symmetry flows.[16–19] So far, however, to the best of our knowledge, there has not been a systematic mathematical theory to explain why there exist τ-symmetries for systems of discrete evolution equations and how we can construct those τ-symmetries when they exist, from the point of discrete zero curvature equations.

Throughout this paper, ''master symmetries'' is used to express the first degree master symmetries that generate τ-symmetries. Our purpose is to give an algebraic explanation of the first question above and to provide a procedure to generate those master symmetries for a given lattice hierarchy. The discrete zero curvature equation is our basic tool to give rise to our answer and procedure. The Volterra lattice hierarchy, the Toda lattice hierarchy, and a sub-KP lattice hierarchy are chosen and analyzed as some illustrative examples, which have one dependent variable, two dependent variables, and three dependent variables, respectively.

Let us now describe our notation. Assume that $u=(u_1,\ldots,u_q)^T$, where $u_i=u_i(t,n)$, $1\leqslant i\leqslant q$, are real functions defined over $\mathbb{R}\times Z$ (in the case of the complex function, the discussion is similar), and let $\mathcal{B}$ denote all real functions $P[u]=P(t,n,u)$, which are C^∞ differentiable with respect to t and n, and C^∞-Gateaux differentiable with respect to u. We always write E as a shift operator and

$$(E^m x)(n)=x^{(m)}(n)=x(m+n),\quad \text{where } x:Z\to\mathbb{R},\quad m,n\in Z. \tag{5}$$

Note that $x^{(m)}$ here does not mean the mth derivative. Set $\mathcal{B}^r=\{(P_1,\ldots,P_r)^T|P_i\in\mathcal{B},1\leqslant i\leqslant r\}$, and denote by $\mathcal{V}^r$ all matrix operators $\Phi=(\Phi_{ij})_{r\times r}$, where the entries $\Phi_{ij}=\Phi_{ij}(t,n,u)\in\mathcal{B}$, and by $\tilde{\mathcal{V}}^r$, all matrix operators depending on a parameter λ: $U=(U_{ij})_{r\times r}$, where the entries $U_{ij}=U_{ij}(t,n,u,\lambda)\in\mathcal{B}$ for all λ, being C^∞ differentiable with respect to λ.

We will need a multiplication operator,

$$[n]:\mathcal{B}\to\mathcal{B},\quad P[u]\mapsto[n]P[u],\quad ([n]P[u])(m)=m(P[u])(m), \tag{6}$$

which is often involved in the construction of master symmetries. This avoids an unclear expression $nP[u]$, which may also mean $(nP[u])(m)=n(P[u])(m)$. For example, it is absolutely clear that $([n]P[u])(m)=mu(m-1)+mu(m)$, when $P[u]=E^{-1}u+u$. We also need a difference operator $\Delta=E-E^{-1}$, whose inverse operator may be defined by

$$(\Delta^{-1}u)(n)=((E-E^{-1})^{-1}u)(n):=\frac{1}{2}\left(\sum_{k=-\infty}^{-1}u(n+1+2k)-\sum_{k=1}^{\infty}u(n-1+2k)\right), \tag{7}$$

where u is required to be rapidly vanishing at the infinity. Moreover, we define

$$(\Delta^{-1}\alpha)=(1/2)\alpha[n], \quad \text{i.e.,} \quad (\Delta^{-1}\alpha)(n)=(1/2)\alpha n, \quad \alpha=\text{const.} \tag{8}$$

Obviously, we can find that

$$(E-1)^{-1}=\Delta^{-1}(1+E^{-1}), \quad (1-E^{-1})^{-1}=\Delta^{-1}(E+1), \tag{9}$$

and thus

$$(E-1)^{-1}\alpha=\alpha[n], \quad (1-E^{-1})^{-1}\alpha=\alpha[n], \quad \alpha=\text{const.}, \tag{10}$$

which may also be viewed as a definition of two inverse operators $(E-1)^{-1}$ and $(1-E^{-1})^{-1}$. Note that here we have used the operator $[n]$ so that two functions $(E-1)^{-1}\alpha$ and $(1-E^{-1})^{-1}\alpha$ have the other clear expressions. The operators Δ^{-1}, $(E-1)^{-1}$, and $(1-E^{-1})^{-1}$ often appears in the expressions of master symmetries, and thus master symmetries are usually nonlocal vector fields belonging to $\mathcal{B}^q$.

In order to carefully analyze algebraic structures related to symmetries, we specify the definition of the Gateaux derivative $X'[S]$ of any vector-valued function $X\in\mathcal{B}^r$ at a direction $S\in\mathcal{B}^q$ as follows:

$$X'[S]=\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} X(u+\epsilon S), \tag{11}$$

which implies that X' is an operator from $\mathcal{B}^q$ to $\mathcal{B}^r$, and need the following two product operations:

$$[K,S]=K'[S]-S'[K], \quad K,S\in\mathcal{B}^q, \tag{12}$$

$$[\![f,g]\!](\lambda)=f'(\lambda)g(\lambda)-f(\lambda)g'(\lambda), \quad f,g\in C^\infty(\mathbb{R}), \tag{13}$$

where $C^\infty(\mathbb{R})$ denotes the space of smooth functions defined over $\mathbb{R}$. It is known that $(\mathcal{B}^q,[\cdot,\cdot])$ and $(C^\infty(\mathbb{R}),[\![\cdot,\cdot]\!])$ are all Lie algebras.

We now assume that $U\in\tilde{\mathcal{V}}^r$ and the Gateaux derivative operator U' is injective throughout the paper. Let us consider the discrete spectral problem,

$$E\phi=U\phi=U(n,u,\lambda)\phi, \quad \phi_t=V\phi=V(n,u,\lambda)\phi, \tag{14}$$

where $V\in\tilde{\mathcal{V}}^r$. Its adjoint system reads as

$$E^{-1}\psi=U\psi=U(n,u,\lambda)\psi, \quad \psi_t=(EV)\psi=(EV(n,u,\lambda))\psi.$$

Their integrability conditions are given by the following discrete zero curvature equation:

$$U_t=(EV)U-UV. \tag{15}$$

If the operator equation (15) is equivalent to a system of discrete evolution equations $u_t=K(n,u)$, $K\in\mathcal{B}^q$, then it is called a discrete zero curvature representation of $u_t=K(n,u)$. Evidently,

$$U_t=U'[u_t]+f(\lambda)U_\lambda, \quad \text{if } \lambda_t=f(\lambda),$$

where $U_\lambda=\partial U/\partial\lambda$. Therefore a system of discrete evolution equations $u_t=K(n,u)$, $K\in\mathcal{B}^q$, is the integrability condition of (14) with the evolution law $\lambda_t=f(\lambda)$ if and only if

$$U'[K]+fU_\lambda=(EV)U-UV. \tag{16}$$

Note that the injective property of U' is indispensable in deriving zero curvature representations of systems of evolution equations. The equation (16) exposes an essential relation between a system of discrete evolution equations and its discrete zero curvature representation. It will play an important role in the context of our construction of master symmetries.

The paper is divided into five sections. The next section will be devoted to a general algebraic structure related to discrete zero curvature equations. Then in the third section we will establish an approach for constructing master symmetries by the use of discrete zero curvature representations, along with an explanation of why there exist master symmetries for systems of discrete evolution equations. In the fourth section, we will go on to illustrate our approach by three concrete examples of lattice hierarchies. Finally, the fifth section provides a conclusion and some remarks.

II. BASIC ALGEBRAIC STRUCTURE

We aim to discuss Lie algebraic structures of symmetries, including master symmetries, by using zero curvature equations. It is natural to ask what algebraic structure exists, related to zero curvature equations. To answer this question, we first plan to expose a Lie algebraic structure for the space $(\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R}))$.

Let $(K,V,f),(S,W,g)\in(\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R}))$, in other words, K,S are vector fields, V,W are $r\times r$ matrix operators, and f,g are smooth functions. We introduce their product:

$$[\![(K,V,f),(S,W,g)]\!]=([K,S],[\![V,W]\!],[\![f,g]\!]), \tag{17}$$

where $[K,S]$, $[\![f,g]\!]$ are defined by (12), (13), respectively, and $[\![V,W]\!]$ is defined by

$$[\![V,W]\!]=V'[S]-W'[K]+[V,W]+gV_\lambda-fW_\lambda, \tag{18}$$

where $[V,W]=VW-WV$. The same product as (18) has been introduced for the continuous case in Ref. 20.

Theorem 1: (Lie algebra) *The space* $((\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R})),[\![\cdot,\cdot]\!])$ *is a Lie algebra, the product* $[\![\cdot,\cdot]\!]$ *being defined by (17), i.e.,*

$$[\![(K,V,f),(S,W,g)]\!]=([K,S],[\![V,W]\!],[\![f,g]\!]),$$

where

$$[K,S]=K'[S]-S'[K],$$

$$[\![V,W]\!]=V'[S]-W'[K]+[V,W]+gV_\lambda-fW_\lambda,$$

$$[\![f,g]\!](\lambda)=f'(\lambda)g(\lambda)-f(\lambda)g'(\lambda).$$

The proof of the theorem will be given in Appendix A. Upon looking at the product a little bit more carefully, we can find that the Lie algebra $((\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R})),[\![\cdot,\cdot]\!])$ has a Lie subalgebra $((\mathcal{B}^q,\tilde{\mathcal{V}}^r,0),[\![\cdot,\cdot]\!])$, for which everything corresponds to the isospectral case. Moreover, the center of an element of this Lie subalgebra is often Abelian.

The above theorem exposes that a Lie algebraic structure hidden in the back of vector fields, Lax operators, and spectral evolution laws. Usually we just touch Lie algebraic structures of vector fields while discussing symmetries. If we analyze symmetries from the point of zero curvature equations, it is natural that we need to find and handle the Lie algebraic structure for all triples $(K,V,f)\in(\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R}))$, where K, V, and f are related to each other by zero curvature equations. In other words, we need to observe how two triples (K,V,f), (S,W,g) that appear in zero curvature equations connect with each other. The following theorem tells us that such a kind of connection can be reflected by the Lie algebraic operation of $(\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R}))$ in Theorem 1. Its proof can be found in Appendix B.

Theorem 2: (Algebraic structure of representations) *Let* $V,W\in\tilde{\mathcal{V}}^r$, $K,S\in\mathcal{B}^q$, *and* $f,g\in C^\infty(\mathbb{R})$. *If two equalities,*

$$(EV)U-UV=U'[K]+fU_\lambda, \tag{19}$$

$$(EW)U-UW=U'[S]+gU_\lambda, \tag{20}$$

hold, then we have a third equality,

$$(E[\![V,W]\!])U-U[\![V,W]\!]=U'[T]+[\![f,g]\!]U_\lambda,\quad T=[K,S], \tag{21}$$

where $[\![V,W]\!]$,[K,S] *and* $[\![f,g]\!]$ *are defined by (18), (12), and (13), respectively.*

According to this theorem, we can easily find that if a system $u_t=K(n,u)$ is isospectral, i.e., $\lambda_t=f=0$, then the product system $u_t=[K,S]$ for any $S\in\mathcal{B}^q$ can be viewed to be still isospectral because we have $[\![f,g]\!]=[\![0,g]\!]=0$, where g is the evolution law corresponding $u_t=S(n,u)$. Actually, the above theorem gives a discrete zero curvature representation for a product system $u_t=[K,S]$, which possesses the same order matrix operators as ones for the original systems $u_t=K(n,u)$ and $u_t=S(n,u)$ (see Refs. 20 and 21 for the continuous case). Combining two theorems above can show the following.

Corollary 1: The space defined by

$$\{(K,V,f)\in(\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R}))\,|\,U'[K]+fU_\lambda=(EV)U-UV\},$$

is a Lie subalgebra of $(\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R}))$ *under the Lie product (17).*

This corollary tells us a Lie algebraic structure about zero curvature equations, which will help us to establish Lie algebraic structures of symmetries, including master symmetries.

However, for zero curvature representations, some interesting problems remain to be solved. For example, assuming that two initial systems $u_t=K(n,u)$ and $u_t=S(n,u)$ have zero curvature representations possessing different-order matrix operators, we want to know whether there exist any zero curvature representations for the product system $u_t=[K,S]$ and what structures the resulting zero curvature representations possess if the answer is yes. It is likely to be helpful in solving this problem to use the Kronecker product, as in Ref. 22.

III. LAX OPERATORS AND MASTER SYMMETRIES

Assume that we already have a hierarchy of isospectral integrable systems of discrete evolution equations of the form

$$u_t=K_k=\Phi^kK_0,\quad \Phi\in\mathcal{V}^q,\quad K_0\in\mathcal{B}^q,\quad k\geqslant 0, \tag{22}$$

or of the form

$$u_t=K_k=JG_k=MG_{k-1},\quad J,M\in V^q,\quad G_{k-1}\in\mathcal{B}^q,\quad k\geqslant 0, \tag{23}$$

associated with a discrete spectral problem,

$$E\phi=U\phi,\quad \phi=(\phi_1,\ldots,\phi_r)^T. \tag{24}$$

The second form (23) occurs more often than the first form (22), although it is simpler to deal with the first form (22). Generally speaking, the operator Φ above is a hereditary symmetry operator (see Ref. 23 for a definition) determined by the spectral problem (24) and J,M constitute a bi-Hamiltonian pair.[24,25] If we choose $\Phi=MJ^{-1}$ when J is invertible, then the form (23) may be changed into the form (22). Usually Φ involves nonlocal operators, for example, Δ^{-1}, but J,M often involves only local operators. Our examples are all local Hamiltonian systems.

A. Structures of Lax operators

For a given $X\in\mathcal{B}^q$ or $G\in\mathcal{B}^q$, let us introduce an operator equation of $\Omega\in\tilde{\mathcal{V}}^r$:

$$(E\Omega(X))U-U\Omega(X)=U'[\Phi X]-\lambda U'[X], \tag{25}$$

in the case of (22), or an operator equation of $\Omega_J\in\tilde{\mathcal{V}}^r$:

$$(E\Omega_J(G))U-U\Omega_J(G)=U'[MG]-\lambda U'[JG], \tag{26}$$

in the case of (23). We call them the characteristic operator equations of U. The introduction of the operator equation (25) [or (26)] is an important step in our manipulation. Obviously, we can choose $\Omega_J(G)=\Omega(JG)$ when $\Phi=MJ^{-1}$. We demand that (25) [or (26)] has solutions, and $\Omega=\Omega(X)$ [or $\Omega_J(G)$] is a particular solution at X (or at G). Usually (25) [or (26)] has infinitely many solutions once one solution exists, because we can construct others $\Omega(X)+fV$ for any $f\in C^\infty(\mathbb{R})$ when $V\in\mathcal{V}^r\otimes C[\lambda,\lambda^{-1}]$ solves the stationary discrete zero curvature equation $(EV)U-UV=0$. The existence of solutions of $(EV)U-UV=0$ may result from the existence of an isospectral hierarchy associated with $E\phi=U\phi$.

Theorem 3: (structure of Lax operators) *Let two matrices $V_0,W_0\in\tilde{\mathcal{V}}^r$ and two vector fields $K_0,\rho_0\in\mathcal{B}^q$ (or $\rho_0=J\gamma_0$, $\gamma_0\in\mathcal{B}^q$) satisfy*

$$(EV_0)U-UV_0=U'[K_0], \tag{27}$$

$$(EW_0)U-UW_0=U'[\rho_0]+\lambda U_\lambda. \tag{28}$$

If we define ρ_l, $l\geqslant 1$, V_k, $k\geqslant 1$, and W_l, $l\geqslant 1$, as follows:

$$\rho_l=\Phi^l\rho_0,\quad l\geqslant 1\ \ [or\ \ \rho_l=J\gamma_l=M\gamma_{l-1},\ \ \gamma_l\in\mathcal{B}^q,\quad l\geqslant 1], \tag{29}$$

$$V_k=\lambda^kV_0+\sum_{i=1}^{k}\lambda^{k-i}\Omega(K_{i-1})\ \ [or\ \ \Omega_J(G_{i-l})],\quad k\geqslant 1, \tag{30}$$

$$W_l=\lambda^lW_0+\sum_{j=1}^{l}\lambda^{l-j}\Omega(\rho_{j-1})\ \ [or\ \ \Omega_J(\gamma_{j-1})],\quad l\geqslant 1, \tag{31}$$

then V_k, W_l, $k,l\geqslant 0$, satisfy

$$(EV_k)U-UV_k=U'[K_k],\quad (EW_l)U-UW_l=U'[\rho_l]+\lambda^{l+1}U_\lambda,\quad k,l\geqslant 0. \tag{32}$$

Therefore for any k,l⩾0, the systems of discrete evolution equations $u_t=K_k$ and $u_t=\rho_l$ possess the isospectral $(\lambda_t=0)$ and nonisospectral $(\lambda_t=\lambda^{l+1})$ *discrete zero curvature representations,*

$$U_t=(EV_k)U-UV_k,\quad U_t=(EW_l)U-UW_l,$$

respectively.

The theorem shows that the Lax operators associated with two hierarchies of interesting vector fields can be constructed simply by a unified form. Its proof is left to Appendix C. We are successful, thanks to introducing a characteristic operator equation. The difficulty is now transferred to seeking a solution to the characteristic operator equation. However, this can automatically be solved on the basis of the structure of Lax operators of isospectral hierarchies, which will be seen in the next Sec. III B.

B. A method for constructing master symmetries

Now we focus our attention on the construction problem of master symmetries. Theorem 3 already shows the structure of Lax operators associated with the isospectral and nonisospectral hierarchies (refer to Ref. 26 for the continuous case). When an isospectral hierarchy (22) [or (23)] is known, the theorem also provides us with a method to construct a nonisospectral hierarchy associated with the discrete spectral problem (24) by solving an initial discrete zero curvature equation (28) and solving a characteristic operator equation (25) [or (26)].

However, a solution to (25) [or (26)] may easily be generated by observing the resulting Lax operators. In fact, we have

$$\Omega(K_k)\ \ [\text{or}\ \ \Omega_J(G_k)]=V_{k+1}-\lambda V_k. \tag{33}$$

This may be checked, say, for the case of (22), as follows:

$$V_{k+1}-\lambda V_k=\left(\lambda^{k+1}V_0+\sum_{i=1}^{k+1}\lambda^{k-i+1}\Omega(K_{i-1})\right)-\lambda\left(\lambda^kV_0+\sum_{i=1}^{k}\lambda^{k-i}\Omega(K_{i-1})\right)=\Omega(K_k),$$

by using (30). Now by the first equality of (32), we may compute the following:

$$\begin{aligned}(E\Omega(K_k))U-U\Omega(K_k)&=(EV_{k+1}-\lambda EV_k)U-U(V_{k+1}-\lambda V_k)\\&=((EV_{k+1})U-UV_{k+1})-\lambda((EV_k)U-UV_k)\\&=U'[K_{k+1}]-\lambda U'[K_k]=U'[\Phi K_k]-\lambda U'[K_k],\end{aligned}$$

for example, for the case of (22). Therefore we see that a possible solution $\Omega(X)$ to (25) [or $\Omega_J(G)$ to (26)] may be generated by replacing the element K_k (or G_k) in the equality (33) with X (or G).

The Lax operator matrices V_{k+1} and V_k are known, when the isospectral hierarchy has already been found. Thus we do not have to directly solve the characteristic operator equations, and then the whole process of construction of the nonisospectral hierarchy becomes *an easy task*: finding ρ_0, W_0 to satisfy (28) and computing $V_{k+1}-\lambda V_k$ to find a solution to (25) [or (26)].

The nonisospectral hierarchy (29) is exactly the master symmetries that we need to find. The reasons are that the product systems between the isospectral hierarchy and the nonisospectral hierarchy are still isospectral by Theorem 2, or as we said before in Sec. II, and that usually all systems of the isospectral hierarchy commute with each other. Therefore it is because there exists a nonisospectral hierarchy that there exist master symmetries for isospectral systems of discrete evolution equations derived from a given discrete spectral problem.

In the next section, we shall in detail illustrate our construction process by three concrete examples and establish the corresponding centerless Virasoro symmetry algebras.

IV. APPLICATIONS

We illustrate only by three examples how to apply the method in the last section to construct master symmetries for various lattice hierarchies.

To make the process clearer, we introduce a conception for a given discrete spectral problem $E\phi=U\phi$, which has an injective Gateaux derivative U'. That is a uniqueness property similar to the one in the continuous case:[27] *if* $(EV)U-UV=U'[K]$, $V\in\mathcal{V}^r\otimes C[\lambda,\lambda^{-1}]$, $K\in\mathcal{B}^q$, *and* $V|_{u=0}=0$, *then* $V=0$, and further $K=0$ by the injective property of U'. It means that if an isospectral $(\lambda_t=0)$ Lax operator V equals zero at $u=0$, then so does V itself. Actually, this property corresponds to the uniqueness of an integrable hierarchy associated with a spectral problem $E\phi=U\phi$. That is to say, when initial conditions and constants of inverse difference operators are fixed [for example, as in (7) and (8)], the associated isospectral hierarchy is uniquely determined. Most of the discrete spectral problems share the uniqueness property. The following three spectral problems are exactly examples that share such a property.

A. The Volterra lattice hierarchy

Let us first consider the following discrete spectral problem:[15]

$$E\phi=U\phi,\quad U=\begin{pmatrix}1 & u\\ \lambda^{-1} & 0\end{pmatrix},\quad \phi=\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}. \tag{34}$$

The corresponding isospectral integrable lattice hierarchy reads as

$$u_t=K_k=\Phi^kK_0=u(a_{k+1}^{(1)}-a_{k+1}^{(-1)}),\quad K_0=u(u^{(-1)}-u^{(1)}),\quad k\geqslant 0. \tag{35}$$

Here the matrix

$$V=\Sigma_{i\geqslant 0}\begin{pmatrix}a_i & uc_{i+1}^{(1)}\\ c_i & a_i\end{pmatrix}\lambda^{-i}$$

solves the stationary discrete zero curvature equation $(EV)U-UV=0$, where we choose the initial conditions

$$a_0=\tfrac{1}{2},\quad c_0=0,\quad a_1=-u,\quad c_1=1,$$

and the hereditary operator Φ is given by

$$\Phi=u(1+E^{-1})(-u^{(1)}E^2+u)(E-1)^{-1}u^{-1}, \tag{36}$$

where $(E-1)^{-1}$ is determined by (9). It is worth pointing out that each system in (35) is local and polynomially dependent on u, although the hereditary operator Φ has nonlocal and nonpolynomially dependent features.

The first discrete evolution equation is the Volterra lattice equation,[28]

$$(u(n))_t=u(n)(u(n-1)-u(n+1)),$$

which is significantly generalized by Bogoyavlensky.[29] The associated Lax operators are as follows:

$$V_k=(\lambda^{k+1}V)_{\geqslant 1}+\begin{pmatrix}a_{k+1} & 0\\ c_{k+1} & a_{k+1}^{(-1)}\end{pmatrix},\quad k\geqslant 0, \tag{37}$$

where $(P)_{\geqslant 1}$ denotes the selection of the terms with degrees of λ no less than 1. In particular, the initial isospectral Lax operator reads as

$$V_0=\begin{pmatrix}\frac{1}{2}\lambda-u & \lambda u\\ 1 & -\frac{1}{2}\lambda-u^{(-1)}\end{pmatrix}. \tag{38}$$

The result until here can be obtained from (34) by using a powerful method in Ref. 30.

We easily obtain the corresponding quantities in the nonisospectral $(\lambda_t=\lambda)$ initial discrete zero curvature equation (28):

$$\rho_0=u,\quad W_0=\begin{pmatrix}\frac{1}{2} & 0\\ 0 & -\frac{1}{2}\end{pmatrix}, \tag{39}$$

and a solution to the characteristic operator equation (25) by (33):

$$\Omega(X)=\begin{pmatrix}\Omega_{11}(X) & \Omega_{12}(X)\\ \Omega_{21}(X) & \Omega_{22}(X)\end{pmatrix}, \tag{40}$$

where $\Omega_{ij}(X)$, $i,j=1,2$, are given by

$$\Omega_{11}(X)=(E-1)^{-1}(-u^{(1)}E^2+u)(E-1)^{-1}u^{-1}X,$$

$$\Omega_{12}(X)=\lambda uE(E-1)^{-1}u^{-1}X,$$

$$\Omega_{21}(X)=(E-1)^{-1}u^{-1}X,$$

$$\Omega_{22}(X)=-\lambda(E-1)^{-1}u^{-1}X+E^{-1}(E-1)^{-1}(-u^{(1)}E^2+u)(E-1)^{-1}u^{-1}X.$$

Now by Theorem 3, we obtain a hierarchy of nonisospectral discrete evolution equations $u_t=\rho_l=\Phi^l\rho_0$, $l\geqslant 0$, associated with the spectral problem (34).

Let us now consider how to compute the corresponding symmetry algebra. The idea below can be applied to other cases. We first make the following computation at $u=0$:

$$K_k|_{u=0}=0,\quad \rho_l|_{u=0}=\Phi^l\rho_0|_{u=0}=0,\quad k,l\geqslant 0,$$

$$V_k|_{u=0}=\lambda^k\begin{pmatrix}\frac{1}{2}\lambda & 0\\ 1 & -\frac{1}{2}\lambda\end{pmatrix},\quad k\geqslant 0,$$

$$W_l|_{u=0}=\lambda^l\begin{pmatrix}\frac{1}{2} & 0\\ 0 & -\frac{1}{2}\end{pmatrix}+(1-\delta_{l0})\lambda^{l-1}\begin{pmatrix}0 & 0\\ [n] & -\lambda[n]\end{pmatrix},\quad l\geqslant 0,$$

$$V_{k\lambda}|_{u=0}=k\lambda^{k-1}\begin{pmatrix}\frac{1}{2}\lambda & 0\\ 1 & -\frac{1}{2}\lambda\end{pmatrix}+\lambda^k\begin{pmatrix}\frac{1}{2} & 0\\ 0 & -\frac{1}{2}\end{pmatrix},\quad k\geqslant 0,$$

$$W_{l\lambda}|_{u=0}=l\lambda^{l-1}\begin{pmatrix}\frac{1}{2} & 0\\ 0 & -\frac{1}{2}\end{pmatrix}+(1-\delta_{l0})\begin{pmatrix}0 & 0\\ (l-1)\lambda^{l-2}[n] & -l\lambda^{l-1}[n]\end{pmatrix},\quad l\geqslant 0,$$

where V_k, W_l, $k,l\geqslant 0$, are given as in Theorem 3 and δ_{l0} represents the Kronecker symbol. While computing $W_l|_{u=0}$, we need to note that $\Omega(\rho_0)|_{u=0}\neq 0$, but $\Omega(\rho_l)|_{u=0}=0, l\geqslant 1$. The other two examples below have a similar character, too. Now we can find by the definition (18) of the product of two Lax operators that

$$[\![V_k,V_l]\!]|_{u=0}=0,\quad k,l\geqslant 0,$$

$$[\![V_k,W_l]\!]|_{u=0}=(k+1)V_{k+l}|_{u=0},\quad k,l\geqslant 0, \tag{41}$$

$$[\![W_k,W_l]\!]|_{u=0}=(k-l)W_{k+l}|_{u=0},\quad k,l\geqslant 0.$$

For example, we can compute that

$$
\begin{aligned}
[\![V_k,W_l]\!]|_{u=0} &= [V_k|_{u=0},W_l|_{u=0}]+\lambda^{l+1}V_{k\lambda}|_{u=0}\\
&= \lambda^{k+l}\left[\begin{pmatrix}\frac12\lambda & 0\\ 1 & -\frac12\lambda\end{pmatrix},\begin{pmatrix}\frac12 & 0\\ 0 & -\frac12\end{pmatrix}+(1-\delta_{l0})\lambda^{-1}\begin{pmatrix}0 & 0\\ [n]-\lambda[n] & \end{pmatrix}\right]\\
&\quad +\lambda^{l+1}\left(k\lambda^{k-1}\begin{pmatrix}\frac12\lambda & 0\\ 1 & -\frac12\lambda\end{pmatrix}+\lambda^{k}\begin{pmatrix}\frac12 & 0\\ 0 & -\frac12\end{pmatrix}\right)\\
&= \lambda^{k+l}\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}+k\lambda^{k+l}\begin{pmatrix}\frac12\lambda & 0\\ 1 & -\frac12\lambda\end{pmatrix}+\lambda^{k+l+1}\begin{pmatrix}\frac12 & 0\\ 0 & -\frac12\end{pmatrix}\\
&= (k+1)\lambda^{k+l}\begin{pmatrix}\frac12\lambda & 0\\ 1 & -\frac12\lambda\end{pmatrix}=(k+1)V_{k+l}|_{u=0}.
\end{aligned}
$$

Because $[\![V_k,V_l]\!]$, $[\![V_k,W_l]\!]-(k+1)V_{k+l}$, $[\![W_k,W_l]\!]-(k-l)W_{k+l}$, $k,l\geqslant 0$, are all isospectral ($\lambda_t=0$) Lax operators belonging to $\mathcal{V}^2\otimes C[\lambda,\lambda^{-1}]$ by Theorem 2, based upon (41) we obtain a Lax operator algebra by the uniqueness property of the spectral problem (34),

$$
\begin{aligned}
&[\![V_k,V_l]\!]=0,\quad k,l\geqslant 0,\\
&[\![V_k,W_l]\!]=(k+1)V_{k+l},\quad k,l\geqslant 0,\\
&[\![W_k,W_l]\!]=(k-1)W_{k+l},\quad k,l\geqslant 0.
\end{aligned}
\tag{42}
$$

Further, due to the injective property of U', we finally obtain a vector field algebra of the isospectral hierarchy and the nonisospectral hierarchy,

$$
\begin{aligned}
&[K_k,K_l]=0,\quad k,l\geqslant 0,\\
&[K_k,\rho_l]=(k+1)K_{k+l},\quad k,l\geqslant 0,\\
&[\rho_k,\rho_l]=(k-l)\rho_{k+l},\quad k,l\geqslant 0.
\end{aligned}
\tag{43}
$$

This implies that ρ_l, $l\geqslant 0$, are all master symmetries of each lattice equation $u_t=K_{k_0}$ in the isospectral hierarchy, and the symmetries,

$$K_k,\quad k\geqslant 0,\quad \text{and}\quad \tau_l^{(k_0)}=t[K_{k_0},\rho_l]+\rho_l,\quad l\geqslant 0,$$

constitute a symmetry algebra of Virasoro type possessing the same commutator relations as (43).

B. The Toda lattice hierarchy

Second, let us consider the discrete spectral problem:[30]

$$E\phi=U\phi,\quad U=\begin{pmatrix}0 & 1\\ -v & \lambda-p\end{pmatrix},\quad u=\begin{pmatrix}p\\ v\end{pmatrix},\quad \phi=\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}. \tag{44}$$

The corresponding isospectral integrable Toda lattice hierarchy[31] reads as

$$u_t=K_k=\Phi^kK_0=\begin{pmatrix}a_{k+2}-a_{k+2}^{(1)}\\ v(b_{k+2}^{(1)}-b_{k+2})\end{pmatrix},\quad K_0=\begin{pmatrix}v-v^{(1)}\\ v(p-p^{(-1)})\end{pmatrix},\quad k\geqslant 0. \tag{45}$$

Here

$$V=\Sigma_{i\geqslant 0}\begin{pmatrix} a_i & b_i \\ -vb_i^{(1)} & -a_i \end{pmatrix}\lambda^{-i}$$

solves $(EV)U-UV=0$, in which we choose

$$a_0=\tfrac{1}{2},\quad b_0=0,\quad a_1=0,\quad b_1=-1,$$

and the hereditary operator Φ is determined by

$$\Phi=\begin{pmatrix} p & (v^{(1)}E^2-v)(E-1)^{-1}v^{-1} \\ v(E^{-1}+1) & v(pE-p^{(-1)})(E-1)^{-1}v^{-1} \end{pmatrix}. \tag{46}$$

The first system of discrete evolution equations is the Toda lattice,[32]

$$(p(n))_t=v(n)-v(n+1),$$

$$(v(n))_t=v(n)(p(n)-p(n-1)),$$

up to a transform of dependent variables. The lattice hierarchy above has a local tri-Hamiltonian structure,

$$u_t=K_k=J\frac{\delta H_{k+2}}{\delta u}=M\frac{\delta H_{k+1}}{\delta u}=N\frac{\delta H_k}{\delta u},\quad k\geqslant 0,$$

where the Hamiltonian operators J,M,N and the conserved quantities H_k, defined by

$$J=\begin{pmatrix} 0 & (1-E)v \\ v(E^{-1}-1) & 0 \end{pmatrix},$$

$$M=J\Phi^{\dagger}=-\Phi J=\begin{pmatrix} Ev-vE^{-1} & p(E-1)v \\ v(1-E^{-1})p & v(E-E^{-1})v \end{pmatrix},$$

$$N=M\Phi^{\dagger}=-\Phi M$$
$$=\begin{pmatrix} p(vE^{-1}-Ev)+(vE^{-1}-Ev)p & p^2(1-E)v+(vE^{-1}-Ev)(1+E)v \\ v(E^{-1}+1)(vE^{-1}-Ev)+v(E^{-1}-1)p^2 & 2v(E^{-1}p-pE)v \end{pmatrix},$$

$$H_0=p+\tfrac{1}{2}\ln v,\quad H_k=-\frac{b_{k+1}}{k},\quad k\geqslant 1,$$

where $\Phi^{\dagger}$ denotes the conjugate operator of Φ. Note that this tri-Hamiltonian structure may be established through a trace identity.[30] The corresponding Lax operators read as

$$V_k=(\lambda^{k+1}V)_+ + \begin{pmatrix} b_{k+2} & 0 \\ 0 & 0 \end{pmatrix},\quad k\geqslant 0, \tag{47}$$

where the subscript $+$ denotes selecting the non-negative part. Hence, in particular,

$$V_0=\begin{pmatrix} \tfrac{1}{2}\lambda-p^{(-1)} & -1 \\ v & -\tfrac{1}{2}\lambda \end{pmatrix}. \tag{48}$$

It is easy to find the corresponding quantities in the nonisospectral ($\lambda_t=\lambda$) initial discrete zero curvature equation (28):

$$\rho_0=\begin{pmatrix} p \\ 2v \end{pmatrix},\quad W_0=\begin{pmatrix} [n]-1 & 0 \\ 0 & [n] \end{pmatrix}, \tag{49}$$

where $[n]$ is the multiplication operator defined by (6), and a solution to the characteristic operator equation (25) by (33):

$$\Omega(X)=\begin{pmatrix} \Omega_{11}(X) & \Omega_{12}(X) \\ \Omega_{21}(X) & \Omega_{22}(X) \end{pmatrix},\quad X=\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \tag{50}$$

where $\Omega_{ij}(X)$, $i,j=1,2$, are given by

$$\Omega_{11}(X)=E^{-1}(E-1)^{-1}X_1+(p^{(-1)}-\lambda)(E-1)^{-1}v^{-1}X_2,$$

$$\Omega_{12}(X)=(E-1)^{-1}v^{-1}X_2,$$

$$\Omega_{21}(X)=vE(E-1)^{-1}v^{-1}X_2,$$

$$\Omega_{22}(X)=(E-1)^{-1}X_1.$$

In this way, we obtain a hierarchy of nonisospectral systems of discrete evolution equations $\rho_l=\Phi^l\rho_0$, $l\geqslant 0$, associated with the spectral problem (44).

In order to construct a vector field algebra, we make a similar computation at $u=0$:

$$K_k|_{u=0}=0,\quad \rho_l|_{u=0}=\Phi^l\rho_0|_{u=0}=0,\quad k,l\geqslant 0,$$

$$V_k|_{u=0}=\lambda^k\begin{pmatrix} \frac{1}{2}\lambda & -1 \\ 0 & -\frac{1}{2}\lambda \end{pmatrix},\quad k\geqslant 0,$$

$$W_l|_{u=0}=\lambda^l\begin{pmatrix} [n]-1 & 0 \\ 0 & [n] \end{pmatrix}+(1-\delta_{l0})\begin{pmatrix} -2\lambda[n] & 2[n] \\ 0 & 0 \end{pmatrix},\quad l\geqslant 0,$$

$$V_{k\lambda}|_{u=0}=k\lambda^{k-1}\begin{pmatrix} \frac{1}{2}\lambda & -1 \\ 0 & -\frac{1}{2}\lambda \end{pmatrix}+\lambda^k\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix},\quad k\geqslant 0,$$

$$W_{l\lambda}|_{u=0}=l\lambda^{l-1}\begin{pmatrix} [n]-1 & 0 \\ 0 & [n] \end{pmatrix}+(1-\delta_{l0})\begin{pmatrix} -2l\lambda^{l-1}[n] & 2(l-1)\lambda^{l-2}[n] \\ 0 & 0 \end{pmatrix},\quad l\geqslant 0.$$

Now we can find through the product definition of $[\![\cdot,\cdot]\!]$ in (18) that

$$[\![V_k,V_l]\!]|_{u=0}=0,\quad k,l\geqslant 0,$$

$$[\![V_k,W_l]\!]|_{u=0}=(k+1)V_{k+1}|_{u=0},\quad k,l\geqslant 0, \tag{51}$$

$$[\![W_k,W_l]\!]|_{u=0}=(k-l)W_{k+l}|_{u=0},\quad k,l\geqslant 0.$$

A similar argument yields a Lax operator algebra by the uniqueness property of the spectral problem (44),

$$
\begin{aligned}
&[\![V_k,V_l]\!]=0, \quad k,l\geqslant 0,\\
&[\![V_k,W_l]\!]=(k+1)V_{k+1}, \quad k,l\geqslant 0,\\
&[\![W_k,W_l]\!]=(k-l)W_{k+l}, \quad k,l\geqslant 0.
\end{aligned}
\tag{52}
$$

And then because of the injective property of U', we obtain a semiproduct Lie algebra of the isospectral hierarchy and the nonisospectral hierarchy,

$$
\begin{aligned}
&[K_k,K_l]=0, \quad k,l\geqslant 0,\\
&[K_k,\rho_l]=(k+1)K_{k+1}, \quad k,l\geqslant 0,\\
&[\rho_k,\rho_l]=(k-l)\rho_{k+l}, \quad k,l\geqslant 0,
\end{aligned}
\tag{53}
$$

which gives rise to a symmetry algebra of the Virasoro type for the isospectral Toda hierarchy (45).

C. A sub-KP lattice hierarchy

Let us finally consider the discrete spectral problem:[33]

$$
E\phi=U\phi, \quad U=\begin{pmatrix} 0 & 1 & 0\\ b-\lambda & a & 1\\ c & 0 & 0\end{pmatrix}, \quad u=\begin{pmatrix} a\\ b\\ c\end{pmatrix}, \quad \phi=\begin{pmatrix}\phi_1\\ \phi_2\\ \phi_3\end{pmatrix},
\tag{54}
$$

which is equivalent to $(-E^2+b+aE+E^{-1}c)\phi_1=\lambda\phi_1$, a sub-KP discrete spectral problem.[34] The corresponding isospectral integrable lattice hierarchy reads as

$$
u_t=K_k=JG_k=MG_{k-1}, \quad k\geqslant 0,
\tag{55}
$$

where a Hamiltonian pair J, M and G_{-1}, G_0, G_1 are defined by

$$
J=\begin{pmatrix} E-E^{-1} & 0 & 0\\ 0 & 0 & (E^{-1}-1)c\\ 0 & -c(E-1) & 0\end{pmatrix},
$$

$$
M=\begin{pmatrix} Eb-bE^{-1}+a\Delta_+\Delta^{-1}\Delta_-a & EcE-E^{-1}c & -a\Delta_+\Delta^{-1}\Delta_-c\\ cE-E^{-1}cE^{-1} & E^{-1}ac-acE & -b\Delta_-c\\ c\Delta_+-\Delta^{-1}\Delta_-a & -c\Delta_+b & c[\Delta_+\Delta^{-1}\Delta_--\Delta_--\Delta_+]c\end{pmatrix},
$$

$$
G_{-1}=\begin{pmatrix}0\\ 1\\ 0\end{pmatrix}, \quad G_0=\begin{pmatrix}c\\ b\\ a\end{pmatrix}, \quad G_1=\begin{pmatrix} c(Eb+b)\\ b^2+ac+E^{-1}ac\\ a(Eb+b)-Ec-E^{-1}c\end{pmatrix},
$$

where Δ_+, Δ_- are the difference operators: $\Delta_+=E-1, \Delta_-=1-E^{-1}$. The first nonlinear system of discrete evolution equations is

$$
\begin{aligned}
&(a(n))_t=c(n+1)-c(n-1),\\
&(b(n))_t=a(n-1)c(n-1)-a(n)c(n),\\
&(c(n))_t=c(n)(b(n)-b(n+1)).
\end{aligned}
$$

We easily find the corresponding quantities in (27) and (28):

$$K_0=\begin{pmatrix}(E-E^{-1})c\\(E^{-1}-1)ac\\c(1-E)b\end{pmatrix},\quad V_0=\begin{pmatrix}0&0&1\\c&0&0\\-E^{-1}ac&E^{-1}c&\lambda-b\end{pmatrix},$$

$$\rho_0=J\gamma_0=M\gamma_{-1}=J\begin{pmatrix}\frac{1}{2}\Delta^{-1}a\\-\frac{3}{2}[n]\\-c^{-1}\Delta_-^{-1}b\end{pmatrix}=M\begin{pmatrix}0\\0\\-([n]+\frac{3}{2})c^{-1}\end{pmatrix}=\begin{pmatrix}\frac{1}{2}a\\b\\\frac{3}{2}c\end{pmatrix},$$

$$W_0=\begin{pmatrix}\frac{1}{2}[n]&0&0\\0&\frac{1}{2}([n]+1)&0\\0&0&\frac{1}{2}([n]+2)\end{pmatrix}.$$

We can also obtain a solution to the characteristic operator equation (26) by (33):

$$\Omega_J(G)=\begin{pmatrix}\Omega_{11}(G)&\Omega_{12}(G)&\Omega_{13}(G)\\\Omega_{21}(G)&\Omega_{22}(G)&\Omega_{23}(G)\\\Omega_{31}(G)&\Omega_{32}(G)&\Omega_{33}(G)\end{pmatrix},\quad G=\begin{pmatrix}G_{(1)}\\G_{(2)}\\G_{(3)}\end{pmatrix},\tag{56}$$

where $\Omega_{ij}(G)$, $i,j=1,2,3$, are determined by

$$\Omega_{11}(G)=-(E^2+E)^{-1}(cG_{(3)}+EaG_{(1)}),$$
$$\Omega_{12}(G)=E^{-1}G_{(1)},\quad \Omega_{13}(G)=G_{(2)},$$
$$\Omega_{21}(G)=cEG_{(2)}+(b-\lambda)G_{(1)},$$
$$\Omega_{22}(G)=-(E+1)^{-1}(cG_{(3)}+EaG_{(1)})+aG_{(1)},\quad \Omega_{23}(G)=G_{(1)},\tag{57}$$
$$\Omega_{31}(G)=E^{-1}cE^{-1}G_{(1)}-E^{-1}acG_{(2)},\quad \Omega_{32}(G)=E^{-1}cG_{(2)},$$
$$\Omega_{33}(G)=-E(E+1)^{-1}(cG_{(3)}+EaG_{(1)})+\Delta_+aG_{(1)}-(b-\lambda G_{(2)}.$$

By Theorem 3, we get a hierarchy of nonisospectral systems of discrete evolution equations $u_t=\rho_l=\Phi^l\rho_0$, $l\geqslant 0$, associated with the spectral problem (54).

In order to generate a vector field algebra of the isospectral hierarchy and the nonisospectral hierarchy, we need the following quantities, which may be directly worked out:

$$K_k|_{u=0}=0,\quad \rho_0|_{u=0}=J\gamma_0|_{u=0}=0,\quad \rho_l|_{u=0}=J\gamma_l|_{u=0}=M\gamma_{l-1}|_{u=0}=0,\quad k\geqslant 0,\quad l\geqslant 1,$$

$$V_k|_{u=0}=\lambda^k\begin{pmatrix}0&0&1\\0&0&0\\0&0&\lambda\end{pmatrix},$$

$$W_l|_{u=0}=\lambda^l\begin{pmatrix}\frac{1}{2}&0&0\\0&\frac{1}{2}([n]+1)&0\\0&0&\frac{1}{2}([n]+2)\end{pmatrix}+(1-\delta_{l0})\lambda^{l-1}\begin{pmatrix}0&0&-\frac{3}{2}[n]\\0&0&0\\0&0&-\frac{3}{2}\lambda[n]\end{pmatrix},$$

$$V_{k\lambda}|_{u=0}=k\lambda^{k-1}\begin{pmatrix}0&0&1\\0&0&0\\0&0&\lambda\end{pmatrix}+\lambda^k\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix},$$

$$W_{l\lambda}|_{u=0}=l\lambda^{l-1}\begin{pmatrix}\frac{1}{2}[n]&0&0\\0&\frac{1}{2}([n]+1)&0\\0&0&\frac{1}{2}([n]+2)\end{pmatrix}+(1-\delta_{l0})\begin{pmatrix}0&0&-\frac{3}{2}(l-1)\lambda^{l-2}[n]\\0&0&0\\0&0&-\frac{3}{2}l\lambda^{l-1}[n]\end{pmatrix}.$$

Now we easily find, according to the product definition of $[\![\cdot,\cdot]\!]$, that

$$[\![V_k,V_l]\!]|_{u=0}=0,\quad k,l\geqslant 0,$$

$$[\![V_k,W_l]\!]|_{u=0}=(k+1)V_{k+l}|_{u=0},\quad k,l\geqslant 0,$$

$$[\![W_k,W_l]\!]|_{u=0}=(k-l)W_{k+l}|_{u=0},\quad k,l\geqslant 0.$$

The same deduction leads to a Lax operator algebra,

$$\begin{aligned}&[\![V_k,V_l]\!]=0,\quad k,l\geqslant 0,\\&[\![V_k,W_l]\!]=(k+1)V_{k+l},\quad k,l\geqslant 0,\\&[\![W_k,W_l]\!]=(k-l)W_{k+l},\quad k,l\geqslant 0,\end{aligned}\tag{58}$$

and further a vector field algebra,

$$\begin{aligned}&[K_k,K_l]=0,\quad k,l\geqslant 0,\\&[K_k,\rho_l]=(k+1)K_{k+l},\quad k,l\geqslant 0,\\&[\rho_k,\rho_l]=(k-l)\rho_{k+l},\quad k,l\geqslant 0,\end{aligned}\tag{59}$$

which may generate a master symmetry algebra possessing the same algebraic structure as (59).

V. CONCLUSION AND REMARKS

We have established an algebraic structure related to discrete zero curvature equations and further introduced a simple but systematic approach for constructing master symmetries of the first degree for isospectral lattice hierarchies associated with discrete spectral problems. The resulting algebraic structures also leads to an explanation of why there exist master symmetries of the first degree. Some complicated calculation in our construction is saved by using a characteristic operator equation (25) [or (26)] and a uniqueness property of discrete spectral problems. The crucial step is the construction of the corresponding nonisospectral lattice hierarchies, which can be found by solving an initial nonisospectral discrete zero curvature equation. Three lattice hierarchies are shown as illustrative examples, and the corresponding master symmetry algebras of the centerless Virasoro type are exhibited. Some of the results in this paper have been reported at SIDE II, UK.[35]

It is worth noting that three examples described in the last section possess the same commutator relations between their isospectral and nonisospectral vector fields. In general, we have $[K_k,\rho_l]=(k+\gamma)K_{k+1}$, γ=const., but the other two equalities of the whole Virasoro algebra do not change. This is also a common phenomenon for continuous integrable hierarchies.[36,37] Furthermore, we may add a nonisospectral master symmetry with $\lambda_t=1$ to the whole Virasoro symmetry algebra, but this often requires additional checking. For example, a nonisospectral master symmetry with $\lambda_t=1$ of the sub-KP lattice hierarchy (55) is $\rho_{-1}=J_{\gamma-1}=(0,1,0)^T$. On the other hand, similar to the theory in Ref. 37, we may also choose an operator solution $\Omega(X)$ [or $\Omega_J(G)$] satisfying $\Omega(X)|_{X=0}=0$ [or $\Omega_J(G)|_{G=0}=0$] (all three examples in the last section have this property), and then we only need to compute $[\![V_0,W_0]\!]|_{u=0}$ so as to give Lax operator algebras at $u=0$ and finally give Lax operator algebras generally.

In our discussion, in fact, we have not used the hereditary property of the recursion operator Φ (or the bi-Hamiltonian property of J and M), while we construct Virasoro symmetry algebras for integrable lattice hierarchies, and thus it can also be applied to lattice hierarchies that possess nonhereditary recursion operators. The advantage of our scheme is to fully utilize discrete zero curvature equations so that the whole process to generate master symmetries of the first degree becomes an easy task. There were also an algorithm implemented in MuPAD[38] and other direct tricks[13–15,39] to compute master symmetries of first degree for systems of discrete evolution equations. However, our theory focuses on seeking an answer to the existence and structure problem of master symmetries of the first degree.

We should mention that there exists a large variety of other theories or methods to discuss integrable properties of systems of nonlinear discrete equations, which include Hamiltonian theory,[40,41] Bäcklund–Darboux transformation,[42,43] The R-matrix method,[34,44] symmetry reduction,[45] etc. Moreover, we can consider the time discretization problem[46] and periodic initial and boundary value problems of time discretizations[47] for symmetry flows of systems of discrete evolution equations. The resulting difference equations and mappings should be useful in discussing the integrability of the underlying systems of discrete evolution equations themselves. We are also curious about the following natural problem: Are there any higher degree master symmetries for systems of discrete evolution equations that do not depend explicitly on the evolution variable? If the answer is yes, can we establish any relations between those higher degree master symmetries and discrete zero curvature equations as we did for the first degree master symmetries?

ACKNOWLEDGMENTS

The authors are indebted to the referee for invaluable comments. One of the authors (W. X. Ma) would like to thank the Alexander von Humboldt Foundation of Germany, the City University of Hong Kong, and the Research Grants Council of Hong Kong for financial support. He is also grateful to J. Leon, W. Oevel, and W. Strampp for their helpful and stimulating discussions, and to R. K. Bullough and P. J. Caudrey for their warm hospitality during his visit at UMIST, UK.

APPENDIX A: PROOF OF THEOREM 1

Let $(K_i,V_i,f_i)\in(\mathcal{B}^q,\tilde{\mathcal{V}}^r,C^\infty(\mathbb{R}))$, $1\leq i\leq 3$. Because the bilinearity and the skew symmetry of the product (17) are self-evident and we already know that the products defined by (12) and (13) are Lie products, we only need to prove the following Jacobi identity:

$$[\![[\![V_1,V_2]\!],V_3]\!]+\text{cycle}(1,2,3)=0. \tag{A1}$$

Let us first compute by (18) that

$$\begin{aligned}[\![[\![V_1,V_2]\!],V_3]\!]&=([\![V_1,V_2]\!])'[K_3]-V_3'[[K_1,K_2]]+[[\![V_1,V_2]\!],V_3]+f_3[\![V_1,V_2]\!]_\lambda-[\![f_1,f_2]\!]V_{3\lambda}\\&=(V_1'[K_2])'[K_3]-(V_2'[K_1])'[K_3]+[V_1,V_2]'[K_3]+f_2(V_{1\lambda})'[K_3]\\&\quad-f_1(V_{2\lambda})'[K_3]-V_3'[[K_1,K_2]]+[V_1'[K_2],V_3]-[V_2'[K_1],V_3]+[[V_1,V_2],V_3]\\&\quad+f_2[V_{1\lambda},V_3]-f_1[V_{2\lambda},V_3]+f_3(V_1'[K_2])_\lambda-f_3(V_2'[K_1])_\lambda+f_3[V_1,V_2]_\lambda\\&\quad+f_{2\lambda}f_3V_{1\lambda}+f_2f_3V_{1\lambda\lambda}-f_{1\lambda}f_3V_{2\lambda}-f_1f_3V_{2\lambda\lambda}-[\![f_1,f_2]\!]V_{3\lambda}.\end{aligned} \tag{A2}$$

We need to use the following fundamental equalities:

$$(V_\lambda)'[K]=(V'[K])_\lambda,\quad V\in\tilde{\mathcal{V}}^r,\quad K\in\mathcal{B}^q,$$

$$[V,W]_\lambda=[V_\lambda,W]+[V,W_\lambda],\quad V,W\in\tilde{\mathcal{V}}^r,$$

$$[V,W]'[K]=[V'[K],W]+[V,W'[K]],\quad V,W\in\tilde{\mathcal{V}}^r,\quad K\in\mathcal{B}^q,$$

$$V'[T]=(V'[K])'[S]-(V'[S])'[K],\quad T=[K,S],\quad V\in\tilde{\mathcal{V}}^r,\quad K,S\in\mathcal{B}^q,$$

which may be shown by a direct computation and the last equality of which is a similar result as in Ref. 21. Now we can go on to compute that

$$\begin{aligned}\Delta_a^{123}&:=(V_1'[K_2])'[K_3]-(V_2'[K_1])'[K_3]-V_3'[[K_1,K_2]]\\&=(V_1'[K_2])'[K_3]-(V_2'[K_1])'[K_3-(V_3'[K_1])'[K_2+(V_3'[K_2])'[K_1],\end{aligned}$$

$$\begin{aligned}\Delta_b^{123}&:=[V_1,V_2]'[K_3]+[V_1'[K_2],V_3]-[V_2'[K_1],V_3]\\&=[V_1'[K_3],V_2]-[V_2'[K_3],V_1]+[V_1'[K_2],V_3]-[V_2'[K_1],V_3],\end{aligned}$$

$$\begin{aligned}\Delta_c^{123}&:=f_2(V_{1\lambda})'[K_3]-f_1(V_{2\lambda})'[K_3]+f_3(V_1'[K_2])_\lambda-f_3(V_2'[K_1])_\lambda\\&=f_2(V_{1\lambda})'[K_3]-f_1(V_{2\lambda})'[K_3]+f_3(V_{1\lambda})'[K_2]-f_3(V_{2\lambda})'[K_1],\end{aligned}$$

$$\begin{aligned}\Delta_d^{123}&:=f_2[V_{1\lambda},V_3]-f_1[V_{2\lambda},V_3]+f_3[V_1,V_2]_\lambda\\&=f_2[V_{1\lambda},V_3]-f_1[V_{2\lambda},V_3]+f_3[V_{1\lambda},V_2]-f_3[V_{2\lambda},V_1],\end{aligned}$$

$$\begin{aligned}\Delta_e^{123}&:=f_{2\lambda}f_3V_{1\lambda}+f_2f_3V_{1\lambda\lambda}-f_{1\lambda}f_3V_{2\lambda}-f_1f_3V_{2\lambda\lambda}-[\![f_1,f_2]\!]V_{3\lambda},\\&=f_{2\lambda}f_3V_{1\lambda}+f_2f_3V_{1\lambda\lambda}-f_{1\lambda}f_3V_{2\lambda}-f_1f_3V_{2\lambda\lambda}-f_{1\lambda}f_2V_{3\lambda}+f_1f_{2\lambda}V_{3\lambda}.\end{aligned}$$

A direct check can result in that

$$\Delta_*^{123}+\text{cycle}(1,2,3)=0,\quad\text{where}\quad *=a,b,c,d\quad\text{or}\quad e.$$

Noting (A2), it follows therefore that

$$[\![[\![V_1,V_2]\!],V_3]\!]+\text{cycle}(1,2,3)=\Delta_a^{123}+\Delta_b^{123}+\Delta_c^{123}+\Delta_d^{123}+\Delta_e^{123}+[[V_1,V_2],V_3]+\text{cycle}(1,2,3)=0,$$

which is exactly the Jacobi identity (A1) and thus completes the proof.

APPENDIX B: PROOF OF THEOREM 2

The proof is an application of the equalities (19) and (20) and the third equality,

$$(U'[K])'[S]-(U'[S])'[K]=U'[T],\quad T=[K,S],\tag{B1}$$

which has been mentioned in the proof of the first theorem. We observe that

$$[\text{Eq. }(19)]'[S]-[\text{Eq. }(20)]'[K]+g[\text{Eq. }(19)]_\lambda-f[\text{Eq. }(20)]_\lambda.$$

The resulting equality reads as

$$\begin{aligned}(U'[K])'[S]&-(U'[S])'[K]+[\![f,g]\!]U_\lambda\\&=(EV'[S])U+(EV)U'[S]-U'[S]V-UV'[S]-(EW'[K])U-(EW)U'[K]\\&\quad+U'[K]W+UW'[K]+g(EV_\lambda)U+g(EV)U_\lambda-gU_\lambda V\\&\quad-gUV_\lambda-f(EW_\lambda)U-f(EW)U_\lambda+fU_\lambda W+fUW_\lambda.\end{aligned}\tag{B2}$$

On the other hand, we have immediately

$$(E[\![V,W]\!])U-U[\![V,W]\!]=(EV'[S])U-(EW'[K])U+(EV)(EW)U-(EW)(EV)U+g(EV_\lambda)U \\ -f(EW_\lambda)U-UV'[S]+UW'[K]-UVW+UWV-gUV_\lambda+fUW_\lambda. \tag{B3}$$

It follows, therefore from (B1), (B2), and (B3) that

$$\begin{aligned}
&(E[\![V,W]\!])U-U[\![V,W]\!]-U'[T]-[\![f,g]\!]U_\lambda\\
&\quad=(E[\![V,W]\!])U-U[\![V,W]\!]-(U'[K])'[S]+(U'[S])'[K]-[\![f,g]\!]U_\lambda\\
&\quad=(EV)\{(EW)U-V'[S]-gU_\lambda\}-(EW)\{(EV)U-U'[K]-fU_\lambda\}\\
&\qquad-UVW+UWV+gU_\lambda V-fU_\lambda W+U'[S]V-U'[K]W\\
&\quad=(EV)UW-(EW)UV-UVW+UWV+gU_\lambda V-fU_\lambda W+U'[S]V-U'[K]W\\
&\quad=\{(EV)U-UV-fU_\lambda-U'[K]\}W-\{(EW)U-UW-gU_\lambda-U'[S]\}V=0,
\end{aligned}$$

which is what we need to prove.

APPENDIX C: PROOF OF THEOREM 3

We prove two equalities in (32). The rest is obvious. We compute that

$$\begin{aligned}
(EV_k)U-UV_k&=\lambda^k[(EV_0)U-UV_0]+\sum_{i=1}^{k}\lambda^{k-i}\{(E\Omega(K_{i-1}))U-U\Omega(K_{i-1})\}\\
&=\lambda^kU'[K_0]+\sum_{i=1}^{k}\lambda^{k-i}\{U'[\Phi K_{i-1}]-\lambda U'[K_{i-1}]\}\\
&=\lambda^kU'[K_0]+\sum_{i=1}^{k}\lambda^{k-i}\{U'[K_i]-\lambda U'[K_{i-1}]\}=U'[K_k],\quad k\geqslant 1;
\end{aligned}$$

$$\begin{aligned}
(EW_l)U-UW_l&=\lambda^l[(EW_0)U-UW_0]+\sum_{j=1}^{l}\lambda^{l-j}\{(E\Omega(\rho_{j-1}))U-U\Omega(\rho_{j-1})\}\\
&=\lambda^l\{U'[\rho_0]+\lambda U_\lambda\}+\sum_{j=1}^{l}\lambda^{l-j}\{U'[\Phi\rho_{j-1}]-\lambda U'[\rho_{j-1}]\}\\
&=\lambda^l\{U'[\rho_0]+\lambda U_\lambda\}+\sum_{j=1}^{l}\lambda^{l-j}\{U'[\rho_j]-\lambda U'[\rho_{j-1}]\}\\
&=U'[\rho_l]+\lambda^{l+1}U_\lambda,\\
&\qquad l\geqslant 1.
\end{aligned}$$

Note that we have used the characteristic operator equation (25), but the situation in the case of (26) is completely similar. The proof is therefore finished.

[1] A. V. Mikhailov, A. B. Shabat, and V. V. Sokolov, in *What is Integrability?*, edited by V. E. Zakhalov (Springer-Verlag, Berlin, 1991), pp. 115–184.
[2] D. Levi and R. I. Yamilov, J. Math. Phys. **38**, 6648–6674 (1997).

[3] B. Fuchssteiner and A. S. Fokas, Physica D **4**, 47–66 (1981).
[4] A. S. Fokas, Stud. Appl. Math. **77**, 253–299 (1987).
[5] B. Fuchssteiner, Prog. Theor. Phys. **70**, 1508–1522 (1983).
[6] W. X. Ma, J. Phys. A **23**, 2707–2716 (1990).
[7] H. H. Chen and Y. C. Lee, in *Advances in Nonlinear Waves*, Research Notes in Mathematics 111, edited by L. Debnath (Pitman, Boston, 1985), pp. 233–239.
[8] A. Y. Orlov and E. I. Schulman, Lett. Math. Phys. **12**, 171–179 (1986).
[9] Y. S. Li and G. C. Zhu, J. Phys. A **19**, 3713–3725 (1986).
[10] W. X. Ma, Phys. Lett. A **179**, 179–185 (1993); *Proceedings of the 21st International Conference on the Differential Geometry Methods in Theoretical Physics*, edited by M. L. Ge (World Scientific, Singapore, 1993), pp. 535–538; J. Phys. A **25**, L719–L726 (1992).
[11] F. Calogero and A. Degasperis, Lett. Nuovo Cimento **22**, 420–424 (1978).
[12] W. L. Chan and Y. K. Zheng, Lett. Math. Phys. **14**, 293–301 (1987).
[13] W. Oevel, H. W. Zhang, B. Fuchssteiner, and O. Ragnisco, J. Math. Phys. **30**, 2664–2670 (1989).
[14] W. Oevel, B. Fuchssteiner, and H. W. Zhang, Prog. Theor. Phys. **81**, 294–308 (1989).
[15] H. W. Zhang, G. Z. Tu, W. Oevel, and B. Fuchssteiner, J. Math. Phys. **32**, 1908–1918 (1991).
[16] M. J. Ablowitz and J. F. Ladik, J. Math. Phys. **16**, 598–603 (1975); **17**, 1011–1018 (1976).
[17] D. Levi, M. Bruschi, and O. Ragnisco, Nuovo Cimento A **58**, 56–66 (1980).
[18] M. Bruschi, D. Levi, and O. Ragnisco, J. Math. Phys. **22**, 2463–2471 (1981).
[19] M. Bruschi and O. Ragnisco, Lett. Nuovo Cimento **31**, 492–496 (1981).
[20] W. X. Ma, J. Phys. A **26**, 2573–2582 (1993).
[21] W. X. Ma, J. Phys. A **25**, 5329–5343 (1992).
[22] W. X. Ma and F. K. Guo, Int. J. Theor. Phys. **36**, 697–704 (1997).
[23] B. Fuchssteiner, Nonlinear Anal. TMA **3**, 849–862 (1979).
[24] F. Magri, J. Math. Phys. **19**, 1156–1162 (1978).
[25] I. M. Gel'fand and I. Y. Dorfman, Funct. Anal. Appl. **13**, 248–262 (1979).
[26] W. X. Ma, Chin. Sci. Bull. **38**, 2025–2031 (1993).
[27] W. X. Ma and W. Strampp, Phys. Lett. A **185**, 277–286 (1994).
[28] V. Volterra, *Lecons sur la Theorie Mathematique de la Lutte Pour la Vie* (Gauthier-Villars, Paris, 1931).
[29] O. I. Bogoyavlensky, Phys. Lett. A **134**, 34–38 (1988); Math. USSR Izv. **31**, 435–454 (1988).
[30] G. Z. Tu, J. Phys. A **23**, 3903–3922 (1990).
[31] B. G. Kupershmidt, ''Discrete Lax equations and differential-difference calculus,'' Asterisque, **123** (Soc. Math. France, Paris, 1985).
[32] M. Toda, *Theory of Nonlinear Lattices*, 2nd enl. ed. (Springer-Verlag, Berlin, 1989).
[33] Y. T. Wu and X. G. Geng, J. Math. Phys. **37**, 2338–2345 (1996).
[34] M. Blaszak and K. Marciniak, J. Math. Phys. **35**, 4661–4682 (1994).
[35] B. Fuchssteiner and W. X. Ma, ''An approach to master symmetries of lattice equations,'' in *Symmetries and Integrability of Difference Equations*, edited by P. A. Clarkson and F. W. Nijhoff (Cambridge University Press, London, 1999).
[36] Y. Cheng and Y. S. Li, Chin. Sci. Bull. **36**, 1428–1433 (1991).
[37] W. X. Ma, J. Math. Phys. **33**, 2464–2476 (1992).
[38] B. Fuchssteiner, S. Ivanov, and W. Wiwianka, Math. Comput. Modell. **25**, 91–100 (1997).
[39] I. Y. Cherdantsev and R. I. Yamilov, Physica D **87**, 140–144 (1995).
[40] M. Bruschi and O. Ragnisco, J. Math. Phys. **24**, 1414–1421 (1983).
[41] O. Ragnisco and P. M. Santini, Inverse Probl. **6**, 441–452 (1990); M Bruschi, O. Ragnisco, P. M. Santini, and G. Z. Tu, Physica D **49**, 273–294 (1991).
[42] M. Bruschi and O. Ragnisco, Phys. Lett. A **129**, 21–25 (1988).
[43] W. Oevel, in *Nonlinear Physics: Theory and Experiment*, edited by E. Alfinito, M. Boiti, L. Martina, and F. Pempinelli (World Scientific, Singapore, 1996), pp. 223–240.
[44] C. Morosi and L. Pizzocchero, J. Math. Phys. **37**, 4484–4513 (1996).
[45] D. Levi and P. Winternitz, Phys. Lett. A **152**, 335–338 (1991); J. Math. Phys. **34**, 3713–3730 (1993).
[46] Yu. B. Suris, J. Phys. A **29**, 451–46 (1996); J. Math. Phys. **37**, 3982–3996 (1996).
[47] V. G. Papageorgiou, F. W. Nijhoff and H. W. Capel, Phys. Lett. A **147**, 106–114 (1990); H. W. Capel, F. W. Nijhoff, and V. G. Papageorgiou, *ibid.* **155**, 377–387 (1991).

On the Fay identity for Korteweg–de Vries tau functions and the identity for the Wronskian of squared solutions of Sturm–Liouville equation

Yordan P. Mishev[a)]
Research Institute for Mathematical Sciences, Kyoto University, Sakyo-ku 606, Kyoto, Japan

(Received 9 April 1998; accepted for publication 20 January 1999)

We show that the well-known identity for the Wronskian of squared solutions of a Sturm–Liouville equation follows from the Fay identity. We also study some odd-order $[(2^n-1)$-order, $n=2,3,...]$ identities which are specific for tau functions, related to the KdV hierarchy.
[S0022-2488(99)00705-7]

I. INTRODUCTION

We began this research as a study of the expression of the Wronskian of squared solutions of the Sturm–Liouville equation by Korteweg–de Vries (KdV) tau functions and the Fay identity. Now, when the desired result is obtained (cf. Theorem 1.1), we realize that this is a story of the common origin of the following three relations for the functions: x, $\sin(x)$, and $\theta_{11}(x)$ respectively $(x,z_1,z_2\in\mathbb{C})$:

$$(z_2-z_1)[(x+z_1+z_2)(x-z_1)(x-z_2)-(x-z_1-z_2)(x+z_1)(x+z_2)]$$
$$=(z_1+z_2)[(x+z_1-z_2)(x-z_1)(x+z_2)-(x-z_1+z_2)(x+z_1)(x-z_2)],$$

$$\sin(z_2-z_1)[\sin(x+z_1+z_2)\sin(x-z_1)\sin(x-z_2)-\sin(x-z_1-z_2)\sin(x+z_1)\sin(x+z_2)]$$
$$=\sin(z_1+z_2)[\sin(x+z_1-z_2)\sin(x-z_1)\sin(x+z_2)$$
$$-\sin(x-z_1+z_2)\sin(x+z_1)\sin(x-z_2)],$$

$$\theta_{11}(z_2-z_1)[\theta_{11}(x+z_1+z_2)\theta_{11}(x-z_1)\theta_{11}(x-z_2)-\theta_{11}(x-z_1-z_2)\theta_{11}(x+z_1)\theta_{11}(x+z_2)]$$
$$=\theta_{11}(z_1+z_2)[\theta_{11}(x+z_1-z_2)\theta_{11}(x-z_1)\theta_{11}(x+z_2)$$
$$-\theta_{11}(x-z_1+z_2)\theta_{11}(x+z_1)\theta_{11}(x-z_2)]$$

(we use the notations for theta functions from Ref. 1).

In the present paper we will prove an identity for general KdV tau functions (it will be a third-order identity for tau). The mentioned three types of functions are, roughly speaking, three types of KdV tau functions for stationary (t-independent) solutions of the KdV equation: $u_t=6uu_x+u_{xxx}$. So, the polynomial relation follows from this cubic identity for KdV taus, but there are some specific problems to translate the identity for the general KdV taus to the cases of trigonometric functions and elliptic theta functions. We will postpone the solution of these problems to some next publication.

This way, the status of the three relations is quite different: the first one is easy to prove directly (and it also follows from the cubic identity for KdV taus); the second one is not difficult

[a)]On leave from Forestry University, Sofia, Bulgaria. Electronic mail: mishevyp@kurims.kyoto-u.ac.jp to prove directly, using the well-known trigonometric identities; and the third one is still conjectural (we could not derive it from the Riemann relations, we could only check it numerically, using the system Mathematica 3.0 at RIMS, Kyoto University).

Let $\tau(t)$, $t\equiv(t_1,t_2,t_3,\dots)\in\mathbb{C}^\infty$, $t_1\equiv x$, be an arbitrary tau function, related to the Kadomtzev–Petviashvili (KP) hierarchy (Ref. 2). Let us denote ($z\in\mathbb{C}$)

$$[z]:=(z,z^2/2,z^3/3,\dots)\in\mathbb{C}^\infty,$$

$$\tau(t+[z]):=\tau(t_1+z,t_2+z^2/2,t_3+z^3/3,\dots).$$

The following identity ($z_0,z_1,z_2,z_3\in\mathbb{C}$),

$$(z_0-z_1)(z_2-z_3)\tau(t+[z_0]+[z_1])\tau(t+[z_2]+[z_3])+(z_0-z_2)(z_3-z_1)\tau(t+[z_0]+[z_2])\tau(t+[z_3]+[z_1])+(z_0-z_3)(z_1-z_2)\tau(t+[z_0]+[z_3])\tau(t+[z_1]+[z_2])=0 \tag{1.1}$$

is called the *Fay identity* (Ref. 3) for the KP tau function τ. It was first obtained (Ref. 4) for theta functions related to Jacobians. In genus $g=1$ case its form is

$$\theta_{11}(z_0-z_1)\theta_{11}(z_2-z_3)\theta_{11}(t+z_0+z_1)\theta_{11}(t+z_2+z_3)+\theta_{11}(z_0-z_2)\theta_{11}(z_3-z_1)\theta_{11}(t+z_0+z_2)\theta_{11}(t+z_3+z_1)+\theta_{11}(z_0-z_3)\theta_{11}(z_1-z_2)\theta_{11}(t+z_0+z_3)\theta_{11}(t+z_1+z_2)=0.$$

Afterwards it was used (Ref. 1) in geometric treatment of soliton equations. Later it was generalized for tau functions (Ref. 3).

The Fay identity is fulfilled at also for tau functions related to the nth ($n=2,3,4,\dots$) Gel'fand–Dickey reduction of the KP hierarchy. In the present paper we will consider only the $n=2$ reduction, i.e., the KdV hierarchy. Such tau functions we will call KdV tau functions. They can be characterized by the conditions ($\partial_{t_{2k}}\equiv\partial/\partial t_{2k}$)

$$\partial_{t_{2k}}\tau(t)=0,\quad k=1,2,3,\dots,$$

which imply for every $z\in\mathbb{C}$

$$\tau(t-[z])=\tau(t+[-z]). \tag{1.2}$$

There are two main goals in the present article. The first aim is to show that the famous identity for the Wronskian $[W(f,g):=fg'-f'g,\ '\equiv\partial_x\equiv\partial/\partial x]$ of squared solutions of the Sturm–Liouville equation (Ref. 5) follows from the Fay identity for KdV tau functions. The second aim is to obtain some specific relations for the KdV tau functions.

We came to these results when studying the problem of finding a dictionary between the tau functions and some formulas related to squared solutions of the Sturm–Liouville equation (especially the mentioned identity for the Wronskian of squared solutions–an important ingredient of this area (Refs. 5 and 6). Such a dictionary will be useful in examining some features of Miura transformations. It is well known that squared solutions span the kernels of the Frechet derivatives $\mathfrak{M}'_\pm(v)=2v\pm\partial_x$ of Miura transformations $u_\pm=\mathfrak{M}_\pm v:=v^2\pm v_x$ ($v_x\equiv\partial_x v$), where v is a solution of the mKdV equation and $u_\pm$ are solutions of the KdV equation. Also well known (Ref. 7) is the interpretation of Miura transformations as projections from flag to corresponding subspace (in Sato Grassmannian), which is intimately connected to tau functions. Some parts of the dictionary were known: e.g., a formula, which expresses the squared solutions by means of $\tau(t)$ and vertex operators (Ref. 2) [in this paper the so-called Λ-operators (Ref. 5) are also mentioned]. There were no relations to the tau functions of the identity for the Wronskian of squared solutions, but there was a well-known expression of the Wronskian of two solutions by means of $\tau(t)$ (Refs. 2 and 8).

Because of the fact that in the proof of the latest formula, the Fay identity was used, we expected that the same identity will be useful in the ‘‘paraphrase’’ of the Wronskian of squared solutions.

We need such a dictionary, because we observed some similarities between Matsuo and Cherednik transformations (Ref. 9):

Knizhniv-Zamolodchikov equation→quantum Calogero-Sutherland system on the one hand and Miura transformation on the other hand. Our opinion is that such similarities will be easier explained on the language of tau functions, flag and Grassmann manifolds, etc. So, the ‘‘paraphrase’’ of the relations for the Wronskian of squared solutions of the Sturm–Liouville equation is only the first step in this direction. We also expect that the presented connections between squared solutions and tau functions will be useful in another areas of the subject (cf. Refs. 10 and 11).

In order to explain the main results of the present article, let us review some notations (Refs. 2 and 8). Let $\psi(x,z)$ and $\psi^*(x,z)$ be two linearly independent solutions (cf. Sec. II) of the Sturm–Liouville equation:

$$(\partial_x^2+u(x))\psi(x,z)=z^2\psi(x,z). \tag{1.3}$$

Then the following relations,

$$\begin{aligned}W(\psi(x,z_1)\psi^*(x,z_1),\psi(x,z_2)\psi^*(x,z_2))&\\ =-(z_1^2-z_2^2)^{-1}\partial_x[W(\psi(x,z_1),\psi(x,z_2))W(\psi^*(x,z_1),\psi^*(x,z_2))]&\\ =-(z_1^2-z_2^2)^{-1}\partial_x[W(\psi(x,z_1),\psi^*(x,z_2))W(\psi^*(x,z_1),\psi(x,z_2))]&\end{aligned} \tag{1.4}$$

$(z_1,z_2\in\mathbb{C})$, we will call the *Faddeev–Tahtajan identity*. This relation has a long history. It was used in the theory of inverse spectral problems for the Sturm–Liouville operators. Afterwards the Faddeev–Tahtajan identity played an important role in the first years of soliton theory. In Ref. 5 the origin of the identity is interpreted in terms of classical *r*-matrixes. Here we will explain the origin of the Faddeev–Tahtajan identity using the language of tau functions.

The first main result in this paper is given in the following.

Theorem 1.1: *The Faddeev–Tahiajan identity (1.4) follows from the Fay identity (1.1) for KdV tau functions.*

The second main result in the present article is given in the following.

Theorem 1.2: *Let $\tau(t)$, $t\in\mathbb{C}^\infty$ be an arbitrary KdV tau function. Then*

(i) for every $z_1,z_2\in\mathbb{C}$,

$$\begin{aligned}(z_2-z_1)[\tau(t+[z_1]+[z_2])\tau(t-[z_1])\tau(t-[z_2])-\tau(t-[z_1]-[z_2])\tau(t+[z_1])\tau(t+[z_2])]&\\ =(z_2+z_1)[\tau(t+[z_1]-[z_2])\tau(t-[z_1])\tau(t+[z_2])-\tau(t-[z_1]+[z_2])\tau(t+[z_1])&\\ \times\tau(t-[z_2])];&\end{aligned}$$

(ii) for every $z\in\mathbb{C}$

$$\begin{aligned}\tau(t+2[z])\tau^2(t-[z])-\tau(t-2[z])\tau^2(t+[z])&\\ =2\sum_{k=0}^{\infty}z^{2k+1}[\tau(t-[z])W_{2k+1}(\tau(t),\tau(t+[z]))+\tau(t+[z])W_{2k+1}(\tau(t),\tau(t-[z]))],&\end{aligned}$$

where we denote $W_{2k+1}(f,g):=f(\partial_{t_{2k+1}}g)-(\partial_{t_{2k+1}}f)g$, $k=0,1,2,\ldots$.

Remark 1.3: Let us mention that the identities from Theorem 1.2 are cubic in τ relations (contrary to the Fay identity, which is quadratic in the τ relation) and they are specific only for the KdV tau functions.

Remark 1.4: The proof of Theorem 1.2 is based *only* on the following three facts:

(i) The Fay identity (1.1) (which is common for all tau functions),
(ii) the relation (1.2) (which is specific only for KdV tau functions), and
(iii) the obvious identity for Wronskians:

$$W(f_1f_2,g_1g_2)=f_1g_1W(f_2,g_2)+f_2g_2W(f_1,g_1)=f_1g_2W(f_2,g_1)+f_2g_1W(f_1,g_2). \qquad (1.5)$$

The paper consists of four sections. In Sec. II we give some preliminary results. The proofs of Theorems 1.1 and 1.2 are given in Sec. III. In Sec. IV we give some examples and comments of the main statements. A preliminary (and from different viewpoint) version of some of the results is presented in Refs. 12 and 13.

II. PRELIMINARY RESULTS

First, let us mention some obvious relations for Wronskians.
Lemma 2.1:

$$\text{(i)}\quad W(e^{z_1x}f,e^{z_2x}g)=e^{(z_1+z_2)x}[W(f,g)-(z_1-z_2)fg],$$

$$\text{(ii)}\quad W\left(\frac{f_1}{g},\frac{f_2}{g}\right)=\frac{W(f_1,f_2)}{g^2},$$

$$\text{(iii)}\quad \partial_x\left(\frac{f_1f_2}{g^2}\right)=-\frac{f_1W(f_2,g)+f_2W(f_1,g)}{g^3}.$$

Instead of Fay identity (1.1) we will use the differential Fay identity of Ref. 8 ($z_1,z_2\in\mathbb{C}$):

$$W(\tau(t+[z_1]),\tau(t+[z_2]))=(z_2^{-1}-z_1^{-1})[\tau(t+[z_1])\tau(t+[z_2])-\tau(t)\tau(t+[z_1]+[z_2])]. \qquad (2.1)$$

Shifting the argument t respectively to $(t-[z_1]-[z_2])$, $(t-[z_2])$, and $(t-[z_1])$ we could obtain expressions respectively for the following Wronskians:

$$W(\tau(t-[z_1],\tau(t-[z_2])),\quad W(\tau(t+[z_1]-[z_2]),\tau(t)),\quad W(\tau(t-[z_1]+[z_2]),\tau(t)).$$

But, shifting t we cannot obtain an expression, e.g., for the Wronskian:

$$W(\tau(t+[z_1]),\tau(t-[z_2])).$$

This is possible for KdV tau functions. Using (2.1) and (1.2), it is easy to see that

$$W(\tau(t+[z_1]),\tau(t-[z_2]))=-(z_2^{-1}+z_1^{-1})[\tau(t+[z_1])\tau(t-[z_2])-\tau(t)\tau(t+[z_1]-[z_2])].$$

This way we obtain the following expressions for the Wronskians of KdV tau functions.
Lemma 2.2: Let $\tau(t)$ be an arbitrary KdV tau function. Then we have

$$\text{(i)}\quad W(\tau(t+[z_1]),\tau(t+[z_2]))=(z_2^{-1}-z_1^{-1})[\tau(t+[z_1])\tau(t+[z_2])-\tau(t)\tau(t+[z_1]+[z_2])],$$

$$W(\tau(t-[z_1]),\tau(t-[z_2]))=-(z_2^{-1}-z_1^{-1})[\tau(t-[z_1])\tau(t-[z_2])-\tau(t)\tau(t-[z_1]-[z_2])],$$

$$W(\tau(t-[z_1]),\tau(t+[z_2]))=(z_2^{-1}+z_1^{-1})[\tau(t-[z_1])\tau(t+[z_2])-\tau(t)\tau(t-[z_1]+[z_2])],$$

$$W(\tau(t+[z_1]),\tau(t-[z_2]))=-(z_2^{-1}+z_1^{-1})[\tau(t+[z_1])\tau(t-[z_2])-\tau(t)\tau(t+[z_1]-[z_2])],$$

$$\text{(ii)}\quad W(\tau(t+[z_1]-[z_2]),\tau(t))=(z_2^{-1}-z_1^{-1})[\tau(t+[z_1]-[z_2])\tau(t)-\tau(t+[z_1])\tau(t-[z_2])],$$

$$W(\tau(t-[z_1]+[z_2]),\tau(t))=-(z_2^{-1}-z_1^{-1})[\tau(t-[z_1]+[z_2])\tau(t)-\tau(t-[z_1])\tau(t+[z_2])],$$

$$W(\tau(t-[z_1]-[z_2]),\tau(t))=(z_2^{-1}+z_1^{-1})[\tau(t-[z_1]-[z_2])\tau(t)-\tau(t-[z_1])\tau(t-[z_2])],$$

$$W(\tau(t+[z_1]+[z_2]),\tau(t))=-(z_2^{-1}+z_1^{-1})[\tau(t+[z_1]+[z_2])\tau(t)-\tau(t+[z_1])\tau(t+[z_2])].$$

Let us define the *wave functions* $\psi(t,z)$ and $\psi^*(t,z)$ in Ref. 8 by expressions

$$\psi(t,z):=\exp\left(\sum_{k=1}^{\infty} t_k z^k\right)\frac{\tau(t-[z^{-1}])}{\tau(t)},$$

$$\psi^*(t,z):=\exp\left(-\sum_{k=1}^{\infty} t_k z^k\right)\frac{\tau(t+[z^{-1}])}{\tau(t)}.$$

For an arbitrary KdV tau function $\tau(t)$, denoting $u(t):=2\partial_x^2 \ln \tau(t)$, it is well-known (Ref. 8) that the wave functions $\psi(t,z)$ and $\psi^*(t,z)$ satisfy the Sturm–Liouville equation (1.3) ($t_1\equiv x$ and $t_3,t_5,\ldots$ are parameters). Using the relations of Lemma 2.2 we can explain the Wronskians of the wave functions $\psi(t,z)$ and $\psi^*(t,z)$ in terms of the tau function $\tau(t)$.

Lemma 2.3: Let $\tau(t)$ be an arbitrary KdV tau function and $\psi(t,z)$ and $\psi^*(t,z)$ are the corresponding wave functions. Then we have ($z_1,z_2\in\mathbb{C}$)

$$\text{(i)}\quad W(\psi(t,z_1),\psi(t,z_2))=(z_1-z_2)\exp\left(\sum_{k=0}^{\infty} t_{2k+1}(z_1^{2k+1}+z_2^{2k+1})\right)\frac{\tau(t-[z_1^{-1}]-[z_2^{-1}])}{\tau(t)},$$

$$\text{(ii)}\quad W(\psi^*(t,z_1),\psi^*(t,z_2))=-(z_1-z_2)\exp\left(-\sum_{k=0}^{\infty} t_{2k+1}(z_1^{2k+1}+z_2^{2k+1})\right)\times\frac{\tau(t+[z_1^{-1}]+[z_2^{-1}])}{\tau(t)},$$

$$\text{(iii)}\quad W(\psi(t,z_1),\psi^*(t,z_2))=(z_1+z_2)\exp\left(\sum_{k=0}^{\infty} t_{2k+1}(z_1^{2k+1}-z_2^{2k+1})\right)\frac{\tau(t-[z_1^{-1}]+[z_2^{-1}])}{\tau(t)},$$

$$\text{(iv)}\quad W(\psi^*(t,z_1),\psi(t,z_2))=-(z_1+z_2)\exp\left(-\sum_{k=0}^{\infty} t_{2k+1}(z_1^{2k+1}-z_2^{2k+1})\right)\frac{\tau(t+[z_1^{-1}]-[z_2^{-1}])}{\tau(t)}.$$

Proof: Let us denote the functions

$$\varphi(t,z):=e^{zx}\frac{\tau(t-[z^{-1}])}{\tau(t)},\qquad \varphi^*(t,z):=e^{zx}\frac{\tau(t+[z^{-1}])}{\tau(t)}.$$

Then we have

$$\psi(t,z)=\exp\left(\sum_{k=1}^{\infty} t_{2k+1}z^{2k+1}\right)\varphi(t,z),\qquad \psi^*(t,z)=\exp\left(-\sum_{k=1}^{\infty} t_{2k+1}z^{2k+1}\right)\varphi^*(t,z),$$

and consequently we have

$$W(\psi(t,z_1),\psi(t,z_2))=\exp\left(\sum_{k=1}^{\infty} t_{2k+1}(z_1^{2k+1}+z_2^{2k+1})\right)W(\varphi(t,z_1),\varphi(t,z_2)),\qquad \text{etc.}$$

Using the relations of Lemmas 2.1 and 2.2 we obtain

$$
\begin{aligned}
W(\varphi(t,z_1),\varphi(t,z_2)) &= W\left(e^{z_1 x}\frac{\tau(t-[z_1^{-1}])}{\tau(t)}, e^{z_2 x}\frac{\tau(t-[z_2^{-1}])}{\tau(t)} \right)\\
&= e^{(z_1+z_2)x}\left[W\left(\frac{\tau(t-[z_1^{-1}])}{\tau(t)}, \frac{\tau(t-[z_2^{-1}])}{\tau(t)} \right) \right.\\
&\quad \left. -(z_1-z_2)\frac{\tau(t-[z_1^{-1}])\tau(t-[z_2^{-1}])}{\tau^2(t)} \right]\\
&= e^{(z_1+z_2)x}\left[\frac{W(\tau(t-[z_1^{-1}]),\tau(t-[z_2^{-1}]))}{\tau^2(t)} \right.\\
&\quad \left. -(z_1-z_2)\frac{\tau(t-[z_1^{-1}])\tau(t-[z_2^{-1}])}{\tau^2(t)} \right]\\
&= \frac{e^{(z_1+z_2)x}}{\tau^2(t)}[(z_1-z_2)(\tau(t-[z_1^{-1}])\tau(t-[z_2^{-1}])-\tau(t)\tau(t-[z_1^{-1}]-[z_2^{-1}]))\\
&\quad -(z_1-z_2)\tau(t-[z_1^{-1}])\tau(t-[z_2^{-1}])]\\
&= (z_1-z_2)e^{x(z_1+z_2)}\frac{\tau(t-[z_1^{-1}]-[z_2^{-1}])}{\tau(t)}.
\end{aligned}
$$

From here follows (i), because we have ($t_1 \equiv x$)

$$
e^{x(z_1+z_2)}\exp\left(\sum_{k=1}^{\infty} t_{2k+1}(z_1^{2k+1}+z_2^{2k+1}) \right) = \exp\left(\sum_{k=0}^{\infty} t_{2k+1}(z_1^{2k+1}+z_2^{2k+1}) \right).
$$

It is easy to proove (ii), (iii), and (iv) in the same way. □

III. PROOF OF THE MAIN RESULTS

Proof of Theorem 1.2: First we will prove the identity (i). Using the identities of Lemma 2.2 and (1.5), let us expand the following Wronskian:

$$
W(\tau(t+[z_1])\tau(t-[z_1]),\tau(t+[z_2])\tau(t-[z_2]))
$$

in two different ways. From the first line of (1.5) we obtain

$$
\begin{aligned}
&(z_2^{-1}-z_1^{-1})\tau(t)[\tau(t-[z_1]-[z_2])\tau(t+[z_1])\tau(t+[z_2])\\
&\quad -\tau(t+[z_1]+[z_2])\tau(t-[z_1])\tau(t-[z_2])],
\end{aligned}
$$

and from the second line of (1.5) we obtain

$$
\begin{aligned}
&(z_2^{-1}+z_1^{-1})\tau(t)[\tau(t+[z_1]-[z_2])\tau(t-[z_1])\tau(t+[z_2])\\
&\quad -\tau(t-[z_1]+[z_2])\tau(t+[z_1])\tau(t-[z_2])].
\end{aligned}
$$

But $(z_2^{-1}-z_1^{-1})=(z_1-z_2/z_1z_2$ and $(z_2^{-1}+z_1^{-1})=(z_1+z_2)/z_1z_2$, so we have

$$
\frac{\tau(t)}{z_1z_2}[\text{lhs of } (i)] = \frac{\tau(t)}{z_1z_2}[\text{rhs of } (i)].
$$

The proof of the first identity of Theorem 1.2 is completed.

Now we will obtain the second identity (ii) of Theorem 1.2 letting z_2 tend to z_1 in the first identity (i) (we will denote $z_1=z_2=z$). The lhs of (ii) is clear. In order to obtain the rhs of (ii) we mention that:

$$\partial_{z_2}(\tau(t+[z_1]-[z_2]))\big|_{z_2=z_1=z}=\partial_{z_2}\left[\tau\left((x+z_1)-z_2,\left(t_3+\frac{z_1^3}{3}\right)-\frac{z_2^3}{3},\left(t_5+\frac{z_1^5}{5}\right)-\frac{z_2^5}{5},\dots\right)\right]\Bigg|_{z_2=z_1=z}$$

$$=-\sum_{k=0}^{\infty} z^{2k}\partial_{t_{2k+1}}\tau(t).$$

In the same way we obtain

$$\partial_{z_2}(\tau(t-[z_1]+[z_2]))\big|_{z_2=z_1=z}=\sum_{k=0}^{\infty} z^{2k}\partial_{t_{2k+1}}\tau(t),$$

$$\partial_{z_2}(\tau(t+[z_2]))\big|_{z_2=z}=\sum_{k=0}^{\infty} z^{2k}\partial_{t_{2k+1}}\tau(t+[z]),$$

$$\partial_{z_2}(\tau(t-[z_2]))\big|_{z_2=z}=-\sum_{k=0}^{\infty} z^{2k}\partial_{t_{2k+1}}\tau(t-[z]),$$

So, from the rhs of (i) we obtain

$$\tau(t-[z])\sum_{k=0}^{\infty} 2z^{2k+1}[\tau(t)\partial_{t_{2k+1}}\tau(t+[z])-\tau(t+[z])\partial_{t_{2k+1}}\tau(t)]$$
$$+\tau(t+[z])\sum_{k=0}^{\infty} 2z^{2k+1}[\tau(t)\partial_{t_{2k+1}}\tau(t-[z])-\tau(t-[z])\partial_{t_{2k+1}}\tau(t)],$$

which gives the rhs of (ii). □

Proof of Theorem 1.1: On the one hand, using the expressions of the wave functions $\psi(t,z)$ and $\psi^*(t,z)$ in terms of tau function $\tau(t)$ (in our case τ is an arbitrary KdV tau function) we obtain from the first line of (1.4)

$$\mathbf{W}\equiv W[\psi(t,z_1)\psi^*(t,z_1),\psi(t,z_2)\psi^*(t,z_2)]$$
$$=W\left(\frac{\tau(t+[z_1^{-1}])\tau(t-[z_1^{-1}])}{\tau^2(t)},\frac{\tau(t+[z_2^{-1}])\tau(t-[z_2^{-1}])}{\tau^2(t)}\right)$$
$$=\frac{1}{\tau^4(t)}W(\tau(t+[z_1^{-1}])\tau(t-[z_1^{-1}]),\tau(t+[z_2^{-1}])\tau(t-[z_2^{-1}])).$$

From the proof of the identity (i) of the Theorem 1.2 we know that this equals either

$$\frac{z_2-z_1}{\tau^3(t)}[\tau(t-[z_1^{-1}]-[z_2^{-1}])\tau(t+[z_1^{-1}])\tau(t+[z_2^{-1}])$$
$$-\tau(t+[z_1^{-1}]+[z_2^{-1}])\tau(t-[z_1^{-1}])\tau(t-[z_2^{-1}])],$$

or

$$\frac{z_2+z_1}{\tau^3(t)}[\tau(t+[z_1^{-1}]-[z_2^{-1}])\tau(t-[z_1^{-1}])\tau(t+[z_2^{-1}])$$
$$-\tau(t-[z_1^{-1}]+[z_2^{-1}])\tau(t+[z_1^{-1}])\tau(t-[z_2^{-1}])].$$

On the other hand, using the relations from Lemmas 2.1–2.3, we obtain from the the second line of (1.4)

$$\mathbf{W}_1\equiv-(z_1^2-z_2^2)^{-1}\partial_x[W[\psi(t,z_1),\psi(t,z_2)]W[\psi^*(t,z_1),\psi^*(t,z_2)]]$$
$$=(z_1-z_2)^2(z_1^2-z_2^2)^{-1}\partial_x\left[\frac{\tau(t-[z_1^{-1}]-[z_2^{-1}])\tau(t+[z_1^{-1}]+[z_2^{-1}])}{\tau^2(t)}\right]-\frac{z_1-z_2}{z_1+z_2}\tau^{-3}(t)$$
$$\times[\tau(t-[z_1^{-1}]-[z_2^{-1}])W(\tau(t+[z_1^{-1}]+[z_2^{-1}]),\tau(t))+\tau(t+[z_1^{-1}]+[z_2^{-1}])$$
$$\times W(\tau(t-[z_1^{-1}]-[z_2^{-1}]),\tau(t))]$$
$$=\frac{z_2-z_1}{z_1+z_2}\tau^{-3}(t)[\tau(t-[z_1^{-1}]-[z_2^{-1}])(-(z_1+z_2)(\tau(t+[z_1^{-1}]+[z_2^{-1}])\tau(t)-\tau(t+[z_1^{-1}])$$
$$\times\tau(t+[z_2^{-1}])))+\tau(t+[z_1^{-1}]+[z_2^{-1}])((z_1+z_2)(\tau(t-[z_1^{-1}]-[z_2^{-1}])\tau(t)-\tau(t-[z_1^{-1}])$$
$$\times\tau(t-[z_2^{-1}])))]$$
$$=\frac{z_2-z_1}{\tau^3(t)}[\tau(t-[z_1^{-1}]-[z_2^{-1}])\tau(t+[z_1^{-1}])\tau(t+[z_2^{-1}])-\tau(t+[z_1^{-1}]+[z_2^{-1}])\tau(t-[z_1^{-1}])$$
$$\times\tau(t-[z_2^{-1}])], \qquad \blacksquare$$

and for the third line of (1.4)

$$\mathbf{W}_2\equiv-(z_1^2-z_2^2)^{-1}\partial_x[W[\psi(t,z_1),\psi^*(t,z_2)]W[\psi^*(t,z_1),\psi^*(t,z_2)]]$$
$$=(z_1+z_2)^2(z_1^2-z_2^2)^{-1}\partial_x\left[\frac{\tau(t-[z_1^{-1}]+[z_2^{-1}])\tau(t+[z_1^{-1}]-[z_2^{-1}])}{\tau^2(t)}\right]$$
$$=-\frac{z_1+z_2}{z_1-z_2}\tau^{-3}(t)[\tau(t-[z_1^{-1}]+[z_2^{-1}])W(\tau(t+[z_1^{-1}]-[z_2^{-1}]),\tau(t))+\tau(t+[z_1^{-1}]$$
$$-[z_2^{-1}])W(\tau(t-[z_1^{-1}]+[z_2^{-1}]),\tau(t))]$$
$$=\frac{z_2+z_1}{z_1-z_2}\tau^{-3}(t)[\tau(t-[z_1^{-1}]+[z_2^{-1}])(-(z_1-z_2)(\tau(t+[z_1^{-1}]-[z_2^{-1}])\tau(t)-\tau(t+[z_1^{-1}])$$
$$\times\tau(t-[z_2^{-1}])))+\tau(t+[z_1^{-1}]-[z_2^{-1}])((z_1-z_2)(\tau(t-[z_1^{-1}]+[z_2^{-1}])\tau(t)-\tau(t-[z_1^{-1}])$$
$$\times\tau(t+[z_2^{-1}])))]$$
$$=\frac{z_2+z_1}{\tau^3(t)}[\tau(t+[z_1^{-1}]-[z_2^{-1}])\tau(t-[z_1^{-1}])\tau(t+[z_2^{-1}])-\tau(t-[z_1^{-1}]+[z_2^{-1}])\tau(t+[z_1^{-1}])$$
$$\times\tau(t-[z_2^{-1}])]. \qquad \blacksquare$$

This way we obtain that **W** equals to $\mathbf{W}_1$ or $\mathbf{W}_2$, i.e., the Faddeev–Tahtajan identity is fulfilled. □

IV. CONCLUSION REMARKS AND EXAMPLES

First we illustrate the identities from Theorem 1.2 by examples with polynomial KdV tau functions. The author thanks F. A. Grünbaum for suggestions that include these examples in the body of the paper.

Example 4.1: The first nontrivial polynomial KdV tau function is $\tau_1(t) := t_1$. In this case the examination of the identities (i) and (ii) of Theorem 1, is easy to do directly and the result is that the both sides of (i) are equal to $2z_1z_2^3 - 2z_1^3z_2$ and the both sides of (ii) are equal to $4z^3$.

Example 4.2: The next polynomial KdV tau function is of degree 3: $\tau_3(t) := t_1^3 - 3t_3$ and as is clear from the results, the examination of the identities (i) and (ii) of Theorem 1, in this case is difficult to do directly. We used the system Maple V Release 4 at RIMS, Kyoto University. So, the both sides of (i) are equal to

$$6(z_1z_2^3 - z_1^3z_2)t_1^6 + 36(z_1^5z_2 - z_1z_2^5)t_1^4 + 126(z_1z_2^3 - z_1^3z_2)t_1^3t_3 + 54(z_1^3z_2^5 - z_1^5z_2^3)t_1^2 + 54(z_1^5z_2 - z_1z_2^5)t_1t_3 + 54(z_1z_2^3 - z_1^3z_2)t_3^2,$$

and the both sides of (ii) are equal to

$$12z^3t_1^6 - 144z^5t_1^4 + 252z^3t_1^3t_3 + 108z^7t_1^2 - 216z^5t_1t_3 + 108z^3t_3^2.$$

There were some problems with fixing the correct form of the KdV tau function $\tau_3(t)$-polynomial of the form $t_1^3 - at_3$. The function $t_1^3 - at_3$ satisfies the Fay identity (1.1) iff $a = 3$.

Remark 4.3: Applying the identities (1.5) to the Wronskian,

$$W(\tau(t+[z_1])\tau(t-[z_1])\tau(t+[z_3])\tau(t-[z_3]), \tau(t+[z_2])\tau(t-[z_2])\tau(t+[z_4])\tau(t-[z_4]))$$

$(z_1, z_2, z_3, z_4 \in \mathbb{C})$, we can obtain eight different (equivalent) expessions where we have Wronskians of two tau functions only (i.e., without any Wronskian of products of tau functions). The expressions are separated in two groups and applying Lemma 2.2 we could see that the resulting identities among the expressions in each group are easily obtained using the result of Theorem 1.2 (i). The equality of the given bellow expressions (from the two groups) is a nontrivial seventh-order (specific for KdV tau functions only) identity:

$$\begin{aligned}
&(z_4^{-1} - z_3^{-1})\tau(t+[z_1])\tau(t-[z_1])\tau(t+[z_2])\tau(t-[z_2])[\tau(t+[z_3])\tau(t+[z_4])\tau(t-[z_3]-[z_4]) \\
&\quad - \tau(t-[z_3])\tau(t-[z_4])\tau(t+[z_3]+[z_4])] + (z_2^{-1} - z_1^{-1})\tau(t+[z_3])\tau(t-[z_3])\tau(t+[z_4])\tau(t \\
&\quad -[z_4])[\tau(t+[z_1])\tau(t+[z_2])\tau(t-[z_1]-[z_2]) - \tau(t-[z_1])\tau(t-[z_2])\tau(t+[z_1]+[z_2])] \\
&\quad = (z_2^{-1} - z_3^{-1})\tau(t+[z_1])\tau(t-[z_1])\tau(t+[z_4])\tau(t-[z_4])[\tau(t+[z_2])\tau(t+[z_3])\tau(t-[z_2] \\
&\qquad -[z_3]) - \tau(t-[z_2])\tau(t-[z_3])\tau(t+[z_2]+[z_3])] \\
&\qquad + (z_4^{-1} - z_1^{-1})\tau(t+[z_2])\tau(t-[z_2])\tau(t+[z_3])\tau(t-[z_3])[\tau(t+[z_1])\tau(t+[z_4])\tau(t-[z_1] \\
&\qquad -[z_4]) - \tau(t-[z_1])\tau(t-[z_4])\tau(t+[z_1]+[z_4])].
\end{aligned}$$

It is clear that this way we can obtain generalized identities of order $2^n - 1$ for any $n = 4,5,\ldots$. The identities from Theorem 1.2 and Remark 4.3 correspond to the cases $n = 2$ and $n = 3$, respectively.

Example 4.4: For the first polynomial tau function $\tau_1(t) = t_1$, both sides of the identity from Remark 4.3 are equal to

$$2(-z_1^2 + z_2^2 - z_3^2 + z_4^2)t_1^4 + 4(z_1^2z_3^2 - z_2^2z_4^2)t_1^2 + 2(-z_1^2z_2^2z_3^2 + z_1^2z_2^2z_4^2 - z_1^2z_3^3z_4^2 + z_2^2z_3^2z_4^2).$$

For the next polynomial KdV tau function $\tau_3(t) = t_1^3 - 3t_3$, both sides of this identity have too many terms (more than 250).

Remark 4.5: As we mentioned in the Introduction, there are some problems to translate the identity (i) from Theorem 1.2 to the cases when KdV tau function is expressed by trigonometric functions or elliptic theta functions. The problems come, roughly speaking, from the fact that in the original Fay identity (i.e., for theta functions related to Jacobians) is used the ''Prime Form'' [e.g., in the $g=1$ case: $\theta_{11}(z_0-z_1)(\theta'_{11}(0))^{-1}$], but in the Fay identity (1.1) for KP tau functions is used the difference (z_0-z_1) instead. Our next task is to fix these problems and to find a ''geometric'' explanation of the identities from the present paper. It will be done in some future article.

Remark 4.6: The ''elliptic version'' of the identity (ii) from Theorem 1.2 is the following relation:

$$\theta'_{11}(0)[\theta_{11}(x+2z)0^2_{11}(x-z)-\theta_{11}(x-2z)\theta^2_{11}(x+z)]$$
$$=\theta_{11}(2z)[\theta_{11}(x-z)W(\theta_{11}(x),\theta_{11}(x+z))+\theta_{11}(x+z)W(\theta_{11}(x),\theta_{11}(x-z))].$$

It is easily obtained from the elliptic version of the identity (i) from Theorem 1.2 (cf. the Introduction) letting $z_1\to z_2$ and denoting $z_1=z_2\equiv z$.

ACKNOWLEDGMENTS

First I would like to express my sincere thanks to my host researcher Professor T. Miwa and to Professor M. Jimbo for making it possible to be at RIMS. I wish to acknowledge the useful discussions with F. A. Grünbaum, M. Jimbo, T. Miwa, M. Ohmiya, E. Previato, T. Shiota, and B. Sturmfels. I am also thankful to all the people at RIMS for warm hospitality. The work was done under JSPS Postdoctoral Fellowship for Foreign Researchers.

[1] D. Mumford, ''Tata lectures on theta I, II,'' Progress in Mathematics Vols. 28 and 43 (Birkhäuser Boston, 1983, 1984).
[2] M. Adler and P. van Moerbeke, ''A matrix integral solution to two-dimensional W_p-gravity,'' Commun. Math. Phys. **147**, 25–56 (1992).
[3] T. Shiota, ''Characterization of Jacobian varieties in terms of soliton equations,'' Invent. Math. **83**(2), 333–382 (1986).
[4] J. Fay, ''Theta functions on Riemann surfaces,'' *Lecture Notes in Mathematics, Vol. 352* (Springer-Verlag, Berlin, 1973).
[5] L. D. Faddeev and L. A. Takhtajan, ''Hamiltonian methods in the theory of solitons,'' *Springer Series in Soviet Mathematics* (Springer-Verlag, Berlin, 1987).
[6] Y. P. Mishev, ''Crum-Kreĭn transforms and Λ-operators for radial Schrödinger equations,'' Inverse Probl. **7**(3), 379–398 (1991).
[7] G. Wilson, ''Habillage et fonctions τ,'' C. R. Acad. Sci., Ser. I: Math. **299**(13), 587–590 (1984).
[8] M. Adler and P. van Moerbeke, ''Birkhoff strata, Bäcklund transformations, and regularization of isospectral operators,'' Adv. Math. **108**(1), 140–204 (1994).
[9] G. Felder and A. Veselov, ''Shift operators for the quantum Calogero-Suthenland problems via Knizhnik-Zamolodchikov equation,'' Commun. Math. Phys. **160**(2), 259–273 (1994).
[10] R. Carroll, ''Some kernels on a Riemann surface,'' Applicable Anal. **65**(3-4), 333–352 (1997).
[11] R. Carroll and J. H. Chang, ''The Whitham equations revisited,'' Applicable Anal. **64**(3-4), 343–378 (1997).
[12] Y. P. Mishev, ''On one cubic identity for KdV tau functions,'' Matematica Balkanika **11**(1–2), 139–142 (1997).
[13] Y. P. Mishev, ''On the Faddeev-Takhtajan identity,'' Matematica Balkanika **11**(3–4), 275–280 (1997).

Differential polynomial expressions related to the Kadomtsev–Petviashvili and Korteweg–de Vries hierarchies

Rainer Schimming
Institut für Mathematik und Informatik, Ernst-Moritz-Arndt-Universität, Friedrich-Ludwig-Jahn-Str. 15a, 17487 Greifswald, Germany

Walter Strampp
Fachbereich 17-Mathematik/Informatik, Universität-GH Kassel, Heinrich-Plett-Str. 40, 34109 Kassel, Germany

(Received 29 October 1998; accepted for publication 20 January 1999)

An integrable nonlinear partial differential equation typically extends to a hierarchy of integrable equations. There exist several recursive schemes for obtaining these hierarchies. Recently, explicit expressions for the KdV hierarchy have been found. We derive explicit expressions for the hierarchy associated with the KP equation. The main tools are Sato's theory, Hirota's formalism, and Bell's polynomials.
 [S0022-2488(99)00805-1]

I. INTRODUCTION

In this paper we derive some new formulas for the Kadomtsev–Petviashvili (KP) and the Korteweg–de Vries (KdV) hierarchies of solitonic partial differential equations. The proofs rely on the properties of the so-called Bell polynomials and their transforms.

The KdV hierarchy[1] consists of a sequence of evolution equations

$$\frac{\partial u}{\partial t_{2n-1}} = K_{2n-1}[u], \quad n \geq 2, \tag{1}$$

the flows of which commute with each other. Here u depends on an infinite set of variables $t_1, t_3, t_5, \ldots$ and $K_n[u]$ is a polynomial in the variable u and its t_1 derivatives up to some higher order. The first equation of the hierarchy, i.e., the KdV equation itself, reads

$$\frac{\partial u}{\partial t_3} = \frac{1}{4} u^{(3)} + 3uu^{(1)}, \tag{2}$$

where we use the notation

$$u^{(k)} = \frac{\partial^k u}{\partial t_1^k}.$$

It is well known that the KdV hierarchy is generated by Lenard's recursion[1]

$$K_{2n+1}[u] = RK_{2n-1}[u],$$

where the recursion operator reads

$$R = \frac{1}{4}\left(\frac{\partial}{\partial t_1}\right)^2 + 2u + u^{(1)}\left(\frac{\partial}{\partial t_1}\right)^{-1}.$$

Hirota[2,3] succeeded in finding bilinear forms for a large number of solitonic equations. Let us recall the definition of Hirota's bilinear operators D_j acting on pairs of functions $\tau(\mathbf{t})$ and $\sigma(\mathbf{t})$:

$$D_j(\tau\cdot\sigma)(\mathbf{t})=\frac{\partial}{\partial z_j}(\tau(\mathbf{t}+\mathbf{z})\sigma(\mathbf{t}-\mathbf{z}))\bigg|_{\mathbf{z}=\mathbf{0}}, \tag{3}$$

where

$$\mathbf{t}=(t_1,t_2,t_3,\ldots)$$

denotes a sequence of real variables t_n and

$$\mathbf{z}=(z_1,z_2,z_3,\ldots)$$

is an auxiliary sequence. More generally, let P denote a polynomial in several variables with constant coefficients. Then Hirota defines

$$P(\mathbf{D})(\tau\cdot\sigma)(\mathbf{t})=P\left(\frac{\partial}{\partial\mathbf{z}}\right)(\tau(\mathbf{t}+\mathbf{z})\sigma(\mathbf{t}\text{-}\mathbf{z}))\bigg|_{\mathbf{z}=\mathbf{0}}, \tag{4}$$

with the symbolic sequences

$$\mathbf{D}=(D_1,D_2,D_3,\ldots)$$

and

$$\frac{\partial}{\partial\mathbf{z}}=\left(\frac{\partial}{\partial z_1},\frac{\partial}{\partial z_2},\frac{\partial}{\partial z_3},\cdots\right).$$

The so-called τ-function is introduced by

$$u=\frac{\partial^2}{\partial t_1^2}\log(\tau).$$

Then the following bilinear expression for the KdV equation holds:

$$(4D_1D_3-D_1^4)(\tau\cdot\tau)=0. \tag{5}$$

The KdV hierarchy can be viewed as a reduction of the more general KP hierarchy[4] which begins with the KP equation itself:

$$\frac{\partial u}{\partial t_3}=\frac{1}{4}u^{(3)}+3uu^{(1)}+\frac{3}{4}\int\frac{\partial^2 u}{\partial t_2^2}dt_1, \tag{6}$$

where u now depends on an infinite set of variables $t_1,t_2,t_3,\ldots$. The Lenard recursion has been extended to some complicated bilocal recursion scheme for the KP equation by Fokas and Santini:[5]

$$\frac{\partial u}{\partial t_n}=K_{n,\mathrm{KP}}[u],\quad n\geqslant 3, \tag{7}$$

where now $K_{n,\mathrm{KP}}[u]$ is a polynomial depending on u, $\int(\partial^2u/\partial t_2^2)dt_1$ and their higher t_1 derivatives up to some order. The higher KP equations generate symmetries for the lower ones.

Sato's theory[4,6] describes the KP hierarchy by means of pseudodifferential operators and a so-called τ-function. This theory allows a compact formulation of this hierarchy in bilinear terms:[7,9]

$$\tfrac{1}{2}D_1 D_n(\tau\cdot\tau)=S_{n+1}(\mathbf{D})(\tau\cdot\tau),\quad n\geqslant 3. \tag{8}$$

The KdV hierarchy emerges from the KP hierarchy by a two-reduction, which means, in particular, that all $t_2,t_4,t_6,\ldots$ dependencies disappear.

The right-hand side of (8) is given by the $(n+1)$-th Schur polynomial S_{n+1}. The sequence of the Schur polynomials $S_n(\mathbf{t})=S_n(t_1,\ldots,t_n)$ is defined through the generating function

$$\exp\left(\sum_{n=1}^{\infty} t_n\frac{h^n}{n}\right)=\sum_{n=0}^{\infty} S_n(t_1,\ldots,t_n)h^n.$$

Let us also introduce a modification $P_n(\mathbf{t})=P_n(t_1,\ldots,t_n)$ of the Schur polynomials which is defined through the generating function

$$\exp\left(\sum_{n=1}^{\infty} t_n h^n\right)=\sum_{n=0}^{\infty} P_n(t_1,\ldots,t_n)h^n.$$

Starting from (8), we have found the following explicit expression of the KP hierarchy:

$$\frac{\partial^2\tilde{u}}{\partial t_1\partial t_n}=\sum_{k=0}^{n} P_k\left(S_1\left(\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}),\cdots,S_k\left(\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u})\right)P_{n-k}\left(S_1\left(-\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}),\cdots,S_{n-k}\left(-\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u})\right),$$

in terms of the function

$$\tilde{u}=\int\int u\,dt_1\,dt_1=\log(\tau)$$

and of the symbolic sequence

$$\frac{\partial}{\partial\mathbf{t}}=\left(\frac{\partial}{\partial t_1},\frac{\partial}{\partial t_2},\frac{\partial}{\partial t_3},\cdots\right).$$

Moreover, in terms of the function u and of two sequences of functions z_n and z_n^*, which are recursively defined by

$$z_n=-\frac{1}{2}\left(\frac{\partial z_{n-1}}{\partial t_1}-\int\frac{\partial z_{n-1}}{\partial t_2}dt_1+\sum_{k=1}^{n-2} z_k z_{n-1-k}\right),\quad z_1=-u,$$

$$z_n^*=-\frac{1}{2}\left(\frac{\partial z_{n-1}^*}{\partial t_1}+\int\frac{\partial z_{n-1}^*}{\partial t_2}dt_1+\sum_{k=1}^{n-2} z_k^* z_{n-1-k}^*\right),\quad z_1^*=u,$$

the KP hierarchy assumes the form:

$$\frac{\partial u}{\partial t_n}=\frac{\partial}{\partial t_1}P_{n+1}\left(\int(z_1+z_1^*)dt_1,\cdots,\int(z_n+z_n^*)dt_1\right),\quad n\geqslant 3.$$

Finally, in the KdV case the analogous explicit expression takes the form

$$\frac{\partial u}{\partial t_{2n-1}}=\frac{\partial}{\partial t_1}P_n\left(2\int z_2\,dt_1,\cdots,2\int z_{2n}\,dt_1\right),\quad n\geqslant 3,$$

where now

$$z_n = -\frac{1}{2}\left(\frac{\partial z_{n-1}}{\partial t_1} + \sum_{k=1}^{n-2} z_k z_{n-1-k}\right), \quad z_1 = -u.$$

II. THE BELL POLYNOMIALS

In this section, we introduce the Bell polynomials together with some of their transforms and list several properties. In particular, we derive a new determinant representation.

The sequence of the Bell polynomials $B_n(\mathbf{t}) = B_n(t_1,\ldots,t_n)$, $n=1,2,3,...$, in real variables $t_1,t_2,t_3,\ldots$ is defined by the bilinear recursion[10,11]

$$B_{n+1}(\mathbf{t}) = \sum_{k=0}^{n} \binom{n}{k} B_k(\mathbf{t}) t_{n+1-k}, \quad B_0(\mathbf{t}) = 1. \tag{9}$$

Equivalently, the Bell polynomials can be introduced through the generating function:[12]

$$\exp\left(\sum_{n=1}^{\infty} t_n \frac{h^n}{n!}\right) = \sum_{n=0}^{\infty} B_n(\mathbf{t}) \frac{h^n}{n!}. \tag{10}$$

The first few Bell polynomials read as follows:

$$B_0(\mathbf{t}) = 1,$$

$$B_1(\mathbf{t}) = t_1,$$

$$B_2(\mathbf{t}) = t_1^2 + t_2,$$

$$B_3(\mathbf{t}) = t_1^3 + 3t_1t_2 + t_3,$$

$$B_4(\mathbf{t}) = t_1^4 + 6t_1^2t_2 + 3t_2^2 + 4t_1t_3 + t_4,$$

$$B_5(\mathbf{t}) = t_1^5 + 10t_1^3t_2 + 15t_1t_2^2 + 10t_1^2t_3 + 10t_2t_3 + 5t_1t_4 + t_5.$$

The following explicit formula goes back to Faá di Bruno:[13]

$$B_n(\mathbf{t}) = \sum_{\|\mathbf{m}\|=n} \frac{n!}{\mathbf{m}!}\left(\frac{t_1}{1!}\right)^{m_1}\left(\frac{t_2}{2!}\right)^{m_2}\cdots\left(\frac{t_n}{n!}\right)^{m_n}, \tag{11}$$

where we use some multiindex notation for non-negative integers $m_1,\ldots,m_n$:

$$\mathbf{m} = (m_1,m_2,\ldots,m_n), \quad \mathbf{m}! = m_1!m_2!\cdots m_n!,$$

$$\|\mathbf{m}\| = 1\cdot m_1 + 2\cdot m_2 + \cdots + n\cdot m_n.$$

We shall also consider some close relatives of the Bell polynomials, namely, the Schur polynomials $S_n(\mathbf{t}) = S_n(t_1,\ldots,t_n)$ defined by

$$\exp\left(\sum_{n=1}^{\infty} t_n \frac{h^n}{n}\right) = \sum_{n=0}^{\infty} S_n(t_1,\ldots,t_n) h^n \tag{12}$$

and the modified Schur polynomials $P_n(\mathbf{t}) = P_n(t_1,\ldots,t_n)$ defined by

$$\exp\left(\sum_{n=1}^{\infty} t_n h^n\right) = \sum_{n=0}^{\infty} P_n(t_1,\ldots,t_n)h^n. \tag{13}$$

Both these are closely related to the Bell polynomials, namely,

$$S_n(t_1,\ldots,t_n) = \frac{1}{n!} B_n(0!t_1,\ldots,(n-1)!t_n), \tag{14}$$

$$P_n(t_1,\ldots,t_n) = \frac{1}{n!} B_n(1!t_1,\ldots,n!t_n). \tag{15}$$

Let us further consider the ''thinned Bell polynomials'' where either the even variables $t_2,t_4,\ldots$ or the odd variables $t_1,t_3,\ldots$ are set to zero.

Proposition 1: For $n \geqslant 0$ there holds

$$B_{2n}(-t_1,0,-t_3,0,\ldots,-t_{2n-1},0) = B_{2n}(t_1,0,t_3,0,\ldots,t_{2n-1},0),$$

$$B_{2n+1}(-t_1,0,-t_3,0,\ldots,-t_{2n+1}) = -B_{2n+1}(t_1,0,t_3,0,\ldots,t_{2n+1}),$$

$$B_{2n+1}(0,t_2,0,t_4,\ldots,0) = 0,$$

$$B_{2n}(0,t_2,0,t_4,\ldots t_{2n}) = \frac{(2n)!}{n!} B_n\left(\frac{1!}{2!}t_2, \frac{2!}{4!}t_4,, \cdots, \frac{n!}{(2n)!}t_{2n}\right).$$

Proof: From

$$\exp\left(\sum_{n=1}^{\infty} (-t_{2n-1}) \frac{h^{2n-1}}{(2n-1)!}\right) = \exp\left(\sum_{n=1}^{\infty} t_{2n-1} \frac{(-h)^{2n-1}}{(2n-1)!}\right)$$

we obtain immediately

$$\sum_{n=0}^{\infty} B_n(-t_1,0,-t_3,0,\cdots,) \frac{h^n}{n!} = \sum_{n=0}^{\infty} B_n(t_1,0,t_3,0,\cdots,) \frac{(-h)^n}{n!}.$$

Further, from the generating function we have on the one hand,

$$\exp\left(\sum_{n=1}^{\infty} t_{2n} \frac{h^{2n}}{(2n)!}\right) = \sum_{n=0}^{\infty} B_n(0,t_2,0,t_4,\cdots) \frac{h^n}{n!},$$

and, on the other hand,

$$\begin{aligned}\exp\left(\sum_{n=1}^{\infty} t_{2n} \frac{h^{2n}}{(2n)!}\right) &= \exp\left(\sum_{n=1}^{\infty} t_{2n} \frac{n!}{(2n)!} \frac{(h^2)^n}{n!}\right)\\ &= \sum_{n=0}^{\infty} B_n\left(\frac{1!}{2!}t_2, \frac{2!}{4!}t_4, \cdots, \frac{n!}{(2n)!}t_{2n}\right) \frac{(h^2)^n}{n!}\\ &= \sum_{n=0}^{\infty} \frac{(2n)!}{n!} B_n\left(\frac{1!}{2!}t_2, \frac{2!}{4!}t_4, \cdots, \frac{n!}{(2n)!}t_{2n}\right) \frac{h^{2n}}{(2n)!}.\end{aligned}$$

The results follow.

Let us also introduce the inverse Bell polynomials $B_n^-(\mathbf{t}) = B_n^-(t_1,\ldots,t_n)$ by the bilinear recursion

$$B_{n+1}^{-}(\mathbf{t})=t_{n+1}-\sum_{k=1}^{n}\binom{n}{k-1}B_{k}^{-}(\mathbf{t})t_{n+1-k}, \quad B_0^{-}(\mathbf{t})=1. \tag{16}$$

They naturally appear if the system (9) is inverted, i.e., solved for the variables $t_1,t_2,t_3,\ldots$. Namely, if we replace in (9) t_n by B_n^- and vice versa we get

$$t_{n+1}=\sum_{k=0}^{n}\binom{n}{k}t_kB_{n+1-k}^{-}(\mathbf{t}),$$

which is nothing but (16). The same replacement transforms (10) into

$$\exp\left(\sum_{n=1}^{\infty}B_n^{-}\frac{h^n}{n!}\right)=\sum_{n=0}^{\infty}t_n\frac{h^n}{n!}, \quad t_0=1, \tag{17}$$

becoming equivalent to

$$\log\left(1+\sum_{n=1}^{\infty}t_n\frac{h^n}{n!}\right)=\sum_{n=1}^{\infty}B_n^{-}(\mathbf{t})\frac{h^n}{n!}. \tag{18}$$

Thus we have found the generating function of the inverse Bell polynomials $B_n^{-}(t_1,\ldots,t_n)$. The first few of them read as follows:

$$B_1^{-}(\mathbf{t})=t_1,$$

$$B_2^{-}(\mathbf{t})=-t_1^2+t_2,$$

$$B_3^{-}(\mathbf{t})=2t_1^3-3t_1t_2+t_3,$$

$$B_4^{-}(\mathbf{t})=-6t_1^4+12t_1^2t_2-3t_2^2-4t_1t_3+t_4,$$

$$B_5^{-}(\mathbf{t})=24t_1^5-60t_1^3t_2+30t_1t_2^2+20t_1^2t_3-10t_2t_3-5t_1t_4+l_5.$$

Finally, we derive a representation of the Bell polynomials in the form of a determinant:

Proposition 2: There holds for $n\geqslant 2$:

$$B_n=\begin{vmatrix} \binom{0}{0}t_1 & -1 & 0 & \cdots & 0 \\ \binom{1}{0}t_2 & \binom{1}{1}t_1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \binom{n-2}{0}t_{n-1} & \binom{n-2}{1}t_{n-2} & \binom{n-2}{2}t_{n-3} & \cdots & -1 \\ \binom{n-1}{0}t_n & \binom{n-1}{1}t_{n-1} & \binom{n-1}{2}t_{n-2} & \cdots & \binom{n-1}{n-1}t_1 \end{vmatrix}.$$

Proof: We shall show that the determinant $|M_n|$ of the matrix M_n given by the right-hand side of the above identity satisfies the bilinear recursion (9). To this end we expand the determinant $|M_n|$ with respect to the last row. The element $\binom{n-1}{k}t_{n-k}$ is multiplied by its algebraic complement

$$(-1)^{n-k-1}\begin{vmatrix} M_k & O_k \\ H_k & T_k \end{vmatrix} = (-1)^{n-k-1}|M_k||T_k| = (-1)^{n-k-1}|M_k|(-1)^{n-k-1} = |M_k|.$$

Here O_k stands for the $k\times(n-k-1)$ zero matrix and H_k is some $(n-k-1)\times k$ matrix of less importance. T_k is a $(n-k-1)\times(n-k-1)$ lower triangular matrix with diagonal elements -1, hence $|T_k|=(-1)^{n-k}$. Furthermore, we set $|M_0|=1$, $|M_1|=t_1$ and $|T_0|=1$. All together, we obtain

$$|M_n|\sum_{k=0}^{n-1}\binom{n-1}{k}|M_k|t_{n-k}, \quad |M_0|=1,$$

and the proof is completed.

Notice that instead of the variables $\mathbf{t}=(t_1,t_2,t_3,\ldots)$, any other commutative variables can be inserted into the Bell polynomials. For instance, inserting the differentiations

$$\frac{\partial}{\partial \mathbf{t}} = \left(\frac{\partial}{\partial t_1},\frac{\partial}{\partial t_2},\frac{\partial}{\partial t_3},\cdots\right),$$

we obtain from the Faá di Bruno formula (11):

$$B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t})) = \sum_{\|\mathbf{m}\|=n}^{\infty}\frac{n!}{\mathbf{m}!}\left(\frac{1}{1!}\frac{\partial}{\partial t_1}\right)^{m_1}\left(\frac{2}{2!}\frac{\partial}{\partial t_2}\right)^{m_2}\cdots\left(\frac{1}{n!}\frac{\partial}{\partial t_n}\right)^{m_n}(\tau(\mathbf{t})). \tag{19}$$

This enables us to handle the formal Taylor expansion:

$$\begin{aligned}\tau(\mathbf{t}+\mathbf{h}) &= \sum_{n=0}^{\infty}\frac{1}{n!}\left(\mathbf{h}\cdot\frac{\partial}{\partial \mathbf{t}}\right)^n(\tau(\mathbf{t}))\\ &= \exp\left(\mathbf{h}\cdot\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))\\ &= \sum_{n=0}^{\infty}B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))\frac{h^n}{n!}\\ &= \sum_{n=0}^{\infty}S_n\left(\frac{1}{0!}\frac{\partial}{\partial t_1},\cdots,\frac{1}{(n-1)!}\frac{\partial}{\partial t_n}\right)(\tau(\mathbf{t}))h^n,\end{aligned}$$

where $\mathbf{h}$ is the sequence

$$\mathbf{h} = \left(\frac{h}{1!},\frac{h^2}{2!},\frac{h^3}{3!},\cdots\right)$$

and

$$\mathbf{h}\cdot\frac{\partial}{\partial \mathbf{t}} = \frac{h}{1!}\frac{\partial}{\partial t_1} + \frac{h^2}{2!}\frac{\partial}{\partial t_2} + \frac{h^3}{3!}\frac{\partial}{\partial t_3} + \cdots.$$

For later convenience we derive the following formula:

Proposition 3: There holds for $n\geqslant 0$,

$$\exp(-\tau(\mathbf{t}))B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\exp(\tau(\mathbf{t}))) = B_n\left(B_1\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t})),\cdots,B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))\right).$$

Proof: Through Taylor expansion we obtain on the one hand

$$\exp(\tau(\mathbf{t}+\mathbf{h}))=\sum_{n=0}^{\infty} B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\exp(\tau(\mathbf{t})))\frac{h^n}{n!},$$

and on the other hand

$$\exp(\tau(\mathbf{t}+\mathbf{h}))=\exp\left(\tau(\mathbf{t})+\sum_{n=1}^{\infty} B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))\frac{h^n}{n!}\right)$$

$$=\exp(\tau(\mathbf{t}))\sum_{n=0}^{\infty} B_n\left(B_1\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t})),\cdots,B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))\right)\frac{h^n}{n!}.$$

Comparison of the coefficients of $h^n/n!$ gives the result.

Next we shall produce some Hirota-type Leibniz rule.

Proposition 4: There holds for $n\geqslant 1$,

$$B_n(\mathbf{D})(\tau\cdot\sigma)(\mathbf{t})=\sum_{k=0}^{n}\binom{n}{k}B_k\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))B_{n-k}\left(-\frac{\partial}{\partial \mathbf{t}}\right)(\sigma(\mathbf{t})).$$

Proof: We obtain by Taylor expansion

$$\tau(\mathbf{t}+(\mathbf{z}+\mathbf{h}))\sigma(\mathbf{t}-(\mathbf{z}+\mathbf{h}))=\sum_{n=0}^{\infty} B_n\left(\frac{\partial}{\partial z_1},\frac{\partial}{\partial z_2},\frac{\partial}{\partial z_3},\cdots\right)(\tau(\mathbf{t}+\mathbf{z})\sigma(\mathbf{t}-\mathbf{z}))\frac{h^n}{n!}.$$

Here we set $\mathbf{z}=\mathbf{0}$:

$$\tau(\mathbf{t}+\mathbf{h})\sigma(\mathbf{t}-\mathbf{h})=\sum_{n=0}^{\infty} B_n(\mathbf{D})(\tau\cdot\sigma)(\mathbf{t})\frac{h^n}{n!}.$$

Multiplication of the series expansions of τ and σ gives, alternatively,

$$\tau(\mathbf{t}+\mathbf{h})\sigma(\mathbf{t}-\mathbf{h})=\sum_{n=0}^{\infty} B_n\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))\frac{h^n}{n!}\sum_{n=0}^{\infty} B_n\left(-\frac{\partial}{\partial \mathbf{t}}\right)(\sigma(\mathbf{t}))\frac{h^n}{n!}$$

$$=\sum_{n=0}^{\infty}\sum_{k=0}^{n} B_k\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))B_{n-k}\left(-\frac{\partial}{\partial \mathbf{t}}\right)(\sigma(\mathbf{t}))\frac{h^n}{k!(n-k)!}$$

$$=\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}\binom{n}{k}B_k\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tau(\mathbf{t}))B_{n-k}\left(-\frac{\partial}{\partial \mathbf{t}}\right)(\sigma(\mathbf{t}))\right)\frac{h^n}{n!}.$$

Comparison of the coefficients of $h^n/n!$ gives the result.

III. HIROTA'S FORMALISM AND THE KdV EQUATION

Let us make explicit some expressions which appear in Hirota's formalism.[2,3] By using

$$\tau(\mathbf{t})=\exp\left(\int\int u(\mathbf{t})\,dt_1\,dt_1\right)$$

we now will express the bilinear quantities $D_1^n(\tau\cdot\tau)(\mathbf{t})$ as differential polynomials in $u(\mathbf{t})=(\partial^2/\partial t_1^2)\log(\tau(\mathbf{t}))$. Here $\int u(\mathbf{t})\,dt_1$ denotes any primitive of $u(\mathbf{t})$ and $\int\int u(\mathbf{t})\,dt_1\,dt_1$ any two-fold primitive.

From the observation $B_n(l_1,0,...,0)=t_1^n$ we obtain the special formulas:

$$B_n\left(\frac{\partial}{\partial t_1},0,\cdots,0\right)(\tau(t_1))=\left(\frac{d}{dt_1}\right)^n(\tau(t_1)),$$

$$B_n\left(-\frac{\partial}{\partial t_1},0,\cdots,0\right)(\tau(t_1))=(-1)^n\left(\frac{d}{dt_1}\right)^n(\tau(t_1)),$$

$$B_n(D_1,0,...,0)(\tau\cdot\tau)(t_1)=D_1^n(\tau\cdot\tau)(t_1).$$

Notice that generally

$$D_1^{2n+1}(\tau\cdot\tau)(\mathbf{t})=0,\quad n\geqslant 0.$$

The even expressions $D_1^{2n}(\tau\cdot\tau)(\mathbf{t})$ do not vanish and we just now calculate these.

Proposition 5: There holds for $n\geqslant 1$,

$$D_1^{2n}(\tau\cdot\tau)=\frac{(2n)!}{n!}(\tau)^2 B_n\left(2\frac{1!}{2!}u,2\frac{2!}{4!}u^{(2)},\cdots,2\frac{n!}{(2n)!}u^{(2n-2)}\right),$$

where

$$u=\frac{\partial^2}{\partial t_1^2}\log(\tau),\quad u^{(k)}=\left(\frac{\partial}{\partial t_1}\right)^k u.$$

Proof: We can confine ourselves to functions depending only on t_1. Then Taylor expansion gives

$$\tau(t_1+h)\tau(t_1-h)=\sum_{n=0}^{\infty}D_1^n(\tau\cdot\tau)(t_1)\frac{h^n}{n!}=\sum_{n=0}^{\infty}D_1^{2n}(\tau\cdot\tau)(t_1)\frac{h^{2n}}{(2n)!}.$$

By insertion of

$$\tau(t_1)=\exp\left(\int\int u(t_1)\,dt_1\,dt_1\right)=\exp(\tilde{u}(t_1))$$

and again by Taylor expansion we obtain

$$\begin{aligned}\tau(t_1+h)\tau(t_1-h)&=\exp(\tilde{u}(t_1+h)+\tilde{u}(t_1-h))\\&=\exp\left(\sum_{n=0}^{\infty}2\tilde{u}^{(2n)}(t_1)\frac{h^{2n}}{(2n)!}\right)\\&=\tau(t_1)^2\exp\left(\sum_{n=1}^{\infty}2\tilde{u}^{(2n)}(t_1)\frac{h^{2n}}{(2n)!}\right)\\&=\tau(t_1)^2\sum_{n=0}^{\infty}B_n(0,2\tilde{u}^{(2)}(t_1),0,2\tilde{u}^{(4)}(t_1),\ldots 0,2\tilde{u}^{(2n)}(t_1))\frac{h^n}{n!}.\end{aligned}$$

The result follows.

As a consequence from Proposition 5 we obtain, by virtue of the inverse Bell polynomials B_n^-, the following propesition.

Proposition 6: There holds for $n\geqslant 1$,

$$u^{(2n-2)}=\frac{(2n)!}{2n!}B_n^{-}\left(\frac{1!}{(2\cdot 1)!}\frac{1}{\tau^2}D^{2\cdot 1}(\tau\cdot\tau),\cdots,\frac{n!}{(2n)!}\frac{1}{\tau^2}D^{2n}(\tau\cdot\tau)\right),$$

where B_n^{-} denotes the nth inverse Bell polynomial.

As a further consequence we obtain the following.

Proposition 7: There holds for $n\geqslant 1$,

$$\frac{n!}{(2n)!}\frac{1}{\tau^2}D_1^{2n}(\tau\cdot\tau)=\begin{vmatrix} \binom{1}{0}2u & -1 & \cdots & 0 \\ \binom{3}{0}2u^{(2)} & \binom{3}{2}2u & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \binom{2n-2}{0}2u^{(2(n-1)-2)} & \binom{2n-2}{2}u^{(2(n-2)-2)} & \cdots & -1 \\ \binom{2n-1}{0}2u^{(2n-2)} & \binom{2n-1}{2}2u^{(2(n-1)-2)} & \cdots & \binom{2n-1}{2n-2}2u \end{vmatrix}.$$

Proof: The proof relies on Proposition 6 and uses determinant rules and proceeds similar to the proof of Proposition 2.

A natural continuation of the KdV equation would be given by the sequence

$$\frac{\partial u}{\partial t_{2n-1}}=\frac{1}{2}\left(\frac{1}{4}\right)^{n-1}\frac{(2n)!}{n!}\frac{\partial}{\partial t_1}B_n\left(2\frac{1!}{2!}u,2\frac{2!}{4!}u^{(2)},\cdots,2\frac{n!}{(2n)!}u^{(2n-2)}\right). \tag{20}$$

It is well known that the stationary KdV equation

$$\tfrac{1}{4}u^{(2)}+\tfrac{3}{2}u^2=c_3$$

is an ordinary differential equation (ODE) of the Painleve type. The question arises whether or not the following ODE's,

$$B_n\left(2\frac{1!}{2!}u,2\frac{2!}{4!}u^{(2)},\cdots,2\frac{n!}{(2n)!}u^{(2n-2)}\right)=c_{2n-1},$$

represent nontrivial integrable equations of the Painleve type.

IV. THE KP AND KdV HIGHER-ORDER EQUATIONS

Let us briefly recall some facts from Sato's theory.[4,6] With the pseudodifferential operator

$$L=\partial_1+u_2\partial_1^{-1}+u_3\partial_1^{-2}+\cdots,\quad \frac{\partial}{\partial t_1},$$

its adjoint

$$L^*=-\partial_1-\partial_1^{-1}u_2-\partial_1^{-2}u_3+\cdots,$$

and a real eigenvalue parameter $h\neq 0$ we build up the linear problems

$$L\psi=h\psi,\quad \frac{\partial\psi}{\partial t_n}=(L^n)_+\psi, \tag{21}$$

$$L^*\psi^* = h\psi, \quad \frac{\partial \psi^*}{\partial t_n} = -(L^n)_+^*\psi^*, \tag{22}$$

where the differential operator $(L^n)_+$ is defined as the non-negative part of the pseudodifferential operator L^n. The first three operators $(L^n)_+$, $(L^n)_+^*$ read as

$$(L^1)_+ = \partial_1,$$

$$(L^2)_+ = \partial_1^2 + 2u_2,$$

$$(L^3)_+ = \partial_1^3 + 3u_2\partial_1 + 3u_3 + 3\frac{\partial u_2}{\partial t_1},$$

$$(L^1)_+^* = -\partial_1,$$

$$(L^2)_+^* = \partial_1^2 + 2u_2,$$

$$(L^3)_+^* = -\partial_1^3 - 3u_2\partial_1 + 3u_3.$$

The linear problems (21) and (22) admit formal solutions, called wave functions,[4,6]

$$\psi(\mathbf{t}) = \exp\left(\sum_{n=1}^{\infty} t_n h^n\right)\frac{\tau(\mathbf{t}-\mathbf{h}^{-1})}{\tau(\mathbf{t})}, \tag{23}$$

$$\psi^*(\mathbf{t}) = \exp\left(-\sum_{n=1}^{\infty} t_n h^n\right)\frac{\tau(\mathbf{t}+\mathbf{h}^{-1})}{\tau(\mathbf{t})}, \tag{24}$$

where

$$\mathbf{h}^{-1} = \left(\frac{h^{-1}}{1}, \frac{h^{-2}}{2}, \frac{h^{-3}}{3}, \cdots\right).$$

We obtain by means of Taylor expansion

$$\log\left(\frac{\tau(\mathbf{t}\mp\mathbf{h}^{-1})}{\tau(\mathbf{t})}\right) = \log(\tau(\mathbf{t}\mp\mathbf{h}^{-1})) - \log(\tau(\mathbf{t}))$$

$$= \sum_{n=0}^{\infty} S_n\left(\mp\frac{\partial}{\partial t}\right)(\log(\tau(\mathbf{t})))h^{-n} - \log(\tau(t))$$

$$= \sum_{n=1}^{\infty} S_n\left(\mp\frac{\partial}{\partial \mathbf{t}}\right)(\log(\tau(\mathbf{t})))h^{-n}.$$

This yields

$$\frac{\tau(\mathbf{t}\mp\mathbf{h}^{-1})}{\tau(\mathbf{t})} = \exp\left(\sum_{n=1}^{\infty} S_n\left(\mp\frac{\partial}{\partial t}\right)(\log(\tau(\mathbf{t})))h^{-n}\right)$$

$$= \exp\left(\sum_{n=1}^{\infty} \int S_n\left(\mp\frac{\partial}{\partial \mathbf{t}}\right)\left(\frac{\partial}{\partial t_1}\log(\tau(\mathbf{t}))\right)h^{-n}\right)$$

and gives the following expression for the wave functions (23) and (24):

$$\psi(\mathbf{t}) = \exp\left(\sum_{n=1}^{\infty} t_n h^n\right) \exp\left(\sum_{n=1}^{\infty} \int S_n\left(-\frac{\partial}{\partial \mathbf{t}}\right)\left(\frac{(\partial\tau/\partial t_1)(\mathbf{t})}{\tau(\mathbf{t})}\right) dt_1 h^{-n}\right)$$

$$= \exp\left(\sum_{n=1}^{\infty} t_n h^n\right) \exp\left(\sum_{n=1}^{\infty} \int z_n(\mathbf{t})\, dt_1 h^{-n}\right), \tag{25}$$

$$\psi^*(\mathbf{t}) = \exp\left(-\sum_{n=1}^{\infty} t_n h^n\right) \exp\left(\sum_{n=1}^{\infty} \int S_n\left(\frac{\partial}{\partial \mathbf{t}}\right)\left(\frac{(\partial\tau/\partial t_1)(\mathbf{t})}{\tau(\mathbf{t})}\right) dt_1 h^{-n}\right)$$

$$= \exp\left(-\sum_{n=1}^{\infty} t_n h^n\right) \exp\left(\sum_{n=1}^{\infty} \int z_n^*(\mathbf{t})\, dt_1 h^{-n}\right), \tag{26}$$

where

$$z_n = S_n\left(-\frac{\partial}{\partial \mathbf{t}}\right)\left(\frac{(\partial\tau/\partial t_1)(\mathbf{t})}{\tau(\mathbf{t})}\right), \quad z_n^* = S_n\left(\frac{\partial}{\partial \mathbf{t}}\right)\left(\frac{(\partial\tau/\partial t_1)(\mathbf{t})}{\tau(\mathbf{t})}\right).$$

In Sato's theory, the KP hierarchy arises from the Lax equations

$$\frac{\partial}{\partial t_n} L = [(L^n)_+, L)],$$

and the so-called τ-function serves as a potential for the function u_n, in particular

$$u = u_2 = \frac{\partial^2}{\partial t_1^2} \log(\tau).$$

Based on the bilinear form (8) we shall now find explicit representations of the polynomials $K_{n,\mathrm{KP}}[u]$ in (7).

Proposition 8: The KP hierarchy can be written as

$$\frac{\partial^2 \tilde{u}}{\partial t_1 \partial t_n} = \sum_{k=0}^{n} P_k\left(S_1\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tilde{u}), \ldots, S_k\left(\frac{\partial}{\partial \mathbf{t}}\right)(\tilde{u})\right) P_{n-k}\left(S_1\left(-\frac{\partial}{\partial \mathbf{t}}\right)(\tilde{u}), \ldots, S_{n-k}\left(-\frac{\partial}{\partial \mathbf{t}}\right)(\tilde{u})\right),$$

where S_n and P_n denote the Schur polynomials and the modified Schur polynomials, respectively. Moreover, there holds

$$\frac{\partial u}{\partial t_n} = \frac{\partial}{\partial t_1} P_{n+1}\left(\int (z_1 + z_1^*)\, dt_1, \cdots, \int (z_n + z_n^*)\, dt_1\right), \quad n \geqslant 3.$$

Proof: We use again

$$\tau(\mathbf{t}) = \exp\left(\int\int u(\mathbf{t})\, dt_1\, dt_1\right) = \exp(\tilde{u}(\mathbf{t})).$$

Then by straightforward calculation we obtain the left-hand side of the bilinear KP equations (8):

$$\tfrac{1}{2}D_1D_n(\tau\cdot\tau)=\tfrac{1}{2}D_1D_n(\exp(\tilde{u})\cdot\exp(\tilde{u}))$$
$$=\exp(2\tilde{u})\int\frac{\partial u}{\partial t_n}\,dt_1.$$

Therefore, the KP hierarchy in bilinear form reads as:

$$\exp(2\tilde{u})\frac{\partial^2\tilde{u}}{\partial t_1\partial t_n}=S_n(\mathbf{D})(\exp(\tilde{u})\cdot\exp(\tilde{u})).$$

From Propositions 3 and 4 the first assertion follows immediately.

The right-hand side of the bilinear KP can be transformed by Taylor expansion similar to the proof of Proposition 3. There holds on the one hand

$$\exp(\tilde{u}(\mathbf{t}+\mathbf{h}^{-1})+\tilde{u}(\mathbf{t}-\mathbf{h}^{-1}))=\exp(2\tilde{u}(\mathbf{t}))\exp\left(\sum_{n=1}^{\infty}\left(S_n\left(\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}(\mathbf{t}))+S_n\left(-\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}(\mathbf{t}))\right)h^{-n}\right)$$
$$=\exp(2\tilde{u}(\mathbf{t}))$$
$$\sum_{n=0}^{\infty}P_n\left(S_1\left(\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}(\mathbf{t}))+S_1\left(-\frac{\partial}{\partial\mathbf{t}}\right)(\tau(\mathbf{t})),\cdots,\right.$$
$$\left.S_n\left(\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}(\mathbf{t}))+S_n\left(-\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}(\mathbf{t}))\right)h^{-n},$$

while we obtain on the other hand

$$\exp(\tilde{u}(\mathbf{t}+\mathbf{h}^{-1})+\tilde{u}(\mathbf{t}-\mathbf{h}^{-1}))=\sum_{n=0}^{\infty}S_n(\mathbf{D})(\exp(\tilde{u}(\mathbf{t}))\cdot\exp(\tilde{u}(\mathbf{t}))h^{-n}).$$

Comparing both expressions gives

$$S_n(\mathbf{D})(\exp(\tilde{u})\cdot\exp(\tilde{u}))$$
$$=\exp(2\tilde{u})P_n\left(S_1\left(\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u})+S_1\left(-\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u}),\cdots,S_n\left(\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u})+S_n\left(-\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u})\right).$$

Now the definition of the quantities z_n and z_n^*, together with

$$S_n\left(\pm\frac{\partial}{\partial\mathbf{t}}\right)(\tilde{u})=S_n\left(\pm\frac{\partial}{\partial\mathbf{t}}\right)\left(\int\int u\,dt_1\,dt_1\right)$$
$$=\int S_n\left(\pm\frac{\partial}{\partial\mathbf{t}}\right)\left(\int u\,dt_1\right)dt_1$$
$$=\int S_n\left(\pm\frac{\partial}{\partial\mathbf{t}}\right)\left(\frac{\partial}{\partial t_1}\log(\tau(\mathbf{t}))\right)dt_1,$$

completes the proof.

Insertion of the wave functions (25) and (26) into

$$\frac{\partial\psi}{\partial t_2}=\frac{\partial^2\psi}{\partial t_1^2}+2u\psi,\quad \frac{\partial\psi^*}{\partial t_n}=-\frac{\partial^2\psi}{\partial t_1^2}-2u\psi^*$$

produces the following recursion formulas:

$$z_n = -\frac{1}{2}\left(\frac{\partial z_{n-1}}{\partial t_1} - \int \frac{\partial z_{n-1}}{\partial t_2}\,dt_1 + \sum_{k=1}^{n-2} z_k z_{n-1-k}\right), \quad z_1 = -u, \tag{27}$$

$$z_n^* = -\frac{1}{2}\left(\frac{\partial z_{n-1}^*}{\partial t_1} + \int \frac{\partial z_{n-1}^*}{\partial t_2}\,dt_1 + \sum_{k=1}^{n-2} z_k^* z_{n-1-k}^*\right), \quad z_1^* = u. \tag{28}$$

These define z_n and z_n^* as polynomials in the variables u, $\int(\partial u/\partial t_2)\,dt_1$ and their higher t_1-derivatives.

Let us finally discuss the KdV case. Here the τ-function and thus u depend only on $t_1, t_3, t_5, \ldots$. Furthermore, the linear problem $\partial\psi/\partial t_2 = B_2\psi$ reads as

$$h^2\psi = \frac{\partial^2\psi}{\partial t_1^2} + 2u\psi,$$

and its adjoint version takes the same form,

$$h^2\psi^* = \frac{\partial^2\psi}{\partial t_1^2} + 2u\psi^*,$$

while the recursion for the z_n becomes

$$z_n = -\frac{1}{2}\left(\frac{\partial z_{n-1}}{\partial t_1} + \sum_{k=1}^{n-2} z_k z_{n-1-k}\right), \quad z_1 = -u. \tag{29}$$

From properties of the B_n (cf. Proposition 1) or equivalently of the P_n and from

$$z_n = S_n\left(-\frac{\partial}{\partial t_1}, 0, -\frac{\partial}{\partial t_3}, 0, \cdots\right)\left(\frac{(\partial\tau/\partial t_1)(\mathbf{t})}{\tau}\right),$$

$$z_n^* = S_n\left(\frac{\partial}{\partial t_1}, 0, \frac{\partial}{\partial t_3}, 0, \cdots\right)\left(\frac{(\partial\tau/\partial t_1)(\mathbf{t})}{\tau}\right),$$

we get

$$z_{2n+1}^* = -z_{2n+1}, \quad z_{2n}^* = z_{2n}.$$

Therefore the KdV equations become for $n \geqslant 1$

$$\frac{\partial u}{\partial t_{2n-1}} = \frac{\partial}{\partial t_1} P_{2n}\left(0, 2\int z_2, \cdots, 0, 2\int z_{2n}\right)$$
$$= \frac{\partial}{\partial t_1} P_n\left(2\int z_2, \ldots, 2\int z_{2n}\right). \tag{30}$$

V. DISCUSSION

The partial differential equations of solitonic type appeared, historically, as physical models. These soliton equations became then seed equations to solitonic hierarchies, i.e., sequencies of PDEs of evolution type with increasing order. The higher-order equations can be interpreted as infinitesimal symmetries to the seed equations, i.e., their flows are one-parameter symmetries to the latter. The most prominent equations or hierarchies are the Korteweg–de Vries (KdV) and the Kadomtsev–Petviashvili (KP) ones. We have studied these here by means of Hirota's bilinear formalism and of pseudodifferential operators of Sato–Gelfand–Dickey type. In Sato's theory the

τ-function is governed by a set of equations arising from the condition $res_\lambda \psi(\mathbf{t})\psi*(\mathbf{t}')=0$ which is just the KP hierarchy. In Ref. 7 this hierarchy has been related to the hierarchy of symmetries of the KP equation studied in Ref. 4 through formula (8). In Refs. 8 and 9 constraints of those hierarchies and their trilinear forms have been considered. In the present paper the representation of the KP hierarchy (Proposition 8) and the KdV hierarchy [Eq. (30)] in terms of the Bell polynomials and their transforms are the main results. These representations are explicit (nonrecursive) expressions since several explicit formulas for the Bell polynomials are known. To these we add some new determinant formulas.

The Bell polynomials and their transforms have been, since the days of Faá di Bruno,[13] studied from different points of view. Recently, one of the authors (RS *et al.*) studied the noncommutative Bell polynomials, i.e., the generalization from real variables $t_1, t_2, t_3, \ldots$ to noncommutative algebra-valued variables.[10,11]

One of the authors (RS) had found an explicit expression for the KdV hierarchy,[14,15] alternative to the expression found in this paper. Progress has been made with the recent paper,[16] where ''full explicitness'' has been achieved: the constant coefficients of the differential polynomials are given by multiple sums and by products of binomial coefficients.

APPENDIX: EXAMPLES BY MATHEMATICA

Considering the generating functions as functions of the variable h shows that the modified Schur polynomials P_n can be calculated according to

$$P_n(t_1,\ldots,t_n)=\frac{1}{n!}\left(\frac{d}{dh}\right)^n \exp(t_1 h+\cdots+t_n h^n)\bigg|_{h=0}.$$

This formula can be easily evaluated by MATHEMATICA and the polynomials $P_1,\ldots,P_6$ are implemented as follows:

$$\xi[\mathrm{t1\_,t2\_,t3\_,t4\_,t5\_,t6\_}]:=\mathrm{t1*h+t2*h\hat{}2+t3*h\hat{}3+t4*h\hat{}4+t5*h\hat{}5+t6*h\hat{}6}$$

$$\mathrm{P[1,t1\_]:=\ Expand}\left[\frac{\partial_{\{h,1\}}\exp[\xi[\mathrm{t1,t2,t3,t4,t5,t6}]]}{1!}/.\mathrm{h}\to 0\right]$$

$$\mathrm{P[2,t1\_,t2\_]:=\ Expand}\left[\frac{\partial_{\{h,2\}}\exp[\xi[\mathrm{t1,t2,t3,t4,t5,t6}]]}{2!}/.\mathrm{h}\to 0\right]$$

$$\mathrm{P[3,t1\_,t2\_,t3\_]:=\ Expand}\left[\frac{\partial_{\{h,3\}}\exp[\xi[\mathrm{t1,t2,t3,t4,t5,t6}]]}{3!}/.\mathrm{h}\to 0\right]$$

$$\mathrm{P[4,t1\_,t2\_,t3\_,t4\_]:=\ Expand}[\partial_{\{h,4\}}\exp[\xi[\mathrm{t1,t2,t3,t4,t5,t6}]]/4!/.\mathrm{h}\to 0]$$

$$\mathrm{P[5,t1\_,t2\_,t3\_,t4\_,t5\_]:=\ Expand}\left[\frac{\partial_{\{h,5\}}\exp[\xi[\mathrm{t1,t2,t3,t4,t5,t6}]]}{5!}/.\mathrm{h}\to 0\right]$$

$$\mathrm{P[6,t1\_,t2\_,t3\_,t4\_,t5\_,t6\_]:=\ Expand}\left[\frac{\partial_{\{h,6\}}\exp[\xi[\mathrm{t1,t2,t3,t4,t5,t6}]]}{6!}/.\mathrm{h}\to 0\right]$$

Next, the recursion formula (29) for the KdV case is programmed:

$$\mathrm{z[n\_]}:=-\frac{1}{2}\left(\partial_x \mathrm{z[n-1]}+\sum_{k=1}^{n-2}\mathrm{z[k]z[n-1-k]}\right);\mathrm{z[1]}:=-\mathrm{u[x]}$$

Let us calculate the differential polynomials z_2, z_4, z_6 needed for the KdV equation itself and for the next higher equation in the hierarchy:

z[2]//Simplify

$$\frac{u'[x]}{2}$$

z[4]//Simplify

$$u[x]u'[x]+\tfrac{1}{8}u^{(3)}[x]$$

z[6]//Simplify

$$\tfrac{1}{32}(64u[x]^2u'[x]+36u'[x]u''[x]+16u[x]u^{(3)}[x]+u^{(5)}[x]).$$

According to (30), we obtain the right-hand side of the first two members of the KdV hierarchy $\partial u/\partial t_3 = K_3[u]$ and $\partial u/\partial t_5 = K_5[u]$:

D[P[2,2∗Integrate[z[2],x],2∗Integrate[z[4],x]],x]//

Simplify

$$3u[x]u'[x]+\tfrac{1}{4}u^{(3)}[x]$$

D[P[3,2∗Integrate[z[2],x],

2∗Integrate[z[4],x],2∗Integrate[z[6],x]],x]//

Simplify

$$\tfrac{1}{16}(120u[x]^2u'[x]+40u'[x]u''[x]+20u[x]u^{(3)}[x]+u^{(5)}[x]).$$

[1] P. Lax, Commun. Pure Appl. Math. 467 (1968).
[2] R. Hirota, in *Bäcklund Transformations*, Lecture Notes in Mathematics (Springer, Berlin, 1976).
[3] R. Hirota, *Direct Methods in Soliton Theory* (Springer, Berlin, 1980), pp. 157–176.
[4] Y. Ohta, J. Satsuma, D. Takahashi, and T. Tokihiro, Prog. Theor. Phys. Suppl. **94**, 219 (1988).
[5] A. Fokas and P. Santini, Stud. Appl. Math. **75**, 179 (1986).
[6] L. A. Dickey, *Soliton Equations and Hamiltonian Systems* (World Scientific, Singapore, 1991).
[7] J. Matsukidaira, J. Satsuma, and W. Strampp, J. Math. Phys. **31**, 1426 (1990).
[8] J. Matsukidaira, J. Satsuma, and W. Strampp, Phys. Lett. A **147**, 467 (1990).
[9] W. Strampp, J. Math. Phys. **35**, 5850 (1994).
[10] R. Schimming and S. Z. Rida, J. Alg. Comput. **6**, 635 (1996).
[11] R. Schimming and S. Z. Rida, in *Application of Fibonacci Numbers*, edited by G. E. Bergum *et al.* (Kluwer, Dordrecht 1998), Vol. 7.
[12] L. Comtet, *Advanced Combinatorics* (Reidel, Dordrecht, 1974).
[13] F. di Bruno, Q. J. Pure Appl. Math. **1**, 359 (1857).
[14] R. Schimming, Z. Anal. Anw. **7**, 203 (1998).
[15] R. Schimming, Acta. Appl. Math. **39**, 489 (1995).
[16] I. G. Avramidi and R. Schimming, Math. Nachr. (to appear).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Attractors for the Klein–Gordon–Schrödinger equation

Bixiang Wang[a)]
Department of Applied Mathematics, Tsinghua University, Beijing 100084, People's Republic of China

Horst Lange
Mathematisches Institut, Universität zu Köln, D-50931 Köln, Germany

(Received 19 October 1998; accepted for publication 1 February 1999)

In this paper we deal with the asymptotic behavior of solutions for the Klein–Gordon–Schrödinger equation. We prove the existence of compact global attractors for this model in the space $H^l \times H^l \times H^{l-1}$ for each integer $l \geqslant 1$. © *1999 American Institute of Physics.* [S0022-2488(99)00405-3]

I. INTRODUCTION

In this paper, we investigate the long time behavior of solutions for the dissipative Klein–Gordon–Schrödinger system in a bounded domain of $\mathbb{R}^n$ with $n \leqslant 3$. This model is concerned with a complex-valued function ψ and a real-valued function ϕ and takes the form

$$i\psi_t + \Delta\psi + i\nu\psi + \phi\psi = f, \tag{1.1}$$

$$\phi_{tt} + \gamma\phi_t - \Delta\phi + \phi - |\psi|^2 = g, \tag{1.2}$$

where ν and λ are positive constants, f and g driving terms. System (1.1)–(1.2) describes the interaction of a nucleon field ψ and a meson field ϕ through the Yukawa coupling. The dissipative mechanism of this model is introduced by the terms $i\nu\psi$ and $\gamma\phi_t$.

The well-posedness problem of system (1.1)–(1.2) has been studied by many authors such as Bachelot;[1] Hayashi and von Wahl;[2] Fukuda and Tsutsumi,[3,4] and the references therein.

The long time behavior of solutions for this model has been studied by Biler;[5] Li;[6] and Guo and Li.[7] In Ref. 5, the author proved the existence of the weak global attractor in the Hilbert space $H^1 \times H^1$. The finite dimensionality of the weak global attractor was also obtained there. The existence of the strong global attractor in $H^2 \times H^2$ was investigated in Refs. 6 and 7.

In this paper, we first intend to establish the continuity property of solutions on initial data in $H^1 \times H^1$, which was left open in Ref. 5. For system (1.1)–(1.2), it seems difficult to show this continuity by usual methods. We here apply an energy equation to achieve our goal. The energy equation method was essentially due to Ball[8] (see also Ref. 9). The continuity property of solutions is needed for us to construct the strong global attractor and also interesting by itself.

The second purpose of this paper is to present the asymptotic compactness of solutions in $H^l \times H^l$ for each $l \geqslant 1$. In general, this kind of compactness is more difficult to obtain for weakly dissipative equations such as (1.1)–(1.2) than strongly dissipative ones, such as the Navier–Stokes equaion; see, e.g., Refs. 10–14. Again, we, here, employ an energy equation to establish the desired compactness for the Klein–Gordon–Schrödinger model.

As a result of the asymptotic compactness, the existence of the strong global attractor in $H^1 \times H^1$ follows. Obviously, this strong global attractor coincides with the weak one. We mention that here we obtain the existence of the compact global attractor by assuming the driving terms f and g only in L^2, which is weaker than the corresponding condition $f \in H^2$ in Ref. 5.

[a)]Electronic mail: wang@math.byu.edu

Our third task in this paper is to show the existence of the strong global attractor in $H^{k+2}\times H^{k+2}$ when driving terms f and g in H^k with $k\geqslant 0$. This is achieved by similar methods as above, except some more complicated computations.

The organization of this paper is as follows. In the next section, we derive *a priori* estimates on the solutions of system (1.1)–(1.2) in $H^{k+2}\times H^{k+2}$ when f and g belong to H^k with $k\geqslant 0$. These estimates are needed for the proof of the existence of bounded absorbing sets and the asymptotic compactness. Then, in Sec. III, we establish the existence of the dynamical systems associated to problem (1.1)–(1.2) in the space $H^l\times H^l$ for each $l\geqslant 1$. Finally, in Sec. IV, we first prove the asymptotic compactness of the dynamical system and then present the existence of the strong global attractor in each space $H^l\times H^l$ with $l\geqslant 1$.

II. *A PRIORI* ESTIMATES

In this section, we formally derive *a priori* estimates on solutions of the Klein–Gordon–Schrödinger equation. These estimates hold for smooth functions and will become rigorous by a limiting process (e.g., the Galerkin method).

We first introduce the transformation $\theta=\phi_t+\delta\phi$ with δ a small positive constant that will be specified below. Then system (1.1)–(1.2) becomes

$$i\psi_t+\Delta\psi+i\nu\psi+\phi\psi=f,\quad \text{in}\ \ \Omega\times\mathbb{R}^+, \tag{2.1}$$

$$\phi_t+\delta\phi=\theta,\quad \text{in}\ \ \Omega\times\mathbb{R}^+, \tag{2.2}$$

$$\theta_t+(\gamma-\delta)\theta-\Delta\phi+(1-\delta(\gamma-\delta))\phi-|\psi|^2=g,\quad \text{in}\ \ \Omega\times\mathbb{R}^+, \tag{2.3}$$

where Ω is a smooth (e.g., C^2) bounded domain in $\mathbb{R}^n$ with $n\leqslant 3$. Problem (2.1)–(2.3) is supplemented with the initial condition

$$\psi(x,0)=\psi_0(x),\quad \phi(x,0)=\phi_0(x),\quad \theta(x,0)=\theta_0(x),\quad x\in\Omega, \tag{2.4}$$

and the boundary condition

$$(\psi,\phi,\theta)|_{\partial\Omega}=0\quad \text{or}\quad (\psi,\phi,\theta)\ \ \text{is}\ \ \Omega\ \ \text{periodic}. \tag{2.5}$$

In the sequel, we denote by $H^s(\Omega)$ both the standard real and complex Sobolev spaces and $H=L^2(\Omega)$. We also use $\|\cdot\|$ and $(\cdot,\cdot)$ for the usual norm and inner product of $L^2(\Omega)$. For any $1\leqslant p\leqslant\infty$, we denote by $\|\cdot\|_p$ the norm of $L^p(\Omega)$ ($\|\cdot\|_2=\|\cdot\|$). In general, $\|\cdot\|_X$ denotes the norm of any Banach space X.

We are now in a position to derive the estimates on solutions of problem (2.1)–(2.5). We start with the estimates in $H^1\times H^1\times H$.

Lemma 2.1: Assume that f and g belong to H. Then there exists a constant δ_1 such that when $\delta\leqslant\delta_1$, every solution (ψ,ϕ,θ) of problem (2.1)–(2.5) satisfies

$$\|\psi(t)\|_{H^1}+\|\phi(t)\|_{H^1}+\|\theta(t)\|\leqslant M,\quad t\geqslant t_2,$$

where M depends on $(\nu,\gamma,\delta,\|f\|,\|g\|)$; t_2 depends on $(\nu,\gamma,\delta,\|f\|,\|g\|)$ and R when $\|(\psi_0,\phi_0,\theta_0)\|_{H^1\times H^1\times H}\leqslant R$.

Proof: The proof of this lemma is similar to Proposition 2.1 in Ref. 5. Therefore, here we only sketch it.

Taking the imaginary part of the inner product of (2.1) with ψ in H, we get

$$\frac{1}{2}\frac{d}{dt}\|\psi\|^2+\nu\|\psi\|^2=\operatorname{Im}\int_\Omega f\overline{\psi}. \tag{2.6}$$

In the sequel, we denote by $\overline{\psi}$ the conjugate of ψ. Then, taking the real part of the inner product of (2.1) with $-\psi_t-\nu\psi$ in H, we get

$$\frac{1}{2}\frac{d}{dt}\|\nabla\psi\|^2+\nu\|\nabla\psi\|^2-\nu\int_\Omega\phi|\psi|^2+\nu\,\mathrm{Re}\int_\Omega f\overline{\psi}-\frac{1}{2}\int_\Omega\phi\frac{\partial}{\partial t}|\psi|^2+\mathrm{Re}\int_\Omega f\overline{\psi_t}=0. \quad (2.7)$$

Using (2.2), we find

$$\int_\Omega\phi\frac{\partial}{\partial t}|\psi|^2=\frac{d}{dt}\int_\Omega\phi|\psi|^2-\int_\Omega\phi_t|\psi|^2=\frac{d}{dt}\int_\Omega\phi|\psi|^2+\delta\int_\Omega\phi|\psi|^2-\int_\Omega\theta|\psi|^2. \quad (2.8)$$

By (2.6)–(2.8), we have

$$\frac{1}{2}\frac{d}{dt}\left(\|\psi\|^2+\|\nabla\psi\|^2-\int_\Omega\phi|\psi|^2+2\,\mathrm{Re}\int_\Omega f\overline{\psi}\right)+\nu\|\psi\|^2+\nu\|\nabla\psi\|^2-\left(\nu+\frac{1}{2}\delta\right)\int_\Omega\phi|\psi|^2$$
$$+\frac{1}{2}\int_\Omega\theta|\psi|^2+\nu\,\mathrm{Re}\int_\Omega f\overline{\psi}-\mathrm{Im}\int_\Omega f\overline{\psi}=0. \quad (2.9)$$

Taking the inner product of (2.3) with θ in H, we have

$$\frac{1}{2}\frac{d}{dt}\|\theta\|^2+(\gamma-\delta)\|\theta\|^2-\int_\Omega\theta\,\Delta\phi+(1-\delta(\gamma-\delta))\int_\Omega\phi\theta-\int_\Omega|\psi|^2\theta=\int_\Omega g\theta. \quad (2.10)$$

Using (2.2), we get from (2.10) that

$$\frac{1}{2}\frac{d}{dt}(\|\theta\|^2+(1-\delta(\gamma-\delta))\|\phi\|^2+\|\nabla\phi\|^2)+(\gamma-\delta)\|\theta\|^2$$
$$+\delta(1-\delta(\gamma-\delta))\|\phi\|^2+\delta\|\nabla\phi\|^2-\int_\Omega|\psi|^2\theta=\int_\Omega g\theta. \quad (2.11)$$

Then, by $4\times(2.9)+2\times(2.11)$, we obtain

$$\frac{d}{dt}E(\psi,\phi,\theta)+2\delta E(\psi,\phi,\theta)=F(\psi,\phi,\theta), \quad (2.12)$$

where

$$E(\psi,\phi,\theta)=2\|\psi\|^2+2\|\nabla\psi\|^2+(1-\delta(\gamma-\delta))\|\phi\|^2+\|\nabla\phi\|^2+\|\theta\|^2-2\int_\Omega|\psi|^2\phi+4\,\mathrm{Re}\int_\Omega f\overline{\psi}, \quad (2.13)$$

and

$$F(\psi,\phi,\theta)=-4(\nu-\delta)(\|\psi\|^2+\|\nabla\psi\|^2)-2(\gamma-2\delta)\|\theta\|^2$$
$$+2(2\nu-\delta)\int_\Omega\phi|\psi|^2+4\,\mathrm{Im}\int_\Omega f\overline{\psi}+4(2\delta-\nu)\mathrm{Re}\int_\Omega f\overline{\psi}+2\int_\Omega g\theta. \quad (2.14)$$

Using (2.12) and proceeding as in Ref. 5, we can easily get the lemma. The details are omitted here and then the proof is finished. □

We remark that the energy equation (2.12) will play a crucial role in the proof of the continuity of solutions on initial data in $H^1\times H^1\times H$ in the next section. Equation (2.12) is also an important tool for the proof of the asymptotic compactness of solutions in Sec. IV.

Below, we improve the estimates in the previous lemma to the space $H^{k+2}\times H^{k+2}\times H^{k+1}$ when f and g belong to H^k with $k\geqslant 0$.

Lemma 2.2: Assume that f and g belong to $H^k(\Omega)$ with $k\geqslant 0$. Then every solution (ψ,ϕ,θ) of problem (2.1)–(2.5) satisfies

$$\|\psi(t)\|_{H^{k+2}}+\|\phi(t)\|_{H^{k+2}}+\|\theta(t)\|_{H^{k+1}}\leqslant M_k,\quad t\geqslant t_k,$$

where M_k depends on $(\nu,\gamma,\delta,\|f\|_{H^k},\|g\|_{H^k})$ and k; t_k depends on $(\nu,\gamma,\delta,\|f\|_{H^k},\|g\|_{H^k})$ and k and R when $\|(\psi_0,\phi_0,\theta_0)\|_{H^{k+2}\times H^{k+2}\times H^{k+1}}\leqslant R$.

Proof: taking the real part of the inner product of (2.1) with $(-1)^k(\Delta^{k+1}\psi_t+\nu\Delta^{k+1}\psi)$ in H, we get that

$$\frac{1}{2}\frac{d}{dt}\|\nabla^{k+2}\psi\|^2+\nu\|\nabla^{k+2}\psi\|^2+\nu\ \mathrm{Re}\int_\Omega\nabla^{k+2}\psi\cdot\nabla^k(\phi\overline{\psi})-\nu\ \mathrm{Re}\int_\Omega\nabla^k f\cdot\nabla^{k+2}\overline{\psi}$$
$$+(-1)^k\ \mathrm{Re}\int_\Omega\phi\psi\ \Delta^{k+1}\overline{\psi_t}-\mathrm{Re}\int_\Omega\nabla^k f\cdot\nabla^{k+2}\overline{\psi_t}=0. \tag{2.15}$$

Note that

$$(-1)^k\ \mathrm{Re}\int_\Omega\phi\psi\,\Delta^{k+1}\overline{\psi_t}=\mathrm{Re}\int_\Omega\nabla^k(\phi\overline{\psi})\cdot\nabla^{k+2}\psi_t$$
$$=\frac{d}{dt}\mathrm{Re}\int_\Omega\nabla^k(\phi\overline{\psi})\cdot\nabla^{k+2}\psi-\mathrm{Re}\int_\Omega\nabla^k(\phi_t\overline{\psi})\cdot\nabla^{k+2}\psi$$
$$-\mathrm{Re}\int_\Omega\nabla^k(\phi\overline{\psi_t})\cdot\nabla^{k+2}\psi. \tag{2.16}$$

By (2.1) and (2.2), we first substitute ψ_t and ϕ_t into (2.16), and then from (2.15) we find that

$$\frac{1}{2}\frac{d}{dt}\left(\|\nabla^{k+2}\psi\|^2+2\ \mathrm{Re}\int_\Omega\nabla^k(\phi\overline{\psi})\cdot\nabla^{k+2}\psi-2\ \mathrm{Re}\int_\Omega\nabla^k f\cdot\nabla^{k+2}\overline{\psi}\right)+\nu\|\nabla^{k+2}\psi\|^2$$
$$+(2\nu+\delta)\mathrm{Re}\int_\Omega\nabla^k(\phi\overline{\psi})\cdot\nabla^{k+2}\overline{\psi}-\nu\ \mathrm{Re}\int_\Omega\nabla^k f\cdot\nabla^{k+2}\overline{\psi}-\mathrm{Re}\int_\Omega\nabla^k(\theta\overline{\psi})\cdot\nabla^{k+2}\psi$$
$$-\mathrm{Im}\int_\Omega\nabla^k(\phi\,\Delta\overline{\psi})\cdot\nabla^{k+2}\psi-\mathrm{Im}\int_\Omega\nabla^k(\phi^2\overline{\psi})\cdot\nabla^{k+2}\psi+\mathrm{Im}\int_\Omega\nabla^k(\phi f)\cdot\nabla^{k+2}\overline{\psi}=0. \tag{2.17}$$

Now, we derive an energy equation for ϕ and θ. Taking the inner product of (2.3) with $(-1)^{k+1}\Delta^{k+1}\theta$ in H, we find

$$\frac{1}{2}\frac{d}{dt}\|\nabla^{k+1}\theta\|^2+(\gamma-\delta)\|\nabla^{k+1}\theta\|^2-(-1)^{k+1}\int_\Omega\Delta\phi\,\Delta^{k+1}\theta+(-1)^{k+1}(1-\delta(\gamma-\delta))$$
$$\times\int_\Omega\phi\,\Delta^{k+1}\theta-(-1)^{k+1}\int_\Omega|\psi|^2\,\Delta^{k+1}\theta=(-1)^{k+1}\int_\Omega g\,\Delta^{k+1}\theta. \tag{2.18}$$

By (2.2) we have

$$-\int_\Omega \Delta\phi\,\Delta^{k+1}\theta+(-1)^{k+1}(1-\delta(\gamma-\delta))\int_\Omega \phi\,\Delta^{k+1}\theta$$
$$=\frac{1}{2}(1-\delta(\gamma-\delta))\frac{d}{dt}\|\nabla^{k+1}\phi\|^2$$
$$+\delta(1-\delta(\gamma-\delta))\|\nabla^{k+1}\phi\|^2+\frac{1}{2}\frac{d}{dt}\|\nabla^{k+2}\phi\|^2+\delta\|\nabla^{k+2}\phi\|^2, \tag{2.19}$$

and

$$(-1)^{k+1}\int_\Omega g\,\Delta^{k+1}\theta=-\frac{d}{dt}\int_\Omega \nabla^k g\cdot\nabla^{k+2}\phi-\delta\int_\Omega \nabla^k g\cdot\nabla^{k+2}\phi. \tag{2.20}$$

Then, it follows from (2.18)–(2.20) that

$$\frac{1}{2}\frac{d}{dt}\left(\|\nabla^{k+1}\theta\|^2+(1-\delta(\gamma-\delta))\|\nabla^{k+1}\phi\|^2+\|\nabla^{k+2}\phi\|^2+2\int_\Omega \nabla^k g\cdot\nabla^{k+2}\phi\right)$$
$$+(\gamma-\delta)\|\nabla^{k+1}\theta\|^2+\delta(1-\delta(\gamma-\delta))\|\nabla^{k+1}\phi\|^2+\delta\|\nabla^{k+2}\phi\|^2$$
$$+\delta\int_\Omega \nabla^k g\cdot\nabla^{k+2}\phi-\int_\Omega \nabla^{k+1}|\psi|^2\cdot\nabla\theta=0. \tag{2.21}$$

Summing up (2.17) and (2.21), we get that

$$\frac{d}{dt}E_k(\psi(t),\phi(t),\theta(t))+\delta E_k(\psi(t),\phi(t),\theta(t))=F_k(\psi(t),\phi(t),\theta(t)), \tag{2.22}$$

where

$$E_k(\psi,\phi,\theta)=\|\nabla^{k+2}\psi\|^2+(1-\delta(\gamma-\delta))\|\nabla^{k+1}\phi\|^2+\|\nabla^{k+2}\phi\|^2+\|\nabla^{k+1}\theta\|^2$$
$$+2\,\mathrm{Re}\int_\Omega \nabla^k(\phi\overline{\psi})\cdot\nabla^{k+2}\psi-2\,\mathrm{Re}\int_\Omega \nabla^k f\cdot\nabla^{k+2}\overline{\psi}+2\int_\Omega \nabla^k g\cdot\nabla^{k+2}\phi,$$

and

$$F_k(\psi,\phi,\theta)=-(2\nu-\delta)\|\nabla^{k+2}\psi\|^2-\delta(1-\delta(\gamma-\delta))\|\nabla^{k+1}\phi\|^2-\delta\|\nabla^{k+2}\phi\|^2$$
$$-(2\gamma-3\delta)\|\nabla^{k+1}\theta\|^2-4\nu\,\mathrm{Re}\int_\Omega \nabla^k(\phi\overline{\psi})\cdot\nabla^{k+2}\psi-2(\gamma-\nu)\mathrm{Re}\int_\Omega \nabla^k f\cdot\nabla^{k+2}\overline{\psi}$$
$$+2\,\mathrm{Re}\int_\Omega \nabla^k(\theta\overline{\psi})\cdot\nabla^{k+2}\psi+2\,\mathrm{Im}\int_\Omega \nabla^k(\phi\,\Delta\overline{\psi})\cdot\nabla^{k+2}\psi$$
$$+2\,\mathrm{Im}\int_\Omega \nabla^k(\phi^2\overline{\psi})\cdot\nabla^{k+2}\psi-2\,\mathrm{Im}\int_\Omega \nabla^k(\phi f)\nabla^{k+2}\overline{\psi}+2\int_\Omega \nabla^{k+1}|\psi|^2\cdot\nabla^{k+1}\theta.$$

Thus, this lemma follows from (2.21) and Gronwall's lemma. The details are omitted here. □

In what follows, we state some estimates on the solutions in a finite time interval that will be used when we establish the existence and uniqueness and continuity of solutions. By (2.12) we can easily get the following result.

Lemma 2.3: Assume that f and g belong to H. Then every solution (ψ,ϕ,θ) *of problem (2.1)–(2.5) satisfies*

$$\|\psi(t)\|_{H^1}+\|\phi(t)\|_{H^1}+\|\theta(t)\|\leqslant L,\quad 0\leqslant t\leqslant T,$$

where L depends on $(\nu,\gamma,\delta,\|f\|,\|g\|)$, *T and* $\|(\psi_0,\phi_0,\theta_0)\|_{H^1\times H^1\times H}$.

The following fact is the analog of Lemma 2.2.

Lemma 2.4: Assume that f and g belong to $H^k(\Omega)$ *with* $k\geqslant 0$. *Then every solution* (ψ,ϕ,θ) *of problem (2.1)–(2.5) satisfies*

$$\|\psi(t)\|_{H^{k+2}}+\|\phi(t)\|_{H^{k+2}}+\|\theta(t)\|_{H^{k+1}}\leqslant L_k,\quad 0\leqslant t\leqslant T,$$

where L_k *depends on* $(\nu,\gamma,\delta,\|f\|_{H^k},\|g\|_{H^k})$ *and T and k and* $\|(\psi_0,\phi_0,\theta_0)\|_{H^{k+2}\times H^{k+2}\times H^{k+1}}$.

III. THE SOLUTION SEMIGROUP

In this section, we establish the existence of the dynamical system associated to problem (2.1)–(2.5). What is more, we shall show the existence and uniqueness and continuity of solutions in the space $H^l\times H^l\times H^{l-1}$ for each $l\geqslant 1$. The existence and uniqueness of solutions follow from standard methods and estimates in Lemmas 2.3 and 2.4. Similar results for $l\leqslant 2$ was also obtained in Ref. 5. Lemmas 2.3 and 2.4 are also sufficient to establish the continuity property of solutions on initial data in $H^l\times H^l\times H^{l-1}$ when $l\geqslant 2$. However, the continuity property in $H^1\times H^1\times H$ is not so easy and needs to be treated separately. The main purpose of this section is to prove the solution is indeed continuous on initial data in $H^1\times H^1\times H$. To this end, the energy equation obtained in the previous section will play a key role; see Theorem 3.4 below.

By Lemma 2.3, we have the following existence and uniqueness result.

Theorem 3.1: *Let f,* $g\in H$ *and* $(\psi_0,\phi_0,\theta_0)\in H^1\times H^1\times H$. *Then problem (2.1)–(2.5) has a unique solution* $(\psi,\phi,\theta)\in C(\mathbb{R}^+,H^1\times H^1\times H)$.

For the proof of this theorem, we refer the readers to Ref. 5 and the references therein. Here we state it just for our purpose below.

The next result can be proved by Lemma 2.4 and the details will be omitted here.

Theorem 3.2: *Let f,* $g\in H^k$ *and* $(\psi_0,\phi_0,\theta_0)\in H^{k+2}\times H^{k+2}\times H^{k+1}$ *with* $k\geqslant 0$. *Then problem (2.1)–(2.5) has a unique solution* $(\psi,\phi,\theta)\in C(\mathbb{R}^+,H^{k+2}\times H^{k+2}\times H^{k+1})$.

We remark that by Theorem 3.2, one has the existence and uniqueness of solutions in $H^2\times H^2\times H^1$ under the assumption of the driving terms in *H*, which is weaker than the corresponding condition $g\in H^1$ in Ref. 5.

The following statement is concerned with the continuity of solutions on initial data in $H^{k+2}\times H^{k+2}\times H^{k+1}$ for each $k\geqslant 0$.

Theorem 3.3: *Assume that the hypothesis of Theorem 3.2 holds. Then the solution* (ψ,ϕ,θ) *of problem (2.1)–(2.5) is continuous with respect to initial data in* $H^{k+2}\times H^{k+2}\times H^{k+1}$.

Proof: We shall sketch the proof below and the details will be omitted.

Consider two solutions (ψ_1,ϕ_1,θ_1) and (ψ_2,ϕ_2,θ_2) of problem (2.1)–(2.5). Then the difference $(\psi,\phi,\theta)=(\psi_1-\psi_2,\phi_1-\phi_2,\theta_1-\theta_2)$ satisfies

$$i\psi_t+\Delta\psi+i\nu\psi+\phi\psi_1+\phi_2\psi=0,\tag{3.1}$$

$$\phi_t+\delta\phi=\theta,\tag{3.2}$$

$$\theta_t+(\gamma-\delta)\theta-\Delta\psi+(1-\delta(\gamma-\delta))\phi+\psi\overline{\psi_1}-\overline{\psi}\psi_2=0.\tag{3.3}$$

Taking the real part of the inner product of (3.1) with $(-1)^k(\Delta^{k+1}\psi_t+\nu\Delta^{k+1}\psi)$ in *H*, and then using Lemma 2.4 to estimate the terms in the resulting identity, after some computations we can get

$$\frac{d}{dt}\|\psi\|^2_{H^{k+2}}\leqslant C(\|\psi\|^2_{H^{k+2}}+\|\phi\|^2_{H^{k+2}}).\tag{3.4}$$

Similarly, taking the inner product of (3.3) with $(-1)^{k+1}\Delta^{k+1}\theta$ in H, then using (3.2) and Lemma 2.4, we can derive the inequality

$$\frac{d}{dt}((1-\delta(\gamma-\delta))\|\phi\|^2_{H^{k+1}}+\|\phi\|^2_{H^{k+2}}+\|\theta\|^2_{H^{k+1}}) \leq C((1-\delta(\gamma-\delta))\|\phi\|^2_{H^{k+1}}+\|\phi\|^2_{H^{k+2}}+\|\theta\|^2_{H^{k+1}}). \tag{3.5}$$

It follows from (3.4) and (3.5) that

$$\frac{d}{dt}(\|\psi\|^2_{H^{k+2}}+(1-\delta(\gamma-\delta))\|\phi\|^2_{H^{k+1}}+\|\phi\|^2_{H^{k+2}}+\|\theta\|^2_{H^{k+1}}) \leq C(\|\psi\|^2_{H^{k+2}}+(1-\delta(\gamma-\delta))\|\phi\|^2_{H^{k+1}}+\|\phi\|^2_{H^{k+2}}+\|\theta\|^2_{H^{k+1}}). \tag{3.6}$$

This and Gronwall's lemma give the Lipschitz continuity of solutions in $H^{k+2}\times H^{k+2}\times H^{k+1}$. The proof is complete. □

We remark that the inequality similar to (3.6) is not easy to get in the space $H^1\times H^1\times H$. So the above method does no longer apply. Below, we shall use the energy equation method to prove the continuity in this case.

Theorem 3.4: *Assume that f and g belong to H. Then the solution* $(\psi,\phi,\theta)\in C(\mathbb{R}^+, H^1\times H^1\times H)$ *of system (2.1)–(2.5) depends continuously on the initial data in* $H^1\times H^1\times H$.

Proof: Assume that $(\psi_{0,n},\phi_{0,n},\theta_{0,n})\to(\psi_0,\phi_0,\theta_0)$ in $H^1\times H^1\times H$; we want to prove $S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n})\to S(t)(\psi_0,\phi_0,\theta_0)$ for each $t>0$ as $n\to\infty$.

Given $t>0$, we choose $T>t$. By the convergence of $(\psi_{0,n},\phi_{0,n},\theta_{0,n})$, we know that $(\psi_{0,n},\phi_{0,n},\theta_{0,n})$ is bounded in $H^1\times H^1\times H$. Thus, it follows from Lemma 2.3 that

$$\|\psi_n(\tau)\|_{H^1}+\|\phi_n(\tau)\|_{H^1}+\|\theta_n(\tau)\|\leq C,\quad 0\leq\tau\leq T, \tag{3.7}$$

where we set $(\psi_n(\tau),\phi_n(\tau),\theta_n(\tau))=S(\tau)(\psi_{0,n},\phi_{0,n},\theta_{0,n})$. By (3.7) and system (2.1)–(2.5) we can also deduce that

$$\left\|\frac{\partial}{\partial\tau}\psi_n\right\|_{L^2(0,T;H^{-1})}+\left\|\frac{\partial}{\partial\tau}\phi_n\right\|_{L^2(0,T;H)}\leq C. \tag{3.8}$$

And, therefore, there exists $(\psi(\tau),\phi(\tau),\theta(\tau))\in L^\infty(0,T;H^1\times H^1\times H)$, such that

$$(\psi_n(\tau),\phi_n(\tau),\theta_n(\tau))\to(\psi(\tau),\phi(\tau),\theta(\tau))\ \text{weakly in}\ L^2(0,T;H^1\times H^1\times H), \tag{3.9}$$

and

$$\frac{\partial}{\partial\tau}\psi_n\to\frac{\partial}{\partial\tau}\psi,\quad\text{weakly in}\ L^2(0,T;H^{-1}), \tag{3.10}$$

$$\frac{\partial}{\partial\tau}\phi_n\to\frac{\partial}{\partial\tau}\phi,\quad\text{weakly in}\ L^2(0,T;H). \tag{3.11}$$

By (3.9)–(3.11) and a standard compactness result, we have

$$(\psi_n,\phi_n)\to(\psi,\phi),\ \text{strongly in}\ L^2(0,T;H\times H). \tag{3.12}$$

Again, for a fixed t, by (3.7) we see that there exists $\xi\in H^1\times H^1\times H$ such that

$$(\psi_n(t),\phi_n(t),\theta_n(t))\to\xi,\ \text{weakly in}\ H^1\times H^1\times H. \tag{3.13}$$

Then, similar to the proof of the existence of solutions (see Ref. 4, for example), we can easily deduce that (ψ,ϕ,θ) is the solution of problem (2.1)–(2.5) with $(\psi(0),\phi(0),\theta(0))=(\psi_0,\phi_0,\theta_0)$ and $(\psi(t),\phi(t),\theta(t))=\xi$ for a fixed t. This fact, together with (3.13), implies the weak convergence of $S(t)$, that is,

$$S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n})\rightarrow S(t)(\psi_0,\phi_0,\theta_0), \quad \text{weakly in } H^1\times H^1\times H. \tag{3.14}$$

We now prove that the weak convergence is, in fact, the strong convergence. To this end, we shall use the energy equation (2.12). In fact, by (2.12) we have

$$E(t)=e^{-2\delta t}E(0)+\int_0^t e^{-2\delta(t-\tau)}F(\psi(\tau),\phi(\tau),\theta(\tau))d\tau, \tag{3.15}$$

where E and F is given by (2.13) and (2.14), respectively. Note that (3.15) means that any solution $(\psi(t),\phi(t),\theta(t))=S(t)(\psi_0,\phi_0,\theta_0)$ of problem (2.1)–(2.5) verifies

$$E(S(t)(\psi_0,\phi_0,\theta_0))=e^{-2\delta t}E(\psi_0,\phi_0,\theta_0)+\int_0^t e^{-2\delta(t-\tau)}F(S(\tau)(\psi_0,\phi_0,\theta_0))d\tau. \tag{3.16}$$

Applying (3.16) to the solution $S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n})$, we get

$$E(S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n}))=e^{-2\delta t}E(\psi_{0,n},\phi_{0,n},\theta_{0,n})+\int_0^t e^{-2\delta(t-\tau)}F(S(\tau)(\psi_{0,n},\phi_{0,n},\theta_{0,n}))d\tau. \tag{3.17}$$

Since $(\psi_{0,n},\phi_{0,n},\theta_{0,n})\rightarrow(\psi_0,\phi_0,\theta_0)$ in $H^1\times H^1\times H$, by the definition of E in (2.13), we can deduce that, as $n\rightarrow\infty$,

$$E(\psi_{0,n},\phi_{0,n},\theta_{0,n})\rightarrow E(\psi_0,\phi_0,\theta_0). \tag{3.18}$$

Below, we deal with the limit of the second term on the right-hand side of (3.17). By (2.14) we have

$$\begin{aligned}
&\int_0^t e^{-2\delta(t-\tau)}F(S(\tau)(\psi_{0,n},\phi_{0,n},\theta_{0,n}))d\tau\\
&\quad=-2\int_0^t e^{-2\delta(t-\tau)}(2(\nu-\delta)(\|S(\tau)\psi_{0,n}\|^2+\|\nabla S(\tau)\psi_{0,n}\|^2)+(\gamma-2\delta)\|S(\tau)\theta_{0,n}\|^2)\\
&\qquad+2\int_0^t e^{-2\delta(t-\tau)}\left((2\nu-\delta)\int_\Omega S(\tau)\phi_{0,n}|S(\tau)\psi_{0,n}|^2+2\,\mathrm{Im}\int_\Omega f\overline{S(\tau)\psi_{0,n}}\right)\\
&\qquad+2\int_0^t e^{-2\delta(t-\tau)}\left(2(2\delta-\nu)\mathrm{Re}\int_\Omega f\overline{S(\tau)\psi_{0,n}}+\int_\Omega gS(\tau)\theta_{0,n}\right).
\end{aligned} \tag{3.19}$$

We now handle the first term on the right-hand side of (3.19). By the weak convergence (3.9), we have

$$e^{-\delta(t-\tau)}S(\tau)\psi_{0,n}\rightarrow e^{-\delta(t-\tau)}S(\tau)\psi_0, \text{ weakly in } L^2(0,t;H^1),$$

and

$$e^{-\delta(t-\tau)}S(\tau)\theta_{0,n}\rightarrow e^{-\delta(t-\tau)}S(\tau)\theta_0, \text{ weakly in } L^2(0,t;H).$$

So we find that

$$\liminf_{n\to\infty} \|e^{-\delta(t-\tau)}S(\tau)\psi_{0,n}\|_{L^2(0,t;H^1)} \geq \|e^{-\delta(t-\tau)}S(\tau)\psi_0\|_{L^2(0,t;H^1)}, \tag{3.20}$$

and

$$\liminf_{n\to\infty} \|e^{-\delta(t-\tau)}S(\tau)\theta_{0,n}\|_{L^2(0,t;H)} \geq \|e^{-\delta(t-\tau)}S(\tau)\theta_0\|_{L^2(0,t;H)}. \tag{3.21}$$

Then, choosing δ small enough such that $\nu-\delta>0$, and $\gamma-2\delta>0$, we have the following estimates for the first term on the right-hand side of (3.19):

$$\limsup_{n\to\infty} -2\int_0^t e^{-2\delta(t-\tau)}(2(\nu-\delta)(\|S(\tau)\psi_{0,n}\|^2+\|\nabla S(\tau)\psi_{0,n}\|^2)+(\gamma-2\delta)\|S(\tau)\theta_{0,n}\|^2)$$
$$\leq -2\int_0^t e^{-2\delta(t-\tau)}(2(\nu-\delta)(\|S(\tau)\psi_0\|^2+\|\nabla S(\tau)\psi_0\|^2)+(\gamma-2\delta)\|S(\tau)\theta_0\|^2). \tag{3.22}$$

For the second term on the right-hand side of (3.19), we want to prove its convergence. We first treat the convergence of

$$\int_0^t e^{-2\delta(t-\tau)}\int_\Omega S(\tau)\phi_{0,n}|S(\tau)\psi_{0,n}|^2 \to \int_0^t e^{-2\delta(t-\tau)}\int_\Omega S(\tau)\phi_0|S(\tau)\psi_0|^2. \tag{3.23}$$

Note that

$$\int_0^t e^{-2\delta(t-\tau)}\int_\Omega S(\tau)\phi_{0,n}|S(\tau)\psi_{0,n}|^2 - \int_0^t e^{-2\delta(t-\tau)}\int_\Omega S(\tau)\phi_0|S(\tau)\psi_0|^2$$
$$=\int_0^t\int_\Omega e^{-2\delta(t-\tau)}(S(\tau)\phi_{0,n}-S(\tau)\phi_0)|S(\tau)\psi_{0,n}|^2$$
$$+\int_0^t\int_\Omega e^{-2\delta(t-\tau)}S(\tau)\phi_0(|S(\tau)\psi_{0,n}|+|S(\tau)\psi_0|)(|S(\tau)\psi_{0,n}|-|S(\tau)\psi_0|)=\mathrm{I}+\mathrm{II},$$

where I and II denotes by the first and second term, respectively. We now show I and II converge to zero as $n\to\infty$. First we have

$$|I| \leq \int_0^t \|S(\tau)\psi_{0,n}\|_4^2\|S(\tau)\phi_{0,n}-S(\tau)\phi_0\| \leq \|S(\tau)\psi_{0,n}\|^2_{L^\infty(0,t;H^1)}\int_0^t\|S(\tau)\phi_{0,n}-S(\tau)\phi_0\|\to 0;$$

the last relation is obtained by (3.12) and Lemma 2.3. Similarly, by (3.12) and Lemma 2.3 again, we can deduce that $\mathrm{II}\to 0$. And then (3.23) follows. Therefore, by (3.12) we can get

$$2\int_0^t e^{-2\delta(t-\tau)}\left((2\nu-\delta)\int_\Omega S(\tau)\phi_{0,n}|S(\tau)\psi_{0,n}|^2+2\,\mathrm{Im}\int_\Omega f\overline{S(\tau)\psi_{0,n}}\right)$$
$$+2\int_0^t e^{-2\delta(t-\tau)}\left(2(2\delta-\nu)\mathrm{Re}\int_\Omega f\overline{S(\tau)\psi_{0,n}}+\int_\Omega gS(\tau)\theta_{0,n}\right)$$
$$\to 2\int_0^t e^{-2\delta(t-\tau)}\left((2\nu-\delta)\int_\Omega S(\tau)\phi_0|S(\tau)\psi_0|^2+2\,\mathrm{Im}\int_\Omega f\overline{S(\tau)\psi_0}\right)$$
$$+2\int_0^t e^{-2\delta(t-\tau)}\left(2(2\delta-\nu)\mathrm{Re}\int_\Omega f\overline{S(\tau)\psi_0}+\int_\Omega gS(\tau)\theta_0\right). \tag{3.24}$$

So, by (3.17)–(3.19), (3.22), and (3.24) and the definition of F in (2.14), we finally get

$$\limsup_{n\to\infty} E(S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n}))$$
$$\leq e^{-2\delta t}E(\psi_0,\phi_0,\theta_0)-2\int_0^t e^{-2\delta(t-\tau)}(2(\nu-\delta)(\|S(\tau)\psi_0\|^2+\|\nabla S(\tau)\psi_0\|^2)$$
$$+(\gamma-2\delta)\|S(\tau)\theta_0\|^2)+2\int_0^t e^{-2\delta(t-\tau)}\left((2\nu-\delta)\int_\Omega S(\tau)\phi_0|S(\tau)\psi_0|^2+2\,\mathrm{Im}\int_\Omega f\overline{S(\tau)\psi_0}\right)$$
$$+2\int_0^t e^{-2\delta(t-\tau)}\left(2(2\delta-\nu)\mathrm{Re}\int_\Omega f\overline{S(\tau)\psi_0}+\int_\Omega gS(\tau)\theta_0\right). \tag{3.25}$$

By (3.16) we see that the right-hand side of (3.25) is exactly $E(S(t)(\psi_0,\phi_0,\theta_0))$, so we have

$$\limsup_{n\to\infty} E(S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n}))\leq E(S(t)(\psi_0,\phi_0,\theta_0)). \tag{3.26}$$

By the weak convergence (3.14) and the compact imbedding $H^1\subset H$, we find

$$S(t)(\psi_{0,n},\phi_{0,n})\to S(t)(\psi_0,\phi_0),\ \text{ strongly in } H,$$

so it is easy to prove

$$\int_\Omega S(t)\phi_{0,n}|S(t)\psi_{0,n}|^2\to\int_\Omega S(t)\phi_0|S(t)\psi_0|^2, \tag{3.27}$$

and

$$\int_\Omega f\overline{S(t)\psi_{0,n}}\to\int_\Omega f\overline{S(t)\psi_0}. \tag{3.28}$$

By (3.26)–(3.28) and the definition of E, we finally get that

$$\limsup_{n\to\infty}\ (2\|S(t)\psi_{0,n}\|_{H^1}^2+(1-\delta(\gamma-\delta))\|S(t)\phi_{0,n}\|^2+\|\nabla S(t)\phi_{0,n}\|^2+\|S(t)\theta_{0,n}\|^2)$$
$$\leq 2\|S(t)\psi_0\|_{H^1}^2+(1-\delta(\gamma-\delta))\|S(t)\phi_0\|^2+\|\nabla S(t)\phi_0\|^2+\|S(t)\theta_0\|^2.$$

We note that the right-hand side of the above is equivalent to the norm of $H^1\times H^1\times H$, so, without loss of generality, we can assume that the norm of $H^1\times H^1\times H$ is defined by it. Then we have

$$\limsup_{n\to\infty}\ \|S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n})\|_{H^1\times H^1\times H}\leq\|S(t)(\psi_0,\phi_0,\theta_0)\|_{H^1\times H^1\times H}. \tag{3.29}$$

On the other hand, the weak convergence (3.14) implies

$$\liminf_{n\to\infty}\ \|S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n})\|_{H^1\times H^1\times H}\geq\|S(t)(\psi_0,\phi_0,\theta_0)\|_{H^1H^1\times H}. \tag{3.30}$$

Then, (3.29) and (3.30) and (3.14) yield

$$S(t)(\psi_{0,n},\phi_{0,n},\theta_{0,n})\to S(t)(\psi_0,\phi_0,\theta_0),\quad \text{strongly in } H^1\times H^1\times H.$$

The proof is complete. □

IV. GLOBAL ATTRACTORS

In this section, we establish the existence of the global attractor for the dynamical system $S(t)$ in the space $H^l \times H^l \times H^{l-1}$ for each $l \geqslant 1$. To this end, we first need to prove the asymptotic compactness of solutions. The technique used here is the energy equation method that was due to Ball.[8] The existence of the global attractor follows from the abstract result[9] (see also Refs. 12 and 13) once the asymptotic compactness of $S(t)$ is established.

In the sequel, we denote by B_1 the ball

$$B_1 = \{(\psi,\phi,\theta) \in H^1 \times H^1 \times H : \|\psi\|_{H^1} + \|\phi\|_{H^1} + \|\theta\| \leqslant M\}, \tag{4.1}$$

where M is the constant in Lemma 2.1, and B_{k+2} the ball

$$B_{k+2} = \{(\psi,\phi,\theta) \in H^{k+2} \times H^{k+2} \times H^{k+1} : \|\psi\|_{H^{k+2}} + \|\phi\|_{H^{k+2}} + \|\theta\|_{H^{k+1}} \leqslant M_k\}, \tag{4.2}$$

for each $k \geqslant 0$, where M_k is the constant in Lemma 2.2.

Then by Lemmas 2.1 and 2.2, we know that B_1 and B_{k+2} $(k \geqslant 0)$ is a bounded absorbing set for $S(t)$ in $H^1 \times H^1 \times H$ and $H^{k+2} \times H^{k+2} \times H^{k+1}$, respectively.

We now prove the asymptotic compactness of $S(t)$ in $H^1 \times H^1 \times H$.

Theorem 4.1: *Assume that f and g belong to H. Then the dynamical system $S(t)$ is asymptotically compact in $H^1 \times H^1 \times H$, that is, if $(\psi_n, \phi_n, \theta_n)$ is bounded in $H^1 \times H^1 \times H$ and $t_n \to \infty$, then $S(t_n)(\psi_n, \phi_n, \theta_n)$ is precompact.*

Proof: Since $(\psi_n, \phi_n, \theta_n)$ is bounded, we can assume that $\|(\psi_n, \phi_n, \theta_n)\|_{H^1 \times H^1 \times H} \leqslant R$ for a suitable constant R. Then by Lemma 2.1, we infer that there exists a constant $T(R)$ depending on R, such that

$$S(t)(\psi_n, \phi_n, \theta_n) \in B_1, \quad \forall t \geqslant T(R), \tag{4.3}$$

where B_1 is the absorbing set in (4.1). Since $t_n \to \infty$, there exists $N_1(R)$ such that if $n \geqslant N_1$, then $t_n \geqslant T(R)$, and hence

$$S(t_n)(\psi_n, \phi_n, \theta_n) \in B_1, \quad \forall n \geqslant N_1. \tag{4.4}$$

By (4.4) we know that there exists $(\psi,\phi,\theta) \in B_1$ such that, up to a subsequence,

$$S(t_n)(\psi_n, \phi_n, \theta_n) \to (\psi,\phi,\theta), \quad \text{weakly in } H^1 \times H^1 \times H. \tag{4.5}$$

For every $T > 0$, again by $t_n \to \infty$, there exists $N_2(R,T)$ such that for $n \geqslant N_2$, we have $t_n - T \geqslant T(R)$. So, by (4.3) we get

$$S(t_n - T)(\psi_n, \phi_n, \theta_n) \in B_1, \quad \forall n \geqslant N_2. \tag{4.6}$$

By (4.6) we infer that there exists $(\psi_T, \phi_T, \theta_T) \in B_1$, such that

$$S(t_n - T)(\psi_n, \phi_n, \theta_n) \to (\psi_T, \phi_T, \theta_T), \quad \text{weakly in } H^1 \times H^1 \times H. \tag{4.7}$$

Following the proof of (3.14), by (4.7), we can get

$$S(T)(S(t_n - T)(\psi_n, \phi_n, \theta_n)) \to S(T)(\psi_T, \phi_T, \theta_T), \quad \text{weakly in } H^1 \times H^1 \times H. \tag{4.8}$$

It follows from (4.5) and (4.8) that

$$(\psi,\phi,\theta) = S(T)(\psi_T, \phi_t, \theta_T). \tag{4.9}$$

By the weak convergence (4.5), we get

$$\liminf_{n\to\infty} \|S(t_n)(\psi_n, \phi_n, \theta_n)\|_{H^1 \times H^1 \times H} \geqslant \|(\psi,\phi,\theta)\|_{H^1 \times H^1 \times H}. \tag{4.10}$$

If we can also prove that

$$\limsup_{n\to\infty} \|S(t_n)(\psi_n,\phi_n,\theta_n)\|_{H^1\times H^1\times H} \leqslant \|(\psi,\phi,\theta)\|_{H^1\times H^1\times H}, \tag{4.11}$$

then (4.5) and (4.10) and (4.11) will imply

$$S(t_n)(\psi_n,\phi_n,\theta_n)\to(\psi,\phi,\theta), \quad \text{strongly in } H^1\times H^1\times H.$$

So, the proof will be finished once (4.11) is verified. In what follows, we apply the energy equation (2.12) to prove (4.11). Here, the idea is analogous to Theorem 3.4. And, therefore, only the main steps will be given below.

Applying (3.16) to the solution $S(T)(S(t_n-T)(\psi_n,\phi_n,\theta_n))$, we find

$$\begin{aligned} E(S(t_n)(\psi_n,\phi_n,\theta_n)) &= E(S(T)(S(t_n-T)(\psi_n,\phi_n,\theta_n))) \\ &= e^{-2\delta T}E(S(t_n-T)(\psi_n,\phi_n,\theta_n)) \\ &\quad + \int_0^T e^{-2\delta(T-\tau)}F(S(\tau)(S(t_n-T)(\psi_n,\phi_n,\theta_n)))\,d\tau. \end{aligned} \tag{4.12}$$

Since (4.6) shows that $S(t_n-T)(\psi_n,\phi_n,\theta_n)$ is bounded, we derive from the definition of E in (2.13) that

$$e^{-2\delta T}E(S(t_n-T)(\psi_n,\phi_n,\theta_n)) \leqslant Ce^{-2\delta T}, \quad \forall n \geqslant N_2. \tag{4.13}$$

For the second term on the right-hand side of (4.12), using (4.7) and proceeding as the proof of (3.22) and (3.24), we can finally get

$$\limsup_{n\to\infty}\int_0^T e^{-2\delta(T-\tau)}F(S(\tau)(S(t_n-T)(\psi_n,\phi_n,\theta_n)))\,d\tau \leqslant \int_0^T e^{-2\delta(T-\tau)}F(S(\tau)(\psi_T,\phi_T,\theta_T))\,d\tau. \tag{4.14}$$

It follows from (4.12)–(4.14) that

$$\limsup_{n\to\infty} E(S(T)(S(t_n-T)(\psi_n,\phi_n,\theta_n))) \leqslant Ce^{-2\delta T} + \int_0^T e^{-2\delta(T-\tau)}F(S(\tau)(\psi_T,\phi_T,\theta_T))\,d\tau. \tag{4.15}$$

Again, applying (3.16) to the solution $(\psi,\phi,\theta)=S(T)(\psi_T,\phi_T,\theta_T)$, we find

$$\begin{aligned} E(\psi,\phi,\theta) &= E(S(T)(\psi_T,\phi_T,\theta_T)) \\ &= e^{-2\delta T}E(\psi_T,\phi_T,\theta_T) + \int_0^T e^{-2\delta(T-\tau)}F(S(\tau)(\psi_T,\phi_T,\theta_T))\,d\tau. \end{aligned} \tag{4.16}$$

It follows from (4.15) and (4.16) that

$$\begin{aligned} \limsup_{n\to\infty} E(S(t_n)(\psi_n,\phi_n,\theta_n)) &\leqslant Ce^{-2\delta T} - e^{-2\delta T}E(\psi_T,\phi_T,\theta_T) + E(\psi,\phi,\theta) \\ &\leqslant C_1e^{-2\delta T} + E(\psi,\phi,\theta) [\text{by } (\psi_T,\phi_T,\theta_T)\in B_1]. \end{aligned}$$

Let $T\to+\infty$, we find

$$\limsup_{n\to\infty} E(S(t_n)(\psi_n,\phi_n,\theta_n)) \leqslant E(\psi,\phi,\theta).$$

Then, by the definition of E in (2.13), and repeating the procedure of (3.27)–(3.29), we can easily obtain (4.11). And then, the proof is complete. □

We now prove that the dynamical system $S(t)$ is asymptotically compact in a more regular space if the driving terms are smoother.

Theorem 4.2: *Assume that f and g belong to* H^k *with* $k \geqslant 0$. *Then the dynamical system* $S(t)$ *is asymptotically compact in* $H^{k+2} \times H^{k+2} \times H^{k+1}$.

Proof: The proof of this theorem is quite similar to Theorem 4.1. In this case, the energy equation (2.22) and the bounded absorbing set B_{k+2} should be used instead of the energy equation (2.12) and the bounded absorbing set B_1. Since the idea is analogous, we will not repeat the details here. □

We are now in a position to state our main result.

Theorem 4.3: *Assume that f and g belong to H. Then problem (2.1)–(2.5) possesses a strong compact global attractor in* $H^1 \times H^1 \times H$.

Theorem 4.4: *Assume that f and g belong to* H^k *with* $k \geqslant 0$. *Then problem (2.1)–(2.5) possesses a strong compact global attractor in* $H^{k+2} \times H^{k+2} \times H^{k+1}$.

The proof of Theorems 4.3 and 4.4 is now obvious. Since we have established the existence of bounded absorbing sets and the asymptotic compactness for $S(t)$ in $H^1 \times H^1 \times H$ and $H^{k+2} \times H^{k+2} \times H^{k+1}$, respectively, then Theorems 4.3 and 4.4 follow from the abstract result (Ref. 9, Theorem 3.3) (see also Refs. 12 and 13).

We remark that by Theorem 4.3 we have the existence of the strong global attractor that is compact in the norm topology of $H^1 \times H^1 \times H$ when f belongs to H, which is weaker than the corresponding condition $f \in H^2$ in Ref. 5.

ACKNOWLEDGMENTS

This work was done when B. Wang was visiting Mathematisches Institut at the Universität zu Köln. He is grateful for their kind hospitality and financial support by a DAAD Fellowship.

[1] A. Bachelot, ''Probleme de Cauchy pour des systemes hyperboliques semi-lineaires,'' Ann. Inst. Henri Poincaré: Anal. Non Linéaire **1**, 453–478 (1984).

[2] N. Hayashi and W. von Wahl, ''On the global strong solutions of coupled Klein–Gordon–Schrödinger equations,'' J. Math. Soc. Jpn. **39**, 489–497 (1987).

[3] I. Fukuda and M. Tsutsumi, ''On coupled Klein–Gordon–Schrödinger equations III,'' Math. Japonica **24**, 307–321 (1979).

[4] I. Fukuda and M. Tsutsumi, ''On coupled Klein–Gordon–Schrödinger equations II,'' J. Math. Anal. Appl. **66**, 358–378 (1978).

[5] P. Biler, ''Attractors for the system of Schrödinger and Klein–Gordon equations with Yukawa coupling,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **21**, 1190–1212 (1990).

[6] Y. Li, ''Finite dimension of global attractor for weakly dissipative Klein–Gordon–Schrödinger equations,'' preprint.

[7] B. Guo and Y. Li, ''Attractor for dissipative Klein–Gordon–Schrödinger equations in R^3,'' J. Diff. Eqns. **136**, 356–377 (1997).

[8] J. M. Ball, ''A proof of the existence of global attractors for damped semilinear wave equations,'' to appear.

[9] J. M. Ball, ''Continuity properties and global attractors of generalized semiflows and Navier–Stokes equations,'' J. Nonlinear Sci. **7**, 475–502 (1997).

[10] J. M. Ghidaglia and R. Temam, ''Attractors for damped nonlinear hyperbolic equations,'' J. Math. Pures Appl. **66**, 273–319 (1987).

[11] A. V. Babin and M. I. Vishik, *Attractors of Evolution Equations* (Nauka, Moscow, 1989).

[12] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics* (Springer-Verlag, New York, 1988).

[13] J. K. Hale, *Asymptotic Behavior of Dissipative Systems* (American Mathematical Society, Providence, RI, 1988).

[14] J. M. Ghidaglia, ''A note on the strong convergence towards attractors of damped forced KdV equations,'' J. Diff. Eqns. **110**, 356–359 (1994).

Anisotropic fluid spherically symmetric space–times admitting a kinematic self-similarity

Patricia M. Benoit[a)]
Department of Mathematics Statistics and Computing Science, St. Francis Xavier University, Antigonish, Nova Scotia B2W 2G5, Canada

Alan A. Coley[b)]
Department of Mathematics Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada

(Received 25 March 1998; accepted for publication 30 September 1998)

Anisotropic fluid spherically symmetric space–times admitting a kinematic self-similar vector are investigated. The geodesic case is considered, and some special subcases in which the anisotropic fluid satisfies additional physical conditions are investigated in detail. A number of other special cases are studied. Particular attention is focused on the possible asymptotic behavior of the models, and it is shown that the models considered always asymptote towards an exact homothetic solution, which is in general either a perfect fluid model or a static solution. [S0022-2488(99)01602-3]

I. INTRODUCTION

In a recent paper spherically symmetric space–times which admit a kinematic self-similarity of the second (or zeroth) kind were studied when the source of the gravitational field was assumed to be a perfect fluid.[1] In that paper several particular subclasses of models were studied in depth, including the subcases ''$M_1=0$'' and ''$M_2=0$'' (which includes the static models as a further subcase). Note that these particular subcases refer to specific forms for the first integral *m(r,t)* of the EFEs. The precise definitions of these subcases in terms of *m(r,t)* are not necessary here; see Benoit and Coley[1] for more details. These subclasses of models, in which exact solutions were obtained, were found to be of particular interest since their qualitative properties were representative of the asymptotic behavior of more general models.

The metric, in comoving coordinates, is given by

$$ds^2=-e^{2\phi}dt^2+e^{2\psi}dr^2+r^2S^2d\Omega^2, \tag{1.1}$$

where the functions ϕ, ψ and S and depend only on the self-similarity coordinate

$$\xi=r(\alpha t)^{-1/\alpha}, \tag{1.2}$$

where α is the self-similar index (and we shall assume henceforward that $\alpha\neq 0$). The kinematic self-similar generator is given by[2]

$$\boldsymbol{\xi}=\xi^a\frac{\partial}{\partial x^a}=\alpha t\frac{\partial}{\partial t}+r\frac{\partial}{\partial r}. \tag{1.3}$$

It follows from (1.1) and (1.3) that

$$\mathcal{L}_{\boldsymbol{\xi}}h_{ab}=2h_{ab}, \tag{1.4}$$

[a)]Electronic mail: pbenoit@stfx.ca
[b)]Electronic mail: aac@mscs.dal.ca

and

$$\mathcal{L}_{\xi} u_a = \alpha u_a, \tag{1.5}$$

where $h_{ab}=g_{ab}+u_a u_b$ is the projection tensor. Hence $\boldsymbol{\xi}$, as given by (1.3) with $\alpha \neq 1$, is a *kinematic self-similar* vector and corresponds to the natural relativistic counterpart of self-similarity of the more general second kind.[3] We note that in the case $\alpha=1$, $\boldsymbol{\xi}$ is a homothetic vector, corresponding to self-similarity of the first kind. We shall adopt the notation and terminology of Benoit and Coley[1] and for brevity we shall not repeat here the motivation or the discussion of self-similarity given in that paper.

In this paper we shall generalize the perfect fluid solutions in Benoit and Coley[4] to the case of an anisotropic fluid, in which stress-energy tensor is given by

$$T_{ab} = \mu u_a u_b + p_{\parallel} n_a n_b + p_{\perp}(g_{ab} + u_a u_b - n_a n_b), \tag{1.6}$$

where u^a is the comoving fluid velocity vector and n^a is a unit spacelike vector orthogonal to u^a (i.e., $u_a n^a = 0$). The stress-energy tensor (1.6) possesses and eigenvalue degeneracy (and hence is not the most general anisotropic fluid stress-energy tensor) consistent with the assumption of spherical symmetry (see metric (1.1)). For the metric (1.1), $\mathbf{n}$ is given by

$$\mathbf{n} = n^a \frac{\partial}{\partial x^a} = e^{-\psi} \frac{\partial}{\partial r}. \tag{1.7}$$

Using Eqs. (1.1)–(1.3), it therefore follows immediately that

$$\mathcal{L}_{\xi} n_a = n_a \tag{1.8}$$

is satisfied identically, so that the form for $\mathbf{n}$ is consistent with the similarity assumption. The scalars $p_{\parallel}$ and $p_{\perp}$ are the pressures parallel to and perpendicular to n^a, respectively, and μ is the energy-density. The perfect fluid case corresponds to the case in which $p_{\parallel} = p_{\perp}$.

Fluids with an anisotropic pressure have been studied for many reasons (see the discussion in Coley and Tupper).[5] For example, in several cases in which the stress-energy tensor is more general than that of a perfect fluid (due to, e.g., a two perfect fluid source, an imperfect fluid source or in the region of interaction of two colliding plane impulsive gravitation waves), the energy-momentum tensor is formally of the form (1.6). In particular, a strong magnetic field in a plasma in which the particle collision density is low can cause the pressure along and perpendicular to the magnetic field lines to be unequal.[6] If the source of the gravitational field can be represented by the sum of a perfect fluid and a local magnetic field $H^a = H n^a$ (as measured by u^a), then the stress-energy tensor can be written in the form (1.6) with

$$\mu = \overline{\mu} + \pi, \quad p_{\parallel} = \overline{p} - \pi, \quad p_{\perp} = \overline{p} + \pi, \tag{1.9}$$

where $\pi = \frac{1}{2}\lambda H^2$ and λ is the magnetic permeability. Other possible sources of anisotropic stresses, in addition to cosmological magnetic and electric fields, include, for example, populations of collisionless particles like gravitons,[7] photons[8] or relativistic neutrinos,[9] Yang–Mills fields,[10] axion fields in low-energy string theory,[11] long wavelength gravitational waves,[12] and topological defects like global monopoles, cosmic strings, and domain walls.[13–15]

Most anisotropic models that have been studied are also spherically symmetric (see references cited in Ref. 5), and have applications especially in relativistic astrophysics (e.g., stellar models); in particular, static anisotropic spheres have received much attention.[5] In addition, such models with additional symmetries, including homothetic vectors and conformal Killing vectors, have also been studied (see Refs. 6, 5, and references within).

For the metric (1.1) the Einstein field equations (EFEs) yield the following expressions for the physical variables:

$$\mu = \frac{W^1(x)}{r^2} + \frac{W^2(x)}{t^2},$$

$$p_\parallel = \frac{P_\parallel^1(x)}{r^2} + \frac{P_\parallel^2(x)}{t^2}, \tag{1.10}$$

$$p_\perp = \frac{P_\perp^1(x)}{r^2} + \frac{P_\perp^2(x)}{t^2},$$

where

$$W^1(x) = \frac{1}{S^2} - e^{-2\psi}[(1+y)^2 + 2y\dot{\phi}],$$

$$W^2(x) = \frac{e^{-2\phi}}{\alpha^2} y[y + 2\dot{\psi}],$$

$$P_\parallel^1(x) = -\frac{1}{S^2} + e^{-2\psi}(1+y)[1+y+2\dot{\phi}],$$

$$P_\parallel^2(x) = -\frac{e^{-2\phi}}{\alpha^2}[2\dot{y} + 2\alpha y + 3y^2 - 2y\dot{\phi}],$$

$$P_\perp^1(x) = e^{-2\psi}[2y\dot{\phi} + \dot{\phi}^2 + \ddot{\phi} - \dot{\phi}\dot{\psi}],$$

$$P_\perp^2(x) = -\frac{e^{-2\phi}}{\alpha^2}[(\alpha-1)y + 2y\dot{\psi} + \dot{\psi} + \alpha\dot{\psi} + \dot{\psi}^2 + \ddot{\psi} - \dot{\phi}\dot{\psi}], \tag{1.11}$$

and where $y \equiv \dot{S}/S$, $x \equiv \ln \xi$ and $\dot{f} = df/dx$. The final EFE (that ensures that the Einstein tensor is diagonal) becomes

$$\dot{y} = y\dot{\phi} + (\dot{\psi} - y)(1+y). \tag{1.12}$$

Clearly there exists a variety of anisotropic fluid spherically symmetric kinematic self-similar space–times satisfying Eqs. (1.10)-(1.12).

If we assume that the physical quantities also obey similarity conditions of the form

$$\mathcal{L}_\xi \mu = a\mu, \quad \mathcal{L}_\xi p_\parallel = b_\parallel p_\parallel, \quad \mathcal{L}_\xi p_\perp = b_\perp p_\perp, \tag{1.13}$$

where a, $b_\parallel$ and $b_\perp$ are constants, then it can easily be shown that:

$$\text{(i)} \quad W^1 = 0 \quad \textit{or} \quad W^2 = 0$$

and

$$\text{(ii)} \quad P_\parallel^1 = 0 \quad \textit{or} \quad P_\parallel^2 = 0$$

and

$$\text{(iii)} \quad P_\perp^1 = 0 \quad \textit{or} \quad P_\perp^2 = 0.$$

The special subcases $W^i=0$ with either $P_\parallel^i\neq 0$ or $P_\perp^i\neq 0$ $(i=1,2)$ are not of physical interest. The special subcase $W^1=P_\parallel^1=P_\perp^1=0$ corresponds to the special subcase ''$M_1=0$.'' Finally, the special subcase $W^2=P_\parallel^2=P_\perp^2=0$ is related to the special subcase ''$M_2=0$,'' and the static models are included within this subclass of models.

It turns out (Benoit and Coley,[1] in particular, see the Appendix therein) that all *static* spherically symmetric kinematic self-similar solutions belong to the subclass ''$M_2=0$,'' regardless of the form of the stress-energy tensor, and, moreover, that all such static spacetimes necessarily admit a *homothetic vector*. Consequently, no new static anisotropic solutions can be obtained that admit a proper kinematic self-similarity. Hence we shall concentrate here on the special subcase ''$M_1=0$.''

II. GEODESIC MODELS

The geodesic case, in which the acceleration of the comoving fluid velocity vector is zero, is characterized by $\dot{\phi}=0$ and is equivalent to the special subcase ''$M_1=0$'' considered in Benoit and Coley.[1] In this model, Eq. (1.12) gives (for $S+\dot{S}\neq 0$)

$$e^{2\phi}=1,\quad \dot{\psi}=\frac{\dot{S}+\ddot{S}}{S+\dot{S}}=\frac{\dot{y}+y^2+y}{1+y},\tag{2.1}$$

whence the metric (1.1) becomes

$$ds^2=-dt^2+(S+\dot{S})^2dr^2+r^2S^2d\Omega^2.\tag{2.2}$$

Assuming the first of conditions (2.1), the second condition guarantees the resulting Einstein tensor is diagonal and hence the remaining EFEs simply yield the following expressions for μ, $p_\parallel$ and $p_\perp$:

$$\mu=W(x)t^{-2},\quad p_\parallel=P_\parallel(x)t^{-2},\quad p_\perp=P_\perp(x)t^{-2},\tag{2.3}$$

(where we have now omitted the index ''2'' for convenience), so that Eqs. (1.13) are automatically satisfied with $a=b_\parallel=b_\perp=-2\alpha$, where

$$W(x)\equiv\frac{y}{\alpha^2(1+y)}(3y+3y^2+2\dot{y}),$$

$$P_\parallel(x)\equiv-(3y^2+2\alpha y+2\dot{y})/\alpha^2,\tag{2.4}$$

$$P_\perp(x)\equiv-\frac{(1+y)(2\dot{y}+3y^2+2\alpha y)+3y\dot{y}+\alpha\dot{y}+\ddot{y}}{\alpha^2(1+y)}.$$

Equations (2.2)–(2.4) represent a class of anisotropic fluid solutions depending upon the arbitrary function $S(x)$.

We note that the following relationships result from the definitions given in Eqs. (2.3):

$$P_\perp=P_\parallel+\frac{\dot{P}_\parallel}{2(1+y)},$$

$$W=\frac{-y[(2\alpha-3)y+\alpha^2P_\parallel]}{\alpha^2(1+y)}.$$

A. Perfect fluid models

In the perfect fluid case we have that $P_{\parallel}=P_{\perp}$, and hence from Eqs. (2.4) we obtain the following differential equation for the function $y(x)$ [and hence $S(x)$] in the metric (2.2):

$$2\dot{y}+3y^2+2\alpha y+\alpha^2 p_0=0. \tag{2.5}$$

In Eq. (2.5) p_0 is an arbitrary integration constant. In the perfect fluid case μ is obtained from Eqs. (2.3) and (2.4) and we have that

$$p=p_0 t^{-2}, \tag{2.6}$$

and hence the significance of p_0 is that it constitutes a dimensional constant (appearing in the pressure) characterizing the physical problem; this property is characteristic of self-similarity of the second kind.[3] It can be shown that these perfect fluid solutions (for $\alpha\neq 1$) cannot, in general, admit any homothetic vectors.[4]

The perfect fluid solutions were studied in detail in Benoit and Coley;[1] in fact, exact solutions were obtained and the qualitative properties of the whole class of models were studied. In particular, in the pressure-free case we obtain the exact dust solution of the Tolman family studied by Lynden-Bell and Lemos[16] and Carter and Henriksen,[2] and we found that all solutions are asymptotic to exact, power-law (flat) FRW models (which admit a homothety).

B. Solutions with $S+\dot{S}=0$

In Benoit and Coley (1998) we showed that the case $S+\dot{S}=0$, which implies that $S=s_0 e^{-x}$, could be factored out of the analysis as it could not lead to a perfect fluid solution. For that reason, we consider it as a special case here. (This case is not contained in the geodesic models studied above.)

When $S=s_0 e^{-x}$ (i.e., $y=-1$), the EFEs yield

$$\dot{\phi}=0, \tag{2.7}$$

whence we can choose coordinates so that $e^{2\phi}=1$, and

$$\mu=s_0^{-2}e^{2x}r^{-2}+(1-2\dot{\psi})\alpha^{-2}t^{-2}, \tag{2.8}$$

$$p_{\parallel}(x)=-s_0^2 e^{2x}r^{-2}+(2\alpha-3)\alpha^{-2}t^{-2}, \tag{2.9}$$

$$p_{\perp}(x)=-[(1-\alpha)(1-\dot{\psi})+\dot{\psi}^2+\ddot{\psi}]\alpha^{-2}t^{-2}. \tag{2.10}$$

The fluid described by these equations will further satisfy Eq. (1.13) in one of two cases. Either (i) $\alpha=1$, and the solution admits a homothetic vector, or (ii) $\dot{\psi}=1/2$, $\alpha=3/2$.

In the first case, i.e., $\alpha=1$, the solution is given by

$$ds^2=-dt^2+e^{2\psi}dr^2+s_0 t^{-2}d\Omega^2, \tag{2.11}$$

with

$$\mu=(s_0^{-2}+1-2\dot{\psi})t^{-2}, \tag{2.12}$$

$$p_{\parallel}=-(s_0^{-2}+1)t^{-2}, \tag{2.13}$$

$$p_{\perp}=-(\dot{\psi}^2+\ddot{\psi})t^{-2}, \tag{2.14}$$

where the function $\psi(x)$ is arbitrary.

In the second case the solution is given (after a coordinate redefinition) by

$$ds^2=-dt^2+t^{-2/3}dr^2+t^{4/3}d\Omega^2, \tag{2.15}$$

with

$$\mu=\mu_0 t^{-4/3}, \tag{2.16}$$

$$p_\parallel=-\mu, \tag{2.17}$$

$$p_\perp=0, \tag{2.18}$$

where μ_0 is a constant. It can be easily shown that the metric (2.15) does not admit a proper homothetic vector. Curiously, cosmic strings satisfy ''equations of state'' of the form $\mu+p_\parallel=0$, $p_\perp=0$.[17]

III. SPECIAL CASES

There are a variety of models which satisfy additional constraints. We consider here two such models.

A. Case A: Dimensional constants

If we assume that $P_\perp=p_0$, a constant, then Eqs. (2.4) yield

$$\dot{P}_\parallel(x)=2(1+y)(p_0-P_\parallel(x)). \tag{3.1}$$

This equation can be integrated to yield

$$P_\parallel(x)=p_0+ce^{-2x}S^{-2}, \tag{3.2}$$

where c is an arbitrary constant. Using this expression for $P_\parallel$, we obtain

$$W(x)=\frac{y}{\alpha^2(1+y)}[y(3-2\alpha)-\alpha^2 p_0-c\alpha^2 e^{-2x}S^{-2}], \tag{3.3}$$

and the differential equation

$$2\dot{y}+3y^2+2\alpha y=-\alpha^2 p_0-\alpha^2 ce^{-2x}S^{-2}. \tag{3.4}$$

Note that when $c=0$ (i.e., $P_\parallel=P_\perp=p_0$, corresponding to a perfect fluid) Eq. (3.4) is related to Eq. (2.56) in Benoit and Coley.[1]

If we had begun the analysis of this section with the assumption that $P_\parallel=p_0$, then Eqs. (2.4) automatically imply that $P_\parallel=P_\perp=p_0$, the perfect fluid case considered in Benoit and Coley.[1]

The pressures $p_\parallel$ and $p_\perp$ are positive if the constants p_0 and c are non-negative. The energy conditions will constrain these constants further (for a given value of α) through (3.3).

B. Case B: Equations of state

We can also consider the subclass of solutions which satisfy equations of state of the form:

$$p_\parallel=f_\parallel(\mu), \quad p_\perp=f_\perp(\mu), \tag{3.5}$$

for arbitrary functions $f_\parallel$ and $f_\perp$. From Eqs. (2.3), conditions (3.5) automatically yield

$$p_\parallel=c_\parallel\mu \quad \text{and} \quad p_\perp=c_\perp\mu, \tag{3.6}$$

where $c_\parallel$ and $c_\perp$ are constants. Substituting these conditions into the definitions (2.4) then yields

$$\mu = \mu_0 t^{-2}[Se^x]^{-2(1-c_\perp/c_\parallel)} \tag{3.7}$$

and the differential equation for y:

$$2\dot{y} + 3y^2 + 2\alpha y = -\alpha^2 c_\parallel \mu_0 [Se^x]^{-2(1-c_\perp/c_\parallel)}. \tag{3.8}$$

Once again we note that when $c_\parallel = c_\perp$ (i.e., the perfect fluid case), we recover Eq. (2.5) as expected.

A positive value for the constant μ_0 guarantees that the energy density is positive. If $|c_\parallel| \leqslant \mu_0$ and $|c_\perp| \leqslant \mu_0$, the energy conditions are satisfied. The pressures are non-negative if $c_\parallel \geqslant 0$ and $c_\perp \geqslant 0$.

IV. ANALYSIS OF SPECIAL CASES

The behavior of each of the special cases derived in Sec. III can be studied qualitatively since each of the ordinary differential equations governing the model is autonomous.

The special cases A(dimensional constants) and B(equations of state) can be considered simultaneously using the following change of variables:

$$\nu = b[Se^x]^{-2n}, \tag{4.1}$$

where b is a non-negative constant. The resulting system is then

$$\dot{y} = -\tfrac{1}{2}(3y^2 + 2\alpha y + k + \nu), \tag{4.2}$$

$$\dot{\nu} = -2n\nu(1+y). \tag{4.3}$$

Using these definitions, case A is characterized by $n=1$, $k=\alpha^2 p_0$, and case B is characterized by $n = 1 + c_\perp/c_\parallel$ and $k=0$.

It is important to note that the invariant set $\nu=0$ of Eqs. (4.2)/(4.3) defines the perfect fluid solutions. We also note that $y=0$ represents the static solutions. Each of these cases is examined in detail in Benoit and Coley.[1]

If we consider only the case of positive pressures and positive energy density, we can impose the necessary (though not necessarily sufficient) condition that the parameters in our equations must satisfy $k \geqslant 0$, $n \geqslant 1$ and $\nu \geqslant 0$. With these restrictions, we find that there are at most three singular points at finite values. We note that $\nu=0$ is an invariant set of the system (4.2)/(4.3), as is the set $\nu>0$. As a result we need only consider the dynamics (and hence the singular points) in the half-plane $\nu \geqslant 0$.

The finite singular points (y_0, ν_0) are given by:

$$Q_1 = (\tfrac{1}{3}(-\alpha + (\alpha^2 - 3k)^{1/2}, 0),$$

$$Q_2 = (\tfrac{1}{3}(-\alpha - (\alpha^2 - 3k)^{1/2}, 0),$$

$$Q_3 = (-1, 2\alpha - 3 - k).$$

The nature of these singular points, which can be determined using standard techniques,[18] depends upon the relationship between the parameters α and k. The results are summarized in Table I. Note that only those singular points which are located in the physical phase space are listed in this table. It is important to note that each of the cases I–IV is possible when considering the Eqs. (4.2)/(4.3) in case A. In case B, however, we find that only the cases labeled (I) and (II) in Table I yield consistent constraints on the parameter α.

TABLE I. Summary of the nature of the finite singular points for the system (4.2)/(4.3). ''N/A'' indicates that the given point is not located in the physical region $\nu \geqslant 0$. The two cases (i) $\alpha^2=3k$, $2\alpha-3>k$ and (ii) $\alpha^2<3k$, $2\alpha-3\geqslant k$ are omitted since they do not give any real solutions for k and α.

	$\alpha^2>3k$		$\alpha^2=3k$	$\alpha^2<3k$
	$2\alpha-3\leqslant k$ I	$2\alpha-3>k$ II	$2\alpha-3\leqslant k$ III	$2\alpha-3<k$ IV
$\mathbf{Q}_1$	sink	sink	saddle-node	N/A
$\mathbf{Q}_2$	source	saddle	($\equiv \mathbf{Q}_1$)	N/A
$\mathbf{Q}_3$	saddle	N/A	N/A	N/A

We can complete the qualitative analysis of these two cases by considering the stability of the singular points at infinity. To perform the analysis at infinity, we apply the following Poincaré transformation to our system (4.2)/(4.3) in order to compactify the phase space:

$$Y=\frac{y}{(1+y^2+\nu^2)^{1/2}}, \quad V=\frac{\nu}{(1+y^2+\nu^2)^{1/2}}. \tag{4.4}$$

In these new variables, the phase space has been compactified to the region $\Theta^2=1-(Y^2+V^2)\geqslant 0$, and all infinite points of the original system are found on the boundary $\Theta=0$. The restriction that $\nu\geqslant 0$ implies that $V\geqslant 0$, and all finite singular points remain at finite values of Y and V and are of the same sign in the new coordinate system (Y,V).

The transformed Eqs. (4.1)/(4.2) are then given by:

$$Y'=\tfrac{1}{2}(4n-3)Y^2V^2+\Theta[\tfrac{1}{2}Y+(2n-\alpha)V]-\tfrac{1}{2}\Theta^2[3Y^2+kV^2]-\Theta^3[\alpha Y+\tfrac{1}{2}V]-\tfrac{1}{2}k\Theta^4, \tag{4.5}$$

$$V'=-\tfrac{1}{2}(4n-3)Y^3V+\Theta YV[\tfrac{1}{2}V-(2n-\alpha)Y]+\tfrac{1}{2}\Theta^2YV[k-4n]-2nV\Theta^3, \tag{4.6}$$

where $f'=\Theta\dot{f}$. There are four singular points at infinity located on the boundary $Y^2+V^2=1$, which are given by

$$R_\pm=(0,\pm 1), \quad S_\pm=(\pm 1,0). \tag{4.7}$$

The points $S_\pm$ correspond to perfect fluid solutions, and $R_\pm$ correspond to static solutions. A local stability analysis shows that the points $S_\pm$ are both saddles. R_+ is a nonhyperbolic point containing both stable and unstable manifolds for all values of α and k. The stable manifold of R_+ lies in an elliptic sector of R_+ and corresponds to homoclinic orbits. The fixed point R_- is not in the physical phase space.

The phase portraits in the compactified phase space $(V^2+Y^2\leqslant 1, V\geqslant 0)$ are given in Fig. 1. From these portraits it is immediately evident that the only stable singular points (both to the past and the future) either lie in the $V=0$ invariant set, occur at the infinite singular point R_+, or occur at $\mathbf{Q}_3$ (when it exists in the phase space). Recall that the invariant set $V=0$ represents the perfect fluid solutions studied previously,[1] where in the equivalent ''$M_1=0$'' case the solutions were shown to asymptote towards a flat FRW model. The fixed point R_+ has $y=0$, and hence is a static solution. Finally, the fixed point $\mathbf{Q}_3$ has the property $y=-1$ (or $S+\dot{S}=0$), which was examined in Sec. II B. Since all of the solutions in the phase space, and in particular those asymptoting to the point $\mathbf{Q}_3$, have the property that $p_\parallel=P_\parallel(x)t^{-2}$, $p_\perp=P_\perp(x)t^{-2}$, and $\mu=W(x)t^{-2}$, by continuity so must the solution at $\mathbf{Q}_3$. Therefore the solution represented by the point Q_3 must be given by the metric (2.11).

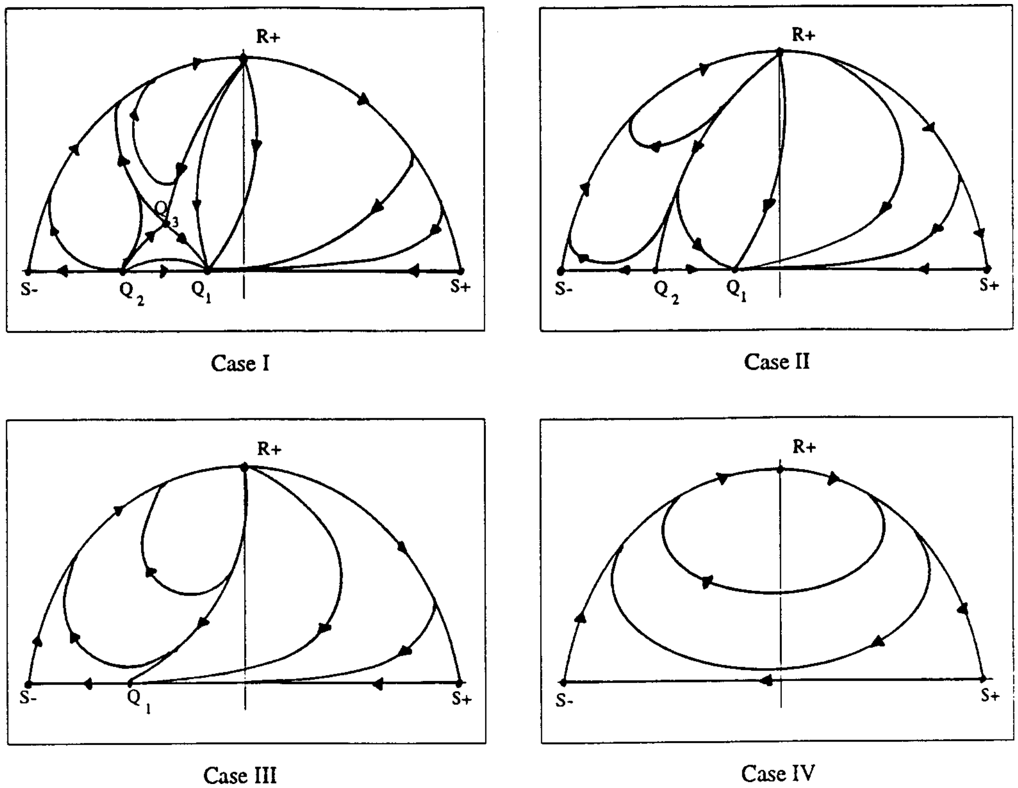

FIG. 1. Phase portraits. The phase portraits for the system (4.5)/(4.6) for various ranges of values of α and k are given, where the particular cases are as listed in Table I.

Consequently we see that in the analysis of the two cases considered in Sec. III A and III B above the asymptotic behavior is described by either a flat FRW perfect fluid model, a static model, or by that of the metric (2.11). In all cases these exact asymptotic models admit a homothetic vector.

V. DISCUSSION

We note that in the cases studied in this paper the dynamics of the models is governed by a system of the form:

$$\dot{y} = -\tfrac{1}{2}(3y^2 + 2\alpha y) + f(\nu), \tag{5.1}$$

$$\dot{\nu} = -2n\nu(1+y). \tag{5.2}$$

The variable ν is defined by Eq. (4.1) and the function $f(\nu)$ depends on the specific case being studied. In the cases considered in Sec. III we had that:

Case A: Dimensional Constants : $f(\nu) = -\tfrac{1}{2}(\nu + \alpha^2 p_0)$.

Case B: Equations of State : $f(\nu) = -\tfrac{1}{2}\nu$.

The system (5.1)/(5.2) results whenever we impose the condition

$$P_{\parallel}(x) = -2\alpha^{-2} f(\nu). \tag{5.3}$$

In the cases examined in Sec. III it was shown that all solutions necessarily asymptote to an exact solution admitting a homothetic vector. It is of interest to consider whether there are any possible asymptotic states for the geodesic anisotropic models which satisfy Eq. (5.3) that do *not* admit a homothetic vector.

As was the case in Sec. IV, the perfect fluid solutions are located in the invariant set $\nu=0$. The definition of ν requires that it be greater than or equal to zero. In the relevant phase space there are then (at most) three finite singular points of the system (5.1)/(5.2). These singular points, equivalent to those studied in Sec. IV, are given by:

$$Q_1=(\tfrac{1}{3}[-\alpha+(\alpha^2+6f(0))^{1/2}],0),$$

$$Q_2=(\tfrac{1}{3}[-\alpha-(\alpha^2+6f(0))^{1/2}],0),$$

$$Q_3=(-1,f^{-1}(3/2-\alpha)).$$

The singular points Q_1 and Q_2 represent perfect fluid models, and Q_3 (as in Sec. IV) is represented by the metric (2.11). In each case the model represented by the finite singular point admits a homothetic vector.

The only possibility for the asymptotic behavior not to be governed by an exact homothetic model is then (i) the model is represented asymptotically by a periodic orbit in the phase space, or (ii) the model is represented by a singular point at infinity not located on one of the coordinate axes $\nu=0$ or $y=0$.

In the first case we can impose necessary conditions for the existence of a periodic orbit. Any periodic orbit in a plane must necessarily enclose a singular point. As a result we must have that the point Q_3 is in the phase space in which case we necessarily have that $f^{-1}(3/2-\alpha)$ is positive. The energy conditions requiring that the pressures and density are positive will result in the further condition that $f(\nu)\leqslant 0$, and therefore $\alpha\geqslant 3/2$ and $y\geqslant 0$. We consider the existence of a periodic orbit which encloses Q_3 by examining the horizontal and vertical isoclines of the system (5.1)/(5.2). The horizontal isoclines are located at (i) $\nu=0$, an invariant line, and (ii) $y=-1$. The second case indicates that if there exists a periodic orbit about the point Q_3 then there must be vertical isoclines on either side of the line $y=-1$. Solving Eq. (5.2), we find that the vertical isoclines are given by

$$y_{\pm}=\tfrac{1}{3}(-\alpha\pm(\alpha^2+3f(\nu))^{1/2}). \tag{5.4}$$

Imposing the energy conditions $f(\nu)\leqslant 0$ and $\alpha\geqslant 3/2$, we find that the y-values of the vertical isoclines must satisfy

$$-1\leqslant y\leqslant 0; \tag{5.5}$$

i.e., $y_{\pm}$ cannot take on values less than -1. Therefore, there can be no periodic orbits enclosing the point Q_3 if the energy conditions are to be satisfied.

If there is an asymptotic solution at infinite values of y and/or ν which is not homothetic then the corresponding singular point at infinity must be such that $y\neq 0$ or $\nu\neq 0$. This will occur when $\lim_{\nu\to\infty}f(\nu)\nu^{-2}\neq 0$. In such cases the infinite fixed point may represent a nonhomothetic asymptotic solution. Therefore, geodesic models for which Eq. (5.3) and the energy conditions are satisfied will not admit a nonhomothetic asymptotic solution whenever $\lim_{\nu\to\infty}f(\nu)\nu^{-2}$ is exactly zero.

VI. OTHER MODELS

Additional anisotropic fluid models can be investigated. For example, we can consider the case in which the source is a combination of *a perfect fluid and a magnetic field* satisfying Eqs. (1.9). Assuming $\overline{p}=(\gamma-1)\overline{\mu}$ (where γ is a constant), in the geodesic case we can immediately derive the governing system as:

$$\dot{y}=-\tfrac{1}{2}(3y^2+2\alpha y)-\tfrac{1}{2}\alpha^2\eta, \tag{6.1}$$

$$\dot{\eta}=-4(1+n)y\eta-4(n-1)(3-2\alpha)\alpha^{-2}y^2, \tag{6.2}$$

where $\eta\equiv-\alpha^{-2}(3y^2+2\alpha y+2\dot{y})=P_\parallel$ and $n\equiv 1/\gamma$. The system (6.1)/(6.2) is of a similar form to Eqs. (5.1) and (5.2) and can be analyzed using similar techniques. In the special cases $\gamma=1$ ($n=1$) and $\alpha=3/2$, Eq. (6.2) can be integrated immediately and exact solutions can be obtained. We note that at the equilibrium points of the system (6.1)/(6.2), $P_\parallel=$ constant ($\dot{P}_\parallel=0$), and hence from Eqs. (1.9), (2.3) and (2.4) we have that

$$\pi=\frac{1}{2}(p_\perp-p_\parallel)=\frac{\dot{P}_\parallel}{2t^2(1+y)}=0; \tag{6.3}$$

hence these equilibrium points correspond to perfect fluid models.

However, in order to study the physics of this particular model we note that $\pi=\lambda H^2/2$ and Eqs. (6.1) and (6.2) need to be supplemented by an additional differential equation (for H, derived from Maxwell's equations) and an assumption on the form of the magnetic permeability, λ.

Finally, we note that in the case in which $\pi=\text{constant}=\pi_0$ (with an unrestricted equation of state) it can be shown that the governing equations reduce to

$$\dot{y}=-\tfrac{1}{2}(3y^2+2\alpha y)-\alpha^2\pi_0\ln(\nu), \tag{6.4}$$

$$\dot{\nu}=-\nu(1+y). \tag{6.5}$$

This system is of the same form as that of (5.1)/(5.2) with $f(\nu)=-\alpha^2\pi_0\ln(\nu)$ and with the constant $n=\frac{1}{2}$. Since (6.3)/(6.4) is of the same form we can immediately conclude that the only asymptotic states of the system necessarily admit a homothetic vector. Note that in this case $f(\nu)$ is not analytic at $\nu=0$; however the physical phase space has $\nu>0$.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Department of Mathematics, Statistics and Computing Science and the Faculty of Graduate Studies at Dalhousie University, and the Sumner Foundation.

[1] P. M. Benoit and A. A. Coley, Class. Quantum Grav. **15**, 2397 (1998).
[2] B. Carter and R. N. Henriksen, Ann. Phys. **14**, 47 (1989).
[3] G. E. Barrenblatt and Ya. B. Zeldovich, Annu. Rev. Fluid Mech. **4**, 285 (1972).
[4] P. M. Benoit and A. A. Coley, in *Proceedings of the Sixth Canadian Conference on General Relativity and Relativistic Astrophysics*, the Fields Institute Communication Series (AMS), edited by S. P. Braham, J. D. Gegenberg, and R. J. McKellar (Providence, 1996).
[5] A. A. Coley and B. O. J. Tupper, Class. Quantum Grav. **11**, 2553 (1994).
[6] R. Maartens, D. P. Mason, and M. Tsamparlis, J. Math. Phys. **27**, 2987 (1986).
[7] N. Lukash and A. A. Starobinskii, Sov. Phys. JETP **39**, 742 (1974).
[8] W. Press, Appl. J. **205**, 311 (1976).
[9] A. G. Doroshkevich, Y. B. Zeldovich, and I. D. Novikov, Sov. Phys. JETP **26**, 408 (1968).
[10] B. K. Darian and H. P. Künzle, Class. Quantum Grav. **12**, 2651 (1995).
[11] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, 1987), Vol. I.
[12] V. N. Lukash, Nuovo Cimento B **35**, 269 (1976).

[13] M. Barriola and A. Vilenkin, Phys. Rev. Lett. **63**, 341 (1989).
[14] A. Vilenkin and P. Shellard, *Cosmic Strings and Other Topological Defects* (Cambridge University Press, Cambridge, 1994).
[15] J. Stachel, Phys. Rev. D **21**, 2171 (1980).
[16] D. Lynden-Bell and J. P. S. Lemos, Mon. Not. R. Astron. Soc. **233**, 197 (1988).
[17] A. Vilenkin, Phys. Rev. D **23**, 851 (1981).
[18] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields* (Springer, New York, 1983).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

The generalized thin-sandwich problem and its local solvability

Domenico Giulini[a)]
Institut für Theoretische Physik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

(Received 5 October 1998; accepted for publication 18 January 1999)

We consider Einstein gravity coupled to matter consisting of a gauge field with any compact gauge group and minimally coupled scalar fields. We investigate under what conditions a free specification of a spatial field configuration and its time derivative allows us to solve the constraints for lapse, shift, and other gauge parameters and hence determine a solution to the field equations (thin-sandwich problem). We establish sufficient conditions under which the thin-sandwich problem can be solved locally in field space. © *1999 American Institute of Physics.* [S0022-2488(99)00305-9]

I. INTRODUCTION

In this paper we consider the initial value problem for Einstein gravity plus matter in space–times $\Sigma\times\mathbf{R}$, where Σ is a closed orientable three-manifold. We are interested in the question of how to find initial data which satisfy the constraints. The most popular approach here is a powerful method devised by Lichnerowicz, Choquet-Bruhat, York, and others, henceforth referred to as the ''conformal method.'' (See Ref. 1 for a brief review and Ref. 2 for more details.) Of the gravitational variables it allows us to freely specify the conformal class of the initial three-metric, the conformally rescaled transverse-traceless components of the extrinsic curvature, and a constant trace thereof (i.e., Σ must have constant mean curvature). Given these data, the constraints turn into a quasilinear elliptic system of second order for the conformal factor (scalar function) and the transverse momentum (vector field), which decouples due to the constant mean-curvature condition. The disadvantages of this method are that it does not easily generalize to data of variable mean curvature and, more important for us, that it does not allow us to control the local scales of the physical quantities initially, since the freely specifiable data (gravitational and non-gravitational) are related to the actual physical quantities by some rescalings with suitable powers of the conformal factor. In particular, one has no control over the conformal part of the initial three-geometry.

In this paper we are concerned with the so-called ''thin-sandwich method,'' which differs from the one just mentioned insofar as it aims to define solutions to the Einstein equations by a *free* specification of the initial field configuration and its coordinate-time-derivative. The constraints are now read as equations for the gauge parameters (lapse, shift,...). From the conformal point of view this means that one tries to trade in the freedom to specify the initial gauge parameters for the freedom to specify the conformal part of the metric and the longitudinal part of the momentum. The disadvantages mentioned above would then be overcome, but unfortunately the equations (for the gauge parameters) turn out to be nonelliptic in general.[3,4] (By ellipticity of nonlinear differential operators we mean the ellipticity of its linearization, which depends on the point (in field space) about which one linearizes. The usual statement that the thin-sandwich equations are not elliptic merely asserts the existence of points where the linearization is not elliptic, but not that the domain of ellipticity is empty.) However, for certain open subsets of initial data they are elliptic and can be locally solved. This was first shown in Ref. 5 and will here be

a)Electronic mail: giulini@physik.unizh.ch

generalized to the full system including dynamical matter in the form of scalar and vector gauge fields.

We note that historically this approach arose from the question (then formulated as a conjecture) of whether the specification of two three-geometries uniquely determine an interpolating Einstein space–time (thick-sandwich problem). For nearby geometries infinitesimally close in time this turns into the thin-sandwich problem (see Ref. 6, Chap. 4, and Ref. 7) which we now describe in more detail.

In a space–time neighborhood $\Sigma\times\mathbf{R}$ of the Cauchy surface Σ, we use the standard parametrization of the space–time metric $g^{(4)}$,

$$g^{(4)}=-\alpha^2 dt\otimes dt+g_{ab}(dx^a+\beta^a dt)\otimes(dx^b+\beta^b dt), \tag{1.1}$$

where g is the (t-dependent) Riemannian metric of Σ and α and β are (t-dependent) scalar and vector fields on Σ, known as lapse and shift. The extrinsic curvature reads

$$K=\frac{1}{2\alpha}(\partial_t-L_\beta)g, \tag{1.2}$$

where L_β denotes the Lie derivative along β. Written in terms of (g,K), the constraints do not depend on α and β and hence constrain the set of allowed values for the data (g,K). The constraints read

$$K_{ab}K^{ab}-(K_a^a)^2-R=-2T_{\perp\perp}, \tag{1.3}$$

$$\nabla^b(K_b^a-\delta_b^a K_c^c)=T_\perp^a, \tag{1.4}$$

where T is the energy-momentum tensor of the matter and $\perp$ denotes the component along the future pointing normal n of Σ.

Alternatively, one may write the constraints in terms of g and $\dot g:=\partial_t g$ by replacing K via (1.2). Then they explicitly involve α and β and one may ask whether it is possible to *freely* specify $(g,\dot g)$ and let the constraints determine α and β. This is the thin-sandwich problem (TSP). If we abbreviate $\Psi:=(g,\dot g)$, $X:=(\alpha,\beta)$, the constraints take the form of the thin-sandwich equation (TSE):

$$F[\Psi,X]=0. \tag{1.5}$$

The TSP now asks for existence and uniqueness of solutions of the TSE, read as equation for X given Ψ. Once this is solved, we can construct (g,K) satisfying (1.3) and (1.4), which uniquely determine space–time via the Einstein evolution equations.[2]

It has long been discussed in the literature that, in general, existence and uniqueness should fail,[3] although some arguments merely showed this under the additional assumption (*a priori*) of constant lapse function α (Ref. 4) (see also Ref. 8 for a related issue). However, more recently it was shown that given a solution (Ψ,X) of the TSE, which satisfies certain bounds on geometric quantities and which admits no nontrivial solutions of the spatially projected Killing equation, there exist unique solutions $X(\Psi')$ for all Ψ' in a neighborhood of Ψ. This was achieved by an implicit function theorem for a reduced version of (1.5) with already eliminated lapse function.

Note that in this formulation of the TSP the right-hand sides of the constraints, that is $T_{\perp\perp}$ and $T_\perp^a$, are assumed given. These are *not* the components of T that an observer along ∂_t would measure. The relation between the two sets of components involve α and β. For example, if the matter is represented by some dynamical field ϕ, the quantities $T_{\perp\perp}$ and $T_\perp^a$ cannot be calculated from the initial data $(\phi,\dot\phi)$ without the use of α and β. Hence there is a certain inconsistency in the traditional formulation of the TSP in that it eliminates any appearance of the normal n in favor of ∂_t, lapse, and shift on the left side (the gravitational part), but not on the right side (matter part) of the constraints. This inconsistency was already felt by others (see, e.g., Sec. IV of Ref. 3), but no alternative formulation was hitherto attempted.

In this paper we consider a generalized thin-sandwich problem (GTSP) for the initial data of the full system of coupled gravitational and matter fields, which avoids the difficulty just mentioned. As matter we shall consider coupled systems of scalar and gauge fields, further specified below. By Φ we shall collectively denote all dynamical fields of the theory. We ask: Under what conditions do input data $\Phi := (\Phi, \dot{\Phi})$ uniquely specify a solution to the Einstein matter equations? In the same fashion as for (1.5), one obtains a now *generalized* thin-sandwich equation (GTSE) from which one tries to determine the "gauge parameters" X given the data Ψ. One nontrivial aspect of our generalization is due to the possible presence of gauge matter fields. In this case X comprises lapse, shift, and *additional* functions with values in the Lie algebra of the gauge group. Our main result will consist of an implicit function theorem for this extended set of variables, which for our GTSP is precisely analogous to the result proven in Ref. 5 for the traditional form of the thin-sandwich problem. But note that the two formulations differ even without gauge fields.

II. THE GENERALIZED FRAMEWORK

The dynamical fields we consider involve the gravitational field, a gauge field with compact gauge group **G** of dimension N and an M-component scalar field with values in an associated $\mathbf{R}^M$-vector-bundle. It couples to both previous fields in the standard minimal fashion. For simplicity we assume the **G**-principal bundle to be trivial. Since the frame bundle of any orientable three-manifold is always trivial, we may choose global trivializations of these bundles and represent fields by their globally defined component fields on space–time. For fixed time, a configuration of fields is given by the $6+3N+M$ component fields on Σ,

$$\Phi^A := (g_{ab}, A_a^\mu, \phi^\alpha), \tag{2.1}$$

where indices $\mu,\nu,...$ denote components in the Lie algebra **LG** and $\alpha,\beta,...$ denote components in $\mathbf{R}^M$. Hence we think of a field configuration as mapping

$$\Phi: \Sigma \to S_2^+(\mathbf{R}^3) \times \mathbf{R}^{3N} \times \mathbf{R}^M, \tag{2.2}$$

where $S_2^+(\mathbf{R}^3)$ denotes the space of symmetric, positive definite, bilinear forms on $\mathbf{R}^3$, in which g_{ab} is valued. Looking ahead, we remark that this space may be identified (via polar decomposition) with the homogeneous manifold GL(3,**R**)/O(3), a fact that is useful when studying the Lorentzian structure (2.11).[8] The total target space, whose dimension is $6+3N+M$, will be denoted by Θ, and the space of mappings $\Sigma \to \Theta$ (to be further specified) by $\mathcal{M}$.

Compactness of the gauge-group **G** implies the existence of **G**-invariant, symmetric, positive definite, bilinear forms $k_{\mu\nu}$ and $h_{\alpha\beta}$ on **LG** and $\mathbf{R}^M$, respectively. The class of models we shall consider here are characterized by the Lagrange four-form

$$\mathcal{L} = \tfrac{1}{2} * R^{(4)} - \tfrac{1}{4} k_{\mu\nu} \Omega^\mu \wedge * \Omega^\nu - \tfrac{1}{2} h^{\alpha\beta} \nabla \phi_\alpha \wedge * \nabla \phi_\beta - W, \tag{2.3}$$

where $*$ is the Hodge-duality map wrt $g^{(4)}$ and Ω is the curvature of A. For notational simplicity we shall denote all the covariant derivatives acting on sections in the various vector bundles by the same symbol ∇. In general, it will therefore involve the Christoffel symbols as well as the gauge connection A in the appropriate representation of **LG**. The potential W depends on the fields and their first spatial derivatives. Its precise form is not important, except that we need to explicitly exclude second (or higher) derivative couplings, in particular, the so-called conformal coupling of the scalar and gravitational field. The reason for this will be explained below. The Hamiltonian constraint for (2.3) has the general form

$$\mathcal{H} = \tfrac{1}{2} G_{AB} V^A V^B + U = 0, \tag{2.4}$$

where the potential is the following sum:

$$U = -R + \tfrac{1}{4} k_{\mu\nu} g^{ac} g^{bd} \Omega^{\mu}_{ab} \Omega^{\nu}_{cd} + h_{\alpha\beta} g^{ab} \nabla_a \phi^{\alpha} \nabla_b \phi^{\beta} + W, \tag{2.5}$$

whose terms represent the contributions of the gravitational, gauge, and scalar fields (two terms), respectively, to the potential energy. R denotes the Ricci scalar for g and Ω now denotes the curvature of the pulled-back **G**-bundle over Σ. The covariant derivative ∇_a refers from now on to the connection in the splicing of the frame bundle of Σ with the pulled-back **G**-bundle. The ''canonical velocities,'' V^A, can be written in terms of the ''coordinate velocities'' $\dot{\Phi}^A := \partial_t \Phi^A$ and the ''gauge-parameters'' α and $\xi := (\beta, \lambda)$. The general structure is [compare (1.2)]

$$V = \frac{1}{\alpha} \Gamma = \frac{1}{\alpha} (\dot{\Phi} + f_{\xi}), \tag{2.6}$$

where f_{ξ} represents the variations generated by the infinitesimal diffeomorphism and gauge-transformation with parameters β and λ, respectively. Resolved in terms of the individual fields (2.1), the components of Γ, and hence of f_{ξ}, read

$$\Gamma_{ab} = \dot{g}_{ab} - 2 \nabla_{(a} \beta_{b)}, \tag{2.7}$$

$$\Gamma^{\mu}_{a} = \dot{A}^{\mu}_{a} - \beta^{b} \Omega^{\mu}_{ba} - \nabla_a \lambda^{\mu}, \tag{2.8}$$

$$\Gamma^{\alpha} = \dot{\phi}^{\alpha} - \beta^{a} \nabla_a \phi^{\alpha} + \lambda^{\mu} \rho^{\alpha}_{\mu\beta} \phi^{\beta}, \tag{2.9}$$

where ρ denotes the representation of **LG** in $\mathrm{gl}(M, \mathbf{R})$.

Finally, following the decomposition of Θ as Cartesian product, the ''kinetic-energy-metric'' G_{AB} on Θ which appears in (2.4) has the following block structure:

$$G_{AB} = G^{ab\,cd} \oplus k_{\mu\nu} g^{ab} \oplus h_{\alpha\beta}, \tag{2.10}$$

where the first 6×6 block is given by the DeWitt metric

$$G^{ab\,cd} = \tfrac{1}{4} (g^{ac} g^{bd} + g^{ad} g^{bc} - 2 g^{ab} g^{cd}), \tag{2.11}$$

which is a Lorentz metric of signature (1.5). Hence G_{AB} itself is a Lorentz metric of signature $(1, 5 + 3N + M)$ on the manifold Θ, which is homeomorphic to $\mathbf{R}^{6+3N+M}$. We shall sometimes denote this metric simply by G and write $G(\cdot, \cdot)$ for the inner product. We will see that the Lorentzian signature of G is *the* important feature on which the proofs of our main results rest. This is also the reason why we had to exclude higher derivative (e.g., conformal) couplings of the scalar and gravitational fields, since they will, in general, destroy this signature structure.[9] On the other hand, our proofs will still apply to more complicated self-couplings of the scalar field. For example, nonlinear σ-models would be allowed, since here the target space metric of the scalar field, $h_{\alpha\beta}$, simply becomes ϕ^{α} dependent, which is unimportant to our proofs as long as it stays positive definite.

The (undensitized) momenta of the field Φ^A are just given by the covariant components—with respect to G—of the velocities: $P_A := G_{AB} V^B$. For the individual fields we write $P_A = (\pi^{ab}, \pi^{a}_{\mu}, \pi_{\alpha})$. In the canonical theory, the phase space function that generates infinitesimal diffeomorphisms and gauge transformations with parameter ξ' is given by

$$\mathcal{P}_{\xi'} := \int_{\Sigma} d\mu \, P_A f^{A}_{\xi'}, \tag{2.12}$$

where here and below we set $d\mu := \sqrt{\det\{g_{ab}\}}d^3x$ for the measure on Σ. For completeness we remark that, using (2.7)–(2.9), a straightforward calculation yields the following familiar Poisson-bracket relation, which involves the curvature tensor Ω of the gauge field on Σ:

$$\{\mathcal{P}_{\xi'}, \mathcal{P}_{\xi''}\} = P_{\xi'''}, \tag{2.13}$$

where $\xi''' = (\beta''', \lambda''')$ reads

$$\beta''' = -[\beta', \beta''], \tag{2.14}$$

$$\lambda''' = [\lambda', \lambda''] - \Omega(\beta', \beta''). \tag{2.15}$$

The diffeomorphism- and Gauss-constraints are just given by $\mathcal{P}_{\xi'} = 0\ \forall \xi'$, which may be expressed by saying that the velocity field V is L^2G-orthogonal to all ''vertical'' vector fields $f_{\xi'}$. Hence we must have

$$0 = \int_\Sigma P_A f^A_{\xi'}\, d\mu = \int_\Sigma \frac{1}{\alpha} G(\Gamma, f_{\xi'}) d\mu =: \int_\Sigma (g_{ab}\beta'^a \mathcal{D}^b + k_{\mu\nu}\lambda'^\mu \mathcal{G}^\nu) d\mu, \tag{2.16}$$

for all, say C^∞, vector fields β' and **LG**-valued functions λ'. This is equivalent to $\mathcal{D}^a = 0$ (diffeomorphism constraint) and $\mathcal{G}^\mu = 0$ (Gauss constraint). Explicitly we get

$$\mathcal{D}^a = 2G^{abcd}\nabla_b\left(\frac{1}{\alpha}\Gamma_{cd}\right) - \frac{1}{\alpha} g^{ab} k_{\mu\nu} \Omega^\mu_{bc} g^{cd} \Gamma^\nu_d - \frac{1}{\alpha} g^{ab} (\nabla_b \phi^\alpha) h_{\alpha\beta} \Gamma^\beta, \tag{2.17}$$

$$\mathcal{G}^\mu = \nabla_a\left(\frac{1}{\alpha} g^{ab} \Gamma^\mu_b\right) + \frac{1}{\alpha} k^{\mu\nu} \rho^\alpha_{\nu\beta} \phi^\beta h_{\alpha\gamma} \Gamma^\gamma. \tag{2.18}$$

Given Φ and $\dot{\Phi}$ we now have the $4+N$ equations $0 = \mathcal{H} = \mathcal{D}^a = \mathcal{G}^\mu$ for the $4+N$ unknowns $\alpha, \beta^a, \lambda^\mu$. The first step consists of inserting (2.6) into (2.4) and solving for α^2:

$$\alpha^2 = -\frac{G(\Gamma, \Gamma)}{2U}. \tag{2.19}$$

For this to make sense the right-hand side must be positive. But for the following analysis it turns out that we need to put the following stronger condition.

Condition 1 (a priori):

$$U > 0, \tag{2.20}$$

$$G(\Gamma, \Gamma) < 0. \tag{2.21}$$

Note that (2.20) just involves the initial data, whereas (2.21) contains as well the unknowns β and λ (hence ''*a priori*''). Note that (2.21) says that the system must, at each point of Σ, move in a ''timelike'' direction with respect to the Lorentz metric G. The need for such an *a priori* bound implies that our results will only be perturbative. Given these bounds, we set α equal to the positive square root of the right-hand side of (2.19). We can then eliminate α from (2.17) and (2.18) and obtain a set of $3+N$ equations for the $3+N$ unknowns $\xi = (\beta, \lambda)$, which we call the generalized *reduced* thin-sandwich equation (GRTSE):

$$F[\Psi, \xi] = 0. \tag{2.22}$$

Now consider the following functional over configurations satisfying Condition 1:

$$S[\Psi,\,\xi]=\int_\Sigma \sqrt{-2UG(\Gamma,\Gamma)}\,d\mu. \tag{2.23}$$

Then solutions to the GRTSE are stationary points with respect to variations in ξ. More precisely, let D_2 denote the partial (functional) derivative in the second argument and denote the L^2 inner product of ξ_1 and ξ_2 by $\langle\xi_1|\xi_2\rangle:=\int_\Sigma d\mu(k_{\mu\nu}\lambda_1^\mu\lambda_2^\nu+g_{ab}\beta_1^a\beta_2^b)$. We then have the following.

Lemma 1:

$$D_2S[\Psi,\,\xi](\xi')=-\langle\xi'|F[\Psi,\,\xi]\rangle. \tag{2.24}$$

Proof: For $s\in(-\epsilon,\epsilon)$ set $\eta(s)=\xi+s\xi'$ and $\Gamma(s)=\dot{\Phi}+f_{\eta(s)}$. Hence $(d/ds)|_{s=0}\Gamma(s)=f_{\xi'}$. Then, recalling (2.16),

$$\frac{d}{ds}\bigg|_{s=0}S[\Psi,\,\eta(s)]=-\int_\Sigma\sqrt{\frac{-2U}{G(\Gamma,\Gamma)}}\,G(\Gamma,f_{\xi'})d\mu=-\langle\xi'|F[\Psi,\,\xi]\rangle. \tag{2.25}$$

III. MAIN RESULTS

In this section we are mainly concerned with the linearization of the GRTSE, except at the end where we will discuss global uniqueness. The corresponding linear operator will be called L, without explicit indication that it depends on Ψ and ξ. It is defined by

$$\langle\xi'|L\xi''\rangle:=\frac{d}{ds}\bigg|_{s=0}\langle\xi'|F[\Psi,\,\xi+s\xi'']\rangle=\int_\Sigma d\mu\,\frac{d}{ds}\bigg|_{s=0}\frac{G(f_{\xi'},\Gamma(s))}{\alpha(s)}, \tag{3.1}$$

where $\Gamma(s)=\dot{\Phi}+f_{\xi+s\xi''}$ and $\alpha(s)=[-G(\Gamma(s),\Gamma(s))/2U]^{1/2}$. Setting $\Gamma(s=0)=\Gamma$, $\alpha(s=0)=\alpha$, and noting that $(d/dt)|_{s=0}\Gamma(s)=f_{\xi''}$, we get

$$\frac{d}{ds}\bigg|_{s=0}\alpha(s)=-\frac{G(f_{\xi''},\Gamma)}{2\alpha U}=\alpha\,\frac{G(f_{\xi''},\Gamma)}{G(\Gamma,\Gamma)}, \tag{3.2}$$

where we used (2.4) to eliminate U in the last step. Hence

$$\frac{d}{ds}\bigg|_{s=0}\frac{G(f_{\xi'},\Gamma(s))}{\alpha(s)}=\frac{1}{\alpha}\left[G(f_{\xi'},f_{\xi''})-\frac{G(f_{\xi'},\Gamma)G(f_{\xi''},\Gamma)}{G(\Gamma,\Gamma)}\right]=\frac{G(f_{\xi'}^\perp,f_{\xi''}^\perp)}{\alpha}, \tag{3.3}$$

where

$$f_{\xi'}^\perp:=f_{\xi'}-\Gamma\,\frac{G(f_{\xi'},\Gamma)}{G(\Gamma,\Gamma)} \tag{3.4}$$

is the G-orthogonal projection of $f_{\xi'}$ pointwise perpendicular to Γ. This leads to the following expression for L's matrix elements:

$$\langle\xi'|L\xi''\rangle=\int_\Sigma d\mu\,\frac{1}{\alpha}\,G(f_{\xi'}^\perp,f_{\xi''}^\perp), \tag{3.5}$$

where α is the square-root of the rhs of (2.19). If (2.21) holds, Γ is ''timelike'' (pointwise on Σ) and hence the $f^\perp$'s are ''spacelike'' or zero. Since the metric G is Lorentzian, it is positive definite on ''spacelike'' vectors. Hence we have shown the following.

Lemma 2: Suppose Condition 1 holds. Then L is self-adjoint and non-negative. Furthermore, $\xi'\in kernel(L)\Leftrightarrow\exists\,\kappa:\Sigma\to\mathbf{R}$ *such that*

$$f_{\xi'} = \kappa \Gamma. \tag{3.6}$$

Remark: Symmetry is expected, for consider the rhs of (2.23) as functional of Ψ and Γ, denoting it by $S[\Psi,\Gamma]$. Then the calculation of the rhs of (3.5) was just that of $-D_2^2 S[\Psi,\Gamma] \times(\Gamma',\Gamma'')$ for $\Gamma'=f_{\xi'}$ and $\Gamma''=f_{\xi''}$.

We will next show that under the same hypotheses L is, in fact, elliptic. For this we will need the following.

Lemma 3: If Condition 1 holds, π^{ab} is a positive or negative definite matrix.

Proof: Condition 1 implies $G^{AB}P_A P_B<0 \Rightarrow G_{ab\,cd}\pi^{ab}\pi^{cd}=2(\pi_{ab}\pi^{ab}-\frac{1}{2}(\pi^a_a)^2)<0$. Choosing a frame where $g_{ab}=\delta_{ab}$ and $\pi^{ab}=\mathrm{diag}\,(p_1,p_2,p_3)$, this is equivalent to the following condition on the eigenvalue vector $\vec{p}:=(p_1,p_2,p_3)$:

$$\left|\frac{\vec{n}\cdot\vec{p}}{\|\vec{p}\|}\right| > \sqrt{\frac{2}{3}}, \tag{3.7}$$

where $\vec{n}=(1,1,1)/\sqrt{3}$, which means that $\vec{p}$ lies in the interior of the double cone with axis along $\vec{n}$ and opening angle $\theta<\cos^{-1}(2/3)^{1/2}$ about its axis. This cone just touches the walls of the positive and negative octants along the bisecting lines, and since $\vec{p}$ must be in its interior, all eigenvalues are either strictly positive or strictly negative. □

Proposition 1: Suppose Condition 1 holds. Then the second-order differential operator L is elliptic.

Proof: For later purposes we shall prove slightly more, namely that the definiteness of π^{ab} is equivalent to the ellipticity of L. This we do by a direct calculation of L's principal symbol. For this we to go back to the explicit formulas (2.17) and (2.18) for the full nonlinear problem and explicitly linearize them, but keeping track only of the highest (second) derivatives. By $\overset{2}{=}$ we shall denote equality in the second derivative terms. We set again $\Gamma(s)=\dot{\Phi}+f_{\xi+s\xi'}$, etc., with $\xi'=(\beta',\lambda')$. It will be convenient to express things in terms of the momenta, using $\alpha P_A=G_{AB}\Gamma^B$, and accordingly write (3.2) in the form

$$\left.\frac{d}{ds}\right|_{s=0}\alpha(s)=-\frac{P_A f^A_{\xi'}}{2U}. \tag{3.8}$$

Then

$$\left.\frac{d}{ds}\right|_{s=0}\mathcal{D}^a \overset{2}{=} \partial_b\left[\frac{P_A f^A_{\xi'}}{\alpha U}\pi^{ab}-\frac{4}{\alpha}G^{ab\,cd}\partial_c\beta'_d\right]$$

$$\overset{2}{=}-\frac{1}{\alpha}\left[\frac{2\pi^{ac}\pi^{de}g_{db}\partial_c\partial_e\beta'^b+\pi^{ab}\pi^c_\nu\partial_b\partial_c\lambda'^\nu}{U}+g^{bc}\partial_b\partial_c\beta'^a-g^{ac}\partial_c\partial_b\beta'^b\right], \tag{3.9}$$

$$\left.\frac{d}{ds}\right|_{s=0}\mathcal{G}^\mu \overset{2}{=} \partial_a\left[\frac{P_A f^A_{\xi'}}{2\alpha U}k^{\mu\nu}\pi^a_\nu-\frac{1}{\alpha}g^{ab}\partial_b\lambda'^\mu\right]$$

$$\overset{2}{=}-\frac{1}{\alpha}\left[\frac{2k^{\mu\nu}\pi^a_\nu g_{bc}\pi^{cd}\partial_a\partial_d\beta'^b+k^{\mu\sigma}\pi^a_\sigma\pi^b_\nu\partial_a\partial_b\lambda'^\nu}{2U}+g^{ab}\partial_a\partial_b\lambda'^\mu\right]. \tag{3.10}$$

Replacing $\partial_a\to k_a$ we can just read off the matrix of the principal symbol $\sigma(k)$ in the general form

$$\begin{bmatrix} \sigma^\mu_\nu & \sigma^\mu_b \\ \sigma^a_\nu & \sigma^a_b \end{bmatrix}, \tag{3.11}$$

where we have chosen to order the $N+3$ rows and columns so that we first count the N components of λ' and then the three components of β'. In order to calculate the determinant we make the following simplifications: We call $\pi^{ab}k_b =: p^a$, $\pi^a_\mu k_a =: \pi_\mu$, $\pi^\mu := k^{\mu\nu}\pi_\nu$, and choose a spatial frame where $g_{ab} = \delta_{ab}$, $k_a = (\|k\|,0,0)$ and $p^a = (p_1, p_2, 0)$. Then (3.11) reads explicitly

$$\sigma(k) = -\frac{1}{\alpha}\begin{bmatrix} \frac{\pi^\mu\pi_\nu}{2U} + \|k\|^2\delta^\mu_\nu & \frac{\pi^\mu p_1}{U} & \frac{\pi^\mu p_2}{U} & 0 \\ \frac{p_1\pi_\nu}{U} & \frac{2p_1^2}{U} & \frac{2p_1p_2}{U} & 0 \\ \frac{p_2\pi_\nu}{U} & \frac{2p_1p_2}{U} & \frac{2p_2^2}{U} + \|k\|^2 & 0 \\ 0 & 0 & 0 & \|k\|^2 \end{bmatrix}. \tag{3.12}$$

Now, for $k \neq 0$, $p_1 = \|k\|\pi(\hat{k},\hat{k})$, where $\hat{k} := k/\|k\|$. Lemma 3 then ensures that $p_1 \neq 0$. In order to calculate det $\{\sigma(k)\}$, we simplify this matrix as follows: We subtract $\pi^\mu/2p_1$ times the $(N+1)$st row from the μth row, for each $1 \leq \mu \leq N$, and also subtract p_2/p_1 times the $(N+1)$st row from the $(N+2)$nd row. The resulting matrix reads

$$-\frac{1}{\alpha}\begin{bmatrix} \|k\|^2\delta^\mu_\nu & 0 & 0 & 0 \\ \frac{p_1\pi_\nu}{U} & \frac{2p_1^2}{U} & \frac{2p_1p_2}{U} & 0 \\ 0 & 0 & \|k\|^2 & 0 \\ 0 & 0 & 0 & \|k\|^2 \end{bmatrix}. \tag{3.13}$$

Its determinant, which equals that of $\sigma(k)$, is now easily calculated:

$$\det\{\sigma(k)\} = \frac{2p_1^2}{U}\|k\|^{2(N+2)}\left[-\frac{1}{\alpha}\right]^{N+3} = 2\left[-\frac{\|k\|^2}{\alpha}\right]^{N+3}\frac{[\pi(\hat{k},\hat{k})]^2}{U}. \tag{3.14}$$

Lemma 3 implies that this is zero $\Leftrightarrow k=0$, which finally proves ellipticity of L. □

Remark: It is interesting to see that all momenta of matter fields in (3.12) drop out when taking the determinant, and that the result is $\propto [\pi(\hat{k},\hat{k})]^2$, just as in the vacuum case.[5] The underlying reason is the fact that $(f_\xi)_{\text{matter}}$ depends ultralocally (i.e., without derivatives) on β [see (2.8) and (2.9)], which implies that the lower-right 3×3-matrix in (3.12) is independent of the momenta for matter fields and hence identical to the principal symbol in the vacuum case. We also note that there is a simple argument that ellipticity of L is implied by the definiteness of π^{ab}. In brief, the argument is simply this: (Here I follow a suggestion made by the referee.) Replacing $\partial_a \to k_a$, any kernel element of the quadratic form (3.5) must, according to (3.6), satisfy $-2k_{(a}\beta_{b)} = \kappa\Gamma_{ab}$, where $\kappa \not\equiv 0$, which immediately implies $\pi(k,k)=0$. Hence a definite π^{ab} implies a trivial kernel. (3.14) shows that this condition is also necessary.

The results obtained so far suffice to deduce an implicit function theorem. To state it precisely, we need to choose appropriate function spaces. It is natural to choose Sobolev spaces which are also used in showing existence of the time evolution.[2] To begin with, it is convenient to summarize the order of spatial differentiation by which the various fields enter the quantities Γ, U, α, and hence the GRTSE, by the following matrix:

	Γ	U	α	GRTSE
g_{ab}	1	2	2	3
A_a^μ	1	1	1	2
ϕ^α	1	1	1	2
$\dot{g}_{ab}$	0	—	0	1
$\dot{A}_a^\mu$	0	—	0	1
$\dot{\phi}^\alpha$	0	—	0	1
β^a	1	—	1	2
λ^μ	1	—	1	2

(3.15)

Note that we assumed that U only contained first derivatives of the matter fields, whereas it contains second derivatives of the gravitational field through the Ricci scalar. Hence g_{ab} enters the GRTSE thrice differentiated. (This seems to have been overlooked in Ref. 5.)

By $H^n(V)$ we shall denote the Sobolev space of V-valued functions on Σ with L^2-norm in the first n derivatives (i.e., generalizing $H^n = W^{n,2}$ using an inner product in V). We shall have $V = T_2^0$ for g_{ab} and $\dot{g}_{ab}$, $V = T_1^0 \otimes \mathbf{LG}$ for A_a^μ and $\dot{A}_a^\mu$, $V = \mathbf{R}^M$ for ϕ^α and $\dot{\phi}^\alpha$, $V = T_0^1$ for β^a, and $V = \mathbf{LG}$ for λ^μ. The inner products for the various V's are just as in the metric G, except for the gravitational field where instead of $G^{ab\,cd}$, which is not positive definite, we choose the positive definite form $g^{ac}g^{bd}$ [compare (2.11)]. Now we define the Sobolev spaces

$$H_\Phi^n := H^{n+3}(T_2^0) \times H^{n+2}(T_1^0 \otimes \mathbf{LG}) \times H^{n+2}(\mathbf{R}^M), \tag{3.16}$$

$$H_{\dot{\Phi}}^n := H^{n+1}(T_2^0) \times H^{n+1}(T_1^0 \otimes \mathbf{LG}) \times H^{n+1}(\mathbf{R}^M), \tag{3.17}$$

$$H_\Psi^n := H_\Phi^n \times H_{\dot{\Phi}}^n, \tag{3.18}$$

$$H_\xi^n := H^n(T_0^1) \times H^n(\mathbf{LG}). \tag{3.19}$$

One may now show that the operator F in the GRTSE, $F[\Psi,\xi]=0$, defines a C^1-map

$$F: H_\Psi^n \times H_\xi^{n+2} \to H_\xi^n \quad \text{for } n \geqslant 2 \tag{3.20}$$

on the domain of fields (Ψ, ξ) which satisfy Condition 1. [The Sobolev embedding theorem for three-dimensional domains and L^2-norms implies a continuous embedding $H^n(V) \hookrightarrow C^k(V)$ for $k < n - 3/2$. $n \geqslant 2$ is needed to guarantee continuity of the functions and gain pointwise control, which is needed in the proof for F being C^1.] For this we need to impose suitable but very mild regularity conditions on the unspecified function W in (2.5). The linear map L is the first derivative of F wrt the second argument: $D_2F[\Psi,\xi]$. Ellipticity implies that Image $(L) := L(H_\xi^{n+2}) \subseteq H_\xi^n$ is closed and hence H_ξ^n splits as orthogonal sum of closed subspaces, given by L's image and the kernel of L's adjoint. Hence, since L is self-adjoint, $H_\xi^n = \text{image}\,(L) \oplus \text{kernel}\,(L)$. We now get an implicit function theorem for the map F if $D_2F[\Psi,\xi]$ is a linear isomorphism, i.e., if kernel $(L) = \{0\}$. But since L is elliptic, any nontrivial element in the kernel may be represented by a C^∞ function ξ' which must then satisfy (3.6). Hence a trivial kernel is equivalent to the following condition for smooth functions:

Condition 2:

$$f_{\xi'} = \kappa\Gamma \quad \text{implies} \quad \xi' = 0, \quad \kappa = 0. \tag{3.21}$$

Hence we arrive at the following formulation of an implicit function theorem for the generalized thin-sandwich problem. It may be seen as generalization of Theorem 2 (or 3) in Ref. 5.

Theorem: *Let $n \geqslant 2$ and $(\Psi,\xi) \in H^n_\Psi \times H^{n+2}_\xi$ be a solution to the GRTSE, $F[\Psi,\xi]=0$, which satisfies Condition 1 [i.e., (2.20 and 2.21)] and Condition 2 [i.e., (3.21)]. Then there exist open neighborhoods $V \subset H^n_\Psi$ of Ψ and $W \subset H^n_\Psi \times H^{n+2}_\xi$ of (Ψ, ξ) and a C^1 map $\sigma: V \to H^{n+2}_\xi$ such that $F[\Psi',\xi']=0$ for $(\Psi',\xi') \in W \Leftrightarrow \Psi' \in V$ and $\xi' = \sigma(\Psi')$.*

Consider the action T of $\dot{\mathbf{R}} := \mathbf{R} - \{0\}$ on $H^n_\Psi \times H^{n+2}_\xi$, given by $T_\delta(\Phi,\dot\Phi,\xi) := (\Phi,\delta\dot\Phi,\delta\xi)$ for $\delta \in \dot{\mathbf{R}}$. It leaves individually invariant the three subsets of points (Ψ, ξ) which (1) obey Condition 1, (2) obey Condition 2, and (3) solve the GRTSE. To see this, recall that f_ξ is linear in ξ, hence $T_\delta \Gamma = \delta\Gamma$. Invariance of the first set is now obvious. Further, if $f_{\xi'} = \kappa\Gamma$ has only the trivial solution, then so does $f_{\xi'} = \kappa T_\delta \Gamma$, since otherwise $(\delta^{-1}\xi',\kappa)$ would be a nontrivial solution to the first equation. Hence the second set is invariant. Finally, since Γ scales with δ and the square root of expression (2.19) for α with $|\delta|$, the GRTSE (2.16) changes at most by an overall sign, which proves invariance of the third set.

We can now repeat the Theorem for each point $T_\delta(\Psi,\xi)$ on the $\dot{\mathbf{R}}$ orbit of (Ψ, ξ) with open sets V_δ, W_δ and solution maps σ_δ. In this way the solution map σ extends to a solution map σ^*: $V^* \to H^{n+2}_\xi$, where $V^* := \cup_{\delta\in\dot{\mathbf{R}}} V_\delta$, which uniquely represents all solutions in $W^* := \cup_{\delta\in\dot{\mathbf{R}}} W_\delta$. By construction it satisfies $\sigma^*(\Phi,\delta\dot\Phi) = \delta\sigma(\Phi,\dot\Phi) \forall \delta \in \dot{\mathbf{R}}$. Dropping the superscript $*$, we formulate this as follows.

Corollary 1: Let (Ψ, ξ) be as in the Theorem. Then there exist open neighborhoods $V \subset H^n_\Psi$ of $\cup_{\delta\in\dot{\mathbf{R}}} T_\delta \Psi$ and $W \subset H^n_\Psi \times H^{n+2}_\xi$ of $\cup_{\delta\in\dot{\mathbf{R}}} T_\delta(\Psi,\xi)$ and a C^1 map $\sigma: V \to H^{n+2}_\xi$ such that $F[\Psi',\xi']=0$ for $(\Psi',\xi') \in W \Leftrightarrow \Psi' \in V$ and $\xi' = \sigma(\Psi')$. Moreover,

$$\sigma(\Phi,\delta\dot\Phi) = \delta\sigma(\Phi,\dot\Phi), \quad \forall \delta \in \dot{\mathbf{R}} \tag{3.22}$$

Finally we prove that Condition 2 not only ensures local but also global uniqueness. It generalizes the analogous result for the traditional RTSE, proven in Ref. 3. [The full statement and proof given in Ref. 3 contains an additional part which is erroneous, as was first pointed out in Ref. 5. If transcribed to our setting, the incorrect part would amount to the claim that (3.23) implied $r \equiv 1$ without using Condition 2.]

Proposition 2: Let (Ψ, ξ) and $(\Psi, \tilde\xi)$ satisfy Condition 1 and (Ψ, ξ) the GRTSE. Then $(\Psi,\tilde\xi)$ satisfies the GRTSE $\Leftrightarrow$ there exists a positive function r: $\Sigma \to \mathbf{R}_+$ such that $\Gamma = \dot\Phi + f_\xi$ and $\tilde\Gamma = \dot\Phi + f_{\tilde\xi}$ are related by

$$\tilde\Gamma = r\Gamma. \tag{3.23}$$

Proof: $\Leftarrow$: This follows trivially from the fact that the GRTSE, i.e., Eqs. (2.17) and (2.18), contain ξ only through the combination $(1/\alpha)\Gamma$.

$\Rightarrow$: For $s \in [0,1]$, consider the convex combinations $\xi(s) := s\xi + (1-s)\tilde\xi$ and $\Gamma_s := \dot\Phi + f_{\xi(s)} = s\Gamma + (1-s)\tilde\Gamma$. In the following it is useful to think of each Γ_s as section in the pulled-back bundle $\Phi^* T(\Theta)$ whose fiber at $p \in \Sigma$ is a Minkowski space $\mathbf{R}^{1,5+3N+M}$ with metric $G_{\Phi(p)}$. Condition 1 requires $\Gamma(p)$ and $\tilde\Gamma(p)$ to be "timelike," so that $s \mapsto \Gamma_s(p)$ is the straight path connecting these two "timelike" vectors. First we show that $\Gamma(p)$ and $\tilde\Gamma(p)$ lie in the interior of the same "light-cone" for some, and hence all, $p \in \Sigma$. To see this, we consider, for each p, the inner product $G(\Gamma - \tilde\Gamma, \mathcal{V})$ with the timelike vector $\mathcal{V} := g_{ab}\partial/\partial g_{ab}$ of constant length-squared $G(\mathcal{V},\mathcal{V}) = -3$. Now,

$$\int_\Sigma G(\Gamma - \tilde\Gamma, \mathcal{V})\, d\mu = 2\int_\Sigma \nabla_a(\beta^a - \tilde\beta^a)\, d\mu = 0, \tag{3.24}$$

so that, because Σ is connected, there exists a point $p\in\Sigma$ where $G(\Gamma-\tilde{\Gamma},\mathcal{V})(p)=0$. Hence $\Gamma(p)$ and $\tilde{\Gamma}(p)$ point in the same half of the ''light-cone,'' and so does $\Gamma_s(p)$, since the interior of the half ''light-cone'' is a convex set. By continuity this must then be true at each point $p\in\Sigma$ so that $G(\Gamma_s,\Gamma_s)$ is a negative-valued function on Σ for each s.

Next we consider the function

$$I(s):=S[\Psi,\,\xi(s)]=\int_\Sigma \sqrt{-2UG(\Gamma_s,\Gamma_s)}\,d\mu. \tag{3.25}$$

We have $I'(0)=0=I'(1)$, where $'=d/ds$, since ξ and $\tilde{\xi}$ solve the GRTSE. Furthermore, a straightforward calculation yields

$$I''(s)=\int_\Sigma \frac{[2U]^{1/2}}{[-G(\Gamma_s,\Gamma_s)]^{3/2}}\,G(\Gamma,\Gamma)G(\tilde{\Gamma}_\perp,\tilde{\Gamma}_\perp)\,d\mu\leqslant 0, \tag{3.26}$$

with

$$\tilde{\Gamma}_\perp:=\tilde{\Gamma}-\Gamma\,\frac{\Gamma(\Gamma,\tilde{\Gamma})}{G(\Gamma,\Gamma)}. \tag{3.27}$$

The inequality in (3.26) results from Γ being ''timelike'' and $\tilde{\Gamma}_\perp$ being ''spacelike'' or zero. But $I'(0)=I'(1)=0$ and $I''(s)\leqslant 0$ imply $I''\equiv 0$. On the other hand, equality in (3.26) can only be achieved for $\tilde{\Gamma}_\perp=0$ which is equivalent to (3.23), where r must be positive valued since Γ and $\tilde{\Gamma}$ point in the same half of the ''light cone.'' □

Now (3.23) implies (3.6) with $\xi'=\tilde{\xi}-\xi$ and $\kappa=r-1$, so that Condition 2 will enforce $r=1$ and $\xi=\tilde{\xi}$. Hence we have the following corollary.

Corollary 2: If $(\Psi,\,\xi)$ *and* $(\Psi,\tilde{\xi})$ *satisfy the GRTSE and Conditions 1 and 2, then* $\xi=\tilde{\xi}$.

IV. DISCUSSION

It is obvious that the strategy of the (generalized) thin-sandwich approach cannot work for all data. Obvious bad data are those for which the Hamiltonian constraint cannot be solved for a nowhere vanishing α. For example, consider fields Φ such that $U>0$ and velocities $\dot{\Phi}$ whose gravitational part is pure gauge: $\dot{g}_{ab}=2\nabla_{(a}\xi'_{b)}$. The Hamiltonian constraint implies $(\nabla_a\eta^a)^2\geqslant U\alpha^2$, where $\eta=\xi'-\xi$, showing that α must vanish somewhere since $\int_\Sigma d\mu\,\nabla_a\eta^a=0$ and Σ is connected. To avoid such situations, Condition 1 or its reversed version, $U<0$ and $G(\Gamma,\Gamma)>0$, may be imposed. However, if the second condition is chosen, formula (3.14) together with the proof of Lemma 3 show that L manifestly fails to be elliptic, thus leaving only Condition 1.

The technical Condition 2 has an interpretation in terms of the ''canonical data'' (Φ,V), where $\alpha V=\dot{\Phi}+f_\xi$ and where we assume $(\Phi,\dot{\Phi},\xi)$ to satisfy Condition 1 in order to have $\alpha\neq 0$. Namely, if $f_{\xi'}=\kappa\Gamma$ for some nonzero $(\dot{\xi}',\kappa)$, then $f_{\xi'}=\alpha\kappa V$ says that the same canonical data admit a representation in terms of the new lapse function $\alpha_{\text{new}}=\kappa\alpha$ (now possibly with zeros), gauge functions $\xi_{\text{new}}=\xi'$, and coordinate-velocities $\dot{\Phi}_{\text{new}}=0$. Conversely, if an α_{new} exists such that $\alpha_{\text{new}}V=f_{\xi_{\text{new}}}$, then $f_{\xi'}=\kappa\Gamma$ with $\kappa=\alpha_{\text{new}}/\alpha$ and $\xi'=\xi_{\text{new}}$. Hence Condition 2 precisely excludes the existence of other representations of the same canonical data with vanishing coordinate velocities $\dot{\Phi}$. We note that Condition 2 may itself be implied by simple geometric conditions on Φ. One such set of conditions is provided by the following.

Proposition 3: Condition 2 is implied by Condition 1 and the following conditions on Φ*:*

$$(i)\quad \mathrm{Ric}<0\quad (\mathrm{Ric}=\textit{Ricci tensor of } g), \tag{4.1}$$

$$(ii) \quad \nabla_a \lambda^\mu = 0 \quad \text{and} \quad \lambda^\mu \rho^\alpha_{\mu\beta} \phi^\beta = 0 \quad \textit{imply} \quad \lambda^\mu = 0. \tag{4.2}$$

Proof: Let $f_{\xi'} = \kappa\Gamma$. Then $G(\Gamma,\Gamma) < 0 \Rightarrow G(f_{\xi'}, f_{\xi'}) \leq 0 \Rightarrow$

$$\int_\Sigma d\mu [G^{ab\,cd} 4\nabla_{(a}\beta'_{b)} \nabla_{(c}\beta'_{d)}] = \int_\Sigma d\mu 2[\nabla_{[a}\beta'_{b]}\nabla^{[a}\beta'^{b]} - R_{ab}\beta'^a\beta'^b] \leq 0 \tag{4.3}$$

$\Rightarrow \beta' = 0$. Since G is positive definite on $f_{\xi'}$'s for which $\beta' = 0$, $G(f_{\xi'}, f_{\xi'}) \leq 0$ implies $f_{\xi'} = 0$ which for $\beta' = 0$ is equivalent to the first two equations in (4.2). □

We recall that metrics with Ric<0 exist on any three-manifold Σ (Ref. 10) (e.g., quite in contrast to Ric>0, which is well known to imply a finite fundamental group). Equation (4.2) should be read as a mild genericity condition for the matter fields. For example, if we have a single U(1) gauge field and a charged scalar field [here represented by a real doublet (ϕ^1, ϕ^2)], then condition (4.2) is satisfied if the scalar field is not identically zero.

Finally we comment on the functional $S[\Psi, \xi]$ defined in (2.23). Given that Condition 1 is satisfied, solutions to the GRTSE are stationary points with respect to variations in ξ (Lemma 1). Lemma 2 asserts that these must be minima which are stable if Condition 2 is satisfied. We now assume Conditions 1 and 2 to hold and evaluate $S[\Psi, \xi]$ at a solution $\xi = \sigma(\Psi)$. We get a C^1-function $S_*: H^n_\Psi \to \mathbf{R}_+$,

$$S_*[\Phi, \dot\Phi] = \int_\Sigma d\mu \sqrt{-2UG(\dot\Phi + f_{\sigma(\Phi,\dot\Phi)}, \quad \dot\Phi + f_{\sigma(\Phi,\dot\Phi)})}, \tag{4.4}$$

which satisfies

$$S_*[\Phi, \delta\dot\Phi] = |\delta| S_*[\Phi, \dot\Phi] \tag{4.5}$$

for all $\delta \in \dot{\mathbf{R}}$. Standard consequences are

$$D_2 S_*[\Phi, \delta\dot\Phi](\dot\Phi) = \text{sign}\,(\delta) S_*[\Phi, \dot\Phi], \tag{4.6}$$

$$D_2^2 S_*[\Phi, \delta\dot\Phi](\dot\Phi, \dot\Phi) = 0, \tag{4.7}$$

where in (4.7) we assumed C^2 smoothness of σ. It is tempting to try and regard S_* as a kind of metric on at least an open subset of the tangent bundle of the space of fields. For pure gravity it generalizes a previously considered expression which is valid only for constant lapse function[8] and also gives rigorous meaning to a formal definition of a distance function given in Ref. 4. However, presently it is unclear to us whether S_* indeed defines an interesting geometric structure. {One may wonder whether it defined a Finsler metric. For this one would have to show that the bilinear form $D_2^2 S_*^2[\Phi, \dot\Phi]$ is (weakly) nondegenerate. But this is not even the case in finite dimensions for functions of the form (4.4) (i.e., sum of square roots rather than square root of sum). Take, e.g., the function $f(y_1, \ldots, y_n) = \Sigma_i \sqrt{y_i^2}$. Then $\partial_i \partial_j f^2 = 2\partial_i f \partial_j f$, which is obviously just of rank one.}

ACKNOWLEDGMENTS

I thank Julian Barbour, Robert Bartnik, Gyula Fodor, and Karel Kuchař for suggestions and clarifying remarks. During the course of this work the author was supported by NSF Grant No. PHY95-14240 at Pennsylvania State University, a visitor's grant from the Albert Einstein Institute at Potsdam, and a grant from the Swiss National Science Foundation.

[1] J. Isenberg, ''Steering the universe,'' Found. Phys. **16**, 651–665 (1986).
[2] Y. Choquet-Bruhat and J. W. York, ''The Cauchy Problem,'' in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 1.
[3] E. P. Belasco and H. C. Ohanian, ''Initial conditions in general relativity: Lapse and shift formulation,'' J. Math. Phys. **10**, 1503–1507 (1969).

[4] D. Christodoulou and M. Francaviglia, ''The geometry of the thin-sandwich problem,'' in *Isolated Gravitating Systems in General Relativity*, edited by J. Ehlers (North-Holland, New York, 1979).
[5] R. Bartnik and G. Fodor, ''On the restricted validity of the thin sandwich conjecture,'' Phys. Rev. D **48**, 3596–3599 (1993).
[6] J. A. Wheeler, ''Geometrodynamics and the issue of the final state,'' in *Relativity, Groups and Topology*, edited by B. DeWitt and C. DeWitt (Gordon and Breach, New York, 1964).
[7] R. F. Baierlein, D. H. Sharp, and J. A. Wheeler, ''Three-dimensional geometry as carrier of information about time,'' Phys. Rev. **126**, 1864–1865 (1962).
[8] D. Giulini, ''What is the geometry of superspace?'' Phys. Rev. D **51**, 5630–5635 (1995).
[9] C. Kiefer, ''Non-minimally coupled scalar fields and the initial value problem in quantum gravity,'' Phys. Lett. B **225**, 227–232 (1989).
[10] Z. L. Gao and S.-T. Yau, ''The existence of Ricci negatively curved metrics on three manifolds,'' Invent. Math. **85**, 637–652 (1986).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Null cones from $\mathcal{I}^+$ and Legendre submanifolds

Mirta Iriondo, Carlos N. Kozameh, and Alejandra T. Rojas
FaMAF, Universidad Nacional de Córdoba, 5000 Córdoba, Argentina

(Received 29 May 1998; accepted for publication 27 January 1999)

It is shown that the main variable Z of the null surface formulation of GR is the generating family of a constrained Lagrange submanifold that lives on the energy surface $H=0$ and that its level surfaces $Z=\text{const}$ yield Legendre submanifolds on that energy surface. Thus, the singularity structure of past null cones with apex at $\mathcal{I}^+$ is obtained by studying the projection map of the Legendre submanifolds to the configuration space. The behavior of the coordinate system defined by the variable Z at the caustic points is analyzed. It is shown that a single function $Z(x^a,\zeta,\overline{\zeta})$ cannot generate the conformal structure of an asymptotically flat space–time that satisfies the generic and weak energy condition. © *1999 American Institute of Physics.* [S0022-2488(99)02105-2]

I. INTRODUCTION

In the last several years a formalism has been developed where null surfaces play a dynamical role replacing the metric as a basic variable.[1–4] The goal of the null surface formulation of General Relativity (GR), or NSF for short, is to introduce a new variable such that from its knowledge one can obtain *all* the conformal structure of the space–time. Field equations equivalent to Einstein's equation then determine the dynamical evolution of those surfaces. By casting GR as a theory of surfaces rather than a theory of fields the NSF provides a completely new point of view with emphasis on the geometrical character of the theory. The basic variable is a function $Z(x^a,\zeta,\overline{\zeta})$ with x^a representing points on the space–time and $(\zeta,\overline{\zeta})$ parametrizing the sphere of null directions. At each point on the space–time the function Z satisfies

$$g^{ab}(x^c)\partial_a Z(x^c,\zeta,\overline{\zeta})\partial_b Z(x^c,\zeta,\overline{\zeta})=0, \tag{1}$$

and the level surfaces of this function, namely $Z=\text{const}$ are null hypersurfaces on the space–time. The reader should be aware that the above construction is done at a local level and that in general it might not be possible to find a single function satisfying these conditions on the whole space–time. Weyl curvature induces self-intersections and caustics on null congruences. Thus, even if one locally obtains a smooth hypersurface, extending such surface along the generators of its null geodesics will fail to be smooth. This generalized null surface is called a *wave front* and cannot be described as the level surface of a single function Z.

In this work we introduce a generalized variable $\hat{Z}$ to describe the wave fronts. By studying the geometry of these fronts we analyze the singularities of our variable. We are also able to determine under what circumstances a single function Z suffices to construct the entire conformal structure, or how many different functions must be given to cover the space–time. We only consider a specific class of space–times which are asymptotically flat along future null directions. For those space–times Z represents past null cones from the future null boundary $\mathcal{I}^+$.

Note that Z can be thought of as the action of the Hamilton–Jacobi equation for the Hamiltonian $H(x,p)=g^{ab}p_a p_b$. Since Eq. (1) can be written as

$$H(x,\partial_a Z)=0,$$

Z is the action of the time-independent Hamilton–Jacobi equation and the study of the unicity of the solution and its global properties can be carried out using the tools of analytical mechanics. It is worth mentioning that the study of the solutions of the Hamilton–Jacobi equations led to the development of the theory of Lagrange submanifolds on cotangent bundles and the loss of unicity on the solutions is directly related to the singularities of the projection map of these submanifolds onto the configuration space.[5–10]

We show that $\hat{Z}$ is the generating family of a constrained Lagrange submanifold that lives on the energy surface $H=0$ of the cotangent bundle T^*M and that its level surfaces are Legendre submanifolds. Thus, the singularity structure of the wave fronts can be obtained by studying the projection maps to the configuration space. We thus, define the caustic set as the points on the Lagrange or Legendre submanifold with singular projection and the projection of those points as the caustics. Since Lagrange and Legendre submanifolds are smooth surfaces in T^*M this work suggests that one can redefine the main variable of the NSF in a way that is free from the singularities and self-intersections that are naturally associated with characteristic wave fronts in GR.

In Sec. II we introduce the necessary mathematical background needed for this work. In this context we also prove that the hypersurfaces of a constrained Lagrange submanifold defined as the restriction of this Lagrange submanifold to the level surfaces of its generating family are Legendre submanifolds on the energy surface $H=\text{const}$.

In Sec. III we study the singularity structure of our variable Z and the main results are found. We show that the caustic points are obtained by choosing the points where $\eth\bar{\eth}Z$, the parameter space Laplacian of Z, blows up. We also show that at those points $(Z,\eth Z,\bar{\eth}Z)$ remain finite whereas $\eth^2 Z$ diverges. Using available singularity theorems we find as a proposition that a single function $Z(x^a,\zeta,\bar{\zeta})$ cannot generate the conformal structure of an asymptotical space–time that satisfies the generic and weak energy condition. Thus, in order to properly study the global behavior of the main variable in the NSF one must abandon the idea of using a single function on the space–time and instead one has to think of our variable as a generating family $\hat{Z}$ of a Lagrange submanifold on the cotangent bundle of the space–time. We close this work with some comments of how to deal with the dynamics of the new variable.

II. LAGRANGE AND LEGENDRE SUBMANIFOLDS

In this section we review the notions of Lagrange and Legendre manifolds in a given cotangent bundle T^*M of an n-dimensional manifold M. Exhaustive treatises at a high mathematical level and/or with applications to different areas in physics can be found in the literature.[5,7–13] We present here a brief review of some basic results needed for the present work. In Secs. II A and II B we introduce the concept of *constrained Lagrange and Legendre submanifolds* in order to reinterpret our variable Z as the generating family of a Lagrange manifold. Moreover, we prove in proposition II.3 that the hypersurfaces of a constrained Lagrange submanifold defined as the restriction of the Lagrange submanifold to the level surfaces of its generating family are Legendre submanifolds on the energy surface $H=\text{const}$.

A. Lagrange manifolds

Recall that (P,ω) is a symplectic manifold if P is an even-dimensional differentiable manifold and ω is a closed, nondegenerate, differential two-form on P. We consider a particular submanifold of P called a *Lagrange manifold*, defined as a manifold whose dimension is the same as the configuration space and where the pullback of the symplectic form vanishes on the manifold.

From now on we will restrict ourselves to a particular class of symplectic manifolds, the cotangent bundle of an n-dimensional manifold M, denoted by T^*M. This bundle can be assigned local coordinates (q^i,p_i), with (q^i) representing points of M and p_i the local coordinates of the covectors at the point (q^i). In these local coordinates the closed nondegenerate differential two-form ω on T^*M can be written as $\omega=dq^i\wedge dp_i$.

If L is a Lagrange submanifold of (T^*M,ω), then the projection map $\pi:T^*M\rightarrow M$ is called the *Lagrange map*. The set of points where the rank of π^* drops are called the *singular set* and the image of this set is called the *caustic*.

Notice also that if $S:M\rightarrow\mathbf{R}$, then the graph of dS is a Lagrange submanifold and π is a diffeomorphism.

The converse is also true, if π is locally a diffeomorphism, then L is the graph of dS, where $S:M\rightarrow\mathbf{R}$.

Now, consider the Hamiltonian system (T^*M,ω,H), where $H:T^*M\rightarrow\mathbf{R}$ is a Hamiltonian function. We introduce the notion of *constrained Lagrange submanifolds* as follows:

Definition II.1: Let $\hat{L}$ be a Lagrange submanifold of T^*M and H a Hamiltonian function, we say that $\hat{L}$ is a constrained Lagrange manifold if $\hat{L}\subset\hat{H}$, where $\hat{H}$ is an energy surface.

Note that constrained Lagrange manifolds are invariant under the flow of the Hamiltonian vector field X_H (Ref. 10, Proposition 5.3.32). This, can be easily proved using the fact that X_H is tangent to the hypersurface $H=\text{const}$ and that $\hat{L}$ is of maximal dimension.

If $\hat{L}$ is the graph of dS, where $S:M\rightarrow\mathbf{R}$, then S must satisfy the time-independent Hamilton–Jacobi equation

$$H\left(q^j,\frac{\partial S}{\partial q^i}\right)=\text{const}, \tag{2}$$

and conversely, a solution of (2) locally defines a constrained Lagrange manifold with a diffeomorphic projection to M.

However, in general (and in the problem we want to address) the projection of $\hat{L}$ will not be globally diffeomorphic to M.

Let us briefly analyze how to construct our constrained Lagrange manifold from the solutions to the time-independent Hamilton–Jacobi equation. From now on we shall assume that $\hat{H}$ is the hypersurface defined by $H=0$.

Since the Hamiltonian flow is tangent to $\hat{L}$, we may solve Hamilton's canonical equations with initial values $(q^i(0),p_i(0))\in\tilde{N}$, where $\tilde{N}$ is constructed as follows. Consider a hypersurface $N\subset M$ and a null covector field on N that is a restriction of a differential of some function on M, then this field defines an $n-1$ dimensional surface $\tilde{N}\subset\hat{H}$ diffeomorphic to N. The initial data shall be the Cauchy data of the Hamilton–Jacobi equation.

To solve Hamilton's canonical equations we look for a generating function of a canonical transformation $g:(q^i,p_i)\rightarrow(Q^i,P_i)$ such that in the new variables the Hamiltonian becomes a function which depends only on the variables Q^i, i.e., $H=K(Q^i)$.

It is easy to show that the difference between the canonical one-forms associated with the two coordinates is exact,[5,7–9] i.e.,

$$p_i dq^i - P_i dQ^i = dS(q^i,Q^j). \tag{3}$$

The function S is called *the generating function* of the canonical transformation. Furthermore, if $S(q^i,Q^j)$ is a solution to the differential equation

$$H\left(q^i,\frac{\partial S}{\partial q^j}\right)=K(Q^i), \tag{4}$$

with Q^i parameters such that $|\partial^2 S/(\partial Q^i \partial q^j)|\neq 0$, then Hamilton's canonical equations can be solved by quadratures and the functions $Q^i=Q^i(q^i,p_j)$ determined by the equations $p_i=\partial S/\partial q^i$ are first integrals of the canonical equations. (Jacobi's Theorem—Refs. 5 and 14).

Given a solution $S(q^i,Q^j)$ to Eq. (4) the corresponding constrained Lagrange manifold in $\hat{H}$, is obtained by first requiring that $K(Q^i)=0$. Then, the function $\hat{S}=S(q^i,Q^j)|_{K(Q^i)=0}$ is a generating family of $\hat{L}$ given by

$$\hat{L}=(q^i,p_j)|p_j=\frac{\partial\hat{S}}{\partial q_j} \text{ and } \frac{\partial\hat{S}}{\partial Q^i}=0.$$

Note that if we can solve for $Q^l=Q^l(q^i)$ from

$$K(Q^i)=0,\quad \frac{\partial S}{\partial Q^l}=0, \tag{5}$$

then $\hat{L}$ is locally the graph of dS, with $S(q^i)=S'(q^i,Q^j(q^i))$ and the *Lagrange map* $\overline{\pi}$ is a diffeomorphism. This can be guaranteed in a neighborhood of a point if the rank of the system built by (5) is $r=n$ in the variables Q^l. If $r=k<n$ then there exists $Q^J=Q^J(q^i)$, for $J=1\cdots k$ and $\hat{S}=\hat{S}(q^i,Q^I)$, for $I=k+1\cdots n$. In this case the rank of $\overline{\pi}^*$ drops at some points and this can be related to the presence of these parameters Q^I, for $I=k+1\cdots n$.

Then $\hat{S}(q^i,Q^I)$, defines a Lagrange submanifold $\hat{L}\subset\hat{H}$ embedded into T^*M by setting:

$$p_i=\frac{\partial\hat{S}}{\partial q^i},\quad 1\leqslant i\leqslant n, \tag{6}$$

and imposing the constraints

$$0=\frac{\partial\hat{S}}{\partial Q^I}. \tag{7}$$

Since the rank of (7) is $n-k$ in the q^I variables, then there exist $q^I=q^I(Q^I,q^J)$, and $\hat{L}$ and $\pi(\hat{L})$ are parametrized by (Q^I,q^J). The derivative $\overline{\pi}^*$ can be written as

$$\begin{pmatrix} \frac{\partial q^I}{\partial Q^I} & \frac{\partial q^I}{\partial q_J} \\ 0 & I \end{pmatrix},$$

where I is the identity matrix $k\times k$, therefore it is clear that the rank of $\pi^*\geqslant k$ and it shall be strictly less than n when

$$\left|\frac{\partial q^I}{\partial Q^I}\right|=0.$$

The set of singular points and thus the caustic set shall be isolated points, curves or in general a set of points of zero measure with respect to the topology of M, since the rank of $\overline{\pi}^*$ cannot drop more than $n-2$.[6]

B. Legendre manifolds

Odd-dimensional manifolds do not admit a *symplectic structure*. The analog of a symplectic structure for odd-dimensional manifolds is a *contact structure*.

We define a *contact manifold* as a pair $(\hat{P},\hat{\omega})$, consisting of an odd-dimensional manifold $\hat{P}$ and a closed two-form $\hat{\omega}$ of maximal rank on this manifold. An exact contact manifold $(\hat{P},\hat{\kappa})$ consists of a $(2n-1)$-dimensional manifold $\hat{P}$ and a one-form $\hat{\kappa}$ on $\hat{P}$ such that $\hat{\omega}=-d\hat{\kappa}$ is of maximal rank on $\hat{P}$.

Moreover, an analogous notion to a Lagrange manifold can also be given. An $(n-1)$-dimensional manifold N of $\hat{P}$ is called a *Legendre submanifold* if the pullback of the contact form $\hat{\kappa}$ vanishes on N.

Now, consider the Hamiltonian system (T^*M,ω,H), the next proposition ensures that we can find, in a natural way, a *contact submanifold* of T^*M.

Proposition II.2: Let (T^*M,ω,H) *be a Hamiltonian system and* $\hat{H}$ *a regular energy surface, defined by* $H=\text{const}$. *Then* $(\hat{H},i^*\omega)$ *is a contact manifold, where* $i:\hat{H}\to T^*M$ *is an inclusion (Ref. 10, Proposition 5.1.7).*

Thus the Legendre submanifolds we will consider are those that are submanifolds of the contact manifold $(\hat{H},i^*\omega)$. Moreover they are hypersurfaces of *constrained Lagrange manifold* of a given Hamiltonian system and the projection map $\pi:T^*M\to M$ will induce a map $\hat{\pi}$ defined as $\hat{\pi}=\pi\circ\hat{e}$ and called the *Legendre map*.

The next proposition gives a description of this kind of manifold.

Proposition II.3: Let $\hat{L}$ *be a constrained Lagrange submanifold of the Hamiltonian system* (T^*M,ω,H) *and* $\hat{S}$ *its generating family, i.e., a solution of the time independent Hamilton–Jacobi equation. Then the hypersurface* $\hat{N}$ *of* $\hat{L}$, *defined as the restriction of* $\hat{L}$ *to* $\hat{S}=\text{const}$, *is a Legendre submanifold of* $\hat{H}$.

Proof: Given a Hamiltonian system, Proposition II.2 ensures that $(\hat{H},i^*\kappa)$ is a contact manifold and since $\hat{L}$ is a constrained Lagrange manifold, the generating family $\hat{S}(q^i,Q^I)$ of $\hat{L}$ satisfies $H(q^i,\partial S/\partial q^j)=0$. Then $\hat{S}$ defines a Legendre submanifold $\hat{N}$ of $\hat{H}$ by setting

$$p_i=\frac{\partial\hat{S}}{\partial q^i},\quad 1\leqslant i\leqslant n, \tag{8}$$

imposing the constraints

$$\hat{S}=u_0,\quad \frac{\partial\hat{S}}{\partial Q^I}=0, \tag{9}$$

and requiring that the rank of (9) shall be $n-k+1$ in the q^I variables.

Recall that (9) is an algebraic nonlinear system of equations. Then if we define the function $\mathbf{G}=(\hat{S}-u_0,\partial\hat{S}/\partial Q^I)$, a solution of (9) satisfies $\mathbf{G}=\mathbf{0}$, that is, it belongs to the kernel of the map $\mathbf{G}$. Therefore demanding that the rank of the derivative of $\mathbf{G}$ be $n-k+1$ in the variables q^I and in one of the q^J variables, the implicit function theorem guarantees that $q^i=q^i(q^j,Q^I)$ for $i\in I+1$ and $j\in J-1$.

Observe that the Legendre manifold constructed in this way becomes an hypersurface of $\hat{L}$ and that both are submanifold of the energy surface $\hat{H}$. □

The image of the Legendre map is called the *wave front* and the image of the constrained Lagrange manifold can be considered as a wave front family. As in the case of the Lagrange manifolds, the set of points where the rank of $\hat{\pi}^*$ drops are called the *singular set* and the image of this set is called the *caustic*. If the singular set of $\hat{L}$ is known then intersection of this set with $\hat{S}=u_0$ yields the singular set of the associated Legendre submanifold.

III. THE FUNCTION Z IN ASYMPTOTICALLY FLAT SPACE–TIME

In this section we will use the results developed in Sec. II to analyze the behavior of null cones on asymptotically flat space–times with a future null boundary diffeomorphic to $S^2\times R$.[15] Those space–times represent compact objects that can emit gravitational radiation.

To define our variable Z we consider the intersection of the future null cone from x^a with the null boundary $\mathcal{I}^+$. This intersection is called a *light cone cut of null infinity* and in general it is a complicated surface with caustics, self-intersections, etc. Introducing Bondi coordinates $(u,\zeta,\bar{\zeta})$ on $\mathcal{I}^+$ [with u representing a Killing time and $(\zeta,\bar{\zeta})$ being stereographic coordinates on the unit sphere] it is possible to give a parametric representation of light cone cuts for points x^a close enough to $\mathcal{I}^+$ so that their future light cones have no caustic points before or at $\mathcal{I}^+$. Under those conditions the light cone cuts can be described as

$$u=Z(x^a,\zeta,\overline{\zeta}). \tag{10}$$

On Minkowski space–time the light cone cuts adopt a very simple form, $u=x^a l_a$ with l_a a null covector constructed from the spherical harmonics $Y_{0,0},Y_{1,-1},Y_{1,0},Y_{1,1}$. On an arbitrary asymptotically flat space–time, the cuts develop singularities and self-intersections. Thus, the function Z becomes, in general, multivalued. However, it can be shown that for regular space–times the index number is always one. Therefore, a light cone cut (as complicated as it might be) is always a continuous deformation of the sphere of null directions above each point x^a. It can also be shown that generically the light cone cuts can only have two kinds of singularities, cusps and swallowtails, since they represent the projection of 2-dim Legendre submanifolds on $\mathcal{I}^+$.[16]

A second meaning can be assigned to our variable $Z(x^a,\zeta,\overline{\zeta})$. Fixing a point $(u,\zeta,\overline{\zeta})$ of $\mathcal{I}^+$, the collection of interior points x^a that satisfy

$$Z(x^a,\zeta,\overline{\zeta})=u=\text{const}, \tag{11}$$

form the past null cone of $(u,\zeta,\overline{\zeta})$. Moreover, from knowledge of Z we can construct a null coordinate system as follows.

Starting with our variable and taking $(\zeta,\overline{\zeta})$ derivatives of $Z(x^a,\zeta,\overline{\zeta})$ we construct the following set of scalars;

$$\theta^i(x^a,\zeta,\overline{\zeta})\equiv(\theta^0,\theta^+,\theta^-,\theta^1)\equiv(u,w,\overline{w},R)\equiv(Z,\eth Z,\bar{\eth}Z,\eth\bar{\eth}Z). \tag{12}$$

For fixed values of $(\zeta,\overline{\zeta})$ they define a coordinate system with the following geometric meaning: $u=\text{const}$ denotes the past null cone from $(u,\zeta,\overline{\zeta})$; $(w,\overline{w})=\text{const}$ single out a null geodesic on that surface; $R=\text{const}$ identifies a point on that geodesic.

However, one knows that null cones can develop caustics and singularities. One also knows that past those singularities the null cone is no longer smooth (it is called a wave front) and thus, a null coordinate system like the one above breaks down past those singular points. Since the main goal of the NSF is to replace the metric with a function Z such that its level surfaces are past null cones from $\mathcal{I}^+$, we immediately face a nontrivial problem: If the null cones develop self-intersections and singularities that cannot be analyzed with a single function Z, then the construction given above is only valid on a neighborhood of $\mathcal{I}^+$. However, we also know that null wave fronts are projections of Legendre submanifolds that live on $T^*(M)$. It would then appear that the best way to deal with this lack of smoothness is to think of our variable as the generating family $\hat{Z}$ of a constrained Lagrange submanifold. An outline of this construction is presented below.

As was done before, we assume that our variable Z describing the past light cone from a point $(u,\zeta,\overline{\zeta})$ at $\mathcal{I}^+$ is a solution to the equation

$$H(x^a,\partial_b Z)=g^{ab}Z_{,a}Z_{,b}=0, \tag{13}$$

with g^{ab} a metric that is asymptotically flat.

In a neighborhood of $\mathcal{I}^+$ this solution is single valued since the (unphysical) metric is ''almost'' conformally flat and thus the past null cones are free from caustics and singularities.

For each value of $(\zeta,\overline{\zeta})$, one then uses this smooth function $Z(x^a,\zeta,\overline{\zeta})$ as the generating family of a constrained Lagrange manifold $\hat{L}$ given by

$$\hat{L}=\left\{\left(x^a,p_b=\frac{\partial Z}{\partial x^b}\right):e^*\kappa=dZ\right\}. \tag{14}$$

Note that Proposition II.3 ensures that the surface $\hat{N}$ defined by $Z=\text{const}$ is a Legendre submanifold of the energy surface given by $H=0$, i.e.,

$$\hat{N}=\left\{\left(x^a,p_b=\frac{\partial Z}{\partial x^b}\right):\hat{e}^*\hat{\kappa}=d(e^*Z)=0\right\}.$$

Thus, we have constructed a constrained Legendre submanifold $\hat{N}$ and a constrained Lagrange submanifold $\hat{L}$ of $\hat{H}$ using our fundamental variable $Z(x^a,\zeta,\overline{\zeta})$.

The idea is to extend this construction to regions where caustics develop. As we mentioned before, in these regions the Lagrange submanifold is not diffeomorphic to its projection. We will thus assume that the generating family describing the past null cone from $(u,\zeta,\overline{\zeta})$ can be written as $\hat{Z}=\hat{Z}(x^a,w,\overline{w},\zeta,\overline{\zeta})$ with $(w,\overline{w})$ parameters labeling the past geodesics. Note that the function $\hat{Z}$ depends on two parameters since the rank of the projection map cannot drop more than two.[5] Note also that $(w,\overline{w})$ are two of our null coordinates θ^i since we are considering $\hat{Z}$ as the generating function of the canonical transformation between x^a and θ^i restricted to $\hat{H}$. The constrained Lagrange submanifold $\hat{L}$ is then given by

$$p_a=\frac{\partial\overline{Z}}{\partial x^a},$$

together with the constraint

$$\frac{\partial\hat{Z}}{\partial w}=0,\quad \frac{\partial\hat{Z}}{\partial\overline{w}}=0. \tag{15}$$

Observe that if we can solve (15) uniquely for $(w,\overline{w})$ in a neighborhood of a point, i.e., $(w,\overline{w})=(w(x^b),\overline{w}(x^b))$, then $\hat{Z}=Z$ and we are back in the previous diffeomorphic region. In general, one will obtain multivalued solutions of (15). Inserting the different solutions of $(w,\overline{w})$ into $\hat{Z}(x^a,w,\overline{w},\zeta,\overline{\zeta})$ one obtains a multiple-valued function $Z(x^a,\zeta,\overline{\zeta})$. The Legendre submanifold is obtained by setting $\hat{Z}=\text{const}$. Conversely, if several functions Z_i are given, one can reconstruct the Lagrange submanifold by imposing (14) on the different Z's. The construction defines the Lagrange submanifold except for the caustic set.

Finally, we would like to determine under what circumstances it is possible to find a single function Z that would yield for us a global coordinate system $(u,R,w,\overline{w})$ on an asymptotically flat space–time. In other words we want to know if there exists space–times that are diffeomorphic to the corresponding Lagrange manifolds. At the same time we would like to know when and how this coordinate system breaks down due to the presence of conjugate points. We are therefore interested in describing the relationship between our fundamental variable Z and the loss of the rank of the derivative of the Legendre map $\hat{\pi}$, i.e., we want to described the singular set in terms of Z.

When the Lagrange manifold is a constrained one, the loss of rank of the Lagrange map indicates the nonexistence of global solutions of the Hamilton–Jacobi equation and the loss of rank of the associated Legendre map is related to the *existence of conjugate points of a congruence of null geodesics*.

In order to clarify this assertion, we consider the local description of the wave front (the projection of the Legendre manifold). We assume that the wave front is locally described by

$$x^a=f^a(u_0,s,w,\overline{w},\zeta_0,\overline{\zeta_0})$$

with s an affine length. The vectors

$$L^a = \frac{\partial f^a}{\partial s}, \quad M^a = \frac{\partial f^a}{\partial w}, \quad \bar{M}^a = \frac{\partial f^a}{\partial \bar{w}}$$

are tangent to the wave front. L^a is directed along the null geodesics whereas M^a and $\bar{M}^a$ are geodesic deviation vectors.

The derivative of the Legendre map losses its rank when these three vectors become linearly dependent. This dependence is related to the existence of conjugate points on the congruence of null geodesic with apex at $\mathcal{I}^+$ and null tangent vector L^a as follows.

We introduce the parallel propagated null triad $\{l^a, m^a, \bar{m}^a\}$, satisfying

$$l^a m_a = 0 \quad m^a \bar{m}_a = -1 \quad l^a \nabla_a m^b = 0. \tag{16}$$

In terms of this triad,

$$L^a = l^a, \quad M^a = \xi m^a + \overline{\eta m^a}, \quad \bar{M}^a = \bar{\xi} \bar{m}^a + \eta m^a, \tag{17}$$

therefore this set of vectors becomes linearly dependent when

$$\begin{vmatrix} \xi & \eta \\ \bar{\eta} & \bar{\xi} \end{vmatrix} = (\xi\bar{\xi} - \eta\bar{\eta}) = 0. \tag{18}$$

On the other hand, this quantity is related to the divergence ρ, and the shear σ of the congruence with apex in $\mathcal{I}^+$. To see this, consider the optical parameters

$$\rho = m^a \bar{m}^b \nabla_a l_b, \quad \sigma = m^a m^b \nabla_a l_b.$$

Using Eq. (17) together with the fact that M^a is Lie propagated along the null direction L^a we get (Ref. 3)

$$\sigma = \frac{\bar{\eta}^2}{A} D\left(\frac{\bar{\xi}}{\bar{\eta}}\right), \quad \rho = \frac{DA}{2A}$$

with $A = (\xi\bar{\xi} - \eta\bar{\eta})$ and where we have used the fact that ρ is real.

Hence, at those points where the Legendre map looses its rank, the divergence of the congruence becomes unbounded, i.e., $\lim \rho_{s \to s_0} = \infty$, where s is the affine parameter and s_0 corresponds to a conjugate point.

Note that $Z(x^a, \zeta, \bar{\zeta})$ is a single valued function for points x^a near $\mathcal{I}^+$. Therefore, $(u, R, w, \bar{w})$ is a well-behaved coordinate system since there are no conjugate points in that region. The question is what happens to our coordinates as we approach a generic conjugate point. Given that our variable has a second meaning, namely, the intersection of the future light cone from x^a with $\mathcal{I}^+$, a related question is what happens to this cut as the apex recedes into the interior of the space–time until a caustic develops at $\mathcal{I}^+$. We first answer the latter question since the results obtained are then used in the analysis of the behavior of the coordinates $(u, R, w, \bar{w})$ near a caustic point.

A. Light cone cuts

We consider the future light cone from a point x^a near $\mathcal{I}^+$ such that the intersection with the null boundary is locally described by a single valued $Z(x^a, \zeta, \bar{\zeta})$.

Lemma III.1. If the apex x^a moves into the interior until the cut develops a generic caustic point, then at this first conjugate point the components of the extrinsic curvature of the cut given by $R = \eth\bar{\eth}Z$ and $\Lambda = \eth^2 Z$ become infinite.

Proof: We first introduce Bondi coordinates $(\Omega, u, \zeta, \bar{\zeta})$ in a neighborhood of $\mathcal{I}^+$ such that this boundary is described by $\Omega = 0$. We then introduce a null geodesic λ that connects the point x^a

with $\mathcal{I}^+$ and an affine length s such that the apex is labeled as s_0. Furthermore, we suppose that the description of the null cone is given by $F=0$, where $F=F(\Omega,u,\zeta,\bar{\zeta},s_0)$. Near $\mathcal{I}^+$ the function F can be written as

$$F=F^0(u,\zeta,\bar{\zeta},s_0)+\Omega F^1(u,\zeta,\bar{\zeta},s_0)+\mathcal{O}(\Omega^2).$$

where $F^0=u-Z$. Then, the divergence and the shear of the light cone congruence with apex at s_0 containing the null geodesic λ and defined by (see Ref. 17)

$$\rho=m^a\bar{m}^b\nabla_aF_b,\quad \sigma=m^am^b\nabla_aF_b,$$

respectively, are calculated at $\mathcal{I}^+$ as

$$\rho(s_0,\mathcal{I}^+)=m^a\bar{m}^b\nabla_a(u-Z)=\rho_B-R(s_0), \tag{19}$$

$$\sigma(s_0,\mathcal{I}^+)=m^am^b\nabla_a(u-Z)=\sigma_B-\Lambda(s_0), \tag{20}$$

where ρ_B and σ_B are the divergence and the shear of a Bondi congruence.

Now, we move the apex x^a (i.e., s_0) into the interior along the null geodesic λ until the cut develops a caustic point. Since at a generic conjugate point at $\mathcal{I}^+$, the divergence and the shear of the light cone congruence become infinite, whereas ρ_B and σ_B are bounded quantities, we prove the statement. □

It is worth mentioning that it is possible to find degenerate conjugate points where the shear goes to zero instead of infinity. Those, however, are not generic singularities since they are removable by a small perturbation of the initial values of the optical parameters in the geodesic deviation equation.

B. Past null cones from $\mathcal{I}^+$

As was shown before, our function $Z(x^a,\zeta,\bar{\zeta})$ is a generating family of past null cone congruences with apex $(u,\zeta,\bar{\zeta})$ at $\mathcal{I}^+$. Furthermore, for fixed values of $(\zeta,\bar{\zeta})$ this function generates a null coordinate system $(u,R,w,\bar{w})$ which is then used in the derivation of the most important results in the NSF formulation. It is therefore very relevant to analyze the range of validity of this coordinate system. To do so we use a reciprocity theorem for null congruences together with the previous lemma.

We first state a reciprocity theorem relating null cone congruences.[18,19]

Theorem: *Given two null cone congruences having a common null geodesic* λ, *denoting by* X_1 *and* X_2 *the matrices whose elements are the tetrad components of the complex deviation vectors associated with the null cone congruences with apex at a point* p_1 *and* p_2 *along* λ, *then*

$$X_1(at\ \ p_2)=-X_2(at\ \ p_1).$$

We now prove the following lemma:

Lemma III.2. Assume $Z=\text{const}$ *describes the past null cone from* $(u,\zeta,\bar{\zeta})$ *at* $\mathcal{I}^+$. *Then, at a conjugate point* $R\to-\infty$ *and* $|\Lambda|\to\infty$.

Proof: Consider the past null cone from $(u,\zeta,\bar{\zeta})$ at $\mathcal{I}^+$, take a geodesic labeled by $(u,w,\bar{w})$ on this congruence, introduce an affine length s on this geodesic, and denote by ρ_1,σ_1 the optical parameters associated with this congruence. If a conjugate point is reached at $s=s_0$ then at this point ρ_1 and σ_1 become infinite.

On the other hand, if we consider the future light cone congruence from s and denote by ρ_2, σ_2 the corresponding optical parameters, then the reciprocity theorem shows that this congruence has a conjugate point at $(u,\zeta,\bar{\zeta})$ when $s\to s_0$. Thus, $\rho_2(s_0,\mathcal{I}^+)\to\infty$, $|\sigma_2(s_0,\mathcal{I}^+)|\to\infty$ and from Lemma III.1 $R(s_0)\to-\infty$ and $|\Lambda(s_0)|\to\infty$. □

The first consequence of this lemma is that our coordinate system is well defined in the domain $R \in (-\infty, \infty)$, in other words from our coordinate system we cannot detect the caustics that arise in the past null cones as we move into the space–time. It is easy to show that $w = \eth Z$ and $\overline{w} = \bar{\eth} Z$ remain finite at a conjugate point. This follows from the fact that both are constants along a null geodesic.

The lemma is also useful to answer the question we posed before, namely, if there exist asymptotically flat space–times that can be covered with a global canonical coordinate system constructed from a single function *Z*. Using proposition 4.4.5 (Ref. 15) which states that on an asymptotically flat space–time that satisfies the generic and weak energy conditions, any null congruence along a geodesic such that the affine length can be extended arbitrarily has a pair of conjugate points and lemma III.1 we conclude that the coordinate system $(u, w, \overline{w}, R)$ derived from *Z* cannot cover such a space–time. We conclude that

Proposition III.3: A single function $Z(x^a, \zeta, \overline{\zeta})$ cannot generate the conformal structure of an asymptotically flat space–time that satisfies the weak energy and generic conditions.

It is also of interest to analyze the behavior of the conformal factor Ω, and the metric components near a caustic point.

It can be shown that the conformal factor can be written as[3]

$$\Omega^2 = g^{01} := g^{ab} Z_{,a} \eth \bar{\eth} Z_{,b} = \frac{dR}{ds}.$$

Since $R(s)$ diverges as s approaches a conjugate point while the affine length is a smooth nonvanishing function along the null geodesic, it follows that g^{01} also blows up at that point.

IV. CONCLUSIONS

We have shown that our main variable $\hat{Z}$ is the generating family of a constrained Lagrange submanifold and that its level surfaces are constrained Legendre submanifolds that project down to past null cones from $\mathcal{I}^+$. For a generic asymptotically flat space–time the projection of this Lagrange submanifold starts diffeomorphic to the configuration space but it later develops caustic sets. Thus, except for Minkowski space, a single function *Z* on configuration space does not give the conformal structure of the space–time. At the caustic points, $\eth \bar{\eth} Z$ diverges. This means that the coordinate system constructed on the null cones is only locally defined but on the other hand one never sees the caustics since they are pushed out to $R = -\infty$.

Although the entire treatment so far has been kinematical, we would like to think of our variable as coming from the solution of a set of field equations given on the space–time.[3,4]

It is clear from the previous results that the solution of those field equations must have multiple valuedness in order to generate the multiple branches needed to construct the generating family $\hat{Z}$ of the Lagrange submanifold. These solutions are defined in a six-dimensional space, four-space–time coordinates and two parameters on the sphere, $(\zeta, \overline{\zeta})$. We demand the solution to be globally defined with respect to the parameters $(\zeta, \overline{\zeta})$, that is, it shall be a piecewise smooth function on the sphere (it could be multiple valued but always finite on the sphere).

Alternatively, we could try to find field equations given on T^*M. In this case the solution would yield a global generating family $\hat{Z}$ of a constrained Lagrange submanifold that coincides with *Z* in a neighborhood of $\mathcal{I}^+$. This last approach will be further explored.

ACKNOWLEDGMENTS

We are indebted to Ted. E. Newman and Simonetta Frittelli for many enlightening conversations and suggestions. We also thank an anonymous referee for carefully reading and improving the manuscript. This research has been partially supported by AIT, CONICET, and CONICOR.

[1] C. N. Kozameh and E. T. Newman, ''Theory of light cone cuts of null infinity,'' J. Math. Phys. **24**, 2481–2489 (1983).
[2] S. Frittelli, E. T. Newman, and C. Kozameh, ''GR via characteristic surfaces,'' J. Math. Phys. **36**, 4984–5004 (1995).
[3] S. Frittelli, E. T. Newman, and C. Kozameh, ''On the dynamics of characteristic surfaces,'' J. Math. Phys. **36**, 6397–6416 (1995).
[4] M. Iriondo, C. Kozameh, and A. Rojas, ''Null surfaces and the Bach equations,'' J. Math. Phys. **38**, 4714–29 (1997).
[5] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, edited by P. R. Halmos, F. W. Gehring, and C. C. Moore (Springer, New York, 1978).
[6] R. Abraham and J. Robbin, *Transversal Mappings and Flows* (Benjamin, New York, 1967), p. 37.
[7] V. I. Arnold, S. M. Gusein-Zade, and A. N. Varchenko, *Singularities of Differentiable Maps* (Birkhauser, Boston, 1985), Vol. I.
[8] V. I. Arnold, *Catastrophe Theory* (Springer, Berlin, 1986).
[9] *Dynamical Systems*, edited by V. I. Arnold and S. P. Novikov (Springer, Berlin, 1992), Vol. IV.
[10] R. Abraham and J. Marsden, *Foundations of Mechanics* (Benjamin, New York, 1978).
[11] H. Friedrich and J. Stewart, ''Characteristic initial data wave front singularities in general relativity,'' Proc. R. Soc. London, Ser. A **385**, 345 (1983).
[12] M. Golubitsky and D. Schaeffer, ''A theory for imperfect bifurcation via singularity theory,'' Commun. Pure Appl. Math. **32**, 21–98 (1979).
[13] I. Stewart, ''Applications of catastrophe theory to the physical sciences,'' Physica D **2**, 245–305 (1981).
[14] L. D. Landau and E. M. Lifshitz, *Curso de Física Teórica, Mecánica* (Reverté, Barcelona, 1970), Vol. I.
[15] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge University Press, Cambridge, 1973).
[16] C. N. Kozameh, P. W. Lamberti, and O. A. Reula, ''Global aspects of light cone cuts,'' J. Math. Phys. **32**, 3423–3426 (1991).
[17] C. Cutler, C. N. Kozameh, and E. T. Newman, *A Two-Surface Encoding of Radiative Space-Times, Twistors in Mathematical Physics*, edited by T. Bailey and R. Baston (Cambridge University Press, Cambridge, 1990).
[18] R. Penrose, *Perspective in Geometry and Relativity*, edited by Hoffman (Indiana University Press, Cambridge, 1984).
[19] I. Etherington, Philos. Mag. **15**, 761 (1933).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Hopf algebra structure of $\mathrm{Gr}_q(1|1)$ related to $\mathrm{GL}_q(1|1)$

Salih Çelik[a)]
Mimar Sinan University, Department of Mathematics, 80690 Besiktas, Istanbul, Turkey

(Received 13 January 1998; accepted for publication 20 November 1998)

We show that the algebra of functions on the Grassmann supergroup $\mathrm{Gr}_q(1|1)$ has a (graded) Hopf algebra structure related to $\mathrm{GL}_q(1|1)$. © *1999 American Institute of Physics.* [S0022-2488(99)01603-5]

In the past few years, quantum groups[1] and q-deformed universal enveloping algebras[2] have been intensively studied both by mathematicians and mathematical physicists. From a mathematical point of view, these algebraic structures are just special classes of noncommutative Hopf algebras.

The algebraic structure underlying quantum groups extends the theory of the supergroups.[3] The simplest quantum supergroup is $\mathrm{GL}_q(1|1)$, i.e., the deformation of the supergroup of 2×2 matrices with two bosonic (even) and two fermionic (odd) matrix entries.

The aim of the present work is to construct the (graded) Hopf algebra structure of the Grassmann supergroup $\mathrm{Gr}_q(1|1)$, the superdual of $\mathrm{GL}_q(1|1)$, which was introduced in Ref. 4. Before discussing the (graded) Hopf algebra structure of $\mathrm{Gr}_q(1|1)$, let us first give some notations and useful formulas about the quantum Grassmann supergroup $\mathrm{Gr}_q(1|1)$.

A Grassmann supermatrix $\hat{T}$ which is an element of $\mathrm{Gr}(1|1)$ is of the form

$$\hat{T}=\begin{pmatrix}\alpha & b\\ c & \delta\end{pmatrix}$$

with two odd (Greek letters) and two even (Latin letters) matrix elements. The symbol *hat* is used to distinguish $\hat{T}$ from an element T of $\mathrm{GL}_q(1|1)$.

The q-deformation of the Grassman supergroup $\mathrm{Gr}(1|1)$ as a quantum matrix supergroup $\mathrm{Gr}_q(1|1)$ is generated by α, b, c, δ with the relations[4]

$$\begin{aligned}&\alpha b=q^{-1}b\alpha,\quad \alpha c=q^{-1}c\alpha,\\ &\delta b=q^{-1}b\delta,\quad \delta c=q^{-1}c\delta,\\ &\alpha\delta+\delta\alpha=0,\quad \alpha^2=0=\delta^2,\\ &bc=cb+(q-q^{-1})\delta\alpha,\end{aligned}\tag{1}$$

where q is a nonzero complex number and $q^2\neq1$. The associative algebra (1) is equivalent to equation[5]

$$R^1\hat{T}_1\hat{T}_2=-\hat{T}_2\hat{T}_1R^2,\tag{2}$$

where

[a)]Electronic mail: scelik@fened.msu.edu.tr

$$R^1=\begin{pmatrix} q & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & q-q^{-1} & -1 & 0 \\ 0 & 0 & 0 & q^{-1} \end{pmatrix}, \quad R^2=\begin{pmatrix} q^{-1} & 0 & 0 & 0 \\ 0 & -1 & q^{-1}-q & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & q \end{pmatrix} \tag{3}$$

are both the solutions of the quantum (graded) Yang–Baxter equation. Here, we used the tensoring convention

$$(\hat{T}_1)^{ij}_{kl}=(\hat{T}\otimes I)^{ij}_{kl}=\hat{T}^i_k\delta^j_l,$$

$$(\hat{T}_2)^{ij}_{kl}=(I\otimes\hat{T})^{ij}_{kl}=(-1)^{i(j+l)}\hat{T}^j_l\delta^i_k. \tag{4}$$

The central element of the algebra (1) is[4]

$$\hat{D}_q=bc^{-1}-\alpha c^{-1}\delta c^{-1}=c^{-1}b-c^{-1}\alpha c^{-1}\delta. \tag{5}$$

We now denote the algebra generated by the elements α, b, c, δ with the relations (1) by $\hat{\mathcal{A}}$. We want to make the algebra $\hat{\mathcal{A}}$ into a (graded) Hopf algebra related to the quantum supergroup $\mathrm{GL}_q(1|1)$. Because of this, we state briefly some properties of the quantum supergroup $\mathrm{GL}_q(1|1)$ we are going to need in this work.

The quantum supergroup $\mathrm{GL}_q(1|1)$ is generated by four generators a, β, γ, d and the q-commutation relations[3]

$$\begin{aligned} a\beta&=q\beta a, \quad d\beta=q\beta d,\\ a\gamma&=q\gamma a, \quad d\gamma=q\gamma d,\\ \beta\gamma+\gamma\beta&=0, \quad \beta^2=0=\gamma^2,\\ ad&=da+(q-q^{-1})\gamma\beta. \end{aligned} \tag{6}$$

The generators satisfying the relations (6) generate the algebra called the algebra of functions on the quantum supergroup $\mathrm{GL}_q(1|1)$ and we shall denote it by $\mathcal{A}$. We know that the algebra $\mathcal{A}$ is a (graded) Hopf algebra whose structure we now discuss. We represent the set of generators a, β, γ, d in the form of a matrix

$$T=\begin{pmatrix} a & \beta \\ \gamma & d \end{pmatrix}.$$

Then the relations (6) are equivalent to equation

$$RT_1T_2=T_2T_1R, \tag{7}$$

where

$$R=\begin{pmatrix} q & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & q-q^{-1} & 1 & 0 \\ 0 & 0 & 0 & q^{-1} \end{pmatrix}. \tag{8}$$

The superinverse of T is given by[3]

$$T^{-1}=\begin{pmatrix} A & \Omega \\ \Gamma & D \end{pmatrix}=\begin{pmatrix} a^{-1}+a^{-1}\beta d^{-1}\gamma a^{-1} & -a^{-1}\beta d^{-1} \\ -d^{-1}\gamma a^{-1} & d^{-1}+d^{-1}\gamma a^{-1}\beta d^{-1} \end{pmatrix}, \tag{9}$$

and the superdeterminant is

$$D_q=ad^{-1}-\beta d^{-1}\gamma d^{-1}. \tag{10}$$

It is easy to verify that D_q commutes with all matrix elements of T. Note that the matrix elements of T with those of T^{-1} satisfy the following relations:

$$\begin{gathered} aA=q^2Aa+1-q^2, \quad dA=Ad, \\ aD=Da, \quad dD=q^2Dd+1-q^2, \\ a\Omega=q\Omega a, \quad d\Omega=q\Omega d, \\ a\Gamma=q\Gamma a, \quad d\Gamma=q\Gamma d, \\ \beta A=qA\beta, \quad \gamma A=qA\gamma, \\ \beta D=qD\beta, \quad \gamma D=qD\gamma, \\ \beta\Omega=\Omega\beta, \quad \gamma\Omega=-q^2\Omega\gamma, \\ \beta\Gamma=-q^2\Gamma\beta, \quad \gamma\Gamma=\Gamma\gamma. \end{gathered} \tag{11}$$

The usual coproduct is given by

$$\Delta:\mathcal{A}\rightarrow\mathcal{A}\otimes\mathcal{A}, \quad \Delta(t^i_j)=t^i_k\otimes t^k_j, \tag{12}$$

where summation over repeated indices is understood. One can rewrite the last formula in the following nice and elegant form,

$$\Delta(T)=T\dot{\otimes}T, \tag{13}$$

where $\otimes$ stands for the usual tensor product and the dot refers to the summation over repeated indices and reminds us about the usual matrix multiplication. The counit is given by

$$\varepsilon:\mathcal{A}\rightarrow\mathcal{C}, \quad \varepsilon(t^i_j)=\delta^i_j. \tag{14}$$

The coinverse (antipode) is given by

$$S:\mathcal{A}\rightarrow\mathcal{A}, \quad S(T)=T^{-1}. \tag{15}$$

It is not difficult to verify the following properties of the co-structures:

$$(\Delta\otimes\mathrm{id})\circ\Delta=(\mathrm{id}\otimes\Delta)\circ\Delta, \tag{16}$$

$$\mu\circ(\varepsilon\otimes\mathrm{id})\circ\Delta=\mu'\circ(\mathrm{id}\otimes\varepsilon)\circ\Delta, \tag{17}$$

$$m\circ(S\otimes\mathrm{id})\circ\Delta=\varepsilon=m\circ(\mathrm{id}\otimes S)\circ\Delta, \tag{18}$$

where id denotes the identity mapping,

$$\mu:\mathcal{C}\otimes\mathcal{A}\rightarrow\mathcal{A}, \quad \mu':\mathcal{A}\otimes\mathcal{C}\rightarrow\mathcal{A} \tag{19}$$

are the canonical isomorphisms, defined by

$$\mu(k\otimes a)=ka=\mu'(a\otimes k), \quad \forall a\in\mathcal{A}, \quad \forall k\in\mathcal{C}, \tag{20}$$

and m is the multiplication map

$$m:\mathcal{A}\otimes\mathcal{A}\rightarrow\mathcal{A}, \quad m(a\otimes b)=ab. \tag{21}$$

The multiplication in $\mathcal{A}\otimes\mathcal{A}$ follows the rule

$$(A\otimes B)(C\otimes D)=(-1)^{p(B)p(C)}AC\otimes BD, \tag{22}$$

where $p(X)$ is the z_2-grade of X, i.e., $p(X)=0$ for even variables and $p(X)=1$ for odd variables.

Since the (graded) Hopf algebra structure of $\mathrm{Gr}_q(1|1)$ is related to those of $\mathrm{GL}_q(1|1)$ it is necessary to obtain the commutation relations of the generators of $\hat{\mathcal{A}}$ with those of $\mathcal{A}$. We define the (mixed) commutation relations between the generators of $\hat{\mathcal{A}}$ and $\mathcal{A}$ as follows:

$$R\hat{T}_1T_2=(-1)^{p(T_2)}T_2\hat{T}_1R', \quad R'=R-(q-q^{-1})P, \tag{23}$$

where P is the superpermutation matrix. The equation (23) gives the mixed relations

$$\begin{aligned}
&a\alpha=q^2\alpha a, \quad \beta\alpha=-q\alpha\beta,\\
&ab=qba+(q^2-1)\alpha\beta, \quad \beta b=b\beta,\\
&ac=qca+(q^2-1)\alpha\gamma, \quad \beta c=c\beta+(q-q^{-1})\alpha d,\\
&a\delta=\delta a+(q-q^{-1})(\beta c-b\gamma), \quad \beta\delta=-q^{-1}\delta\beta+(1-q^{-2})bd,\\
&d\alpha=\alpha d, \quad \gamma\alpha=-q\alpha\gamma,\\
&db=q^{-1}bd, \quad \gamma b=b\gamma-(q-q^{-1})\alpha d,\\
&dc=q^{-1}cd, \quad \gamma c=c\gamma,\\
&d\delta=q^{-2}\delta d, \quad \gamma\delta=-q^{-1}\delta\gamma+(1-q^{-2})cd.
\end{aligned} \tag{24}$$

Using these relations, it is easy to verify that $\hat{D}_q$, which is given by (5), is still a central element, i.e., $\hat{D}_q$ also commutes with the generators of $\mathcal{A}$.

After some algebra the commutation relations between the matrix elements of $\hat{T}$ with T^{-1} are obtained to be

$$\begin{aligned}
&\alpha A=q^2A\alpha, \quad \delta A=A\delta,\\
&\alpha D=D\alpha, \quad \delta D=q^{-2}D\delta+(q-q^{-1})^2A\alpha+(q^{-2}-1)(\Omega c-\Gamma b),\\
&\alpha\Omega=-q\Omega\alpha, \quad \delta\Omega=-q^{-1}\Omega\delta+(q^{-1}-q)Ab,\\
&\alpha\Gamma=-q\Gamma\alpha, \quad \delta\Gamma=-q^{-1}\Gamma\delta+(q^{-1}-q)Ac,\\
&bA=qAb, \quad cA=qAc,\\
&bD=q^{-1}Db+(q-q^{-1})\Omega\alpha, \quad cD=q^{-1}Dc+(q-q^{-1})\Gamma\alpha.\\
&b\Omega=\Omega b, \quad c\Omega=\Omega c+(q^2-1)A\alpha,
\end{aligned} \tag{25}$$

$$b\Gamma=\Gamma b+(1-q^2)A\alpha,\quad c\Gamma=\Gamma c,$$

and

$$D_q u=q^2 u D_q,\quad u\in\{\alpha,b,c,\delta\}. \tag{26}$$

Before defining a coproduct on the algebra $\hat{\mathcal{A}}$, let us note the following facts. Let $\hat{T}$ and $\hat{T}'$ be any two supercommuting matrices that satisfy (1). We denote a product $\hat{T}\hat{T}'$ by T. Then, it can be verified that the matrix elements of T satisfy the commutation relations (6) of $\mathrm{GL}_q(1|1)$, i.e., if

$$T=\begin{pmatrix}\alpha & b\\ c & \delta\end{pmatrix}\begin{pmatrix}\alpha' & b'\\ c' & \delta'\end{pmatrix},$$

then we have the relations (6). In short,

$$\hat{T},\hat{T}'\in \mathrm{Gr}_q(1|1)\Rightarrow T=\hat{T}\hat{T}'\in \mathrm{GL}_q(1|1).$$

In view of these facts, we can say that there may be no coproduct of the usual form $\Delta(\hat{T})=\hat{T}\dot{\otimes}\hat{T}$. For this coproduct, if it existed, would be invariant under the q-commutation relations (6) of $\mathrm{GL}_q(1|1)$. But we can define a map on the algebra $\hat{\mathcal{A}}$ as follows:

$$\hat{\Delta}:\hat{\mathcal{A}}\to\hat{\mathcal{A}}\otimes\hat{\mathcal{A}},\quad \hat{\Delta}(\hat{T})=\hat{T}\dot{\otimes}T+(-1)^{p(T)}T\dot{\otimes}\hat{T}. \tag{27}$$

Explicitly,

$$\begin{aligned}
\hat{\Delta}(\alpha)&=\alpha\otimes a+b\otimes\gamma+a\otimes\alpha-\beta\otimes c,\\
\hat{\Delta}(b)&=b\otimes d+\alpha\otimes\beta+a\otimes b-\beta\otimes\delta,\\
\hat{\Delta}(c)&=c\otimes a+\delta\otimes\gamma-\gamma\otimes\alpha+d\otimes c,\\
\hat{\Delta}(\delta)&=\delta\otimes d+c\otimes\beta-\gamma\otimes b+d\otimes\delta.
\end{aligned} \tag{28}$$

The action on the generators of $\hat{\mathcal{A}}$ of $\hat{\varepsilon}:\hat{\mathcal{A}}\to\mathcal{C}$ is

$$\hat{\varepsilon}(\alpha)=\hat{\varepsilon}(b)=\hat{\varepsilon}(c)=\hat{\varepsilon}(\delta)=0. \tag{29}$$

Finally, we define the coinverse as

$$\hat{S}:\hat{\mathcal{A}}\to\hat{\mathcal{A}},\quad \hat{S}(\hat{T})=-(-1)^{p(T^{-1})}T^{-1}\hat{T}T^{-1}. \tag{30}$$

The action of $\hat{S}$ on the generators of $\hat{\mathcal{A}}$ is

$$\begin{aligned}
\hat{S}(\alpha)&=-(\alpha A+b\Gamma)A+q(cA+\delta\Gamma)\Omega,\\
\hat{S}(b)&=-(\alpha A-c\Omega)\Omega-q(bA+\delta\Omega)D,\\
\hat{S}(c)&=-(\alpha A+b\Gamma)\Gamma-q(cA+\delta\Gamma)D,\\
\hat{S}(\delta)&=q^2(\alpha A-c\Omega)D+q(\alpha\Omega+q^2 bD)\Gamma-(\alpha A+q^2\delta D)D.
\end{aligned} \tag{31}$$

It is not difficult to check that the maps $\hat{\Delta}$ and $\hat{\varepsilon}$ are both algebra homomorphisms and $\hat{S}$ is an algebra anti-homomorphism and also the three maps satisfy the properties (16)–(18), and they preserve the relations (24) provided that the action on the generators of $\mathcal{A}$ of $\hat{\Delta}$ is the same with (13).

The coproduct, counit, and coinverse which are specified above supply $\mathrm{Gr}_q(1|1)$ with a structure, which can be called a quasi-Hopf algebra.

It is interesting to note that there is a close connection with the differential calculus[6] on the quantum supergroup $\mathrm{GL}_q(1|1)$ via the equation (23). In fact, we have observed that the matrix elements of $\hat{T} \in \mathrm{Gr}_q(1|1)$ are just the differentials of the matrix elements of $T \in \mathrm{GL}_q(1|1)$. In other words, we can interpret the generating elements of $\mathrm{Gr}_q(1|1)$ as differentials of coordinate functions on $\mathrm{GL}_q(1|1)$. In this case, we can write $\hat{T} = \mathsf{d}T$ (more information on these issues are given in Ref. 6). Then the extended algebra can be interpreted as an algebra of differential forms on $\mathrm{GL}_q(1|1)$. Thus the coproduct is interpreted as a (left and right) coaction of the quatum supergroup $\mathrm{GL}_q(1|1)$ on differential forms. To this end, we consider the two maps

$$\Delta_R : \Gamma \to \Gamma \otimes \mathcal{A}, \quad \Delta_R \circ \mathsf{d} = (\mathsf{d} \otimes \mathrm{id}) \circ \Delta \tag{32a}$$

and

$$\Delta_L : \Gamma \to \mathcal{A} \otimes \Gamma, \quad \Delta_L \circ \mathsf{d} = (\tau \otimes \mathsf{d}) \circ \Delta, \tag{32b}$$

where Γ denotes the differential algebra of $\mathcal{A}$. Here τ: $\Gamma \to \Gamma$ is the linear map of degree zero which gives $\tau(a) = (-1)^{p(a)} a$. We now define a map ϕ_R as follows,

$$\phi_R(u_1 \mathsf{d} v_1 + \mathsf{d} v_2 u_2) = \Delta(u_1) \Delta_R(\mathsf{d} v_1) + \Delta_R(\mathsf{d} v_2) \Delta(u_2), \tag{33}$$

and another map ϕ_L by replacing L with R. The following identities are satisfied.

$$(\phi_R \otimes \mathrm{id}) \circ \phi_R = (\mathrm{id} \otimes \Delta) \circ \phi_R, \quad (\mathrm{id} \otimes \epsilon) \circ \phi_R = \mathrm{id}, \tag{34a}$$

and

$$(\mathrm{id} \otimes \phi_L) \circ \phi_L = (\Delta \otimes \mathrm{id}) \circ \phi_L, \quad (\epsilon \otimes \mathrm{id}) \circ \phi_L = \mathrm{id}. \tag{34b}$$

Consequently, we define the map $\hat{\Delta}$, in (27), as

$$\hat{\Delta} = \phi_R + \phi_L. \tag{35}$$

ACKNOWLEDGMENTS

This work was supported in part by **T.B.T.A.K.**, the Turkish Scientific and Technical Research Council.

I would like to express my deep gratitude to the referee for critical comments and suggestions on the manuscript.

[1] N. Y. Reshetikhin, L. A. Takhtajan, and L. D. Faddeev, Leningrad Math. J. **1**, 193 (1990).
[2] V. G. Drinfeld, ''Quantum groups,'' in Proc. IMS, Berkeley (1986).
[3] Yu. I. Manin, Commun. Math. Phys. **123**, 163 (1989); E. Corrigan, D. Fairlie, P. Fletcher, and R. Sasaki, J. Math. Phys. **31**, 776 (1990); S. Schwenk, B. Schmidke, and S. Vokos, Z. Phys. C **46**, 643 (1990); B. Schmidke, S. Vokos, and B. Zumino, *ibid.* **48**, 249 (1990).
[4] S. Çelik and S. A. Çelik, Balkan Phys. Lett. **3**, 188 (1995).
[5] S. Çelik, J. Math. Phys. **37**, 3568 (1996); S. Çelik, Balkan Phys. Lett. **5**, 149 (1997).
[6] S. Çelik and S. A. Çelik, ''On the differential geometry of $\mathrm{GL}_q(1|1)$,'' J. Phys. A **31**, 9685 (1998).

SU(N)- and SO(N)-invariant chiral fields: One- and two-dimensional subspaces

Tonatiuh Matos[a] and Ulises Nucamendi[b]
Departamento de Física, Centro de Investigación y de Estudios Avanzados del I.P.N., A. P. 14-740, 07000, D. F., México

Petra Wiederhold[c]
Departamento de Ingenieria Eléctrica, Centro de Investigación y de Estudios Avanzados del I.P.N., A. P. 14-740, 07000, D. F., México

(Received 9 July 1997; accepted for publication 15 October 1998)

Applying the harmonic map ansatz we find new classes of SU(N) and SO(N) chiral fields. We write the corresponding one-dimensional subgroups of SU(N) and SO(N) chiral fields in terms of one and two harmonic maps and the corresponding two-dimensional subgroups in terms of SU(2) and SO(3) chiral fields. In other words, we reduce the SU(N) and SO(N) chiral equations to harmonic maps in one- and two-dimensional spaces.
[S0022-2488(99)01204-9]

I. INTRODUCTION

One of the most important and interesting models in mathematical physics are the models called nonlinear σ models or principal chiral models. These are important for different reasons: they have similar properties to the four-dimensional Yang–Mills theories,[1] they have an infinite number of conservation laws,[1,2] they contain soliton solutions,[3,4] and they are used as approximate models in particle physics. Their applications as field equations in various theories also makes these models very attractive and interesting. In general relativity and gravitational theories, the Einstein's field equations in a space–time with two Killing vectors reduce to a nonlinear σ model with invariant group SU(1,1) for the vacuum case, and with invariant group SU(2,1) for the electrovacuum case.[5] The n-dimensional Einstein's equations with a $(n-2)$-dimensional isometry group reduce to a nonlinear σ model with invariant group $\mathrm{SL}(n-2,\mathcal{R})$.[6] But one of the most interesting features of the σ models is that self-dual Yang–Mills fields and self-dual general relativity reduce to the field equations of the σ models[5] (see Refs. 7 and 8). In the first case the field equations reduce to a SU(N) nonlinear σ model, and in the last case the field equations reduce to a $\mathrm{SU}(\infty)\sim\mathrm{Diff}(\Sigma^2)$-$\sigma$ model.[7,8] In this work we deal with SU(N) and SO(N)-σ models. These are defined by means of the following Lagrangian density:

$$\mathcal{L}=\frac{\alpha}{4}\mathrm{Tr}(g_{,z}g^{-1}g_{,\bar{z}}g^{-1}), \tag{1}$$

where $g\in\mathrm{SU}(N)$, $(\mathrm{SO}(N))$, i.e.,

$$gg^{\dagger}=1 \quad (gg^{T}=1), \quad \mathrm{Det}\, g=\alpha^2=1, \tag{2}$$

[a]Electronic mail: tmatos@fis.cinvestav.mx
[b]Electronic mail: ulises@fis.cinvestav.mx
[c]Electronic mail: biene@ctrl.cinvestav.mx

where z is a complex coordinate and $\overline{z}$ its complex conjugate, g^T is the transpose of g, and $g^\dagger = \overline{g}^T$. Even if $\alpha = 1$ for the groups studied here, we shall write it explicitly; the only difference appears in the harmonic map equations, but our results are written in terms of these harmonic maps, which will be supposed to be arbitrary in this work. The corresponding Euler–Lagrange equations of (2) are given by

$$(\alpha g_{,z} g^{-1})_{,\overline{z}} + (\alpha g_{,\overline{z}} g^{-1})_{,z} = 0, \tag{3}$$

which are the chiral equations.

Solutions to chiral equations have been found by the application of various methods, e.g., by Bäcklund transformations (see Ref. 1), by the inverse scattering method (see Refs. 3 and 4), and recently, Uhlenbeck has deduced the form of all classical solutions of the chiral equations with constraints from a factorization theorem proved by herself.[9] In this work we will apply the method developed by Matos in Ref. 10, which permits us to suppose g to be an arbitrary element of a Lie group G, where the dimension of G can even be infinite. The most important feature of this method is that we can write the chiral fields, i.e., the solutions of the chiral equations, in terms of explicit functions which are solutions of a well-known differential equation like the Laplace equation or minimal surfaces equation in two dimensions. We obtain classes of solutions in terms of these arbitrary harmonic maps. The solutions can be then interpreted physically. In this work we will give a set of exact solutions of the chiral equations for the Lie groups SU(N) and SO(N) in terms of harmonic maps. In order to make this paper self-contained, we will resume the harmonic map ansatz method of Ref. 10 in Sec. II. In Sec. III we will analyze one-dimensional subspaces, and in Sec. IV the two-dimensional ones.

II. THE HARMONIC MAP ANSATZ METHOD

In this section we briefly give a general treatment of the harmonic map ansatz. For full details we refer the reader to Ref. 10. Let G be a Lie group and

$$g:\mathcal{C}\times\mathcal{C}\to G, \quad (z,\overline{z})\mapsto g(z,\overline{z}) \in G, \tag{4}$$

where g is a solution of (3) and it depends on the complex variables z and $\overline{z}$. Let w_R be the Maurer–Cartan form in G:

$$w_R = dg\, g^{-1}, \quad g \in G, \tag{5}$$

then w_R is right invariant on G, that is

$$R_c^*(dg\, g^{-1}) = dg\, g^{-1}. \tag{6}$$

R_c is the right translation associated with the element $c \in G$. Furthermore, we define the functions A_z and $A_{\overline{z}}$ by

$$g\mapsto A_z(g) = g_{,z} g^{-1}, \quad g\mapsto A_{\overline{z}}(g) = g_{,\overline{z}} g^{-1}, \tag{7}$$

where $A_z, A_{\overline{z}} \in \mathcal{G}$, $\mathcal{G}$ being the corresponding Lie algebra of G. Then, using (7), the Maurer–Cartan form in G can be written in terms of the Lie algebra valued functions A_z and $A_{\overline{z}}$ as

$$w_R = A_z\, dz + A_{\overline{z}}\, d\overline{z} = g_{,z} g^{-1}\, dz + g_{,\overline{z}} g^{-1}\, d\overline{z}. \tag{8}$$

Now, let us suppose that the matrix g depends on a set of functions $\lambda^i(z,\overline{z})$, which are local coordinates of an arbitrary Riemannian manifold M_r, that is

$$\lambda^i: M_r \to \Re, \quad \forall i = 1,2,...,r, \quad g = g(\lambda^i), \quad \forall i = 1,2,...,r. \tag{9}$$

The harmonic map ansatz consists of assuming that the parameters λ^i build minimal surfaces in the Riemannian manifold M_r, that is, they satisfy the following equation:

$$(\alpha\lambda^i{}_{,z})_{,\bar{z}}+(\alpha\lambda^i{}_{,\bar{z}})_{,z}+2\alpha\Gamma^i_{jk}\lambda^j{}_{,z}\lambda^k{}_{,\bar{z}}=0, \quad \forall i=1,2,...,r, \tag{10}$$

where Γ^i_{jk} represents the Christoffel's symbols in the Riemannian manifold M_r. If we do so, the chiral equations reduce to a Killing equation in the Riemannian manifold M_r.

In order to see this, we substitute (9) in the chiral equations (3), getting

$$\alpha[(g_{,i}g^{-1})_{,j}+(g_{,j}g^{-1})_{,i}]\lambda^i_{,z}\lambda^i_{,\bar{z}}+g_{,i}g^{-1}[(\alpha\lambda^i_{,z})_{,\bar{z}}+(\alpha\lambda^i_{,\bar{z}})_{,z}]=0. \tag{11}$$

The substitution of (10) in (11) gives

$$[(g_{,i}g^{-1})_{,j}+(g_{,j}g^{-1})_{,i}-2\Gamma^k_{ij}g_{,k}g^{-1}]\lambda^j{}_{,\bar{z}}\lambda^i{}_{,z}=0. \tag{12}$$

From Eq. (12) it follows

$$(g_{,i}g^{-1})_{,j}+(g_{,j}g^{-1})_{,i}-2\Gamma^k_{ij}g_{,k}g^{-1}=0. \tag{13}$$

Analogously, we can define the functions

$$A_i:G\rightarrow\mathcal{G}, \quad g\mapsto A_i(g)=g_{,i}g^{-1}, \quad \forall i=1,2,...,r. \tag{14}$$

Again, the matrices $A_i(g)$ belong to the Lie algebra of G. Then, the Maurer–Cartan matrix one-form w can be written as

$$w(g)=A_i(g)d\lambda^i=g_{,i}g^{-1}\,d\lambda^i. \tag{15}$$

The covariant derivative in the Riemannian manifold M_r is given by

$$\nabla_jA_i=A_{i,j}-\Gamma^k_{ij}A_k. \tag{16}$$

Using (14) and (16) in (13), it is easy to see that the chiral equations (3) become

$$\nabla_jA_i+\nabla_iA_j=0. \tag{17}$$

The matrices $A_i(g)$ also satisfy the relation

$$\nabla_jA_i-\nabla_iA_j-[A_j,A_i]=0; \tag{18}$$

the last relation means that the matrices $A_i(g)$ are pure gauge potentials. (17) and (18) together reduce to

$$\nabla_jA_i=\tfrac{1}{2}[A_j,A_i]. \tag{19}$$

Let the matrices σ^j be a basis of the Lie algebra $\mathcal{G}$. We can write the matrices A_i in this basis,

$$A_i(g)=\epsilon_i^j\sigma_j. \tag{20}$$

Using (20) and (17), we obtain

$$[\nabla_j\epsilon_i^k+\nabla_i\epsilon_j^k]\sigma_k=0. \tag{21}$$

Because the matrices σ^k are linearly independent we can conclude that

$$\nabla_j\epsilon_i^k+\nabla_i\epsilon_j^k=0. \tag{22}$$

Equation (22) is the Killing equation on the Riemannian manifold M_r. Using the following relation, which is satisfied for any Killing vector field ϵ^k:

$$\nabla_l \nabla_j \epsilon_i^k = R_{lji}^m \epsilon_m^k, \tag{23}$$

we find that the covariant derivative of the Riemann tensor on the Riemannian manifold M_r vanishes, then M_r is a symmetric manifold. Any symmetric manifold contains $r(r+1)/2$ linearly independent Killing vector fields.[11] Then if we know the Riemannian manifold M_r we have the Killing vector space. Let $\{\epsilon^k\}$ be a basis of Killing fields on M_r, and write the matrices A_i in terms of this basis. The substitution of (20) in (19) yields the algebra of the $\{\sigma^k\}$ matrices. The next step is to choose a representation of this algebra and to solve the following system of first-order differential equations:

$$A_i(g) = g_{,i} g^{-1}, \quad \forall i = 1,2,...,r. \tag{24}$$

Observe that the field equations (3) are invariant under the transformation

$$g \mapsto g' = L_h g = hg, \quad h \in G_c, \tag{25}$$

where $G_c = \{c \in G | c \text{ is a constant matrix}\}$ and L_h is the left action of G_c on G. In other words, if g is solution of the field equations (3), then g' is solution of them too. Under the transformation (25), the matrices $A_i(g)$ transform into A_i':

$$A_i \mapsto A_i' = h A_i h^{-1}. \tag{26}$$

The relation (26) separates the set of matrices $\{A_i\}$ in equivalent classes. Now let us work with a representation of each class. After integrating the equations (24) for this representation, we can find all the solutions in this class to the field equations (3) by means of the transformation (25). The problem of solving the chiral equations is then reduced to finding harmonic maps, i.e., to solve the system of coupled linear differential equations (10) for the parameters λ^i.

III. ONE-DIMENSIONAL SUBSPACES

One-dimensional subspaces are characterized by a one-dimensional manifold M_1, with only one parameter λ, and a metric given by

$$ds^2 = d\lambda^2. \tag{27}$$

The metric (27) is symmetric and has a constant Killing vector field. The harmonic map ansatz is then

$$g = g(\lambda), \quad \lambda = \lambda(z, \bar{z}). \tag{28}$$

From (13) we can see that A_λ is a constant matrix [A_λ belongs to the corresponding Lie algebra of SU(N)]. The Lie algebra su(N), corresponding to SU(N), is the set of traceless anti-Hermitian complex $N \times N$ matrices

$$\mathrm{su}(N) = \{A \in GL(N,C) | A = -A^\dagger, \ \mathrm{Tr}\, A = 0\}. \tag{29}$$

For the su(N) algebra, the Jordan normal form is always diagonal. We use this form by integrating the matrix differential equation $g_{,\lambda} = A_\lambda g$. We have only the following representation for N even, $N = 2n$:

$$A_\lambda = \begin{pmatrix} a_1 i & 0 & \cdot & \cdot & 0 & 0 \\ 0 & -a_1 i & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & a_n i & 0 \\ 0 & 0 & \cdot & \cdot & 0 & -a_n i \end{pmatrix}, \tag{30}$$

and the following one for N odd, $N=2n+1$:

$$A_\lambda = \begin{pmatrix} a_1 i & 0 & \cdot & \cdot & 0 & 0 & 0 \\ 0 & -a_1 i & \cdot & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & a_n i & 0 & 0 \\ 0 & 0 & \cdot & \cdot & 0 & -a_n i & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 0 & 0 \end{pmatrix}, \tag{31}$$

where $a_1,\ldots,a_n \in \mathcal{R}$, $a_i \neq 0$, $i=1,...,n$.

For the case $N=2n$, we integrate the differential equation $g_{,\lambda}=A_\lambda g$, and obtain

$$g = \begin{pmatrix} u_{1,1}\exp(a_1 i\lambda) & \cdot & \cdot & \cdot & \cdot & u_{1,2n}\exp(a_1 i\lambda) \\ u_{2,1}\exp(-a_1 i\lambda) & \cdot & \cdot & \cdot & \cdot & u_{2,2n}\exp(-a_1 i\lambda) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_{2n-1,1}\exp(a_n i\lambda) & \cdot & \cdot & \cdot & \cdot & u_{2n-1,2n}\exp(a_n i\lambda) \\ u_{2n,1}\exp(-a_n i\lambda) & \cdot & \cdot & \cdot & \cdot & u_{2n,2n}\exp(-a_n i\lambda) \end{pmatrix}. \tag{32}$$

Under the conditions $g^{-1}=g^\dagger$, $\det g=1$, the $(2n)^2$ complex constants $u_{i,j}$, $i,j=1,...,2n$ satisfy

$$\begin{pmatrix} u_{1,1} & \cdot & \cdot & \cdot & \cdot & u_{1,2n} \\ u_{2,1} & \cdot & \cdot & \cdot & \cdot & u_{2,2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_{2n-1,1} & \cdot & \cdot & \cdot & \cdot & u_{2n-1,2n} \\ u_{2n,1} & \cdot & \cdot & \cdot & \cdot & u_{2n,2n} \end{pmatrix} \in \mathrm{SU}(N). \tag{33}$$

Analogously, for the case $N=2n+1$ we obtain

$$g = \begin{pmatrix} u_{1,1}\exp(a_1 i\lambda) & \cdot & \cdot & \cdot & \cdot & \cdot & u_{1,2n+1}\exp(a_1 i\lambda) \\ u_{2,1}\exp(-a_1 i\lambda) & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2,2n+1}\exp(-a_1 i\lambda) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_{2n-1,1}\exp(a_n i\lambda) & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2n-1,2n+1}\exp(a_n i\lambda) \\ u_{2n,1}\exp(-a_n i\lambda) & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2n,2n+1}\exp(-a_n i\lambda) \\ u_{2n+1,1} & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2n+1,2n+1} \end{pmatrix}. \tag{34}$$

Under the restrictions $g^{-1}=g^\dagger$, $\det g=1$, the $(2n+1)^2$ complex constants $u_{i,j}$, $i,j=1,...,2n+1$ satisfy

$$\begin{pmatrix} u_{1,1} & \cdot & \cdot & \cdot & \cdot & \cdot & u_{1,2n+1} \\ u_{2,1} & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2,2n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_{2n-1,1} & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2n-1,2n+1} \\ u_{2n,1} & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2n,2n+1} \\ u_{2n+1,1} & \cdot & \cdot & \cdot & \cdot & \cdot & u_{2n+1,2n+1} \end{pmatrix} \in \mathrm{SU}(N). \tag{35}$$

From (10) we see that the parameter λ must satisfy the Laplace equation, that is

$$(\alpha\lambda_{,z})_{,\bar{z}} + (\alpha\lambda_{,\bar{z}})_{,z} = 0. \tag{36}$$

For each solution λ of the Laplace equation, we obtain a different matrix solution of the chiral equations by means of matrices (32) and (34).

IV. TWO-DIMENSIONAL SUBSPACES

A. Two-dimensional subalgebras of su(N)

Two-dimensional subspaces are characterized by a two-dimensional manifold M_2 with parameters $\lambda^1 = \lambda$, $\lambda^2 = \tau$, and a metric given by

$$ds^2 = \frac{d\lambda\, d\tau}{(1 + k\lambda\tau)^2}. \tag{37}$$

The metric (37) is symmetric and has a constant curvature k. The harmonic map ansatz is then

$$g = g(\lambda, \tau), \quad \lambda = \lambda(z, \bar{z}), \quad \tau = \tau(z, \bar{z}). \tag{38}$$

The matrices A_λ, and A_τ, satisfy Eqs. (17) and (19), this is,

$$\begin{aligned} &\nabla_\lambda A_\tau + \nabla_\tau A_\lambda = 0, \\ &\Delta_\lambda A_\tau = \tfrac{1}{2}[A_\lambda, A_\tau], \\ &\nabla_\tau A_\lambda = \tfrac{1}{2}[A_\tau, A_\lambda]. \end{aligned} \tag{39}$$

The Killing vector space of the metric (37) can be obtained by solving the Killing equation (22)

$$\begin{aligned} &\epsilon_{\lambda,\lambda} + \frac{2k\tau}{(1 + k\lambda\tau)} \epsilon_\lambda = 0, \\ &\epsilon_{\tau,\tau} + \frac{2k\lambda}{(1 + k\lambda\tau)} \epsilon_\tau = 0, \\ &\epsilon_{\lambda,\tau} + \epsilon_{\tau,\lambda} = 0. \end{aligned} \tag{40}$$

The components of the Killing vector fields are

$$\epsilon_\lambda = \frac{1}{V^2}(ak\tau^2 + b\tau + c),$$

$$\epsilon_\tau = \frac{1}{V^2}(ck\lambda^2 - b\lambda + a), \tag{41}$$

with

$$V = (1 + k\lambda\tau),$$

We choose three independent Killing vectors on the manifold M_2:

$$\phi = \frac{1}{V^2}(1, k\lambda^2), \quad \xi = \frac{1}{V^2}(-2\tau, 2\lambda), \quad \zeta = \frac{1}{V^2}(k\tau^2, 1). \tag{42}$$

Using Eq. (20) we can define the matrix vector $A = (A_\lambda, A_\tau)$ by

$$A = \phi\sigma_1 + \xi\sigma_2 + \zeta\sigma_3. \tag{43}$$

The matrices σ_1, σ_2, σ_3 form a basis of a Lie subalgebra of su(N). These matrices have the following expressions:

$$A_\lambda = \frac{1}{V^2}[\sigma_1 - 2\tau\sigma_2 + k\tau^2\sigma_3], \quad A_\tau = \frac{1}{V^2}[k\lambda^2\sigma_1 + 2\lambda\sigma_2 + \sigma_3]. \tag{44}$$

Introducing Eqs. (44) into (39) we obtain the commutation relations of the subalgebra of σ_k matrices

$$[\sigma_1, \sigma_2] = 2k\sigma_1, \quad [\sigma_2, \sigma_3] = 2k\sigma_3, \quad [\sigma_3, \sigma_1] = -4\sigma_2. \tag{45}$$

We take the case $k=0$ and $\sigma_2 = 0$. The manifold M_2 is therefore flat. The commutation relations of the two-dimensional Abelian subalgebras of su(N) are

$$[\sigma_3, \sigma_1] = 0. \tag{46}$$

From (44), the matrices A_λ, A_τ are reduced to

$$g_{,\lambda}g^{-1} = A_\lambda = \sigma_1, \quad g_{,\tau}g^{-1} = A_\tau = \sigma_3. \tag{47}$$

Since the group SU(N) has range $(N-1)$, its algebra has $(N-1)$ diagonal commuting elements. The two-dimensional Abelian subalgebras of su(N) can be written in the basis σ_1, σ_3, where

$$\sigma_1 = \begin{pmatrix} a_1 i & 0 & \cdot & \cdot & 0 & 0 \\ 0 & a_2 i & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & a_N i \end{pmatrix},$$

and

$$\sigma_3=\begin{pmatrix} b_1 i & 0 & \cdot & \cdot & 0 & 0 \\ 0 & b_2 i & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & b_N i \end{pmatrix}, \tag{48}$$

here $a_1+\cdots+a_N=0$, $a_1,\ldots,a_N\in\mathfrak{R}$, $b_1+\cdots+b_N=0$, $b_1,\ldots,b_N\in\mathfrak{R}$.

The constants $b_1,\ldots,b_N,a_1,\ldots,a_N$ must be chosen in such a way that the matrices σ_1 and σ_3 are linearly independent. We integrate the differential equations $g_{,\lambda}=A_\lambda g$, and $g_{,\tau}=A_\tau g$, we obtain

$$g=\begin{pmatrix} u_{1,1}\exp(a_1 i\lambda+b_1 i\tau) & \cdot & \cdot & \cdot & u_{1,N}\exp(a_1 i\lambda+b_1 i\tau) \\ u_{2,1}\exp(a_2 i\lambda+b_2 i\tau) & \cdot & \cdot & \cdot & u_{2,N}\exp(a_2 i\lambda+b_2 i\tau) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ u_{N,1}\exp(a_N i\lambda+b_N i\tau) & \cdot & \cdot & \cdot & u_{N,N}\exp(a_N i\lambda+b_N i\tau) \end{pmatrix},$$

where the complex constants satisfy the condition

$$\begin{pmatrix} u_{1,1} & \cdot & \cdot & \cdot & u_{1,N} \\ u_{2,1} & \cdot & \cdot & \cdot & u_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ u_{N,1} & \cdot & \cdot & \cdot & u_{N,N} \end{pmatrix}\in \mathrm{SU}(N). \tag{49}$$

From Eq. (10), we see that the parameters λ and τ satisfy the Laplace equation

$$(\alpha\lambda_{,z})_{,\bar{z}}+(\alpha\lambda_{,\bar{z}})_{,z}=0, \quad (\alpha\tau_{,z})_{,\bar{z}}+(\alpha\tau_{,\bar{z}})_{,z}=0. \tag{50}$$

For each solution λ and τ of the Laplace equation we obtain a new matrix solution for the chiral equations by applying the matrix (49).

B. Three-dimensional subalgebras of su(*N*) and so(*N*)

Let us consider now a nonflat manifold M_2 with constant curvature $k\neq 0$. Let $\lambda^1=x$, $\lambda^2=y$, be the local coordinates of M_2. Then we choose the following metric on M_2,

$$ds^2=\frac{dx^2+dy^2}{[1+k(x^2+y^2)]^2}. \tag{51}$$

The matrices g depend on the coordinates x,y,

$$g=g(x,y), \quad x=x(z,\bar{z}), \quad y=y(z,\bar{z}). \tag{52}$$

For convenience, we perform a transformation from the variables x,y in the complex variables $w,\bar{w}$, defined by

$$w=x+iy, \quad \bar{w}=x-iy. \tag{53}$$

Then the matrices g are functions of the variables $w,\bar{w}$,

$$g=g(w,\bar{w}), \quad w=w(z,\bar{z}), \quad \bar{w}=\bar{w}(z,\bar{z}). \tag{54}$$

Equations (17) and (19), rewritten in terms of the complex variables w and $\overline{w}$, become

$$\nabla_w A_{\overline{w}} + \nabla_{\overline{w}} A_w = 0, \quad \nabla_w A_{\overline{w}} = \tfrac{1}{2}[A_w, A_{\overline{w}}], \quad \nabla_{\overline{w}} A_w = \tfrac{1}{2}[A_{\overline{w}}, A_w]. \tag{55}$$

Metric (51) can be written in terms of the new variables $w, \overline{w}$ as

$$ds^2 = \frac{dw\, d\overline{w}}{(1+kw\overline{w})^2}. \tag{56}$$

The Killing equations for the metric (56) are

$$\epsilon_{w,w} + \frac{2k\overline{w}}{(1+kw\overline{w})}\epsilon_w = 0, \quad \epsilon_{\overline{w},\overline{w}} + \frac{2kw}{(1+kw\overline{w})}\epsilon_{\overline{w}} = 0, \quad \epsilon_{w,\overline{w}} + \epsilon_{\overline{w},w} = 0. \tag{57}$$

Solving the above differential equations, we obtain

$$\epsilon_w = \frac{1}{V^2}(yk\overline{w}^2 + ai\overline{w} + \overline{\gamma}), \quad \epsilon_{\overline{w}} = \frac{1}{V^2}(\overline{\gamma}kw^2 - aiw + \gamma), \tag{58}$$

with

$$V = (1+kw\overline{w}), \quad \gamma \in C, \quad a \in \mathfrak{R}.$$

Now, we choose three linearly independent Killing vectors

$$\phi = \frac{1}{V^2}(\gamma k\overline{w}^2 + \overline{\gamma}, \overline{\gamma}kw^2 + \gamma),$$

$$\xi = \frac{1}{V^2}(-ai\overline{w}, aiw), \tag{59}$$

$$\zeta = \frac{1}{V^2}(\overline{\gamma}k\overline{w}^2 + \gamma, \gamma kw^2 + \overline{\gamma}).$$

Then the matrix vector $A = (A_w, A_{\overline{w}})$ can be written in the form

$$A = \phi\sigma_1 + \xi\sigma_2 + \zeta\sigma_3, \tag{60}$$

where the matrices $A = (A_w, A_{\overline{w}})$ read

$$A_w = \frac{1}{V^2}[(\gamma k\overline{w}^2 + \overline{\gamma})\sigma_1 - ai\overline{w}\sigma_2 + (\overline{\gamma}k\overline{w}^2 + \gamma)\sigma_3],$$

$$A_{\overline{w}} = \frac{1}{V^2}[(\overline{\gamma}kw^2 + \gamma)\sigma_1 + aiw\sigma_2 + (\gamma kw^2 + \overline{\gamma})\sigma_3]. \tag{61}$$

Introducing Eq. (60) in Eq. (55), we obtain the commutation relations for the σ_k matrices,

$$[\sigma_1, \sigma_2] = \sigma_3, \quad [\sigma_2, \sigma_3] = \sigma_3, \quad [\sigma_3, \sigma_1] = \sigma_2, \tag{62}$$

where we have chosen the constants $\gamma = 1 + i$, $a = 2$, $k = \frac{1}{2}$.

The three-dimensional subalgebra of su(N), fulfilling relations (62), is the three-dimensional algebra su(2). A vector basis of the three-dimensional subalgebra of su(N) in terms of the su(2) basis is given by $\{\sigma_1, \sigma_2, \sigma_3\}$, where

$$\sigma_1=\frac{1}{2}\begin{pmatrix} 0 & i & 0 & \cdot & \cdot & 0\\ i & 0 & 0 & \cdot & \cdot & 0\\ 0 & 0 & 0 & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix}, \quad \sigma_2=\frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & \cdot & \cdot & 0\\ -1 & 0 & 0 & \cdot & \cdot & 0\\ 0 & 0 & 0 & \cdot & \cdot & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix},$$
$$\sigma_3=\frac{1}{2}\begin{pmatrix} -i & 0 & 0 & \cdot & \cdot & 0\\ 0 & i & 0 & \cdot & \cdot & 0\\ 0 & 0 & 0 & \cdot & \cdot & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix}. \tag{63}$$

Introducing this basis in (61) we obtain for $A=(A_w,A_{\bar{w}})$,

$$A_w=\frac{i}{2V^2}\begin{pmatrix} -(\bar{\gamma}k\bar{w}^2+\gamma) & (\gamma k\bar{w}^2+\bar{\gamma}-a\bar{w}) & 0 & \cdot & \cdot & 0\\ (\gamma k\bar{w}^2+\bar{\gamma}+a\bar{w}) & (\bar{\gamma}k\bar{w}^2+\gamma) & 0 & \cdot & \cdot & 0\\ 0 & 0 & 0 & \cdot & \cdot & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix},$$
$$A_{\bar{w}}=\frac{i}{2V^2}\begin{pmatrix} -(\gamma kw^2+\bar{\gamma}) & (\bar{\gamma}kw^2+\gamma+aw) & 0 & \cdot & \cdot & 0\\ (\bar{\gamma}kw^2+\gamma-aw) & (\gamma kw^2+\bar{\gamma}) & 0 & \cdot & \cdot & 0\\ 0 & 0 & 0 & \cdot & \cdot & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix}. \tag{64}$$

Now we go into the group SU(N). This can be done by means of the exponential map or by integration. Let us integrate them. We must solve the matrix differential system

$$g_{,w}g^{-1}=A_w, \quad g_{,\bar{w}}g^{-1}=A_{\bar{w}}, \tag{65}$$

writing (65) in components, it is easy to obtain the next $2N^2$-differential equations

$$g_{1i,w}=\frac{i}{2V^2}[-(\bar{\gamma}k\bar{w}^2+\gamma)g_{1i}+(\gamma k\bar{w}^2+\bar{\gamma}-a\bar{w})g_{2i}],$$
$$g_{2i,w}=\frac{i}{2V^2}[(\gamma k\bar{w}^2+\bar{\gamma}+a\bar{w})g_{1i}+(\bar{\gamma}k\bar{w}^2+\gamma)g_{2i}],$$
$$g_{ji,w}=0,$$
$$g_{1i,\bar{w}}=\frac{i}{2V^2}[-(\gamma kw^2+\bar{\gamma})g_{1i}+(\bar{\gamma}kw^2+\gamma+aw)g_{2i}],$$
$$g_{2i,\bar{w}}=\frac{i}{2V^2}[(\bar{\gamma}kw^2+\gamma-aw)g_{1i}+(\gamma kw^2+\bar{\gamma})g_{2i}],$$
$$g_{ji,\bar{w}}=0, \tag{66}$$

where $i=1,2,...,N$, and $j=3,4,...,N$.

This system has the solution

$$g_{1i}(w,\overline{w})=\frac{[2d_{0i}+(\gamma f_{0i}-\overline{\gamma}d_{0i})i\overline{w}+(\overline{\gamma}f_{0i}-\gamma d_{0i})iw-d_{0i}w\overline{w}]}{2(1+kw\overline{w})},$$

$$g_{2i}(w,\overline{w})=\frac{[2f_{0i}+(\gamma f_{0i}+\overline{\gamma}d_{0i})iw+(\overline{\gamma}f_{0i}+\gamma d_{0i})i\overline{w}-f_{0i}w\overline{w}]}{2(1+kw\overline{w})}, \tag{67}$$

$$g_{ji}(w,\overline{w})=c_{ji}, \quad c_{ji}\in\mathbf{C},$$

for $i=1,2,...,N$, and $j=3,4,...,N$.

Because this solution (67) must fulfill the constraint $g^{-1}=g^{\dagger}$, $\det g=1$, the constants d_{0i},f_{0i},c_{ji}, must satisfy,

$$\begin{pmatrix} d_{01} & d_{02} & \cdot & \cdot & \cdot & d_{0N} \\ f_{01} & f_{02} & \cdot & \cdot & \cdot & f_{0N} \\ c_{31} & c_{32} & \cdot & \cdot & \cdot & c_{3N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{N1} & c_{N2} & \cdot & \cdot & \cdot & c_{NN} \end{pmatrix} \in \mathrm{SU}(N). \tag{68}$$

For the so(N) algebra, the three-dimensional subalgebra that fulfills the relations (62), is the so(3) algebra which has the three basis elements

$$\sigma_1=\begin{pmatrix} 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & -1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix}, \quad \sigma_2=\begin{pmatrix} 0 & 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ -1 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix}, \tag{69}$$

$$\sigma_3=\begin{pmatrix} 0 & -1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \end{pmatrix}.$$

The introduction of the basis (69) in Eq. (61) gives for the matrix vector $A=(A_w,A_{\overline{w}})$

$$A_w=\frac{1}{V^2}\begin{pmatrix} 0 & -(\overline{\gamma}k\overline{w}^2+\gamma) & (-ai\overline{w}) & 0 & \cdot & \cdot & 0 \\ (\overline{\gamma}k\overline{w}^2+\gamma) & 0 & -(\gamma k\overline{w}^2+\overline{\gamma}) & 0 & \cdot & \cdot & 0 \\ (ai\overline{w}) & (\gamma k\overline{w}^2+\overline{\gamma}) & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix},$$

$$(70)$$

$$A_{\overline{w}}=\frac{1}{V^2}\begin{pmatrix} 0 & -(\gamma kw^2+\overline{\gamma}) & (aiw) & 0 & \cdot & \cdot & 0 \\ (\gamma kw^2+\overline{\gamma}) & 0 & -(\overline{\gamma}kw^2+\gamma) & 0 & \cdot & \cdot & 0 \\ (-aiw) & (\overline{\gamma}kw^2+\gamma) & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix}.$$

Again, we must integrate Eq. (65) in order to obtain the elements of the group. We obtain the $2N^2$-differential equations

$$g_{1i,w}=\frac{1}{V^2}[-(\overline{\gamma}k\overline{w}^2+\gamma)g_{2i}-(ai\overline{w})g_{3i}],$$

$$g_{2i,w}=\frac{1}{V^2}[(\overline{\gamma}k\overline{w}^2+\gamma)g_{1i}-(\gamma k\overline{w}^2+\overline{\gamma})g_{3i}],$$

$$g_{3i,w}=\frac{1}{V^2}[(ai\overline{w})g_{1i}+(\gamma k\overline{w}^2+\overline{\gamma})g_{2i}],$$

$$g_{ji,w}=0,$$

$$g_{1i,\overline{w}}=\frac{1}{V^2}[-(\gamma kw^2+\overline{\gamma})g_{2i}+(aiw)g_{3i}],$$

$$g_{2i,\overline{w}}=\frac{1}{V^2}[(\gamma kw^2+\overline{\gamma})g_{1i}-(\overline{\gamma}kw^2+\gamma)g_{3i}],$$

$$g_{3i,\overline{w}}=\frac{1}{V^2}[-(aiw)g_{1i}+(\overline{\gamma}kw^2+\gamma)g_{2i}],$$

$$g_{ji,\overline{w}}=0,$$

where $i=1,2,...,N$, and $j=4,...,N$.

This system has a solution given by

$$g_{1i}(w,\overline{w})=\Bigg[d_{0i}-d_{0i}w\overline{w}-f_{0i}\gamma w-\overline{\gamma}f_{0i}\overline{w}+(e_{0i}-id_{0i})w^2+(e_{0i}+id_{0i})\overline{w}^2+\frac{\gamma f_{0i}}{2}w^2\overline{w}$$
$$+\frac{\overline{\gamma}f_{0i}}{2}\overline{w}^2w+\frac{d_{0i}}{4}\overline{w}^2w^2\Bigg]\Bigg/(1+kw\overline{w})^2,$$

$$g_{2i}(w,\overline{w})=\Bigg[f_{0i}-3f_{0i}w\overline{w}-(e_{0i}-id_{0i})\overline{\gamma}w-(e_{0i}+id_{0i})\gamma\overline{w}+\frac{\overline{\gamma}}{2}(e_{0i}-id_{0i})\overline{w}w^2$$
$$+\frac{\gamma}{2}(e_{0i}+id_{0i})w\overline{w}^2+\frac{f_{0i}}{4}\overline{w}^2w^2\Bigg]\Bigg/(1+kw\overline{w})^2,$$
$$(71)$$
$$g_{3i}(w,\overline{w})=\Bigg[e_{0i}-e_{0i}w\overline{w}+\overline{\gamma}f_{0i}w+\gamma f_{0i}\overline{w}+i(e_{0i}-id_{0i})w^2-i(e_{0i}+id_{0i})\overline{w}^2-\frac{\overline{\gamma}}{2}f_{0i}\overline{w}w^2$$
$$-\frac{\gamma}{2}f_{0i}w\overline{w}^2+\frac{e_{0i}}{4}\overline{w}^2w^2\Bigg]\Bigg/(1+kw\overline{w})^2,$$

$$g_{ji}(w,\overline{w})=c_{ji},\quad c_{ji}\in\Re$$

for $i=1,2,...,N$, and $j=4,...,N$.

The solution (71) must fulfill the constraints $g^{-1}=g^T$, $\det g=1$, which implies that the constants d_{0i}, f_{0i}, c_{ji}, must satisfy

$$\begin{pmatrix} d_{01} & d_{02} & \cdot & \cdot & \cdot & d_{0N}\\ f_{01} & f_{02} & \cdot & \cdot & \cdot & f_{0N}\\ e_{01} & e_{02} & \cdot & \cdot & \cdot & e_{0N}\\ c_{41} & c_{42} & \cdot & \cdot & \cdot & c_{4N}\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ c_{N1} & c_{N2} & \cdot & \cdot & \cdot & c_{NN}\end{pmatrix}\in SO(N). \qquad (72)$$

The complex parameters w, $\overline{w}$ must satisfy the harmonic equation in M_2 given by

$$(\alpha w_{,z})_{,\overline{z}}+(\alpha w_{,\overline{z}})_{,z}-\frac{4\alpha k\overline{w}}{(1+kw\overline{w})}w_{,z}w_{,\overline{z}}=0,$$
$$(73)$$
$$(\alpha\overline{w}_{,z})_{,\overline{z}}+(\alpha\overline{w}_{,\overline{z}})_{,z}-\frac{4\alpha kw}{(1+kw\overline{w})}\overline{w}_{,z}\overline{w}_{,\overline{z}}=0.$$

For each solution of Eq. (73), we get a new solution, g, of the chiral equations given by (67) for SU(N), and by (71) for SO(N).

V. CONCLUSIONS

We have found three classes of exact solutions of the SU(N)- and SO(N)-chiral equations (3). The first class depends on a harmonic map which corresponds to a one-dimensional Riemannian space. Any one-dimensional space is flat, and its harmonic functions are solutions of the Laplace equation (36). For the two-dimensional Riemannian spaces we found two classes depending on two harmonic maps, which are the local coordinates of the two-dimensional symmetric Riemannian space M_2. All the two-dimensional Riemannian spaces are conformally flat, and all the two-dimensional symmetric spaces have constant curvature. The first two-dimensional class has zero curvature and it is flat, while the second class has nonzero curvature. This last class corresponds to the three-dimensional subalgebras of su(N) and so(N), they can be put in terms of the su(2) and so(3) respective vector space basis. For the flat two-dimensional class, the harmonic maps correspond to a system of two decoupled Laplace equations, while for the nonflat class, the harmonic maps are minimal surfaces on a two-dimensional Riemannian space with positive curvature. The corresponding group elements can be found by the exponential map or by integration. We have integrated all the group elements to obtain explicit expressions of the chiral field in terms of the harmonic maps.

ACKNOWLEDGMENT

This work was partially supported by CONACYT-México.

[1] L. L. Chau (unpublished).
[2] E. Brezin, C. Itzykson, J. Zinn-Justin, and B. J. Zuber, Phys. Lett. **82B**, 442 (1979).
[3] V. E. Zakharov and A. V. Mikhailov, Sov. Phys. JETP **47**, 1017 (1978).
[4] D. Kramer, G. Neugebauer, and T. Matos, J. Math. Phys. **32**, 2727 (1991).
[5] N. Sanchez, in *Proceedings of the Scheveningen (1984) International Conference on Geometric Aspects of the Einstein Equations and Integrable Systems*, edited by R. Martini, Lecture Notes in Physics Vol. 239 (Springer, Berlin, 1984).
[6] T. Matos, J. Math. Phys. **35**, 1302 (1994).
[7] V. Husain, Phys. Rev. Lett. **72**, 800 (1994).
[8] H. Garcia-Compean and T. Matos, Phys. Rev. D **52**, 4425 (1995).
[9] K. Uhlenbeck, J. Diff. Geom. **30**, 1 (1989).
[10] T. Matos, Math. Notes **58**, 1178 (1995).
[11] M. Nakahara, *Geometry, Topology and Physics,* Graduate Student Series In Physics (I.P, Bristol, 1992).

q-Trinomial identities

S. Ole Warnaar[a)]
Instituut voor Theoretische Fysica, Universiteit van Amsterdam, Valckenierstraat 65, 1018 XE Amsterdam, The Netherlands

(Received 10 November 1998; accepted for publication 8 December 1998)

We obtain connection coefficients between q-binomial and q-trinomial coefficients. Using these, one can transform q-binomial identities into q-trinomial identities and back again. To demonstrate the usefulness of this procedure we rederive some known trinomial identities related to partition theory and prove many of the conjectures of Berkovich, McCoy and Pearce, which have recently arisen in their study of the $\phi_{2,1}$ and $\phi_{1,5}$ perturbations of minimal conformal field theory. © *1999 American Institute of Physics.* [S0022-2488(99)01105-6]

I. INTRODUCTION

The q-binomial coefficients can be defined by the q-analog of Newton's binomial expansion,

$$(1+x)(1+qx)\cdots(1+q^{n-1}x)=\sum_{a=0}^{n} x^a q^{a(a-1)/2}\begin{bmatrix} n \\ a \end{bmatrix}. \tag{1}$$

An explicit expression for the q-binomial coefficients is given by

$$\begin{bmatrix} n \\ a \end{bmatrix}_q=\begin{bmatrix} n \\ a \end{bmatrix}=\begin{cases} \dfrac{(q)_n}{(q)_a(q)_{n-a}} & \text{for } 0\leqslant a\leqslant n, \\ 0, & \text{otherwise,} \end{cases}$$

where

$$(q)_n=\prod_{j=1}^{n}(1-q^j), \quad n\geqslant 1 \quad \text{and} \quad (q)_0=1.$$

q-Binomials play an essential role in combinatorics, partition theory, and statistical mechanics; see, e.g., Refs. 1–4, and one of MacMahon's famous results is that $\begin{bmatrix} n+m \\ m \end{bmatrix}$ is the generating function of partitions with no more than m parts, no part exceeding n. Less well understood are the q-trinomial coefficients, defined as q-analogs of the numbers appearing in the generalized Pascal triangle

$$\begin{array}{ccccccccc} & & & & 1 & & & & \\ & & & 1 & 1 & 1 & & & \\ & & 1 & 2 & 3 & 2 & 1 & & \\ & 1 & 3 & 6 & 7 & 6 & 3 & 1 & \\ . & . & . & . & . & . & . & . & . \end{array}. \tag{2}$$

Andrews, Baxter, and Forrester[5,6] were the first to encounter q-trinomial coefficients, and in Ref. 6 Andrews and Baxter defined

[a)]Electronic mail: warnaar@wins.uva.nl

$$\begin{bmatrix} L,b;q \\ a \end{bmatrix}_2 = \begin{bmatrix} L,b \\ a \end{bmatrix}_2 = \sum_{k\geqslant 0} q^{k(k+b)} \begin{bmatrix} L \\ k \end{bmatrix} \begin{bmatrix} L-k \\ k+a \end{bmatrix} \tag{3}$$

and

$$T_n(L,a;q) = T_n(L,a) = q^{(L-a)(L+a-n)/2} \begin{bmatrix} L,a-n;q^{-1} \\ a \end{bmatrix}_2 . \tag{4}$$

The q-trinomial T_n can be expressed explicitly as

$$T_n(L,a) = \sum_{\substack{r=0 \\ L-a-r \text{ even}}}^{L-|a|} \frac{q^{r(r-n)/2}(q)_L}{(q)_{(L-a-r)/2}(q)_{(L+a-r)/2}(q)_r} . \tag{5}$$

Clearly, the q-trinomial coefficients are nonzero for $a=-L,-L+1,...,L$ only and satisfy the symmetries

$$\begin{bmatrix} L,b;q \\ a \end{bmatrix}_2 = q^{a(a-b)} \begin{bmatrix} L,b-2a \\ -a \end{bmatrix}_2 \quad \text{and} \quad T_n(L,a) = T_n(L,-a).$$

To see that (3) indeed defines q-analogs of the trinomial coefficients, set $q=1$ and twice apply the binomial formula to find that

$$\sum_{a=-L}^{L} x^a \begin{bmatrix} L,b;1 \\ a \end{bmatrix}_2 = (1+x+x^{-1})^L,$$

in accordance with (2). The only further properties of q-trinomials needed in this paper are the limiting formulas[6]

$$\lim_{\substack{L\to\infty \\ L-a \text{ even}}} T_0(L,a) = \frac{(-q^{1/2})_\infty + (q^{1/2})_\infty}{2(q)_\infty}, \tag{6}$$

$$\lim_{\substack{L\to\infty \\ L-a \text{ odd}}} T_0(L,a) = \frac{(-q^{1/2})_\infty - (q^{1/2})_\infty}{2(q)_\infty}, \tag{7}$$

and

$$\lim_{L\to\infty} \begin{bmatrix} L,a \\ a \end{bmatrix}_2 = \frac{1}{(q)_\infty} . \tag{8}$$

Finally, we introduce the abbreviation

$$\begin{bmatrix} L,a \\ a \end{bmatrix}_2 = \begin{bmatrix} L \\ a \end{bmatrix}_2 .$$

Since their discovery about a decade ago, q-trinomials have found numerous applications in, again, combinatorics, partition theory, and statistical mechanics.[5–23] Among the most striking results is a q-trinomial proof of Schur's partition theorem and Capparelli's (then) conjecture,[9] a q-trinomial proof of the Göllnitz–Gordon partition theorem[7] and their Andrews–Bressoud generalizations,[13,16] the proof of an E_8 Rogers–Ramanujan-type identity,[10] and a trinomial analog of Bailey's lemma.[19]

Most of the above-cited papers contain q-trinomial identities. Upon close inspection of many of these identities, one is struck by their similarity with well-known q-binomial identities. This strongly suggests that many q-trinomial identities can be simply viewed as corollaries of q-binomial identities. In an earlier paper[23] we made a first, only partially successful, attempt to relate q-trinomial identities to q-binomial identities, showing that each Bailey pair (which implies a q-binomial identity) implies a trinomial Bailey pair (which implies a q-trinomial identity). The problem with the idea of Ref. 23 is that it applies to q-trinomial identities in which the parameter a in (3) and (4) takes even values only. Therefore, q-trinomial identities in which a takes arbitrary integer values remained irreducible to q-binomial identities.

In this paper we intend to deal with this problem, and in the next section connection coefficients between q-binomial and q-trinomial coefficients are obtained. Using these coefficients and the idea of Ref. 23, many q-trinomial identities are derived from known q-binomial identities. In Sect. III, several q-trinomial identities related to partitions are obtained and in Sec. IV general classes of q-trinomial identities are proved, including many of the recent conjectures of Berkovich, McCoy, and Pearce.[21] To make contact with the recently discovered trinomial analog of Bailey's lemma, our results are finally formulated in the language of Bailey pairs in Sec. V. In the Appendix some necessary formulas for q-binomial coefficients are collected.

II. CONNECTION COEFFICIENTS

To relate q-binomials and q-trinomials, we consider the simple problem of finding the coefficients $C_{L,k}$ and $C'_{L,k}$, such that

$$T_0(L,a)=\sum_{k=0}^{L} C_{L,k}(a)\begin{bmatrix}2k\\ k-a\end{bmatrix} \tag{9}$$

and

$$\begin{bmatrix}2L\\ L-a\end{bmatrix}=\sum_{k=0}^{L} C'_{L,k}(a)T_0(k,a). \tag{10}$$

Of course, the two equations imply that

$$\sum_{k=M}^{L} C_{L,k}(a)C'_{k,M}(a)=\delta_{L,M}. \tag{11}$$

The answer to the above connection coefficient problem is given by the following lemma.

Lemma II.1: For $C_{L,k}$ and $C'_{L,k}$ as above,

$$C_{L,k}(a)=(-1)^{L-k}q^{\binom{L-k}{2}+(a^2-L^2)/2}\begin{bmatrix}L\\ k\end{bmatrix}, \tag{12}$$

$$C'_{L,k}(a)=q^{(k^2-a^2)/2}\begin{bmatrix}L\\ k\end{bmatrix}. \tag{13}$$

Proof: Substitution of the expression for $C'_{L,k}$ into the right-hand side of (10) and using Eq. (5) for T_0 gives

$$\sum_{k=0}^{L} C'_{L,k}(a)T_0(k,a)=\sum_{k=0}^{L}\sum_{\substack{r=0\\ k-a-r\ \text{even}}}^{k-|a|}\frac{q^{(k^2-a^2+r^2)/2}(q)_L}{(q)_{L-k}(q)_{(k-a-r)/2}(q)_{(k+a-r)/2}(q)_r}.$$

To proceed, we introduce new summation variables i,j defined by $k=i+j+a$ and $r=i-j$, and apply the q-Chu–Vandermonde sum, i.e.,

$$\sum_{k=0}^{L} C'_{L,k}(a)T_0(k,a)=\sum_{i=0}^{L}\sum_{j=0}^{i} q^{i(i+a)+j(j+a)}\begin{bmatrix}L\\i\end{bmatrix}\begin{bmatrix}i\\j\end{bmatrix}\begin{bmatrix}L-i\\j+a\end{bmatrix} \overset{\text{by (A1)}}{=} \sum_{i=0}^{L} q^{i(i+a)}\begin{bmatrix}L\\i\end{bmatrix}\begin{bmatrix}L\\i+a\end{bmatrix}$$

$$\overset{\text{by (A1)}}{=} \begin{bmatrix}2L\\L-a\end{bmatrix}.$$

This settles (13), and to prove (12) we show that (11) holds. Taking the left-hand side of (11) and substituting the claim of the lemma, we find

$$\sum_{k=M}^{L} C_{L,k}(a)C'_{k,M}(a)=\sum_{k=M}^{L}(-1)^{L-k}q^{\binom{L-k}{2}+(M^2-L^2)/2}\begin{bmatrix}L\\k\end{bmatrix}\begin{bmatrix}k\\M\end{bmatrix}$$

$$=q^{(M^2-L^2)/2}\begin{bmatrix}L\\M\end{bmatrix}\sum_{k=0}^{L-M}(-1)^k q^{\binom{k}{2}}\begin{bmatrix}L-M\\k\end{bmatrix}=\delta_{L,M},$$

where in the last step we have used (1) with $x=-1$. □

We note that a proof of (12) that does not rely on (13) is implied by Eqs. (2.12) and (2.35) of Ref. 6.

The analogous result involving T_1 instead of T_0 can be stated as follows. Define $D_{L,k}$ and $D'_{L,k}$ by

$$T_1(L,a)=\sum_{k=0}^{L} D_{L,k}(a)\begin{bmatrix}2k\\k-a\end{bmatrix}$$

and

$$\begin{bmatrix}2L\\L-a\end{bmatrix}=\sum_{k=0}^{L} D'_{L,k}(a)T_1(k,a). \tag{14}$$

Lemma II.2: For $D_{L,k}$ and $D'_{L,k}$ as above,

$$D_{L,k}(a)=(-1)^{L-k}q^{\binom{L-k}{2}+\binom{a}{2}-\binom{L}{2}}\frac{1+q^a}{1+q^k}\begin{bmatrix}L\\k\end{bmatrix}, \tag{15}$$

$$D'_{L,k}(a)=q^{\binom{k}{2}-\binom{a}{2}}\frac{1+q^L}{1+q^a}\begin{bmatrix}L\\k\end{bmatrix}. \tag{16}$$

Proof: Following the proof of Lemma II.1 with T_0 replaced by T_1, one finds after application of the q-Chu–Vandermonde sum (A1), that the right-hand side of (14) is equal to

$$\frac{1+q^L}{1+q^a}\sum_{i=0}^{L} q^{i(i+a-1)}\begin{bmatrix}L\\i\end{bmatrix}\begin{bmatrix}L\\i+a\end{bmatrix}.$$

Before (A1) can again be applied, the recurrence (A5) is needed to rewrite this as

$$\frac{1+q^L}{1+q^a}\left\{\sum_{i=0}^{L} q^{i(i+a-1)}\begin{bmatrix}L\\i\end{bmatrix}\begin{bmatrix}L-1\\i+a-1\end{bmatrix}+q^a\sum_{i=0}^{L} q^{i(i+a)}\begin{bmatrix}L\\i\end{bmatrix}\begin{bmatrix}L-1\\i+a\end{bmatrix}\right\}.$$

Using (A1) and combining terms gives $\left[\begin{smallmatrix}2L\\L-a\end{smallmatrix}\right]$. To prove (15) it again suffices to consider $\Sigma_{k=M}^{L} D_{L,k}(a)D'_{k,M}(a)$. After substituting the results for D and D' and replacing $k\to L-k$, one finds that this becomes $\delta_{L,M}$ after using (1) with $x=-1$. □

To conclude this section, we note that the representations (3) and (5) for the q-trinomial coefficients can also be written as a relation between q-trinomials and q-binomials. That is,

$$T_n(L,2a)=\sum_{k\geqslant 0} q^{(L-2k)(L-2k-n)/2}\begin{bmatrix} L \\ 2k \end{bmatrix}\begin{bmatrix} 2k \\ k-a \end{bmatrix} \tag{17}$$

and

$$\begin{bmatrix} L,b \\ 2a \end{bmatrix}_2=\sum_{k\geqslant 0} q^{(k-a)(k-a+b)}\begin{bmatrix} L \\ 2k \end{bmatrix}\begin{bmatrix} 2k \\ k-a \end{bmatrix}. \tag{18}$$

These results, which, unlike the previous transformations are not invertible, will be needed later.

III. SIMPLE EXAMPLES FROM PARTITION THEORY

Before proving general series of q-trinomial identities using the results of the previous section, we treat some simple examples related to partition identities first.

The first example concerns the following result of Andrews[7] (see also Ref. 21). It is well known[24] that the (first) Rogers–Ramanujan identity can be obtained as a limiting case of the polynomial identity,

$$\sum_{n\geqslant 0} q^{n^2}\begin{bmatrix} L-n \\ n \end{bmatrix}=\sum_{j=-\infty}^{\infty} (-1)^j q^{j(5j+1)/2}\begin{bmatrix} L \\ (L-5j)/2 \end{bmatrix}. \tag{19}$$

Here the polynomials appearing on either side are known to be the generating function of partitions with the difference between parts of at least two and the largest part not exceeding $L-1$.[4,25] In Ref. 7, Andrews remarks that it is "most surprising and intriguing" that the following also holds:

$$\sum_{n\geqslant 0} q^{n^2}\begin{bmatrix} L-n \\ n \end{bmatrix}=\sum_{j=-\infty}^{\infty}\left\{ q^{j(10j+1)}\begin{bmatrix} L \\ 5j \end{bmatrix}_2 - q^{(2j+1)(5j+2)}\begin{bmatrix} L \\ 5j+2 \end{bmatrix}_2\right\}. \tag{20}$$

We now show that (20) is a corollary of (19), or for those who prefer to decrease instead of increase complexity, that (19) is a corollary of (20). Replacing $q\to 1/q$ in (9) and (12), using (4) and (A7), we find that [see also Ref. 6, Eqs. (2.12) and (2.35)]

$$\begin{bmatrix} L \\ a \end{bmatrix}_2=\sum_{k=0}^{L} (-1)^{L-k} q^{(L-k)(L+k+1)/2}\begin{bmatrix} L \\ k \end{bmatrix}\begin{bmatrix} 2k \\ k-a \end{bmatrix}.$$

If we thus take (19) with L replaced by $2k$, multiply by $(-1)^{L-k}q^{(L-k)(L+k+1)/2}\left[\begin{smallmatrix} L \\ k \end{smallmatrix}\right]$ and sum over k, we arrive at

$$\sum_{k\geqslant 0}\sum_{n\geqslant 0} (-1)^{L-k} q^{(L-k)(L+k+1)/2+n^2}\begin{bmatrix} L \\ k \end{bmatrix}\begin{bmatrix} 2k-n \\ n \end{bmatrix}$$
$$=\sum_{j=-\infty}^{\infty}\left\{ q^{j(10j+1)}\begin{bmatrix} L \\ 5j \end{bmatrix}_2 - q^{(2j+1)(5j+1)}\begin{bmatrix} L \\ 5j+2 \end{bmatrix}_2\right\}.$$

To simplify the left-hand side, we set $k=L-m+n$ followed by $n\to m-n$ to get

$$\sum_{m\geqslant 0} q^{m^2}\sum_{n\geqslant 0} (-1)^n q^{\binom{n}{2}+n(L-2m+1)}\begin{bmatrix} L \\ n \end{bmatrix}\begin{bmatrix} 2L-m-n \\ m-n \end{bmatrix}=\sum_{m\geqslant 0} q^{m^2}\begin{bmatrix} L-m \\ m \end{bmatrix},$$

where the sum over n has been performed using the q-Chu–Vandermonde summation (A3). As remarked before, one can equally well take the reverse route and starting from (20), using Lemma II.1, one readily obtains (19). We leave this to the reader.

Our second example concerns the following identity of Slater[26] related to the (first) Göllnitz–Gordon partition identity:[27,28]

$$\sum_{n=0}^{\infty} \frac{q^{n^2}(-q;q^2)_n}{(q^2;q^2)_n} = \prod_{n=0}^{\infty} \frac{1}{(1-q^{8j+1})(1-q^{8j+4})(1-q^{8j+7})}. \tag{21}$$

A polynomial identity that implies this equation is given by[13,16]

$$\sum_{m,n\geqslant 0} q^{(m^2+n^2)/2} \begin{bmatrix} L-m \\ n \end{bmatrix} \begin{bmatrix} n \\ m \end{bmatrix} = \sum_{j=-\infty}^{\infty} (-1)^j q^{2j^2+j/2} \{T_0(L,4j)+T_0(L,4j+1)\}. \tag{22}$$

It was observed in Ref. 7 that for fixed L the polynomial appearing on the right-hand side with q replaced by q^2 is the generating function of partitions $\lambda=(\lambda_1,\lambda_2,\ldots)$ with $\lambda_i-\lambda_{i+1}\geqslant 2$ for λ_i odd, $\lambda_i-\lambda_{i+1}\geqslant 3$ for λ_i even, and with the largest part not exceeding $2L-1$. To see that (22) indeed implies (21), let L tend to infinity using (6), (7), and (A6). Hence,

$$\sum_{m,n\geqslant 0} \frac{q^{(m^2+n^2)/2}}{(q)_n} \begin{bmatrix} n \\ m \end{bmatrix} = \frac{(-q^{1/2})_\infty}{(q)_\infty} \sum_{j=-\infty}^{\infty} (-1)^j q^{2j^2+j/2}.$$

Using Jacobi's triple product identity [Eq. (2.2.10) of Ref. 1] and Eq. (1) with $x=q^{1/2}$ gives

$$\sum_{n\geqslant 0} \frac{q^{n^2/2}(-q^{1/2})_n}{(q)_n} = \frac{(-q^{1/2})_\infty (q^{3/2};q^4)_\infty (q^{5/2};q^4)_\infty (q^4;q^4)_\infty}{(q)_\infty}.$$

Letting $q\to q^2$ and cleaning up the right-hand side finally yields (21).

The companion q-binomial identity of (22) is given by the following identity of Refs. 29 and 30:

$$\sum_{\substack{m_1,m_2\geqslant 0\\ m_1+m_2 \text{ even}}} q^{(m_1^2+m_2^2)/4} \begin{bmatrix} L+\frac{1}{2}(m_1-m_2) \\ m_1 \end{bmatrix} \begin{bmatrix} \frac{1}{2}(m_1+m_2) \\ m_2 \end{bmatrix} = \sum_{j=-\infty}^{\infty} (-1)^j \left\{ q^{j(20j+1)/2} \begin{bmatrix} 2L \\ L-4j \end{bmatrix} + q^{(4j+1)(5j+1)/2} \begin{bmatrix} 2L \\ L-4j-1 \end{bmatrix} \right\}.$$

To prove this we replace L by k, multiply by $q^{-a^2/2}C_{L,k}(a)$ as given by (12), and sum over k using (9). The resulting equation is

$$\begin{aligned} &\sum_{j=-\infty}^{\infty} (-1)^j q^{2j^2+j/2} \{T_0(L,4j)+T_0(L,4j+1)\} \\ &= \sum_{\substack{m_1,m_2\geqslant 0\\ m_1+m_2 \text{ even}}} q^{(m_1^2+m_2^2-2L^2)/4} \begin{bmatrix} \frac{1}{2}(m_1+m_2) \\ m_2 \end{bmatrix} \sum_{k=0}^{L} (-1)^k q^{\binom{k}{2}} \begin{bmatrix} L \\ k \end{bmatrix} \begin{bmatrix} L-k+\frac{1}{2}(m_1-m_2) \\ m_1 \end{bmatrix} \\ &\overset{\text{by (A2)}}{=} \sum_{\substack{m_1,m_2\geqslant 0\\ m_1+m_2 \text{ even}}} q^{((m_1-L)^2+(m_2-L)^2)/4} \begin{bmatrix} \frac{1}{2}(m_1+m_2) \\ m_2 \end{bmatrix} \begin{bmatrix} \frac{1}{2}(m_1-m_2) \\ m_1-L \end{bmatrix}. \end{aligned}$$

Making the variable change $m_1\to L+n-m$ and $m_2\to L-n-m$, we find Eq. (22).

IV. *q*-TRINOMIAL IDENTITIES

After the previous examples, we now derive general classes of q-trinomial identities, as stated in Propositions IV.1–IV.5 below. The setup will be as follows. First we describe a family of q-binomial identities for bounded analogs of Virasoro characters, based on continued fraction expansions. We then transform these identities into q-trinomial identities, by either using (9) or (18). Many of the q-trinomial identities available in the literature are contained in Propositions IV.1–IV.5 or can be derived in a completely analogous fashion.

A. *q*-binomial identities for bounded Virasoro characters

Using the inclusion–exclusion construction of Feigin and Fuchs,[31] the (normalized) characters of the Virasoro algebra of central charge $c=1-6(p'-p)^2/pp'$, with p,p' integers such that $1<p<p'$ and $\gcd(p,p')=1$, are given by[32,33]

$$\chi_{r,s}^{(p,p')}(q)=\frac{1}{(q)_\infty}\sum_{j=-\infty}^{\infty}\{q^{j(pp'j+p'r-ps)}-q^{(pj+r)(p'j+s)}\}.$$

Here $r=1,...,p-1$ and $s=1,...,p'-1$ label the highest weight representations.

For simplicity we only deal with the ''vacuum'' character, determined by $|p'r-ps|=1$. The following polynomial analogs of the vacuum Virasoro characters have arisen in the context of statistical mechanics[34,35] and partition theory,[36]

$$B_L(p,p';q)=\sum_{j=-\infty}^{\infty}\left\{q^{j(pp'j+1)}\begin{bmatrix}2L\\L-p'j\end{bmatrix}-q^{(pj+r)(p'j+s)}\begin{bmatrix}2L\\L-p'j-s\end{bmatrix}\right\}. \tag{23}$$

The polynomials $B_L(p,p')$ are known to be related to the minimal conformal field theory $M(p,p')$ perturbed by the operator $\phi_{1,3}$.

Recently, very different, so-called fermionic representations for the above polynomials have been obtained by Berkovich, McCoy and Schilling using continued fractions.[29,30] Assume $p<p'<2p$, $\gcd(p,p')=1$ and define non-negative integers n and $\nu_0,\ldots,\nu_n$ by the continued fraction expansion

$$\frac{p}{p'-p}=\nu_0+\cfrac{1}{\nu_1+\cfrac{1}{\nu_2+\cdots+\cfrac{1}{\nu_n+2}}}=[\nu_0,\ldots,\nu_{n-1},\nu_n+2].$$

Using n and ν_j, set

$$t_m=\sum_{j=0}^{m-1}\nu_j,\quad 1\le m\le n\quad\text{and}\quad d=\sum_{j=0}^{n}\nu_j. \tag{24}$$

The t_m and d are used to define a fractional incidence matrix $\mathcal{I}$ and a fractional Cartan-type matrix $2B=2I-\mathcal{I}$ (with I the d by d unit matrix) as follows:

$$\mathcal{I}_{i,j}=\begin{cases}\delta_{i,j+1}+\delta_{i,j-1}, & \text{for } 1\le i<d,\quad i\ne t_m,\\ \delta_{i,j+1}+\delta_{i,j}-\delta_{i,j-1}, & \text{for } i=t_m,\quad 1\le m\le n-\delta_{\nu_n,0},\\ \delta_{i,j+1}+\delta_{\nu_n,0}\delta_{i,j}, & \text{for } i=d.\end{cases} \tag{25}$$

When $p'=p+1$, the incidence matrix $\mathcal{I}$ has components $\mathcal{I}_{i,j}=\delta_{|i-j|,1}$ $(i,j=1,...,p-2)$, so that $2B$ corresponds to the Cartan matrix of the Lie algebra A_{p-3}. When $p=2k-1$ and $p'=2k+1$ the matrix $\mathcal{I}$ has components $\mathcal{I}_{i,j}=\delta_{|i-j|,1}+\delta_{i,j}\delta_{i,k-1}$ $(i,j=1,...,k-1)$, so that $2B$ corresponds to the Cartan-type matrix of the tadpole graph of $k-1$ nodes.

Using the above definition, the fermionic representation for the bounded Virasoro characters with $p<p'<2p$ can be given as

$$F_L(p,p';q)=\sum_{m\in 2\mathbb{Z}^d} q^{mBm/2}\prod_{j=1}^{d}\begin{bmatrix} L\,\delta_{j,1}+\frac{1}{2}(\mathcal{I}m)_j \\ m_j\end{bmatrix}. \tag{26}$$

Here we use the notations $vMw=\Sigma_{j,k}v_jM_{j,k}w_k$, $(Mv)_j=\Sigma_k M_{j,k}v_k$ and $(vM)_j=\Sigma_k v_k M_{k,j}$. These conventions are important since, generally, $M(=\mathcal{I},B)$ is not a symmetric matrix. The general form (26) for $F_L(p,p')$ can be found in Refs. 29 and 30 (see also Ref. 37). The important special cases $(p,p')=(p,p+1)$ and $(2k-1,2k+1)$ were proven prior to this is in Refs. 38, 39 and Ref. 40, respectively.

The expression for $F_L(p,p';q)$ with $p'>2p$ follows from the duality transformation

$$F_L(p,p';1/q)=q^{-L^2}F_L(p'-p,p';q). \tag{27}$$

To obtain fermionic character formulas for $\chi_{r,s}^{(p,p')}(q)$ with $|p'r-ps|=1$, one simply lets L tend to infinity in (26).

Before we proceed to use the identity,

$$F_L(p,p';q)=B_L(p,p';q), \tag{28}$$

to derive trinomial identities, let us comment on the convention of writing $2B$ for a Cartan-type matrix in the above formulas. This has its origin in the work of Ref. 41, where, in more general situations, the matrix B has a (nontrivial) tensor product structure, $B=b_1\otimes b_2$. In the identities of this section the matrix b_1 is simply the inverse of the A_1 Cartan matrix, $(b_1)=(\frac{1}{2})$. In Sec. IV D, however, we indeed encounter a different situation, b_1 being the (still trivial) Cartan-type matrix of the tadpole graph with just a single node, so that $b_1=(1)$.

B. *q*-trinomial identities I

We start with the q-binomial identity (28) for $p<p'<2p$, assuming that $d\geqslant 2$. Applying Eq. (9), with $C_{L,k}$ given by (12), we find

$$\begin{aligned}
&\sum_{j=-\infty}^{\infty}\{q^{(p'(2p-p')j+2)j/2}T_0(L,p'j)-q^{((2p-p')j+2r-s)(p'j+s)/2}T_0(L,p'j+s)\}\\
&\quad=\sum_{k=0}^{L}(-1)^{L-k}q^{\binom{L-k}{2}-L^2/2}\begin{bmatrix}L\\k\end{bmatrix}F_k(p,p';q)\\
&\quad=\sum_{m\in 2\mathbb{Z}^d}q^{(mBm-L^2)/2}\left(\prod_{j=2}^{d}\begin{bmatrix}\frac{1}{2}(\mathcal{I}m)_j\\ m_j\end{bmatrix}\right)\sum_{k=0}^{L}(-1)^k q^{\binom{k}{2}}\begin{bmatrix}L\\k\end{bmatrix}\begin{bmatrix}L-k+\frac{1}{2}(\mathcal{I}m)_1\\ m_1\end{bmatrix}\\
&\quad\overset{\text{by (A2)}}{=}q^{L^2/2}\sum_{m\in 2\mathbb{Z}^d}q^{mBm/2-L(Bm)_1}\begin{bmatrix}\frac{1}{2}(\mathcal{I}m)_1\\ m_1-L\end{bmatrix}\prod_{j=2}^{d}\begin{bmatrix}\frac{1}{2}(\mathcal{I}m)_j\\ m_j\end{bmatrix}\\
&\quad=q^{L^2\mathcal{I}_{1,1}/2}\sum_{m+Le_1\in 2\mathbb{Z}^d}q^{mBm/2+L(mB-Bm)_1/2}\prod_{j=1}^{d}\begin{bmatrix}\frac{1}{2}L\mathcal{I}_{j,1}+\frac{1}{2}(\mathcal{I}m)_j\\ m_j\end{bmatrix},
\end{aligned}$$

with e_j $(j=1,...,d)$ the standard unit vectors in $\mathbb{Z}^d$. We now have to distinguish two cases according to whether $\nu_0=1$ (so that $3p/2<p'<2p$) or $\nu_0>1$ (so that $p<p'\leqslant 3p/2$). In the latter case $\mathcal{I}_{1,j}=\mathcal{I}_{j,1}=\delta_{1,j-1}$, and we obtain the following polynomial identities.

Proposition IV.1: For integers p,p′ with $p<p'\leqslant 3p/2$ *and* $\gcd(p,p')=1$, *let integers* $1\leqslant r<p$ *and* $1\leqslant s<p'$ *be fixed by* $|p'r-ps|=1$ *and let* $\mathcal{I}$ *and B be defined by (24) and (25). Then the following polynomial identity holds for* $L\in\mathbb{Z}$:

$$\sum_{m+Le_1\in 2\mathbb{Z}^d} q^{mBm/2}\prod_{j=1}^{d}\begin{bmatrix}\frac{1}{2}L\,\delta_{j,2}+\frac{1}{2}(\mathcal{I}m)_j\\ m_j\end{bmatrix}$$
$$=\sum_{j=-\infty}^{\infty}\{q^{(p'(2p-p')j+2)j/2}T_0(L,p'j)-q^{((2p-p')j+2r-s)(p'j+s)/2}T_0(L,p'j+s)\}.$$

The admissible pairs $(p,p')=(3,4)$ and $(p,p')=(2,3)$ have been neglected in our derivation due to the assumption that $d\geqslant 2$. These two cases can be treated in a similar fashion, and when $(p,p')=(3,4)$ the left-hand side is 1 for L even and 0 for L odd. When $(p,p')=(2,3)$, in which case $F_L(2,3;q)=1$, the left-hand side becomes $\delta_{L,0}$. All of the identities of Proposition IV.1 have been derived before, and for $p'=p+1$ they were first found by Schilling.[42,14] The more general case can be found in Ref. 22.

Next we treat the case $\nu_0=1$. When this occurs $\mathcal{I}_{1,j}=\delta_{j,1}-\delta_{1,j-1}$ and $\mathcal{I}_{j,1}=\delta_{j,1}+\delta_{1,j-1}$, and we obtain the following polynomial identities.

Proposition IV.2: For integers p,p′ with $3p/2<p'<2p$ *and* $\gcd(p,p')=1$ *let integers* $1\leqslant r<p$ *and* $1\leqslant s<p'$ *be fixed by* $|p'r-ps|=1$ *and let* $\mathcal{I}$ *and B defined by (24) and (25). Then the following polynomial identity holds for* $L\in\mathbb{Z}$:

$$\sum_{m+Le_1\in 2\mathbb{Z}^d} q^{L(L-2m_2)/4+mBm/2}\prod_{j=1}^{d}\begin{bmatrix}\frac{1}{2}L(\delta_{j,1}+\delta_{j,2})+\frac{1}{2}(\mathcal{I}m)_j\\ m_j\end{bmatrix}$$
$$=\sum_{j=-\infty}^{\infty}\{q^{(p'(2p-p')j+2)j/2}T_0(L,p'j)-q^{((2p-p')j+2r-s)(p'j+s)/2}T_0(L,p'j+s)\}.$$

The case $(p,p')=(3,5)$ has again escaped a proper derivation, but has, in fact, been treated previously, corresponding to identity (20) with q replaced by $1/q$. Apart from this special case due to Andrews,[7] the identities of Proposition IV.2 have been proved by Berkovich, McCoy, and Orrick[13,16] for $(p,p')=(2\nu+1,4\nu)$ and were conjectured for general p and p' by Berkovich, McCoy, and Pearce [Eq. (8.8) of Ref. 21].

C. *q*-trinomial identities II

Our starting point for deriving q-trinomial identities is again Eq. (28), but this time we rely on (18). This implies that (28) with L replaced by k, multiplied by $q^{k^2}\begin{bmatrix}L\\2k\end{bmatrix}$, and summed over k yields

$$\sum_{k\geqslant 0} q^{k^2}\begin{bmatrix}L\\2k\end{bmatrix}F_k(p,p';q)=\sum_{j=-\infty}^{\infty}\left\{q^{j(p'(p+p')j+1)}\begin{bmatrix}L\\2p'j\end{bmatrix}_2-q^{(p'j+s)((p+p')j+r+s)}\begin{bmatrix}L\\2p'j+2s\end{bmatrix}_2\right\}. \tag{29}$$

To transform this into explicit polynomial identities we need to distinguish between $p<p'<2p$ and $p'>2p$.

First, assume that $p<p'<2p$. After substituting expression (26) for F_L, the left side of (29) is

$$\sum_{k\geqslant 0}\sum_{m\in 2\mathbb{Z}^d} q^{k^2+mBm/4}\begin{bmatrix}L\\2k\end{bmatrix}\prod_{j=1}^{d}\begin{bmatrix}k\,\delta_{j,1}+\frac{1}{2}(\mathcal{I}m)_j\\ m_j\end{bmatrix}.$$

By the q-Chu–Vandermonde summation (A1), with $L\to L-k+m_1/2$, $a\to k-m_1/2$, and $b\to -k-m_1/2$, this can be rewritten as

$$\sum_{i,k\geqslant 0}\sum_{m\in 2\mathbb{Z}^d} q^{i(i-k-m_1/2)+k^2+mBm/4}\begin{bmatrix} L-k+\frac{1}{2}m_1 \\ i\end{bmatrix}\begin{bmatrix} k-\frac{1}{2}m_1 \\ 2k-i\end{bmatrix}\prod_{j=1}^{d}\begin{bmatrix} k\delta_{j,1}+\frac{1}{2}(\mathcal{I}m)_j \\ m_j\end{bmatrix}.$$

Replacing $m_j\to m_{j+2}$, $j=1,...,d$, followed by $k\to(m_1+m_2)/2$ and $i\to m_1$ yields

$$\sum{}' q^{(3m_1^2+m_2^2-2m_1m_3)/4+\Sigma_{j,k=1}^{d}m_{j+2}B_{j,k}m_{k+2}/4}$$
$$\times\begin{bmatrix} L-\frac{1}{2}(m_1+m_2-m_3) \\ m_1\end{bmatrix}\begin{bmatrix} \frac{1}{2}(m_1+m_2-m_3) \\ m_2\end{bmatrix}\prod_{j=1}^{d}\begin{bmatrix} \frac{1}{2}(m_1+m_2)\delta_{j,1}+\frac{1}{2}\Sigma_{k=1}^{d}\mathcal{I}_{j,k}m_{k+2} \\ m_j+2\end{bmatrix},$$

where the primed sum denotes a sum over $m\in\mathbb{Z}^{d+2}$ such that m_1+m_2 and $m_3,\ldots,m_{d+2}$ are all even.

Now define a new incidence matrix $\mathcal{I}'$ and Cartan-type matrix $2B'=2I-\mathcal{I}'$ of dimension $d'=d+1$ by replacing the continued fraction expansion $[\nu_0,\ldots\nu_n+2]$ by $[1,\nu_0,\ldots,\nu_n+2]$, so that $\mathcal{I}'$ becomes the incidence matrix corresponding to the continued fraction expansion of p'/p. Also define $\mathcal{I}''$ and $2B''=2I-\mathcal{I}''$ of dimension $d''=d+2$ as

$$\mathcal{I}''_{i,j}=\begin{cases} -\delta_{i,1}\delta_{j,1}+\delta_{i,2}+\delta_{i,3}-\delta_{j,2}+\delta_{j,3}, & \text{for } i=1 \text{ or } j=1,\\ \mathcal{I}'_{i-1,j-1}, & \text{for } i,j=2,...,d+2.\end{cases} \tag{30}$$

Then the above sequence of transformations implies the following proposition.

Proposition IV.3: For integers p,p' with $p<p'<2$ and $\gcd(p,p')=1$ let integers $1\leqslant r<p$ and $1\leqslant s<p'$ be fixed by $|p'r-ps|=1$ and let $\mathcal{I}''$ and B'' be defined by (30). Then the following polynomial identity holds for $L\in\mathbb{Z}$:

$$\sum{}' q^{mB''m/4}\prod_{j=1}^{d''}\begin{bmatrix} L\delta_{j,1}+\frac{1}{2}(\mathcal{I}''m)_j \\ m_j\end{bmatrix}$$
$$=\sum_{j=-\infty}^{\infty}\left\{q^{j(p'(p+p')j+1)}\begin{bmatrix} L \\ 2p'j\end{bmatrix}_2-q^{(p'j+s)((p+p')j+r+s)}\begin{bmatrix} L \\ 2p'j+2s\end{bmatrix}_2\right\}.$$

The identities of Proposition IV.3 are the $n=0$ case of the conjectured equation (8.11) [which contains the $n=0$ instances of (6.19) and (8.3)] of Ref. 21, and are related to the $\phi_{2,1}$ perturbation of the minimal conformal field theory $M(p',p+p')$.

When $p'>2p$ we replace $p\to p'-p$ in (29) and use the duality property (27). Hence

$$\sum_{k\geqslant 0} q^{2k^2}\begin{bmatrix} L \\ 2k\end{bmatrix}F_k(p,p';q^{-1})$$
$$=\sum_{j=-\infty}^{\infty}\left\{q^{j(p'(2p'-p)j+1)}\begin{bmatrix} L \\ 2p'j\end{bmatrix}_2-q^{(p'j+s)((2p'-p)j+r+s)}\begin{bmatrix} L \\ 2p'j+2s\end{bmatrix}_2\right\}. \tag{31}$$

Observe that the transformation carried out above implies $p<p'<2p$ and $|p'(r-s)+ps|=1$.

Substituting expression (26) for F_L and using (A7), the left side of (31) yields

$$\sum_{k\geqslant 0}\sum_{m\in 2\mathbb{Z}^d} q^{k(2k-m_1)+mBm/4}\begin{bmatrix} L \\ 2k\end{bmatrix}\prod_{j=1}^{d}\begin{bmatrix} k\delta_{j,1}+\frac{1}{2}(\mathcal{I}m)_j \\ m_j\end{bmatrix}.$$

By the q-Chu–Vandermonde summation (A1), with $L\to L-m_1/2$, $a\to m_1/2$, $b\to m_1/2-2k$, this can be rewritten as

$$\sum_{i,k\geqslant 0}\sum_{m\in 2\mathbb{Z}^d} q^{i(i-2k+m_1/2)+k(2k-m_1)+mBm/4}\begin{bmatrix} L-\frac{1}{2}m_1 \\ i\end{bmatrix}\begin{bmatrix}\frac{1}{2}m_1\\ 2k-1\end{bmatrix}\prod_{j=1}^{d}\begin{bmatrix}k\delta_{j,1}+\frac{1}{2}(\mathcal{I}m)_j\\ m_j\end{bmatrix}.$$

Replacing $m_j\to m_{j+2}$, $j=1,...,d$, followed by $t\to m_1+m_2$ and $i\to m_1$, gives

$$\sum{}' q^{(m_1^2+m_2^2-m_2m_3)/2+\Sigma_{j,k=1}^{d}m_{j+2}B_{j,k}m_{k+2}/4}\begin{bmatrix}L-\frac{1}{2}m_3\\ m_1\end{bmatrix}$$
$$\times\begin{bmatrix}\frac{1}{2}m_3\\ m_2\end{bmatrix}\prod_{j=1}^{d}\begin{bmatrix}\frac{1}{2}(m_1+m_2)\delta_{j,1}+\frac{1}{2}\Sigma_{k=1}^{d}\mathcal{I}_{j,k}m_{k+2}\\ m_{j+2}\end{bmatrix},$$

where the primed sum again denotes a sum over $m\in\mathbb{Z}^{d+2}$ such that m_1+m_2 and $m_3,\ldots,m_{d+2}$ are all even.

Now define a new incidence matrix $\mathcal{I}$ and Cartan-type matrix $2B'=2I-\mathcal{I}$ of dimension $d'=d+1$ by replacing the continued fraction expansion $[\nu_0,\ldots,\nu_n+2]$ by $[\nu_0+1,\nu_1,\ldots,\nu_n+2]$, so that $\mathcal{I}'$ becomes the incidence matrix corresponding to the continued fraction expansion of $p'/(p'-p)$. Also define $\mathcal{I}''$ and $2B''=2I-\mathcal{I}''$ of dimension $d''=d+2$ as

$$\mathcal{I}''_{i,j}=\begin{cases}\delta_{i,3}-\delta_{j,3}, & \text{for } i=1 \text{ or } j=1,\\ \mathcal{I}'_{i-1,j-1}, & \text{for } i,j=2,...,d+2.\end{cases} \tag{32}$$

Then the above sequence of transformations implies the following proposition.

Proposition IV.4: For integers p,p' with $p<p'<3p/2$ and $gcd(p,p')=1$ let integers $1\leqslant r<p$ and $1\leqslant s<p'$ be fixed by $|p'(r-s)r+ps|=1$ and let $\mathcal{I}$ and B be defined by (32). Then the following polynomial identity holds for $L\in\mathbb{Z}$:

$$\sum{}' q^{mB''m/4}\prod_{j=1}^{d''}\begin{bmatrix}L\delta_{j,1}+\frac{1}{2}(\mathcal{I}''m)_j\\ m_j\end{bmatrix}$$
$$=\sum_{j=-\infty}^{\infty}\left\{q^{j(p'(2p'-p)j+1)}\begin{bmatrix}L\\ 2p'j\end{bmatrix}_2-q^{(p'j+s)((2p'-p)j+r+s)}\begin{bmatrix}L\\ 2p'j+2s\end{bmatrix}_2\right\}.$$

The identities of Proposition IV. 4, which are related to the $\phi_{2,1}$ perturbation of the conformal field theory $M(p',2p'-p)$, were conjectured in Ref. 21 [as Eq. (6.9)]. For $p=p'-1$ a proof using recurrences was recently given in Ref. 20.

D. *q*-trinomial identities III

There are, of course, many more q-trinomial identities that can be derived using the techniques of the previous sections. Our final application is to show that in some cases a bit more ingenuity is required to arrive at the desired result. The identities we set out to prove here were again conjectured by Berkovich, McCoy, and Pearce [Eq. (9.4) of Ref. 21] and are interesting, as they contain the (polynomial) Rogers–Ramanujan identity (20) as the simplest case. It also provides an example for which the matrix $B=b_1\otimes b_2$ (in the proposition below denoted as C_n) of Sec. IV A has $b_1=(1)$ and not $(\frac{1}{2})$.

Proposition IV.5: For $n\geqslant 1$, let C_n be the Cartan matrix of A_n. Then for all $L\in\mathbb{Z}$,

$$\sum_{m\in\mathbb{Z}^n} q^{mC_nm/2}\prod_{j=1}^{n}\begin{bmatrix}L\delta_{j,1}+m_j-(C_nm)_j\\ m_j\end{bmatrix}$$
$$=\sum_{j=-\infty}^{\infty}\left\{q^{((n+3)(n+4)j+2)j/2}\begin{bmatrix}L\\ (n+4)_j\end{bmatrix}_2-q^{((n+3)j+2)((n+4)j+2)/2}\begin{bmatrix}L\\ (n+4)j+2\end{bmatrix}_2\right\}. \tag{33}$$

Letting L tend to infinity using (8) and (A6), this yields the following Virasoro-character identities.

Corollary IV.1: For $n\geqslant 1$ and $|q|<1$,

$$\sum_{m\in\mathbb{Z}^n}\frac{q^{mC_n m/2}}{(q)_{m_1}}\prod_{j=2}^{n}\begin{bmatrix} m_j-(C_n m)_j \\ m_j\end{bmatrix}=\begin{cases}\chi_{1,2}^{((n+3)/2,n+4)}(q), & n \text{ odd},\\ \chi_{1,2}^{((n+4)/2,n+3)}(q), & n \text{ even}.\end{cases} \tag{34}$$

In Ref. 21, the identities (33) and (34) were associated with the $\phi_{2,1}$ perturbation of the conformal field theories $M((n+4)/2,n+3)$ when n is odd and the $\phi_{1,5}$ perturbation of $M((n+3)/2,n+4)$ when n is even.

Proof: The corollary betrays a hidden parity dependence of (33), which also plays a role in the proof. Treating n being odd first, we set $n=2k-1$. The left-hand side of (33) then reads as

$$\sum_{m\in\mathbb{Z}^{2k-1}} q^{mC_{2k-1}m/2}\prod_{j=1}^{2k-1}\begin{bmatrix}\frac{1}{2}L\,\delta_{j,1}+m_{j-1}-m_j+m_{j+1}\\ m_j\end{bmatrix}, \tag{35}$$

with the convention that $m_0=L/2$ and $m_{2k}=0$. We eliminate the variables m_{2j-1}, $j=1,...,k$ in favor of new variables $M_1,\ldots,M_k$, defined as

$$m_{2j-1}=m_{2j-2}-\tfrac{1}{2}(M_j-M_{j+1}),$$

where $M_{k+1}=0$. If after this replacement we relabel m_{2j} to m_j for $j=1,...,k$ (so that $m_k=0$), expression (35) becomes

$$\sum_{M+Le_1\in 2\mathbb{Z}^k} q^{(L(L-2M_1)+M_1^2+\Sigma_{i,j=2}^{k}M_i(C_{k-1})_{i,j}M_j)/4}$$
$$\times\sum_{m_1,\ldots,m_{k-1}} q^{\Sigma_{j=1}^{k-1}(M_{j+1}-m_j)(m_{j-1}-m_j-(M_j-M_{j+2})/2)}$$
$$\times\begin{bmatrix}m_0+m_1+\frac{1}{2}(M_1-M_2)\\ m_0-\frac{1}{2}(M_1-M_2)\end{bmatrix}\prod_{j=1}^{k-1}\begin{bmatrix}m_{j-1}-\frac{1}{2}(M_j-M_{j+2})\\ m_j\end{bmatrix}\begin{bmatrix}m_{j+1}+\frac{1}{2}(M_{j+1}-M_{j+2})\\ M_j-\frac{1}{2}(M_{j+1}-M_{j+2})\end{bmatrix}. \tag{36}$$

This allows for successive summation over $m_{k-1},\ldots,m_1$ by the q-Saalschütz sum (A4). When summing over m_j, we take (A4) with $L\to m_{j-1}-(M_j-M_{j+2})/2$, $a\to(M_{j+1}-M_{j+2})/2$, $b\to -(M_{j+1}+M_{j+2})/2$, $c\to(M_j-M_{j+1})/2$ (for $j\geqslant 2$), and $c\to m_0+(M_1-M_2)/2$ (for $j=1$). As a result, (36) collapses into

$$\sum_{M+Le_1\in 2\mathbb{Z}^k} q^{L(L-2M_1)/4+MBM/2}\prod_{j=1}^{k}\begin{bmatrix}\frac{1}{2}L(\delta_{j,1}+\delta_{j,2})+\frac{1}{2}(\mathcal{I}M)_j\\ M_j\end{bmatrix}, \tag{37}$$

with matrices $\mathcal{I}$ and $2B=2I-\mathcal{I}$ defined in Eqs. (24) and (25) corresponding to the continued fraction expansion of $(k+2)/(k+1)=[1,k-1]$, i.e.,

$$\mathcal{I}_{i,j}=\begin{cases}\delta_{i,1}\delta_{j,1}+\delta_{i,2}-\delta_{j,2}, & \text{for } i=1 \text{ or } j=1,\\ \delta_{i,j-1}+\delta_{i,j+1}, & \text{for } i,j=2,...,k.\end{cases}$$

The last part of the proof consists of the observation that the identity obtained by equating (37) with the right-hand side of (33) (with $n=2k-1$) is nothing but the identity of Proposition IV.2 with $(p,p')=(k+2,2k+3)$ after letting $q\to 1/q$. This is readily seen using (4) and (A7).

Next, we deal with n being even, setting $n=2k$. The left-hand side of (33) then is

$$\sum_{m\in\mathbb{Z}^{2k}} q^{mC_{2k}m/2}\prod_{j=1}^{2k}\begin{bmatrix} L\delta_{j,1}+m_{j-1}-m_j+m_{j+1} \\ m_j\end{bmatrix}, \tag{38}$$

where $m_0=m_{2k+1}=0$. We eliminate the variables m_{2j}, $j=1,...,k$, introducing new variables $m_0,\ldots,M_{k-1}$ by

$$m_{2j}=m_{2j-1}-\tfrac{1}{2}(M_{j-1}-M_j),$$

where $M_k=0$. After this replacement we shift $m_{2j-1}\to m_j$ for $j=1,...,k$ so that expression (38) becomes

$$\sum_{\substack{M_0,\ldots,M_{k-1}\\ M_j\ \text{even}}} q^{(M_0^2+\Sigma_{i,j=1}^{k-1}M_i(C_{k-1})_{i,j}M_j)/4}$$
$$\times\sum_{m_1,\ldots,m_k} q^{m_1(m_1-(M_0+M_1)/2)+\Sigma_{j=2}^{k}(M_{j-1}-m_j)(m_{j-1}-m_j-(M_{j-2}-M_j)/2)}$$
$$\times\begin{bmatrix} L-\frac{1}{2}(M_0-M_1) \\ m_1\end{bmatrix}\left(\prod_{j=2}^{k}\begin{bmatrix} m_{j-1}-\frac{1}{2}(M_{j-2}-M_j) \\ m_j\end{bmatrix}\right)\prod_{j=1}^{k}\begin{bmatrix} m_{j+1}+\frac{1}{2}(M_{j-1}-M_j) \\ m_j-\frac{1}{2}(M_{j-1}-M_j)\end{bmatrix}.$$

We now sum over $m_k,\ldots,m_3$ by successive application of the q-Saalschütz sum (A4). When summing over m_j we take (A4) with $L\to m_{j-1}-(M_{j-2}-M_j)/2$, $a\to(M_{j-1}-M_j)/2$, $b\to-(M_{j-1}+M_j)/2$, and $c\to(M_{j-2}-M_{j-1})/2$. The final sum over m_1 follows from (A1) with $L\to L-(M_0-M_1)/2$, $a\to(M_0-M_1)/2$, and $b\to-(M_0+M_1)/2$. Setting $M_0\to 2i$, the resulting expression is

$$\sum_{i\geqslant 0} q^{i^2}\begin{bmatrix} L \\ 2i\end{bmatrix}\sum_{M\in 2\mathbb{Z}^{k-1}} q^{MC_{k-1}M/4}\prod_{j=1}^{k-1}\begin{bmatrix} i\delta_{j,1}+M_j-\frac{1}{2}(C_{k-1}M)_j \\ M_j\end{bmatrix}.$$

Equating this with the right-hand side of (33) for $n=2k$, we recognize identity (29) with $(p,p')=(k+1,k+2)$. □

V. THE TRINOMIAL BAILEY LEMMA

In this final section of our paper we formulate some of our results in the language of Bailey pairs. As we will see, the connection coefficients obtained in Sec. II provide a very elementary proof of the trinomial analog of Bailey's lemma recently obtained by Andrews and Berkovich.[19]

First, some definitions are needed. In subsequent formulas, $T_n(L,a)/(q)_L$ will be abbreviated to $Q_n(L,a)$.

Definition V.1: A pair of sequences $\alpha=\{\alpha_L\}_{L\geqslant 0}$ *and* $\beta=\{\beta_L\}_{L\geqslant 0}$ *that satisfies*

$$\beta_L=\sum_{r=0}^{L}\frac{\alpha_r}{(q)_{L-r}(aq)_{L+r}},$$

forms a (binomial) Bailey pair relative to a.

Definition V.2: A pair of sequences $A=\{A_L\}_{L\geqslant 0}$ *and* $B=\{B_L\}_{L\geqslant 0}$ *that satisfies*

$$B_L=\sum_{r=0}^{L} Q_n(L,r)A_r,$$

forms a trinomial Bailey pair relative to n.

The Bailey lemma[43] and trinomial Bailey lemma[19] can now be stated as the following summation formulas.

Lemma V.1. Let (α,β) be a Bailey pair relative to a. Then

$$\sum_{L=0}^{M} \frac{(\rho_1)_L(\rho_2)_L(aq/\rho_1\rho_2)^L\alpha_L}{(aq/\rho_1)_L(aq/\rho_2)_L(q)_{M-L}(aq)_{M+L}} = \sum_{L=0}^{M} \frac{(\rho_1)_L(\rho_2)_L(aq/\rho_1\rho_2)^L(aq/\rho_1\rho_2)_{M-L}\beta_L}{(aq/\rho_1)_M(aq/\rho_2)_M(q)_{M-L}}.$$

Lemma V.2. Let (A,B) form a trinomial Bailey pair relative to 0. Then

$$\sum_{L=0}^{M} (-1)_L q^{L/2} B_L = (-1)_{M+1} \sum_{L=0}^{M} q^{L/2} A_L \frac{Q_1(M,L)}{1+q^L}. \tag{39}$$

If (A,B) is a trinomial Bailey pair relative to 1, then

$$\sum_{L=0}^{M} (-q^{-1})_L q^L B_L = (-1)_M \sum_{L=0}^{M} A_L \left\{ Q_1(M,L) - \frac{Q_1(M-1,L+1)}{1+q^{-L-1}} - \frac{Q_1(M-1,L-1)}{1+q^{L-1}} \right\}.$$

Before we translate the results of Sec. II in the language of Bailey pairs, let us point out that the connection coefficients between q binomials and q trinomials can be applied to yield a very simple proof of the trinomial Bailey lemma. At the heart of the proof of Lemma V.2 is the following identity derived in Ref. 19 by a considerable amount of work,

$$T_0(L,a) = q^{(a-L)/2} \left\{ \frac{1+q^L}{1+q^a} T_1(L,a) - \frac{1-q^L}{1+q^a} T_1(L-1,a) \right\}. \tag{40}$$

To see, for example, that this implies (39), we multiply (40) by $q^{L/2}(-1)_L/(q)_L$ and sum over L from a to M. On the right-hand side all but one term cancels, so that

$$\sum_{L=a}^{M} q^{L/2}(-1)_L Q_0(L,a) = \frac{q^{a/2}}{1+q^a}(-1)_{M+1} Q_1(M,a),$$

which obviously implies (39).

By Eqs. (9)–(13), Eq. (40) is proved if we can show its validity when multiplied by $C_{M,L}(a)$ and summed over L. Doing this and using (10), one finds (replacing $L \to k$ and $M \to L$)

$$\begin{aligned}
\begin{bmatrix} 2L \\ L-a \end{bmatrix} &= \sum_{k=0}^{L} q^{\binom{k}{2}-\binom{a}{2}} \begin{bmatrix} L \\ k \end{bmatrix} \left\{ \frac{1+q^k}{1+q^a} T_1(k,a) - \frac{1-q^k}{1+q^a} T_1(k-1,a) \right\} \\
&= \sum_{k=0}^{L} q^{\binom{k}{2}-\binom{a}{2}} \left\{ \frac{1+q^k}{1+q^a} \begin{bmatrix} L \\ k \end{bmatrix} - q^k \frac{1-q^{k+1}}{1+q^a} \begin{bmatrix} L \\ k+1 \end{bmatrix} \right\} T_1(k,a) \\
&= \sum_{k=0}^{L} q^{\binom{k}{2}-\binom{a}{2}} \frac{1+q^L}{1+q^a} \begin{bmatrix} L \\ k \end{bmatrix} T_1(k,a).
\end{aligned}$$

But the extremes of this string of equations is nothing but Eq. (14), with $D'_{L,k}(a)$ given by Eq. (16) of Lemma II.2, establishing (40).

We now give a series of lemmas that are all straightforward consequences of the results of Sec. II.

Lemma V.3: Let (α,β) be a Bailey pair relative to 1. Then

$$A_L = q^{-L^2/2}\alpha_L, \quad B_L = \sum_{k=0}^{L} \frac{(-1)^{L-k} q^{\binom{L-k}{2} - L^2/2}(q)_{2k}}{(q)_k(q)_{L-k}} \beta_k$$

is a trinomial Bailey pair relative to 0 and

$$A_L=\frac{q^{-\binom{L}{2}}}{1+q^L}\alpha_L\,,\quad B_L=\sum_{k=0}^{L}\frac{(-1)^{L-k}q^{\binom{L-k}{2}-\binom{L}{2}}(q)_{2k}}{(1+q^k)(q)_k(q)_{L-k}}\beta_k$$

is a trinomial Bailey pair relative to 1.

The converse statement is as follows.

Lemma V.4: Let $(A(n),B(n))$ *be a trinomial Bailey pair relative to n. Then,*

$$\alpha_L=q^{L^2/2}A_L(0),\quad \beta_L=\frac{(q)_L}{(q)_{2L}}\sum_{k=0}^{L}\frac{q^{k^2/2}}{(q)_{L-k}}B_k(0)$$

and

$$\alpha_L=q^{\binom{L}{2}}(1+q^L)A_L(1),\quad \beta_L=\frac{(q)_L}{(q)_{2L}}(1+q^L)\sum_{k=0}^{L}\frac{q^{\binom{k}{2}}}{(q)_{L-k}}B_k(1),$$

are Bailey pairs relative to 1.

Lemma V.3 is to be compared with the following result of Ref. 23.

Lemma V.5: Let l be a non-negative integer and (α,β) a Bailey pair relative to $a=q^l$. *Then*

$$A_L=\begin{cases}\alpha_{(L-l)/2}\,, & \text{for } L=l,l+2,\ldots,\\ 0, & \text{otherwise;}\end{cases}$$

$$B_L=\begin{cases}\displaystyle\sum_{k=0}^{[(L-l)/2]}\frac{q^{(L-l-2k)(L-l-2k-n)/2}}{(q)_l(q)_{L-l-2k}}\beta_k\,, & \textit{for } L\geqslant l;\\ 0, & \text{otherwise,}\end{cases}$$

forms a trinomial Bailey pair relative to n.

ACKNOWLEDGMENTS

I thank George E. Andrews, Alexander Berkovich, and Anne Schilling for helpful discussions. This work has been supported by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

APPENDIX: SOME *q*-BINOMIAL FORMULAS

In this appendix we list some standard q-binomial identities that are repeatedly used in the main text.

The following three formulas all hold for integers a, b, L such that a, $L\geqslant 0$,

$$\sum_{k=0}^{L}q^{k(k+b)}\begin{bmatrix}L\\k\end{bmatrix}\begin{bmatrix}a\\k+b\end{bmatrix}=\begin{bmatrix}a+L\\b+L\end{bmatrix}, \tag{A1}$$

$$\sum_{k=0}^{L}(-1)^k q^{\binom{k}{2}}\begin{bmatrix}L\\k\end{bmatrix}\begin{bmatrix}L+a-k\\b\end{bmatrix}=q^{L(L+a-b)}\begin{bmatrix}a\\b-L\end{bmatrix}, \tag{A2}$$

$$\sum_{k=0}^{L}(-1)^k q^{\binom{k}{2}+k(b-L+1)}\begin{bmatrix}L\\k\end{bmatrix}\begin{bmatrix}L+a-k\\b\end{bmatrix}=\begin{bmatrix}a\\b-L\end{bmatrix}. \tag{A3}$$

The first two equations are specializations of the q-Chu–Vandermonde sum (II.7) of Ref. 3 and the last equation is a specialization of the q-Chu–Vandermonde sum (II.6) of Ref. 3. Identity (A2) is also given in Ref. 1 as Eq. (3.3.10). A useful specialization of the q-Saalschütz sum [(II.12) of Ref. 3] is given by

$$\sum_{k=0}^{L} q^{(a-b-k)(L-k)}\begin{bmatrix} L \\ k \end{bmatrix}\begin{bmatrix} a \\ k+b \end{bmatrix}\begin{bmatrix} k+c \\ a+L \end{bmatrix}=\begin{bmatrix} c \\ b+L \end{bmatrix}\begin{bmatrix} c-b \\ a-b \end{bmatrix}, \tag{A4}$$

true for integers a, b, c, L such that a, c, $L \geqslant 0$. This is Eq. (3.3.11) of Ref. 1. Finally, we list the elementary results:

$$\begin{bmatrix} L \\ a \end{bmatrix}=\begin{bmatrix} L-1 \\ a-1 \end{bmatrix}+q^{a}\begin{bmatrix} L-1 \\ a \end{bmatrix}, \quad \text{for } L,a \geqslant 0, \quad L+a \neq 0, \tag{A5}$$

$$\lim_{L\to\infty}\begin{bmatrix} L \\ a \end{bmatrix}=\frac{1}{(q)_a} \tag{A6}$$

and

$$\begin{bmatrix} L \\ a \end{bmatrix}_{1/q}=q^{-a(L-a)}\begin{bmatrix} L \\ a \end{bmatrix}_{q}. \tag{A7}$$

[1] G. E. Andrews, ''The theory of partitions,'' *Encyclopedia of Mathematics and its Applications* (Addison-Wesley, Reading, MA, 1976), Vol. 2.
[2] G. E. Andrews, ''q-Series: Their development and application in analysis, number theory, combinatorics, physics, and computer algebra,'' in *CBMS Regional Conference Series in Mathemancs* (AMS, Providence, RI, 1985), Vol. 66.
[3] G. Gasper and M. Rahman, ''Basic hypergeometric series,'' *Encyclopedia of Mathematics and its Applications* (Cambridge Univiversity Press, Cambridge, 1990).
[4] P. A. MacMahon, *Combinatory Analysis* (Cambridge University Press, London and New York, 1916), Vol. 2.
[5] P. J. Forrester and G. E. Andrews, ''Height probabilities in solid-on-solid models. I,'' J. Phys. A **19**, L923–L926 (1986).
[6] G. E. Andrews and R. J. Baxter, ''Lattice gas generalization of the hard hexagon model. III. q-Trinomial coefficients,'' J. Stat. Phys. **47**, 297–330 (1987).
[7] G. E. Andrews, ''q-Trinomial coefficients and Rogers–Ramanujan type identities,'' in *Analytic Number Theory*, edited by B. C. Berndt, H. G. Diamond, and A. J. Hildebrand (Birkhäuser, Boston, 1990), pp. 1–11.
[8] G. E. Andrews, ''Euler's 'Exemplum memorabile inductionis fallacis' and q-trinomial coefficients,'' J. Am. Math. Soc. **3**, 653–669 (1990).
[9] G. E. Andrews, ''Schur's theorem, Capparelli's conjecture and q-trinomial coefficients,'' Contemp. Math. **166**, 141–154 (1994).
[10] S. O. Warnaar and P. A. Pearce, ''Exceptional structure of the dilute A_3 model: E_8 and E_7 Rogers–Ramanujan identities,'' J. Phys. A **27**, L891–L897 (1994).
[11] S. O. Warnaar, P. A. Pearce, K. A. Seaton, and B. Nienhuis, ''Order parameters of the dilute A models,'' J. Stat. Phys. **74**, 469–531 (1994).
[12] G. E. Andrews, ''Rogers–Ramanujan polynomials for modulus 6,'' in *Analytic Number Theory: Proceedings of a Conference in Honor of Heini Halberstam*, edited by B. C. Berndt, H. G. Diamond, and A. J. Hildebrand (Birkhäuser, Boston, 1996), Vol. 1, pp. 17–30.
[13] A. Berkovich, B. M. McCoy, and W. P. Orrick, ''Polynomial identities, indices, and duality for the $N=1$ superconformal model $SM(2,4\nu)$,'' J. Stat. Phys. **83**, 795–837 (1996).
[14] A. Schilling, ''Multinomials and polynomial bosonic forms for the branching functions of the $\widehat{su}(2)_M \times \widehat{su}(2)_N/\widehat{su}(2)_{N+M}$ conformal coset models,'' Nucl. Phys. B **467**, 247–271 (1996).
[15] S. O. Warnaar, ''Fermionic solution of the Andrews–Baxter–Forrester model. II. Proof of Melzer's polynomial identities,'' J. Stat. Phys. **84**, 49–83 (1996).
[16] A. Berkovich and B. M. McCoy, ''Generalizations of the Andrews–Bressoud identities for the $N=1$ superconformal model $SM(2,2\nu)$,'' Math. Comput. Modell. **26**, 37–49 (1997).
[17] K. A. Seaton and L. C. Scott, ''q-Trinomial coefficients and the dilute A model,'' J. Phys. A **30**, 7667–7676 (1997).
[18] S. O. Warnaar, ''The Andrews–Gordon identities and q-multinomial coefficients,'' Commun. Math. Phys. **184**, 203–232 (1997).
[19] G. E. Andrews and A. Berkovich, ''A trinomial analogue of Bailey's lemma and $N=2$ superconformal invariance,'' Commun. Math. Phys. **192**, 245–260 (1998).
[20] A. Berkovich and B. M. McCoy, ''The perturbation $\phi_{2,1}$ of the $M(p,p+1)$ models of conformal field theory and related polynomial character identities,'' preprint ITP-SB-98-49, math.QA/9809066. To appear in the Ramanujan Journal.

[21] A. Berkovich, B. M. McCoy, and P. A. Pearce, ''The perturbations $\phi_{2,1}$ and $\phi_{1,5}$ of the minimal models $M(p,p')$ and the trinomial analogue of Bailey's lemma,'' Nucl. Phys. B **519**, 597–625 (1998).
[22] A. Schilling and S. O. Warnaar, ''Supernomial coefficients, polynomial identities and *q*-series,'' Ramanujan J. **2**, 459–494 (1998).
[23] S. O. Warnaar, ''A note on the trinomial analogue of Bailey's lemma,'' J. Comb. Theory, Ser. A **81**, 114–118 (1998).
[24] G. E. Andrews, ''A polynomial identity which implies the Rogers–Ramanujan identities,'' Scr. Math. **28**, 297–305 (1970).
[25] I. J. Schur, ''Ein Beitrag zur additiven Zahlentheorie und zur Theorie der Kettenbrüche,'' S.-B. Preuss. Akad. Wiss. Phys.-Math. Kl. **X**, 302–321 (1917).
[26] L. J. Slater, ''Further identities of the Rogers–Ramanujan type,'' Proc. London Math. Soc. **54**, 147–167 (1952).
[27] H. Göllnitz, ''Partitionen mit differenzenbedienungen,'' J. Reine Angew. Math. **225**, 154–190 (1967).
[28] B. Gordon, ''Some continued fractions of the Rogers–Ramanujan type,'' Duke Math. J. **31**, 741–748 (1965).
[29] A. Berkovich and B. M. McCoy, ''Continued fractions and fermionic representations for characters of $M(p,p')$ minimal models,'' Lett. Math. Phys. **37**, 49–66 (1996).
[30] A. Berkovich, B. M. McCoy, and A. Schilling, ''Rogers–Schur–Ramanujan type identities for the $M(p,p')$ minimal models of conformal field theory,'' Commun. Math. Phys. **191**, 325–395 (1998).
[31] B. L. Feigin and D. B. Fuchs, ''Verma modules over a Virasoro algebra,'' Funct. Anal. Appl. **17**, 241–242 (1983).
[32] A. Rocha-Caridi, ''Vacuum vector representation of the Virasoro algebra,'' in *Vertex Operators in Mathematics and Physics*, edited by J. Lepowsky, S. Mandelstam, and I. M. Singer (Springer-Verlag, Berlin, 1985), pp. 451–473.
[33] V. K. Dobrev, ''Characters of the irreducible highest weight modules over the Virasoro and super-Virasoro algebras,'' Rend. Circ. Mat. Palermo (2) Suppl. **14**, 25–42 (1987).
[34] G. E. Andrews, R. J. Baxter, and P. J. Forrester, ''Eight-vertex SOS model and generalized Rogers–Ramanujan-type identities,'' J. Stat. Phys. **35**, 193–266 (1984).
[35] R. J. Baxter and P. J. Forrester, ''Further exact solutions of the eight-vertex SOS model and generalizations of the Rogers–Ramanujan identities,'' J. Stat. Phys. **38**, 435–472 (1985).
[36] G. E. Andrews, R. J. Baxter, D. M. Bressoud, W. H. Burge, P. J. Forrester, and G. Viennot, ''Partitions with prescribed hook differences,'' Eur. J. Comb. **8**, 341–350 (1987).
[37] O. Foda, K. S. M. Lee, and T. A. Welsh, ''A Burge tree of Virasoro-type polynomial identities,'' Int. J. Mod. Phys. A **13**, 4967–5012 (1998).
[38] A. Berkovich, ''Fermionic counting of RSOS-states and Virasoro character formulas for the unitary minimal series $M(\nu,\nu+1)$. Exact results,'' Nucl. Phys. B **431**, 315–348 (1994).
[39] S. O. Warnaar, ''Fermionic solution of the Andrews–Baxter–Forrester model. I. Unification of CTM and TBA methods,'' J. Stat. Phys. **82**, 657–685 (1996).
[40] O. Foda and Y.-H. Quano, ''Polynomial identities of the Rogers–Ramanujan type,'' Int. J. Mod. Phys. A **10**, 2291–2315 (1995).
[41] S. Dasmahapatra, R. Kedem, T. R. Klassen, B. M. McCoy, and E. Melzer, ''Quasi-particles, conformal field theory, and *q*-series,'' Int. J. Mod. Phys. B **7**, 3617–3648 (1993).
[42] A. Schilling, ''Polynomial fermionic forms for the branching functions of the rational coset conformal field theories $\widehat{su}(2)_M\times\widehat{su}(2)_N/\widehat{su}(2)_{N+M}$,'' Nucl. Phys. B **459**, 393–436 (1996).
[43] W. N. Bailey, ''Identities of the Rogers–Ramanujan type,'' Proc. London Math. Soc. **50**, 1–10 (1949).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 5 MAY 1999

Comment on phase-space representation of quantum state vectors

Klaus B. Møller[a)]
Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, Pasadena, California 91125

(Received 13 July 1998; accepted for publication 1 February 1999)

A simple approach to phase-space representation of quantum state vectors using the displacement-operator formalism is presented. Although the resulting expressions for the fundamental operators (position and momentum) are equivalent to those obtained by other methods, this approach provides both alternative mathematical foundation as well as physical interpretation of phase-space representation of quantum state vectors.
[S0022-2488(99)01605-9]

I. INTRODUCTION

Over the past few years there has been a renewed interest in phase-space descriptions of quantum systems. In a recent paper by Ban[1] a novel approach to phase-space representation of quantum state vectors is obtained within the relative-state formulation, and in this Comment we make a few remarks on the physical contents of this construction. Also, we relate it to another approach to phase-space representation of quantum state vectors, the so-called displacement-operator approach.

The idea of phase-space representation of quantum state vectors, i.e., representation of a quantum state as a probability amplitude depending on *two real* variables related to the position and momentum coordinates goes back to the works of Fock[2] and Bargmann.[3] In their formulation, a quantum state is represented as a complex function depending on *one complex* coordinate whose real and imaginary part is proportional to the position and momentum coordinate, respectively. This is a result of regarding the bosonic creation and annihilation operators as the fundamental operators.

The relative-state formulation, on the other hand, treats the position and momentum operators themselves as the fundamental operators and is therefore more closely related to the works of Torres-Vega and Frederick[4] and Harriman.[5] Both of these works rely to a certain extent on Dirac's representation theory of quantum mechanics,[6] either as a Hilbert-space-vector approach postulating the existence of a complete set of states depending on two real parameters that can be used as a basis in phase space or a linear transformation onto phase space from position or momentum space.

In fact, the relative-state representation of Ban[1] becomes, under certain conditions, equivalent to those of Torres-Vega and Frederick[4] and Harriman.[5] The relative-state formulation may therefore serve as a mathematical and physical foundation for the representations presented by these authors since it is derived from first principles without assumptions or transformations from other representations.

However, the relative-state formulation is not the only way to construct a phase-space representation of quantum state vectors from first principles that becomes equivalent to those of Torres-Vega and Frederick[4] and Harriman.[5] Below, we present an alternative construction, using the displacement operators, and discuss the mathematical and physical differences between this method the relative-state approach.

[a)]Electronic mail: klaus@cco.caltech.edu

The displacement-operator approach is essentially equivalent to the coherent-state formalism as put forward by, for instance, Klauder and Skagerstam[7] and studied in some detail by the present author.[8] Hence, the presentation given here is extracted from these earlier works and put in a form relevant for the present discussion. For a thorough review and analysis of the use of displacement operators, the reader is referred to Refs. 7 and 8 and the references therein.

II. RELATIVE-STATE FORMULATION

The relative-state formulation is presented in great detail by Ban[1] and here we only include a few results relevant for the further discussion. The key of this approach is to enlarge the Hilbert space $\mathcal{H}$ of a quantum system by introducing an auxiliary (reference) quantum system and treat quantum state vectors in the extended Hilbert space $\tilde{\mathcal{H}}=\mathcal{H}\otimes\mathcal{H}_r$, where $\mathcal{H}_r$ is the Hilbert space of the reference system. A state vector in the extended Hilbert space $\tilde{\mathcal{H}}$ then becomes $|\Psi\rangle\rangle\equiv|\psi\rangle\otimes|\phi\rangle_r$ where $|\phi\rangle_r$ is the reference state.

A set of state vectors $\{|\omega(r,k;s)\rangle\rangle|r,k\in\mathbf{R}\}$ may be introduced[1] that becomes a complete orthonormal system in $\tilde{\mathcal{H}}$. These state vectors, which can be written on the following form:

$$|\omega(r,k;s)\rangle\rangle\equiv\frac{1}{\sqrt{2\pi}}e^{-i(1+s)kr/2}\int_{-\infty}^{\infty}dx\, e^{ikx}|x\rangle\otimes|x-r\rangle_r \tag{1}$$

(as in Ref. 1, we set $\hbar=1$ throughout this Comment) are simultaneous eigenstates of the operators $\hat{x}-\hat{x}_r$, and $\hat{p}+\hat{p}_r$,

$$(\hat{x}-\hat{x}_r)|\omega(r,k;s)\rangle\rangle=r|\omega(r,k;s)\rangle\rangle, \tag{2}$$

$$(\hat{p}+\hat{p}_r)|\omega(r,k;s)\rangle\rangle=k|\omega(r,k;s)\rangle\rangle. \tag{3}$$

However, when we investigate the properties of the relevant quantum system, we only need a description of this system in the Hilbert space $\mathcal{H}$. Thus, the extended Hilbert space is reduced again by fixing the state vector of the reference system. For any fixed state vector $|\phi\rangle_r$ of the reference system, the set $\{|\omega(r,k;s)\rangle|r,k\in\mathbf{R}\}$, where

$$|\omega(r,k;s)\rangle\equiv{}_r\langle\phi|\omega(r,k;s)\rangle\rangle=\frac{1}{\sqrt{2\pi}}e^{-i(1+s)kr/2}\int_{-\infty}^{\infty}dx\, e^{ikx}\phi^*(x-r)|x\rangle, \tag{4}$$

becomes an overcomplete system in the Hilbert space $\mathcal{H}$.[1] Therefore, the relevant quantum system can be represented by an $\mathcal{L}^2(2)$ normalized wave function $\psi_\omega(r,k;s)\equiv\langle\omega(r,k;s)|\psi\rangle$ depending of the two real parameters k and r. In this representation, the fundamental operators $\hat{x}$ and $\hat{p}$ take the form

$$\langle\omega(r,k;s)|\hat{x}|\psi\rangle=\left[\frac{1}{2}(1+s)r+i\frac{\partial}{\partial k}\right]\psi_\omega(r,k;s), \tag{5}$$

$$\langle\omega(r,k;s)|\hat{p}|\psi\rangle=\left[\frac{1}{2}(1-s)k-i\frac{\partial}{\partial r}\right]\psi_\omega(r,k;s). \tag{6}$$

Apart from some notational differences these are essentially the expressions given by Torres-Vega and Frederick[4] and Harriman[5] in their representations. Thus, the construction by Ban[1] may serve as a mathematical foundation for the work of Torres-Vega and Frederick[4] and Harriman.[5] Furthermore, the relative-state formulation provides a physical interpretation of the wave function $\psi_\omega(r,k;s)$ and the parameters k and r as phase-space coordinates. In light of Eqs. (2) and (3), the function $|\psi_\omega(r,k;s)|^2$ represents the probability distributions of the eigenvalues of the operators $\hat{x}-\hat{x}_r$ and $\hat{p}+\hat{p}_r$ in the extended Hilbert space $\tilde{\mathcal{H}}$.

Alternatively, one may utilize $|\psi_\omega(r,k;s)|^2$ as a combined probability distribution directly in the r,k-parametrized space as follows:

$$\overline{r} \equiv \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk\, r |\psi_\omega(r,k;s)|^2 = x_\psi - x_\phi, \tag{7}$$

$$\overline{k} \equiv \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk\, k |\psi_\omega(r,k;s)|^2 = p_\psi + p_\phi. \tag{8}$$

Here, $x_\psi = \langle \psi | \hat{x} | \psi \rangle$ and so on. Thus, r and k may be interpreted as phase-space coordinates in the sense that the average value of r equals the relative position between the relevant and the reference system, and the average value of k equals the sum of the momenta of the relevant and the reference system. Hence, the physical interpretation of the wave function depends on the reference state, although the operator expressions, Eqs. (5) and (6), do not, and from this point of view the most satisfactory representation is obtained using a reference state with $\langle \phi | \hat{x} | \phi \rangle = 0$ and $\langle \phi | \hat{p} | \phi \rangle = 0$. In general, also the physical interpretation of higher momenta of r and k depend on the reference system.[1]

III. DISPLACEMENT-OPERATOR APPROACH

Here we present an alternative derivation from first principles of the phase-space representation of quantum state vectors that also becomes equivalent to the ones of Torres-Vega and Frederick[4] and Harriman[5] and therefore to the result of Ban,[1] as well. However, the derivation presented here differs from the one obtained in the relative-state formulation in both the mathematical foundation and in the physical interpretation of the phase-space wave functions. In fact, it resembles closely Dirac's construction of the usual position and momentum representations.[6]

Two things are important for the definitions of these representations. First, the basis states, denoted by $|r\rangle_x$ and $|k\rangle_p$, are eigenstates of the position and momentum operator, respectively,

$$\hat{x}|r\rangle_x = r|r\rangle_x \quad \text{and} \quad \hat{p}|k\rangle_p = k|k\rangle_p. \tag{9}$$

Second, the position (momentum) eigenstate with eigenvalue $r(k)$ can be generated from the eigenstate with eigenvalue $r=0$ $(k=0)$ by a displacement operator,

$$|r\rangle_x = \hat{D}_x(r)|0\rangle_x \quad \text{and} \quad |k\rangle_p = \hat{D}_p(k)|0\rangle_p, \tag{10}$$

where the displacement operators are given as $\hat{D}_x(r) = \exp(-ir\hat{p})$ and $\hat{D}_p(k) = \exp(ik\hat{x})$.[6] The wave function in position (momentum) space is then obtained by projection onto a position (momentum) eigenstate, $\psi(r) \equiv \langle r | \psi \rangle$ $(\psi(k) \equiv \langle k | \psi \rangle)$. This implies that the displacement operators when acting on a state displace the expectation value of the position or momentum by r and k, respectively.

An identical approach to a phase-space representation of quantum state vectors would require the existence of an Hermitian operator representing a point in phase space. Torres-Vega and Frederick[4] claim that such an operator exist but without proof and, in fact, the existence of such an operator would violate the Heisenberg uncertainty relation. In the relative-state formulation, a close resemblance is obtained for the basis states $|\omega(r,k;s)\rangle\rangle$ in the extended Hilbert space; cf. Eqs. (2) and (3).

Nevertheless, an r,k-parametrized basis *can* be constructed utilizing displacement operators. In general, a displacement operator that displaces the expectation values of the position and momentum for any state by r and k simultaneously, can be defined as[7,8]

$$\hat{D}_s(r,k) = \exp[i(k\hat{x} - r\hat{p} - skr/2)], \tag{11}$$

where s is real number determining the phase such that $\hat{D}_1(r,k)=\hat{D}_q(r)\hat{D}_p(k)$, $\hat{D}_{-1}(r,k)=\hat{D}_p(k)\hat{D}_q(r)$, and $\hat{D}_0(r,k)$ is a symmetric combination. An r,k-parametrized state vector may then be defined as $|\Omega(r,k;s)\rangle\equiv(2\pi)^{-1/2}\hat{D}_s(r,k)|\chi\rangle$, where $|\chi\rangle$ is an arbitrary normalized state, and the set $\{|\Omega(r,k;s)\rangle|r,k\in\mathbf{R}\}$ becomes an overcomplete set of normalized vectors.[7] The set $\{|\Omega(r,k;s)\rangle|r,k\in\mathbf{R}\}$ can therefore be used as a basis and the relevant quantum system represented by the $\mathcal{L}^2(2)$ normalized wave function $\psi_\Omega(r,k;s)\equiv\langle\Omega(r,k;s)|\psi\rangle$, depending on the two real parameters k and r. These basis vectors obviously satisfy the displacement relation

$$|\Omega(r,k;s)\rangle=\hat{D}_s(r,k)|\Omega(0,0;s)\rangle. \tag{12}$$

Using that

$$i\frac{\partial}{\partial k}\hat{D}(r,k;s)=\left[\frac{1}{2}(1+s)r-\hat{x}\right]\hat{D}(r,k;s), \tag{13}$$

$$i\frac{\partial}{\partial r}\hat{D}(r,k;s)=-\left[\frac{1}{2}(1-s)k-\hat{p}\right]\hat{D}(r,k;s), \tag{14}$$

it is seen that in this representation, the fundamental operators $\hat{x}$ and $\hat{p}$ take the same form as in the relative-state formulation, given by Eqs. (5) and (6).

Therefore, the displacement-operator approach provides an alternative derivation from first principles to the results obtained within the relative-state formalism. Here, the state of the relevant system is projected onto an auxiliary (reference) state $|\chi\rangle$, displaced by r and k, whereas the auxiliary state $|\phi\rangle$ in the relative-state formulation is utilized to project the orthonormal basis in the extended Hilbert space onto a reduced Hilbert space. Thus, the auxiliary states play different physical roles, as can also be seen from the relations

$$\overline{r}\equiv\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk\, r|\psi_\Omega(r,k;s)|^2=x_\psi-x_\chi, \tag{15}$$

$$\overline{k}\equiv\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk\, k|\psi_\Omega(r,k;s)|^2=p_\psi-p_\chi. \tag{16}$$

Hence, r and k may here be interpreted as phase space coordinates, in the sense that the average values obtained using $|\psi_\Omega(r,k;s)|^2$ as a combined probability distribution equal the relative position and momentum, respectively, between the relevant and the auxiliary system. Hence, the displacement-operator approach provides a more symmetrical interpretation of the r,k-parametrized representation of the quantum state vector.

Since

$$\hat{D}_s(r,k)|\chi\rangle=e^{-i(1+s)kr/2}\int_{-\infty}^{\infty}dx\, e^{ikx}\chi(x-r)|x\rangle, \tag{17}$$

we see that the displacement-operator approach and the phase-space representation obtained within the relative-state formulation become formally identical if $\chi(x)=\phi^*(x)$; cf. Eq. (4), which implies that $p_\phi=-p_\chi$, as expected [compare Eqs. (8) and (16)].

In conclusion, we have shown that the two different mathematical approaches to a phase-space representation of quantum state vectors lead to identical expressions for the fundamental operators. However, usage of the well-known technique of displacement operators is in spirit closer to the construction of the usual position and momentum representations and, also, it provides a more transparent physical interpretation of the auxiliary state as a ‘‘probe’’ state in phase space.[8] With this interpretation, phase-space representation of quantum state vectors becomes a powerful tool and has been applied recently in the study of quantum dynamics directly in phase space[9] or as a route to semiclassical approximations.[10]

ACKNOWLEDGMENTS

The author thanks Professor Ahmed H. Zewail for his kind hospitality at Caltech and the Carlsberg Foundation of Denmark for financial support.

[1] M. Ban, J. Math. Phys. **39**, 1744 (1998).
[2] V. Fock, Z. Phys. **49**, 339 (1928).
[3] V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961).
[4] Go. Torres-Vega and J. H. Frederick, J. Chem. Phys. **98**, 3103 (1993).
[5] J. E. Harriman, J. Chem. Phys. **100**, 3651 (1994).
[6] P. A. M. Dirac, *The Principles of Quantum Mechanics*, 4th ed. (Oxford University Press, Oxford, 1958); W. H. Louisell, *Quantum Statistical Properties of Radiation* (Wiley, New York, 1973), Chap. 1.
[7] J. R. Klauder and B.-S. Skagerstam, *Coherent States* (World Scientific, Singapore, 1985).
[8] K. B. Møller, T. G. Jørgensen, and Go. Torres-Vega, J. Chem. Phys. **106**, 7228 (1997).
[9] Go. Torres-Vega, K. B. Møller, and A. Zúñiga-Segundo, Phys. Rev. A **57**, 771 (1998).
[10] S. Jang, M. Zhao, and S. A. Rice, Chem. Phys. **230**, 237 (1998).

Moyal–Nahm equations

L. M. Baker[a] and D. B. Fairlie[b]
Department of Mathematical Sciences, Science Laboratories, University of Durham, Durham DH1 3LE, England

(Received 28 January 1999; accepted for publication 10 March 1999)

Various aspects of the Nahm equations in three and seven dimensions are investigated. The residues of the variables at simple poles in the seven-dimensional case form an algebra. A large class of matrix representations of this algebra is constructed. The large N limit of these equations is taken by replacing the commutators by Moyal brackets, and a set of nontrivial solutions in a generalized form of Wigner distribution functions is obtained. © *1999 American Institute of Physics.* [S0022-2488(99)03706-8]

I. INTRODUCTION

When the study of self-dual gauge fields was very fashionable, there was some interest in extending the theory to higher dimensions.[1–3] In recent years interest in this subject has been revived[4] partly because of the occurrence of Yang–Mills gauge actions in the M(atrix) theory approximation to string theories.[5–7] The present article is a study of a class of solutions to the Nahm equations[8] in seven dimensions, which are a particular form of the self-dual Yang–Mills equations in Euclidean eight-dimensional space, where the gauge fields A^μ, $\mu=1,...,7$ depend only upon the eighth coordinate, τ, and we work in a gauge where $A^8=0$. These equations take the form

$$\begin{aligned}
&\frac{\partial A^1}{\partial\tau}-[A^2,A^7]-[A^6,A^3]-[A^5,A^4]=0,\\
&\frac{\partial A^2}{\partial\tau}-[A^7,A^1]-[A^5,A^3]-[A^4,A^6]=0,\\
&\frac{\partial A^3}{\partial\tau}-[A^1,A^6]-[A^2,A^5]-[A^4,A^7]=0,\\
&\frac{\partial A^4}{\partial\tau}-[A^1,A^5]-[A^6,A^2]-[A^7,A^3]=0,\\
&\frac{\partial A^5}{\partial\tau}-[A^4,A^1]-[A^3,A^2]-[A^6,A^7]=0,\\
&\frac{\partial A^6}{\partial\tau}-[A^3,A^1]-[A^2,A^4]-[A^7,A^5]=0,\\
&\frac{\partial A^7}{\partial\tau}-[A^1,A^2]-[A^3,A^4]-[A^5,A^6]=0.
\end{aligned} \tag{1}$$

[a]Electronic mail: l.m.baker@durham.ac.uk
[b]Electronic mail: david.fairlie@durham.ac.uk

The sum of the squares of the left-hand sides of these equations give the Lagrangian density for Yang–Mills theory in eight dimensions dependent only upon τ in the gauge $A^8=0$, up to divergence terms. This is just the Bogomol'nyi property of Eq. (1). Alternatively, the equations may be iterated to obtain the equations of motion for the Yang–Mills field. Note that when the τ dependence of A^μ is a simple pole, then the equations for the residues are just algebraic. The first objective of this paper is to initiate a study of the algebraic solution to such equations. It is known that a solution to (1) takes the form

$$A^i=\frac{-1}{2\tau}e_i, \quad i=1,...,7, \tag{2}$$

where the e_i form a basis for the imaginary octonions, but since these do not possess a matrix representation, as they are nonassociative, one might wonder whether a matrix valued solution exists. Indeed many such solutions exist, and we determine the allowable τ dependence of some of them. We find a particular solution, modeled upon the structure constants for octonionic multiplication and a more general solution in the form of a direct sum of SU(2) representations.

The second objective of this paper is to consider the infinite limit of the Nahm equations in both three and seven dimensions in the limit of large N for the gauge group SU(N). The motivation for this is a possible application to string theory and matrix models. In this case the matrices A^μ go over to functions $X^\mu(x,p)$ on phase space (x,p). The matrix elements may be regarded as the Fourier components of X^μ. The commutator goes over to the Moyal bracket.[9] This we recall is the antisymmetric part of the star product, which acts on functions in phase space (x,p). The star product of two functions $f(x,p)$ and $g(x,p)$ is defined as

$$f(x,p)\star g(x,p)=f(x,p)e^{i\lambda(\overleftarrow{\partial}_x\overrightarrow{\partial}_p-\overleftarrow{\partial}_p\overrightarrow{\partial}_x)}g(x,p), \tag{3}$$

where λ is a parameter. The Moyal bracket is proportional to the antisymmetric part of the star product and so the Moyal bracket of two functions $f(x,p)$ and $g(x,p)$ is written as

$$\{f,g\}_{\rm MB}=\frac{1}{2i}(f\star g-g\star f). \tag{4}$$

It is the unique one-parameter associative deformation of the Poisson bracket.[10–12] As $N=2\pi/\lambda$ passes through the odd integers the Moyal bracket $\{X^\mu,X^\nu\}_{\rm MB}$ degenerates into an infinite direct sum of copies of the commutator $[A^\mu,A^\nu]$, where A^μ, A^ν are SU(N) matrices, the large N limit of which is the Poisson bracket,

$$\{X^\mu,X^\nu\}_{\rm PB}=\frac{\partial X^\mu}{\partial x}\frac{\partial X^\nu}{\partial p}-\frac{\partial X^\mu}{\partial p}\frac{\partial X^\nu}{\partial x}. \tag{5}$$

The square of this quantity is just the Schild[13,14] form of the string Lagrangian, giving the same classical equations of motion as does the NambuGoto string. Thus the large N limit of the Yang–Mills Lagrangian in the strong coupling limit, where it is simply proportional to the square of the trace of the commutator, is equivalent to the string Lagrangian.[14] This may be viewed as a primitive form of a type of Maldacena conjecture,[15] relating a field theory in the large N limit to a string theory. The phase space coordinates x,p are to be interpreted as coordinate parametrizations on the world sheet of the string. The idea had been advanced that the target space coordinates X^μ (D_0 branes) may be represented by a generalized form of the Wigner function familiar from quantum mechanics.[16,17] We shall demonstrate solutions of the Moyal–Nahm equations (where the commutators are replaced by Moyal brackets) which take the form of a generalized Wigner function.

II. SEVEN-DIMENSIONAL NAHM EQUATIONS

Equation (1) can be written more succinctly with the aid of totally antiysmmetric structure constants C_{ijk} which, in fact, define the multiplication table of octonions of different index;

$$e_i \times e_j = C_{ijk} e_k \tag{6}$$

in the same way as the ϵ_{ijk} symbol does for the quaternions. The Nahm equations may then be written in the form

$$\frac{\partial A_i}{\partial \tau} = \frac{1}{2} C_{ijk} [A_j, A_k] \tag{7}$$

with a solution for the residues B_i of the simple poles in $A_i = -(1/\tau) B_i$ given by

$$B_i = C_{ijk}. \tag{8}$$

An explicit set of matrices B_i is listed in the Appendix. No more general τ-dependent solution of this type has been found. However, a very large class of solutions whose matrices take the form of a direct sum of representations of the SU(2) algebra, but which have a nontrivial τ dependence and are different for each B_i does exist and we proceed to explain their construction. Let

$$\begin{gathered}
B_1 = -i \begin{pmatrix} \sigma_3 & 0 & 0 \\ 0 & \sigma_3 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}, \quad B_2 = -i \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & b\sigma_2 & 0 \\ 0 & 0 & ic\sigma_3 \end{pmatrix}, \\
B_3 = -i \begin{pmatrix} a\sigma_3 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & ic\sigma_2 \end{pmatrix}, \quad B_4 = -i \begin{pmatrix} ia\sigma_2 & 0 & 0 \\ 0 & b\sigma_3 & 0 \\ 0 & 0 & \sigma_2 \end{pmatrix}, \\
B_5 = -i \begin{pmatrix} a\sigma_2 & 0 & 0 \\ 0 & -ib\sigma_3 & 0 \\ 0 & 0 & \sigma_1 \end{pmatrix}, \quad B_6 = -i \begin{pmatrix} ia\sigma_3 & 0 & 0 \\ 0 & \sigma_1 & 0 \\ 0 & 0 & c\sigma_2 \end{pmatrix}, \\
B_7 = -i \begin{pmatrix} \sigma_2 & 0 & 0 \\ 0 & ib\sigma_2 & 0 \\ 0 & 0 & c\sigma_3 \end{pmatrix}.
\end{gathered} \tag{9}$$

It is somewhat surprising to find that this solution involves three arbitrary parameters, yet none of these parameters can be set to zero in such a way that the solution remains faithful, i.e., each matrix is different. Obviously many such solutions of direct sum form can be constructed where the σ's in each row can be replaced by representations of SU(2) of arbitrary dimension.

In particular, a representation of the Nahm algebra with 4×4 matrices can be found, by simply omitting the last two rows in the above matrices (9). Our next task will be to search for the presence of solutions with nontrivial τ dependence. An attempt was made to find a solution to the Nahm equations using the matrices B_i (setting $a=1$, $b=1$, $c=1$). Each of these matrices was multiplied by a function of τ, $f_i(\tau)$. Therefore,

$$A_i = f_i(\tau) B_i, \quad i = 1,...,7. \tag{10}$$

However this ansatz is not sufficiently flexible and the previous result, $f_i(\tau) = -1/2\tau$ is all that is recovered. A more general ansatz for a solution is to multiply each matrix B_i by a diagonal matrix C_i given by

$$C_i=\begin{pmatrix} f_i(\tau)\mathbb{1} & 0 & 0 \\ 0 & g_i(\tau)\mathbb{1} & 0 \\ 0 & 0 & h_i(\tau)\mathbb{1} \end{pmatrix}. \tag{11}$$

(This amounts to multiplying each σ matrix entry in each matrix B_i by a different τ-dependent function.) It is easiest to consider each row of σ matrices separately. First, we shall look at the top row of sigma matrices and put in the τ dependence by multiplying each matrix by a function $f_i(\tau)$. Putting these 2×2 matrices into the Nahm equations gives the following set of differential equations:

$$\frac{\partial f_2}{\partial \tau}=2\, f_1 f_7+2\, f_5 f_3-2\, f_4 f_6,$$

$$\frac{\partial f_1}{\partial \tau}=2\, f_2 f_7, \quad \frac{\partial f_3}{\partial \tau}=2\, f_2 f_5, \quad \frac{\partial f_4}{\partial \tau}=2\, f_2 f_6, \tag{12}$$

$$\frac{\partial f_5}{\partial \tau}=2\, f_2 f_3, \quad \frac{\partial f_6}{\partial \tau}=2\, f_2 f_4, \quad \frac{\partial f_7}{\partial \tau}=2\, f_2 f_1$$

and the following constraints:

$$f_7 f_3=f_1 f_5, \quad f_6 f_7=f_1 f_4, \quad f_3 f_4=f_5 f_6. \tag{13}$$

Note that all of the differential equations involve f_2, but none of the constraints do. These can be solved in terms of elliptic functions. It was found that

$$f_6=K_1 f_3=K_1 M_1 f_1=\tfrac{1}{2} K_1 M_1 Q_1 \,\mathrm{sn}(q_1\tau+d_1),$$

$$f_4=K_1 f_5=K_1 M_1 f_7=\frac{-i}{2} K_1 M_1 Q_1 \,\mathrm{cn}(q_1\tau+d_1), \tag{14}$$

$$f_2=\frac{i}{2} q_1 \,\mathrm{dn}(q_1\tau+d_1),$$

where cn, sn, dn are elliptic functions and K_1, M_1, q_1, Q_1, d_1 are all constants. The elliptic functions are related to each other by a parameter k_1 as follows:

$$\mathrm{sn}^2(x)+\mathrm{cn}^2(x)=1, \quad \mathrm{dn}^2(x)+k_1^2\,\mathrm{sn}^2(x)=1, \tag{15}$$

where $k_1=(Q_1/q_1)\sqrt{1+M_1^2(1-k_1^2)}$.

The following set of matrices solve Nahm's equations:

$$A_1=-i\begin{pmatrix} f_1\sigma_3 & 0 & 0 \\ 0 & g_1\sigma_3 & 0 \\ 0 & 0 & h_1\sigma_3 \end{pmatrix}, \quad A_2=-i\begin{pmatrix} f_2\sigma_1 & 0 & 0 \\ 0 & g_2\sigma_2 & 0 \\ 0 & 0 & ih_2\dot{\sigma}_3 \end{pmatrix},$$

$$A_3=-i\begin{pmatrix} f_3\sigma_3 & 0 & 0 \\ 0 & g_3\sigma_2 & 0 \\ 0 & 0 & ih_3\sigma_2 \end{pmatrix}, \quad A_4=-i\begin{pmatrix} if_4\sigma_2 & 0 & 0 \\ 0 & g_4\sigma_3 & 0 \\ 0 & 0 & h_4\sigma_2 \end{pmatrix}, \tag{16}$$

$$A_5=-i\begin{pmatrix} f_5\sigma_2 & 0 & 0 \\ 0 & -ig_5\sigma_3 & 0 \\ 0 & 0 & h_5\sigma_1 \end{pmatrix}, \quad A_6=-i\begin{pmatrix} if_6\sigma_3 & 0 & 0 \\ 0 & g_6\sigma_1 & 0 \\ 0 & 0 & h_6\sigma_2 \end{pmatrix},$$

$$A_7=-i\begin{pmatrix} f_7\sigma_2 & 0 & 0 \\ 0 & ig_7\sigma_2 & 0 \\ 0 & 0 & h_7\sigma_3 \end{pmatrix},$$

where the other τ-dependent functions are given by

$$g_5=K_2g_4=K_2M_2g_1=\tfrac{1}{2}K_2M_2Q_2\,\mathrm{sn}(q_2\tau+d_2),$$

$$g_7=K_2g_2=K_2M_2g_3=\frac{-i}{2}K_2M_2Q_2\,\mathrm{cn}(q_2\tau+d_2), \tag{17}$$

$$g_6=\frac{i}{2}q_2\,\mathrm{dn}(q_2\tau+d_2),$$

$$h_6=K_3h_3=K_3M_3h_1=\tfrac{1}{2}K_3M_3Q_3\,\mathrm{sn}(q_3\tau+d_3),$$

$$h_4=K_3h_5=K_3M_3h_7=\frac{-i}{2}K_3M_3Q_3\,\mathrm{cn}(q_3\tau+d_3), \tag{18}$$

$$h_2=\frac{i}{2}q_3\,\mathrm{dn}(q_3\tau+d_3).$$

III. MOYAL–NAHM EQUATIONS

Consider a field X^k $(k=0,1,2,3)$ in four dimensions where X^k depends upon only one coordinate (in this case t) and phase space (x,p). The gauge is fixed so X^0 is a constant. The Moyal–Nahm equations in three dimensions are:

$$\frac{\partial X^1}{\partial t}=\{X^2,X^3\}_{\mathrm{MB}}, \quad \frac{\partial X^2}{\partial t}=\{X^3,X^1\}_{\mathrm{MB}}, \quad \frac{\partial X^3}{\partial t}=\{X^1,X^2\}_{\mathrm{MB}}. \tag{19}$$

If the Moyal brackets were to be replaced by commutators and the functions $X^k(t,x,p)$ were replaced by matrices $X^k(t)$ then the equations would become the Nahm equations for a self-dual field. The main idea to solve these Moyal–Nahm equations is to use the following ansatz:

$$X^i=i\int_{-\infty}^{\infty}\psi_j^{\dagger}(x-y,t)\,\epsilon^{ijk}\psi_k(x+y,t)\,e^{2\pi ipy/\lambda}\,\mathrm{d}y, \tag{20}$$

where $\psi(x,t)$ are three component wave functions. These wave functions were chosen to be of the following form:

$$\psi(x,t)=\begin{pmatrix}\psi_1(x,t)\\ \psi_2(x,t)\\ \psi_3(x,t)\end{pmatrix}=\begin{pmatrix}f_1(t)\phi_1(x)\\ f_2(t)\phi_2(x)\\ f_3(t)\phi_3(x)\end{pmatrix}, \tag{21}$$

where the $\phi_i(x)$ are orthonormal wave functions. The star product of X^j and X^k is calculated as follows:

$$
\begin{aligned}
X^j\star X^k &= -\int\int \psi_i^\dagger(x-y,t)\epsilon^{jil}\psi_l(x+y,t)e^{2\pi ipy/\lambda}\star\psi_m^\dagger(x-y',t)\\
&\quad\times\epsilon^{kmn}\psi_n(x+y',t)e^{2\pi ipy'/\lambda}\,\mathrm{d}y\,\mathrm{d}y'\\
&= -\int\int \psi_i^\dagger(x-y+y',t)\epsilon^{jil}\psi_l(x+y+y',t)\\
&\quad\times e^{2\pi ipy/\lambda}\psi_m^\dagger(x-y'-y,t)\epsilon^{kmn}\psi_n(x+y'-y,t)e^{2\pi ipy'/\lambda}\,\mathrm{d}y\,\mathrm{d}y'\\
&= -\frac{1}{2}\int \psi_m^\dagger(x-y,t)\epsilon^{kmn}Z(t)\epsilon^{jil}\psi_l(x+y,t)e^{2\pi ipy/\lambda}\,\mathrm{d}y\\
&= -\frac{1}{2}\int \phi_m^\dagger(x-y)f_{ms}^\dagger(t)\epsilon^{ksn}Z(t)\epsilon^{jir}f_{rl}(t)\phi_l(x+y)e^{2\pi ipy/\lambda}\,\mathrm{d}y,
\end{aligned}
\tag{22}
$$

where orthogonality of the $\phi_k(x)$ is assumed to be of the form

$$
\int_{-\infty}^{\infty}\phi_j^\dagger(x)\phi_k(x)\mathrm{d}x=\delta_{jk} \tag{23}
$$

and

$$
Z(t)=f^\dagger f,
$$

$$
\text{where } f=\begin{pmatrix} f_1(t) & 0 & 0\\ 0 & f_2(t) & 0\\ 0 & 0 & f_3(t)\end{pmatrix}. \tag{24}
$$

The partial derivative $\partial X^i/\partial t$ can be written as

$$
\frac{\partial X^i}{\partial t}=i\int \phi_j^\dagger(x-y)\frac{\partial}{\partial t}(f^\dagger(t)\epsilon^{ijk}f(t))\phi_k(x+y)e^{2\pi ipy/\lambda}\,\mathrm{d}y. \tag{25}
$$

By putting these into the Moyal–Nahm equations one obtains three matrix equations of the form

$$
i\frac{\partial}{\partial t}(f^\dagger(t)\epsilon^1 f(t))=\frac{-1}{4i}(f^\dagger(t)\epsilon^3 Z(t)\epsilon^2 f(t)-f^\dagger(t)\epsilon^2 Z(t)\epsilon^3 f(t)), \tag{26}
$$

where ϵ^i is a 3×3 matrix with jkth entry ϵ^{ijk}. Equating the entries in the matrices gives differential equations of the form

$$
\frac{\partial}{\partial t}(f_2^* f_3)=-\frac{1}{4}|f_1|^2(f_2^* f_3),\quad \frac{\partial}{\partial t}(f_3^* f_2)=-\frac{1}{4}|f_1|^2(f_3^* f_2) \tag{27}
$$

and cycle combinations of these. These can be used to create the following set of three differential equations:

$$\frac{\partial}{\partial t}(|f_2|^2|f_3|^2) = -\frac{1}{2}|f_1|^2|f_2|^2|f_3|^2,$$

$$\frac{\partial}{\partial t}(|f_3|^2|f_1|^2) = -\frac{1}{2}|f_1|^2|f_2|^2|f_3|^2, \tag{28}$$

$$\frac{\partial}{\partial t}(|f_1|^2|f_2|^2) = -\frac{1}{2}|f_1|^2|f_2|^2|f_3|^2.$$

Note that for each of the above, the right-hand side of the equations is always the same.

A. Simplest solution

The simplest solution is to set all the f_i equal to each other. This gives the solution

$$|f_1|^2 = |f_2|^2 = |f_3|^2 = \frac{4}{t+K}, \tag{29}$$

so that

$$f_1(t) = f_2(t) = f_3(t) = \frac{2}{\sqrt{t+K}}, \tag{30}$$

where K is an arbitrary constant.

B. Another simple solution

By setting two of the f_i equal to each other then a solution in terms of the hyperbolic functions can be found.

$$|f_1|^2 = |f_2|^2 = 4q \coth(qt+K), \quad |f_3|^2 = 8q \operatorname{csch}(2qt+2K), \tag{31}$$

so that

$$f_1(t) = f_2(t) = 2\sqrt{q \coth(qt+K)}, \quad f_3(t) = 2\sqrt{2q \operatorname{csch}(2qt+2K)}, \tag{32}$$

where K and q are both real constants.

C. General solution

However, ideally we want a general solution to these equations. In this case the solutions are written in terms of elliptic functions sn, cn, and dn. The most general solution was found to be:

$$|f_1|^2 = 4qk \operatorname{sn}(qt+c),$$

$$|f_2|^2 = 2qk \operatorname{sn}(qt+c) + \frac{2q(\operatorname{dn}(qt+c)\operatorname{cn}(qt+c)+1)}{\operatorname{sn}(qt+c)}, \tag{33}$$

$$|f_3|^2 = 2qk \operatorname{sn}(qt+c) + \frac{2q(\operatorname{dn}(qt+c)\operatorname{cn}(qt+c)-1)}{\operatorname{sn}(qt+c)}.$$

k, q, and c are all constants but may have to be carefully chosen in order to ensure that all the $|f_i|^2$ are positive. k depends on the elliptic functions.

A more aesthetically pleasing form of solution is as follows:

$$|f_1|^2 = 4qk^2 \frac{\mathrm{sn}(qt+c)\mathrm{cn}(qt+c)}{\mathrm{dn}(qt+c)},$$

$$|f_2|^2 = -4q \frac{\mathrm{cn}(qt+c)\mathrm{dn}(qt+c)}{\mathrm{sn}(qt+c)}, \tag{34}$$

$$|f_3|^2 = 4q \frac{\mathrm{dn}(qt+c)\mathrm{sn}(qt+c)}{\mathrm{cn}(qt+c)}.$$

q, c, and k are all constants. Again, k depends on the elliptic functions.

IV. SEVEN-DIMENSIONAL MOYAL EQUATIONS AGAIN

The Moyal–Nahm equations in seven dimensions are:

$$\frac{\partial X^1}{\partial t} = \{X^2, X^7\}_{\mathrm{MB}} + \{X^6, X^3\}_{\mathrm{MB}} + \{X^5, X^4\}_{\mathrm{MB}},$$

$$\frac{\partial X^2}{\partial t} = \{X^7, X^1\}_{\mathrm{MB}} + \{X^5, X^3\}_{\mathrm{MB}} + \{X^4, X^6\}_{\mathrm{MB}},$$

$$\frac{\partial X^3}{\partial t} = \{X^1, X^6\}_{\mathrm{MB}} + \{X^2, X^5\}_{\mathrm{MB}} + \{X^4, X^7\}_{\mathrm{MB}},$$

$$\frac{\partial X^4}{\partial t} = \{X^1, X^5\}_{\mathrm{MB}} + \{X^6, X^2\}_{\mathrm{MB}} + \{X^7, X^3\}_{\mathrm{MB}}, \tag{35}$$

$$\frac{\partial X^5}{\partial t} = \{X^4, X^1\}_{\mathrm{MB}} + \{X^3, X^2\}_{\mathrm{MB}} + \{X^6, X^7\}_{\mathrm{MB}},$$

$$\frac{\partial X^6}{\partial t} = \{X^3, X^1\}_{\mathrm{MB}} + \{X^2, X^4\}_{\mathrm{MB}} + \{X^7, X^5\}_{\mathrm{MB}},$$

$$\frac{\partial X^7}{\partial t} = \{X^1, X^2\}_{\mathrm{MB}} + \{X^3, X^4\}_{\mathrm{MB}} + \{X^5, X^6\}_{\mathrm{MB}}.$$

The set of matrices B_i can also be used to find a solution to Eq. (35). Using the ansatz

$$A_i = i \int_{-\infty}^{\infty} \psi^{\dagger}(x-y,\tau) B_i \psi(x+y,\tau) e^{2\pi i p y/\lambda}\, \mathrm{d}y, \tag{36}$$

where $\psi(x,t)$ are six component wave functions of the form

$$\psi_j = f_j(\tau)\phi(x)_j, \quad j \ \text{not summed} \tag{37}$$

we find a rather simple solution in terms of this ansatz of the form

$$\{f_1, f_2, f_3, f_4, f_5, f_6\} = \left\{ \frac{2\sqrt{K_1} e^{i\theta_1}}{\sqrt{1-e^{-K_1\tau}}}, \frac{2\sqrt{K_1} e^{i\theta_2}}{\sqrt{e^{K_1\tau}-1}}, \frac{2\sqrt{K_2} e^{i\theta_3}}{\sqrt{1-e^{-K_2\tau}}}, \frac{2\sqrt{K_2} e^{i\theta_4}}{\sqrt{e^{K_2\tau}-1}}, \frac{2\sqrt{K_3} e^{i\theta_5}}{\sqrt{1-e^{-K_3\tau}}}, \frac{2\sqrt{K_3} e^{i\theta_6}}{\sqrt{e^{K_3\tau}-1}} \right\}.$$

All K_i and θ_j are constants.

The previous solution used the two-dimensional σ matrix representation of SU(2). If instead the three-dimensional representation of SU(2), which involves the completely antisymmetric matrices ϵ_{ijk}, is used to construct the B_i a more general τ-dependent solution can be found along similar lines to that in Sec. III. This time the matrices B_i were taken to be

$$B_1=-\begin{pmatrix}\epsilon_3 & 0\\ 0 & \epsilon_3\end{pmatrix},\quad B_2=-\begin{pmatrix}\epsilon_1 & 0\\ 0 & \epsilon_2\end{pmatrix},\quad B_3=-\begin{pmatrix}\epsilon_3 & 0\\ 0 & \epsilon_2\end{pmatrix},\quad B_4=-\begin{pmatrix}i\epsilon_2 & 0\\ 0 & \epsilon_3\end{pmatrix},$$
$$B_5=-\begin{pmatrix}\epsilon_2 & 0\\ 0 & i\epsilon_3\end{pmatrix},\quad B_6=-\begin{pmatrix}i\epsilon_3 & 0\\ 0 & \epsilon_1\end{pmatrix},\quad B_7=-\begin{pmatrix}\epsilon_2 & 0\\ 0 & i\epsilon_2\end{pmatrix}, \tag{38}$$

where the jkth entry of the matrix ϵ_i is given by the totally antisymmetric tensor ϵ_{ijk}.

The same ansatz (36), (37) was used as before but with the new B_i matrices. The τ-dependent functions f_i were found to be of the same form as the solution to the three-dimensional Moyal–Nahm equations when solved using ϵ matrices (34). They are as follows:

$$|f_1|^2=4qk^2\frac{\mathrm{sn}(q\tau+c)\mathrm{cn}(q\tau+c)}{\mathrm{dn}(q\tau+c)},\quad |f_4|^2=4QK^2\frac{\mathrm{sn}(Q\tau+b)\mathrm{cn}(Q\tau+b)}{\mathrm{dn}(Q\tau+b)},$$
$$|f_2|^2=-4q\,\frac{\mathrm{cn}(q\tau+c)\mathrm{dn}(q\tau+c)}{\mathrm{sn}(q\tau+c)},\quad |f_5|^2=-4Q\,\frac{\mathrm{cn}(Q\tau+b)\mathrm{dn}(Q\tau+b)}{\mathrm{sn}(Q\tau+b)}, \tag{39}$$
$$|f_3|^2=4q\,\frac{\mathrm{dn}(q\tau+c)\mathrm{sn}(q\tau+c)}{\mathrm{cn}(q\tau+c)},\quad |f_6|^2=4Q\,\frac{\mathrm{dn}(Q\tau+b)\mathrm{sn}(Q\tau+b)}{\mathrm{cn}(Q\tau+b)}.$$

These solutions can be extended for direct sums of more than two ϵ_i matrices.

ACKNOWLEDGMENTS

L.M.B. is grateful to EPSRC for a postgraduate research award. D.B.F. wishes to thank Tatsuya Ueno for discussions.

APPENDIX: MATRIX REPRESENTATION OF STRUCTURE CONSTANTS

The following 7×7 matrices which solve the Nahm algebra are created using the octonionic structure constants c_{ijk} which are taken to be:

$$c_{127}=c_{631}=c_{541}=c_{532}=c_{246}=c_{347}=c_{567}=1. \tag{A1}$$

These are totally antisymmetric. All other c_{ijk} are zero. The jkth entry of the matrix m_i is given by $[m_i]_{jk}=c_{ijk}$,

$$m_1=\begin{pmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&1\\0&0&0&0&0&-1&0\\0&0&0&0&-1&0&0\\0&0&0&1&0&0&0\\0&0&1&0&0&0&0\\0&-1&0&0&0&0&0\end{pmatrix},\quad m_2=\begin{pmatrix}0&0&0&0&0&0&-1\\0&0&0&0&0&0&0\\0&0&0&0&-1&0&0\\0&0&0&0&0&1&0\\0&0&1&0&0&0&0\\0&0&0&-1&0&0&0\\1&0&0&0&0&0&0\end{pmatrix},$$

$$m_3=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1\\ 0 & -1 & 0 & 0 & 0 & 0 & 0\\ -1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix}, \quad m_4=\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ -1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$m_5=\begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 0\\ 0 & 0 & -1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \quad m_6=\begin{pmatrix} 0 & 0 & -1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & -1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

$$m_7=\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0\\ -1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & -1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & -1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

[1] E. Corrigan, C. Devchand, D. Fairlie, and J. Nuyts, Nucl. Phys. B **214**, 452 (1982).
[2] R. Dundarer, F. Gursey, and C. H. Tze, J. Math. Phys. **25**, 1496 (1984).
[3] R. S. Ward, Nucl. Phys. B **236**, 381 (1984).
[4] L. Baulieu, H. Kanno, and I. M. Singer, Commun. Math. Phys. **194**, 149 (1988); B. S. Acharya, M. O'Laughlin, and B. Spence, Nucl. Phys. B **503**, 657 (1997).
[5] M. Claudson and M. B. Halpern, Nucl. Phys. B **250**, 689 (1985).
[6] T. Banks, W. Fischler, S. Shenker, and L. Susskind, Phys. Rev. D **55**, 5112 (1997).
[7] R. Dijkgraaf, E. Verlinde, and H. Verlinde, Nucl. Phys. B **500**, 43 (1997).
[8] W. Nahm, *Monopoles in Quantum Field Theory*, edited by N. S. Craigie, P. Goddard, and W. Nahm (World Scientific, Singapore, 1982), pp. 87–94.
[9] J. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).
[10] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, Ann. Phys. (N.Y.) **111**, 61 (1978).
[11] W. Arveson, Commun. Math. Phys. **89**, 77 (1983).
[12] P. Fletcher, Phys. Lett. A **248**, 323 (1990).
[13] A. Schild, Phys. Rev. D **16**, 1722 (1977); T. Eguchi, Phys. Rev. Lett. **44**, 126 (1980).
[14] D. B. Fairlie, P. Fletcher, and C. K. Zachos, J. Math. Phys. **31**, 1088 (1990).
[15] J. Maldacena, Adv. Theor. Math. Phys. **2**, 231 (1998).
[16] T. Curtright, D. B. Fairlie, and C. K. Zachos, Phys. Lett. B **405**, 37 (1997).
[17] D. B. Fairlie, Mod. Phys. Lett. A **13**, 263 (1998); D. B. Fairlie, Chaos Solitons Fractals **10**, 365 (1999).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

A rigorous path integral for quantum spin using flat-space Wiener regularization

Bernhard Bodmann
Department of Mathematics, University of Florida, 358 Little Hall, Gainesville, Florida 32611

Hajo Leschke and Simone Warzel[a)]
Institut für Theoretische Physik, Universität Erlangen-Nürnberg, Staudtstr. 7, D-91058 Erlangen, Germany

(Received 20 November 1998; accepted for publication 8 February 1999)

Adapting ideas of Daubechies and Klauder [J. Math. Phys. **26**, 2239 (1985)] we derive a rigorous continuum path-integral formula for the semigroup generated by a spin Hamiltonian. More precisely, we use spin coherent vectors parametrized by complex numbers to relate the coherent representation of this semigroup to a suitable Schrödinger semigroup on the Hilbert space $L^2(\mathbb{R}^2)$ of Lebesgue square-integrable functions on the Euclidean plane $\mathbb{R}^2$. The path-integral formula emerges from the standard Feynman–Kac–Itô formula for the Schrödinger semigroup in the ultradiffusive limit of the underlying Brownian bridge on $\mathbb{R}^2$. In a similar vein, a path-integral formula can be constructed for the coherent representation of the unitary time evolution generated by the spin Hamiltonian. © *1999 American Institute of Physics.* [S0022-2488(99)02005-8]

I. INTRODUCTION

Even 50 years after the appearance of Feynman's celebrated paper[1] that introduced the path-integral formalism[2–6] into quantum theory in a heuristic but convincing manner, there is no general consensus on how to treat a quantum spin within this framework. To the best of our knowledge, among the various approaches over the years, see, for example, Refs. 7–22, the only rigorous expression for the dynamics of a quantum spin in terms of an integral over continuous paths is due to Daubechies and Klauder.[12] These authors were able to write the coherent representation of the unitary time-evolution operator of a spin with a definite quantum number as a Wiener-regularized path integral, more precisely, as the ultradiffusive limit of a well-defined integral over spherical Brownian-motion paths.

The main goal of the present paper is to show that one may equally well perform the Wiener regularization by employing planar Brownian motion. In this way also a closer contact to symbolic continuum path-integral formulas widely discussed in the recent literature[23–28] is established. One may hope that the wealth of analytical tools associated with the flat-space Wiener measure helps clarifying some subtle points there.

II. BASIC DEFINITIONS, RESULT, AND COMMENTS

We consider a single spin with fixed *quantum number* $j\in\{0,1/2,1,3/2,...\}$, that is, using physical units where Planck's constant $2\pi\hbar$ equals 2π,

$$\tfrac{1}{2}(\mathcal{J}_+\mathcal{J}_-+\mathcal{J}_-\mathcal{J}_+)+\mathcal{J}_3^2=j(j+1)\mathbf{1}. \tag{1}$$

The *spin operators* $\mathcal{J}_+$, $\mathcal{J}_-$, and $\mathcal{J}_3$ obey the usual angular-momentum commutation relations $\mathcal{J}_+\mathcal{J}_--\mathcal{J}_-\mathcal{J}_+=2\mathcal{J}_3$, $\mathcal{J}_3\mathcal{J}_\pm-\mathcal{J}_\pm\mathcal{J}_3=\pm\mathcal{J}_\pm$ and are viewed as acting on the $(2j+1)$-

a)Electronic mail: simone@theorie1.physik.uni-erlangen.de

TABLE I. Contravariant symbols for selected operators on $\mathbb{C}^{2j+1}$, which are bounded and continuous.

Operator	Contravariant symbol	Operator	Contravariant symbol
$\mathcal{J}_+$	$2(j+1)\dfrac{z^*}{1+\lvert z\rvert^2}$	$\mathcal{J}_+\mathcal{J}_-$	$-2(j+1)\dfrac{1-2(j+1)\lvert z\rvert^2}{(1+\lvert z\rvert^2)^2}$
$\mathcal{J}_-$	$2(j+1)\dfrac{z}{1+\lvert z\rvert^2}$	$\mathcal{J}_-\mathcal{J}_+$	$2(j+1)\dfrac{2(j+1)\lvert z\rvert^2-\lvert z\rvert^4}{(1+\lvert z\rvert^2)^2}$
$\mathcal{J}_3$	$-(j+1)\dfrac{1-\lvert z\rvert^2}{1+\lvert z\rvert^2}$	$\mathcal{J}_3^2$	$(j+1)(j+\frac{3}{2})\left(\dfrac{1-\lvert z\rvert^2}{1+\lvert z\rvert^2}\right)^2-\dfrac{j+1}{2}$

dimensional complex Hilbert space $\mathbb{C}^{2j+1}$. Its standard scalar product is denoted as $\langle\cdot|\cdot\rangle$ and, by convention, antilinear in the first argument. The unit operator on $\mathbb{C}^{2j+1}$ is denoted by **1**.

Non-normalized so-called *coherent vectors*[9,29] in this Hilbert space,

$$|z\rangle := g(z)e^{z\mathcal{J}_+}|j,-j\rangle, \quad z\in\mathbb{C}, \tag{2}$$

are parametrized by complex numbers z. Henceforth, z^* will refer to their complex conjugates, $z_1:=(z+z^*)/2$ and $z_2:=(z-z^*)/2i$ to their real and imaginary parts, and we write $f^*(z) := (f(z))^*$ for the values of complex-conjugated functions f^*. For later notational convenience the strictly positive prefactor is taken as

$$g(z) := \left(\frac{2j+1}{\pi}\right)^{1/2}(1+|z|^2)^{-j-1}, \tag{3}$$

and a normalized *spin-down vector* $|j,-j\rangle\in\mathbb{C}^{2j+1}$, obeying $\mathcal{J}_-|j,-j\rangle=0$ and $\langle j,-j|j,-j\rangle=1$, serves as the reference vector. Every vector $|\psi\rangle\in\mathbb{C}^{2j+1}$ is characterized by its so-called *coherent representation* $\langle z|\psi\rangle$, a function of the form $g(z)$ times a polynomial in z^* of maximal degree $2j$. The scalar product of two coherent vectors $\langle z|z'\rangle = g(z)g(z')(1+z^*z')^{2j}$ is an example. Given an arbitrary operator $\mathcal{B}$ on $\mathbb{C}^{2j+1}$, the scalar product $\langle z|\mathcal{B}|z'\rangle$ of $|z\rangle$ and $\mathcal{B}|z'\rangle$ is called the *coherent representation* of $\mathcal{B}$. The mapping $(z,z')\mapsto\langle z|\mathcal{B}|z'\rangle$ is continuous, because $z\mapsto|z\rangle$ is continuous, every operator $\mathcal{B}$ on $\mathbb{C}^{2j+1}$ is bounded, and the scalar product $(|\varphi\rangle,|\psi\rangle)\mapsto\langle\varphi|\psi\rangle$ is continuous. An example is $\langle z|e^{2\lambda\mathcal{J}_3}|z'\rangle = g(z)g(z')(e^{-\lambda}+z^*z'e^{\lambda})^{2j}$, $\lambda\in\mathbb{C}$.

In what follows, it is a comforting fact that whatever the *spin Hamiltonian* $\mathcal{H}$ may be—given as a (self-adjoint) operator on $\mathbb{C}^{2j+1}$—it is polynomial in the spin operators $\mathcal{J}_+$, $\mathcal{J}_-$, and $\mathcal{J}_3$, and it is always possible to write it in *pseudodiagonal form*,

$$\mathcal{H} = \int_{\mathbb{C}} d^2z\, h(z)|z\rangle\langle z|. \tag{4}$$

Here the (real-valued) function h on $\mathbb{C}\cong\mathbb{R}\times\mathbb{R}$ may be chosen bounded and continuous,[30–32] the operator $|z\rangle\langle z|/\langle z|z\rangle$ denotes the orthogonal projection onto the one-dimensional subspace spanned by $|z\rangle\in\mathbb{C}^{2j+1}$, and $d^2z := dz_1\,dz_2$ is the two-dimensional Lebesgue measure on the Euclidean plane $\mathbb{R}\times\mathbb{R}=:\mathbb{R}^2$. Following Ref. 33, we call h a *contravariant symbol* of $\mathcal{H}$, elsewhere called an upper[34] or lower[35] symbol. In particular, the unit operator **1** has the constant 1 as a contravariant symbol. In this sense, the coherent vectors are *unity-resolving* and hence (over-)complete. Other examples for contravariant symbols are listed in Table I; confer Ref. 36.

After these preparations we are able to state the main result of the present paper, namely, a rigorous expression for the *spin semigroup* $\{e^{-t\mathcal{H}}\}_{t\geqslant 0}$ as the *ultradiffusive limit* of a Wiener type of integral over Brownian-motion paths $\{s \mapsto b(s) = b_1(s) + i b_2(s)\}_{s \geq 0}$ on the complex plane $\mathbb{C} \cong \mathbb{R} \times \mathbb{R}$. More precisely, the coherent representation of $e^{-t\mathcal{H}}$ may, for all $z, z' \in \mathbb{C}$ and $t > 0$, be written as

$$\langle z | e^{-t\mathcal{H}} | z' \rangle = \lim_{\nu \to \infty} \int d\mu^{(\nu)}_{z,0;z',t}(b) \exp\left\{ 4(j+1)\nu \int_0^t \frac{ds}{(1+|b(s)|^2)^2} \right\} \times \exp\left\{ (j+1) \int_0^t ds \, \frac{\dot{b}(s) b^*(s) - \dot{b}^*(s) b(s)}{1+|b(s)|^2} - \int_0^t ds \, h(b(s)) \right\}. \tag{5}$$

Here for given z, $z' \in \mathbb{C}$, $t > 0$, and $\nu > 0$ the path integration is defined by

$$\int d\mu^{(\nu)}_{z,0;z',t}(b)(\cdot) := \frac{1}{4\pi t \nu} e^{-|z-z'|^2/4t\nu} \mathbb{E}(\cdot), \tag{6}$$

where $\mathbb{E}(\cdot)$ indicates the probabilistic expectation with respect to the *two-dimensional Brownian bridge*, with diffusion constant ν starting in $z = b(0)$ and arriving at $z' = b(t)$ a time t later.[3,6,37–39] As a Gaussian stochastic process with continuous paths on $\mathbb{C} \cong \mathbb{R} \times \mathbb{R}$ the Brownian bridge, in its turn, is uniquely determined by its mean,

$$\mathbb{E}(b(s)) = z + (z' - z) \frac{s}{t}, \quad s \in [0, t], \tag{7}$$

and covariances,

$$\mathbb{E}(b^*(r) b(s)) - \mathbb{E}(b^*(r)) \mathbb{E}(b(s)) = 4\nu \left(\min\{r, s\} - \frac{rs}{t} \right), \tag{8}$$

$$\mathbb{E}(b(r) b(s)) - \mathbb{E}(b(r)) \mathbb{E}(b(s)) = 0, \quad r, s \in [0, t]. \tag{9}$$

The second integral in the exponent on the right-hand side of (5) is a purely imaginary stochastic (line) integral,[37–39] which is understood in the sense of Fisk and Stratonovich and to which one is therefore allowed to apply the rules of ordinary calculus,[40] although the time derivative $\dot{b}$ does not exist.

Several comments apply.

(i) By the Itô formula[3,37–39] it can be seen that the stochastic integral in (5) may equally well be interpreted as a stochastic integral in the sense of Itô. Moreover, using the Itô formula in a different way, the sum of this integral and the first (Lebesgue) integral in the exponent of the right-hand side of (5) can be converted[41] according to

$$4\nu \int_0^t \frac{ds}{(1+|b(s)|^2)^2} + \int_0^t ds \, \frac{\dot{b}(s) b^*(s) - \dot{b}^*(s) b(s)}{1+|b(s)|^2} = \ln\left(\frac{1+|b(t)|^2}{1+|b(0)|^2} \right) - 2 \int_0^t \frac{db^*(s) b(s)}{1+|b(s)|^2}. \tag{10}$$

Here the complex stochastic integral $\int_0^t db^*(s) b(s) / [1+|b(s)|^2]$ has to be understood in the sense of Itô. It contains the only ν-dependence of the right-hand side. By using (10) in the path integrand in (5), the logarithmic term results in the prefactor $[(1+|z'|^2)/(1+|z|^2)]^{j+1} = g(z)/g(z')$.

(ii) The stochastic integral in (5) is of kinematical origin and reflects the symplectic structure, which renders the complex plane a phase space for the so-called classical spin;[42,43,31] also see the concluding remarks.

(iii) If one wants to use (5) to express the trace $\int_{\mathbb{C}} d^2 z \langle z | e^{-t\mathcal{H}} | z \rangle$ of $e^{-t\mathcal{H}}$ as a path integral, one should resist the temptation to interchange the integration with respect to z with the ultradiffusive limit $\nu \to \infty$, because the resulting prelimit expression would be infinite.

(iv) Instead of taking the ultradiffusive limit, one may perform the regularization also by a *long-time limit*, in the sense that

$$\langle z|e^{-t\mathcal{H}}|z'\rangle=\lim_{u\to\infty}\int d\mu^{(\nu)}_{z,0;z',u}(b)\exp\left\{4(j+1)\nu\int_0^u\frac{ds}{(1+|b(s)|^2)^2}\right\}$$
$$\times\exp\left\{(j+1)\int_0^u ds\,\frac{\dot b(s)b^*(s)-\dot b^*(s)b(s)}{1+|b(s)|^2}-\frac{t}{u}\int_0^u ds\,h(b(s))\right\}. \qquad (11)$$

This formula can be deduced from (5) by suitably scaling the Brownian bridge, holds for all $\nu>0$, and, in contrast to (5), makes sense as it stands even for $t\leqslant 0$, hence for all $t\in\mathbb{R}$. One should notice that the time-parameter set of the Brownian bridge used in (11) is the closed interval $[0,u]$ and not $[0,t]$.

(v) Replacing h by ih in (5) or (11) yields analogous expressions for the coherent representation of the (unitary) *spin time-evolution* operator $e^{-it\mathcal{H}}$. A rigorous justification relies on the boundedness and continuity of h and requires extending the subsequent proof by showing analyticity of both sides of (5) in a coupling parameter $\lambda\in\mathbb{C}$ multiplying h. The left-hand side and the prelimit expression in (5) are easily seen to be analytic in λ. Analyticity in λ in the limit $\nu\to\infty$ is then proved with the help of an equation analogous to (29) and uniform convergence in $\nu>2\nu_0>0$ of the perturbation series in λ of the relevant operator and functions there.

(vi) The flat-space Wiener-regularized path-integral expression (5) for the spin semigroup is an alternative to a result first given and proved in Ref. 12. There the authors integrate over Brownian-motion paths on the unit-sphere in the three-dimensional Euclidean space $\mathbb{R}^3$ to obtain the coherent representation of $e^{-it\mathcal{H}}$. Unlike in Ref. 12, the regularizing path measure $d\mu^{(\nu)}_{z,0;z',t}(b)\exp\{4(j+1)\nu\int_0^t ds(1+|b(s)|^2)^{-2}\}$ used in (5) is not invariant under the full special unitary group SU(2) when the latter is realized by suitable Möbius transformations on the (extended) complex plane. Yet in the limit $\nu\to\infty$ all symmetries of a given spin Hamiltonian are restored. Contrary to what one might expect, Eq. (5) cannot be obtained from the corresponding result in Ref. 12 merely by stereographically projecting the paths from the sphere onto the (extended) plane. Nevertheless, the proof given in the next section shows that the key ideas behind both constructions are the same; also see the concluding remarks.

(vii) So far we have considered a fixed spin quantum number j. In order to make contact with the Wiener-regularized path-integral expression associated with a canonical degree of freedom, also proved in Ref. 12, one has to contract[44,45] the algebra of SU(2) to the Heisenberg–Weyl algebra by taking the *high-spin limit* $j\to\infty$. More explicitly, in the given (polynomial) spin Hamiltonian $\mathcal{H}$ on $\mathbb{C}^{2j+1}$, one has to replace $\mathcal{J}_+$, $\mathcal{J}_-$, and $\mathcal{J}_3$ by $\mathcal{J}_+/\sqrt{2j}$, $\mathcal{J}_-/\sqrt{2j}$, and $\mathcal{J}_3+j\mathbf{1}$, respectively. If $\mathcal{H}_j=\int_{\mathbb{C}}d^2z\,h_j(z)|z\rangle\langle z|$ denotes the resulting operator, one then finds the relation

$$\lim_{j\to\infty}\frac{\pi}{2j}\langle z/\sqrt{2j}|e^{-t\mathcal{H}_j}|z'/\sqrt{2j}\rangle=\langle\langle z|e^{-t\mathsf{H}}|z'\rangle\rangle, \qquad (12)$$

where $|z\rangle\rangle\in L^2(\mathbb{R})$ is a normalized canonical coherent vector[30–32] and the Hamiltonian H on $L^2(\mathbb{R})$, the Hilbert space of Lebesgue square-integrable complex-valued functions on the real line $\mathbb{R}$, is defined by

$$\mathsf{H}:=\int_{\mathbb{C}}\frac{d^2z}{\pi}\mathsf{h}(z)|z\rangle\rangle\langle\langle z|, \quad \text{with } \mathsf{h}(z):=\lim_{j\to\infty}h_j(z/\sqrt{2j}). \qquad (13)$$

By using (5) for the prelimit expression in (12), suitably rescaling the Brownian bridge, and interchanging the order of the limits $j\to\infty$ and $\nu\to\infty$, one arrives at the path-integral formula

$$\langle\langle z|e^{-t\mathsf{H}}|z'\rangle\rangle=\pi\lim_{\nu\to\infty}e^{2t\nu}\int d\mu^{(\nu)}_{z,0;z',t}(b)\exp\left\{\frac{1}{2}\int_0^t ds[\dot b(s)b^*(s)-\dot b^*(s)b(s)]\right\}$$
$$\times\exp\left\{-\int_0^t ds\,\mathsf{h}(b(s))\right\}, \qquad (14)$$

in agreement with Eq. (1.3) in Ref. 12; also see Refs. 46 and 35. Formula (14) can be shown to hold not only for the polynomial Hamiltonians $\mathcal{H}$ resulting from the contraction, but for a wider class of operators whose conditions are stated in Theorem 2.4 of Ref. 12.

(viii) With regard to some of the symbolic path-integral expressions for spin systems frequently encountered in the literature, see, for example, Refs. 10, 11, and 23–28, it might be illuminating to recognize certain formal similarities between these expressions and the above result (5). While the kinematical and dynamical terms in the exponents of all the corresponding path integrands look essentially the same, only the above result is based on a genuine path measure, namely, $d\mu^{(\nu)}_{z,0;z',t}(b)\exp\{4(j+1)\nu\int_0^t ds(1+|b(s)|^2)^{-2}\}$, but requires taking the limit $\nu\to\infty$. Here, the Wiener type of measure $d\mu^{(\nu)}_{z,0;z',t}(b)$ is often symbolically written as $\delta^2 b\ \delta(b(0)-z)\,\delta(b(t)-z')\exp\{-(1/4\nu)\int_0^t ds|\dot{b}(s)|^2\}$, or similarly. In any case, the necessity to regularize by some ultradiffusive limit was observed several times also in nonrigorous works.[10,17–19,27]

III. PROOF

The proof of (5) consists of three major steps, adapting key ideas of Ref. 12. First, the spin Hilbert space $\mathbb{C}^{2j+1}$ is embedded into $L^2(\mathbb{C})$, the Hilbert space of Lebesgue square-integrable complex-valued functions on $\mathbb{C}$. Next, it is identified with the $(2j+1)$-dimensional ground-state eigenspace of a suitable Schrödinger operator R acting on $L^2(\mathbb{C})$. Then the spin semigroup, now realized on $L^2(\mathbb{C})$, is shown to be the limit $\nu\to\infty$ of a Schrödinger semigroup generated by a suitably perturbed νR. Rewriting this Schrödinger semigroup with the help of the standard Feynman–Kac–Itô path-integral formula finally gives (5).

A. The embedding of the spin Hilbert space

The embedding of the spin Hilbert space $\mathbb{C}^{2j+1}$ into the infinite-dimensional Hilbert space $L^2(\mathbb{C})$, equipped with the standard scalar product $(\varphi|\psi):=\int_{\mathbb{C}} d^2z\ \varphi^*(z)\psi(z)$, is accomplished by interpreting the coherent representation as a linear isometric mapping,

$$I:\ \mathbb{C}^{2j+1}\to L^2(\mathbb{C}),\quad |\psi\rangle\mapsto\psi, \tag{15}$$

where the function ψ on $\mathbb{C}\cong\mathbb{R}\times\mathbb{R}$ is defined by its values $\psi(z):=\langle z|\psi\rangle$.

The (Hilbert) adjoint $I^\dagger$ of I explicitly reads as

$$I^\dagger:\ L^2(\mathbb{C})\to\mathbb{C}^{2j+1},\quad \varphi\mapsto\int_{\mathbb{C}} d^2z\ \varphi(z)|z\rangle, \tag{16}$$

and the isometric property is simply stated as $I^\dagger I=\mathbf{1}$. The orthogonal projection from $L^2(\mathbb{C})$ onto $I(\mathbb{C}^{2j+1})$ is the operator $II^\dagger=:E_0$.

Every operator $\mathcal{B}$ on $\mathbb{C}^{2j+1}$ can be realized by the unitary equivalent $I\mathcal{B}I^\dagger$ on $E_0(L^2(\mathbb{C}))=I(\mathbb{C}^{2j+1})$, which trivially extends to the whole of $L^2(\mathbb{C})$. In particular, it follows from (4) that

$$I\mathcal{H}I^\dagger=E_0HE_0, \tag{17}$$

where H is the bounded multiplication operator on $L^2(\mathbb{C})$ defined by the function h, that is, $(H\varphi)(z):=h(z)\varphi(z)$ for all $\varphi\in L^2(\mathbb{C})$. Furthermore, the embedded operator $I\mathcal{H}I^\dagger$ possesses a continuous integral kernel $I\mathcal{H}I^\dagger(z,z')$ (also known as its position representation) given by the coherent representation of $\mathcal{H}$, that is,

$$I\mathcal{H}I^\dagger(z,z')=\langle z|\mathcal{H}|z'\rangle. \tag{18}$$

Using (17), one can now verify the identify $Ie^{-t\mathcal{H}}I^\dagger=E_0e^{-tE_0HE_0}$ to all orders in t, which, analogous to (18), shows that $E_0e^{-tE_0HE_0}$ has a continuous integral kernel given by the equation

$$E_0e^{-tE_0HE_0}(z,z')=\langle z|e^{-t\mathcal{H}}|z'\rangle. \tag{19}$$

B. A Schrödinger operator and its ground-state eigenspace

Consider on $L^2(\mathbb{C})$ the ‘‘magnetic’’ Schrödinger operator,

$$R := (i\partial_1 + A_1)^2 + (i\partial_2 + A_2)^2 + V, \tag{20}$$

with the partial differential operators $\partial_1 := \partial/\partial z_1$, $\partial_2 := \partial/\partial z_2$ and the vector and scalar potentials $\binom{A_1}{A_2}$ and V acting as multiplication operators defined by the bounded and continuous functions,

$$\begin{pmatrix} a_1(z) \\ a_2(z) \end{pmatrix} := \begin{pmatrix} -\partial_2 \\ \partial_1 \end{pmatrix} \ln g(z) = \frac{2(j+1)}{1+|z|^2} \begin{pmatrix} z_2 \\ -z_1 \end{pmatrix}, \tag{21}$$

$$v(z) := \partial_1 a_2(z) \text{-} \partial_2 a_1(z) = -\frac{4(j+1)}{(1+|z|^2)^2}. \tag{22}$$

The self-adjoint operator R is tailored such that its ground-state eigenspace is identical to $E_0(L^2(\mathbb{C}))$ and the corresponding eigenvalue vanishes. In essence, this follows from a result of Aharonov and Casher[47] on zero-energy eigenstates. Since the proof is quite short, we will give it, thereby closely following the presentation in Ref. 48. Factorized like $R = D^\dagger D$, where $D := i\partial_1 + \partial_2 + A_1 - iA_2$, the positivity of R becomes manifest. Its null space consists of all those functions ψ in $L^2(\mathbb{C})$ with $D\psi = 0$. The general solution of this differential equation is a product $\psi = g\phi$, where ϕ is any function analytic in z^*, that is, $(\partial_1 - i\partial_2)\phi = 0$. Due to (3), square integrability then requires ϕ to be any polynomial in z^* of maximal degree $2j$, which proves that the ground-state eigenspace of R and the subspace $E_0(L^2(\mathbb{C})) = I(\mathbb{C}^{2j+1})$ are identical.

Two remarks are in order.

(i) The spectrum of R coincides with the positive half-line, as can be inferred from Theorem 6.1 in Ref. 48. Following arguments as in the proof of Theorem 6.2 in Ref. 48, one sees that zero is the only eigenvalue. Therefore the nature of the spectrum and the ground-state eigenfunctions are explicitly known. However, we are not aware of explicit results on generalized eigenfunctions corresponding to strictly positive spectral values.

(ii) Employing the spectral theorem, one proves that the semigroup generated by νR converges strongly to the ground-state projection E_0, in the sense that

$$\lim_{\nu\to\infty} \| e^{-t\nu R}\varphi - E_0\varphi \| = 0, \quad \text{for all} \quad \varphi \in L^2(\mathbb{C}) \quad \text{and} \quad t > 0, \tag{23}$$

where the norm $\|\cdot\| := (\cdot|\cdot)^{1/2}$ corresponds to the standard scalar product on $L^2(\mathbb{C})$.

C. The spin semigroup as the limit of a Schrödinger semigroup

With the material gathered in Secs. III A and III B we can isolate the central reason for the validity of the main result (5) of the present paper. The point is that the spin semigroup, now realized on $L^2(\mathbb{C})$, can be understood as the limit $\nu \to \infty$ of the Schrödinger semigroup generated by $\nu R + H$. More precisely, we will show that the continuous integral kernel given in (19) is the pointwise limit

$$E_0 e^{-tE_0HE_0}(z,z') = \lim_{\nu\to\infty} e^{-t(\nu R + H)}(z,z'), \quad \text{for all} \quad z,z' \in \mathbb{C} \quad \text{and} \quad t > 0, \tag{24}$$

where the prelimit expression is the continuous integral kernel of $\exp\{-t(\nu R + H)\}$. By expressing this integral kernel in terms of the *Feynman–Kac–Itô formula*[49,50] (observing $\partial_1 a_1 + \partial_2 a_2 = 0$)

$$e^{-t(\nu R+H)}(z,z')=\int d\mu^{(\nu)}_{z,0;z',t}(b)\exp\left\{(j+1)\int_0^t ds\,\frac{\dot{b}(s)b^*(s)-\dot{b}^*(s)b(s)}{1+|b(s)|^2}\right\}$$
$$\times\exp\left\{4(j+1)\nu\int_0^t\frac{ds}{(1+|b(s)|^2)^2}-\int_0^t ds\,h(b(s))\right\},\tag{25}$$

the right-hand sides of (24) and (5) are seen to coincide.

The proof of (24) makes essential use of the semigroup property of $e^{-t(\nu R+H)}$. Throughout the proof we fix $t>0$ and pick some reference diffusion constant $\nu_0>0$. As a starting point we define

$$\eta_w^{(\lambda)}(z):=e^{-t(\nu_0 R+\lambda H)}(z,w),\tag{26}$$

for all $\lambda\in\mathbb{R}$ and $w,z\in\mathbb{C}$. We assert that the function $\eta_w^{(\lambda)}:z\mapsto\eta_w^{(\lambda)}(z)$ is continuous, bounded, and lies in $L^2(\mathbb{C})$. The continuity follows from that of the integral kernel in (25). Boundedness and square integrability result from the inequality

$$|\eta_w^{(\lambda)}(z)|\leqslant e^{4(j+1)t\nu_0}e^{t|\lambda|\|h\|_\infty}(4\pi t\nu_0)^{-1}e^{-|z-w|^2/4t\nu_0},\tag{27}$$

where $\|h\|_\infty:=\sup_{z\in\mathbb{C}}|h(z)|<\infty$ denotes the supremum norm of h. This inequality, in turn, is found by estimating the path integral in (25). We also state that the mappings $w\mapsto\eta_w^{(\lambda)}$ and $\lambda\mapsto\eta_w^{(\lambda)}$ are strongly continuous. The first statement holds because of $(\eta_w^{(\lambda)}|\eta_{w'}^{(\lambda)})=e^{-2t(\nu_0 R+\lambda H)}(w,w')$ and the continuity of the integral kernel. The second one is a consequence of the inequality

$$\|\eta_w^{(\lambda)}-\eta_w^{(\lambda')}\|\leqslant\sqrt{\frac{t}{8\pi\nu_0}}|\lambda-\lambda'|\|h\|_\infty e^{4(j+1)t\nu_0}e^{t\max\{|\lambda|,|\lambda'|\}\|h\|_\infty},\tag{28}$$

which is derived by estimating the difference of two path integrals of type (25) using the elementary inequality $|e^x-e^y|\leqslant|x-y|e^{\max\{x,y\}}$, for $x,y\in\mathbb{R}$.

The following two steps of the proof are based on writing the integral kernel for $\nu>2\nu_0$ as a scalar product,

$$e^{-t(\nu R+H)}(z,z')=(\eta_z^{(\nu_0/\nu)}|e^{-t(\nu-2\nu_0)(R+H/\nu)}\eta_{z'}^{(\nu_0/\nu)}).\tag{29}$$

In the first step, we claim that

$$\lim_{\nu\to\infty}(\eta_z^{(\nu_0/\nu)}|e^{-t(\nu-2\nu_0)(R+H/\nu)}\eta_{z'}^{(\nu_0/\nu)})=(\eta_z^{(0)}|E_0e^{-tE_0HE_0}\eta_{z'}^{(0)}),\tag{30}$$

for all $z,z'\in\mathbb{C}$.

Due to the strong continuity of $\lambda\mapsto\eta_w^{(\lambda)}$, the boundedness of $e^{-t(\nu-2\nu_0)(R+H/\nu)}$, which is uniform in ν, and the continuity of the scalar product $(\cdot|\cdot)$, it suffices to show that

$$\lim_{\nu\to\infty}\|e^{-t(\nu-2\nu_0)(R+H/\nu)}\varphi-E_0e^{-tE_0HE_0}\varphi\|=0,\quad\text{for all}\quad\varphi\in L^2(\mathbb{C}).\tag{31}$$

To prove this strong operator convergence we employ the Duhamel–Dyson–Phillips perturbation expansion,

$$e^{-t(\nu-2\nu_0)(R+H/\nu)}\varphi=e^{-t(\nu-2\nu_0)R}\varphi+\sum_{n=1}^{\infty}\left(\frac{2\nu_0-\nu}{\nu}\right)^n\int_0^t ds_n\cdots\int_0^{s_3}ds_2\int_0^{s_2}ds_1\,e^{-(t-s_n)(\nu-2\nu_0)R}$$
$$\times H\times\cdots\times e^{-(s_2-s_1)(\nu-2\nu_0)R}He^{-s_1(\nu-2\nu_0)R}\varphi,\tag{32}$$

which converges uniformly in $\nu\in]2\nu_0,\infty[$ with respect to the norm on $L^2(\mathbb{C})$. This holds because the norm of the series is dominated by the exponential series $\Sigma_{n=0}^{\infty}(t^n/n!)\|h\|_\infty^n\|\varphi\|$, independent of ν. Thus, the limit $\nu\rightarrow\infty$ can be interchanged with the summation and, using (23) and the dominated-convergence theorem, we obtain the expansion of $E_0 e^{-tE_0HE_0}\varphi$.

In the second and final step we claim that the right-hand side of (30) is already the desired integral kernel, that is,

$$(\eta_z^{(0)}|E_0 e^{-tE_0HE_0}\eta_{z'}^{(0)})=E_0 e^{-tE_0HE_0}(z,z'),\quad \text{for all } z,z'\in\mathbb{C}. \tag{33}$$

This is verified by checking that the mapping $(z,z')\mapsto(\eta_z^{(0)}|E_0 e^{-tE_0HE_0}\eta_{z'}^{(0)})$ constitutes an integral kernel of $E_0 e^{-tE_0HE_0}$ and is, in fact, continuous. The former is true since $e^{-t\nu_0 R}E_0=E_0$. The latter holds because the mapping $w\mapsto\eta_w^{(0)}$ is strongly continuous, the operator $E_0 e^{-tE_0HE_0}$ is bounded, and the scalar product $(\cdot|\cdot)$ is continuous.

IV. CONCLUDING REMARKS

We conclude the paper with six remarks.

(i) As already mentioned in Sec. II, the main result (5) cannot be obtained from a result in Ref. 12 merely by stereographically projecting the Brownian paths from the two-sphere S^2 onto the (extended) Euclidean plane $\mathbb{R}^2$. The reason can be traced back to the different operators, or equivalently path measures, used for regularization. The stereographic projection corresponds to reexpressing the differential operator on $L^2(S^2)$ used by the authors of Ref. 12 in flat Cartesian coordinates. The resulting operator is not of the standard Schrödinger form, acts on a weighted Hilbert space, and is not related to planar Brownian motion.

(ii) In contrast to Ref. 12 the regulatizing operator R used in the proof of (24), and hence of (5), has no spectral gap above its ground-state eigenvalue. Accordingly, $e^{-t\nu R}$ only converges strongly, and not in operator norm, to the corresponding eigenprojection E_0 as $\nu\rightarrow\infty$. As a consequence, the foregoing proof of the pointwise convergence of integral kernels required a strategy different from that in Ref. 12.

(iii) From a fundamental point of view, it is gratifying that a spin system can be related to a limit of a well-defined integral over continuous Brownian-motion paths. From a practical point of view, it would be desirable to apply to (5) the well-established theory and computational possibilities associated with the flat-space Wiener measure,[3,39,51] in order to attack specific spin problems of physical interest. One such problem, which has been extensively discussed in the recent literature,[23–28] is to understand the nature of the saddle-point approximation for the evaluation of continuum path integrals connected with simple spin Hamiltonians. Looking at Table I and the resulting j-dependence of the path integrand in (5), this approximation is expected to be the more reliable the larger the spin quantum number is. Moreover, for Hamiltonians $\mathcal{H}$ linear in the spin operators, the saddle-point approximation is believed[23–27] to give the (explicitly known) exact result already for given finite j. In this context, when dealing with symbolic continuum path integrals one has to overcome the so-called overspecification problem due to missing regularizing terms in the action functionals of those path integrals.[10,27] Rigorous continuum path integrals as used in (5) do not suffer from this problem by their very construction. Of course, the details for the saddle-point approximation of the Wiener type of path integral in the ultradiffusive limit still have to be worked out.

(iv) In Refs. 52 and 53, the ground-state eigenspace of a charged point mass under the influence of a certain magnetic field on an even-dimensional Riemannian manifold is studied, thereby extending the Aharonov–Casher theorem.[47,48] This result lies at the heart of the quantization procedure proposed in Refs. 53–55. A quantum system is hereby represented on the ground-state eigenspace of such a generalized Landau Hamiltonian on the Hilbert space of functions over its classical phase space. The symplectic structure of the latter determines the magnetic field. In this sense, Eq. (5) read from right to left can be viewed as a *quantization prescription* for a classical spin system.

In this context it is worth mentioning that the path integral in (5) is well-defined for all values of j taken from the positive half-line. Even more, in the limit $\nu\to\infty$ it manages to single out the set of allowed spin quantum numbers, {0,1/2,1,3/2,...}, from the ''classical continuum $[0,\infty[$.

More precisely, for a given bounded and continuous h: $\mathbb{C}\to\mathbb{C}$ and $j\in[0,\infty[$ we assert that the right-hand side of (5) is equal to $\langle\psi(z)|e^{-t\mathcal{H}_\psi}|\psi(z')\rangle$. Here the set of vectors,

$$|\psi(z)\rangle := g(z)\sum_{n=0}^{2(j)}\sqrt{\binom{2j}{n}}z^n|\psi_n\rangle,\quad z\in\mathbb{C}, \tag{34}$$

is unity-resolving in $\mathbb{C}^{2(j)+1}$, where (j) denotes the smallest integer or half-integer equal to or larger than j, and $\{|\psi_n\rangle\}$ is a fixed but arbitrary orthonormal basis in $\mathbb{C}^{2(j)+1}$. The binomial coefficient can be defined recursively by $\binom{2j}{0}:=1$ and $\binom{2j}{n+1}:=[(2j-n)/(n+1)]\binom{2j}{n}$, and $g(z)$ is defined by (3) as it stands for general $j\in[0,\infty[$. Finally, $\mathcal{H}_\psi$ is an operator on $\mathbb{C}^{2(j)+1}$ associated to the given h by the definition

$$\mathcal{H}_\psi := \int_{\mathbb{C}} d^2z\,h(z)|\psi(z)\rangle\langle\psi(z)|. \tag{35}$$

This association can be viewed as a quantization, which maps the pair (j,h) to the pair $((j),\mathcal{H}_\psi)$ with $\mathcal{H}_\psi$ being interpreted as the Hamiltonian of a spin with quantum number (j). While $\mathcal{H}_\psi$ in general depends on the chosen basis $\{|\psi_n\rangle\}$, the expression $\langle\psi(z)|e^{-t\mathcal{H}_\psi}|\psi(z')\rangle$ does not because of unitary invariance.

For the proof of the above assertion we remark that the latter is identical to (5) in the case $j=(j)$, because then $|\psi(z)\rangle=|z\rangle$ when choosing $|\psi_n\rangle=|j,n-j\rangle$, the usual orthonormal eigenbasis of $\mathcal{J}_3$. In the case $j<(j)$, the proof follows from (25), equations analogous to (24) and (19), and the Aharonov-Casher theorem, which in our setting states that the ground-state eigenspace of the ''magnetic'' Schödinger operator R [stemming from g, confer (20)–(22)] has a dimension equal to the largest integer strictly smaller than $|\int_{\mathbb{C}}d^2z\,v(z)|/2\pi=2j+2$ and is spanned by the set of orthonormal functions $z\mapsto\langle\psi(z)|\psi_n\rangle$, $n=0,1,\ldots,2(j)$.

(v) It is straightforward to generalize formula (5) to systems where the Hamiltonian $\mathcal{H}$ depends explicitly on time and/or several (coupled) spins. The formula in the latter case, like its older ''spherical relative'' in Ref. 12, may then serve as a rigorous starting point for the derivation of effective field theories, which aim to describe the low-energy excitations of quantum lattice models for magnetism. Confer, for example, Refs. 17, 56, 57, and references therein.

(vi) Following the reasoning of the present paper it should also be straightforward to derive flat-space Wiener-regularized path integrals also for physical systems with degrees of freedom that are neither of the canonical nor of the spin type.

ACKNOWLEDGMENTS

The authors are grateful to G. Junker, J. R. Klauder, and E. A. Kochetov for stimulating discussions and helpful remarks. One of the authors (B.B.) would like to thank the Institut für Theoretische Physik at the Universität Erlangen–Nürnberg for the kind hospitality during the final stage of the preparation of this work.

[1] R. P. Feynman, ''Space–time approach to non-relativistic quantum mechanics,'' Rev. Mod. Phys. **20**, 367–387 (1948).
[2] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).
[3] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[4] L. S. Schulman, *Techniques and Applications of Path Integration*, corr. reprint of the 1981 edition (Wiley, New York, 1996).
[5] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, and Polymer Physics*, 2nd ed. (World Scientific, Singapore, 1995).
[6] G. Roepstorff, *Path Integral Approach to Quantum Physics* (Springer-Verlag, Berlin, 1994).
[7] J. L. Martin, ''Generalized classical dynamics, and the 'classical analogue' of a Fermi oscillator,'' Proc. R. Soc. London, Ser. A **251**, 536–542 (1959).
[8] J. L. Martin, ''The Feynman principle for a Fermi system,'' Proc. R. Soc. London, Ser. A **251**, 543–549 (1959).

[9] J. R. Klauder, ''The action option and a Feynman quantization of spinor fields in terms of ordinary c-numbers,'' Ann. Phys. (N.Y.) **11**, 123–168 (1960).
[10] J. R. Klauder, ''Path integrals and stationary-phase approximations,'' Phys. Rev. D **19**, 2349–2356 (1979).
[11] H. Kuratsuji and T. Suzuki, ''Path integral in the representation of SU(2) coherent state and classical dynamics in a generalized phase space,'' J. Math. Phys. **21**, 472–476 (1980).
[12] I. Daubechies and J. R. Klauder, ''Quantum-mechanical path integrals with Wiener measure for all polynomial Hamiltonians II,'' J. Math. Phys. **26**, 2239–2256 (1985).
[13] L. Schulman, ''A path integral for spin,'' Phys. Rev. **176**, 1558–1569 (1968).
[14] A. Inomata, G. Junker, and C. Rösch, ''Remarks on the magnetic top,'' Found. Phys. **28**, 735–739 (1998).
[15] H. B. Nielsen and D. Rohrlich, ''A path integral to quantize spin,'' Nucl. Phys. B **299**, 471–483 (1988).
[16] K. Johnson, ''Functional integrals for spin,'' Ann. Phys. (N.Y.) **192**, 104–118 (1989).
[17] E. Fradkin and M. Stone, ''Topological terms in one- and two-dimensional quantum Heisenberg antiferromagnets,'' Phys. Rev. B **38**, 7215–7218 (1988).
[18] M. Stone, ''Supersymmetry and the quantum mechanics of spin,'' Nucl. Phys. B **314**, 557–586 (1989).
[19] T. Jaroszewicz and P. S. Kurzepa, ''Spin, statistics, and geometry of random walks,'' Ann. Phys. (N.Y.) **210**, 255–322 (1991).
[20] D. C. Cabra, A. Dobry, A. Greco, and G. L. Rossini, ''On the path integral representation for spin systems,'' J. Phys. A **30**, 2699–2704 (1997).
[21] G. F. De Angelis and G. Jona-Lasinio, ''A stochastic description of a spin-1/2 particle in a magnetic field,'' J. Phys. A **15**, 2053–2061 (1982).
[22] L. F. Lemmens, ''The Ehrenfest model: A path-integral for spin,'' Phys. Lett. A **222**, 419–423 (1996).
[23] H. Kuratsuji, ''Path integrals in the SU(2) coherent state representation and related topics,'' in *Path Integrals and Coherent States of SU(2) and SU(1,1)*, edited by A. Inomata, H. Kuratsuji, and C. C. Gerry (World Scientific, Singapore, 1992), pp. 139–218.
[24] K. Funahashi, T. Kashiwa, S. Sakoda, and K. Fujii, ''Coherent states, path integral, and semi-classical approximation,'' J. Math. Phys. **36**, 3232–3253 (1995).
[25] K. Funahashi, T. Kashiwa, S. Nima, and S. Sakoda, ''More about path integrals for spin,'' Nucl. Phys. B **453**, 508–528 (1995).
[26] E. A. Kochetov, ''SU(2) coherent-state path integral,'' J. Math. Phys. **36**, 4667–4679 (1995).
[27] E. Ercolessi, G. Morandi, F. Napoli, and P. Pieri, ''Path integrals for spinning particles, stationary phase and the Duistermaat–Heckman theorem,'' J. Math. Phys. **37**, 535–553 (1996).
[28] J. Shibata and S. Takagi, ''A note on (spin-) coherent-state path integral,'' quant-ph/9807005.
[29] J. M. Radcliffe, ''Some properties of coherent spin states,'' J. Phys. A **4**, 313–323 (1971).
[30] J. R. Klauder and B.-S. Skagerstam, *Coherent States—Applications in Physics and Mathematical Physics* (World Scientific, Singapore, 1985).
[31] A. Perelomov, *Generalized Coherent States and their Applications* (Springer-Verlag, Berlin, 1986).
[32] W.-M. Zhang, D. H. Feng, and R. Gilmore, ''Coherent states: Theory and some applications,'' Rev. Mod. Phys. **62**, 867–927 (1990).
[33] F. A. Berezin, ''Covariant and contravariant symbols of operators,'' Math. USSR Izv. **6**, 1117–1151 (1972), Russian original, Izv. Akad. Nauk SSSR, Ser. Mat. **36**, 1134–1167 (1972).
[34] B. Simon, ''The classical limit of quantum partition functions,'' Commun. Math. Phys. **71**, 247–276 (1980).
[35] J. R. Klauder, ''Coherent state path integrals at (nearly) 40,'' quant-ph/9810043, to appear in *Path Integrals from peV to TeV—Firenze'98*, edited by V. Tognetti, R. Vaia, R. Casalbuoni, R. Giachetti, and P. Verrucchi (World Scientific, Singapore, 1999).
[36] E. H. Lieb, ''The classical limit of quantum spin systems,'' Commun. Math. Phys. **31**, 327–340 (1973).
[37] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales*, Vol. 2, *Itô Calculus* (Wiley, Chichester, 1987).
[38] Ph. Protter, *Stochastic Integration and Differential Equations—A New Approach*, 3rd printing (Springer-Verlag, Berlin, 1995).
[39] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed. (Springer-Verlag, Berlin, 1999).
[40] Theorem IV.46.4 in Ref. 37.
[41] Exercise IV.2.18 in Ref. 39.
[42] F. A. Berezin, ''Quantization,'' Math. USSR Izv. **8**, 1109–1165 (1974), Russian original, Izv. Akad. Nauk SSSR, Ser. Mat. **38**, 1116–1175 (1974).
[43] F. A. Berezin, ''General concept of quantization,'' Commun. Math. Phys. **40**, 153–174 (1975).
[44] F. T. Arecchi, E. Courtens, R. Gilmore, and H. Thomas, ''Atomic coherent states in quantum optics,'' Phys. Rev. A **6**, 2211–2237 (1972).
[45] R. Gilmore, *Lie groups, Lie algebras, and Some of their Applications* (Wiley, New York, 1974).
[46] I. Daubechies and J. R. Klauder, ''True measures for real time path integrals,'' in *Path integrals from meV to MeV*, edited by M. C. Gutzwiller, A. Inomata, J. R. Klauder, and L. Streit (World Scientific, Singapore, 1986), pp. 425–432.
[47] Y. Aharonov and A. Casher, ''Ground state of spin-1/2 charged particle in a two-dimensional magnetic field,'' Phys. Rev. A **19**, 2461–2462 (1979).
[48] H. L. Cycon, R. G. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators* (Springer-Verlag, Berlin, 1987).
[49] Theorems 14.5 and 15.5 in Ref. 3.
[50] K. Broderix, D. Hundertmark, and H. Leschke, ''Continuity properties of Schrödinger semi-groups with magnetic fields,'' math-ph/9808004 and mp_arc 98-564, to appear in Rev. Math. Phys.

[51] A. D. Egorov, P. I. Sobolevski, and L. A. Yanovich, *Functional Integrals: Approximate Evaluation and Applications* (Kluwer, Dordrecht, 1993).
[52] P. Maraner, ''Landau ground state on Riemannian surfaces,'' Mod. Phys. Lett. A **7**, 2555–2558 (1992).
[53] R. Alicki, J. R. Klauder, and J. Lewandowski, ''Landau-level ground state and its relation for a general quantization procedure,'' Phys. Rev. A **48**, 2538–2548 (1993).
[54] R. Alicki and J. R. Klauder, ''Quantization of systems with a general phase space equipped with a Riemannian metric,'' J. Phys. A **29**, 2475–2483 (1996).
[55] J. R. Klauder and E. Onofri, ''Landau levels and geometric quantization,'' Int. J. Mod. Phys. A **4**, 3939–3949 (1989).
[56] E. Fradkin, *Field Theories of Condensed Matter Systems* (Addison-Wesley, Redwood City, CA, 1991).
[57] A. Auerbach, *Interacting Electrons and Quantum Magnetism*, corr. 2nd printing (Springer-Verlag, New York, 1998).

On the geometric equivalence of certain discrete integrable Heisenberg ferromagnetic spin chains

M. Daniel and K. Manivannan
Centre for Nonlinear Dynamics, Department of Physics, Bharathidasan University, Tiruchirapalli 620 024, India

(Received 24 November 1997; accepted for publication 1 February 1999)

Using a discrete curve mapping procedure, we reformulate the problem of nonlinear spin dynamics in three different discrete Heisenberg ferromagnetic spin chain models with different magnetic interactions. The procedure helps to rewrite the Landau–Lifshitz equations that govern the dynamics of spins in these ferromagnetic spin systems as equivalent to the integrable discretization of the completely integrable nonlinear Schrödinger family of equations. The elementary spin excitations in these spin systems are governed by lattice solitons. © *1999 American Institute of Physics.* [S0022-2488(99)00706-9]

I. INTRODUCTION

The one-dimensional classical continuum Heisenberg ferromagnetic spin chain models with different magnetic interactions have been identified as one of the interesting class of nonlinear dynamical systems exhibiting complete integrability and soliton spin excitations on many occasions.[1–12] Notable among them are the one-dimensional isotropic Heisenberg ferromagnetic spin chain with bilinear[3,4] and also biquadratic[5] exchange interactions, anisotropic interaction,[6] interaction with external magnetic fields,[7] spin-phonon coupling,[8–10] weak interaction,[11,12] etc. The dynamics of spins in these classical continuum Heisenberg ferromagnetic spin systems are governed by the Landau–Lifshitz (L–L) equations, a highly nontrivial vector nonlinear partial differential equation. In addition to linear magnons an interesting class of nonlinear elementary spin excitations that occur are solitons, domain walls, kinks, breathers, etc.[1–12] in one-dimensional chains and line solitons, lump solitons, dromions, etc.[13] in higher dimensional spin systems. Solving the L–L equations for classical spin systems in its natural vector form for understanding the underlying nonlinear dynamics is a difficult task in general. Hence, attempts were made to relate the L–L equations as equivalent to nonlinear equations which have been already studied. Notable among them are the geometric equivalence or mapping the spin chain with a moving continuous space curve[14] and the gauge equivalence technique.[15,16] Both these methods have helped to rewrite the L–L equations as equivalent to the nonlinear Schrödinger (NLS) family of equations depending on the nature of magnetic interactions involved. For instance, the nonlinear spin dynamics of the isotropic bilinear ferromagnetic spin chain can be represented equivalently in terms of the completely integrable cubic NLS equation through geometric and gauge equivalence methods[3,15] and the biquadratic ferromagnetic chain to the fourth order completely integrable NLS equation.[5] As the above integrable NLS family of equations admit N-soliton solutions, the elementary spin excitations in these magnetic systems can then be expressed in terms of soliton modes. Later, these methods have been extensively used in the case of several other physically interesting and mathematically generalized ferromagnetic spin models, which resulted in the identification of new integrable spin systems.[17] The dynamics of a continuous moving space curve has played an important role in identifying integrable models not only in the case of classical continuum ferromagnets but also in several other contexts. We have many physical examples in which geometric considerations yield NLS family of equations and also the motion of the continuous space curve and its elementary geometric properties were proved to select integrable dynamics of several soliton equations.[18–24]

Though several classical continuum Heisenberg ferromagnetic spin models have been identified as completely integrable systems through continuous space curve mapping, this procedure has not been adapted in the case of more realistic discrete or lattice spin chain models so far. In a recent paper Doliwa and Santini[25] from the elementary geometric properties of a discrete moving curve showed that the motion of it on a sphere selects integrable dynamics of the Ablowitz–Ladik[26] hierarchy of evolution equations. The theory of discrete curves found in Ref. 25 motivated us to develop a discrete generalized mapping procedure[27] for the lattice models of few integrable classical Heisenberg ferromagnetic spin chains which can help to rewrite the discrete dynamical equations in the L–L form in terms of the integrable discretized version of the NLS family of equations thus making the analysis of the underlying dynamics more simple. After presenting a brief account of this discrete mapping procedure in Sec. II, we employ the procedure in the case of three different integrable discrete classical Heisenberg ferromagnetic spin chain models with different magnetic interactions. In Sec. III, we map the integrable discrete isotropic bilinear Heisenberg spin chain model originally proposed by Ishimori to the integrable discretized version of the completely integrable cubic NLS equation using the discrete curve mapping procedure developed in Sec. II. Also, using the same procedure we map a higher order integrable discrete ferromagnetic spin chain model to the completely integrable discrete Hirota equation and a new generalized isotropic Heisenberg ferromagnetic discrete spin chain model involving higher order and next nearest neighbor exchange interactions to the integrable discretization of the fourth order integrable NLS equation. The results are concluded in Sec. IV.

II. MOTION OF A DISCRETE CURVE

Before attempting to map integrable discrete ferromagnetic spin systems to integrable discretization of the completely integrable nonlinear Schrödinger family of equations, in this section, we present a brief account of a generalization of the theory of discrete moving space curves closely following the procedure presented by Doliwa and Santini in Ref. 25. The motion of the discrete curve is considered to be taking place on a sphere with origin at O and $\mathbf{r}_n$ is the unit vector at the point n of the sequence pointing along the direction of the position vector $\mathbf{R}_n$ (see Fig. 1). The unit vectors $\mathbf{t}_n$ and $\mathbf{t}'_n$ are tangent to the big circle passing through the points $\mathbf{r}_n$ and $\mathbf{r}_{n+1}$ at the points, respectively, and the angle between $\mathbf{r}_n$ and $\mathbf{r}_{n+1}$ is defined as θ_n. A third unit vector $\mathbf{b}_n$ is defined as normal to both $\mathbf{r}_n$ and $\mathbf{t}_n$ (i.e) $\mathbf{b}_n=\mathbf{r}_n\wedge\mathbf{t}_n$. The angle between $\mathbf{t}'_{n-1}$ and $\mathbf{t}_n$ is denoted by ϕ_{n-1}. The transition from the basis $(\mathbf{r}_n,\mathbf{t}_n,\mathbf{b}_n)$ at the nth point to the next basis $(\mathbf{r}_{n+1},\mathbf{t}_{n+1},\mathbf{b}_{n+1})$ at the neighboring $(n+1)$th point can be represented by[25,27]

$$(\mathbf{r}_{n+1}\ \ \mathbf{t}_{n+1}\ \ \mathbf{b}_{n+1})^T=[M](\mathbf{r}_n\ \ \mathbf{t}_n\ \ \mathbf{b}_n)^T, \tag{2.1a}$$

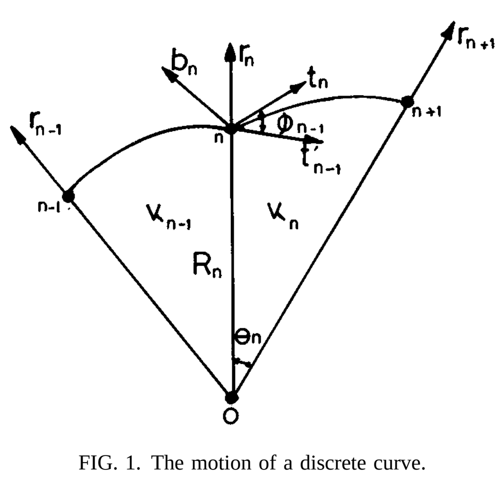

FIG. 1. The motion of a discrete curve.

where

$$[M]=\begin{pmatrix} \cos\theta_n & \sin\theta_n & 0 \\ -\cos\phi_n \sin\theta_n & \cos\phi_n\cos\theta_n & \sin\phi_n \\ \sin\phi_n\sin\theta_n & -\sin\phi_n\cos\theta_n & \cos\phi_n \end{pmatrix}, \tag{2.1b}$$

and T means the transpose of the matrix. Similarly, the basis $(\mathbf{r}_{n-1},\mathbf{t}_{n-1},\mathbf{b}_{n-1})$ associated with the point $(n-1)$ of the sequence can be obtained from $(\mathbf{r}_n,\mathbf{t}_n,\mathbf{b}_n)$ by using the matrix $[M]^{-1}$ (with n replaced by $(n-1)$). In a similar way the next nearest bases, namely, $(\mathbf{r}_{n+2},\mathbf{t}_{n+2},\mathbf{b}_{n+2})$ and $(\mathbf{r}_{n-2},\mathbf{t}_{n-2},\mathbf{b}_{n-2})$ can be generated in terms of $(\mathbf{r}_n,\mathbf{t}_n,\mathbf{b}_n)$.

As the motion of the discrete curve takes place on the surface of the sphere, the velocity field must always be tangent to the surface. The evolution of the orthonormal frame $(\mathbf{r}_n,\mathbf{t}_n,\mathbf{b}_n)$ can then be described by the antisymmetric matrix equation[25]

$$\frac{d}{dt}\begin{pmatrix}\mathbf{r}_n\\ \mathbf{t}_n\\ \mathbf{b}_n\end{pmatrix}=\begin{pmatrix} 0 & V_n & U_n \\ -V_n & 0 & W_n \\ -U_n & -W_n & 0\end{pmatrix}\begin{pmatrix}\mathbf{r}_n\\ \mathbf{t}_n\\ \mathbf{b}_n\end{pmatrix}. \tag{2.2}$$

Here V_n and U_n are the velocity field components parallel to $\mathbf{t}_n$ and $\mathbf{b}_n$, respectively, and W_n is an unknown function to be determined. The compatibility between Eqs. (2.1) and (2.2) then provides the following evolution equations[25] for θ_n and ϕ_n,

$$\frac{d\theta_n}{dt}=\cos\phi_n V_{n+1}-\sin\phi_n U_{n+1}-V_n, \tag{2.3a}$$

$$\frac{d\phi_n}{dt}=W_{n+1}-\cos\theta_n W_n+\sin\theta_n U_n, \tag{2.3b}$$

and the unknown function W_n as

$$W_n=\frac{1}{\sin\theta_n}[\cos\phi_n U_{n+1}+\sin\phi_n V_{n+1}-\cos\theta_n U_n]. \tag{2.3c}$$

Thus the motion of the discrete curve on the surface of the sphere is governed by the coupled evolution equations for the two angles θ_n and ϕ_n. Specific values of the field components V_n and U_n will give rise to different nonlinear evolution equations. Experience gained in the case of continuous space curve mapping suggests to rewrite the coupled evolution equations for θ_n and ϕ_n given in Eqs. (2.3a) and (2.3b) as a single evolution equation in terms of a new variable using a suitable transformation involving θ_n and ϕ_n. This may be advantageous because it leads to identification of the resultant equation with standard nonlinear discrete evolution equations. Now in the forthcoming sections, we employ our discrete mapping procedure developed in this section to find geometrically equivalent representations of few integrable discrete classical Heisenberg ferromagnetic spin chain models.

III. GEOMETRIC EQUIVALENCE OF CERTAIN DISCRETE HEISENBERG SPIN CHAINS

Integrable classical Heisenberg ferromagnetic spin chain models are of great interest because they exhibit an interesting class of nonlinear elementary spin excitations namely solitons. Though real ferromagnets exist in lattice, soliton excitations could be identified mostly in the classical continuum limit. In the following, we consider three one-dimensional discrete classical Heisenberg ferromagnetic spin chain models and employ our discrete curve mapping procedure to identify them with standard discrete integrable nonlinear evolution equations so that solitons can be constructed readily.

A. Ishimori model of the discrete spin chain

An integrable discrete classical Heisenberg ferromagnetic spin chain model originally proposed by Ishimori[28] is governed by the equation of motion

$$\frac{d\mathbf{S}_n(t)}{dt}=2\mathbf{S}_n\wedge\left[\frac{\mathbf{S}_{n+1}}{1+\mathbf{S}_n\cdot\mathbf{S}_{n+1}}+\frac{\mathbf{S}_{n-1}}{1+\mathbf{S}_n\cdot\mathbf{S}_{n-1}}\right], \tag{3.1a}$$

$$\mathbf{S}_n=(S_n^x,S_n^y,S_n^z), \tag{3.1b}$$

$$\mathbf{S}_n^2=1. \tag{3.1c}$$

Here $\mathbf{S}_n$ represents the three component classical spin vector at the lattice site n, the evolution of which is governed by the interaction of it with its nearest neighbors in the form of a bilinear exchange. Now, we employ our mapping procedure developed in the previous section on the Ishimori model by identifying the three-component classical spin vector $\mathbf{S}_n(t)$ of Eq. (3.1) with the unit vector $\mathbf{r}_n(t)$ of the discrete curve. Because of this identification, Eq. (3.1) can be rewritten as

$$\frac{d\mathbf{r}_n}{dt}=2\mathbf{r}_n\wedge\left[\frac{\mathbf{r}_{n+1}}{1+\mathbf{r}_n\cdot\mathbf{r}_{n+1}}+\frac{\mathbf{r}_{n-1}}{1+\mathbf{r}_n\cdot\mathbf{r}_{n-1}}\right]. \tag{3.2}$$

Substituting $\mathbf{r}_{n+1}$ and $\mathbf{r}_{n-1}$ from Eqs. (2.1) in Eq. (3.2), we obtain

$$\frac{d\mathbf{r}_n}{dt}=-2\left(\tan\frac{\theta_{n-1}}{2}\sin\phi_{n-1}\right)\mathbf{t}_n-2\left(\tan\frac{\theta_{n-1}}{2}\cos\phi_{n-1}-\tan\frac{\theta_n}{2}\right)\mathbf{b}_n\,. \tag{3.3}$$

On comparing Eq. (3.3) with Eq. (2.2) the velocity fields V_n and U_n can be identified as

$$V_n=-2\tan\frac{\theta_{n-1}}{2}\sin\phi_{n-1}, \tag{3.4a}$$

$$U_n=2\left(\tan\frac{\theta_n}{2}-\tan\frac{\theta_{n-1}}{2}\cos\phi_{n-1}\right). \tag{3.4b}$$

Substituting the above values of V_n and U_n in Eqs. (2.3), the evolution equations for θ_n and ϕ_n can be written as

$$\frac{d\theta_n}{dt}=2\left(\tan\frac{\theta_{n-1}}{2}\sin\phi_{n-1}-\tan\frac{\theta_{n+1}}{2}\sin\phi_n\right), \tag{3.5a}$$

$$\begin{aligned}\frac{d\phi_n}{dt}=&\frac{2}{\sin\theta_{n+1}}\left(\tan\frac{\theta_{n+2}}{2}\cos\phi_{n+1}+\tan\frac{\theta_n}{2}\cos\phi_n\right)\\&-\frac{2}{\sin\theta_n}\left(\tan\frac{\theta_{n+1}}{2}\cos\phi_n+\tan\frac{\theta_{n-1}}{2}\cos\phi_{n-1}\right).\end{aligned} \tag{3.5b}$$

Defining a new angle Ψ_n as the difference between ϕ_{n-1} and ϕ_n (i.e.),

$$\phi_n=\Psi_{n-1}-\Psi_n, \tag{3.6}$$

Eqs. (3.5) can be rewritten as

$$\frac{d\theta_n}{dt}=-2\tan\frac{\theta_{n+1}}{2}\sin(\Psi_{n-1}-\Psi_n)+2\tan\frac{\theta_{n-1}}{2}\sin(\Psi_{n-2}-\Psi_{n-1}), \tag{3.7a}$$

$$\frac{d\Psi_n}{dt}=2\left(1-\frac{1}{\sin\theta_{n+1}}\tan\frac{\theta_{n+2}}{2}\cos(\Psi_n-\Psi_{n+1})\right)-\left(\frac{2}{\sin\theta_{n+1}}\tan\frac{\theta_n}{2}\cos(\Psi_{n-1}-\Psi_n)\right). \tag{3.7b}$$

In order to indentify Eqs. (3.7) with standard nonlinear differential-difference equations, we make the transformation

$$q_n=\tan\frac{\theta_n}{2}\exp(i\Psi_{n-1}), \tag{3.8}$$

and hence Eqs. (3.7) can be rewritten as

$$i\frac{dq_n}{dt}=(1+|q_n|^2)[q_{n+1}+q_{n-1}]-2q_n. \tag{3.9}$$

Equation (3.9) is the integrable discretization of the completely integrable cubic NLS equation,

$$i\frac{\partial q}{\partial t}+\frac{\partial^2 q}{\partial x^2}+2|q|^2q=0. \tag{3.10}$$

Thus the discrete isotropic classical Heisenberg ferromagnetic spin chain model proposed by Ishimori which is governed by the equation of motion (3.1) is proved to be geometrically equivalent to the integrable discrete NLS equation (3.9) via our mapping procedure for discrete curves. Equation (3.9) can be solved by the inverse scattering transform (IST) method and found to admit N-soliton solutions.[28,29]

B. A discrete higher order spin chain

Another integrable classical discrete Heisenberg ferromagnetic spin chain model with higher order magnetic interactions is described by the equation of motion[30]

$$\begin{aligned}\frac{d\mathbf{S}_n(t)}{dt}&=2a\left[\frac{(\mathbf{S}_n\cdot\mathbf{S}_{n+1})\mathbf{S}_n-\mathbf{S}_{n+1}}{1+\mathbf{S}_n\cdot\mathbf{S}_{n+1}}-\frac{(\mathbf{S}_n\cdot\mathbf{S}_{n-1})\mathbf{S}_n-\mathbf{S}_{n-1}}{1+\mathbf{S}_n\cdot\mathbf{S}_{n-1}}\right]\\&\quad+2b\mathbf{S}_n\wedge\left[\frac{\mathbf{S}_{n+1}}{1+\mathbf{S}_n\cdot\mathbf{S}_{n+1}}+\frac{\mathbf{S}_{n-1}}{1+\mathbf{S}_n\cdot\mathbf{S}_{n-1}}\right],\\ \mathbf{S}_n^2&=1.\end{aligned} \tag{3.11}$$

Here a and b are constant parameters. In Eq. (3.11) higher order interactions are represented by terms proportional to a. It may be noted that in the continuum limit using the continuous space curve mapping procedure and also by the gauge transformation,[30] Eq. (3.11) is shown to be equivalent to the continuous third order NLS equation (Hirota equation),

$$\frac{\partial q}{\partial t}=a\left[\frac{\partial^3 q}{\partial x^3}+6|q|^2\frac{\partial q}{\partial x}\right]+ib\left[\frac{\partial^2 q}{\partial x^2}+2|q|^2q\right]. \tag{3.12}$$

Thus a natural question arises as to see whether using our mapping procedure for discrete curves developed in Sec. II, Eq. (3.11) can be mapped to the integrable discretization of the completely integrable Hirota equation (3.12). For this we map the discrete higher order ferromagnetic spin chain governed by the equation of motion (3.11) onto the discrete moving curve on the sphere and identify the spin vector $\mathbf{S}_n(t)$ with the unit vector $\mathbf{r}_n(t)$ of the discrete curve. Following the procedure adapted in the case of Ishimori model, we finally arrive at the following equation:

$$i\frac{dq_n}{dt}=-ia(1+|q_n|^2)[q_{n+1}-q_{n-1}]+b[(1+|q_n|^2)(q_{n+1}+q_{n-1})-2q_n]. \tag{3.13}$$

Equation (3.13) is the integrable discretization of the Hirota equation (3.12) which is a completely integrable system possessing *N*-soliton solutions.[30] Again when $a=0$, Eq. (3.13) reduces to the integrable discretization of the completely integrable cubic NLS Eq. (3.9).

Thus the dynamics of the discrete higher order ferromagnetic spin model represented by Eq. (3.11) is found equivalent to the integrable discretization of the completely integrable Hirota equation given by Eq. (3.13). As the integrable discrete Eq. (3.13) admits multisoliton solutions the spin excitations can also be expressed in terms of solitons.[30]

C. A discrete generalized Heisenberg spin chain with higher order and next nearest neighbor interactions

Finally, we consider a classical discrete generalized Heisenberg ferromagnetic spin chain characterized by higher order and next nearest neighbor spin–spin interactions in addition to the interaction corresponding to the Ishimori model described by the generalized new Hamiltonian,

$$H=-2(1-2K)\sum_i \log(1+\mathbf{S}_i\cdot\mathbf{S}_{i+1})-4K\sum_i\left[\frac{1+\mathbf{S}_i\cdot\mathbf{S}_{i+1}+\mathbf{S}_i\cdot\mathbf{S}_{i+2}+\mathbf{S}_{i+1}\cdot\mathbf{S}_{i+2}}{(1+\mathbf{S}_i\cdot\mathbf{S}_{i+1})(1+\mathbf{S}_{i+1}\cdot\mathbf{S}_{i+2})}\right]. \tag{3.14}$$

When $K=0$, Hamiltonian (3.14) will reduce to the isotropic bilinear exchange Hamiltonian of the Ishmori model.[28] The equation of motion representing the spin dynamics of the model corresponding to the above Hamiltonian (3.14) can be written as

$$\frac{d\mathbf{S}_n}{dt}=\mathbf{S}_n\wedge[A\mathbf{S}_{n+1}+B\mathbf{S}_{n-1}+C\mathbf{S}_{n+2}+D\mathbf{S}_{n-2}],\quad \mathbf{S}_n^2=1, \tag{3.15a}$$

where *A*, *B*, *C* and *D* are spin functions as given below,

$$A=\left[\frac{2(1-2K)}{(1+\mathbf{S}_n\cdot\mathbf{S}_{n+1})}-\frac{4K}{(1+\mathbf{S}_n\cdot\mathbf{S}_{n+1})^2}\left\{\frac{(\mathbf{S}_n+\mathbf{S}_{n+1})\cdot\mathbf{S}_{n+2}}{(1+\mathbf{S}_{n+1}\cdot\mathbf{S}_{n+2})}+\frac{(\mathbf{S}_n+\mathbf{S}_{n+1})\cdot\mathbf{S}_{n-1}}{(1+\mathbf{S}_n\cdot\mathbf{S}_{n-1})}\right\}\right], \tag{3.15b}$$

$$B=\left[\frac{2(1-2K)}{(1+\mathbf{S}_n\cdot\mathbf{S}_{n-1})}-\frac{4K}{(1+\mathbf{S}_n\cdot\mathbf{S}_{n-1})^2}\left\{\frac{(\mathbf{S}_n+\mathbf{S}_{n-1})\cdot\mathbf{S}_{n-2}}{(1+\mathbf{S}_{n-1}\cdot\mathbf{S}_{n-2})}+\frac{(\mathbf{S}_n+\mathbf{S}_{n-1})\cdot\mathbf{S}_{n+1}}{(1+\mathbf{S}_n\cdot\mathbf{S}_{n+1})}\right\}\right], \tag{3.15c}$$

$$C=\frac{4K}{(1+\mathbf{s}_n\cdot\mathbf{S}_{n+1})(1+\mathbf{S}_{n+1}\cdot\mathbf{S}_{n+2})}, \tag{3.15d}$$

$$D=\frac{4K}{(1+\mathbf{S}_n\cdot\mathbf{S}_{n-1})(1+\mathbf{S}_{n-1}\cdot\mathbf{S}_{n-2})}. \tag{3.15e}$$

When the angle between the orientation of the neighboring spins are maintained small and in the continuum limit, Eq. (3.15) represents the dynamics of spins in a classical Heisenberg ferromagnetic spin chain.[5] For specific values of *K* and lattice parameter, in the small angle and continuum limits, through the methods of continuous space curve mapping and gauge transformation, Eq. (3.15) has been proven to be equivalent to the completely integrable fourth order NLS equation[5] given by

$$i\frac{\partial q}{\partial t}+\frac{\partial^2 q}{\partial x^2}+2|q|^2q+\gamma_1\left[\frac{\partial^4 q}{\partial x^4}+8|q|^2\frac{\partial^2 q}{\partial x^2}+2q^2\frac{\partial^2 q^*}{\partial x^2}+4q\left|\frac{\partial q}{\partial x}\right|^2+6q^*\left(\frac{\partial q}{\partial x}\right)^2+6|q|^4q\right]=0. \tag{3.16}$$

Now we consider the discrete generalized Heisenberg spin chain described by the Hamiltonian (3.14) and the equation of motion (3.15) and see whether Eq. (3.15) can be mapped to the integrable discretization of the complete integrable fourth order NLS equation given by Eq. (3.16). For this as before we rewrite Eq. (3.15) by replacing $\mathbf{S}_n$ by $\mathbf{r}_n$ and by substituting for $\mathbf{r}_{n+1}$, $\mathbf{r}_{n-1}$, $\mathbf{r}_{n+2}$, and $\mathbf{r}_{n-2}$ from Eqs. (2.1), we obtain the velocity field components V_n and U_n in the form

$$V_n=-2(1-2K)\tan\frac{\theta_{n-1}}{2}\sin\phi_{n-1}-2K\left\{\tan\frac{\theta_{n-2}}{2}\sec^2\frac{\theta_{n-1}}{2}\sin(\phi_{n-1}+\phi_{n-2})\right.$$
$$+\tan\frac{\theta_n}{2}\tan^2\frac{\theta_{n-1}}{2}\sin 2\phi_{n-1}+\tan\frac{\theta_{n+1}}{2}\sec^2\frac{\theta_n}{2}\sin\phi_n$$
$$\left.+\tan\frac{\theta_{n-1}}{2}\sec^2\frac{\theta_n}{2}\sin\phi_{n-1}-4\tan\frac{\theta_{n-1}}{2}\sin\phi_{n-1}\right\}, \tag{3.17a}$$

$$U_n=2(1-2K)\left(\tan\frac{\theta_n}{2}-\tan\frac{\theta_{n-1}}{2}\cos\phi_{n-1}\right)-2K\left\{\tan\frac{\theta_{n-2}}{2}\sec^2\frac{\theta_{n-1}}{2}\cos(\phi_{n-1}+\phi_{n-2})\right.$$
$$+\tan\frac{\theta_n}{2}\tan^2\frac{\theta_{n-1}}{2}\cos 2\phi_{n-1}-\tan\frac{\theta_{n+1}}{2}\sec^2\frac{\theta_n}{2}\cos\phi_n$$
$$\left.-\tan\frac{\theta_{n-1}}{2}\sec^2\frac{\theta_n}{2}\cos\phi_{n-1}+3\tan\frac{\theta_n}{2}-2\tan\frac{\theta_{n-1}}{2}\cos\phi_{n-1}\right\}. \tag{3.17b}$$

Substituting Eqs. (3.17) in Eqs. (2.3) and using Eqs. (3.6) and (3.8), the set of coupled equations for θ_n and ϕ_n can be rewritten as a single equation given by

$$i\frac{dq_n}{dt}=(1+|q_n|^2)(q_{n+1}+q_{n-1})-2q_n+K\{(1+|q_n|^2)[(1+|q_{n+1}|^2)q_{n+2}+(1+|q_{n-1}|^2)q_{n-2}$$
$$+(q_{n+1}^2+q_{n-1}^2)q_n^*+(q_{n+1}^*q_{n-1}+q_{n-1}^*q_{n+1})q_n-4(q_{n+1}+q_{n-1})]+6q_n\}. \tag{3.18}$$

Equation (3.18) is the integrable discretized version of the completely integrable fourth order NLS Eq. (3.16). When $K=0$, Eq. (3.18) will reduce to the integrable discrete cubic NLS Eq. (3.9). Equation (3.18) can also be derived through the AKNS formulation (for details see Refs. 29, 31).

For all the above three completely integrable spin models the Lax pair and soliton solutions can be constructed at the discrete level and in the continuum limit in terms of the spin (S_n) and the q_n variables. Hence the elementary spin excitations in all these three models are governed by solitons. More details on this aspect can be found in Refs. 5, 28–31.

IV. CONCLUSIONS

In this paper, we have reformulated the problem of nonlinear spin dynamics of certain discrete classical Heisenberg ferromagnetic spin chain models using a mapping procedure for discrete curves. As the L–L equations that govern the spin dynamics are normally highly nontrivial vector nonlinear partial differential equations, the purpose of the paper was to find equivalent representations for discrete spin models so that the analysis of the spin dynamics becomes easier in the new representation and identification of integrable lattice spin models in a more direct way. Here we specifically considered three one-dimensional discrete Heisenberg spin chain models namely the simple isotropic spin chain with bilinear exchange interaction (Ishimori model), a higher order discrete spin chain, and finally a new generalized spin chain involving next nearest neighbor and higher order spin–spin interactions. Our discrete curve mapping procedure helped to rewrite the L–L equations representing the spin dynamics of the isotropic bilinear chain as equivalent to the integrable discretization of the completely integrable cubic NLS equation. Similarly, the dynamics of the higher order spin chain is equivalently expressed in terms of the integrable discretization of the completely integrable Hirota equation. Finally, the generalized discrete spin chain considered here is found to be equivalent to the integrable discretization of the completely integrable fourth order NLS equation. It was interestingly noted that under the small angle approximation and in the continuum limit the above generalized spin model is found to correspond to the one-dimensional continuum biquadratic Heisenberg spin chain for which the integrability properties are known and the elementary spin excitations are governed by solitons. In all the three models the elementary spin excitations are found to be governed by soliton modes. Thus the discrete curve mapping procedure helps to indentify integrable discrete spin models admitting soliton spin excitations. There are few other ferromagnetic lattice spin models available which are integrable in the continuum limit and not at the discrete level. Our attempt on mapping these discrete spin models (which are not integrable) onto discrete curves on spheres does not fetch equivalent representations in terms of generalized or perturbed discrete NLS family of equations through the transformation (3.8). Hence work is under progress to map these lattice spin chain models to the discrete curve in the space (not on the sphere) and the results will be reported elsewhere.

ACKNOWLEDGMENTS

The work of M.D. forms part of a major research project sponsored by the DST, Government of India. K.M. thanks the Bharathidasan University for the award of a research fellowship.

[1] A. M. Kosevich, B. A. Ivanov, and A. S. Kovalov, Phys. Rep. **194**, 117 (1990).
[2] H. J. Mikeska and M. Steiner, Adv. Phys. **40**, 191 (1990).
[3] M. Lakshmanan, Phys. Lett. **61A**, 53 (1977).
[4] L. A. Takhtajan, Phys. Lett. **64A**, 235 (1977).
[5] K. Porsezian, M. Daniel, and M. Lakshmanan, J. Math. Phys. **33**, 1807 (1992).
[6] K. Nakamura and T. Sasada, Phys. Lett. **48A**, 321 (1974).
[7] M. Daniel, M. D. Kruskal, M. Lakshmanan, and K. Nakamura, J. Math. Phys. **33**, 771 (1992).
[8] R. Balakrishnan, Phys. Lett. **92A**, 243 (1982); J. Phys. C **15C**, L1305 (1982).
[9] R. Ferrer and J. Pozo, J. Phys.: Condens. Matter **3**, 3527 (1991).
[10] R. Myrzakulov, M. Daniel, and R. Amuda, Physica (Utrecht) **234**, 715 (1997).
[11] M. Daniel and R. Amuda, J. Phys. A **28**, 5529 (1995).
[12] M. Daniel and R. Amuda, Phys. Rev. B **53**, R2930 (1996).
[13] V. G. Dubrovsky and B. G. Konopelchenko, Physica D **48D**, 367 (1991); **55D**, 1 (1992).
[14] M. Lakshmanan, Th. W. Ruijgrok, and C. J. Thompson, Physica A **84**, 577 (1976).
[15] V. E. Zakharov and L. A. Takhtajan, Theor. Math. Phys. **38**, 17 (1979).
[16] A. Kundu, J. Math. Phys. **25**, 3439 (1984).
[17] R. Amuda, ''Nonlinear excitations in certain ordered magnetic spin systems,'' Ph.D. thesis, Bharathidasan University, 1996 (unpublished).
[18] H. Hasimoto, J. Fluid Mech. **51**, 477 (1972).
[19] G. L. Lamb, Jr., J. Math. Phys. **18**, 1654 (1977).
[20] R. E. Goldstein and D. M. Petrich, Phys. Rev. Lett. **67**, 3203 (1991); **69**, 555 (1992).
[21] K. Nakayama, H. Segur, and M. Wadati, Phys. Rev. Lett. **69**, 2603 (1992).
[22] A. Doliwa and P. M. Santini, Phys. Lett. A **185A**, 373 (1994).
[23] L. Uby, Proc. R. Soc. London, Ser. A **452**, 1531 (1996).
[24] K. Porsezian, Phys. Rev. E **55**, 3785 (1997).
[25] A. Doliwa and P. M. Santini, J. Math. Phys. **36**, 1259 (1995).
[26] M. J. Ablowitz and J. F. Ladik, J. Math. Phys. **17**, 1011 (1976).
[27] M. Daniel and K. Manivannan, Phys. Rev. B **57**, 1 (1998).
[28] Y. Ishimori, J. Phys. Soc. Jpn. **51**, 3417 (1982).
[29] M. J. Ablowitz and P. A. Ckarkson, *Solitons, Nonlinear Evolution Equations, and Inverse Scattering* (Cambridge University Press, Cambridge, 1992).
[30] K. Porsezian and M. Lakshmanan, Inverse Probl. **5**, L15 (1989).
[31] M. Daniel and M. M. Latha, Physica A **240**, 526 (1997).

The geometry of coherent states

Timothy R. Field
DERA Malvern, WR14 3PS, United Kingdom

Lane P. Hughston
King's College London, The Strand, London WC2R 2LS, United Kingdom

(Received 11 September 1998; accepted for publication 23 November 1998)

We examine the geometry of the state space of a relativistic quantum field. The mathematical tools used involve complex algebraic geometry and Hilbert space theory. We consider the Kähler geometry of the state space of any quantum field theory based on a linear classical field equation. The state space is viewed as an infinite-dimensional complex projective space. In the case of boson fields, a special role is played by the coherent states, the totality of which constitutes a nonlinear submanifold $\mathcal{C}$ of the projective Fock space $P\mathcal{F}$. We derive the metric on $\mathcal{C}$ induced from the ambient Fubini–Study metric on $P\mathcal{F}$. Arguments from differential geometric, algebraic, and Kählerian points of view are presented, leading to the result that the induced metric is flat, and that the intrinsic geometry of $\mathcal{C}$ is Euclidean. The coordinates for the single-particle Hilbert space of solutions are shown to be complex Euclidean coordinates for $\mathcal{C}$. A transversal intersection property of complex projective lines in $P\mathcal{F}$ with $\mathcal{C}$ is derived, and it is shown that the intrinsic geodesic distance between any two coherent states is strictly greater than the corresponding geodesic distance in the ambient Fubini–Study geometry. The functional metric norm of a difference field is shown to give the intrinsic geodesic distance between two coherent states, and the metric overlap expression is shown to measure the angle subtended by two coherent states at the vacuum, which acts as a preferred origin in the Euclidean geometry of $\mathcal{C}$. Using the flatness of $\mathcal{C}$ we demonstrate the relationship between the manifold complex structure on $P\mathcal{F}$ and the quantum complex structure viewed as an active transformation on the single-particle Hilbert space. These properties of $\mathcal{C}$ hold independently of the specific details of the single-particle Hilbert space. We show how $\mathcal{C}$ arises as the affine part of its compactification obtained by setting the vacuum part of the state vector to zero. We discuss the relationship between unitary orbits and geodesics on $\mathcal{C}$ and on $P\mathcal{F}$. We show that for a Fock space in which the expectation of the total number operator is bounded above, the coherent state submanifold is Kähler and has finite conformal curvature. [S0022-2488(99)00506-X]

I. INTRODUCTION

The existence of complex structures in quantum theory, and their possible role in a theory of quantum gravity, is a subject of much interest. Such structures are essential mechanics to express the idea that the state vector evolves unitarily according to appropriate dynamical equations. One of the main applications of complex structures in physics recently has been the subject of *geometric quantum mechanics*, an area that has been developed by a number of authors (Refs. 1–11). In this picture the natural geometry of the quantum phase space is shown to be characterized not merely by a symplectic structure, but also by a compatible complex structure and a Riemannian metric, from which one derives the concepts of quantum mechanical uncertainty and expectation.

In this paper we explore geometric aspects of the complex structures on function spaces that arise in quantum theory. By a complex structure J we mean a linear operator acting on the Hilbert space V of solutions to a linear field equation, satisfying $J^2 = -1$. The operator J is an endomorphism of V, and if we complexify V to define the single-particle Hilbert space $\mathcal{H}$ then J has eigenspaces with eigenvalues $\pm\sqrt{-1}$, which we refer to as the positive and negative frequency single-particle Hilbert spaces. From this $\mathcal{H}$ we define the Fock space $\mathcal{F}$ via tensor products of $\mathcal{H}$ with itself. In addition to the complex structure J we also have the quantum mechanical symplectic form Ω and the real Hilbert space metric g, all of which are mutually compatible. Then the triple $\{J,\Omega,g\}$ constitutes a *quantum Kähler structure*. Such structures have received much attention in the literature, and form the basis of the relationship between classical and quantum field theories. On a classical level, the J operation bridges the symplectic and metric structures, and for real boson fields enables one to obtain a classical expression for the norms of fields. Considerations of quantum field theory and the introduction of Planck's constant then show that the squared norm is equal to the expectation of the quantum number operator in a corresponding coherent state (see, e.g., Ref. 11).

In classical Hamiltonian mechanics the symplectic structure Ω, when taken together with a preferred Hamiltonian function H on the phase space, determines the system completely. It is in the passage to quantum theory that additional structure is required, namely, the quantum complex structure. When compatibility of the elements of the quantum Kähler structure is assumed, any two of its elements imply the other, and thus quantization can be viewed as the addition of either J or g to the classical system. In geometric quantum mechanics it is the complex projective space CP^n, the quantum mechanical state space, that replaces the classical phase space. This is a Kähler manifold, endowed with the Fubini–Study metric. It is also a symplectic manifold, and thus the geometry of Hamiltonian classical mechanics carries over to the quantum state space. In particular, the equation governing the evolution of a quantum mechanical state, Schrödinger's equation, becomes Hamilton's equations of classical mechanics. In other words, Schrödinger trajectories are given by the orbits of a Hamiltonian vector field on the quantum state space. The extra ingredient of quantum theory is the quantum metric g, which takes the form of a real Riemannian structure on the state manifold, enabling one to calculate quantum mechanical transition amplitudes and uncertainties.

A further modification that quantum theory entails is that uncertainty terms appear in the Hamiltonian function of the system, in addition to the function obtained by replacing classical phase space coordinates by the expectations of their associated operators. The harmonic oscillator is an example of this phenomenon, for which the quantum Hamiltonian function is equal to $x^2+p^2+(\Delta x)^2+(\Delta p)^2$. In this context we are naturally led to consider the *coherent states* (Refs. 12–14; cf. also Refs. 15, 30), which are in some respects classical in behavior. These states saturate the quantum mechanical uncertainty inequality $\Delta x\,\Delta p \geqslant \hbar/2$, and their quantum mechanical evolution corresponds closely to classical evolution.

In this paper we address various aspects of the quantum Kähler structure, with an emphasis on the geometric features of Fock space. Our main results are as follows. We begin by developing geometric quantum mechanics for relativistic quantum field theory, thus extending existing results in the literature to projective Fock space $P\mathcal{F}$. We see how this applies to coherent states and derive geometric properties of the coherent state submanifold $\mathcal{C}$. In particular, in Theorem 1 we show that the intrinsic geometry of $\mathcal{C}$ is complex Euclidean, and that the single-particle state vectors serve as complex Euclidean coordinates for this submanifold. This enables one to deduce other properties of $\mathcal{C}$, and provides a relationship between the complex structure on the single-particle Hilbert space $\mathcal{H}$ and that on the ambient state manifold. We discuss the geodesic and unitary orbits of $\mathcal{C}$, observing that $\mathcal{C}$ and the state space $P\mathcal{F}$ share no geodesics, and that the geodesic orbits of $\mathcal{C}$ are nonunitary. We discuss the corresponding situation for truncated Fock spaces, for which the expectation of the total number operator is bounded above. In particular, we deduce that the associated submanifold of coherent states has finite conformal curvature.

We conclude with a discussion of some physical issues. We indicate the relationship between the theory of positive operator measures (POMs) and coherent states in the context of our results, and in this connection remark on prospects for a model of stochastic state vector reduction.

In the generation of a quantum Fock space, we consider the Hilbert spaces of real and complex solutions to a set of linear classical field equations. In the former case, there is only one available complex structure, since this maps real solutions of the field equations to themselves. In terms of the Fourier transform $\hat{\varphi}(k)$ of the field, multiplication by a complex number $\alpha=a+ib$ is replaced by the action of the operator $a+Jb$ on φ, where J acts linearly and multiplies the positive and negative frequency parts of the field by $+i$ and $-i$, respectively. We say that J is the *quantum complex structure* acting on the Hilbert space. In the case of a complex-valued field there are two possible complex structures, namely (a) the one given above, and (b) multiplication of $\hat{\varphi}$ by the complex number α. It is only the former choice that leads to positive values for the expectation of energy (Ref. 16).

II. THE QUANTUM KÄHLER STRUCTURE

Let V denote the Hilbert space of normalizable real solutions to a classical linear field equation, and define the single-particle Hilbert space $\mathcal{H}=V\otimes C$. We adopt the notation $\mathcal{H}^n\equiv\mathcal{H}^{\otimes n}$ for the n-fold tensor product of $\mathcal{H}$ with itself, where $\otimes$ denotes the symmetric tensor product for bosons and antisymmetric tensor product for fermions. Then $\mathcal{H}^n$ is said to be the n-particle Hilbert space, and we define Fock space as the Hilbert space $\mathcal{F}=C\oplus\mathcal{H}\oplus\mathcal{H}^2\oplus\cdots\oplus\mathcal{H}^n\oplus\cdots$ (see e.g., Ref. 17 for various technical details of this construction, which are not as relevant for the present discussion). We introduce the *quantum complex structure J*, a linear map $J:V\rightarrow V$ such that $J^2\equiv-\mathbf{1}$. Then, if we extend V to $V\otimes C$, the map J admits eigenspaces with eigenvalues $+i$ and $-i$ that we call the spaces of positive and negative frequency solutions, denoted by $\mathcal{H}_+$ and $\mathcal{H}_-$, respectively, so $\mathcal{H}=\mathcal{H}_+\oplus\mathcal{H}_-$. The spaces $\mathcal{H}_\pm$ give rise naturally to the positive and negative frequency Fock spaces $\mathcal{F}_\pm$ according to the scheme $\mathcal{F}_\pm=C\oplus\mathcal{H}_\pm\oplus\mathcal{H}_\pm^2\oplus\cdots\oplus\mathcal{H}_\pm^n\oplus\cdots$. We use an abstract index notation (Ref. 18) for elements of $\mathcal{H}$, so for a typical element of $\mathcal{H}$ we write ξ^a. See Refs. 16, 19, and 20 for more details of the abstract index notation in a Hilbert space context. We take $\mathcal{H}$ to have a countably infinite basis, so the index on ξ can be thought of as running over the natural numbers. There are some technicalities associated with the fact that one deals with an infinite-dimensional Fock space, which is itself built up from an infinite-dimensional single particle Hilbert space $\mathcal{H}^1$. However, in practice it is reasonable to assume that the underlying single-particle Hilbert space is separable[21] so any vector can be decomposed along countably many basis states, as occurs in the Fourier series analysis of an oscillator with boundary conditions. In the case of a state not built up in this way one usually argues via continuity in the relevant function space (for further discussion of these issues, see Ref. 22).

We now develop in further detail the abstract index notation. For a positive frequency field we use an unprimed Greek index, so, for example, $\xi^\alpha\in\mathcal{H}_+^1$; for a negative frequency field we use a primed Greek index. Then we can regard the italic index a as "composite" by writing $a\equiv\alpha,\alpha'$, where ξ^α and $\xi^{\alpha'}$ have eigenvalues $+i$ and $-i$ under the action of J, respectively, so $J^\alpha{}_\beta\xi^\beta=i\xi^\alpha$ and $J^{\alpha'}{}_{\beta'}\xi^{\beta'}=-i\xi^{\alpha'}$. A general complex field ξ decomposes into independent positive and negative frequency parts $\xi^a\equiv\xi^{\alpha,\alpha'}=\xi^\alpha\oplus\xi^{\alpha'}$. We refer to ξ^α and $\xi^{\alpha'}$, respectively, as the positive and negative frequency parts of the complex field ξ^a. For all ϕ^α the action of complex conjugation will be denoted $\phi^\alpha\mapsto\overline{\phi^\alpha}=\bar{\phi}^{\alpha'}$, and thus it interchanges the index type. For a real field we have $\phi^\alpha=\overline{\phi^{\alpha'}}$ and thus $\phi^a=\phi^\alpha\oplus\bar{\phi}^{\alpha'}$, indicating that in this case the positive and negative frequency parts are complex conjugates of each other.

We can also represent the metric, the symplectic structure, and the complex structure in matrix form according to this scheme. For the metric we have

$$g_{ab}=\begin{pmatrix} 0 & g_{\alpha\beta'} \\ g_{\alpha'\beta} & 0 \end{pmatrix}, \tag{2.1a}$$

in which the tensor $g_{\alpha\beta'}$ is Hermitian (or real) in a sense explained below. Note that, as in the case of the algebra of two-component spinors, a primed index "commutes" with any unprimed index, so that for any multivalence tensor we have $T_{\alpha\cdots\delta\rho'\cdots\tau'}\equiv T_{\rho'\cdots\tau'\alpha\cdots\delta}$. Now complex conjugation interchanges index type according to the scheme $\overline{T_{\alpha\beta\cdots\delta\rho'\sigma'\cdots\tau'}}=\overline{T}_{\alpha'\beta'\cdots\delta'\rho\sigma\cdots\tau}$. Again, as in the case of two-component spinors, a tensor can only be Hermitian if it has equal numbers of primed and unprimed indices, in which case the Hermitian condition is $\overline{T_{\alpha\beta\cdots\gamma\alpha'\beta'\cdots\gamma'}} = T_{\alpha\beta\cdots\gamma\alpha'\beta'\cdots\gamma'}$. For the metric $g_{\alpha\beta'}$, in particular, we have $\overline{g_{\alpha\beta'}} \equiv \bar{g}_{\alpha'\beta} = g_{\alpha'\beta} \equiv g_{\beta\alpha'}$. Similarly, we write the symplectic structure

$$\Omega_{ab} = \begin{pmatrix} 0 & ig_{\alpha\beta'} \\ -ig_{\alpha'\beta} & 0 \end{pmatrix}. \tag{2.1b}$$

This is compatible with the complex structure in that $\Omega(\phi,\psi) = \Omega(J[\phi],J[\psi])$. The action of the metric is given by $g(\phi,\psi) = g_{ab}\phi^a\psi^b = g_{\alpha\beta'}\phi^\alpha\psi^{\beta'} + g_{\alpha'\beta}\phi^{\alpha'}\psi^\beta$, and thus $g(\phi,\psi)$ is real for all real fields ϕ and ψ. The action of the symplectic structure is given by $\Omega(\phi,\psi) = ig_{\alpha\beta'}\phi^\alpha\psi^{\beta'} - ig_{\alpha'\beta}\phi^{\alpha'}\psi^\beta$. We write the quantum complex structure in matrix form as

$$J^a{}_b = \begin{pmatrix} i\delta^\alpha_\beta & 0 \\ 0 & -i\delta^{\alpha'}_{\beta'} \end{pmatrix}. \tag{2.1c}$$

The Dirac or ''quantum mechanical'' scalar product between the fields ϕ and ψ is defined by $F(\phi,\psi) \equiv \frac{1}{2}(g+i\Omega)(\phi,\psi) = g_{\alpha'\beta}\phi^{\alpha'}\psi^\beta$ and is Hermitian in the sense that $F(\phi,\psi) = \overline{F(\psi,\phi)}$ for all real fields ϕ and ψ. We can use $g_{\alpha'\beta}$ to lower indices by writing $\phi_\alpha = g_{\alpha\beta'}\phi^{\beta'}$, and $\phi_{\alpha'} = g_{\alpha'\beta}\phi^\beta$. The complex structure satisfies $J^a{}_cJ^c{}_b = -\delta^a_b$. The compatibility of the set $\{g,\Omega,J\}$ can be represented by the following conditions:

$$g_{ac} = \Omega_{ab}J^b{}_c, \tag{2.2a}$$

$$g_{ab}J^a{}_cJ^b{}_d = g_{cd}, \tag{2.2b}$$

$$\Omega_{ab}J^a{}_cJ^b{}_d = \Omega_{cd}, \tag{2.2c}$$

$$g^{ap}g^{bq}\Omega_{ab} = \Omega^{pq}. \tag{2.2d}$$

Condition (2.2a) is the expression for the positive definite quantum metric in terms of the symplectic form and complex structure, i.e., $g(\phi,\psi) \equiv \Omega(\phi,J\psi)$. A straightforward consequence is that $g(\phi,J\phi) \equiv 0$ for any ϕ. Condition (2.2b) states that the metric g is compatible with the complex structure J, i.e., that the metric g is Hermitian (Ref. 23), so $g(\phi,\psi) \equiv g(J\phi,J\psi)$. We note that the term ''Hermitian'' is used conventionally in two related but rather different ways. First, it can be used to indicate a reality condition on noncomposite tensors of equally mixed rank, as indicated above, i.e., for tensors belonging to the subspace $\mathcal{H}^n_+ \otimes \mathcal{H}^n_-$. On the other hand, we say that a tensor of general even rank, belonging to $\mathcal{H}^{2n}$, is Hermitian if $J^p_aJ^q_b\cdots J^r_cJ^s_dT_{pq\cdots rs} = T_{ab\cdots cd}$. The context will usually indicate in exactly which sense the term is being applied. Thus (2.2c) states that Ω is Hermitian in the latter sense. Some authors use a convention for the symplectic tensor Ω_{ab} that differs by a factor of 2 from the one used here (e.g., Ref. 24). We adopt conventions such that the tensor Ω with its indices raised according to (2.2d) above is the inverse to Ω_{ab}, so $\Omega^{ab}\Omega_{cb} = \delta^a_c$. We refer to the compatible set $\{g,\Omega,J\}$ as the *quantum Kähler structure* (cf. Ref. 25). Observe that condition (2.2a) above implies that any two of $\{g,\Omega,J\}$ are sufficient to determine the quantum Kähler structure.

We now proceed briefly to examine various operations on Fock space in this spirit. A state vector in Fock space $\mathcal{F}$ can be written $|\xi\rangle = (\xi,\xi^\alpha,\xi^{\alpha\beta},\ldots)$, where $\xi^\alpha \in \mathcal{H}^1$, $\xi^{\alpha\beta} \in \mathcal{H}^2$, and so on. The evaluation of the squared Hilbert space norm of a vector $|\xi\rangle$ in $\mathcal{F}$ is given by $\|\xi\|^2 = \xi\bar{\xi} + \xi^\alpha\bar{\xi}_\alpha + \xi^{\alpha\beta}\bar{\xi}_{\alpha\beta} + \cdots$. For any $\sigma^\alpha \in \mathcal{H}^1$ we define the annihilation and creation operators $\hat{A}^\alpha$ and $\hat{C}_\alpha$ according to the prescription

$$\hat{A}^{\alpha}\overline{\sigma}_{\alpha}|\xi\rangle=\hat{A}(\overline{\sigma})|\xi\rangle=(\xi^{\mu}\overline{\sigma}_{\mu},\sqrt{2}\xi^{\mu\alpha}\overline{\sigma}_{\mu},\sqrt{3}\xi^{\mu\alpha\beta}\overline{\sigma}_{\mu},\ldots),\tag{2.3a}$$

$$\hat{C}_{\alpha}\sigma^{\alpha}|\xi\rangle=\hat{C}(\sigma)|\xi\rangle=(0,\sigma^{\alpha}\xi,\sqrt{2}\sigma^{(\alpha}\xi^{\beta)},\sqrt{3}\sigma^{(\alpha}\xi^{\beta\gamma)},\ldots).\tag{2.3b}$$

These obey the commutation relations $[\hat{C}(\sigma),\hat{C}(\sigma')]=0$, $[\hat{A}(\overline{\sigma}),\hat{A}(\overline{\sigma}')]=0$, and $[\hat{A}^{\alpha},\hat{C}_{\beta}]=\delta^{\alpha}_{\beta}$ or, equivalently, $[\hat{A}(\overline{\sigma}),\hat{C}(\sigma)]=(\sigma\cdot\overline{\sigma})I$. These operators are adjoints of each other in the sense that for any vector $\phi\in\mathcal{F}$ we have $\langle\hat{C}(\sigma)\psi,\phi\rangle=\langle\psi,\hat{A}(\overline{\sigma})\phi\rangle$.

We then define the number operator associated with $\sigma\in\mathcal{H}^1$ $(\sigma\neq 0)$ by $\hat{N}(\sigma)=\|\sigma\|^{-2}\hat{C}(\sigma)\hat{A}(\overline{\sigma})$, from which it follows that $\hat{N}(\sigma)\xi=\|\sigma\|^2(0,\sigma^{\alpha}\xi^{\mu}\overline{\sigma}_{\mu},2\sigma^{(\alpha}\xi^{\beta)\mu}\overline{\sigma}_{\mu},\ldots)$. Summing over an orthonormal basis $\{\sigma^{\alpha}\}$ using the identity $\Sigma_N\sigma^{\alpha}\overline{\sigma}_{\mu}=\delta^{\alpha}_{\mu}$ we obtain the total number operator $\hat{N}$, given by $\hat{N}\xi=(0,\xi^{\alpha},2\xi^{\alpha\beta},3\xi^{\alpha\beta\gamma},\ldots)$. The number operators satisfy the following commutation relations: $[\hat{N}(\sigma),\hat{C}(\sigma)]=[\hat{N},\hat{C}(\sigma)]=\hat{C}(\sigma)$, $[\hat{N}(\sigma),\hat{A}(\overline{\sigma})]=[\hat{N},\hat{A}(\overline{\sigma})]=-\hat{A}(\overline{\sigma})$, and $[\hat{N}(\sigma),\hat{N}]=0$. For further details of these and other relations see Refs. 16 and 19.

III. COHERENT STATES

There are various characterizations of coherent states in quantum field theory. First, we give a definition via exponentiation of the single-particle Hilbert space $\mathcal{H}^1$. We begin with a vector $\xi^a\in\mathcal{H}^1$ and construct from this a unique element of $\mathcal{F}$, obtained by exponentiating ξ^a, denoted $|\xi_c\rangle$. Explicitly, we have

$$\xi^{\alpha}\mapsto\mathcal{E}(\xi^{\alpha})=\exp(\xi^{\alpha}\hat{C}_{\alpha})|0\rangle(1,\xi^{\alpha},\xi^{\alpha}\xi^{\beta}/\sqrt{2!},\ldots,\xi^{\alpha}\xi^{\beta}\cdots\xi^{\delta}/\sqrt{n!},\ldots)=:|\xi_c\rangle\in\mathcal{F},\tag{3.1}$$

where the term containing $\sqrt{n!}$ has n indices [cf. the discussion of Perelomov's ''generalized coherent states'' in proof (b) of Theorem 1 below]. The element $|\xi_c\rangle$ is said to be a coherent state vector. Now we introduce the projection operator P from the Fock space $\mathcal{F}$ down to the state space $P\mathcal{F}$ so the totality of coherent states form a submanifold $\mathcal{C}$ of $P\mathcal{F}$. It is important to note that if $\xi^a\neq 0$ then $P\circ\mathcal{E}(\lambda\xi^a)=P\circ\mathcal{E}(\mu\xi^a)$ if and only if $\lambda=\mu$. Note, however, that although $\lambda\xi^a,\mu\xi^b$ define different vectors in $\mathcal{H}^1$ for $\lambda\neq\mu$, they define the same single-particle state for all nonzero λ, μ, because the single-particle states are elements of $P\mathcal{H}^1$, not $\mathcal{H}^1$. Therefore changing the phase or scale of a single-particle state vector changes the associated coherent state. We can consider the *universal bundle* $\mathcal{U}$ over the single-particle state space with projection $\Pi:\mathcal{U}\to P\mathcal{H}^1$, defined so the fiber above any point or state is the ray in the Hilbert space that it represents. The map $\mathcal{E}:\mathcal{U}\to\mathcal{F}$ defined in (3.1) is a map from this bundle to Fock space, and this is nonconstant along the fibers $\Pi^{-1}(s)$ for all $s\in P\mathcal{H}^1$.

The action of the creation and annihilation operators on coherent states is as follows. Let τ^{α} be an element of $\mathcal{H}^1_+$ and $|\psi_c\rangle$ a coherent state vector defined as before. Then from (2.3a) we have

$$\hat{A}(\overline{\tau})|\psi_c\rangle=(\xi\cdot\overline{\tau})|\psi_c\rangle,\tag{3.2a}$$

or, equivalently, $\hat{A}^{\alpha}|\psi_c\rangle=\psi^{\alpha}|\psi_c\rangle$, from which it follows that *coherent states are eigenstates of the annihilation operator* $\hat{A}(\overline{\tau})$ for any vector $\tau^a\in\mathcal{H}^1$. On the other hand, from (2.3b) it follows that the action of the creation operator on a coherent state vector is given by differentiation with respect to $\mathcal{H}^1$, so

$$\hat{C}_{\alpha}|\psi_c\rangle=\frac{d|\psi_c\rangle}{d\psi^{\alpha}}.\tag{3.2b}$$

For convenience we set $\Lambda:=\xi^{\alpha}\overline{\xi}_{\alpha}$. Then $\langle\psi_c|\psi_c\rangle=e^{\Lambda}$, so we have to divide by this factor to calculate the expectation of any operator from its matrix element with a coherent state vector. For the expectation of the number operator $\hat{N}(\sigma)$ in a coherent state $P|\xi_c\rangle$ we obtain $\langle\hat{N}(\sigma)\rangle=\|\sigma\|^{-2}(\sigma\cdot\overline{\xi})(\xi\cdot\overline{\sigma})$, and for the total number operator $\hat{N}$ we have $\langle\hat{N}\rangle=\Lambda$. More generally, we observe that for any vector $\xi^a \in \mathcal{H}^1$ with norm $\Lambda = \xi^\alpha \xi_\alpha$ the probability distribution for the total number of particles in the coherent state associated with ξ^a is $\mathrm{Prob}[\hat{N}=n] = \Lambda^n e^{-\Lambda}/n!$, the *Poisson distribution* with mean and variance Λ. By a resolution of the identity we mean an expansion of the unit operator of the form

$$\int_{t\in\mathcal{A}} p_t |c_t\rangle\langle c_t| = 1, \tag{3.3}$$

where t is an element of some (generally multidimensional) index set $\mathcal{A}$, endowed with a notion of continuity, and p_t is positive. In the case of coherent states the choice $\mathcal{A}=\mathcal{H}$ determines p_t uniquely (see Ref. 14). This uniqueness property holds, even though the coherent states are not mutually orthogonal and form an over complete basis [in our Hilbert space notation $\langle \xi_c | \psi_c \rangle = \exp(\psi\cdot\bar{\xi})$, which is never zero]. The resolution of the identity (3.3) leads to an important concept in quantum measurement theory and quantum optics known as a *positive operator measure* (POM). We return to a discussion of POMs in Sec. VI.

In the case of the quantum harmonic oscillator the space of coherent states evolves into itself under the unitary evolution associated with the Hamiltonian operator $\hat{H}$. The orbits in classical phase space are identical to the quantum mechanical orbits in the ‘‘expectation phase space’’ coordinatized by the expectation values of the position and momentum operators $\langle \hat{Q} \rangle$ and $\langle \hat{P} \rangle$. In this sense the coherent state description of the quantum harmonic oscillator is suggestively classical. Moreover, the uncertainty relation $(\Delta\hat{Q})(\Delta\hat{P}) \geqslant \hbar/2$ is saturated by the coherent states, and $(\Delta\hat{Q}),(\Delta\hat{P})$ remain constant under unitary Hamiltonian time evolution. These properties hold in the more general case of a system of coupled harmonic oscillators (see Ref. 13).

Now consider the fiber bundle $\mathcal{V}=\Gamma\times F$, where Γ is the classical phase space of elementary Hamiltonian mechanics, and the fiber F above any point $(x,p)\in\Gamma$ is the set of quantum mechanical states $P|\psi\rangle$ such that $\langle\psi|(\hat{X},\hat{P})|\psi\rangle = (x,p)$. In geometric quantum mechanics the evolution can be viewed as taking place in $\mathcal{V}$. The quantum Hamiltonian operator $\hat{H}$ is obtained by promoting x and p to the corresponding position and momentum operators $\hat{X}$ and $\hat{P}$, so that $\hat{H}=\hat{X}^2 + \hat{P}^2$ in appropriate physical units. However, the quantum Hamiltonian function $h=\langle\psi|\hat{H}|\psi\rangle$ is equal to $[x^2+p^2]+[(\Delta\hat{X})^2+(\Delta\hat{P})^2]$. Let us label the two bracketed terms h_0, h_Δ, respectively. The term h_Δ comes from the quantum metric g of Sec. I and is the essential extra ingredient in quantum theory. We can separate the Hamiltonian vector field $X_h^a = \Omega^{ab}\nabla_b h$ uniquely into horizontal and vertical parts X_0, X_Δ, respectively, in the bundle $\mathcal{V}$. For the quantum harmonic oscillator the projection to Γ of any trajectory in $\mathcal{V}$ is always the classical orbit in Γ, corresponding to X_0. The coherent states are characterized by the fact that the Heisenberg inequality $(\Delta x)(\Delta p) \geqslant \hbar/2$ is saturated, and this fixes the values of $\Delta x, \Delta p$. The coherent state evolution has $X_\Delta = 0$, and thus the trajectory is purely horizontal in $\mathcal{V}$ (cf. Ref. 22).

In the case of the quantum electrodynamics of a free photon field, the quantum field is essentially an infinite collection of harmonic oscillators (see Ref. 26), and the properties of coherent states apply. Thus, a ‘‘classical’’ state of the quantum electromagnetic field can be represented by a coherent state in $P\mathcal{F}$, where the Fock space $\mathcal{F}$ is built up in the standard way from $\mathcal{H}_{\mathrm{M}}$, the space of square integrable real solutions of Maxwell's equations. The creation and annihilation operators are based on the electromagnetic 4-potential A^μ, and the field operator is defined as $\hat{A}+\hat{A}^\dagger$. The expectation of this operator in a coherent state $|\xi_c\rangle$ is the corresponding ‘‘ancestor’’ classical solution of Maxwell's equations $\xi^a \in \mathcal{H}_{\mathrm{M}}$ (cf. the discussion of Sec. VI).

IV. THE FUBINI–STUDY GEOMETRY

The projective form of the Fubini–Study metric on CP^n that one usually encounters in quantum theory (see, e.g., Refs. 23, 27, and 7) can be written in the elegant form

$$ds^2 = \frac{8Z^\alpha\, dZ^\beta \overline{Z}_{[\alpha}\, d\overline{Z}_{\beta]}}{k(Z^\gamma \overline{Z}_\gamma)^2}, \tag{4.1a}$$

where Z^α are homogeneous coordinates for CP^n with $\alpha = 0,1,...,n$, and k is the holomorphic sectional curvature, which in subsequent calculations we take to equal one. In application to coherent states it will be useful also to have this metric expressed in nonhomogeneous coordinates, since if Z^0 is the coordinate of the vacuum part of a coherent state then it is necessarily nonvanishing. Thus, the coherent states form a submanifold $\mathcal{G}$ of the affine part of the projective Fock space, the latter consisting of elements of the form $\{(1,\xi^a,\xi^{ab},\ldots)\} =: \mathcal{A} \supset \mathcal{C} \cong C^\infty$ and whose compactification is $\{P(0,\psi^a,\psi^{ab},\ldots)\} =: \mathcal{B} \cong CP^\infty$, i.e., states for which the probability of no quanta being present is zero. Observe that the image of any state vector $|\psi\rangle$ under the creation operator $\hat{C}(\sigma)$ lies within the compactification, i.e., $\hat{C}(\sigma)|\psi\rangle \in \mathcal{B}_+$, $\forall|\psi\rangle \in \mathcal{F}_+$, $\sigma^\alpha \in \mathcal{H}^1_+$. From now on we deal with coherent states viewed as forming a submanifold of $\mathcal{A}$ and we shall make use of the following result (Ref. 23).

Lemma 2: The nonprojective form of the Fubini–Study metric on CP^n is given by

$$ds^2 = 4\,\frac{(1+\zeta^\alpha \overline{\zeta}_\alpha) d\zeta^\alpha\, d\overline{\zeta}_\alpha - (\zeta^\alpha\, d\overline{\zeta}_\alpha)(\overline{\zeta}_\alpha\, d\zeta^\alpha)}{k(1+\zeta^\gamma \overline{\zeta}_\gamma)^2}, \tag{4.1b}$$

where $\zeta^\alpha = Z^\alpha/Z^0$, $\alpha = 1,2,...,n$ are inhomogeneous coordinates.

Proof: We have $d\zeta^\alpha = (Z^0\, dZ^\alpha - Z^\alpha\, dZ^0)/(Z^0)^2$, together with its complex conjugate, and thus $d\zeta^\alpha\, d\overline{\zeta}_\alpha = (Z^0 \overline{Z}_0\, dZ^\alpha\, d\overline{Z}_\alpha - \overline{Z}_0 Z^\alpha\, dZ^0\, d\overline{Z}_\alpha - Z^0 \overline{Z}_\alpha\, d\overline{Z}_0\, dZ^\alpha + Z^\alpha \overline{Z}_\alpha\, dZ^0\, d\overline{Z}_0)/(Z^0 \overline{Z}_0)^2$. Also, $\overline{\zeta}_\alpha\, d\zeta^\alpha = [1/Z_0(Z^0)^2](Z^0 \overline{Z}_\alpha\, dZ^\alpha - Z^\alpha \overline{Z}_\alpha\, dZ^0)$ and similarly for its complex conjugate. Straightforward algebra then shows that the numerator of the right-hand side of (4.1b) is equal to $[(Z^\alpha \overline{Z}_\alpha) \times (dZ^\alpha\, d\overline{Z}_\alpha) + (Z^0 \overline{Z}_0)(dZ^\alpha\, d\overline{Z}_\alpha) - (Z^\beta\, d\overline{Z}_\beta)(dZ^\alpha\, \overline{Z}_\alpha) - (Z^0\, d\overline{Z}_0)(dZ^\alpha\, \overline{Z}_\alpha)]/(Z^0 \overline{Z}_0)^2$. Further manipulation shows that the bracket in the above expression is equal to $(Z^\alpha \overline{Z}_\alpha)(dZ^\alpha\, d\overline{Z}_\alpha) - (Z^\alpha\, d\overline{Z}_\alpha)(dZ^\alpha\, \overline{Z}_\alpha) \equiv 2Z^{[\alpha}\, dZ^{\beta]}\, \overline{Z}_\alpha\, d\overline{Z}_\beta$, and dividing through by $(1+\zeta^\alpha \overline{\zeta}_\alpha)^2$ completes the proof. ■

We shall use this lemma to provide a differential geometric proof of Theorem 1 below, which concerns the intrinsic geometry of the submanifold of coherent states. Before stating this result, we make some general remarks on the curvature tensor of the Fubini–Study geometry and of Kähler manifolds in general.

Consider again the Fubini–Study line element (4.1a). This provides a one-parameter family of Fubini–Study metrics on CP^n. The Riemann tensor derived from the associated metric connection is given by (Refs. 23, 27, and 7) $R_{abcd} = -\frac{1}{2}k(g_{a[c}g_{b|d]} + \Omega_{a[c}\Omega_{b|d]} + \Omega_{ab}\Omega_{cd})$, the Ricci tensor by $R_{ab} = -\frac{1}{2}k(n+1)g_{ab}$, and the Ricci scalar by $R = -kn(n+1)$. We adopt the convention that $k=1$, and in this case for $n=1$ the Fubini–Study metric becomes the intrinsic distance measure on the 2-sphere of the unit radius.

Kähler geometries have a special curvature property that relates to Theorem 1 below. First, we give the definition of a Kähler manifold (cf. Ref. 28).

Definition 2: A complex manifold M is said to be Kähler if it comes equipped with a Hermitian metric $h_{\alpha\beta'}$ with $ds^2 = h_{\alpha\beta'}\, dz^\alpha \otimes d\overline{z}^{\beta'}$ such that the real 2-form $\Omega = i h_{\alpha\beta'}\, dz^\alpha \wedge d\overline{z}^{\beta'}$ is closed. Then Ω is said to be a Kähler form for M.

This is equivalent (Ref. 28) to the existence on M of a real-valued function K, the Kähler scalar function, such that $\Omega = i\partial\overline{\partial}K$, or equivalently $h_{\alpha\beta'} = \partial_\alpha \overline{\partial}_{\beta'} K$. Consequently, a complex submanifold N of a Kähler manifold M is itself Kähler, since the restriction of the function K to N provides the intrinsic Kähler form by applying the operator $\sqrt{-1}\partial\overline{\partial}$. In the case of the Fubini–Study metric on CP^n, the Kähler scalar function takes the form $K = 4k^{-1}\log[1+k(|\zeta^1|^2 + |\zeta^2|^2 + \cdots + |\zeta^n|^2)]$, where ζ^α are inhomogeneous coordinates on CP^n.

Proposition 1: Let Ω be a positive (1,1)-form on a complex manifold M. Then Ω is a Kähler form for M if and only if for all $x_0 \in M$ there exist holomorphic ''Euclidean'' coordinates $z^1,\ldots,z^n$ around x_0, such that $\Omega = i h_{\alpha\beta'}\, dz^\alpha \wedge d\bar{z}^{\beta'}$ and $h_{\alpha\beta'} = \delta_{\alpha\beta'} + O(|z|^2)$ at x_0, and thus the Kähler metric osculates the flat Euclidean metric to second order.

Proof: The implication toward Ω being a Kähler form is clear. To prove the reverse implication, begin with holomorphic coordinates $z^1,\ldots,z^n$, such that $dz^1,\ldots,dz^n$ give an orthonormal basis of T^*_{M,x_0}, the dual tangent space to M at x_0. This implies that $\Omega = i\tilde{h}_{\alpha\beta'}\, dz^\alpha \wedge d\bar{z}^{\beta'}$, where $\tilde{h}_{\alpha\beta'} = \delta_{\alpha\beta'} + \Sigma_{1\leq\gamma\leq n}(a_{\gamma\alpha\beta'} z^\gamma + a'_{\gamma'\alpha\beta'}\bar{z}^{\gamma'}) + O(|z|^2)$. That Ω is real implies $a'_{\gamma'\alpha\beta'} = \bar{a}_{\gamma'\beta'\alpha}$. The Kähler condition $\partial h_{\alpha\beta'}/\partial z^\gamma = \partial h_{\gamma\beta'}/\partial z^\alpha$ at x_0 implies that $a_{\alpha\gamma\beta'} = a_{(\alpha\gamma)\beta'}$. To complete the proof, we define holomorphic Euclidean coordinates $\hat{z}^\alpha$ as $\hat{z}^\beta := z^\beta + \frac{1}{2}\Sigma_{\gamma,\alpha} a_{\gamma\alpha\beta'}\, \delta^{\beta\beta'} z^\gamma z^\alpha$. ■

V. THE GEOMETRY OF COHERENT STATES

We begin our discussion with a result concerning the ''nonlinear'' geometry of the coherent state submanifold.

Lemma 3: Given a pair of distinct coherent states, the complex projective line $L \subset P\mathcal{F}$ joining them intersects $\mathcal{C}$ exactly twice, at the coherent states themselves.

Proof: Suppose that L intersects $\mathcal{C}$ in three or more distinct points. This implies a relation of the form $|\alpha_c\rangle + \lambda|\beta_c\rangle = \mu|\gamma_c\rangle$ for $|\alpha\rangle$, $|\beta\rangle$, $|\gamma\rangle$ normalized, with $\lambda\mu \neq 0$. In the harmonic oscillator case we expand this equation according to $|\alpha_c\rangle = \Sigma_{n=0}^\infty(\alpha^n/\sqrt{n!})\cdot\exp(-\frac{1}{2}\alpha\bar{\alpha})|n\rangle$ for energy eigenstates $|n\rangle$. Taking the Dirac product with $\langle m|$ for all m gives infinitely many linear equations for the two unknowns λ and μ, whose solution space is empty for distinct α, β, γ. A similar argument applies if we generalize the coherent states according to their algebraic characterization, via the exponential map $\mathcal{E}$ of (3.1) (i.e., we do not specialize to the case of the harmonic oscillator). Then suppose $\mathcal{E}(\xi^a) + \lambda\mathcal{E}(\theta^a) = (1+\lambda)\mathcal{E}(\eta^a)$ with ξ^a, θ^a, η^a nonzero distinct elements of $\mathcal{H}^1$. This relation implies infinitely many equations to be satisfied by λ. The square of the $\mathcal{H}^n$ part of the relation gives $(\xi^2)^n + \lambda^2(\theta^2)^n + 2(\xi\cdot\theta)^n\lambda = (1+\lambda)^2(\eta^2)^n$. If all of ξ^2, θ^2, η^2 are nonzero, and not all equal to unity, then these equations have no solution. So, for a solution in λ to exist, we must have ξ^2, θ^2, η^2 all equal to unity. Thus, ξ^a, η^a, θ^a are unit vectors and $1+\lambda^2+2(\xi\cdot\theta)^n\lambda = (1+\lambda)^2$, $\forall n$. This has no solution for nonzero λ unless $(\xi\cdot\theta)=1$, in which case ξ^a, θ^a are the same unit vectors. ■

However, any normalized coherent state vector $|\alpha_c\rangle$ is decomposable as a continuous integral over the states of $\mathcal{C}$ via the resolution of unity (3.3). Provided an analyticity assumption holds, this decomposition is unique (see Sec. IV of Ref. 12 for a proof of this result).

We now state our main result concerning the geometry of $\mathcal{C}$ (cf. also Refs. 29 and 11).

Theorem 1a: *The metric induced on the coherent state submanifold $\mathcal{C}$ from the ambient Fubini–Study metric on the quantum state space is intrinsically flat. The coordinates ξ^a on the single-particle Hilbert space $\mathcal{H}^1$ are complex Euclidean coordinates for $\mathcal{C}$.*

If instead we begin with the coherent state submanifold and decide, *a priori*, to place on it the complex Euclidean metric, giving us the manifold with metric $\mathcal{C}_E$, then we have the following equivalent result.

Theorem 1b: *The Euclidean coherent state submanifold has an isometric embedding into the Fubini–Study state manifold.*

We remark that this theorem relates to the work of Rawnsley,[30] which discusses the geometric quantization of a Kähler manifold K, and how the resulting coherent states give a map $\mathcal{E}$ from K into the quantum mechanical projective state space. In particular, following Corollary 6, it is remarked that, by Kodaira's theorem on Hodge manifolds (Refs. 31 and 28), there exists a holomorphic line bundle connection over K for which the map $\mathcal{E}$ is an embedding. This result is independent of the curvature of K, which is determined by the Poisson bracket of the classical field theory being quantized. In our theorem we have effectively adopted complex Cartesian coordinates for K so that K is complex Euclidean space and so trivially a Kähler manifold (cf. Corollary 1 below, and also Lemma 6 to follow on truncated Fock spaces). Our approach differs from that of Rawnsley, in that we have focused on geometric quantum mechanics rather than geometric quantization. It is nevertheless noteworthy that the two approaches yield descriptions of the coherent states with essentially the same underlying geometry. With regard to the choice of Cartesian coordinates, we note the remarks of Dirac (p. 114 of Ref. 26). For further background to the material covered in Ref. 30 see Ref. 24. Observe that our theorem is independent of the details of the single-particle Hilbert space $\mathcal{H}^1$.

Proof: We give three independent proofs of this result, from the points of view of differential geometry, operator algebra, and the Kähler form.

(a) Differential geometry. From the way that we defined coherent state vectors, we can regard $\xi^a \in \mathcal{H}^1$ as complex coordinate functions for the coherent state submanifold. It will be helpful to introduce some further notation. We define $\xi^{(n)} := \xi^{(\alpha}\xi^{\beta}\cdots\xi^{\delta)}/\sqrt{n!}$ so that $\xi^{(n)}$ is the tensor contribution to the coherent state vector $|\xi_c\rangle$, which lies in $\mathcal{H}^n$, there being n factors in the symmetrized tensor product; we define $\overline{\xi}_{(n)}$ similarly. Then setting $\Lambda = \xi^\alpha \overline{\xi}_\alpha$, we find $\xi^{(n)}\overline{\xi}_{(n)} = \Lambda^n/n!$. Restricted down to the coherent state submanifold $\mathcal{C}$, we can calculate the tangent vector to a coherent state induced by an element $d\overline{\xi}$ of $T^*\mathcal{H}$. The component in $\mathcal{H}^n$ of the (dual) tangent vector is given by $d\xi^{(n)} = n\, d\xi^{(\alpha}\,\xi^{\beta}\xi^{\gamma}\cdots\xi^{\delta)}/\sqrt{n!}$ and similarly for the complex conjugate expression. To calculate the Fubini–Study line element we find the coordinate inner product of a tangent vector with itself. The contribution to this of any pair of vectors lying in distinct $\mathcal{H}^n$ vanishes, as follows from our expression for the Hilbert space norm given earlier. Thus, when evaluating the inner product of two Fock space vectors in the abstract index notation, we contract over vectors and their conjugates with the same number of indices. Hence $d\xi^{(n)}\, d\overline{\xi}_{(m)} \equiv \delta^{(n)}_{(m)} \cdot [1/(n-1)!] \times [\Lambda^{n-1}\, d\xi^\alpha\, d\overline{\xi}_\alpha + (n-1)\Lambda^{n-2}|\xi^\alpha\, d\overline{\xi}_\alpha|^2]$ for all $n \geqslant 1$. We also need the coordinate inner product expression $\overline{\xi}_{(m)} d\xi^{(n)} = \delta^{(n)}_{(m)} \cdot \overline{\xi}_{(\alpha}\overline{\xi}_{\beta}\cdots\overline{\xi}_{\delta)}/\sqrt{n!}\cdot(n/\sqrt{n!})\, d\xi^{(\alpha}\xi^{\beta}\xi^{\gamma}\cdots\xi^{\delta)} \equiv \delta^{(n)}_{(m)} \cdot [1/(n-1)!] \times (\overline{\xi}_\alpha\, d\xi^\alpha)\Lambda^{n-1}$, and, similarly, $\xi^{(n)}\, d\overline{\xi}_{(m)} \equiv \delta^{(n)}_{(m)} \cdot [1/(n-1)!](\xi^\alpha\, d\overline{\xi}_\alpha)\Lambda^{n-1}$ for all $n \geqslant 1$. In (4.1b) above a vector ξ is given by the collection $\{\xi^{(n)}\}$ for all values of n, and thus to evaluate the line element induced on the coherent state submanifold $\mathcal{C}$ we must sum over all $1 \leqslant m, n \leqslant \infty$ in the above identities. This yields $\Sigma_{m,n}\, d\xi^{(n)}\, d\overline{\xi}_{(m)} = e^{\Lambda}(d\xi^\alpha\, d\overline{\xi}_\alpha + |d\xi^\alpha\, \overline{\xi}_\alpha|^2)$ and $\Sigma_{m,n}\overline{\xi}_{(m)}\, d\xi^{(n)} = e^{\Lambda} \cdot \overline{\xi}_\alpha\, d\xi^\alpha$, together with its complex conjugate. The denominator in the Fubini–Study line element equals $1 + \Sigma_{m,n}\xi^{(n)}\overline{\xi}_{(m)} = e^{\Lambda}$, and thus the induced line element reduces to $ds^2 = 4\, d\xi^\alpha\, d\overline{\xi}_\alpha$, as required. ■

(b) Operator algebra. Here we shall assume only the canonical commutation relations (CCR) for the creation and annihilation operators $[\hat{A}^\alpha, \hat{C}_\beta] = \delta^\alpha_\beta$, together with (3.2a) and (3.2b), for these properties characterize the coherent states up to unitary transformations (Refs. 14, 32). The proof that follows thus applies in the case of Perelomov's ''generalized coherent states'' (Ref. 32) since the CCR and properties (3.2) are preserved under the action of the unitary group. We shall adopt the Dirac notation for state vectors according to $Z^\alpha \leftrightarrow |\psi\rangle$, $\overline{Z}_\alpha \leftrightarrow \langle\psi|$. In this notation the Fubini–Study line element becomes

$$ds^2_{\text{F.S.}} = 4\left\{\frac{\langle d\psi|d\psi\rangle}{\langle\psi|\psi\rangle} - \frac{\langle\psi|d\psi\rangle\langle d\psi|\psi\rangle}{\langle\psi|\psi\rangle^2}\right\}. \tag{5.1}$$

We abbreviate so that $|\psi\rangle \in \mathcal{F}$ denotes the coherent state vector associated with $\psi^\alpha \in \mathcal{H}^1$. Then by (3.2a) and (3.2b), we have $|d\psi\rangle = \hat{C}_\alpha|\psi\rangle d\phi^\alpha$ and $\langle d\psi| = d\overline{\phi}_\beta\langle\psi|\hat{A}^\beta$. Using the relations $[\hat{A}^\alpha, \hat{C}_\beta] = \delta^\alpha_\beta$ we calculate $\langle d\psi|d\psi\rangle = d\overline{\phi}_\beta\, d\phi^\alpha\langle\psi|\hat{A}^\beta\hat{C}_\alpha|\psi\rangle$. Rewriting the operator in the matrix element in terms of a commutator gives $d\overline{\phi}_\beta\, d\phi^\alpha\langle\psi|[\hat{A}^\beta, \hat{C}_\alpha] + \hat{C}_\alpha\hat{A}^\beta|\psi\rangle$, which by the CCR and (3.2) equals $[d\phi^\alpha\, d\overline{\phi}_\alpha + (\phi^\beta\, d\overline{\phi}_\beta)(\overline{\phi}_\alpha\, d\phi^\alpha)]\langle\psi|\psi\rangle$. Similarly, $\langle\psi|d\psi\rangle = \langle\psi|\hat{C}_\alpha|\psi\rangle d\phi^\alpha = \langle\psi|\overline{\phi}_\alpha|\psi\rangle d\phi^\alpha = \langle\psi|\psi\rangle(\overline{\phi}_\alpha\, d\phi^\alpha)$. Thus, the line element induced on $\mathcal{C}$ reduces to $ds^2 = 4(d\phi^\alpha\, d\overline{\phi}_\alpha)^2$, as required. The proof is illustrated below. ■

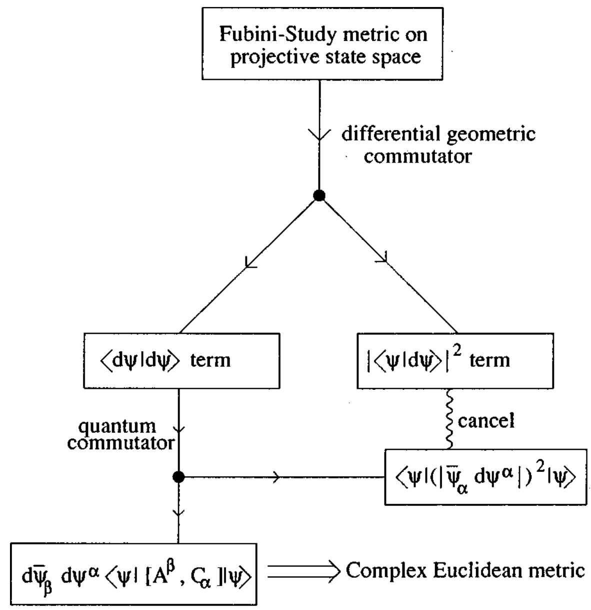

(c) Kähler form. Recall the Kähler scalar function for CP^n, where now we take n to be countable infinity $\aleph_0$. For $\mathcal{A}\subset\mathcal{F}$ defined in Sec. IV, we have $K=4\log(1+|\xi^{(1)}|^2+\cdots+|\xi^{(j)}|^2+\cdots)$, where $\xi^{(j)}=\xi^{\alpha_1\cdots\alpha_j}$ as before. For $\mathcal{C}$ we have the coherent state vector associated with $\psi^\alpha\in\mathcal{H}^1$ given by $\psi^{(j)}=(\psi^\alpha)^{\otimes_s^j}/\sqrt{j!}$, and thus $|\psi^{(j)}|^2=\Lambda^j j!$ with $\Lambda=\psi^\alpha\overline{\psi}_\alpha$. Summing over j to infinity (to sum to infinity is, in fact, necessary for flatness, as we discuss below) we obtain the simple relation

$$K|_{\mathcal{C}}=4\Lambda. \tag{5.2}$$

Then the induced metric on $\mathcal{C}$ is given by $h_{\alpha\beta'}=4\partial_\alpha\overline{\partial}_{\beta'}(\psi^\gamma\overline{\psi}_\gamma)=4\delta_{\alpha\beta'}$, as required. ■

The theorem has the following immediate consequence.

Corollary 1: The coherent state submanifold is a Kähler manifold.

We see therefore that the theorem asserts a global geometric property of the coherent state submanifold (in itself Kähler), which, by comparison with Proposition 1, is a special case of a *second-order* flatness property that applies locally to any Kähler manifold.

We have seen in Lemma 3 that the coherent state submanifold $\mathcal{C}$ is nonlinear, in the sense that the complex projective line joining two distinct coherent states lies in the complement of $\mathcal{C}$, except at its two intersection points. This is an algebraic result whose proof relies upon the uniqueness of decomposition of any given state into coherent states. It suggests the following geometric property of the coherent state submanifold (Refs. 29 and 11).

Proposition 2: Given any two distinct coherent states, the complex projective line joining them intersects $\mathcal{C}$ transversally; that is to say, the line joining the two coherent states does not lie in the tangent space to $\mathcal{C}$ at either intersection point.

Proof: Since $\mathcal{C}$ is homogeneous, we can assume that one of the coherent states is the vacuum state, that is, $P|0\rangle$, where $|0\rangle$ is the element of Fock space that is the exponential of the origin in the vector space $\mathcal{H}^1$. Then from Theorem 1 the intrinsic geodesic distance s from $P|0\rangle$ to $P|\xi_c\rangle(\xi\neq 0)$ is given by $s=2\Lambda^{1/2}$. Recall (e.g., Ref. 7) that the geodesic distance θ between the two states in $P\mathcal{F}$ with respect to the ambient Fubini–Study metric on $P\mathcal{F}$ is determined by the cross ratio

$$\frac{1}{2}(1+\cos\theta)=\frac{\langle\xi_c|0\rangle\langle 0|\xi_c\rangle}{\langle 0|0\rangle\langle\xi_c|\xi_c\rangle}, \tag{5.3}$$

where we take θ to be the principal value determined by this equation, $0\leqslant\theta\leqslant\pi$. Clearly $\langle 0|0\rangle=\langle 0|\xi_c\rangle=\langle\xi_c|0\rangle=1$, and thus $\theta=\cos^{-1}(2e^{-\Lambda}-1)$. It follows that $d\theta/d\Lambda=(e^{\Lambda}-1)^{-1/2}$, and so $d\theta/ds=[\Lambda/(e^{\Lambda}-1)]^{1/2}$. Thus, $d\theta/ds$ is a monotone decreasing function of Λ, beginning at $d\theta/ds=1$, where $\Lambda=0$, and decaying to zero as Λ tends to infinity. Tangency at $P|\xi_c\rangle$ would require $d\theta/ds=1$ for some $\Lambda\neq 0$, and this is not possible given the form of the function $d\theta/ds$. ■

The method above proves another result that is intuitive from the nonlinear geometric property of $\mathcal{C}$ derived in Lemma 3.

Corollary 2: The geodesic distance along the projective line in $P\mathcal{F}$ joining two distinct coherent states is strictly less than the intrinsic geodesic distance within $\mathcal{C}$.

Theorem 1 also establishes the following simple geometric properties of $\mathcal{C}$.

Lemma 4: The intrinsic $\mathcal{C}$ geodesic distance between two coherent states is given by the Hilbert space norm of the difference field of the two corresponding vectors in $\mathcal{H}^1$. The corresponding distance of a coherent state from the vacuum is equal to its Hilbert space norm.

Lemma 5: The overlap $\mathrm{Re}\langle\xi_c|\psi_c\rangle$ *of two normalized coherent state vectors is the cosine of the angle that these states subtend at the vacuum state in the intrinsic geometry of $\mathcal{C}$.*

These results illustrate the geometric character of two emergent linear structures in quantum theory. On the one hand, the addition of elements of the Hilbert space $\mathcal{H}^1$ of solutions to some classical linear field equation yields a new classical field. As we have seen, the intrinsic geometry of the associated coherent states is Euclidean, with the elements of $\mathcal{H}^1$ serving as Euclidean coordinates. On the other hand, any two distinct coherent states can be superposed in the quantum mechanical sense of joining them with the unique complex projective line in the ambient Fubini–Study geometry of the underlying state space $P\mathcal{F}$; we have seen that this superposition is ''noncoherent.'' These two features are present in a linear theory of gravity (cf. the discussion in Ref. 33). The coherent states provide a natural preferred basis, together with a unique probability distribution for the associated resolution of unity. The state space geometry illustrates how a quantum superposition of distinct classical field configurations is outside the classical domain (see Fig. 1).

Theorem 1 has an important consequence for the relationship between the quantum complex structure J of Sec. I and the manifold complex structure on the state space $P\mathcal{F}$. This theorem shows that, in a suitable sense, these two complex structures are identical. Since the coherent state submanifold is Euclidean, and has as Euclidean coordinates the single-particle state vectors themselves, the active transformation J of (2.1c) induces a corresponding transformation on the tangent space to $\mathcal{C}$ at the vacuum state $P|0\rangle$. (This is because the notion of finite displacement from some origin in any Euclidean space is vectorial.) Thus, to find the action of the manifold complex

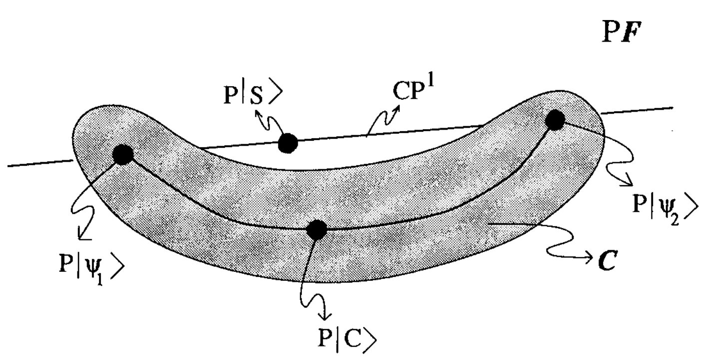

FIG. 1. Coherent state submanifold, embedded inside projective Fock space. For the states shown, $|S\rangle=\lambda|\psi_1\rangle+\mu|\psi_2\rangle$, $\lambda\mu\neq 0$, and $|C\rangle=\mathcal{E}[u\psi_1^a+(1-u)\psi_2^a]$, $u\neq 0,1$, $\psi_i^a\in\mathcal{H}^1$.

structure of the state space on any tangent vector $\mathbf{V}$ to $\mathcal{C}$ at the vacuum, we regard this vector as a displacement in the Euclidean space $\mathcal{C}$ and accordingly as a coherent state given by this displacement from the vacuum. Then we act on the corresponding $\mathcal{H}^1$ element with J and exponentiate this according to (3.1) to give a new displacement vector in $\mathcal{C}$. In turn, this is associated with a unique tangent vector to $\mathcal{C}$ at $P|0\rangle$, which gives the required action of the manifold complex structure on the original vector $\mathbf{V}$.

It is profitable to think in terms of the geometry of the embedding of $\mathcal{C}$ into $P\mathcal{F}$ as an affine subspace whose compactification is the set of states for which the vacuum entry is zero. In this way symmetry is broken in affine $\mathcal{C}$ via the specification of the location of the vacuum state. The geodesics in $\mathcal{C}$, which are straight lines, pass through infinity, or, more precisely, they pass through the infinite-dimensional complex projective space $P\Pi$, where $\Pi=\{(0,\psi^\alpha,\psi^{\alpha\beta},\ldots)\in\mathcal{F}\}$. The only geodesics in the ambient geometry that pass through a given point $P_\infty\in P\Pi$ are circles that pass through an antipodal point, for example the vacuum. Observe however that $\mathcal{C}$ and $P\mathcal{F}$ share no geodesics in common, since any geodesic in the state space has length 2π, whereas Theorem 1 implies that geodesics of $\mathcal{C}$ have infinite length with respect to the same distance measure. Motion of a coherent state along such a geodesic corresponds to scaling the amplitude of its associated single-particle state vector. As this amplitude tends to infinity, the compactification of $\mathcal{C}$ is approached, and the expectation of the total number operator approaches infinity.

The isometry group of $\mathcal{C}$ corresponds to Killing vector fields of the induced Euclidean metric, and consists of rigid rotations and translations, giving rise to orbits that are circles or straight lines inside $\mathcal{C}$. Observe that the projective unitary group $PSU(\infty)$ is isomorphic to the isometry group of the Fubini–Study metric. Hence, the Killing orbits of $\mathcal{C}$ will not, in general, be unitary orbits of the ambient state space.

We examine to what extent the infinite dimensionality of the situation described above affects the result of Theorem 1. Suppose we truncate the Fock space at the N particle states, and define $\mathcal{F}_T=C\oplus\mathcal{H}^1\oplus\cdots\oplus\mathcal{H}^N$ for some finite positive integer N. We define analogs of coherent states as previously, where now the coordinates $\xi^{(n)}$ for $n>N$ are assumed to vanish. The above procedure for calculating the induced metric on the resulting submanifold can be followed closely, and some interesting features of the truncated coherent state submanifold $\mathcal{C}_T$ emerge. We define the real function $S_N(\Lambda):=1+\Sigma_{n=1}^N(1/n!)\Lambda^n$ with Λ as above. In physical terms, the truncated Fock space $\mathcal{F}_T$ is appropriate for a situation in which the expectation of the total number operator $\hat{N}$ is bounded above by N. For example, one might consider the quantum electrodynamics of a photon field constrained inside a finite spatial volume with a saturation density. For a coherent state vector $|\xi_c\rangle$ we have $\langle\hat{N}\rangle=\Lambda S_{N-1}(\Lambda)/S_N(\Lambda)$. Thus, $\langle\hat{N}\rangle<N$, since $\Lambda S_{N-1}(\Lambda)<NS_N(\Lambda)$, as is straightforward to verify. In the limit $\xi^a\mapsto\lambda\xi^a$, $|\lambda|\to\infty$ we find $\langle\hat{N}\rangle\to N$, and thus the distribution among the n particle states becomes more strongly peaked at the N particle states, as the amplitude increases without bound.

Following the same argument as given in proof (a) of Theorem 1, we obtain the following expression for the induced metric on $\mathcal{C}_T$:

$$ds^2=4\left\{\frac{S_{N-1}(\Lambda)}{S_N(\Lambda)}d\xi^a\,d\overline{\xi}_a+\left(\frac{S_{N-2}(\Lambda)}{S_N(\Lambda)}-\frac{S_{N-1}^2(\Lambda)}{S_N^2(\Lambda)}\right)|d\xi^a\overline{\xi}_a|^2\right\}. \tag{5.4}$$

The case of Theorem 1 is given by setting $N=\aleph_0$, and then $S_\infty(\Lambda)=e^\Lambda$ so that the above line element reduces to the flat Euclidean line element. It is natural to ask, in the case of finite N, whether the submanifold $\mathcal{C}_T$ is Kähler, and whether it possesses intrinsic curvature. As regards the Kähler geometry, we have the following result.

Lemma 6: The induced metric on $\mathcal{C}_T$ is Kähler, with a Kähler scalar function given by $K_N=\log S_N(\Lambda)$.

Proof: There are two ways of proceeding. Simplest is to use the same argument as in the Kählerian proof (c) of Theorem 1, summing only to finite N in the expression for K. This incorporates a proof of Theorem 1 as a special case in the limit $N\to\infty$, since $S_{\aleph_0}=e^\Lambda$, and thus

$K_{\mathfrak{N}_0}=4\Lambda$ as in (5.2). Alternatively, it suffices to evaluate the 2-form associated with K_N and show that this is identical to the line element (5.4) upon replacement $\wedge\mapsto\otimes$. We have $\partial K=(S_{N-1}(\Lambda)/S_N(\Lambda))\overline{\xi}_a\,d\xi^a$, and thus $\overline{\partial}\partial K=S_{N-1}(\Lambda)/S_N(\Lambda)d\overline{\xi}_a\wedge d\xi^a+[S_N(\Lambda)S_{N-2}(\Lambda)-S_{N-1}^2(\Lambda)]/S_N^2(\Lambda)(\xi^b\,d\overline{\xi}_b)\wedge(\overline{\xi}_a\,d\xi^a)$, which is of the required form. ■

As regards the intrinsic curvature, we adopt coordinates $\xi^a=(\xi^\alpha,\overline{\xi}^{\alpha'})$ on $\mathcal{C}_T$ and write the induced metric as $g_{\alpha\beta'}=K_N'\delta_{\alpha\beta'}+K_N''\xi_{\beta'}\overline{\xi}_\alpha$. Here the prime operation is $d/d\Lambda$, and we adopt the index notation, as explained in Sec. II. That g is Hermitian implies $\overline{g_{\alpha\beta'}}\equiv\overline{g}_{\alpha'\beta}=g_{\alpha'\beta}=K_N'\delta_{\alpha'\beta}+K_N''\overline{\xi}_\beta\xi_{\alpha'}$. To calculate the Christoffel connection, we need the inverse metric to g_{ab}, written as in (2.1a) with all raised indices, and, by inspection, we find $g^{\alpha\beta'}=(1/K_N')\delta^{\alpha\beta'}-(H_N/K_N')\xi^\alpha\overline{\xi}^{\beta'}$, where the function H_N is defined by $H_N:=K_N''/(K_N'+\Lambda K_N'')$. The derivatives of Λ are $\partial_\alpha\Lambda=\overline{\xi}_\alpha$, $\overline{\partial}_{\alpha'}\Lambda=\xi_{\alpha'}$, and so $\overline{\partial}_{\gamma'}g_{\alpha\beta'}=2K_N''\delta_{\alpha(\beta'}\xi_{\gamma')}+K_N'''\overline{\xi}_\alpha\xi_{\beta'}\xi_{\gamma'}$ and $\partial_\gamma g_{\alpha\beta'}=2K_N''\delta_{\beta'(\alpha}\overline{\xi}_{\gamma)}+K_N'''\xi_{\beta'}\overline{\xi}_\alpha\overline{\xi}_\gamma$. Observe that these relations imply the symmetries $\overline{\partial}_{\gamma'}g_{\beta'\alpha}\equiv\overline{\partial}_{(\gamma'}g_{\beta')\alpha}$, $\partial_\gamma g_{\alpha\beta'}\equiv\partial_{(\gamma}g_{\alpha)\beta'}$, which, in fact, hold for an arbitrary Kähler manifold (as follows directly from $g_{\alpha\beta'}\propto\partial_\alpha\partial_{\beta'}K$ and that $\partial_\alpha,\overline{\partial}_{\beta'}$ commute among one another). The Christoffel symbols are defined by $\Gamma^a{}_{bc}=\frac{1}{2}g^{ad}(\partial_c g_{bd}+\partial_b g_{cd}-\partial_d g_{bc})$ and the calculation of these proceeds using the above identities. We find $\Gamma^\alpha{}_{\beta\gamma}=g^{\alpha\delta'}\partial_{(\gamma}g_{\beta)\delta'}=(1/K_N')(2K_N''\delta^\alpha_{(\gamma}\overline{\xi}_{\beta)}+\Theta_N\xi^\alpha\overline{\xi}_\gamma\overline{\xi}_\beta)$, where the function Θ_N is defined by $\Theta_N:=(K_N'''-2H_NK_N''-\Lambda H_NK_N''')$. The calculation of $\Gamma^{\alpha'}{}_{\beta'\gamma'}$ proceeds in the same way with the roles of ξ and $\overline{\xi}$ interchanged, so that $\Gamma^{\alpha'}{}_{\beta'\gamma'}=(1/K_N')(2K_N''\delta^{\alpha'}_{(\beta'}\xi_{\gamma')}+\Theta_N\overline{\xi}^{\alpha'}\xi_{\beta'}\xi_{\gamma'})$. The ''mixed'' symbol $\Gamma^\alpha{}_{\beta\gamma'}$ vanishes, since $\Gamma^\alpha{}_{\beta\gamma'}=g^{\alpha\delta'}\overline{\partial}_{[\gamma'}g_{\delta']\beta}\equiv 0$. Thus, $\Gamma^\alpha{}_{\beta\gamma'}$, $\Gamma^{\alpha'}{}_{\beta'\gamma}\equiv 0$. The remaining Christoffel symbols $\Gamma^{\alpha'}{}_{\beta\gamma}$ and $\Gamma^\alpha{}_{\beta'\gamma'}$ vanish by the identities $g_{\alpha\beta}=g_{\alpha'\beta'}\equiv 0$. In summary, only the symbols $\Gamma^\alpha{}_{\beta\gamma}$ and $\Gamma^{\alpha'}{}_{\beta'\gamma'}$ are nonzero. (This property holds for a general Kähler manifold, since the above symmetries of the derivatives of g hold in the general case cf. remark following Definition 2 above.) For a general Kähler manifold there exist simple identities for the Riemann curvature tensor in terms of the Christoffel connection. We have in particular (Ref. 34) $\Gamma^\alpha{}_{\beta\gamma}=g^{\alpha\delta'}\,\partial^3K/\partial\xi^\beta\,\partial\xi^\gamma\,\partial\overline{\xi}^{\delta'}$, $\Gamma^{\alpha'}{}_{\beta'\gamma'}=g^{\delta\alpha'}\,\partial^3K/\partial\xi^\delta\,\partial\overline{\xi}^{\beta'}\,\partial\overline{\xi}^{\gamma'}$; and $\Gamma^\alpha{}_{\beta\gamma'}=\Gamma^\alpha{}_{\beta'\gamma}=\Gamma^{\alpha'}{}_{\beta\gamma}=\Gamma^\alpha{}_{\beta'\gamma'}=\Gamma^{\alpha'}{}_{\beta'\gamma}=\Gamma^{\alpha'}{}_{\beta\gamma'}=0$. It follows that for a general Kähler manifold the only nonvanishing components of the Riemann tensor are $R_{\alpha\beta'\gamma\delta'}$, $R_{\alpha\beta'\gamma'\delta}$, $R_{\alpha'\beta\gamma\delta'}$, $R_{\alpha'\beta\gamma'\delta}$, and these possess the Hermitian and symmetry properties $R_\lambda{}^\alpha{}_\mu{}^\beta=R_\lambda{}^{(\alpha}{}_\mu{}^{\beta)}=R_{(\lambda}{}^{(\alpha}{}_{\mu)}{}^{\beta)}=\overline{R}_\lambda{}^\alpha{}_\mu{}^\beta=\overline{R_{\lambda'}{}^{\alpha'}{}_{\mu'}{}^{\beta'}}$. As a consequence, the Ricci tensor is determined by a Hermitian $R_{\alpha\beta'}$, and the Ricci scalar R is real. For a general Kähler manifold we have (Ref. 34) the useful simplifying identity $R_{\alpha\beta'\lambda}{}^\mu\equiv\overline{\partial}_{\beta'}\Gamma^\mu{}_{\alpha\lambda}$, so that in our case we find

$$R_\alpha{}^\beta{}_\lambda{}^\mu=-R_\alpha{}^{\beta\mu}{}_\lambda=-\left(\frac{2K_N''}{K_N'}\right)'\delta^\mu_{(\alpha}\overline{\xi}_{\lambda)}\xi^\beta-\frac{2K_N''}{K_N'}\delta^\mu_{(\alpha}\delta^\beta_{\lambda)}-\frac{2\Theta_N}{K_N'}\xi^\mu\overline{\xi}_{(\alpha}\delta^\beta_{\lambda)}-\left(\frac{\Theta_N}{K_N'}\right)'\xi^\beta\xi^\mu\overline{\xi}_\alpha\overline{\xi}_\lambda. \tag{5.5a}$$

Thus, with indices in any position, the Riemann curvature vanishes in the limit that $N\to\mathfrak{N}_0$, for then $K'=1$ and Θ, K'' both vanish. This is in accordance with Theorem 1. (It makes sense to discuss the limit $N\to\mathfrak{N}_0$, since the underlying $\mathcal{H}^1$ serves to coordinatize $\mathcal{C}_T$ for each N.) We calculate the Ricci tensor of $\mathcal{C}_T$ by taking the trace on β and λ in the expression for $R_\alpha{}^{\beta\mu}{}_\lambda$ from (5.5a). We obtain

$$R_\alpha{}^\mu=\xi^\mu\overline{\xi}_\alpha[(K_N''/K_N')'+(n+1)(\Theta_N/K_N')+\Lambda(\Theta_N/K_N')']+\delta^\mu_\alpha[\Lambda(K_N''/K_N')'+(n+1)(K_N''/K_N')], \tag{5.5b}$$

where n is the holomorphic dimension of $\mathcal{H}^1$, and this $R_\alpha{}^\mu$ is Hermitian. For the Ricci scalar R we find

$$R=(n+1)\Lambda\left(\frac{2K_N''}{K_N'}\right)'+n(n+1)\frac{K_N''}{K_N'}+(n+1)\Lambda\frac{\Theta_N}{K_N'}+\Lambda^2\left(\frac{\Theta_N}{K_N'}\right)', \tag{5.5c}$$

which is real. Observe that both the Ricci tensor and Ricci scalar diverge as the holomorphic dimension n approaches $\aleph_0$. Nevertheless, the Weyl curvature is well defined in this limit, as we demonstrate. The Weyl tensor (e.g., Ref. 18) of a $2n$-real dimensional Kähler manifold is determined by $C_\alpha{}^\beta{}_\gamma{}^\delta - R_\alpha{}^\beta{}_\gamma{}^\delta = [1/(2n-2)](\delta_\alpha^\delta R_\gamma{}^\beta + \delta_\gamma^\beta R_\alpha{}^\delta) - [1/(2n-1)(2n-2)]R\,\delta_\alpha^\delta \delta_\gamma^\beta$ so that in our case, as $n = \dim_{\mathbf{C}}(\mathcal{H}^1)$ approaches $\aleph_0$, we obtain

$$C_\alpha{}^\beta{}_\gamma{}^\delta - R_\alpha{}^\beta{}_\gamma{}^\delta = \frac{1}{2}\,\frac{\Theta_N}{K_N'}\,(\delta_\alpha^\delta \xi^\beta \overline{\xi}_\gamma + \delta_\gamma^\beta \xi^\delta \overline{\xi}_\alpha) + \frac{3}{4}\,\frac{K_N''}{K_N'}\,\delta_\alpha^\delta \delta_\gamma^\beta. \tag{5.6}$$

We observe that this quantity is finite, which has the following consequences.

Proposition 3: For the complete Fock space $(N = \aleph_0)$, *the Weyl curvature of the coherent state submanifold vanishes for all values of* $\dim \mathcal{H}^1$, *including countable infinity* $\aleph_0$. *This is despite the divergence of the Ricci tensor in the limit* $\dim \mathcal{H}^1 \to \aleph_0$. *Thus,* $\mathcal{C}$ *is conformally flat for all single-particle Hilbert spaces* $\mathcal{H}^1$.

Proposition 4: For the Fock space $\mathcal{F}_T$ *(N finite), the submanifold* $\mathcal{C}_T$ *has finite conformal curvature for any single-particle Hilbert space* $\mathcal{H}^1$.

By comparison, the Fubini–Study geometry of the underlying state manifold $P\mathcal{F}$ has a divergent Ricci tensor and scalar for $\dim \mathcal{H}^1 = \aleph_0$, and finite nonzero Riemann and Weyl curvatures proportional to the holomorphic sectional curvature k, for all $\mathcal{H}^1$.

VI. DISCUSSION

Theories of stochastic state vector reduction processes have recently been extensively studied (see, e.g., Refs. 35 and articles cited therein). In particular, a proposal of this kind has been given in the context of geometric quantum mechanics). (See Ref. 7.) It would be interesting to see whether such an approach is viable in the infinite-dimensional case that arises in the context of quantum field theory, with the geometry we have described above. The submanifold of coherent states suggests itself as a natural geometric object to study in the context of a stochastic model, in which the coherent states provide the preferred basis for state vector reduction (cf. Ref. 7, where the energy eigenstates feature in this respect). The physical motivation for such a model is that the coherent states are precisely those from which a unique classical field configuration can be inferred in a quantum field theoretic context. In this way the coherent states play an important role in many quantization procedures (cf. Refs. 30, 24, and 36). We take the view therefore that coherent states are central in understanding the results of quantum measurement. Evidence for this exists in the theory of quantum optics, in which the empirically determined photon numbers obey the Poisson statistics of coherent states (cf. Ref. 14). The theory of positive operator-valued measures (Ref. 37) is important in this regard. As remarked in Sec. III, a POM provides a resolution of the unit operator among nonorthogonal states, and constitutes the spectrum for an associated measurement (a consistent theory of POM measurements has been developed in Refs. 37, 38, and 39). We have observed that the space of coherent states provides a natural POM, and that the underlying geometry of the coherent state manifold relates to the notion of quantum versus classical superposition in a physically intuitive way, as discussed in Sec. V. We envisage a stochastic process on the state manifold, with a drift oriented toward the coherent state submanifold $\mathcal{C}$, and a vanishing diffusion tensor on $\mathcal{C}$. An initial incoherent state is driven toward $\mathcal{C}$ via the stochastic evolution combined with drift, until $\mathcal{C}$ is reached, when ordinary unitary evolution proceeds according to the Hamiltonian flow. Thus $\mathcal{C}$ becomes an invariant attractor for the stochastic evolution (see Ref. 13, which discusses the preservation of coherent states under unitary evolution for a wide class of Hamiltonians). The stochastic differential geometry of processes on CP^n is well understood (cf. Ref. 7) and appears to generalize naturally to the infinite dimensional case. A suggestion for a stochastic differential equation governing this type of evolution is given in Chap. 8 of Ref. 11.

We have not addressed the issue of boundedness of the field operators associated with a given $\mathcal{H}^1$, and their kinematical representation. While the former can be dealt with by introducing the corresponding Weyl form of the canonical commutation relations (Refs. 40 and 41), the latter amounts to the choice of vacuum state together with a representation of the quantum complex structure J on some kinematical background, e.g., spacetime. For massless fields on Minkowski space and linearized gravity in vacuum, various representations of J and the associated metric g are given in Ref. 11.

An open problem is that of constructing a geometric phase space formulation of quantum field theory for which the state space is an arbitrary Kähler manifold. Our present view is that the study of stochastic processes on CP^{∞} may suggest a consistent measurement theory in which certain modifications of the state space geometry become necessary. We have seen that field theory based on a linear $\mathcal{H}^1$ can be formulated naturally in geometric terms. As occurs with spacetime in the passage from Maxwell's theory to Einstein's general relativity, the quantum state space may require global departures from the standard complex projective geometry and maximal isometry group, in order to accommodate the quantum description of nonlinear phenomena such as gravitation.

ACKNOWLEDGMENTS

The authors are grateful to B. D. Bramson, D. C. Brody, R. Penrose, H. Urbantke, and N. M. J. Woodhouse for helpful discussions.

[1] T. W. B. Kibble, Commun. Math. Phys. **64**, 73 (1978).
[2] T. W. B. Kibble, Commun. Math. Phys. **65**, 189 (1979).
[3] J. Anandan and Y. Aharonov, Phys. Rev. Lett. **65**, 1697 (1990).
[4] R. Cirelli, A. Manin, and L. Pizzocchero, J. Math. Phys. **31**, 2891 (1990).
[5] G. W. Gibbons, J. Geom. Phys. **8**, 147 (1992).
[6] L. P. Hughston, ''Geometric aspects of quantum mechanics,'' in *Twistor Theory*, edited by S. A. Huggett (Marcel Dekker, New York, 1995), Chap. 6.
[7] L. P. Hughston, ''Geometry of stochastic state vector reduction,'' Proc. R. Soc. London, Ser. A **452**, 953 (1996).
[8] A. Ashtekar and T. A. Schilling, ''Geometrical formulation of quantum mechanics,'' in *On Einstein's Path*, edited by A. Harvey (Springer-Verlag, Berlin, 1998).
[9] D. C. Brody and L. P. Hughston, Phys. Rev. Lett. **77**, 2851 (1996).
[10] D. C. Brody and L. P. Hughston, ''Geometric models for quantum statistical inference,'' in *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose*, edited by S. A. Huggett (Oxford University Press, Oxford, 1998), Chap. 18.
[11] T. R. Field, ''The quantum complex structure,'' D. Phil thesis, Oxford University, 1997.
[12] R. J. Glauber, ''Coherent and incoherent states of the radiation field,'' Phys. Rev. **131**, 2767 (1963).
[13] R. J. Glauber, ''Classical behaviour of systems of quantum oscillators,'' Phys. Lett. **21**, 650 (1966).
[14] J. R. Klauder and B. Skagerstam, *Coherent States* (World Scientific, New York, 1985), Chap. 1.
[15] E. Schrödinger, Naturwissenschaften **14**, 664 (1926).
[16] R. P. Geroch, ''Special topics in particle physics,'' unpublished lecture notes, University of Texas at Austin, 1971.
[17] R. P. Geroch, *Mathematical Physics* (Chicago University Press, Chicago and London, 1984).
[18] R. Penrose and W. Rindler, *Spinors and Space–Time* (Cambridge University Press, Cambridge, 1984), Vol. 1.
[19] R. P. Geroch, Ann. Phys. (N.Y.) **62**, 582 (1971).
[20] R. M. Wald, *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics* (Chicago University Press, City, 1994).
[21] R. F. Streater and A. S. Wightman, *PCT Spin and Statistics and All That* (Addison-Wesley, New York, 1980).
[22] T. Schilling, ''Geometry of quantum mechanics,'' Ph.D. thesis, The Pennsylvania State University, 1996.
[23] S. Kobayashi and K. Nomizu, ''Foundations of differential geometry,'' Wiley Classics Library, reissue (Wiley, New York, 1996), Vol. 2.
[24] N. M. J. Woodhouse, *Geometric Quantization* (Oxford University Press, New York, 1992).
[25] A. Ashtekar and A. Magnon-Ashtekar, ''A geometrical approach to external potential problems in quantum field theory,'' Gen. Relativ. Gravit. **12**, 205 (1980).
[26] P. A. M. Dirac, *The Principles of Quantum Mechanics*, 4th ed. (Oxford University Press, Oxford, 1958).
[27] D. A. Page, Phys. Rev. A **36**, 3479 (1987).
[28] J. A. Morrow and K. Kodaira, *Complex Manifolds* (Holt, Rinehart and Winston, New York, 1971).
[29] T. R. Field, ''Coherent states and Fubini–Study geometry,'' *Twistor Newsletter, Mathematical Institute, Oxford University* 1996, Vol. 41, p. 6.
[30] J. H. Rawnsley, Q. J. Math. Oxford, Ser. 2 **28**, 403 (1977).
[31] P. A. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1978).
[32] A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[33] R. Penrose, in *General Relativity and Gravitation 13. Part 1: Plenary Lectures*, edited by R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi (IOPP, Bristol/Philadelphia, 1993), p. 179.

[34] M. Ko, E. T. Newman, and R. Penrose, ''The Kähler structure of asymptotic twistor space,'' J. Math. Phys. **18**, 58 (1977).
[35] I. C. Percival, Proc. R. Soc. London, Ser. A **447**, 189 (1994).
[36] J. R. Klauder, ''Is quantization geometry?,'' IHES preprint, quant-ph/9604032, 1996.
[37] A. Peres, *Quantum Theory: Concepts and Methods* (Kluwer Academic, New York, 1995).
[38] A. Peres, ''How to differentiate between non-orthogonal states,'' Phys. Lett. A **128**, 19 (1988).
[39] I. D. Ivanovic, ''How to differentiate between non-orthogonal states, 2'' Phys. Lett. A **123**, 257 (1988).
[40] M. Reed and B. Simon, ''Methods of modern mathematical physics,'' *Functional Analysis* (Academic, New York, 1972), Vol. I, Chap. VIII.
[41] M. Reed and B. Simon, ''Methods of modern mathematical physics,'' *Fourier Analysis, Self-Adjointness* (Academic, New York, 1975), Vol. 2, appendix to Sec. X.7.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

A new type of loop independence and SU(N) quantum Yang–Mills theory in two dimensions

Christian Fleischhack[a)]
Institut für Theoretische Physik, Fakultät für Physik und Geowissenschaften, Universität Leipzig, Augustusplatz 10/11, 04109 Leipzig, Germany and Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstrasse 22-26, 04103 Leipzig, Germany

(Received 3 September 1998; accepted for publication 9 February 1999)

The expectation values of Wilson loop products for the pure Euclidean Yang–Mills theory on $\mathbb{R}\times\mathbb{R}$ given by Ashtekar *et al.* in the article ''SU(N) Quantum Yang–Mills Theory in Two Dimensions: A Complete Solution'' [J. Math. Phys. **38**, 5453 (1997)] are determined directly for all piecewise analytic loops. For that purpose we enlarge their calculations from quadratic lattices to general floating lattices introducing a new kind of loop independence and slightly modifying the regularization scheme. © *1999 American Institute of Physics.* [S0022-2488(99)01606-0]

I. INTRODUCTION

For quite a long time the quantization of Yang–Mills theories has been investigated. One of the main emphases is the approach via functional integration. The crucial point is the definition of an appropriate measure $d\mu$ on the space $\mathcal{A}/\mathcal{G}$ of all connections modulo gauge transformations. Heuristically one sets simply $d\mu := e^{-S(A)}\mathcal{D}A$, where $S(A)$ is the Yang–Mills action and $\mathcal{D}A$ is a kinematical measure on $\mathcal{A}/\mathcal{G}$, but the resulting mathematical problems are enormous. Some years ago, Ashtekar and Isham[1] developed an interesting idea to overcome these difficulties. They considered a certain completion of $\mathcal{A}/\mathcal{G}$, the compact Hausdorff space $\overline{\mathcal{A}/\mathcal{G}}$. Now, Ashtekar and Lewandowski[2] were able to construct a natural kinematical measure $d\mu_0$ corresponding to $\mathcal{D}A$, but the extension of S onto the whole $\overline{\mathcal{A}/\mathcal{G}}$ remained difficult. This problem was circumvented using the duality between measures on $\overline{\mathcal{A}/\mathcal{G}}$ and positive linear functionals on the space of all Wilson loop products. Using the lattice regularization, Thiemann[3] and Ashtekar *et al.*[4] defined these expectation values and received the measure $d\mu$.

Nevertheless, some technical problems remained open. The authors of Ref. 4 did not specify the type of hoop independence used for the projection $\overline{\mathcal{A}/\mathcal{G}}\to\mathbf{G}^n$. Both the strong independence and the weak independence[2] are not applicable—the former because obviously the lattice loops $\beta_{x,y}=\rho_{x,y}\Box_{x,y}\rho_{x,y}^{-1}$ cannot be strongly independent for lattices with more than two rows and columns, and the latter because then the integral would become illdefined.[2] Furthermore, the authors of Ref. 4 used the completeness of the plaquette loops $\beta_{x,y}$, i.e., that the subgroup of the hoop group generated by the $\beta_{x,y}$ coincides with the subgroup generated by all loops in the lattice. But, in general, the completeness is not guaranteed if one chooses arbitrary paths $\rho_{x,y}$ from the base point to the plaquette (x,y). So we will prove that there *exists* a choice for the $\rho_{x,y}$ such that the plaquette loops $\beta_{x,y}$ are complete. For the same reasons, the proof of the decomposition lemma, which ensures that any loop α without self-intersections can be expressed by a product of the loops corresponding to the plaquettes in the interior of α, has to be modified.

The present article is intended to provide these missing mathematical details. Moreover, we drop the restriction on quadratic lattices. We admit now any finite connected graph—a ''floating'' lattice—for the regularization. For this we slightly modify the regularization of the Yang–Mills action simply replacing a^2 (a...lattice spacing) by the area $|G|$ of the plaquette (see also Ref. 5) and

---

a)Electronic mail: Christian.Fleischhack@itp.uni-leipzig.de *or* Christian.Fleischhack@mis.mpg.de adapting the regularization to the given loops and not, as usual, the opposite. Thus, the use of floating lattices allows us to calculate the Wilson loop expectation values for all sets of hoops directly, i.e., without approximating them in a certain sense by loops in a quadratic lattice and without a subsequent (naive) limit. On the other hand we need a little bit more sophisticated—and, unfortunately, more technical—analysis, even if we would consider only quadratic lattices. At the beginning we define a new type of independence—the so-called moderate independence, which stands between the strong and the weak independence and is well suited to making the calculations mathematically rigorous. We prove that it is strong enough to make the integration calculus still applicable. Then we generalize the propositions in Ref. 4 to the case of floating lattices. The loops $\beta_{x,y}$ correspond now to the so-called flags f_G, i.e., loops that run from the base point m to the interior domain G—the generalized plaquette—traverse G once, and return to m. Choosing a flag to each interior domain we get a flag world. The crucial point is now the proof that there is a (moderately) independent and complete flag world for any graph. Moreover, the generalized decomposition lemma yields that, if one refines the underlying graph, any flag world can be naturally refined to a new (again moderately independent and complete) flag world and each flag f of the old flag world is a product of exactly the flags of the new one that correspond to domains in the interior of f.

By means of these propositions we can finally compute the Wilson loop expectation values reusing the calculations of Thiemann and Ashtekar *et al.*

II. PRELIMINARIES

In this section we summarize the basic facts about the space $\overline{\mathcal{A}/\mathcal{G}}$ of generalized connections modulo gauge transformations following Refs. 1, 2, and 4.

Let P be a fixed principal fiber bundle over the base manifold M with structure group $\mathbf{G}$ and m any fixed point in M. Furthermore, let $\{U_i\}$ be a covering of M,$\{\chi_i\}$ a trivialization of P over $\{U_i\}$ and j a fixed index with $m \in U_j$. In the following we suppose $\mathbf{G}$ to be either $\mathrm{SU}(N)$, $N \geqslant 2$, or $\mathrm{U}(1)$. Connections on P are described by their connection one-form A on P or, equivalently, their localized forms A_i on U_i. Similarly, we describe a gauge transformation by its corresponding equivariant map $\rho: P \to \mathbf{G}$ or its localized forms $\rho_i : U_i \to \mathbf{G}$. We will only consider C^∞ connections and C^∞ gauge transformations. The spaces of all connections and all gauge transformations are denoted by $\mathcal{A}$ and $\mathcal{G}$, respectively, and their quotient with respect to the natural action of $\mathcal{G}$ on $\mathcal{A}$ is denoted by $\mathcal{A}/\mathcal{G}$.

Next, we define $\mathcal{L}_m$ to be the set of all piecewise analytic loops in M with base point m, i.e., all piecewise analytic maps $\alpha:[0,1] \to M$, $\alpha(0) = \alpha(1) = m$. Two loops α_1 and α_2 are multiplied by

$$\alpha_1 \circ \alpha_2(t) := \begin{cases} \alpha_1(2t) & \text{for } t \in [0, \frac{1}{2}] \\ \alpha_2(2t-1) & \text{for } t \in [\frac{1}{2}, 1] \end{cases}.$$

Note, that $\circ$ is not associative. For any $\alpha \in \mathcal{L}_m$ and $A \in \mathcal{A}$ we define the holonomy $h_\alpha(A) = h_A(\alpha) = h(\alpha, A) \in \mathbf{G}$ as the group element, which corresponds to the parallel transport with respect to A of $\chi_j^{-1}(m, e_{\mathbf{G}})$ along α. In the trivialization χ_j we have $h(\alpha, A) = \mathcal{P} \exp - (\oint_\alpha A_j)$ if α is completely contained in U_j. A change of the trivialization yields only a conjugation of $h(\alpha, A)$ independent of α. (Since such a conjugation is irrelevant for our purpose, we fix now a certain chart U_j for the computation of the holonomies.) Moreover, we have $h_\alpha h_\beta = h_{\alpha \circ \beta}$ for all $\alpha, \beta \in \mathcal{L}_m$.

The fundamental idea of Ashtekar and Isham was to use the description of connections by the traces of their holonomies, the so-called Wilson loops. First, they defined an equivalence relation on $\mathcal{L}_m$. Two loops $\alpha_1, \alpha_2 \in \mathcal{L}_m$ are said to be holonomically equivalent $\alpha_1 \sim \alpha_2$ iff $h_{\alpha_1}(A) = h_{\alpha_2}(A)$ for any $A \in \mathcal{A}$. The equivalence classes $[\alpha]$ are called hoops. [In the following we often drop the brackets. Then the equal sign (=) means the equality of loops and the symbol $\sim$ means equality of hoops.] The hoop group $\mathcal{HG}$ is the set of all hoops with the well-defined projected multiplication of $\mathcal{L}_m:[\alpha_1]\circ[\alpha_2]=[\alpha_1\circ\alpha_2]$ and $[\alpha]^{-1}=[\beta]$ with $\beta(t)=\alpha(1-t)$. For instance, two loops are holonomically equivalent if they can be obtained from each other by reparametrization or insertion of retracings. Second, Ashtekar and Isham made use of the so-called Wilson loops $T_\alpha:\mathcal{A}\to\mathbb{C}$ defined by $T(\alpha,A)=T_\alpha(A)=(1/N)\operatorname{tr} h_\alpha(A)$. Obviously, T factorizes over $\sim$ and $\mathcal{G}$, i.e., $T:\mathcal{HG}\times\mathcal{A}/\mathcal{G}\to\mathbb{C}$. Next, they defined the algebra $\mathcal{HA}:=\{f:\mathcal{A}/\mathcal{G}\to\mathbb{C}|f=\Sigma_{i=1}^n c_i\Pi_{j_i=1}^{n_i} T_{\alpha_{j_i}}|n,n_i\in\mathbb{N},c_i\in\mathbb{C}\}$ of all finite linear combinations of finite products of Wilson loops and called its completion $\overline{\mathcal{HA}}$ with respect to the sup-norm on $\mathcal{A}/\mathcal{G}$ holonomy algebra. Clearly, $\overline{\mathcal{HA}}$ is a commutative $\mathcal{C}^*$ algebra. This allows to use the powerful tools provided by the theory of $\mathcal{C}^*$ algebras. Due to the Gelfand–Naimark theorem there exists a compact Hausdorff space $\mathcal{M}(\overline{\mathcal{HA}})$, the space of all characters of $\overline{\mathcal{HA}}$, i.e., all nontrivial, linear, multiplicative functionals on $\overline{\mathcal{HA}}$, such that $\overline{\mathcal{HA}}\cong\mathcal{C}(\mathcal{M}(\overline{\mathcal{HA}}))$. Giles[6] had proved that given all Wilson loops one can reconstruct the corresponding connection up to a gauge transformation. Rendall[7] observed that, therefore, $\mathcal{A}/\mathcal{G}$ can be densely embedded into $\mathcal{M}(\overline{\mathcal{HA}})$. This justifies the Ashtekar–Isham definition $\overline{\mathcal{A}/\mathcal{G}}:=\mathcal{M}(\overline{\mathcal{HA}})$ of the space of the generalized connections modulo gauge transformations. The elements of $\overline{\mathcal{A}/\mathcal{G}}$ are denoted by $\bar{A}$. The isomorphism between $\overline{\mathcal{HA}}$ and $\mathcal{C}(\mathcal{M}(\overline{\mathcal{HA}}))$ is given by the Gelfand transformation

$$\sim:\overline{\mathcal{HA}}\to C(\overline{\mathcal{A}/\mathcal{G}})\quad\text{with}\quad \tilde{f}:\overline{\mathcal{A}/\mathcal{G}}\to\mathbb{C},$$

$$f\mapsto\tilde{f},\quad \bar{A}\mapsto\bar{A}(f).$$

The theory of $\mathcal{C}^*$ algebras yields also the measure theory and representation theory on $\overline{\mathcal{A}/\mathcal{G}}$. There is a one-to-one correspondence between Borel measures μ on $\overline{\mathcal{A}/\mathcal{G}}$, linear continuous positive functionals F on $\overline{\mathcal{HA}}$, and continuous cyclic Hilbert space representations ϕ of $\overline{\mathcal{HA}}$. More precisely, any such functional F can be obtained by $F(f)=\int_{\overline{\mathcal{A}/\mathcal{G}}}\tilde{f}\,d\mu_F$ with a certain unique Borel measure μ_F and any such ϕ is unitary equivalent to the representation φ of $\overline{\mathcal{HA}}$ on $L^2(\overline{\mathcal{A}/\mathcal{G}},d\mu_\phi)$ by multiplication operators $\varphi(f)\psi=\tilde{f}\cdot\psi$ with a certain measure μ_ϕ.

Ashtekar and Lewandowski[2] (in the following denoted by AL) discovered a second description of $\overline{\mathcal{A}/\mathcal{G}}$ via the hoop group $\mathcal{HG}$. (Marolf and Mourão[8] obtained a third description of $\overline{\mathcal{A}/\mathcal{G}}$ via projective limits. However, this approach is unimportant for our purpose and we only mention it for completeness.) They defined two kinds of independence on $\mathcal{L}_m$. A finite subset $\boldsymbol{\beta}:=\{\beta_i\}$ of $\mathcal{L}_m$ is called strongly independent iff each β_i contains an open segment which is traced once and only once by β_i and which is intersected by the remaining β_j at most in a finite set of points. $\boldsymbol{\beta}$ is weakly independent iff to any $(g_1,\ldots,g_n)\in\mathbf{G}^n$ there exists an $A\in\mathcal{A}$ such that $h_{\beta_i}(A)=g_i$ for all i. They proved that strong independence implies weak independence. Then they could give a bijection between $\overline{\mathcal{A}/\mathcal{G}}$ and the space $\operatorname{Hom}(\mathcal{HG},\mathbf{G})/\operatorname{Ad}$ of *all* homomorphisms from $\mathcal{HG}$ to $\mathbf{G}$ modulo a hoop independent conjugation. More precisely, any $h\in\operatorname{Hom}(\mathcal{HG},\mathbf{G})/\operatorname{Ad}$ yields an $\bar{A}_h\in\overline{\mathcal{A}/\mathcal{G}}$ via $\bar{A}_h(T_\alpha):=(1/N)\operatorname{tr} h(\alpha)$ and vice versa.

This graph-theoretical approach was used by AL to define a natural integration measure, the so-called induced Haar measure.[2] They introduced an equivalence relation on $\overline{\mathcal{A}/\mathcal{G}}$ for finitely generated subgroups $\mathcal{HG}(\boldsymbol{\beta})\subseteq\mathcal{HG}$:$\bar{A}_1\sim\bar{A}_2$ with respect to $\mathcal{HG}(\beta)$ iff $h_{\bar{A}_1}(\gamma)=g^{-1}h_{\bar{A}_2}(\gamma)g$ for all $\gamma\in\mathcal{HG}(\boldsymbol{\beta})$ with a (hoop independent) $g\in\mathbf{G}$. $\pi_\beta\overline{\mathcal{A}/\mathcal{G}}:\to\overline{\mathcal{A}/\mathcal{G}}\sim$ is the corresponding projection. Thus, there is a bijection $\overline{\mathcal{A}/\mathcal{G}}\sim\leftrightarrow\operatorname{Hom}(\mathcal{HG}(\boldsymbol{\beta}),\mathbf{G})/\operatorname{Ad}$ as for $\overline{\mathcal{A}/\mathcal{G}}$ and $\operatorname{Hom}(\mathcal{HG},\mathbf{G})/\operatorname{Ad}$. $\operatorname{Hom}(\mathcal{HG}(\beta),\mathbf{G})/\operatorname{Ad}$ itself is isomorphic to $\mathbf{G}^{\#\boldsymbol{\beta}}/\operatorname{Ad}$ if $\boldsymbol{\beta}$ is weakly independent. Therefore AL could reduce the integration over $\overline{\mathcal{A}/\mathcal{G}}$ under certain circumstances to the case of the integration over a finite dimensional Lie group. In detail, they defined cylindrical functions, i.e., functions f being pullbacks $\pi_{\boldsymbol{\beta}}f_{\boldsymbol{\beta}}$ of continuous functions $f_{\boldsymbol{\beta}}$ on $\operatorname{Hom}(\mathcal{HG}(\beta),\mathbf{G})/\operatorname{Ad}=\mathbf{G}^{\#\boldsymbol{\beta}}/\operatorname{Ad}$ with strongly independent β and showed that the set $\mathcal{C}$ of all such functions is dense in $\overline{\mathcal{HA}}=\mathcal{C}(\overline{\mathcal{A}/\mathcal{G}})$. Now,

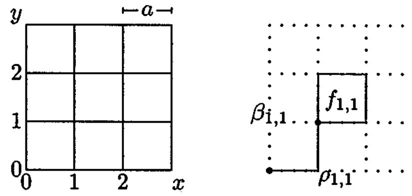

FIG. 1. Example of a lattice ($l=3$) and the loop $\beta_{1,1}$.

they defined $\int_{\overline{\mathcal{A}/\mathcal{G}}} f\, d\mu_0 := \int_{\mathbf{G}^{\#\beta}/\mathrm{Ad}} f_{\boldsymbol{\beta}} d\mu_{\boldsymbol{\beta}}$ and chose $d\mu_{\boldsymbol{\beta}}$ to be the Haar measure for each $\boldsymbol{\beta}$. Thus they got a well-defined, regular and positive measure μ_0 on $\overline{\mathcal{A}/\mathcal{G}}$, the so-called induced Haar measure.

Ashtekar and Lewandowski realized that μ_0 could serve as a kinematical measure of physical theories in the functional integral approach. Since the elements of $\mathcal{A}/\mathcal{G}$ are classical potential configurations, the completion $\overline{\mathcal{A}/\mathcal{G}}$ seems to be a candidate for the space of histories in the quantum regime and the physical measure is built from $d\mu_0$ by multiplication with e^{-S}, where S is the physical action of the theory. The crucial point was to choose such an S defined not only on $\mathcal{A}/\mathcal{G}$ but on $\overline{\mathcal{A}/\mathcal{G}}$. Neglecting that fact, one could compute via $\langle f\rangle = \int e^{-S} f\, d\mu_0$ any expectation value of the theory supposed f to be a function on $\overline{\mathcal{A}/\mathcal{G}}$. Thiemann[3] and Ashtekar *et al.*[4] (in the following denoted by TA^+) proposed a solution of that problem in the case of the two-dimensional quantum Yang–Mills theory using lattice regularization. The main problem was the replacement of the Yang–Mills action $S_{\mathrm{YM}} = \frac{1}{4}\int_M F_{\mu\nu}F^{\mu\nu}\, dx$ by an expression whose domain is $\overline{\mathcal{A}/\mathcal{G}}$. The only *a priori* available quantities are the generalized holonomies. This indicates the use of Wilson's lattice regularization. For this one places a finite quadratic lattice with spacing a and length R on the 2-plane and defines $S_{\mathrm{YM}}^{\mathrm{req}} = (N/g^2a^2)\Sigma_\Box(1-(1/N)\mathrm{Re}\,\mathrm{tr}\, h_\Box)$ where the sum goes over all plaquettes of the lattice. $h_\Box$ denotes the holonomy around the plaquette $\Box$. In the limit $a\to 0$ and $R\to\infty$ one can show naively the regularized action to converge to S_{YM}. The advantage of $S_{\mathrm{YM}}^{\mathrm{req}}$ is its natural extendability to $\overline{\mathcal{A}/\mathcal{G}}$. Now, TA^+ could compute the expectation values of the Wilson loops expected to determine the whole pure quantum YM_2 theory:

$$\langle T_{\alpha_1}\cdots T_{\alpha_n}\rangle = \frac{1}{Z}\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_0\, e^{-\lim_{\alpha\to 0,R\to\infty} S_{\mathrm{YM}}^{\mathrm{reg}}} T_{\alpha_1}\cdots T_{\alpha_n} = \lim_{a\to 0,R\to\infty}\frac{1}{Z_{a,R}}\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_0\, e^{-S_{\mathrm{YM}}^{\mathrm{reg}}} T_{\alpha_1}\cdots T_{\alpha_n} \tag{1}$$

after exchanging limit and integral. (The factor $1/Z$ guarantees $\langle T_1\rangle = 1$.) Afterwards they expressed each loop $\alpha_1,\ldots,\alpha_n$ and each plaquette loop $\Box$ by a product of ‘‘simple’’ loops (i.e., loops traversing exactly one plaquette and connecting it with the base point m by conjugation), provided, however, $\alpha_1,\ldots,\alpha_n$ are contained in the lattice. Under the assumption that these loops are independent they could reduce the integration over $\overline{\mathcal{A}/\mathcal{G}}$ to the integration over G^n, n finite. Finally, they computed the integrals explicitly and got an algebraic expression depending only on the areas enclosed by the loops. For general $\alpha_1,\ldots,\alpha_n$ they suggested to approximate these loops naively by lattice loops and to consider the limit of the expectation values, but this is simply given by the limit of the enclosed areas.

III. MODERATE INDEPENDENCE

In this section we will introduce a new type of independence being crucial for the considerations below—the so-called moderate independence.

A. Why a new type of independence?

We consider a quadratic lattice with spacing a and length $R = la$, $l\in\mathbb{N}^+$, i.e., with l^2 plaquettes, see, e.g., Fig. 1. Now we assign (see Ref. 4) a loop $\beta_{x,y} := \rho_{x,y}\circ f_{x,y}\circ\rho_{x,y}^{-1}$ to each plaquette $\Box_{x,y}$. x,y indicates the position of the plaquette, as follows: First, choose a path $\rho_{x,y}$

from the base point m to the bottom left-hand corner (x,y) and then define $\beta_{x,y}:=\rho_{x,y}\circ f_{x,y}\circ\rho_{x,y}^{-1}$ where $f_{x,y}$ is a path traversing $\Box_{x,y}$ counterclockwise. For our example, we choose $\rho_{x,y}$ to consist of a horizontal and a subsequent vertical path as in Fig. 1.

Obviously, the set $\boldsymbol{\beta}$ of all these loops $\beta_{x,y}$ is not strongly independent (for the exact definition see Sec. III B) because, e.g., $\beta_{1,1}$ does not have a segment which is intersected by any other $\beta_{x,y}$ at most in a finite number of points. Of course, one can prove that $\boldsymbol{\beta}$ is weakly independent, but this is not sufficient to allow the application of the integration calculus. Therefore we need a third type of independence between these two ones; this will be the moderate independence.

B. Moderate independence: Definition and position among the independencies

In the following, $\boldsymbol{\beta}$ denotes any finite subset $\{\beta_i\}$ of $\mathcal{L}_m$ (or $\mathcal{HG}$) and $\mathcal{HG}(\boldsymbol{\beta})$ the subgroup of $\mathcal{HG}$ generated by $\boldsymbol{\beta}$. (To avoid technical complications we set $\mathcal{HG}(\emptyset)=\{[1]\}$.) First, we recall the definition of the strong independence.[2]

Definition 3.1: Strong independence in $\mathcal{L}_m$. $\boldsymbol{\beta}\subseteq\mathcal{L}_m$ is *strongly independent* iff any $\beta_i\in\boldsymbol{\beta}$ contains an open segment e_i, the so-called *free segment*, traced exactly once by β_i and intersected by any β_j, $j\neq i$, in at most a finite number of points. (The intersection condition can be replaced by "$e_i\cap\beta_j=\emptyset \quad \forall_j\neq i$." However, this yields to an equivalent definition.)

Our definition of the moderate independence differs very little from the previous one. We only replace $j\neq i$ by $j<i$.

Definition 3.2: Moderate Independence in $\mathcal{L}_m$. $\boldsymbol{\beta}\subseteq\mathcal{L}_m$ is *moderately independent* iff any $\beta_i\in\boldsymbol{\beta}$ contains an open segment e_i, the so-called *free segment*, traced exactly once by β_i and intersected by any β_j, $j<i$, in at most a finite number of points. (The remark in Definition 3.1 holds analogously in the case of moderate independence: "$e_i\cap\beta_j=\emptyset \quad \forall_j<i$".)

We have simply replaced the rigid condition of a simultaneous freeness of segments by the flexible condition of an iterative freeness. We will see that this keeps the integration calculus valid and makes the set of all plaquette loops (cf. Fig. 1) independent.

We mention that the simple specification of the elements of a moderately independent set $\boldsymbol{\beta}$ is not sufficient. If we say "$\boldsymbol{\beta}$ is moderately independent" then there is an order of the elements $\beta_i\in\boldsymbol{\beta}$, such that the above criterion is valid. Analogously, the specification "$\{\beta_1,\beta_2\}$ or $\{\beta_2,\beta_1\}$, respectively, are moderately independent" should be clear.

Finally, we recall the definition of weak independence.[2]

Definition 3.3: Weak Independence in $\mathcal{L}_m$. $\boldsymbol{\beta}\subseteq\mathcal{L}_m$ is *weakly independent* iff for any $(g_1,\ldots,g_n)\in\mathbf{G}^n$, $n=\#\boldsymbol{\beta}$, there is an $A\in\mathcal{A}$, such that $h_{\beta_i}(A)=g_i$ for all $i=1,\ldots,n$.

Obviously, this kind of independence can be extended from $\mathcal{L}_m$ to $\mathcal{HG}$.

Instead of the previous two definitions being graph-theoretical we have here an algebraic condition. Weak independence of $\boldsymbol{\beta}$ means no relations between the holonomies h_{β_i} and so it ensures the freeness of the corresponding subgroup $\mathcal{HG}(\boldsymbol{\beta})\subseteq\mathcal{HG}$, see Sec. III C.

The position of the moderate independence clarifies the next

Proposition 3.1: $\boldsymbol{\beta}$ strongly independent $\Rightarrow\boldsymbol{\beta}$ moderately independent $\Rightarrow\boldsymbol{\beta}$ weakly independent.

Proof: (1) The first implication is obvious. (2) The proof of the second implication is technical and can be found in the Appendix. **qed**

C. Algebraic consequences of the weak independence

Proposition 3.2: Let $\boldsymbol{\beta}\subseteq\mathcal{HG}$ be weakly independent. Then the following holds:

(1) $\mathcal{HG}(\boldsymbol{\beta})$ is freely generated by $\boldsymbol{\beta}$. [In the case $\mathbf{G}=U(1)$ we understand by "free" anytime "Abelian free."]

(2) Let there be given a $\boldsymbol{\gamma}\subseteq\mathcal{HG}$, such that $\mathcal{HG}(\beta)=\mathcal{HG}(\boldsymbol{\gamma})$. Then we have: $\boldsymbol{\gamma}$ is weakly independent$\Leftrightarrow\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have the same cardinality.

Proof: (1) See Ref. 2.

(2) $\Leftarrow$

(a) For $\mathcal{HG}(\boldsymbol{\gamma})=\mathcal{HG}(\boldsymbol{\beta})$ there are expressions

$$\gamma_i \sim \prod_{k_i=1}^{K_i} \beta_{j(i,k_i)}^{\epsilon(i,k_i)} \quad \text{and} \quad \beta_j \sim \prod_{l_j=1}^{L_j} \gamma_{i(j,l_j)}^{\eta(j,l_j)}$$

for any $i,j \in [1,n]$, $n := \#\boldsymbol{\beta} = \#\boldsymbol{\gamma}$, and thus

$$\gamma_i \sim \prod_{k_i=1}^{K_i} \left(\prod_{l_{j(i,k_i)}=1}^{L_{j(i,k_i)}} \gamma_{i(j(i,k_i)l_{j(i,k_i)})}^{\eta(j(i,k_i)l_{j(i,k_i)})} \right)^{\epsilon(i,k_i)} \qquad \forall i \in [1,n].$$

(b) Due to the first point $\boldsymbol{\beta}$ is a free system of generators for $\mathcal{HG}(\boldsymbol{\beta})$. Since $\boldsymbol{\gamma}$ also generates $\mathcal{HG}(\boldsymbol{\beta})=\mathcal{HG}(\boldsymbol{\gamma})$ and $\#\boldsymbol{\gamma}=\#\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ is a free system of generators for $\mathcal{HG}(\boldsymbol{\beta})=\mathcal{HG}(\boldsymbol{\gamma})$ (Ref. 9).

(c) Let there be given $(g_1,\ldots,g_n) \in \mathbf{G}^n$ and let $\mathbf{H}$ be the group generated by $\{g_1,\ldots,g_n\}$. Since $\mathcal{HG}(\boldsymbol{\gamma})$ has the free rank n there is (Ref. 9) an epimorphism $\pi:\mathcal{HG}(\boldsymbol{\gamma}) \rightarrow \mathbf{H}$ with $\pi(\gamma_i)=g_i$.

(d) Since $\boldsymbol{\beta}$ is weakly independent, there exists an $A \in \mathcal{A}$ with $h_{\beta_j}(A)=\prod_{l_j=1}^{L_j} g_{i(j,l_j)}^{\eta(j,l_j)} \forall_j$, i.e., we have for all $i \in [1,n]$

$$\begin{aligned} h_{\gamma_i}(A) = h_{\prod_{k_i=1}^{K_i} \beta_{j(i,k_i)}^{\epsilon(i,k_i)}}(A) &= \prod_{k_i=1}^{K_i} \left(\prod_{l_j(i,k_i)=1}^{L_{j(i,k_i)}} g_{i(j(i,k_i),l_{j(i,k_i)})}^{\eta(j,(i,k_i),l_{j(i,k_i)})} \right)^{\epsilon(i,k_i)} \\ &= \pi\left(\prod_{k_i=1}^{K_i} \left(\prod_{l_{j(i,k_i)}=1}^{L_{j(i,k_i)}} \gamma_{i(j(i,k_i),l_{j(i,k_i)})}^{\eta(j(i,k_i),l_{j(i,k_i)})} \right)^{\epsilon(i,k_i)} \right) = \pi(\gamma_i) = g_i . \end{aligned}$$

Thus, $\boldsymbol{\gamma}$ is weakly independent.

$\Rightarrow$

Let $\boldsymbol{\gamma}$ be weakly independent, i.e., $\mathcal{HG}(\boldsymbol{\gamma})=\mathcal{HG}(\boldsymbol{\beta})$ is free. Consequently, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have the same cardinality.[9] **qed**

D. Graphs and loops

We recall some fundamental facts about graphs (see, e.g., Ref. 10).

A graph (X,X_0) consists of a Hausdorff space X and a discrete subspace X_0, the space of the so-called vertices. $X \backslash X_0$ is a disjoint union of edges, i.e., open subsets e_i isomorphic to the interval (0, 1). e_i can connect one or two vertices. In the first case e_i is called sling. Two vertices are connected by a multiple edge iff there are at least two different edges connecting these vertices. Iff a graph has neither slings nor multiple edges, it is called ordinary. Furthermore, (X,X_0) is finite iff both the set of edges and the set of vertices are finite. A graph (X',X_0') is called subgraph (or refinement) of a graph (X,X_0) iff $X' \subseteq X$ and $X_0' \subseteq X_0$. Obviously, any (finite) graph is subgraph of an ordinary (finite) graph. In the following we will briefly denote a graph by X instead of (X,X_0). Additionally, $X \leqslant X'$ means that X is a subgraph of X'.

In a natural way one can choose an orientation to any edge. The initial (terminal) vertex of an edge e is denoted by $v_e^-(v_e^+)$. A path f in a graph is a finite sequence of (oriented) edges $(e_1,\ldots,e_n)$, $n \leqslant 0$, such that the terminal vertex of e_i coincides with the initial vertex of e_{i+1} $(1 \leqslant i < n)$ with respect to the chosen orientation. Iff $n=0$, f is called trivial. Iff the initial vertex v_f^- and the terminal vertex v_f^+ of f are equal, f is called closed path or loop with base point $v_f = v_f^\pm$. f is called reduced iff no edge is retraced immediately and is called genuine iff no vertex is traced twice (exception: initial and terminal vertex can be equal). Finally, a tree T is a graph without any nontrivial genuine closed path.

Obviously, any graph contains trees. If we partially order the set of all trees in a graph using the inclusion, i.e., subgraph relation, we get

Lemma 3.3: Any tree in a graph X is contained in a maximal tree in X. If X is connected, then a tree T in X is maximal if and only if T contained all vertices of X.
Using this lemma one can construct explicitly the fundamental group of a connected graph. First choose a vertex v_0 and a maximal tree. Let $\{e_\lambda | \lambda \in \Lambda\}$ be the set of all edges of X not contained in T and choose an orientation for each e_λ. Now denote by t_λ^- and t_λ^+ the (unique) reduced path along T from v_0 to the initial vertex of e_λ and, respectively, from the terminal vertex of e_λ to v_0. Finally, define α_λ to be the product of t_λ^-, e_λ and t_λ^+. We have

Proposition 3.4: The fundamental group $\pi(X,v_0)$ is the free group generated by $\{\alpha_\lambda | \lambda \in \Lambda\}$, where α_λ denotes here not the loop itself, but its homotopy class.
The Euler–Poincaré characteristic $\chi(X)$ of a finite graph is per definition the difference of the number of vertices and the number of edges.

Proposition 3.5: Let X be finite and connected. Then $\pi(X,v_0)$ is a free group with $1-\chi(X)$ generators and X is a tree iff $\chi(X)=1$.
Let there be given now a finite set of loops $\boldsymbol{\beta}=\{\beta_i\}\subseteq\mathcal{L}_m$ in a manifold M. Note that $\mathcal{L}_m$ contains only piecewise analytic loops. The image of $\boldsymbol{\beta}$ in M defines naturally a finite connected graph $\Gamma_{\boldsymbol{\beta}}$ via the following (see also Ref. 2).

Construction 3.4: (1) Mark all end points of overlapping intervals of two loops and all intersection points outside those overlapping intervals. These points become the vertices of $\Gamma_{\boldsymbol{\beta}}$. Due to the piecewise analyticity the number of vertices is finite.

(2) Divide any β_i into paths between ''neighboring'' vertices and call these paths edges of $\Gamma_{\boldsymbol{\beta}}$. Again due to the piecewise analyticity the set of edges is finite.

(3) Since any β_i is a loop with base point m, $\Gamma_{\boldsymbol{\beta}}$ is connected.

E. Relations between the fundamental group and the hoop group of a graph

In this section Γ is a finite connected graph and m an arbitrary, but fixed vertex of Γ. Furthermore, we denote by $\mathcal{HG}(\Gamma)$ the subgroup of $\mathcal{HG}$ generated by all loops in Γ.

It was an important observation of Ashtekar and Lewandowski[2] that there is a close relation between the representation of a loop as a hoop and as an equivalence class with respect to the homotopy in a graph. In detail, they got

Lemma 3.6: Two homotopically equivalent loops are holonomically equivalent, i.e., there is an epimorphism $\phi:\pi(\Gamma,m)\rightarrow\mathcal{HG}(\Gamma)$. ϕ is an isomorphism if $\mathbf{G}=\mathrm{SU}(N)$. For $\mathbf{G}=\mathrm{U}(1)$ we have $\ker\phi=[\pi(\Gamma,m),\pi(\Gamma,m)]$.
Consequently, in the case $\mathbf{G}=\mathrm{SU}(N)$ two loops are holonomically equivalent if and only if they can be obtained from each other by reparametrizations or (if necessary successively) canceling retracings. Obviously, we have

Lemma 3.7: Let T be a maximal tree and $\{\alpha_\lambda\}$ the set of the corresponding generators of $\pi(\Gamma,m)$ as in Proposition 3.4. Then $\{\alpha_\lambda\}$ is strongly independent and complete in Γ, i.e., we have $\mathcal{HG}(\{\alpha_\lambda\})=\mathcal{HG}(\Gamma)$.
The free segments are the edges e_λ not contained in T. Additionally, one can express any finite set of hoops by a finite set of strongly independent hoops.[2]

Lemma 3.8: For any finite set $[\boldsymbol{\beta}]$ of hoops there is a set $\boldsymbol{\alpha}\subseteq\mathcal{L}_m$, such that

(1) $\mathcal{HG}(\boldsymbol{\beta})\subseteq\mathcal{HG}(\boldsymbol{\alpha})$,
(2) $\boldsymbol{\alpha}$ is strongly independent, and
(3) $\#\boldsymbol{\alpha}=\mathrm{rank}\,\pi(\Gamma_{\boldsymbol{\beta}},m)$.

For this choose the natural graph $\Gamma_{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Choose now some generating set $\boldsymbol{\alpha}$ of the fundamental group $\pi(\Gamma_{\boldsymbol{\beta}},m)$. Obviously, $\boldsymbol{\alpha}$ fulfills the required conditions.

Now we want to investigate the independence of loops.

Lemma 3.9: Let n be the rank of $\pi(\Gamma,m)$. Then any $\boldsymbol{\beta}\subseteq\mathcal{L}_m$ with $\#\boldsymbol{\beta}=n$ and $\mathcal{HG}(\boldsymbol{\beta})=\mathcal{HG}(\Gamma)$ is weakly independent.

Proof: Choose any maximal tree T in Γ and a corresponding system $\{\alpha_\lambda\}$ of generators of $\pi(\Gamma,m)$. $\{\alpha_\lambda\}$ has n elements and is a free generating system. Due to Lemma 3.7 $\{\alpha_\lambda\}$ is strongly independent and thus weakly independent. Proposition 3.2 finishes the proof. **qed**

Generally, one cannot conclude that $\boldsymbol{\beta}$ is even moderately independent. To see this let $\pi(\Gamma,m)$ be generated by two loops α_1,α_2 as in Proposition 3.4. Set $\beta_1:=\alpha_1\alpha_2\alpha_1^{-1}$ and $\beta_2:=\alpha_1\alpha_2$.

(1) We have $\mathcal{HG}(\boldsymbol{\beta})=\mathcal{HG}(\Gamma)=\mathcal{HG}(\{\alpha_1,\alpha_2\})$, because $\alpha_1=\beta_1^{-1}\beta_2$ and $\alpha_2=\beta_2^{-1}\beta_1\beta_2$.

(2) Suppose $\{\beta_1,\beta_2\}$ are moderately independent. Any segment of β_2 is already traced by β_1. This is a contradiction to the assumption β_2 has a free segment.

The case ''$\{\beta_2,\beta_1\}$ are moderately independent'' yields an analogous contradiction. Thus $\boldsymbol{\beta}$ is not moderately independent.

We finish this section with a criterion for the completeness of loops in a given graph.

Proposition 3.10: Let Γ be a finite connected graph and $\boldsymbol{\beta}$ a set of moderately independent loops in Γ. Then is $\boldsymbol{\beta}$ complete with respect to Γ if and only if the cardinality of $\boldsymbol{\beta}$ equals the rank of $\pi(\Gamma,m)$.

Proof: The $\Rightarrow$ direction is simple. Due to Lemma 3.8 there is a set $\boldsymbol{\alpha}$ with $\mathcal{HG}(\alpha)=\mathcal{HG}(\Gamma)=\mathcal{HG}(\boldsymbol{\beta})$, whose cardinality is just equal to the rank of $\pi(\Gamma,m)$. Proposition 3.2 yields that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have the same cardinality.

The $\Leftarrow$ direction is a little bit technical.

The free segments of the β_i are as usual denoted by e_i and the cardinality of $\boldsymbol{\beta}$ by n. Suppose first that no β_i has a retracing interval.

(1) W.l.o.g. the free segments e_i of β_i are edges of Γ. Otherwise, if necessary, restrict any e_i, such that it is still contained in only one edge k_i. Since β_i has no retracing intervals, the whole k_i is a free segment of β_i. Thus one can set $e_i:=k_i$.

(2) The graph $T:=\Gamma\setminus\cup_{i=1}^n\{e_i\}$ created by removing all free segments is again a connected graph.

Set $\Gamma_j:=\Gamma\setminus\cup_{i=j}^n\{e_i\}$. Then $\Gamma_{n+1}=\Gamma$, $\Gamma_1=T$. Due to the moderate independence of the β_i we have $\beta_i\cap e_{\tau'}=\emptyset\ \ \forall i'>i$, i.e., β_i is a loop in Γ_{i+1}. Suppose T is not connected. Then there would exist a $j\in[1,n]$, such that all Γ_i with $i>j$ are connected, but Γ_j is not connected. Since β_j is a loop in Γ_{j+1} and β_j passes e_j,β_j has to pass vertices of both connected components of $\Gamma_j=\Gamma_{j+1}\setminus\{e_j\}$. Thus e_j must be passed at least once in each direction by β_j, i.e., we have a contradiction to the assumption that e_j is a free segment. Thus, Γ_i is connected for all $i\in[1,n+1]$.

(3) T is a maximal tree in Γ.

Due to Proposition 3.5 we have $n=\operatorname{rank}\pi(\Gamma,m)=1-\chi(\Gamma)=1-\epsilon_\Gamma+\kappa_\Gamma$, where ϵ_Γ and κ_Γ are the numbers of vertices and edges of Γ, respectively. Since $T=\Gamma\setminus\cup_{i=1}^n\{e_i\}$ we have $\kappa_T=\kappa_\Gamma-n$ and obviously $\epsilon_T=\epsilon_\Gamma$. For T connected, we have $\chi(T)=\epsilon_T-\kappa_T=\epsilon_\Gamma-\kappa_\Gamma+n=\chi(\Gamma)+n=1$. Thus T is a tree in Γ due to Proposition 3.5. T is even maximal because T contains all vertices of Γ.

(4) Let $\boldsymbol{\alpha}:=\{\alpha_i\}$ be a free system of generators of $\pi(\Gamma,m)$ due to Proposition 3.4 for the just constructed maximal tree T and the edges $\{e_i\}$. Thus, $\boldsymbol{\alpha}$ fulfills $\mathcal{HG}(\boldsymbol{\alpha})=\mathcal{HG}(\Gamma)$. W.l.o.g. α_i traces the edge e_i in the same direction as β_i. We show that $\boldsymbol{\beta}$ is complete in Γ.

(a) β_1 is a loop in $T\cup\{e_1\}=\Gamma_{1+1}$, where e_1 is traced once and in the same direction as α_1 is. Thus $\beta_1=t_+e_1t_-\sim\alpha_1$ with certain paths $t_\pm$ in T, i.e., $\mathcal{HG}(\{\beta_1\})=\mathcal{HG}(\{\alpha_1\})$, i.e., $\{\beta_1\}$ is complete in Γ_{1+1}.

(b) Let $\mathcal{HG}(\{\beta_1,\ldots,\beta_i\})=\mathcal{HG}(\Gamma_{i+1})=\mathcal{HG}(\{\alpha_1,\ldots,\alpha_i\})$ hold for all $i<j$. We have now $\beta_j=k_{j,+}e_jk_{j,-1}$, where $k_{j,\pm}$ are some paths in $\Gamma_{j+1}\setminus\{e_j\}=\Gamma_j$. Furthermore, we have $\alpha_j=t_{j,+}e_jt_{j,-}$ with $t_{j,\pm}\subseteq T\subseteq\Gamma_j$. Thus $\beta_j\sim k_{j,+}t_{j,+}^{-1}\alpha_jt_{j,-}^{-1}k_{j,-}$. Since $k_{j,+}t_{j,+}^{-1}$ and $t_{j,-}^{-1}k_{j,-}$ are loops in Γ_j, we have $[k_{j,+}t_{j,+}^{-1}],[t_{j,-}^{-1}k_{j,-}]\in\mathcal{HG}(\Gamma_j)=\mathcal{HG}(\{\alpha_1,\ldots,\alpha_{j-1}\})=\mathcal{HG}(\{\beta_1,\ldots,\beta_{j-1}\})$. Due to $\alpha_j\sim t_{j,+}k_{j,+}^{-1}\beta_jk_{j,-}^{-1}t_{j,-}\in\mathcal{HG}(\{\beta_1,\ldots,\beta_j\})$ we have $\mathcal{HG}(\{\alpha_1,\ldots,\alpha_{j-1}\}\cup\{\alpha_j\})\subseteq\mathcal{HG}(\{\beta_1,\ldots,\beta_j\})$. Since β_j is a loop in Γ_{j+1}, we get immediately the $\supseteq$ relation, i.e., $\mathcal{HG}(\Gamma_{j+1})=\mathcal{HG}(\{\alpha_1,\ldots,\alpha_j\})=\mathcal{HG}(\{\beta_1,\ldots,\beta_j\})$. Thus $\{\beta_1,\ldots,\beta_j\}$ is complete in Γ_{j+1}.

The induction yields also $\mathcal{HG}(\boldsymbol{\beta})=\mathcal{HG}(\boldsymbol{\alpha})=\mathcal{HG}(\Gamma_{n+1})$, i.e., $\boldsymbol{\beta}$ is complete in $\Gamma_{n+1}=\Gamma$. We allow now the β_i to have retracing intervals. Denote by β_i' the loop that remains after canceling all these intervals in β_i. Obviously, β_i' lies in the same hoop class as β_i, i.e., $\mathcal{HG}(\boldsymbol{\beta})=\mathcal{HG}(\boldsymbol{\beta}')$. Thus, since we have already proven the proposition for the retracing-free $\boldsymbol{\beta}'$, we get immediately the claim for arbitrary $\boldsymbol{\beta}$. **qed**

IV. FLAG WORLDS

This section provides some facts about the hoop group of a graph (''lattice'') Γ in the two-dimensional manifold $M=\mathbb{R}^2$. For this we can specialize the facts of Sec. III E to the case of planar graphs (see, e.g., Ref. 11). These have a crucial advantage: one can define domains enclosed by the graph edges. The set of all these domains induces a basis of the corresponding hoop group $\mathcal{HG}(\Gamma)$. Finally, we will investigate the behavior of that set under refinement of the graph Γ generalizing the results of TA^+.

A. Planar graphs

This section collects some basic and simple facts about planar graphs and is intended to clarify the notations. We call a graph *X planar* iff there exists a homomorphism $\iota:X\rightarrow\Gamma\subseteq\mathbb{R}^2$. We identify X and Γ in the sequel. Furthermore, in the following any graph is supposed to be planar, finite, and connected.

Any graph is the complement of a disjoint union of domains. Exactly one of them if unbounded—the so-called exterior domain G_{ext}. The set of the remaining domains, the so-called interior domains, is denoted by $L_{\mathrm{int}}(\Gamma)$ and we set $L(\Gamma):=L_{\mathrm{int}}(\Gamma)\cup\{G_{\mathrm{ext}}\}$. We say that a domain G is contained in Γ iff its boundary ∂G is in Γ and $G\cap G_{\mathrm{ext}}=\emptyset$.

One easily proves Euler's polyhedron formula $\epsilon-\kappa+\lambda=2$, where ϵ, κ, and λ are the numbers of vertices, edges, and domains, respectively, of the graph. Since $\lambda-1=1-(\epsilon-\kappa)=1-\chi(\Gamma)$, we have using Proposition 3.5

Lemma 4.1: The number of interior domains of a graph Γ is equal to the rank of $\pi(\Gamma,m)$. We are now interested in the behavior of $L(\Gamma')$ under refinement of Γ'. Clearly, if we refine a graph Γ' to a graph Γ, then any domain of Γ' is refined into a certain set of domains in Γ (see, e.g., Fig. 2). We have in detail the simple

Proposition 4.2: Let $\Gamma'\leq\Gamma$. Then the following holds:

(1) For any $G\in L(\Gamma),G'\in L(\Gamma')$ we have $G\cap G'\neq\emptyset\Rightarrow G\subseteq G'$. Especially, two interior domains of one and the same graph are disjoint or equal.
(2) For any $G\in L(\Gamma)$ there exists exactly one $G'\in L(\Gamma')$ with $G\cap G'\neq\emptyset$.
(3) For any $G'\in L(\Gamma')$ there exists exactly one $L_{G'}\subseteq L(\Gamma)$, such that $G\cap G'\neq\emptyset\Leftrightarrow G\in L_{G'}$ and $\bigcup_{G\in L_{G'}}G\supseteq G'$.
(4) Now let G' be any domain in Γ, not necessarily an interior domain. There is exactly one set $L_{G'}(\Gamma)\subseteq L_{\mathrm{int}}(\Gamma)$, such that for all interior domains G holds: $G\in L_{G'}(\Gamma)\Leftrightarrow G\cap G'\neq\emptyset$ and $\bigcup_{G\in L_{G'}}G\supseteq G'$.

We call $L(\Gamma)$ a refinement of $L(\Gamma')$ (and analogously for L_{int}) iff Γ is a refinement if Γ'.

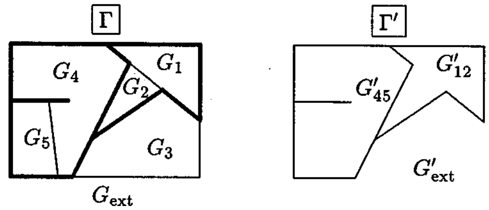

FIG. 2. Example for the decomposition of domains.

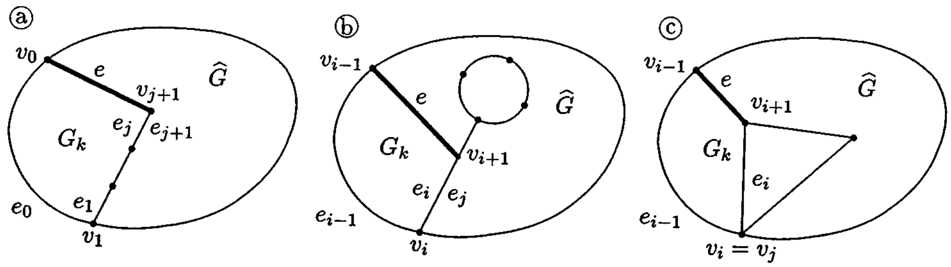

FIG. 3. Canceling (a) retracings, (b) repetition of edges, (c) repetition of vertices.

Definition 4.1: A domain $G \subseteq \mathbb{R}^2$ is called *simple* iff it is the interior of a Jordan curve. A graph Γ is called *simple* iff each of its interior domains is simple.

Finally we need

Proposition 4.3: Any ordinary graph Γ is subgraph of a simple, ordinary graph Γ' whose exterior domain coincides with that of Γ.

The proof is quite easy. First one eliminates the retracing, then the repetitions of edges, and finally the repetitions of vertices by inserting appropriate edges as demonstrated in Fig. 3.

B. Boundary loops and flags

We start with a simple

Lemma 4.4: For any simple domain $G \subseteq \Gamma$ and any $\tilde{m} \in \Gamma_0 \cap \partial G$ there is exactly one genuine loop $\alpha_{G,\tilde{m}}$ in Γ with base point $\tilde{m}$, such that (we recall that we do not distinguish between loops and hoops in the sequel—the equal sign means equality of loops and the symbol $\sim$ means equality of hoops)

(1) $\alpha_{G,\tilde{m}} = \partial G$ and
(2) $\alpha_{G,\tilde{m}}$ traverses the domain G counterclockwise.

By contrast, any such loop determines exactly one simple $G \subseteq \Gamma$. We call $\alpha_{G,\tilde{m}}$ the *boundary loop* of G with base point $\tilde{m}$.

Analogously, for any $G \subseteq \Gamma$ and any $\tilde{m} \in \Gamma_0 \cap \partial G$ there exists a loop $\alpha_{G,\tilde{m}}$ in Γ with base point $\tilde{m}$ and the properties above.

Now we are interested in loops with base point m, that traverse only one domain G in Γ. This is provided by

Definition 4.2: Flag. Let G be a simple domain in a graph Γ. We call a loop $f_{G,m,\tilde{m}}$ *flag* with base point m, flag point $\tilde{m}$, and domain G iff

(1) $f = \rho_{m\tilde{m}} \alpha_{G,\tilde{m}} \rho_{m\tilde{m}}^{-1}$,
(2) $\alpha_{G,\tilde{m}}$ is a boundary loop of G with base point $\tilde{m}$ and
(3) $\rho_{m\tilde{m}}$ is a path from m to $\tilde{m}$ in Γ;
(4) there is a $v \in \partial G$, such that
 (a) $\rho_{m\tilde{m}} = \rho_{mv} \rho_{v\tilde{m}}$,
 (b) $\rho_{mv} \cap \partial G = \{v\}$,
 (c) ρ_{mv} traces neither an edge nor a vertex twice and
 (d) $\rho_{v\tilde{m}} \subseteq \partial G$ holds.

Then $\rho_{m\tilde{m}}$ is called *flagpole*. We call $f_{G,m,\tilde{m}}$ *minimal* iff $v = \tilde{m}$.

Since Γ is connected, we get from Lemma 4.4

Lemma 4.5: For any triple $\{G, m, \tilde{m}\}$ with the above properties there exists a corresponding flag $f_{G,m,\tilde{m}}$ (Fig. 4).

Remark: (1) To any simple domain G and any $m\in\Gamma_0$ there exists a minimal flag. For this choose a maximal tree T and an $m'\in\partial G\cap\Gamma_0$. Furthermore, choose the shortest path ρ from m to m' along T. Let $\tilde{m}$ be the nearest to m (with respect to the up to there traced edges of ρ) point in $\partial G\cap\rho$ and $\rho_{m\tilde{m}}$ the corresponding initial path of ρ from m to $\tilde{m}$. Obviously, $f_{G,m,\tilde{m}} :=\rho_{m,\tilde{m}}\alpha_{G,\tilde{m}}\rho_{m\tilde{m}}^{-1}$ with the boundary loop $\alpha_{G,\tilde{m}}$ is a minimal flag for G.

(2) All flags beginning with the same ρ_{mv} are equal modulo holonomy equivalence, especially any flag is holonomically equivalent to a minimal flag.

(3) For $\mathbf{G}=\mathrm{U}(1)$ all flags to one and the same domain are holonomically equivalent.

Let $f_i=\rho_{m\tilde{m}_i}\alpha_{G,\tilde{m}_i}\rho_{m\tilde{m}_i}^{-1}$, $i=1,2$. We have

$$\begin{aligned} f_1=\rho_{m\tilde{m}_i}\alpha_{G,\tilde{m}_i}\rho_{m\tilde{m}_1}^{-1}&\sim\rho_{m\tilde{m}_1}\rho_{\tilde{m}_1\tilde{m}_2}\alpha_{G,\tilde{m}_2}\rho_{\tilde{m}_1\tilde{m}_2}^{-1}\rho_{m\tilde{m}_1}^{-1}\\ &\sim\rho_{m\tilde{m}_1}\rho_{\tilde{m}_1\tilde{m}_2}\rho_{m\tilde{m}_2}^{-1}\rho_{m\tilde{m}_2}\alpha_{G,\tilde{m}_2}\rho_{m\tilde{m}_2}^{-1}\rho_{m\tilde{m}_2}\rho_{\tilde{m}_1\tilde{m}_2}^{-1}\rho_{m\tilde{m}_1}^{-1}\\ &\sim\rho_{m\tilde{m}_1}\rho_{\tilde{m}_1\tilde{m}_2}\rho_{m\tilde{m}_2}^{-1}\rho_{m\tilde{m}_2}\rho_{\tilde{m}_1\tilde{m}_2}^{-1}\rho_{m\tilde{m}_1}^{-1}\rho_{m\tilde{m}_2}\alpha_{G,\tilde{m}_2}\rho_{m\tilde{m}_2}^{-1}\sim f_2. \end{aligned}$$

$\rho_{\tilde{m}_1\tilde{m}_2}$ is any path from $\tilde{m}_1$ to $\tilde{m}_2$ along ∂G. In the last but one step we used the commutativity of $\mathcal{HG}\supseteq\mathcal{HG}(\Gamma)$ induced by the commutativity of U(1).

(4) Two flags to disjoint domains are nonoverlapping.

C. Flag worlds: Definition and existence

In this section and Sec. IV D we only consider simple graphs, i.e., graphs with only simple interior domains, to avoid technical complications.

We are looking for a set $\boldsymbol{\beta}$ of hoops, such that any hoop in Γ can be expressed by a product of elements of $\boldsymbol{\beta}$, i.e., $\mathcal{HG}(\beta)=\mathcal{HG}(\Gamma)$ holds. Furthermore, we are interested in integrating cylindrical functions over $\mathcal{HG}(\boldsymbol{\beta})$. For this we need the moderate independence of $\boldsymbol{\beta}$, that means at least the weak independence. Due to Proposition 3.2 that is guaranteed only if the number of elements of $\boldsymbol{\beta}$ equals the number of generators of $\mathcal{HG}(\Gamma)$, i.e., equals the number of generators of the fundamental group $\pi(\Gamma,m)$. With this in mind one could choose $\boldsymbol{\beta}$ to be a system of generators as in Proposition 3.4. But, because of our regularization we need loops enclosing an area being as tiny as possible, i.e., enclosing only one interior domain. For this the above defined flags are well suited. We already know that the number of interior domains of Γ equals the rank of the fundamental group (cf. Euler's polyhedron formula in Sec. IV A). Thus the following definition is obvious.

Definition 4.3: Flag World. A set $\mathcal{F}$ of flags is called *flag world* to the simple graph Γ (with base point m) iff $\mathcal{F}=\{f_G|G\in L_{\mathrm{int}}(\Gamma)\}$, where f_G is any flag to the domain G and to the base point m. $\mathcal{F}$ is called *complete* iff $\mathcal{HG}(\mathcal{F})=\mathcal{HG}(\Gamma)$.

Using Proposition 3.2 we have immediately

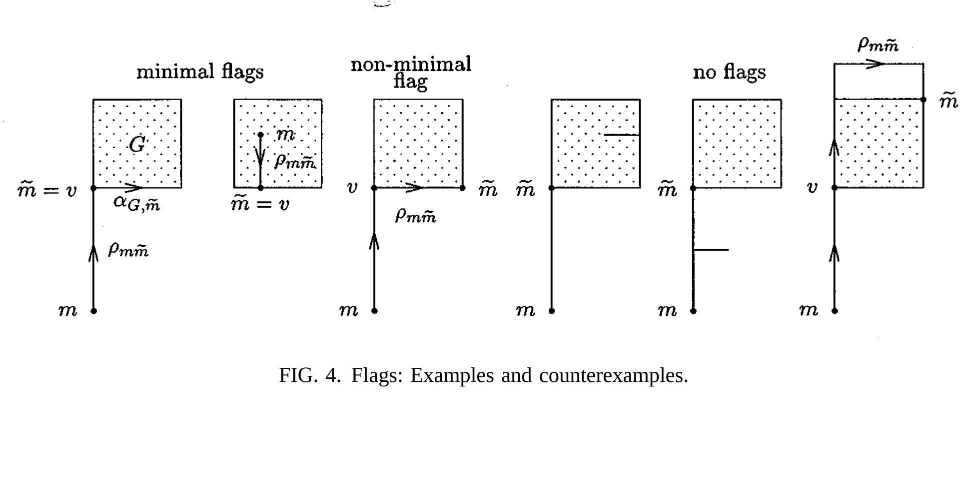

FIG. 4. Flags: Examples and counterexamples.

Corollary 4.6: The flags of a complete flag world are weakly independent.

Now we are interested in moderately independent flag worlds because they are necessary for the integration calculus and because of

Proposition 4.7: Let $\mathcal{F}$ be a moderately independent flag world in a simple graph Γ. Then $\mathcal{F}$ is complete.

Proof: Γ is a finite, connected graph and $\mathcal{F}$ is a moderately independent set of loops in Γ, whose cardinality is equal to the rank of $\pi(\Gamma,m)$ due to Lemma 4.1. Proposition 3.10 finishes the proof. **qed**

We can construct naturally a flag world to any tree as follows.

Definition 4.4: Let T be a maximal tree in a simple graph Γ.

$\mathcal{F}$ is called T-flag world for Γ iff the following holds for all flags $f\in\mathcal{F}$:

(1) f is a minimal flag.
(2) The flagpole of f is a path in T.

Proposition 4.8: Let T be a maximal tree in a simple graph Γ.

(1) There is a T-flag world for Γ.
(2) Any T-flag world for Γ is moderately independent.

From this we get the crucial

Corollary 4.9: For any simple graph Γ there exists a moderately independent, i.e., also complete flag world.

Corollary 4.10: Any loop in Γ is holonomically equivalent to a product of mutually nonoverlapping loops.

Proof (Proposition 4.8):

(i) First, let Γ be a tree, i.e., $\Gamma=T$. Then there is no interior domain and therefore no flag, too. We have $\mathcal{F}=\emptyset$ and $\mathcal{HG}(\mathcal{F})=\{1\}=\mathcal{HG}(\Gamma)$.

(ii) Now, Γ is not a tree. Let T be a maximal tree in Γ and $E:=\{e_\lambda\}$ the corresponding set of edges of Γ not contained in T. Now we can construct Γ from T inserting successively edges e_λ. The intermediate graphs are denoted by Γ_λ. This allows us to use induction on the number of interior domains increased exactly by 1 in each step. We can insert these edges, such that any new edge e_λ lies on the boundary of the corresponding graph Γ_λ. (Suppose there is a tree T' with $\partial\Gamma\subseteq T'$. Then $\partial\Gamma$ is a tree itself and $\partial\Gamma$ has no interior domain. Consequently, Γ has no interior domain, i.e., Γ is a tree. Thus, there is no tree T' with $\partial\Gamma\subseteq T'$ and so for any tree T in Γ there is an edge $e_\lambda\subseteq\partial\Gamma$ that is not contained in T.) Thus the interior domains of the intermediate graphs are simple due to $L_{\mathrm{int}}(\Gamma_\lambda)\subseteq L_{\mathrm{int}}(\Gamma)$. Obviously, any Γ_λ is finite, planar, and connected.

(iii) Suppose the proposition holds for any graph with $k-1\geqslant 0$ interior domains. Now, Γ has k interior domains, T and E are chosen as above and $e\in E$ is an edge in $\partial\Gamma$. We set $\Gamma':=\Gamma\backslash\{e\}$ and $E':=E\backslash\{e\}$. By inserting e in Γ' we get a new (simple) interior domain G, i.e., $L_{\mathrm{int}}(\Gamma)=L_{\mathrm{int}}(\Gamma')\cup\{G\}$. Obviously, T is also a maximal tree in Γ' and E' is the set of all edges of Γ' not contained in T. Γ' has exactly $k-1$ interior domains and we have by induction:

(1) There exists a T-flag world for Γ'.
(2) Any T-flag world for Γ' is moderately independent.

(1) Existence of a T-flag world for Γ

We construct a flag for G. Since any vertex of Γ is contained in T, there is a path in T from m to a vertex of ∂G. We choose from among these paths a path ρ which is minimal with respect to the number of traced edges. The terminal vertex of ρ is denoted by $\tilde{m}$, $\tilde{m}\in\partial G$. Due to Lemma 4.4 we choose a boundary loop α of G with base point $\tilde{m}$. $f:=\rho\alpha\rho^{-1}$ is now a minimal flag for G and $\mathcal{F}:=\mathcal{F}'\cup\{f\}$ is a T-flag world for Γ.

(2) Moderate independence of any T-flag world for Γ

Γ', E', and G are still chosen as above. Set $\mathcal{F}':=\mathcal{F}\backslash\{f_G\}$, where $f_G\in\mathcal{F}$, $f_G=\rho\alpha\rho^{-1}$, is the flag

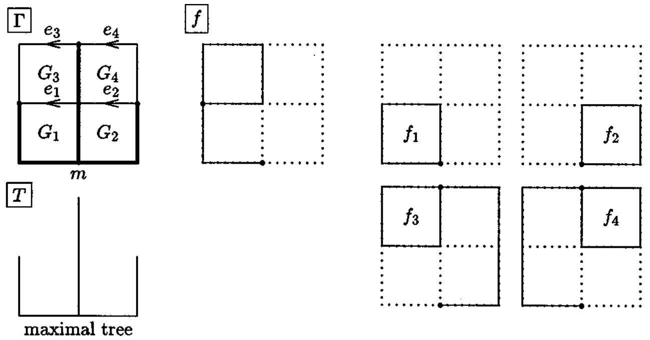

FIG. 5. Example of a noncomplete flag world.

for G with flagpole $\rho \subseteq T$. Obviously, $\mathcal{F}'$ is a T-flag world for Γ', and therefore moderately independent by induction.

Since f_G is minimal, e is traced exactly once by f_G, and because $\mathcal{F}'$ is a flag world in $\Gamma' = \Gamma \setminus e$, not any $f_i \in \mathcal{F}'$ traces e. Therefore e is now a free segment of f_G.

Finally, $\mathcal{F}'$ itself is moderately independent with the free segments e_i of the corresponding $f_i \in \mathcal{F}'$. Thus, $\mathcal{F} = \mathcal{F}' \cup \{f_G\}$ is moderately independent with the free segments $\{e_1,\ldots,e_{k-1},e\}$. **qed**

Remark: For $\mathbf{G} = \mathrm{U}(1)$ even any flag world $\mathcal{F}$ is complete.

To prove this choose any complete flag world $\mathcal{F}'$ for Γ. Since [for $\mathbf{G} = \mathrm{U}(1)$] all flags belonging to one and the same domain are equal up to holonomy equivalence, we have $\mathcal{HG}(\mathcal{F}) = \mathcal{HG}(\mathcal{F})' = \mathcal{HG}(\Gamma)$, i.e., $\mathcal{F}$ is complete.

In other words, for U(1) all flag worlds to one and the same graph Γ are equal modulo holonomy equivalence.

The completeness of a flag world is not at all trivial for the SU(N) because of

Proposition 4.11: Let $\mathbf{G} = \mathrm{SU}(N)$. Then there exists a simple graph Γ, such that a noncomplete (and so also not moderately independent) flag world exists to Γ.

Proof: It is sufficient to give an example. Due to $\mathbf{G} = \mathrm{SU}(N)$ holonomy equivalence equals homotopy equivalence and we will indentify hoops and the corresponding elements of the fundamental group $\pi(\Gamma,m)$. It is sufficient to construct a flag world $\mathcal{F}$, such that there is a loop $f \in \pi(\Gamma,m) = \mathcal{HG}(\Gamma)$ not contained in the subgroup $\mathcal{HG}(\mathcal{F})$ of the fundamental group generated by $\mathcal{F}$.

Let Γ be the graph in Fig. 5 with the flag world $\mathcal{F} := \{f_1,f_2,f_3,f_4\}$, the maximal tree T, and the corresponding free edges e_1,e_2,e_3,e_4. We constuct from T and e_i the free generators $\alpha_1,\ldots,\alpha_4$ of $\pi(\Gamma,m)$ as in Proposition 3.4. We will prove, that $\mathcal{F}$ is not complete showing that $f \notin \mathcal{HG}(\mathcal{F})$, where f is the loop defined in Fig. 5.

A simple calculation shows:

$$f \sim \alpha_1^{-1}\alpha_3,$$

$$f_1 \sim \alpha_1,$$

$$f_2 \sim \alpha_2,$$

$$f_3 \sim \alpha_4\alpha_3\alpha_1^{-1}\alpha_4^{-1},$$

$$f_4 \sim \alpha_3^{-1}\alpha_2^{-1}\alpha_4\alpha_3.$$

Suppose $f \in \mathcal{HG}(\mathcal{F})$, i.e., $f \sim \Pi_{j=1}^{J} f_{i_j}^{\eta_j} \sim \Pi_k \alpha_{\lambda_k}^{\epsilon_k}$ with $\eta_j \in \mathbb{Z}$ and $\epsilon_k \in \{-1,+1\}$. Choose this decomposition, such that the number J of used factors $f_{i_j}^{\eta_j}$ is minimal. Due to the freeness of $\mathcal{HG}(\Gamma) = \pi(\Gamma,m)$ there must exist a j' with $i_{j'} = 3$ and $\eta_{j'} \geq +1$, i.e.,

$$\alpha_1^{-1}\alpha_3 \sim f \sim \prod_{j=1}^{j'-1} f_{i_j}^{\eta_j} f_3^{\eta_{j'}} \prod_{j=j'+1}^{J} f_{i_j}^{\eta_j} \sim \prod_k \alpha_{\lambda_k}^{\epsilon_k} \alpha_4 \prod^{\eta_{j'}} (\alpha_3 \alpha_1^{-1}) \alpha_4^{-1} \prod_{k'} \alpha_{\lambda_{k'}}^{\epsilon_{k'}}.$$

In the last step f_j^{η} has been replaced by the corresponding reduced representation in the α_λ (see above), e.g., f_3^{η} by $\alpha_4(\alpha_3\alpha_1^{-1})^{\eta}\alpha_4^{-1}$ [i.e., not by $(\alpha_4\alpha_3\alpha_1^{-1}\alpha_4^{-1})^{\eta}$, since here (for $|\eta|>1$) the $\alpha_4\alpha_4^{-1}$ terms are not reduced].

The right-hand decomposition of f in α_λ is (with respect to the number of used factors) longer than the left-hand one. Again by the freeness of $\mathcal{HG}(\Gamma)$ there must exist in the right-hand decomposition of f in α_λ a k with $\alpha_{\lambda_k}^{\epsilon_k} = \alpha_{\lambda_{k+1}}^{-\epsilon_{k+1}}$. This case does not occur in the decompositions of the f_i in α_λ above, thus this must occur during the multiplication $f_{i_j}^{\eta_j} f_{i_{j+1}}^{\eta_{j+1}}$ of two flags. From the decompositions above we see that such a collision of α_λ is only possible, if $i_j = i_{j+1}$. This is a contradiction to the minimality of the decomposition of f into a product of flags $f_i^{\eta} \in \mathcal{F}$.
Thus, $f \notin \mathcal{HG}(\mathcal{F})$, and $\mathcal{F}$ is not complete. **qed**

Remark: (1) Up to now, we do not know, whether noncomplete flag worlds can be constructed for graphs with less than four interior domains.

(2) Simultaneously, we have constructed an example for the fact that from $\mathcal{HG}(\boldsymbol{\beta}) \subseteq \mathcal{HG}(\boldsymbol{\alpha})$ and the equality of the cardinalities of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ not generally follows, that $\mathcal{HG}(\boldsymbol{\beta}) = \mathcal{HG}(\boldsymbol{\alpha})$.

But, obviously, $\mathcal{HG}(\mathcal{F})$ is freely generated by $\{f_1, f_2, f_3, f_4\}$. Thus, we have constructed a genuine (free) subgroup of $\mathcal{HG}(\Gamma)$ having the same rank as $\mathcal{HG}(\Gamma)$.

D. Refinement of flag worlds

Now we want to investigate the behavior of flag worlds under refinement of the underlying graph. We need the following:

Lemma 4.12: Let Γ be a simple graph and G a simple domain in $\Gamma(m \notin G)$ with corresponding refinement $\{G_i | i \in I\} \subseteq L_{\text{int}}(\Gamma)$. Let f be a minimal flag belonging to G with base point m. Furthermore, e is an arbitrary edge of Γ on ∂G. Then, there exist minimal flags f_i with base point m, such that:

(1) f_i is a flag to domain G_i for all $i \in I$;
(2) f is holonomically equivalent to the product of all f_i in a certain order;
(3) $\{f_i\}$ is moderately independent and any of the free segments lies in $\text{int}\, G \cup e$.

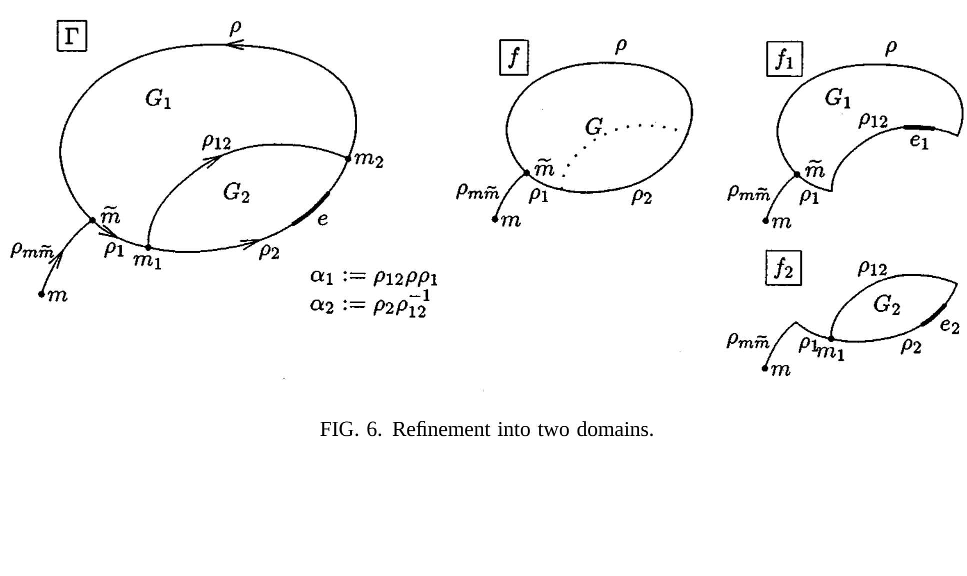

FIG. 6. Refinement into two domains.

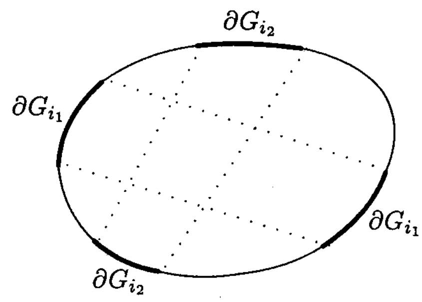

FIG. 7. Existence of a G_k with connected $\partial G_k \cap \partial G$.

Proof: Induction on the cardinality of I. [I is finite, since Γ is finite and thus $L_{\text{int}}(\Gamma)$ is finite.]

(1) $I=1$ is trivial, i.e., $G=G_i$ is an interior domain itself.

(2) First, we consider the case $I=\{1,2\}$.

We consider the case that e and $\tilde{m}$ do not lie on the boundary of one and the same interior domain. Topologically, we have the situation of Fig. 6; if necessary, one has to exchange the domains 1 and 2. Let $\rho_{m\tilde{m}}$ be the flagpoles of f from m to $\tilde{m},\rho,\rho_1,\rho_2,\rho_{12}$ as in Fig. 6 and α_i the corresponding boundary loop for G_i with base point m_1.

We set $f_i := \rho_{m\tilde{m}}\rho_1\alpha_i\rho_1^{-1}\rho_{m\tilde{m}}^{-1}$ for $i=1,2$, after canceling possible retracings, i.e., we consider f_i to be minimal.

(a) Obviously, f_i is a flag for G_i.

(b) We have $f \sim f_2 f_1$.

(c) Choose an edge $e_1 \subseteq \partial G_1 \cap \partial G_2$, i.e., $e_1 \subseteq \rho_{12}$, and set $e_2 := e$. Then $\{f_1,f_2\}$ is obviously moderately independent with the free segments $e_1,e_2 \subseteq \text{int}\, G \cup e$. In the case that e and $\tilde{m}$ lie on the boundary of one and the same domain, one has to exchange, if necessary, ρ_1 and ρ in the construction above, such that $e \cap \rho_1 = \emptyset$. The rest of the proof is completely analogous.

(3) Suppose the lemma is proven for refinements by $k-1 \geqslant 2$ domains and now let $\{G_i\}$ be a refinement of G by $k \geqslant 3$ domains.

(a) Choose any $i \in I$, such that $\overline{G_i} \cap \partial G$ contains at least one edge of Γ and the domain $\tilde{G}$ built from the remaining G_j is again simple (W.l.o.g. we set $i=k$ and j runs in the following from 1 to $k-1$.) More precisely: ∂G, ∂G_k, and $\rho_{m\tilde{m}}$ span a finite and for $\overline{G_k} \cap \partial G \neq \emptyset$ again connected graph. We demand that the set of the interior domains in this graph is equal to $\{\tilde{G},G_k\}$ and that $\tilde{G}$ is simple.

It remains the question, whether such a G_k exists. The first condition is trivial. To prove the second one it is sufficient to choose a domain G_k, such that $\partial G_k \cap \partial G$ is connected.

To see this let α be a boundary loop of G. One gets an $\tilde{\alpha}$ from this, if one replaces the subpath α_k of α belonging to ∂G_k by the path $\tilde{\alpha}_k$ corresponding to the boundary $\partial G_k \setminus \partial G$. Obviously, $\tilde{\alpha}$ is a path in Γ. $\tilde{\alpha}$ has neither repetitions of vertices nor of edges, because neither α nor α_k have the like and because $\tilde{\alpha}_k$ touches α only in its initial and terminal vertex (these are distinct). Otherwise, we would have a contradiction to the connectivity of $\partial G_k \cap \partial G$. Therefore $\tilde{\alpha}$ is a Jordan path, i.e., a boundary of exactly one simple interior domain $\tilde{G}$.

It remains now to ask for the existence of such a domain. Suppose not any $\partial G_i \cap \partial G$ is connected. Then there would exist a pair of indices (i_1,i_2), such that we have the situation in Fig. 7. Obviously, this is a contradiction to the connectivity of G_{i_1} and G_{i_2}.

Thus, there is a refinement of G into two simple domains $\{\tilde{G},G_k\}$, such that $\tilde{G}$ itself has a refinement into $\{G_j\}$ in Γ.

(b) Due to point (2) there are minimal flags $\tilde{f},f_k$ for $\tilde{G}$ and G_k, respectively, such that

(i) $f \sim \tilde{f} f_k$ (or $f_k \tilde{f}$);

(ii) $\{\tilde{f}, f_k\}$ or $\{f_k, \tilde{f}\}$ is moderately independent, where the free segments $\tilde{e}$ and e_k lie in $\operatorname{int} G \cup e$.

(c) Let $\tilde{e}$ be the free segment of $\tilde{f}$. It is obviously an edge in $\Gamma \cap \partial \tilde{G}$. Thus, by induction there are minimal flags f_j, such that:

(i) $\tilde{f}$ is a product of the f_j in a certain order;

(ii) $\{f_j\}$ is moderately independent, where any of the free segments e_j lies in $\operatorname{int} \tilde{G} \cup \tilde{e}$.

(d) Thus, f can be represented as a hoop product of the f_i in a certain order.

(e) The proof of the moderate independence of $\{f_j\} \cup \{f_k\}$ goes completely analogously to the case of two domains.

(i) Let $\{\tilde{f}, f_k\}$ be moderately independent. Then e_k lies in $(\operatorname{int} G \cup e) \setminus \tilde{f}$, otherwise e_k would already be traced by $\tilde{f}$. Thus, $e_k = e$, since $(\operatorname{int} G \setminus \tilde{f}) \cap f_k = \emptyset$, and so $\tilde{e} \subseteq \operatorname{int} G$. Due to $e_j \subseteq \operatorname{int} \tilde{G} \cup \tilde{e} \subseteq \operatorname{int} G$ we have e_j, $e_k \subseteq \operatorname{int} G \cup e$.
$\{f_1, \ldots, f_{k-1}, f_k\}$ is moderately independent because $e_1, \ldots, e_{k-1}$ are free segments of $f_1, \ldots, f_{k-1}$, and e_k is free, because $e_k \cap f_j \subseteq e_k \cap (\tilde{f} \cup \operatorname{int} \tilde{G}) = (e_k \cap \tilde{f}) \cup (e_k \cap \operatorname{int} \tilde{G}) = \theta$. The second intersection vanishes obviously and the first one does because $\{\tilde{f}, f_k\}$ are moderately independent.

(ii) Let $\{f_k, \tilde{f}\}$ be moderately independent. The argumentation is analogous to the other case, however, here $\{f_k, f_1, \ldots, f_{k-1}\}$ is moderately independent. **qed**

We have now

Proposition 4.13: Let Γ, Γ' be simple graphs, Γ' a refinement of Γ and $m \in \Gamma$. Then there exists for any moderately independent flag world $\mathcal{F}$ of Γ a moderately independent flag world $\mathcal{F}'$ of Γ', such that the following holds for all interior domains G_I of Γ: The flag $f_I \in \mathcal{F}$ to G_I is the hoop product of exactly these flags $f_{I,i_I} \in \mathcal{F}'$, that belong to the interior domains G_{I,i_I} with $G_{I,i_I} \subseteq G_I$, in a certain order.

Proof: Obviously, we have $m \notin G_I$ for all $G_I \in L_{\text{int}}(\Gamma)$ because $m \in \Gamma$. First, we define Γ'' to be the graph built from all interior domains of Γ' that are contained in the exterior domain of Γ and from all interior domains of Γ. Obviously, Γ'' is simple, $\Gamma \leqslant \Gamma'' \leqslant \Gamma'$ and the exterior domains of Γ'' and Γ' coincide. Now let $\mathcal{F} = \{f_I\} = \{f_1, \ldots, f_\Lambda\}$ be moderately independent with the free segments e_I. We can refine $\mathcal{F}$ to a moderately independent flag world $\mathcal{F}'' = \{f_1, \ldots, f_{\Lambda''}\} \supseteq \mathcal{F}$ of Γ'', where Λ'' is the number of interior domains of Γ'', analogous to the proof of Proposition 4.8. Next, we consider for any interior domain of Γ the corresponding refinement of G_I into the $G_{I,i_I} \in L_{\text{int}}(\Gamma')$. Due to Lemma 4.12 there exist minimal flags f_{I,i_I} with base point m, such that:

(1) f_{I,i_I} is a flag to the domain G_{I,i_I}.
(2) f_I is holonomically equivalent to the product of all f_{I,i_I} in a certain order.
(3) $\{f_{I,i_I}\}$ is moderately independent and any free segment e_{I,i_I} is contained in $\operatorname{int} G_I \cup e_I$. (W.l.o.g. e_I is an edge of Γ on ∂G_I.)

The flags f_I in $\mathcal{F}'' \setminus \mathcal{F}$, i.e., those flags that belong to the interior domain of Γ', but are contained in the exterior domain of Γ, are left untouched. We only set $\lambda_I := 1$ and $f_{I,i_I} := f_I$. Now (λ_I is the number of domains, that the G_I are refined into) $\mathcal{F}' = \{f_{1,1}, \ldots, f_{1,\lambda_1}, f_{2,1}, \ldots, f_{2,\lambda_2}, \ldots, f_{\Lambda'',1}, \ldots, f_{\Lambda'',\lambda_{\Lambda''}}\}$ is a moderately independent flag world of Γ' because:

(i) e_{I,i_I} is traced exactly once by f_{I,i_I} per constructionem and is not traced by any $f_{I,j}$ with $j < i_I$ due to the just stated point (3). $\{f_{I,j} | j \in [1, \lambda_I]\}$ is moderately independent with the free segments $e_{I,j}$ for a fixed I.
(ii) But, e_{I,i_I} is also not traced by $f_{J,j}$ with $J < I$:

$f_{J,j}$ traces only $f_J \cup \operatorname{int} G_J$ and we have $e_{I,i_I} \subseteq \operatorname{int} G_I \cup e_I$. Since the domains of Γ are disjoint, we have $\operatorname{int} G_J \cap \operatorname{int} G_I = \emptyset$, $f_J \cap \operatorname{int} G_I = \emptyset$ and $\operatorname{int} G_J \cap e_I = \emptyset$. Finally, we have $f_J \cap e_I = \emptyset$ since $\mathcal{F}'' = \{f_1, \ldots, f_{\Lambda''}\}$ itself is moderately independent. Thus, $f_{J,j} \cap e_{I,i_I} = \emptyset$.

Thus, e_{I,i_I} fits all conditions for a free segment. Since $\mathcal{F}'$ is obviously a flag world of Γ', we get the proof. **qed**

E. Conclusions

We collect the most important facts with regard to the applications in Sec. VI sometimes neglecting mathematical details. Any graph is finite, planar, connected, and nonempty.

Let there be given an arbitrary graph Γ.

(1) There is a refinement of Γ to an ordinary graph Γ'.

(2) Any graph can be naturally associated with a finite set of connected interior domains and an exterior domain (Sec. IV A). By a refinement of Γ this set is refined.

(3) A graph is called simple iff its interior domains are simple, i.e., are bounded by Jordan loops.

(4) Any ordinary graph Γ' is subgraph of a simple, ordinary graph Γ''. The exterior domains of both graphs are the same (Proposition 4.3).

(5) Any simple domain G in a graph can be naturally associated with a flag, i.e., a loop running from a base point m to ∂G, traversing G exactly once and running back to m (Definition 4.2).

(6) By choosing a flag to each interior domain one gets a flag world (Definition 4.3). It is called complete iff it spans the full hoop group of Γ.

(7) We are looking for moderately independent and complete flag worlds. The completeness ensures that any loop in Γ can be expressed by elements of a flag world. The moderate independence is necessary for the integration of cylindrical functions. Fortunately, the moderate independence implies the completeness (Proposition 4.7).

(8) One can naturally construct flag worlds to any simple graph. For this one chooses a maximal tree in this graph and then for any interior domain a flag consisting of a path along the tree, a boundary loop of the corresponding domain and the inverse initial path. Any such flag world is moderately independent (Proposition 4.8).

(9) There is a moderately independent flag world for any simple graph (Corollary 4.9). Thus, any hoop can be represented as a hoop product of mutually nonoverlapping loops.

(10) Under refinement of a simple domain G with a flag f one can choose flags f_i to the new domains G_i such that these generate all hoops ''in G'' and that f can be expressed as a hoop product of the f_i in a certain order (Lemma 4.12).

(11) In simple graphs Γ' any moderately independent flag world $\mathcal{F}$ of a simple subgraph Γ can be refined to a moderately independent flag world $\mathcal{F}'$ von Γ' such that any flag $f_G \in \mathcal{F}$ is a product of the flags $f_{G'} \in \mathcal{F}'$ to the interior domains $G' \subseteq G$ in a certain order (Proposition 4.13).

In Sec. VI we will see that especially the last point is crucial for the regularization of the Wilson loop functionals. We can now decompose the ''banner'' of a given flag in smaller ''banners.'' But all small ''banners'' have ''equal rights'' since $f_I \sim f_{I,1} \cdots f_{I,\lambda_I}$. That is why they give identical contributions if we integrate cylindrical functions in f_I.

V. INTEGRATION ON $\overline{\mathcal{A}/\mathcal{G}}$

In this section we slightly generalize the integration calculus on $\overline{\mathcal{A}/\mathcal{G}}$ which was in detail investigated by Ashtekar and Lewandowski.[2] Their key idea was to first define an equivalence relation on $\overline{\mathcal{A}/\mathcal{G}}$ which identifies two connections iff their holonomies on a certain finite set $\boldsymbol{\beta}$ of hoops are equal (up to conjugation), i.e., factorizing with respect to that relation they extracted the properties of a generalized connection on that finite set. But, if one knows these properties for all finite sets of hoops, one can reconstruct via $\overline{\mathcal{A}/\mathcal{G}} \sim \mathrm{Hom}(\mathcal{HG},\mathbf{G})/\mathrm{Ad}$ the generalized connection in $\overline{\mathcal{A}/\mathcal{G}}$. The main advantage of the factorization is the reduction of the infinite-dimensional problem to a finite-dimensional one, since $\overline{\mathcal{A}/\mathcal{G}}/\sim \cong \mathrm{Hom}(\mathcal{HG}(\beta),\mathbf{G})/\mathrm{Ad} \cong \mathbf{G}^{\#\boldsymbol{\beta}}/\mathrm{Ad}$. Comparing that situation with the case of infinite-dimensional topological vector spaces, AL defined first cylindrical functions as functions on $\overline{\mathcal{A}/\mathcal{G}}/\sim$ and second the integral of cylindrical functions $f = \pi_{\boldsymbol{\beta}}^* f_{\boldsymbol{\beta}}$ via

$\int_{\overline{\mathcal{A/G}}} f, d\mu = \int_{\overline{\mathcal{A/G}}/\sim} f_{\boldsymbol{\beta}} d\mu_{\boldsymbol{\beta}}$, where $d\mu_{\boldsymbol{\beta}}$ is a measure on $\overline{\mathcal{A/G}}/\sim \cong \mathbf{G}^{\#\boldsymbol{\beta}}/\mathrm{Ad}$. The main problem is to guarantee that this integral is well defined. AL could prove this for the choice that $d\mu_{\boldsymbol{\beta}}$ is the Haar measure on $\mathbf{G}^{\#\boldsymbol{\beta}}/\mathrm{Ad}$, and if only *strongly* independent $\boldsymbol{\beta}$ are allowed for calculating the integral above. The use of merely weakly independent $\boldsymbol{\beta}$ leads to contradicitons. Our task is now to prove that the use of *moderately* independent $\boldsymbol{\beta}$ keeps instead the definition valid. This point is crucial for the calculation of the Wilson loop expectation values using the not strongly, but moderately independent flag worlds.

A. Equivalence of connections

We recall[2] the following.

Definition 5.1: Equivalence of Connections. Let $\mathcal{HG}(\boldsymbol{\beta}) \subseteq \mathcal{HG}$ be a finitely generated subgroup of the hoop group $\mathcal{HG}$ with weakly independent $\boldsymbol{\beta}$. Two (generalized) connections $\overline{A}_1$ and $\overline{A}_2$ are called *equivalent with respect to* $\mathcal{HG}(\boldsymbol{\beta})$ iff

$$h_{\overline{A}_1}(\gamma) = g^{-1} h_{\overline{A}_2}(\gamma) g, \quad \forall \gamma \in \mathcal{HG}(\boldsymbol{\beta})$$

with a fixed (hoop independent) $g \in \mathbf{G}$.
Furthermore, let $\pi_{\boldsymbol{\beta}}: \overline{\mathcal{A/G}} \to \overline{\mathcal{A/G}}/\sim$ be the corresponding canonical projection.
Using the bijection $\overline{\mathcal{A/G}} \leftrightarrow \mathrm{Hom}(\mathcal{HG}, \mathbf{G})/\mathrm{Ad}$ Ashtekar and Lewandowski[2] could easily analyze the structure of $\overline{\mathcal{A/G}}/\sim$.

Lemma 5.1: (1) There is a bijection $\overline{\mathcal{A/G}}/\sim \to \mathrm{Hom}(\mathcal{HG}(\boldsymbol{\beta}), \mathbf{G})/\mathrm{Ad}$. That means, two generalized connections are equivalent if and only if they coincide mod Ad on $\mathcal{HG}(\boldsymbol{\beta})$.

(2) Any choice of n weakly independent generators $\beta_i \in \mathcal{HG}(\boldsymbol{\beta})$ yields a bijection $\phi_{\boldsymbol{\beta}}: \overline{\mathcal{A/G}}/\sim \to \mathbf{G}^n/\mathrm{Ad}$.

(3) Given $\mathcal{HG}(\boldsymbol{\beta}) \subseteq \mathcal{HG}$ the topology on $\overline{\mathcal{A/G}}/\sim$ induced by the last point is independent of the choice of generators.
Furthermore, we have[2]

Corollary 5.2: Let $\mathcal{HG}(\boldsymbol{\beta})$ be a finitely generated subgroup of the hoop group and $\sim$ the induced equivalence relation on $\overline{\mathcal{A/G}}$. Then any equivalence class $[\overline{A}] \in \overline{\mathcal{A/G}}/\sim$ contains a regular connection.

B. Cylindrical functions

In the following we set $\mathcal{B}_{\boldsymbol{\beta}} := \mathrm{Hom}(\mathcal{HG}(\boldsymbol{\beta}), \mathbf{G})/\mathrm{Ad} \cong \mathbf{G}^n/\mathrm{Ad}$ with $\boldsymbol{\beta}$ a weakly independent set of n hoops. Furthermore we usually do not distinguish between a function $f \in \overline{\mathcal{HA}}$ and its Gelfand transform $\tilde{f}: \overline{\mathcal{A/G}} \to \mathbb{C}$.

We now provide a slightly modified version of the Ashtekar–Lewandowski definition of a cylindrical function.

Definition 5.2: Cylindrical Function. $f: \overline{\mathcal{A/G}} \to \mathbb{C}$ is called *cylindrical function* iff there is a finite set $\boldsymbol{\beta}$ of weakly independent hoops and a continuous $f_{\boldsymbol{\beta}}: \mathcal{B}_{\boldsymbol{\beta}} \to \mathbb{C}$, such that $f = \pi^*_{\boldsymbol{\beta}} f_{\boldsymbol{\beta}}$. Iff f can be obtained that way for a given $\boldsymbol{\beta}$, f is called *cylindrical* with respect to $\boldsymbol{\beta}$.

The set of all cylindrical functions is denoted by $\mathcal{C}$.
It is very simple to verify

Lemma 5.3: Let f be cylindrical with respect to $\boldsymbol{\beta}$. Then f is cylindrical with respect to $\boldsymbol{\alpha}$, if the following holds:

(1) $\boldsymbol{\alpha}$ is weakly independent.
(2) $\mathcal{HG}(\boldsymbol{\alpha}) \supseteq \mathcal{HG}(\boldsymbol{\beta})$.

Remark: In contrast to Ref. 2 we define cylindrical functions not only on strongly independent, but also on weakly independent $\boldsymbol{\beta}$. For the present the set of cylindrical functions seems to be enlarged. But, it is easy to see, that given an $f \in \mathcal{C}$ there is a set $\boldsymbol{\alpha}$ of strongly independent loops, such that f is cylindrical with respect to $\mathcal{B}_{\boldsymbol{\alpha}}$.

Let $f\in\mathcal{C}$, i.e., there is a finite set $\boldsymbol{\beta}$ of weakly independent hoops with respect to that f is cylindrical. Following Lemma 3.8 there is a set $\boldsymbol{\alpha}$ of strongly independent loops, such that $\mathcal{HG}(\boldsymbol{\beta})\subseteq\mathcal{HG}(\boldsymbol{\alpha})$. Due to the just proven lemma f is cylindrical with respect to the strongly independent set $\boldsymbol{\alpha}$. Thus, our definition is equivalent to that one in Ref. 2.
Finally, we quote[2]

Proposition 5.4: $\mathcal{C}$ is a normed $*$-algebra and $\overline{\mathcal{C}}$ is isomorphic to $\overline{\mathcal{HA}}$.

C. The induced Haar measure on $\overline{\mathcal{A}/\mathcal{G}}$

Definition 5.3: Let be $f\in\mathcal{C}$ and $\boldsymbol{\beta}\subseteq\mathcal{HG}$ be a moderately independent set of n hoops, such that f is cylindrical with respect to $\boldsymbol{\beta}$, i.e., $f=\pi^*_{\boldsymbol{\beta}}f_{\boldsymbol{\beta}}$ with a continuous function $f_{\boldsymbol{\beta}}:\boldsymbol{\mathcal{B}}_\beta\to\mathbb{C}$. Furthermore, $d\mu_{\boldsymbol{\beta}}$ is an arbitrary measure on $\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}}$. Then we define $\int_{\overline{\mathcal{A}/\mathcal{G}}}f\,d\mu:=\int_{\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}}}f_{\boldsymbol{\beta}}\,d\mu_{\boldsymbol{\beta}}$.
We have to guarantee that the measures on the distinct $\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}}$ are compatible in order to make the integral in the definition above well-defined.

Ashtekar and Lewandowski suggested to choose the Haar measure on each $\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}}$, $\boldsymbol{\beta}$ strongly independent, induced from $\mathbf{G}^n/\mathrm{Ad}$ with n the cardinality of $\boldsymbol{\beta}$. Indeed, they could prove that the definition above provides a well-defined integral on $\overline{\mathcal{A}/\mathcal{G}}$. We are only left with the proof that the integral is still well-defined if we allow $\boldsymbol{\beta}$ to be merely moderately independent instead of strongly independent. Fortunately, for this we can reuse the AL proof with slight modifications. Thus, we have

Theorem 5.5: Let $\int_{\overline{\mathcal{A}/\mathcal{G}}}f\,d\mu_0$ be defined as in Definition 5.3, where the measure on $\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}}$ is in each case the Haar measure on $d^n\mu_{\mathrm{Haar}}$.

(1) The integral $\int_{\overline{\mathcal{A}/\mathcal{G}}}f\,d\mu_0$ is well-defined.
(2) The functional $F:\overline{\mathcal{HA}}\to\mathbb{C}$, $f\mapsto\int_{\overline{\mathcal{A}/\mathcal{G}}}f(\overline{A})\,d\mu_0(\overline{A})$ is linear, continuous, positive and Diff(M) invariant.
(3) The cylindrical measure $d\mu_0$ is a regular, positive and Diff(M)-invariant measure on $\overline{\mathcal{A}/\mathcal{G}}$.

Proof: It remains to prove the integral to be well defined. If it is, then our measure coincides with the AL measure defined only by the use of strongly independent hoops, since the AL measure is unique and we did not remove any of the conditions the integral has to fulfill—because any strongly independent $\boldsymbol{\beta}$ is moderately independent. Consequently, all the other assertions of the theorem already proven in Ref. 2 using the strong independence can be generalized to our problem.

(1) Let there be given an $f\in\mathcal{C}$ and two sets $\beta',\beta''\subseteq\mathcal{L}_m$ of moderately independent loops, such that f is cylindrical with respect to $\mathcal{B}_{\boldsymbol{\beta}'}$ and $\mathcal{B}_{\boldsymbol{\beta}''}$.

(2) W.l.o.g. choose the free segments $e'_{i'},e''_{i''}$ of $\beta'_{i'},\beta''_{i''}$, such that they are in each case completely contained in an edge of $\Gamma_{\boldsymbol{\beta}'\cup\boldsymbol{\beta}''}$. (If necessary, free $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}''$ from retracings and then use the argumentation of Proposition 3.10. For the definition of $\Gamma_{\boldsymbol{\beta}'\cup\boldsymbol{\beta}''}$ see Construction 3.4.) Now connect any vertex $v\neq m$ of $\Gamma_{\boldsymbol{\beta}'\cup\boldsymbol{\beta}''}$ with the base point m by a piecewise analytic Jordan path h_v, such that $h_v\cap h_{v'}$, $\forall v\neq v'$ and $h_v\cap\Gamma_{\boldsymbol{\beta}'\cup\boldsymbol{\beta}''}$ consist of at most a finite number of points. Construct all paths $\beta_i:=h_{e_i^-}e_ih_{e_i^+}^{-1}$, where e_i runs over all edges of $\Gamma_{\boldsymbol{\beta}'\cup\boldsymbol{\beta}''}$. Obviously, $\boldsymbol{\beta}',\boldsymbol{\beta}''\subseteq\mathcal{HG}(\boldsymbol{\beta}),\boldsymbol{\beta}:=\{\beta_i|i=1,...,n\}$, and also $\mathcal{HG}(\boldsymbol{\beta}'),\mathcal{HG}(\boldsymbol{\beta}'')\subseteq\mathcal{HG}(\boldsymbol{\beta})$. More precisely: Let $\beta'_j=\prod_{k_j=1}^{K_j}e_{i(j,k_j)}^{\epsilon(j,k_j)}$ be a (minimal) decomposition of β'_j into a sequence of edges, so $\beta'_j\sim\prod_{k_j=1}^{K_j}\beta_{i(j,k_j)}^{\epsilon(j,k_j)}$ is a (minimal) decomposition of β'_j in β_i. The same holds for β''_j.
Next, $\boldsymbol{\beta}$ is strongly independent with the free segments e_i.
Since $\mathcal{HG}(\boldsymbol{\beta}'),\mathcal{HG}(\boldsymbol{\beta}'')\subseteq\mathcal{HG}(\boldsymbol{\beta})$, f is also cylindrical with respect to $\mathcal{B}_{\boldsymbol{\beta}}$. Thus, it is sufficient to prove $\int_{\mathcal{B}_{\boldsymbol{\beta}'}}f_{\boldsymbol{\beta}''}\,d\mu_{\mathcal{B}_{\boldsymbol{\beta}'}}=\int_{\mathcal{B}_{\boldsymbol{\beta}}}f_{\boldsymbol{\beta}}\,d\mu_{\mathcal{B}_{\boldsymbol{\beta}}}$.

(3) Now we can express any $\beta'_i\in\boldsymbol{\beta}'$ by a product of $\beta_i\in\boldsymbol{\beta}$, such that for all $i\in[1,n']$ there exists a $K(i')\in[1,n]$ and that the following holds:

(a) $i\neq j'\Leftrightarrow K(i')\neq K(j')$;

(b) $\beta_{K(i')}$ is not used in any decomposition of the $\beta'_{j'}$, $j'<i'$, into elements of $\boldsymbol{\beta}$;

(c) $\beta_{K(i')}$ (or $\beta_{K(i')}^{-1}$) is used in any decomposition of $\beta_{i'}$ exactly once.
To see this choose $K(i')$, such that $e_{K(i')}$ contains the free segment of $\beta'_{i'}$. Since there is a bijection $e_i \leftrightarrow \beta_i$, these three conditions are only a reformulation of the criteria for the moderate independence of the $\beta'_{i'}$.

(4) Since f is as well cylindrical with respect to β as with respect to $\boldsymbol{\beta}'$, $f = \pi_{\boldsymbol{\beta}}^* f_{\boldsymbol{\beta}} = \pi_{\boldsymbol{\beta}'}^* f_{\boldsymbol{\beta}'}$. Analogously to Lemma 5.3 we have $\pi_{\boldsymbol{\beta}'} = \pi \pi_{\boldsymbol{\beta}}$, where

$$\begin{array}{rccc} \pi: & \mathbf{G}^n/\mathrm{Ad} & \to & \mathbf{G}^{n'}/\mathrm{Ad} \\ & [g_1,\ldots,g_n]_{\mathrm{Ad}} & \mapsto & \left[\prod_{k_1=1}^{K_1} g_{i(1,k_1)}^{\epsilon(1,k_1)}, \ldots, \prod_{k_{n'}=1}^{K_{n'}} g_{i(n',k_{n'})}^{\epsilon(n',k_{n'})} \right]_{\mathrm{Ad}} \end{array}$$

is defined due to the decompositions

$$\beta'_{i'} = \prod_{k'_i=1}^{K_{i'}} e_{i(i',k_{i'})}^{\epsilon(i',k_{i'})}.$$

Thus, we have

$$f_\beta = \pi^* f_{\boldsymbol{\beta}'}, \quad \text{i.e.,} \quad f_\beta([g_1,\ldots,g_n]_{\mathrm{Ad}}) = (\pi^* f_{\boldsymbol{\beta}'})([g_1,\ldots,g_n]_{\mathrm{Ad}})$$
$$= f_{\beta'}\left(\left[\prod_{k_1=1}^{K_1} g_{i(1,k_1)}^{\epsilon(1,k_1)}, \ldots, \prod_{k_{n'}=1}^{K_{n'}} g_{i(n',k_{n'})}^{\epsilon(n',k_{n'})} \right]_{\mathrm{Ad}} \right).$$

(5) Since we fixed the generators of $\mathcal{HG}(\boldsymbol{\beta})$, we can interpret the integration on $\boldsymbol{\mathcal{B}}_\beta$ as an integration on $\mathbf{G}^n/\mathrm{Ad}$. Since the Haar measure is Ad invariant, we can pull back any function of $\mathbf{G}^n/\mathrm{Ad}$ onto the whole $\mathbf{G}^n$ and integrate hereon. The analogon holds for $\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}'}$.

(6) Now we can integrate (considering $f_{\boldsymbol{\beta}}$ to be both a function on $\mathbf{G}^n$ and $\mathbf{G}^n/\mathrm{Ad}$):

$$\int_{\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}}} f_{\boldsymbol{\beta}} \, d\mu_{\boldsymbol{\mathcal{B}}_{\boldsymbol{\beta}}} = \int_{\mathbf{G}^n} \prod_{i=1}^{n} d\mu_i \, f_{\boldsymbol{\beta}}(g_1,\ldots,g_n)$$
$$= \int_{\mathbf{G}^n} \prod_{i=1}^{n} d\mu_i \, f_{\boldsymbol{\beta}'}\left(\prod_{k_1=1}^{K_1} g_{i(1,k_1)}^{\epsilon(1,k_1)}, \ldots, \prod_{k_{n'}=1}^{K_{n'}} g_{i(n',k_{n'})}^{\epsilon(n',k_{n'})} \right)$$
$$= \int_{\mathbf{G}^{n-n'}} \prod_{i=1, i \notin K([1,n'])}^{n} d\mu_i \int_{\mathbf{G}} d\mu_{K(1)} \cdots$$
$$\times \int_{\mathbf{G}} d\mu_{K(n')} f_{\boldsymbol{\beta}}{}'(\ldots, g_{K(1)} \ldots, \ldots, \ldots, g_{K(n')}, \ldots)$$

(permutation of the order of integration. The three dots in $\ldots, g_{K(i')} \ldots$ denote a product of g_i, which because of the construction above does not contain a $g_{K(j')}$ with $j' \geqslant i'$.)

$$= \int_{\mathbf{G}^{n-n'}} \prod_{i=1, i \notin K([1,n'])}^{n} d\mu_i \int_{\mathbf{G}} d\mu_{K(n')} \int_{\mathbf{G}} d\mu_{K(1)} \cdots$$
$$\times \int_{\mathbf{G}} d\mu_{K(n'-1)} f_{\boldsymbol{\beta}'}(\ldots, g_{K(1)} \ldots, \ldots, \ldots, g_{K(n'-1)}, \ldots, g_{K(n')})$$

(results from the translation invariance of the Haar measure, since for all $j'<n',\ldots,g_{K(j')},\ldots$ does not contain a factor $g_{K(n')}$ and since $g_{K(n')}$ appears in ..., $g_{K(n')},\ldots$ exactly once.)

$$\vdots$$

$$=\int_{\mathbf{G}^{n-n'}}\prod_{i=1,i\notin K([1,n'])}^{n} d\mu_i \int_{\mathbf{G}} d\mu_{K(1)}\cdots$$

$$\times\int_{\mathbf{G}} d\mu_{K(n')} f_{\beta'}(g_{K(1)},\ldots,g_{K(n')})$$

(We used successively the translation invariance of the Haar measure in order to eliminate the $\cdots$ products as in the step above.)

$$=\int_{\mathbf{G}^{n'}}\prod_{i=1}^{n'} d\mu_i\, f_{\boldsymbol{\beta}'}(g_1,\ldots,g_{n'})$$

(Normalization of the Haar measure and bijection $i'\leftrightarrow K(i')$)

$$=\int_{\mathcal{B}_{\beta'}} f_{\boldsymbol{\beta}'}\, d\mu_{\mathcal{B}_{\beta'}}.$$

(7) Thus, $\int_{\overline{\mathcal{A}/\mathcal{G}}} f\, d\mu_0$ is well-defined. (We assumed in our calculation, that in any case $\ldots,g_{K(i')},\ldots$ appears and $\ldots,g_{K(i')}^{-1},\ldots$ does not. Otherwise in all steps but the but the third we get a function partially in $g_{K(i')}^{-1}$. The claim remains valid since the Haar measure is invariant under inversions, i.e., we have $\int_{\mathbf{G}} d\mu_{\mathrm{Haar}} f(g)=\int_{\mathbf{G}} d\mu_{\mathrm{Haar}} f(g^{-1})$.)

Remark: The proof that the integral is well-defined gives us the earlier mentioned importance of moderate independence. Though the flag worlds in Sec. IV are usually not strongly independent, they can be used for the integration calculus. If one instead demanded only the weak independence for the definition of the integral, the integral would become *ill-defined*. Let, e.g., $\mathbf{G}=\mathrm{SU}(2)$ and β be a strongly or, equivalently, a moderately independent loop. $\gamma:=\beta^2$ is no longer moderately independent, but, of course, still weakly independent, since extracting the square root is possible in SU(2). Let now $f=\mathrm{tr}\, h_\gamma=\mathrm{tr}\, h_\beta^2$. f is cylindrical with respect to γ and with respect to β. We integrate f with respect to γ and receive $\int_{\overline{\mathcal{A}/\mathcal{G}}} f\, d\mu_0=\int_{\mathbf{G}}\mathrm{tr}\, g\, d\mu_{\mathrm{Haar}}=0$. But, with respect to β we have $\int_{\overline{\mathcal{A}/\mathcal{G}}} f\, d\mu_0=\int_{\mathbf{G}} tr\, g^2\, d\mu_{\mathrm{Haar}}=-1$, i.e., the integral is ill-defined. Thus, the moderate independence is best-suited for the mathematically rigorous calculation of the Wilson loop expectation values in Sec. VI.

VI. CALCULATION OF THE WILSON LOOP EXPECTATION VALUES

In this section the expectation values of the Wilson loop products $\langle T_{\alpha_1}\cdots T_{\alpha_n}\rangle$ of the pure Yang–Mills theory are computed. Thiemann[3] and Ashtekar *et al.*[4] were the first who succeeded in calculating $\langle T_{\alpha_1}\cdots T_{\alpha_n}\rangle$—at least for loops α_i that lie in a certain quadratic lattice—in the Ashtekar framework. Our goal is now to generalize their results for arbitrary α_i.

It is well known[2] that given the expectation values $\langle T_{\alpha_1}\cdots T_{\alpha_n}\rangle$ for all α_i one can reconstruct the measure $d\mu_{\mathrm{YM}}$ of the theory and vice versa. A direct definition $d\mu_{\mathrm{YM}}:=e^{-S[\overline{A}]}d\mu_0$ is difficult since one has to define the action S not only on $\mathcal{A}/\mathcal{G}$ but on the whole $\overline{\mathcal{A}/\mathcal{G}}$. The first step to overcome this problem is an appropriate regularization $S^W_{\mathrm{reg}}(A)$ of $S_{\mathrm{YM}}(A)=\int_M \frac{1}{4}\,\mathrm{tr}\,F_{\mu\nu}F^{\mu\nu}\,dx$. Since the only variables used *a priori* in the Ashtekar approach are the Wilson loops, it seems very likely to use the lattice regularization. Strictly speaking, TA$^+$ set

$$S^W_{\mathrm{reg}}(A):=\frac{N}{g^2a^2}\sum_{\Box}\left(1-\frac{1}{N}\mathrm{Re}\,\mathrm{tr}\,h_{\Box}(A)\right), \tag{2}$$

where a denotes the lattice spacing, $\Box$ runs over all plaquettes of the lattice, and $h_{\Box}(A)$ is the holonomy around the boundary of $\Box$. On the one hand, S^W_{reg} converges naively to S_{YM}, when the lattice grows ad infinitum and a goes to zero, and on the other hand, S^W_{reg} is a function of Wilson loops, i.e., it can be naturally extended from $\mathcal{A}/\mathcal{G}$ onto the whole $\overline{\mathcal{A}/\mathcal{G}}$. The second step of TA$^+$ was now the definition of $\langle T_{\alpha_1}\cdots T_{\alpha_n}\rangle$ exchanging limit and integration (L is the length of the lattice):

$$\langle T_{\alpha_1}\cdots T_{\alpha_n}\rangle:=\lim_{a\to 0,L\to\infty}\frac{1}{Z_{a,L}}\int_{\overline{\mathcal{A}/\mathcal{G}}}d\mu_0\,e^{-S^W_{\mathrm{reg}}}T_{\alpha_1}\cdots T_{\alpha_n}\text{``}=\text{''}\frac{1}{Z}\int_{\overline{\mathcal{A}/\mathcal{G}}}d\mu_0\,e^{-\lim_{a\to 0,L\to\infty}S^W_{\mathrm{reg}}}T_{\alpha_1}\cdots T_{\alpha_n}.$$

Now they were able to calculate explicitly the expectation values for all $\alpha_1,\ldots,\alpha_n$ contained in a quadratic lattice. Finally, they suggested to compute these values for general α_i by approximating them by certain lattice loops.

We avoid this problem using a slightly modified regularization. The idea is to adapt the regularization to the given loops and not vice versa. We consider any finite lattice with certain interior domains G generalizing the quadratic plaquettes $\Box$. Then we replace in (2) $\Box$ by G and also a^2 by $|G|$, the area of the interior domain G, in the denominator. Following the calculations of TA$^+$ we get an explicit formula for $\langle T_{\alpha_1}\cdots T_{\alpha_n}\rangle$ with arbitrary $\alpha_1,\ldots,\alpha_n$ that coincides with the naive limit of TA$^+$.

A. Regularization of the Wilson loop functionals

In this section we want to introduce our regularization.

Definition 6.1: Generalized Yang–Mills action. Let G be a simple domain in $\mathbb{R}^2$,$|G|$ its area, α_G a boundary loop of G and $[A]\in\mathcal{A}/\mathcal{G}$. Then we set

$$S_G([A]):=\frac{N}{g^2}\frac{1}{|G|}\left(1-\frac{1}{N}\mathrm{Re}\,\mathrm{tr}\,h_{\alpha_G}(A)\right)=\frac{N}{g^2}\frac{1}{|G|}(1-\mathrm{Re}\,T_{\alpha_G}(A)).$$

(This definition is obviously independent of the choice of the boundary loop and the chosen $A\in[A]$.)

Now let Γ be a finite simple graph in $\mathbb{R}^2$ with interior domains G. $\Gamma\to\mathbb{R}^2$ means that the supremum $\sup_{G\in\Gamma}\mathrm{diam}\,G$ of the diameters of the interior domains goes to 0 and the supremum of the diameters of all circles with center in m, which are completely contained in $\cup\overline{G}$ goes to ∞. (The choice of m is arbitrary. One can choose any point in $M=\mathbb{R}^2$, but one has to fix that point once for all.)

We set the *regularized Yang–Mills action* to be $S_\Gamma([A]):=\Sigma_G\,S_G([A])$ and define $S(A)\equiv S([A]):=\lim_{\Gamma\to\mathbb{R}^2}S_\Gamma([A])$.

Finally let $\overline{A}\in\overline{\mathcal{A}/\mathcal{G}}$. We define $S(\overline{A}):=\lim_{\Gamma\to\mathbb{R}^2}\Sigma_G\,S_G(\overline{A})$ to be the *generalized Yang–Mills action*.

Remark: While we have written the present paper the article ‘‘Study of Wilson loop functionals in 2D Yang–Mills theories’’ of Aroca and Kubyshin[5] appeared. They used an analogous regularization, i.e., they also permitted arbitrarily bounded domains instead of the usual quadratic plaquettes. They even considered a more general class of actions $S_\Gamma(A):=\Sigma_G\,S_1(h_{\alpha_G}(A))$, where G runs over all plaquettes which the lattice on the (compact) two-dimensional manifold is divided into and where S_1 has to fulfill the following axioms

(1) $S_1(g)=S_1(g^{-1})$ for all $g\in\mathbf{G}$;
(2) $S_1(g)$ has an absolute minimum in $g=e_{\mathbf{G}}$;
(3) $\lim_{G\to\{x\}}(1/|G|)S_1(h_{\alpha_G}(A))=\frac{1}{2}\operatorname{tr}F_{\mu\nu}(x)F^{\mu\nu}(x)$.

Our definition reduces to that of Thiemann, Ashtekar *et al.*[3,4,12] if all domains G are quadratic and congruent with area $|G|=a^2$ (a... lattice spacing). One can prove—at least in a naive limit—that $S(A)$ converges pointwise to $S_{\mathrm{YM}}:=\frac{1}{4}\int_M \operatorname{tr}F_{\mu\nu}F^{\mu\nu}\,dx$ with $F_{\mu\nu}=\partial_{[\mu}A_{\nu]}-ig[A_\mu,A_\nu]$.

Note that the limit $S(\overline{A})$ is very formal because one can easily prove that this limit neither is unique nor exists for general $\overline{A}\in\overline{\mathcal{A}/\mathcal{G}}$ However, why should we need the existence or uniqueness of the limit $S(\overline{A})$? Actually, we only have to calculate terms like

$$\int_{\overline{\mathcal{A}/\mathcal{G}}}\exp\left[-\lim_{\Gamma\to\mathbb{R}^2}\sum_G S_G(\overline{A})\right]\widetilde{T}_{\alpha_1}(\overline{A})\cdots\widetilde{T}_{\alpha_n}(\overline{A})\,d\mu_0.$$

In order to use the integration calculus one has to exchange the limit and the integral. *A priori* we do not know, whether this is—at least mathematically—correct. Astonishingly, one can prove that such an exchange makes the limit *of the integrals* independent of the limiting process. By now, we do not really know which effect is responsible for that behavior.

B. Results

Given a finite set $\boldsymbol{\alpha}=\{\alpha_1,\ldots,\alpha_n\}$ of loops. We have to calculate the following expressions:

$$\chi(\boldsymbol{\alpha}):=\lim_{\Gamma\to\mathbb{R}^2}\frac{1}{Z}\int_{\overline{\mathcal{A}/\mathcal{G}}}\exp\left[-\sum_G S_G(\overline{A})\right]\widetilde{T}_{\alpha_1}(\overline{A})\cdots\widetilde{T}_{\alpha_n}(\overline{A})\,d\mu_0.$$

Z is chosen here so that we have $\chi(1)=1$. (Strictly speaking, Z actually depends on Γ, but we suppress this here and in the sequel.)

Due to the analyticity of the loops the set $\boldsymbol{\alpha}$ generates a finite, nonempty, planar and connected graph $\Gamma_{\boldsymbol{\alpha}}$. We enlarge $\Gamma_{\boldsymbol{\alpha}}$ to an ordinary graph (Sec. III D) and afterwards to a simple, ordinary graph (Proposition 4.3) again denoted by $\Gamma_{\boldsymbol{\alpha}}$ with the interior domains G_I, $I=1,\ldots,\lambda$. Furthermore we choose any moderately independent flag world $\mathcal{F}=\{f_I\}$ for $\Gamma_{\boldsymbol{\alpha}}$ existing due to Proposition 4.9. Now, due to Corollary 4.10 any hoop in $\Gamma_{\boldsymbol{\alpha}}$ can be expressed by a hoop product of flags in $\mathcal{F}$, i.e., by a product of nonoverlapping loops:

$$\alpha_i=\prod_{j=1}^{j_i} f_{I(i,j)}^{\epsilon(i,j)},\quad \forall i=1,\ldots,n;\epsilon(i,j)=\pm 1.$$

Finally, we demand that any graph Γ in the limiting process is a refinement of Γ_α.

Thus, we arrived at the point where we can reuse the calculations of TA^+. We have to replace simply the plaquette loops in TA^+ by the flags and, analogously, the corresponding refinements. Therefore we skip the technical details (for that purpose, see, e.g., Ref. 13 and present only the result:

$$\chi(\boldsymbol{\alpha})=\frac{1}{N^n}\prod_{I=1}^{\lambda}\left(\sum_{(m)}e^{-(1/2)g^2c_{(m)}|G_I|}p_{(m)}^{(n)}\otimes^n 1\right)_{\vec{A}_I}^{\vec{B}_I}\mathcal{C}_{\vec{C}}^{\vec{D}}\mathcal{E}_{\vec{E}}^{\vec{F}},$$

which for $\alpha_1,\ldots,\alpha_n$ contained in a quadratic lattice coincides with the results of TA^+. Here, $p_{(m)}^{(n)}$ are the projectors built from the Young tableaux (m) for $\otimes^n g$ and $c_{(m)}$ is the eigenvalue of the corresponding quadratic Casimir operator. n depends only on the decomposition of the α_i into the flags f_I. Finally, $\mathcal{C}_{\vec{C}}^{\vec{D}}$ and $\mathcal{E}_{\vec{E}}^{\vec{F}}$ are certain tensors and $\vec{A}_I$ and $\vec{B}_I$ indices whose structure is also determined by the algebraic structure of that decomposition.

For $\mathbf{G}=\mathrm{U}(1)$ we get

$$\chi(\boldsymbol{\alpha})=\prod_{I=1}^{\lambda} e^{-(1/2)g^2 n_I^2 |G_I|},$$

where n_I is the ''effective'' winding number of $\alpha_1\cdots\alpha_n$ around the domain G_I.

In conclusion we emphasize that the limits above are completely independent of the limiting process $\Gamma\to\mathbb{R}^2$ supposed $\Gamma_{\boldsymbol{\alpha}}$ is a restriction of any graph in the limiting process. Thus, the limit exists and is unique.

VII. DISCUSSION

In Sec. VI we ''derived'' the expectation values of the Wilson loop products. Actually, the word ''derived'' is an exaggeration—de facto we *defined* the values even if the Yang–Mills action on $\mathcal{A}/\mathcal{G}$ influenced the definition of χ. But we did not deduce the values of χ from S_{YM} in a mathematically correct way. Formally we got χ by

$$\chi(\boldsymbol{\alpha})=\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_0\, e^{-S_{\mathrm{YM}}(\bar{A})}T_{\alpha_1}\cdots T_{\alpha_n}=\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_0 e^{-\lim S_{\mathrm{reg}}(\bar{A})}T_{\alpha_1}\cdots T_{\alpha_n},$$

i.e., by extending S_{YM} onto $\overline{\mathcal{A}/\mathcal{G}}$, and subsequently by exchanging the limiting process and the integration

$$\chi(\boldsymbol{\alpha}):=\lim \int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_0\, e^{-S_{\mathrm{reg}}(\bar{A})}T_{\alpha_1}\cdots T_{\alpha_n}.$$

Consequently, this definition is the actual start of our considerations. In principle, that approach is a kind of constructive quantum field theory that needs a physical justification only a posteriori.

In Sec. VI A we already noted that the regularization of S_{YM} by

$$S:=\lim_{\Gamma\to\mathbb{R}^2}\sum_G \frac{N}{g^2}\frac{1}{|G|}\left(1-\frac{1}{N}\mathrm{Re}\,\mathrm{tr}\,h_{\alpha_G}\right)$$

makes no problems on $\mathcal{A}/\mathcal{G}$, but breaks down on $\overline{\mathcal{A}/\mathcal{G}}$, because the limit does not exist in general. Thus, S cannot be in $\overline{\mathcal{HA}}$. But, surprisingly the exchange of limit and integral yields very regular results. We have even shown[13] that the limit $\chi(\boldsymbol{\alpha})$ for our choice of regularization exists for all finite $\alpha\subseteq\mathcal{L}_m$ and is independent of the limiting process. Is there a deeper reason behind that?

However, we know that the given expectation values define a unique Borel measure μ on $\overline{\mathcal{A}/\mathcal{G}}$[2] because we can extend these values to a linear continuous positive functional on $\overline{\mathcal{HA}}$. Note that originally the expectation values are not mutually independent, but subjected to the so-called Mandelstam relations. Since we defined the expectation values using integrals on T_α, these relations are indeed implemented. What properties does μ have? Is μ *strictly* positive or is μ absolutely continuous with respect to the induced Haar measure μ_0? Is it even possible to define an action S on $\overline{\mathcal{A}/\mathcal{G}}$ directly, i.e., without regularization, and is it therefore possible to get the desired measure by $d\mu:=e^{-S}d\mu_0$?

The choice of regularization is also worth being discussed. In the present case the regularization of S_{YM} depends crucially on the dimension two. It cannot be extended to three or more dimensions because it uses—roughly speaking—the chance that for two-dimensional manifolds a loop has both dimension and codimension one. But the codimension is decisive. To avoid renormalization one has to regularize the d-dimensional Yang–Mills theory by

$$\lim_{\sup\ \mathrm{diam}\, G\to 0} \frac{N}{g^2}\frac{1}{\mathrm{vol}\, G}\sum_{\{G\}}\left(1-\frac{1}{N}\mathrm{Re}\,\mathrm{tr}\,\mathcal{P}\exp\left(-ig\int_{\partial G}A\right)\right),$$

where $\{G\}$ is a decomposition of the base manifold into certain d-dimensional objects. How to connect $\mathcal{P}\exp(-ig\int_{\partial G}A)$ and $\mathcal{P}\exp(-ig\int_{\alpha}A)$? Moreover, the used propositions for planar graphs cannot be applied to higher dimensions. Thus, from dimension three on problems of knot theory will be important and so will methods of the topological quantum field theory. Perhaps using algebraic topology or invariant theory one can specify a class of constructible models.

Let us return finally to the concrete generalization of the two-dimensional Yang–Mills theory within the Ashtekar approach. In the last few years some papers were published that calculated the expectation values of the Wilson loops in $\mathcal{A}/\mathcal{G}$ (e.g., Refs. 14–16) and performed the continuum limit. They provided an area law, an indication for the confinement in the theory. All in all these papers delivered the same result as the Ashtekar approach does today. Thus, we get a little justification for our choice of the regularization. Perhaps it is possible to translate further models into the new approach and to confirm that way the results got on $\mathcal{A}/\mathcal{G}$. However, it seems to be unlikely that one gets—at least in the next time—general assertions for the equivalence of the ''classical'' and the Ashtekar approach. But, from the mathematical point of view this would be very interesting because some problems of the classical approach could be circumvented.

ACKNOWLEDGMENTS

I am very grateful to Gerd Rudolph for his great support while I wrote my diploma thesis and the present paper. I thank the Studienstiftung des deutschen Volkes for its long-time grant and the Max-Planck-Institut für Mathematik in den Naturwissenschaften for its generous promotion.

APPENDIX: PROOF OF PROPOSITION 3.1

Lemma A.1.: Let γ be a Jordan path in M contained completely in a chart U_i with trivialization χ_i and p any point in the fiber over $\gamma(0)$. Furthermore, let $\mathbf{G}$ be compact and connected and $\epsilon \in]0,\frac{1}{2}[$ arbitrary.

Then $\tau_{\gamma,\mathcal{A}_{\epsilon,\gamma,i}}(p)=P_{\gamma(1)}$, where $\mathcal{A}_{\epsilon,\gamma,i}$ is defined by $\mathcal{A}_{\epsilon,\gamma,i}:=\{A\in\mathcal{A}|A_i(\gamma(t))\equiv 0$ for $t\notin[\epsilon,1-\epsilon]\}$, i.e., any point of $P_{\gamma(1)}$ can be reached by parallel transport starting in p with respect to connections in $\mathcal{A}_{\epsilon,\gamma,i}$. $[\tau_{\gamma,A}(p):P\to P$ is the parallel transport of p along γ with respect to the connection A and $\tau_{\gamma,p}(A)$ the corresponding group element.]

Proof: (1) Let $p\equiv p_i:=s_i(\gamma(0)):=\chi_i^{-1}(\gamma(0),e_{\mathbf{G}})$. Then $\tau_{\gamma,p}(A)=\mathcal{P}\exp[-\int_\gamma A_i(\dot\gamma)dt]$, where $\dot\gamma$ is the tangential vector field to γ and A_i is the connection A in the local trivialization χ_i. (We dropped the factor ig.)

(2) Obviously, there is a one-form $a_i:TU_i\to\mathbb{C}$ with $a_i(\dot\gamma)|_{\gamma([0,\epsilon]\cup[1-\epsilon,1])}\equiv 0$ and $-\int_\gamma a_i(\dot\gamma)dt=1\neq 0$.

(3) Set $A_{\lambda,i}:=a_i\otimes\lambda$ for any $\lambda\in\mathbf{g}$ and extend $A_{\lambda,i}$ to a connection A_λ on TP. (This is possible, see, e.g., Ref. 17, p. 67.) Obviously, $A_\lambda\in\mathcal{A}_{\epsilon,\gamma,i}$ for any $\lambda\in\mathbf{g}$.

(4) For λ constant, we have $\tau_{\gamma,p}(A_\lambda)=\mathcal{P}\exp[-(\int_\gamma a_i(\dot\gamma)dt)\lambda]=e^\lambda$.

(5) Since the image of the Lie algebra $\mathbf{g}$ under the exponential map is the connected component of unity of the Lie group $\mathbf{G}$, we have $\mathbf{G}\supseteq\tau_{\gamma,p}(\mathcal{A}_{\epsilon,\gamma,i})\supseteq\{\tau_{\gamma,p}(A_\lambda)|\lambda\in\mathbf{g}\}=\{e^\lambda|\lambda\in\mathbf{g}\}=\mathbf{G}$, i.e., $\mathbf{G}=\tau_{\gamma,p}(\mathcal{A}_{\epsilon,\gamma,i})$ and thus $\tau_{\gamma,\mathcal{A}_{\epsilon,\gamma,i}}(p)=P_{\gamma(1)}$.

(6) Now let p be arbitrary. Since the parallel transport commutes with the right action, we have $\tau_{\gamma,\mathcal{A}_{\epsilon,\gamma,i}}(p)=(\mathrm{Ad}\,\psi_g)\tau_{\gamma,\mathcal{A}_{\epsilon,\gamma,i}}(p_i)=(\mathrm{Ad}\,\psi_g)=P_{\gamma(1)}=P_{\gamma(1)}$ because $\mathbf{G}$ acts freely on P. We chose g, such that $p=p_i\cdot g$. **qed**

Proof: (Proposition 3.1) Let $\boldsymbol{\alpha}:=\{\alpha_1,\ldots,\alpha_n\}$ be a set of moderately independent loops. We have to show that for any n-tupel $(g_1,\ldots,g_n)\in\mathbf{G}^n$ there is an $A\in\mathcal{A}$ with $h_{\alpha_i}(A)=g_i\ \forall 1\leqslant i\leqslant n$. (Note, that we have fixed a trivialization χ and therefore a base point p in P_m from the very beginning.)

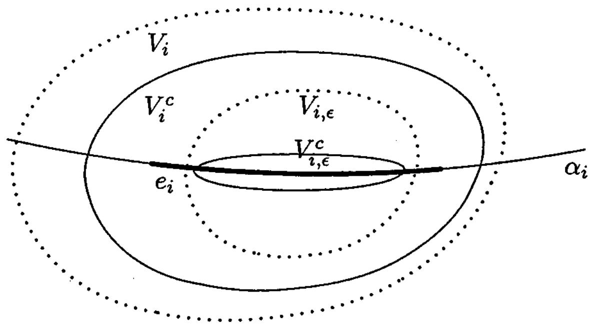

FIG. 8. The domains V_i, V_i^c, $V_{i,\epsilon}$, and $V_{i,\epsilon}^c$.

Fix a covering $\{U_k\}$ of M. Choose a free segment e_i to any $\alpha_i \in \boldsymbol{\alpha}$ due to Definition 3.2, such that

(1) $\alpha_i = f_i^- e_i f_i^+$ with $f_i^\pm \cap e_i = \emptyset$ and $\alpha_j \cap e_i = \emptyset, \forall_j < i$ and
(2) any free segment lies completely in a chart U_i.

Next, choose open neighborhoods V_i of e_i, such that $V_i \subseteq U_i$ are mutually disjoint and that $\alpha_j \cap V_i = \emptyset$ for any $j < i$, and modify the covering of M in that way, that V_i lies in exactly one chart (denoted again by U_i). Furthermore, choose open $V_{i,\epsilon}$ and compact sets $V_i^c, V_{i,\epsilon}^c$ with some $\epsilon \in]0,\frac{1}{2}[$, $e_i \subseteq V_i^c \subseteq V_i$ and $V_{i,\epsilon}^c \subseteq V_{i,\epsilon} \subseteq V_i^c$, $V_{i,\epsilon} \cap f_i^\pm = \emptyset$ and $\gamma_i(t) \in V_{i,\epsilon}^c \Leftrightarrow t \in [\epsilon, 1-\epsilon]$, where $\gamma_i : [0,1] \to e_i$ is a parametrization of e_i. (See Fig. 8.)
It is a well-known fact that there exists a $\phi \in C^\infty(M)$ with $\phi \equiv 1$ on $\cup V_i^c$ and $\phi \equiv 0$ on $M \setminus \cup V_i$ and analogously a $\phi_{i,\epsilon} \in C^\infty(M)$ with $\phi_{i,\epsilon} \equiv 1$ auf $V_{i,\epsilon}^c$ and $\phi_{i,\epsilon} \equiv 0$ on $M \setminus V_{i,\epsilon}$ for all i.
Let $B \in \mathcal{A}$ be some connection.

(1) $i = 0$.

$A^{(0)} := B - \phi B$ is again a connection. (This simple notation means: There is an $A^{(0)}$ for that B, such that $A_i^{(0)} = (1-\phi) B_i$ on $\cup V_i$ and elsewhere $B = A^{(0)}$ since because of the special selection of V_i the compatibility conditions of chart changes are not touched.) We have now $A_j^{(i)} \equiv 0$ on e_j for all $j > i = 0$ (and obviously $h_{\alpha_j}(A^{(i)}) = g_j$ for all $j \leqslant i = 0$).

(2) $i > 0$.

Let $p_{i,-} := \tau_{f_i^-, A^{(i-1)}}(p) \in P_{\gamma_i(0)}$ be the parallel transport to $A^{(i-1)}$ of p along f_i^- and $p_{i,+} := \tau^{-1}_{f_i^+, A^{(i-1)}}(p \cdot g_i)$ the "inverse" parallel transport with respect to $A^{(i-1)}$ along f_i^+ leading from $P_{\gamma_i(1)}$ to $p \cdot g_i$. Due to the lemma above there is an $A' \in \mathcal{A}_{\epsilon,\gamma_i,i}$ with $p_{i,+} = \tau_{e_i,A'}(p_{i,-})$ and we have

$$p \cdot g_i = \tau_{f_i^+, A^{(i-1)}}(\tau_{e_i, A'}(\tau_{f_i^-, A^{(i-1)}}(p)))$$

$$= \tau_{f_i^+, A^{(i-1)} + \phi_{i,e} A'}(\tau_{e_i, A'}(\tau_{f_i^{-1}, A^{(i-1)} + \phi_{i,e} A'}(p))) \quad (\text{due to } \phi_{i,\epsilon} \equiv 0 \text{ on } f_i^\pm)$$

$$= \tau_{f_i^+, A^{(i-1)} + \phi_{i,e} A'}(\tau_{e_i A^{(i-1)} + \phi_{i,e} A'}(\tau_{f_i^-, A^{(i-1)} + \phi_{i,e} A'}(p)))$$

$$(\text{due to } A_i^{(i-1)}|_{e_i \equiv 0} \text{ and } \phi_{i,\epsilon}|_{\text{supp}\, A_i' \cap e_i} \equiv 1)$$

$$= \tau_{f_i^- e_i f_i^+, A^{(i-1)} + \phi_{i,e} A'}(p) = \tau_{\alpha_i, A^{(i)}}(p),$$

where we set $A^{(i)}:=A^{(i-1)}+\phi_{i,\epsilon}A'$. Obviously, $A^{(i)}$ is a connection, and we get $h_{\alpha_i}(A^{(i)})=g_i$. Since $A^{(i)}=A^{(i-1)}$ outside V_i and $V_i\cap\alpha_j=\emptyset$ $\forall_j<i$, we have also $h_{\alpha_j}(A^{(i)})=h_{\alpha_j}(A^{(i-1)})$ $=g_j$, $\forall_j<i$ by induction. Furthermore, we have $A_j^{(i)}\equiv 0$ on e_j for all $j>i$.

The proof ends setting $A:=A^{(n)}$. **qed**

[1] A. Ashtekar and C. J. Isham, Class. Quantum Grav. **9**, 1433 (1992).
[2] A. Ashtekar and J. Lewandowski, in *Knots and Quantum Gravity*, edited by J. Baez, Oxford Lecture Series in Mathematics and its Applications (Oxford University Press, Oxford, 1994).
[3] T. Thiemann, ''An axiomatic approach to quantum gauge field theory,'' hep-th/9511122.
[4] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourao, and T. Thiemann, J. Math. Phys. **38**, 5453 (1997).
[5] J. M. Aroca and Yu. A. Kubyshin, ''Study of Wilson loop functionals in 2-D Yang–Mills theories,'' hep-th/9802048.
[6] R. Giles, Phys. Rev. D **24**, 2160 (1981).
[7] A. D. Rendall, Class. Quantum Grav. **10**, 605 (1993).
[8] D. Marolf and J. M. Mourao, Commun. Math. Phys. **170**, 583 (1995).
[9] A. G. Kurosch, *Gruppentheorie* (Akademie, Berlin, 1970), Vol. 1.
[10] W. S. Massey, *Algebraic Topology: An Introduction* (Springer, New York, 1989).
[11] R. Halin, *Graphentheorie* (Akademie, Berlin, 1989).
[12] A. Ashtekar and J. Lewandowski, J. Math. Phys. **36**, 2170 (1995).
[13] C. Fleischhack, Diplomarbeit, Universität Leipzig, 1998.
[14] S. Klimek and W. Kondracki, Commun. Math. Phys. **113**, 389 (1987).
[15] B. K. Driver, Commun. Math. Phys. **123**, 575 (1989).
[16] L. Gross, C. King, and A. Sengupta, Ann. Phys. (N.Y.) **194**, 65 (1989).
[17] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1963), Vol. 1.

Quantization of massive vector fields in curved space–time

Edward P. Furlani
Department of Physics, State University of New York at Buffalo, Buffalo, New York 14260

(Received 7 December 1998; accepted for publication 12 March 1999)

We develop a cannonical quantization for massive vector fields on a globally hyperbolic Lorentzian manifold. © *1999 American Institute of Physics.* [S0022-2488(99)04506-5]

I. INTRODUCTION

Rigorous theories have been developed for scalar, Dirac, and electromagnetic quantum fields on globally hyperbolic manifolds.[1–6] However, apparently few rigorous results exist for massive vector fields on such manifolds.[7] A particle described by such a field is the ω-meson with a mass $m_\omega = 783$ MeV.[8] In this article, we develop a canonical quantization for massive vector fields on a globally hyperbolic Lorentzian manifold. The analysis is divided into two parts, the classical problem and the quantum problem.

For the classical problem, we start with the equations of motion $(\delta^{(4)}d^{(4)} + m^2)\mathcal{A} = 0$ which are the curved space–time generalizations of Proca's equations. We reduce these to a hyperbolic system $(\Box + m^2)\mathcal{A} = 0$, and a constraint $\delta^{(4)}\mathcal{A} = 0$. The hyperbolic system has global fundamental solutions $E_m^{\pm(1)}$ and a propagator $E_m^{(1)} = E_m^{+(1)} - E_m^{-(1)}$.[9–12] We introduce a series of operators $\rho_{(0)}$, $\rho_{(d)}$, $\rho_{(\delta)}$, and $\rho_{(n)}$ that map a solution $\mathcal{A}$ to its Cauchy data $A_{(0)}$, $A_{(d)}$, $A_{(\delta)}$, and $A_{(n)}$, respectively. These operators have transposes $\rho'_{(0)}$, $\rho'_{(d)}$, $\rho'_{(\delta)}$, and $\rho'_{(n)}$. We construct a series of operators $E_m^{(1)}\rho'_{(0)}$, $E_m^{(1)}\rho'_{(d)}$, $E_m^{(1)}\rho'_{(\delta)}$, and $E_m^{(1)}\rho'_{(n)}$ and show that these operators collectively map Cauchy data to a unique field solution. We apply these operators to obtain a global solution for Proca's equations wherein we satisfy the constraint by restricting the Cauchy data. Our representation of the solution is apparently new, and especially useful for field quantization as we demonstrate. This method also applies to Maxwell's equations. Since these equations are also of interest, but not needed for our quantum problem, we treat them separately in an appendix.

For the quantum problem, we start with a representation $(\phi, \pi, \mathcal{H})$ of the CCRs on an arbitrary Cauchy surface. We construct a space–time field operator $\mathcal{A}$ in terms of the data (ϕ, π) in accordance with the classical initial value problem. We then pass to the Weyl form W of the CCRs and study the C^* algebra of observables $\mathfrak{A}$ generated by W. We show that $\mathfrak{A}$ is independent of the representation on a given Cauchy surface, and also independent of the Cauchy surface. This work generalizes Dimock's treatment of scalar fields.[3]

II. PRELIMINARY CONCEPTS

Let $(\mathcal{M}, g)$ be a globally hyperbolic, orientable, time-orientable, space–time consisting of a smooth four-dimensional manifold $\mathcal{M}$ endowed with a smooth Lorentzian metric g with signature $(-1,1,1,1)$. As a consequence of global hyperbolicity, there is a (nonunique) smooth time coordinate t, and $(\mathcal{M}, g)$ can be foliated by a one-parameter family of Cauchy surfaces $\{t\} \times \Sigma_t$ giving it the topology $\mathcal{M} \approx \mathbb{R} \times \Sigma$.[2,13] Each Cauchy surface Σ inherits a smooth, proper Riemannian metric γ. We label events in $\mathcal{M}$ using (t,x) where $x \in \Sigma$, and adopt the standard convention in which Greek subscripts apply to $(\mathcal{M}, g)$ taking values from 0 to 4, and Latin subscripts apply to (Σ, γ) and range from 1 to 3.

Let $\mathcal{E}^{(p)}(\mathcal{M})$ denote the space of smooth, real-valued p-forms on $(\mathcal{M}, g)$ and $\mathfrak{D}^{(p)}(\mathcal{M})$ specify such forms with compact support. These spaces have duals $\mathcal{E}'^{(p)}(\mathcal{M})$ and $\mathfrak{D}'^{(p)}(\mathcal{M})$ with topologies similar to those defined for the corresponding spaces on $\mathbb{R}^n$.[10–12] We use the standard notation $\langle \mathcal{T},\mathcal{F}\rangle$ to denote the action of a distribution $\mathcal{T}$ on a test function $\mathcal{F}\in\mathfrak{D}^{(p)}(\mathcal{M})$. Specializing to $\mathcal{T}\in\mathfrak{D}^{(p)}(\mathcal{M})\subset\mathfrak{D}'^{(p)}(\mathcal{M})$ we have $\langle \mathcal{T},\mathcal{F}\rangle=\langle \mathcal{T},\mathcal{F}\rangle_{\mathcal{M}}$, where

$$\langle \mathcal{T},\mathcal{F}\rangle_{\mathcal{M}}\equiv\int_{\mathcal{M}}\mathcal{T}\wedge *^{(4)}\mathcal{F}, \tag{1}$$

is the global inner product, and $*^{(4)}:\mathfrak{D}^{(p)}(\mathcal{M})\rightarrow\mathfrak{D}^{(4-p)}(\mathcal{M})$ is the Hodge star operator with respect to g $((*^{(4)})^2=(-1)^{p+1})$. On $(\mathcal{M},g)$ we have an exterior derivative $d^{(4)}$ and codifferential $\delta^{(4)}=*^{(4)}d^4*^{(4)}$ which are the formal adjoints of one another $\langle d^{(4)}\mathcal{A},\mathcal{F}\rangle_{\mathcal{M}}=\langle\mathcal{A},\delta^{(4)}\mathcal{F}\rangle_{\mathcal{M}}$, where $\mathcal{A}\in\mathfrak{D}^{(p-1)}(\mathcal{M})$ and $\mathcal{F}\in\mathfrak{D}^{(p)}(\mathcal{M})$. There is also the D'Alembertian $\Box=(\delta^{(4)}d^{(4)}+d^{(4)}\delta^{(4)})$ which is formally self-adjoint (on Minkowski space $\Box=\partial_0^2-\nabla^2$).[14,15]

Let $\mathcal{E}^{(k)}(\Sigma)$, $\mathfrak{D}^{(k)}(\Sigma)$, $\mathcal{E}'^{(k)}(\Sigma)$ and $\mathfrak{D}'^{(k)}(\Sigma)$ denote the corresponding spaces on (Σ,γ). For $T\in\mathfrak{D}^{(k)}(\Sigma)\subset\mathfrak{D}'^{(k)}(\Sigma)$ we have $\langle T,F\rangle=\langle T,F\rangle_{\Sigma}$, where

$$\langle T,F\rangle_{\Sigma}\equiv\int_{\Sigma}T\wedge *^{(3)}F, \tag{2}$$

is the global inner product on (Σ,γ), and $*^{(3)}$: $\mathfrak{D}^{(k)}(\Sigma)\rightarrow\mathfrak{D}^{(3-k)}(\Sigma)$ is the Hodge star operator with respect to γ $((*^{(3)})^2=1)$. The exterior derivative $d^{(3)}$ and codifferential $\delta^{(3)}=(-1)^k*^{(3)}d^{(3)}*^{(3)}$ on (Σ,γ) are formal adjoints of one another $\langle d^{(3)}A,F\rangle_{\Sigma}=\langle A,\delta^{(3)}F\rangle_{\Sigma}$, with $A\in\mathfrak{D}^{(k-1)}(\Sigma)$ and $F\in\mathfrak{D}^{(k)}(\Sigma)$. The Laplace–Beltrami operator $\triangle=\delta^{(3)}d^{(3)}+d^{(3)}\delta^{(3)}$ is formally self-adjoint.[14,15] Finally, we work in units in which $c=1$.

III. PROCA'S EQUATIONS

We study the initial value problem for a vector field $\mathcal{A}$ of mass $m\in(0,\infty)$ satisfying the curved space–time generalization of Proca's equations

$$(\delta^{(4)}d^{(4)}+m^2)\mathcal{A}=0. \tag{3}$$

These equations reduce to the hyperbolic system

$$(\Box+m^2)\mathcal{A}=0, \tag{4}$$

and a constraint

$$\delta^{(4)}\mathcal{A}=0. \tag{5}$$

The system (4) has unique fundamental solutions $E_m^{\pm^{(1)}}(p,q)\in\mathfrak{D}'^{(1)}(\mathcal{M}_p\times\mathcal{M}_q)$ where $E_m^{\pm^{(1)}}(p,q):\mathfrak{D}^{(1)}(\mathcal{M}_q)\rightarrow\mathcal{E}^{(1)}(\mathcal{M}_p)$ ($p,q\in\mathcal{M}$ represent space–time events and $\mathcal{M}_p$, $\mathcal{M}_q$ identify the action of $E_m^{\pm^{(1)}}(p,q)$).[9,10,16] These solutions satisfy

$$(\Box_p+m^2)E_m^{\pm^{(1)}}(p,q)=\delta^{(1)}(p,q), \tag{6}$$

where $\delta^{(1)}(p,q)$ is the Dirac 1-tensor kernel, and supp $E_m^{\pm^{(1)}}(\cdot,q)\subset J^{\pm}(q)$. Also $\mathcal{A}^{\pm}(p)=\langle E_m^{\pm^{(1)}}(p,q),\mathcal{F}(q)\rangle_{\mathcal{M}_q}$ satisfies $(\Box+m^2)\mathcal{A}^{\pm}=\mathcal{F}$, with

$$\mathrm{supp}(\mathcal{A}^{\pm})\subset J^{\pm}(\mathrm{supp}(\mathcal{F})), \tag{7}$$

where $J^{\pm}(S)$ is the set of point in $(\mathcal{M},g)$ that can be reached from the set $S\subset\mathcal{M}$ by a future/past directed causal curve.[10]

The kernels $E_m^{\pm^{(1)}}(p,q)$ are identified with operators $E_m^{\pm^{(1)}}:\mathfrak{D}^{(1)}(\mathcal{M})\rightarrow\mathcal{E}^{(1)}(\mathcal{M})$ and from (6),

$$E_m^{\pm(1)}(\Box+m^2)=(\Box+m^2)E_m^{\pm(1)}=I. \tag{8}$$

The $E_m^{\pm(1)}$, which are linear and continuous, give rise to transpose operators $E_m^{\pm\prime(1)}:\mathcal{E}'^{(1)}(\mathcal{M})\to\mathfrak{D}'^{(1)}(\mathcal{M})$ which are also continuous.[17] Since $\Box+m^2$ is formally self-adjoint $E_m^{\pm\prime(1)}=E_m^{\mp(1)}$ on $\mathfrak{D}^{(1)}(\mathcal{M})$. Moreover, since $E_m^{\pm\prime(1)}$ are linear and continuous on $\mathcal{E}'^{(1)}(\mathcal{M})$ we extend $E_m^{\pm(1)}$ from $\mathfrak{D}^{(1)}(\mathcal{M})\subset\mathcal{E}'^{(1)}(\mathcal{M})$ to $\mathcal{E}'^{(1)}(\mathcal{M})$, i.e.,

$$E_m^{\mp(1)}=E_m^{\pm\prime(1)}:\mathcal{E}'^{(1)}(\mathcal{M}_p)\to\mathfrak{D}'^{(1)}(\mathcal{M}_q). \tag{9}$$

Lastly, we introduce the propagator

$$E_m^{(1)}=E_m^{+(1)}-E_m^{-(1)},\quad \text{(propagator)},$$

where $(\Box+m^2)E_m^{(1)}=0$. This operator has a transpose $E_m'^{(1)}=E_m^{+\prime(1)}-E_m^{-\prime(1)}$ and from (9) we have

$$E_m'^{(1)}=-E_m^{(1)}:\mathcal{E}'^{(1)}(\mathcal{M}_p)\to\mathfrak{D}'^{(1)}(\mathcal{M}_q). \tag{10}$$

Next, we introduce a series of operators that collectively map a solution of (4) to its data. Choose any Cauchy surface say $\{0\}\times\Sigma$, and let $i:\Sigma\to\mathcal{M}$ be the inclusion operator with pullback i^*. We define the operators

$$\begin{aligned}\rho_{(0)}&\equiv i^* \quad \text{(pullback)}\\ \rho_{(d)}&\equiv -*^{(3)}i^**^{(4)}d^{(4)} \quad \text{(forward normal derivative)}\\ \rho_{(\delta)}&\equiv i^*\delta^{(4)} \quad \text{(pullback of divergence)}\\ \rho_{(n)}&\equiv -*^{(3)}i^**^{(4)} \quad \text{(forward normal)},\end{aligned} \tag{11}$$

where $\rho_{(0)}$, $\rho_{(d)}:\mathcal{E}^{(1)}(\mathcal{M})\to\mathcal{E}^{(1)}(\Sigma)$, and $\rho_{(\delta)}$, $\rho_{(n)}:\mathcal{E}^{(1)}(\mathcal{M})\to\mathcal{E}^{(0)}(\Sigma)$. These operators can be applied to any smooth p-form. The motivation for (11) comes from an analysis of Green's identity for $\Box+m^2$ (Appendix A).

Next, let $\mathcal{A}\in\mathcal{E}^{(1)}(\mathcal{M})$ and define

$$A_{(0)}\equiv\rho_{(0)}\mathcal{A}, \tag{12}$$

$$A_{(d)}\equiv\rho_{(d)}\mathcal{A}, \tag{13}$$

$$A_{(n)}\equiv\rho_{(n)}\mathcal{A}, \tag{14}$$

and

$$A_{(\delta)}\equiv\rho_{(\delta)}\mathcal{A}, \tag{15}$$

with $A_{(0)}$, $A_{(d)}\in\mathcal{E}^{(1)}(\Sigma)$ and $A_{(n)}$, $A_{(\delta)}\in\mathcal{E}^{(0)}(\Sigma)$. Specifying $A_{(0)}$, $A_{(d)}$, $A_{(n)}$, and $A_{(\delta)}$ is equivalent to specifying the Cauchy data for $\mathcal{A}$. To see this, let (n,e_i) be a surface normal reference frame for (Σ,γ), then (12) and (14) specify $\mathcal{A}_\mu(0,x)$ from which we obtain $\nabla_k\mathcal{A}_\mu(0,x)$. Given $\nabla_k\mathcal{A}_\mu(0,x)$, (13) specifies $\nabla_0\mathcal{A}_k(0,x)$ and finally, (15) gives $\nabla_0\mathcal{A}_0(0,x)$. Thus, the operators (11) collectively map a solution $\mathcal{A}$ to its Cauchy data $(\mathcal{A}_\mu, n^\alpha\nabla_\alpha\mathcal{A}_\mu)$. Alternatively, we view $A_{(0)}$, $A_{(d)}$, $A_{(\delta)}$ and $A_{(n)}$ as Cauchy data for (4).

The operators (11) which are continuous and linear, give rise to continuous transpose operators $\rho'_{(0)}$, $\rho'_{(d)}:\mathcal{E}'^{(1)}(\Sigma)\to\mathcal{E}'^{(1)}(\mathcal{M})$, and $\rho'_{(n)}$, $\rho'_{(\delta)}:\mathcal{E}'^{(0)}(\Sigma)\to\mathcal{E}'^{(1)}(\mathcal{M})$.[17] We construct the following continuous operators $E_m^{(1)}\,\rho'_{(d)}$, $E_m^{(1)}\,\rho'_{(n)}$, $E_m^{(1)}\,\rho'_{(\delta)}$, and $E_m^{(1)}\,\rho'_{(0)}$, where

$$E_m^{(1)}\rho_{(0)}',E_m^{(1)}\rho_{(d)}':\mathfrak{D}^{(1)}(\Sigma)\subset\mathcal{E}'^{(1)}(\Sigma)\to\mathfrak{D}'^{(1)}(\mathcal{M}),\tag{16}$$

and

$$E_m^{(1)}\rho_{(\delta)}',E_m^{(1)}\rho_{(n)}':\mathfrak{D}^{(0)}(\Sigma)\subset\mathcal{E}'^{(0)}(\Sigma)\to\mathfrak{D}'^{(1)}(\mathcal{M}).\tag{17}$$

These play a crucial role in the classical Cauchy problem as is now shown.

Theorem 1: *Let $A_{(0)}$, $A_{(d)}\in\mathfrak{D}^{(1)}(\Sigma)$, and $A_{(n)}$, $A_{(\delta)}\in\mathfrak{D}^{(0)}(\Sigma)$ specify Cauchy data on (Σ,γ), and let $m\in[0,\infty)$. Then,*

$$\mathcal{A}'=-E_m^{(1)}\rho_{(d)}'A_{(0)}-E_m^{(1)}\rho_{(n)}'A_{(\delta)}+E_m^{(1)}\rho_{(\delta)}'A_{(n)}+E_m^{(1)}\rho_{(0)}'A_{(d)}\tag{18}$$

is the unique smooth solution of $(\Box+m^2)\mathcal{A}'=0$ with these data. Moreover, $\mathcal{A}'\in\mathcal{E}^{(1)}(\mathcal{M})$ is continuously dependent on the data.

Proof: First, from (16) and (17) we know that (18) makes sense. Let $\mathcal{F}\in\mathfrak{D}^{(1)}(\mathcal{M})$ be any 1-form test function, and consider

$$\langle\mathcal{A}',\mathcal{F}\rangle=\langle A_{(0)},\rho_{(d)}E_m^{(1)}\mathcal{F}\rangle_\Sigma+\langle A_{(\delta)},\rho_{(n)}E_m^{(1)}\mathcal{F}\rangle_\Sigma-\langle A_{(n)},\rho_{(\delta)}E_m^{(1)}\mathcal{F}\rangle_\Sigma-\langle A_{(d)},\rho_{(0)}E_m^{(1)}\mathcal{F}\rangle_\Sigma .$$

From Theorem 5 we know that if $\mathcal{A}$ is a smooth solution with data (12)–(15), then $\langle\mathcal{A}',\mathcal{F}\rangle=\langle\mathcal{A},\mathcal{F}\rangle_{\mathcal{M}}$ which implies that $\mathcal{A}'=\mathcal{A}$ in a distributional sense. Thus $\mathcal{A}'$ is identified with the unique smooth solution $\mathcal{A}\in\mathcal{E}^{(1)}(\mathcal{M})$.

It remains to show that $\mathcal{A}'$ is continuously dependent on the data. For this result, we assume that Σ is compact. From the first part of the proof we have $E_m^{(1)}\rho_{(0)}'$, $E_m^{(1)}\rho_{(d)}':\mathfrak{D}^{(1)}(\Sigma)\to\mathcal{E}^{(1)}(\mathcal{M})$ and $E_m^{(1)}\rho_{(\delta)}',E_m^{(1)}\rho_{(n)}':\mathfrak{D}^{(0)}(\Sigma)\to\mathcal{E}^{(1)}(\Sigma)$. It suffices to show that these restrictions are continuous. The same analysis applies to both sets of operators so we need only consider the former. Recall that $E_m^{(1)}\rho_{(0)}'$, $E_m^{(1)}\rho_{(d)}':\mathcal{E}'^{(1)}(\Sigma)\to\mathfrak{D}'^{(1)}(\mathcal{M})$ are continuous with respect to the weak topologies of dual spaces. Thus, the graphs of the restrictions $E_m^{(1)}\rho_{(0)}'$, $E_m^{(1)}\rho_{(d)}'$ $\mathfrak{D}^{(1)}(\Sigma)\to\mathcal{E}^{(1)}(\mathcal{M})$ are closed with respect to these weak topologies and it follows that they are closed with respect to the topologies of $\mathfrak{D}^{(1)}(\Sigma)$ and $\mathcal{E}^{(1)}(\mathcal{M})$ as well. Finally, since $\mathfrak{D}^{(1)}(\Sigma)$ and $\mathcal{E}^{(1)}(\mathcal{M})$ are Frechet spaces (assuming Σ is compact) it follows from the closed graph theorem that the restrictions of $E_m^{(1)}\rho_{(0)}'$ and $E_m^{(1)}\rho_{(d)}'$ are continuous. As noted, a similar analysis applies to the restrictions of $E_m^{(1)}\rho_{(\delta)}'$ and $E_m^{(1)}\rho_{(n)}'$ and therefore $\mathcal{A}'\in\mathcal{E}^{(1)}(\mathcal{M})$ is continuously dependent on the data. This presumably holds for noncompact Σ as well. ■

Thus, we obtain a global solution to (4) as a mapping of Cauchy data. Notice that Theorem 1 applies for $m=0$ which we use in our study of Maxwell's equations (Appendix B).

Corollary 1: On $\mathfrak{D}^{(1)}(\Sigma)$ we have

$$\rho_{(0)}E_m^{(1)}\rho_{(d)}'=-I,\quad \rho_{(d)}E_m^{(1)}\rho_{(d)}'=0,\quad \rho_{(\delta)}E_m^{(1)}\rho_{(d)}'=0,\quad \rho_{(n)}E_m^{(1)}\rho_{(d)}'=0,\tag{19}$$

$$\rho_{(0)}E_m^{(1)}\rho_{(0)}'=0,\quad \rho_{(d)}E_m^{(1)}\rho_{(0)}'=I,\quad \rho_{(\delta)}E_m^{(1)}\rho_{(0)}'=0,\quad \rho_{(n)}E_m^{(1)}\rho_{(0)}'=0,\tag{20}$$

and on $\mathfrak{D}^{(0)}(\Sigma)$ we have

$$\rho_{(0)}E_m^{(1)}\rho_{(n)}'=0,\quad \rho_{(d)}E_m^{(1)}\rho_{(n)}'=0,\quad \rho_{(\delta)}E_m^{(1)}\rho_{(n)}'=-I,\quad \rho_{(n)}E_m^{(1)}\rho_{(n)}'=0,\tag{21}$$

$$\rho_{(0)}E_m^{(1)}\rho_{(\delta)}'=0,\quad \rho_{(d)}E_m^{(1)}\rho_{(\delta)}'=0,\quad \rho_{(\delta)}E_m^{(1)}\rho_{(\delta)}'=0,\quad \rho_{(n)}E_m^{(1)}\rho_{(\delta)}'=I.\tag{22}$$

Proof: These identities follow from (18). For example, the first identity in (19) is obtained by applying $\rho_{(0)}$ to (18) with $A_{(\delta)}$, $A_{(n)}$, $A_{(d)}=0$. The remaining indentities are obtained in a similar fashion. □

We are finally ready for the main result of this section. Notice that Theorem (1) gives a global solution to (4). We apply it to Proca's equations and obtain a global solution to the initial value problem by restricting the data.

Theorem 2: *Let* $A_{(0)}$, $A_{(d)} \in \mathfrak{D}^{(1)}(\Sigma)$, *and* $A_{(n)}$, $A_{(\delta)} \in \mathfrak{D}^{(0)}(\Sigma)$ *specify Cauchy data on* Σ *and let* $m \in [0,\infty)$. *Set*

$$A_{(\delta)} = 0, \tag{23}$$

and

$$\delta^{(3)} A_{(d)} = m^2 A_{(n)} . \tag{24}$$

Then,

$$\mathcal{A} = -E_m^{(1)} \rho'_{(d)} A_{(0)} + E_m^{(1)} \rho'_{(\delta)} A_{(n)} + E_m^{(1)} \rho'_{(0)} A_{(d)} \tag{25}$$

is the unique smooth solution of $(\delta^{(4)} d^{(4)} + m^2)\mathcal{A} = 0$ *with these data. Moreover,* $\mathcal{A} \in \mathcal{E}^{(1)}(\mathcal{M})$ *is continuously dependent on the data. When* $m > 0$ (*Proca's equations) we satisfy (24) with arbitrary* $A_{(d)}$ *and* $A_{(n)} = [\delta^{(3)} A_{(d)} / m^2]$, *thus*

$$\mathcal{A} = -E_m^{(1)} \rho'_{(d)} A_{(0)} + E_m^{(1)} \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) \rho'_{(0)} A_{(d)} \quad (m > 0), \tag{26}$$

where we have used $\rho'_{(\delta)} \delta^{(3)} = d^{(4)} \delta^{(4)} \rho'_{(0)}$ *on* $\mathfrak{D}^{(1)}(\Sigma)$. *When* $m = 0$ *we satisfy (24) with arbitrary* $A_{(n)}$, *and with* $A_{(d)}$ *satisfying* $\delta^{(3)} A_{(d)} = 0$ (*see Maxwell's equations in Appendix B).*

Proof: We show that $\mathcal{A}$ is a smooth solution of $(\Box + m^2)\mathcal{A} = 0$ with $\delta^{(4)}\mathcal{A} = 0$. The former follows from Theorem 1. For the latter, it suffices to show that $\langle \delta^{(4)}\mathcal{A}, f\rangle_M = 0$, for all $f \in \mathfrak{D}^{(0)}(\mathcal{M})$. Consider,

$$\begin{aligned} \langle \delta^{(4)} \mathcal{A}, f \rangle_{\mathcal{M}} &= \langle \mathcal{A}, d^{(4)} f \rangle_{\mathcal{M}} , \\ &= \langle A_{(0)} , \rho_{(d)} E_m^{(1)} d^{(4)} f \rangle_\Sigma - \langle A_{(n)} , \rho_{(\delta)} E_m^{(1)} d^{(4)} f \rangle_\Sigma , \\ &\quad - \langle A_{(d)} , \rho_{(0)} E_m^{(1)} d^{(4)} f \rangle_\Sigma . \end{aligned} \tag{27}$$

Now, since $E_m^{(1)} d^{(4)} = d^{(4)} E_m^{(0)}$ on $\mathfrak{D}^{(0)}(\mathcal{M})$ we have $\rho_{(d)} E_m^{(1)} d^{(4)} f = 0$ and the first term on the right-hand side of (27) is zero. Consider the second term on the right-hand side of (27),

$$\begin{aligned} \rho_{(\delta)} E_m^{(1)} d^{(4)} f &= i^* \delta^{(4)} E_m^{(1)} d^{(4)} f, \\ &= i^* \delta^{(4)} d^{(4)} E_m^{(0)} f, \\ &= i^* (\Box - d^{(4)} \delta^{(4)}) E_m^{(0)} f, \\ &= -\rho_{(0)} m^2 E_m^{(0)} f, \end{aligned} \tag{28}$$

where we have used $(\Box + m^2) E_m^{(0)} f = 0$, and $\delta^{(4)} E_m^{(0)} f = 0$ ($E_m^{(0)} f$ is a function). Substituting (28) into (27) and making use of $\rho_{(0)} E_m^{(1)} d^{(4)} f = d^{(3)} \rho_{(0)} E_m^{(0)} f$ we have

$$\begin{aligned} \langle \delta^{(4)} \mathcal{A}, f \rangle_{\mathcal{M}} &= \langle A_{(n)} , \rho_{(0)} m^2 E_m^{(0)} f \rangle_\Sigma - \langle A_{(d)} , d^{(3)} \rho_{(0)} E_m^{(0)} f \rangle_\Sigma \\ &= \langle m^2 A_{(n)} - \delta^{(3)} A_{(d)} , \rho_{(0)} E_m^{(0)} f \rangle_\Sigma , \\ &= 0, \end{aligned}$$

where, in the last step we have used (24). Thus $\delta^{(4)}\mathcal{A}=0$ which is compatible with (23).

Lastly, from Theorem 1 we know that $\mathcal{A}$ is unique. We also have that $\mathcal{A}$ is continuously dependent on the data when Σ is compact. ■

From Theorem 2 we have the following additional result which is need for the quantum problem.

Corollary 2: The operator identity

$$\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)E_m^{(1)}=-E_m^{(1)}\rho'_{(d)}\rho_{(0)}\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)E_m^{(1)}+E_m^{(1)}\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)\rho'_{(0)}\rho_{(d)}E_m^{(1)}, \quad (29)$$

holds on $\mathfrak{D}^{(1)}(\mathcal{M})$ *for* $m>0$.

Proof: Let $\mathcal{F}\in\mathfrak{D}^{(1)}(\mathcal{M})$. It is easy to check that $(d^{(4)}\delta^{(4)}/m^2+I)E_m^{(1)}\mathcal{F}$ is a smooth solution to Proca's equations. Thus, (29) follows from Theorem 2 with $\mathcal{A}=[(d^{(4)}\delta^{(4)}/m^2)+I]E_m^{(1)}\mathcal{F}$ and the fact that $\rho_{(d)}[d^{(4)}\delta^{(4)}/m^2]E_m^{(1)}\mathcal{F}=0$. ■

This last result completes the prerequisite classical work. We proceed to the quantum problem.

IV. THE QUANTUM PROBLEM

Our approach to field quantization closely follows the work of Dimock.[3] We start with a representation $(\phi,\Pi,\mathcal{H})$ of the CCRs on an arbitrary Cauchy surface Σ. Let $\mathfrak{h}$ denote the completion of smooth complex-valued 1-forms on Σ with respect to the norm $\|\cdot\|_{\mathfrak{h}}^2=\langle\cdot,\cdot\rangle_{\mathfrak{h}}$, where

$$\langle F,G\rangle_{\mathfrak{h}}=\int(\overline{F},G)_x\,d\tau, \quad (30)$$

and

$$(F,G)_x=F_n(x)G^n(x), \quad (n=1,2,3) \quad (31)$$

with $d\tau=\sqrt{\gamma}dx^1\wedge dx^2\wedge dx^3$, where $x=(x^1,x^2,x^2)$ are local coordinates for (Σ,γ). Next, construct the Bose–Fock space $\mathcal{H}$ over $\mathfrak{h}$,

$$\mathcal{H}=\mathbb{C}\oplus\left(\bigoplus_{n=1}^{\infty}\mathfrak{h}^{(n)}\right), \quad (32)$$

where $\mathfrak{h}^{(n)}=\otimes_s^n\mathfrak{h}$, and the subscript s denotes the symmetric tensor product. Let $\mathbf{a}(\cdot)$, $\mathbf{a}^*(\cdot)$ denote the usual creation and annihilation operators defined on finite particle vectors in $\mathcal{H}$ with $[\mathbf{a}(F),\mathbf{a}^*(G)]=\langle F,G\rangle_{\mathfrak{h}}$.[18] Let

$$\phi(F)\equiv\frac{1}{\sqrt{2}}[\mathbf{a}(F)+\mathbf{a}^*(F)], \quad (33)$$

and

$$\Pi(F)\equiv\frac{i}{\sqrt{2}}[\mathbf{a}^*(F)-\mathbf{a}(F)], \quad (34)$$

and then take the closure of (33) and (34) (keeping the same notation) to obtain self-adjoint ϕ and Π on $\mathcal{H}$ with $[\phi(F),\Pi(G)]=i\langle F,G\rangle_{\mathfrak{h}}$ for $F,G\in\mathfrak{D}^{(1)}(\Sigma)$.[18] This gives the representation $(\phi,\Pi,\mathcal{H})$.

Now, given a representation $(\phi,\Pi,\mathcal{H})$ on Σ, not necessarily as above, we define a space–time field operator

$$\mathcal{A} \equiv -E_m^{(1)} \rho'_{(d)} \phi + E_m^{(1)} \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) \rho'_{(0)} \Pi , \tag{35}$$

in accordance with the classical initial value problem (Theorem 2 with $m > 0$). This holds in a distributional sense, therefore,

$$\mathcal{A}(\mathcal{F}) = \phi(\rho_{(d)} E_m^{(1)} \mathcal{F}) - \Pi \left(\rho_{(0)} \left[\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right] E_m^{(1)} \mathcal{F} \right)$$

for $\mathcal{F} \in \mathfrak{D}^{(1)}(\mathcal{M})$. This makes sense because $E_m^{(1)} \mathcal{F} \in \mathcal{E}^{(1)}(\mathcal{M})$ and $J^{\pm}(\text{supp}(\mathcal{F})) \cap \Sigma$ is compact.[3]

Theorem 3: *Let* $(\phi, \Pi, \mathcal{H})$ *be a representation of the CCRs over* $\mathfrak{D}^{(1)}(\Sigma)$, *and let* $\mathcal{A}$ *be the field operator (35) with test functions* $\mathcal{F} \in \mathfrak{D}^{(1)}(\mathcal{M})$. *Then* $\mathcal{A}$ *satisfies the Proca's equations in a distributional sense,*

$$(\delta^{(4)} d^{(4)} + m^2) \mathcal{A} = 0, \tag{36}$$

and

$$[\mathcal{A}(\mathcal{F}), \mathcal{A}(\mathcal{F}')] = -i \left\langle \mathcal{F}, \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) E_m^{(1)} \mathcal{F}' \right\rangle_{\mathcal{M}} . \tag{37}$$

Proof: First we verify (36). Consider,

$$\begin{aligned}
(\delta^{(4)} d^{(4)} + m^2) \mathcal{A}(\mathcal{F}) &= \mathcal{A}((\delta^{(4)} d^{(4)} + m^2) \mathcal{F}), \\
&= \phi(\rho_{(d)} E_m^{(1)} (\delta^{(4)} d^{(4)} + m^2) \mathcal{F}) \\
&\quad - \Pi \left(\rho_{(0)} \left[\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right] (\delta^{(4)} d^{(4)} + m^2) E_m^{(1)} \mathcal{F} \right), \\
&= -\phi(\rho_{(d)} d^{(4)} \delta^{(4)} E_m^{(1)} \mathcal{F}) - \Pi(\rho_{(0)} [\Box + m^2] E_m^{(1)} \mathcal{F}), \\
&= 0,
\end{aligned}$$

where we have used $\rho_{(d)} d^{(4)} = 0$ and $(\Box + m^2) E_m^{(1)} = 0$ on $\mathcal{E}^{(1)}(\mathcal{M})$.

Next, we verify (37). Consider,

$$\begin{aligned}
[\mathcal{A}(\mathcal{F}), \mathcal{A}(\mathcal{F}')] &= -i \left\langle \rho_{(d)} E_m^{(1)} \mathcal{F}, \rho_{(0)} \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) E_m^{(1)} \mathcal{F}' \right\rangle_{\Sigma} \\
&\quad + i \left\langle \rho_{(0)} \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) E_m^{(1)} \mathcal{F}, \rho_{(d)} E_m^{(1)} \mathcal{F}' \right\rangle_{\Sigma} , \\
&= i \left\langle \mathcal{F}, E_m^{(1)} \rho'_{(d)} \rho_{(0)} \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) E_m^{(1)} \mathcal{F}' \right\rangle_{\mathcal{M}} \\
&\quad - i \left\langle \mathcal{F}, E_m^{(1)} \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) \rho'_{(0)} \rho_{(d)} E_m^{(1)} \mathcal{F}' \right\rangle_{\mathcal{M}} , \\
&= -i \left\langle \mathcal{F}, \left(\frac{d^{(4)} \delta^{(4)}}{m^2} + I \right) E_m^{(1)} \mathcal{F}' \right\rangle_{\mathcal{M}} ,
\end{aligned}$$

where we have applied Corollary 2. ■

Notice from (7) that

$$\operatorname{supp}\left(\left[\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right]E_m^{(1)}\mathcal{F}'\right)\subset J^+(\operatorname{supp}(\mathcal{F}'))\cup J^-(\operatorname{supp}(\mathcal{F}')),$$

and therefore (37) implies causality. Also, (37) reduces to the usual relation on Minkowski space.[19]

We now show that a representation $(\phi,\Pi,\mathcal{H})$ on Σ induces a representation on any other Cauchy surface.

Corollary 3: Let $(\phi,\Pi,\mathcal{H})$ *be a representation of the CCRs over* $\mathfrak{D}^{(1)}(\Sigma)$, *and let* $\mathcal{A}$ *be the field operator (35). Let* $\hat{\Sigma}$ *be another Cauchy surface in* $\mathcal{M}$, *and let* $\hat{\phi}$ *and* $\hat{\Pi}$ *be data of* $\mathcal{A}$ *on* $\hat{\Sigma}$, *i.e.,*

$$\hat{\phi}\equiv\hat{\rho}_{(0)}\mathcal{A},$$

and,

$$\hat{\Pi}\equiv\hat{\rho}_{(d)}\mathcal{A}.$$

Then $(\hat{\phi},\hat{\Pi},\mathcal{H})$ *is a representation of the CCRs over* $\mathfrak{D}^{(1)}(\hat{\Sigma})$. *Furthermore, let*

$$\hat{\mathcal{A}}\equiv-E_m^{(1)}\hat{\rho}'_{(d)}\hat{\phi}+E_m^{(1)}\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)\hat{\rho}'_{(0)}\hat{\Pi},\tag{38}$$

then

$$\hat{\mathcal{A}}(\mathcal{F})=\mathcal{A}(\mathcal{F}).\tag{39}$$

Proof: We first show that $(\hat{\phi},\hat{\Pi},\mathcal{H})$ is a representation. Let $F\in\mathfrak{D}^{(1)}(\hat{\Sigma})$, then

$$\begin{aligned}\hat{\phi}(F)&=\mathcal{A}(\hat{\rho}'_{(0)}F),\\&=\phi(\rho_{(d)}E_m^{(1)}\hat{\rho}'_{(0)}F)-\Pi\left(\rho_{(0)}\left[\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right]E_m^{(1)}\hat{\rho}'_{(0)}F\right),\end{aligned}\tag{40}$$

and

$$\begin{aligned}\hat{\Pi}(F)&=\mathcal{A}(\hat{\rho}'_{(d)}F),\\&=\phi(\rho_{(d)}E_m^{(1)}\hat{\rho}'_{(d)}F)-\Pi\left(\rho_{(0)}\left[\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right]E_m^{(1)}\hat{\rho}'_{(d)}F\right).\end{aligned}\tag{41}$$

These make sense because $E_m^{(1)}\hat{\rho}'_{(0)},E_m^{(1)}\hat{\rho}'_{(d)}:\mathfrak{D}^{(1)}(\hat{\Sigma})\to\ \mathcal{E}^{(1)}(\mathcal{M})$ (Theorem 1) and $J^{\pm}(\operatorname{supp}(F))\cap\Sigma$ is compact.[3] Consider,

$$[\hat{\phi}(F),\hat{\Pi}(F')] = -\left[\phi(\rho_{(d)}E_m^{(1)}\hat{\rho}'_{(0)}F),\Pi\left(\rho_{(0)}\left[\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right]E_m^{(1)}\hat{\rho}'_{(d)}F'\right)\right]$$
$$+\left[\phi(\rho_{(d)}E_m^{(1)}\hat{\rho}'_{(d)}F'),\Pi\left(\rho_{(0)}\left[\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right]E_m^{(1)}\hat{\rho}'_{(0)}F\right)\right]$$
$$= -i\langle\rho_{(d)}\mathcal{F},\rho_{(0)}\mathcal{F}'\rangle_\Sigma + i\langle\rho_{(0)}\mathcal{F},\rho_{(d)}\mathcal{F}'\rangle_\Sigma$$
$$\equiv i\Omega_\Sigma(\mathcal{F},\mathcal{F}') \tag{42}$$

where $\mathcal{F}\equiv\{[d^{(4)}\delta^{(4)}/m^2]+I\}E_m^{(1)}\hat{\rho}'_{(0)}F$ and $\mathcal{F}'\equiv\{[d^{(4)}\delta^{(4)}/m^2]+I\}E_m^{(1)}\hat{\rho}'_{(d)}F'$ are smooth solutions of Proca's equations. It follows from Green's identity that $\Omega_\Sigma(\cdot,\cdot)$ is independent of the Cauchy surface for such solutions (see the vector potential formulation of Maxwell's equations).[2,4] Thus, (42) is independent of the Cauchy surface and we have

$$[\hat{\phi}(F),\hat{\Pi}(F')] = i\Omega_{\hat{\Sigma}}(\mathcal{F},\mathcal{F}')$$
$$= -i\langle\hat{\rho}_{(d)}\mathcal{F},\hat{\rho}_{(0)}\mathcal{F}'\rangle_{\hat{\Sigma}} \tag{43}$$
$$+ i\langle\hat{\rho}_{(0)}\mathcal{F},\hat{\rho}_{(d)}\mathcal{F}'\rangle_{\hat{\Sigma}}$$
$$= i\langle F,F'\rangle_{\hat{\Sigma}}, \tag{44}$$

where in the last step we have applied Corollary (1) to $\hat{\Sigma}$. Thus $(\hat{\phi},\hat{\Pi},\mathcal{H})$ is a representation.

We now verify (39). Substitute (40) and (41) into (38) and obtain

$$\hat{\mathcal{A}}(\mathcal{F}) \equiv \phi(\rho_{(d)}\hat{\mathcal{F}}) - \Pi\left(\rho_{(0)}\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)\hat{\mathcal{F}}\right),$$
$$= \phi(\rho_{(d)}\tilde{\mathcal{F}}) - \Pi\left(\rho_{(0)}\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)\tilde{\mathcal{F}}\right)$$
$$= \phi(\rho_{(d)}E_m^{(1)}\mathcal{F}) - \Pi\left(\rho_{(0)}\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)E_m^{(1)}\mathcal{F}\right)$$
$$= \mathcal{A}(\mathcal{F}),$$

where

$$\hat{\mathcal{F}} = \left[E_m^{(1)}\hat{\rho}'_{(0)}\hat{\rho}_{(d)}E_m^{(1)} - \hat{\rho}'_{(d)}\hat{\rho}_{(0)}\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)E_m^{(1)}\right]\mathcal{F}$$

and

$$\tilde{\mathcal{F}} = \left[\left(\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right)E_m^{(1)} - \frac{d^{(4)}\delta^{(4)}}{m^2}E_m^{(1)}\hat{\rho}'_{(0)}\hat{\rho}_{(d)}E_m^{(1)}\right]\mathcal{F}.$$

In the second step we have used Corollary 2 and in subsequent steps we have used $\rho_{(d)}d^{(4)}=0$ on $\mathcal{E}^{(1)}(\mathcal{M})$ and $\delta^{(4)}\mathcal{G}=0$ for any smooth solution $\mathcal{G}$ of Proca's equations. ■

At this point we have a field operator $\mathcal{A}$ defined in terms of a representation $(\phi,\Pi,\mathcal{H})$ on an arbitrary Cauchy surface Σ. Unlike Minkowski space–time, there is, in general, no preferred representation (i.e., no natural positive and negative frequency decomposition).[2] Thus, we must consider a class of representations, and show that the theory is independent of the representation, and independent of the Cauchy surface. To this end, we pass to the Weyl form of the CCRs, construct an algebra of observables $\mathfrak{A}$, and exploit known results on the equivalence of such algebras.

Given a representation $(\phi,\Pi,\mathcal{H})$ let

$$W(F,F')=\exp(i(\phi(F)-\Pi(F'))).$$

Then $W(\cdot,\cdot)$is a map from $\mathfrak{D}^{(1)}(\Sigma)\times\mathfrak{D}^{(1)}(\Sigma)$ to unitary operators on $\mathcal{H}$ satisfying

$$W(F,F')W(G,G')=W(F+G,F'+G')\exp(-\,i/2\,\sigma_{\Sigma}((F,F'),(G,G'))), \tag{45}$$

where

$$\sigma_{\Sigma}((F,F'),(G,G'))\equiv\langle F,G'\rangle_{\Sigma}-\langle F',G\rangle_{\Sigma}, \tag{46}$$

is symplectic on $\mathfrak{D}^{(1)}(\Sigma)\times\mathfrak{D}^{(1)}(\Sigma)$. Also, $t\rightarrow W(tF,tF')$ is strongly continuous. Thus we have the Weyl system $(W,\mathcal{H})$ with the Weyl form of the CCRs (45). As an aside, note that $\Omega_{\Sigma}(\mathcal{F},\mathcal{F}')=\sigma_{\Sigma}((\rho_{(0)}\mathcal{F},\rho_{(d)}\mathcal{F}),(\rho_{(0)}\mathcal{F}',\rho_{(d)}\mathcal{F}'))$ for smooth solutions $\mathcal{F}$ and $\mathcal{F}'$ of Proca's equations.

Alternatively, given a representation $(W,\mathcal{H})$ we recover self-adjoint ϕ and Π via Stone's Theorem, i.e.,

$$e^{i\phi(F)t}=W(tF,0), \tag{47}$$

and

$$e^{-i\mathbf{\Pi}(F)t}=W(0,tF). \tag{48}$$

We also obtain a self-adjoint field operator $\mathcal{A}(\mathcal{F})$ via Stone's Theorem,

$$e^{i\mathcal{A}(\mathcal{F})t}=W\left(t\rho_{(d)}E_m^{(1)}\mathcal{F},t\rho_{(0)}\left[\frac{d^{(4)}\delta^{(4)}}{m^2}+I\right]E_m^{(1)}\mathcal{F}\right), \tag{49}$$

where $\mathcal{F}\in\mathfrak{D}^{(1)}(\mathcal{M})$.

Next, we define an algebra of observables $\mathfrak{A}$. Given a Weyl system $(W,\mathcal{H})$ take the set of all finite sums of the form

$$\sum_{\alpha} c_{\alpha}W(F_{\alpha},F'_{\alpha}),\quad c_{\alpha}\in\mathbb{C},$$

where $F_{\alpha},F'_{\alpha}\in\mathfrak{D}^{(1)}(\Sigma)$ and define $\mathfrak{A}$ to be the norm closure of this set in the Banach space of all bounded operators on $\mathcal{H}$.

We now exploit known results on the equivalence of such algebras.

Theorem 4: *Let* $(W,\sigma_{\Sigma},\mathfrak{A},\mathcal{H})$ *and* $(\tilde{W},\sigma_{\hat{\Sigma}},\tilde{\mathfrak{A}},\tilde{\mathcal{H}})$ *be representations on Cauchy surfaces* Σ *and* $\hat{\Sigma}$, *respectively. There is a unique* $*$*-isomorphism* $\alpha:\mathfrak{A}\rightarrow\tilde{\mathfrak{A}}$ *with* $\alpha:e^{i\mathcal{A}(\mathcal{F})}\rightarrow e^{i\tilde{\mathcal{A}}(\mathcal{F})}$.

Proof: First, consider the case $\Sigma=\hat{\Sigma}$. Given $(W,\sigma_{\Sigma},\mathfrak{A},\mathcal{H})$ and $(\tilde{W},\sigma_{\Sigma},\tilde{\mathfrak{A}},\tilde{\mathcal{H}})$ over $\mathfrak{D}^{(1)}(\Sigma)\times\mathfrak{D}^{(1)}(\Sigma)$ there is a unique $*-$isomorphism $\alpha:\mathfrak{A}\rightarrow\tilde{\mathfrak{A}}$ such that $\alpha(W(F,F'))=\tilde{W}(F,F')$ (see Theorem 5.2.8 in 18). It follows from (49) that $\alpha:e^{i\mathcal{A}(\mathcal{F})}\rightarrow e^{i\tilde{\mathcal{A}}(\mathcal{F})}$. Thus $\mathfrak{A}$ is independent of the representation on Σ in this sense.

Next, let Σ and $\hat{\Sigma}$ be different, and let $(\phi,\Pi,\mathcal{H})$ and $(\hat{\phi},\hat{\Pi},\mathcal{H})$ be respective representations as defined in Corollary 3. It follows that $e^{i\mathcal{A}(\mathcal{F})}=e^{i\hat{\mathcal{A}}(\mathcal{F})}$ and therefore $\mathfrak{A}=\hat{\mathfrak{A}}$. Moreover, from the first part of the proof we have $\alpha:\hat{\mathfrak{A}}\rightarrow\tilde{\mathfrak{A}}$ and therefore $\alpha:\mathfrak{A}\rightarrow\tilde{\mathfrak{A}}$ with $\alpha:e^{i\mathcal{A}(\mathcal{F})}\rightarrow e^{i\tilde{\mathcal{A}}(\mathcal{F})}$. Thus $\mathfrak{A}$ is independent of the Cauchy surface in this sense. ∎

From this last result, we see that the Fock representation $(\phi,\Pi,\mathcal{H})$ defined by (33) and (34) gives rise to an algebra of observables $\mathfrak{A}$ that is unique up to $*-$isomorphism. This concludes the quantum problem.

V. CONCLUSION

We have obtained classical and quantum results for the propagation of massive vector fields on a globally hyperbolic Lorentzian manifold. Our classical results include solutions of the initial value problem for Proca's equations (3), the vector Klein Gordon equation (4), and Maxwell's equations (Appendix B). The form of these solutions is apparently new and useful for field quantization. Our quantum results include a causal field operator constructed from a representation of the CCRs on an arbitrary Cauchy surface. The algebra of observables generated by this field operator is independent of the representation and independent of the Cauchy surface.

ACKNOWLEDGMENT

The author would like to thank Professor J. Dimock for numerous helpful discussions.

APPENDIX A: GREEN'S IDENTITY

In this section we develop Green's identity for $\Box + m^2$. Start with Stoke's Theorem,

$$\int_{\mathcal{O}} d^{(4)}\mathcal{G} = \int_{\partial\mathcal{O}} i^*\mathcal{G}, \tag{A1}$$

where $\mathcal{O}\subset\mathcal{M}$, $\partial\mathcal{O}$ is the boundary of $\mathcal{O}$, $i:\partial\mathcal{O}\to\mathcal{O}$ is the natural inclusion, i^* is the pullback, and $\mathcal{G}\in\mathfrak{D}^{(3)}(\mathcal{M})$.[20–22] Let $\mathcal{G}^{(1)} = \delta^{(4)}\mathcal{F}\wedge *^{(4)}\mathcal{A} - \delta^{(4)}\mathcal{A}\wedge *^{(4)}\mathcal{F}$ and $\mathcal{G}^{(2)} = \mathcal{A}\wedge *^{(4)}d^{(4)}\mathcal{F} - \mathcal{F}\wedge *^{(4)}d^{(4)}\mathcal{A}$, with $\mathcal{A}$, $\mathcal{F}\in\mathfrak{D}^{(1)}(\mathcal{M})$. We apply (A1) with $\mathcal{G}=\mathcal{G}^{(1)}+\mathcal{G}^{(2)}$ and obtain

$$\int_{\mathcal{O}} \mathcal{A}\wedge *^{(4)}\Box\mathcal{F} - \mathcal{F}\wedge *^{(4)}\Box\mathcal{A} = -\int_{\partial\mathcal{O}} i^*(\mathcal{A}\wedge *^{(4)}d^{(4)}\mathcal{F} + \delta^{(4)}\mathcal{A}\wedge *^{(4)}\mathcal{F}) + \int_{\partial\mathcal{O}} i^*(\delta^{(4)}\mathcal{F}\wedge *^{(4)}\mathcal{A} + \mathcal{F}\wedge *^{(4)}d^{(4)}\mathcal{A}). \tag{A2}$$

Finally, add $\mathcal{A}\wedge *^{(4)}m^2\mathcal{F} - \mathcal{F}\wedge *^{(4)}m^2\mathcal{A}$ to the left-hand side of (A2) and obtain Green's identity for $(\Box + m^2)$,

$$\int_{\mathcal{O}} \mathcal{A}\wedge *^{(4)}(\Box + m^2)\mathcal{F} - \mathcal{F}\wedge *^{(4)}(\Box + m^2)\mathcal{A} = -\int_{\partial\mathcal{O}} i^*(\mathcal{A}\wedge *^{(4)}d^{(4)}\mathcal{F} + \delta^{(4)}\mathcal{A}\wedge *^{(4)}\mathcal{F}) + \int_{\partial\mathcal{O}} i^*(\delta^{(4)}\mathcal{F}\wedge *^{(4)}\mathcal{A} + \mathcal{F}\wedge *^{(4)}d^{(4)}\mathcal{A}). \tag{A3}$$

Next, apply (A3) to the regions $\mathcal{O}=\Sigma^{\pm}\equiv J^{\pm}(\Sigma)\setminus\Sigma$, with $\partial\mathcal{O}=\Sigma$ and obtain

$$\int_{\Sigma^{\pm}} \mathcal{A}\wedge *^{(4)}(\Box + m^2)\mathcal{F} - \mathcal{F}\wedge *^{(4)}(\Box + m^2)\mathcal{A} = \mp\{\langle\rho_{(0)}\mathcal{A},\rho_{(d)}\mathcal{F}\rangle + \langle\rho_{(\delta)}\mathcal{A},\rho_{(n)}\mathcal{F}\rangle_{\Sigma} - \langle\rho_{(n)}\mathcal{A},\rho_{(\delta)}\mathcal{F}\rangle_{\Sigma} - \langle\rho_{(d)}\mathcal{A},\rho_{(0)}\mathcal{F}\rangle_{\Sigma}, \tag{A4}$$

where $\rho_{(0)}$, $\rho_{(n)}$, $\rho_{(d)}$, and $\rho_{(\delta)}$ are as defined in (11), and, for example, $\int_{\Sigma} i^*\delta^{(4)}\mathcal{F}\wedge *^{(3)}(-*^{(3)}i^**^{(4)})\mathcal{A} = \langle\rho_{(\delta)}\mathcal{F},\rho_{(n)}\mathcal{A}\rangle_{\Sigma}$ (the standard orientation is used for both regions $\Sigma^{\pm}$).

The action of $\rho_{(0)}$ and $\rho_{(\delta)}$ is obvious, however, $\rho_{(n)}$ and $\rho_{(d)}$ are more subtle. Specifically, $\rho_{(n)}$ and $\rho_{(d)}$ are the forward normal, and pullback of the forward normal derivative operators, respectively. To see this, let $x^{\mu}=(t,x^{i})$, then $e_{\mu}=\partial_{x^{\mu}}$ represents the standard basis for the tangent space of $(\mathcal{M},g)$. Let (Σ,γ) be a Cauchy surface with forward normal n. Following the presentation of Misner, Thorne and Wheeler we have $n=n^{\mu}e_{\mu}$,

$$n=N^{-1}(\partial_t-N^{i}\partial_{x^{i}})$$

where $N=(-g^{00})^{-(1/2)}$ is the lapse function, and $N^{i}=g^{ik}g_{ok}$ are the components of the shift vector.[14] Notice that $n=-N\ dt$ in covariant form, and recall that $\sqrt{|g|}=N\sqrt{\gamma}$. Now, let $\mathcal{A}=\mathcal{A}_{\mu}dx^{\mu}$ and consider,

$$\begin{aligned}\rho_{(n)}\mathcal{A}&=-*^{(3)}i**^{(4)}\mathcal{A}_{\mu}dx^{\mu},\\&=-*^{(3)}i*\frac{1}{3!}\sqrt{|g|}\mathcal{A}^{\mu}\epsilon_{\mu\alpha\beta\eta}dx^{\alpha}\wedge dx^{\beta}\wedge dx^{\eta},\\&=-*^{(3)}\frac{1}{3!}N\sqrt{\gamma}\mathcal{A}^{0}\epsilon_{0ijk}dx^{i}\wedge dx^{j}\wedge dx^{k},\\&=-\frac{1}{3!}N\gamma\mathcal{A}^{0}\epsilon_{0lmn}\epsilon_{ijk}\gamma^{il}\gamma^{jm}\gamma^{kn},\\&=-N\mathcal{A}^{0},\\&=\mathcal{A}(n),\end{aligned}$$

where $\epsilon_{\mu\alpha\beta\eta}$ is the Levi Cevita symbol.[14]

For the analysis of $\rho_{(d)}$ it is convenient to work with a basis $\hat{e}_{\mu}$ where $\hat{e}_0=n$ and $\hat{e}_i=\partial_{x^{i}}$. The dual for this basis is $\hat{\omega}^{\mu}$ where $\hat{\omega}^{0}=Ndt$ and $\hat{\omega}^{i}=dx^{i}+N^{i}dt$.[14] We also have $|\hat{g}|=\gamma$. Let $\mathcal{A}=\hat{\mathcal{A}}_{\mu}\hat{\omega}^{\mu}$, and $\mathcal{F}=d\mathcal{A}=(1/2!)\hat{\mathcal{F}}_{\mu\nu}\hat{\omega}^{\mu}\wedge\hat{\omega}^{\nu}$. It follows that $\hat{\mathcal{F}}_{0s}=\mathcal{F}(\hat{e}_0,\hat{e}_s)=\mathcal{F}(n^{\mu}e_{\mu},e_s)=n^{\mu}\mathcal{F}_{\mu s}$ and therefore,

$$n^{\mu}\mathcal{F}_{\mu s}dx^{s}=\rho_{(0)}\mathcal{F}(n,\cdot). \tag{A5}$$

Consider,

$$\begin{aligned}\rho_{(d)}\mathcal{A}&=-*^{(3)}i**^{(4)}d^{(4)}\hat{\mathcal{A}}_{\mu}\hat{\omega}^{\mu},\\&=-*^{(3)}i**^{(4)}\frac{1}{2!}\hat{\mathcal{F}}_{\mu\nu}\hat{\omega}^{\mu}\wedge\hat{\omega}^{\nu},\\&=-*^{(3)}i*\frac{1}{2!}\sqrt{|\hat{g}|}\frac{1}{2!}\hat{g}^{\mu\alpha}\hat{g}^{\nu\beta}\hat{\mathcal{F}}_{\alpha\beta}\epsilon_{\mu\nu\sigma\tau}\hat{\omega}^{\sigma}\wedge\hat{\omega}^{\tau},\\&=-*^{(3)}\sqrt{\gamma}\hat{g}^{0\alpha}\hat{g}^{i\beta}\hat{\mathcal{F}}_{\alpha\beta}\epsilon_{0ijk}dx^{j}\wedge dx^{k},\\&=\gamma\frac{1}{2!}\hat{\mathcal{F}}_{0s}\gamma^{is}\gamma^{jm}\gamma^{kn}\epsilon_{0imn}\epsilon_{jkp}dx^{p},\\&=\frac{1}{2!}\hat{\mathcal{F}}_{0s}\epsilon_{sjk}\epsilon_{jkp}dx^{p},\\&=\hat{\mathcal{F}}_{0s}dx^{s},\\&=n^{\alpha}\mathcal{F}_{\alpha s}dx^{s}\\&=\rho_{(0)}(d^{(4)}\mathcal{A})(n,\cdot),\end{aligned}$$

where in the last step we have used (A5).

We are finally ready to prove

Theorem 5: *Let* $\mathcal{A}$ *be a smooth solution of* $(\Box+m^2)\mathcal{A}=0$ *with Cauchy data* $A_{(0)}\equiv\rho_{(0)}\mathcal{A}$, $A_{(d)}\equiv\rho_{(d)}\mathcal{A}\in\mathfrak{D}^{(1)}(\Sigma)$, *and* $A_{(n)}\equiv\rho_{(n)}\mathcal{A}$, $A_{(\delta)}\equiv\rho_{(\delta)}\mathcal{A}\in\mathfrak{D}^{(0)}(\Sigma)$. *Then,*

$$\langle\mathcal{A},\mathcal{F}\rangle_{\mathcal{M}}=\langle A_{(0)},\rho_{(d)}E_m^{(1)}\mathcal{F}\rangle_\Sigma+\langle A_{(\delta)},\rho_{(n)}E_m^{(1)}\mathcal{F}\rangle_\Sigma-\langle A_{(n)},\rho_{(\delta)}E_m^{(1)}\mathcal{F}\rangle_\Sigma-\langle A_{(d)},\rho_{(0)}E_m^{(1)}\mathcal{F}\rangle_\Sigma, \tag{A6}$$

for any test function $\mathcal{F}\in\mathfrak{D}^{(1)}(\mathcal{M})$.

Proof: Since $(\Box+m^2)\mathcal{A}=0$, the second term on the left-hand side of (A4) is zero. For the remaining integrals, substitute $\mathcal{F}=E_m^{\mp(1)}\mathcal{F}'$ for the regions $\Sigma^\pm$, respectively. All integrals are well-defined because they entail integrations of smooth functions over compact sets. Specifically, for the left-hand side of (A4) we have

$$\mathrm{supp}(E_m^{\mp(1)}\mathcal{F}')\subset J^{\mp}(\mathrm{supp}(\mathcal{F}')),$$

with $J^{\mp}(\mathrm{supp}(\mathcal{F}'))\cap J^{\pm}(\Sigma)$ compact ($\mathcal{M}$ is globally hyperbolic), and for the right-hand side, Σ is compact by assumption.[3] Next, sum the integrations over the $\Sigma^\pm$ regions substituting $(\Box+m^2)E_m^{\pm(1)}=I$ in $\Sigma^\pm$ integrals and $E_m^{(1)}=E_m^{+(1)}-E_m^{-(1)}$ in Σ integrals. The sum of the $\Sigma^\pm$ integrals gives an integral over $\mathcal{M}$ (Σ constitutes a set of measure zero relative to this integration). Finally, relabel $\mathcal{F}'\rightarrow\mathcal{F}$ and obtain (A6). This completes the proof. ■

APPENDIX B: MAXWELL'S EQUATIONS

In this section we study Maxwell's equations,

$$d^{(4)}\mathcal{F}=0 \tag{B1}$$

and

$$\delta^{(4)}\mathcal{F}=0, \tag{B2}$$

where $\mathcal{F}$ is the field strength 2-form (not a test function as above). We pose an initial value problem for these equations following the presentation in Wald.[13] Specifically, we specify the initial data for $\mathcal{F}$ in terms of the electric and magnetic fields $E\equiv\rho_{(n)}\mathcal{F}$ and $B\equiv\rho_{(0)}\mathcal{F}$, respectively, which are regarded as a 1-form and 2-form on Σ, respectively. This data satisfies additional constraints $\delta^{(3)}E=0$ and $d^{(3)}B=0$. Given these data, we obtain a field solution $\mathcal{F}$ that satisfies (B1) and (B2) on $\mathcal{M}$. Our approach is similar to Dimock's; the differences being that our emphasis is on the fields rather than the vector potential, and we give an explicit representation for $\mathcal{F}$.[4]

Before we proceed, we make a further restriction on $\mathcal{M}$. Specifically, we assume that Σ is compact and contractible. If Σ is contractible then any closed p-form on Σ is exact, that is, $K\in\mathfrak{D}^{(p)}(\Sigma)$ with $d^{(3)}K=0\Rightarrow K=d^{(3)}H$ for some $H\in\mathfrak{D}^{(p-1)}(\Sigma)$ $(p>0)$.[20,21]

Theorem 6: *Let* $E\in\mathfrak{D}^{(1)}(\Sigma)$ *and* $B\in\mathfrak{D}^{(2)}(\Sigma)$ *be data for the field strength* $\mathcal{F}$, *i.e.,*

$$\rho_{(n)}\mathcal{F}=E, \tag{B3}$$

and

$$\rho_{(0)}\mathcal{F}=B, \tag{B4}$$

where

$$\delta^{(3)}E=0, \tag{B5}$$

and

$$d^{(3)}B=0. \tag{B6}$$

Then, given these data, there is a smooth potential $\mathcal{A}$ such that $\mathcal{F}=d^{(4)}\mathcal{A}$ satisfies Maxwell's equations $d^{(4)}\mathcal{F}=0$ and $\delta^{(4)}\mathcal{F}=0$, as well as (B3)–(B6). Moreover, any two such potentials are gauge equivalent i.e., they differ by the exterior derivative of a scalar. Here we assume that Σ is compact and contractible.

Proof: Existence: We choose data for $\mathcal{A}$ as follows:

$$A_{(d)}=E, \tag{B7}$$

$$d^{(3)}A_{(0)}=B, \tag{B8}$$

with

$$\delta^{(3)}A_{(d)}=0, \tag{B9}$$

$A_{(\delta)}=0$, and $A_{(n)}$ arbitrary. These choices of data are compatible with the field constraints (B5), and (B6) in that (B7) and (B9) imply (B5), and (B8) implies (B6). Regarding (B8), we know that such an $A_{(0)}$ exists because Σ is contractible and B is exact. These choices of data satisfy (23) and (24) of Theorem 2 for the $m=0$ case. Consequently, there is a unique smooth $\mathcal{A}$ that satisfies

$$\delta^{(4)}d^{(4)}\mathcal{A}=0, \tag{B10}$$

and $\mathcal{F}=d^{(4)}\mathcal{A}$ is our desired field strength. Next we show that $\mathcal{F}$ renders the data (B3) and (B4). From Theorem 2 we have

$$\mathcal{A}=-E_0^{(1)}\rho'_{(d)}A_{(0)}+E_0^{(1)}\rho'_{(\delta)}A_{(n)}+E_0^{(1)}\rho'_{(0)}A_{(d)}. \tag{B11}$$

Consider,

$$\begin{aligned}\rho_{(n)}\mathcal{F}&=-\rho_{(n)}d^{(4)}E_0^{(1)}\rho'_{(d)}A_{(0)}+\rho_{(n)}d^{(4)}E_0^{(1)}\rho'_{(\delta)}A_{(n)}+\rho_{(n)}d^{(4)}E_0^{(1)}\rho'_{(0)}A_{(d)},\\ &=-\rho_{(d)}E_0^{(1)}\rho'_{(d)}A_{(0)}+\rho_{(d)}E_0^{(1)}\rho'_{(\delta)}A_{(n)}+\rho_{(d)}E_0^{(1)}\rho'_{(0)}A_{(d)},\\ &=A_{(d)},\\ &=E,\end{aligned}$$

and

$$\begin{aligned}\rho_{(0)}\mathcal{F}&=-\rho_{(0)}d^{(4)}E_0^{(1)}\rho'_{(d)}A_{(0)}+\rho_{(0)}d^{(4)}E_0^{(1)}\rho'_{(\delta)}A_{(n)}+\rho_{(0)}d^{(4)}E_0^{(1)}\rho'_{(0)}A_{(d)},\\ &=-d^{(3)}\rho_{(0)}E_0^{(1)}\rho'_{(d)}A_{(0)}+d^{(3)}\rho_{(0)}E_0^{(1)}\rho'_{(\delta)}A_{(n)}+d^{(3)}\rho_{(0)}E_0^{(1)}\rho'_{(0)}A_{(d)},\\ &=d^{(3)}A_{(0)},\\ &=B,\end{aligned}$$

where we have used the results of Corollary 1.

Uniqueness: We want to show that any two potentials $\mathcal{A}'$ and $\tilde{\mathcal{A}}$ that satisfy (B10) with data (B7)–(B9) are gauge equivalent, i.e.,

$$\mathcal{A}'=\tilde{\mathcal{A}}+d^{(4)}f,$$

where $f\in\mathfrak{D}^{(0)}(\mathcal{M})$.[4] It suffices to show that any solution is gauge equivalent to the unique solution $\mathcal{A}$ above. Consider a solution $\mathcal{A}'$, we want to show that f' exists such that

$$\mathcal{A}=\mathcal{A}'+d^{(4)}f'.$$

Since $\mathcal{A}$ is unique, it suffices to construct $f'\in\mathfrak{D}^{(0)}(\mathcal{M})$ such that $\mathcal{A}'+d^{(4)}f'$ satisfies

$$\Box(\mathcal{A}'+d^{(4)}f')=0 \tag{B12}$$

on $\mathcal{M}$ with the same data as $\mathcal{A}$ on Σ. First, notice that (B12) is equivalent to

$$\delta^{(4)}d^{(4)}\mathcal{A}'+d^{(4)}\delta^{(4)}\mathcal{A}'+d^{(4)}\delta^{(4)}d^{(4)}f'=0. \tag{B13}$$

The first term in (B13) is zero because $\mathcal{A}'$ satisfies (B10). Therefore (B12) reduces to

$$\Box f'=-\delta^{(4)}\mathcal{A}'. \tag{B14}$$

Next, we study the data. Recall from Theorem 2 that the unique solution $\mathcal{A}$ has data $A_{(\delta)}$, $A_{(n)}$, $A_{(0)}$, and $A_{(d)}$, where

$$A_{(\delta)}=0, \tag{B15}$$

$$\delta^{(3)}A_{(d)}=0, \tag{B16}$$

and $A_{(0)}$ and $A_{(n)}$ are arbitrary. We need to construct f' so that $\mathcal{A}'+d^{(4)}f'$ renders the same data. To this end, we specify

$$\rho_{(\delta)}(\mathcal{A}'+d^{(4)}f')=0, \tag{B17}$$

$$\rho_{(n)}\mathcal{A}=\rho_{(n)}(\mathcal{A}'+d^{(4)}f'), \tag{B18}$$

$$\rho_{(0)}\mathcal{A}=\rho_{(0)}(\mathcal{A}'+d^{(4)}f'), \tag{B19}$$

and

$$\rho_{(d)}\mathcal{A}=\rho_{(d)}(\mathcal{A}'+d^{(4)}f'). \tag{B20}$$

To satisfy (B17) we impose (B14). The conditions (B18) and (B19) are satisfied when

$$\rho_{(d)}f'=\rho_{(n)}(\mathcal{A}-\mathcal{A}'), \tag{B21}$$

and

$$d^{(3)}\rho_{(0)}f'=\rho_{(0)}(\mathcal{A}-\mathcal{A}'), \tag{B22}$$

respectively. Regarding (B22), we know that such a $\rho_{(0)}f'$ exists because Σ is contractible and, by assumption, $\rho_{(0)}(\mathcal{A}-\mathcal{A}')$ is exact, i.e., $d^{(3)}\rho_{(0)}(\mathcal{A}-\mathcal{A}')=B-B=0$. Finally, (B20) is satisfied because, by assumption, $\mathcal{A}$ and $\mathcal{A}'$ satisfy (B7), and we know that $\rho_{(d)}d^{(4)}f'=0$.

Now, by assumption, we are given $\mathcal{A}$ and $\mathcal{A}'$ so we view (B14), (B21) and (B22) as specifying a Cauchy problem for the scalar field f'. That is, (B21) and (B22) specify the Cauchy data $\rho_{(d)}f'$ and $\rho_{(0)}f'$ for the nonhomogeneous linear hyperbolic equation (B14). A unique solution to this problem is known to exist which gives us the desired f'.[13] This shows that any solution $\mathcal{A}'$ is gauge equivalent to the unique $\mathcal{A}$ and therefore, given two different solutions $\mathcal{A}'$ and $\tilde{\mathcal{A}}$ we have $\mathcal{A}=\mathcal{A}'+d^{(4)}f'$, and $\mathcal{A}=\tilde{\mathcal{A}}+d^{(4)}\tilde{f}$ which shows that $\mathcal{A}'=\tilde{\mathcal{A}}+d^{(4)}f$, where $f=\tilde{f}-f'$. Thus any two such solutions give rise to the same field strength $\mathcal{F}$. This completes the proof. ■

We obtain an explicit expression for $\mathcal{F}$ as follows:

Corollary 4: Let $\mathcal{A}$ be a vector potential with data $A_{(0)}$, $A_{(d)}$, $A_{(n)}$, and $A_{(\delta)}$ satisfying the conditions of Theorem 6. The field strength is given by

$$\mathcal{F} = -d^{(4)} E_0^{(1)} \rho'_{(d)} A_{(0)} + d^{(4)} E_0^{(1)} \rho'_{(0)} A_{(d)}, \tag{B23}$$

where $\mathcal{F} \in \mathcal{E}^{(2)}(\mathcal{M})$ *is continuously dependent on* $A_{(0)}$, $A_{(d)} \in \mathfrak{D}^{(1)}(\Sigma)$.

Proof: From Theorem 2 we have

$$\mathcal{A} = -E_0^{(1)} \rho'_{(d)} A_{(0)} + E_0^{(1)} \rho'_{(\delta)} A_{(n)} + E_0^{(1)} \rho'_{(0)} A_{(d)} \tag{B24}$$

with

$$\delta^{(3)} A_{(d)} = 0. \tag{B25}$$

We show that (B23) equals $d^{(4)}\mathcal{A}$. Let $\mathcal{G} \in \mathfrak{D}^{(2)}(\mathcal{M})$ be a 2-form test function, and consider,

$$\begin{aligned}
\langle d^{(4)}\mathcal{A}, \mathcal{G}\rangle_{\mathcal{M}} &= \langle \mathcal{A}, \delta^{(4)}\mathcal{G}\rangle_{\mathcal{M}}, \\
&= \langle A_{(0)}, \rho_{(d)} E_0^{(1)} \delta^{(4)}\mathcal{G}\rangle_{\Sigma} - \langle A_{(n)}, \rho_{(\delta)} E_0^{(1)} \delta^{(4)}\mathcal{G}\rangle_{\Sigma} \\
&\quad - \langle A_{(d)}, \rho_{(0)} E_0^{(1)} \delta^{(4)}\mathcal{G}\rangle_{\Sigma} \\
&= \langle -d^{(4)} E_0^{(1)} \rho'_{(d)} A_{(0)} + d^{(4)} E_0^{(1)} \rho'_{(0)} A_{(d)}, \mathcal{G}\rangle_{\Sigma},
\end{aligned}$$

where, in the last step we have used $\rho_{(\delta)} E_0^{(1)} \delta^{(4)}\mathcal{G} = 0$. Thus, (B23) is satisfied in a distributional sense. From Theorem 1 we know that $E_0^{(1)} \rho'_{(0)}$, $E_0^{(1)} \rho'_{(d)} : \mathfrak{D}^{(1)}(\Sigma) \rightarrow \mathcal{E}^{(1)}(\mathcal{M})$ are continuous, and we also know that $d^{(4)} : \mathcal{E}^{(1)}(\mathcal{M}) \rightarrow \mathcal{E}^{(2)}(\mathcal{M})$ is continuous, therefore $\mathcal{F} \in \mathcal{E}^{(2)}(\mathcal{M})$ is continuously dependent on $A_{(0)}$ and $A_{(d)}$. ■

From this final result, we see that $\mathcal{F}$ depends only on the data $A_{(0)}$ and $A_{(d)}$. Thus, we can set $A_{(n)} = 0$ in (B24). This choice of data is useful when quantizing the electromagnetic field.[4]

[1] S. Fulling, *Aspects of Quantum Field Theory on Curved Space–times* (Cambridge University Press, Cambridge, 1989).
[2] R. M. Wald, *Quantum Field Theory in Curved Space–time and Black Hole Thermodynamics* (Chicago University Press, Chicago, 1994).
[3] J. Dimock, ''Algebras of local observables on a manifold,'' Commun. Math. Phys. **77**, 219 (1980).
[4] J. Dimock, ''Quantized electromagnetic field on a manifold,'' Rev. Math. Phys. **4**, 223 (1992).
[5] J. Dimock, ''Dirac quantum fields on a manifold,'' Trans. Am. Math. Soc. **269**, 133 (1982).
[6] E. P. Furlani, ''Quantization of the electromagnetic field on static space–times,'' J. Math. Phys. **3**, 1063 (1995).
[7] E. P. Furlani, ''Quantization of massive vector fields on ultrastatic space–times,'' Class. Quantum Grav. **14**, 1665 (1997).
[8] U. Mosel, *Fields, Symmetries, and Quarks* (McGraw-Hill, New York, 1989).
[9] Y. Choquet-Bruhat, ''Hyperbolic partial differential equations on a manifold,'' in *Battelle Rencontres: Lectures in Mathematics and Physics*, edited by C. M. DeWitt and J. A. Wheeler (Benjamin, New York, 1967), pp. 84–106.
[10] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis Manifolds and Physics Part I: Basics*, revised edition (North–Holland, New York, 1982).
[11] J. Leray, *Hyperbolic Differential Equations*, lecture notes (Institute for Advanced Study, Princeton, 1953).
[12] F. G. Friedlander, *The Wave Equation on a Curved Space–Time* (Cambridge University Press, Cambridge, 1975).
[13] R. M. Wald, *General Relativity* (Chicago University Press, Chicago, 1984).
[14] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, New York, 1973).
[15] C. W. Misner and J. A. Wheeler, ''Classical physics as geometry: Gravitation, electromagnetism, unquantized charge, and mass as properties of curved space,'' Ann. Phys. (San Diego) **2**, 525 (1957).
[16] A. Lichnerowicz, Propagateurs, Commutateurs et Anticommutateurs en Relativite Generale, Publication I.H.E.S. No. 10, 1961.
[17] F. Treves, *Topological Vector Spaces, Distributions, and Kernels* (Academic, New York, 1969).
[18] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics II* (Springer, New York, 1981).
[19] C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw–Hill, New York, 1980).
[20] M. Gockeler and T. Schucker, *Differential Geometry: Gauge Theories and Gravity*, Cambridge Monographs on Mathematical Physics (Cambridge University Press, Cambridge, 1990).
[21] C. Westenholz, *Differential Forms in Mathematical Physics*, revised edition (North–Holland, Amsterdam, 1986), Vol. 3.
[22] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2nd ed. (Springer, New York, 1983).

Exact spectral values for discrete quantum pendulum-integrals

Johannes Kellendonk[a)]
Fachbereich Mathematik, Technische Universität Berlin, 10623 Berlin, Germany

(Received 1 July 1998; accepted for publication 19 January 1999)

For specific choice of parameters the spectrum of the discrete quantum pendulum-integral contains the eigenvalues of a finite matrix which depends analytically on the flux. Under natural continuity assumptions these eigenvalues include the spectral values which may be obtained by the algebraic Bethe ansatz. © *1999 American Institute of Physics.* [S0022-2488(99)01205-0]

I. INTRODUCTION

Discrete quantum pendulum-integrals (QP-integrals) are integrals of motion of a (1+0)-dimensional discrete integrable quantum field theory. This quantum field theory is the discrete sine–Gordon field theory with shortest possible (symmetric) periodicity in space. With space compatified in that way to one point, the model becomes that of the discrete quantum (mathematical) pendulum.[1] Coincidently, QP-integrals arise in solid state physics as well: they may be interpreted as magnetic Schrödinger operators on $\mathbb{Z}^2$ generalizing the Hofstadter Hamiltonian[2] which describes an elementary quantum–Hall system.[3] The QP-integrals so arise as self-adjoint elements of the discrete Weyl–Heisenberg algebra, which is the C^*-algebra generated by two unitary elements u and v which are subject to the relation $uv=q^{-1}vu$, $q=e^{i\gamma}$. Here γ plays here the role of the flux of the magnetic field. To define QP-integrals in this context, consider the family

$$H(a,b,k):=au+\overline{a}u^*+bv+\overline{b}v^*+k(q^{1/2}uv+q^{-1/2}v^*u^*)+k^{-1}(q^{-1/2}uv^*+q^{1/2}vu^*) \quad (1)$$

parametrized by complex continuous functions $a=a(q^{1/2})$ and $b=b(q^{1/2})$ and a strictly positive real number k. We will right away focus our attention on particular choices for a and b and write more briefly

$$\tilde{H}(z,n,k):=H(z+z^{-1},q^{n/2}+q^{-n/2},k), \quad (2)$$

$z=z(q^{1/2})$, $n\in\mathbb{N}$. In case $z(q^{1/2})$ is of modulus 1, $\tilde{H}(z,n,k)$ arises as an integral of motion of discrete sine–Gordon field theory with shortest possible periodicity in space.[4,5] For $z(q^{1/2})=q^{n/2}$, i.e., $a=b=2\cos(n\gamma/2)$, $\tilde{H}(z,n,k)$ is referred to as a QP-integral. We shall prove here that, for $z+z^{-1}\in\mathbb{R}$, the spectrum of $\tilde{H}(z,n,k)$ contains the eigenvalues of the tridiagonal matrix

$$\tilde{A}(z,n,k):=\begin{pmatrix} \tilde{x}_0 & \tilde{y}_0 & & & \\ \tilde{z}_0 & \tilde{x}_1 & \tilde{y}_1 & & \\ & \ddots & \ddots & \ddots & \\ & & & \tilde{z}_{n-2} & \tilde{x}_{n-1} \end{pmatrix}, \quad (3)$$

where

$$\tilde{x}_j=-(z+z^{-1})(kq^{(1-n)/2+j}+k^{-1}q^{(n-1)/2-j}), \quad (4)$$

[a)]Electronic mail: kellen@math.tu-berlin.de

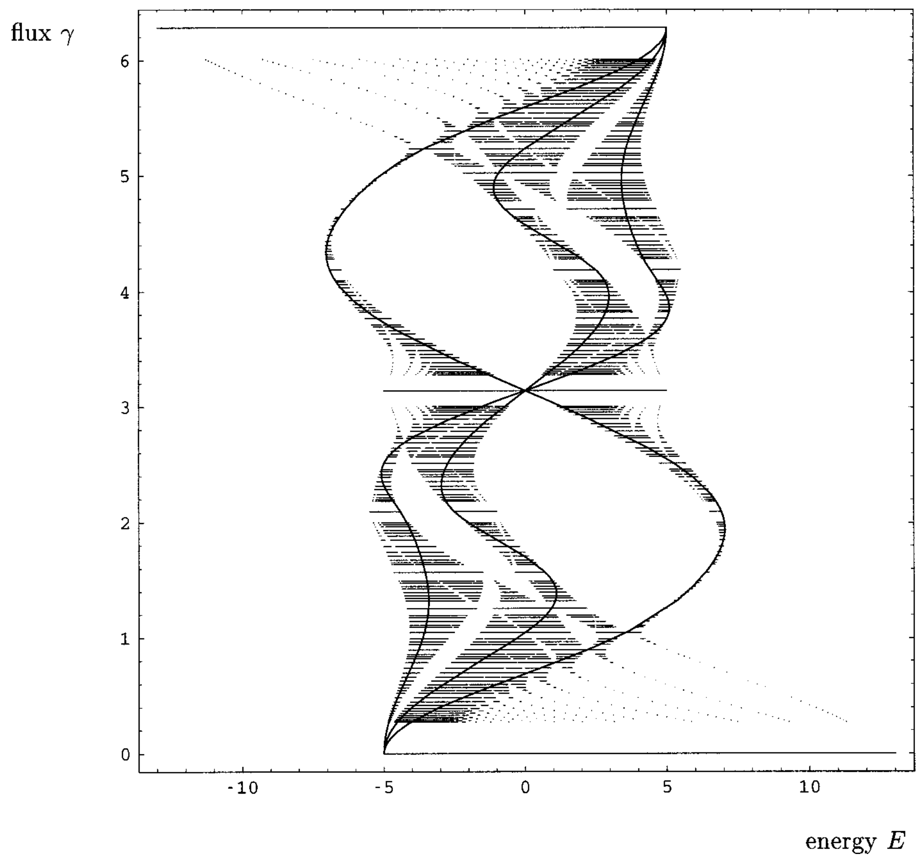

FIG. 1. Spectrum of $\tilde{H}(q^{3/2},3,2)$ (horizontal lines, for $\gamma/2\pi = M/N$ with small N) and eigenvalues of $\tilde{A}(q^{3/2},3,2)$ (continuous curves).

$$\tilde{y}_j = q^{n/2} + q^{-n/2} - (q^{j+1-n/2} + q^{n/2-j-1}), \tag{5}$$

$$\tilde{z}_j = q^{n/2} + q^{-n/2} - (k^2 q^{j+1-n/2} + k^{-2} q^{n/2-j-1}). \tag{6}$$

Figure 1 shows a plot of the spectrum of $\tilde{H}(q^{3/2},3,2)$ against γ (for rational $\gamma/2\pi$ with small denominator) together with the three curves of eigenvalues of $\tilde{A}(q^{3/2},3,2)$.

Although diagonalization of the above matrix furnishes only finitely many spectral values, we find this result surprising, since one expects the spectrum of $H(a,b,k)$ to be a Cantor set for $\gamma/2\pi \notin \mathbb{Q}$ and not much is known about spectral values of operators with Cantor spectrum.

As already mentioned, in case $z(q^{1/2})$ is of modulus 1, $\tilde{H}(z,n,k)$ is an integral of motion of an integrable discrete quantum field theory. The integrability means in particular that one may determine (generalized) eigenvalues and functions of its integrals of motion, among which are the QP-integrals. This is achieved by use of the algebraic Bethe ansatz,[6] which was adapted to the present case by Nadja Kutz.[4] It leads to the following result.[4,5]

Theorem 1: *Let $\delta \in S^1 = \mathbb{R}/2\pi\mathbb{Z}$, $n \in \mathbb{N}$, and $k > 0$. If the set of $n-1$ spectral parameters $\eta_1, \ldots, \eta_{n-1}$, which ought to be complex, pairwise different numbers, satisfies the so-called Bethe ansatz equations*

$$\prod_{j \neq i=1}^{n-1} \frac{\eta_j q - \eta_i}{\eta_j - \eta_i q} = e^{2i\delta} \frac{q^{-1/2}\eta_j^2 + \eta_j(k+k^{-1}) + q^{1/2}}{q^{1/2}\eta_j^2 + \eta_j(k+k^{-1}) + q^{-1/2}}, \tag{7}$$

then, with $q = e^{i\gamma}$,

$$E(\eta_1,\ldots,\eta_{n-1}) = -2(k+k^{-1})\cos\left(\frac{n-1}{2}\gamma+\delta\right)+4\sin\left(\frac{n-1}{2}\gamma+\delta\right)\sin\frac{\gamma}{2}\sum_{j=1}^{n-1}\eta_j^{-1}(\gamma) \quad (8)$$

lies in the spectrum of $\tilde{H}(e^{i\delta},n,k)$.

A formula for eigenfunctions (in the generalized sense, i.e., not neccessarily normalizable) corresponding to these so-called Bethe ansatz eigenvalues will be recalled later.

We expect that the so-called Bethe ansatz eigenvalues (8) coincide with the eigenvalues of $\tilde{A}(e^{i\delta},n,k)$. Indeed, we shall derive that the Bethe ansatz eigenvalues are among the eigenvalues of $\tilde{A}(e^{i\delta},n,k)$ in case $k=1$ and extend this result to $k\neq 1$ under natural assumptions on the behavior of the Bethe ansatz eigenvalues on k, z, and γ. Therefore, our result may also be understood as a simpler approach to the determination of the Bethe ansatz eigenvalues and we expect that a similar simplification occurrs in other models, too, in particular those coming from the sine–Gordon model. In fact, the Bethe ansatz equations are rather difficult to solve. Already for $n=3$ they lead to polynomial equations which have too high orders to be tractable by analytic means. In Ref. 5, solutions for $k>1$ were for that reason only obtained in case $n\leqslant 2$.

The key to compare the Bethe ansatz eigenvalues with the eigenvalues of $\tilde{A}(e^{i\delta},n,k)$ is to look first at $k=1$, a case in which the solution of the Bethe ansatz eigenvector simplifies enormously. It is shown in Ref. 5 that the Bethe ansatz eigenvector is then not only a true eigenvector, or, more precisely, a square summable vector in a particular representation on $\ell^2(\mathbb{Z})$, but it even vanishes in that representation outside $\{0,\ldots,n-1\}\subset\mathbb{Z}$. This leads us to a remark on yet another approach to determine spectral values of $\tilde{H}(z,n,k)$ which we will, however, not follow here. That approach is essentially an ansatz for eigenfunctions of $\tilde{H}(z,n,k)$ which have finite support. Formulated in a representation on $L^2(S^1)$, the Fourier space of $\ell^2(\mathbb{Z})$, the above ansatz is one for polynomial eigenfunctions. It may then also be reformulated in terms of Bethe ansatz equations[7,5] [different from (7)] which, however, are again rather difficult to solve.

After recalling some important facts about the discrete Weyl–Heisenberg algebra, which we do in the next section following mainly Ref. 5, we prove in Sec. III that the eigenvalues of $\tilde{A}(z,n,k)$ are spectral values of $\tilde{H}(z,n,k)$ by showing that the characteristic polynomial factorizes in the appropriate way. In Sec. IV we relate the Bethe ansatz eigenvalues to these eigenvalues.

II. PRELIMINARIES

The discrete Weyl–Heisenberg or rotation algebra $\mathcal{A}_\gamma$ with angle γ is the C^*-envelope of the $*$-algebra generated by two elements u and v which are subject to the relations[8]

$$uu^*=1,\quad u^*u=1,\quad vv^*=1,\quad v^*v=1,\quad uv=q^{-1}vu,$$

with $q=e^{i\gamma}$. Any element of the discrete Weyl–Heisenberg algebra $\mathcal{A}_\gamma$ may be approximated in norm by finite sums of the form $\Sigma\, c_n u^{n_1}v^{n_2}$, $n\in\mathbb{Z}^2$, $c_n\in\mathbb{C}$. A faithful representation of $\mathcal{A}_\gamma$ on the Hilbert space $\ell^2(\mathbb{Z}^2)$ is given by

$$u\cdot\psi(n)=e^{iA_1(n)}\psi(n_1-1,n_2), \quad (9)$$

$$v\cdot\psi(n)=e^{iA_2(n)}\psi(n_1,n_2-1), \quad (10)$$

where A: $\mathbb{Z}^2\to\mathbb{R}^2$ is a gauge potential with discrete rotation γ:

$$(A_2(n)-A_2(n_1-1,n_2))-(A_1(n)-A_1(n_1,n_2-1))=\gamma,$$

and we have used the convention to denote the action of an element $x\in\mathcal{A}_\gamma$ in some representation space simply by $\cdot$. In other words, $\mathcal{A}_\gamma$ is isomorphic to the norm-closure of the subalgebra of bounded operators of $\ell^2(\mathbb{Z}^2)$ generated by u and v in the above representation. The above representation yields the framework for the discretized version of the Landau model which describes an electric particle in the discretized plane $\mathbb{Z}^2$ to which a constant magnetic field is perpendicularly applied. More precisely, $\mathcal{A}_\gamma$ is the algebra of observables for that model, and the famous Hofstadter Hamiltonian, which plays the role of the discrete Landau Hamiltonian, equals $u+v+u^*+v^*$ acting in the above representation. The angle γ is proportional to the magnetic field or the flux per unit cell. The function A is for given γ unique up to a discrete gradient. Different gauges lead, in general, to different representations which are, however, unitarily equivalent. For simplicity we will call $\gamma/2\pi$ the flux.

The above representation is faithful but not irreducible. Its decomposition into irreducible components leads, for rational $\gamma/2\pi$, to a family of representations labeled by a two-torus, the two-torus of quasimomenta, or Bloch parameters. The first step of this decomposition, which applies as well to irrational $\gamma/2\pi$, leads to the family of Weyl–Schrödinger representations, labeled by an angle θ, with representation space $\ell^2(\mathbb{Z})$. They are given by

$$u\cdot\psi(n)=e^{i\theta}q^{-n}\psi(n), \tag{11}$$

$$v\cdot\psi(n)=\psi(n-1). \tag{12}$$

For irrational $\gamma/2\pi$ the Weyl–Schrödinger representations are irreducible and still faithful; for rational $\gamma/2\pi$ this is not the case. In fact, for $\gamma/2\pi=M/N$, $(M,N)=1$ (meaning that M and N are coprime), the Weyl–Schrödinger representations are neither irreducible nor faithful. In that case, the Weyl–Schrödinger representation with angle θ decomposes into a direct integral of N-dimensional irreducible representations parametrized by a second angle. Denoting the second angle by φ, u has in the irreducible representation labeled by (θ,φ) matrix representation

$$u_{(\theta,\varphi)}=e^{i\theta}\begin{pmatrix} q^{-1} & & & & \\ & \ddots & & & \\ & & q^{-1} & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}, \tag{13}$$

which in particular is independent of φ, and v

$$v_{(\theta,\varphi)}=\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 \\ e^{i\varphi} & 0 & \cdots & & 0 \end{pmatrix}. \tag{14}$$

Two such matrix representations, one for (θ,φ) and one for (θ',φ'), are unitarily equivalent whenever $\theta-\theta'=n\gamma$ for some $n\in\mathbb{Z}$ and $\varphi=\varphi'$, that is, if $u_\theta^N=u_{\theta'}^N$ and $v_\varphi^N=v_{\varphi'}^N$. The asymmetry appearing in the parameters (θ,φ) labeling the irreducible components is, from a physical point of view, artificial. In the (discrete) Landau model, u^N and v^N are translation operators along N sites in the one- and two-directions, respectively, and the parameters $N\theta$ and φ have the physical interpretation of quasimomentum in the x_1- and the x_2-direction, respectively. We denote by $x_{(\theta,\varphi)}$ the matrix representation of $x\in\mathcal{A}_\gamma$ in the N-dimensional representation labeled by (θ,φ).

It follows from the above consideration that the spectrum $\sigma(x)$ of any element $x\in\mathcal{A}_\gamma$ is, for irrational $\gamma/2\pi$, equal to the spectrum of x in any Weyl–Schrödinger representation, and for rational $\gamma/2\pi$ given by

$$\sigma(x)=\bigcup_{(\theta,\varphi)\in S^1\times S^1}\sigma(x_{(\theta,\varphi)}).$$

Since $x_{(\theta,\varphi)}$ is a finite-dimensional matrix, its spectrum consists of eigenvalues, the zeroes of its characteristic polynomial. It turns out that, for our family of operators, the characteristic polynomial has a very special form, namely, with $\gamma/2\pi = M/N$, $(M,N)=1$,

$$\det(H_{(\theta,\varphi)}(a,b,k)-E)=p(a,b,k;E)+h(a,b,k;\theta,\varphi),$$

where $p(a,b,k;E)$ is a polynomial in E of degree N which is independent of (θ,φ), and $h(a,b,k;\theta,\varphi)$, the so-called off-set function, is independent of E. Such a relation is called a Chambers relation. While the polynomial p may look rather complicated, the off-set function h is explicitly known.[2,5] With $T_N(z+z^{-1})/2)=(z^N+z^{-N})/2$ for any complex $z\neq 0$, it is

$$h(a,b,k;\theta,\varphi)=-2\left(T_N\left(-\frac{a}{2}\right)e^{iN\theta}+T_N\left(-\frac{b}{2}\right)e^{i\varphi}\right)+(-1)^M(k^N e^{i(N\theta+\varphi)}+k^{-N}e^{i(N\theta-\varphi)})+\text{c.c.}$$

We also write

$$\tilde{h}(z,n,k;\zeta_1,\zeta_2)=h\left(z+z^{-1},2\cos\frac{n\gamma}{2},k;\theta,\varphi\right),\quad \zeta_1=e^{iN\theta},\quad \zeta_2=e^{i\varphi}. \tag{15}$$

For the above result we have normalized $p(a,b,k;0)=0$ and c.c. stands for complex conjugate. T_N is, up to constants, the Nth Chebychev polynomial of the second kind.

Despite the irregular behavior of $p(a,b,k;E)$ on γ, the spectrum of $H(a,b,k)$ depends, in a certain sense, continuously on the flux. The collection of all $\mathcal{A}_\gamma$, $\gamma\in S^1$, can be combined into a continuous field of C^*-algebras over S^1.[9] Its continuous sections, which are in particular functions $S^1\to\cup_\gamma\mathcal{A}_\gamma$: $\gamma\mapsto x_\gamma\in\mathcal{A}_\gamma$ for which $\gamma\mapsto\|x_\gamma\|$ is continuous, form a C^*-algebra, and it is shown[10] that, under the hypothesis that all x_γ are normal, the following continuity property holds: for each open $U\subset\mathbb{C}$, the set of $\gamma\in S^1$ for which $\sigma(x_\gamma)\cap U\neq\emptyset$ and the set of $\gamma\in S^1$ for which $\sigma(x_\gamma)\in U$ are both open. In the present case, $H(a,b,k)=H_\gamma(a(\gamma),b(\gamma),k)$, which depends on γ also through the relation between u and v, is a continuous section of self-adjoint elements provided we remove one point (let's say π) from S^1. We are forced to remove one point, because $H_\gamma(a(\gamma),b(\gamma),k)$ contains square roots of $q=e^{i\gamma}$. An immediate consequence of the above property is therefore that any continuous function E: $S^1\backslash\{\pi\}\to\mathbb{C}$ which satisfies $E(\gamma)\in\sigma(H_\gamma(a(\gamma),b(\gamma),k))$ on a dense subset of S^1 must already satisfy $E(\gamma)\in\sigma(H_\gamma(a(\gamma),b(\gamma),k))$ for all $\gamma\in S^1\backslash\{\pi\}$.

III. FACTORIZATION OF THE CHARACTERISTIC POLYNOMIAL

As already mentioned in the Introduction, if $k=1$, a different ansatz to determine the eigenvalues (8) of $H(a,b,k)$ is one for eigenfunctions of finite support in a Weyl–Schrödinger representation of $\mathcal{A}_\gamma$. Denoting by $H_\theta(a,b,k)$ the operator $H(a,b,k)$ acting in such a representation with angle θ, we have

$$H_\theta(a,b,k)\psi(m)=z_{m-1}\psi(m-1)+x_m\psi(m)+y_m\psi(m+1),$$

where

$$x_m=aq^{-m}e^{i\theta}+\overline{aq}^m e^{-i\theta}, \tag{16}$$

$$y_m=b+(kq^{-m-1/2}e^{i\theta}+\text{inv.}), \tag{17}$$

$$z_m=\overline{b}+(k^{-1}q^{-m-1/2}e^{i\theta}+\text{inv.}), \tag{18}$$

$x+\text{inv.}$ standing for $x+x^{-1}$ and θ labeling the representation. Note that $H_\theta(a,b,k)$ is self-adjoint only if $k\in\mathbb{R}$. We have required this above and in Ref. 5 for the definition of QP-integrals, but below we allow k to be complex.

Lemma 1: Let $\theta\in\mathbb{C}$ *and* $n\in\mathbb{N}$. *If* $k=-q^{(n-1)/2}e^{-i\theta}$ *and* $b=2\cos(n\gamma/2)$, *then* $y_{-1}=y_{n-1}=0$.

Proof: Setting $\kappa=kq^{1/2}e^{i\theta}$ one obtains that $y_{-1}=y_{n-1}=0$ is equivalent to $-b=\kappa+\kappa^{-1}=\kappa q^{-n}+\kappa^{-1}q^{n}$. This, in turn, is equivalent to $\kappa=\epsilon q^{n/2}$ and $b=-\epsilon 2\cos(n\gamma/2)$, $\epsilon=\pm 1$, the statement of the lemma being one choice of sign. q.e.d.

In view of the above lemma we shall from now on restrict the parameter b to be

$$b=2\cos\frac{n\gamma}{2}.$$

Let P_n be the orthogonal projection onto the subspace of $\ell^2(\mathbb{Z})$ of all wave functions which vanish outside $\{0,...,n-1\}$. The operator

$$A_\theta(z,n,k):=P_n\tilde{H}_\theta(z,n,k)P_n$$

has, when restricted to the image of P_n, (tridiagonal) matrix representation

$$A_\theta(z,n,k)=\begin{pmatrix} x_0 & y_0 & & & \\ z_0 & x_1 & y_1 & & \\ & \ddots & \ddots & \ddots & \\ & & & z_{n-2} & x_{n-1} \end{pmatrix}.$$

The following result was also obtained in Ref. 5.

Theorem 2: *If* $k=1$ *and* $\theta=\pi+[(n-1)/2]\gamma$, *then the eigenvalues of* $\tilde{A}(z,n,1)$ *coincide with those eigenvalues of* $H_\theta(z,n,1)$ *whose corresponding eigenvectors vanish outside* $\{0,...,n-1\}\subset\mathbb{Z}$.

Proof: If $k=1$ and $\theta=\pi+[(n-1)/2]\gamma$, then $\tilde{A}(z,n,k)=A_\theta(z,n,k)$ and $y_{-1}=y_{n-1}=0$ and $y_j=z_j$. This implies that $H_{\pi+[(n-1)/2]\gamma}(z,n,1)$ commutes with P_n from which the statement follows. q.e.d.

Let

$$\tilde{\mathcal{B}}(z,n,k):=\left\{(\zeta_1,\zeta_2)\in\mathbb{C}\times\mathbb{C}\,\middle|\,\zeta_2=-z^N\frac{(-1)^{Mn}\zeta_1-(-1)^{M+N}k^N}{1-(-1)^{M(n-1)+N}\zeta_1k^N}\right\} \tag{19}$$

and

$$\mathcal{B}(z,n,k):=\{(\theta,\varphi)\in S^1\times S^1|(e^{iN\theta},e^{i\varphi})\in\tilde{\mathcal{B}}\}. \tag{20}$$

In Ref. 5 it was shown that $\mathcal{B}(z,n,k)$ furnishes, for stictly positive k and $|z|=1$, the set of labels (Bloch parameters) for the irreducible representations of $\mathcal{A}_\gamma$ in which the Bethe ansatz eigenvalues (8) are spectral values of $\tilde{H}_{(\theta,\varphi)}(z,n,k)$. The following lemma, which follows for stictly positive k and $|\zeta_1|=|\zeta_2|=|z|=1$ from that result, is straightforwardly verified:

Lemma 2: The off-set function $\tilde{h}(z,n,k;\zeta_1,\zeta_2)$ *is for* $z+z^{-1}\in\mathbb{R}$, *nonzero* $k\in\mathbb{C}$, *and* $n\in\mathbb{N}$ *constant on* $\tilde{\mathcal{B}}(z,n,k)$.

In the irreducible representation of $\mathcal{A}_{2\pi M/N}$ labeled by (θ,φ), $\tilde{H}(z,n,k)$ is represented by the matrix (tridiagonal apart from the entries in the corners)

$$\tilde{H}_{(\theta,\varphi)}(z,n,k)=\begin{pmatrix} x_0 & y_0 & & & z_{-1}e^{-i\varphi} \\ z_0 & x_1 & y_1 & & \\ & \ddots & \ddots & \ddots & \\ y_{-1}e^{i\varphi} & & & z_{N-2} & x_{N-1} \end{pmatrix}.$$

Theorem 3: *Let* $z+z^{-1}\in\mathbb{R}$ *and, for* $\gamma/2\pi=M/N$, $(M,N)=1$,

$$\chi(z,n,k,E):=p\left(z+z^{-1},2\cos\frac{n\gamma}{2},k;E\right)+h(z,n,k;\zeta_1,\zeta_2),$$

where $(\zeta_1,\zeta_2)\in\tilde{\mathcal{B}}(z,n,k)$. *Then*

$$\chi(z,n,k,E)=\det(\tilde{A}(z,n,k)-E)\det(\tilde{B}(z,n,k)-E), \tag{21}$$

where

$$\tilde{B}(z,n,k):=\begin{pmatrix}\tilde{x}_n & \tilde{y}_n & & & \\ \tilde{z}_n & \tilde{x}_{n+1} & \tilde{y}_{n+1} & & \\ & \ddots & \ddots & \ddots & \\ & & & \tilde{z}_{N-2} & \tilde{x}_{N-1}\end{pmatrix}.$$

In particular, the eigenvalues of $\tilde{A}(z,n,k)$ *are, even for arbitary* $\gamma/2\pi$, *spectral values of* $\tilde{H}(z,n,k)$, *their corresponding (generalized) eigenvectors occurring in irreducible representations which, for* $\gamma/2\pi=M/N$, $(M,N)=1$, *have Bloch parameters in* $\mathcal{B}(z,n,k)$.

Proof: Consider first the rational case $\gamma/2\pi=M/N$, $(M,N)=1$, in which $\chi(z,n,k,E)$ is the characteristic polynomial of $\tilde{H}_{(\theta,\varphi)}(z,n,k)$ with $(\theta,\varphi)\in\mathcal{B}(z,n,k)$. Since $\chi(z,n,k,E)$ is independent on the value of $(e^{iN\theta},e^{i\varphi})$ as long as the latter belongs to $\tilde{\mathcal{B}}(z,n,k)$, we may consider complex $k\neq 0$ and complex θ such that $k=-q^{(n-1)/2}e^{-i\theta}$. Then, using Lemma 1, the characteristic polynomial can be written as

$$\chi(z,n,k,E)=\det(A_\theta(z,n,k)-E)\det(B_\theta(z,n,k)-E)-(-1)^N e^{-i\varphi}\prod_{j=0}^{N-1} z_j,$$

where $B_\theta(z,n,k):=(1-P_n)\tilde{H}_{(\theta,\varphi)}(z,n,k)(1-P_n)$ and $(e^{iN\theta},e^{i\varphi})\in\tilde{\mathcal{B}}(z,n,k)$. (Here P_n has to be understood as the projection onto the first n components of the vectors in $\mathbb{C}^N$.) But $(e^{iN\theta},e^{i\varphi})\in\tilde{\mathcal{B}}(z,n,k)$ implies

$$e^{-i\varphi}=-z^{-N}\frac{1-(-1)^{M(n-1)+N}e^{iN\theta}(-e^{-i\theta}q^{(n-1)/2})^N}{(-1)^{Mn}e^{iN\theta}-(-1)^{M+N}(-e^{-i\theta}q^{(n-1)/2})^N}=0.$$

Furthermore, at the above value for θ we have $A_\theta(z,n,k)=\tilde{A}(z,n,k)$ and $B_\theta(z,n,k)=\tilde{B}(z,n,k)$. This proves Eq. (21). Thus all eigenvalues of $\tilde{A}(z,n,k)$ are spectral values of $\tilde{H}(z,n,k)$, at least if $\gamma/2\pi$ is rational. By the continuity property of the spectrum with respect to γ, which we mentioned in the last section, this is also the case for irrational $\gamma/2\pi$. q.e.d.

IV. COMPARISON WITH BETHE ANSATZ EIGENVALUES

As mentioned in the Introduction, the algebraic Bethe ansatz can be used to determine exact spectral values of $\tilde{H}(z,n,k)$ as well. In this section we argue that the spectral values determined in Sec. III contain the Bethe ansatz eigenvalues, the argumentation being based on natural continuity assumptions which we cannot prove yet. Much of the work has already been done, because Theorem 3 states in particular that the Bloch parameters, which label, for rational flux, the irreducible representations in which spectral values of $\tilde{H}(z,n,k)$ coincide with the eigenvalues of $\tilde{A}(z,n,k)$, belong to $\mathcal{B}(z,n,k)$. On the other hand, it has been shown in Ref. 5 that the Bethe ansatz eigenvalues (8) are spectral values of $\tilde{H}(z,n,k)$ in these representations as well. Therefore, to assert our claim in the rational case we just need to show that the spectral values of $\tilde{H}(z,n,k)$ coinciding with the eigenvalues of $\tilde{A}(z,n,k)$ on the one hand and its Bethe ansatz eigenvalues on the other belong to the same h-bands. By an h-band we mean the closure of a connected component of $p^{-1}((h_{\min},h_{\max}))$, the preimage under p understood as a polynomial in E [i.e., $p(E) = p(z+z^{-1},2\cos n\gamma/2,k;E)$] of the open interval $(h_{\min},h_{\max})$, $h_{\min}$ and $h_{\max}$ denoting the absolute minimum and maximum, respectively, of the off-set function as a function of (θ,φ). The strategy is to show this for $k=1$ by comparison of the eigenfunctions, and then to extend this result to arbitrary $k\neq 0$ by topological arguments assuming continuity.

A detailed derivation of the Bethe ansatz for the discrete sine–Gordon model and the quantum pendulum goes beyond the scope of this article, so we give only a very short description and refer the reader to the original literature[11,12,1,4] and to Ref. 5 for more information on that subject. We use mainly the notation of Ref. 5.

Doubly discrete Minkowski space–time in 1+1 dimensions is modeled by $\mathbb{Z}^2$ as a light cone lattice. This means that the two elements of its standard base, here denoted by e_l and e_r, point in left and in right moving directions of the light cone, respectively. Quantum fields of the (doubly) discrete sine–Gordon model are fields on that lattice whose values are unitary operators $Q(n)$, $n\in\mathbb{Z}^2$, which commute exept if they are neighbors on the light cone, namely in this case

$$Q(n+e_r)Q(n)=q^{-1}Q(n)Q(n+e_r),$$

$$Q(n+e_l)Q(n)=q^{-1}Q(n)Q(n+e_l).$$

Here $q=e^{i\gamma}$, where γ is a real number proportional to Planck's constant. Moreover, the fields are subject to the field equation

$$Q_{n+e_r+e_l}=\frac{k+q^{1/2}Q_{n+e_l}}{1+q^{1/2}kQ_{n+e_l}}\,\frac{k+q^{1/2}Q_{n+e_r}}{1+q^{1/2}kQ_{n+e_r}}\,Q_n^{-1}, \tag{22}$$

the discrete equation of motion. The strictly positive real number k is a parameter of the model. In particular, the quantum fields on all of the light cone lattice are uniquely determined by their values on the set $\{n\in\mathbb{Z}^2|n_1-n_2\in\{0,1\}\}$, a so-called Cauchy zig-zag. The discrete quantum pendulum arrises upon taking specific periodic boundary conditions (periodic in space). First, let us take two-periodic boundary conditions, i.e., $Q(n+2(e_r-e_l))=Q(n)$. This then implies that

$$z(n)\coloneqq Q(n)Q(n+e_r-e_l)^{-1}$$

is a unitary which commutes with all Q's. Furthermore, the $z(n)$ are invariant under time evolution. Thus the (operator-valued) degrees of freedom of the discrete sine–Gordon model which is two-periodic in space may be taken to be $\{Q(0),Q(e_l),z(0),z(e_l)\}$, in particular they generate the C^*-algebra $\mathcal{C}\coloneqq\mathcal{A}_\gamma\otimes\mathcal{C}(S^1\times S^1)$, and time evolution is given by an automorphism of $\mathcal{A}_\gamma$ extended trivially on the second factor in the tensor product. Second, fixing eigenvalues for $z(0)$ and $z(e_l)$ amounts to taking boundary conditions which make the periodicity even shorter (as short as possible), periodicity being understood here as one up to an alternating phase. The discrete quantum pendulum is the discrete sine–Gordon model with such boundary conditions but with an additional symmetry requirement that both eigenvalues are equal. How do we obtain integrals of motion for these models?

The answer to this question is based on the observation that the fields of the discrete sine–Gordon model are the gauge-invariant and spectral-parameter-independent part of a larger field algebra in which a field is assigned to each edge of the light cone lattice (or strictly speaking its dual) and the values of the field are 2×2 matrices: An edge is simply a link between n and $n+e_l$ or between n and $n+e_r$. The equivalent of a Cauchy zig-zag is therefore a sequence of edges which are every other time parallel to e_l or to e_r, respectively. Indexing these edges by integer numbers so that the sth edge is parallel to e_r if s is odd and parallel to e_l if s is even, the value of the above-mentioned field on edge s is given by

$$\mathcal{V}_s(\eta)=\begin{pmatrix} \hat{u}_s & -\eta^{1/2}k^{(-1)^s/2}\hat{v}_s^{-1} \\ \eta^{1/2}k^{(-1)^s/2}\hat{v}_s & \hat{u}_s^{-1} \end{pmatrix},$$

where $\hat{u}_s$ and $\hat{v}_s$ are unitaries satisfying $\hat{u}_s\hat{v}_s=q^{-1/2}\hat{v}_s\hat{u}_s$, i.e., generating $\mathcal{A}_{\gamma/2}$, and η is a spectral parameter. Such a matrix is called Volterra L matrix,[11] in fact, the discrete Volterra model and the discrete sine–Gordon model are related by a simple coordinate transformation. The above-mentioned sine–Gordon fields on the Cauchy zig-zag are quartic expressions in the $\hat{u}_s$ and $\hat{v}_s$ which are independent of η and their time evolution follows from a quantum analog of a zero curvature condition stated in terms of the Volterra matrices. We do not repeat the formulas here, but refer the reader to Ref. 5. The use of that formulation is that for $2p$-periodic boundary conditions (in space, i.e., there are $2p$ independent edges in the Cauchy zig-zag), the matrix trace of the monodromy matrix $\mathcal{M}(\eta)=\mathcal{V}_{2p}(\eta)\cdots\mathcal{V}_1(\eta)$ is conserved under time evolution and therefore are the coefficients of its expansion in the spectral parameter integrals of motion. In the case of the sine–Gordon model with periodicity 2 in space there are four edges in a Cauchy zig-zag, i.e., $p=2$. In that case, the relation between the sine–Gordon variables and the entries of the four Volterra L matrices is

$$Q(0)=\hat{u}_3^{-1}\hat{u}_2^{-1}\hat{v}_3^{-1}\hat{v}_2,\quad Q(e_l)=\hat{u}_2^{-1}\hat{u}_1^{-1}\hat{v}_2^{-1}\hat{v}_1,$$

$$Q(e_r)=\hat{u}_4\hat{u}_3\hat{v}_4\hat{v}_3^{-1},\quad Q(e_r-e_l)=\hat{u}_1\hat{u}_4\hat{v}_1\hat{v}_4^{-1}.$$

We now describe the monodromy matrix in this case in detail.

Its diagonal elements turn out to depend only on the sine–Gordon variables $\{Q(0),Q(e_l),z(0),z(e_l)\}$, and moreover $\mathcal{M}_{22}(\eta)^*=\mathcal{M}_{11}(\eta)$, where we treat the spectral paramater as a variable which is independent under the $*$-operation. Expansion in the spectral parameter results in

$$\mathcal{M}_{11}(\eta)=A^{(0)}-\eta A^{(1)}+\eta^2A^{(2)},$$

where $A^{(2)}$ is a unitary depending on $z(0)$, $z(e_l)$ which is the identity if $z(0)=z(e_l)$,

$$A^{(0)}=w^2A^{(2)}$$

with another unitary w depending on $z(0)$, $z(e_l)$, and

$$A^{(1)}+A^{(1)*}=H(wA^{(2)}+(wA^{(2)})^*,w+w^*,k).$$

(The precise form of $A^{(1)}$ is not of importance here.) $H(wA^{(2)}+(wA^{(2)})^*,w+w^*,k)$, which is (1) with operator-valued a and b (and real k), should now be understood as an element of $\mathcal{C}$. The unitaries u and v of (1) are, up to factors depending on $z(0)$, $z(e_l)$, equal to $Q(e_l)^{-1}$ and $Q(0)^{-1}$, respectively. This change of variables is convenient for what follows; its precise form a bit cumbersome but straightforward and we supress it here. It can be found in Ref. 5. To summarize, there are three integrals of motion for the discrete sine–Gordon model with periodicity 2 in space, namely $A^{(0)}+A^{(0)*}$, $A^{(1)}+A^{(1)*}$, and $A^{(2)}+A^{(2)*}$. The spectra of $A^{(0)}$ and of $A^{(2)}$ are both $\{z\in\mathbb{C}||z|=1\}$ and describe the possible shortest boundary conditions. The remaining task is to determine the spectrum of $A^{(1)}+A^{(1)*}$. This can be partly achieved by the algebraic Bethe ansatz for which we need to know also the off-diagonal entries of the monodromy matrix.

The off-diagonal entries satisfy $\mathcal{M}_{21}(\eta)^*=-\mathcal{M}_{12}(\eta)$, and one may write

$$\mathcal{M}_{21}(\eta)=SC(\eta)$$

where $C(\eta)\in\mathcal{C}$ and $\mathcal{S}$ is a unitary element which satisfies the following relations with the generators of $\mathcal{C}$:

$$A^{(2)}S=q^{1/2}SA^{(2)}, \tag{23}$$

$$wS=q^{-1/2}Sw, \tag{24}$$

$$uS=q^{1/2}Su, \tag{25}$$

$$vS=qSv. \tag{26}$$

Expansion of $C(\eta)$ in powers of the spectral parameter results in

$$C(\eta)=\eta^{1/2}(C^{(1/2)}-\eta C^{(3/2)})$$

with

$$C^{(1/2)}=q^{-1/2}((k^{-1/2}+k^{1/2}q^{-1/2}wu^*)w^2A^{(2)}+(k^{-1/2}u^*+k^{1/2}q^{1/2}w)v),$$

$$C^{(3/2)}=q^{-3/2}wu^*((k^{-1/2}+k^{1/2}q^{-1/2}wu^*)qA^{(2)}+(k^{-1/2}u^*+k^{1/2}q^{1/2}w)v).$$

In distinction to Ref. 5 we have formulated here the operators in terms of the generators of $\mathcal{C}$ and, as above, written u and v for P_1 and P_2, respectively. What should be kept in mind is that $C(\eta)$ has the form

$$C(\eta)=f_1(u,w,A^{(2)},\eta)+f_2(u,w,A^{(2)},\eta)v \tag{27}$$

with two continuous functions f_1 and f_2.

The starting point of the algebraic Bethe ansatz is the construction of a so-called Bethe ansatz ground state, which is a null vector of $\mathcal{M}_{21}(\eta)$. One proceeds to construct ''excited'' states by applying the ''ladder operator'' $\mathcal{M}_{12}(\eta)$ one or several times. A Yang-Baxter equation gives rise to a commutation relation between the trace $\mathcal{M}_{11}(\eta)+\mathcal{M}_{22}(\eta)$ and $\mathcal{M}_{12}(\eta)$ which may be used to determine more eigenvalues of the trace in a purely algebraic way, i.e., just upon using these relations. Their corresponding eigenvectors arise upon application of products like $\mathcal{M}_{12}(\eta_1)\cdots\mathcal{M}_{12}(\eta_n)$ to the Bethe ansatz ground state; however, due to the nature of the above commutation relations, only with a special choice of spectral parameters $\eta_1,\ldots,\eta_n$. The system of equations which determines these spectral values is called Bethe ansatz equations. For the present case, the Bethe ansatz equations and Bethe ansatz eigenvalues are stated in Theorem 1.

A. Bethe ansatz ground state

The Bethe ansatz ground state is a null vector Ω of $\mathcal{M}_{21}(\eta)$ and therefore a solution of the two equations

$$((k^{-1/2}+k^{1/2}q^{-1/2}wu^*)w^2A^{(2)}+(k^{-1/2}u^*+k^{1/2}q^{1/2}w)v)\Omega=0, \tag{28}$$

$$(w^2-q)A^{(2)}\Omega=0. \tag{29}$$

To solve for such a null vector one has to choose a representation for $\mathcal{C}$. But since w is central in $\mathcal{C}$, we may as well take a representation of the quotient algebra which is $\mathcal{C}$ modulo the ideal generated by the element w^2-q. We choose a representation with representation space $\ell^2(\mathbb{Z})$, labeled by two angles θ and δ:

$$u\cdot\phi(n)=e^{i\theta}q^{-n}\phi(n), \tag{30}$$

$$v\cdot\phi(n)=\phi(n-1), \tag{31}$$

$$A^{(2)}\cdot\phi(n)=q^{-1/2}e^{i\delta}\phi(n). \tag{32}$$

Note that this representation extends the Weyl–Schrödinger representation for the algebra generated by u and v. Upon substituting $q^{1/2}$ for w, Eq. (28) looks in the above representation, where Ω is a function over $\mathbb{Z}$, like

$$(k^{-1/2}+k^{1/2}e^{-i\theta}q^n)q^{1/2}e^{i\delta}\Omega(n)+(k^{-1/2}e^{-i\theta}q^n+k^{1/2}q)\Omega(n-1)=0. \tag{33}$$

To solve (33) there are two cases to distinguish. The first case is that, for all $n\in\mathbb{Z}$, $ke^{\pm i\theta}q^n\neq -1$ (we require real k here). Equation (33) gives then rise to a recursion relation for Ω which can easily be solved. If $\gamma/2\pi=M/N$, then the solution of this recursion relation is N-periodic up to a phase and vanishes nowhere. For general $\gamma/2\pi$ one finds that

$$\left|\frac{\Omega(m)}{\Omega(0)}\right|=\left|\frac{k+e^{i\theta}}{k+e^{i\theta}q^m}\right|,$$

from which we see that, for $k\neq 1$, Ω is what is called an extended state, it is bounded but it does not decay at infinity. Moreover, (33) gives rise to the equation for the Bloch parameters (θ,φ) at which, for rational flux, the Bethe ansatz eigenvalue corresponding to the ground state exists, or, stated differently, which label the irreducible representations in which Ω decomposes. It is that consideration–it has been carried out in a slightly different representation in Ref. 5—which leads to the result mentioned earlier, namely that for rational flux, the Bethe ansatz eigenvalues (8) occurr in irreducible representations which are labeled by $(\theta,\varphi)\in\mathcal{B}(e^{i\delta},n,k)$.

The other case, $k=1$ and $\theta=\pi$, is of even more importance in this section. In that case $\Omega(n)=\delta_{n0}$ is a solution of (33), where δ_{mn} is the Kronecker symbol, $\delta_{mn}=1$ if $n=m$, and 0 otherwise.

B. ''Excited'' Bethe ansatz eigenstates

Since $w+w^*$ and $wA^{(2)}+(wA^{(2)})^*$ are central elements in $\mathcal{C}$, they act as scalar operators in irreducible representations of $\mathcal{C}$. We mentioned already that these scalars are integrals of motion which define boundary conditions. They may also be interpreted as quantum numbers which label a super selection sector of the two-periodic sine–Gordon model. Since the ''ladder'' operator $\mathcal{M}_{12}(\eta)$ has the form $\mathcal{M}_{12}(\eta)=-C^*(\eta)S^*$ and S does not commute with the elements of $\mathcal{C}$, it cannot be implemented as an operator in the above representation and its application to a state in a specific sector will change the sector. It is convenient to combine all these sectors into one large representation. As representation space we take $\mathcal{H}=\oplus_{n\in\mathbb{Z}}\mathcal{H}_n$, where each $\mathcal{H}_n$ is a copy of $\ell^2(\mathbb{Z})$ which is preserved by $\mathcal{C}$ and plays the role of a sector. Identifying $\mathcal{H}$ with $\ell^2(\mathbb{Z}^2)$ by denoting the component of $\Psi\in\mathcal{H}$ which belongs to $\mathcal{H}_n$ by $\Psi(\cdot,n)$ we represent the algebra generated by $\mathcal{C}$ and the element S as

$$u\cdot\Psi(n_1,n_2)=e^{i\vartheta}q^{n_2/2-n_1}\Psi(n_1,n_2), \tag{34}$$

$$v\cdot\Psi(n_1,n_2)=\Psi(n_1-1,n_2), \tag{35}$$

$$w\cdot\Psi(n_1,n_2)=q^{n_2/2}\Psi(n_1,n_2), \tag{36}$$

$$A^{(2)}\cdot\Psi(n_1,n_2)=e^{i\delta}q^{-n_2/2}\Psi(n_1,n_2), \tag{37}$$

$$S\cdot\Psi(n_1,n_2)=q^{-n_1}\Psi(n_1,n_2-1). \tag{38}$$

The angles ϑ and δ label the representation. It is straightforward to check that this definition is compatible with the relations among the elements of $\mathcal{C}$ and S.

Lemma 3: Consider the operator $H(wA^{(2)}+(wA^{(2)})^*,w+w^*,1)$ *acting in the above representation with* $\vartheta=\pi-\gamma/2$. *It preserves the subspaces* $\mathcal{H}_n$ *and its restriction to* $\mathcal{H}_1$ *coincides with*

$\tilde{H}_\pi(e^{i\delta},1,1)$. *Moreover, a Bethe ansatz ground state for* $\tilde{H}_\pi(e^{i\delta},1,1)$ *may be identified with the vector* $\Omega(n_1,n_2)=\delta_{n_10}\delta_{n_21}$ *in* $\mathcal{H}$ *and any vector of the form* $\mathcal{M}_{12}(\eta_1)\cdots\mathcal{M}_{12}(\eta_{n-1})\Omega$ *vanishes outside the subset* $\{(0,n),\ldots,(n-1,n)\}\subset\mathbb{Z}^2$.

Proof: That $H(wA^{(2)}+(wA^{(2)})^*,w+w^*,1)$ preserves $\mathcal{H}_n$ is straightforwardly verified. Its restriction to $\mathcal{H}_1$ clearly coincides with $\tilde{H}(e^{i\delta},1,1)$ acting in the Weyl–Schrödinger representation with angle $\theta=\vartheta+\gamma/2$. Hence, if $\vartheta=\pi-\gamma/2$, then $\Omega(n)=\delta_{n0}$ is a Bethe ansatz ground state for $\tilde{H}(e^{i\delta},1,1)$ in that Weyl–Schrödinger representation. In representation (34)–(38) Ω has to be identified with $\Omega(n_1,n_2)=\delta_{n_10}\delta_{n_21}$, i.e., it has support $\{(0,1)\}$. The lemma follows therefore from the special form (27) of $C(\eta)$ and Eqs. (35) and (38). q.e.d.

Now consider a Bethe ansatz eigenvector in the above representation on $\mathcal{H}$. It is a vector of the form $\mathcal{M}_{12}(\eta_1)\cdots\mathcal{M}_{12}(\eta_{n-1})\Omega$, $n\in\mathbb{N}$, ''excited'' when $n>1$, where $\eta_1,\ldots,\eta_{n-1}$ satisfy the Bethe ansatz equations (7). Although there are, in general, infinitely many such vectors for $H(wA^{(2)}+(wA^{(2)})^*,w+w^*,k)$, their corresponding eigenvalues do not exhaust the spectrum of this operator. In particular, (36) and Lemma 3 imply that the eigenvalue of w on the above Bethe ansatz eigenvector $\mathcal{M}_{12}(\eta_1)\cdots\mathcal{M}_{12}(\eta_{n-1})\Omega$ has to be $q^{n/2}$. Thus, such a Bethe ansatz eigenvector is a (generalized) eigenvector of the restriction of $H(wA^{(2)}+(wA^{(2)})^*,w+w^*,k)$ to $\mathcal{H}_n$, and this restriction coincides with the operator $H(2\cos\delta,2\cos(n\gamma/2),k)=\tilde{H}(e^{i\delta},n,k)$ acting in the Weyl–Schrödinger representation with angle $\theta+(n/2)\gamma$.

Theorem 4: *Suppose that there exist, an open interval I of the real line containing 1 and a neighborhood* $U\subset S^1$ *of* δ *such that the Bethe ansatz eigenvalues (8) depend continuously on k and* δ *when* (k,δ) *is varied inside* $I\times U$. *Then, for rational values of the flux, the Bethe ansatz eigenvalues (8) for* $\tilde{H}(e^{i\delta},n,k)$, $\delta\in S^1$, $n\in\mathbb{N}$, $k\in I$, *are eigenvalues of* $\tilde{A}(e^{i\delta},n,k)$.

Proof: First, let $k=1$ and consider $H(wA^{(2)}+(wA^{(2)})^*,w+w^*,1)$ in a representation (34)–(38) with $\theta=\pi-\gamma/2$. Then, by the above lemma, Bethe ansatz eigenvectors of the form $\mathcal{M}_{12}(\eta_1)\cdots\mathcal{M}_{12}(\eta_{n-1})\Omega$ belong to $\mathcal{H}_n$ and have support contained in $\{(0,n),\ldots,(n-1,n)\}\subset\mathbb{Z}^2$. Restricting $H(wA^{(2)}+(wA^{(2)})^*,w+w^*,1)$ to $\mathcal{H}_n$ we obtain the operator $\tilde{H}(e^{i\delta},n,1)$ acting in the Weyl–Schrödinger representation with $\theta=\pi+[(n-1)/2]\gamma$, and we have just shown that the Bethe ansatz eigenvectors are among its eigenvectors belonging to the image of P_n. For $k=1$, the statement of the theorem follows therefore from Theorem 2. To extend this statement to $k\in I$ recall that we only have to show that the eigenvalues of which we want to show equality lie in the same h-bands. To see that this is the case, first note that the eigenvalues of $\tilde{A}(e^{i\delta},n,k)$ depend continuously on k and δ. By assumption, the Bethe ansatz eigenvalues depend continuously on (k,δ) when varied inside $I\times U$. Therefore, if the offset function does not have an absolute extremum at $(\theta,\varphi)\in\mathcal{B}(z,n,k)$, then the eigenvalues in question are spectral values in the interior of h-bands and hence cannot change bands under variation of k inside I. This is the case for an open dense set of values for δ. Using continuity in δ, the theorem follows. q.e.d.

Continuity of the Bethe ansatz eigenvalues in $\gamma\in S^1\backslash\{\pi\}$ would imply the above result to hold true for all values of γ.

V. CONCLUDING REMARKS

We have found n values in the spectrum of the operator $\tilde{H}(z,n,k)$ which are exact even for irrational flux and given evidence that these coincide with the eigenvalues obtained by the algebraic Bethe ansatz. Moreover, these spectral values depend analytically on the flux. Which role do they play for the model of our consideration? We conclude this article with two remarks which give partial answers to this question.

(1) We have seen that for $k\neq 1$ the Bethe ansatz ground state is an extended state. First results indicate that this remains true for the ''exited'' states as well. In fact, one can show that if $\gamma/2\pi$ is approximated by a sequence of rationals $(p_n/q_n)_n$ for which $\lim_{n\to\infty} q_n^2|\gamma/2\pi - p_n/q_n|=0$ (equivalently, its continued fraction expansion is unbounded), then $\Omega(n)$ is an almost periodic sequence provided $k\neq 1$. (The set of irrational numbers for which this is the case has full Lebesgue measure and is nowadays also referred to as Last-admissible numbers.)

From the specific form of the ''ladder'' operator $\mathcal{M}_{12}$ one sees that it acts on a state as a sum of two operators, one being multiplication by an almost periodic sequence and the other being the shift followed by multiplication with another almost periodic sequence. Hence, if a state is an almost periodic sequence, then its image under $\mathcal{M}_{12}$ is one, too. In particular, if the image does not vanish, then it is extended.

A similar phenomenon has been observed for tight binding models on codimension one quasicrystals[13] like the Fibonacci chain. Also there, exact spectral values could be obtained and their corresponding eigenstates are extended. Extended states are a peculiarity in these models in which, in general, an eigenstate is expected to be critical due to the self-similarity of the chain. The nature of the (generalized) eigenstates can be related to transport properties.[13]

(2) For the symmetric case of the QP-integral [i.e., $a=b=2\cos(n\gamma/2)$] at rational flux $\gamma/2\pi=M/N$ with $N\geqslant 2n$ it can be shown that the spectral values obtained in this article are points in the spectrum at which h-bands touch. This effect occurs also in Ref. 13. For $a\neq b$, however, we do not observe this band touching.

ACKNOWLEDGMENTS

I thank Ruedi Seiler and Alexander Dranov for stimulating discussions. This work was supported by the Deutsche Forschungsgemeinschaft.

[1] A. Bobenko, N. Kutz, and U. Pinkall, ''The discrete quantum pendulum,'' Phys. Lett. A **177**, 399–404 (1993).
[2] C. Kreft and R. Seiler, ''Models of the Hofstadter type,'' J. Math. Phys. **37**, 5207–5243 (1996).
[3] T. Richter and R. Seiler, ''Charge transport in the discretized Landau model,'' in *Discrete Integrable Geometry and Physics* (Oxford U.P., Oxford, to appear).
[4] N. Kutz, ''On the spectrum of the quantum pendulum,'' Phys. Lett. A **187**, 365–372 (1994).
[5] J. Kellendonk, N. Kutz, and R. Seiler, ''Spectra of quantum integrals,'' in *Discrete Integrable Geometry and Physics*, edited by A. Bobenko and R. Seiler (Oxford U.P., Oxford, to appear).
[6] L. A. Takhtadzhan and L. D. Faddeev, ''The quantum method of the inverse problem and the Heisenberg XYZ model,'' Russian Math. Surveys **34(5)**, 11–68 (1979).
[7] P. B. Wiegmann and A. V. Zabrodin, ''Quantum group and magnetic translations, Bethe ansatz for the Asbel-Hofstadter problem,'' Nucl. Phys. B **422**, 495–514 (1994).
[8] M. A. Rieffel, ''C^*-algebras associated with irrational rotations,'' Pac. J. Math. **93**, 415–429 (1981).
[9] M. A. Rieffel, ''Continuous fields of C^*-algebras coming from group cocycles and actions,'' Math. Ann. **283(4)**, 631–643 (1989).
[10] G. Elliot, ''Gaps in the spectrum of an almost periodic Schrödinger operator,'' C. R. Math. Acad. Sci. (Waterloo) **4**, 255–259 (1982).
[11] A. Yu. Volkov, ''Quantum volterra model,'' Phys. Lett. A **167**, 345–355 (1992).
[12] A. Yu. Volkov and L. D. Faddeev, ''Quantum inverse scattering method,'' Theor. Math. Phys. **92(2)**, 207–214 (1992).
[13] C. Sire, ''Excitation spectrum, extended states, gap closing: some exact results for codimension one quasicrystals,'' J. Phys. (France) **51**, 1569–1583 (1990).

Exact ground state of several *N*-body problems with an *N*-body potential

Avinash Khare[a)]
Institute of Physics, Sachivalaya Marg, Bhubaneswar 751005, India

(Received 31 December 1997; accepted for publication 1 February 1999)

The exact (bosonic) ground state and a class of excited states are obtained for several Calogero-type *N*-body problems in *D* dimensions when the *N* bodies are also interacting via an *N*-body potential of the form $-e^2/\sqrt{\Sigma \mathbf{r}_i^2}$. © *1999 American Institute of Physics.* [S0022-2488(99)01505-4]

I. INTRODUCTION

Over the years, the exact solutions of *N*-body problems have attracted considerable attention because of their possible relevance in statistical mechanics as well as in atomic, nuclear, and gravitational many-body problems. Whereas the exact solution of several *N*-body problems in one dimension is known by now,[1,2] a class of exact solutions including the bosonic ground state have been obtained for several Calogero-type *N*-body problems in higher dimensions, when they are also interacting via an oscillator potential.[3–7] Clearly, it is of considerable interest to discover other *N*-body problems that are either completely solvable or for which at least the ground state is exactly known.

The purpose of this paper is to show that a class of exact solutions including the bosonic ground state of all the *N*-body problems in *D* dimensions discussed in Refs. 3–7 can also be obtained when the oscillator potential is replaced by an *N*-body potential of the form

$$V(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N) = -\frac{e^2}{\sqrt{\Sigma_i \mathbf{r}_i^2}}. \tag{1}$$

We further show that one can also add an *N*-body potential of the form

$$V(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N) = -\frac{\delta^2}{\Sigma_i \mathbf{r}_i^2}, \tag{2}$$

and the problem is still analytically solvable with that the degeneracy in the bound state spectrum being much reduced. In this context, we may add that recently we have obtained[8] a class of exact solutions of the *N*-anyon problem (in two dimensions) when they are interacting via the *N*-body potential (1). Furthermore, some time ago we also obtained the complete bound state spectrum of the *N*-particle problem in one dimension when they are interacting via a variant of the above potential.[9–11] Subsequently Gurappa *et al.*[12] showed that the complete bound state spectrum can also be obtained in one dimension when an *N*-body potential of the form $\beta^2/\Sigma_{i<j}(x_i-x_j)^2$ is added either to an oscillator or to a variant of the *N*-body potential (1).

Actually, the basic idea of this paper is quite elementary. There are several quantum mechanical *N*-body problems whose solution is ultimately reduced to solving the Schrödinger equation in an appropriate radial variable for the harmonic oscillator potential. These problems remain solvable if the oscillator potential is replaced by the ''Coulomb'' potential (1). Furthermore, both the oscillator and the Coulomb problem remain solvable with the addition of the ''centrifugal'' (in-

[a)]Electronic mail: khare@iopb.res.in

verse square) potential (2). Thus, what we are really showing in this paper is that this is exactly what happens in several N-body problems. These additional terms are susceptible of being interpreted as rather artificial many-body potentials in the original physical problem. The hope is that some of these new many-body problems may be of some applicative relevance.

The plan of the paper is as follows. In Sec. II we discuss the N-body Calogero–Marchioro[3] model in D dimensions[4] and show that one can analytically obtain some exact eigenstates, including the (bosonic) ground state when the N bodies are also interacting via the potential (1). In Sec. III we consider a model of Murthy *et al.*[5,6] in two dimensions and show that exact solutions with novel correlations can also be obtained when one replaces the oscillator potential by the potential (1). We also show that the corresponding two-body problem is completely solvable. In Sec. IV we consider a Calogero-type model in D dimensions[8] which has only two-body (and no three-body) interactions, and show that one can obtain some exact eigenstates including the (bosonic) ground state when the N bodies are interacting via the N-body potential (1). In all the above cases we also show that some exact states including the bosonic ground states can still be obtained analytically if one adds the N-body potential (2) to the N-body potential (1) (or to the oscillator potential), and that the degeneracy in the discrete spectrum is then much reduced.

II. CALOGERO–MARCHIORO MODEL WITH *N*-BODY POTENTIAL

A long time ago, in an effort to generalize the original Calogero model[2] to dimensions higher than one, Calogero and Marchioro[3] considered a model in three dimensions in which the N particles are interacting via two-body and the three-body inverse square interactions as well as by pairwise harmonic oscillator potentials, and obtained some exact eigenstates, including the bosonic ground state of the system. They also mentioned that these results could be easily extended to higher-dimensional spaces. This generalization was explicitly carried out by us recently,[4] and in the special case of two dimensions we were able to obtain all the correlation functions of a many-body theory by mapping the problem to complex random matrices. In particular, we considered the following N-body Hamiltonian in D-dimensions:

$$H=-\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2+\frac{g\hbar^2}{2m}\sum_{i<j}\frac{1}{\mathbf{r}_{ij}^2}+\frac{G\hbar^2}{2m}\sum_{\substack{i,j,k\\ i\neq k,j\neq k}}\frac{\mathbf{r}_{ki}\cdot\mathbf{r}_{kj}}{\mathbf{r}_{ki}^2\mathbf{r}_{kj}^2}+\frac{m\omega^2}{4}\sum_{i}^{N}\mathbf{r}_i^2, \tag{3}$$

and obtained some eigenstates, including the bosonic ground state of the system provided G and g are related to each other [see Eq. (6) below]. Here $\mathbf{r}_i$ is the D-dimensional position vector of the ith particle and $\mathbf{r}_{ij}=\mathbf{r}_i-\mathbf{r}_j$ denotes the relative separation of the ith and jth particles while r_{ij} denotes its magnitude.

Let us consider the same many-body problem, as given by Eq. (3), but with the oscillator potential being replaced by the N-body potential (1). Throughout this paper, whenever we discuss the exact solutions with the N-body potential (1), we shall rescale all distances $\mathbf{r}_i\to\hbar^2\mathbf{r}_i/me^2$ and measure energy in units of $me^4/\hbar^2$ so that m, e, $\hbar$ are all scaled away. On the other hand, in the oscillator case we shall rescale all distances $\mathbf{r}_i\to\sqrt{\hbar/m\omega}\mathbf{r}$; and measure energy in units of $\hbar\omega$. On substituting the ansatz,

$$\psi=\left(\prod_{i<j}|\mathbf{r}_i-\mathbf{r}_j|^{\Lambda_D}\right)\phi(\rho) \tag{4}$$

in the Schrödinger equation, one obtains

$$\rho\phi''(\rho)+(2\Gamma_D+1)\phi'(\rho)+(2-2\rho|E|)\phi(\rho)=0. \tag{5}$$

Here Λ_D and Γ_D are given by

$$\Lambda_D\equiv\sqrt{G}=\tfrac{1}{2}[\sqrt{(D-2)^2+4g}-(D-2)], \tag{6}$$

$$\Gamma_D=\tfrac{1}{2}[DN-2+\Lambda_D N(N-1)], \tag{7}$$

while, here and always below,

$$\rho^2=\sum_i^N \mathbf{r}_i^2. \tag{8}$$

It is easily seen that the solution to this equation is

$$\phi(\rho)=\exp(-\sqrt{2|E|}\rho)L_{n_r}^{2\Gamma_D}(2\sqrt{2|E|}\rho), \tag{9}$$

with the corresponding energy eigenvalues being (in units of $me^4/\hbar^2$)

$$E_{n_r}=-\frac{1}{2(n_r+\Gamma_D+\frac{1}{2})^2}. \tag{10}$$

Several comments are in order at this stage.

(1) The exact solutions are obtained when G and g are related by Eq. (6). As $g\to 0$, we see from Eq. (6) that also $G\to 0$, and the wave function as given by Eqs. (4) and (9) becomes the ground state eigenfunction of the hyperspherical ''Coulomb problem'' in D dimensions without a centrifugal barrier and with Bose statistics. Thus, the situation is different from the one-dimensional problem,[2,10] where, as $g\to 0$, the eigenfunction is the ground state of the ''Coulomb problem,'' but with Fermi statistics. We shall see that in all the higher-dimensional many-body problems $(D>1)$, unlike in the one-dimensional case, as the coupling is switched off the eigenfunction corresponds to that of Bose statistics.

(2) The N-body problem is still solvable if, in addition to replacing the oscillator potential with the N-body potential (1), we also add the potential (2) to the Hamiltonian (3). In this case the ansatz (4) in the Schrödinger equation yields

$$\rho\phi''(\rho)+(2\Gamma_D+1)\phi'(\rho)+\left(2-\frac{\delta^2}{\rho}-2\rho|E|\right)\phi(\rho)=0, \tag{11}$$

so that the exact eigenstates are given by (in units of $me^4/\hbar^2$)

$$E_{n_r}=-\frac{1}{2(n_r+\gamma+\frac{1}{2})^2}, \tag{12}$$

$$\phi(\rho)=\rho^{(\gamma-\Gamma_D)}\exp(-\sqrt{2|E|}\rho)L_{n_r}^{2\gamma}(2\sqrt{2|E|}\rho), \tag{13}$$

where

$$\gamma=\sqrt{\Gamma_D^2+\delta^2}. \tag{14}$$

(3) In the same way, the N-body problem with the oscillator potential as given by Eq. (3) is also solvable if we add the N-body potential (2) to it. In particular, it is easily shown that the corresponding exact eigenstates are then

$$E_{n_r}=[2n_r+1+\gamma], \tag{15}$$

$$\psi_{n_r}=\left(\prod_{i<j}|\mathbf{r}_i-\mathbf{r}_j|^{\Lambda_D}\right)\exp\left(-\frac{1}{2}\rho^2\right)\rho^{(\gamma-\Gamma_D)}L_{n_r}^{\gamma}(\rho^2). \tag{16}$$

III. NOVEL CORRELATIONS WITH AN *N*-BODY POTENTIAL

In a recent paper, Murthy *et al.*[5,6] have proposed a model in two dimensions with two-body and three-body interactions. They were able to obtain the bosonic ground state and a class of excited states of the system by adding an external harmonic oscillator potential. The model was constructed in such a way that the solutions have a novel correlation of the form

$$X_{ij}=x_iy_j-x_jy_i \tag{17}$$

built into them. Note that X_{ij} is a pseudoscalar. Furthermore, unlike the Laughlin–Jastrow type of correlation, this correlation is not translationally invariant. We now show that the bosonic ground state and the radial excitations over it can also be obtained when the oscillator potential is replaced by the *N*-body potential (1). Furthermore, we also obtain the complete solution of the two-body problem. Notice that the two-body problem is nontrivial since the center of mass motion cannot be separated.

Following Murthy *et al.*,[5,6] we start with the *N*-particle Hamiltonian *H* (in the scaled variables), as given by

$$2H=-\sum_{i=1}^{N}\nabla_i^2+g_1\sum_{i\neq j}^{N}\frac{\mathbf{r}_j^2}{X_{ij}^2}+g_2\sum_{i\neq j\neq k}^{N}\frac{\mathbf{r}_j\cdot\mathbf{r}_k}{X_{ij}X_{ik}}-\frac{2}{\rho}, \tag{18}$$

where X_{ij} is given by Eq. (17) while g_1 and g_2 are the dimensionless coupling constants of the two-body and the three-body interactions, respectively. Note that this Hamiltonian differs from those considered in Refs. 5, 6 by the replacement of an external harmonic oscillator potential with the external *N*-body potential (1) [the last term on the rhs of Eq. (18)].

It is easily seen that

$$\psi_0(x_i,y_i)=\left(\prod_{i<j}^{N}|X_{ij}|^g\right)\exp(-\sqrt{2|E_0|}\rho) \tag{19}$$

is the exact ground state of this system with the corresponding ground-state energy being (in units of $me^4/\hbar^2$)

$$E_0=-\frac{1}{2[gN(N-1)+N-\frac{1}{2}]^2}, \tag{20}$$

provided g_1 and g_2 are related to g by

$$g_1=g(g-1),\quad g_2=g^2. \tag{21}$$

It may be noted that this ψ_0 is regular for $g\geqslant 0$ which implies that $g_1\geqslant -1/4$, $g_2\geqslant 0$. There is a neat way of proving that this is indeed the ground state, by using the method of operators.[13]

A. A class of excited states

As in the oscillator case, also in our case a class of excited states can be obtained analytically. To that end we consider the ansatz

$$\psi(x_i,y_i)=\left(\prod_{i<j}^{N}|X_{ij}|^g\right)\exp(-\alpha\rho)\phi(x_i,y_i). \tag{22}$$

On using Eqs. (18) and (22) it is easily seen that $\alpha^2=-2E$, while ϕ satisfies the eigenvalue equation

$$\left[-\frac{1}{2}\sum_i \nabla_i^2 + \frac{\alpha}{\rho}\sum_i \mathbf{r}_i\cdot\boldsymbol{\nabla}_i + \frac{A}{\rho} + g\sum_{i\neq j}\left(x_j\frac{\partial}{\partial y_i} - y_j\frac{\partial}{\partial x_i}\right)\right]\phi = 0, \tag{23}$$

where

$$A = \alpha[gN(N-1) + N - \tfrac{1}{2}] - 1. \tag{24}$$

It is worth remarking that ϕ is also an eigenstate of the total angular momentum operator i.e., $L\Phi = l\Phi$. One can now run through the steps of Ref. 5 and classify some exact solutions according to their angular momentum. In particular, one can show that one has a tower of excited states whose exact energy eigenvalues may be written in the form

$$E_{n_r,l} = -\frac{1}{2[n_r + gN(N-1) + N + |l| - \frac{1}{2}]^2}. \tag{25}$$

The existence of a tower is a general result applicable to all excited states of which the exact solutions shown above form a subset. This is easily proved by following the arguments given in Bhaduri *et al.*[6] Finally, the *N*-body problem is still solvable if we add the *N*-body potential (2) to the Hamiltonian (18).

B. The two-body problem: Complete solution

As in the oscillator case,[6] we now show that the two-body problem is integrable and exactly solvable when the two bodies are interacting via the *N*-body potential (1). Let us again emphasize that the two-body problem is nontrivial here since the center of mass cannot be separated.

The two-body Hamiltonian is given by [see Eq. (18)]

$$H = -\frac{1}{2}(\nabla_1^2 + \nabla_2^2) - \frac{1}{\sqrt{\mathbf{r}_1^2 + \mathbf{r}_2^2}} + \frac{g_1}{2}\frac{(\mathbf{r}_1^2 + \mathbf{r}_2^2)}{X^2}, \tag{26}$$

where $X = x_1y_2 - x_2y_1$. This two-body problem is best solved in the hyperspherical coordinates. To that end, let us parametrize the coordinates $\mathbf{r}_1$, $\mathbf{r}_2$ in terms of three angles and one length, (R,θ,ϕ,ψ).[14,15] In terms of the hyperspherical coordinates the Hamiltonian (26) is given by

$$H = -\frac{1}{2}\left[\frac{\partial^2}{\partial R^2} + \frac{3}{R}\frac{\partial}{\partial R} - \frac{\Lambda^2}{R^2} + \frac{2}{R}\right] + \frac{2g_1}{R^2\sin^2(2\theta)}, \tag{27}$$

where the operator Λ^2 is the Laplacian on the sphere S^3 and is given by

$$-\Lambda^2 = \frac{\partial}{\partial\theta^2} - \frac{2\sin(2\theta)}{\cos(2\theta)}\frac{\partial}{\partial\theta} + \frac{1}{\cos^2(2\theta)}\left[\frac{\partial^2}{\partial\phi^2} + 2\sin(2\theta)\frac{\partial^2}{\partial\phi\partial\psi} + \frac{\partial^2}{\partial\psi^2}\right]. \tag{28}$$

It is easily seen that if we write

$$\Psi(R,\theta,\phi,\psi) = F(R)\Phi(\theta,\phi,\psi), \tag{29}$$

then the eigenvalue equation $H\Psi = E\Psi$ separates into angular and radial equations. In particular, the angular equation is given by

$$\left[\Lambda^2 + \frac{4g_1}{\sin^2 2\theta}\right]\Phi = \beta(\beta+2)\Phi, \tag{30}$$

where $\beta \geqslant -1$, while the radial equation is given by

$$F''(R)+\frac{3}{R}F'(R)+\left(2E+\frac{2}{R}-\frac{\beta(\beta+2)}{R^2}\right)F(R)=0. \tag{31}$$

The radial equation is easily solved, yielding

$$F(R)=R^{\beta}\exp(-\sqrt{2|E|}R)M(a,b;2\sqrt{2|E|}R), \tag{32}$$

where $b=2\beta+3$, $a=\frac{3}{2}+\beta-1/\sqrt{2|E|}$, and $M(a,b;x)$ is the confluent hypergeometric function. Demanding $a=-n_r$, where n_r is a non-negative integer, yields the bound state energy eigenvalues

$$E=-\frac{1}{2(n_r+\beta+\frac{3}{2})^2}. \tag{33}$$

The tower structure of the eigenvalues built on the radial excitations of the ground state is obvious from this formula.

It must be emphasized here that β is still unknown and has to be obtained by solving the angular equation (30). We now note that the angular equation is, in fact, identical in the oscillator and our case, and it has been analyzed in great detail by Bhaduri *et al.*[6] We can therefore borrow their results and draw conclusions about the value of β and hence the spectrum, as given by Eq. (33).

IV. CALOGERO-TYPE MODELS IN HIGHER DIMENSIONS WITH *N*-BODY INTERACTION

In Sec II, we have considered one possible generalization of the Calogero-type models in higher dimensions. The key point there was to have a long-ranged three-body interaction term. This is in addition to the long-ranged two-body interaction term that is also present in one dimension. Only then was it possible to obtain a class of exact solutions including the bosonic ground state. Another possible generalization of the Calogero model to higher dimensions was considered recently by Ghosh.[7] In a remarkable paper[7] he introduced two models with purely two-body long-ranged interactions and in both cases he was able to obtain the exact bosonic ground state and radial excitations over it. In this section we show that the exact ground state and radial excitations over it can also be obtained if the oscillator potential in Ref. 7 is replaced by the N-body potential (1). Following Ghosh,[7] let us consider the Hamiltonian

$$H=-\frac{1}{2}\sum_{k=1}^{N}\nabla_k^2-\frac{1}{\rho}+V_1(\beta)+V_2(\beta)+W_3(\beta), \tag{34}$$

where

$$V_1(\beta)=\frac{\beta^2}{2}g(g-1)\sum_{k\neq j}\frac{|\mathbf{r}_k|^{2(\beta-1)}}{(|\mathbf{r}_k|^{\beta}-(|\mathbf{r}_j|^{\beta})^2},$$

$$V_2(\beta)=\frac{g\beta}{2}(D+\beta-2)\sum_{k\neq j}\frac{|\mathbf{r}_k|^{(\beta-2)}}{(|\mathbf{r}_k|^{\beta}-|\mathbf{r}_j|^{\beta})}, \tag{35}$$

$$W_3(\beta)=\frac{\beta^2}{2}g(g-1)\sum_{i\neq j\neq k}\frac{|\mathbf{r}_i|^{2(\beta-1)}}{(|\mathbf{r}_i|^{\beta}-|\mathbf{r}_j|^{\beta})(|\mathbf{r}_i|^{\beta}-|\mathbf{r}_k|^{\beta})},$$

with g a dimensionless constant. Note that for β=1 and β=2 the three-body interaction term W_3 vanishes. Also note that the Hamiltonian as given by Eq. (34) differs from Ref. 7 by the replacement of an external harmonic oscillator potential with the external N-body potential (1) [the second term on the rhs of (34)].

It is straightforward to obtain the bosonic ground state and the radial excitation spectrum over it. In particular, it is easily seen that the exact eigenfunctions are

$$\psi_{n_r}=\left(\prod_{i<j}(|\mathbf{r}_i|^\beta-|\mathbf{r}_j|^\beta)^g\right)M(a=-n_r,b;2\alpha\rho)\exp(-\sqrt{2|E|\rho}),\tag{36}$$

with the corresponding eigenvalues being

$$E_{n_r}=-\frac{1}{2\left[n_r+\frac{ND}{2}+\frac{g\beta}{2}N(N-1)-1\right]^2}.\tag{37}$$

Here $a=A/\alpha$, $b=2(A+1)/\alpha$ while $E=-\alpha^2/2$ and $A=\alpha[ND/2+g\beta N(N-1)/2-1]$. Similar exact solutions are also obtained if one adds the N-body potential (2) to the Hamiltonian (34).

V. SUMMARY AND OPEN PROBLEMS

In this paper, we have discussed several N-body problems in two and higher dimensions and in every case we have shown that if exact eigenstates including the N-boson ground state can be obtained when the N-particles are also interacting by an oscillator potential, then similar exact eigenstates can also be obtained when the oscillator potential is replaced by the ''Coulomblike'' N-body potential (1). Furthermore, we have shown that in both cases one can also add an inverse square N-body potential, and the problems are still analytically solvable while the degeneracy in the spectrum is much reduced. It is clearly of interest to examine other solvable many-body problems with an external harmonic oscillator potential and see if exact results can still be obtained with the replacement of the oscillator potential by the N-body potential (1).

[1] See, for example, D. C. Mattis, *The Many Body Problem: An Encyclopedia of Exactly Solved Models in One Dimension* (World Scientific, Singapore, 1995).
[2] F. Calogero, J. Math. Phys. **10**, 2191, 2197 (1969); **12**, 419 (1971). For an exhaustive review of various exactly solvable N-body problems see M. A. Olshanetsky and A. M. Perelomov, Phys. Rep. **71**, 314 (1981); **94**, 315 (1983).
[3] F. Calogero and C. Marchioro, J. Math. Phys. **14**, 182 (1973).
[4] A. Khare and K. Ray, Phys. Lett. A **230**, 139 (1997).
[5] M. V. N. Murthy, R. K. Bhaduri, and D. Sen, Phys. Rev. Lett. **76**, 4103 (1996).
[6] R. K. Bhaduri, A. Khare, J. Law, M. V. N. Murthy, and D. Sen, J. Phys. A **30**, 2557 (1997).
[7] P. K. Ghosh, Phys. Lett. A **229**, 203 (1997).
[8] A. Khare, Phys. Lett. A **221**, 365 (1996). For the solution of the N-anyon problem in an oscillator potential see, for example, A. Khare, *Fractional Statistics and Quantum Theory* (World Scientific, Singapore, 1997), Chap. 3 and references therein.
[9] A. Khare, J. Phys. A **29**, L45 (1996).
[10] F. Calogero, J. Phys. A **29**, 6455 (1996).
[11] A. Khare, J. Phys. A **29**, 6459 (1996).
[12] N. Gurappa, C. Nagaraja Kumar, and P. K. Panigrahi, Mod. Phys. Lett. A **11**, 1737 (1996).
[13] A. P. Polychronakos, Phys. Rev. Lett. **69**, 703 (1992).
[14] J. E. Kilpatrick and S. Y. Larsen, Few-Body Syst. **3**, 75 (1987).
[15] A. Khare and J. McCabe, Phys. Lett. B **269**, 330 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Eigenvalue problems of Ginzburg–Landau operator in bounded domains

Kening Lu
Department of Mathematics, Brigham Young University, Provo, Utah 84602

Xing-Bin Pan
Center for Mathematical Sciences, Zhejiang University, Hangzhou 310027, People's Republic of China

(Received 21 September 1998; accepted for publication 5 March 1999)

In this paper we study the eigenvalue problems for the Ginzburg–Landau operator with a large parameter in bounded domains in $\mathbb{R}^2$ under gauge invariant boundary conditions. The estimates for the eigenvalues are obtained and the asymptotic behavior of the associated eigenfunctions is discussed. These results play a key role in estimating the critical magnetic field in the mathematical theory of superconductivity. © *1999 American Institute of Physics.* [S0022-2488(99)02806-6]

I. INTRODUCTION AND STATEMENT OF MAIN RESULTS

This paper is devoted to the asymptotic estimates, for large parameter σ, of the *first eigenvalue* $\mu(\sigma A)$ and the associated eigenfunctions of the *Ginzburg–Landau operator* $-\nabla^2_{\sigma A}$ in a smooth bounded domain Ω in $\mathbb{R}^2$. Given a real vector field $A=(A^1,A^2)$, the Ginzburg–Landau operator $-\nabla^2_A$ associated with A is defined by

$$-\nabla^2_A\psi=-\nabla_A\cdot(\nabla_A\psi)=-\nabla\psi+i[2A\cdot\nabla\psi+\psi\,\mathrm{div}\,A]+|A|^2\psi,$$

where $i=\sqrt{-1}$. We denote $\nabla_A\psi=\nabla\psi-i\psi A$, $\mathrm{curl}\,A=\partial_1A^2-\partial_2A^1$, and $\mathrm{curl}^2A=(\partial_2(\mathrm{curl}\,A),-\partial_1(\mathrm{curl}\,A))$, here $\partial_j=\partial/\partial x_j$.

Let $\mu=\mu(A)$ be the first eigenvalue of the following problem:

$$\begin{aligned}-\nabla^2_A\psi&=\mu\psi \quad \text{in } \Omega,\\ (\nabla_A\psi)\cdot\nu+\gamma\psi&=0 \quad \text{on } \partial\Omega,\end{aligned} \tag{1.1}$$

where ψ is a complex-valued function, ν is the unit outer normal to $\partial\Omega$, and $\gamma\geqslant 0$ is a given constant. Then,

$$\mu(A)=\inf_{\psi\in W^{1,2}(\Omega)}\frac{\int_\Omega|\nabla_A\psi|^2dx+\gamma\int_{\partial\Omega}|\psi|^2\,ds}{\int_\Omega|\psi|^2\,dx}. \tag{1.2}$$

It is well-known that the Ginzburg–Landau operator has the *gauge invariance* property

$$\nabla_{A+\nabla\chi}(e^{i\chi}\psi)=e^{i\chi}\nabla_A\psi,\quad \nabla^2_{A+\nabla\chi}(e^{i\chi}\psi)=e^{i\chi}\nabla^2_A\psi$$

for every real smooth function χ. The equation and the boundary condition in (1.1) as well as the functional in (1.2) are invariant under the gauge transformation $A\to A+\nabla_\chi$, $\psi\to e^{i\chi}\psi$. Therefore, $\mu(A+\nabla_\chi)=\mu(A)$. By a gauge transformation if necessary, we may assume

$$\mathrm{div}\,A=0 \quad \text{in } \Omega,\quad A\cdot\nu=0 \quad \text{on } \partial\Omega.$$

Our main result is the following

Theorem 1: *There exists a universal constant* β_0, $0<\beta_0<1$, *such that for all* $A\in C^2(\overline{\Omega})$

$$\lim_{\sigma\to\infty}\frac{\mu(\sigma A)}{|\sigma|}=\min\{\min_{x\in\Omega}|\mathrm{curl}\,A(x)|,\beta_0\min_{x\in\partial\Omega}|\mathrm{curl}\,A(x)|\}. \tag{1.3}$$

Remark 1.1: As a consequence of Theorem 1 we see that, if $\mathrm{curl}\,A(x)\equiv H$, *a nonzero constant, then*

$$\lim_{\sigma\to\infty}\frac{\mu(\sigma A)}{|\sigma|}=\beta_0|H|.$$

The universal constant β_0 stated in Theorem 1 is the number $\beta(1)$ given in Lemma 2.2. We have an estimate for β_0:

$$0.5<\beta_0<0.76,$$

see Ref. 1. It has been expected that $\beta_0\simeq 0.59$, see for instance Refs. 2 and 3. If $\mathrm{curl}\,A$ vanishes at some points, the estimates can be greatly improved, see Sec. VI. It is interesting to see that the distribution of minimum points of $|\mathrm{curl}\,A|$ determines the magnitude of $\mu(\sigma A)$ and the location of the concentration points of the eigenfunctions for large σ. This is partially due to the gauge invariance of the Ginzburg–Landau operator and due to the invariance of $\mathrm{curl}\,A$ under the gauge transformations.

To prove Theorem 1 we shall establish two estimates for $\mu(\sigma A)$, the upper bound estimate (given in Sec. VI) and the lower bound estimate (given in Sec. VII). The gauge invariance of the Ginzburg–Landau operator, the local decomposition formula of vector fields obtained in Sec. III, and the results obtained in Ref. 4 concerning the eigenvalue problems of Ginzburg–Landau operator in the entire plane and on the half plane will play essential roles to obtain these estimates. To derive the lower bound estimate we also need to show the local convergence, as $\sigma\to\infty$, of the rescaled eigenfunctions (after a series of gauge transformations). Since the eigenfunctions may concentrate either in the interior of Ω or at the boundary, both interior and boundary *a priori* estimates established in Secs. IV and V are needed to obtain the local convergence. We mention that most of the estimates given in this paper are gauge invariant. As a by-product, the asymptotic behavior of the eigenfunctions as σ goes to ∞ will also be obtained.

The technical difficulty in our problem comes from the boundary effects, which is our main concern in this paper. One may see in Sec. VI that when the eigenfunctions concentrate in the interior of Ω, the limiting equation obtained after rescaling is an eigenvalue problem in the entire plane $\mathbb{R}^2$, see (2.3). All the eigenvalues of (2.3) have been obtained in Ref. 4. However, when the concentration happens at the boundary, very technical analyses are required to get the boundary estimates and to prove the local convergence of the rescaled eigenfunctions near the boundary. In this case, the limiting equation is an eigenvalue problem in the half plane $\partial\mathbb{R}^2_+$, see (2.5). The first eigenvalue $\beta(h)$ of (2.5) was obtained in Ref. 4 after lengthy analyses, which is the difficult part of Ref. 4. Comparing Lemma 2.1 with Lemma 2.2 in Sec. II, one may see the significant difference between the problems in the domain without or with boundary.

The motivation to study such type eigenvalue problems is to estimate the value of the upper critical magnetic field at which superconductivity can nucleate.

In the mathematical theory of superconductivity, the following Ginzburg–Landau equation for (ψ, A) was proposed as a macroscopic model (see Ref. 5)

$$\begin{aligned}-\nabla^2_{\kappa A}\psi&=\kappa^2(1-|\psi|^2)\psi,\\ \mathrm{curl}^2 A&=-\frac{i}{2\kappa}(\overline{\psi}\nabla\psi-\psi\nabla\overline{\psi})-|\psi|^2A+\mathrm{curl}\,H\quad\text{in}\ \ \Omega.\end{aligned} \tag{1.4}$$

Here Ω is the region occupied by the superconducting specimen, ψ is a complex-valued function called *order parameter*, A is a real vector field called *magnetic potential*, H is the applied magnetic field, κ is the so-called *Ginzburg–Landau parameter*. The natural boundary conditions for a superconductor-other material junction are (see Ref. 6)

$$(\nabla_{\kappa A}\psi)\cdot\nu+\gamma\psi=0, \quad (\mathrm{curl}\,A-H)\times\nu=0 \quad \text{on } \partial\Omega, \tag{1.5}$$

where ν is the unit out-normal vector at the boundary of Ω and γ is a positive constant.

It is well-known that a superconductor placed in an applied magnetic field may change its phase when the field varies. Consider a *spatially homogeneous field*. If the field is sufficiently strong, it penetrates through the entire sample and the superconductor is in a normal state. As the field is gradually reduced to a certain value H_{C_3} called the *upper critical field*, the nucleation of superconductivity at surface occurs. If the field is further reduced to another value H_{C_2}, the nucleation in the interior occurs. It is important in both theory and applications to estimate the values of the critical fields, especially for type 2 superconductors with large value of κ.

The physicists Saint-James and De Gennes were the first to study the surface nucleation phenomenon for semi-infinite superconductor occupying the half space (see Ref. 2). The most amazing result they obtained was the relation $H_{C_3}/H_{C_2}=1/0.59$. The argument for this relation was nontrivial, even though they studied only the superconductor which occupies the half space and is subjected to a *spatially homogeneous* applied magnetic field.

We have been interested in estimating the value of the upper critical field for superconducting specimen occupying an *arbitrary* bounded smooth domain. In Ref. 1, to get such estimate, we considered the applied field having the form $H=\sigma H_0$ and estimated the maximal value of σ, say σ^*, so that under the applied field $\sigma^* H_0$ the nucleation of superconductivity occurs. Choosing a vector field F so that $\mathrm{curl}\,F=H_0$, we found that when κ is large, the value of σ^* is close to the number σ_* for which $\mu(\sigma_*\kappa F)=\kappa^2$. This led us to study the asymptotic estimates of $\mu(\sigma F)$ for large value of σ. In Ref. 1, by using the results in this paper, we obtained the asymptotic estimate for H_{C_3} for large κ and the location of nucleation of superconductivity.

There have been many recent works on the mathematical theory of superconductivity, see Refs. 3, 7–19, and the references therein. The works[3] by Chapman,[7] by Bauman, Phillips, and Tang, and by Bernoff and Sternberg[10] are closely related to our present paper, while Refs. 7 and 10 were found after this work had been completed. In Ref. 3, Chapman studied the half-plane problem on H_{C_3} by using formal mathematical analysis. In Ref. 7, Bauman, Phillips, and Tang rigorously estimated H_{C_3} and found the location of nucleation for a sample occupying a cylinder with two-dimensional cross section consisting of a disk. The sample is adjacent to a vacuum and is subject to a homogeneous applied magnetic field pointing in the axial direction. From the bifurcation point of view, they studied small solutions bifurcating from the eigenfunctions. In Ref. 10, Bernoff and Sternberg considered a sample occupying an infinite cylinder with two-dimensional cross section consisting of an arbitrary simply connected smooth bounded region in $\mathbb{R}^2$. The sample is adjacent to a vacuum and is subject to a homogeneous applied magnetic field pointing in the axial direction. They estimated H_{C_3} and found the location of nucleation by using formal asymptotic expansions. In this paper we study eigenvalue problems in bounded smooth domains with nonhomogeneous applied magnetic fields under the boundary conditions for a superconductor–other material junction. The result obtained in this paper was used in Ref. 1 to obtain rigorously estimates for H_{C_3} and locations of nucleation for a cylindrical sample which is placed in an applied magnetic field being parallel to the lateral surface but not necessarily spatially homogeneous and is adjacent to other material.

II. PRELIMINARIES

In this section we give some basic lemmas which will be used later to establish our main result. Throughout this paper, we let

$$\omega(x)=\tfrac{1}{2}(-x_2,x_1). \tag{2.1}$$

Note that $\operatorname{curl}\omega=1$ and $\operatorname{div}\omega=0$. Denote, for a nonzero real number h,

$$\alpha(h)=\inf_{\psi\in\mathcal{W}(\mathbb{R}^2)}\frac{\int_{\mathbb{R}^2}|\nabla_{h\omega}\psi|^2\,dx}{\int_{\mathbb{R}^2}|\psi|^2\,dx}, \tag{2.2}$$

where $\mathcal{W}(\mathbb{R}^2)=W^{1,2}_{\mathrm{loc}}(\mathbb{R}^2)\cap L^2(\mathbb{R}^2)$. Clearly, the minimizers are the L^2 eigenfunctions of the following problem associated with $\alpha=\alpha(h)$:

$$-\nabla^2_{h\omega}\psi=\alpha\psi \quad \text{in } \mathbb{R}^2. \tag{2.3}$$

Let

$$\beta(h)=\inf_{\psi\in\mathcal{W}(\mathbb{R}^2_+)}\frac{\int_{\mathbb{R}^2_+}|\nabla_{h\omega}\psi|^2\,dx}{\int_{\mathbb{R}^2_+}|\psi|^2\,dx}, \tag{2.4}$$

where $\mathcal{W}(\mathbb{R}^2_+)=W^{1,2}_{\mathrm{loc}}(\mathbb{R}^2_+)\cap L^2(\mathbb{R}^2_+)$. The associated eigenvalue problem is

$$\begin{aligned}-\nabla^2_{h\omega}\psi&=\beta\psi \quad \text{in } \mathbb{R}^2_+,\\ (\nabla_{h\omega}\psi)\cdot\nu&=0 \quad \text{on } \partial\mathbb{R}^2_+,\end{aligned} \tag{2.5}$$

where $\nu(x)=(0,-1)$ is the outer normal to $\mathbb{R}^2_+$.

Lemma 2.1: (Ref. 4). *For every $h\neq 0$, $\alpha(h)=|h|$. The associated eigenfunctions are given by*

$$\psi(x)=\begin{cases} f(x)\exp(-|h|r^2/4) & \text{if } h>0\\ \overline{f(x)}\exp(-|h|r^2/4) & \text{if } h<0,\end{cases}$$

where $r=|x|$, $f(x)$ is any function analytic in $\mathbb{R}^2$ such that $f(x)\exp(-|h|r^2/4)\in L^2(\mathbb{R}^2)$. For all $\alpha<\alpha(h)$, (2.3) *has no nontrivial bounded solution.* □

Lemma 2.2: (Ref. 4). *There exists a positive constant β_0, $0<\beta_0<1-1/\sqrt{2e\pi}$, such that $\beta(h)=\beta_0|h|$. For all $h\neq 0$, $\beta(h)$ is not achieved in $\mathcal{W}(\mathbb{R}^2_+)$, i.e., there is no L^2 eigenfunction associated with $\beta(h)$. For all $\beta<\beta(h)$,* (2.5) *has no nontrivial bounded solution.* □

III. LOCAL DECOMPOSITION OF VECTOR FIELDS

In the proof of the convergence of the rescaled eigenfunctions in later sections, we use the gauge transformations frequently. Thus, we need to decompose a vector field into a gradient part and a curl part near a given point P. When P is an interior point, this decomposition follows directly from the Taylor expansion (see Lemma 3.1). When P is a boundary point, we need to decompose the vector field in new coordinates which straighten a portion of boundary (see Lemma 3.2).

Let $A(x)=(A^1(x),A^2(x))\in C^2(\overline{B_R})$ and denote

$$a^i_j=\frac{\partial A^i}{\partial x_j}(0),\quad a^i_{jk}=\frac{\partial^2 A^i}{\partial x_j\partial x_k}(0),\quad a^1=A^1(0),\quad a^2(0)=A^2(0).$$

Let $H(x)=\operatorname{curl}A(x)$. Then, $\underline{\operatorname{curl}}^2A(x)=(\partial_2H,-\partial_1H)$.

Lemma 3.1: Let $A\in C^2(\overline{B_R})$. Then,

$$A(x)=A(0)+\nabla\xi(x)+\nabla\zeta(x)+\operatorname{curl}A(0)\,\omega(x)-\tfrac{1}{2}|x|^2\underline{\operatorname{curl}}^2A(0)+D(x), \tag{3.1}$$

where

$$\xi(x)=\tfrac{1}{2}[a_1^1x_1^2+(a_2^1+a_1^2)x_1x_2+a_2^2x_2^2],$$
$$\zeta(x)=\tfrac{1}{6}[c_1x_1^3+3c_2x_1^2x_2+3c_3x_1x_2^2+c_4x_2^3] \tag{3.2}$$

with

$$c_1=a_{11}^1+\partial_2H(0),\quad c_2=a_{12}^1,\quad c_3=a_{12}^2,\quad c_4=a_{22}^2-\partial_1H(0),$$

and $|D(x)|=o(|x|^2)$ *as* $x\to 0$. *If* $A\in C^3(\overline{B_R})$, *then* $|D(x)|\leqslant C(R)|x|^3$ *in* B_R. □

In the following we assume that Ω is a smooth (say, C^k for some $k\geqslant 3$) bounded domain in $\mathbb{R}^2$ and $0\in\partial\Omega$. Then, $\partial\Omega$ consists of a finite number of simple closed C^k curves disintersecting with each other. Every component Γ of $\partial\Omega$ can be represented as $z=z(s)$, where s is the arclength of Γ, and $\tau(s)=(\tau_1,\tau_2)=z'(s)$ is the unit tangent vector. Let $\nu(s)=(\nu_1,\nu_2)$ be the unit outer normal. We choose the positive direction of Γ in such a way that the orientation of (ν,τ) is coincident with the orientation of the x_1x_2 coordinates. Then, $\tau_1=-\nu_2$, $\tau_2=\nu_1$. From the Frenet formula we have

$$\tau'=-\kappa_r\nu,\quad \tau''=-\kappa_r'\nu-\kappa_r^2\tau,\quad \nu'=\kappa_r\tau,\quad \nu''=\kappa_r'\tau-\kappa_r^2\nu, \tag{3.3}$$

where κ_r is the relative curvature of Γ under the given orientation. Obviously, there exists a positive constant $\mu_0=\mu_0(\Omega)$ such that $|\kappa_r|\leqslant 1/\mu_0$ on $\partial\Omega$.

Fix $0<\mu<\mu_0$. Denote by $d(x)=\mathrm{dist}(x,\partial\Omega)$ the distance function, and denote $\Omega(\mu)=\{x\in\overline{\Omega}:d(x)<\mu\}$. Then, $d\in C^{k-1}(\overline{\Omega(\mu_0)})$. For every $x\in\Omega(\mu_0)$ there exists a unique point $z=z(x)\in\partial\Omega$ such that $x=z-d(x)\nu(z)$, $\nabla d(x)=-\nu(z)$. The mapping

$$x=\mathcal{F}(s,t)=z(s)-t\nu(s) \tag{3.4}$$

determines a C^1 transformation of coordinates. Set

$$g(s,t)=|\det D\mathcal{F}|=|\mathcal{F}_s\times\mathcal{F}_t|=1-t\kappa_r(s).$$

After rotating the coordinate system we may assume $\tau(0)=(1,0)$, $\nu(0)=(0,-1)$. Denote $\mathbf{e}_1=\tau$, $\mathbf{e}_2=-\nu$, $y_1=s$, $y_2=t$, $y=(y_1,y_2)$. y is the new coordinate straightening the boundary. Using (3.3) we get

$$\mathbf{e}_1(y)=\tau(y_1)=\begin{pmatrix}1\\0\end{pmatrix}+\begin{pmatrix}0\\ \kappa_r(0)y_1\end{pmatrix}+\frac{1}{2}y_1^2\begin{pmatrix}-\kappa_r^2(0)\\ \kappa_r'(0)\end{pmatrix}+O(|y_1|^3),$$

$$\mathbf{e}_2(y)=-\nu(y_1)=\begin{pmatrix}0\\1\end{pmatrix}-\begin{pmatrix}\kappa_r(0)y_1\\ 0\end{pmatrix}-\frac{1}{2}y_1^2\begin{pmatrix}\kappa_r'(0)\\ \kappa_r^2(0)\end{pmatrix}+O(|y_1|^3),$$

$$\mathcal{F}(y)=y+\frac{\kappa_r(0)}{2}\begin{pmatrix}-2y_1y_2\\ y_1^2\end{pmatrix}+O(|y|^3)\quad \text{as } y\to 0, \tag{3.5}$$

$$g(y)=1-\kappa_r(0)y_2-\kappa_r'(0)y_1y_2+O(|y|^3).$$

Denote the inverse map of $\mathcal{F}$ by $\mathcal{G}(x)$. At the point $x=\mathcal{F}(y)$ we have

$$D\mathcal{G}(x)=\begin{pmatrix}\mathcal{G}_1^1 & \mathcal{G}_2^1\\ \mathcal{G}_1^2 & \mathcal{G}_2^2\end{pmatrix}=\frac{1}{1-y_2\kappa_r(y_1)}\begin{pmatrix}-\nu_2 & \nu_1\\ -(1-y_2\kappa_r(y_1))\tau_2 & (1-y_2\kappa_r(y_1))\tau_1\end{pmatrix}. \tag{3.6}$$

For a given vector field $A(x)$ we define a new vector field $\mathbf{a}(y)$ associated with $A(x)$ by

$$\mathbf{a}(y)=\mathbf{a}^1(y)\mathbf{e}_1+\mathbf{a}^2(y)\mathbf{e}_2, \tag{3.7}$$

where

$$\mathbf{a}^1(y)=g(y)A(\mathcal{F}(y))\cdot\mathbf{e}_1(y),\quad \mathbf{a}^2(y)=A(\mathcal{F}(y))\cdot\mathbf{e}_2(y). \tag{3.8}$$

Then,

$$\begin{aligned}\mathbf{a}^1(y)=&a^1+[a_1^1+\kappa_r(0)a^2]y_1+[a_2^1-\kappa_r(0)a^1]y_2+\tfrac{1}{2}\{[a_{11}^1+\kappa_r(0)a_2^1+2\kappa_r(0)a_1^2\\&-\kappa_r(0)^2a^1+\kappa_r'(0)a^2]y_1^2+2[a_{12}^1-2\kappa_r(0)a_1^1+\kappa_r(0)a_2^2-\kappa_r^2(0)a^2-\kappa_r'(0)a^1]y_1y_2\\&+[a_{22}^1-2\kappa_r(0)a_2^1]y_2^2\}+o(|y|^2),\end{aligned}$$

$$\begin{aligned}\mathbf{a}^2(y)=&a^2+(a_1^2-\kappa_r(0)a^1)y_1+a_2^2y_2+\tfrac{1}{2}\{[a_{11}^2+\kappa_r(0)a_2^2-2\kappa_r(0)a_1^1-\kappa_r^2(0)a^2-\kappa_r'(0)a^1]y_1^2\\&+2[a_{12}^2-\kappa_r(0)a_1^2-\kappa_r(0)a_2^1]y_1y_2+a_{22}^2y_2^2\}+o(|y|^2).\end{aligned}$$

Summarizing the above discussion, we obtain

Lemma 3.2: Let Ω be a smooth domain in $\mathbb{R}^2$ with $0\in\partial\Omega$. Assume that $A\in C^2(\overline{\Omega\cap\mathcal{F}(B_R)})$. Then, in the new coordinates y straightening the boundary, the vector field $\mathbf{a}(y)$ associated with $A(x)$ has the following decomposition for $y\in B_R$:

$$\begin{aligned}(\mathbf{a}^1(y),\mathbf{a}^2(y))=&A(0)+\nabla\tilde{\xi}(y)+\nabla\tilde{\zeta}(y)+\operatorname{curl}A(0)\omega(y)\\&-\frac{|y|^2}{2}[\operatorname{curl}^2A(0)-\kappa_r(0)\operatorname{curl}A(0)\tau(0)]+\tilde{D}(y),\end{aligned} \tag{3.9}$$

where

$$\tilde{\xi}(y)=\tfrac{1}{2}[(a_1^1+\kappa_r(0)a^2)y_1^2+(a_2^1+a_1^2-2\kappa_r(0)a^1)y_1y_2+a_2^2y_2^2],$$

$$\tilde{\zeta}(y)=\tfrac{1}{6}[\tilde{c}_1y_1^3+3\tilde{c}_2y_1^2y_2+3\tilde{c}_3y_1y_2^2+\tilde{c}_4y_2^3]$$

with

$$\tilde{c}_1=a_{11}^1-a_{22}^1+a_{12}^2+\kappa_r(0)(a_1^2+a_2^1)-\kappa_r^2(0)a^1+\kappa_r'(0)a^2,$$

$$\tilde{c}_2=a_{12}^1+\kappa_r(0)(a_2^2-2a_1^1)-\kappa_r^2(0)a^2-\kappa_r'(0)a^1,$$

$$\tilde{c}_3=a_{12}^2-\kappa_r(0)(a_1^2+a_2^1),$$

$$\tilde{c}_4=a_{12}^1-a_{11}^2+a_{22}^2.$$

$|\tilde{D}(y)|=o(|y|^2)$ as $y\to0$. *If $A\in C^3(\overline{\Omega\cap\mathcal{F}(B_R)})$, then $|\tilde{D}(y)|\le C(R)|y|^3$ in B_R^+.* □

Note that in (3.9) $\nabla\tilde{\xi}(y)=(\partial_{y1}\tilde{\xi},\partial_{y2}\tilde{\xi})$. In the following we denote

$$\nabla_yf=\left(\frac{\partial f}{\partial y_1}\right)\mathbf{e}_1+\left(\frac{\partial f}{\partial y_2}\right)f\mathbf{e}_2. \tag{3.10}$$

From (3.7) and (3.10), we can write (3.9) as follows:

$$\begin{aligned}\mathbf{a}(y)=&A(0)+\nabla_y\tilde{\xi}(y)+\nabla_y\tilde{\zeta}(y)+\operatorname{curl}A(0)\tilde{\omega}(y)\\&-\frac{|y|^2}{2}[\operatorname{curl}^2A(0)-\kappa_r(0)\operatorname{curl}A(0)\tau(0)]+\tilde{D}(y),\end{aligned} \tag{3.11}$$

where $\tilde{\omega}(y) = -(y_2/2)\mathbf{e}_1 + (y_1/2)\mathbf{e}_2$. The decomposition in the form of (3.11) is more closely related to the gauge invariance of the operators involving the vector **a**, and will be used often in later sections.

IV. INTERIOR ESTIMATES

In this section we shall derive *a priori* interior estimates for the solutions ψ of the equation

$$\nabla_A^2 \psi = g \quad \text{in} \ \ \Omega, \tag{4.1}$$

where the vector field A and the function g are given. We shall establish the gauge invariant estimates which depend on $\operatorname{curl} A$ instead of A itself.

Theorem 4.1: *Assume that ψ is a smooth solution of Eq. (4.1) and $\operatorname{curl} A \in L^2(\Omega)$. Then, for any compact subset K of Ω, there exists a constant C depending only on Ω and K such that*

$$\begin{aligned} \| |\nabla_A \psi| \|^2_{H^1(K)} \leqslant \sum_{j,k} \|\nabla_{A^j}\nabla_{A^k}\psi\|^2_{L^2(K)} \leqslant 2\|\psi \operatorname{curl} A\|^2_{L^2(\Omega)} + 6\|g\|^2_{L^2(\Omega)} + C[1 + \|\operatorname{curl} A\|^4_{L^2(\Omega)}] \\ \times [\|\nabla_A \psi\|^2_{L^2(\Omega)} + \|\psi\|^2_{L^2(\Omega)}]. \end{aligned} \tag{4.2}$$

Before proving Theorem 4.1 we mention that $\|\nabla_A \psi\|_{L^2}$ can be controlled by $\|\psi\|_{L^2}$, as shown in the following

Lemma 4.2: Assume that $\psi \in W^{1,2}_{\text{loc}}(\Omega)$ is a weak solution of Eq. (4.1) for $g \in L^2_{\text{loc}}(\Omega)$. Then, for every $R > 0$ such that $B_{2R} \subset \Omega$ we have

$$\int_{B_R} |\nabla_A \psi|^2 \, dx \leqslant 2 \int_{B_{2R}} |g\psi| dx + \frac{16}{R^2} \int_{R_{2R}} |\psi|^2 \, dx.$$

Proof: Let η be a smooth cutoff function supported in B_{2R} such that $\eta = 1$ on B_R and $|\nabla \eta| \leqslant 2/R$. Multiplying Eq. (4.1) by $\eta^2 \psi$ and integrating by parts we get the conclusion. □

For convenience we denote $F_{j,k}\psi = (\nabla_{A^j}\nabla_{A^k} - \nabla_{A^k}\nabla_{A^j})\psi$, $\nabla_{A^j}\psi = (\partial_j - iA^j)\psi$ and $(\phi,\psi) = \int_\Omega \phi \overline{\psi} \, dx$.

Proposition 4:3: Let $A \in C^1(\overline{\Omega})$ and $\psi \in C^2(\overline{\Omega})$. Then,

$$\sum_{j,k} \|\nabla_{A^j}\nabla_{A^k}\psi\|^2_{L^2(\Omega)} = \|\psi \operatorname{curl} A\|^2_{L^2(\Omega)} + \|\nabla_A^2 \psi\|^2_{L^2(\Omega)} + 2\mathfrak{J}(\overline{\nabla_{A^1}\psi}\nabla_{A^2}\psi, \operatorname{curl} A) + I(\partial\Omega), \tag{4.3}$$

where

$$\begin{aligned} I(\partial\Omega) &= \int_{\partial\Omega} \left\{ \sum_{j,k} [\overline{\nabla_{A^k}\nabla_{A^j}\psi}\nabla_{A^k}\psi\nu_j] - \overline{\nabla_A^2\psi}(\nabla_A\psi)\cdot\nu \right\} ds \\ &= \int_{\partial\Omega} \left\{ \frac{1}{2}\frac{\partial}{\partial\nu}|\nabla_A\psi|^2 + (\operatorname{curl} A)[\mathfrak{J}(\overline{\psi}\nabla\psi) - |\psi|^2 A]\cdot\tau - \mathfrak{R}(\nabla_A\psi\cdot\nu)\overline{\nabla_A^2\psi} \right\} ds. \end{aligned} \tag{4.4}$$

Here τ is the unit tangent vector to $\partial\Omega$ such that the orientation of (ν,τ) is the same as the orientation of x_1x_2 coordinates.

Proof: Let $\psi_j = \nabla_{A^j}\psi$. Then,

$$\begin{aligned} (\nabla_{A^j}\nabla_{A^k}\psi, \nabla_{A^j}\nabla_{A^k}\psi) = (\nabla_{A^j}\psi_k, F_{j,k}\psi) - (\psi_k, F_{j,k}\psi_j) + (\nabla_{A^k}\psi_k, \nabla_{A^j}\psi_j) \\ + \int_{\partial\Omega} \overline{\nabla_{A^k}\psi_j}\psi_k\nu_j \, ds - \int_{\partial\Omega} \overline{\nabla_{A^j}\psi_j}\psi_k\nu_k \, ds. \end{aligned}$$

Taking summation over $1\leqslant j,\ k\leqslant 2$ we obtain

$$\begin{aligned}\sum_{j,k}\|\nabla_{A^j}\nabla_{A^k}\psi\|^2_{L^2(\Omega)}&=\sum_{j,k}[(\nabla_{A^j}\psi_k,F_{j,k}\psi)-(\psi_k,F_{j,k}\psi_j)]+\|\nabla_A^2\psi\|^2_{L^2(\Omega)}\\&\quad+\int_{\partial\Omega}\left\{\sum_{j,k}[\overline{\nabla_{A^k}\psi_j}\psi_k\nu_j]-\overline{\nabla_A^2\psi}(\nabla_A\psi)\cdot\nu\right\}ds\\&=i(\overline{\psi}[\nabla_{A^1}\psi_2-\nabla_{A^2}\psi_1]+\psi_1\overline{\psi_2}-\psi_2\overline{\psi_1},\operatorname{curl}A)\\&\quad+\|\nabla_A^2\psi\|^2_{L^2(\Omega)}+\int_{\partial\Omega}\left\{\sum_{j,k}[\overline{\nabla_{A^k}\psi_j}\psi_k\nu_j]-\overline{\nabla_A^2\psi}(\nabla_A\psi)\cdot\nu\right\}ds\\&=(|\psi|^2\operatorname{curl}A+2\mathfrak{F}(\partial_2\psi\partial_1\overline{\psi})+2\mathfrak{R}(A^1\overline{\psi}\partial_2\psi-A^2\psi\partial_1\overline{\psi}),\operatorname{curl}A)\\&\quad+\|\nabla_A^2\psi\|^2_{L^2(\Omega)}+\int_{\partial\Omega}\left\{\sum_{j,k}[\overline{\nabla_{A^k}\psi_j}\psi_k\nu_j]-\overline{\nabla_A^2\psi}(\nabla_A\psi)\cdot\nu\right\}ds\\&=\|\psi\operatorname{curl}A\|^2_{L^2(\Omega)}+\|\nabla_A^2\psi\|^2_{L^2(\Omega)}+2\mathfrak{F}(\partial_2\psi\partial_1\overline{\psi},\operatorname{curl}A)\\&\quad+2\mathfrak{R}(A^1\overline{\psi}\partial_2\psi-A^2\psi\partial_1\overline{\psi},\operatorname{curl}A)\\&\quad+\int_{\partial\Omega}\left\{\sum_{j,k}[\overline{\nabla_{A^k}\psi_j}\psi_k\nu_j]-\overline{\nabla_A^2\psi}(\nabla_A\psi)\cdot\nu\right\}ds,\end{aligned}$$

which gives (4.3). (4.3) implies that $I(\partial\Omega)$ is real and

$$I(\partial\Omega)=\mathfrak{R}\int_{\partial\Omega}\left\{\sum_{j,k}[\overline{F_{j,k}}+\overline{\nabla_{A^j}\psi_k}]\psi_k\nu_j-\overline{\nabla_A^2\psi}(\nabla_A\psi)\cdot\nu\right\}ds.$$

A computation shows

$$\begin{aligned}\mathfrak{R}\int_{\partial\Omega}\sum_{j,k}\overline{F_{j,k}}\psi_k\nu_j&=\mathfrak{R}\int_{\partial\Omega}\overline{F_{1,2}}(\nabla_{A^1}\psi\nu_2-\nabla_{A^2}\psi\nu_1)ds\\&=\mathfrak{R}\int_{\partial\Omega}i\overline{\psi}(\operatorname{curl}A)(\nabla_{A^1}\psi\nu_2-\nabla_{A^2}\psi\nu_1)ds\\&=\mathfrak{R}\int_{\partial\Omega}i\overline{\psi}(\operatorname{curl}A)(\partial_1\psi\nu_2-\partial_2\psi\nu_1)ds+\int_{\partial\Omega}|\psi|^2(\operatorname{curl}A)(A^1\nu_2-A^2\nu_1)ds\\&=\int_{\partial\Omega}(\operatorname{curl}A)[\mathfrak{F}(\overline{\psi}\nabla\psi)-|\psi|^2A]\cdot\tau\,ds;\end{aligned}$$

and

$$\begin{aligned}\mathfrak{R}\int_{\partial\Omega}\sum_{j,k}[\overline{\nabla_{A^j}\psi_k}]\psi_k\nu_j\,ds&=\mathfrak{R}\int_{\partial\Omega}\sum_{j,k}(\partial_j+iA^j)\overline{\nabla_{A^k}}\nabla_{A^k}\psi\nu_j\,ds\\&=\mathfrak{R}\int_{\partial\Omega}\sum_{j,k}\partial_j\overline{\nabla_{A^k}}\nabla_{A^k}\psi\nu_j\,ds=\frac{1}{2}\int_{\partial\Omega}\frac{\partial}{\partial\nu}|\nabla_A\psi|^2\,ds.\end{aligned}$$

So (4.4) is true. □

Proof of Theorem 4.1: For a given compact subset K of Ω, we choose a smooth real cutoff function η such that spt $\eta\subset\Omega$ and $\eta=1$ on K. Denote $\psi_j=\nabla_{A^j}\psi$. From (4.3) and the Kato's inequality it follows that

$$\begin{aligned}\|\nabla|\nabla_A(\eta\psi)|\|^2_{L^2(\Omega)}&\leqslant\sum_{j,k}\|\nabla_{A^j}\nabla_{A^k}(\eta\psi)\|^2_{L^2(\Omega)}\\&=\|\eta\psi\,\mathrm{curl}\,A\|^2_{L^2(\Omega)}+\|\nabla^2_A(\eta\psi)\|^2_{L^2(\Omega)}+2\mathfrak{F}(\overline{\nabla_{A^1}(\eta\psi)}\nabla_{A^2}(\eta\psi),\mathrm{curl}\,A)\\&\leqslant\|\eta\psi\,\mathrm{curl}\,A\|^2_{L^2(\Omega)}+\|\eta\nabla^2_A\psi+2\nabla\eta\cdot\nabla_A\psi+\psi\Delta\eta\|^2_{L^2(\Omega)}\\&\quad+2\int_\Omega|\mathrm{curl}\,A||\nabla_A(\eta\psi)|^2\,dx.\end{aligned}$$

Next, we estimate

$$\begin{aligned}\int_\Omega|\mathrm{curl}\,A||\nabla_A(\eta\psi)|^2\,dx&\leqslant\|\mathrm{curl}\,A\|_{L^2(\Omega)}\left\{\int_\Omega|\nabla_A(\eta\psi)||\nabla_A(\eta\psi)|^3\,dx\right\}^{1/2}\\&\leqslant\|\mathrm{curl}\,A\|_{L^2(\Omega)}\|\nabla_A(\eta\psi)\|^{1/2}_{L^2(\Omega)}\|\nabla_A(\eta\psi)\|^{3/2}_{L^6(\Omega)}\\&\leqslant C_1\|\mathrm{curl}\,A\|_{L^2(\Omega)}\|\nabla_A(\eta\psi)\|^{1/2}_{L^2(\Omega)}\|\nabla|\nabla_A(\eta\psi)|\|^{3/2}_{L^2(\Omega)}\\&\qquad(\text{by Sobolev inequality})\\&\leqslant\tfrac{1}{2}\|\nabla|\nabla_A(\eta\psi)|\|^2_{L^2(\Omega)}+C_2\|\mathrm{curl}\,A\|^4_{L^2(\Omega)}\|\nabla_A(\eta\psi)\|^2_{L^2(\Omega)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\tfrac{1}{2}\|\nabla|\nabla_A(\eta\psi)|\|^2_{L^2(\Omega)}&\leqslant\|\eta\psi\,\mathrm{curl}\,A\|^2_{L^2(\Omega)}+3\|\eta\nabla^2_A\psi\|^2_{L^2(\Omega)}+12\|\nabla\eta\cdot\nabla_A\psi\|^2_{L^2(\Omega)}\\&\quad+3\|\psi\Delta\eta\|^2_{L^2(\Omega)}+C_2\|\mathrm{curl}\,A\|^4_{L^2(\Omega)}\|\nabla_A(\eta\psi)\|^2_{L^2(\Omega)}\\&\leqslant\|\psi\,\mathrm{curl}\,A\|^2_{L^2(\Omega)}+3\|g\|^2_{L^2(\Omega)}+C_3[\|\nabla_A\psi\|^2_{L^2(\Omega)}+\|\psi\|^2_{L^2(\Omega)}]\\&\quad+C_2\|\mathrm{curl}\,A\|^4_{L^2(\Omega)}\|\nabla_A(\eta\psi)\|^2_{L^2(\Omega)},\end{aligned}$$

where C_1,C_2,C_3 are constants depending only on Ω and K. The proof is complete. □

V. ESTIMATES NEAR BOUNDARIES

In this section we establish the boundary estimates for the solutions of the equation

$$\begin{aligned}-\nabla^2_A\psi=g\quad&\text{in }\Omega,\\(\nabla_A\psi)\cdot\nu+\gamma\psi=0\quad&\text{on }\partial\Omega.\end{aligned}\tag{5.1}$$

As mentioned in Sec. I, by making a gauge transformation if necessary, we may assume that

$$\mathrm{div}\,A=0\quad\text{in }\Omega,\quad A\cdot\nu=0\quad\text{on }\partial\Omega.$$

Of course, under the gauge transformation, function g in (5.1) will be changed to a new function $\tilde{g}$. However, since it does not effect the estimation given below, we still denote the new function $\tilde{g}$ by g.

To obtain the estimates we shall straighten a portion of boundary and study the new equation in the half ball B_R^+. We also need to extend the solutions to the entire ball. For this purpose we transform Eq. (5.1) to an equation having homogeneous boundary condition. Let u be the positive eigenfunction associated with the first eigenvalue λ of the following eigenvalue problem:

$$
\begin{aligned}
&-\Delta u=\lambda u \quad \text{in } \Omega,\\
&\frac{\partial u}{\partial \nu}+\gamma u=0 \quad \text{on } \partial\Omega.
\end{aligned}
\tag{5.2}
$$

u is smooth and positive on $\bar{\Omega}$. Set $\psi=u\phi$, $v=\log(u^2)$, $f=g/u$. Then, ϕ satisfies the equation

$$
-\nabla_A^2\phi=\nabla v\cdot\nabla_A\phi-\lambda\phi+f \quad \text{in } \Omega,
$$

$$
\frac{\partial\phi}{\partial\nu}=0 \quad \text{on } \partial\Omega.
$$

In the following we denote $\tilde{\phi}(y)=\phi(\mathcal{F}(y))$, $\tilde{v}(y)=v(\mathcal{F}(y))$, where $\mathcal{F}(y)$ is the diffeomorphism defined on B_{R_0}, see (3.4). We shall always assume $R<R_0/2$. Let $\mathbf{a}(y)$ be the vector field associated with $A(x)$ defined by (3.7). We define the following differential operators:

$$
D(g)w=D(g)_1 w\mathbf{e}_1+D(g)_2 w\mathbf{e}_2,
$$

$$
\text{where } D(g)_1=\frac{1}{g}\partial_1, \quad D(g)_2=\partial_2,
$$

$$
D(g)_{\mathbf{a}}w=[D(g)_{\mathbf{a}^1}w]\mathbf{e}_1+[D(g)_{\mathbf{a}^2}w]\mathbf{e}_2,
$$

$$
\text{where } D(g)_{\mathbf{a}^1}w=\frac{1}{g}(\partial_1-i\mathbf{a}^1)w, \quad D(g)_{\mathbf{a}^2}w=(\partial_2-i\mathbf{a}^2)w,
$$

$$
D(g)_{\mathbf{a}}^*w=[D(g)_{\mathbf{a}^1}^*w]\mathbf{e}_1+[D(g)_{\mathbf{a}^2}^*w]\mathbf{e}_2,
$$

$$
\text{where } D(g)_{\mathbf{a}^1}^*w=D(g)_{\mathbf{a}^1}w, \quad D(g)_{\mathbf{a}^2}^*w=\frac{1}{g}[\partial_2(gw)-i\mathbf{a}^2 gw],
$$

$$
\begin{aligned}
\Delta(g)_{\mathbf{a}}w&=D(g)_{\mathbf{a}^1}^*D(g)_{\mathbf{a}^1}w+D(g)_{\mathbf{a}^2}^*D(g)_{\mathbf{a}^2}w\\
&=\frac{1}{g}\left\{\partial_1\left[\frac{1}{g}(\partial_1 w-i\mathbf{a}^1 w)\right]-\frac{i\mathbf{a}^1}{g}(\partial_1 w-i\mathbf{a}^1 w)\right\}\\
&\quad+\frac{1}{g}\{\partial_2[g(\partial_2 w-i\mathbf{a}^2 w)]-i\mathbf{a}^2 g(\partial_2 w-i\mathbf{a}^2 w)\}.
\end{aligned}
$$

As in Sec. III we denote $\nabla_y\chi=(\partial_1\chi)\mathbf{e}_1+(\partial_2\chi)\mathbf{e}_2$. The operators $D(g)_{\mathbf{a}}$ and $\Delta(g)_{\mathbf{a}}$ have the following *gauge invariant properties:*

$$
D(g)_{\mathbf{a}+\nabla_y\chi}(e^{i\chi}\varphi)=e^{i\chi}D(g)_{\mathbf{a}}\varphi, \quad \Delta(g)_{\mathbf{a}+\nabla_y\chi}(e^{i\chi}\varphi)=e^{i\chi}\Delta(g)_{\mathbf{a}}\varphi.
\tag{5.3}
$$

Note that, in the above notations, $\nabla_A\phi=D(g)_{\mathbf{a}}\tilde{\phi}$, $\nabla_A^2\phi=\Delta(g)_{\mathbf{a}}\tilde{\phi}$. Thus, $\tilde{\phi}$ satisfies the equation

$$-\Delta(g)_{\mathbf{a}}\overline{\phi} = D(g)\tilde{v}\cdot D(g)_{\mathbf{a}}\tilde{\phi} - \lambda\tilde{\phi} + \tilde{f} \quad \text{on } B_R^+,$$
$$\frac{\partial\tilde{\phi}}{\partial y_2} = 0 \quad \text{on } \Gamma_R, \tag{5.4}$$

where $\Gamma_R = \{(y_1,0): |y_1| < R\}$.

Next, we extend the solution $\tilde{\phi}$ of Eq. (5.4) onto the entire ball. Note that

$$\mathbf{a}^2 = 0, \quad \partial_2\tilde{\phi} = 0 \quad \text{when } y_2 = 0.$$

Hence, we can evenly extend $\mathbf{a}^1$ and $\tilde{\phi}$ in y_2 and oddly extend $\mathbf{a}^2$ in y_2. Note that although $g(y)$ is defined on the entire ball, it is not even in y_2. Therefore, for $y_2 < 0$ we define

$$\tilde{\phi}(y_1,y_2) = \tilde{\phi}(y_1,-y_2), \quad \tilde{v}(y_1,y_2) = \tilde{v}(y_1,-y_2),$$

$$g(y_1,y_2) = g(y_1,-y_2),$$

$$\mathbf{a}^1(y_1,y_2) = \mathbf{a}^1(y_1,-y_2), \quad \mathbf{a}^2(y_1,y_2) = -\mathbf{a}^2(y_1,-y_2).$$

After such extensions, $\tilde{\phi}\in C^1(B_R)$, $\mathbf{a}\in C(B_R)$, and $\partial_j\mathbf{a}^j\in C^1(B_R)$. We further notice that $D(g)_{\mathbf{a}^1}\tilde{\phi}$ is continuous and even in y_2, $D(g)_{\mathbf{a}^2}\tilde{\phi}$ is continuous and odd in y_2, and $\Delta(g)_{\mathbf{a}}\tilde{\phi}$ is even in y_2. Although $\partial_2\tilde{v}$ is not continuous at $y_2 = 0$, it is bounded, and $D(g)_{\mathbf{a}^2}\tilde{\phi} = 0$ at $y_2 = 0$. Hence, $D(g)\tilde{v}\cdot D(g)_{\mathbf{a}}\tilde{\phi}$ is continuous.

The main result in this section is the following

Theorem 5.1: *Assume that $\tilde{\phi}$ is a solution of Eq. (5.4) and is extended as the above. Then,*

$$\begin{aligned}\sum_{j,k}\|D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}\tilde{\phi}\|^2_{L^2_g(B_R)} \leq{}& 6\|\tilde{f}\|^2_{L^2_g(B_R)} + 6\int_{B_R}|D(g)\tilde{v}\cdot D(g)_{\mathbf{a}}\tilde{\phi}|^2 g\,dy + C(g,R)\|\tilde{\phi}\|^2_{L^2_g(B_R)}\\ &+ C(g,R)\{1 + \|\partial_1\mathbf{a}^2 - \partial_2\mathbf{a}^1\|^4_{L^2_g(B_R)}\}\|D(g)_{\mathbf{a}}\tilde{\phi}\|^2_{L^2_g(B_R)}\\ &+ C(g,R)\int_{B_R}|\partial_1\mathbf{a}^2 - \partial_2\mathbf{a}^1|^2|\tilde{\phi}|^2\,dy.\end{aligned} \tag{5.5}$$

To prove Theorem 5.1 we need an identity, see Proposition 5.3 below. Define

$$G_{jk}w = D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}w - D(g)_{\mathbf{a}^k}D(g)_{\mathbf{a}^j}w,$$

$$G^*_{jk} = D(g)^*_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}w - D(g)_{\mathbf{a}^k}D(g)^*_{\mathbf{a}^j}w.$$

Denote

$$[w]_{\Gamma_R} = \int_{\Gamma_R^+} w\,dy_1 - \int_{\Gamma_R^-} w\,dy_1 = \lim_{\epsilon_1,\epsilon_2\to 0}\left\{\int_{-R}^{R}[w(y_1,\epsilon_1) - w(y_1,-\epsilon_2)]dy_1\right\}.$$

If w is continuous in B_R, then $[w]_{\Gamma_R} = 0$. Note that $g = 1$ on Γ_R. We have

Lemma 5.2: Assume that ϕ and $\psi\in C^1(B_R\backslash\Gamma_R)$ and the support $\mathrm{spt}(\phi)\subset B_R$. *Then,*

$$\int_{B_R}(D(g)_{\mathbf{a}^j}\phi)\overline{\psi}g\,dy=-\int_{B_R}\phi\overline{D(g)^*_{\mathbf{a}^j}\psi}g\,dy+[\nu_j\phi\overline{\psi}]_{\Gamma_R},$$

where $\nu_1=0,\ \nu_2=-1$. *Moreover if* $\psi\in C^2(B_R\backslash\Gamma_R)$, *then*

$$\int_{B_R}(D(g)_{\mathbf{a}^j}\phi)\overline{D(g)_{\mathbf{a}^k}\psi}g\,dy=-\int_{B_R}\phi\overline{D(g)^*_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}\psi}g\,dy+[\nu_j\phi\overline{D(g)_{\mathbf{a}^k}\psi}]_{\Gamma_R}.$$

For our convenience we denote by $\|\cdot\|_{L^2_g}$ the L^2 norm with the weight g, and $(\phi,w)_g=\int_{B_R}\phi\overline{w}g\,dy$.

Proposition 5.3: Assume that $\psi\in C^1(B_R)\cap C^2(B_R\backslash\Gamma_R)$ *with its support* $\mathrm{spt}(\psi)\subset B_R$, *and* ψ *is even in* y_2. *Then,*

$$\sum_{j,k}\|D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}\psi\|^2_{L^2_g(B_R)}=\|\Delta(g)_{\mathbf{a}}\psi\|^2_{L^2_g(B_R)}+\|G_{12}\psi\|^2_{L^2_g(B_R)}+\int_{B_R}|D(g)_{\mathbf{a}^2}\psi|^2g\,\partial_2\left(\frac{\partial_2 g}{g}\right)dy$$
$$-2\mathfrak{F}\int_{B_R}(\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1)(D(g)_{\mathbf{a}^1}\psi)\overline{D(g)_{\mathbf{a}^2}\psi}\,dy-2\mathfrak{R}\int_{B_R}(\partial_2 g)$$
$$\times(D(g)_{\mathbf{a}^2}\psi)\overline{D(g)^2_{\mathbf{a}^1}\psi}\,dy+\int_{B_R}\left[\partial_1\left(\frac{\partial_2 g}{g}\right)-\frac{\partial_1 g\,\partial_2 g}{g}\right]$$
$$\times(D(g)_{\mathbf{a}^1}\psi)\overline{D(g)_{\mathbf{a}^2}\psi}\,dy. \tag{5.6}$$

Remark 5.1: Note that $\mathrm{curl}\,\mathbf{a}=(\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1)/g$. Although it is not continuous at $y_2=0$, it remains bounded. The term $\partial_1((\partial_2 g)/g)-(\partial_1 g)(\partial_2 g)/g$ is also bounded. Therefore, the integrals involving such terms make sense. Also note that when $y_2\neq 0$, $g\partial_2((\partial_2 g)/g)=-[\kappa_r(y_1)]^2/g\leqslant 0$. So

$$\int_{B_R}|D(g)_{\mathbf{a}^2}\psi|^2g\,\partial_2\left(\frac{\partial_2 g}{g}\right)dy\leqslant 0.$$

Proof of Proposition 5.3: The proof is similar to one of Proposition 4.3, but involves more computations. Set $\psi_j=D(g)_{\mathbf{a}^j}\psi$. Using Lemma 5.2 we have

$$\begin{aligned}&(D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}\psi,D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}\psi)_g\\&=(D(g)_{\mathbf{a}^j}\psi_k,G_{jk}\psi+D(g)_{\mathbf{a}^k}\psi_j)_g\\&=(D(g)_{\mathbf{a}^j}\psi_k,G_{jk}\psi)_g+(D(g)_{\mathbf{a}^j}\psi_k,D(g)_{\mathbf{a}^k}\psi_j)_g\\&=(D(g)_{\mathbf{a}^j}\psi_k,G_{jk}\psi)_g-(\psi_k,G^*_{jk}\psi_j)_g+(D(g)^*_{\mathbf{a}^k}\psi_k,D(g)^*_{\mathbf{a}^j}\psi_j)_g\\&\quad+[\nu_j\psi_k\overline{D(g)_{\mathbf{a}^k}\psi_j}]_{\Gamma_R}-[\nu_k\psi_k\overline{D(g)^*_{\mathbf{a}^j}\psi_j}]_{\Gamma_R}.\end{aligned}$$

Summing up the above over $1\leqslant j,\ k\leqslant 2$ we have

$$\sum_{j,k}\|D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}\psi\|^2_{L^2_g(B_R)}=\|\Delta(g)_{\mathbf{a}}\psi\|^2_{L^2_g(B_R)}+\sum_{j,k}\{(D(g)_{\mathbf{a}^j}\psi_k,G_{jk}\psi)_g-(\psi_k,G^*_{jk}\psi_j)_g\}$$
$$+\sum_{jk}\{[\nu_j\psi_k\overline{D(g)_{\mathbf{a}^k}\psi_j}]_{\Gamma_R}-[\nu_k\psi_k\overline{D(g)^*_{\mathbf{a}^j}\psi_j}]_{\Gamma_R}\}.$$

Since $\nu_1=0$, $\nu_2=-1$, $\psi_2=D(g)_{\mathbf{a}^2}\psi=0$ on Γ_R, we have $\nu_k\psi_k=0$ on Γ_R, and

$$\sum_{jk}\{[\nu_j\psi_k\overline{D(g)_{\mathbf{a}^k}\psi_j}]_{\Gamma_R}-[\nu_k\psi_k\overline{D(g)^*_{\mathbf{a}^j}\psi_j}]_{\Gamma_R}\}=-[(D(g)_{\mathbf{a}^1}\psi)\overline{D(g)_{\mathbf{a}^1}\psi_2}]_{\Gamma_R}=0.$$

Here the following fact is used:

$$D(g)_{\mathbf{a}^1}\psi_2=D(g)_{\mathbf{a}^1}D(g)_{\mathbf{a}^2}\psi=\frac{1}{g}[\partial_1\partial_2\psi-i\mathbf{a}^1\partial_2\psi-i\partial_1(\mathbf{a}^2\psi)-\mathbf{a}^1\mathbf{a}^2\psi]\to 0\quad\text{as }\ y_2\to 0.$$

Therefore,

$$\begin{aligned}\sum_{j,k}\|D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}\psi\|^2_{L^2_g(B_R)}&=\|\Delta(g)_{\mathbf{a}}\psi\|^2_{L^2_g(B_R)}+(D(g)_{\mathbf{a}^1}\psi_2-D(g)_{\mathbf{a}^2}\psi_1,G_{12}\psi)_g-\sum_{jk}(\psi_k,G^*_{jk}\psi_j)_g\\&=\|\Delta(g)_{\mathbf{a}}\psi\|^2_{L^2_g(B_R)}+\|G_{12}\psi\|^2_{L^2_g(B_R)}-\sum_{jk}(\psi_k,G_{jk}\psi_j)_g+\sum_k\left(\psi_k,\psi_2D(g)_k\left[\frac{\partial_2 g}{g}\right]\right)_g.\end{aligned}\tag{5.7}$$

By computation we get

$$\sum_k\left(\psi_k,\psi_2D(g)_k\left[\frac{\partial_2 g}{g}\right]\right)_g=\int_{B_R}|D(g)_{\mathbf{a}^2}\psi|^2g\,\partial_2\left(\frac{\partial_2 g}{g}\right)dy+\int_{B_R}\partial_1\left(\frac{\partial_2 g}{g}\right)(D(g)_{\mathbf{a}^1}\psi)\overline{D(g)_{\mathbf{a}^2}\psi}\,dy,\tag{5.8}$$

$$\begin{aligned}\sum_{jk}(\psi_k,G_{jk}\psi_j)_g=2\mathfrak{I}\int_{B_R}(\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1)(D(g)_{\mathbf{a}^1}\psi)\overline{D(g)_{\mathbf{a}^2}\psi}\,dy+2\mathfrak{R}\int_{B_R}(\partial_2 g)\psi_2\overline{D(g)_{\mathbf{a}^1}\psi_1}\,dy\\+\int_{B_R}\frac{\partial_1 g\,\partial_2 g}{g}(D(g)_{\mathbf{a}^1}\psi)\overline{D(g)_{\mathbf{a}^2}\psi}\,dy.\end{aligned}\tag{5.9}$$

For instance, to obtain (5.9), we note that

$$\begin{aligned}\sum_{jk}(\psi_k,G_{jk}\psi_j)_g&=-(\psi_1,G_{12}\psi_2)_g+(\psi_2,G_{12}\psi_2)_g\\&=\int_{B_R}g[\psi_2\overline{G_{12}\psi_1}-\psi_1\overline{G_{12}\psi_2}]dy\\&=2\mathfrak{I}\int_{B_R}(\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1)(D(g)_{\mathbf{a}^1}\psi)\overline{D(g)_{\mathbf{a}^2}\psi}\,dy\\&\quad+\int_{B_R}(\partial_2 g)[\psi_2\overline{D(g)_{\mathbf{a}^1}\psi_1}-\psi_1\overline{D(g)_{\mathbf{a}^1}\psi_2}]dy.\end{aligned}$$

Since

$$\int_{B_R}(\partial_2 g)[\psi_2\overline{D(g)_{\mathbf{a}^1}\psi_1}-\psi_1\overline{D(g)_{\mathbf{a}^1}\psi_2}]dy$$

$$=\int_{B_R}(\partial_2 g)[\psi_2\overline{D(g)_{\mathbf{a}^1}\psi_1}+\overline{\psi_2}(D(g)_{\mathbf{a}^1}\psi_1)]dy+\int_{B_R}\frac{\partial_1 g\,\partial_2 g}{g}\,\psi_1\overline{\psi_2}\,dy$$

$$=2\Re\int_{B_R}(\partial_2 g)\,\psi_2\overline{D(g)_{\mathbf{a}^1}\psi_1}+\int_{B_R}\frac{\partial_1 g\,\partial_2 g}{g}\,\psi_1\overline{\psi_2}\,dy,$$

so (5.9) holds.

Now, (5.6) follows from (5.7) to (5.9). □

Proof of Theorem 5.1: In the proof, for simplicity we denote $\tilde{\phi}$ by ϕ and denote a constant depending only on g and R by C. Let η be a smooth cutoff function supported in B_{2R} such that $\eta=1$ on $B_{R/2}$ and η is even in y_2. Using Proposition 5.3 we have

$$\sum_{j,k}\|D(g)_{\mathbf{a}^j}D(g)_{\mathbf{a}^k}(\eta\phi)\|^2_{L^2_g(B_R)}\leqslant\|\Delta(g)_{\mathbf{a}}(\eta\phi)\|^2_{L^2_g(B_R)}+\|G_{12}\psi\|^2_{L^2_g(B_R)}$$
$$+C\{\|D(g)_{\mathbf{a}}\phi\|^2_{L^2_g(B_R)}+\|\phi\|^2_{L^2_g(B_R)}\}+J_1+2J_2, \quad (5.10)$$

where

$$J_1=\int_{B_R}|\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1||D(g)_{\mathbf{a}}(\eta\phi)|^2\,dy,$$

$$J_2=\int_{B_R}|\partial_2 g||D(g)_{\mathbf{a}^2}(\eta\phi)||D(g)^2_{\mathbf{a}^1}(\eta\phi)|dy.$$

Now, we estimate each term on the right of (5.10),

$$\|\Delta(g)_{\mathbf{a}}(\eta\phi)\|^2_{L^2_g(B_R)}\leqslant 3\|\tilde{f}\|^2_{L^2_g(B_R)}+C\{\|\phi\|^2_{L^2_g(B_R)}+\|D(g)_{\mathbf{a}}\phi\|^2_{L^2_g(B_R)}\}$$
$$+3\int_{B_R}|D(g)\tilde{v}\cdot D(g)_{\mathbf{a}}\phi|^2 g\,dy,$$

$$\|G_{12}\psi\|^2_{L^2_g(B_R)}\leqslant C\left\{\|D(g)_{\mathbf{a}}\phi\|^2_{L^2_g(B_R)}+\int_{B_R}|\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1||\phi|^2\,dy\right\},$$

$$J_1\leqslant\|\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1\|_{L^2_g(B_R)}\left\{\int_{B_R}|D(g)_{\mathbf{a}}(\eta\phi)||D(g)_{\mathbf{a}}(\eta\phi)|^3\,dy\right\}^{1/2}$$
$$\leqslant C\|\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1\|_{L^2_g(B_R)}\|D(g)_{\mathbf{a}}(\eta\phi)\|^{1/2}_{L^2_g(B_R)}\|\nabla|D(g)_{\mathbf{a}}(\eta\phi)|\|^{3/2}_{L^2_g(B_R)}$$

(by Sobolev inequality)

$$\leqslant\epsilon\|\nabla|D(g)_{\mathbf{a}}(\eta\phi)|\|^2_{L^2_g(B_R)}+\frac{C}{\epsilon}\|\partial_1\mathbf{a}^2-\partial_2\mathbf{a}^1\|^4_{L^2_g(B_R)}\|D(g)_{\mathbf{a}}(\eta\phi)\|^2_{L^2_g(B_R)},$$

and

$$J_2 \leqslant \epsilon \|D(g)^2_{\mathbf{a}^1}(\eta\phi)\|^2_{L^2_g(B_R)} + \frac{C}{\epsilon}\|D(g)_{\mathbf{a}}(\eta\phi)\|^2_{L^2_g(B_R)}$$

$$\leqslant \epsilon \sum_{jk} \|D(g)_{\mathbf{a}^j} D(g)_{\mathbf{a}^k}(\eta\phi)\|^2_{L^2_g(B_R)} + \frac{C}{\epsilon}\|D(g)_{\mathbf{a}}(\eta\phi)\|^2_{L^2_g(B_R)}.$$

Plugging the above inequalities back in (5.10), using the following

$$\|\nabla|D(g)_{\mathbf{a}}(\eta\phi)|\|^2_{L^2_g(B_R)} \leqslant C(g,R)\sum_{jk} \|D(g)_{\mathbf{a}^j} D(g)_{\mathbf{a}^k}(\eta\phi)\|^2_{L^2_g(B_R)},$$

and choosing ϵ small enough, we obtain the estimate (5.5). □

In the same fashion as the above, one can also prove the following

Lemma 5.4: Assume that $\tilde{\phi}$ *is a solution of Eq. (5.4) and is extended even in* y_2*. Then,*

$$\int_{B_{R/2}} |D(g)_{\mathbf{a}}\tilde{\phi}|^2 g\, dy \leqslant C(g,R)\int_{B_R} \{|\tilde{\phi}|^2 + |\tilde{f}|^2 + |D(g)\tilde{v}|^2 |\tilde{\phi}|^2\} g\, dy.$$

Note that after extension $\partial_1 \mathbf{a}^2 - \partial_2 \mathbf{a}^1$ is not continuous at Γ_R. Therefore, the estimates depending only on the data given on B_R^+ are needed. As a direct corollary of Theorem 5.1 we have

Theorem 5.5: *Assume that* $\tilde{\phi}$ *is a solution of Eq. (5.4). Then,*

$$\sum_{jk} \|D(g)_{\mathbf{a}^j} D(g)_{\mathbf{a}^k}\tilde{\phi}\|^2_{L^2_g(B_R^+)} \leqslant C\{\|\tilde{f}\|^2_{L^2_g(B_R^+)} + \|\tilde{\phi}\|^2_{L^2_g(B_R^+)}\}, \tag{5.11}$$

where the constant C depends on R, g, $\|D(g)\tilde{v}\|_{L^2_g(B_R^+)}$ and $\|\partial_1 \mathbf{a}^2 - \partial_2 \mathbf{a}^1\|_{L^2_g(B_R^+)}$. □

VI. UPPER-BOUND ESTIMATES

In this section, we give an upper bound for $\mu(\sigma A)/|\sigma|$. Throughout this section we assume $A \in C^2(\Omega)$.

Lemma 6.1: Assume that $A \in C^2(\Omega)$*. Then,*

$$\limsup_{\sigma\to\infty} \frac{\mu(\sigma A)}{|\sigma|} \leqslant \min_{x\in\Omega} |\mathrm{curl}\, A(x)|. \tag{6.1}$$

Proof: Let $H(x) = \mathrm{curl}\, A(x)$. First, we note that $\mu(-\sigma A) = \mu(\sigma A)$. In fact, for every $\psi \in W^{1,2}(\Omega)$ we set $\phi = \bar{\psi}$. Then, $|\nabla_{-\sigma A}\phi| = |\overline{\nabla_{\sigma A}\psi}| = |\nabla_{\sigma A}\psi|$. Therefore, we may assume $\sigma > 0$. We shall show that for every $x_0 \in \Omega$,

$$\limsup_{\sigma\to+\infty} \frac{\mu(\sigma A)}{\sigma} \leqslant |H(x_0)|.$$

Without loss of generality we may assume $x_0 = 0$.

Set $h = H(0)$. When $h = 0$ the conclusion is obvious, see Proposition 6.3 below. So, we assume $h \neq 0$. Denote $\delta = 1/\sqrt{\sigma}$. Let $R > 0$ be fixed such that $B_R \subset \Omega$. For any $\psi \in W^{1,2}(\Omega)$, we let $\psi_\delta(x) = \psi(\delta x)$ and $A_\delta(x) = A(\delta x)/\delta$. Then,

$$\frac{\mu(\sigma A)}{\sigma} \leqslant \frac{1}{\sigma} \inf_{\psi \in W_0^{1,2}(B_R)} \frac{\int_{B_R} |\nabla_{\sigma A}\psi|^2\, dx}{\int_{B_R}|\psi|^2},$$

$$= \inf_{\phi \in W_0^{1,2}(B_{R/\delta})} \frac{\int_{B_{R/\delta}} |\nabla_{A_\delta}\phi|^2\, dx}{\int_{B_{R/\delta}}|\phi|^2\, dx}.$$

Using (3.1) and noting that $\nabla\xi(\delta x) = \delta\nabla\xi(x)$, $\omega(\delta x) = \delta\omega(x)$, we have

$$A_\delta(x) = \nabla\chi_\delta(x) + h\omega(x) + B_\delta(x),$$

where

$$\chi_\delta(x) = \frac{1}{\delta}A(0)\cdot x + \xi(x) + \delta\zeta(x),$$

$$B_\delta(x) = -\frac{\delta}{2}|x|^2 \operatorname{curl}^2 A(0) + \frac{1}{\delta}D(\delta x),$$

$$|B_\delta(x)| \leqslant \frac{\delta}{2}|\operatorname{curl}^2 A(0)||x|^2[1 + o(\delta R)] \quad \text{in } B_{R/\delta}.$$

Therefore,

$$|\nabla_{A_\delta} e^{i\chi_\delta}\phi|^2 = |\nabla_{h\omega + B_\delta}\phi|^2 = |\nabla_{h\omega}\phi - iB_\delta\phi|^2 \leqslant (1+\lambda)|\nabla_{h\omega}\phi|^2 + \frac{(1+\lambda)\delta^2}{4\lambda}(1 + o(\delta R))|x|^4|\phi|^2,$$

where $0 \leqslant \lambda \leqslant 1$. So,

$$\frac{\mu(\sigma A)}{\sigma} \leqslant \inf_{\phi \in W_0^{1,2}(B_{R/\delta})} \frac{1}{\int_{B_{R/\delta}}|\phi|^2\, dx}\Bigg\{(1+\lambda)\int_{R_{R/\delta}} |\nabla_{h\omega}\phi|^2\, dx + \frac{(1+\lambda)\delta^2}{4\lambda}(1 + o(\delta R))|\operatorname{curl}^2 A(0)|^2 \int_{R_{R/\delta}} |x|^4|\phi|^2\, dx\Bigg\}.$$

Choose $\phi = \phi_m = u\,\eta_m$, where $u(x) = u(|x|) = \exp(-h^2|x|^2/4)$ and η_m is a smooth cutoff function supported in B_m such that $\eta_m \equiv 1$ on $B_{m/2}$. For fixed R and for all small δ,

$$\frac{\mu(\sigma A)}{\sigma} \leqslant \frac{1}{\int_{B_m}|\phi_m|^2\, dx}\Bigg\{(1+\lambda)\int_{R_m} |\nabla_{h\omega}\phi_m|^2\, dx + \frac{(1+\lambda)\delta^2}{4\lambda}(1 + o(\delta R))|\operatorname{curl}^2 A(0)|^2 \int_{R_m} |x|^4|\phi_m|^2\, dx\Bigg\}.$$

We first fix $m > 1$, $\lambda \in (0,1)$ and let σ approach $+\infty$ (so $\delta \to 0$), then we fix m and send λ to 0, finally we send m to $+\infty$. By using Lemma 2.1, we obtain

$$\limsup_{\sigma \to +\infty} \frac{\mu(\sigma A)}{\sigma} \leqslant \frac{\int_{\mathbb{R}^2}|\nabla_{h\omega}u|^2\, dx}{\int_{\mathbb{R}^2}|u|^2\, dx} = \alpha(h) = |h| = |H(0)|.$$

This completes the proof. □

Lemma 6.2: Assume that $A \in C^2(\overline{\Omega})$. *Then,*

$$\limsup_{\sigma\to\infty}\frac{\mu(\sigma A)}{|\sigma|}\leqslant\beta_0\min_{x\in\partial\Omega}|\mathrm{curl}\,A(x)|,\tag{6.2}$$

where β_0 is given in Lemma 2.2.

Proof: As in the proof of Lemma 6.1, we only need to show that, if $0\in\partial\Omega$ and $h=H(0)$, then

$$\limsup_{\sigma\to+\infty}\frac{\mu(\sigma A)}{|\sigma|}\leqslant\beta_0|h|.$$

Now, we need to use the local decomposition of A in the new variables which straighten a portion of boundary of $\partial\Omega$ near the point 0. We shall use the notations presented in Sec. III. For a fixed small $R>0$, we have

$$\mu(\sigma A)\leqslant\inf_{\phi\in W^*(B_R^+)}\frac{\int_{B_R^+}|D(g)_{\sigma\mathbf{a}}\phi|^2g(y)dy+\gamma\int_{-R}^{R}|\phi|^2\,dy_1}{\int_{B_R^+}|\phi|^2g(y)dy},$$

where $W^*(B_R^+)=\{\phi\in W^{1,2}(B_R^+):\mathrm{spt}(\phi)\subset\overline{B_R}\}$, $\mathbf{a}$ is the vector field associated with A given by (3.7).

For $\sigma>0$ we set $\delta=1/\sqrt{\sigma}$, $\mathbf{a}_\delta(y)=(1/\delta)\mathbf{a}(\delta y)$, $g_\delta(y)=g(\delta y)$. Then, for all small δ,

$$\frac{\mu(\sigma A)}{\sigma}\leqslant\inf_{\phi\in W^*(B_{R/\delta}^+)}\frac{\int_{B_{R/\delta}^+}|D(g_\delta)_{\mathbf{a}_\delta}\phi|^2g_\delta(y)dy+\gamma\delta\int_{-R/\delta}^{R/\delta}|\phi|^2\,dy_1}{\int_{B_{R/\delta}^+}|\phi|^2g_\delta(y)dy}.$$

From (3.11) we have

$$\mathbf{a}_\delta(y)=\nabla_y\tilde{\chi}_\delta(y)+h\tilde{\omega}(y)+\tilde{B}_\delta(y),$$

where

$$\tilde{\chi}_\delta(y)=\frac{1}{\delta}A(0)\cdot y+\tilde{\xi}(y)+\delta\tilde{\zeta}(y),$$

$$\tilde{\omega}(y)=(-y_2/2)\mathbf{e}_1+(y_1/2)\mathbf{e}_2,$$

$$\tilde{B}_\delta(y)=-\frac{\delta|y|^2}{2}[\mathrm{curl}^2A(0)-h\kappa_r(0)\tau(0)]+\frac{1}{\delta}\tilde{D}(\delta y).$$

Here $\kappa_r(0)$ is the relative curvature of $\partial\Omega$ at the point 0. Since the operator $D(g)_{\mathbf{a}}$ is gauge invariant, see (5.3), so

$$|D(g_\delta)_{\mathbf{a}_\delta}\exp(i\tilde{\chi}_\delta)\phi|^2=|D(g_\delta)_{h\omega+\tilde{B}_\delta}\phi|^2\leqslant(1+\lambda)|D(g_\delta)_{h\omega}\phi|^2+\frac{1+\lambda}{\lambda|g_\delta|^2}|\tilde{B}_\delta\phi|^2,$$

where λ is an arbitrary number lying between 0 and 1.

Choose $\phi=\psi\eta_m$, where $\psi\in\mathcal{W}(\mathbb{R}_+^2)$, η_m is the cutoff function used in the proof of Lemma 6.1. Note that $g_\delta\to1$ uniformly on each B_m^+ as $\delta\to0$. Therefore, by the same argument as in the proof of Lemma 6.1, we obtain

$$\limsup_{\sigma\to+\infty}\frac{\mu(\sigma A)}{\sigma}\leqslant\frac{\int_{\mathbb{R}_+^2}|\nabla_{h\omega}\psi|^2\,dy}{\int_{\mathbb{R}_+^2}|\psi|^2\,dy}.$$

Since the above is true for all $\psi \in \mathcal{W}(\mathbb{R}^2_+)$, using Lemma 2.2 we have

$$\limsup_{\sigma \to +\infty} \frac{\mu(\sigma A)}{\sigma} \leqslant \beta(h) = \beta_0 |h| = \beta_0 |H(0)|.$$

The proof is complete. □

If curl A vanishes at some point in $\overline{\Omega}$, the estimates (6.1), (6.2) can be greatly improved. Denote

$$\mathcal{Z}(A,\Omega) = \{x \in \Omega : \operatorname{curl} A(x) = 0\}, \quad \mathcal{Z}(A,\partial\Omega) = \{x \in \partial\Omega : \operatorname{curl} A(x) = 0\}.$$

Define, for $\tau > 0$,

$$p(\tau) = \inf_{u \in \mathcal{W}(\mathbb{R}^2)} \frac{\int_{\mathbb{R}^2} \{|\nabla u|^2 + \frac{1}{4}\tau^2 |x|^4 |u|^2\} dx}{\int_{\mathbb{R}^2} |u|^2 \, dx}. \tag{6.3}$$

Using the rescaling method we can show that $p(\tau) = p(1)|\tau|^{2/3}$ and $p(\tau)$ is achieved for every $\tau \neq 0$. Choosing $u = \exp(-|x|^3/6)$ as a test function we see that $p(1) \leqslant \sqrt[3]{3}/\Gamma(\frac{5}{3})$. Define, for a constant vector $\mathbf{a}$,

$$q(\mathbf{a}) = \inf_{\phi \in \mathcal{W}(\mathbb{R}^2_+)} \frac{\int_{\mathbb{R}^2_+} \left| \nabla \phi - \frac{i}{2} |y|^2 \mathbf{a} \phi \right|^2 dy}{\int_{\mathbb{R}^2_+} |\phi|^2 \, dy}. \tag{6.4}$$

Obviously,

$$q(\mathbf{a}) \leqslant \inf_{\phi \in \mathcal{W}(\mathbb{R}^2_+)} \frac{\int_{\mathbb{R}^2_+} \{|\nabla \phi|^2 + \frac{1}{4}|\mathbf{a}|^2 |y|^2 |\phi|^2\} dy}{\int_{\mathbb{R}^2_+} |\phi|^2 \, dy} \leqslant p(|\mathbf{a}|) = p(1)|\mathbf{a}|^{2/3} \leqslant \frac{\sqrt[3]{3}}{\Gamma(\frac{5}{3})} |\mathbf{a}|^{2/3}.$$

Proposition 6.3: Assume that $A \in C^2(\overline{\Omega})$. If $\mathcal{Z}(A,\Omega) \neq \emptyset$, *then*

$$\limsup_{\sigma \to \infty} \frac{\mu(\sigma A)}{|\sigma|^{2/3}} \leqslant p(1) \inf_{x \in \mathcal{Z}(A,\Omega)} |\operatorname{curl}^2 A(x)|^{2/3}. \tag{6.5}$$

If $\mathcal{Z}(A,\partial\Omega) \neq \emptyset$, *then*

$$\limsup_{\sigma \to \infty} \frac{\mu(\sigma A)}{|\sigma|^{2/3}} \leqslant \inf_{x \in \mathcal{Z}(A,\partial\Omega)} q(\operatorname{curl}^2 A(x)) \leqslant p(1) \inf_{x \in \mathcal{Z}(A,\partial\Omega)} |\operatorname{curl}^2 A(x)|^{2/3}. \tag{6.6}$$

Here $p(1)$ *and* $q(\mathbf{a})$ *are defined in (6.3), (6.4).*

Remark 6.1: Note that $|\operatorname{curl}^2 A(x)| = |\nabla H(x)|$, *where* $H(x) = \operatorname{curl} A(x)$.

Proof of Proposition 6.3: Assume that $0 \in \mathcal{Z}(A,\Omega)$. For $\sigma > 0$ we set $\delta = 1/\sqrt[3]{\sigma}$, $\psi_\delta(x) = \psi(\delta x)$, $A_\delta(x) = A(\delta x)/\delta^2$. From Lemma 3.1, we have

$$A_\delta(x) = \nabla \hat{\chi}_\delta(x) - \tfrac{1}{2}|x|^2 \operatorname{curl}^2 A(0) + D_\delta(x),$$

where

$$\hat{\chi}_\delta(x)=\frac{1}{\delta^2}A(0)\cdot x+\frac{1}{\delta}\xi(x)+\zeta(x),$$

$$D_\delta(x)=\frac{1}{\delta^2}D(\delta x),\quad |D_\delta(x)|=o(\delta|x|^3)\quad \text{in } B_{R/\delta}.$$

Set $\phi_m=u\eta_m$, where η_m is a smooth cutoff function as we used above and u is a real function to be determined later. Then, we have

$$\frac{\mu(\sigma A)}{\sigma^{2/3}}\leqslant\frac{1}{\int_{B_m}|\phi_m|^2\,dx}\left\{(1+\lambda)\int_{R_m}\left[|\nabla\phi_m|^2+\frac{1}{4}|\mathrm{curl}^2A(0)|^2|x|^4|\phi_m|^2\right]dx\right.$$
$$\left.+O(\delta)\left(1+\frac{1}{\lambda}\right)\int_{R_m}|x|^6|\phi_m|^2\,dx\right\}.$$

First sending δ to 0, then sending λ to 0, finally sending m to $+\infty$, we conclude that

$$\limsup_{\sigma\to+\infty}\frac{\mu(\sigma A)}{\sigma^{2/3}}\leqslant\inf_{u\in\mathcal{W}(\mathbb{R}^2)}\frac{\int_{\mathbb{R}^2}\{|\nabla u|^2+\frac{1}{4}|\mathrm{curl}^2A(0)|^2|x|^4|u|^2\}dx}{\int_{\mathbb{R}^2}|u|^2\,dx}\leqslant p(1)|\mathrm{curl}^2A(0)|^{2/3}.$$

So, (6.5) is true.

Now, we assume $0\in\mathcal{Z}(A,\partial\Omega)$. From (3.11) it follows that

$$\mathbf{a}(y)=A(0)+\nabla_y\tilde{\xi}(y)+\nabla_y\tilde{\zeta}(y)-\tfrac{1}{2}|y|^2\,\mathrm{curl}^2A(0)+\tilde{D}(y).$$

Using the similar argument we obtain (6.6). □

Remark 6.2: If there exist a smooth open subdomain $D\subset\Omega$ such that $\mathrm{curl}\,A(x)$ *vanishes in D, then*

$$\mu(\sigma A)\leqslant\inf_{\phi\in W_0^{1,2}(D)}\frac{\|\nabla_{\sigma A}\phi\|^2_{L^2(D)}}{\|\phi\|^2_{L^2(D)}}=\inf_{\phi\in W_0^{1,2}(D)}\frac{\|\nabla\phi\|^2_{L^2(D)}}{\|\phi\|^2_{L^2(D)}}=\lambda_1(D),$$

where $\lambda_1(D)$ is the first Dirichlet eigenvalue of $-\Delta$ on D.

VII. LOWER-BOUND ESTIMATES

In this section we give an lower bound of $\mu(\sigma A)/|\sigma|$ for large σ. The asymptotic behavior of the eigenfunctions as $\sigma\to\infty$ will also be discussed.

Lemma 7.1: Let Ω be a smooth bounded domain in $\mathbb{R}^2$ and $A\in C^2(\bar{\Omega})$. Then,

$$\liminf_{\sigma\to\infty}\frac{\mu(\sigma A)}{|\sigma|}\geqslant\min\{\min_{x\in\Omega}|\mathrm{curl}\,A(x)|,\beta_0\min_{x\in\partial\Omega}|\mathrm{curl}\,A(x)|\},\tag{7.1}$$

where β_0 is the positive constant given in Lemma 2.2.

Proof: Let $H(x)=\mathrm{curl}\,A(x)$,

$$m(\Omega)=\min_{\bar{\Omega}}|H(x)|,\quad \Omega_m=\{x\in\Omega:|H(x)|=m(\Omega)\},$$

$$m(\partial\Omega)=\min_{\partial\Omega}|H(x)|,\quad (\partial\Omega)_m=\{x\in\partial\Omega:|H(x)|=m(\partial\Omega)\}$$

and $m=\min\{m(\Omega),m(\partial\Omega)\beta_0\}$. We shall show that $\liminf_{\sigma\to\infty}\mu(\sigma A)/|\sigma|\geq m$. As in Sec. VI we assume $\sigma>0$ and denote $\delta=1/\sqrt{\sigma}$. Let ψ^δ be the eigenfunction associated with $\mu(\sigma A)$ satisfying $\max_{x\in\bar\Omega}|\psi^\delta(x)|=1$. Then, ψ^δ satisfies

$$-\nabla^2_{\sigma A}\psi=\mu(\sigma A)\psi \quad \text{in} \quad \Omega, \qquad (\nabla_{\sigma A}\psi)\cdot\nu+\gamma\psi=0 \quad \text{on} \quad \partial\Omega. \tag{7.2}$$

Denote by x^δ the maximum point of $|\psi^\delta|$.

Now, we assume that $\{\sigma_k\}$ is a given sequence, $\sigma_k\to+\infty$. We choose a subsequence σ_{k_j} such that

$$x^{\delta_{k_j}}\to x^0, \quad \frac{\mu(\sigma_{k_j}A)}{\sigma_{k_j}}\to a$$

for some non-negative number a. Lemma 6.1 implies $a\leq m$. We shall show $a\geq m$. Then it follows that $a=m$ for any sequence $\{\sigma_k\}$. For simplicity, we denote σ_{k_j} by σ. Let $h_\delta=H(x^\delta)$, $h=H(x^0)$.

Case 1: $x^0\in\Omega$. We shall show $a\geq m(\Omega)$. Let $\Omega_\delta=(\Omega-x^\delta)/\delta$, $\psi_\delta(x)=\psi^\delta(x^\delta+\delta x)$, $A_\delta(x)=(1/\delta)A(x^\delta+\delta x)$, $H_\delta(x)=H(x^\delta+\delta x)$. Note that $\operatorname{curl}A_\delta(x)=H_\delta(x)$. Using (7.2) we check that ψ_δ satisfies

$$-\nabla^2_{A_\delta}\psi_\delta=\frac{\mu(\sigma A)}{\sigma}\psi_\delta \quad \text{in} \quad \Omega_\delta \tag{7.3}$$

and $|\psi_\delta(0)|=1=\|\psi_\delta\|_{L^\infty}$. We shall show that $\{\psi_\delta\}$ locally converges up to gauge transformations.

Let $R>0$ be a fixed constant. Then, for δ small enough we have $B_{3R}\subset\Omega_\delta$. Since $\{|\psi_\delta|\}$ is uniformly bounded in L^2_{loc}, Lemma 4.2 implies that $\{|\nabla_{A_\delta}\psi_\delta|\}$ is also uniformly bounded in L^2_{loc}. Applying Theorem 4.1 to Eq. (7.3) we have

$$\begin{aligned}\||\nabla_{A_\delta}\psi|\|^2_{H^1(B_R)}&\leq 2\|\psi_\delta H_\delta\|^2_{L^2(2R)}+6\left[\frac{\mu(\sigma A)}{\sigma}\right]^2\|\psi_\delta\|^2_{L^2(B_{2R})}\\&\quad+C(R)[1+\|H_\delta\|^4_{L^2(B_{2R})}][\|\nabla_{A_\delta}\psi_\delta\|^2_{L^2(B_{2R})}+\|\psi_\delta\|^2_{L^2(B_{2R})}]\\&\leq C(R,\|H\|_{L^\infty})\|\psi_\delta\|^2_{L^2(B_{3R})}.\end{aligned}$$

So, $\{|\nabla_{A_\delta}\psi_\delta|\}$ is uniformly bounded in $W^{1,2}_{\text{loc}}$, hence, is relatively compact in L^2_{loc}. Since $|\nabla|\psi_\delta||\leq|\nabla_{A_\delta}\psi|$, $\{|\nabla|\psi_\delta||\}$ is uniformly bounded in L^2_{loc}. Thus, $\{|\psi_\delta|\}$ is relatively compact in L^2_{loc}. Passing to a subsequence we may assume that $|\psi_\delta|$ converges in $L^2_{\text{loc}}(\mathbb{R}^2)$ as $\delta\to 0$. It follows from Lemma 3.1 that

$$A_\delta(x)=\nabla\chi_\delta(x)+h_\delta\omega(x)+B_\delta(x),$$

where

$$\chi_\delta(x)=\frac{1}{\delta}A(x^\delta)x+\frac{1}{2}[\partial_1A^1(x^\delta)x_1^2+(\partial_1A^2(x^\delta)+\partial_2A^1(x^\delta))x_1x_2+\partial_2A^2(x^\delta)x_2^2].$$

Set $\phi_\delta(x)=\exp(-i\chi_\delta)\psi_\delta(x)$. Then, ϕ_δ satisfies

$$-\nabla^2_{h\omega}\phi_\delta=\frac{\mu(\sigma A)}{\sigma}\phi_\delta+f_\delta(x), \tag{7.4}$$

where $h=H(x^0)$ and

$$f_\delta(x)=-[i\,\mathrm{div}\,B_\delta+2h(h_\delta-h)+2h\omega\cdot B_\delta+|(h_\delta-h)\omega+B_\delta|^2]\phi_\delta \\ -2i[(h_\delta-h)\omega+B_\delta(x)]\cdot\nabla\phi_\delta .$$

Since $|\nabla_{h_\delta\omega+B_\delta}\phi_\delta|=|\nabla_{A_\delta}\psi_\delta|$ and

$$|\nabla\phi_\delta|^2\leqslant|\nabla_{h_\delta\omega+B_\delta}\phi_\delta+i(h_\delta\omega+B_\delta)\phi_\delta|^2\leqslant2|\nabla_{h_\delta\omega+B_\delta}\phi_\delta|^2+2|(h_\delta\omega+B_\delta)\phi_\delta|^2,$$

$\{|\nabla\phi_\delta|\}$ is also uniformly bounded in L^2_{loc}. Passing to another subsequence we have $\phi_\delta\rightarrow\phi_0$ weakly in $W^{1,2}_{\mathrm{loc}}$ and strongly in L^2_{loc}. Since $\mathrm{div}\,B_\delta(x)=(\mathrm{div}\,A)(x^\delta+\delta x)-(\mathrm{div}\,A)(x^\delta)\rightarrow0$, $|B_\delta(x)|\leqslant C\delta|x|^2$ and $h_\delta\rightarrow h$, we have $f_\delta\rightarrow0$ in L^2_{loc}. Hence, the limiting function ϕ_0 satisfies

$$-\nabla^2_{h\omega}\phi_0=a\phi_0 \quad \text{in } \mathbb{R}^2 \tag{7.5}$$

and $|\phi_0(x)|\leqslant1$. Applying Theorem 4.1 to Eq. (7.5) yields that ϕ_0 is smooth.

Denote $\hat\phi_\delta(x)=\phi_\delta(x)-\phi_0(x)$. From (7.4) and (7.5),

$$-\nabla^2_{h\omega}\hat\phi_\delta=a\hat\phi_\delta+\hat f_\delta, \tag{7.6}$$

where

$$\hat f_\delta=f_\delta+\left[\frac{\mu(\sigma A)}{\sigma}-a\right]\phi_\delta\rightarrow0 \quad \text{in } L^2_{\mathrm{loc}}$$

and $\hat\phi_\delta\rightarrow0$ in L^2_{loc}. Applying Lemma 4.2 to (7.6) we get $|\nabla_{h\omega}\hat\phi_\delta|\rightarrow0$ in L^2_{loc}. Since $|\nabla\hat\phi_\delta|^2\leqslant2|\nabla_{h\omega}\hat\phi_\delta|^2+2|h\omega\hat\phi_\delta|^2$ we have $|\nabla\hat\phi_\delta|\rightarrow0$ in L^2_{loc}. So,

$$\hat\phi_\delta\rightarrow0 \quad \text{in } W^{1,2}_{\mathrm{loc}}. \tag{7.7}$$

Denote $\omega=(\omega^1,\omega^2)$, $\nabla_{h\omega^j}=\partial_j-ih\omega^j$. Applying Theorem 4.1 to Eq. (7.6) we have

$$\nabla_{h\omega^j}\nabla_{h\omega^k}\hat\phi_\delta\rightarrow0 \quad \text{in } L^2_{\mathrm{loc}}. \tag{7.8}$$

Note that, for example,

$$\nabla_{h\omega^1}\nabla_{h\omega^1}\hat\phi_\delta=\frac{\partial^2}{\partial x_1}\hat\phi_\delta+ihx_2\frac{\partial}{\partial x_1}\hat\phi_\delta-\frac14|hx_2|^2\hat\phi_\delta .$$

Therefore, (7.7) and (7.8) imply that $\partial_j\partial_k\hat\phi_\delta\rightarrow0$ in L^2_{loc}. So, $\hat\phi_\delta\rightarrow0$ in L^2_{loc}.

Now, we apply the classical C^α estimates to (7.6) and conclude that $\hat\phi_\delta\rightarrow0$ in C^α_{loc}, that is, $\phi_\delta\rightarrow\phi_0$ in C^α_{loc}. Especially, we get $\phi_0(0)=\lim_{\delta\rightarrow0}\phi_\delta(0)=1$. Therefore, ϕ_0 is a nonzero bounded smooth solution of Eq. (7.5) in $\mathbb{R}^2$. From Lemma 2.1 we have

$$a\geqslant\alpha(h)=|h|=|H(x^0)|\geqslant\min_{x\in\Omega}|H(x)|=m(\Omega).$$

Since $a\leqslant m$, we conclude that $a=m$. We also see that if Case 1 happens then $m=m(\Omega)$, $x^0\in\Omega_m$ and (7.1) holds.

Case 2: $x^0\in\partial\Omega$. Now, we shall prove $a\geqslant m(\partial\Omega)$. Let $d_\delta=\mathrm{dist}(\partial\Omega,x^\delta)$, the distance between x^δ and $\partial\Omega$. Then, $B_{d_\delta/\delta}\subset\Omega_\delta$. If there exists a subsequence $\delta_j\rightarrow0$ such that $d_{\delta_j}/\delta_j\rightarrow\infty$, then the argument in Case 1 also gives that $\mu(\sigma_jA)/|\sigma_j|\rightarrow|H(x_0)|$. Therefore, we assume that d_δ/δ is bounded. Passing to a subsequence, we may assume that $d_\delta/\delta\rightarrow d_0$.

Let $\hat{x}^\delta \in \partial\Omega$ such that $|\hat{x}^\delta - x^\delta| = \mathrm{dist}(x^\delta, \partial\Omega) = d_\delta$. At each point $\hat{x}^\delta$ we take a diffeomorphism $\mathcal{F}_\delta : B^+_{R_0} \to \Omega \cap \mathcal{F}_\delta(B_{R_0})$ to straighten a portion of boundary around the point $\hat{x}^\delta$ such that $\mathcal{F}_\delta(0) = \hat{x}^\delta$. For simplicity, we denote $\mathcal{F}_\delta$ by $\mathcal{F}$. We keep in mind that the diffeomorphism $\mathcal{F}$ depends on δ. However, the constant R_0 can be chosen to be independent of δ, thus, we have uniform estimates on $\mathcal{F}$ for all small δ. Let $y^\delta \in B^+_{R_0}$ be such that $\mathcal{F}(y^\delta) = x^\delta$. Then, $|y^\delta| \leqslant C d_\delta \leqslant C\delta$.

Let $\psi^\delta(\mathcal{F}(y)) = u(\mathcal{F}(y)) \tilde{\psi}^\delta(y)$, where u is the positive eigenfunction of Eq. (5.2) associated with the first eigenvalue λ and $\|u\|_{L^\infty} = 1$. Then, $\{|\tilde{\psi}^\delta|\}$ is uniformly bounded from above and $|\tilde{\psi}^\delta(y^\delta)|$ is uniformly bounded away from zero. As in Sec. V we can check that $\tilde{\psi}^\delta$ satisfies the following equation:

$$-\Delta(g)_{\sigma\mathbf{a}} \tilde{\psi}^\delta = D(g)\tilde{v} \cdot D(g)_{\sigma\mathbf{a}} \tilde{\phi}^\delta + [\mu(\sigma A) - \lambda] \tilde{\psi}^\delta \quad \text{in } B^+_{R_0},$$
$$(D(g)_{\sigma\mathbf{a}} \tilde{\psi}^\delta) \cdot \nu = 0 \quad \text{on } \Gamma_{R_0}. \tag{7.9}$$

Here the notations involved are the same as in Secs. III and V.

Define the following rescaled functions and vector fields: $\tilde{\psi}_\delta(y) = \tilde{\psi}^\delta(\delta y)$, $\tilde{v}_\delta(y) = \tilde{v}(\delta y)$, $g_\delta(y) = g(\delta y)$, $\mathbf{a}_\delta(y) = (1/\delta)\mathbf{a}(\delta y)$. Then,

$$-\Delta(g_\delta)_{\mathbf{a}_\delta} \tilde{\psi}_\delta = D(g_\delta)\tilde{v}_\delta \cdot D(g_\delta)_{\mathbf{a}_\delta} \tilde{\psi}_\delta + \frac{\mu(\sigma A) - \lambda}{\sigma} \tilde{\psi}_\delta \quad \text{in } B^+_{R_0/\delta},$$

$$(D(g_\delta)_{\mathbf{a}_\delta} \tilde{\psi}_\delta) \cdot \nu = 0 \quad \text{on } \Gamma_{R_0/\delta}.$$

Recall that $h_\delta = H(x^\delta)$, $h = H(x^0)$. From (3.11) we have

$$\mathbf{a}_\delta(y) = \nabla_y \tilde{\chi}_\delta(y) + h_\delta \tilde{\omega}(y) + \tilde{B}_\delta(y),$$

which holds in $B^+_{R_0/\delta}$, but not in the entire ball. Set $\tilde{\phi}_\delta(y) = \exp(-i\tilde{\chi}_\delta)\tilde{\psi}_\delta$. Then,

$$-\Delta(g_\delta)_{h_\delta\tilde{\omega} + \tilde{B}_\delta} \tilde{\phi}_\delta = D(g_\delta)\tilde{v}_\delta \cdot D(g_\delta)_{h_\delta\tilde{\omega} + \tilde{B}_\delta} \tilde{\phi}_\delta + \frac{\mu(\sigma A) - \lambda}{\sigma} \tilde{\phi}_\delta \quad \text{in } B^+_{R_0/\delta}.$$

Using Theorem 5.5 we obtain

$$\tilde{\phi}_\delta \to \tilde{\phi}_0 \ \text{ weakly in } W^{1,2}_{\mathrm{loc}} \text{ and strongly in } L^2_{\mathrm{loc}}.$$

Write $\tilde{B}_\delta = \tilde{B}^1_\delta \mathbf{e}_1 + \tilde{B}^2_\delta \mathbf{e}_2$, and write the equation for $\tilde{\phi}$ as follows:

$$-\Delta(g_\delta)_{h\tilde{\omega}} \tilde{\phi}_\delta = \frac{\mu(\sigma A)}{\sigma} \tilde{\phi}_\delta + \tilde{f}_\delta \quad \text{in } B^+_{R_0/\delta},$$
$$(D(g_\delta)_{h_\delta\tilde{\omega}} \tilde{\phi}_\delta) \cdot \nu = -i\tilde{B}^2_\delta \tilde{\phi}_\delta \quad \text{on } \Gamma_{R_0/\delta}. \tag{7.10}$$

Note that $D(g_\delta)\tilde{v}_\delta \cdot D(g_\delta)_{h_\delta\tilde{\omega} + \tilde{B}_\delta} \tilde{\phi}_\delta \to 0$ in L^2_{loc} and $\tilde{B}_\delta \to 0$ in L^2_{loc}. So, $\tilde{f}_\delta \to 0$ in L^2_{loc}. We also note that $g_\delta(y) = 1 - \delta y_2 \kappa_r(\delta y_1) \to 1$. Hence, $\tilde{\phi}_0$ satisfies

$$-\nabla^2_{h\omega} \tilde{\phi}_0 = \alpha \tilde{\phi}_0 \quad \text{in } \mathbb{R}^2_+,$$
$$\nabla_{h\omega} \cdot \tilde{\phi}_0 = 0 \quad \text{on } \partial\mathbb{R}^2_+. \tag{7.11}$$

We apply Theorem 5.5 to the equation for $\tilde{\phi}-\tilde{\phi}_0$, then use the classical elliptic estimates to obtain that $\tilde{\phi}_\delta\to\tilde{\phi}_0$ in C^α_{loc}.

Recall that $|y^\delta|/\delta\leqslant C$. By passing to a subsequence, we may assume that $z^\delta\equiv y^\delta/\delta\to z^0$, hence, $\tilde{\phi}_0(z_0)=\lim\tilde{\phi}_\delta(z^\delta)\neq 0$. Therefore, $\tilde{\phi}_0\not\equiv 0$, that is, $\tilde{\phi}_0$ is a nonzero bounded smooth solution of Eq. (7.11). Using Lemma 2.2 we conclude that

$$a\geqslant\beta(h)=\beta_0|h|=\beta_0|H(x^0)|\geqslant\beta_0\min_{x\in\partial\Omega}|H(x)|=m(\partial\Omega)\beta_0.$$

Since $m(\partial\Omega)\beta_0\geqslant m\geqslant a$, we have $m=a$. We also see that if Case 2 happens then $m=m(\partial\Omega)\beta_0$, $x^0\in(\partial\Omega)_m$ and (7.1) holds. □

The proof of Lemma 7.1 has the following consequence.

Proposition 7.2: Assume that $A\in C^2(\bar{\Omega})$. *Let* ψ_σ *be the eigenfunction of Eq. (7.2) associated with the first eigenvalue* $\mu(\sigma A)$ *such that* $\|\psi_\sigma\|_{L^\infty}\leqslant C$. *Then,*

$$|\psi_\sigma|\to 0 \quad \textit{in} \ C^\alpha_{\mathrm{loc}}(\Omega\backslash\Omega_m) \ \textit{and in} \ C^\alpha_{\mathrm{loc}}(\partial\Omega\backslash(\partial\Omega)_m) \ \textit{as} \ \sigma\to\infty.$$

Proof of Theorem 1: Combining Lemmas 6.1, 6.2, 7.1 *yields Theorem 1.* □

As a corollary of Theorem 1 and Proposition 7.2, we have

Theorem 7.3: *Assume* $A\in C^2(\bar{\Omega})$ *and* $\mathrm{curl}\,A(x)\equiv H$, *a nonzero constant. Then*

$$\lim_{\sigma\to\infty}\frac{\mu(\sigma A)}{|\sigma|}=\beta_0|H|. \tag{7.12}$$

Let ψ_σ *be the eigenfunction of (7.2) satisfying* $\|\psi_\sigma\|_{L^\infty}=1$, *then* ψ_σ *concentrates at some points on* $\partial\Omega$, *that is,* $\|\psi_\sigma\|_{L^\infty(\partial\Omega)}\to 1$ *and*

$$\psi_\sigma\to \ \textit{in} \ C^\alpha(\Omega) \quad \textit{as} \ \sigma\to\infty.$$

Proposition 7.2 says that as $\sigma\to\infty$ the eigenfunctions concentrate at some points in $\Omega_m\cup(\partial\Omega)_m$. From the proof of Lemma 7.1 one easily see that, after rescaling near the maximum points and making gauge transitions, the eigenfunctions exhibit profiles of either the eigenfunction of (2.3) in the entire plane $\mathbb{R}^2$ (when interior concentration happens), or the eigenfunction of (2.5) in the half plane $\mathbb{R}^2_+$ (when boundary concentration happens). It will be interesting to find the exact location of the concentration points. In Ref. 1 the concentration behavior of minimal solutions of Ginzburg–Landau equations is studied and the location of concentration is investigated. The arguments used in Ref. 1 can be applied in a similar way to obtain the location of concentration of the eigenfunctions. We should mention that in Ref. 10 Bernoff and Sternberg obtained the location of surface nucleation of superconductivity by using the asymptotic analysis.

ACKNOWLEDGMENTS

This work was partially done when the second author (X.-B.P.) visited the Department of Mathematics at Brigham Young University in the academic year of 1995. He would like to thank the Department for hospitality. This work was partially supported by National Science Foundation Grant No. DMS-9622853 (to K.L.) and by Chinese National Natural Science Foundation, Science Foundation of the Ministry of Education of China, and Zhejiang Provincial Natural Science Foundation of China (to X.-B.P.)

[1] K. Lu and X.-B. Pan, ''Estimates of upper critical external magnetic field for the Ginzburg–Landau equation,'' Physica D **127**, 73–104 (1999).
[2] D. Saint-James and P. de Gennes, ''Onset of superconductivity in decreasing fields,'' Phys. Lett. **7**, 306–308 (1963).
[3] S. J. Chapman, ''Nucleation of superconductivity in decreasing fields, 1, 2,'' Eur. J. Appl. Math. **5**, 449–468 (1994); **5**, 469–494 (1994).
[4] K. Lu and X.-B. Pan, ''Gauge invariant eigenvalue problems in $\mathbb{R}^2$ and in $\mathbb{R}^2_+$,'' Trans. Am. Math. Soc. (to be published).

[5] V. Ginzburg and L. Landau, ''On the theory of superconductivity,'' Sov. Phys. JETP **20**, 1064–1082 (1950).
[6] P. De Gennes, *Superconductivity of Metals and Alloys* (Benjamin, New York, 1966).
[7] P. Bauman, D. Phillips, and Q. Tang, ''Stable nucleation for the Ginzburg–Landau system with an applied magnetic field,'' Arch. Ration. Mech. Anal. **142**, 1–43 (1998).
[8] F. Bethuel, H. Brézis, and F. Hélein, *Ginzburg–Landau Vortices* (Birkhöuser, Boston, 1994).
[9] F. Bethuel and T. Riviére, ''Vortices for a variational problem related to superconductivity,'' Ann. Inst. Henri Poincaré Anal. Non Linéaire **12**, 243–303 (1995).
[10] A. Bernoff and P. Sternberg, ''Onset of superconductivity in decreasing fields for general domains,'' J. Math. Phys. **39**, 1272–1284 (1998).
[11] S. J. Chapman, S. D. Howison, and J. R. Ockendon, ''Macroscopic models for superconductivity,'' SIAM (Soc. Ind. Appl. Math.) Rev. **34**, 529–560 (1992).
[12] S. Chanillo and M. Kiessling, ''Symmetry of solutions of Ginzburg–Landau equations,'' Comp. Rend. Acad. Sci., Paris, **32**, 1023–1026 (1995).
[13] Q. Du, M. Gunzburger, and J. Peterson, ''Analysis and approximation of the Ginzburg–Landau model of superconductivity,'' SIAM (Soc. Ind. Appl. Math.) Rev. **34**, 45–81 (1992).
[14] A. Jaffe and C. Taubes, *Vortices and Monopoles* (Birkhäuser, Boston, 1980).
[15] F.-H. Lin, ''Solutions of Ginzburg–Landau equations and critical points of the renormalized energy,'' Ann. Inst. Henri Poincaré, Anal. Non Linéaire **12**, 599–622 (1995).
[16] F.-H. Lin and T.-C. Lin, ''Minimax solutions of the Ginzburg–Landau equations,'' Selecta Math. (N.S.) **3**, 99–113 (1997).
[17] K. Lu and X.-B. Pan, ''Ginzburg–Landau equation with DeGennes boundary condition,'' J. Diff. Eqns. **129**, 136–165 (1996).
[18] P. Mironescu, ''On the stability of radial solutions of the Ginzburg–Landau equations,'' J. Funct. Anal. **130**, 334–347 (1995).
[19] M. Struwer, ''On the asymptotical behavior of minimizers of the Ginzburg–Landau model in 2-dimensions,'' J. Diff. Eqns. **7**, 1613–1627 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Slater sum for central field problems characterized by its *s*-wave component alone

N. H. March
University of Oxford, Oxford, England

L. M. Nieto
Departamento de Física Teórica, Universidad de Valladolid, 47011 Valladolid, Spain

C. Amovilli
Department of Chemistry and Industrial Chemistry, University of Pisa, 56126 Pisa, Italy

L. C. Balbás
Departamento de Física Teórica, Universidad de Valladolid, 47011 Valladolid, Spain

M. L. Glasser
Department of Physics, Clarkson University, Potsdam, New York 13676

(Received 22 December 1998; accepted for publication 25 February 1999)

For the hydrogenlike atom, with central potential $-Z/r$, partial differential equations exist for the Slater sum $\mathcal{Z}(r,\beta)$ $[\beta=(k_\mathrm{B}T)^{-1}]$ and for its *s*-wave $(l=0)$ component $\mathcal{Z}_0(r,\beta)$. It is shown that Z can be eliminated, to lead to a result in which $\mathcal{Z}(r,\beta)$ is solely characterized by $\mathcal{Z}_0(r,\beta)$. A similar situation is exhibited for the three-dimensional isotropic harmonic oscillator, for which closed forms of both $\mathcal{Z}(r,\beta,\omega)$ and $\mathcal{Z}_0(r,\beta,\omega)$ can be obtained explicitly. Finally, a third central field problem is considered in which independent electrons are confined within a sphere of radius R, but are otherwise free. We are able to derive explicitly for this model the *s*-wave component $\mathcal{Z}_0(r,\beta,R)$. The full Slater sum $\mathcal{Z}(r,\beta,R)$ then is also analyzed in some detail. © *1999 American Institute of Physics.*
[S0022-2488(99)01706-5]

I. BACKGROUND

The single particle level density, or density of states $g(E)$, of a quantum system described by a one-particle Hamiltonian with a local effective potential $V(\mathbf{r})$ is given by the inverse Laplace transform of the canonical partition function $Z(\beta)$ divided by β, $Z(\beta)$ being obtained as the trace of the canonical density matrix

$$C(\mathbf{r},\mathbf{r}_0,\beta)=\sum_n \chi_n(\mathbf{r})\chi_n^*(\mathbf{r}_0)e^{-\beta\epsilon_n}. \tag{1}$$

Here ϵ_n are the eigenvalues of the energy for the states $\chi_n(\mathbf{r})$, and $\beta=(k_\mathrm{B}T)^{-1}$, with k_B Boltzmann's constant, and T the absolute temperature. In the context of semiclassical periodic orbit theory, the so-called "trace formula" [1] for the oscillating part of $g(E)$ provides the basis for many interesting interpretations of quantum phenomena in terms of classical orbits. Examples are given in Ref. 1 of applications in atomic nuclei, metal clusters, and semiconductor quantum dots.

In this paper we are interested in the diagonal part of the canonical density matrix, the so-called Slater sum

$$\mathcal{Z}(\mathbf{r},\beta)=C(\mathbf{r},\mathbf{r},\beta). \tag{2}$$

In particular, we construct this Slater sum for several illustrative central field problems. In the early work of March and Murray,[2] central field problems were considered for the Slater sum

$\mathcal{Z}(r,\beta)$, and for its analysis into its angular momentum components $\mathcal{Z}_l(r,\beta)$. Here, if $V(r)$ denotes a general central potential energy, and it generates radial eigenfunctions $\chi_{nl}(r)$ and corresponding eigenvalues ϵ_{nl}, then following Ref. 2

$$4\pi\mathcal{Z}_l(r,\beta)=\sum_n \chi_{nl}(r)\chi_{nl}^*(r)\exp(-\beta\epsilon_{nl}). \tag{3}$$

The full Slater sum $\mathcal{Z}(r,\beta)$ is constructed from $Z_l(r,\beta)$ as[2]

$$\mathcal{Z}(r,\beta)=\sum_{l=0}^{\infty}(2l+1)\mathcal{Z}_l(r,\beta). \tag{4}$$

In the present work, we shall focus on the *s*-wave component $l=0$, $\mathcal{Z}_0(r,\beta)$ in Eq. (3), and the way in which this can be used in specific central field problems to construct the full Slater sum $\mathcal{Z}(r,\beta)$. We turn immediately in Sec. II below to the first of three central field examples, namely the three-dimensional isotropic harmonic oscillator. In Sec. III some results concerning the Coulomb potential are derived. In Sec. IV we deal with the case of a spherical cavity of radius *R*. The *s*-wave component $\mathcal{Z}_0$ of the Slater sum is explicitly evaluated. In Sec. V we gather some discussion and a summary, while more technical details are relegated to the Appendices.

II. THREE-DIMENSIONAL ISOTROPIC OSCILLATOR

For the three-dimensional isotropic harmonic oscillator $V(r)=\omega^2r^2/2$, the evaluation of the Slater sum $\mathcal{Z}(r,\beta,\omega)$ essentially goes back to the pioneering work by Sondheimer and Wilson[3] on free electrons in an external magnetic field. These workers, in fact, obtained for this problem the canonical density matrix $C(\mathbf{r},\mathbf{r}_0,\beta,\omega)$, which is essentially equivalent to the oscillator example under discussion here (see also Ref. 4). In Appendix A, we record first the full results for the canonical density matrix from which the Slater sum $\mathcal{Z}(r,\beta,\omega)$ is immediately obtained as its diagonal element $C(\mathbf{r},\mathbf{r}_0=\mathbf{r},\beta,\omega)$:

$$\mathcal{Z}(r,\beta,\omega)=\left[\frac{\omega}{2\pi\sinh(\beta\omega)}\right]^{3/2}\exp\left[-\omega r^2\tanh\left(\frac{\beta\omega}{2}\right)\right]. \tag{5}$$

It is important to mention that Amovilli and March[5] found a differential equation satisfied by the Slater sum for the harmonic oscillator [see Eqn. (2.10) of Ref. 5], namely

$$\frac{1}{8}\frac{\partial^3\mathcal{Z}}{\partial r^3}+\frac{1}{4r}\frac{\partial^2\mathcal{Z}}{\partial r^2}-\left[\frac{1}{4r^2}+V+\frac{\partial}{\partial\beta}\right]\frac{\partial\mathcal{Z}}{\partial r}+\frac{1}{2}\frac{dV}{dr}\mathcal{Z}=0. \tag{6}$$

Going back to Eq. (5), note that as $\omega\to 0$, it reduces to

$$\mathcal{Z}_\infty(\beta)=(2\pi\beta)^{-3/2}, \tag{7}$$

which is the correct partition function for free particles, per unit volume.

What, however, is important for the present consideration is that the knowledge of the canonical density matrix $C(\mathbf{r},\mathbf{r}_0,\beta,\omega)$ allows the *s*-state component of the Slater sum $Z_0(r,\beta,\omega)$ to be directly extracted, and this is also accomplished in Appendix A. The final result for $\mathcal{Z}_0(r,\beta,\omega)$ is given there by Eqs. (A9) and (A10):

$$\mathcal{Z}_0(r,\beta,\omega)=\left[\frac{\omega}{2\pi\sinh(\beta\omega)}\right]^{3/2}\frac{e^{r^2\mu}-e^{-r^2\mu}}{2r^2\mu}\exp\left(-\frac{\omega}{2}r^2\left[\coth\frac{\beta\omega}{2}+\tanh\frac{\beta\omega}{2}\right]\right), \tag{8}$$

where

$$\mu = \frac{\omega}{2}\left[\coth\frac{\beta\omega}{2} - \tanh\frac{\beta\omega}{2}\right].$$

This reduces to the free-particle result of March and Murray,[2] when the ''force constant'' ω is reduced to zero, i.e.,

$$\mathcal{Z}_0(r,\beta,\omega\to 0) = \frac{1}{2(2\pi)^{3/2}\beta^{1/2}}\frac{[1-\exp(-2r^2/\beta)]}{r^2}. \tag{9}$$

We discuss the relation between $\mathcal{Z}(r,\beta)$ and its s-state component $\mathcal{Z}_0(r,\beta)$ in Appendix A, Sec. 2. We shall also return briefly to this example after discussing the hydrogenlike atom. However, we can generalize $\mathcal{Z}_0(r,\beta,\omega)$ in Eq. (8) to a general orbital quantum number l, the result for $\mathcal{Z}_l(r,\beta,\omega)$ being given also in Appendix A.

III. COULOMB CASE

March and Murray[2] derive for the s-state component $\mathcal{Z}_0(r,\beta)$ of the Slater sum $\mathcal{Z}(r,\beta)$ in a hydrogenlike atom with central potential energy $-Z/r$ the partial differential equation [see Ref. 2, equation (4.7)]

$$\frac{1}{8}\frac{\partial^3}{\partial r^3}(r^2\mathcal{Z}_0) - \frac{\partial^2}{\partial r\partial\beta}(r^2\mathcal{Z}_0) + \frac{Z}{r}\frac{\partial}{\partial r}(r^2\mathcal{Z}_0) - \frac{Z\mathcal{Z}_0}{2} = 0. \tag{10}$$

Let us now use in this Eq. (10) the ''generalized Kato'' result of March,[6] namely (see also Refs. 5, 7, and 8)

$$\frac{\partial\mathcal{Z}}{\partial r} = -2Z\mathcal{Z}_0. \tag{11}$$

By replacing $Z\mathcal{Z}_0$ in Eq. (10) by half the lhs of Eq. (11), one finds

$$\frac{1}{8}\frac{\partial^3}{\partial r^3}(r^2\mathcal{Z}_0) - \frac{\partial^2}{\partial r\partial\beta}(r^2\mathcal{Z}_0) = \frac{1}{2r}\frac{\partial}{\partial r}\left(r^2\frac{\partial\mathcal{Z}}{\partial r}\right) - \frac{1}{4}\frac{\partial\mathcal{Z}}{\partial r}. \tag{12}$$

This, then, is an equation relating the total Slater sum for the hydrogenlike atom directly to its s-state component, the nuclear potential energy $-Z/r$ no longer appearing explicitly. Thus Eq. (10) is the counterpart of the relation between $\mathcal{Z}$ and $\mathcal{Z}_0$ referred to in Sec. II for the harmonic oscillator [see also Eq. (6) below]. However, neither $\mathcal{Z}$ nor $\mathcal{Z}_0$ is yet known in closed form for the Coulomb case.

We can expect Eq. (12) to apply even when we ''switch off'' Z, to reach the ''free-electron limit'' in, initially, an infinite volume. In this limiting case, the function $\mathcal{Z}(r,\beta)$ will take the constant value $(2\pi\beta)^{-3/2}$ [see also Eq. (7) above], which in the units adopted here, is simply the partition function per unit volume of free electrons. Then the terms involving $\mathcal{Z}(r,\beta)$ in Eq. (12) become zero and we regain the free-particle s-state-only equation already solved by March and Murray,[2] namely

$$\frac{1}{8}\frac{\partial^3}{\partial r^3}(r^2\mathcal{Z}_0) - \frac{\partial^2}{\partial r\partial\beta}(r^2\mathcal{Z}_0) = 0, \tag{13}$$

with explicit solution given in Eq. (9) above. In fact, Eq. (13) is readily integrated with respect to r to yield

$$\frac{1}{2}\frac{\partial^2}{\partial r^2}(r^2\mathcal{Z}_0) - 4\frac{\partial}{\partial\beta}(r^2\mathcal{Z}_0) = (2\pi\beta)^{-3/2} = \mathcal{Z}_\infty(\beta), \tag{14}$$

again relating the Slater sum $\mathcal{Z}(r,\beta)\equiv\mathcal{Z}_\infty(\beta)$ directly to its s-state component $\mathcal{Z}_0(r,\beta)$ of Eq. (9).

IV. SPHERICAL BARRIER WITH FINITE RADIUS *R*

We turn to the third central field problem to be treated in the present study: namely noninteracting electrons confined by an infinite spherical barrier to move inside a sphere of radius R, but which are otherwise free. Fortunately, in this case, wave functions and energy levels can be obtained explicitly for the s states as the radial functions are given by

$$\chi_{n0}(\mathbf{r})=\sqrt{\frac{2}{R}}\,\frac{\sin(n\pi r/R)}{r}, \tag{15}$$

and the corresponding eigenvalues in atomic units $(\hbar=m=1)$ are

$$\epsilon_n=\frac{\pi^2}{2R^2}n^2. \tag{16}$$

Hence, inserting Eqs. (15) and (16) into Eq. (3), and using the definition of the Jacobi function ϑ_3 (see Ref. 9)

$$\vartheta_3(z,q)=1+2\sum_{n=1}^{\infty} q^{n^2}\cos(2nz), \quad |q|<1, \tag{17}$$

Eq. (18) below is found for $\mathcal{Z}_0(r,\beta,R)$:

$$\mathcal{Z}_0(r,\beta,R)=\frac{1}{8\pi r^2 R}\left[\vartheta_3\left(0,\exp\left(-\frac{\beta\pi^2}{2R^2}\right)\right)-\vartheta_3\left(\frac{\pi r}{R},\exp\left(-\frac{\beta\pi^2}{2R^2}\right)\right)\right]. \tag{18}$$

The off-diagonal form $C_0(r,r_0,\beta,R)$ is derived in Appendix B.

From Eq. (18) it is straightforward to derive the result

$$\frac{1}{8}\,\frac{\partial^2(r^2\mathcal{Z}_0)}{\partial r^2}-\frac{\partial(r^2\mathcal{Z}_0)}{\partial\beta}=\frac{\pi}{8}\sum_{n=0}^{\infty}\frac{1}{R}\left(\frac{n}{R}\right)^2\exp\left(-\frac{\beta\pi^2}{2}\left(\frac{n}{R}\right)^2\right), \tag{19}$$

which is just a function of β and R, and therefore is constant for $r<R$ (for $r>R$ it is zero). There is a discontinuity at $r=R$, reflecting the fact that the potential we are dealing with has a strong singularity at this boundary. Observe that the function $r^2\mathcal{Z}_0$ satisfies Eq. (10) with $Z=0$, or equivalently (13), but with a strong singularity at the border $r=R$.

We note that for large R, the limiting process included in Eq. (18) can be carried out essentially by replacing the summation involved for the theta function in Eq. (17) by an integration; such a replacement is valid only as $R\to\infty$. Then Eq. (18) can be rewritten as

$$\begin{aligned}\lim_{R\to\infty}\mathcal{Z}_0(r,\beta,R)&=\frac{1}{2\pi r^2}\lim_{R\to\infty}\sum_{n=1}^{\infty}\frac{\sin^2(\pi rn/R)}{R}e^{-(\beta\pi^2/2)(n/R)^2}\\&=\frac{1}{2\pi r^2}\int_0^\infty e^{-(\beta\pi^2/2)x^2}\sin^2(\pi rx)\,dx,\end{aligned} \tag{20}$$

which can be shown to be equivalent to

$$\mathcal{Z}_0(r,\beta,\infty)=\sum_{n=1}^{\infty}\frac{(-4r^2)^{n-1}}{(2n)!\,\pi^2}\left[\frac{2}{\beta}\right]^{n+1/2}\int_0^\infty e^{-y^2}y^{2n}\,dy. \tag{21}$$

The integration appearing in (21) is a gamma function. Therefore, after some straightforward calculation the previous result can be written as

$$\mathcal{Z}_0(r,\beta,\infty)=-\frac{1}{8\pi^2r^2}\sqrt{\frac{2\pi}{\beta}}\sum_{n=1}^{\infty}\frac{1}{n!}\left(-\frac{2r^2}{\beta}\right)^n, \tag{22}$$

which can be summed to regain the limiting form (9).

In Appendix B, some progress is made in setting up a differential equation for the Slater sum $\mathcal{Z}(r,\beta,R)$. However, in contrast to the earlier examples of the harmonic oscillator and the hydrogen atom, near-diagonal behavior of $C(\mathbf{r},\mathbf{r}_0,\beta,R)$ is required in the formalism available to date. Therefore, although the *s*-wave component $\mathcal{Z}_0(r,\beta,R)$ is completely solved, full information relating $\mathcal{Z}$ and $\mathcal{Z}_0$ is still lacking except in the limit as R tends to infinity.

V. DISCUSSION AND SUMMARY

As is evident from Eq. (12) for the hydrogenlike atom, the knowledge of the *s*-state component for the Slater sum $\mathcal{Z}_0(r,\beta)$ determines the entire sum $\mathcal{Z}(r,\beta)$ by integration of an ordinary differential equation to yield

$$\frac{\partial\mathcal{Z}}{\partial r}=f(\beta)+e^{-3r^2/4}\int^r 2re^{3r^2/4}\left[\frac{1}{8}\frac{\partial^3(r^2\mathcal{Z}_0)}{\partial r^3}-\frac{\partial^2(r^2\mathcal{Z}_0)}{\partial r\partial\beta}\right]dr, \tag{23}$$

$f(\beta)$ being an arbitrary function. Evidently then, $\mathcal{Z}(r,\beta)$ can be written explicitly by a further quadrature of Eq. (23). Naturally, physical boundary conditions must be imposed. Equation (23) makes explicit the idea motivating the present investigation: namely the characterization of the Slater sum by its *s*-state component.

For the other examples considered here, the more complex analog of (23) is embodied in Eqs. (A11) and (A12) for the three-dimensional isotropic harmonic oscillator. However, in this specific example, both $\mathcal{Z}$, already given in the work of Ref. 3, and $\mathcal{Z}_0$ derived in the present paper and given in (8), are known explicitly. Finally, new results are given for the *s*-state component $\mathcal{Z}_0$ in Eq. (18) although a satisfactory integro-differential equation for the full Slater sum $\mathcal{Z}(r,\beta,R)$ for the spherical barrier problem is not yet found (see Appendix B).

ACKNOWLEDGMENTS

One of us (NHM) made his contribution to this investigation during visits to the Theoretical Physics Department of the University of Valladolid. He wishes to thank especially Professor J. A. Alonso and his colleagues for much stimulating discussion and for generous hospitality. NHM also acknowledges partial financial support from the Leverhulme Trust, UK. MLG acknowledges support by Junta de Castilla y León and the kind hospitality at University of Valladolid. This work has been partially supported by Ministerio de Educación y Cultura (DGES Project No. PB94-1115) and by Junta de Castilla y León (CO2/197).

APPENDIX A: SOME RESULTS FOR THE ISOTROPIC THREE-DIMENSIONAL HARMONIC OSCILLATOR

1. Canonical density matrix

As used, for example, in Eq. (2.4) of Ref. 4, if the potential energy of the oscillator is $\frac{1}{2}\omega^2r^2$, then

$$C(\mathbf{r},\mathbf{r}_0,\beta,\omega)=\left[\frac{\omega}{2\pi\sinh(\beta\omega)}\right]^{3/2}\exp\left[-\frac{\omega|\mathbf{r}-\mathbf{r}_0|^2}{4}\coth\frac{\beta\omega}{2}\right]\exp\left[-\frac{\omega|\mathbf{r}+\mathbf{r}_0|^2}{4}\tanh\frac{\beta\omega}{2}\right]. \tag{A1}$$

In the limit $\omega\rightarrow 0$ we have

$$C_{\text{free}}(\mathbf{r},\mathbf{r}_0,\beta)=\frac{1}{(2\pi\beta)^{3/2}}\exp\left[-\frac{|\mathbf{r}-\mathbf{r}_0|^2}{2\beta}\right]. \tag{A2}$$

This is translationally invariant, coming essentially from plane waves $\exp(i\mathbf{k}\cdot\mathbf{r})$.

Evidently, from Eq. (A1), the result (5) of the main text is recovered by setting $\mathbf{r}_0=\mathbf{r}$. In the limit $\omega\to 0$ this equation gives the correct result (7) above.

To get the l-state matrix $C_l(r,r_0,\beta,\omega)$, let us introduce in Eq. (A1)

$$C(\mathbf{r},\mathbf{r}_0,\beta,\omega)=\sum_{k=0}^{\infty}(2k+1)C_k(r,r_0,\beta,\omega)P_k(\cos\gamma), \tag{A3}$$

where γ is the angle between $\mathbf{r}$ and $\mathbf{r}_0$, and $P_k(x)$ are Legendre polynomials.

Now extract the "l" state C_l from (A3) by multiplying both sides by $P_l(\cos\gamma)$, and by $\sin\gamma$, and integrating from $\gamma=0$ to $\gamma=\pi$:

$$C_l(r,r_0,\beta,\omega)=\frac{\mathbb{L}_l(a)}{2}\left[\frac{\omega}{2\pi\sinh(\beta\omega)}\right]^{3/2}\exp\left[-\frac{\omega(r^2+r_0^2)}{4}\left(\coth\frac{\beta\omega}{2}+\tanh\frac{\beta\omega}{2}\right)\right]. \tag{A4}$$

The essential integral involving γ to be evaluated is

$$\mathbb{L}_l(a)=\int_0^{\pi}\exp(a\cos\gamma)\sin\gamma P_l(\cos\gamma)d\gamma=\int_{-1}^{1}\exp(as)P_l(s)ds, \tag{A5}$$

with

$$a=\frac{rr_0\omega}{2}\left[\coth\frac{\beta\omega}{2}-\tanh\frac{\beta\omega}{2}\right].$$

The last integral (A5) can be evaluated in different forms. For example, we can give an expression in terms of the derivatives of Legendre polynomials $P_l^{(n)}(s)$ evaluated at ± 1:

$$\mathbb{L}_l(a)=\left[\frac{e^{as}}{a}\sum_{n=0}^{l}\frac{(-1)^n}{a^n}P_l^{(n)}(s)\right]_{-1}^{1}=\frac{1}{a^{l+1}}\sum_{n=0}^{l}(-1)^n a^{l-n}\{e^aP_l^{(n)}(1)-e^{-a}P_l^{(n)}(-1)\}. \tag{A6}$$

However, an equivalent and more interesting expression is obtained when using the explicit form of Legendre polynomials:

$$\mathbb{L}_l(a)=\frac{1}{2^l}\sum_{m=0}^{[l/2]}(-1)^m\binom{l}{m}\binom{2l-2m}{l}\frac{d^{l-2m}\mathbb{L}_0(a)}{da^{l-2m}}. \tag{A7}$$

For example, the first terms are

$$\mathbb{L}_0(a)=\frac{e^a-e^{-a}}{a},\quad \mathbb{L}_1(a)=\frac{d\mathbb{L}_0(a)}{da}. \tag{A8}$$

Using the previous results, C_0 can be evaluated explicitly. In particular, on the diagonal $r=r_0$, and $C_0(r,r,\beta,\omega)=\mathcal{Z}_0(r,\beta,\omega)$. Hence

$$\mathcal{Z}_0(r,\beta,\omega)=\left[\frac{\omega}{2\pi\sinh(\beta\omega)}\right]^{3/2}\frac{e^{r^2\mu}-e^{-r^2\mu}}{2r^2\mu}\exp\left(-\frac{\omega}{2}r^2\left[\coth\frac{\beta\omega}{2}+\tanh\frac{\beta\omega}{2}\right]\right), \tag{A9}$$

where

$$\mu=\frac{\omega}{2}\left[\coth\frac{\beta\omega}{2}-\tanh\frac{\beta\omega}{2}\right]. \tag{A10}$$

As $\omega\to 0$, the infinite volume free particle result for $\mathcal{Z}_0$ is regained.

2. Relation between Slater sum $\mathcal{Z}(r,\beta,\omega)$ and *s*-wave component $\mathcal{Z}_0(r,\beta,\omega)$

We add here that an explicit, though somewhat complicated, relation exists between $\mathcal{Z}$ and $\mathcal{Z}_0$ for the harmonic oscillator follows by using (i) the general relation between $V(r)$ and $\mathcal{Z}_0$ and (ii) the specific relation between $V(r)$ and $\mathcal{Z}$ given in (6).

The analog of Eq. (10) for the case of the harmonic oscillator is immediately found from Eq. (4.7) of Ref. 2 by putting $V(r)=\frac{1}{2}\omega^2 r^2$. However, in the context of the present work it is helpful to note that Eq. (41) of Ref. 2 can be viewed as a first-order differential equation for $V(r)$, given the *s* state component $\mathcal{Z}_0(r,\beta)$ of the Slater sum. The integral of this equation then takes the form

$$V(r)=\frac{1}{Q^2}\int^r\left[\frac{1}{4}QQ'''-2Q\frac{\partial Q'}{\partial\beta}\right]dr,\quad Q=r^2\mathcal{Z}_0, \tag{A11}$$

which, in fact, is a quite general central field equation characterizing V and $\mathcal{Z}_0$.

The corresponding equation for $\mathcal{Z}$, but now specific for the harmonic oscillator, can be obtained from Ref. 5 as

$$V(r)=\mathcal{Z}^2\int^r\frac{2}{\mathcal{Z}^3}\left[\frac{\partial\mathcal{Z}'}{\partial\beta}-\frac{\mathcal{Z}'}{4r^2}-\frac{\mathcal{Z}''}{2r}-\frac{\mathcal{Z}'''}{8}\right]dr. \tag{A12}$$

Clearly, equating Eqs. (A11) and (A12) gives the relation between the Slater sum $\mathcal{Z}$ and its *s*-wave component $\mathcal{Z}_0$ for the three-dimensional harmonic oscillator.

APPENDIX B: CANONICAL DENSITY MATRIX FOR SPHERICAL BARRIER

1. $l=0$ component of canonical density matrix

Let us consider the $l=0$ solutions of the Schrödinger equation in three dimensions for a free particle inside a spherical barrier of radius R. Using the normalized radial eigenfunctions and the eigenvalues given in (15) and (16), we can evaluate the *s*-state component of the canonical density matrix

$$\begin{aligned}C_0(r,r_0,\beta,R)&=\frac{1}{4\pi}\sum_{n=1}^{\infty}\chi_{n0}(\mathbf{r})\chi^*_{n0}(\mathbf{r})e^{-\beta\epsilon_n}\\&=\frac{1}{4\pi rr_0R}\sum_{n=1}^{\infty}e^{-\beta n^2\pi^2/2R^2}\left(\cos\left[\frac{2n\pi(r-r_0)}{2R}\right]-\cos\left[\frac{2n\pi(r+r_0)}{2R}\right]\right)\\&=\frac{1}{8\pi rr_0R}\left[\vartheta_3\left(\frac{\pi(r-r_0)}{2R},e^{-\beta\pi^2/2R^2}\right)-\vartheta_3\left(\frac{\pi(r+r_0)}{2R},e^{-\beta\pi^2/2R^2}\right)\right].\end{aligned} \tag{B1}$$

We have used here the standard definition of one of the theta functions[9] given in (17). If we consider the limit $q\to 1$, we get a Fourier series expansion that is precisely a ''Dirac comb''[10]

$$\vartheta_3(z,1)=\pi\sum_{k=-\infty}^{\infty}\delta(z-k\pi). \tag{B2}$$

Therefore, taking the limit $\beta\to 0$ in (B1) we find

$$4\pi r r_0 C_0(r,r_0,0,R)=\sum_{k=-\infty}^{\infty}[\delta(r-r_0-2Rk)-\delta(r+r_0-2Rk)];\tag{B3}$$

for the relevant range of variation of the radial variables, only one of the Dirac delta functions contributes. Therefore,

$$4\pi r r_0 C_0(r,r_0,0,R)=\delta(r-r_0),\tag{B4}$$

and one of the required conditions on the canonical density matrix is satisfied.

By taking $r=r_0$ in (B1), we obtain the s-wave component of the Slater sum for this problem, the result being given in (18). From this expression, it is easy to prove that

$$\mathcal{Z}_0(0,\beta,R)=\frac{\pi}{2R^3}\sum_{n=1}^{\infty} n^2 e^{-\beta n^2\pi^2/2R^2}=\mathcal{Z}(0,\beta,R).\tag{B5}$$

This function presents a divergence for $\beta/R^2\rightarrow 0$, which is represented explicitly by

$$R^3\mathcal{Z}(0,\beta,R)\approx\frac{1}{(2\pi)^{3/2}}\left(\frac{\beta}{R^2}\right)^{-3/2}.\tag{B6}$$

In addition, one can easily check that the function $\mathcal{Z}_0(r,\beta,R)$ of (18) satisfies Eq. (13). This is an immediate consequence of the fact that the ϑ_3 function obeys the heat equation.[11]

2. Additional comments on the canonical density matrix

From Eq. (4.6) of March and Murray[2] for $V=0$, and after multiplication by $2r^2$ we get

$$r^2\frac{\partial^2}{\partial r^2}(rC_l)\bigg|_{r'=r}-l(l+1)(r\mathcal{Z}_l)-2r^2\frac{\partial}{\partial\beta}(r\mathcal{Z}_l)=0.\tag{B7}$$

Retaining the ‘‘pathological’’ dV/dr, we have for the diagonal quantity $\mathcal{Z}_l$, again with $V=0$:

$$\frac{1}{8r}\frac{\partial^3}{\partial r^3}(r^2\mathcal{Z}_l)-\frac{1}{2r^2}\frac{\partial}{\partial r}(rl(l+1)\mathcal{Z}_l)-\frac{1}{2}\frac{\partial V}{\partial r}(r^2\mathcal{Z}_l)-\frac{1}{r}\frac{\partial^2}{\partial\beta\partial r}(r^2\mathcal{Z}_l)=0.\tag{B8}$$

Substitute from Eq. (B7) for $l(l+1)r\mathcal{Z}_l$ into Eq. (B8) to find

$$\frac{1}{8r}\frac{\partial^3}{\partial r^3}(r^2\mathcal{Z}_l)-\frac{1}{2r^2}\left\{\frac{\partial}{\partial r}\left[r^2\frac{\partial^2}{\partial r^2}(rC_l)\bigg|_{r'=r}-2r^2\frac{\partial}{\partial\beta}(r\mathcal{Z}_l)\right]\right\}-\frac{1}{r}\frac{\partial^2}{\partial\beta\partial r}(r^2\mathcal{Z}_l)=\frac{1}{2r}\frac{\partial V}{\partial r}(r^2\mathcal{Z}_l).\tag{B9}$$

We note here that the great simplification for $l=0$, fully solved in the present work, is that the square bracket in Eq. (B9) is identically zero from Eq. (B7).

To make progress with Eq. (B9) we now appeal to the relation utilized earlier for the harmonic oscillator:

$$C_l=\int_0^{\pi} CP_l(\cos\gamma)\sin\gamma\, d\gamma,\tag{B10}$$

where P_l are Legendre polynomials. Use of Eq. (B10) moves the (as yet unknown) l dependence of C_l [expect for $R\rightarrow\infty$ where we have the March and Murray result for $\mathcal{Z}_l(r,\beta,R\rightarrow\infty)$ in terms of Bessel functions $I_{l+1/2}$, namely $4\pi r\beta\mathcal{Z}_l(r,\beta)=\exp(-r^2/\beta)I_{l+1/2}(r^2/\beta)$] entirely into the P_l's. Inserting Eq. (B10) into (B9) we can now multiply throughout by $(2l+1)$ and then sum over all l to find

$$\frac{1}{2r}\frac{\partial V}{\partial r}(r^2\mathcal{Z})=\frac{1}{8r}\frac{\partial^3}{\partial r^3}(r^2\mathcal{Z})-\frac{1}{r}\frac{\partial^2}{\partial\beta\partial r}(r^2\mathcal{Z})-\frac{1}{2r^2}\frac{\partial}{\partial r}\times\left[r^2\frac{\partial^2}{\partial r^2}\left(r\int_0^\pi\sum_{l=0}^\infty(2l+1)CP_l(\cos\gamma)\sin\gamma\,d\gamma\right)\bigg|_{r'=r}-2r^2\frac{\partial}{\partial\beta}(r\mathcal{Z})\right]. \tag{B11}$$

To our knowledge, Eq. (B11) is new. But to get an explicit differential equation for the Slater sum $\mathcal{Z}(r,\beta,R)$ for the spherical barrier, such as Amovilli and March give for both the harmonic oscillator and the hydrogen atom, one must be able to perform the sum over all l in the square brackets of Eq. (B11) and then express solely in terms of the diagonal of C, i.e., $\mathcal{Z}(r,\beta,R)$. Or, within the philosophy of the present work, one could consider relating the off-diagonal term in Eq. (B11) to the known $\mathcal{Z}_0(r,\beta,R)$ for the spherical barrier. We have, so far, not achieved such further simplification for the barrier problem.

[1] M. Brack and R. K. Bhaduri, *Semiclassical Physics* (Addison-Wesley, Reading, MA, 1997).
[2] N. H. March and A. M. Murray, Phys. Rev. **120**, 830 (1960).
[3] E. H. Sondheimer and A. H. Wilson, Proc. R. Soc. London, Ser. A **210**, 173 (1951).
[4] N. H. March, J. Math. Phys. **38**, 2037 (1997).
[5] C. Amovilli and N. H. March, Phys. Chem. Liq. **30**, 135 (1995).
[6] N. H. March, Phys. Lett. A **111**, 47 (1985).
[7] I. L. Cooper, Phys. Rev. A **50**, 1040 (1994).
[8] S. Pfalzner, H. Lehmann, and N. H. March, J. Math. Chem. **16**, 9 (1994).
[9] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).
[10] D. Mumford, *Tata Lectures on Theta I* (Birkhäuser, Boston, 1983).
[11] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids* (Clarendon, Oxford, 1986), 2nd ed.

Flavor symmetry of the tensor Dirac theory

Frank Reifler and Randall Morris
Lockheed Martin Corporation, Government Electronic Systems 137-227, 199 Borton Landing Road, Moorestown, New Jersey 08057

(Received 7 July 1998; accepted for publication 12 January 1999)

Recently, the Dirac and Einstein equations were unified in a tetrad formulation of a Kaluza–Klein model with gauge group $\mathrm{SL}(2,R)\times\mathrm{U}(1)$. In this model, the self-adjoint modes of the tetrad describe gravity, whereas the isometric modes of the tetrad together with a scalar field describe fermions. This model gives precisely the usual Dirac–Einstein Lagrangian. In this paper we generalize the tensor Dirac theory to the larger gauge group $\mathrm{SL}(2,C)\times\mathrm{U}(1)$ acting on bispinors. We show that each $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ subgroup of $\mathrm{SL}(2,C)\times\mathrm{U}(1)$ corresponds to a different factorization of the second-order Klein–Gordon equation into a first-order Dirac equation. Since the Noether currents are different for each factorization, the solutions describe different flavors of fermions. We show that electric charge, lepton number, and baryon number are conserved in this generalization of the Dirac theory. © *1999 American Institute of Physics.* [S0022-2488(99)02904-7]

I. INTRODUCTION

Recently, the Dirac and Einstein equations were unified in a tetrad formulation of a Kaluza–Klein model which gives precisely the usual Dirac–Einstein Lagrangian.[1,2] In this model, the self-adjoint modes of the tetrad describe gravity, whereas the isometric modes of the tetrad together with a scalar field describe fermions. An analogy can be made between the tetrad modes and the elastic and rigid modes of a deformable body.[1] For a deformable body, the elastic modes are self-adjoint and the rigid modes are isometric with respect to the Euclidean metric on R^3. This analogy extends into the quantum realm since rigid modes satisfying Euler's equation can be Fermi quantized.[3]

The use of tetrads to describe gravity has a long history,[4] which includes coupling with the Dirac field as a source.[5] However, introducing a tetrad to describe both fermion and gravitational fields solves an important problem posed by current theories of fermion–graviton interaction. To define bispinors, reference tetrad fields or their equivalent must be defined on the space–time manifold.[6] These reference fields have been treated as purely boson fields in supersymmetric theories.[7] We found that the reference tetrads themselves can be taken as fundamental fields that describe both fermions and gravity.[1]

Boson gauge fields were added to the unified Dirac–Einstein Lagrangian by defining the gauge group of the Kaluza–Klein model to be a semidirect product.[2] It was shown that the semidirect product structure of the gauge group produces precisely the usual ''minimal coupling'' between bosons and fermions.

The tetrad Kaluza–Klein model is based on a constrained Yang–Mills formulation of the Dirac Theory.[1–3,8] In this formulation a bispinor field Ψ is mapped to a set of $\mathrm{SL}(2,C)\times\mathrm{U}(1)$ gauge potentials A_α^K and a complex scalar field ρ. The gauge potentials A_α^K in the image of the map vanish, except for an $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ subgroup of $\mathrm{SL}(2,C)\times\mathrm{U}(1)$. The restriction to $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ arises because the Dirac equation has only $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ gauge symmetry, while the larger $\mathrm{SL}(2,C)\times\mathrm{U}(1)$ gauge group acts on bispinors.

In this paper we generalize the tensor Dirac theory to the larger $\mathrm{SL}(2,C)\times\mathrm{U}(1)$ gauge group acting on bispinors, and discuss its possible application to describing fermion flavors (e.g., quark and lepton flavors). We show that each $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ subgroup of $\mathrm{SL}(2,C)\times\mathrm{U}(1)$ corresponds to a different factorization of the second-order Klein–Gordon equation into a first-order Dirac equation. Thus, the Yang–Mills Lagrangian (see Sec. II) for the $SL(2,C)\times U(1)$ gauge potentials A_α^K and scalar field ρ, describes solutions for different flavors of fermions. An assignment of flavor parameters to first generation leptons and quarks shows that electric charge, lepton number, and baryon number are conserved.

In Sec. II we review the derivation, which demonstrates that the Dirac bispinor Lagrangian equals a constrained Yang–Mills Lagrangian in the limit of an infinitely large coupling constant. In Sec. III we then derive the generalization that gives rise to fermion flavors. We show how the flavor symmetric tensor theory includes both the Dirac and Majorana bispinor theories as special cases. In Appendices A and B we provide a new derivation of the tensor Dirac theory from first principles, starting from trace formulas for the Pauli matrices, and ending with formulas for differentiating bispinor Fierz identities. This gauge symmetric derivation greatly simplifies computations with bispinors used in previous work on Fierz identities.[9,10]

Other authors have investigated the Fierz identities using more powerful Clifford algebra techniques.[11,12] Clifford algebra provides significant generalizations of the simple matrix algebra presented in Appendices A and B. It is clear from this work that the $SL(2,C)\times U(1)$ theory presented here can be extended to a wider class of gauge groups acting on bispinors by different representations.[13–16] A full generalization of the tensor Dirac theory with Clifford algebra techniques merits further study, as discussed in the following brief review of developments in the study of flavor symmetry.

Numerous books and journal articles have noted that different factorizations of the Klein–Gordon equation can give rise to flavor symmetry. For example, our earlier work represented a bispinor field Ψ as a triplet of complex antisymmetric tensors $\mathbf{F}_{\alpha\beta}$ of Carmeli class (G), defined in terms of A_α^K and ρ, as follows:[8]

$$\mathbf{F}_{\alpha\beta}=\rho(A_\alpha^0\mathbf{A}_\beta-\mathbf{A}_\alpha A_\beta^0+i\mathbf{A}_\alpha\times\mathbf{A}_\beta), \tag{1.1}$$

where $A_\alpha^K=(A_\alpha^0,\mathbf{A}_\alpha)$ and $\mathbf{A}_\alpha=(A_\alpha^1,A_\alpha^2,A_\alpha^3)$. Note that in this paper we denote gauge triplets, such as $\mathbf{F}_{\alpha\beta}=(F_{\alpha\beta}^1,F_{\alpha\beta}^2,F_{\alpha\beta}^3)$ and $\mathbf{A}_\alpha=(A_\alpha^1,A_\alpha^2,A_\alpha^3)$ by bolding and use $\cdot$ and $\times$ to denote the dot and cross products with respect to the gauge indices 1,2,3. Class (G) triplets $\mathbf{F}_{\alpha\beta}$ were studied by Carmeli, who classified the algebraic properties of Yang–Mills curvature tensors.[17] The $SL(2,C)\times U(1)$ gauge symmetry of the triplets $\mathbf{F}_{\alpha\beta}$ revealed the flavor symmetry of bispinors.[9,18,19] The following composite map from bispinors to Carmeli class (G) triplets $\mathbf{F}_{\alpha\beta}$:

$$\Psi\rightarrow(A_\alpha^K,\rho)\rightarrow\mathbf{F}_{\alpha\beta}, \tag{1.2}$$

which commutes with both the space–time and flavor symmetries, is an extension of the Cartan map from spinors to complex isotropic vectors.[8,20–22] Note that, similar to the Cartan map, the extension (1.2) is a holomorphic, branched double covering map with a single branch point at $\mathbf{F}_{\alpha\beta}=0$. For physically realizable solutions of the Dirac equation, the covering map (1.2) is unobservable.[8,23]

Using Clifford algebra techniques, Keller and Rodriguez-Romo generalized all of the constructions of the last paragraph, including flavor symmetry, to a wider class of gauge groups used in elementary particle physics.[11,12,24,25] We believe that the Kaluza–Klein model for the Dirac theory[1,2] also extends to this wider class of gauge groups. As a step in this direction, a specific generalization of the tensor Dirac theory to bispinor multiplets $\Psi=(\Psi_1,\ldots,\Psi_n)$ with the gauge group $SL(2n,C)\times U(1)$ will be presented in a forthcoming paper.

The tensor form of Dirac's bispinor equation itself has a long history.[1] Most noteworthy was the derivation of the tensor form of Dirac's bispinor Lagrangian by Takahashi, who ascribed a similar derivation to Zhelnorovich.[10,26] However, at that time, the flavor symmetry of Takahashi's formula [see formula (2.8) in Sec. II] was not recognized. Clifford algebra techniques could generalize Takahashi's formula (2.8) to the wider class of gauge groups investigated by Keller and

Rodriguez-Romo.[13] Moreover, studies of such generalizations could expand the number of Kaluza–Klein models suitable for the unification of the fundamental fields of elementary particle physics.

We conclude this Introduction with several remarks on the development of quantum mechanics. As discussed earlier, all the fundamental particles can be derived through the ''second'' quantization of fields unified at the classical level in a tetrad Kaluza–Klein model.[2] In this model the classical modes of a tetrad can be separated into isometric and self-adjoint modes, except near a Schwarzschild horizon.[1,27] Fermion statistics can be assumed for the isometric modes, whose classical Hamiltonian structure consists of noncanonical, unitary Lie–Poisson brackets, while only boson statistics can be assumed for the self-adjoint modes with canonical Hamiltonian structure.[1,3,28]

As previously mentioned, isometric modes of the reference frame fields, which in the tetrad Kaluza–Klein model represent fermion degrees of freedom, are treated as boson degrees of freedom in supersymmetric theories.[7] An alternative supersymmetric theory would treat only the self-adjoint modes of the reference frame fields as bosons, and these would describe gravity. The isometric modes would describe fermions. Thus, the tensor Dirac theory leads to a new formulation of the fermion–graviton interaction, which requires the modification of supersymmetric theories.

Consider first quantization. At the inception of quantum mechanics, with the realization of ''wave–particle duality,'' particles, classically conceived as points of mass and charge, were given wavelike properties through first quantization. Thus, Schrodinger's wave function Ψ for a single particle describes a matter field similar to the classical fields describing electromagnetism and gravity. Through ''second'' quantization, particles associated with the matter, electromagnetic, and gravitational fields are then represented uniformly by occupation states acted upon by creation and annihilation operators. From a practical viewpoint the occupation states and their precedent ''classical'' fields are sufficient to describe all observed phenomena.

The tetrad Kaluza–Klein model supports this viewpoint, since it shows that the ''classical'' precedents for fermions and bosons are the same *kind* of field, i.e., a classical Yang–Mills field with a classical Hamiltonian structure. Specifically, the Schrodinger wave function Ψ for a single particle (which Schrodinger originally conceived as a classical field), is the nonrelativistic limit of the Dirac field (A_α^K,ρ), which satisfies a classical Yang–Mills equation. Thus, Ψ inherits from the Dirac field (A_α^K,ρ) a classical Hamiltonian structure that admits Fermi quantization.[3] Therefore, quantum mechanics need only consist of three parts: the classical field equations, field quantization, and rules for applying the formalism to experiments. A fundamental model must account for all three parts of the theory in a unified framework. Unification of fermion and boson fields at a classical level is a step in this direction.

Supersymmetric fields have been investigated, in part, because they describe fermion fields with an anticanonical Hamiltonian structure, and hence unify fermions with bosons at a superclassical level.[7] However if supersymmetry is fundamental, a motivation other than unification at the prequantum level, is required, since the tensor Dirac theory, with its noncanonical Hamiltonian structure, achieves this unification at a classical level without resorting to supernumbers.

II. TENSOR FORM OF THE DIRAC LAGRANGIAN

In this section we review the derivation that demonstrates that the Dirac bispinor Lagrangian (2.4) equals the constrained Yang–Mills Lagrangian (2.14) in the limit of an infinitely large coupling constant. The derivation exploits the $SL(2,R)\times U(1)$ gauge symmetry of Dirac's bispinor Lagrangian.

Consider the $SL(2,R)\times U(1)$ gauge transformations, acting on the bispinor field Ψ, with infinitesimal generators τ_K for $K=0,1,2,3$, defined by

$$\begin{aligned} \tau_0\Psi &= -i\Psi, \quad \tau_1\Psi = i\Psi^C, \\ \tau_2\Psi &= \Psi^C, \quad \tau_3\Psi = i\gamma^5\Psi, \end{aligned} \tag{2.1}$$

where (using bispinor notation) Ψ^C denotes the charge conjugate of Ψ and γ^5 is the fifth Dirac matrix.[29] Note that the action of $SL(2,R)\times U(1)$ on Ψ is real linear, whereas usually only complex linear gauge transformations of bispinors are considered. The infinitesimal gauge generators τ_0, τ_1, and τ_2 generate SL(2,R) and τ_3 generates U(1).

The $SL(2,R)\times U(1)$ gauge transformations generated by τ_K commute with Lorentz transformations.[29] From formula (2.1) the commutation relations of the gauge generators τ_K are given by

$$[\tau_0,\tau_1]=2\tau_2, \quad [\tau_0,\tau_2]=-2\tau_1, \quad [\tau_1,\tau_2]=-2\tau_0, \tag{2.2}$$

and τ_3 commutes with all the τ_K. Formula (2.2) can be written more compactly as

$$[\tau_J,\tau_K]=2f_{JK}^{\mathrm{L}}\tau_{\mathrm{L}}, \tag{2.3}$$

which defines the Lie algebra structure constants f_{JK}^{L} for the gauge group $SL(2,R)\times U(1)$.

By formula (2.2), the Minkowski metric g_{JK} (with diagonal elements $\{1,-1,-1,-1\}$ and zeros off the diagonal) is an invariant metric[30] for the gauge group $SL(2,R)\times U(1)$. Gauge indices J, K, L will be lowered and raised using the Minkowski metric g_{JK} and its inverse g^{JK}. As in formula (2.3), repeated indices are to be summed from 0 to 3.

Dirac's bispinor Lagrangian L is given by

$$L=\mathrm{Re}[i\bar{\Psi}\gamma^\alpha\partial_\alpha\Psi-m_0 s], \tag{2.4}$$

where s is the complex scalar field defined by

$$\mathrm{Re}[s]=\bar{\Psi}\Psi, \qquad \mathrm{Im}[s]=i\bar{\Psi}\gamma^5\Psi, \tag{2.5}$$

where (using bispinor notation) $\bar{\Psi}=\Psi^+\gamma^0$, where Ψ^+ denotes the transpose conjugate of Ψ, and γ^α for α=0,1,2,3 are Dirac matrices.[29] Moreover, in formula (2.4), m_0 denotes the fermion mass, and ∂_α denote partial derivatives with respect to space–time coordinates. Tensor indices α, β, γ are lowered and raised using the Minkowski space–time metric, which we denote as $g_{\alpha\beta}$, and its inverse $g^{\alpha\beta}$.

Apart from the mass term, Dirac's bispinor Lagrangian is invariant under the $SL(2,R)\times U(1)$ gauge transformations (2.1). From formula (2.5), the scalar s is invariant under SL(2,R) gauge transformations, and transforms as a complex scalar under the U(1) gauge transformations generated by τ_3. To make the Lagrangian (2.4) invariant for all $SL(2,R)\times U(1)$ gauge transformations, it suffices that m_0 transform like $\bar{s}$ (the complex conjugate of s). Since m_0 appears in the Lagrangian (2.4) without derivatives, the assumption that m_0 transform like $\bar{s}$ under U(1) gauge transformations, has no effect on the Dirac equation.

From the Dirac Lagrangian (2.4) we can derive the following $SL(2,R)\times U(1)$ Noether currents:

$$j_\alpha^K=\mathrm{Re}[i\bar{\Psi}\gamma_\alpha\tau^K\Psi]. \tag{2.6}$$

The Noether currents j_α^K and scalar s satisfy an orthogonal constraint known as a Fierz identity[10] (see Appendix A for an elementary derivation):

$$j_\alpha^K j_{K\beta}=|s|^2 g_{\alpha\beta}. \tag{2.7}$$

Takahashi[10] derived the following formula for the kinetic part of the Dirac Lagrangian (2.4):

$$\mathrm{Re}[i\overline{\Psi}\gamma^\alpha\,\partial_\alpha\Psi]=-\frac{1}{4|s|^2}\mathrm{Re}[(\partial_\alpha\mathbf{J}_\beta)\cdot\mathbf{J}^\alpha\times\mathbf{J}^\beta-2i\overline{s}J^0_\alpha\,\partial^\alpha s],\tag{2.8}$$

which uses the following notation (which differs from Takahashi's):

$$J^K_\alpha=(J^0_\alpha,\mathbf{J}_\alpha)=(-j^3_\alpha,-ij^2_\alpha,ij^1_\alpha,-j^0_\alpha).\tag{2.9}$$

Thus, from Takahashi's formula (2.8), we see that the Dirac Lagrangian (2.4) can be expressed entirely in terms of the Noether currents j^K_α and the complex scalar field s, satisfying the orthogonal constraint (2.7).

Takahashi[10] derived formulas (2.7) and (2.8), along with 15 general and 75 specific Fierz identities, without exploiting the $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ gauge symmetry. The derivation of formula (2.8) from first principles given in Appendices A and B of this paper exploits this symmetry. Once the $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ gauge symmetry of formula (2.8) is recognized, the demonstration that Dirac's bispinor Lagrangian (2.4) equals a constrained Yang–Mills Lagrangian in the limit of an infinitely large coupling constant, is fairly obvious.

Indeed, we can map a subset of $\mathrm{SL}(2,C)\times\mathrm{U}(1)$ gauge potentials A^K_α and a complex scalar field ρ into (J^K_α,s) by setting

$$\begin{aligned}J^K_\alpha&=4|\rho|^2A^K_\alpha,\\ s&=4|\rho|^2\overline{\rho}.\end{aligned}\tag{2.10}$$

Since we regard the Lie algebra of $\mathrm{SL}(2,C)$ as the complexification of the Lie algebra of SU(2), the $\mathrm{SL}(2,C)$ gauge potentials $\mathbf{A}_\alpha=(A^1_\alpha,A^2_\alpha,A^3_\alpha,)$ are complex, while the U(1) gauge potential A^0_α is real. By formula (2.9) the gauge potentials A^K_α are restricted to the subset for which

$$\mathrm{Re}[A^1_\alpha]=\mathrm{Re}[A^2_\alpha]=\mathrm{Im}[A^3_\alpha]=0.\tag{2.11}$$

This subset corresponds precisely to a $\mathrm{SL}(2,R)\times\mathrm{U}(1)$ subgroup of the gauge group $\mathrm{SL}(2,C)\times\mathrm{U}(1)$. On substituting formula (2.10) into Takahashi's formula (2.8), Dirac's Lagrangian (2.4) becomes

$$L=-\mathrm{Re}[(\partial_\alpha\mathbf{A}_\beta)\cdot\mathbf{A}^\alpha\times\mathbf{A}^\beta+2i\overline{\rho}A^0_\alpha\,\partial^\alpha\rho+4m_0|\rho|^2\overline{\rho}],\tag{2.12}$$

and the orthogonal constraint (2.7) becomes

$$A^K_\alpha A_{K\beta}=-|\rho|^2g_{\alpha\beta}.\tag{2.13}$$

Consider the following Yang–Mills Lagrangian L_g for the gauge potentials A^K_α and the complex scalar field ρ:

$$L_g=-\tfrac{1}{4}\mathrm{Re}[A^K_{\alpha\beta}A^{\alpha\beta}_K]+\overline{D_\alpha(\rho+m)}D^\alpha(\rho+m)-\tfrac{1}{2}g^2|\rho|^4,\tag{2.14}$$

where $A^K_{\alpha\beta}=(A^0_{\alpha\beta},\mathbf{A}_{\alpha\beta})$ and

$$\begin{aligned}A^0_{\alpha\beta}&=\partial_\alpha A^0_\beta-\partial_\beta A^0_\alpha,\\ \mathbf{A}_{\alpha\beta}&=\partial_\alpha\mathbf{A}_\beta-\partial_\beta\mathbf{A}_\alpha-g\mathbf{A}_\alpha\times\mathbf{A}_\beta,\\ D_\alpha(\rho+m)&=\partial_\alpha\rho+igA^0_\alpha(\rho+m),\end{aligned}\tag{2.15}$$

where g denotes the Yang–Mills coupling constant and $m_0=\frac{1}{2}mg$ is the fermion mass. From formulas (2.12) and (2.13), Dirac's bispinor Lagrangian (2.4) equals

$$L=\lim_{g\to\infty} g^{-1}L_g . \tag{2.16}$$

Note that the Euler–Lagrange equation for the Lagrangian (2.14) with the constraint (2.13) expressed using Lagrange multipliers, commutes with the restriction (2.11). Hence, the $\mathbf{A}_\alpha$ can be used to denote either $\mathrm{SL}(2,C)$ or the subset of $\mathrm{SL}(2,R)$ gauge potentials. By regarding $\mathrm{SL}(2,R)$ as embedded in the complex analytic group $\mathrm{SL}(2,C)$, we are able to use familiar vector operations to express the Lie algebra structure constants in formulas (2.12) and (2.15). The vector operations greatly simplify derivations.

Also, note that the part of the Lagrangian (2.14) for the scalar field ρ is not uniquely determined by Eq. (2.16). However, the Lagrangian (2.14) has unique additional properties discussed in Refs. 1 and 31. Specifically, the Yang–Mills equation derived from the Lagrangian (2.14) with the constraints (2.11) and (2.13) has exact plane wave solutions in one-to-one correspondence with the plane wave solutions of Dirac's bispinor equation. Because the Yang–Mills equation is nonlinear, the mass of each plane wave could depend on its amplitude, which would cause velocity splitting of wave packets.[32] However, the mass of each plane wave equals m_0, and hence is constant. Wave packets are identical to the wave packets derived from Dirac's equation, and do not exhibit velocity splitting.[31] Thus, the constrained Yang–Mills equation has solutions similar to Dirac's bispinor equation, which is a limiting case of it by formula (2.16).

However, if the coefficient g^2 in the quartic potential $V(\rho)=\frac{1}{2}g^2|\rho|^4$ of the Lagrangian (2.14) is varied, then the limit in formula (2.16) fails to exist, and the mass of each plane wave becomes dependent on amplitude. Unless the amplitude vanishes, the mass becomes infinite in the limit of an infinitely large coupling constant g. Wave packets lose their elementary character with severe velocity splitting and unobservably high mass. Thus, in the Kaluza–Klein tetrad model,[1] assuming a fixed mass parameter m, the quartic potential $V(\rho)$ determines both the coupling constant g and the fermion mass $m_0=\frac{1}{2}mg$. This property will be exploited as an alternative to Higgs fields[33] for obtaining fermion masses in a future paper, which generalizes the tensor theory to multiplets of n bispinors $\Psi=(\Psi_1,\ldots,\Psi_n)$, where each bispinor component of Ψ is associated with a different fermion mass.

Another unique property of the Lagrangian (2.14) may have significance for astrophysics. For Yang–Mills Lagrangians in general, the energy–momentum tensor $T^{\alpha\beta}$ for wave packets exhibits both mass density and pressure. However, for plane waves and wave packets derived from the Lagrangian (2.14), we can show that

$$T^{\alpha\beta}=m_0|j|v^\alpha v^\beta, \tag{2.17}$$

where m_0 is the fermion mass, $j^\alpha=|j|v^\alpha$ is the electric current, which is also called the particle current, and $v^\alpha=j^\alpha/|j|$ with $v^\alpha v_\alpha=1$ is the velocity field.[31] As for Dirac's bispinor Lagrangian, formula (2.17) exhibits an energy–momentum tensor $T^{\alpha\beta}$ for a fluid with velocity field v^α, mass density $m_0|j|$, and zero pressure. Since for a Yang–Mills Lagrangian the pressure generally grows as the square of the coupling constant g, which must be large to obtain the Dirac Lagrangian, a nonzero pressure would be highly observable, for example, in the internal dynamics of stars.

For the Lagrangian (2.14), neither the exact plane wave solutions nor their wave packets provide any information about the magnitude of the coupling constant g. Consider, however, the single-particle (unquantized) Dirac equation applied to the scattering of a single fermion particle in an external electromagnetic potential.[29] The computation of scattering cross sections with these simplifying assumptions may provide a way to estimate the magnitude of the coupling constant g.

III. FLAVOR SYMMETRY

We will show in future papers that theories using multiplets of n bispinors $\Psi=(\Psi_1,\ldots,\Psi_n)$ can be formulated as tensor theories. Such formulations are based on extensions of the Fierz identities given in Appendix A to bispinor multiplets Ψ, with the gauge group $\mathrm{SL}(2n,C)\times \mathrm{U}(1)$. We will show that any given bispinor multiplet maps to fermion gauge potentials A_α^K that vanish, except for an $SL(2,R)\times U(1)$ subgroup of $SL(2n,C)\times U(1)$. The specific $SL(2,R)\times U(1)$ subgroup for which the A_α^K are nonvanishing, determines the fermion flavor. That is, only the components of A_α^K associated with the $SL(2,R)\times U(1)$ subgroup are nonzero, and hence A_α^K reduces to the Dirac theory for a single flavor of fermion. We will also show, in particular, that the standard electroweak model can be embedded in a $SL(2n,C)\times U(1)$ tensor theory in a manner consistent with the gauge symmetries used to derive the tensor Dirac equation.

Even for a bispinor singlet $(n=1)$, which we treat in this paper, current bispinor theories do not utilize the full $SL(2,C)\times U(1)$ gauge symmetry of the tensor Dirac Lagrangian (2.14). We will show in this section that the $SL(2,C)\times U(1)$ symmetric tensor Dirac theory includes both the Dirac and Majorana bispinor theories as special cases. Furthermore, we will show that the additional symmetry can be used to model the observed conservation of electric charge, lepton number, and baryon number.

A bispinor Ψ consists of a spinor and a dual conjugate spinor [see formula (A30) in Appendix A]. Equivalently, Ψ can be represented by a spinor doublet Φ consisting of two spinors, as in formula (A18). Acting on spinor doublets Φ is the gauge group $SL(2,C)\times U(1)$ whose generators are defined in formula (A19). The bijective map $\Phi\to\Psi$ defined by formula (A30), induces an equivalent representation of the gauge group $SL(2,C)\times U(1)$ acting on bispinors Ψ.

Unitary gauge transformations form a subgroup $U(2)=SU(2)\times U(1)$ of the $SL(2,C)\times U(1)$ gauge transformations acting on spinor doublets Φ. Using the representation induced on bispinors Ψ by the bijective map $\Phi\to\Psi$, the generators T_K of the unitary group $SU(2)\times U(1)$ become

$$\begin{aligned} T_0&=\tau_3, \quad T_1=-\tau_3\tau_2, \\ T_2&=\tau_3\tau_1, \quad T_3=\tau_0, \end{aligned} \tag{3.1}$$

where the τ_K are the $SL(2,R)\times U(1)$ gauge generators acting on bispinors Ψ defined in formula (2.1). Note from formulas (A19) and (A30) in Appendix A, that the full set of generators for the gauge group $SL(2,C)\times U(1)$ acting on bispinors Ψ consists of the T_K $(K=0,1,2,3)$, which generate $SU(2)\times U(1)$ and three additional gauge generators:

$$T_1'=\tau_2, \quad T_2'=-\tau_1, \quad T_3'=\tau_3\tau_0. \tag{3.2}$$

From formula (2.1), the gauge generators T_K satisfy the following commutation relations:

$$[T_1,T_2]=2T_3, \quad [T_2,T_3]=2T_1, \quad [T_3,T_1]=2T_2. \tag{3.3}$$

T_0 commutes with all the T_K. Thus, T_1, T_2, and T_3 generate SU(2), and T_0, which is τ_3, generates U(1) as before.

Dirac's bispinor Lagrangian (2.4) can now be expressed by

$$L=\mathrm{Re}[i\bar{\Psi}\Gamma^\alpha\partial_\alpha\Psi-m_0\bar{\Psi}\Psi], \tag{3.4}$$

where $\Gamma^\alpha=i\gamma^\alpha T_3=\gamma^\alpha$. Hence, Dirac's equation can be expressed as

$$i\Gamma^\alpha\partial_\alpha\Psi=m_0\Psi. \tag{3.5}$$

We may replace T_3 with an equivalent gauge generator $T=c^K T_K$, where $c^K\in R^4$ satisfies $c^K c_K=-1$. That is, defining $\Gamma^\alpha=i\gamma^\alpha T$ generalizes the Dirac matrices γ^α. From formulas (2.1) and (3.1) we see that $T^K\gamma^\alpha=-\gamma^\alpha T_K$, and from a standard identity for Dirac matrices γ^α we get[29]

$$\Gamma_\alpha\Gamma_\beta+\Gamma_\beta\Gamma_\alpha=2g_{\alpha\beta}. \tag{3.6}$$

Formulas (3.5) and (3.6) imply that the bispinor field Ψ satisfies the Klein–Gordon equation:

$$\partial_\alpha\partial^\alpha\Psi=-m_0^2\Psi. \tag{3.7}$$

Thus, each generator $T = c^K T_K$, with $c^K c_K = -1$, gives a different factorization of the Klein–Gordon equation into a Dirac equation (3.5), with Lagrangian (3.4). The following theorem generalizes Takahashi's formula (2.8).

Theorem 1: Let Ψ be a bispinor field. Then

$$\mathrm{Re}[i\bar{\Psi}\Gamma^\alpha \partial_\alpha \Psi] = -\frac{1}{4|s|^2}\mathrm{Re}[(\partial_\alpha \mathbf{J}_\beta)\cdot \mathbf{J}^\alpha \times \mathbf{J}^\beta - 2i\bar{s}J^0_\alpha \partial^\alpha s], \tag{3.8}$$

where [see formula (2.9)]

$$J^0_\alpha = c_K j^K_\alpha, \\ \mathbf{J}_\alpha = c^0 \mathbf{j}_\alpha - \mathbf{c} j^0_\alpha + i\mathbf{c}\times \mathbf{j}_\alpha, \tag{3.9}$$

and $c^K = (c^0, \mathbf{c})$ with $\mathbf{c} = (c^1, c^2, c^3)$, and $j^K_\alpha = (j^0_\alpha, \mathbf{j}_\alpha)$ with $\mathbf{j}_\alpha = (j^1_\alpha, j^2_\alpha, j^3_\alpha)$.

Proof: A proof of Theorem 1 from first principles using simple trace formulas is provided in Appendices A and B. See formula (B15). Alternatively, Theorem 1 follows from formulas (2.8) and (2.9) by a symmetry argument. Note that the J^K_α defined by formula (3.9) are the SL$(2,C)\times$U(1) Noether currents of the bispinor Lagrangian (3.4). Formulas (2.8) and (2.9) are a special case of formulas (3.8) and (3.9) with $c^K = (0,0,0,1)$. As in formula (2.9), the nonvanishing J^K_α are associated with an SL$(2,R)\times$U(1) subgroup of SL$(2,C)\times$U(1). Since the gauge group SL$(2,C)\times$U(1) acts transitively on the Noether currents $\mathbf{J}_\alpha$ and leaves J^0_α invariant, the SL$(2,C)\times$U(1) Noether currents in formula (3.9) can be obtained from the SL$(2,R)\times$U(1) Noether currents (2.9) by a gauge transformation. The proof is completed by showing the gauge invariance of formula (3.8).

Note in Theorem 1 that c^K, j^K_α, and J^K_α are all invariant under U(1) gauge transformations. The SL$(2,C)$ gauge transformations act as SO(1,3) transformations on c^K and j^K_α, and as complex orthogonal transformations on $\mathbf{J}_\alpha$. As previously stated, J^0_α is invariant for all gauge transformations. Since the complex scalar s is invariant for all but U(1) gauge transformations, the right-hand side of formula (3.8) is invariant for all gauge transformations. The left-hand side of formula (3.8) is also invariant for all gauge transformations, as can be seen in formula (B4) in Appendix B. Q.E.D.

As a consequence of Theorem 1, we prove the following theorem.

Theorem 2: Set $J^K_\alpha = 4|\rho|^2 A^K_\alpha$ and $s = 4|\rho|^2\bar{\rho}$. Then (A^K_α, ρ) satisfy the orthogonal constraint (2.13), and the bispinor Lagrangian (3.4) equals

$$L = \lim_{g\to\infty} g^{-1} L_g, \tag{3.10}$$

where L_g is the Yang–Mills Lagrangian (2.14).

Proof: To establish the orthogonal constraint (2.13), we define an associative binary operation, denoted as $\otimes$ on C^4 as follows. For a^K, b^K, $c^K \in C^4$, we define $c = a\otimes b$ if and only if

$$c^0 = a^0 b^0 + \mathbf{a}\cdot\mathbf{b}, \\ \mathbf{c} = a^0\mathbf{b} + \mathbf{a} b^0 + i\mathbf{a}\times\mathbf{b}, \tag{3.11}$$

where $a^K = (a^0, \mathbf{a})$ with $\mathbf{a} = (a^1, a^2, a^3)$, and similarly for b^K and c^K. It is straightforward to show that the binary operation $\otimes$ is associative. Denote $\tilde{a}^K = a_K = (a^0, -\mathbf{a})$. Then formula (3.9) can be written as

$$J^K_\alpha = [j_\alpha \otimes \tilde{c}]^K, \quad \tilde{J}^K_\alpha = [c\otimes \tilde{j}_\alpha]^K, \tag{3.12}$$

where $[a\otimes b]^K$ denotes the Kth component of $a\otimes b$. Since from $c^K c_K = -1$ we have $\tilde{c}\otimes c = (-1,0,0,0)$, formulas (2.7), (3.11), and (3.12) give

$$J^K_\alpha J_{K\beta} = [J_\alpha \otimes \tilde{J}_\beta]^0 = [j_\alpha \otimes \tilde{c} \otimes c \otimes \tilde{j}_\beta]^0 = -[j_\alpha \otimes \tilde{j}_\beta]^0 = -j^K_\alpha j_{K\beta} = -|s|^2 g_{\alpha\beta}. \tag{3.13}$$

The orthogonal constraint (2.13) now follows from formulas (2.10) and (3.13). On substituting formulas (2.5) and (3.8) into the bispinor Lagrangian (3.4), formula (3.10) is a straightforward evaluation of formulas (2.12), (2.14), and (2.15). Q.E.D.

Each factorization of the Klein–Gordon equation (3.7) gives a different Dirac equation (3.5), parametrized by the choice of $c^K \in R^4$, such that $c^K c_K = -1$. The usual Dirac equation has the parameter $c^K = (0,0,0,1)$. Thus, the Yang–Mills Lagrangian (2.14) for the $\mathrm{SL}(2,C) \times \mathrm{U}(1)$ gauge potentials A^K_α and scalar field ρ describes solutions from different Dirac equations corresponding to different factorizations of the Klein–Gordon equation. We can show using formulas (2.10), (2.13), and (3.12) that A^K_α and ρ uniquely determine both the parameter c^K and the bispinor field Ψ (except for the unobservable sign of Ψ). Since the Noether currents (3.9) are different for each value of the parameter c^K, the parameter c^K describes the fermion flavor.

In the tensor theory, fermion gauge potentials A^K_α, associated with a gauge group G, can be unified with boson gauge potentials V^K_α, associated with a gauge group H that acts on G via a homomorphism $\varphi: H \rightarrow \mathrm{Aut}(G)$, by regarding (A^K_α, V^K_α) as gauge potentials associated with the semidirect product group $G_\varphi \times H$. As previously shown, the semidirect product structure of the gauge group $G_\varphi \times H$ uniquely prescribes the usual ''minimal coupling'' between fermions and bosons.[2] For flavor gauge groups let

$$\begin{aligned} G &= \mathrm{SL}(2,C) \times \mathrm{U}(1), \\ H &= \mathrm{SU}(2) \times \mathrm{U}(1), \end{aligned} \tag{3.14}$$

with φ defined as the adjoint action of H on G.

Note that the boson gauge group $\mathrm{SU}(2)\times\mathrm{U}(1)$ is a compact subgroup of the noncompact fermion gauge group $\mathrm{SL}(2,C) \times \mathrm{U}(1)$. The U(1) gauge transformations leave the flavor parameter $c^K \in R^4$ invariant; whereas the SU(2) gauge transformations leave c^0 invariant and rotate $\mathbf{c} = (c^1, c^2, c^3)$. Since for the electron we choose $c^0 = 0$ and $\mathbf{c} = (0,0,1)$, we will assume that $c^0 = 0$ for all fermion flavors. That is, we restrict the flavor parameter c^K to a single $\mathrm{SU}(2)\times\mathrm{U}(1)$ orbit for which $c^0 = 0$.

From formula (3.9) with $c^0 = 0$, the $\mathrm{SU}(2)\times\mathrm{U}(1)$ Noether currents $\mathrm{Re}[J^K_\alpha]$ can be expressed with the vector and axial currents, j^0_α and $j^5_\alpha = \mathbf{c} \cdot \mathbf{j}_\alpha$, as follows:

$$\mathrm{Re}[J^K_\alpha] = (-j^5_\alpha, -\mathbf{c} j^0_\alpha). \tag{3.15}$$

Both j^0_α and j^5_α are invariant under all $\mathrm{SU}(2)\times\mathrm{U}(1)$ gauge transformations. Thus, from formulas (2.1) and (2.6), j^0_α and j^5_α are gauge-invariant vector and axial currents, respectively, which can be identified with electroweak vector and axial currents.[33] It has been observed that the electromagnetic current is purely vector and that weak currents have both vector and axial components.

Table I shows a possible assignment of flavors c^K to first generation leptons and quarks. Note that in Table I antifermions are obtained by replacing c^K with $-c^K$. With these assignments of

TABLE I. Flavor parameters for leptons and quarks.

FERMION	$c^K = (c^0,\ c^1,\ c^2,\ c^3)$
Electron (e)	(0,0,0,1)
Neutrino (ν)	(0,1,0,0)
Up-quark (u)	$(0, \frac{1}{3}, -\frac{2}{3}, -\frac{2}{3})$
Down-quark (d)	$(0, -\frac{2}{3}, -\frac{2}{3}, \frac{1}{3})$

flavors, the following theorem identifies the Standard Model electromagnetic and neutral currents, denoted as $J_\alpha^{(Q)}$ and $J_\alpha^{(Z)}$, respectively, by expressing them as linear combinations of the SU(2)×U(1) Noether currents (3.15).

Theorem 3: With the assignment of flavors c^K to first generation leptons and quarks shown in Table I, the Standard Model electromagnetic and neutral currents are uniquely determined as follows:

$$J_\alpha^{(Q)} = \mathrm{Re}[J_\alpha^3],$$
$$J_\alpha^{(Z)} = -\frac{1}{2\sin\varphi}\mathrm{Re}[\pm J_\alpha^0 + J_\alpha^1 + (1-2\cos\varphi)J_\alpha^3], \tag{3.16}$$

where φ is twice the Weinberg angle.[33] The plus sign occurs with the U(1) current J_α^0 for an electron and down-quark and a minus sign for the neutrino and up-quark.

Proof: This is a straightforward verification by substituting values of c^K from Table I into formulas (3.15) and (3.16), and showing that the currents agree with the Standard Model.[33] Q.E.D.

One can deduce from formulas (2.1) and (3.1) that flavor symmetry generalizes the Majorana representation used in quantum field theory to massive fermions.[33] The neutrino in Table I for the case of zero mass is precisely a Majorana neutrino.

By a standard argument, conservation of electric current implies the additivity of electric charge. By a similar argument, conservation of the SU(2) vector currents, $-\mathbf{c}j_\alpha^0$, implies the additivity of the flavor parameter $c^K=(0,\mathbf{c})$. To illustrate this additivity consider the beta decay $d \rightarrow u + e + \overline{\nu}$ in which a down-quark d decays into an up-quark u, an electron e, and an antineutrino $\overline{\nu}$. Since $c^K(\overline{\nu}) = -c^K(\nu)$, we have from Table I,

$$c^K(d) = c^K(u) + c^K(e) + c^K(\overline{\nu}). \tag{3.17}$$

In all known interactions between e, ν, u, and d fermions, the c^K as defined in Table I are additive, and hence can be regarded as conserved electroweak charges associated with the SU(2) vector currents $-\mathbf{c}j_\alpha^0$.

The additivity of the c^K leads directly to the three familiar laws stated in Theorem 4 that follows.

Theorem 4: Additivity of the flavor parameter implies the conservation of electric charge, lepton number, and baryon number.

Proof: Using Table I we derive the following relations:

$$\tilde{B} = -\tfrac{1}{2}c^2 = \text{baryon number},$$
$$\tilde{L} = c^1 - \tfrac{1}{2}c^2 + c^3 = \text{lepton number}, \tag{3.18}$$
$$\tilde{Q} = -c^3 = \text{electric charge}.$$

Since $\tilde{B}$, $\tilde{L}$, and $\tilde{Q}$ are linear functions of c^K, additivity of c^K implies additivity (conservation) of $\tilde{B}$, $\tilde{L}$, and $\tilde{Q}$. Q.E.D.

Hence, the tensor theory models the conservation of electric charge, lepton number, and baryon number.

APPENDIX A: FIERZ IDENTITIES

In this appendix we will present an elementary derivation of the Fierz identities, from which the tensor form of the Dirac Lagrangian will be derived in Appendix B. The spinor Fierz identity

(A13) is a consequence of trace formulas (A3) and (A4) satisfied by the Pauli matrices. We extend the spinor Fierz identity first to pairs of spinors (spinor doublets), and then to bispinors, which are defined in formula (A30) as spinor and dual conjugate spinor pairs.

A spinor is a two-dimensional complex vector, denoted as

$$\xi=\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}\in C^2. \tag{A1}$$

Acting on spinors ξ are the 2×2 complex Pauli matrices $\sigma^\alpha=(\sigma^0,\sigma^1,\sigma^2,\sigma^3)$ defined by

$$\sigma^0=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},\quad \sigma^1=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},$$

$$\sigma^2=\begin{bmatrix}0 & -i\\ i & 0\end{bmatrix},\quad \sigma^3=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}. \tag{A2}$$

We define $\sigma_\alpha=(\sigma^0,-\sigma^1,-\sigma^2,-\sigma^3)$ and denote $\tilde{\sigma}^\alpha=\sigma_\alpha$ and $\tilde{\sigma}_\alpha=\sigma^\alpha$. A straightforward evaluation of the Pauli matrices gives the following trace formula:

$$\mathrm{Tr}[\sigma_\alpha\tilde{\sigma}_\beta]=2g_{\alpha\beta}, \tag{A3}$$

where $g_{\alpha\beta}$ denotes the Minkowski metric tensor (with diagonal elements $\{1,-1,-1,-1\}$ and zeros off the diagonal). A further trace formula is expressed by:

$$\mathrm{Tr}[\sigma_\alpha\tilde{\sigma}_\delta\sigma_\beta\tilde{\sigma}_\gamma]=2C_{\alpha\beta\gamma\delta}, \tag{A4}$$

where, as will be seen in formula (A13), $C_{\alpha\beta\gamma\delta}$ is a Lorentz tensor. Such a tensor is a linear combination of $g_{\alpha\beta}g_{\gamma\delta}$, $g_{\alpha\gamma}g_{\beta\delta}$, $g_{\alpha\delta}g_{\beta\gamma}$, and $\epsilon_{\alpha\beta\gamma\delta}$, where $\epsilon_{\alpha\beta\gamma\delta}$ is the permutation tensor. A straightforward derivation shows that

$$C_{\alpha\beta\gamma\delta}=g_{\alpha\gamma}g_{\beta\delta}+g_{\alpha\delta}g_{\beta\gamma}-g_{\alpha\beta}g_{\gamma\delta}-i\epsilon_{\alpha\beta\gamma\delta}. \tag{A5}$$

The tensor $C_{\alpha\beta\gamma\delta}$ satisfies numerous identities, chief of which are the symmetries:

$$C_{\alpha\beta\gamma\delta}=C_{\beta\alpha\delta\gamma}=C_{\gamma\delta\alpha\beta}=C_{\delta\gamma\beta\alpha}, \tag{A6}$$

and the inversion formula:

$$C_{\alpha\beta\gamma\delta}C^{\gamma\delta\lambda\mu}=4\delta_\alpha^\lambda\delta_\beta^\mu, \tag{A7}$$

where δ_α^β equals one if $\alpha=\beta$ and zero otherwise. Note that the tensor indices α, β, γ, δ, λ, μ are lowered and raised using the Minkowski metric tensor $g_{\alpha\beta}$ and its inverse $g^{\alpha\beta}$, and all repeated indices are to be summed from 0 to 3.

Formulas equivalent to (A6) and (A7) are discussed by Penrose and Rindler.[34] Note that formulas vary depending on how one defines the matrices σ^α. Our choice of σ^α in formula (A2) is made in order to define the standard Dirac matrices γ^α in a consistent manner [see formula (A32)].

If we set $\delta=0$ in formula (A4), noting that $\tilde{\sigma}_0=\sigma^0=I$ where I is the 2×2 identity matrix, we have

$$\mathrm{Tr}[\sigma_\alpha\sigma_\beta\tilde{\sigma}_\gamma]=2C_{\alpha\beta\gamma 0}. \tag{A8}$$

Since the Pauli matrices σ_γ are a basis for 2×2 complex matrices, the product $\sigma_\alpha\sigma_\beta$ is a linear combination of the matrices σ_γ. From formulas (A3) and (A8), this linear combination is given by

$$\sigma_\alpha \sigma_\beta = C^{\gamma 0}_{\alpha\beta} \sigma_\gamma . \tag{A9}$$

By a similar argument, setting $\alpha=0$ in formula (A4) gives

$$\tilde{\sigma}_\gamma \tilde{\sigma}_\delta = C^{0\beta}_{\gamma\delta} \tilde{\sigma}_\beta . \tag{A10}$$

Next, we consider a pair of spinors ξ and η to which we associate a complex Lorentz four-vector j^α, whose components are defined by the 2×2 matrix:

$$j = 2\,\eta\xi^+ = \begin{bmatrix} j^0 + j^3 & j^1 - i j^2 \\ j^1 + i j^2 & j^0 - j^3 \end{bmatrix}, \tag{A11}$$

where $\xi^+ = (\overline{\xi_1}, \overline{\xi_2})$ denotes the transpose conjugate of ξ. (The overbar denotes ordinary complex conjugation.) The spin group of 2×2 complex matrices with determinant one, denoted SL$(2,C)$ or Spin (1,3), acts on the spinors ξ and η. Acting on $\eta\xi^+$ in formula (A11), the spin group leaves invariant the determinant of j, and hence the Minkowski norm of j^α. Thus, j^α becomes a Lorentz four-vector.

We can solve for j^α in formula (A11) by first noting that $j = j^\beta \tilde{\sigma}_\beta$, multiplying by σ_α, and then using the trace formula (A3). This defines a map $j_\alpha : C^2 \times C^2 \to C^4$, mapping each pair of spinors ξ and η to a complex Lorentz four-vector $j_\alpha(\xi,\eta)$, given by

$$j_\alpha(\xi,\eta) = \xi^+ \sigma_\alpha \eta. \tag{A12}$$

We now derive the following Fierz identity.

Proposition 1: For all $\xi,\ \eta,\ \kappa,\ \nu \in C^2$,

$$2 j_\alpha(\xi,\eta) j_\beta(\kappa,\nu) = C^{\gamma\delta}_{\alpha\beta} j_\gamma(\xi,\nu) j_\delta(\kappa,\eta). \tag{A13}$$

(Note that since j_α, j_β, j_γ, and j_δ are Lorentz four-vectors, $C_{\alpha\beta\gamma\delta}$ is a tensor.)

Proof: From formula (A11) we have

$$2\,\eta\xi^+ = j^\alpha(\xi,\eta) \tilde{\sigma}_\alpha . \tag{A14}$$

Then the trace formula (A4) gives

$$\begin{aligned} 2 j_\alpha(\xi,\eta) j_\beta(\kappa,\nu) &= 2(\xi^+ \sigma_\alpha \eta)(\kappa^+ \sigma_\beta \nu) = 2\,\mathrm{Tr}[\sigma_\alpha(\eta\kappa^+)\sigma_\beta(\nu\xi^+)] \\ &= \tfrac{1}{2}\mathrm{Tr}[\sigma_\alpha \tilde{\sigma}_\delta \sigma_\beta \tilde{\sigma}_\gamma] j^\gamma(\xi,\nu) j^\delta(\kappa,\eta) = C_{\alpha\beta\gamma\delta} j^\gamma(\xi,\nu) j^\delta(\kappa,\eta), \end{aligned} \tag{A15}$$

which proves formula (A13). Q.E.D.

The parity map $P : C^2 \to C^2$ sends a spinor ξ, as defined in formula (A1), to its dual conjugate $\tilde{\xi}$:

$$P\xi = \tilde{\xi} = \begin{bmatrix} \overline{\xi_2} \\ -\overline{\xi_1} \end{bmatrix}. \tag{A16}$$

Formula (A2) gives $P\sigma_\alpha \xi = \tilde{\sigma}_\alpha P\xi$. Hence, $\tilde{\xi} = P\xi$ transforms under the conjugate representation of the spin group SL$(2,C)$. Since $P^2\xi = -\xi$, the parity map P is a bijection. From formulas (A11) and (A12) we have

$$j^\alpha(P\xi, P\eta) = j_\alpha(\eta,\xi) = \overline{j_\alpha(\xi,\eta)}. \tag{A17}$$

Note that a ''parity operation'' can be defined for spinor fields $\xi(x^\alpha)$, which combines the parity map P with the space reflection, sending the space–time point $x^\alpha \in R^4$ to x_α. However, as defined here, the ''parity map'' $P : C^2 \to C^2$ transforms only the spinor components $\xi \in C^2$.

A spinor doublet is a four-dimensional complex vector, denoted as

$$\Phi=\begin{bmatrix}\xi\\ \eta\end{bmatrix}\in C^4, \tag{A18}$$

where $\xi,\ \eta\in C^2$ are spinors as defined in formula (A1). Acting on spinor doublets Φ are the Pauli matrices σ^α, which extend to 4×4 matrices by the usual direct sum representation. Also, acting on Φ are the 4×4 gauge matrices $t^K=(t^0,t^1,t^2,t^3)$ defined as follows. First, note from formula (A2) that, apart from $\pm$signs, all entries of the Pauli matrices σ^α consist of 0, 1, and i. The 4×4 complex matrices t^K are similar to the 2×2 Pauli matrices σ^α, with I and iI replacing 1 and i, where I is the 2×2 identity matrix. That is,

$$\begin{gathered} t^0=\begin{bmatrix}I & 0\\ 0 & I\end{bmatrix},\quad t^1=\begin{bmatrix}0 & I\\ I & 0\end{bmatrix},\\ t^2=\begin{bmatrix}0 & -iI\\ iI & 0\end{bmatrix},\quad t^3=\begin{bmatrix}I & 0\\ 0 & -I\end{bmatrix}. \end{gathered} \tag{A19}$$

The gauge matrices t^K satisfy trace formulas similar to the Pauli matrices σ^α. We define $t_K=(t^0,-t^1,-t^2,-t^3)$ and denote $\tilde{t}^K=t_K$ and $\tilde{t}_K=t^K$. For 4×4 matrices σ_α and t_K we have, similar to formula (A3),

$$\mathrm{Tr}[\sigma_\alpha\tilde{\sigma}_\beta]=4g_{\alpha\beta},\quad \mathrm{Tr}[t_J\tilde{t}_K]=4g_{JK}, \tag{A20}$$

where g_{JK} is a Minkowski metric similar to $g_{\alpha\beta}$. Similar to formula (A4), we have

$$\begin{gathered} \mathrm{Tr}[\sigma_\alpha\tilde{\sigma}_\delta\sigma_\beta\tilde{\sigma}_\gamma]=4C_{\alpha\beta\gamma\delta},\\ \mathrm{Tr}[t_J\tilde{t}_M t_K\tilde{t}_L]=4C_{JKLM}, \end{gathered} \tag{A21}$$

where C_{JKLM} is a tensor similar to $C_{\alpha\beta\gamma\delta}$.

The 4×4 complex matrices σ_α and t_K commute for each index α and K. Moreover,

$$\mathrm{Tr}[\sigma_\alpha]\mathrm{Tr}[t_K]=4\ \mathrm{Tr}[\sigma_\alpha t_K]=16\delta^0_\alpha\delta^0_K. \tag{A22}$$

As with σ_α in formulas (A9) and (A10), products of the gauge matrices t_K (or $\tilde{t}_K$) can be expressed as linear combinations of the t_K (or $\tilde{t}_K$), e.g.,

$$\tilde{t}_L\tilde{t}_M=C^{0K}_{LM}\tilde{t}_K. \tag{A23}$$

From this observation and formula (A22), if σ is a product of matrices σ_α and t is a product of matrices t_K, then

$$\mathrm{Tr}[\sigma]\mathrm{Tr}[t]=4\ \mathrm{Tr}[\sigma t]. \tag{A24}$$

Note that, as in formula (A23), gauge indices J,K,L,M are lowered and raised using the Minkowski metric g_{JK} and its inverse g^{JK}, and all repeated indices are to be summed from 0 to 3. Now consider the map $j^K_\alpha:C^4\times C^4\to C^{16}$, mapping each pair of spinor doublets Φ and Θ to a set of four complex Lorentz four-vectors $j^K_\alpha(\Phi,\Theta)$, defined by

$$j^K_\alpha(\Phi,\Theta)=\Phi^+\sigma_\alpha t^K\Theta, \tag{A25}$$

where Φ^+ denotes the transpose conjugate of Φ. We will derive the following Fierz identity for spinor doublets.

Proposition 2: For all Φ, Θ, Φ', $\Theta' \in C^4$,

$$C^{LM}_{JK} j^J_\alpha(\Phi,\Theta) j^K_\beta(\Phi',\Theta') = C^{\gamma\delta}_{\alpha\beta} j^L_\gamma(\Phi,\Theta') j^M_\delta(\Phi',\Theta). \tag{A26}$$

Proof: The proof is similar to the proof of Proposition 1, using the definition (A25) with the trace formulas (A20), (A21), and (A24). In particular, similar to formula (A14), we derive

$$4\Theta\Phi^+ = j^{\alpha K}(\Phi,\Theta)\tilde{\sigma}_\alpha \tilde{t}_K. \tag{A27}$$

An equation similar to formula (A15) is then obtained. The inversion formula (A7) is used in the final step. Q.E.D.

The parity map $P:C^4 \to C^4$ sends a spinor doublet Φ, as defined in formula (A18), to its dual conjugate doublet:

$$P\Phi = \begin{bmatrix} P\eta \\ -P\xi \end{bmatrix}. \tag{A28}$$

Note that we use the same symbol P for both the parity map (A16) defined on spinors as well as its extension (A28) to spinor doublets. It should be clear from the context which map $P:C^2 \to C^2$ or $P:C^4 \to C^4$ is intended. Since from formula (A16), $P^2\xi = -\xi$ and $P^2\eta = -\eta$, it follows from formula (A28) that $P^2\Phi = \Phi$. Thus, the parity map $P:C^4 \to C^4$ equals its inverse, i.e., $P = P^{-1}$. From formulas (A25) and (A27), we have

$$j^K_\alpha(P\Phi, P\Theta) = j^\alpha_K(\Theta,\Phi) = \overline{j^\alpha_K(\Phi,\Theta)}. \tag{A29}$$

A bispinor Ψ consists of a spinor ξ and a dual conjugate spinor $\tilde{\eta} = P\eta$. That is,

$$\Psi = B\Phi = \begin{bmatrix} \xi \\ P\eta \end{bmatrix}, \tag{A30}$$

where $B:C^4 \to C^4$ sends the spinor doublet Φ, as defined in formula (A18), to the bispinor Ψ. The charge conjugate of Ψ is defined by

$$\Psi^C = \begin{bmatrix} \eta \\ P\xi \end{bmatrix}, \tag{A31}$$

and the Dirac matrices γ^α and γ^5 acting on Ψ are defined by

$$\gamma^\alpha \Psi = \begin{bmatrix} P\sigma^\alpha \eta \\ \sigma^\alpha \xi \end{bmatrix}, \quad \gamma^5 \Psi = \begin{bmatrix} \xi \\ -P\eta \end{bmatrix}. \tag{A32}$$

Note from formulas (A28) and (A30) that the parity map P acting on bispinors Ψ is given by $\gamma^0 = BPB^{-1}$.

From formulas (A30), (A31), and (A32), the four-vectors (A25) expressed in terms of bispinors Ψ and Ξ become

$$j^K_\alpha(\Psi,\Xi) = \mathrm{Re}[i\bar{\Psi}\gamma_\alpha \tau^K \Xi] - i\,\mathrm{Im}[i\bar{\Psi}\gamma_\alpha T^K \Xi], \tag{A33}$$

where the gauge generators τ_K and T_K are defined in formulas (2.1) and (3.1). Note that in formula (A33) we employ the bispinor notation $\bar{\Psi} = (\gamma^0\Psi)^+ = \Psi^+\gamma^0$, and Ψ^+ denotes the transpose conjugate of Ψ. Also, note that for simplicity we have written $j^K_\alpha(\Psi,\Xi)$ instead of $j^K_\alpha(B^{-1}\Psi, B^{-1}\Xi)$. That is, we suppress the bispinor map $B^{-1}:C^4 \to C^4$, which sends the bispinors Ψ and Ξ to spinor doublets Φ and Θ in definition (A25).

The Cartan map[8] is defined on bispinors by

$$b_\alpha^K(\Psi,\Xi)=j_\alpha^K(\gamma^0\Psi,\Xi). \tag{A34}$$

The map $b_\alpha^K: C^4\times C^4\to C^{16}$ generalizes the bilinear spinor map originally discovered in 1913 by Cartan.[20] The Cartan map b_α^K, whose useful properties we exploit in Appendix B, is summarized in the following proposition.

Proposition 3: For all Ψ, Ξ, Ψ', $\Xi'\in C^4$,

$$C_{JK}^{LM} b_\alpha^J(\Psi,\Xi) b_\beta^K(\Psi',\Xi')=C_{\alpha\beta}^{\gamma\delta} b_\gamma^L(\Psi,\Xi') b_\delta^M(\Psi',\Xi). \tag{A35}$$

Moreover, the Cartan map b_α^K has the following symmetry property:

$$b_\alpha^K(\Psi,\Xi)=b_K^\alpha(\Xi,\Psi) \tag{A36}$$

and the parity property:

$$b_\alpha^K(\gamma^0\Psi,\gamma^0\Xi)=\overline{b_K^\alpha(\Psi,\Xi)}. \tag{A37}$$

Proof: Set $\Phi=B^{-1}(\gamma^0\Psi)$, $\Phi'=B^{-1}(\gamma^0\Psi')$, $\Theta=B^{-1}(\Xi)$, and $\Theta'=B^{-1}(\Xi')$ in formula (A26) to obtain formula (A35). Formulas (A36) and (A37) are consequences of formula (A29) and the identity $B^{-1}\gamma^0=PB^{-1}$. Q.E.D.

Note from the symmetry property (A36) that the components of $b_\alpha^K(\Psi,\Xi)$ are either symmetric or antisymmetric in the arguments Ψ and Ξ. For example,

$$b_0^0(\Psi,\Xi)=b_0^0(\Xi,\Psi) \tag{A38}$$

is symmetric; whereas for $\mathbf{b}_0=(b_0^1,b_0^2,b_0^3)$,

$$\mathbf{b}_0(\Psi,\Xi)=-\mathbf{b}_0(\Xi,\Psi) \tag{A39}$$

is antisymmetric. When $\Psi=\Xi$, the antisymmetric components of the Cartan map vanish.

A reduced form of the Fierz identity (A35) exploits a binary associative operation, denoted as $\otimes$, defined on C^4 as follows: Let p^K, q^K, $r^K\in C^4$, and define $r=p\otimes q$ if and only if

$$r^K=C_{LM}^{0K}p^L q^M. \tag{A40}$$

From formula (A5), writing $p^K=(p^0,\mathbf{p})$, where $\mathbf{p}=(p^1,p^2,p^3)$ and similarly for q^K and r^K, formula (A40) becomes

$$\begin{gathered} r^0=p^0q^0+\mathbf{p}\cdot\mathbf{q}, \\ \mathbf{r}=p^0\mathbf{q}+\mathbf{p}q^0+i\mathbf{p}\times\mathbf{q}. \end{gathered} \tag{A41}$$

It is straightforward to show that the bilinear operation $\otimes$ is associative [compare formulas (A23) and (A40)], and with this operation C^4 becomes the algebra of complex quaternions. Setting the index $L=0$ in the Fierz identity (A35) gives for all bispinors Ψ, Ξ, Ψ', $\Xi'\in C^4$,

$$[b_\alpha(\Psi,\Xi)\otimes b_\beta(\Psi',\Xi')]^M=C_{\alpha\beta}^{\gamma\delta} b_\gamma^0(\Psi,\Xi') b_\delta^M(\Psi',\Xi), \tag{A42}$$

where the Mth component of $p\otimes q$ is denoted as $[p\otimes q]^M$. We will see that quaternion algebra simplifies computations with bispinors in Appendix B.

As a first application of the Fierz identity (A42), we will derive the following identity:

$$j_\alpha^K j_{K\beta}=|s|^2 g_{\alpha\beta}, \tag{A43}$$

where

$$j_\alpha^K = j_\alpha^K(\Psi,\Psi) = b_\alpha^K(\gamma^0\Psi,\Psi),$$
$$s = b_0^0(\gamma^0\Psi,\gamma^0\Psi). \tag{A44}$$

Note using formulas (A30), (A32), and (A33) that the definitions (2.5) and (2.6) for the complex scalar field s and the Noether currents j_α^K are consistent with formula (A44). From formulas (A36) and (A37),

$$j_K^\alpha = b_\alpha^K(\Psi,\gamma^0\Psi), \quad \overline{s} = b_0^0(\Psi,\Psi). \tag{A45}$$

From formula (A36), $\overline{s}$ and s are the only nonvanishing components of $b_0^K(\Psi,\Psi)$ and $b_0^K(\gamma^0\Psi,\gamma^0\Psi)$, respectively. Thus, on substituting $\gamma^0\Psi$, Ψ, Ψ, $\gamma^0\Psi$ for Ψ, Ξ, Ψ', Ξ', respectively, in Fierz identity (A42), and setting $M=0$, formula (A43) is obtained.

APPENDIX B: PROOF OF THEOREM 1

In this appendix we prove Theorem 1 of Sec. III using the Fierz identity (A42). Let $\Psi(x)$ denote a smooth bispinor field defined at each space–time point $x \in R^4$. For all pairs of points x, $y \in R^4$, define the following two-point functions using formulas (A44) and (A45):

$$j_\alpha^K(x,y) = b_\alpha^K(\gamma^0\Psi(x),\Psi(y)),$$
$$q^K(x,y) = b_0^K(\Psi(x),\Psi(y)),$$
$$Q_\alpha(x,y) = b_\alpha^0(\Psi(x),\Psi(y)), \tag{B1}$$
$$r(x,y) = b_0^0(\Psi(x),\Psi(y)),$$
$$\overline{r(x,y)} = b_0^0(\gamma^0\Psi(x),\gamma^0\Psi(y)).$$

As previously defined in formula (A25), the $j_\alpha^K(x,y)$ for $K=0,1,2,3$ are a tetrad of complex Lorentz four-vectors, and as will become apparent, the $q^K(x,y)$ and $Q_\alpha(x,y)$ are complex quaternions with scalar part $r(x,y)$. We also denote, setting $y=x$,

$$j_\alpha^K(x) = j_\alpha^K(x,x), \quad r(x) = r(x,x). \tag{B2}$$

Note that by formula (A36), components of $q^K(x,y)$ and $Q_\alpha(x,y)$ with $K\neq 0$ or $\alpha \neq 0$ are antisymmetric in the variables x and y, and hence vanish for $y=x$. The only nonvanishing component of $q^K(x,x)$ and $Q_\alpha(x,x)$ is $r(x) = q^0(x,x) = Q_0(x,x)$. As in formula (3.12) we define

$$J_\alpha^K(x,y) = [j_\alpha(x,y)\otimes\tilde{c}]^K,$$
$$J_\alpha^K(x) = [j_a(x)\otimes\tilde{c}]^K, \tag{B3}$$

where $c^K \in R^4$ satisfies $c^K c_K = -1$, and $\tilde{c}^K = c_K$. Since in formula (3.8), $\Gamma_\alpha = i\gamma_\alpha c^K T_K$, we have from formulas (A33) and (B3),

$$\mathrm{Re}[i\overline{\Psi(x)}\Gamma_\alpha\,\partial_\beta\Psi(x)] = -\frac{\partial}{\partial y^\beta}\mathrm{Re}[iJ_\alpha^0(x,y)]\Big|_{y=x}. \tag{B4}$$

Thus, Theorem 1 follows from the following proposition.

Proposition 4:

$$\frac{\partial}{\partial y^{\alpha}} J_{\beta}^{K}(x,y)\bigg|_{y=x} = -\frac{1}{4|s|^2}[(\partial_\alpha J_\gamma)\otimes \tilde{J}^{\gamma}\otimes J_\beta + 2\bar{s}J_\beta\,\partial_\alpha s]^K, \tag{B5}$$

where $J_\alpha^K = J_\alpha^K(x)$ and $s=\overline{r(x)}$.

Proof: We will suppress gauge indices and simply write j_α for j_α^K and $\tilde{j}_\alpha$ for $j_{K\alpha}$, etc. The Fierz identity (A42) is then expressed as

$$b_\alpha(\Psi,\Xi)\otimes b_\beta(\Psi',\Xi') = C_{\alpha\beta}^{\gamma\delta} b_\gamma^0(\Psi,\Xi')b_\delta(\Psi',\Xi). \tag{B6}$$

On substituting $\Psi(x)$, $\Psi(y)$, $\gamma^0\Psi(x)$, $\Psi(x)$ for Ψ, Ξ, Ψ', Ξ', respectively, in the Fierz identity (B6), and setting the index $\alpha=0$, we get from formulas (B1) and (B2) and the vanishing of $Q_\gamma(x,x)$, for $\gamma\neq 0$,

$$q(x,y)\otimes j_\beta(x) = r(x) j_\beta(x,y). \tag{B7}$$

Quaternion multiplication of formula (B7) on the right by $\tilde{c}^K\in R^4$ satisfying $c^K c_K=-1$, gives, using formula (B3),

$$q(x,y)\otimes J_\beta(x) = r(x) J_\beta(x,y). \tag{B8}$$

Using the inversion formula (A7), the Fierz identity (B6) can also be expressed as

$$C_{\gamma\delta}^{\alpha\beta} b_\alpha(\Psi,\Xi)\otimes b_\beta(\Psi',\Xi') = 4 b_\gamma^0(\Psi,\Xi')b_\delta(\Psi',\Xi). \tag{B9}$$

Setting $\gamma=\delta=0$ and using formulas (A5) and (A36), we have

$$b_\alpha(\Psi,\Xi)\otimes \tilde{b}^{\alpha}(\Xi',\Psi') = 4 b_0^0(\Psi,\Xi')b_0(\Psi',\Xi). \tag{B10}$$

On substituting $\gamma^0\Psi(y)$, $\Psi(y)$, $\Psi(x)$, $\gamma^0\Psi(x)$ for Ψ, Ξ, Ψ', Ξ' in formula (B10) using formulas (B1) and (B2) we get

$$j_\alpha(y)\otimes \tilde{j}^{\alpha}(x) = 4\overline{r(y,x)}q(x,y). \tag{B11}$$

Since by formula (A41) $J_\alpha = j_\alpha\otimes\tilde{c}$ implies $\tilde{J}_\alpha = c\otimes\tilde{j}_\alpha$ and, furthermore, $\tilde{c}\otimes c=(-1,0,0,0)$, formulas (B3) and (B11) give

$$J_\gamma(y)\otimes \tilde{J}^{\gamma}(x) = -4\overline{r(y,x)}q(x,y). \tag{B12}$$

Quaternion multiplication of formula (B12) on the right by $J_\beta(x)$, and using formula (B8), gives

$$J_\gamma(y)\otimes \tilde{J}^{\gamma}(x)\otimes J_\beta(x) = -4\overline{r(y,x)}r(x)J_\beta(x,y). \tag{B13}$$

Since $r(x,y)=r(y,x)$ by formulas (A36) and (B1), we get

$$\frac{\partial}{\partial y^{\alpha}} r(y,x)\bigg|_{y=x} = \frac{1}{2}\frac{\partial}{\partial x^{\alpha}} r(x). \tag{B14}$$

Partially differentiating formula (B13) with respect to y^α, and setting $y=x$ using formula (B14), we obtain formula (B5). Q.E.D.

Corollary (Theorem 1):

$$\mathrm{Re}[i\bar{\Psi}\Gamma^{\alpha}\,\partial_\alpha\Psi] = -\frac{1}{4|s|^2}\mathrm{Re}[(\partial_\alpha\mathbf{J}_\beta)\cdot\mathbf{J}^{\alpha}\times\mathbf{J}^{\beta} - 2i\bar{s}J_\alpha^0\,\partial^\alpha s], \tag{B15}$$

where $J_\alpha^K = (J_\alpha^0, \mathbf{J}_\alpha)$.

Proof: Substitute formula (B5) with $K=0$ into formula (B4) and evaluate the associative operation $\otimes$ using formula (A41), noting by formulas (2.6) and (3.9) that both J_α^0 and $\mathbf{J}_\alpha \cdot \mathbf{J}_\beta$ are real. Q.E.D.

[1] F. Reifler and R. Morris, ''Unification of the Dirac and Einstein Lagrangians in a tetrad model,'' J. Math. Phys. **36**, 1741–1752 (1995).
[2] F. Reifler and R. Morris, ''Inclusion of gauge bosons in the tensor formulation of the Dirac theory,'' J. Math. Phys. **37**, 3630–3640 (1996).
[3] F. Reifler and R. Morris, ''Fermi quantization of tensor systems,'' Int. J. Mod. Phys. A **9**, 5507–5515 (1994).
[4] M. Carmeli, E. Leibowitz, and N. Nissani, *Gravitation: SL(2,C) Gauge Theory and Conservation Laws* (World Scientific, Singapore, 1990), pp. 21–27.
[5] R. Hammond, ''Dirac coupling in gravity with torsion potential,'' Class. Quantum Grav. **12**, 279–285 (1995).
[6] A. Ashtekar and R. Geroch, ''Quantum theory of gravitation,'' Rep. Prog. Phys. **37**, 1211–1256 (1974).
[7] B. DeWitt, *Supermanifolds* (Cambridge University Press, Cambridge, 1985), pp. 229–230 and 235–237.
[8] F. Reifler and R. Morris, ''Unobservability of bispinor two-valuedness in Minkowski space–time,'' Ann. Phys. (N.Y.) **215**, 264–276 (1992).
[9] F. Reifler and R. Morris, ''A gauge symmetric approach to Fierz identities,'' J. Math. Phys. **27**, 2803–2806 (1986).
[10] Y. Takahashi, ''The Fierz identities—a passage between spinors and tensors,'' J. Math. Phys. **24**, 1783–1790 (1983).
[11] S. Rodriguez-Romo, F. Viniegra, and J. Keller, ''Geometrical content of the Fierz identities,'' *Clifford Algebras and Their Applications in Mathematical Physics*, edited by A. Micali *et al.* (Kluwer, Dordrecht, 1992), pp. 479–497.
[12] S. Rodriguez-Romo, ''An analysis of Fierz identities, factorization and inversion theorems,'' Found. Phys. **23**, 1535–1542 (1993).
[13] J. Keller and S. Rodriguez-Romo, ''Multivector representation of Lie groups,'' Int. J. Theor. Phys. **30**, 185–196 (1991).
[14] W. Krolikowski, ''A sequence of Clifford algebras and three replicas of a Dirac particle,'' Acta Phys. Pol. B **21**, 871–879 (1990).
[15] J. Keller, ''Dirac equations with electroweak and color symmetry,'' *Differential Geometric Methods in Theoretical Physics 1*, edited by S. Catto and A. Rocha (World Scientific, Singapore, 1992), pp. 355–361.
[16] J. P. Crawford, ''Hypergravity I and II,'' *Clifford (Geometric) Algebras*, edited by W. E. Baylis (Birkhauser, Boston, 1996), pp. 341–363.
[17] M. Carmeli, Kh. Huleihil, and E. Leibowitz, *Gauge Fields: Classification and Equations of Motion* (World Scientific, Singapore, 1989), p. 40.
[18] F. Reifler, ''A vector model for electroweak interactions,'' J. Math. Phys. **26**, 542–550 (1985).
[19] F. Reifler and R. Morris, ''A prediction of the Cabibbo angle in the vector model for electroweak interactions,'' J. Math. Phys. **26**, 2059–2066 (1985).
[20] E. Cartan, *The Theory of Spinors* (Dover, New York, 1981), 41–42.
[21] F. Reifler, ''A vector wave equation for neutrinos,'' J. Math. Phys. **25**, 1088–1092 (1984).
[22] J. Mickelsson, ''The vector form of the neutrino equation and the photon neutrino duality,'' J. Math. Phys. **26**, 2346–2349 (1985).
[23] F. Reifler and A. Vogt, ''Unique continuation of some dispersive waves,'' Commun. Partial Diff. Eqns. **19**, 1203–1215 (1994).
[24] J. Keller and S. Rodriguez-Romo, ''Multivectorial generalization of the Cartan map,'' J. Math. Phys. **32**, 1591–1598 (1991).
[25] J. Keller and S. Rodriguez-Romo, ''A multivectorial Dirac equation,'' J. Math. Phys. **31**, 2501–2510 (1990).
[26] V. A. Zhelnorovich, ''Complex vector triplets in the spinor theory in Minkowski space,'' Proc. Acad. Sci. USSR **311**, 590–593 (1990).
[27] F. Reifler and R. Morris, ''The Hamiltonian structure of Dirac's equation in tensor form and its Fermi quantization,'' *Workshop on Squeezed States and Uncertainty Relations*, edited by D. Han, Y. S. Kim, and W. W. Zachary, NASA 3135, 1992, pp. 381–383.
[28] F. Reifler and R. Morris, ''Particle localization, spinor two-valuedness, and Fermi quantization of tensor systems,'' *3rd International Workshop on Squeezed States and Uncertainty Relations*, edited by D. Han, Y. S. Kim, N. H. Rubin, Y. Shih, and W. W. Zachary, NASA 3270, 1994, pp. 281–286.
[29] B. Thaller, *The Dirac Equation* (Springer-Verlag, Berlin, 1992).
[30] M. P. DoCarmo, *Riemannian Geometry* (Birkhauser, Boston, 1992), pp. 40–41.
[31] F. Reifler and R. Morris, ''Nonlinear modes of the tensor Dirac equation and CPT violation,'' *Workshop on Harmonic Oscillators*, edited by D. Han, Y. S. Kim, and W. W. Zachary, NASA 3197, 1993, pp. 289–294.
[32] G. B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974), pp. 485–510.
[33] C. W. Kim and A. Pevsner, *Neutrinos in Physics and Astrophysics* (Harwood Academic, Langhorne, 1993), pp. 22–30.
[34] R. Penrose and W. Rindler, *Spinors and Space–Time, Volume 1* (Cambridge University Press, Cambridge, 1986), pp. 153–155.

Superconnections and the Higgs field

G. Roepstorff[a)]
Institute for Theoretical Physics, RWTH Aachen, D-52062 Aachen, Germany

(Received 5 October 1998; accepted for publication 7 January 1999)

Within the mathematical framework of Quillen, one interprets the Higgs field as part of the superconnection $\mathbb{D}$ on a superbundle. We propose to take as superbundle the exterior algebra $\wedge V$ obtained from a Hermitian vector bundle V of rank n with structure group U(n) and study the curvature $\mathbb{F}=\mathbb{D}^2$. The Euclidean action, at most quadratic in $\mathbb{F}$ and invariant under gauge transformations, depends on $n+1$ central charges. Spontaneous symmetry breaking is related to a nonvanishing constant scalar curvature in the ground state, $\mathbb{F}=L_c^2$, where L_c is the Higgs condensate. The $U(1)$ Higgs model is nothing but the familiar Ginzburg–Landau theory, whereas the $U(2)$ Higgs model relates to the electro-weak theory (without matter fields). The present formulation leads to the relation $g^2=3g'^2$ for the coupling constants, the formula $\sin^2\theta=1/4$ for the Weinberg mixing angle, and the ratio $m_W^2:m_Z^2:m_H^2=3:4:12$ for the masses of $W^\pm$, Z^0, and the Higgs boson. Experimentally observed deviations are attributed to loop corrections. © *1999 American Institute of Physics.* [S0022-2488(99)01006-3]

I. INTRODUCTION

It has been generally accepted that spontaneously broken local gauge symmetries provide the correct framework for understanding the electro-weak interactions of elementary particles.[1] The mechanism that gives masses to the $W^\pm$, Z^0, and leptons however needs the introduction of a doublet of scalar fields, the so-called Higgs field, with many puzzling features, physically as well as mathematically. The concepts of the Higgs field and the related Higgs mechanism, over the years, have triggered many investigations, either from the supersymmetry or the differential geometry point of view.

Most attempts were a response to the fact that the Lagrangian of the standard model contains a large number of free parameters, among them various gauge coupling constants, the parameters of the Higgs potential, coupling constants of matter fields, and the elements of the quark mixing matrix. Some of these constants are expected to come out of some kind of symmetry breaking mechanism occurring in some yet unknown theory while others can be chosen at our will. One therefore feels that at present one is actually dealing with an effective (low energy, long range) field theory where only some degrees of freedom appear explicitly. Consequently, no explanation for most of the constants, chosen to fit the experimental data, is offered.

As a normal mathematical setting one would perhaps regard the theory of fiber bundles[2,3] that emerged as a primary tool for studying Yang–Mills systems. Then the question may be raised: Is the Higgs field an object of geometry? Below we shall briefly survey some of the attempts to extend the formalism of gauge theory to Yang–Mills–Higgs systems before trying to give new answers.

A popular approach to the problem of assigning a geometrical role to the Higgs field comes under the heading *dimensional reduction*. Witten,[4] Manton,[5] and Fairlie[6] were first to provide interesting model theories in higher dimension. The reduction technique has been taken up again and used as a guiding principle by other authors.[7] In its simplest version it uses one extra dimension, flat space, and translational invariance. Thus, one starts from a Yang–Mills connection on

[a)]Electronic mail: roep@physik.rwth-aachen.de

the trivial principal bundle $\mathbb{R}^{n+1}\times G$ with compact semisimple Lie group G, considers the splitting $dx^0A_0+dx^iA_i$ ($i=1,...,n$) of the connection one-form, and identifies the Higgs field with A_0. The drawback is twofold; (1) the gauge field is not allowed to depend on the extra variable x^0, (2) the Higgs field is always in the adjoint representation.

As a precursor of Quillen's superconnection theory, one may regard Ne'eman's proposal[8] to make use of the supergroup SU(2|1) for an algebraically irreducible electro-weak unification. Supergroups are formal objects obtained from super-Lie algebras where commutators are replaced by supercommutators. At first, the model appeared to suffer from spin-statistics complications. The final treatment with Sternberg[9] however took full advantage of Quillen's formalism. Super-Lie algebras are also at the heart of an attempt[10] to construct a renormalizable model of gravity as a broken gauge theory.

Another approach borrows from the framework of noncommutative geometry (NCG) (Ref. 11) and leads to what has been called *algebraic Yang–Mills–Higgs theories*[12] with obvious links to the supergroup formalism. The idea is to replace the exterior algebra of differential forms (the de Rham complex) by some noncommutative $\mathbb{Z}_2$-graded differential algebra. To start with, one replaces $C^\infty(M)$ by $A\otimes C^\infty(M)$, where A is some matrix algebra together with a grading automorphism and M is spacetime. As there is a generalized notion of what should be called a connection, by a proper choice of the algebra A one can accommodate a Higgs field in the connection, be it one multiplet or several multiplets. By now, many versions of the NCG approach have appeared which successfully reformulate the standard model. In the Connes–Lott approach,[13] $A=\mathbb{C}\oplus\mathbb{C}$ whereas the Mainz–Marseille group (see Ref. 12 for details) prefers $A=M_2(\mathbb{C})$ with grading automorphism diag$(1,-1)$ in both cases. Recently, Okumura[14] proposed yet another formulation. When calculating the curvature, these authors get different results which influence the Weinberg angle, the Higgs mass and the quartic Higgs coupling. Therefore, the predictive power of the NCG approach has come under intense scrutiny.

Recently, generalized Dirac operators have been proposed as a ''model building kit'' for action functionals that include the full standard model as well as gravity by Ackermann and Tolksdorf.[15] In their scheme the bosonic part of the action is given by the Wodzicki residue and the Higgs field is intimately related to the gravitational potential. Though this approach hardly uses NCG concepts, it somehow parallels the Connes–Lott approach.

Last but not least there are attempts to add a fifth ''discrete dimension''[16] to space–time with possible relation to parity and chiral symmetry breaking. We feel, though cannot prove, that such an approach, once fully worked out, will provide but another reformulation of a specific model within the territory of noncommutative geometry.

In 1985, Quillen described his concept of a superconnection[17] (see also Ref. 18), thereby abandoning the traditional $\mathbb{Z}$-grading (of the exterior algebra of differential forms) in favor of a $\mathbb{Z}_2$-grading, giving thus more freedom to constructions in (commutative) differential geometry. Bundles carrying a $\mathbb{Z}_2$-graded structure are termed superbundles. Quillen aimed at the construction of invariants of a superbundle (Chern–Weil forms) and the definition of the Chern character of a superconnection. A serious attempt to extend the formalism of gauge theories using Quillen's concept of superconnections has been launched in 1990 by Coquereaux *et al.*[19] It still borrows from the NCG formalism. So does the work of Lee[20] and Figueroa *et al.*[21]

In the present paper, we do not rely on the NCG approach but strictly follow the guidelines of Quillen and try to paint a coherent picture of U(n) Higgs systems whose ground states have constant generalized curvature (to be defined below). The role we assign to the Higgs field is similar to the one of the NCG and the Ackermann–Tolksdorf approach. But the choice of the superbundle is new to the best of our knowledge. Here we restrict ourselves to discussing the bosonic part of the action. In a second paper we shall include fermions and introduce the Dirac operator of a Clifford superconnection to construct the action of the electro-weak theory.

The formalism we propose has applications to Ginzburg–Landau theory[22] and topological field theory.[23] We shall leave out that aspect here.

II. SUPERBUNDLES

We assume that M is an oriented Riemannian manifold of dimension m which may be arbitrary. Later we shall be more specific and think of M as the four-dimensional Euclidean space–time with flat Levi-Civita connection. We let Γ denote the algebra of smooth complex functions on M and, if B is some bundle with base M, write $\Gamma(M,B)$ for the space of smooth sections $s:M \to B$.

Pure Yang–Mills theory starts from a principal G bundle P over M, and an invariant action functional on the set $\mathcal{A}$ of connections on P. The compact semisimple Lie group G is called the *gauge group* of the theory. The set $\mathcal{A}$ is modeled on the vector space of gauge fields, i.e., the space of 1-forms A taking values in the bundle

$$\mathrm{ad}\, P = P \times_G \mathbf{g},$$

where $\mathbf{g}$ is the Lie algebra of G, and G acts on $\mathbf{g}$ via the adjoint representation. A gauge field A can locally be written as $dx^\mu A_\mu(x)$ with $A_\mu(x) \in \mathbf{g}$. The electro-weak theory, however, is based on the nonsemisimple gauge group $\mathrm{SU}(2)\times\mathrm{U}(1)$ and thus admits two independent gauge couplings. In what follows the focus will be on the nonsemisimple case. Though we work with the gauge group $\mathrm{U}(n)$, we argue in favor of only *one* coupling constant.

The notion of a gauge transformation is more subtle. Though, locally, gauge transformations may be thought of as maps from the base manifold M into the group G, they cannot be extended globally to sections of P (unless P is a trivial bundle). Instead, gauge transformations are *bundle automorphisms* of P. Automorphisms commute with the group action on P by definition. Let $\mathcal{G} = \mathrm{Aut}(P)$ denote the group of bundle automorphisms of P. A more explicit description of $\mathcal{G}$, which is closer to the physicists' notion, uses sections of the adjoint bundle,

$$\mathcal{G} = \Gamma(M, \mathrm{Ad}\, P), \quad \mathrm{Ad}\, P = P \times_G G.$$

The bundle $\mathrm{Ad}(P)$ is the associated bundle whose fibers are copies of the group G. But the group action on G is the adjoint action. The Lie algebra of $\mathcal{G}$ can now be easily constructed; $\mathrm{Lie}\,\mathcal{G} = \Gamma(M, \mathrm{ad}\, P)$.

From now on we shall assume that P is a principal bundle with structure group $G = \mathrm{U}(n)$. Let the complex vector space $\mathbb{C}^n$ be equipped with the standard scalar product so that its group of automorphisms is G. We may construct the associated bundle

$$V = P \times_G \mathbb{C}^n$$

which is a complex Hermitian vector bundle of rank n with structure group G. It is always understood that G acts on the right of P and on the left of $\mathbb{C}^n$, and the notation $\times_G$ means that we identify $(pg,z) \sim (p,gz)$ for $p \in P$, $z \in \mathbb{C}^n$, and $g \in G$.

Since algebraic constructions on vector spaces carry over to associated bundles, we may consider the exterior algebra $\wedge V$ which is a Hermitian vector bundle of rank 2^n acted upon by gauge transformations $u \in \mathcal{G}$ via the representation $\wedge$, namely, at $x \in M$ we have, for $u(x) \in \mathrm{U}(n)$,

$$\wedge u : \wedge V_x \to \wedge V_x, \quad \wedge u(v_1 \wedge \cdots \wedge v_i) = (uv_1) \wedge \cdots \wedge (uv_i). \tag{1}$$

Recall now that a *superspace* is a $\mathbb{Z}_2$-graded vector space whose elements are said to have even or odd degree (or parity). Likewise, a *superbundle* is a vector bundle whose fibers are superspaces. Furthermore, a *superalgebra* has a superspace as underlying vector space, and a product that respects the $\mathbb{Z}_2$-grading. The exterior algebra $\wedge V$ is both a superbundle and a superalgebra with grading

$$\wedge V = \wedge^+ V \oplus \wedge^- V, \quad \wedge^\pm V = \sum_{(-1)^p = \pm 1} \wedge^p V.$$

Though the subbundles $\wedge^{\pm}V$ have the same rank 2^{n-1}, there exists no natural isomorphism between them. It will soon become apparent that only a spontaneous symmetry breaking connects $\wedge^{+}V$ with $\wedge^{-}V$.

The remainder of this section is devoted to reviewing basic facts about the representation $\wedge = \Sigma_k \wedge^k$ of $G = \mathrm{U}(n)$ on $\wedge\mathbb{C}^n$. Since the subrepresentations $\wedge^k$ are irreducible and inequivalent, the commutant

$$(\wedge G)' = \{\Sigma_{k=0}^{n} c_k P_k | c_k \in \mathbb{C}\},$$

where P_k projects onto $\wedge^k\mathbb{C}^n$, is an Abelian algebra. Particular elements $C \in (\wedge G)'$ will enter the Euclidean action. If $C = \Sigma c_k P_k$, we shall refer to the numbers c_k as *central charges*.

Another consequence is that the representation $\wedge$ of G respects the $\mathbb{Z}_2$-grading of $\wedge\mathbb{C}^n$ and decomposes as $\wedge^{+} \oplus \wedge^{-}$. We thus write

$$\wedge u = \begin{pmatrix} \wedge^{+}u & 0 \\ 0 & \wedge^{-}u \end{pmatrix}, \quad u \in U(n).$$

Because the operator $\wedge u$ does not change the parity of vectors, it is said to be *even*.

Similar properties may be established for the induced representation $a \mapsto \hat{a}$ of the Lie algebra $\mathbf{g} = \mathbf{u}(n)$ given by

$$\hat{a} = \frac{d}{dt} \wedge \exp(ta)|_{t=0} = \begin{pmatrix} \hat{a}^{+} & 0 \\ 0 & \hat{a}^{-} \end{pmatrix}, \quad a \in \mathbf{u}(n), \quad \hat{a}^{\pm} \in \mathrm{End} \wedge^{\pm}\mathbb{C}^n.$$

In fact, $\hat{a}$ is the unique extension of $a \in \mathrm{End}\,\mathbb{C}^n$ to an even derivation of the algebra $\wedge\mathbb{C}^n$, i.e.,

$$\hat{a}(z \wedge z') = \hat{a}z \wedge z' + z \wedge \hat{a}z', \qquad z, z' \in \wedge\mathbb{C}^n.$$

In particular,

$$\hat{a}z = 0 \qquad z \in \wedge^0\mathbb{C}^n \cong \mathbb{C},$$

$$\hat{a}z = az \qquad z \in \wedge^1\mathbb{C}^n \cong \mathbb{C}^n.$$

An operator L on $\wedge\mathbb{C}^n$ is even (odd) if it preserves (changes) parity. This gives $\mathrm{End} \wedge\mathbb{C}^n$ the structure of superalgebra,

$$\mathrm{End} \wedge\mathbb{C}^n = \mathrm{End}^{+} \wedge\mathbb{C}^n \oplus \mathrm{End}^{-} \wedge\mathbb{C}^n.$$

Note that $\hat{a} \in \mathrm{End}^{+} \wedge\mathbb{C}^n$.

Up to normalization there exists a unique bilinear form $q(a,b)$, or equivalently a quadratic form $q(a) = q(a,a)$, on the Lie algebra $\mathbf{su}(n)$, known as the Killing form, which is invariant and nondegenerate. By contrast, the Lie algebra $\mathbf{u}(n) = \mathbf{su}(n) \oplus \mathbf{u}(1)$, where $n \geqslant 2$, has a two-parameter family of such forms (we require that they be positive definite) parameterized by g and g',

$$q(a) = -\frac{n}{g^2}\,\mathrm{tr}\left(a - \frac{1}{n}\,\mathrm{tr}\,a\right)^2 - \frac{1}{g'^2}(\mathrm{tr}\,a)^2, \quad a \in \mathbf{u}(n). \tag{2}$$

When restricted to the subalgebra $\mathbf{su}(n)$, any member of this family reduces to a multiple of the Killing form as it should,

$$\mathrm{tr}\,a = 0 \Rightarrow q(a) \sim \mathrm{tr}\,\mathrm{ad}(a)^2 = 2n\,\mathrm{tr}\,a^2.$$

In the context of the electro-weak theory, g and g' are known as the two independent gauge coupling constants. Unless one is committed to a specific representation of the Lie algebra, there will be no *a priori* relation between g and g'. On the other hand, given a distinguished faithful representation ρ, the condition $q(a)\sim -\mathrm{tr}\,\rho(a)^2$ fixes the ratio g/g' once and for all.

In fact, according to the point of view taken in this paper, the U(n) Higgs model starts from a distinguished representation, namely $\rho(a)=\hat{a}$, and thus provides a canonical choice for the value of the ratio g/g'. Writing Tr for the trace on End $\wedge\mathbb{C}^n$ (we reserve tr for traces in other circumstances) and setting

$$q(a)=-\mathrm{Tr}\,\hat{a}^2,\quad a\in\mathbf{u}(n), \tag{3}$$

we also fix the value of g which is solely a matter of convenience without intrinsic meaning.

This choice of $q(a)$ may appear as an ''article of faith'' and is certainly questionable; the matter is not being debated here. Instead, we will demonstrate how g and g' are related. Expanding both sides of the well-known formula

$$\log \mathrm{Tr}\exp\hat{a}=\mathrm{tr}\log(\mathbb{1}+e^a),$$

we get

$$\mathrm{Tr}\,\mathbb{1}=2^n,\quad \mathrm{Tr}\,\hat{a}=2^{n-1}\,\mathrm{tr}\,a,\quad \mathrm{Tr}\,\hat{a}^2=2^{n-2}((\mathrm{tr}\,a)^2+\mathrm{tr}\,a^2). \tag{4}$$

Comparison with (2) shows that $(g/g')^2=n+1$. For the electro-weak theory $(n=2)$, we get the equation $g^2=3g'^2$ apart from the value $g^2=2$ owing to our choice of normalization in (3).

To prepare for later work, we introduce a basis e_i $(i=1,...,n^2)$ in $\mathbf{u}(n)$ such that $q(e_i,e_k)=\delta_{ik}$. It is also assumed that e_i $(i=1,...,n^2-1)$ is a basis for $\mathbf{su}(n)$.

Let us now investigate the two simplest situations U(1) and U(2). In the U(1) case, it is obvious that

$$\wedge u=\begin{pmatrix}1 & 0\\ 0 & u\end{pmatrix},\quad \hat{a}=\begin{pmatrix}0 & 0\\ 0 & a\end{pmatrix},\quad u=e^a\in U(1),\quad a\in i\mathbb{R}.$$

As expected, $e_1=i$, since $q(i)=-i^2=1$.

We treat the U(2) case in greater detail. Here $2^n=4$ and so $\wedge u$ is a unitary 4×4 matrix which is block diagonal (blocks are 2×2 matrices),

$$\wedge u=\begin{pmatrix}\wedge^+u & 0\\ 0 & \wedge^-u\end{pmatrix},\quad \wedge^+u=\begin{pmatrix}1 & 0\\ 0 & \det u\end{pmatrix},\quad \wedge^-u=u\in\mathrm{U}(2).$$

It follows that

$$\hat{a}=\begin{pmatrix}\hat{a}^+ & 0\\ 0 & \hat{a}^-\end{pmatrix},\quad \hat{a}^+=\begin{pmatrix}0 & 0\\ 0 & \mathrm{tr}\,a\end{pmatrix},\quad \hat{a}=a\in\mathbf{u}(2),$$

and $q(a)=-(\mathrm{tr}\,a)^2-\mathrm{tr}(a^2)$ by (3) and (4). Setting

$$a=a^ke_k=\frac{i}{\sqrt{2}}\begin{pmatrix}\frac{1}{\sqrt{3}}a^4+a^3 & a^1-ia^2\\ a^1+ia^2 & \frac{1}{\sqrt{3}}a^4-a^3\end{pmatrix}\quad (a^k\in\mathbb{R}),$$

we obtain a basis in $\mathbf{u}(2)$ with the required property $q(a)=\Sigma_k(a^k)^2$.

III. DIFFERENTIAL FORMS

We let T^*M denote the complexified cotangent bundle of the manifold M. Sections of T^*M are said to be (complex) 1-forms. The exterior algebra $\wedge T^*M$ is a superbundle and so is $\wedge(T^*M\oplus V)$. The latter construction relates to the dimensional reduction formalism in an obvious way. For, if N is another manifold and $V=T^*N$, then $T^*M\oplus V\cong T^*(M\times N)$. Whether or not some manifold N of dimension n is lurking behind the scene, there exists a natural isomorphism

$$\wedge(T^*M\oplus V)\cong\wedge T^*M\otimes\wedge V \tag{5}$$

between $\mathbb{Z}_2$-graded algebras. The tensor product on the right-hand side of (5) is special for graded algebras. It is often called a *skew tensor product*. Generally speaking, if X and Y are $\mathbb{Z}_2$-graded algebras, the multiplication in $X\otimes Y$ is given by

$$(x\otimes y)(x'\otimes y)=(-1)^r xx'\otimes yy',\quad r=\begin{cases}1 & \text{if } x' \text{ and } y \text{ are odd}\\ 0 & \text{otherwise.}\end{cases}$$

The fact that the isomorphism (5) respects the grading means

$$\wedge^{\pm}(T^*M\oplus V)\cong\sum_{(-1)^{p+q}=\pm 1}\wedge^p T^*M\otimes\wedge^q V,$$

or stated equivalently,

$$\wedge^{+}(T^*M\oplus V)\cong(\wedge^{+}T^*M\otimes\wedge^{+}V)\oplus(\wedge^{-}T^*M\otimes\wedge^{-}V),$$

$$\wedge^{-}(T^*M\oplus V)\cong(\wedge^{-}T^*M\otimes\wedge^{+}V)\oplus(\wedge^{+}T^*M\otimes\wedge^{-}V).$$

The grading carries over to the space of $\wedge V$-valued differential forms,

$$\Omega:=\Gamma(M,\wedge T^*M\otimes\wedge V)=\Omega^{+}\oplus\Omega^{-},$$

and to the algebra of sections of the endomorphism bundle,

$$\mathcal{B}:=\Gamma(M,\wedge T^*M\otimes\mathrm{End}\wedge V)=\mathcal{B}^{+}\oplus\mathcal{B}^{-}. \tag{6}$$

Note that $\mathrm{End}\wedge V$ is a superbundle, and the tensor product in (6) is between graded algebras. Elements of $A\in\mathcal{B}$ act on Ω and are called *local operators* since they leave fibers intact. Equivalently, they commute with the multiplication by functions $f\in\Gamma$.

A local operator $A\in\mathcal{B}^{\pm}$ is said to have parity ± 1—or is referred to as an even (odd) operator—where parity is defined as follows:

$$\mathrm{par}(A)=+1 \Leftrightarrow A\Omega^{\pm}\subset\Omega^{\pm},$$

$$\mathrm{par}(A)=-1 \Leftrightarrow A\Omega^{\pm}\subset\Omega^{\mp}.$$

A different decomposition of $\mathcal{B}$ arises from the $\mathbb{Z}$-grading of $\wedge T^*M$,

$$\mathcal{B}=\sum_{p=0}^{m}\mathcal{B}^p,\quad \mathcal{B}^p=\Gamma(M,\wedge^p T^*M\otimes\mathrm{End}\wedge V).$$

Notice that $1\otimes id$ serves as the unit in the algebra $\mathcal{B}$ and that there are two natural embeddings,

$$\Gamma(M,\wedge T^*M)\to\mathcal{B},\quad \omega\mapsto\omega\otimes id,$$

$$\Gamma(M,\mathrm{End}\wedge V)\to\mathcal{B},\quad A\mapsto 1\otimes A.$$

Owing to these embeddings, various constructions on $\Gamma(M,\wedge T^*M)$ and $\Gamma(M,\mathrm{End}\wedge V)$ have extensions to $\mathcal{B}$.

For instance, the operator of exterior differentiation d on $\Gamma(M,\wedge T^*M)$ may be extended.

The trace $\mathrm{Tr}:\Gamma(M,\mathrm{End}\wedge V)\to\Gamma$ can be extended in an obvious manner,

$$\mathrm{Tr}:\Gamma(M,\wedge T^*M\otimes\mathrm{End}\wedge V)\to\Gamma(M,\wedge T^*M),\omega\otimes A\mapsto\omega\,\mathrm{Tr}\,A.$$

Any local operator A may be decomposed into homogeneous components (p-forms),

$$A=A_{[0]}+A_{[1]}+A_{[2]}+\cdots,\quad A_{[p]}\in\mathcal{B}^p.$$

The series truncates at $p=m$ where m is the dimension of the manifold M, and taking the trace of the top form, the integral

$$\mathrm{Int}(A)=\int_M \mathrm{Tr}\,A_{[m]}\in\mathbb{C},\tag{7}$$

assigns complex numbers to local operators of compact support.

The Hodge star operator on $\Gamma(M,\wedge T^*M)$ can uniquely be extended to a real-linear operator on $\mathcal{B}$ so as to satisfy

$$*(AB)=B^**A,\quad A\in\mathcal{B},B\in\mathcal{B}^0.\tag{8}$$

Let f_i $(i=1,...,m)$ be an oriented frame of the tangent bundle TM and f^i the dual frame of T^*M. For any multi-index $I\subset\{1,...,m\}$ we form the exterior product,

$$f^I=f^{i_1}f^{i_2}\cdots f^{i_p},\quad I=\{i_1,i_2,\ldots,i_p\},\quad i_1<i_2\cdots<i_p,\quad p=|I|\tag{9}$$

to obtain a frame of $\wedge T^*M$. It is assumed that $f^{\varnothing}=1$. Using (8), we have

$$*(f^I\otimes A_I)=*((f^I\otimes id)(1\otimes A_I))=(1\otimes A_I^*)(*f^I\otimes id)=(\pm 1)^{m-p}*f^I\otimes A_I^*,\quad A_I\in\Gamma(M,\mathrm{End}^{\pm}\wedge V).$$

Let $d\tau=*1=f^1f^2\cdots f^m$ denote the volume element. Then there are functions $g^I\in\Gamma$ such that $f^I*f^I=g^I\cdot d\tau$ and from (9), $g^I=\det(f^{i_k},f^{i_l})_{k,l=1,\ldots,p}>0$.

The algebra $\mathcal{B}$ may be equipped with a scalar product,

$$(A,B)=\mathrm{Int}(B*A),\quad A,B\in\mathcal{B},\tag{10}$$

and, by a straightforward calculation,

$$\|A\|^2=(A,A)=\int_M d\tau\sum_I g^I\,\mathrm{Tr}(A_IA_I^*)\geqslant 0,\quad A=\sum_I A_If^I.$$

The norm $\|\cdot\|$ on $\mathcal{B}$ will be used in Sec. V to construct the Euclidean action.

IV. SUPERCONNECTIONS

We start with a few remarks about connections. With P a principal G bundle, where $G=U(n)$, the space $\mathcal{A}$ of connections is an affine space with nontrivial topology if $n\geqslant 2$, e.g., $\pi_0(\mathcal{A})=\mathbb{Z}$. With $\mathcal{G}$ acting on $\mathcal{A}$, it seems natural to pass to the quotient

$$B\mathcal{G}=\mathcal{A}/\mathcal{G}$$

to obtain the classifying space for $\mathcal{G}$ bundles. In physics, $B\mathcal{G}$ is known as the space of gauge orbits. It corresponds to the phase space of classical mechanics. The passage to statistical mechanics is mirrored, in Euclidean field theory, by the process of quantization, i.e., the introduction of path integrals over $\mathcal{A}$. Granted the absence of anomalies, path integrals project onto $B\mathcal{G}$. The calculation, however, requires gauge fixing and the introduction of Faddeev–Popov ghosts. Gribov's discovery, also known as the *Gribov ambiguity*, may be rephrased by saying that there is no continuous global choice of gauge or, stated more formally, $\mathcal{A}$ does not admit a smooth global section. Though these intricacies are not the subject of the present paper, we should be aware that some of the formulas below hold only on local coordinate patches without explicit mentioning.

The advantage of giving a connection A on the principal bundle P is that it determines a connection on every associated bundle and thus provides covariant derivatives d_A on various vector bundles. We use the terms *connection* and *covariant derivative* interchangeably. A connection on the bundle V simply is a linear map

$$d_A:\Gamma(M,V)\rightarrow\Gamma(M,T^*M\otimes V)$$

satisfying the Leibniz rule $d_A(fs)=dfs+fd_As$ for all functions f and sections s. The connection extends in a unique way to an operator d_A on $\Gamma(M,\wedge T^*M\otimes V)$ sending p-forms to $(p+1)$-forms. Locally, $d_A=d+A$, where $A\in\Gamma(M,T^*M\otimes \mathrm{End}\,V)$ is the connection 1-form or gauge field. The 2-form $F=d_A^2\in\Gamma(M,\wedge^2T^*M\otimes\mathrm{End}\,V)$ is said to be the *curvature* of the connection d_A. In terms of physics, F is the field strength of a gauge theory. Under a gauge transformation,

$$^uA=uAu^{-1}+udu^{-1},\quad {}^uF=uFu^{-1},\quad u\in\mathcal{G}.$$

We may pass now to the superbundle $\wedge V$ and lift the fields A and F to certain local operators on Ω of definite parity,

$$\hat{A}=\begin{pmatrix}\hat{A}^+ & 0\\ 0 & \hat{A}^-\end{pmatrix}\in\mathcal{B}^-,\quad \hat{F}=\begin{pmatrix}\hat{F}^+ & 0\\ 0 & \hat{F}^-\end{pmatrix}\in\mathcal{B}^+.$$

Of course, $\hat{A}$ and $\hat{F}$ are still 1 and 2-forms, respectively. Recall that the matrix representation refers to the $\mathbb{Z}_2$-grading of $\wedge V$. In the same manner, d_A can be lifted to a connection D on the superbundle,

$$D=\begin{pmatrix}D^+ & 0\\ 0 & D^-\end{pmatrix},\quad \hat{F}=D^2.$$

Locally, we have $D=d+\hat{A}$ and $D^\pm=d+\hat{A}^\pm$. When acting on Ω, the differential operator D changes the parity, and so is of odd type.

To extend the connection D to a superconnection $\mathbb{D}=D+L$ we introduce a skew selfadjoint operator L on $\Gamma(M,\wedge V)$ of odd type,

$$L=\begin{pmatrix}0 & i\Phi^*\\ i\Phi & 0\end{pmatrix},$$

formally a section of the bundle

$$\wedge^0T^*M\otimes\mathrm{End}^-\wedge V\cong\mathrm{End}^-\wedge V,$$

and hence an element of $\mathcal{B}^-\cap\mathcal{B}^0$. The complex scalar field $\Phi(x)$ is said to be the *Higgs field* of the system. It has the following characteristic properties:

(a) At $x \in M$, the Higgs field $\dot{\Phi}(x)$ is a linear map from $\wedge^+ V_x$ to $\wedge^- V_x$. Consequently, $\Phi^*(x)$ maps $\wedge^- V_x$ to $\wedge^+ V_x$.

(b) Under a change of the gauge,

$$^u\Phi = (\wedge^- u)\Phi(\wedge^+ u)^{-1}, \quad ^u\Phi^* = (\wedge^+ u)\Phi^*(\wedge^- u)^{-1},$$

which is summarized by

$$^uL = (\wedge u)L(\wedge u)^{-1}.$$

(c) Like any section of the bundle $\mathrm{End}^- \wedge V, L$ extends to an odd operator on Ω. In more detail, L acts on $\wedge V$-valued p-forms by

$$L(dx^{\mu_1}\wedge\cdots\wedge dx^{\mu_p} v_{\mu_1\cdots\mu_p}(x)) = (-1)^p dx^{\mu_1}\wedge\cdots\wedge dx^{\mu_p} L(x) v_{\mu_1\cdots\mu_p}(x)$$

so as to be in accord with the skew tensor product $\wedge T^*M \otimes \mathrm{End}\wedge V$. To put it differently, L satisfies the rule $\{L, dx^\mu\} = 0$ or, equivalently, L anticommutes with the multiplication by Γ-valued 1-forms. Thus $L: \Omega^\pm \to \Omega^\mp$ is parity changing and hence $L \in \mathcal{B}^-$ by construction.

(d) Since both d and L are odd degree operators, their anticommutator (or supercommutator) $dL := \{d, L\}$ is an even operator (and a 1-form) called the covariant derivative of L. Similarly, the anticommutator,

$$DL := \{D, L\} = \begin{pmatrix} 0 & i(D\Phi)^* \\ iD\Phi & 0 \end{pmatrix},$$

provides the covariant derivatives of the Higgs field and its adjoint,

$$D\Phi := D^-\Phi + \Phi D^+ = d\Phi + \hat{A}^-\Phi + \Phi\hat{A}^+ = dx^\mu(\partial_\mu\Phi + \hat{A}^-_\mu\Phi - \Phi\hat{A}^+_\mu),$$
$$(D\Phi)^* := D^+\Phi^* + \Phi^* D^- = d\Phi^* + \hat{A}^+\Phi^* + \Phi^*\hat{A}^- = dx^\mu(\partial_\mu\Phi^* + \hat{A}^+_\mu\Phi^* - \Phi^*\hat{A}^-_\mu).$$

Here, we used the fact that Φ and Φ^* anticommute with dx^μ.

Finally, the operator

$$\mathbb{D} = D + L = \begin{pmatrix} D^+ & i\Phi^* \\ i\Phi & D^- \end{pmatrix}$$

defines a *superconnection* on the superbundle $\wedge V$ in the sense of Quillen; $\mathbb{D}$ is a differential operator of odd type on Ω, hence acts on $\wedge V$-valued differential forms. It no longer sends p-forms to $(p+1)$-forms, but sends odd elements of Ω to even elements and vice versa so as to satisfy a Leibniz formula.

In Physics, fields are viewed as varying objects. Varying both the gauge field and the Higgs field means passage from one superconnection $\mathbb{D}$ to another, say $\mathbb{D}'$, such that the difference $\mathbb{D} - \mathbb{D}'$ comes out as a local operator built upon 1-forms (the diagonal parts) and 0-forms (the off-diagonal parts). Hence the notion of a *superconnection on a superbundle* is in accordance with the requirement that, whatever the context, connections form an affine space modeled on some set of local operators.

From $\mathbb{F} = (D+L)^2 = D^2 + \{D, L\} + L^2$ we obtain the decomposition

$$\mathbb{F} = \mathbb{F}_{[0]} + \mathbb{F}_{[1]} + \mathbb{F}_{[2]} \in \mathcal{B}^+, \quad \mathbb{F}_{[p]} \in \mathcal{B}^p$$

for the curvature $\mathbb{F}$ of the superbundle $\wedge V$. In particular, the curvature is a local operator (not a differential operator). Note that the Bianchi identity $[\mathbb{D}, \mathbb{F}] = 0$ is a trivial consequence of the definition of $\mathbb{F}$.

As indicated, the curvature $\mathbb{F}$ has homogeneous components for $p=0,1,2$. The 0-form is bilinear in the Higgs field,

$$\mathbb{F}_{[0]}=L^2=\begin{pmatrix} -\Phi^*\Phi & 0 \\ 0 & -\Phi\Phi^* \end{pmatrix},$$

while the 1-form is linear in the covariant derivatives of the Higgs field,

$$\mathbb{F}_{[1]}=DL=\begin{pmatrix} 0 & i(D\Phi)^* \\ iD\Phi & 0 \end{pmatrix}.$$

Finally, the 2-form

$$\mathbb{F}_{[2]}=D^2=\hat{F}=\begin{pmatrix} \hat{F}^+ & 0 \\ 0 & \hat{F}^- \end{pmatrix}$$

gives the curvature when the Higgs field is absent.

V. EUCLIDEAN ACTION AND STATIONARY POINTS

We shall always stay within the realm of Euclidean field theory. For the remainder of this paper, M denotes the four-dimensional Euclidean flat spacetime with standard orientation, $d\tau = dx^1\wedge dx^2\wedge dx^3\wedge dx^4$.

Before describing the field equations of the Higgs model, we motivate the construction of a gauge invariant Euclidean action based on the superbundle $\wedge V$ and the gauge group $\mathrm{U}(n)$. With F the curvature of a Yang–Mills connection, one takes $S=\frac{1}{2}\|F\|^2$ as the action so that the global minimum is attained for the flat connection. Similarly, the superbundle is flat if $\mathbb{F}=0$. However, the definition $S=\frac{1}{2}\|\mathbb{F}\|^2$ gives us models that show no sign of spontaneous symmetry breaking. To our rescue comes the abelian algebra $(\wedge G)'$ of gauge invariant operators C, each of them constant on M. If C is self-adjoint, the following definition of the Euclidean action serves the purpose,

$$S=\tfrac{1}{2}\|\mathbb{F}+\mu^2 C\|^2, \quad C\in(\wedge G)'. \tag{11}$$

Euclidean actions that differ by the choice of C are said to be *phases* of the same model. As an element of an abelian algebra, C can always be written in terms of central charges $c_k, k=0,\ldots,n$. Self-adjointness of C makes these charges real numbers. We may write

$$C=\begin{pmatrix} C^+ & 0 \\ 0 & C^- \end{pmatrix}, \quad C^\pm=\sum_{(-1)^k=\pm 1} c_k P_k$$

and split the action into different parts for easier interpretation,

$$S=\tfrac{1}{2}\|\hat{F}\|^2+\tfrac{1}{2}\|DL\|^2+\tfrac{1}{2}\|L^2+\mu^2 C\|^2. \tag{12}$$

The last term involves the Higgs potential $V(\Phi)$,

$$\tfrac{1}{2}\|L^2+\mu^2 C\|^2=\int_M d\tau V(\Phi)$$

$$\begin{aligned} V(\Phi)&=\tfrac{1}{2}\,\mathrm{Tr}(L^2+\mu^2 C)^2 \\ &=\tfrac{1}{2}\,\mathrm{Tr}(\Phi^*\Phi-\mu^2 C^+)^2+\tfrac{1}{2}\,\mathrm{Tr}(\Phi\Phi^*-\mu^2 C^-)^2. \end{aligned} \tag{13}$$

In (12) we encounter the term $\frac{1}{2}\|\hat{F}\|^2$ as part of the action. To analyze it we introduce the components of the curvature F with respect to the basis dx^μ in T^*M and the basis e_k in $\mathbf{u}(n)$,

$$F=F^k e_k, \quad F^k=\tfrac{1}{2}dx^\mu\wedge dx^\nu F^k_{\mu\nu}(x).$$

The coefficients $F^k_{\mu\nu}(x)$ are real functions on M. It follows that $\hat{F}=-\hat{F}^*=F^k\hat{e}_k$ and, in view of the definitions (10) and (7),

$$\frac{1}{2}\|\hat{F}\|^2=\int_M d\tau \frac{1}{4}\sum_{k,\mu\nu}(F^k_{\mu\nu})^2.$$

This reveals that $\frac{1}{2}\|\hat{F}\|^2$ is the Euclidean action, correctly normalized, of a conventional gauge theory without Higgs field.

The positive parameter μ sets the mass scale of the model while the central charges c_k control the expectation value of the Higgs field (the so-called condensate) on the classical (or tree) level. The role of these parameters is similar when the theory is quantized using path integrals. The latter procedure replaces the classical Higgs potential by an effective potential whose minima describe the vacuum states.

One remark is in order. Recall that we have deliberately put g^2 equal to 2. It may later be necessary to work with an arbitrary value of the gauge coupling g. The introduction of g as a parameter can be achieved by an appropriate scaling,

$$S\rightarrow\lambda^{-2}S(\lambda A,\lambda\Phi), \quad 2\lambda^2=g^2.$$

Since scaling has no influence on second order terms of the action, it will not alter the results of the present paper.

Let us concentrate on the term $\frac{1}{2}\|DL\|^2$ of the action and its dependence on the connection (i.e., the gauge potential A),

$$S_L(A):=\tfrac{1}{2}\|DL\|^2=\int_M d\tau\ \mathrm{Tr}\sum_\mu (DL)^*_\mu(DL)_\mu,$$

where $DL=dx^\mu(DL)_\mu$ and $(DL)_\mu=\partial_\mu L+[\hat{A}_\mu,L]$. As usual, the response to a small change of the connection determines the current of the model,

$$S_L(A+a)=S_L(A)+(\hat{a},\{DL,L\})+O(a^2).$$

From $\{DL,L\}=dx^\mu[(DL)_\mu,L]$ and $a=dx^\mu a^k_\mu e_k$ we obtain $\hat{a}=dx^\mu a^k_\mu\hat{e}_k$ and

$$(\hat{a},\{DL,L\})=\int_M d\tau\sum_\mu a^k_\mu\,\mathrm{Tr}(\hat{e}_k[(DL)_\mu,L]).$$

Thus the current $j=dx^\mu j^k_\mu e_k$, which is Lie algebra valued 1-form, has components

$$j^k_\mu=\mathrm{Tr}(\hat{e}_k[(DL)_\mu,L]),$$

while the structure of the commutator in terms of the Higgs field Φ is as follows:

$$[(DL)_\mu,L]=\begin{pmatrix}\Phi^*(D\Phi)_\mu-(D\Phi)^*_\mu\Phi & 0\\ 0 & \Phi(D\Phi)^*_\mu-(D\Phi)_\mu\Phi^*\end{pmatrix}.$$

Let us now consider the formal adjoint operators d^*, d_A^*, and D^*. They belong to the standard repertoire of Yang–Mills systems. Adjoints are always formed with respect to the scalar product of sections. Each of the above adjoint operators maps p-forms into $(p-1)$-forms. The operator $\delta := -d^*$ is called the *coderivative* and $\Delta = -\{d, d^*\}$ the *Laplacian.*

From the condition that the action S be stationary one obtains the field equations of the U(n) Higgs model,

$$D^*\hat{F} + \hat{j} = 0, \quad D^*DL = \{L, L^2 + \mu^2 C\}.$$

For this and similar calculations, it is useful to keep in mind that $D^*\hat{F}$ is short for $[D^*, \hat{F}]$ and D^*DL is short for $[D^*, \{D, L\}]$. The current j has been lifted to obtain $\hat{j} = dx^\mu j_\mu^k \hat{e}_k$.

The field equations may also be put into a form reminiscent of previous Yang–Mills–Higgs models,

$$D_A^* F + j = 0, \quad D^*D\Phi = -2\Phi\Phi^*\Phi + \mu^2(\Phi C^+ + C^-\Phi). \tag{14}$$

Again, $d_A^* F$ is short for $[d_A^*, F]$, and $D^*D\Phi$ is obtained from

$$[D^*, \{D, L\}] = \begin{pmatrix} 0 & i(D^*D\Phi)^* \\ iD^*D\Phi & 0 \end{pmatrix}.$$

Any solution (A, Φ) of the second order field Eqs. (14) is said to be a stationary point of the action. We should be aware that not every stationary point corresponds to a local or global minimum of the action.

The global minimum is attained if $A = 0$ and if L solves of the variational problem

$$\mathrm{Tr}(L^2 + \mu^2 C)^2 = \text{minimum}. \tag{15}$$

The solutions are said to describe (classical) vacua or ground states. Granted that M is connected, any solution L_c of (15) is constant on M and is referred to as the *Higgs condensate*. The group U(n) acts upon the set of solutions, though not always freely; the residual gauge group

$$G_0 = \{u \in \mathrm{U}(n) \,|\, (\wedge u) L_c = L_c (\wedge u)\} \tag{16}$$

may well be nontrivial. Ground states that lie on the same gauge orbit are physically equivalent. We must not expect the group U(n) to act transitively; there may exist many gauge orbits.

If L_c^2 is unique, we obtain a U(n) invariant vacuum. Nonuniqueness is characteristic of a broken phase. Note also that each ground state has constant scalar curvature, $\mathbb{F} = L_c^2$. The structure of L_c is that of a constant matrix,

$$L_c = i\begin{pmatrix} 0 & \Phi_c^* \\ \Phi_c & 0 \end{pmatrix}, \quad \Phi_c = \mu v.$$

There are two special cases where the variational problem (15) can be solved with ease. First, $C = 0$ implies $L_c = 0$ giving a U(n) invariant vacuum. Second, $C = 1$ implies $L_c^2 = \mu^2$, hence $v^*v = vv^* = 1$, and v establishes an isomorphism between the spaces $\wedge^+\mathbb{C}^n$ and $\wedge^-\mathbb{C}^n$. Conversely, any isomorphism v gives us a solution of (15).

Suppose we look for exitations from some ground state, but ignore the Higgs degrees of freedom. Then L is kept constant, i.e., $L = L_c$ and $DL = \{\hat{A}, L_c\} = dx^\mu [\hat{A}_\mu, L_c]$. Provided $[\hat{A}_\mu, L_c] \neq 0$, the gauge particles acquire masses. Indeed, the mass term of the action may be written

$$\frac{1}{2}\|DL_c\|^2 = \int_M d\tau \frac{1}{2} \sum_\mu Q(A_\mu),$$

where Q is a positive semidefinite quadratic form on the Lie algebra,

$$Q(a) = -\mathrm{Tr}[\hat{a}, L_c]^2 = a^i a^k m_{ik}^2, \quad a = a^i e_i \in \mathbf{u}(n). \tag{17}$$

The eigenvalues of the matrix (m_{ik}^2) are the masses (squared) of gauge fields given by the eigenvectors where it is assumed that the eigenvectors are orthonormal with respect to the bilinear form $q(a,b)$ on the Lie algebra obtained from (3).

Suppose now that L_c' is another Higgs condensate giving rise to the quadratic form $Q'(a)$. Both L_c and L_c' lie on the same gauge orbit if $(\wedge u)L_c = L_c'(\wedge u)$ for some $u \in U(n)$. Owing to the invariance property $Q(a) = Q'(uau^{-1})$, the eigenvalues of the mass matrix stay constant along any gauge orbit.

VI. THE U(1) HIGGS MODEL

A very simple situation arises when $n=1$ since there is only one basis element $e_1 = i$ in $\mathbf{u}(1) = i\mathbb{R}$. We may thus write $A = i dx^\mu A_\mu(x)$ and $F = i\frac{1}{2} dx^\mu \wedge dx^\nu F_{\mu\nu}(x)$ with real-valued components A_μ and $F_{\mu\nu}$. In the two-dimensional cap representation of $\mathbf{u}(1)$ we have

$$\hat{A} = \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix}, \quad \hat{F} = \begin{pmatrix} 0 & 0 \\ 0 & F \end{pmatrix}.$$

The Higgs field is simply some complex scalar field Φ. With c_0 and c_1 the central charges, the Higgs potential becomes

$$V(\Phi) = \tfrac{1}{2}(|\Phi|^2 - \mu^2 c_0)^2 + \tfrac{1}{2}(|\Phi|^2 - \mu^2 c_1)^2 = (|\Phi|^2 - \mu^2 c)^2 + \tfrac{1}{4}\mu^2 (c_0 - c_1)^2, \quad c = \tfrac{1}{2}(c_0 + c_1).$$

Provided that $c > 0$, the minimum is attained for $\Phi = \mu c e^{i\alpha}$. Otherwise, the minimum is attained for $\Phi = 0$. There is no restriction in assuming that $c_0 = c_1$ and $c = \pm 1$. From

$$\mathbb{D} = \begin{pmatrix} d & \Phi^* \\ -\Phi & d + A \end{pmatrix}, \quad \mathbb{F} = \begin{pmatrix} -|\Phi|^2 & i(d_A \Phi)^* \\ i d_A \Phi & F - |\Phi|^2 \end{pmatrix},$$

we obtain the action of the Ginzburg–Landau theory,

$$S = \int_M d\tau \left(\frac{1}{4} \sum_{\mu\nu} F_{\mu\nu}^2 + \sum_\mu |(\partial_\mu + iA_\mu)\Phi|^2 + (|\Phi|^2 - \mu^2 c)^2 \right),$$

whose current is given by

$$j = i dx^\mu j_\mu(x), \quad j_\mu = 2\, \mathrm{Im}(\Phi^* (\partial_\mu + iA_\mu)\Phi).$$

For $c=1$, the system is in the superconducting phase, the residual gauge group is trivial, and ground states differ by a constant phase, $\Phi(x) = \mu e^{i\alpha}$. However, these states belong to a single gauge orbit and hence are equivalent. Provided Φ is kept at its ground state value, S reduces to the action of a massive photon $(m_\gamma^2 = 2\mu^2)$,

$$S = \int_M d\tau \left(\frac{1}{4} \sum_{\mu\nu} F_{\mu\nu}^2 + \mu^2 \sum_\mu A_\mu^2 \right).$$

For $c = -1$, the system is in the Coulomb phase. Exitations from the ground state show that the vector particle (i.e., the photon) has zero mass.

VII. THE U(2) HIGGS MODEL

We now come to the $n=2$ situation. The U(2) connection 1-form A may be written in the basis e_k, $k=1,...,4$ as introduced in Sec. II,

$$A=\frac{i}{\sqrt{2}}\begin{pmatrix}\frac{1}{\sqrt{3}}A^4+A^3 & A^1-iA^2\\ A^1+iA^2 & \frac{1}{\sqrt{3}}A^4-A^3\end{pmatrix},\quad A^k=dx^\mu A^k_\mu(x),$$

where $A^k_\mu(x)$ $(k=1,...,4)$ are real gauge fields. A similar decomposition holds for the curvature $F=d_A^2$. Notice that $\operatorname{tr} A=i\sqrt{2/3}A^4$. We move on to write the superconnection $\mathbb{D}$ (a 4×4 matrix) in block form,

$$\mathbb{D}=\begin{pmatrix} d+\hat{A}^+ & i\Phi^* \\ i\Phi & d+\hat{A}^-\end{pmatrix},\quad \hat{A}^+=\begin{pmatrix}0 & 0\\ 0 & i\sqrt{2/3}A^4\end{pmatrix},\quad \hat{A}^-=A.$$

Subgroups of U(2) have a specific interpretation in the context of the electro-weak theory. For instance, the U(1) subgroup consisting of phase transformations leads to the conservation of the weak hypercharge in the unbroken phase. The Higgs field Φ is some 2×2-matrix whose columns represent Higgs doublets in the fundamental representation of SU(2),

$$\Phi=\begin{pmatrix}\Phi_1 & \Phi_3\\ \Phi_2 & \Phi_4\end{pmatrix}.$$

The two doublets have opposite weak hypercharge. For an account of the physical implications of two-doublet models see the review article by Sher.[24]

There are five real second-order forms, invariant under U(2), that one may construct from the Higgs field,

$$R_1(\Phi)=|\Phi_1|^2+|\Phi_2|^2,\quad R_3(\Phi)=|\Phi_1^*\Phi_3+\Phi_2^*\Phi_4|,$$

$$R_2(\Phi)=|\Phi_3|^2+|\Phi_4|^2,\quad R_4(\Phi)+iR_5(\Phi)=\Phi_1\Phi_4-\Phi_2\Phi_3.$$

Since they satisfy the relation

$$R_1R_2=R_3^2+R_4^2+R_5^2, \tag{18}$$

only four of them are algebraically independent. It is natural to think of the manifold (18) as some moduli space related to superconnections.

In principle, any gauge invariant Higgs potential $V(\Phi)$, be it the classical or the effective potential, can be written as a function of the above invariants. Such a representation is convenient because the problem of minimizing the action is then reduced to solving a simpler problem in lower dimension. Each solution provides certain constants

$$r_i=R_i(\Phi_c),\quad i=1,...5,$$

which characterize the gauge orbit of Φ_c, and the moduli space of vacua becomes a submanifold of

$$r_1r_2=r_3^2+r_4^2+r_5^2. \tag{19}$$

Given the numbers r_i, the next step would be to determine the eigenvalues of the mass matrix m^2 of the vector bosons as defined by (17). Since the matrix elements depend on the choice of Φ_c, hence on the point of the gauge orbit chosen, we better look at the characteristic polynomial of that matrix which is gauge invariant and thus can be written entirely in terms of the variables r_i. By a tedious but straightforward calculation one finds

$$\det(m^2-\lambda)=(\lambda-r)^2(\lambda^2-\tfrac{4}{3}r\lambda+\tfrac{4}{3}r_3^2),\quad (0\leqslant 2r_3\leqslant r),\tag{20}$$

where $r=r_1+r_2=-\frac{1}{2}\operatorname{Tr}L_c^2$. An immediate consequence is the following alternative:

(1) If $r=0$, the eigenvalue zero of the mass matrix is fourfold degenerate: all vector bosons are massless.
(2) If $r>0$, the eigenvalue r of the mass matrix is twofold degenerate. We interpret r as the mass (squared) of the $W^{\pm}$ bosons. If $r_3=0$, there is an eigenvalue zero naturally associated to the photon and an eigenvalue $\frac{4}{3}r$ interpreted as the mass (squared) of the Z boson so that $m_W^2:m_Z^2=3:4$. In the latter case, the residual gauge group is isomorphic to U(1), leading to the notion of the electric charge. The $W^{\pm}$ bosons receive the charge ± 1, while the Z boson is neutral.

So far we have not fixed the value of the mass parameter μ. From now on we shall always assume that $\mu^2=r$ so that μ coincides with the W mass. In other words, it is the W mass that sets the mass scale. The U(2) Higgs model conforms to the existing empirical data only if $r_3=0$. It is therefore important to show that $r_3=0$ is not an extra assumption but follows from the Higgs potential (13). Though the Higgs potential depends on arbitrary constants c_0, c_1, and c_2, it is special among U(2) invariant fourth-order polynomials. For a simple calculation reveals that

$$V(\Phi)=(R_1(\Phi)-b_1)^2+(R_2(\Phi)-b_2)^2+2R_3(\Phi)^2+b_3^2.\tag{21}$$

Here we passed from the set c_i of constants to another more convenient set b_i given by

$$b_1=\tfrac{1}{2}\mu^2(c_0+c_1),\quad b_2=\tfrac{1}{2}\mu^2(c_2+c_1),\quad b_3=\tfrac{1}{2}\mu^2((c_0-c_1)^2+(c_2-c_1)^2)^{1/2}.$$

While b_3 is physically irrelevant, b_1 and b_2 are essential for spontaneous symmetry breaking. It is now obvious that any ground state has coordinates

$$r_i=\max(0,b_i),\quad (i=1,2),\quad r_3=0,\quad r_4^2+r_5^2=r_1r_2=\text{const},$$

which establishes two things; (1) our claim that $r_3=0$ and (2) the moduli space of vacua is the sphere S^1 provided $r_1r_2>0$, or simply a point if either $r_1=0$ or $r_2=0$. Granted the condition $r_3=0$ we can always perform a U(2) gauge transformation so that the Higgs condensate assumes the form

$$\Phi_c=\begin{pmatrix} r_1^{1/2}e^{i\alpha} & 0 \\ 0 & r_2^{1/2}\end{pmatrix},\quad r_1+r_2=\mu^2,\quad r_4+ir_5=(r_1r_2)^{1/2}e^{i\alpha},\tag{22}$$

with α parametrizing the sphere S^1. The choice of such a standard form is essential for getting a standard set of eigenvectors of the mass matrix. In fact it follows at once from (17) and (22) that

$$Q(a)=\mu^2\left(|a^1+ia^2|^2+\left(\frac{1}{\sqrt{3}}a^4+a^3\right)^2\right),$$

and thus the eigenvectors of the mass matrix are

mass^2	eigenvector
0	$\frac{1}{2}(\sqrt{3}a^4-a^3)$
$\frac{4}{3}\mu^2$	$\frac{1}{2}(\sqrt{3}a^3+a^4)$
μ^2	$a^1,\ a^2$

The residual gauge group, isomorphic to U(1), is

$$G_0=\left\{\begin{pmatrix}1 & 0\\ 0 & e^{i\alpha}\end{pmatrix}\in \mathrm{U}(2)\,\middle|\,0\leqslant\alpha\leqslant 2\pi\right\}. \tag{23}$$

We relate the photon, the Z boson, and the W boson to the above eigenvectors and thus work with the following fields:

$$A^0=\tfrac{1}{2}(\sqrt{3}A^4-A^3),\quad Z=\tfrac{1}{2}(\sqrt{3}A^3+A^4),\quad W^{\pm}=\frac{1}{\sqrt{2}}(A^1\mp iA^2). \tag{24}$$

It is common practice to write $Z=\cos\theta\,A^3+\sin\theta\,A^4$ with θ the Weinberg angle and to determine $\sin^2\theta$ by experiment. Comparison with (24) shows that $\sin^2\theta=1/4$ in the present theory while the generally accepted value obtained from experiment is $\sin^2\theta=0.231$. The value 1/4, however, has previously been predicted on different grounds (see, for instance, Refs. 6 and 8).

As a matrix, the connection 1-form may now be written in terms of A^0, Z, and W,

$$A=i\begin{pmatrix}\sqrt{\frac{2}{3}}Z & W^+\\ W^- & \sqrt{\frac{1}{2}}A^0-\sqrt{\frac{1}{6}}Z\end{pmatrix}. \tag{25}$$

To summarize, the choice of the central charges c_i does not seem to matter as long as we keep r at a fixed value, say μ^2. That this impression is false will become clear as soon as the coupling to matter is taken into account. In a forthcoming paper, the Yukawa interaction of fundamental fermions with the Higgs field results from a widening of the concept of Dirac operators and so is viewed as integral part of the gauge coupling. It will then become clear that the invariant parameters r_1 and r_2 are proportional to the masses (squared) of the pair (ν_e,e) in a purely leptonic model (one generation only). Vanishing of the neutrino mass requires that $r_1=0$. In this situation, the moduli space of vacua shrinks to a point. This puts another constraint on the parameters c_i, namely $c_0+c_1\leqslant 0$ or $b_1\leqslant 0$.

Let us now discuss the Higgs field itself. The assignment of electric charges to the four complex degrees of freedom can be read off from $\left(\begin{smallmatrix}0 & +1\\ -1 & 0\end{smallmatrix}\right)$. The one-doublet model of Salam and Weinberg is recovered if

$$\Phi=\begin{pmatrix}0 & \Phi^+\\ 0 & \Phi^0\end{pmatrix} \tag{26}$$

which is consistent with $r_1=0$ but not with $r_1>0$.

The two-doublet model we propose has eight real degrees of freedom. Three of them can be gauged away, giving an extra polarization degree of freedom to each massive gauge field. The remaining five degrees can be arranged as follows:

$$\Phi = \begin{pmatrix} X^0 & 0 \\ X^- & \sqrt{\frac{1}{2}}\phi \end{pmatrix} + \Phi_c .$$

There are two complex fields X^0 and X^-. The real field ϕ describes the Higgs particle of the conventional theory. To determine the (bare) masses of these fields we expand the Higgs potential to second order assuming $r_1=0$,

$$V(\Phi)=b_1^2+b_3^2-2b_1|X^0|^2+2(\mu^2-b_1)|X^-|^2+2\mu^2\phi^2+\cdots .$$

Recall now that $2b_1=\mu^2(c_0+c_1)<0$. Therefore,

$$m_{X^0}^2=\mu^2|c_0+c_1|, \quad m_{X^-}^2=\mu^2(2+|c_0+c_1|), \quad m_H^2=4\mu^2.$$

In this scenario, the hypthetical X particles have masses that depend on the constants c_i while the mass of the Higgs particle is not influenced by their values. We may state our results as

$$m_W^2:m_Z^2:m_H^2=3:4:12, \quad m_{X^-}^2=m_{X^0}^2+2m_W^2, \tag{27}$$

with the prediction $m_H=2m_W=161$ GeV. A value of the Higgs mass near 160 GeV has also been predicted by Okumura.[25]

To summarize, gauge potentials, Higgs fields, and the Higgs condensate can be accommodated in a single Hermitian 4×4 matrix,

$$\hat{A}+L=i\begin{pmatrix} 0 & 0 & \overline{X^0} & \overline{X^-} \\ 0 & \sqrt{\frac{1}{2}}A^0+\sqrt{\frac{1}{6}}Z & 0 & \sqrt{\frac{1}{2}}\phi+\mu \\ X^0 & 0 & \sqrt{\frac{2}{3}}Z & W^+ \\ X^- & \sqrt{\frac{1}{2}}\phi+\mu & W^- & \sqrt{\frac{1}{2}}A^0-\sqrt{\frac{1}{6}}Z \end{pmatrix}. \tag{28}$$

The electric charges 0, ±1 attributed to the entries of such a matrix may be read off from the scheme,

$$\begin{pmatrix} 0 & +1 & 0 & +1 \\ -1 & 0 & -1 & 0 \\ 0 & +1 & 0 & +1 \\ -1 & 0 & -1 & 0 \end{pmatrix}.$$

It should be kept in mind that we rely here on a classical approximation. Quantization changes the Higgs potential to some effective potential which is expected to considerably differ from the classical potential. The same proviso applies to the computation of masses since they also depend on the effective potential. To include loop corrections is one way to change predictions, perhaps not in a reliable way. Such corrections depend on the mass matrices of matter fields and thus are outside the scope of this paper. Another way is to apply renormalization group methods which also rely on loop calculations. It should also be kept in mind that the relation $\tan\theta=g'/g$ holds for g and g' defined on a sliding energy scale. Therefore, the Weinberg angle θ cannot be a constant over a large energy range. The values to be used here should come from energies comparable to the mass parameter μ.

The superconnection formalism can be extended to include matter fields as will be shown in a subsequent paper. Left-handed and right-handed leptons (or quarks) of one generation are modelled by the vector space $\wedge\mathbb{C}^2$ of dimension four. The $\mathbb{Z}_2$ grading of that space coincides with one given by the handedness of fields. To obtain the correct action of the electro-weak structure group, one associates the subspaces $\wedge^0\mathbb{C}^2$, $\wedge^1\mathbb{C}^2$, and $\wedge^2\mathbb{C}^2$ with ν_{eR}, the pair (ν_{eL}, e_L), and e_R, respectively (taking the first generation of leptons as an example).

It would also be desirable to push the theory further, so as to obtain a unified theory of weak, electromagnetic, and strong interactions as a gauge theory based on a larger group incorporating both the vector bosons of the electroweak theory and the gluons of QCD. It is not clear at the moment whether such an approach will give reasonable results.

ACKNOWLEDGMENTS

The author wishes to thank Stephan Korden and Thomas Strobl for numerous discussions.

[1] S. Weinberg, Rev. Mod. Phys. **92**, 515 (1980); *The Quantum Theory of Fields*, Vol. II in *Modern Applications* (Cambridge University Press, Cambridge, 1996).
[2] K. B. Marathe and G. Martucci, *The Mathematical Foundation of Gauge Theories* (North–Holland, Amsterdam, 1992).
[3] C. Nash, *Differential Topology and Quantum Field Theory* (Academic, New York, 1991).
[4] E. Witten, Phys. Rev. Lett. **38**, 121 (1977).
[5] N. Manton, Nucl. Phys. B **158**, 141 (1979).
[6] D. B. Fairlie, Phys. Lett. B **82**, 97 (1979).
[7] P. Forgács and N. Manton, Commun. Math. Phys. **72**, 15 (1980); J. Harnad, S. Shnider, and J. Tafel, Lett. Math. Phys. **4**, 107 (1980).
[8] Y. Ne'eman, Phys. Lett. B **81**, 190 (1979).
[9] Y. Ne'eman and S. Sternberg, Proc. Natl. Acad. Sci. USA **87**, 7875 (1990).
[10] Y. Ne'eman, hep-th/9708052.
[11] A. Connes, *Noncommutative Geometry* (Academic, New York, 1994).
[12] F. Scheck, The Standard Model within Noncommutative Geometry: A Comparison of Models, hep-th/9701073, and references therein.
[13] A. Connes and J. Lott, Nucl. Phys. B **18** (Proc Suppl), 29 (1990); in *Proceedings of the Cargèse Summer School 1991*, edited by J. Fröhlich *et al.* (Plenum, New York, 1992).
[14] Y. Okumura, Prog. Theor. Phys. **65**, 969 (1996).
[15] Th. Ackermann and J. Tolksdorf, hep-th/9503180; J. Tolksdorf, J. Math. Phys. **39**, 2213 (1998).
[16] S. Balakrishna, F. Gürsey, and K. C. Wali, Phys. Lett. B **254**, 430 (1991); R. Erdem, hep-th/9701067.
[17] D. Quillen, Topology **24**, 89 (1985).
[18] V. Mathai and D. Quillen, Topology **25**, 85 (1986); see also, N. Berline, E. Getzler, and M. Vergne, *Heat Kernels and Dirac Operators* (Springer, Berlin, 1996).
[19] R. Coquereaux, ''Higgs Fields and Superconnections, Lecture Notes in Physics,'' Vol. 375; R. Coquereaux, G. Esposito-Farese, and G. Vailant, Nucl. Phys. B **353**, 689 (1991).
[20] C.-Y. Lee, hep-th/9602159.
[21] H. Figueroa, J. M. Gracia-Bondía, F. Lizzi, and J. C. Várilly, hep-th/9701179.
[22] C. H. Taubes, Commun. Math. Phys. **75**, 207 (1980).
[23] M. Blau, J. Geom. Phys. **11**, 95 (1993).
[24] M. Sher, Phys. Rep. **179**, 273 (1989).
[25] Y. Okumura, hep-ph/9707350.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Momentum and spin of a particle with spin unity

C. B. van Wyk[a)]
Department of Mathematics and Applied Mathematics, University of the Orange Free State, Bloemfontein, South Africa

(Received 14 January 1998; accepted for publication 19 February 1999)

The first part of this paper deals with the problem of expressing a given boost as the product of an operator which leaves the momentum of a given particle, massive or massless, invariant and a helicity conserving operator. This is done by multiplying boosts and rotations only. In this way more information is obtained than by applying the operators to the relevant vectors. An operator with the property of conserving spin while perhaps changing momentum is developed. The second part of the paper defines a spin angular momentum operator leaving two four-vectors invariant. These vectors are interpreted as the momentum and spin vectors of a massive particle and the theory is extended to accommodate a massless particle. This theory emphasizes the close association of spin and momentum in a relativistic theory. The spin angular momentum operator of a massless particle has invariance properties which resemble some of those of the gauge invariant electromagnetic theory.
 [S0022-2488(99)02306-3]

I. INTRODUCTION

The first part of this paper deals with the problem of expressing an arbitrary boost as the product of an operator conserving the momentum of a given massive particle and of an operator conserving the helicity of that particle.[1–3]

The product of two boosts can be expressed as the product of a third boost preceded or followed by a rotation, the angle of rotation being the Wigner angle. The third boost can also be preceded by a rotation and followed by another rotation. The sum of the two angles of rotation must equal the Wigner angle. The choice of these angles introduces a flexibility into the theory which facilitates the accommodation of both the momentum and helicity conserving operators in the kinematics.

We construct a momentum conserving operator for a massive particle that can be adapted to be applicable to a massless particle. We then solve our problem also for a massless particle.

In the course of our discussions we encounter products of boosts and rotations that have physical meaning. Examples are the operator conserving momentum and the one conserving spin. We show how to distinguish in a unique way between different classes of these products.

The problem discussed in Sec. III has also been studied by other authors.[2,3] Their methods differ radically from ours. For instance the angle that we call the Wigner angle plays an important part in our methods while this angle does not occur in their discussions.

In the second part of this paper we regard the momentum conserving operator having the form of a product of a boost and its inverse with a rotation between them as the Lorentz transformation of a tensor of the second rank. In this case the tensor being transformed is the rotation operator. Proceeding along these lines we define a spin angular momentum operator for a spin 1 particle. A study of this operator reveals the interrelation between the spin and momentum of a relativistic particle.

Finally we derive a spin angular momentum operator for a massless particle. Under Lorentz transformations this operator has gauge invariant properties. The operator for a massive particle does not have these properties.

[a)]Present address: Helderberg Village 210, Private Bag X19, Somerset West, 7129, South Africa.

Almost all our calculations are in coordinate free form. By multiplying operators we obtain more information than by applying these operators to the relevant vectors. We demonstrate this in Sec. IV. Another advantage of this procedure is that some of our results are valid for particles of spin unity and spin half.

II. SOME ELEMENTARY DECOMPOSITIONS

In this section we prepare for later sections by briefly restating results derived earlier but used here in another context. The decomposition of the product of two boosts into a product of a third boost and a rotation can be written as[4,5]

$$L(\hat{\mathbf{u}},\phi_1)L(\hat{\mathbf{p}},\phi)=L(\hat{\mathbf{p}}',\phi_3)R(\hat{\mathbf{n}},\theta_W). \tag{1}$$

Here $L(\hat{\mathbf{u}},\phi_1)$ denotes the boost associated with the relative velocity $\hat{\mathbf{u}}\tanh\phi_1$ while $R(\hat{\mathbf{n}},\theta)$ denotes a rotation about the direction $\hat{\mathbf{n}}$ through the angle θ. The parameters on the right-hand side of (1) are determined in terms of those on the left by the multiplication of the 2×2 Pauli matrices. This procedure consists of writing (1) in the form

$$\begin{aligned}(\cosh(\phi_1/2)+\sinh(\phi_1/2)\sigma.\hat{\mathbf{u}})(\cosh(\phi/2)+\sinh(\phi/2)\sigma.\hat{\mathbf{p}})\\=(\cosh(\phi_3/2)+\sinh(\phi_3/2)\sigma.\hat{\mathbf{p}})(\cos(\theta_W/2)+i\sin(\theta_W/2)\sigma.\hat{\mathbf{n}})\end{aligned}$$

carrying out the multiplications and then comparing coefficients of corresponding sigmas. We find

$$\cos(\theta_W/2)\cosh(\phi_3/2)=\cosh(\phi_1/2)\cosh(\phi/2)+\sinh(\phi_1/2)\sinh(\phi/2)\hat{\mathbf{u}}.\hat{\mathbf{p}},$$

$$\hat{\mathbf{p}}'\sinh\phi_3=\hat{\mathbf{u}}[\sinh\phi_1\cosh\phi+\sinh\phi(\cosh\phi_1-1)\hat{\mathbf{u}}.\hat{\mathbf{p}}]+\hat{\mathbf{p}}\sinh\phi, \tag{2}$$

$$\sin(\theta_W/2)=\sinh(\phi_1/2)\sinh(\phi/2)|\hat{\mathbf{u}}_\wedge\hat{\mathbf{p}}|/\cosh(\phi_3/2), \tag{3}$$

$$\hat{\mathbf{n}}=\hat{\mathbf{u}}_\wedge\hat{\mathbf{p}}/|\hat{\mathbf{u}}_\wedge\hat{\mathbf{p}}|, \tag{4}$$

$$\cosh\phi_3=\cosh\phi\cosh\phi_1+\sinh\phi\sinh\phi_1\hat{\mathbf{u}}.\hat{\mathbf{p}}. \tag{5}$$

The angle θ_W is of course the Wigner angle associated with the left-hand side of (1). For reasons that will become evident later we regard p and p' as the initial and final momentum of a particle, respectively.

All the parameters on the right of (1) are uniquely determined by those on the left.

An equation like (1) introduces a hyperbolic as well as an addition triangle[5] (see Fig. 1). For the hyperbolic triangle $A'B'D'$ of (1) we have

$$\sin A'/\sinh\phi_1=\sin B'/\sinh\phi=\sin D'/\sinh\phi_3. \tag{6}$$

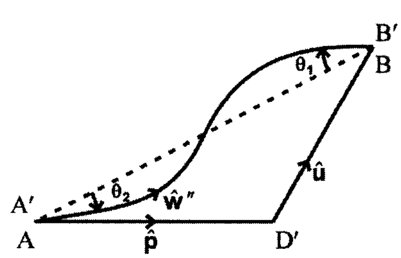

FIG. 1. The momentum diagram for (12) consists of the addition triangle ABD', the (solid) sides of which represent the momentum vectors occurring in (12). The triangle $A'B'D'$ with the dotted side $A'B'$ represents the hyperbolic triangle of this equation. In this figure we have chosen $\theta_1>0$ and $\theta_2<0$. Equation (1) has the same hyperbolic triangle as (12) but its addition triangle is obtained from Fig. 1 by replacing θ_1 by 0, θ_2 by $\theta_w>0$, and $\hat{\mathbf{w}}''$ by $\hat{\mathbf{p}}'$.

ABD' is the addition triangle of (1) and the measures of its sides and those of the hyperbolic triangle satisfy

$$A'B' = AB = \phi_3, \quad A'D' = AD' = \phi, \quad BD' = \phi_1.$$

Hence (5) holds for both the hyperbolic and the addition triangles. However the angles satisfy

$$A = A' + \theta_W, \quad B = B, \quad D' = D'. \tag{7}$$

For the hyperbolic triangle, A' is the angle between between $A'D'$ and $A'B'$. By (6)

$$\sin A' = \sinh \phi_1 |\hat{\mathbf{u}}_\wedge \hat{\mathbf{p}}| / \sinh \phi_3. \tag{8}$$

For (1) the angle between $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}'$, the directions of the initial and final momentum respectively, are given by

$$\Omega = A' + \theta_W. \tag{9}$$

where

$$\sin \Omega = Y / \sinh \phi_3, \tag{10}$$

and

$$Y = |\hat{\mathbf{u}}_\wedge \hat{\mathbf{p}}| [\sinh \phi_1 \cosh \phi + \sinh \phi (\cosh \phi_1 - 1) \hat{\mathbf{u}}.\hat{\mathbf{p}}]. \tag{11}$$

In the decomposition (12) the left-hand side is the same as in (1) but on the right we have a boost preceded and followed by a rotation[5]

$$L(\hat{\mathbf{u}}, \phi_1) L(\hat{\mathbf{p}}, \phi) = R(\hat{\mathbf{n}}, \theta_1) L(\hat{\mathbf{w}}'', \phi_3) R(\hat{\mathbf{n}}, \theta_2), \tag{12}$$

where (5) holds and

$$\sin[(\theta_1 + \theta_2)/2] = \sinh(\phi_1/2) \sinh(\phi/2) |\hat{\mathbf{u}}_\wedge \hat{\mathbf{p}}| / \cosh(\phi_3/2). \tag{13}$$

Note that (13) has the same structure as (3) whence

$$\theta_1 + \theta_2 = \theta_W \tag{14}$$

can be called the Wigner angle for decomposition (12). This decomposition can be considered as a generalization of (1). It is important to note that if the left-hand sides of (1) and (12) are the same then the same ϕ_3 and θ occur in the right-hand sides. Hence (1) and (12) have the same hyperbolic triangle $A'B'D'$ and therefore we would expect (5) to hold for (12). The angles satisfy

$$A = A' + \theta_2, \quad B = B' + \theta_1, \quad D' = D'. \tag{15}$$

The unique determination of the parameters on the right of (1) does not hold for those of (12). Equation (14) allows a choice of θ_j and this arbitrariness affects $\hat{\mathbf{w}}''$. This situation will be used in our search for the operators conserving helicity and momentum in Sec. III.

For future reference we present a formula describing the effect of a rotation on a boost. We have

$$L(\hat{\mathbf{p}}', \phi) R(\hat{\mathbf{n}}, \theta) = R(\hat{\mathbf{n}}, \theta) L(\hat{\mathbf{p}}, \phi). \tag{16}$$

For arbitrary $\hat{\mathbf{n}}$ and $\hat{\mathbf{p}}$,

$$\hat{\mathbf{p}}' = \hat{\mathbf{p}} \cos \theta + \hat{\mathbf{p}}_\wedge \hat{\mathbf{n}} \sin \theta + (1 - \cos \theta)(\hat{\mathbf{p}}.\hat{\mathbf{n}}) \hat{\mathbf{n}}.$$

If we choose $\hat{\mathbf{n}}$ and $\hat{\mathbf{p}}$ orthogonal then θ is the angle between $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}'$.

If $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}'$ are the directions of the initial and final momentum of a particle and θ is the angle between them, then the two sides of (16) are operators that leave helicity invariant and then we have two forms of the operator required in Sec. III.

The literature contains two meanings of the term Wigner angle. One group of authors[5,6] use this term for the angle denoted by θ_W in this article. Another group[2,3] use this term for the angle appearing in the usual momentum conservation operator denoted here by A'. Towards the end of Sec. III we refer to some of the properties of A'.

III. CONSERVATION OF HELICITY AND INITIAL MOMENTUM

We consider a massive particle with spin unity and momentum

$$p=(\hat{\mathbf{p}}\sinh\phi, i\cosh\phi) \tag{17}$$

and with spin along $\hat{\mathbf{p}}$ and a boost $L(\hat{\mathbf{u}},\phi_1)$. We try to factorize this operator in the form

$$L(\hat{\mathbf{u}},\phi_1)=\Lambda_h\Lambda_p, \tag{18}$$

where the operator Λ_p leaves the four-momentum (17) of the particle invariant according to

$$\Lambda_p p=p. \tag{19}$$

For the present we assume that

$$\Lambda_p=L(\hat{\mathbf{p}},\phi)R(\hat{\mathbf{n}},\theta)L^{-1}(\hat{\mathbf{p}},\phi). \tag{20}$$

Elementary considerations show that if this operator operates on a particle with momentum p the last factor reduces it to rest, R acts on the stationary particle, changing at most its spin direction while L restores the original momentum. We postpone further discussion of this momentum conservation operator to Sec. V.

The operator Λ_h leaves the helicity of the particle unchanged in the sense that if initially the spin and the momentum of the particle are parallel then after the action of Λ_h they are still parallel. Since the most general Lorentz operator can be expressed as a product of a rotation and a boost, the conservation of helicity can be achieved by an operator of the type *RL* or *LR*. We assume that

$$\Lambda_h=R(\hat{\mathbf{n}},\theta_1)L(\hat{\mathbf{p}},\phi_2). \tag{21}$$

Even if the axis $\hat{\mathbf{n}}$ and the angle θ of the rotation $R(\hat{\mathbf{n}},\theta)$ in (20) are completely arbitrary we have (19). However, if in (21) $\hat{\mathbf{n}}$ is arbitrary, then Λ_h is a very general operator containing six independent parameters.[5] Our problem can be solved by limiting $\hat{\mathbf{n}}$ in (21) to be orthogonal to $\hat{\mathbf{p}}$ in accordance with (4). Now Λ_h contains five independent parameters.

Since, according to (16), the right-hand sides of (21) and (22) are equal we can also have

$$\Lambda_h=L(\hat{\mathbf{p}}',\phi_2)R(\hat{\mathbf{n}},\theta_1). \tag{22}$$

Here $\hat{\mathbf{p}}'$ and $\hat{\mathbf{p}}$ are the vectors of Sec. II and θ_1 is the angle between them.

From (18), (20), and (21) we have

$$L(\hat{\mathbf{u}},\phi_1)=R(\hat{\mathbf{n}},\theta_1)L(\hat{\mathbf{p}},\phi_3)R(\hat{\mathbf{n}},\theta)L^{-1}(\hat{\mathbf{p}},\phi),$$

where

$$\phi_3=\phi+\phi_2. \tag{23}$$

Written in the form of (12) this equation becomes

$$L(\hat{\mathbf{u}},\phi_1)L(\hat{\mathbf{p}},\phi)=R(\hat{\mathbf{n}},\theta_1)L(\hat{\mathbf{p}},\phi_3)R(\hat{\mathbf{n}},\theta). \tag{24}$$

The only material differences between (12) and (24) are the directions of the velocities on the right-hand sides. We expect that their ϕ_3, their Wigner angles, and their hyperbolic triangles will be the same. This will be confirmed below.

We have to solve (24) for θ, θ_1, and ϕ_3 when $\hat{\mathbf{u}}$, $\hat{\mathbf{p}}$, ϕ_1, and ϕ are given. We find

$$\cosh(\phi_3/2)\cos((\theta_1+\theta)/2)=\cosh(\phi_1/2)\cosh(\phi/2)+\sinh(\phi_1/2)\sinh(\phi/2)\hat{\mathbf{u}}.\hat{\mathbf{p}}, \tag{25a}$$

$$\cosh(\phi_3/2)\sin((\theta_1+\theta)/2)=\sinh(\phi_1/2)\sinh(\phi/2)|\hat{\mathbf{u}}_{\wedge}\hat{\mathbf{p}}|, \tag{25b}$$

$$\sinh(\phi_3/2)\cos((\theta_1-\theta)/2)=\cosh(\phi_1/2)\sinh(\phi/2)+\sinh(\phi_1/2)\cosh(\phi/2)\hat{\mathbf{u}}.\hat{\mathbf{p}}, \tag{25c}$$

$$\sinh(\phi_3/2)\sin((\theta_1-\theta)/2)=\sinh(\phi_1/2)\cosh(\phi/2)|\hat{\mathbf{u}}_{\wedge}\hat{\mathbf{p}}|, \tag{25d}$$

where we have used

$$\hat{\mathbf{u}}=\hat{\mathbf{p}}(\hat{\mathbf{u}}.\hat{\mathbf{p}})-|\hat{\mathbf{u}}_{\wedge}\hat{\mathbf{p}}|\hat{\mathbf{n}}_{\wedge}\hat{\mathbf{p}} \tag{26}$$

which follows from (4). Equations (25a) and (25b) yield (5) while (25b) confirms (13) and (14) in the form

$$\theta_1+\theta=\theta_W. \tag{27}$$

Using the formulas for θ_W, A', and Ω from Sec. II, we find after some manipulation that

$$\tan(A'+\theta_W/2)=(\text{RHS of (25d)})/(\text{RHS of (25c)}) \tag{28}$$

$$=\tan((\theta_1-\theta)/2), \tag{29}$$

when considering also the left-hand sides of (25c) and (25d). Here RHS is an abbreviation for right-hand side. By (28) and (29),

$$\theta_1-\theta=2A'+\theta_W, \tag{30}$$

whence by (27) and (30),

$$\theta=-A',\quad \theta_1=A'+\theta_W=\Omega.$$

Hence (24) can be written

$$L(\hat{\mathbf{u}},\phi_1)=R(\hat{\mathbf{n}},\Omega)L(\hat{\mathbf{p}},\phi_2)L(\hat{\mathbf{p}},\phi)R(\hat{\mathbf{n}},-A')L^{-1}(\hat{\mathbf{p}},\phi), \tag{31}$$

which has the required form (18). Here we have used the notation of Sec. II and (23).

Equation (31) illustrates two points. First, the factorization (18) was achieved by using the freedom of choice of θ_j allowed by (14) to replace $\hat{\mathbf{w}}''$ in (12) by $\hat{\mathbf{p}}$. Second, the operator Λ_p contains an arbitrary parameter θ which is determined in (31) as A', an angle of an associated hyperbolic triangle. By (9) A' is linked to the Wigner angle and the angle between the relevant initial and final momentum.

The momentum diagram for (31) can be obtained from Fig. 1 by putting $\theta_2=-A'$ and $\theta_1=\Omega$ thereby satisfying (9) and (14).

The above theory enables us to replace the factorization (18) by

$$L(\hat{\mathbf{u}},\phi_1)=\Lambda_{p'}\Lambda_h,$$

where the order of the two operators on the right has been changed and the first conserves the final momentum p' of Sec. II. Using (1), (9), (16), and (23) we find in terms of the above parameters

$$L(\hat{\mathbf{u}},\phi_1)=L(\hat{\mathbf{p}}',\phi_3)R(\hat{\mathbf{n}},-A')L^{-1}(\hat{\mathbf{p}}',\phi_3)L(\hat{\mathbf{p}}',\phi_2)R(\hat{\mathbf{n}},\Omega).$$

Note that one of the factors on the right of (18) conserves momentum and the other helicity but the product does not conserve both.

IV. THE EFFECT OF A BOOST ON SPIN AND THE POINT DIAGRAM

In an effort to obtain a maximum of information we limited ourselves thus far to the multiplication of operators and tried to avoid their application to vectors. For instance, if we apply the two sides of (1) to the momentum vector of a stationary particle we only obtain (2) and (5). Since $R(\hat{\mathbf{n}},\theta)$ leaves the above-mentioned vector invariant, (3), (4) and the equation preceding (2) and therefore the Wigner angle and the rotation axis, do not appear in the calculations. When dealing with spin we have to use the spin vector.

The first topic of this section is the effect of a boost on the spin vector of a spin 1 particle. This vector is discussed in greater detail in Sec. IX. For the present we assume it to have the form

$$t=(\hat{\mathbf{t}}\cosh\phi_9,\ \ i\sinh\phi_9)$$

which is a spacelike vector with direction $\hat{\mathbf{t}}$. The parameter ϕ_9 is linked to the momentum of the particle by (32). The spin and momentum four-vectors must be orthogonal and if (17) is the momentum vector of the particle then

$$pt=\hat{\mathbf{p}}.\hat{\mathbf{t}}\cosh\phi_9\sinh\phi-\sinh\phi_9\cosh\phi=0. \tag{32}$$

When applying a boost to a four-vector we use the formula[7]

$$L(\hat{\mathbf{p}},\phi)=I+U^D\sinh\phi+(\cosh\phi-1)(U^D)^2,$$

where U^D is obtained from U of (49) by putting $\mathbf{a}=0$ and $\mathbf{b}=-i\hat{\mathbf{p}}$. This is a special case of a more general operator to be introduced in (69).

The left-hand side of (1) applied to the spin vector of a particle at rest and with spin along $\hat{\mathbf{p}}$ yields

$$L(\hat{\mathbf{u}},\phi_1)(\hat{\mathbf{p}}\cosh\phi,i\sinh\phi)=(\hat{\mathbf{t}}'\cosh\phi_5,i\sinh\phi_5),$$

where

$$\hat{\mathbf{t}}'\cosh\phi_5=\hat{\mathbf{p}}\cosh\phi+\hat{\mathbf{u}}[\sinh\phi_1\sinh\phi+(\cosh\phi_1-1)\cosh\phi(\hat{\mathbf{u}}.\hat{\mathbf{p}})],$$

$$\sinh\phi_5=\cosh\phi_1\sinh\phi+\sinh\phi_1\cosh\phi(\hat{\mathbf{u}}.\hat{\mathbf{p}}). \tag{33}$$

For the angle Ω' between $\hat{\mathbf{p}}$ and $\hat{\mathbf{t}}'$, the directions of the initial and final spin, we have by (33),

$$\sin\Omega'=|\hat{\mathbf{u}}_\wedge\hat{\mathbf{p}}|[\sinh\phi_1\sinh\phi+(\cosh\phi_1-1)\cosh\phi(\hat{\mathbf{u}}.\hat{\mathbf{p}})]/\cosh\phi_5. \tag{34}$$

By (9), Ω is the angle between the initial and final momentum in (1). Then the angle $\Omega-\Omega'$ between the final momentum $\hat{\mathbf{p}}'$ and the final spin $\mathbf{t}'$ is given by

$$\sin(\Omega-\Omega')=|\hat{\mathbf{p}}'_\wedge\hat{\mathbf{t}}'|=\sin A'/\cosh\phi_5, \tag{35}$$

where we have used (8), (10), (33), and (34). It is evident that $\Omega'<\Omega$ and that $\Omega-\Omega'<A'$.

The second topic discussed in this section is the abundance of angles generated by the equations we have studied in Secs. II and III. Almost every operator in these equations produces an angle. We distinguish between two types of angle. Angles like θ_1 and θ_2 result from the rotation

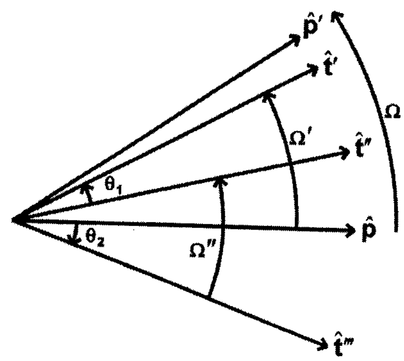

FIG. 2. This point diagram represents the effect of the right-hand side of (12) on the spin of a spin 1 particle at rest with spin along $\hat{\mathbf{p}}$. As in Fig. 1 we choose $\theta_2<0$ whence $R(\hat{\mathbf{n}},\theta_2)$ rotates the spin to the position $\hat{\mathbf{t}}'''$ ''below'' $\hat{\mathbf{p}}$. $L(\hat{\mathbf{w}}'',\phi_3)$ rotate it through Ω'' and then $R(\hat{\mathbf{n}},\theta_1)$ rotates the spin to its final position $\hat{\mathbf{t}}'$. The angles produced by rotations and boosts are indicated by small and large arcs, respectively.

factors while angles like A' and B' are caused by the boost factors. Note that the first type of angle appears in the equation while the second type does not. These ''hidden'' angles play an important part in the discussion of helicity.

We try to provide a representation of the angles which is more convenient than the momentum diagram. We introduce the point diagram in which unit vectors radiating from a common origin indicate the spin and momentum directions at each stage. To construct a point diagram like Fig. 2 we need merely consult equations, like (33).

The point diagram of the right-hand side of (12) operating on the momentum of a particle at rest can be obtained from Fig. 2. Since θ_2 does not appear, the diagram will consist merely of $\hat{\mathbf{p}}$, $\hat{\mathbf{w}}''$, and $\hat{\mathbf{p}}'$ the last two making angles $\Omega-\theta_1$ and θ, respectively, with $\hat{\mathbf{p}}$.

V. ANOTHER APPROACH TO CONSERVATION OF MOMENTUM

In Sec. III we mentioned the conservation of momentum rather superficially. We now discuss this topic in a more comprehensive way.

Since any Lorentz operator can be expressed as the product of a boost and a rotation, we look for a momentum conserving operator having the form

$$\Lambda_p = R^{-1}(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{q}},\phi_4) \tag{36}$$

and satisfying (19). Here θ_2 and ϕ_4 are real parameters to be determined later. We have

$$L(\hat{\mathbf{q}},\phi_4)p = p' = (\hat{\mathbf{p}}' \sinh\phi_6, i\cosh\phi_6),$$

where the last vector denotes an as yet unknown vector which is timelike for real ϕ_6. Since R changes the direction of the momentum vector only we must have $\phi_6=\phi$ whence by (2),

$$\hat{\mathbf{p}}' \sinh\phi = \hat{\mathbf{p}}\sinh\phi + \hat{\mathbf{q}}[\sinh\phi_4\cosh\phi + (\cosh\phi_4 - 1)\sinh\phi\,\hat{\mathbf{p}}.\hat{\mathbf{q}}], \tag{37}$$

and by (5) the unknown parameter ϕ_4 is given by

$$\sinh(\phi_4/2)/\cosh(\phi_4/2) = -(\sinh\phi/\cosh\phi)(\hat{\mathbf{p}}.\hat{\mathbf{q}}). \tag{38}$$

By (37) and (38),

$$\hat{\mathbf{p}}' = \hat{\mathbf{p}} - 2(\hat{\mathbf{p}}.\hat{\mathbf{q}})\hat{\mathbf{q}} \tag{39}$$

and the angle Ω between the initial and final momentum is given by

$$\cos\Omega = \hat{\mathbf{p}}'.\hat{\mathbf{p}} = 1 - 2(\hat{\mathbf{p}}.\hat{\mathbf{q}})^2 \quad \text{or} \quad \hat{\mathbf{p}}.\hat{\mathbf{q}} = -\sin(\Omega/2). \tag{40}$$

The invariance of p under (36) demands that R must rotate $\hat{\mathbf{p}}'$ back to $\hat{\mathbf{p}}$ through the angle

$$\theta_2 = \Omega. \tag{41}$$

This will be the case if $\hat{\mathbf{n}}$ is defined by

$$\cos(\theta_2/2)\hat{\mathbf{n}} = -\hat{\mathbf{p}}_\wedge\hat{\mathbf{q}}. \tag{42}$$

By (38), (40), and (41),

$$\cos(\theta_2/2)\cosh(\phi_4/2) < 1. \tag{43}$$

Now (36), with its parameters specified by (38), (42), and (43), is our momentum conserving operator. By (40) and (42) the vector left invariant by (36) has the form (17) with

$$\hat{\mathbf{p}} = -\sin(\theta_2/2)\hat{\mathbf{q}} + \cos(\theta_2/2)\hat{\mathbf{n}}_\wedge\hat{\mathbf{q}}. \tag{44}$$

Note that for given $\hat{\mathbf{p}}$ we can choose any $\hat{\mathbf{q}}$ which makes an obtuse angle with $\hat{\mathbf{p}}$.

Since (20) and (36) have the same meaning we must have

$$L(\hat{\mathbf{p}},\phi)R(\hat{\mathbf{n}},\theta)L^{-1}(\hat{\mathbf{p}},\phi) = R^{-1}(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{q}},\phi_4). \tag{45}$$

The axis of rotation need not be the same on the two sides of (45). However we limit ourselves to the special case where (42) holds. We find

$$\cos(\theta/2) = \cos(\theta_2/2)\cosh(\phi_4/2),$$

$$-\sin(\theta/2)\cosh\phi\,\hat{\mathbf{n}} = \sin(\theta_2/2)\cosh(\phi_4/2)\hat{\mathbf{n}}, \tag{46}$$

$$-\sin(\theta/2)\sinh\phi\,\hat{\mathbf{p}}_\wedge\hat{\mathbf{n}} = \sinh(\phi_4/2)(\cos(\theta_2/2)\hat{\mathbf{q}} + \sin(\theta_2/2)\hat{\mathbf{n}}_\wedge\hat{\mathbf{q}}), \tag{47}$$

and then (38), (40), (42), and (44) follow. The only new relations involve θ and are given by

$$\cosh^2(\phi_4/2) = \cos^2(\theta/2) + \sin^2(\theta/2)\cosh^2\phi,$$

$$\tan(\theta_2/2) = \cosh\phi\tan(\theta/2).$$

All the parameters on the left (right) of (45) are determined by those on the right (left).

Equation (45) has close links with the topics already discussed. Written in the form

$$L(\hat{\mathbf{q}},\phi_4)L(\hat{\mathbf{p}},\phi) = R(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{p}},\phi)R^{-1}(\hat{\mathbf{n}},\theta),$$

it is a special case of (12).

We conclude this approach to momentum conservation by mentioning that each of the two sides of (45) equals the operator[7]

$$\Lambda(\mathbf{a},\mathbf{b},\theta,0) = \cos(\theta/2) + i\sin(\theta/2)\sigma.(\mathbf{a} + i\mathbf{b})$$

in spinor form or in vector form

$$\Lambda(\mathbf{a},\mathbf{b},\theta,0) = \exp(iU\theta) = I + iU\sin\theta + (\cos\theta - 1)U^2. \tag{48}$$

Here

$$U=\begin{pmatrix} 0 & -ia_3 & ia_2 & b_1 \\ ia_3 & 0 & -ia_1 & b_2 \\ -ia_2 & ia_1 & 0 & b_3 \\ -b_1 & -b_2 & -b_3 & 0 \end{pmatrix}, \tag{49}$$

where

$$\mathbf{a}.\mathbf{a}-\mathbf{b}.\mathbf{b}=1, \quad \mathbf{a}.\mathbf{b}=0, \tag{50}$$

and then

$$U^3=U.$$

U has the eigenvalues 0, 0, ± 1 and the eigenvectors belonging to the zero eigenvalues are A and D, where

$$A=(\hat{\mathbf{a}},0) \quad D=(\hat{\mathbf{a}}_\wedge \mathbf{b}, ia). \tag{51}$$

The operator $\Lambda(\mathbf{a},\mathbf{b},\theta,0)$ leaves the vector

$$A \sinh\Phi + D\cosh\Phi$$

invariant. For real Φ this is a timelike vector which can be interpreted as a momentum and therefore the operator qualifies as a momentum conserving operator.

The parameters $\mathbf{a}\sin(\theta/2)$ and $\mathbf{b}\sin(\theta/2)$ of (48) for (45) are provided by (46) and (47), respectively.

We can use (36) as the momentum conserving operator to solve the problem of Sec. III. Then it takes the form

$$L(\hat{\mathbf{u}},\phi_1)=R(\hat{\mathbf{n}},\theta_1)L(\hat{\mathbf{p}},\phi_2)R^{-1}(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{q}},\phi_4). \tag{52}$$

Relations similar to (2), (5), and (10) can be derived by applying both sides of (52) to p yielding ϕ_2 and θ_1. If we apply both sides to $(-\hat{\mathbf{q}}\sinh\phi_4, i\cosh\phi_4)$ and use the properties of the hyperbolic triangle we obtain ϕ_4 and θ_2.

VI. THE CASE OF ZERO MASS

The Lorentz operator on the left of (45) which conserves the momentum of a massive particle, presupposes a transformation to the rest frame. Such a frame does not exist for a massless particle which means that we need a new momentum conservation operator for such a particle.

In our search for the desired operator we could apply the limiting process $\phi\rightarrow\infty$ to the relevant formulas of Sec. V and such a procedure could be rewarding.

However, we proceed to a more satisfactory derivation by concentrating on the operator on the right of (45) which now must leave a null vector invariant. Following the procedure of Sec. V we must have

$$L(\hat{\mathbf{q}},\phi_4)(\hat{\mathbf{p}},i)=(\hat{\mathbf{p}}',i), \tag{53}$$

where we have used the fact that $R^{-1}(\hat{\mathbf{n}},\theta_2)$ must rotate the vector on the right of (53) to the one on the left. From (53) follows that

$$\cosh\phi_4+(\hat{\mathbf{p}}.\hat{\mathbf{q}})\sinh\phi_4=1 \quad \text{or} \quad \hat{\mathbf{p}}.\hat{\mathbf{q}}=-\sinh(\phi_4/2)/\cosh(\phi_4/2) \tag{54}$$

and (39). After some calculation we find that (43) is replaced by

$$\sin(\theta_2/2)=\sinh(\phi_4/2)/\cosh(\phi_4/2) \quad \text{or} \quad \cos(\theta_2/2)\cosh(\phi_4/2)=1. \tag{55}$$

Hence the operator (36), with $\hat{\mathbf{n}}.\hat{\mathbf{q}}=0$, leaves the null vector on the left of (53) invariant provided that (55) holds.

In Sec. V we have given a meaning to the operator on the right of (36) in the case $\cos(\theta_2/2)\cosh(\phi_4/2)<1$ and in Sec. VII we shall do so for the case $\cos(\theta_2/2)\cosh(\phi_4/2)>1$.

We study another form of the operator leaving a null vector invariant.[7] Instead of (50) we now choose

$$\mathbf{a}.\mathbf{a}-\mathbf{b}.\mathbf{b}=0, \quad \mathbf{a}.\mathbf{b}=0, \tag{56}$$

whence

$$U^3=0$$

and then we have the singular operator, resembling (48),

$$\exp(iU)=I+iU-\tfrac{1}{2}U^2 \tag{57}$$

for the spin 1 case. In the case of the spinor the operator is

$$\exp[(i/2)\sigma.(\mathbf{a}+\mathbf{ib})]=I+(i/2)\sigma.(\mathbf{a}+\mathbf{ib}).$$

Hence we must have

$$R^{-1}(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{q}},\phi_4)=I+\tfrac{1}{2}i\sigma.(\mathbf{a}+\mathbf{ib}), \tag{58}$$

whence

$$\cos(\theta_2/2)\cosh(\phi_4/2)=1,$$

$$\mathbf{a}=-2\sin(\theta_2/2)\cosh(\phi_4/2)\hat{\mathbf{n}},$$

$$\mathbf{b}=-2\sinh(\phi_4/2)(\sin(\theta_2/2)\hat{\mathbf{n}}_\wedge\hat{\mathbf{q}}+\cos(\theta_2/2)\hat{\mathbf{q}}). \tag{59}$$

The operator (57) contains four independent parameters and leaves the null vector

$$p=(\hat{\mathbf{a}}_\wedge\hat{\mathbf{b}},i)=[(-\sin(\theta_2/2)\hat{\mathbf{q}}+\cos(\theta_2/2)\hat{\mathbf{n}}_\wedge\hat{\mathbf{q}}),i] \tag{60}$$

invariant. When (60) is multiplied by a suitable factor p_0, the resulting vector can be interpreted as the momentum vector of the massless particle.

Hence the operator (36) conserves momentum for a massive particle if (43) holds and for a zero mass particle if (55) holds.

Having found a general form for the operator Λ_p that leaves the momentum of a zero mass particle invariant we now consider

$$L(\hat{\mathbf{u}},\phi_1)=R(\hat{\mathbf{n}},\theta_1)L(\hat{\mathbf{p}},\phi_2)R^{-1}(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{q}},\phi_4), \tag{61}$$

where the last two factors on the right conserve the momentum of a massless particle having momentum p and the first two factors must be determined such that (61) holds for given $L(\hat{\mathbf{u}},\phi_1)$. Here p, $\hat{\mathbf{u}}$, and ϕ_1 are given while ϕ_2, ϕ_4, and θ_1 must be found.

We solve our problem by replacing the last two factors on the right of (61) by the operator on the right of (58). Here $\mathbf{a}$, $\mathbf{b}$, and $\hat{\mathbf{p}}$ are given by (59) and (60). We now have

$$L(\hat{\mathbf{u}},\phi_1)=P_0+\sigma.(\mathbf{P}+\mathbf{iQ}),$$

where

$$P_0=\cosh(\phi_1/2)=\cos(\theta_1/2)\cosh(\phi_2/2)+\sin(\theta_1/2)\sinh(\phi_4/2)(\cosh(\phi_2/2)+\sinh(\phi_2/2)), \tag{62}$$

$$\begin{aligned}\mathbf{P}&=\sinh(\phi_1/2)\hat{\mathbf{u}}\\&=\hat{\mathbf{p}}[\cos(\theta_1/2)\sinh(\phi_2/2)-\sin(\theta_1/2)(\cosh(\phi_2/2)+\sinh(\phi_2/2))\sinh(\phi_4/2)]\\&\quad+\hat{\mathbf{n}}_\wedge\hat{\mathbf{p}}[-\cos(\theta_1/2)\sinh(\phi_4/2)(\cosh(\phi_2/2)+\sinh(\phi_2/2))-\sin(\theta_1/2)\sinh(\phi_2/2)],\end{aligned} \tag{63}$$

$$\mathbf{Q}=\hat{\mathbf{n}}[\sin(\theta_1/2)\cosh(\phi_2/2)-\cos(\theta_1/2)\sinh(\phi_4/2)(\cosh(\phi_2/2)+\sinh(\phi_2/2))]=0. \tag{64}$$

Equation (64) follows because $\hat{\mathbf{u}}S_1$ the coefficient of σ in $L(\hat{\mathbf{u}},\phi_1)$ is real. We use (62), (63), and (64) supplemented by (26) to solve for θ_1, ϕ_2, and ϕ_4. We find

$$\tan(\tfrac{1}{2}\theta_1)=\sinh(\phi_1/2)|\hat{\mathbf{u}}_\wedge\hat{\mathbf{p}}|/[\cosh(\phi_1/2)+\sinh(\phi_1/2)\hat{\mathbf{u}}.\hat{\mathbf{p}}], \tag{65}$$

$$\exp(\phi_2)=\cosh\phi_1+\sinh\phi_1\hat{\mathbf{u}}.\hat{\mathbf{p}},$$

$$\sinh(\phi_4/2)=\sinh(\phi_1/2)\cosh(\phi_1/2)|\hat{\mathbf{u}}_\wedge\hat{\mathbf{p}}|/[\cosh\phi_1+\sinh\phi_1\hat{\mathbf{u}}.\hat{\mathbf{p}}]=\tan(\tfrac{1}{2}\theta_2), \tag{66}$$

where we have used (55).

Here we have solved for θ_1, θ_2, ϕ_2, and ϕ_4 in terms of p, $\hat{\mathbf{u}}$, and ϕ_1. In doing so we have regarded the product of the last two factors of (61) as a singular operator conserving the momentum null vector. We can however multiply both sides of (61) from the right by $L^{-1}(\hat{\mathbf{q}},\phi_4)$ yielding an equation of the type of (12) and with a Wigner angle

$$\theta_W=\theta_1-\theta_2$$

given by (65) and (66).

If $\phi\rightarrow\infty$ in (10) and (11) then $\Omega\rightarrow\Omega_\infty$ which is the same as θ_1 in (65) as can be expected.

VII. SPIN CONSERVATION AND THE OPERATOR $\Lambda(\mathbf{a},\mathbf{b},0,\phi)$

As was done for momentum we look for an operator that conserves the spin t of a particle with spin unity. We proceed as in Sec. V and we try to construct an operator

$$\Lambda=R^{-1}(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{q}},\phi_4)$$

resembling (36), that leaves invariant the given spacelike vector

$$t=(\hat{\mathbf{t}}\cosh\phi,i\sinh\phi).$$

The arguments used and the results obtained are almost identical to those of Sec. V. However, (38) is replaced by

$$\sinh(\phi_4/2)/\cosh(\phi_4/2)=-(\hat{\mathbf{q}}.\hat{\mathbf{t}})\cosh\phi/\sinh\phi$$

which implies that now

$$\sinh(\phi_4/2)/\cosh(\phi_4/2)>\sin(\theta_2/2)\quad\text{or}\quad\cos(\theta_2/2)\cosh(\phi_4/2)>1. \tag{67}$$

An operator having these properties has been discussed before.[7] It has the form

$$\Lambda(\mathbf{a},\mathbf{b},0,\phi)=\exp(U^D\phi)=I+U^D\sinh\phi+(\cosh\phi-1)(U^D)^2, \tag{68}$$

where, in terms of (49)

$$(U^D)_{12}=U_{34}, \quad (U^D)_{23}=U_{14}, \tag{69}$$

the other elements following by cyclic permutation of 1, 2, and 3.

The operator

$$\Lambda=L(\hat{\mathbf{p}},\phi)L(\hat{\mathbf{w}},\phi_5)L^{-1}(\hat{\mathbf{p}},\phi) \tag{70}$$

can also be expressed in the form (68) and the vectors left invariant by (70) can be found by means of the general procedure developed elsewhere.[7] However we can find, in a direct way, two linearly independent vectors by writing $\Lambda t=t$ in the form

$$L(\hat{\mathbf{w}},\phi_5)T=T, \quad T=L^{-1}(\hat{\mathbf{p}},\phi)t.$$

Now $T=(\mathbf{W},0)$ provided that $\hat{\mathbf{w}}.\mathbf{W}=0$. Then the spacelike vectors left invariant by (70) are given by

$$t=L(\hat{\mathbf{p}},\phi)T=[\mathbf{W}+(\cosh\phi-1)(\hat{\mathbf{p}}.\mathbf{W})\hat{\mathbf{p}},i(\mathbf{W}.\hat{\mathbf{p}})\sinh\phi]. \tag{71}$$

An obvious choice for $\mathbf{W}$ is $\hat{\mathbf{p}}_\wedge\hat{\mathbf{w}}$ yielding

$$t=t_1=(\hat{\mathbf{p}}_\wedge\hat{\mathbf{w}},0).$$

For the second vector we can choose $\mathbf{W}=\hat{\mathbf{p}}-(\hat{\mathbf{p}}.\hat{\mathbf{w}})\hat{\mathbf{w}}$ in (71).

In the special case where $\hat{\mathbf{w}}.\hat{\mathbf{p}}=0$ we can choose $\mathbf{W}=\hat{\mathbf{p}}$ whence

$$t=(\hat{\mathbf{p}}\cosh\phi,i\sinh\phi). \tag{72}$$

Hence if this special case applies and initially a particle has momentum (17) and spin (72) then (70) leaves the spin invariant while the momentum is transformed to $p'=(\hat{\mathbf{p}}\sinh\phi_3,i\cosh\phi_3)$.

According to (2),

$$\hat{\mathbf{p}}'\sinh\phi_3=\hat{\mathbf{w}}\sinh\phi_5+\hat{\mathbf{p}}\sinh\phi\cosh\phi_5, \quad \cosh\phi_3=\cosh\phi\cosh\phi_5,$$

and the angle Ω''' between final spin and momentum is given by

$$\sin\Omega'''=|\hat{\mathbf{p}}'_\wedge\hat{\mathbf{p}}|=\sinh\phi_5/\sinh\phi_3, \tag{73}$$

where we have used $\hat{\mathbf{w}}.\hat{\mathbf{p}}=0$.

To summarize, the operator we look for takes the general form $\Lambda(\mathbf{a},\mathbf{b},0,\phi)$ which can be expressed as (36) with (67) as well as (68) and (70).

The two operators $\Lambda(\mathbf{a},\mathbf{b},\theta,0)$ in (48) and $\Lambda(\mathbf{a},\mathbf{b},0,\phi)$ in (68), although different, have a number of corresponding properties examples of which are (43) and (67) as well as (20) and (70). Both contain five independent parameters.

VIII. A SPIN ANGULAR MOMENTUM OPERATOR

In the discussion thus far the momentum conserving operator (20) played an important part and its main properties were discussed in Secs. III and V. However this operator may also be regarded as the Lorentz transformation of the rotation operator $R(\hat{\mathbf{n}},\theta)$ when the latter is regarded as a second rank tensor.

We consider the generator for an elementary spatial rotation about the unit vector $\hat{\mathbf{m}}$ which is an antisymmetric 3×3 matrix. We generalize it to space–time to yield the 4×4 matrix,

$$J_0=\begin{pmatrix} 0 & -i\hat{m}_3 & i\hat{m}_2 & 0 \\ i\hat{m}_3 & 0 & -i\hat{m}_1 & 0 \\ -i\hat{m}_2 & i\hat{m}_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{74}$$

We interpret this as the spin angular momentum tensor for a spin 1 particle at rest. To obtain the tensor for such a particle with momentum p in a direction other than $\hat{\mathbf{m}}$ we transform from the rest frame to find

$$J=L(\hat{\mathbf{p}},\phi)J_0L^{-1}(\hat{\mathbf{p}},\phi)=U, \tag{75}$$

where U is given by (49). We can find $\mathbf{a}$ and $\mathbf{b}$ without matrix multiplication in the usual sense by using the Pauli σ-matrix procedure. We have

$$L(\hat{\mathbf{p}},\phi)(\sigma.\hat{\mathbf{m}})L^{-1}(\hat{\mathbf{p}},\phi)=\sigma.(\mathbf{a}+\mathbf{ib}),$$

where

$$\mathbf{a}=\hat{\mathbf{m}}\cosh\phi-(\cosh\phi-1)(\hat{\mathbf{m}}.\hat{\mathbf{p}})\hat{\mathbf{p}},\quad \mathbf{b}=\hat{\mathbf{p}}_\wedge\hat{\mathbf{m}}\sinh\phi, \tag{76}$$

which are of course the parameters required in (75). They satisfy (50) leaving only four independent parameters.

To confirm that we are indeed dealing with the Lorentz transformation of a six-vector we allow $\mathbf{m}$ to be complex and not necessarily a unit vector and we replace $\mathbf{a}+\mathbf{ib}$ by $\mathbf{P}$, where

$$\mathbf{P}=\mathbf{m}\cosh\phi-(\cosh\phi-1)(\mathbf{m}.\hat{\mathbf{p}})\hat{\mathbf{p}}+i\hat{\mathbf{p}}_\wedge\mathbf{m}\sinh\phi. \tag{77}$$

Now we replace $\mathbf{P}$ by $\mathbf{E}'+\mathbf{iB}'$, $\mathbf{m}$ by $\mathbf{E}+\mathbf{iB}$. Then (77) represents the well-known Lorentz transformation linking the initial components $\mathbf{E}$ and $\mathbf{B}$ of the six-vector to the final ones $\mathbf{E}'$ and $\mathbf{B}'$.

IX. THE SPIN VECTOR

In the discussion of Sec. IV we found it convenient to introduce the spin vector of a particle in an intuitive way. Now we are able to develop this vector from the spin angular momentum operator and it is closely associated with the momentum of the particle.

By (76),

$$p_0\mathbf{b}=\mathbf{p}_\wedge\mathbf{a},$$

which implies that

$$Jp=0. \tag{78}$$

J has another eigenvector belonging to the eigenvalue zero which can be chosen orthogonal to p. The form

$$t_j=(1/2)\epsilon_{jkrs}J_{kr}p_s,\quad j,k=1,2,3,4 \tag{79}$$

for this vector clearly exhibits its transformation properties. The coefficient of p_s on the right of (79) is of course U^D defined by (69) which means that a more convenient way of introducing t is provided by

$$t=U^Dp,\quad Ut=0$$

from which follows that

$$p=U^D t, \quad tt=1,$$

where we have used $U^2+(U^D)^2=I$. We find

$$t=(\mathbf{a}p_0+\mathbf{p}_\wedge\mathbf{b},i(\mathbf{a}.\mathbf{p}))=(\hat{\mathbf{t}}\cosh\phi_9,i\sinh\phi_9). \tag{80}$$

By (76) and (80) the form for t when the particle is at rest suggests that t is its spin vector. We also have

$$\hat{\mathbf{t}}\cosh\phi_9=\hat{\mathbf{m}}+(\cosh\phi-1)(\hat{\mathbf{m}}.\hat{\mathbf{p}})\hat{\mathbf{p}},$$
$$\sinh\phi_9=(\hat{\mathbf{m}}.\hat{\mathbf{p}})\sinh\phi. \tag{81}$$

This means that

$$L(\hat{\mathbf{p}},\phi)(\hat{\mathbf{m}},0)=t. \tag{82}$$

Since the four-vector on the left is the spin vector of the particle at rest, we can interpret t as its spin when the momentum is p. Hence, the momentum and spin of a particle with spin 1 are eigenvectors of its spin angular momentum operator with eigenvalues zero.

Furthermore, for arbitrary Φ the linear combinations

$$p=D\cosh\Phi+A\sinh\Phi, \quad t=D\sinh\Phi+A\cosh\Phi,$$

with A and D given by (51) are also eigenvectors of U belonging to the eigenvalue zero. This means that U or **a** and **b** do not define p and t uniquely. Hence two particles with different but compatible spin and momentum may have the same spin angular momentum operator. For example, J_0 of (74) can be the spin angular momentum operator of a particle at rest or of a particle with spin and momentum direction $\hat{\mathbf{m}}$.

Since the fourth component of A vanishes we have

$$a^2=p_0^2-t_0^2. \tag{83}$$

If in addition to **a** and **b** we also know p_0 then we have the unique expressions

$$p=(p_0/a)D+(t_0/a)A, \quad t=(p_0/a)A+(t_0/a)D, \tag{84}$$

satisfying (32). By (84),

$$\mathbf{a}=p_0\mathbf{t}-t_0\mathbf{p}, \quad \mathbf{b}=\mathbf{p}_\wedge\mathbf{t}. \tag{85}$$

If $\hat{\mathbf{m}}.\hat{\mathbf{p}}\neq 0$ then by (81) momentum and spin have the same direction for very large momentum. In the special case where $\hat{\mathbf{p}}.\hat{\mathbf{m}}=1$ we have $\mathbf{b}=0$ and $t_0=\sinh\phi$, where $p_0=\cosh\phi$. Note that here $\hat{\mathbf{p}}=\hat{\mathbf{a}}$ while p and t have the forms (17) and (72), respectively. We have often used this form for t.

If $\hat{\mathbf{m}}.\hat{\mathbf{p}}=0$ it follows from (81) and (84) that

$$t=A, \quad p=D. \tag{86}$$

Hence in this case spin and momentum are not parallel for large momentum.

In the first few sections we discussed the effect of various operators on momentum and spin by direct application of these operators to the relevant vectors. In this way we could find for example, the angle between final spin and momentum in terms of the initial angle between them as in (35). Another way of studying momentum and spin is to find the effect of these operators on the spin angular momentum tensor.

In this regard we consider the general transformation

$$(M_0+\boldsymbol{\sigma}.\mathbf{M})\sigma.\hat{\mathbf{p}}(M_0-\sigma.\mathbf{M})=\sigma.\mathbf{P}, \tag{87}$$

where

$$M_0^2-\mathbf{M}.\mathbf{M}=1.$$

We find

$$\mathbf{P}=(M_0^2+\mathbf{M}.\mathbf{M})\hat{\mathbf{p}}-2iM_0\hat{\mathbf{p}}_\wedge\mathbf{M}-2(\hat{\mathbf{p}}.\mathbf{M})\mathbf{M} \tag{88}$$

and with

$$\mathbf{M}=\boldsymbol{\alpha}+\mathbf{i}\boldsymbol{\beta},\quad \mathbf{P}=\mathbf{a}+\mathbf{ib},$$

we find

$$\begin{aligned}\mathbf{a}&=(2M_0^2-1)\hat{\mathbf{p}}+2M_0\hat{\mathbf{p}}_\wedge\boldsymbol{\beta}-2(\hat{\mathbf{p}}.\boldsymbol{\alpha})\boldsymbol{\alpha}+2(\hat{\mathbf{p}}.\boldsymbol{\beta})\boldsymbol{\beta},\\ \mathbf{b}&=-2[M_0\hat{\mathbf{p}}_\wedge\boldsymbol{\alpha}+(\hat{\mathbf{p}}.\boldsymbol{\alpha})\boldsymbol{\beta}+(\hat{\mathbf{p}}.\boldsymbol{\beta})\boldsymbol{\alpha}].\end{aligned} \tag{89}$$

With this **b** we can use (85) in the form

$$|\hat{\mathbf{p}}_\wedge\hat{\mathbf{t}}|=b/[|\mathbf{p}||t|]=b/[\{p_0^2-1\}^{1/2}\{p_0^2-b^2\}^{1/2}]. \tag{90}$$

Here we have expressed the sine of the angle between spin and momentum in terms of the parameters **a**, **b**, and p_0 which determine p and t uniquely.

We use (90) to determine the angle between final spin and momentum in a few cases where an operator acts on the spin angular momentum operator of a particle with initial spin along its momentum.

First, we consider the effect of the momentum conserving operator with $\hat{\mathbf{n}}.\hat{\mathbf{p}}=0$ on spin. In the notation of (87) we have

$$L(\hat{\mathbf{p}},\phi)R(\hat{\mathbf{n}},\theta)L^{-1}(\hat{\mathbf{p}},\phi)=M_0+\sigma.\mathbf{M},$$

where

$$M_0=\cos(\theta/2),\quad \mathbf{M}=\sin(\theta/2)(-\hat{\mathbf{p}}_\wedge\hat{\mathbf{n}}\sinh\phi+i\hat{\mathbf{n}}\cosh\phi).$$

By (89)

$$\mathbf{b}=-\hat{\mathbf{n}}\sin\theta\sinh\phi,$$

whence by (90) the angle θ' between final momentum and spin is given by

$$\sin\theta'=\sin\theta/[1+\sinh^2\phi\cos^2\theta]^{1/2}.$$

Second, Eq. (35) applies when the above particle is boosted by $L(\hat{\mathbf{u}},\phi_1)$. By using (89) and (90) we can confirm (35).

Third, we consider the spin conserving operator of Sec. VII and try to confirm (73). We have

$$L(\hat{\mathbf{p}},\phi)L(\hat{\mathbf{w}},\phi_5)L^{-1}(\hat{\mathbf{p}},\phi)=\cosh(\phi_5/2)+\sinh(\phi_5/2)\sigma.(\hat{\mathbf{w}}\cosh\phi+i\sinh\phi\hat{\mathbf{p}}_\wedge\hat{\mathbf{w}}),$$

whence

$$M_0=\cosh(\phi_5/2),\quad \mathbf{M}=\sinh(\phi_5/2)(\hat{\mathbf{w}}\cosh\phi+i\hat{\mathbf{p}}_\wedge\hat{\mathbf{w}}\sinh\phi)$$

and if $\hat{\mathbf{p}}.\hat{\mathbf{w}}=0$, then

$$\mathbf{b} = -\sinh\phi_5 \cosh\phi \hat{\mathbf{p}}_\wedge \hat{\mathbf{w}},$$

and

$$|\hat{\mathbf{p}}_\wedge \hat{\mathbf{t}}| = \sinh\phi_5 / \sinh\phi_3$$

confirming (73).

X. GAUGE INVARIANCE

Having discussed some of the important properties of the spin angular momentum operator we now turn to its invariance properties and to the properties of the relevant transformation operators. We consider (87) in the form

$$(M_0 + \sigma.\mathbf{M})\sigma.\mathbf{m}(M_0 - \sigma.\mathbf{M}) = \sigma.\mathbf{m}. \tag{91}$$

With appropriate adjustments (88) still holds for (91). Hence for given $\mathbf{m}$ to be invariant, M_0 and $\mathbf{M}$ must satisfy

$$M_0^2 + \mathbf{M}.\mathbf{M} = 1,$$
$$\mathbf{m}.\mathbf{M} = 0, \tag{92}$$
$$\mathbf{m}_\wedge \mathbf{M} = 0.$$

We ignore the trivial case $\mathbf{M} = \mathbf{m}$ and proceed to the third equation of (92). We put

$$\mathbf{m} = \mathbf{a} + \mathbf{i}\mathbf{b}, \quad \mathbf{M} = \boldsymbol{\alpha} + \mathbf{i}\boldsymbol{\beta},$$

where these four three-vectors are real and

$$\mathbf{a}.\mathbf{b} = 0 = \boldsymbol{\alpha}.\boldsymbol{\beta}.$$

Then (92) requires that

$$\mathbf{a}_\wedge \boldsymbol{\alpha} = \mathbf{b}_\wedge \boldsymbol{\beta}, \quad \mathbf{a}_\wedge \boldsymbol{\beta} = -\mathbf{b}_\wedge \boldsymbol{\alpha},$$

which means that these four vectors must be coplanar in such a way that

$$\hat{\mathbf{a}}_\wedge \hat{\mathbf{b}} = \boldsymbol{\alpha}_\wedge \boldsymbol{\beta}. \tag{93}$$

Hence

$$a^2 = b^2, \quad \alpha^2 = \beta^2, \quad \mathbf{m}.\mathbf{m} = 0, \quad \mathbf{M}.\mathbf{M} = 0, \quad M_0 = 1, \tag{94}$$

and the second equation of (92) is also satisfied.

In terms of the operators for spin 1 particles we are therefore dealing with U and U^D for which (56) holds. We denote these singular operators, for which (94) holds, by $U(\mathbf{a},\mathbf{b})$, $U^D(\alpha,\beta)$, etc.

Provided that (93) holds we have

$$[U(\mathbf{a},\mathbf{b}), U(\alpha,\beta)] = 0.$$

Hence (91) becomes

$$\Lambda(\alpha,\beta) U(\mathbf{a},\mathbf{b}) \Lambda^{-1}(\alpha,\beta) = U(\mathbf{a},\mathbf{b}), \tag{95}$$

where

$$\Lambda(\alpha,\beta)=I+iU(\alpha,\beta)-\tfrac{1}{2}U^2(\alpha,\beta), \tag{96}$$

which resembles (57).

Therefore our study of (91) leads to the following conclusions. First, both the spin angular momentum operator and the Lorentz operator under which it is invariant must be singular operators. Second, guided by (56) and (76) we regard $U(\mathbf{a},\mathbf{b})$ as the spin angular momentum operator of a massless spin 1 particle. Third, since the operators for which (50) holds do not lead to invariance, the invariance expressed by (95) does not apply to massive particles.

The operator $U(\mathbf{a},\mathbf{b})$ has four zero eigenvalues and two linearly independent eigenvectors, A as in (51) which is spacelike and

$$D'=(\hat{\mathbf{a}}_\wedge\hat{\mathbf{b}},i)$$

a null vector uniquely representing the momentum of a zero mass particle in the form

$$p=p_0D', \quad pp=0. \tag{97}$$

It is left invariant by (96).

We try to give physical meaning to A. We find

$$U(\alpha,\beta)A=-i(\alpha.\hat{\mathbf{b}})D',$$

$$\Lambda(\alpha,\beta)A=A+(\boldsymbol{\alpha}.\hat{\mathbf{b}})D'.$$

We define

$$t=A+t_0D', \quad t_0=\alpha.\hat{\mathbf{b}}, \quad tt=1, \quad pt=0. \tag{98}$$

In this linear combination of A and D' the factor $\alpha.\hat{\mathbf{b}}$ is arbitrary in the sense that α can be any vector in the plane of $\mathbf{a}$ and $\mathbf{b}$.

Despite this arbitrariness we find that for given p_0, $\mathbf{a}$, and $\mathbf{b}$, subject of course to (56),

$$p_0\hat{\mathbf{a}}=p_0\mathbf{t}-t_0\mathbf{p}, \quad p_0\hat{\mathbf{b}}=\mathbf{p}_\wedge\mathbf{t} \tag{99}$$

which resembles (85). Here the magnitudes of $\mathbf{a}$ and $\mathbf{b}$ appear to be unimportant.

Hence for given p_0 and six-vector $\mathbf{a}$, $\mathbf{b}$ we can find the three-vector $\mathbf{p}$ and the four-vector t for which (99) holds. It is important to note that t is not defined uniquely. This situation is analogous to gauge invariance in electromagnetic theory where a given six-vector $\mathbf{E}$, $\mathbf{B}$ defines the four-potential $\mathbf{A}$, ϕ in a way which is not unique.

Hence $U(\mathbf{a},\mathbf{b})$, the spin angular momentum operator of a massless spin 1 particle, has gauge invariance properties under the operator (96) which can be called the gauge operator.

To arrive at the linear combination (98) we could have used the fact that A and D' are eigenvectors of $U(\mathbf{a},\mathbf{b})$ belonging to a degenerate eigenvalue. However the procedure used above is more useful.

In electromagnetic theory no direct physical significance is attached to the four-potential. In this case the mathematics presented above seems to suggest that t can be interpreted as the spin of the massless particle. The question whether such an interpretation could be supported by the physical evidence remains open.

Since

$$[U(\mathbf{a},\mathbf{b}),U^D(-\beta,\alpha)]=0$$

provided that (93) holds, the operator

$$I+U^D(\alpha,\beta)+\tfrac{1}{2}(U^D)^2,$$

can also be regarded as a gauge operator which leaves $U(\mathbf{a},\mathbf{b})$ invariant.

The papers by Weinberg[8] and Shnerb and Horwitz[9] also deal with massless systems but are mostly concerned with field theoretical aspects.

XI. DISCUSSION

In the above discussion of products of rotations and boosts we distinguish between three different types of product which we denote by $\Lambda(\mathbf{a},\mathbf{b},0,\phi)$, $\Lambda(\mathbf{a},\mathbf{b},\theta,0)$, and $\Lambda(\mathbf{a},\mathbf{b})$. Each of the first two types contain five independent parameters while the last one contains four.

The first type includes operators like

$$L(\hat{\mathbf{u}},\phi_1)L(\hat{\mathbf{p}},\phi), L(\hat{\mathbf{p}},\phi)L(\hat{\mathbf{w}},\phi_5)L^{-1}(\hat{\mathbf{p}},\phi) \quad \text{and} \quad \exp(U^D\phi)$$

which leave two spacelike vectors invariant. In Sec. VII we have associated spin with one of these vectors.

The product on the left of

$$L(\hat{\mathbf{u}},\phi_1)L(\mathbf{v},\phi_2)=\Lambda(\mathbf{a},\mathbf{b},0,\phi)$$

contains six independent parameters with only five on the right. This means that for every set of parameters on the right there exists an infinite number of sets of parameters on the left.[5]

$\Lambda(\mathbf{a},\mathbf{b},\theta,0)$ is a general name for operators like $L(\hat{\mathbf{p}},\phi)R(\hat{\mathbf{n}},\theta)L^{-1}(\hat{\mathbf{p}},\phi)$ and $\exp(iU\theta)$ and these operators leave a timelike vector $D\cosh\Phi + A\sinh\Phi$ invariant. Under suitable circumstances this vector can be interpreted as the momentum vector of a particle.

The product on the left of

$$L(\hat{\mathbf{p}},\phi)R(\hat{\mathbf{n}},\theta)L^{-1}(\hat{\mathbf{p}},\phi)=\Lambda(\mathbf{a},\mathbf{b},\theta,0)$$

contains six independent parameters with only five on the right. Arguments similar to those in Ref. 5 show that for every set of parameters on the right we have an infinite number of sets of parameters on the left. As is well-known the operator on the left belongs to a group.

Operators of the type

$$\Lambda(\mathbf{a},\mathbf{b})=I+(i/2)\sigma.(\mathbf{a}+i\mathbf{b}), \quad \mathbf{a}.\mathbf{a}=\mathbf{b}.\mathbf{b}, \quad \mathbf{a}.\mathbf{b}=0$$

in its spinor form, or

$$\Lambda(\mathbf{a},\mathbf{b})=I+iU(\mathbf{a},\mathbf{b})-\tfrac{1}{2}U^2(\mathbf{a},\mathbf{b})$$

in its vector form, leave the spacelike vector A and the null vector D' invariant. For any λ the spacelike vector $A+\lambda D'$ is of course also left invariant. Provided that the four three-vectors $\mathbf{a}$, $\mathbf{b}$, α, and β are coplanar and (93) is satisfied we have

$$\Lambda(\mathbf{a},\mathbf{b})\Lambda(\alpha,\beta)=\Lambda(\mathbf{a}',\mathbf{b}'),$$

where

$$\mathbf{a}'=\mathbf{a}+\alpha, \quad \mathbf{b}'=\mathbf{b}+\beta.$$

Hence these operators belong to what can be called the gauge group.

We have seen that, depending on the value of $\cos(\theta/2)\cosh(\phi/2)$, the operator $R(\hat{\mathbf{n}},\theta)L(\hat{\mathbf{q}},\phi)$ with $\hat{\mathbf{n}}.\hat{\mathbf{q}}=0$ can take any of the above three forms.

The product of any two boosts always has the form $\Lambda(\mathbf{a},\mathbf{b},0,\phi)$. To classify the product of three Lorentz operators is more complicated. We can write (1) in the form

$$L^{-1}(\hat{\mathbf{p}}',\phi_3)L(\hat{\mathbf{u}},\phi_1)L(\hat{\mathbf{p}},\phi)=R(\hat{\mathbf{n}},\theta)=\Lambda(\mathbf{a},0,\theta,0).$$

Therefore with the same notation and meaning

$$L^{-1}(\hat{\mathbf{p}}',\phi_4)L(\hat{\mathbf{u}},\phi_1)L(\hat{\mathbf{p}},\phi)=L(\hat{\mathbf{p}}',\phi_3-\phi_4)R(\hat{\mathbf{n}},\theta),$$

whence, for given θ, the magnitude of $\phi_3-\phi_4$ determines the classification according to (43), (55), and (67).

If the directions of three or more boosts are independent then we must expect their product to contain more than five independent parameters.

The operator discussed in Sec. VI is also a gauge operator. When expressed in the form

$$\Lambda(\mathbf{a},\mathbf{b})=R^{-1}(\hat{\mathbf{n}},\theta_2)L(\hat{\mathbf{q}},\phi_4),\quad \cos(\theta/2)\cosh(\phi_4/2)=1,\quad \hat{\mathbf{n}}.\hat{\mathbf{q}}=0$$

the gauge operator has structure in the sense that it is expressible as a product of a rotation and a boost. The four independent parameters on the right are determined uniquely in terms of **a** and **b**. According to the statement following (99) the magnitudes of **a** and **b** seem to be unimportant. However by (59),

$$a=b=2\sinh(\phi_4/2).$$

We conclude this paper by stating that special relativity appears to make provision for an interrelation between the spin and the momentum of a particle. This statement is supported by using (85) to express the generator U^D of any $\Lambda(\mathbf{a},\mathbf{b},0,\phi)$ in the form

$$U^D=|p\rangle\langle t|-|t\rangle\langle p|,$$

where p and t are the momentum and spin, respectively, of a fictitious particle having the spin angular momentum operator U associated with U^D.

A similar form can be obtained for a massless particle.

If it is correct to interpret the t of (98) as the spin of a massless particle then, as in (86), the spin and momentum directions of a massless particle need not coincide.

ACKNOWLEDGMENT

It is a pleasure to thank the referee for a number of valuable suggestions.

[1] E. P. Wigner, Rev. Mod. Phys. **29**, 255 (1957).
[2] D. Han, Y. S. Kim, and D. Son, J. Math. Phys. **28**, 2373 (1987).
[3] D. V. Vassiliades, J. Math. Phys. **30**, 2177 (1989).
[4] C. B. van Wyk, Am. J. Phys. **52**, 853 (1984).
[5] C. B. van Wyk, J. Math. Phys. **32**, 425 (1991).
[6] N. A. Salingaros, J. Math. Phys. **29**, 1265 (1988).
[7] C. B. van Wyk, J. Math. Phys. **27**, 1306 (1986).
[8] S. Weinberg, Phys. Rev. B **138**, 998 (1965).
[9] N. Shnerb and L. P. Horwitz, J. Phys. A **27**, 3565 (1994).

Zero curvature formalism of the four-dimensional Yang–Mills theory in superspace

L. C. Q. Vilar[a)] and S. P. Sorella
UERJ, Universidade do Estado do Rio de Janeiro, Departamento de Física Teórica, Instituto de Física, Rua São Francisco Xavier, 524 20550-013, Maracanã, Rio de Janeiro, Brazil

C. A. G. Sasaki
C.B.P.F, Centro Brasileiro de Pesquisas Físicas, Rua Xavier Sigaud 150, 22290-180 Urca, Rio de Janeiro, Brazil

(Received 28 April 1998; accepted for publication 9 February 1999)

The supersymmetric descent equations in $N=1$ superspace are discussed by means of the introduction of two operators $\zeta^\alpha, \overline{\zeta}^{\dot\alpha}$, which allow to decompose the supersymmetric covariant derivatives $D^\alpha, \overline{D}^{\dot\alpha}$ as BRS commutators. © *1999 American Institute of Physics.* [S0022-2488(99)02006-X]

I. INTRODUCTION

It is well known nowadays that the problem of finding the anomalies and the invariant counterterms that arise in the renormalization of local field theories can be handled in a purely algebraic way by means of the BRS technique (For a recent account on the so-called *Algebraic Renormalization* see Ref. 1). This amounts to looking at the nontrivial solution of the integrated consistency condition,

$$s\int \omega_D^g = 0, \tag{I.1}$$

where s is the BRS operator and g and D denote, respectively, the ghost number and the space–time dimension. Condition (I.1), when translated at the nonintegrated level, yields a system of equations usually called descent equations (see Ref. 1 and references therein),

$$\begin{aligned}
&s\omega_D^g + d\omega_{D-1}^{g+1} = 0,\\
&s\omega_{D-1}^{g+1} + d\omega_{D-2}^{g+2} = 0,\\
&\cdots\\
&\cdots\\
&s\omega_1^{g+D-1} + d\omega_0^{g+D} = 0,\\
&s\omega_0^{g+D} = 0,
\end{aligned} \tag{I.2}$$

where $d = dx^\mu \partial_\mu$ is the exterior space–time derivative and $\omega_i^{g+D-i} (0 \leq i \leq D)$ are local polynomials in the fields of ghost number $(g+D-i)$ and form degree i. The cases $g=0,1$ correspond, respectively, to the invariant counterterms and to the anomalies. The operators s, d obey the algebraic relations

[a)]Electronic mail: LCQVILAR@SYMBCOMP.UERJ.BR

$$s^2=d^2=sd+ds=0. \tag{I.3}$$

The problem of solving the descent equations (I.2) is a problem of cohomology of s modulo d,[2,3] the corresponding cohomology classes being given by solutions of (I.2) that are not of the type

$$\omega_m^{g+D-m}=s\hat{\omega}_m^{g+D-m-1}+d\hat{\omega}_{m-1}^{g+D-m}, \quad 1\leq m\leq D,$$

$$\omega_0^{g+D}=s\hat{\omega}_0^{g+D-1},$$

with $\hat{\omega}$'s local polynomials. Notice also that at the nonintegrated level one loses the property of making integration by parts. This implies that the fields and their derivatives have to be considered as independent variables.

Of course, the knowledge of the most general nontrivial solution of the descent equations (I.2) yields the integrated cohomology classes of the BRS operator. Indeed, once the full system (I.2) has been solved, integration on space–time gives the general solution of the consistency condition (I.1).

Recently, a new method of obtaining nontrivial solutions of the tower (I.2) has been proposed[4] and successfully applied to a large number of field models such as Yang–Mills theories,[4,5] gravity,[6] topological field theories,[7–9] string[10] and superstring[11] theories as well as W_3 algebras.[12] The method relies on the introduction of an operator δ, which allows us to decompose the exterior derivative as a BRS commutator,

$$d=-[s,\delta]. \tag{I.4}$$

It is easily proven, in fact, that repeated applications of the operator δ on the cocycle ω_0^{g+D} that solves the last of the equations (I.2) will provide an explicit nontrivial solution for the higher cocycles ω_i^{g+D-i}.

One has to note that solving the last equation of the tower (I.2) is a problem of local cohomology instead of a modulo-d one. The former can be systematically analyzed by using several methods as, for instance, the spectral sequences technique.[13] It is also worth mentioning that in the case of the Yang–Mills-type gauge theories, the solutions of the descent equations (I.2) obtained via the decomposition (I.4) have been proven to be equivalent to those provided by the so-called *Russian Formula*.[14,15]

Another important geometrical aspect related to the existence of the operator δ is the possibility of encoding all the relevant informations concerning the BRS transformations of the fields and the solutions of the system (I.2) into a unique equation that takes the form of a generalized zero curvature condition,[16] i.e.,

$$\tilde{\mathcal{F}}=\tilde{d}\tilde{\mathcal{A}}-\tilde{\mathcal{A}}^2=0. \tag{I.5}$$

The operator $\tilde{d}$ and the generalized gauge connection $\tilde{\mathcal{A}}$ in Eq. (I.5) turn out to be, respectively, the δ transform of the BRS operator s and of the ghost field c corresponding to the Maurer–Cartan form of the underlying gauge algebra,

$$\tilde{d}=e^{\delta}se^{-\delta}, \quad \tilde{d}^2=0,$$

$$\tilde{\mathcal{A}}=e^{\delta}c.$$

As discussed in detail in Refs. 16, the zero curvature formulation allows us to obtain straightforwardly the cohomology classes of the operator $\tilde{d}$. The latters are deeply related to the solutions of the descent equations (I.2).

The BRS algebraic procedure can be easily adapted to include the case of the renormalizable $N=1$ superspace supersymmetric gauge theories in four space–time dimensions, for which a set of superspace descent equations have been established.[17–19] The solution of these equations as much as in the nonsupersymmetric case yields directly all the manifestly supersymmetric gauge anomalies as well as the manifestly supersymmetric BRS-invariant counterterms. One has to remark, however, that in the supersymmetric case both the derivation and the construction of a solution of the superspace version of the descent equations are more involved than the nonsupersymmetric case, due to the algebra of the spinorial covariant derivatives D_α and $\bar{D}_{\dot\alpha}$ and to the (anti)chirality constraints of some of the superfields characterizing the theory.

In order to have an idea of the differences between the superspace and the ordinary case, let us briefly consider the integrated superspace $N=1$ BRS consistency condition corresponding to the supersymmetric chiral U(1) Yang–Mills axial anomaly,[19]

$$s\int d^4x\, d^2\bar{\theta}\, K^0=0, \tag{I.6}$$

with K^0 a local power series in the gauge vector superfield with ghost number zero and dimension two. It can be proven[18,19] that condition (I.6) implies that the BRS variation of the integrand, i.e., sK^0, is a total derivative in superspace,

$$sK^0=\bar{D}_{\dot\alpha}\bar{K}^{1\dot\alpha}, \tag{I.7}$$

with $\bar{K}^{1\dot\alpha}$ local power series with ghost number one. [The absence of the term $D^\alpha K^1_\alpha$ in Eq. (I.7) is actually due to the chirality nature of the consistency condition (I.6).] Acting now on both sides of Eq. (I.7) with the nilpotent BRS operator s, we get

$$\bar{D}_{\dot\alpha} s\bar{K}^{1\dot\alpha}=0.$$

This equation admits a superspace solution (see Sec. V and Appendix A for details), which, as in the standard nonsuperspace case (I.2), entails a set of new conditions, which, together with the equation (I.7), gives the whole set of the superspace descent equations for the U(1) axial anomaly,[19] namely

$$\begin{aligned}
sK^0&=\bar{D}_{\dot\alpha}\bar{K}^{1\dot\alpha},\\
s\bar{K}^1_{\dot\alpha}&=(2D^\alpha\bar{D}_{\dot\alpha}+\bar{D}_{\dot\alpha}D^\alpha)K^2_\alpha,\\
sK^{2\alpha}&=D^\alpha K^3,\\
sK^3&=0,
\end{aligned} \tag{I.8}$$

with K^2_α and K^3 local power series of ghost number two and three.

From now on, the operator s in Eqs. (I.8), whose explicit form will be given later in Sec. II [see Eqs. (II.14)], refers to the BRS operator acting on the space of superfields of $N=1$ Yang–

Mills theories in superspace. Therefore, its integrated cohomology in the sectors of ghost number 1 and 0 will yield the possible gauge anomalies and gauge-invariant counterterms, which, being given in terms of superfields, are manifestly supersymmetric. In particular, the Eqs. (I.8) are the superspace analogous of the Wess–Zumino consistency condition (I.2) for the U(1) axial anomaly.

For the sake of clarity, it is worth emphasizing here that the existence of anomalies for the supersymmetry itself has been deeply investigated by Refs. 20 and 21, who showed, in fact, that the $N=1$ supersymmetric algebra with the usual content of superfields (i.e., scalar, chiral, and vector superfields) does not allow for anomalies of supersymmetry.

Our aim in this paper is to extend the previous works[16] to the case of the $N=1$ four-dimensional supersymmetric Yang–Mills theory, thus yielding a simple way of solving the superspace descent equations. This means that we will introduce two operators ζ_α and $\overline{\zeta}_{\dot\alpha}$, which, in analogy with the case of the operator δ of Eq. (I.4), allow us to decompose the supersymmetric covariant derivatives D_α and $\overline{D}_{\dot\alpha}$ as BRS commutators, according to

$$[\zeta_\alpha,s]=D_\alpha, \quad [\overline{\zeta}_{\dot\alpha},s]=\overline{D}_{\dot\alpha}, \tag{I.9}$$

with

$$D_\alpha\overline{D}_{\dot\alpha}+\overline{D}_{\dot\alpha}D_\alpha=2i\sigma^\mu_{\alpha\dot\alpha}\partial_\mu, \tag{I.10}$$

$\sigma^\mu_{\alpha\dot\alpha}$ being the Pauli matrices.

Moreover, as we shall see in Sec. IV [see Eq. (IV.39)], the decomposition (I.9) will lead to an algebraic structure that will close on shell, i.e., on the equations of motion of the standard $N=1$ Yang–Mills theory in superspace. In other words, we shall work without introducing the so-called antifields, which are known not to contribute to anomalies and nontrivial counterterms.[3]

We now have to remark that in the last few years the study of the BRS consistency condition for supersymmetric $N\geqslant 1$ models has been undertaken by several authors,[20–23] who faced and solved aspects that were still open. We mention, for instance, the results concerning the potential existence of susy anomalies in the presence of nonstandard constrained multiplets,[20,21] and the very useful possibility of performing a purely algebraic regularization-independent analysis of the susy gauge theories in a fully off-shell version of the Wess–Zumino gauge.[22,23] In particular, this last result has allowed for a simple discussion of the renormalization of models with extended supersymmetry (i.e., $N\geqslant 2$), yielding an algebraic proof of the ultraviolet finiteness of the $N=4$ gauge theories.[22]

However, although much is already known about the BRS cohomology of the $N=1$ gauge theories, we emphasize here that the decomposition formulas (I.9) will allow us to cast both the susy algebra and the supersymmetric BRS transformations into a unique equation, which, in complete analogy with the nonsupersymmetric case, takes the form of a generalized zero curvature condition. Moreover, by means of this zero curvature equation, we shall be able to derive the full set of superspace descent equations for the invariant action and for the U(1) axial anomaly from a unique equation of the type

$$\tilde{d}\tilde{\omega}=0, \tag{I.11}$$

$\tilde{\omega}$ being a suitable superspace cocycle and $\tilde{d}$ the generalized nilpotent operator entering the zero curvature condition. In addition, a modified version of Eq. (I.11) will allow us to also include the more complex case of the $N=1$ supersymmetric gauge anomaly, thus improving our understanding of the well-known nonpolynomial character of this anomaly.

The zero curvature equation and the related possibility of collecting the superspace descent equations into a unique condition represent the main results of this paper. Their relevance is due to the fact that, besides the possibility of recovering the solutions of the BRS consistency condition in a simple way, they provide an interesting pure geometrical framework in $N=1$ superspace.

The work is organized as follows. In Sec. II we introduce the general notations and we discuss the supersymmetric decomposition (I.9). Section III is devoted to the analysis of the algebraic relations entailed by the operators ζ_α and $\overline{\zeta}_{\dot\alpha}$. In Sec. IV we present the zero cuvature formulation of the superspace BRS transformations and of the descent equations corresponding to the invariant super Yang–Mills Lagrangian. In Sec. V we discuss the descent equations for the superspace version of the U(1) axial anomaly. In Sec. VI we deal with the case of the supersymmetric chiral gauge anomaly appearing in the quantum extension of the supersymmetric Slavnov–Taylor identity. In order to make the paper self-contained, in the final Appendices A, B, and C we collect a short summary of the main results concerning the Yang–Mills superspace BRS cohomology, as well as the solution of certain equations relevant for the superspace version of the descent equations.

II. GENERAL NOTATIONS AND DECOMPOSITION FORMULAS

In order to present the general algebraic setup, let us being by fixing the notations. (The superspace conventions used here are those of Ref. 24.) We shall work in a four-dimensional space–time with $N=1$ supersymmetry. The superfield content that will be used throughout is the standard set of the superfields of the pure $N=1$ super-Yang–Mills theories, i.e., the vector superfield ϕ and the gauge superconnections φ_α and $\overline{\varphi}_{\dot\alpha}$. They are defined as

$$\varphi_\alpha \equiv e^{-\phi} D_\alpha e^{\phi}, \quad \overline{\varphi}_{\dot\alpha} \equiv e^{\phi} \overline{D}_{\dot\alpha} e^{-\phi}, \tag{II.12}$$

where D_α and $\overline{D}_{\dot\alpha}$ are the usual supersymmetric derivatives:

$$\begin{gathered} \{D_\alpha, D_\beta\} = \{\overline{D}_{\dot\alpha}, \overline{D}_{\dot\beta}\} = 0, \\ D_\alpha \overline{D}_{\dot\alpha} + \overline{D}_{\dot\alpha} D_\alpha = 2i\sigma^\mu_{\alpha\dot\alpha}\partial_\mu . \end{gathered} \tag{II.13}$$

Introducing now the chiral and antichiral Faddeev–Popov ghosts c and $\overline{c}$,

$$\overline{D}_{\dot\alpha} c = D_\alpha \overline{c} = 0,$$

for the superspace nilpotent BRS transformations, one has

$$\begin{gathered} se^{\phi} = e^{\phi} c - \overline{c} e^{\phi}, \quad sc = -c^2, \quad s\overline{c} = -\overline{c}^2, \\ s\varphi_\alpha = -D_\alpha c - \{c, \varphi_\alpha\}, \quad s\overline{\varphi}_{\dot\alpha} = -\overline{D}_{\dot\alpha}\overline{c} - \{\overline{c}, \overline{\varphi}_{\dot\alpha}\}, \end{gathered} \tag{II.14}$$

and

$$\{s, D_\alpha\} = \{s, \overline{D}_{\dot\alpha}\} = 0.$$

Let us also give, for further use, the BRS transformations of the chiral and antichiral superfield strengths F_α and $\overline{F}_{\dot\alpha}$,

$$\begin{gathered} F_\alpha \equiv \overline{D}^2 \varphi_\alpha, \quad \overline{D}_{\dot\alpha} F_\alpha = 0, \\ \overline{F}_{\dot\alpha} \equiv D^2 \overline{\varphi}_{\dot\alpha}, \quad D_\alpha \overline{F}_{\dot\alpha} = 0, \\ sF_\alpha = -\{c, F_\alpha\}, \quad s\overline{F}_{\dot\alpha} = -\{\overline{c}, \overline{F}_{\dot\alpha}\}.‘ \end{gathered} \tag{II.15}$$

The quantum numbers, i.e., the dimensions, the ghost numbers, and the $\mathcal{R}$ weights of all the fields are assigned as follows (Table I).

TABLE I. Dim., ghost numb., and R weights.

	s	D_α	$\bar{D}_{\dot\alpha}$	ϕ	c	$\bar{c}$	φ_α	$\bar{\varphi}_{\dot\alpha}$	F_α	$\bar{F}_{\dot\alpha}$
dim	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{3}{2}$
N_g	1	0	0	0	1	1	0	0	0	0
$\mathcal{R}$	0	1	−1	0	0	0	1	−1	−1	1

The fields will be treated as commuting or anticommuting according to the fact that their total degree, here chosen to be the sum of the ghost number and of the spinorial indices, is even or odd. Otherwise stated, all the fields are Lie-algebra valued, the gauge group $\mathcal{G}$ being assumed to be a semisimple Lie group with anti-hermitian generators T^a.

The set of fields $(c,\bar{c},\phi,\varphi_\alpha,\bar{\varphi}_{\dot\alpha})$ and their covariant derivatives will define therefore the basic local space for studying the superspace descent equations. Let us also observe that due to the fact that D, $\bar{D}$ have dimension $\frac{1}{2}$, the number of covariant derivatives turns out to be limited by power counting requirements. For instance, as we shall see in the explicit examples considered in the next sections, the analysis of the superspace consistency condition for both the U(1) axial anomaly and the gauge anomaly requires the use of local formal power series in the variables $(c,\bar{c},\phi,\varphi_\alpha,\bar{\varphi}_{\dot\alpha})$ of dimension 2. We recall here that the nonpolynomial character of certain $N=1$ superspace expressions is due to the fact that the vector superfield ϕ is dimensionless. Finally, whenever the space time derivatives ∂_μ appear they are meant to be replaced by the covariant derivatives $D,\bar{D}$, according to the supersymmetric algebra (II.13).

Let us introduce now the two operators ζ_α and $\bar{\zeta}_{\dot\alpha}$ of ghost number -1, defined by

$$\zeta_\alpha c=\varphi_\alpha, \quad \bar{\zeta}_{\dot\alpha}\bar{c}=\bar{\varphi}_{\dot\alpha},$$
$$\zeta_\alpha\bar{c}=\bar{\zeta}_{\dot\alpha}c=\zeta_\alpha\phi=\bar{\zeta}_{\dot\alpha}\phi=0, \tag{II.16}$$
$$\zeta_\alpha\varphi_\beta=\bar{\zeta}_{\dot\alpha}\varphi_\beta=0.$$

Thus, it is almost immediate to check that they are of total degree zero and that they obey the following algebraic relations:

$$[\zeta_\alpha,s]=D_\alpha, \quad [\bar{\zeta}_{\dot\alpha},s]=\bar{D}_{\dot\alpha},$$
$$[\zeta_\alpha,\zeta_\beta]=[\zeta_\alpha,\bar{\zeta}_{\dot\beta}]=[\bar{\zeta}_{\dot\alpha},\bar{\zeta}_{\dot\beta}]=0, \tag{II.17}$$

yielding then the supersymmetric decomposition (I.9) we are looking for. As we shall see later on, the operators ζ_α and $\bar{\zeta}_{\dot\alpha}$ will turn out to be very useful in order to solve the superspace descent equations. Let us focus, for the time being, on the analysis of the consequences stemming from the equations (II.17).

III. ALGEBRAIC RELATIONS

To study the algebra entailed by the two operators ζ_α and $\bar{\zeta}_{\dot\alpha}$, let us first observe that they do not commute with the supersymmetric covariant derivatives D, $\bar{D}$. Instead, as one can easily check by using the equations (11.16) we have, in complete analogy with the nonsupersymmetric case,[4]

$$[\bar{\zeta}_{\dot\beta},D_\alpha]=[\zeta_\alpha,\bar{D}_{\dot\beta}]=-G_{\alpha\dot\beta}, \tag{III.18}$$

$$[\bar{\zeta}_{\dot\alpha},\bar{D}_{\dot\beta}]=[\zeta_\alpha,D_\beta]=0, \tag{III.19}$$

where the new operator $G_{\alpha\dot\beta}$ has negative ghost number -1 and acts on the fields as

TABLE II. Dim., ghost numb., and R weights.

	ζ^{α}	$\overline{\zeta}^{\dot{\alpha}}$	$G^{\alpha\dot{\alpha}}$	R^{α}	$\overline{R}^{\dot{\alpha}}$
Dim	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{3}{2}$	$\frac{3}{2}$
N_g	-1	-1	-1	-1	-1
$\mathcal{R}$	1	-1	0	-1	1

$$G_{\alpha\dot{\alpha}}c=\overline{D}_{\dot{\alpha}}\varphi_{\alpha}, \quad G_{\alpha\dot{\alpha}}\overline{c}=D_{\alpha}\overline{\varphi}_{\dot{\alpha}},$$
$$G_{\alpha\dot{\alpha}}\phi=G_{\alpha\dot{\alpha}}\varphi_{\beta}=G_{\alpha\dot{\alpha}}\overline{\varphi}_{\dot{\beta}}=0, \tag{III.20}$$

and

$$\{G_{\alpha\dot{\alpha}},s\}=\{D_{\alpha},\overline{D}_{\dot{\alpha}}\},$$
$$[\zeta_{\alpha},G_{\beta\dot{\beta}}]=[\overline{\zeta}_{\dot{\alpha}},G_{\beta\dot{\beta}}]=\{G_{\alpha\dot{\alpha}},G_{\beta\dot{\beta}}\}=0. \tag{III.21}$$

Again, the operator $G_{\alpha\dot{\beta}}$ does not anticommute with the covariant derivatives $D,\overline{D}$. It yields, in fact,

$$\{G_{\alpha\dot{\alpha}},D_{\beta}\}=-\tfrac{1}{2}\epsilon_{\alpha\beta}\overline{R}_{\dot{\alpha}}, \quad \{G_{\alpha\dot{\alpha}},\overline{D}_{\dot{\beta}}\}=\tfrac{1}{2}\epsilon_{\dot{\alpha}\dot{\beta}}R_{\alpha}, \tag{III.22}$$

with R_{α} and $\overline{R}_{\dot{\alpha}}$ of ghost number -1 and defined as

$$R_{\alpha}c=F_{\alpha}, \quad \overline{R}_{\dot{\alpha}}\overline{c}=\overline{F}_{\dot{\alpha}},$$
$$R_{\alpha}\overline{c}=2\overline{D}_{\dot{\alpha}}D_{\alpha}\overline{\varphi}^{\dot{\alpha}}+D_{\alpha}\overline{D}_{\dot{\alpha}}\overline{\varphi}^{\dot{\alpha}}+(D_{\alpha}\overline{\varphi}_{\dot{\alpha}})\overline{\varphi}^{\dot{\alpha}}+\overline{\varphi}_{\dot{\alpha}}(D_{\alpha}\overline{\varphi}^{\dot{\alpha}}),$$
$$\overline{R}_{\dot{\alpha}}c=2D^{\alpha}\overline{D}_{\dot{\alpha}}\varphi_{\alpha}+\overline{D}_{\dot{\alpha}}D^{\alpha}\varphi_{\alpha}+(\overline{D}_{\dot{\alpha}}\varphi^{\alpha})\varphi_{\alpha}+\varphi^{\alpha}(\overline{D}_{\dot{\alpha}}\varphi_{\alpha}), \tag{III.23}$$
$$R_{\alpha}\phi=R_{\alpha}\varphi_{\beta}=R_{\alpha}\overline{\varphi}_{\dot{\beta}}=R_{\alpha}F_{\beta}=R_{\alpha}\overline{F}_{\dot{\beta}}=0,$$
$$\overline{R}_{\dot{\alpha}}\phi=\overline{R}_{\dot{\alpha}}\varphi_{\beta}=\overline{R}_{\dot{\alpha}}\overline{\varphi}_{\dot{\beta}}=\overline{R}_{\dot{\alpha}}F_{\beta}=\overline{R}_{\dot{\alpha}}\overline{F}_{\dot{\beta}}=0.$$

In addition, we have

$$[R_{\alpha},s]=[R_{\alpha},D_{\beta}]=[R_{\alpha},\overline{D}_{\dot{\beta}}]=[R_{\alpha},G_{\alpha\dot{\beta}}]=0,$$
$$[R_{\alpha},\zeta_{\beta}]=[R_{\alpha},\overline{\zeta}_{\dot{\beta}}]=[R_{\alpha},R_{\beta}]=[R_{\alpha},\overline{R}_{\dot{\beta}}]=0. \tag{III.24}$$

Let us display the quantum numbers of the operators entering the algebraic relations (II.17), (III.18), (III.22) (Table II).

In the next section, it will be shown how the operators in the Table II can be combined into a unique generalized operator by means of the introduction of a set of global parameters. These parameters will be required to fulfill a certain number of suitable conditions [see Eqs. (IV.26) in the next section], which will project the algebra (II.17), (III.18), (III.22) on the equations of motion of $N=1$ super-Yang–Mills, and will allow us to cast the BRS transformations of the superfields in the form of a zero curvature condition in superspace. In particular, the introduction of the aforementioned global parameters will have the effect of realizing the decomposition (II.17) on all the elementary superfields $(c,\overline{c},\phi,\varphi_{\alpha},\overline{\varphi}_{\dot{\alpha}})$ and their covariant derivatives on shell, i.e., modulo the equations of motion. Therefore, we can assume as the basic functional space for the forthcoming analysis that of the polynomials in the elementary superfields and their covariant

TABLE III. Dim., ghost numb., and *R* weights.

	e^{α}	$\overline{e^{\dot\alpha}}$	$\tilde{e}^{\alpha\dot\alpha}$
dim	$-\frac{1}{2}$	$-\frac{1}{2}$	-1
N_g	1	1	1
$\mathcal{R}$	0	0	0

derivatives bounded by dimension two. This limitation comes directly from superspace power-counting considerations. Moreover, all possible nontrivial counterterms and anomalies allowed by the power counting will be included.

IV. THE ZERO CURVATURE CONDITION

Having characterized all the relevant operators entailed by the consistency of the supersymmetric decomposition (II.17), let us pay attention to the geometrical aspects of the algebraic relations so far obtained. To this purpose it is useful to introduce a set of global parameters e^{α}, $\overline{e^{\dot\alpha}}$, and $\tilde{e}^{\alpha\dot\alpha}$, naturally associated to the operators ζ_{α}, $\overline{\zeta}_{\dot\alpha}$, and $G_{\alpha\dot\beta}$, of ghost number one, and obeying the relations (Table III)

$$e^{\alpha}e^{\beta}=\overline{e^{\dot\alpha}}\,\overline{e^{\dot\beta}}=\tilde{e}^{\alpha\dot\alpha}\tilde{e}^{\beta\dot\beta}=0,$$
$$[e^{\alpha},\overline{e^{\dot\beta}}]=[e^{\alpha},\tilde{e}^{\beta\dot\beta}]=[\tilde{e}^{\alpha\dot\alpha},\overline{e^{\dot\beta}}]=0. \tag{IV.25}$$

In addition, the global parameters $(e^{\alpha},\overline{e^{\dot\alpha}},\tilde{e}^{\alpha\dot\alpha})$ will be required to obey the following conditions (see also Appendix D):

$$e^{\alpha}\tilde{e}^{\beta\dot\alpha}=-\tfrac{1}{2}\epsilon^{\alpha\beta}e^{\gamma}\tilde{e}_{\gamma}^{\ \dot\alpha},\qquad \tilde{e}^{\alpha\dot\alpha}\overline{e^{\dot\beta}}=\tfrac{1}{2}\epsilon^{\dot\alpha\dot\beta}\tilde{e}^{\alpha}_{\ \dot\gamma}\overline{e^{\dot\gamma}},$$
$$e^{\alpha}\tilde{e}^{\beta\dot\alpha}\overline{e^{\dot\beta}}=-\tfrac{1}{4}\epsilon^{\alpha\beta}\epsilon^{\dot\alpha\dot\beta}e^{\gamma}\tilde{e}_{\gamma\dot\gamma}\tilde{e}^{\dot\gamma}, \tag{IV.26}$$

fixing the symmetry properties of the product of two parameters with respect to their spinorial indices. Defining now the nilpotent dimensionless operators $\zeta,\overline{\zeta}$, and G as

$$\zeta=\zeta^{\alpha}e_{\alpha},\qquad \overline{\zeta}=\overline{\zeta}_{\dot\alpha}\overline{e^{\dot\alpha}},\qquad G=G^{\alpha}_{\dot\alpha}\tilde{e}^{\dot\alpha}_{\ \alpha},$$

it is straightforward to verify that they have zero ghost number and $\mathcal{R}$ weight, respectively, 1, -1, 0, and that the subalgebra generated by ζ_{α}, $\overline{\zeta}_{\dot\alpha}$, and $G_{\alpha\dot\beta}$, i.e.,

$$[\zeta_{\alpha},\zeta_{\beta}]=[\zeta_{\alpha},\overline{\zeta}_{\dot\beta}]=[\overline{\zeta}_{\dot\alpha},\overline{\zeta}_{\dot\beta}]=0,$$

$$[\zeta_{\alpha},G_{\beta\dot\beta}]=[\overline{\zeta}_{\dot\alpha},G_{\beta\dot\beta}]=\{G_{\alpha\dot\alpha},G_{\beta\dot\beta}\}=0,$$

can be simply rewritten as

$$[\zeta,\overline{\beta}]=[\zeta,G]=[\overline{\zeta},G]=0.$$

Analogously, introducing the nilpotent operators $\tilde{G},D,\bar{D},R,\bar{R},\partial,\tilde{\partial}$

$$
\begin{aligned}
&\tilde{G}=G^{\alpha}_{\dot{\alpha}}e_{\alpha}\bar{e}^{\dot{\alpha}}, \quad G=G^{\alpha}_{\dot{\alpha}}\tilde{e}^{\dot{\alpha}}_{\alpha},\\
&D=D^{\alpha}e_{\alpha}, \quad \bar{D}=\bar{D}_{\dot{\alpha}}\bar{e}^{\dot{\alpha}},\\
&R=R^{\alpha}\tilde{e}_{\alpha\dot{\alpha}}\bar{e}^{\dot{\alpha}}, \quad \bar{R}=\bar{R}_{\dot{\alpha}}e^{\alpha}\tilde{e}^{\dot{\alpha}}_{\alpha},\\
&\tilde{\partial}=\{D^{\alpha},\bar{D}_{\dot{\alpha}}\}e_{\alpha}\bar{e}^{\dot{\alpha}}, \quad \partial=\{D^{\alpha},\bar{D}_{\dot{\alpha}}\}\tilde{e}^{\dot{\alpha}}_{\alpha},
\end{aligned}
\tag{IV.27}
$$

it is immediate to check that all the algebraic relations and field transformations of eqs. (II.16)–eqs. (III.24) may be cast into the following free index notation

$$
\begin{aligned}
&[\zeta,s]=D, \quad [\bar{\zeta},s]=\bar{D}, \quad \{\tilde{G},s\}=\tilde{\partial}, \quad [G,s]=\partial,\\
&\{s,D\}=0, \quad \{s,\bar{D}\}=0, \quad [s,\tilde{\partial}]=0, \quad \{s,\partial\}=0,\\
&[D,\tilde{\partial}]=0, \quad [\bar{D},\tilde{\partial}]=0, \quad \{D,\partial\}=0, \quad \{\bar{D},\partial\}=0,\\
&[D,\zeta]=0, \quad [\bar{D},\bar{\zeta}]=0, \quad [\bar{D},\zeta]=\tilde{G}, \quad [D,\bar{\zeta}]=\tilde{G},\\
&[\partial,\tilde{\partial}]=0, \quad [G,\tilde{G}]=0, \quad [\zeta,\tilde{G}]=0, \quad [\bar{\zeta},\tilde{G}]=0,\\
&[G,\partial]=0, \quad [\tilde{G},\tilde{\partial}]=0, \quad [G,\tilde{\partial}]=0, \quad \{\tilde{G},\partial\}=0,\\
&\{\tilde{G},D\}=0, \quad \{\tilde{G},\bar{D}\}=0, \quad 2[D,G]=\bar{R}, \quad 2[G,\bar{D}]=R,\\
&[\zeta,\tilde{\partial}]=0, \quad [\bar{\zeta},\tilde{\partial}]=0, \quad 2[\zeta,\partial]=\bar{R}, \quad 2[\partial,\bar{\zeta}]=R,\\
&[\zeta,R]=0, \quad [\zeta,\bar{R}]=0, \quad [\bar{\zeta},R]=0, \quad [\bar{\zeta},\bar{R}]=0,\\
&[R,\tilde{\partial}]=0, \quad [\bar{R},\tilde{\partial}]=0, \quad \{R,\partial\}=0, \quad \{\bar{R},\partial\}=0,\\
&\{D,R\}=0, \quad \{\bar{D},R\}=0, \quad \{D,\bar{R}\}=0, \quad \{\bar{D},\bar{R}\}=0,\\
&\{\tilde{G},R\}=0, \quad \{\tilde{G},\bar{R}\}=0, \quad [G,R]=0, \quad [G,\bar{R}]=0,\\
&\{s,R\}=0, \quad \{s,\bar{R}\}=0, \quad \{R,\bar{R}\}=0.
\end{aligned}
\tag{IV.28}
$$

Let us proceed now by showing that, as announced in the Introduction, the supersymmetric BRS transformations (II.14), (II.15) can be obtained by means of a generalized zero curvature condition. To this aim let us introduce the operator δ,

$$\delta=\zeta+\bar{\zeta}-G, \tag{IV.29}$$

from which one easily obtains the following decomposition:

$$[s,\delta]=-D-\bar{D}-\partial.$$

Defining now the δ transform of the BRS operator s as

$$\tilde{d}=e^{\delta}se^{-\delta}, \tag{IV.30}$$

one gets

$$\tilde{d} = s + D + \bar{D} + \partial - \tilde{G} + \tfrac{1}{2}\bar{R} - \tfrac{1}{2}R,$$
$$\tilde{d}\tilde{d} = 0, \tag{IV.31}$$

so that, calling $\tilde{A}$ and $\tilde{\bar{A}}$ the δ transform of the chiral and antichiral ghosts $(c,\bar{c})$,

$$\tilde{A} = e^{\delta}c = c + \varphi + \bar{D}\varphi, \quad \varphi = \varphi^{\alpha}e_{\alpha}, \tag{IV.32}$$

$$\tilde{\bar{A}} = e^{\delta}\bar{c} = \bar{c} + \bar{\varphi} + D\bar{\varphi}, \quad \bar{\varphi} = \bar{\varphi}_{\dot{\alpha}}\bar{e}^{\dot{\alpha}}, \tag{IV.33}$$

it follows that the BRS transformations of $(c,\bar{c})$ imply the zero curvature equations,

$$e^{\delta}se^{-\delta}e^{\delta}c = -e^{\delta}c^{2} \Rightarrow \tilde{d}\tilde{A} + \tilde{A}^{2} = 0 \tag{IV.34}$$

and

$$e^{\delta}se^{-\delta}e^{\delta}\bar{c} = -e^{\delta}\bar{c}^{2} \Rightarrow \tilde{d}\tilde{\bar{A}} + \tilde{\bar{A}}^{2} = 0. \tag{IV.35}$$

Equations (IV.34) and (IV.35) are easily checked to reproduce all the BRS transformations (II.14), (II.15), as well as the whole set of the equations (III.20)–(III.23). One sees, thus, that, in complete analogy with the nonsupersymmetric case,[16] the zero curvature equations (IV.34) and (IV.35) deeply rely on the existence of the operators ζ_{α} and $\bar{\zeta}_{\dot{\alpha}}$. Let us underline here that the nilpotent operator $\tilde{d}$ in Eq. (IV.31) will play a rather important role in the discussion of the superspace descent equations. For instance, as we shall see explicitly in the example given in the next section, it turns out that the superspace descent equations corresponding to the BRS-invariant counterterms can be remarkably obtained from the single equation,

$$\tilde{d}\tilde{\omega} = 0, \tag{IV.36}$$

where $\tilde{\omega}$ is an appropriate cocycle of dimension zero and ghost number three, whose components are the superspace field polynomials of the Taylor expansion of $\tilde{\omega}$ in the global parameters $(e^{\alpha},\bar{e}^{\dot{\alpha}},\tilde{e}^{\alpha\dot{\alpha}})$. Equation (IV.36) can also be applied to characterize the descent equations of the U(1) anomaly. In Sec. VI we shall see that a slight modification of Eq. (IV.36) will allow us to treat the case of the Yang–Mills gauge anomaly, as well. In all these cases the components of $\tilde{\omega}$ will not exceed dimension two, this dimension being taken as the upper limit of our superspace analysis of the descent equations. In other words, in what follows we shall limit ourselves to the study of the solutions of the superspace descent equations in the space of local functionals with dimension less or equal to two. In particular, according to Table I, this implies that the maximum number of covariant derivatives $D,\bar{D}$ present in each component of $\tilde{\omega}$ is four.

Let us conclude this section with the following important remark. Being interested in the descent equations involving superspace functionals of dimension less or equal to two, we should have checked the closure of the algebra (IV.28) built up by the operators $(s,\zeta,\bar{\zeta},G,\tilde{G},D,\bar{D},R,\bar{R},\partial,\tilde{\partial})$ on all the fields and their covariant derivatives up to reaching dimension two. It is not difficult to convince oneself that actually there is a breakdown of the closure of this algebra in the highest level of dimension two. However, as it usually happens in supersymmetry, the breaking terms turn out to be nothing but the equations of motion corresponding to the pure $N=1$ susy Yang–Mills action, thus implying an on-shell closure of the algebra. Evaluating, in fact, the commutator between the operators ζ and s on the superfield strength F, one gets

$$[\zeta,s]F = -[\varphi,F]. \tag{IV.37}$$

TABLE IV. Dim. and ghost numb.

	ω^3	$\omega^{2\alpha}$	$\overline{\omega}^2_{\dot\alpha}$	$\tilde\omega^{2\alpha}_{\dot\alpha}$	$\tilde\omega^{1\alpha}_{\dot\alpha}$	$\omega^{1\alpha}$	$\overline{\omega}^1_{\dot\alpha}$	ω^0
dim	0	$\frac{1}{2}$	$\frac{1}{2}$	1	1	$\frac{3}{2}$	$\frac{3}{2}$	2
N_g	3	2	2	2	1	1	1	0

The right-hand side of the equation (IV.37) can be rewritten as

$$[\zeta,s]F=DF-(DF+[\varphi,F]),$$

so that, recalling that

$$DF+[\varphi,F]=-\tfrac{1}{2}e^{\gamma}\tilde{e}_{\gamma\dot\gamma}\overline{e}^{\dot\gamma}(D^{\alpha}F_{\alpha}+\{\varphi^{\alpha},F_{\alpha}\})=0, \tag{IV.38}$$

are precisely the equations of motion of the pure $N=1$ susy Yang–Mills action, one obtains

$$[\zeta,s]F=DF-\text{equation of motion}. \tag{IV.39}$$

It is worth underlining here that the on-shell closure of the algebra relies precisely on the introduction of the global parameters $e^{\alpha},\overline{e}^{\dot\alpha},\tilde{e}^{\alpha\dot\alpha}$ and on the relations (IV.26). Nevertheless, this on-shell closure does not represent a real obstruction in order to solve the superspace consistency conditions. In fact, from Eq. (IV.38) and from Table I, one can observe that the equations of motion of $N=1$ super-Yang–Mills are of dimension two. Therefore, they could eventually contribute only to the highest level of the descent equations. Rather, the above on-shell closure (IV.28) is related to the absence of the so-called BRS external fields (i.e., the Batalin–Vilkoviski antifields), which are known to properly take care of the equations of motion. However, as shown by Refs. 17, 18, and 24, these external fields do not contribute to the superspace BRS cohomology in the cases considered here of the U(1) chiral anomaly, of the gauge anomaly as well as of the invariant counterterms. This is the reason why we have discarded them. In Appendix C it will be shown how the introduction of an appropriate external field takes care in a simple way of the Yang–Mills equations of motion, thus closing the algebra off shell.

A. Nonchiral descent equations for the invariant action

In order to apply the supersymmetric decomposition (II.17) to the analysis of the superspace descent equations, let us begin by considering the BRS consistency condition corresponding to the nonchiral Yang–Mills invariant action, i.e.,

$$s\int d^4x\, d^2\theta\, d^2\overline{\theta}\,\mathcal{L}^0=0\Rightarrow s\mathcal{L}^0=D^{\alpha}\mathcal{L}^1_{\alpha}+\overline{D}_{\dot\alpha}\mathcal{L}^{1\dot\alpha}, \tag{IV.40}$$

where $\mathcal{L}^0$ is a local power series of dimension two and ghost number zero. According to what was mentioned in the previous section, the full set of the superspace descent equations characterizing $\mathcal{L}^0$ can be obtained directly from the generalized equation,

$$\tilde{d}\tilde{\omega}=0, \tag{IV.41}$$

with $\tilde\omega$ a generalized cocycle of ghost number three and dimension zero, whose Taylor expansion in the global parameters $(e^{\alpha},\tilde{e}^{\alpha\dot\alpha},\overline{e}^{\dot\alpha})$ reads as

$$\tilde\omega=\omega^3+\omega^{2\alpha}e_{\alpha}+\overline{\omega}^2_{\dot\alpha}\overline{e}^{\dot\alpha}+\tilde\omega^{2\alpha}_{\dot\alpha}\tilde{e}^{\dot\alpha}_{\alpha}+\tilde\omega^{1\alpha}_{\dot\alpha}e_{\alpha}\overline{e}^{\dot\alpha}+\omega^{1\alpha}\tilde{e}_{\alpha\dot\alpha}\overline{e}^{\dot\alpha}+\overline{\omega}^1_{\dot\alpha}e^{\alpha}\tilde{e}^{\dot\alpha}_{\alpha}+\omega^0e^{\alpha}\tilde{e}_{\alpha\dot\alpha}\overline{e}^{\dot\alpha}. \tag{IV.42}$$

The coefficients $(\omega^3,\omega^{2\alpha},\overline{\omega}^2_{\dot\alpha},\tilde{\omega}^{2\alpha}_{\dot\alpha},\tilde{\omega}^{1\alpha}_{\dot\alpha},\omega^{1\alpha},\overline{\omega}^1_{\dot\alpha},\omega^0)$ are local power series in the superfields with the following quantum numbers (Table IV).

In particular, one observes that the coefficient ω^0 in the expression (IV.42) has the same dimension of the invariant action we are looking for, thus justifying the choice of the quantum numbers of $\tilde{\omega}$ in Eq. (IV.41).

The generalized condition (IV.41) is easily worked out and yields the following set of equations:

$$
\begin{gathered}
s\omega^0=-\tfrac{1}{2}D^\alpha\omega^1_\alpha+\tfrac{1}{2}\overline{D}_{\dot\alpha}\overline{\omega}^{1\dot\alpha}+\tfrac{1}{4}\overline{R}_{\dot\alpha}\overline{\omega}^{2\dot\alpha}+\tfrac{1}{4}R^\alpha\omega^2_\alpha-\tfrac{1}{4}\{D^\alpha,\overline{D}_{\dot\alpha}\}\tilde{\omega}^{1\dot\alpha}_\alpha-\tfrac{1}{4}G^{\alpha}_{\dot\alpha}\tilde{\omega}^{2\dot\alpha}_\alpha,\\
s\overline{\omega}^1_{\dot\alpha}=-\tfrac{1}{2}\{D^\alpha,\overline{D}_{\dot\alpha}\}\omega^2_\alpha-\tfrac{1}{2}D^\alpha\tilde{\omega}^2_{\alpha\dot\alpha}+\tfrac{1}{2}\overline{R}_{\dot\alpha}\omega^3,\\
s\tilde{\omega}^{1\alpha}_{\dot\alpha}=-D^\alpha\overline{\omega}^2_{\dot\alpha}-\overline{D}_{\dot\alpha}\omega^{2\alpha}+G^{\alpha}_{\dot\alpha}\omega^3,\quad s\omega^{1\alpha}=\tfrac{1}{2}\{D^\alpha,\overline{D}_{\dot\alpha}\}\overline{\omega}^{2\dot\alpha}+\tfrac{1}{2}\overline{D}_{\dot\alpha}\tilde{\omega}^{2\alpha\dot\alpha}-\tfrac{1}{2}R^\alpha\omega^3,\\
s\overline{\omega}^2_{\dot\alpha}=-\overline{D}_{\dot\alpha}\omega^3,\quad s\tilde{\omega}^{2\alpha}_{\dot\alpha}=\{D^\alpha,\overline{D}_{\dot\alpha}\}\omega^3,\quad s\omega^{2\alpha}=-D^\alpha\omega^3,\quad s\omega^3=0.
\end{gathered}
\tag{IV.43}
$$

These equations do not yet represent the final version of the superspace descent equations, due to the presence of the operators $(G_{\alpha\dot\alpha},R_\alpha,\overline{R}_{\dot\alpha})$ on their right-hand sides. However, we shall prove that these undesired terms can be rewritten as pure BRS cocycles or as total superspace derivatives, meaning that they can be eliminated by means of a redefinition of the ω's cocycles entering the equations (IV.43). Let us first observe that a particular solution of the tower (IV.43) can be fully expressed in terms of the BRS invariant cocycle ω^3. In fact, owing to the zero curvature equations (IV.30), (IV.34), and (IV.35), it is apparent that the system (IV.43) is solved by

$$\tilde{\omega}=e^\delta\omega^3, \tag{IV.44}$$

which, when written in components, yields the following expressions:

$$
\begin{gathered}
\omega^{2\alpha}=\zeta^\alpha\omega^3,\quad \tilde{\omega}^{2\alpha}_{\dot\alpha}=G^{\alpha}_{\dot\alpha}\omega^3,\\
\overline{\omega}^2_{\dot\alpha}=\overline{\zeta}_{\dot\alpha}\omega^3,\quad \omega^{1\alpha}=\tfrac{1}{2}G^{\alpha}_{\dot\alpha}\overline{\zeta}^{\dot\alpha}\omega^3,\\
\tilde{\omega}^{1\alpha}_{\dot\alpha}=\zeta^\alpha\overline{\zeta}_{\dot\alpha}\omega^3,\quad \overline{\omega}^1_{\dot\alpha}=-\tfrac{1}{2}G^{\alpha}_{\dot\alpha}\zeta_\alpha\omega^3,\\
\omega^0=\tfrac{1}{4}\zeta^\alpha G_{\alpha\dot\alpha}\overline{\zeta}^{\dot\alpha}\omega^3.
\end{gathered}
\tag{IV.45}
$$

In particular, from the results on the superspace BRS cohomology[18,19,25] (see Appendix B), it turns out that the most general form for ω^3 can be identified with the invariant ghost monomial,

$$\mathrm{Tr}\left(\frac{c^3}{3}\right), \tag{IV.46}$$

which, of course, is determined modulo a trivial exact BRS cocycle. Recalling then (Appendix B) that the difference $(\mathrm{Tr}\,c^3-\mathrm{Tr}\,\overline{c}^3)$ is cohomologically trivial, i.e.,

$$\mathrm{Tr}\,c^3-\mathrm{Tr}\,\overline{c}^3=s(\cdots),$$

we can choose for ω^3 the following symmetric expression [one should observe that due to the anti-hermiticity property of the group generators T^a, the cocycle $(\mathrm{Tr}\,c^3+\mathrm{Tr}\,\overline{c}^3)$ is real]:

$$\omega^3=\mathrm{Tr}\left(\frac{c^3}{3}\right)+\mathrm{Tr}\left(\frac{\overline{c}^3}{3}\right). \tag{IV.47}$$

On the other hand, it is easily established that all the terms $R^{\alpha}\omega^3$, $\overline{R}_{\dot{\alpha}}\omega^3$, $R^{\alpha}\omega^2_{\alpha}$, $\overline{R}_{\dot{\alpha}}\overline{\omega}^{2\dot{\alpha}}$ on the right-hand side of Eqs. (IV.43) are trivial BRS cocycles. Considering, for instance, the first term, we have, from Eqs. (III.24),

$$sR^{\alpha}\omega^3 = R^{\alpha}s\omega^3 = 0, \tag{IV.48}$$

which implies that $R^{\alpha}\omega^3$ belongs to the cohomology of s in the sector of ghost number two and dimension one-half. Therefore, being the BRS cohomology empty in this sector, it follows that

$$R^{\alpha}\omega^3 = s\Lambda^{1\alpha}, \tag{IV.49}$$

as well as

$$\overline{R}^{\dot{\alpha}}\omega^3 = s\overline{\Lambda}^{1\dot{\alpha}}. \tag{IV.50}$$

In fact, from

$$R^{\alpha}\,\mathrm{Tr}\left(\frac{c^3}{3}\right) = s\,\mathrm{Tr}(cR^{\alpha}c) = s\,\mathrm{Tr}(cF^{\alpha}),$$

$$\overline{R}^{\dot{\alpha}}\,\mathrm{Tr}\left(\frac{c^3}{3}\right) = s\,\mathrm{Tr}(c\overline{R}^{\dot{\alpha}}c),$$

and

$$R^{\alpha}\,\mathrm{Tr}\left(\frac{\overline{c}^3}{3}\right) = s\,\mathrm{Tr}(\overline{c}R^{\alpha}\overline{c}),$$

$$\overline{R}^{\dot{\alpha}}\,\mathrm{Tr}\left(\frac{\overline{c}^3}{3}\right) = s\,\mathrm{Tr}(\overline{c}\overline{R}^{\dot{\alpha}}\overline{c}) = s\,\mathrm{Tr}(\overline{c}\overline{F}^{\dot{\alpha}}),$$

we have that $\Lambda^{1\alpha}$ and $\overline{\Lambda}^{1\dot{\alpha}}$ can be identified, modulo trivial terms, with

$$\Lambda^{1\alpha} = \mathrm{Tr}(cF^{\alpha}) + \mathrm{Tr}(\overline{c}R^{\alpha}\overline{c}), \quad \overline{\Lambda}^{1\dot{\alpha}} = \mathrm{Tr}(\overline{c}\overline{F}^{\dot{\alpha}}) + \mathrm{Tr}(c\overline{R}^{\dot{\alpha}}c), \tag{IV.51}$$

where $\overline{R}^{\dot{\alpha}}c, R^{\alpha}\overline{c}$ are given in Eqs. (III.23).

In the same way, we have

$$R^{\alpha}\omega^2_{\alpha} = R^{\alpha}\zeta_{\alpha}\omega^3 = \zeta_{\alpha}R^{\alpha}\omega^3 = \zeta_{\alpha}s\Lambda^{1\alpha} = s(\zeta_{\alpha}\Lambda^{1\alpha}) + D_{\alpha}\Lambda^{1\alpha}, \tag{IV.52}$$

showing that $R^{\alpha}\omega^2_{\alpha}$ is a trivial BRS cocycle plus a total superspace derivative. The same conclusions hold for $\overline{R}_{\dot{\alpha}}\omega^3$ and $\overline{R}_{\dot{\alpha}}\overline{\omega}^{2\dot{\alpha}}$ and can be extended by similar arguments to include the G terms $G^{\alpha}_{\dot{\alpha}}\omega^3$ and $G^{\alpha}_{\dot{\alpha}}\tilde{\omega}^{2\dot{\alpha}}_{\alpha}$.

The final result is that the equations (IV.43) can be rewritten without the explicit presence of the operators R and G, thus yielding the final version of the superspace descent equations for the invariant action, i.e.,

$$\begin{aligned}
s\left(\omega^0+\tfrac{1}{4}\bar{\zeta}_{\dot\alpha}\bar{\Lambda}^{1\dot\alpha}+\tfrac{1}{4}\zeta^\alpha\Lambda^1_\alpha\right)&=-\tfrac{1}{2}D^\alpha\left(\omega^1_\alpha+\tfrac{1}{2}\Lambda^1_a\right)+\tfrac{1}{2}\bar{D}_{\dot\alpha}\left(\bar{\omega}^{1\dot\alpha}-\tfrac{1}{2}\bar{\Lambda}^{1\dot\alpha}\right),\\
s\left(\bar{\omega}^1_{\dot\alpha}-\tfrac{1}{2}\bar{\Lambda}^1_{\dot\alpha}\right)&=-\tfrac{1}{2}\bar{D}_{\dot\alpha}D^\alpha\omega^2_\alpha-D^\alpha\bar{D}_{\dot\alpha}\omega^2_\alpha-\tfrac{1}{2}D^2\bar{\omega}^2_{\dot\alpha},\\
s\left(\omega^{1\alpha}+\tfrac{1}{2}\Lambda^{1\alpha}\right)&=\tfrac{1}{2}D^\alpha\bar{D}_{\dot\alpha}\bar{\omega}^{2\dot\alpha}+\bar{D}_{\dot\alpha}D^\alpha\bar{\omega}^{2\dot\alpha}+\tfrac{1}{2}\bar{D}^2\omega^{2\alpha},\\
s\bar{\omega}^2_{\dot\alpha}&=-\bar{D}_{\dot\alpha}\omega^3,\\
s\omega^{2\alpha}&=-D^\alpha\omega^3,\\
s\omega^3&=0.
\end{aligned}\tag{IV.53}$$

In particular, the first equation of the above system explicitly shows that the invariant action $\mathcal{L}^0$ can be identified with

$$\mathcal{L}^0=\omega^0+\tfrac{1}{4}\bar{\zeta}_{\dot\alpha}\bar{\Lambda}^{1\dot\alpha}+\tfrac{1}{4}\zeta^\alpha\Lambda^1_\alpha.\tag{IV.54}$$

The above expression has to be understood modulo an exact BRS cocycle or a total superspace derivative. Its nontriviality relies on the nontriviality of the ghost cocycle (IV.47), as one can show by using a well-known standard cohomological argument.[14,15] Recalling then the expressions (IV.45), (IV.51), for $\mathcal{L}^0$, we get

$$\mathcal{L}^0=\tfrac{1}{4}\mathrm{Tr}(\varphi^\alpha F_\alpha)+\tfrac{1}{4}\mathrm{Tr}(\bar{\varphi}_{\dot\alpha}\bar{F}^{\dot\alpha}),$$

which when integrated on the full superspace $d^4x\,d^2\theta\,d^2\bar{\theta}$, yields the familiar $N=1$ supersymmetric invariant Yang–Mills Lagrangian (here we recall the useful superspace identity $\int d^4x\,d^2\theta\,d^2\bar{\theta}=\int d^4x\,d^2\theta\,\bar{D}^2$):

$$S_{\mathrm{YM}}=\int d^4x\,d^2\theta\,d^2\bar{\theta}\,\mathcal{L}^0=\frac{1}{4}\int D^4x\,d^2\theta\,\mathrm{Tr}\,F^\alpha F_\alpha+\frac{1}{4}\int d^4x\,d^2\bar{\theta}\,\mathrm{Tr}\,\bar{F}_{\dot\alpha}\bar{F}^{\dot\alpha}.$$

V. DESCENT EQUATIONS FOR THE U(1) ANOMALY

As already remarked in the Introduction the BRS consistency condition for the chiral U(1) axial anomaly reads[19,24] as

$$s\int d^4x\,d^2\bar{\theta}\,K^0=0\Rightarrow sK^0=\bar{D}_{\dot\alpha}\bar{K}^{1\dot\alpha},\tag{V.55}$$

where K^0 and $\bar{K}^{1\dot\alpha}$ have dimensions two and three half and ghost numbers zero and one, respectively. K^0 has thus the same quantum numbers of the invariant action considered in the previous section, the only difference lying in the fact that the superspace measure, i.e., $d^4x\,d^2\bar{\theta}$, is now chiral instead of the vector one $d^4x\,d^2\bar{\theta}\,d^2\theta$. Therefore the descent equations for K^0 are obtained by performing the chiral limit of the vector equations (IV.43). Acting indeed with the BRS operator on the second equation of the condition (V.55), we obtain

$$\bar{D}_{\dot\alpha}s\bar{K}^{1\dot\alpha}=0.\tag{V.56}$$

Using then the results given in Appendix A, it follows that the general solution of the equation (V.56) is given by

$$
\begin{gathered}
s\bar{K}^{1\dot{\alpha}}=(\bar{D}^{\dot{\alpha}}D^{\alpha}+2D^{\alpha}\bar{D}^{\dot{\alpha}})K^2_{\alpha},\\
\bar{D}^2K^2_{\alpha}=0,
\end{gathered}
\tag{V.57}
$$

where K^2_{α} is of dimension one-half and ghost number two. Again, acting with the BRS operator in Eq. (V.57), one gets

$$
(\bar{D}^{\dot{\alpha}}D^{\alpha}+2D^{\alpha}\bar{D}^{\dot{\alpha}})sK^2_{\alpha}=0,
\tag{V.58}
$$

which, according to Appendix A, implies that

$$
\begin{gathered}
sK^2_{\alpha}=D^{\alpha}K^3,\\
\bar{D}^2D^{\alpha}K^3=0,\quad D^2\bar{D}^{\dot{\alpha}}K^3=0,
\end{gathered}
$$

with K^3 of dimension zero and ghost number three. Finally, from

$$
D^{\alpha}sK^3=0,
$$

it follows that

$$
sK^3=0.
$$

Summarizing, the superspace descent equations for the U(1) chiral axial anomaly are

$$
\begin{gathered}
sK^0=\bar{D}_{\dot{\alpha}}\bar{K}^{1\dot{\alpha}},\\
s\bar{K}^1_{\dot{\alpha}}=(2D^{\alpha}\bar{D}_{\dot{\alpha}}+\bar{D}_{\dot{\alpha}}D^{\alpha})K^2_{\alpha},\\
sK^{2\alpha}=D^{\alpha}K^3,\quad sK^3=0,
\end{gathered}
\tag{V.59}
$$

with the constraints

$$
\begin{gathered}
\bar{D}^2K^2_{\alpha}=0,\\
\bar{D}^2D^{\alpha}K^3=D^2\bar{D}^{\dot{\alpha}}K^3=0.
\end{gathered}
\tag{V.60}
$$

Recalling then the result of the previous section, for K^3 we have

$$
K^3=\left(\mathrm{Tr}\frac{c^3}{3}+\mathrm{Tr}\frac{\bar{c}^3}{3}\right)+s\Delta^2,
\tag{V.61}
$$

for some local power series Δ^2. It is interesting to observe that in this case the constraints (V.60) fix completely the trivial part of K^3, giving, for instance,

$$
s\Delta^2=0.
$$

Acting with the operator ζ_{α} on both sides of the last of Eqs. (V.59), and making use of the decomposition (II.17), for K^2_{α} one gets

$$
K^2_{\alpha}=-\zeta_{\alpha}K^3+s\Delta^1_{\alpha}.
$$

Once more, it is not difficult to prove that the imposition of the constraints (V.60) yields a unique expression for Δ^1_α, i.e.,

$$\Delta^1_\alpha = \mathrm{Tr}(c\varphi_\alpha),$$

so that for K^2_α we get

$$K^2_\alpha = \mathrm{Tr}(cD_\alpha c).$$

One sees thus that in the chiral case, due to the constraints (V.60), the trivial BRS contributions are uniquely fixed at the lowest levels of the descent equations. Repeating now the same procedure and making use of the relations (III.18) for $\overline{K}^{1\dot\alpha}$, one obtains

$$\overline{K}^{1\dot\alpha} = G^{\alpha\dot\alpha}K^2_\alpha - \overline{\Lambda}^{1\dot\alpha} + D^\alpha\overline{D}^{\dot\alpha}\,\mathrm{Tr}(c\varphi_\alpha) + \mathrm{Tr}(\overline{c}\overline{F}^{\dot\alpha}) + s\Delta^{0\dot\alpha}, \tag{V.62}$$

where the cocycle $\overline{\Lambda}^{1\dot\alpha}$ is the same as in Eq. (IV.48), i.e.,

$$\overline{\Lambda}^{1\dot\alpha} = \mathrm{Tr}(c\overline{R}^{\dot\alpha}c) + \mathrm{Tr}(\overline{c}\overline{F}^{\dot\alpha}).$$

Thus, that it follows

$$\overline{K}^{1\dot\alpha} = -2\,\mathrm{Tr}(D^\alpha c\overline{D}^{\dot\alpha}\varphi_\alpha) + s\Delta^{0\dot\alpha}. \tag{V.63}$$

Finally, acting with the operator $\overline{\zeta}_{\dot\alpha}$ on both sides of the equation,

$$s\overline{K}^1_{\dot\alpha} = (2D^\alpha\overline{D}_{\dot\alpha} + \overline{D}_{\dot\alpha}D^\alpha)K^2_\alpha,$$

for the last level K^0, we find

$$K^0 = -\overline{\zeta}_{\dot\alpha}\overline{K}^{1\dot\alpha} + \mathrm{Tr}(2\varphi^\alpha F_\alpha + \overline{D}_{\dot\alpha}\varphi^\alpha\overline{D}^{\dot\alpha}\varphi_\alpha),$$

reproducing the well-known expression for the U(1) supersymmetric chiral anomaly,

$$K^0 = \mathrm{Tr}(2\varphi^\alpha F_\alpha - \overline{D}_{\dot\alpha}\varphi^\alpha\overline{D}^{\dot\alpha}\varphi_a) - \overline{D}_{\dot\alpha}\Delta^{0\dot\alpha}.$$

Let us conclude by remarking that the expressions of the cocycles K^3, K^2_α, $\overline{K}^{1\dot\alpha}$, and K^0 found here are completely equivalent to those of Ref. 19, i.e., the difference is an exact BRS cocycle or a total superspace derivative.

VI. THE SUPERSYMMETRIC GAUGE ANOMALY

As the last example of our superspace analysis, let us consider the case of the supersymmetric gauge anomaly. As usual, let us first focus on the derivation of the corresponding descent equations. The latter, as mentioned in the Introduction and in Sec. IV, can be obtained by adding to the right-hand side of the generalized equation (IV.41) an appropriate extra term. The presence of this term actually stems from the BRS triviality[18] of the pure ghost cocycles $(\mathrm{Tr}\,c^{2n+1} - \mathrm{Tr}\,\overline{c}^{2n+1})$, $n \geqslant 1$,

$$s\Omega^{2n} = \mathrm{Tr}\frac{c^{2n+1}}{2n+1} - \mathrm{Tr}\frac{\overline{c}^{2n+1}}{2n+1}, \tag{VI.64}$$

Ω^{2n} being a local dimensionless functional of $(\phi, c, \overline{c})$ with ghost number $2n$. Acting in fact with the operator e^δ on both sides of Eq. (VI.64) and recalling the definitions (IV.32) and (IV.33), we get the desired modified version of the generalized superspace equation (IV.41) we are looking for,

$$
\begin{aligned}
\tilde{d}\tilde{\Omega} &= \mathrm{Tr}\frac{\tilde{A}^{2n+1}}{2n+1} - \mathrm{Tr}\frac{\tilde{\tilde{A}}^{2n+1}}{2n+1}, \\
\tilde{\Omega} &= e^{\delta}\Omega^{2n}.
\end{aligned}
\tag{VI.65}
$$

The descent equations for the gauge anomaly follows then from Eq. (VI.65) when $n=2$, i.e.,

$$
\begin{aligned}
\tilde{d}\tilde{\Omega} &= \tfrac{1}{5}\,\mathrm{Tr}(\tilde{A}^5 - \tilde{\tilde{A}}^5), \\
\tilde{\Omega} &= e^{\delta}\Omega^4.
\end{aligned}
\tag{VI.66}
$$

To see that the above equation characterizes indeed the gauge anomaly, let us write it in components. Expanding $\tilde{\Omega}$ in the global parameters $(e^{\alpha},\overline{e}^{\dot{\alpha}},\tilde{e}^{\dot{\alpha}}_{\alpha})$,

$$
\tilde{\Omega} = \Omega^4 + \Omega^{3\alpha}e_{\alpha} + \overline{\Omega}^3_{\dot{\alpha}}\overline{e}^{\dot{\alpha}} + \tilde{\Omega}^{3\alpha}_{\dot{\alpha}}\tilde{e}^{\dot{\alpha}}_{\alpha} + \tilde{\Omega}^{2a}_{\dot{\alpha}}e_{\alpha}\overline{e}^{\dot{\alpha}} + \Omega^{2\alpha}\tilde{e}_{\alpha\dot{\alpha}}\overline{e}^{\dot{\alpha}} + \overline{\Omega}^2_{\dot{\alpha}}e^{\alpha}\tilde{e}^{\dot{\alpha}}_{\alpha} + \Omega^1 e^{\alpha}\tilde{e}_{\alpha\dot{\alpha}}\overline{e}^{\dot{\alpha}},
\tag{VI.67}
$$

and eliminating the G and R terms as done in Sec. IV A, we get the known descent equations for the superspace gauge anomaly,[18,25]

$$
\begin{aligned}
s\Omega^1 &= D^{\alpha}\Omega^2_{\alpha} + \overline{D}_{\dot{\alpha}}\Omega^{2\dot{\alpha}}, \\
s\Omega^2_{\alpha} &= -\overline{D}^2\Omega^3_{\alpha} + (2\overline{D}_{\dot{\alpha}}D_{\alpha} + D_{\alpha}\overline{D}_{\dot{\alpha}})\overline{\Omega}^{3\dot{\alpha}} + 2\,\mathrm{Tr}((D_{\alpha}\overline{D}_{\dot{\alpha}}\overline{c})(\overline{c}\overline{D}^{\dot{\alpha}}\overline{c} + \overline{D}^{\dot{\alpha}}\overline{cc})), \\
s\overline{\Omega}^{2\dot{\alpha}} &= D^2\overline{\Omega}^{3\dot{\alpha}} - (2D^{\alpha}\overline{D}_{\dot{\alpha}} + \overline{D}_{\dot{\alpha}}D^{\alpha})\Omega^3_{\alpha} - 2\,\mathrm{Tr}((\overline{D}^{\dot{\alpha}}D^{\alpha}c)(cD_{\alpha}c + D_{\alpha}cc)), \\
s\Omega^3_{\alpha} &= D_{\alpha}\Omega^4 + \mathrm{Tr}(c^3 D_{\alpha}c), \\
s\overline{\Omega}^{3\dot{\alpha}} &= -\overline{D}^{\dot{\alpha}}\Omega^4 + \mathrm{Tr}(\overline{c}^3\overline{D}^{\dot{\alpha}}\overline{c}), \\
s\Omega^4 &= \tfrac{1}{5}\mathrm{Tr}(c^5 - \overline{c}^5).
\end{aligned}
\tag{VI.68}
$$

One sees, in particular, that integrating the first equation of (VI.68) on superspace, the cocycle Ω^1 obeys exactly the BRS consistency condition corresponding to the possible gauge breakings,

$$
s\int d^4x\, d^2\theta\, d^2\overline{\theta}\,\Omega^1 = 0,
$$

identifying therefore Ω^1 with the supersymmetric Yang–Mills anomaly.

In order to find a solution of the descent equations (VI.68), we use the same climbing procedure of the previous examples, obtaining the following nontrivial expressions:

$$
\begin{aligned}
\Omega^3_{\alpha} &= -\zeta^{\alpha}\Omega^4 - \mathrm{Tr}(\varphi^{\alpha}c^3), \\
\overline{\Omega}^{3\dot{\alpha}} &= \overline{\zeta}_{\dot{\alpha}}\Omega^4 - \mathrm{Tr}(\overline{\varphi}_{\dot{\alpha}}\overline{c}^3), \\
\Omega^2_{\alpha} &= G_{\alpha\dot{\alpha}}\overline{\zeta}^{\alpha}\Omega^4 + \overline{D}_{\dot{\alpha}}\tilde{\overline{\zeta}}^{\dot{\alpha}}\zeta_{\alpha}\Omega^4 - \mathrm{Tr}(\overline{\varphi}_{\dot{\alpha}}(D_{\alpha}\overline{\varphi}^{\dot{\alpha}})\overline{c}^2 - \overline{\varphi}_{\dot{\alpha}}\overline{c}(D_{\alpha}\overline{\varphi}^{\dot{\alpha}})\overline{c} + \overline{\varphi}_{\dot{\alpha}}\overline{c}^2 D_{\alpha}\overline{\varphi}^{\dot{\alpha}}) \\
&\quad + 2\,\mathrm{Tr}((D_{\alpha}\overline{\varphi}_{\dot{\alpha}})(\overline{c}\overline{D}^{\dot{\alpha}}\overline{c} + \overline{D}^{\dot{\alpha}}\overline{cc})), \\
\overline{\Omega}^{2\dot{\alpha}} &= G^{\alpha\dot{\alpha}}\zeta_{\alpha}\Omega^4 + D^{\alpha}\zeta^{\dot{\alpha}}\zeta_{\alpha}\Omega^4 + \mathrm{Tr}(\varphi^{\alpha}(\overline{D}^{\dot{\alpha}}\varphi_{\alpha})\overline{c}^2 - \varphi^{\alpha}\overline{c}(\overline{D}^{\dot{\alpha}}\varphi_{\alpha})\overline{c} + \varphi^{\alpha}\overline{c}^2\overline{D}^{\dot{\alpha}}\varphi_{\alpha}) \\
&\quad - 2\,\mathrm{Tr}((\overline{D}^{\dot{\alpha}}\varphi_{\alpha})(cD_{\alpha}c + D_{\alpha}cc)),
\end{aligned}
\tag{VI.69}
$$

and for the gauge anomaly,

$$\Omega^1 = 2\zeta^\alpha G_{\alpha\dot\alpha}\overline{\zeta}^{\dot\alpha}\Omega^4 + 2\,\mathrm{Tr}(F^\alpha c\varphi_\alpha - F^\alpha\varphi_\alpha c + (\overline{D}_{\dot\alpha}\varphi^\alpha)(\overline{D}^{\dot\alpha}\varphi_\alpha)c)$$
$$-2\,\mathrm{Tr}(\overline{F}_{\dot\alpha}\overline{c}\overline{\varphi}^{\dot\alpha} - \overline{F}_{\dot\alpha}\overline{\varphi}^{\dot\alpha}\overline{c} + (D^\alpha\overline{\varphi}_{\dot\alpha})(D_\alpha\overline{\varphi}^{\dot\alpha})\overline{c}). \tag{VI.70}$$

One should observe that the explicit final expression for the gauge anomaly depends on the knowledge of the cocycle Ω^4 solution of the last of the descent equations (VI.68). This point is particularly important and deserves some further clarifying remarks.

A. Nonpolynomial character of the gauge anomaly

It is known that due to a theorem by Ferrara, Girardello, Piguet, and Stora,[26] the superspace gauge anomaly cannot be expressed as a polynomial in the variables $(\varphi_\alpha, \lambda_\alpha \equiv e^\varphi D_\alpha e^{-\varphi})$ and their covariant derivatives. In fact, all the known superspace closed expressions of the gauge anomaly so far obtained by means of homotopic transgression procedures[27–30] show up as a highly nonpolynomial character in the gauge superconnection. On the other hand, in our approach the simple knowledge of the cocycle Ω^4 would produce a closed expression for the supersymmetric gauge anomaly without any homotopic integral. Of course, this would imply a deeper understanding of this anomaly. It is not difficult, however, to convince oneself that solving the equation

$$s\Omega^4 = \tfrac{1}{5}\mathrm{Tr}(c^5 - \overline{c}^5) \tag{VI.71}$$

is not an easy task. This is actually due to the BRS transformation of the vector superfield ϕ,

$$se^\phi = e^\phi c - \overline{c}e^\phi,$$

which when written in terms of ϕ, takes the highly complex form[24]

$$s\phi = \frac{1}{2}\mathcal{L}_\phi(c+\overline{c}) + \frac{1}{2}\mathcal{L}_\phi\left[\coth\left(\frac{\mathcal{L}_\phi}{2}\right)\right](c-\overline{c}), \tag{VI.72}$$

where

$$\mathcal{L}_\phi \cdot = [\phi, \cdot],$$

and

$$\coth\left(\frac{\mathcal{L}_\phi}{2}\right) = \frac{e^{\mathcal{L}_\phi/2} + e^{-\mathcal{L}_\phi/2}}{e^{\mathcal{L}_\phi/2} - e^{-\mathcal{L}_\phi/2}}.$$

The formula (VI.72) can be expanded in powers of ϕ, allowing us to solve the equation (VI.71) order by order in the vector superfield ϕ. For instance, in the first approximation, which corresponds to the Abelian limit of retaining only the linear terms of the BRS transformations, i.e.,

$$s \to s_{ab},$$

with

$$s_{ab}\phi = c - \overline{c},$$

$$s_{ab}c = s_{ab}\overline{c} = 0,$$

one easily checks that

$$\mathrm{Tr}(c^5 - \overline{c}^5) = s_{ab}\,\mathrm{Tr}(\phi(c^4 + c^3\overline{c} + c^2\overline{c}^2 + c\overline{c}^3 + \overline{c}^4)), \tag{VI.73}$$

which shows indeed the BRS triviality[1] of $\mathrm{Tr}(c^5-\overline{c}^5)$.

To our knowledge a closed exact form for Ω^4 has not yet been established. In other words, due to the theorem of Ferrara, Girardello, Piguet, and Stora,[26] the nonpolynomiality of the supersymmetric gauge anomaly directly relies on the nonpolynomial nature of the cocycle Ω^4. Any progress in this direction will be reported as soon as possible.

Let us conclude this section by giving the explicit expression of the gauge anomaly (VI.70) up to the second order in the vector field ϕ, i.e.,

$$\Omega^1=-2\,\mathrm{Tr}(D^\alpha\phi\overline{D}^2D_\alpha\phi c+\overline{D}^2D^\alpha\phi D_\alpha\phi c-(\overline{D}_{\dot\alpha}D^\alpha\phi)(\overline{D}^{\dot\alpha}D_\alpha\phi)c)+2\,\mathrm{Tr}(\overline{D}_{\dot\alpha}\phi D^2\overline{D}^{\dot\alpha}\phi\overline{c}+D^2\overline{D}_{\dot\alpha}\phi\overline{D}^{\dot\alpha}\phi\overline{c}-(D^\alpha\overline{D}_{\dot\alpha}\phi)(D_\alpha\overline{D}^{\dot\alpha}\phi)\overline{c}), \tag{VI.74}$$

which is easily recognized to be equivalent to that of Ref. 18. One should also observe that the above expressions do not receive contributions from the term Ω^4 since they are at least of the order three in ϕ, as it can be checked by applying the combination $\zeta^\alpha G_{\alpha\dot\alpha}\overline{\zeta}^{\dot\alpha}$ on the cocycle of Eq. (VI.73).

VII. CONCLUSION

The supersymmetric version of the descent equations for the four-dimensional $N=1$ super-Yang–Mills gauge theories can be analyzed by means of the introduction of two operators, ζ^α and $\overline{\zeta}^{\dot\alpha}$, which decompose the supersymmetric derivatives D^α and $\overline{D}^{\dot\alpha}$ as BRS commutators. These operators provide an algebraic setup for a systematic derivation of the superspace descent equations. In addition, they allow us to cast both the supersymmetric BRS transformations and the descent equations into a very suggestive zero curvature formalism in superspace.

ACKNOWLEDGMENTS

We are grateful to Olivier Piguet for many valuable discussions and private communications. The Conselho Nacional de Desenvolvimento Científico e Technológico CNPq-Brazil and SR2-UERJ are gratefully acknowledged for the financial support.

APPENDIX A

We list here the superspace algebraic solutions[18,19,24,31] of some equations needed for the analysis of the supersymmetric descent equations. All these solutions are built up by superfields. They have always to be understood modulo terms that automatically solve the corresponding equations but cannot be written in the same algebraic form as the solutions. The existence of such particular terms strongly depends on the superfield content of the particular model under consideration.

The first result states that the solution of the superspace equation,

$$\overline{D}^2Q=0,$$

can be generically written as

$$Q=\overline{D}_{\dot\alpha}\mathcal{M}^{\dot\alpha},$$

for some superfield $\mathcal{M}^{\dot\alpha}$.

The second important result concerns the solution of the following equation:

$$(2\overline{D}_{\dot\alpha}D_\alpha+D_\alpha\overline{D}_{\dot\alpha})\overline{Q}^{\dot\alpha}=\overline{D}^2Q_\alpha.$$

For the superfields $\overline{Q}^{\dot\alpha}$ and Q_α, we have now

$$\overline{Q}^{\dot\alpha}=\overline{D}^{\dot\alpha}\mathcal{M}, \quad Q_a=-D_\alpha\mathcal{M},$$

with $\mathcal{M}$ an arbitrary superfield. Let us observe that in this case the term $\mathrm{Tr}(cD_\alpha c)$, due to the fact that the ghost c is a chiral superfield, is automatically annihilated by the operator $\overline{D}^2$. Therefore it must be included in the expression given for Q_α, although it cannot be written as a total superspace derivative.

Considering now the equation

$$D^\alpha Q_\alpha=\overline{D}_{\dot\alpha}\overline{Q}^{\dot\alpha},$$

we have

$$\begin{aligned} Q_\alpha&=-\overline{D}^2\mathcal{P}_\alpha+(2\overline{D}_{\dot\alpha}D_\alpha+D_\alpha\overline{D}_{\dot\alpha})\overline{\mathcal{P}}^{\dot\alpha}+D^\beta\mathcal{N}_{(\alpha\beta)},\\ \overline{Q}^{\dot\alpha}&=-D^2\overline{\mathcal{P}}^{\dot\alpha}+(2D^\alpha\overline{D}^{\dot\alpha}+\overline{D}^{\dot\alpha}D^\alpha)\mathcal{P}_\alpha+\overline{D}_{\dot\beta}\overline{\mathcal{N}}^{(\dot\alpha\dot\beta)}, \end{aligned} \tag{A1}$$

with $\mathcal{P}_\alpha$ and $\mathcal{N}_{(\alpha\beta)}$ appropriate superfields. Of course, the existence of the symmetric superfield $\mathcal{N}_{(\alpha\beta)}$ depends on the dimension and on the ghost number of Q_α. For instance, in the case of the vector descent equations (IV.53) in which Q_α corresponds to $s(\omega^1_\alpha+\frac{1}{2}\Lambda^1_\alpha)$, it is not difficult to check that $\mathcal{N}_{(\alpha\beta)}$ is automatically absent due to the quantum numbers of the problem.

In particular, in the case of the chiral descent equations considered in Sec. V, Eqs. (A1) imply that the most general solution of Eq. (V.56) is given indeed by

$$s\overline{K}^{1\dot\alpha}=(\overline{D}^{\dot\alpha}D^\alpha+2D^\alpha\overline{D}^{\dot\alpha})K^2_\alpha,$$

with the constraint

$$\overline{D}^2K^2_\alpha=0.$$

APPENDIX B

In this appendix we summarize some useful results concerning the BRS superspace cohomology for the $N=1$ supersymmetric Yang–Mills gauge theories. The various BRS cohomolgy classes are labeled by the ghost number g and by the spinor indices.

The following results hold.[18,19,25]

(1) The BRS cohomology is empty in the space of the invariant local power series A^g with dimension 2 and positive ghost number g.

(2) The cohomology classes corresponding to local BRS invariant cocycles A^g_α or $\overline{A}^g_{\dot\alpha}$ with dimension $\frac{3}{2}$ and ghost number $g=1$, 2, or 3 are empty.

(3) The cohomology classes in the space of the BRS-invariant local power series A^g_α or $\overline{A}^g_{\dot\alpha}$ with dimension $\frac{1}{2}$ and ghost number g greater than zero are empty.

(4) The BRS cohomology classes in the space of the local power series A^g with dimension 0, ghost number g, and at least of order $g+1$ in the fields are empty.

(5) Any invariant object A^g with dimension 0 and even ghost number g greater than zero and of order g in the fields is BRS trivial.

In particular, it turns out that in the pure ghost sector the BRS cohomology classes are given by polynomials built up with monomials of the type

$$\mathrm{Tr}\frac{c^{2n+1}}{2n+1}, \quad n\geqslant 1, \tag{B1}$$

or

$$\mathrm{Tr}\frac{\bar{c}^{2n+1}}{2n+1}, \quad n \geqslant 1. \tag{B2}$$

We remark also that the two expressions above (B1) and (B2) do not actually define different cohomology classes. Instead they are equivalent, due to the triviality[18,24] of the combination,

$$\mathrm{Tr}\frac{c^{2n+1}}{2n+1} - \mathrm{Tr}\frac{\bar{c}^{2n+1}}{2n+1} = s\Omega^{2n},$$

for some local power series Ω^{2n}. This result implies that the expressions (B1) and (B2) are related to each other by means of an exact BRS cocycle.

APPENDIX C

In this appendix we show that the off-shell closure of the algebra (IV.28) can be recovered in a simple way by introducing an appropriate external field η. Indeed, let η be a superfield with dimension 2 and ghost number -1, whose BRS transformation reads as

$$s\eta = [\eta, c] + 2(DF + [\varphi, F]),$$

$$s^2\eta = 0.$$

Modifying now the operator ζ in such a way that

$$\zeta F = -\tfrac{1}{2}\eta,$$

it is easily verified that the commutator (IV.37)

$$[\zeta, s]F = -\zeta[c, F] + \tfrac{1}{2}s\eta = DF,$$

gives now the covariant derivative of F without making use of the equations of motion, closing therefore the algebra (IV.28) off shell. Let us conclude by also remarking that the external field η cannot contribute to the BRS cohomology classes relevant for the examples considered in the previous sections due to its ghost number and to its dimension.

APPENDIX D

This appendix is devoted to some technical details concerning the introduction of the global parameters e^{α}, $\overline{e^{\dot{\alpha}}}$, and $\tilde{e}^{\alpha\dot{\alpha}}$. As it has been already underlined, these parameters have been introduced in order to project the algebraic relations (II.16)–(III.24) on the equations of motion of $N=1$ super-Yang–Mills. This has been achieved by requiring that the conditions (IV.26) are fulfilled. In addition, as it is apparent from the construction of Sec. IV, the introduction of these global parameters, while collecting all the algebraic relations (II.16)–(III.24) into a unique extended generalized operator $\tilde{d}$ [see Eq. (IV.31)], is of great relevance in order to cast the BRS transformations of the superfields and the full system of superspace consistency conditions (IV.53), (V.59), and (VI.68) into a unique equation. The latter has the meaning of a zero-curvature condition [see Eqs. (IV.34)–(IV.36)). Therefore, the expansion of the equation,

$$\tilde{d}\tilde{\omega} = 0,$$

in powers of the global parameters $(e^{\alpha}, \overline{e}^{\dot{\alpha}}, \tilde{e}^{\alpha\dot{\alpha}})$, will automatically provide the full set of the superspace consistency conditions for the cocycle $\tilde{\omega}$. In this sense, the parameters $(e^{\alpha}, \overline{e^{\dot{\alpha}}}, \tilde{e}^{\alpha\dot{\alpha}})$ can be seen as a suitable basis in superspace for the zero curvature formulation of $N=1$ supersymmetry.

It is also worth remarking that this construction does not have the meaning of collecting into an extended BRS operator the various classical symmetries of a given gauge-fixed action, as done, for instance, in the case of the extended ($N \geqslant 2$) supersymmetric Yang–Mills theories.[22,23] [It is indeed rather simple to convince oneself that the operators $(\zeta_\alpha, \bar{\zeta}_{\dot{\alpha}})$ do not actually represent invariances of the gauge-fixed $N=1$ super-Yang–Mills action.] Rather, it is closer to the zero curvature formulation of the topological field theories discussed by Refs. 32 and 16 in terms of the so-called universal bundle.

[1] O. Piguet and S. P. Sorella, *Algebraic Renormalization*, Monographs Series, m 28 (Springer-Verlag, Berlin, 1995).
[2] F. Brandt, N. Dragon, and M. Kreuzer, Phys. Lett. B **231**, 263 (1989); Nucl. Phys. B **332**, 224 (1990); **332**, 250 (1990); **340**, 187 (1990); M. Dubois-Violette, M. Henneaux, M. Talon, and C. M. Viallet, Phys. Lett. B **289**, 361 (1992).
[3] G. Barnich, F. Brandt, and M. Henneaux, Commun. Math. Phys. **174**, 57 (1995); **174**, 93 (1995)
[4] S. P. Sorella, Commun. Math. Phys. **157**, 231 (1993).
[5] S. P. Sorella and L. Tataru, Phys. Lett. B **324**, 351 (1994).
[6] M. Werneck de Oliveira and S. P. Sorella, Int. J. Mod. Phys. A **92**, 979 (1994); O. Moritsch, M. Schweda, and S. P. Sorella, Class. Quantum Grav. **11**, 1225 (1994); O. Moritsch and M. Schweda, Helv. Phys. Acta **67**, 289 (1994); P. A. Blaga, O. Moritsch, M. Schweda, T. Sommer, L. Tataru, and H. Zerrouki, Phys. Rev. D **61**, 2792 (1995); O. Moritsch, M. Schweda, T. Sommer, L. Tataru, and H. Zerrouki, ''BRST cohomology of Yang–Mills gauge fields in the presence of gravity in Ashtekar variables,'' REP TUW 94-17.
[7] E. Guadagnini, N. Maggiore, and S. P. Sorella, Phys. Lett. B **255**, 65 (1991); N. Maggiore and S. P. Sorella, Int. J. Mod. Phys. A **8**, 929 (1993); C. Lucchesi, O. Piguet, and S. P. Sorella, Nucl. Phys. B **395**, 325 (1993); D. Birmingham and M. Rakowski, Phys. Lett. B **275**, 289 (1992); **289**, 271 (1992).
[8] F. Delduc, F. Gieres, and S. P. Sorella, Phys. Lett. B **225**, 367 (1989); F. Delduc, C. Lucchesi, O. Piguet, and S. P. Sorella, Nucl. Phys. B **346**, 313 (1990); O. Piguet, ''On the role of vector supersymmetry in topological field theory,'' UGVA-DPT 1995/02-880, hep-th/9502033;
[9] A. Brandhuber, O. Moritsch, M. W. de Oliveira, O. Piguet, and M. Schweda, Nucl. Phys. B **341**, 173 (1994);
[10] M. Werneck de Oliveira, M. Schweda, and S. P. Sorella, Phys. Lett. B **315**, 93 (1993); G. Bandelloni and S. Lazzarini, J. Math. Phys. **34**, 5413 (1993); L. Tataru and I. V. Vancea, Int. J. Mod. Phys. A **11**, 375 (1996).
[11] A. Boresch, M. Schweda, and S. P. Sorella, Phys. Lett. B **328**, 36 (1994).
[12] M. Carvalho, L. C. Queiroz Vilar, and S. P. Sorella, Int. J. Mod. Phys. A **27**, 3877 (1995).
[13] J. A. Dixon, Commun. Math. Phys. **139**, 495 (1991).
[14] J. Thierry-Mieg, J. Math. Phys. **21**, 2834 (1980); Nuovo Cimento A **56**, 396 (1980); L. Baulieu, Nucl. Phys. B **241**, 557 (1984); L. Baulieu and J. Thierry-Mieg, Phys. Lett. B **145**, 53 (1984); J. Manes, R. Stora, and B. Zumino, Commun. Math. Phys. **102**, 157 (1985); R. Stora, ''Algebraic structure and topological origin of anomalies,'' in *Progress in Gauge Field Theory*, edited by 't Hofft *et al.* (Plenum, New York, 1984); R. D. Ball, Phys. Rep. **182**, 1 (1989).
[15] M. Dubois-Violette, M. Talon, and C. M. Viallet, Commun. Math. Phys. **102**, 105 (1985).
[16] M. Carvalho, L. C. Q. Vilar, C. A. G. Sasaki, and S. P. Sorella, J. Math. Phys. **37**, 5310 (1996); M. Carvalho, L. C. Q. Vilar, C. A. G. Sasaki, and S. P. Sorella, *ibid.* **37**, 5325 (1996).
[17] O. Piguet and K. Sibold, Nucl. Phys. B **197**, 257 (1982); **197**, 272 (1982).
[18] O. Piguet and K. Sibold, Nucl. Phys. B **247**, 484 (1984).
[19] O. Piguet and K. Sibold, Int. J. Mod. Phys. A **4**, 913 (1986); C. Lucchesi, O. Piguet, and K. Sibold, Helv. Phys. Acta **61**, 321 (1988).
[20] O. Piguet, K. Sibold, and M. Schweda, Nucl. Phys. B **174**, 183 (1980); J. A. Dixon, R. Minasian, and J. Rahmfeld, Commun. Math. Phys. **171**, 459 (1995).
[21] F. Brandt, Nucl. Phys. B **392**, 428 (1993); F. Brandt and J. París, Phys. Rev. D **55**, 2398 (1997); F. Brandt, Ann. Phys. **259**, 253 (1997).
[22] P. L. White, Class. Quantum Grav. **9**, 413, 1663 (1992).
[23] N. Maggiore, Int. J. Mod. Phys. A **10**, 3781, 3937 (1995); N. Maggiore, O. Piguet, and S. Wolf, Nucl. Phys. B **458**, 403 (1996).
[24] O. Piguet and K. Sibold, ''Renormalized supersymmetry,'' series *Progress in Physics* (Birkhäuser, Boston, 1986), Vol. 12.
[25] M. Porrati, Z. Phys. C **32**, 65 (1986).
[26] S. Ferrara, L. Girardello, O. Piguet, and R. Stora, Phys. Lett. B **157**, 179 (1985).
[27] N. K. Nielsen, Nucl. Phys. B **244**, 499 (1984).
[28] G. Girardi, R. Grimm, and R. Stora, Phys. Lett. B **156**, 203 (1985).
[29] L. Bonora, P. Pasti, and M. Tonin, Nucl. Phys. B **261**, 249 (1985); Phys. Lett. B **156**, 341 (1985).
[30] E. Guadagnini and M. Mintchev, Nucl. Phys. B **269**, 543 (1986); E. Guadagnini, K. Konishi, and M. Mintchev, Phys. Lett. B **157**, 37 (1985).
[31] O. Piguet (private communications).
[32] D. Birmingham, M. Blau, M. Rakowski, and G. Thompson, Phys. Rep. **209**, 129 (1991); Phys. Lett. B **272**, 217 (1991); **272**, 223 (1991); **275**, 289 (1992).

Complete sets of Bloch and Wannier functions composed of oscillator eigenfunctions

P. Zeiner[a] and R. Dirl
Institut für Theoretische Physik, & Center for Computational Materials Science, Technische Universität Wien A-1040 Wien, Wiedner Hauptstraße 8-10, Austria

B. L. Davies
School of Mathematics, University of Wales, Bangor, Gwynedd LL57 1UT, Wales, United Kingdom

(Received 23 October 1998; accepted for publication 1 February 1999)

We discuss Bloch and Wannier functions related to oscillator eigenfunctions. In particular, we construct complete sets of mutually orthogonal Bloch and Wannier functions. We show that they can be expressed in several ways in terms of theta functions and their derivatives. We also analyze their localization properties and discuss expectation values for specifically chosen Hamiltonians. © *1999 American Institute of Physics.* [S0022-2488(99)02406-8]

I. INTRODUCTION

Localized but nonorthogonal functions, like Gaussian orbitals, are widely used in physical applications (see, for instance, Ref. 1). In the case of periodic crystals, the most prominent and important functions for describing their electronic properties are Bloch[2,3] and Wannier[4] functions. Wannier was the first to consider Bloch and Wannier functions composed of Gaussian orbitals.[5] He showed that these Bloch functions can be expressed in terms of θ functions, which makes a detailed analysis of their properties possible.[6,7] Another recent work on these topics is Ref. 8.

However, in many applications it is not enough to use only Gaussian orbitals centered at the atomic sites. One approach is to use in addition Gaussian orbitals that are centered in between. Alternatively, one could use oscillator eigenfunctions that are all centered at the sites of the atoms. Thus, it is natural to study not only Bloch functions composed of Gaussian orbitals, but to investigate the more general case of Bloch and Wannier functions that are composed of oscillator eigenfunctions. This is particularly interesting since one can construct a complete set of mutually orthogonal Bloch and Wannier functions. Thus they may be used to expand any arbitrary Bloch function, which might go far beyond the usual expansion in terms of Gaussian orbitals. Moreover, this complete set of orthonormalized Bloch functions provides an alternative to the expansions in terms of plane waves. In fact, the plane waves can be seen as the limits of these orthonormal Bloch functions in the case that the oscillator frequency ω goes to zero (instead of ω we use the parameter $\alpha \sim \sqrt{\omega}$ in the following). Thus, particularly for small α the Bloch functions $\psi_n^\alpha(k,x)$, which are composed of oscillator eigenfunctions, are an alternative to the plane wave basis. On the other hand, for large α the Bloch functions $\psi_n^\alpha(k,x)$ are typical Bloch functions for tight-binding approximations. Thus, depending on the choice of α, these Bloch functions could be useful both in tight-binding approximations as well as in the case of nearly free electrons. The idea of constructing this complete set of mutually orthogonal Bloch and Wannier functions is the following: Since the (properly normalized) eigenfunctions form a complete orthonormal system of functions, the corresponding Bloch and Wannier functions should also form a complete basis. However, these functions are, in general, not mutually orthogonal. By applying an appropriate orthogonalization procedure, we construct a complete set of mutually orthonormal Bloch and Wannier functions.

[a]Electronic mail: zeiner@tph.tuwien.ac.at

All these considerations confine to the one-dimensional case. However, it is not difficult to generalize these ideas to the three-dimensional case. In the simplest case of primitive lattices the three-dimensional Bloch functions composed of the eigenfunctions of a three-dimensional harmonic oscillator are just products of three one-dimensional Bloch functions composed of eigenfunctions of one-dimensional harmonic oscillators.

In the first section of this paper we transform the oscillator eigenfunctions $u_n(x)$ into Bloch functions $\phi_n(k,x)$ and discuss their properties as well as the corresponding Wannier functions. In particular, we show that these Bloch and Wannier functions are, in general, not orthogonal for different values of n. In the following sections we construct complete orthonormal systems of Bloch and Wannier functions. We show that these Bloch functions can be represented in various ways in terms of θ functions and their derivatives. In particular, the k dependence can be expressed in terms of θ functions only. In addition, we discuss the localization properties of the corresponding Wannier functions and their symmetry properties, as well as certain limits and several expectation values.

II. BLOCH AND WANNIER FUNCTIONS COMPOSED OF OSCILLATOR EIGENFUNCTIONS

A. Harmonic oscillator

First we want to recall some basic properties of the harmonic oscillator. The normalized eigenfunctions of the Hamiltonian $H=(1/2m)P^2+(m\omega^2/2)X^2$ are given by

$$u_n^\alpha(x)=\sqrt{\alpha}(\sqrt{\pi}2^n n!)^{-1/2}H_n(\alpha x)e^{-\alpha^2x^2/2}, \tag{1}$$

where $\alpha=\sqrt{m\omega/\hbar}$ is the inverse of the classical amplitude for a harmonic oscillator whose energy is equal to the ground state energy $E_0=\frac{1}{2}\hbar\omega$. Here the symbols H_n denote the Hermite polynomials,[9] which are defined as follows:

$$H_n(x)=e^{x^2/2}\left(x-\frac{d}{dx}\right)^n e^{-x^2/2}=(-1)^n e^{x^2}\frac{d^n}{dx^n}e^{-x^2}. \tag{2}$$

Recall that the u_n^α are real, fulfilling the symmetry relations $u_n^\alpha(-x)=(-1)^n u_n^\alpha(x)$, and satisfying the orthonormality condition $\int_{-\infty}^{\infty}u_m^\alpha(x)u_n^\alpha(x)dx=\delta_{mn}$, respectively. For sake of simplicity, we set $\hbar=1$.

B. Bloch functions

Now we can investigate the Bloch functions that are composed of the oscillator eigenfunctions $u_n^\alpha(x)$, namely,

$$\phi_n^\alpha(k,x)=N_n^\alpha(k)\sum_{m=-\infty}^{\infty}e^{ikm}u_n^\alpha(x-m), \tag{3}$$

where $N_n^\alpha(k)$ is chosen positive and such that $\langle\phi_n^\alpha(k),\phi_n^\alpha(k)\rangle_1=\int_0^1 dx|\phi_n^\alpha(k,x)|^2=1/2\pi$. Recall that the Bloch functions for $n=0$ are given by[6]

$$\phi_0^\alpha(k,x)=\left(\frac{\alpha}{\sqrt{\pi}}\right)^{1/2}\frac{1}{\sqrt{2\pi}}e^{-\alpha^2x^2/2}\frac{\theta_3\left(\frac{k}{2}-i\frac{\alpha^2x}{2}\middle|\frac{i\alpha^2}{2\pi}\right)}{\sqrt{\theta_3\left(\frac{k}{2}\middle|\frac{i\alpha^2}{4\pi}\right)}}=\frac{1}{\sqrt{2\pi}}e^{ikx}\frac{\theta_3\left(\frac{i\pi}{\alpha^2}k+\pi x\middle|\frac{i2\pi}{\alpha^2}\right)}{\sqrt{\theta_3\left(\frac{i2\pi}{\alpha^2}k\middle|\frac{4i\pi}{\alpha^2}\right)}}, \tag{4}$$

$$N_0^\alpha(k)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\theta_3\left(\frac{k}{2}\middle|\frac{i\alpha^2}{4\pi}\right)}}=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\alpha}{\sqrt{4\pi}}}e^{k^2/2\alpha^2}\frac{1}{\sqrt{\theta_3\left(\frac{i2\pi}{\alpha^2}k\middle|\frac{4i\pi}{\alpha^2}\right)}},\tag{5}$$

where θ_3 is one of the well-known θ functions.[10–12] The Bloch functions $\phi_n^\alpha(k,x)$ can be expressed either in terms of theta functions or by the special Bloch functions $\phi_0^\alpha(k,x)$ and their derivatives:

$$\phi_n^\alpha(k,x)=N_n^\alpha(k)\left(\frac{\alpha}{\sqrt{\pi}2^n n!}\right)^{1/2}\sum_{m=-\infty}^{\infty}e^{ikm}H_n(\alpha(x-m))e^{-\alpha^2(x-m)^2/2}\tag{6a}$$

$$=N_n^\alpha(k)\left(\frac{\alpha}{\sqrt{\pi}2^n n!}\right)^{1/2}\left[\alpha\left(x+i\frac{\partial}{\partial k}\right)-\frac{1}{\alpha}\frac{\partial}{\partial x}\right]^n e^{-\alpha^2x^2/2}\theta_3\left(\frac{k}{2}-i\frac{\alpha^2x}{2}\middle|\frac{i\alpha^2}{2\pi}\right)\tag{6b}$$

$$=N_n^\alpha(k)(2^n n!)^{-1/2}\left[\alpha\left(x+i\frac{\partial}{\partial k}\right)-\frac{1}{\alpha}\frac{\partial}{\partial x}\right]^n\frac{\phi_0^\alpha(k,x)}{N_0^\alpha(k)}\tag{6c}$$

$$=N_n^\alpha(k)(2^n n!)^{-1/2}e^{ikx}\left[i\alpha\frac{\partial}{\partial k}-\frac{1}{\alpha}\left(\frac{\partial}{\partial x}+ik\right)\right]^n e^{-ikx}\frac{\phi_0^\alpha(k,x)}{N_0^\alpha(k)}\tag{6d}$$

$$=N_n^\alpha(k)\left(\frac{2\sqrt{\pi}}{\alpha 2^n n!}\right)^{1/2}e^{ikx}\left[i\alpha\frac{\partial}{\partial k}-\frac{1}{\alpha}\left(\frac{\partial}{\partial x}+ik\right)\right]^n e^{-k^2/2\alpha^2}\theta_3\left(\frac{i\pi}{\alpha^2}k+\pi x\middle|\frac{i2\pi}{\alpha^2}\right)\tag{6e}$$

$$=N_n^\alpha(k)\left(\frac{2\sqrt{\pi}}{\alpha 2^n n!}\right)^{1/2}(-i)^n e^{ikx}\sum_{m=-\infty}^{\infty}H_n\left(\frac{k+2\pi m}{\alpha}\right)e^{-(k+2\pi m)^2/2\alpha^2}e^{i2\pi mx}.\tag{6f}$$

The normalization constants $N_n^\alpha(k)$ read as

$$(N_n^\alpha(k))^{-2}=4\pi\frac{\sqrt{\pi}}{\alpha}(2^n n!)^{-1}\sum_{m=-\infty}^{\infty}H_n\left(\frac{k+2\pi m}{\alpha}\right)^2 e^{-(k+2\pi m)^2/\alpha^2}\tag{7a}$$

$$=2\pi\sum_{r=-\infty}^{\infty}L_n^0\left(\frac{\alpha^2r^2}{2}\right)e^{-\alpha^2r^2/4}e^{-ikr},\tag{7b}$$

where $L_n^\gamma(x)$ are the Laguerre polynomials[9]

$$L_n^\gamma(x)=\frac{1}{n!}e^x x^{-\gamma}\frac{d^n}{dx^n}x^{n+\gamma}e^{-x}=\sum_{m=0}^{n}\binom{n+\gamma}{n-m}\frac{(-x)^m}{m!}.\tag{8}$$

Note that Eq. (7b) is just the Fourier series of Eq. (7a).

The inner product of Bloch functions referring to different n values is easily calculated:

$$\langle\phi_m^\alpha(k),\phi_n^\alpha(k)\rangle_1=\frac{i^{m-2}2\sqrt{\pi}N_m^\alpha(k)N_n^\alpha(k)}{\alpha(2^{m+n}m!n!)^{1/2}}\sum_{r=-\infty}^{\infty}H_m\left(\frac{k+2\pi r}{\alpha}\right)H_n\left(\frac{k+2\pi r}{\alpha}\right)e^{-(k+2\pi r)^2/\alpha^2}\tag{9a}$$

$$=\frac{N_m^\alpha(k)N_n^\alpha(k)(2^n n!)^{1/2}}{(2^m m!)^{1/2}}\sum_{r=-\infty}^{\infty}(-\alpha r)^{m-n}L_n^{m-n}\left(\frac{\alpha^2r^2}{2}\right)e^{-\alpha^2r^2/4}e^{-ikr}.\tag{9b}$$

Again the second equation is the Fourier series of the first one. From Eq. (9b) we infer that the Bloch functions cannot be orthogonal for all k values. For special k, however, the Bloch functions may be orthogonal. For instance, if $k=0$ than $\langle \phi_m^\alpha(k), \phi_n^\alpha(k)\rangle_1 = 0$ if $m-n$ is odd.

Let us briefly discuss the symmetry properties of the Bloch functions ϕ_n^α. They have the following transformation properties:

$$(\phi_n^\alpha(k,x))^* = \phi_n^\alpha(-k,x), \tag{10}$$

$$\phi_n^\alpha(k,-x) = (-1)^n \phi_n^\alpha(-k,x), \tag{11}$$

which correspond to time reversal and inversion symmetry, respectively. These equations follow immediately from the fact that the oscillator eigenfunctions are real and even or odd, depending on $n=2m$ or $n=2m+1$, respectively. In terms of band representations (see Refs. 13–15) the symmetry of the Bloch functions ϕ_n^α is described as follows: For n even the Bloch functions ϕ_n^α transform according to the band representation of the only nontrivial one-dimensional space group $\mathbb{Z} \circledS C_2$, where the former is characterized by the Wyckoff position $w=0$ and the trivial one-dimensional irreducible unitary representation of the corresponding site symmetry group. For n odd the Bloch functions ϕ_n^α transform according to the band representation that is characterized by the Wyckoff position $w=0$ and the one-dimensional irreducible unitary representation $\{1,-1\}$ of C_2.

C. Wannier functions

We adopt the following convention for the Wannier functions:

$$W_n^\alpha(m;x) = W_n^\alpha(x-m) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} dk\, e^{-ikm} \phi_n^\alpha(k,x), \tag{12}$$

which leads to the following symmetry relations $W_n^\alpha(x) = (-1)^n W_n^\alpha(-x)$. Unfortunately there is no explicit expression for the Wannier functions. Since the Bloch functions are unique only up to a phase factor, we could have also chosen

$$W_n^\alpha(m;x) = W_n^\alpha(x-m) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} dk\, e^{-ikm} e^{i\gamma(k)} \phi_n^\alpha(k,x) \tag{13}$$

to define the Wannier functions, where $e^{i\gamma(k)}$ is a phase factor. This freedom of definition of the Wannier functions is exploited in Ref. 5 for $n=0$ to construct Wannier functions that can be stated explicitly. However, this is no longer possible for $n \geqslant 1$, and thus we use Eq. (12) to define the Wannier functions. For $n=0$ the Wannier functions have a remarkable property: The uncertainty of position $\Delta x = 1/2\alpha^2$ is the same for the Wannier functions and the original Gaussians.[6,7] Thus, it is interesting to see whether this result can be generalized for all n. Obviously $\langle x \rangle = 0$ holds due to the symmetry relations $W_n^\alpha(x) = (-1)^n W_n^\alpha(-x)$, and we need to compute $\langle x^2 \rangle$ only. Using several partial integrations and some well-known formulas for the Hermite polynomials,[9] we get

$$\langle x^2\rangle_{W_n^\alpha}=\frac{2n+1}{2\alpha^2}-\frac{n}{2\pi\alpha^2}\int_{-\pi}^{\pi}dk\,\frac{\sum_{m=-\infty}^{\infty}e^{-(k+2\pi m)^2/\alpha^2}H_{n-1}\left(\frac{k+2\pi m}{\alpha}\right)H_{n+1}\left(\frac{k+2\pi m}{\alpha}\right)}{\sum_{m=-\infty}^{\infty}e^{-(k+2\pi m)^2/\alpha^2}H_n\left(\frac{k+2\pi m}{\alpha}\right)^2} \tag{14a}$$

$$=\frac{2n+1}{2\alpha^2}+\frac{1}{4\pi\alpha^2}\int_{-\pi}^{\pi}dk\,\frac{\sum_{m=-\infty}^{\infty}(\alpha m)^2L_{n-1}^2\left(\frac{\alpha^2m^2}{2}\right)e^{-\alpha^2m^2/4}e^{-ikm}}{\sum_{m=-\infty}^{\infty}L_n^0\left(\frac{\alpha^2m^2}{2}\right)e^{-\alpha^2m^2/4}e^{-ikm}}, \tag{14b}$$

where the second equation is obtained by expanding the integrand into a Fourier series. Apparently $\langle x^2\rangle_{W_n^\alpha}\neq\langle x^2\rangle_{u_n^\alpha}$ and, in fact, numerical calculations suggest that $\langle x^2\rangle_{W_n^\alpha}>\langle x^2\rangle_{u_n^\alpha}$ is valid for arbitrary α, but it is an open question how to prove this statement analytically.

For large α we get the following approximation:

$$\langle x^2\rangle_{W_n^\alpha}=\frac{2n+1}{2\alpha^2}-L_{n-1}^2\left(\frac{\alpha^2}{2}\right)L_n^0\left(\frac{\alpha^2}{2}\right)e^{-\alpha^2/2}+O(\alpha^{8n-2}e^{-\alpha^2}), \tag{15}$$

hence we have, in fact, $\langle x^2\rangle_{W_n^\alpha}>\langle x^2\rangle_{u_n^\alpha}$ for sufficiently large α. For sufficiently small α the following heuristic arguments apply. The main problem are the zeros of the Hermite polynomials. Consider the integral in Eq. (14a). Of course, it would be quite tentative to neglect all terms with $m\neq0$, but in this approximation the denominator would have a zero of second order, and thus the integral would diverge. So we have to take into account also the terms for $m=\pm1$, and the denominator reads in this approximation as

$$e^{-k^2/\alpha^2}H_n\left(\frac{k}{\alpha}\right)^2+e^{-(k-2\pi)^2/\alpha^2}H_n\left(\frac{k-2\pi}{\alpha}\right)^2+e^{-(k+2\pi)^2/\alpha^2}H_n\left(\frac{k+2\pi}{\alpha}\right)^2.$$

Let us assume $n=1$ for simplicity. Then the only zero of the Hermite polynomial $H_1(x)=x$ is $x=0$. Hence, for $|k/\alpha|\lesssim e^{-2\pi^2/\alpha^2}$ the denominator is of order $e^{-(2\pi)^2/\alpha^2}$. On the other hand, for $|k/\alpha|\gtrsim e^{-2\pi^2/\alpha^2}$ we may neglect all terms with $m\neq0$, and hence the integrand may be approximated by $H_{n-1}(k/\alpha)H_{n+1}(k/\alpha)/H_n(k/\alpha)^2$, which should contribute less to the integral than the interval $|k/\alpha|\lesssim e^{-2\pi^2/\alpha^2}$. Thus, we infer that the integral in question is of order $\sim O(e^{2\pi^2/\alpha^2})$. In fact, the term of order $\sim O(e^{2\pi^2/\alpha^2})$ is positive, since $H_0(0)H_2(0)<0$. Hence $\langle x^2\rangle_{W_1^\alpha}\sim O(e^{2\pi^2/\alpha^2})\gg3/2\alpha^2$. Similarly, for all $n>1$ the main contributions to the integral arise from the zeros of the Hermite polynomial H_n, and are again approximately of order $O(e^{2\pi^2/\alpha^2})$. Again the integral is negative, since $H_{n-1}(x)H_{n+1}(x)<0$ for $H_n(x)=0$.

D. The limit $\alpha\to0$

The limits of the Bloch functions ϕ_n^α read as

$$\phi_n^0(k,x):=\lim_{\alpha\to0}\phi_n^\alpha(k,x)=\begin{cases}(-1)^{n/2}\frac{1}{\sqrt{2\pi}}e^{ikx}, & \dot{k}\in(-\pi,\pi)\setminus\{0\},\\ \frac{1}{\sqrt{2\pi}}, & k=0,\\ (-1)^{n/2}\frac{1}{\sqrt{\pi}}\cos(\pi x), & k=\pm\pi,\end{cases} \tag{16}$$

if n is even, and

$$\phi_n^0(k,x) := \lim_{\alpha\to 0} \phi_n^\alpha(k,x) = \begin{cases} -i^n \frac{1}{\sqrt{2\pi}} e^{ikx}, & k\in(0,\pi), \\ i^n \frac{1}{\sqrt{2\pi}} e^{ikx}, & k\in(-\pi,0), \\ (-1)^{(n-1)/2} \frac{1}{\sqrt{\pi}} \sin(2\pi x), & k=0, \\ (-1)^{(n-1)/2} \frac{1}{\sqrt{\pi}} \sin(\pi x), & k=\pm\pi, \end{cases} \tag{17}$$

for n odd. Convergence is uniform with respect to x and uniform with respect to k in the intervals $[-\pi+\epsilon,-\epsilon]$ and $[\epsilon,\pi-\epsilon]$. Note that except for $k=0,\pi$ the Bloch functions $\phi_n^\alpha(k,x)$ converge for all n to the same free electron wave function (up to a phase factor). Thus the limit Bloch functions do not form a complete set of Bloch functions, whereas for all $\alpha>0$ the Bloch functions $\{\phi_n^\alpha(k)\}$ form a complete (but not orthogonal) basis.

The limits of the corresponding Wannier functions are given by

$$W_{2m}^0(x) := \lim_{\alpha\to 0} W_{2m}^\alpha(x) = (-1)^m \frac{\sin \pi x}{\pi x}, \tag{18}$$

$$W_{2m+1}^0(x) := \lim_{\alpha\to 0} W_{2m+1}^\alpha(x) = 2(-1)^m \frac{\left(\sin \frac{\pi x}{2}\right)^2}{\pi x}. \tag{19}$$

III. ORTHONORMALIZED BLOCH FUNCTIONS

A. Bloch functions

We have discussed the Bloch functions composed of oscillator eigenfunctions in the preceding sections and we have seen that Bloch functions of different n values are for any given $k\in(-\pi,\pi)$, in general, not orthogonal, contrary to the oscillator eigenfunctions. In the following sections we want to construct and discuss a complete set of orthonormal Bloch functions composed of oscillator eigenfunctions. One way of constructing an orthonormal set is to apply the Gram–Schmidt procedure onto the Bloch functions $\phi_n^\alpha(k,x)$ of the preceding sections. Since the Hermite polynomials H_n are polynomials of degree n, the Bloch functions $\phi_n^\alpha(k,x)$ are just linear combinations of the following (not normalized) Bloch functions:

$$\chi_\ell^\alpha(k,x) = (-i)^\ell \sum_{m=-\infty}^{\infty} \left(\frac{k+2\pi m}{\alpha}\right)^\ell e^{-(k+2\pi m)^2/2\alpha^2} e^{i(k+2\pi m)x} \tag{20a}$$

$$= \left(-\frac{1}{\alpha}\frac{\partial}{\partial x}\right)^\ell \chi_0^\alpha(k,x) = \left(-\frac{1}{\alpha}\frac{\partial}{\partial x}\right)^n e^{ikx} e^{-k^2/2\alpha^2} \theta_3\left(\frac{i\pi^k}{\alpha^2}+\pi x \middle| \frac{i2\pi}{\alpha^2}\right), \tag{20b}$$

with $\ell=0,...,n$. Thus, we can alternatively apply the Gram–Schmidt procedure to the functions $\chi_n^\alpha(k,x)$. Then the orthogonalized Bloch functions read as

$$\eta_n^\alpha(k,x) = \det Q_n^\alpha(k,x), \tag{21}$$

where $Q_n^\alpha(k,x)$ is the $(n+1)\times(n+1)$ matrix with matrix elements $q_{n;m,\ell}^\alpha(k)$,

$$q^{\alpha}_{n;m,\ell}(k)=\langle \chi^{\alpha}_{m}(k),\chi^{\alpha}_{\ell}(k)\rangle_1, \quad \text{for } m=0,...,n-1, \quad \ell=0,...,n, \tag{22a}$$

$$q^{\alpha}_{n;n,\ell}(k,x)=\chi^{\alpha}_{\ell}(k,x), \quad \ell=0,\dots,n, \tag{22b}$$

where $\langle f,g\rangle_1$ denotes the inner product $\int_0^1 f^*(x)g(x)dx$. The norm of $\eta^{\alpha}_n(k,x)$ is easily calculated:

$$\|\eta^{\alpha}_n(k)\|=\sqrt{\Delta^{\alpha}_{n-1}(k)\Delta^{\alpha}_n(k)}, \tag{23}$$

where $\Delta^{\alpha}_n(k)=\det(\langle\chi^{\alpha}_j(k),\chi^{\alpha}_{\ell}(k)\rangle_1)_{j,\ell=0,...,n}$ denotes the Gramian determinant. To obtain an explicit expression for $\psi^{\alpha}_n(k,x)$, we note that

$$\langle \chi^{\alpha}_j(k),\chi^{\alpha}_{\ell}(k)\rangle_1=(-i)^{\ell-j}\sum_{m=-\infty}^{\infty}\left(\frac{k+2\pi m}{\alpha}\right)^{j+\ell} e^{-(k+2\pi m)^2/\alpha^2} \tag{24a}$$

$$=(-1)^j\left(-\frac{1}{\alpha}\frac{\partial}{\partial y}\right)^{j+\ell}\sum_{m=-\infty}^{\infty} e^{-(k+2\pi m)^2/\alpha^2}e^{i(k+2\pi m)y}\Bigg|_{y=0} \tag{24b}$$

is valid. Now $\eta^{\alpha}_n(k,x)$ can be written as

$$\eta^{\alpha}_n(k,x)=(-i)^n\frac{1}{n!}\left(\frac{2\pi}{\alpha}\right)^{n^2}\sum_{m_0=-\infty}^{\infty}\cdots\sum_{m_n=-\infty}^{\infty}\left(\prod_{j=0}^{n-1} e^{-(k+2\pi m_j)^2/\alpha^2}\right)e^{-(k+2\pi m_n)^2/2\alpha^2}$$
$$\times\prod_{n-1\geqslant j>\ell\geqslant 0}(m_j-m_\ell)\prod_{n\geqslant j>\ell\geqslant 0}(m_j-m_\ell)e^{i(k+2\pi m_n)^x}, \tag{25}$$

where we have made use of Vandermonde's determinant, which reads as follows:

$$\det\begin{pmatrix} 1 & 1 & \cdots & 1\\ x_0 & x_1 & \cdots & x_n\\ x_0^2 & x_1^2 & \cdots & x_n^2\\ \vdots & \vdots & & \vdots\\ x_0^n & x_1^n & \cdots & x_n^n\end{pmatrix}=\prod_{i>j}(x_i-x_j). \tag{26}$$

Similarly, we obtain

$$\Delta^{\alpha}_n(k)=\frac{(2\pi)^{n(n+1)}}{(n+1)!\,\alpha^{n(n+1)}}\sum_{m_0=-\infty}^{\infty}\cdots\sum_{m_n=-\infty}^{\infty}\left(\prod_{j=0}^{n} e^{-(k+2\pi m_j)^2/\alpha^2}\right)\left[\prod_{n\geqslant j>\ell\geqslant 0}(m_j-m_\ell)\right]^2. \tag{27}$$

There are several transformation rules for the Bloch functions $\eta^{\alpha}_n(k,x)$. For instance, by changing the summation indices appropriately we see immediately that $\eta^{\alpha}_n(k,x)$ are periodic functions:

$$\eta^{\alpha}_n(k+2\pi,x)=\eta^{\alpha}_n(k,x). \tag{28}$$

Also, the following quasiperiodicity properties can be easily proved:

$$\eta^{\alpha}_n(k+i\alpha^2,x)=e^{(n+1/2)\alpha^2}e^{-(2n+1)ik}e^{-\alpha^2 x}\eta^{\alpha}_n(k,x), \tag{29}$$

$$\eta^{\alpha}_n(k,x+1)=e^{ik}\eta^{\alpha}_n(k,x), \tag{30}$$

$$\eta_n^\alpha\left(k+i\,\frac{\alpha^2}{2},x+\frac{1}{2}\right)=e^{(n+1/2)(\alpha^2/4)}e^{-ink}e^{-(\alpha^2/2)(x+1/2)}\,\eta_n^\alpha(k,x). \tag{31}$$

In addition, $\eta_n^\alpha(k,x)$ have the following symmetry properties:

$$\eta_n^\alpha(k,-x)=(-1)^n\,\eta_n^\alpha(-k,x), \tag{32}$$

$$(\eta_n^\alpha(k,x))^*=\eta_n^\alpha(-k,x), \tag{33}$$

which are the analogs of Eqs. (11) and (10). Thus, the Bloch functions $\eta_n^\alpha(k,x)$ transform according to the same band representations as the nonorthogonalized Bloch functions $\phi_n^\alpha(k,x)$.

The equations (28) and (29) are crucial since they determine the k dependence of the Bloch functions nearly completely. It is shown in Appendix A 1 that a function with the properties (28) and (29) can be written as

$$\eta_n^\alpha(k,x)=e^{ikx}e^{-[(2n+1)/2\alpha^2]k^2}\sum_{r=-n}^{n}d_r^{\alpha,n}(x)\,\theta_3\left(\frac{i\pi}{\alpha^2}k+\frac{\pi}{2n+1}(x-r)\,\middle|\,i\,\frac{2\pi}{(2n+1)\alpha^2}\right), \tag{34}$$

where the functions $d_r^{\alpha,n}(x)$ are independent of k. Note the simple structure of $\eta_n^\alpha(k,x)$ with respect to k: It is a sum of $2n+1$ theta functions multiplied by a Gaussian $e^{-[(2n+1)/2\alpha^2]k^2}$ and the plane wave factor e^{ikx}. Only the dependence on x is rather complicated. Next we take Eq. (30) into account and infer

$$d_r^{\alpha,n}(x+1)=d_{r-1}^{\alpha,n}(x). \tag{35}$$

Hence $d_r^{\alpha,n}(x)$ may be expressed by $d_0^{\alpha,n}(x)$, and thus we have

$$\eta_n^\alpha(k,x)=e^{ikx}e^{-[(2n+1)/2\alpha^2]k^2}\sum_{r=-n}^{n}d_0^{\alpha,n}(x-r)\,\theta_3\left(\frac{i\pi}{\alpha^2}k+\frac{\pi}{2n+1}(x-r)\,\middle|\,i\,\frac{2\pi}{(2n+1)\alpha^2}\right). \tag{36}$$

From Eqs. (31), (32), and (33) we infer the following properties of $d_0^{\alpha,n}(x)$:

$$d_0^{\alpha,n}(x+n+\tfrac{1}{2})=d_0^{\alpha,n}(x), \tag{37}$$

$$d_0^{\alpha,n}(-x)=(-1)^n d_0^{\alpha,n}(x), \tag{38}$$

$$(d_0^{\alpha,n}(x))^*=d_0^{\alpha,n}(x). \tag{39}$$

Let us determine the function $d_0^{\alpha,n}(x)$. To this end, note that $\eta_n^\alpha(k,x)$ is a linear combination of the $2n+1$ functions $\theta_3((i\pi/\alpha^2)k+[\pi/(2n+1)](x-r)|i[2\pi/(2n+1)\alpha^2])$ with the coefficients $d_0^{\alpha,n}(x-r)$, which are independent of k. Determining the function $d_0^{\alpha,n}(x)$ is, of course, equivalent to compute the $2n+1$ coefficients $d_0^{\alpha,n}(x-r)$. Thus, we need $2n+1$ equations for the coefficients $d_0^{\alpha,n}(x-r)$. These can be obtained from Eq. (36) by choosing $2n+1$ values for k. A convenient choice is $k=i\alpha^2[s/(2n+1)]$ for $s=-n,\ldots,n$. The resulting convolution equation,

$$\eta_n^\alpha\left(i\alpha^2\frac{s}{2n+1},x\right)e^{\alpha^2[s/(2n+1)]x}e^{-\alpha^2s^2/2(2n+1)}$$
$$=\sum_{r=-n}^{n}d_0^{\alpha,n}(x-r)\,\theta_3\left(\frac{\pi}{2n+1}(x-r-s)\,\middle|\,i\,\frac{2\pi}{(2n+1)\alpha^2}\right), \tag{40}$$

can be solved easily:

$$d_0^{\alpha,n}(x-r)=\frac{1}{2n+1}\sum_{\ell=-n}^{n}\frac{e^{i[2\pi r\ell/(2n+1)]}\Sigma_{s=-n}^{n}e^{i[2\pi s\ell/(2n+1)]}e^{\alpha^2[s/(2n+1)]x}e^{-\alpha^2s^2/2(2n+1)}\eta_n^{\alpha}\left(i\alpha^2\frac{s}{2n+1},x\right)}{\Sigma_{s=-n}^{n}e^{i[2\pi s\ell/(2n+1)]}\theta_3\left(\frac{\pi}{2n+1}(x-s)\middle|i\frac{2\pi}{(2n+1)\alpha^2}\right)} \tag{41}$$

$$=\frac{1}{(2n+1)^2}\sum_{\ell=-n}^{n}\frac{e^{[2\pi^2/(2n+1)\alpha^2]\ell^2}\Sigma_{s=-n}^{n}e^{-i[2\pi(x-r-s)\ell/(2n+1)]}e^{\alpha^2[s/(2n+1)]x}e^{-\alpha^2s^2/2(2n+1)}\eta_n^{\alpha}\left(i\alpha^2\frac{s}{2n+1},x\right)}{\theta_3\left(\pi x+\frac{i2\pi^2}{\alpha^2}\ell\middle|i\frac{2\pi(2n+1)}{\alpha^2}\right)}. \tag{42}$$

Various alternative expressions for $d_0^{\alpha,n}(x)$ can be obtained if one chooses, for instance, $k=\kappa+i\alpha^2[s/(2n+1)]$ in the derivation above. In particular, the choice $k=i\alpha^2[(x+s)/(2n+1)]$ yields the Fourier series for $d_0^{\alpha,n}(x)$.

A formula analogous to Eq. (36) can also be derived for the functions $\Delta_n^{\alpha}(k)$. The periodicity and quasiperiodicity properties,

$$\Delta_n^{\alpha}(k+2\pi)=\Delta_n^{\alpha}(k), \tag{43}$$

$$\Delta_n^{\alpha}\left(k+i\frac{\alpha^2}{2}\right)=e^{(n+1)(\alpha^2/4)}e^{-i(n+1)k}\Delta_n^{\alpha}(k), \tag{44}$$

imply that $\Delta_n^{\alpha}(k)$ may be written as follows:

$$\Delta_n^{\alpha}(k)=\sum_{r=0}^{n}b_r^{\alpha,n}e^{-[(n+1)/\alpha^2]k^2}\theta_3\left(\frac{i2\pi}{\alpha^2}k-\frac{\pi}{n+1}r\middle|i\frac{4\pi}{(n+1)\alpha^2}\right). \tag{45}$$

The coefficients $b_r^{\alpha,n}$ can be determined as before. Analogously the norm square $\|\eta_n^{\alpha}(k)\|^2$ can be expressed as a linear combination of θ functions:

$$\|\eta_n^{\alpha}(k)\|^2=\sum_{r=-n}^{n}c_r^{\alpha,n}e^{-[(2n+1)/\alpha^2]k^2}\theta_3\left(\frac{i2\pi}{\alpha^2}k-\frac{\pi}{2n+1}r\middle|i\frac{4\pi}{(2n+1)\alpha^2}\right), \tag{46}$$

with appropriate coefficients $c_r^{\alpha,n}$, respectively.

B. Another representation of the Bloch functions

To get more insight into the x dependence of the Bloch functions $\eta_n^{\alpha}(k,x)$, we derive another expression for them. To this end we need, in addition to Eqs. (28)–(31), the information how $\eta_n^{\alpha}(k,x)$ transforms if we replace x by $x+i(\alpha^2/2)$. This transformation law is more complicated than the other ones, since $\eta_n^{\alpha}(k,x)$ is not quasiperiodic in the complex x direction. We may write the Bloch functions $\eta_n^{\alpha}(k,x)$ as follows:

$$\eta_n^{\alpha}(k,x)=(-i)^n\frac{1}{n!}\left(\frac{2\pi}{\alpha}\right)^{n^2}e^{ikx}e^{-(n+1/2)k^2/\alpha^2}\sum_{\ell=0}^{n}a^{\alpha,n,\ell}(k)b^{\alpha,n,n-\ell}(k,x), \tag{47}$$

with the factors

$$b^{\alpha,n,\ell}(k,x) := e^{k^2/2\alpha^2} e^{-ikx} \sum_{m=-\infty}^{\infty} e^{-(k+2\pi m)^2/2\alpha^2} e^{i(k+2\pi m)x} m^{\ell}$$

$$= \frac{1}{(2i)^{\ell}} \frac{\partial^{\ell}}{\partial z^{\ell}} \theta_3\left(\frac{i\pi k}{\alpha^2} + \pi x + z \middle| \frac{2\pi i}{\alpha^2} \right) \Bigg|_{z=0}, \tag{48}$$

$$a^{\alpha,n,\ell}(k) := (-1)^{\ell} e^{n(k^2/\alpha^2)} \sum_{m_0=-\infty}^{\infty} \cdots \sum_{m_{n-1}=-\infty}^{\infty} \left(\prod_{j=0}^{n-1} e^{-(k+2\pi m_j)^2/\alpha^2} \right)$$

$$\times \left[\prod_{n-1 \geqslant j > \ell \geqslant 0} (m_j - m_{\ell}) \right]^2 \sum_{n-1 \geqslant j_1 > \cdots j_{\ell} \geqslant 0} m_{j_1} \cdots m_{j_{\ell}}. \tag{49}$$

One verifies easily that $a^{\alpha,n,\ell}(k)$ transforms according to the following rules:

$$a^{\alpha,n,\ell}\left(k + i\frac{\alpha^2}{2} \right) = a^{\alpha,n,\ell}(k), \tag{50}$$

$$a^{\alpha,n,\ell}(k+2\pi) = e^{[(2\pi)^2 n/\alpha^2]} e^{(4\pi nk/\alpha^2)} \sum_{s=0}^{\ell} \binom{n-\ell+s}{s} a^{\alpha,n,\ell-s}(k). \tag{51}$$

Thus, $a^{\alpha,n,\ell}(k)$ may be written as the following linear combination of θ functions and their derivatives (see Appendix A 2):

$$a^{\alpha,n,\ell}(k) = \sum_{p=0}^{n-1} e^{-[(2\pi)^2/n\alpha^2]p^2} e^{(4\pi/\alpha^2)kp} \sum_{s=0}^{\ell} \binom{n-\ell+s}{s} d_p^{\alpha,n,\ell-s}$$

$$\times \frac{1}{(2i)^s} \frac{\partial^s}{\partial z^s} \theta_3\left(-\frac{2\pi ink}{\alpha^2} + \frac{i4\pi^2 p}{\alpha^2} + z \middle| \frac{4\pi in}{\alpha^2} \right) \Bigg|_{z=0}. \tag{52}$$

Hence, $\eta_n^{\alpha}(k,x)$ reads as

$$\eta_n^{\alpha}(k,x) = (-i)^n \frac{1}{n!} \left(\frac{2\pi}{\alpha} \right)^{n^2} e^{ikx} e^{-k^2/2\alpha^2} \sum_{p=0}^{n-1} e^{-(n/\alpha^2)(k-2\pi p/n)^2} \sum_{q=0}^{n} d_p^{\alpha,n,n-q} \frac{1}{(2i)^q}$$

$$\times \frac{\partial^q}{\partial z^q} \left(\theta_3\left(-\frac{2\pi ink}{\alpha^2} + \frac{i4\pi^2 p}{\alpha^2} + z \middle| \frac{4\pi in}{\alpha^2} \right) \theta_3\left(\frac{i\pi k}{\alpha^2} + \pi x + z \middle| \frac{2\pi i}{\alpha^2} \right) \right) \Bigg|_{z=0}, \tag{53}$$

which is obtained by changing the summation indices appropriately: $\Sigma_{\ell=0}^{n} \Sigma_{s=0}^{\ell} = \Sigma_{q=0}^{n} \Sigma_{r=0}^{q}$, with $q = n-(\ell-s)$, $r = n-\ell$. The coefficients $d_p^{\alpha,n,\ell}$ read as

$$d_p^{\alpha,n,\ell} = e^{(2\pi p)^2/n\alpha^2} \sum_{m_0+\cdots+m_{n-1}=p}^{\infty} \left(\prod_{j=0}^{n-1} e^{-(2\pi m_j)^2/\alpha^2} \right)$$

$$\times \left[\prod_{n-1 \geqslant j > \ell \geqslant 0} (m_j - m_{\ell}) \right]^2 \sum_{n-1 \geqslant j_1 > \cdots > j_{\ell} \geqslant 0} m_{j_1} \cdots m_{j_{\ell}}, \tag{54}$$

which we can derive by comparing the Fourier coefficients of Eqs. (A17) and (49).

C. Expressions for Bloch functions suitable for large α

In the preceding sections we have derived various expressions for the Bloch functions $\eta_n^{\alpha}(k,x)$, but for numerical calculations these expressions are suitable only for small values of α. Thus, we want to derive some expressions for the orthogonalized Bloch functions, which are suitable for large α. This can be achieved by using Jacobi's transformation for θ functions. Note that it is not sufficient to apply Jacobi's transformation to the θ functions in Eqs. (36) and (53) since we still need Eq. (25) to determine the coefficients $d_n^{\alpha,n}(x-r)$ and $d_p^{\alpha,n,\ell}$, respectively. Using Jacobi's transformation we get the equations

$$\chi_n^\alpha(k,x)=\frac{\alpha}{\sqrt{2\pi}}\left(-\frac{1}{\alpha}\frac{\partial}{\partial x}\right)^n e^{-(\alpha^2/2)x^2}\theta_3\left(-\frac{k}{2}+i\frac{\alpha^2}{2}x\middle|\frac{i\alpha^2}{2\pi}\right), \tag{55}$$

$$=2^{-n/2}\frac{\alpha}{\sqrt{2\pi}}\sum_{m=-\infty}^{\infty} e^{ikm}H_n\left(\frac{\alpha}{\sqrt{2}}(x-m)\right)e^{-(\alpha^2/2)(x-m)^2} \tag{56}$$

and

$$\langle\chi_j^\alpha(k),\chi_\ell^\alpha(k)\rangle_1=(-1)^j\frac{\alpha}{\sqrt{4\pi}}\left(-\frac{1}{\alpha}\frac{\partial}{\partial y}\right)^{j+\ell} e^{-(\alpha^2/4)y^2}\theta_3\left(-\frac{k}{2}+i\frac{\alpha^2}{4}y\middle|\frac{i\alpha^2}{4\pi}\right)\Bigg|_{y=0}, \tag{57}$$

$$=(-1)^\ell\frac{\alpha}{\sqrt{4\pi}}2^{-j-\ell}\sum_{m=-\infty}^{\infty} e^{ikm}H_{j+\ell}\left(\frac{\alpha}{2}m\right)e^{-(\alpha^2/4)m^2}. \tag{58}$$

Thus, the orthonormalized Bloch functions $\psi_n^\alpha(k,x)=(1/\sqrt{2\pi})[\eta_n^\alpha(k,x)/\sqrt{\Delta_{n-1}^\alpha(k)\Delta_n^\alpha(k)}]$ read as

$$\psi_n^\alpha(k,x)=\frac{\sqrt{\alpha}}{\sqrt[4]{\pi}}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\Lambda_{n-1}^\alpha(k)\Lambda_n^\alpha(k)}}\sum_{m_0=-\infty}^{\infty}\cdots\sum_{m_n=-\infty}^{\infty}\left(\prod_{j=0}^{n-1} e^{ikm_j}e^{-(\alpha^2/4)m_j^2}\right)$$
$$\times e^{ikm_n}e^{-(\alpha^2/2)(x-m_n)^2}\det M_{n;m_0,\ldots,m_{n-1}}^\alpha(x-m_n), \tag{59}$$

where the $(n+1)\times(n+1)$ matrix $M_{n;m_0,\ldots,m_{n-1}}^\alpha(x-m_n)$ is defined by its matrix elements $m_{n;m_0,\ldots,m_{n-1};j,\ell}^\alpha(x-m_n)$:

$$m_{n;m_0,\ldots,m_{n-1};j,\ell}^\alpha(x-m_n)=(-1)^\ell 2^{-j-\ell}H_{j+\ell}\left(\frac{\alpha}{2}m_j\right), \tag{60}$$

for $j=0,\ldots,n-1$, $\ell=0,\ldots,n$ and

$$m_{n;m_0,\ldots,m_{n-1};n,\ell}^\alpha(x-m_n)=2^{-(\ell/2)}H_\ell\left(\frac{\alpha}{\sqrt{2}}(x-m_n)\right), \tag{61}$$

for $\ell=0,\ldots,n$, and $\Lambda_n^\alpha(k)$ is given by

$$\Lambda_n^\alpha(k)=(-1)^{n(n+1)/2}2^{-n(n+1)}$$
$$\times\sum_{m_0=-\infty}^{\infty}\cdots\sum_{m_n=-\infty}^{\infty}\left(\prod_{j=0}^{n} e^{ikm_j}e^{-(\alpha^2/4)m_j^2}\right)\det\left(H_{j+\ell}\left(\frac{\alpha}{2}m_j\right)\right)_{j,\ell=0,\ldots,n}. \tag{62}$$

It might be surprising that $\psi_n^\alpha(k,x)$ is of the form

$$\psi_n^\alpha(k,x)=\sum_{\ell=0}^{n} b_{n;\ell}^\alpha(k)\sum_{m=-\infty}^{\infty} e^{ikm}H_\ell\left(\frac{\alpha}{\sqrt{2}}(x-m)\right)e^{-(\alpha^2/2)(x-m)^2}, \tag{63}$$

whereas one would have expected that $\psi_n^\alpha(k,x)$ reads as

$$\psi_n^\alpha(k,x)=\sum_{\ell=0}^{n} c_{n;\ell}^\alpha(k)\phi_l^\alpha(k,x)=\sum_{\ell=0}^{n} c_{n;\ell}^\alpha(k)\sum_{m=-\infty}^{\infty} e^{ikm}H_\ell(\alpha(x-m))e^{-(\alpha^2/2)(x-m)^2}, \quad (64)$$

with appropriate coefficients $c_{n;l}^\alpha(k)$. This is due to the fact that we have applied the orthogonalization procedure on the functions $\chi_j^\alpha(k)$ instead on $\phi_l^\alpha(k,x)$. But Eq. (63) can be easily transformed into Eq. (64) by using the addition theorem[9]

$$\sum_{\ell=0}^{n}\binom{n}{\ell}H_\ell(2^{1/2}x)H_{n-\ell}(2^{1/2}y)=2^{n/2}H_n(x+y). \quad (65)$$

Thus the coefficients $c_{n;\ell}^\alpha(k)$ in Eq. (64) are given by

$$c_{n;\ell}^\alpha(k)=\frac{1}{\sqrt{2\pi}}\frac{(-1)^{n(n+1)/2}}{2^{n^2}}\frac{(2^\ell\ell!)^{1/2}}{N_n^\alpha(k)}\frac{1}{\sqrt{\Lambda_{n-1}^\alpha(k)\Lambda_n^\alpha(k)}}$$
$$\times\sum_{m_0=-\infty}^{\infty}\cdots\sum_{m_{n-1}=-\infty}^{\infty}\left(\prod_{j=0}^{n-1}e^{ikm_j}e^{-(\alpha^2/4)m_j^2}\right)D_{n;\ell;m_0,\ldots,m_{n-1}}^\alpha, \quad (66)$$

where $D_{n;\ell;m_0,\ldots,m_{n-1}}^\alpha$ denotes the determinant of the $(n+1)\times(n+1)$ matrix $B^\alpha(n;\ell;m_0,\ldots,m_{n-1})$, with the matrix elements

$$b_{pq}^\alpha(n,\ell,m_p)=H_{p+q}\left(\frac{\alpha}{2}m_p\right),\quad \text{for } p=0,\ldots,n-1,\quad q=0,\ldots,n, \quad (67)$$

$$b_{nq}^\alpha(n,\ell)=0,\quad \text{for } q=0,\ldots,\ell-1, \quad (68)$$

$$b_{nq}^\alpha(n,\ell)=(-1)^q\binom{q}{\ell}H_{q-\ell}(0),\quad \text{for } q=\ell,\ldots,n. \quad (69)$$

One expects that the orthogonalized Bloch functions $\psi_n^\alpha(k,x)$ are approximately the same as the nonorthogonalized Bloch functions $\phi_n^\alpha(k,x)$ in the limit of large α, and this is, in fact, true. More precisely, we have

$$c_{n;\ell}^\alpha(k)=O(\alpha^{2n-1}e^{-\alpha^2/4}),\quad \text{for } \ell<n, \quad (70)$$

$$c_{n;n}^\alpha(k)=1+O(\alpha^{2n}e^{-\alpha^2/4}). \quad (71)$$

For proving the first equation, it is sufficient to show that the determinant $D_{n;\ell;m_0,\ldots,m_{n-1}}^\alpha$ is zero for $m_0=\cdots=m_{n-1}=0$. To this end, we show that the last row of the matrix $B^\alpha(n;\ell;0,\ldots,0)$ is a linear combination of the first ℓ rows. Let $r_p:=\{H_{p+q}(0)\}$ be the pth row and let $\rho_\ell=\{0,\ldots,0,(-1)^q\binom{q}{\ell}H_{q-\ell}(0)\}$ be the last one. In the case of even ℓ, one can prove

$$\sum_{p'=0}^{\ell'}\binom{\ell'}{p'}\frac{p'!}{(2p')!}r_{2p'}=2^{2\ell'}\ell'!\rho_{2\ell'}, \quad (72)$$

where we have set $\ell=2\ell'$, $p=2p'$ and $q=2q'$. The case for odd ℓ is very similar. Defining $\ell=2\ell'+1$, $p=2p'+1$, $q=2q'+1$, the essential equation is

$$\sum_{p'=0}^{\ell'}\binom{\ell'}{p'}\frac{(p'+1)!}{(2p'+2)!}r_{2p'+1}=2^{2\ell'}\ell'!\rho_{2\ell'+1}. \quad (73)$$

Equation (71) can be proved by using the facts

$$\Lambda_n^\alpha(k)=(-1)^{n(n+1)/2}2^{-n(n+1)}\det A_n+O(\alpha^{2n}e^{-\alpha^2/4}), \tag{74}$$

$$N_n^\alpha(k)=\frac{1}{\sqrt{2\pi}}+O(\alpha^{2n}e^{-\alpha^2/4}), \tag{75}$$

$$D^\alpha_{n;n;0,\dots,0}=(-1)^{n(n-1)/2}2^{-n^2}\det A_{n-1}, \tag{76}$$

$$\det A_n=(-1)^n 2^n n!\det A_{n-1}, \tag{77}$$

where A_n is the $(n+1)\times(n+1)$ matrix $A_n := (H_{j+l}(0))_{j,l=0,\dots,n}$. Thus

$$\psi_n^\alpha(k,x)=\phi_n^\alpha(k,x)+\epsilon_n^\alpha(k,x) \tag{78}$$

hold true, where $\epsilon_n^\alpha(k,x)$ are linear combinations of $\phi_0^\alpha(k,x),\dots,\phi_{n-1}^\alpha(k,x)$, whose coefficients are all of order $O(\alpha^{2n}e^{-\alpha^2/4})$.

D. Limit $\alpha\to 0$

We know already that the Bloch functions $\psi_0^\alpha(k,x)$ converge with $\alpha\to 0$ to the free electron wave functions.[6] In fact, this is also valid for general n, i.e., the Bloch functions $\psi_n^\alpha(k,x)$ converge to the free electron wave functions corresponding to the k vectors of the $(n+1)$th Brillouin zone. To prove this, recall Eqs. (25) and (27). Obviously, in Eqs. (25) and (27) only the terms where all indices m_j are different do not vanish. In other words, only those terms are nonzero for which the values $k+2\pi m_j$ lie all in different Brillouin zones. Thus, the leading terms are those for which $k+2\pi m_0,\dots,k+2\pi m_{n-1}$ are elements of the first n Brillouin zones, and $k+2\pi m_n$ is a member of the $(n+1)$th Brillouin zone. For $n=2m$ the limit $\alpha\to 0$ of the orthonormalized Bloch function $\psi_n^\alpha(k,x)$ reads as

$$\psi_{2m}^0(k,x)=\lim_{\alpha\to 0}\psi_{2m}^\alpha(k,x)=\frac{(-1)^m}{\sqrt{2\pi}}\begin{cases} e^{i(k+2\pi m)x}, & k\in(0,\pi),\\ e^{i(k-2\pi m)x}, & k\in(-\pi,0),\\ \sqrt{2}\cos(2\pi m x), & k=0,\\ \sqrt{2}\cos(\pi(2m+1)x), & k=\pm\pi. \end{cases} \tag{79}$$

For $n=2m+1$, we have

$$\psi_{2m+1}^0(k,x)=\lim_{\alpha\to 0}\psi_{2m+1}^\alpha(k,x)=\frac{(-1)^m}{\sqrt{2\pi}}\begin{cases} ie^{i(k-2\pi(m+1))x}, & k\in(0,\pi),\\ -ie^{i(k+2\pi(m+1))x}, & k\in(-\pi,0),\\ \sqrt{2}\sin(2\pi(m+1)x), & k=0\\ \sqrt{2}\sin(\pi(2m+1)x), & k=\pm\pi. \end{cases} \tag{80}$$

Note that these functions are Bloch functions for free electrons. In particular, the limit of the Bloch function $\psi_n^\alpha(k,x)$ is a free electron wave function of the $n+1$th band. The corresponding Wannier functions are given by

$$W_{2m}^0(x)=(-1)^m\frac{2}{\pi x}\cos\left(2\pi m+\frac{\pi}{2}\right)x\sin\frac{\pi}{2}x, \tag{81}$$

$$W_{2m+1}^0(x)=(-1)^m\frac{2}{\pi x}\sin\left(2\pi m+\frac{3\pi}{2}\right)x\sin\frac{\pi}{2}x. \tag{82}$$

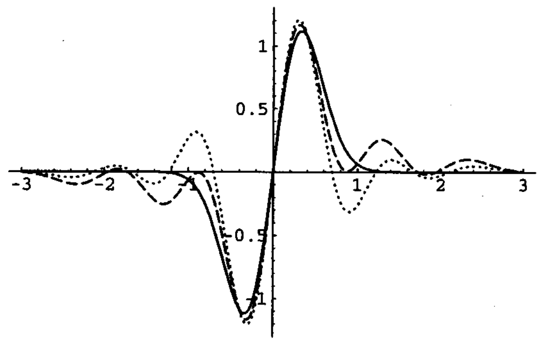

FIG. 1. Oscillator eigenfunction and Wannier functions for $n=1$ and $\alpha=3$.

E. Wannier functions

Again we use the convention

$$W_n^\alpha(m;x)=W_n^\alpha(x-m)=\frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi}dk\, e^{-ikm}\psi_n^\alpha(k,x), \tag{83}$$

for the definition of the Wannier functions. The transformation properties,

$$(\psi_n^\alpha(k,x))^*=\psi_n^\alpha(-k,x), \tag{84}$$

$$\psi_n^\alpha(k,-x)=(-1)^n\psi_n^\alpha(-k,x), \tag{85}$$

imply again the reality and (anti-)symmetry of the Wannier functions:

$$(W_n^\alpha(x))^*=W_n^\alpha(x), \tag{86}$$

$$W_n^\alpha(x)=(-1)^n W_n^\alpha(-x). \tag{87}$$

Unfortunately there is no explicit expression for the Wannier functions. In Figs. 1–6 we compare the nonorthogonal (dashed) and the orthonormalized (dotted) Wannier functions with the original oscillator eigenfunctions (solid) for $n=1$ and $n=2$. Figures for $n=0$ can be found in Ref. 6.

The orthonormalized Wannier functions have the remarkable property that the uncertainty of position $(\Delta x)_{W_n^\alpha}$ is the same as the uncertainty of the original oscillator eigenfunction $u_n^\alpha(x)$, namely, we have $(\Delta x)^2_{W_n^\alpha}=(2n+1)/2\alpha^2$. The proof is rather lengthy and thus we omit most of the

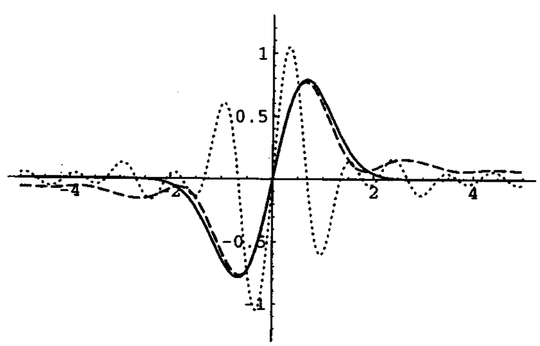

FIG. 2. Oscillator eigenfunction and Wannier functions for $n=1$ and $\alpha=1.5$.

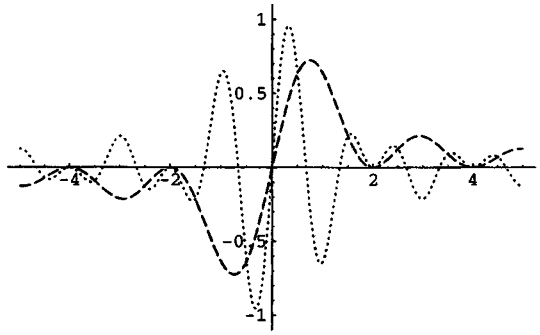

FIG. 3. Wannier functions for $n=1$ and $\alpha=0$.

details. Obviously we have $\langle x\rangle_{W_n^\alpha}=0$ and hence we need to calculate only $\langle x^2\rangle_{W_n^\alpha}$. For this purpose we express the expectation value $\langle x^2\rangle_{W_n^\alpha}$ in terms of the Bloch factor:

$$\langle x^2\rangle_{W_n^\alpha}=\int_0^1 dx\int_{-\pi}^{\pi} dk\left|\frac{\partial}{\partial k}\,\omega_n^\alpha(k,x)\right|^2, \tag{88}$$

$$\omega_n^\alpha(k,x)=e^{-ikx}\psi_n^\alpha(k,x). \tag{89}$$

According to Sec. III A the Bloch factor $\omega_n^\alpha(k,x)$ may be expressed in terms of the functions $\xi_\ell^\alpha(k,x)$,

$$\xi_\ell^\alpha(k,x)=e^{-ikx}\chi_\ell^\alpha(k,x)=(-i)^\ell\sum_{m=-\infty}^{\infty}\left(\frac{k+2\pi m}{\alpha}\right)^\ell e^{-(k+2\pi m)^2/2\alpha^2}e^{i2\pi mx}, \tag{90}$$

whose inner products are given by

$$\langle\xi_j^\alpha(k),\xi_\ell^\alpha(k)\rangle_1=(-i)^{\ell-j}\sum_{m=-\infty}^{\infty}\left(\frac{k+2\pi m}{\alpha}\right)^{j+\ell}e^{-(k+2\pi m)^2/\alpha^2}. \tag{91}$$

For writing the Bloch factor as convenient as possible we use the following notions and notations: Let $\xi_1,\ldots,\xi_n\in L^2([0,1])$; then

$$\xi_1\otimes\xi_2\otimes\cdots\otimes\xi_n\in(L^2([0,1]))^n \tag{92}$$

denotes the direct product and

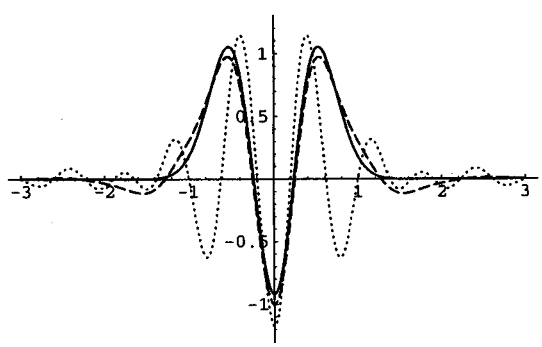

FIG. 4. Oscillator eigenfunction and Wannier functions for $n=2$ and $\alpha=3$.

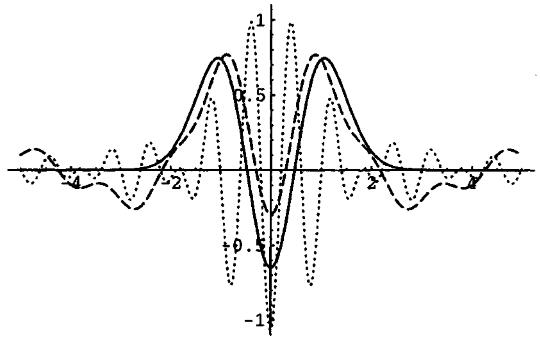

FIG. 5. Oscillator eigenfunction and Wannier functions for $n=2$ and $\alpha=1.5$.

$$\xi_1\wedge\xi_2\wedge\cdots\wedge\xi_n=\sum_{\pi\in P_n}\operatorname{sign}(\pi)\xi_{\pi(1)}\otimes\xi_{\pi(2)}\otimes\cdots\otimes\xi_{\pi(n)} \tag{93}$$

is the antisymmetrized product of the vectors $\xi_1,\ldots,\xi_n$. Here the inner product $(\cdot,\cdot)$ means the inner product of $(L^2([0,1]))^n$, where the value of n should be clear from the context. Note that

$$(\zeta_1\otimes\cdots\otimes\zeta_n,\xi_1\wedge\cdots\wedge\xi_n)=\frac{1}{n!}(\zeta_1\wedge\cdots\wedge\zeta_n,\xi_1\wedge\cdots\wedge\xi_n) \tag{94}$$

$$=(\zeta_1\wedge\cdots\wedge\zeta_n,\xi_1\otimes\cdots\otimes\xi_n). \tag{95}$$

In addition, we introduce the notation

$$(\zeta_1\otimes\cdots\otimes\zeta_n,\xi_1\wedge\cdots\wedge\xi_n)\wedge\xi_{n+1}:=\sum_{\pi\in P_{n+1}}\operatorname{sign}(\pi)(\zeta_1\otimes\cdots\otimes\zeta_n,\xi_{\pi(1)}\otimes\cdots\otimes\xi_{\pi(n)})\xi_{\pi(n+1)} \tag{96}$$

$$=\sum_{\pi\in P_{n+1}}\operatorname{sign}(\pi)\langle\zeta_1,\xi_{\pi(1)}\rangle_1\cdots\langle\zeta_n,\xi_{\pi(n)}\rangle_1\xi_{\pi(n+1)}. \tag{97}$$

In case we want to stress that $(\zeta_1\otimes\cdots\otimes\zeta_n,\xi_1\wedge\cdots\wedge\xi_n)\wedge\xi_{n+1}\in L^2([0,1])$ is a function of x, we write

$$(\zeta_1\otimes\cdots\otimes\zeta_n,\xi_1\wedge\cdots\wedge\xi_n)\wedge\xi_{n+1}(x)=[(\zeta_1\otimes\cdots\otimes\zeta_n,\xi_1\wedge\cdots\wedge\xi_n)\wedge\xi_{n+1}](x). \tag{98}$$

Thus, $\omega_n^\alpha(k,x)$ can be written as follows:

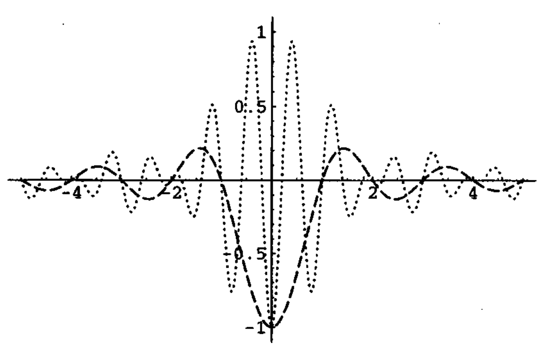

FIG. 6. Wannier functions for $n=2$ and $\alpha=0$.

$$\omega_n^\alpha(k,x)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\Delta_{n-1}^\alpha(k)\Delta_n^\alpha(k)}}(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_{n-1}^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\xi_{n-1}^\alpha(k))\wedge\xi_n^\alpha(k,x), \tag{99}$$

$$\Delta_n^\alpha(k)=(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_n^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\wedge\xi_n^\alpha(k)). \tag{100}$$

Note that $(\partial/\partial k)\xi_\ell^\alpha(k,x)$ is given by

$$\frac{\partial}{\partial k}\xi_\ell^\alpha(k,x)=-\frac{i}{\alpha}\xi_{\ell+1}^\alpha(k,x)-\frac{i\ell}{\alpha}\xi_{\ell-1}^\alpha(k,x). \tag{101}$$

Now a lengthy calculation with appropriate partial integrations yields

$$\langle x^2\rangle_{W_n^\alpha}=\frac{1}{2\pi}\int_{-\pi}^{\pi}dk(I_n^\alpha(k)+I_{n-1}^\alpha(k)), \tag{102}$$

where

$$I_n^\alpha(k):=\frac{1}{\Delta_n^\alpha(k)}\left(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_{n-1}^\alpha(k)\otimes\frac{\partial}{\partial k}\xi_n^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\wedge\xi_{n-1}^\alpha(k)\wedge\frac{\partial}{\partial k}\xi_n^\alpha(k)\right)-\frac{\dfrac{\partial^2}{\partial k^2}\Delta_n^\alpha(k)}{4\Delta_n^\alpha(k)} \tag{103}$$

$$=\frac{n+1}{2\alpha^2}+\frac{1}{2\alpha^2}\frac{1}{\Delta_n^\alpha(k)}J_n^\alpha(k). \tag{104}$$

Now one shows easily that

$$\begin{aligned}J_n^\alpha(k)=&(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_{n-1}^\alpha(k)\otimes\xi_{n+1}^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\wedge\xi_{n-1}^\alpha(k)\wedge\xi_{n+1}^\alpha(k))\\&+(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_{n-2}^\alpha(k)\otimes\xi_n^\alpha(k)\otimes\xi_{n+1}^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\wedge\xi_n^\alpha(k))\\&+(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_{n-1}^\alpha(k)\otimes\xi_{n+2}^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\wedge\xi_n^\alpha(k))=0\end{aligned} \tag{105}$$

is valid, where one takes into account

$$\det\begin{pmatrix}1&1&\cdots&1\\x_0&x_1&\cdots&x_n\\x_0^2&x_1^2&\cdots&x_n^2\\\vdots&\vdots&&\vdots\\x_0^{n-1}&x_1^{n-1}&\cdots&x_n^{n-1}\\x_0^{n+1}&x_1^{n+1}&\cdots&x_n^{n+1}\end{pmatrix}=\left(\sum_{i=0}^n x_i\right)\prod_{n\geqslant i>j\geqslant 0}(x_i-x_j), \tag{106}$$

$$\det\begin{pmatrix}1&1&\cdots&1\\x_0&x_1&\cdots&x_n\\x_0^2&x_1^2&\cdots&x_n^2\\\vdots&\vdots&&\vdots\\x_0^{n-1}&x_1^{n-1}&\cdots&x_n^{n-1}\\x_0^{n+2}&x_1^{n+2}&\cdots&x_n^{n+2}\end{pmatrix}=\left(\sum_{i=0}^n x_i^2+\sum_{n\geqslant i>j\geqslant 0}^n x_ix_j\right)\prod_{n\geqslant i>j\geqslant 0}(x_i-x_j), \tag{107}$$

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ x_0^2 & x_1^2 & \cdots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_0^{n-2} & x_1^{n-2} & \cdots & x_n^{n-2} \\ x_0^n & x_1^n & \cdots & x_n^n \\ x_0^{n+1} & x_1^{n+1} & \cdots & x_n^{n+1} \end{pmatrix} = \left(\sum_{n \geqslant i > j \geqslant 0}^{n} x_i x_j \right) \prod_{n \geqslant i > j \geqslant 0} (x_i - x_j). \tag{108}$$

Hence, $I_n^\alpha(k) = (n+1)/2\alpha^2$, and thus we have the simple result

$$\langle x^2 \rangle_{W_n^\alpha} = \frac{2n+1}{2\alpha^2}. \tag{109}$$

Note that this is also the expectation value of x^2 for the nth excited eigenstate of the harmonic oscillator. Recall that the oscillator eigenfunctions can be obtained by applying the Gram–Schmidt procedure onto the set $\{x^\ell e^{\alpha^2 x^2/2}\}_{\ell=0}^\infty$, whereas the orthonormalized Wannier functions are obtained in the following way: First construct the Wannier functions corresponding to $\{x^\ell e^{\alpha^2 x^2/2}\}_{\ell=0}^\infty$ and then orthonormalize them, a procedure that is best done in terms of Bloch functions. Although the oscillator eigenfunctions and the corresponding orthonormalized Wannier functions have been constructed in a different way and are completely different for small values of α, their uncertainty Δx is the same. This remarkable feature is quite astonishing, and naturally the question arises whether this is only pure chance or whether there is a more sophisticated reason behind it.

IV. EXPECTATION VALUES FOR THE ORTHONORMALIZED BLOCH FUNCTIONS

One of the most interesting properties of a physical system is its energy. Here we want to discuss briefly the expectation values of the kinetic energy $(1/2m)P^2$ and the potential energy V, where $V(x) = V(x+1)$ is a periodic potential.

A. Kinetic energy

1. Expectation value

The expectation value of the kinetic energy reads as

$$\begin{aligned} \langle P^2 \rangle_n^\alpha(k) &= 2\pi \langle \psi_n^\alpha(k), P^2 \psi_n^\alpha(k) \rangle_1 \\ &= -\frac{\alpha^2}{\Delta_{n-1}^a(k)} (\chi_0^\alpha(k) \otimes \cdots \otimes \chi_{n-1}^\alpha(k), \chi_0^\alpha(k) \wedge \cdots \wedge \chi_{n-3}^\alpha(k) \wedge \chi_{n-1}^\alpha(k) \wedge \chi_n^\alpha(k)) \\ &\quad + \frac{\alpha^2}{\Delta_{n-1}^\alpha(k) \Delta_n^\alpha(k)} (\chi_0^\alpha(k) \otimes \cdots \otimes \chi_{n-1}^\alpha(k), \chi_0^\alpha(k) \wedge \cdots \wedge \chi_{n-2}^\alpha(k) \wedge \chi_n^\alpha(k)) \\ &\quad \times (\chi_0^\alpha(k) \otimes \cdots \otimes \chi_n^\alpha(k), \chi_0^\alpha(k) \wedge \cdots \wedge \chi_{n-1}^\alpha(k) \wedge \chi_{n+1}^\alpha(k)) \\ &\quad - \frac{\alpha^2}{\Delta_n^\alpha(k)} (\chi_0^\alpha(k) \otimes \cdots \otimes \chi_n^\alpha(k), \chi_0^\alpha(k) \wedge \cdots \wedge \chi_{n-1}^\alpha(k) \wedge \chi_{n+2}^\alpha(k)). \end{aligned} \tag{110}$$

Of course, $\langle P^2 \rangle_n^\alpha(k) = \langle P^2 \rangle_n^\alpha(-k)$ is symmetric. For $n=0$ one can show that $\langle P^2 \rangle_0^\alpha(k)$ is monotone increasing for $k \in [-\pi, \pi]$. For general n, numerical calculations suggest that $\langle P^2 \rangle_n^\alpha(k)$ is monotone increasing or decreasing, depending on whether n is even or odd. However, the proof of the case $n=0$ cannot be easily generalized.

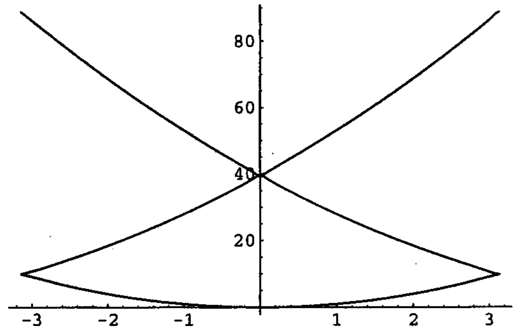

FIG. 7. Expectation values $\langle P^2\rangle_n^\alpha$ for $n=0,1,2$ and $\alpha=1$.

2. Further matrix elements

One can easily compute all the matrix elements $\langle\psi_m^\alpha(k),P^2\psi_n^\alpha(k)\rangle_1$. We have $\langle\psi_m^\alpha(k),P^2\psi_n^\alpha(k)\rangle_1=0$ for $|m-n|>2$. For $m=n+1$ and $m=n+2$ the matrix elements read as

$$\langle\psi_{n+1}^\alpha(k),P^2\psi_n^\alpha(k)\rangle_1 = \frac{1}{2\pi}\frac{\alpha^2}{\Delta_{n+1}^\alpha(k)}\Bigg[\sqrt{\frac{\Delta_{n+2}^\alpha(k)}{\Delta_n^\alpha(k)}}(\chi_0^\alpha(k)\otimes\cdots\otimes\chi_{n-1}^\alpha(k),\chi_0^\alpha(k)\wedge\cdots\wedge\chi_{n-2}^\alpha(k)\wedge\chi_n^\alpha(k)) - \sqrt{\frac{\Delta_n^\alpha(k)}{\Delta_{n+2}^\alpha(k)}}(\chi_0^\alpha(k)\otimes\cdots\otimes\chi_{n+1}^\alpha(k),\chi_0^\alpha(k)\wedge\cdots\wedge\chi_n^\alpha(k)\wedge\chi_{n+2}^\alpha(k))\Bigg], \tag{111}$$

$$\langle\psi_{n+2}^\alpha(k),P^2\psi_n^\alpha(k)\rangle_1 = -\alpha^2\sqrt{\frac{\Delta_n^\alpha(k)\Delta_{n+3}^\alpha(k)}{\Delta_{n+1}^\alpha(k)\Delta_{n+2}^\alpha(k)}}. \tag{112}$$

Figures 7–10 show the expectation values of the kinetic energy $\langle P^2\rangle_n^\alpha(k)$ and the first two or three eigenvalues of the matrices $\{\langle\psi_m^\alpha(k),P^2\psi_n^\alpha(k)\rangle\}$, where $m,n=0,1$ (2×2 matrix), $m,n=0,1,2$ (3×3 matrix), and $m,n=0,1,2,3$ (4×4 matrix), respectively. In the limit $\alpha\to0$, $\langle P^2\rangle_n^\alpha(k)$ converges to the free electron energy parabola, and hence the expectation values $\langle P^2\rangle_n^\alpha(k)$ are a very good approximation for the exact eigenvalues for small α; see Fig. 7 for $\alpha=1$. With increasing α, the difference between the correct eigenvalues of P^2 and the expectation value $\langle P^2\rangle_n^\alpha(k)$ becomes larger (see Figs. 8–10), and thus more and more matrix elements have to be taken into account if one wants to get a good approximation for the correct eigenvalues of P^2; see Figs. 8–10. Note the band gap structure for the approximate eigenvalues.

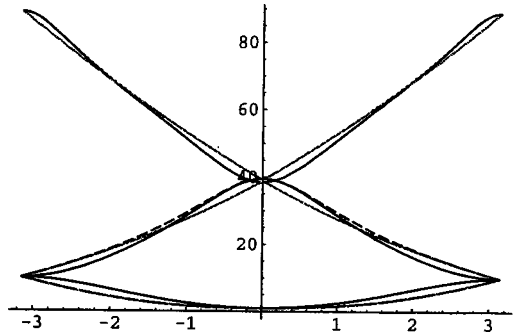

FIG. 8. Expectation values $\langle P^2\rangle_n^\alpha$ (black) for $n=0,1,2$ and $\alpha=3$ the eigenvalues of the corresponding 2×2- (dashed), 3×3- (dotted) and 4×4-matrix (grey).

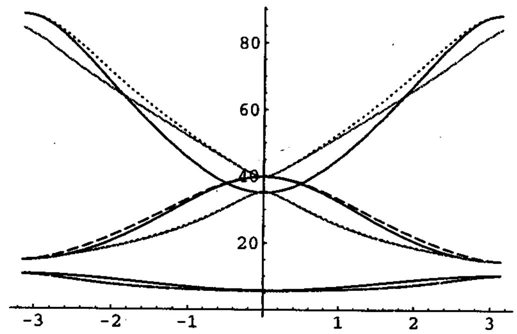

FIG. 9. Expectation values $\langle P^2\rangle_n^\alpha$ (black) for $n=0,1,2$ and $\alpha=4$ the eigenvalues of the corresponding 2×2- (dashed), 3×3- (dotted) and 4×4-matrix (grey).

B. Potential energy

Another interesting quantity is the expectation value for a periodic potential V. Let $V(x) =\Sigma_{p=-\infty}^{\infty} c_p e^{i2\pi px}$ be the Fourier series of the potential V. Then we have

$$\langle \chi_j^\alpha(k), V\chi_\ell^\alpha(k)\rangle = \sum_{s=0}^{j}\sum_{t=0}^{\ell}\binom{j}{s}\binom{\ell}{t} i^{t+s}[\langle \chi_{j-s}^\alpha(k),\chi_{\ell-t}^\alpha(k)\rangle_1 V_{s+t}^+(\alpha) + \langle \chi_{j-s}^\alpha(k+\pi),\chi_{\ell-t}^\alpha(k+\pi)\rangle_1 V_{s+t}^-(\alpha)], \tag{113}$$

where we have employed the definitions

$$V_s^+(\alpha) := \sum_{q=-\infty}^{\infty} \frac{(2\pi q)^s}{\alpha^s} c_{2q} e^{-(2\pi q)^2/\alpha^2}, \tag{114}$$

$$V_s^-(\alpha) := \sum_{q=-\infty}^{\infty} \frac{(\pi(2q+1))^s}{\alpha^s} c_{2q+1} e^{-(\pi(2q+1))^2/\alpha^2}. \tag{115}$$

Thus, the expectation value of the potential energy reads as

$$\langle V\rangle_n^\alpha = 2\pi\langle \psi_n^\alpha(k), V\psi_n^\alpha(k)\rangle_1 = \sum_{s=0}^{2n} (d_s^+(k,\alpha,n)V_s^+(\alpha) + d_s^-(k,\alpha,n)V_s^-(\alpha)), \tag{116}$$

where the coefficients $d_s^+(k,\alpha,n)$ and $d_s^-(k,\alpha,n)$ are given by

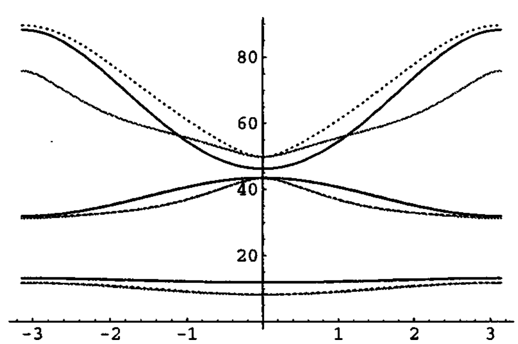

FIG. 10. Expectation values $\langle P^2\rangle_n^\alpha$ (black) for $n=0,1,2$ and $\alpha=5$ the eigenvalues of the corresponding 2×2- (dashed), 3×3- (dotted) and 4×4-matrix (grey).

$$d_s^+(k,\alpha,n):=\frac{1}{\Delta_{n-1}^\alpha(k)\Delta_n^\alpha(k)}\sum_{j,\ell=0}^{n}(-1)^{j+\ell}\sum_{t=\max(s-j,0)}^{\min(\ell,s)}\binom{j}{s-t}\binom{\ell}{t}i^s$$
$$\times(\chi_0^\alpha(k)\otimes\cdots\otimes\chi_{j-1}^\alpha(k)\otimes\chi_{j+1}^\alpha(k)\otimes\cdots\otimes\chi_n^\alpha(k),\chi_0^\alpha(k)\wedge\cdots\wedge\chi_{n-1}^\alpha(k))$$
$$\times(\chi_0^\alpha(k)\otimes\cdots\otimes\chi_{n-1}^\alpha(k),\chi_0^\alpha(k)\wedge\cdots\wedge\chi_{\ell-1}^\alpha(k)\wedge\chi_{\ell+1}^\alpha(k)\wedge\cdots\wedge\chi_n^\alpha(k))$$
$$\times\langle\chi_{j-s+t}^\alpha(k),\chi_{\ell-t}^\alpha(k)\rangle_1, \tag{117}$$

$$d_s^-(k,\alpha,n):=\frac{1}{\Delta_{n-1}^\alpha(k)\Delta_n^\alpha(k)}\sum_{j,\ell=0}^{n}(-1)^{j+\ell}\sum_{t=\max(s-j,0)}^{\min(\ell,s)}\binom{j}{s-t}\binom{\ell}{t}i^s$$
$$\times(\chi_0^\alpha(k)\otimes\cdots\otimes\chi_{j-1}^\alpha(k)\otimes\chi_{j+1}^\alpha(k)\otimes\cdots\otimes\chi_n^\alpha(k),\chi_0^\alpha(k)\wedge\cdots\wedge\chi_{n-1}^\alpha(k))$$
$$\times(\chi_0^\alpha(k)\otimes\cdots\otimes\chi_{n-1}^\alpha(k),\chi_0^\alpha(k)\wedge\cdots\wedge\chi_{\ell-1}^\alpha(k)\wedge\chi_{\ell+1}^\alpha(k)\wedge\cdots\wedge\chi_n^\alpha(k))$$
$$\times\langle\chi_{j-s+t}^\alpha(k+\pi),\chi_{\ell-t}^\alpha(k+\pi)\rangle_1. \tag{118}$$

Note that the potential enters into the expression for the expectation value $\langle V\rangle_n^\alpha$ only via the $2n+2$ constants $V_s^+(\alpha)$, $V_s^-(\alpha)$, whereas only $d_s^+(k,\alpha,n)$ and $d_s^-(k,\alpha,n)$ depend on k. For special values of s, the functions $d_s^+(k,\alpha,n)$ and $d_s^-(k,\alpha,n)$ simplify considerably:

$$d_{2n}^+(k,\alpha,n)=(-1)^n\frac{\Delta_{n-1}^\alpha(k)}{\Delta_n^\alpha(k)}\langle\chi_0^\alpha(k),\chi_0^\alpha(k)\rangle_1, \tag{119}$$

$$d_{2n}^-(k,\alpha,n)=(-1)^n\frac{\Delta_{n-1}^\alpha(k)}{\Delta_n^\alpha(k)}\langle\chi_0^\alpha(k+\pi),\chi_0^\alpha(k+\pi)\rangle_1, \tag{120}$$

$$d_0^+(k,\alpha,n)=1. \tag{121}$$

V. FURTHER PROPERTIES OF THE WANNIER FUNCTIONS

In a previous section we have seen that the uncertainty $\Delta x=\sqrt{(2n+1)/2\alpha^2}$ is the same for the orthonormalized Wannier functions and the original oscillator eigenfunctions. Here we show that an analogous result is not true for the uncertainty of momentum Δp.

Of course, $\langle P\rangle_{W_n^\alpha}=0$ due to symmetry. The expectation value $\langle P^2\rangle_{W_n^\alpha}$ can be computed with the same methods used in Sec. III E, if one starts with Eq. (110) and expresses $\langle P^2\rangle_n^\alpha(k)$ in terms of ξ_n^α instead of $\chi_n^\alpha(k)$. Then we get the result

$$\langle P^2\rangle_{W_n^\alpha}=\alpha^2\left(n+\frac{1}{2}\right)+\frac{1}{2\pi}\frac{\alpha^4}{8}\int_{-\pi}^{\pi}dk\left[\frac{\frac{\partial}{\partial k}\Delta_{n-1}^\alpha(k)}{\Delta_{n-1}^\alpha(k)}-\frac{\frac{\partial}{\partial k}\Delta_n^\alpha(k)}{\Delta_n^\alpha(k)}\right]^2$$
$$+\frac{1}{2\pi}\frac{\alpha^4}{8}\int_{-\pi}^{\pi}dk\left\{\left[\frac{\frac{\partial}{\partial k}\Delta_{n-1}^\alpha(k)}{\Delta_{n-1}^\alpha(k)}\right]^2+\left[\frac{\frac{\partial}{\partial k}\Delta_n^\alpha(k)}{\Delta_n^\alpha(k)}\right]^2\right\}+\frac{1}{2\pi}\int_{-\pi}^{\pi}dk\,K_n^\alpha(k), \tag{122}$$

where $K_n^\alpha(k)$ is defined by

$$K_n^\alpha(k)=\frac{\alpha^2}{\Delta_{n-1}^\alpha(k)}(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_{n-2}^\alpha(k)\otimes\xi_{n+1}^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\wedge\xi_{n-1}^\alpha(k))$$
$$+\frac{\alpha^2}{\Delta_n^\alpha(k)}(\xi_0^\alpha(k)\otimes\cdots\otimes\xi_{n-2}^\alpha(k)\otimes\xi_n^\alpha(k)\otimes\xi_{n+1}^\alpha(k),\xi_0^\alpha(k)\wedge\cdots\wedge\xi_n^\alpha(k)). \tag{123}$$

The first term in Eq. (122) is the expectation value $\langle P^2\rangle_{u_n^\alpha}$ for the original oscillator eigenfunctions, the second and third term are obviously positive, and also the last term seems to be positive, since numerical calculations suggest that $K_n^\alpha(k)\geqslant 0$. Hence

$$\langle P^2\rangle_{W_n^\alpha}>\alpha^2(n+\tfrac{1}{2}), \tag{124}$$

which means that the uncertainty of momentum Δp for the orthonormalized Wannier functions is larger than the corresponding uncertainty for the original oscillator eigenfunctions. Of course, an analogous statement is valid for the uncertainty product:

$$(\Delta x\,\Delta p)_{W_n^\alpha}>(\Delta x\,\Delta p)_{u_n^\alpha}=n+\tfrac{1}{2}. \tag{125}$$

In the special case $n=0$, we can calculate the expectation value $\langle P^2\rangle_{W_0^\alpha}$ explicitly:

$$\langle P^2\rangle_{W_0^\alpha}=\frac{\alpha^2}{2}+\frac{1}{2\pi}\frac{\alpha^4}{4}\int_{-\pi}^{\pi}dk\left[\frac{\frac{\partial}{\partial k}\theta_3\left(\frac{k}{2}\middle|\frac{i\alpha^2}{4\pi}\right)}{\theta_3\left(\frac{k}{2}\middle|\frac{i\alpha^2}{4\pi}\right)}\right]^2=\frac{\alpha^2}{2}+\frac{\alpha^4}{192}\left[\frac{\theta_1'''\left(0\middle|\frac{i\alpha^2}{4\pi}\right)}{\theta_1'\left(0\middle|\frac{i\alpha^2}{4\pi}\right)}+1\right]. \tag{126}$$

For the limits of large and small α we obtain the following simple formulas:

$$\langle P^2\rangle_{W_0^\alpha}=\frac{\alpha^2}{2}+O(\alpha^4e^{-\alpha^2/2}),\quad\text{for }\alpha\to\infty, \tag{127}$$

$$\langle P^2\rangle_{W_0^\alpha}=\frac{\pi^2}{3}+O(\alpha^4),\quad\text{for }\alpha\to 0. \tag{128}$$

The corresponding uncertainty products read as

$$(\Delta x\,\Delta p)_{W_0^\alpha}=\tfrac{1}{2}+O(\alpha^2e^{-\alpha^2/2}),\quad\text{for }\alpha\to\infty, \tag{129}$$

$$(\Delta x\,\Delta p)_{W_0^\alpha}=\frac{\pi}{\sqrt{6}\,\alpha}+O(\alpha^2),\quad\text{for }\alpha\to 0. \tag{130}$$

VI. CONCLUSIONS

We have used the eigenfunctions of the harmonic oscillator to construct a complete set of mutually orthonormal Bloch and Wannier functions. The interesting aspect for solid state physics is that these Bloch functions provide an expansion for arbitrary Bloch functions, which might go far beyond the usual expansions in terms of Gaussian orbitals. In the limit $\alpha\to 0$, these Bloch functions converge to the plane wave eigenfunctions of the free electron. Thus, the free electron wave functions may be viewed as a special case of the Bloch functions $\psi_n^\alpha(k,x)$. Hence the Bloch functions $\psi_n^\alpha(k,x)$ represent an alternative complete set of mutually orthonormal Bloch functions, which is appropriate for nearly free electrons in the case of small α. On the other hand, for large α the Bloch functions $\psi_n^\alpha(k,x)$ are typical Bloch functions for tight-binding approximations. Thus, depending on the choice of α, these Bloch functions could be useful both in tight-binding approximations as well as in the case of nearly free electrons. In addition, we have derived several expressions for these Bloch functions and calculated certain expectation values. Moreover, we have discussed the localization properties of the corresponding Wannier functions.

ACKNOWLEDGMENT

The present work was financially supported by Österreichische Nationalbank under Grant No. 5917.

APPENDIX A: SOME FACTS RELATED TO THETA FUNCTIONS

1. A function related to theta functions

In this section we show that an analytic function $\eta_n^\alpha(k,x)$ with the quasiperiodicity properties,

$$\eta_n^\alpha(k+2\pi,x)=\eta_n^\alpha(k,x), \tag{A1}$$

$$\eta_n^\alpha(k+i\alpha^2,x)=e^{(n+1/2)\alpha^2}e^{-\alpha^2 x}e^{-(2n+1)ik}\eta_n^\alpha(k,x), \tag{A2}$$

can be expressed in terms of theta functions. Due to periodicity $\eta_n^\alpha(k,x)$ can be expanded into a Fourier series:

$$\eta_n^\alpha(k,x)=\sum_{m=-\infty}^{\infty} c_m^{\alpha,n}(x)e^{ikm}. \tag{A3}$$

Taking Eq. (A2) into account we obtain the following relation for the Fourier coefficients:

$$e^{-\alpha^2 m}c_m^{\alpha,n}(x)=e^{(n+1/2)\alpha^2}e^{-\alpha^2 x}c_{m+2n+1}^{\alpha,n}(x). \tag{A4}$$

The general solution of these equations is given by

$$c_{r+s(2n+1)}^{\alpha,n}(x)=e^{[-\alpha^2/2(2n+1)](r+s(2n+1)-x)^2}d_r^{\alpha,n}(x), \tag{A5}$$

where $r=-n,\dots,n$. Thus, $\eta_n^\alpha(k,x)$ can be expressed as a sum of $2n+1$ θ functions:

$$\eta_n^\alpha(k,x)=\sum_{r=-n}^{n} d_r^{\alpha,n}(x)e^{-[\alpha^2/2(2n+1)](r-x)^2}e^{irk}\theta_3\left(\frac{2n+1}{2}k+i\frac{\alpha^2}{2}(r-x)\middle| i\frac{(2n+1)\alpha^2}{2\pi}\right). \tag{A6}$$

By applying Jacobi's transformation, we get

$$\eta_n^\alpha(k,x)=\sqrt{\frac{2\pi}{(2n+1)\alpha^2}}e^{ikx}\sum_{r=-n}^{n} d_r^{\alpha,n}(x)$$
$$\times e^{-[(2n+1)/2\alpha^2]k^2}\theta_3\left(\frac{i\pi}{\alpha^2}k-\frac{\pi}{2n+1}(r-x)\middle| i\frac{2\pi}{(2n+1)\alpha^2}\right). \tag{A7}$$

2. A function related to derivatives of theta functions

Let $a^{\alpha,n,\ell}(k)$ be an analytic function obeying the following transformation rules:

$$a^{\alpha,n,\ell}\left(k+i\frac{\alpha^2}{2}\right)=a^{\alpha,n,\ell}(k), \tag{A8}$$

$$a^{\alpha,n,\ell}(k+2\pi)=e^{(2\pi)^2 n/\alpha^2}e^{4\pi nk/\alpha^2}\sum_{s=0}^{\ell}\binom{n-\ell+s}{s}a^{\alpha,n,\ell-s}(k). \tag{A9}$$

Then this function may be expressed in terms of θ functions and their derivatives. From Eq. (A8), we infer that $a^{\alpha,n,\ell}(k)$ may be expanded into the following Fourier series:

$$a^{\alpha,n,\ell}(k)=\sum_{m=-\infty}^{\infty}c_m^{\alpha,n,\ell}e^{(4\pi/\alpha^2)km}. \tag{A10}$$

Due to Eq. (A9) we have:

$$c_{m+n}^{\alpha,n,\ell}=e^{-(4\pi^2/\alpha^2)(2m+n)}\sum_{s=0}^{\ell}\binom{n-\ell+s}{s}c_m^{\alpha,n,\ell-s}. \tag{A11}$$

These equations can be solved recursively and it is easily seen that the solution of these equations is unique for given starting values $c_m^{\alpha,n,\ell-s}$, $s=0,...,\ell$, $m=0,...,n-1$. Using the ansatz

$$c_m^{\alpha,n,\ell}=e^{-[(2\pi)^2/n\alpha^2]m^2}d_m^{\alpha,n,\ell}, \tag{A12}$$

we obtain the simpler equations

$$d_{m+n}^{\alpha,n,\ell}=\sum_{s=0}^{\ell}\binom{n-\ell+s}{s}d_m^{\alpha,n,\ell-s}. \tag{A13}$$

Now we use the ansatz

$$d_{p+qn}^{\alpha,n,\ell}=\sum_{s=0}^{\ell}\binom{n-\ell+s}{s}q^s f_p^{\alpha,n,\ell,s}, \tag{A14}$$

where $f_p^{\alpha,n,\ell,0}=d_p^{\alpha,n,\ell}$. Then Eq. (A13) is fulfilled if we set

$$f_p^{\alpha,n,\ell,s}=f_p^{\alpha,n,\ell-s,0}=d_p^{\alpha,n,\ell-s}. \tag{A15}$$

Since the solution is unique, we have

$$d_{p+qn}^{\alpha,n,\ell}=\sum_{s=0}^{\ell}\binom{n-\ell+s}{s}q^s d_p^{\alpha,n,\ell-s}, \tag{A16}$$

where $p=0,...,n-1$. Hence, $a^{\alpha,n,\ell}(k)$ may be written in the following way:

$$a^{\alpha,n,\ell}(k)=\sum_{p=0}^{n-1}\sum_{q=-\infty}^{\infty}\sum_{s=0}^{\ell}\binom{n-\ell+s}{s}d_p^{\alpha,n,\ell-s}q^s e^{-[(2\pi)^2/n\alpha^2](p+qn)^2}e^{(4\pi/\alpha^2)k(p+qn)}$$

$$=\sum_{p=0}^{n-1}e^{-[(2\pi)^2/n\alpha^2]p^2}e^{(4\pi/\alpha^2)kp}\sum_{s=0}^{\ell}\binom{n-\ell+s}{s}d_p^{\alpha,n,\ell-s} \tag{A17}$$

$$\times\left(\frac{\alpha^2}{4\pi n}\frac{\partial}{\partial k}\right)^s\theta_3\left(-\frac{2\pi ink}{\alpha^2}+\frac{i4\pi^2 p}{\alpha^2}\middle|\frac{4\pi in}{\alpha^2}\right). \tag{A18}$$

[1] P. W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).
[2] F. Bloch, Z. Phys. **52**, 555 (1928).
[3] E. E. Wittmer and L. Rosenfeld Z. Phys. **48**, 530 (1928).
[4] G. H. Wannier, Phys. Rev. **52**, 191 (1937).

[5] G. H. Wannier, J. Math. Phys. **19**, 131 (1978).
[6] P. Zeiner, R. Dirl, and B. L. Davies, J. Math. Phys. **39**, 2437 (1998).
[7] P. Zeiner, R. Dirl, and B. L. Davies, Phys. Rev. B **58**, 7681 (1998).
[8] H. Karabulut and E. L. Sibert, J. Math. Phys. **38**, 4815 (1997).
[9] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953).
[10] D. F. Lawden, *Elliptic Functions and Applications* (Springer-Verlag, New York, 1989).
[11] H. Rademacher, *Topics in Analytic Number Theory* (Springer-Verlag, Berlin, 1973).
[12] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, Cambridge, 1962).
[13] J. Zak, Phys. Rev. B **26**, 3010 (1982).
[14] H. Bacry, L. Michel, and J. Zak, *Symmetry and Classification of Energy Bands in Crystals*, in Lecture Notes in Physics (Springer-Verlag, Berlin, 1988), Vol. 313, p. 291.
[15] L. Michel and J. Zak, Europhys. Lett. **18**, 239 (1992), and references therein.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Symmetries of discrete dynamical systems involving two species

D. Gómez-Ullate,[a)]
Departamento de Física Teórica II, Facultad de Ciencias Físicas, 28040 Universidad Complutense, Madrid, Spain

S. Lafortune[b)] and P. Winternitz,[c)]
Centre de Recherches Mathématiques, Université de Montréal, C. P. 6128, Succ. Centre-ville, Montréal (QC) H3C 3J7, Canada

(Received 24 December 1998; accepted for publication 4 March 1999)

The Lie point symmetries of a coupled system of two nonlinear differential-difference equations are investigated. It is shown that in special cases the symmetry group can be infinite dimensional, in other cases up to ten dimensional. The equations can describe the interaction of two long molecular chains, each involving one type of atoms. © *1999 American Institute of Physics.* [S0022-2488(99)03206-5]

I. INTRODUCTION

Our purpose in this article is to perform a symmetry analysis of a system of two coupled differential-difference equations of the form

$$E_1=\ddot{u}_n-F_n(t,u_{n-1},u_n,u_{n+1},v_{n-1},v_n,v_{n+1})=0,$$
$$E_2=\ddot{v}_n-G_n(t,u_{n-1},u_n,u_{n+1},v_{n-1},v_n,v_{n+1})=0. \tag{1.1}$$

The overdots denote time derivatives. The discrete variable n plays the role of a space variable; it labels positions along a one-dimensional lattice. The functions F_n and G_n represent interactions, e.g., between different atoms along a double chain of molecules (see Fig. 1). The functions F_n and G_n are *a priori* unspecified; our aim is to classify equations of the type (1.1) according to the Lie point symmetries that they allow. The interactions in such a model depend on up to six neighboring particles. For instance, we can interpret u_n and v_n as deviations from equilibrium positions of two different types of atoms, say type U and type V. The accelerations $\ddot{u}_n$ and $\ddot{v}_n$ depend on the deviations u and v of both types of atoms at the neighboring sites $n-1$, n, and $n+1$. We do not restrict to two-body forces, nor do we impose translational invariance for the chain. We do, however, assume there is no dissipation, i.e., system (1.1) does not involve first derivatives with respect to time.

Such differential-difference equations typically arise when modeling phenomena in molecular physics, biophysics, or simply coupled oscillations in classical mechanics.[1–3]

A recent article[4] was devoted to a similar problem, but was concerned with a single species, i.e., one dependent variable $u_n(t)$. The approach adopted here is similar to that of Ref. 4. Thus, we shall consider only symmetries acting on the continuous variables t, u_n, and v_n. Transformations of the discrete variable n must then be studied separately.

Several different treatments of Lie symmetries of difference and differential-difference equations exist in the literature.[4–13] The one adopted in this article is that of Refs. 4–6. It has been

[a)]Electronic mail: dgu@eucmos.sim.ucm.es
[b)]Electronic mail: lafortus@crm.umontreal.ca
[c)]Electronic mail: wintern@crm.umontreal.ca

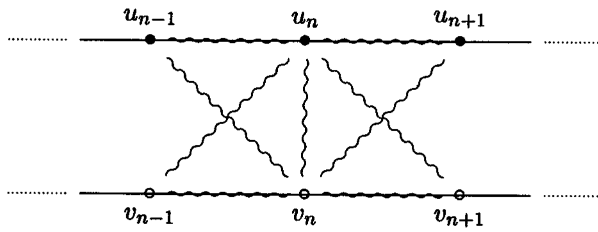

FIG. 1. Double molecular chain with two types of atoms.

called the ''intrinsic method,'' makes use of a Lie algebraic approach, and is entirely algorithmic. The Lie algebra of the symmetry group, the ''symmetry algebra'' for short, is realized by vector fields of the form

$$\hat{X}=\tau(t,u_n,v_n)\partial_t+\phi_n(t,u_n,v_n)\partial_{u_n}+\psi_n(t,u_n,v_n)\partial_{v_n}. \tag{1.2}$$

The algorithm for finding the functions τ, ϕ_n, and ψ_n in (1.2) is to construct the appropriate prolongation $\mathrm{pr}\,\hat{X}$ of $\hat{X}$ (see Refs. 4–6 and Sec. II) and to impose that it should annihilate the studied system of equations on their solution set,

$$\mathrm{pr}\,\hat{X}E_1|_{E_1=E_2=0}=0, \quad \mathrm{pr}\,\hat{X}E_2|_{E_1=E_2=0}=0. \tag{1.3}$$

Our first step is to find and classify all interactions (F_n,G_n) for which the system (1.1) allows at least a one-dimensional symmetry algebra. The next step is to specify the interactions further and to find all those that allow a higher-dimensional, possibly infinite-dimensional, symmetry algebra.

As in previous articles,[4,14] our classification will be up to conjugacy under a group of ''allowed transformations.'' These are fiber preserving locally invertible point transformations,

$$u_n=\Omega_n(\tilde{u}_n,\tilde{v}_n,\tilde{t}), \quad v_n=\Gamma_n(\tilde{u}_n,\tilde{v}_n,\tilde{t}), \quad t=t(\tilde{t}), \tag{1.4}$$

which preserve the form of Eqs. (1.1), but not necessarily the functions F_n and G_n (they go into new functions $\tilde{F}_n$ and $\tilde{G}_n$ of the new arguments).

Throughout the article we assume that both F_n and G_n depend on at least one of the quantities $u_{n-1},u_{n+1},v_{n-1},v_{n+1}$, so that nearest neighbors are genuinely involved. In the bulk of the article the interaction is assumed to be nonlinear.

In Sec. II we formulate the problem, establish the general form of the elements of the symmetry algebra, and present the determining equations for the symmetries. We also derive the ''allowed transformations'' under which we classify the interactions and their symmetries. Section III is devoted to a classification of interactions F_n, G_n, allowing at least a one-dimensional symmetry algebra. Ten classes of such interactions exist, each involving two arbitrary functions of six variables. In Sec. IV we study higher-dimensional symmetry algebras and introduce an important restriction. We first prove that four equivalence classes of symmetry algebras isomorphic to $\mathrm{sl}(2,\mathbb{R})$ exist. Then we restrict to just one of them, $\mathrm{sl}(2,\mathbb{R})_1$ generating a gauge group acting only on the fields u_n and v_n (in a global, coordinate-independent manner). We describe all symmetry algebras, containing the chosen $\mathrm{sl}(2,\mathbb{R})$ as a subalgebra. In Sec. V we obtain the invariant interactions for all algebras containing $\mathrm{sl}(2,\mathbb{R})_1$. The results are summed up and discussed in Sec. VI, where we also outline future work to be done.

II. FORMULATION OF THE PROBLEM

To find the Lie point symmetries of the system (1.1), we write the second prolongation of the vector field (1.2) in the form[4–6]

$$\mathrm{pr}^{(2)}\hat{X}=\tau(t,u_n,v_n)\partial_t+\sum_{k=n-1}^{n+1}\phi_k(t,u_n,v_n)\partial_{u_k}+\sum_{k=n-1}^{n+1}\psi_k(t,u_n,v_n)\partial_{v_k}+\phi_n^{tt}\partial_{\ddot{u}_n}+\psi_n^{tt}\partial_{\ddot{v}_n}, \tag{2.1}$$

with

$$\begin{aligned}\phi_n^{tt}&=D_t^2\phi_n-(D_t^2\tau)\dot{u}_n-2(D_t\tau)\ddot{u}_n,\\ \psi_n^{tt}&=D_t^2\psi_n-(D_t^2\tau)\dot{v}_n-2(D_t\tau)\ddot{v}_n,\end{aligned} \tag{2.2}$$

where D_t is the total time derivative. The determining equations for the symmetries are obtained by requiring that Eq. (1.3) be satisfied. The obtained equations will involve terms like $\dot{u}^k$, $\dot{v}^k$, and $\dot{u}^k\dot{v}^l$. The coefficients of each linearly independent term must vanish and this provides 16 linear differential equations that are easy to solve and do not involve the interaction functions F_n,G_n. The result is that an element $\hat{X}$ of the symmetry algebra must have the form

$$\hat{X}=\tau(t)\partial_t+\left[\left(\frac{\dot{\tau}}{2}+a_n\right)u_n+b_nv_n+\lambda_n(t)\right]\partial_{u_n}+\left[c_nu_n+\left(\frac{\dot{\tau}}{2}+d_n\right)v_n+\mu_n(t)\right]\partial_{v_n}, \tag{2.3}$$

where the overdots denote time derivatives. The functions $\tau(t)$, $\lambda_n(t)$, $\mu_n(t)$, a_n, b_n, c_n, and d_n satisfy the two remaining determining equations, namely,

$$\begin{aligned}&\frac{\dddot{\tau}}{2}u_n+\ddot{\lambda}_n+\left(a_n-\frac{3}{2}\dot{\tau}\right)F_n+b_nG_n-\tau F_{n,t}-\sum_{k=n-1}^{n+1}F_{n,u_k}\left[\left(\frac{\dot{\tau}}{2}+a_k\right)u_k+b_kv_k+\lambda_k(t)\right]\\ &\quad-\sum_{k=n-1}^{n+1}F_{n,v_k}\left[\left(\frac{\dot{\tau}}{2}+d_k\right)v_k+c_ku_k+\mu_k(t)\right]=0,\end{aligned} \tag{2.4}$$

$$\begin{aligned}&\frac{\dddot{\tau}}{2}v_n+\ddot{\mu}_n+\left(d_n-\frac{3}{2}\dot{\tau}\right)G_n+c_nF_n-\tau G_{n,t}-\sum_{k=n-1}^{n+1}G_{n,u_k}\left[\left(\frac{\dot{\tau}}{2}+a_k\right)u_k+b_kv_k+\lambda_k(t)\right]\\ &\quad-\sum_{k=n-1}^{n+1}G_{n,v_k}\left[\left(\frac{\dot{\tau}}{2}+d_k\right)v_k+c_ku_k+\mu_k(t)\right]=0.\end{aligned} \tag{2.5}$$

In Eqs. (2.3), (2.4), and (2.5) the quantities a_n, b_n, c_n, and d_n are independent of t. To proceed further, one could specify the interaction functions F_n and G_n. Instead, we shall assume that at least one symmetry generator (2.3) exists and make use of allowed transformations to simplify this vector. The second step is to find interactions F_n and G_n compatible with such a symmetry.

Substituting (1.4) into Eq. (1.1) and requiring that the form of these two equations be preserved, we find that the allowed transformations are quite restricted, namely,

$$\begin{pmatrix}u_n(t)\\ v_n(t)\end{pmatrix}=\begin{pmatrix}Q_n & R_n\\ S_n & T_n\end{pmatrix}\dot{\tilde{t}}^{-1/2}\begin{pmatrix}\tilde{u}_n(\tilde{t})\\ \tilde{v}_n(\tilde{t})\end{pmatrix}+\begin{pmatrix}\alpha_n(t)\\ \beta_n(t)\end{pmatrix},\quad \tilde{t}=\tilde{t}(t),\quad \frac{d\tilde{t}}{dt}\neq 0. \tag{2.6}$$

The entries Q_n, R_n, S_n, and T_n are independent of t; $\tilde{t}(t)$ is an arbitrary locally invertible function of t; α_n,β_n are arbitrary functions of n and t, and the matrix

$$M_n=\begin{pmatrix}Q_n & R_n\\ S_n & T_n\end{pmatrix},\quad \det M_n\neq 0, \tag{2.7}$$

is nonsingular.

It will be convenient to use a shorthand notation for the vector field X_n of Eq. (2.3), namely,

$$\left\{ \tau(t), A_n, \begin{pmatrix} \lambda_n(t) \\ \mu_n(t) \end{pmatrix} \right\}, \quad A_n = \begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix}. \tag{2.8}$$

If we perform an allowed transformation (2.6), then Eq. (1.1) goes into an equation of the same form, with F_n and G_n replaced by

$$\begin{pmatrix} \tilde{F}_n \\ \tilde{G}_n \end{pmatrix} = \dot{\tilde{t}}^{-3/2} M_n^{-1} \left[\begin{pmatrix} F_n \\ G_n \end{pmatrix} - \begin{pmatrix} \ddot{\alpha}_n \\ \ddot{\beta}_n \end{pmatrix} \right] + \left(\frac{1}{2} \dddot{\tilde{t}} \dot{\tilde{t}}^{-3} - \frac{3}{4} \ddot{\tilde{t}}^2 \dot{\tilde{t}}^{-4} \right) \begin{pmatrix} \tilde{u}_n \\ \tilde{v}_n \end{pmatrix}, \tag{2.9}$$

where $\tilde{F}_n$ and $\tilde{G}_n$ are functions of the new variables.

The vector field characterized by the triplet (2.3) goes into a new one of the same form,

$$\left\{ \tilde{\tau}(\tilde{t}), \tilde{A}_n, \begin{pmatrix} \tilde{\lambda}_n(\tilde{t}) \\ \tilde{\mu}_n(\tilde{t}) \end{pmatrix} \right\}, \tag{2.10}$$

with

$$\tilde{\tau}(\tilde{t}) = \tau(t(\tilde{t}))\dot{\tilde{t}},$$

$$\tilde{A}_n = M_n^{-1} A_n M_n,$$

$$\begin{pmatrix} \tilde{\lambda}_n(\tilde{t}) \\ \tilde{\mu}_n(\tilde{t}) \end{pmatrix} = M_n^{-1} \dot{\tilde{t}}^{1/2} \left[\left(A_n + \frac{\dot{\tau}}{2} \right) \begin{pmatrix} \alpha_n \\ \beta_n \end{pmatrix} - \tau \begin{pmatrix} \dot{\alpha}_n \\ \dot{\beta}_n \end{pmatrix} + \begin{pmatrix} \lambda_n \\ \mu_n \end{pmatrix} \right].$$

We shall use the allowed transformations to simplify the vector field, rather than the equation itself.

III. SYSTEMS WITH ONE-DIMENSIONAL SYMMETRY GROUPS

Let us now assume that the system (1.1) has at least a one-dimensional symmetry group, generated by a vector field of the type (2.3). Using allowed transformations (2.6), we take $\hat{X}$ into one of ten inequivalent classes.

Indeed, for $\tau \neq 0$ we can choose the function $\tilde{t}(t)$ so as to transform $\tau(t)$ into $\tau=1$, the functions $\alpha_n(t)$ and $\beta_n(t)$ so as to annul $\lambda_n(t)$, and $\mu_n(t)$ and the matrix M_n so as to take A_n into its canonical Jordan form.

For $\tau=0$ the standardized form of $\hat{X}$ depends on the rank of the matrix A_n. For rank $A_n = 2$, we can again transform λ_n and μ_n into $\lambda_n = \mu_n = 0$ and take A_n into one of three canonical forms. For rank $A_n = 1$, only one of the functions λ_n or μ_n can be annulled. We choose it to be $\lambda_n(t) = 0$. Then A_n can be taken into one of the two standard matrices of rank 1 in $\mathbb{R}^{2\times 2}$. For rank $A_n = 0$ both $\lambda_n(t)$ and $\mu_n(t)$ survive.

We thus obtain ten mutually inequivalent one-dimensional symmetry algebras, listed below. The statement now is that any single vector field $\hat{X}$ of the form (2.3) can be transformed by an allowed transformation into precisely one of these vector fields.

The next step is to determine the interactions for which a one-dimensional symmetry group exists. To do this, we run through the canonical vector fields just obtained, substitute the corresponding τ (=1 or 0), A_n, $\lambda_n(t)$, and $\mu_n(t)$ into Eqs. (2.4) and (2.5), and solve these equations for F_n and G_n.

Following this procedure, we arrive at the following list of interactions and their one-dimensional symmetry algebras:

$A_{1,1}$
$$\hat{X}=\partial_t+a_nu_n\partial_{u_n}+d_nv_n\partial_{v_n},$$
$$F_n=e^{a_nt}f_n(\xi_k,\eta_k),$$
$$G_n=e^{d_nt}g_n(\xi_k,\eta_k),$$
$$\xi_k=u_ke^{-a_kt},\quad \eta_k=v_ke^{-d_kt},$$
$$k=n-1,n,n+1;$$

$A_{1,2}$
$$\hat{X}=\partial_t+(a_nu_n+v_n)\partial_{u_n}+a_nv_n\partial_{v_n},$$
$$F_n=e^{a_nt}[f_n(\xi_k,\eta_k)+tg_n(\chi_k,\eta_k)],$$
$$G_n=e^{a_nt}g_n(\xi_k,\eta_k),$$
$$\xi_k=(u_k-tv_k)e^{-a_kt},\quad \eta_k=v_ke^{-a_kt},$$
$$k=n-1,n,n+1;$$

$A_{1,3}$
$$\hat{X}=\partial_t+(a_nu_n+b_nv_n)\partial_{u_n}+(-b_nu_n+a_nv_n)\partial_{v_n},\quad b_n>0,$$
$$\begin{pmatrix}F_n\\G_n\end{pmatrix}=e^{a_nt}\begin{pmatrix}\cos b_nt & \sin b_nt\\ -\sin b_nt & \cos b_nt\end{pmatrix}\begin{pmatrix}f_n(\xi_k,\eta_k)\\ g_n(\xi_k,\eta_k)\end{pmatrix},$$
$$\xi_k=r_ke^{-a_kt},\quad \eta_k=\theta_k+b_kt,$$
$$u_k=r_k\cos\theta_k,\quad v_k=r_k\sin\theta_k,$$
$$k=n-1,n,n+1;$$

$A_{1,4}$
$$\hat{X}=a_nu_n\partial_{u_n}+d_nv_n\partial_{v_n},\quad |a_n|\geqslant|d_n|,$$
$$F_n=u_nf_n(\xi_\alpha,\eta_k,t),$$
$$G_n=v_ng_n(\xi_\alpha,\eta_k,t),$$
$$\xi_a=u_\alpha^{a_n}u_n^{-a_\alpha},\quad \eta_k=v_k^{a_n}u_n^{-d_k},$$
$$k=n-1,n,n+1,\quad \alpha=n-1,n+1;$$

$A_{1,5}$
$$\hat{X}=(a_nu_n+v_n)\partial_{u_n}+a_nv_n\partial_{v_n},\quad a_n\neq 0,$$
$$F_n=v_nf_n(\eta_\alpha,\xi_k,t)+v_n\ln(v_n)g_n(\eta_\alpha,\xi_k,t),$$
$$G_n=a_nv_ng_n(\eta_\alpha,\xi_k,t),$$
$$\xi_k=a_k\frac{u_k}{v_k}-\ln(v_k),\quad \eta_\alpha=v_\alpha^{a_n}v_n^{-a_\alpha},$$
$$k=n-1,n,n+1,\quad \alpha=n-1,n+1;$$

$A_{1,6}$
$$\hat{X}=v_n\partial_{u_n},$$
$$F_n=f_n(v_k,\xi_\alpha,t)+u_ng_n(v_k,\xi_\alpha,t),$$
$$G_n=v_ng_n(v_k,\xi_\alpha,t),$$
$$\xi_\alpha=-v_\alpha u_n+v_nu_\alpha,$$
$$k=n-1,n,n+1,\quad \alpha=n-1,n+1;$$

$A_{1,7}$
$$\hat{X}=(a_nu_n+b_nv_n)\partial_{u_n}+(-b_nu_n+a_nv_n)\partial_{v_n},\quad b_n>0,$$
$$\begin{pmatrix}F_n\\G_n\end{pmatrix}=e^{-(a_n/b_n)\theta_n}\begin{pmatrix}\cos\theta_n & -\sin\theta_n\\ \sin\theta_n & \cos\theta_n\end{pmatrix}\begin{pmatrix}f_n(\xi_k,\eta_\alpha,t)\\ g_n(\xi_k,\eta_\alpha,t)\end{pmatrix},$$
$$\xi_k=r_k^{b_n}e^{a_k\theta_n},\quad \eta_\alpha=b_n\theta_\alpha-b_\alpha\theta_n,$$
$$u_k=r_k\cos\theta_k,\quad v_k=r_k\sin\theta_k,$$
$$k=n-1,n,n+1,\quad \alpha=n-1,n+1;$$

$A_{1,8}$
$$\hat{X}=a_nu_n\partial_{u_n}+\mu_n(t)\partial_{v_n},\quad \mu_n\neq 0,$$
$$F_n=u_nf_n(\eta_\alpha,\xi_k,t),$$
$$G_n=\frac{\ddot{\mu}_n}{\mu_n}v_n+g_n(\eta_\alpha,\xi_k,t),$$
$$\eta_\alpha=\mu_nv_\alpha-\mu_\alpha v_n,\quad \xi_k=u_ke^{-\alpha_kv_n/\mu_n},$$
$$k=n-1,n,n+1,\quad \alpha=n-1,n+1;$$

$A_{1,9}$
$$\hat{X}=v_n\partial_{u_n}+\mu_n(t)\partial_{v_n},\quad \mu_n\neq 0,$$
$$F_n=\frac{1}{2}\frac{\ddot{\mu}_n}{\mu_n^2}v_n^2+v_n g_n(\eta_\alpha,\eta_n,\xi_\alpha,t)+f_n(\eta_\alpha,\eta_n,\xi_\alpha,t),$$
$$G_n=\frac{\ddot{\mu}_n}{\mu_n}v_n+\mu_n g_n(\eta_\alpha,\eta_n,\xi_\alpha,t),$$
$$\eta_\alpha=\mu_n^2u_\alpha+\tfrac{1}{2}\mu_\alpha v_n^2-\mu_n v_n v_\alpha,\quad \xi_\alpha=\mu_\alpha v_n-\mu_n v_\alpha,$$
$$\eta_n=\mu_n u_n-\tfrac{1}{2}v_n^2,\quad \alpha=n-1,n+1;$$

$A_{1,10}$
$$\hat{X}=\lambda_n(t)\partial_{u_n}+\mu_n(t)\partial_{v_n},\quad \lambda_n,\ \mu_n\neq 0,$$
$$F_n=\frac{\ddot{\lambda}_n}{\lambda_n}u_n+f_n(\eta_k,\xi_\alpha,t),$$
$$G_n=\frac{\ddot{\mu}_n}{\mu_n}u_n+g_n(\eta_k,\xi_\alpha,t),$$
$$\xi_\alpha=\lambda_n u_\alpha-\lambda_\alpha u_n,\quad \eta_k=\mu_k u_n-\lambda_n v_k,$$
$$k=n-1,n,n+1,\quad \alpha=n-1,n+1.$$

We mention that the variables ξ_k and η_k are to be taken exactly as above. For instance, ξ_{n+1} is not an upshift of ξ_n.

The above results are summed up quite simply. Namely, the existence of a one-dimensional symmetry algebra restricts the interaction terms F_n and G_n to two arbitrary functions of six variables, rather than the original seven variables. The algebras $A_{1,1}$, $A_{1,2}$ and $A_{1,3}$ involve time translations. Hence, the time dependence in these cases is restricted: F_n and G_n depend on time explicitly and via invariant variables ξ_k and η_k that, in turn, depend explicitly on t. The algebras $A_{1,4},\ldots,A_{1,10}$ correspond to gauge transformations: the group transformations act on dependent variables only. The time variable figures in the arbitrary functions, as does the discrete independent variable n.

IV. HIGHER-DIMENSIONAL SYMMETRY ALGEBRAS

A. General strategy

The commutator of two symmetry operators (2.3) is an operator $X_3=[X_1,X_2]$ of the same form, satisfying

$$\tau_3=\tau_1\dot{\tau}_2-\tau_2\dot{\tau}_1,\quad A_{n,3}=-[A_{n,1},A_{n,2}],$$
$$\begin{pmatrix}\lambda_{n,3}\\ \mu_{n,3}\end{pmatrix}=\tau_1\begin{pmatrix}\dot{\lambda}_{n,2}\\ \dot{\mu}_{n,2}\end{pmatrix}-\tau_2\begin{pmatrix}\dot{\lambda}_{n,1}\\ \dot{\mu}_{n,1}\end{pmatrix}-\left(A_{n,1}+\frac{\dot{\tau}_1}{2}\right)\begin{pmatrix}\lambda_{n,2}\\ \mu_{n,2}\end{pmatrix}+\left(A_{n,2}+\frac{\dot{\tau}_2}{2}\right)\begin{pmatrix}\lambda_{n,1}\\ \mu_{n,1}\end{pmatrix}. \tag{4.1}$$

To obtain a finite-dimensional Lie algebra of symmetry operators, we see that the ''differential components'' $\tau_i(t)\partial_t$ must form a Lie algebra L_d, the ''matrix components'' $A_{n,i}$ must also form a Lie algebra L_m, homomorphic to L_d. Moreover, Eq. (4.1) shows that the ''functional components'' $(\lambda_{n,i}(t),\mu_{n,i}(t))$ must satisfy certain cohomology conditions.

The algebra of diffeomorphisms of $\mathbb{R}^1$, $\{\tau(t)\partial_t\}$ has only three mutually nondiffeomorphic finite-dimensional subalgebras, namely sl$(2,\mathbb{R})$ and its subalgebras, realized, e.g., as

$$\{\partial_t,t\partial_t,t^2\partial_t\},\quad \{\partial_t,t\partial_t\},\quad \text{and } \{\partial_t\}, \tag{4.2}$$

respectively.

For n fixed, the matrices A_n generate the Lie algebra of gl$(2,\mathbb{R})$. However, since the dependence on n is arbitrary, an unlimited number of copies of gl$(2,\mathbb{R})$ and its subalgebras is available.

We shall not perform a complete classification of possible symmetry algebras here. Instead, we shall first concentrate on sl(2,$\mathbb{R}$) symmetry algebras and show that, up to allowed transformations, four different sl(2,$\mathbb{R}$) symmetry algebras can be constructed. We then consider just one of these four and study its extensions to higher-dimensional Lie algebras.

B. Equivalence classes of sl(2,$\mathbb{R}$) symmetry algebras

Since sl(2,$\mathbb{R}$) is a simple Lie algebra, it has no ideals. Hence, a homomorphism between sl(2,$\mathbb{R}$) algebras is either an isomorphism, or one of the algebras is mapped onto zero. Correspondingly, we have three possibilities to explore: we shall call them sl$(2,\mathbb{R})_d$, sl$(2,\mathbb{R})_m$, and sl$(2,\mathbb{R})_c$ (where *d* stands for ''differential,'' *m* for ''matrix,'' and *c* for ''combined'').

1. The algebra sl$(2,\mathbb{R})_d$

We have *a priori*

$$\begin{aligned} X_1&=\partial_t+\lambda_n(t)\partial_{u_n}+\mu_n(t)\partial_{v_n},\\ X_2&=t\partial_t+(\tfrac{1}{2}u_n+\rho_n(t))\partial_{u_n}+(\tfrac{1}{2}v_n+\sigma_n(t))\partial_{v_n},\\ X_3&=t^2\partial_t+(tu_n+\omega_n(t))\partial_{u_n}+(tv_n+\kappa_n(t))\partial_{v_n}. \end{aligned}\tag{4.3}$$

Using allowed transformations we transform $\lambda_n\to0$, $\mu_n\to0$. The commutation relation $[X_1,X_2]=X_1$ then requires $\dot\rho_n=\dot\sigma_n=0$. A further allowed transformation (2.6) with $\tilde t(t)=t$, $M_n=I$, and (α_n,β_n) constant will not change X_1, but take $\rho_n\to0$, $\sigma_n\to0$ (while leaving $\lambda_n=\mu_n=0$). The commutation relations $[X_2,X_3]=X_3$ and $[X_1,X_3]=2X_2$ then imply $\omega_n=\kappa_n=0$.

2. The algebra sl$(2,\mathbb{R})_m$

A priori we have

$$\begin{aligned} X_1&=b_nv_n\partial_{u_n}+\lambda_n(t)\partial_{u_n}+\mu_n(t)\partial_{v_n},\\ X_2&=a_n(u_n\partial_{u_n}-v_n\partial_{v_n})+\rho_n(t)\partial_{u_n}+\sigma_n(t)\partial_{v_n},\\ X_3&=c_nu_n\partial_{v_n}+\omega_n(t)\partial_{u_n}+\kappa_n(t)\partial_{v_n}. \end{aligned}\tag{4.4}$$

The structure constants cannot depend on *n*, so the commutation relations imply

$$a_n=a,\quad b_nc_n=bc.\tag{4.5}$$

Given that the product b_nc_n does not depend on *n*, we can use an allowed transformation to take $b_n\to b$, $c_n\to c$. A further allowed transformation will take $\rho_n\to0$, $\sigma_n\to0$. The commutation relations then imply $\lambda_n=\mu_n=0$ and $\omega_n=\kappa_n=0$.

3. The combined algebra sl$(2,\mathbb{R})_c$

In view of the above results, we can write a ''combined'' algebra as

$$\begin{aligned} X_1&=\partial_t+\alpha v_n\partial_{u_n}+\xi_n\partial_{u_n}+\eta_n\partial_{v_n},\quad \alpha\neq0,\\ X_2&=t\partial_t+[(\tfrac{1}{2}+\beta)u_n+\lambda_n]\partial_{u_n}+[(\tfrac{1}{2}-\beta)v_n+\mu_n]\partial_{v_n},\\ X_3&=t^2\partial_t+(tu_n+\rho_n)\partial_{u_n}+(\gamma u_n+tv_n+\sigma_n)\partial_{v_n}. \end{aligned}\tag{4.6}$$

We use allowed transformations to set $\alpha=1$, $\xi_n=\eta_n=0$. The commutation relations then determine $\beta=\frac{1}{2}$, $\gamma=-1$. The functions $\lambda_n(t)$, $\mu_n(t)$, $\rho_n(t)$, and $\sigma_n(t)$ are greatly restricted by the commutation relations. As a matter of fact, we either have $\lambda_n=\mu_n=\rho_n=\sigma_n=0$, or we can use allowed transformations to obtain $\lambda_n=t$, $\mu_n=1$, $\rho_n=2t^2$, $\sigma_n=2t$.

We arrive at the following result.

Theorem 1: *Precisely four classes of* sl(2,ℝ) *algebras can be realized by vector fields of the form (2.3). Any such* sl(2,ℝ) *algebra can be taken by an allowed transformation (2.6) into precisely one of the following algebras:*

$$\mathrm{sl}(2,\mathbb{R})_1: \quad \begin{aligned} &X_1=v_n\partial_{u_n},\\ &X_2=\tfrac{1}{2}(u_n\partial_{u_n}-v_n\partial_{v_n}),\\ &X_3=u_n\partial_{v_n}, \end{aligned} \tag{4.7}$$

$$\mathrm{sl}(2,\mathbb{R})_2: \quad \begin{aligned} &X_1=\partial_t,\\ &X_2=t\partial_t+\tfrac{1}{2}(u_n\partial_{u_n}+v_n\partial_{v_n}),\\ &X_3=t^2\partial_t+t(u_n\partial_{u_n}+v_n\partial_{v_n}), \end{aligned} \tag{4.8}$$

$$\mathrm{sl}(2,\mathbb{R})_3: \quad \begin{aligned} &X_1=\partial_t+v_n\partial_{u_n},\\ &X_2=t\partial_t+u_n\partial_{u_n},\\ &X_3=t^2\partial_t+tu_n\partial_{u_n}+(tv_n-u_n)\partial_{v_n}, \end{aligned} \tag{4.9}$$

$$\mathrm{sl}(2,\mathbb{R})_4: \quad \begin{aligned} &X_1=\partial_t+v_n\partial_{u_n}\\ &X_2=t\partial_t+(u_n+t)\partial_{u_n}+\partial_{v_n}\\ &X_3=t^2\partial_t+(tu_n+2t^2)\partial_{u_n}+(tv_n-u_n+2t)\partial_{v_n}. \end{aligned} \tag{4.10}$$

C. Indecomposable Lie algebras containing sl(2,ℝ)$_1$

A Lie algebra L is called indecomposable if it cannot be written as a direct sum, $L=L_1\oplus L_2$. A Lie algebra over ℝ containing sl(2,ℝ) is either simple or it allows a nontrivial Levi decomposition,[15]

$$L=S\triangleright R, \tag{4.11}$$

where S is a semisimple Lie algebra and R is the radical, that is, the maximal solvable ideal of L.

It follows from the results of Sec. IV A that the only simple Lie algebras that can be constructed from operators of the form (2.3) are the four sl(2,ℝ) algebras obtained in Sec. IV B. We can hence concentrate on Lie algebras of the form (4.11).

The algebra S is either sl(2,ℝ)$_1$ itself, or the direct sum of sl(2,ℝ)$_1$ with one or more other sl(2,ℝ) algebras.

Requiring that a symmetry operator Y should commute with all elements of sl(2,ℝ)$_1$, we find that Y must have the form

$$Y_0=\tau\partial_t+(\tfrac{1}{2}\dot{\tau}+a_n)(u_n\partial_{u_n}+v_n\partial_{v_n}). \tag{4.12}$$

It is hence possible to construct precisely one semisimple Lie algebra properly containing sl(2,ℝ)$_1$, namely, the direct sum sl(2,ℝ)$_1\oplus$sl(2,ℝ)$_2$ with sl(2,ℝ)$_2$ defined in Eq. (4.8).

Let us introduce some notations for vector fields, to be used below. We put

$$V(a_n)=a_n(u_n\partial_{u_n}+v_n\partial_{v_n}), \tag{4.13}$$

$$T(a_n)=\partial_t+a_n(u_n\partial_{u_n}+v_n\partial_{v_n}),\tag{4.14}$$

$$D(a_n)=t\partial_t+(\tfrac{1}{2}+a_n)(u_n\partial_{u_n}+v_n\partial_{v_n}),\tag{4.15}$$

$$P(a_n)=t^2\partial_t+(t+a_n)(u_n\partial_{u_n}+v_n\partial_{v_n}),\tag{4.16}$$

$$R(a_n)=(t^2+1)\partial_t+(t+a_n)(u_n\partial_{u_n}+v_n\partial_{v_n}),\tag{4.17}$$

$$Y_u(\lambda_n)=\lambda_n(t)\partial_{u_n},\quad Y_v(\lambda_n)=\lambda_n(t)\partial_{v_n}.\tag{4.18}$$

In all cases we have $\dot{a}_n=0$, but $\lambda_n(t)$ can be a function of t. Both a_n and $\lambda_n(t)$ can be functions of n.

Let us consider $S=\mathrm{sl}(2,\mathbb{R})_1$ and $S=\mathrm{sl}(2,\mathbb{R})_1\oplus\mathrm{sl}(2,\mathbb{R})_2$ in Eq. (4.11) separately.

1. $S=\mathrm{sl}(2,\mathbb{R})_1$

The considered Lie algebras will have a basis $\{X_1,X_2,X_3,Y_1,\ldots,Y_n\}$ with X_i given in Eq. (4.7). The basis elements $\{Y_1,\ldots,Y_n\}$ span the radical R. The algebra S acts on R according to some linear, not necessarily irreducible, finite-dimensional representation.

We start with the Cartan subalgebra $\{X_2\}$ of $\mathrm{sl}(2,\mathbb{R})$. It can be represented by a diagonal matrix in any finite-dimensional representation. Consider $Y\in R$. We have

$$[X_2,Y]=pY,\tag{4.19}$$

with Y as in Eq. (2.3). Equation (4.19) implies

$$\begin{gathered}p\tau=0,\\ p\left(\frac{\dot\tau}{2}+a_n\right)=0,\quad \left(p+\frac{1}{2}\right)\lambda_n=0,\quad (p+1)b_n=0,\\ p\left(\frac{\dot\tau}{2}+d_n\right)=0,\quad \left(p-\frac{1}{2}\right)\mu_n=0,\quad (p-1)c_n=0.\end{gathered}\tag{4.20}$$

For $p=0$ we obtain an operator that commutes not only with X_2, but with all of $\mathrm{sl}(2,\mathbb{R})_1$, namely, Y_0 of Eq. (4.12). This is a singlet representation of $\mathrm{sl}(2,\mathbb{R})$.

For $p=1$, or $p=-1$, Eq. (4.19) forces Y to be an element of $\mathrm{sl}(2,\mathbb{R})_1$, in other words, no such $Y\in R$ exists.

For $p=\pm\frac{1}{2}$ we obtain $Y_1=\lambda_n(t)\partial_{u_n}$ and $Y_2=\mu_n(t)\partial_{v_n}$, respectively. Acting with X_1 and X_3 on these operators, we find that the only representation of $\mathrm{sl}(2,\mathbb{R})_1$ that can be realized is a doublet one, namely $\{Y_u(\lambda_n),Y_v(\lambda_n)\}$ of Eq. (4.18), with $\lambda_n(t)$ an arbitrary function of n and t. The indecomposable Lie algebra $\{X_1,X_2,X_3,Y_u(\lambda_n),Y_v(\lambda_n)\}$ is isomorphic to the special affine Lie algebra $\mathrm{saff}(2,\mathbb{R})$.

All further indecomposable symmetry algebras containing $\mathrm{sl}(2,\mathbb{R})_1$ must be extensions of $\mathrm{saff}(2,\mathbb{R})$. The objects that we can add to $\mathrm{saff}(2,\mathbb{R})$ are either $\mathrm{sl}(2,\mathbb{R})$ doublets or singlets. Let us run through all possibilities.

(1) We can add an arbitrary number k of doublets of the form (4.18), where the k functions $\{\lambda_n^1(t),\lambda_n^2(t),\ldots,\lambda_n^k(t)\}$ must be linearly independent. However, we shall see in Sec. V that the presence of three such pairs forces the functions F_n and G_n in Eq. (1.1) to be linear. Moreover, even two such pairs are compatible with a nonlinear interaction only if they are of the form (or transformable into)

$$\lambda_n^1(t)=1,\quad \lambda_n^2(t)=t.\tag{4.21}$$

(2) We can add a singlet of the form (4.12). If we have $\tau=0$, then the commutation relations $[Y_0,Y_u]$ and $[Y_0,Y_v]$ imply $a_n=a_{n+1}$ and we can set $a_n=1$. We obtain an affine Lie algebra $\mathrm{gaff}(2,\mathbb{R})_1$ consisting of $\mathrm{saff}(2,\mathbb{R})$ and $V(1)$ of Eq. (4.13).
If we have $\tau\neq0$ in Eq. (4.12) and only one operator of this type, then we can use allowed transformations to take $\tau(t)$ into $\tau(t)=1$. The commutation relations $[Y_0,Y_u]$ and $[Y_0,Y_v]$ then imply

$$\lambda_n(t)=R_n e^{(a_n+k)t}, \quad \dot{R}_n=0.$$

For $k=0$, the algebra is decomposable. For $k\neq0$ we can use allowed transformations to put $k=-1$ and $R_n=1$. We obtain a second algebra isomorphic to $\mathrm{gaff}(2,\mathbb{R})$, but not conjugate to the previous one. We have

$$\mathrm{gaff}(2,\mathbb{R})_2\sim\{X_1,X_2,X_3,Y_u(e^{(a_n-1)t}),Y_v(e^{(a_n-1)t}),T(a_n)\}. \tag{4.22}$$

In the special case of $a_n=a_{n+1}$ in Eq. (4.22), a further extension is possible. We transform $\lambda=e^{(a-1)t}$ into $\lambda=1$; then $T(a_n)$ goes into $D(b_n)$ with $b_n=b_{n+1}\equiv b\neq-\frac{1}{2}$, since for $b=-\frac{1}{2}$ the algebra is decomposable.

(3) We can add two singlets of the form (4.12). If they commute, they must be $\{V(1),T(0)\}$. The obtained algebra is decomposable. If they do not commute, they must form a two-dimensional Lie algebra, namely, $\{T(0),D(a),a_n=a_{n+1}\equiv a\}$. This implies $\lambda_n(t)\sim1$, i.e., the entire radical is $\{Y_u(1),Y_v(1),T(0),D(a)\}$ with $a\neq\frac{1}{2}$ (the case $a=\frac{1}{2}$ corresponds to a decomposable algebra).

(4) If we add three singlets, the only case corresponds to the radical $\{Y_u(1),Y_v(1),V(1),T(0),D(0)\}$. There will then be no invariant interaction (see below).

(5) Let us consider the special case of two doublets of the form (4.18), namely,

$$Y_u(1)=\partial_{u_n}, \quad Y_v(1)=\partial_{v_n}, \quad Y_u(t)=t\partial_{u_n}, \quad Y_v(t)=t\partial_{v_n}. \tag{4.23}$$

This algebra can be extended by a further element, namely,

$$Z=(\tau_0+\tau_1 t+\rho_2 t^2)\partial_t+(\tfrac{1}{2}\tau_1+\tau_2 t+a)(u_n\partial_{u_n}+v_n\partial_{v_n}),$$
$$a_n=a_{n+1}\equiv a, \tag{4.24}$$

where τ_0, τ_1, and τ_2 are constants. By allowed transformations we can take Z into one of the four operators $V(1)$, $T(a)$, $D(a)$, or $R(a)$ of (4.13), (4.14), (4.15), and (4.17), respectively.

(6) We can add a two-dimensional algebra to (4.23), namely,

$$\{T(0),D(a)\}, \quad \{T(0),V(1)\}, \quad \{V(1),D(0)\}, \text{ or } \{V(1),R(0)\}.$$

(7) We can add only one three-dimensional algebra to (4.23), namely,

$$\{T(0), \quad D(0), \quad V(1)\}.$$

This completes the list of indecomposable symmetry algebras of the form (4.11) with $S=\mathrm{sl}(2,\mathbb{R})_1$.

2. $S=\mathrm{sl}(2,\mathbb{R})_1\oplus\mathrm{sl}(2,\mathbb{R})_2$

The algebra S is itself decomposable. It gives rise to precisely two indecomposable symmetry algebras. First, we have the one obtained by adding the Abelian ideal (4.23) to $\mathrm{sl}(2,\mathbb{R})_1\oplus\mathrm{sl}(2,\mathbb{R})_2$. Second, we get an 11-dimensional algebra by adding $V(1)$ to the first case.

D. Decomposable Lie algebras containing $\mathrm{sl}(2,\mathbb{R})_1$

All decomposable Lie algebras L_D can be obtained from the indecomposable L_I ones, by adding their centralizers,

$$L_D = L_I \oplus C, \quad [C, L_I] = 0. \tag{4.25}$$

The centralizer C must commute with all elements of $\mathrm{sl}(2,\mathbb{R})_1$ and hence all of its elements will have the form of Y_0 of Eq. (4.12).

Let us consider the individual indecomposable algebras L_I.

1. $L_I = \mathrm{sl}(2,\mathbb{R})_1$

The centralizer C can be Abelian. Then we have the following possibilities: $C = \{V(a_{i,n}), i = 1,...,k\}$ or $C = \{V(a_{i,n}), T(b_n), i = 1,...,k\}$. The quantities $a_{1,n},\ldots,a_{k,n}$ must form a set of k linearly independent functions of n. If the centralizer is non-Abelian, then we have either $C \sim \mathrm{sl}(2,\mathbb{R})_2$ or $C = \{T(0), D(a)\}$. Both of these centralizers can be further extended by adding $V(a_{i,n})$, $i = 1,...,k$, (with $a_{1,n},\ldots,a_{k,n}$ linearly independent).

2. $L_I = \mathrm{saff}(2,\mathbb{R})$

We must require Y_0 of Eq. (4.12) to commute with $Y_u(\lambda_n)$ and $Y_v(\lambda_n)$ of Eq. (4.18). We obtain

$$\lambda_n\left(\tfrac{1}{2}\dot{\tau} + a_n\right) - \tau\dot{\lambda}_n = 0. \tag{4.26}$$

For $\tau = 0$, Eq. (4.26) implies $\lambda_n a_n = 0$, and this is not allowed. For $\tau \neq 0$ we take $\tau \to 1$ by an allowed transformation, and Eq. (4.26) then implies $\lambda_n(t) = \gamma_n e^{a_n t}$. A further allowed transformation will take $\gamma_n \to 1$. We obtain the decomposable Lie algebra $\mathrm{saff}(2,\mathbb{R}) \oplus T(a_n)$. In the special case $a_n = a_{n+1}$ we transform $\lambda_n(t) \to 1$ and obtain a larger centralizer, namely, $\{T(0), D(-\tfrac{1}{2})\}$.

3. $L_I = \mathrm{gaff}(2,\mathbb{R})_1$

A nontrivial centralizer exists only if we have $\lambda_n(t) = e^{a_n t}$ in $\mathrm{saff}(2,\mathbb{R})$. In the case $a_n \neq 0$, the centralizer is $C = \{T(a_n)\}$. If $a_n = 0$ the centralizer is $C = \{T(0), D(-\tfrac{1}{2})\}$.

4. $L_I = \mathrm{gaff}(2,\mathbb{R})_2$

The centralizer is $C = \{T(a_n) - V(1)\}$. This algebra corresponds to the first one obtained in the case $L_I = \mathrm{gaff}(2,\mathbb{R})_1$ above.

E. Summary of possible symmetry algebras containing $\mathrm{sl}(2,\mathbb{R})_1$

The classification of possible symmetry algebras can now be summed up rather simply. In addition to $\mathrm{sl}(2,\mathbb{R})_1$ of Eq. (4.7), we have a further algebra L_C (the ''complementary'' algebra). The structure of each symmetry algebra is

$$L = \mathrm{sl}(2,\mathbb{R})_1 \dotplus L_C, \quad [\mathrm{sl}(2,\mathbb{R})_1, L_C] \subseteq L_C, \quad [L_C, L_C] \subseteq L_C. \tag{4.27}$$

The symbol $\dotplus$ denotes a direct sum of vector spaces. Moreover, Eq. (4.27) shows that L is either a direct sum or a semidirect one. The algebra L_C is also a representation space for $\mathrm{sl}(2,\mathbb{R})_1$. Irreducible representations in this case can be of dimension 1 or 2. All higher-dimensional representations are completely reducible into sums of one- and two-dimensional representations.

For further use it is convenient to split the symmetry algebras into four series, according to the structure of the Lie algebra L_C. In all cases L contains $\mathrm{sl}(2,\mathbb{R})_1$. We shall just specify L_C.

1. Series A

L_C is solvable and each element is a $\mathrm{sl}(2,\mathbb{R})_1$ singlet. There exist three different infinite-dimensional Lie algebras of this type:

$$A_1. \quad \{V(a_{k,n})\}, \tag{4.28}$$

$$A_2. \quad \{T(b_n), V(a_{k,n})\}, \tag{4.29}$$

$$A_3.\ \{T(0), D(b_n), V(a_{k,n})\}. \tag{4.30}$$

In each case we have $k=1,2,\ldots,$ and the expressions a_k must be linearly independent functions of n. Taking $1 \leqslant k \leqslant N$ for some finite N, we obtain finite-dimensional subalgebras.

2. Series B

L_C is solvable and contains precisely one $\mathrm{sl}(2,\mathbb{R})_1$ doublet,

$$B_1 = \{Y_u(\lambda_n), Y_v(\lambda_n)\}. \tag{4.31}$$

This is the indecomposable algebra $\mathrm{saff}(2,\mathbb{R})$ [B_1 together with $\mathrm{sl}(2,\mathbb{R})_1$]. We have $\dim L=5$,

$$B_2 = \{Y_u(\lambda_n), Y_v(\lambda_n), V(1)\}. \tag{4.32}$$

B_2 corresponds to the indecomposable algebra $\mathrm{gaff}(2,\mathbb{R})_1$ with $\dim L=6$,

$$B_3 = \{Y_u(e^{(a_n-1)t}), Y_v(e^{(a_n-1)t}), T(a_n)\}. \tag{4.33}$$

B_3 corresponds to the Lie algebra $\mathrm{gaff}(2,\mathbb{R})_2$, isomorphic but not conjugate to B_2,

$$B_4 = \{Y_u(e^{a_n t}), Y_v(e^{a_n t}), T(a_n)\}. \tag{4.34}$$

This algebra is $\mathrm{saff}(2,\mathbb{R}) \oplus T(a_n)$,

$$B_5 = \{Y_u(1), Y_v(1), T(0), D(a)\}. \tag{4.35}$$

The algebra B_5 is indecomposable (except if $a=-\frac{1}{2}$),

$$B_6 = \{Y_u(e^{(a_n-1)t}), Y_v(e^{(a_n-1)t}), T(a_n), V(1)\}. \tag{4.36}$$

The algebra B_6 is decomposable,

$$B_7 = \{Y_u(1), Y_v(1), T(0), D(0), V(1)\}. \tag{4.37}$$

The algebra B_7 is indecomposable.

3. Series C

L_C contains two $\mathrm{sl}(2,\mathbb{R})$ doublets. The doublets could be characterized by any two functions $\lambda_{1,n}(t)$ and $\lambda_{2,n}(t)$. However, we shall only be interested in the case $\lambda_1=1$, $\lambda_2=t$. The others do not lead to invariant interactions. Similarly, we do not need algebras containing three or more doublets. In all cases the algebra L_C contains the elements (4.23). For $\dim L_C \geqslant 5$ it contains further elements. We have

$$C_1 = \{Y_u(1), Y_v(1), Y_u(t), Y_v(t)\}. \tag{4.38}$$

Further, we just list the additional elements,

$$C_2.\ \{T(a)\},\quad a=0 \ \text{ or } \ 1, \tag{4.39}$$

$$C_3.\ \{D(a)\}, \tag{4.40}$$

$$C_4.\ \{R(a)\}, \tag{4.41}$$

$$C_5.\ \{V(1)\}, \tag{4.42}$$

$$C_6.\ \{T(0), D(a)\}. \tag{4.43}$$

In all cases above, a does not depend on $n(a_{n+1}=a_n)$,

$$C_7.\ \{V(1),T(0)\}, \tag{4.44}$$

$$C_8.\ \{V(1),D(0)\}, \tag{4.45}$$

$$C_9.\ \{V(1),R(0)\}, \tag{4.46}$$

$$C_{10}.\ \{T(0),D(0),P(0)\}\sim \mathrm{sl}(2,\mathbb{R})_2, \tag{4.47}$$

$$C_{11}.\ \{T(0),D(0),V(1)\}, \tag{4.48}$$

$$C_{12}.\ \{T(0),D(0),P(0),V(1)\}. \tag{4.49}$$

4. Series D

L_C contains $\mathrm{sl}(2,\mathbb{R})_2$ and (possibly) further elements, namely,

$$D_1.\ \text{None}, \tag{4.50}$$

$$D_2.\ \{V(a_n)\}, \tag{4.51}$$

$$D_3.\ \{V(a_{1,n}),V(a_{2,n})\}, \tag{4.52}$$

$$D_4.\ \{Y_u(1),Y_v(1),Y_u(t),Y_v(t)\}, \tag{4.53}$$

$$D_5.\ \{Y_u(1),Y_v(1),Y_u(t),Y_v(t),V(1)\} \tag{4.54}$$

(D_4 coincides with C_{10} and D_5 with C_{12}).

V. THE INVARIANT INTERACTIONS

A. General procedure and interactions invariant under $\mathrm{SL}(2,\mathbb{R})_1$

In this section we shall find all interaction functions, invariant under symmetry groups, containing $\mathrm{SL}(2,\mathbb{R})_1$. We make use of the subalgebra classification provided in Sec. IV.

We first establish the form of the interaction, invariant under $\mathrm{SL}(2,\mathbb{R})_1$ itself. To do this we set $\tau(t)=\lambda_n(t)=\mu_n(t)=0$ in the determining equations (2.4) and (2.5) and consider the equations obtained for $a_n=-d_n=1$, $b_n=c_n=0$, then $b_n=1$, $a_n=-d_n=c_n=0$, and, finally, $c_n=1$, $a_n=-d_n=b_n=0$. The general solution of the obtained system of six equations can be written in the following form:

$$F_n=u_{n+1}f_n+u_ng_n, \quad G_n=v_{n+1}f_n+v_ng_n, \tag{5.1}$$

where f_n and g_n are functions of four variables each, namely,

$$t,\ \ \xi_n=u_{n+1}v_{n-1}-u_{n-1}v_{n+1}, \quad \xi_\alpha=u_\alpha v_n-u_nv_\alpha, \quad \alpha=n\pm 1. \tag{5.2}$$

Note that ξ_n, ξ_{n+1}, and ξ_{n-1} are as given in Eq. (5.2). They are not upshifts or downshifts of each other.

We shall proceed further by dimension of the symmetry algebra and by its structure. Thus, we can successively add $\mathrm{sl}(2,\mathbb{R})$ singlets of the form (4.12) or doublets of the form (4.18). We continue adding symmetry elements until the interaction is completely specified, i.e., it involves no further arbitrary functions. We then solve the ''inverse problem.'' That is, we substitute the functions F_n and G_n back into the determining equations and solve for the symmetries. This provides a verification of previous calculations. More important, this procedure will find the largest symmetry algebra allowed by any given interaction.

Obviously, all invariant interactions will have the form (5.1). It is the functions f_n and g_n that will be further refined, and their dependence on the variables ξ_k and t will be restricted.

For future convenience we write down two further forms of the $\mathrm{SL}(2,\mathbb{R})_1$ invariant interaction functions, equivalent to (5.1). The first is

$$F_n = u_{n+1}\frac{\xi_{n-1}}{\xi_n}h_n + u_n k_n, \quad G_n = v_{n+1}\frac{\xi_{n-1}}{\xi_n}h_n + v_n k_n, \tag{5.3}$$

where h_n and k_n are arbitrary functions of the variables (5.2). The second convenient form is

$$\begin{aligned} F_n &= (\lambda_{n-1}u_{n+1} - \lambda_{n+1}u_{n-1})\phi_n + (\lambda_{n+1}u_n - \lambda_n u_{n+1})\psi_n + \frac{\ddot{\lambda}_n}{\lambda_{n+1}}u_{n+1}, \\ G_n &= (\lambda_{n-1}v_{n+1} - \lambda_{n+1}v_{n-1})\phi_n + (\lambda_{n+1}v_n - \lambda_n v_{n+1})\psi_n + \frac{\ddot{\lambda}_n}{\lambda_{n+1}}v_{n+1}, \end{aligned} \tag{5.4}$$

where $\lambda_n(t)$ is some arbitrary function of n and t and ϕ_n and ψ_n depend in an unspecified manner on the variables (5.2).

B. Interactions invariant under four-dimensional symmetry groups

As was shown in Sec. IV, two types of four-dimensional symmetry algebras containing $\mathrm{sl}(2,\mathbb{R})_1$ can exist. Both are decomposable according to the pattern $4=3+1$. Here and below we shall always list the operators that we can add to $\mathrm{sl}(2,\mathbb{R})_1$.

1. $V(a_n) = a_n(u_n\partial_{u_n} + v_n\partial_{v_n})$

The invariant interactions will have the form (5.3), but h_n and k_n will depend on three variables only.

(i) $a_{n-1}+a_{n+1}\neq 0$. The variables are

$$t, \quad \eta_\alpha = (\xi_\alpha)^{a_{n-1}+a_{n+1}}(\xi_n)^{-a_n-a_\alpha}, \quad \alpha = n\pm 1. \tag{5.5}$$

(ii) $a_{n-1}+a_{n+1}=0$. The variables are

$$t, \quad \xi_n, \quad \eta = (\xi_{n+1})^{a_{n+1}-a_n}(\xi_{n-1})^{a_{n+1}+a_n}. \tag{5.6}$$

2. $T(b_n) = \partial_t + b_n(u_n\partial_{u_n} + v_n\partial_{v_n})$

The invariant interaction will again have the form (5.3), however, in this case h_n and k_n are arbitrary functions of the three variables,

$$\zeta_n = \xi_n e^{-(b_{n-1}+b_{n+1})t}, \quad \zeta_\alpha = \xi_\alpha e^{-(b_n+b_\alpha)t}, \quad \alpha = n\pm 1. \tag{5.7}$$

We see that adding further singlets of the type $V(a_n)$ will restrict the variables in the functions h_n and k_n, not, however, the general form of Eq. (5.3).

C. Five-dimensional symmetry groups

From the results of Sec. IV, we know that three decomposable and one indecomposable symmetry algebras of dimension 5 can exist. Let us run through all four possibilities.

1. Decomposition 5=3+1+1

a. $V(a_{i,n}) = a_{i,n}(u_n\partial_{u_n} + v_n\partial_{v_n})$ $i=1,2$, $a_{2,n}\neq \lambda a_{1,n}$. The interaction is of the form (5.3). The functions h_n and k_n depend on two variables each, namely, time t and

$$\eta = (\xi_{n-1})^A(\xi_{n+1})^B(\xi_n)^C, \tag{5.8}$$

$$A=a_{1,n}(a_{2,n+1}+a_{2,n-1})+a_{1,n+1}(a_{2,n-1}-a_{2,n})-a_{1,n-1}(a_{2,n+1}+a_{2,n}),$$

$$B=-a_{1,n}(a_{2,n+1}+a_{2,n-1})+a_{1,n+1}(a_{2,n-1}+a_{2,n})-a_{1,n-1}(a_{2,n+1}-a_{2,n}), \tag{5.9}$$

$$C=a_{1,n}(a_{2,n+1}-a_{2,n-1})-a_{1,n+1}(a_{2,n-1}+a_{2,n})+a_{1,n-1}(a_{2,n+1}+a_{2,n}).$$

Note that the variable η always exists since the condition $A=B=C=0$ (and hence η=const) only occurs for $a_{1,n-1}a_{2,n}-a_{1,n}a_{2,n-1}=0$, which implies $a_{2,n}=\lambda a_{1,n}$, λ=const, and this is not allowed.

b. $V(a_n)=a_n(u_n\partial_{u_n}+v_n\partial_{v_n})$, $T(b_n)=\partial_t+b_n(u_n\partial_{u_n}+v_n\partial_{v_n})$. The invariant interaction is as in Eq. (5.3) with h_n and k_n functions of two variables each. Namely, the following.

(i) $a_{n+1}+a_{n-1}\neq 0$:

$$\rho_\alpha=(\zeta_\alpha)^{a_{n+1}+a_{n-1}}(\zeta_n)^{-a_\alpha-a_n}, \quad \alpha=n\pm 1, \tag{5.10}$$

with ζ_α, ζ_n as in Eq. (5.7).

(ii) $a_{n+1}+a_{n-1}=0$:

$$\rho_n=\zeta_n, \quad \sigma_n=(\zeta_{n-1})^{a_{n+1}+a_n}(\zeta_{n+1})^{a_{n+1}-a_n}. \tag{5.11}$$

2. Decomposition 5=3+2

a. $T(0)=\partial_t$, $D(b_n)=t\partial_t+(\frac{1}{2}+b_n)(u_n\partial_{u_n}+v_n\partial_{v_n})$. We impose $b_n\neq -\frac{1}{2}$; otherwise we have no invariant interaction. We must distinguish two subcases here.

(1) $b_{n+1}+b_{n-1}+1\neq 0$. The interaction as in Eq. (5.3), with

$$h_n=(\xi_n)^{-2/(b_{n+1}+b_{n-1}+1)}p_n, \quad k_n=(\xi_n)^{-2/(b_{n+1}+b_{n-1}+1)}q_n, \tag{5.12}$$

where p_n and q_n depend on two variables, namely,

$$\chi_\alpha=(\xi_\alpha)^{b_{n+1}+b_{n-1}+1}(\xi_n)^{-b_n-b_\alpha-1}, \quad \alpha=n\pm 1. \tag{5.13}$$

(2) $b_{n+1}+b_{n-1}+1=0$, $b_{n+1}+b_n+1\neq 0$:

$$h_n=(\xi_{n+1})^{-2/(b_{n+1}+b_n+1)}p_n, \quad k_n=(\xi_{n+1})^{-2/(b_{n+1}+b_n+1)}q_n, \tag{5.14}$$

where p_n and q_n depend on

$$\chi_n=(\xi_{n-1})^{b_{n+1}+b_n+1}(\xi_{n+1})^{-b_{n-1}-b_n-1}, \quad \xi_n. \tag{5.15}$$

Note that for $b_{n+1}+b_{n-1}+1=0$, $b_{n+1}+b_n+1=0$, we have $b_n=-\frac{1}{2}$, and there is no invariant interaction.

3. Indecomposable Lie algebra

$$Y_u(\lambda_n)=\lambda_n(t)\partial_{u_n}, \quad Y_v(\lambda_n)=\lambda_n(t)\partial_{v_n}. \tag{5.16}$$

The invariant interaction is as in Eq. (5.4), but the functions ϕ_n and ψ_n depend on only two variables, namely,

$$t, \quad \omega=\lambda_{n-1}\xi_{n+1}-\lambda_n\xi_n-\lambda_{n+1}\xi_{n-1}. \tag{5.17}$$

D. Six-dimensional symmetry groups

1. Decomposition 6=3+1+1+1

a. $V(a_{i,n})=a_{i,n}(u_n\partial_{u_n}+v_n\partial_{v_n})$, $i=1,2,3$. The invariant interaction is as in Eq. (5.3), but h_n and k_n are functions of t only. We see that the coefficients $a_{i,n}$ do not figure in the interaction.

Hence, we can add an arbitrary number of vector fields $V(a_{i,n})$, $i\in\mathbb{Z}$ to the symmetry algebra. In other words, the symmetry algebra for the interaction (5.3) with h_n and k_n depending on t alone is infinite dimensional.

b. $V(a_{i,n})=a_{i,n}(u_n\partial_{u_n}+v_n\partial_{v_n}),\ i=1,2,\ T(b_n)=\partial_t+b_n(u_n\partial_{u_n}+v_n\partial_{v_n})$. The invariant interaction is as in Eq. (5.3), but h_n and k_n depend on one variable only, namely,

$$\omega=\eta e^{-2t|M|},\quad M=\begin{pmatrix} b_{n-1} & b_n & b_{n+1} \\ a_{1,n-1} & a_{1,n} & a_{1,n+1} \\ a_{2,n-1} & a_{2,n} & a_{2,n+1}\end{pmatrix}, \tag{5.18}$$

with η as in Eq. (5.8).

2. Decomposition 6=3+2+1

a. $V(a_n)=a_n(u_n\partial_{u_n}+v_n\partial_{v_n}),\ T(0)=\partial_t,\ D(c_n)=t\partial_t+(\frac{1}{2}+c_n)(u_n\partial_{u_n}+v_n\partial_{v_n})$. We start from Eq. (5.3). The presence of $T(0)=\partial_t$ implies that h_n and k_n do not depend on t. We first notice that if we have

$$\gamma_n=c_n+\tfrac{1}{2}=0 \quad \text{or} \quad \gamma_n=c_n+\tfrac{1}{2}=\lambda a_n, \tag{5.19}$$

then we must have $h_n=k_n=0$ (no invariant interaction). In all other cases, invariance under $V(a_n)$ and $D(c_n)$ implies

$$\begin{gathered} h_n=(\xi_n)^{\mu}(\xi_{n+1})^{\nu}(\xi_{n-1})^{\rho}p_n(\omega),\quad k_n=(\xi_n)^{\mu}(\xi_{n+1})^{\nu}(\xi_{n-1})^{\rho}q_n(\omega),\\ \omega=(\xi_{n-1})^{A}(\xi_{n+1})^{B}(\xi_n)^{C}, \end{gathered} \tag{5.20}$$

with A, B, and C as in Eq. (5.9), with the substitutions

$$a_{1,n}\rightarrow c_n+\tfrac{1}{2},\quad a_{2,n}\rightarrow a_n.$$

The constants μ, ν, and ρ in Eq. (5.20) satisfy

$$\begin{gathered} (a_{n+1}+a_{n-1})\mu+(a_{n+1}+a_n)\nu+(a_{n-1}+a_n)\rho=0,\\ (\gamma_{n+1}+\gamma_{n-1})\mu+(\gamma_{n+1}+\gamma_n)\nu+(\gamma_{n-1}+\gamma_n)\rho=-2. \end{gathered} \tag{5.21}$$

Thus, for $C\neq 0$ we can put

$$\mu=0,\quad \nu=2\,\frac{a_n+a_{n-1}}{C},\quad \rho=-2\,\frac{a_n+a_{n+1}}{C}.$$

For $C=0$, $A\neq 0$,

$$\mu=2\,\frac{a_n+a_{n+1}}{A},\quad \nu=-2\,\frac{a_{n+1}+a_{n-1}}{A},\quad \rho=0.$$

For $C=A=0$, $B\neq 0$,

$$\mu=-2\,\frac{a_{n-1}+a_n}{B},\quad \nu=0,\quad \rho=2\,\frac{a_{n+1}+a_{n-1}}{B}.$$

The case $A=B=C=0$ corresponds to Eq. (5.19) and hence to the absence of an invariant interaction.

3. Decomposition 6=3+3

a. $\mathrm{sl}(2,\mathbb{R})_1 \oplus \mathrm{sl}(2,\mathbb{R})_2$. The algebra $\mathrm{sl}(2,\mathbb{R})_2$ is as in Eq. (4.8) and the invariant interaction is

$$F_n=\frac{1}{(\xi_n)^2}\left[u_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n(\chi_{n+1},\chi_{n-1})+u_nq_n(\chi_{n+1},\chi_{n-1})\right],$$

$$G_n=\frac{1}{(\xi_n)^2}\left[v_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n(\chi_{n+1},\chi_{n-1})+v_nq_n(\chi_{n+1},\chi_{n-1})\right], \tag{5.22}$$

$$\chi_{n+1}=\frac{\xi_{n+1}}{\xi_n},\quad \chi_{n-1}=\frac{\xi_{n-1}}{\xi_n}.$$

4. Decomposition 6=5+1

a. $\mathrm{saff}(2)\oplus A_1$. We have

$$Y_u(e^{a_nt})=e^{a_nt}\partial_{u_n},\quad Y_v(e^{a_nt})=e^{a_nt}\partial_{v_n},\quad T(a_n)=\partial_t+a_n(u_n\partial_{u_n}+v_n\partial_{v_n}).$$

The invariant interaction will be as in Eq. (5.4) with $\lambda_n=e^{a_nt}$. The functions ϕ_n and ψ_n will satisfy

$$\phi_n=e^{(a_n-a_{n-1}-a_{n+1})t}K_n(\omega),\quad \psi_n=e^{-a_{n+1}t}L_n(\omega),$$
$$\omega=e^{-(a_n+a_{n+1})t}\xi_{n+1}-e^{-(a_{n+1}+a_{n-1})t}\xi_n-e^{-(a_{n-1}+a_n)t}\xi_{n-1}. \tag{5.23}$$

5. Indecomposable symmetry algebras

It was shown in Sec. IV that two inequivalent gaff(2) symmetry algebras exist.

a. $\mathrm{gaff}(2,\mathbb{R})_1$.

$$Y_u(\lambda_n)=\lambda_n(t)\partial_{u_n},\quad Y_v(\lambda_n)=\lambda_n(t)\partial_{v_n},\quad V(1)=u_n\partial_{u_n}+v_n\partial_{v_n}.$$

The interaction is as in Eq. (5.4), however, ϕ_n and ψ_n depend only on t. This means that the equations are linear and, moreover, the equations (1.1) for u_n and v_n are decoupled.

b. $\mathrm{gaff}(2,\mathbb{R})_2$. The algebra is as in Eq. (4.22) [or (4.33)], the interaction as in Eq. (5.4) with $\lambda_n(t)=e^{(a_n-1)t}$. The functions ϕ_n and ψ_n satisfy

$$\phi_n=e^{-(a_{n+1}+a_{n-1}-a_n-1)t}K_n(\omega),\quad \psi_n=e^{(-a_{n+1}+1)t}L_n(\omega), \tag{5.24}$$

with ω as in Eq. (5.23).

E. Seven-dimensional symmetry groups

1. Decomposition 7=3+1+1+1+1

We exclude the case

$$V(a_{i,n})=a_{i,n}(u_n\partial_{u_n}+v_n\partial_{v_n}),\quad i=1,...,4,$$

since the only invariant interaction is (5.3) with h_n and k_n functions of t. We already know that the symmetry algebra is infinite dimensional.

a. $V(a_{i,n})=a_{i,n}(u_n\partial_{u_n}+v_n\partial_{v_n}),\ i=1,2,3,\ T(b_n)=\partial_t+b_n(u_n\partial_{u_n}+v_n\partial_{v_n})$. The interaction is as in Eq. (5.3) with h_n and k_n constants (depending on n). The algebra is actually infinite dimensional: we can take any number of operators $V(a_{i,n})$.

2. Decomposition 7=3+2+1+1

a. $V(a_{i,n})=a_{i,n}(u_n\partial_{u_n}+v_n\partial_{v_n})$, $i=1,2$, $T(0)=\partial_t$, $D(c_n)=t\partial_t+(\frac{1}{2}+c_n)(u_n\partial_{u_n}+v_n\partial_{v_n})$. We put $\gamma_n=c_n+\frac{1}{2}$. An invariant interaction exists if and only if we have

$$\Delta=\det\begin{pmatrix}\gamma_n & \gamma_{n+1} & \gamma_{n-1}\\ a_{1,n} & a_{1,n+1} & a_{1,n-1}\\ a_{2,n} & a_{2,n+1} & a_{2,n-1}\end{pmatrix}\neq 0. \tag{5.25}$$

The invariant interaction is that of Eq. (5.3), with

$$h_n=\eta^k p_n,\quad k_n=\eta^k q_n,\quad k=-\frac{2}{\Delta}. \tag{5.26}$$

The variable η is as in Eq. (5.8); p_n and q_n are constants.

3. Decomposition 7=3+3+1

a. $\mathrm{sl}(2,\mathbb{R})_1\oplus\mathrm{sl}(2,\mathbb{R})_2\oplus A_1$. We have $A_1=\{V(a_n)\}$. The invariant interaction can be obtained from Eq. (5.22). The additional invariance implied by the presence of $V(a_n)$ restricts p_n and q_n to

$$p_n=\left(\frac{\xi_{n+1}}{\xi_n}\right)^{2(a_{n+1}+a_{n-1})/(a_n-a_{n-1})} r_n(\omega),$$

$$q_n=\left(\frac{\xi_{n+1}}{\xi_n}\right)^{2(a_{n+1}+a_{n-1})/(a_n-a_{n-1})} s_n(\omega), \tag{5.27}$$

$$\omega=(\xi_{n+1})^{a_{n+1}-a_n}(\xi_{n-1})^{a_n-a_{n-1}}(\xi_n)^{a_{n-1}-a_{n+1}},$$

and we must impose $a_n\neq a_{n-1}$ (otherwise we have $F_n=G_n=0$).

4. Decomposition 7=6+1

The algebra $\mathrm{gaff}(2,\mathbb{R})_1$ does not allow any nonlinear interactions. Let us consider $\mathrm{gaff}(2,\mathbb{R})_2$ of Eq. (4.22).

a. $\mathrm{gaff}(2,\mathbb{R})_2\oplus\{U=u_n\partial_{u_n}+v_n\partial_{v_n}\}$. The interaction is as in Eq. (5.4), with ϕ_n and ψ_n as in Eq. (5.24). Invariance under the dilations corresponding to U implies that ϕ_n and ψ_n do not depend on ω. Hence, the interaction is linear and decoupled.

5. Indecomposable Lie algebras

a. $Y_u(\lambda_n)=\lambda_n(t)\partial_{u_n}$, $Y_v(\lambda_n)=\lambda_n(t)\partial_{v_n}$, $Y_u(\mu_n)=\mu_n(t)\partial_{u_n}$, $Y_v(\mu_n)=\mu_n(t)\partial_{v_n}$. We start from Eq. (5.4) with ϕ_n and ψ_n functions of t and ω as in Eq. (5.17). If ϕ_n and ψ_n do not depend on ω, the interaction is already linear and decoupled. Hence, ω must be invariant under the transformations corresponding to $Y_u(\mu_n)$ and $Y_v(\mu_n)$. This implies that λ_n and μ_n are independent of n. Further, invariance implies

$$\frac{\ddot{\lambda}_n}{\lambda_n}=\frac{\ddot{\mu}_n}{\mu_n}=\tilde{k}, \tag{5.28}$$

with $\tilde{k}=\mathrm{const}$. Equation (5.28) allows solutions,

$$\begin{pmatrix}\lambda_n\\ \mu_n\end{pmatrix}=\begin{pmatrix}\sin kt\\ \cos kt\end{pmatrix},\quad \begin{pmatrix}\sinh kt\\ \cosh kt\end{pmatrix},\quad \begin{pmatrix}1\\ t\end{pmatrix}. \tag{5.29}$$

These solutions are all equivalent under allowed transformations. We choose $\lambda_n=1$, $\mu_n=t$, i.e.,

$$Y_u(1)=\partial_{u_n},\quad Y_v(1)=\partial_{v_n},\quad Y_u(t)=t\partial_{u_n},\quad Y_v(t)=t\partial_{v_n}. \tag{5.30}$$

The invariant interaction is

$$\begin{aligned} F_n&=(u_{n+1}-u_{n-1})\phi_n(\omega,t)+(u_n-u_{n+1})\psi_n(\omega,t),\\ G_n&=(v_{n+1}-v_{n-1})\phi_n(\omega,t)+(v_n-v_{n+1})\psi_n(\omega,t), \end{aligned} \tag{5.31}$$

with

$$\omega=\xi_{n+1}-\xi_{n-1}-\xi_n. \tag{5.32}$$

b. $Y_u(1)=\partial_{u_n}$, $Y_v(1)=\partial_{v_n}$, $T(0)=\partial_t$, $D(b)=t\partial_t+(\frac{1}{2}+b)(u_n\partial_{u_n}+v_n\partial_{v_n})$, $b\neq-\frac{1}{2}$, $b=\text{const}$. The invariant interaction is as in Eq. (5.31), with

$$\phi_n=k_n\omega^{-2/(2b+1)},\quad \psi_n=p_n\omega^{-2/(2b+1)}, \tag{5.33}$$

with k_n and p_n constants, ω as in Eq. (5.32). For $b=-\frac{1}{2}$ there is no invariant interaction. For $b\neq-\frac{1}{2}$ the symmetry algebra is actually larger and includes $Y_u(t)=t\partial_{u_n}$ and $Y_v(t)=t\partial_{v_n}$.

F. Symmetry groups of dimensions 8, 9, and 10

By now, all invariant interactions have been specified up to arbitrary constants (depending on n), except those involving symmetry algebras containing the subalgebra $\mathrm{sl}(2,\mathbb{R})_1\oplus\mathrm{sl}(2,\mathbb{R})_2$, or the subalgebra $\{Y_u(1),Y_v(1),Y_u(t),Y_v(t)\}$ of Eq. (5.30). Let us consider the remaining nonlinear interactions.

1. $\mathrm{sl}(2,\mathbb{R})_1\oplus\mathrm{sl}(2,\mathbb{R})_2\oplus\{V(a_{1,n})\}\oplus\{V(a_{2,n})\}$

The invariant interaction is obtained from Eq. (5.27) by specifying $r_n(\omega)$ and $s_n(\omega)$ to be specific powers of ω. The result is

$$\begin{aligned} F_n&=\xi_n^{-2}\left[u_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n+u_nq_n\right](\xi_{n-1})^{-2A/D}(\xi_{n+1})^{-2B/D}(\xi_n)^{2[(A+B)/D]},\\ G_n&=\xi_n^{-2}\left[v_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n+v_nq_n\right](\xi_{n-1})^{-2A/D}(\xi_{n+1})^{-2B/D}(\xi_n)^{2[(A+B)/D]}. \end{aligned} \tag{5.34}$$

Here p_n and q_n are constants, A and B are as in Eq. (5.9), and

$$D=a_{1,n}(a_{2,n+1}-a_{2,n-1})+a_{1,n+1}(a_{2,n-1}-a_{2,n})+a_{1,n-1}(a_{2,n}-a_{2,n+1}). \tag{5.35}$$

We assume $D\neq0$; otherwise there is no invariant interaction. In particular, we have $a_{1,n}\neq a_{1,n+1}$, $a_{2,n}\neq a_{2,n+1}$.

2. Algebras containing $(Y_u(1),Y_v(1),Y_u(t),Y_v(t))$ *of (5.30) plus one additional operator* Z

The interaction is as in Eq. (5.31) with a restriction on ϕ_n and ψ_n.

(i) $Z=T(a)=\partial_t+a(u_n\partial_{u_n}+v_n\partial_{v_n})$, $a\equiv a_n=a_{n+1}$,

$$\phi_n=\phi_n(\eta),\quad \psi_n=\psi_n(\eta),\quad \eta=\omega e^{-2at}. \tag{5.36}$$

(ii) $Z=D(a)=t\partial_t+(\frac{1}{2}+a)(u_n\partial_{u_n}+v_n\partial_{v_n})$, $a\equiv a_n=a_{n+1}$,

$$\phi_n=\frac{1}{t^2}r_n(\eta),\quad \psi_n=\frac{1}{t^2}s_n(\eta),\quad \eta=\omega t^{-(2a+1)}. \tag{5.37}$$

(iii) $Z=R(b)=(t^2+1)\partial_t+(t+b)(u_n\partial_{u_n}+v_n\partial_{v_n}),\ b\equiv b_n=b_{n+1}$,

$$\phi_n=\frac{1}{(t^2+1)^2}r_n(\eta),\quad \psi_n=\frac{1}{(t^2+1)^2}s_n(\eta),$$
$$\eta=\frac{\omega}{1+t^2}e^{-2b\arctan t}, \tag{5.38}$$

with ω as in Eq. (5.32) in all cases.

(iv) $Z=V(1)$. Then ϕ_n and ψ_n depend only on t and the interaction is linear.

We can add two operators to those of Eq. (5.30)

$$T(0)=\partial_t,\quad D(b)=t\partial_t+(\tfrac{1}{2}+b)(u_n\partial_{u_n}+v_n\partial_{v_n}).$$

The invariant interaction coincides with that of Eq. (5.33).

Finally, the interaction (5.31) is invariant under a ten-dimensional symmetry algebra of the form

$$(\mathrm{sl}(2,\mathbb{R})_1\oplus \mathrm{sl}(2,\mathbb{R})_2)\rhd\{Y_u(1),Y_v(1),Y_u(t),Y_v(t)\},$$

for

$$\phi_n=k_n\omega^{-2},\quad \psi_n=p_n\omega^{-2}, \tag{5.39}$$

i.e., $b=0$ in Eq. (5.33).

VI. SUMMARY AND CONCLUSIONS

Let us first sum up the results on invariant interactions and the corresponding symmetry algebras. We shall follow the summary of possible symmetry algebras outlined in Sec. IV E. The results are presented in the following tables.

Table I. The Series A of symmetry algebras. The algebra L_C of Eq. (4.27) consists entirely of $\mathrm{sl}(2,\mathbb{R})_1$ singlets. In the first column of Table I we list the symmetry algebras. The number in brackets [e.g., $A_1(3)$] denotes the dimension of the symmetry algebra. The notation for basis elements in column 2 are as in Eqs. (4.13)–(4.18). Note that if the functions h_n and k_n in the interaction (5.3) depend only on t or are constants, then the symmetry algebra is infinite dimensional, although the interaction is nonlinear.

The case $A_3(7)$ corresponds to an algebra L with $\dim L=7$ and the interaction is completely specified [see (5.3), (5.25)–(5.26)]. In other cases the functions h_n and k_n depend on one, two, or three variables involving u_k and v_k.

Table II. The Series B of symmetry algebras. The symmetry algebras are either five or six dimensional. The interactions are as in Eq. (5.4) and involve two arbitrary functions, ϕ_n and ψ_n. A B-type symmetry allows ϕ_n and ψ_n to depend on just one variable involving u_k and v_k. Any extension of the B-type algebras will restrict $\lambda_n(t)$ to be $\lambda_n=1$ and will involve a further pair with $\lambda_n=t$. This takes us into the series C of symmetry algebras.

The algebras B_2, B_6, and B_7 of Eqs. (4.32), (4.36), and (4.37) lead to linear interactions. Any interaction invariant with respect to B_5 will be invariant under a larger group, corresponding to a Lie algebra in the series C. We do not include linear interactions in the tables and we list interactions together with their *maximal* symmetry algebras.

Table III. The Series C of symmetry algebras. The interaction will be as in Eq. (5.31), involving a variable ω as in Eq. (5.32). The algebras $C_5(8),C_7(9),C_8(9),C_9(9),C_{11}(10),C_{12}(11)$, absent in the table, lead to a linear interaction.

TABLE I. Series A of symmetry algebras. The interaction has the form (5.3).

No.	L_C	Restrictions on h_n and k_n	Variables and comments
$A_1(3)$	$\cdots$	$\cdots$	$t,\xi_{n+1},\xi_{n-1},\xi_n$ (5.2)
$A_1(4)$	$V(a_n)$	$\cdots$	$\begin{cases} t,\eta_{n+1},\eta_{n-1} & (5.5) \\ t,\xi_n,\eta & (5.6)\end{cases}$
$A_1(5)$	$V(a_{1,n}),V(a_{2,n})$	$\cdots$	t, η (5.8)
$A_1(\infty)$	$V(a_{i,n}),i\in\mathbb{Z}^{>}$	$\cdots$	t
$A_2(4)$	$T(b_n)$	$\cdots$	$\zeta_{n+1},\zeta_{n-1},\zeta_n$ (5.7)
$A_2(5)$	$T(b_n),V(a_n)$	$\cdots$	$\begin{cases} \rho_{n-1},\rho_{n+1} & (5.10) \\ \rho_n,\sigma_n & (5.11)\end{cases}$
$A_2(6)$	$T(b_n),V(a_{1,n}),V(a_{2,n})$	$\cdots$	η (5.18)
$A_2(\infty)$	$T(b_n),V(a_{k,n}),k\in\mathbb{Z}^{>}$	h_n,k_n constants	None
$A_3(5)$	$T(0),D(b_n)$	(5.12) or (5.14)	(5.13) or (5.15)
$A_3(6)$	$T(0),D(c_n),V(a_n)$	(5.20)	ω (5.20)
$A_3(7)$	$T(0),D(c_n),V(a_{1,n})V(a_{2,n})$	(5.26)	None

For $C_6(9)$ and $C_{10}(10)$ the interactions are specified up to constants (that can depend on n). In all other cases, the arbitrary functions depend on one variable, involving u_k and v_k.

Table IV. The Series D of symmetry algebras. There are three such algebras of dimension 6, 7, and 8, respectively. They all lead to nontrivial invariant interactions of the form (5.22). For $D_3(8)$, the interaction is completely specified. We do not list $D_4(10)$ in Table IV since it coincides with $C_{10}(10)$ of Table III. The algebra $D_5(11)$ corresponds to a linear interaction.

For each interaction we have verified that the given symmetry algebra is the maximal one.

A few words about the interpretation of the invariant interactions. From Eq. (5.1) and the variables (5.2) we see that invariance under $\mathrm{sl}(2,\mathbb{R})_1$ imposes very strong restrictions.

(1) In particular, if the interaction is linear and $\mathrm{sl}(2,\mathbb{R})_1$ invariant, we must have

$$F_n=\sum_{k=n-1}^{n+1} A_k(t)u_k, \quad G_n=\sum_{k=n-1}^{n+1} A_k(t)v_k, \tag{6.1}$$

i.e., the equations (1.1) for u_k and v_k decouple (into identical equations for u_n and v_n separately).

(2) If the interaction terms F_n and G_n in Eq. (5.1) are nonlinear, they always involve many-body forces. That is, they cannot be written as sums of terms of the type $h_n(u_n,v_n)$ or $h_n(u_n,v_{n+1})$, etc. Indeed, each invariant variable $\xi_n,\xi_{n+1},\xi_{n-1}$ itself involves four of the original variables u_i,v_i simultaneously. This many-body character becomes more pronounced when the invariance algebra is larger.

(3) The operators $V(a_n)$ correspond to site-depending dilations,

TABLE II. Series B of symmetry algebras. The algebra includes one pair $Y_u(\lambda_n)$, $Y_v(\lambda_n)$. The interaction has the form (5.4).

No.	Restrictions on λ_n, additional Elements of L_C	Restrictions on ϕ_n and ψ_n	Variables and comments
$B_1(5)$	$\cdots$	$\cdots$	t, ω as in (5.17)
$B_4(6)$	$\lambda_n=e^{a_n t},T(a_n)$	(5.23)	ω (5.23)
$B_3(6)$	$\lambda_n=e^{(a_n-1)t},T(a_n)$	(5.24)	ω (5.23)

TABLE III. Series C symmetry algebras. The algebras contain $\mathrm{sl}(2,\mathbb{R})_1, Y_u(1), Y_v(1), Y_u(t), Y_v(t)$, and possibly additional elements. The interaction is as in Eq. (5.31).

No.	Additional elements	Conditions on ϕ_n and ψ_n	Variables
$C_1(7)$	–	$\cdots$	ω, t (5.32)
$C_2(8)$	$T(a)$	$\cdots$	$\eta=\omega e^{-2at}$
$C_3(8)$	$D(a)$	$\phi_n=t^{-2}r_n(\eta),\ \psi_n=t^{-2}s_n(\eta)$	$\eta=\omega t^{-(2a+1)}$
$C_4(8)$	$R(b)$	$\phi_n=(t^2+1)^{-2}r_n(\eta)$, $\psi_n=(t^2+1)^{-2}s_n(\eta)$	$\eta=\omega(t^2+1)^{-1}$ $e^{-2b\ \arctan t}$
$C_6(9)$	$T(0),D(a)$	$\phi_n=k_n\omega^{-2/(2a+1)},\ \psi_n=p_n\omega^{-2(2a+1)}$ $k_n,\ p_n$ constants, $2a+1\neq 0$	None
$C_{10}(10)$	$T(0),D(0),P(0)$	$\phi_n=k_n\omega^{-2},\ \psi_n=p_n\omega^{-2}$	None

$$\tilde{u}_n=e^{\epsilon a_n}u_n, \quad \tilde{v}_n=e^{\epsilon a_n}v_n. \tag{6.2}$$

Invariance under two such one-dimensional symmetry groups, generated by $\{V(a_{1,n}),V(a_{2,n})\}$, where $a_{1,n}$ and $a_{2,n}$ are two linearly independent functions of n, introduces the symmetry variable

$$\omega_D\equiv(\xi_{n-1})^A(\xi_{n+1})^B(\xi_n)^C, \tag{6.3}$$

as in Eq. (5.8). Here all six variables are coupled together.

(4) The pair of operators $Y_u(\lambda_n),Y_v(\lambda_n)$ induces site-dependent (and time-dependent) shifts of the dependent variables,

$$\tilde{u}_n=u_n+\epsilon\lambda_n(t), \quad \tilde{v}_n=v_n+\epsilon\lambda_n(t). \tag{6.4}$$

The corresponding invariant variable again involves all six variables [see Eq. (5.17)],

$$\omega_T\equiv\lambda_{n-1}\xi_{n+1}-\lambda_n\xi_n-\lambda_{n+1}\xi_{n-1}. \tag{6.5}$$

A special case of the variable ω_T is obtained setting $\lambda_n=\lambda_{n-1}=\lambda_{n+1}=1$. This is the case of Eq. (5.32), where

$$\omega=\omega_S=\xi_{n+1}-\xi_n-\xi_{n-1} \tag{6.6}$$

is invariant with respect to two such translations:

$$\tilde{u}_n=u_n+\epsilon_1+\epsilon_2 t, \quad \tilde{v}_n=v_n+\epsilon_1+\epsilon_2 t \tag{6.7}$$

(ϵ_1 and ϵ_2 are group parameters and hence constants).

A continuation of this study is in progress. It involves several aspects.

The first is a study of the integrability properties of the equations that are completely specified by their symmetries. These are, first of all, those with infinite-dimensional symmetry groups, namely

$$\ddot{u}_n=u_{n+1}\frac{\xi_{n-1}}{\xi_n}h_n+u_nk_n, \quad \ddot{v}_n=v_{n+1}\frac{\xi_{n-1}}{\xi_n}h_n+v_nk_n, \tag{6.8}$$

TABLE IV. Series D of symmetry algebras. The algebra contains $\mathrm{sl}(2,\mathbb{R})_1\oplus\mathrm{sl}(2,\mathbb{R})_2$. The interaction has the form (5.22).

No.	Additional elements in L_C	Conditions on p_n and q_n	Variables
$D_1(6)$	$\cdots$	$\cdots$	χ_{n+1}, χ_{n-1} as in (5.22)
$D_2(7)$	$V(a_n)$	(5.27)	η as in (5.27)
$D_3(8)$	$V(a_{1,n}),\ V(a_{2,n})$	(5.34)	$\cdots$

with h_n and k_n functions of t or constants [see $A_1(\infty)$ and $A_2(\infty)$ in Table I].

Completely specified equations with finite-dimensional symmetry algebras L are the following ones.

(i)

$$\ddot{u}_n=\left(u_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n+u_nq_n\right)\omega_D^{-2/\Delta}, \quad \ddot{v}_n=\left(v_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n+v_nq_n\right)\omega_D^{-2/\Delta}, \tag{6.9}$$

with ω_D as in Eq. (6.3), Δ as in Eq. (5.25). This is case $A_3(7)$ of Table I.

(ii)

$$\ddot{u}_n=[(u_{n+1}-u_{n-1})p_n+(u_n-u_{n+1})q_n]\omega_S^{-2/(2a+1)},$$

$$\ddot{v}_n=[(v_{n+1}-v_{n-1})p_n+(v_n-v_{n+1})q_n]\omega_S^{-2/(2a+1)}, \tag{6.10}$$

with ω_S as in Eq. (6.6), $p_n, q_n, a\neq -\frac{1}{2}$ const. This is case $C_6(9)$ of Table III.

(iii) For $a=0$, Eq. (6.10) is invariant under a ten-dimensional symmetry algebra, namely $C_{10}(10)$ of Table III.

(iv)

$$\ddot{u}_n=(\xi_{n-1})^{-2A/k}(\xi_{n+1})^{-2B/D}(\xi_n)^{[2(A+B-D)/D]}\left[u_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n+u_nq_n\right],$$

$$\ddot{v}_n=(\xi_{n-1})^{-2A/D}(\xi_{n+1})^{-2B/D}(\xi_n)^{[2(A+B-D)/D]}\left[v_{n+1}\frac{\xi_{n-1}}{\xi_n}p_n+v_nq_n\right],$$

with p_n and q_n depending only on n. The constants A and B are given in Eq. (5.9), D in Eq. (5.35).

A further task is to complete the classification, that is, to treat the cases of other sl(2,$\mathbb{R}$) algebras and also of solvable symmetry algebras.

ACKNOWLEDGMENTS

The authors thank D. Levi and M. A. Rodriguez for helpful discussions. The research of S.L. and P.W. was partly supported by the NSERC of Canada and FCAR du Québec. S. L. would like to thank the Departamento de Física Teórica II de la Universidad Complutense for their hospitality during his stay in Madrid. D.G.U.'s work was partly supported by DGES Grant No. PB95-0401. He would like to express his gratitude to the Centre de Recherches Mathématiques for their kind hospitality.

[1] A. Campa, A. Giansanti, A. Tenenbaum, D. Levi, and O. Ragnisco, Phys. Rev. B **48**, 10 168 (1993).
[2] A. C. Scott, Phys. Rep. **217**, 1 (1992).
[3] S. Pneumaticos, N. Flytzanis, and M. Remoissenent, Phys. Rev. B **33**, 2308 (1986).
[4] D. Levi and P. Winternitz, J. Math. Phys. **37**, 5551 (1996).
[5] D. Levi and P. Winternitz, Phys. Lett. A **152**, 335 (1991).
[6] D. Levi and P. Winternitz, J. Math. Phys. **34**, 3713 (1993).
[7] D. Levi, L. Vinet, and P. Winternitz, J. Phys. A **30**, 633 (1997).
[8] S. Maeda, Math. Japonica **25**, 405 (1980); **26**, 85 (1981).
[9] R. Quispel, H. W. Capel, and R. Sahadevan, Phys. Lett. A **170**, 379 (1992).
[10] V. A. Dorodnitsyn, J. Sov. Math. **55**, 1490 (1991).
[11] V. A. Dorodnitsyn, in *Symmetries and Integrability of Difference Equations*, edited by D. Levi, L. Vinet, and P. Winternitz (AMS, Providence, RI, 1995).
[12] R. Floreanini, J. Negro, L. M. Nieto, and L. Vinet, Lett. Math. Phys. **36**, 351 (1996).
[13] R. Floreanini and L. Vinet, J. Math. Phys. **36**, 7024 (1995).
[14] J. P. Gazeau and P. Winternitz, Phys. Lett. A **167**, 246 (1992); J. Math. Phys. **33**, 4087 (1992).
[15] N. Jacobson, *Lie Algebras* (Dover, New York, 1979).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Dissipative canonical flows in classical and quantum mechanics

John Gough
Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland

(Received 20 February 1998; accepted for publication 14 February 1999)

A theory of stochastic flows over the algebra of observables of a dynamical system is presented in which the main objective is to ensure that the overall canonical/symplectic structure on the algebra is preserved. We study both classical and quantum systems and the importance of physical interpretation in the Stratonovich interpretation is stressed. We find the natural formulation of quantum dissipative systems to be given in terms of quantum stochastic calculus. This treatment allows for a physically meaningful treatment of both constant and nonlinear dissipation. As an application, we quantize a mechanical system with the same nonlinear damping mechanism as the van der Pol oscillator. © *1999 American Institute of Physics.* [S0022-2488(99)00206-6]

I. INTRODUCTION

The theory of open systems has been developed widely in both classical and quantum situations. To quantize a classical dynamical system, it is necessary to have an underlying Hamiltonian mechanism and there is a problem whenever the dynamical system is dissipative. By dissipative, we mean that the Poisson structure is not preserved under the dynamical evolution. Typically, an open system is viewed as a subcomponent of a larger Hamiltonian system and the dynamical equations of the system (observed on its own) are approximations to the actual regular motion. In this paper we consider classical stochastic flows wherein the fluctuation-dissipation has a geometric content, or perhaps more exactly a symplecto-geometric content. That is, the fluctuations balance the Poisson structure dissipation. In doing so, we discover the natural way in which to compare classical open systems to quantum ones and, in the process, learn how to quantize some standard classical dissipative models.

The program to quantize dissipative mechanical systems is now at least easy to state; first quantize the system-reservoir and then trace out the reservoir degrees of freedom. Under suitable assumptions one may obtain Markovian approximations for the reduced system evolutions, the correct approach to this is detailed in Ref. 1 and also Chap. XVI of Ref. 2 for a review and further references.

Before pursuing this program further, it is important to take a closer look at the difficulties encountered in the classical case. In particular, we find that many of the mathematical treatments of quantum dissipative processes designed to parallel the Langevin approach fall down at the most fundamental level of physical interpretation. The Langevin approach, as emphasized by van Kampen especially in Sec. IX.5 of Ref. 2, cannot be applied indiscriminately to problems where the noise coefficient, denoted by σ in this paper, is nonlinear in the variables of the system. A common mistake is to construct a Langevin equation by taking a deterministic macroscopic term and adding on a noise term with coefficient σ arrived at by independent considerations. Van Kampen blames this on the disregard of physicists for possible physical origins of the fluctuations and on a belief that ''...*stochasticity is part and parcel of the mathematics and requires no physical cause*.'' This is, if anything doubly true in the quantum domain. The essential problem is, of course, that Langevin approach is justifiable only in the Stratonovich interpretation of stochastic calculus while it is in the Ito interpretation where most research is formulated. In point of fact, the

Ito version is the one appropriate for mathematical analysis while the Stratonovich is the one appropriate for physical insight.

In this paper we examine the notion of canonical flows as the correct description of the evolution of physical functions of phase variables. The classical situation is described first and we introduce both deterministic and stochastic canonical flows and study the latter in both Ito and Stratonovich interpretations. We show that the quantum analog of this is given in terms of quantum stochastic processes in the sense of Hudson and Parthasarathy,[3] in particular we employ the notion of a quantum stochastic flows and derivations which was introduced by Accardi and Hudson.[4] As an application of our ideas we construct a quantum process with the same nonlinear damping mechanism as the van der Pol oscillator. We conclude with several remarks on the physical status of our investigations.

Notations: Let $\mathcal{A}$ be a unital $*$-algebra throughout. The algebra of observables of a dynamical system is described as a Poisson manifold, that is, a pair $(\mathcal{A}, \wp)$ where $\wp$ gives a Poisson structure. In classical systems, the geometry (kinematics) is distinct from Poisson structure (dynamics), however, for quantum systems this is not so as the Poisson structure is typically algebraic. We outline some definitions below to fix our ideas.

Definition 1: A Poisson Structure $\wp$ on $\mathcal{A}$, compatible with the $*$-involution, is a bilinear map $\wp:\mathcal{A}\times\mathcal{A}\mapsto\mathcal{A}$ satisfying

(i) $\wp(a,1)=0$;
(ii) $\wp(a,b)=-\wp(b,a)$;
(iii) $\wp(a,bc)=\wp(a,b)c+b\wp(a,c)$;
(iv) $\wp(a,\wp(b,c))+\wp(b,\wp(c,a))+\wp(c,\wp(a,b))=0$;
(v) $\wp(a,b)^*=\wp(a^*,b^*)$.

Definition 2: The distension of a $*$-linear map λ on $\mathcal{A}$ is the bilinear map $\mathcal{D}(\lambda)$ defined by

$$\mathcal{D}(\lambda;a,b):=\lambda(ab)-\lambda(a)b-a\lambda(b). \tag{1.1}$$

Definition 3: A $*$-linear map μ: $\mathcal{A}\mapsto\mathcal{A}$ is said to be canonical with respect to a Poisson structure $\wp$ on $\mathcal{A}$ if

$$\mu\wp(a,b)=\wp(\mu a,b)+\wp(a,\mu b). \tag{1.2}$$

The dissipation of a $*$-linear map λ with respect to the Poisson structure $\wp$ on $\mathcal{A}$ is the bilinear map $D(\lambda)$ defined by

$$D(\lambda;a,b):=\lambda\wp(a,b)-\wp(\lambda a,b)-\wp(a,\lambda b). \tag{1.3}$$

The distension measures how far a map is from being a derivation and as such is an algebraic concept independent of whether a Poisson structure exists. The dissipation measures how far a map deviates from being a canonical with respect to a given Poisson structure and as such is a dynamical concept. The set of derivations is $\mathrm{Ker}\,\mathcal{D}$. We take $\mathrm{Can}(\Gamma, \wp)$ to be the kernel space of D when restricted to $\mathrm{Ker}\,\mathcal{D}$, that is, the set of canonical derivations.

Definition 4: For each $H\in\mathcal{A}_S$, the Hamiltonian map X_H generated by H is defined to be

$$X_H(a):=\wp(a,H). \tag{1.4}$$

By property (iii) of the Poisson structure, Hamiltonian maps are derivations. By property (iv), they are also canonical with respect to $\wp$. We denote the set of Hamiltonian maps by $\mathrm{Ham}(\Gamma, \wp)$.

Definition 5: A Poisson algebra $\mathcal{A}$ is said to be $\wp$-simple if every canonical map is Hamiltonian (wrt. $\wp$).

II. CLASSICAL FLOWS

Let $\mathcal{A}$ denote the $*$-algebra of observables of a dynamical system. A stationary flow $(u_t)_{t\geqslant 0}$ is a continuous one-parameter group of $*$-homomorphisms on $\mathcal{A}$ which, by Stone's theorem, possesses a generator $\mathcal{L}$; that is $u_t=\exp\{t\mathcal{L}\}$. Let $u_t^0=\exp\{t\mathcal{L}^0\}$ be a second stationary flow which we take as reference or unperturbed dynamics. The perturbation is then described by $\mathcal{L}-\mathcal{L}^0$. A standard device is to transfer to the interaction picture of the flow u_t with respect to u_t^0 which is given in terms of the family $(j_t)_{t\geqslant 0}$ of automorphisms defined by

$$j_t:=u_t\circ(u_t^0)^{-1}. \tag{2.1}$$

The family $(j_t)_{t\geqslant t}$ does not form a semigroup but instead is a u^0-cocycle, that is,

$$j_{t+s}=j_s\circ u_s^0\circ j_t\circ(u_s^0)^{-1}. \tag{2.2}$$

The dynamical equations of motion are $du_tf=u_t(\mathcal{L}f)dt$ which, when transferred to the interaction picture, become

$$dj_tf=j_t(\mathcal{L}_t'f)dt, \tag{2.3}$$

where $\mathcal{L}_t':=u_t^0\circ\mathcal{L}'\circ(u_t^0)^{-1}$.

Stochastic flows on phase space

Let Γ be the phase space (finite dimensional manifold) of a classical dynamical system and $(\Gamma, \mathcal{G}, \mu)$ a probability space. Let B_t be the Wiener process with standard representation on the Wiener space $(\Omega,\mathcal{F},P)$ and let $\mathcal{F}_{t]}$, $\mathcal{F}_{(t}$ denote the past, respectively, future filtrations generated by the process at time t. Consider the SDE on Γ given by

$$dX_t=b(X_t)dt+\sigma(X_t)dB_t, \tag{2.4}$$

where b and σ are vector fields on Γ satisfying global Lipschitz conditions so as to guarantee uniqueness of the solution.[5] For Y a random variable independent of $\mathcal{F}_{(s}$, we shall denote by $X_t^{(Y,s)}$ the solution to the SDE for times $t>s$ which satisfies the initial condition $X_s=Y$. By uniqueness of the solution we have the identity P-a.s.

$$X_{t+s}^{(x_0,0)}=X_t^{(X_s^{(x_0,0)},s)}, \quad \forall x_0\in\Gamma, \tag{2.5}$$

however, it is more useful to refer only to diffusions starting at initial time zero and so in place of (2.5) we have P-a.s.,

$$X_{t+s}^{(x_0,0)}(\omega)=X_t^{(X_s^{(x_0,0)}(\omega),0)}(\theta_s\omega), \tag{2.6}$$

where θ_s is the time shift map on the Ω give by $(\theta_s\omega)_t=\omega_{t+s}-\omega_s$ for each sample path $\omega\in\Omega$.

A family of $*$-homomorphisms $j_t:L^\infty(\Gamma,\mu)\mapsto L^\infty(\Gamma,\mu)\otimes\mathcal{B}(L^2(\Omega,\mathcal{F},P))$ is defined for $t>0$ by

$$j_tf(x_0,\omega):=f\circ X_t^{(x_0,0)}(\omega); \tag{2.7}$$

j_t is then trivially extended from $L^\infty(\Gamma,\mu)$ to $L^\infty(\Gamma,\mu)\otimes\mathcal{B}(L^2(\Omega,\mathcal{F},P))$ by setting $j_t(f\otimes\xi_{(t)}):=(j_tf)\otimes\xi_{(t}$ for $\xi_{(t}\in\mathcal{B}(L^2(\Omega,\mathcal{F}_{(t},P))$.

Introducing the time shift u_t^0 induced by θ_t, that is $u_t^0Z(\omega):=Z(\theta_t\omega)$, the following property is readily obtained

$$j_{t+s}f=j_s(u_s^0(j_t(f))). \tag{2.8}$$

Bearing in mind that the u^0 act trivially on $L^\infty(\Gamma,\mu)$, it follows that $(j_t)_{t\geqslant 0}$ is a u_0-cocycle.

III. CLASSICAL CANONICAL FLOWS

Let $\mathcal{A}=\mathcal{C}^\infty(\Gamma)$ be the algebra of smooth complex functions on phase space with complex conjugation being involution. We shall denote a Poisson structure $\wp$ on $\mathcal{A}$ using bracket notation. The existence of a Poisson structure is equivalent to the existence of a second order tensor Λ such that

$$\{f,g\}=\Lambda(df,dg). \tag{3.1}$$

In terms of local coordinates (x^i) on Γ, $\{f,g\}=\Lambda^{ij}f_{,i}g_{,j}$ and the components Λ^{ij} are required to satisfy the conditions

$$\Lambda^{ij}+\Lambda^{ji}=0; \quad \Lambda^{il}\Lambda^{jk}{}_{,l}+\Lambda^{jl}\Lambda^{ki}{}_{,l}+\Lambda^{kl}\Lambda^{ij}{}_{,l}=0. \tag{3.2}$$

In general $\mathrm{Ham}(\mathcal{A},\wp)\subseteq\mathrm{Can}(,\wp)\subseteq T\Gamma$. For $\wp\equiv 0$, every tangent vector field is canonical but only the zero vector field is Hamiltonian. We, however, shall only deal with the situation where the algebra is simple with respect to Poisson bracket structure. Two important examples are given below.[6]

(i) *Symplectic manifolds.* Let $(\Gamma,\ \omega)$ be a $2n$-dimensional symplectic manifold. Poisson brackets are given by $\{f,g\}=\omega(X_f,X_g)$ and in this way the mapping $f\mapsto X_f$ defines a Lie algebra homomorphism. The Hamiltonian vector fields span $T\Gamma$ since the symplectic form ω is, by definition, nondegenerate. The prototype for a symplectic manifold is the cotangent space T^*M over an n-dimensional manifold (configuration space) and there always exist local coordinates $(q^1,\dots,q^n,p_1,\dots,p_n)$, called canonical coordinates, such that $\omega=dq^i\wedge dp_i$.

(ii) *Angular momentum algebra.* Take Γ to be three-dimensional Euclidean space with Cartesian coordinates (x^1,x^2,x^3) and define Λ as the tensor with Cartesian components $\Lambda^{ij}=\epsilon_{ijk}x^k$. This leads to a degenerate Poisson structure since any function Φ of $r=\sqrt{(x^1)^2+(x^2)^2+(x^3)^2}$ has the property that $\{f,\Phi(r)\}=0$, for all $f\in C^\infty(\Gamma)$. In fact, the Hamiltonian vector fields span a two-dimensional subspace of the tangent space to Γ at any point.

In general, let (x^i) be a local coordinate system on Γ, then the map $v=v^i\partial_i\in T\Gamma$ is characterized by the dissipation tensor $D^{ij}(v)$ given by

$$D^{ij}(v):=D(v;x^i,x^j)\equiv v^k\Lambda^{ij}{}_{,k}-v^i{}_{,k}\Lambda^{kj}-v^j{}_{,k}\Lambda^{ik}. \tag{3.3}$$

A. Stochastic symplectic flows

A stochastic flow on Γ can be described by the system of SDEs,

$$d_IX_t^i=\tilde{v}^i(X_t)\quad dt+\sigma^i(X_t)d_IB_t. \tag{3.4}$$

This equation is termed nonlinear if components of σ depend on the phase point x. Here, B_t is the Wiener process and we take the differentials to be the Ito type, that is $d_IZ(t):=Z(t+dt)-Z(t)$.

For $f\in\mathcal{A}$, we obtain for the corresponding cocycle j_t,

$$d_Ij_tf=j_t(d_ILf), \tag{3.5}$$

where, by the Ito rules of stochastic calculus,

$$d_Ij_tf:=j_t(\tilde{v}^if_{,i}+\tfrac{1}{2}f_{,ij}\sigma_i\sigma_j)dt+j_t(\sigma^if_{,i})d_IB_t, \tag{3.6}$$

that is,

$$d_Ij_t=j_t\circ l\,dt+j_t\circ\sigma d_IB_t, \tag{3.7}$$

where

$$l := \tilde{v} + \tfrac{1}{2}\sigma^i \sigma^j \partial_i \partial_j , \quad \tilde{v} := \tilde{v}^i \partial_i , \text{ and } \sigma = \sigma^i \partial_i . \tag{3.8}$$

It is convenient to introduce a shifted vector field v defined by

$$v^i = \tilde{v}^i - \tfrac{1}{2}\sigma^j \sigma^i{}_{,j} \tag{3.9}$$

in which case the second order partial differential operator l becomes

$$l = v + \tfrac{1}{2}\sigma \circ \sigma. \tag{3.10}$$

Definition 6: A stochastic flow j_t is canonical if

$$d_I j_t \{f,g\} = \{d_I j_t f, j_t g\}_t + \{j_t f, d_I j_t g\}_t + \{d_I j_t f, d_I j_t g\}_t ; \quad \forall f,g \in \mathcal{A}, \tag{3.11}$$

where $\{a,b\} := j_t \{j_t^{-1} a, j_t^{-1} b\}$.

A formal infinitesimal generator $d_I L$ is then given by

$$d_I L := j_t^{-1} d_I j_t \equiv l\,dt + \sigma B_{dt} , \tag{3.12}$$

where we used the identity $j_t^{-1} d_I B_t = (u_t^0)^{-1}(B_{t+dt} - B_t) = B_{dt}$. As such the condition reduces to

$$d_I L\{f,g\} = \{d_I L f, g\} + \{f, d_I L g\} + \{d_I L f, d_I L g\} \tag{3.13}$$

or

$$(D(v;f,g) + \tfrac{1}{2}\sigma \circ \sigma\{f,g\} - \tfrac{1}{2}\{\sigma \circ \sigma f, g\} - \tfrac{1}{2}\{f, \sigma \circ \sigma g\} + \{\sigma f, \sigma g\})dt + D(\sigma; f,g) B_{dt} = 0. \tag{3.14}$$

Both the dt and B_{dt} coefficients must vanish. Clearly the dissipation of σ must be zero and so σ is Hamiltonian, say $\sigma = \{.,F\} \equiv X_f$. Next of all, if we note the identity

$$X_F \circ X_F \{f,g\} = \{X_F \circ X_F f, g\} + 2\{X_F f, X_F g\} + \{f, X_F \circ X_F g\},$$

then the vanishing of the dt coefficient reduces to the requirement $D(v;f,g) = 0$ and so v must likewise be Hamiltonian, say $v = X_H$.

Theorem: A stochastic flow j_t, driven by a single Wiener process, is canonical if and only if there exist $H,F \in C^\infty(\Gamma)$ such that

$$d_I j_t(f) = j_t(\{f,H\} + \tfrac{1}{2}\{\{f,F\},F\})dt + j_t(\{f,F\}) d_I B_t . \tag{3.15}$$

B. Remarks

(1) If $d_I L_{(\alpha)}$ are infinitesimal generators of canonical stochastic flows, then so too is $\Sigma_\alpha d_I L_{(\alpha)} + \Sigma_{\alpha \neq \beta} d_I L_{(\alpha)} \circ d_I L_{(\beta)}$. In this way the generators form an algebra. The nontensorality and nonadditivity of generators comes from the Ito calculus. If we switch to the Stratonovich calculus we find that the SDE (3.4) becomes

$$d_S X_t^i = v^i(X_t) dt + \sigma^i(X_t) d_S B_t , \tag{3.16}$$

while in place of (3.6) we get

$$d_S j_t(f) = j_t(vf) dt + j_t(\sigma f) d_S B_t . \tag{3.17}$$

Thus we see that v is the phase velocity in the Stratonovich calculus. The infinitesimal generator in Stratonovich form then reduce to first order form, they can then be added vectorially, and so the apparent nonadditivity emerges only when we convert back to Ito form.

Moreover, the condition (3.11) for a flow to be canonical now reads as $d_S\{f_t,g_t\}_t=\{d_Sf_t,g_t\}_t+\{f_td_Sg_t\}_t$ since the ordinary rules of calculus now hold. In fact, we may write

$$d_Sj_t=j_t\circ X_H dt+j_t\circ X_F d_SB_t \tag{3.18}$$

which suggests that we may partially recover the picture of deterministic flows by introducing a formal stochastic Hamiltonian $F_t=H+Fd_SB_t/dt$ were we able to give an appropriate meaning to the derivative of the Wiener process in Stratonovich calculus.

From the physical point of view this is reasonable. The SDEs using the Wiener process can only possibly be approximations to a regular random dynamical evolution. As is well-known,[12] if the system was governed by the actual Hamiltonian $H+FdB_t^{(\lambda)}/dt$ with $B_t^{(\lambda)}$ a regular stochastic process which converged to the Wiener process as, say, $\lambda\to 0$, then this limit would lead to the Stratonovich SDEs (3.16–3.17). This gives strong hope that the stochastic flows can be derived from microscopic dynamics.

(2) Define the forward derivative of an adapted stochastic process Z_t to be $D_t^+Z_t:=\lim_{r\to 0}E_{t]}[(Z_{t+r}-Z_t)/r]$. We have $D_t^+j_t=j_t\circ l$ and specifically $D_t^+X_t^i=\tilde{v}^i(X_t)$. As remarked above, $\tilde{v}$ is typically not covariant with respect to (deterministic) canonical transformations.

(3) The differential operator l is an example of what is known as a second order tangent vector field and their relevance to stochastic processes on manifolds is well-known.[7]

(4) So far we have worked exclusively with diffusions, however, the generalization to continuous semimartingales, in the present context, does not change our findings.[7]

(5) In a local coordinate system (x^i), a second order vector λ can be uniquely decomposed as $\lambda_I+\lambda_{II}$ where $\lambda_I=\lambda_I^i\partial_i$ and $\lambda_{II}=\lambda_{II}^{ij}\partial_i\partial_j$ (with λ_{II}^{ij} symmetric). We have $D(\lambda)=D(\lambda_I)+D(\lambda_{II})$, however, the dissipation tensor only gives characteristics of λ_I since $D^{ij}(\lambda_{II})\equiv 0$.

As a result, for l and $\tilde{v}$ as above, we have $D^{ij}(\tilde{v})=D^{ij}(l)$. Now $D^{ij}(l)=\frac{1}{2}D^{ij}(\sigma\circ\sigma)\equiv\{\sigma^i,\sigma^j\}$. Therefore, we have a fluctuation-dissipation relation

$$D^{ij}(\tilde{v})=\{\sigma^i,\sigma^j\} \tag{3.19}$$

but this is not a complete characterization. In reality, for a given $\tilde{v}$, there can exist several inequivalent choices for H and F such that $\tilde{v}^i=l(x^i)$ with $l=X_H+\frac{1}{2}X_F\circ X_F$.

(6) The dissipation of $l=X_H+\frac{1}{2}X_F\circ X_F$ is given explicitly as

$$D(l;f,g)=\{\{f,F\},\{g,F\}\}. \tag{3.20}$$

Note that $D(l;f^*,f)=|\{f,F\}|^2$ is then positive.

IV. SYSTEMS WITH ONE DEGREE OF FREEDOM

Let (q,p) be local canonical coordinates on a symplectic manifold (Γ,ω) and consider the fields $\tilde{v}=\tilde{v}_q\partial_q+\tilde{v}_p\partial_p$, $\sigma=\sigma_q\partial_q+\sigma_p\partial_p$. The dissipation tensor has only $\gamma:=D^{qp}(\tilde{v})=-D^{pq}(\tilde{v})$ as entries which are possibly nonzero on account of antisymmetry. In terms of the field $\tilde{v}$ we have

$$\gamma=-\tilde{v}_{q,q}-\tilde{v}_{p,p}, \tag{4.1}$$

that is, minus the divergence of $\tilde{v}$ on Γ.

Our requirement is that $d\{q_t,p_t\}=\{dq_t,p_t\}+\{q_t,dp_t\}+\{dq_t,dp_t\}\equiv 0$, or equivalently

$$(\{\sigma_q,\sigma_p\}-\gamma)dt+(\sigma_{q,q}+\sigma_{p,p})dB_t=0. \tag{4.2}$$

Therefore, in the case that Γ is just two-dimensional, the conditions that the Poisson structure is preserved is that, if M is the 2×2-matrix with entries $\sigma_{\alpha,\beta}$, then

$$\text{(i)}\det M=\gamma, \quad \text{(ii)}\ \operatorname{tr} M=0. \tag{4.3}$$

Equivalently, by the Cayley–Hamilton theorem, the condition is that $M^2=-\gamma I_2$.

The noise coefficients σ are derivable from a Hamiltonian F; $\sigma_q \equiv F_{,p}$, $\sigma_p \equiv -F_{,q}$. The fluctuation-dissipation relation (3.19) then reduces to

$$\gamma(q,p) = -(F_{,qp})^2 + F_{,qq}F_{,pp}. \tag{4.4}$$

Likewise the Stratonovich fields v are given by (3.9),

$$v_q = \tilde{v}_q - \tfrac{1}{2}F_{,qp}F_{,p} + \tfrac{1}{2}F_{,pp}F_{,q} \equiv H_{,p}; \quad v_p = \tilde{v}_p + \tfrac{1}{2}F_{,qq}F_{,p} - \tfrac{1}{2}F_{,qp}F_{,q} \equiv -H_{,q}. \tag{4.5}$$

V. EXAMPLES OF QUANTUM SYMPLECTIC FLOWS

A. Linear damping

Starting from the equation $\ddot{q} + \xi\dot{q} + V'(q) = 0$, $(\xi > 0)$, which is linearly damped we see that

$$\dot{q} = \tilde{v}_q \equiv p; \quad \dot{p} = \tilde{v}_p \equiv -\xi p - V'(q). \tag{5.1}$$

The dissipation is $\gamma = \xi$ and it suffices to take the noise coefficients σ_α linear in the phase coordinates

$$\sigma_\alpha \equiv M_{\alpha\beta}x_\beta. \tag{5.2}$$

Here the matrix M is the same as above except now that its entries are constant. The Stratonovich version of the fields are given by $v_\alpha = \tilde{v}_\alpha - \frac{1}{2}M_{\alpha\beta}M_{\beta\mu}x_\mu$, however, since we require the stochastic flow to be canonical we have $M^2 = -\gamma I_2$ and so

$$v_\alpha = \tilde{v}_\alpha + \tfrac{1}{2}\xi x_\alpha. \tag{5.3}$$

The corresponding functions H and F introduced in Sec. III are then given by

$$H = \tfrac{1}{2}p^2 + \tfrac{1}{2}\xi qp + V(q); \quad F = \tfrac{1}{2}x^T JMx; \tag{5.4}$$

where $x = \binom{q}{p}$ and $J := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. As possible choice we may take $M = \begin{pmatrix} 0 & \xi^{-1/2} \\ -\xi^{3/2} & 0 \end{pmatrix}$, in which case $F \equiv \frac{1}{2}\xi^{-1/2}(p^2 + \xi^2 q^2)$. Note that we have chosen the units of mass to be unity; it therefore follows that ξ has dimensions s^{-1} and F has dimensions $m^2 s^{-3/2}$.

The canonical transformation $(q,p) \mapsto (Q,P)$ given by $Q := q$, $P := p + \frac{1}{2}\xi q$ leads to

$$H = \tfrac{1}{2}P^2 + V(Q) - \tfrac{1}{8}\xi^2 Q^2 \tag{5.5}$$

and

$$F = \tfrac{1}{2}\xi^{-1/2}(P^2 - \xi PQ + \tfrac{5}{4}\xi^2 Q^2). \tag{5.6}$$

B. Nonlinear damping

We consider as an example the equation $\ddot{q} + \xi(q^2 - a^2)\dot{q} + V'(q)$, $(\xi > 0, a > 0)$, which has the same damping mechanism as the van der Pol oscillator.[8] In terms of phase space coordinates we have

$$\dot{q} = \tilde{v}_q \equiv p; \quad \dot{p} = \tilde{v}_p \equiv -\xi(q^2 - a^2)p - V'(q). \tag{5.7}$$

The dissipation is

$$\gamma = -\tilde{v}_{q,q} - \tilde{v}_{p,p} = \xi(q^2 - a^2), \tag{5.8}$$

which is positive for $|q| > a$ and negative for $|q| < a$ and moreover independent of p. We shall look for a candidate for F of the form $\frac{1}{2}cp^2 + \phi(q)$. From dimensional arguments, here ξ has dimensions $m^{-2}s^{-1}$, we find that a solution to (5.4) is given by

$$F=\frac{1}{a}\xi^{-1/2}\left[\frac{1}{2}p^2+a^2\xi^2\left(\frac{q^4}{12}-\frac{q^2a^2}{2}\right)\right]. \tag{5.9}$$

The v fields are from (5.5),

$$v_q=p+\frac{1}{2}\xi\left(\frac{q^2}{3}-qa^2\right),\quad v_p=-\frac{1}{2}\xi(q^2-a^2)p-V'(q); \tag{5.10}$$

which are derivable from the Hamiltonian

$$H=\frac{1}{2}p^2+V(q)+\frac{1}{2}\xi\left(\frac{q^3}{3}-qa^2\right)p. \tag{5.11}$$

As in the linear case, we can introduce a canonical transformation, this time $(q,p)\mapsto(Q,P)$ where $Q:=q,\ P:=p+\frac{1}{2}\xi[(q^3/3)-qa^2]$, and so we have

$$H=\frac{1}{2}P^2+V(Q)-\frac{1}{8}\xi^2\left(\frac{Q^3}{3}-Qa^2\right)^2 \tag{5.12}$$

and

$$F=\frac{1}{a}\xi^{-1/2}\left[\frac{1}{2}P^2-\frac{1}{2}\xi P\left(\frac{Q^3}{3}-Qa^2\right)+\xi^2\left(\frac{Q^6}{72}-\frac{3Q^2a^4}{8}\right)\right]. \tag{5.13}$$

VI. QUANTUM STOCHASTIC SYMPLECTIC EVOLUTIONS

Let $\mathcal{A}$ be a C^*-algebra. The extension of a map l on $\mathcal{A}$ to the matrix algebra $M_n(\mathcal{A})$ is denoted by $l^{(n)}$ and l is called positive if it and its extensions are positive. For commutative algebras, positivity automatically implies complete positivity. Following Lindblad,[9] a bounded $*$-map l is called completely dissipative if $\mathcal{D}(l^{(n)};a^*,a)\geqslant 0$ for all n and $a\in M_n(\mathcal{A})$. We remark that Lindblad refers to our $\mathcal{D}$ rather than D as the dissipation, however, for quantum algebras the Poisson structure is given by

$$\wp(a,b):=\frac{1}{i\hbar}[a,b] \tag{6.1}$$

and so

$$D(l;a,b)\equiv\frac{1}{i\hbar}\{\mathcal{D}(l;a,b)-\mathcal{D}(l;b,a)\}. \tag{6.2}$$

The distension $\mathcal{D}(l)$ determines l up to a Hamiltonian map $X_H\equiv(1/i\hbar)[.,H]$. It was proved by Gorini, Kossakowski, and Sudarshan[10] (for finite dimensional algebras) and by Lindblad[9] (for the more general case of hyperfinite factors) that, for $\Phi_t=e^{tl}$ a norm-continuous, completely positive identity preserving semigroup on $\mathcal{A}$, its generator l must be completely dissipative and that the space of such generators is convex with extremal (pure) elements of the type

$$l(a)=\frac{1}{i\hbar}[a,H]+\frac{1}{\hbar^2}\left(F^*aF-\frac{1}{2}aF^*F-\frac{1}{2}F^*Fa\right)\equiv X_H(a)+\frac{1}{2}(F^\dagger X_F(a)-X_{F^\dagger}(a)F), \tag{6.3}$$

for H (self-adjoint) and F,F^* in $\mathcal{A}$.

The distension for the pure Lindblad generator above is then

$$\mathcal{D}(l;a,b)=\frac{1}{\hbar^2}[F^*,a][b,F]=(X_{F^*})(X_F b), \tag{6.4}$$

and so $\mathcal{D}(l,a^*,a)=(1/\hbar^2)[F,a]^*[F,a]\geqslant 0$. The dissipation (in the sense of our definition 3) is then

$$D(l;a,b)=\frac{1}{i\hbar}\{X_{F^*}aX_Fb-X_{F^*}bX_Fa\}. \tag{6.5}$$

Notice that taking F to be also self-adjoint leads to $l=X_H+\frac{1}{2}X_F\circ X_F$ which is formally the same as what we have obtained for classical canonical flows.

Quantum stochastic flows

We now describe how quantum stochastic calculus can be used to construct quantum stochastic flows j_t such that $\Phi_t=E_{0]}\circ j_t$ has Lindblad generator.

In the following $\Gamma_B(h)$ shall denote the Bose Fock space over a Hilbert space h and $A^\sharp(f)$ the creation and annihilation operators with test function $f\in h$. Let $\mathcal{H}_0$ be the Hilbert space of states for a quantum mechanical system. The system is to be considered as open and its evolution operator V_t satisfying the following QSDE on $\mathcal{H}_0\otimes\Gamma_B(L^2[0,\infty))$ driven by the Hudson–Parthasarathy quantum Wiener processes $A_t^\dagger=A^\sharp(\chi_{[0,t]})$,

$$d_IV_t=\frac{1}{i\hbar}\left(F\otimes d_IA_t^\dagger+F^*\otimes d_IA_t+\left(H-\frac{1}{2\hbar}F^*F\right)\otimes dt\right)V_t, \tag{6.6}$$

with initial condition $V_0=1$. Now, if K were also self-adjoint then we could combine the noise terms into $K\otimes(d_IA_t^\dagger+d_IA_t)$ in which case it would be much simpler to deal with just the standard Wiener process, however as this is an exceptional case, we are quite generally forced to use the quantum stochastic processes instead. It should be remarked that there are alternative representations, for instance a single quantum Wiener process can be described by two standard ones; however the choice of quantum stochastic processes is the most natural.

The quantum stochastic flow is given on the set of observables $\mathcal{A}=\mathcal{B}(\mathcal{H})$ by the family of maps $(j_t)_{t\geqslant 0}$ defined by

$$j_tX:=V_t^\dagger(X\otimes 1)V_t,\quad \forall X\in\mathcal{A}. \tag{6.7}$$

The differential form of the noisy equations of motion is the given by the quantum Langevin equation

$$d_Ij_tX=j_t\left(\frac{1}{i\hbar}[X,F]\right)\otimes d_IA_t^\dagger+j_t\left(\frac{1}{i\hbar}[X,F^*]\right)\otimes d_IA_t+j_t(l(X))\otimes dt, \tag{6.8}$$

where l is the generator (6.3). The quantum stochastic process V_t is unitary on the combined system-noise Hilbert space. Moreover it respects the canonical commutation structure; to see this let $d_IL:=j_t^{-1}d_Ij_t=A_{dt}^\dagger X_F+A_{dt}X_{F^*}+dtl$ then we must show that, for arbitrary $X,Y\in\mathcal{A}$, that

$$d_IL[X,Y]:=[d_ILX,Y]+[X,d_ILY]+[d_ILX,d_ILY]. \tag{6.9}$$

To this end, we can use the Jacobi identity for commutators at several places. We also have to calculate the Ito shift and here we use the fact that $A_{dt}A_{dt}^\dagger=dt$ with all other products of differentials equal to zero,[3]

$$[d_ILa,d_ILb]=(X_{F^*}aX_Fb-X_{F^*}bX_Fa)dt=i\hbar D(l;a,b)dt. \tag{6.10}$$

Here we see explicitly the quantum fluctuation-dissipation relations.

VII. QUANTIZATION OF DISSIPATIVE EVOLUTIONS

In Sec. VI we introduced a scheme for finding the classical Hamiltonians H and F, while in Sec. V we did this for the case of linear and nonlinear damping. Both these models are relatively easy to quantize. We use the Weyl quantization procedure to produce self-adjoint operators H and F which we can insert into the QSDE (6.6). For the situations in Sec. V, we end up with the equations $E_{t]}d_I q_t = \tilde{v}_q(q_t, p_t)$ and $E_{t]}d_I p_t = \tilde{v}_p(q_t, p_t)$ where now $\tilde{v}_i$ are the symmetrically-ordered versions of the originals.

The operator ordering ambiguities are not apparent here yet because the special form of the Hamiltonians encountered in Sec. V. However, this will in general be a complicating feature which will in general mean that F will have to have a modified form from the Weyl quantization of its classical counterpart. We also mention that canonical transformations will not generally respect given orderings.

VIII. CONCLUSIONS

At the heart of our presentation is the notion that a physical flow, whether classical stochastic, quantum mechanical, or quantum stochastic, should be canonical in some sense. The QSDE (6.1) is of the most general type defining a unitary process driven by the quantum stochastic process $A_t^\dagger$ and A_t.[3] We now wish to make several remarks.

(1) The Ornstein–Uhlenbeck process, and consequently any approach which seeks to generalize it, is not symplectic. Here the Langevin equations are $\dot{q} = p$, $dp = b_p dt = \sigma_p dB_t$; that is $\sigma_q \equiv 0$. The matrix M introduced in Sec. V therefore has first row equal to zero and so condition (i) of (5.3) cannot be satisfied.

(2) In the quantized models in Sec. V, we meet a Hamiltonian $\tilde{H} := \frac{1}{2}p^2 + V(q)$ however we eventually deal with the Hamiltonian H and we have from (5.5) and (5.12) that $H = \tilde{H} - U(q)$, where $U(q) = \frac{1}{8}\xi^2 q^2$ in the linear case and $U(q) = \frac{1}{8}\xi^2[(q^3/3) - q]^2$ in the nonlinear case. We therefore have three Hamiltonians, F,H, and $\tilde{H}$, at our disposal. From our point of view it is H which is the physical Hamiltonian and not $\tilde{H}$! The discrepancy between $\tilde{H}$ and H is, as already in the classical theory, a Stratonovich–Ito dilemma. Therefore, although we would agree with the approach of Sinha[11] to construct a symplectic quantum stochastic flow for the linearly damped model, we disagree with the arguments introduced to justify the use of a QSDE driven by two independent quantum Wiener processes wherein the second serves to account for the convective drift, see (3.9). There the argument was that, if $\tilde{H}$ had bounded-below spectrum, then this property would not generally follow for $\tilde{H} - U$.

(3) Returning to our program to deduce the stochastic flow from a microscopic Hamiltonian flow, we make the following observation. Suppose we obtain ODEs describing the evolution of the system observables driven by a stochastic process relating to the reservoir. In general, this process, though noisy, is not white. However, taking τ to measure the autocorrelation time for the reservoir processes we may expect to obtain SDEs of the type (3.7) in the limit τ goes to zero. If indeed this is the case and the reservoir forcing term does converge to white noise, then the result of Wong and Zakai[12] tells us that the limit SDEs will have the identical form as the pre-limit ODEs provided we take the Stratonovich interpretation. Therefore, it is the Hamiltonians H and F which are physically relevant. Recall the stochastic derivation (3.18) which identified $H + FW_S(t)$ as the effective physical Hamiltonian. The Hamiltonian $\tilde{H}$ is therefore to be viewed as derived from H and not vice versa.

(4) In Refs. 13, 14 we have introduced a treatment of Stratonovich calculus in a noncommutative setting and discussed the quantum version of the Wong–Zakai theorem. The Stratonovich QSDE corresponding to (6.6) is of the form

$$d_S V_t = \frac{1}{i\hbar}(F a_t^\dagger + F^\dagger a_t + H dt)V_t. \tag{8.1}$$

The $a_t^\sharp$ are formal operators with action on exponential states $\Psi(f)$ of $\Gamma_B(L^2[0,\infty))$ given by

$$a_t\Psi(f) := \tfrac{1}{2}(f(t^+) + f(t^-))\Psi(f). \tag{8.2}$$

However it should be noted that the integrators $a_t^\sharp dt$ do not commute with adapted integrands, contrary to the Ito situation, and the QSDEs can only be compared when (8.2) is placed into normal-ordered form. The passage from Stratonovich to Ito (Hudson–Parthasarathy) form leads, however, to (8.1) being equivalent to (6.6). This can be extended to include the so-called conservation process also.

Moreover, it is reasonable to expect (8.2) as the weak coupling limit of a physical system-reservoir dynamics, typically the emission-absorption-type interaction, as detailed in Ref. 15.

ACKNOWLEDGMENTS

The work presented here was carried out while the author was supported by the EU Presidency Fellowship (Ireland). In particular, the author would like to thank Professor D. Heffernan (Maynooth) for many stimulating discussions and comments during the writing of this paper. Also several important comments and suggestions made by the referee are gratefully acknowledged.

[1] R. Alicki and K. Lendi, *Quantum Dynamical Semigroups and Applications, Lecture Notes in Physics 286* (Springer, Berlin, 1987).
[2] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North–Holland Personal Library, Amsterdam, 1992).
[3] R. L. Hudson and K. R. Parthasarathy, ''Quantum lto's formula and stochastic evolutions,'' Commun. Math. Phys. **93**, 301–323 (1984).
[4] L. Accardi and R. L. Hudson, ''The Non-Abelian Cohomology of Stochastic Derivations,'' in *Quantum Probability and Applications V, Lecture Notes in Mathematics 1442* (Springer, Berlin, 1990).
[5] K. D. Elworthy, *Stochastic Differential Equations on Manifolds* (Cambridge University Press, Cambridge, 1982).
[6] The Poisson-simplicity for symplectic manifolds is well-known. For the angular momentum algebra (it is also true) however, it is possible to contruct local flows which are canonical but non-Hamiltonian. Such flows are in fact related to Dirac monopoles and have line-type singularities known as ''Dirac strings.''
[7] P. A. Meyer, ''A Differential Geometric Formalism for the Ito Calculus,'' *LNM 851, Stochastic Integrals*, edited by D. Williams (Springer, Berlin, 1980).
[8] D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations, Oxford Applied Mathematics and Computing Science Series* (Clarendon, Oxford, 1977).
[9] G. Lindblad, ''On the generators of quantum dynamical semigroups,'' Commun. Math. Phys. **48**, 119–130 (1976).
[10] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan, ''Completely positive dynamical semigroups of *n*-level systems,'' J. Math. Phys. **17**, 821–825 (1976).
[11] K. B. Sinha, ''Quantum mechanics of dissipative systems,'' J. Ind. Inst. Sci. **77**, 275–279 (1997).
[12] E. Wong and M. Zakai, ''Riemann–Stieltjes approximations of stochastic integrals,'' Z. Wahrscheinlichkeitstheor. Verwandte Geb. **12**, 87–97 (1969).
[13] J. Gough, ''The causal structure of quantum stochastic integrators,'' Theor. Math. Phys. **111**, 218–233 (1997).
[14] J. Gough, ''Noncommutative Ito and Stratonovich noise and stochastic evolutions,'' Theor. Math. Phys. **113**, 276–284 (1997).
[15] J. Gough, ''The Stratonovich interpretation for quantum stochastic approximations,'' J. Potential Analysis (to be published).

A new first integral for a binary rigid body collision of arbitrarily short duration

Patrick L. Nash[a]
Division of Earth and Physical Sciences, The University of Texas at San Antonio, San Antonio, Texas 78249-0663

(Received 21 September 1998; accepted for publication 22 February 1999)

A standard classical model of a so-called rigid two-body collision that employs the dynamic Coulomb friction law to model friction is studied. For arbitrary object geometries and initial conditions it is known that the direction of the relative sliding velocity continuously changes during the impact. A (new) exact analytical solution for the relative sliding speed u_{tr} of the two objects in terms of initial conditions and sliding direction is derived. This solution is formulated in terms of a first integral, which is used to rigorously prove that the dynamic Coulomb friction law does not allow either instantaneous sticking or stable sticking to evolve from an initially nonzero u_{tr}, except for certain very special cases. The first integral also yields a new procedure for accurately and efficiently computing the final center of mass velocity and the final angular velocity of each of the objects in the model two-body collision. Accurate solutions such as these are essential for analyzing and controlling impacts, which is important, for example, in robot manipulation. Efficient solutions are critically important for producing real-time simulations of rigid two-body collisions. © *1999 American Institute of Physics.* [S0022-2488(99)01206-2]

I. INTRODUCTION

Well over a century ago Routh[1,2] laid out a formalism for analyzing a rigid two-body collision of arbitrarily short duration. Routh showed that during an impact the point of contact must, in general, slide. He argued that if the friction forces are strong enough then sliding results in sticking, possibly followed by rolling. Otherwise, the contact point slides continuously during the collision, possibly reversing direction during the course of the impact. Long after Routh's work the problems of: (i) formulating a fundamental model describing the basic classical collision process; (ii) increasing the accuracy and efficiency of the numerical solution to a particular model; and (iii) extending the known analytical results; continue to receive a great deal of attention.[3–17] With regard to the analytical solution to the general problem, Routh suggested that there was, in fact, no closed form solution, and currently there appears to be a consensus that there may be no nonperturbative solution to this scattering problem.[12] However much progress has been made. Recently in remarkable papers, Stronge[4,6] has shown how to generalize both Newton's Impact Law and that of Poisson by giving a new definition of the coefficient of restitution that is consistent with conservation of energy (previous impact laws did not always conserve energy). In addition, Bhatt and Koechling,[12] among others, have clearly shown for arbitrary object shapes and initial conditions, that the direction of the relative sliding velocity continuously changes throughout the duration of sliding, and have offered a classification of the possible motions.

In this paper we consider a small part of the unsolved problem, that of analytically calculating the relationship between the transverse relative speed (sliding speed) and the direction of the transverse relative velocity (sliding velocity) during the impact of two rigid bodies that suffer an arbitrarily short collision. *Arbitrarily short* so that during the collision we may neglect all external forces other than the impulsive forces of impact that the two colliding objects exert on one

[a]Electronic mail: nsh@susan.ep.utsa.edu

another, and so that the collision matrix [defined below in Eq. (6)] has constant matrix elements. In order to produce exact analytical results the dynamic Coulomb friction law is assumed to describe friction during the scattering. This model asserts that the force $\mathbf{f}=\mathbf{f}_{\text{normal}}+\mathbf{f}_{\perp}$ that one body exerts on the other has a transverse component $\mathbf{f}_{\perp}$ that is directed opposite to the transverse component of the relative velocity, and has a magnitude given by μf_{normal}, where μ is the coefficient of kinetic friction. We derive an exact analytical expression for the transverse relative speed of the two objects in terms of initial conditions and sliding direction, which is formulated as a first integral [see Eq. (26)]. Using it we prove a somewhat surprising result, namely, that the dynamic Coulomb friction law, strictly imposed, does not allow either instantaneous sticking or stable sticking to evolve from an initially nonzero u_{tr}, except possibly for certain special cases. We also show that this does not contradict the conventional view of possible (stable or unstable) sticking during a collision.

The dynamic Coulomb friction law is assumed to model sliding with friction because it (i) is widely used in computations; (ii) is a valid starting point for more detailed investigation; and (iii) admits an analytical solution. On a macroscopic level the dynamic Coulomb friction law may be justified by the Bowden–Tabor adhesion model, which is based on the elastic and plastic properties of the colliding objects.[8,7] However, Coulomb's laws of friction often oversimplify very complex phenomena involving elastomechanical, plastic, and even chemical interactions that usually operate on very different scales of time and length.[15] In fact, deviations from the dynamic Coulomb friction law are often encountered in practice.[8,10,11] For example, the force of kinetic friction is known to depend on the sliding speed. Heslot *et al.*[10] have demonstrated that under some circumstances the force of dry friction first decreases, passes through a minimum, and then finally increases. Bhushan *et al.*,[11] working with a thin lubrication layer a few monolayers thick, found that the force of kinetic friction first increases, goes through a maximum, and then finally decreases. Notwithstanding these observations, since the dynamic Coulomb friction law is widely used as a starting point in modeling friction, we believe that there is value in knowing exactly what the dynamic Coulomb friction law predicts when applied to a rigid two-body collision. The possibility of performing reproducible friction experiments on mesoscopic and nanoscopic scales increases the interest in comparing the predictions of an exact (albeit incomplete) theory with the results of experiment.[7,15]

An outline of the content of this paper is as follows. We first develop our notation by briefly reviewing the basic theory; although the basic theory is very well known, the notation used to formulate this theory has not been standardized, and we would like to reduce possible ambiguities. Next, the equations for the relative velocity of the contact point of the impact are reformulated, followed by an examination of the projection of the motion into the natural tangent plane (defined below). This leads, after a series of steps, to the statement of the new first integral (conservation law). Finally we apply these results to specific examples.

II. OVERVIEW AND KINEMATICS

We shall assume that the impact begins at $t_{\text{initial}}=0$ and contact terminates at t_{final}, assuming that the objects do not stick together. When two physical bodies collide they always experience deformation over time comprised generally of compression and restoration phases. Conceptually therefore, this two-body collision is resolved into two periods: a compression phase of quasielastic deformation of each of the bodies that begins with first contact at t_{initial} and terminates at the time of maximum compression $t_{\text{max compression}}$, and a restitution phase that begins at $t_{\text{max compression}}$ and ends at time t_{final}.

The standard model approximates this collision in the zeroth order as being an instantaneous impact between *rigid* bodies with the relative velocity being discontinuous with respect to time, *but not with respect to other parameters such as the normal component of the total impulse*. We let J_z denote the normal component of the total impulse. The zeroth-order approximation invokes a phenomenological law such as Newton's Impact Law or Poisson's Hypothesis[3] to relate initial and final normal components of the relative velocity to one another, and employs integrations with respect to J_z to calculate the transverse component discontinuity (with respect to time, not J_z) of the relative velocity. This approach bears a definite resemblance to a ''Born approximation'' for the scattering in which we approximate the contact point of the impact as fixed with respect to time in the zeroth order, and the moments of inertia of the two bodies as constant during the collision.

Position: The two colliding bodies are labeled by $\alpha=1,2$. Let us mentally divide each object up into a large number of macroscopically small but microscopically large volume elements. Let N_α = number of volume elements of the αth body. $\mathbf{r}_{(\alpha)i}$ locates the ith volume element of the αth body and $m_{(\alpha)i}$ denotes its mass, where $i=0,\ldots,N_\alpha-1$. $\mathbf{r}_{(1)0}$ and $\mathbf{r}_{(2)0}$ locate the (fixed, in the zeroth-order approximation) point of impact of the two bodies. The z axis is perpendicular to the plane of contact, directed from body 2 to body 1. The $x-y$ plane is tangent to the surface of each object at the point of contact. Transverse (i.e., tangential) components of vectors lie in the $x-y$ plane.

The center of mass and position vector are defined as usual: let m_α denote the mass of the αth body, $m_\alpha=\Sigma_{i=0}^{N_\alpha-1} m_{(\alpha)i}$. The center of mass $\mathbf{R}_\alpha$ is defined through $m_\alpha\mathbf{R}_\alpha=\Sigma_{i=0}^{N_\alpha-1} m_{(\alpha)i}\mathbf{r}_{(\alpha)i}$. The relative position vector $\delta\mathbf{r}_{(\alpha)i}$ of $m_{(\alpha)i}$ with respect to $\mathbf{R}_\alpha$ is $\mathbf{r}_{(\alpha)i}=\mathbf{R}_\alpha+\delta\mathbf{r}_{(\alpha)i}$, and we find that the usual constraint identity holds:

$$m_\alpha\mathbf{R}_\alpha=\sum_{i=0}^{N_\alpha-1} m_{(\alpha)i}\mathbf{r}_{(\alpha)i}\Rightarrow 0=\sum_{i=0}^{N_\alpha-1} m_{(\alpha)i}\delta\mathbf{r}_{(\alpha)i}. \tag{1}$$

Velocity: Let $\dot{\mathbf{R}}_\alpha$ denote the velocity of the center of mass of the αth object, where the dot denotes differentiation with respect to time. A rigid body rotation about the center of mass is specified according to $\delta\dot{\mathbf{r}}_{(\alpha)i}=\omega_\alpha\times\delta\mathbf{r}_{(\alpha)i}$, where ω_α is the angular velocity of the αth body. Define the three standard 3×3 antisymmetric matrices $\mathbf{S}=(S_{(1)},S_{(2)},S_{(3)})=(S_x,S_y,S_z)$ as follows: let $\mathbf{A}$ and $\mathbf{B}$ be vectors, and define $\mathbf{S}$ by $(\mathbf{A}\cdot\mathbf{S})B=\mathbf{A}\times\mathbf{B}$. In terms of matrix elements, $S_{(j)ik}=\epsilon_{ijk}$, where the Levi-Civita tensor, $\epsilon_{123}=+1$.

Let $\mathbf{S}_{(\alpha)i}=\delta\mathbf{r}_{(\alpha)i}\cdot\mathbf{S}$. Since $\delta\dot{\mathbf{r}}_{(\alpha)i}=\omega_\alpha\times\delta\mathbf{r}_{(\alpha)i}=-\delta\mathbf{r}_{(\alpha)i}\times\omega_\alpha=-\mathbf{S}_{(\alpha)i}\omega_\alpha$, the velocity of the ith volume element of the αth body is $\dot{\mathbf{r}}_{(\alpha)i}=\dot{\mathbf{R}}_\alpha+\delta\dot{\mathbf{r}}_{(\alpha)i}=\dot{\mathbf{R}}_\alpha-\mathbf{S}_{(\alpha)i}\omega_\alpha$.

For future reference we note that

$$\mathbf{P}_{(\alpha)i}\equiv-\mathbf{S}_{(\alpha)i}\mathbf{S}_{(\alpha)i}=\mathbf{I}_{3\times 3}|\delta\mathbf{r}_{(\alpha)i}|^2-\delta\mathbf{r}_{(\alpha)i}\otimes\delta\mathbf{r}_{(\alpha)i}, \tag{2}$$

where $\mathbf{I}_{3\times 3}$ is the 3×3 unit matrix, defines a projection operator $\mathbf{P}_{(\alpha)i}/|\delta\mathbf{r}_{(\alpha)i}|^2$ onto the plane perpendicular to $\delta\mathbf{r}_{(\alpha)i}$.

Let $\mathbf{u}=\dot{\mathbf{r}}_{(1)0}-\dot{\mathbf{r}}_{(2)0}$ be the relative velocity of the ''fixed'' point of impact. We let $\mathbf{u}_\perp=u_x\hat{\mathbf{x}}+u_y\hat{\mathbf{y}}$ denote the transverse component of the relative velocity, and $u_{\text{transverse}}=u_{\text{tr}}$ denote the magnitude of the transverse (tangential) component of the relative velocity. In this coordinate system $\mathbf{u}=u_z\hat{\mathbf{z}}+u_x\hat{\mathbf{x}}+u_y\hat{\mathbf{y}}=u_z\hat{\mathbf{z}}+u_\perp=u_z\hat{\mathbf{z}}+u_{\text{tr}}[\hat{\mathbf{x}}\cos(\theta)+\hat{\mathbf{y}}\sin(\theta)]$, where $u_x=u_{\text{tr}}\cos(\theta)$ and $u_y=u_{\text{tr}}\sin(\theta)$.

The z component of the relative velocity is subject to several conditions: $u_z(t_{\text{initial}})<0$; $u_z(t_{\text{max compression}})=0$; and $u_z(t_{\text{final}})\geqslant 0$.

Angular momentum: The angular momentum of the ith volume element of the αth body is $\mathbf{L}_{(\alpha)i}=m_{(\alpha)i}\mathbf{r}_{(\alpha)i}\times\dot{\mathbf{r}}_{(\alpha)i}=m_{(\alpha)i}(\mathbf{R}_\alpha+\delta\mathbf{r}_{(\alpha)i})\times(\dot{\mathbf{R}}_\alpha+\delta\dot{\mathbf{r}}_{(\alpha)i})$. We all know that the angular momentum of the αth body decomposes into two pieces corresponding to the center of mass motion and rotation about an axis through the center of mass: $\mathbf{L}_\alpha=\Sigma_{i=0}^{N_\alpha-1}\mathbf{L}_{(\alpha)i}=\Sigma_{i=0}^{N_\alpha-1} m_{(\alpha)i}(\mathbf{R}_\alpha+\delta\mathbf{r}_{(\alpha)i})\times(\dot{R}_\alpha+\delta\dot{\mathbf{r}}_{(\alpha)i})=m_\alpha\mathbf{R}_\alpha\times\dot{\mathbf{R}}_\alpha+\Sigma_{i=0}^{N_\alpha-1} m_{(\alpha)i}\delta\mathbf{r}_{(\alpha)i}\times\delta\dot{\mathbf{r}}_{(\alpha)i}=m_\alpha\mathbf{R}_\alpha\times\dot{\mathbf{R}}_\alpha+\Sigma_{i=0}^{N_\alpha-1} m_{(\alpha)i}\mathbf{S}_{(\alpha)i}(-\mathbf{S}_{(\alpha)i}\omega_\alpha)=m_\alpha\mathbf{R}_\alpha\times\dot{\mathbf{R}}_\alpha+\Sigma_{i=0}^{N_\alpha-1} m_{(\alpha)i}\mathbf{P}_{(\alpha)i}\omega_\alpha=m_\alpha\mathbf{R}_\alpha\times\dot{\mathbf{R}}_\alpha+\mathbf{I}_\alpha\omega_\alpha$, where we have used Eq. (1) twice. Here $\mathbf{I}_\alpha$ is the inertia tensor of the αth body defined by $\mathbf{I}_\alpha=\Sigma_{i=0}^{N_\alpha-1} m_{(\alpha)i}\mathbf{P}_{(\alpha)i}$.

III. EQUATIONS OF MOTION

Let $\mathbf{f}_{(\alpha)i}^{\text{external}}$ be the external force on the ith volume element of the αth body. In the standard approach it is assumed that during an arbitrarily short collision we may neglect all external forces other than the impulsive forces of impact that the two colliding objects exert on one another. Let $\mathbf{f}_{(1)0}^{\text{external}}=-\mathbf{f}_{(2)0}^{\text{external}}\equiv\mathbf{f}$ denote these forces.

Newton: The net force on the ith volume element of the αth body is $m_{(\alpha)i}\ddot{\mathbf{r}}_{(\alpha)i}=\Sigma_{j=0,j\neq i}^{N_\alpha-1}\mathbf{f}_{(\alpha)j\to i}+\mathbf{f}_{(\alpha)i}^{\text{external}}$, where $\mathbf{f}_{(\alpha)j\to i}$ is the internal force that the jth volume element exerts on the ith volume element of the αth body. $\mathbf{f}_{(\alpha)j\to i}$ is assumed to obey Newton's Third Law. The net force on the αth body is $m_\alpha\ddot{\mathbf{R}}_\alpha=\Sigma_{i=0}^{N_\alpha-1}m_{(\alpha)i}\ddot{\mathbf{r}}_{(\alpha)i}=\Sigma_{i=0}^{N_\alpha-1}\mathbf{f}_{(\alpha)i}^{\text{external}}=\mathbf{f}_{(\alpha)}^{\text{external}}$. Moreover a well-known simple calculation yields $\dot{\mathbf{L}}_\alpha=\mathbf{R}_\alpha\times\mathbf{f}_{(\alpha)}^{\text{external}}+\Sigma_{i=0}^{N_\alpha-1}\mathbf{S}_{(\alpha)i}\mathbf{f}_{(\alpha)i}^{\text{external}}$, which implies that $(d/dt)\mathbf{I}_\alpha\omega_\alpha=\Sigma_{i=0}^{N_\alpha-1}\mathbf{S}_{(\alpha)i}\mathbf{f}_{(\alpha)i}^{\text{external}}=\mathbf{S}_{(\alpha)0}\mathbf{f}_{(\alpha)0}^{\text{external}}$.

A. Impulses and velocity changes

The external impulse $\mathbf{J}_{(\alpha)i}$ on the ith volume element of the αth body is defined as $\mathbf{J}_{(\alpha)i}=\int_{t_{\text{initial}}}^{t_{\text{final}}}\mathbf{f}_{(\alpha)i}^{\text{external}}\,dt$. Let $\mathbf{J}_{(1)0}=-\mathbf{J}_{(2)0}\equiv\mathbf{J}$ denote the nonzero external impulses. Upon integrating the equations of motion we find that $\Delta m_\alpha\dot{\mathbf{R}}_\alpha=\Sigma_{i=0}^{N_\alpha-1}\mathbf{J}_{(\alpha)i}=\mathbf{J}_{(\alpha)0}$ so that

$$\Delta\dot{\mathbf{R}}_\alpha=\frac{\mathbf{J}_{(\alpha)0}}{m_\alpha}. \tag{3}$$

In addition, since the objects are assumed not to move during the collision, $\Delta(\mathbf{I}_\alpha\omega_\alpha)=\mathbf{I}_\alpha\Delta\omega_\alpha=\Sigma_{i=0}^{N_\alpha-1}\int\mathbf{S}_{(\alpha)i}\mathbf{f}_{(\alpha)i}dt=\int\mathbf{S}_{(\alpha)0}\mathbf{f}_{(\alpha)0}\,dt=\mathbf{S}_{(\alpha)0}\int\mathbf{f}_{(\alpha)0}\,dt=\mathbf{S}_{(\alpha)0}\mathbf{J}_{(\alpha)0}$. Therefore

$$\Delta\omega_\alpha=\mathbf{I}_\alpha^{-1}\mathbf{S}_{(\alpha)0}\mathbf{J}_{(\alpha)0}. \tag{4}$$

The impulse $\mathbf{J}$ is not yet determined. An approximation for it may be obtained using a standard approach outlined in Sec. III B.

B. Calculation of J

One may generate an equation for the impulse $\mathbf{J}$ in terms of $\Delta\mathbf{u}$ as follows: The change in the velocity of the ith volume element of the αth body is $\Delta\dot{\mathbf{r}}_{(\alpha)i}=\Delta\dot{\mathbf{R}}_\alpha+\Delta\,\delta\dot{\mathbf{r}}_{(\alpha)i}=\Delta\dot{\mathbf{R}}_\alpha-\Delta\mathbf{S}_{(\alpha)i}\omega_\alpha=\Delta\dot{\mathbf{R}}_\alpha-\mathbf{S}_{(\alpha)i}\Delta\omega_\alpha=\mathbf{J}_{(\alpha)0}/m_\alpha-\mathbf{S}_{(\alpha)i}\mathbf{I}_\alpha^{-1}\mathbf{S}_{(\alpha)0}\mathbf{J}_{(\alpha)0}=(\mathbf{I}_{3\times3}/m_\alpha-\mathbf{S}_{(\alpha)i}\mathbf{I}_\alpha^{-1}\mathbf{S}_{(\alpha)0})\mathbf{J}_{(\alpha)0}$. Therefore $\Delta\mathbf{u}=\Delta\dot{\mathbf{r}}_{(1)0}-\Delta\dot{\mathbf{r}}_{(2)0}=(\mathbf{I}_{3\times3}/m_1-\mathbf{S}_{(1)0}\mathbf{I}_1^{-1}\mathbf{S}_{(1)0})\mathbf{J}_{(1)0}-(\mathbf{I}_{3\times3}/m_2-\mathbf{S}_{(2)0}\mathbf{I}_2^{-1}\mathbf{S}_{(2)0})\mathbf{J}_{(2)0}$ or

$$\Delta\mathbf{u}=\left[\mathbf{I}_{3\times3}\left(\frac{1}{m_1}+\frac{1}{m_2}\right)-(\mathbf{S}_{(1)0}\mathbf{I}_1^{-1}\mathbf{S}_{(1)0}+\mathbf{S}_{(2)0}\mathbf{I}_2^{-1}\mathbf{S}_{(2)0})\right]\mathbf{J}\equiv\mathbf{KJ}, \tag{5}$$

where the collision matrix $\mathbf{K}$ is defined by

$$\mathbf{K}=\mathbf{I}_{3\times3}\left(\frac{1}{m_1}+\frac{1}{m_2}\right)-(\mathbf{S}_{(1)0}\mathbf{I}_1^{-1}\mathbf{S}_{(1)0}+\mathbf{S}_{(2)0}\mathbf{I}_2^{-1}\mathbf{S}_{(2)0}). \tag{6}$$

It is well known[14] that $\mathbf{K}$ is a real, positive definite, symmetric nonsingular matrix.

It is convenient to decompose $\mathbf{K}$ into row vectors $\mathbf{k}_{(a)}$, $a,b,c=1,2,3=x,y,z$, according to $\mathbf{k}_{(a)b}=\mathbf{K}_{ab}$. In this notation

$$\mathbf{K}=\begin{pmatrix}\mathbf{k}_{(1)}\\ \mathbf{k}_{(2)}\\ \mathbf{k}_{(3)}\end{pmatrix},$$

the determinant of $\mathbf{K}$ is $\mathbf{k}_{(a)}\cdot\mathbf{k}_{(b)}\times\mathbf{k}_{(c)}$ where $(a,b,c)=$even permutation of (1,2,3), and $\mathbf{K}^{-1}=1/(\mathbf{k}_{(3)}\cdot\mathbf{k}_{(1)}\times\mathbf{k}_{(2)})(\mathbf{k}_{(2)}\times\mathbf{k}_{(3)},\mathbf{k}_{(3)}\times\mathbf{k}_{(1)},\mathbf{k}_{(1)}\times\mathbf{k}_{(2)})$.

Solving for $\mathbf{J}$ in Eq. (5) yields

$$\mathbf{J}=\mathbf{K}^{-1}\Delta\mathbf{u}. \tag{7}$$

$\Delta\dot{\mathbf{R}}_\alpha$ and $\Delta\omega_\alpha$ may then be calculated using Eqs. (3) and (4).

The evaluation of $\Delta\mathbf{u}$ employs, in part, energy methods that are dependent on a phenomenological model of restitution. Relevant energy definitions and rebound models are briefly discussed in the next subsections.

Energy: Let $dW_{(\alpha)i}=(\Sigma_{j=0,j\neq i}^{N_\alpha-1}\mathbf{f}_{(\alpha)j\to i}+\mathbf{f}_{(\alpha)i}^{\text{external}})\cdot d\mathbf{r}_{(\alpha)i}$ denote the infinitesimal work done on $m_{(\alpha)i}$ in a displacement $d\mathbf{r}_{(\alpha)i}$. The total infinitesimal work done on the system is $dW=\Sigma_{\alpha=1}^{2}\Sigma_{j=0}^{N_\alpha-1}dW_{(\alpha)i}$. Since $d\mathbf{r}_{(\alpha)i}$ is tangent to the trajectory, during a binary collision we approximate this as $dW=dW_{(1)0}+dW_{(2)0}\approx\mathbf{f}_{(1)0}^{\text{external}}\cdot d\mathbf{r}_{(1)0}+\mathbf{f}_{(2)0}^{\text{external}}\cdot d\mathbf{r}_{(2)0}=\mathbf{f}\cdot(d\mathbf{r}_{(1)0}/dt-d\mathbf{r}_{(2)0}/dt)dt=\mathbf{f}\cdot\mathbf{u}dt=\mathbf{u}\cdot d\mathbf{J}$.

Impact law: A practical problem encountered in integrating the equations of motion is determining when to terminate the restitution phase of the collision. Newton, Poisson, and others have postulated different impact laws that govern this judgment. We shall adopt the hypothesis put forward by Stronge.[4,6] Stronge has given a new definition of the coefficient of restitution that is consistent with conservation of energy (previous impact laws did not always conserve energy).

Stronge decomposes dW into three scalar contributions, $dW=dW_x+dW_y+dW_z$, with $dW_z=f_zu_zdt=u_zdJ_z$ and so on. Stronge postulates that

$$W_z(t_{\text{final}})-W_z(t_{\text{max compression}})=-e^2W_z(t_{\text{max compression}}), \tag{8}$$

where e is the coefficient of restitution and $W_z(t_{\text{initial}})=0$.

IV. GOVERNING DIFFERENTIAL FORMS

Consider $d\Delta\mathbf{u}=d(\mathbf{u}(t_{\text{final}})-\mathbf{u}(t_{\text{initial}}))=d\mathbf{u}(t_{\text{final}})\equiv d\mathbf{u}$. Here $\mathbf{u}=u_z\hat{\mathbf{z}}+u_x\hat{\mathbf{x}}+u_y\hat{\mathbf{y}}$, and $d\mathbf{u}=(\hat{\mathbf{z}}(\partial u_z/\partial\lambda)+\hat{\mathbf{x}}(\partial u_x/\partial\lambda)+\hat{\mathbf{y}}(\partial u_y/\partial\lambda))d\lambda$, where λ can be any one of $\{u_z,W_z,J_z\}$, or any other arbitrary monotonically increasing (during the collision) parameter.[17] W_z is usually included as an allowed parameter because it is monotonically decreasing (respectively, increasing) on $t_{\text{initial}}\leqslant t\leqslant t_{\text{max compression}}$ (respectively, $t_{\text{max compression}}\leqslant t\leqslant t_{\text{final}}$). For completeness we mention that there are exceptional occasions in which u_z initially decreases but then monotonically increases. There are standard methods[18] for handling this case, and as this case does not affect our results, we do not consider it further.

Using Eq. (5) and $d\mathbf{J}=\mathbf{f}dt$ we find that $d\Delta\mathbf{u}=d\mathbf{u}=\mathbf{K}\,\mathbf{f}dt$. We also recall that $dW_z=f_zu_zdt$. Eliminating dt from these equations gives $dJ_z/f_z=dW_z/f_zu_z=du_z/\mathbf{k}_{(3)}\cdot\mathbf{f}=du_x/\mathbf{k}_{(1)}\cdot\mathbf{f}=du_y/\mathbf{k}_{(2)}\cdot\mathbf{f}$ or

$$dJ_z=\frac{dW_z}{u_z}=\frac{du_z}{\mathbf{k}_{(3)}\cdot(\mathbf{f}/f_z)}=\frac{du_x}{\mathbf{k}_{(1)}\cdot(\mathbf{f}/f_z)}=\frac{du_y}{\mathbf{k}_{(2)}\cdot(\mathbf{f}/f_z)}. \tag{9}$$

These differential forms and Stronge's postulate enable us to calculate $\Delta\mathbf{u}$. $\mathbf{J}$ then follows from Eq. (7).

We shall only consider the case in which u_{tr} is not identically zero and f_z does not vanish and employ the dynamic Coulomb friction law to model friction. Hence the force $\mathbf{f}$ has a transverse component that is directed opposite to the transverse component of the relative velocity, and has a magnitude given by μf_z, where μ is the coefficient of kinetic friction.

If we put

$$\xi=\frac{\mathbf{f}}{f_z}=\begin{pmatrix}-\mu\cos(\theta)\\-\mu\sin(\theta)\\1\end{pmatrix} \tag{10}$$

then, using Eq. (9), the differential system becomes

$$dJ_z=\frac{dW_z}{u_z}=\frac{du_z}{\mathbf{k}_{(3)}\cdot\xi}=\frac{du_x}{\mathbf{k}_{(1)}\cdot\xi}=\frac{du_y}{\mathbf{k}_{(2)}\cdot\xi}. \tag{11}$$

The calculation of $\Delta\mathbf{u}$ now follows from Eq. (11). We know that $u_z(t_{\text{max compression}})=0$. Formally integrating the differential system, we find that at time $t=t_{\text{max compression}}$, $W_z(t_{\text{max compression}})=\int_{W(t_{\text{initial}})}^{W(t_{\text{max compression}})} dW_z=\int_{u_z(t_{\text{initial}})}^{0}(u_z/\mathbf{k}_{(3)}\cdot\xi)du_z$ while $u_x(t_{\text{max compression}})-u_x(t_{\text{initial}})=\int_{u_z(t_{\text{initial}})}^{0}(\mathbf{k}_{(1)}\cdot\xi/\mathbf{k}_{(3)}\cdot\xi)du_z$ and $u_y(t_{\text{max compression}})-u_y(t_{\text{initial}})=\int_{u_z(t_{\text{initial}})}^{0}[(\mathbf{k}_{(2)}\cdot\xi)/\mathbf{k}_{(3)}\cdot\xi)]du_z$. Moreover, using Eq. (11) we find (formally) that at time $t=t_{\text{final}}$, $\mathbf{u}(t_{\text{final}})-\mathbf{u}(t_{\text{max compression}})=\int_{W(t_{\text{max compression}})}^{W(t_{\text{final}})=(1-e^2)W(t_{\text{max compression}})}\mathbf{K}\xi(dW_z/u_z)$, where we have used Stronge's hypothesis in the upper limit of the last integration.

A. Transverse relative velocity

Let ''pseudoforce'' components $\{F_x,F_y\}$ be defined by

$$F_x=\mathbf{k}_{(1)}\cdot\xi=-\mu\cos(\theta)K_{11}-\mu\sin(\theta)K_{12}+K_{13} \tag{12}$$

and

$$F_y=\mathbf{k}_{(2)}\cdot\xi=-\mu\cos(\theta)K_{21}-\mu\sin(\theta)K_{22}+K_{23}, \tag{13}$$

where we recall that $K_{21}=K_{12}$. Inspection of Eq. (11) shows that the equations of motion for u_x and u_y may be written as

$$\frac{du_x}{dJ_z}=a_x=F_x, \quad \frac{du_y}{dJ_z}=a_y=F_y, \tag{14}$$

where J_z is monotonically increasing during the collision. Here $\{a_x,a_y\}$ denotes the components of the ''pseudotransverse acceleration.''

The locus of all possible $\{a_x,a_y\}$ is an ellipse. To see this, substitute in Eq. (14) using Eqs. (12) and (13), solve for $\cos(\theta)$ and $\sin(\theta)$, then square and add. This yields

$$[K_{22}(a_x-K_{13})-K_{12}(a_y-K_{23})]^2+[-K_{12}(a_x-K_{13})+K_{11}(a_y-K_{23})]^2=\mu^2[K_{11}K_{22}-K_{12}K_{21}]^2. \tag{15}$$

This is the equation for an ellipse centered at (K_{13},K_{23}).

In terms of polar coordinates (u_{tr},θ), $u_x=u_{\text{tr}}\cos(\theta)$ and $u_y=u_{\text{tr}}\sin(\theta)$. Substituting this into Eq. (14) yields

$$\frac{du_{\text{tr}}}{dJ_z}=F_x\cos(\theta)+F_y\sin(\theta), \quad u_{\text{tr}}\frac{d\theta}{dJ_z}=-F_x\sin(\theta)+F_y\cos(\theta). \tag{16}$$

We denote the ''pseudotorque'' on the right-hand side of Eq. (11) by

$$\begin{aligned}D_0(\mathbf{K},\mu,\theta)&\equiv-F_x\sin(\theta)+F_y\cos(\theta)\\&=\mu\cos(\theta)\sin(\theta)K_{11}-\mu\cos(\theta)^2K_{12}+\mu\sin(\theta)^2K_{12}\\&\quad-\sin(\theta)K_{13}-\mu\cos(\theta)\sin(\theta)K_{22}+\cos(\theta)K_{23}.\end{aligned} \tag{17}$$

One observes that $D_0\equiv0$ when either $u_{\text{tr}}\equiv0$ (sticking occurs) or sliding is along a straight line given by

$$\tan(\theta)=\frac{F_y}{F_x}=\frac{-\mu\cos(\theta)K_{21}-\mu\sin(\theta)K_{22}+K_{23}}{-\mu\cos(\theta)K_{11}-\mu\sin(\theta)K_{12}+K_{13}}. \tag{18}$$

Putting $T=\tan(\theta/2)$ in Eq. (18) and solving for T in the resulting quartic equation gives the possible distinguished directions of sliding along a straight line. When sliding is along a straight line the transverse relative acceleration is either parallel or antiparallel to the transverse relative velocity, and the cosine of the included angle ϕ between $\mathbf{u}_\perp$ and $\mathbf{a}_\perp$,

$$\cos(\phi)=\frac{\mathbf{u}_\perp\cdot\mathbf{a}_\perp}{u_\perp a_\perp}=\frac{F_x\cos(\theta)+F_y\sin(\theta)}{\sqrt{F_x^2+F_y^2}}, \tag{19}$$

is ±1. If the transverse relative velocity is parallel to the transverse relative acceleration then u_{tr} is increasing and the flow of $\mathbf{u}_\perp$ in the u_x-u_y plane is called a *diverging* ray. If the transverse relative velocity is antiparallel to the transverse relative acceleration then u_{tr} is decreasing and the flow of $\mathbf{u}_\perp$ in the u_x-u_y plane is called a *converging* ray. Bhatt and Koechling[12] have shown that the total number of converging and diverging rays is 2 or 4.

We consider the case in which D_0 is not identically zero. Dividing the second of equations (16) into the first yields

$$\frac{1}{u_{\mathrm{tr}}}\frac{du_{\mathrm{tr}}}{d\theta}=\frac{F_x\cos(\theta)+F_y\sin(\theta)}{D_0(\mathbf{K},\mu,\theta)},$$

or

$$d\ln(u_{\mathrm{tr}})=\frac{F_x\cos(\theta)d\theta+F_y\sin(\theta)d\theta}{D_0(\mathbf{K},\mu,\theta)}. \tag{20}$$

The numerator of Eq. (20) may be recast using $F_x\cos(\theta)d\theta+F_y\sin(\theta)d\theta=F_x\,d\sin(\theta)-F_y\,d\cos(\theta)=d(F_x\sin(\theta)-F_y\cos(\theta))-\sin(\theta)dF_x+\cos(\theta)dF_y$. Hence

$$\begin{aligned}d\ln(u_{\mathrm{tr}})&=\frac{d(F_x\sin(\theta)-F_y\cos(\theta))-\sin(\theta)dF_x+\cos(\theta)dF_y}{-F_x\sin(\theta)+F_y\cos(\theta)}\\&=-d\ln(-F_x\sin(\theta)+F_y\cos(\theta))+\frac{-\sin(\theta)dF_x+\cos(\theta)dF_y}{-F_x\sin(\theta)+F_y\cos(\theta)}.\end{aligned}$$

Therefore

$$d\ln(u_{\mathrm{tr}})=-d\ln(D_0)+\frac{-\sin(\theta)dF_x+\cos(\theta)dF_y}{D_0(\mathbf{K},\mu,\theta)}. \tag{21}$$

The numerator of the last term of the right-hand side of Eq. (21) is $-\sin(\theta)dF_x+\cos(\theta)dF_y=(-\sin(\theta)^2K_{11}+2\cos(\theta)\sin(\theta)K_{12}-\cos(\theta)^2K_{22})d\theta$.

In terms of $T=\tan(\theta/2)$, the differential system Eq. (21) governing the dynamical evolution of the transverse speed is

$$d\ln(u_{\mathrm{tr}})=-d\ln(D_0)+\left(\frac{-\sin(\theta)dF_x+\cos(\theta)dF_y}{D_0(\mathbf{K},\mu,T)}\right)\frac{d\theta}{dT}dT.$$

After some calculation we find that

$$\left(\frac{-\sin(\theta)dF_x+\cos(\theta)dF_y}{D_0(\mathbf{K},\mu,T)}\right)\frac{d\theta}{dT}dT=-2\,\frac{T}{1+T^2}\,dT+2\,\frac{N(\mathbf{K},\mu,T)}{D(\mathbf{K},\mu,T)}\,dT$$

$$=-d\,\ln(1+T^2)+2\,\frac{N(\mathbf{K},\mu,T)}{D(\mathbf{K},\mu,T)}\,dT$$

where

$$N(\mathbf{K},\mu,T)\equiv\mu K_{22}-T(3\mu K_{12}+K_{23})+T^2(2(\mu K_{11}+K_{13})-\mu K_{22})+T^3(\mu K_{12}+K_{23}) \quad (22)$$

and

$$D(\mathbf{K},\mu,T)\equiv-(1+T^2)^2D_0(\mathbf{K},\mu,T)$$
$$=\mu K_{12}-K_{23}+2T(K_{13}+\mu(-K_{11}+K_{22}))-6T^2\mu K_{12}$$
$$+2T^3(K_{13}+\mu(K_{11}-K_{22}))+T^4(\mu K_{12}+K_{23}). \quad (23)$$

Therefore,

$$d\,\ln(u_{\mathrm{tr}})=-d\,\ln(D_0)-d\,\ln(1+T^2)+2\,\frac{N(\mathbf{K},\mu,T)}{D(\mathbf{K},\mu,T)}\,dT=-d\,\ln\frac{-D(\mathbf{K},\mu,T)}{(1+T^2)}+2\,\frac{N(\mathbf{K},\mu,T)}{D(\mathbf{K},\mu,T)}\,dT. \quad (24)$$

B. A first integral of the motion

The term $2[N(\mathbf{K},\mu,T)/D(\mathbf{K},\mu,T)]dT$ is an integrable rational form with a simple exact integral. The general form for the integral is well known and is given in, for example, Gradshteyn and Ryzhik,[19] Sec. 2.1. The precise functional form of the integral depends on the degree of $D(\mathbf{K},\mu,T)$ and the multiplicities of the roots of $D(\mathbf{K},\mu,T)$. We explicitly consider one important case here, the case when $D(\mathbf{K},\mu,T)$ has four distinct roots. Let T_r $r=1,2,3,4$ denote the simple roots of $D(\mathbf{K},\mu,T)$, $D(\mathbf{K},\mu,T)=c_4\Pi_{r=1}^{r=4}(T-T_r)$, where $c_4=\mu K_{12}+K_{23}\neq 0$ is the coefficient of T^4 in $D(\mathbf{K},\mu,T)$. Define $\psi(T)=2[N(\mathbf{K},\mu,T)/(d/dT)D(\mathbf{K},\mu,T)]$. We find that $[2N(\mathbf{K},\mu,T)/D(\mathbf{K},\mu,T)]dT=\Sigma_{r=1}^{r=4}\,\psi(T_r)d\,\ln|T-T_r|$ so that Eq. (24) becomes

$$d\,\ln(u_{\mathrm{tr}})=-d\,\ln\frac{-D(\mathbf{K},\mu,T)}{(1+T^2)}+2\,\frac{N(\mathbf{K},\mu,T)}{D(\mathbf{K},\mu,T)}\,dT=-d\,\ln\frac{-D(\mathbf{K},\mu,T)}{(1+T^2)}-\sum_{r=1}^{r=4}\,d\,\ln|T-T_r|^{-\psi(T_r)}. \quad (25)$$

This integrates to

$$\Xi(\mathbf{K},\mu,T)=-u_{\mathrm{tr}}(T)D_0(\mathbf{K},\mu,T)(1+T^2)\prod_{r=1}^{r=4}|T-T_r|^{-\psi(T_r)}$$
$$=u_{\mathrm{tr}}(T)\,\frac{D(\mathbf{K},\mu,T)}{1+T^2}\prod_{r=1}^{r=4}|T-T_r|^{-\psi(T_r)}$$
$$=u_{\mathrm{tr}}(T)\,\frac{\mu K_{12}+K_{23}}{1+T^2}\prod_{r=1}^{r=4}\frac{(T-T_r)}{|T-T_r|^{\psi(T_r)}}=\mathrm{constant}. \quad (26)$$

This is a new first integral (conservation law) for a binary rigid body collision of arbitrarily short duration, and the main result of this paper.

In Eq. (26) two of the T_r may be complex conjugates of each other, say T_3 and $T_4=T_3^*$. Then two of the $\psi(T_r)$ are complex conjugates of each other, $\psi(T_3)$ and $\psi(T_4)=\psi(T_3)^*$. Put $T_3=x+iy$ and $\psi(T_3)=\alpha+i\beta$, where x, y, α, and β are real. Consider $\Sigma_{\mathrm{complex}}\,\psi(T_r)\ln|T-T_r|$

$=\Sigma_{r=3}^{4}\psi(T_r)\ln|T-T_r|=(\alpha+i\beta)\ln[T-(x+iy)]+(\alpha-i\beta)\ln[T-(x-iy)]=\alpha\ln[(T-x)^2+y^2]+\beta(2\tan^{-1}\times([y/(T-x)])+2n\pi)$. Also $\Pi_{\text{complex}}|T-T_r|^{\psi(T_r)}=\Pi_{\text{complex}}\exp(\psi(T_r)\ln|T-T_r|)=\exp(\Sigma_{\text{complex}}\times\psi(T_r)\ln|T-T_r|)=\exp(\alpha\ln[(T-x)^2+y^2]+\beta(2\tan^{-1}([y/(T-x)]+2n\pi))=|(T-x)^2+y^2|^{\alpha}\exp[\beta\times(2\tan^{-1}([y/(T-x)])+2n\pi)]$, where n is an integer (the branch cut for the natural logarithm is on the negative real axis; if it is crossed then the phase may change discontinuously). These obvious identities may be used as needed to recast expressions involving $\Pi_{\text{complex}}|T-T_r|^{\psi(T_r)}$ into a manifestly real form.

We may use Eq. (26) to eliminate u_{tr} in the second of Eq. (16), $u_{\text{tr}}(d\theta/dJ_z)=u_{\text{tr}}(d\theta/dT)\times(dT/dJ_z)=u_{\text{tr}}[2/(1+T^2)(dT/dJ_z)]=-F_x\sin(\theta)+F_y\cos(\theta)=D_0(\mathbf{K},\mu,T)$. Let $\Xi=\Xi(\mathbf{K},\mu,T\times(\theta_{\text{initial}}))$. Multiplying on each side with $(1+T^2)^2D_0(\mathbf{K},\mu,T)\Pi_{r=1}^{r=4}|T-T_r|^{-\psi(T_r)}$ yields

$$2\Xi(dT/dJ_z)=[(1+T^2)D_0(\mathbf{K},\mu,\theta)]^2\prod_{r=1}^{r=4}[T-T_r]^{-\psi(T_r)}=\left[\frac{D(\mathbf{K},\mu,T)}{1+T^2}\right]^2\prod_{r=1}^{r=4}|T-T_r|^{-\psi(T_r)}.$$

Therefore

$$2\Xi\left(\left[\frac{1+T^2}{D(\mathbf{K},\mu,T)}\right]^2\prod_{r=1}^{r=4}|T-T_r|^{\psi(T_r)}\right)dT=2\Xi\left[\left(\frac{1+T^2}{c_4}\right)^2\prod_{r=1}^{r=4}\frac{|T-T_r|^{\psi(T_r)}}{(T-T_r)^2}\right]dT=dJ_z,$$

or

$$\frac{2\Xi}{c_4^2}\left[(1+T^2)^2\prod_{r=1}^{r=4}|T-T_r|^{\psi(T_r)-2}\right]dT=dJ_z=\frac{du_z}{\mathbf{k}_{(3)}\cdot\xi}=\frac{dW_z}{u_z}. \tag{27}$$

This result can be (numerically) integrated to determine the dynamical evolution of θ during the collision.

V. DISCUSSION

Physically, $u_{\text{tr}}=0$ corresponds to (perhaps instantaneous) sticking. During the impact u_{tr} usually changes, and the conventional wisdom is that it may become zero. This idea is based on physical insight into the problem, but not on analytical results derived from a specific analytical model of friction (the dynamic Coulomb friction law, in this case). However the conservation law of Eq. (26) guarantees that if the constant in the law is not initially zero, then u_{tr} is never zero. Analytically, therefore, the dynamic Coulomb friction law does not formally allow an initially nonzero u_{tr} to vanish, except possibly for the special cases of $\mathbf{u}_\perp$ flow along either a converging or a diverging ray. This may bear repeating: using Eq. (26) we have proven that *the dynamic Coulomb friction law does not allow either instantaneous sticking or stable sticking to evolve from an initially nonzero* u_{tr}, except possibly for the special cases of $\mathbf{u}_\perp$ flow along either a converging or a diverging ray.

This somewhat surprising prediction is certainly not a dramatic refutation of the traditional view of sticking during a two-body collision. The traditional view is based largely on observation and heuristic arguments, but not on exact analytical results founded on a mathematical model of friction (of limited validity). There is no reason for a calculation based on Eq. (26) and an accurate calculation based on the heuristic approach to significantly disagree with one another.

VI. EXAMPLE

The result of Eq. (26) may be employed to compute the transverse relative velocity during a two-body collision. Before focusing on concrete examples, let us first review the concepts of stable and unstable sticking. Consider an arbitrary model of friction and the consequences of $u_{\text{tr}}=0$. If sticking is *stable* then the governing differential system implies that $du_x=0=du_y$, which evidently requires that $\mathbf{k}_{(1)}\cdot\mathbf{f}=0=\mathbf{k}_{(2)}\cdot\mathbf{f}$. This implies that $\mathbf{f}\propto\mathbf{k}_{(1)}\times\mathbf{k}_{(2)}\propto\mathbf{K}_{(3)}^{-1}$ $[\mathbf{k}_{(1)}\times\mathbf{k}_{(2)}\neq 0$ since $\det(K)\neq 0]$. Another way to see this is to note that for the case of instantaneous sticking, the relative velocity $\mathbf{u}=u_z\hat{\mathbf{z}}$. If sticking is stable, an arbitrary time variation δt in this system preserves sticking. In order that u_x and u_y remain zero it is necessary that $\delta u_x=\dot{u}_x\delta t=0=\dot{u}_y\delta t=\delta u_y$, so that $\dot{\mathbf{u}}=\dot{u}_z\hat{\mathbf{z}}$. Therefore

$$\delta\mathbf{J}=\mathbf{f}\delta t=\mathbf{K}^{-1}\delta\Delta\mathbf{u}=\mathbf{K}^{-1}\delta\mathbf{u}=\delta u_z\mathbf{K}_{(3)}^{-1}=\mathbf{K}_{(3)}^{-1}\frac{du_z}{dJ_z}\frac{dJ_z}{dt}\delta t=\mathbf{K}_{(3)}^{-1}\frac{1}{(dJ_z/du_z)}f_z\delta t=\frac{\mathbf{K}_{(3)}^{-1}}{\mathbf{K}_{33}^{-1}}f_z\delta t.$$

We conclude that

$$\frac{\mathbf{f}}{f_z}=\frac{\mathbf{K}_{(3)}^{-1}}{\mathbf{K}_{33}^{-1}}=\frac{\mathbf{k}_{(1)}\times\mathbf{k}_{(2)}}{K_{11}K_{22}-K_{12}K_{21}} \tag{28}$$

for such a system.

The static friction force components can maintain sticking only if $(f_x)^2+(f_y)^2\leqslant\mu_{\text{static}}^2(f_z)^2$ (e.g., in 1-dim, suppose that $|f_x|$ is required for sticking, while $\mu_{\text{static}}|f_z|$ is available), which implies that $(\mathbf{K}_{31}^{-1})^2+(\mathbf{K}_{32}^{-1})^2\leqslant\mu_{\text{static}}^2(\mathbf{K}_{33}^{-1})^2$, in light of the previous result for $\mathbf{f}$. This is the well-known condition for *stable* sticking. If the pseudotransverse acceleration ellipse encircles the origin then it is known that, in principle, the frictional forces are strong enough to preserve sticking if $u_{\text{tr}}=0$, and that in this case there is no diverging ray of constant sliding direction.[12,14] This means that sliding would permanently cease if u_{tr} should ever become zero. The theory embodied in Eq. (26) tells us that once u_{tr} gets ''small'' the flow of $\mathbf{u}_\perp$ is trapped by converging rays and is driven to even smaller values of u_{tr} toward some converging ray. Thereafter changes in θ, which are proportional to $D(\mathbf{K},\mu,T)$, are also very small since $D(\mathbf{K},\mu,T)=0$ on a ray. Effectively the system is ''stuck.'' Therefore our analytical results admit stable sticking.

Otherwise, if frictional forces are not strong enough to preserve sticking and if, in some particular physical model of friction, u_{tr} evolves to zero, then slipping immediately resumes along a diverging ray. This is called *instable* sticking. It is known that if frictional forces are not strong enough to preserve sticking,[14] then there is exactly one diverging ray.[12,14] The theory described in this paper accommodates instable sticking in a natural manner, as can be seen in the following example.

In this first example the impact is characterized by a coefficient of kinetic friction $\mu=0.2$ and a collision matrix

$$\mathbf{K}=\begin{pmatrix}+8 & -2 & +1\\ -2 & +3 & -1\\ +1 & -1 & +5\end{pmatrix}. \tag{29}$$

These values are used by Mirtich in an example in Ref. 14 and we would like to compare our results with his, since our method of computing results and Mirtich's method are completely different. The determinant of $\mathbf{K}$ is $\Delta=93$ and

$$\mathbf{K}^{-1}=\frac{1}{\Delta}\begin{pmatrix}14 & 9 & -1\\ 9 & 39 & 6\\ -1 & 6 & 20\end{pmatrix}.$$

Notice that $(\mathbf{K}_{31}^{-1})^2+(\mathbf{K}_{32}^{-1})^2=(1/\Delta^2)37>\mu_{\text{static}}^2(\mathbf{K}_{33}^{-1})^2=(1/\Delta^2)16$ so that sticking is not stable. We shall not specify the coefficient of restitution because we are not trying to find θ as a function of J_z, but are instead studying the behavior of (u_x,u_y) as functions of θ.

In Fig. 1 we plot the flow lines of the transverse relative velocity as θ varies for an initial transverse relative speed of ten units; its particular magnitude is immaterial as long as it is nonzero because conservation law is a function of u_{tr} times a function of angle. The flow lines are computed using the first integral of Eq. (26) for selected values of θ^{initial}, where the constant in the

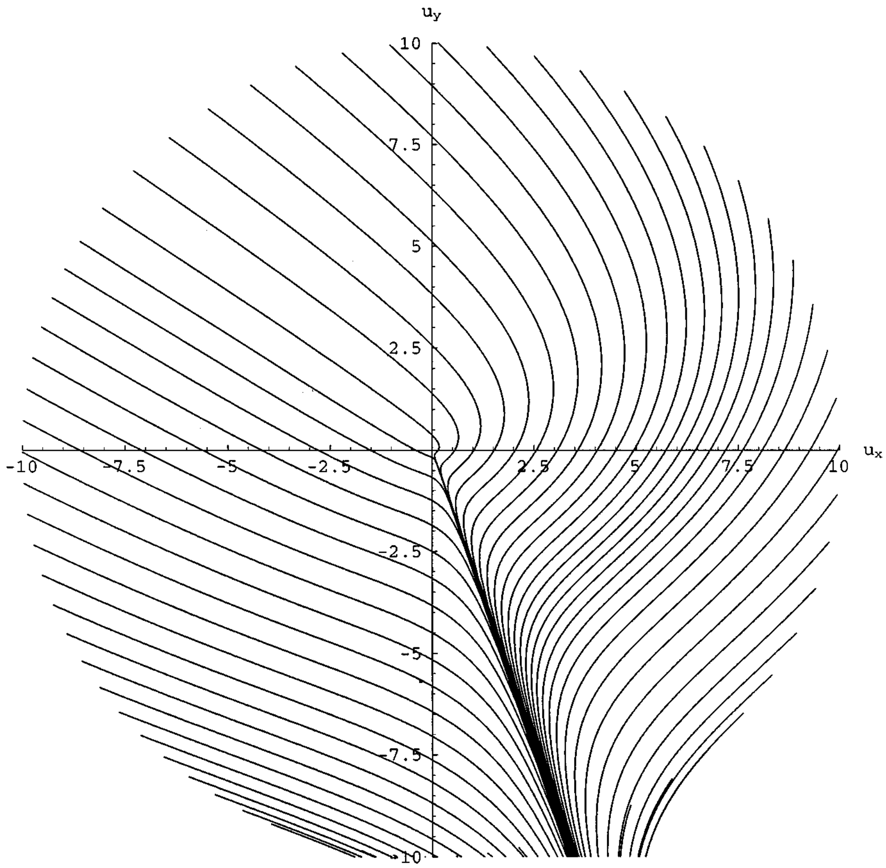

FIG. 1. Scale invariant transverse relative velocity flows for unstable sticking.

conservation law is evaluated using $u_{\mathrm{tr}}^{\mathrm{initial}}=10$ and $T^{\mathrm{initial}}=\tan(\frac{1}{2}\theta^{\mathrm{initial}})$ for each individual case. Of course, the flow lines appear qualitatively the same when plotted for other initial transverse relative speeds such as 1 unit, 0.1, 0.01, and so on. The flow exhibits a type of scale invariance that is a consequence of the fact that the conservation law is a function of u_{tr} times a function of angle.

Figure 1 confirms that there is a diverging ray (transverse relative velocity parallel to transverse relative acceleration) at $\theta_{\mathrm{diverging}}=-71.1127\,\mathrm{deg}$ that attracts outgoing trajectories and a converging ray (transverse relative velocity antiparallel to transverse relative acceleration) at $\theta_{\mathrm{converging}}=147.001\,\mathrm{deg}$. Using Eq. (19) we verify that there is a single converging ray and a single diverging ray at these angles. If the transverse relative velocity lies on one of these rays then the direction of sliding remains constant, until $u_{\mathrm{tr}}=0$.

We find that in this case there are two simple real roots and a complex conjugate pair of roots. The roots of $D(\mathbf{K},\mu,T)$ are

$$T_r=\begin{pmatrix} -0.714\,777\,790\,129\,410\,917 \\ 3.376\,057\,857\,830\,754\,29 \\ 0.097\,931\,394\,720\,756\,870\,2+0.409\,889\,210\,722\,295\,516\;i \\ 0.097\,931\,394\,720\,756\,870\,2-0.409\,889\,210\,722\,295\,516\;i \end{pmatrix},$$

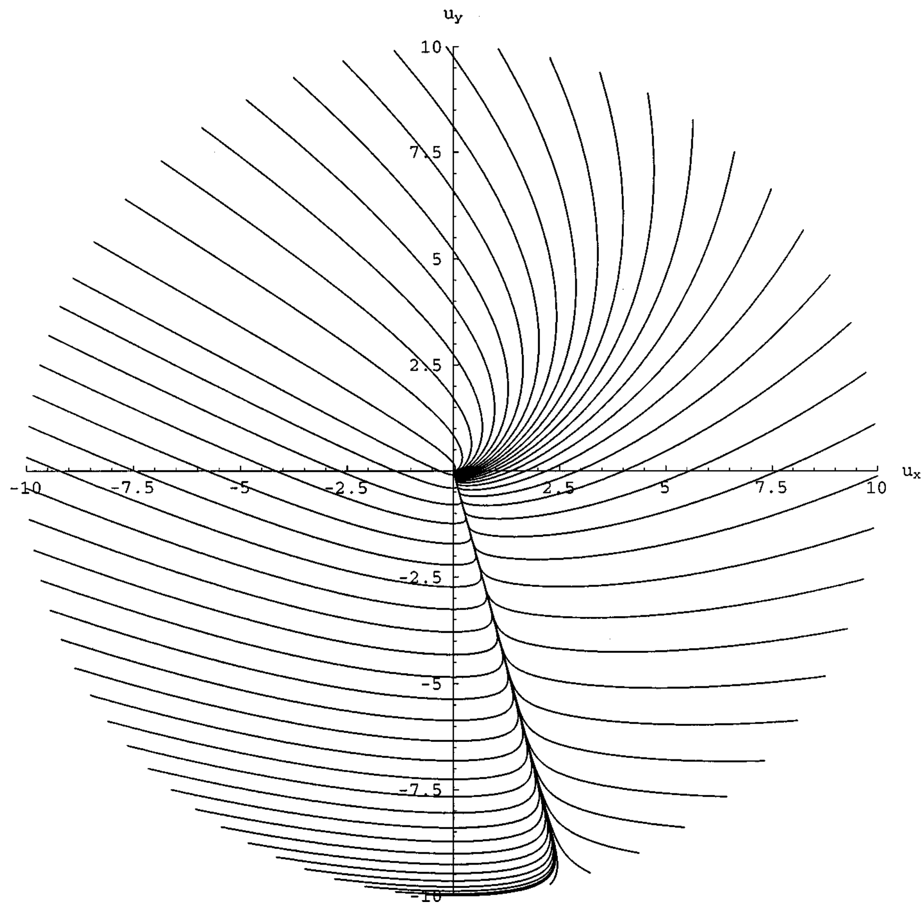

FIG. 2. Scale invariant transverse relative flows for stable sticking.

$$\psi(T_r)=\begin{pmatrix} 0.796\,175\,001\,065\,219\,056 \\ -0.210\,719\,532\,590\,733\,082 \\ 0.707\,272\,265\,762\,756\,902-0.347\,433\,352\,914\,288\,784\ i \\ 0.707\,272\,265\,762\,756\,902+0.347\,433\,352\,914\,288\,784\ i \end{pmatrix}.$$

Using the notation defined in the paragraph preceding Eq. (26)

$$\begin{pmatrix} x \\ y \\ \alpha \\ \beta \end{pmatrix}=\begin{pmatrix} 0.097\,931\,394\,720\,756\,870\,2 \\ 0.409\,889\,210\,722\,295\,516 \\ 0.707\,272\,265\,762\,756\,902 \\ -0.347\,433\,352\,914\,288\,784 \end{pmatrix}.$$

The conservation law Eq. (26) is equivalent to

$$u_{\mathrm{tr}}(T)\,\frac{\mathrm{sign}[T-T_1]\mathrm{sign}[T-T_2]}{1+T^2}|T-T_1|^{1-\psi(T_1)}|T-T_2|^{1-\psi(T_2)}\frac{[(T-x)^2+y^2]^{1-\alpha}}{\exp\left[\beta\left(2\tan^{-1}\left(\frac{y}{T-z}\right)+2\pi n\right)\right]}$$

$$=\mathrm{constant},$$

where $n=0$ for $-\pi<\theta_{\mathrm{diverging}}<\theta<\theta_{\mathrm{converging}}<\pi$ and

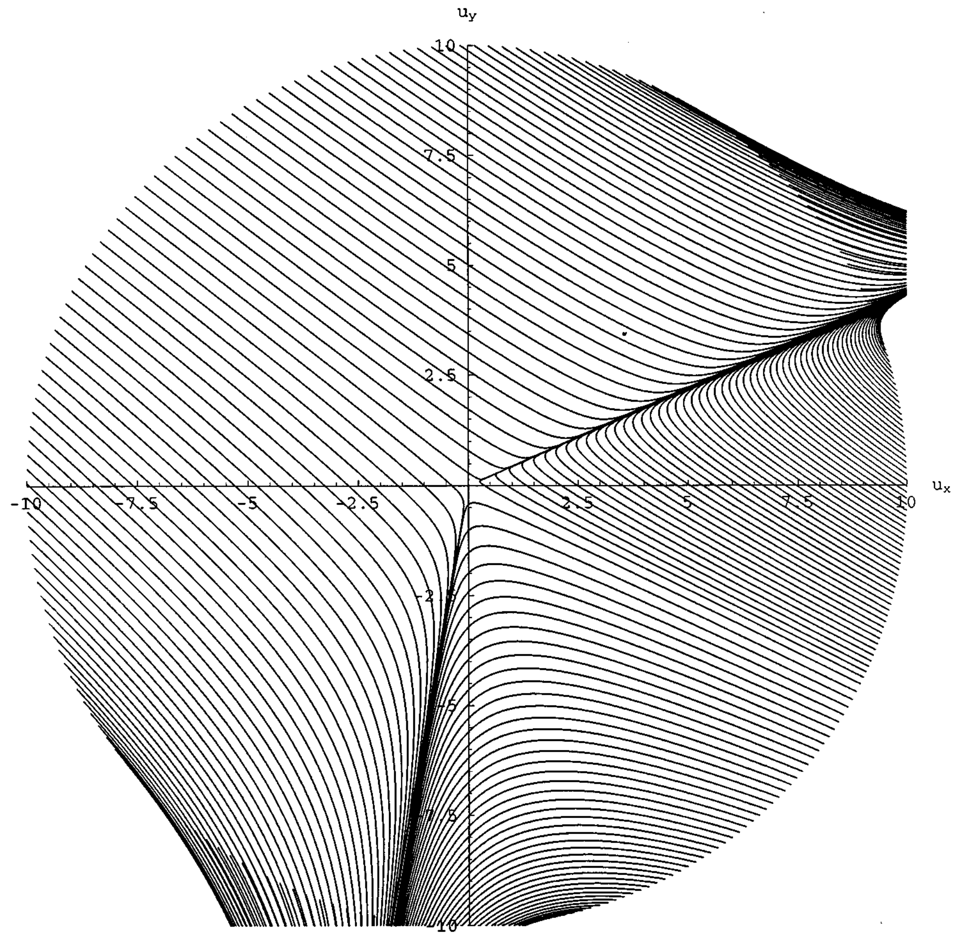

FIG. 3. Transverse relative velocity flows for unphysical **K**.

$$n=\begin{cases} 0, & u_y>0 \\ -1, & u_y<0 \end{cases}$$

for $\frac{1}{2}\pi<\theta_{\text{converging}}<\theta<2\pi+\theta_{\text{diverging}}$ (there is a branch cut for the logarithm on the negative real axis; we find that there is a change in phase of -2π across the cut). Notice that if we relabel roots and label $T_{3,4}$ as $T_{4,3}$, then $y\to -y$ and $\beta\to -\beta$; in this case $n=+1$ gives the correct change in phase of across the cut.

As previously stated, these same values for **K** and μ are used by Mirtich in an example in Ref. 14, where, however, the flow lines are computed by numerically integrating a system of ordinary differential equations equivalent to Eq. (11). The flows graphed in Fig. 1 computed using analytical methods and the flows presented in Mirtich's Figure[3.3] computed using numerical methods are in agreement.

Next, modify the collision matrix so that sticking is stable. Consider

$$\mathbf{K}=\begin{pmatrix} +12 & -2 & +1 \\ -2 & +5 & -1 \\ +1 & -1 & +1 \end{pmatrix}. \tag{30}$$

The determinant of **K** is $\Delta=43$ and

$$\mathbf{K}^{-1}=\frac{1}{\Delta}\begin{pmatrix} 4 & 1 & -3 \\ 1 & 1 & 10 \\ -3 & 10 & 56 \end{pmatrix}.$$

Since $(\mathbf{K}_{31}^{-1})^2+(\mathbf{K}_{32}^{-1})^2=(1/\Delta^2)109<\mu_{\text{static}}^2(\mathbf{K}_{33}^{-1})^2=(1/\Delta^2)125.44$, sticking is stable. In this case we find that there are two real simple roots. By Eq. (19), there are two converging rays.

In Fig. 2 we plot the flow lines of the transverse relative velocity as a function of θ. The initial transverse relative speed is again (arbitrarily) ten units and μ is unchanged. The flow lines are again computed using the conservation law Eq. (26) for selected values of θ^{initial}. Figure 2 confirms that there are two converging rays at $\theta_{\text{converging}}=-74.6983$ deg and 150.832 deg, respectively. We see that if $\mathbf{u}_{\text{tr}}$ becomes small, then it is swept into ever decreasing values as the collision progresses. Effectively, sticking occurs just as in the heuristic model.

For purposes of illustration, we modify the collision matrix so that it is not positive definite, and hence unphysical. Let

$$\mathbf{K}=\begin{pmatrix} +8 & -7 & +1 \\ -7 & +3 & -1 \\ +1 & -1 & +5 \end{pmatrix}. \tag{31}$$

The determinant of $\mathbf{K}$ is $\Delta=-122$ and

$$\mathbf{K}^{-1}=\frac{1}{\Delta}\begin{pmatrix} 14 & 34 & 4 \\ 34 & 39 & 1 \\ 4 & 1 & -25 \end{pmatrix}.$$

In this unphysical case we find that there are four real simple roots. Equation (19) reveals that there are two converging rays and two diverging rays.

In Fig. 3 we plot the flow lines of the transverse relative velocity as a function of θ. The initial transverse relative speed is again ten units and μ is unchanged. The flow lines are again computed using the conservation law Eq. (26) for selected values of θ^{initial}. Figure 3 confirms that there are a diverging rays at $\theta_{\text{diverging}}=-99.7051$ deg and 23.6415 deg and converging rays at $\theta_{\text{converging}}=-26.2886$ deg and 141.66 deg. However, this does not illustrate a possible physical motion since $\mathbf{K}$ is not positive definite.

[1] E. J. Routh, *Dynamics of a System of Rigid Bodies: Elementary Part*, 7th ed. (Dover, New York, 1960).
[2] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed. (Cambridge University Press, Cambridge, 1937).
[3] J. B. Keller, J. Appl. Mech. **53**, 1 (1986).
[4] W. J. Stronge, Proc. R. Soc. London, Ser. A **431**, 168 (1990).
[5] J. A. Battle and A. B. Condomines, Mech. Mach. Theory **26**, 565 (1991).
[6] W. J. Stronge, J. Appl. Mech. **58**, 1049 (1991).
[7] *Fundamentals of Friction: Macroscopic and Microscopic processes*, edited by I. L. Singer and H. M. Pollock (Kluwer, Dordrecht, 1992).
[8] Y. Wang and M. T. Mason, J. Appl. Mech. **59**, 635 (1992).
[9] W. J. Stronge, J. Appl. Mech. **60**, 149 (1993).
[10] F. Heslot, T. Baumberger, B. Perrin, B. Caroli, and C. Caroli, Phys. Rev. E **49**, 4973 (1994).
[11] B. Bhushan, J. N. Israelachvili, and U. Landman, Nature (London) **374**, 607 (1995).
[12] V. Bhatt and J. Koechling, J. Appl. Mech. **62**, 740 (1995).
[13] J. M. Carlson and A. A. Batista, Phys. Rev. E **53**, 4153 (1996).
[14] B. V. Mirtich, Ph.D. thesis, U.C. Berkely, 1996; http://www.merl.com/people/mirtich/papers/thesis/thesis.html.
[15] M. Weiss and F. J. Elmer, Phys. Rev. B **53**, 7539 (1996).
[16] S. Nasuno, A. Kudrolli, and J. P. Gollub, Phys. Rev. Lett. **79**, 949 (1997).
[17] F. P. Bowden and D. Tabor, *Friction and Lubrication* (Oxford University Press, New York, 1954).
[18] D. Baraff Comput. Graph. **23**, 223 (1989).
[19] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 5th ed. (Academic, New York, 1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Towards a classification of Euler–Kirchhoff filaments

Michel Nizette[a)]
Université Libre de Bruxelles, Departement de Physique Statistique CP231, Campus Plaine, 1050 Brussels, Belgium

Alain Goriely[b)]
University of Arizona, Department of Mathematics and Program in Applied Mathematics, Tucson, Arizona 85721

(Received 2 September 1998; accepted for publication 1 February 1999)

Euler–Kirchhoff filaments are solutions of the static Kirchhoff equations for elastic rods with circular cross sections. These equations are known to be formally equivalent to the Euler equations for spinning tops. This equivalence is used to provide a classification of the different shapes a filament can assume. Explicit formulas for the different possible configurations and specific results for interesting particular cases are given. In particular, conditions for which the filament has points of self-intersection, self-tangency, vanishing curvature or when it is closed or localized in space are provided. The average properties of generic filaments are also studied. They are shown to be equivalent to helical filaments on long length scales. © *1999 American Institute of Physics.* [S0022-2488(99)05006-9]

I. INTRODUCTION

The study of elastic deformations in rods has a long tradition in mathematics, physics, and engineering, dating back to Euler and Lagrange. Engineers have been confronted to the problem of coiling in sub-oceanic cables and have tried to understand the process of loop formation in twisted wires.[1–3] In chemistry and biology, increasing interest is taken in the elastic character of filamentary structures such as polymers[4–7] (such as DNA molecules[8–15]) and bacterial fibers,[16–18] for which the macroscopic theory of rods provides an idealized model. Long, twisted structures also play an important role in hydrodynamic models[19] such as scroll wave propagation,[20] vortex tube motion,[21] or sun spots formation and solar corona heating.[22,23]

The Kirchhoff model (1859) (Ref. 24) provides the basic framework for the theory of elastic filaments. A remarkable feature of this model, known as the Kirchhoff kinetic analogy, is that the equations governing the static phenomena are formally equivalent to the Euler equations describing the motion of a rigid body with a fixed point under an external force field. The statics of rods is thus intimately connected to the dynamics of spinning tops, a problem to which innumerable work has been devoted. For instance, the most studied case where the filament has a circular cross section is shown to correspond to a top having an axis of revolution, in which case the equations are fully integrable. There is a rich mathematical literature on the statics of rods (see for instance, Ref. 26). In the particular case of circular cross sections and linear elasticity, various researchers have considered particular equilibrium filament shapes (helices,[25,27] rings,[28,3] localizing buckling modes,[1,29] solutions having points with vanishing curvature,[30] supercoiled helices,[31] see also Ref. 32 for a Hamiltonian formulation and Ref. 33 for a group theory approach). Recently, departing from the traditional Euler angles approach, Shi and Hearst (1994) have obtained a closed form of the general solution of the static Kirchhoff equations for circular cross sections.[34] Despite this achievement, it remains highly nontrivial to obtain a global picture of all possible static filament shapes. A step towards a general geometric classification of the equilibrium solutions is provided here by considering in more detail the analogy between filaments and spinning tops. In Sec. III, we

[a)]Electronic mail: mnizette@ulb.ac.be
[b)]Electronic mail: goriely@math.arizona.edu

rederive Shi and Hearst's general solution, keeping an explicit dependence in the constants of the motion of the spinning top, and we discuss in detail various filament configurations on the basis of the corresponding orbits of the top, recovering the aforementioned shapes as particular cases.

In this paper, we provide new results on the different configurations of Kirchhoff filaments in 2D and 3D. We give explicit formulas for the centerline coordinate and find conditions for self-tangency and self-intersection of planar filaments, conditions for vanishing curvature and average behavior of long filaments. Explicit formulas for periodic or localized (homoclinic) filaments in space are also given.

II. THE KIRCHHOFF MODEL

We first introduce the Kirchhoff model. It accounts for the dynamics of a thin elastic filament subject to internal stresses and boundary constraints. A filament is a unidimensional piece of elastic material which can be mathematically modelized by a curve in space, together with extra information about its twist, that is, how longitudinal material lines on the edge of the filament wind around it. This curve-plus-twist concept is formalized in the notion of ribbon discussed in Sec. II A as a preliminary.

A. Space curves and ribbons

We define a dynamical *space curve* $\mathbf{R}(s,t)$ as a smooth function mapping $\mathbb{R}^2$ into the physical space $\mathbb{R}^3$, and taking as variables the arc length s and the time t. For every s and t we define the Frenet basis $(\mathbf{n},\mathbf{b},\mathbf{t})$ to be the normal, binormal and tangent vectors to the curves. These vectors form the Frenet basis. The tangent vector is a unit vector given by $\mathbf{t}=(\partial\mathbf{R}/\partial s)$. The curvature κ of the curve at the point s is then given by

$$\kappa=\left|\frac{\partial\mathbf{t}}{\partial s}\right|. \tag{1}$$

At points where the curvature does not vanish, the normal vector is defined by

$$\frac{\partial\mathbf{t}}{\partial s}=\kappa\mathbf{n}. \tag{2}$$

The third unit vector $\mathbf{b}$ is

$$\mathbf{b}=\mathbf{t}\times\mathbf{n}. \tag{3}$$

Therefore, the Frenet basis is a right-handed orthonormal basis on the space curve $\mathbf{R}$. As a consequence, one has

$$\frac{\partial\mathbf{n}}{\partial s}=\tau\mathbf{b}-\kappa\mathbf{t}. \tag{4}$$

This relation defines the torsion τ, which measures the amount of rotation of the Frenet triad around the tangent $\mathbf{t}$ as the arc length increases. Finally the derivative of the binormal $\mathbf{b}$ is given in terms of the normal and tangent vector,

$$\frac{\partial\mathbf{b}}{\partial s}=-\tau\mathbf{n}. \tag{5}$$

The coupled equations (2), (4), and (5) are the Frenet–Serret equations. If the curvature κ and the torsion τ are known for all s, the Frenet triad $(\mathbf{n},\mathbf{b},\mathbf{t})$ can be obtained as the unique solution of the Frenet–Serret equations. It is then possible to reconstruct the space curve $\mathbf{R}$ by integrating the tangent vector $\mathbf{t}$.

A ribbon is a space curve $\mathbf{R}(s,t)$ supplied with a smooth unit vector field $\mathbf{d}_2(s,t)$ orthogonal to the curve. Let $\mathbf{d}_3=\mathbf{t}$ be the unit tangent vector. A third unit vector field $\mathbf{d}_1=\mathbf{d}_2\times\mathbf{d}_3$ is introduced so that the triad $(\mathbf{d}_1,\mathbf{d}_2,\mathbf{d}_3)$ forms a right-handed orthonormal basis. This basis is a generalization of the Frenet triad $(\mathbf{n},\mathbf{b},\mathbf{t})$.

The components of the derivatives of the local basis vectors $\mathbf{d}_1$, $\mathbf{d}_2$, and $\mathbf{d}_3$ with respect to the arc length s and the time t expressed in the local basis form the twist vector $\kappa(s,t)=\kappa_1\mathbf{d}_1+\kappa_2\mathbf{d}_2+\kappa_3\mathbf{d}_3$ and the spin vector $\boldsymbol{\omega}(s,t)=\omega_1\mathbf{d}_1+\omega_2\mathbf{d}_2+\omega_3\mathbf{d}_3$, defined as follows:

$$\frac{\partial \mathbf{d}_i}{\partial s}=\boldsymbol{\kappa}\times\mathbf{d}_i, \quad i=1,2,3 \tag{6a}$$

$$\frac{\partial \mathbf{d}_i}{\partial t}=\boldsymbol{\omega}\times\mathbf{d}_i. \quad i=1,2,3 \tag{6b}$$

Equations (6a) constitute the generalization of the Frenet–Serret equations for the ribbon. These equations can also be expressed in terms of the twist matrix $\mathbf{K}(s,t)$ and the spin matrix $\mathbf{W}(s,t)$, which we define as follows:

$$\mathbf{K}=\begin{pmatrix} 0 & -\kappa_3 & \kappa_2 \\ \kappa_3 & 0 & -\kappa_1 \\ -\kappa_2 & \kappa_1 & 0 \end{pmatrix}, \quad \mathbf{W}=\begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}. \tag{7}$$

The Frenet–Serret equations then read

$$\left(\frac{\partial \mathbf{d}_1}{\partial s}\ \frac{\partial \mathbf{d}_2}{\partial s}\ \frac{\partial \mathbf{d}_3}{\partial s}\right)=(\mathbf{d}_1\mathbf{d}_2\mathbf{d}_3)\mathbf{K}, \tag{8}$$

$$\left(\frac{\partial \mathbf{d}_1}{\partial t}\ \frac{\partial \mathbf{d}_2}{\partial t}\ \frac{\partial \mathbf{d}_3}{\partial t}\right)=(\mathbf{d}_1\mathbf{d}_2\mathbf{d}_3)\mathbf{W}. \tag{9}$$

B. The Kirchhoff equations

1. Main assumptions and derivation of the Kirchhoff equations

A thin filament, or rod, can be modelized by a ribbon constituted of a space curve $\mathbf{R}$ joining the loci of the centroids of the cross sections, together with a vector field $\mathbf{d}_2$ attached to the filament material. The space curve $\mathbf{R}$ is referred to as the centerline of the rod.

The Kirchhoff equations describe the dynamical evolution of the filament under the effect of internal elastic stresses and boundary constraints, in the absence of external force fields such as gravity. Also, only local interactions, between adjacent cross sections, are considered, ignoring the possibility for two remote segments of the filament to intersect with each other. Let $\mathbf{F}(s)$ and $\mathbf{M}(s)$ be the total force and total moment exerted on the back side of a cylinder $C(s)$ by the cylinder $C(s+ds)$ whose cross-section shape is defined by $\mathcal{S}$, the set of all the values of the couple (x_1,x_2) corresponding to a material point inside a given cross section. The conservation of linear and angular momentum yields[35]

$$\frac{\partial \mathbf{F}}{\partial s}=\rho\mathcal{A}\frac{\partial^2\mathbf{R}}{\partial t^2}, \tag{10a}$$

$$\frac{\partial \mathbf{M}}{\partial s}+\mathbf{d}_3\times\mathbf{F}=\rho\left(I_2\mathbf{d}_1\times\frac{\partial^2\mathbf{d}_1}{\partial t^2}+I_1\mathbf{d}_2\times\frac{\partial^2\mathbf{d}_2}{\partial t^2}\right), \tag{10b}$$

where $\mathcal{A}$ denotes the area of the cross section and the quantities I_1 and I_2 are the principal moments of inertia of the cross section,

$$I_1=\int_{\mathcal{S}} dx_1 dx_2 x_2^2, \quad I_2=\int_{\mathcal{S}} dx_1 dx_2 x_1^2. \tag{11}$$

Equations (10) are closed by using the constitutive relation of linear elasticity relating the torque **M** to the twist vector $\boldsymbol{\kappa}$,

$$\mathbf{M}=EI_1\kappa_1\mathbf{d}_1+EI_2\kappa_2\mathbf{d}_2+\mu J\kappa_3\mathbf{d}_3, \tag{12}$$

where E is Young's modulus, μ is the shear modulus, and J is a function of the shape $\mathcal{S}$ of the cross section. In the particular case of a circular cross section, one has

$$I_1=I_2=\frac{J}{2}=\frac{\pi R^4}{4}, \tag{13}$$

where R is the radius of the cross section. The combinations EI_1 and EI_2 are called the principal bending stiffnesses of the rod and measure how strong an applied torque must be in order to bend it, whereas the combination μJ is called the torsional stiffness and measures how large an applied torsional moment must be in order to twist the rod.

The coupled Eqs. (10) and (12) constitute the dynamical Kirchhoff equations. These are three vector equations involving the local basis $(\mathbf{d}_1,\mathbf{d}_2,\mathbf{d}_3)$ and its derivatives, the tension **F** and the torque **M**, which add up to nine degrees of freedom, hence the system is closed. In the static case (if the time dependencies is dropped), the term in **R** vanishes from Eq. (10a),

$$\mathbf{F}'=0, \tag{14a}$$

$$\mathbf{M}'+\mathbf{d}_3\times\mathbf{F}=0, \tag{14b}$$

$$\mathbf{M}=EI_1\kappa_1\mathbf{d}_1+EI_2\kappa_2\mathbf{d}_2+\mu J\kappa_3\mathbf{d}_3. \tag{14c}$$

2. The Kirchhoff equations in scaled form

In order to restrict the number of independent constants in (14), we scale the variables, by choosing combination of the length $[L]$, time $[T]$, and mass $[M]$ units in the following way:

$$[M]=\rho\sqrt{AI_1}, \quad \frac{[M][L]^3}{[T]^2}=EI_1, \tag{15}$$

with

$$a=\frac{I_2}{I_1}, \quad b=\frac{\mu J}{EI_1}=\frac{J}{2I_1(1+\sigma)}, \tag{16}$$

where σ denotes the Poisson ratio. The constant a measures the asymmetry of the cross section. Our convention is to orient the vector fields $\mathbf{d}_1$ and $\mathbf{d}_2$ such that I_1 and I_2 are, respectively, the larger and smaller bending stiffnesses. In this case, we have

$$0<a\leqslant 1, \tag{17}$$

the value 1 being reached in the symmetric case where the moments of inertia are identical. The constant b is the scaled torsional stiffness. It involves the constant $1/(1+\sigma)$ which ranges from $\frac{2}{3}$, corresponding to incompressible media (if the volume is unchanged as the material is stretched), to 1, corresponding to hyperelasticity (if there is no striction as the material is stretched). In the particular case of a circular cross section, one has

TABLE I. Analogy between rigid bodies and static filaments.

Symbol	Meaning for rods	Meaning for rigid bodies
$\boldsymbol{d}_3$	Unit tangent vector	Unit vector joining the fixed point to the center of mass
$(\boldsymbol{d}_1,\boldsymbol{d}_2,\boldsymbol{d}_3)$	Basis attached to the rod	Basis attached to the solid body
s	Arc length	Time
$\boldsymbol{F}$	Tension	Force equal and opposite to gravity
$\boldsymbol{M}$	Moment	Angular momentum
κ	Twist vector	Angular velocity vector
EI_1, EI_2	Principal bending stiffnesses	Principal moments of inertia in directions orthogonal to $\boldsymbol{d}_3$
μJ	Torsional stiffness	Principal moment of inertia along $\boldsymbol{d}_3$
a	Bending stiffnesses ratio	Ratio of the moments of inertia in directions orthogonal to $\boldsymbol{d}_3$
b	Scaled torsional stiffness	Scaled moment of inertia along $\boldsymbol{d}_3$

$$b=\frac{1}{1+\sigma}\in\left[\frac{2}{3},1\right]. \tag{18}$$

The scaled system reads now

$$\mathbf{F}'=0, \tag{19a}$$

$$\mathbf{M}'+\mathbf{d}_3\times\mathbf{F}=0, \tag{19b}$$

$$\mathbf{M}=\kappa_1\mathbf{d}_1+a\kappa_2\mathbf{d}_2+b\kappa_3\mathbf{d}_3. \tag{19c}$$

The quantities involved still have a dimension. From (19), we see that the scaled moment $\mathbf{M}$ has the dimension of the inverse of a length, and that the scaled tension $\mathbf{F}$ has the dimension of the inverse of a squared length, hence every variable involved is a length to a power. However the system cannot be further simplified by choosing the length unit $[L]$, because the remaining constants a and b are already dimensionless. It is still possible to choose a convenient length scale for a given problem. For example, if we consider a finite rod, a natural choice for the length unit will be the length of the rod. In Sec. III, we choose the length unit $[L]$ such that the norm of $\mathbf{F}$ has a given value.

The fact that the length unit $[L]$ is undetermined has yet another implication. Considering the static system (14) or (19), we see that every known solution actually determines a one-parameter family of solutions. More precisely, if $\{\mathbf{F}(s),\mathbf{M}(s),\boldsymbol{\kappa}(s)\}$ is a solution of the system, then $\{\lambda^{-2}\mathbf{F}(\lambda s),\lambda^{-1}\mathbf{M}(\lambda s),\lambda^{-1}\boldsymbol{\kappa}(\lambda s)\}$ is another solution of the system for every real nonvanishing λ. That is, the system is scale-invariant. Furthermore, if such a transformation is performed on the solution together with a rescaling of the length unit $[L]$ by a factor λ^{-1}, the solution remains unchanged, although the rod thickness is modified by a factor λ^{-1}. Hence, the statics of a filament, in the limit of the Kirchhoff model, does not depend on the rod thickness.

3. Integrability of the static Kirchhoff equations

The static Kirchhoff equations (19) are formally identical to the Euler equations describing the motion of a rigid body with a fixed point under gravity, in the particular case where the axis joining the fixed point to the center of mass lies along a principal direction of inertia. (The correspondence between rigid body variables and rod variables is shown on Table I). Therefore, they are fully integrable in two cases of interest, namely, the Lagrange and Kowalevskaya cases. Actually, the Euler equations are integrable in three cases. The third one corresponds to the motion of a rigid body with arbitrary shape in free fall and is known as the Euler case. However, one cannot speak of full integrability in the context of filaments since this case corresponds to a vanishing tension vector $\mathbf{F}$ and represents only particular boundary conditions, hence a subset of all possible configurations which can be adopted by a filament for given values of the material parameters a and b.

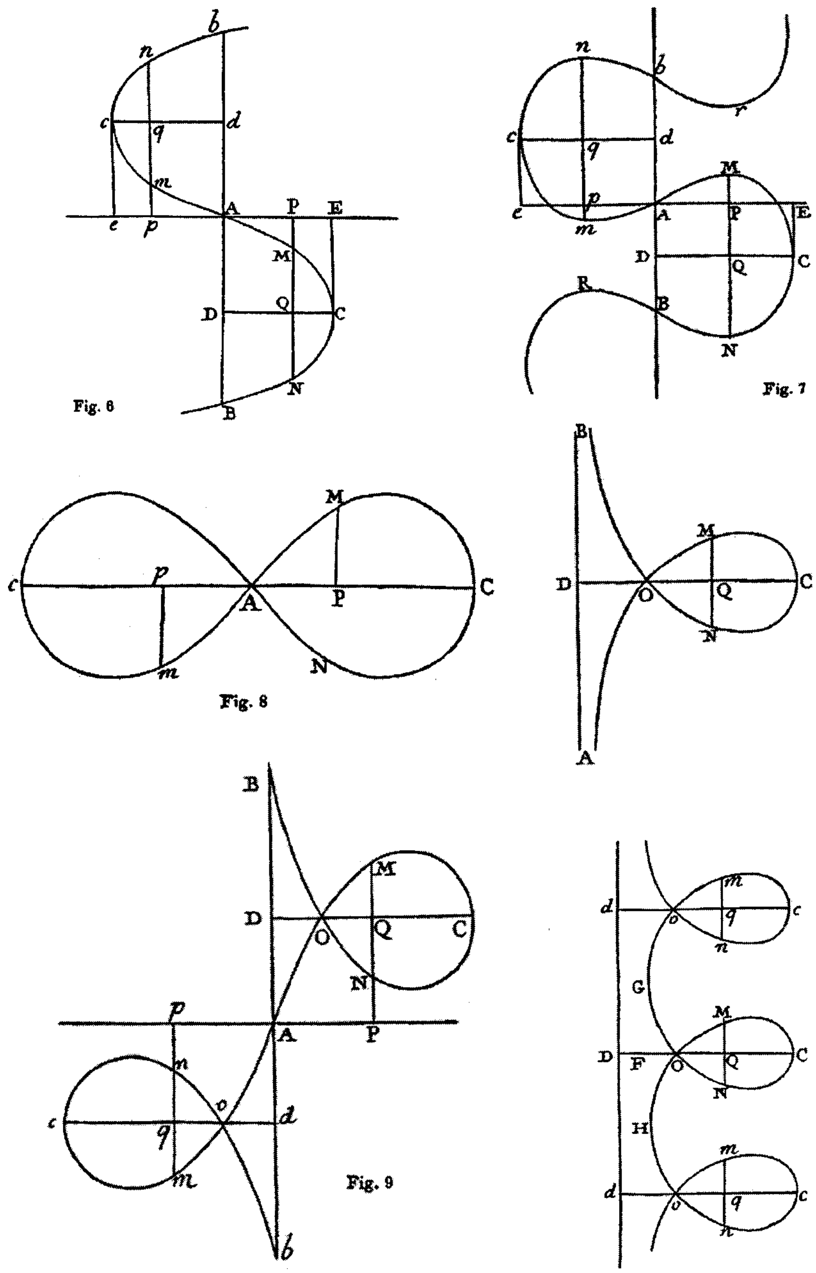

FIG. 1. Euler's drawings of planar filaments.

In the Lagrange case, the rigid body has identical moments of inertia along every direction perpendicular to the axis joining the fixed point to the center of mass, as for a symmetric top. The corresponding rods have identical bending stiffnesses in every direction, so that $a=1$. This is realized, for instance, in the very common situation where the filament cross section is circular, showing the importance of the Lagrange case for filaments. Much work has been devoted to it (Euler classified the planar solutions of the equations, see Fig. 1), and recently, Shi and Hearst (1994) (Ref. 34) have obtained a closed form for the general solution. Their results are rederived with special emphasis on the correspondence between rod and rigid body variables and developed in detail in Sec. III.

In the Kowalevskaya case, there exists an axis D, originating from the fixed point, along which the moment of inertia is half as large as the moments of inertia along directions perpendicular to D. In addition, the center of mass lies in the plane perpendicular to D. In the context of rods, this corresponds to asymmetric cross sections with $a=\frac{1}{2}$ and $b=1$. However, it turns out that the Kowalevskaya case, having a very high torsional stiffness, lies far beyond the region covered by the possible physical values of the parameters. For instance, an elliptic cross section with $a=\frac{1}{2}$ and $b=1$ would have a Poisson ratio $\sigma=-\frac{1}{3}$. This value describes a material which inflates transversally as it is stretched in length, and although this is not precluded in theory, it is not realized in practice.

III. SYMMETRIC RODS AND LAGRANGE TOPS

A. Generalities

The static Kirchhoff equations, written in terms of an appropriate set of variables, are formally equivalent to the Euler equations describing the dynamics of a heavy top. To every possible motion of the top, one can associate a particular static solution of the Kirchhoff equations. Here we focus exclusively on the analogy between tops and static rods, in one of the three cases where the equations are fully integrable; the Lagrange case, where the top has two identical principal moments of inertia. This condition is satisfied if it has a symmetry of revolution. The corresponding filament has identical bending stiffnesses in all directions, that is, we must set $a=1$ in the scaled form of the Kirchhoff equations (19).

The solutions of the Euler equations in the Lagrange case are well known and can be written as combinations of elliptic functions.[36–38] In order to obtain the centerline $\mathbf{R}$ of the corresponding static filament, we must identify the tangent vector $\mathbf{d}_3$ to a unit vector lying along the axis of revolution of the top. The centerline is then obtained by integrating $\mathbf{d}_3$ over the arc length s which, in the context of tops, corresponds to time. An obvious difficulty arises: in order to describe the behavior of real filaments under given external constraints, the space curve $\mathbf{R}$ must be available in a form that allows for boundary conditions at two distinct points. That is, it is essential to explicitly carry the integration of the tangent vector. Although it is not obvious that this integration can be performed, Shi and Hearst (1994) (Ref. 34) recently obtained expressions for the centerline $\mathbf{R}$ in cylindric coordinates in a closed analytic form involving elliptic functions. Despite this achievement, the problem is still partially unsolved. Indeed, explicit forms of the solutions are by no means sufficient to get a global insight on the large variety of possible filament shapes. Furthermore, the detailed analysis of the correspondence between spinning Lagrange tops and static symmetric rods provides a way of establishing an exhaustive geometric classification of the solutions. The aim of this paper is to provide a first step towards such a classification.

Shi and Hearst obtained their solutions by first solving the Kirchhoff equations for the curvature and torsion and then solving the Frenet–Serret equations to obtain the centerline. The resulting expressions depend on integration constants which do not have a clear meaning. Here, we depart from their approach and work consistently with variables and integration constants relevant to the top. Namely, in Sec. III C, we express the Kirchhoff–Euler equations in terms of the Euler angles. Then we compute the centerline $\mathbf{R}$ of the filament as a function of the constants of the motion for the spinning top. Once the expressions for the centerline have been obtained, we study various classes of shapes of the rod and their correspondence with the motion of the top. This analysis can be thought of an extension of Euler's work, who classified planar shapes of filaments.

In this section, we show illustrations of the spinning top orbits together with the corresponding filament shapes. A top orbit is displayed as a curve on the unit sphere which represents the extremity of the unit vector $\mathbf{d}_3$. The filament shapes are represented with a circular cross section, hence b must be set to a value lying between $\frac{2}{3}$ and 1. We have chosen $b=\frac{3}{4}$. The radius of the cross section and the zoom factor vary from one figure to another and are chosen for clarity (as seen in Sec. II, the radius is arbitrary in the static case).

In the remaining of this section, we write down the static Kirchhoff equations (19) in the case $a=1$, and identify a set of three first integrals necessary to guarantee the integrability of the system. With $a=1$, the system (19) reads

$$\mathbf{F}'=\mathbf{0}, \tag{20a}$$

$$\mathbf{M}'+\mathbf{d}_3\times\mathbf{F}=\mathbf{0}, \tag{20b}$$

$$\mathbf{M}=\kappa_1\mathbf{d}_1+\kappa_2\mathbf{d}_2+b\kappa_3\mathbf{d}_3. \tag{20c}$$

The first equation expresses the fact that the tension $\mathbf{F}$ is constant. We choose it oriented along the third vector of the fixed basis,

$$\mathbf{F}=F\mathbf{e}_Z. \tag{21}$$

In terms of spinning tops, the tension corresponds to the opposite of the top weight, $-m\mathbf{g}$, which describes an external force field. In the same spirit, we consider the tension as having a fixed value. As a consequence, we consider F as a parameter rather than a first integral. Inserting (21) into (20b) and projecting along the fixed basis vector $\mathbf{e}_Z$, we have

$$\mathbf{M}'\cdot\mathbf{e}_Z=0. \tag{22}$$

The basis vector $\mathbf{e}_Z$ being independent of s, we can extend the derivative in (22) to take effect over the whole left-hand side, leading to

$$M_Z'=0, \tag{23}$$

where M_Z denotes the component of the moment along $\mathbf{e}_Z$. It is a first integral that represents the vertical component of the angular momentum of a spinning top. By projecting (20b) along $\mathbf{d}_3$, we obtain

$$\mathbf{M}'\cdot\mathbf{d}_3=(\mathbf{M}\cdot\mathbf{d}_3)'-\mathbf{M}\cdot\mathbf{d}_3'=0. \tag{24}$$

Using the fact that $\mathbf{d}_3'=\boldsymbol{\kappa}\times\mathbf{d}_3=\mathbf{M}\times\mathbf{d}_3$, we see that the second term of (24) vanishes identically, leading to

$$M_3'=0. \tag{25}$$

That is, the torsional moment M_3 is another first integral. From (20c), we see that it corresponds to a constant twist density. For the spinning top, M_3 represents the component of the angular momentum along the axis of revolution of the top.

Finally, taking the dot product of both sides of (20b) with $\boldsymbol{\kappa}$, we have

$$\mathbf{M}'\cdot\boldsymbol{\kappa}+(\mathbf{d}_3\times\mathbf{F})\cdot\boldsymbol{\kappa}=\mathbf{M}'\cdot\boldsymbol{\kappa}+\mathbf{F}\cdot(\boldsymbol{\kappa}\times\mathbf{d}_3)=\mathbf{0}. \tag{26}$$

Using (20c) and the expressions (6a) of the derivatives of the local basis $(\mathbf{d}_1,\mathbf{d}_2,\mathbf{d}_3)$ in terms of $\boldsymbol{\kappa}$, (26) reduces to

$$\kappa_1\kappa_1'+\kappa_2\kappa_2'+\mathbf{F}\cdot\mathbf{d}_3'=(\tfrac{1}{2}\mathbf{M}\cdot\boldsymbol{\kappa}+\mathbf{F}\cdot\mathbf{d}_3)'=0, \tag{27}$$

which provides the last first integral,

$$\tfrac{1}{2}\mathbf{M}\cdot\boldsymbol{\kappa}+\mathbf{F}\cdot\mathbf{d}_3=H. \tag{28}$$

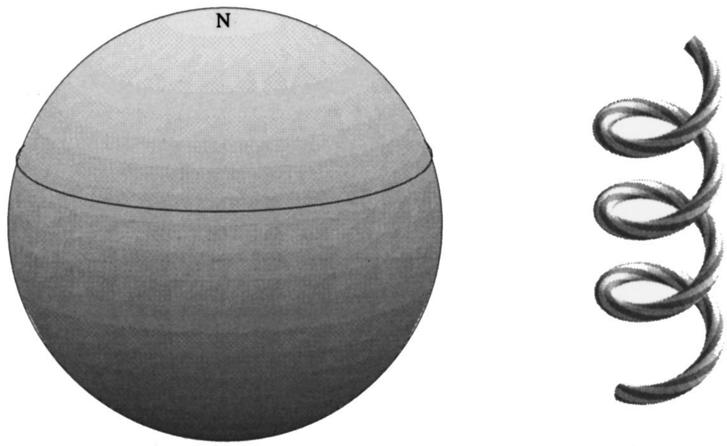

FIG. 2. An overtwisted helix ($\kappa=3$, $\tau=1$, $\kappa_3=6$).

This constant quantity H is the total elastic-plus-strain energy density of the filament and corresponds to the total kinetic-plus-potential mechanical energy of the spinning top. We now proceed to analyze different solutions of the system (20).

B. Helical filaments

1. The general helical solution

Helical solutions have constant Frenet curvature κ and torsion τ. Taking into account the fact that the twist density κ_3 is also constant, we introduce the following definitions:

(a) A Frenet helix is a helix with pure torsion, that is, $\kappa_3=\tau$.
(b) An overtwisted helix is a helix such that $(\kappa_3-\tau)$ has the same sign as τ.
(c) An undertwisted helix is a helix such that $(\kappa_3-\tau)$ and τ have opposite signs.

These three types of helices are represented in Figs. 2, 3, and 4, respectively. The undertwisted and overtwisted helices can be distinguished by the relation between the handedness of the helix itself and the handedness of the apparent twist pattern on the helix. In the case of an overtwisted helix, both hands are identical, whereas an undertwisted helix has opposite hands.

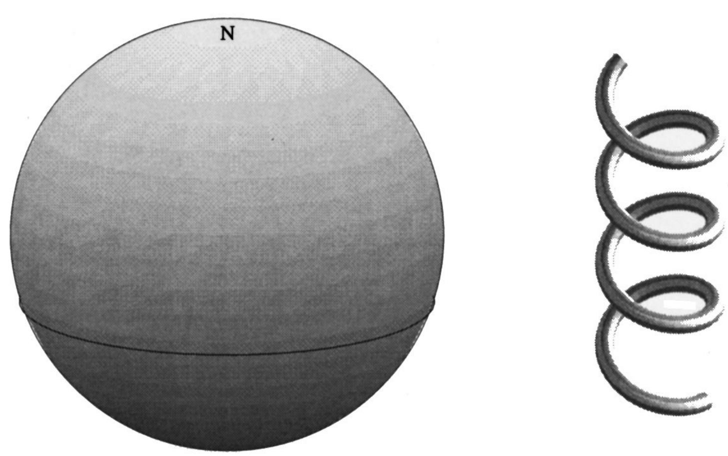

FIG. 3. A Frenet helix ($\kappa=3$, $\tau=-1$, $\kappa_3=-1$).

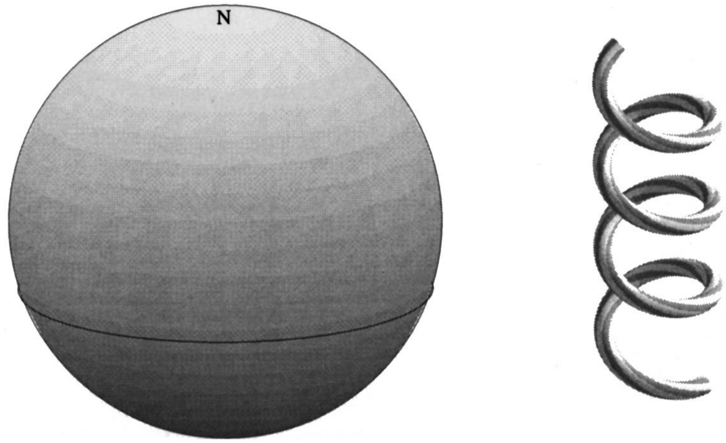

FIG. 4. An undertwisted helix ($\kappa=3$, $\tau=-1$, $\kappa_3=4$).

It is convenient to treat the cases of helical, circular (ring-like) and straight solutions of the Kirchhoff equations independently. The rings and the helical rods are easily identified as corresponding to spinning tops with the extremity describing a circle (centered at the fixed point in the case of rings), while the straight solutions correspond to the cases where the extremity of the top is at rest (the so-called sleeping tops); pointing upwards in the case of positive tension F_3 and downwards in the case of negative tension. Illustrations of these top orbits are given in Figs. 2–7. For the sake of simplicity, helical, circular, and straight filaments will be referred to as helical filaments.

We do not introduce Euler angles for the helical solutions. Instead, we introduce a fixed Frenet curvature κ and a fixed torsion τ into the expressions for the twist vector,

$$\kappa_1=\kappa \sin[(\kappa_3-\tau)(s-s_0)], \tag{29a}$$

$$\kappa_2=\kappa \cos[(k_3-\tau)(s-s_0)]. \tag{29b}$$

Substituting this into (20), we obtain one single nontrivial vector condition,

$$\mathbf{F}=(b\kappa_3-\tau)[\boldsymbol{\kappa}-(\kappa_3-\tau)\mathbf{d}_3], \tag{30}$$

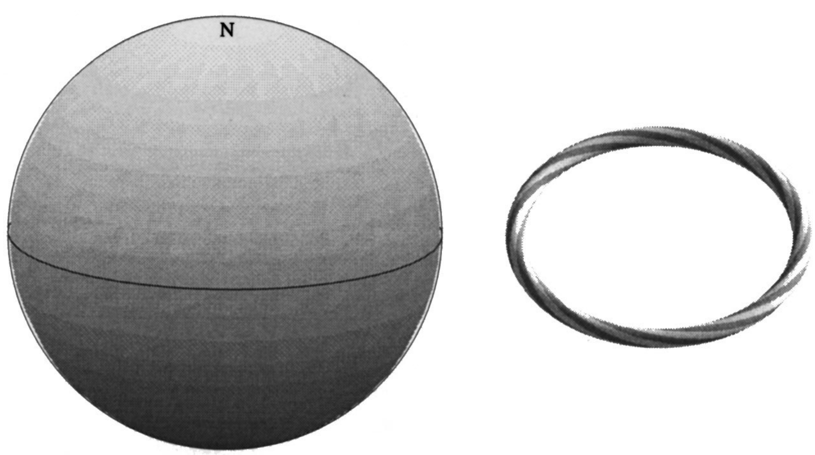

FIG. 5. A twisted ring ($\kappa=1$, $\tau=0$, $\kappa_3=3$).

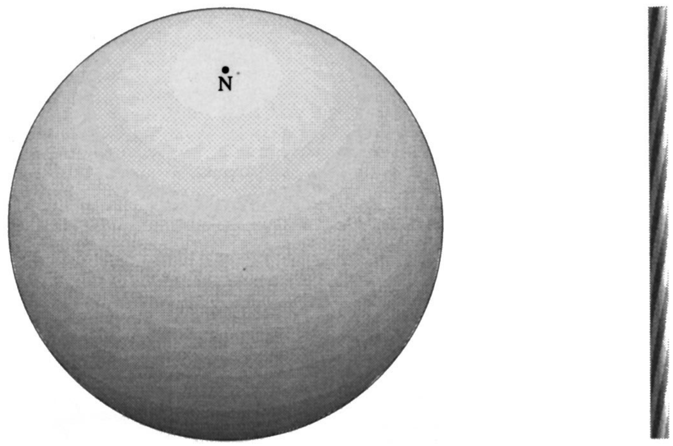

FIG. 6. A straight rod subject to extensive tension ($\kappa=0$, $\kappa_3=1$, $F_3=1$).

which simply gives the tension in the local basis as a constant combination of the other quantities. Hence, there exist helical solutions with arbitrary curvature κ, torsion τ, and twist density κ_3. This is unique to the case of symmetric rods. Indeed, in the regular case $a\neq 1$ the only possible helices are Frenet helices with either $\kappa_1=0$ or $\kappa_2=0$.[39]

Three more facts can be noticed from (30). First, in the case of rings ($\tau=0$), the tension has no longitudinal component ($\mathbf{F}\cdot\mathbf{d}_3=0$). Second, in the case of a noncircular filament ($\tau\neq 0$), a vanishing longitudinal tension implies $b\kappa_3=\tau$, hence the tension vector $\mathbf{F}$ itself vanishes. Finally, in the case of circular cross section, one has $b\leqslant 1$. This means that the Frenet helices ($\kappa_3=\tau$) have a negative longitudinal tension, and that the helices with null tension are overtwisted.

For straight solutions ($\kappa=0$), (30) does not hold and is replaced by

$$\mathbf{F}=F_3\mathbf{d}_3, \tag{31}$$

where F_3 is constant and arbitrary. Hence, there exist straight solutions with arbitrary longitudinal tension and twist density. The torsion has no meaning for straight rods. Figures 2–7 show representations of helical, ringlike, and straight solutions.

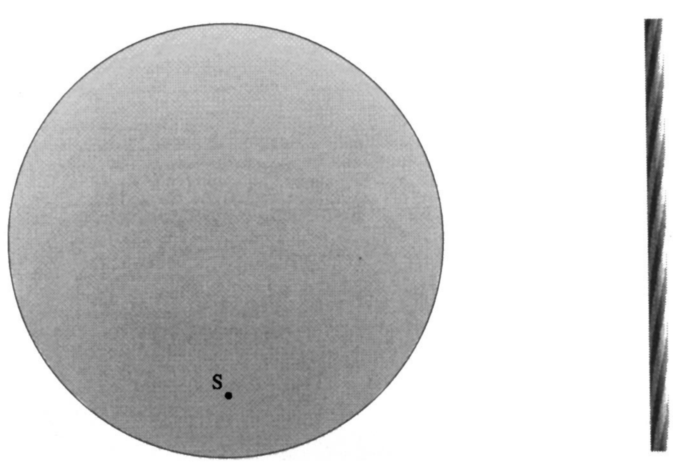

FIG. 7. A straight rod subject to compressive tension ($\kappa=0$, $\kappa_3=1$, $F_3=-1$).

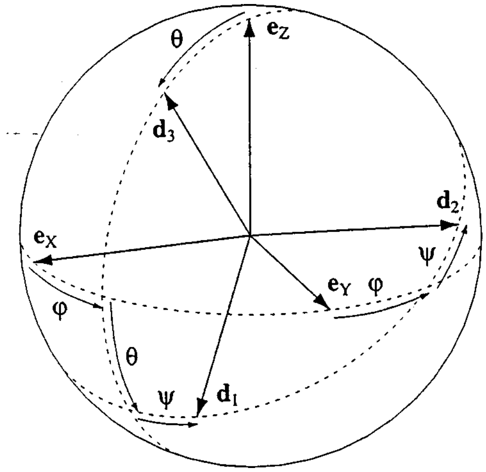

FIG. 8. The Euler angles.

2. The general solution with null tension

Substituting $\mathbf{F}=\mathbf{0}$ into (20b), we see that the moment $\mathbf{M}$ is constant. Differentiating (20c) with respect to s and projecting the resulting equation successively along $\mathbf{d}_1$, $\mathbf{d}_2$, and $\mathbf{d}_3$, we obtain

$$\kappa_1' + (b-1)\kappa_3\kappa_2 = 0, \tag{32a}$$

$$\kappa_2' + (1-b)\kappa_3\kappa_1 = 0, \tag{32b}$$

$$0 = 0, \tag{32c}$$

which, taking into account the fact that κ_3 is constant, can be integrated,

$$\kappa_1 = \kappa \sin[(1-b)\kappa_3(s-s_0)], \tag{33a}$$

$$\kappa_2 = \kappa \cos[(1-b)\kappa_3(s-s_0)], \tag{33b}$$

where κ and s_0 are integration constants. We see that (33) assumes the form (29), with $\tau = b\kappa_3$. In other words, (33) is a helix.

We conclude that all the solutions with vanishing tension are helices. In the following sections, we consider nonhelical filaments. Therefore, from now on, we assume $\mathbf{F}\neq\mathbf{0}$. This gives a precise sense to the vertical unit vector $\mathbf{e}_Z$.

C. General solution for the local basis

1. Equations for the Euler angles

We now introduce the Euler angles (φ,θ,ψ) for the local basis $(\mathbf{d}_1,\mathbf{d}_2,\mathbf{d}_3)$. The angles φ, θ, and ψ denote, respectively, the precession, nutation, and self-rotation angles (see Fig. 8). In matrix form, the local basis is obtained from the fixed trihedron $(\mathbf{e}_X,\mathbf{e}_Y,\mathbf{e}_Z)$ as follows:

$$(\mathbf{d}_1\mathbf{d}_2\mathbf{d}_3) = (\mathbf{e}_X\mathbf{e}_Y\mathbf{e}_Z)\mathbf{E}, \tag{34}$$

where the general rotation matrix $\mathbf{E}$ reads

$$\mathbf{E} = \begin{pmatrix} \cos\varphi\cos\theta\cos\psi - \sin\varphi\sin\psi & -\cos\varphi\cos\theta\sin\psi - \sin\varphi\cos\psi & \cos\varphi\sin\theta \\ \sin\varphi\cos\theta\cos\psi + \cos\varphi\sin\psi & -\sin\varphi\cos\theta\sin\psi + \cos\varphi\cos\psi & \sin\varphi\sin\theta \\ -\sin\theta\cos\psi & \sin\theta\sin\psi & \cos\theta \end{pmatrix}. \tag{35}$$

From (34) and (35) we extract the fixed vector $\mathbf{e}_Z$ as a function of the local basis,

$$\mathbf{e}_Z = \sin\theta(\sin\psi\mathbf{d}_2 - \cos\psi\mathbf{d}_1) + \cos\theta\mathbf{d}_3. \tag{36}$$

We now express the twist matrix **K** in terms of the rotation matrix **E**. Differentiating (34) yields

$$\frac{\partial}{\partial s}(\mathbf{d}_1\mathbf{d}_2\mathbf{d}_3) = (\mathbf{e}_X\mathbf{e}_Y\mathbf{e}_Z)\frac{\partial}{\partial s}\mathbf{E} = (\mathbf{d}_1\mathbf{d}_2\mathbf{d}_3)\mathbf{E}^{\mathsf{T}}\mathbf{E}', \tag{37}$$

hence, from the definition (7) of the twist matrix, we see that

$$\mathbf{K} = \mathbf{E}^{\mathsf{T}}\mathbf{E}'. \tag{38}$$

Therefore, we can express the twist vector in terms of Euler angles,

$$\kappa_1 = \theta'\sin\psi - \varphi'\sin\theta\cos\psi, \tag{39a}$$

$$\kappa_2 = \theta'\cos\psi + \varphi'\sin\theta\sin\psi, \tag{39b}$$

$$\kappa_3 = \varphi'\cos\theta + \psi'. \tag{39c}$$

Using (36) and (39), we now write the three first integrals in terms of the Euler angles,

$$M_Z = \varphi'[1 + (b-1)\cos^2\theta] + b\psi'\cos\theta, \tag{40a}$$

$$M_3 = b(\varphi'\cos\theta + \psi'), \tag{40b}$$

$$H = \frac{1}{2}\left(\theta'^2 + \varphi'^2\sin^2\theta + \frac{M_3^2}{b}\right) + F\cos\theta. \tag{40c}$$

It is convenient to introduce the following auxiliary constants:

$$h = \frac{1}{F}\left(H - \frac{M_3^2}{2b}\right), \tag{41a}$$

$$\tilde{h} = \frac{1}{F}\left(H - \frac{M_3^2}{2b} + \frac{M_3^2 - M_Z^2}{2}\right), \tag{41b}$$

which are well defined since $\mathbf{F} \neq \mathbf{0}$. We also carry out the following change of variables in Eqs. (40a)–(40c),

$$z = \cos\theta. \tag{42}$$

Note that a solution with constant z corresponds to a helical rod excluded in this discussion, hence z is not constant. We can solve (40a) to (40c) for φ', θ', and z' to obtain the final form of our equations in the Euler angles,

$$\varphi' = \frac{M_Z - M_3 z}{1 - z^2}, \tag{43a}$$

$$\psi' = \left(\frac{1}{b} - 1\right)M_3 + \frac{M_3 - M_Z z}{1 - z^2}, \tag{43b}$$

$$z'^2 = 2F(h - z)(1 - z^2) - (M_Z - M_3 z)^2 \tag{44a}$$

$$\Leftrightarrow z'^2 = 2F(\tilde{h} - z)(1 - z^2) - (M_3 - M_Z z)^2. \tag{44b}$$

2. A better choice of first integrals

The differential Eq. (44a) or (44b) is an identity between z'^2 and a cubic polynomial in z. In order to solve this equation, we must know the roots z_1, z_2, z_3 of this polynomial. Rather than computing the roots in terms of the constants M_Z, M_3, h or $\tilde{h}$, the classical approach consists in considering the roots z_1, z_2, and z_3 as three independent first integrals, and then expressing the constants M_Z, M_3, h, and $\tilde{h}$ as functions of z_1, z_2, and z_3. This is achieved by rewriting the cubic polynomial in (44a) and (44b) as a product of three factors involving z_1, z_2, and z_3,

$$z'^2 = 2F(h - z)(1 - z^2) - (M_Z - M_3 z)^2 \tag{45a}$$

$$= 2F(\tilde{h} - z)(1 - z^2) - (M_3 - M_Z z)^2 \tag{45b}$$

$$= 2F(z - z_1)(z_2 - z)(z_3 - z). \tag{45c}$$

By setting $z = \pm 1$ in (45a) or (45b), the right-hand side assumes a nonpositive value. Furthermore, if we choose F to be positive (which we can always do by defining adequately the vertical unit vector $\mathbf{e}_Z$), we see that the right-hand side of (45a) or (45b) tends to $+\infty$ as $z \to +\infty$. This means that one of the roots (conventionally, z_3) lies in the interval $[1, +\infty[$. Finally, in order to obtain solutions with real values of θ, we require that z'^2 be positive for z ranging in some interval contained in $[-1,1]$. Hence the other two roots of the polynomial must be real and lie between -1 and 1. Conventionally, we choose $z_1 \leqslant z_2$. We conclude that the physical values of our independent constants z_1, z_2 and z_3 must satisfy

$$-1 \leqslant z_1 \leqslant z_2 \leqslant 1 \leqslant z_3. \tag{46}$$

In the previous section, we showed that in order to obtain the scaled form of the Kirchhoff equations in the static case, it was not necessary to perform a complete scaling. Namely, at this level, every variable involved in the equations is a length raised to a given power, and we still have the freedom to choose an arbitrary length unit $[L]$. In the following, it is convenient to choose the length unit to be

$$[L] = \sqrt{\frac{2}{F(z_3 - z_1)}}, \tag{47}$$

which is equivalent to the substitution,

$$F = \frac{2}{z_3 - z_1}. \tag{48}$$

The right-hand side of (48) is well-defined as long as $z_1 \neq z_3$. The condition $z_1 = z_3$ implies $z_1 = z_2 = z_3 = 1$, which corresponds to a straight rod with $z = 1$, a case excluded from this discussion.

Expressions for the constants M_Z, M_3, h, and $\tilde{h}$ in terms of the roots can be obtained by identifying the coefficients of the powers of z in (45a)–(45c), or equivalently, by considering the equalities between the right-hand sides of (45a)–(45c) for three well-chosen values of z. Setting $z = \pm 1$ leads, together with (48), to

$$-(M_Z \mp M_3)^2 = \frac{4}{z_3 - z_1}(\pm 1 - z_1)(z_2 \mp 1)(z_3 \mp 1). \tag{49}$$

Equation (49) gives $M_Z - M_3$ and $M_Z + M_3$ up to sign determination. A third equality of type (45) corresponding to another value of z would lead to an equation involving h or $\tilde{h}$ and could not provide additional knowledge on M_Z or M_3 by themselves. We conclude that given values of z_1, z_2, and z_3 do not yield unique values of M_Z and M_3. Instead they give these two constants with a complete sign indetermination. By performing a mirror reflection in space, we can map any filament with $M_Z+M_3<0$ onto a filament with $M_Z+M_3>0$. Hence, we restrict our analysis to the case $M_Z+M_3\geqslant 0$. Nevertheless, we still have to supply the values of z_1, z_2, and z_3 with extra information, namely, the sign S of M_Z-M_3,

$$S=+ \quad \text{or} \quad S=-. \tag{50}$$

Now, if we set

$$M_+=\frac{M_Z+M_3}{2}\geqslant 0, \tag{51a}$$

$$M_-=\frac{|M_Z-M_3|}{2}, \tag{51b}$$

we have

$$M_Z=M_+ + SM_-, \tag{52a}$$

$$M_3=M_+ - SM_-, \tag{52b}$$

with

$$M_\pm=\sqrt{\frac{(1\pm z_1)(1\pm z_2)(z_3\pm 1)}{z_3-z_1}}. \tag{53}$$

The constants h and $\tilde{h}$ in (44a) and (44b) are obtained from suitable combinations of equalities between the coefficients of z in (45a)–(45c),

$$h=\tfrac{1}{2}[z_1+z_2+z_3-z_1z_2z_3+S(z_3-z_1)M_+M_-], \tag{54a}$$

$$\tilde{h}=\tfrac{1}{2}[z_1+z_2+z_3-z_1z_2z_3-S(z_3-z_1)M_+M_-]. \tag{54b}$$

3. General solution for the Euler angles

The solutions of Eqs. (43a)–(44) involve elliptic functions (see Appendix). With a suitable origin for the arc length s, the solution of Eq. (44) assumes the form,

$$z=z_1+(z_2-z_1)\operatorname{sn}^2(s|k), \tag{55}$$

where the modulus k ranges between 0 and 1 and is given by

$$k^2=\frac{z_2-z_1}{z_3-z_1}. \tag{56}$$

We can use the fact that $z=\cos\theta$ to obtain θ as a function of s. In the generic case where $z_1\neq -1$ and $z_2\neq 1$, the cosine never reaches its extreme values ± 1, and there is a bijective correspondence between z and θ. In this case, we can take $\theta=\arccos z$. In the degenerate cases where $z_1=-1$ or $z_2=1$, we must take care of the behavior of θ as $\cos\theta$ reaches its limiting values. This is achieved by substituting $z=\cos\theta$ in (44) and examining the behavior of θ' around $z=-1$ or $z=1$. The results are as follows:

(a) If $z_1=-1$ and $z_2\neq 1$, θ' has a nonvanishing limit as $z\to -1$, hence the sign of $\theta-\pi$ changes at $z=-1$.

(b) If $z_1 \neq -1$, $z_2 = 1$, and $z_3 \neq 1$, θ' has a nonvanishing limit as $z \to 1$, hence the sign of θ changes at $z = 1$.

(c) If $z_1 \neq -1$, $z_2 = 1$, and $z_3 = 1$, z never reaches its extreme value 1, hence we can take $\theta = \arccos z$.

(d) If $z_1 = -1$, $z_2 = 1$, and $z_3 \neq 1$, θ' has a nonvanishing limit in both cases $z \to \pm 1$, hence θ is a monotonous function of s. In this case, $M_Z = M_3 = 0$, that is, the top behaves like a plane pendulum. This corresponds to planar rods studied by Euler.

(e) If $z_1 = -1$, $z_2 = 1$, and $z_3 = 1$, θ' has a nonvanishing limit as $z \to -1$ and z never reaches its extreme value 1, hence z assumes the value -1 only at the single point $s = 0$, and θ covers the interval $]0,2\pi[$ crossing every value only once. This corresponds to the homoclinic orbit of the plane pendulum.

In order to obtain φ and ψ as functions of s, we must integrate (43a) and (43b), which, using (52a) and (52b), can be rewritten as

$$\varphi' = M_+ \frac{1}{1+z} + S M_- \frac{1}{1-z}, \tag{57a}$$

$$\psi' = \left(\frac{1}{b} - 1\right) M_3 + M_+ \frac{1}{1+z} - S M_- \frac{1}{1-z}. \tag{57b}$$

Next, we define

$$n_\pm = \mp \frac{z_2 - z_1}{1 \pm z_1}. \tag{58}$$

In the cases where $n_\pm$ have a nonvanishing denominator, we can express (57a) and (57b) using (55) as

$$\varphi' = \frac{M_+}{1+z_1} \frac{1}{1 - n_+ \mathrm{sn}^2(s|k)} + S \frac{M_-}{1-z_1} \frac{1}{1 - n_- \mathrm{sn}^2(s|k)}, \tag{59a}$$

$$\psi' = \left(\frac{1}{b} - 1\right) M_3 + \frac{M_+}{1+z_1} \frac{1}{1 - n_+ \mathrm{sn}^2(s|k)} - S \frac{M_-}{1-z_1} \frac{1}{1 - n_- \mathrm{sn}^2(s|k)}. \tag{59b}$$

Notice that we have the freedom to perform global rotations of the local basis around $\mathbf{d}_3$ and of the fixed trihedron around $\mathbf{e}_Z$, in such a way that $\varphi = 0$ and $\psi = 0$ for $s = 0$. Equations (59a) and (59b) can be integrated to yield

$$\varphi = \frac{M_+}{1+z_1} \Pi(s|n_+, k) + S \frac{M_-}{1-z_1} \Pi(s|n_-, k), \tag{60a}$$

$$\psi = \left(\frac{1}{b} - 1\right) M_3 s + \frac{M_+}{1+z_1} \Pi(s|n_+, k) - S \frac{M_-}{1-z_1} \Pi(s|n_-, k), \tag{60b}$$

where Π is the incomplete elliptic integral of the third kind in ''practical'' form, as defined in the Appendix.

In the degenerate cases where n_+ or n_- has a vanishing denominator, the correct limits of (60a) and (60b) are obtained by setting the term involving the ill-defined quantity to zero.

Expressions (60a) and (60b) together with (55) constitute the general solution of the spinning symmetric top problem. We can then use (34) to obtain the nonfixed basis $(\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3)$ as a function of the Euler angles.

4. Particular orbits of the spinning top

The generic orbits of the spinning top are those for which the extremity of $\mathbf{d}_3$ oscillates vertically between two parallels on the unit sphere, while it revolves horizontally, either monotonously as in Figs. 21, 20, 22, and 23, or making loops as in Figs. 16 and 24. The looping orbits arise if $S=-$ and $z_3>(1-z_1z_2)/(z_2-z_1)$. This condition is obtained by allowing φ' to vanish for some value of z.

The degenerate case $S=-$, $z_3=(1-z_1z_2)/(z_2-z_1)$ separates looping orbits from monotonously precessing orbits. This corresponds to trajectories which present turnback points, as shown in Fig. 19. We shall see in the following section that the condition $z_3=(1-z_1z_2)/(z_2-z_1)$ has a clear meaning in terms of rods for both values of S.

The case $z_1=-1$ with $z_2\neq 1$ and $z_3\neq 1$ corresponds to orbits which cross periodically the south pole of the unit sphere, as shown in Fig. 18, while the case $z_2=1$ with $z_1\neq -1$ and $z_3\neq 1$ corresponds to orbits which cross the north pole, as shown in Fig. 17.

The case $z_2=z_3=1$ with arbitrary z_1 represents homoclinic orbits such as those shown in Figs. 13–15.

Next, there are the orbits for which $M_Z=M_3=0$, that is, for which the top behaves like a plane pendulum (the extremity of $\mathbf{d}_3$ is restricted to a vertical grand circle on the unit sphere). They correspond to $z_1=-1$ and either $z_2=1$ or $z_3=1$. The case $z_2=1$ describes an oscillating pendulum, whereas the case $z_3=1$ describes a revolving pendulum. The case $z_2=z_3=1$ corresponds to the homoclinic orbit of the pendulum.

Finally, there are orbits with constant z which we took apart from our preceding analysis, and which correspond to the helical, circular, and straight rods examined in Sec. III B (see Figs. 2–7).

D. Centerline in cylindrical coordinates, curvature, and torsion

1. Polar coordinates, complex curvature, and complex centerline radius

Following Shi and Hearst (1994), we introduce cylindrical coordinates R,Φ,Z for the filament centerline $\mathbf{R}$,

$$\mathbf{R}=R\cos\Phi\,\mathbf{e}_X+R\sin\Phi\,\mathbf{e}_Y+Z\mathbf{e}_Z. \tag{61}$$

Rather than adopting the method of Shi and Hearst to obtain R, Φ, and Z as functions of the arc length, we lead the calculations in a way which highlights the remarkable correspondence between the expressions for the radius R and the Frenet curvature κ, as well as between the polar angle Φ and the angle $\zeta=\int ds(\kappa_3-\tau)$ giving the orientation of the local basis $(\mathbf{d}_1,\mathbf{d}_2,\mathbf{d}_3)$ with respect to the Frenet triad $(\mathbf{n},\mathbf{b},\mathbf{d}_3)$. The computations are quite analogous and can be led in parallel.

First, we introduce the complex centerline radius $\hat{R}$, the complex curvature $\hat{\kappa}$ and the complex horizontal component of the moment $\hat{M}$,

$$\hat{R}=R\exp i\Phi, \tag{62a}$$

$$\hat{\kappa}=\kappa_1+i\kappa_2, \tag{62b}$$

$$\hat{M}=M_X+iM_Y, \tag{62c}$$

where M_X and M_Y are the components of the moment along $\mathbf{e}_X$ and $\mathbf{e}_Y$. Using (34), (35), and (39a), $\hat{\kappa}$ and $\hat{M}$ can be expressed in Euler angles as

$$\hat{\kappa}=-\frac{M_Z-M_3z+iz'}{\pm\sqrt{1-z^2}}\exp -i\psi, \tag{63a}$$

$$\hat{M}=\frac{M_3-M_Z z-iz'}{\pm\sqrt{1-z^2}}\exp i\varphi. \tag{63b}$$

The correct sign to put in front of the root in (63a) and (63b) is the one of $\sin\theta$. The key to obtain the complex centerline $\hat{R}$ is the moment equation (20b). Taking into account the fact that $\mathbf{F}=F\mathbf{e}_Z$ is constant and that $\mathbf{R}=\int ds\,\mathbf{d}_3$, we can integrate both sides of this equation, leading to

$$\mathbf{M}+F\mathbf{R}\times\mathbf{e}_Z=M_Z\mathbf{e}_Z \tag{64}$$

for an appropriate choice of origin in the tridimensional space. Taking the dot product of (64) with $\mathbf{e}_X+i\mathbf{e}_Y$, we obtain

$$\hat{R}=\frac{-i}{F}\hat{M}=\frac{-i}{F}\,\frac{M_3-M_Z z-iz'}{\pm\sqrt{1-z^2}}\exp i\varphi. \tag{65}$$

2. Frenet curvature and centerline radius

Using (63a) and (65), the Frenet curvature κ and the radius R are

$$\kappa^2=|\hat{\kappa}|^2=\frac{(M_Z-M_3z)^2+z'^2}{1-z^2}, \tag{66a}$$

$$R^2=|\hat{R}|^2=\frac{(M_3-M_Zz)^2+z'^2}{F^2(1-z^2)}. \tag{66b}$$

We now substitute the expressions (45a) and (45b) for z'^2, respectively, in (66a) and (66b) to yield the final forms of κ and R. They depend on s only through the variable z,

$$\kappa^2=2F(h-z), \tag{67a}$$

$$R^2=\frac{2}{F}(\tilde{h}-z). \tag{67b}$$

Notice that the left-hand sides of these equations are positive for all z, hence h and $\tilde{h}$ are both greater than or equal to z_2. Also, ignoring the case of straight rods, (67a) and (67b) show that κ and R can only vanish at isolated points where $z=z_2$. A natural question is: for which values of z_1, z_2, z_3 and S do the equalities $h=z_2$ or $\tilde{h}=z_2$ hold? Using expressions (48) for F and (54a)–(54b) for h and $\tilde{h}$, one can easily obtain the following results:

$$h=z_2\Leftrightarrow S=-\quad\text{and}\quad z_3=\frac{1-z_1z_2}{z_2-z_1}, \tag{68a}$$

$$\tilde{h}=z_2\Leftrightarrow S=+\quad\text{and}\quad z_3=\frac{1-z_1z_2}{z_2-z_1}. \tag{68b}$$

Condition (68a) is necessary and sufficient for the curvature to vanish at some isolated points. Remarkably, it is identical to the condition for the orbit of the spinning top to present turnback points. In the same way, (68b) is a necessary and sufficient condition for the radius R to vanish at isolated points.

In order to have a continuous dependence in s for the polar angle Φ across these isolated points where the radius R vanishes, we must change the sign of R. In the same way, the sign of κ must change wherever κ vanishes in order for the angle ζ to be a continuous function of s.

3. Frenet torsion and polar angle of the centerline

The argument of the right-hand side of (65) can be used as an expression for the polar angle Φ. However, this is not convenient. A more tractable expression can be obtained by first computing the derivative of the polar angle from (65),

$$\Phi' = \frac{\partial}{\partial s} \arg \hat{R} = \frac{\partial}{\partial s} \arctan \frac{-z'}{M_3 - M_Z z} + \varphi'. \tag{69}$$

Similarly, the Frenet torsion τ can be computed,

$$\tau = \kappa_3 + \frac{\partial}{\partial s} \arg \hat{\kappa} = \frac{M_3}{b} + \frac{\partial}{\partial s} \arctan \frac{z'}{M_Z - M_3 z} - \psi'. \tag{70}$$

Taking the derivatives in (69) and (70) leads to expressions for Φ' and τ involving z, z'^2, z'', φ', and ψ'. Using the expressions (43a)–(44) to express everything in terms of z only, we obtain

$$\Phi' = \frac{1}{2}\left(M_Z + \frac{M_3 - M_Z \tilde{h}}{\tilde{h} - z} \right), \tag{71a}$$

$$\tau = \frac{1}{2}\left(M_3 + \frac{M_3 h - M_Z}{h - z} \right). \tag{71b}$$

In the case of a nonconstant z, we see from (71) that the condition for a constant Φ is identical to the condition for the rod to be planar ($\tau = 0$); M_Z and M_3 must both vanish. This means that noncircular planar filaments correspond to the case where the spinning top behaves like a plane pendulum ($z_1 = -1$ and either $z_2 = 1$ or $z_3 = 1$).

For nonplanar rods, we define

$$n = \frac{z_2 - z_1}{h - z_1}, \tag{72a}$$

$$\tilde{n} = \frac{z_2 - z_1}{\tilde{h} - z_1}. \tag{72b}$$

We can then integrate (71a) and (71b) to yield

$$\Phi = \frac{1}{2}\left(M_Z s + \frac{M_3 - M_Z \tilde{h}}{\tilde{h} - z_1} \Pi(s|\tilde{n},k) \right) - \frac{\pi}{2}, \tag{73a}$$

$$\zeta = \frac{M_3}{b} s - \frac{1}{2}\left(M_3 s + \frac{M_3 h - M_Z}{h - z_1} \Pi(s|n,k) \right) - \frac{\pi}{2}. \tag{73b}$$

The integration constant $-(\pi/2)$ in (73a) has been determined from the complex radius $\hat{R}$ (65) in the limit $s \to 0$. The integration constant $-(\pi/2)$ in (73b) is determined from the fact that $\mathbf{d}_1$ and the binormal $\mathbf{b}$ are opposite when $\psi = 0$.

4. Vertical cartesian coordinate of the centerline

To complete the filament description, we need an expression for Z, the vertical coordinate of the centerline. It is given by

$$Z' = z, \tag{74}$$

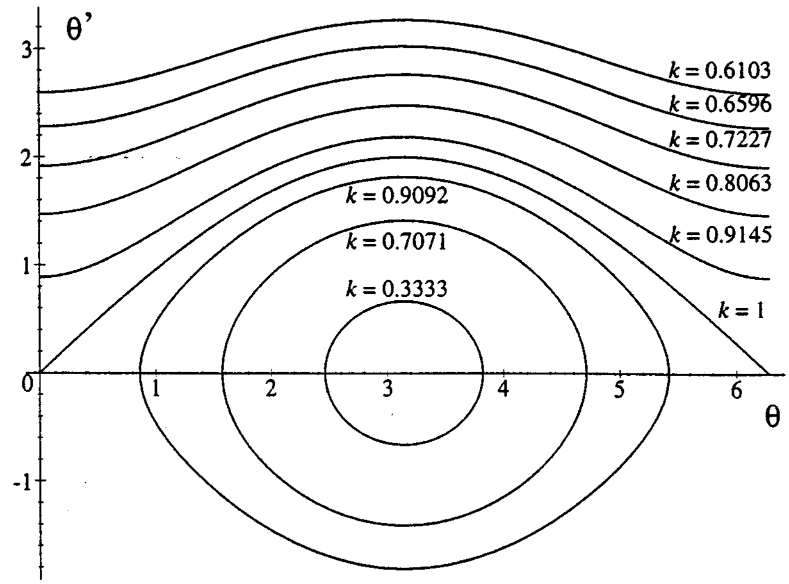

FIG. 9. Phase portrait for the plane pendulum. The closed curves represent oscillating orbits, while the open curves represent revolving orbits. The homoclinic orbit corresponds to $k=1$.

which, using (55) and (56), can be written as

$$Z'=z_3-(z_3-z_1)[1-k^2\mathrm{sn}^2(s|k)]. \tag{75}$$

The last expression is readily integrated to yield, for an appropriate choice of origin of the vertical coordinate,

$$Z=z_3s-(z_3-z_1)E(s|k), \tag{76}$$

where E is the incomplete elliptic integral of the second kind in ''practical'' form, as defined in the Appendix.

For very large z_3, Z is expressed as a difference between two large quantities, although Z itself is finite for $z_3\to\infty$. Hence, (76) could be tricky to handle numerically for large z_3.

Finally, we note that the expression for the radius R is bounded while, in general, the expression for Z is not. As a consequence, the vector $\mathbf{d}_3$, averaged over all s, has no component in the $(\mathbf{e}_X,\mathbf{e}_Y)$ plane. This means that, on average, a spinning top is vertical whatever the constants of the motion are.

E. Filament shapes

1. Planar shapes

This section is dedicated to planar filaments other than the circular and straight ones. This is essentially a modern restatement of Euler's results. As seen in the previous section, the noncircular and nonstraight planar filaments correspond to values of the constants z_1, z_2, and z_3 for which the top behaves like a plane pendulum. In this case, the sign variable S is meaningless. There are two one-parameter families of planar solutions. The first one is obtained by setting $z_1=-1$ and $z_3=1$ and keeping z_2 arbitrary; it corresponds to oscillating orbits of the pendulum. The second one is obtained by setting $z_1=-1$ and $z_2=1$ and keeping z_3 arbitrary; it corresponds to revolving orbits of the pendulum. In both cases, it is convenient to adopt the modulus k as the arbitrary parameter.

The two families of pendulum orbits are displayed on the phase portrait shown in Fig. 9. Each

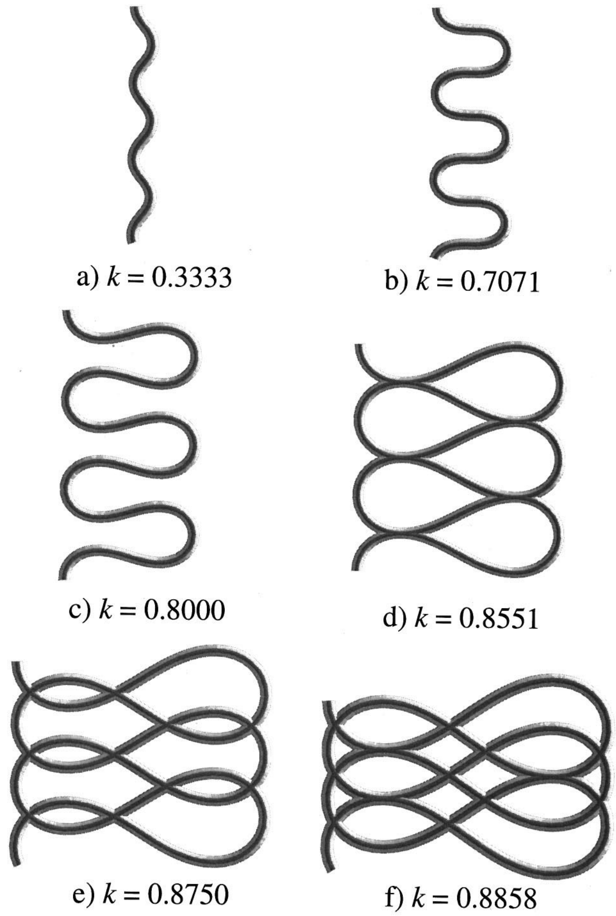

FIG. 10. Planar filaments corresponding to low amplitude oscillating orbits of the plane pendulum.

curve represents a different orbit with a given value of k in the (θ,θ') space. The closed orbits correspond to oscillating states of the pendulum, while the open orbits correspond to revolving states of the pendulum. The corresponding filament shapes are represented in Figs. 10–12. Since the scaling defined in Eq. (47) depends on k through z_3-z_1, so does the length unit for the filament, or the time unit for the pendulum. This is inadequate to draw a phase portrait, hence, in Fig. 9, we have exceptionally chosen the scaled force unit $[L]^{-2}$ to be F. This is identical to the scaling defined in (47) in the oscillating case.

We introduce the following notations for the complete elliptic integrals of the first and second kinds defined in the Appendix:

$$K=K(k), \tag{77a}$$

$$E=E(k). \tag{77b}$$

a. Oscillating orbits of the pendulum: In the case $z_3=1$, the expressions for the relevant variables *z, R,* and *Z,* reduce to

$$z=2k^2\mathrm{sn}^2(s|k)-1, \tag{78a}$$

$$R=2k\mathrm{cn}(s|k), \tag{78b}$$

$$Z=s-2E(s|k). \tag{78c}$$

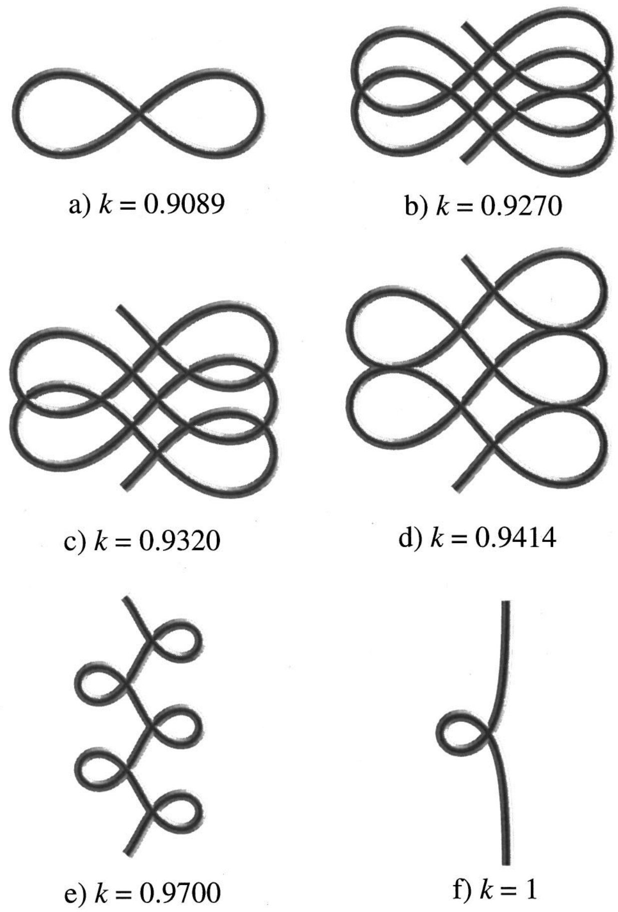

FIG. 11. Planar filaments corresponding to high amplitude oscillating orbits of the plane pendulum. The homoclinic orbit is reached in the limit $k=1$.

In the following, we define

$$f = F^A\left(\frac{1}{\sqrt{2}k}\middle| k\right), \tag{79a}$$

$$e = E^A\left(\frac{1}{\sqrt{2}k}\middle| k\right), \tag{79b}$$

where F^A and E^A are the incomplete elliptic integrals of the first and second kinds in algebraic form (see Appendix).

The extrema of Z are given by the condition $z=0$, which is satisfied for

$$\mathrm{sn}^2(s|k) = \frac{1}{2k^2} \Leftrightarrow s = \pm f + 2mK, \tag{80}$$

where m is an arbitrary integer. We see from (79a) that f is real only if $k^2 \geqslant \frac{1}{2}$. Hence, for $k^2 < \frac{1}{2}$, Z is a decreasing function of s [see Fig. 10(a)]. The limiting case $k^2 = \frac{1}{2}$ corresponds to an oscillating pendulum which has just enough energy to reach the horizontal position, as shown in Fig. 10(b). Above this critical value of k^2, Z is not monotonous, and for large enough k, the

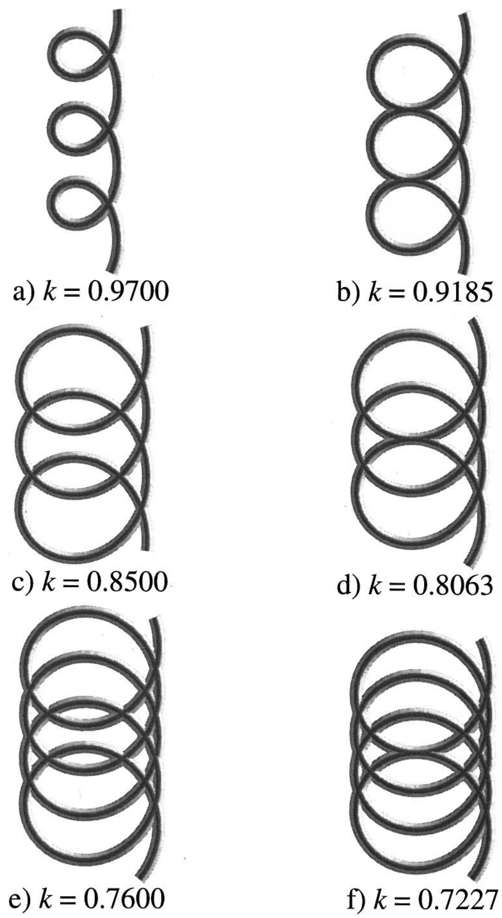

FIG. 12. Planar filaments corresponding to revolving orbits of the plane pendulum.

filaments can have points of self-tangency or self-intersection [see Figs. 10(d)–10(f)]. The cases of self-tangency are obtained by imposing the equality of Z at different points where z vanishes. This leads to the condition

$$f-2e+2m(2E-K)=0, \tag{81}$$

where m is a positive integer which denotes the number of pendulum oscillations before a self-tangency occurs. The solutions k_m of Eq. (81) for the lowest values of m are given in the first column of Table II.

Setting $m=\infty$ in Eq. (81) yields

TABLE II. Values of k corresponding to self-tangency.

$k_1=0.8551$	$k_{-1}=0.9414$	$\tilde{k}_1=0.9145$
$k_2=0.8858$	$k_{-2}=0.9270$	$\tilde{k}_2=0.8063$
$k_3=0.8942$	$k_{-3}=0.9214$	$\tilde{k}_3=0.7227$
$k_4=0.8981$	$k_{-4}=0.9185$	$\tilde{k}_4=0.6596$
$k_5=0.9004$	$k_{-5}=0.9167$	$\tilde{k}_5=0.6103$
...	...	...
$k_\infty=0.9089$	$k_{-\infty}=0.9089$	$\tilde{k}_\infty=0$

$$k_\infty = 0.9089, \tag{82}$$

for which the filament shape is the lemniscate represented in Fig. 11(a). Every value of $k < k_\infty$ corresponds to a solution Z which, on average, is a decreasing function of s, while for $k > k_\infty$, Z is, on average, an increasing function of s. The increasing solutions can also present points of intersection and self-tangency for given values of k [see Figs. 11(b)–11(e)]. The condition of self-tangency in the case $k > k_\infty$ also assumes the form (81), provided that we set m to be negative. The positive integer $-m$ counts the number of pendulum oscillations before a self-tangency occurs. The solutions k_m of Eq. (81) for the lowest values of $-m$ are given in the second column of Table II.

b. Homoclinic orbit of the pendulum: The homoclinic orbit is obtained by taking the limit $k \to 1$ in either the oscillating case or the revolving case. It is shown in Fig. 11(f). The expressions for z, R, and Z then reduce to

$$z = 1 - 2\,\mathrm{sech}^2\, s, \tag{83a}$$

$$R = 2\,\mathrm{sech}\, s, \tag{83b}$$

$$Z = s - 2\tanh s. \tag{83c}$$

c. Revolving orbits of the pendulum: In this case, we set $z_1 = -1$ and $z_2 = 1$. The expressions for θ, R, and Z read

$$\theta = \pi - 2\,\mathrm{am}(s|k), \tag{84a}$$

$$\rho = 2k^{-2}\mathrm{dn}(s|k), \tag{84b}$$

$$Z = (2k^{-2} - 1)s - 2k^{-2}E(s|k). \tag{84c}$$

Next, we define

$$\tilde{f} = F\left(\frac{1}{\sqrt{2}}\middle| k\right), \tag{85a}$$

$$\tilde{e} = E\left(\frac{1}{\sqrt{2}}\middle| k\right). \tag{85b}$$

The extrema of Z correspond to the values of s for which θ is an integer multiple of π, hence $\mathrm{am}(s|k)$ is an integer multiple of $\pi/2$, or

$$s = \pm 2\tilde{f} + 2mK, \tag{86}$$

where m is an integer. Notice that $\tilde{f}$ and $\tilde{e}$ are real for all k. As in the case of oscillating orbits, we can find a condition for the existence of points of self-tangency,

$$(2-k^2)\tilde{f} - 2\tilde{e} + m[(2-k^2)K - 2E] = 0, \tag{87}$$

where the positive integer m counts the number of pendulum revolutions before a self-tangency occurs. The solutions $\tilde{k}_m$ of Eq. (87) for the lowest values of m are given in the third column of Table II. The sequence $(\tilde{k}_m)$ has a vanishing limit for $m \to \infty$ and, in the limit $k \to 0$, the filament is a vertical ring. Some filaments corresponding to revolving orbits are shown in Fig. 12.

2. Nonplanar localizing solutions

The homoclinic orbits constitute a one-parameter family of solutions which are obtained by setting $z_2 = z_3 = 1$ while z_1 is kept arbitrary. Here again, the sign parameter S is meaningless. In terms of rods, these orbits correspond to the localizing solutions studied by Coyne (1990).[1] They connect continuously the straight state $(z_1 = 1)$ to the planar loop $(z_1 = -1)$. These solutions have constant torsion,

$$\tau = \sqrt{\frac{1+z_1}{1-z_1}}. \tag{88}$$

We see that τ is bijective in z_1 and can assume any non-negative value, from 0 in the case $z_1 = -1$ (plane pendulum) to infinity in the limit $z_1 \rightarrow 1$. Actually, for fixed tension, the torsion cannot be arbitrarily large. One must keep in mind that the parameter τ used here is the scaled torsion and that the scaling defined in (47) in turn depends on τ through z_1. As a consequence, the unscaled torsion $\tilde{\tau}$ is not simply proportional to τ, but instead is given by $\tilde{\tau}^2 = [\tilde{F}/EI_1(1+\tau^{-2})]$, where $\tilde{F}$ is the unscaled tension. Hence, the torsion has actually an upper bound proportional to the square root of the tension. In the following, we adopt the torsion τ rather than z_1 as the arbitrary parameter.

The top variables φ, z, and ψ assume the form

$$\varphi = \arctan\left(\frac{1}{\tau}\tanh s\right) + \tau s, \tag{89a}$$

$$z = 1 - \frac{2}{1+\tau^2}\operatorname{sech}^2 s, \tag{89b}$$

$$\psi = \arctan\left(\frac{1}{\tau}\tanh s\right) + \left(3 - \frac{2}{b}\right)\tau s. \tag{89c}$$

Notice that the boundary values $\frac{2}{3}$ and 1 for the variable b in the case of circular cross section take a new sense in view of expressions (89). The value $b = \frac{2}{3}$ makes the self-rotation angle ψ bounded, while $b = 1$ is the value for which $\varphi = \psi$.

The centerline variables R, Φ, and Z read

$$R = \frac{2}{1+\tau^2}\operatorname{sech} s, \tag{90a}$$

$$\Phi = \tau s - \frac{\pi}{2}, \tag{90b}$$

$$Z = s - \frac{2}{1+\tau^2}\tanh s. \tag{90c}$$

Three typical nonplanar $(\tau \neq 0)$ homoclinic orbits of the spinning top together with the corresponding filament shapes are shown in Figs. 13–15.

3. Generic filament shapes

In the most generic case, the spinning top oscillates between two parallels z_1 and z_2, while it precesses either monotonously or with backward-and-forward motion. The corresponding filament centerline $\mathbf{R}$ behaves, on average, as a helix around which it is wound.

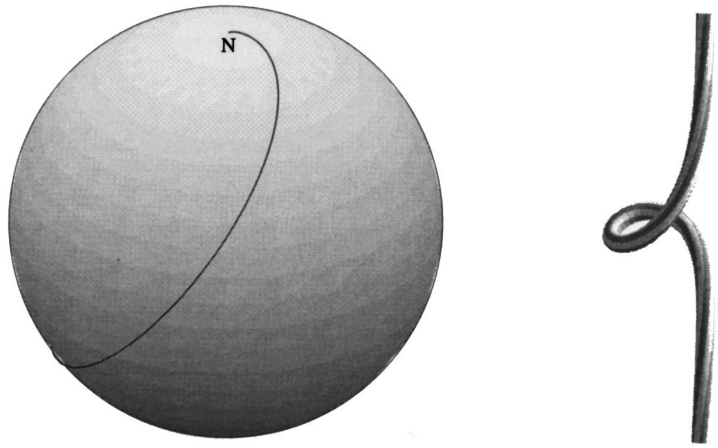

FIG. 13. A locally buckled filament with low torsion and high curvature ($\tau = \frac{1}{2}$).

a. Average helix: We define the mean helical centerline $\langle \mathbf{R} \rangle$ in the following way: We introduce cylindrical coordinates $\langle R \rangle, \langle \Phi \rangle, \langle Z \rangle$ for the helix $\langle \mathbf{R} \rangle$, and we take the mean vertical coordinate $\langle Z \rangle$ to be a linear function of s, namely,

$$\langle Z \rangle = \langle z \rangle s, \tag{91}$$

where $\langle z \rangle$ is the average of z over a period, $2K(k)$. Using (76), we have

$$\langle z \rangle = \frac{1}{2K(k)} \int_0^{2K(k)} z ds = z_3 - (z_3 - z_1) \frac{E(k)}{K(k)}. \tag{92}$$

Now, consider the expression (65) for the complex centerline radius $\hat{R}$. The polar angle Φ being the argument of $\hat{R}$, we have

$$\Phi = \arg \hat{\rho} + \varphi - \frac{\pi}{2}, \tag{93}$$

with

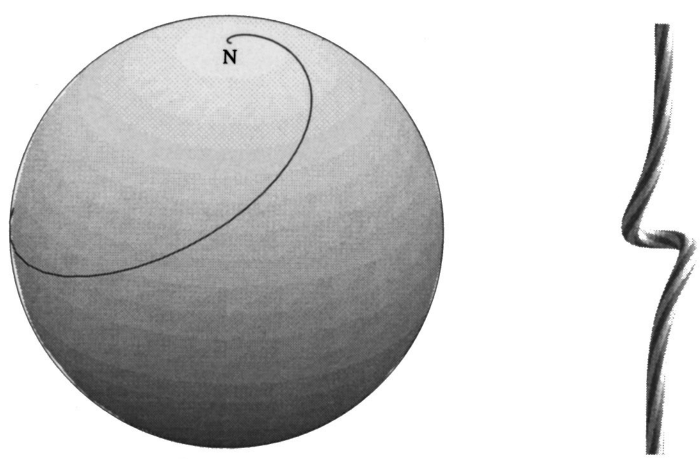

FIG. 14. A locally buckled filament with intermediate torsion and curvature ($\tau = 1$).

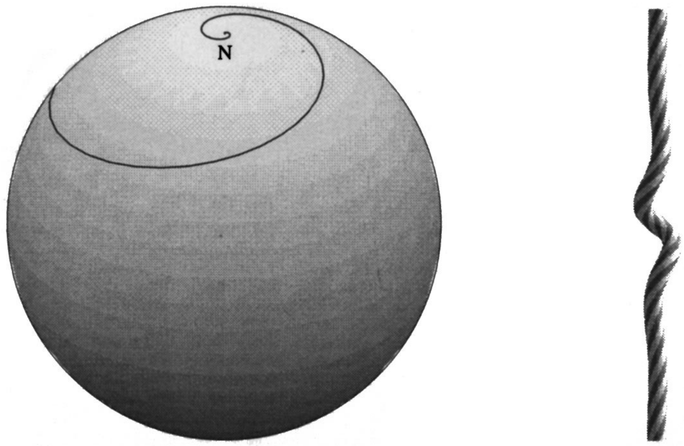

FIG. 15. A locally buckled filament with high torsion and low curvature ($\tau=2$).

$$\hat{\rho}=\frac{1}{F}\,\frac{M_3-M_Z z-iz'}{\pm\sqrt{1-z^2}}. \tag{94}$$

The function $\hat{\rho}$ depends on s through z and z' only, hence it describes a closed curve in the complex plane. If $\hat{\rho}$ does not vanish anywhere, as s increases by a period $2K(k)$, the argument of $\hat{\rho}$ increases by $2\pi m$, where m is some integer. Moreover, the real and imaginary parts of the numerator in (94) being decreasing functions of z and z', respectively, the curve defined by $\hat{\rho}$ has no self-crossing (and is parameterized clockwise by s). Hence, it turns around the origin at most once clockwise over a period $2K(k)$, so that m is restricted to the values 0 and -1. The domain in the (z_1,z_2,z_3,S) space where either value of m holds is delimited by the condition (68b) ensuring the existence of points where $\hat{\rho}$ vanishes. We find that $m=-1$ if $S=+$ and $z_3>(1-z_1z_2)/(z_2-z_1)$, and $m=0$ otherwise. As a consequence, we have

$$\varphi(s+2K(k))-\varphi(s)=\Phi(s+2K(k))-\Phi(s)+\begin{cases}2\pi & \text{if } S=+ \text{ and } z_3>\dfrac{1-z_1z_2}{z_2-z_1},\\[2ex] 0 & \text{if } S=- \text{ or } z_3<\dfrac{1-z_1z_2}{z_2-z_1}.\end{cases} \tag{95}$$

Hence, over a period of z, the precession angle φ of the top covers either the same angular distance as the polar angle Φ of the rod, or the same angular distance plus one complete revolution.

Now, using (60a) and (73a), we define the mean angular velocities $\langle\varphi'\rangle$ and $\langle\Phi'\rangle$ for the corresponding angles φ and Φ as

$$\langle\varphi'\rangle=\frac{1}{2K(k)}\int_0^{2K(k)}ds\,\varphi'=\frac{M_+}{1+z_1}\,\frac{\Pi(n_+,k)}{K(k)}+S\,\frac{M_-}{1-z_1}\,\frac{\Pi(n_-,k)}{K(k)}, \tag{96a}$$

$$\langle\Phi'\rangle=\frac{1}{2K(k)}\int_0^{2K(k)}ds\,\Phi'=\frac{1}{2}\left(M_Z+\frac{M_3-M_Z\tilde{h}}{\tilde{h}-z_1}\,\frac{\Pi(\tilde{n},k)}{K(k)}\right). \tag{96b}$$

As a consequence, these two quantities either are equal or differ from $\pi/K(k)$. The question arises then: which mean angular velocity, $\langle\varphi'\rangle$ or $\langle\Phi'\rangle$, should we use to define the polar angle $\langle\Phi\rangle$ of our average helix? The most natural choice seems to take $\langle\Phi\rangle=\langle\Phi'\rangle s$. However, expression (96b) is not continuous through the boundary (68b), making the definition of the mean polar angle $\langle\Phi\rangle$ ambiguous at that point. Therefore, we define

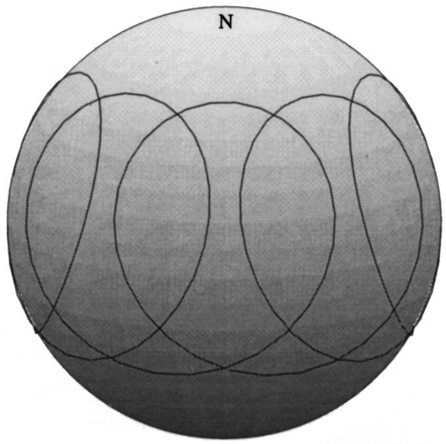

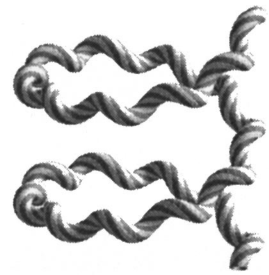

FIG. 16. A slowly precessing, nearly circular top orbit corresponds to a supercoiled helix ($z_1=-\frac{1}{2}$, $z_2=\frac{3}{4}$, $z_3=4$, $S=-$).

$$\langle\Phi\rangle=\langle\varphi'\rangle s-\frac{\pi}{2}\quad\text{for } S=-, \tag{97a}$$

$$\langle\Phi\rangle=\left(\langle\varphi'\rangle-\frac{\pi}{K(k)}\right)s-\frac{\pi}{2}\quad\text{for } S=+. \tag{97b}$$

These expressions reduce to $\langle\Phi'\rangle s$ for sufficiently large z_3, and are continuous across the boundary (68b). As we shall see, this definition is the most adequate in the case where the filament shape is a supercoiled helix (this case has a great importance in biochemical applications, in particular for DNA supercoiling.[31,40,41]) The supercoiled helices are defined and discussed below. It remains now to define the mean helix radius $\langle R\rangle$. A definition consistent with (97) is

$$\langle R\rangle=\frac{1}{2K(k)}\int_0^{2K(k)}ds\hat{R}\exp-i\langle\Phi\rangle. \tag{98}$$

Therefore, the generic filament behaves on average like an helical filament.

b. Supercoiled helices: A supercoiled helix is a curve which looks like a helix on short length scales, with the central axis itself shaped like a helix on large length scales. The condition for the centerline $\mathbf{R}$ to be a supercoiled helix, in terms of spinning tops, is that the vector $\mathbf{d}_3$ describes slowly precessing, nearly circular, oblique loops on the unit sphere. Two such examples are shown in Figs. 16 and 17. If any of these three conditions (slow precession, near-circularity, and oblicity of the top orbit) is not fulfilled, the filament shape will not look like a supercoiled helix, but rather like a deformed helix, as shown in Figs. 18–21.

The supercoiled helices can be studied systematically as solutions close to the oblique circular orbits of the spinning top. There are two ways to obtain circular orbits. The first one consists in setting $z_1=z_2$, in which case the orbit is horizontal, and not oblique. The second one is to take the limit $z_3\to\infty$. Indeed, in view of (48), the tension vanishes as z_3 grows without bound, and, as we mentioned in Sec. III B, every filament with null tension is a helix, hence every top orbit with $F=0$ is a circle. Furthermore, these asymptotic circular orbits can be arbitrarily oriented since the vertical direction $\mathbf{e}_Z$ cannot be distinguished from the other directions in the limit $F=0$. As a consequence, we can redefine a supercoiled helix as a solution with a large (''close to infinity'') value of z_3. In practice, however, z_3 does not need to be very large, so that the threshold value appearing in (95) can be reached with the centerline $\mathbf{R}$ still reasonably looking like a supercoiled helix.

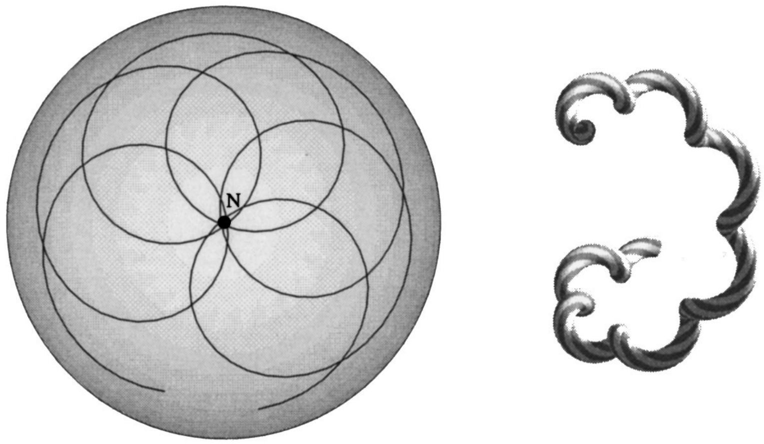

FIG. 17. A top orbit crossing periodically the north pole. The corresponding filament has a periodically vertical tangent ($z_1=\frac{1}{2}$, $z_2=1$, $z_3=3$).

In the limit of large z_3, the vertical axis joining the poles of the unit sphere is interior to precessing circular orbit in the case $S=+$, and exterior to it in the case $S=-$. As a consequence, one passes continuously from the supercoiled helices with $S=-$ to the supercoiled helices with $S=+$ by enlarging the slowly precessing orbit so that it passes through one of the poles, as in Fig. 17.

We mentioned above that our definitions of the average helix coordinates $\langle R\rangle,\langle\Phi\rangle,\langle Z\rangle$ are well adapted to supercoiled helices, in the sense that they describe consistently the large scale helical behavior of the axis around which the centerline is wound. Namely, in the limit $z_3\to\infty$, the supercoiled helix tends towards an (infinitely remote) ordinary helix, with a straight axis given by the limit of the average helix $\langle\mathbf{R}\rangle$.

c. Deformed helices: In many cases, although the centerline $\mathbf{R}$ winds around the average helix $\langle\mathbf{R}\rangle$, it does not quite look like a supercoiled helix. This happens if the criterion discussed above is not fulfilled, which can result in the following:

(1) The short scale spatial period $2K(k)$ and the large scale spatial period $2\pi/\langle\Phi\rangle'$ can be too close to each other, in which case the two orders of helicity cannot be clearly distinguished. Such a situation is shown in Fig. 18.

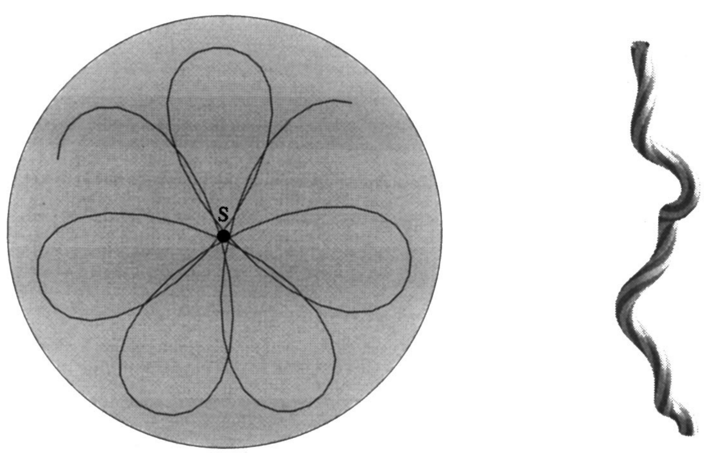

FIG. 18. A top orbit crossing periodically the south pole. The corresponding filament has a periodically vertical tangent ($z_1=-1$, $z_2=-\frac{1}{2}$, $z_3=\frac{14}{10}$, $S=-$).

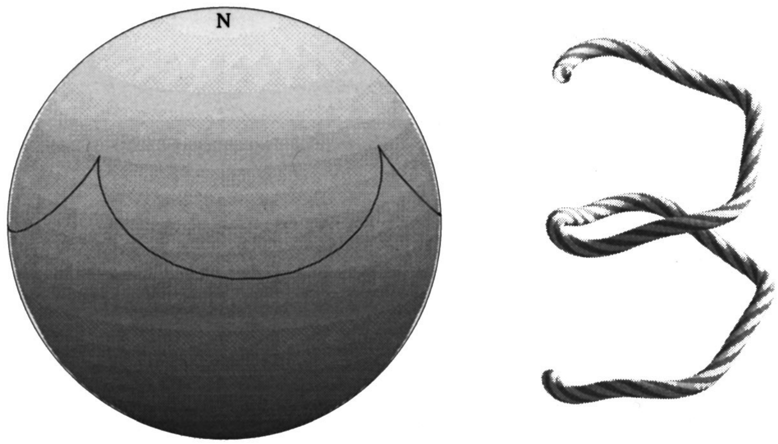

FIG. 19. A top orbit presenting turn-back points corresponds to a filament with periodically vanishing curvature ($z_1=0$, $z_2=\frac{1}{2}$, $z_3=2$, $S=$—).

(2) If the top orbit is too far away from a circle, the short scale pattern does not look like a helix. As an example, Fig. 19 shows a large-scale helix with curvature varying on short scale.
(3) For some values of the constants, the top orbit can be periodic although quite different from a circle. In this case, the filament shape is periodic in space, but different from a helix (see the ''oblique helix'' in Fig. 20).

In addition, there is the possibility for the amplitude of the short scale pattern to be large enough for two consecutive turns of the large-scale helix to overlap each other. In this case, the topology of the solution is different from the topology of an ordinary helix. Such a ''knotted helix'' is shown in Fig. 21.

4. Bounded and closed filament shapes

The filament shapes discussed so far are all unbounded in space, except for the twisted ring shown in Fig. 5 and the lemniscate shown in Fig. 11(a). In general, bounded shapes are obtained by imposing the coordinate Z to be a periodic function of s, that is, $Z=0$ for $s=2K(k)$. Using (76), the condition for boundedness reads

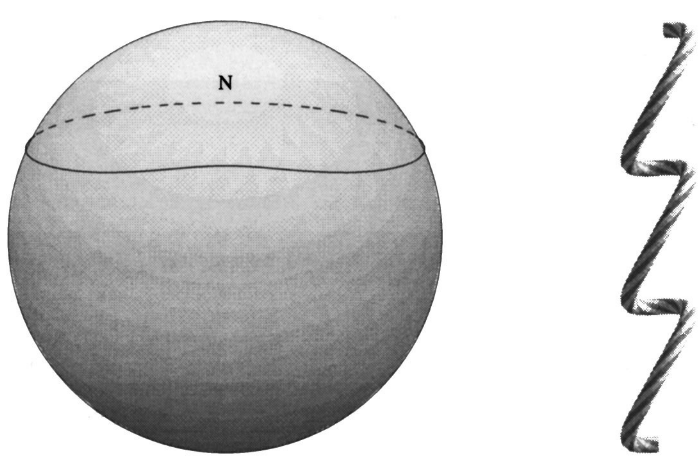

FIG. 20. A periodic top orbit results in a spatially periodic filament shape ($z_1=-0.7822$, $z_2=0.8782$, $z_3=1$, $S=$—).

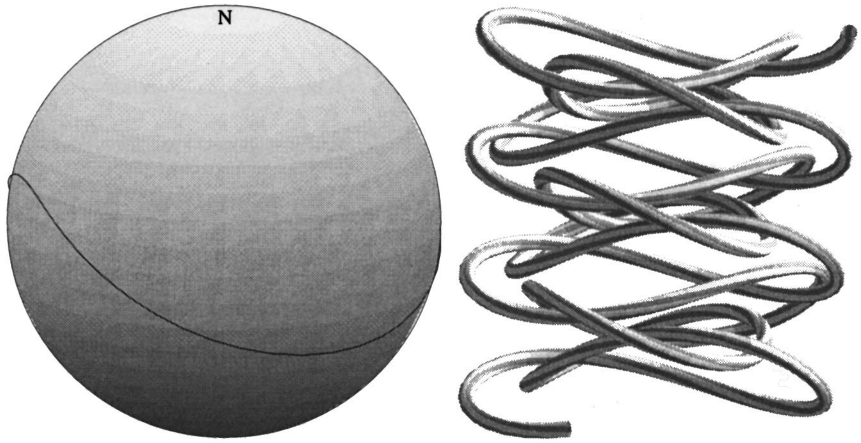

FIG. 21. An apparently simple top orbit resulting in a surprisingly complex ''knotted helix'' ($z_1=-0.4152$, $z_2=0.2800$, $z_3=1.026$, $S=$—).

$$z_3K(k)-(z_3-z_1)E(k)=0. \tag{99}$$

The condition for the space curve **R** to be closed is obtained by requiring, in addition, that the periods of the variables Z and Φ be in a ratio of integer numbers. Namely, using (73a),

$$M_ZK(k)+\frac{M_3-M_Z\tilde{h}}{\tilde{h}-z_1}\Pi(\tilde{n},k)=2\pi\frac{m_\Phi}{m_Z}, \tag{100}$$

where the integers m_Φ and m_Z denote, respectively, the number of periods of the variables Z and Φ in a complete covering of the centerline **R**.

Finally, one can impose the ribbon associated to the filament to be closed by requiring that the periods of the variables Z and ζ be in a ratio of integer numbers, or, using (73b),

$$\frac{2M_3}{b}K(k)-\left(M_3K(k)+\frac{M_3h-M_Z}{h-z_1}\Pi(n,k)\right)=2\pi\frac{m_\zeta}{m_Z}, \tag{101}$$

where m_ζ denotes the number of periods of ζ in a complete covering of the centerline **R**. The condition (101) is useful if one has, for instance, an octagonal cross section, in which case m_ζ must be set to an integer multiple of $\frac{1}{8}$, in order for the octagons at $s=0$ and at $s=2m_ZK(k)$ to match each other.

This together makes three conditions from which the constants z_1, z_2, and z_3 can be determined. However, the conditions (100) and (101) are not very tractable, because the left-hand side of (100) is proportional to the average angular velocity $\langle\Phi'\rangle$ which, as we mentioned, is not a continuous function of the constants z_1, z_2, and z_3. This holds too for the left-hand side of (101). As a consequence, numerical root solvers are inefficient in solving the systems (99)–(101). In practice, it is well advised to replace Eqs. (100) and (101) by equivalent conditions on the Euler angles φ and ψ. Remember that, over a period of Z, the angles φ and Φ cover angular distances which differ by integer multiples of 2π. A similar relation holds for the angles ψ and ζ. The conditions on φ and ψ analogous to (100) and (101) read

$$\frac{M_+}{1+z_1}\Pi(n_+,k)+S\frac{M_-}{1-z_1}\Pi(n_-,k)=\pi\frac{m_\varphi}{m_Z}, \tag{102a}$$

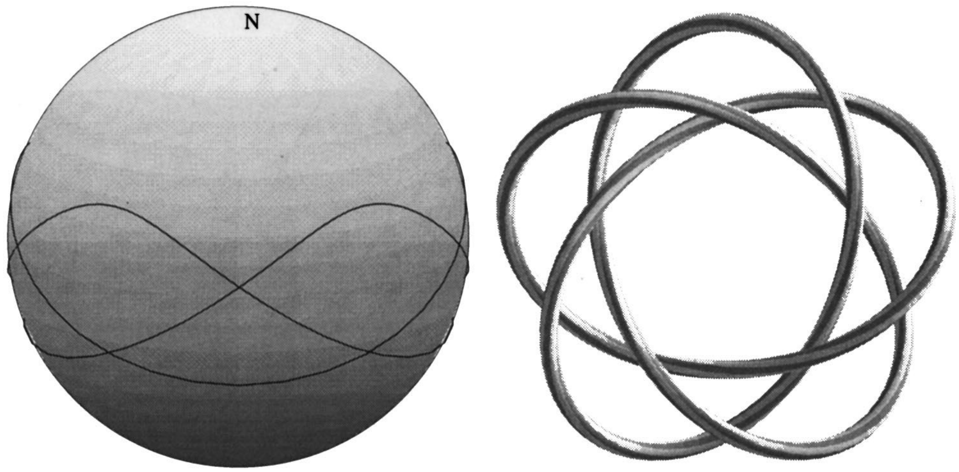

FIG. 22. A torus knot with $m_Z=5$, $m_\Phi=3$, and $m_\zeta=1$ ($z_1=-0.4152$, $z_2=0.3446$, $z_3=1.026$, $S=-$).

$$\left(\frac{1}{b}-1\right)M_3K(k)+\frac{M_+}{1+z_1}\Pi(n_+,k)-S\frac{M_-}{1-z_1}\Pi(n_-,k)=\pi\frac{m_\psi}{m_Z}, \tag{102b}$$

where m_φ and m_ψ differ from m_Φ and m_ζ, respectively, by an integer.

Examples of closed filaments are displayed in Figs. 22–24. Figures 22 and 23 show torus knots, that is, knotted curves which are topologically equivalent to closed curves lying on a torus. Figure 24 shows a supercoiled ring (notice the large value of z_3).

IV. CONCLUSIONS

In this paper we have shown how to classify the shapes of Kirchhoff filaments based on the geometry of the spinning top solutions. To do so, we have pushed the Kirchhoff analogy to its extreme and systematically obtained interesting properties of filaments based on the corresponding solutions of the Euler equations. We showed that the solutions of Kirchhoff equations can be extremely varied and that many interesting cases can be distinguished. In particular, we found explicit conditions on the boundary values for filaments to have points of self-tangency and multiple self-intersection. We also studied the case where filaments have points of vanishing

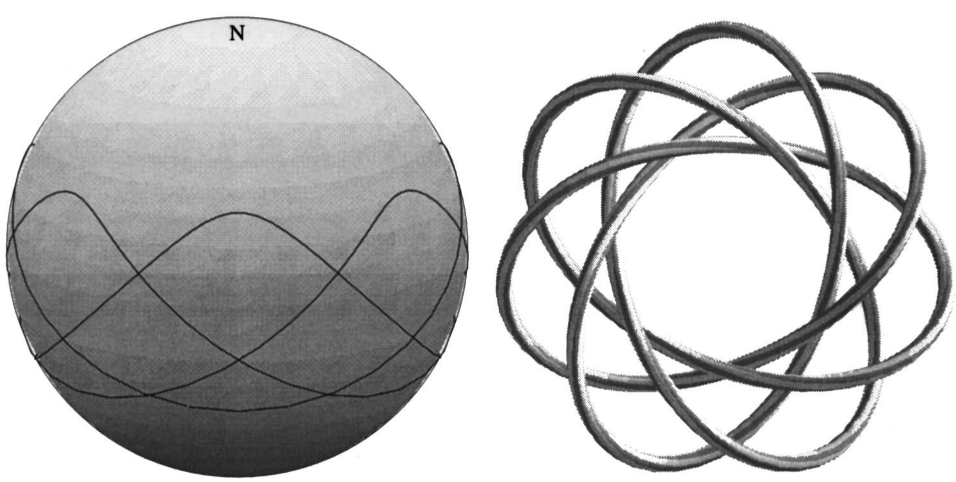

FIG. 23. A torus knot with $m_Z=7$, $m_\Phi=4$, and $m_\zeta=1$ ($z_1=-0.4997$, $z_2=0.4013$, $z_3=1.037$, $S=-$).

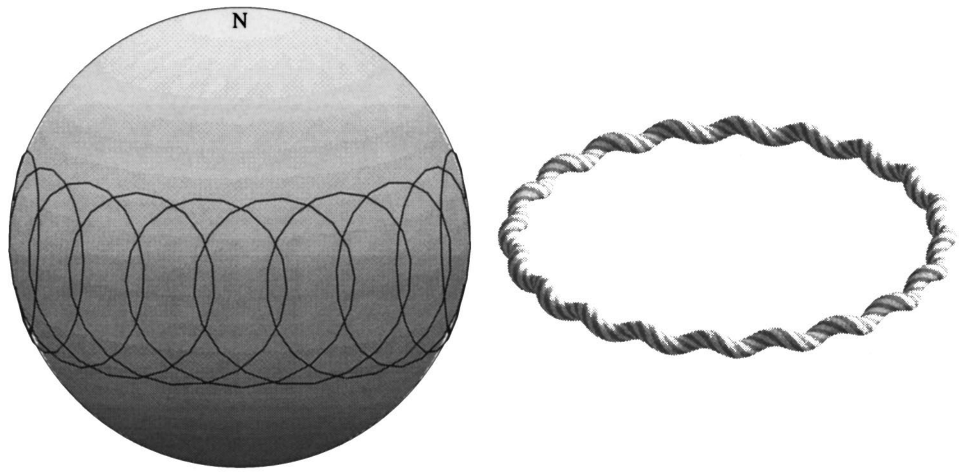

FIG. 24. A supercoiled ring with $m_Z=20$, $m_\Phi=1$, and $m_\zeta=6$ ($z_1=-0.4269$, $z_2=0.4171$, $z_3=9.084$, $S=$—).

curvature and show that they correspond to orbits of spinning tops with turnback points. We gave a complete description of localizing solutions, that is solutions which are homoclinic in the curvature-torsion space; these filaments are in space asymptotic to a straight line. In the same way, we found conditions to obtain filaments which have the topology of torus knots, that is bounded and periodic filaments (in the physical space). Finally, we studied the behavior of generic filaments and show that on long length scales they always behave like helical filaments.

Some of the particular solutions presented here have been obtained in various places by different authors. In this paper we have stressed on the geometry of these solutions and presented them in a unifying way based on the familiar framework of the spinning top.

The solutions of the Kirchhoff equations for rods with circular cross sections are often used as a first guess to study numerically physical filaments with different properties (e.g., noncircular cross sections, intrinsic curvature or torsion,...). We hope that the explicit solutions given in this paper together with their geometric classification will be useful in this context.

ACKNOWLEDGMENTS

M.N. acknowledges support from the Fonds National de la Recherche Scientifique (F.N.R.S). M.N. is an aspirant at the FNRS. This work is supported by NATO-CRG 97/037.

APPENDIX: ELLIPTIC FUNCTIONS

Remark: In the following, the constant k is called the modulus of the elliptic functions and ranges between 0 and 1, while the constant n is called the characteristic and is a real number less than 1.

1. Elliptic integrals

A. Incomplete elliptic integrals in standard form

The incomplete elliptic integrals of the first, second, and third kind in standard form are, respectively, defined by

$$F^S(\Phi|k)=\int_0^\Phi \frac{d\Phi'}{\sqrt{1-k^2\sin^2\Phi'}}, \tag{A1a}$$

$$E^S(\Phi|k)=\int_0^\Phi d\Phi'\sqrt{1-k^2\sin^2\Phi'}, \tag{A1b}$$

$$\Pi^S(\Phi|n,k)=\int_0^{\Phi}\frac{d\Phi'}{(1-n\sin^2\Phi')\sqrt{1-k^2\sin^2\Phi'}}. \tag{A1c}$$

B. Incomplete elliptic integrals in algebraic form

The algebraic forms are obtained by carrying out the following change of variable into the standard forms:

$$u=\sin\Phi. \tag{A2}$$

This yields

$$F^A(u|k)=\int_0^u\frac{du'}{\sqrt{(1-u'^2)(1-k^2u'^2)}}, \tag{A3a}$$

$$E^A(u|k)=\int_0^u du'\sqrt{\frac{1-k^2u'^2}{1-u'^2}}, \tag{A3b}$$

$$\Pi^A(u|n,k)=\int_0^u\frac{du'}{(1-nu'^2)\sqrt{(1-u'^2)(1-k^2u'^2)}}. \tag{A3c}$$

The algebraic forms are those which are implemented in the symbolic calculus software Maple. Notice that for the change of variable (A2) to be bijective, one must restrict Φ to the interval $[-(\pi/2),(\pi/2)]$.

C. Incomplete elliptic integrals in practical form

We call the following forms of the elliptic integrals of the second and third kind ''practical'' because these are the forms under which they appear the most naturally in the problems in Sec. III. These forms are obtained by carrying out the following change of variable into the standard forms:

$$s=F^S(\Phi|k). \tag{A4}$$

This yields

$$E(s|k)=\int_0^s ds'\,\mathrm{dn}^2(s'|k), \tag{A5a}$$

$$\Pi(s|n,k)=\int_0^s ds'\frac{ds'}{1-n\,\mathrm{sn}^2(s'|k)}, \tag{A5b}$$

where the functions sn and dn are defined in Sec. 2.

D. Complete elliptic integrals

The complete elliptic integrals are defined by the expressions for the corresponding incomplete elliptic integrals evaluated at $u=1$ in algebraic form. They are denoted in the following way:

$$K(k)=F^A(1|k), \tag{A6a}$$

$$E(k)=E^A(1|k), \tag{A6b}$$

$$\Pi(n,k)=\Pi^A(1|n,k). \tag{A6c}$$

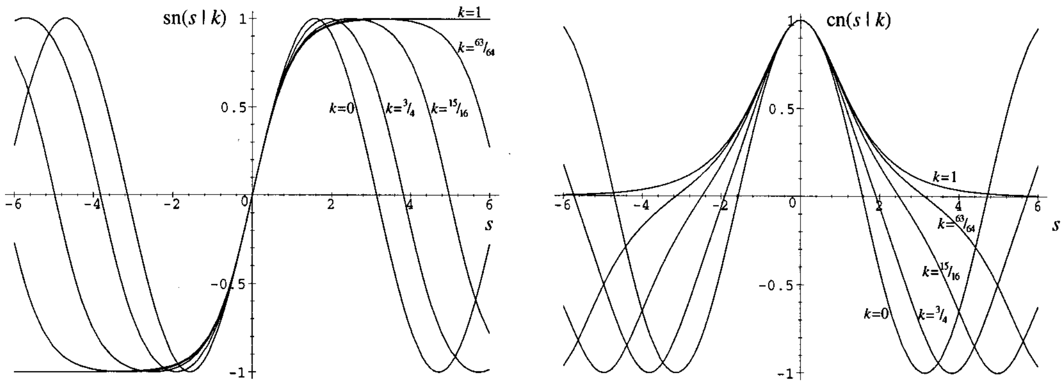

FIG. 25. The function sn and cn for various values of k.

2. Jacobi's elliptic functions

A. Definitions

The incomplete elliptic integrals, having positive integrands, define monotonous, hence invertible, functions. The function am is defined as the inverse of the standard form F^S of the incomplete elliptic integral of the first kind,

$$\mathrm{am}(s|k)=(F^S)^{-1}(s|k). \tag{A7}$$

We then define Jacobi's elliptic functions sn, cn, and dn as

$$\mathrm{sn}(s|k)=\sin \mathrm{am}(s|k), \tag{A8a}$$

$$\mathrm{cn}(s|k)=\cos \mathrm{am}(s|k), \tag{A8b}$$

$$\mathrm{dn}(s|k)=\sqrt{1-k^2\mathrm{sn}^2(s|k)}. \tag{A8c}$$

Notice that the function sn itself, restricted to the interval $[-K(k),K(k)]$, is the inverse of the algebraic form F^A of the incomplete elliptic integral of the first kind. The four functions sn, cn, dn, and am are represented in Figs. 25–26 for various values of k.

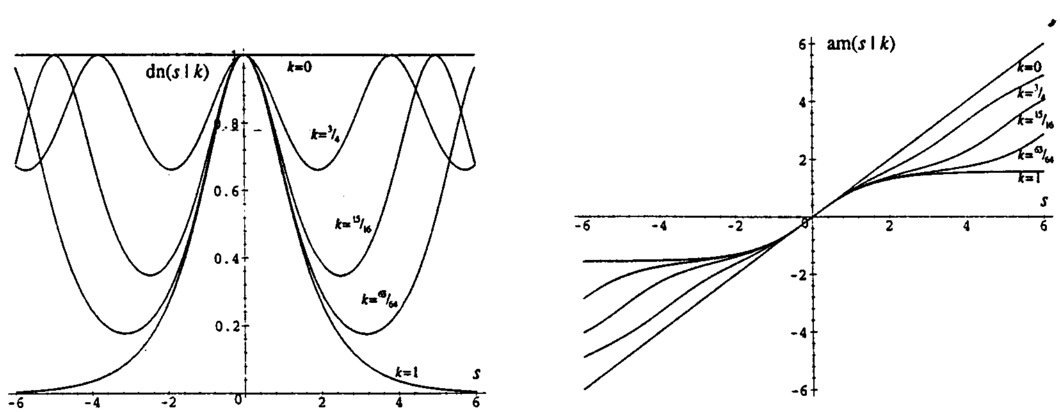

FIG. 26. The function dn and am for various values of k.

B. Elementary properties of am, sn, cn, and dn

The function am obeys the relations

$$\mathrm{am}(2nK(k)\pm s|k)=n\pi\pm\mathrm{am}(s|k), \tag{A9}$$

for any real s and integer n. Hence, the knowledge of am over the interval $[0,K(k)]$, together with (111), is sufficient to reconstruct the function over the whole real line. This holds too for the periodic functions sn, cn, and dn, which obey the following relations:

$$\mathrm{sn}(2nK(k)\pm s|k)=\pm(-1)^n\mathrm{sn}(s|k), \tag{A10a}$$

$$\mathrm{cn}(2nK(k)\pm s|k)=(-1)^n\mathrm{cn}(s|k), \tag{A10b}$$

$$\mathrm{dn}(2nK(k)\pm s|k)=\mathrm{dn}(s|k). \tag{A10c}$$

C. Limits of am, sn, cn, and dn for $k=0$ and $k=1$

These limits are obtained by considering, for $k=0$ and $k=1$, expression (A1a) defining the function F^S, and then taking the limit of am to be the inverse function. For $k=0$, one has

$$\mathrm{am}(s|0)=s, \tag{A11a}$$

$$\mathrm{sn}(s|0)=\sin s, \tag{A11b}$$

$$\mathrm{cn}(s|0)=\cos s, \tag{A11c}$$

$$\mathrm{dn}(s|0)=1, \tag{A11d}$$

whereas $k=1$ leads to

$$\mathrm{am}(s|1)=\arcsin\tanh s, \tag{A12a}$$

$$\mathrm{sn}(s|1)=\tanh s, \tag{A12b}$$

$$\mathrm{cn}(s|1)=\mathrm{sech}\, s, \tag{A12c}$$

$$\mathrm{dn}(s|1)=\mathrm{sech}\, s. \tag{A12d}$$

D. Derivatives of am, sn, cn, and dn

The derivatives of am, sn, cn, and dn are obtained by differentiating (A1a) and the definition Eqs. (A8a)–(A8c). This leads to the following differential relations:

$$\frac{d}{ds}\mathrm{am}(s|k)=\mathrm{dn}(s|k), \tag{A13a}$$

$$\frac{d}{ds}\mathrm{sn}(s|k)=\mathrm{cn}(s|k)\mathrm{dn}(s|k), \tag{A13b}$$

$$\frac{d}{ds}\mathrm{cn}(s|k)=-\mathrm{sn}(s|k)\mathrm{dn}(s|k), \tag{A13c}$$

$$\frac{d}{ds}\mathrm{dn}(s|k)=-k^2\mathrm{sn}(s|k)\mathrm{cn}(s|k). \tag{A13d}$$

[1] J. Coyne, ''Analysis of the formation and elimination of loops in twisted cable,'' IEEE J. Ocean Eng. **15**, 72–83 (1990).
[2] Z. Tan and J. A. Witz, *Loop Formation of Marine Cables and Umbilicals during Installation*, edited by M. H. Patel and R. Gibbins (BBP Technical Services, London, 1992), Vol. 92, pp. 1270–1285.
[3] E. E. Zajac, ''Stability of two loop elasticas,'' Trans. ASME (March), 136–142 (1962).
[4] M. D. Barkley and B. H. Zimm, ''Theory of twisting and bending of chain macromolecules: Analysis of the fluorescence depolarization of DNA,'' J. Chem. Phys. **70**, 2991–3006 (1979).
[5] R. E. Goldstein and S. A. Langer, ''Nonlinear dynamics of stiff polymers,'' Phys. Rev. Lett. **75**, 1094–1097 (1995).
[6] W. Helfrich, ''Elastic theory of helical fibers,'' Langmuir **7**, 567–568 (1991).
[7] J. V. Selinger, F. C. MacKintosh, and J. M. Schnur, ''Theory of cylindrical tubules and helical ribbons of chiral lipid membranes,'' Phys. Rev. E **53**, 3804–3818 (1996).
[8] W. R. Bauer, R. A. Lund, and J. H. White, ''Twist and writhe of a DNA loop containing intrinsic bends,'' Proc. Natl. Acad. Sci. USA **90**, 833–837 (1993).
[9] D. Bensimon, A. J. Simon, V. Croquette, and A. Bensimon, ''Stretching DNA with a receding meniscus: Experiments and models,'' Phys. Rev. Lett. **74**, 4754–4757 (1995).
[10] Ph. Cluzel, A. Lebrun, Ch. Heller, R. Lavery, J.-L. Viovy, D. Chatenay, and F. Caron, ''DNA: An extensible molecule,'' Science **271**, 792–794 (1996).
[11] N. G. Hunt and J. E. Hearst, ''Elastic model of DNA supercoiling in the infinite length limit,'' J. Chem. Phys. **12**, 9329–9336 (1991).
[12] T. Schlick and W. K. Olson, ''Trefoil knotting revealed by molecular dynamics simulations of supercoiled DNA,'' Science **257**, 1110–1114 (1992).
[13] S. B. Smith, Y. Cui, and C. Bustamante, ''Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules,'' Science **271**, 795–799 (1996).
[14] T. R. Strick, J.-F. Allemand, D. Bensimon, and V. Croquette, ''The elasticity of a single supercoiled DNA molecule,'' Science **271**, 1835–1837 (1996).
[15] Y. Yang, I. Tobias, and W. K. Olson, ''Finite element analysis of DNA supercoiling,'' J. Chem. Phys. **98**, 1673–1686 (1993).
[16] A. Goriely and M. Tabor, Nonlinear dynamics of filaments, nonlinear dynamics (to be published).
[17] N. H. Mendelson, ''Bacterial microfibers: The morphogenesis of complex multicellular bacterial forms,'' Sci. Prog. **74**, 425–441 (1990).
[18] J. J. Thwaites and N. H. Mendelson, ''Mechanical behavior of bacterial cell walls,'' Adv. Microbial. Physiol. **32**, 174–222 (1991).
[19] R. L. Ricca and M. A. Berger, ''Topological ideas and fluid mechanics,'' Phys. Today (12), 28–34 (1996).
[20] J. J. Tyson and S. H. Strogatz, ''The differential geometry of scroll waves,'' Int. J. Bifurcation Chaos Appl. Sci. Eng. **1**, 723–744 (1991).
[21] J. P. Keener, ''Knotted vortex filament in an ideal fluid,'' J. Fluid Mech. **211**, 629–651 (1990).
[22] S. Da Silva and A. R. Chouduri, ''A theoretical model for tilts of bipolar magnetic regions,'' Astron. Astrophys. **272**, 621 (1993).
[23] H. C. Spruit, ''Motion of magnetic flux tubes in the solar convection zone and chromosphere,'' Astron. Astrophys. **98**, 155 (1981).
[24] G. Kirchhoff, ''Über das gleichgewicht und die bewegung eines unendlich dünnen elastischen stabes,'' J. Reine Anglew. Math. **56**, 285–313 (1859).
[25] A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, 4th ed. (Dover, New York, 1944).
[26] S. S. Antman, *A Nonlinear Problems of Elasticity* (Springer, Berlin, 1995).
[27] A. Goriely and M. Tabor, ''Nonlinear dynamics of filaments III: Instabilities of helical rods,'' Proc. R. Soc. London, Ser. A **453**, 2583–2601 (1997).
[28] A. Goriely and M. Tabor, ''Nonlinear dynamics of filaments I: Dynamical instabilities,'' Physica D **105**, 20–24 (1997).
[29] A. Goriely and M. Tabor, ''Nonlinear dynamics of filaments IV: Spontaneous looping of elastic rods,'' Proc. R. Soc. London, Ser. A **454**, 3183–3202 (1998).
[30] R. L. Ricca, ''The energy spectrum of a twisted flexible string under elastic relaxation,'' J. Phys. A **28**, 2335–2352 (1995).
[31] C. J. Benham, ''An elastic model of the large-scale structure of duplex DNA,'' Biopolymers **18**, 609–623 (1979).
[32] S. Kehrbaum and J. H. Maddocks, ''Elastic rods, rigid bodies, quaternions, and the last quadrature,'' Philos. Trans. **355**, 2117–2136 (1997).
[33] G. Domokos, ''A group-theoretic approach to the geometry of elastic rings,'' J. Nonlinear Sci. **5**, 453–478 (1995).
[34] Y. Shi and J. E. Hearst, ''The Kirchhoff elastic rod, the nonlinear Schrödinger equation and DNA supercoiling,'' J. Chem. Phys. **101**, 5186–5200 (1994).
[35] B. D. Coleman, E. H. Dill, M. Lembo, Z. Lu, and I. Tobias, ''On the dynamics of rods in the theory of Kirchoff and Clebsch,'' Arch. Ration. Mech. Anal. **121**, 339–359 (1993).
[36] J. L. Lagrange, *Mécanique Analytique* (Chez la Veuve Desaint, Paris, 1788).
[37] W. D. MacMillan, *Dynamics of Rigid Bodies* (McGraw–Hill, New York, 1936).
[38] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed. (Cambridge University Press, London, 1965).
[39] A. Goriely, M. Nizette, and M. Tabor, ''Nonlinear dynamics of strips'' (preprint, 1999).
[40] C. J. Benham, ''Theoretical analysis of conformational equilibria in superhelical DNA,'' Annu. Rev. Biophys. Chem. **14**, 23–45 (1985).
[41] W. F. Pohl, ''DNA and differential geometry,'' Math. Intelligencer **3**, 20–27 (1980).

An exact solution for several charges in classical electrodynamics

R. Rivera[a] and D. Villarroel
Departamento de Física, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile

(Received 31 December 1998; accepted for publication 23 February 1999)

An exact solution for an arbitrary number of identical charges that are equally spaced along a circumference and that rotate at constant angular velocity is presented. The solution is valid for any velocity of the charges less than the speed of light and considers the radiation reaction effects as well as the retarded interaction between the charges. The external field that allows this motion consists of a time-independent electric field tangent to the charges' orbit and a homogeneous time-independent magnetic field that points orthogonally to the orbit plane. A detailed analytical study of the total power of radiation associated with this system of charges is carried out, and it is shown that it is in perfect agreement with the energy that the external electric field supplies to the charges. In particular, in the limit when N goes to infinity only static fields remain. © *1999 American Institute of Physics.* [S0022-2488(99)01906-4]

I. INTRODUCTION

Unlike the equation of motion of a relativistic charge,[1,2] the equations of motion for more than one charge have received little attention in the literature. In this case, besides the radiation reaction forces, the retarded interaction between the charges must also be taken into account. The equations of motion for several charges were obtained by Dirac.[3] If the mass and charge of a particle are denoted by m_k and e_k, respectively, the corresponding equation of motion is the following:

$$m_k a_k^{\mu} = (e_k/c) F_{\text{ext}}^{\mu\alpha} v_{k\alpha} + (e_k/c)\left(\sum_{i\neq k} F_{i\text{ret}}^{\mu\alpha} \right) v_{k\alpha} + (2e_k^2/3c^3)(\dot{a}_k^{\mu} - (1/c^2) a_k^{\lambda} a_{\lambda}^{k} v_k^{\mu}). \tag{1.1}$$

Here $v_{k\mu}$ and $a_{k\mu}$ denote the four-velocity and four-acceleration of charge e_k, respectively; c is the speed of light; $\dot{a}_{k\mu}$ is the proper time derivative of $a_{k\mu}$. Moreover, greek indices range from 0 to 3 and the diagonal metric of Minkowski space is $(-1, +1, +1, +1)$.

The second term on the right-hand side of Eq. (1.1) represents the force on the charge e_k due to the retarded fields of the rest of the charges. It is precisely this term that links the equations of motion of the different charges. Equations (1.1) are usually named Lorentz–Dirac equations, and, in spite of the troubles associated with the existence of preacceleration effects and the existence of unphysical runaway solutions, they are the most widely accepted equations of motion for point charges in classical electrodynamics.

In this paper we will show that Eqs. (1.1) admit as an exact solution the motion of N identical charges that are equally spaced over a circumference of radius a rotating as a rigid body at constant angular velocity ω. The external fields that make possible this motion are the same as those of a previous paper;[4] that is, a time-independent electric field tangent to the orbit circle with a fixed value on it, and a homogeneous time-independent magnetic field orthogonal to the plane where the charges are rotating. As it will be shown, these external fields exist for arbitrary values of the parameters that characterize the motion, that is, charge e, mass m, orbit radius a, angular velocity ω (such that $a\omega < c$), and a number of charges N.

---

[a]Electronic mail: rivera@fis.utfsm.cl

Although Eqs. (1.1) are derived starting from the conservation law for the energy–momentum tensor of the electromagnetic field, the ambiguities associated with the divergent self-energy of a point charge make it necessary to verify the consistency of the solution with the energy that it is radiated away by the system of charges. To this end, we derived an exact formula for the total power of radiation, starting from the far-retarded field generated by the charges. On the other hand, because of the physical symmetries of the charges' motion, in this case it is easy to obtain the total power of radiation directly from the solution of Eqs. (1.1). It turns out that both ways lead to the same formula; so the Lorentz–Dirac equations correctly take into account the energy that escapes to infinity.

Let us point out that, unlike the one charge case, an exact formula for the total power of radiation of a system of several charges in the literature does not exist. The difficulty in obtaining such a formula can be illustrated by the nonexistence of a Lorentz frame, where all the charges are at rest. In general, it is also impossible to obtain any clear answer for the total power of radiation, starting from the equations of motion (1.1). Our success in obtaining an exact formula is mainly due to the physical symmetries that the charges' motion has in the present case. Although the radiation that escapes from the system of charges is determined by their far-retarded fields, the striking point of our calculation is that the physical symmetries allow us to calculate the total power of radiation by means of the fields near the charges. The formula for the total power of radiation takes, of course, full account of the interference effects between the fields of the different charges. In particular, the interference part of the total power of radiation is equal to the number of charges N times the sum of the powers due to the forces on each charge associated with the retarded electric fields of the rest of the charges.

The radiation of the system of charges under consideration here has been studied previously by Comay,[5] who used numerical techniques and considered up to $N=16$ charges. Our treatment is, however, analytic and does not have any restriction on the number of charges. Jackson, in his well-known book,[6] suggests studying the radiation emitted by these charges, starting from the spectral components of the radiation. In this paper, instead of the time-integrated energy flux, we work directly with the instantaneous energy flux.

Our formula for the total power of radiation allows us to study in detail the interference effects on the radiation as the number of charges increases. When the number of charges is very large, the system of charges under study resembles a uniformly charged ring. Now, in the case of a uniformly charged ring that rotates at constant angular velocity, the charge distribution, as well as the corresponding current, is time independent. This means that the electric and magnetic fields are static, and therefore the charges do not radiate energy. This is precisely what happens with the exact solution of the Lorentz–Dirac equation constructed in this paper when the number of charges increases without limit. Incidentally, the mathematical techniques that allow us to carry out the proof of this property are the same as those already developed by Bessel and Watson, in connection with the old Kepler's problem, which can be found in the classic book by Watson.[7] This happens because the equations that define the retarded times of the different charges are practically the same as the equation of motion in the Kepler's problem.

II. THE SOLUTION

In this section it will be shown that for an external time-independent electric field tangent to a circumference with a value fixed on it, and a homogeneous time-independent external magnetic field that points orthogonally to the plane of the circumference, the equations (1.1) admit as a solution the motion of N identical charges equally spaced on the circumference and such that the charges rotate at constant angular velocity ω. For definiteness, the electric charge of the particles is chosen as positive, and it is assumed that the orbit circle lies on the X-Y plane in such a way that the center of the circular orbit of radius a coincides with the origin of coordinates. In addition, the charges are supposed to be rotating counterclockwise, so the external magnetic field $\mathbf{B}^{\text{ext}}$ points in the negative direction of the Z axis and the external electric field $\mathbf{E}^{\text{ext}}$ is tangent to the orbit and points in the direction of the charges' motion.

The retarded electric field generated by a charge e is the following:[8]

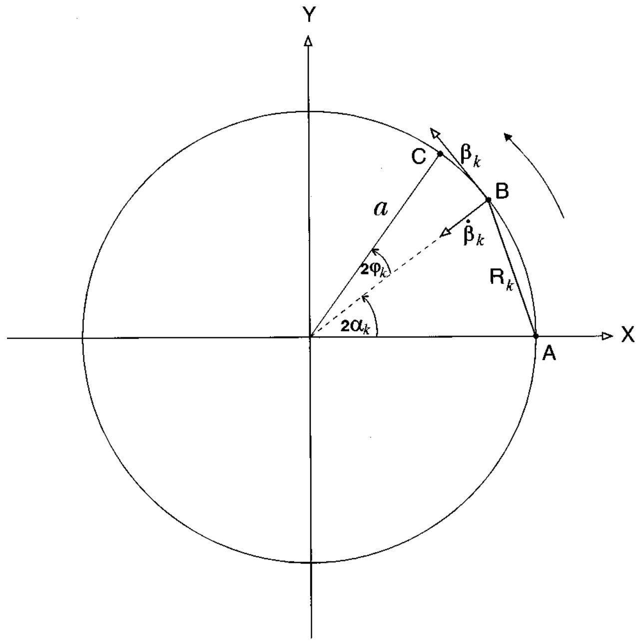

FIG. 1. Points A and C represent the positions at time $t=0$ of particles N and k, respectively. Point B corresponds to the retarded position of particle k associated to the position of particle N at $t=0$. The angular separation between the actual and retarded positions of particle k is denoted by $2\varphi_k$; and the angular separation between the retarded position of particle k and the actual position of particle N is denoted by $2\alpha_k$. Also shown are the retarded quantities $\boldsymbol{\beta}_k$, $\dot{\boldsymbol{\beta}}_k$, and R_k.

$$\mathbf{E}^{\text{ext}}(\mathbf{x},t)=e\left[\frac{(\hat{\mathbf{n}}-\boldsymbol{\beta})(1-\beta^2)}{s^3R^2}\right]+\frac{e}{c}\left[\frac{\hat{\mathbf{n}}\times\{(\hat{\mathbf{n}}-\boldsymbol{\beta})\times\dot{\boldsymbol{\beta}}\}}{s^3R}\right]. \tag{2.1}$$

In this equation, $\hat{\mathbf{n}}$ is the unit vector that points from the retarded position $\mathbf{r}(t')$ of the charge to the point $\mathbf{x}$ where the field is being considered at time t; R is the distance from the retarded position of the charge to point $\mathbf{x}$, that is, $R=|\mathbf{x}-\mathbf{r}(t')|$; $\boldsymbol{\beta}$ and $\dot{\boldsymbol{\beta}}$ are defined by $(1/c)(d\mathbf{r}(t')/dt')$ and $d\boldsymbol{\beta}/dt'$, respectively, and are both evaluated at the retarded time t', implicitly defined by $t=t'+|\mathbf{x}-\mathbf{r}(t')|/c$; and s is the positive number defined by

$$s=1-\hat{\mathbf{n}}\cdot\boldsymbol{\beta}. \tag{2.2}$$

The retarded magnetic field $\mathbf{B}(\mathbf{x},t)$ is given by

$$\mathbf{B}(\mathbf{x},t)=\hat{\mathbf{n}}\times\mathbf{E}(\mathbf{x},t). \tag{2.3}$$

It is easy to see that for the motion under discussion, the external and retarded forces over the charges in Eqs. (1.1) do not have components along the Z axis. Thus, the equations with $\mu=3$ in Eqs. (1.1) are identically satisfied for the motion of the charges in the X-Y plane. Moreover, from the symmetries of the external fields and the charges' motion, it is immediate that the radial and tangential forces are the same for all the charges and that the magnitude of these forces are time independent. In other words, it is enough to consider the forces on only one charge in order to get the values for the external electric field and external magnetic field that make possible the motion under consideration. In Fig. 1 we have drawn the positions of the charges at the actual time, which is chosen as $t=0$. Let us analyze the forces on the charge N, which is located on the X axis at $t=0$. For this charge the radial direction is along the X axis and the tangential direction is along the Y axis. Each charge has associated a number that goes from 1 to N, and where the radial direction that determines the position of the kth charge at time $t=0$ makes an angle $k(2\pi/N)$ with the positive X axis. Besides the actual position of the charge k, in Fig. 1 the retarded position of this charge with respect to the position of charge N at time $t=0$ is also shown. The angle between the radial direction that determines the retarded position of the charge k with the positive X axis is denoted by $2\alpha_k$. It is also convenient to introduce the angle $2\varphi_k$ as the angle between the radial directions of the retarded and actual position of the charge k.

All the retarded quantities associated with the charge k can be expressed in terms of the angle α_k. In fact, the retarded time of the charge k, that is, the time that it takes to go from the retarded position determined by the angle $2\alpha_k$ to its actual position determined by the angle $k(2\pi/N)$, is $2\varphi_k/\omega$. Now, since this time is the same as that light takes to travel the distance $R_k = 2a\sin\alpha_k$, we obtain the following equation that determines the angle α_k as a function of β and parameter k:

$$k\pi/N=\alpha_k+\beta\sin\alpha_k. \tag{2.4}$$

From Fig. 1 it is also easy to obtain for the quantity s_k of (2.2) the following expression:

$$s_k=1+\beta\cos\alpha_k. \tag{2.5}$$

The component with $\mu=2$ in Eq. (1.1) for the charge N located on the X axis at $t=0$ gives rise to the following equation for the determination of E^{ext}:

$$E^{\text{ext}}=\frac{2}{3}\frac{e}{a^2}\beta^3\gamma^4-\sum_{k=1}^{N-1}E_{ky}, \tag{2.6}$$

where $\gamma=(1-\beta^2)^{-1/2}$ and E_{ky} is the component along the Y axis of the retarded electric field of charge k over charge N. According to Eq. (2.1), E_{ky} turns out to be

$$E_{ky}=\frac{e}{4a^2}\left\{\frac{2\beta}{s_k^2}-\frac{\beta+\cos\alpha_k}{\gamma^2 s_k^3\sin^2\alpha_k}\right\}. \tag{2.7}$$

For some charges $E_{ky}>0$, but there are also charges for which $E_{ky}<0$. In general, the sign of E_{ky} has a rather complicated dependence on k and the charge's velocity β. However, as will be shown in Sec. IV, the sum of the Y components of the retarded fields in Eq. (2.6) is positive, as the Y component of the external electric field $\mathbf{E}^{\text{ext}}$ is. On the other hand, the radiation reaction force, that is, the first term on the right-hand side of (2.6), has a negative Y component.

The component along the X axis E_{kx} of the retarded electric field of charge k over charge N turns out to be

$$E_{kx}=\frac{e}{4a^2 s_k^3\sin\alpha_k}\{s_k^2-\beta^2\sin^2\alpha_k\}. \tag{2.8}$$

From this equation, it is easy to see that E_{kx}, in contradistinction with E_{ky}, is positive for any k; which is rather obvious since all charges are positive.

The component $\mu=1$ of Eq. (1.1) gives the following equation, which determines the external magnetic field:

$$B_z^{\text{ext}}=-\frac{mc^2\gamma\beta}{ae}-\frac{1}{\beta}\sum_{k=1}^{N-1}\{E_{kx}+\beta(E_{kx}\cos\alpha_k+E_{ky}\sin\alpha_k)\}. \tag{2.9}$$

Using Eqs. (2.7) and (2.8), the retarded field of the charge k that appears in the { } bracket in (2.9) can be written as follows:

$$\{\ \} = \frac{e}{4a^2 s_k^3 \sin\alpha_k}\left[2\beta(\beta+\cos\alpha_k)s_k + \frac{1}{\gamma^4}\right]. \tag{2.10}$$

Now, since $0<\alpha_k<\pi$ and $s_k \geqslant 1-\beta$ it is easy to see that the quantity (2.10) is positive for any k. In other words, according to Eq. (2.9) the magnitude of the external magnetic field increases as the number of charges increases. This is obviously a consequence of the fact that the repulsive forces between the charges increase as the number of charges increases.

Finally, the component $\mu=0$ in Eq. (1.1), that is, the component associated with energy conservation, reproduces once again Eq. (2.6). In order to analyze the energy conservation equation, let us multiply Eq. (2.6) by the charge e, the velocity v, and the number of charges N. In this way, (2.6) becomes

$$N(evE^{\text{ext}}) = N\left(\frac{2e^2c}{3a^2}\beta^4\gamma^4\right) - N\sum_{k=1}^{N-1} evE_{ky}. \tag{2.11}$$

The left-hand side of this equation represents the power that the external electric field supplies to the system of N charges. Now, in this case the kinetic energy of the charges remains unchanged, so, according to energy conservation, all the power supplied by E^{ext} must be radiated away. In other words, the right-hand side of Eq. (2.11) must represent the total radiated power due to the system of charges under study. But $(2e^2c/3a^2)\beta^4\gamma^4$ is the total radiated power of just one charge in circular motion. So the first term on the right-hand side of (2.11) is the rate of radiation of the system of charges, neglecting all the interference effects. Therefore, the second term on the right-hand side of (2.11) must represent that part of the radiation associated with the interference of the fields of the different charges. In the next section the total power of radiation is studied in a way independent of the equations of motion (1.1), and it will be shown that the radiation emitted does not have any connection with the troubles of mass renormalization. As a result of this calculation, the formula obtained for the total power of radiation of this system of charges is precisely the right-hand side of Eq. (2.11); so the Lorentz–Dirac equations (1.1) are in perfect agreement with energy conservation.

III. THE TOTAL POWER OF RADIATION

The total power of radiation at time t is given by

$$\left(\frac{c}{4\pi}\right)\int_{\Sigma}(\mathbf{E}\times\mathbf{B})\cdot\mathbf{d\Sigma}, \tag{3.1}$$

where Σ is the surface of a sphere of a very large radius, centered at the orbit's center; and $\mathbf{E}$, $\mathbf{B}$ are the retarded electric and magnetic fields generated by the system of charges, which are evaluated over $\mathbf{\Sigma}$ at time t. Although only the far field contributes to the instantaneous energy flux across Σ, it will be shown now that, on account of the symmetries of the charges' motion, the total power of radiation can be calculated using the electric and magnetic fields near the charge distribution. In order to do that, let us consider the instantaneous energy flux across a spherical surface Σ_r centered at the orbit's center and with a radius $r>a$, but otherwise arbitrary. Then, if (θ, φ) denote the usual spherical angles, the integral (3.1) for Σ_r can be performed by considering first the contribution of the ribbon parallel to the orbit plane defined between the angles θ and $\theta + d\theta$. On a fixed point of this ribbon the electric and magnetic fields are obviously changing with time, but since the charges are rotating at constant angular velocity, the positions of the system of charges at different times look as completely equivalent for the ribbon as a whole. This means that the instantaneous energy flux across the ribbon is time independent. Such a property is valid, of course, for any ribbon of Σ_r, so the instantaneous energy flux across Σ_r is time independent.

Let us consider now two spherical surfaces Σ_1 and Σ_2 of radii r_1 and r_2, respectively, and such that $r_2>r_1>a$. If the volume bounded by Σ_1 and Σ_2 is denoted by Ω, then at different times the charge distribution is indistinguishable with respect to Ω; and therefore the total field energy contained in Ω is time independent. The Poynting vector $\mathbf{S}=(c/4\pi)\mathbf{E}\times\mathbf{B}$ satisfies the following conservation equation in Ω:

$$\nabla\cdot\mathbf{S}+\frac{\partial u}{\partial t}=0, \tag{3.2}$$

where u is the energy density $(1/8\pi)(\mathbf{E}^2+\mathbf{B}^2)$. If Eq. (3.2) is integrated over Ω, then it follows, on account of the time independence of the total energy contained in Ω, that the instantaneous energy flux across the surface of the sphere of radius r is not only time independent, but also it does not depend on the radius of the sphere, provided $r>a$. In other words, the total radiated power (3.1) can be calculated using any spherical surface centered at the orbit's center and with a radius r larger than the orbit radius a.

Let us consider, besides the spherical surface Σ_r of radius $r>a$, a torus that encloses the charges' orbit and such that it is contained inside Σ_r. Then, by the same arguments that led to the independence of (3.1) on the radius of the spherical surface, it can be concluded that the instantaneous energy flux across the surface of the torus is precisely the total radiated power associated with the system of charges under study. Moreover, the instantaneous energy flux across the surface of the torus obviously does not depend on the radius of the torus, and therefore it has a perfectly well-defined limit when the radius of the torus goes to zero. In what follows we will use this result in order to obtain an explicit formula for the total power of radiation for the system of charges under consideration. The total power of radiation will be calculated at the laboratory time $t=0$, which is the time when the charge characterized by the number N is on the X axis.

If the electric and magnetic fields of the charge characterized by the number k are denoted by $\mathbf{E}_k$ and $\mathbf{B}_k$, respectively, the total power of radiation can be written as follows:

$$\frac{c}{4\pi}\int\{(\mathbf{E}_1+\mathbf{E}_2+\cdots+\mathbf{E}_N)\times(\mathbf{B}_1+\mathbf{B}_2+\cdots+\mathbf{B}_N)\}\cdot\mathbf{d\Sigma}, \tag{3.3}$$

where the integral is carried out over the surface of a torus that encloses the charges' orbit. In Eq. (3.3) there are N terms that involve only the fields of one charge. Now, since the above discussion about the symmetries applies also to the case of only one charge in circular motion, the integral over the surface of the torus of the Poynting vector associated with the field of one charge gives, of course, the total power of radiation of a monoenergetic electron in circular orbit, which is $(2e^2c/3a^2)\beta^4\gamma^4$. Let us emphasize that the corresponding calculation does not present any trouble in the limit when the radius of the torus goes to zero.[9]

The interference part of the rate of radiation (3.3) associated with the field of the charge N appears only in the terms

$$\frac{c}{4\pi}\int(\mathbf{E}_N\times\mathbf{B}_k+\mathbf{E}_k\times\mathbf{B}_N)\cdot\mathbf{d\Sigma}, \tag{3.4}$$

where k takes the values $1,2,...,N-1$. In Fig. 2 the surface of the torus over which the instantaneous energy flux (3.4) is evaluated is shown. This torus is characterized by the radius b of the circumference determined by the intersection of the torus surface with the planes that contain the Z axis, which, in turn, are determined by means of the azimuthal angle φ. A point over a circumference of radius b is defined by means of the angle θ, which is measured from the $X-Y$ plane and ranges between 0 and 2π.

The surface element of the torus is given by

$$\mathbf{d\Sigma}=a^2\epsilon(1+\epsilon\cos\theta)\hat{\mathbf{b}}\,d\theta\,d\varphi, \tag{3.5}$$

where $\epsilon=b/a<1$, and $\hat{\mathbf{b}}$ is the outer unit normal to the torus, given by

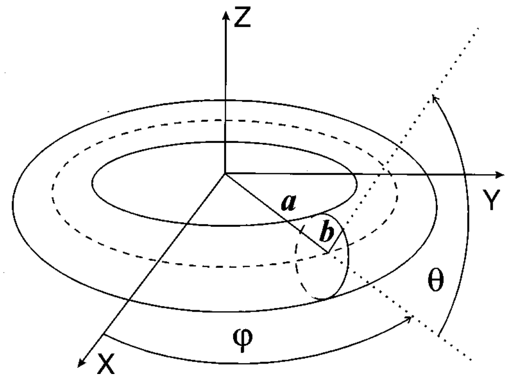

FIG. 2. The toroidal surface of radius b enclosing the electron orbit. Point P over the surface of the torus is characterized by the azimuthal angle φ that specifies a particular cross section, and the angle θ that locates P over the circumference of the latter.

$$\hat{\mathbf{b}}=(\cos\theta\cos\varphi)\hat{\mathbf{x}}+(\cos\theta\sin\varphi)\hat{\mathbf{y}}+(\sin\theta)\hat{\mathbf{z}}. \tag{3.6}$$

The detection point $\mathbf{x}$ over the surface of the torus can be written as follows:

$$\mathbf{x}=[a(1+\epsilon\cos\theta)\cos\varphi]\hat{\mathbf{x}}+[a(\mathbf{1}+\epsilon\cos\theta)\sin\varphi]\hat{\mathbf{y}}+[b\sin\theta]\hat{\mathbf{z}}. \tag{3.7}$$

Therefore the retarded times t'_N and t'_k of the charges N and k, respectively, that are associated with the point $\mathbf{x}$ of (3.7) and time $t=0$ are such that

$$\omega t=0=\omega t'_N+\beta\rho_N=\omega t'_k+\beta\rho_k, \tag{3.8}$$

where

$$\rho_N=\{\epsilon^2+2(1+\epsilon\cos\theta)(1-\cos x)\}^{1/2}, \tag{3.9}$$

and

$$\rho_k=\{\epsilon^2+2(1+\epsilon\cos\theta)(1-\cos y)\}^{1/2}. \tag{3.10}$$

In Eqs. (3.9) and (3.10), the variables x and y are the following:

$$x=\varphi-\omega t'_N, \tag{3.11}$$

$$y=\varphi-\omega t'_k-k(2\pi/N). \tag{3.12}$$

According to Eq. (3.8), the variables x and y are linked by

$$y-x+k(2\pi/N)=\beta(\rho_k-\rho_N). \tag{3.13}$$

In the limit when the radius of the torus goes to zero, that is, when the parameter ϵ tends to zero, the surface element (3.5), as well as the surface of the torus, goes to zero. But, as is already known, the integral (3.4) cannot be zero; so $\mathbf{E}_N\times\mathbf{B}_k+\mathbf{E}_k\times\mathbf{B}_N$ must be singular in some points in this limit. From a physical point of view this behavior is rather trivial because the fields of the charges N and k are singular at the points where these charges are located at $t=0$. From a mathematical point of view, the singular behavior can be easily seen, since the distances to the charges are measured by means of the quantities ρ_N and ρ_k given in (3.9) and (3.10), respectively; but ρ_N is zero for $\epsilon=0$ and $x=0$, while ρ_k is zero for $\epsilon=0$ and $y=0$. Now, for points very close to the charge N the retarded time t'_N is almost equal to the actual time t, which is zero; so according to (3.11), $x=0$ corresponds to $\varphi=0$, that it is precisely the point where the charge N is located at $t=0$. Similarly, from (3.12) it follows that y is zero for $\varphi=k(2\pi/N)$, that it is the point of the orbit where the charge k is located at $t=0$. Therefore, for a very small ϵ the integrand of (3.4) is very close to zero everywhere, with the only exception of two small vicinities around $\varphi=0$ and $\varphi=k(2\pi/N)$.

In what follows we will study the contribution of (3.4) around $\varphi=0$ for ϵ very small. The electric and magnetic fields of charge k, that is, $\mathbf{E}_k$ and $\mathbf{B}_k$, respectively, are smooth functions of φ around $\varphi=0$; so the contribution of (3.4) around $\varphi=0$ can be written in the form

$$\left(\frac{c}{4\pi}\right)\mathbf{B}_k\cdot\int_{\varphi\approx 0}(\hat{\mathbf{b}}\times\mathbf{E}_N)d\Sigma+\left(\frac{c}{4\pi}\right)\mathbf{E}_k\cdot\int_{\varphi\approx 0}(\mathbf{B}_N\times\hat{\mathbf{b}})d\Sigma. \tag{3.14}$$

Here $\hat{\mathbf{b}}\times\mathbf{E}_N$ and $\mathbf{B}_N\times\hat{\mathbf{b}}$ are sharply defined around $\varphi=0$, and thus the integrals must be carefully carried out. Since in the limit ϵ going to zero only a small vicinity around $\varphi=0$ needs to be considered, it is evident that its contribution is fully taken into account if the variable φ covers the whole range $0\leqslant\varphi<2\pi$. Moreover, instead of φ it is convenient to use the variable x defined in (3.11) to carry out the integration around the ribbon of the torus. From (3.11) and (3.8), it follows that

$$\frac{d\varphi}{dx}=s_N>0, \tag{3.15}$$

which shows that φ is a strictly increasing function of x. From (3.11) it also follows that the limits of integration in the variable x are x_0 and $x_0+2\pi$, with $x_0=-\omega t_N'(\varphi=0)=-\omega t_N'(\varphi=2\pi)$. But, as it is easy to see, the variable x appears in the integrand only as $\sin x$ and $\cos x$; thus, the integrand is a periodic function of x and, therefore, x_0 can be put equal to zero. Finally, in order to get a systematic procedure to study the different contributions in (3.14), it is convenient to replace $\sin x$ in terms of the corresponding s_N of (2.2) and $\cos x$ in terms of ρ_N given in (3.9), that is,

$$\sin x=\frac{\rho_N(1-s_N)}{\beta(1+\epsilon\cos\theta)},$$
$$\cos x=1-\frac{\rho_N-\epsilon^2}{2(1+\epsilon\cos\theta)}. \tag{3.16}$$

In this way, all the contributions to the energy flux across the ribbon of the torus between θ and $\theta+d\theta$ in (3.14) are expressed in terms of integrals of the type

$$I(n,p,q)=\epsilon^n\int_0^{2\pi}\frac{dx}{(s_N)^p(\rho_N)^q}, \tag{3.17}$$

with $n\geqslant 1$ due to Eq. (3.5) for the surface element.

Before going to the explicit evaluation of (3.17), let us remember that when (3.14) is evaluated for different values of k, that is, for $k=1,2,...,N-1$, the result will represent that part of the total power of radiation associated with the interference of the field of charge N with the field of the rest of the charges. We are not considering the second contribution that comes from the peak associated with $y=0$ in (3.4), because this will be part of the power of radiation associated with the interference of the field of charge k with the field of the rest of the charges, that is, with the field of the charges $k+1,k+2,...,N,1,2,...,k-1$. But because of the symmetries, the contribution to the total power of radiation of the interference between the field of the charge k with the rest is, of course, identical with the contribution that comes from the interference of the field of charge N with the rest of the charges.

In what follows we will study (3.17) in order to obtain the leading terms of a power series representation of the integral in the parameter ϵ, and where all the terms that vanish in the limit $\epsilon\to 0$ are thrown away. To this end, note that the quantity $\hat{\mathbf{n}}_N\cdot\boldsymbol{\beta}_N=\beta(1+\epsilon\cos\theta)\sin x/\rho_N$ in (2.2) is less than one, which allows us to expand the s_N factor in Eq. (3.17):

$$I(n,p,q)=\sum_{m=0}^{\infty}\frac{a_m^p\beta^{2m}(1+\epsilon\cos\theta)^{2m}}{[\epsilon^2+2(1+\epsilon\cos\theta)]^{(2m+q)/2}}\cdot 2\epsilon^n\int_0^{\pi}\frac{(\sin x)^{2m}dx}{(1-\delta\cos x)^{m+q/2}}, \tag{3.18}$$

where we have written Eq. (3.9) in the form

$$\rho_N=[\epsilon^2+2(1+\epsilon\cos\theta)]^{1/2}\{1-\delta\cos x\}^{1/2}, \tag{3.19}$$

with

$$\delta=\frac{2(1+\epsilon\cos\theta)}{\epsilon^2+2(1+\epsilon\cos\theta)}<1, \tag{3.20}$$

and where

$$a_m^p=\frac{(2m+p-1)!}{(2m)!(p-1)!}. \tag{3.21}$$

Note that, if we exclude a small vicinity of $x=0$ in the integral of (3.18), and since $\delta<1$ and $\cos\alpha>\cos x$ for $x>\alpha$, then

$$\epsilon^n\int_{\alpha}^{\pi}\frac{(\sin x)^{2m}dx}{(1-\delta\cos x)^r}<\frac{\epsilon^n}{(1-\cos\alpha)^r}\int_{\alpha}^{\pi}(\sin x)^{2m}dx. \tag{3.22}$$

The right-hand side of Eq. (3.22) goes to 0 when ϵ goes to 0, which shows that it is sufficient to consider an arbitrarily small vicinity around $x=0$. The use of the whole range $0\leqslant x<2\pi$ is advantageous because it makes it possible to work with analytic expressions for the integral. Thus, the following result will be used:[10]

$$\int_0^{\pi}\frac{(\sin x)^{2m}\,dx}{(1-\delta\cos x)^r}=\frac{\pi^{1/2}\Gamma(m+\frac{1}{2})}{\Gamma(m+1)}\,{}_2F_1((r+1)/2,r/2,m+1,\delta^2), \tag{3.23}$$

where $\Gamma(z)$ is the gamma function and ${}_2F_1(a,b,c,x)$ denotes the hypergeometric function.

Let us study (3.17) in the specific case $q=3$. Then we must use $r=m+\frac{3}{2}$ in (3.23). In order to see clearly the leading terms in (3.23), it is convenient to use the transformation formula[11] for the hypergeometric function,

$$F(a,b,a+b-l,\delta^2)=\frac{\Gamma(l)\Gamma(a+b-l)}{\Gamma(a)\Gamma(b)}(1-\delta^2)^{-l}\sum_{n=0}^{l-1}\frac{(a-l)_n(b-l)_n}{n!(1-l)_n}(1-\delta^2)^n$$
$$+\text{terms irrelevant to the flux}, \tag{3.24}$$

where $(z)_n$ is the Pochhammer symbol,

$$(z)_n=z(z+1)\cdots(z+n-1)=\Gamma(z+n)/\Gamma(z). \tag{3.25}$$

The advantage of using Eq. (3.24) is that, instead of the parameter δ^2 that appears in (3.23) on the right-hand side of (3.24), it appears as

$$1-\delta^2=\frac{\epsilon^2[\epsilon^2+4(1+\epsilon\cos\theta)]}{[\epsilon^2+2(1+\epsilon\cos\theta)]^2}, \tag{3.26}$$

which goes to zero as ϵ^2 when $\epsilon\to 0$. The irrelevant terms that are not explicitly given in (3.24) vanish in the limit $\epsilon\to 0$, except one of them that diverges as $\log\epsilon$. However, this is of no importance in the calculation of the energy flux because of the overall factor ϵ^n in (3.17) with $n\geqslant 1$.

Therefore, the leading term in Eq. (3.23) is

$$\epsilon^n\,{}_2F_1(m/2+\tfrac{5}{4},m/2+\tfrac{3}{4},m+1,\delta^2)=\frac{\epsilon^n\Gamma(m+1)}{\Gamma(m/2+\frac{3}{4})\Gamma(m/2+\frac{5}{4})}(1-\delta^2)^{-1}. \tag{3.27}$$

Now, using

$$\Gamma(m/2+\tfrac{3}{4})\Gamma(m/2+\tfrac{5}{4})=(2\pi)^{1/2}2^{-(m+2)}(2m+1)\Gamma(m+\tfrac{1}{2}), \tag{3.28}$$

replacing Eqs. (3.23), (3.27), and (3.26) in Eq. (3.18), and expanding the expressions containing ϵ, we get, for the leading terms of $I(n,p,3)$,

$$I(n,p,3)=2\sum_{m=0}^{\infty}\frac{a_m^p\beta^{2m}}{(2m+1)}\epsilon^{n-2}\left\{1+\frac{(2m-1)\cos\theta}{2}\epsilon+\cdots\right\}. \tag{3.29}$$

The same ideas as above are easily applied to Eq. (3.17) in the cases $q=2$ and $q=1$. In the case $q=2$, we obtain

$$I(n,p,2)=\pi\sum_m a_m^p\beta^{2m}\frac{(2m-1)!!}{(2m)!!}\epsilon^{n-1}\left\{1+\frac{(2m-1)\cos\theta-2m}{2}\epsilon+\cdots\right\}, \tag{3.30}$$

while in the case $q=1$, and since $n\geqslant 1$, it is found that $I(n,p,1)$ always vanishes in the limit when ϵ goes to zero. Of course, the cases $q\geqslant 0$ also vanish in this limit.

By using the above rules, it follows that in the limit $\epsilon\to 0$ the first term of (3.14) vanishes. In the second term, in order to simplify more easily the results, it is convenient to introduce the components of the field $\mathbf{E}_k$ under the integral sign. The following nonvanishing terms result in the limit $\epsilon\to 0$,

$$\left(\frac{e^2c}{4\pi a^2}\right)\left\{\frac{-\beta\gamma^{-4}\sin y_0}{s_k^3\rho_k^3}+\frac{\beta^2\gamma^{-4}\cos y_0}{s_k^3\rho_k^2}+\frac{\beta^3\gamma^{-2}\sin y_0}{s_k^2\rho_k}+\frac{\beta^4\gamma^{-2}\cos y_0(1-\cos y_0)}{s_k^3\rho_k^2}\right.$$
$$\left.-\frac{\beta^3\gamma^{-2}\sin y_0(1-\cos y_0)}{s_k^3\rho_k^3}\right\}\cdot\int_0^{2\pi}d\theta\int_0^{2\pi}\frac{\epsilon^2\,dx}{s_N^2\rho_N^3}, \tag{3.31}$$

where it is understood that the factors s_k and ρ_k in the curly brackets of Eq. (3.31) are evaluated at $y=y_0$, which is given by the retardation condition, Eq. (3.13), with $x=0$ and $\epsilon=0$.

The integral in Eq. (3.31) has $q=3$, $n=2$, $p=2$, so only the factor 1 in the curly brackets of (3.29) must be retained. In addition, Eq. (3.21) implies $a_m^p=(2m+1)$ and therefore the integral in Eq. (3.31) becomes

$$\int_0^{2\pi}d\theta\cdot 2\sum_{m=0}^{\infty}\beta^{2m}=4\pi\gamma^2.$$

As mentioned above, the value of y_0 in Eq. (3.31) is determined by the implicit equation

$$y_0+2k\pi/N=\beta\{2(1-\cos y_0)\}^{1/2}. \tag{3.32}$$

By plotting both sides of (3.32) as a function of y_0, it is easily seen that $-2k\pi/N < y_0 < 0$, and therefore Eq. (3.32) can be written as

$$k\pi/N = -\frac{y_0}{2} - \beta \sin\frac{y_0}{2}, \tag{3.33}$$

which, upon comparison with Eq. (2.4), shows that

$$y_0 = -2\alpha_k. \tag{3.34}$$

Then, by substituting Eq. (3.34) into Eq. (3.31) and comparing with Eq. (2.7) for the retarded field of particle k over particle N, it is easily verified that the flux over the torus in Eq. (3.4) takes the value $-evE_{ky}$, and, therefore, when the whole system of charges is considered, Eq. (2.11), derived from the Lorentz–Dirac equations, is obtained again.

IV. THE EXTERNAL ELECTRIC FIELD FOR LARGE *N*

In this section we are going to study in detail the dependence of the external electric field E^{ext} given by Eq. (2.6), as a function of the number of charges N for large values of N. For this purpose it is convenient to write the retarded electric field E_{ky} of Eq. (2.6) in a slightly different form that is more appropriate for obtaining a representation of it as a power series in the parameter β.

The retarded field E_{ky} depends, of course, in an explicit way of β, but since according to Eq. (2.4) the angle α_k is a function of β, E_{ky} also depends implicitly on β. From this last equation it is easy to obtain

$$\frac{\partial \alpha_k}{\partial \beta} = -\frac{\sin \alpha_k}{s_k}, \tag{4.1}$$

which shows that α_k is a decreasing function of β, on account of the fact that $0 < \alpha_k < \pi$ and $s_k > 0$. The partial derivation symbol is used in (4.1) because the angle α_k depends also on the variable $k\pi/N$, which is completely independent of β. With the help of (4.1) it is easy to see that (2.7) can be written in the following form:

$$E_{ky} = -\frac{e}{4a^2}\left\{(1-\beta^2)\frac{\partial}{\partial \beta}\cdot\left(\frac{1}{s_k \sin^2 \alpha_k}\right) + 2\beta^2 \frac{\partial}{\partial \beta}\left(\frac{1}{s_k}\right)\right\}. \tag{4.2}$$

The right-hand side of Eq. (4.2) can be written as a power series in β with the help of the formalism of Watson's book.[7] For this purpose it is convenient to replace the discrete variable $k\pi/N$ of Eq. (2.4) by a continuous variable ψ, that is,

$$\psi = \alpha + \beta \sin \alpha. \tag{4.3}$$

If in Eq. (4.3) β is changed by minus the parameter of eccentricity associated with the elliptical orbit of the inverse square law force, Eq. (4.3) becomes identical with the equation of motion of Kepler's problem. Watson presents a detailed treatment of Eq. (4.3) in his book.[7] The pertinent part of Watson's book will be sketched here. From Eq. (4.3), it follows that

$$\frac{\partial \alpha}{\partial \psi} = \frac{1}{s} > 0, \tag{4.4}$$

where $s = 1 + \beta \cos \alpha$ is the same that appears in Eq. (2.5), but without the subindex k. Equation (4.4) shows that α is an increasing function of ψ and the effect of increasing ψ by 2π is to increase α by 2π. Therefore, $1/(1+\beta\cos\alpha)$ is an even periodic function of ψ, and so it can be expanded in a Fourier cosines series, which turns out to be

$$\frac{1}{s}=1+2\sum_{n=1}^{\infty} J_n(-n\beta)\cos(n\psi), \tag{4.5}$$

where $J_n(x)$ is the Bessel function. When Eq. (4.5) is introduced on the right-hand side of

$$\frac{\partial}{\partial\beta}\left(\frac{1}{s\sin\alpha}\right)=-\frac{\partial}{\partial\psi}\left(\frac{1}{s}\right), \tag{4.6}$$

the following expansion is obtained for the function $1/(s\sin\alpha)$:

$$\frac{1}{s\sin\alpha}=\frac{1}{\sin\psi}+2\sum_{n=1}^{\infty} n\sin(n\psi)\int_0^{\beta} J_n(-nx)dx. \tag{4.7}$$

Now, if this equation is differentiated with respect to ψ, and taking into account that

$$\frac{\cos\alpha}{s^2\sin^2\alpha}-\frac{\beta}{s^3}=\frac{\partial}{\partial\beta}\left(\frac{1}{s\sin^2\alpha}\right), \tag{4.8}$$

the following expansion is obtained:

$$\frac{\partial}{\partial\beta}\left(\frac{1}{s\sin^2\alpha}\right)=\frac{\cos\psi}{\sin^2\psi}-2\sum_{n=1}^{\infty} n^2\cos(n\psi)\int_0^{\beta} J_n(-nx)dx. \tag{4.9}$$

Equations (4.2), (4.5), and (4.9) are reduced to those previously given by Gordeyev[12] for the case of two charges rotating at a constant angular velocity at opposite ends of a diameter.

Equations (4.5) and (4.9), with $\psi=k\pi/N$, allow us to write the tangential component of the retarded electric field E_{ky} of (4.2) as a power series in β. When (4.5), (4.9), and the power series representation for $J_n(x)$ are introduced in (4.2), the following expression for the sum of the retarded tangential electric field is obtained:

$$\sum_{k=1}^{N-1} E_{ky}=-\left(\frac{e}{2a^2}\right)\sum_{n=1}^{\infty}(-1)^n\left[\sum_{k=1}^{N-1}\cos(nk\pi/N)\right] \times\left\{(-n^2D_{n,0}+2C_{n,0})\beta^{n+1}+\sum_{m=0}^{\infty}(-n^2D_{n,m+1}+2C_{n,m+1}+n^2D_{n,m})\beta^{n+2m+3}\right\}, \tag{4.10}$$

where we used the result

$$\sum_{k=1}^{N-1}\frac{\cos(k\pi/N)}{\sin^2(k\pi/N)}=0,$$

and the coefficients C_{nm} and D_{nm} come from the power series representation of $J_n(x)$, and are the following:

$$C_{n,m}=\frac{(-1)^m(n+2m)n^{n+2m}}{2^{n+2m}m!(n+m)!}, \tag{4.11}$$

$$D_{n,m}=\frac{(-1)^m n^{n+2m}}{2^{n+2m}m!(n+m)!(n+2m+1)}. \tag{4.12}$$

The sum over k that appears in (4.10) can be explicitly carried out, since

$$\sum_{k=1}^{N-1} \cos(nk\pi/N) = \begin{cases} 0, & n \text{ odd}, \\ -1, & n \text{ even} < 2N, \\ N-1, & n = 2N, 4N, \ldots . \end{cases} \tag{4.13}$$

From this formula it follows that only odd powers of β remain in (4.10), where, because of (4.13), it is convenient to split the sum over the index n in a part with $n \leqslant 2N-2$, and the rest that contains powers of β higher than $2N+1$. In this way, (4.10) can be written as

$$\sum_{k=1}^{N-1} E_{ky} = \left(\frac{e}{a^2}\right)\Bigg\{[-2D_{2,0}+C_{2,0}]\beta^3 + \sum_{n=1}^{N-2}\Bigg[-2(n+1)^2 D_{2n+2,0} + C_{2n+2,0} + \sum_{m=0}^{n-1}(-2(n-m)^2 D_{2n-2m,m+1} + C_{2n-2m,m+1} + 2(n-m)^2 D_{2n-2m,m})\Bigg]\beta^{2n+3}\Bigg\} + O(\beta^{2N+1}), \tag{4.14}$$

which, when (4.11) and (4.12) are used, is reduced to

$$\sum_{k=1}^{N-1} E_{ky} = \left(\frac{e}{a^2}\right)\Bigg\{\frac{2}{3}\beta^3 + \sum_{n=1}^{N-2}\Bigg[\frac{2(n+2)(n+1)^{2n+3}}{(2n+3)!} + 2\sum_{m=0}^{n-1}\frac{(-1)^m (n-m)^{2n+2}(-7n^2-2n^3-2m^2-3n+4nm)}{(2n+1)(2n+3)(m+1)!(2n+1-m)!}\Bigg]\beta^{2n+3}\Bigg\} + O(\beta^{2N+1}). \tag{4.15}$$

The sums in Eq. (4.15) can be cast in a simpler way by using the new dumb indices $i=n+1$ and $l=n-m$, in terms of which Eq. (4.15) reads as

$$\sum_{k=1}^{N-1} E_{ky} = \left(\frac{e}{a^2}\right)\Bigg\{\frac{2}{3}\beta^3 + 2\sum_{i=2}^{N-1}\sum_{l=1}^{i}\frac{(-1)^{i-l} l^{2i}[2i^3-i^2-i+2l^2]}{(2i-1)(2i+1)(i-l)!(i+l)!}\beta^{2i+1}\Bigg\} + O(\beta^{2N+1}). \tag{4.16}$$

The sum over the index l that appears in the above expression can be carried out with the help of the formulas in the Appendix, and the result is

$$\sum_{k=1}^{N-1} E_{ky} = \frac{2e}{3a^2}\sum_{n=1}^{N-1} n\beta^{2n+1} + O(\beta^{2N+1}). \tag{4.17}$$

Therefore, from Eq. (2.6) the following estimate is obtained for the external electric field:

$$E^{\text{ext}} = O(\beta^{2N+1}). \tag{4.18}$$

Equation (4.18) is valid for arbitrary N, and since $\beta < 1$, it shows that the external electric field tends to zero when the number of charges increases. For a nonrelativistic motion, the interference part of the total power of radiation is very effective in supressing the energy radiated away, even with a moderate number of charges N. But for a highly relativistic motion, a very large number of charges is needed to supress the radiation. This can be understood on account of the high directionality of synchrotron radiation.[8]

ACKNOWLEDGMENTS

We wish to thank to Comisión Nacional de Ciencia y Tecnología de Chile, CONICYT, for support through project FONDECYT 1960179 and 1990297 that has been a determinant to the development of this research. R. Rivera is thankful to Fundación Andes for a doctoral grant.

APPENDIX: SUMS TO USE IN EQ. (4.16)

In this appendix, we will show that

$$S_1=\sum_{l=1}^{i}\frac{(-1)^{i-l}l^{2i}}{(i-l)!(i+l)!}=\frac{1}{2}, \tag{A1}$$

$$S_2=\sum_{l=1}^{i}\frac{(-1)^{i-l}l^{2i+2}}{(i-l)!(i+l)!}=\frac{i(i+1)(2i+1)}{12}. \tag{A2}$$

Let us first rewrite Eq. (A1) in terms of $v=i-l$,

$$S_1=\sum_{v=0}^{i-1}\frac{(-1)^{v}(i-v)^{2i}}{(2i-v)!v!},$$

and note that

$$S_1=\frac{1}{2}\cdot\sum_{v=0}^{2i}\frac{(-1)^{v}(i-v)^{2i}}{(2i-v)!v!},$$

which can be written as

$$S_1=\frac{1}{2(2i)!}\sum_{v=0}^{2i}(-1)^{v}\binom{2i}{v}(i-v)^{2i}. \tag{A3}$$

Therefore, the identity[13]

$$\sum_{k=0}^{n}(-1)^{k}\binom{n}{k}(\alpha+k)^{n}=(-1)^{n}n! \tag{A4}$$

directly leads to Eq. (A1).

Equation (A2) can be proved by induction on i. Here it also proves convenient to write S_2 in the form

$$S_2=\frac{1}{2(2i)!}\sum_{v=0}^{2i}(-1)^{v}\binom{2i}{v}(i-v)^{2i+2}.$$

Then, we want to prove

$$\frac{1}{2(2i+2)!}\sum_{v=0}^{2i+2}(-1)^{v}\binom{2i+2}{v}(i+1-v)^{2i+4}=\frac{(i+1)(i+2)(2i+3)}{12}, \tag{A5}$$

by using the hypothesis

$$\frac{1}{2(2i)!}\sum_{v=0}^{2i}(-1)^{v}\binom{2i}{v}(i-v)^{2i+2}=\frac{i(i+1)(2i+1)}{12}. \tag{A6}$$

To this end, the identity

$$\binom{a}{b}=\binom{a-1}{b}+\binom{a-1}{b-1} \tag{A7}$$

is used twice on the left-hand side of Eq. (A5), which after relabeling the dumb indices reads as

$$\frac{1}{2(2i+2)!}\sum_{v=0}^{2i}(-1)^v\binom{2i}{v}(i+1-v)^{2i+4}-\frac{1}{(2i+2)!}\sum_{v=0}^{2i}(-1)^v\binom{2i}{v}(i-v)^{2i+4}$$
$$+\frac{1}{2(2i+2)!}\sum_{v=0}^{2i}(-1)^v\binom{2i}{v}(i-1-v)^{2i+4}. \tag{A8}$$

Here the binomial theorem is used in the first and third term of Eq. (A8) to expand the powers of $(i-v\pm1)$ into powers of $(i-v)$. After collecting terms and because of the identity[13]

$$\sum_{k=0}^{N}(-1)^k\binom{N}{k}(\alpha+k)^{n-1}=0, \quad n\leqslant N, \tag{A9}$$

only two nonvanishing sums remain. They are easily evaluated by using Eqs. (A6) and (A4), leading directly to the right-hand side of Eq. (A5).

[1] G. N. Plass, Rev. Mod. Phys. **33**, 37 (1961).
[2] F. Rohrlich, *Classical Charged Particles* (Addison-Wesley, Reading, MA, 1965).
[3] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **167**, 148 (1938).
[4] R. Rivera and D. Villarroel, J. Math. Phys. **38**, 5690 (1997).
[5] E. Comay, Phys. Lett. A **123**, 425 (1987).
[6] J. D. Jackson, *Classical Electrodynamics*, 2nd ed. (Wiley, New York, 1975), Chap. 14, Prob. 14.12.
[7] G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge University Press, Cambridge, 1922), Chap. 17.
[8] See Ref. 6, Chap. 14.
[9] R. Rivera and D. Villarroel, to appear in Am. J. Phys.
[10] I. S. Gradshtein and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, Orlando, 1980); put $\cos x=u$ and use formulas (8.712) and (8.703).
[11] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1964).
[12] A. N. Gordeyev, Am. J. Phys. **65**, 1060 (1997)
[13] See Ref. 10, formulas (0.154).

Three-dimensional formulation of the Maxwell equations for stationary space–times

G. F. Torres del Castillo[a)]
Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, 72000 Puebla, Pue., México

J. Mercado-Pérez
Instituto Tecnológico de Toluca, Apartado postal 890, 50000 Metepec, Méx., México

(Received 7 October 1998; accepted for publication 1 February 1999)

Using the correspondence between a stationary space–time and a curved three-dimensional space with a static magnetic field and a conservative field of force, the Maxwell equations in a stationary space–time are expressed in terms of fields associated with the corresponding three-dimensional geometry. The Maxwell equations take a form analogous to the one they have in flat space–time. The Ricci tensor of the space–time is also written in terms of the three-dimensional fields. © *1999 American Institute of Physics.* [S0022-2488(99)01405-X]

I. INTRODUCTION

In the study of the null[1,2] and timelike[2] geodesics of a stationary space–time, a certain decomposition of the metric arises in a natural way in such a manner that, by a suitable change in the parametrization, the problem of finding the geodesics of a stationary space–time is equivalent to finding the orbits of a charged (nonrelativistic) particle in the three-dimensional space of trajectories of the timelike Killing vector field, with a static magnetic field and a velocity-independent potential determined by the space–time metric. When the space–time is static, this magnetic field vanishes and the metric defined on the three-dimensional space mentioned above reduces to the so-called ''optical metric,'' which has many remarkable properties (see, e.g., Refs. 3, 4 and the references cited therein). One of these properties is the fact that the Maxwell equations in a static space–time take their simplest form (the one they have in an inertial frame in flat space–time) when expressed in terms of the optical metric.[4] In this paper we show that, in a similar manner, the decomposition of the metric of a stationary space–time found in Refs. 1 and 2 is not only distinguished by the dynamics of test particles, but also by the fact that the Maxwell equations can be written in the same form as in an inertial frame in flat space–time.

In Sec. II we write down explicitly the geodesic equation in a stationary space–time and we show its equivalence with the equations of motion of a charged particle in a three-dimensional space subject to a magnetic field and a velocity-independent potential. We also show that from the symmetries of this three-dimensional space that leave the magnetic field and the velocity-independent potential invariant one can obtain symmetries of the space–time metric. In Sec. III the Maxwell equations in a stationary space–time are written in a $3+1$ form, making use of the decomposition of the metric given in Sec. II. We find that the Maxwell equations can be expressed in the same form as in flat space-time, though, among other things, in the present case the three-dimensional space can be curved. The signature of the space–time metric is taken as $(-\,+\,+\,+)$, lower case Greek indices run from 0 to 3, and lower case Latin indices run from 1 to 3.

II. DECOMPOSITION OF THE SPACE–TIME METRIC

As shown in Ref. 2, the geodesic equation for a test particle with a nonvanishing rest mass in a stationary space–time is equivalent to the equations of motion of a charged (nonrelativistic)

[a)]Electronic mail: gtorres@fcfm.buap.mx

particle in a, possibly curved, three-dimensional space in the presence of a static magnetic field and a conservative field of force. Indeed, if $ds^2 = g_{\alpha\beta}\, dx^\alpha\, dx^\beta$ is the metric of a stationary space–time expressed in a coordinate system such that

$$\frac{\partial g_{\alpha\beta}}{\partial x^0} = 0, \tag{1}$$

we have

$$\begin{aligned} ds^2 &= g_{00}(dx^0)^2 + 2g_{0i}\, dx^0\, dx^i + g_{ij}\, dx^i\, dx^j \\ &= g_{00}\left(dx^0 + \frac{g_{0i}}{g_{00}} dx^i \right)^2 + \left(g_{ij} - \frac{g_{0i}g_{0j}}{g_{00}} \right) dx^i\, dx^j \\ &= g_{00}[(dx^0 - \mathcal{A}_i\, dx^i)^2 - \gamma_{ij}\, dx^i\, dx^j], \end{aligned} \tag{2}$$

with

$$\mathcal{A}_i \equiv \frac{g_{0i}}{(-g_{00})}, \quad \gamma_{ij} \equiv (-g_{00})^{-2}(g_{0i}g_{0j} - g_{00}g_{ij}). \tag{3}$$

The metric tensor,

$$d\sigma^2 \equiv \gamma_{ij}\, dx^i\, dx^j \tag{4}$$

is conformal to the metric in the three-dimensional space of trajectories of the timelike Killing vector field $\partial/\partial x^0$.[5,6,1] When the space–time is static, there exists a coordinate system such that Eq. (1) holds and $g_{0i} = 0$ (i.e., $\mathcal{A}_i = 0$); then the metric $d\sigma^2$ [Eq. (4)] reduces to the so-called *optical metric* (see, e.g., Ref. 4 and the references cited therein).

Making use of Eqs. (2) and (3), one finds that

$$g_{ij} = (-g_{00})(\gamma_{ij} - \mathcal{A}_i\mathcal{A}_j), \quad g_{0i} = (-g_{00})\mathcal{A}_i\,, \tag{5}$$

and

$$g^{ij} = \frac{\gamma^{ij}}{(-g_{00})}, \quad g^{0i} = \frac{\gamma^{ij}\mathcal{A}_j}{(-g_{00})}, \quad g^{00} = \frac{\gamma^{ij}\mathcal{A}_i\mathcal{A}_j - 1}{(-g_{00})}, \tag{6}$$

where (γ^{ij}) denotes the inverse of (γ_{ij}). Then, a straightforward computation shows that the Christoffel symbols corresponding to the metric tensor $g_{\alpha\beta}$, $\Gamma^\alpha_{\beta\gamma}$, are given by

$$\Gamma^0_{00} = \tfrac{1}{2}\gamma^{ij}\mathcal{A}_i\partial_j \ln\phi, \quad \Gamma^i_{00} = \tfrac{1}{2}\gamma^{ij}\partial_j \ln\phi, \quad \Gamma^i_{0j} = \tfrac{1}{2}\gamma^{ik}(\mathcal{B}_{jk} - \mathcal{A}_j\partial_k \ln\phi), \tag{7}$$

where

$$\phi \equiv \tfrac{1}{2}(-g_{00}), \quad \mathcal{B}_{ij} \equiv \partial_i\mathcal{A}_j - \partial_j\mathcal{A}_i\,, \tag{8}$$

and

$$\begin{aligned} \Gamma^0_{0i} &= \tfrac{1}{2}(\partial_i \ln\phi + \gamma^{jk}\mathcal{B}_{ij}\mathcal{A}_k - \mathcal{A}_i\gamma^{jk}\mathcal{A}_k\partial_j \ln\phi), \\ \Gamma^0_{ij} &= -\tilde{\boldsymbol{\nabla}}_{(i}\mathcal{A}_{j)} + \gamma^{km}\mathcal{A}_k\mathcal{B}_{m(i}\mathcal{A}_{j)} + \tfrac{1}{2}(\mathcal{A}_i\mathcal{A}_j - \gamma_{ij})\gamma^{km}\mathcal{A}_k\partial_m \ln\phi, \\ \Gamma^i_{jk} &= \gamma^i_{jk} + \delta^i_{(j}\partial_{k)}\ln\phi + \tfrac{1}{2}(\mathcal{A}_j\mathcal{A}_k - \gamma_{jk})\gamma^{im}\partial_m \ln\phi + \gamma^{im}\mathcal{B}_{m(j}\mathcal{A}_{k)}\,, \end{aligned} \tag{9}$$

where $\tilde{\boldsymbol{\nabla}}$ denotes the covariant derivative compatible with γ_{ij}, γ^i_{jk} denotes the Christoffel symbols for the metric γ_{ij} (e.g., $\tilde{\boldsymbol{\nabla}}_i\mathcal{A}_j=\partial_i\mathcal{A}_j-\gamma^k_{ij}\mathcal{A}_k$), and the parentheses denote symmetrization on the indices enclosed. Therefore, according to Eqs. (7) and (9), the geodesic equations,

$$0=\frac{d^2x^i}{d\lambda^2}+\Gamma^i_{\alpha\beta}\frac{dx^\alpha}{d\lambda}\frac{dx^\beta}{d\lambda}=\frac{d^2x^i}{d\lambda^2}+\Gamma^i_{00}\left(\frac{dx^0}{d\lambda}\right)^2+2\Gamma^i_{0j}\frac{dx^0}{d\lambda}\frac{dx^j}{d\lambda}+\Gamma^i_{jk}\frac{dx^j}{d\lambda}\frac{dx^k}{d\lambda},$$

where λ is the proper time of the particle, amount to

$$0=\frac{d^2x^i}{d\lambda^2}+\tfrac{1}{2}\gamma^{ij}\partial_j\ln\phi\left(\frac{dx^0}{d\lambda}\right)^2+\gamma^{ik}(\mathcal{B}_{jk}-\mathcal{A}_j\partial_k\ln\phi)\frac{dx^0}{d\lambda}\frac{dx^j}{d\lambda}$$
$$+[\gamma^i_{jk}+\delta^i_j\partial_k\ln\phi+\tfrac{1}{2}(\mathcal{A}_j\mathcal{A}_k-\gamma_{jk})\gamma^{im}\partial_m\ln\phi+\gamma^{im}\mathcal{B}_{mj}\mathcal{A}_k]\frac{dx^j}{d\lambda}\frac{dx^k}{d\lambda}. \tag{10}$$

On the other hand, the condition $g_{\alpha\beta}(dx^\alpha/d\lambda)(dx^\beta/d\lambda)=-1$ (which is equivalent to the definition of the proper time), takes the form [see Eq. (2)]

$$(-g_{00})\left[\left(\frac{dx^0}{d\lambda}-\mathcal{A}_i\frac{dx^i}{d\lambda}\right)^2-\gamma_{ij}\frac{dx^i}{d\lambda}\frac{dx^j}{d\lambda}\right]=1, \tag{11}$$

and since $\partial/\partial x^0$ is a Killing vector field [see Eq. (1)], $g_{0\alpha}(dx^\alpha/d\lambda)\equiv-\epsilon$ is a constant of the motion. Making use of Eq. (5), we find that

$$\epsilon=(-g_{00})\left(\frac{dx^0}{d\lambda}-\mathcal{A}_i\frac{dx^i}{d\lambda}\right), \tag{12}$$

and substituting Eq. (12) into Eq. (11), it follows that

$$\gamma_{ij}\frac{dx^i}{d\lambda}\frac{dx^j}{d\lambda}=\frac{\epsilon^2}{(-g_{00})^2}-\frac{1}{(-g_{00})}. \tag{13}$$

Now, with the aid of Eq. (12), we can eliminate $dx^0/d\lambda$ from Eq. (10), and making use of Eq. (13) we obtain

$$0=\frac{d^2x^i}{d\lambda^2}+\frac{\epsilon}{(-g_{00})}\gamma^{ik}\mathcal{B}_{jk}\frac{dx^j}{d\lambda}+\gamma^i_{jk}\frac{dx^j}{d\lambda}\frac{dx^k}{d\lambda}+\frac{dx^i}{d\lambda}\partial_k\ln\phi\frac{dx^k}{d\lambda}+\frac{1}{2(-g_{00})}\gamma^{im}\partial_m\ln\phi. \tag{14}$$

Introducing a new parameter, τ, in place of the proper time λ by means of

$$d\tau\equiv\frac{d\lambda}{(-g_{00})}, \tag{15}$$

one can get rid of the fourth term on the right-hand side of Eq. (14) and this last equation can be written in the form

$$\frac{d^2x^i}{d\tau^2}+\gamma^i_{jk}\frac{dx^j}{d\tau}\frac{dx^k}{d\tau}=\epsilon\gamma^{ik}\mathcal{B}_{kj}\frac{dx^j}{d\tau}-\gamma^{im}\partial_m\phi, \tag{16}$$

which are the equations of motion of a nonrelativistic particle of unit mass and electric charge ϵ in a three-dimensional space with metric (4), in the presence of a magnetic field with vector potential $\mathcal{A}_i$ and a potential ϕ. If one has found the solution of the equations of motion (16), $x^i=x^i(\tau)$, then Eqs. (12) and (15) yield $dx^0/d\tau=\epsilon+\mathcal{A}_i\,dx^i/d\tau$, which determines x^0 as a function of τ, and the original affine parameter λ is given by $\lambda=\int-g_{00}(x^i(\tau))d\tau$.

In the case of the equations for the null geodesics of ds^2, the right-hand side of Eq. (11) is equal to zero, where λ is now an affine parameter of the geodesic. Then, the last term on the right-hand side of Eqs. (13), (14), and (16) is absent and one recovers the result of Ref. 1. When $\mathcal{A}_i\,dx^i$ is exact, by replacing dx^0 by $dx'^0 \equiv dx^0 - \mathcal{A}_i\,dx^i$, one finds explicitly that the space–time is static and the ''magnetic field,'' $\mathcal{B}_{ij}$, vanishes.

The symmetries of the space–time metric $g_{\alpha\beta}$ do not necessarily correspond to symmetries of the three-dimensional metric γ_{ij}. In fact, with the aid of Eqs. (5), one finds that the Killing equations, $K^\mu\,\partial_\mu g_{\alpha\beta} + 2g_{\mu(\alpha}\partial_{\beta)}K^\mu = 0$, amount to the set of equations

$$\begin{gathered} K^i\partial_i \ln\phi = -2\partial_0(K^0 - \mathcal{A}_iK^i), \\ K^j\,\partial_j\mathcal{A}_i + \mathcal{A}_j\,\partial_iK^j = \partial_0(K^0 - \mathcal{A}_jK^j)\mathcal{A}_i + \partial_iK^0 - \gamma_{ij}\,\partial_0K^j, \\ K^k\,\partial_k\gamma_{ij} + 2\gamma_{k(i}\partial_{j)}K^k = 2\partial_0(K^0 - \mathcal{A}_kK^k)\gamma_{ij} - 2\mathcal{A}_{(i}\gamma_{j)k}\,\partial_0K^k. \end{gathered} \tag{17}$$

However, if $K^i\partial_i$ is a Killing vector field of γ_{ij}, with $\partial_0K^i = 0$, and the Lie derivatives of ϕ and $\mathcal{B}_{ij}\,dx^i \wedge dx^j$ along $K^i\partial_i$ vanish, then Eqs. (17) are satisfied if we take K^0 in such a way that $\partial_0K^0 = 0$ and

$$K^j\,\partial_j\mathcal{A}_i + \mathcal{A}_j\,\partial_iK^j = \partial_iK^0. \tag{18}$$

The integrability conditions of these equations for K^0 are satisfied by virtue of the assumed invariance of $\mathcal{B}_{ij}\,dx^i \wedge dx^j$ under the transformatious generated by $K^i\partial_i$.

For example, the Taub-NUT metric,[7]

$$ds^2 = -\left[\frac{r^2 - 2mr - l^2}{r^2 + l^2}\right]\left\{(dt + 2l\cos\theta\,d\varphi)^2 - \frac{(r^2+l^2)^2\,dr^2}{(r^2 - 2mr - l^2)^2} - \frac{(r^2+l^2)^2(d\theta^2 + \sin^2\theta\,d\varphi^2)}{r^2 - 2mr - l^2}\right\}, \tag{19}$$

where m and l are arbitrary constants, is a stationary solution of the Einstein vacuum field equations in the region where $r^2 - 2mr - l^2 > 0$, which reduces to the Schwarzschild metric when $l = 0$. A comparison of Eqs. (2) and (19) shows that we can take $\mathcal{A}_i\,dx^i = -2l\cos\theta\,d\varphi$ and that the corresponding metric $\gamma_{ij}\,dx^i\,dx^j$ is invariant under rotations, i.e., the three-dimensional vector fields,

$$\begin{gathered} K^i_{(1)}\partial_i = -\sin\varphi\,\partial_\theta - \cot\theta\cos\varphi\,\partial_\varphi, \\ K^i_{(2)}\partial_i = \cos\varphi\,\partial_\theta - \cot\theta\sin\varphi\,\partial_\varphi, \\ K^i_{(3)}\partial_i = \partial_\varphi, \end{gathered} \tag{20}$$

which are the generators of rotations on the two-sphere, are Killing vector fields of $d\sigma^2 = \gamma_{ij}\,dx^i\,dx^j$. Furthermore, the Lie derivative of $\mathcal{B}_{ij}\,dx^i \wedge dx^j = 4l\sin\theta\,d\theta \wedge d\varphi$ and ϕ (which depends on r only) with respect to the vector fields (20) vanish and, therefore, by integrating Eq. (18) one obtains three Killing vector fields of the metric (19), induced by the Killing vector fields of $d\sigma^2$ given by Eqs. (20). It is easy to see that the substitution of Eqs. (20) into Eq. (18) yields (setting the constant of integration equal to zero), $K^0_{(1)} = 2l\,\mathrm{cosec}\,\theta\cos\varphi$, $K^0_{(2)} = 2l\,\mathrm{cosec}\,\theta\sin\varphi$, $K^0_{(3)} = 0$. [This computation is simplified by noticing that on the left-hand side of Eq. (18) are the components of the Lie derivative along $K^i\partial_i$ of the 1-form $\mathcal{A}_i\,dx^i$, hence, in the present case, Eq. (18) reduces to $\pounds_{K^i\partial_i}(-2l\cos\theta\,d\varphi) = dK^0$.] Thus, the vector fields,

$$K^{\alpha}_{(1)}\partial_\alpha = -\sin\varphi\,\partial_\theta - \cot\theta\cos\varphi\,\partial_\varphi + 2l\,\mathrm{cosec}\,\theta\cos\varphi\,\partial_0,$$

$$K^{\alpha}_{(2)}\partial_\alpha = \cos\varphi\,\partial_\theta - \cot\theta\sin\varphi\,\partial_\varphi + 2l\,\mathrm{cosec}\,\theta\sin\varphi\,\partial_0, \tag{21}$$

$$K^{\alpha}_{(3)}\partial_\alpha = \partial_\varphi,$$

are Killing vector fields of the space–time metric (19).

A straightfoward computation shows that the Killing vector fields (21) satisfy the same commutation relations as those of (20); however, for other stationary metrics, the Lie algebra of the Killing vector fields induced in this manner is a central extension of the Lie algebra of the three-dimensional vector fields that leave $d\sigma^2$, $\mathcal{B}_{ij}\,dx^i\wedge dx^j$, and ϕ invariant. As a matter of fact, assuming that $K^i_{(a)}\partial_i (a=1,...,r)$ are Killing vector fields of $d\sigma^2$ that leave $\mathcal{B}_{ij}\,dx^i\wedge dx^j$ and ϕ invariant with $[K^i_{(a)}\partial_i, K^j_{(b)}\partial_j] = c^d_{ab}K^i_{(d)}\partial_i$, one finds

$$[K^\alpha_{(a)}\partial_\alpha, K^\beta_{(b)}\partial_\beta] = c^d_{ab}K^\alpha_{(d)}\partial_\alpha + f_{ab}\partial_0, \tag{22}$$

where

$$f_{ab} \equiv K^i_{(a)}\partial_i K^0_{(b)} - K^i_{(b)}\partial_i K^0_{(a)} - c^d_{ab}K^0_{(d)}. \tag{23}$$

One can show that the f_{ab} are constant making use of Eq. (18).

A simple example of a stationary metric for which the f_{ab} are not all equal to zero is provided by

$$ds^2 = -\frac{Q(y)}{y^2+a^2}\left\{(du - 2ax\,dv)^2 - \frac{(y^2+a^2)^2}{Q(y)^2}[dy^2 + Q(y)(dx^2+dv^2)]\right\}, \tag{24}$$

where $Q(y) = -2My + b$, and a, b and M are real constants.[8] Taking $x^0 = u$, $\phi = Q(y)/[2(y^2+a^2)]$ and $\mathcal{A}_i\,dx^i = 2ax\,dv$ [see Eqs. (2) and (8)], we find that $\mathcal{B}_{ij}\,dx^i\wedge dx^j = 4a\,dx\wedge dv$ and, evidently, $K^i_{(1)}\partial_i = \partial_x$, $K^i_{(2)}\partial_i = \partial_v$, and $K^i_{(3)}\partial_i = x\partial_v - v\partial_x$ are three Killing vector fields of $d\sigma^2$ that leave $\mathcal{B}_{ij}\,dx^i\wedge dx^j$ and ϕ invariant. The only nonvanishing commutators between these vector fields are $[K^i_{(1)}\partial_i, K^j_{(3)}\partial_j] = K^i_{(2)}\partial_i$, and $[K^i_{(2)}\partial_i, K^j_{(3)}\partial_j] = -K^i_{(1)}\partial_i$. Then, from Eq. (18) one obtains $K^0_{(1)} = 2av$, $K^0_{(2)} = 0$, and $K^0_{(3)} = a(x^2 - v^2)$ (setting the constants of integration equal to zero) and, according to Eq. (23), $f_{12} = \partial_x K^0_{(2)} - \partial_v K^0_{(1)} = -2a$; similarly, one finds that $f_{13} = f_{23} = 0$. Thus, $K^\alpha_{(1)}\partial_\alpha = \partial_x + 2av\,\partial_u$, $K^\alpha_{(2)}\partial_\alpha = \partial_v$, and $K^\alpha_{(3)}\partial_\alpha = x\partial_v - v\partial_x + a(x^2-v^2)\partial_u$ are Killing vector fields of the metric (24) induced by Killing vector fields of $d\sigma^2$.

III. FIELD EQUATIONS

The Maxwell equations,

$$\nabla_\beta F^{\alpha\beta} = 4\pi J^\alpha \tag{25}$$

and

$$\partial_\alpha F_{\beta\gamma} + \partial_\beta F_{\gamma\alpha} + \partial_\gamma F_{\alpha\beta} = 0, \tag{26}$$

can be written in terms of the three-dimensional quantities,

$$D^i \equiv (-g_{00})^2 F^{0i}, \quad H^{ij} \equiv (-g_{00})^2 F^{ij}, \tag{27}$$

and

$$E_i \equiv F_{i0}, \quad B_{ij} \equiv F_{ij}. \tag{28}$$

Recalling that $\nabla_\beta F^{\alpha\beta}=(1/\sqrt{|g|})\partial_\beta(\sqrt{|g|}F^{\alpha\beta})$, where $g\equiv\det(g_{\alpha\beta})$, and noting that from Eqs. (5) it follows that

$$g=-(-g_{00})^4\gamma, \tag{29}$$

where $\gamma\equiv\det(\gamma_{ij})$, one finds that Eq. (25) amounts to

$$\frac{1}{\sqrt{\gamma}}\partial_i(\sqrt{\gamma}D^i)=4\pi\tilde{\rho}, \quad \frac{1}{\sqrt{\gamma}}\partial_i(\sqrt{\gamma}H^{ki})-\partial_0 D^k=4\pi\tilde{j}^k, \tag{30}$$

where we have introduced the definitions

$$\tilde{\rho}\equiv(-g_{00})^2J^0, \quad \tilde{j}^k\equiv(-g_{00})^2J^k. \tag{31}$$

Note that $(1/\sqrt{\gamma})\partial_i(\sqrt{\gamma}D^i)=\tilde{\nabla}_i D^i$ and $(1/\sqrt{\gamma})\partial_i(\sqrt{\gamma}H^{ki})=\tilde{\nabla}_i H^{ki}$. According to Eqs. (29) and (31), $\sqrt{\gamma}\tilde{\rho}=\sqrt{|g|}J^0$, therefore $\tilde{\rho}$ indeed corresponds to the charge density, according to the volume defined by (γ_{ij}).

The remaining Maxwell equations [Eqs. (26)] are equivalent to $\epsilon^{ijk}\partial_i B_{jk}=0$ and $\partial_0 B_{ij}+\partial_i E_j-\partial_j E_i=0$ [see Eqs. (28)] or, in terms of the dual of B_{ij} and E_i,

$$B^k\equiv\frac{1}{2\sqrt{\gamma}}\epsilon^{kij}B_{ij}, \quad E^{ij}\equiv\frac{1}{\sqrt{\gamma}}\epsilon^{ijk}E_k, \tag{32}$$

we have

$$\frac{1}{\sqrt{\gamma}}\partial_i(\sqrt{\gamma}B^i)=0, \quad \frac{1}{\sqrt{\gamma}}\partial_i(\sqrt{\gamma}E^{ki})+\partial_0 B^k=0. \tag{33}$$

Making use of Eqs. (5), (6), (27), and (28) one easily finds that

$$E^i=D^i+\mathcal{A}_j H^{ij}, \quad H_{ij}=B_{ij}+E_i\mathcal{A}_j-E_j\mathcal{A}_i, \tag{34}$$

with the indices of the three-dimensional quantities being raised and lowered by means of γ^{ij} and γ_{ij}. Expressions (34) are somewhat similar to those given in Ref. 9, p. 257 (cf. also Ref. 10 and the references cited therein); but when $\mathcal{A}_i=0$ (static case) our formulas yield the simple relations $E_i=D_i$, $H_{ij}=B_{ij}$, and Eqs. (30) and (33) reduce to the expressions obtained in Ref. 4.

It should be stressed that, from the physical point of view, our results are equivalent to those of, e.g., Ref. 9, or of any other formulation of the Maxwell equations. The difference in the mathematical appearance of the equations comes from the use of the three-dimensional metric (γ_{ij}), which does not define the spatial lengths [see Eq. (2)].

If $\mathcal{A}$, $\tilde{\mathbf{j}}$, $\mathbf{E}$, $\mathbf{B}$, $\mathbf{D}$, and $\mathbf{H}$ are the three-dimensional vector fields with components $\mathcal{A}^i$, $\tilde{j}^i$, E^i, B^i, D^i, and H^i (where H^i is the dual of H_{ij}), respectively, then Eqs. (30), (33), and (34) take the form

$$\begin{aligned}\operatorname{div}\mathbf{D}&=4\pi\tilde{\rho}, \quad \operatorname{curl}\mathbf{H}-\partial_0\mathbf{D}=4\pi\tilde{\mathbf{j}},\\ \operatorname{div}\mathbf{B}&=0, \quad \operatorname{curl}\mathbf{E}+\partial_0\mathbf{B}=0\end{aligned} \tag{35}$$

and

$$\mathbf{E}=\mathbf{D}+\mathcal{A}\times\mathbf{H}, \quad \mathbf{H}=\mathbf{B}-\mathcal{A}\times\mathbf{E}. \tag{36}$$

In terms of the complex combinations $\mathbf{E}+i\mathbf{H}$ and $\mathbf{D}+i\mathbf{B}$, Eqs. (35) and (36) can be rewritten as

$$\operatorname{div}(\mathbf{D}+i\mathbf{B})=4\pi\tilde{\rho}, \quad \operatorname{curl}(\mathbf{E}+i\mathbf{H})-i\partial_0(\mathbf{D}+i\mathbf{B})=4\pi i\tilde{\mathbf{j}} \tag{37}$$

and

$$\mathbf{D}+i\mathbf{B}=\mathbf{E}+i\mathbf{H}+i\mathcal{A}\times(\mathbf{E}+i\mathbf{H}), \tag{38}$$

respectively.

As pointed out in Ref. 4, by expressing the Maxwell equations in their usual form one is able to apply some standard techniques to solve them. For instance, for a time-independent solution of the source-free Maxwell equations, from the second equation (37) it follows that there exists locally a complex potential ψ, such that

$$\mathbf{E}+i\mathbf{H}=-\operatorname{grad}\psi, \tag{39}$$

then, from the first equation (37) and Eq. (38), one finds that

$$\tilde{\nabla}^2\psi+i\mathcal{B}\cdot\operatorname{grad}\psi=0, \tag{40}$$

where $\tilde{\nabla}^2$ is the Laplace operator corresponding to the metric $d\sigma^2$ [i.e., $\tilde{\nabla}^2\psi=(1/\sqrt{\gamma})\partial_i(\sqrt{\gamma}\gamma^{ij}\partial_j\psi)$] and the vector field $\mathcal{B}^i$ is the dual of $\mathcal{B}_{ij}$. The presence of the factor i in the second term of the left-hand side of the last equation implies that, if the space–time is not static (i.e., $\mathcal{B}\neq 0$), it may not be possible to have an electrostatic field alone, without a magnetostatic field, and vice versa (see below).

In the case of the space–time metric (19), one finds that Eq. (40) admits separable solutions of the form $\psi=f_j(r)Y_{jm}(\theta,\varphi)$, where the Y_{jm} are spherical harmonics, and f_j obeys the condition

$$(r^2-2mr-l^2)\frac{d}{dr}\left[(r-il)^2\frac{df_j}{dr}\right]-j(j+l)(r-il)^2f_j=0. \tag{41}$$

When $j=0$ (static monopolar field), Eq. (41) immediately gives $df_0/dr=(\alpha+i\beta)/(r-il)^2$, where α and β are real constants, and the nonvanishing components of the field generated by this solution (with $x^1=r$) are given by

$$E^1+iH^1=\frac{1}{\sqrt{4\pi}}\frac{(r^2-2mr-l^2)^2}{(r^2+l^2)^4}[\alpha(r^2-l^2)-2\beta lr+i(\beta(r^2-l^2)+2\alpha lr)], \tag{42}$$

hence, there are no values of α and β for which the field is purely electric or purely magnetic. If one looks for a solution with $j=1$, which could correspond to a uniform field, one finds that the solution of Eq. (41) grows faster than r as $r\to\infty$ if $l\neq 0$; thus, by contrast with the Schwarzschild metric,[4] the presence of the NUT parameter does not allow the existence of an asymptotically uniform field.

In the case of the Schwarzschild metric, which is given by Eq. (19) with $l=0$, we can find solutions to the source-free Maxwell equations analogous to the usual multipole fields in flat space–time (cf. Ref. 10). Since, in the present case, $\mathcal{A}=0$, we have $\mathbf{D}=\mathbf{E}$ and $\mathbf{B}=\mathbf{H}$, therefore, looking for solutions of the form

$$\mathbf{E}=\operatorname{curl}\left(\psi_{\omega j}(r)e^{-i\omega t}Y_{jm}(\theta,\varphi)\frac{\partial}{\partial r}\right), \tag{43}$$

where $\psi_{\omega j}(r)$ is a function to be determined and ω is a constant, div $\mathbf{D}$ is identically equal to zero and from the last equation (35), assuming that $\mathbf{B}$ also has a time dependence of the form $e^{-i\omega t}$, it follows that

$$\mathbf{B}=\frac{1}{i\omega}\operatorname{curl}\mathbf{E}, \tag{44}$$

and div **B** is equal to zero. The nonvanishing components of the fields (43) and (44), with $(x^1,x^2,x^3)=(r,\theta,\varphi)$, are given explicitly by

$$E^2=\frac{\gamma_{11}}{\sqrt{\gamma}}\psi_{\omega j}\,\partial_\varphi Y_{jm}\,e^{-i\omega t},\quad E^3=-\frac{\gamma_{11}}{\sqrt{\gamma}}\psi_{\omega j}\,\partial_\theta Y_{jm}\,e^{-i\omega t} \tag{45}$$

and

$$B^1=\frac{j(j+1)}{i\omega\gamma_{22}}\psi_{\omega j}Y_{jm}e^{-i\omega t},$$

$$B^2=\frac{1}{i\omega\sqrt{\gamma}}\partial_r\left(\frac{\sqrt{\gamma}}{\gamma_{22}}\psi_{\omega j}\right)\partial_\theta Y_{jm}\,e^{-i\omega t}, \tag{46}$$

$$B^3=\frac{1}{i\omega\sqrt{\gamma}}\partial_r\left(\frac{\sqrt{\gamma}}{\gamma_{22}}\psi_{\omega j}\right)\frac{1}{\sin^2\theta}\partial_\varphi Y_{jm}\,e^{-i\omega t}.$$

Then, from curl $\mathbf{H}-\partial_0\mathbf{D}=0$, one obtains the only condition,

$$\frac{d}{dr}\left[\frac{r-2m}{r}\frac{d}{dr}\left(\frac{r}{r-2m}\psi_{\omega j}\right)\right]-\frac{j(j+1)}{r(r-2m)}\psi_{\omega j}+\frac{\omega^2r^2}{(r-2m)^2}\psi_{\omega j}=0. \tag{47}$$

This radial equation, written in terms of $r\psi_{\omega j}/(r-2m)$, coincides with the Schrödinger-type equation found by other procedures [see, e.g., Ref. 11, Eq. (21) and Ref. 12, Eq. (20)]. The electromagnetic field given by Eqs. (43)–(46) is a TE multipole field (since $E^1=0$). The TM multipole fields are obtained from Eqs. (43)–(46), replacing **E** by **B** and **B** by −**E**.

Since the ''magnetic field'' $\mathcal{B}_{ij}$ and the potential ϕ are made out of the components of the metric [see Eqs. (8)], the differential relations satisfied by $\mathcal{B}_{ij}$ and ϕ depend on the conditions imposed on the metric. By means of a straightforward computation, making use of Eqs. (7)–(9), one finds that the components of the Ricci tensor of the space–time metric (2) are given by

$$R_{00}=\tfrac{1}{2}\phi^{-1}\tilde{\nabla}^2\phi+\tfrac{1}{4}\mathcal{B}_{ij}\mathcal{B}^{ij}, \tag{48}$$

$$(-g_{00})R^i_0=\tfrac{1}{2}\phi^{-1}\tilde{\boldsymbol{\nabla}}_j(\phi\mathcal{B}^{ij}), \tag{49}$$

$$(-g_{00})^2R^{ij}=\gamma^{ik}\gamma^{jl}[{}^{(3)}R_{kl}-\phi^{-1}\tilde{\boldsymbol{\nabla}}_k\tilde{\boldsymbol{\nabla}}_l\phi+\tfrac{3}{2}\phi^{-2}(\partial_k\phi)(\partial_l\phi)-\tfrac{1}{2}\phi^{-1}(\tilde{\nabla}^2\phi)\gamma_{kl}+\tfrac{1}{2}\mathcal{B}_{km}\mathcal{B}_l{}^m], \tag{50}$$

where ${}^{(3)}R_{ij}$ is the Ricci tensor of the metric $d\sigma^2$. Equations (48)–(50) are analogous to the expressions given in Ref. 5, which are based on a generalized Lewis–Papapetrou form of the space–time metric, and these equations may also be useful in the search for exact solutions of the Einstein field equations (cf. Ref. 6).

From Eqs. (48), (49), and (36) it follows that if $R_{\alpha\beta}=0$, then the electromagnetic field defined by

$$\mathbf{H}=\phi\mathcal{B},\quad \mathbf{E}=-\operatorname{grad}\phi, \tag{51}$$

satisfies the Maxwell equations (35) with $\tilde{\rho}=0$, $\tilde{\mathbf{j}}=0$. This result is a special case of the well-known fact that if K^α is a Killing vector field and $R_{\alpha\beta}=0$, then $F_{\alpha\beta}=\boldsymbol{\nabla}_\alpha K_\beta$ satisfies the source–free Maxwell equations. The electromagnetic field (51) is associated with the Killing vector field $\partial/\partial x^0$. In the case of the Taub-NUT metric, the field (51) coincides with that given by Eq. (42) if we take $\alpha+i\beta=-m+il$.

IV. CONCLUSIONS

The decomposition of the metric of a stationary space–time given by Eq. (2) allows us to deal with the particle dynamics and the Maxwell equations in terms of three-dimensional objects in a convenient way. Owing to the conformal invariance of the Maxwell equations, a similar result applies in the case of the conformally stationary space-times.

As we have shown, some symmetries of the stationary space–times can also be easily obtained from the symmetries of the equivalent three-dimensional objects. The expressions for the components of the Ricci tensor [Eqs. (48)–(50)] are relatively simple and may be useful in the integration of the Einstein equations as they stand or in combination with the spinor formalism (cf. Ref. 5).

ACKNOWLEDGMENTS

One of the authors (J.M.P.) thanks the Departmento de Física Matemática of the Universidad Autónoma de Puebla for its hospitality and the Academia Mexicana de Ciencias for the financial support. The authors would like to thank the referee for helpful comments.

[1] V. Perlick, Class. Quantum Grav. **7**, 1849 (1990).
[2] G. F. Torres del Castillo, Rev. Mex. Fís. **44**, 540 (1998).
[3] M. A. Abramowicz, B. Carter, and J. P. Lasota, Gen. Relativ. Gravit **20**, 1173 (1988).
[4] S. Sonego and M. A. Abramowicz, J. Math. Phys. **39**, 3158 (1998).
[5] Z. Perjés, J. Math. Phys. **11**, 3383 (1970).
[6] M. Marklund and Z. Perjés, J. Math. Phys. **38**, 5280 (1997).
[7] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973).
[8] G. F. Torres del Castillo and J. C. Pérez-Pérez, Phys. Rev. D **58**, 47 501 (1998).
[9] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 4th ed. (Pergamon, Oxford, 1975).
[10] I. Brevik and S. Straumsnes, Gen. Relativ. Gravit. **29**, 837 (1997).
[11] G. F. Torres del Castillo, J. Math. Phys. **37**, 5684 (1996).
[12] G. Cognola and P. Lecca, Phys. Rev. D **57**, 1108 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Hydrodynamic behavior of Brownian particles in a position-dependent constant force field

C. Barbachoux and F. Debbasch
Laboratoire de Radioastronomie, E.N.S., 24 rue Lhomond, F-75231 Paris Cedex 05, France

J. P. Rivet
C.N.R.S., Laboratoire G. D. Cassini, Observatoire de Nice, F-06304 Nice Cedex 04, France

(Received 6 October 1998; accepted for publication 26 January 1999)

The diffusion equation in physical space–time for a Brownian particle driven by an external force field has been derived by Smoluchowski in the two particular cases where the external field is uniform or varies linearly with position (elastic force). In more general cases, correction terms must be added to the Smoluchowski equation. We show here how to use a multi-scale Chapman–Enskog expansion to obtain, in the hydrodynamic limit, the first corrective terms to the Smoluchowski equation, without any restriction on the friction coefficient, and for any sufficiently small position-dependent constant force field. We also compare our approach with the works of Wilemski, Titulaer, and van Kampen. © *1999 American Institute of Physics.* [S0022-2488(99)00806-3]

I. INTRODUCTION

In 1915, Smoluchowski[1,2] raised the important question of how the usual diffusion equation should be modified in order to describe properly the collective motion of Brownian particles under the influence of a given external force field. He proved that, for at least two special choices of the force field, the diffusion process could be described by a common new equation, which now carries the name of the Austrian physicist. Some 60 years later, the problem was reconsidered by various authors,[3–9] who applied to Kramers' equation an asymptotic procedure similar in spirit to the Chapman–Enskog expansion introduced for solving the Boltzmann equation. The Smoluchowski equation could then be recovered as the first approximation to a more general diffusion equation and higher-order correction terms have also been derived.[3,4]

We believe that these important results may be improved in at least one significant direction. First, for the problem at hand, the most general asymptotics susceptible to a treatment by a Chapman–Enskog expansion actually involve, in the one-dimensional case, three *a priori* distinct small parameters; however, the results presented so far in the literature have always been obtained under the tacit assumption that a single infinitesimal quantity is sufficient to properly address the problem. The principal aim of this article is to propose a fresh investigation of this matter. In particular, we will present, using as examples cases where the external force field is sufficiently small, a new implementation of the Chapman–Enskog procedure which will hopefully make clearer the number and physical significance of the involved small parameters. An approach similar in spirit to the one adopted here has already been used successfully to derive a diffusion equation from the relativistic Ornstein–Uhlenbeck process.[10,11] We will also deduce from our investigations the first correction terms to the Smoluchowski equation for at least one possible set of asymptotics (see Secs. IV and V).

This article is organized as follows. In Sec. II, we review rapidly some fundamentals about the dynamical model of diffusion which will be used in the rest of this work. In Sec. III we elaborate on the general philosophy underlying the Chapman–Enskog expansion and present the various asymptotics considered in this article. Section IV is devoted to obtaining perturbative solutions to

Kramers' equation along the lines presented in the preceding section. In Sec. V we derive the actual ''reduced'' transport equation corresponding to all solutions obtained in Sec. IV. The work in Secs. IV and V is restricted to the few perturbation orders which are necessary to obtain the first correction terms to the Smoluchowski equation and to compare them, in Sec. VI, with those presented in the existing literature. In the conclusion of the article, we review rapidly our main results and mention some problems left open for further study.

II. FUNDAMENTALS

A. The microscopic model of diffusion

To describe the motion of a Brownian particle of mass m under the influence of a given force field $\mathbf{f}$, we start with the Langevin-like system:

$$\frac{d}{dt}\mathbf{x}=\mathbf{v},$$
$$\frac{d}{dt}\mathbf{v}=-\alpha\mathbf{v}+\frac{1}{m}\mathbf{G}+\frac{1}{m}\mathbf{f}. \tag{1}$$

Here, $\mathbf{x}$ and $\mathbf{v}$ are respectively the instantaneous position and velocity of the particle, α is a constant positive coefficient, and $\mathbf{G}$ is a stochastic force. The frictionlike term $-\alpha\mathbf{v}$ and the stochastic force $\mathbf{G}$ model the collisional interaction of the Brownian particle with the particles of the surrounding fluid. As usual, we will suppose that $\mathbf{G}$ is actually a centered Gaussian white noise or, somewhat more precisely, the derivative of a Wiener process multiplied by a constant coefficient $(m^2\alpha^2\chi)$, so that

$$\langle G_i(t_1)G_j(t_2)\rangle=+2m^2\alpha^2\chi\delta(t_2-t_1)\delta_{ij}, \quad \chi>0.$$

(The coefficient χ is the diffusion coefficient in physical space; see Sec. IV.)

Let $\Pi(t,\mathbf{x},\mathbf{v})$ be the probability distribution function (in phase space) associated to the stochastic process defined by (1). It can be shown,[5,10,12] that Π satisfies the following differential equation, known as Kramers' equation:

$$\partial_t\Pi+\nabla_{\mathbf{x}}\cdot(\mathbf{v}\Pi)+\nabla_{\mathbf{v}}\cdot(-\alpha\mathbf{v}\Pi)+\nabla_{\mathbf{v}}\cdot\left(\frac{1}{m}\mathbf{f}\Pi\right)=\alpha^2\chi\Delta_{\mathbf{v}}\Pi. \tag{2}$$

[In Ref. 10, the derivation of Kramers' equation has been carried out in the (special) relativistic framework, which straightforwardly degenerates, in the proper limit, into the Galilean case.]

B. Dimensionless Kramers' equation

To lighten further algebraic manipulations, we feel it convenient to express the external force $\mathbf{f}$ and the variables t, $\mathbf{x}$, and $\mathbf{v}$ in terms of natural units of force, time, position, and velocity.

The natural time unit that comes directly out of (1) is α^{-1}. It represents the typical microscopic relaxation time of the stochastic process. We therefore choose as dimensionless time variable:

$$\underline{t}\equiv\alpha t.$$

The typical ''thermal'' velocity $\sqrt{\alpha\chi}$ will be chosen as velocity unit. The dimensionless velocity variable is, consequently,

$$\underline{\mathbf{v}}\equiv\frac{1}{\sqrt{\alpha\chi}}\mathbf{v}.$$

A temperature T can be defined by $k_{\mathrm{B}}T = m\alpha\chi$, so that the thermal velocity $\sqrt{\alpha\chi}$ is given by the usual expression $\sqrt{k_B T/m}$ (k_{B} is the Boltzmann constant).

The natural space unit $\sqrt{\chi/\alpha}$ is simply the ratio of the velocity unit to the time unit. The dimensionless position variable is then

$$\underline{\mathbf{x}} \equiv \sqrt{\frac{\alpha}{\chi}}\,\mathbf{x}.$$

From the mass m and the natural units of space and time defined above, it is straightforward to obtain the natural unit of force: $m\sqrt{\alpha^3\chi}$. The dimensionless external force is thus defined as

$$\underline{\mathbf{f}} \equiv \frac{1}{m\sqrt{\alpha^3\chi}}\,\mathbf{f}.$$

In terms of these dimensionless variables, Kramers' equation reads

$$\partial_{\underline{t}}\Pi + \nabla_{\underline{\mathbf{x}}}\cdot(\underline{\mathbf{v}}\Pi) - \nabla_{\underline{\mathbf{v}}}\cdot(\underline{\mathbf{v}}\Pi) + \nabla_{\underline{\mathbf{v}}}\cdot(\underline{\mathbf{f}}\Pi) = \Delta_{\underline{\mathbf{v}}}\Pi. \tag{3}$$

C. Hypotheses and restrictions

Throughout this article, we will make the following assumptions:

(1) When $|\mathbf{v}|$ tends to infinity, the probability distribution $\Pi(t,\mathbf{x},\mathbf{v})$ and all its derivatives with respect to $\mathbf{v}$ vanish more rapidly than any power of $\mathbf{v}$, for all time and position:

$$\forall k,l,t,\mathbf{x},\quad \lim_{|\mathbf{v}|\to\infty} \mathbf{v}^k \partial_{v^l}\Pi(t,\mathbf{x},\mathbf{v}) = 0.$$

(2) For technical simplicity reasons, we restrict our study to the one-dimensional case. We therefore work with a one-dimensional probability distribution $\Pi(t,x,v)$ and with a one-dimensional version of (2):

$$\partial_t\Pi + \partial_x(v\Pi) - \partial_v(\alpha v\Pi) + \partial_v\left(\frac{1}{m} f\Pi\right) - \alpha^2\chi\partial_{vv}\Pi = 0. \tag{4}$$

The dimensionless form of (4) reads

$$\partial_{\underline{t}}\Pi + \partial_{\underline{x}}(\underline{v}\Pi) - \partial_{\underline{v}}(\underline{v}\Pi) + \partial_{\underline{v}}(\underline{f}\Pi) - \partial_{\underline{vv}}\Pi = 0. \tag{5}$$

(3) The typical linear size $\mathcal{L}$ of the accessible region in physical space will be assumed to be finite although very large compared to any physically relevant length scale of the problem. This gives a well-defined meaning to the notion of uniform particle-density in physical space.

(4) The probability distribution $\Pi(t,x,v)$, the external force f, and all their derivatives are supposed to exist for any value of $(t>0,x,v)$.

The main points of this article are, first, to present a new implementation of the Chapman–Enskog method applied to Kramers' equation, and, second, to derive from (4), under the above-listed hypotheses, an evolution equation for the particle density in physical space, in the so-called ''hydrodynamic'' limit.

D. The momentum hierarchy

Let us now consider the hierarchy of (evolution-) equations obtained by multiplying (5) by $\underline{v}^k$, $k\in\mathbb{N}$ and integrating the result over velocity space. Using the hypotheses presented in Sec. II C, the level k of the hierarchy takes the form

$$\partial_{\underline{t}}(n\langle\underline{v}^k\rangle) + \partial_{\underline{x}}(n\langle\underline{v}^{k+1}\rangle) + kn\langle\underline{v}^k\rangle - k\underline{f}n\langle\underline{v}^{k-1}\rangle - k(k-1)n\langle\underline{v}^{k-2}\rangle = 0,\quad \text{for } k\geqslant 0, \tag{6}$$

where the symbol $\langle\rangle$ designates the average over $\underline{v}$:

$$n\langle\psi\rangle\equiv\int_{\mathbb{R}}\psi(\underline{v})\Pi(\underline{t},\underline{x},\underline{v})\,d\underline{v}.$$

In particular, the levels $k=0$ and $k=1$ provide the balance equations for the particle and momentum density:

$$\begin{gathered}\partial_{\underline{t}}(n)+\partial_{\underline{x}}(n\langle\underline{v}\rangle)=0,\\ \partial_{\underline{t}}(n\langle\underline{v}\rangle)+\partial_{\underline{x}}(n\langle\underline{v}^2\rangle)+n\langle\underline{v}\rangle-\underline{f}n=0.\end{gathered}\tag{7}$$

The hierarchy (6) will be of crucial importance in establishing the possible asymptotic behaviors of the system in the next section.

III. THE PRINCIPLE BEHIND THE CHAPMAN–ENSKOG EXPANSION AND THE VARIOUS POSSIBLE ASYMPTOTICS

The general idea behind the Chapman–Enskog expansion as we implement it in this article is to solve perturbatively the dimensionless Kramers' equation (5) by searching for ''slowly varying'' solutions in space and time which correspond to a given spatial probability density $n(\underline{t},\underline{x})$. To be more specific, let us choose a ''sufficiently regular'' function $n(\underline{t},\underline{x})$ and try to find solutions of (5) which give back this density when integrated over the whole velocity space. The evolution equations for the various macroscopic ''hydrodynamical'' quantities will appear in this perspective as solvability conditions.

It is to be noted that this whole approach of solving Kramers' equation differs fundamentally from the more usual ''physical'' one which consists in fixing initial and boundary conditions and using them to generate the solution at all (subsequent) time. If one follows this more traditional approach, one is naturally led to distinguish, for example, between a transient regime and long time behavior. Both types of solutions appear, however, on equal footing if one uses the Chapman–Enskog method, since the very notion of ''initial condition'' is absent from the whole formalism. From the Chapman–Enskog point of view the data of the problem are Kramers' equation, and a given spatial probability density *at all positive time*.

In the special case where $\underline{f}$ vanishes identically, a possible solution of (5) corresponding to the constant density $n(\underline{t},\underline{x})=n_0$ is the global equilibrium distribution:

$$\Pi^{(\mathrm{eq})}(\underline{v})=n_0\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\underline{v}^2}{2}\right).\tag{8}$$

It seems therefore reasonable, for densities $n(\underline{t},\underline{x})$ which vary sufficiently slowly in space and time, to search for possible solutions of Kramers' equation in the form of an expansion around the local equilibrium distribution (9), at least if the force field $\underline{f}$ is sufficiently small and has also slow spatial variations:

$$\Pi^{(\mathrm{loc})}(\underline{v})=n(\underline{t},\underline{x})\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\underline{v}^2}{2}\right).\tag{9}$$

Without restricting for the moment the choice of $\underline{f}$, let us pick two ''small parameters'' (i.e., infinitesimal quantities) ϵ and η, and consider a spatial density $n(\underline{t},\underline{x})$ which, in the domain of space–time under consideration, verifies the scaling relations:

$$\frac{\partial_{\underline{x}}n}{n}=\mathcal{O}(\epsilon),\quad \frac{\partial_{\underline{t}}n}{n}=\mathcal{O}(\eta).\tag{10}$$

These relations state that the spatial variation scale of n is large compared to the mean free path $\sqrt{k_B T/m\alpha^2}$, and that its time variation scale is large compared to the mean flight time α^{-1}. The two parameters ϵ and η are *a priori* independent. However, in the degenerated case of free diffusion $(\mathbf{f}=0)$, both parameters are naturally linked by $\eta=\epsilon^2$ (see the discussion in Sec. VI B).

To make proper use of these relations, it is best to turn our attention again to the hierarchy (6). Because $n\langle \underline{v}^{2k+1}\rangle, k\in\mathbb{N}$, vanishes for the local equilibrium distribution (9), let us introduce a new ''small parameter'' ϵ' and search for solutions of Kramers' equation which verify

$$n\langle \underline{v}\rangle=\mathcal{O}(\epsilon'). \tag{11}$$

If one supposes, in agreement with the preceding paragraph, that the sought-for distribution function can be written as an expansion around (9), a simple integration by part shows that the scaling (11) is actually valid for any odd power of the velocity:

$$n\langle \underline{v}^{2k+1}\rangle=\mathcal{O}(\epsilon'), \quad k\in\mathbb{N}. \tag{12}$$

The level $k=0$ in the hierarchy (6) then delivers immediately that $\eta=\epsilon\epsilon'$ and the level $k=1$ implies that $\underline{f}$ has also to be a vanishingly small quantity, whose order will be hereafter denoted by ν. One is therefore left with the following possibilities for balancing properly the terms in (7):

$$\begin{aligned}&\text{(i)}\quad \epsilon'=\epsilon \quad\text{and}\quad \nu\leqslant\epsilon,\\ &\text{(ii)}\quad \nu=\epsilon \quad\text{and}\quad \epsilon'\leqslant\epsilon,\\ &\text{(iii)}\quad \nu=\epsilon' \quad\text{and}\quad \epsilon\leqslant\nu.\end{aligned} \tag{13}$$

Each one of these alternatives defines a particular family of solutions of Eq. (5). It is already apparent at this stage that each family subdivides into at least two subfamilies; the first one corresponds to a strict inequality in (13) and the second one to the case in which all three small parameters ϵ, ϵ', and ν are actually identical. The situation is, however, more complicated than this because ϵ, ϵ', $\eta=\epsilon\epsilon'$ and ν are not the only infinitesimal quantities *a priori* involved in the problem. Indeed, (7) also constraints the spatial variations of n with respect to those of $\underline{f}$. To investigate this matter further, it is convenient to introduce a fifth (infinitesimal) quantity ϵ'' such that

$$\frac{\partial_{\underline{x}}\underline{f}}{\underline{f}}=\mathcal{O}(\epsilon''), \tag{14}$$

and to derive the level k of the hierarchy (6) with respect to $x^p, p\in\mathbb{N}$:

$$\begin{aligned}&\partial_{\underline{t}}\partial_{\underline{x}^p}(n\langle \underline{v}^k\rangle)+\partial_{\underline{x}^{p+1}}(n\langle \underline{v}^{k+1}\rangle)-k\partial_{\underline{x}^p}(\underline{f}n\langle \underline{v}^{k-1}\rangle)\\ &\quad =k(k-1)\partial_{\underline{x}^p}(n\langle \underline{v}^{k-2}\rangle)-k\partial_{\underline{x}^p}(n\langle \underline{v}^{k}\rangle)\quad \text{for } k\geqslant 0.\end{aligned} \tag{15}$$

If k is even, then both terms on the right-hand side of (15) are of order ϵ^p and the first two terms on the left-hand side are clearly of a higher order. Let us suppose that ϵ'' is strictly superior to ϵ. Then, the main contribution to the third term on the left-hand side of (15) is of order $\nu\epsilon''^p$ and this quantity has to be inferior to ϵ^p, for all even integers p. Let us now specialize the discussion according to the family under consideration. For family (i), one can always find a non-negative integer q such that

$$\epsilon^{q+1}\leqslant\nu\leqslant\epsilon^q.$$

This implies that, for all even integers p,

TABLE I. The definition of all the infinitesimal parameters introduced in this article are recalled here.

Quantity	$\partial_x \Pi/\Pi$	$\partial_t \Pi/\Pi$	$n\langle v^{2k+1}\rangle$	f	$\partial_x f/f$
Order	$\mathcal{O}(\epsilon)$	$\mathcal{O}(\eta)$	$\mathcal{O}(\epsilon')$	$\mathcal{O}(\nu)$	$\mathcal{O}(\epsilon'')$

$$\epsilon'' \leqslant \epsilon^{(p-q-1)/p},$$

which obviously contradicts the hypothesis $\epsilon'' > \epsilon$. Similarly, for family (iii), one can again introduce a non-negative integer q verifying

$$\epsilon^{-q} \leqslant \nu \leqslant \epsilon^{-(q+1)},$$

so that, for all even integers p, one has

$$\epsilon'' \leqslant \epsilon^{(p+q)/p},$$

which again contradicts the hypothesis $\epsilon'' > \epsilon$. Finally, for family (ii), one has directly that, for all even integers p,

$$\epsilon'' \leqslant \epsilon^{(p-1)/p},$$

which delivers the same result as the one obtained for the other families. A similar argument for odd values of k delivers the same conclusion. We can therefore conclude that in all cases ϵ has to be superior or equal to ϵ''.

Taking into account all four parameters ϵ, ϵ', ϵ'', and ν as well as their possible relationships to one another, we can sum up the preceding discussion in the following way. There are three main families of solutions which are *a priori* susceptible to a treatment by the Chapman–Enskog method. Each of these families can be subdivided into four different subfamilies of solutions. In each family, the first subfamily encompasses three-parameter solutions, the second and third subfamilies both represent two-parameter solutions, the second subfamily being actually common to all main families. The fourth subfamily is also common to the three families and involves a single infinitesimal quantity. These conclusions are displayed in a more compact form in Tables I and II.

One of the advantages of the Chapman–Enskog procedure is to provide an elegant way to recover the Smoluchowski equation and its various corrections. Indeed, it actually turns out that, for a given spatial density, the whole expansion is only feasible if, at any order, the involved coefficients satisfy various constraints or solvability conditions in the form of partial differential equations. The density and the force f are the only ''data'' of the problem. It therefore follows that all coefficients in the expansion depend only on the density, the imposed force field f, and their various time and space derivatives. Consequently, the constraints satisfied by these coefficients can be transcribed, at any order, into differential equations which are to be verified by the density,

TABLE II. The characteristics of the three families of solutions and of the four subfamilies are summed up here. Each column of this table corresponds to a family, and each line to a subfamily.

	(i)	*(ii)*	*(iii)*
(−.1)	$\epsilon' = \epsilon, \quad \nu < \epsilon, \quad \epsilon'' < \epsilon$	$\nu = \epsilon, \quad \epsilon' < \epsilon, \quad \epsilon'' < \epsilon$	$\nu = \epsilon', \quad \epsilon < \epsilon', \quad \epsilon'' < \epsilon$
(−.2)		$\nu = \epsilon, \quad \epsilon' = \epsilon, \quad \epsilon'' < \epsilon$	
(−.3)	$\epsilon' = \epsilon, \quad \nu < \epsilon, \quad \epsilon'' = \epsilon$	$\nu = \epsilon, \quad \epsilon' < \epsilon, \quad \epsilon'' = \epsilon$	$\nu = \epsilon', \quad \epsilon < \epsilon', \quad \epsilon'' = \epsilon$
(−.4)		$\nu = \epsilon', \quad \epsilon = \epsilon', \quad \epsilon'' = \epsilon$	

the force *f*, and their various derivatives for the whole expansion procedure to exist. These differential equations are actually the evolution equations for the spatial density one is usually looking for.

IV. PERTURBATIVE RESOLUTION OF KRAMERS' EQUATION FOR FAMILY (II)

A. The general framework

Let us now present, in the ''hydrodynamic'' limit, a direct resolution of Kramers' equation by a generalized multi-scale Chapman–Enskog expansion. The dimensionless Kramers' equation reads

$$\partial_{\underline{t}}\Pi+\underline{v}\,\partial_{\underline{x}}\Pi+\underline{f}\,\partial_{\underline{v}}\Pi=\partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})\Pi). \tag{16}$$

Each family and/or subfamily envisaged in Sec. III corresponds to a different physical situation (see Table I). A complete examination of the problem at hand should therefore involve a detailed study of eight different cases. This lies clearly outside the scope of the present article; we will now present the perturbative solutions to Kramers' equation belonging to one of the three main families only. Let us point out, however, that preliminary calculations seem to indicate that the major steps in the Chapman–Enskog procedure are similar for all three families. There is no physical or mathematical *a priori* reason to prefer one family to the other two. In what follows, we will concentrate on family (ii).

B. The asymptotic expansion for subfamily (ii.1)

Using the small parameters presented in Sec. III, we introduce new rescaled space, time, and force variables $X=\epsilon\underline{x}, Y=\epsilon''\underline{x}, T=\epsilon\epsilon'\underline{t}, \underline{f}=\nu F$ and rewrite (16) as

$$\epsilon\epsilon'\,\partial_T\Pi+\underline{v}\,(\epsilon\partial_X\Pi+\epsilon''\partial_Y\Pi)+\nu F\,\partial_{\underline{v}}\Pi=\partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})\Pi). \tag{17}$$

A reasonable form for the Chapman–Enskog expansion corresponding to subfamily (ii.1) is

$$\Pi=\sum_{k,l,m\in\mathbb{N}^3}\epsilon^k\left(\frac{\epsilon'}{\epsilon}\right)^l\left(\frac{\epsilon''}{\epsilon}\right)^m\Pi_{klm}, \tag{18}$$

where Π_{klm} is a function of *X, Y, T*, and $\underline{v}$. Substituting expression (18) in Eq. (17) and collecting all terms of order $\epsilon^k(\epsilon'/\epsilon)^l(\epsilon''/\epsilon)^m$, we obtain

$$\partial_T\Pi_{k-2l-1m}+\underline{v}\,(\partial_X\Pi_{k-1lm}+\partial_Y\Pi_{k-1lm-1})+F\,\partial_{\underline{v}}\Pi_{k-1lm}=\partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})\Pi_{klm}), \tag{19}$$

with the convention that Π_{klm} vanishes for any strictly negative value of either *k, l*, or *m*. We rewrite Eq. (19) in the more condensed form:

$$\partial_T(P\cdot{}^\circ\Pi_{k-2l-1m})+\underline{v}\,\partial\cdot{}^\circ\Pi_{k-1lm}+F\,\partial_{\underline{v}}(P\cdot{}^\circ\Pi_{k-1lm})=\partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})P\cdot{}^\circ\Pi_{klm}), \tag{20}$$

where the vectorlike quantity ${}^\circ\Pi_{klm}$ is defined by: ${}^\circ\Pi_{klm}=\binom{\Pi_{klm}}{\Pi_{klm-1}}$. In (20), P and ∂ represent respectively the adjoint vectors (1; 0) and $(\partial_X;\partial_Y)$. The usefulness of this rather abstract vectorial formalism and of the covariant derivative $\mathcal{D}$ to be introduced in Sec. IV B 2 may not be quite apparent at this stage. The principle advantage provided by these notations is to furnish cumbersome results obtained in this article in the most possible tractable form. The conversion back to more usual notations will be carried out at the end of all calculations in Sec. V A. We will now present in full detail the resolution of (20) for $0\leqslant k\leqslant 4$.

1. Order k=0

Setting $k=0$ in Eq. (20) gives

$$\partial_{\underline{v}}((\partial_{\underline{v}}+\underline{v})P\cdot{}^{\circ}\Pi_{0lm})=0. \tag{21}$$

The only solutions of (21) compatible with hypothesis 1 are of the form

$${}^{\circ}\Pi_{0lm}={}^{\circ}A_{0lm}e^{-\underline{v}^2/2}, \tag{22}$$

where ${}^{\circ}A_{0lm}$ is a function of *X, Y*, and *T*. The A_{0lm}'s and the other similar functions to be introduced below are not arbitrary. Their link to the spatial density n will be discussed in Sec. IV F.

2. Order k=1

Setting k equal to 1 in (5) and using the expression (22) for ${}^{\circ}\Pi_{0lm}$, we obtain

$$\underline{v}\mathcal{D}\cdot{}^{\circ}A_{0lm}e^{-\underline{v}^2/2}=\partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})P\cdot{}^{\circ}\Pi_{1lm}), \tag{23}$$

where $\mathcal{D}$ is a ''covariant'' derivative defined by $\mathcal{D}=\partial-FP$. Hypothesis 1 leads us to retain, as only solutions of (23),

$${}^{\circ}\Pi_{1lm}=\{{}^{\circ}A_{1lm}-\underline{v}{}^{\circ}(\mathcal{D}\cdot{}^{\circ}A_{0lm})\}e^{-\underline{v}^2/2}, \tag{24}$$

where ${}^{\circ}A_{1lm}$ is a function of *X, Y*, and *T*. Up to this order, no solvability condition has to be imposed to obtain solutions of (20) verifying hypothesis 1.

3. Order k=2

We set k equal to 2 in (20) and use for ${}^{\circ}\Pi_{0lm}$ and ${}^{\circ}\Pi_{1lm}$ the expressions (22) and (24) to obtain

$$\begin{aligned}&\{\partial_T(P\cdot{}^{\circ}A_{0l-1m})-F(P\cdot{}^{\circ}(\mathcal{D}\cdot{}^{\circ}A_{0lm}))+\underline{v}(\mathcal{D}\cdot{}^{\circ}A_{1lm})-\underline{v}^2\mathcal{D}\cdot{}^{\circ}(\mathcal{D}\cdot{}^{\circ}A_{0lm})\}e^{-\underline{v}^2/2}\\&\quad=\partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})P\cdot{}^{\circ}\Pi_{2lm}).\end{aligned} \tag{25}$$

The commutation relation $[\mathcal{D};P]=0$ has also been used in deriving (25). To satisfy hypothesis 1, the integral over $\underline{v}$ on the left-hand side must vanish. Indeed, let us take the primitive of both sides of (25). The primitive of the right-hand side clearly vanishes when $\underline{v}$ tends to infinity if hypothesis 1 is verified. Thus, the primitive of the left-hand side must also vanish when $\underline{v}$ tends to infinity. Hence, the integral over $\underline{v}$ of the left-hand side must be zero. We thus find that a necessary (and sufficient) condition for the integral over $\underline{v}$ of the left-hand side of (25) to vanish is

$$\partial_T(P\cdot{}^{\circ}A_{0l-1m})-\partial\cdot{}^{\circ}(\mathcal{D}\cdot{}^{\circ}A_{0lm})=0. \tag{26}$$

Assuming this solvability condition to be fulfilled, the only solutions of (25) which verify hypothesis 1 are

$${}^{\circ}\Pi_{2lm}=\left\{{}^{\circ}A_{2lm}-\underline{v}{}^{\circ}(\mathcal{D}\cdot{}^{\circ}A_{1lm})+\frac{\underline{v}^2}{2}{}^{\circ}(\mathcal{D}\cdot{}^{\circ}(\mathcal{D}\cdot{}^{\circ}A_{0lm}))\right\}e^{-\underline{v}^2/2}, \tag{27}$$

where ${}^{\circ}A_{2lm}$ is again a function of *X, Y*, and *T*.

4. Order k=3

With $k=3$ in (20) and the expressions (24) and (27) for ${}^{\circ}\Pi_{1lm}$ and ${}^{\circ}\Pi_{2lm}$, respectively, we obtain

$$\left\{ \partial_T(P\cdot{}^\circ A_{1l-1m}) - F(P\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{1lm})) + \underline{v}(\mathcal{D}\cdot{}^\circ A_{2lm} - {}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}) - \mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}))) \right.$$
$$\left. -\underline{v}^2(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{1lm})) + \frac{\underline{v}^3}{2}(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}))) \right\} e^{-\underline{v}^2/2} = \partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})P\cdot{}^\circ\Pi_{3lm}), \tag{28}$$

where the condensed notation ${}^\circ[\mathcal{D};\partial]\cdot{}^\circ A$ stands for

$$^\circ[\mathcal{D};\partial]\cdot{}^\circ A = \mathcal{D}\cdot{}^\circ(\partial\cdot{}^\circ A) - \partial\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A).$$

Note that the straightforward commutation relations $[\mathcal{D};P]=0$ and $[\partial_T;\mathcal{D}]=0$ have been used in deriving (28). The solvability condition associated to (28) is

$$\partial_T(P\cdot{}^\circ A_{1l-1m}) - \partial\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{1lm}) = 0. \tag{29}$$

The corresponding solutions compatible with hypothesis 1 are

$$^\circ\Pi_{3lm} = \left\{ {}^\circ A_{3lm} - \underline{v}({}^\circ(\mathcal{D}\cdot{}^\circ A_{2lm}) - {}^\circ({}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}))) + \frac{\underline{v}^2}{2}{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{1lm})) \right.$$
$$\left. - \frac{\underline{v}^3}{6}{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}))) \right\} e^{-(\underline{v}^2/2)}, \tag{30}$$

where ${}^\circ A_{3lm}$ is a function of *X*, *Y*, and *T*.

5. Order k=4

If we set $k=4$ in (20) and give to ${}^\circ\Pi_{2lm}$ and ${}^\circ\Pi_{3lm}$ the forms (27) and (30), we have

$$\left\{ \partial_T(P\cdot{}^\circ A_{2l-1m}) - F(P\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{2lm})) + \partial\cdot{}^\circ({}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm})) - \mathcal{D}\cdot{}^\circ({}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm})) \right.$$
$$+\underline{v}(\mathcal{D}\cdot{}^\circ A_{3lm} - {}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{1lm}) - \mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{1lm}))) - \underline{v}^2\Big(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{2lm})$$
$$-\mathcal{D}\cdot{}^\circ({}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm})) - \tfrac{1}{2}{}^\circ[\mathcal{D}^2;\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}) - \tfrac{1}{2}\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm})))\Big)$$
$$\left. + \frac{\underline{v}^3}{2}(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{1lm}))) - \frac{\underline{v}^4}{6}(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm})))) \right\} e^{-(\underline{v}^2/2)}$$
$$= \partial_{\underline{v}}((\underline{v}+\partial_{\underline{v}})P\cdot{}^\circ\Pi_{4lm}). \tag{31}$$

In deriving this expression, we have used the solvability conditions (26) and (29) and the commutation relations $[\mathcal{D};P]=0$ and $[\partial_T;\mathcal{D}]=0$. Following again the logic of Secs. IV B 3 and IV B 4, we obtain the solvability condition:

$$\partial_T(P\cdot{}^\circ A_{2l-1m}) - \partial\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{2lm}) = -\partial\cdot({}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm})) - \tfrac{1}{2}{}^\circ[\mathcal{D}^2;\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}). \tag{32}$$

To obtain the desired corrections terms to the Smoluchowski equation, we will only need at this order the solvability condition (32) and no explicit expression for ${}^\circ\Pi_{4lm}$.

C. The asymptotic expansion for subfamily (ii.2)

The subfamily (ii.1) formally degenerates into (ii.2) if $\epsilon' = \epsilon$. The Chapman–Enskog expansion of Π takes therefore the form

$$\Pi = \sum_{k,m \in \mathbb{N}^2} \epsilon^k \left(\frac{\epsilon''}{\epsilon} \right)^m \Pi_{km}, \tag{33}$$

where

$$\Pi_{km} = \sum_{l \in \mathbb{N}} \Pi_{klm}.$$

The expressions for Π_{klm}, $0 \leqslant k \leqslant 3$, obtained in the preceding section furnish directly by summation over l the form of the desired Π_{km}, $0 \leqslant k \leqslant 3$. The solvability conditions associated to (33) can also be straightforwardly deduced from those given in Sec. IV B and do not need to be explicitly presented here.

D. The asymptotic expansion for subfamily (ii.3)

Contrary to (ii.2), subfamily (ii.3) cannot be directly deduced from subfamily (ii.1). The main reason for this is simply that (ii.1) and (ii.2) involve two spatial scales whereas (ii.3) only involves one. The corresponding dimensionless Kramers' equation is

$$\epsilon\epsilon' \partial_T \Pi + \underline{v}\, \epsilon \partial_X \Pi + \epsilon F \partial_{\underline{v}} \Pi = \partial_{\underline{v}}((\underline{v} + \partial_{\underline{v}})\Pi), \tag{34}$$

where Π is now a function of X, T, and $\underline{v}$ only. As in Secs. IV B and IV C, solutions of (34) are to be sought for under the form

$$\Pi = \sum_{k,l \in \mathbb{N}^2} \epsilon^k \left(\frac{\epsilon'}{\epsilon} \right)^l \Pi_{kl}. \tag{35}$$

Substituting (35) in Eq. (34) and collecting all terms of order $\epsilon^k(\epsilon'/\epsilon)^l$, we obtain

$$\partial_T \Pi_{k-2\,l-1} + \underline{v}\, \partial_X \Pi_{k-1\,l} + F \partial_{\underline{v}} \Pi_{k-1\,l} = \partial_{\underline{v}}((\underline{v} + \partial_{\underline{v}})\Pi_{kl}), \tag{36}$$

with the convention that Π_{kl} vanishes if k or l is strictly negative. There is, however, an important technical difference between (36) and (19). In (19), the index m traced back the presence of two spatial scales and motivated the introduction of the vectorial formalism used in Secs. IV B and IV C. Because (34) involves only one spatial scale, this formalism is of no use for solving (36). Solving (36) is consequently an essentially similar but simpler process than solving (19) for subfamily (ii.1). It seems therefore un-necessary to present here the corresponding cumbersome algebra.

E. The asymptotic expansion for subfamily (ii.4)

Because subfamily (ii.3) degenerates into subfamily (ii.4) for $\epsilon' = \epsilon$, all results pertaining to this last subfamily can be formally deduced from those discussed in Sec. IV D by a summation over l.

F. Choice of the coefficients in the Chapman–Enskog expansion

One usually fixes the coefficients in the Chapman–Enskog expansion by normalizing the lowest-order term to the given spatial density n and by requiring consequently that all superior orders are normalized to zero.[11,13] When applied to subfamily (ii.4), this delivers

$$\int_{\mathbb{R}} \Pi_0 \, d\underline{v} = n, \tag{37}$$

and

$$\int_{\mathbb{R}} \Pi_k \, d\underline{v} = 0 \quad \text{for } k > 0. \tag{38}$$

Equation (37) clearly determines A_0 in terms of n and (38) delivers A_k once the lower-order coefficients A_j, $j<k$, are known.

The implementation of this procedure for subfamilies (ii.3), (ii.2), and (ii.1) is more subtle. Let us start with subfamily (ii.3). There are *a priori* two different ways of fixing the coefficients of the expansion. The first one would be to normalize Π_{00} to n and all other Π_{kl}'s to 0. This would have the direct consequence of actually setting to zero all Π_{k0}'s, for $k>0$, and the corresponding expansion would not involve ϵ' but only ϵ. It is easy to check that such an expansion cannot satisfy Eq. (34). This means that the retained normalization condition is too restrictive and therefore unsuitable to subfamily (ii.3). The correct way of relating the A_{kl}'s to n is to impose

$$\int_{\mathbb{R}} \sum_{l \in \mathbb{N}} \left(\frac{\epsilon'}{\epsilon}\right)^l \Pi_{0l} \, d\underline{v} = n \tag{39}$$

and

$$\int_{\mathbb{R}} \sum_{l \in \mathbb{N}} \left(\frac{\epsilon'}{\epsilon}\right)^l \Pi_{kl} \, d\underline{v} = 0 \quad \text{for } k > 0. \tag{40}$$

Equations (39) and (40) solve the problem at the price of not fixing unambiguously every A_{kl} but only their weighed sum over l: $\Sigma_{l \in \mathbb{N}} (\epsilon'/\epsilon)^l A_{kl}$.

A similar reasoning leads for subfamily (ii.2) to the conditions

$$\int_{\mathbb{R}} \sum_{m \in \mathbb{N}} \left(\frac{\epsilon''}{\epsilon}\right)^m \Pi_{0m} \, d\underline{v} = n \tag{41}$$

and

$$\int_{\mathbb{R}} \sum_{m \in \mathbb{N}} \left(\frac{\epsilon''}{\epsilon}\right)^m \Pi_{km} \, d\underline{v} = 0 \quad \text{for } k > 0, \tag{42}$$

and, for subfamily (ii.1), to

$$\int_{\mathbb{R}} \sum_{l,m \in \mathbb{N}^2} \left(\frac{\epsilon'}{\epsilon}\right)^l \left(\frac{\epsilon''}{\epsilon}\right)^m \Pi_{0lm} \, d\underline{v} = n, \tag{43}$$

and

$$\int_{\mathbb{R}} \sum_{l,m \in \mathbb{N}^2} \left(\frac{\epsilon'}{\epsilon}\right)^l \left(\frac{\epsilon''}{\epsilon}\right)^m \Pi_{klm} \, d\underline{v} = 0 \quad \text{for } k > 0. \tag{44}$$

V. THE FIRST CORRECTION TERMS TO THE SMOLUCHOWSKI EQUATION

In classical Brownian motion theory without external force field, the spatial density n verifies, in the long-time limit, the (dimensionless) diffusion equation:

$$\partial_t n - \partial_{\underline{xx}} n = 0. \tag{45}$$

When an external (time-independent) force field is present, (45) obviously has to be modified. Smoluchowski has proved that, if $\underline{f}$ is uniform[1] or varies linearly with position,[2] the exact transport equation for n is

$$\partial_{\underline{t}} n - \partial_{\underline{xx}} n + \partial_{\underline{x}} (n \underline{f}) = 0. \tag{46}$$

For a more general force field, (46) is only valid approximately. Wilemski[3] and Titulaer[4] have obtained, for a sufficiently small force field, the lowest-order correction terms to (46), which becomes

$$\partial_{\underline{t}} n - \partial_{\underline{xx}} n + \partial_{\underline{x}} (n \underline{f}) = -\partial_{\underline{x}} \{ \partial_{\underline{x}} \underline{f} (\partial_{\underline{x}} n - \underline{f} n) \}. \tag{47}$$

Our approach, at least for family (ii), leads to the same equation for n. We will now present in full detail how to reach (47) for subfamily (ii.1). Subfamilies (ii.2), (ii.3), and (ii.4) will be also briefly discussed.

Let us define Δ as the correction to the Smoluchowski equation (46). We can write

$$\Delta = \partial_{\underline{t}} n - \partial_{\underline{xx}} n + \partial_{\underline{x}} (n \underline{f}). \tag{48}$$

Since all preceding developments involve the phase-space distribution Π, it is convenient to introduce the analog of Δ in phase space, δ, defined by

$$\delta = \partial_{\underline{t}} \Pi - \partial_{\underline{xx}} \Pi + \partial_{\underline{x}} (\Pi \underline{f}). \tag{49}$$

Here Δ can be recovered by a direct integration of δ over the velocity $\underline{v}$.

A. Subfamily (ii.1) and (ii.2)

Using the rescaled space, time, and force variables defined in Sec. IV A, Eq. (49) becomes

$$\delta = \epsilon \epsilon' \partial_T \Pi - (\epsilon^2 \partial_{X^2} \Pi + 2 \epsilon \epsilon'' \partial_{XY} \Pi + (\epsilon'')^2 \partial_{Y^2} \Pi) + \epsilon (\epsilon \partial_X (\Pi F) + \epsilon'' \partial_Y (\Pi F)). \tag{50}$$

The Chapman–Enskog expansion for Π introduced in Sec. IV B leads naturally to the following form for δ:

$$\delta = \sum_{k,l,m \in \mathbb{N}^3} \epsilon^k \left(\frac{\epsilon'}{\epsilon} \right)^l \left(\frac{\epsilon''}{\epsilon} \right)^m \delta_{klm}. \tag{51}$$

Replacing in (50) Π and δ by their respective expressions (18) and (51), we obtain

$$\begin{aligned} \delta_{klm} = {} & \partial_T \Pi_{k-2l-1m} - (\partial_{X^2} \Pi_{k-2lm} + 2 \partial_{XY} \Pi_{k-2lm-1} + \partial_{Y^2} \Pi_{k-2lm-2}) \\ & + F (\partial_X \Pi_{k-2lm} + \partial_Y \Pi_{k-2lm-1}) + F_Y \Pi_{k-2lm-1}. \end{aligned} \tag{52}$$

Because Π_{klm} vanishes for negative values of k, so do δ_{0lm} and δ_{1lm}. With the help of the vectorial formalism introduced in Sec. IV B, Eq. (52) takes the more compact form

$$\delta_{klm} = \partial_T (P \cdot {}^{\circ}\Pi_{k-2l-1m}) - \partial \cdot {}^{\circ} (\mathcal{D} \cdot {}^{\circ}\Pi_{k-2lm}). \tag{53}$$

Making use of Eq. (22) and of the solvability condition (26), Eq. (53) implies that δ_{2lm} vanishes identically for all l and m. This already proves that, at this order, the density n verifies indeed the Smoluchowski equation (46). The correction terms in Equation (47) will now be obtained by evaluating δ_{3lm} and δ_{4lm}. Considering the solvability condition (29) and the expression (24) for ${}^{\circ}\Pi_{1lm}$, we obtain from (53)

$$\delta_{3lm} = -\underline{v} (\partial_T (P \cdot {}^{\circ} (\mathcal{D} \cdot {}^{\circ} A_{0lm})) - \partial \cdot {}^{\circ} (\mathcal{D} \cdot {}^{\circ} (\mathcal{D} \cdot {}^{\circ} A_{0lm}))) e^{-(\underline{v}^2/2)}. \tag{54}$$

After integration over $\underline{v}$, this term provides an identically vanishing contribution to Δ. Finally, the substitution of (27) in (53) leads to

$$\delta_{4lm}=\Big\{\partial_T(P\cdot{}^\circ A_{2l-1m})-\partial\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{2lm})+\frac{\underline{v}^2}{2}(\partial_T(P\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0l-1m})))$$
$$-\partial\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm}))))+K(\underline{v})\Big\}e^{(-\underline{v}^2/2)}, \tag{55}$$

where $K(\underline{v})$ represents a polynomial expression in $\underline{v}$ which only involves odd powers of this variable. $K(\underline{v})$ will therefore not contribute to Δ. With the help of the solvability conditions (26) and (32) and considering the commutation relations $[P;\mathcal{D}]=0$ and $[\partial_T;\mathcal{D}]=0$, (55) becomes, after integration over $\underline{v}$,

$$\int_{\mathbb{R}}\delta_{4lm}\,d\underline{v}=-\sqrt{2\pi}\partial\cdot{}^\circ({}^\circ[\mathcal{D};\partial]\cdot{}^\circ(\mathcal{D}\cdot{}^\circ A_{0lm})). \tag{56}$$

Multiplying (56) by $\epsilon^4(\epsilon'/\epsilon)^l(\epsilon''/\epsilon)^m$ and summing over l and m yields the following expression for Δ:

$$\Delta=-\sqrt{2\pi}\epsilon^4\partial\cdot{}^\circ\left({}^\circ[\mathcal{D};\partial]\cdot{}^\circ\left(\sum_{l,m\in\mathbb{N}^2}\left(\frac{\epsilon'}{\epsilon}\right)^l\left(\frac{\epsilon''}{\epsilon}\right)^m\mathcal{D}\cdot{}^\circ A_{0lm}\right)\right)+\mathcal{O}(\epsilon^5). \tag{57}$$

According to the discussion at the end of Sec. IV, the spatial density is

$$n=\int_{\mathbb{R}}\sum_{l,m\in\mathbb{N}^2}\left(\frac{\epsilon'}{\epsilon}\right)^l\left(\frac{\epsilon''}{\epsilon}\right)^m\Pi_{0lm}\,d\underline{v}. \tag{58}$$

Consequently, the summation over l and m in Eq. (57) can be expressed in terms of the density n:

$$\epsilon\sqrt{2\pi}\sum_{l,m\in\mathbb{N}^2}\left(\frac{\epsilon'}{\epsilon}\right)^l\left(\frac{\epsilon''}{\epsilon}\right)^m(\mathcal{D}\cdot{}^\circ A_{0lm})=\underline{D}n, \tag{59}$$

where the operator $\underline{D}$ is

$$\underline{D}=\partial_{\underline{x}}-\underline{f}.$$

Substituting (59) in (57), we have

$$\Delta=-\partial_{\underline{x}}[\underline{D};\partial_{\underline{x}}](\underline{D}n). \tag{60}$$

Using the identity $[\underline{D};\partial_{\underline{x}}]=\partial_{\underline{x}}f$, (60) leads immediately to the expression (47) for the corrected Smoluchowski equation. We have seen in Sec. IV C that the Chapman–Enskog expansion for subfamily (ii.2) can be formally deduced from the one obtained in Sec. IV B for subfamily (ii.1) by a simple summation over l. By its very definition (51), δ involves a summation over l and therefore Δ does also. The correction terms to the Smoluchowski equation for subfamily (ii.2) are consequently identical to those obtained in this section for subfamily (ii.1).

B. Subfamily (ii.3) and (ii.4)

Contrary to the preceding case, it is not necessary here to introduce a vectorial formalism. The physical reason is that subfamily (ii.1) involves two space scales, whereas subfamily (ii.3) only involves one. The corresponding expression for δ in terms of Π is then

$$\delta=\epsilon\epsilon'\partial_T\Pi-\epsilon^2\partial_{X^2}\Pi+\epsilon^2\partial_X(\Pi F). \tag{61}$$

Considering the form of the Chapman–Enskog expansion for Π introduced in Sec. IV D, we write for δ

$$\delta = \sum_{k,l \in \mathbb{N}^2} \epsilon^k \left(\frac{\epsilon'}{\epsilon} \right)^l \delta_{kl}, \tag{62}$$

With the help of expressions (35) for Π and (62) for δ, (61) becomes

$$\delta_{kl} = \partial_T \Pi_{k-2l-1} - \partial_{X^2} \Pi_{k-2l} + F \partial_X \Pi_{k-2l} + F_X \Pi_{k-2l}. \tag{63}$$

The solution of (36) leads, through (63), to an expression for Δ identical to the one obtained in the preceding section [Eq. (60)]. Finally, since all results obtained in Sec. IV E can be formally deduced from those concerning subfamily (ii.3) by summation over l, Eq. (47) is also valid for subfamily (ii.4).

C. A brief summary of the preceding results

Using expression (60) for the correction Δ, the dimensionless Smoluchowski equation with the first correction terms reads

$$\partial_{\underline{t}} n - \partial_{\underline{xx}} n + \partial_{\underline{x}} (n \underline{f}) + \partial_{\underline{x}} \{ \partial_{\underline{x}} \underline{f} (\partial_{\underline{x}} n - \underline{f} n) \} = 0. \tag{64}$$

In terms of the original physical variables, Eq. (64) can be rewritten as

$$\partial_t n - \chi \partial_{xx} n + \frac{1}{m\alpha} \partial_x (nf) + \frac{\chi}{m\alpha^2} \partial_x \left\{ \partial_x f \left(\partial_x n - \frac{1}{\chi m \alpha} nf \right) \right\} = 0. \tag{65}$$

VI. DISCUSSION

A. Comparison with previous works

We would like now to discuss thoroughly the results presented in both preceding sections and compare them with those already available in the literature. Some typical recent and frequently quoted references on the topic are Refs. 3–9. Actually, Ref. 4 proposes a systematic generalization of the work presented in Ref. 3 and discusses most of the literature before 1978 at great length. Reference 5 is essentially based on Ref. 4; it proposes a derivation of the Smoluchowski equation only, without any correction terms, but it also contains some original very important physical discussions which render its reading essential to any proper evaluation of the issues raised by the present article. References 6–9 elaborate on a relatively different perturbation scheme, where the unperturbed phase space distribution is actually a *drifting* Maxwellian, in contradistinction to the unperturbed local equilibrium retained in the present article and Refs. 3–5. References 6–9 will therefore be discussed at the end of this section, after having compared carefully our results with those of Titulaer[4] and van Kampen.[5]

Titulaer envisages the problem of finding the correct solution to Kramers' equation (2) for given initial conditions. To this end, he considers different classes of solutions, each class being labeled by a non-negative integer n (which, naturally, must not be confused with the notation n used in the present article for the spatial density). In any of these classes, each solution has to be obtained by a Chapman–Enskog expansion about which more will be said later on. [Each class is actually associated to an eigenvalue of the differential operator which appears, e.g., on the right-hand side of Eq. (16); n labels the various eigenvalues of this operator.] After having constructed or explained how to construct at least the first expansion terms for solutions belonging to each class, Titulaer argues rather convincingly that the solutions which belong to a class characterized by a strictly positive integer n will decay exponentially in time on a typical time-scale $(n\alpha)^{-1}$; this apparently remains true even for spatially homogeneous solutions in homogeneous force field. This makes clear that the solutions considered in the present article, which describe the system in the hydrodynamic limit, have to be compared only with the solutions of Ref. 4 labeled by $n=0$.

As already alluded to before, these solutions are obtained by Titulaer (and van Kampen) by means of a Chapman–Enskog expansion which involves a single ''small parameter,'' the inverse of the friction coefficient α. While the corresponding expansion is formally well defined, its physical interpretation does not seem straightforward to us.

Indeed, if one wants to interpret the asymptotics presented in Refs. 4 and 5 by introducing the scale over which the distribution function Π varies in space, one can follow van Kampen (Ref. 5, Remark at the end of p. 218) and obtain the following inequality (recast in our notation):

$$\sqrt{\frac{\chi}{\alpha}}\left|\frac{\partial_x \Pi_0}{\Pi_0}-\frac{f}{m\propto\chi}\right| \ll 1. \tag{66}$$

This essentially states that the small parameter used by Titulaer and van Kampen may be considered as some kind of mixture of two fundamentally different quantities, which are respectively the ratio of the ''mean free path'' $\sqrt{\chi/\alpha}$ to the scale over, which Π varies spatially, and the ratio of the external force field to the natural force unit $m\sqrt{\alpha^3\chi}$. Moreover, both these quantities may be large and their difference may still verify (66). To our eyes, the uneasiness one might feel in trying to interpret physically (66) any further is a sign that the general asymptotics to be used in a Chapman–Enskog expansion applied to Kramers' equation does not *a priori* depend on a single small parameter only. Indeed, (66) alone strongly suggests that it might be wise to consider at least the two independent small parameters $\sqrt{\chi/\alpha}\partial_x\Pi/\Pi$ and $f/m\sqrt{\alpha^3\chi}$, namely, ϵ and ν introduced in Sec. III. As it turns out from the investigation presented in that section and from the results of Sec. IV, the most general collective behavior of point particles diffusing in a time-independent vanishingly small force field involves, in the hydrodynamic limit, three independent small parameters. This general situation can naturally degenerate into various two-parameter problems and even into a one-parameter case. Except in this last situation, the corresponding Chapman–Enskog expansions, presented in part in Sec. IV, are naturally more cumbersome than those proposed by Titulaer and van Kampen.

Let us now compare briefly our results with the work presented in Refs. 6–9. If Refs. 6 and 7 contain very interesting original material, a most useful source for the procedure they introduced seems to us to be Refs. 8 and 9, where some mathematical and physical points are more extensively discussed. What is envisaged in these references is a Chapman–Enskog expansion around a *drifting* Maxwellian, with a drift velocity v_d related to the force f by $v_d=f/\alpha$ (in our notation). Following the procedure of Refs. 3 and 4, the expansion in Refs. 8 and 9 still involves, as single small parameter, the inverse of the friction coefficient α. The possible limitations imposed by this last choice have already been discussed at great length in the preceding paragraph and will not be mentioned again. The choice of a drifting Maxwellian as local ''equilibrium'' around which the expansion is carried out seems to be useful in at least two potentially different (nonexclusive) situations. The first one would involve a non vanishingly small external force field. It is not obvious to us that the choice of v_d retained in Refs. 8 and 9 would then exhaust all physically interesting solutions. In particular, this choice seems to correspond only to an overdamped mean microscopic motion of the particle. Anyhow, the corresponding physics cannot be compared to the one presented in this article, where the force field is supposed to be vanishingly small, as in Refs. 3–5. According to Ref. 9, the choice of a drifting Maxwellian may also be useful in describing transient regimes of at least some solutions of Kramers' equation. From this point of view, it might be interesting to compare the solutions presented in Ref. 9 to those of families (iii.1) and (iii.3) of the present article. As was already discussed before, this cannot be done in the framework of the Chapman–Enskog formalism, and lies therefore outside the scope of this work.

B. An interpretation of the various subfamilies

The aim we shared with Titulaer[4] in performing these voluminous expansions was to derive from them (for a sufficiently small force field) the Smoluchowski equation and at least the first correction terms to it. The fact that Eq. (47) formally agrees with the one obtained by Titulaer (Ref. 4) hides that our derivation is more general than his, and that Eq. (47) is also valid in various

TABLE III. This table gives the orders of magnitude of all the terms in the corrected Smoluchowski equation for all the subfamilies. Each column corresponds to a subfamily, each line to a term.

Term	$(i.1)$	$(ii.1)$	$(iii.1)$	$(-.2)$	$(i.3)$	$(ii.3)$	$(iii.3)$	$(-.4)$
$\partial_t n$	ϵ^2	$\epsilon'\nu$	$\epsilon\nu$	ν^2	ϵ''^2	$\epsilon'\nu$	$\nu\epsilon''$	ν^2
$\partial_{xx}\nu$	ϵ^2	ν^2	ϵ^2	ν^2	ϵ''^2	ν^2	ϵ''^2	ν^2
$f\partial_x n$	$\epsilon\nu$	ν^2	$\epsilon\nu$	ν^2	$\nu\epsilon''$	ν^2	$\nu\epsilon''$	ν^2
$n\partial_x f$	$\nu\epsilon''$	$\nu\epsilon''$	$\nu\epsilon''$	$\nu\epsilon''$	$\nu\epsilon''$	ν^2	$\nu\epsilon''$	ν^2
$\partial_{xx}f\partial_x n$	$\epsilon\nu\epsilon''^2$	$\nu^2\epsilon''^2$	$\epsilon\nu\epsilon''^2$	$\nu^2\epsilon''^2$	$\nu\epsilon''^3$	ν^4	$\nu\epsilon''^3$	ν^4
$nf\partial_{xx}f$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	ν^4	$\nu^2\epsilon''^2$	ν^4
$\partial_x f\partial_{xx}n$	$\epsilon^2\nu\epsilon''$	$\nu^3\epsilon''$	$\epsilon^2\nu\epsilon''$	$\nu^3\epsilon''$	$\nu\epsilon''^3$	ν^4	$\nu\epsilon''^3$	ν^4
$n(\partial_x f)^2$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	$\nu^2\epsilon''^2$	ν^4	$\nu^2\epsilon''^2$	ν^4
$f\partial_x f\partial_x n$	$\epsilon\nu^2\epsilon''$	$\nu^3\epsilon''$	$\epsilon\nu^2\epsilon''$	$\nu^3\epsilon''$	$\nu^2\epsilon''^2$	ν^5	$\nu^2\epsilon''^2$	ν^5

physically different situations not investigated in Ref. 4 [see in particular Eq. (66) and the discussion thereafter]. The physical meaning of expression (60) for Δ can be investigated by evaluating the order of magnitude of the various terms which appear in Eq. (65) for the four different subfamilies (ii.1)–(ii.4). The results are gathered in Table III. This table also indicates the order of magnitude of the same terms for family (i) and (iii), because preliminary calculations indicate that (47) is indeed valid for all three families. We will thus be able to discuss them together via (47) only.

This clearly reveals that, although formally identical for the three families, Eq. (65) actually describes substantially different physics in each case.

There are two main groups of subfamilies. The first one (group A) encompasses the subfamilies (see Table IV) for which the spatial variations of the density $n(\underline{x},\underline{t})$ occur on much smaller scales than those of the external force field, which can therefore be considered as nearly homogeneous ($\epsilon''<\epsilon$). The other group (group B) contains the subfamilies for which the density $n(\underline{x},\underline{t})$ and the external force vary on the same spatial scale ($\epsilon''=\epsilon$).

Each subfamily will be discussed with the help of Table III, retaining in each case the dominant terms in Eq. (64).

1. Free diffusion regime

Subfamilies (i.1) in group A and (i.3) in group B represent cases where the magnitude of the force is so small that (64) practically degenerates into the usual diffusion equation:

$$\partial_{\underline{t}}n-\partial_{\underline{xx}}n=0. \tag{67}$$

TABLE IV. This table sums up the physical discussion presented in Sec. VI B. Each row corresponds to one of the four physical regimes; columns one and two help distinguish between cases where the force can be considered nearly homogeneous or not.

	Quasi-homogeneous force group A ($\epsilon''<\epsilon$)	Heterogeneous force group B ($\epsilon''=\epsilon$)
Free diffusion regime ($\epsilon=\epsilon'$; $\epsilon>\nu$)	Family $(i.1)$ $\partial_t n-\partial_{xx}n=0$	Family $(i.3)$ $\partial_t n-\partial_{xx}n=0$
Barostatic regime ($\epsilon'<\epsilon$; $\epsilon=\nu$)	Family $(ii.1)$ $-\partial_{xx}n+f\partial_x(n)=0$	Family $(ii.3)$ $-\partial_{xx}n+\partial_x(nf)=0$
Over-damped regime ($\epsilon<\epsilon'$; $\epsilon'=\nu$)	Family $(iii.1)$ $\partial_t n+f\partial_x(n)=0$	Family $(iii.3)$ $\partial_t n+\partial_x(nf)=0$
Driven diffusion regime ($\epsilon'=\epsilon$; $\epsilon=\nu$)	Family $(-.2)$ $\partial_t n-\partial_{xx}n+f\partial_x(n)=0$	Family $(-.4)$ $\partial_t n-\partial_{xx}n+\partial_x(nf)=0$

This means that these regimes essentially describe free diffusion phenomena.

2. Barostatic regime

For subfamilies (ii.1) in group A and (ii.3) in group B, Eq. (64) degenerates respectively into

$$-\partial_{\underline{xx}} n + \underline{f} \partial_{\underline{x}}(n) = 0 \tag{68}$$

and

$$-\partial_{\underline{xx}} n + \partial_{\underline{x}}(n \underline{f}) = 0. \tag{69}$$

Quite logically, Eq. (68) is essentially identical to (69) specialized to a situation where the force field is nearly homogeneous. Equation (69) admits as solution

$$n(\underline{x}, \underline{t}) = n_0 \exp(-\Phi(x)),$$

where $\Phi(x) = -\int^x f(y) dy$ is the potential associated to f, and n_0 normalizes n to unity. We therefore call this regime ''barostatic.''

3. Overdamped regime

For subfamilies (iii.1) in group A and (iii.3) in group B, Eq. (64) degenerates respectively into

$$\partial_{\underline{t}} n + \underline{f} \partial_{\underline{x}}(n) = 0 \tag{70}$$

and

$$\partial_{\underline{t}} n + \partial_{\underline{x}}(n \underline{f}) = 0. \tag{71}$$

As before, (70) can be considered as a special case of (71) for nearly uniform force fields. Mathematically, Eq. (71) can be obtained from the microscopic equation of motion (1) by neglecting $(d/dt)\mathbf{v}$. The mean motion of the particle is then an overdamped motion in the force field $\mathbf{f}$.

4. Driven diffusion regime

For subfamily $(-.4)$ in group B, Eq. (64) essentially becomes the standard Smoluchowski equation:

$$\partial_{\underline{t}} n - \partial_{\underline{xx}} n + \partial_{\underline{x}}(\underline{f} n) = 0, \tag{72}$$

and subfamily $(-.2)$ in group A is characterized by a nearly homogeneous force field version of (72):

$$\partial_{\underline{t}} n - \partial_{\underline{xx}} n + \underline{f} \partial_{\underline{x}}(n) = 0. \tag{73}$$

These equations obviously describe regimes where both diffusion and forcing effects are of comparable importance. This justifies the terminology ''driven diffusion.''

VII. CONCLUSION

We have proposed a fresh investigation of the collective motion of point particles which diffuse stochastically under the influence of a time-independent force field. In the hydrodynamic limit, there are three different families of solutions susceptible to a treatment by the Chapman and Enskog expansion method. Each of these families subdivides into four subfamilies characterized by the nature and number of the small parameters involved in the expansion. More precisely, each family contains a class of three- (small) parameter solutions, two different classes of two-parameter solutions and a single class of one-parameter solutions which is actually common to the three main families. We have presented in full the Chapman–Enskog expansions corresponding to the four subcases for one of the three main families. We have then derived from our results, for each subcase, the Smoluchowski equation and the first correction terms to it. These are actually identical for each of the studied subfamilies. We have also discussed at great length the physical significance of our results and compared them with the ones already existing in the literature.

Various extensions of this work are possible and are currently the object of active investigation. One should first of all perform the Chapman–Enskog expansions for all subfamilies corresponding to the two main families which were not dealt with in Sec. IV. Preliminary results indicate that these expansions also lead to the Smoluchowski equation and to the same correction terms as those derived in this article. Considering the apparent ''genericity'' of these terms as well as the natural emergence of covariant derivatives ($\underline{D}$ and $\mathcal{D}$) and commutators in the Chapman–Enskog expansions, it seems to us quite possible that an elegant geometric structure exists behind all the results derived in this paper. A better understanding of the problem in this direction should prove most enlightening. This could be linked with an extension of our results to the case of a time-dependent force field or to the general relativistic realm with the help of the relativistic Ornstein–Uhlenbeck process introduced in Ref. 10.

As already noted, the Chapman–Enskog method does not permit an investigation of the possible transient nature and dynamical stability of the various regimes envisaged in this article. This could be at least partially achieved by direct numerical simulations.

[1] M. Von Smoluchowski, Bulletin international de l'académie des sciences de Cracovie, Mat.-naturw. Klasse A, 418 (1913).
[2] M. Von Smoluchowski, Ann. Phys. **48**, 1103 (1915).
[3] G. Wilemski, J. Stat. Phys. **14** (2), 153 (1976).
[4] U. M. Titulaer, Physica A **91**, 321 (1978).
[5] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1992), 2nd ed.
[6] K. Kaneko, Prog. Theor. Phys. **66**, 129 (1981).
[7] F. Haake, Z. Phys. B **48**, 31 (1982).
[8] U. M. Titulaer, Z. Phys. B **50**, 71 (1983).
[9] W. Theiß and U. M. Titulaer, Z. Phys. B **52**, 75 (1983).
[10] F. Debbasch, K. Mallick, and J. P. Rivet, J. Stat. Phys. **88** (3/4), 945 (1997).
[11] F. Debbasch and J. P. Rivet, J. Stat. Phys. **90** (5/6), 1179 (1998).
[12] M. C. Mackey, *Time's Arrow: The Origins of Thermodynamic Behavior* (Springer Verlag, Berlin, 1992).
[13] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-uniform Gases* (Cambridge U.P., Cambridge, 1970), 3rd ed.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Integral-difference collision operators: Analytical and numerical spectral analysis

Yu. Melnikov
International Solvay Institutes for Physics and Chemistry, Campus Plaine ULB C.P. 231, Bd.du Triomphe, Brussels 1050, Belgium and Laboratory of Complex Systems Theory, Institute for Physics, St. Petersburg State University, Uljanovskaya 1, Petrodvoretz, St. Petersburg 198904, Russia and Theoretische Natuurkunde, Free University of Brussels, C.P. 231, Brussels 1050, Belgium

E. Yarevsky
International Solvay Institutes for Physics and Chemistry, Campus Plaine ULB C.P. 231, Bd. du Triomphe, Brussels 1050, Belgium and Laboratory of Complex Systems Theory, Institute for Physics, St. Petersburg State University, Uljanovskaya 1, Petrodvoretz, St. Petersburg 198904, Russia

(Received 2 November 1998; accepted for publication 17 March 1999)

We analyze spectral properties of a class of integral-difference collision operators arising in some nonequilibrium statistical physics models. We present analytical estimates and numerical results for the operators defined on finite intervals and corresponding to the truncated Gaussian equilibrium distribution function. Some conclusions are drawn about the spectrum of operators on whole axis. Physical limitations for these kinds of models are discussed. [S0022-2488(99)04006-2]

I. INTRODUCTION

In this paper we study analytically and numerically spectral properties of a class of integral-difference operators. These operators arise as collision operators in nonequilibrium gas models[1] in the frame of approach developed in numerous papers (see Refs. 1–5, and references therein). A rigorous spectral analysis of these operators is important for the study of the thermodynamic limit and the problem of integrability in such models. On the other hand, these operators are mathematical objects with interesting properties which need an accurate analytical treatment as well as a numerical analysis based on rigorous approximation theory.

The collision operator acts in the Hilbert space $L_2(\mathbf{R})$ as[1,6,7]

$$(\mathcal{K}_{\varphi}u)(x) \overset{\text{def}}{=} \int_{-\infty}^{\infty} \frac{\varphi(s)u(x)-\varphi(x)u(s)}{|x-s|}\, ds, \tag{1}$$

where $\varphi(x)$ stands for the function of the equilibrium distribution. It has the sense of probability, therefore $\|\varphi\|_{L_1(\mathbf{R})}=1$. In our previous papers[6,7] we proved the basic spectral properties of this operator for a wide class of the equilibrium distributions $\varphi(x)$. However, the results of papers 6 and 7 cannot be applied, for example, to the physical Gaussian equilibrium distribution $\varphi(x) = C_\beta e^{-\beta x^2}$, because they demand the condition that supp $\varphi(x)$ is compact. Actually, together with the normalization $\|\varphi\|_{L_1}=1$ that is a necessary condition to have such $\epsilon>0$ that $\varphi(x)>\epsilon$ $\forall x \in \operatorname{supp} \varphi(x)$, which is very essential for the results proved in Ref. 7. In the present paper we study analytically and numerically the spectral properties of the operator K_φ for the Gaussian distribution, which allow one to establish physical limitations for such models.

The paper is organized as follows. In Sec. II we quote preliminary analytic results. In Sec. III we formulate the main problem of this paper. In Sec. IV we discuss some numerical results toward a solution of this problem. In Sec. V we present analytical results for an asymptotic estimate of the lowest eigenvalues of operator (1). Analytical results for the higher eigenvalues are not yet obtained. Finally, Sec. VI contains a discussion of the results and induced physical limitations for the models generating such collision operators.

II. PRELIMINARY RESULTS

First of all, we need to quote some previous results.[6,7]

Lemma 1: For any real-valued function $\varphi(x)\in L_1(\mathbf{R})\cap L_2(\mathbf{R})$ *the operator* $\mathcal{K}_\varphi$ *acting in the space* $L_2(\mathbf{R})$ *in accordance with formula (1) obeys the relation*

$$\mathcal{K}_\varphi\circ\varphi=\varphi\circ\mathcal{K}_\varphi^*, \tag{2}$$

where $\mathcal{K}_\varphi^*$ *is the adjoint operator and* φ *stands for the operator of multiplication by the function* $\varphi(x)$.

The proof was obtained[6] through the Fourier transform and study of the kernels of the corresponding operators. This result obviously leads to

Corollary 1: For any positive real-valued function $\varphi(x)\in L_1(\mathbf{R})\cap L_2(\mathbf{R})$, *the operator* $\mathcal{K}_\varphi$ *is self-adjoint in the space* $L_2(\mathbf{R};\varphi^{-1}(x)dx)$.

We introduce some notations. Let $[a,b]$ be an interval on the real axis. We denote as $\chi_-(x)$, $\chi_{ab}(x)$, and $\chi_+(x)$ the indicators of the intervals $(-\infty,a)$, $[a,b]$, and (b,∞), respectively. We use the notations P_-, P_{ab}, and P_+ for the projection operators on the subspaces $L_2(-\infty,a)$, $L_2[a,b]$, and $L_2(b,\infty)$, respectively. Actually, for any function $f(x)\in L_2(\mathbf{R})$ we have $(P_{ab}f)(x)=\chi_{ab}(x)f(x)$, and similarly for $P_\pm$. By $\mathcal{K}_0$ we denote operator (1) with (non-normalized) equilibrium distribution $\varphi(x)=\chi_{-1,1}(x)$. We also use the notation $K_\varphi=P_{ab}\mathcal{K}_\varphi P_{ab}$. The operator $K_0=P_{-1,1}\mathcal{K}_0P_{-1,1}$ acts as

$$(K_0u)(x)=\int_{-1}^{1}\frac{u(x)-u(s)}{|x-s|}ds.$$

The following statement is valid.[6]

Theorem 1: *The reduced operator* $K_0=P_{-1,1}\mathcal{K}_0P_{-1,1}$ *in the Hilbert space* $L_2[-1,1]$ *is self-adjoint,* $K_0=K_0^*$. *Its spectrum* $\sigma(K_0)$ *is discrete and equal to the set of simple eigenvalues* $\sigma(K_0)=\{\mu_n\}_{n=0}^{\infty}$, *where*

$$\mu_0=0,\quad \mu_n=2\sum_{j=1}^{n}\frac{1}{j},\quad n\geqslant 1. \tag{3}$$

The corresponding eigenfunctions are Legendre polynomials $P_n(x)$.

The proof[6] is based on straightforward calculations and employs Lemma 1.

Another necessary result is the following.[7]

Theorem 2: *Let the equilibrium distribution function* $\varphi(x)$ *satisfy the following conditions*:

(i) $\varphi(x)$ *has a compact support*: $\operatorname{supp}\varphi(x)\subset[a,b]$;

(ii) $\varphi(x)$ *is bounded, positive and separated from zero on* $[a,b]$: $\exists\epsilon,A:0<\epsilon\leqslant\varphi(x)\leqslant A\,\forall x\in[a,b]$;

(iii) $\varphi(x)\in\mathrm{Lip}(\alpha)$ *for some* $\alpha>0$, *i.e.,* $\exists\alpha,C>0:|\varphi(x)-\varphi(s)|\leqslant C|x-s|^\alpha\,\forall x,s\in[a,b]$.

Then the spectrum $\sigma(\mathcal{K}_\varphi)$ *of the operator* $\mathcal{K}_\varphi$ *given by formula (1) fills the positive semiaxis* $\mathbf{R}_+$. *Additionally, the operator* $\mathcal{K}_\varphi$ *has a discrete real spectrum* $\sigma_d(\mathcal{K}_\varphi)=\{\lambda_n\}$, *semibounded from below,* $\lambda_n\to+\infty$ *when* $n\to\infty$. *This discrete spectrum* $\sigma_d(\mathcal{K}_\varphi)$ *coincides with the spectrum* $\sigma(K_\varphi)$ *of the reduced operator* $K_\varphi=P_{ab}\mathcal{K}_\varphi P_{ab}$ *(which is purely discrete). The corresponding eigenfunctions are the eigenfunctions of the operator* K_φ *multiplied by the indicator* $\chi_{ab}(x)$. *If* $\lambda\in\mathbf{R}_+\backslash\sigma(K_\varphi)$, *then it is a spectral point of double multiplicity. The corresponding generalized eigenfunctions are*

$$u_\lambda^\pm(x) = \delta(x - y_\lambda^\pm) + \chi_{ab}(x)(K_\varphi - \lambda)^{-1}\frac{\varphi(x)}{|x - y_\lambda^\pm|}, \tag{4}$$

where $y_\lambda^\pm$ *are the inverse images of the function* $q_\varphi(x) = \int_a^b \varphi(s)/|x-s|ds$ *in the point* $\lambda: q_\varphi(y_\lambda^\pm) = \lambda,\ y_\lambda^+ > b,\ y_\lambda^- < a$.

Proof of this theorem[7] is based on the decomposition of the space $L_2(\mathbf{R})$ into orthogonal sum $L_2(\mathbf{R}) = L_2(-\infty, a) \oplus L_2[a,b] \oplus L_2(b,\infty)$. It allows for reducing the problem essentially to the study of the reduced operator $K_\varphi = P_{ab}\mathcal{K}_\varphi P_{ab}$. Linear change of variables leads to transformation of the interval $[a,b]$ into the interval $[-1, 1]$ which gives a possibility to consider the operator K_φ on the space $L_2[-1,1]$. Next, we use the formula which can be obtained by a straightforward calculation:

$$K_\varphi = \varphi \circ K_0 - (K_0\varphi), \tag{5}$$

where $(K_0\varphi)$ stands for the operator of multiplication by the function $(K_0\varphi)(x)$. Indeed,

$$\begin{aligned}
(K_\varphi u)(x) &= \int_a^b \frac{u(x)\varphi(s) - u(s)\varphi(x)}{|x-s|} ds \\
&= \int_a^b \frac{u(x)\varphi(x) - u(s)\varphi(x) + u(x)\varphi(s) - u(x)\varphi(x)}{|x-s|} ds \\
&= \varphi(x)\int_a^b \frac{u(x) - u(s)}{|x-s|} ds - u(x)\int_a^b \frac{\varphi(x) - \varphi(s)}{|x-s|} ds \\
&= \varphi(x)(K_0 u)(x) - u(x)(K_0\varphi)(x).
\end{aligned}$$

Using this formula, we compare the resolvents of the operators K_0 and K_φ. Finally, spectral properties of the operator K_0 known from Theorem 1 allow one to finish the proof of Theorem 2.

III. THE PROBLEM

Our main concerns here are, on one hand, conditions (i)–(iii) of Theorem 2, and, on the other hand, the statement of Theorem 2 on the spectrum of the collision operator $\mathcal{K}_\varphi$.

Conclusions on the character of the discrete spectrum of the collision operator $\mathcal{K}_\phi$ are within physical expectations.[1] But there is not, to our knowledge, any reasonable physical interpretation of the continuous spectrum. On the other hand when used by physicists equilibrium distribution functions $\varphi(x)$ do not always satisfy conditions (i)–(iii) of Theorem 2. Namely, condition (iii) (smoothness) and positivity of $\varphi(x)$ are natural physical conjectures. At the same time condition (i) (compactness of the support) is not satisfied even for the most natural Gaussian distribution $\varphi(x) = C_\beta \exp\{-\beta x^2\}$, where C_β is the normalizing constant. Together with the normalization condition $\|\varphi\|_{L_1} = 1$ the fact that $\operatorname{supp}\varphi = \mathbf{R}$ leads to the violation of condition (ii), namely there cannot exist such $\epsilon > 0$ that $\varphi(x) > \epsilon \forall x \in \operatorname{supp}\varphi$. The latter condition is very essential for Theorem 2 (see Ref. 7). Therefore, three interrelated questions appear:

(1) What happens if condition (i) of Theorem 2 is violated? Namely, will the spectral properties of the collision operator $\mathcal{K}_\varphi$ be changed drastically?
(2) How to calculate numerically the spectrum of the operator $\mathcal{K}_\varphi$?
(3) If violation of condition (i) of Theorem 2 (e.g., using the Gaussian distribution function) does not eliminate unnatural spectral properties of the collision operator, what are possible modifications of the mathematical object under investigation and to which physical limitations in the corresponding models they lead?

The rest of our paper is devoted to answering these questions.

IV. NUMERICAL RESULTS FOR THE GAUSSIAN EQUILIBRIUM DISTRIBUTION

Let us consider the Gaussian equilibrium distribution function. Infinite integration limits in Eq. (1) are always understood as the limit of the integral over the interval $[-a,a]$ when $a\to\infty$. In fact, the truncated functions

$$\varphi_a(x)\overset{\text{def}}{=}C_a\chi_{-a,a}(x)e^{-\beta x^2}, \tag{6}$$

are very similar for different but large enough values of the parameter a. Therefore, it may seem that, for large enough a, the truncation parameter a does not influence essentially the spectral properties of the corresponding collision operator. However, that is not true. Namely, there is no regular limit of the collision operator $\mathcal{K}_{\varphi_a}$ on the truncation parameter a when $a\to\infty$. Therefore, there is no way to develop a successful perturbation theory for the spectrum of the operator $\mathcal{K}_{\varphi_a}$ with respect to the parameter $1/a$. We shall show this both numerically and analytically.

Let us make a simple remark. By change of variables one can show that the spectral problem

$$K_{\varphi_a}u=\lambda u \tag{7}$$

on the space $L_2[-a,a]$ is equivalent to the spectral problem

$$K_{\tilde{\varphi}_a}\tilde{u}=\lambda\tilde{u} \tag{8}$$

on the space $L_2[-1,1]$, where $\tilde{\varphi}_a(x)=\tilde{C}_a e^{-\beta a^2x^2}$, i.e., operators K_{φ_a} and $K_{\tilde{\varphi}_a}$ have the same (discrete) spectrum. Indeed, introducing the notations $s'=s/a$, $x'=x/a$, and $\tilde{u}(x)=u(ax)$, we can calculate the action of the operator:

$$\begin{aligned}K_{\varphi_a}u&=\int_{-a}^{a}\frac{u(x)e^{-\beta s^2}-u(s)e^{-\beta x^2}}{|x-s|}ds\\&=a\int_{-1}^{1}\frac{u(x)e^{-\beta a^2s'^2}-u(as')e^{-\beta x^2}}{|x-as'|}ds'\\&=\int_{-1}^{1}\frac{u(ax')e^{-\beta a^2s'^2}-u(as')e^{-\beta a^2x'^2}}{|x'-s'|}ds'\\&=\int_{-1}^{1}\frac{\tilde{u}(x')e^{-\beta a^2s'^2}-\tilde{u}(s')e^{-\beta a^2x'^2}}{|x'-s'|}ds'\\&=K_{\tilde{\varphi}_a}\tilde{u}.\end{aligned}$$

Furthermore, the function $u(x)$ in Eq. (7) transforms with changing of the interval into $u(ax)=\tilde{u}(x)$.

The problem for the equilibrium distribution function $\exp(-\beta x^2)$ with an arbitrary β can be reduced to one with $\beta=1$ by changing the interval. In the same manner as above, we introduce the notations $s'=\sqrt{\beta}s$, $x'=\sqrt{\beta}x$, $\tilde{u}(x)=u(x/\sqrt{\beta})$ and calculate the operator:

$$K_{\varphi_a}u=\int_{-a}^{a}\frac{u(x)e^{-\beta s^2}-u(s)e^{-\beta x^2}}{|x-s|}ds$$
$$=\frac{1}{\sqrt{\beta}}\int_{-a\sqrt{\beta}}^{a\sqrt{\beta}}\frac{u(x)e^{-s'^2}-u(s'/\sqrt{\beta})e^{-\beta x^2}}{|x-s'/\sqrt{\beta}|}ds'$$
$$=\int_{-a\sqrt{\beta}}^{a\sqrt{\beta}}\frac{\tilde{u}(x')e^{-s'^2}-\tilde{u}(s')e^{-x'^2}}{|x'-s'|}ds'.$$

Again, the function $u(x)$ on the interval $[-a,a]$ transforms into $\tilde{u}(x')$ on the interval $[-a\sqrt{\beta},a\sqrt{\beta}]$. Hence, in the following we assume $\beta=1$.

A. Numerical method

In this section we study the spectral problem (7). The equilibrium distribution function $\varphi_a(x)$ is defined in Eq. (6). The normalizing constant C_a is calculated as $\|\varphi_a(x)\|_{L_1([-a,a])}=1$ and is equal to

$$C_a^{-1}=\int_{-a}^{a}e^{-x^2}\,dx=\sqrt{\pi}\,\mathrm{erf}(a).$$

It is worth noticing that the normalizing constant has an asymptotics $C_a\sim 1/2a$ at origin and $C_a\sim 1/\sqrt{\pi}$ at infinity.

It is convenient to construct a numerical method on the interval $[-1, 1]$. As noticed above, in order to keep the same spectrum, we have to change the equilibrium distribution function to $\tilde{\varphi}_a(x)=C_a\exp(-a^2x^2)$. Being led by the spectral analysis of the operator K_0 in Theorem 1, we find eigenfunctions of $K_{\tilde{\varphi}_a}$ as an expansion with respect to the Legendre polynomials:

$$\psi(x)=\sum_{k=0}^{N}v_kP_k(x). \tag{9}$$

Acting by operator (5) on the function ψ and calculating the scalar product in $L_2([-1,1],dx/\tilde{\varphi}_a(x))$, we obtain for spectral problem (7):

$$\int_{-1}^{1}\psi(x)\int_{-1}^{1}\frac{\psi(x)-\psi(s)}{|x-s|}ds\,dx-\int_{-1}^{1}\frac{\psi^2(x)}{\tilde{\varphi}_a(x)}(K_0\tilde{\varphi}_a)(x)dx=\lambda\int_{-1}^{1}\frac{\psi^2(x)}{\tilde{\varphi}_a(x)}dx.$$

Substituting expansion (9) and using Theorem 1 for calculating the first integral, we obtain a generalized spectral problem $\tilde{K}v=\lambda\tilde{S}v$. Matrix elements K_{nm} and S_{nm} of the operators $\tilde{K}$ and $\tilde{S}$ are defined as

$$K_{nm}=\frac{2\mu_n}{2n+1}\delta_{nm}-\int_{-1}^{1}P_n(x)P_m(x)\frac{(K_0\tilde{\varphi}_a)(x)}{\tilde{\varphi}_a(x)}dx, \tag{10}$$

and

$$S_{nm}=\int_{-1}^{1}\frac{P_n(x)P_m(x)}{\tilde{\varphi}_a(x)}dx.$$

To calculate integrals in Eq. (10), we use the relationship:[8]

$$P_n(x)P_m(x)=\sum_{k=0}^{\min(n,m)} A_{nm}^k P_{n+m-2k}(x), \tag{11}$$

where

$$A_{nm}^k=\frac{a_{m-k}a_k a_{n-k}}{a_{n+m-k}}\left(\frac{2n+2m-4k+1}{2n+2m-2k+1}\right), \quad a_k=\frac{(2k-1)!!}{k!}.$$

As the function $\tilde{\varphi}_a(x)$ is symmetric, one can easily see that $(K_0\tilde{\varphi}_a)(x)$ is also symmetric. Thus, symmetric and antisymmetric solutions are separated and one can sum in Eq. (9) over even and odd indices separately. Therefore, all matrix elements in Eq. (10) can be expressed as a linear combination of integrals

$$\int_0^1 P_{2l}(x)\,\frac{f_0(x)}{\tilde{\varphi}_a(x)}\,dx, \tag{12}$$

where $f_0(x)=(K_0\tilde{\varphi}_a)(x)$ for the matrix elements of $\tilde{K}$ and $f_0(x)=1$ for those of $\tilde{S}$.

Special care must be taken concerning numerical calculations of integrals in Eq. (12) since the integrand is highly oscillating for large l and has a huge derivative for large a. These integrals can hardly be calculated with any usual quadrature method. We thus derive a special representation for them.

Let us first discuss the calculation of S_{nm}, $f_0(x)=1$. Using the Taylor expansion for $\exp(a^2x^2)$ and the orthogonality relation

$$\int_{-1}^1 x^m P_k(x)=0, \quad m<k,$$

we find that

$$I_l^{(1)}(a^2)=\frac{1}{C_a}\int_{-1}^1 P_{2l}(x)\exp(a^2x^2)\,dx=\frac{2}{C_a}\sum_{k=l}^{\infty}\frac{a^{2k}}{k!}\int_0^1 x^{2k}P_{2l}(x)\,dx. \tag{13}$$

The latter integral is calculated analytically,[8] and we can finally write

$$I_l^{(1)}(a^2)=\frac{(-1)^l}{C_a}\sum_{k=l}^{\infty}\frac{a^{2k}}{k!}\,\frac{\Gamma(l-k)\Gamma(\frac{1}{2}+k)}{\Gamma(-k)\Gamma(l+\frac{3}{2}+k)}. \tag{14}$$

The residual term of expansion (14) can be easily estimated, and the calculation of a finite sum in (14) is done by using two successive recursion relations with respect to indices l and k.

A similar idea is implemented to the calculation of integrals in K_{nm}:

$$I_l^{(2)}=\int_{-1}^1 P_{2l}\frac{(K_0\tilde{\varphi}_a)(x)}{\tilde{\varphi}_a(x)}\,dx. \tag{15}$$

We expand $\tilde{\varphi}_a(x)$ in terms of Legendre polynomials:

$$\tilde{\varphi}_a(x)=C_a\sum_{k=0}^{\infty}\frac{4k+1}{2}P_{2k}(x)I_k^{(1)}(-a^2). \tag{16}$$

As we know the action of K_0 on Legendre polynomials, we find an expression for $I_l^{(2)}$ in terms of matrix elements of the operator $\tilde{S}$:

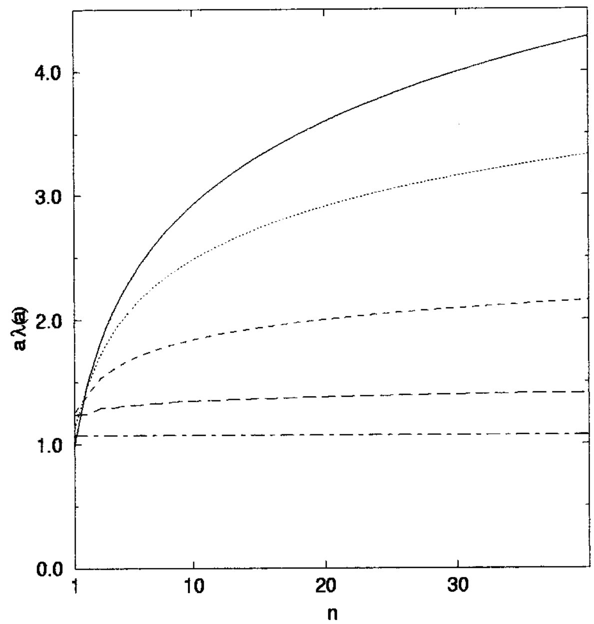

FIG. 1. Eigenvalues $a\lambda_n$ for different values of a. The solid, dotted, dashed, long-dashed, and dot-dashed lines correspond to $a=0$, 1, 1.5, 2, and 3, respectively.

$$I_l^{(2)}=C_a\sum_{k=0}^{\infty}\frac{4k+1}{2}\mu_{2k}I_k^{(1)}(-a^2)S_{2k,2l}. \tag{17}$$

Eigenvalues and eigenfunctions of the generalized matrix problem (10) are computed with standard LAPACK routines in double precision.

B. Eigenvalues and their asymptotics

In Fig. 1 we present a set of a few tens first eigenvalues for different a. (We excluded zeroth eigenvalue since $\lambda_0=0$ for all a.) One can see that an estimation $\lambda_n \geqslant 1/a$ is valid for the eigenvalues. The numerical asymptotics can be found for large n values: $\lambda_n \sim 1/a(1+\kappa(a)\ln n)$. Calculating eigenvalues with rather high accuracy, values of $\kappa(a)$ can be determined. We have found that the function $\kappa(a)$ excellently fits to the exponent. Summarizing the results of this numerical study, we obtain the following asymptotics:

$$\lambda_n \sim \frac{1}{a}(1+\tau e^{-\alpha a^2}\ln n), \quad \tau\approx 0.62, \quad \alpha\approx 0.92, \tag{18}$$

which is valid when both $n,a\to\infty$. Asymptotics (18) is numerically checked up to $n\simeq 100$ and $a\simeq 10$. We would like to say, however, that this is very accurate already for rather small values of n and a. It is also worth pointing out the analytical asymptotics of λ_n for small a:

$$\lambda_n \sim \frac{1}{2a}\mu_n, \quad a\to 0, \tag{19}$$

where μ_n are given by Eq. (3).

Let us now discuss a dependence of eigenfunctions on a. For n being fixed, any eigenfunction is ''pushed off'' the interval and is concentrated very closely to its boundaries with increase in a.

To illustrate this, we first plot three eigenfunctions for different a in Fig. 2. We can see that already for $a=2$ the eigenfunctions are localized nearby the boundary. The eigenfunctions with $a>2$ cannot be plotted in a linear scale.

As the eigenfunctions for small n and big a are nearly zero in a vicinity $x=0$, a difference in pair of symmetric and antisymmetric eigenvalues is getting small, $\lambda_n \approx \lambda_{n+1}$, $n=1,3,5,\ldots$. Evidently, this is not the case for eigenfunctions with large n which are not close to zero anywhere in the interval. To illustrate this ''pairing'' effect, we plot in Fig. 3 the first few eigenvalues as a function of parameter a.

The asymptotics (18) gives us a tool to investigate a limit of the spectrum of $K_{\tilde{\varphi}_a}$ when $a \to \infty$. Let us calculate the difference $\lambda_{n+1}-\lambda_n$ in a vicinity of a positive σ. If n_0, $\sigma=\lambda_{n_0}$, is sufficiently large, we may approximate the finite difference as the derivative, and we find

$$\left.\frac{d\lambda}{dn}\right|_{\lambda=\sigma} = \tau \frac{e^{-\alpha a^2}}{a} \exp\left(-\frac{(a\sigma-1)}{\tau} e^{\alpha a^2} \right). \tag{20}$$

One can see that $(d\lambda/dn)|_{\lambda=\sigma} \to 0$ when $a\to\infty$ for any $\sigma>0$, and $\sigma=0$ also belongs to the spectrum of the operator. Hence, the spectrum condenses into $[0,+\infty)$ when $a\to\infty$. We should note, however, that this does not necessarily mean that the interval $[0,+\infty)$ constitutes the spectrum of the operator $K_{\tilde{\varphi}_a}$ defined on the whole axis. The corresponding limiting procedure cannot be investigated numerically and needs an additional analytical support. However, the numerical results have stimulated analytical calculations presented in Sec. V.

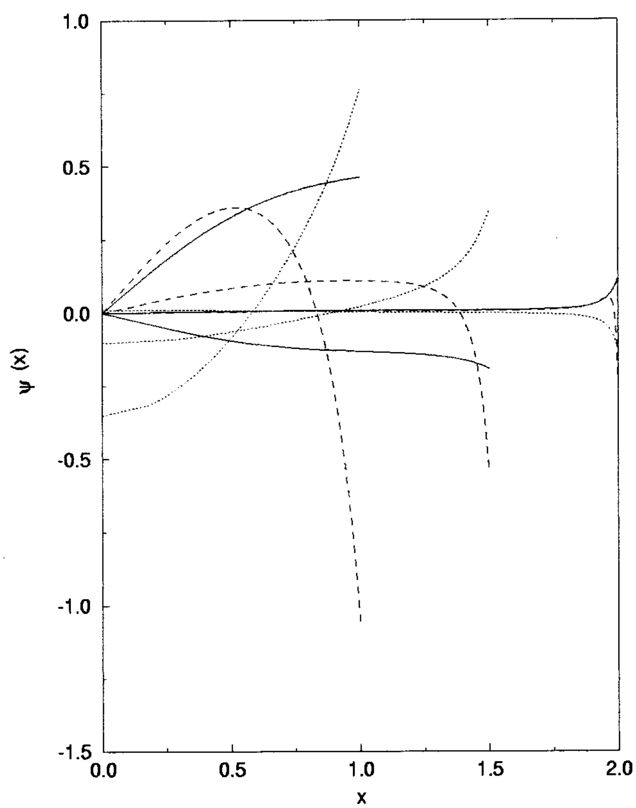

FIG. 2. Normalized eigenfunctions for $a=1$ ([0,1]), $a=1.5$([0,1.5]), and $a=2$ ([0,2]). The solid, dotted, and dashed lines correspond to $n=1$, 2, and 3, respectively.

V. ANALYTIC ESTIMATE FOR THE FIRST EIGENVALUES OF THE COLLISION OPERATOR $\mathcal{K}_{\varphi_a}(x)$ FOR $a \to \infty$

As in the previous section, we consider the collision operator K_{φ_a} restricted to the interval $[-a,a]$ and determined by the truncated Gaussian equilibrium distribution function $\varphi_a(x)$ given

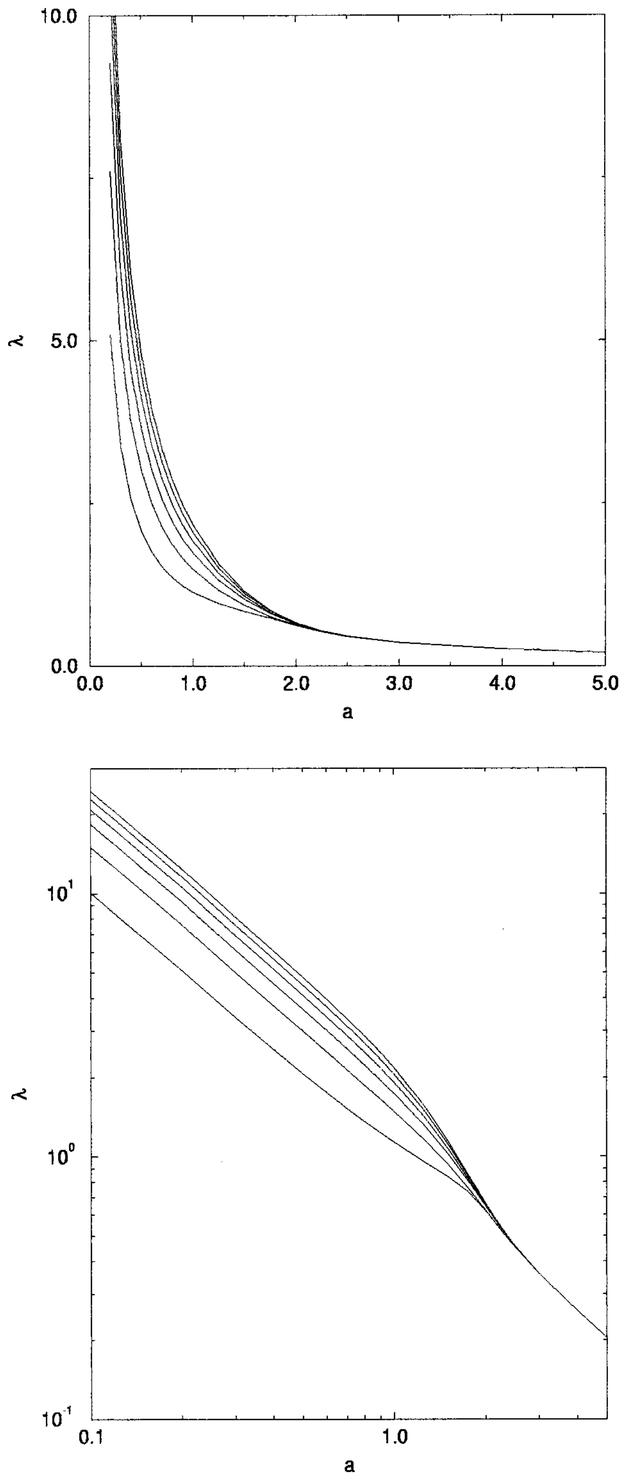

FIG. 3. First six eigenvalues as the function of a (a). The same in the double logarithmic scale (b).

by Eq. (6). We introduce the operator $\mathcal{K}_a$ which is given by formula (1) with the equilibrium distribution function $\varphi(x)=\chi_{-a,a}(x)$ and the reduced operator $K_a=P_{-a,a}\mathcal{K}_a P_{-a,a}$. Theorem 1 can obviously be generalized for the operator K_a on the space $L_2[-a,a]$. Equation (5) reads as

$$K_\varphi=\varphi\circ K_a-(K_a\varphi). \tag{21}$$

Here we prove the following analytic result which partially confirms our numerical calculations.

Theorem 3: *First eigenvalues of the operator K_{φ_a} have the following asymptotic estimates as $a\to\infty$:*

$$\lambda_{1,2}(a)\leqslant \mathrm{Const}\frac{1}{a}(1+o(1)). \tag{22}$$

Proof: The lowest eigenvalue for any a is obviously $\lambda_0=0$ with the eigenfunction $\psi_0^a(x)=\varphi_a(x)$. Operator K_{φ_a} is self-adjoint in the space $L_2([-a,a];\varphi_a^{-1}(x)dx)$ (see Theorem 2), therefore for the next eigenvalue we have:

$$\lambda_1(a)=\min_{g\perp\varphi_a}\frac{\langle K_{\varphi_a}g,g\rangle_{\varphi_a^{-1}}}{\langle g,g\rangle_{\varphi_a^{-1}}}, \tag{23}$$

where $\langle\cdot,\cdot\rangle_{\varphi_a^{-1}}$ stands for the inner product in $L_2([-a,a];\varphi_a^{-1}(x)dx)$,

$$\langle u,v\rangle_{\varphi_a^{-1}}\overset{\text{def}}{=}\int_{-a}^{a}u(x)\overline{v(x)}\varphi_a^{-1}(x)dx,$$

and $g\perp\varphi_a$ means orthogonality with respect to this inner product. The notation $(\cdot,\cdot)$ we reserve for the inner product in $L_2[-a,a]$:

$$(u,v)\overset{\text{def}}{=}\int_{-a}^{a}u(x)\overline{v(x)}dx.$$

Let us choose a test function

$$h_a(x)=x\varphi_a^{1/2}(x).$$

It is antisymmetric and obviously $h_a\perp\varphi_a$ in $L_2([-a,a];\varphi_a^{-1}(x)dx)$. Due to Eq. (23) we have

$$\lambda_1(a)\leqslant\frac{\langle K_{\varphi_a}h_a,h_a\rangle_{\varphi_a^{-1}}}{\langle h_a,h_a\rangle_{\varphi_a^{-1}}}. \tag{24}$$

First, let us calculate the denominator:

$$\langle h_a,h_a\rangle_{\varphi_a^{-1}}=\langle x^2\varphi_a,1\rangle_{\varphi_a^{-1}}=(x^2,1)=\tfrac{2}{3}a^3. \tag{25}$$

Now we shall estimate the nominator using representation (21). We shall also use the fact that operators K_{φ_a} and K_a are self-adjoint on the spaces $L_2([-a,a];\varphi_a^{-1}(x)dx)$ and $L_2[-a,a]$, respectively:

$$\begin{aligned}\langle K_{\varphi_a}h_a,h_a\rangle_{\varphi_a^{-1}}&=\langle K_{\varphi_a}(x\varphi_a^{1/2}),x\varphi_a^{1/2}\rangle_{\varphi_a^{-1}}\\&=(K_{\varphi_a}(x\varphi_a^{1/2}),x\varphi_a^{-1/2})\\&=(\varphi_aK_a(x\varphi_a^{1/2}),x\varphi_a^{-1/2})-(x\varphi_a^{1/2}K_a\varphi_a,x\varphi_a^{-1/2})\\&=(K_a(x\varphi_a^{1/2}),x\varphi_a^{1/2})-(K_a\varphi_a,x^2)\\&=(K_a(x\varphi_a^{1/2}),x\varphi_a^{1/2})-(\varphi_a,K_ax^2).\end{aligned}\tag{26}$$

Therefore, we have to estimate

$$\begin{aligned}(K_a(x\varphi_a^{1/2}),x\varphi_a^{1/2})&=C_a\int_{-a}^{a}\int_{-a}^{a}\frac{x\exp\{-x^2/2\}-s\exp\{-s^2/2\}}{|x-s|}xe^{-x^2/2}\,ds\,dx\\&=a^3C_a\int_{-1}^{1}\int_{-1}^{1}\frac{x\exp\{-a^2x^2/2\}-s\exp\{-a^2s^2/2\}}{|x-s|}xe^{-a^2x^2/2}\,ds\,dx\\&=a^3C_a\int_{-1}^{1}xf_a(x)e^{-a^2x^2/2}\,dx,\end{aligned}\tag{27}$$

where

$$f_a(x)=\int_{-1}^{1}\frac{x\exp\{-a^2x^2/2\}-s\exp\{-a^2s^2/2\}}{|x-s|}\,ds.$$

Using integration by parts, we find that the leading order of expression (27) is

$$(K_a(x\varphi_a^{1/2}),x\varphi_a^{1/2})\sim aC_a\int_{-1}^{1}f_a'(x)e^{-a^2x^2/2}\,dx.$$

Using the Laplace method[9] one can get at $a\to\infty$:

$$\int_{-1}^{1}f_a'(x)e^{-a^2x^2/2}\,dx=\sqrt{\pi}f_a'(0)a^{-1}(1+o(1)).\tag{28}$$

On the other hand,

$$\begin{aligned}f_a'(0)&=2\int_0^1\frac{1-\exp\{-a^2s^2/2\}}{s}\,ds\\&=2\int_0^{a/\sqrt{2}}\frac{1-\exp\{-t^2\}}{t}\,dt\\&=2\int_0^1\frac{1-\exp\{-t^2\}}{t}\,dt+2\int_1^{a/\sqrt{2}}\frac{1-\exp\{-t^2\}}{t}\,dt\\&\leqslant\text{Const}+2\int_1^{a/\sqrt{2}}\frac{dt}{t}\\&=\text{Const}+2\ln a.\end{aligned}\tag{29}$$

Equations (27)–(29) lead to

$$(K_a\varphi_a^{1/2},\varphi_a^{1/2})\leqslant\text{Const}\ln a(1+o(1))\quad\text{as}\quad a\to\infty.\tag{30}$$

The second term in expression (26) is easy to evaluate:

$$(\varphi_a, K_a x^2) = C_a \int_{-a}^{a} dx\, e^{-x^2} \int_{-a}^{a} \frac{x^2 - s^2}{|x-s|}\, ds = a^3 C_a \int_{-1}^{1} dx\, e^{-a^2 x^2} \int_{-1}^{1} \frac{x^2 - s^2}{|x-s|}\, ds.$$

Again using (28), we find at $a \to \infty$:

$$(\varphi_a, K_a x^2) = -\sqrt{\pi} C_a a^2. \tag{31}$$

Combining results (30), (31) and Eqs. (24)–(26), we have

$$\lambda_1(a) \leqslant \frac{3}{2} \frac{1}{a} (1 + o(1)) \quad \text{as} \quad a \to \infty. \tag{32}$$

Now let us notice that subspaces of symmetric and antisymmetric functions are invariant with respect to the operator K_{φ_a}. Therefore, to estimate the second eigenvalue

$$\lambda_2(a) = \min_{g \perp \varphi_a, g \perp \psi_1^a} \frac{\langle K_{\varphi_a} g, g \rangle_{\varphi_a^{-1}}}{\langle g, g \rangle_{\varphi_a^{-1}}}$$

we only should care about the orthogonality to the function $\varphi_a(x)$. Hence, it is natural to choose a test function

$$g(x) = x^2 \varphi_a^{1/2}(x) - \alpha_a \varphi_a(x),$$

where

$$\alpha_a = \langle x^2 \varphi_a^{1/2}, \varphi_a \rangle_{\varphi_a^{-1}} = (x^2 \varphi_a^{1/2}, 1) = C_a^{1/2} \int_{-a}^{a} x^2 e^{-x^2/2}\, dx.$$

Obviously, $g(x)$ is orthogonal both to $\varphi_a(x)$ and to the (unknown explicitly) first antisymmetric eigenfunction. Calculations with the function $g(x)$ are done in the same way as described above, and yield estimate (22). The theorem is proved.

It is interesting to consider as a test function

$$g(x) = x^k \varphi_a^{1/2}(x),$$

properly orthogonalized to φ_a for even k (the symmetric case). Repeating calculations of Theorem 3, we find the asymptotic

$$\lambda_{1,2}(a) \leqslant \frac{2k+1}{2k} \frac{1}{a} (1 + o(1)) \quad \text{as} \quad a \to \infty.$$

This result is well consistent with numerical asymptotics (18). It also confirms the ''pushing off'' effect: The more a test function is concentrated in the vicinity of interval's boundaries, the more it resembles the eigenfunction.

VI. CONCLUSIONS AND DISCUSSION

Results of the present paper directly lead to the following conclusions.

(1) As it was shown in our previous papers,[6,7] in the case when the equilibrium distribution function $\varphi(x)$ satisfies the condition of Theorem 2, collision operator (1) has a branch of continuous spectrum. To our knowledge, this has no physical interpretation.

(2) In the natural case of the Gaussian equilibrium distribution, if the truncation parameter a goes to infinity in improper integral (1), due to the normalization condition $\|\varphi\|_{L_1}=1$, this leads to the violation of condition (ii) of Theorem 2. Namely, there cannot exist such $\epsilon>0$ that $\varphi(x)>\epsilon \forall x \in \operatorname{supp} \varphi$, because $\epsilon \to 0$ when $a \to \infty$. In this case eigenvalues condense in accordance with formula (18) when the truncation parameter a goes to infinity in improper integral (1). In our paper this is proved analytically for the first eigenvalues and numerically for the others. Analytical proof for $n \geqslant 3$ is still an open question. It means that in this case we also do not have physically reasonable spectrum.

(3) The reduced operator $K_\varphi = P_{-a,a}\mathcal{K}_\varphi P_{-a,a}$ with any fixed a has a purely discrete spectrum. The eigenvalues depend essentially on the truncation parameter a.

Summarizing these results one can suggest the following. In order to get physically reasonable spectral properties of collision operator (1), one may consider it on the interval $[-a,a]$ from the very beginning. Namely, the operator K_φ instead of $\mathcal{K}_\varphi$ may be called the collision operator. The truncation parameter a should be chosen taking two things into account. First, it has to provide correct spectrum, which can be done using our analysis. Second, it has to be in agreement with the physical model. The derivation of expression (1) for the collision operator employed some physical assumptions and approximations.[1] Given the interval $[-a,a]$, where the collision operator has a reasonable spectrum, one can clarify what are the physical limitations which restrict our consideration to this interval. However, this question is rather addressed to physicists working in the field of the nonequilibrium statistical physics. Our aim was to demonstrate from the mathematical point of view the presence of such limitations and to provide a rigorous analytical and numerical tool for the investigation of the spectral properties of this class of operators.

ACKNOWLEDGMENTS

We are thankful to Professor I. Prigogine and Professor T. Petrosky who attracted our attention to the problem and to Professor B. Pavlov who emphasized the mathematical interest of this class of problems. The work was supported by the European Commission under the ESPRIT Contract No. 21042 CTIAC.

[1] T. Petrosky and G. Ordonez, Phys. Rev. A **56**, 3507 (1997).
[2] I. Prigogine, *Non-Equilibrium Statistical Mechanics* (Wiley Interscience, New York, 1962).
[3] T. Petrosky and I. Prigogine, Chaos Solitons Fractals **4**, 311 (1994).
[4] T. Petrosky and I. Prigogine, Adv. Chem. Phys. **99**, 1 (1997).
[5] I. Antoniou and S. Tasaki, Int. J. Quantum Chem. **46**, 425 (1993).
[6] Yu. Melnikov, Lett. Math. Phys. **42**, 379 (1997); **43**, 379 (1998).
[7] Yu. Melnikov, *Generalized Functions, Operator Theory and Dynamical Systems*, edited by I. Antoniou and G. Lumer, Research Notes in Mathematics Vol. 399 (Chapman and Hall, London, 1998), p. 354.
[8] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, London, 1980).
[9] M. V. Fedoruk, *Asymptotics: Integrals and Series* (Moscow, Nauka, 1987) (in Russian).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Supersymmetric Kadomtsev–Petviashvili hierarchy: ''Ghost'' symmetry structure, reductions, and Darboux–Bäcklund solutions

H. Aratyn
Department of Physics, University of Illinois at Chicago, 845 West Taylor Street, Chicago, Illinois 60607-7059

E. Nissimov and S. Pacheva
Institute of Nuclear Research and Nuclear Energy, Boul. Tsarigradsko Chausee 72, BG-1784 Sofia, Bulgaria and Department of Physics, Ben-Gurion University of the Negev, Box 653, IL-84105 Beer Sheva, Israel

(Received 19 February 1998; accepted for publication 29 January 1999)

This paper studies Manin–Radul supersymmetric Kadomtsev–Petviashvili hierarchy (MR-SKP) in three related aspects: (i) We find an infinite set of additional (''ghost'') symmetry flows spanning the same (anti)commutation algebra as the ordinary MR-SKP flows. (ii) The latter are used to construct consistent reductions $\mathrm{SKP}_{r/2,m/2}$ of the initial unconstrained MR-SKP hierarchy which involves a nontrivial modification for the fermionic flows. (iii) For the simplest constrained MR-SKP hierarchy $\mathrm{SKP}_{\frac{1}{2},\frac{1}{2}}$ we show that the orbit of Darboux–Bäcklund transformations lies on a supersymmetric Toda lattice being a square root of the standard one-dimensional Toda lattice, and also we find explicit Wronskian-ratio solutions for the super-tau function. © *1999 American Institute of Physics.*
[S0022-2488(99)00906-8]

I. INTRODUCTION

Supersymmetric integrable hierarchies of nonlinear evolution (''super-soliton'') equations were originally proposed[1] from purely mathematical motivations, but soon they attracted active interest also in theoretical physics mainly due to their close connections with superstring theory[2] (for related studies of supersymmetric integrable systems of Korteveg–de Vries or nonlinear-Schrödinger type, see Ref. 3).

The scope of the present paper is the supersymmetric Manin–Radul Kadomtsev–Petviashvili (MR-SKP) hierarchy[1] of integrable supersoliton nonlinear equations within the super-pseudo-differential operator formulation (see also Ref. 4; for other formulations see Ref. 5). We study extensions of the MR-SKP hierarchy incorporating additional (anti)commuting ''ghost'' symmetries, as well as reductions of MR-SKP. We use supersymmetric generalization of several basic concepts in the theory of integrable systems which up to now have been most actively pursued in the context of the ordinary (bosonic) KP hierarchy: Baker–Akhiezer wave functions and tau-functions,[6,7] eigenfunctions, and squared eigenfunction potentials (see Refs. 8 and 9, and references therein).

The advantage of constructing an infinite set of (anti)commuting ''ghost'' symmetries in the supersymmetric context (see Sec. IV below) is twofold. On the one hand, it allows us to double the original supersymmetric hierarchy according to the ''duality'' concept, recently introduced in the context of the ordinary KP hierarchy.[10] On the other hand, using the ''ghost'' symmetries we are able to define systematic reductions of the original MR-SKP model to a broad class of constrained supersymmetric KP hierarchies denoted as $\mathrm{SKP}_{r/2,m/2}$ [see Eq. (5.2) below]. These hierarchies posses correct evolution under both even and odd isospectral flows. The latter turns out to be a nontrivial problem since reductions to $\mathrm{SKP}_{r/2,m/2}$ hierarchies are *incompatible* with the original MR-SKP fermionic flows. We provide a solution to this problem by appropriately modifying

MR-SKP fermionic flows while preserving their original (anti)commutation algebra, i.e., preserving the integrability of the constrained $SKP_{r/2,m/2}$ systems.

The second part of the paper contains a detailed discussion of the simplest constrained MR-SKP hierarchy—$SKP_{1/2,1/2}$ [Eq. (5.3) below], for which we construct Darboux–Bäcklund (DB) transformations preserving both types (even and odd) of the isospectral flows. This again is achieved thanks to the above-mentioned modification of the original MR-SKP fermionic flows. Further, we study the pertinent DB orbit and discover a new supersymmetric Toda (*s*-Toda) lattice structure on it. As a consequence of this result we are able to find explicit Wronskian-ratio representation for corresponding super tau function.

Let us mention that several interesting reduced models of the supersymmetric KP hierarchy have been previously constructed in the literature in terms of super-pseudo-differential operators.[11–14] In particular, the supersymmetric version of AKNS hierarchy was found which allows a description in terms of a bosonic[13] as well as fermionic[14] super-Lax operators. The various properties and superspace formulation of these models were worked out, however, their evolution equations involve only even time flows defining them effectively as reductions of the SKP_2 hierarchy,[11] where only even time flows are present by construction.

II. BACKGROUND ON MANIN–RADUL SUPER-KP HIERARCHY

We shall use throughout the super-pseudo-differential calculus[1] with the following notations: ∂ and $\mathcal{D}=\partial/\partial\theta+\theta\partial$ denote operators, whereas the symbols ∂_x and $\mathcal{D}_\theta$ will indicate application of the corresponding operators on superfield functions. As usual, (x,θ) denote superspace coordinates. For any super-pseudo-differential operator $\mathcal{A}=\Sigma_j a_{j/2}\mathcal{D}^j$ the subscripts $(\pm)$ denote its purely differential part $(\mathcal{A}_+=\Sigma_{j\geqslant 0} a_{j/2}\mathcal{D}^j)$ or its purely pseudodifferential part $(\mathcal{A}_-=\Sigma_{j\geqslant 1} a_{-j/2}\mathcal{D}^{-j})$, respectively. For any $\mathcal{A}$ the super-residuum is defined as $\operatorname{Res}\mathcal{A}=a_{-1/2}$. The rules of conjugation within the super-pseudo-differential formalism are as follows:[13] $(\mathcal{A}\mathcal{B})^*=(-1)^{|A||B|}\mathcal{B}^*\mathcal{A}^*$ for any two elements with gradings $|A|$ and $|B|$; $(\partial^k)^*=(-1)^k\partial^k$, $(\mathcal{D}^k)^*=(-1)^{k(k+1)/2}\mathcal{D}^k$ and $u^*=u$ for any coefficient superfield.

Finally, in order to avoid confusion we shall also employ the following notations: for any super-(pseudo-)differential operator $\mathcal{A}$ and a superfield function f, the symbol $\mathcal{A}(f)$ will indicate application (action) of $\mathcal{A}$ on f, whereas the symbol $\mathcal{A}f$ will denote just operator product of $\mathcal{A}$ with the zero-order (multiplication) operator f.

MR-SKP hierarchy is defined through the *fermionic* Lax operator $\mathcal{L}$:

$$\mathcal{L}=\mathcal{D}+f_0+\sum_{j=1}^{\infty} b_j\partial^{-j}\mathcal{D}+\sum_{j=1}^{\infty} f_j\partial^{-j} \tag{2.1}$$

expressed in terms of a *bosonic* ''dressing'' operator $\mathcal{W}$:

$$\mathcal{L}=\mathcal{W}\mathcal{D}\mathcal{W}^{-1},\quad \mathcal{W}=1+\sum_{j=1}^{\infty}\alpha_j\partial^{-j}\mathcal{D}+\sum_{j=1}^{\infty}\beta_j\partial^{-j}, \tag{2.2}$$

where b_j,β_j are bosonic superfield functions whereas f_j,α_j are fermionic ones and where

$$f_0=2\alpha_1,\quad b_1=-\mathcal{D}_\theta\alpha_1,\quad f_1=2\alpha_2-\alpha_1\mathcal{D}_\theta\alpha_1-2\alpha_1\beta_1-\mathcal{D}_\theta\beta_1. \tag{2.3}$$

Remark: The square of MR-SKP Lax operator (2.1) is an even operator of the form

$$\mathcal{L}^2=\partial+\mathcal{D}_\theta b_1\partial^{-1}\mathcal{D}+(2b_2+b_1^2+\mathcal{D}_\theta f_1+b_1\mathcal{D}_\theta f_0)\partial^{-1}+\cdots. \tag{2.4}$$

Note that the zero-order term in $\mathcal{L}^2$ vanishes $\mathcal{D}_\theta f_0+2b_1=0$ due to (2.3).
The Lax evolution equations for MR-SKP read[1]

$$\frac{\partial}{\partial t_l}\mathcal{L}=-[\mathcal{L}_-^{2l},\mathcal{L}]=[\mathcal{L}_+^{2l},\mathcal{L}], \tag{2.5}$$

$$D_n\mathcal{L}=-\{\mathcal{L}_-^{2n-1},\mathcal{L}\}=\{\mathcal{L}_+^{2n-1},\mathcal{L}\}-2\mathcal{L}^{2n}, \tag{2.6}$$

$$\frac{\partial}{\partial t_l}\mathcal{W}=-(\mathcal{W}\partial^l\mathcal{W}^{-1})_-\mathcal{W}, \quad D_n\mathcal{W}=-(\mathcal{W}\mathcal{D}^{2n-1}\mathcal{W}^{-1})_-\mathcal{W}, \tag{2.7}$$

with the short-hand notations

$$D_n=\frac{\partial}{\partial\theta_n}-\sum_{k=1}^{\infty}\theta_k\frac{\partial}{\partial t_{n+k-1}}, \quad \{D_k,D_l\}=-2\frac{\partial}{\partial t_{k+l-1}}, \tag{2.8}$$

$$(t,\theta)\equiv(t_1\equiv x,t_2,\ldots;\theta,\theta_1,\theta_2,\ldots). \tag{2.9}$$

Accordingly, the super-Zakharov–Shabat (super-ZS) equations take the following form:

$$\frac{\partial}{\partial t_k}\mathcal{L}_+^{2l}-\frac{\partial}{\partial t_l}\mathcal{L}_+^{2k}-[\mathcal{L}_+^{2k},\mathcal{L}_+^{2l}]=0, \quad \frac{\partial}{\partial t_k}\mathcal{L}_+^{2l-1}-D_l\mathcal{L}_+^{2k}-[\mathcal{L}_+^{2k},\mathcal{L}_+^{2l-1}]=0, \tag{2.10}$$

$$D_k\mathcal{L}_+^{2l-1}+D_l\mathcal{L}_+^{2k-1}-\{\mathcal{L}_+^{2k-1},\mathcal{L}_+^{2l-1}\}+2\mathcal{L}_+^{2(k+l-1)}=0. \tag{2.11}$$

Remark: Let us stress that, unlike the possibility to identify $t_1\equiv x$ [since the zero-order term in $\mathcal{L}^2$ (2.4) vanishes], we *cannot* identify $\theta_1\equiv\theta$. Therefore, there is a nontrivial ''evolution'' already with respect to the lowest fermionic flow D_1 (which cannot in general be identified with $\mathcal{D}$).

The super-Baker–Akhiezer (super-BA) and the adjoint super-BA wave functions are defined as

$$\psi_{\mathrm{BA}}(t,\theta;\lambda,\eta)=\mathcal{W}(\psi_{\mathrm{BA}}^{(0)}(t,\theta;\lambda,\eta)), \quad \psi_{\mathrm{BA}}^{*}(t,\theta;\lambda,\eta)=\mathcal{W}^{*-1}(\psi_{\mathrm{BA}}^{*(0)}(t,\theta;\lambda,\eta)) \tag{2.12}$$

(with η being a fermionic ''spectral'' parameter), in terms of the ''free'' super-BA functions

$$\psi_{\mathrm{BA}}^{(0)}(t,\theta;\lambda,\eta)\equiv e^{\xi(t,\theta;\lambda,\eta)}, \quad \psi_{\mathrm{BA}}^{*(0)}(t,\theta;\lambda,\eta)\equiv e^{-\xi(t,\theta;\lambda,\eta)}, \tag{2.13}$$

$$\xi(t,\theta;\lambda,\eta)=\sum_{l=1}^{\infty}\lambda^l t_l+\eta\theta+(\eta-\lambda\theta)\sum_{n=1}^{\infty}\lambda^{n-1}\theta_n \tag{2.14}$$

for which it holds

$$\frac{\partial}{\partial t_k}\psi_{\mathrm{BA}}^{(0)}=\partial_x^k\psi_{\mathrm{BA}}^{(0)}, \quad D_n\psi_{\mathrm{BA}}^{(0)}=\mathcal{D}_\theta^{2n-1}\psi_{\mathrm{BA}}^{(0)}=\partial_x^{n-1}\mathcal{D}_\theta\psi_{\mathrm{BA}}^{(0)}. \tag{2.15}$$

Accordingly, (adjoint) super-BA wave functions satisfy

$$(\mathcal{L}^2)^{(*}\psi_{\mathrm{BA}}^{(*)}=\pm\lambda\,\psi_{\mathrm{BA}}^{(*)}, \quad \frac{\partial}{\partial t_l}\psi_{\mathrm{BA}}^{(*)}=\pm(\mathcal{L}^{2l})_+^{(*)}(\psi_{\mathrm{BA}}^{(*)}), \quad D_n\psi_{\mathrm{BA}}^{(*)}=\pm(\mathcal{L}^{2n-1})_+^{(*)}(\psi_{\mathrm{BA}}^{(*)}). \tag{2.16}$$

Correspondingly, the defining equations for arbitrary (adjoint-) super-eigenfunctions (sEFs) are

$$\frac{\partial}{\partial t_l}\Phi=\mathcal{L}_+^{2l}(\Phi), \quad D_n\Phi=\mathcal{L}_+^{2n-1}(\Phi), \quad \frac{\partial}{\partial t_l}\Psi=-(\mathcal{L}^{2l})_+^{*}(\Psi), \quad D_n\Psi=-(\mathcal{L}^{2n-1})_+^{*}(\Psi) \tag{2.17}$$

with supersymmetric ''spectral'' representations (cf. Ref. 9)

$$\Phi(t,\theta)=\int d\lambda\, d\eta\, \varphi(\lambda,\eta)\psi_{\rm BA}(t,\theta;\lambda,\eta), \quad \Psi(t,\theta)=\int d\lambda\, d\eta\, \varphi^{*}(\lambda,\eta)\psi^{*}_{\rm BA}(t,\theta;\lambda,\eta). \tag{2.18}$$

For later use let us write down the explicit expression for the ''free'' sEF $\Phi^{(0)}$ of the ''free'' $\mathcal{L}^{(0)}=\mathcal{D}$. Namely, taking into account (2.13)–(2.15) and (2.18) we get (for definiteness, consider bosonic $\Phi^{(0)}$)

$$\frac{\partial}{\partial t_k}\Phi^{(0)}=\partial_x^k\Phi^{(0)}, \quad D_n\Phi^{(0)}=\mathcal{D}_\theta^{2n-1}\Phi^{(0)}, \tag{2.19}$$

$$\begin{aligned}\Phi^{(0)}(t,\theta)&=\int d\lambda\, d\eta\, \varphi^{(0)}(\lambda,\eta)e^{\xi(t,\theta;\lambda,\eta)}\\ &=\int d\lambda\left[\left(1-\theta\sum_{n\geqslant 1}\lambda^n\theta_n\right)\varphi_B(\lambda)+\left(\theta+\sum_{n\geqslant 1}\lambda^{n-1}\theta_n\right)\varphi_F(\lambda)\right]\exp\left(\sum_{l\geqslant 1}\lambda^l t_l\right),\end{aligned} \tag{2.20}$$

where $\varphi^{(0)}(\lambda,\eta)=\varphi_F(\lambda)+\eta\varphi_B(\lambda)$ is arbitrary ''spectral'' density.

The super-tau-function $\tau(t,\theta)$ is related with the super-residues of powers of the super-Lax operator (2.1) as follows:

$$\mathrm{Res}\,\mathcal{L}^{2k}=\frac{\partial}{\partial t_k}\mathcal{D}_\theta \ln\tau, \quad \mathrm{Res}\,\mathcal{L}^{2k-1}=D_k\mathcal{D}_\theta\ln\tau. \tag{2.21}$$

Equation (2.21) follows from the identities

$$\begin{gathered}\frac{\partial}{\partial t_l}\mathrm{Res}\,\mathcal{L}^{2k}=\frac{\partial}{\partial t_k}\mathrm{Res}\,\mathcal{L}^{2l}, \quad \frac{\partial}{\partial t_l}\mathrm{Res}\,\mathcal{L}^{2k-1}=D_k\,\mathrm{Res}\,\mathcal{L}^{2l},\\ D_l\,\mathrm{Res}\,\mathcal{L}^{2k-1}+D_k\,\mathrm{Res}\,\mathcal{L}^{2l-1}+2\,\mathrm{Res}\,\mathcal{L}^{2(k+l-1)}=0,\end{gathered} \tag{2.22}$$

which in turn are easily derived from Eqs. (2.5) to (2.6). In particular, for the coefficients of $\mathcal{L}$ and $\mathcal{W}$ we have

$$b_1=\frac{\partial}{\partial t_1}\ln\tau\equiv\partial_x\ln\tau, \quad \alpha_1=D_1\ln\tau. \tag{2.23}$$

In what follows we shall encounter objects of the form $\mathcal{D}_\theta^{-1}(\Phi\Psi)=\mathcal{D}_\theta\partial_x^{-1}(\Phi\Psi)$ where Φ,Ψ is a pair of sEF and adjoint-sEF. Similarly to the purely bosonic case[15] one can show that application of the inverse derivative on such products is well-defined [up to an overall (t,θ)-independent constant]. Namely, there exists a unique superfield function—supersymmetric ''squared eigenfunction potential'' (super-SEP) $S(\Phi,\Psi)$ such that: $\mathcal{D}_\theta S(\Phi,\Psi)=\Phi\Psi$. More precisely the super-SEP satisfies the relations

$$\frac{\partial}{\partial t_k}S(\Phi,\Psi)=\mathrm{Res}(\mathcal{D}^{-1}\Psi\mathcal{L}^{2k}\Phi\mathcal{D}^{-1}), \quad D_kS(\Phi,\Psi)=\mathrm{Res}(\mathcal{D}^{-1}\Psi\mathcal{L}^{2n-1}\Phi\mathcal{D}^{-1}) \tag{2.24}$$

whose consistency follows from the super-ZS Eqs. (2.10) and (2.11). In particular, Eq. (2.24) for $k=1$ and $n=1$ read

$$\partial_x S(\Phi,\Psi)=\mathrm{Res}(\mathcal{D}^{-1}\Psi\mathcal{L}^2\Phi\mathcal{D}^{-1})=\mathcal{D}_\theta(\Phi\Psi), \quad D_1 S(\Phi,\Psi)=\mathrm{Res}(\mathcal{D}^{-1}\Psi\mathcal{L}\Phi\mathcal{D}^{-1})=\Phi\Psi. \tag{2.25}$$

III. ISSUE OF DARBOUX–BÄCKLUND TRANSFORMATIONS IN MR-SKP HIERARCHY

Consider the ''gauge'' transformation of $\mathcal{L}$ (2.1) of the form

$$\tilde{\mathcal{L}}=\mathcal{T}\mathcal{L}\mathcal{T}^{-1}, \quad \mathcal{T}=\chi\mathcal{D}\chi^{-1}, \tag{3.1}$$

which parallels the familiar DB transformation in the purely bosonic case.[15,16] Requiring the transformed Lax operator $\tilde{\mathcal{L}}$ to obey MR-SKP evolution equation of the same form (2.5)–(2.6) as $\mathcal{L}$ implies that $\mathcal{T}$ must satisfy

$$\frac{\partial}{\partial t_l}\mathcal{T}\mathcal{T}^{-1}+(\mathcal{T}\mathcal{L}_+^{2l}\mathcal{T}^{-1})_-=0, \quad D_n\mathcal{T}\mathcal{T}^{-1}-(\mathcal{T}\mathcal{L}_+^{2n-1}\mathcal{T}^{-1})_-=-2(\tilde{\mathcal{L}}^{2n-1})_-. \tag{3.2}$$

The first Eq. (3.2) is exactly analogous to the purely bosonic case and implies that χ must be a sEF (2.17) of $\mathcal{L}$ with respect to the even MR-SKP flows. However, there is a problem with the second Eq. (3.2). Namely, for the general (unconstrained) MR-SKP hierarchy it does not have solutions for χ. In particular, if χ would be a sEF also with respect to fermionic flows [cf. the second Eq. (2.17)], then the left-hand side of second Eq. (3.2) would become zero whereupon we would get the contradictory relation: $(\tilde{\mathcal{L}}^{2n-1})_-=0$.

Thus, we conclude that the DB transformations of the general MR-SKP hierarchy preserve only the bosonic flow equations. In what follows we shall look for consistent solutions of (3.2) in the framework of *constrained* MR-SKP systems which will be achieved thanks to a nontrivial modification of the fermionic MR-SKP flows preserving their anticommutation algebra (2.8).

There is a further essential distinction of DB transformations for MR-SKP hierarchy and its purely bosonic counterpart. Calculating the super-residues of the powers of the DB-transformed Lax operator we obtain

$$\mathrm{Res}\,\tilde{\mathcal{L}}^s=\mathcal{D}_\theta(\chi^{-1}\mathcal{L}_+^s(\chi))+(-1)^{s+1}\,\mathrm{Res}\,\mathcal{L}^s. \tag{3.3}$$

Note the crucial sign factor in front of the second term on the right-hand side of Eq. (3.3). Together with the first Eq. (2.21) it implies for the DB-transformed super-τ function

$$\tilde{\tau}=\chi\tau^{-1} \tag{3.4}$$

in contrast with the bosonic case (where we have $\tilde{\tau}=\chi\tau$).

IV. SUPER-''GHOST'' SYMMETRIES OF MR-SKP HIERARCHY

Consider an infinite set $\{\Phi_{j/2},\Psi_{j/2}\}_{j=0}^{\infty}$ of pairs of (adjoint-)sEFs of $\mathcal{L}$ where those with integer indices are bosonic, whereas those with half-integer indices are fermionic. Next, let us introduce the following infinite set of super-pseudo-differential operators:

$$\mathcal{M}_{s/2}=\sum_{k=0}^{s-1}\Phi_{(s-1-k)/2}\mathcal{D}^{-1}\Psi_{k/2}, \quad s=1,2,..., \tag{4.1}$$

which generate an infinite set of flows $\bar{\partial}_{s/2}(\bar{\partial}_{n-1/2}\equiv\bar{D}_n,\bar{\partial}_k\equiv\partial/\partial\bar{t}_k)$:

$$\bar{\partial}_{s/2}\mathcal{W}=\mathcal{M}_{s/2}\mathcal{W}, \quad \bar{D}_n\mathcal{L}=\{\mathcal{M}_{n-1/2},\mathcal{L}\}, \quad \frac{\partial}{\partial\bar{t}_k}\mathcal{L}=[\mathcal{M}_k,\mathcal{L}]. \tag{4.2}$$

On (adjoint-)sEFs entering $\mathcal{M}_{s/2}$ we allow a *nonhomogeneous* action of the superflows (4.2) which parallels the construction of generalized ''ghost'' symmetry flows in the bosonic case[10] (nonhomogeneous terms are absent in the traditional approach to ''ghost'' symmetry flows[17]):

$$\bar{\partial}_{s/2}\Phi_{l/2}=\mathcal{M}_{s/2}(\Phi_{l/2})-\Phi_{s+l/2}, \quad \bar{\partial}_{s/2}\Psi_{l/2}=-\mathcal{M}^{*}_{s/2}(\Psi_{l/2})+(-1)^{sl}\Psi_{s+l/2}, \tag{4.3}$$

$$\bar{\partial}_{s/2}F^{(*}=\pm\mathcal{M}^{(*)}_{s/2}(F^{(*}), \tag{4.4}$$

where $F^{(*)}$ is a generic (adjoint-)sEF not belonging to the set $\{\Phi_{j/2},\Psi_{j/2}\}$.

Using (4.3) we arrive at the following:

Proposition 1: The infinite set of superflows $\bar{\partial}_{s/2}$ *(4.1) (anti)commute both with the ordinary superflows of MR-SKP (2.5)–(2.6) as well as among themselves:*

$$\left[\frac{\partial}{\partial \bar{t}_s},\frac{\partial}{\partial t_l}\right]=\left[\frac{\partial}{\partial \bar{t}_s},D_n\right]=0, \quad \left[\bar{D}_s,\frac{\partial}{\partial t_l}\right]=\{\bar{D}_s,D_n\}=0, \tag{4.5}$$

$$\left[\frac{\partial}{\partial \bar{t}_s},\frac{\partial}{\partial \bar{t}_k}\right]=\left[\frac{\partial}{\partial \bar{t}_s},\bar{D}_n\right]=0, \quad \{\bar{D}_i,\bar{D}_j\}=-2\frac{\partial}{\partial \bar{t}_{i+j-1}} \tag{4.6}$$

meaning that $\mathcal{M}_{s/2}$ *obey the following equations:*

$$\frac{\partial}{\partial t_k}\mathcal{M}_{s/2}=[\mathcal{L}^{2k}_{+},\mathcal{M}_{s/2}]_{-}, \quad D_n\mathcal{M}_k=[\mathcal{L}^{2n-1}_{+},\mathcal{M}_k]_{-}, \quad D_n\mathcal{M}_{k-1/2}=\{\mathcal{L}^{2n-1}_{+},\mathcal{M}_{k-1/2}\}_{-}, \tag{4.7}$$

$$\frac{\partial}{\partial \bar{t}_k}\mathcal{M}_l-\frac{\partial}{\partial \bar{t}_l}\mathcal{M}_k-[\mathcal{M}_k,\mathcal{M}_l]=0, \quad \frac{\partial}{\partial \bar{t}_k}\mathcal{M}_{l-1/2}-\bar{D}_l\mathcal{M}_k-[\mathcal{M}_k,\mathcal{M}_{l-1/2}]=0, \tag{4.8}$$

$$\bar{D}_k\mathcal{M}_{l-1/2}+\bar{D}_l\mathcal{M}_{k-1/2}-\{\mathcal{M}_{k-1/2},\mathcal{M}_{l-1/2}\}=-2\mathcal{M}_{k+l-1}. \tag{4.9}$$

In checking Eqs. (4.7)–(4.9) we make use of several useful identities for super-pseudo-differential operators:

$$[\mathcal{B}_b,\Phi_{s/2}\mathcal{D}^{-1}\Psi_{k/2}]_{-}=\mathcal{B}_b(\Phi_{s/2})\mathcal{D}^{-1}\Psi_{k/2}-\Phi_{s/2}\mathcal{D}^{-1}\mathcal{B}^{*}_b(\Psi_{k/2}), \tag{4.10}$$

$$[\mathcal{B}_f,\Phi_{s/2}\mathcal{D}^{-1}\Psi_{k/2}]^{(\pm)}_{-}=\mathcal{B}_f(\Phi_{s/2})\mathcal{D}^{-1}\Psi_{k/2}+(-1)^{s}\Phi_{s/2}\mathcal{D}^{-1}\mathcal{B}^{*}_f(\Psi_{k/2}), \tag{4.11}$$

$$(\Phi_{s/2}\mathcal{D}^{-1}\Psi_{k/2})(\Phi_{j/2}\mathcal{D}^{-1}\Psi_{l/2})=\mathcal{X}_{(s,k)}(\Phi_{j/2})\mathcal{D}^{-1}\Psi_{l/2}+(-1)^{k(l+j+1)}\Phi_{s/2}\mathcal{D}^{-1}\mathcal{X}^{*}_{(j,l)}(\Psi_{k/2}), \tag{4.12}$$

$$(\Phi_{j/2}\mathcal{D}^{-1}\Psi_{l/2})^{*}=(-1)^{lj+j+l}\Psi_{l/2}\mathcal{D}^{-1}\Phi_{j/2}, \quad \mathcal{X}_{(s,k)}(\Phi)\equiv\Phi_{s/2}\mathcal{D}^{-1}_{\theta}(\Psi_{k/2}\Phi), \tag{4.13}$$

where $\mathcal{B}_b,\mathcal{B}_f$ indicate arbitrary bosonic/fermionic purely differential super-operators, and $[\cdot,\cdot]^{(\pm)}$ denotes commutator or anticommutator whenever the second element is bosonic/fermionic.

V. CONSTRAINED MR-SKP HIERARCHIES

The super-''ghost''-symmetry flows and the corresponding generating operators $\mathcal{M}_{s/2}$ (4.1) and (4.2) can be used to construct reductions of the full (unconstrained) MR-SKP hierarchy.

Namely, since according to Proposition 1 the super-''ghost'' flows obey the same algebra (4.6) as the original MR-SKP flows, we can identify an infinite subset of the latter with a corresponding infinite subset of the former:

$$\partial_{\ell r/2} = -\bar{\partial}_{\ell m/2}, \quad \ell = 1,2,\ldots, \quad \partial_k \equiv \frac{\partial}{\partial t_k}, \quad \partial_{k-1/2} \equiv D_k, \quad \bar{\partial}_k \equiv \frac{\partial}{\partial \bar{t}_k}, \quad \bar{\partial}_{k-1/2} \equiv \bar{D}_k, \tag{5.1}$$

where (r,m) are some fixed positive integers of equal parity, and retain only these flows as Lax evolution flows (this is a supersymmetric extension of the usual reduction procedure in the purely bosonic case[18]). Equation (5.1) implies the identification $(\mathcal{L}^{r\ell})_- = \mathcal{M}_{\ell m/2}$ for any ℓ and, therefore, the corresponding reduced MR-SKP hierarchy denoted as $\mathrm{SKP}_{r/2,m/2}$ is described by the following constrained super-Lax operator:

$$\mathcal{L}_{(r/2,m/2)} = \mathcal{L}_+^r + \sum_{j=0}^{m-1} \Phi_{m-1-j/2} \mathcal{D}^{-1} \Psi_{j/2}. \tag{5.2}$$

The two simplest constrained MR-SKP Lax operators read

$$\mathcal{L}_{(1/2,1/2)} \equiv \mathcal{L} = \mathcal{D} + f_0 + \Phi_0 \mathcal{D}^{-1} \Psi_0, \tag{5.3}$$

$$\mathcal{L}_{(1,1)} = \partial + \Phi_0 \mathcal{D}^{-1} \Psi_{1/2} + \Phi_{1/2} \mathcal{D}^{-1} \Psi_0, \tag{5.4}$$

where Φ_0, Ψ_0 and $\Phi_{1/2}$, $\Psi_{1/2}$ are pairs of bosonic and fermionic (adjoint-)sEFs with respect to the bosonic flows (about the fermionic flows, see below).

In what follows we shall consider in some detail the simplest constrained $\mathrm{SKP}_{1/2,1/2}$ hierarchy (5.3), and henceforth we shall skip the subscript $(\frac{1}{2}, \frac{1}{2})$ of (5.3) for brevity.

Using identities (4.10)–(4.12) we find the identity for any integer power N (for an analogous formula in the purely bosonic case, see Ref. 19):

$$(\mathcal{L}^N)_- = \sum_{j=0}^{N-1} \mathcal{L}^{N-j-1}(\Phi_0) \mathcal{D}^{-1} \mathcal{L}^{j*}(\Psi_0). \tag{5.5}$$

In particular, for the square of (5.3) we get

$$\mathcal{L}^2 = \partial + \mathcal{L}(\Phi_0) \mathcal{D}^{-1} \Psi_0 + \Phi_0 \mathcal{D}^{-1} \mathcal{L}^*(\Psi_0), \tag{5.6}$$

where again the zero-order term $\mathcal{D}_\theta f_0 + 2\Phi_0 \Psi_0 = 0$ as a particular case of (2.3).

The constrained MR-SKP Lax operator (5.3) satisfies consistently the bosonic flow Eq. (2.5). However, we need to make a nontrivial modification of the original fermionic flows (2.6) in order to keep them compatible with the reduction from the general to the constrained MR-SKP hierarchy. Indeed, taking the $(-)$ part of Eq. (2.6) for the constrained $\mathcal{L}$ (5.3) and using identity (4.11) together with (5.5) we obtain

$$\begin{aligned} &(D_n \Phi_0 - \mathcal{L}_+^{2n-1}(\Phi_0)) \mathcal{D}^{-1} \Psi_0 - \Phi_0 \mathcal{D}^{-1} (D_n \Psi_0 + (\mathcal{L}^{2n-1})_+^*(\Psi_0)) \\ &= -2(\mathcal{L}^{2n})_- = -2 \sum_{j=0}^{2n-1} \mathcal{L}^{2n-1-j}(\Phi_0) \mathcal{D}^{-1} \mathcal{L}^{j*}(\Psi_0), \end{aligned} \tag{5.7}$$

which leads to apparent contradiction.

In Ref. 8 we solved the problem of incompatibility of the standard Orlov–Schulman additional nonisospectral symmetry flows[20] with the reductions of the full bosonic KP hierarchy by appropriately modifying the original Orlov–Schulman flows. Motivated by this work[8] we arrive at the following important proposition:

Proposition 2: There exists the following consistent modification of MR-SKP flows D_n (2.6) *for constrained* $\mathrm{SKP}_{1/2,1/2}$ *hierarchy:*

$$D_n\mathcal{L}=-\{\mathcal{L}_-^{2n-1}-X^{(2n-1)},\mathcal{L}\}=\{\mathcal{L}_+^{2n-1},\mathcal{L}\}+\{X^{(2n-1)},\mathcal{L}\}-2\mathcal{L}^{2n}, \tag{5.8}$$

$$X^{(2n-1)}\equiv 2\sum_{l=0}^{n-2}\mathcal{L}^{2(n-l)-3}(\Phi_0)\mathcal{D}^{-1}(\mathcal{L}^{2l+1})^*(\Psi_0), \tag{5.9}$$

$$D_n\Phi_0=\mathcal{L}_+^{2n-1}(\Phi_0)-2\mathcal{L}^{2n-1}(\Phi_0)+X^{(2n-1)}(\Phi_0), \tag{5.10}$$

$$D_n\Psi_0=-(\mathcal{L}^{2n-1})_+^*(\Psi_0)+2(\mathcal{L}^{2n-1})^*(\Psi_0)-(X^{(2n-1)})^*(\Psi_0). \tag{5.11}$$

The modified D_n *flows obey the same anticommutation algebra (2.8) as in the original unconstrained case.*

In checking the correct anticommutation algebra for D_n (5.8) one has to verify the identities

$$D_kX^{(2l-1)}+D_lX^{(2k-1)}-\{X^{(2k-1)},X^{(2l-1)}\}-\{X^{(2k-1)},\mathcal{L}^{2l-1}\}_--\{X^{(2l-1)},\mathcal{L}^{2k-1}\}_-=0, \tag{5.12}$$

which in turn follow from the definition of $X^{(2n-1)}$ (5.9) together with identities (4.10)–(4.13).

Remark: It is straightforward to generalize Proposition 2 for arbitrary constrained $\mathrm{SKP}_{r/2,m/2}$ hierarchy (5.2). Namely, the modified fermionic flows have the same form as in (5.8) where in the expression for $X^{(2n-1)}$ [cf. (5.9)] one has to sum over all pairs of (adjoint-) sEFs entering the purely pseudodifferential part of $\mathcal{L}_{(r/2,m/2)}$ in (5.2).

Let us now consider DB transformations on $\mathcal{L}\equiv\mathcal{L}_{(1/2,1/2)}$ (5.3) preserving its constrained form:

$$\tilde{\mathcal{L}}=\mathcal{T}\mathcal{L}\mathcal{T}^{-1}=\mathcal{D}+\tilde{f}_0+\tilde{\Phi}_0\mathcal{D}^{-1}\tilde{\Psi}_0,\quad \mathcal{T}=\Phi_0\mathcal{D}\Phi_0^{-1}, \tag{5.13}$$

$$\tilde{f}_0=-f_0-2\mathcal{D}_\theta\ln\Phi_0,\quad \tilde{\Phi}_0=\mathcal{T}\mathcal{L}(\Phi_0)=\Phi_0\partial_x\ln\Phi_0+\Phi_0\mathcal{D}_\theta f_0+\Phi_0^2\Psi_0,\quad \tilde{\Psi}_0=\Phi_0^{-1}. \tag{5.14}$$

We have the following useful identities for DB-transformed quantities:

$$\tilde{\mathcal{L}}^s(\tilde{\Phi}_0)=\mathcal{T}\mathcal{L}^{s+1}(\Phi_0),$$

$$(\tilde{\mathcal{L}}^{s+1})^*(\tilde{\Psi}_0)=(-1)^{s+1}\mathcal{T}^{*-1}\mathcal{L}^{s*}(\Psi_0)=(-1)^s\Phi_0^{-1}\mathcal{D}_\theta^{-1}(\Phi_0\mathcal{L}^{s*}(\Psi_0)). \tag{5.15}$$

There is a further crucial property of the modified D_n flows (5.8)–(5.9):

Proposition 3: The conditions for preserving the fermionic flow Eqs. (5.8)–(5.9) by the Darboux–Bäcklund transformations on $\mathcal{L}\equiv\mathcal{L}_{1/2,1/2}$ *(5.3) [cf. second Eq. (3.2)]:*

$$D_n\mathcal{T}\mathcal{T}^{-1}-(\mathcal{T}\mathcal{L}_+^{2n-1}\mathcal{T}^{-1})_-=-2(\tilde{\mathcal{L}}^{2n-1})_-+\tilde{X}^{(2n-1)}+\mathcal{T}X^{(2n-1)}\mathcal{T}^{-1}, \tag{5.16}$$

where $\mathcal{T}=\Phi_0\mathcal{D}\Phi_0^{-1}$ *and the "tilde" refers to DB-transformed objects, are now satisfied.* The proof of (5.16) proceeds by using the modified D_n flow definitions (5.9)–(5.11) together with identities (4.10)–(4.13) and (5.15).

VI. THE DARBOUX–BÄCKLUND ORBIT OF THE CONSTRAINED MR-SKP HIERARCHY

The recursive expression for the chain of the DB-transformations (5.13)–(5.14) of the constrained $\mathrm{SKP}_{1/2,1/2}$ hierarchy, starting from the "free" initial $\mathcal{L}_0=\mathcal{D}$, reads (the subscript k indicating the step of DB iteration)

$$\mathcal{L}_{k+1}=\mathcal{T}_k\mathcal{L}_k\mathcal{T}_k^{-1}=\mathcal{D}+f_{k+1}+\Phi_{k+1}\mathcal{D}^{-1}\Psi_{k+1},\quad \mathcal{T}_k=\Phi_k\mathcal{D}\Phi_k^{-1}, \tag{6.1}$$

$$\mathcal{L}_1=\mathcal{T}_0\mathcal{D}\mathcal{T}_0^{-1}=\mathcal{D}-2\mathcal{D}_\theta \ln \Phi_0+\Phi_0(\partial_x \ln \Phi_0)\mathcal{D}^{-1}\Phi_0^{-1}, \tag{6.2}$$

where

$$f_{k+1}=-2\mathcal{D}_\theta \ln \Phi_k - f_k, \quad \Psi_{k+1}=\Phi_k^{-1}, \tag{6.3}$$

$$\Phi_{k+1}=\Phi_k\partial_x \ln \Phi_k+\Phi_k\mathcal{D}_\theta f_k+\Phi_k^2\Psi_k \tag{6.4}$$

and where Φ_0 is a sEF of the initial ''free'' $\mathcal{L}_0=\mathcal{D}$ satisfying the ''free'' version of Eq. (5.10) (no $X^{(2n-1)}$ term). Therefore, its explicit expression is given by Eq. (2.20) with substituting $\theta_n \to -\theta_n$. Further we have

$$\Phi_1=\partial_x\Phi_0, \quad \Psi_1=\Phi_0^{-1}, \quad f_1=-2\mathcal{D}_\theta \ln \Phi_0. \tag{6.5}$$

Note, that from (6.3) to (6.4) we find

$$2\Phi_{k+1}\Psi_{k+1}+\mathcal{D}_\theta f_{k+1}=2\Phi_k\Psi_k+\mathcal{D}_\theta f_k=\cdots=0, \tag{6.6}$$

which is consistent with the absence of a zero-order term in the square of $\mathcal{L}_k$ in (6.1).

Equation (6.3) can easily be rewritten as follows:

$$f_{k+1}=-2\mathcal{D}_\theta \sum_{i=0}^{k} (-1)^{k-i} \ln \Phi_i. \tag{6.7}$$

Recalling identity (6.6) we can alternatively rewrite Eq. (6.4) as

$$\Phi_{k+1}=-\tfrac{1}{2}\Phi_k\mathcal{D}_\theta f_{k+1}=\Phi_k\partial_x \ln \Phi_k-\Phi_k^2\Psi_k \tag{6.8}$$

from which we obtain

$$\Phi_{k+1}=\Phi_k\sum_{i=0}^{k} (-1)^{k-i}\partial_x \ln \Phi_i. \tag{6.9}$$

After making the standard substitution $\Phi_k=e^{\varphi_k}$, we find from the second equation in (6.8) a *new* super-Toda (**s**-Toda) lattice equation:

$$\partial_x\varphi_k=e^{\varphi_{k+1}-\varphi_k}+e^{\varphi_k-\varphi_{k-1}}. \tag{6.10}$$

Note, that by acting on (6.10) with ∂_x we get

$$\partial_x^2\varphi_k=e^{\varphi_{k+2}-\varphi_k}-e^{\varphi_k-\varphi_{k-2}}, \tag{6.11}$$

which has the form of the ordinary one-dimensional Toda lattice equation but with a *doubled* lattice spacing and, of course, the Toda variables $\varphi_k=\varphi_k(x,t_2,\ldots;\theta,\theta_1,\ldots)$ are now superfields. Equation (6.10) can also be rewritten as

$$e^{\varphi_{k+1}-\varphi_k}=\sum_{i=0}^{k} (-1)^{k-i}\partial_x\varphi_i \tag{6.12}$$

or

$$\varphi_{k+1}=\varphi_k+\ln\left(\sum_{i=0}^{k} (-1)^{k-i}\partial_x\varphi_i\right). \tag{6.13}$$

We now discuss the Wronskian representation for the sEFs Φ_k. The **s**-Toda lattice (6.10) can apparently be thought of as the square root of the standard Toda lattice. We can use this idea to proceed without any technical calculations. According to the construction given in Ref. 21 the EFs Φ_{2n} associated with even lattice points can be given the usual Wronskian expressions with the starting ''point'' Φ_0. For the same reason, EFs Φ_{2n+1} associated with odd lattice points of the **s**-Toda lattice will have the usual Wronskian expressions with the starting ''point'' $\Phi_1=\partial_x\Phi_0 \equiv \Phi_0^{(1)}$ (6.5).

Generally, for $n=0,1,\ldots$ we find by the above arguments

$$\Phi_{2n}=\frac{W_{n+1}[\Phi_0,\Phi_0^{(1)},\ldots,\Phi_0^{(n)}]}{W_n[\Phi_0,\Phi_0^{(1)},\ldots,\Phi_0^{(n-1)}]},\quad \Phi_{2n+1}=\frac{W_{n+1}[\Phi_0^{(1)},\Phi_0^{(2)},\ldots,\Phi_0^{(n+1)}]}{W_n[\Phi_0^{(1)},\Phi_0^{(2)},\ldots,\Phi_0^{(n)}]}, \tag{6.14}$$

where $W_k[f_1,\ldots,f_k]\equiv\det\|\partial_x^{i-1}f_j\|$, $i,j=1,\ldots,k$, denotes standard Wronskian determinant [however, with superfield entries in (6.14)] and where $\Phi_0^{(k)}\equiv\partial_x^k\Phi_0$ with Φ_0 as in (2.20) (with $\theta_n\to-\theta_n$).

Using (3.4) and the above Wronskians expressions (6.14) we find by iteration the super-tau functions obtained by $2n$ recursive steps of the DB transformations:

$$\tau^{(2n)}=\frac{\Phi_{2n-1}\Phi_{2n-3}\cdots\Phi_1}{\Phi_{2n-2}\Phi_{2n-4}\cdots\Phi_0}=\frac{W_n[\Phi_0^{(1)},\ldots,\Phi_0^{(n)}]}{W_n[\Phi_0,\Phi_0^{(1)},\ldots,\Phi_0^{(n-1)}]}, \tag{6.15}$$

$$\tau^{(2n+1)}=\frac{\Phi_{2n}\Phi_{2n-2}\cdots\Phi_0}{\Phi_{2n-1}\Phi_{2n-3}\cdots\Phi_1}=\frac{W_{n+1}[\Phi_0,\Phi_0^{(1)},\ldots,\Phi_0^{(n)}]}{W_n[\Phi_0^{(1)},\ldots,\Phi_0^{(n)}]}. \tag{6.16}$$

Moreover, since for (5.3) $\partial_x \ln\tau=\Phi_0\Psi_0$, for the k-step DB iteration we have $\partial_x \ln\tau^{(k)}=\Phi_k/\Phi_{k-1}$ by taking into account (3.4). The latter equation together with the relation $\tau^{(k+1)}=\Phi_k/\tau^{(k)}$ true for any DB-step k [cf. (3.4)] yields an alternative super-tau-function form of **s**-Toda lattice:

$$\partial_x \ln\tau^{(k)}(t,\theta)=\frac{\tau^{(k+1)}(t,\theta)}{\tau^{(k-1)}(t,\theta)} \tag{6.17}$$

with the short-hand notation (2.9).

In a subsequent paper we plan to discuss several interesting issues connected with extending the present results: (a) construction of a ''doubled'' MR-SKP hierarchy by providing a super-Lax formulation for the super-''ghost'' symmetry flows [cf. (4.5)–(4.6)]–a supersymmetric extension of the double-KP construction of Ref. 10; (b) general treatment of arbitrary constrained $\mathrm{SKP}_{r/2,m/2}$ hierarchies, including derivation of more general Wronskian-type solutions for the super-tau function and elucidating their Berezinian origin; (c) obtaining consistent formulation of supersymmetric two-dimensional Toda lattice as Darboux–Bäcklund orbit on the ''doubled'' MR-SKP hierarchy (similar to the purely bosonic case[10]) and of supersymmetric analogs of random (multi)matrix models; (d) study of possible connections of super-tau functions, on one hand, and partition functions and joint distribution functions in random matrix models in condensed matter physics (cf. Ref. 22), on the other hand.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support by NSF Grant No. INT-9724747. The work of H.A. was supported in part by the U.S. Department of Energy Grant No. DE-FG02-84ER40173 and the work of E.N. and S.P. was supported in part by Bulgarian NSF Grant No. Ph-401.

[1] Yu. Manin and A. Radul, Commun. Math. Phys. **98**, 65 (1985).
[2] L. Alvarez-Gaumé, H. Itoyama, J. Mañes, and A. Zadra, Int. J. Mod. Phys. A **7**, 5337 (1992); L. Alvarez-Gaumé, K. Becker, M. Becker, R. Emperan, and J. Mañes, *ibid.* **8**, 2297 (1993); S. Stanciu, Commun. Math. Phys. **165**, 261 (1994), hep-th/9407189; J. M. Figueroa-O'Farrill and S. Stanciu, Phys. Lett. B **316**, 282 (1993).
[3] P. Mathieu, J. Math. Phys. **29**, 2499 (1988); S. Belluci, E. Ivanov, S. Krivonos, and A. Pichugin, Phys. Lett. B **312**, 463

(1993), hep-th/9305078; F. Delduc, E. Ivanov, and S. Krivonos, J. Math. Phys. **37**, 1356 (1996), hep-th/9510033; F. Toppan, Int. J. Mod. Phys. A **11**, 3257 (1996), hep-th/9506133; Q. P. Liu and M. Mañas, solv-int/9711002.
[4] L. Martinez Alonso and E. Medina Reus, J. Math. Phys. **36**, 4898 (1995); A. Ibort, L. Martinez Alonso, and E. Medina Reus, *ibid.* **37**, 6157 (1996).
[5] V. Kac and J. van de Leur, Ann. Inst. Fourier **37**, 99 (1987); V. Kac and E. Medina, Lett. Math. Phys. **37**, 435 (1996); A. LeClair, Nucl. Phys. B **314**, 425 (1989); M. Mulase, J. Diff. Geom. **34**, 651 (1991); J. Rabin, Commun. Math. Phys. **137**, 533 (1991).
[6] L. Dickey, *Soliton Equations and Hamiltonian Systems* (World Scientific, Singapore 1991); Acta Applicandae Math. **47**, 243 (1997).
[7] P. van Moerbeke, in *Lectures on Integrable Systems*, edited by O. Babelon *et al.* (World Scientific, Singapore, 1994).
[8] H. Aratyn, E. Nissimov, and S. Pacheva, Int. J. Mod. Phys. A **12**, 1265 (1997), hep-th/9607234; Phys. Lett. A **228**, 164 (1997), hep-th/9602068.
[9] H. Aratyn, E. Nissimov, and S. Pacheva, Commun. Math. Phys. **193**, 493 (1998), solv-int/9701017.
[10] H. Aratyn, E. Nissimov, and S. Pacheva, Phys. Lett. A **244**, 245 (1998), solv-int/9712012.
[11] J. M. Figueroa-O'Farrill, J. Mas, and E. Ramos, Rev. Mod. Phys. **3**, 479 (1991); W. Oevel and Z. Popowicz, Commun. Math. Phys. **139**, 441 (1991); F. Yu, Nucl. Phys. B **375**, 173 (1992).
[12] Z. Popowicz, J. Phys. A **29**, 1281 (1996), hep-th/9510185; J. C. Brunelli and A. Das, Phys. Lett. B **337**, 303 (1994), hep-th/9406214; L. Bonora, S. Krivonos, and A. Sorin, Nucl. Phys. B **477**, 835 (1996), hep-th/9604165; F. Delduc and L. Gallot, Commun. Math. Phys. **190**, 395 (1997), solv-int/9609008; J.-C. Shaw and M.-H. Tu, solv-int/9712009.
[13] H. Aratyn and C. Rasinariu, Phys. Lett. B **391**, 99 (1997), hep-th/9608107; H. Aratyn, A. Das, and C. Rasinariu, Mod. Phys. Lett. A **12**, 2623 (1997), solv-int/9704119; H. Aratyn, A. Das, C. Rasinariu, and A. H. Zimerman, in *Supersymmetry and Integrable Models*, Lecture Notes in Physics Vol. 502, edited by H. Aratyn *et al.* (Springer-Verlag, Berlin, 1998).
[14] H. Aratyn and A. Das, Mod. Phys. Lett. A **13**, 1185 (1998), solv-int/9710026.
[15] W. Oevel, Physica A **195**, 533 (1993); W. Oevel and W. Schief, Rev. Math. Phys. **6**, 1301 (1994).
[16] L.-L. Chau, J. C. Shaw, and H. C. Yen, Commun. Math. Phys. **149**, 263 (1992).
[17] A. Orlov, in *Nonlinear Processes in Physics: Proceedings of the III Potsdam-V Kiev Workshop*, edited by A. S. Fokas *et al.* (Springer, Berlin, 1993); Y. Cheng, W. Strampp, and B. Zhang, Commun. Math. Phys. **168**, 117 (1995).
[18] W. Oevel, Physica A **195**, 533 (1993); Y. Cheng, W. Strampp, and B. Zhang, Commun. Math. Phys. **168**, 117 (1995).
[19] B. Enriquez, A. Yu. Orlov, and V. N. Rubtsov, Inverse Probl. **12**, 241 (1996), solv-int/9510002.
[20] A. Orlov and E. Schulman, Theor. Math. Phys. **64**, 323 (1985); Lett. Math. Phys. **12**, 171 (1986); A. Orlov, in *Plasma Theory and Nonlinear and Turbulent Processes in Physics* (World Scientific, Singapore, 1988).
[21] H. Aratyn, E. Nissimov, and S. Pacheva, Phys. Lett. A **201**, 293 (1995), hep-th/9501018.
[22] Y. Avishai, Y. Hatsugai, and M. Kamoto, Phys. Rev. B **53**, 8369 (1996).

Hamiltonian formalism for the nonlinear Schrödinger equation in physical space–time

Zhi-De Chen
Department of Physics, Wuhan University, Wuhan 430072, China and Department of Physics, Guangzhou Normal University, Guangzhou 510400, China

Xiang-Jun Chen
Department of Physics, Jinan University, Guangzhou 510632, China

Nian-Ning Huang
Department of Physics, Wuhan University, Wuhan 430072, China

(Received 10 March 1998; accepted for publication 10 March 1999)

Hamiltonian formalism for the nonlinear Schrödinger equation in physical space–time is developed. Owing to the fact that the equation involves the second partial derivative with respect to time, the canonical variables are shown to be u, $\overline{u}$, $-\overline{u_t}$, and $-u_t$. The first Lax equation and its variations with respect to canonical variables are very complicated, but the Poisson brackets of transition coefficients are shown to be simple, and thus the Hamiltonian formalism in terms of action-angle variables has been achieved. A peculiarity is that the continuous spectrum consists of real λ as well as pure imaginary λ. In the case of pure imaginary λ the full Jost solutions shall tend to be infinite or vanish as $|t|\to\infty$. This problem needs further investigation. [S0022-2488(99)04106-7]

I. INTRODUCTION

In the last two decades the nonlinear Schrödinger (NLS) equation has been thoroughly investigated from the inverse scattering transform,[1] Hamiltonian formalism[2–4] to the quantum inverse scattering transform.[5,6] As is well known, to describe, for example, a short temporal pulse propagation in nonlinear optical fibers the NLS equation takes a form[7–9]

$$iu_x+u_{tt}+2|u|^2u=0, \tag{1}$$

where t and x denote the time and the space coordinate in the frame of reference moving with the group velocity. This equation will be referred to as the NLS equation in physical space–time. Nevertheless, since a short temporal pulse is a soliton in time, usually the equation used to describe the propagation of the so-called optical soliton in a fiber is the conventional NLS equation[8,10]

$$iu_T+u_{XX}+2|u|^2u=0, \tag{2}$$

where T and X are not true physical time and space though we read them as time and space. Most of the investigations start from (2). In finding classical solutions, for example, soliton solutions, t, x, and T, X are just pure parameters, the solutions to (1) can be obtained from the solutions to (2) by simply exchanging variables and *vice versa*.[11,12] However, for developing Hamiltonian formalism as to the quantum inverse scattering transform, one should start from (1) where x and t are true space and time. On the other hand, Eq. (1) is also called the unstable nonlinear Schrödinger equation which has been studied in detail by Wadati *et al.*[11,13–15] It was shown that (2) can be used to describe the soliton phenomena in unstable media. (1) has been derived for two physics systems: the Rayleigh–Taylor instability and electron beam plasma. The soliton solutions of (1) and its application to the physics problem have been discussed. As is known, a Hamiltonian formalism of Eq. (1) has not yet been developed, and the quantum inverse scattering transform for (1) is still an open question.

In this paper the Hamiltonian formalism for the NLS equation in physical space–time is developed in an exact manner which provides a basis for further development of the quantum inverse scattering transform and its potential application in the physical systems mentioned above. For convenience, the boundary condition will be chosen as:

$$u \to 0 \quad \text{as} \quad |x| \to \infty. \tag{3}$$

II. GENERAL FORMALISM

The Lagrangian density $\mathcal{L}$ for (1) is given by

$$\mathcal{L} = -\frac{i}{2}\{\bar{u}_x u - \bar{u} u_x\} - \bar{u}_t u_t + \bar{u}^2 u^2. \tag{4}$$

Since Eq. (1) involves the second derivative with respect to t, the generalized coordinates are chosen as u and $\bar{u}$. This can be easily shown to be right since the Lagrangian equation

$$\frac{\partial \mathcal{L}}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial \mathcal{L}}{\partial u_x}\right) - \frac{\partial}{\partial t}\left(\frac{\partial \mathcal{L}}{\partial u_t}\right) = 0, \tag{5}$$

and that for $\bar{u}$ reproduce Eq. (1) and its complex conjugate. The momentum densities conjugated to u and $\bar{u}$, π and $\bar{\pi}$, are given by

$$\pi \equiv \frac{\partial \mathcal{L}}{\partial u_t} = -\bar{u}_t, \quad \bar{\pi} \equiv \frac{\partial \mathcal{L}}{\partial \bar{u}_t} = -u_t. \tag{6}$$

Then the Hamiltonian density $\mathcal{H}$ is given by

$$\mathcal{H} \equiv \pi u_t + \overline{\pi u_t} - \mathcal{L} = \frac{i}{2}\{\bar{u}_x u - \bar{u} u_x\} - \bar{u}_t u_t - \bar{u}^2 u^2. \tag{7}$$

Since u and π, as well as $\bar{u}$ and $\bar{\pi}$, are conjugated with each other, the basic Poisson brackets are

$$\begin{aligned} &\{u(x), -\bar{u}_t(y)\} = \delta(x-y), \quad \{u(x), \bar{u}(y)\} = 0, \\ &\{\bar{u}(x), -u_t(y)\} = \delta(x-y), \quad \{u_t(x), \bar{u}_t(y)\} = 0. \end{aligned} \tag{8}$$

Suppose that S and T are functionals of u, $\bar{u}$, $-u_t$, and $-\bar{u}_t$, then the Poisson bracket of S and T is given by

$$\{S,T\} = -\int_{-\infty}^{\infty}\left(\frac{\delta S}{\delta u(x)}\frac{\delta T}{\delta \bar{u}_t(x)} + \frac{\delta S}{\delta \bar{u}(x)}\frac{\delta T}{\delta u_t(x)} - \frac{\delta S}{\delta \bar{u}_t(x)}\frac{\delta T}{\delta u(x)} - \frac{\delta S}{\delta u_t(x)}\frac{\delta T}{\delta \bar{u}(x)}\right)dx. \tag{9}$$

III. THE JOST SOLUTIONS

The first Lax equation of (1) is given by

$$\partial_x \Phi(x,\lambda) = L(x,\lambda)\Phi(x,\lambda), \tag{10}$$

where

$$L(x,\lambda) = -i2\lambda^2\sigma_3 + 2\lambda U(x,t) - i\{U^2(x,t) + U_t(x,t)\}\sigma_3, \tag{11}$$

and

$$U(x,t)=\begin{pmatrix} 0 & u(x,t) \\ -\overline{u(x,t)} & 0 \end{pmatrix}. \tag{12}$$

We define the Jost solutions by boundary condition (3),

$$\Phi(x,\lambda)\to e^{-i2\lambda^2 x\sigma_3} \quad \text{as} \quad x\to-\infty, \tag{13}$$

and then find

$$\Phi(x,\lambda)=e^{-i2\lambda^2 x\sigma_3}+\int_{-\infty}^{x} dy\, e^{-i2\lambda^2(x-y)\sigma_3}\{L(y,\lambda)+i2\lambda^2\sigma_3\}\Phi(y,\lambda). \tag{14}$$

Similarly, we define

$$\Psi(x,\lambda)\to e^{-i2\lambda^2 x\sigma_3} \quad \text{as} \quad x\to\infty. \tag{15}$$

The monodromy matrix is given by

$$T(\lambda)=\Psi^{-1}(x,\lambda)\Phi(x,\lambda) \quad \text{or} \quad \Phi(x,\lambda)=\Psi(x,\lambda)T(\lambda). \tag{16}$$

We write

$$\Phi(x,\lambda)=(\phi(x,\lambda)\tilde{\phi}(x,\lambda)), \quad \Psi(x,\lambda)=(\tilde{\psi}(x,\lambda)\psi(x,\lambda)), \quad T(\lambda)=\begin{pmatrix} a(\lambda) & -\tilde{b}(\lambda) \\ b(\lambda) & \tilde{a}(\lambda) \end{pmatrix}, \tag{17}$$

where $\phi(x,\lambda)$, etc., have two components, respectively. Since $\det\Psi(x,\lambda)=1$, from (16), we have

$$a(\lambda)=\psi_2(x,\lambda)\phi_1(x,\lambda)-\psi_1(x,\lambda)\phi_2(x,\lambda), \quad b(\lambda)=-\tilde{\psi}(x,\lambda)\phi_1(x,\lambda)+\tilde{\psi}_1(x,\lambda)\phi_2(x,\lambda), \tag{18}$$

etc., and

$$\phi(x,\lambda)=a(\lambda)\tilde{\psi}(x,\lambda)+b(\lambda)\psi(x,\lambda), \tag{19}$$

etc. These functions are defined in real λ^2, i.e., on the real axis and on the imaginary axis of complex λ plane.

From (16) we show that $\phi(x,\lambda)$ can be analytically continued into the region of $\operatorname{Im}\lambda^2>0$, i.e., in the first and the third quadrants and that $\tilde{\phi}(x,\lambda)$ can be analytically continued into the region of $\operatorname{Im}\lambda^2<0$, i.e., in the second and the fourth quadrants. Zeros of $a(\lambda)$ are located in the first and the third quadrants and are assumed to be simple. From (10) it can be shown

$$\tilde{\phi}(x,\bar{\lambda})=i\sigma_2\overline{\phi(x,\lambda)}, \quad \tilde{\psi}(x,\bar{\lambda})=-i\sigma_2\overline{\psi(x,\lambda)}, \tag{20}$$

$$\tilde{a}(\bar{\lambda})=\overline{a(\lambda)}, \quad \tilde{b}(\bar{\lambda})=\overline{b(\lambda)}, \tag{21}$$

noting that the second one of the last equation holds only for real λ^2. In the first and the third quadrants $a(\lambda)$ may have zeros. $a(\lambda)$ is assumed to have N simple zeros, λ_n, $n=1,2,...,N$. From (18) we have

$$\phi(x,\lambda_n)=b_n\psi(x,\lambda_n), \tag{22}$$

where b_n is independent of x.

IV. AN INVERSE SCATTERING TRANSFORM

Introducing

$$\Theta(x,\lambda)=\begin{cases} a(\lambda)^{-1}\phi(x,\lambda) & \text{as Im}\,\lambda^2>0 \\ \tilde{\psi}(x,\lambda) & \text{as Im}\,\lambda^2<0. \end{cases} \tag{23}$$

By using the standard procedure, we obtain the equation of inverse scattering transform in the form of Zakharov–Shabat:

$$\tilde{\psi}(x,\lambda)e^{i2\lambda^2 x}=\begin{pmatrix}1\\0\end{pmatrix}+R(x,\lambda)+J(x,\lambda), \tag{24}$$

where

$$R(x,\lambda)=\sum_n \frac{1}{\lambda-\lambda_n}c_n\psi(x,\lambda_n)e^{i2\lambda_n^2 x}, \tag{25}$$

$$J(x,\lambda)=\frac{1}{i2\pi}\int_\Gamma \frac{1}{\lambda'-\lambda}r(\lambda')\psi(x,\lambda')e^{i2\lambda'^2 x}\,d\lambda', \tag{26}$$

$$c_n=\frac{b_n}{a_n}, \quad a_n\equiv\dot{a}(\lambda_n)=\frac{d}{d\lambda}a(\lambda)\bigg|_{\lambda=\lambda_n}, \quad r(\lambda)=\frac{b(\lambda)}{a(\lambda)}, \tag{27}$$

and where Γ denotes the integral path which contains two parts.[12] One part is along the real axis from 0 to $-\infty$ and from 0 to $+\infty$, the other is along the imaginary axis from 0 to $-i\infty$ and from 0 to $+i\infty$, i.e.,

$$\int_\Gamma \mathcal{F}(x,\lambda)d\lambda=\left(\int_0^{+\infty}-\int_0^{-\infty}\right)\mathcal{F}(x,\mu)d\mu-\left(\int_0^{+\infty}-\int_0^{-\infty}\right)\mathcal{F}(x,i\nu)d(i\nu). \tag{28}$$

The time dependence is determined by the second Lax equation,

$$\partial_t\Phi(x,t,\lambda)=M(x,t,\lambda)\Phi(x,t,\lambda), \tag{29}$$

where

$$M(x,t,\lambda)=-i\lambda\sigma_3+U(x,t). \tag{30}$$

By using the second Lax equation, the time dependence is achieved by simple replacements:

$$a(\lambda)\to a(t,\lambda)=a(0,\lambda), \quad b(\lambda)\to b(t,\lambda)=b(0,\lambda)e^{i2\lambda t}, \quad b_n\to b_n(t)=b_n(0)e^{i2\lambda_n t}. \tag{31}$$

The solution is then determined by

$$\overline{u(x,t)}=\lim_{|\lambda|\to\infty}(i2\lambda)\tilde{\psi}_2(x,t,\lambda)e^{i2\lambda^2 x+i\lambda t}. \tag{32}$$

In the case of no reflections, the Zakharov–Shabat equation (24) including the time dependence (31) has the same form as that of Eq. (2) by simply exchanging[12]

$$x,t\leftrightarrow T,X, \tag{33}$$

while the locations of zeros of $a(\lambda)$ are somewhat different. In the latter case, all the zeros are in the upper half plane, while in the former case, those are in the first or third quadrants. However, as we have shown,[16] one can change a pole from the lower half plane to upper half plane in the case that these two poles are symmetric with respect to the real axis, and the final results are the same. This implies that, for finding soliton solutions, Eqs. (1) and (2) are equivalent.

V. CONSERVATIVE QUANTITIES

Since $a(\lambda)$ is assumed to have N simple poles in the first and third quadrants, $a(\lambda)$ can be expressed as

$$a(\lambda)=\prod_{n=1}^{N}\frac{\lambda-\lambda_n}{\lambda-\overline{\lambda_n}}\check{a}(\lambda), \tag{34}$$

where $\check{a}(\lambda)$ is analytical in the first and the third quadrants and has no zeros. Taking into account that

$$\ln a(\lambda)\to 0 \quad \text{as } |\lambda|\to\infty,$$

the general expression of $a(\lambda)$ is

$$a(\lambda)=\prod_{n=1}^{N}\frac{\lambda-\lambda_n}{\lambda-\overline{\lambda_n}}\exp\left\{\frac{1}{i2\pi}\int_\Gamma\frac{\ln|a(\lambda')|}{\lambda'-\lambda}d\lambda'\right\}. \tag{35}$$

The fact that $a(t,\lambda)$ is independent of t leads to infinite conservation laws. An expansion of $\ln a(\lambda)$ as $|\lambda|\to\infty$ is similar to that for (2), e.g., the second conservative quantity I_2 is[10]

$$I_2=\sum_{n=1}^{n}(\lambda_n^2-\overline{\lambda_n^2})-i\frac{2}{\pi}\int_\Gamma \ln|a(\lambda)|2\lambda\,d\lambda. \tag{36}$$

The expressions related to u for Eq. (1) are different from those related to u for (2), e.g., I_2 in the present case is (for details, see Appendix A)

$$I_2=\int_{-\infty}^{\infty}dx\{i|u_t|^2-i|u|^4+(|u|^2)_x-\overline{u}u\}. \tag{37}$$

VI. VARIATIONS WITH RESPECT TO $u(x)$ AND $\overline{u(x)}$

From (10) we have

$$\partial_x\delta\Phi(x,\lambda)=L(x,\lambda)\,\delta\Phi(x,\lambda)+\delta L(x,\lambda)\Phi(x,\lambda). \tag{38}$$

The solution is obviously

$$\delta\Phi(x,\lambda)=\int_{-\infty}^{x}dz\,\Phi(x,\lambda)\Phi^{-1}(z,\lambda)\,\delta L(z,\lambda)\Phi(z,\lambda). \tag{39}$$

We have

$$\frac{\delta L(z,\lambda)}{\delta u(x)}=\{2\lambda\sigma_+ + i\overline{u(x)}\sigma_3\}\delta(x-z), \tag{40}$$

etc., in the case of $x>z$,

$$\frac{\delta\Phi(x,\lambda)}{\delta u(z)}=\Phi(x,\lambda)\Phi^{-1}(z,\lambda)\{2\lambda\sigma_+ + i\overline{u(z)}\sigma_3\}\Phi(z,\lambda), \tag{41}$$

and then setting $z=x^-\equiv x-0$,

$$\frac{\delta\Phi(x,\lambda)}{\delta u(x^-)}=\{2\lambda\sigma_+ + i\overline{u(x)}\sigma_3\}\Phi(x,\lambda). \tag{42}$$

Similarly,

$$\frac{\delta\Phi(x,\lambda)}{\delta\overline{u(x^-)}}=\{-2\lambda\sigma_- + iu(x)\sigma_3\}\Phi(x,\lambda), \tag{43}$$

$$\frac{\delta\Phi(x,\lambda)}{\delta\overline{u_t(x^-)}}=i\sigma_-\Phi(x,\lambda), \tag{44}$$

and

$$\frac{\delta\Phi(x,\lambda)}{\delta u_t(x^-)}=-i\sigma_+\Phi(x,\lambda), \tag{45}$$

where

$$\sigma_+=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad \sigma_-=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}. \tag{46}$$

In the same way, we obtain

$$\frac{\delta\Psi^{-1}(x,\lambda)}{\delta u(x^+)}=\Psi^{-1}(x,\lambda)\{2\lambda\sigma_+ + i\overline{u(x)}\sigma_3\}, \tag{47}$$

$$\frac{\delta\Psi^{-1}(x,\lambda)}{\delta\overline{u(x^+)}}=\Psi^{-1}(x,\lambda)\{-2\lambda\sigma_- + iu(x)\sigma_3\}, \tag{48}$$

$$\frac{\delta\Psi^{-1}(x,\lambda)}{\delta\overline{u_t(x^+)}}=\Psi^{-1}(x,\lambda)i\sigma_-, \tag{49}$$

and

$$\frac{\delta\Psi^{-1}(x,\lambda)}{\delta u_t(x^+)}=-\Psi^{-1}(x,\lambda)i\sigma_+, \tag{50}$$

where $x^+\equiv x+0$.

From (16) we have

$$\delta T(\lambda)=\delta\Psi^{-1}(z,\lambda)\Phi(z,\lambda)+\Psi^{-1}(z,\lambda)\delta\Phi(z,\lambda), \tag{51}$$

and then

$$\frac{\delta T(\lambda)}{\delta u(x)}=\frac{\delta\Psi^{-1}(z,\lambda)}{\delta u(x)}\Phi(z,\lambda)+\Psi^{-1}(z,\lambda)\frac{\delta\Phi(z,\lambda)}{\delta u(x)}. \tag{52}$$

Since z is arbitrary in the above formula, we can set $z\rightarrow x^-$,

$$\frac{\delta T(\lambda)}{\delta u(x)}=\Psi^{-1}(x,\lambda)\{2\lambda\sigma_+ + i\overline{u(x)}\sigma_3\}\Phi(x,\lambda). \tag{53}$$

Similarly,

$$\frac{\delta T(\lambda)}{\delta u(x)} = \Psi^{-1}(x,\lambda)\{-2\lambda\sigma_- + iu(x)\sigma_3\}\Phi(x,\lambda), \tag{54}$$

$$\frac{\delta T(\lambda)}{\delta u_t(x)} = \Psi^{-1}(x,\lambda) i\sigma_- \Phi(x,\lambda), \tag{55}$$

and

$$\frac{\delta T(\lambda)}{\delta u_t(z)} = -\Psi^{-1}(x,\lambda) i\sigma_+ \Phi(x,\lambda). \tag{56}$$

Since

$$\delta T^{-1}(\lambda) = -T^{-1}(\lambda)\,\delta T(\lambda) T^{-1}(\lambda), \tag{57}$$

we can find variation of $T^{-1}(\lambda)$ with respect to $u(x)$, etc.

VII. BASIC POISSON BRACKETS (CONTINUOUS SPECTRUM)

Consider Poisson bracket of $\{T_{ij}(\lambda), T_{kl}^{-1}(\lambda')\}$. From (9), the integrand of the right-hand side is

$$\frac{\delta T_{ij}(\lambda)}{\delta u(x)}\frac{\delta T_{kl}^{-1}(\lambda')}{\delta u_t(x)} + \frac{\delta T_{ij}(\lambda)}{\delta u(x)}\frac{\delta T_{kl}^{-1}(\lambda')}{\delta u_t(x)} - \frac{\delta T_{ij}(\lambda)}{\delta u_t(x)}\frac{\delta T_{kl}^{-1}(\lambda')}{\delta u(x)} - \frac{\delta T_{ij}(\lambda)}{\delta u_t(x)}\frac{\delta T_{kl}^{-1}(\lambda')}{\delta u(x)}. \tag{58}$$

Substituting (52), etc., into (58) we can obtain its explicit expression. With the help of the first Lax equation, we find that (58) can be expressed as (for details, see Appendix B)

$$\frac{1}{2(\lambda-\lambda')}\partial_x\{[\Psi^{-1}(x,\lambda)\Psi(x,\lambda')]_{il}[\Phi^{-1}(x,\lambda')\Phi(x,\lambda)]_{kj}\}. \tag{59}$$

Hence we obtain

$$\{T_{ij}(\lambda), T_{kl}^{-1}(\lambda')\} = \lim_{L\to\infty}\frac{1}{2(\lambda-\lambda')}[\Psi^{-1}(x,\lambda)\Psi(x,\lambda')]_{il}[\Phi^{-1}(x,\lambda')\Phi(x,\lambda)]_{kj}\Big|_{-L}^{L}. \tag{60}$$

Taking account of (13) and (15), the right-hand side is equal to

$$\lim_{L\to\infty}\frac{1}{2(\lambda-\lambda')}[e^{i2(\lambda^2-\lambda'^2)L\sigma_3}]_{il}[T^{-1}(\lambda')e^{-i(\lambda^2-\lambda'^2)L\sigma_3}T(\lambda)]_{kj}, \tag{61}$$

plus

$$-\lim_{L\to\infty}\frac{1}{2(\lambda-\lambda')}[T(\lambda)e^{-i2(\lambda^2-\lambda'^2)L\sigma_3}T^{-1}(\lambda')]_{il}[e^{i2(\lambda^2-\lambda'^2)L\sigma_3}]_{kj}. \tag{62}$$

By employing the following relations:

$$\lim_{L\to\infty}\frac{e^{-i4(\lambda^2-\lambda'^2)L}}{2(\lambda-\lambda')} = -i\frac{1}{2}\pi\delta(\lambda-\lambda'), \tag{63}$$

and

$$\frac{1}{\lambda \pm i0} = \text{p.v.}\frac{1}{\lambda} \mp \pi i \delta(\lambda), \tag{64}$$

where p.v. denotes the Cauchy principal value.[3] Finally we obtain

$$\{a(\lambda),a(\lambda')\}=0, \quad \{a(\lambda),\tilde{a}(\lambda')\}=0, \quad \{\tilde{a}(\lambda),\tilde{a}(\lambda')\}=0, \tag{65}$$

$$\{b(\lambda),b(\lambda')\}=0, \quad \{\tilde{b}(\lambda),\tilde{b}(\lambda')\}=0, \tag{66}$$

$$\{a(\lambda),b(\lambda')\}=\frac{1}{2}\frac{1}{\lambda-\lambda'+i0}a(\lambda)b(\lambda'), \quad \{a(\lambda),\tilde{b}(\lambda')\}=-\frac{1}{2}\frac{1}{\lambda-\lambda'+i0}a(\lambda)\tilde{b}(\lambda'), \tag{67}$$

$$\{\tilde{a}(\lambda),\tilde{b}(\lambda')\}=\frac{1}{2}\frac{1}{\lambda-\lambda'-i0}\tilde{a}(\lambda)\tilde{b}(\lambda'), \quad \{\tilde{a}(\lambda),b(\lambda')\}=-\frac{1}{2}\frac{1}{\lambda-\lambda'-i0}\tilde{a}(\lambda)b(\lambda'), \tag{68}$$

and

$$\{b(\lambda),\tilde{b}(\lambda')\}=i\pi\delta(\lambda-\lambda')|a(\lambda)|^2. \tag{69}$$

In last three formulas λ and λ' are simultaneously real or simultaneously imaginary.

VIII. BASIC POISSON BRACKETS (DISCRETE SPECTRUM)

From (65) we know that $\{\ln a(\lambda),a(\lambda')\}=0$, then by (35), we have

$$\{\ln \breve{a}(\lambda),a(\lambda')\}+\sum_{n=1}^{N}\left(\frac{\{\lambda_n,a(\lambda')\}}{\lambda-\lambda_n}-\frac{\{\overline{\lambda_n},a(\lambda')\}}{\lambda-\overline{\lambda_n}}\right)=0. \tag{70}$$

This expression cannot have pole at λ_n, hence we obtain

$$\{\lambda_n,a(\lambda')\}=0. \tag{71}$$

Similarly, we also have

$$\{\overline{\lambda_n},a(\lambda')\}=0, \quad \{\lambda_n\tilde{a}(\lambda')\}=0, \quad \{\overline{\lambda_n},\tilde{a}(\lambda')\}=0. \tag{72}$$

With the same procedure from $\{\lambda_n,\ln a(\lambda')\}=0$, we obtain

$$\{\lambda_n,\lambda_m\}=0, \quad \{\lambda_n,\overline{\lambda_m}\}=0. \tag{73}$$

From (67), we obtain

$$\{\ln a(\lambda),b(\lambda')\}=-b(\lambda')\frac{1}{2}\frac{1}{\lambda-\lambda'+i0}. \tag{74}$$

Substituting (34), the left-hand side becomes

$$\{\ln \breve{a}(\lambda),b(\lambda')\}+\sum_{n=1}^{N}\left(\frac{\{\lambda_n,b(\lambda')\}}{\lambda-\lambda_n}-\frac{\{\overline{\lambda_n},b(\lambda')\}}{\lambda-\overline{\lambda_n}}\right). \tag{75}$$

Since $\mathrm{Im}\,\lambda^2>0$, λ' is real or pure imaginary, $\lambda-\lambda'+i0\neq 0$, (74) is finite, so that (74) has no poles, thus

$$\{\lambda_n, b(\lambda')\}=0. \tag{76}$$

Similarly, we have

$$\{\overline{\lambda_n}, b(\lambda')\}=0. \tag{77}$$

Poisson brackets involving b_n or $\tilde{b}_m$ can be obtained from those involving $b(\lambda)$ or $\tilde{b}(\lambda)$ by simply assuming $b_n=b(\lambda_n)$ and $\tilde{b}_m=\tilde{b}(\overline{\lambda_m})$, which are the results of the assumption of compact support[3] of u (for the derivation without compact support see Appendix C). From (67), etc., we have

$$\{a(\lambda),b_n\}=\frac{1}{2}\frac{1}{\lambda-\lambda_n}a(\lambda)b_n,\quad \{a(\lambda),\tilde{b}_m\}=-\frac{1}{2}\frac{1}{\lambda-\overline{\lambda_m}}a(\lambda)\tilde{b}_m, \tag{78}$$

$$\{\tilde{a}(\lambda),\tilde{b}_m\}=\frac{1}{2}\frac{1}{\lambda-\overline{\lambda_m}}\tilde{a}(\lambda)\tilde{b}_m,\quad \{\tilde{a}(\lambda),b_n\}=-\frac{1}{2}\frac{1}{\lambda-\lambda_n}\tilde{a}(\lambda)b_n. \tag{79}$$

Corresponding to (69), we have

$$\{b(\lambda),b_n\}=0,\quad \{b(\lambda),\tilde{b}_m\}=0,\quad \{b_n,b_m\}=0,\quad \{b_n,\tilde{b}_m\}=0. \tag{80}$$

By using a similar procedure leading to (76), etc., we have

$$\{\lambda_m,b_n\}=\delta_{mn}\tfrac{1}{2}b_n,\quad \{\overline{\lambda_m},b_n\}=0, \tag{81}$$

and

$$\{\lambda_m,\tilde{b}_n\}=\delta_{mn}\tfrac{1}{2}\tilde{b}_n,\quad \{\lambda_m\tilde{b}_n\}=0. \tag{82}$$

IX. HAMILTONIAN FORMALISM

After integration by parts, we can see that I_2 in (37) is proportional to the Hamiltonian H which is the integral of $\mathcal{H}$ in (7),

$$H=-iI_2. \tag{83}$$

Hence we have

$$H=i2\sum_{n=1}^{N}(\overline{\lambda_n^2}-\lambda_n^2)-\frac{2}{\pi}\int_\Gamma \ln|a(\lambda)|^2\lambda\, d\lambda. \tag{84}$$

The NLS equation in physical space–time can be expressed in the Hamiltonian form

$$\partial_t\pi(x,t)=\{H,\pi(x,t)\}=\frac{\delta H}{\delta u(x,t)},\quad \partial_t u(x,t)=\{H,u(x,t)\}=-\frac{\delta H}{\delta\pi(x,t)}, \tag{85}$$

etc. Substituting (84), the former leads to the complex conjugate of (1) and the latter gives an equality.

Similarly,

$$\partial_t a(t,\lambda)=\{H,a(t,\lambda)\},\quad \partial_t b(t,\lambda)=\{H,b(t,\lambda)\}. \tag{86}$$

We obtain

$$\partial_t a(t,\lambda)=0, \quad a(t,\lambda)=a(0,\lambda). \tag{87}$$

Owing to

$$\{|a(\lambda)|^2,b(\lambda')\}=i\pi\delta(\lambda-\lambda')|a(\lambda)|^2 b(\lambda'), \tag{88}$$

and then

$$\{\ln|a(\lambda)|^2,b(\lambda')\}=i\pi\delta(\lambda-\lambda')b(\lambda'), \tag{89}$$

we obtain

$$\{H,b(\lambda)\}=-\frac{2}{\pi}\int_\Gamma\{\ln|a(\lambda')|^2,b(\lambda)\}\lambda'\,d\lambda'. \tag{90}$$

This implies

$$\partial_t b(\lambda)=\{H,b(\lambda)\}=-i2\lambda b(\lambda), \tag{91}$$

and hence

$$b(\lambda,t)=b(\lambda,0)e^{-i2\lambda t}. \tag{92}$$

Similarly, since we know

$$\{|a(\lambda)|^2,b_n\}=0, \tag{93}$$

and

$$\{\lambda_m^2,b_n\}=2\lambda_m\{\lambda_m,b_n\}=\delta_{mn}\lambda_m b_n, \tag{94}$$

we can find

$$\partial_t b_n=\{H,b_n\}=-i2\lambda_n b_n, \tag{95}$$

and then

$$b_n(t)=b_n(0)e^{-i2\lambda_n t}. \tag{96}$$

X. ACTION-ANGLE VARIABLES

Now we introduce the action-angle variables for the NLS equation in physical space–time (1), for continuous spectrum,

$$P(\lambda)=\frac{1}{\pi}\ln\frac{1}{|a(\lambda)|^2}, \quad Q(\lambda)=\arg b(\lambda), \tag{97}$$

which are real, and for the discrete spectrum,

$$P_n=2\lambda_n, \quad Q_n=\ln\frac{1}{b_n}, \tag{98}$$

which are complex, and hence

$$P_n'=2\,\mathrm{Re}\,\lambda_n, \quad P_n''=2\,\mathrm{Im}\,\lambda_n, \tag{99}$$

$$Q_n' = \ln|b_n|, \quad Q_n'' = -\arg b_n. \tag{100}$$

Since they are also generalized coordinates and conjugated momentum, we must show

$$\{P(\lambda), Q(\lambda')\} = \delta(\lambda - \lambda'), \tag{101}$$

$$\{P_n', Q_m'\} = \{P_n'', Q_m''\} = \delta_{nm}, \tag{102}$$

and all other Poisson brackets vanish.

By definition (97), we know

$$\{e^{i2Q(\lambda)}, e^{i2Q(\lambda')}\} = \left\{ \frac{b(\lambda)}{\overline{\overline{b(\lambda)}}}, \frac{b(\lambda')}{\overline{\overline{b(\lambda')}}} \right\}, \tag{103}$$

the right-hand side obviously vanishes since

$$-\frac{b(\lambda')}{\overline{\overline{b(\lambda)}}\cdot\overline{\overline{b(\lambda')^2}}}\{b(\lambda), \overline{b(\lambda')}\} - \frac{b(\lambda)}{\overline{\overline{b(\lambda)^2}}\cdot\overline{\overline{b(\lambda')}}}\{\overline{b(\lambda)}, b(\lambda)\} = 0. \tag{104}$$

Hence we obtain

$$\{Q(\lambda), Q(\lambda')\} = 0. \tag{105}$$

For a function of $|a(\lambda)|^2$ as $f(|a(\lambda|^2)$, we have

$$\{f(|a(\lambda)|^2), Q(\lambda')\} = \frac{f'(|a(\lambda)|^2)}{i2} \frac{\overline{b(\lambda')}}{b(\lambda')} \left\{ |a(\lambda')|^2, \frac{b(\lambda')}{\overline{\overline{b(\lambda')}}} \right\}. \tag{106}$$

By using (88), etc., we obtain

$$\{f(|a(\lambda)|^2), Q(\lambda')\} = \pi\delta(\lambda - \lambda') \frac{f'(|a(\lambda)|^2)}{2} |a(\lambda)|^2. \tag{107}$$

Setting

$$f(|a(\lambda)|^2) = \frac{1}{\pi} \ln \frac{1}{|a(\lambda)|^2}, \tag{108}$$

(107) leads to (101). On the other hand, it is obvious

$$\{2\lambda_m, \ln b_n\} = \delta_{mn}, \tag{109}$$

and

$$\{2\lambda_m, \ln \tilde{b}_n\} = 0. \tag{110}$$

Furthermore, we must show that action variables are constants while angle variables are dependent on time periodically. Substituting (97) and (99) into (84) we obtain

$$H = 2\sum_n P_n' P_n'' + 2\int_\Gamma P(\lambda')\lambda' \, d\lambda'. \tag{111}$$

By this expression, it is easy to know that the Poisson brackets $\{H, P(\lambda)\}$, $\{H, P_n'\}$, and $\{H, P_n''\}$ vanish, so that $P(\lambda)$, P_n', and P_n'' are constants. On the other hand, since

$$\partial_t Q(\lambda)=\{H,Q(\lambda)\}=2\int_\Gamma \{P(\lambda'),Q(\lambda\}\lambda' d\lambda'=2\lambda, \tag{112}$$

we have

$$Q(\lambda,t)=Q(\lambda,0)+2\lambda t. \tag{113}$$

Similarly, one can find

$$\partial_t Q_n'=\{H,Q_n'\}=2P_n'', \quad \partial_t Q_n''=\{H,Q_n''\}=2P_n', \tag{114}$$

and hence

$$Q_n'(t)=Q_n'(0)+2P_n''t, \quad Q_n''(t)=Q_n'(0)+2P_n't. \tag{115}$$

We thus see $Q(\lambda)$, Q_n', and Q_n'' are indeed angle variables.

XI. CONCLUSION AND DISCUSSION

In previous sections, the Hamiltonian formalism for the NLS equation in physical space–time is developed. Owing to the fact that the equation involves the second partial derivative with respect to time, the generalized coordinates are chosen as u and $\overline{u}$, the canonically conjugated momentum are $\pi=-\overline{u_t}$ and $\overline{\pi}=-u_t$. The first Lax equation in the present case is more complicated, it contains the generalized coordinates as well as the canonically conjugated momentums. We find that the continuous spectrum consists of real λ as well as pure imaginary λ. Though variations of the first Lax equation with respect to u ($\overline{u}$, π, and $\overline{\pi}$) are complicated, the Poisson brackets of transition coefficients are simple, and thus the Hamiltonian formalism in terms of action-angle variables can be achieved.

As we have shown, for dealing with the soliton solution, (1) and (2) are equivalent. This is easy to understand because in the frame moving with the soliton, $u\to 0$ when x or $t\to\infty$. Nevertheless, the initial condition becomes relevant when dealing with a physics system. For example, as perturbation is involved, the initial problem will play an important role.[17] The importance of the initial condition has been pointed out by Wadati *et al.*[11,13–15] In the case of pure imaginary λ, the full Jost solutions shall tend to infinite or vanish as $|t|\to\infty$. The control of instability in physics system has been discussed by Wadati *et al.*[13,15] This problem needs further investigations.

APPENDIX A

The expression of the conservative quantity can be obtained from the asymptotic behavior of the Jost solutions. The first Lax equation (10) can be expressed as

$$\phi_{1x}=(-i2\lambda^2+i|u|^2)\phi_1+(iu_t+2\lambda u)\phi_2, \tag{A1}$$

$$\phi_{2x}=(i2\lambda^2-i|u|^2)\phi_2+(-i\overline{u_t}+2\lambda\overline{u})\phi_1. \tag{A2}$$

Here

$$\phi=\begin{pmatrix}\phi_1\\ \phi_1\end{pmatrix}.$$

Taking the derivative with respect to x in (A1), one can get an equation that just contains ϕ_1 by substituting (A2). Introducing

$$\phi_1(x,\lambda)=e^{(-i2\lambda^2 x+\hat{\phi})}, \tag{A3}$$

we have

$$\mu^2 - i4\lambda^2\mu + |iu_t + 2\lambda u|^2 - |u|^2(4\lambda^2 - |u|^2) = (iu_t + 2\lambda u)\left(\frac{i|u|^2 - \mu}{iu_t + 2\lambda u}\right)_x, \tag{A4}$$

here

$$\mu = \hat{\phi}_x. \tag{A5}$$

From (13), we know that

$$\text{as } |\lambda| \to \infty, \quad \hat{\phi} \to 0 \quad \hat{\phi}_x \to 0. \tag{A6}$$

Consequently, μ can be expanded as

$$\mu = \sum_{j=1}^{\infty} \frac{\mu_j}{(i2\lambda)^j}. \tag{A7}$$

Substituting (A7) back to (A4), one can find

$$\mu_1 = -i(u_t\bar{u} - u\bar{u}_t), \tag{A8}$$

$$\mu_2 = i|u_t|^2 - i|u|^4 + (|u|^2)_x - \bar{u}u_x, \tag{A9}$$

etc. From (A7), we can express $\hat{\phi}$ as

$$\hat{\phi} = \sum_{j=1}^{\infty} \frac{\hat{\phi}_j}{(i2\lambda)^j}. \tag{A10}$$

Here

$$\hat{\phi}_i = \int_{-\infty}^{x} dy\, \mu_i. \tag{A11}$$

On the other hand, from (16), (17), and (13), we have

$$a(t,\lambda) = \lim_{x \to \infty} \phi_1(x,t,\lambda) e^{i2\lambda^2 x}. \tag{A12}$$

Taking account of (A3), one can find

$$\ln a(t,\lambda) = \lim_{x \to \infty} \hat{\phi}(x,t,\lambda). \tag{A13}$$

By using (A10) and (A11), we have

$$\ln a(t,\lambda) = \sum_{j=1}^{\infty} \frac{I_j}{(i2\lambda)^j}. \tag{A14}$$

Here

$$I_j = \int_{-\infty}^{\infty} \mu_j(x,t)\, dx. \tag{A15}$$

Owing to the fact that $a(t,\lambda)$ is t independent, all the I_j are conservative. By substituting (A9) into (A15), we get Eq. (37).

APPENDIX B

From (53) to (57), we can find the explicit expression of (58)

$$\overline{u(x)}\Phi_{k2}^{-1}(x,\lambda')\Psi_{1l}(x,\lambda')[\Psi^{-1}(x,\lambda)\sigma_3\Phi(x,\lambda)]_{ij}-u(x)\Psi_{i1}^{-1}(x,\lambda)\Phi_{2j}(x,\lambda)$$
$$\times[\Phi^{-1}(x,\lambda')\sigma_3\Psi(x,\lambda')]_{kl}+\overline{u(x)}\Psi_{i2}^{-1}(x,\lambda)\Phi_{1j}(x,\lambda)[\Phi^{-1}(x,\lambda')\sigma_3\Psi(x,\lambda')]_{kl}$$
$$-u(x)\Phi_{k1}^{-1}(x,\lambda')\Psi_{2l}(x,\lambda')[\Psi^{-1}(x,\lambda)\sigma_3\Psi(x,\lambda)]_{ij}+2i(\lambda+\lambda')$$
$$\times\{\Psi_{i1}^{-1}(x,\lambda)\Psi_{1l}(x,\lambda')\Phi_{k2}^{-1}(x,\lambda')\Phi_{2j}(x,\lambda)-\Psi_{i2}^{-1}(x,\lambda)\Psi_{2l}(x,\lambda')\Phi_{k1}^{-1}(x,\lambda')\Phi_{1j}(x,\lambda)\}. \tag{B1}$$

On the other hand, by employing the first Lax equation (10), we have

$$\partial_x[\Psi^{-1}(x,\lambda)\Psi(x,\lambda')]_{il}=(\lambda-\lambda')\{[2i(\lambda+\lambda')[\Psi_{i1}^{-1}(x,\lambda)\Psi_{1l}(x,\lambda')$$
$$-\Psi_{i2}^{-1}(x,\lambda)\Psi_{2l}(x,\lambda')]-2[\Psi^{-1}(x,\lambda)U(x,t)\Psi(x,\lambda')]_{il}\} \tag{B2}$$

and

$$\partial_x[\Phi^{-1}(x,\lambda)\Phi(x,\lambda')]_{kj}=(\lambda-\lambda')\{[2i(\lambda+\lambda')[\Phi_{k1}^{-1}(x,\lambda)\Phi_{1j}(x,\lambda')$$
$$-\Phi_{k2}^{-1}(x,\lambda)\Phi_{2j}(x,\lambda')]-2[\Phi^{-1}(x,\lambda)U(x,t)\Phi(x,\lambda')]_{kj}\}. \tag{B3}$$

By using the above equations, we can find out the explicit expression of

$$\partial_x\{[\Psi^{-1}(x,\lambda)\Psi(x,\lambda')]_{il}[\Phi^{-1}(x,\lambda')\Phi(x,\lambda)]_{kj}\}. \tag{B4}$$

One can reach the result of (59) by comparing the expression of (B4) and (B1).

APPENDIX C

In the appendix, we provide the derivation of the Poisson bracket (78) without introducing the assumption of compact support of u. From (22), we have

$$b_n=\frac{\phi_1(x,\lambda_n)}{\psi_1(x,\lambda_n)}=\frac{\phi_2(x,\lambda_n)}{\psi_2(x,\lambda_n)}. \tag{C1}$$

Taking account of this expression and (18), to evaluate the Poisson bracket $\{a(\lambda),b_n\}$ we need to know the Poisson brackets of all components of the Jost solutions.

Since we know that

$$\frac{\delta\Phi(z,\lambda)_{ij}}{\delta u(x)}\neq 0 \quad \text{for } x<z, \tag{C2}$$

etc., and

$$\frac{\delta\Psi(z,\lambda)_{kl}}{\delta u(x)}\neq 0 \quad \text{for } x>z, \tag{C3}$$

etc., the integrand of the Poisson bracket $\{\Phi(x,\lambda)_{ij},\Psi(x,\lambda')_{kl}\}$ vanishes because two factors of each term cannot be unequal to zero simultaneously, this implies that,

$$\{\Phi(x,\lambda)_{ij},\Psi(x,\lambda')_{kl}\}=0. \tag{C4}$$

By the Jacobi property the Poisson bracket of the same components of the Jost solutions vanishes, that is,

$$\{\Phi(x,\lambda)_{ij},\Phi(x,\lambda')_{ij}\}=\{\Psi(x,\lambda)_{ij},\Psi(x,\lambda')_{ij}\}=0. \tag{C5}$$

The nonvanishing Poisson brackets of the components of the Jost solutions needed to evaluate the Poisson bracket $\{a(\lambda),b_n\}$ are $\{\phi_1(x,\lambda),\phi_2(x,\lambda')\}$ and $\{\psi_1(x,\lambda),\psi_2(x,\lambda')\}$. By using the expressions of elements of (42), etc., we find

$$\{\phi_1(x,\lambda),\phi_2(x,\lambda')\}=\frac{1}{2}\frac{1}{\lambda-\lambda'}[\phi_1(x,\lambda)\phi_2(x,\lambda')-\phi_1(x,\lambda')\phi_2(x,\lambda)], \tag{C6}$$

and

$$\{\psi_1(x,\lambda),\psi_2(x,\lambda')\}=-\frac{1}{2}\frac{1}{\lambda-\lambda'}[\psi_1(x,\lambda)\psi_2(x,\lambda')-\psi_1(x,\lambda')\psi_2(x,\lambda)]. \tag{C7}$$

The above formulas are valid when λ or λ' is equal to λ_n.

We now have

$$\{a(\lambda),b_n\}=\{a(\lambda),\phi_1(x,\lambda_n)\}\frac{1}{\psi_1(\lambda_n)}-\{a(\lambda),\psi_1(x,\lambda_n)\}\frac{\phi_1(\lambda_n)}{\psi_1(x,\lambda_n)^2}. \tag{C8}$$

By using the above formulas, we obtain

$$\{a(\lambda),\phi_1(x,\lambda_n)\}=-\psi_1(x,\lambda)\{\phi_2(x,\lambda),\phi_1(x,\lambda_n)\}, \tag{C9}$$

and

$$\{a(\lambda),\psi_1(x,\lambda_n)\}=\phi_1(x,\lambda)\{\psi_2(x,\lambda),\psi_1(x,\lambda_n)\}. \tag{C10}$$

Substituting (C6)–(C10) into (C8), we obtain

$$\{a(\lambda),b_n\}=-\frac{1}{2}\frac{1}{\lambda-\lambda_n}a(\lambda)b_n-\frac{1}{2}\frac{1}{\lambda-\lambda_n}\psi_1(x,\lambda)\phi_1(x,\lambda)\left[\frac{\phi_2(x,\lambda_n)}{\psi_1(x,\lambda_n)}-\frac{\psi_2(x,\lambda_n)\phi_1(x,\lambda_n)}{\psi_1(x,\lambda_n)^2}\right]. \tag{C11}$$

The last term vanishes by (C1). Hence we finally obtain the Poisson bracket of $\{a(\lambda),b_n\}$. Other Poisson brackets involving b_n can be found in the same way.

[1] V. E. Zakharov and A. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[2] L. D. Faddeev, *Mathematical Physics Review Section C.: Mathematical Physical Review* (Harwood, Academic, 1980), pp. 107
[3] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer, Berlin, 1987), and references therein.
[4] B. A. Dubrovin and S. P. Novikov, Sov. Math. Dokl. **26**, 760 (1983).
[5] E. K. Sklyanin, Sov. Phys. Dokl. **24**, 107 (1979).
[6] E. K. Sklyanin, J. Sov. Math. **19**, 1546 (1982).
[7] G. P. Agrawal, *Nonlinear Fiber Optics* (Academic, New York, 1989).
[8] A. Hasegawa and F. Tappert, Appl. Phys. Lett. **23**, 142 (1973).
[9] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. D. Morris, *Solitons and Nonlinear Wave Equation* (Academic, New York, 1982).
[10] N.-N. Huang, *Theory of Soltitons and Method of Pertubations* (Shanghai Scientific and Technological Education Publishing House, Shanghai, 1996) (in Chinese).
[11] T. Yajima and M. Wadati, J. Phys. Soc. Jpn. **59**, 41 (1990); **59**, 3237 (1990).

[12] Z.-d. Chen, X.-J. Chen, and N.-N. Huang, Commun. Theor. Phys. (to be published).
[13] M. Wadati, T. Iizuka, and T. Yajima, Physica D **51**, 388 (1991).
[14] M. Wadati, T. Yajima, and T. Iizuka, Chaos Solitons Fractals **1**, 249 (1991).
[15] M. Wadati, H. Segur, and M. J. Ablowitz, J. Phys. Soc. Jpn. **61**, 1187 (1992).
[16] N.-N. Huang and G. J. Liao, Phys. Lett. A **154**, 373 (1991).
[17] Z.-d. Chen and X.-J. Chen, and N.-N. Huang, Acta Phys. Sin. (Overseas Edition) **8**, 118 (1999).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

The nonlinear Schrödinger equation on the half line

Mario Gattobigio
Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, Dipartimento di Fisica dell'Università di Pisa, Piazza Torricelli 2, 56100 Pisa, Italy

Antonio Liguori
International School for Advanced Studies, 34014 Trieste, Italy

Mihail Mintchev
Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, Dipartimento di Fisica dell'Università di Pisa, Piazza Torricelli 2, 56100 Pisa, Italy

(Received 16 November 1998; accepted for publication 5 January 1999)

The nonlinear Schrödinger equation on the half line with mixed boundary condition is investigated. After a brief introduction to the corresponding classical boundary value problem, the exact second quantized solution of the system is constructed. The construction is based on a new algebraic structure, which is called in what follows boundary algebra and which substitutes, in the presence of boundaries, the familiar Zamolodchikov–Faddeev algebra. The fundamental quantum field theory properties of the solution are established and discussed in detail. The relative scattering operator is derived in the Haag–Ruelle framework, suitably generalized to the case of broken translation invariance in space. © *1999 American Institute of Physics.* [S0022-2488(99)01406-1]

I. INTRODUCTION

The general interest in quantization on the half line $\mathbf{R}_+ = \{x \in \mathbf{R}: x > 0\}$ stems from the recently growing number of applications in different physical areas, including open string theory, dissipative quantum mechanics, and quantum impurity problems. In the last few years, important progress has been made in this subject by means of conformal field theory. Focusing on the nonlinear Schrödinger (NLS) model on $\mathbf{R}_+$, in the present paper we explore the possibility to employ integrability.

Let us recall that when considered on the whole line $\mathbf{R}$, the NLS model represents one of the most extensively studied nonrelativistic integrable systems (see, e.g., Ref. 1). The corresponding equation of motion is

$$(i\partial_t + \partial_x^2)\Phi(t,x) = 2g|\Phi(t,x)|^2\Phi(t,x), \tag{1.1}$$

where $\Phi(t,x)$ is a classical complex field. The model on the half line is obtained restricting Eq. (1.1) on $\mathbf{R}_+$, supplemented with the boundary condition

$$\lim_{x\downarrow 0}(\partial_x - \eta)\Phi(t,x) = 0. \tag{1.2}$$

Here η is a dimensionful parameter of the theory. For $\eta = 0$ and in the limit $\eta \to \infty$ one recovers from Eq. (1.2) the familiar Neumann and Dirichlet boundary conditions respectively. To our knowledge, the boundary value problem (1.1)–(1.2) has been first investigated by Sklyanin[2] and Fokas,[3] who have shown that the integrability, which holds for the system on the whole line, persists also on the half line. Our main goal below will be to construct the exact second quantized solution of Eqs. (1.1) and (1.2), in the case $g \geqslant 0$, $\eta \geqslant 0$. Concretely, this means

(1) To construct a Hilbert space $\mathcal{H}_{g,\eta}$ describing the states of the system;

(2) To define on an appropriate dense domain in $\mathcal{H}_{g,\eta}$ an operator valued distribution $\Phi(t,x)$, $x>0$, satisfying, in a sense that will be made precise below, the equation of motion (1.1), the boundary condition (1.2) and the equal time canonical commutation relations

$$[\Phi(t,x),\Phi(t,y)]=[\Phi^*(t,x),\Phi^*(t,y)]=0, \tag{1.3}$$

$$[\Phi(t,x),\Phi^*(t,y)]=\delta(x-y), \tag{1.4}$$

where Φ^* is the Hermitian conjugate of Φ;

(3) To show the existence of a vacuum state Ω in the above mentioned domain, which is cyclic with respect to the field Φ^*.

The analogous construction in the case of the whole real line has been carried out some years ago[4–9] by means of the quantum inverse scattering transform. The basic algebraic tool of this approach is the Zamolodchikov–Faddeev[10] (ZF) algebra $\mathcal{A}_R$—an appropriate generalization of the canonical commutation relations which incorporates the two-body scattering matrix R. We will show below that the half line system can be treated in the framework of inverse scattering as well, the relevant algebraic structure being now the so called boundary algebra $\mathcal{B}_R$. In the same way as the ZF algebra has been conceived[10] to represent the factorized scattering of integrable systems on the line, the general concept of boundary algebra[11] is inspired by Cherednik's scattering theory[12] of integrable systems on the half line. The fundamental feature of $\mathcal{B}_R$ is that it encodes both the nontrivial scattering between particles and the reflection from the boundary at $x=0$.

A preliminary account without proofs, which partially covers the results presented below, is given in Ref. 13. This paper is organized as follows. In the next section we summarize some known, but useful facts, about the classical NLS model both on $\mathbf{R}$ and $\mathbf{R}_+$. Section III represents a summary of those fundamental properties of $\mathcal{B}_R$ and its Fock representations, which are needed in the quantization. In Sec. IV we define the quantum field $\Phi(t,x)$ and establish its kinematic properties, verifying the canonical commutation relations (1.3)–(1.4). The dynamics is investigated in Sec. V, where it is shown that Eqs. (1.1) and (1.2) are indeed satisfied. We sketch there also the derivation of the correlation functions. Section VI is devoted to the asymptotic theory of the NLS model on $\mathbf{R}_+$. The last section contains our conclusions.

II. THE CLASSICAL NLS MODEL

The study of the classical NLS equation has a long story. Without entering the details, we will collect in this section some basic facts providing useful hints for the quantization.

A. NLS on the real line

The equation of motion (1.1) on $\mathbf{R}$ is obtained by varying the action

$$A[\Phi,\bar{\Phi}]=\int_{\mathbf{R}} dt \int_{\mathbf{R}} dx [i\bar{\Phi}(t,x)\partial_t\Phi(t,x)-|\partial_x\Phi(t,x)|^2-g|\Phi(t,x)|^4]. \tag{2.1}$$

The system admits an infinite number of integrals of motion, the energy

$$E[\Phi,\bar{\Phi}]=\int_{\mathbf{R}} dx[|\partial_x\Phi(t,x)|^2+g|\Phi(t,x)|^4] \tag{2.2}$$

being one of them. Notice that $E[\Phi,\bar{\Phi}]$ is non-negative as long as $g\geqslant 0$. This constraint has an important role in the quantum version of the theory.

About 20 years ago Rosales[14] discovered that Eq. (1.2) on $\mathbf{R}$ admits solutions of the form

$$\Phi(t,x)=\sum_{n=0}^{\infty}(-g)^n\Phi^{(n)}(t,x), \tag{2.3}$$

where

$$\Phi^{(0)}(t,x)\equiv\tilde{\lambda}(t,x)=\int_{\mathbf{R}}\frac{dq}{2\pi}\lambda(q)e^{ixq-itq^2}, \tag{2.4}$$

solves the free Schrödinger equation and

$$\Phi^{(n)}(t,x)=\int_{\mathbf{R}^{2n+1}}\prod_{\substack{i=1\\j=0}}^{n}\frac{dp_i}{2\pi}\frac{dq_j}{2\pi}\overline{\lambda}(p_1)\cdots\overline{\lambda}(p_n)\lambda(q_n)\cdots\lambda(q_0)\frac{e^{i\Sigma_{j=0}^{n}(xq_j-tq_j^2)-i\Sigma_{i=1}^{n}(xp_i-tp_i^2)}}{\Pi_{i=1}^{n}[(p_i-q_{i-1})(p_i-q_i)]}. \tag{2.5}$$

The integration in (2.5) is defined by the principal value prescription and one assumes that $\lambda(k)$ is a function for which the integrals (2.4) and (2.5) exist and the series (2.3) converges uniformly in x for sufficiently small g. It is not difficult to argue that there is a large set of such functions; any λ belonging to the Schwartz test function space $\mathcal{S}(\mathbf{R})$ meets for instance the above requirements. In fact, expressing $\Phi^{(n)}(t,x)$ in terms of $\tilde{\lambda}(t,x)$, one finds

$$\Phi^{(n)}(t,x)=\int_{\mathbf{R}^{2n}}\left[\prod_{i=1}^{n}dy_i dz_i\bar{\tilde{\lambda}}(t,y_i)\tilde{\lambda}(t,z_i)\right]\tilde{\lambda}\left(t,x+\sum_{i=1}^{n}y_i-z_i\right)\sigma(x;y_1,z_1,\ldots,y_n,z_n), \tag{2.6}$$

where

$$\sigma(x;y_1,z_1,\ldots,y_n,z_n)=4^{-n}\prod_{i=1}^{n}\epsilon\left(x+\sum_{j=1}^{i-1}y_j-\sum_{k=1}^{i}z_k\right)\epsilon\left(\sum_{j=1}^{i}(y_j-z_j)\right), \tag{2.7}$$

and $\epsilon(x)$ denotes the sign of x. Therefore,

$$|\Phi^{(n)}(t,x)|\leq\int_{\mathbf{R}^{2n}}\left[\prod_{i=1}^{n}dy_i dz_i|\tilde{\lambda}(t,y_i)\tilde{\lambda}(t,z_i)|\right]\left|\tilde{\lambda}\left(t,x+\sum_{i=1}^{n}y_i-z_i\right)\right|. \tag{2.8}$$

At the other hand, using standard estimates one can deduce that for any $\lambda(k)\in\mathcal{S}(\mathbf{R})$ there exist two positive constants Λ_1 and Λ_2 such that

$$\int_{\mathbf{R}}dx|\tilde{\lambda}(t,x)|\leq\Lambda_1(1+|t|),\quad \sup_{x\in\mathbf{R}}|\tilde{\lambda}(t,x)|\leq\Lambda_2. \tag{2.9}$$

Combining Eqs. (2.8) and (2.9) we conclude that the series (2.3) converges uniformly in x for

$$g<[\Lambda_1(1+|t|)]^{-2}. \tag{2.10}$$

The main reason for focusing on the result of Rosales is because it turns out[5–9] that the general structure of the solutions (2.3)–(2.5) is preserved by the quantization. From this point of view it is instructive to investigate the behavior of (2.3)–(2.5) when the system is restricted on $\mathbf{R}_+$.

B. NLS on the half line

The relative action, giving rise both to the equation of motion (1.1) on $\mathbf{R}_+$ and the boundary condition (1.2) is

$$A[\Phi,\overline{\Phi}]=\int_{\mathbf{R}}dt\int_{\mathbf{R}_+}dx[i\overline{\Phi}(t,x)\partial_t\Phi(t,x)-|\partial_x\Phi(t,x)|^2-g|\Phi(t,x)|^4]-\eta\int_{\mathbf{R}}dt|\Phi(t,0)|^2. \tag{2.11}$$

This action is invariant under time translations, which leads to conservation of the energy

$$E[\Phi,\overline{\Phi}]=\int_{\mathbf{R}_+}dx[|\partial_x\Phi(t,x)|^2+g|\Phi(t,x)|^4]+\eta|\Phi(t,0)|^2. \tag{2.12}$$

Positivity implies $g\geqslant 0$ and $\eta\geqslant 0$, which is the case we are going to analyze below.

The series (2.3), being a solution of the NLS equation on **R**, is *a fortiori* a solution when restricted on $\mathbf{R}_+$. In general however, it does not satisfy the boundary condition (1.2). In this respect, one has the following:

Proposition 1: $\Phi(t,x)$ obeys the boundary condition (1.2), provided that $\lambda(k)$ satisfies

$$\lambda(k)=B(k)\lambda(-k), \tag{2.13}$$

where

$$B(k)=\frac{k-i\eta}{k+i\eta}. \tag{2.14}$$

Proof: Using (2.13), we will show that $\Phi^{(n)}(t,x)$ satisfies (1.2) for any $n\geqslant 0$. For $n=0$ the statement is obvious. So, let us focus on $\Phi^{(n)}(t,x)$ with $n\geqslant 1$. Changing variables in Eq. (2.5) according to

$$k_{2i-1}=p_i\,,\quad k_{2j}=-q_j\,,\quad i=1,...,n,\quad j=0,...,n, \tag{2.15}$$

one finds

$$\lim_{x\downarrow 0}(\partial_x-\eta)\Phi^{(n)}(t,x)=\int_{\mathbf{R}^{2n+1}}\prod_{j=0}^{2n}\frac{dk_j}{2\pi}f^{(n)}(k_0,\ldots,k_{2n})\overline{\lambda}(k_1)\cdots\overline{\lambda}(k_{2n-1})$$
$$\times\lambda(-k_{2n})\cdots\lambda(-k_0)e^{-it\Sigma_{j=0}^{2n}(-1)^jk_j^2}, \tag{2.16}$$

where

$$f^{(n)}(k_0,\ldots,k_{2n})=\frac{\Sigma_{j=0}^{2n}k_j-i\eta}{i\Pi_{j=1}^{2n}(k_j+k_{j-1})}. \tag{2.17}$$

Using the simple relations

$$B(k)B(-k)=B(k)\overline{B}(k)=1, \tag{2.18}$$

one concludes that $f^{(n)}$ in Eq. (2.16) can be equivalently replaced by its B-symmetrized counterpart

$$f_B^{(n)}(k_0,\ldots,k_{2n})=\sum_{\sigma_0,\ldots,\sigma_{2n}\in\{-1,1\}}\frac{1}{4^n}\left(\prod_{j=0}^{2n}\frac{k_j+i\sigma_j\eta}{k_j+i\eta}\right)\frac{\Sigma_{j=0}^{2n}\sigma_jk_j-i\eta}{i\Pi_{j=1}^{2n}(\sigma_jk_j+\sigma_{j-1}k_{j-1})}. \tag{2.19}$$

We shall show now that $f_B^{(n)}$ vanishes identically. Equation (2.19) can be given the more convenient form

$$f_B^{(n)}(k_0,\ldots,k_{2n})=\frac{N^{(n)}(k_0,\ldots,k_{2n})}{4^ni\Pi_{j=0}^{2n}(k_j+i\eta)\Pi_{j=1}^{2n}(k_j^2-k_{j-1}^2)},$$

where

$$N^{(n)}(k_0,\ldots,k_{2n})=\sum_{\sigma_0,\ldots,\sigma_{2n}\in\{-1,1\}}\prod_{j=0}^{2n}(k_j+i\sigma_j\eta)\prod_{j=1}^{2n}(\sigma_jk_j-\sigma_{j-1}k_{j-1})\left(\sum_{j=0}^{2n}\sigma_jk_j-i\eta\right). \tag{2.20}$$

The final step is to prove then that the numerator $N^{(n)}$ vanishes. One way to show the validity of this quite remarkable identity, is to introduce the auxiliary function

$$M^{(n)}(k_0,\ldots,k_{2n})=\sum_{\sigma_0,\ldots,\sigma_{2n}\in\{-1,1\}}(\sigma_0k_0-i\eta)\prod_{j=0}^{2n}(k_j+i\sigma_j\eta)\prod_{j=1}^{2n}(\sigma_jk_j-\sigma_{j-1}k_{j-1}). \tag{2.21}$$

Now, after some algebra one derives the recurrence relations

$$N^{(n)}(k_0,\ldots,k_{2n})=-4k_0k_1(k_1^2+\eta^2)N^{(n-1)}(k_2,\ldots,k_{2n})+4k_0k_1(k_1^2-k_0^2)M^{(n-1)}(k_2,\ldots,k_{2n}), \tag{2.22}$$

$$M^{(n)}(k_0,\ldots,k_{2n})=-4k_0k_1(k_0^2+\eta^2)M^{(n-1)}(k_2,\ldots,k_{2n}). \tag{2.23}$$

Since $N^{(0)}(k_0)=M^{(0)}(k_0)=0$, Eqs. (2.22) and (2.23) imply by induction that

$$N^{(n)}(k_0,\ldots,k_{2n})=0,\quad M^{(n)}(k_0,\ldots,k_{2n})=0, \tag{2.24}$$

which completes the argument.

We conclude here the brief introduction to the classical boundary value problem (1.1)–(1.2). Our next step will be to establish the quantum counterparts of the solution (2.3)–(2.5) and the constraint (2.13).

III. THE BOUNDARY ALGEBRA

As already mentioned in the Introduction, our basic algebraic tool will be a particular associative algebra $\mathcal{B}_R$, whose generators satisfy specific quadratic relations.

A. Definition of $\mathcal{B}_R$

The concept of boundary algebra has been introduced and investigated in a general context in Ref. 11. Here we will consider the following special case. Let $R:\mathbf{R}\times\mathbf{R}\to\mathbf{C}$ be a measurable function satisfying

$$R(k_1,k_2)R(k_2,k_1)=R(k_1,k_2)\overline{R}(k_1,k_2)=1. \tag{3.1}$$

The boundary algebra $\mathcal{B}_R$ is generated by the operator valued distributions $\{a(k),a^*(k),b(k):k\in\mathbf{R}\}$, satisfying quadratic exchange relations, which can be conveniently grouped in two sets. The first one is

$$a(k_1)a(k_2)-R(k_2,k_1)a(k_2)a(k_1)=0, \tag{3.2}$$

$$a^*(k_1)a^*(k_2)-R(k_2,k_1)a^*(k_2)a^*(k_1)=0, \tag{3.3}$$

$$a(k_1)a^*(k_2)-R(k_1,k_2)a^*(k_2)a(k_1)=2\pi\delta(k_1-k_2)+b(k_1)2\pi\delta(k_1+k_2). \tag{3.4}$$

The second set of constraints describes the exchange relations of $b(k)$ and reads

$$a(k_1)b(k_2)=R(k_2,k_1)R(k_1,-k_2)b(k_2)a(k_1), \tag{3.5}$$

$$b(k_2)a^*(k_1)=R(k_2,k_1)R(k_1,-k_2)a^*(k_1)b(k_2), \tag{3.6}$$

$$b(k_1)b(k_2)=b(k_2)b(k_1). \tag{3.7}$$

Notice that if we formally set $b(k)\to 0$, the relations (3.5)–(3.7) trivialize, while (3.2)–(3.4) reproduce the defining relations of the ZF algebra $\mathcal{A}_R$. As it is well known, the factorized scattering of $1+1$ dimensional integrable systems is encoded in $\mathcal{A}_R$, i.e., in a boundary algebra in which the so called boundary operator $b(k)$ is trivially implemented. On the contrary, it turns out[11] that whenever there is a reflecting boundary, one needs a *reflection* boundary algebra, i.e., a boundary algebra with the additional constraint

$$b(k)b(-k)=1, \tag{3.8}$$

which obviously prevents the boundary operator from being zero. In the case of the NLS on the half line, we shall need a reflection boundary algebra $\mathcal{B}_R$ with exchange factor

$$R(k_1,k_2)=\frac{k_1-k_2-ig}{k_1-k_2+ig}, \tag{3.9}$$

where $g\geqslant 0$ is the coupling constant of the NLS model. $R(k_1,k_2)$ is actually the two-body bulk scattering matrix of the NLS model[4–9] and satisfies (3.1).

B. Fock representations

Following some basic ideas of Ref. 15, we have constructed in Ref. 11 the Fock representations of $\mathcal{B}_R$. These representations are characterized by the existence of a vacuum state Ω, which is cyclic with respect $a^*(k)$ and satisfies

$$a(k)\Omega=0. \tag{3.10}$$

In the reflection case (3.8), the vacuum is[11] always an eigenvector of the boundary operator $b(k)$, i.e.,

$$b(k)\Omega=B(k)\Omega, \tag{3.11}$$

where $B(k)$ is a measurable function obeying Eq. (2.18). Conversely, any $B(k)$ of this type defines a Fock representation on a Hilbert space $\mathcal{F}_{R,B}$, whose vacuum satisfies (3.11). We will show below that the state space $\mathcal{H}_{g,\eta}$ of the NLS model on $\mathbf{R}_+$ is

$$\mathcal{H}_{g,\eta}=\mathcal{F}_{R,B}, \tag{3.12}$$

with B and R given by (2.14) and (3.9), respectively. The mere fact that our system has a boundary shows up at the algebraic level, turning the ZF algebra into a reflection boundary algebra $\mathcal{B}_R$, i.e., forcing a nonzero boundary operator $b(k)$. The details of the boundary condition (the value of the parameter η) enter at the representation level through the reflection coefficient $B(k)$. In the Fock space $\mathcal{F}_{R,B}$ one has

$$a(k)=b(k)a(-k), \tag{3.13}$$

$$a^*(k)=a^*(-k)b(-k), \tag{3.14}$$

which descend from a peculiar automorphism of $\mathcal{B}_R$, established in Ref. 11. The relation (3.13) turns out to be the correct quantum analogue of Eq. (2.13). Let us stress once more that the c-number reflection coefficient $B(k)$ must be distinguished from the boundary generator $b(k)$, which according to Eqs. (3.5) and (3.6) does not even commute with $\{a(k),a^*(k)\}$.

To the end of this section we will give some details about the structure of $\mathcal{F}_{R,B}$ which are needed for our construction. One has

$$\mathcal{F}_{R,B} \equiv \bigoplus_{n=0}^{\infty} \mathcal{H}_{R,B}^n, \tag{3.15}$$

where $\mathcal{H}_{R,B}^0 \equiv \mathbf{C}$ and the n-particle space $\mathcal{H}_{R,B}^n$ with $n \geqslant 1$ is a subspace of $L^2(\mathbf{R}^n)$ defined as follows:

(i) a L^2-function $\varphi(p_1)$ belongs to $\mathcal{H}_{R,B}^1$ if and only if

$$\varphi(p_1) = B(p_1)\varphi(-p_1); \tag{3.16}$$

(ii) a L^2-function $\varphi(p_1,\ldots,p_n)$ with $n \geqslant 2$ belongs to $\mathcal{H}_{R,B}^n$ if and only if

$$\varphi(p_1,\ldots,p_{n-1},p_n) = B(p_n)\varphi(p_1,\ldots,p_{n-1},-p_n), \tag{3.17}$$

and

$$\varphi(p_1,\ldots,p_i,p_{i+1},\ldots,p_n) = R(p_i,p_{i+1})\varphi(p_1,\ldots,p_{i+1},p_i,\ldots,p_n), \tag{3.18}$$

for any $1 \leqslant i \leqslant n-1$.

Equations (3.16)–(3.18) define a closed subspace $\mathcal{H}_{R,B}^n \subset L^2(\mathbf{R}^n)$. We will denote by $P_{R,B}^{(n)}$ the corresponding orthogonal projection operator. We introduce also the finite particle space $\mathcal{F}_{R,B}^0 \subset \mathcal{F}_{R,B}$, generated by $\{\mathcal{H}_{R,B}^n : n = 0,1,\ldots\}$. We recall that $\mathcal{F}_{R,B}^0$ is the linear space of sequences $\varphi = (\varphi^{(0)}, \varphi^{(1)}, \ldots, \varphi^{(n)}, \ldots)$ with $\varphi^{(n)} \in \mathcal{H}_{R,B}^n$ and $\varphi^{(n)} = 0$ for n large enough. The vacuum state is $\Omega = (1,0,\ldots,0,\ldots)$. The L^2-scalar product on $\mathcal{H}_{R,B}^n$ defines in the standard way the scalar product $\langle \cdot,\cdot \rangle$ in the (Hilbert) direct sum (3.15).

At this point we are in position to define on $\mathcal{F}_{R,B}^0$ the annihilation and creation operators $\{a(f), a^*(f) : f \in L^2(\mathbf{R})\}$. We set $a(f)\Omega = 0$ and

$$[a(f)\varphi]^{(n)}(p_1,\ldots,p_n) = \sqrt{n+1}\int_{\mathbf{R}} \frac{dp}{2\pi}\, \overline{f}(p)\, \varphi^{(n+1)}(p,p_1,\ldots,p_n), \tag{3.19}$$

$$[a^*(f)\varphi]^{(n)}(p_1,\ldots,p_n) = \sqrt{n}[P_{R,B}^{(n)} f \otimes \varphi^{(n-1)}](p_1,\ldots,p_n), \tag{3.20}$$

for all $\varphi \in \mathcal{F}_{R,B}^0$. The operators $a(f)$ and $a^*(f)$ are in general unbounded on $\mathcal{F}_{R,B}^0$. One can easily see however that $a(f)$ and $a^*(f)$ are bounded on each $\mathcal{H}_{R,B}^n$. In fact, for all $\varphi \in \mathcal{H}_{R,B}^n$ one has the estimates

$$\|a(f)\varphi\| \leqslant \sqrt{n}\|f\|\|\varphi\|, \quad \|a^*(f)\varphi\| \leqslant \sqrt{n+1}\|f\|\|\varphi\|, \tag{3.21}$$

$\|\cdot\|$ being the L^2-norm. Notice also that $a^*(f)$ is linear in f, whereas $a(f)$ is antilinear. The operator-valued distributions $a(p)$ and $a^*(p)$, generating the Fock representation of $\mathcal{B}_R$, are defined by

$$a(f) = \int_{\mathbf{R}} \frac{dp}{2\pi}\, \overline{f}(p) a(p), \quad a^*(f) = \int_{\mathbf{R}} \frac{dp}{2\pi}\, f(p) a^*(p), \tag{3.22}$$

and are related by Hermitian conjugation, namely

$$\langle \varphi, a(f)\psi \rangle = \langle a^*(f)\varphi, \psi \rangle, \quad \forall \varphi, \psi \in \mathcal{F}_{R,B}^0. \tag{3.23}$$

Finally, the action of the boundary generator $b(p)$ on $\mathcal{F}_{R,B}^0$ is defined by Eq. (3.11) and

$$[b(p)\varphi]^{(n)}(p_1,\ldots,p_n)$$
$$=[R(p,p_1)R(p,p_2)\cdots R(p,p_n)B(p)R(p_n,-p)\cdots R(p_2,-p)R(p_1,-p)]\varphi^{(n)}(p_1,\ldots,p_n). \tag{3.24}$$

One can show[11] that $\{a(p),a^*(p),b(p)\}$, defined above, indeed satisfy the exchange relations (3.2)–(3.7) and the reflection condition (3.8). Moreover, the vacuum Ω obeys the requirements formulated in the beginning of this subsection.

It is convenient to introduce here a domain $\mathcal{D}\subset\mathcal{F}_{R,B}$, which will be frequently used in what follows. Setting

$$\mathcal{D}^0\equiv\mathbf{C},\quad \mathcal{D}^n\equiv\left\{\int_{\mathbf{R}^n}dp_1\cdots dp_n f(p_1,\ldots,p_n)a^*(p_1)\cdots a^*(p_n)\Omega : f\in\mathcal{S}(\mathbf{R}^n), n\geqslant 1\right\}, \tag{3.25}$$

we define $\mathcal{D}$ to be the linear space of sequences $\varphi=(\varphi^{(0)},\varphi^{(1)},\ldots,\varphi^{(n)},\ldots)$, where $\varphi^{(n)}\in\mathcal{D}^n$ and $\varphi^{(n)}$ vanish for n large enough. By construction $\mathcal{D}$ is a proper subspace of $\mathcal{F}^0_{R,B}$. Nevertheless, $\mathcal{D}$ is dense in $\mathcal{F}_{R,B}$ as well. Indeed, using that the factors R and B are smooth (i.e., C^∞) bounded functions, one has that $\mathcal{D}^n$ is dense in $\mathcal{H}^n_{R,B}$, which implies the statement. We observe that

$$a(f)\mathcal{D}\subset\mathcal{D},\quad a^*(f)\mathcal{D}\subset\mathcal{D},\quad \forall f\in\mathcal{S}(\mathbf{R}). \tag{3.26}$$

Notice also that the matrix elements of $a^*(k)$ between states from $\mathcal{D}$ are smooth functions of k. More generally, one has

$$\langle\varphi,a^*(k_1)\cdots a^*(k_n)\psi\rangle\in\mathcal{S}(\mathbf{R}^n),\quad \forall\varphi,\psi\in\mathcal{D}. \tag{3.27}$$

Summarizing, we introduced in this section the boundary algebra $\mathcal{B}_R$ and its Fock representation $\mathcal{F}_{R,B}$, which are the main ingredients in the construction of the quantum solution of the boundary value problem (1.1)–(1.2).

IV. QUANTIZATION

A. The quantum field $\Phi(t,x)$

Our first step will be to introduce the quantum analog of $\Phi^{(n)}(t,x)$. For this purpose we consider

$$\Phi^{(0)}(t,x)\equiv\tilde{a}(t,x)=\int_{\mathbf{R}}\frac{dq}{2\pi}a(q)e^{ixq-itq^2}, \tag{4.1}$$

$$\Phi^{(n)}(t,x)=\int_{\mathbf{R}^{2n+1}}\prod_{\substack{i=1\\ j=0}}^{n}\frac{dp_i}{2\pi}\frac{dq_j}{2\pi}a^*(p_1)\cdots a^*(p_n)a(q_n)\cdots a(q_0)$$
$$\times\frac{e^{i\Sigma_{j=0}^n(xq_j-tq_j^2)-i\Sigma_{i=1}^n(xp_i-tp_i^2)}}{\Pi_{i=1}^n[(p_i-q_{i-1}-i\epsilon)(p_i-q_i-i\epsilon)]}, \tag{4.2}$$

thus replacing formally $\{\lambda(p),\overline{\lambda}(p)\}$ in Eqs. (2.4) and (2.5) by the generators $\{a(p),a^*(p)\}$ of $\mathcal{B}_R$ in the Fock representation $\mathcal{F}_{R,B}$ and fixing an $i\epsilon$ prescription to contour poles. Our first task will be to give meaning of $\Phi^{(n)}(t,x)$ as a quadratic form in $\mathcal{D}$.

Proposition 2: For any $\varphi,\psi\in\mathcal{D}$, the expectation value

$$\langle\varphi,\Phi^{(n)}(t,x)\psi\rangle, \tag{4.3}$$

is a C^∞ function of t,x.

Proof: The case $n=0$ is trivial. For $n\geqslant 1$ it is enough to take $\varphi\in\mathcal{D}^m$ and $\psi\in\mathcal{D}^{m+1}$ with $m>n$. Some elementary algebra leads to

$$\langle\varphi,\Phi^{(n)}(t,x)\psi\rangle=\int_{\mathbf{R}^{m+n+1}}\prod_{i_1=1}^{n}\frac{dp_{i_1}}{2\pi}\prod_{i_2=0}^{n}\frac{dq_{i_2}}{2\pi}\prod_{i_3=n+1}^{m}\frac{dk_{i_3}}{2\pi}\overline{\varphi}(p_1,\ldots,p_n,k_{n+1},\ldots,k_m)$$
$$\times\frac{e^{i\Sigma_{j=0}^{n}(xq_j-tq_j^2)-i\Sigma_{i=1}^{n}(xp_i-tp_i^2)}}{\Pi_{i=1}^{n}[(p_i-q_{i-1}-i\epsilon)(p_i-q_i-i\epsilon)]}\psi(q_0,\ldots,q_n,k_{n+1},\ldots,k_m),\qquad(4.4)$$

which, using that φ and ψ are Schwartz test functions, implies the proposition.

Taking into account that $\mathcal{D}$ contains only finite particle vectors, we conclude that also $\Phi(t,x)$ is a quadratic form on $\mathcal{D}$, smooth in both t and x. The conjugate $\Phi^*(t,x)$ is defined by

$$\langle\varphi,\Phi^*(t,x)\psi\rangle=\overline{\langle\psi,\Phi(t,x)\varphi\rangle},\qquad(4.5)$$

which is of course smooth in t and x as well. The counterparts of Eqs. (4.1) and (4.2) read

$$\Phi^{*(0)}(t,x)\equiv\tilde{a}^*(t,x)=\int_{\mathbf{R}}\frac{dq}{2\pi}a^*(q)e^{-ixq+itq^2},\qquad(4.6)$$

$$\Phi^{*(n)}(t,x)=\int_{\mathbf{R}^{2n+1}}\prod_{\substack{i=1\\ j=0}}^{n}\frac{dp_i}{2\pi}\frac{dq_j}{2\pi}a^*(q_0)\cdots a^*(q_n)a(p_n)\cdots a(p_1)$$
$$\times\frac{e^{i\Sigma_{i=1}^{n}(xp_i-tp_i^2)-i\Sigma_{j=0}^{n}(xq_j-tq_j^2)}}{\Pi_{i=1}^{n}[(p_i-q_{i-1}+i\epsilon)(p_i-q_i+i\epsilon)]}.\qquad(4.7)$$

Since the system we are considering is in $\mathbf{R}_+$, we adopt the smearing

$$\Phi(t,f)=\int dx\overline{f}(x)\Phi(t,x),\quad \Phi^*(t,f)=\int dx f(x)\Phi^*(t,x),\quad f\in C_0^\infty(\mathbf{R}_+),\qquad(4.8)$$

where $C_0^\infty(\mathbf{R}_+)$ is the set of infinitely differentiable functions with compact support in $\mathbf{R}_+$. Again, $\Phi(t,f)$ and $\Phi^*(t,f)$ have meaning as quadratic forms on $\mathcal{D}$, which are related by

$$\langle\varphi,\Phi^*(t,f)\psi\rangle=\overline{\langle\psi,\Phi(t,f)\varphi\rangle}.\qquad(4.9)$$

In order to formulate some other less obvious properties of $\Phi(t,f)$ and $\Phi^*(t,f)$, we have to introduce the following partial ordering relation in $C_0^\infty(\mathbf{R}_+)$. Let $f_1,\ f_2\in C_0^\infty(\mathbf{R}_+)$. Then

$$f_1<f_2\Leftrightarrow x_1<x_2\forall x_1\in\operatorname{supp}f_1,\quad\forall x_2\in\operatorname{supp}f_2.\qquad(4.10)$$

Instead of $f_1<f_2$, we will also write $f_2>f_1$. Denoting by $\tilde{a}^*(t,f)$ the operator

$$\tilde{a}^*(t,f)=\int dx f(x)\tilde{a}^*(t,x),\qquad(4.11)$$

one can prove the following technical

Lemma 1: Let $\varphi,\psi\in\mathcal{D}$.

(a) *The identity*

$$\langle\varphi,\Phi^*(t,h)\tilde{a}^*(t,f)\psi\rangle=\langle\varphi,\tilde{a}^*(t,f)\Phi^*(t,h)\psi\rangle,\qquad(4.12)$$

holds if $h<f$;

(b) *One has*

$$\langle\varphi,\Phi^*(t,h)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega\rangle=\langle\varphi,\tilde{a}^*(t,h)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega\rangle, \tag{4.13}$$

provided that $h>f_j$ *for any* $j=1,...,n$;

(c) *For any* $f_1>f_2>\cdots>f_n$, *one has*

$$\langle\varphi,\Phi(t,h)\tilde{a}^*(t,f_1)\tilde{a}^*(t,f_2)\cdots\tilde{a}^*(t,f_n)\Omega\rangle=\sum_{j=1}^{n}(h,f_j)\langle\varphi,\tilde{a}^*(t,f_1)\cdots\hat{\tilde{a}}^*(t,f_j)\cdots\tilde{a}^*(t,f_n)\Omega\rangle, \tag{4.14}$$

where $(\cdot,\cdot)$ *denotes the* L^2*-scalar product and the hat indicates that the corresponding field must be omitted.*

Proof: The proof of the identities (4.12)–(4.14) is analogous to that given by Davies[8] for the NLS on $\mathbf{R}$, so we skip it. We only remark that the novelty on $\mathbf{R}_+$ consists in evaluating the contributions of the boundary generator b, which stem from the exchange of a and a^*. It is easy to see that these contributions actually vanish, due to the support requirements imposed on the test functions and the condition $\eta\geqslant 0$.

Summarizing, $\Phi(t,f)$ and $\Phi^*(t,f)$ have been so far defined as quadratic forms on $\mathcal{D}$ and are Schwartz distributions with respect to f. Our main goal to the end of this subsection will be to show that $\Phi(t,f)$ and $\Phi^*(t,f)$ are actually well defined operators. In order to construct a common invariant domain for these operators, we introduce the subspace

$$\mathcal{D}_0^n\equiv \mathrm{sp}\{\tilde{a}^*(t,f_1)\tilde{a}^*(t,f_2)\cdots\tilde{a}^*(t,f_n)\Omega: f_1>f_2>\cdots>f_n\}\subset\mathcal{H}_{R,B}^n,\quad n\geqslant 1, \tag{4.15}$$

where sp indicates the linear span and $t\in\mathbf{R}$ is arbitrary but fixed. Setting $\mathcal{D}_0^0=\mathbf{C}$, we define $\mathcal{D}_0$ to be the linear space of sequences $\varphi=(\varphi^{(0)},\varphi^{(1)},\ldots,\varphi^{(n)},\ldots)$ with $\varphi^{(n)}\in\mathcal{D}_0^n$ and $\varphi^{(n)}=0$ for n large enough. Both $\mathcal{D}$ and $\mathcal{D}_0$ are subspaces of the finite particle space $\mathcal{F}_{R,B}^0$. We know already that $\mathcal{D}$ is dense in $\mathcal{F}_{R,B}$. Although it is less obvious, the same is true for $\mathcal{D}_0$.

Proposition 3: $\mathcal{D}_0$ *is dense in* $\mathcal{F}_{R,B}$.

Proof. It is enough to demonstrate that the space $\mathcal{D}_0^n$ is dense in $\mathcal{H}_{R,B}^n$ for any $t\in\mathbf{R}$ and $n\geqslant 1$. So, let us consider the matrix element

$$\tilde{A}_{t,\varphi}(x_1,\ldots,x_n)\equiv\langle\varphi,\tilde{a}^*(t,x_1)\cdots\tilde{a}^*(t,x_n)\Omega\rangle, \tag{4.16}$$

where $\varphi\in\mathcal{D}^n$ is arbitrary. According to Eq. (3.27), $\tilde{A}_{t,\varphi}\in\mathcal{S}(\mathbf{R}^n)$. In order to prove the statement, it is sufficient to show that

$$\tilde{A}_{t,\varphi}(x_1,\ldots,x_n)=0,\quad \forall x_1>x_2>\cdots>x_n>0, \tag{4.17}$$

implies $\varphi=0$. It is convenient for this purpose to investigate

$$A_{t,\varphi}(p_1,\ldots,p_n)\equiv\int_{\mathbf{R}^n}\prod_{j=1}^{n}dx_j e^{i\Sigma_{j=1}^n p_j x_j}\tilde{A}_{t,\varphi}(x_1,\ldots,x_n)$$
$$=e^{it\Sigma_{j=1}^n p_j^2}\langle\varphi,a^*(p_1)\cdots a^*(p_n)\Omega\rangle\in\mathcal{S}(\mathbf{R}^n). \tag{4.18}$$

The behavior of this function under the reflection of one of its arguments or the exchange of two consecutive arguments is determined by Eqs. (3.3), (3.6), (3.11), (3.14). Using this fact, one can verify that the function

$$B_{t,\varphi}(p_1,\ldots,p_n)\equiv\Lambda(p_1,\ldots,p_n)A_{t,\varphi}(p_1,\ldots,p_n), \tag{4.19}$$

where

$$\Lambda(p_1,\ldots,p_n)\equiv\prod_{j=1}^{n}\left[(p_j-i\eta)\prod_{\substack{k=1\\k>j}}^{n}(p_j-p_k-ig)(p_j+p_k-ig)\right] \tag{4.20}$$

satisfies

$$B_{t,\varphi}(p_1,\ldots,p_j,\ldots,p_n)=-B_{t,\varphi}(p_1,\ldots,-p_j,\ldots,p_n),\quad \forall j=1,\ldots,n, \tag{4.21}$$

$$B_{t,\varphi}(p_1,\ldots,p_j,p_{j+1},\ldots,p_n)=-B_{t,\varphi}(p_1,\ldots,p_{j+1},p_j\ldots,p_n),\quad \forall j=1,\ldots,n-1. \tag{4.22}$$

By construction $B_{t,\varphi}\in\mathcal{S}(\mathbf{R}^n)$ and

$$\widetilde{B}_{t,\varphi}(x_1,\ldots,x_n)=\int_{\mathbf{R}^n}\prod_{j=1}^{n}\frac{dp_j}{2\pi}e^{-i\Sigma_{j=1}^{n}p_jx_j}B_{t,\varphi}(p_1,\ldots,p_n)=\Lambda(i\partial_1,\ldots,i\partial_n)\widetilde{A}_{t,\varphi}(x_1,\ldots,x_n), \tag{4.23}$$

admits the same antisymmetry properties as $B_{t,\varphi}$. Therefore, using the smoothness of $\widetilde{A}_{t,\varphi}$ and Eq. (4.17), we deduce that $\widetilde{B}_{t,\varphi}$ vanishes identically, or equivalently,

$$B_{t,\varphi}(p_1,\ldots,p_n)=0,\quad \forall p_j\in\mathbf{R}. \tag{4.24}$$

Combining Eqs. (4.18), (4.19), (4.24) with the fact that $\Lambda(p_1,\ldots,p_n)\neq 0$ for any $p_j\in\mathbf{R}$, one gets

$$\langle\varphi,a^*(p_1)\cdots a^*(p_n)\Omega\rangle=0,\quad \forall p_j\in\mathbf{R}, \tag{4.25}$$

which, because of the cyclicity of Ω with respect to a^*, implies $\varphi=0$. This concludes the argument.

It is convenient in what follows to have an explicit formula for the scalar product in $\mathcal{D}_0$. It is provided by the following:

Lemma 2: Let $f_1\succ f_2\succ\cdots\succ f_n$ and $h_1\succ h_2\succ\cdots\succ h_n$. Then

$$\langle\tilde{a}^*(t,h_1)\cdots\tilde{a}^*(t,h_n)\Omega,\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega\rangle=(h_1\otimes\cdots\otimes h_n,f_1\otimes\cdots\otimes f_n). \tag{4.26}$$

Proof: It is enough to expand the left hand side, using the algebraic relations (3.4) and Eq. (3.10). Taking into account the support properties of the test functions involved, all terms, except the one in the right-hand side of (4.26), vanish.

A simple corollary of the previous lemma is now in order. Since any $\varphi\in\mathcal{D}_0^n$ can be represented as

$$\varphi=\sum_{\alpha\in A}\tilde{a}^*(t,f_1^\alpha)\cdots\tilde{a}^*(t,f_n^\alpha)\Omega,$$

where A is a finite set and $f_1^\alpha\succ f_2^\alpha\succ\cdots\succ f_n^\alpha$ for all $\alpha\in A$, one has that

$$\langle\varphi,\varphi\rangle^2\equiv\|\varphi\|^2=\left\|\sum_{\alpha\in A}f_1^\alpha\otimes\cdots\otimes f_n^\alpha\right\|^2. \tag{4.27}$$

We are now in position to show the following:

Proposition 4: The estimate

$$|\langle\varphi,\Phi(t,f)\psi\rangle|\leqslant(n+1)\|f\|\|\varphi\|\|\psi\| \tag{4.28}$$

holds for any $\varphi\in\mathcal{D}_0^n$, $\psi\in\mathcal{D}_0^{n+1}$, and $f\in C_0^\infty(\mathbf{R}_+)$.

Proof: Let

$$\varphi = \sum_{\alpha \in A} \tilde{a}^*(t,f_1^\alpha)\cdots\tilde{a}^*(t,f_n^\alpha)\Omega, \quad \psi = \sum_{\beta \in B} \tilde{a}^*(t,h_0^\beta)\cdots\tilde{a}^*(t,h_n^\beta)\Omega, \tag{4.29}$$

with $f_1^\alpha > f_2^\alpha > \cdots > f_n^\alpha$ and $h_0^\beta > h_1^\beta > \cdots > h_n^\beta$. Then

$$\begin{aligned}\langle \varphi, \Phi(t,f)\psi\rangle &= \sum_{\alpha \in A}\sum_{\beta \in B} \langle \tilde{a}^*(t,f_1^\alpha)\cdots\tilde{a}^*(t,f_n^\alpha)\Omega, \Phi(t,f)\tilde{a}^*(t,h_0^\beta)\cdots\tilde{a}^*(t,h_n^\beta)\Omega\rangle \\ &= \sum_{\alpha \in A}\sum_{\beta \in B}\sum_{j=0}^{n} (f,h_j^\alpha)(f_1^\alpha\otimes\cdots\otimes f_n^\alpha, h_0^\beta\otimes\cdots\otimes\hat{h}_j^\beta\otimes\cdots\otimes h_n^\beta) \\ &= \sum_{j=0}^{n}\left(\sum_{\alpha\in A} f_1^\alpha\otimes\cdots\otimes f_{j-1}^\alpha\otimes f\otimes f_j^\alpha\otimes\cdots\otimes f_n^\alpha, \sum_{\beta\in B} h_0^\beta\otimes\cdots\otimes h_j^\beta\otimes\cdots\otimes h_n^\beta\right),\end{aligned}$$

where use has been made of point (c) of Lemma 1. Applying now the Minkowski inequality, one finds

$$|\langle \varphi, \Phi(t,f)\psi\rangle| \leq \sum_{j=0}^{n} \|f\| \left\| \sum_{\alpha\in A} f_1^\alpha\otimes\cdots\otimes f_n^\alpha\right\| \left\| \sum_{\beta\in B} h_0^\beta\otimes\cdots\otimes h_n^\beta\right\| \leq (n+1)\|f\|\|\varphi\|\|\psi\|. \tag{4.30}$$

The above proposition shows that $\Phi(t,f)$, considered as quadratic form, is bounded on $\mathcal{D}_0^n \times \mathcal{D}_0^{n+1}$ and defines therefore a bounded operator $\mathcal{H}_{R,B}^{n+1} \to \mathcal{H}_{R,B}^n$. Since this occurs for any $n \geq 0$, we recover an operator $\Phi(t,f):\mathcal{F}_{R,B}^0 \to \mathcal{F}_{R,B}^0$, whose properties are collected in

Theorem 1: *$\Phi(t,f):\mathcal{F}_{R,B}^0 \to \mathcal{F}_{R,B}^0$ is a linear operator, satisfying*

$$\Phi(t,f)\Omega = 0, \quad \Phi(t,f):\mathcal{H}_{R,B}^{n+1} \to \mathcal{H}_{R,B}^n, \quad n \geq 0. \tag{4.31}$$

Moreover, for any $\varphi,\psi \in \mathcal{F}_{R,B}^0$, the matrix element $\langle \varphi, \Phi(t,f)\psi\rangle$ has the following properties:

(i) *It is antilinear and L^2-continuous in f;*
(ii) *It is continuous in $t \in \mathbf{R}$;*
(iii) *It is smooth in $t \in \mathbf{R}$, provided that $\varphi,\psi \in \mathcal{D}$.*

Proof: All the statements are simple corollaries of the above propositions.

The operator $\Phi(t,f)$ is densely defined and admits therefore a Hermitian conjugate $\Phi^*(t,f)$.

Theorem 2: *The field $\Phi^*(t,f)$ satisfies*

$$\Phi^*(t,f)\Omega = \tilde{a}^*(t,f)\Omega, \quad \Phi^*(t,f):\mathcal{H}_{R,B}^n \to \mathcal{H}_{R,B}^{n+1}, \quad n \geq 0, \tag{4.32}$$

and therefore leaves $\mathcal{F}_{R,B}^0$ invariant. Moreover

$$\langle \varphi, \Phi(t,f)\psi\rangle = \langle \Phi^*(t,f)\varphi, \psi\rangle, \tag{4.33}$$

holds for any $\varphi,\psi \in \mathcal{F}_{R,B}^0$.

Proof: One uses the fact that $\Phi(t,f)$ is bounded on each $\mathcal{H}_{R,B}^n$.

We will show now that the operators $\Phi(t,f)$ and $\Phi^*(t,f)$ satisfy the basic requirements for nonrelativistic quantum fields.

B. Cyclicity of Ω and commutation relations

We start with

Theorem 3 (Cyclicity)**:** *The vacuum Ω is a cyclic vector for the field Φ^*. More precisely the space*

$$\mathcal{E}_0^n \equiv \text{sp}\{\Phi^*(t,f_1)\Phi^*(t,f_2)\cdots\Phi^*(t,f_n)\Omega : f_1 \prec f_2 \prec \cdots \prec f_n\},$$

is dense in $\mathcal{H}_{R,B}^n$.

Proof: Using Eqs. (4.12)–(4.13) of Lemma 1, one easily proves by induction that

$$\Phi^*(t,f_1)\Phi^*(t,f_2)\cdots\Phi^*(t,f_n)\Omega = \tilde{a}^*(t,f_n)\cdots\tilde{a}^*(t,f_1)\Omega, \tag{4.34}$$

as long as $f_1 \prec f_2 \prec \cdots \prec f_n$. Thus $\mathcal{E}_0^n = \mathcal{D}_0^n$, and the statement follows directly from Proposition 3.

Remark: Theorem 3 is slightly stronger than the standard cyclicity,[16] because of the ordering among the functions $f_1,\ldots,f_n$ required in the definition of $\mathcal{E}_0^n$.

Let us consider now the canonical commutation relations (1.3) and (1.4). We shall prove

Theorem 4: *The equal time canonical commutation relations*

$$[\Phi(t,h_1),\Phi(t,h_2)] = [\Phi^*(t,h_1),\Phi^*(t,h_2)] = 0, \tag{4.35}$$

$$[\Phi(t,h_1),\Phi^*(t,h_2)] = (h_1,h_2), \tag{4.36}$$

hold on $\mathcal{F}_{R,B}^0$ *for any* $h_1,h_2 \in \mathcal{S}(\mathbf{R}_+)$.

Proof: In order to demonstrate Eq. (4.35), we observe that Eq. (4.14) implies

$$\Phi(t,h_2)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega = \sum_{j=1}^{n} (h_2,f_j)\tilde{a}^*(t,f_1)\cdots\hat{\tilde{a}}^*(t,f_j)\cdots\tilde{a}^*(t,f_n)\Omega,$$

where $f_1 \succ \cdots \succ f_n$. Therefore,

$$\Phi(t,h_1)\Phi(t,h_2)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega$$
$$= \sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} (h_2,f_j)(h_1,f_k)\tilde{a}^*(t,f_1)\cdots\hat{\tilde{a}}^*(t,f_j)\cdots\hat{\tilde{a}}^*(t,f_k)\cdots\tilde{a}^*(t,f_n)\Omega, \tag{4.37}$$

which, being symmetric under the exchange of h_1 with h_2, implies the vanishing of $[\Phi(t,h_1),\Phi(t,h_2)]$ on $\mathcal{D}_0^n$. Then one extends by continuity to $\mathcal{H}_{R,B}^n$ and by linearity to $\mathcal{F}_{R,B}^0$. The validity of $[\Phi^*(t,h_1),\Phi^*(t,h_2)] = 0$ follows applying Hermitian conjugation.

We turn now to Eq. (4.36). Let $f_1 \succ \cdots \succ f_n$ and $h_1,h_2 \in \mathcal{S}(\mathbf{R}_+)$. Assume that

$$f_k \succ h_2 \succ f_{k+1}. \tag{4.38}$$

Using Lemma 1, one gets

$$\Phi(t,h_1)\Phi^*(t,h_2)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega$$
$$= \Phi(t,h_1)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_k)\Phi^*(t,h_2)\tilde{a}^*(t,f_{k+1})\cdots\tilde{a}^*(t,f_n)\Omega$$
$$= \Phi(t,h_1)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_k)\tilde{a}^*(t,h_2)\tilde{a}^*(t,f_{k+1})\cdots\tilde{a}^*(t,f_n)\Omega$$
$$= (h_1,h_2)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega$$
$$+ \sum_{j=1}^{n} (h_1,f_j)\tilde{a}^*(t,f_1)\cdots\hat{\tilde{a}}^*(t,f_j)\cdots\tilde{a}^*(t,f_k)\tilde{a}^*(t,h_2)\tilde{a}^*(t,f_{k+1})\cdots\tilde{a}^*(t,f_n)\Omega.$$

Analogously,

$$\Phi^*(t,h_2)\Phi(t,h_1)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega$$
$$= \sum_{j=1}^{n} (h_1,f_j)\tilde{a}^*(t,f_1)\cdots\hat{\tilde{a}}^*(t,f_j)\cdots\tilde{a}^*(t,f_k)\tilde{a}^*(t,h_2)\tilde{a}^*(t,f_{k+1})\cdots\tilde{a}^*(t,f_n)\Omega.$$

Therefore,

$$[\Phi(t,h_1),\Phi^*(t,h_2)]\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega=(h_1,h_2)\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega. \tag{4.39}$$

So, Eq. (4.36) holds on states of the type $\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega$, which satisfy the condition (4.38). Observing that the couples $\{h_2,\tilde{a}^*(t,f_1)\cdots\tilde{a}^*(t,f_n)\Omega\}$ obeying (4.38) are norm dense in $L^2(\mathbf{R}_+)\otimes\mathcal{H}^n_{RB}$, Eq. (4.36) follows by continuity.

As a consequence of the commutation relations (4.35) and (4.36) one has the following useful estimate:

Proposition 5: Let A be a finite set and let $f_1^\alpha,\ldots,f_n^\alpha\in C_0^\infty$ *for any* $\alpha\in A$. *Then the norm of the operator*

$$\sum_{\alpha\in A}\Phi(t,f_1^\alpha)\Phi(t,f_2^\alpha)\cdots\Phi(t,f_n^\alpha),$$

restricted to $\mathcal{H}^m_{R,B}$ *with* $m\geq n$, *satisfies*

$$\left\|\sum_{\alpha\in A}\Phi(t,f_1^\alpha)\Phi(t,f_2^\alpha)\cdots\Phi(t,f_n^\alpha)\right\|\leq\sqrt{m(m-1)\cdots(m-n+1)}\left\|\sum_{\alpha\in A}f_1^\alpha\otimes\cdots\otimes f_n^\alpha\right\|. \tag{4.40}$$

Proof: Let $\psi\in\mathcal{D}_0^n$. Then there is some finite set B, such that ψ can be written in the form

$$\psi=\sum_{\beta\in B}\Phi^*(t,h_1^\beta)\cdots\Phi^*(t,h_m^\beta)\Omega,\quad h_1^\beta<\cdots<h_m^\beta.$$

Now, by means of the commutation relations (4.35) and (4.36) one finds

$$\left\|\sum_{\alpha\in A}\Phi(t,f_1^\alpha)\cdots\Phi(t,f_n^\alpha)\psi\right\|\leq\sqrt{m(m-1)\cdots(m-n+1)}\left\|\sum_{\alpha\in A}f_1^\alpha\otimes\cdots\otimes f_n^\alpha\right\|\|\psi\|, \tag{4.41}$$

implying Eq. (4.40) by continuity.

V. TIME EVOLUTION

In order to investigate the time evolution in the NLS model on $\mathbf{R}_+$, we consider the mapping

$$\alpha_t(a(k))=e^{-ik^2t}a(k),\quad \alpha_t(a^*(k))=e^{ik^2t}a^*(k),\quad \alpha_t(b(k))=b(k),\quad t\in\mathbf{R}. \tag{5.1}$$

It is straightforward to verify that α_t defines a 1-parameter group of automorphisms of the boundary algebra $\mathcal{B}_R$. Using the relations (3.2)–(3.6), (3.13), (3.14), one can easily check that this group is unitarily implemented in the Fock space $\mathcal{F}_{R,B}$ by means of the operator

$$U(t)=\exp(iHt),\quad H=\frac{1}{2}\int_{\mathbf{R}}\frac{dk}{2\pi}k^2a^*(k)a(k). \tag{5.2}$$

The Hamiltonian H acts on $\mathcal{D}$ according to

$$[H\varphi]^{(n)}(k_1,\ldots,k_n)=(k_1^2+\cdots+k_n^2)\varphi^{(n)}(k_1,\ldots,k_n), \tag{5.3}$$

which implies that the domain $\mathcal{D}$ is invariant both under $U(t)$ and H. Moreover, since

$$-i\frac{d}{dt}U(t)\bigg|_{t=0}=H, \tag{5.4}$$

on $\mathcal{D}$, the latter is a domain of essential self-adjointness for H.

The crucial point now is that the time evolution of the field $\Phi(t,f)$ is given by

$$\Phi(t,f)=U(t)\Phi(0,f)U(t)^{-1}. \tag{5.5}$$

This fact follows directly from the time dependence encoded in Eqs. (4.1) and (4.2) and is quite remarkable. It shows the power of both the quantum inverse scattering transform (4.2) and the algebra $\mathcal{B}_R$, which combined together allow to write down the Hamiltonian of an interacting field theory as a simple quadratic expression in a and a^*. In this form H depends only implicitly on the coupling constant g through the exchange factor R. Notice also that the boundary generator b does not evolve in time.

A. The quantum equation of motion

A preliminary problem to be faced here is to give a precise meaning on the quantum level of the cubic term $|\Phi(t,x)|^2\Phi(t,x)$ present in Eq. (1.1). For this purpose we will follow the standard approach, introducing the concept of a normal ordered :...: product involving Φ and Φ^*. As usually assumed, in such a product all creation operators a^* stand to the left of all annihilation operators a. In view of Eqs. (3.2) and (3.3) in our case one must further specify the ordering of creators and annihilators themselves. We define :...: to preserve the original order of the creators. The original order of two annihilators is preserved if both belong to the same Φ or Φ^* and inverted otherwise. The quantum version of Eq. (1.1) is then obtained by the substitution

$$|\Phi(t,x)|^2\Phi(t,x)\mapsto :\Phi\Phi^*\Phi:(t,x). \tag{5.6}$$

Concerning the relation between the above way of defining the normal product and the alternative point-splitting procedure, we observe that

$$:\Phi\Phi^*\Phi:(t,x)=\lim_{\sigma\downarrow 0}\Phi(t,x+2\sigma)\Phi^*(t,x+\sigma)\Phi(t,x), \tag{5.7}$$

holds in mean value on $\mathcal{D}$. Following Ref. 6, Eq. (5.7) can be derived by using the analyticity properties of the commutator between $a(p)$ and $\Phi(t,x)$. One can formulate at this point

Theorem 5: *The nonlinear Schrödinger equation,*

$$(i\partial_t+\partial_x^2)\langle\varphi,\Phi(t,x)\psi\rangle=2g\langle\varphi,:\Phi\Phi^*\Phi:(t,x)\psi\rangle, \tag{5.8}$$

is satisfied for any $\varphi,\psi\in\mathcal{D}$.

Proof: The first step is analogous to the proof of Proposition 2 and consists in showing that the matrix element $\langle\varphi,:\Phi\Phi^*\Phi:(t,x)\psi\rangle$ is smooth in t and x for any $\varphi,\psi\in\mathcal{D}$. The next step is to compare $(i\partial_t+\partial_x^2)\langle\varphi,\Phi^{(n)}(t,x)\psi\rangle$ with the $(n-1)$th order term in the expansion of $\langle\varphi,:\Phi\Phi^*\Phi:(t,x)\psi\rangle$ in terms of g. A straightforward computation, similar to that performed in Ref. 8 for the NLS model on **R**, shows that these terms indeed coincide.

B. Boundary conditions

We shall demonstrate now

Theorem 6: *The following boundary conditions hold for any* $\varphi,\psi\in\mathcal{D}$, *and* $t\in\mathbf{R}$,

$$\lim_{x\downarrow 0}(\partial_x-\eta)\langle\varphi,\Phi(t,x)\psi\rangle=0, \tag{5.9}$$

$$\lim_{x\to\infty}\langle\varphi,\Phi(t,x)\psi\rangle=0. \tag{5.10}$$

Let us first prove

Lemma 3: Let $\varphi,\psi\in\mathcal{F}^0_{R,B}$. *There exists a vector* $\chi\in\mathcal{H}^1_{R,B}$ *such that*

$$\langle \varphi, \Phi(t,f)\psi\rangle = \langle \Omega, \Phi(t,f)\chi\rangle. \tag{5.11}$$

Proof: Without loss of generality one can take $\varphi \in \mathcal{H}^n_{R,B}$, $\psi \in \mathcal{H}^{n+1}_{R,B}$. Suppose first that $\varphi \in \mathcal{E}^n_0 = \mathcal{D}^n_0$. Then φ is of the form

$$\varphi = \sum_{\alpha\in A} \Phi^*(t,f_1^\alpha)\Phi^*(t,f_2^\alpha)\cdots\Phi^*(t,f_n^\alpha)\Omega, \tag{5.12}$$

where A is a finite set and $f_1^\alpha \prec f_2^\alpha \prec \cdots \prec f_n^\alpha$ for all $\alpha\in A$. Using the commutation relations (4.35) and (4.36) one easily obtains

$$\langle \varphi, \Phi(t,f)\psi\rangle = \sum_{\alpha\in A} \langle \Omega, \Phi(t,f)\Phi(t,f_n^\alpha)\phi(t,f_{n-1}^\alpha)\cdots\Phi(t,f_1^\alpha)\psi\rangle. \tag{5.13}$$

In order to solve (5.11), it is then sufficient to define

$$\chi = \Phi(t,f_n^\alpha)\Phi(t,f_{n-1}^\alpha)\cdots\Phi(t,f_1^\alpha)\psi, \tag{5.14}$$

which belongs to $\mathcal{H}^1_{R,B}$ since $\psi\in\mathcal{H}^{n+1}_{R,B}$. Take now a general $\varphi\in\mathcal{H}^n_{R,B}$. By cyclicity (Theorem 3), there exists a sequence $\{\varphi_k\}\subset\mathcal{D}^n_0$ converging to φ. By Proposition 5, the corresponding vectors $\{\chi_k\}$ given by Eq. (5.14) form a Cauchy sequence, which converges to a vector $\chi\in\mathcal{H}^1_{R,B}$, satisfying (5.11) by continuity.

We can now prove Theorem 6.

Proof: Let $\varphi,\psi\in\mathcal{D}_0\subset\mathcal{F}^0_{R,B}$. From the lemma above there exists $\chi\in\mathcal{H}^1_{R,B}$ such that

$$\langle \varphi, \Phi(t,x)\psi\rangle = \langle \Omega, \Phi(t,x)\chi\rangle = \int_{\mathbf{R}} \frac{dk}{2\pi} e^{ikx - ik^2 t}\chi(k). \tag{5.15}$$

Since $\chi\in L^2$, the matrix element $\langle\varphi,\Phi(t,x)\psi\rangle$, which by Proposition 2 is smooth, is also square integrable with respect to x. Therefore it vanishes at infinity and Eq. (5.10) is satisfied. Moreover, taking the derivative with respect to x, the B-symmetry (3.16) of χ, immediately leads to Eq. (5.9).

C. Correlation functions

From the general structure of our solution it follows that

(i) the nonvanishing correlation functions involve equal number of Φ and Φ^*;
(ii) for computing the exact $2n$-point function one does not need all terms in the expansion (2.3), but at most the $(n-1)$th order contribution.

One has, for instance,

$$\langle \Omega, \Phi(t_1,x_1)\Phi^*(t_2,x_2)\Omega\rangle = \langle \Omega, \Phi^{(0)}(t_1,x_1)\Phi^{*(0)}(t_2,x_2)\Omega\rangle, \tag{5.16}$$

$$\begin{aligned}\langle \Omega, \Phi(t_1,x_1)\Phi(t_2,x_2)\Phi^*(t_3,x_3)\Phi^*(t_4,x_4)\Omega\rangle \\ = \langle \Omega, \Phi^{(0)}(t_1,x_1)\Phi^{(0)}(t_2,x_2)\Phi^{*(0)}(t_3,x_3)\Phi^{*(0)}(t_4,x_4)\Omega\rangle \\ + g^2\langle \Omega, \Phi^{(0)}(t_1,x_1)\Phi^{(1)}(t_2,x_2)\Phi^{*(1)}(t_3,x_3)\Phi^{*(0)}(t_4,x_4)\Omega\rangle.\end{aligned} \tag{5.17}$$

Since the vacuum expectation value of any number of $\{a(k),a^*(k),b(k)\}$ is known explicitly,[11] employing Eqs. (4.1), (4.2), (4.6), (4.7) one can derive integral representations for the NLS correlation functions on $\mathbf{R}_+$. For example,

$$\langle \Omega, \Phi(t_1,x_1)\Phi^*(t_2,x_2)\Omega\rangle = \int_{\mathbf{R}} \frac{dp}{2\pi} e^{-ip^2(t_1-t_2)}[e^{ip(x_1-x_2)} + B(p)e^{ip(x_1+x_2)}], \tag{5.18}$$

which coincides with that of the nonrelativistic free field on the half line. In spite of this fact, the four-point function (5.17) differs from the free one. We would like to recall in this respect that according to Jost's theorem (see, e.g., Ref. 16), such a phenomenon is forbidden in relativistic invariant models.

VI. SCATTERING THEORY

As it is well known, integrable quantum systems on the real line are characterized by a factorized scattering matrix. This means that multiparticle scattering is described by an appropriate product of two-particle scattering matrices, which in turn are subject to physical constraints like unitarity, crossing symmetry, etc.

Some years ago, Cherednik[12] proposed a version of factorized scattering, adapted to the half line case. The following physical picture emerges from his investigation. Let $|k_1,\ldots,k_n\rangle^{\text{in}}$ be an in-state, representing n particles coming from $x=+\infty$ and thus having negative momenta $k_1<k_2<\cdots<k_n<0$. These particles interact among themselves before and after being reflected by the wall at $x=0$, giving rise to an out-state $|p_1,\ldots,p_m\rangle^{\text{out}}$ composed of particles traveling towards $x=+\infty$ and thus having positive momenta $p_1>p_2>\cdots>p_m>0$. The transition amplitude between these states vanishes unless $n=m$ and $p_i=-k_i$, $i=1,\ldots,n$. Therefore, not only the total momentum, but each momentum is separately reflected. According to Ref. 12, the scattering amplitude is

$$^{\text{out}}\langle p_1,\ldots,p_m|k_1,\ldots,k_n\rangle^{\text{in}}=\delta_{mn}\prod_{i=1}^{n} 2\pi\delta(p_i+k_i)B(p_i)\prod_{\substack{i,j=1\\ i<j}}^{n} R(p_i,p_j)R(p_i,-p_j). \tag{6.1}$$

The R-factors describe the interactions among the particles in the bulk, while the B-factors take into account the reflection from the wall.

The main goal of this section is to prove that the NLS model on $\mathbf{R}_+$ perfectly fits the scheme of Cherednik. In order to do that, we must develop first the scattering theory corresponding to the off-shell quantum field $\Phi^*(t,f)$. Our framework will be the conventional Haag–Ruelle approach,[17] suitably adapted to the nonrelativistic case.

A first relation between the quantum solutions (4.6), (4.7) and Cherednik's scattering amplitude (6.1) is obtained through the identification

$$|p_1,\ldots,p_n\rangle^{\text{out}}=a^*(p_1)\cdots a^*(p_n)\Omega,\quad p_1>\cdots>p_n>0, \tag{6.2}$$

$$|k_1,\ldots,k_n\rangle^{\text{in}}=a^*(k_1)\cdots a^*(k_n)\Omega,\quad k_1<\cdots<k_n<0. \tag{6.3}$$

We recall in fact that $\mathcal{B}_R$ has been designed in such a way, that the amplitudes

$$\langle a^*(p_1)\cdots a^*(p_m)\Omega,\, a^*(k_1)\cdots a^*(k_n)\Omega\rangle, \tag{6.4}$$

precisely reproduce the right-hand side of Eq. (6.1). What is still missing therefore is the construction of suitable states, expressed in terms of $\Phi^*(t,h)$ and Ω, which approach the out-states (6.2) for $t\to\infty$ and the in-states (6.3) for $t\to-\infty$. We are now going to fill this gap.

Proposition 5 shows that $\Phi^*(t,f)$, restricted on $\mathcal{H}^n_{R,B}$ is a bounded operator of norm

$$\|\Phi^*(t,f)\|\leq\sqrt{n+1}\|f\|, \tag{6.5}$$

which in turn implies that it can be extended to any $f\in L^2(\mathbf{R}_+)$. From the estimates (3.21) we know that also $a^*(h)$ is bounded on $\mathcal{H}^n_{R,B}$, where

$$\|a^*(h)\|\leq\sqrt{n+1}\|h\|,\quad \forall h\in L^2(\mathbf{R}). \tag{6.6}$$

Combining this inequality with the definition (4.6), one finds,

$$\left\| \int_{\mathbf{R}^n} dx_1 \cdots dx_n f(x_1,\ldots,x_n) \tilde{a}^*(t,x_1)\cdots\tilde{a}^*(t,x_n)\Omega \right\| \leqslant \sqrt{n!}\|f\|, \quad \forall f \in L^2(\mathbf{R}^n). \tag{6.7}$$

In order to develop the Haag–Ruelle formalism, we will need also the following notations. Let $h(k) \in \mathcal{S}(\mathbf{R})$. Then we set

$$h^t(x) \equiv \int_{\mathbf{R}} \frac{dk}{2\pi} e^{ikx - ik^2 t} h(k), \quad h^t_+(x) \equiv \theta(x)[h^t(x) + h^t(-x)], \quad \tilde{h}(k) \equiv h(-k), \tag{6.8}$$

where $\theta(x)$ is the Heaviside step function. Notice that

$$\tilde{h}^t_+(x) = \theta(x)[\tilde{h}^t(x) + \tilde{h}^t(-x)] = \theta(x)[h^t(-x) + h^t(x)] = h^t_+(x). \tag{6.9}$$

We are now in position to formulate

Theorem 7: (Asymptotic states) *Let*

$$h_1 \succ h_2 \succ \cdots \succ h_n, \quad h_j \in \mathcal{S}(\mathbf{R}_+), \quad j = 1,\ldots,n.$$

Then one has the following strong limits:

$$\lim_{t\to+\infty} \Phi^*(t,h^t_{1+})\Phi^*(t,h^t_{2+})\cdots\Phi^*(t,h^t_{n+})\Omega = a^*(h_1)a^*(h_2)\cdots a^*(h_n)\Omega, \tag{6.10}$$

$$\lim_{t\to-\infty} \Phi^*(t,h^t_{1+})\Phi^*(t,h^t_{2+})\cdots\Phi^*(t,h^t_{n+})\Omega = a^*(\tilde{h}_1)a^*(\tilde{h}_2)\cdots a^*(\tilde{h}_n)\Omega. \tag{6.11}$$

For proving this statement, we need some preliminary results.

Lemma 4: Let $h \in \mathcal{S}(\mathbf{R}_+)$. *Then,*

$$\lim_{t\to+\infty} \|h^t_+ - h^t\| = 0, \quad \lim_{t\to-\infty} \|h^t_+ - \tilde{h}^t\| = 0. \tag{6.12}$$

Proof: A direct computation gives

$$\|h^t_+ - h^t\|^2 = 2i \int_{\mathbf{R}^2} \frac{dk}{2\pi}\frac{dp}{2\pi} \overline{h}(k)h(p) \frac{e^{it(k+p)(k-p)}}{k-p+i\epsilon}, \tag{6.13}$$

$$\|h^t_+ - \tilde{h}^t\|^2 = -2i \int_{\mathbf{R}^2} \frac{dk}{2\pi}\frac{dp}{2\pi} \overline{h}(k)h(p) \frac{e^{it(k+p)(k-p)}}{k-p-i\epsilon}. \tag{6.14}$$

Now, for proving Eq. (6.12), it is enough to take into account that supp $h > 0$ and to use the weak limit

$$\lim_{t\to\pm\infty} \frac{e^{itk}}{k \pm i\epsilon} = 0. \tag{6.15}$$

Corollary 1: Let $h_1, h_2, \ldots, h_n \in \mathcal{S}(\mathbf{R}_+)$. *Then,*

$$\lim_{t\to+\infty} \|h^t_{1+}\otimes\cdots\otimes h^t_{n+} - h^t_1\otimes\cdots\otimes h^t_n\| = 0, \tag{6.16}$$

$$\lim_{t\to-\infty} \|h^t_{1+}\otimes\cdots\otimes h^t_{n+} - \tilde{h}^t_1\otimes\cdots\otimes \tilde{h}^t_n\| = 0. \tag{6.17}$$

Lemma 5: Let $h_1, h_2 \in \mathcal{S}(\mathbf{R}_+)$ are such that $h_1 \succ h_2$. Then, the functions

$$H^t(x_1,x_2)=h_1^t(x_1)h_2^t(x_2)\,\theta(x_2-x_1), \tag{6.18}$$

$$\tilde{H}^t(x_1,x_2)=\tilde{h}_1^t(x_1)\tilde{h}_2^t(x_2)\,\theta(x_2-x_1), \tag{6.19}$$

satisfy

$$\lim_{t\to+\infty}\|H^t\|=0, \quad \lim_{t\to-\infty}\|\tilde{H}^t\|=0. \tag{6.20}$$

Proof: Let us consider for instance H^t. One has

$$\begin{aligned}\|H^t\|^2&=\int_{-\infty}^{\infty}dx_1\int_{x_1}^{\infty}dx_2|h_1^t(x_1)h_2^t(x_2)|^2\\&=\int_{\mathbf{R}^4}\frac{dk_1}{2\pi}\,\frac{dp_1}{2\pi}\,\frac{dk_2}{2\pi}\,\frac{dp_2}{2\pi}\overline{h_1}(k_1)h_1(p_1)\overline{h_2}(k_2)h_2(p_2)I(k_1,p_1,k_2,p_2)e^{(k_1^2-p_1^2+k_2^2-p_2^2)t},\end{aligned} \tag{6.21}$$

with

$$I(k_1,p_1,k_2,p_2)\equiv\int_{-\infty}^{\infty}dx_1\int_{x_1}^{\infty}dx_2e^{i[(p_1-k_1)x_1+(p_2-k_2)x_2]}.$$

The integration in x_1 and x_2 gives

$$I(k_1,p_1,k_2,p_2)=\frac{2\pi i\,\delta(k_1-p_1+k_2-p_2)}{p_2-k_2+i\epsilon}. \tag{6.22}$$

Therefore,

$$\|H^t\|^2=i\int_{\mathbf{R}^3}\frac{dp_1}{2\pi}\,\frac{dp_2}{2\pi}\,\frac{dk_2}{2\pi}\overline{h_1}(p_1-p_2+k_2)h_1(p_1)\overline{h_2}(k_2)h_2(p_2)\,\frac{e^{2i(p_1-k_2)(p_2-k_2)t}}{p_2-k_2+i\epsilon}. \tag{6.23}$$

The support properties of the function h_1 and h_2 imply that the integrand vanishes unless $p_1>k_2>0$, which completes the argument because of Eq. (6.15). Analogous considerations apply to $\tilde{H}^t$.

Corollary 2: Let

$$G^t(x_1,x_2)=h_{1+}^t(x_1)h_{2+}^t(x_2)\,\theta(x_2-x_1). \tag{6.24}$$

Then,

$$\lim_{t\to\pm\infty}\|G^t\|=0. \tag{6.25}$$

Proof: One has to combine Eqs. (6.9), (6.16), (6.17), (6.20).

The statement of Corollary 2 has the following generalization to the case of $n\geqslant 2$ variables. Suppose that $h_1,\ldots,h_n\in\mathcal{S}(\mathbf{R}_+)$ and $h_1\succ\cdots\succ h_n$. Let $\mathcal{P}_n$ be the group of all permutations of the indices $\{1,2,...,n\}$. For any $\sigma\in\mathcal{P}_n$ we define the function

$$G_\sigma^t(x_1,\ldots,x_n)\equiv h_{1+}^t(x_1)\cdots h_{n+}^t(x_n)\,\theta(x_{\sigma_1},\ldots,x_{\sigma_n}), \tag{6.26}$$

where

$$\theta(x_{\sigma_1},\ldots,x_{\sigma_n})\equiv\prod_{\substack{i,j=1\\ i<j}}^{n}\theta(x_{\sigma_i}-x_{\sigma_j}). \tag{6.27}$$

Corollary 3: For any $\sigma\in\mathcal{P}_n$ different from the identity $e=(1,2,...,n)$, one has

$$\lim_{t\to\pm\infty}\|G^t_\sigma\|=0. \tag{6.28}$$

We are now ready to prove Theorem 7.

Proof: The case $n=1$ is quite simple. Using the identities

$$\Phi^*(t,f)\Omega=\tilde{a}^*(t,f)\Omega,\quad a^*(h)=\tilde{a}^*(t,h^t), \tag{6.29}$$

one finds

$$\|\Phi^*(t,h^t_+)\Omega-a^*(h)\Omega\|=\|\tilde{a}^*(t,h^t_+)\Omega-\tilde{a}^*(t,h^t)\Omega\|\leqslant\|h^t_+-h^t\|, \tag{6.30}$$

which according to Lemma 5 tends to 0 in the limit $t\to+\infty$. Let us consider now the case $n\geqslant 2$. Applying Eq. (4.34) and

$$\theta(x_1,\ldots,x_n)=1-\sum_{\substack{\sigma\in\mathcal{P}_n\\ \sigma\neq e}}\theta(x_{\sigma_1},\ldots,x_{\sigma_n}), \tag{6.31}$$

we get

$$\begin{aligned}
&\Phi^*(t,h^t_{1+})\cdots\Phi^*(t,h^t_{n+})\Omega\\
&\quad=\int_{\mathbf{R}^n}dx_1\cdots dx_n h^t_{1+}(x_1)\cdots h^t_{n+}(x_n)\sum_{\sigma\in\mathcal{P}_n}\theta(x_{\sigma_1},\ldots,x_{\sigma_n})\tilde{a}^*(t,x_{\sigma_1})\cdots\tilde{a}^*(t,x_{\sigma_n})\Omega\\
&\quad=\tilde{a}^*(t,h^t_{1+})\cdots\tilde{a}^*(t,h^t_{n+})\Omega+\sum_{\substack{\sigma\in\mathcal{P}_n\\ \sigma\neq e}}\int_{\mathbf{R}^n}dx_1\cdots dx_n G^t_\sigma(x_1,\ldots,x_n)[\tilde{a}^*(t,x_{\sigma_1})\cdots\tilde{a}^*(t,x_{\sigma_n})\Omega\\
&\qquad-\tilde{a}^*(t,x_1)\cdots\tilde{a}^*(t,x_n)\Omega].
\end{aligned} \tag{6.32}$$

The estimate (6.7) then leads to

$$\begin{aligned}
&\|\Phi^*(t,h^t_{1+})\cdots\Phi^*(t,h^t_{n+})\Omega-a^*(h_1)\cdots a^*(h_n)\Omega\|\\
&\quad=\|\Phi^*(t,h^t_{1+})\cdots\Phi^*(t,h^t_{n+})\Omega-\tilde{a}^*(t,h^t_1)\cdots\tilde{a}^*(t,h^t_n)\Omega\|\\
&\quad\leqslant\|\tilde{a}^*(t,h^t_{1+})\cdots\tilde{a}^*(t,h^t_{n+})\Omega-\tilde{a}^*(t,h^t_1)\cdots\tilde{a}^*(t,h^t_n)\Omega\|+2\sqrt{n!}\sum_{\substack{\sigma\in\mathcal{P}_n\\ \sigma\neq e}}\|G^t_\sigma\|\\
&\quad\leqslant\sqrt{n!}\|h^t_{1+}\otimes\cdots\otimes h^t_{n+}-h^t_1\otimes\cdots\otimes h^t_n\|+2\sqrt{n!}\sum_{\substack{\sigma\in\mathcal{P}_n\\ \sigma\neq e}}\|G^t_\sigma\|,
\end{aligned} \tag{6.33}$$

which implies the strong limit (6.10). Analogous considerations give

$$\|\Phi^*(t,h_{1+}^t)\cdots\Phi^*(t,h_{n+}^t)\Omega - a^*(\tilde{h}_1)\cdots a^*(\tilde{h}_n)\Omega\| \leqslant \sqrt{n!}\|h_{1+}^t\otimes\cdots\otimes h_{n+}^t - \tilde{h}_1^t\otimes\cdots\otimes\tilde{h}_n^t\| + 2\sqrt{n!}\sum_{\substack{\sigma\in\mathcal{P}_n\\ \sigma\neq e}}\|G_\sigma^t\|, \tag{6.34}$$

which proves (6.11).

We proceed with the construction of the scattering operator S, following the general strategy developed in Ref. 18. According to Theorem 7, the asymptotic spaces $\mathcal{F}^{\rm out}$ and $\mathcal{F}^{\rm in}$ are generated by finite linear combinations of the vectors $(n\geqslant 1)$,

$$\mathcal{E}^{\rm out}=\{\Omega, a^*(h_1)\cdots a^*(h_n)\Omega : h_1 > \cdots > h_n, h_j\in\mathcal{S}(\mathbf{R}_+)\} \tag{6.35}$$

and

$$\mathcal{E}^{\rm in}=\{\Omega, a^*(\tilde{h}_1)\cdots a^*(\tilde{h}_n)\Omega : h_1 > \cdots > h_n, h_j\in\mathcal{S}(\mathbf{R}_+)\}, \tag{6.36}$$

respectively. One can show[11] moreover, that $\mathcal{F}^{\rm out}$ and $\mathcal{F}^{\rm in}$ are separately dense $\mathcal{F}_{R,B}$. This property of asymptotic completeness allows to demonstrate[11] that the mapping $S:\mathcal{E}^{\rm out}\to\mathcal{E}^{\rm in}$, defined by

$$S\Omega=\Omega, \tag{6.37}$$

$$Sa^*(h_1)a^*(h_2)\cdots a^*(h_n)\Omega = a^*(\tilde{h}_1)a^*(\tilde{h}_2)\cdots a^*(\tilde{h}_n)\Omega, \tag{6.38}$$

extends to a unitary scattering operator on $\mathcal{F}_{R,B}$. We stress that S is nontrivial, in spite of the fact that the quantum fields Φ and Φ^* realize a Fock representation of the canonical commutation relations. This feature is not in contradiction with Haag's theorem,[16] because we are dealing with a nonrelativistic system, which does not satisfy in particular relativistic local commutativity.

The construction of the scattering operator S completes the picture and concludes our quantum field theory description of the NLS model on $\mathbf{R}_+$.

VII. OUTLOOK AND CONCLUSIONS

We studied the nonlinear Schrödinger equation on the half line with mixed boundary condition. After a brief discussion of some aspects of the corresponding classical boundary value problem, we constructed the exact second quantized solution of the system, establishing its basic properties. The explicit form of our solution shows that the quantum inverse scattering transform works also on the half line, provided that the Zamolodchikov–Faddeev algebra is replaced by the boundary algebra $\mathcal{B}_R$. This is one of the main results of the present paper. It demonstrates that besides being a useful tool in scattering theory,[11] the concept of boundary algebra is essential also for the construction of off-shell interacting fields in integrable systems on $\mathbf{R}_+$. We emphasize in this respect, that our results have a straightforward generalization to all elements of the NLS hierarchy (e.g., the complex modified Korteveg–de Vries equation) on the half line. The case with internal SU(N) symmetry can also be treated analogously.

As for future extensions of the present work, it would be interesting to investigate the range $\eta<0$. The new phenomenon, which can be expected on general grounds, is the presence of boundary bound states. Taking into account that one can describe by $\mathcal{B}_R$ also degrees of freedom residing on the boundary (see the Appendix in Ref. 11), we strongly believe that our framework extends to the case $\eta<0$ as well.

[1] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer, Berlin, 1987).
[2] E. Sklyanin, J. Phys. A **21**, 2375 (1988).
[3] A. S. Fokas, Physica D **35**, 167 (1989).
[4] E. Sklyanin and L. D. Faddeev, Sov. Phys. Dokl. **23**, 902 (1978); E. Sklyanin, *ibid.* **24**, 107 (1979).

[5] H. B. Thacker and D. Wilkinson, Phys. Rev. D **19**, 3660 (1979); D. B. Creamer, H. B. Thacker, and D. Wilkinson, *ibid.* **21**, 1523 (1980).
[6] H. B. Thacker, in *Integrable Quantum Field Theories*, edited by J. Hietarinta and C. Montonen (Springer, Berlin, 1982).
[7] J. Honerkamp, P. Weber, and A. Wiesler, Nucl. Phys. B **152**, 266 (1979).
[8] B. Davies, J. Phys. A **14**, 2631 (1981); Inverse Probl. **4**, 47 (1988).
[9] E. Gutkin, Phys. Rep. **167**, 1 (1988).
[10] A. B. Zamolodchikov and A. B. Zamolodchikov, Ann. Phys. **120**, 253 (1979); L. D. Faddeev, Sov. Sci. Rev. Sec. C **1**, 107 (1980).
[11] A. Liguori, M. Mintchev, and L. Zhao, Commun. Math. Phys. **194**, 569 (1998).
[12] I. V. Cherednik, Theor. Math. Phys. **61**, 977 (1984).
[13] M. Gattobigio, A. Liguori, and M. Mintchev, Phys. Lett. B **428**, 143 (1998).
[14] R. R. Rosales, Stud. Appl. Math. **59**, 117 (1978).
[15] A. Liguori and M. Mintchev, Commun. Math. Phys. **169**, 635 (1995); Lett. Math. Phys. **33**, 283 (1995).
[16] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Addison–Wesley, Reading, 1980).
[17] M. Reed and B. Simon, *Methods of Modern Mathematical Physics III: Scattering Theory* (Accademic, New York, 1979).
[18] A. Liguori, M. Mintchev, and M. Rossi, J. Math. Phys. **38**, 2888 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Finite-band solutions of the classical Boussinesq–Burgers equations

Xianguo Geng
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China and Department of Mathematics, Zhengzhou University, Zhengzhou Henan 450052, People's Republic of China

Yongtang Wu
Department of Computer Science, Hong Kong Baptist University, 224 Waterloo Road, Kowloon, Hong Kong, People's Republic of China

(Received 20 April 1998; accepted for publication 5 February 1999)

The classical Boussinesq–Burgers hierarchy is decomposed into two systems of solvable ordinary differential equations with the help of the Lax representations of stationary evolution equations, from which the finite-band solutions of the higher-order classical Boussinesq–Burgers equations are obtained. © *1999 American Institute of Physics.* [S0022-2488(99)00406-5]

I. INTRODUCTION

It is well known that finite-band solutions of soliton equations are important, which also can be reduced to the multisoliton solutions and elliptic function solutions.[1] Dubrovin and Its-Matveev developed an analog of the inverse scattering theory for Hill's equation and gave an explicit formula of the periodic potentials with the finite number of gaps in the spectrum.[2,3] The periodic N-soliton solution of the KdV equation was constructed explicitly. Date[4] obtained the quasiperiodic solutions of the field equation of classical massive Thirring model via using the Lax equations of squared eigenfunctions (also see Ref. 5). In Ref. 6, Matveev and Yavor gave complex finite-band multiphase solutions of the Kaup–Boussinesq (KB) equation and used the degeneracy procedure to find multisoliton solutions. In Ref. 7, Smirnov continued the investigation of finite-band solutions of the KB equation by finding smooth real solutions and found simple reductions of the general smooth two and three-band solutions to one-dimensional Riemann theta functions. Recently, nonlinearization approach of Lax pairs[8,9] has been developed and applied to construct the finite-band solutions of soliton equations.[10,11] Very recently, Gesztesy and Ratnaseelan have proposed an alternative systematic approach based on elementary algebraic methods to get algebro-geometric solutions of the AKNS hierarchy.[12]

In this paper, we shall develop a direct method to construct finite-band solutions of the higher-order classical Boussinesq–Burgers (CBB) equations based on the Lax pairs of the stationary evolution equations and the ideas in Refs. 4, 5. Although the CBB equation can be respectively changed into the KB equation and the coupled nonlinear Schrödinger equation in the AKNS hierarchy under certain transformations,[13] yet the finite-band solutions of the higher-order KB equations are still open, and it is also nontrivial that finite-band solutions of the higher-order CBB equations are obtained with the help of the AKNS equations. The present paper is organized as follows. In Sec. II, we derive the hierarchy of the CBB equations and the corresponding stationary CBB equations. In Sec. III, solutions of the CBB hierarchy are reduced to solving two systems of solvable ordinary differential equations. In Sec. IV, a hyperelliptic Riemann surface of genus N and Abel–Jacobi coordinates are introduced to straighten the associated flows. The Jacobi's inversion problem is discussed, from which the finite-band solutions of the CBB equation and two high-order CBB equations are expressed explicitly by the Riemann theta functions.

II. THE HIERARCHY AND STATIONARY EVOLUTION EQUATIONS

Let us consider the spectral problem with a constant spectral parameter λ,

$$y_x = Uy, \quad U = \begin{pmatrix} \lambda - v & u + \beta v_x \\ -1 & -\lambda + v \end{pmatrix}, \tag{1}$$

where u and v are two potentials and β is a constant. To drive the CBB hierarchy associated with the spectral problem (1), we first define the Lenard's gradient sequence S_j, $-1 \leqslant j \in \mathbb{Z}$, by the following recursion relation:

$$KS_{j-1} = JS_j, \quad S_j|_{(u,v)=0} = 0, \quad S_{-1} = (1,0)^T, \quad j \geqslant 0, \tag{2}$$

with two skew-symmetric operators

$$K = \begin{pmatrix} \beta(1-\frac{1}{2}\beta)\partial^3 + u\partial + \partial u & \frac{1}{2}(1-\beta)\partial^2 + v\partial \\ -\frac{1}{2}(1-\beta)\partial^2 + \partial v & \frac{1}{2}\partial \end{pmatrix}, \quad J = \begin{pmatrix} 0 & \partial \\ \partial & 0 \end{pmatrix}.$$

It is easy to see that S_j is uniquely determined by the recursion relation (2). The initial condition in Eq. (2) equivalently selects constants of integration to be zero. A direct calculation gives from the recursion relation (2) that

$$S_0 = \begin{pmatrix} v \\ u \end{pmatrix}, \quad S_1 = \begin{pmatrix} \frac{1}{2}u + v^2 + \frac{1}{2}(\beta - 1)v_x \\ \beta(1-\frac{1}{2}\beta)v_{xx} + 2uv + \frac{1}{2}(1-\beta)u_x \end{pmatrix}, \tag{3a}$$

$$S_2 = \begin{pmatrix} \frac{1}{4}v_{xx} + \frac{3}{2}(\beta-1)vv_x + \frac{3}{2}uv + v^3 \\ \frac{1}{4}u_{xx} + \beta(1-\frac{1}{2}\beta)(\frac{3}{2}v_x^2 + 3vv_{xx}) - \frac{3}{2}(\beta-1)vu_x + \frac{3}{4}u^2 + 3uv^2 \end{pmatrix}. \tag{3b}$$

Assume that the time evolution of the eigenfunction y of Eq. (1) obeys the differential equation

$$y_{t_m} = V^{(m)}y, \quad V^{(m)} = \begin{pmatrix} V_{11}^{(m)} & V_{12}^{(m)} \\ V_{21}^{(m)} & -V_{11}^{(m)} \end{pmatrix}, \tag{4}$$

where

$$V_{11}^{(m)} = \sum_{j=0}^{m} \left(\frac{1}{2} S_{j-1,x}^{(1)} - vS_{j-1}^{(1)} - v_x S_{j-1}^{(1)} \right) \lambda^{m-j} + \sum_{j=0}^{m} S_{j-1}^{(1)} \lambda^{m+1-j},$$

$$V_{12}^{(m)} = \sum_{j=0}^{m} \left[\frac{\beta}{2} S_{j-1,xx}^{(1)} + (u + \beta v_x) S_{j-1}^{(1)} + \frac{1}{2} S_{j-1,x}^{(2)} \right] \lambda^{m-j}, \quad V_{21}^{(m)} = -\sum_{j=0}^{m} S_{j-1}^{(1)} \lambda^{m-j}.$$

Then the compatibility condition between Eqs. (1) and (4) yields the zero-curvature equation $U_{t_m} - V_x^{(m)} + [U, V^{(m)}] = 0$, which is equivalent to the hierarchy of nonlinear evolution equations

$$(u_{t_m}, v_{t_m})^T = X_m, \quad m \geqslant 0, \tag{5}$$

where $X_j = KS_{j-1} = JS_j$ is called the CBB vector field. If $m_0 > 1$, then $(u_{t_{m_0}}, v_{t_{m_0}})^T = X_{m_0}$ is called a higher-order CBB equation. The first two nontrivial equations in the hierarchy (5) are the CBB equation[14–16]

$$u_{t_1} = \beta\left(1 - \frac{\beta}{2}\right) v_{xxx} + \frac{1}{2}(1-\beta)u_{xx} + 2(uv)_x, \tag{6a}$$

$$v_{t_1}=\tfrac{1}{2}(\beta-1)v_{xx}+2vv_x+\tfrac{1}{2}u_x \tag{6b}$$

and a higher-order CBB equation,

$$u_{t_2}=\tfrac{1}{4}u_{xxx}+2\beta(1-\tfrac{1}{2}\beta)(3v_xv_{xx}+vv_{xxx})+\tfrac{3}{2}(1-\beta)(v_xu_x+vu_{xx})+\tfrac{3}{2}uu_x+6uvv_x+3u_xv^2, \tag{7a}$$

$$v_{t_2}=\tfrac{1}{4}v_{xxx}+\tfrac{3}{2}(\beta-1)(vv_x)_x+\tfrac{3}{2}(uv)_x+3v^3v_x. \tag{7b}$$

Assume that Eqs. (1) and (4) have two basic solutions, $\psi=(\psi_1,\psi_2)^T$ and $\phi=(\phi_1,\phi_2)^T$, which satisfy the different boundary conditions. We introduce a matrix W of three functions f, g, h by

$$W=\frac{1}{2}(\phi\psi^T+\psi\phi^T)\sigma=\begin{pmatrix} f & g \\ h & -f\end{pmatrix},\quad \sigma=\begin{pmatrix} 0 & -1 \\ 1 & 0\end{pmatrix}. \tag{8}$$

A direct calculation shows that

$$W_x=[U,W],\quad W_{t_m}=[V^{(m)},W], \tag{9}$$

which imply that the function $\det W$ is an independent constant of x and t_m. Equation (9) can be written as

$$\begin{aligned} f_x&=(u+\beta v_x)h+g,\\ g_x&=2(\lambda-v)g-2(u+\beta v_x)f,\\ h_x&=-2f-2(\lambda-v)h, \end{aligned} \tag{10}$$

and

$$\begin{aligned} f_{t_m}&=hV_{12}^{(m)}-gV_{21}^{(m)},\\ g_{t_m}&=2gV_{11}^{(m)}-2fV_{12}^{(m)},\\ h_{t_m}&=2fV_{21}^{(m)}-2hV_{11}^{(m)}. \end{aligned} \tag{11}$$

Now we suppose that the functions f, g, and h are finite-order polynomials in λ,

$$\begin{aligned} f&=\tfrac{1}{2}b_x+(\lambda-v)b,\quad h=-b,\\ g&=\frac{1}{2}c_x+ub+\beta v_xb+\frac{\beta}{2}b_{xx},\\ b&=\sum_{j=0}^{N} b_{j-1}\lambda^{N-j},\quad c=\sum_{j=0}^{N} c_{j-1}\lambda^{N-j}. \end{aligned} \tag{12}$$

Substituting Eq. (12) into Eq. (10) yields

$$KG=\lambda JG,\quad G=(b,c)^T, \tag{13}$$

that is

$$KG_{j-1}=JG_j,\quad JG_{-1}=0, \tag{14a}$$

$$KG_{N-1}=0,\ \ G_j=(b_j,c_j)^T. \tag{14b}$$

It is easy to see that the equation $JG_{-1}=0$ has the general solution,

$$G_{-1}=\alpha_0 S_{-1}+\beta_0 S_{-2},\ \ S_{-2}=(0,1)^T, \tag{15}$$

where α_0 and β_0 are integral constants. Therefore, if we take Eq. (15) as a starting point, then G_j can be recursively determined by the relation (14a). In fact, noticing $\ker J=\{a_0 S_{-1}+a_{-1}S_{-2}|\forall a_0,a_{-1}\}$, $S_{-2}\in\ker K$ and acting with the operator $(J^{-1}K)^k$ upon G_{-1} in Eq. (15), we obtain from Eqs. (14a) and (2) that

$$G_k=\sum_{j=0}^{k+1}\alpha_j S_{k-j}+\beta_{k+1}S_{-2},\ \ -1\leqslant k\leqslant N-1, \tag{16}$$

where $\alpha_0,\ldots,\alpha_{k+1}$ and β_{k+1} are integral constants. Substituting Eq. (16) into Eq. (14b) yields the following higher-order stationary CBB equation:

$$\alpha_0 X_N+\alpha_1 X_{N-1}+\cdots+a_N X_0=0. \tag{17}$$

Thus u and v satisfy certain stationary CBB equation associated with the spectral problem (1). This means that these solutions u and v are finite-band solutions.

III. THE SOLVABLE ORDINARY DIFFERENTIAL EQUATIONS

In this section, we shall show how the CBB hierarchy is decomposed into two systems of solvable ordinary differential equations. Without any loss of generality we can set $\alpha_0=1$ and $\beta_0=0$ since changing b_{-1} simple results in multiplying *f, g,* and *h* by a constant, and the constant c_{-1} vanishing by the differential operator in Eq. (12). From Eq. (16), we have

$$G_{-1}=\begin{pmatrix}1\\0\end{pmatrix},\ \ G_0=\begin{pmatrix}v+\alpha_1\\u+\beta_1\end{pmatrix}, \tag{18a}$$

$$G_1=\begin{pmatrix}\frac{1}{2}(\beta-1)v_x+\frac{1}{2}u+v^2+\alpha_1 v+\alpha_2\\ \beta\left(1-\frac{\beta}{2}\right)v_{xx}+2uv+\frac{1}{2}(1-\beta)u_x+\alpha_1 u+\beta_2\end{pmatrix}. \tag{18b}$$

By using the above expressions and Eq. (12), we get

$$f=\sum_{j=0}^{N+1}f_j\lambda^{N+1-j},\ \ g=\sum_{j=0}^{N}g_j\lambda^{N-j},\ \ h=\sum_{j=0}^{N}h_j\lambda^{N-j}, \tag{19}$$

with

$$\begin{gathered}f_0=1,\ \ f_1=\alpha_1,\ \ f_2=\tfrac{1}{2}(u+\beta v_x)+\alpha_2,\\ g_0=u+\beta v_x,\ \ h_0=-1,\ \ h_1=-v-\alpha_1,\\ h_2=\tfrac{1}{2}(1-\beta)v_x-\tfrac{1}{2}u-v^2-\alpha_1 v-\alpha_2.\end{gathered} \tag{20}$$

This means if we write h as a finite product it takes the form

$$h=-\prod_{i=1}^{N}(\lambda-\mu_i), \tag{21}$$

which implies by comparing the coefficients of the same power for λ that

$$h_1=\sum_{j=1}^{N}\mu_j,\quad h_2=-\sum_{i<j}\mu_i\mu_j=\frac{1}{2}\sum_{j=1}^{N}\mu_j^2-\frac{1}{2}\left(\sum_{j=1}^{N}\mu_j\right)^2,\ldots, \tag{22a}$$

$$h_l=(-1)^{l+1}\sum_{j_1<j_2<\cdots<j_l}\mu_{j_1}\mu_{j_2}\cdots\mu_{j_l},\quad 1\leqslant l\leqslant N. \tag{22b}$$

From Eqs. (20) and (22), we have

$$v=-\sum_{j=1}^{N}\mu_j-\alpha_1, \tag{23a}$$

$$u=\left(\sum_{j=1}^{N}\mu_j\right)^2-\sum_{j=1}^{N}\mu_j^2-2v^2+(1-\beta)v_x-2\alpha_1 v-2\alpha_2. \tag{23b}$$

If we look at the terms of λ^{N+m} and of λ^{N+m-1} for the third expression in Eq. (11), we find also Eq. (23) in view of Eq. (22). Consider the function det W, which is a $(2N+2)$th-order polynomial in λ with constant coefficients

$$-\det W=f^2+gh=\prod_{j=1}^{2N+2}(\lambda-\lambda_j)=R(\lambda). \tag{24}$$

Substituting Eq. (19) into Eq. (24) and comparing the coefficients of $\lambda^{2N+2},\lambda^{2N+1},\ldots,\lambda^{N+2}$ yield

$$2f_0f_1=-\sum_{j=1}^{2N+2}\lambda_j,$$

$$2f_l+\sum_{j=0}^{l-2}f_{j+1}f_{l-j-1}+\sum_{j=0}^{l-2}g_jh_{l-2-j}=(-1)^l\sum_{j_1<j_2<\cdots<j_l}\lambda_{j_1}\lambda_{j_2}\cdots\lambda_{j_l},\quad 2\leqslant l\leqslant N, \tag{25}$$

which shows if g_j,h_j $(j\leqslant l-2)$ and f_j $(j\leqslant l-1)$ are given then f_l can be also obtained from Eq. (25). On the other hand, noticing Eqs. (12) and (19), we have

$$h_j=-b_{j-1},\quad f_j=\tfrac{1}{2}b_{j-2,x}+b_{j-1}-vb_{j-2}, \tag{26a}$$

$$g_j=\frac{1}{2}c_{j-1,x}+(u+\beta v_x)b_{j-1}+\frac{\beta}{2}b_{j-1,xx},\quad 1\leqslant j\leqslant N-2. \tag{26b}$$

By using Eqs. (26) and (16), we obtain the following equalities:

$$f_j|_{(u,v)=0}=\alpha_j,\quad h_j|_{(u,v)=0}=\alpha_j,\quad g_j|_{(u,v)=0}=0,\quad 1\leqslant j\leqslant N, \tag{27}$$

which together with Eq. (25) implies

$$\sum_{j=0}^{l}\alpha_j\alpha_{l-j}=(-1)^l\sum_{j_1<j_2<\cdots<j_l}\lambda_{j_1}\lambda_{j_2}\cdots\lambda_{j_l},\quad(\alpha_0=1),\quad 1\leqslant l\leqslant N. \tag{28}$$

From Eq. (28), it is easy to see that α_j $(1\leqslant j\leqslant N)$ can be explicitly rerepresented by the constants $\lambda_1,\ldots,\lambda_{2N+2}$, for example,

$$\alpha_1=-\frac{1}{2}\sum_{j=1}^{2N+2}\lambda_j,\quad \alpha_2=\frac{1}{2}\sum_{i<j}\lambda_i\lambda_j-\frac{1}{8}\left(\sum_{j=1}^{2N+2}\lambda_j\right)^2, \tag{29a}$$

$$\alpha_3=-\frac{1}{2}\sum_{j_1<j_2<j_3}\lambda_{j_1}\lambda_{j_2}\lambda_{j_3}+\frac{1}{4}\left(\sum_{j=1}^{2N+2}\lambda_j\right)\left[\sum_{i<j}\lambda_i\lambda_j-\frac{1}{4}\left(\sum_{j=1}^{2N+2}\lambda_j\right)^2\right]. \tag{29b}$$

Noticing Eq. (24), we have

$$f|_{\lambda=\mu_k}=\epsilon_k\sqrt{R(\mu_k)}, \tag{30}$$

where ϵ_k is a sheet index that indicates which sheet of the Riemann surface associated with $\sqrt{R(\mu_k)}$ that μ_k lies on. By Eq. (21) and the third expression of Eq. (10), we get

$$h_x|_{\lambda=\mu_k}=-2\,f|_{\lambda=\mu_k}=\mu_{kx}\prod_{i=1,i\neq k}^{N}(\mu_k-\mu_i),\quad 1\leqslant k\leqslant N, \tag{31}$$

which together with Eq. (30) gives

$$\mu_{kx}=-\frac{2\epsilon_k\sqrt{R(\mu_k)}}{\Pi_{i=1,i\neq k}^{N}(\mu_k-\mu_i)},\quad 1\leqslant k\leqslant N. \tag{32}$$

By using Eqs. (16), (19), and (12), we have

$$b_k=-h_{k+1},\quad b_{-1}=1,\quad h_0=-1,\quad S_{-1}^{(1)}=1,$$

$$h_{k+1}=-S_k^{(1)}-\alpha_1S_{k-1}^{(1)}-\alpha_2S_{k-2}^{(1)}-\cdots-\alpha_kS_0^{(1)}-\alpha_{k+1},\quad k\geqslant 0,$$

which implies

$$S_k^{(1)}=-h_{k+1}-\gamma_1h_k-\gamma_2h_{k-1}-\cdots-\gamma_kh_1-\gamma_{k+1}h_0,\quad k\geqslant 0, \tag{33}$$

where

$$\gamma_1=-\alpha_1,\quad \gamma_2=-\alpha_1\gamma_1-\alpha_2,\quad \gamma_3=-\alpha_1\gamma_1-\alpha_2\gamma_2-\alpha_3,\quad \ldots,$$

$$\gamma_k=-\alpha_1\gamma_{k-1}-\alpha_2\gamma_{k-2}-\cdots-\alpha_{k-2}\gamma_2-\alpha_{k-1}\gamma_1-\alpha_k.$$

Therefore, we obtain from Eqs. (4) and (33) that

$$V_{21}^{(m)}|_{\lambda=\mu_k}=\sum_{n=0}^{m}\sum_{s=0}^{n}\gamma_{n-s}h_s\mu_k^{m-n}\quad(\gamma_0=1). \tag{34}$$

In a way similar to the calculation of Eq. (32), we arrive at

$$\mu_{kt_m}=\frac{2\Sigma_{n=0}^{m}\Sigma_{s=0}^{n}\gamma_{n-s}h_s\mu_k^{m-n}\epsilon_k\sqrt{R(\mu_k)}}{\Pi_{i=1,i\neq k}^{N}(\mu_k-\mu_i)}, \tag{35}$$

where $h_0=-1$, and h_s's are determined by Eq. (22). Therefore, if the $2N+2$ distinct parameters $\lambda_1,\ldots,\lambda_{2N+2}$ are given and let $\mu_k(x,t)$ be a solution of ordinary differential Eqs. (32) and (35), then $(u,\,v)$ determined by Eq. (23) is a solution of the CBB Eqs. (5).

IV. THE FINITE-BAND SOLUTIONS

In order to obtain the finite-band solutions of the higher-order CBB equations, we first introduce the Riemann surface Γ of the hyperelliptic curve $\zeta^2=R(\lambda)$. It is easy to see that the genus of this surface is N. On Γ there are two infinite points ∞_1, ∞_2, which are not branch points of Γ. Equip Γ with canonical basis cycles, $a_1,\ldots,a_N$; $b_1,\ldots,b_N$, which are independent and have intersection numbers as follows:

$$a_i \circ a_j = 0, \quad b_i \circ b_j = 0, \quad a_i \circ b_j = \delta_{ij}.$$

For the present, we will choose as our basis the following set

$$\varpi_l = \frac{\lambda^{l-1} d\lambda}{\sqrt{R(\lambda)}}, \quad 1 \leqslant l \leqslant N,$$

which are N linearly independent homomorphic differentials on Γ. By using the cycles a and b, the period matrices A and B can be constructed from

$$A_{ij} = \int_{a_j} \varpi_i, \quad B_{ij} = \int_{b_j} \varpi_i.$$

It is possible to show that the matrices A and B are invertible.[17,18] Now we define the matrices C and τ by $C = A^{-1}$, $\tau = A^{-1}B$. The matrix τ can be shown to be symmetric ($\tau_{ij} = \tau_{ji}$) and it has positive definite imaginary part (Im $\tau > 0$). If we normalize ϖ into the new basis ω_j,

$$\omega_j = \sum_{l=1}^{N} C_{jl} \varpi_l, \quad 1 \leqslant j \leqslant N,$$

then we have

$$\int_{a_i} \omega_j = \sum_{l=1}^{N} C_{jl} \int_{a_i} \varpi_l = \sum_{l=1}^{N} C_{jl} A_{li} = \delta_{ji},$$

and

$$\int_{b_i} \omega_j = \tau_{ji}.$$

Now we introduce Abel–Jacobi coordinates as follows:

$$\rho_j(x,t_m) = \sum_{k=1}^{N} \int_{p_0}^{\mu_k(x,t_m)} \omega_j = \sum_{k=1}^{N} \sum_{l=1}^{N} C_{jl} \int_{p_0}^{\mu_k} \frac{\lambda^{l-1} d\lambda}{\sqrt{R(\lambda)}}, \quad 1 \leqslant j \leqslant N, \tag{36}$$

where p_0 is chosen a base point on Γ. Noticing Eq. (32), we obtain

$$\partial_x \rho_j^{(m)} = \sum_{k=1}^{N} \sum_{l=1}^{N} C_{jl} \frac{\mu_k^{l-1} \mu_{kx}}{\epsilon_k \sqrt{R(\mu_k)}} = \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{-2\mu_k^{l-1} C_{jl}}{\Pi_{i \neq k}(\mu_k - \mu_i)}, \quad \rho_j^{(m)} = \rho_j(x,t_m), \tag{37}$$

which implies

$$\partial_x \rho_j^{(m)} = -2 C_{jN}, \quad 1 \leqslant j \leqslant N, \tag{38}$$

in view of the following equality

$$\sum_{k=1}^{N} \frac{\mu_k^{l-1}}{\Pi_{i \neq k}(\mu_k - \mu_i)} = \delta_{lN}, \quad 1 \leqslant l \leqslant N. \tag{39}$$

In a similar way, we have from Eqs. (35) and (36) that

$$\partial_{t_m}\rho_j^{(m)}=\sum_{l=1}^{N}\sum_{k=1}^{N}\frac{C_{jl}\mu_k^{l-1}\mu_{kt_m}}{\epsilon_k\sqrt{R(\mu_k)}}=2\sum_{l=1}^{N}\sum_{k=1}^{N}\frac{C_{jl}\Sigma_{n=0}^{m}\Sigma_{s=0}^{n}\gamma_{n-s}h_s\mu_k^{m-n+l-1}}{\Pi_{i\neq k}(\mu_k-\mu_i)}. \tag{40}$$

For simplicity, we shall discuss the several special cases as follows:

(I) For $m=1$, Eq. (40) is reduced to

$$\partial_{t_1}\rho_j^{(1)}=2\sum_{l=1}^{N}C_{jl}\sum_{k=1}^{N}\frac{(\Sigma_{s=0}^{N}\mu_s-\gamma_1)\mu_k^{l-1}-\mu_k^{l}}{\Pi_{i\neq k}(\mu_k-\mu_i)},$$

which implies

$$\partial_{t_1}\rho_j^{(1)}=-2C_{j,N-1}-2C_{jN}\gamma_1=\Omega_j^{(1)}, \tag{41}$$

resorting to the equality

$$\sum_{k=1}^{N}\frac{\mu_k^{s}}{\Pi_{i\neq k}(\mu_k-\mu_i)}=\begin{cases}\delta_{s,N-1}, & s\leqslant N-1\\ \displaystyle\sum_{j_1+\cdots+j_N=s-N+1,j_i\geqslant 0}\mu_1^{j_1}\mu_2^{j_2}\cdots\mu_N^{j_N}, & s\geqslant N.\end{cases} \tag{42}$$

Equation (42) can be proved with the help of the following contour integral

$$\frac{1}{2\pi\sqrt{-1}}\int_C H(\lambda)d\lambda,\quad H(\lambda)=\frac{\lambda^{s}}{\Pi_{i=1}^{N}(\lambda-\mu_i)},$$

where the contour C encloses all of the poles μ_j in a counterclockwise manner. It is not difficult to calculate that

$$\begin{aligned}H(\lambda)&=\lambda^{s-N}\left(1-\frac{\mu_1}{\lambda}\right)^{-1}\left(1-\frac{\mu_2}{\lambda}\right)^{-1}\cdots\left(1-\frac{\mu_N}{\lambda}\right)^{-1}\\&=\lambda^{s-N}\sum_{j_1\geqslant 0}\sum_{j_2\geqslant 0}\cdots\sum_{j_N\geqslant 0}\lambda^{-(j_1+j_2+\cdots+j_N)}\mu_1^{j_1}\mu_2^{j_2}\cdots\mu_N^{j_N}\\&=\lambda^{s-N}\sum_{l\geqslant 0}\lambda^{-l}\sum_{j_1+j_2+\cdots+j_N=l}\mu_1^{j_1}\mu_2^{j_2}\cdots\mu_N^{j_N},\end{aligned}$$

which implies

$$-\operatorname*{Res}_{\lambda=\infty}H(\lambda)d\lambda=\begin{cases}\delta_{s,N-1}, & s\leqslant N-1\\ \displaystyle\sum_{j_1+\cdots+j_N=s-N+1,j_i\geqslant 0}\mu_1^{j_1}\mu_2^{j_2}\cdots\mu_N^{j_N}, & s\geqslant N.\end{cases}$$

Therefore, it is easily seen by the residue theorem that Eq. (42) holds.

(II) For $m=2$, Eq. (40) can be written as

$$\partial_{t_2}\rho_j^{(2)}=2\sum_{l=1}^{N}C_{jl}\sum_{k=1}^{N}\frac{-\mu_k^{l+1}+(\Sigma_{s=0}^{N}\mu_s-\gamma_1)\mu_k^{l}-(\Sigma_{i<s}\mu_i\mu_s-\gamma_1\Sigma_{s=0}^{N}\mu_s+\gamma_2)\mu_{k-1}^{l}}{\Pi_{i\neq k}(\mu_k-\mu_i)},$$

which together with Eqs. (42), (22) yields

$$\partial_{t_2}\rho_j^{(2)} = -2C_{j,N-2} - 2C_{j,N-1}\gamma_1 - 2C_{jN}\gamma_2 = \Omega_j^{(2)}. \tag{43}$$

(III) For $m=3$, applying the same approach, we have

$$\partial_{t_3}\rho_j^{(3)} = -2C_{j,N-3} - 2C_{j,N-2}\gamma_1 - 2C_{j,N-1}\gamma_2 - 2C_{jN}\gamma_3 = \Omega_j^{(3)}. \tag{44}$$

On the basis of the above results we get the following:

$$\rho_j^{(m)} = -2C_{jN}x + \Omega_j^{(m)}t_m + \gamma_j, \quad 1 \leqslant j \leqslant N, \quad 1 \leqslant m \leqslant 3, \tag{45}$$

where γ_j's are constants,

$$\gamma_j = \sum_{k=1}^{N} \int_{p_0}^{\mu_k(0,0)} \omega_j.$$

Let $\mathcal{T}$ be the lattice generated by $2N$ vectors $\{\delta_j, \tau_j\}$, where

$$\delta_j = (\underbrace{0,\ldots,0}_{j-1},1,\underbrace{0,\ldots,0}_{N-j})^T$$

and $\tau_j = \tau\delta_j$. The complex torus $\mathcal{J} = \mathbb{C}^N/\mathcal{T}$ is called the Jacobian variety of Γ. An Abel map on Γ is defined as

$$\mathcal{A}(p) = \int_{p_0}^{p} \omega, \qquad \omega = (\omega_1, \ldots, \omega_N)^T,$$

where p_0 is a fixed point on Γ, whose domain of definition can be linearly extended into the factor group $\mathrm{Div}(\Gamma)$,

$$\mathcal{A}\left(\sum n_k p_k\right) = \sum n_k \mathcal{A}(p_k).$$

Now we consider a special divisor $\Sigma_{j=1}^{N} p_k$ and assume that

$$\mathcal{A}\left(\sum_{k=1}^{N} p_k\right) = \sum_{k=1}^{N} \mathcal{A}(p_k) = \sum_{k=1}^{N} \int_{p_0}^{p_k} \omega = \rho^{(m)},$$

with $p_k = (\mu_k(x,t_m), \zeta(\mu_k))$, whose component is

$$\sum_{k=1}^{N} \int_{p_0}^{p_k} \omega_j = \rho_j^{(m)}.$$

Therefore, according to the Riemann theorem,[17,18] there exists a constant vector $M^{(m)} = (M_1^{(m)}, \ldots, M_N^{(m)})^T \in \mathbb{C}^N$ such that the function

$$F^{(m)}(\lambda) = \theta(\mathcal{A}(p) - \rho^{(m)} - M^{(m)})$$

has the only N zero points $p_1, \ldots, p_N$, with $p = (\lambda, \zeta)$. Here $M^{(m)}$ is the Riemann constant determined by Γ itself and θ is the Riemann theta function defined by

$$\theta(\xi|\tau) = \sum_{n \in \mathbb{Z}^N} \exp(\pi\sqrt{-1}\langle \tau n, \; n\rangle + 2\pi\sqrt{-1}\langle \xi, n\rangle),$$

in which $\xi = (\xi_1, \ldots, \xi_N)^T \in \mathbb{C}^N$, $\langle \cdot , \cdot \rangle$ represents the standard inner-product.

To make the function single valued the surface Γ is cut along all a_k, b_k to form a simple connected region, whose boundary is denoted by γ. Notice the fact that the integrals[16,17]

$$\frac{1}{2\pi\sqrt{-1}}\int_\gamma \lambda^k d\ln F^{(m)}(\lambda)=I_k(\Gamma)$$

are constants independent of ρ with

$$I_k(\Gamma)=\sum_{j=1}^N \int_{a_j}\lambda^k\omega_j.$$

By the residue theorem, we have

$$I_k(\Gamma)=\sum_{l=1}^N \mu_l^k+\sum_{s=1}^2 \operatorname*{Res}_{\lambda=\infty_s}\lambda^k d\ln F^{(m)}(\lambda). \tag{46}$$

In what follows, we shall compute the residues in Eq. (46) as $k=1,2$. To this end we first introduce local coordinates $z=\lambda^{-1}$ at infinity. Then the hyperelliptic curve $\zeta^2=R(\lambda)$ in the neighborhood of infinity is expressed as $\hat{\zeta}^2=\hat{R}(z)$ with $\hat{\zeta}=z^{N+1}\zeta$, $\hat{R}(z)=\Pi_{j=1}^{2N+2}(1-\lambda_j z)$, and $\infty_s=(z,(-1)^{s-1}\sqrt{\hat{R}(\lambda)})|_{z=0}=(0,(-1)^{s-1})$, $s=1,2$. It is easy to see that

$$\mathcal{A}(p)_j=-\eta_{sj}+\int_{\infty_s}^p\omega_j=-\eta_{sj}+\sum_{l=1}^N C_{jl}\int_{\infty_s}^p\frac{\lambda^{l-1}d\lambda}{(-1)^{s-1}\sqrt{R(\lambda)}},$$

with

$$\eta_{sj}=\int_{\infty_s}^{p_0}\omega_j,\quad 1\leqslant j\leqslant N,\quad s=1,2.$$

Noting the local coordinates $z=\lambda^{-1}$, we have

$$\mathcal{A}(p)_j=-\eta_{sj}-(-1)^{s-1}\sum_{l=1}^N C_{jl}\int_0^z\frac{z^{N-l}dz}{\sqrt{\hat{R}(z)}}.$$

Since the theta function is an even function, $F^{(m)}(\lambda)$ is written as

$$\begin{aligned}F^{(m)}(z^{-1})&=\theta\left(\dots,\rho_j^{(m)}+M_j^{(m)}+\eta_{sj}+(-1)^{s-1}\sum_{l=1}^N C_{jl}\int_0^z\frac{z^{N-l}dz}{\sqrt{\hat{R}(z)}},\dots\right)\\&=\theta(\dots,\rho_j^{(m)}+M_j^{(m)}+\eta_{sj}+(-1)^{s-1}[C_{jN}z+\tfrac{1}{2}(C_{j(N-1)}-\alpha_1C_{jN})z^2+O(z^3)],\dots),\end{aligned}$$

which together with Eq. (45) leads to

$$\begin{aligned}F^{(m)}(z^{-1})&=\theta_s^{(m)}+z(-1)^{s-1}\sum_{j=1}^N C_{jN}D_j\theta_s^{(m)}\\&\quad+\frac{z^2}{2}\left[(-1)^{s-1}\sum_{j=1}^N(C_{j(N-1)}-\alpha_1C_{jN})D_j\theta_s^{(m)}+\sum_{j=1}^N\sum_{k=1}^N C_{jN}C_{kN}D_jD_k\theta_s^{(m)}\right]+O(z^3),\end{aligned} \tag{47}$$

where $\theta_s^{(m)}=\theta(\rho^{(m)}+M^{(m)}+\eta_s)=\theta(\dots,\rho_j^{(m)}+M_j^{(m)}+\eta_{sj},\dots)$, D_j signifies a derivative with respect to the jth argument of $\theta_s^{(m)}$. It is easy to calculate that

$$\partial_x\theta_s^{(m)}=-2\sum_{j=1}^N C_{jN}D_j\theta_s^{(m)},\quad \partial_x^2\theta_s^{(m)}=4\sum_{k=1}^N\sum_{j=1}^N C_{jN}C_{kN}D_jD_k\theta_s^{(m)},\quad s=1,2.$$

Therefore we have

$$F^{(m)}(z^{-1})=\theta_s^{(m)}+\frac{z}{2}(-1)^s\partial_x\theta_s^{(m)}+\frac{z^2}{2}\left[(-1)^{s-1}\sum_{j=1}^N C_{j(N-1)}D_j\theta_s^{(m)}+\frac{1}{2}(-1)^{s-1}\alpha_1\partial_x\theta_s^{(m)}+\frac{1}{4}\partial_x^2\theta_s^{(m)}\right]+O(z^3), \tag{48}$$

which gives

$$\frac{d}{dz}\ln F^{(m)}(z^{-1})=r_{0s}^{(m)}+r_{1s}^{(m)}+z+O(z^2) \tag{49}$$

with

$$r_{0s}^{(m)}=\tfrac{1}{2}(-1)^s\partial_x\ln\theta_s^{(m)},$$

$$r_{1s}^{(m)}=(-1)^{s-1}\sum_{j=1}^N C_{j(N-1)}D_j\ln\theta_s^{(m)}+\frac{1}{2}(-1)^{s-1}\alpha_1\partial_x\ln\theta_s^{(m)}+\frac{1}{4}\partial_x^2\ln\theta_s^{(m)}.$$

Noticing the equality

$$\operatorname*{Res}_{\lambda=\infty_s}\lambda^k d\ln F^{(m)}(\lambda)=\operatorname*{Res}_{z=0}z^{-k}d\ln F^{(m)}(z^{-1}),$$

we have

$$\operatorname*{Res}_{\lambda=\infty_s}\lambda d\ln F^{(m)}(\lambda)=r_{0s}^{(m)},\quad \operatorname*{Res}_{\lambda=\infty_s}\lambda^2 d\ln F^{(m)}(\lambda)=r_{1s}^{(m)}. \tag{50}$$

By using Eq. (46), we have

$$\sum_{l=1}^N\mu_l(x,t_m)=I_1(\Gamma)+\frac{1}{2}\partial_x\ln\frac{\theta_1^{(m)}}{\theta_2^{(m)}},\quad 1\leqslant m\leqslant 3, \tag{51a}$$

$$\sum_{l=1}^N\mu_l^2(x,t_m)=I_2(\Gamma)+\sum_{j=1}^N C_{j(N-1)}D_j\ln\frac{\theta_2^{(m)}}{\theta_1^{(m)}}+\frac{1}{2}\alpha_1\partial_x\ln\frac{\theta_2^{(m)}}{\theta_1^{(m)}}-\frac{1}{4}\partial_x^2\ln\theta_1^{(m)}\theta_2^{(m)}. \tag{51b}$$

Substituting Eq. (51) into Eq. (23), we obtain

$$u(x,t_m)=\frac{1}{4}\partial_x^2\ln\theta_1^{(m)}\theta_2^{(m)}-\frac{1}{2}(1-\beta)\partial_x^2\ln\frac{\theta_1^{(m)}}{\theta_2^{(m)}}-\frac{1}{4}\left[\partial_x\ln\frac{\theta_1^{(m)}}{\theta_2^{(m)}}\right]^2+\kappa_1\partial_x\ln\frac{\theta_1^{(m)}}{\theta_2^{(m)}}+\sum_{j=1}^N C_{j(N-1)}D_j\ln\frac{\theta_1^{(m)}}{\theta_2^{(m)}}+\kappa_2, \tag{52a}$$

$$v(x,t_m)=-\frac{1}{2}\partial_x \ln\frac{\theta_1^{(m)}}{\theta_2^{(m)}}+\kappa_3, \quad 1\leqslant m\leqslant 3, \tag{52b}$$

with constants

$$\kappa_1=-\tfrac{1}{2}\alpha_1-I_1(\Gamma), \quad \kappa_2=-2\alpha_2-2\alpha_1 I_1(\Gamma)-I_1^2(\Gamma)-I_2(\Gamma), \quad \kappa_3=-\alpha_1-I_1(\Gamma).$$

The expressions (52) are, respectively, the finite-band solutions of the CBB equation and the first two higher-order CBB equations $(u_{t_m},v_{t_m})^T=X_m$, $1\leqslant m\leqslant 3$.

ACKNOWLEDGMENTS

Project 19671074 was supported by National Natural Science Foundation of China. One of the authors (X. G. Geng) would like to thank the Henan Science Foundation Committee of China for financial support. This work was partially supported by Hong Kong RGC/97-98/21.

[1] S. P. Novikov, S. V. Manakov, L. P. Pitaevskii, and V. E. Zakharov, *Theory of Solitons, The Inverse Scattering Methods*, Nauka, Moskva, 1980 (in Russian) [English translation (Plenum, New York, 1984)].
[2] B. A. Dubrovin, Funct. Anal. Appl. **9**, 61 (1975).
[3] A. R. Its and V. B. Matveev, Theor. Math. Phys. **23**, 51 (1975).
[4] E. Date, Prog. Theor. Phys. **59**, 265 (1978).
[5] E. R. Tracy, H. H. Chen, and Y. C. Lee, Phys. Rev. Lett. **53**, 218 (1984).
[6] V. B. Matveev and M. I. Yavor, Ann. Inst. Henri Poincaré, Sect. A **31**, 25 (1979).
[7] A. O. Smirnov, Theor. Math. Phys. **66**, 19 (1986).
[8] C. W. Cao, Sci. China, Ser. A **33**, 528 (1990).
[9] C. W. Cao and X. G. Geng, J. Phys. A **23**, 4117 (1990).
[10] C. W. Cao, ''Integrable structure of the TD-Bargmann system through the generating function approach'' (to be published).
[11] R. G. Zhou, J. Math. Phys. **38**, 2535 (1997).
[12] F. Gesztesy and R. Ratnaseelan, Rev. Math. Phys. **10**, 345 (1998).
[13] R. L. Sachs, Physica D **30**, 1 (1988).
[14] M. Ito, Phys. Lett. **104A**, 248 (1984).
[15] S. Kawamoto, J. Phys. Soc. Jpn. **53**, 469 (1984).
[16] Z. Q. Gu, J. Math. Phys. **31**, 1374 (1990).
[17] C. L. Siegel, *Topics in Complex Function Theory* (Wiley, New York, 1971), Vol. 2.
[18] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1994).

New direct linearizations for KdV and solutions of the other Cauchy problem

Pierre C. Sabatier[a)]
Laboratoire de Physique Mathématique et Théorique—URA CNRS 768, Université Montpellier II—34095 Montpellier Cedex 05, France

(Received 14 December 1998; accepted for publication 24 December 1998)

The author showed previously that there are several equations that ''directly linearize'' the Korteweg de Vries equation. They may give different classes of solutions and they correspond either to generalized Gelfand Levitan or generalized Marchenko inversion equations. The idea is developed here with a precise goal: solving by linear methods the ''other Cauchy problem'' for KdV, i.e., the boundary value problem where the solution is known at fixed x, together with its two first derivatives. After several new direct linearization equations are given and analyzed, the one that solves the problem is eventually obtained. It also corresponds to a new inverse spectral problem, whose scalar equations are fourth order, and that is first studied in the Gelfand Levitan form, and then studied and completely solved in the Marchenko form. The methods given here can be extended probably to most nonlinear integrable equations, and suggest several new problems. [S0022-2488(99)01004-X]

I. INTRODUCTION

In a previous paper,[1] we showed how the direct linearization of the Korteweg de Vries equation can be managed in such a way that the relevant inversion equation is not of the Marchenko form. As a remarkable example, we showed a case where the radial Gelfand Levitan equation is relevant. Our study also showed that various ''direct linearization equations'' (DLE) generate various classes of solutions for the Korteweg de Vries equation, so that it is sound to seek an adapted DLE for solving a given problem. So as to first remind on a simple case how we handle a DLE, we rederive rapidly this example in Sec. II (the derivation here is simpler than the previous one because it is not part of a more general derivation). In the following, we call this example (G).

The classical direct linearization of the Korteweg de Vries equation could also be considered an example of the matricial direct linearization studied in our previous paper, and it leads to the Marchenko inversion equation. We shall call it example (M). Both examples share a very important common property: they rely on one spectral measure only related to $V(x,t)$—a ''Gelfand Levitan spectral measure'' in the case of (G), ''scattering data'' in the case of (M). And so does also the matricial direct linearization we also introduced. These remarks explain our concern in the present paper, which is to derive new DLE, able to generate new solutions of KdV, with the ultimate goal of solving boundary values problem of KdV that were not solved by ''linear analysis'' up to now. Recall, in particular, the ''Cauchy Problems:''

Space Cauchy problem. Derive $V(x,t)$, $(x,t\in\mathbb{R})$ from $[V(x,0), x\in\mathbb{R}]$.

Mixed Cauchy problem. Derive $V(x,t)$ $(x\in\mathbb{R}^+, t\geqslant 0)$ from $[V(x,0), x\in\mathbb{R}^+]$ and $[V(0,t), t\geqslant 0]$.

Time Cauchy problem. Derive $V(x,t)$, $(x,t\in\mathbb{R})$ from $V(0,t)$, $V'(0,t)$, $V''(0,t)$, $t\in\mathbb{R}$.

Of course, it is well known[2] that the space Cauchy problem can be solved by the linearization (M), and also well known that it is the only boundary value problem of the Korteweg de Vries ''linearized'' up to now! What about the others: In particular, what about the time Cauchy

[a)]Electronic mail: SABATIER@LPM.UNIV-MONTP2.FR problem? It really deserves to be called ''the other Cauchy problem,'' since one would never say that the linearized Korteweg de Vries equation is integrable if one did not know how to solve both the space and the time Cauchy problems, and yet, people currently say that Korteweg de Vries is integrable by means of inverse scattering without knowing, up to now, how this ''other'' Cauchy problem can be managed!

Section III is also an introductory section, where we recall the inversion procedure and generalized inversion equations of our previous paper and discuss their relevance for handling a Cauchy problem. Although we feel that cases may be done by using this method, we do not go further because it is very clear that the method is in any case much more complicated and less complete than that using (M) to manage the space Cauchy problem, and we are much less interested by a case study than by the time Cauchy problem. What we do learn, in fact, from this section is that a Gelfand Levitan-type approach leads to a new DLE useful for generating new classes of solutions, but with spectral measures not easy to handle.

In Sec. IV, we wonder if methods showing only one spectral measure are able to deal with the time Cauchy problem. So as to get a feeling of it, we first set down the Lax pair of spectral problems associated to the Korteweg de Vries equation as matrix first-order equations, derive the Born approximations for $V(x,t)$, and thus easily see that in an example where the time Cauchy problem can be analyzed, formulas showing only one spectral measure are not convenient in the linear approximation, and there is no serious reason to be different in exact analyses. Hence, we learn from this analysis that three spectral measures formally present in the DLE is probably necessary in the general case.

After our strategy is defined by the previous sections, we begin the attack to the problem in Sec. V, where we primarily seek general representations of Lax pair solutions based on their analytic properties. The results readily lead us to a generalized Gelfand Levitan formalism and thus to a new direct linearization of KdV, involving for the first time three spectral measures. It can be used to generate classes of solutions solving ''their own'' time Cauchy problem exactly as physicists do when they use the inverse scattering method to derive, for instance multisoliton solutions solving their own ''space Cauchy problem.'' But for solving the time Cauchy problem with initial conditions $V(0,t)V'(0,t)V''(0,t)$ that are arbitrary functions in convenient sets, this approach suffers from the fact that in any Gelfand Levitan formalism, spectral measures are not directly related to (sort of) physical properties, and we shall not use it further.

Thus, in Sec. VI, we definitely rely on the Scattering theory for handling the time Cauchy problem. First we try to guess, by means of extrapolating analogy, what might be a DLE of the Marchenko form with three spectral measures (oddly enough, we obtain one!), and how scattering data may appear on time paths. After these appetizers, we derive a Marchenko-type formalism for the direct and the inverse scattering problems on time paths! It turns out that they are spectral problems with sparse matrices (not many ones in the literature) but that can be related to a class of spectral problems with ''good'' matrices already studied—this meaning that we can associate to the time scattering problem a ''transformed'' time scattering problem, and we must go back and forth from one to the other for solving eventually the inverse scattering problem. Hence, we obtain a generalized Marchenko formalism which relates at time zero the values of $V(0,t)$, $V'(0,t)$, $V''(0,t)$, to conveniently defined ''time path reflection coefficients,'' which behave as spectral measures in the DLE associated to this formalism. This DLE is given in Sec. VII. It enables the evolution of the problem as x varies, and hence the time Cauchy problem is completely solved by this new ''inverse time scattering transform,'' similar, but more complicated than the ordinary one. It turns out that this ''final'' DLE given in Sec. VII is, in fact, a set of coupled integral equations, therefore more complicated that the equation we ''guessed'' at the beginning of Sec. VI. Therefore, for the latter one should work only for a restricted class of initial conditions (but we shall not try to identify them.)

The paper contains at least three new ''direct linearizations'' of the Korteweg de Vries equation, two new Inverse Problems analyses, and the solution by Inverse Scattering of the Korteweg de Vries ''other'' Cauchy problem. Yet, it suggests considerably more work, for instance.

(1) One has to identify the classes of solutions reached by the DLE's. Some of them may be tools for deriving new solutions in the Painlevé analysis.
(2) The method (and the same sort of problems) obviously extend to all nonlinear partial differential equations integrable by Inverse Scattering Transform and which can be directly linearized (for instance, by means of an ansatz we proposed a few years ago).[1–3]
(3) The ''solitonic'' solutions that correspond to discrete spectra in the time scattering problems should be studied in general.
(4) The method yields generalized Gelfand Levitan or Marchenko formalisms in two variables. Can the linear partial differential equations of the generalized kernels be geometrically transformed so as to manage other kinds of boundary value problems? Can these formalisms be used in term as a tool for generating three variables integrable nlpde? Of course, we do not guess that any kind of boundary value problem for an integrable nonlinear partial differential equation can be linearized by similar methods. Direct linearization methods are most often available,[1,3] but the boundary value problems may be very intricate, even for the linearized equation,[4] together with the Lax pairs,[4] and, in most cases, we shall have to deal with quite unusual inverse scattering problems.[5] But I must say that this is for me the most interesting aspect of this subject.

II. A SKETCHY REDERIVATION OF FORMER RESULTS

The starting point was the spectral integral equation,

$$f_1(k,x,t)=\sin[kx+k^3t]+\int \frac{d\mu(\lambda)}{k+\lambda}\sin[(k+\lambda)x+(k^3+\lambda^3)t]f_1(\lambda,x,t). \tag{2.1}$$

We assume throughout that the support of $d\mu(\lambda)$ is contained in $\mathrm{Im}\,\lambda\geqslant 0$, and $d\mu(\lambda)$ is ''symmetric'' with respect to the imaginary axis, i.e., if it is a Stieljes measure $\mu'(\lambda)d\lambda$, $\mu'(\lambda)$ is even. We assume also that the properties of this measure are such that (2.1) is a Fredholm integral equation, for which the Fredholm alternative holds. In the generic case and/or at generic values of x and t, we use the working assumption (WA) that the homogeneous equation,

$$\varphi(k,x,t)=\int T(k,\lambda,x,t)\varphi(\lambda,x,t)d\mu(\lambda), \tag{2.2}$$

where T is the factor of f_1 in (2.1), has no solution but $\varphi=0$. As in our previous paper, we can show that if φ is the odd part of φ as a function of k, the working assumption (WA) implies that only 0 is a solution of

$$\varphi^-(k,x,t)=\int d\mu(\lambda)T^-(k,\lambda,x,t)\varphi^-(\lambda,x,t), \tag{2.3}$$

where

$$T^-(k,\lambda,x,t)=\frac{1}{2}\left\{\frac{\sin[k+\lambda)x+(k^3+\lambda^3)t]}{k+\lambda}-\frac{\sin[(k-\lambda)x+(k^3-\lambda^3)t]}{k-\lambda}\right\} \tag{2.4}$$

and that f_1 is the unique solution of

$$f_1(k,x,t)=\sin[kx+k^3t]+\int d\mu(\lambda)T^-(k,\lambda,x,t)f_1(\lambda,x,t). \tag{2.5}$$

In the method of direct linearization, which is recalled in our previous paper, differential equations for f_1 are obtained by applying differential operators to both sides of (2.1) [or (2.5)]. When the operator applies inside the integral, ''*parasite terms*'' may appear: we call so those that are not similar (modulo the exchange k, λ) to those of the left-hand side of the equation. Parasite terms are ''*reduced*'' if algebraic calculations kill for them the denominator $(k+\lambda)$, so that, thanks to the separability of $\sin[(k+\lambda)x+(k^3+\lambda^3)t]$, the former parasite terms become a combination of ''*free terms*'' (i.e., go outside the integral). The game is to obtain an operator D that produces a zero free term, so that, according to the assumption (WA), Df_1 vanishes. Thus, following exactly our previous derivations, we readily obtain the two equations for f_1:

$$\left[\frac{\partial^2}{\partial x^2}+k^2-V(x,t)\right]f_1(k,x,t)=0 \tag{2.6}$$

and

$$\left[\frac{\partial}{\partial t}-k^2\frac{\partial}{\partial x}-\frac{1}{2}V(x,t)\,\frac{\partial}{\partial x}+\frac{1}{4}V'(x,t)\right]f_1(k,x,t)=0, \tag{2.7}$$

where the ''prime'' is used to mean a differentiation with respect to x, and

$$V(x,t)=-2\,\frac{\partial}{\partial x}\int d\mu(\lambda)\sin(\lambda x+\lambda^3 t)f_1(\lambda,x,t). \tag{2.8}$$

At this point, we go away from the lines of our previous paper, where we derived the Korteweg de Vries equation from the Lax pair (2.6), (2.7) and the usual study of a consistency formula as $k\rightarrow\pm\infty$. Here, let us rather set

$$g(k,x,t,u)=f_1(k,x,t)\sin(kx+k^3u)=:fs. \tag{2.9}$$

From (2.6), written for f and for s, we get

$$g'''=\frac{\partial}{\partial x}(Vg)-4k^2g'+2Vfs'. \tag{2.10}$$

From (2.7), written for f and for s, we get

$$k^2g'=\left(s\,\frac{\partial f}{\partial t}+f\,\frac{\partial s}{\partial u}\right)-\frac{1}{2}Vf'+\frac{1}{4}V'f. \tag{2.11}$$

From (2.10) and (2.11) we readily obtain

$$g'''+4\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial u}\right)g+3Vg'=0. \tag{2.12}$$

In particular, if $t=u$, (2.12) reduces to the simple equation for

$$h(k,x,t)=g(k,x,t,t), \tag{2.13}$$

$$\frac{\partial^3 h}{\partial x^3}+4\,\frac{\partial h}{\partial t}-3V(x,t)\,\frac{\partial h}{\partial x}=0. \tag{2.14}$$

If we multiply both sides of (2.14) by $d\mu(k)$ and integrate, we obtain the following equation for

$$U(x,t)=\int d\mu(\lambda)\sin(\lambda x+\lambda^3 t)f_1(\lambda,x,t), \tag{2.15}$$

$$\frac{\partial^3 U}{\partial x^3}+4\,\frac{\partial U}{\partial t}-3V\,\frac{\partial U}{\partial x}=0. \tag{2.16}$$

But according to (2.8), $V=-2U'$, so that we obtain for U and V the nonlinear evolution equation,

$$\frac{\partial U}{\partial t}+\frac{1}{4}\frac{\partial^3 U}{\partial x^3}+\frac{3}{2}\left(\frac{\partial U}{\partial x}\right)^2=0, \tag{2.17}$$

$$\frac{\partial V}{\partial t}+\frac{1}{4}\frac{\partial^3 V}{\partial x^3}-\frac{3}{2}V\frac{\partial V}{\partial x}=0; \tag{2.18}$$

Eq. (2.18) is nothing but the usual Korteweg de Vries equation.

III. REPRESENTATION OF SOLUTIONS AND INVERSION EQUATION

Again, we first recall results obtained in our previous paper. Writing down (2.4) as a sum of integrals, we obtained a generalized Povzner–Levitan representation of $f_1(k,x,t)$:

$$f_1(k,x,t)=\sin(kx+k^3t)+\int_0^x dy\,\sin(ky+k^3t)L_1(x,y,t,t)+k^2\int_0^t du\,\sin k^3u\,l_1(x,t,u)$$
$$+k\int_0^t du\,\cos k^3u\,l_1'(x,t,u)-\int_0^t du\,\sin k^3u\,l_1''(x,t,u), \tag{3.1}$$

where

$$L_1(x,y,t,u)=-\int d\mu(\lambda)f_1(\lambda,x,t)\sin(\lambda y+\lambda^3u), \tag{3.2}$$

$$l_1(x,t,u)=L_1(x,0,t,u)=-\int d\mu(\lambda)f_1(\lambda,x,t)\sin\lambda^3u, \tag{3.3}$$

$$l_1'(x,t,u)=\left[\frac{\partial}{\partial y}L_1(x,y,t,u)\right]_{y=0}=-\int d\mu(\lambda)f_1(\lambda,x,t)\lambda\,\cos(\lambda^3u), \tag{3.4}$$

$$l_1''(x,t,u)=\left[\frac{\partial^2}{\partial y^2}L_1(x,y,t,u)\right]_{y=0}=\int d\mu(\lambda)f_1(\lambda,x,t)\lambda^2\sin\lambda^3u. \tag{3.5}$$

It follows from (2.8) that

$$V(x,t)=2\frac{\partial}{\partial x}L_1(x,x,t,t)=2\left[\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)L_1(x,y,t,t)\right]_{x=y}. \tag{3.6}$$

Thus,

$$V(0,t)=\left[\frac{\partial}{\partial x}l_1(x,t,t)+l_1'(x,t,t)\right]_{x=0}. \tag{3.7}$$

We derive an integral equation from (3.1), by multiplying by $[d\mu(k)\sin(kx+k^3t)]$ and integrating.

With the notations (3.3)–(3.5), the equation reads as

$$L_1(x,y,t,u)+G_1(x,y,t,u)+\int_0^x G_1(z,y,t,u)L_1(x,z,t,t)dz-\int_0^t g_1''(y,s,u)l_1(x,t,s)ds$$
$$+\int_0^t g_1'(y,s,u)l_1'(x,t,s)ds-\int_0^t g_1(y,s,u)l_1''(x,t,s)ds=0, \tag{3.8}$$

where

$$G(x,y,t,u)=\int d\mu(\lambda)\sin(\lambda x+\lambda^3 t)\sin(\lambda y+\lambda^3 u), \tag{3.9}$$

$$g_1(y,s,u)=G(0,y,s,u)=\int d\mu(\lambda)\sin\lambda^3 s\,\sin(\lambda y+\lambda^3 u), \tag{3.10}$$

$$g_1'(y,s,u)=\left[\frac{\partial}{\partial x}G(x,y,s,u)\right]_{x=0}=\int d\mu(\lambda)\lambda\,\cos\lambda^3 s\,\sin(\lambda y+\lambda^3 u), \tag{3.11}$$

$$g_1''(y,s,u)=\left[\frac{\partial^2}{\partial x^2}G(x,y,s,u)\right]_{x=0}=-\int d\mu(\lambda)\lambda^2\sin\lambda^3 s\,\sin(\lambda y+\lambda^3 u). \tag{3.12}$$

Now this Gelfand Levitan scheme enables us to study the following incomplete Cauchy problem.

We want to construct a solution $V(x,t)$ of the Korteweg de Vries equation (2.18) such that $V(x,0)$ has prescribed values for $x\geqslant 0$. The equations (3.1), (3.2), (3.6), and (3.8) simplify. In particular, if we write $L(x,y)$ for $L(x,y,0,0)$ and $G(x,y)$ for $G(x,y,0,0)$,

$$L(x,y)=-\int d\mu(\lambda)f_1(\lambda,x,0)\sin\lambda y, \tag{3.13}$$

$$G(x,y)=\int d\mu(\lambda)\sin\lambda x\,\sin\lambda y, \tag{3.14}$$

$$L(x,y)+G(x,y)+\int_0^x L(x,z)G(z,y)dz=0. \tag{3.15}$$

These equations are the usual Gelfand Levitan equations, and it follows from (3.2) and (2.6) that $L(x,y)$ obeys the partial differential equations,

$$\left(\frac{\partial^2}{\partial x^2}-V(x)-\frac{\partial^2}{\partial y^2}\right)L(x,y)=0, \tag{3.16}$$

with $L(x,0)=0$ and $L(x,x)$ related to $V(x,0)$ by (3.6). The usual integration on characteristics yields the Volterra equation for $L(x,y)(x\geqslant y\geqslant 0)$ from $V(y,0)$ [which we write here $V(y)$]:

$$L(x,y)=\frac{1}{2}\int_{(x-y)/2}^{(z+y)/2}V(s)ds+\int_{(x-y)/2}^{(x+y)/2}ds\int_0^{(x-y)/2}du\,V(s+u)L(s+u,s-u). \tag{3.17}$$

Hence, giving $V(y)$ for $y\geqslant 0$ enables us to construct $L(x,y)$ for $x\geqslant y\geqslant 0$. Then we can construct $f_1(k,x,0)$ from $\sin kx$ by using the transformation formula (3.1); here

$$f_1(k,x,0)=\sin kx+\int_0^x L(x,y)\sin ky\,dy. \tag{3.18}$$

If $V(x,0)$ is a scattering potential, for which the Gelfand Levitan method applies, we can then construct the Jost function $F(k)$:

$$F(k)=1+\int_0^\infty e^{ikr}V(r,0)k^{-1}f_1(k,r,0)dr, \tag{3.19}$$

and $d\mu(k)$ from the modulus of the Jost function: the part corresponding to the continuous spectrum is proportional to $k(|F(k)|^{-2}-1)$, and it is easy to see that a weak condition of regularity on V [a derivative in $L_1(\mathbb{R}^+)$ is sufficient] guarantees the working assumptions we imposed to $d\mu(\lambda)$. There are also potentials that are not regular, and whose $d\mu(\lambda)$ is such that the method works. We shall not study them. In all cases, giving $d\mu(\lambda)$ gives (2.5), from which the problem is solved. In our previous paper we gave another Gelfand Levitan equation, which corresponds to a different equation (2.5) and also enables us to find (another) solution of KdV from its value on $(0,t)$. The ambiguity in this determination should not be surprising since it is obvious on the linearized problem, and it is related to our selecting from the beginning a symmetric measure.

Hence, it appears that the usual Faddeev–Marchenko scheme is probably better than the Gelfand Levitan one, since it takes the exact amount of information required for solving the space Cauchy problem without any ambiguity. But the scheme we gave here may produce solutions of the Cauchy problems in classes that are not managed by the Marchenko scheme!

IV. MATRICIAL LAX EQUATIONS AND A LINEAR ANALYSIS

The spectral equations (2.6) and (2.7) are fulfilled by a function $f_1(k,x,t)$ defined by (1). It is easy to replace (2.6) by a first-order equation for the vector

$$\mathbf{f}=\begin{pmatrix} f_1(k,x,t) \\ \dfrac{\partial}{\partial x}f_1(k,x,t)\end{pmatrix}. \tag{4.1}$$

In the following, we shall say that the vector **f** written in (3.1) ''corresponds'' to $f_1(k,x,t)$. On the other hand, if we look for solutions of (2.7) that are twice differentiable functions of x and t, we can also derive from (2.7) and (2.6) another first-order equation for **f**. These two first-order equations read as

$$\frac{\partial}{\partial x}\mathbf{f}(k,x,t)=\mathbf{M}\mathbf{f}(k,x,t), \tag{4.2}$$

$$\frac{\partial}{\partial t}\mathbf{f}(k,x,t)=\mathbf{N}\mathbf{f}(k,x,t), \tag{4.3}$$

where

$$\mathbf{M}=\begin{pmatrix} 0 & 1 \\ -k^2+V & 0\end{pmatrix},\qquad \mathbf{N}=\begin{pmatrix} -\frac{1}{4}V' & k^2+\frac{1}{2}V \\ -\frac{1}{4}V''-(k^2-V)(k^2+\frac{1}{2}V) & \frac{1}{4}V'\end{pmatrix} \tag{4.4}$$

(again we used the prime for x derivatives).

They are consistent with each other if

$$\frac{\partial \mathbf{M}}{\partial t}-\frac{\partial \mathbf{N}}{\partial x}+[\mathbf{M},\mathbf{N}]=0, \tag{4.5}$$

which holds *if and only if*

$$\frac{\partial V}{\partial t}+\frac{1}{4}V'''-\frac{3}{2}VV'=0, \tag{4.6}$$

i.e., if the Korteweg de Vries equation holds! If it does, (4.2) and (4.3) can be readily transformed into the Volterra equation,

$$\mathbf{f}(k,x,t)=\mathbf{f}(k,0,0)+\int_0^x \mathbf{M}(y,0)\mathbf{f}(k,y,0)dy+\int_0^t \mathbf{N}(0,u)\mathbf{f}(k,0,u)du+\int_0^x\int_0^t \mathbf{P}(y,u)\mathbf{f}(k,y,u)dy\,du, \tag{4.7}$$

where

$$\mathbf{P}(k,x,t)=\frac{\partial \mathbf{M}}{\partial t}+\mathbf{MN}=\frac{\partial \mathbf{N}}{\partial x}+\mathbf{NM}. \tag{4.8}$$

One can derive approximate solutions of (4.8) that are linear in V (a norm $\|V\|$ may denote the orders of a series expansion). It can be done by setting

$$\mathbf{g}(k,x,t)=\exp[-\mathbf{M}_0(x+k^2t)]\mathbf{f}(k,x,t), \tag{4.9}$$

where **M** is the matrix

$$\begin{pmatrix} 0 & 1 \\ -k^2 & 0 \end{pmatrix},$$

whose eigenvectors

$$\begin{pmatrix} 1 \\ \pm ik \end{pmatrix}$$

correspond to eigenvalues $\pm ik$. Then one notices that

$$\mathbf{M}-\mathbf{M}_0=\mathbf{m}, \quad \mathbf{N}-k^2\mathbf{M}_0=\mathbf{n} \tag{4.10}$$

are matrices with all elements of order $\|V\|$, and one derives the Volterra equation for $\mathbf{g}(k,x,t)$:

$$\mathbf{g}(k,x,t)=\mathbf{g}(k,0,0)+\int_0^x dy\,\exp[-\mathbf{M}_0y]\mathbf{m}(y)\exp[\mathbf{M}_0y]\mathbf{g}(k,y,0)$$
$$+\int_0^t du\,\exp[-k^2\mathbf{M}_0u]\mathbf{n}(u)\exp[k^2\mathbf{M}_0u]\mathbf{g}(k,0,u)+\int_0^x\int_0^t dy\,du\,\mathbf{p}(k,y,u)\mathbf{g}(k,y,u), \tag{4.11}$$

where

$$\mathbf{p}(k,x,t)=\exp[-\mathbf{M}_0(x+k^2t)][\mathbf{q}(k,x,t)]\exp[\mathbf{M}_0(x+k^2t)],$$

$$\mathbf{q}(k,x,t)=\frac{\partial \mathbf{m}}{\partial t}+\mathbf{mN}-k^2\mathbf{M}_0\mathbf{m}=\frac{\partial \mathbf{n}}{\partial x}+\mathbf{nM}-\mathbf{M}_0\mathbf{n}. \tag{4.12}$$

By using (4.12), with the parity assumption on $d\mu(\lambda)$, the reader may easily check that in the V-linear approximation, the function $f_1(k,x,0)$ that reduces at order 0 to $\sin kx$, is indeed an approximate solution of (2.5):

$$f_1(k,x,0) \simeq \sin kx + \frac{1}{2}\int d\mu(\lambda)\left[\frac{\sin(k+\lambda)x}{k+\lambda} - \frac{\sin(k-\lambda)x}{k-\lambda}\right]\sin\lambda x. \tag{4.13}$$

In this approximation, $V(x,0)$ is given by

$$V(x,0) \simeq -2\frac{\partial}{\partial x}\int d\mu(\lambda)\sin^2\lambda x\, dx, \tag{4.14}$$

from which we learn that (a) V is known for any x if it is known for $x \geqslant 0$; (b) $d\mu(\lambda)$ is known from the knowledge of $V(x,0)$ for $x \geqslant 0$, so that the solution $V(x,t)$ of KdV in this linear approximation (and with this method) is known from $V(x,0)$, $x \geqslant 0$.

On the other hand, we can study the V-linear approximation of (4.12) at $x=0$:

$$\mathbf{g}(k,0,t) = \mathbf{g}(k,0,0) + \int_0^t du\, \exp[-k^2\mathbf{M}_0 u]\mathbf{n}(u)\exp[k^2\mathbf{M}_0 u]\mathbf{g}(k,0,0), \tag{4.15}$$

with, say,

$$\mathbf{g}(k,0,0) = \begin{pmatrix} b \\ ka \end{pmatrix} = \alpha\begin{pmatrix} 1 \\ ik \end{pmatrix} + \beta\begin{pmatrix} 1 \\ -ik \end{pmatrix},$$

$$\mathbf{n}(t) = \begin{pmatrix} -\frac{1}{4}V'(0,t) & \frac{1}{2}V(0,t) \\ -\frac{1}{4}V''(0,t) + \frac{1}{2}k^2 V(0,t) & \frac{1}{4}V'(0,t) \end{pmatrix}. \tag{4.16}$$

Setting $\mathbf{g}(k,0,0) = \binom{0}{k}$, we obtain, after some direct calculations on (4.16),

$$f_1(k,0,t) \simeq \sin k^3 t + \int_0^t \left\{ -\frac{1}{4}V''(0,u)\frac{\sin k^3 u}{k}\sin k^3(t-u) \right.$$
$$\left. -\tfrac{1}{4}V'(0,u)\sin(2k^3u - k^3t) + \tfrac{1}{2}V(0,u)\cos(2k^3u - k^3t) \right\} du. \tag{4.17}$$

In the V-linear approximation, V, V', V'' should be themselves the Born approximations:

$$V(x,t) \simeq -2\int d\mu(\lambda)\lambda\, \sin[2(\lambda x + \lambda^3 t)], \tag{4.18}$$

$$V(0,t) \simeq -2\int d\mu(\lambda)\lambda\, \sin[2\lambda^3 t], \tag{4.19}$$

$$V'(0,t) \simeq -4\int d\mu(\lambda)\lambda^2 \cos[2\lambda^3 t], \tag{4.20}$$

$$V''(0,t) \simeq 8\int d\mu(\lambda)\lambda^3 \sin[2\lambda^3 t]. \tag{4.21}$$

If we substitute these results into (4.17), and take into account the symmetry of $d\mu(\lambda)$, we obtain the V-linear approximation of $f_1(k,0,t)$ as it is given by (2.5):

$$f_1(k,0,t) \simeq \sin k^3 t + \frac{1}{2}\int_0^t d\mu(\lambda)\left\{\frac{\sin(k^3+\lambda^3)t}{k+\lambda} - \frac{\sin(k^3-\lambda^3)t}{k-\lambda}\right\}\sin\lambda^3 t. \tag{4.22}$$

Hence, the linearized solution of the Lax-pair equations (4.2), (4.3) corresponding at zero order to $\sin(kx+k^3t)$ and that of (2.5) are identical. Therefore, it is essential to see what can be freely chosen in (4.19)–(4.21). As a first guess, one may hope that V'',V',V, are independent parameters. But since the derivation has been made with the linear conditions, they are given by (4.19), (4.20), and (4.21). This means that, not only V,V',V'' are determined by their values for $t\geqslant 0$ (which is a particular feature of the method), but also they must be consistent altogether with a representation using only one measure, and with at least another one involving three measures (which are here trivially dependent on each other because of the linear equation).

Indeed, let us try to solve ''linearized Korteweg de Vries equation,'' for instance,

$$\frac{\partial V}{\partial t}+\frac{1}{4}V'''=0; \tag{4.23}$$

we see that (4.18) is a solution, but using a full axis x-Fourier transform would enable us, choosing freely $V(x,0)$, whereas using the representation

$$V(x,t)=\int_{-\infty}^{+\infty}d\lambda\; e^{2ik^3t}[\gamma_0(\lambda)e^{2ikx}+\gamma_1(\lambda)e^{2ikjx}+\gamma_2(\lambda)e^{2ikj^2x}], \tag{4.24}$$

where

$$j=\exp[2i\pi/3] \tag{4.25}$$

is convenient to solve the time Cauchy problem.

Hence, methods showing one spectral measure only, and/or assuming that it is symmetric, are not convenient to solve general boundary value problems.

V. REPRESENTATION OF $f(k,x,t)$ ON TIME PATHS

A. Consistency on paths and analytic properties

The known linearizing methods all started from a representation of the wave function [and thus of the function $\mathbf{f}(k,x,t)$ defined by (4.1)] in terms of one spectral measure. Since it seems that it is not enough to deal with the second Cauchy problem, we try now to derive the most general representations of $\mathbf{f}(k,x,t)$, which are of Levitan's form *and* consistent with Eqs. (4.2) and (4.3). First, notice that we can go from $\mathbf{f}(k,x_0,t_0)$ to $\mathbf{f}(k,x,t)$ by following successively an x path and a t path (k being fixed, and dropped for a time from the notations):

$$\mathbf{f}(x,t_0)=\mathbf{f}(x_0,t_0)+\int_{x_0}^{x}dy\;\mathbf{M}(y,t_0)\mathbf{f}(y,t_0), \tag{5.1}$$

$$\mathbf{f}(x,t)=\mathbf{f}(x,t_0)+\int_{0}^{t}du\;\mathbf{N}(x,u)\mathbf{f}(x,u). \tag{5.2}$$

We could as well follow successively a t path and an x path, or, as in (4.7), combine both ones.

The solution $\mathbf{f}(x,t)$ of (5.2) is the solution of

$$\frac{\partial}{\partial t}\mathbf{f}(x,t)=\mathbf{N}(x,t)\mathbf{f}(x,t), \tag{5.3}$$

$\mathbf{f}(\mathbf{x},t_0)$ being given. Now let $\mathbf{S}(x,t)$ be the matrix solution of

$$\mathbf{S}(x,t)=\mathbf{I}+\int_{t_0}^{t}\mathbf{N}(x,u)\mathbf{S}(x,u)du, \tag{5.4}$$

where **I** is the 2×2 unit matrix.

It is clear that $\mathbf{S}(x,t)\mathbf{f}(x,t_0)$ is equal to $\mathbf{f}(x,t)$.

Suppose that the x path (5.1) has produced a function $\mathbf{f}(x,t_0)$ that satisfies the following property:

$$\frac{\partial}{\partial x}\mathbf{f}(x,t_0)=\mathbf{M}(x,t_0)\mathbf{f}(x,t_0), \tag{5.5}$$

and remind that **M** and **N** must fulfill the consistency condition (4.5). It is a matter of simple derivations, and using the solution uniqueness property of Volterra equations (5.2) and (5.4), to show that the function $\mathbf{S}(x,t)\mathbf{f}(x,t_0)$ [or $\mathbf{f}(x,t)$] satisfies for any value of t the equation

$$\frac{\partial}{\partial x}\mathbf{f}(x,t)=\mathbf{M}(x,t)\mathbf{f}(x,t). \tag{5.6}$$

Hence, if we get on the x path a representation of $\mathbf{f}(x,t_0)$ consistent with (5.6), and if we solve (5.4) on the t path, **M** and **N** fulfilling (4.5), we construct $\mathbf{f}(x,t)$, which satisfies simultaneously (5.3) and (5.6). Since we already know Gelfand Levitan or Marchenko representations of $\mathbf{f}(x,t_0)$ on the x path, we have now only to study the t path.

We shall use the convenient notations:

$$\mathbf{N}(k,x,t)=-k^4n_2+k^2n_1+n_0=N_2(k)+N_1(k,x,t), \tag{5.7}$$

where

$$n_2=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix},\quad n_1=\begin{pmatrix}0 & 1\\ V_0 & 0\end{pmatrix},\quad n_0=\begin{pmatrix}V_1 & V_0\\ V_2 & -V_1\end{pmatrix}, \tag{5.8}$$

$$V_0=\tfrac{1}{2}V(x,t),\quad V_1=-\tfrac{1}{4}V'(x,t),\quad V_2=(\tfrac{1}{2}V^2-\tfrac{1}{4}V'')(x,t), \tag{5.9}$$

$$N_2=\begin{pmatrix}0 & k^2\\ -k^4 & 0\end{pmatrix},\quad N_1=k^2V_0n_2+n_0=\mathbf{n}, \tag{5.10}$$

and, for the sake of simplicity, we choose $t_0=0$. Notice that on the path $(0,t)$, x is fixed and only t varies in V_0,V_1,V_2. We suppose that the $V_{i'}$'s are bounded for the value of x and any $t\in[-T,T]$, and let $\overline{V}(x,t)$ be their common absolute upper bound. We suppose that $\int_{-\infty}^{\infty}\overline{V}(x,t)dt$ is finite. We first prove the following.

T1. The solution $\mathbf{S}(k,t)$ of (5.4) at fixed x is an entire function of k^2. If it is written as

$$\mathbf{S}(k,t)=\mathbf{S}_0(k^3,t)+k^2\mathbf{S}_1(k^3,t)+k^4\mathbf{S}_2(k^3,t), \tag{5.11}$$

where

$$\mathbf{S}_l(k^3,t)=\tfrac{1}{3}[\mathbf{S}(k,t)+j^l\mathbf{S}(jk,t)+j^{2l}\mathbf{S}(j^2k,t)], \tag{5.12}$$

then $\mathbf{S}_0,\mathbf{S}_1,\mathbf{S}_2$ are even entire functions of k^3 of exponential type t.

Proof: From (5.4), (5.7), (5.11), and (5.12), we derive the integral equation for the vector $\hat{\sigma}$ (in fact, a 6×2 matrix) whose components are the 2×2 matrices $\mathbf{S}_0,\mathbf{S}_1,\mathbf{S}_2$,

$$\hat{\sigma}(k,t)=\widehat{\sigma_0}+\int_0^t\nu(k,u)\hat{\sigma}(k,u)du, \tag{5.13}$$

where

$$\nu(k,t)=\begin{pmatrix} n_0 & -k^6 n_2 & k^6 n_1 \\ n_1 & n_0 & -k^6 n_2 \\ -n_2 & n_1 & n_0 \end{pmatrix}, \tag{5.14}$$

$$\hat{\sigma}_0=\begin{pmatrix} \mathbf{I} \\ 0 \\ 0 \end{pmatrix}. \tag{5.15}$$

The solution of the Volterra equation (5.13) is given by the uniformly convergent series,

$$\hat{\sigma}_{n+1}(k,t)=\int_0^t \nu(k,u)\hat{\sigma}_n(k,u)du,$$

starting at $\widehat{\sigma_0}$. It follows that $\hat{\sigma}(k,t)$, like $\nu(k,t)$, is an entire function of k^6, and that each element of the 6×2 matrix $\hat{\sigma}(k,t)$ is absolutely bounded by the corresponding one of the matrix $\overline{\sigma_n}(k,t)$ that appears in the series expansion of the solution $\overline{\sigma_n}(k,t)$ of the equation,

$$\overline{\sigma}(k,t)=\widehat{\sigma_0}+\int_0^t \overline{\nu}(|k|,u)\overline{\sigma}(k,u)du, \tag{5.16}$$

where

$$\overline{\nu}|k|,t)=\begin{pmatrix} \overline{n_0} & |k|^6 n_2 & |k|^6\,\overline{n_1} \\ \overline{n_1} & \overline{n_0} & |k^6|n_2 \\ n_2 & \overline{n_1} & \overline{n_0} \end{pmatrix}, \tag{5.17}$$

$$\overline{n_0}=K^2\begin{pmatrix} K & I \\ 2K^2 & K \end{pmatrix}, \quad \overline{n_1}=\begin{pmatrix} 0 & I \\ K^2 & 0 \end{pmatrix}, \tag{5.18}$$

and the positive number K has been chosen to yield absolute bounds for the coefficients containing V and its derivatives in n_0 and n_1. Notice that $\overline{\sigma}(k,t)$ is an entire function of $|k|^6$.

Going back from $\overline{\sigma}(k,t)$ to the 2×2 matrix,

$$\tilde{\sigma}(k,t)=\overline{\sigma}(k,t)+|k|^2\overline{\sigma_1}(k,t)+|k|^4\,\overline{\sigma_2}(k,t), \tag{5.19}$$

we readily derive

$$\tilde{\sigma}(k,t)=I+\int_0^t \tilde{n}(|k|,u)\tilde{\sigma}(k,u)du, \tag{5.20}$$

where

$$\tilde{n}(|k|,u)=\begin{pmatrix} K^3 & \chi^2 \\ \chi^4 & K^3 \end{pmatrix} \tag{5.21}$$

and

$$\chi=\sqrt{|k|^2+K^2}, \tag{5.22}$$

and the solution of (5.20) is

$$\tilde{\sigma}=e^{K^3 t}\begin{pmatrix} \cosh \chi^3 t & \sinh \dfrac{\chi^3 t}{X} \\ \chi \sinh \chi^3 t & \cosh \chi^3 t \end{pmatrix}. \tag{5.23}$$

From (5.23) one readily derives bounds for the components of $\mathbf{S}_0$, $\mathbf{S}_1$, $\mathbf{S}_2$ as $|k|\to\infty$, and from the standard definitions of Boas,[6] it follows that the functions of k^3 that appear are of order not larger than 1 and type t. Q.E.D.

T2. With the assumptions of **T1** and assuming also that $(\partial/\partial t)V(x,t)$ and $(\partial/\partial t)V'(x,t)$ are absolutely bounded for the fixed value of x and any $t\in[-t_1,t_1]$, we claim that $\mathbf{S}_i(k^3,t)$ $(i=0,1,2)$ is absolutely bounded for $k^3\in\mathbb{R}$ and $|t|<t_1$.

Proof: From (5.7) we get

$$\exp[-N_2 t]=\begin{pmatrix} \cos k^3 t & -k^{-1}\sin k^3 t \\ k \sin k^3 t & \cos k^3 t \end{pmatrix}. \tag{5.24}$$

Now set

$$T(k,t)=\begin{pmatrix} 1 & 0 \\ 0 & k^{-1} \end{pmatrix}\exp[-N_2 t]\mathbf{S}(k,t). \tag{5.25}$$

It follows from (5.4) (dropping the fixed x in notations) that

$$T(k,t)=T(k,0)+\sum_{-1}^{1}\int_0^t k^l v^{(l)}(u)n^{(l)}(k^3u)T(k,u)du, \tag{5.26}$$

where

$$v^{(-1)}(t)=-\tfrac{1}{2}V_2(t);v^{(0)}(t)=-\tfrac{1}{4}V'(t);v^{(1)}(t)=\tfrac{1}{2}V(t), \tag{5.27}$$

$$n^{(-1)}(z)=\begin{pmatrix} \sin 2z & 2\sin^2 z \\ -2\cos^2 z & -\sin 2z \end{pmatrix}, \tag{5.28}$$

$$n^{(0)}(z)=\begin{pmatrix} \cos 2z & \sin 2z \\ \sin 2z & -\cos 2z \end{pmatrix}, \tag{5.29}$$

$$n^{(1)}(z)=\begin{pmatrix} -\sin 2z & \cos 2z \\ \cos 2z & \sin 2z \end{pmatrix}. \tag{5.30}$$

It is readily seen that

$$2k^3 n^{(1)}(k^3u)=\frac{\partial}{\partial u}n^{(0)}(k^3u),\quad 2k^3n^{(0)}(k^3u)=-\frac{\partial}{\partial u}n^{(1)}(k^3u). \tag{5.31}$$

Using (5.31) for integrating by parts the term with $l=1$, and noticing that $\partial T/\partial u$ can be expressed by means of (5.26) and that

$$[n^0(z)]^2=I,\quad n^0(z)n^+(z)=\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{5.32}$$

we obtain

$$[1-\tfrac{1}{4}k^{-4}(v^{(1)}(t))^2]T(k,t)=[1+\tfrac{1}{2}k^{-2}n^{(0)}(k^3t)v^{(1)}(t)]$$
$$\times\Bigg\{\bigg[1-\frac{1}{2}k^{-2}v^{(1)}(0)\bigg]T(k,0)+\sum_{-1}^{0}k^l\int_0^l v^{(l)}(u)n^{(l)}(k^3u)T(k,u)du$$
$$-\frac{1}{2}\int_0^t du\Bigg\{-k^{-3}v^{(-1)}(u)n^{(-1)}(k^3u)+k^{-2}$$
$$\times\bigg[v^{(0)}(u)+\frac{\partial}{\partial u}v^{(1)}(u)n^{(0)}(k^3u)\bigg]+k^{-1}v^{(1)}(u)$$
$$\times\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\Bigg\}T(k,u)\Bigg\}. \tag{5.33}$$

This Volterra equation can be studied by means of a majorant equation, all the coefficients being trivially bounded as $|k|$ becomes large, and it follows that the elements of $T(k,t)$, at finite t, are absolutely bounded by those of

$$\overline{T}(k,t)=C(t)\begin{pmatrix}1 & k^{-1}\\ 1 & k^{-1}\end{pmatrix}, \tag{5.34}$$

where C is an appropriate function. We can then improve the asymptotic behavior of $T(k,t)$ as $|k|\to\infty$ on $\mathbb{R}$ by integrating by parts also the term $\int_0^t v^{(0)}(u)n^{(0)}(k^3u)T(k,u)du$, obtaining

$$T(k,t)\begin{pmatrix}1 & 0\\ 0 & k^{-1}\end{pmatrix}=\begin{pmatrix}O(|k|^{-1}) & O(k^{-2})\\ O(|k|^{-1}) & O(k^{-2})\end{pmatrix}, \tag{5.35}$$

from which it follows that

$$\left|\mathbf{S}(k,t)-\begin{pmatrix}\cos k^3t & k^{-1}\sin k^3t\\ -k\sin k^3t & \cos k^3t\end{pmatrix}\right| \tag{5.36}$$

is bounded for fixed t, real k, and so are therefore (Q.E.D.)

$$\mathbf{S}_0(k,t),\quad \mathbf{S}_1(k,t),\quad \mathbf{S}_2(k,t). \tag{5.37}$$

Since $\mathbf{S}(k,t)$ enables us to write down any solution of (5.3) whose value and derivative are fixed at $t=0$, it follows from **T1** and **T2** that we can derive generalized Povzner–Levitan representations of such solutions.

B. Representation of the wave function

As a matter of fact, if φ is the first component of a vector $\mathbf{f}$ solution of

$$\frac{\partial \mathbf{f}}{\partial t}=\mathbf{N}f \tag{5.38}$$

(which, in standard scalar form, would be a fourth-order spectral problem), φ is a solution of

$$D_t(k)\varphi=:\bigg[(k^2+V_0)\bigg(\frac{\partial^2}{\partial t^2}+k^6+\alpha k^2+\beta\bigg)-\frac{\partial V_0}{\partial t}\bigg(\frac{\partial}{\partial t}-V_1\bigg)\bigg]\varphi(k,t)=0, \tag{5.39}$$

where V_0, V_1, V_2, are defined by (5.9), and

$$\alpha=-(V_0^2+V_2),\quad \beta=-\bigg(V_1^2+V_0V_2+\frac{\partial V_1}{\partial t}\bigg). \tag{5.40}$$

We can define two independent solutions of the time spectral problem (5.39) by initial (i.e., $t=0$) conditions and, after some elementary algebra, write down their generalized Povzner–Levitan representations,

$$\varphi_1(k,t)=k^{-1}\sin k^3 t+\int_0^t du\{k^{-1}\sin k^3 u\,\psi_-(t,u)+\cos k^3 u\,\psi_0(t,u)+k\sin k^3 u\,\psi_+(t,u)\},\tag{5.41}$$

$$\varphi_2(k,t)=\cos k^3 t+\chi(t)k^{-1}\sin k^3 t+\int_0^t du\{k^{-1}\sin k^3 u\,\chi_-(t,u)+\cos k^3 u\,\chi_0(t,u)+k\sin k^3 u\,\chi_+(t,u)\}.\tag{5.42}$$

Inserting (5.41) into (5.39), we find boundary conditions to be satisfied by the functions ψ:

$$\psi_-(t,0)=\psi_+(t,0)=\left[\frac{\partial}{\partial u}\psi_0(t,u)\right]_{u=0}=\left[\frac{\partial^2}{\partial t^2}\psi_+(t,u)\right]_{u=0}=\left[\frac{\partial^2}{\partial u^2}\psi_+(t,u)\right]_{u=0}=0,\tag{5.43}$$

$$2\frac{d}{dt}\psi_+(t,t)+\alpha=0,\tag{5.44}$$

$$2\frac{d}{dt}\psi_0(t,t)-\alpha\psi_+(t,t)-\frac{\partial V_0}{\partial t}=0,\tag{5.45}$$

$$2\frac{d}{dt}\psi_-(t,t)+\alpha\psi_0(t,t)-\frac{\partial V_0}{\partial t}\psi_+(t,t)+\beta=0,\tag{5.46}$$

$$\left[\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial u^2}\right)\psi_+(t,u)\right]_{t=u}+\alpha\psi_-(t,t)+\frac{\partial V_0}{\partial t}\psi_0(t,t)+(\beta+\alpha V_0)\psi_+(t,t)-2V_0\frac{d}{dt}\psi_0(t,t)=0,\tag{5.47}$$

$$\left[\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial u^2}\right)\psi_0(t,u)\right]_{t=u}+\alpha\left[\frac{\partial}{\partial u}\psi_+(t,u)\right]_{t=u}+(\alpha V_0+\beta)\psi_0(t,t)-\frac{\partial V_0}{\partial t}\psi_-(t,t)+2V_0\frac{d}{dt}\psi_-(t,t)=0,\tag{5.48}$$

and we also find the partial differential equations to be satisfied by the functions ψ:

$$\left(\frac{\partial^3}{\partial u^3}-\frac{\partial^3}{\partial t^2\,\partial u}\right)\psi_0(t,u)-\alpha\frac{\partial^2}{\partial u^2}\psi_+(t,u)+V_0\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial u^2}\right)\psi_-(t,u)-(\beta+V_0\alpha)\frac{\partial}{\partial u}\psi_0(t,u)-\frac{\partial V_0}{\partial t}\frac{\partial}{\partial t}\psi_-(t,u)+\left(V_0\beta+V_1\frac{\partial V_0}{\partial t}\right)\psi_-(t,u)=0,\tag{5.49}$$

$$\left(\frac{\partial^3}{\partial t^2\,\partial u}-\frac{\partial^3}{\partial u^3}\right)\psi_+(t,u)+V_0\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial u^2}\right)\psi_0(t,u)+(\beta+V_0\alpha)\frac{\partial}{\partial u}\psi_+(t,u)+\alpha\frac{\partial}{\partial u}\psi_-(t,u)-\frac{\partial V_0}{\partial t}\frac{\partial}{\partial t}\psi_0(t,u)+\left(\beta V_0+V_1\frac{\partial V_0}{\partial t}\right)\psi_0(t,u)=0,\tag{5.50}$$

$$\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial u^2}\right)\psi_-(t,u)+V_0\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial u^2}\right)\psi_+(t,u)-\alpha\frac{\partial}{\partial u}\psi_0(t,u)$$
$$-\frac{\partial V_0}{\partial t}\frac{\partial}{\partial t}\psi_+(t,u)+(\beta+\alpha V_0)\psi_-(t,u)+\left(\beta V_0+V_1\frac{\partial V_0}{\partial t}\right)\psi_+(t,u)=0. \tag{5.51}$$

Similar results hold for $\varphi_2(k,t)$. It is clear that all these results remind us of those of the Gelfand Levitan approach to the space spectral problem, and it is likely they can be used to construct solvable examples of Korteveg de Vries equation, but it is also clear that they are much too complicated for doing more. We can simplify their management if we notice that setting

$$\psi_-(t,u)=\int_{-\infty}^{+\infty}d\lambda\,\sin\lambda^3u\lambda\,\tilde{\psi}_{-1}(t,\lambda), \tag{5.52}$$

$$\psi_0(t,u)=\int_{-\infty}^{+\infty}d\lambda\,\cos\lambda^3u\,\tilde{\psi}_0(t,\lambda), \tag{5.53}$$

$$\psi_+(t,u)=\int_{-\infty}^{+\infty}d\lambda\,\sin\lambda^3u\lambda^{-1}\tilde{\psi}_{+1}(t,\lambda), \tag{5.54}$$

where the $\tilde{\psi}_i(t,\lambda)$ are even functions of λ, and setting

$$\Psi_0(t,\lambda)=\tilde{\psi}_{-1}(t,\lambda)+\tilde{\psi}_0(t,\lambda)+\tilde{\psi}_{+1}(t,\lambda), \tag{5.55}$$

$$\Psi_-(t,\lambda)=j^2\tilde{\psi}_{-1}(t,\lambda)+\tilde{\psi}_0(t,\lambda)+j\tilde{\psi}_{+1}(t,\lambda), \tag{5.56}$$

$$\Psi_+(t,\lambda)=j\tilde{\psi}_{-1}(t,\lambda)+\tilde{\psi}_0(t,\lambda)+j^2\tilde{\psi}_{+1}(t,\lambda), \tag{5.57}$$

we readily prove that ψ_-, ψ_0, and ψ_+ are solutions of (5.49), (5.50), (5.51) if and only if Ψ_0, Ψ_-, Ψ_+, are solutions of

$$D_t(\lambda)\Psi_0=0=D_t(j\lambda)\Psi_-=D_t(j^2\lambda)\Psi_+. \tag{5.58}$$

Hence, we can express the Ψ's (and the similar functions obtained from the χ's), by means of linear combinations of $\varphi_1(\lambda,t)$, $\varphi_2(\lambda,t)$ (and their values for $j\lambda$ and $j^2\lambda$). Taking into account the parity, the coefficients of these linear combinations give a way for representing a solution of the time spectral problem (5.39) in terms of three measures. For instance, suppose we stick at $\varphi_1(k,t)$, as it should be done (we guess it) in a Gelfand Levitan approach of the Cauchy problem. Setting Ψ_0, Ψ_-, Ψ_+, respectively, proportional to $\varphi_1(\lambda,t)$, $\varphi_1(j\lambda,t)$, $\varphi_1(j^2\lambda,t)$, we obtain for $\tilde{\Psi}_0$, $\tilde{\Psi}_{-1}$, $\tilde{\Psi}_{+1}$, the expansions

$$\tilde{\psi}_i(t,\lambda)=\sum_{l=-1}^{l=+1}a_l(\lambda)j^{il}\varphi_1(j^{2l}\lambda,t), \tag{5.59}$$

where the a_l's are even functions of λ. On the other hand, inserting (5.52), (5.53), and (5.54) into (5.41), we obtain

$$\varphi_1(k,t)-k^{-1}\sin k^3t=\int_{-\infty}^{+\infty}d\lambda\Bigg\{\left[\tilde{\psi}_{-1}(t,\lambda)+\frac{k}{\lambda}\tilde{\psi}_{+1}(t,\lambda)\right]\int_0^t du\,\sin\lambda^3u\,\sin k^3u$$
$$+\tilde{\psi}_0(t,\lambda)\int_0^t du\,\cos\lambda^3u\,\cos k^3u\Bigg\}. \tag{5.60}$$

Setting now

$$f(k,t)=k\varphi_1(k,t), \tag{5.61}$$

and inserting (5.59) into (5.60) (with this new notation), we get, after elementary algebra,

$$f(k,t)=\sin k^3t+\sum_{l=-1}^{l=+1}\int d\mu_l(\lambda)T^-(k,j^l\lambda,0,t)f(j^l\lambda,t), \tag{5.62}$$

where T^- has been defined by (2.4), and the measures $d\mu(\lambda)$ are symmetric (i.e., they reduce to even functions of λ times $d\lambda$ when they are Stieltjes measures on the real axis). It is possible to derive from (5.62) a representation valid for any x,t by combining the x path and the t path (as we shall do in Sec. VI). It is better to use (5.62) as the cornerstone of a guess and to prove this guess by setting a new direct linearization of the Korteveg de Vries equation.

C. A new direct linearization of KdV

Guess: The following integral equation yields a direct linearization of the Korteweg de Vries equation:

$$f(k,x,t)=\sin(kx+k^3t)+\sum_{l=-1}^{l=+1}\int d\mu_l(\lambda)T^-(k,j^l\lambda,x,t)f(j^l\lambda,x,t), \tag{5.63}$$

where $T^-(k,\lambda,x,t)$ is defined by (2.4).

Proof: (1) As in all direct linearizations, we assume that for a given free term [here $\sin(kx+k^3t)$], (5.63) has only one solution. As a matter of fact, (5.63) can be gathered with the two other equations, respectively, obtained by substituting jk and j^2k to k. The system that is obtained is a Fredholm system, which can be used to derive f. It is clear that the uniqueness property holds for the system only if it does for Eq. (5.63).

(2) We need five formulas, readily derived from (2.4):

$$\frac{\partial}{\partial x}T^-(k,\lambda,x,t)=-\sin(kx+k^3t)\sin(\lambda x+\lambda^3t), \tag{5.64}$$

$$\frac{\partial^2}{\partial x^2}T^-(k,\lambda,x,t)=-k\cos(kx+k^3t)\sin(\lambda x+\lambda^3t)-\lambda\sin(kx+k^3t)\cos(\lambda x+\lambda^3t), \tag{5.65}$$

$$(k^2-\lambda^2)T^-(k,\lambda,x,t)=k\cos(kx+k^3t)\sin(\lambda x+\lambda^3t)-\lambda\sin(kx+k^3t)\cos(\lambda x+\lambda^3t), \tag{5.66}$$

$$\left(\frac{\partial}{\partial t}-k^2\frac{\partial}{\partial x}\right)T^-(k,\lambda,x,t)=-\lambda k\cos(kx+k^3t)\cos(\lambda x+\lambda^3t)-\lambda^2\sin(kx+k^3t)\sin(\lambda x+\lambda^3t), \tag{5.67}$$

$$\left(\frac{\partial^3}{\partial x^3}+k^2\frac{\partial}{\partial x}\right)T^-(k,\lambda,x,t)=-2\lambda k\cos(kx+k^3t)\cos(\lambda x+\lambda^3t)+\lambda^2\sin(kx+k^3t)\sin(\lambda x+\lambda^3t). \tag{5.68}$$

(3) Let us now apply $(\partial^2/\partial x^2+k^2)$ to $f(k,x,t)$ in (5.63). Using (5.64)–(5.66) and their transforms as $\lambda\to j\lambda$, $j^2\lambda$, we see that a function $V(x,t)$ [given below in (5.72)] is such that

$$\left(\frac{\partial^2}{\partial x^2}+k^2\right)f(k,x,t)=V(x,t)\sin(kx+k^3t)+\sum_l\int d\mu_l\,T(k,j^l\lambda,x,t)\left(\frac{\partial^2}{\partial x^2}+j^{2l}\lambda^2\right)f(j\lambda,x,t), \tag{5.69}$$

whereas it follows from (5.63) that for any function V that does not depend on k, we can write down

$$V(x,t)f(k,x,t)=V(x,t)\sin(kx+k^3t)+\sum_l d\mu_l(\lambda)T(k,j^l\lambda,x,t)V(x,t)f(j\lambda,x,t). \quad (5.70)$$

Subtracting (5.70) from (5.69) and thanks to the uniqueness property we get

$$\left[\frac{\partial^2}{\partial x^2}+k^2-V(x,t)\right]f(k,x,t)=0, \quad (5.71)$$

whereas

$$V(x,t)=-2\frac{\partial}{\partial x}\sum_l \int d\mu_l(\lambda)\sin(j^l\lambda x+\lambda^3 t)f(j^l\lambda,x,t). \quad (5.72)$$

(4) Let us apply to f in (5.63) the operator $(\partial/\partial t-\frac{3}{2}k^2(\partial/\partial x)-\frac{1}{2}\partial^3/\partial x^3)$. Using (5.65), (5.67) and their transforms as $\lambda\rightarrow j\lambda$, $j^2\lambda$, we obtain

$$\left(\frac{\partial}{\partial t}-\frac{3}{2}k^2\frac{\partial}{\partial x}-\frac{1}{2}\frac{\partial^3}{\partial x^3}\right)f(k,x,t)=-\frac{3}{4}\left[\frac{\partial}{\partial x}V(x,t)\right]\sin(kx+k^3t)+\sum_l \int_{-\infty}^{+\infty}d\mu_l(\lambda)T(k,j^l\lambda,x,t)$$
$$\times\left(\frac{\partial}{\partial t}-\frac{3}{2}j^{2l}\lambda^2\frac{\partial}{\partial x}-\frac{1}{2}\frac{\partial^3}{\partial x^3}\right)f(j^l\lambda,x,t). \quad (5.73)$$

As in (5.70), it follows that

$$\left[\frac{\partial}{\partial t}-\frac{3}{2}k^2\frac{\partial}{\partial x}-\frac{1}{2}\frac{\partial^3}{\partial x^3}+\frac{3}{4}\frac{\partial}{\partial x}V(x,t)\right]f(k,x,t)=0, \quad (5.74)$$

or, combining this result with the x derivative of (5.71),

$$\left[\frac{\partial}{\partial t}-k^2\frac{\partial}{\partial x}-\frac{1}{2}V(x,t)+\frac{1}{4}V'(x,t)\right]f(k,x,t)=0, \quad (5.75)$$

where we use the "prime" notation for x derivatives.

(5) It is clear that (5.71) and (5.75) are the Lax pair already met in (2.6) and (2.7) for the "potential" $V(x,t)$ given by (5.72), which therefore yields a solution of the Korteweg de Vries equation (2.18). Q.E.D.

The integral equation (5.63) can lead us back to the generalized Povzner Levitan representation (5.41) as its simplified sister did it in our previous paper. It can also lead us to the corresponding, and very complicated, system, generalizing the Gelfand Levitan equation as in Sec. III. There is no need to write down these results, because it is clear that either (5.63) or these equations can be used for constructing solutions of KdV, but (5.63) is a simpler tool.

It is clear also that the solutions that are thus obtained satisfy parity constraints. Using also $\varphi_2(k,t)$ would get rid of this constraint. But the result would still be much more complicated than that obtained by sticking at a Marchenko approach, which we shall do in the next section.

VI. AN INVERSE "SCATTERING METHOD" ON TIME PATHS

A. A new visit to the traditional inverse method

The traditional inverse method is related to the following integral equation of direct linearization:

$$f(k,x,t)=\exp[i(kx+k^3t)]+\int d\rho(\lambda)\tau(k,\lambda,x,t)f(\lambda,x,t), \tag{6.1}$$

where

$$\tau(k,\lambda,x,t)=-i\frac{\exp[i(k+\lambda)x+i(k^3+\lambda^3)t]}{k+\lambda+i0^+}, \tag{6.2}$$

the linearization property follows from the formulas

$$\frac{\partial\tau}{\partial x}=\exp[i(kx+k^3t)+i(\lambda x+\lambda^3t)], \tag{6.3}$$

$$\frac{\partial^2\tau}{\partial x^2}=i(k+\lambda)\exp[i(kx+k^3t)+i(\lambda x+\lambda^3t)], \tag{6.4}$$

$$(k^2-\lambda^2)\tau=i(\lambda-k)\exp[i(kx+k^3t)+i(\lambda x+\lambda^3t)], \tag{6.5}$$

$$\left(\frac{\partial}{\partial t}-k^2\frac{\partial}{\partial x}\right)\tau=(\lambda^2-k\lambda)\exp[i(kx+k^3t)+i(\lambda x+\lambda^3t)], \tag{6.6}$$

$$\left(\frac{\partial^3}{\partial x^3}+k^2\frac{\partial}{\partial x}\right)\tau=-(\lambda^2+2k\lambda)\exp[i(kx+k^3t)+i(\lambda x+\lambda^3t)], \tag{6.7}$$

by proceeding exactly as in Sec. V. It satisfies the two Lax Spectral equations with

$$V(x,t)=2\frac{\partial}{\partial x}\int d\rho(\lambda)\exp[i(\lambda x+\lambda^3t)]f(\lambda,x,t). \tag{6.8}$$

We wish to generalize this approach, as we did in Sec. V. An obvious guess for a new DLE is

$$f(k,x,t)=\exp[i(kx+k^3t)]+\sum_{l=-1}^{l=+1}\int d\rho_l(\lambda)\tau(k,j^l\lambda,x,t)f(j^l\lambda,x,t). \tag{6.9}$$

Again, assuming that the homogeneous form of (6.9) has only the zero solution, and using (6.3)–(6.7) enables us to show that the solution $f(k,x,t)$ of (6.9) satisfies the Lax pair for

$$V(x,t)=2\frac{\partial}{\partial x}\sum_{l=-1}^{l=+1}\int d\rho_l(\lambda)\exp[i(j^l\lambda x+\lambda^3t)]f(j^l\lambda,x,t). \tag{6.10}$$

But two questions arise. First is (6.9), the most general approach using Jost solutions? Next, can the measures $d\rho_l(\lambda)$ be obtained as scattering data derived from $V(0,t)$ and its first two x derivatives, as $d\rho(\lambda)$ of (6.1) is obtained as scattering data derived from $V(x,0)$ in the traditional inverse method?

Some light can be put on these questions if we understand how the measure $d\rho(\lambda)$ of the traditional method is related to the scattering data of $V(x,0)$ by using only the Lax pair. We are reminded that (5.1), and (5.2) give ways, defined from the Lax pair, for going from (x_0,t_0) to (x,t)—say, using the notation F for vector solutions:

$$t-\text{paths: }\frac{\partial F}{\partial t}=\mathbf{N}F, \tag{6.11}$$

$$x-\text{paths: } \frac{\partial F}{\partial x}=\mathbf{M}F. \tag{6.12}$$

In the traditional inverse method we consider an x path, say $t=t_0$, and define the following Jost solutions $\vec{F}_\pm$ of (6.12) by their asymptotic behavior $\vec{J}_\pm$ as $x\to+\infty$, $k\in\mathbb{R}$.

$$\vec{F}_\pm(k,x,t_0) \xrightarrow{x\to\infty} \vec{J}_\pm=\begin{pmatrix} \exp[\pm ikx] \\ \pm ik \exp[\pm ikx]\end{pmatrix}, \tag{6.13}$$

and in the same way, as $x\to-\infty$,

$$\overleftarrow{F}_\mp(k,x,t_0) \xrightarrow{x\to-\infty} \overleftarrow{J}_\mp=\begin{pmatrix} \exp[\mp ikx] \\ \mp ik \exp[\mp ikx]\end{pmatrix}. \tag{6.14}$$

Now, since the trace of $\mathbf{M}$ vanishes, so does the x derivative of the determinant of any matrix made out of two vector solutions. Hence, we can write down

$$\overleftarrow{F}_-(k,x,t_0)=b(k,t_0)\vec{F}_+(k,x,t_0)+a(k,t_0)\vec{F}_-(k,x,t_0), \tag{6.15}$$

where

$$2ika(k,t_0)=\det[\overleftarrow{F}_-,\vec{F}_+], \tag{6.16}$$

$$-2ikb(k,t_0)=\det[\overleftarrow{F}_-,\vec{F}_-] \tag{6.17}$$

define the ''scattering problem.'' As a matter of fact,

$$R_x(k,t_0)=b(k,t_0)/a(k,t_0), \tag{6.18}$$

$$T_x(k,t_0)=[a(k,t_0)]^{-1}, \tag{6.19}$$

are the ''Scattering coefficients'' the index x recalls that it is on an x path. The existence of the Jost solutions of course requires that $V(x,t_0)$ goes to zero rapidly enough as $|x|\to\infty$. Assuming that this assumption holds for any real t, let us now follow a t path with $|x|$ very large, $x=x_0$ so that $V(x_0,t)$ is very small. It is clear that, asymptotically,

$$F(k,x_0,t)\underset{|x|\gg 1}{\widetilde{}} \exp[N_2(t-t_0)]F(k,x_0,t_0). \tag{6.20}$$

Hence, the evolution of the solutions at $x=x_0$ as t varies is

$$\vec{F}_\pm(k,x_0,t_0)\Rightarrow\vec{F}_\pm(k,x_0,t_0)\exp[\pm ik^3(t-t_0)], \tag{6.21}$$

$$\overleftarrow{F}_\mp(k,x_0,t_0)\Rightarrow\overleftarrow{F}_\pm(k,x_0,t_0)\exp[\mp ik^3(t-t_0)]. \tag{6.22}$$

They have no longer the right asymptotic behavior (6.13) and (6.14), and so as to recover the true Jost solutions of the x-scattering problem, say $J\vec{F}_\pm$ and $J\overleftarrow{F}_\pm$, we must set (omitting for simplicity k and x_0 in the notations)

$$J\vec{F}_\pm(t)=\exp[\mp ik^3(t-t_0)]\vec{F}_\pm(t), \tag{6.23}$$

$$J\overleftarrow{F}_\mp(t)=\exp[\pm ik^3(t-t_0)]\overleftarrow{F}_\mp(t). \tag{6.24}$$

On the other hand, since $\mathbf{N}$ is also a matrix of zero trace, the determinant of the matrix that is made of two vector solutions is an invariant. Hence, we can write from (6.23) and (6.24), and from the invariance property,

$$\det[J\overleftarrow{F}_-(t),J\vec{F}_+(t)]=\det[\overleftarrow{F}_-(t),\vec{F}_+(t)]=\det[\overleftarrow{F}_-(t_0),\vec{F}_+(t_0)]. \tag{6.25}$$

Hence, $a(k,t)$ is invariant, and so is $T(k)$. On the other hand, the same formulas yield

$$\det[J\overleftarrow{F}_-(t),J\vec{F}_-(t)]=e^{2ik^3(t-t_0)}\det[\overleftarrow{F}_-(t_0),\vec{F}_-(t_0)], \tag{6.26}$$

and so $R_x(k,t)/R_x(k,t_0)$ is equal to the factor $\exp[2ik^3(t-t_0)]$. Inserting it in the Marchencko equation yields the traditional inverse method and (6.1).

Now suppose convenient assumptions on V guarantee on any t path the existence of time Jost solutions $\vec{G}^{\pm}$ and $\overleftarrow{G}^{\mp}$ (real k):

$$\vec{G}^{\pm}(k,x_0,t)\xrightarrow{t\to\infty}e^{\pm ik^3t}\begin{pmatrix}\mp i\\ k\end{pmatrix}, \tag{6.27}$$

$$\overleftarrow{G}^{\mp}(k,x_0,t)\xrightarrow{t\to-\infty}e^{\mp ik^3t}\begin{pmatrix}\pm i\\ k\end{pmatrix}, \tag{6.28}$$

they define a scattering problem where $\det[\overleftarrow{F}_-,\vec{F}_+]$ and $\det[\overleftarrow{F}_-,\vec{F}_-]$ are ''time scattering'' coefficients. Again, for very large t, $V(x,t)$ ''small'' implies

$$G(k,x,t)\sim\exp[\mathbf{M}_0(x-x_0)]G(k,x_0,t), \tag{6.29}$$

and coefficients $\exp[\pm ik(x-x_0)]$ appear in the same manner $\exp[\pm ik^3(t-t_0)]$ did it in the analysis above. Hence, the x evolution of the measures in (6.9) can be understood on the same grounds as was the t evolution of $d\rho(\lambda)$ in (6.1). Thus, the way to check (6.9) or another DLE and to relate it to Cauchy data is by solving the inverse problem from ''time scattering'' coefficients.

B. The direct problem on time paths, put to second order

For a fixed value of x, say x_0, we study the equation (6.11), where $\mathbf{N}$ is given by (5.7)–(5.10) as N_2+N_1 (and we drop x_0 in the notations). We assume that V and its x and t derivatives go to zero as $t\to\pm\infty$ ''rapidly enough'' to warrant that the ''leading'' equation at infinite times is

$$\frac{\partial F}{\partial t}\simeq N_2F. \tag{6.30}$$

It is easy to check that the asymptotic behaviors $\vec{G}^{\infty}_{\pm}$ and $\overleftarrow{G}^{\infty}_{\mp}$ that appear as limits in (6.27) and (6.28) are indeed solutions of (6.30).

In order to study the existence and properties of Jost solutions, it is convenient to first consider the second-order problem,

$$\left(\frac{\partial}{\partial t}+\mathbf{N}\right)\left(\frac{\partial}{\partial t}-\mathbf{N}\right)F=0=\left[\frac{\partial^2}{\partial t^2}-\left(\frac{\partial\mathbf{N}}{\partial t}+\mathbf{N}^2\right)\right]F. \tag{6.31}$$

Clearly, all solutions of (6.31) are not necessarily solutions of (6.11). We shall see below in Sec. VI C how the latter are identified, and we give first a study of the second-order problem, writing (6.31) as

$$\frac{\partial^2F}{\partial t^2}+k^6F-k^2(B_1I+B_3n_2)F-AF, \tag{6.32}$$

where

$$B_1=\frac{3}{4}V^2-\frac{1}{4}V'',\quad B_3=\frac{1}{2}\frac{\partial V}{\partial t}, \tag{6.33}$$

$$A = n_0^2 + \frac{\partial n_0}{\partial t} = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}, \tag{6.34}$$

$$B = B_1 I + B_3 n_2 = \begin{pmatrix} B_1 & 0 \\ B_3 & B_1 \end{pmatrix}. \tag{6.35}$$

Except that F is a two-vector (or a 2×2 matrix), (6.32) is similar to equations studied by Jaulent and Jean,[7] and where the inverse problem can be managed by a method analogous to one introduced by the author[8] more than 30 years ago. So as to handle it, we first gather (6.32) and its "j transformations," ($j = e^{2i\pi/3}$):

$$\frac{\partial^2 F_l}{\partial t^2} + (k^6 - j^{2l} k^2 B - A) F_l = 0, \tag{6.36}$$

and define the three "components" of a vector $\mathbf{\Phi}$:

$$\Phi_0 = \frac{1}{3} \sum_l F_l, \tag{6.37}$$

$$\Phi_1 = \frac{1}{3} k^{-1} \sum_l j^{2l} F_l, \tag{6.38}$$

$$\Phi_2 = \frac{1}{3} k^{-2} \sum_l j^l F_l. \tag{6.39}$$

If the F_l's are 2-vectors, Φ is a 6-vector. It is a solution of the equation

$$\frac{\partial^2 \mathbf{\Phi}}{\partial t^2} + k^6 \mathbf{\Phi} - k^3 \begin{pmatrix} 0 & B & 0 \\ 0 & 0 & B \\ 0 & 0 & 0 \end{pmatrix} \mathbf{\Phi} - \begin{pmatrix} A & 0 & 0 \\ 0 & A & 0 \\ B & 0 & A \end{pmatrix} \mathbf{\Phi} = 0. \tag{6.40}$$

The notation that is used for $\mathbf{\Phi}$ in (6.40) is that of a 3-vector whose elements are 2-vectors, so that it is, in fact, a 6-vector, and of 3×3 matrices, say, $\mathcal{B}$ and $\mathcal{A}$ whose elements are 2×2 matrices so that they are, in fact, 6×6 matrices. If we wish to compare a with Jaulent Jean presentation, we would write $\mathbf{\Phi}_+$ instead of $\mathbf{\Phi}$ and introduce $\mathbf{\Phi}_-$ as a solution of the equation obtained by making $k \to -k$ in (6.40). This is avoided here by defining four Jost solutions, as we did in (6.27), (6.28).

Now, in order to study analytic properties, we introduce the "Jost matrices" of (6.40) as the four 6×6 matrices $\mathcal{M}$, which are solutions of (6.40) and have the asymptotic behaviors ($k^3 \in \mathbb{R}$):

$$\vec{\mathcal{M}}^{\pm} \to \exp[\pm i k^3 t] I (t \to \infty), \tag{6.41}$$

$$\overleftarrow{\mathcal{M}} \to \exp[\mp i k^3 t] I (t \to -\infty), \tag{6.42}$$

where I is the unit matrix. We first study the existence and properties of them, and after that we use them to construct the functions (6.27), (6.28).

From (6.40) and (6.41), we get

$$\vec{\mathcal{M}}^{\pm}(k^3,t) = I e^{\pm i k^3 t} - \int_t^{\infty} \frac{\sin k^3 (t-u)}{k^3} [k^3 \mathcal{B}(u) + \mathcal{A}(u)] \vec{\mathcal{M}}^{\pm}(k^3,u) du, \tag{6.43}$$

or setting

$$\vec{m}^{\pm}(k^3,t) = e^{\mp i k^3 t} \vec{\mathcal{M}}^{\pm}(k^3,t), \tag{6.44}$$

$$\vec{m}^{\pm}(k^3,t)=I-\int_t^{\infty}du\,\frac{e^{\mp ik^3(t-u)}\sin k^3(t-u)}{k^3}[k^3\mathcal{B}(u)+\mathcal{A}(u)]\vec{m}^{\pm}(k^3,u). \tag{6.45}$$

For any finite k, (6.45) is a Volterra integral equation whose kernel can be absolutely bounded, respectively, for $\pm\operatorname{Im}k\geqslant 0$, by $C|k|^3$ (C is a positive number), and by $(u-t)(t<u)$ which can be also replaced in these bounds by $I(t)J(u)$, where

$$I(t)=J(-t)=1+|x|\,\theta(-t), \tag{6.46}$$

as we did it for a quite similar opportunity, p. 326, Ref. 5. The result is that $|\vec{m}^{\pm}|$ can be constructed by the usual series expansion of the solution of (6.45) and that it is uniformly bounded on the real k axis, provided

$$\int_{-\infty}^{+\infty}t|B_i(t)|dt<\infty,\quad \int_{-\infty}^{+\infty}t|A_i(t)|dt<\infty,\quad i=1,2,3. \tag{6.47}$$

Going to the k^3 complex plane, a similar approach with the same assumptions shows that $\vec{\mathcal{M}}^{+}(k^3,t)$ can be extended into a holomorphic function in $\operatorname{Im}k^3>0$, with continuity to the real axis, and that $\vec{\mathcal{M}}^{-}$ has the same property in $\operatorname{Im}k^3\leqslant 0$. By the same token, we extend these results to $\overleftarrow{\mathcal{M}}^{-}$ in $\operatorname{Im}k^3\geqslant 0$ and $\overleftarrow{\mathcal{M}}^{+}$ in $\operatorname{Im}k^3\leqslant 0$. It is convenient to set hereafter $\lambda=k^3$, and to look more carefully to the behavior at large $|\lambda|$. This is done by seeking the solution of (6.43) from the ansatz

$$\vec{\mathcal{M}}^{\pm}(\lambda,t)=e^{\pm i\lambda t}\vec{P}^{\pm}(t)+\lambda^{-1}\vec{Q}^{\pm}(\lambda,t), \tag{6.48}$$

where

$$\vec{P}^{\pm}(t)=\exp\left[\mp\frac{1}{2i}\int_t^{\infty}\mathcal{B}(u)du\right], \tag{6.49}$$

has been chosen in order to match the leading asymptotic behavior in (6.43). With this ansatz and if the constraints (6.47) extend to the time derivatives of $\mathcal{B}$ elements, we can show that $|\lambda\vec{Q}^{\pm}(\lambda,t)|$ is uniformly bounded, respectively, in $\{\pm\operatorname{Im}\lambda\geqslant 0\}$. Let us sketch the proof for $\vec{Q}^{+}(\lambda,t)$. From (6.43) and (6.48) we obtain

$$\begin{aligned}\vec{Q}^{+}(\lambda,t)=&-\int_t^{\infty}\sin\lambda(t-u)\mathcal{A}(u)[e^{i\lambda u}\vec{P}^{+}(u)+\lambda^{-1}\vec{Q}^{+}(\lambda,u)]du\\&+(2i)^{-1}e^{-i\lambda t}\int_t^{\infty}(\lambda\mathcal{B}(u)+\mathcal{A}(u))[e^{2i\lambda u}\vec{P}^{+}(u)+\lambda^{-1}e^{i\lambda u}\vec{Q}^{+}(\lambda,u)]du.\end{aligned} \tag{6.50}$$

An integration by parts readily yields

$$\int_t^{\infty}e^{2i\lambda u}\mathcal{B}(u)\vec{P}^{+}(u)du=-\frac{e^{2i\lambda t}}{2i\lambda}\mathcal{B}(t)\vec{P}^{+}(t)-\int_t^{\infty}du\,\frac{e^{2i\lambda u}}{2\lambda}(\mathcal{B}'(u)+\mathcal{B}^2(u)). \tag{6.51}$$

Gathering the terms that contain Q in (6.50), and evaluating those where an integration by parts inserts $o(\lambda^{-1})$, we obtain a Volterra integral equation whose free term and kernel are absolutely bounded, except maybe at $\lambda=0$. The boundedness of Q follows for $\lambda\neq 0$, and at $\lambda=0$ it follows from our previous results.

If the bounds (6.47) also extend to second time derivatives, we can go further by the same method and show that

$$\vec{\mathcal{M}}^{\pm}(\lambda,t)=e^{\pm i\lambda t}[\vec{P}^{\pm}(t)+\lambda^{-1}\vec{Q}^{\pm}(t)+\lambda^{-2}\vec{R}^{\pm}(\lambda,t)], \tag{6.52}$$

where $\vec{P}^{\pm}(t)$ is given by (6.49), $\vec{Q}^{\pm}(t)$ is the (exactly calculable) solution of the equation:

$$\vec{Q}^{\pm}(t)=\frac{1}{4}\mathcal{B}(t)\vec{P}^{\pm}(t)\mp\frac{1}{2i}\int_t^{\infty}du[\mathcal{A}(u)\vec{P}^{\pm}(u)+\mathcal{B}(u)\vec{Q}^{\pm}(u)], \tag{6.53}$$

and $\vec{R}^{\pm}$ has its elements absolutely bounded in the adequate λ half-plane.

By the same token, starting from the equation,

$$\overleftarrow{\mathcal{M}}^{\mp}(\lambda,t)=e^{\mp i\lambda t}+\int_{-\infty}^{t}du\,\frac{\sin\lambda(t-u)}{\lambda}[\lambda\mathcal{B}(u)+\mathcal{A}(u)]\overleftarrow{\mathcal{M}}^{\mp}(\lambda,u), \tag{6.54}$$

we can show that if (6.47) extends to second-order derivatives, and if, respectively, $\pm\operatorname{Im}\lambda\geqslant 0$,

$$\overleftarrow{\mathcal{M}}^{\mp}(\lambda,t)=e^{\mp i\lambda t}[\overleftarrow{P}^{\mp}(t)+\lambda^{-1}\overleftarrow{Q}^{\mp}(t)+\lambda^{-2}\overleftarrow{R}^{\mp}(\lambda,t)], \tag{6.55}$$

where

$$\overleftarrow{P}^{\mp}(t)=\exp\left[\mp\frac{1}{2i}\int_{-\infty}^{t}\mathcal{B}(u)du\right], \tag{6.56}$$

and $\overleftarrow{Q}^{\mp}(t)$ is the (exactly calculable) solution of the equation,

$$\overleftarrow{Q}^{\mp}(t)=\frac{1}{4}\mathcal{B}(t)\overleftarrow{P}^{\mp}(t)\mp\frac{1}{2i}\int_{-\infty}^{t}du[\mathcal{A}(u)\overleftarrow{P}^{\mp}(u)+\mathcal{B}(u)\overleftarrow{Q}^{\mp}(u)]. \tag{6.57}$$

If (6.47) extends only to first-order derivatives, the first term only in (6.55) is guaranteed and the remainder is of order $O(|\lambda|^{-1})$ (in the adequate half-plane, with continuity to $\mathbb{R}$). This, however, implies that it is in $L^2(\mathbb{R})$, so that we can construct its Fourier transform, obtaining in this way the transformation formulas for the Jost matrices:

$$\vec{\mathcal{M}}^{\pm}(\lambda,t)=\vec{P}^{\pm}(t)\exp[\pm i\lambda t]+\int_t^{\infty}\vec{K}^{\pm}(t,u)\exp(\pm i\lambda u)du, \tag{6.58}$$

$$\overleftarrow{\mathcal{M}}^{\mp}(\lambda,t)=\overleftarrow{P}^{\mp}(t)\exp(\mp i\lambda t)+\int_{-\infty}^{t}\overleftarrow{K}^{\mp}(t,u)\exp(\mp i\lambda u)du. \tag{6.59}$$

C. Back to the first-order time equations

As we see below, for any fixed 6-vector $\mathbf{v}^{\pm}$, a matrix solution $\vec{\mathcal{M}}^{\pm}(t)$ of (6.40), yields a 6-vector solution $\vec{\mathcal{M}}^{\pm}\mathbf{v}^{\mp}$, denoted below as $\vec{\boldsymbol{\Psi}}^{\pm}(t)$, which is asymptotic to $\mathbf{v}^{\mp}\exp[\pm ik^3t]$ as $t\to\infty$. In the same way as F was related to $\boldsymbol{\Phi}$, we can derive from $\boldsymbol{\Psi}$ a 2-vector solution $\vec{H}^{\pm}(t)$ of (6.31), which is asymptotic to $\mathbf{w}^{\mp}\exp[\pm ik^3t]$, and (6.31) can be written as

$$\left(\frac{\partial}{\partial t}+\mathbf{N}\right)\left(\frac{\partial}{\partial t}-\mathbf{N}\right)\vec{H}^{\pm}(t)=0, \tag{6.60}$$

which implies that $(\partial/\partial t-\mathbf{N})\vec{H}^{\pm}(t)$ is a solution $y(t)$ of the equation

$$\left(\frac{\partial}{\partial t}+\mathbf{N}\right)y(t)=0. \tag{6.61}$$

So as to get a solution of (6.11), we want $y(t)=0$. It cannot be so unless its asymptotic behavior $y_{\infty}(t)$ is so, and y_{∞} is not zero unless $\mathbf{w}^{\mp}e^{\pm ik^3t}$ is a solution of

$$\left(\frac{\partial}{\partial t}-N_2\right)(\mathbf{w}^{\mp}\exp[\pm ik^3t])=0, \tag{6.62}$$

where N_2 [see (5.10)] is the leading term of $\mathbf{N}$ as $t\to\infty$. It follows after some algebra that the Jost solutions $\vec{G}^{\pm}(t)$ of (6.11) are derived from $\vec{\mathcal{M}}^{\pm}\mathbf{v}$ if (6.62) holds, i.e., if

$$\mathbf{w}^{\mp}=\begin{pmatrix}\mp i\\ k\end{pmatrix}, \tag{6.63}$$

which, indeed, defines $\vec{G}^{\pm}(t)$ in (6.27). A similar analysis holds as $t\to-\infty$, where (6.28) properly defines $\overleftarrow{G}^{\mp}(t)$, and these results pave the way going from second-order to first-order results.

We also wish appraising asymptotic behaviors as $|k|\to\infty$, fixed t, sufficiently large that V and derivatives are small compared to convenient powers of $|k|$. For this purpose, we found it useful to have exact reference solutions of a first-order problem as (6.11), whose matrix $\tilde{N}$ goes ''close'' to $\mathbf{N}$ as either t or $|k|$ goes to ∞. Here they are

$$\vec{E}^{\pm}=\omega^{\mp}\exp[\pm i\vec{\varphi}];\quad \overleftarrow{E}^{\mp}=\omega^{\pm}\exp[\mp i\overleftarrow{\varphi}], \tag{6.64}$$

where

$$\omega^{\pm}=\begin{pmatrix}\pm i(1+V_0/2k^2)\\ k(1-V_0/2k^2)\end{pmatrix}, \tag{6.65}$$

$$\vec{\varphi}=k^3t+(2k)^{-1}\int_t^{\infty}V_0^2(u)\,du, \tag{6.66}$$

$$\overleftarrow{\varphi}=\vec{\varphi}-(2k)^{-1}\int_{-\infty}^{+\infty}V_0^2(u)\,du=:\vec{\varphi}-\omega_0/k. \tag{6.67}$$

The time derivative of any solution E is $\tilde{N}E$, with

$$\tilde{N}=\begin{pmatrix}\dfrac{-2V_1}{2k^2+V_0} & \dfrac{k^2\left(1-\frac{1}{2}k^{-4}V_0^2\right)\left(1+\frac{1}{2}k^{-2}V_0\right)}{1-\frac{1}{2}k^{-2}V_0}\\[2ex] \dfrac{-k^4\left(1-\frac{1}{2}k^{-4}V_0^2\right)\left(1-\frac{1}{2}k^{-2}V_0\right)}{1+\frac{1}{2}k^{-2}V_0} & \dfrac{2V_1}{2k^2-V_0}\end{pmatrix}, \tag{6.68}$$

to be compared with $\mathbf{N}$, written as

$$\mathbf{N}=\begin{pmatrix}0 & k^2+V_0\\ -k^2(k^2-V_0) & 0\end{pmatrix}+\begin{pmatrix}V_1 & 0\\ V_2 & -V_1\end{pmatrix}. \tag{6.69}$$

One easily sees that if t is sufficiently large that $|V_0|$ be smaller than $2|k|^2$, $\mathbf{N}-\tilde{N}$ is bounded uniformly in $|k|$ by an integrable function of t, and goes to the k-independent matrix in (6.69), called U in the following, as rapidly as $|k|^{-2}$ times an integrable function of t, but the remainder is not a zero trace matrix (its trace goes to zero as $|k|^{-4}$).

We can now proceed to the first-order scattering problem. The Jost solutions G that correspond to the first-order time equation (6.11) were defined by (6.27) and (6.28). The time scattering coefficients (which are numbers) are defined from them by

$$\overleftarrow{G}^{-}(k,t)=c^{+}(k)\vec{G}^{-}(k,t)+d^{+}(k)\vec{G}^{+}(k,t), \tag{6.70}$$

$$\overleftarrow{G}^{+}(k,t)=c^{-}(k)\vec{G}^{+}(k,t)+d^{-}(k)\vec{G}^{-}(k,t). \tag{6.71}$$

By means of two ''j transformations'' we define the ''Jost vectors'' $\mathbf{G}$:

$$\mathbf{G}=\begin{pmatrix} G(k) \\ G(jk) \\ G(j^2k) \end{pmatrix}, \tag{6.72}$$

which are, in fact, 6-vectors, solutions of

$$\frac{\partial}{\partial t}\mathbf{G}(k,t)=\hat{\mathbf{N}}(k,t)\mathbf{G}(k,t), \tag{6.73}$$

with

$$\hat{\mathbf{N}}(k,)=\begin{pmatrix} \mathbf{N}(k,) & 0 & 0 \\ 0 & \mathbf{N}(jk,) & 0 \\ 0 & 0 & \mathbf{N}(j^2k,) \end{pmatrix}, \tag{6.74}$$

and, according to (6.70) and (6.71),

$$\overleftarrow{\mathbf{G}}^{\mp}(k,t)=C^{\pm}(k)\vec{\mathbf{G}}^{\mp}(k,t)+D^{\pm}(k)\vec{\mathbf{G}}^{\pm}(k,t), \tag{6.75}$$

where C and D are diagonal matrices obtained from c and d by two ''j transformations,'' for instance,

$$C^{\pm}(k)=\begin{pmatrix} c^{\pm}(k) & 0 & 0 \\ 0 & c^{\pm}(jk) & 0 \\ 0 & 0 & c^{\pm}(j^2k) \end{pmatrix}, \tag{6.76}$$

it being understood that each element that is a number is multiplied by the 2×2 unit matrix (we call this agreement the ''duplex rule''). We define now the ''transformed'' Jost vectors $\mathbf{\Psi}$ from the Jost solutions G as we defined in (6.37)–(6.39) the solutions $\mathbf{\Phi}$ of (6.40) from the solutions F of (6.32). Using (6.72), we can write

$$\mathbf{\Psi}=\tfrac{1}{3}T\mathbf{G}, \tag{6.77}$$

where

$$T=\begin{pmatrix} 1 & 1 & 1 \\ k^{-1} & j^2k^{-1} & jk^{-1} \\ k^{-2} & jk^{-2} & j^2k^{-2} \end{pmatrix} \tag{6.78}$$

(with duplex rule). The inverse matrix,

$$T^{-1}=\frac{1}{3}\begin{pmatrix}1 & k & k^2\\ 1 & jk & j^2k^2\\ 1 & j^2k & jk^2\end{pmatrix}\tag{6.79}$$

(with duplex rule), enables us to go back from $\boldsymbol{\Psi}$ to G through (6.77). The $\boldsymbol{\Psi}$'s are solutions of

$$\frac{\partial}{\partial t}\boldsymbol{\Psi}=\tilde{\mathbf{N}}\boldsymbol{\Psi},\tag{6.80}$$

$$\tilde{\mathbf{N}}=T\hat{\mathbf{N}}T^{-1}=\begin{pmatrix}n_0 & k^3n_1 & -k^6n_2\\ -k^3n_2 & n_0 & k^3n_1\\ n_1 & -k^3n_2 & n_0\end{pmatrix}.\tag{6.81}$$

It is easy to check that $\boldsymbol{\Psi}$ is also a solution of (6.40). The asymptotic behaviors of the G's can be derived from (6.27), (6.28), and (6.72), and that of the $\boldsymbol{\Psi}$'s from that of the $\mathbf{G}$'s by means of (6.77):

$$\vec{\boldsymbol{\Psi}}^{\pm}(k,t)\xrightarrow{t\to\infty}e^{\pm ik^3t}v_0^{\mp},\tag{6.82}$$

$$\overleftarrow{\boldsymbol{\Psi}}^{\mp}(k,t)\xrightarrow{t\to-\infty}e^{\mp ik^3t}v_0^{\pm},\tag{6.83}$$

where

$$v_0^{\pm}=\begin{pmatrix}\pm i\\ 0\\ 0\\ 1\\ 0\\ 0\end{pmatrix}.\tag{6.84}$$

Since the matrix Jost solutions $\mathcal{M}$ are the solutions of (6.40) that go to the identity matrix at infinite t, it follows from (6.32) that

$$\vec{\boldsymbol{\Psi}}^{\pm}(k,t)=\vec{\mathcal{M}}^{\pm}(k^3,t)v_0^{\mp},\tag{6.85}$$

$$\overleftarrow{\boldsymbol{\Psi}}^{\mp}(k,t)=\overleftarrow{\mathcal{M}}^{\mp}(k^3,t)v_0^{\pm}.\tag{6.86}$$

Hence, the Ψ's have the analytic and asymptotic properties of the corresponding $\mathcal{M}$'s. Readers may notice that in the Jaulent Jean second-order problem analysis, a vector showing only one nonzero component was used instead of v_0. It was so because such vectors were a basis of solutions of the asymptotic form of their equation replacing (6.40). Here, $v_0^{\pm}$ comes in as an eigenvector of the asymptotic form of their (6.77). Multiplying the G's in (6.27) and (6.28) by k and k^2, then deriving the $\mathbf{G}$'s and the $\boldsymbol{\Psi}$'s would lead to two other eigenvectors, $v_1^{\pm}$ and $v_2^{\pm}$, linearly independent of $v_0^{\pm}$, and the 6-dimensions space would have a basis.

From (6.75) and (6.77), we define the ''transformed scattering problem,''

$$\overleftarrow{\boldsymbol{\Psi}}^{\mp}(k,t)=\tilde{C}^{\pm}(k)\vec{\boldsymbol{\Psi}}^{\mp}(k,t)+\tilde{D}^{\pm}(k)\vec{\boldsymbol{\Psi}}^{\pm}(k,t),\tag{6.87}$$

where

$$\tilde{C}^{\pm}=TC^{\pm}T^{-1},\tag{6.88}$$

and similarly for $\tilde{D}^{\pm}$. So as to derive the analytic and asymptotic properties of $\tilde{C}^{\pm}$ and $\tilde{D}^{\pm}$, we have now to go back and forth several times between the scattering problem and the transformed scattering problem. Let Ψ_0, Ψ_1, Ψ_2 the three 2-vectors ''components'' of a $\boldsymbol{\Psi}$. From (6.77) we get in a domain of holomorphy of $k^3 \to \boldsymbol{\Psi}$,

$$G(k,t)=\Psi_0(k^3,t)+k\Psi_1(k^3,t)+k^2\Psi_2(k^3,t). \tag{6.89}$$

If we want to study, for instance, $\tilde{C}^+(k)$, we know that $\vec{\Psi}_l^+, \overleftarrow{\Psi}_l^-$, are holomorphic functions of k^3 in $\operatorname{Im} k^3 \geqslant 0$. From (6.70), we derive

$$2ikc^+=\det[\overleftarrow{G}^-,\vec{G}^+], \tag{6.90}$$

and substituting in (6.90) the respective expansions (6.89) of $\overleftarrow{G}^-$ and $\vec{G}^+$, we obtain that $c^+(k)$ is equal to

$$c^+(k)=k^{-1}[c_0^+ + c_1^+ k + c_2^+ k^2], \tag{6.91}$$

where c_0^+, c_1^+, c_2^+ are holomorphic functions of k^3 in $\operatorname{Im} k^3 \geqslant 0$, with

$$2ic_0^+=\det[\overleftarrow{\Psi}_0^-,\vec{\Psi}_0^+]. \tag{6.92}$$

It follows from (6.76) and (6.88) that

$$\tilde{C}^+=\begin{pmatrix} c_1^+ & c_0^+ & c_2^+k^3 \\ c_2^+ & c_1^+ & c_0^+ \\ c_0^+k^{-3} & c_2^+ & c_1^+ \end{pmatrix}. \tag{6.93}$$

All the coefficients of $\tilde{C}^+$ are holomorphic functions of k^3 in $\operatorname{Im} k^3 \geqslant 0$, except maybe one, which has a pole at $k=0$ unless $c_0^+(0)=0$. We can study this condition by making $\lambda \to 0$ in (6.43) and (6.54). It is exactly similar to the generic condition of no zero energy bound state in the theory of (ordinary) potential scattering, and for the sake of simplicity, we henceforth assume it, together with the symmetric one on c^-. It would not be difficult to generalize the following theory to the case $c_0^+(0)c_0^-(0)\neq 0$. The study of c^- is symmetric to the study of c^+, and yields similar results in $\operatorname{Im} k \leqslant 0$. But the study of $d^{\pm}$ can be done only on the real k axis, where arguments similar to those above show that $\tilde{D}(\lambda)$ has continuous coefficients on $\lambda \in \mathbb{R}$.

Let us now look at the asymptotic behavior of $\tilde{C}^+$ as $|\lambda| \to \infty$ in $\operatorname{Im} k \geqslant 0$ and of $\tilde{D}^+(\lambda)$ on $\mathbb{R}$. For doing it, again we go back to solutions of (6.11). We first give formulas that generalize Wronskian results of potential theory.

Let $G=\binom{G_1}{G_2}$ be any solution of (6.11), where $\mathbf{N}$ stands, and E any solution of the same equation with $\tilde{N}$, instead of $\mathbf{N}$. From (6.68) and (6.69), we can write down

$$\mathbf{N}=N^{(1)}+U, \quad \tilde{N}=\tilde{N}^{(1)}-k^{-4}\nu(k,t)I, \tag{6.94}$$

where $\nu(k,t)$ is t integrable uniformly in k for $|k|$ or t large. All the other matrices in (6.94) can be written in the general form

$$N=\begin{pmatrix} N_1 & N_2 \\ N_3 & -N_1 \end{pmatrix} \tag{6.95}$$

(notice that, more specifically, $N_1=U_2=0$).

The time derivative of

$$\Delta=\det[G,E], \tag{6.96}$$

can be calculated from (6.94), and it yields

$$\Delta(+\infty)-\Delta(-\infty)=\int_{-\infty}^{+\infty}\{G_1E_1[-U_3+\tilde{N}_3^{(1)}-N_3^{(1)}]+G_2E_2[N_2^{(1)}-\tilde{N}_2^{(1)}]$$

$$+(G_1E_2+G_2E_1)(N_1-\tilde{N}_1+U_1)+(G_1E_2-G_2E_1)k^{-4}\nu\}, \tag{6.97}$$

where $\tilde{N}_i-N_i^{(1)}$ is equal for large $|k|$ to $k^{-2}\nu_i(t)+k^{-4}\nu_i(k,t)$, the functions ν being integrable.

As for the functions G of interest, they can be appraised uniformly in t for large $|k|$ from the formula (6.52), which yields through (6.48) and (6.56) the asymptotics of $\vec{G}^+$, for instance, as

$$\vec{G}^+=\begin{pmatrix} -i+\vec{g}_1^+(t)k^{-3}+O(|k|^{-6}) \\ k-\dfrac{1}{2i}\displaystyle\int_t^\infty B_1(u)du+\vec{g}_2^+(t)k^{-2}+O(|k|^{-5}) \end{pmatrix}, \tag{6.98}$$

and similar formulas hold for $\vec{G}^-$ and $\overleftarrow{G}^{\pm}$. The undefined quantities g_1, g_2 are integrable and the numbers c,d,d_∞, which will appear below, are bounded. Setting now $\overleftarrow{G}^-$, $\overleftarrow{E}^-$ in (6.96), and using the constant ratio between $\overleftarrow{E}^-$ and $\vec{E}^-$, we obtain for large real $|k|$,

$$-2ikd^+(k)=\exp[-i\omega_0/k][d_\infty^+ +k^{-1}d+O(|k|^{-3})]. \tag{6.99}$$

A similar work on $\det[\vec{G}^+,\overleftarrow{E}^- -\overleftarrow{G}^-]$ yields

$$c^+(k)-\exp[i\omega_0/k]=(2ik)^{-1}\exp[i\omega_0/k][c_\infty^+ +k^{-2}c+O(|k|^{-4})], \tag{6.100}$$

an estimate that extends to $\operatorname{Im}k^3\geqslant 0$, and where

$$c_\infty^+=-\int_{-\infty}^{+\infty}V_2(t)dt. \tag{6.101}$$

Hence, the leading asymptotic behavior of the diagonal matrix $C^+(k)$ as $|k|\to\infty$ is the unit matrix, and since its determinant is that of the holomorphic function $\tilde{C}^+$ in $\operatorname{Im}k^3\geqslant 0$, it can vanish there only at finitely many separate points. Notice that integrating by parts in the derivation of $d^+(k)$ by means of the formula (6.97) is possible because $\overleftarrow{G}^-$ and $\overleftarrow{E}^-$ both contain the same factor $\exp[-ik^3t]$, and if the convenient assumptions on time derivatives of V hold, it yields

$$d^+(k)=O(|k|^{-4}). \tag{6.102}$$

Similar derivations hold for $c^-(k)$, $d^-(k)$, and the related 3×3 ''duplex rule'' matrices. The final result of these estimates is that

$$(\tilde{C}^{\pm}(k)]^{-1}\tilde{D}^{\pm}(k)=O(|k|^{-6}),\quad k\in\mathbb{R},\quad |k|\to\infty, \tag{6.103}$$

and that there exists upper triangular matrices, with 1 as diagonal elements, for instance,

$$\tilde{C}^+(\infty)=\begin{pmatrix} 1 & i(\omega_0-\frac{1}{2}c_\infty^+) & -\omega_0^2 \\ 0 & 1 & i(\omega_0-\frac{1}{2}c_\infty^+) \\ 0 & 0 & 1 \end{pmatrix}, \tag{6.104}$$

such that

$$\tilde{C}^{\pm}(k)-\tilde{C}^{\pm}(\infty)=0(|k|^{-3})\quad(\pm\operatorname{Im}k^3\geqslant 0). \tag{6.105}$$

D. Inversion equations

We write down (6.87) as

$$[\tilde{C}^{\pm}(\lambda)]^{-1}\overleftarrow{\boldsymbol{\Psi}}^{\mp}(\lambda,t)=\vec{\boldsymbol{\Psi}}^{\mp}(\lambda,t)+\tilde{R}^{\pm}(\lambda)\vec{\boldsymbol{\Psi}}^{\pm}(\lambda,t), \tag{6.106}$$

where the notation $\lambda=k^3$ is used instead of k, without modifying the function's name, in order to remind that they depend on λ only, and where

$$\tilde{R}^{\pm}(\lambda)=[\tilde{C}^{\pm}(\lambda)]^{-1}\tilde{D}^{\pm}(\lambda) \tag{6.107}$$

is the ''reflection matrix,'' $[\tilde{C}^{\pm}]^{-1}$ being the ''transmission matrix'' of the transformed scattering problem. Let us first focus on the upper indices, which label functions $\boldsymbol{\Psi}$ and $\tilde{C}$ that are holomorphic in $\mathrm{Im}\,\lambda\geqslant 0$. Using (6.58) and (6.85), we can write down (6.106) as

$$\mathcal{G}^{+}(\lambda,t)-\mathcal{H}^{+}(\lambda,t)=\int_t^{\infty}\vec{K}^{-}(t,w)e^{-i\lambda w}\,dw\,v_0^{+}, \tag{6.108}$$

where

$$\mathcal{G}^{+}(\lambda,t)=[\tilde{C}^{+}(\lambda)]^{-1}\overleftarrow{\boldsymbol{\Psi}}^{-}(\lambda,t)-e^{-i\lambda t}\vec{P}^{-}(t)v_0^{+}, \tag{6.109}$$

$$\mathcal{H}^{+}(\lambda,t)=\tilde{R}^{+}(\lambda)[\vec{\boldsymbol{\Psi}}^{+}(\lambda,t)-e^{i\lambda t}\vec{P}^{+}(t)v_0^{-}]+\tilde{R}^{+}(\lambda)e^{i\lambda t}\vec{P}^{+}(t)v_0^{-}. \tag{6.110}$$

From (6.103), (6.104), and (6.105), the asymptotic behaviors of $\mathcal{G}^{+}$ and $\mathcal{H}^{+}$ on the real axis are easily shown to be $O(|\lambda|^{-1})$ or less. Hence, we can make the Fourier transform of both sides of (6.108), obtaining

$$\vec{K}^{-}(t,u)v_0^{+}=S^{+}(t+u)\vec{P}^{+}(t)v_0^{-}+\int_t^{\infty}S^{+}(u+u')\vec{K}^{+}(t,u')du'v_0^{-}+g^{+}(t,u), \tag{6.111}$$

where

$$S^{+}(u)=-\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{i\lambda u}\tilde{R}^{+}(\lambda)d\lambda, \tag{6.112}$$

and $g^{+}(t,u)$ is the Fourier transform of $\mathcal{G}^{+}(\lambda,t)$ [it is defined as in (6.112)]. Now, the function $\mathcal{G}^{+}(\lambda)$ being $e^{-i\lambda t}0(\lambda^{-1})$ in $\mathrm{Im}\geqslant 0$, where it is a meromorphic function, we can evaluate $g^{+}(t,u)$ for $u\geqslant t$ by a contour integration along the real axis completed by a half-circle centered at $\lambda=0$, whose radius goes to ∞. Hence, $g^{+}(t,u)$ is equal to a finite sum of residues, each of them involving values of $\overleftarrow{\boldsymbol{\Psi}}^{-}(\lambda,t)$ or derivatives at a pole λ_m, or using (6.106), values of $\vec{\boldsymbol{\psi}}^{+}(\lambda,t)$ and derivatives at this point. In Sec. VII, we shall see how the vanishing of $\det[\overleftarrow{G}^{-},\vec{G}^{+}]$ at this point can be used to proceed, in a simplified case, and to derive an extra term for $S^{+}(t+u)$. Here we prefer, assuming for the sake of simplicity that there is no singularity of $\tilde{C}^{-1}$, so that $g^{\pm}(t,u)$ vanishes (i.e., the ''no bound state'' case of usual potential theory).

It follows that $\vec{K}^{-}$, $\vec{K}^{+}$, and S^{+}, are related by the inversion equation

$$\vec{K}^{-}(t,u)v_0^{+}=S^{+}(t+u)\vec{P}^{+}(t)v_0^{-}+\int_t^{\infty}S^{+}(u+w)\vec{K}^{+}(t,w)dw\,v_0^{-}, \tag{6.113}$$

where

$$S^{+}(u) = -\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{i\lambda u}\tilde{R}^{+}(\lambda)d\lambda. \tag{6.114}$$

Sticking to the scattering problem on the right, we can do similar derivations with the functions of the lower index in (6.106), obtaining with similar assumptions the equation

$$\vec{K}^{+}(t,u)v_0^{-} = S^{-}(t+u)\vec{P}^{-}(t)v_0^{+} + \int_t^{\infty} S^{-}(u+w)\vec{K}^{-}(t,w)dw\, v_0^{+}, \tag{6.115}$$

where

$$S^{-}(u) = -\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-i\lambda u}\tilde{R}^{-}(\lambda)d\lambda. \tag{6.116}$$

These derivations are similar to those of Jaulent Jean. We shall see in Sec. VII that they also produce a constraint coupling $\vec{K}^{-},\vec{K}^{+},\vec{P}^{-},\vec{P}^{+}$. Constraints are nuisances in true inverse problems. Fortunately, here, either constraints or consistency conditions are not a nuisance, since we shall start from the values of V and derivatives at a point, solve the direct problem to get $K^{\pm}$, and we already know from the analysis of formula (5.5) that the consistencies are preserved in the evolution. These remarks will become more clear when we come to the DLE. Here, let us only notice that if we tried to solve the true inverse scattering problem an obvious algebraic constraints on $\tilde{C}$, or $\tilde{C}^{-1}$, $\tilde{D}$, and $\tilde{R}$, would be that they are transformed from diagonal (duplex) 3-matrices as in (6.88), so that, if *a,b,c*, are its elements on the diagonal, the tilde matrix, for instance $\tilde{R}$, should be of the form

$$\tilde{R} = \begin{pmatrix} r_0 & r_2 & r_1 \\ k^{-3}r_1 & r_0 & r_2 \\ k^{-3}r_2 & k^{-3}r_1 & r_0 \end{pmatrix}, \tag{6.117}$$

where all the elements are functions of k^3 only, and are related to *a,b,c*, by

$$r_0 = \tfrac{1}{3}(a+b+c); \quad k^{-2}r_1 = \tfrac{1}{3}(a+j^2b+jc); \quad k^{-1}r_2 = \tfrac{1}{3}(a+jb+j^2c), \tag{6.118}$$

in agreement with the fact that, in the direct problem, b and c are j transforms of $a=r(k)$.

VII. THE FINAL INTEGRAL EQUATIONS, SOLVING THE "OTHER CAUCHY PROBLEM"

A. Inversion integral equations at fixed *x*

We first derive the inversion integral equation at fixed x and the corresponding spectral measures from (6.58) and (6.113)–(6.114). For the sake of simplicity, we assume that there is no discrete term in (6.114). Equation (6.58) yields the transformation formula for $\vec{\Psi}$:

$$\vec{\Psi}^{\pm}(k^3,t) = \vec{P}^{\pm}(t)\exp[\pm ik^3t]v_0^{\mp} + \int_t^{\infty}\vec{K}^{\pm}(t,u)\exp[\pm ik^3u]du v_0^{\mp}, \tag{7.1}$$

whereas Eqs. (6.115)–(6.116) yield

$$\vec{K}^{\pm}(t,u)v_0^{\mp} = -\frac{1}{2\pi}\int_{-\infty}^{+\infty} d\lambda\, e^{\mp i\lambda u}\tilde{R}^{\mp}(\lambda)\left[\vec{P}^{\mp}(t)e^{\mp i\lambda t} + \int_t^{\infty} dw\,\vec{K}^{\mp}(t,w)e^{\mp i\lambda w}\right]v_0^{\pm}, \tag{7.2}$$

which is obviously the generalized Marchenko equation of the problem on a t path, at fixed x. As in standard inverse theories, we can insert (7.1) into (6.40) to get after some algebra the linear partial differential equation (for $t<u$) and the relations at $t=u$ relative to the kernels $K^{\pm}v_0^{\mp}$ (which vanish for $t>u$, according to the analytic properties):

$$K^{\pm}(t,u)\to 0, \quad u\to\infty, \tag{7.3}$$

$$\left\{\left[\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial u^2}-\mathcal{A}(t)\right]\vec{K}^{\pm}(t,u)\mp i\mathcal{B}(t)\frac{\partial}{\partial u}\vec{K}^{\pm}(t,u)\right\}v_0^{\mp}=0, \tag{7.4}$$

$$\left[\frac{d}{dt}\vec{K}^{\pm}(t,t)\mp\frac{i}{2i}\mathcal{B}(t)\vec{K}^{\pm}(t,t)-\mathcal{E}^{\pm}(t)\vec{P}^{\pm}(t)\right]v_0^{\mp}=0, \tag{7.5}$$

where

$$\mathcal{E}^{\pm}(t)=-\frac{1}{2}\mathcal{A}(t)\pm\frac{1}{4i}\mathcal{B}'(t)-\frac{1}{8}\mathcal{B}^2(t). \tag{7.6}$$

The equations (7.2) and (7.1) also yield a representation of $\vec{K}^{\pm}(t,u)v_0^{\mp}$:

$$\vec{K}^{\pm}(t,u)v_0^{\mp}=-\frac{1}{2\pi}\int_{-\infty}^{+\infty}d\lambda\, e^{\mp i\lambda u}\tilde{R}^{\mp}(\lambda)\vec{\Psi}^{\mp}(\lambda,t). \tag{7.7}$$

Inserting (7.7) into (7.1) yields a set of coupled integral equations for $\vec{\Psi}^{\pm}$:

$$\vec{\Psi}^{\pm}(k^3,t)=\vec{P}^{\pm}(t)e^{\pm ik^3t}v_0^{\mp}-\frac{1}{2\pi}\int_{-\infty}^{+\infty}d\lambda\,\tilde{R}^{\mp}(\lambda)\vec{\Psi}^{\mp}(\lambda,t)\int_t^{\infty}du\, e^{\pm i(k^3-\lambda\pm i0^+)u}. \tag{7.8}$$

The equations (7.7)–(7.8) are, of course, the inversion equations of the transformed time-path scattering problem. Here (7.4) and (7.5) are consistency relations involving V. Going to the true time-path scattering problem is done by using (6.118) and (6.77) for deriving **G**, using also (6.49) and (6.35) for handling the first term on the rhs of (7.8), (6.88), and (6.107) for the second one. Setting $\lambda=\mu^3$ in the integral, we obtain

$$\vec{\mathbf{G}}^{\pm}(k,t)=\vec{\mathbf{G}}_0^{\pm}(k,t)\mp\frac{i}{2\pi}T^{-1}(k)\int_{-\infty}^{+\infty}d\mu^3\,T(\mu)R^{\mp}(\mu)\vec{\mathbf{G}}^{\mp}(\mu,t)\frac{e^{\pm i(k^3-\mu^3)t}}{k^3-\mu^3\pm i0^+}, \tag{7.9}$$

where

$$\vec{\mathbf{G}}_0^{\pm}(k,t)=\begin{pmatrix}\vec{g}_0^{\pm}(k,t)\\ \vec{g}_0^{\pm}(jk,t)\\ \vec{g}_0^{\pm}(j^2k,t)\end{pmatrix}e^{\pm ik^3t}, \tag{7.10}$$

$$\vec{g}_0^{\pm}(k,t)=\begin{pmatrix}\mp i\\ k\mp(2i)^{-1}\int_t^{\infty}B_1(u)du\end{pmatrix}=:\begin{pmatrix}\mp i\\ k\pm ig(t)\end{pmatrix}, \tag{7.11}$$

and $R^{\mp}(k)$ was given by (6.117), (6.118). Notice that the values of the first ''duplex'' element of the diagonal, $r^{\pm}(k)$, yield the others by j transformation. Now, the first ''duplex'' element of the vector $\vec{\mathbf{G}}^{\pm}$ in (7.9) is easily derived:

$$\vec{G}^{\pm}(k,t)=\vec{g}_0^{\pm}(k,t)e^{\pm ik^3t}\mp\frac{i}{2\pi}\int_{-\infty}^{+\infty}d\mu\ \mu^2\sum_{l=0}^{l=2}\left(1+j^{2l}\frac{k}{\mu}+j^l\frac{k^2}{\mu^2}\right)r^{\mp}(j^l\mu)$$
$$\times\vec{G}^{\mp}(j^l\mu,t)(k^3-\mu^3\pm i0^+)^{-1}\exp[\pm i(k^3-\mu^3)t]. \tag{7.12}$$

B. The direct linearization equation

Equation (7.12) should be the fixed x reduction of the direct linearization equation we seek. It is an equation for two-vectors, but those we studied previously might also have been put in this form. What is unexpected and new is that two coupled integral equations come in. What is excellent is that the spectral measures that come in can be calculated as scattering data of the time-path scattering problem from $V(x_0,t)$, $V'(x_0,t)$, $V''(x_0,t)$, and/or time derivatives related to them by the Korteweg de Vries equation (x_0 is the fixed value of x, omitted in notations). By the way, notice that since $\mathbf{N}$ is a real matrix, $\vec{G}^-(k,t)$, $c^-(k)$, $d^-(k)r^-(k)$ are for a real k conjugate of $\vec{G}^+(k,t)$, $c^+(k)$, $d^+(k)$, $r^+(k)$, so that ''in fine,'' there are (in some way) three (complex) spectral measures r_0, r_1, r_2.

We have now to write down a direct linearization equation as x evolves. We find in (7.12) quantities inside $\vec{g}_0$ where x dependence was given from the beginning, and the scattering data. For the latter ones, we use the t-scattering analysis already sketched with the equation (6.29). More precisely, if t is so large that the Born formula (6.29) applies, we can write down, for $x_0=0$,

$$\vec{G}^{\pm}(k,x,t)=\exp[\mathbf{M}_0x]\vec{G}^{\pm}(k,0,t)\xrightarrow{t\to\infty}\begin{pmatrix}\mp i\\ k\end{pmatrix}e^{\pm i(kx+k^3t)}. \tag{7.13}$$

Hence, the ''time problem Jost solution'' $\vec{G}_J^{\pm}(k,x,t)$ is related to the solution of the time Lax equation that has evolved from 0 to x according to the space Lax equation by the restandardizing relation:

$$\vec{G}_J^{\pm}(k,x,t)=e^{\mp ikx}\vec{G}^{\pm}(k,x,t), \tag{7.14}$$

and, by the same token,

$$\overleftarrow{G}_J^{\mp}(k,x,t)=e^{\pm ikx}\overleftarrow{G}^{\mp}(k,x,t). \tag{7.15}$$

If we notice that $\mathbf{M}$ being a zero trace matrix, the determinant of two solutions of the space Lax equation is independent of x, we deduce from (7.14) (7.15), (6.70), and (6.71) that

$$c^{\pm}(k,x)=c^{\pm}(k,0);\quad d^{\pm}(k,x)=e^{\pm 2ikx}d^{\pm}(k,0), \tag{7.16}$$

$$r^{\pm}(k,x)=e^{\pm 2ikx}r^{\pm}(k,0). \tag{7.17}$$

From (7.12) and (7.17), we finally propose the direct linearization equation, which should be convenient for solving the second Cauchy problem:

$$\vec{G}^{\pm}(k,x,t)=\begin{pmatrix}\mp i\\ k\pm ig\end{pmatrix}e^{\pm iZ(k,x,t)}+\frac{1}{2\pi}\int_{-\infty}^{+\infty}d\mu\sum_{l=0}^{2}j^lr^{\mp}(j^l\mu)\vec{G}^{\mp}(j^l\mu,x,t)\sigma^{\pm}(k,j^l\mu,x,t), \tag{7.18}$$

where

$$g(x,t)=\frac{1}{2}\int_t^{\infty}B_1(u)du=\frac{1}{2}\int_x^{\infty}dy\ V(y,t), \tag{7.19}$$

$$Z(k,x,t)=kx+k^3t, \tag{7.20}$$

$$\sigma^{\pm}(k,\mu,x,t)=\mp i(k-\mu\mp i0^+)^{-1}e^{\pm i[Z(k,x,t)-Z(\mu,x,t)]}. \tag{7.21}$$

Notice that the equality between the two values of g in (7.19) is the consistency formula KdV between the two equations whose $G^{\pm}$ must be solutions. One could also say that it is one of the constraints plugged in the method to enforce that $\vec{G}^{\pm}$ solve the two Lax equations. We check now that they do it using the standard assumptions of the direct linearization methods [see them in other sections and, in particular, those after (5.63)]. Indeed, straightforward calculations show that

$$\begin{aligned}\left[\frac{\partial}{\partial x}-\begin{pmatrix}0 & 1\\ -k^2+V & 0\end{pmatrix}\right]G^{\pm}(k,x,t)\\ =\left[\begin{pmatrix}\mp ig\\ \pm i(g'+V)-kg\end{pmatrix}+\begin{pmatrix}1 & 0\\ \mp ik & 1\end{pmatrix}W^{0\mp}(x,t)\mp i\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}W^{1\mp}(x,t)\right]e^{\pm iZ(k,x,t)}\\ +\frac{1}{2\pi}\int_{-\infty}^{+\infty}d\mu\sum_0^2 j^l r^{\mp}(j^l\mu)\left[\frac{\partial}{\partial x}-\begin{pmatrix}0 & 1\\ j^{2l}\mu^2-V & 0\end{pmatrix}\right]\vec{G}^{\mp}(j^l\mu,x,t)\sigma^{\pm}(k,j^l\mu,x,t),\end{aligned} \tag{7.22}$$

where

$$W^{p\mp}(x,t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}d\mu(j^l\mu)^p\sum_{l=0}^{2} j^l r^{\mp}(j^l\mu)\vec{G}^{\mp}(j^l\mu,x,t)e^{\mp iZ(\mu,x,t)}. \tag{7.23}$$

The free terms vanish and the space Lax equation is solved if and only if the component $\vec{G}_2^{\pm}$ is the derivative of $\vec{G}_1^{\pm}$ and the two (i.e., $\pm$) following sums satisfy

$$\vec{G}_2^{\pm}(k,x,t)=\frac{\partial}{\partial x}\vec{G}_1^{\pm}(k,x,t), \tag{7.24}$$

$$W_1^{0\mp}(x,t)=\pm ig. \tag{7.25}$$

As for the time Lax equation, we similarly proceed through

$$\begin{aligned}\left[\frac{\partial}{\partial t}-\mathbf{N}(k)\right]G^{\pm}(k,x,t)=\left[\frac{\partial}{\partial t}-\mathbf{N}(k)\right]\left[\begin{pmatrix}\mp i\\ k\pm ig\end{pmatrix}e^{\pm iZ(k,x,t)}\right]\\ +\frac{1}{2\pi}\int d\mu\sum_0^2 j^l r^{\mp}(j^l\mu)[\mathbf{N}(j^l\mu)-\mathbf{N}(k)][G^{\mp}(j^l\mu,x,t)]\sigma^{\pm}(k,\mu,x,t)\\ +\frac{1}{2\pi}\int d\mu\sum_0^2 j^l r^{\mp}(j^l\mu)G^{\mp}(j^l\mu,x,t)\frac{\partial}{\partial t}\sigma^{\pm}(k,\mu,x,t)\\ +\frac{1}{2\pi}\int d\mu\sum_0^2 j^l r^{\mp}(j^l\mu)\left[\left(\frac{\partial}{\partial t}-\mathbf{N}(j^l\mu)\right)G^{\mp}(j^l\mu,x,t)\right]\sigma^{\pm}(k,\mu,x,t),\end{aligned} \tag{7.26}$$

and the free term we obtain is equal to

$$\text{F.T.}=\begin{pmatrix} \mp ik^2 g \\ -k^3 g \pm ik^2 V_0 \pm i\dfrac{\partial g}{\partial t}\end{pmatrix} - n_0\begin{pmatrix}\mp i \\ k\pm ig\end{pmatrix} - \{(I\mp ikn_2)[k^2 W^{0\mp}(x,t)+kW^{1\mp}(x,t)+W^{2\mp}(x,t)] \pm in_1[kW^{0\mp}(x,t)+W^{1\mp}(x,t)]\mp in_2[W^{3\mp}(x,t)]\}. \tag{7.27}$$

It vanishes if (7.25) holds, and

$$W_2^{0\mp}\mp iW_1^{1\mp}\pm iV_0=0, \tag{7.28}$$

$$W_2^{1\mp}\mp iW_1^{2\mp}-V_0 g+V_1=0, \tag{7.29}$$

$$W_2^{2\mp}\pm iV_0 W_1^{1\mp}\mp iW_1^{3\mp}\pm i\frac{\partial g}{\partial t}+iV_2\pm igV_{1=0}. \tag{7.30}$$

Equations (7.25), (7.28), (7.29), (7.30), yield $V(x,t)$ and its first and second derivatives as functions of x,t, and since the two equations of the Lax pair are satisfied for these functions, they give the solution of the Korteweg de Vries equation. It is interesting and not difficult (but lengthy) to check that for $x=0$, the values of the W's can be derived from the equations (7.5), (7.6) by applying the transformations that lead to $G^{\pm}(k,t)$ and to check that (7.25) and (7.28)–(7.30) are satisfied, together with the equations

$$\frac{\partial}{\partial x}W_1^{p\mp}=\mp iW_1^{(p+1)\mp}+W_2^{p\mp}, \tag{7.31}$$

$$\frac{\partial}{\partial x}W_2^{p\mp}=-W^{(p+2)\mp}\mp iW_2^{(p+1)\mp}, \tag{7.32}$$

which are derived readily from (7.24) and the space Lax equation.

Hence, the coupled vectorial equations (7.18) have almost all properties of a DLE, but one: they cannot be written readily from the values of $V(x,t)$ at $x=0$, since $g(x,t)$ [see (7.19)] contains V, and if we knew $V(x,t)$ the problem would be solved. Obviously, this point is easy to circumvent because the kernel of (7.18) is scalar. Hence, (7.18) can be separated into two couples of scalar equations: the first one for $\vec{G}_1^{\pm}(k,x,t)$, the second one for $\vec{G}_2^{\pm}(k,x,t)$. The first one can be written and solved as soon as we know the time scattering data at $x=0$, and it yields $g(x,t)$, and $V(x,t)$, by means of the consistency condition (7.26). Using $g(x,t)$ defines the second couple of equations, whose solutions give readily various quantities, as, for instance, the other W's, and enables one to check that the method works. Hence, the following couple of scalar integral equations is really the DLE of our ''other Cauchy problem:''

$$\vec{G}_1^{\pm}(k,x,t)=\mp ie^{\pm iZ(k,x,t)}+\frac{1}{2\pi}\int_{-\infty}^{+\infty}d\mu\sum_{l=0}^{2}j^l r^{\mp}(j^l\mu)\vec{G}_1^{\mp}(j^l\mu,x,t)\sigma^{\pm}(k,j^l\mu,x,t), \tag{7.33}$$

and the potential is derived from its solution by means of the formula (7.25). The ''time inverse scattering method'' can be summarized as so

(1) At $x=0$ (say) where V and its first two x derivatives are known for any t, solve the direct problem (6.11), so as to derive the Jost solutions $\vec{G}^{\pm}(k,0,t)$ and $\overleftarrow{G}^{\mp}(k,0,t)$ defined by (6.27), (6.28).

(2) Derive det$[\overleftarrow{G}^{-},\vec{G}^{+}]$ and det$[\overleftarrow{G}^{+},\vec{G}^{-}]$ to get $c^{\pm}(k)$, $d^{\pm}(k)$ and $r^{\pm}(k)$. Check (if it is the case) that there is no pole of $\tilde{T}^{\pm}(\lambda)$ in Im $\lambda\geqslant 0$.

(3) Insert the scattering data $r^{\pm}$ into (7.33) and solve it, you get $V(x,t)$ by means of (7.25).

C. The case of poles in the transmission coefficient, with zero reflection coefficient

From (6.108), and the equation obtained by twisting indices $+$ and $-$, we saw that poles can contribute functions $g^{\pm}(t,u)$ in (6.111) and the twisted equation. The general case is complicated. We introduce three simplifying assumptions.

Assumption A: We assume that $[\tilde{C}^{\pm}(\lambda)]^{-1}$ have only simple poles in $\pm \mathrm{Im}\,\lambda > 0$.

These poles correspond to zeros of $\det \tilde{C}^{\pm}(\lambda)$ and thus of $\det C^{\pm}(k)$. For such a zero λ_0, according to the determination of the cubic root k_0, the product $c(k)$, $c(jk)c(j^2k)$ vanishes!

Assumption B: We assume that if k_n is such that $c(k_n)$ vanishes, $c(jk_n)$ and $c(j^2k_n)$ do not.

Finally, in order to obtain relatively simple formulas, and because it is not difficult to remove it by taking into account the preceding results, we also introduce the following.

Assumption C: The reflection coefficients [or $d^{\pm}(k)$] vanish for any real k. Thanks to Assumptions A and C, $\vec{K}^{-}(t,u)$ can then be directly derived from (6.108) and (6.109) by Fourier transforming $\mathcal{G}^{+}(\lambda,t)$, and the Fourier transform can be calculated for $u>t$ as a sum of the residues of $[C^{+}(\lambda)]^{-1}\overleftarrow{\boldsymbol{\Psi}}^{-}(\lambda,t)e^{i\lambda u}$ at the poles $\lambda_1^{+}, \lambda_2^{+}, \lambda_n^{+}$ [the second term in (6.109) only implies that the contribution on the infinite half-circle vanishes]. The same method (in $\mathrm{Im}\,\lambda<0$) applies to the quantities with the other $\pm$ index. The result is

$$\vec{K}^{\mp}(t,u)v_0^{\pm} = \pm i \sum_{n=1}^{N} e^{\pm i\lambda_n^{\pm}} \Gamma_n^{\pm} \overleftarrow{\boldsymbol{\Psi}}^{\mp}(\lambda_n^{\pm},t), \tag{7.34}$$

where

$$\Gamma_n^{\pm} = \lim_{\lambda\to\lambda_0^{\pm}} (\lambda-\lambda_0^{\pm})[\tilde{C}^{\pm}(\lambda)]^{-1}. \tag{7.35}$$

Inserting (7.33) into (7.1), we obtain

$$\overleftarrow{\boldsymbol{\Psi}}^{\pm}(k^3,t) = \vec{P}^{\pm}(t)\exp(\pm ik^3t]v_0^{\mp} \mp i\sum_{n=1}^{N}\int_t^{\infty} du\, \exp[\pm i(k^3-\lambda_n^{\mp})u]\Gamma_n^{\mp}\overleftarrow{\boldsymbol{\Psi}}^{\pm}(\lambda_n^{\mp},t). \tag{7.36}$$

Let us now notice that if for each $\lambda_n^{\pm}$, we define $k_n^{\pm}$ as the cubic root of $\lambda_n^{\pm}$ whose phase is such that $c^{+}(k_n^{+})$ [resp., $c^{-}(k_n^{-})$] vanishes, and thanks to the assumption B, and to the formula (6.88), we derive from (7.35),

$$T^{-1}(k_n^{\mp})\Gamma_n^{\mp}T(k_n^{\mp}) = \begin{pmatrix} 3(k_n^{\mp})^2 \lim_{k\to k_n^{\mp}} \dfrac{k-k_n^{\mp}}{c^{\mp}(k)} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{7.37}$$

Applying $3T^{-1}(k)$ to both sides of (7.36), we derive an equation for $\mathbf{G}^{\pm}(k,t)$ and thanks to (7.37) we see at once that the three duplex components of the six vector carry on the same information (except for trivial j transformations). Hence, we write down

$$\vec{G}^{\pm}(k,t) = \vec{G}_0^{\pm}(k,t) + \sum_{n=1}^{N} \frac{\exp[\pm i(k^3-\lambda_n^{\mp})t]}{k^3-\lambda_n^{\mp}} (k_n^{\mp})^2 T^{-1}(k)T(k_n^{\mp})\gamma^{\mp}(k_n^{\mp})\overleftarrow{G}^{\pm}(k_n^{\mp},t), \tag{7.38}$$

where

$$\gamma^{\pm}(k_n^{\pm}) = \lim_{k\to k_n^{\pm}}\left[\frac{k-k_n^{\pm}}{c^{\pm}(k)}\right]. \tag{7.39}$$

Now, since $k_n^{\pm}$ is a zero of $c^{\pm}(k)$, $\det[\overleftarrow{G}^{\mp},\vec{G}^{\pm}]$ vanishes there, so that the functions $\overleftarrow{G}^{-}$ and $\vec{G}^{+}$, $\overleftarrow{G}^{+}$ and $\overleftarrow{G}$ are proportional when $k=k_n^{+}$, resp. $k=k_n^{-}$. Multiplying $\gamma^{\mp}(k_n^{\mp})$ by the corresponding factors yields, say, $\rho_n^{\mp}$, and expanding $T^{-1}(k)T(k_n^{\mp})$ kills partially the denominators in (7.38).

Hence, we obtain the coupled equations,

$$\vec{G}^{\pm}(k,t)=\vec{G}_0^{\pm}(k,t)+\sum_{n=1}^{N}\frac{\exp[\pm i(k^3-\lambda_n^{\mp})t]}{k-k_n^{\mp}}\rho_n^{\mp}\vec{G}^{\mp}(k_n^{\mp},t). \tag{7.40}$$

As an example, the well-known ''soliton,''

$$V(x,t)=-\frac{2\kappa^2}{\cosh^2[\kappa(x-x_0)-\kappa^3 t]}, \tag{7.41}$$

is related at fixed x, say $x=x_0$, to the following functions and numbers:

$$\vec{G}^{+}(k,t)=-i(k+i\kappa)(k-i\kappa)^{-1}F(k,t), \tag{7.42}$$

$$\vec{G}^{-}(k,t)=i(k-i\kappa)(k+i\kappa)^{-1}\tilde{F}(k,t), \tag{7.43}$$

$$\overleftarrow{G}^{+}(k,t)=-iF(k,t), \tag{7.44}$$

$$\overleftarrow{G}^{-}(k,t)=i\tilde{F}(k,t), \tag{7.45}$$

where

$$F(k,t)=e^{ik^3t}\begin{pmatrix}\dfrac{k+i\kappa(1-\alpha)}{k+i\kappa}\\ \dfrac{ik^2+k\kappa(\alpha-1)+i\kappa^2\beta}{k+i\kappa}\end{pmatrix}, \tag{7.46}$$

$$\tilde{F}(k,t)=e^{-ik^3t}\begin{pmatrix}\dfrac{k-i\kappa(1-\alpha)}{k-i\kappa}\\ \dfrac{ik^2+k\kappa(\alpha-1)-i\kappa^2\beta}{k-i\kappa}\end{pmatrix}, \tag{7.47}$$

$$\alpha=e^{\kappa^3 t}/\cosh\kappa^3 t \quad \beta=2\cosh^2\kappa^3 t, \tag{7.48}$$

$$c^{\pm}(k)=\frac{k\pm i\kappa}{k\mp i\kappa},\quad d^{\pm}(k)=0, \tag{7.49}$$

κ is positive, and it is not difficult to derive the value of $\rho_0^{\pm}$ for the (unique) pole,

$$k_0^{\pm}=\mp i\kappa,\quad \rho_0^{\mp}=\pm 2i\kappa, \tag{7.50}$$

and to check (7.40).

From (7.38) we can derive the direct linearization equation for solutions of this kind. We write readily the ''first component'' and the equations that yields the potential:

$$\vec{G}_1^{\pm}(k,x,t)=\mp ie^{\pm iZ(k,x,t)}+\sum_{n=1}^{N}\frac{\exp[\pm i(k^3t-(k_n^{\mp})^3t+kx-k_n^{\mp}x)]}{k-k_n^{\mp}}\rho_n^{\mp}\vec{G}_1^{\mp}(k_n^{\mp},x,t), \tag{7.51}$$

$$g(x,t)=\sum_{n=1}^{N} \exp[\mp i((k_n^{\mp})^3 t+k_n^{\mp}x)]\rho_n^{\mp}G_1^{\mp}(k_n^{\mp},x,t), \tag{7.52}$$

$$V(x,t)=-2\frac{\partial g(x,t)}{\partial x}. \tag{7.53}$$

They could be used to solve the ''other Cauchy problem'' for a multisolitonic solution. The reader will easily check that using (7.51) in these three formulas give back $V(x,t)$, as given by (7.41) (for $x_0=0$).

D. Final remarks

The linearization equations (7.33) and (7.51) could also be used to construct generalized two variable Marchenko equations like we did in the Gelfand Levitan case in our previous paper and recall it in Sec. III, since they can also be understood as a generalized two-variable Povzner Levitan representation for $\vec{G}^{\pm}(k,x,t)$. We shall not write down these equations, which may be a step to further generalizations.

ACKNOWLEDGMENTS

It is my pleasure to acknowledge some discussions on this work with Professor Marcel Jaulent and with the participants of the Guardamar meeting on nonlinear equations, organized in July 1998 by Professor Luis Martinez Alonso and with participants of the Vietri meeting on Inverse Problems, organized in October 1998 by Professor Mario Bertero.

[1] P. C. Sabatier, ''On local Inverse Transforms,'' J. Math. Phys. **10**, 5396–5405 (1998).
[2] F. Calogero and A. Degasperis, *Spectral Transform and Solitons* (North Holland, Amsterdam, 1982).
[3] P. C. Sabatier, ''Spectral transform for nonlinear evolution equations with *N* space-dimensions,'' Phys. Lett. A **161**, 345–351 (1992); M. Boiti, F. Pempinelli, and P. C. Sabatier, ''First and second order non linear partial differential evolution equations from an inverse problem,'' Inverse Probl. **9**, 1–37 (1993); P. C. Sabatier, ''Inverse problems related to Integrable Nonlinear Partial Differential Equations,'' pp. 193–202, in *Inverse Problems in Mathematical Physics*, Lecture Notes in Physics, edited by L. Paivarinta and E. Somersalo (Springer-Verlag, Berlin, 1993), Vol. 422; P. C. Sabatier, ''Constructing nonlinear integrable equations by direct linearization,'' Teor. Mat. Fiz. , **99**, 517–527 (1994). P. C. Sabatier, ''On algebra of direct linearization,'' pp. 208–215 in *Nonlinear Evolution Equations and Dynamical Systems*, edited by V. Makhankov, A. Bishop, and D. Holm (World Scientific, Singapore, 1995).
[4] A. S. Fokas and B. Pelloni, ''The solution of certain initial boundary-value problems for the linearized Korteweg-de Vries equation,'' Proc. R. Soc. London, Ser. A **454**, 645–657 (1998); A. S. Fokas and B. Pelloni, ''A spectral method for linear and integrable nonlinear PDE's in arbitrary domains,'' preprint, May 1998.
[5] K. Chadan and P. C. Sabatier, ''*Inverse Problems in Quantum Scattering Theory*,'' 2nd ed., revised and expanded (Springer Verlag, Berlin, 1989).
[6] R. P. Boas, *Entire Functions* (Academic, New York, 1954), p. 15 & ff.
[7] M. Jaulent and C. Jean, ''Solution of a Schrodinger inverse scattering problem with a polynomial spectral dependence in the potential,'' J. Math. Phys. **23**, 258–266 (1982).
[8] P. C. Sabatier, ''Approach to scattering problems through interpolation formulas and application to spin-orbit potentials,'' J. Math. Phys. **9**, 1241–1258 (1968).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Hamiltonian structures of generalized Manin–Radul super-KdV and constrained super KP hierarchies

Ming-Hsien Tu[a)]
Department of Mathematics, National Chung Cheng University, Mingshiung, Chiayi 621, Taiwan

Jiin-Chang Shaw[b)]
Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan

(Received 20 January 1999; accepted for publication 14 February 1999)

A study of Hamiltonian structures associated with supersymmetric Lax operators is presented. Following a constructive approach, the Hamiltonian structures of Inami–Kanno super-KdV hierarchy and constrained modified super-KP hierarchy are investigated from the reduced supersymmetric Gelfand–Dickey brackets. By applying a gauge transformation on the Hamiltonian structures associated with these two nonstandard super-Lax hierarchies, we obtain the Hamiltonian structures of generalized Manin–Radul super-KdV and constrained super-KP hierarchies. We also work out a few examples and compare them with the known results. © *1999 American Institute of Physics.* [S0022-2488(99)02206-9]

I. INTRODUCTION

In the past decade and more, the supersymmetric integrable systems have received much attention in the literature (for recent reviews, see Refs. 1–3 and references therein), especially in the explorations of the relationship to the supersymmetric conformal field theories and string theories. On the one hand, in superconformal/superstring theories, correlation functions are governed by supersymmetric extensions of the Korteweg–de Vries (KdV) [or Kadomtsev–Petviashvili (KP)] systems. On the other hand, the knowledge of super-KdV/KP systems have motivated people to study nonperturbative properties of superstrings. These superintegrable systems share many features in common: they have supersymmetric Lax representations, infinitely many conserved quantities and soliton solutions, etc. Furthermore, it is a common belief that they also possess bi-Hamiltonian structures that define the dynamical flows on the corresponding Poisson supermanifolds. In particular, for the super-KdV-type systems, the Poisson brackets relative to their associated second Hamiltonian structures provide extended superconformal algebras (*W* superalgebras) whose quantum versions serve as the highest weight representations of some infinite-dimensional symmetries in string theories.

The main purpose of this paper is to construct the Hamiltonian structures of the generalized Manin–Radul super-KdV (MR sKdV) and constrained super-KP (csKP) hierarchies (for the definitions of these hierarchies, see Sec. IV) using the method of gauge transformation. Although the Hamiltonian structures for the simplest cases have been obtained in Refs. 4 and 5, however, to our knowledge, those for the general cases are still unexplored. Our motivation comes from the fact that, for two gauge-equivalent integrable systems, the gauge transformation between them transforms not only the Lax formulations but also the Hamiltonian structures of the corresponding hierarchies. Hence, the preparation of suitable superintegrable systems that are gauge equivalent to the generalized MR sKdV and csKP hierarchies is the key in this approach. Our strategy is the following: First, for an odd-order super-Lax operator $\hat{L}$, we consider its associated supersymmet-

[a)]Electronic mail: mhtu@math.ccu.edu.tw
[b)]Electronic mail: shaw@math.nctu.edu.tw

ric Gelfand–Dickey (GD) bracket[6] defined by the Hamiltonian map J. We then consider a usual reduction that modifies the Hamiltonian map J to J_c. Second, we construct out two nonstandard superhierarchies from $(\hat{L},J_c)$ that have super-Lax operators defined by $K_A=\hat{L}D$ and $K_B=D^{-1}\hat{L}$, respectively. The former is referred to the Inami–Kanno sKdV (IK sKdV) hierarchy,[7] whereas the latter to the constrained modified sKP (cmsKP) hierarchy.[8–10] The Hamiltonian structures associated with K_i can also be constructed from J_c and are denoted by $\Omega^{(i)}$ $(i=A,B)$. Finally, we perform a gauge transformation on the systems $(K_i,\Omega^{(i)})$ and denote the resulting systems by $(\tilde{L}_i,\Theta^{(i)})$, which describe the Lax operators and the Hamiltonian structures of the generalized MR sKdV and csKP hierarchies.

In summary, we shall follow the following steps to achieve the goal:

$$(\hat{L},J_c)\rightarrow(K_i,\Omega^{(i)})\rightarrow(\tilde{L}_i,\Theta^{(i)}). \tag{1.1}$$

It will be shown below that each step described above automatically guarantees the requirement that the associated Hamiltonian structures should obey the super-Jacobi identity.

We organize this paper as follows: In Sec. II, we recall some basic facts concerning superpseudodifferential operators (SPDOs). We then introduce the second supersymmetric GD bracket and its reduction from a Miura transformation viewpoint. In Sec. III, the IK sKdV and the cmsKP hierarchies are defined. We give a detailed construction of their associated Hamiltonian structures from the reduced supersymmetric GD bracket. We find that, up to a sign, the Poisson brackets defined by their corresponding Lax operators have the same form. In Sec. IV, we define the generalized MR sKdV and csKP hierarchies by applying a gauge transformation to the IK sKdV and cmsKP hierarchies, respectively. We also show that this gauge transformation enables us to obtain the Hamiltonian structures associated with the generalized MR sKdV and csKP hierarchies. In Sec. V, we give several examples to compare them with the known results. We present our concluding remarks in Sec. VI.

II. SUPERSYMMETRIC GELFAND–DICKEY BRACKETS

To begin with, we consider the supersymmetric Lax operator of the form

$$L=D^n+U_{n-1}D^{n-1}+\cdots+U_0, \tag{2.1}$$

where the supercovariant derivative $D\equiv\partial_\theta+\theta\partial(\partial\equiv\partial/\partial x)$ satisfies $D^2=\partial$, θ is the Grassmann variable $(\theta^2=0)$, which together with the even variable $x\equiv t_1$ defines the $(1|1)$ superspace with coordinate (x,θ). The coefficients U_i are superfields that depend on the variables θ, t_i and can be represented by $U_i=u_i(t)+\theta v_i(t)$. The parity of a superfield U is denoted by $|U|$, which is zero for U being even and one for U being odd. Since L is assumed to be homogeneous under Z_2 grading, thus $|U_i|=n+i(\mathrm{mod}\,2)$. We will introduce the Poisson bracket associated with L on functionals of the form

$$F(U)=\int_B f(U), \tag{2.2}$$

where $f(U)$ is a homogeneous differential polynomial of U_i and $\int_B\equiv\int dx\,d\theta$ is the Berezin integral, such that if $f(U)=a(u,v)+\theta b(u,v)$, then $\int_B f(U)=\int b$. The supercovariant derivative D satisfies the supersymmetric version of the Leibniz rule;[11]

$$D^iU=\sum_{k=0}^{\infty}(-1)^{|U|(i-k)}\begin{bmatrix} i \\ k \end{bmatrix}U^{[k]}D^{i-k}, \tag{2.3}$$

where $U^{[k]}\equiv(D^kU)$ and the superbinomial coefficients $\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]$ are defined by

$$\begin{bmatrix} i \\ k \end{bmatrix} = \begin{cases} \begin{pmatrix} [i/2] \\ [k/2] \end{pmatrix} & \text{for } 0 \leqslant k \leqslant i \quad \text{and} \quad (i,k) \neq (0,1) \bmod 2, \\ (-1)^{[k/2]} \begin{bmatrix} -i+k-1 \\ k \end{bmatrix}, & \text{for } i<0, \\ 0, \quad \text{otherwise.} \end{cases} \tag{2.4}$$

For a given SPDO $P=\Sigma p_i D^i$, it is convenient to separate P into a direct sum of two linear spaces: $P=P_{\geqslant k} \oplus P_{<k}$ with $P_{\geqslant k}=\Sigma_{i \geqslant k} p_i D^i$ and $P_{<k}=\Sigma_{i<k} p_i D^i$. In particular, we denote $P_+ = \Sigma_{i \geqslant 0} p_i D^i$, $P_- = \Sigma_{i<0} p_i D^i$, and $(P)_0 = p_0$. We also define its super-residue as sres $P=p_{-1}$ and its supertrace as Str $P=\int_B \text{sres}\, P$. It can be shown that, for any two SPDOs P and Q, $\text{Str}[P,Q]=0$ for $[P,Q] \equiv PQ-(-1)^{|P||Q|}QP$, and hence $\text{Str}\, PQ=(-1)^{|P||Q|}\,\text{Str}\, QP$. Given a functional $F(U)=\int_B f(U)$, we define its gradient as

$$d_L F = \sum_{k=0}^{n-1} (-1)^k D^{-k-1} \frac{\delta f}{\delta U_k}, \tag{2.5}$$

and its variation as

$$\delta F = (-1)^{|F|+|L|+1}\, \text{Str}(\delta L\, d_L F), \tag{2.6}$$

where the variational derivative is defined by

$$\frac{\delta f}{\delta U_k} = \sum_{i=0}^{\infty} (-1)^{|U_k|i + i(i+1)/2} \left(\frac{\partial f}{\partial U_k^{[i]}} \right)^{[i]}. \tag{2.7}$$

The supersymmetric second GD bracket associated with L is given by[6,12,13]

$$\{F,G\}(L) = (-1)^{|F|+|G|+|L|+1}\, \text{Str}[J(d_L F) d_L G], \tag{2.8}$$

where the Hamiltonian map J is defined by

$$J(X) = (LX)_+ L - L(XL)_+, \tag{2.9}$$

where $X=\Sigma_k X_k D^k$. It has been shown[6,13] that (2.8) indeed defines a Hamiltonian structure, namely, it is antisymmetric and satisfies the super-Jacobi identity.

If we factorize $L=(D-\Phi_n)(D-\Phi_{n-1})\cdots(D-\Phi_1)$, which defines a supersymmetric Miura transformation between the coefficient functions U_i and the Miura fields Φ_i, then the second GD bracket (2.8) becomes

$$\{F,G\}(L) = \int_B \sum_{i=1}^{n} (-1)^i \left(D \frac{\delta f}{\delta \Phi_i} \right) \frac{\delta g}{\delta \Phi_i}, \tag{2.10}$$

which implies that the fundamental brackets of the Miura fields Φ_i are given by[6,12]

$$\{\Phi_i(X), \Phi_j(Y)\} = (-1)^i \delta_{ij} D\, \delta(X-Y), \tag{2.11}$$

where $X=(x,\theta)$, $Y=(y,\omega)$ and $\delta(X-Y) \equiv \delta(x-y)(\theta-\omega)$. This result is what we called the supersymmetric Kupershmidt–Wilson theorem. Equation (2.11) enables us to write down the fundamental brackets of U_k through the super-Miura transformation.

Next, let us consider the case when the constraint $U_{n-1}=0$ is imposed in (2.1). It can be easily shown that such constraint for odd n is second class, which will modify the Hamiltonian structure J. On the other hand, for even n, the constraint is first class and hence the induced

Poisson brackets can not be well defined. Therefore, for the odd-order operator $\hat{L}=D^{2k+1}+U_{2k-1}D^{2k-1}+\cdots+U_0$, we shall consider the factorization $\hat{L}=(D-\Phi_{2k+1})(D-\Phi_{2k})\cdots(D-\Phi_1)$. Then the modified Poisson bracket defined by $\hat{L}$ becomes

$$\{F,G\}_c=(-1)^{|F|+|G|}\,\mathrm{Str}(J_c(\hat{d}F)\hat{d}G), \tag{2.12}$$

where $\hat{d}F\equiv d_{\hat{L}}F=\Sigma_{i=0}^{2k-1}(-1)^iD^{-i-1}(\delta f/\delta U_i)$ and

$$J_c(\hat{d}F)=J(\hat{d}F)+\left[\hat{L},\int^x D\ \mathrm{sres}[\hat{L},\hat{d}F]\right]. \tag{2.13}$$

We remark that the second term is called the third GD structure, which is compatible with the second structure. Equation (2.12) yields that the modified Poisson brackets for the Miura fields Φ_i are given by

$$\{\Phi_i(X),\Phi_j(Y)\}_c=[1+(-1)^i\delta_{ij}]D\delta(X-Y), \tag{2.14}$$

which provide the free-field realizations of classical W superalgebras associated with the odd-order Lax operator $\hat{L}$.[12,14,15] Besides the usual reduction described above, there are other reductions that have been discussed in Refs. 13 and 16. Since the first Hamiltonian structure can be obtained from the second Hamiltonian structure by replacing L by $L+\lambda$, where λ is called the spectral parameter, we shall focus only on the second structure.

III. TWO NONSTANDARD SUPER-LAX HIERARCHIES

There are several superintegrable hierarchies whose Lax operators are related to the modifications or reductions of the supersymmetric Lax operator (2.1) in the literature. Here, for our purpose, we consider the following two Lax systems:

$$\frac{dK_i}{dt_k}=[(K_i^{k/n})_{\geqslant 1},K_i] \quad (i=A,B), \tag{3.1}$$

with the Lax operators K_i defined by

$$K_A=D^{2n}+V_{2n-2}D^{2n-2}+\cdots+V_1D, \tag{3.2}$$

$$K_B=D^{2n}+V_{2n-2}D^{2n-2}+\cdots+V_0+D^{-1}V_{-1}. \tag{3.3}$$

The Lax equation for K_A is referred to the IK sKdV hierarchy.[7] The simplest example in this case is just the Laberge–Mathieu super KdV (LM sKdV) hierarchy $(n=2)$, which was constructed from a $N=2$sKdV hierarchy.[17] On the other hand, the Lax equation for K_B is the generalization of the super two-boson hierarchy (sTB) $(n=1)$,[18] which we call the cmsKP hierarchy. In particular, from (3.1) it is easy to show that the coefficient function V_{-1} obeys the evolution equation

$$\frac{dV_{-1}}{dt_k}=-((K_B^{k/n})^*_{\geqslant 1}V_{-1}), \tag{3.4}$$

which implies that V_{-1} is an adjoint eigenfunction associated with the Lax operator K_B.

In general, the second Poisson brackets associated with the Lax operators K_i can be written as

$$\{F,G\}^{(i)}(K_i)=(-1)^{|F|+|G|+1}\,\mathrm{Str}(\Omega^{(i)}(d_iF)d_iG), \tag{3.5}$$

where $d_iF\equiv d_{K_i}F$, and the Hamiltonian maps $\Omega^{(i)}$ are defined by

$$\Omega^{(A)}(d_AF)=(K_Ad_AF)_+K_A-K_A(d_AFK_A)_+ +[K_A,(d_AFK_A)_0] + (-1)^{|F|}\left[\int^x D\,\mathrm{sres}[d_AF,K_A],K_A\right]+(-1)^{|F|}K_AD^{-1}\,\mathrm{sres}[d_AF,K_A], \quad (3.6)$$

$$\Omega^{(B)}(d_BF)=(K_Bd_BF)_+K_B-K_B(d_BFK_B)_+ +[K_B,(K_Bd_BF)_0] + (-1)^{|F|}\left[K_B,\int^x D\,\mathrm{sres}[d_BF,K_B]\right]+(-1)^{|F|}D^{-1}\,\mathrm{sres}[d_BF,K_B]K_B. \quad (3.7)$$

Notice that the map $\Omega^{(A)}$, in operator form, is similar to but different from $\Omega^{(B)}$. Instead of giving $\Omega^{(i)}$ by other methods,[8–10,19] we will follow a constructive approach, analogous to that of the supersymmetric GD structure,[6] to verify the Hamiltonian maps $\Omega^{(i)}$ from a supersymmetric Miura transformation point of view. To show that the maps $\Omega^{(i)}$ are indeed Hamiltonian, we have to check that the Poisson brackets defined in (3.5) are antisymmetric and obey the super-Jacobi identity. For antisymmetry, by direct computation, it can be easily shown that

$$\{F,G\}^{(i)}=-(-1)^{|F||G|}\{G,F\}^{(i)}. \quad (3.8)$$

For the super-Jacobi identity, instead of direct computation, we rewrite the Lax operator K_i as

$$K_A=\hat{L}_AD, \quad K_B=D^{-1}\hat{L}_B, \quad (3.9)$$

where $\hat{L}_A$ and $\hat{L}_B$ are superdifferential operators with order $2n-1$ and $2n+1$, respectively. Furthermore, from the relation

$$\delta F=(-1)^{|F|+1}\,\mathrm{Str}(\delta K_id_iF)=(-1)^{|F|}\,\mathrm{Str}(\delta\hat{L}_i\hat{d}_iF), \quad (3.10)$$

where $\hat{d}_i\equiv d_{\hat{L}_i}$, we have

$$\hat{d}_AF=-Dd_AF, \quad \hat{d}_BF=(-1)^{|F|}d_BFD^{-1}. \quad (3.11)$$

Substituting (3.9) and (3.11) into (3.6) and (3.7), we find

$$\Omega^{(A)}(d_AF)=-J_c(\hat{d}_AF)D, \quad \Omega^{(B)}(d_BF)=(-1)^{|F|}D^{-1}J_c(\hat{d}_BF), \quad (3.12)$$

which imply that the Poisson brackets defined by K_i can be transformed to those defined by $\hat{L}_i$ as follows:

$$\{F,G\}^{(i)}(K_i)=\eta_i\{F,G\}_c(\hat{L}_i), \quad (3.13)$$

where $\eta_A=-1$ and $\eta_B=+1$. Hence, the super-Jacobi identity associated with the maps $\Omega^{(i)}$ is automatically satisfied due to the fact that the reduced supersymmetric GD brackets defined by $\hat{L}_i$ admit Miura representations (2.14).

Therefore the maps $\Omega^{(i)}$ provide the Hamiltonian formulation for the Lax equations (3.1):

$$\frac{dK_i}{dt_k}=\{H_k^{(i)},K_i\}^{(i)}=\Omega^{(i)}(d_iH_k^{(i)}), \quad (3.14)$$

where the Hamiltonian functionals $H_k^{(i)}$ are given by

$$H_k^{(i)}=-\frac{n}{k}\,\mathrm{Str}(K_i^{k/n}). \quad (3.15)$$

Notice that the relative signs in the Hamiltonian maps $\Omega^{(i)}$ are crucial. It is this choice so that $\Omega^{(i)}(d_i H_k^{(i)})$ are differential operators of order less than $2n-2$, and Eq. (3.14) makes sense.

Before ending this section, two remarks are in order. First, we note that both Piosson brackets defined by K_i, up to a sign, are mapped to the same reduced supersymmetric GD bracket defined by $\hat{L}_i$, which is different from the situation in the bosonic case, where type A is mapped to the *difference* of the second and the third GD structures,[20] whereas type B is the *sum* of the second and the third ones.[20,21] Second, both Lax operators K_A and K_B can be factorized into multiplicative forms, i.e.,

$$K_A=(D-\Phi_{2n-1})(D-\Phi_{2n-2})\cdots(D-\Phi_1)D,$$
$$K_B=D^{-1}(D-\Phi_{2n+1})(D-\Phi_{2n})\cdots(D-\Phi_1), \tag{3.16}$$

where the Miura fields Φ_i obey the Poisson brackets,

$$\{\Phi_j(X),\Phi_k(Y)\}^{(i)}=\eta_i[1+(-1)^j\delta_{jk}]D\delta(X-Y). \tag{3.17}$$

IV. GENERALIZED MR SKDV AND CONSTRAINED SKP HIERARCHIES

Having constructed the Hamiltonian structures of two nonstandard super-Lax hierarchies in the previous section, we are now ready to discuss gauge equivalences related to these two nonstandard hierarchies. Based on the fact that gauge transformations are canonical transformations, we can use them to obtain new integrable Hamiltonian systems from the known ones. In the following, we will show that the second Hamiltonian structures of the generalized MR sKdV and csKP hierarchies are just the ones that can be obtained in this way.

Let us perform the following gauge transformation to the Lax operators K_i:

$$\tilde{L}_i=T^{-1}K_iT \quad (i=A,B), \tag{4.1}$$

where the gauge operator T is defined by $T=\exp(-\int^x V_{2n-2}/n)$, and hence the next leading term of K_i can be gauged away. The resulting differential operators $\tilde{L}_i$ are thus given by

$$\tilde{L}_A=D^{2n}+U_{2n-3}D^{2n-3}+\cdots+U_0,$$
$$\tilde{L}_B=D^{2n}+U_{2n-3}D^{2n-3}+\cdots+U_0+\phi D^{-1}\psi, \tag{4.2}$$

where $\phi\equiv T^{-1}$ and $\psi\equiv V_{-1}T$. It can be proved that T^{-1} is an even eigenfunction associated with the operator $\tilde{L}_i$, i.e., $\partial T^{-1}/\partial t_k=((\tilde{L}_i^{k/n})_+T^{-1})_0$, and the nonstandard Lax equations in (3.1) are then transformed to the standard ones,

$$\frac{d\tilde{L}_i}{dt_k}=[(\tilde{L}_i^{k/n})_+,\tilde{L}_i]. \tag{4.3}$$

Therefore the gauge transformation (4.1) provides a connection between K_i and $\tilde{L}_i$ in the Lax formulation. For $\tilde{L}_A$, the Lax equation (4.3) gives the generalization of the MR sKdV hierarchy $(n=2)$, which was originally constructed from the MR sKP hierarchy by reduction.[11] On the other hand, the Lax equation (4.3) for $\tilde{L}_B$ describes the csKP hierarchy that contains the sAKNS hierarchy $(n=1)$[5,22] as the simplest example. It can be easily shown that the Lax equation (4.3) for $\tilde{L}_B$ is consistent with the following equations:

$$\frac{\partial\phi}{\partial t_k}=((\tilde{L}_B^{k/n})_+\phi)_0, \quad \frac{\partial\psi}{\partial t_k}=-((\tilde{L}_B^{k/n})_+^*\psi)_0, \tag{4.4}$$

and thus ϕ and ψ are an even eigenfunction and an odd adjoint eigenfunction of the csKP hierarchy, respectively.

Moreover, since the hierarchy flows associated with K_i have Hamiltonian descriptions, it is quite natural to ask whether we can use such gauge equivalence to obtain the second Hamiltonian structures of the generalized MR sKdV and csKP hierarchies. The answer is yes. To see this, consider an infinitesimal gauge transformation $K_i \to K_i + Q$, where Q is a homogeneous superdifferential operator of order, at most, $2n-2$. Then, in view of (4.1), we can read off the linearized map T' and its transposed map $T'^{\dagger}$ as

$$T': Q \to T^{-1} Q T + \frac{1}{n}\left[\int^x q_{2n-2}, \tilde{L}_i\right], \tag{4.5}$$

$$T'^{\dagger}: P \to T P T^{-1} + \frac{(-1)^{|P|+1}}{n}\int^x \mathrm{sres}[P, \tilde{L}_i], \tag{4.6}$$

where P is an arbitrary SPDO, $q_{2n-2} \equiv \mathrm{sres}(Q D^{-2n+1})$, and the adjoint of an operator R is defined by $\mathrm{Str}(PRQ) = (-1)^{|R||P|}\,\mathrm{Str}(R^{\dagger} P Q)$. Using T' and $T'^{\dagger}$, a straightforward but tedius calculation (see Appendix A) shows that

$$\begin{aligned} T'\Theta^{(i)} T'^{\dagger}(P) = {} & (\tilde{L}_i P)_+ \tilde{L}_i - \tilde{L}_i (P \tilde{L}_i)_+ + \frac{1}{n}\left[\int^x \mathrm{res}[P, \tilde{L}_i], \tilde{L}_i\right] + \frac{1}{n}\left[\left(\int^x \mathrm{sres}[P, \tilde{L}_i]\right) D, \tilde{L}_i\right] \\ & - \frac{2}{n^2}\left[\int^x \left(\left(\int^{x'} \mathrm{sres}[P, \tilde{L}_i]\right) U_{2n-3}\right), \tilde{L}_i\right] \equiv \Theta^{(i)}(P). \end{aligned} \tag{4.7}$$

That means the Hamiltonian maps $\Theta^{(A)}$ and $\Theta^{(B)}$, in terms of their own Lax operators, have the same form. Since $\Theta^{(i)}$ are canonical equivalent to the Hamiltonian map $\Omega^{(i)}$, the Poisson brackets defined by $\Theta^{(i)}$ are also antisymmetric and obey the super-Jacobi identity. As a result, $\Theta^{(A)}(\Theta^{(B)})$ can be defined as the Hamiltonian map of the generalized MR sKdV (csKP) hierarchy. A further consistent check shows that $\Theta^{(i)}$ map the Hamiltonian one-forms $\tilde{d}_i \tilde{H}_k^{(i)}$ to (pseudo-) superdifferential operators of order, at most, $2n-3$. Now we can write down the Hamiltonian flows associated with the Lax operators $\tilde{L}_i$ as

$$\frac{d\tilde{L}_i}{dt_k} = \{\tilde{H}_k^{(i)}, \tilde{L}_i\} = \Theta^{(i)}(\tilde{d}_i \tilde{H}_k^{(i)}), \tag{4.8}$$

where the Hamiltonian functionals, in view of (3.15) and (4.1), are defined by

$$\tilde{H}_k^{(i)} = -\frac{n}{k}\,\mathrm{Str}\,\tilde{L}_i^{k/n}. \tag{4.9}$$

From the Hamiltonian flows (4.8) we can read off the Poisson brackets for the coefficient functions of $\tilde{L}_i$.

In fact, for $\tilde{L}_B$, we can express the associated Poisson brackets for U_i, ϕ, and ψ more precisely. Let us rewrite $\tilde{L}_B = l + \phi D^{-1}\psi$ and denote $H = \int_B h$ as one of the Hamiltonian functionals $\tilde{H}_k^{(B)}$. Then the Hamiltonian one-form can be expressed as

$$\tilde{d}_B H = d_l H + X, \tag{4.10}$$

where X is a superdifferential operator and

$$d_l H = \sum_{k=0}^{2n-3} (-1)^k D^{-k-1} \frac{\delta h}{\delta U_k}. \tag{4.11}$$

Then, from the relation

$$\delta H=-\mathrm{Str}((\delta l+\delta\phi D^{-}\psi+\phi D^{-1}\delta\psi)(d_l H+X))=-\mathrm{Str}(\delta l d_l H)+\int_B\left(\delta\phi\frac{\delta h}{\delta\phi}+\delta\psi\frac{\delta h}{\delta\psi}\right), \tag{4.12}$$

we have the following identifications:

$$\frac{\delta h}{\delta\phi}=(X^*\psi)_0, \quad \frac{\delta h}{\delta\psi}=(X\phi)_0. \tag{4.13}$$

Inserting (4.10) with X satisfying (4.13) into the Hamiltonian map $\Theta^{(B)}$ gives

$$\begin{aligned}\frac{dl}{dt}=&(ld_lH)_+l-l(d_lHl)_++((ld_lH)_+\phi D^{-1}\psi)_+-(\phi D^{-1}\psi(d_lHl)_+)_++\left(l\frac{\delta h}{\delta\psi}D^{-1}\psi\right)_+\\&-\left(\phi D^{-1}\frac{\delta h}{\delta\phi}l\right)_++\frac{1}{n}\left[\int^x\mathrm{res}[\tilde d_BH,\tilde L_B],l\right]-\frac{2}{n}\phi\psi\int^x\mathrm{sres}[\tilde d_BH,\tilde L_B]\\&+\frac{1}{n}\left[\int^x\mathrm{sres}[\tilde d_BH,\tilde L_B],l\right]+\frac{2}{n^2}\left[\int^x\left(U_{2n-3}\int^{x'}\mathrm{sres}[\tilde d_BH,\tilde L_B]\right),l\right],\\\frac{d\phi}{dt}=&((ld_lH)_+\phi)_0+\left(l\frac{\delta h}{\delta\psi}\right)_0+\phi\left[\int^x\left(D\psi\frac{\delta h}{\delta\psi}\right)-\int^x\left(D\phi\frac{\delta h}{\delta\phi}\right)\right]+\frac{1}{n}\phi\int^x\mathrm{res}[\tilde d_BH,\tilde L_B]\\&-\frac{1}{n}(D\phi)\int^x\mathrm{sres}[\tilde d_BH,\tilde L_B]+\frac{2}{n^2}\phi\int^x\left(U_{2n-3}\int^{x'}\mathrm{sres}[\tilde d_BH,\tilde L_B]\right),\\\frac{d\psi}{dt}=&-((l^*(d_lH)^*)_+\psi)_0-\left(l^*\frac{\delta h}{\delta\phi}\right)_0+\psi\left[\int^x\left(D\phi\frac{\delta h}{\delta\phi}\right)-\int^x\left(D\psi\frac{\delta h}{\delta\psi}\right)\right]-\frac{1}{n}\psi\int^x\mathrm{res}[\tilde d_BH,\tilde L_B]\\&+\frac{1}{n}\left(D\psi\int^x\mathrm{sres}[\tilde d_BH,\tilde L_B]\right)-\frac{2}{n^2}\psi\int^x\left(U_{2n-3}\int^{x'}\mathrm{sres}[\tilde d_BH,\tilde L_B]\right),\end{aligned} \tag{4.14}$$

where

$$\begin{aligned}\mathrm{res}[\tilde d_BH,\tilde L_B]&=\mathrm{res}[d_lH,l]+(D\psi)\frac{\delta h}{\delta\psi}-\phi\left(D\frac{\delta h}{\delta\phi}\right)-\mathrm{sres}(d_lH\phi\psi)-\phi(D(d_lH)^*\psi),\\\mathrm{sres}[\tilde d_BH,\tilde L_B]&=\mathrm{sres}[d_lH,l]-\psi\frac{\delta h}{\delta\psi}+\phi\frac{\delta h}{\delta\phi}.\end{aligned} \tag{4.15}$$

Equation (4.14) can be regarded as the supersymmetric generalization of the second Hamiltonian structures of constrained KP hierarchy derived by Oevel and Strampp.[23]

V. EXAMPLES

In this section we work out a number of examples to illustrate the previous results explicitly. We write down the Poisson brackets for these systems according to the formulas given above and compare them with the known results.

A. Laberge–Mathieu super-KdV hierarchy

For $K_A=\partial^2+v_2\partial+v_1D$, the first equations in (3.1) are given by

$$\frac{d}{dt_0}\begin{pmatrix}v_1\\ v_2\end{pmatrix}=\begin{pmatrix}v_{1x}\\ v_{2x}\end{pmatrix},$$
$$\frac{d}{dt_1}\begin{pmatrix}v_1\\ v_2\end{pmatrix}=\frac{1}{4}\begin{pmatrix}(v_{1xx}+3v_1(Dv_1)-\frac{3}{2}v_1v_2^2-3v_1v_{2x})_x\\ (v_{2xx}-\frac{1}{2}v_2^3+3v_1(Dv_2))_x\end{pmatrix}, \tag{5.1}$$

which represents the first equations of the LM sKdV hierarchy. The Hamiltonian formulation for these equations is given by (3.14), where the second Poisson structure can be obtained by substituting $d_AH_k^{(A)}=-D^{-2}(\delta h_k^{(A)}/\delta v_1)+D^{-3}(\delta h_k^{(A)}/\delta v_2)$ into (3.6). We find

$$\frac{d}{dt_k}\begin{pmatrix}v_1\\ v_2\end{pmatrix}=\begin{pmatrix}-2v_1\partial-v_{1x} & -\partial^2-v_2\partial+v_1D-(Dv_1)\\ \partial^2-v_2\partial+v_1D-v_{2x} & -2D^3+(Dv_2)-2v_1\end{pmatrix}\begin{pmatrix}\dfrac{\delta h_k^{(A)}}{\delta v_1}\\ \dfrac{\delta h_k^{(A)}}{\delta v_2}\end{pmatrix}, \tag{5.2}$$

where the first Hamiltonian functionals are given by

$$H_0^{(A)}=-2\,\mathrm{Str}\,K_A^{1/2}=-\int_B v_1,$$
$$H_1^{(A)}=-\frac{2}{3}\mathrm{Str}\,K_A^{3/2}=-\frac{3}{8}\int_B\left[\frac{1}{2}v_1v_2^2+v_1v_{2x}-v_1(Dv_1)\right]. \tag{5.3}$$

To compare with the known result, we consider the change of variables as follows:

$$(v_1,v_2)\to(-(Du)-\tau,-2u), \tag{5.4}$$

then the Poisson structure in (5.2) becomes

$$\frac{1}{2}\begin{pmatrix}-D\partial+\tau & 2u\partial-(Du)D+2u_x\\ 2u\partial-(Du)D+u_x & -D\partial^2+3\tau\partial+(D\tau)D+2\tau_x\end{pmatrix}, \tag{5.5}$$

which is just the form presented in Ref. 24.

B. Super-two-boson hierarchy

For $K_B=\partial+v_0+D^{-1}v_{-1}$ the first Lax equations in (3.1) are given by

$$\frac{d}{dt_1}\begin{pmatrix}v_0\\ v_{-1}\end{pmatrix}=\begin{pmatrix}v_{0x}\\ v_{-1x}\end{pmatrix},$$
$$\frac{d}{dt_2}\begin{pmatrix}v_0\\ v_{-1}\end{pmatrix}=\begin{pmatrix}v_{0xx}+2(Dv_{-1})_x+(v_0^2)_x\\ -v_{-1xx}+2(v_0v_{-1})_x\end{pmatrix}, \tag{5.6}$$

which represents the first equations of the sTB hierarchy. The Hamiltonian description for these equations are given by (3.14), where the second Poisson structure can be obtained by substituting $d_BH_k^{(B)}=D^{-1}(\delta h_k^{(B)}/\delta v_0)+(\delta h_k^{(B)}/\delta v_{-1})$ into (3.7). It turns out that

$$\frac{d}{dt_k}\begin{pmatrix} v_0 \\ v_{-1} \end{pmatrix} = \begin{pmatrix} 2D^3+(Dv_0)+2v_{-1} & \partial^2+v_0\partial+v_{-1}D+v_{0x} \\ -\partial^2+v_0\partial+v_{-1}D-(Dv_{-1}) & 2v_{-1}\partial+v_{-1x} \end{pmatrix} \begin{pmatrix} \dfrac{\delta h_k^{(B)}}{\delta v_0} \\ \dfrac{\delta h_k^{(B)}}{\delta v_{-1}} \end{pmatrix}, \tag{5.7}$$

where the first Hamiltonian functionals are given by

$$H_1^{(B)} = -\operatorname{Str} K_B = -\int_B v_{-1},$$

$$H_2^{(B)} = -\frac{1}{2}\operatorname{Str} K_B^2 = \int_B v_0 v_{-1}. \tag{5.8}$$

Equation (5.7) provides the second Hamiltonian formulation of the sTB hierarchy.

If we make the following identification:

$$(v_0, v_{-1}) \to (-(DJ_0), J_1), \tag{5.9}$$

then the second Poisson structure in (5.7) becomes

$$\begin{pmatrix} 2D+2D^{-1}J_1D^{-1}-D^{-1}J_{0x}D^{-1} & -D^3+D(DJ_0)-D^{-1}J_1D \\ D^3+(DJ_0)D+DJ_1D^{-1} & J_1D^2+D^2J_1 \end{pmatrix}, \tag{5.10}$$

which is the form of the second Poisson structure discussed in Ref. 18.

C. Manin–Radul super-KdV hierarchy

For $\tilde{L}_A = \partial^2 - \varphi D + a$, the first Lax equations in (4.3) are given by

$$\frac{d}{dt_0}\begin{pmatrix} a \\ \varphi \end{pmatrix} = \begin{pmatrix} a_x \\ \varphi_x \end{pmatrix},$$

$$\frac{d}{dt_1}\begin{pmatrix} a \\ \varphi \end{pmatrix} = \frac{1}{4}\begin{pmatrix} \varphi_{xxx}-3(\varphi(D\varphi))_x+6(a\varphi)_x \\ a_{xxx}-3(\varphi(Da))_x+3(a^2)_x \end{pmatrix}, \tag{5.11}$$

which represents the first equations of the MP sKdV hierarchy. The Hamiltonian formulation of these equations are given by (4.8), in which the first Hamiltonian functionals are given by

$$\tilde{H}_0^{(A)} = -2\operatorname{Str}\tilde{L}_A^{1/2} = \int_B \varphi,$$

$$\tilde{H}_1^{(A)} = -\frac{2}{3}\operatorname{Str}\tilde{L}_A^{3/2} = -\frac{1}{4}\int_B [\varphi(D\varphi)-2\varphi a], \tag{5.12}$$

and the second Poisson structure can be obtained by substituting $\tilde{d}_A\tilde{H}_k^{(A)} = D^{-1}(\delta\tilde{h}_k^{(A)}/\delta a) + D^{-2}(\delta\tilde{h}_k^{(A)}/\delta\varphi)$ into (4.7). It turns out that

$$\frac{d}{dt_k}\begin{pmatrix} a \\ \varphi \end{pmatrix} = \begin{pmatrix} P_{aa} & P_{a\varphi} \\ P_{\varphi a} & P_{\varphi\varphi} \end{pmatrix} \begin{pmatrix} \dfrac{\delta\tilde{h}_k^{(A)}}{\delta a} \\ \dfrac{\delta\tilde{h}_k^{(A)}}{\delta\varphi} \end{pmatrix}, \tag{5.13}$$

where the second Poisson matrix is given by

$$
\begin{aligned}
P_{aa}=&\tfrac{1}{2}[D\partial^3-3\varphi\partial^2+4aD\partial+(2(Da)-3\varphi_x)\partial+2a_xD+3\varphi(D\varphi)+(D^3a)-4a\varphi-\varphi_{xx}\\
&+\varphi D^{-1}(Da)-(Da)D^{-1}\varphi-\varphi D^{-1}\varphi D^{-1}\varphi-\varphi D^{-1}\varphi_x+\varphi_x D^{-1}\varphi],\\
P_{\alpha\varphi}=&\tfrac{1}{2}[\partial^3-2\varphi D\partial+4a\partial-\varphi_xD+2a_x+\varphi D^{-1}(D\varphi)],\\
P_{\varphi a}=&\tfrac{1}{2}[\partial^3+2\varphi D\partial+(4a-2(D\varphi))\partial+\varphi_xD+2a_x-(D^3\varphi)+(D\varphi)D^{-1}\varphi],\\
P_{\varphi\varphi}=&\tfrac{1}{2}[4\varphi\partial+2\varphi_x].
\end{aligned}
\tag{5.14}
$$

Equation (5.13) provides the second Hamiltonian formulation of the MR sKdV hierarchy reported in Ref. 4.

Starting from the Lax operator $K_A=\partial^2+v_2\partial+v_1D$ associated with the LM sKdV hierarchy, one can perform the gauge transformation $T=\exp(-\int^x v_2/2)$[7] on the Lax operator K_A as follows:

$$
K_A\rightarrow\tilde{L}_A=e^{\int^x v_2/2}K_Ae^{-\int^x v_2/2}=\partial^2+v_1D-\left(\frac{v_2^2}{4}+\frac{v_{2x}}{2}+\frac{v_1(D^{-1}v_2)}{2}\right). \tag{5.15}
$$

Then the Lax operator $\tilde{L}_A=\partial^2-\phi D+a$ associated with the MR sKdV hierarchy is related to the Lax operator K_A as

$$
\phi=-v_1,\quad a=-\left(\frac{v_2^2}{4}+\frac{v_{2x}}{2}+\frac{v_1(D^{-1}v_2)}{2}\right), \tag{5.16}
$$

which provides the gauge equivalence between the LM sKdV hierarchy (5.1) and the MR sKdV hierarchy (5.11). Moreover, it has been shown[24] that the second Hamiltonian structure (5.5) of the LM sKdV hierarchy can be transformed to the second Hamiltonian structure (5.14) of the MR sKdV hierarchy via this gauge transformation.

D. Super-AKNS hierarchy

For $\tilde{L}_B=\partial+\phi D^{-1}\psi$, the first equations in (4.3) are given by

$$
\begin{aligned}
\frac{d}{dt_1}\begin{pmatrix}\phi\\ \psi\end{pmatrix}&=\begin{pmatrix}\phi_x\\ \psi_x\end{pmatrix},\\
\frac{d}{dt_2}\begin{pmatrix}\phi\\ \psi\end{pmatrix}&=\begin{pmatrix}\phi_{xx}+2\phi(D\phi\psi)\\ -\psi_{xx}-2\psi(D\phi\psi)\end{pmatrix},
\end{aligned}
\tag{5.17}
$$

which are the first equations in the sAKNS hierarchy. Hamiltonian formulations for these equations are given by (4.14), where the first Hamiltonian functions are given by

$$
\begin{aligned}
\tilde{H}_1^{(B)}&=-\mathrm{Str}\,\tilde{L}_B=\int_B\phi\psi,\\
\tilde{H}_2^{(B)}&=-\frac{1}{2}\mathrm{Str}\,\tilde{L}_B^2=\int_B\phi_x\psi.
\end{aligned}
\tag{5.18}
$$

From (4.14), the Hamiltonian flow can be expressed as

$$\frac{d}{dt_k}\begin{pmatrix}\phi\\ \psi\end{pmatrix}=\begin{pmatrix}P_{\phi\phi} & P_{\phi\psi}\\ P_{\psi\phi} & P_{\psi\psi}\end{pmatrix}\begin{pmatrix}\dfrac{\delta\tilde{h}_k^{(B)}}{\delta\phi}\\ \dfrac{\delta\tilde{h}_k^{(B)}}{\delta\psi}\end{pmatrix}, \tag{5.19}$$

where the Poisson brackets are given by

$$\begin{aligned}P_{\phi\phi}&=-\phi D^{-1}\phi-\phi D^{-2}\phi D-(D\phi)D^{-2}\phi-2\phi D^{-2}\phi\psi D^{-2}\phi,\\ P_{\phi\psi}&=D^2+\phi D^{-1}\psi+\phi D^{-2}(D\psi)+(D\phi)D^{-2}\psi+2\phi D^{-2}\phi\psi D^{-2}\psi,\\ P_{\psi\phi}&=D^2+\psi D^{-2}\phi D+(D\psi)D^{-2}\phi+2\psi D^{-2}\phi\psi D^{-2}\phi,\\ P_{\psi\psi}&=-(D\psi)D^{-2}\psi-\psi D^{-2}(D\psi)-2\psi D^{-2}\phi\psi D^{-2}\psi,\end{aligned} \tag{5.20}$$

which is just the second Poisson structure obtained in Ref. 5. Equation (5.19) provides the second Hamiltonian formulation of the sAKNS hierarchy.

Starting from the Lax operator $K_B=\partial+v_0+D^{-1}v_{-1}$ associated with the sTB hierarchy, one can perform the gauge transformation $T=\exp(-\int^x v_0)$[22,25] to the Lax operator K_B as follows:

$$K_B\to\tilde{L}_B=e^{\int^x v_0}K_B e^{-\int^x v_0}=\partial+e^{\int^x v_0}D^{-1}e^{-\int^x v_0}. \tag{5.21}$$

Then the Lax operator $\tilde{L}_B=\partial+\phi D^{-1}\psi$ associated with the sAKNS hierarchy is related to the Lax operator K_B as

$$\phi=e^{\int^x v_0},\quad a=v_{-1}e^{-\int^x v_0}, \tag{5.22}$$

which provides the gauge equivalence between the sTB hierarchy (5.6) and the sAKNS hierarchy (5.17). Moreover, it can be proved[25] that the second Hamiltonian structure (5.10) of the sTB hierarchy can be transformed to the second Hamiltonian structure (5.20) of the sAKNS hierarchy via this gauge transformation.

VI. CONCLUDING REMARKS

In this paper, we investigate the Hamiltonian structures associated with several supersymmetric extensions of the KdV hierarchy. Starting with the reduced super-GD bracket, the Hamiltonian structures of two nonstandard super-KdV hierarchies can be constructed via supersymmetric Miura transformations. We then perform a gauge transformation on these two nonstandard Lax hierarchies to obtain the Hamiltonian structures of the generalized MR sKdV hierarchy and constrained sKP hierarchy in a unified fashion. To compare the obtained Hamiltonian structures with the known results, we work out a few examples, including the LM sKdV, sTB, MR sKdV, and sAKNS hierarchies.

Our approach on the gauge transformation relies on the algebra of superpseudodifferential operators, which provides an effective method to achieve the goal. In fact, the gauge transformation (4.1) that maps $\Omega^{(i)}$ to $\Theta^{(i)}$ is by no means unique. There is another gauge transformation triggered by $S=D^{-1}T$[25,26] that also brings $\Omega^{(i)}$ to $\Theta^{(i)}$. Since the parity of S is odd, the gauge equivalence of the Hamiltonian maps given by (4.7) should be replaced by $S'\Omega^{(i)}S'^{\dagger}=-\Theta^{(i)}$, where the minus sign will be compensated by that induced from the transformation of the Hamiltonians such that the hierarchy flows (3.14) are transformed to (4.8).

Finally, we would like to comment briefly on the algebraic structures associated with the Poisson brackets defined by the Hamiltonian maps $\Omega^{(i)}$ and $\Theta^{(i)}$. As we shows in Eq. (3.13), the Poisson brackets defined by $\Omega^{(i)}$ are encoded by the Poisson bracket defined by J_c. However, it has been shown[12,15] that in the space of the supersymmetric Lax operator of odd order, the reduced supersymmetric GD bracket (2.12) defines an infinite series of classical $N=2W$ superalgebras, which contain $N=2$ super-Virasoro algebra as a subalgebra. Therefore, through the Miura transformation, the differential polynomials of the coefficient functions V_i of K_i can be identified as the $N=2$ supermultiplets, and Eq. (3.17) provides the free-field realizations of the corresponding W superalgebras. On the other hand, for the MR sKdV and csKP hierarchies, the Poisson algebras defined by $\Theta^{(i)}$ are not quite clear so far, even for the simplest cases. It seems not so obvious to construct the super-Virasoro generator by covariantizing the supersymmetric Lax operator $\widetilde{L}_i$ due to the fact that $U_{2n-1}=U_{2n-2}=0$. Therefore, to explore the algebraic structures associated with $\Theta^{(i)}$, the decompositions of coefficient functions U_i into primary fields remain to be worked out. Work in this direction is still in progress.

ACKNOWLEDGMENTS

We would like to thank W. J. Huang for a number of helpful discussions and M. C. Chang for reading the manuscript. This work is supported by the National Science Council of Taiwan under Grants No. NSC 88-2811-M-194-0003(MHT) and No. NSC 88-2112-M-009-008 (JCS).

APPENDIX: PROOF FOR (4.7)

To prove (4.7), let P be an arbitrary superpseudodifferential operator; then

$$T'\Omega^{(A)}T'^{\dagger}P=T'Q, \tag{A1}$$

where

$$\begin{aligned} Q\equiv\Omega^{(A)}T'^{\dagger}P=&(K_AT'^{\dagger}P)_+K_A-K_A(T'^{\dagger}PK_A)_++[K_A,(T'^{\dagger}PK_A)_0]\\ &+(-1)^{|P|}\left[\int^x D\,\mathrm{sres}[T'^{\dagger}P,K_A],K_A\right]\\ &+(-1)^{|P|}K_AD^{-1}\,\mathrm{sres}[T'^{\dagger}P,K_A]. \end{aligned}\tag{A2}$$

Using (4.6), each term in Q can be calculated as follows:

$$(1)=(TLPT^{-1})_+K_A+\frac{(-1)^{|P|+1}}{n}D\left(\int^x \mathrm{sres}[P,L]\right)K_A,$$

$$(2)=-K_A(TPLT^{-1})_++\frac{(-1)^{|P|}}{n}K_A\left(D\int^x \mathrm{sres}[P,L]\right)-\frac{1}{n}\left(\int^x \mathrm{sres}[P,L]\right)D,$$

$$(3)=[K_A,(TPLT^{-1})_0]+\frac{(-1)^{|P|+1}}{n}\left[K_A,\left(D\int^x \mathrm{sres}[P,L]\right)\right],$$

$$(4)=(5)=0,$$

which imply that

$$Q=(TLPT^{-1})_+K_A-K_A(TPLT^{-1})_++[K_A,(TPLT^{-1})_0]+\frac{1}{n}\left[\left(\int^x \mathrm{sres}[P,L]\right)D,K_A\right] \tag{A3}$$

and

$$\frac{1}{n}\int^x q_{2n-2} = \frac{1}{n}\int^x \mathrm{sres}(QD^{-2n+1})$$
$$= (TPLT^{-1})_0 + \frac{1}{n}\int^x \mathrm{res}(T[P,L]T^{-1})$$
$$+ \frac{1}{n}\int^x \left[\left(\int^{x'} \mathrm{sres}[P,L]\right)\frac{(DV_{2n-2})}{n}\right] - \frac{2}{n^2}\int^x\left[\left(\int^{x'} \mathrm{sres}[P,L]\right)V_{2n-3}\right]. \tag{A4}$$

Substituting (A3) and (A4) into (4.5), we obtain the desired result (4.7).

Since the proof for K_B is parallel to the above one, we hence omit it here.

[1] B. A. Kupershmidt, *Elements of Super Integrable Systems* (Kluwer, Dordrecht, 1987).
[2] S. Stanciu, ''Supersymmetric integrable hierarchies and string theory,'' (hepth/9403129).
[3] H. Aratyn *et al.*, *Supersymmetry and Integrable Models* (Springer-Verlag, Berlin, 1998).
[4] W. Oevel and Z. Popowicz, Commun. Math. Phys. **139**, 441 (1991).
[5] H. Aratyn and A. Das, Mod. Phys. Lett. A **13**, 1185 (1998).
[6] J. M. Figueroa-O'Farrill and E. Ramos, Phys. Lett. B **262**, 265 (1991).
[7] T. Inami and H. Kanno, Int. J. Mod. Phys. A **7**, 419 (1992).
[8] Q. P. Liu, Phys. Lett. A **235**, 335 (1997).
[9] F. Delduc and L. Gallot, Commun. Math. Phys. **190**, 395 (1997).
[10] F. Delduc and L. Gallot, ''Supersymmetric Drinfeld–Sokolov reduction,'' solv-int/9802013.
[11] Y. I. Manin and A. O. Radul, Commun. Math. Phys. **98**, 65 (1985).
[12] J. M. Figueroa-O'Farrill and E. Ramos, Nucl. Phys. B **368**, 361 (1992).
[13] J. M. Figueroa-O'Farrill and E. Ramos, Commun. Math. Phys. **145**, 43 (1992).
[14] K. Huitu and D. Nemeschansky, Mod. Phys. Lett. A **6**, 3179 (1991).
[15] W. J. Huang, J. Math. Phys. **35**, 2570 (1994).
[16] J. M. Figueroa-O'Farrill, J. Mas, and E. Ramos, Rev. Math. Phys. **3**, 479 (1991).
[17] C. Laberge and P. Mathieu, Phys. Lett. B **215**, 718 (1988).
[18] J. C. Brunelli and A. Das, Phys. Lett. B **337**, 303 (1994); *ibid.* **354**, 307 (1995); Int. J. Mod. Phys. A **10**, 4563 (1995).
[19] A. Das and S. Panda, Mod. Phys. Lett. A **11**, 723 (1996).
[20] Q. P. Liu, Lett. Math. Phys. **43**, 65 (1998).
[21] J. C. Shaw and M. H. Tu, J. Phys. A **30**, L725 (1997).
[22] H. Aratyn and C. Rasinariu, Phys. Lett. B **391**, 99 (1997).
[23] W. Oevel and W. Strampp, Commun. Math. Phys. **157**, 51 (1993).
[24] C. Morosi and L. Pizzocchero, J. Math. Phys. **35**, 2397 (1994).
[25] J. C. Shaw and M. H. Tu, J. Phys. A **31**, 6517 (1998).
[26] J. C. Shaw and M. H. Tu, J. Phys. A **31**, 4805 (1998).

Large amplitude gravitational waves

G. Alì
Institute for Applications of Mathematics, Consiglio Nazionale delle Ricerche, Napoli, Italy

John K. Hunter
Department of Mathematics and Institute of Theoretical Dynamics, University of California at Davis, Davis, California 95616

(Received 30 November 1998; accepted for publication 17 March 1999)

We derive an asymptotic solution of the Einstein field equations which describes the propagation of a thin, large amplitude gravitational wave into a curved space–time. The resulting equations have the same form as the colliding plane wave equations without one of the usual constraint equations. © *1999 American Institute of Physics.* [S0022-2488(99)04306-6]

I. INTRODUCTION

Gravitational wave propagation is one of the most important features of Einstein's general theory of relativity. The Einstein field equations are highly nonlinear, and a question of fundamental interest is how nonlinearity affects the propagation of gravitational waves.

Small amplitude gravitational waves are well described by the linearized Einstein equations. Large amplitude, unidirectional gravitational plane waves are described by the exact Brinkmann–Rosen solution of the vacuum Einstein equations.[1,2] Despite the nonlinearity of the Einstein equations, a unidirectional plane wave propagates into flat space–time without distortion, and there are no dynamic nonlinear effects. Thus, one of the simplest situations in which nonlinear effects are significant is when a large amplitude gravitational wave propagates into curved space–time.

If the space–time ahead of the wave is that of a counterpropagating gravitational plane wave, then the resulting space–time has a two-parameter family of spacelike isometries, and the metric is given by the exact colliding plane wave solution of the vacuum Einstein equations.[3–6] Exact solutions do not exist for more general space–time ahead of the wave.

In this paper, we derive an asymptotic solution of the Einstein equations that describes the propagation of a thin, large-amplitude, pulselike gravitational wave into a general curved space–time. The solution applies when the metric inside the wave varies much more rapidly than the metric on either side of the wave. As a result, the wave can be approximated locally by a nonlinear plane wave, which slowly distorts as it propagates into the curved space–time. For plane-polarized waves, the asymptotic solution is given by Eqs. (3.1), (3.6), (3.7), and (3.9)–(3.11) below. For nonpolarized waves, the asymptotic solution is given by (3.1), (6.3), (6.5), and (6.8)–(6.11). The asymptotic solution satisfies the colliding plane wave equations without one of the usual constraints. The colliding plane wave equations are therefore canonical equations for nonlinear gravitational waves which describe a much larger class of solutions than the ones with exact plane-wave symmetry.

The nonlinearity of the asymptotic equations may result in the development of a space–time singularity. The mechanism of singularity formation in gravitational waves is the mutual focusing of the gravitational wave and the curved space–time into which it propagates. This mechanism differs from the nonlinear steepening of waves in quasilinear hyperbolic systems that leads to the formation of shocks. A second effect of nonlinearity is the distortion of space–time by the passage of a nonplanar gravitational wave. Moreover, the wave generates a slowly varying, backscattered gravitational wave.

Isaacson [7] and Choquet-Bruhat[8] derived a short-wave asymptotic expansion of the Einstein equations that describes the propagation of a small amplitude, oscillatory gravitational wave through a slowly varying background space–time. The mean energy density of the gravitational wave curves the background space–time, and this curvature refracts the null geodesics along which the wave propagates. Some further developments of this work are described in Ref. 9. The short-wave expansion derived here applies to large amplitude, impulsive gravitational waves, and is completely new.

In Sec. II, we summarize the exact colliding plane wave solution of the Einstein equations. In Sec. III, we give an overview of the asymptotic expansion. In Sec. IV, we write out expansions of the metric components, the connection coefficients, and the Ricci curvature components. In Sec. V, we construct a coordinate system in which the metric adopts its simplest form. In Sec. VI, we complete the derivation of the asymptotic equations. In Sec. VII, we show that the same equations follow from an expansion of the variational principle for the Einstein equations. In Sec. VIII, we explain how to derive boundary conditions for the asymptotic equations, and in Sec. IX, we consider some specific physical examples.

II. COLLIDING PLANE WAVES

The vacuum Einstein field equations imply that

$$\mathbf{Ricci}=0, \tag{2.1}$$

where **Ricci** is the Ricci tensor associated with the metric tensor **g**. The plane-polarized, colliding plane wave solution of (2.1) is given by

$$\mathbf{g}=-2e^{-M}\,du\,dv+e^{-U}(e^{V}\,dy^2+e^{-V}\,dz^2), \tag{2.2}$$

where the functions $M(u,v),U(u,v),V(u,v)$ satisfy the colliding plane wave equations,

$$U_{uv}=U_uU_v\,, \tag{2.3}$$

$$V_{uv}=\tfrac{1}{2}(U_uV_v+U_vV_u), \tag{2.4}$$

$$M_{uv}=\tfrac{1}{2}(-U_uU_v+V_uV_v), \tag{2.5}$$

$$U_{uu}=\tfrac{1}{2}(U_u^2+V_u^2)-U_uM_u\,, \tag{2.6}$$

$$U_{vv}=\tfrac{1}{2}(U_v^2+V_v^2)-U_vM_v\,. \tag{2.7}$$

Equations (2.3)–(2.5) are wave equations for M,U,V in characteristic coordinates (u,v). Equations (2.6) and (2.7) are constraints which are preserved by (2.3)–(2.5). To specify a unique solution, the wave equations can be supplemented by characteristic initial data for M,U,V on the lines $u=0$ and $v=0$, which satisfy the appropriate constraint equation.

The metric which describes the collision of nonpolarized plane waves is

$$\mathbf{g}=-2e^{-M}\,du\,dv+e^{-U}(e^{V}\cosh W\,dy^2-2\sinh W\,dy\,dz+e^{-V}\cosh W\,dz^2),$$

where the functions $M(u,v),U(u,v),V(u,v),W(u,v)$ satisfy

$$U_{uv}=U_uU_v\,, \tag{2.8}$$

$$V_{uv}=\tfrac{1}{2}(U_uV_v+U_vV_u)-(V_uW_v+V_vW_u)\tanh W, \tag{2.9}$$

$$W_{uv}=\tfrac{1}{2}(U_uW_v+U_vW_u)+V_uV_v\sinh W\cosh W, \tag{2.10}$$

$$M_{uv}=\tfrac{1}{2}(-U_uU_v+V_uV_v\cosh^2 W+W_uW_v), \tag{2.11}$$

$$U_{uu}=\tfrac{1}{2}(U_u^2+V_u^2\cosh^2 W+W_u^2)-U_uM_u, \tag{2.12}$$

$$U_{vv}=\tfrac{1}{2}(U_v^2+V_v^2\cosh^2 W+W_v^2)-U_vM_v. \tag{2.13}$$

When $W=0$, this solution reduces to the plane-polarized solution. When all functions are independent of v, the solution reduces to the exact Rosen solution for a unidirectional plane wave.

III. OVERVIEW OF THE EXPANSION

In this section, we outline the main ideas in the derivation of the asymptotic solution. For simplicity, we describe the case of plane-polarized waves. The algebraic details are given in the following sections.

We consider metrics of the form

$$\begin{gathered}\mathbf{g}=\mathbf{g}\left(\frac{u(x)}{\epsilon},x;\epsilon\right),\\ \mathbf{g}(\theta,x;\epsilon)=\overset{0}{\mathbf{g}}(\theta,x)+\epsilon\overset{1}{\mathbf{g}}(\theta,x)+O(\epsilon^2),\end{gathered} \tag{3.1}$$

where ϵ is a small parameter, and u is a scalar-valued phase function with $du\neq 0$. This ansatz corresponds to a metric that varies rapidly and strongly in the u direction. The phase u is a null function of the metric, at least up to the order ϵ. That is, it satisfies

$$\mathbf{g}^\sharp(du,du)=O(\epsilon^2), \tag{3.2}$$

where $\mathbf{g}^\sharp$ is the contravariant form of the metric tensor. The component form of this equation is written out in (4.5) below. The scaled variable

$$\theta=\frac{u}{\epsilon} \tag{3.3}$$

is a ''stretched'' coordinate inside the wave. We assume that the derivatives of $\mathbf{g}(\theta,x;\epsilon)$ with respect to θ decay to zero sufficiently quickly as $\theta\to\infty$. Thus, the solution (3.1) describes a thin, pulselike gravitational wave which is located near the null surface $u=0$. For example, if the metric is independent of θ when $|\theta|$ is sufficiently large, then the solution describes a thin ''sandwich'' wave that separates slowly varying metrics on either side.

The Ricci tensor associated with the metric (3.1) has an expansion of the form

$$\mathbf{Ricci}=\frac{1}{\epsilon^2}\overset{-2}{\mathbf{Ricci}}+\frac{1}{\epsilon}\overset{-1}{\mathbf{Ricci}}+O(1). \tag{3.4}$$

At the order ϵ^{-2}, the Einstein equations (2.1) imply that

$$\overset{-2}{\mathbf{Ricci}}=0.$$

This equation is a nonlinear, second-order ordinary differential equation in ∂_θ for the leading order term of the metric in which the ''slow'' variables x occur as parameters. We write it symbolically as

$$N(\partial_\theta^2)[\overset{0}{\mathbf{g}}]=0. \tag{3.5}$$

In suitable coordinates (u,v,y,z), a solution of this equation is the plane-polarized plane wave metric

$$\overset{0}{\mathbf{g}}=-2e^{-M}\,du\,dv+e^{-U}(e^{V}\,dy^2+e^{-V}\,dz^2), \tag{3.6}$$

where M,U,V are functions of (θ,v,y,z). For a metric of the form (3.6), Eq. (3.5) reduces to the θ-constraint equation,

$$U_{\theta\theta}=\tfrac{1}{2}(U_\theta^2+V_\theta^2)-U_\theta M_\theta. \tag{3.7}$$

At the order ϵ^{-1}, the Einstein equations imply that

$$\overset{-1}{\mathbf{Ricci}}=0.$$

This is a linear equation for $\overset{1}{\mathbf{g}}$ of the form

$$L(\partial_\theta^2)\left[\overset{1}{\mathbf{g}}\right]=F(\partial_\theta,\partial_v,\partial_y,\partial_z)\left[\overset{0}{\mathbf{g}}\right], \tag{3.8}$$

where L is a second-order linear ordinary differential operator in ∂_θ acting on $\overset{1}{\mathbf{g}}$, whose coefficients depend on $\overset{0}{\mathbf{g}}$, and F is a nonlinear partial differential operator acting on $\overset{0}{\mathbf{g}}$. The equations in (3.8) are not independent. The requirement that (3.8) can be solved for $\overset{1}{\mathbf{g}}$ implies that M, U, and V satisfy the equations

$$U_{\theta v}=U_\theta U_v, \tag{3.9}$$

$$V_{\theta v}=\tfrac{1}{2}(U_\theta V_v+U_v V_\theta), \tag{3.10}$$

$$M_{\theta v}=\tfrac{1}{2}(-U_\theta U_v+V_\theta V_v). \tag{3.11}$$

Equations (3.9)–(3.11) are identical to the evolution equations (2.8)–(2.10) for the exact colliding plane wave solution, with $\theta=u/\epsilon$. The leading order solution (3.6) satisfies the constraint equation (3.7) in the ‘‘fast’’ phase variable θ, but need not satisfy the constraint equation (2.7) in the ‘‘slow’’ variable v. If the v-constraint equation does not hold, then the asymptotic expansion of the metric contains higher order terms which are absent from the exact colliding plane wave solution.

Equation (3.6) implies that $\partial_v=-e^M\mathbf{g}^\sharp\cdot du$. Thus, ∂_v is a vector on the light cone which is tangent to the null surface $u=0$, and the ‘‘slow’’ derivative with respect to v which appears in (3.9)–(3.11) is a derivative along the bicharacteristic null geodesics associated with u. The transverse variables y and z occur as parameters. Therefore, in the short-wave limit considered here, the $(1+3)$-dimensional field equations reduce to $(1+1)$-dimensional asymptotic equations along the set of null geodesics associated with the phase u. The parametric dependence of the solution on y and z allows the pulse to be compactly supported in the transverse directions, so the wave need not have infinite extent. Moreover, the asymptotic solution need not have any special exact symmetries.

The asymptotic equations for nonpolarized gravitational waves are obtained in a similar way. They consist of the general colliding plane wave equations (2.8)–(2.12) with u replaced by θ. The v-constraint equation (2.13) is not required to hold.

Since the asymptotic equations follow from the order ϵ^{-2} and order ϵ^{-1} components of the field equations, the asymptotic solution remains valid in the presence of matter with a slowly varying, order one energy–momentum tensor, $\mathbf{T}=\mathbf{T}(x)$.

One subtle point in carrying out the expansion concerns the choice of the phase function u. In order for (3.5) to have a nontrivial solution, the phase u must be a null function of the leading order metric, but u need not be a null function of the entire metric. However, it follows from the analysis in Sec. V that we can use a transformation of the form

$$u \to \epsilon \Psi\left(\frac{u}{\epsilon}, x; \epsilon\right) \tag{3.12}$$

to choose a phase which satisfies (3.2). The asymptotic solutions obtained with the use of the old and the new phases can be shown to be equivalent. When the phase satisfies (3.2), variations in the metric propagate along the null geodesics associated with the phase, and the asymptotic equations adopt their simplest form.

IV. EXPANSION OF THE METRIC AND THE CURVATURE

In this section, we write out expansions of the metric components, the connection coefficients, and the Ricci curvature components.

We use local coordinates x^α in which

$$\mathbf{g} = g_{\alpha\beta}\, dx^\alpha\, dx^\beta. \tag{4.1}$$

Here and below, Greek indices $\alpha,\beta,\mu,\nu,\ldots$ take on the values 0,1,2,3. We look for an expansion of the metric components as $\epsilon \to 0$ of the form

$$\begin{gathered} g_{\alpha\beta} = g_{\alpha\beta}\left(\frac{u(x)}{\epsilon}, x; \epsilon\right), \\ g_{\alpha\beta}(\theta,x;\epsilon) = \overset{0}{g}_{\alpha\beta}(\theta,x) + \epsilon \overset{1}{g}_{\alpha\beta}(\theta,x) + O(\epsilon^2). \end{gathered} \tag{4.2}$$

The contravariant metric components $g^{\alpha\beta}$ satisfy

$$g^{\alpha\mu} g_{\mu\beta} = \delta^\alpha_\beta.$$

Expansion of this equation in a power series in ϵ gives

$$g^{\alpha\beta} = \overset{0}{g}{}^{\alpha\beta} - \epsilon \overset{1}{g}{}^{\alpha\beta} + O(\epsilon^2). \tag{4.3}$$

In (4.3), $\overset{0}{g}{}^{\alpha\beta}$ is the inverse of $\overset{0}{g}_{\alpha\beta}$, and we use the leading order metric components to raise indices, so that

$$\overset{1}{g}{}^{\alpha\beta} = \overset{0}{g}{}^{\alpha\mu} \overset{0}{g}{}^{\beta\nu} \overset{1}{g}_{\mu\nu}. \tag{4.4}$$

With this notation, the order ϵ term in the expansion of the contravariant metric component $g^{\alpha\beta}$ is $-\overset{1}{g}{}^{\alpha\beta}$, not $\overset{1}{g}{}^{\alpha\beta}$.

In terms of the metric components, we have

$$\mathbf{g}^\sharp(du,du) = g^{\alpha\beta} \frac{\partial u}{\partial x^\alpha} \frac{\partial u}{\partial x^\beta} = \overset{0}{g}{}^{\alpha\beta} \frac{\partial u}{\partial x^\alpha} \frac{\partial u}{\partial x^\beta} - \epsilon \overset{1}{g}{}^{\alpha\beta} \frac{\partial u}{\partial x^\alpha} \frac{\partial u}{\partial x^\beta} + O(\epsilon^2). \tag{4.5}$$

Thus, the null condition (3.2) holds provided that

$$\overset{0}{g}{}^{\alpha\beta} \frac{\partial u}{\partial x^\alpha} \frac{\partial u}{\partial x^\beta} = 0, \quad \overset{1}{g}{}^{\alpha\beta} \frac{\partial u}{\partial x^\alpha} \frac{\partial u}{\partial x^\beta} = 0. \tag{4.6}$$

The first condition in (4.6) states that u is a null function of $\overset{0}{\mathbf{g}}$. The second condition is required in order for the phase to be a null function of the perturbed metric up to the order ϵ.

The Ricci tensor components $R_{\alpha\beta}$ are given by

$$R_{\alpha\beta}=\frac{\partial\Gamma^{\lambda}{}_{\alpha\beta}}{\partial x^{\lambda}}-\frac{\partial\Gamma^{\lambda}{}_{\beta\lambda}}{\partial x^{\alpha}}+\Gamma^{\lambda}{}_{\alpha\beta}\Gamma^{\mu}{}_{\lambda\mu}-\Gamma^{\mu}{}_{\alpha\lambda}\Gamma^{\lambda}{}_{\beta\mu}, \tag{4.7}$$

where $\Gamma^{\lambda}{}_{\alpha\beta}$ are the connection coefficients

$$\Gamma^{\lambda}{}_{\alpha\beta}=\frac{1}{2}g^{\lambda\mu}\left(\frac{\partial g_{\beta\mu}}{\partial x^{\alpha}}+\frac{\partial g_{\alpha\mu}}{\partial x^{\beta}}-\frac{\partial g_{\alpha\beta}}{\partial x^{\mu}}\right). \tag{4.8}$$

From (3.3), the derivative of a function $f_{\alpha\beta}(\theta,x)$, with respect to x^{μ} is given by

$$\frac{\partial f_{\alpha\beta}}{\partial x^{\mu}}=\frac{1}{\epsilon}f_{\alpha\beta,\theta}u_{\mu}+f_{\alpha\beta,\mu}, \tag{4.9}$$

where

$$u_{\mu}=\frac{\partial u}{\partial x^{\mu}},\quad f_{\alpha\beta,\theta}=\left.\frac{\partial f_{\alpha\beta}}{\partial\theta}\right|_{x},\quad f_{\alpha\beta,\mu}=\left.\frac{\partial f_{\alpha\beta}}{\partial x^{\mu}}\right|_{\theta}.$$

We use (4.2), (4.3), and (4.9) in (4.7) and (4.8) and expand the result with respect to ϵ. After some algebra, we find that

$$\begin{aligned}\Gamma^{\lambda}{}_{\alpha\beta}&=\frac{1}{\epsilon}\overset{-1}{\Gamma}{}^{\lambda}{}_{\alpha\beta}+\overset{0}{\Gamma}{}^{\lambda}{}_{\alpha\beta}+O(\epsilon),\\ R_{\alpha\beta}&=\frac{1}{\epsilon^{2}}\overset{-2}{R}{}_{\alpha\beta}+\frac{1}{\epsilon}\overset{-1}{R}{}_{\alpha\beta}+O(1),\end{aligned} \tag{4.10}$$

where

$$\begin{aligned}\overset{-1}{\Gamma}{}^{\lambda}{}_{\alpha\beta}&=\tfrac{1}{2}\overset{0}{g}{}^{\lambda\mu}(\overset{0}{g}{}_{\beta\mu,\theta}u_{\alpha}+\overset{0}{g}{}_{\alpha\mu,\theta}u_{\beta}-\overset{0}{g}{}_{\alpha\beta,\theta}u_{\mu}),\\ \overset{0}{\Gamma}{}^{\lambda}{}_{\alpha\beta}&=\tfrac{1}{2}\overset{0}{g}{}^{\lambda\mu}(\overset{0}{g}{}_{\beta\mu,\alpha}+\overset{0}{g}{}_{\alpha\mu,\beta}-\overset{0}{g}{}_{\alpha\beta,\mu})+\tfrac{1}{2}\overset{0}{g}{}^{\lambda\mu}(\overset{1}{g}{}_{\beta\mu,\theta}u_{\alpha}+\overset{1}{g}{}_{\alpha\mu,\theta}u_{\beta}-\overset{1}{g}{}_{\alpha,\beta,\theta}u_{\mu})\\ &\quad-\tfrac{1}{2}\overset{1}{g}{}^{\lambda\mu}(\overset{0}{g}{}_{\beta\mu,\theta}u_{\alpha}+\overset{0}{g}{}_{\alpha\mu,\theta}u_{\beta}-\overset{0}{g}{}_{\alpha\beta,\theta}u_{\mu}),\\ \overset{-2}{R}{}_{\alpha\beta}&=\overset{-1}{\Gamma}{}^{\mu}{}_{\alpha\beta,\theta}u_{\mu}-\overset{-1}{\Gamma}{}^{\mu}{}_{\beta\mu,\theta}u_{\alpha}+\overset{-1}{\Gamma}{}^{\mu}{}_{\alpha\beta}\overset{-1}{\Gamma}{}^{\nu}{}_{\mu\nu}-\overset{-1}{\Gamma}{}^{\mu}{}_{\alpha\nu}\overset{-1}{\Gamma}{}^{\nu}{}_{\beta\mu},\\ \overset{-1}{R}{}_{\alpha\beta}&=\overset{-1}{\Gamma}{}^{\mu}{}_{\alpha\beta,\mu}-\overset{-1}{\Gamma}{}^{\mu}{}_{\beta\mu,\alpha}+\overset{0}{\Gamma}{}^{\mu}{}_{\alpha\beta,\theta}u_{\mu}-\overset{0}{\Gamma}{}^{\mu}{}_{\beta\mu,\theta}u_{\alpha}+\overset{-1}{\Gamma}{}^{\mu}{}_{\alpha\beta}\overset{0}{\Gamma}{}^{\nu}{}_{\mu\nu}\\ &\quad+\overset{0}{\Gamma}{}^{\mu}{}_{\alpha\beta}\overset{-1}{\Gamma}{}^{\nu}{}_{\mu\nu}-\overset{-1}{\Gamma}{}^{\mu}{}_{\alpha\nu}\overset{0}{\Gamma}{}^{\nu}{}_{\beta\mu}-\overset{0}{\Gamma}{}^{\mu}{}_{\alpha\nu}\overset{-1}{\Gamma}{}^{\nu}{}_{\beta\mu}.\end{aligned} \tag{4.11}$$

The component form of the field equations (2.1) is

$$R_{\alpha\beta}=0. \tag{4.12}$$

Using (4.10) in (4.12) and equating coefficients of ϵ^{-2} and ϵ^{-1} to zero, we get that

$$\overset{-2}{R}{}_{\alpha\beta}=0, \tag{4.13}$$

$$\overset{-1}{R}{}_{\alpha\beta}=0. \tag{4.14}$$

In order to solve these equations, we first use a coordinate transformation to simplify the form of the metric.

V. COORDINATE TRANSFORMATIONS

In this section, we show that there is a choice of a local coordinate system x^α in which $u = x^0$ and the metric has the form

$$\mathbf{g} = 2\overset{0}{g}_{01}\, dx^0\, dx^1 + \overset{0}{g}_{ab}\, dx^a\, dx^b + \epsilon\{2\overset{1}{g}_{1a}\, dx^1\, dx^a + \overset{1}{g}_{ab}\, dx^a\, dx^b\} + O(\epsilon^2). \tag{5.1}$$

Here and below, indices $a,b,c,\ldots$ take on the values 2,3, while indices $i,j,k,\ldots$ take on the values 1,2,3.

The corresponding expansion of the contravariant form of the metric tensor is

$$\mathbf{g}^\sharp = 2\overset{0}{g}{}^{01}\partial_0\partial_1 + \overset{0}{g}{}^{ab}\partial_a\partial_b - \epsilon\{2\overset{0}{g}{}^{01}\overset{0}{g}{}^{ab}\overset{1}{g}_{1b}\partial_0\partial_a + \overset{0}{g}{}^{ac}\overset{0}{g}{}^{bd}\overset{1}{g}_{cd}\partial_a\partial_b\} + O(\epsilon^2). \tag{5.2}$$

For this metric, we have

$$\overset{0}{g}{}^{00} = 0, \quad \overset{1}{g}{}^{00} = 0. \tag{5.3}$$

Thus, the phase $u = x^0$ satisfies (4.6), and hence (3.2).

The most general coordinate transformation which is compatible with an expansion of the form (4.2) is

$$\frac{x^0}{\epsilon} \to \overset{1}{\Psi}{}^0\left(\frac{x^0}{\epsilon}, x\right) + \epsilon\overset{2}{\Psi}{}^0\left(\frac{x^0}{\epsilon}, x\right) + O(\epsilon^2), \tag{5.4}$$

$$x^i \to \overset{0}{\Psi}{}^i(x) + \epsilon\overset{1}{\Psi}{}^i\left(\frac{x^0}{\epsilon}, x\right) + \epsilon^2\overset{2}{\Psi}{}^i\left(\frac{x^0}{\epsilon}, x\right) + O(\epsilon^3). \tag{5.5}$$

Here, we suppose that the phase is given by $u = x^0$ in both the old and the new coordinates. Thus, the change of coordinates (5.4) implies a change in the phase of the form (3.12).

First, we simplify the leading order metric components by means of a transformation

$$x^0 \to x^0, \quad x^i \to x^i + \epsilon\overset{1}{\Psi}{}^i\left(\frac{x^0}{\epsilon}, x\right). \tag{5.6}$$

Expansion of the transformation law for the change in covariant tensor components implies that the leading order metric components transform under (5.6) according to

$$\overset{0}{g}_{00} \to \overset{0}{g}_{00} + 2\overset{1}{\Psi}{}^k_{,\theta}\overset{0}{g}_{0k} + \overset{1}{\Psi}{}^k_{,\theta}\overset{1}{\Psi}{}^l_{,\theta}\overset{0}{g}_{kl}, \tag{5.7}$$

$$\overset{0}{g}_{0i} \to \overset{0}{g}_{0i} + \overset{1}{\Psi}{}^k_{,\theta}\overset{0}{g}_{ki}, \tag{5.8}$$

$$\overset{0}{g}_{ij} \to \overset{0}{g}_{ij}. \tag{5.9}$$

If the matrix $\overset{0}{g}_{ij}$ is nonsingular, then (5.8) implies that we can transform $\overset{0}{g}_{0i}$ to zero. This contradicts the requirement that x^0 is null (cf. Ref. 2, Sec. 109). Hence, we must have

$$\det \overset{0}{g}_{ij} = 0. \tag{5.10}$$

By an appropriate renumbering of the i coordinates, we can suppose without loss of generality that

$$\det \overset{0}{g}_{ab} \neq 0. \tag{5.11}$$

From (5.7) and (5.8), we can then choose the transformation (5.6) so that

$$\overset{0}{g}_{00} = \overset{0}{g}_{02} = \overset{0}{g}_{03} = 0. \tag{5.12}$$

Solving Eq. (5.10) for $\overset{0}{g}_{11}$, we get

$$\overset{0}{g}_{11} = \overset{0}{g}{}^{ab} \overset{0}{g}_{1a} \overset{0}{g}_{1b}, \tag{5.13}$$

where $\overset{0}{g}{}^{ab}$ is the inverse of $\overset{0}{g}_{ab}$. We define

$$g^a = \overset{0}{g}{}^{ab} \overset{0}{g}_{1b}. \tag{5.14}$$

From (5.13) and (5.14), it follows that

$$\overset{0}{g}_{11} = \overset{0}{g}_{cd} g^c g^d, \quad \overset{0}{g}_{1a} = \overset{0}{g}_{ac} g^c. \tag{5.15}$$

Using (5.12)–(5.15) in (4.1), we find that, in the transformed coordinate system, the metric is

$$\mathbf{g} = 2\overset{0}{g}_{01}\, dx^0\, dx^1 + \overset{0}{g}_{ab} (dx^a + g^a\, dx^1)(dx^b + g^b\, dx^1) + O(\epsilon). \tag{5.16}$$

From (4.13), the metric (5.16) satisfies the condition

$$\overset{-2}{R}_{ab} = 0. \tag{5.17}$$

Using (5.16) in (4.11), we find that

$$\overset{-2}{R}_{ab} = -\tfrac{1}{2} (\overset{0}{g}{}^{01})^2 \overset{0}{g}_{ac} g^c_{,\theta} \overset{0}{g}_{bd} g^d_{,\theta}. \tag{5.18}$$

Equations (5.17) and (5.18) imply that

$$g^a_{,\theta} = 0,$$

so g^a is independent of θ. This fact allows us to remove g^a by a transformation

$$x^a \to \Psi^a(x^1, x^c). \tag{5.19}$$

The form of the metric (5.16) is unchanged by (5.19), and

$$\begin{gathered} \overset{0}{g}_{ab} \to \Psi^c_{,a} \Psi^d_{,b} \overset{0}{g}_{cd}, \\ g^a \to (A^{-1})^a_c (g^c + \Psi^c_{,1}), \end{gathered} \tag{5.20}$$

where $(A^a_c) = (\Psi^a_{,c})$. From (5.20), we can set $g^a = 0$. The metric (5.16) then reduces to

$$\mathbf{g} = 2\overset{0}{g}_{01}\, dx^0\, dx^1 + \overset{0}{g}_{ab}\, dx^a\, dx^b + O(\epsilon). \tag{5.21}$$

Next, we simplify the form of $\overset{1}{\mathbf{g}}$. We consider the transformation of coordinates

$$x^0 \to \epsilon \overset{1}{\Psi}{}^0\left(\frac{x^0}{\epsilon}, x\right) + \epsilon^2 \overset{2}{\Psi}{}^0\left(\frac{x^0}{\epsilon}, x\right), \quad x^i \to x^i + \epsilon^2 \overset{2}{\Psi}{}^i\left(\frac{x^0}{\epsilon}, x\right). \tag{5.22}$$

Under the action of (5.22), the form (5.21) of the metric is unchanged at order zero and

$$\overset{0}{g}_{01}\to\overset{1}{\Psi}{}^0_{,\theta}\overset{0}{g}_{01}.$$

At order one, the components transform according to

$$\overset{1}{g}_{00}\to\overset{1}{\Psi}{}^0_{,\theta}(\overset{1}{\Psi}{}^0_{,\theta}\overset{1}{g}_{00}+2\overset{2}{\Psi}{}^1_{,\theta}\overset{0}{g}_{01}),$$

$$\overset{1}{g}_{01}\to\overset{1}{\Psi}{}^0_{,\theta}\overset{1}{g}_{01}+(\overset{1}{\Psi}{}^0_{,0}+\overset{2}{\Psi}{}^0_{,\theta})\overset{0}{g}_{01},$$

$$\overset{1}{g}_{0a}\to\overset{1}{\Psi}{}^0_{,\theta}\overset{1}{g}_{0a}+\overset{2}{\Psi}{}^b_{,\theta}\overset{0}{g}_{ab},$$

$$\overset{1}{g}_{11}\to\overset{1}{g}_{11}+2\overset{1}{\Psi}{}^0_{,1}\overset{0}{g}_{01},$$

$$\overset{1}{g}_{1a}\to\overset{1}{g}_{1a}+\overset{1}{\Psi}{}^0_{,a}\overset{0}{g}_{01},$$

$$\overset{1}{g}_{ab}\to\overset{1}{g}_{ab}.$$

These transformations can be used to make

$$\overset{1}{g}_{11}=\overset{1}{g}_{0\alpha}=0. \tag{5.23}$$

The resulting metric then has the form given in (5.1).

Use of (5.1) and (5.2) in (4.11) implies that the nonzero connection coefficients at the orders ϵ^{-1} and ϵ^0 are

$$\overset{-1}{\Gamma}{}^0{}_{00}=\overset{0}{g}{}^{01}\overset{0}{g}_{01,\theta},\quad \overset{-1}{\Gamma}{}^1{}_{ab}=-\tfrac12\overset{0}{g}{}^{01}\overset{0}{g}_{ab,\theta},\quad \overset{-1}{\Gamma}{}^a{}_{0b}=\tfrac12\overset{0}{g}{}^{ac}\overset{0}{g}_{bc,\theta},$$

$$\overset{0}{\Gamma}{}^0{}_{00}=\overset{0}{g}{}^{01}\overset{0}{g}_{01,0},\quad \overset{0}{\Gamma}{}^0{}_{0a}=\tfrac12\overset{0}{g}{}^{01}(\overset{0}{g}_{01,a}+\overset{1}{g}_{1a,\theta})-\tfrac12\overset{1}{g}{}^{0b}\overset{0}{g}_{ab,\theta},$$

$$\overset{0}{\Gamma}{}^0{}_{ab}=-\tfrac12\overset{0}{g}{}^{01}\overset{0}{g}_{ab,1},\quad \overset{0}{\Gamma}{}^1{}_{11}=\overset{0}{g}{}^{01}\overset{0}{g}_{01,1},\quad \overset{0}{\Gamma}{}^1{}_{1a}=\tfrac12\overset{0}{g}{}^{01}(\overset{0}{g}_{01,a}-\overset{1}{g}_{1a,\theta}),$$

$$\overset{0}{\Gamma}{}^1{}_{ab}=-\tfrac12\overset{0}{g}{}^{01}(\overset{0}{g}_{ab,0}+\overset{1}{g}_{ab,\theta}),\quad \overset{0}{\Gamma}{}^a{}_{01}=-\tfrac12\overset{0}{g}{}^{ac}(\overset{0}{g}_{01,c}-\overset{1}{g}_{1c,\theta}),$$

$$\overset{0}{\Gamma}{}^a{}_{0b}=\tfrac12\overset{0}{g}{}^{ac}(\overset{0}{g}_{bc,0}+\overset{1}{g}_{bc,\theta})-\tfrac12\overset{1}{g}{}^{ac}\overset{0}{g}_{bc,\theta},$$

$$\overset{0}{\Gamma}{}^a{}_{1b}=\tfrac12\overset{0}{g}{}^{ac}\overset{0}{g}_{bc,1},\quad \overset{0}{\Gamma}{}^a{}_{bc}=\tfrac12\overset{0}{g}{}^{ad}(\overset{0}{g}_{bd,c}+\overset{0}{g}_{cd,b}-\overset{0}{g}_{bc,d})+\tfrac12\overset{1}{g}{}^{0a}\overset{0}{g}_{bc,\theta}.$$

The nonzero components of the Ricci curvature at the orders ϵ^{-2} and ϵ^{-1} are

$$\overset{-2}{R}_{00}=-\tfrac12(\overset{0}{g}{}^{ab}\overset{0}{g}_{ab,\theta})_{,\theta}-\tfrac14\overset{0}{g}{}^{ac}\overset{0}{g}_{bc,\theta}\overset{0}{g}{}^{bd}\overset{0}{g}_{ad,\theta}+\tfrac12\overset{0}{g}{}^{01}\overset{0}{g}_{01,\theta}\overset{0}{g}{}^{ab}\overset{0}{g}_{ab,\theta}, \tag{5.24}$$

$$\overset{-1}{R}_{01}=-(\overset{0}{g}{}^{01}\overset{0}{g}_{01,1})_{\theta}-\tfrac12(\overset{0}{g}{}^{ab}\overset{0}{g}_{ab,1})_{\theta}-\tfrac14\overset{0}{g}{}^{ac}\overset{0}{g}_{bc,\theta}\overset{0}{g}{}^{bd}\overset{0}{g}_{ad,1}, \tag{5.25}$$

$$\overset{-1}{R}_{ab}=-\overset{0}{g}{}^{01}(\overset{0}{g}_{ab,1\theta}-\tfrac12\overset{0}{g}{}^{cd}(\overset{0}{g}_{ac,\theta}\overset{0}{g}_{bd,1}+\overset{0}{g}_{ac,1}\overset{0}{g}_{bd,\theta})+\tfrac14\overset{0}{g}{}^{cd}(\overset{0}{g}_{cd,1}\overset{0}{g}_{ab,\theta}+\overset{0}{g}_{cd,\theta}\overset{0}{g}_{ab,1})), \tag{5.26}$$

$$\begin{aligned}\overset{-1}{R}_{00} = &-\tfrac{1}{2}\overset{1}{g}{}^{a}_{a,\theta\theta} - \tfrac{1}{2}\overset{0}{g}{}^{ac}\overset{0}{g}_{bc,\theta}\overset{1}{g}{}^{b}_{a,\theta} + \tfrac{1}{2}\overset{0}{g}{}^{01}\overset{0}{g}_{01,\theta}\overset{1}{g}{}^{a}_{a,\theta} - (\overset{0}{g}{}^{ab}\overset{0}{g}_{ab,\theta})_0 \\ &- \tfrac{1}{2}\overset{0}{g}{}^{ac}\overset{0}{g}_{bc,\theta}\overset{0}{g}{}^{bd}\overset{0}{g}_{ad,0} + \tfrac{1}{2}\overset{0}{g}{}^{01}\overset{0}{g}{}^{ab}(\overset{0}{g}_{01,\theta}\overset{0}{g}_{ab,0} + \overset{0}{g}_{01,0}\overset{0}{g}_{ab,\theta}),\end{aligned} \tag{5.27}$$

$$\begin{aligned}\overset{-1}{R}_{0a} = &\tfrac{1}{2}(\overset{0}{g}_{ab}\overset{0}{g}{}^{01}(\overset{0}{g}_{01}\overset{1}{g}{}^{0b})_\theta)_\theta + \tfrac{1}{4}\overset{0}{g}{}^{cd}\overset{0}{g}_{cd,\theta}\overset{0}{g}_{ab}\overset{0}{g}{}^{01}(\overset{0}{g}_{01}\overset{1}{g}{}^{0b})_\theta \\ &- \tfrac{1}{2}(\overset{0}{g}{}^{01}\overset{0}{g}_{01,a} + \overset{0}{g}{}^{cd}\overset{0}{g}_{cd,a})_\theta + \tfrac{1}{2}(\overset{0}{g}{}^{bc}\overset{0}{g}_{ab,\theta})_c + \tfrac{1}{4}\overset{0}{g}{}^{bc}\overset{0}{g}_{ab,\theta}\overset{0}{g}{}^{de}\overset{0}{g}_{de,c} + \tfrac{1}{4}\overset{0}{g}{}^{01}\overset{0}{g}_{01,a}\overset{0}{g}{}^{cd}\overset{0}{g}_{cd,\theta} \\ &- \tfrac{1}{4}\overset{0}{g}{}^{bd}\overset{0}{g}_{cd,\theta}\overset{0}{g}{}^{ce}\overset{0}{g}_{be,a}.\end{aligned} \tag{5.28}$$

VI. THE ASYMPTOTIC EXPANSION

We choose coordinates

$$(x^0,x^1,x^2,x^3) = (u,v,y,z) \tag{6.1}$$

in which the metric has the form (5.1). We introduce functions M,U,V,W of (θ,v,y,z) such that

$$\begin{aligned}\overset{0}{g}_{01} &= -e^{-M}, \\ (\overset{0}{g}_{ab}) &= \begin{pmatrix} e^{-U+V}\cosh W & -e^{-U}\sinh W \\ -e^{-U}\sinh W & e^{-U-V}\cosh W \end{pmatrix}.\end{aligned} \tag{6.2}$$

It follows from (5.1), (6.1), and (6.2) that the leading order metric has the form of the colliding plane wave metric,

$$\overset{0}{\mathbf{g}} = -2e^{-M}\,du\,dv + e^{-U}(e^{V}\cosh W\,dy^2 - 2\sinh W\,\mathrm{dy}\,dz + e^{-V}\cosh W\,dz^2). \tag{6.3}$$

From (5.24), the only component of the leading order perturbation equation (4.13) which is not identically satisfied is

$$\overset{-2}{R}_{00} = 0. \tag{6.4}$$

Using (5.24) and (6.2) in (6.4), we obtain the θ-constraint equation,

$$U_{\theta\theta} = \tfrac{1}{2}(U_\theta^2 + V_\theta^2\cosh^2 W + W_\theta^2) - U_\theta M_\theta. \tag{6.5}$$

From (5.25) to (5.28), the only components of the first-order perturbation equation (4.14) which are not identically satisfied are

$$\overset{-1}{R}_{01} = 0, \quad \overset{-1}{R}_{ab} = 0, \tag{6.6}$$

$$\overset{-1}{R}_{00} = 0, \quad \overset{-1}{R}_{0a} = 0. \tag{6.7}$$

Using (5.25)–(5.26) and (6.2) in (6.6), we get the evolution equations in the colliding plane wave equations,

$$U_{\theta v} = U_\theta U_v, \tag{6.8}$$

$$V_{\theta v} = \tfrac{1}{2}(U_\theta V_v + U_v V_\theta) - (V_\theta W_v + V_v W_\theta)\tanh W, \tag{6.9}$$

$$W_{\theta v} = \tfrac{1}{2}(U_\theta W_v + U_v W_\theta) + V_\theta V_v \sinh W \cosh W. \tag{6.10}$$

$$M_{\theta v}=\tfrac{1}{2}(-U_\theta U_v+V_\theta V_v\cosh^2 W+W_\theta W_v). \tag{6.11}$$

From (5.2) and (5.27)–(5.28), we find that the remaining equations (6.7) are satisfied by a suitable choice of the first-order metric components $\overset{1}{g}_{ab},\overset{1}{g}_{1a}$.

VII. VARIATIONAL PRINCIPLE

The variational principle for the vacuum Einstein field equations is

$$\delta S=0,\quad S=\int L\,d^4x,$$
$$L=R\sqrt{-\det g}, \tag{7.1}$$

where R is the scalar curvature,

$$R=g^{\alpha\beta}R_{\alpha\beta}.$$

Using (4.2), (4.3), and (4.10) to expand the scalar curvature, we obtain that

$$R=\frac{1}{\epsilon^2}\overset{-2}{R}+\frac{1}{\epsilon}\overset{-1}{R}+O(1),$$
$$\overset{-2}{R}=\overset{0}{g}{}^{\alpha\beta}\overset{-2}{R}_{\alpha\beta}, \tag{7.2}$$
$$\overset{-1}{R}=\overset{0}{g}{}^{\alpha\beta}\overset{-1}{R}_{\alpha\beta}-\overset{1}{g}{}^{\alpha\beta}\overset{-2}{R}_{\alpha\beta}.$$

For a metric of the form (5.21), we find that

$$\overset{-2}{R}=0,$$
$$\overset{-1}{R}=\overset{0}{g}{}^{ab}\overset{-1}{R}_{ab}+2\overset{0}{g}{}^{01}\overset{-1}{R}_{01}-\overset{1}{g}{}^{00}\overset{-2}{R}_{00}. \tag{7.3}$$

The only order one metric component which appears in (7.3) is

$$\lambda=-\overset{1}{g}{}^{00}.$$

In the derivation of the asymptotic equations, we used a coordinate system in which $\lambda=0$—see (5.3). In the variational principle, λ acts as a Lagrange multiplier for the constraint equation, so we will not set it to zero until after we take variations.

We use (7.3) in (7.1), expand the result with respect to ϵ, and write the expanded Lagrangian in terms of λ and the functions M,U,V,W, defined in (6.2). This gives

$$L=\frac{1}{\epsilon}\overset{-1}{L}+O(1),$$

$$\overset{-1}{L}=\{-2M_{\theta v}-4U_{\theta v}+3U_\theta U_v+V_\theta V_v\cosh^2 W+W_\theta W_v\}e^{-U}$$
$$+\lambda\{U_{\theta\theta}-\tfrac{1}{2}(U_\theta^2+V_\theta^2\cosh^2 W+W_\theta^2)+U_\theta M_\theta\}e^{-M-U}.$$

We make a change of variables in the integration,

$$d^4x=du\,dv\,dy\,dz=\epsilon\,d\theta\,dv\,dy\,dz,$$

and omit the integration with respect to the parametric variables (y,z). The leading order asymptotic variational principle then becomes

$$\delta \overset{0}{S}=0, \quad \overset{0}{S}=\int \overset{-1}{L}\, d\theta\, dv.$$

Variations of $\overset{0}{S}$ with respect to the first-order metric component λ give the constraint (6.5). Variations with respect to M,U,V,W give the evolution equations (6.8)–(6.11), after we set $\lambda=0$. It is permissible to set $\lambda=0$ because the constraint is a gauge-type constraint which is preserved by the evolution equations.

VIII. BOUNDARY CONDITIONS

In this section, we discuss the derivation of boundary conditions for the asymptotic equations. For simplicity, we consider a ''sandwich'' wave located near the null surface $u=0$ that varies rapidly in a thin strip

$$\theta_- \leqslant \frac{u}{\epsilon} \leqslant \theta_+ .$$

We denote the slowly varying metrics on either side of the wave by

$$\mathbf{g}=\begin{cases} \mathbf{g}_+ & \text{in } u>0 \\ \mathbf{g}_- & \text{in } u<0. \end{cases} \tag{8.1}$$

We consider a coordinate patch around a point on the surface $u=0$ with local coordinates (u,v,y,z) chosen as in the derivation of the asymptotic equations. In order for the metric outside the wave to join continuously with the metric inside, we must have

$$\mathbf{g}_\pm \rightarrow -2e^{-M_\pm}\, du\, dv + e^{-U_\pm}(e^{V_\pm}\cosh W_\pm\, dy^2 - 2\sinh W_\pm\, dy\, dz + e^{-V_\pm}\cosh W_\pm\, dz^2), \tag{8.2}$$

as $u \rightarrow 0^\pm$, where $M_\pm, U_\pm, V_\pm, W_\pm$ are functions of (v,y,z). From (6.3), (8.2), and the continuity of the metric, it follows that the solution of (6.8)–(6.11) satisfies the characteristic boundary conditions,

$$M=M_\pm, \quad U=U_\pm, \quad V=V_\pm, \quad W=W_\pm \quad \text{when} \;\; \theta=\theta_\pm . \tag{8.3}$$

This data need not satisfy the constraint (2.13).

The asymptotic equations must be supplemented by a condition which specifies the profile of the wave. For example, we can impose a characteristic initial condition

$$M=M_0, \quad U=U_0, \quad V=V_0, \quad W=W_0 \quad \text{when} \;\; v=0, \tag{8.4}$$

where M_0, U_0, V_0, W_0 are functions of (θ,y,z) which satisfy the constraint (6.5). The characteristic initial data must be compatible with the characteristic boundary data, meaning that

$$M_0(\theta_\pm, y,z)=M_\pm(0,y,z),$$

together with the analogous conditions for the other variables.

Equations (6.8)–(6.11), the characteristic initial condition (8.4) on $v=0$, and the characteristic boundary condition (8.3) on $\theta=\theta_-$ form a well-posed problem. Provided that singularities do not form, the problem has a unique solution, so the solution at $\theta=\theta_+$ is uniquely determined. Thus, in principle, the asymptotic equations (6.8)–(6.11) and the characteristic initial data (8.4) determine a set of jump relations that connect the minus and plus metrics ahead of and behind the wave, respectively. If the metric ahead of the wave is known, then the jump conditions provide characteristic boundary conditions on $u=0$ for the space–time behind the wave. Together with a characteristic initial condition on $v=0$ and $u>0$, for example, this gives a characteristic initial value problem[10] for the full field equations. This problem determines the slowly varying metric behind the wave, at least locally. The solution of this problem typically includes a slowly varying gravitational wave component which propagates in the opposite direction into the space–time behind the rapidly varying ''sandwich'' wave.

For instance, in the case of a plane polarized wave, the solution of (3.9) for U is[6]

$$U(\theta,v)=-\log[f(\theta)+g(v)]. \tag{8.5}$$

Here f and g are functions of integration, and we do not explicitly show the parametric dependence of the functions on (y,z). The solution is nonsingular provided that $f(\theta)+g(v)>0$. From (8.3), (8.4), and (8.5) we have

$$f(\theta)+g(0)=e^{-U_0(\theta)}, \quad f(\theta_-)+g(v)=e^{-U_-(v)}. \tag{8.6}$$

It follows from (8.5) to (8.6) that the jump relation for U is

$$e^{-U_+(v)}-e^{-U_-(v)}=e^{-U_0(\theta_+)}-e^{-U_0(\theta_-)}.$$

Use of (8.5) in (3.10) gives a linear wave equation for V,

$$(f+g)V_{\theta v}=\tfrac{1}{2}(g_v V_\theta+f_\theta V_v).$$

Solution of this equation with the characteristic initial data $V=V_0$ on $v=0$ and the characteristic boundary data $V=V_-$ on $\theta=\theta_-$ determines, in principle, the solution $V=V_+$ on $\theta=\theta_+$. Finally, for $W=0$, we define the v-constraint function G by

$$G=U_{vv}-\tfrac{1}{2}(U_v^2+V_v^2)+U_v M_v\,. \tag{8.7}$$

It follows from (8.7) and (3.9)–(3.11) that

$$G_\theta=U_\theta G.$$

Integration of this equation with respect to θ implies that

$$\log G_+(v)-\log G_-(v)=U_+(v)-U_-(v).$$

This equation provides a jump condition for M.

One difficulty which arises in the formulation of boundary conditions ahead of the wave is that the metric $\mathbf{g}_-$ may not be given in a coordinate system which is compatible with the coordinate system used in the derivation of the asymptotic equations. It is then necessary to construct compatible coordinates (u,v,y,z). The u coordinate is the phase, so it is a null coordinate of the metric, which can be found by solving an eikonal equation, subject to appropriate initial conditions. The v coordinate is a null coordinate that is orthogonal to u, while the y and z coordinates parametrize the null geodesics on the surface $u=v=0$.

If the gravitational wave front $u=0$ forms a caustic, then the solution of the eikonal equation becomes multivalued. When this happens, the local plane-wave approximation breaks down, and the asymptotic solution is not valid. However, the focusing at a caustic of the congruence of null geodesics associated with the phase does not necessarily imply the formation of a space–time singularity.

IX. EXAMPLES

In this section, we derive boundary conditions for the asymptotic equations which describe the propagation of a gravitational wave into Minkowski space–time, the exterior Schwarzschild space–time, and Robertson–Walker space–time. In each example, we consider the case of spherical waves, where the boundary data can be explicitly computed.

A. Nonplanar wave propagation into Minkowski space–time

We suppose that the space–time ahead of the wave is flat. In inertial coordinates $(t,\vec{x})$, with $t=x^0$ and $\vec{x}=(x^1,x^2,x^3)$, the metric is

$$\mathbf{g}_- = -dt^2 + d\vec{x}^2.$$

We consider a wave with phase

$$u=\frac{t-w(\vec{x})}{\sqrt{2}}.$$

The phase u is a null function of $\mathbf{g}_-$ if

$$|\nabla w|^2 = 1,$$

where ∇ is the gradient with respect to $\vec{x}$. We define

$$v=\frac{t+w(\vec{x})}{\sqrt{2}},$$

and choose coordinates $y(\vec{x}),z(\vec{x})$ such that $\nabla w,\nabla y,\nabla z$ are orthogonal. In the (u,v,y,z) coordinates, we have

$$\mathbf{g}_- = -2\,du\,dv + \frac{1}{|\nabla y|^2}dy^2 + \frac{1}{|\nabla z|^2}dz^2. \tag{9.1}$$

A comparison of (8.2) and (9.1) implies that the minus boundary data are given by

$$M_- = 0, \quad e^{-U_-} = \frac{1}{|\nabla y||\nabla z||_{u=0}}, \quad e^{-V_-} = \left.\frac{|\nabla y|}{|\nabla z|}\right|_{u=0}, \quad W_- = 0.$$

For example, in the case of an outgoing spherical wave, suitable coordinates are

$$u=\frac{t-r}{\sqrt{2}}, \quad v=\frac{t+r}{\sqrt{2}}, \quad y=\vartheta, \quad z=\varphi, \tag{9.2}$$

where (r,ϑ,φ) are spherical polar coordinates and $t>0$. In (u,v,y,z) coordinates, the flat space–time metric is

$$\mathbf{g}_- = -2\,du\,dv + \tfrac{1}{2}(u-v)^2(dy^2+\sin^2 y\,dz^2).$$

Evaluation of this metric at $u=0$ and a comparison with (8.2) gives the minus boundary data

$$M_-=0, \quad e^{-U_-}=\tfrac{1}{2}v^2\sin y, \quad e^{-V_-}=\sin y, \quad W_-=0,$$

where $v>0$. In this case, M_-, V_-, and W_- are independent of v, while $U=U_-$ satisfies the equation

$$U_{vv}=\tfrac{1}{2}U_v^2.$$

Thus, the boundary data satisfy the v-constraint equation (2.13). The solution is therefore identical to an exact solution for the collision of outgoing and incoming spherical waves, with an additional slow parametric dependence on the polar angles (ϑ, φ). Some exact solutions for spherical wave propagation into flat space–time are constructed in Ref. 11.

For an incoming spherical wave, we use

$$u=\frac{t+r}{\sqrt{2}},\quad v=\frac{t-r}{\sqrt{2}},$$

where $t<0$. This leads to the same boundary data as in the case of an outgoing spherical wave, but with $v<0$, instead of $v>0$.

B. Gravitational waves incident on a black hole

The exterior Schwarzschild metric is

$$\mathbf{g}_-=-a\,dt^2+\frac{1}{a}dr^2+r^2(d\vartheta^2+\sin^2\vartheta\,d\varphi^2), \tag{9.3}$$

where $r>2m$ and

$$a(r)=1-\frac{2m}{r}.$$

The contravariant metric tensor is

$$\mathbf{g}_-^\sharp=-\frac{1}{a}\partial_t^2+a\partial_r^2+\frac{1}{r^2}\left(\partial_\vartheta^2+\frac{1}{\sin^2\vartheta}\partial_\varphi^2\right).$$

For simplicity, we consider an axially symmetric phase of the form

$$u=\frac{t-w(r,\vartheta)}{\sqrt{2}}.$$

The function u is null if w satisfies the eikonal equation

$$aw_r^2+\frac{1}{r^2}w_\vartheta^2=\frac{1}{a}.$$

We define the orthogonal null coordinate v by

$$v=\frac{t+w(r,\vartheta)}{\sqrt{2}},$$

and choose a coordinate $y(r,\vartheta)$ whose gradient is orthogonal to the gradient of $w(r,\vartheta)$. That is,

$$y_r=-\frac{hw_\vartheta}{r^2},\quad y_\vartheta=haw_r,$$

where $h(r,\vartheta)$ is a suitable integrating factor. We take $z=\varphi$. In (u,v,y,z) coordinates, the Schwarzschild metric (9.3) is given by

$$\mathbf{g}_{-} \mp -2a\,du\,dv + \frac{r^2}{h^2}dy^2 + r^2\sin^2\vartheta\,dz^2. \tag{9.4}$$

Inversion of the change of coordinates $(t,r,\vartheta,\varphi)\mapsto(u,v,y,z)$ implies that $r=r_{-}(v,y)$ and $\vartheta=\vartheta_{-}(v,y)$ on $u=0$ for suitable functions r_{-} and ϑ_{-}. A comparison of (8.2) and (9.4) implies that the boundary data are given by

$$e^{-M_{-}}=a_{-}, \quad e^{-U_{-}}=\frac{r_{-}^2\sin\vartheta_{-}}{h_{-}}, \quad e^{-V_{-}}=h_{-}\sin\vartheta_{-}, \quad W_{-}=0,$$

where $a_{-}=a(r_{-})$ and $h_{-}=h(r_{-},\vartheta_{-})$.

In the case of an incoming spherical wave incident on the black hole, suitable coordinates are

$$u=\frac{t+A(r)}{\sqrt{2}}, \quad v=\frac{t-A(r)}{\sqrt{2}}, \quad y=\vartheta, \quad z=\varphi, \tag{9.5}$$

where

$$A_r=\frac{1}{a}.$$

Integration of this equation implies that

$$A(r)=r+\log(r-2m).$$

In (u,v,y,z) coordinates, the exterior Schwarzschild metric is

$$\mathbf{g}_{-}=-2a\,du\,dv+r^2(dy^2+\sin^2 y\,dz^2). \tag{9.6}$$

From (9.5), we have $r=r_{-}(v)$ on $u=0$, where

$$A(r_{-})=\frac{v}{\sqrt{2}}. \tag{9.7}$$

A comparison of (8.2) and (9.6) implies that the boundary data ahead of the incoming spherical wave are given by

$$e^{-M_{-}}=a_{-}, \quad e^{-U_{-}}=r_{-}^2\sin y, \quad e^{-V_{-}}=\sin y, \quad W_{-}=0. \tag{9.8}$$

Dropping the minus subscripts, we find that the constraint function G in (8.7) for the boundary data (9.8) is given by

$$G=2\left(\frac{a_v r_v}{ar}-\frac{r_{vv}}{r}\right).$$

Differentiation of (9.7) with respect to v implies that

$$r_v=-\frac{a}{\sqrt{2}}, \quad r_{vv}=-\frac{a_v}{\sqrt{2}}.$$

Use of this equation in the expression for G implies that $G=0$. Thus, the boundary data (9.8) satisfies the v-constraint equation (2.13).

Numerical solutions of the interaction of a spherical gravitational wave with a black hole appear in Ref. 12.

C. Gravitational waves in a Robertson–Walker space–time

The Robertson–Walker metric is

$$\mathbf{g}_- = -dt^2 + \frac{1}{R^2}\left\{ \frac{1}{1-kr^2} dr^2 + r^2 (d\vartheta^2 + \sin^2 \vartheta \, d\varphi^2) \right\}, \tag{9.9}$$

where $R(t)$ is the scale factor, and $k=-1,0,1$.

As in the Schwarzschild example, we consider an axially symmetric phase for simplicity, given by

$$u = \frac{I(t) - w(r,\vartheta)}{\sqrt{2}},$$

where

$$I_t = R, \quad (1-kr^2) w_r^2 + \frac{1}{r^2} w_\vartheta^2 = 1. \tag{9.10}$$

We define an orthogonal null coordinate v by

$$v = \frac{I(t) + w(r,\vartheta)}{\sqrt{2}}.$$

We choose a coordinate $y(r,\vartheta)$ whose gradient is orthogonal to the gradient of $w(r,\vartheta)$, so that

$$y_r = -\frac{h w_\vartheta}{r^2 \sqrt{1-kr^2}}, \quad y_\vartheta = h\sqrt{1-kr^2} w_r,$$

where $h(r,\vartheta)$ is a suitable integrating factor, and take $z=\varphi$. In (u,v,y,z) coordinates, the Robertson–Walker metric (9.9) is given by

$$\mathbf{g}_- = -\frac{2}{R^2} du\, dv + \frac{r^2}{h^2 R^2} dy^2 + \frac{r^2}{R^2} \sin^2 \vartheta \, dz^2. \tag{9.11}$$

A comparison of (8.2) and (9.11) implies that the boundary data are given by

$$e^{-M_-} = \frac{1}{R_-^2}, \quad e^{-U_-} = \frac{r_-^2 \sin \vartheta_-}{h_- R_-^2}, \quad e^{-V_-} = h_- \sin \vartheta_-, \quad W_- = 0,$$

where $r = r_-(v,y)$, $\vartheta = \vartheta_-(v,y)$, $R = R_-(v,y)$, and $h = h_-(v,y)$ on $u=0$.

For an outgoing spherical wave in a Robertson–Walker space–time, suitable coordinates are

$$u = \frac{I(t) - w(r)}{\sqrt{2}}, \quad v = \frac{I(t) + w(r)}{\sqrt{2}}, \quad y = \vartheta, \quad z = \varphi,$$

where

$$w_r = \frac{1}{\sqrt{1-kr^2}}.$$

Integration of this equation implies that

$$w(r)=\begin{cases}\sin^{-1} r & \text{if } k=1\\ r & \text{if } k=1\\ \sinh^{-1} r & \text{if } k=-1.\end{cases}$$

The corresponding boundary data are given by

$$e^{-M_-}=\frac{1}{R_-^2},\quad e^{-U_-}=\frac{r_-^2\sin y}{R_-^2},\quad e^{-V_-}=\sin y,\quad W_-=0, \tag{9.12}$$

where $t_-(v)$ and $r_-(v)$ are given by

$$I(t_-)=\frac{v}{\sqrt{2}},\quad r_-=\begin{cases}\sin(v/\sqrt{2}) & \text{if } k=1\\ v/\sqrt{2} & \text{if } k=0\\ \sinh(v/\sqrt{2}) & \text{if } k=-1\end{cases} \tag{9.13}$$

and $R_-=R(t_-)$.

Dropping the minus subscripts, we find that the constraint function G in (8.7) for the boundary data (9.12) is given by

$$G=2\left(\frac{R_{vv}}{R}-\frac{r_{vv}}{r}\right). \tag{9.14}$$

From (9.10) and (9.13), we find that

$$r_{vv}=-\tfrac{1}{2}kr,\quad R_{vv}=\frac{RR_{tt}-R_t^2}{2R^3}.$$

Use of these expressions in (9.14) gives

$$G=\frac{RR_{tt}-R_t^2}{R^4}+k.$$

Thus, in general, the boundary data (9.12) do not satisfy the v-constraint equation (2.13).

ACKNOWLEDGMENTS

The work of J.K.H. was partially supported by the NSF under Grant No. DMS-9704152. The work of G.A. was partially supported by the CNR 1996/97 short-term fellowships program.

[1] M. W. Brinkmann, Proc. Natl. Acad. Sci. USA **9**, 1 (1923).
[2] L. D. Landau and E. M. Lifschitz, *The Classical Theory of Fields* (Pergamon, Oxford, 1975).
[3] K. A. Khan and R. Penrose, Nature (London) **229**, 185 (1971).
[4] P. Szekeres, Nature (London) **228**, 1183 (1970).
[5] P. Szekeres, J. Math. Phys. **13**, 286 (1972).
[6] J. B. Griffiths, *Colliding Plane Waves in General Relativity* (Oxford University Press, Oxford, 1991).
[7] R. A. Isaacson, Phys. Rev. **166**, 1263 (1968).
[8] Y. Choquet-Bruhat, Commun. Math. Phys. **12**, 16 (1969).
[9] A. M. Anile, *Relativistic Fluids and Magneto-fluids* (Cambridge University Press, Cambridge, 1989).
[10] A. D. Rendall, Proc. R. Soc. London, Ser. A **427**, 221 (1990).
[11] G. A. Alekseev and J. B. Griffiths, Class. Quantum Grav. **13**, 2191 (1996).
[12] A. Abrahams, D. Bernstein, D. Hobill, E. Seidel, and L. Smarr, Phys. Rev. D **45**, 3544 (1992).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Super-energy tensor for space–times with vanishing scalar curvature

Miguel Á. G. Bonilla[a),b)]
Departament de Física Fonamental, Universitat de Barcelona, Avda. Diagonal 647, E-08028 Barcelona, Spain

Carlos F. Sopuerta[a),c)]
Institut for Theoretical Physics, FSU Jena, Max-Wien-Platz 1, D-07743 Jena, Germany

(Received 22 May 1998; accepted for publication 23 February 1999)

A four-index tensor is constructed with terms both quadratic in the Riemann tensor and linear in its second derivatives, which has zero divergence for space–times with vanishing scalar curvature. This tensor reduces in vacuum to the Bel–Robinson tensor. Furthermore, the completely timelike component referred to any observer is positive, and zero if and only if the space–time is flat (excluding some unphysical space–times). We also show that this tensor is the unique one that can be constructed with these properties. Such a tensor does not exist for general gravitational fields. Finally, we study this tensor in several examples: the Friedmann–Lemaître–Robertson–Walker space–times filled with radiation, the plane–fronted gravitational waves, and the Vaidya radiating metric. © *1999 American Institute of Physics.* [S0022-2488(99)04206-1]

I. INTRODUCTION

The investigation of conservation laws in general relativity has a long history. From its very beginning much of this research was based on pseudotensors instead of fully covariant methods. The aim of many of these works was to find differential laws which, once transformed into integral ones, were interpreted as energy balances in such a way that expressions for the energy and momentum densities of the gravitational field could be identified.

The covariant approach to this problem is mainly based on the analogy of the Bel–Robinson tensor[1] with the energy–momentum tensor of the electromagnetic field (there are other approaches based on this analogy, see e.g., Ref. 2). The Bel–Robinson tensor is conserved *in vacuum*, completely symmetric and traceless. Moreover, the completely timelike component referred to any observer (described by a timelike unit vector field) is non-negative, and its vanishing implies that the space–time is conformally flat (flat in vacuum). This is a desirable positivity property for any candidate to gravitational energy density. In spite of these good properties, the Bel–Robinson tensor has dimensions of energy density square and this fact makes its interpretation somewhat unclear. Nevertheless, it has been revealed as a very useful tool in many kinds of studies, which has led to some efforts in finding extensions of the Bel–Robinson tensor to more general cases than vacuum. Therefore, the question arises whether generalizations of the Bel–Robinson tensor exist for space–times not necessarily empty.

The Bel tensor[3] was the first attempt on this problem. It is a tensor whose completely timelike component is positive and zero only when the space–time is Minkowski. In vacuum, it reduces to the Bel–Robinson tensor, but in the general case it is no longer conserved.

For general space–times, Sachs[4] found a divergence-free tensor that coincides with the Bel–Robinson tensor in vacuum. Unfortunately, this tensor does not satisfy any positivity property and it is neither completely symmetric nor traceless.

[a)]Also at: Laboratori de Física Matemàtica, Societat Catalana de Física, IEC, Barcelona.
[b)]Electronic mail: mangel@ffn.ub.es
[c)]Electronic mail: cfs@tpi.uni-jena.de

The systematic treatment of this problem was made by Collinson,[5] who found all four-index divergence-free tensors with terms either quadratic in the Riemann tensor or linear in its second derivatives. The discovered result is that any such tensor can be derived from only *one* tensor, namely $T_{10}^{\alpha\beta\lambda\mu}$ [see Eq. (A1)], whose divergence with respect to the first index vanishes. However, there is not enough freedom to construct a tensor with its time component positive.

In this paper we show that, unlike the general case, for space–times with zero scalar curvature $(R=0)$ it is possible to construct a unique generalization of the Bel–Robinson tensor. We begin in Sec. II by proving that, when $R=0$, there exists another conserved tensor which cannot be derived from the Collinson one. In Sec. III we show that, demanding symmetry in the three free indices, there is not any other tensor independent from these two. This new tensor allows us to construct (Sec. IV) a divergence-free tensor which has the completely timelike component non-negative and zero only when the space–time is flat (excluding some cases that, via Einstein's field equations, have an unphysical matter content). This tensor is completely symmetric in its last three indices, but it is impossible to get a similar tensor symmetric in all their indices. We remark that it is not possible to construct any other tensor with such characteristics.

In order to illustrate this development, we study in Sec. V some examples in which we can define this tensor, namely: the Friedmann–Lemaître–Robertson–Walker (FLRW) models with an energy–momentum content of (incoherent) radiation $(p=\varrho/3)$, the plane-fronted gravitational waves with parallel rays (*pp waves*), and the Vaidya radiating space–time.

Finally, we recall that for purely electromagnetic space–times, and supposing that the Einstein field equations hold, Penrose and Rindler[6] also gave a generalization of the Bel–Robinson tensor by using spinor methods. This tensor is conserved, completely symmetric, and traceless in its last three indices. In Appendix B we find its tensorial expression and a new (to our knowledge) positivity property.

II. DEDUCTION OF THE NEW CONSERVED TENSOR FOR $R=0$

Unless otherwise stated, throughout this paper we will consider the metric tensor $g_{\alpha\beta}$ to have signature $(-,+,+,+)$. The convention for indices on the Riemann tensor that will be used is defined through the Ricci identities:

$$(\nabla_\alpha\nabla_\beta-\nabla_\beta\nabla_\alpha)v_\lambda=-R^\sigma{}_{\lambda\alpha\beta}v_\sigma, \tag{1}$$

where v_α is an arbitrary 1-form. The Ricci tensor and the scalar curvature are defined as usual: $R_{\alpha\beta}\equiv R^\sigma{}_{\alpha\sigma\beta}$ and $R\equiv R^\sigma_\sigma$. We also recall the Riemann symmetries and the first and second Bianchi identities:

$$\begin{gathered} R_{\alpha\beta\lambda\mu}=R_{[\alpha\beta][\lambda\mu]}=R_{\lambda\mu\alpha\beta}, \\ R_{[\alpha\beta\lambda]\mu}=0, \\ \nabla_{[\nu}R_{\alpha\beta]\lambda\mu}=0. \end{gathered} \tag{2}$$

The procedure we are going to use to find the conserved tensor starts from the expression for the divergence of the Bel tensor:[1]

$$\nabla_\alpha T^{\alpha\beta\lambda\mu}=R^\beta{}_\rho{}^\lambda{}_\sigma J^{\mu\sigma\rho}+R^\beta{}_\rho{}^\mu{}_\sigma J^{\lambda\sigma\rho}-\tfrac{1}{2}g^{\lambda\mu}R^\beta{}_{\rho\sigma\gamma}J^{\sigma\gamma\rho}, \tag{3}$$

where the Bel tensor $T^{\alpha\beta\lambda\mu}$ and $J^{\alpha\beta\lambda}$ are defined as follows:

$$\begin{gathered} T^{\alpha\beta\lambda\mu}\equiv\tfrac{1}{2}(R^{\alpha\rho\lambda\sigma}R^\beta{}_\rho{}^\mu{}_\sigma+{*R*}^{\alpha\rho\lambda\sigma}{*R*}^\beta{}_\rho{}^\mu{}_\sigma+{*R}^{\alpha\rho\lambda\sigma}{*R}^\beta{}_\rho{}^\mu{}_\sigma+R*^{\alpha\rho\lambda\sigma}R*^\beta{}_\rho{}^\mu{}_\sigma), \\ J^{\lambda\mu\beta}\equiv\nabla^\lambda R^{\mu\beta}-\nabla^\mu R^{\lambda\beta}=\nabla_\sigma R^{\mu\lambda\beta\sigma}, \end{gathered} \tag{4}$$

with the asterisk being the usual dual operator acting over any pair of antisymmetric indices:

$$*R_{\alpha\beta\lambda\mu}\equiv\tfrac{1}{2}\eta_{\alpha\beta\sigma\rho}R^{\sigma\rho}{}_{\lambda\mu},\quad R*_{\alpha\beta\lambda\mu}\equiv\tfrac{1}{2}\eta_{\lambda\mu\sigma\rho}R_{\alpha\beta}{}^{\sigma\rho},$$

$$*R*_{\alpha\beta\lambda\mu}\equiv\tfrac{1}{4}\eta_{\alpha\beta\gamma\delta}\eta_{\lambda\mu\sigma\rho}R^{\gamma\delta\sigma\rho},$$

and $\eta_{\alpha\beta\lambda\mu}$ is the canonical volume 4-form. Our purpose now is to work out the right-hand side (r.h.s.) of Eq. (3) in order to convert it into a global divergence. To that end, we will repeatedly integrate by parts and make use of Eqs. (1) and (2).

To begin with, the last term in Eq. (3) can be easily transformed into a divergence, by means of the Ricci and Bianchi identities (1) and (2):

$$-\tfrac{1}{2}g^{\lambda\mu}R^{\beta}{}_{\rho\sigma\gamma}J^{\sigma\gamma\rho}=-g^{\lambda\mu}\nabla_{\alpha}\nabla_{\rho}J^{\alpha\rho\beta}. \tag{5}$$

Next we expand the leading terms of the r.h.s. of (3) by using the definition of $J^{\alpha\beta\lambda}$ (4) and integrating by parts:

$$R^{\beta}{}_{\rho}{}^{\lambda}{}_{\sigma}\nabla^{\mu}R^{\sigma\rho}-\nabla_{\alpha}(R^{\mu\rho}R^{\beta}{}_{\rho}{}^{\lambda\alpha})+R^{\mu\rho}\nabla_{\alpha}R^{\beta}{}_{\rho}{}^{\lambda\alpha}+[\lambda\leftrightarrow\mu]. \tag{6}$$

The first term in the previous expression can be rewritten by means of the Ricci identities (1) as follows:

$$R^{\beta}{}_{\rho}{}^{\lambda}{}_{\sigma}\nabla^{\mu}R^{\sigma\rho}=(\nabla^{\lambda}\nabla_{\sigma}-\nabla_{\sigma}\nabla^{\lambda})\nabla^{\mu}R^{\sigma\beta}+[-\nabla^{\alpha}(R^{\mu}{}_{\alpha}{}^{\lambda}{}_{\sigma}R^{\sigma\beta})+R^{\sigma\beta}\nabla^{\alpha}R^{\lambda}{}_{\sigma}{}^{\mu}{}_{\alpha}]+R^{\lambda}_{\sigma}\nabla^{\mu}R^{\sigma\beta}. \tag{7}$$

Thus, we have converted the r.h.s. of Eq. (3) into a divergence plus the following terms:

$$\begin{aligned}&R^{\sigma\beta}\nabla^{\alpha}R^{\lambda}{}_{\sigma}{}^{\mu}{}_{\alpha}+R^{\sigma\lambda}\nabla^{\mu}R^{\beta}_{\sigma}+R^{\mu\rho}\nabla_{\alpha}R^{\beta}{}_{\rho}{}^{\lambda\alpha}+[\lambda\leftrightarrow\mu]\\&\quad=R^{\sigma\beta}(\nabla_{\sigma}R^{\lambda\mu}-\nabla^{\lambda}R^{\mu}_{\sigma})+R^{\sigma\lambda}\nabla^{\mu}R^{\beta}_{\sigma}+R^{\mu\rho}(\nabla_{\rho}R^{\beta\lambda}-\nabla^{\beta}R^{\lambda}_{\rho})+[\lambda\leftrightarrow\mu].\end{aligned} \tag{8}$$

Now, taking into account the contracted Bianchi identities ($\nabla_{\mu}R^{\mu\nu}=\tfrac{1}{2}\nabla^{\nu}R$), these terms can be transformed into the following expression:

$$\begin{aligned}&-\nabla^{\beta}(R^{\lambda}_{\sigma}R^{\sigma\mu})+\nabla^{\mu}(R^{\lambda}_{\sigma}R^{\sigma\beta})+\nabla^{\lambda}(R^{\mu}_{\sigma}R^{\sigma\beta})-2R^{\beta\sigma}(\nabla^{\mu}R^{\lambda}_{\sigma}+\nabla^{\lambda}R^{\mu}_{\sigma})\\&\quad+\nabla_{\sigma}(R^{\beta\lambda}R^{\mu\sigma}+R^{\beta\mu}R^{\lambda\rho}+2R^{\beta\sigma}R^{\lambda\mu})-\tfrac{1}{2}(R^{\beta\lambda}\nabla^{\mu}R+R^{\beta\mu}\nabla^{\lambda}R+2R^{\lambda\mu}\nabla^{\beta}R).\end{aligned} \tag{9}$$

The last three terms of this expression vanish when R is constant, so we are finally left with $-2R^{\beta\sigma}(\nabla^{\mu}R^{\lambda}_{\sigma}+\nabla^{\lambda}R^{\mu}_{\sigma})$. Nevertheless, notice that our final purpose is to find a conserved tensor $T''^{\alpha\beta\lambda\mu}$ whose completely timelike component referred to an observer $\vec{u},T''^{\alpha\beta\lambda\mu}u_{\alpha}u_{\beta}u_{\lambda}u_{\mu}$, is positive, which means that we are only interested in the symmetric part. Therefore, without lost of generality, we can symmetrize the whole expression and, as a consequence, the remaining terms transform themselves into a divergence:

$$-4R^{\sigma(\beta}\nabla^{\lambda}R^{\mu)}_{\sigma}=-2\nabla^{(\beta}(R^{\lambda}_{\sigma}R^{\mu)\sigma}). \tag{10}$$

So, we have finally achieved a conserved tensor if the scalar curvature vanishes (in fact, if it is constant). Collecting all the previous terms (5)–(10) we get the final result:

$$\nabla_{\alpha}T''^{\alpha\beta\lambda\mu}=-2R^{(\beta\lambda}\nabla^{\mu)}R,$$

where we have defined

$$\begin{aligned}T''^{\alpha\beta\lambda\mu}\equiv{}&T^{\alpha(\beta\lambda\mu)}-4R^{\alpha(\beta}R^{\lambda\mu)}+g^{\alpha(\beta}R^{\lambda}_{\sigma}R^{\mu)\sigma}-2\nabla^{(\beta}\nabla^{\lambda}R^{\mu)\alpha}+2\nabla^{(\beta}\nabla^{|\alpha|}R^{\lambda\mu)}-2\nabla^{\alpha}\nabla^{(\beta}R^{\lambda\mu)}\\&-2g^{\alpha(\beta}\nabla_{\sigma}\nabla^{\lambda}R^{\mu)\sigma}-\nabla_{\sigma}\nabla^{\sigma}R^{\alpha(\beta}g^{\lambda\mu)}+\nabla_{\sigma}\nabla^{\alpha}R^{\sigma(\beta}g^{\lambda\mu)}.\end{aligned} \tag{11}$$

It is a matter of checking that this tensor cannot be obtained from Collinson tensor $T_{10}^{\alpha\beta\lambda\mu}$ when it is restricted to the case $R=0$. Now, we are left with the question of its uniqueness.

III. UNIQUENESS

In this section we will prove that, in the case we are concerned with ($R=0$), no other conserved tensor exists, symmetric in its last three indices, independent from $T''^{\alpha\beta\lambda\mu}$ and the Collinson tensor, $T_{10}^{\alpha\beta\lambda\mu}$. The reasoning is the following. Suppose you are given a tensor $T'''^{\alpha\beta\lambda\mu}$ which is conserved when $R=0$. The divergence computed in the general case will be a combination of the following type (taking into account symmetries and unit dimensions):

$$\nabla_\alpha T'''^{\alpha\beta\lambda\mu}=aR^{(\beta\lambda}\nabla^{\mu)}R+bR\nabla^{(\beta}R^{\lambda\mu)}+cR^{\sigma(\beta}\nabla_\sigma Rg^{\lambda\mu)}+dRg^{(\beta\lambda}\nabla^{\mu)}R+e\nabla^{(\beta}\nabla^{\lambda}\nabla^{\mu)}R$$
$$+f\nabla^\sigma\nabla_\sigma\nabla^{(\beta}Rg^{\lambda\mu)}+h\nabla^\sigma\nabla^{(\beta}\nabla_\sigma Rg^{\lambda\mu)}+i\nabla^{(\beta}\nabla^{|\sigma|}\nabla_\sigma Rg^{\lambda\mu)},$$

a, b, c, d, e, f, h, and *i* being constants. This can be immediately cast in the following form:

$$\nabla_\alpha T'''^{\alpha\beta\lambda\mu}=(a-b)R^{(\beta\lambda}\nabla^{\mu)}R+\nabla_\alpha\tau^{\alpha\beta\lambda\mu},$$

where $\tau^{\alpha\beta\lambda\mu}$ stands for:

$$\tau^{\alpha\beta\lambda\mu}\equiv bRg^{\alpha(\beta}R^{\lambda\mu)}+c(RR^{\alpha(\beta}g^{\lambda\mu)}-\tfrac{1}{4}g^{\alpha(\beta}g^{\lambda\mu)}R^2)+d\tfrac{1}{2}g^{(\beta\lambda}g^{\mu)\alpha}R^2+eg^{\alpha(\beta}\nabla^{\lambda}\nabla^{\mu)}R$$
$$+fg^{\alpha\sigma}\nabla_\sigma\nabla^{(\beta}Rg^{\lambda\mu)}+hg^{\alpha\sigma}\nabla^{(\beta}\nabla_\sigma Rg^{\lambda\mu)}+ig^{\alpha(\beta}\nabla^{|\sigma|}\nabla_\sigma Rg^{\lambda\mu)},$$

so clearly it is a tensor that vanishes when R does. That is, if a tensor of the kind we are considering is divergence free when $R=0$, in the general case its divergence should be a multiple of $R^{(\beta\lambda}\nabla^{\mu)}R$ plus the divergence of a tensor of the type $\tau^{\alpha\beta\lambda\mu}$. If we had two such tensors, a suitable combination of them removing the term $R^{(\beta\lambda}\nabla^{\mu)}R$ would give a conserved tensor for the general case and, therefore, due to the Collinson result,[5] it could be constructed from $T_{10}^{\alpha\beta\lambda\mu}$. Given that the tensor $\tau^{\alpha\beta\lambda\mu}$ vanishes when $R=0$, the three tensors would not be independent in that case.

On the other hand, this reasoning shows that, from the very beginning, we were able to know that in the $R=0$ case at most one more conserved tensor could exist apart from Collinson's one, as finally has been the case.

IV. POSITIVITY

As it has been pointed out above, in the general case all the conserved tensors can be constructed from $T_{10}^{\alpha\beta\lambda\mu}$. This construction is based in two procedures. First, it is clear that if we perform any permutation on the last three indices we will still have a conserved tensor. Second, by taking the two traces $T_{10}^{\alpha\beta\rho}{}_{\rho}$, $T_{10}^{\alpha\rho\beta}{}_{\rho}$ and multiplying them by $g^{\lambda\mu}$ we obtain new conserved tensors. Actually, there is not any other two-index divergence-free tensor independent from them. These two tensors can be taken to be (as usually obtained by Hamiltonian differentiation):

$$t_1^{\alpha\beta}=2\nabla^\alpha\nabla^\beta R-2g^{\alpha\beta}\nabla_\mu\nabla^\mu R+\tfrac{1}{2}g^{\alpha\beta}R^2-2RR^{\alpha\beta},$$

$$t_2^{\alpha\beta}=2\nabla_\sigma\nabla^\alpha R^{\sigma\beta}-\nabla_\sigma\nabla^\sigma R^{\alpha\beta}-2R^{\alpha\sigma}R_\sigma^\beta+\tfrac{1}{2}R_{\sigma\rho}R^{\sigma\rho}g^{\alpha\beta}-\tfrac{1}{2}g^{\alpha\beta}\nabla_\mu\nabla^\mu R.$$

This is all the freedom we have in the general case.

In the $R=0$ case, there exists only one two-index conserved tensor, which is that obtained from $t_2^{\alpha\beta}$. On the other hand, as we consider tensors which are symmetric in the last three indices, from the Collinson tensor we will only have one independent tensor (apart from the two-index tensor), namely $T_{10}^{\alpha(\beta\lambda\mu)}$ or, equivalently, the Sachs tensor $T'^{\alpha\beta\lambda\mu}$,[4] which is symmetric in its last three indices.

Therefore, to construct the super-energy tensor in the $R=0$ case, we are led with three tensors: (i) the Sachs tensor $T'^{\alpha\beta\lambda\mu}$ (restricted to the $R=0$ case):

$$T'^{\alpha\beta\lambda\mu} \equiv T^{\alpha(\beta\lambda\mu)} + 2R^{\alpha\sigma}R^{(\beta}_{\sigma}g^{\lambda\mu)} - \tfrac{1}{3}g^{\alpha(\beta}R^{\lambda}_{\sigma}R^{\mu)\sigma} - \tfrac{1}{2}R_{\sigma\rho}R^{\sigma\rho}g^{\alpha(\beta}g^{\lambda\mu)}$$
$$+2\nabla^{(\beta}\nabla^{\lambda}R^{\mu)\alpha} - \tfrac{2}{3}\nabla^{(\beta}\nabla^{|\alpha|}R^{\lambda\mu)} - \tfrac{2}{3}\nabla^{\alpha}\nabla^{(\beta}R^{\lambda\mu)}$$
$$+\tfrac{4}{3}\nabla_{\sigma}\nabla^{\sigma}R^{(\beta\lambda}g^{\mu)\alpha} - 2g^{\alpha(\beta}\nabla_{\sigma}\nabla^{\lambda}R^{\mu)\sigma} - \nabla_{\sigma}\nabla^{\alpha}R^{\sigma(\beta}g^{\lambda\mu)},$$

(ii) the two-index tensor $t^{\alpha\beta}$:

$$t^{\alpha\beta} = 2\nabla_{\sigma}\nabla^{\alpha}R^{\sigma\beta} - \nabla_{\sigma}\nabla^{\sigma}R^{\alpha\beta} - 2R^{\alpha\sigma}R^{\beta}_{\sigma} + \tfrac{1}{2}R_{\sigma\rho}R^{\sigma\rho}g^{\alpha\beta},$$

and (iii) the tensor $T''^{\alpha\beta\lambda\mu}$ previously found in (11).

Now, we have to combine these three tensors in such a way that any observer measures a positive quantity. Moreover, we would like that this completely timelike component vanishes if and only if the space–time is flat.

First of all, we have to take into account that terms made of derivatives of the Ricci tensor do not have a definite sign, so it would be necessary to eliminate their contributions. This aim can only be achieved by means of the following combination:

$$A^{\alpha\beta\lambda\mu} \equiv \tfrac{1}{2}(3T'^{\alpha\beta\lambda\mu} - T''^{\alpha\beta\lambda\mu}) + \tfrac{5}{2}t^{\alpha(\beta}g^{\lambda\mu)},$$

which more explicitly reads:

$$A^{\alpha\beta\lambda\mu} = T^{\alpha(\beta\lambda\mu)} + 2R^{\alpha(\beta}R^{\lambda\mu)} - 2R^{\alpha\sigma}R^{(\beta}_{\sigma}g^{\lambda\mu)} - g^{\alpha(\beta}R^{\lambda}_{\sigma}R^{\mu)\sigma} + \tfrac{1}{2}R_{\sigma\rho}R^{\sigma\rho}g^{\alpha(\beta}g^{\lambda\mu)}$$
$$+2\nabla^{(\beta}\nabla^{\lambda}R^{\mu)\alpha} - 2\nabla^{(\beta}\nabla^{|\alpha|}R^{\lambda\mu)} + 3\nabla_{\sigma}\nabla^{\alpha}R^{\sigma(\beta}g^{\lambda\mu)}$$
$$-2g^{\alpha(\beta}\nabla_{\sigma}\nabla^{\lambda}R^{\mu)\sigma} + 2\nabla_{\sigma}\nabla^{\sigma}R^{(\beta\lambda}g^{\mu)\alpha} - 2\nabla_{\sigma}\nabla^{\sigma}R^{\alpha(\beta}g^{\lambda\mu)}. \quad (12)$$

Since we have already exhausted all the freedom, we finally examine the completely timelike component referred to any timelike unit vector $\vec{u}$:

$$A(\vec{u}) \equiv A_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu}$$
$$= T_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu} + \tfrac{1}{2}R_{\sigma\rho}R^{\sigma\rho} + 2(R_{\alpha\beta}u^{\alpha}u^{\beta})^2$$
$$+3(R^{\alpha\sigma}R^{\beta}_{\sigma})u_{\alpha}u_{\beta} - (\nabla_{\sigma}\nabla^{\alpha}R^{\sigma\beta})u_{\alpha}u_{\beta}. \quad (13)$$

To check the positivity of $A(\vec{u})$ it is convenient to write out the last term of Eq. (13) in the following form:

$$\nabla_{\sigma}\nabla^{\alpha}R^{\sigma\beta} = C^{\alpha}{}_{\sigma\rho}{}^{\beta}R^{\sigma\rho} + 2R^{\alpha\sigma}R^{\beta}_{\sigma} - \tfrac{1}{2}g^{\alpha\beta}R_{\sigma\rho}R^{\sigma\rho}.$$

At this point, we introduce four spatial tensors, namely $E_{\alpha\beta}(\vec{u})$, $H_{\alpha\beta}(\vec{u})$, $M_{\alpha\beta}(\vec{u})$, and $N_{\alpha\beta}(\vec{u})$, that (together with R) wholly characterize the Riemann tensor.[7] Their definitions, properties, and some useful formulas are given in Appendix A.

Introducing the previous definitions in (13) we obtain, after some calculations:

$$A(\vec{u}) = (E_{\sigma\rho} - M_{\sigma\rho})(E^{\sigma\rho} - M^{\sigma\rho}) + H_{\sigma\rho}H^{\sigma\rho} + 3N_{\sigma\rho}N^{\sigma\rho} + (R_{\alpha\beta}u^{\alpha}u^{\beta})^2, \quad (14)$$

which is a sum of square terms (all the tensors appearing here are spatial). Therefore it is manifestly positive and its vanishing implies:

$$H_{\alpha\beta} = 0, \quad N_{\alpha\beta} = 0,$$

$$R_{\alpha\beta}u^{\alpha}u^{\beta} = 0, \quad E_{\alpha\beta} = M_{\alpha\beta}.$$

The previous expressions lead to $R_{\alpha\beta}u^{\alpha}=0$ (see Appendix A). This condition, when considering the Einstein field equations, immediately drives to an unphysical energy–momentum tensor. Hence, if we eliminate these unphysical space–times (for instance, adding any energy condition), the vanishing of $A(\vec{u})$ finally implies the Minkowski space–time.

V. SOME EXAMPLES

In this section we are going to study the tensor $A^{\alpha\beta\lambda\mu}$ (12) in some space–times with vanishing scalar curvature. In particular, we are going to consider the following examples: (i) the radiation FLRW cosmological models, (ii) the *pp waves*, and (iii) the Vaidya radiating metric. In examples (i) and (ii) we will give the expression for $A^{\alpha\beta\lambda\mu}$ and its completely timelike component $A(\vec{u})$ (14) for an arbitrary observer $\vec{u}$. In example (iii), for the sake of brevity we will give only the expression of $A(\vec{u})$, also for an arbitrary observer.

In the first example we study the case of the FLRW models (see, for instance, Ref. 8) with vanishing scalar curvature, the radiation models, whose energy–momentum tensor (of perfect-fluid type) is given by the following expression (throughout this section we will use units in which $8\pi G=c=1$):

$$T_{\alpha\beta}=\varrho U_{\alpha}U_{\beta}+ph_{\alpha\beta}, \quad p=\tfrac{1}{3}\varrho,$$

where $\vec{U}$ is the fluid velocity ($U^{\alpha}U_{\alpha}=-1$), ϱ the energy density, p the pressure, and $h_{\alpha\beta}=g_{\alpha\beta}+U_{\alpha}U_{\beta}$ the orthogonal projector to the fluid velocity. The line element of these conformally flat models can be written as

$$ds^2=-dt^2+a^2(t)\{d\chi^2+\Sigma^2(\epsilon,\chi)(d\theta^2+\sin^2\theta\delta\varphi^2)\},$$

where $\Sigma(\epsilon,\chi)$ is given by

$$\Sigma(\epsilon,\chi)=\begin{cases}\sin\chi & \text{if } \epsilon=1\\ \chi & \text{if } \epsilon=0\\ \sinh\chi & \text{if } \epsilon=-1.\end{cases}$$

The fluid velocity $\vec{U}$, the scale factor $a(t)$, and the energy density $\varrho(t)$ are

$$\vec{U}=\frac{\partial}{\partial t}, \quad a^2(t)=(t-t_0)[2A-\epsilon(t-t_0)], \quad \varrho(t)=\frac{3A^2}{a^4(t)},$$

respectively, and A and t_0 are arbitrary constants.

After some straightforward calculations, and using the special properties of these space–times, we arrive at the following expression for $A^{\alpha\beta\lambda\mu}$:

$$A^{\alpha\beta\lambda\mu}=\tfrac{4}{3}\varrho^2\{U^{\alpha}U^{\beta}U^{\lambda}U^{\mu}+U^{\alpha}U^{(\beta}h^{\lambda\mu)}+\tfrac{5}{3}h^{\alpha(\beta}U^{\lambda}U^{\mu)}+\tfrac{1}{3}h^{\alpha(\beta}h^{\lambda\mu)}\}. \tag{15}$$

As we can see, it is proportional to the energy density squared. We can also check that it is indeed divergence free. Now, let us compute the completely timelike component (14) of this tensor with respect to an arbitrary observer $\vec{u}$. To that end, we decompose $\vec{u}$ in the next way

$$\vec{u}=\gamma(\vec{U}+\vec{v}), \quad v^{\alpha}U_{\alpha}=0, \quad v^{\alpha}v_{\alpha}=v^2\geqslant 0, \quad \gamma\equiv(1-v^2)^{-1/2},$$

where the case $\vec{v}=0$ corresponds to an observer comoving with the fluid ($\vec{u}=\vec{U}$). Then, from (14), (15) we find that $A(\vec{u})$ is given by

$$A(\vec{u})=\tfrac{4}{3}\varrho^2\gamma^4\{1+\tfrac{8}{3}v^2+\tfrac{1}{3}v^4\}.$$

That is, it is function of ϱ and v only. Moreover, it increases monotonically as v increases and its minimum corresponds to the case $v=0$, in which the observer is comoving with the fluid.

Now, we are going to study the tensor $A^{\alpha\beta\lambda\mu}$ in the case of the *pp waves* space–times. The corresponding line element can be written in null coordinates $\{u,v,\zeta,\overline{\zeta}\}$ as follows (see Ref. 8 for details)

$$ds^2=-2\,du\,dv+2d\zeta\,d\overline{\zeta}-2H\,du^2,$$

where H is an arbitrary function which does not depend on $v[H=H(u,\zeta,\overline{\zeta})]$. Using the following Newman–Penrose basis $\{\ell,k,m,\overline{m}\}$,

$$\vec{\ell}=\frac{\partial}{\partial v},\quad \vec{k}=\frac{\partial}{\partial u}-H\frac{\partial}{\partial v},\quad \vec{m}=\frac{\partial}{\partial \zeta},$$

the Ricci and self-dual Weyl tensors are

$$R_{\alpha\beta}=2\Phi\ell_\alpha\ell_\beta,\quad \hat{C}_{\alpha\beta\lambda\mu}\equiv C_{\alpha\beta\lambda\mu}+i\overset{*}{C}_{\alpha\beta\lambda\mu}=2\Psi_4V_{\alpha\beta}V_{\lambda\mu}, \tag{16}$$

respectively, where the quantities Φ, Ψ_4, and $V_{\alpha\beta}$ are given by

$$\Phi\equiv H_{,\zeta\overline{\zeta}},\quad \Psi_4=H_{,\overline{\zeta}\overline{\zeta}},\quad V_{\alpha\beta}\equiv 2\ell_{[\alpha}m_{\beta]}.$$

From (16) we can see that the energy–momentum content can correspond with vacuum, Einstein–Maxwell, or pure radiation fields. Moreover, the Petrov type is *N*, with $\vec{\ell}$ being the repeated principal direction of the Weyl tensor, which in fact is a constant vector field $(\nabla_\alpha\ell_\beta=0)$.

After some calculations, we have found that the Bel tensor $T^{\alpha\beta\lambda\mu}$ and our tensor $A^{\alpha\beta\lambda\mu}$ are

$$T^{\alpha\beta\lambda\mu}=4(\Phi^2+\Psi_4\overline{\Psi}_4)\ell^\alpha\ell^\beta\ell^\lambda\ell^\mu,$$

$$\tfrac{1}{4}A^{\alpha\beta\lambda\mu}=(3\Phi^2+\Psi_4\overline{\Psi}_4)\ell^\alpha\ell^\beta\ell^\lambda\ell^\mu+\ell^\alpha\ell^{(\beta}\nabla^\lambda\nabla^{\mu)}\Phi-\ell^{(\beta}\ell^\lambda\nabla^{\mu)}\nabla^\alpha\Phi$$
$$+[g^{\alpha(\beta}\ell^\lambda\ell^{\mu)}-\ell^\alpha\ell^{(\beta}g^{\lambda\mu)}]\nabla^\sigma\nabla_\sigma\Phi.$$

From this expression, the completely timelike component is given by

$$A(\vec{u})=4(3\Phi^2+\Psi_4\overline{\Psi}_4)(\ell^\alpha u_\alpha)^4.$$

Then, the vanishing of $A(\vec{u})$ implies the Minkowski space–time.

Finally, we are going to consider the Vaidya radiating space–time (see for instance Ref. 8). For the sake of brevity we only give here the completely timelike component (14) of the tensor $A^{\alpha\beta\lambda\mu}$. The line element of this spherically-symmetric metric can be written as follows:

$$ds^2=-2F^2(u,v)du\,dv+r^2(u,v)(d\theta^2+\sin^2\theta\,d\varphi^2),$$

where

$$F^2(u,v)=f(u)\frac{\partial r}{\partial v},\quad \frac{\partial r}{\partial u}=\frac{1}{2}f(u)\left(\frac{2m(u)}{r}-1\right).$$

Here, $m(u)$ is the invariantly defined mass function. As is well known, the Petrov type of this metric is *D*. Then, taking the following Newman–Penrose adapted basis:

$$\vec{\ell}=\frac{-1}{F}\frac{\partial}{\partial v},\quad \vec{k}=\frac{-1}{F}\frac{\partial}{\partial u},\quad \vec{m}=\frac{1}{\sqrt{2}r}\left(\frac{\partial}{\partial \theta}+\frac{i}{\sin\theta}\frac{\partial}{\partial \varphi}\right),$$

where $\vec{\ell}$ and $\vec{k}$ are aligned with the principal directions of the Weyl tensor, the Ricci tensor is given by

$$R_{\alpha\beta}=2\Phi\ell_{\alpha}\ell_{\beta}, \quad \Phi\equiv-\frac{m_{,u}}{r^2 r_{,u}}.$$

And the only nonzero component of the Weyl tensor in this basis (see Ref. 8) is

$$\Psi_2=-\frac{m(u)}{r^3(u,v)},$$

which in this case is real.

In terms of these quantities, we have found the following expression for $A(\vec{u})$:

$$A(\vec{u})=2[\Psi_2+\Phi(\ell^{\alpha}u_{\alpha})^2]^2+4\Psi_2^2[36(\ell^{\alpha}u_{\alpha})^2(k^{\beta}u_{\beta})^2-18(\ell^{\alpha}u_{\alpha})(k^{\beta}u_{\beta})+1]+10\Phi^2(\ell^{\alpha}u_{\alpha})^4,$$

and again, taking into account that

$$2(\ell^{\alpha}u_{\alpha})(k^{\beta}u_{\beta})\geqslant 1 \Rightarrow 36(\ell^{\alpha}u_{\alpha})^2(k^{\beta}u_{\beta})^2-18(\ell^{\alpha}u_{\alpha})(k^{\beta}u_{\beta})+1\geqslant 1,$$

$A(\vec{u})$ vanishes if and only if the space–time is the Minkowski space–time. When we restrict ourselves to observers lying on the 2-planes generated by the principal directions $[2(\ell^{\alpha}u_{\alpha})\times(k^{\beta}u_{\beta})=1]$, which are precisely the observers that minimize $A(\vec{u})$, the result is

$$A(\vec{u})=2[\Psi_2+\Phi(\ell^{\alpha}u_{\alpha})^2]^2+4\Psi_2^2+10\Phi^2(\ell^{\alpha}u_{\alpha})^4.$$

ACKNOWLEDGMENTS

M.Á.G.B. gratefully acknowledges the Comissionat per a Universitats i Recerca de la Generalitat de Catalunya for financial support. C.F.S. gratefully acknowledges financial support in the form of a fellowship from the Alexander von Humboldt Foundation.

APPENDIX A: USEFUL DEFINITIONS

We first write down the Collinson tensor:[5]

$$T_{10}^{\alpha\beta\lambda\mu}=6Q_1^{\alpha\beta\lambda\mu}+Q_2^{\alpha\beta(\lambda\mu)}+Q_3^{\alpha\beta(\lambda\mu)}, \tag{A1}$$

where

$$Q_1^{\alpha\beta\lambda\mu}=R^{\lambda}_{\sigma}R^{\beta\mu\alpha\sigma}+R^{\beta}_{\sigma}R^{\mu\alpha\lambda\sigma}+\nabla^{\mu}\nabla_{\sigma}R^{\alpha\sigma\beta\lambda}-\nabla^{\beta}\nabla_{\sigma}R^{\alpha\lambda\mu\sigma}-g^{\alpha\beta}\nabla_{\rho}\nabla_{\sigma}R^{\mu\sigma\lambda\rho},$$

$$Q_2^{\alpha\beta\lambda\mu}=-4\nabla^{\lambda}\nabla_{\sigma}R^{\alpha\sigma\beta\mu}-6\nabla^{\mu}\nabla_{\sigma}R^{\alpha\beta\lambda\sigma}-6g^{\alpha\mu}\nabla_{\sigma}\nabla^{\lambda}R^{\sigma\beta}+6g^{\alpha\mu}\nabla_{\rho}\nabla_{\sigma}R^{\sigma\lambda\beta\rho}+6\nabla^{\mu}\nabla^{\lambda}R^{\alpha\beta},$$

$$\begin{aligned}Q_3^{\alpha\beta\lambda\mu}=&-8g^{\alpha\lambda}R_{\sigma\rho}{}^{\mu}{}_{\gamma}R^{\rho\gamma\beta\sigma}+8R_{\sigma}{}^{\lambda\mu}{}_{\rho}R^{\beta\sigma\alpha\rho}+8R^{\mu}_{\sigma}R^{\beta\sigma\lambda\alpha}+8R_{\sigma}{}^{\beta\mu}{}_{\rho}R^{\lambda\sigma\alpha\rho}+2R^{\mu}_{\sigma}R^{\beta\alpha\sigma\lambda}\\&+2R_{\sigma}{}^{\mu\beta}{}_{\rho}R^{\lambda\sigma\alpha\rho}+3g^{\alpha\beta}R^{\mu}_{\sigma}R^{\lambda\sigma}-g^{\alpha\beta}R_{\sigma\rho}{}^{\mu}{}_{\gamma}R^{\rho\gamma\lambda\sigma},\end{aligned}$$

which is divergence free in the index α and whose only symmetry on the indices β, λ, and μ is $T_{10}^{\alpha\beta[\lambda\mu]}-T_{10}^{\alpha\mu[\beta\lambda]}=0$.

Next, in order to introduce other useful definitions, recall the well-known decomposition of the Riemann tensor into its irreducible parts under the full Lorentz group:

$$R_{\alpha\beta\lambda\mu}=C_{\alpha\beta\lambda\mu}+E_{\alpha\beta\lambda\mu}+G_{\alpha\beta\lambda\mu},$$

where $C_{\alpha\beta\lambda\mu}$ is the Weyl tensor and

$$E_{\alpha\beta\lambda\mu} \equiv \tfrac{1}{2}(\tilde{R}_{\alpha\lambda}g_{\beta\mu} - \tilde{R}_{\alpha\mu}g_{\beta\lambda} + \tilde{R}_{\beta\mu}g_{\alpha\lambda} - \tilde{R}_{\beta\lambda}g_{\alpha\mu}), \quad \tilde{R}_{\alpha\beta} \equiv R_{\alpha\beta} - \tfrac{1}{4}Rg_{\alpha\beta},$$

$$G_{\alpha\beta\lambda\mu} \equiv \frac{R}{12}(g_{\alpha\lambda}g_{\beta\mu} - g_{\alpha\mu}g_{\beta\lambda}),$$

with $R \equiv R^{\mu}_{\mu}$ being the scalar curvature.

The electric and magnetic parts of the Weyl tensor associated to a timelike vector field $\vec{u}$ are:

$$E_{\alpha\lambda}(\vec{u}) \equiv C_{\alpha\beta\lambda\mu}u^{\beta}u^{\mu}, \quad H_{\alpha\lambda}(\vec{u}) \equiv -\overset{*}{C}_{\alpha\beta\lambda\mu}u^{\beta}u^{\mu}.$$

These tensors are spatial (orthogonal to $\vec{u}$), symmetric, and traceless, and they fully determine the Weyl tensor.

We can proceed analogously with the tensor $E_{\alpha\beta\lambda\mu}$ and define (see Ref. 7 for a more detailed study of these matters):

$$M_{\alpha\lambda}(\vec{u}) \equiv E_{\alpha\beta\lambda\mu}u^{\beta}u^{\mu}, \quad N_{\alpha\lambda}(\vec{u}) \equiv -{*E}_{\alpha\beta\lambda\mu}u^{\beta}u^{\mu}.$$

We give here some of their properties, since they are less well known than the electric and magnetic parts of the Weyl tensor:

$$M_{\alpha\lambda} = M_{\lambda\alpha}, \quad M_{\alpha\lambda}u^{\lambda} = 0, \quad M^{\mu}_{\mu} = \tilde{R}_{\mu\nu}u^{\mu}u^{\nu},$$

$$N_{\alpha\lambda} = -N_{\lambda\alpha}, \quad N_{\alpha\lambda}u^{\lambda} = 0, \quad N^{\mu}_{\mu} = 0.$$

The tensor $M_{\alpha\lambda}$ has six independent components, while $N_{\alpha\lambda}$ has only three. Actually, they completely characterize the traceless Ricci tensor:

$$\tilde{R}_{\alpha\beta} = -2M_{\alpha\beta} - 4{*N}_{\sigma(\alpha}u^{\sigma}u_{\beta)} + M^{\sigma}_{\sigma}(g_{\alpha\beta} + 2u_{\alpha}u_{\beta}).$$

From the previous definitions, it is clear that

$$N_{\alpha\beta} = R_{\alpha\beta}u^{\alpha}u^{\beta} = R = 0 \Rightarrow R_{\alpha\beta} = -2M_{\alpha\beta},$$

and this implies, in particular, that $R_{\alpha\beta}u^{\beta} = 0$.

Let us finally give some formulas which are useful for the derivations of some expressions in Sec. IV.

$$T^{\alpha\beta\lambda\mu}u_{\alpha}u_{\beta}u_{\lambda}u_{\mu} = E_{\alpha\beta}E^{\alpha\beta} + H^{\alpha\beta}H_{\alpha\beta} + M_{\alpha\beta}M^{\alpha\beta} + N_{\alpha\beta}N^{\alpha\beta} + \frac{R^2}{48},$$

$$(\tilde{R}_{\alpha\rho}u^{\rho})(\tilde{R}^{\alpha}_{\sigma}u^{\sigma}) = 2N_{\sigma\rho}N^{\sigma\rho} - (M^{\sigma}_{\sigma})^2,$$

$$\tilde{R}_{\sigma\rho}\tilde{R}^{\sigma\rho} = 4M_{\sigma\rho}M^{\sigma\rho} - 4N_{\sigma\rho}N^{\sigma\rho}.$$

APPENDIX B: ELECTROMAGNETIC CASE

For space–times with an electromagnetic energy–momentum content, Penrose and Rindler[6] gave the following modification of the Bel–Robinson tensor (see Ref. 6 for the spinor notations and conventions):

$$t_{\alpha\beta\lambda\mu} = \Psi_{ABCD}\bar{\Psi}_{A'B'C'D'} - 2\gamma\nabla_{CD'}\varphi_{AB}\nabla_{C'D}\bar{\varphi}_{A'B'} + 6\gamma\nabla_{D(A'}\varphi_{(AB}\nabla_{C)|D'|}\bar{\varphi}_{B'C')}, \tag{B1}$$

where Ψ_{ABCD} and φ_{AB} are the Weyl and the electromagnetic spinors, respectively, and γ is the gravitational constant. This tensor is symmetric and traceless in the first three indices and has zero covariant derivative with respect to the last one, provided that the Einstein field equations hold:

$$\nabla^{\mu} t_{\alpha\beta\lambda\mu} = 0,$$

$$t_{\alpha\beta\lambda\mu} = t_{(\alpha\beta\lambda)\mu}, \quad t^{\alpha}_{\ \alpha\lambda\mu} = 0.$$

It is important to notice that the second part of this tensor is formed with $(\nabla F)^2$ terms, so it cannot be expressed by means of the Ricci tensor and its derivatives. Therefore it is independent from the tensors considered above.

The tensorial expression for (18) is rather involved but, in this case, it is useful in order to prove a positivity property. It should be noted here that in the final result we have returned to the initial signature metric $(-,+,+,+)$, and units such that $8\pi\gamma = c = 1$. With these conventions:

$$\begin{aligned} t_{\alpha\beta\lambda\mu} = {} & \tfrac{1}{4}\mathcal{T}_{\alpha\beta\lambda\mu} + 2\nabla_{(\alpha}F_{|\sigma|\beta}\nabla_{\lambda)}F^{\sigma}_{\ \mu} + 2\nabla_{(\alpha}F_{|\sigma|\beta}\nabla^{\sigma}F_{\lambda)\mu} \\ & - \tfrac{1}{2}\nabla_{(\alpha}F_{|\sigma\rho|}\nabla_{\beta}F^{\sigma\rho}g_{\lambda)\mu} + \nabla_{\sigma}F_{\rho(\alpha}\nabla^{\rho}F^{\sigma}{}_{\beta}g_{\lambda)\mu} - g_{(\alpha\beta}\nabla^{\sigma}F^{\rho}{}_{\lambda)}\nabla_{\rho}F_{\sigma\mu}, \end{aligned} \tag{B2}$$

where $\mathcal{T}_{\alpha\beta\lambda\mu}$ is the Bel–Robinson tensor and $F_{\alpha\beta}$ is the electromagnetic tensor. From this expression it is easily seen that (19) satisfies the following positivity property:

$$t_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu} \geqslant 0, \quad \forall \vec{u}, \quad u_{\mu}u^{\mu} < 0,$$

$$t_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu} = 0 \Leftrightarrow \begin{cases} \mathcal{T}_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu} = 0 \Leftrightarrow C_{\alpha\beta\lambda\mu} = 0 \\ \dot{F}_{\lambda\mu} \equiv u^{\alpha}\nabla_{\alpha}F_{\lambda\mu} = 0. \end{cases}$$

To prove this, let us introduce the orthogonal projector to $\vec{u}$, $h_{\alpha\beta} \equiv g_{\alpha\beta} + u_{\alpha}u_{\beta}$. Then, we compute $t_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu}$:

$$\begin{aligned} t_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu} &= \tfrac{1}{4}\mathcal{T}_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu} + 2u^{\alpha}u^{\beta}\dot{F}_{\sigma\alpha}\dot{F}^{\sigma}{}_{\beta} + \tfrac{1}{2}(\dot{F}_{\sigma\rho}\dot{F}^{\sigma\rho}) \\ &= \tfrac{1}{4}\mathcal{T}_{\alpha\beta\lambda\mu}u^{\alpha}u^{\beta}u^{\lambda}u^{\mu} + (h^{\alpha\sigma}\dot{F}_{\alpha\beta}u^{\beta})(h_{\lambda\sigma}\dot{F}^{\lambda\mu}u_{\mu}) \\ &\quad + \tfrac{1}{2}(h^{\alpha\sigma}h^{\beta\rho}\dot{F}_{\alpha\beta})(h_{\lambda\sigma}h_{\mu\rho}\dot{F}^{\lambda\mu}) \\ &\geqslant 0, \end{aligned}$$

and the equality holds only when $C_{\alpha\beta\lambda\mu} = 0$ and

$$h^{\alpha\lambda}h^{\beta\mu}\dot{F}_{\lambda\mu} = 0, \quad h^{\alpha\sigma}\dot{F}_{\alpha\beta}u^{\beta} = 0 \Leftrightarrow \dot{F}_{\alpha\beta} = 0.$$

[1] L. Bel, ''Sur la radiation gravitationnelle,'' C. R. Acad. Sci. Paris **247**, 1094–1096 (1958).
[2] B. Mashhoon, J. C. McClune, and H. Quevedo, ''Gravitational superenergy tensor,'' Phys. Lett. A **231**, 47–51 (1997).
[3] L. Bel, ''Introduction d'un tenseur du quatrième ordre,'' C. R. Acad. Sci. Paris **248**, 1297–1300 (1959).
[4] R. K. Sachs, ''Propagation laws for null and type III gravitational waves,'' Z. Phys. **157**, 462–477 (1960).
[5] C. D. Collinson, ''A class of conserved tensors in an arbitrary gravitational field,'' Proc. Cambridge Philos. Soc. **58**, 346–362 (1962).
[6] R. Penrose and W. Rindler, *Spinors and Space–time* (Cambridge University Press, Cambridge, 1984).
[7] M. Á. G. Bonilla and J. M. M. Senovilla, ''Some properties of the Bel and Bel–Robinson Tensors,'' Gen. Relativ. Gravit. **29**, 91–116 (1997).
[8] D. Kramer, H. Stephani, E. Herlt, and M. A. H. MacCallum, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Variational and conformal structure of nonlinear metric-connection gravitational Lagrangians

Spiros Cotsakis[a] and John Miritzis[b]
Department of Mathematics, University of the Aegean, Karlovassi 83200 Samos, Greece

Laurent Querella[c]
Institut d'Astrophysique et de Géophysique, Université de Liège,
Avenue de Cointe 5, B-4000 Liège, Belgium

(Received 15 May 1998; accepted for publication 5 March 1999)

We examine the variational and conformal structures of higher-order theories of gravity that are derived from a metric-connection Lagrangian that is an arbitrary function of the curvature invariants. We show that the constrained first-order formalism when applied to these theories may lead consistently to a new method of reduction of order of the associated field equations. We show that the similarity of the field equations that are derived from appropriate actions via this formalism to those produced by Hilbert varying purely metric Lagrangians is not merely formal but is implied by the diffeomorphism covariant property of the associated Lagrangians. We prove that the conformal equivalence theorem of these theories with general relativity plus a scalar field, holds in the extended framework of Weyl geometry with the same forms of field and self-interacting potential but, in addition, there is a new ''source term'' that plays the role of a stress. We point out how these results may be further exploited and address a number of new issues that arise from this analysis. © *1999 American Institute of Physics.* [S0022-2488(99)03906-7]

I. INTRODUCTION

According to the standard variational principle (Hilbert variation) that leads to the Einstein field equations of general relativity, the gravitational action $\int R\sqrt{-g}$ is varied with respect to the metric tensor of a spacetime manifold that is taken to be a four-dimensional Lorentz manifold $(\mathcal{M},\mathbf{g})$ with metric $\mathbf{g}$ and the Levi-Civita connection ∇. However, in many instances (see Ref. 1 for a complete review), one considers a Lorentz manifold with an arbitrary connection $\tilde{\nabla}$ that is incompatible with the metric, i.e., $\tilde{\nabla}\mathbf{g}\neq 0$. A motivation for such a generalization was initially inspired by the early work of Weyl (Ref. 2). In this case, one considers the variation of an appropriate action with respect to both the metric components g_{ab} and the connection coefficients Γ^a_{bc} without imposing from the beginning that Γ^a_{bc} be the usual Christoffel symbols. In the current literature this variational principle, where the metric and the connection are considered as independent variables, is referred to as the *first-order* or *metric-connection*, or simply *Palatini variation*. (For a historical commentary of this principle of variation and related issues, we refer to Ref. 3.)

Such alternative variational methods were first analyzed in the framework of nonlinear gravitational Lagrangians by Weyl (Ref. 2), Eddington (Ref. 4), and others (Refs. 5–20). In an effort to obtain second-order field equations different from Einstein's, Stephenson (Refs. 5, 6) and Higgs (Ref. 7) applied the first-order formalism to the quadratic Lagrangians R^2, $R_{ab}R^{ab}$, $R_{abcd}R^{abcd}$ and Yang (Ref. 8) investigated the Lagrangian $R_{abcd}R^{abcd}$, by analogy with the Yang–Mills Lagrangian. However, Buchdahl (Ref. 9) pointed out a difficulty associated with this version of

[a]Electronic mail: skot@aegean.gr
[b]Electronic mail: john@env.aegean.gr
[c]Electronic mail: querella@astro.ulg.ac.be

the metric-connection variation, which is related to imposing the metricity condition, i.e., the connection coefficients equal to the Christoffel symbols, *after* completing the variation, and subsequently constructed specific examples, showing that this version of the first-order formalism is not a reliable method, in general (Ref. 10). Van den Bergh (Ref. 11) arrived at a similar conclusion in the context of general scalar-tensor theories. The $R+\alpha R^2$ theory including matter was investigated in this framework by Shahid-Saless (Ref. 12) and generalized to the $f(R)$ case by Hamity and Barraco (Ref. 13). These authors also studied conservation laws and the weak field limit of the resulting equations. More recently, Ferraris *et al.* (Ref. 14) showed that the first-order formalism applied to general $f(R)$ vacuum Lagrangians leads to a series of Einstein spaces with cosmological constants determined by the explicit form of the function *f*. Similar results were obtained in the case of $f(\mathrm{Ric}^2)$ theories by Borowiec *et al.* (Ref. 15).

A consistent way to consider independent variations of the metric and connection in the context of Riemannian geometry is to add a compatibility condition between the metric and the connection as a constraint with Lagrange multipliers. In vacuum general relativity, this *constrained first-order formalism* results in the Lagrange multipliers vanishing identically as a consequence of the field equations (Ref. 16). This method was applied to quadratic Lagrangians, with the aim of developing a Hamiltonian formulation for these theories in Ref. 17.

Consider a Lorentzian manifold $(\mathcal{M},\mathbf{g},\tilde{\nabla})$ of dimension *D*, where $\tilde{\nabla}$ is an arbitrary symmetric connection. Hence, $\tilde{\nabla}\mathbf{g}\neq 0$ that is, the connection coefficients, are functions independent of the metric components, and the Ricci tensor is a function of the connection only. In the case of general relativity without matter fields, varying the corresponding action,

$$S=\int L\sqrt{-g}\,\mathrm{d}^D x,\quad L=g^{mn}R_{mn}, \tag{1}$$

one arrives at the well-known result that variation with respect to the metric produces the vacuum Einstein's equations, whereas variation with respect to the connection reveals that the connection is necessarily the Levi-Civita connection (provided that $D\neq 2$). The integral (1) is taken over a compact region $\mathcal{U}$ of the spacetime $(\mathcal{M},\mathbf{g},\tilde{\nabla})$, and we assume that the metric and the connection are held constant on the boundary of $\mathcal{U}$. In the sequel we omit the symbol $\mathrm{d}^D x$ under the integral sign and set $w:=\sqrt{-g}$. Gothic characters denote tensor densities, for example $\mathfrak{g}_{ab}:=wg_{ab}$.

In the presence of matter fields, there is an ambiguity because the compatibility condition between the metric and the connection does not hold. The matter Lagrangian depends primarily on the field variables ψ, and assumes a form that is a generalization of its special relativistic form, which is achieved via the strong principle of equivalence and the principle of minimal coupling, according to the scheme $\eta_{ab}\rightarrow g_{ab}$ and $\partial\rightarrow\tilde{\nabla}$ (the order of the two steps being irrelevant as long as the connection is the Levi-Civita one). Variation of the total action,

$$S=\int[\mathfrak{R}(g,\Gamma)+\mathfrak{L}_m(g,\psi,\tilde{\nabla}\psi)], \tag{2}$$

gives the following pair of equations:

$$G_{ab}=T_{ab}:=-\frac{2}{w}\frac{\delta\mathfrak{L}_m}{\delta g^{ab}}, \tag{3a}$$

$$\delta_c^b\,\tilde{\nabla}_d\mathfrak{g}^{ad}+\delta_c^a\,\tilde{\nabla}_d\mathfrak{g}^{bd}-2\,\tilde{\nabla}_c\mathfrak{g}^{ab}=2\,\frac{\delta\mathfrak{L}_m}{\delta\Gamma_{ab}^c}. \tag{3b}$$

These equations are inconsistent in general unless the matter Lagrangian does not depend explicitly on the connection, i.e., $\delta\mathfrak{L}_m/\delta\Gamma_{ab}^c=0$. [It is interesting to note that in the case $D=2$ in vacuum, since the equation $\Gamma^c_{ab}=\{^{c}_{ab}\}+1/2(\delta^c_a Q_b+\delta^c_b Q_a-g_{ab}Q^c)$, with $Q_a:=-\bar{\nabla}_a \ln w=-\partial_a \ln w+\Gamma_a$, and $\Gamma_a:=\Gamma^b_{ab}$, has a vanishing trace, $(1-D/2)(\partial_a \ln\sqrt{-g}-\Gamma_a)=0$, the Γ_a part of the connection is undetermined (Ref. 18).]

In Sec. II, after a brief review of the unconstrained metric-connection variational results in higher-order gravity theories, we show how the constrained first-order formalism is used to prove that the field equations of these theories can be given in a reduced form that makes the comparison to the usual Hilbert equations direct and point out that such a correspondence is due to the diffeomorphism covariance property of the associated Lagrangians. In Sec. III, we prove that in Weyl geometry, higher-order gravity theories are conformally equivalent to general relativity plus a scalar field matter source with the same self-interacting potential as in the standard Riemannian case, together with a new ''source term'' that arises as a result of the presence of the Weyl covariant vector field. In the last section, we comment on the usefulness of the extended form of the conformal equivalence and point out how these results may be further exploited to develop this framework.

II. CONSTRAINED AND UNCONSTRAINED VARIATIONS

We begin with a Lagrangian that is a smooth function of the scalar curvature R and vary the corresponding action,

$$S=\int wf(R), \tag{4}$$

with respect to the metric tensor and the connection to obtain, respectively,

$$f'R_{(ab)}-\tfrac{1}{2}fg_{ab}=0, \tag{5a}$$

$$\bar{\nabla}_a(wf'g^{bc})=0. \tag{5b}$$

Explicitly, the Γ equation (5b) reads as

$$(\partial_a \ln w+(\ln f')'\partial_a R-\Gamma_a)g_{bc}-\partial_a g_{bc}+\Gamma^m_{ba}g_{mc}+\Gamma^m_{ca}g_{mb}=0, \tag{6}$$

and so we can solve for Γ_a and substitute back in (6) to find

$$\partial_a\tilde{g}_{bc}=\Gamma^m_{ba}\tilde{g}_{mc}+\Gamma^m_{ca}\tilde{g}_{mb}, \tag{7}$$

where we have introduced a new metric $\tilde{g}_{ab}:=f'g_{ab}$, with conformal factor f'. This means that Γ is the Levi-Civita connection for the metric $\tilde{\mathbf{g}}$.

Equation (5a) is more straightforward. On the one hand, its trace $f'(R)R=2f(R)$ is satisfied identically if $f(R)=\alpha R^2$ (up to a constant rescaling factor α), and so (5a) becomes $R_{ab}-(1/4)Rg_{ab}=0$ [provided that $f'(R)\neq 0$], which finally gives $\tilde{R}_{ab}-(1/2\alpha)\tilde{g}_{ab}=0$ so that the underlined manifold is an Einstein space with constant scalar curvature $\tilde{R}=\tilde{g}^{ab}\tilde{R}_{ab}=2/\alpha$. On the other hand, one could regard the above trace as an algebraic equation in R with roots $\rho_1,\rho_2,\ldots$. This situation was analyzed by Ferraris *et al.* (Ref. 14), who showed that such an analysis leads to a series of Einstein spaces, each having a constant scalar curvature (see also Ref. 10).

By a completely analogous procedure, one finds Einstein spaces for the choice $L=f(r)$ where $r=Q_{ab}Q^{ab}$, and Q_{ab} is the symmetric part of the Ricci tensor, and also for the Lagrangian $L=f(K)$, where $K=R_{abcd}R^{abcd}$. Note that in this last case, varying the corresponding action with respect to the metric and the connection, one obtains

$$-\tfrac{1}{2}fg_{ab}-f'R_a{}^{klm}R_{bklm}+f'R^k{}_{alm}R_{kb}{}^{lm}+2f'R^k{}_{lam}R_k{}^l{}_b{}^m=0 \tag{8}$$

and

$$\overset{\Gamma}{\nabla}_d(wf' R_a{}^{(bc)d})=0, \tag{9}$$

with trace $f'(K)K=f(K)$. Hence, either $f(K)=\alpha K$ identically or, given a function f, the trace is solved algebraically for K. In contrast to the previous cases, there exists no natural way to derive a metric $\tilde{\mathbf{g}}$ from the field equation (9) unless the Weyl tensor vanishes (Ref. 19). Notice that the field equations derived from the Lagrangian $R_{[ab]}R^{ab}$ by the Palatini variation impose only four conditions upon the 40 connection coefficients and leave the metric components entirely undetermined (Ref. 10).

Let us now introduce a vector field Q_c called the Weyl covariant vector field and assume a linear metric-connection relation,

$$\overset{\Gamma}{\nabla}_c g_{ab}=-Q_c g_{ab}, \tag{10}$$

and define

$$C^c_{ab}=\Gamma^c_{ab}-\left\{ {c \atop ab} \right\}, \tag{11}$$

where

$$\Gamma^c_{ab}=\left\{ {c \atop ab} \right\}+\tfrac{1}{2}g^{cm}(Q_b g_{am}+Q_a g_{mb}-Q_m g_{ab}). \tag{12}$$

The *constrained* first-order formalism consists of adding to the original Lagrangian the following term as a constraint (with Lagrange multipliers Λ):

$$L_c(g,\Gamma,\Lambda)=\Lambda_r{}^{mn}\left[\Gamma^r_{mn}-\left\{ {r \atop mn} \right\}-C^r_{mn}\right]. \tag{13}$$

For instance, in Riemannian geometry (13) takes the form

$$L_c(g,\Gamma,\Lambda)=\Lambda_r{}^{mn}\left[\Gamma^r_{mn}-\left\{ {r \atop mn} \right\}\right],$$

while, in Weyl geometry,

$$L_c(g,\Gamma,\Lambda)=\Lambda_r{}^{mn}\left[\Gamma^r_{mn}-\left\{ {r \atop mn} \right\}-\frac{1}{2}g^{rs}(Q_n g_{ms}+Q_m g_{sn}-Q_s g_{mn})\right].$$

As an example, consider any of the previous test Lagrangians $L(g,\Gamma,\psi)$ and vary the resulting action,

$$S=\int w[L(g,\Gamma,\psi)+L_c(g,\Gamma,\Lambda)], \tag{14}$$

with respect to the independent fields g, Γ, Λ, and ψ. We find the g equations

$$\left.\frac{\delta(wL)}{\delta g^{ab}}\right|_\Gamma+wB_{ab}=0, \tag{15}$$

where B_{ab} is defined by

$$B_{ab}:=-\tfrac{1}{2}\overset{\Gamma}{\nabla}{}^m[\Lambda_{bam}+\Lambda_{amb}-\Lambda_{mab}], \tag{16}$$

and the Γ equations

$$\left.\frac{\delta L}{\delta \Gamma^{c}{}_{ab}}\right|_{g} + \Lambda_{c}{}^{ab} = 0. \tag{17}$$

Variation with respect to the matter fields ψ yields their respective equations of motion. In the Riemannian case, in particular, we have (some of the expressions given below appear incorrectly in Ref. 17): for the Lagrangian $L=R^2$,

$$\tfrac{1}{2}R^2 g^{ab} - 2RR^{ab} + B^{ab} = 0, \tag{18a}$$

$$\Lambda_{c}{}^{ab} = (2g^{ab}\delta^{m}_{c} - g^{am}\delta^{b}_{c} - g^{mb}\delta^{a}_{c})\bar{\nabla}_{m}R, \tag{18b}$$

$$B^{ab} = -2g^{ab}\bar{\Box}R + 2\bar{\nabla}^{b}\bar{\nabla}^{a}R; \tag{18c}$$

for the Lagrangian $L=R_{mn}R^{mn}$,

$$\tfrac{1}{2}R_{mn}R^{mn}g^{ab} - R^{am}R^{b}{}_{n} - R_{m}{}^{b}R^{ma} + B^{ab} = 0, \tag{19a}$$

$$\Lambda_{c}{}^{ab} = 2\bar{\nabla}_{c}R^{ab} - \delta^{a}_{c}\bar{\nabla}_{m}R^{mb} - \delta^{b}_{c}\bar{\nabla}_{m}R^{am}, \tag{19b}$$

$$B^{ab} = -\bar{\Box}R^{ab} + 2\bar{\nabla}_{m}\bar{\nabla}^{b}R^{am} - g^{ab}\bar{\nabla}_{n}\bar{\nabla}_{m}R^{mn}; \tag{19c}$$

for the Lagrangian $L=R_{mnrs}R^{mnrs}$,

$$\tfrac{1}{2}R_{mnrs}R^{mnrs}g^{ab} - 2R^{amnr}R^{b}{}_{mnr} + B^{ab} = 0, \tag{20a}$$

$$\Lambda_{a}{}^{bc} = 2\bar{\nabla}_{m}R_{a}{}^{bcm} + 2\bar{\nabla}_{m}R_{a}{}^{cbm}, \tag{20b}$$

$$B^{ab} = 4\bar{\nabla}_{n}\bar{\nabla}_{m}R^{ambn}; \tag{20c}$$

and for the Lagrangian $L=f(R)$,

$$\tfrac{1}{2}fg_{ab} - f'R_{(ab)} + B_{ab} = 0, \tag{21a}$$

$$(2g^{bc}\delta^{m}_{a} - g^{mc}\delta^{b}_{a} - g^{bm}\delta^{b}_{a})\bar{\nabla}_{m}f' = \Lambda_{a}{}^{bc}, \tag{21b}$$

$$B^{ab} = -g^{ab}\bar{\Box}f' + \bar{\nabla}^{b}\bar{\nabla}^{a}f'. \tag{21c}$$

It is straightforward to obtain the correspondence with the Hilbert case by substituting the B_{ab}'s in the first equation in each of these above cases. They read, respectively, as

$$\tfrac{1}{4}R^2 g^{ab} - RR^{ab} + \nabla^{b}\nabla^{a}R - g^{ab}\Box R = 0, \tag{22a}$$

$$\tfrac{1}{2}R_{mn}R^{mn}g^{ab} - 2R^{bman}R_{mn} + \nabla^{b}\nabla^{a}R - \Box R^{ab} - \tfrac{1}{2}\Box R g^{ab} = 0, \tag{22b}$$

$$\tfrac{1}{2}R_{mnrs}R^{mnrs}g^{ab} - 2R^{mnrb}R_{mnr}{}^{a} + 4\nabla_{n}\nabla_{m}R^{amnb} = 0, \tag{22c}$$

$$f'R_{(ab)} - \tfrac{1}{2}fg_{ab} - \nabla_{a}\nabla_{b}f' + g_{ab}\Box f' = 0. \tag{22d}$$

Strictly speaking, compared to the usual Hilbert variation, in all cases considered so far using the first-order formalism, one starts from a *different* Lagrangian defined in a *different* function space, follows a *different* method, but nonetheless ends up in the same set of field equations. This means that although the gravitational Lagrangians in a Hilbert type of variational procedure are treated as functions of $(x^{\mu}, g, \partial g, \ldots)$, whereas in a metric-connection formalism the corresponding function spaces become sets of points of the form $(x^{\mu},g,\partial g,\Gamma,\partial\Gamma,\ldots)$, the field equations are the same. We shall now see this is not a mere formal coincidence but that the reason lies in the fact that all our test Lagrangians are *diffeomorphism covariant*.

All previous cases can be considered as specializations derived from a very general Lagrangian n-form constructed locally as follows:

$$L=L(g_{ab},\tilde{\nabla}_{a_1}g_{ab},\ldots,\tilde{\nabla}_{(a_1}\cdots\tilde{\nabla}_{a_k)}g_{ab},\psi,\tilde{\nabla}_{a_1}\psi,\ldots,\tilde{\nabla}_{(a_1}\cdots\tilde{\nabla}_{a_l)}\psi,\gamma), \tag{23}$$

that is, L is a function of dynamical fields g,ψ and also of other fields collectively referred to as γ. Referring to g and ψ as ϕ, L is called f-covariant, $f\in\mathrm{Diff}(M)$, or simply diffeomorphism covariant, if

$$L(f^*(\phi))=f^*L(\phi), \tag{24}$$

where f^* denotes the induced action of f on the fields ϕ. Note that this definition excludes the action of f^* on $\tilde{\nabla}$ or the other fields γ. It immediately follows that our test Lagrangians considered previously satisfy the above definition and as a result are diffeomorphism covariant.

It is a very interesting result, first shown by Iyer and Wald (Ref. 21), that if L in (23) is diffeomorphism covariant, then L can be reexpressed in the form

$$L=L(g_{ab},R_{bcde},\nabla_{a_1}R_{bcde},\ldots,\nabla_{(a_1}\cdots\nabla_{a_m)}R_{bcde},\psi,\nabla_{a_1}\psi,\ldots,\nabla_{(a_1}\cdots\nabla_{a_l)}\psi), \tag{25}$$

where ∇ denotes the Levi-Civita connection of g_{ab}, R_{abcd} denotes the Riemann curvature of g_{ab}, and $m=\max(k-2,l-2)$. Notice that everything is expressed in terms of the Levi-Civita connection of the metric tensor and also that all other fields γ are absent.

Applying the Iyer–Wald theorem in our test Lagrangians, we immediately see that we could have reexpressed them from the beginning in a form that involves only the Levi-Civita connection and not the original arbitrary connection $\tilde{\nabla}$, and vary them to obtain the corresponding ''Hilbert'' equations. As we showed above, we arrived at this result by treating the associated Lagrangians as *different* (indeed they are!) according to whether or not they involved an arbitrary (symmetric) or a Levi-Civita connection.

III. CONFORMAL STRUCTURE AND WEYL GEOMETRY

For the more general nonlinear Lagrangians of the form $f(q)$, where $q=R$, $R_{ab}R^{ab}$, or $R_{abcd}R^{abcd}$, where f is an arbitrary smooth function considered in the previous section, the field equations obtained by the metric-connection formalism are of second order while the corresponding ones obtained via the usual metric variation are of fourth order. This result sounds very interesting since it could perhaps lead to an alternative way to ''cast'' the field equations of these theories in a more tractable, reduced form than the one that is usually used for this purpose, namely, the conformal equivalence theorem (Ref. 22). In this way, certain interpretational issues related to the question of the physicality of the two metrics (Refs. 23, 24) associated with the conformal transformation would perhaps be avoided.

As discussed in the previous section, the constraint (13) for Weyl geometry becomes

$$L_c(g,\Gamma,\Lambda)=\Lambda_c{}^{ab}\left[\Gamma^c_{ab}-\left\{\begin{matrix}c\\ ab\end{matrix}\right\}-\frac{1}{2}g^{cm}(Q_a g_{mb}+Q_b g_{am}-Q_m g_{ab})\right]. \tag{26}$$

In order to examine the consequences of the Weyl constraint (26), we now apply the constrained first-order formalism to the Lagrangian $L=f(R)$. Variation with respect to the Lagrange multipliers recovers the expression (12) of the Weyl connection. Variation with respect to the metric yields the g equations,

$$f'R_{(ab)}-\tfrac{1}{2}fg_{ab}+B_{ab}=0, \tag{27}$$

where B_{ab} is defined by (16). Variation with respect to the connection yields the explicit form of the Lagrange mutlipliers, namely,

$$\Lambda_c{}^{ab}=\tfrac{1}{2}\delta_c^b(Q^a f'-\bar\nabla^a f')+\tfrac{1}{2}\delta_c^a(Q^b f'-\bar\nabla^b f')-g^{ab}(Q_c f'-\bar\nabla_c f'). \tag{28}$$

(We note that since we are considering only variations with respect to the dependent variables g and Γ and also the Lagrange multipliers, and since the Γ's are functions of Q_a [from Eq. (12)], the Weyl vector field Q_a essentially plays the role of an independent variable and so we refrain from considering variations with respect to Q_a.) Substituting back this last result into Eq. (16), we find

$$B_{ab}=2Q_{(a}\bar\nabla_{b)}f'-\bar\nabla_{(a}\bar\nabla_{b)}f'+f'\bar\nabla_{(a}Q_{b)}-f'Q_aQ_b-g_{ab}(2Q_m\bar\nabla^m f'-Q^2f'-\bar\Box f'+f'\bar\nabla^m Q_m). \tag{29}$$

Inserting this result into Eq. (27) we obtain the full field equations for the Lagrangian $L=f(R)$ in the framework of Weyl geometry, namely,

$$f'R_{(ab)}-\tfrac{1}{2}fg_{ab}-\bar\nabla_a\bar\nabla_b f'+g_{ab}\bar\Box f'=M_{ab}, \tag{30}$$

where M_{ab} is defined by

$$M_{ab}=-2Q_{(a}\bar\nabla_{b)}f'-f'\bar\nabla_{(a}Q_{b)}+f'Q_aQ_b+g_{ab}(2Q_m\bar\nabla^m f'-Q^2f'+f'\bar\nabla^m Q_m). \tag{31}$$

It is interesting to note that the degenerate case $Q_a=0$ corresponds to the usual field equations obtained by the Hilbert variation in the framework of Riemann geometry, namely,

$$f'R_{ab}-\tfrac{1}{2}fg_{ab}-\nabla_a\nabla_b f'+g_{ab}\Box f'=0.$$

It is known that these equations are conformally equivalent to Einstein equations with a self-interacting scalar field as the matter source (Ref. 22). In what follows, we generalize this property of the $f(R)$ field equations in Weyl geometry. To this end, we define the metric $\tilde g$ conformally related to g with f' as the conformal factor. Under a conformal transformation, the Weyl vector field transforms as

$$\tilde Q_a=Q_a-\bar\nabla_a \ln f',$$

and the field equations (30) in the conformal frame read as

$$f'\tilde R_{(ab)}-\frac{1}{2}\frac{f}{f'}\tilde g_{ab}-\tilde{\bar\nabla}_a\tilde{\bar\nabla}_b f'+\tilde g_{ab}\bar\Box f'=\tilde M_{ab},$$

where $\tilde{\bar\nabla}=\bar\nabla$, $\bar\Box\tilde g^{ab}\tilde{\bar\nabla}_a\tilde{\bar\nabla}_b=(f')^{-1}\bar\Box$, and $\tilde M_{ab}$ is given by

$$\tilde M_{ab}=f'\tilde Q_a\tilde Q_b-f'\tilde{\bar\nabla}_{(a}\tilde Q_{b)}-\tilde{\bar\nabla}_a\tilde{\bar\nabla}_b f'+\tilde g_{ab}(f'\tilde{\bar\nabla}^m\tilde Q_m-f'\tilde Q^2+\bar\Box f'). \tag{32}$$

Introducing the scalar field $\varphi=\ln f'$ and the potential $V(\varphi)$ in the ''usual'' form (Ref. 22),

$$V(\varphi)=\tfrac{1}{2}(f'(R))^{-2}(Rf'(R)-f(R)), \tag{33}$$

we find that the field equations take the final form

$$\tilde G_{ab}=\tilde M_{ab}^{Q}-\tilde g_{ab}V(\varphi), \tag{34}$$

where

$$\tilde G_{ab}=\tilde R_{(ab)}-\tfrac{1}{2}\tilde R\tilde g_{ab},$$

and

$$\tilde{M}^{Q}_{ab}=\tilde{Q}_a\tilde{Q}_b-\tilde{\nabla}_{(a}\tilde{Q}_{b)}+\tilde{g}_{ab}(\tilde{\nabla}^m\tilde{Q}_m-\tilde{Q}^2).$$

The field equations (34) are Einstein equations for a self-interacting scalar field matter source with a potential $V(\varphi)$ and a source term $\tilde{M}^{Q}_{ab}$ depending on the field $\tilde{Q}_a$. If the geometry is Riemannian, i.e., $\tilde{Q}_a=0$, one recovers the standard unconstrained variation result. This will be the case only if the original Weyl vector field is a gradient, $Q_a=\nabla_a\Phi$. Then it can be gauged away by the conformal transformation $\tilde{g}_{ab}=(\exp\Phi)g_{ab}$, and therefore the original space is not a general Weyl space but a Riemann space with an undetermined gauge (Ref. 25). We saw an example of this previously when we applied the unconstrained method to the Lagrangian $L=f(R)$. Here, the Weyl vector was deduced using Eq. (6) and turned out to be $Q_a=\nabla_a(\ln f')$. (In Ref. 20, this peculiarity is used in order to find out a subclass of theories based on a general *D*-dimensional dilaton gravity action, for which both unconstrained method and Hilbert variation yield dynamically equivalent systems.) This fact shows that unconstrained variations cannot deal with a general Weyl geometry and correspond to a degenerate case of the constrained method—the field equations obtained from the former can be recovered only by choosing specific forms of the Weyl vector field (Ref. 26).

IV. DISCUSSION

The results obtained in Sec. II and Sec. III have the interpretation that a consistent way to investigate generalized theories of gravity without imposing from the beginning that the geometry is Riemannian, is the constrained first-order formalism. Applications to quadratic and $f(R)$ Lagrangians in the framework of Riemannian and Weyl geometry reveal that unconstrained variational methods are degenerate cases corresponding to a particular gauge and that the usual conformal structure can be recovered in the limit of a vanishing Weyl vector.

The generalization of the result stated above to include arbitrary connections with torsion can be an interesting exercise. The physical interpretation of the source term [Eqs. (30) and (31)] is closely related to the choice of the Weyl vector field *Q*. However, it cannot be interpreted as a genuine stress–energy tensor in general since, for instance, choosing *Q* to be a unit timelike, hypersurface-orthogonal vector field, the sign of $M_{ab}Q^aQ^b$ depends on the signs of $f'(R)$ and the ''expansion'' ∇_aQ^a.

The generalization of the conformal equivalence theorem presented in Sec. III opens the way to analyzing cosmology in the framework of these Weyl $f(R)$ theories by methods such as those used in the traditional Riemannian case. The first steps in such a program may be as follows (Ref. 27).

(a) Analyze the structure and properties of Friedmann cosmologies, find their singularity structure, and examine the possibility of inflation.
(b) Consider the past and future asymptotic states of Bianchi cosmologies. Examine isotropization and recollapse conjectures in such universes. Look for chaotic behavior in the Bianchi VIII and IX spacetimes.
(c) Formulate and prove singularity theorems in this framework. This will differ from the analysis in the Riemannian case (cf. Ref. 22) because of the presence of the source term M_{ab}.

All the problems discussed above can be tackled by leaving the conformal Weyl vector field $\tilde{Q}_a$ undetermined while setting it to zero at the end will lead to detailed comparisons with the results already known in the Riemannian case.

ACKNOWLEDGMENTS

The work of S.C. and J.M. was supported by Grants No. 3516075-2 and No. 1361/4.1, received from the General Secretariat of Science and Technology and the Research Commission of the University of the Aegean, respectively, which are gratefully acknowledged. J.M. is grateful to the ''Institut d'Astrophysique et de Géophysique de l'Université de Liège'' for their kind hospitality while part of this work was done. L.Q. received support from a F.R.I.A. scholarship, from F.N.R.S. and A.R.C. grants. L.Q. wishes to acknowledge the kind hospitality of the Department of Mathematics of the University of the Aegean, where part of this work was done.

[1] F. W. Hehl, J. D. McCrea, E. W. Mielke, and Y. Ne'eman, Phys. Rep. **258**, 1 (1995).
[2] H. Weyl, *Space–Time-Matter* (Dover, New York 1952).
[3] M. Ferraris, M. Francaviglia, and C. Reina, Gen. Relativ. Gravit. **14**, 243 (1982).
[4] A. S. Eddington, *The Mathematical Theory of Relativity* (Cambridge University Press, Cambridge, 1922).
[5] G. Stephenson, Nuovo Cimento **9**, 1959 (1958).
[6] G. Stephenson, Nuovo Cimento **9**, 2085 (1958).
[7] P. Higgs, Nuovo Cimento **11**, 816 (1958).
[8] C. N. Yang, Phys. Rev. Lett. **33**, 445 (1974).
[9] H. Buchdahl, Proc. Cambridge Philos. Soc. **56**, 396 (1960).
[10] H. Buchdahl, J. Phys. A **12**, 1229 (1979).
[11] N. Van den Bergh, J. Math. Phys. **22**, 2245 (1981).
[12] B. Shahid-Saless, Phys. Rev. D **35**, 467 (1987).
[13] V. Hamity and D. Barraco, Gen. Relativ. Gravit. **25**, 461 (1993).
[14] M. Ferraris, M. Francaviglia, and I. Volovich, Class. Quantum Grav. **11**, 1505 (1994).
[15] A. Borowiec, M. Ferraris, M. Francaviglia, and I. Volovich, Class. Quantum Grav. **15**, 43 (1998).
[16] J. R. Ray, Nuovo Cimento B **25**, 706 (1975).
[17] J. Safko and F. Elston, J. Math. Phys. **17**, 1531 (1976).
[18] S. Deser, ''Inequivalence of first and second order formulations in $D=2$ gravity models,'' gr-qc/9512022, 1995. See also references therein.
[19] S. Davis, Gen. Relativ. Gravit. **30**, 345 (1998).
[20] H. Burton and R. B. Mann, Class. Quantum Grav. **15**, 1375 (1998).
[21] V. Iyer and R. M. Wald, Phys. Rev. D **50**, 846 (1994) (cf. Lemma 2.1).
[22] J. Barrow and S. Cotsakis, Phys. Lett. B **214**, 515 (1988).
[23] S. Cotsakis, Phys. Rev. D **47**, 1437 (1993); **49**, 1145 (1994) (erratum).
[24] S. Cotsakis, Phys. Rev. D **52**, 6199 (1995).
[25] J. A. Schouten, *Ricci Calculus* (Springer-Verlag, Berlin, 1954), p. 134.
[26] L. Querella, ''Constrained variations and nonlinear gravitational Lagrangians,'' Talk given at GR 15, The 15th Meeting of the International Society on General Relativity and Gravitation, Poona, India, 16–21 December 1997. See also R. Mann, ''Summary of session A6, Alternative theories of gravity,'' gr-qc/9803051.
[27] S. Cotsakis, ''Mathematical problems in higher order gravity and cosmology,'' talk given at the 8th Marcel Grossmann Meeting on Recent Developments in Theoretical and Experimental General Relativity, Gravitation and Relativistic Field Theories (MG 8), Jerusalem, Israel, 22–27 June 1997, gr-qc/9712046. See also V. N. Melnikov, ''Multidimensional gravity and Cosmology: Exact solutions,'' gr-qc/9801037.

Random walk with memory

Ryszard Rudnicki
Institute of Mathematics, Polish Academy of Sciences, Staromiejska 8, 40-013 Katowice, Poland and Institute of Mathematics, Silesian University, Bankowa 14, 40-007 Katowice, Poland

Marek Wolf[a)]
Institute of Theoretical Physics, University of Wrocław, PI. Maxa Borna 9, PL-50-204 Wrocław, Poland

(Received 16 June 1998, accepted for publication 17 November 1998)

A reinforced random walk on the d-dimensional lattice is considered. It is shown that this walk is equivalent to an iterated function system (IFS). Criteria for the existence of limit cycles are given. Numerical results and conjectures about the quantitative behavior of the walk are stated. © *1999 American Institute of Physics.* [S0022-2488(99)00905-6]

I. INTRODUCTION

There are a large number of different modifications and variants of the usual symmetrical random walk (RW).[1–3] Let us mention only Levy flights, biased diffusions, self-avoiding walk (SAW for short), etc. Let us confine ourselves to the random walks on the discrete lattices. In SAW a walking particle is choosing its trajectory in such a way that it does not step down onto the already visited site. If a particle runs into such a node that all neighboring sites were already visited, it stops. In Ref. 4 the interacting RW was discussed in which the parameter $0<p<\infty$ has influenced probabilities of visiting a given site and $p=1$ corresponds to usual RW. For $p\rightarrow\infty$ this RW goes on into the SAW.

In 1987 Coppersmith and Diaconis[5] introduced reinforced random walk (RRW). This walk, opposite to SAW, prefers earliest visited paths. Pemantle[6] discussed a related process on trees and proved that this process is equivalent to a random walk in a random environment. He also gave criteria for transience and recurrence of RRW. Davis[7] considered a variety of types of RRW on the integers **Z**. One of them was RRW of sequence type. This process is defined in the following way. Let w_k be an increasing sequence of non-negative numbers. Let (X_n) be a random motion on **Z**. If some interval was traversed k-times, then its weights is w_k. If $X_n=i$, then the probability that $X_{n+1}=i-1$ or $X_{n+1}=i+1$ is proportional to the weights at time n of the intervals $(i-1,i)$ and $(i,i+1)$. Davis proved that the moving point visits a finite number of integers and eventually oscillates between two adjacent integers if and only if $\Sigma_{k=0}^{\infty}w_k^{-1}<\infty$. This result was generalized to RRW sequence type on the d-dimensional lattice by Sellke.[8]

In this paper we consider another type of reinforced random walk on the d-dimensional lattice. The random point moves according to the following reinforcement convention. Let the moving point be found at time $t=n$ at a certain point $A\in\mathbf{Z}^n$. Let $p_1,\ldots,p_N$ be the probabilities of choosing one of the adjacent points $A_1,\ldots,A_N$. Assume that we choose the point A_{i_0}. If after some time the moving point returns to A, then the probabilities that at the next step it can be found at the adjacent points are equal to $p_1',\ldots,p_N'$. The values of $p_1',\ldots,p_N'$ depend on the previous values $p_1,\ldots,p_N$ and i_0. We assume that the probability of choosing a given path will increase when it was already traversed and probabilities of remaining paths emanating from a given site will decrease. In other words, the fact that some sites were already visited will be remembered. The memory of passing particular edges will be encoded in the change of probabilities. At some time

[a)]Electronic mail: mwolf@ift.uni.wroc.pl

the probability of going in some direction from a given site will reach almost 1, while probabilities to go in other directions will be practically zero. It will result in closed paths: a random walker will oscillate between a few sites with practically zero probability to escape from such a limiting cycle. We will treat such a final behavior as stopping of the random walk.

Our walk is not Markovian because the probability of choosing any direction changes in time. If we extend the phase space by adding the distributions of probabilities of passing particular edges, we obtain a Markov process which is also an iterated function system.[9] In Refs. 10 and 11 were introduced self-attracting diffusions: processes attracted by their own trajectories. It is interesting that these processes and our walk have similar features. For example, self-attracting diffusions are not Markovian, but jointly with their occupation measures are Markov processes. Moreover, their trajectories converge almost surely.

The paper is organized as follows. First we define our random walk and the notion of the limit cycle is introduced. Next we prove the theorem that this walk reaches the limit cycle. In Sec. IV the result of the Monte Carlo simulations for a particular ''memorizing'' function are presented. These computer experiments allow us to make some conjecture about the quantitative behavior of some characteristics of the walk.

II. MATHEMATICAL MODEL

A. Description of the random walk with memory

Let $\mathcal{Z}$ denote the set of all points of the d-dimensional Euclidean space which have integer coordinates, i.e., $\mathcal{Z}$ is the d-dimensional lattice. A point moves randomly over this lattice. It starts at point $\mathbf{0}=(0,...,0)$. If at time $t=n$ the moving point can be found at a certain point $x=(x_1,\ldots,x_d)$, then at the time $t=n+1$ it can be found at one of the $N=2^d$ adjacent points $y=(y_1,\ldots,y_d)$, where $y_i=x_i+1$ or $y_i=x_i-1$ for $i=1,2,...,d$. By $\mathcal{K}$ we denote the set of all possible ''steps'' during the walk, i.e.,

$$\mathcal{K}=\{(x,y)\in\mathcal{Z}\times\mathcal{Z}:|y_i-x_i|=1 \quad \text{for } i=1,...,d\}.$$

If $z=(x,y)\in\mathcal{K}$, then the points x and y are, respectively, the beginning and the end of the step z. Let $\mathcal{S}=\{1,-1\}^d$ be the set of all N steps to the nearest neighbors. If $x\in\mathcal{Z}$, $s\in\mathcal{S}$, and $y=x+s$, then $(x,y)\in\mathcal{K}$.

At time $t=0$ the probability of choosing of any adjacent point equals 2^{-d}. During the walk the point ''memorizes'' its path in the following way. Assume that the moving random walker can be found at some time $t=n$ at a certain point $x\in\mathcal{Z}$. Let $p_{x,x+s}$, $s\in\mathcal{S}$, be the probabilities that at the time $t=n+1$ it can be found at one of the adjacent points $x+s$, $s\in\mathcal{S}$. Assume that we choose the point $x+s_0$ to shift the particle from the point x. If after some time the moving point returns to x, then the probabilities that at the next step it can be found at the adjacent points are equal to $p'_{x,x+s}$, $s\in\mathcal{S}$. The numbers $p'_{x,x+s}$ are related to the previous values $p_{x,x+s}$ in the following way. Let $\mathbf{p}_x=(p_{x,x+s})_{s\in\mathcal{S}}$, $\mathbf{p}'_x=(p'_{x,x+s})_{s\in\mathcal{S}}$, and

$$\mathbf{P}=\left\{\mathbf{p}_x\in[0,1]^{\mathcal{S}}:\sum_{s\in\mathcal{S}}p_{x,x+s}=1\right\}.$$

Then

$$\mathbf{p}'_x=f_{s_0}(\mathbf{p}_x), \tag{1}$$

where $f_{s_0}:\mathbf{P}\to\mathbf{P}$ is a continuous function.

If $z\in\mathcal{K}$ and $z=(x,y)$, we will often write p_z instead of $p_{x,y}$.

B. Iterated function system

The state of the random walk with memory is described at any time t by the position of the moving point and the probability p_z for every $z \in \mathcal{K}$. Let $\mathcal{P}$ denote the set of all admissible distributions of probabilities p_z, $z \in \mathcal{K}$, i.e.,

$$\mathcal{P} = \left\{ p \in [0,1]^{\mathcal{K}} : \sum_{s \in \mathcal{S}} p_{x,x+s} = 1 \quad \text{for each} \quad x \in \mathcal{Z} \right\}.$$

Then the phase space is the set $\mathcal{X} = \mathcal{Z} \times \mathcal{P}$. Since $\mathbf{p}_x = (p_{x,x+s})_{s \in \mathcal{S}}$ for every $x \in \mathcal{Z}$, we have $\mathcal{P} = \mathbf{P}^{\mathcal{Z}}$ and $\mathcal{X} = \mathcal{Z} \times \mathbf{P}^{\mathcal{Z}}$.

Now, we define an iterated function[9] system on the phase space $\mathcal{X}$. It consists of N transformations $T_s : \mathcal{X} \to \mathcal{X}$, $s \in \mathcal{S}$. These transformations are defined as follows. Let $x \in \mathcal{Z}$ and $s \in \mathcal{S}$ be given. Then $T_s(x,p) = (x+s, p')$, where

$$\mathbf{p}'_z = \begin{cases} \mathbf{p}_z, & \text{if } z_1 \neq x, \\ f_s(\mathbf{p}_z), & \text{if } z_1 = x, \end{cases}$$

for each $z = (z_1, z_2) \in \mathcal{K}$. If (x,p) is the state of the random walk at a time t and if the next position of the moving point is $x+s$, then $T_s(x,p)$ is the next state of the random walk. The probability that at a point $\chi = (x,p)$ we choose the transformation T_s is equal $p_s(\chi) = p_{x,x+s}$.

C. Markov process on $\mathcal{X}$

Now we construct a Markov process corresponding to the iterated function system given in Sec. II B.

The phase space $\mathcal{X}$ is a metric space with some metric ρ defined as follows. The set $\mathcal{K}$ is countable, that is, $\mathcal{K} = \{z_1, z_2, z_3, \ldots\}$, where $\{z_n\}_{n \in \mathbf{N}}$, $z_n \in \mathcal{K}$, is the sequence of all possible steps. If $x \in \mathcal{X}$, $y \in \mathcal{X}$, then $x = (u, p_{z_1}, p_{z_2}, \ldots)$ and $y = (u', p'_{z_1}, p'_{z_2}, \ldots)$, where u, $u' \in \mathcal{Z}$ and $p_{z_k}, p'_{z_k} \in [0,1]$ for each $k \in \mathbf{N}$. The metric ρ is given by

$$\rho(x,y) = |u - u'| + \sum_{k=1}^{\infty} 2^{-k} |p_{z_k} - p'_{z_k}|.$$

Let $\mathcal{B}$ be the σ-algebra of Borel subsets of $\mathcal{X}$. For any $x \in \mathcal{X}$ and $A \in \mathcal{B}$ we set

$$I(x,A) = \{s \in \mathcal{S} : T_s(x) \in A\}, \quad P(x,A) = \sum_{s \in I(x,A)} p_s(x).$$

Then $P(x,A)$ is a transition probability function, i.e.,

(a) for each $x \in \mathcal{X}$ the function $A \mapsto P(x,A)$ is a probabilistic measure and
(b) for each $A \in \mathcal{B}$ the function $x \mapsto P(x,A)$ is $\mathcal{B}$-measurable.

Since the space $\mathcal{X}$ is σ-compact, there exists a homogeneous Markov process $\{\xi_n\}_{n=0}^{\infty}$ which corresponds to the transition function $P(x,A)$.[12] It means that we have some probability space $(\Omega, \mathcal{A}, \text{Prob})$ and a sequence $\{\xi_n\}_{n=0}^{\infty}$ of random elements $\xi_n : \Omega \to \mathcal{X}$ such that the sequence ξ_n is a Markov process and

$$\text{Prob}\,(\xi_{n+1} \in A | \xi_n = x) = P(x,A)$$

for each $x \in \mathcal{X}$, $A \in \mathcal{B}$, $n \geqslant 0$. Since the initial state of the system is $x_0 = (\mathbf{0}, 2^{-d}, 2^{-d}, \ldots)$, we assume that $\xi_0 = x_0$.

We assume that the probability space $(\Omega, \mathcal{A}, \text{Prob})$ is complete, i.e., if A is a measurable set and $\text{Prob}\,(A) = 0$, then every subset of A is measurable.

D. Limit cycle

A sequence $(u_0,u_1,\ldots,u_{m-1})$ of different elements of $\mathcal{Z}$ is called a cycle if there exists a sequence $(s_0,s_1,\ldots,s_{m-1})$ of elements of $\mathcal{S}$ such that $u_{k+1}=u_k+s_k$ for $k=0,...,m-1$, where $u_m=u_0$. Let $\Pi_0:\mathcal{X}\to\mathcal{Z}$ be the operator given by $\Pi_0(u,p)=u$. We say that a sequence $(x_n)_{n=0}^{\infty}$ of elements of $\mathcal{X}$ has a limit cycle if there exist a cycle $(u_0,u_1,\ldots,u_{m-1})$ and an integer $n_0\geqslant 0$ such that for every $n\geqslant n_0$ we have $\Pi_0(x_n)=u_k$, where $k=n(\mathrm{mod}\, m)$.

Let $\psi^0(t)=l$ and $\psi^n(t)=\psi\circ\psi^{n-1}(t)$.

Now we can formulate the following theorem.

Theorem 1: *Let* $\psi:[0,1]\to[0,1]$ *be a continuous nondecreasing function such that* $\psi(1)=1$ *and*

$$\sum_{n=0}^{\infty}\left[1-\psi^n\left(\frac{1}{N}\right)\right]<\infty. \tag{2}$$

We assume that there exists a continuous function $\psi:[0,1]\to[0,1]$ *such that*

$$f_{s,s}(\mathbf{p}_x)\geqslant\psi(p_{x,x+s}) \tag{3}$$

for $s\in\mathcal{S}$, *where* $f_{s,s}$ *is the s-th coordinate of* f_s. *Then there exists a measurable subset* $\Omega_0\subset\Omega$ *such that* $\mathrm{Prob}\,(\Omega_0)=1$ *and for each* $\omega\in\Omega_0$ *the sequence* $\{\xi_n(\omega)\}_{n=0}^{\infty}$ *has a limit cycle.*

The proof of Theorem 1 is given in Sec. III.

Remark 1: Let Π_z, $z\in\mathcal{K}$, be the operator $\Pi_z:\mathcal{X}\to[0,1]$ given by $\Pi_z(u,p)=p_z$. Assume that the sequence $\{\xi_n(\omega)\}_{n=0}^{\infty}$ has the limit cycle $(u_0,u_1,\ldots u_{m-1})$. Then from Theorem 1 it follows that $\lim_{n\to\infty}\Pi_z(\xi_n(\omega))=1$ for each $z=(u_k,u_{k+1})$, $k=0,...,m-1$.

Remark 2: If $\psi:[0,1]\to[0,1]$ is a continuous nondecreasing function such that $\psi(x)>x$ for $x\in(0,1)$ and $\psi'(1)>1$, then ψ satisfies (2)

III. PROOF OF THEOREM 1

A. Boundedness of trajectories

The thread of the proof of Theorem 1 is as follows. First we check that almost all paths are bounded. From this it follows that a point performing a random walk returns infinitely often to some points of the lattice $\mathcal{Z}$. Then we show that if a point $u\in\mathcal{Z}$ is visited infinitely often, then after some time the random walker chooses a fixed adjacent point to u. This implies that the random walk has a limit cycle.

Let $\eta_n(\omega)=\Pi_0(\xi_n(\omega))$. Then the random variable η_n describes the position of the moving point at time $t=n$.

Proposition 1: For almost all ω *the sequence* $\{\eta_n(\omega)\}$ *is bounded.*

We precede the proof of Proposition 1 with the following lemmas.

Lemma 1: Let $\varphi(t)=\Pi_{n=0}^{\infty}\psi^n(t)$. *Then* $\varphi(t)>0$ *for* $t\geqslant 1/N$ *and*

$$\lim_{t\to 1}\varphi(t)=1. \tag{4}$$

Proof: Since ψ is a nondecreasing function, from (2) it follows that

$$\varphi(t)\geqslant\varphi\left(\frac{1}{N}\right)<\infty \quad \text{for } t\in\left[\frac{1}{N},1\right]. \tag{5}$$

Let $\varepsilon>0$ be given. Since $\varphi(1/N)>0$ there is an integer n_0 such that

$$\prod_{n=n_0}^{\infty}\psi^n(t)\geqslant\prod_{n=n_0}^{\infty}\psi^n\left(\frac{1}{N}\right)>1-\varepsilon \quad \text{for } t\in\left[\frac{1}{N},1\right]. \tag{6}$$

The function ψ is continuous and $\psi(1)=1$. This implies that there is $\delta>0$ such that

$$\prod_{n=0}^{n_0-1} \psi^n(t)>1-\varepsilon \quad \text{for } t\in[1-\delta,1]. \tag{7}$$

From (6) and (7) it follows that $\varphi(t)>(1-\varepsilon)^2$ for $t\in[1-\delta,1]$. Consequently, $\lim_{t\to 1}\varphi(t)=1$.□

Lemma 2: Let $x\in\mathcal{Z}$ *and* $y\in\mathcal{Z}$ *be two adjacent points. Denote by A the event that the point x is visited for the first at time* $t=n_0$ *and the point y was not visited earlier. Let*

$$B=\{\omega\in\Omega:\eta_{n_0+2i-1}(\omega)=y,\quad \eta_{n_0+2i}(\omega)=x,\quad \text{for } i=1,2,...\}.$$

Then the conditional probability $\text{Prob}\,(B|A)$ *satisfies*

$$\text{Prob}\,(B|A)\geqslant\left[\varphi\left(\frac{1}{N}\right)\right]^2.$$

Proof: Let $B_0=A$ and

$$B_k=\{\omega\in\Omega:\eta_{n_0+2i-1}(\omega)=y,\quad \eta_{n_0+2i}(\omega)=x,\quad \text{for } i=1,2,...,k\}.$$

If $\omega\in B_k\cap A$, then at each time $n_0\leqslant t<n_0+2k$ and at each visit at x we have chosen the next point y and at each visit at y we have chosen the next point x. Let $s=y-x$ and $s'=x-y$. Then

$$\text{Prob}\,(\eta_{n_0+2k+1}=y|B_k\cap A)=f_{s,s}^k\left(\frac{1}{N}\right),$$

$$\text{Prob}\,(\eta_{n_0+2k+2}=x|B_k\cap A\cap\{\eta_{n_0+2k+1}=y\})=f_{s',s'}^k\left(\frac{1}{N}\right),$$

where f_s^k is the kth iterate of f_s and $f_{s,s}^k$ is the sth coordinate of f_s^k. Since ψ is a nondecreasing function, from (3) we obtain

$$f_{s,s}^k\left(\frac{1}{N}\right)\geqslant\psi\left(f_{s,s}^{k-1}\left(\frac{1}{N}\right)\right)\geqslant\psi^2\left(f_{s,s}^{k-2}\left(\frac{1}{N}\right)\right)\geqslant\cdots\geqslant\psi^k\left(\frac{1}{N}\right),$$

$$f_{s',s'}^k\left(\frac{1}{N}\right)\geqslant\psi^k\left(\frac{1}{N}\right).$$

Consequently,

$$\text{Prob}\,(B_{k+1}|B_k\cap A)\geqslant\left[\psi^k\left(\frac{1}{N}\right)\right]^2. \tag{8}$$

From (8) it follows that

$$\text{Prob}\,(B_{k+1}\cap A)\geqslant\text{Prob}\,(A)\left[\prod_{i=0}^{k}\psi^i\left(\frac{1}{N}\right)\right]^2. \tag{9}$$

If $k\to\infty$, then we obtain

$$\text{Prob}\,(B\cap A)\geqslant\text{Prob}\,(A)\left[\varphi\left(\frac{1}{N}\right)\right]^2,$$

and finally

$$\text{Prob}\,(B|A) \geqslant \left[\varphi\left(\frac{1}{N}\right)\right]^2 > 0. \qquad \square$$

Proof of Proposition 1: For $x \in \mathbf{R}^d$ we set

$$\|x\| = \max\{|x_i| : i = 1,...,d\}.$$

For $k = 1,2,...$, we define

$$C_k = \{\omega \in \Omega : \sup_n \|\eta_n \omega)\| \geqslant k\}.$$

According to Lemma 2,

$$\text{Prob}\,(C_{k+1}|C_k) \leqslant 1 - \left[\varphi\left(\frac{1}{N}\right)\right]^2. \tag{10}$$

Indeed, let $n_0(\omega)$ be the first time such that $\|\eta_{n_0}(\omega)\| = k$. If $x = \eta_{n_0}$ and y is an adjacent point to x such that $\|y\| = k+1$, then with probability $p \geqslant [\varphi(1/N)]^2$ the moving point visits only x and y at any time $t > n_0$. This implies (10). From (10) it follows that

$$\text{Prob}\,(C_k) \leqslant \left\{1 - \left[\varphi\left(\frac{1}{N}\right)\right]^2\right\}^{k-1}. \tag{11}$$

Let $C = \cap_{k=1}^{\infty} C_k$. Since a trajectory $\eta_n(\omega)$ is unbounded if and only if $\omega \in C$, from (11) it follows that almost all trajectories are bounded. $\square$

B. Stabilization of directions

From Proposition 1 it follows that almost all trajectories are bounded. Consequently, the moving point visits some points of the lattice $\mathcal{Z}$ infinitely often. Let a point $x \in \mathcal{Z}$ be given. By A we denote the event that the point x is visited infinitely often. For any $\omega \in A$ we denote by $\{k_n(\omega)\}_{n=1}^{\infty}$ successive times of visits at point x. Let $x^n(\omega)$ be the adjacent point to x visited at time $t = k_n(\omega) + 1$. We show that for almost every $\omega \in A$ there exists a point $y(\omega) \in \mathcal{Z}$ such that $x^n(\omega) = y(\omega)$ for $n > n_0(\omega)$.

The process of choosing the adjacent points can be described as an iterated function system $(f_s)_{s \in \mathcal{S}}$ on the space $\mathbf{P}$ and the probability that at the point $\mathbf{p}_x \in \mathbf{P}$ we choose the transformation f_{s_0} equals p_{x,s_0}. Indeed, let us assume that we visit the point x and let $\mathbf{p}_x = (p_{x,s})_{s \in \mathcal{S}}$ be the distribution of probability of choosing adjacent points $x+s$, $s \in \mathcal{S}$. If we choose the point $x + s_0$, then at the next visit at x, $\mathbf{p}'_x = f_{s_0}(\mathbf{p}_x)$ is the new distribution of probability of choosing adjacent points. Since x is a given point we will write $\mathbf{p}$ instead of $\mathbf{p}_x$ and p_s instead of $p_{s,x}$.

Let $\{\zeta_n\}_{n=1}^{\infty}$ be a homogeneous Markov process on the phase space $\mathbf{P}$ corresponding to the iterated function system $(f_s)_{s \in \mathcal{S}}$. The transition probability function for the process (ζ_n) is given by the formula

$$P(\mathbf{p},A) = \sum_{s \in I} \mathbf{p}_s, \quad \text{where } I = \{s \in \mathcal{S} : f_s(\mathbf{p}) \in A\}. \tag{12}$$

Denote by $x + s_n(\omega)$ the adjacent point chosen at time $t = n$. Then $\zeta_1 = (1/N,\dots,1/N)$, $\zeta_{n+1} = f_{s_n}(\zeta_n)$, and

$$\text{Prob}\,(s_n = s | \zeta_n = \mathbf{p}) = p_s. \tag{13}$$

for $\mathbf{p} \in \mathbf{P}$, $n = 1,2,...$, and $s \in \mathcal{S}$.

Proposition 2: For almost every ω *there is* $s(\omega)\in\mathcal{S}$ *and an integer* $n_0(\omega)$ *such that* $s_n(\omega)=s(\omega)$ *for* $n>n_0(\omega)$.

We precede the proof of Proposition 2 by a lemma. Let

$$\mathbf{P}_\epsilon^s=\{\mathbf{p}\in\mathbf{P}:p_s\geqslant\varepsilon\},$$

$$\mathbf{P}_\epsilon=\bigcup_{s\in\mathcal{S}}\mathbf{P}_\epsilon^s.$$

Lemma 3: Let $P^n(\mathbf{p},A)$ *be the n-step transition probability function. Then for every* $\delta<1$ *and* $\mathbf{p}\in\mathbf{P}$ *we have*

$$\lim_{n\to\infty}P^n(\mathbf{p},\mathbf{P}_\delta)=1. \tag{14}$$

Proof: Let $\varepsilon<1$ be given. First we check that for $\mathbf{p}\in\mathbf{P}_\varepsilon$ we have

$$P(\mathbf{p},\mathbf{P}_{\psi(\varepsilon)})\geqslant\varepsilon. \tag{15}$$

Indeed, if $\mathbf{p}\in\mathbf{P}_\varepsilon$, then for some s we have $\mathbf{p}\in\mathbf{P}_\varepsilon^s$ and $\mathbf{p}_s\geqslant\varepsilon$. Consequently, $f_s(\mathbf{p})\in\mathbf{P}^s_{\psi(\varepsilon)}$ and inequality (15) follows immediately from (12). From (15) it follows that

$$P^{i+1}(\mathbf{p},\mathbf{P}_{\psi^{i+1}(\varepsilon)})\geqslant\int_{\mathbf{P}_{\psi^i(\varepsilon)}}P(\mathbf{q},\mathbf{P}_{\psi^{i+1}(\varepsilon)})P^i(\mathbf{p},d\mathbf{q})\geqslant\psi^i(\varepsilon)P^i(\mathbf{p},\mathbf{P}_{\psi^i(\varepsilon)}),$$

which gives

$$P^n(\mathbf{p},\mathbf{P}_{\psi^n(\varepsilon)})\geqslant\prod_{k=0}^{n-1}\psi^k(\varepsilon)\geqslant\varphi(\varepsilon)\quad\text{for }\mathbf{p}\in\mathbf{P}_\varepsilon. \tag{16}$$

If $\varepsilon\geqslant 1/N$, then $\lim_{n\to\infty}\psi^n(\varepsilon)=1$. From (16) it follows that for every $\delta<1$, $\varepsilon\geqslant 1/N$, and $\mathbf{p}\in\mathbf{P}_\varepsilon$ we have

$$\liminf_{n\to\infty}P^n(\mathbf{p},\mathbf{P}_\delta)\geqslant\varphi(\varepsilon). \tag{17}$$

If $\mathbf{p}\in\mathbf{P}$, then $\mathbf{p}\in\mathbf{P}_{1/N}$ and, consequently,

$$\liminf_{n\to\infty}P^n(\mathbf{p},\mathbf{P}_\delta)\geqslant\varphi\left(\frac{1}{N}\right). \tag{18}$$

Since

$$P^{n+m}(\mathbf{p},\mathbf{P}_\delta)=\int_{\mathbf{P}}P^n(\mathbf{q},\mathbf{P}_\delta)P^m(\mathbf{p},d\mathbf{q})=\int_{\mathbf{P}_\delta}P^n(\mathbf{q},\mathbf{P}_\delta)P^m(\mathbf{p},d\mathbf{q})+\int_{\mathbf{P}\setminus\mathbf{P}_\delta}P^n(\mathbf{q},\mathbf{P}_\delta)P^m(\mathbf{p},d\mathbf{q}),$$

the inequalities (17) and (18) imply

$$\liminf_{n\to\infty}P^{n+m}(\mathbf{p},\mathbf{P}_\delta)\geqslant\varphi(\delta)P^m(\mathbf{p},\mathbf{P}_\delta)+\varphi\left(\frac{1}{N}\right)P^m(\mathbf{p},\mathbf{P}\setminus\mathbf{P}_\delta)=\left(\varphi(\delta)-\varphi\left(\frac{1}{N}\right)\right)P^m(\mathbf{p},\mathbf{P}_\delta)+\varphi\left(\frac{1}{N}\right). \tag{19}$$

Set

$$\alpha(\delta)=\liminf_{n\to\infty}P^n(\mathbf{p},\mathbf{P}_\delta).$$

Then from (19) it follows

$$\alpha(\delta) \geqslant \left(\varphi(\delta) - \varphi\left(\frac{1}{N}\right)\right)\alpha(\delta) + \varphi\left(\frac{1}{N}\right). \tag{20}$$

Since $\alpha(\delta)$ is a nonincreasing function there exists the limit $\lim_{\delta\to 1}\alpha(\delta) = \alpha_0$. According to Lemma 1, $\lim_{\delta\to 1}\varphi(\delta) = 1$. A passage to the limit $\delta \to 1$ in inequality (20) gives

$$\alpha_0 \geqslant \left(1 - \varphi\left(\frac{1}{N}\right)\right)\alpha_0 + \varphi\left(\frac{1}{N}\right). \tag{21}$$

From (21) we conclude that $\alpha_0 \geqslant 1$ and (14) holds. □

Proof of Proposition 2: Let $\delta \in (1/N,1)$ be a given number. Since

$$\text{Prob}\,(\zeta_n \in \mathbf{P}_\delta) = P^{n-1}(\zeta_1, \mathbf{P}_\delta),$$

from Lemma 3 it follows that there exists n_0 such that

$$\text{Prob}\,(\zeta_{n_0} \in \mathbf{P}_\delta) \geqslant \delta. \tag{22}$$

Let $A = \{\omega : \zeta_{n_0} \in \mathbf{P}_\delta\}$ and $A_s = \{\omega : \zeta_{n_0} \in \mathbf{P}_\delta^s\}$. Then $A = \cup_{s \in \mathcal{S}} A_s$ and the sets A_s, $s \in \mathcal{S}$, are pair disjoint. From (13) it follows that

$$\text{Prob}\,(s_{n_0+1} = s, \ldots, s_{n_0+m} = s \,|\, A_s) \geqslant \prod_{i=0}^{m} \psi^i(\delta) \geqslant \varphi(\delta). \tag{23}$$

Let

$$B_s = \{\omega : s_n(\omega) = s \quad \text{for } n \geqslant n_0(\omega)\},$$

$$B = \bigcup_{s \in \mathcal{S}} B_s.$$

Inequality (23) implies that

$$\text{Prob}\,(B_s | A_s) \geqslant \varphi(\delta)$$

and consequently

$$\text{Prob}\,(B_s) \geqslant \varphi(\delta)\ \text{Prob}\,(A_s). \tag{24}$$

The sets A_s, $s \in \mathcal{S}$, are pair disjoint and the sets B_s, $s \in \mathcal{S}$, are pair disjoint. From (22) and (24) we obtain

$$\text{Prob}\,(B) \geqslant \varphi(\delta)\text{Prob}\,(A) \geqslant \varphi(\delta)(1-\delta).$$

Letting $\delta \to 0$ we have $\text{Prob}\,(B) = 1$, which completes the proof. □

C. Existence of the limit cycles

Now we are ready to complete the proof of Theorem 1.

According to Proposition 1 almost every trajectory is bounded and goes through some point $u_0 \in \mathcal{Z}$ infinitely often. Let D be a bounded subset of $\mathcal{Z}$ and $u_0 \in \mathcal{Z}$ be a given point. Denote by A_0 the subset of Ω which consists of all $\omega \in \Omega$ such that the trajectory $\{\eta_n(\omega)\}$ is contained in D and goes through u_0 infinitely often. According to Proposition 2, for almost every $\omega \in A_0$ there exists $s(\omega) \in \mathcal{S}$ such that the random walker going through u_0 chooses the direction $s(\omega)$ for sufficiently

TABLE I. The numbers RW of the cycle for a few values of α.

	0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94
2	9554	9555	9592	9640	9645	9633	9723	8973
4	397	410	363	329	316	343	247	199
6	38	29	38	23	27	19	17	10
8	7	5	3	5	6	1	4	6
10	0	0	1	0	3	2	0	2
12	0	0	1	0	1	0	0	0
14	1	1	0	0	0	0	0	0
16	0	0	0	0	0	0	0	2
18	0	0	0	0	0	0	1	0

large times. Now, we can divide the set A_0 into N disjoint subsets $B_1,\ldots,B_N$ in such a way that in each set B_k the step (u_0,u_1) is determined uniquely. Denote one of these sets by A_1. Then we can divide the set A_1 into N subsets related to the next step (u_1,u_2), etc. After some steps the moving point returns to u_0 and in this way we obtain a limit cycle. □

IV. NUMERICAL SIMULATIONS

A. Details of the studied models

The theorem proved in previous sections gives only general information about the RW with memory. To gain some insights into the more quantitative characteristics of RW with memory we have performed Monte Carlo simulations for the function

$$f_s(x)=\psi_\alpha(x)=x[2-\alpha-(1-\alpha)x], \tag{25}$$

where α is a parameter from the interval (0,1). Since for $\alpha=1$ the function $\psi_\alpha(x)$ is equal to the identical mapping, we expect that for $\alpha\to 1$, the RW with memory will tend to the usual symmetrical RW. In particular, for $\alpha\approx 1$ it should never fall into the limiting cycle. In other words some critical slowing down in reaching the limit cycle will occur for $\alpha\approx 1$.

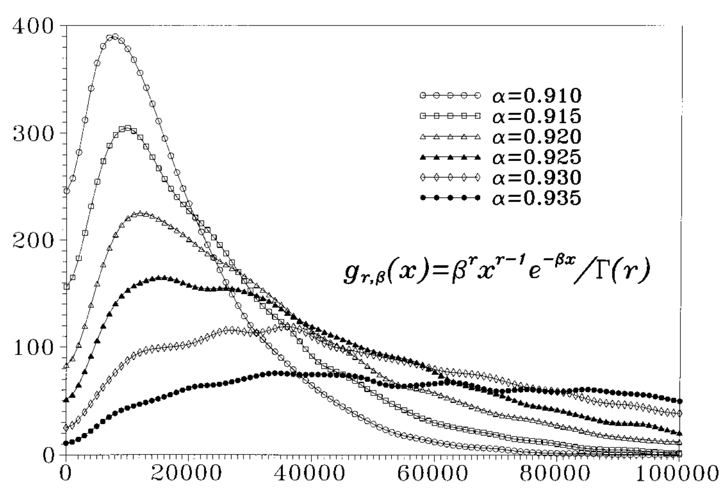

FIG. 1. The plot of the histograms of the number of steps performed by random walkers before the stop of the process for a sample of the values of α.

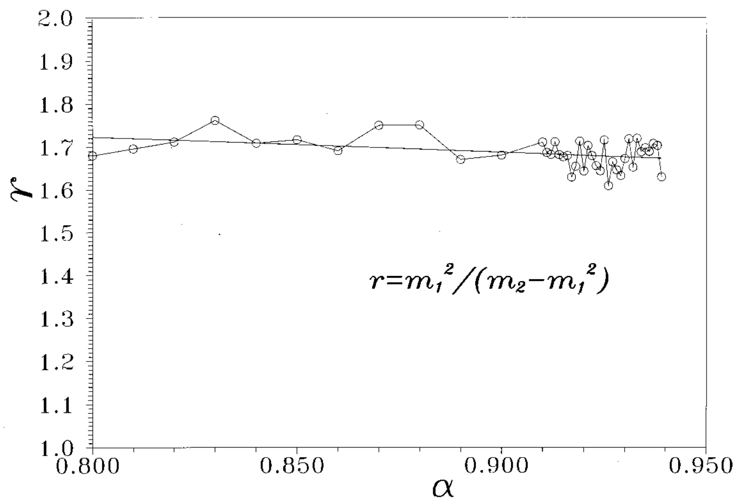

FIG. 2. The plot of the dependence of the parameter r in (26) obtained from the moments of actual data. Remarkably, this figure suggests that r does not depend on α.

B. Obtained results

We have performed the simulations only on the two-dimensional, square lattice of the size 1300×1300. In each node of the lattice we have stored probabilities of making a step in one of the four directions. Initially all p_z were set to be equal $\frac{1}{4}$. The random walker started from the origin of the lattice and after each step the probabilities were updated according to Eq. (1). A given particular simulation of the RW was finished when one p_z reached the value of 0.999 or else the total number of steps was equal to 2 000 000. We have imposed the periodic boundary conditions on the RW and we recorded the facts of crossing by RW the edges of the torus. There were rare cases of such events, most of them occurred, of course, for larger values of α. We have performed simulations for α in the range (0.8, 0.94). In the subrange (0.8, 0.91) α was changed with the step $\Delta\alpha=0.01$, while in the subinterval (0.91, 0.94) with the step $\Delta\alpha=0.001$, because the number of steps performed by random walker before the stop was increasing very rapidly with growing α. We did not continue to larger values of α, because the number of steps needed to stop the RW was too large. For each α there were 10 000 separate random walks performed. We have stored the number of steps N at which for the first time one of the probabilities reached $p_z=0.999$. The path of RW falls into the cycle (see Sec. II D) and the length of the limiting trajectory was also stored. Table I gives a sample of this data for a few values of α for the length of the cycle 2, 4,..., 18—larger cycles have occurred very randomly. The numbers in this table do not sum up to 10 000 because some RW had limit cycles larger than 18, and for large α rare samples did not fall into the limit cycle in less than 2 000 000 steps. These lengths of cycles do not follow the Poisson distribution and we do not have any conjecture describing these numbers.

The numbers of steps, for each α, varied considerably from one sample RW to another. For example, for $\alpha=0.8$ there was a RW which stabilized after $N_{\min}=63$ steps, while the largest number of steps was $N_{\max}=4114$. This gap between smallest and largest number of steps needed to stop RW increased with α, for example, for $\alpha=0.93$ the minimal and maximal number of steps before RW stopped was $N_{\min}=1034$ and $N_{\max}=590\,176$, respectively. We claim that the number of steps N is governed by the gamma distribution with parameters r, β:

$$g_{r,\beta}(x)=\beta^r x^{r-1} e^{-\beta x}/\Gamma(r), \tag{26}$$

where $\Gamma(r)$ is a generalization of the factorial:

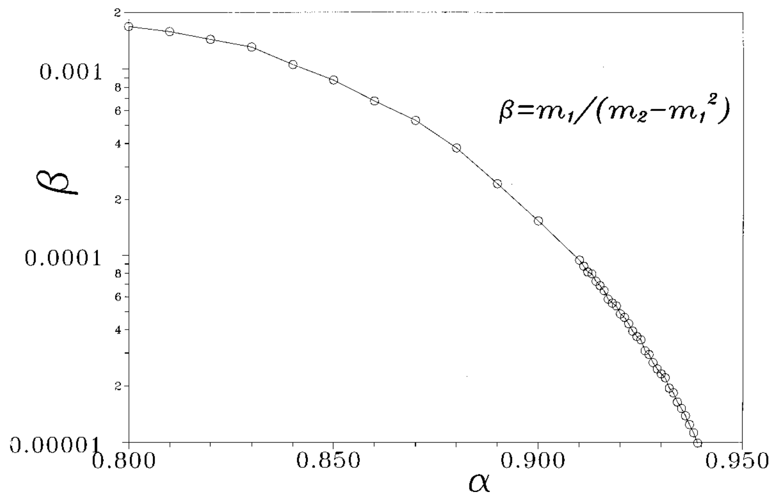

FIG. 3. The plot of the dependence of the parameter β in (26) obtained from the moments of actual data.

$$\Gamma(r)=\int_0^{\infty} e^{-t}t^{r-1}\,dt. \tag{27}$$

In Fig. 1 we present plots of the histograms of the number of steps for a few of values of α. The size of the bins was 1000, so the y axis gives the number of random walks with the number of steps in the range $(1000\times k, 1000\times k+1000)$. In Figs. 2 and 3 the values of the fitted parameters r and β for all investigated values of α are shown. Remarkably, the parameter r takes values around 1.72 and it seems not to depend on α. It is probably linked with the special choice of the function (25).

Despite the large fluctuation of N between different realizations of RW, there seems to be a simple formula describing the median μ value of N. Here μ is defined as such a value of N that the same number of sample random walks stopped in smaller than μ, as well in larger than μ steps. Since for $\alpha=1$ the RW with memory passes into the usual RW, N should diverge to infinity for

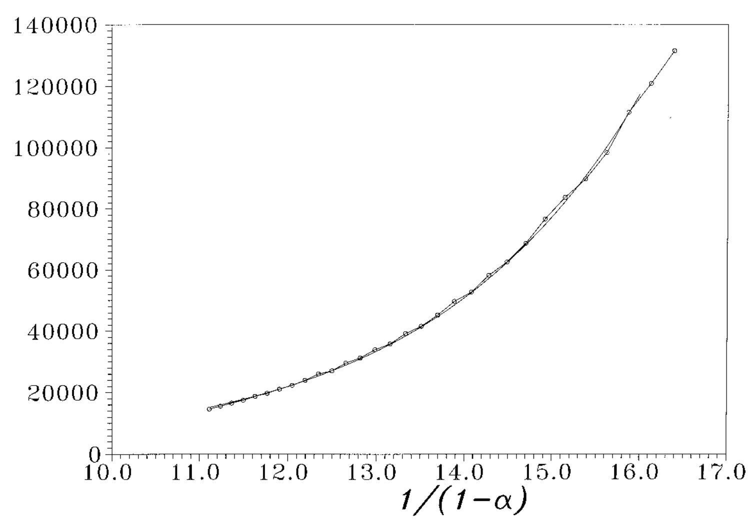

FIG. 4. The plot of dependence of the median $\mu(\alpha)$ versus $u=1/(1-\alpha)$ is shown. The solid line presents the least-square fit to the points obtained from the Monte Carlo simulations, represented by circles, under the assumption that fit is made by the exponential function.

$\alpha \to 1$, thus we guessed that μ is a function of $1/(1-\alpha)$. Hence, in Fig. 4 the plot of the median μ versus $u = 1/(1-\alpha)$ is shown. This figure suggests that $\mu(\alpha)$ grows exponentially with u—the dashed line presents the exponential fit to the actual values obtained by the least-square method:

$$\mu(\alpha) \sim \exp\left(\frac{1}{1-\alpha}\right). \tag{28}$$

Summarizing, the Monte Carlo simulations suggest that there seems to be strict, quantitative rules governing the behavior of some characteristics of the RW with memory for the function $\psi_\alpha(x)$.

ACKNOWLEDGMENTS

This research was supported by the State Committee for Scientific Research (Poland) Grant No. 2 P03A 042 09 (RR) and No. 2 P302 057 07 (MW).

[1] F. Spitzer, *Principles of Random Walk* (Springer-Verlag, New York, 1976).
[2] J. W. Haus and K. W. Kehr, Phys. Rep. **150**, 263 (1987).
[3] S. Havlin and D. Ben-Avraham, Adv. Phys. **36**, 695 (1987).
[4] H. E. Stanley, K. Kang, S. Redner, and R. L. Blumberg, Phys. Rev. Lett. **51**, 1223 (1983).
[5] Two talks by P. Diaconis at the 1987 Midwest Probability Conference, unpublished.
[6] R. Pemantle, ''Phase transition in reinforced random walk and RWRE on trees,'' Ann. Prob. **16**, 1229 (1988).
[7] B. Davis, Probab. Th. Rel. Fields **84**, 203 (1990).
[8] T. Sellke, ''Reinforced random walk on d-dimensional integer lattice,'' Tech. Rept. No. #94-26, Dept. of Statistics, Purdue Univ.
[9] M. F. Barnsley, *Fractals Everywhere* (Academic, New York, 1988).
[10] M. Cranston and Y. Le Jan, Math. Ann. **303**, 87 (1995).
[11] O. Raimond, Probab. Th. Rel. Fields **107**, 177 (1997).
[12] E. B. Dynkin, *Markov Processes, Vol. I* (Springer-Verlag, New York, 1965).

Curvature conditions for the occurrence of a class of spacetime singularities

Wiesław Rudnicki[a] and Paweł Zięba[b]
Institute of Physics, University of Rzeszów, ul. Rejtana 16A, PL-35-310 Rzeszów, Poland

(Received 15 October 1998; accepted for publication 21 January 1999)

It has recently been shown [W. Rudnicki, Phys. Lett. A **224**, 45–50 (1996)] that a generic gravitational collapse cannot result in a naked singularity accompanied by closed timelike curves. An important role in this result plays the so-called *inextendibility condition,* which is required to hold for certain incomplete null geodesics. In this paper, a theorem is proved that establishes some relations between the inextendibility condition and the rate of growth of the Ricci curvature along incomplete null geodesics. This theorem shows that the inextendibility condition may hold for a much more general class of singularities than only those of the strong curvature type. It is also argued that some earlier cosmic censorship results obtained for strong curvature singularities can be extended to singularities corresponding to the inextendibility condition. [S0022-2488(99)02106-4]

I. INTRODUCTION

Recently, one of us[1] has shown that, under certain physically reasonable conditions, a generic gravitational collapse developing from a regular initial state cannot lead to the formation of a final state resembling the Kerr solution with $a^2 > m^2$—i.e., of a naked singularity accompanied by closed timelike curves. This result supports the validity of Penrose's cosmic censorship hypothesis[2] and suggests that there may exist some deeper connection between cosmic censorship and the chronology protection conjecture put forward by Hawking.[3] An important role in this result plays the so-called *inextendibility condition* (see Sec. II), which is assumed to be satisfied for certain incomplete null geodesics. This condition enables one to rule out artificial naked singularities that could easily be created by simply removing points from otherwise well-behaved spacetimes. The inextendibility condition is based on the idea that physically essential singularities should always be associated with large curvature strengths, which are in turn usually associated with the focusing of Jacobi fields along null geodesics.

It is easily seen that the inextendibility condition will always hold for null geodesics terminating at the so-called *strong curvature singularities* defined by Tipler[4] (see below). Singularities of this type are sometimes considered to be the *only* physically reasonable singularities (cf., e.g., Refs. 5 and 6). However, strong curvature singularities can exist only if the curvature in their neighborhood diverges strong enough,[7] while it is not unlikely that some singularities occurring in generic collapse situations will involve a weaker divergence of the curvature. In fact, one cannot *a priori* exclude the existence of some ''real'' singularities near which the curvature would remain even bounded (such singularities occur, for example, in Taub-NUT space). Accordingly, since we still have no fully accepted *necessary* condition on the behavior of the curvature near generic singularities, one should try to prove any cosmic censorship result under as weak a curvature condition as possible. It would be therefore of interest, in view of the mentioned censorship result,[1] to know what are curvature conditions for the occurrence of singularities corresponding to the inextendibility condition. Furthermore, the inextendibility condition has also been used in

[a]Electronic mail: rudnicki@atena.univ.rzeszow.pl
[b]Electronic mail: pzieba@atena.univ.rzeszow.pl

proving some other recent results[8,9] that restrict a class of possible causality violations in classical general relativity.

In this paper, we formulate and prove a theorem that establishes some relations between the inextendibility condition and the rate of growth of the Ricci curvature along incomplete null geodesics. This theorem shows that the inextendibility condition may hold for a much more general class of possible singularities than only those of the strong curvature type. Our theorem will be stated in Sec. II of the paper. In Sec. III, we present a proof of the theorem; our main mathematical tool in this proof is a Sturm-type comparison lemma for nonoscillatory solutions of second-order differential equations. In Sec. IV we give a few concluding remarks; in particular, we argue that some earlier cosmic censorship results obtained for strong curvature singularities can be extended to singularities corresponding to the inextendibility condition.

II. THE THEOREM

To begin with, we clearly need to recall the precise formulation of the inextendibility condition. Let $\eta(t)$ be an affinely parametrized null geodesic, and let Z_1 and Z_2 be two linearly independent spacelike vorticity-free Jacobi fields along $\eta(t)$. The exterior product of these Jacobi fields defines a spacelike area element, whose magnitude at affine parameter value t we denote by $A(t)$. If we now introduce the function $z(t)$ defined by $A(t)\equiv z^2(t)$, then one can show[4] that $z(t)$ satisfies the following equation:

$$\frac{d^2z}{dt^2}+\frac{1}{2}(R_{ab}K^aK^b+2\sigma^2)z=0, \tag{1}$$

where K^a is the tangent vector to $\eta(t)$ and σ^2 is a non-negative function of t defined as follows: $2\sigma^2\equiv\sigma_{mn}\sigma^{mn}$ $(m,n=1,2)$. Here σ_{mn} is the shear tensor (see Ref. 10, p. 88) that satisfies the equation[4]

$$\frac{d}{dt}\sigma_{mn}=-C_{manb}K^aK^b-\frac{2}{z}\frac{dz}{dt}\sigma_{mn}. \tag{2}$$

In the following, by M we shall denote a spacetime, i.e., a smooth, boundaryless, connected, four-dimensional Hausdorff manifold with a globally defined C^{2-} Lorentz metric.

Definition (cf. Refs. 1 and 8): Let $\eta:(0,a]\rightarrow M$ *be an affinely parametrized, incomplete null geodesic. Assume also that* $\eta(t)$ *generates an achronal set, i.e., a set such that no two points of it can be joined by a timelike curve. Then* $\eta(t)$ *is said to satisfy the inextendibility condition if for some affine parameter value* $t_1\in(0,a)$ *there exists a solution* $z(t)$ *of Eq. (1) along* $\eta(t)$ *such that* $z(t_1)=0$, $dz/dt|_{t_1}\neq 0$ *and* $\lim_{t\rightarrow 0}z(t)=0$.

The key idea behind the inextendibility condition is based on the fact that any two zeros of any solution of Eq. (1), which is not identically zero along a given null geodesic, correspond to a pair of conjugate points along the geodesic (see Ref. 4). From Proposition 4.5.12 of Ref. 10, it follows that incomplete null geodesics generating achronal sets cannot contain any pairs of conjugate points. One can thus easily show[8] that if a geodesic $\eta:(0,a]\rightarrow M$ satisfies the inextendibility condition, then there is *no* extension of the spacetime M, preserving all the above mentioned properties of M, in which $\eta(t)$ could be extended beyond a point $\eta(0)$. This means, according to the standard interpretation, that $\eta(t)$ should then approach a genuine singularity of the spacetime M at affine parameter value 0. [Formally, this singularity has the same status as those predicted by the familiar singularity theorems,[10] because these theorems predict, in fact, the existence of incomplete causal (usually null) geodesics in maximally extended spacetimes satisfying just the same topological and smoothness conditions as those imposed on M.]

Let us now compare the inextendibility condition with the concept of a strong curvature singularity.[4] Consider a null geodesic $\lambda:(0,a]\rightarrow M$ that terminates in a strong curvature singularity at affine parameter value 0. This means that every solution $z(t)$ of Eq. (1) along $\lambda(t)$, which vanishes for, at most, finitely many points in $(0,a]$, satisfies $\lim_{t\rightarrow 0}z(t)=0$ (cf. Ref. 5, p. 160).

Suppose now that $\lambda(t)$ generates an achronal set; then any solution of Eq. (1), which is not identically zero along $\lambda(t)$, cannot vanish for any two points in $(0,a]$ by the argument with conjugate points mentioned above. Thus, for *all* $t_1 \in (0,a]$ and for *all* solutions $z(t)$ of Eq. (1) along $\lambda(t)$ with initial conditions $z(t_1)=0$, we will have $\lim_{t\to 0} z(t)=0$. It is thus clear that any null geodesic terminating in Tipler's strong curvature singularity and generating an achronal set must always satisfy the inextendibility condition. Notice also that the terms ''all'' emphasized above imply, via Eqs. (1) and (2), that $\lambda(t)$ can terminate in the strong curvature singularity only if the curvature diverges strong enough along $\lambda(t)$ as $t\to 0$, while the inextendibility condition could actually be satisfied for $\lambda(t)$, even if the curvature along it would remain bounded. Indeed, the theorem stated below makes it clear [see condition (i)] that the curvature need not necessarily diverge along geodesics satisfying the inextendibility condition.

Theorem: *Let $\eta:(0,a]\to M$ be an affinely parametrized, incomplete null geodesic generating an achronal set. Suppose that the Ricci tensor term $r(t)\equiv R_{ab}K^aK^b$ along $\eta(t)$, where t is the affine parameter and K^a is the tangent vector to $\eta(t)$, obeys at least one of the following conditions.*

(i) *There exists an affine parameter value $b\in(0,a)$ such that* $\inf\{r(t)|0<t\leqslant b\}\geqslant 2(\pi/b)^2$.

(ii) *There exist an affine parameter value $c\in(0,a)$ and a constant $\mu\in(0,2)$ such that $r(t)\geqslant \kappa t^{-\mu}$ for all $t\in(0,c]$, where $\kappa=(\frac{2}{3})(33-26\mu+5\mu^2)c^{\mu-2}$.*

Then $\eta(t)$ satisfies the inextendibility condition.

Remark 1: From the proof of this theorem, which is given below, it may be seen that the parameter values b and c mentioned above in conditions (i) and (ii) correspond to the parameter value t_1 occurring in the definition of the inextendibility condition.

Remark 2: Since in the theorem $\eta(t)$ is assumed to be a generator of an achronal set, $\eta(t)$ cannot contain any pair of conjugate points, and so one can expect that there should exist an *upper* limit on the rate of growth of the curvature along $\eta(t)$. Indeed, from Theorems (3) and (4) of Ref. 11, it follows immediately that the Ricci tensor term $r(t)$ along $\eta(t)$ must satisfy the following two conditions: (1) there is no affine parameter value $b'\in(0,a]$ such that $\inf\{r(t)|0<t<b'\}>8(\pi/b')^2$; and (2) if $r(t)\geqslant 0$ on $\eta(t)$, then $\lim_{t\to 0}\inf t^2 r(t)\leqslant \frac{1}{2}$. Similar restrictions on the growth of the Weyl part of the curvature along $\eta(t)$ can be obtained from Proposition 1.2 of Ref. 12.

In the context of our theorem, it is worth recalling the analogous results obtained by Clarke and Królak[7] for singularities of the strong curvature type. They have been obtained for two definitions of a strong curvature singularity: the original one formulated by Tipler[4] and its modification proposed by Królak.[6] According to these results, if a null geodesic $\eta:(0,a]\to M$ terminates at affine parameter value 0 in a strong curvature singularity defined by Tipler (resp., by Królak), then there must exist some affine parameter value $c\in(0,a]$ such that $R_{ab}K^aK^b>At^{-2}$ (resp., $R_{ab}K^aK^b>At^{-1}$) on $(0,c]$, where K^a is the tangent vector to $\eta(t)$, t is the affine parameter, and A is some fixed positive constant. [Or very similar conditions on the rate of growth of the Weyl part of the curvature along $\eta(t)$ must be satisfied; see Corollary 2 of Ref. 7.] Comparing these results with condition (ii) of our theorem, we see that singularities of the strong curvature type involve a considerably stronger divergence of the Ricci tensor term $R_{ab}K^aK^b$ than singularities corresponding to the inextendibility condition. There may thus exist a large class of curvature singularities that are not strong in the sense of the definition of Tipler or Królak, but they may still satisfy the inextendibility condition. Note also that the above conditions for strong curvature singularities are the *necessary* ones, whereas conditions (i) and (ii) of our theorem are only *sufficient* to ensure that the inextendibility condition does hold for a given geodesic. This implies that the inextendibility condition might be satisfied in more general situations than only those characterized by conditions (i) and (ii).

III. PROOF OF THE THEOREM

Now we shall prove the theorem; our main tool in this proof will be the following comparison lemma.

Lemma (The comparison lemma): Suppose that $u(s)$ is a solution of the equation

$$\frac{d^2u}{ds^2}+F(s)u(s)=0,$$

on an interval (a,b] with initial conditions: $u(b)=0$ and $du/ds|_b\neq 0$. Let $v(s)$ be a solution of

$$\frac{d^2v}{ds^2}+G(s)v(s)=0,$$

on (a,b], such that $v(b)=0$, $dv/ds|_b=du/ds|_b$ and $v(s)>0$ on (a,b). Assume also that $F(s)$ and $G(s)$ are piecewise continuous on (a,b], and let $G(s)\geqslant F(s)$ on (a,b]. Then $u(s)\geqslant v(s)$ on (a,b].

Proof: The proof of this lemma is based essentially on Theorem 1.2 of Ref. 13, p. 210. To apply this theorem in its original form, it is convenient to reparametrize both of the equations in the lemma, introducing the parameter $t=-s$ instead of s. Note that this reparametrization does not change the form of the equations. Clearly, we shall now have established the lemma if we show that for any $c\in(a,b)$, $u(t)\geqslant v(t)$ on $[-b,-c]$.

Consider the ratio $u(t)/v(t)$. Since $v(t)>0$ on $(-b,-a)$, it is well defined on $(-b,-c]$. Using l'Hospital's rule, we get

$$\lim_{t\to -b}\frac{u(t)}{v(t)}=1.$$

Therefore, as $v(t)>0$ on $(-b,-c]$, to show that $u(t)\geqslant v(t)$ on $[-b,-c]$, it suffices to show that

$$\frac{d}{dt}\left[\frac{u(t)}{v(t)}\right]\geqslant 0,$$

on $(-b,-c]$. It is easy to see that this inequality holds if

$$\frac{v(t)}{\dot{v}(t)}\geqslant\frac{u(t)}{\dot{u}(t)}, \tag{3}$$

on $(-b,-c]$, where the overdot denotes the first derivative with respect to t. By Theorem 1.2 of Ref. 13, p. 210, we have

$$\tan^{-1}\left[\frac{v(t)}{\dot{v}(t)}\right]\geqslant\tan^{-1}\left[\frac{u(t)}{\dot{u}(t)}\right],$$

for all $t\in[-b,-c]$. Thus, as $\tan^{-1}$ is an increasing function, the inequality (3) does hold, as it is desirable. □

Proof of the theorem: (Part I) Suppose the condition (i) is satisfied. Let $z_0(t)$ be a solution of Eq. (1) along $\eta(t)$ such that $z_0(t)$ is not identically zero on $(0,b]$ and $z_0(b)=0$, where b is the parameter value mentioned in condition (i). Clearly, such a solution will always exist. Since $\eta(t)$ generates an achronal set, $z_0(t)$ can vanish nowhere in $(0,b)$; otherwise $\eta(t)$ would have a pair of conjugate points in $(0,b]$ (see Ref. 4), which would contradict, by Proposition 4.5.12 of Ref. 10, the achronality of $\eta(t)$. Notice also that Eq. (1) is linear, and so the function $-z_0(t)$ will be a solution of Eq. (1) as well. Thus, as $z_0\neq 0$ on $(0,b)$, without loss of generality we can assume that $z_0(t)>0$ on $(0,b)$. This implies, as $z_0(b)=0$, that $dz_0/dt|_b\leqslant 0$. Since $z_0(t)>0$ on $(0,b)$, and condition (i) holds, from Eq. (1) we see at once that $z_0(t)$ must be a concave function on $(0,b]$. This makes it obvious that $dz_0/dt|_b\neq 0$, and so we must have $dz_0/dt|_b=\alpha<0$. Let us now define the function $z_1(t)\equiv-(1/\alpha)z_0(t)$. As Eq. (1) is linear, it is clear that $z_1(t)$ will be a solution of Eq. (1) along $\eta(t)$; notice also that $z_1(t)>0$ on $(0,b)$, $z_1(b)=0$ and $dz_1/dt|_b=-1$.

Consider now the equation

$$\frac{d^2x}{dt^2}+\omega x(t)=0, \tag{4}$$

where $\omega=\frac{1}{2}\inf\{r(t)|0<t\leqslant b\}$ and $r(t)$ is the function defined in the theorem. Notice that $\omega>0$ by condition (i). Let $x_1(t)$ be a solution of Eq. (4) on $(0,b]$ with initial conditions $x_1(b)=0$ and $dx_1/dt|_b=-1$. It is a simple matter to see that $x_1(t)=\omega^{-1/2}\sin[\omega^{1/2}(b-t)]$. Let us now apply the comparison lemma to the equations (1) and (4) and their solutions $z_1(t)$ and $x_1(t)$. Since $\omega\leqslant\frac{1}{2}r(t)$ on $(0,b]$, by the comparison lemma we must have $x_1(t)\geqslant z_1(t)$ on $(0,b]$. Consequently, as $z_1(t)>0$ on $(0,b)$, we obtain $x_1(t)>0$ on $(0,b)$. This implies, by the above form of $x_1(t)$, that $\omega\leqslant(\pi/b)^2$. But $\omega\geqslant(\pi/b)^2$ by condition (1). We must thus have $\omega=(\pi/b)^2$, which gives $\lim_{t\to 0}x_1(t)=0$. Therefore $\lim_{t\to 0}z_1(t)=0$ since $x_1(t)\geqslant z_1(t)>0$ on $(0,b)$. This means that $\eta(t)$ does satisfy the inextendibility condition.

(Part II) The task is now to prove the theorem in the case when condition (ii) holds. For this purpose, let us consider the following equation:

$$\frac{d^2y}{dt^2}+Bt^{-\mu}y(t)=0, \tag{5}$$

on $(0,c]$, where $B=\kappa/2$, and κ, μ, and c are some fixed constants mentioned in the condition (ii). Let $y_1(t)$ be a solution of this equation with initial conditions $y_1(c)=0$ and $dy_1/dt|_c=-1$. Let $z_2(t)$ be a solution of Eq. (1) along $\eta(t)$, such that $z_2(c)=0$ and $dz_2/dt|_c=-1$. [There is no loss of generality in assuming $z_2(t)$ to exist; the existence of $z_2(t)$ can be established in the same manner as the existence of the analogous solution $z_1(t)$ considered in the first part of the proof.] Clearly, the solution $z_2(t)$, just as $z_1(t)$, can vanish nowhere in $(0,c)$ by the argument with conjugate points. Therefore, as $dz_2/dt|_c=-1$, we must have $z_2(t)>0$ on $(0,c)$. Let us now apply the comparison lemma to the equations (1) and (5) and their solutions $z_2(t)$ and $y_1(t)$. By condition (ii) we have $r(t)\geqslant\kappa t^{-\mu}$ on $(0,c]$. Thus, by the comparison lemma, we must have $y_1(t)\geqslant z_2(t)$ on $(0,c]$. Of course, in order to prove the theorem, it suffices to show that $\lim_{t\to 0}z_2(t)=0$. Thus, as $y_1(t)\geqslant z_2(t)>0$ on $(0,c)$, to complete the proof it suffices to show that $\lim_{t\to 0}y_1(t)=0$. We shall show below that $y_1(t)$ does possess this property.

To this end, let us first find the general solution of Eq. (5). It is easy to check that if one puts $x=t$, $\alpha=1/2$, $\beta=2\sqrt{B}(2-\mu)^{-1}$, $\gamma=(2-\mu)/2$, and $n=(2-\mu)^{-1}$ into the equation (4.1) of Ref. 14, p. 138, then this equation reduces to our equation (5). Thus, according to the solution (4.3) of Eq. (4.1) of Ref. 14, our equation (5) has the following general solution

$$y(t)=t^{1/2}[C_1J_n(\beta t^\gamma)+C_2Y_n(\beta t^\gamma)], \tag{6}$$

where C_1 and C_2 are arbitrary constants of integration, and $J_n(\beta t^\gamma)$ and $Y_n(\beta t^\gamma)$ are the Bessel functions of order n, of the first and second kind, respectively. Since $\mu\in(0,2)$, from the above relations it follows that $\frac{1}{2}<n<\infty$, $\sqrt{B}<\beta<\infty$ and $0<\gamma<1$.

Let us recall that any Bessel function of the first kind has infinitely many positive zeros (cf., e.g., Ref. 15, p. 29). Let $j_{n,1}$ be the first positive zero of the function $J_n(\beta t^\gamma)$, i.e., $J_n(j_{n,1})=0$ and $J_n(\beta t^\gamma)\neq 0$ as long as $0<\beta t^\gamma<j_{n,1}$. Since $n>\frac{1}{2}$, $j_{n,1}$ must satisfy the following relation [see Eq. (2) of Ref. 15, p. 29]:

$$j_{n,1}<2[(n+1)(n+5)/3]^{1/2}. \tag{7}$$

For $J_n(\beta t^\gamma)$ we now define L to be the number such that $j_{n,1}=L\beta c^\gamma$. Putting this into (7), and taking into account the fact that $\beta=(2\kappa)^{1/2}(2-\mu)^{-1}$, $\kappa=3^{-1}(66-52\mu+10\mu^2)c^{\mu-2}$, $\gamma=(2-\mu)/2$ and $n=(2-\mu)^{-1}$, we readily find that $L^2<1$.

Consider now Eq. (5) with B replaced by $B'=L^2B$. Let $y_2(t)$ be a solution of this equation on $(0,c]$ with initial conditions $y_2(c)=0$ and $dy_2/dt|_c=-1$. The general form of this solution is given by (6), where β should be replaced by $\beta'=2\sqrt{B'}(2-\mu)^{-1}$ (notice that $\beta'=L\beta$). Let us now insert the initial conditions for $y_2(t)$ into this general solution in order to determine for $y_2(t)$ the constants C_1 and C_2 occurring in (6). To find the first derivative of the general solution (6), we use the following recurrence formula:

$$\frac{dJ_n(x)}{dx}=-J_{n+1}(x)+\frac{n}{x}J_n(x),$$

which is also valid for $Y_n(x)$ (see Ref. 15, p. 197). We can now easily calculate the constants C_1 and C_2; the result is as follows:

$$C_1=\frac{Y_n(\beta' c^\gamma)}{\beta'\gamma c^{\gamma-1/2}[Y_n(\beta' c^\gamma)J_{n+1}(\beta' c^\gamma)-Y_{n+1}(\beta' c^\gamma)J_n(\beta' c^\gamma)]} \tag{8}$$

and

$$C_2=\frac{-J_n(\beta' c^\gamma)}{\beta'\gamma c^{\gamma-1/2}[Y_n(\beta' c^\gamma)J_{n+1}(\beta' c^\gamma)-Y_{n+1}(\beta' c^\gamma)J_n(\beta' c^\gamma)]}. \tag{9}$$

As $\beta'=L\beta$, from the above definition of L it is clear that $\beta' c^\gamma=j_{n,1}$. Thus $J_n(\beta' c^\gamma)=0$, and the numerator in (9) must vanish. As $J_n(\beta' c^\gamma)=0$, the denominator in (9) can vanish only if $Y_n(\beta' c^\gamma)J_{n+1}(\beta' c^\gamma)=0$. But the Bessel functions J_{n+1} and Y_n cannot have any common zeros with the Bessel function J_n (see Ref. 15, pp. 29–32), and so the denominator in (9) cannot vanish. We thus have $C_2=0$ and, by (6) and (8), the solution $y_2(t)$ can be written as follows:

$$y_2(t)=C_1 t^{1/2}J_n(\beta' t^\gamma), \tag{10}$$

where $C_1=[\beta'\gamma c^{\gamma-1/2}J_{n+1}(\beta' c^\gamma)]^{-1}$.

Let us now compare the solutions $y_1(t)$ and $y_2(t)$ by means of the comparison lemma. Recall that $y_1(t)$ is a solution of Eq. (5) with $B=\kappa/2$, while $y_2(t)$ is a solution of the same equation with B replaced by $B'=L^2\kappa/2$. Since $L^2<1$, by the comparison lemma we must have $y_2(t)\geqslant y_1(t)$ for all $t\in(0,c]$. We recall that any Bessel function $J_k(x)$ of the first kind with real x and $k>0$ is continuous at $x=0$ (cf. Ref. 15, p. 182). Thus, as $n>\frac{1}{2}$ and $0<\gamma<1$, from (10) it follows immediately that $\lim_{t\to 0} y_2(t)=0$. Therefore, as $y_2(t)\geqslant y_1(t)>0$ on $(0,c)$, we obtain $\lim_{t\to 0} y_1(t)=0$, which completes the proof. □

IV. CONCLUDING REMARKS

We have been concerned in this paper with the problem of determining what are curvature conditions for the occurrence of singularities corresponding to the inextendibility condition. We have found two such sufficient conditions concerning the behavior of the Ricci tensor term $R_{ab}K^aK^b$ along incomplete null geodesics—these are conditions (i) and (ii) of the theorem stated in Sec. II. This theorem shows that the inextendibility condition may hold for a considerably larger class of possible singularities than only those of the strong curvature type. In particular, condition (i) of the theorem shows that the inextendibility condition may hold, even if the curvature along incomplete geodesics would remain bounded. In this context, it is worth recalling that singularities predicted by the famous singularity theorems[10] can be interpreted as regions of the universe at which the normal classical spacetime picture and/or certain energy conditions break down, and this may occur in regions where the curvature, though extremely large, still remains finite. Accordingly, if one attempts to establish, for example, whether or not these singular regions will conform to any cosmic censorship principle, it would be well to try to characterize, if necessary, incomplete geodesics terminating in these regions by a condition that may hold even if the curvature along the geodesics would remain bounded. One possible candidate for such a condition may thus be the inextendibility condition.

It should also be stressed here that some earlier cosmic censorship theorems[6,16,17] proved for strong curvature singularities can be extended to singularities corresponding to the inextendibility condition. To see this, let us first recall that these theorems show, briefly, that under certain restrictions imposed on the causal structure, strong curvature singularities are censored (see Refs. 6, 16, and 17 for details). Proofs of these theorems are, in essence, alike. In a brief outline, they run as follows. First, one shows that if the theorem under consideration were false, then there would have to exist a sequence $\{\mu_i\}$ of future endless, future complete null geodesics converging to a null geodesic μ that terminates in the future at a strong curvature singularity. One also shows that μ and all the μ_i must be generators of achronal sets. As all μ_i are achronal, none of them can have a pair of conjugate points, and so any irrotational congruence of Jacobi fields along any μ_i cannot be refocused. As $\{\mu_i\}$ converges to μ, this must then imply, by continuity, that any irrotational congruence of Jacobi fields along μ cannot be refocused as well. However, as μ terminates in a strong curvature singularity, all irrotational congruences of Jacobi fields along μ should be refocused. This gives the required contradiction. It is not difficult to see, however, that this contradiction can equally well be obtained if μ would be assumed to satisfy the inextendibility condition, for this condition holds if at least one irrotational congruence of Jacobi fields along a given geodesic is refocused. It is thus clear that the censorship theorems given in Refs. 6, 16, and 17 are unnecessarily restricted to strong curvature singularities and they can be extended to singularities corresponding to the inextendibility condition.

ACKNOWLEDGMENT

This research was supported in part by the Polish State Committee for Scientific Research (KBN) under Grant No. 2 P03B 073 15.

[1] W. Rudnicki, ''Cosmic censorship for Kerr-like collapse,'' Phys. Lett. A **224**, 45–50 (1996).
[2] R. Penrose, ''Gravitational collapse: the role of general relativity,'' Riv. Nuovo Cimento **1**, 252–276 (1969).
[3] S. W. Hawking, ''Chronology protection conjecture,'' Phys. Rev. D **46**, 603–611 (1992).
[4] F. J. Tipler, ''Singularities in conformally flat spacetimes,'' Phys. Lett. A **64**, 8–10 (1977).
[5] F. J. Tipler, C. J. S. Clarke, and G. F. R. Ellis, ''Singularities and horizons—a review article,'' in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 2.
[6] A. Królak, ''Towards the proof of the cosmic censorship hypothesis,'' Class. Quantum Grav. **3**, 267–280 (1986).
[7] C. J. S. Clarke and A. Królak, ''Conditions for the occurrence of strong curvature singularities,'' J. Geom. Phys. **2**, 127–143 (1985).
[8] W. Rudnicki, ''Black hole interiors cannot be totally vicious,'' Phys. Lett. A **208**, 53–58 (1995).
[9] K. Maeda, A. Ishibashi, and M. Narita, ''Chronology protection and non-naked singularity,'' Class. Quantum Grav. **15**, 1637–1651 (1998).
[10] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973).
[11] F. J. Tipler, ''On the nature of singularities in general relativity,'' Phys. Rev. D **15**, 942–945 (1977).
[12] L. B. Szabados, ''On singularity theorems and curvature growth,'' J. Math. Phys. **28**, 142–145 (1987).
[13] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* (McGraw-Hill, New York, 1955).
[14] J. Irving and N. Mullineux, *Mathematics in Physics and Engineering* (Academic, New York, 1959).
[15] N. W. McLachlan, *Bessel Functions for Engineers* (Clarendon, Oxford, 1955).
[16] A. Królak, ''Towards the proof of the cosmic censorship hypothesis in cosmological space–times,'' J. Math. Phys. **28**, 138–141 (1987).
[17] R. P. A. C. Newman and P. S. Joshi, ''Constraints on the structure of naked singularities in classical general relativity,'' Ann. Phys. (N.Y.) **182**, 112–119 (1988).

Two-parametric extension of *h*-deformation of Gr(1|1)

Sultan A. Çelik
Department of Mathematics, Yildiz Technical University, 80270 Sisli, Istanbul, Turkey

Emanullah Hizel
Department of Mathematics, Istanbul Technical University, 80626 Maslak, Istanbul, Turkey

Salih Çelik
Department of Mathematics, Mimar Sinan University, 80690 Besiktas, Istanbul, Turkey

(Received 10 November 1998; accepted for publication 19 February 1999)

The two-parametric quantum deformation of the algebra of coordinate functions on the (dual) supergroup Gr(1|1) via a contraction of $\mathrm{Gr}_{p,q}(1|1)$ is presented. Although the quantum superdeterminant of any element of $\mathrm{Gr}_{p,q}(1|1)$ is not central, in the two parametric Jordanian deformation of Gr(1|1) the quantum superdeterminant belongs to the center. The Hopf algebra structure of $\mathrm{Gr}_{h_1,h_2}(1|1)$ is discussed.
 [S0022-2488(99)00606-4]

I. INTRODUCTION

The discussion of the *h*-deformation of the matrix Lie group GL(2) via a contraction[1] of the quantum matrix group $\mathrm{GL}_q(2)$ has led to renewed interest in the *h*-deformation of the simplest supergroup GL(1|1) (Ref. 2) and the (dual) supergroup Gr(1|1) (Ref. 3) which appears as the Jordanian deformation of the supergroups.

In recent years this deformation [*h*-deformation], which appears as a new class of quantum deformations of matrix Lie groups and algebras, has been intensively studied by many authors.[1–13] An interesting property of the *h*-deformations of GL(1|1) and Gr(1|1), is that in both cases, the deformation parameter *h* is an anticommuting Grassmann number.[2,3] In the two-parametric extension of *h*-deformation of GL(1|1), it is seen that both deformation parameters are again anticommuting Grassmann numbers.[4] Another interesting point is that the quantum superdeterminant of a matrix in $\mathrm{Gr}_{p,q}(1|1)$ does not commute with all the matrix elements, i.e., it does not belong to the center of $\mathrm{Gr}_{p,q}(1|1)$. However in the two-parametric Jordanian deformation of Gr(1|1) the quantum superdeterminant belongs to the center of the (h_1,h_2)-deformed supergroup Gr(1|1). Hence the general message of this paper may be that multiparameter deformations are more natural for *h*-deformations as compared to *q*-deformations.

The purpose of this paper is to present the two-parametric extension of the *h*-deformation of the (dual or Grassmann) supergroup Gr(1|1) using the approaches of Refs. 4 and 3. The paper is organized as follows. In Sec. II we give some notations and useful formulas which will be used in this work. In Sec. III we present the two-parameter deformation of Gr(1|1) as related to superplanes. In the following section we get a two-parameter *R*-matrix which deforms the supergroup Gr(1|1). Since the Hopf algebra structure of $\mathrm{Gr}_{h_1,h_2}(1|1)$ is related to $\mathrm{GL}_{h_1,h_2}(1|1)$ it is presented in Sec. V, as a separate section.

II. REVIEW OF $\mathrm{Gr}_{p,q}(1|1)$

A Grassmann supermatrix $\hat{T}$ which is an element of Gr(1|1) is of the form

$$\hat{T}=\begin{pmatrix} \alpha & b \\ c & \delta \end{pmatrix}$$

with two odd (Greek letters) and two even (Latin letters) matrix elements. As usual, the even elements commute with everything and the odd elements anticommute among themselves. We state briefly the properties of the (p,q)-deformed Grassmann supergroup[14] we are going to need in this work.

In this paper we denote (p,q)-deformed objects by primed quantities. Unprimed quantities represent transformed coordinates.

Let us consider the Manin[15] quantum superplane A_p and its dual A_q^*,

$$U'=\begin{pmatrix} x' \\ \xi' \end{pmatrix} \in A_p \Leftrightarrow x'\xi' - p\xi' x' = 0, \quad \xi'^2 = 0 \tag{2.1}$$

and its dual

$$V'=\begin{pmatrix} \eta' \\ y' \end{pmatrix} \in A_q^* \Leftrightarrow \eta'^2 = 0, \quad \eta' y' - q^{-1} y' \eta' = 0. \tag{2.2}$$

Suppose that the matrix elements of $\hat{T}'$ (anti-) commute with the coordinates of A_p and A_q^*. Then, the endomorphisms

$$\hat{T}':A_p \rightarrow A_q^* \quad \text{and} \quad \hat{T}':A_q^* \rightarrow A_p \tag{2.3}$$

impose the following bilinear product relations among the matrix elements:[14]

$$\begin{gathered} \alpha' b' = p^{-1} b' \alpha', \quad \alpha' c' = q^{-1} c' \alpha', \\ \delta' b' = p^{-1} b' \delta', \quad \delta' c' = q^{-1} c' \delta', \\ \alpha' \delta' + \delta' \alpha' = 0, \quad \alpha'^2 = 0 = \delta'^2, \\ b' c' = pq^{-1} c' b' + (p - q^{-1}) \delta' \alpha', \end{gathered} \tag{2.4}$$

where p and q are nonzero complex numbers and $pq \pm 1 \neq 0$. These relations define a two-parameter deformation of the algebra of coordinate functions on the Grassmann matrix supergroup $\mathrm{Gr}(1|1)$ as an associative algebra with unit, generated by the generators α, b, c, and δ.

The above relations are equivalent to the equation[14]

$$R_{p,q} \hat{T}_1' \hat{T}_2' = -\hat{T}_2' \hat{T}_1' R_{p,q}, \tag{2.5}$$

where $\hat{T}_1' = \hat{T}' \otimes I$, $\hat{T}_2' = I \otimes \hat{T}'$ and

$$R_{p,q} = \begin{pmatrix} p+q^{-1} & 0 & 0 & 0 \\ 0 & -2 & q^{-1}-p & 0 \\ 0 & p-q^{-1} & -2pq^{-1} & 0 \\ 0 & 0 & 0 & p+q^{-1} \end{pmatrix}. \tag{2.6}$$

Here we employ the convenient grading notation

$$(\hat{T}_1)_{kl}^{ij} = (\hat{T} \otimes I)_{kl}^{ij} = (-1)^{k(j+l)} \hat{T}_k^i \delta_l^j, \tag{2.7}$$

$$(\hat{T}_2)_{kl}^{ij} = (I \otimes \hat{T})_{kl}^{ij} = (-1)^{i(j+l)} \hat{T}_l^j \delta_k^i. \tag{2.8}$$

Note that although the algebra (2.4) is an associative algebra of the matrix elements of $\hat{T}'$, $R_{p,q}$ does not satisfy the graded quantum Yang–Baxter equation (QYBE),

$$R_{12}R_{13}R_{23}=R_{23}R_{13}R_{12}. \tag{2.9}$$

However, an interesting point is that if we decompose the matrix $R_{p,q}$ in the form

$$R_{p,q}=R^1_{p,q}+R^2_{p,q}, \tag{2.10}$$

where

$$R^1_{p,q}=\begin{pmatrix} p & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & p-q^{-1} & -pq^{-1} & 0 \\ 0 & 0 & 0 & q^{-1} \end{pmatrix}, \quad R^2_{p,q}=\begin{pmatrix} q^{-1} & 0 & 0 & 0 \\ 0 & -1 & q^{-1}-p & 0 \\ 0 & 0 & -pq^{-1} & 0 \\ 0 & 0 & 0 & p \end{pmatrix}, \tag{2.11}$$

then both matrices $R^1_{p,q}$ and $R^2_{p,q}$ do satisfy the graded QYBE. Also Eq. (2.5) can be written of the form

$$R^1_{p,q}T'_1T'_2=-T'_2T'_1R^2_{p,q}. \tag{2.12}$$

This equation will be used in Sec. IV.

III. THE TWO-PARAMETRIC *h*-DEFORMATION OF Gr(1|1)

We introduce new coordinates x and ξ by

$$U=g^{-1}_{h_1}U', \quad U=\begin{pmatrix} x \\ \xi \end{pmatrix}, \tag{3.1}$$

where

$$g_{h_1}=\begin{pmatrix} 1 & 0 \\ f_1 & 1 \end{pmatrix}, \quad f_1=\frac{h_1}{p-1}. \tag{3.2}$$

Here the deformation parameter h_1 is a Grassmann number which has the following properties:

$$h_1^2=0, \quad h_1\xi=-\xi h_1. \tag{3.3}$$

Now, in the limit $p\rightarrow 1$ we get the following exchange relations:

$$x\xi=\xi x+h_1x^2, \quad \xi^2=-h_1x\xi. \tag{3.4}$$

These relations define a new deformation, which we called the h_1-deformation, of the algebra of functions on the Manin superplane generated by x and ξ, and we denote it by A_{h_1}.

Let us consider dual coordinates η and y with

$$V=g^{-1}_{h_2}V', \quad V=\begin{pmatrix} \eta \\ y \end{pmatrix}, \tag{3.5}$$

where[4]

$$g_{h_2}=\begin{pmatrix} 1 & f_2 \\ 0 & 1 \end{pmatrix}, \quad f_2=\frac{h_2}{q-1}. \tag{3.6}$$

The deformation parameter h_2 is again a Grassmann number and it has the following properties:

$$h_2^2=0, \quad h_2\eta=-\eta h_2. \tag{3.7}$$

Next, taking the $q \to 1$ limit we obtain the following relations, which define the dual h_2-superplane $A^*_{h_2}$ as generated by η and y with the exchange relations[4]

$$\eta^2 = -h_2 \eta y, \quad \eta y = y\eta - h_2 y^2. \tag{3.8}$$

We now consider the endomorphisms

$$\hat{T}: A_{h_1} \to A^*_{h_2} \quad \text{and} \quad \hat{T}: A^*_{h_2} \to A_{h_1}. \tag{3.9}$$

Then, we define the corresponding (h_1, h_2)-deformation of the Grassmann supergroup Gr$(1|1)$ as a quantum matrix group $\mathrm{Gr}_{h_1,h_2}(1|1)$ generated by α, b, c, δ which satisfy the following (h_1, h_2)-commutation relations:

$$\begin{aligned}
\alpha b &= b\alpha + h_1 b^2 - h_2(\alpha\delta + bc) - h_1 h_2 b\delta,\\
\alpha c &= c\alpha + h_1(cb - \alpha\delta) - h_2 c^2 - h_1 h_2 c\delta,\\
\alpha^2 &= h_1 \alpha b - h_2 \alpha c + h_1 h_2 \alpha\delta,\\
\delta b &= b\delta - h_1 b^2 + h_2(\alpha\delta + cb) - h_1 h_2 b\alpha,\\
\delta c &= c\delta + h_1(\alpha\delta - bc) + h_2 c^2 - h_1 h_2 c\alpha,\\
\delta^2 &= -h_1 b\delta + h_2 c\delta - h_1 h_2 \alpha\delta,\\
\alpha\delta &= -\delta\alpha + h_1(\delta b - b\alpha) + h_2(c\alpha - c\delta) - h_1 h_2(\delta\alpha + cb),\\
bc &= cb + h_1(\delta b + \alpha b) - h_2(c\delta + \alpha c) + h_1 h_2(\alpha\delta - cb),
\end{aligned} \tag{3.10}$$

provided that α and δ anticommute with ξ, η, h_1, and h_2, and

$$h_1 h_2 = -h_2 h_1. \tag{3.11}$$

When we take $h_2 = 0$, we obtain the quantum Grassmann supergroup with one parameter.[3]

These relations can be obtained from the requirement that A_{h_1} and $A^*_{h_2}$ have to be covariant under the left coactions

$$\mu: A_{h_1} \to Gr_{h_1,h_2}(1|1) \otimes A^*_{h_2}, \quad \mu^*: A^*_{h_2} \to Gr_{h_1,h_2}(1|1) \otimes A_{h_1}, \tag{3.12}$$

such that

$$\mu(U) = \hat{T} \otimes V, \quad \text{i.e.,} \quad \mu(U^i) = \hat{T}^i_j \otimes V^j, \tag{3.13}$$

$$\mu^*(V) = \hat{T} \otimes U, \quad \text{i.e.,} \quad \mu^*(V^i) = \hat{T}^i_j \otimes U^j. \tag{3.14}$$

Recall that the multiplication in the tensor product space follows the rule

$$(A \otimes B)(C \otimes D) = (-1)^{P(B)P(C)} AC \otimes BD, \tag{3.15}$$

where $P(X)$ is the z_2-grade of X. Note that Eqs. (3.13) and (3.14) do not conform to standard definitions because of (2.3).

Alternatively, the relations (3.10) can be obtained using the following similarity transformation which was given by Aghamohammadi *et al.*:[1]

$$\hat{T}' = g\hat{T}g^{-1}, \tag{3.16}$$

where in our case[4]

$$g = g_{h_1} g_{h_2}. \tag{3.17}$$

To do this, we use the relations (2.4) and then take the limits $p \to 1$, $q \to 1$. But in that case, the required steps are rather complicated and tedious.

The quantum (dual) superdeterminant of $\hat{T}$ is defined as

$$\hat{D}_{h_1,h_2} = bc^{-1} - \alpha c^{-1} \delta c^{-1} = c^{-1} b - c^{-1} \alpha c^{-1} \delta, \tag{3.18}$$

which is independent of the relations (3.10). As an interesting case, it can be verified that $\hat{D}_{h_1,h_2}$ commutes with all matrix elements of $\hat{T}$, that is, $\hat{D}_{h_1,h_2}$ belongs to the center of the algebra

$$\hat{T}\hat{D} = \hat{D}\hat{T},$$

although the quantum (dual) superdeterminant of a matrix in $\mathrm{Gr}_{p,q}(1|1)$ does not commute with all the matrix elements.

IV. *R*-MATRIX FOR $\mathrm{Gr}_{h_1,h_2}(1|1)$

We shall obtain an *R*-matrix for the quantum Grassmann supergroup $\mathrm{Gr}_{h_1,h_2}(1|1)$ from the *R*-matrix of $\mathrm{Gr}_{p,q}(1|1)$.

The algebra (2.4) is associative under multiplication and the relations (2.4) may be expressed in terms of a graded *R*-matrix condition, which we repeat here,

$$R^1_{p,q} \hat{T}'_1 \hat{T}'_2 = -\hat{T}'_2 \hat{T}'_1 R^2_{p,q} \tag{4.1}$$

[see Eqs. (2.10)–(2.12)]. Now substituting (3.16) into (4.1) and defining the *R*-matrix R_{h_1,h_2} as

$$R_{h_1,h_2} = \lim_{p \to 1} \lim_{q \to 1} (g_{h_1,h_2} R'_{p,q} g_{h_1,h_2}), \tag{4.2}$$

where

$$g_{h_1,h_2} = g_1 g_2^{-1}, \quad R'_{p,q} \in \{R^1_{p,q}, R^2_{p,q}\}, \tag{4.3}$$

we get the following *R*-matrix R_{h_1,h_2}, as the two parameter extension of the *R*-matrix in Ref. 3,

$$R_{h_1,h_2} = \begin{pmatrix} 1 - h_1 h_2 & h_2 & h_2 & 0 \\ -h_1 & -1 & -h_1 h_2 & -h_2 \\ -h_1 & -h_1 h_2 & -1 & h_2 \\ 0 & h_1 & -h_1 & 1 + h_1 h_2 \end{pmatrix}, \tag{4.4}$$

which gives the (h_1,h_2)-deformed algebra of functions on $\mathrm{Gr}_{h_1,h_2}(1|1)$ with the equation

$$R_{h_1,h_2} \hat{T}_1 \hat{T}_2 = -\hat{T}_2 \hat{T}_1 R_{-h_1,-h_2}. \tag{4.5}$$

V. HOPF ALGEBRA STRUCTURE OF $\mathrm{Gr}_{h_1,h_2}(1|1)$

We denote the algebra generated by the matrix elements α, b, c, δ of $\hat{T}$ with the relations (3.10) by $\hat{\mathcal{A}}$. To make the algebra $\hat{\mathcal{A}}$ into a Hopf algebra related to the quantum supergroup $\mathrm{GL}_{h_1,h_2}(1|1)$, we state briefly some properties of the quantum supergroup $\mathrm{GL}_{h_1,h_2}(1|1)$ we are going to need in this section.

The quantum supergroup $\mathrm{GL}_{h_1,h_2}(1|1)$ generated by four generators a, β, γ, d and the (h_1,h_2)-commutation relations[4]

$$\begin{gathered} a\beta=\beta a-h_2(a^2-\beta\gamma-ad), \quad d\beta=\beta d+h_2(d^2+\beta\gamma-da), \\ a\gamma=\gamma a+h_1(a^2+\gamma\beta-ad), \quad d\gamma=\gamma d-h_1(d^2-\gamma\beta-da), \\ \beta^2=h_2\beta(a-d), \quad \gamma^2=h_1\gamma(a-d), \\ \beta\gamma=-\gamma\beta+(h_1\beta-h_2\gamma)(d-a), \quad ad=da+h_1\beta(a-d)+h_2(a-d)\gamma. \end{gathered} \tag{5.1}$$

The generators satisfying the relations (5.1) generate the algebra called the algebra of functions on the quantum supergroup $\mathrm{GL}_{h_1,h_2}(1|1)$ and we shall denote it by $\mathcal{A}$. We know that[4] the algebra $\mathcal{A}$ is a graded Hopf algebra with the following structure: the coproduct

$$\Delta:\mathcal{A}\to\mathcal{A}\otimes\mathcal{A}, \quad \Delta(t^i_j)=t^i_k\otimes t^k_j, \tag{5.2}$$

the counit

$$\epsilon:\mathcal{A}\to\mathcal{C}, \quad \epsilon(t^i_j)=\delta^i_j, \tag{5.3}$$

and the antipode

$$S:\mathcal{A}\to\mathcal{A}, \quad S(t^i_j)=(t^i_j)^{-1}, \tag{5.4}$$

where $t^i_j\in\{a,\beta,\gamma,d\}$.

It is not difficult to check the Hopf algebra axioms.[16] If we represent the set of generators a, β, γ, d in the form of a matrix

$$T=\begin{pmatrix} a & \beta \\ \gamma & d \end{pmatrix}, \tag{5.5}$$

then the relations (5.1) are equivalent to the equation

$$RT_1T_2=T_2T_1R, \tag{5.6}$$

where[4]

$$R=\begin{pmatrix} 1-h_1h_2 & -h_2 & h_2 & 0 \\ -h_1 & 1 & -h_1h_2 & h_2 \\ h_1 & -h_1h_2 & 1 & h_2 \\ 0 & h_1 & h_1 & 1+h_1h_2 \end{pmatrix} \tag{5.7}$$

is the solution of the quantum Yang–Baxter equation.

Since the (graded) Hopf algebra structure of $\mathrm{Gr}_{h_1,h_2}(1|1)$ will be related to those of $\mathrm{GL}_{h_1,h_2}(1|1)$, it is necessary to obtain the commutation relations of the generators of $\hat{\mathcal{A}}$ with those of $\mathcal{A}$. We obtain the commutation relations between the generators of $\hat{\mathcal{A}}$ and $\mathcal{A}$ as follows:

$$\hat{T}_1 T_2 = (-1)^{p(T_2)} R^{-1} T_2 \hat{T}_1 R. \tag{5.8}$$

Equation (5.8) gives

$$\begin{aligned}
\alpha a &= a\alpha + h_1(ab+\beta\alpha) - h_2(ac+\gamma\alpha) - h_1h_2(a\delta - \beta c + \gamma b + d\alpha + 2a\alpha),\\
\alpha\beta &= -\beta\alpha + h_1\beta b + h_2(a\alpha - \beta c + d\alpha) - h_1h_2(ab - \beta\alpha - \beta\delta + db),\\
\alpha\gamma &= -\gamma\alpha + h_1(a\alpha + \gamma b + d\alpha) - h_2\gamma_c - h_1h_2(ac - \gamma\alpha - \gamma\delta + dc),\\
\alpha d &= d\alpha - h_1(\beta\alpha - db) + h_2(\gamma\alpha - dc) + h_1h_2(a\alpha - \beta c + \gamma b - d\delta),\\
ba &= ab + h_1\beta b - h_2(a\alpha + a\delta + \gamma b) - h_1h_2(ab - \beta\alpha - \beta\delta + db),\\
b\beta &= \beta b - h_2(ab - \beta\alpha - \beta\delta + db),\\
b\gamma &= \gamma b - h_1(ab + db) - h_2(\gamma\alpha - \gamma\delta) + h_1h_2(a\alpha + a\delta + d\alpha + d\delta),\\
bd &= db - h_1\beta b + h_2(\gamma b - d\alpha - d\delta) + h_1h_2(ab - \beta\alpha - \beta\delta + db),\\
ca &= ac - h_1(a\alpha + a\delta - \beta c) - h_2\gamma c - h_1h_2(ac - \gamma\alpha - \gamma\delta + dc),\\
c\beta &= \beta c + h_1(\beta\alpha + \beta\delta) - h_2(ac + dc) - h_1h_2(a\alpha + a\delta + d\alpha + d\delta),\\
c\gamma &= \gamma c - h_1(ac - \gamma\alpha - \gamma\delta + dc),\\
cd &= dc - h_1(\beta c + d\alpha + d\delta) + h_2\gamma c + h_1h_2(ac - \gamma\alpha - \gamma\delta + dc),\\
\delta a &= a\delta - h_1(ab - \beta\delta) + h_2(ac - \gamma\delta) + h_1h_2(a\alpha - \beta c + \gamma b - d\delta),\\
\delta\beta &= -\beta\delta - h_1\beta b + h_2(a\delta + \beta c + d\delta) + h_1h_2(ab - \beta\alpha - \beta\delta + db),\\
\delta\gamma &= -\gamma\delta + h_1(a\delta - \gamma b + d\beta) + h_2\gamma c + h_1h_2(ac - \gamma\alpha - \gamma\delta + dc),\\
\delta d &= d\delta - h_1(\beta\delta + db) - h_2(\gamma\delta + dc) + h_1h_2(a\delta + \beta c - \gamma b + d\alpha + 2d\delta).
\end{aligned} \tag{5.9}$$

Using these relations, it can be checked that $\hat{D}_{h_1,h_2}$, which is given by (3.18), is still a central element, i.e., $\hat{D}_{h_1,h_2}$ also commutes with the generators of $\mathcal{A}$.

Before defining a coproduct on the algebra $\hat{\mathcal{A}}$, let us note the following facts. Let $\hat{T}$ and $\hat{T}'$ be any two supercommuting quantum matrices whose elements satisfy (3.10). We denote a product $\hat{T}\hat{T}'$ by T. Then, it can be verified that the matrix elements of T satisfy the commutation relations (5.1), i.e., if

$$\hat{T}, \hat{T}' \in \mathrm{Gr}_{h_1,h_2}(1|1) \Rightarrow T = \hat{T}\hat{T}' \in \mathrm{GL}_{h_1,h_2}(1|1).$$

In view of these facts, we can say that there may be no coproduct of the form $\Delta(\hat{T}) = \hat{T}\dot{\otimes}\hat{T}$. For, this coproduct, if it existed, would be invariant under the (h_1,h_2)-commutation relations (5.1) of $\mathrm{GL}_{h_1,h_2}(1|1)$. But we can define a coproduct on the algebra $\hat{\mathcal{A}}$ as follows (see Ref. 17, for more details):

$$\hat{\Delta}: \hat{\mathcal{A}} \to \hat{\mathcal{A}} \otimes \hat{\mathcal{A}}, \quad \hat{\Delta}(\hat{T}) = \hat{T}\dot{\otimes} T + (-1)^{p(T)} T \dot{\otimes} \hat{T}. \tag{5.10}$$

Explicity,

$$\begin{aligned}
\hat{\Delta}(\alpha) &= \alpha \otimes a + b \otimes \gamma + a \otimes \alpha - \beta \otimes c, \\
\hat{\Delta}(b) &= b \otimes d + \alpha \otimes \beta + a \otimes b - \beta \otimes \delta, \\
\hat{\Delta}(c) &= c \otimes a + \delta \otimes \gamma - \gamma \otimes \alpha + d \otimes c, \\
\hat{\Delta}(\delta) &= \delta \otimes d + c \otimes \beta - \gamma \otimes b + d \otimes \delta.
\end{aligned} \tag{5.11}$$

The action on the generators of $\hat{\mathcal{A}}$ of $\hat{\epsilon}:\hat{\mathcal{A}} \to \mathcal{C}$ is

$$\hat{\epsilon}(\alpha) = \hat{\epsilon}(b) = \hat{\epsilon}(c) = \hat{\epsilon}(\delta) = 0. \tag{5.12}$$

Finally, we define the coinverse as

$$\hat{S}:\hat{\mathcal{A}} \to \hat{\mathcal{A}}, \quad \hat{S}(\hat{T}) = -(-1)^{p(T^{-1})} T^{-1} \hat{T} T^{-1}. \tag{5.13}$$

It can be checked that the maps $\hat{\Delta}$ and $\hat{\epsilon}$ are both algebra homomorphisms and $\hat{S}$ is an algebra antihomomorphism and the three maps satisfy the following properties of the costructures:

$$\begin{aligned}
(\hat{\Delta} \otimes \mathrm{id}) \circ \hat{\Delta} &= (\mathrm{id} \otimes \hat{\Delta}) \circ \hat{\Delta}, \\
\mu \circ (\hat{\epsilon} \otimes \mathrm{id}) \circ \hat{\Delta} &= \mu' \circ (\mathrm{id} \otimes \hat{\epsilon}) \circ \hat{\Delta}, \\
m \circ (\hat{S} \otimes \mathrm{id}) \circ \hat{\Delta} = \hat{\epsilon} &= m \circ (\mathrm{id} \otimes \hat{S}) \circ \hat{\Delta},
\end{aligned} \tag{5.14}$$

where id denotes the identity mapping, $\mu:\mathcal{C} \otimes \hat{\mathcal{A}} \to \hat{\mathcal{A}}$, $\mu':\hat{\mathcal{A}} \otimes \mathcal{C} \to \hat{\mathcal{A}}$ are the canonical isomorphisms, defined by $\mu(k \otimes a) = ka = \mu'(a \otimes k)$, $\forall a \in \hat{\mathcal{A}}$, $\forall k \in \mathcal{C}$, and m is the multiplication map $m:\hat{\mathcal{A}} \otimes \hat{\mathcal{A}} \to \hat{\mathcal{A}}$, $m(a \otimes b) = ab$.

ACKNOWLEDGMENTS

This work was supported in part by T.B.T.A.K., the Turkish Scientific and Technical Research Council. We would like to express our deep gratitude to the referee for suggestions on the manuscript.

[1] A. Aghamohammadi, M. Khorrami, and A. Shariati, J. Phys. A **28**, L225 (1995).
[2] L. Dabrowski and P. Parashar, Lett. Math. Phys. **38**, 331 (1996).
[3] S. Çelik, Balkan Phys. Lett. **5**, 149 (1997).
[4] S. Çelik, Lett. Math. Phys. **42**, 299 (1997).
[5] E. Demidov, Yu I. Manin, E. Muhkin, and D. Zhdanovich, Prog. Theor. Phys. Suppl. **102**, 203 (1990).
[6] H. Ewen, O. Ogievetsky, and J. Wess, Lett. Math. Phys. **22**, 297 (1991).
[7] S. Zakrzewski, Lett. Math. Phys. **22**, 287 (1991).
[8] S. L. Woronowicz, Rep. Math. Phys. **30**, 259 (1991).
[9] C. H. Ohn, Lett. Math. Phys. **25**, 85 (1992).
[10] V. Karimipour, Lett. Math. Phys. **30**, 87 (1994).
[11] B. A. Kupershmidt, J. Phys. A **25**, L1239 (1992).
[12] A. Aghamohammadi, Mod. Phys. Lett. A **8**, 2607 (1993).
[13] S. Çelik and E. Hizel, J. Phys. A **30**, 4677 (1997).
[14] S. Çelik, J. Math. Phys. **37**, 3568 (1996).
[15] Yu I. Manin, Commun. Math. Phys. **123**, 163 (1989).
[16] E. Abe, *Hopf Algebras* (Cambridge University Press, Cambridge, 1980).
[17] S. Çelik, "Hopf algebra structure of $\mathrm{Gr}_q(1|1)$ related to $\mathrm{GL}_q(1|1)$," J. Math. Phys. (to be published).

Instability of vector fields induced by first integrals

F. G. Gascón and D. Peralta Salas
Departamento de Física Teórica II, Fac. CC. Físicas, Universidad Complutense, Madrid 28040, Spain

(Received 11 September 1998; accepted for publication 9 March 1999)

It is shown that when a first integral of a $\mathbb{R}^3$ vector field **X** is known, instabilities are induced on the equilibrium points of **X**. © *1999 American Institute of Physics.* [S0022-2488(99)03106-0]

I. INTRODUCTION

Let **X** be an analytic (C^w) $\mathbb{R}^n$ vector field (v.f.) with an isolated singularity at the origin, i.e., $\mathbf{X}(\mathbf{0})=\mathbf{0}$. We are concerned here with establishing criteria for the instability of **X** at **0** (the origin).

It is a classical result that when the linear part $\mathbf{X}_L$ of **X** at **0** has an eigenvalue of positive real part then **0** is an unstable equilibrium point of **X** (Ref. 1). This criterion gives no information concerning instability when there are not eigenvalues of $\mathbf{X}_L$ to the right of the imaginary axis.

When $\mathbf{X}_\mathrm{H}$ is a Hamiltonian v.f. and H is an analytic function of the form

$$H=\sum_{i,j=1}^{m} p_i p_j a_{ij}(\mathbf{q})+V(\mathbf{q}) \quad (\mathbf{q}\in\mathbb{R}^m, \quad n=2m), \tag{1}$$

and (i) $a_{ij}(\mathbf{q})$ is definite positive for any **q**, (ii) **0** is a critical point of V, (iii) **0** is not a strict minimum of V, and (iv) $m=1,2$. Then **0** is an unstable equilibrium point of $\mathbf{X}_\mathrm{H}$ (Ref. 2).

When $m>2$, the instability of $\mathbf{X}_\mathrm{H}$ at **0**, under the above assumptions, is an unproved conjecture. Nevertheless, the unstable behavior of $\mathbf{X}_\mathrm{H}$ at **0** has been obtained under additional requirements on $V(\mathbf{q})$ (Ref. 3).

The stability of periodic solutions of Hamiltonian v.f. when first integrals are known has also been recently investigated (Ref. 4).

The technique proposed in this paper is valid for $\mathbb{R}^3$ v.f. with an isolated singularity (equilibrium point) at **0** and with a known C^w first integral I. The technique is illustrated with examples that show that the method is valid, even in the case of trivial center (that is, when all the eigenvalues of $\mathbf{X}_L$ lie on the imaginary axis).

The method proposed here is based on the well-known fact that the w limit of a bounded trajectory of a planar vector field must include either a singularity or a closed trajectory (Bendixon–Poincaré theorem).

The possibilities of extending the new technique to $\mathbb{R}^n$ v.f. ($n>3$) are also discussed.

II. INSTABILITY INDUCED BY FIRST INTEGRALS

Let **X** be a $\mathbb{R}^3$ dynamical system with an isolated singularity at **0** and I a C^w first integral of **X**. Assume that either

$$\text{(i)} \quad \nabla I_{|\mathbf{0}}\neq\mathbf{0},$$

or (2)

$$\text{(ii)} \quad \nabla I(\mathbf{P})=\mathbf{0},\ \ \mathbf{P}\in N_0\Rightarrow\mathbf{P}=(0,0,0),$$

and I has a saddle at the origin.

Then **0** is an unstable equilibrium point of **X**.

Remember that ∇I stands for the gradient of I. On the other hand, I by definition has a saddle at **0** if $\nabla I_{|\mathbf{0}}=\mathbf{0}$ and there are points P and Q arbitrarily near **0** on which I takes values of opposite signs. Remember that we assume in this paper that the first integral I has the value 0 at **0**.

Proof: Assume that $\nabla I_{|\mathbf{0}}\neq\mathbf{0}$. In this case local coordinates (u_1,u_2,u_3) can be introduced on N_0 (a neighborhood of **0**) on which I takes the canonical form $I=u_1$. Therefore, if **X** is assumed stable at **0** its trajectories will lie in N_0 and on the local planes $u_l=k_1$. The w limits of these trajectories must be, on account of the Poincaré–Bendixon theorem,[5] singular points of **X**, polygonals whose vertices are singular points of **X** or closed trajectories.

In any of these three cases, singular points of **X**, lying on the planes $u_1=k_1$ and arbitrarily near **0**, are obtained. But since **0** was assumed to be an isolated singularity of **X**, we get a contradiction. Therefore **X** cannot be stable at **0**.

Assume now that ∇I vanishes on N_0 just at **0** and that I has a saddle at **0**.

These assumptions imply (as we now explain) that on a certain domain $Z_0\subset N_0$ the level sets of I resemble locally topological planes, to which the above reasoning can be applied, getting again a contradiction if **0** is assumed to be a stable singularity of **X**. Therefore **0** must be an unstable singularity of **X**, as we desired to prove.

We now show that if I is an $\mathbb{R}^3$ analytic function with a saddle at **0** and $\nabla I_{|N_0}$ vanishes just at **0**, then a domain $Z_0\subset N_0$ exists on which the sets $I^{-1}(c)\cap Z_0$ are local planes (disks).

In fact, the analiticity of I implies that $I^{-1}(0)\cap N_0$ is the finite union of the surfaces C_i, $i\in J$, through **0**. Condition (ii) of Eq. (2) implies that the surfaces C_i do not intersect each other on $N_0-\{\mathbf{0}\}$. The surfaces C_i divide N_0 into solid zones Z_j, whose boundary is made up of one or several of the surfaces C_i.

By topological reasons it is not too difficult to show that one at least (say Z_0) of the zones Z_j is diffeomorphic to $\mathbb{R}^3$. This is due to the fact that C_i is, inside N_0, either a topological plane (if C_i has a tangent at **0**) or a topological cone (if C_i has not a tangent plane at **0**); in any case, each C_i separates N_0 into zones, one of which is clearly diffeomorphic to $\mathbb{R}^3$. This property, valid for any of the surfaces C_i, is the geometric reason underlying the existence of the zone Z_0.

For example, consider the functions $I_1=(x^2+y^2-z^2)z$, $I_2=(x^2+y^2-z^2)(x^2+y^2-4z^2)$. I_1 and I_2 have clearly a saddle at **0**, and it is easy to check that ∇I_i $(i=1,2)$ vanishes just at **0**. The set Z_0 diffeomorphic to $\mathbb{R}^3$ can be chosen to be

$$Z_0=\{(x,y,z)|x^2+y^2<z^2,\quad z>0\}. \tag{3}$$

Consider now the C^w curves φ_α, defined either by

$$\varphi_\alpha=I^{-1}(c)\cap Z_0\cap\pi_\alpha, \tag{4}$$

π_α standing for a family of planes through **0**, intersecting Z_0, or by

$$I_{|\pi_\alpha\cap Z_0}=c. \tag{5}$$

Calling $I_{|\pi_\alpha}$ by I^*_α, we have the following.

(1) (0,0) is a saddle of I^*_α. This is a consequence of the fact that the sign of I changes on the surfaces C_i, since otherwise $\nabla I=\mathbf{0}$ on points of $N_0-\{\mathbf{0}\}$.

(2) ∇I^*_α has an isolated zero at (0,0). In fact, if $\nabla I^*_{\alpha|\varphi}=\mathbf{0}$, where φ is a curve through (0,0) we would get $I_{|\varphi}=0$, in contradiction with the fact that $I\neq0$ inside Z_0. A similar contradiction is obtained if ∇I^*_α vanishes on a succession of points tending to (0,0).

Summarizing, the curves φ_α are the zeros of plane C^w functions with a saddle at (0,0) and an isolated critical point at (0,0). Therefore (Ref. 6), φ_α is just an open segment. The union of these segments, when the plane π_α varies is, given the topology of Z_0, a local plane (a disk).

Therefore $I^{-1}(c)\cap Z_0$ is locally a plane.

The reasoning above is sketchy and probably can be improved.

We have not been able to improve it by consultations with professional mathematicians. We now give some examples of $\mathbb{R}^3$ v.f. whose instability at $\mathbf{0}$ can be detected with the above-mentioned techniques. To our knowledge they cannot be integrated via quadratures and they are interesting since most of them have a vanishing linear part.

A. Consider the $\mathbb{R}^3$ v.f.

Here

$$\mathbf{X}=(y(1+z^2)-x-z)\partial_x+(-y-x(1+z^2))\partial_y+((x^2+y^2+xz)(1+z^2))\partial_z. \tag{6}$$

It is easy to check that this v.f. has an isolated zero at (0,0,0) and the eigenvalues of $\mathbf{X}_L$ at $\mathbf{0}$ are 0 and $-1\pm i$. Therefore the eigenvalues cannot decide between stability and instability at $\mathbf{0}$.

This v.f. has the first integral $I=\frac{1}{2}(x^2+y^2)-\arctan(z)$. Note that $\nabla I_{|\mathbf{0}}\neq\mathbf{0}$. Therefore by (i) of Eq. (2), $\mathbf{X}$ is unstable at $\mathbf{0}$.

B. Consider the v.f.

Here

$$\mathbf{X}=(x^4(y^2+z^2)+x^2+y^2)\partial_x-(2x^3(y^2+z^2)(1+y)+x^2+y^2+z^2)\partial_y+(2x(x^2+y^2)(1+y)$$
$$-x^2(x^2+y^2+z^2))\partial_z. \tag{7}$$

It is easy to check that (i) $\mathbf{0}$ is an isolated zero of $\mathbf{X}$ and that $\mathbf{X}_L$ (the linear part of $\mathbf{X}$ at $\mathbf{0}$) is identically zero; (ii) $I=x^2(1+y)-z$ is a first integral of $\mathbf{X}$. (iii) $\nabla I_{|\mathbf{0}}\neq\mathbf{0}$.

Therefore, according to (i) of Eq. (2), $\mathbf{0}$ is an unstable singular point of $\mathbf{X}$.

C. Let X be the v.f.

Here

$$\mathbf{X}=(2x(y-z)(x^2+y^2+z^2))\partial_x-((3x^2+y^2+z^2)(x^2+y^2+z^2)2x^2yz)\partial_y$$
$$+((3x^2+y^2+z^2)(x^2+y^2+z^2)+2x^2y^2)\partial_z. \tag{8}$$

It is easy to check that (i) $\mathbf{0}$ is an isolated zero of $\mathbf{X}$ and $\mathbf{X}_L=0$; (ii) $I=x(x^2+y^2+z^2)$ is a first integral of $\mathbf{X}$. The first integral has a saddle at $\mathbf{0}$ and its gradient vanishes just at $\mathbf{0}$.

According to (ii) of Eq. (2), $\mathbf{0}$ is an unstable singularity of $\mathbf{X}$.

D. Consider the $\mathbb{R}^3$ v.f.

Here

$$\mathbf{X}=-2(y^2+zx^2+zxy^2+xz^3)\partial_x+(-(z^2+3x^2)y+2xyz(x^2+y^2+z^2))\partial_y$$
$$+((x^2+y^2+z^2)(3x^2+2y^2+z^2))\partial_z. \tag{9}$$

It is easy to check that (i) $\mathbf{0}$ is an isolated zero of $\mathbf{X}$ and $\mathbf{X}_L=\mathbf{0}$; (ii) $I=z^2x+x^3-y^2$ is a first integral of $\mathbf{X}$. In addition, I has a saddle at $\mathbf{0}$ and ∇I vanishes just at $\mathbf{0}$.
Therefore, by applying (ii) of Eq. (2), we can conclude that $\mathbf{X}$ is unstable at $\mathbf{0}$.

We conclude by noting that our instability criterion can be applied to $\mathbb{R}^n$ v.f. $(n>3)$ when $(n-2)$ first integrals of $\mathbf{X}$ are known and $\operatorname{rank}(\nabla I_1,\ldots,\nabla I_{n-2})_{|\mathbf{0}}=n-2$. This can be seen by introducing local coordinates $(u_1,\ldots,u_n)$ in N_0 on which the first integrals take the local form $I_1=u_1,\ldots,I_{n-2}=u_{n-2}$.

Therefore the local level sets of $I_1,\ldots,I_{n-2}$ will be local planes (two-dimensional disks). By applying to them the considerations used to demonstrate of Eq. (2), we get instability of $\mathbf{X}$ at $\mathbf{0}$.

When $\text{rank}(\nabla I_1,\ldots,\nabla I_{n-2})_{|\mathbf{0}}<n-2$, a general criterion for instability seems difficult to get. We list now several partial results in this direction.

(i) Let $\mathbf{X}$ be a $\mathbb{R}^4$ v.f. with an isolated zero at $\mathbf{0}$ and the two first integrals:

$$I_1=(1+x^2)y^2+(1+x^4)z^2-(1+e^x)u^4,$$
$$I_2=x. \tag{10}$$

Note that $\text{rank}(\nabla I_1,\nabla I_2)_{|\mathbf{0}}=1$.

On the level set $I_2=0$, I_1 and $\mathbf{X}$ become

$$I_1^*=y^2+z^2-2u^4,$$
$$\mathbf{X}^*=a\partial_y+b\partial_z+c\partial_u. \tag{11}$$

It is clear that ∇I_1^* vanishes just at (0,0,0), that $\mathbf{X}^*$ has an isolated zero at (0,0,0), and that I_1^* has a saddle at $\mathbf{0}$. Therefore the couple $(\mathbf{X}^*,I_1^*)$ satisfies the assumptions of (ii) of Eq. (2), and we conclude that $\mathbf{X}^*$, and therefore $\mathbf{X}$, is unstable at (0,0,0,0).

Examples of this type are not only academic, since they appear in the study of systems of the type

$$\ddot{x}=V_{,x}(x,y),$$
$$\ddot{y}=V_{,y}(x,y), \tag{12}$$

whenever a pair of first integrals of a $\mathbb{R}^4$ v.f. are known and the gradient of one of them does not vanish at $\mathbf{0}$ (Ref. 7). The second first integral is, usually, linear in the components of the velocity.

In fact, via a local change of variables this first integral can be reduced to a canonical form similar to the function I_2 of (10). This fact gives generality to the couple of first integrals chosen in (10).

(ii) Let I_1 and I_2 be defined by

$$I_1=u^n-P(x,y,z),$$
$$I_2=x^m-Q(y,z), \tag{13}$$

where n and m are positive integers $(n,m>1)$, P and Q non-negative polynomials and $\text{rank}(\nabla I_1,\nabla I_2)_{|\mathbf{0}}=0$.

It is immediate to check that the level sets

$$I_1=C_1,$$
$$I_2=C_2, \tag{14}$$

are planes when $C_1,C_2>0$ (one has just to get u and x as global functions of y and z). Therefore, by using similar arguments to those given in the proof of (2), any $\mathbb{R}^4$ v.f. with an isolated zero at $\mathbf{0}$ and these first integrals is unstable at $\mathbf{0}$.

(iii) Let I_1 and I_2 be defined by

$$I_1=y^2-f(x),$$
$$I_2=xu-zy, \tag{15}$$

$f(x)$ being a non-negative function and $f'(0)=0$.

Note that I_2 has the form of an angular momentum and that $\text{rank}(\nabla I_1,\nabla I_2)_{|\mathbf{0}}=0$. On the other hand, the level sets,

$$y^2 - f(x) = C,$$
$$xu - zy = D, \tag{16}$$
$$C > 0,$$

can be globally parametrized in the form

$$\left(x, \pm\sqrt{C+f(x)}, \frac{xu-D}{\pm\sqrt{C+f(x)}}, u \right), \tag{17}$$

and they are a couple of two-dimensional planes (note that the parameters x and u are free). Therefore any v.f. with an isolated zero at **0** and the two first integrals (15) is unstable at **0**.

[1] L. Perko, *Differential Equations and Dynamical Systems* (Springer-Verlag, New York, 1996); S. Wiggins, *Introduction to Applied Non Linear Dynamical Systems and Chaos* (Springer-Verlag, New York, 1990); F. Gantmacher, *Lectures on Analytical Mechanics* (Mir. Moscow, 1975).
[2] V. Rumiantsev and S. Sosmiksmi, J. Appl. Math. Mech. **57**, 1101 (1993); V. Palamodov, Funct. Anal. Appl. **11**, 277 (1977).
[3] S. Tagliaferro, Arch. Ration. Mech. Anal. **109**, 183 (1989); **73**, 183 (1980); V. Kozlov, Appl. Math. Mech. (in Chinese) **50**, 719 (1986).
[4] V. Kozlov, Russ. Math. Surv. **38**, 1 (1983); H. Yoshick, Physica A **210**, 163 (1986).
[5] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* (McGraw-Hill, New York, 1955). See also Perko in Ref. 1.
[6] A. Buno, *Local Methods in Nonlinear Differential Equations* (Springer-Verlag, Berlin, 1989).
[7] G. D. Birkhoff, *Dynamical Systems*, Am. Math. Soc. Colloq. Publ., 1927; S. Ichtiaroglov, J. Phys. A **20**, 5079 (1987); H. Yehia, *ibid.* **25**, 197 (1992).

Nonlinear superposition formulas based on imprimitive group action

M. Havlíček[a)] and S. Pošta[b)]
Department of Mathematics and Doppler Institute, FNSPE, Czech Technical University, Trojanova 13, CZ-120 00 Prague 2, Czech Republic

P. Winternitz[c)]
Centre de Recherches Mathématiques, Université de Montréal, C. P. 6128, Succ. Centre-ville, Montréal, Québec H3C 3J7, Canada

(Received 2 November 1998; accepted for publication 14 February 1999)

Systems of nonlinear ordinary differential equations are constructed for which the general solution is expressed algebraically in terms of a finite number of particular solutions. The equations and the corresponding nonlinear superposition formula are based on a nonlinear action of the Lie group $SL(N,\mathbb{C})$ on a homogeneous space M. The isotropy group of the origin of this space is a nonmaximal parabolic subgroup of $SL(N,\mathbb{C})$. Such equations can occur as Bäcklund transformations for soliton equations on flag manifolds.
[S0022-2488(99)00306-0]

I. INTRODUCTION

Let us consider a system of n first order ordinary differential equations (ODEs),

$$\dot{y}^{\mu}=\eta^{\mu}(\mathbf{y},t) \quad \mu=1,...,n, \tag{1.1}$$

where the dot denotes differentiation with respect to time t.

If the equations are linear we have a linear superposition formula; the general solution is a linear combination of n linearly independent particular solutions. More interestingly, even if the system (1.1) is nonlinear, it may allow a nonlinear superposition formula

$$\begin{gathered}\mathbf{y}(t)=\mathbf{F}(\mathbf{y}_1,\ldots,\mathbf{y}_m,c_1,\ldots,c_n),\\ \mathbf{y}\in\mathbb{C}^n \quad (\text{or } \ \mathbf{y}\in\mathbb{R}^n),\end{gathered} \tag{1.2}$$

where $\mathbf{y}_1,\ldots,\mathbf{y}_m$ are particular solutions, $c_1,\ldots,c_n$ are arbitrary constants and $\mathbf{y}(t)$ is the general solution.

Lie[1] established the conditions under which the system (1.1) allows a superposition formula (1.2), i.e., the general solution can be expressed as a function of a finite number m of particular solutions. Lie's result can be summed up as follows:

The system (1.1) allows a superposition formula (1.2) if and only if

(1) It has the form

$$\dot{\mathbf{y}}=\sum_{k=1}^{r} Z_k(t)\boldsymbol{\xi}_k(\mathbf{y}). \tag{1.3}$$

(2) The vector functions $\boldsymbol{\xi}_k(\mathbf{y})$ (independent of t) are such that the vector fields

---

[a)]Electronic mail: havlicek@km1.fjfi.cvut.cz
[b)]Electronic mail: severin@km1.fjfi.cvut.cz
[c)]Electronic mail: wintern@CRM.UMontreal.CA

$$\hat{X}_k = \sum_{\mu=1}^{n} \xi_k^{\mu}(\mathbf{y})\partial_{y^{\mu}} \tag{1.4}$$

generate a finite dimensional Lie algebra L,

$$[X_k, X_l] = \sum_{j=1}^{r} f_{klj} X_j . \tag{1.5}$$

The number of solutions m needed in the superposition formula (1.2) satisfies the relation

$$mn \geqq r, \tag{1.6}$$

where n is the number of equations and r is the dimension of the Lie algebra L.

Somewhat unexpectedly, it turned out that nonlinear ordinary differential equations with superposition formulas play an important role in soliton theory[2] where they occur as Bäcklund transformations.

In turn, Bäcklund transformations provide soliton superposition formulas, i.e., explicit formulas for multisoliton solutions that asymptotically correspond to a combination of independent solitons. Thus two very different types of nonlinear superposition formulas become linked via Bäcklund transformations for integrable nonlinear partial differential equations.

The classification and construction of all systems of n nonlinear ODEs with superposition formulas amounts to a classification of all finite dimensional subalgebras of diff(n), the infinite dimensional Lie algebra of vector fields in n dimensions. For $n=1$ such a classification is quite simple. Indeed the only finite-dimensional subalgebras of diff$(1,\mathbb{C})$ are sl$(2,\mathbb{C})$ and its subalgebras. For $n=2$ a complete classification exists and this is again due to Lie.[3] For $n \geqq 3$ the problem becomes intractable.

A more resctricted problem has however been solved[4–6] and that is the classificaton of indecomposable systems of ODEs with superposition formulas. These are systems satisfying Eqs. (1.3), (1.4), and (1.5) from which it is not possible to split off a subset of $l<n$ equations that also satisfy Lie's criteria and hence have a superposition formula of their own.

The classification of these indecomposable systems is best formulated in a geometric manner. Thus, let us view the variables $\{y_1,\ldots,y_n\}$ (for fixed t), as local coordinates on some manifold M. Let G be a Lie group acting transitively and effectively on M. We can then identify M with a quotient space $M \sim G/G_0$, where $G_0 \subset G$ is the isotropy group of the origin in M.

The system (1.1) is decomposable if local coordinates on M exist that can be divided into two subsets, $\{y_1,\ldots,y_n\} \sim \{x_1,\ldots,x_l,z_{l+1},\ldots,z_n\}$ such that the vector fields (1.4) all simultaneously have the form

$$X_k = \sum_{a=1}^{l} f_k(x)\partial_{x_a} + \sum_{b=l+1}^{n} g_k(x,z)\partial_{z_b} \tag{1.7}$$

(i.e., the coefficients of ∂_{x_a} depend on the coordinates x only). Such coordinates exist if there exists a G-invariant foliation of the space M. The action of G on M is called primitive (in addition to being transitive and effective) if no such foliation exists. Locally, this can be expressed in terms of the Lie algebras L and L_0, corresponding to G and G_0. The system (1.1) is indecomposable if the pair of algebras (L_0, L) determines a transitive primitive Lie algebra. This means that L_0, the subalgebra of vector fields vanishing at the origin, must be a maximal subalgebra of L, and must not contain an ideal of L.

Transitive primitive Lie algebras have been classified.[7–11] In turn, this classification was used to classify indecomposable systems of ODEs with superposition formulas.[4–6] Several articles have been devoted to constructing systems of equations with superposition formulas, and to the superposition formulas themselves.[12–17] Supersymmetric versions of these equations have been constructed[18] as well as difference equations with superposition formulas.[19,20] All cases considered so far correspond to transitive primitive Lie algebras.

The purpose of this article is to investigate the nonprimitive case and to show how the previously studied indecomposable systems serve as building blocks for decomposable ones.

We restrict ourselves to the Lie algebras $\mathrm{sl}(N,\mathbb{C})$ (for arbitrary finite N) and make use of realizations of these algebras, constructed earlier.[21]

II. FORMULATION OF THE PROBLEM AND EXAMPLE OF SL(2,$\mathbb{C}$)

A. General formulation

Let us consider a system of ODEs as in Eq. (1.3), allowing a superposition formula. In view of Eqs. (1.4) and (1.5) the right-hand side of Eq. (1.3) determines an element of the Lie algebra L for any fixed value of time t. As t varies, this element varies along some path in L. The general form of the solution is obtained by integrating the vector fields (1.4) and composing the results. Thus the solution will be given by the corresponding group action

$$\mathbf{y}(t)=g(t)\cdot\mathbf{u}, \tag{2.1}$$

where $g(t)$ is an element of the Lie group $G=\langle\exp L\rangle$ and $\mathbf{u}$ is a constant vector, related to the initial conditions for $\mathbf{y}(t)$. The group element $g(t)$ depends on $r=\dim L$ group parameters. These in turn depend on time t in such a way that $g(t)$ follows a path in the group G, corresponding to the path in L, determined by the equation, i.e., by the coefficients $Z_k(t)$, $k=1,...,r$.

The superposition formula (1.2) is obtained from the group action (2.1), once the time dependence in $g(t)$ is established. To do this, we assume that we know m solutions $\mathbf{y}_k(t)$, $k=1,...,m$ corresponding to the initial values u_k in (2.1). The m relations

$$\mathbf{y}_k(t)=g(t)\mathbf{u}_k \tag{2.2}$$

are then used to express all parameters in $g(t)$ in terms of the known solutions. Each solution provides n equations. The condition (1.6) simply means that we must have at least as many equations as unknowns. The actual number m of different solutions needed is obtained from the requirement that the only transformation $g(t_0)$ that simultaneously stabilizes all initial conditions u_k ($u_k=g(t_0)u_k$, $k=1,...,m$) is the identity transformation $g(t_0)=I$. This requirement determines the minimal number m and also the independence conditions on $u_1,\ldots,u_m$.

B. Example of the algebra sl(2,$\mathbb{C}$)

For $n=1$ the situation is very simple. The only finite dimensional subalgebras of $\mathrm{diff}(1,\mathbb{C})$ are $\mathrm{sl}(2,\mathbb{C})$ and its one and two-dimensional subalgebras. Up to local diffeomorphisms there exists just one realization of $\mathrm{sl}(2,\mathbb{C})$ as a subalgebra of $\mathrm{diff}(1,\mathbb{C})$, namely,

$$X_1=\partial_y,\quad X_2=y\partial_y,\quad X_3=y^2\partial_y. \tag{2.3}$$

Equation (1.3) in this case is the Riccati equation

$$\dot{y}=Z_1(t)+Z_2(t)y+Z_3(t)y^2. \tag{2.4}$$

The $\mathrm{SL}(2,\mathbb{C})$ group transformations corresponding to the realization (2.3) are projective transformations of $\mathbb{C}^1$. Thus, Eq. (2.1) in this case is

$$y(t)=\frac{g_{11}u+g_{21}}{g_{12}u+g_{22}},\quad g_{11}g_{22}-g_{21}g_{12}=1. \tag{2.5}$$

In this case we have $n=1$ (one equation), $r=\dim \mathrm{sl}(2,\mathbb{C})=3$, hence the number of solutions m needed to reconstruct the group element $g(t)=\{g_{ik}(t)\}$ satisfies $m\geqq 3$. In fact we have $m=3$.

Indeed let $y_1(t)$, $y_2(t)$, and $y_3(t)$ be any 3 different solutions of Eq. (2.4), corresponding, e.g., to the choices $u_1=0$, $u_2\to\infty$, $u_3=1$, respectively. Substituting into Eq. (2.5) we express g_{12}, g_{21}, and g_{22} in terms of $y_i(t)$ and g_{11} (which cancels out) and obtain the (well known) nonlinear superposition formula

$$y(t)=\frac{uy_2(y_1-y_3)+y_1(y_3-y_2)}{u(y_1-y_3)+y_3-y_2}. \tag{2.6}$$

Choosing $u=0$ as the origin of the space M, we see that it is stabilized by the maximal parabolic subgroup of SL(2,$\mathbb{C}$) and we obtain the identification

$$M\sim\mathbb{C}\sim G/G_0,\quad G\sim \mathrm{SL}(2,\mathbb{C}),$$
$$G_0\sim\begin{pmatrix} g_{11} & g_{12}\\ 0 & g_{11}^{-1}\end{pmatrix}. \tag{2.7}$$

Now let us turn to a case that has not been considered before, namely $n=2$ when the realization of the algebra sl(2,$\mathbb{C}$) is not primitive. The group SL(2,$\mathbb{C}$) has two inequivalent one dimensional subgroups,

$$G_{0,1}\sim\begin{pmatrix} g_{11} & 0\\ 0 & g_{11}^{-1}\end{pmatrix},\quad G_{0,2}\sim\begin{pmatrix} 1 & g_{12}\\ 0 & 1\end{pmatrix}, \tag{2.8}$$

i.e., the maximal torus $G_{0,1}$ and the unipotent group $G_{0,2}$. Hence we must obtain 2 inequivalent $n=2$ realizations of sl(2,$\mathbb{C}$). In appropriate coordinates (y_1,y_2) the coefficients of ∂_{y_1} in all vector fields will depend on y_1 only, those of y_2 can depend on both y_1 and y_2 [see Eq. (1.7)]. The coefficients of ∂_{y_1} will hence be as in Eq. (2.3). The analysis is easy to perform and we simply present the result

$$X_1=\partial_{y_1},\quad X_2=y_1\partial_{y_1}+y_2\partial_{y_2},\quad X_3=y_1^2\partial_{y_1}+(2y_1y_2+ky_2^2)\partial_{y_2},$$
$$k=0 \ \text{ or } \ 1. \tag{2.9}$$

The two inequivalent realizations correspond to $k=0$ and $k=1$, respectively. Lie in his classification of the $n=2$ case gave these two reaizations in a different, but equivalent form.[3] The form (2.9) was already used by Krause and Michel.[22]

The system of ODEs (1.3) corresponding to (2.9) are

$$\dot y_1=Z_1(t)+Z_2(t)y_1+Z_3(t)y_1^2,$$
$$\dot y_2=Z_2(t)y_2+Z_3(t)(2y_1y_2+ky_2^2). \tag{2.10}$$

We now wish to obtain superposition formulas for the above system, separately for $k=0$ and $k=1$.

Integrating the vector fields (2.9) we obtain the group action

$$y_1=\frac{g_{11}u_1+g_{21}}{g_{12}u_1+g_{22}},\quad y_2=\frac{u_2}{(g_{12}u_1+g_{22})[g_{12}(u_1+ku_2)+g_{22}]}. \tag{2.11}$$

Let us choose the point $(u_1,u_2)=(0,1)$ as the origin. For $k=0$ we see that it is stabilized by the unipotent group $G_{0,2}$ of Eq. (2.8). For $k=1$ the stabilizer of the origin is the subgroup

$$\tilde{G}_{0,1} \sim \begin{pmatrix} g_{11} & g_{11} - \dfrac{1}{g_{11}} \\ 0 & \dfrac{1}{g_{11}} \end{pmatrix}. \tag{2.12}$$

This is conjugate to $G_{0,1}$ of Eq. (2.8), however if we transform $\tilde{G}_{0,1}$ into $G_{0,1}$, the origin is shifted to $(u_1, u_2) = (0, \infty)$. We prefer to stay with the more convenient origin (0,1).

The relation $mn \geqq r$ for Eq. (2.10) is $2m \geqq 3$, hence $m \geqq 2$, and it turns out that two solutions are indeed sufficient to reconstruct the group element $g(t)$.

Let us assume that (v_1, v_2) and (w_1, w_2) are solutions of Eq. (2.10). The four components of **v** and **w** are then not independent, but satisfy the relation

$$\frac{(v_1 - w_1)[v_1 - w_1 + k(v_2 - w_2)]}{v_2 w_2} = R(k), \quad k = 0,1, \tag{2.13}$$

where $R(k)$ is a constant. This relation is the analog of the famous unharmonic ratio of four solutions of the Riccati equation,

$$\frac{(y_1 - y_2)}{(y_1 - y_3)} \frac{(y_3 - y_4)}{(y_2 - y_4)} = K = \text{const.} \tag{2.14}$$

We assume

$$v_1(0) \neq w_1(0), \quad v_2(0) w_2(0) \neq 0, \tag{2.15}$$

and for $k = 1$,

$$R(1) \neq 1.$$

With no loss of generality we can assume that the initial conditions for the two known solutions are

$$\begin{pmatrix} v_1(0) \\ v_2(0) \end{pmatrix} = \begin{pmatrix} 1 \\ a \end{pmatrix}, \quad \begin{pmatrix} w_1(0) \\ w_2(0) \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}, \quad ab \neq 0. \tag{2.16}$$

Substituting these two solutions and their initial conditions into Eq. (2.11), we can solve for $g_{ik}(t)$. We mention that the reconstruction is more efficient in the imprimitive case than in the primitive one. Indeed for the Riccati equation we need to know three solutions, for the system (2.10) only one.

Let us consider the two cases separately.

(1) $k = 0$:

We obtain

$$g_{11} = v_1 \sqrt{\frac{a}{v_2}} - w_1 \sqrt{\frac{b}{w_2}}, \quad g_{12} = \sqrt{\frac{a}{v_2}} - \sqrt{\frac{b}{w_2}}, \quad g_{21} = w_1 \sqrt{\frac{b}{w_2}}, \quad g_{22} = \sqrt{\frac{b}{w_2}}. \tag{2.17}$$

Furthermore, the invariant $R(0)$ of Eq. (2.13) is equal to

$$R(0) = \frac{(v_1 - w_1)^2}{v_2 w_2} = \frac{1}{ab}, \tag{2.18}$$

and hence

$$g_{11}g_{22}-g_{12}g_{21}=\frac{v_1-w_2}{\sqrt{v_2w_2}}\sqrt{ab}=1. \tag{2.19}$$

(2) $k=1$:
We use the invariant (2.13)

$$R(1)=\frac{(v_1-w_1)(v_1-w_1+v_2-w_2)}{v_2w_2}=\frac{a-b+1}{ab} \tag{2.20}$$

to express v_2 in terms of v_1, w_1, and w_2 and obtain

$$\begin{aligned} g_{11}&=\frac{b[w_1(v_1-w_2)-w_1^2]+(b-1)v_1w_2}{[b(1-b)w_2(w_1-v_1)(w_1+w_2-v_1)]^{1/2}},\\ g_{22}&=\frac{b(w_1+w_2-v_1)}{[b(1-b)w_2(w_1-v_1)(w_1+w_2-v_1)]^{1/2}},\\ g_{21}&=g_{22}w_1,\quad g_{12}=\frac{1}{v_1}[g_{11}+g_{22}(v_1-u_1)]. \end{aligned} \tag{2.21}$$

The results of this section can be summed up as follows:

Theorem 1: Two inequivalent realizations of sl(2,$\mathbb{C}$) by vector fields in two dimensions exist, given by Eq. (2.9) with $k=0$ or $k=1$, respectively. The corresponding group actions on the homogeneous space SL(2,$\mathbb{C}$)/$G_{0,k}$ is given in Eq. (2.11).

Theorem 2: The nonlinear ODEs (2.10) for $k=0$ and $k=1$ have superposition formulas given by the imprimitive group action (2.11). The group elements $g_{ik}(t)$, $i,k\in\{1,2\}$ are reconstructed from any two solutions $\mathbf{v}=(v_1,v_2)$ and $\mathbf{w}=(w_1,w_2)$ with the initial conditions satisfying Eq. (2.15). The explicit reconstruction formulas are given in Eq. (2.17) for $k=0$ and (2.21) for $k=1$, respectively.

III. INDUCED REPRESENTATIONS OF SL(N,$\mathbb{C}$) AND PARABOLIC SUBGROUPS

A. General theory

Let us consider the Lie algebra sl(N,$\mathbb{C}$) realized by matrices $X\in\mathbb{C}^{N\times N}$, Tr $X=0$. We shall make use of several subalgebras of sl(N,$\mathbb{C}$). The **Borel** subalgebra is the maximal solvable subalgebra and it can be realized by the set of all traceless upper triangular matrices. A **parabolic** subalgebra of sl(N,$\mathbb{C}$) is any subalgebra containing the Borel subalgebra. A **maximal** parabolic subalgebra is one that is not contained in any proper subalgebra of sl(N,$\mathbb{C}$). All parabolic subalgebras can be realized by block triangular matrices. Maximal parabolic subalgebras correspond to the case of precisely two blocks on the diagonal.

A classification of all maximal parabolic subalgebras of sl(N,$\mathbb{C}$) is obtained by taking all decompositions of N into $N=r_1+r_2$, min(r_1,r_2)$\geqq 1$ and writing all sets of matrices of the form

$$\begin{gathered} p(r_1,r_2)=\left\{X\in\mathbb{C}^{N\times N}\,\middle|\,\mathrm{Tr}\,X=0,\ X=\begin{pmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{pmatrix}\right\},\\ N=r_1+r_2,\\ A_{11}\in\mathbb{C}^{r_1\times r_1},\ A_{22}\in\mathbb{C}^{r_2\times r_2},\ A_{12}\in\mathbb{C}^{r_1\times r_2}. \end{gathered} \tag{3.1}$$

Similarly, a maximal parabolic subgroup of the group corresponds to the same partition of N and satisfies

$$P(r_1,r_2)=\left\{G\in\mathbb{C}^{N\times N}\,\middle|\,G=\begin{pmatrix}G_{11}&G_{12}\\0&G_{22}\end{pmatrix},\ \det G_{11}\det G_{22}=1\right\}. \tag{3.2}$$

The homogeneous space $M\sim\mathrm{SL}(N,\mathbb{C})/P(r_1,r_2)$ was constructed in an earlier article[14] (as a Grassmanian). Local coordinates on this space were introduced as matrix elements of a matrix $W\in\mathbb{C}^{r_1\times r_2}$. The group action on this space is represented by matrix fractional linear transformations

$$\tilde{W}=(G_{11}W+G_{12})(G_{21}W+G_{22})^{-1}. \tag{3.3}$$

The corresponding Lie algebra $\mathrm{sl}(N,\mathbb{C})$ is represented by vector fields with (specific) quadratic nonlinearities. The group action is primitive, the corresponding nonlinear ordinary differential equations with superposition formulas are matrix Riccati equations,

$$\dot{W}=A+BW+WC+WDW,$$

$$W,A\in\mathbb{C}^{r_1\times r_2},\ \ B\in\mathbb{C}^{r_1\times r_1},\ \ C\in\mathbb{C}^{r_2\times r_2},\ \ D\in\mathbb{C}^{r_2\times r_1}. \tag{3.4}$$

In particular for $r_2=1$ we obtain projective Riccati equations corresponding to the projective action of $\mathrm{sl}(N,\mathbb{C})$ on $\mathbb{C}^{N-1}$. The corresponding superposition formula is given by Eq. (3.3) in which $W=\mathrm{const}$ represents the initial data, $\tilde{W}=\tilde{W}(t)$ the general solution and the matrices $G_{ik}(t)$ can be reconstructed from $N+1$ known solutions.[13,14]

Series of nonprimitive realizations of the Lie algebra $\mathrm{sl}(N,\mathbb{C})$ for $N\geqq 3$ can be obtained using the theory of induced representations.[23,24] The homogeneous spaces that we construct are $M\sim G/G_0$, where G is $\mathrm{SL}(N,\mathbb{C})$ and G_0 is a (nonmaximal) parabolic subgroup of $\mathrm{SL}(N,\mathbb{C})$. More specifically, in this article $P(N)$ is group of matrices

$$G_0\sim g=\begin{pmatrix}g_{11}&g_{12}&\cdots&\cdots&g_{1N}\\ \cdots&\cdots&\cdots&\cdots&\cdots\\ g_{N-2,1}&\cdots&g_{N-2,N-2}&g_{N-2,N-1}&g_{N-2,N}\\ 0&\cdots&0&g_{N-1,N-1}&g_{N-1,N}\\ 0&\cdots&0&0&g_{NN}\end{pmatrix}. \tag{3.5}$$

We use the Borel subgroup $B\subset\mathrm{SL}(N,\mathbb{C})$ to induce representations of $\mathrm{SL}(N,\mathbb{C})$ on spaces of functions $f(Z)$, where Z is a point on the space $\mathrm{SL}(N,\mathbb{C})/B$.

To obtain the group action explicitly, we use the defining representation of $\mathrm{SL}(N,\mathbb{C})$ by $N\times N$ matrices and write the Gauss decomposition

$$g=\xi DZ,\quad \xi=\begin{pmatrix}1&\xi_{12}&\xi_{13}&\cdots&\xi_{1N}\\0&1&\xi_{23}&\cdots&\xi_{2N}\\0&0&\ddots&\cdots&\cdots\\0&\cdots&0&1&\xi_{N-1,N}\\0&\cdots&&0&1\end{pmatrix},\quad Z=\begin{pmatrix}1&0&&\cdots&0\\z_{21}&1&0&\cdots&0\\z_{31}&z_{32}&\ddots&\cdots&\cdots\\&\cdots&&1&0\\z_{N1}&\cdots&&z_{N,N-1}&1\end{pmatrix},$$

$$D=\mathrm{diag}(d_{11},\dots,d_{NN}). \tag{3.6}$$

The matrix elements z_{ij}, $1\leqq j<i\leqq N$ are local coordinates on $\tilde{M}\sim\mathrm{SL}(N,\mathbb{C})/B$ (the Borel subgroup is represented by the matrices ξD). The action of the group $\mathrm{SL}(N,\mathbb{C})$ on $\tilde{M}$ is given in local coordinates as

$$Zg=\xi D\tilde{Z}. \tag{3.7}$$

Explicitly we obtain the group action

$$\tilde{Z}=F(g,Z) \tag{3.8}$$

by eliminating d_{ii} and ξ_{ik} from Eq. (3.7) using a subset of these N^2 equations and substituting into the remaining $[N(N-1)]/2$ equations.

B. Example of the group SL(3,$\mathbb{C}$)

Let us illustrate the procedure for the simplest nontrivial case, namely SL(3,$\mathbb{C}$). Equation (3.7) in this case is

$$\begin{pmatrix} 1 & 0 & 0 \\ z_{21} & 1 & 0 \\ z_{31} & z_{32} & 1 \end{pmatrix}\begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{pmatrix}=\begin{pmatrix} 1 & \xi_{12} & \xi_{13} \\ 0 & 1 & \xi_{23} \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ \tilde{z}_{21} & 1 & 0 \\ \tilde{z}_{31} & \tilde{z}_{32} & 1 \end{pmatrix}. \tag{3.9}$$

We rewrite Eq. (3.9) symbolically as

$$R_{ik}=0.$$

Equation $R_{i3}=0$ gives us d_{33}, ξ_{23}, and ξ_{13} in terms of z_{ik} and g_{ik}. Equation $R_{3k}=0$ for $k=1$ and 2 give us $\tilde{z}_{32}$ and $\tilde{z}_{31}$ in terms of z_{ik} and g_{ik}, eq. $R_{22}=0$ and $R_{21}=0$ give us d_{22} and finally $\tilde{z}_{21}$.

Defining $z_{31}=x_1$, $z_{32}=x_2$, and $z_{21}=x_3$ and similarly for $\tilde{z}_{31}$, $\tilde{z}_{32}$, and $\tilde{z}_{21}$ we obtain the group transformations, namely,

$$\begin{aligned} \tilde{x}_1&=\frac{g_{11}x_1+g_{21}x_2+g_{31}}{g_{13}x_1+g_{23}x_2+g_{33}}, \\ \tilde{x}_2&=\frac{g_{12}x_1+g_{22}x_2+g_{32}}{g_{13}x_1+g_{23}x_2+g_{33}}, \\ \tilde{x}_3&=\frac{(-x_1+x_2x_3)A_{1123}+x_3A_{1133}+A_{2133}}{(-x_1+x_2x_3)A_{1223}+x_3A_{1233}+A_{2233}}, \end{aligned} \tag{3.10}$$

where

$$A_{ijkl}=\begin{vmatrix} g_{ij} & g_{il} \\ g_{kj} & g_{kl} \end{vmatrix}=g_{ij}g_{kl}-g_{il}g_{kj}. \tag{3.11}$$

We see that the elements of the last row in Z, namely, $z_{31}=x_1$, $z_{32}=x_2$ transform independently of the second row ($z_{21}=x_3$). Thus, we have a realization of SL(3,$\mathbb{C}$) on a flag manifold. Indeed $\{x_1,x_2\}$ are local coordinates on the space SL(3,$\mathbb{C}$)/$P(2,1)$ and we have a primitive action on this subspace (the projective action). Further $\{x_1,x_2,x_3\}$ are local coordinates on SL(3,$\mathbb{C}$)/B and on this space the action is imprimitive (we have $P(3)\sim B$).

The linear representations of SL(3,$\mathbb{C}$) on functions $f(Z)$ is given as

$$T_g f(x_1,x_2,x_3)=f(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3), \tag{3.12}$$

with $\tilde{x}_i$ as in Eq. (3.10). Calculating the infinitesimal operators in this representation we find

$$\hat{E}_{31}=\partial_1,\quad \hat{E}_{32}=\partial_2,\quad \hat{E}_{21}=x_2\partial_1+\partial_3,$$

$$\hat{E}_{11}=x_1\partial_1+x_3\partial_3,\quad \hat{E}_{22}=x_2\partial_2-x_3\partial_3,\quad \hat{E}_{12}=x_1\partial_2-x_3^2\partial_3,$$

$$\hat{E}_{13}=x_1(x_1\partial_1+x_2\partial_2)-x_3(x_3x_2-x_1)\partial_3, \tag{3.13}$$

$$\hat{E}_{23}=x_2(x_1\partial_1+x_2\partial_2)-(x_3x_2-x_1)\partial_3,$$

$$\partial_i\equiv\partial_{x_i}.$$

The nonlinear ordinary differential equations with a superposition formula corresponding to the action of SL(3,ℂ) on SL(3,ℂ)/B can be read off from (3.13) using Lie's result (1.3) and (1.4). They are

$$\dot{x}_1=Z_{31}+Z_{11}x_1+Z_{21}x_2+Z_{13}x_1^2+Z_{23}x_1x_2,$$

$$\dot{x}_2=Z_{32}+Z_{12}x_1+Z_{22}x_2+Z_{13}x_1x_2+Z_{23}x_2^2, \tag{3.14}$$

$$\dot{x}_3=Z_{21}+(Z_{11}-Z_{22})x_3-Z_{12}x_3^2-Z_{13}x_3(x_3x_2-x_1)-Z_{23}(x_3x_2-x_1),$$

where $Z_{ik}=Z_{ik}(t)$ are arbitrary functions of time t.

We mention that the origin of the coordinate system in SL(3,ℂ)/$P(2,1)$ is the point $(x_1,x_2)=(0,0)$ and in SL(3,ℂ)/$P(3)$ it is $(x_1,x_2,x_3)=(0,0,0)$.

Notice that the first two equations in Eq. (3.14) involve quadratic nonlinearities, while the third equation involves cubic ones. However, if x_1 and x_2 are known (from the first two equations), then the equation for $\dot{x}_3$ reduces to a Riccati equation.

C. The group SL(4,ℂ)

For SL(4,ℂ) we no longer have $P(4)\sim B$ and the flag manifold consists of three spaces

$$\mathrm{SL}(4,\mathbb{C})/P(3,1)\subset\mathrm{SL}(4,\mathbb{C})/P(4)\subset\mathrm{SL}(4,\mathbb{C})/B. \tag{3.15}$$

We restrict ourselves to the space SL(4,ℂ)/$P(4)$, put $x_1=z_{41}$, $x_2=z_{42}$, $x_3=z_{43}$, $x_4=z_{31}$, and $x_5=z_{32}$ and use Eq. (3.7) to obtain

$$\tilde{x}_1=\frac{x_1g_{11}+x_2g_{21}+x_3g_{31}+g_{41}}{x_1g_{14}+x_2g_{24}+x_3g_{34}+g_{44}},$$

$$\tilde{x}_2=\frac{x_1g_{12}+x_2g_{22}+x_3g_{32}+g_{42}}{x_1g_{14}+x_2g_{24}+x_3g_{34}+g_{44}},$$

$$\tilde{x}_3=\frac{x_1g_{13}+x_2g_{23}+x_3g_{33}+g_{43}}{x_1g_{14}+x_2g_{24}+x_3g_{34}+g_{44}}, \tag{3.16}$$

$$\tilde{x}_4=\frac{(x_1x_5-x_2x_4)A_{1124}+(x_1-x_3x_4)A_{1134}+(x_2-x_3x_5)A_{2134}-x_4A_{1144}-x_5A_{2144}-A_{3144}}{(x_1x_5-x_2x_4)A_{1324}+(x_1-x_3x_4)A_{1334}+(x_2-x_3x_5)A_{2334}-x_4A_{1344}-x_5A_{2344}-A_{3344}},$$

$$\tilde{x}_5=\frac{(x_1x_5-x_2x_4)A_{1224}+(x_1-x_3x_4)A_{1234}+(x_2-x_3x_5)A_{2234}-x_4A_{1244}-x_5A_{2244}-A_{3244}}{(x_1x_5-x_2x_4)A_{1324}+(x_1-x_3x_4)A_{1334}+(x_2-x_3x_5)A_{2334}-x_4A_{1344}-x_5A_{2344}-A_{3344}}.$$

The infinitesimal operators can again be obtained as vector fields, using the representation $T_gf(Z)=f(\tilde{Z})$ with Z and $\tilde{Z}$ related as in Eq. (3.16). Instead of writing them out we give the corresponding nonlinear ordinary differential equations, namely,

$$\dot{x}_1=Z_{41}+Z_{11}x_1+Z_{21}x_2+Z_{31}x_3+x_1(Z_{14}x_1+Z_{24}x_2+Z_{34}x_3),$$

$$\dot{x}_2=Z_{42}+Z_{12}x_1+Z_{22}x_2+Z_{32}x_3+x_2(Z_{14}x_1+Z_{24}x_2+Z_{34}x_3),$$

$$\dot{x}_3 = Z_{43} + Z_{13}x_1 + Z_{23}x_2 + Z_{33}x_3 + x_3(Z_{14}x_1 + Z_{24}x_2 + Z_{34}x_3), \tag{3.17}$$

$$\dot{x}_4 = Z_{31} + (Z_{11} - Z_{33})x_4 + Z_{21}x_5 - x_4(Z_{13}x_4 + Z_{23}x_5) + (x_1 - x_3x_4)(Z_{14}x_4 + Z_{24}x_5 + Z_{34}),$$

$$\dot{x}_5 = Z_{32} + Z_{12}x_4 + (Z_{22} - Z_{33})x_5 - x_5(Z_{13}x_4 + Z_{23}x_5) + (x_2 - x_3x_5)(Z_{14}x_4 + Z_{24}x_5 + Z_{34}),$$

where Z_{ik} are arbitrary functions of t. Notice that $\{x_1, x_2, x_3\}$ satisfy projective Riccati equations. The last two equations involve cubic nonlinearities. However, similarly as in the $\mathrm{sl}(3,\mathbb{C})$ case the equations "decompose." If $\{x_1, x_2, x_3\}$ are known, then the equations for $\{x_4, x_5\}$ satisfy projective Riccati equations, based on algebra $\mathrm{sl}(3,\mathbb{C})$.

D. Finite transformations and vector fields for general *N*

For general $N \geqq 3$ the formulas are quite similar to the above cases $N=3,4$ only somewhat more cumbersome to write. Dropping the calculations, we just present $\mathrm{SL}(N,\mathbb{C})$ group action on the space $\mathrm{SL}(N,\mathbb{C})/P(N)$,

$$\tilde{x}_i = \frac{(\sum_{j=1}^{N-1} g_{ji}x_j) + g_{Ni}}{(\sum_{j=1}^{N-1} g_{jN}x_j) + g_{NN}}, \quad i = 1,\ldots,N-1,$$
$$\tilde{x}_{i+N-1} = F(i)/F(N-1), \quad i = 1,\ldots,N-2, \tag{3.18}$$

where

$$F(i) = \left(\sum_{k=1}^{N-3} \sum_{l=1}^{N-2-k} (x_k x_{l+N+k-1} - x_{k+l}x_{k+N-1}) A_{ki(k+l)N} \right) + \left(\sum_{k=1}^{N-2} (x_k - x_{N-1}x_{k+N-1}) A_{ki(N-1)N} \right) + \left(\sum_{k=1}^{N-2} (-x_{k+N-1}) A_{kiNN} \right) - A_{(N-1)iNN}.$$

We see that $\{x_1,\ldots,x_{N-1}\}$ transform according to the projective realization of $\mathrm{SL}(N,\mathbb{C})$. The remaining $N-2$ variables $\{x_N,\ldots,x_{2N-3}\}$ transform in a manner involving quadratic polynomials in the denominator and numerator of a fraction.

The nonlinear ODEs with superposition formulas can be read off from the vector fields representing the Lie algebra $\mathrm{sl}(N,\mathbb{C})$ in this realization, namely,

$$\hat{E}_{Nj} = \partial_j, \quad 1 \leqq j \leqq N-1,$$

$$\hat{E}_{ij} = x_i\partial_j + x_{N-1+i}\partial_{N-1+j}, \quad 1 \leqq i \leqq N-2, \quad 1 \leqq j \leqq N-2,$$

$$\hat{E}_{N-1j} = x_{N-1}\partial_j + \partial_{N+j-1}, \quad 1 \leqq j \leqq N-2,$$

$$\hat{E}_{N-1N-1} = x_{N-1}\partial_{N-1} - \sum_{m=N}^{2N-3} x_m\partial_m, \tag{3.19}$$

$$\hat{E}_{iN-1} = x_i\partial_{N-1} - x_{N-1+i}\sum_{m=N}^{2N-3} x_m\partial_m, \quad 1 \leqq i \leqq N,$$

$$\hat{E}_{iN} = x_i\sum_{j=1}^{N-1} x_j\partial_j - x_{N-1+i}\sum_{m=N}^{2N-3} (x_m x_{N-1} - x_{m-N+1})\partial_m, \quad 1 \leqq i \leqq N-2,$$

$$\hat{E}_{N-1N}=x_{N-1}\sum_{j=1}^{N-1} x_j\partial_j-\sum_{m=N}^{2N-3} (x_m x_{N-1}-x_{m-N+1})\partial_m .$$

The sl(N,ℂ) equations can now be written in a quite compact form, namely,

$$\dot{x}_j=Z_{Nj}+\sum_{i=1}^{N-1} Z_{ij}x_i+x_j\sum_{i=1}^{N-1} Z_{iN}x_i , \quad 1\leqq j\leqq N-1, \tag{3.20}$$

$$\dot{x}_{N-1+j}=Z_{N-1j}+\sum_{i=1}^{N-2} Z_{ij}x_{N-1+i}-Z_{N-1N-1}x_{N-1+j}-x_{N-1+j}\sum_{i=1}^{N-2} Z_{iN-1}x_{N-1+i}$$
$$+(x_j-x_{N-1+j}x_{N-1})\sum_{i=1}^{N-2} (Z_{iN}x_{N-1+i}+Z_{N-1N}), \quad 1\leqq j\leqq N-2. \tag{3.21}$$

In (3.20) and (3.21) Z_{ik} are arbitrary functions of t. For $N=3$ and $N=4$ these formulas coincide with (3.14) and (3.17), respectively. The same comments pertain, namely, Eq. (3.20) are projective Riccati equations based on sl(N,ℂ). Equation (3.21) are projective Riccati equations based on sl($N-1$,ℂ) if $x_1,\ldots,x_{N-1}$ are known.

IV. SUPERPOSITION FORMULAS

A. General comments

The superposition formula for the imprimitive SL(N,ℂ) equations is given by the group action formula (3.18) in which x_i, $i=1,\ldots,2N-3$ is a constant vector, related to the initial conditions. The matrix elements $g_{ik}(t)$ must be expressed in terms of m particular solutions. The number m satisfies Eq. (1.6). In our case that means that we must have

$$m(2N-3)\geqq N^2-1. \tag{4.1}$$

The actual number m, as well as conditions that must be imposed on the known solutions, is obtained from the requirement that m be the smallest number of solutions such that their joint isotropy group consists only of the identity transformation. We shall call such set a ''fundamental set of solutions.''

B. Example of SL(3,ℂ)

Equation (4.1) in this case is $3m\geqq 8$ and hence $m\geqq 3$.

The equations under consideration are given in Eq. (3.14), the superposition formula has the form (3.10).

1. The fundamental set of solutions

Let us assume that we know three solutions of Eq. (3.14), $\mathbf{x}(t)$, $\mathbf{y}(t)$, and $\mathbf{z}(t)$. Each solution is a three component vector. The group SL(3,ℂ) acts on the space $M\times M\times M$ of these three solutions. Since we have dim SL(3,ℂ)$=8$ the group can sweep out at most an eight-dimensional orbit, so there must exist at least one SL(3,ℂ) invariant in this nine-dimensional space. We denote this invariant

$$R(x_1,x_2,x_3,y_1,y_2,y_3,z_1,z_2,z_3)=K, \tag{4.2}$$

where **x**, **y**, and **z** are three solutions of Eq. (3.14). We calculate $\dot{R}=(\mathrm{d}R/\mathrm{d}t)$, replace $\dot{\mathbf{x}},\dot{\mathbf{y}},\dot{\mathbf{z}}$ using Eq. (3.14) and require $\dot{R}=0$ for all functions $Z_{ik}(t)$. This provides us with eight linear first order partial differential equations for R. These equations can be solved and we obtain a single elementary invariant, namely,

$$R=\frac{(XZ)_{321}(ZY)_{321}(YX)_{321}}{(XY)_{321}(YZ)_{321}(ZX)_{321}}, \tag{4.3}$$

where we have defined

$$(XY)_{321}=x_3(x_2-y_2)-x_1+y_1, \tag{4.4}$$

etc. The quantity $(XY)_{321}$ and similarly defined $(XZ)_{321}$, $(YX)_{321}$, $(YZ)_{321}$, $(ZX)_{321}$, and $(ZY)_{321}$ play a crucial role in the reconstruction of the group elements $g_{ik}(t)$. For SL$(3,\mathbb{C})$ the notation is somewhat redundant. However for SL$(N,\mathbb{C})$, $N\geqq 4$ we will have quantities like $(XY)_{431}, (XY)_{532}$, etc. so for uniformity we keep the subscripts for $N=3$ as well.

The initial condition for three solutions are $\mathbf{x}(0)$, $\mathbf{y}(0)$, and $\mathbf{z}(0)$. We shall use the transformation (3.10) (with constant coefficients) to standardize the initial conditions. Let us do this transformation in two steps,

$$g=g_2g_1,\quad g_1=\begin{pmatrix} g_{11} & g_{12} & 0\\ g_{21} & g_{22} & 0\\ g_{31} & g_{32} & g_{33}\end{pmatrix},\quad g_2=\begin{pmatrix} a_{11} & 0 & a_{11}-a_{33}\\ 0 & a_{22} & a_{22}-a_{33}\\ 0 & 0 & a_{33}\end{pmatrix}. \tag{4.5}$$

Let us now assume that the first two components of $\mathbf{x}(0)$, $\mathbf{y}(0)$, and $\mathbf{z}(0)$ satisfy

$$\Delta=\begin{vmatrix} x_1(0)-z_1(0) & y_1(0)-z_1(0)\\ x_2(0)-z_2(0) & y_2(0)-z_2(0)\end{vmatrix}\neq 0. \tag{4.6}$$

This condition can be rewritten as

$$\Delta=x_1(0)(y_2(0)-z_2(0))+y_1(0)(z_2(0)-x_2(0))+z_1(0)(x_2(0)-y_2(0))\neq 0,$$

so that it is actually symmetric in the three two-dimensional vectors,

$$x_{\mathrm{T}}(0)=\begin{pmatrix} x_1(0)\\ x_2(0)\end{pmatrix},\quad y_{\mathrm{T}}(0)=\begin{pmatrix} y_1(0)\\ y_2(0)\end{pmatrix},\quad z_{\mathrm{T}}(0)=\begin{pmatrix} z_1(0)\\ z_2(0)\end{pmatrix}. \tag{4.7}$$

If Eq. (4.6) is satisfied we can use the transformation (3.10) with the matrix g_1 to transform the three initial vectors into a more convenient form

$$g_1:\{\mathbf{x}(0),\mathbf{y}(0),\mathbf{z}(0)\}\to\left\{\begin{pmatrix} 1\\ 0\\ \tilde{x}_3\end{pmatrix},\begin{pmatrix} 0\\ 1\\ \tilde{y}_3\end{pmatrix},\begin{pmatrix} 0\\ 0\\ \tilde{z}_3\end{pmatrix}\right\} \tag{4.8}$$

with

$$\tilde{x}_3=\frac{(XY)_{321}-(XZ)_{321}}{(XZ)_{321}},\quad \tilde{y}_3=-\frac{(YZ)_{321}}{(YZ)_{321}-(YX)_{321}},\quad \tilde{z}_3=-\frac{(ZY)_{321}}{(ZX)_{321}}$$

(all components evaluated at $t=0$).

The initial conditions (4.8) imply that $(XY)_{321}$ and $(XZ)_{321}$ cannot vanish simultaneously (and similarly for $(YZ)_{321}$ and $(YX)_{321}$, or $(ZX)_{321}$ and $(ZY)_{321}$). To proceed further we need a stronger condition, namely, that at least one of the following relations holds:

$$(XY)_{321}(XZ)_{321}\neq 0,\quad (YZ)_{321}(YX)_{321}\neq 0,\quad (ZX)_{321}(ZY)_{321}\neq 0. \tag{4.9}$$

With no loss of generality we assume that the first of the above relations holds. We then transform further using the matrix g_2 and obtain

$$g_2 g_1 : \{\mathbf{x}(0),\mathbf{y}(0),\mathbf{z}(0)\} \rightarrow \{\mathbf{x}_S(0),\mathbf{y}_S(0),\mathbf{z}_S(0)\},$$

$$\mathbf{x}_S(0)=\begin{pmatrix}1\\0\\0\end{pmatrix},\quad \mathbf{y}_S(0)=\begin{pmatrix}0\\1\\\alpha\end{pmatrix},\quad \mathbf{z}_S(0)=\begin{pmatrix}0\\0\\\beta\end{pmatrix}, \tag{4.10}$$

where

$$\alpha=\left[-1+\frac{a_{33}}{a_{11}}\frac{(YX)_{321}}{(YZ)_{321}}\right]^{-1},\quad \beta=-\frac{a_{11}}{a_{22}}\frac{(ZY)_{321}}{(ZX)_{321}}. \tag{4.11}$$

To standardize $\mathbf{x}_S(0)$ we have already fixed the ratio a_{33}/a_{22} hence only one of the ratios a_{33}/a_{11} and a_{11}/a_{22} can be chosen freely. From Eq. (4.11) we immediately see four special cases,

$$\begin{aligned}(YX)_{321}&=0\Rightarrow\alpha=-1,\\ (YZ)_{321}&=0\Rightarrow\alpha=0,\\ (ZY)_{321}&=0\Rightarrow\beta=0,\\ (ZX)_{321}&=0\Rightarrow\beta\rightarrow\infty.\end{aligned} \tag{4.12}$$

Thus, if $(YX)_{321}(YZ)_{321}\neq 0$ we can standardize $\alpha\rightarrow 1$. Alternatively, if $(ZY)_{321}(ZX)_{321}\neq 0$ we can standardize $\beta\rightarrow 1$. The invariant R of Eq. (4.3) for the standardized initial conditions (4.10) is

$$R=-\frac{(\alpha+1)}{\alpha}\beta. \tag{4.13}$$

We recall that R is time independent and cannot be changed by group transformation of the initial conditions.

To see whether a reconstruction is possible from three solutions with initial conditions (4.10), let us calculate the stabilizer of these ''standard'' initial conditions. The stabilizer of the vector $\mathbf{x}_S(0)$ and of the first two components of the other two vectors is

$$g=\begin{pmatrix}a_{11} & 0 & a_{11}-a_{22}\\ 0 & a_{22} & 0\\ 0 & 0 & a_{22}\end{pmatrix}. \tag{4.14}$$

The remaining conditions for $\mathbf{y}_S(0)$ and $\mathbf{z}_S(0)$ to be stabilized are

$$\alpha(\alpha+1)(g_{11}-g_{22})=0,\quad \beta g_{11}=\beta g_{22}. \tag{4.15}$$

Thus, the stabilizer is $g\sim I$ if we have at least one of the following conditions:

$$\alpha(\alpha+1)\neq 0,\quad 0<|\beta|<\infty. \tag{4.16}$$

All other cases must be excluded.

Thus, if $\Delta=0$, the stabilizer is always too large. The same is true if all the products in Eq. (4.9) vanish (at $t=0$).

Thus, a reconstruction is possible if and only if

(1) $\Delta\neq 0$,
(2) at least 2 of the products in Eq. (4.9) are nonzero.

2. Reconstruction of the group element

Let us reconstruct the SL(3,ℂ) group element from the fundamental set of solutions, corresponding to the initial data (4.10) satisfying one of the conditions (4.16). We first make use of the first two components of the three solutions $\mathbf{x}(t),\mathbf{y}(t),\mathbf{z}(t)$. The first two equations in (3.10) then imply

$$x_1=\frac{g_{11}+g_{31}}{g_{13}+g_{33}},\quad y_1=\frac{g_{21}+g_{31}}{g_{23}+g_{33}},\quad z_1=\frac{g_{31}}{g_{33}},$$
$$x_2=\frac{g_{12}+g_{32}}{g_{13}+g_{33}},\quad y_2=\frac{g_{22}+g_{32}}{g_{23}+g_{33}},\quad z_2=\frac{g_{32}}{g_{33}}. \tag{4.17}$$

This allows us to express all off-diagonal elements $g_{ik}(t),\ i\neq k$ in terms of the diagonal ones and the first two components of the three known solutions,

$$g_{12}=\frac{1}{x_1}[g_{11}x_2+g_{33}(z_1x_2-x_1z_2)],\quad g_{13}=\frac{1}{x_1}[g_{11}+g_{33}(z_1-x_1)],$$
$$g_{21}=\frac{1}{y_2}[g_{22}y_1+g_{33}(y_1z_2-z_1y_2)],\quad g_{23}=\frac{1}{y_2}[g_{22}+g_{33}(z_2-y_2)], \tag{4.18}$$
$$g_{31}=g_{33}z_1,\quad g_{32}=g_{33}z_2.$$

Next, we substitute the known solutions and initial conditions (4.10) into the third equation in (3.10). We replace the off-diagonal elements using Eq. (4.18) and obtain three linear equations for the diagonal elements g_{11}, g_{22}, and g_{33},

$$\begin{pmatrix} 0 & -(XY)_{321} & y_2(XZ)_{321}-z_2(XY)_{321} \\ -\alpha(YX)_{321} & 0 & (1+\alpha)x_1(YZ)_{321}-\alpha z_1(YX)_{321} \\ \beta y_2(ZX)_{321} & x_1(ZY)_{321} & \beta z_1y_2(ZX)_{321}+x_1z_2(ZY)_{321}\end{pmatrix}\begin{pmatrix} g_{11}\\ g_{22}\\ g_{33}\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}. \tag{4.19}$$

The rank r_M of the above matrix is $r_M\leqq 2$. A reconstruction is possible if $r_M=2$; then we can express g_{11} and g_{22} linearly in terms of g_{33}. The rank is $r_M=1$ if and only if both conditions (4.16) are violated (i.e., $\alpha=0$ or $\alpha=-1$ and simultaneously $\beta=0$ or $\beta\to\infty$).

The above results are particularly obvious at $t=0$ when Eq. (4.19) reduce to

$$\begin{pmatrix} 0 & 1 & -1 \\ -\alpha(\alpha+1) & 0 & \alpha(\alpha+1) \\ \beta & -\beta & 0\end{pmatrix}\begin{pmatrix} g_{11}\\ g_{22}\\ g_{33}\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}. \tag{4.20}$$

Let us now sum up the results of this section as follows.

Theorem 3: (1) The general solution of the system of nonlinear ordinary differential equations (3.14) associated with the imprimitive action of SL(3,ℂ) on the space SL(3,ℂ)/P(3) is given by the formula

$$v_1(t)=\frac{g_{11}u_1+g_{21}u_2+g_{31}}{g_{13}u_1+g_{23}u_2+g_{33}},\qquad v_2(t)=\frac{g_{12}u_1+g_{22}u_2+g_{32}}{g_{13}u_1+g_{23}u_2+g_{33}},$$
$$v_3(t)=\frac{(u_2u_3-u_1)A_{11,23}+u_3A_{11,33}+A_{21,33}}{(u_2u_3-u_1)A_{12,23}+u_3A_{12,33}+A_{22,33}}, \tag{4.21}$$

$$A_{ij,kl}(t)=\begin{vmatrix} g_{ij} & g_{il} \\ g_{kj} & g_{kl} \end{vmatrix}=g_{ij}g_{kl}-g_{il}g_{kj}\,.$$

The three constants u_1, u_2, and u_3 are related to the initial conditions for the solution $\mathbf{v}(t)$.

(2) The group elements $g_{ik}(t)$ (and hence also the quantities A_{ijkl}) are reconstructed from a set of three solutions $\mathbf{x}(t)$, $\mathbf{y}(t)$, and $\mathbf{z}(t)$. This fundamental set of solutions is quite generic and is subject to just two conditions:

(i) $\Delta \neq 0$ with Δ defined in Eq. (4.6),
(ii) At lest two of the inequalities (4.9) hold with $(XY)_{321}$ defined in Eq. (4.4).

(3) The off-diagonal elements $g_{ik}(t)$, $i\neq k$ are given in Eq. (4.18). The diagonal ones are obtained by solving Eq. (4.19).

Comments: (1) The reconstruction is linear in that we only need to solve linear algebraic relations (once three solutions are known). All elements g_{ik} are proportional to $g_{33}(t)$ which cancels out from Eq. (4.21).

(2) A fundamental set of solutions for projective Riccati equations based on the primitive action of SL(3,$\mathbb{C}$) consists of four solutions. In the imprimitive case we only need three solutions.

C. The group SL(4,$\mathbb{C}$)

The sl(4,$\mathbb{C}$) ODEs are given in Eq. (3.17), the general form of the superposition formula in Eq. (3.16). The number of equations is $n=5$, the dimension of the Lie algebra is $r=15$, hence $nm \geqq r$ implies $m\geqq 3$.

We shall show that we do actually need precisely $m=3$ generically chosen particular solutions and give an explicit reconstruction of the group element. For the group SL(4,$\mathbb{C}$) we have $nm=r$ (for $m=3$) and no SL(4,$\mathbb{C}$) invariant can be formed out of three solutions.

Let us assume that we know three solutions **x**, **y**, **z**, each of them a five-component vector.

Let us assume that the first three components of these vectors satisfy an independence condition for $t=0$, namely,

$$\operatorname{rank}\begin{pmatrix} x_1(0)-z_1(0) & y_1(0)-z_1(0) \\ x_2(0)-z_2(0) & y_2(0)-z_2(0) \\ x_3(0)-z_3(0) & y_3(0)-z_3(0) \end{pmatrix}=2. \tag{4.22}$$

We can then use a constant coefficient SL(4,$\mathbb{C}$) transformation to take the initial conditions into

$$\mathbf{x}(0)=\begin{pmatrix} 1\\0\\0\\x_4(0)\\x_5(0)\end{pmatrix},\quad \mathbf{y}(0)=\begin{pmatrix} 0\\1\\0\\y_4(0)\\y_5(0)\end{pmatrix},\quad \mathbf{z}(0)=\begin{pmatrix}0\\0\\0\\0\\0\end{pmatrix}. \tag{4.23}$$

Let us make a further assumption, namely, that transformed initial conditions satisfy

$$x_5(0)y_4(0)\neq 0,\quad y_4(0)+y_5(0)\neq x_4(0)+x_5(0). \tag{4.24}$$

We can then standardize the initial conditions further, namely take them into

$$\mathbf{x}(0)=\begin{pmatrix} 1\\0\\0\\0\\1\end{pmatrix},\quad \mathbf{y}(0)=\begin{pmatrix} 0\\1\\0\\\alpha\\0\end{pmatrix},\quad \mathbf{z}(0)=\begin{pmatrix}0\\0\\0\\0\\0\end{pmatrix}. \tag{4.25}$$

The stabilizer of these three vectors in SL(4,ℂ) consists of matrices satisfying

$$g(0)=\begin{pmatrix} g_{11} & 0 & 0 & g_{11}-g_{44} \\ 0 & g_{22} & 0 & g_{22}-g_{44} \\ 0 & 0 & g_{22} & g_{44}-g_{22} \\ 0 & 0 & 0 & g_{44} \end{pmatrix}, \quad \alpha(g_{11}-g_{22})=0, \quad (\alpha-1)(g_{11}-g_{44})=0. \tag{4.26}$$

Thus, for $\alpha\neq 0$, $\alpha\neq 1$ we have $g\sim I$ and a unique reconstruction of the group elements $g_{ik}(t)$ is possible. We mention that even though α has no invariant meaning, the values $\alpha=0$ and $\alpha=1$ correspond to degenerate orbits of triplets of vectors. These values of α must be excluded from further considerations. Indeed, the stabilizer (4.26) for $\alpha=0$ or $\alpha=1$ is larger, since g_{11} and either g_{22} or g_{44} remain free.

Let us now perform a reconstruction, using 3 solutions satisfying Eq. (4.25) with $\alpha(1-\alpha)\neq 0$. Substituting the components x_i, y_i, and z_i, $i=1,2,3$ into Eq. (3.16) we obtain

$$\begin{gathered} g_{41}=g_{44}z_1, \quad g_{42}=g_{44}z_2, \quad g_{43}=g_{44}z_3, \\ g_{12}=\frac{1}{x_1}[g_{11}x_2+g_{44}(x_2z_1-x_1z_2)], \quad g_{13}=\frac{1}{x_1}[g_{11}x_3+g_{44}(x_3z_1-x_1z_3)], \\ g_{14}=\frac{1}{x_1}[g_{11}+g_{44}(z_1-x_1)], \quad g_{21}=\frac{1}{y_2}[g_{22}y_1+g_{44}(y_1z_2-y_2z_1)], \\ g_{23}=\frac{1}{y_2}[g_{22}y_3+g_{44}(y_3z_2-y_2z_3)], \quad g_{24}=\frac{1}{y_2}[g_{22}+g_{44}(z_2-y_2)]. \end{gathered} \tag{4.27}$$

Using the relations (3.16) w_4, w_5, and x_4, we obtain the remaining off-diagonal elements in terms of the diagonal ones (and known solutions),

$$\begin{gathered} g_{31}+(z_3z_4-z_1)g_{34}=g_{33}w_4, \\ g_{32}+(z_3z_5-z_2)g_{34}=g_{33}w_5, \\ g_{34}=\frac{-(a_{22}+z_2a_{44})(XY)_{431}+a_{33}y_2(x_4-z_4)+a_{44}y_2(XZ)_{431}}{y_2[x_3x_4-x_1-z_3z_4+z_1]}. \end{gathered} \tag{4.28}$$

The expressions $(XY)_{431}$ and $(XZ)_{431}$ are defined as in Eq. (4.4).

Finally using the expressions for $x_5(t)$, $y_4(t)$, and $y_5(t)$ we obtain three linear relations for the diagonal elements g_{ii}, making it possible to express g_{11}, g_{22}, and g_{33} linearly in terms of g_{44},

$$\begin{aligned} & a_{22}\{(x_3x_5-x_2-z_3z_5+z_2)(XY)_{431}-(x_3x_4-x_1-z_3z_4+z_1)(XY)_{532}\} \\ & \quad +a_{33}y_2\{(x_3x_5-x_2-z_3z_5+z_2)(-x_4+z_4)+(x_3x_4-x_1-z_3z_4+z_1)(x_5-z_5)\} \\ & \quad +a_{44}\{(x_3x_5-x_2-z_3z_5+z_2)[-y_2(XZ)_{431}+z_2(XY)_{431}] \\ & \quad +(x_3x_4-x_1-z_3z_4+z_1)[y_2(XZ)_{532}-z_2(XY)_{532}]\}=0, \end{aligned} \tag{4.29}$$

$$\begin{aligned} & -a_{11}\alpha y_2(x_3x_4-x_1-z_3z_4+z_1)(YX)_{431}+a_{22}x_1(y_3y_4-y_1-z_3z_4+z_1)(XY)_{431} \\ & \quad +a_{33}x_1y_2[(y_4-z_4)(x_3x_4-x_1-z_3z_4+z_1)-(x_4-z_4)(y_3y_4-y_1-z_3z_4+z_1)] \\ & \quad +a_{44}\{\alpha y_2(x_3x_4-x_1-z_3z_4+z_1)[-z_1(YX)_{431}+x_1(YZ)_{431}] \\ & \quad -x_1(y_3y_4-y_1-z_3z_4+z_1)[y_2(XZ)_{431}-z_2(XY)_{431}]\}=0, \end{aligned} \tag{4.30}$$

$$-a_{11}\alpha y_2(x_3x_4-x_1-z_3z_4+z_1)(YX)_{532}+a_{22}x_1(y_3y_5-y_2-z_3z_5+z_2)(XY)_{431}$$
$$+a_{33}x_1y_2[(x_3x_4-x_1-z_3z_4+z_1)(y_5-z_5)-(y_3y_5-y_2-z_3z_5+z_2)(x_4-z_4)]$$
$$+a_{44}\{\alpha y_2(x_3x_4-x_1-z_3z_4+z_1)[-z_1(YX)_{532}+x_1(YZ)_{532}]$$
$$-x_1(y_3y_5-y_2-z_3z_5+z_2)[y_2(XZ)_{431}-z_2(XY)_{431}]\}=0. \tag{4.31}$$

Again, we sum up the results as a theorem.

Theorem 4: The general solution of Eqs. (3.17) based on the imprimitive action of SL(4,$\mathbb{C}$) can be expressed in terms of three generically chosen particular solutions $\mathbf{x}(t)$, $\mathbf{y}(t)$, $\mathbf{z}(t)$. The general solution $u_a(t)$, $a=1,...,5$ is given by Eq. (3.16), where $z_1,\ldots,z_5$ are constants representing the initial conditions for $\mathbf{u}_a(t)$. The matrix elements $g_{ik}(t)$ are expressed in Eq. (4.27),...,(4.31) in terms of three solutions with initial conditions (4.25), satisfying $\alpha(\alpha-1)\neq 0$, α finite. □

Comments: (1) As in the case of SL(3,$\mathbb{C}$) the reconstruction of $g_{ik}(t)$ is linear. (2) In the primitive case we need five solutions for SL(4,$\mathbb{C}$), in the imprimitive case only three.

D. The group SL(N,$\mathbb{C}$) for $N\geqq 2$

Let us sum up without proof the main results valid for all N.

Theorem 5: (1) The nonlinear ODEs with superposition formulas, based on the action of SL(N,$\mathbb{C}$) on the space SL(N,$\mathbb{C}$)/$P(N)$ are given for all $N\geqq 3$ in Eqs. (3.20) and (3.21). The general form of the solution is given by Eq. (3.18).

(2) The number of equations for SL(N,$\mathbb{C}$) is $n=2N-3$. The group elements $g_{ij}(t)$ can be reconstructed from m particular generically chosen solutions with

$$m=\begin{cases} k+2 & \text{for } N=2k+1 \\ k+1 & \text{for } N=2k \end{cases}. \tag{4.32}$$

(3) Such a fundamental set of m solutions satisfies τ constraints with

$$\tau=mn-N^2+1=\begin{cases} 3k-2 & \text{for } N=2k+1 \\ k-2 & \text{for } N=2k \end{cases}. \tag{4.33}$$

(4) The reconstruction of the group action is linear in the sense that it requires the solution of $2N-3$ linear algebraic equations.

V. CONCLUSIONS

We mentioned in the Introduction that nonlinear ordinary equations with superposition formulas occur in soliton theory as Bäcklund transformations. For equations in the AKNS family[25] the underlying Lie algebra is sl(2,$\mathbb{R}$) and hence the Bäcklund transformations are essentially Riccati equations.

As an example consider the sine-Gordon equation and the Bäcklund transformation, relating two solutions, z_1 and z_2,

$$z_{i,xy}=\sin z_i, \quad i=1,2, \tag{5.1}$$

$$z_{1x}-z_{2x}=2a\sin\frac{z_1+z_2}{2}, \qquad z_{1y}+z_{2y}=\frac{2}{a}\sin\frac{z_1-z_2}{2}. \tag{5.2}$$

The point transformation

$$u_i=\tan\frac{z_i}{4}, \quad i=1,2 \tag{5.3}$$

takes the above system into

$$u_{i,xy}=\frac{u_i}{1+u_i^2}(2u_{i,x}^2-1+u_i^2), \tag{5.4}$$

$$u_{1,x}=\frac{1}{1+u_2^2}[au_2+u_{2,x}+a(1-u_2^2)u_1+(u_{2,x}-au_2)u_1^2],$$

$$u_{1,y}=\frac{1}{1+u_2^2}\left[-\frac{u_2}{a}-u_{2,y}+\frac{1}{a}(1-u_2^2)u_1+\left(\frac{u_2}{a}-u_{2,y}\right)u_1^2\right]. \tag{5.5}$$

Equations (5.5) are Riccati equations for u_1, once u_2 is given.

Wahlquist and Estabrook have proposed[26] a method for finding Bäcklund transformations for a given nonlinear partial differential equation. They introduced a prolongation structure, essentially an additional system of matrix equations. The compatibility conditions are solved by requiring that the right-hand sides of these equations lie in a finite-dimensional Lie algebra. This is the same condition that is required in Lie's theorem on ODEs with superposition formulas.

For integrable multifield equations the Bäcklund transformations are based on other Lie algebras and groups. Thus, for *n*-dimensional generalizations of the sine-Gordon equation and also the wave equation[27,28] the Bäcklund transformations are matrix Riccati equations. Similarly, for Toda field theories (two-dimensional generalized Toda lattices) Bäcklund transformations are given[29,30] that can be transformed into projective Riccati equations.[13] The Bäcklund transformations for nonlinear σ-models are again various types of matrix Riccati equations.[31–33]

All the above examples, and all other Bäcklund transformations for soliton equations that we are aware of, share a common feature. Namely, they have the form of nonlinear ODEs with superposition formulas based on transitive and primitive group actions.

The equations presented in this article correspond to imprimitive actions. We have a group G, in this article $SL(N,\mathbb{C})$. We have a chain of subgroups $G_1\subset G_2\subset G_3\subset\ldots\subset G_{n-1}\subset G_n\equiv G$, where G_{n-1} is maximal in G. Correspondingly, we have a flag of subspaces $G/G_n\subset G/G_{n-1}\subset\cdots\subset G/G_1$. The action of G on G/G_{n-1} is primitive, on the other spaces in the series it is not primitive.

The integrable systems discussed above ''live'' either on Lie groups, like the σ-model, or on the Grassmannians, on which the group acts primitively. Now, there also exist integrable systems on flag manifolds.[34,35] While the manifolds, in particular Grassmannians involved are *a priori*, infinite-dimensional, various reduction schemes lead to finite-dimensional ones.

It is in this direction that we hope that the equations obtained in this article will appear as Bäcklund transformations.

ACKNOWLEDGMENTS

The authors are indebted to O. Navratil and A. Turbiner for helpful discussions. They thank each others institutions for hospitality during mutual visits that made their collaboration possible. The research of P.W. was supported by research grants from the NSERC of Canada and FCAR du Québec, and that of M.H. and S.P. by research grants FRUS 1267/1998 and GaCR 202/96/0218 (1304-1222).

[1] S. Lie, *Vorlesungen über Continuirliche Gruppen mit Geometrischen und Anderen Anwendungen*, Bearbeitet und herausgegeben von Dr. G. Scheffers (Teubner, Leipzig, 1983).

[2] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, London, 1991).

[3] S. Lie, *Theorie der Transformationsgruppen III*, Unter Mitwirkung von F. Engel (Teubner, Leipzig, 1893).

[4] S. Shnider and P. Winternitz, Lett. Math. Phys. **8**, 69 (1984).

[5] S. Shnider and P. Winternitz, J. Math. Phys. **25**, 3155 (1984).

[6] L. Michel and P. Winternitz, ''Families of Transitive Primitive Maximal Simple Lie Subalgebras of diff$_n$,'' in Advances in Math Sciences, CRM's 25 years, *CRM Proceedings and Lecture Notes* (AMS, Providence, R.I., 1997), Vol. 11, pp. 451–479.

[7] V. W. Guillemin and S. Sternberg, Bull. Am. Math. Soc. **70**, 16 (1964).
[8] S. Kobayashi and T. Nagano, J. Math. Mech. **13**, 875 (1964); **14**, 513 (1965).
[9] T. Ochiai, Trans. Am. Math. Soc. **124**, 313 (1966).
[10] M. Golubitsky, J. Diff. Geom. **7**, 175 (1972).
[11] M. Demazure, *Seminaire Bourbaki* (Benjamin, New York, 1968), Vol. 1966–1967, pp. 326–401.
[12] R. L. Anderson, Lett. Math. Phys. **4**, 1 (1982).
[13] R. L. Anderson, J. Harnad, and P. Winternitz, Physica D **4**, 164 (1982).
[14] J. Harnad, P. Winternitz, and R. L. Anderson, J. Math. Phys. **24**, 1062 (1983).
[15] D. W. Rand and P. Winternitz, Comput. Phys. Commun. **33**, 305 (1984).
[16] M. A. del Olmo, M. A. Rodriguez, and P. Winternitz, J. Math. Phys. **27**, 14 (1986); **28**, 530 (1987).
[17] J. Beckers, V. Hussin, and P. Winternitz, Lett. Math. Phys. **11**, 81 (1986); J. Math. Phys. **27**, 2217 (1986); **28**, 520 (1987).
[18] J. Beckers, L. Gagnon, V. Hussin, and P. Winternitz, J. Math. Phys. **31**, 2528 (1996).
[19] B. Grammaticos, A. Ramani, and P. Winternitz, Phys. Lett. A **245**, 382 (1998).
[20] C. Rogers, W. Schief, and P. Winternitz, J. Math. Anal. Appl. **216**, 246 (1997).
[21] M. Havlíček and W. Lassner, Rep. Math. Phys. **8**, 391 (1975).
[22] J. Krause and L. Michel, *Classification of Symmetries of Ordinary Differential Equations, Lecture Notes in Physics* (Springer, Berlin, 1991), Vol. 382, pp. 251–262.
[23] G. W. Mackey, *The Theory of Unitary Group Representations* (University of Chicago Press, Chicago, 1976).
[24] A. O. Barut and R. Racka, *Theory of Group Representations and Applications* (World Scientific, Singapore, 1986).
[25] M. J. Ablowitz, D. Kaup, A. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[26] H. D. Wahlquist and F. B. Estabrook, J. Math. Phys. **16**, 1 (1975); **17**, 1293 (1976).
[27] K. Tenenblat and C. T. Terng, Ann. Math. **111**, 477 (1980).
[28] M. J. Ablowitz, R. Beals, and K. Tenenblat, Stud. Appl. Math. **74**, 177 (1986).
[29] A. V. Mikhailov, Pis'ma Zh. Eksp. Teor. Fiz. **30**, 443 (1979).
[30] A. P. Fordy and J. Gibbons, Commun. Math. Phys. **77**, 21 (1980).
[31] A. T. Ogielski, M. K. Prasad, A. Sinha, and L. L. Chau Wang, Phys. Lett. B **91**, 387 (1980).
[32] J. Harnad, Y. Saint-Aubin, and S. Shnider, Commun. Math. Phys. **93**, 33 (1984).
[33] S. Lafortune and P. Winternitz, J. Math. Phys. **37**, 1539 (1996).
[34] M. Adler, L. Haine, and P. van Moerbeke, Math. Ann. **296**, 1 (1993).
[35] K. Ueno and K. Takasaki, ''Toda lattice hierarchy,'' in *Advanced Studies in Pure Mathematics 4* (North-Holland, Amsterdam, 1984).

Center and universal *R*-matrix for quantized Borcherds superalgebras

Jin Hong[a)]
Department of Mathematics, Seoul National University, Seoul 151-742, Korea

(Received 27 October 1998; accepted for publication 24 December 1998)

We construct a nondegenerate symmetric bilinear form on quantized enveloping algebras associated to Borcherds superalgebras. With this, we study its center and its universal R matrix. © *1999 American Institute of Physics.* [S0022-2488(99)00106-1]

I. INTRODUCTION

Quantized enveloping algebras for Kac–Moody algebras were introduced independently by Drinfel'd (Ref. 1) and Jimbo (Ref. 2) in their studies of the Yang–Baxter equation. The Kac–Moody algebras were generalized (Ref. 3) by Borcherds to accommodate for his study of the monstrous moonshine (Ref. 4). And a quantized version of the enveloping algebras for Borcherds algebras (Ref. 5) was soon studied. There are also superalgebra versions of these algebras (Ref. 6). We shall study the structure of the center and find the R matrix for the quantized Borcherds superalgebras.

Much work has been done on the center of quantized enveloping algebras for finite-dimensional semisimple Lie algebras (Refs. 7–13), and there are Kac–Moody (Ref. 14) and Borcherds (Ref. 15) versions also. We will mainly follow Refs. 13 and 15 to find the center for quantized Borcherds superalgebras.

As for the universal R matrix, the quantum double construction by Drinfel'd (Ref. 16) gives its existence for any Hopf algebra satisfying some conditions. Even though there is a quantum double construction for $\mathbf{Z}_2$-graded Hopf algebras (Ref. 17), we do not use it in this paper. Instead, we explicitly construct a universal R matrix and show that it satisfies the Yang–Baxter equation.

The paper is organized as follows. In Sec. II, we define the quantized Borcherds superalgebras and give it a Hopf algebra structure. The triangular decomposition will also be mentioned. In Sec. III, the character formula for highest weight representations will be given and we prove a lemma that will be used in later sections. The next section is devoted to providing the quantized Borcherds superalgebras with a bilinear form and proving its nondegeneracy. In Sec. V, we define the Harish–Chandra homomorphism, show its injectivity, and prove some properties concerning its image. Information on the center of the quantized Borcherds superalgebra will be obtained in Sec. VI. In the last section we will give the universal R matrix and show that it satisfies the Yang–Baxter equation.

II. QUANTUM DEFORMATION OF BORCHERDS SUPERALGEBRAS

In this section, we define the quantized Borcherds superalgebras and give it a Hopf algebra structure.

Let I be a countable index set. A matrix $A=(a_{i,j})_{i,j\in I}$ with entries in the real numbers is a *Borcherds–Cartan* matrix if (i) $a_{i,i}=2$ or $a_{i,i}\leqslant 0$ for all $i\in I$; (ii) $a_{i,j}\leqslant 0$ if $i\neq j$ and $a_{i,j}\in\mathbf{Z}$ if $a_{i,i}=2$; (iii) $a_{i,j}=0$ if and only if $a_{j,i}=0$.

If there exists a diagonal matrix $D=\mathrm{diag}(s_i|i\in I,s_i>0)$ such that DA is symmetric, then A is said to be *symmetrizable*. If a symmetrizable Borcherds–Cartan matrix A further satisfies the

[a)]Electronic mail: jhong@math.snu.ac.kr

constraints, $a_{i,j}\in\mathbf{Z}$, $s_i\in\mathbf{Z}_{>0}$, for all $i,j\in I$, then it is said to be *integral*. We will be citing many results from Ref. 6. The condition $a_{i,i}\in 2\mathbf{Z}$ appearing therein is superfluous.

A complex matrix $C=(\theta_{i,j})_{i,j\in I}$ is a *coloring matrix* if $\theta_{i,j}\theta_{j,i}=1$ for all $i,j\in I$. Necessarily, $\theta_{i,i}=\pm 1$, and we say i is *even* when $\theta_{i,i}=1$, *odd* when $\theta_{i,i}=-1$. A Borcherds–Cartan matrix A is *colored by C* if for every $i\in I$ such that $a_{i,i}=2$ and $\theta_{i,i}=-1$ we have, $a_{i,j}\in 2\mathbf{Z}$ for all $j\in I$.

Throughout this paper, we shall assume that A is a symmetrizable integral Borcherds–Cartan matrix that is colored by a coloring matrix C.

Let $I^{\mathrm{re}}=\{i\in I|a_{i,i}=2\}$ and $I^{\mathrm{im}}=\{i\in I|a_{i,i}\leqslant 0\}$. Also, let $\underline{m}=(m_i|i\in I)$ be a collection of positive integers such that $m_i=1$ for all $i\in I^{\mathrm{re}}$. We call $\underline{m}$ a charge of the Borcherds–Cartan matrix A.

For a symmetrizable integral Borcherds–Cartan matrix A, which is colored by a coloring matrix C, we denote by $\mathfrak{g}(A,\underline{m},C)$ the Borcherds superalgebra of charge $\underline{m}$. (See Ref. 6.)

We set $P^\vee=(\oplus_{i\in I}\mathbf{Z}h_i)\oplus(\oplus_{i\in I}\mathbf{Z}d_i)$ and let $\mathfrak{h}=\mathbf{C}\otimes_{\mathbf{Z}}P^\vee$ be the complex vector space with basis $\{h_i,d_i|i\in I\}$. For $i\in I$, we define α_i in the dual space $\mathfrak{h}^*$ of $\mathfrak{h}$ by setting $\alpha_i(h_j)=a_{j,i}$ and $\alpha_i(d_j)=\delta_{i,j}$. Since A is assumed to be symmetrizable, there exists a nondegenerate symmetric bilinear form $(|)$ on $\mathfrak{h}$, given by $(s_ih_i|h)=\alpha_i(h)$ and $(d_i|d_j)=0$ for $i,j\in I,h\in\mathfrak{h}$.

The free Abelian group $Q=\oplus_{i\in I}\mathbf{Z}\alpha_i$ generated by the $\alpha_i(i\in I)$ is called the *root lattice* associated to A. Let $Q^+=\Sigma_{i\in I}\mathbf{Z}_{\geqslant 0}\alpha_i$ and $Q^-=-Q^+$. The coloring matrix $C=(\theta_{i,j})$ gives rise to a complex-valued mapping $\theta:Q\times Q\to\mathbf{C}^\times$, satisfying $\theta(\alpha_i,\alpha_j)=\theta_{i,j}$, $\theta(\alpha,\beta+\gamma)=\theta(\alpha,\beta)\theta(\alpha,\gamma)$, $\theta(\alpha+\beta,\gamma)=\theta(\alpha,\gamma)\theta(\beta,\gamma)$, for all $\alpha,\beta,\gamma\in Q$.

We define the binomial coefficients by setting $\{n\}_{q_i}=[(\theta_{i,i}^nq_i^n-q_i^{-n})/(\theta_{i,i}q_i-q_i^{-1})]$, $\{n\}_{q_i}!=\Pi_{t=1}^n\{t\}_{q_i}$, and $\left\{{m \atop n}\right\}_{q_i}=\{m\}_{q_i}!/\{n\}_{q_i}!\{m-n\}_{q_i}!$, where $\{0\}_{q_i}!=1$ and $q_i=q^{s_i}$. Let $\xi_i=q_i-q_i^{-1}$ and $K_i=q^{s_ih_i}$.

Definition II.1 (Ref. 6): Suppose $\mathfrak{g}=\mathfrak{g}(A,\underline{m},C)$ is the Borcherds superalgebra of charge $\underline{m}$ determined by the symmetrizable integral Borcherds–Cartan matrix A that is colored by a coloring matrix C. Let q be an indeterminant. Then the *quantized Borcherds superalgebra* $U_q(\mathfrak{g})$ associated to $\mathfrak{g}$ is the associative algebra over $\mathbf{C}(q)$ with 1, generated by the elements $q^h(h\in P^\vee)$, $e_{i,k}$, $f_{i,k}$ $(i\in I,\ k=1,2,...,m_i)$ with the defining relations:

(R1) $q^0=1,\ q^hq^{h'}=q^{h+h'},\quad \text{for } h,h'\in P^\vee,$

(R2) $q^he_{i,k}q^{-h}=q^{\alpha_i(h)}e_{i,k},\quad \text{for } h\in P^\vee,\quad i\in I,k=1,2,...,m_i,$

(R3) $q^hf_{i,k}q^{-h}=q^{-\alpha_i(h)}f_{i,k},\quad \text{for } h\in P^\vee,\quad i\in I,\quad k=1,2,...,m_i,$

(R4) $e_{i,k}f_{j,l}-\theta_{j,i}f_{j,l}e_{i,k}=\delta_{i,j}\delta_{k,l}\frac{1}{\xi_i}(K_i-K_i^{-1}),\quad \text{for } i,j\in I,\ k=1,2,...,m_i,\quad l=1,2,...,m_j,$

(R5) $\displaystyle\sum_{n=0}^{1-a_{i,j}}(-1)^n\theta_{i,j}^n\theta_{i,i}^{n(n-1)/2}\left\{{l-a_{i,j} \atop n}\right\}_{q_i}e_{i,k}^{1-a_{i,j}-n}e_{j,l}e_{i,k}^n=0,\quad \text{if } a_{i,i}=2 \text{ and } i\neq j,$

(R6) $\displaystyle\sum_{n=0}^{1-a_{i,j}}(-1)^n\theta_{i,j}^n\theta_{i,i}^{n(n-1)/2}\left\{{l-a_{i,j} \atop n}\right\}_{q_i}f_{i,k}^{1-a_{i,j}-n}f_{j,l}f_{i,k}^n=0,\quad \text{if } a_{i,i}=2 \text{ and } i\neq j,$

(R7) $e_{i,k}e_{j,l}-\theta_{i,j}e_{j,l}e_{i,k}=0,\quad \text{if } a_{i,j}=0,$

(R8) $f_{i,k}f_{j,l}-\theta_{i,j}f_{j,l}f_{i,k}=0,\quad \text{if } a_{i,j}=0.$

Proposition II.2 (Ref. 6): The algebra $U_q(\mathfrak{g})$ has a Hopf superalgebra structure with comultiplication Δ, counit ϵ, and antipode S, defined by

$$\Delta(q^h)=q^h\otimes q^h, \tag{1}$$

$$\Delta(e_{i,k})=e_{i,k}\otimes 1+K_i\otimes e_{i,k}, \tag{2}$$

$$\Delta(f_{i,k})=f_{i,k}\otimes K_i^{-1}+1\otimes f_{i,k}, \tag{3}$$

$$\epsilon(q^h)=1, \tag{4}$$

$$\epsilon(e_{i,k})=0, \tag{5}$$

$$\epsilon(f_{i,k})=0, \tag{6}$$

$$S(q^h)=q^{-h}, \tag{7}$$

$$S(e_{i,k})=-K_i^{-1}e_{i,k}, \tag{8}$$

$$S(f_{i,k})=-f_{i,k}K_i, \tag{9}$$

for $h\in P^\vee$, $i\in I$, $k=1,2,...,m_i$.

We denote by U^0 the subalgebra of $U=U_q(\mathfrak{g})$ generated by q^h for $h\in P^\vee$ and U^+ (respectively, U^-) the subalgebra of U generated by the elements $e_{i,k}$ (respectively, $f_{i,k}$) for $i\in I$, $k=1,2,...,m_i$. We also denote by $U^{\geqslant 0}$ (respectively, $U^{\leqslant 0}$) the subalgebra of U generated by the elements q^h and $e_{i,k}$ (respectively, q^h and $f_{i,k}$) for $h\in P^\vee$, $i\in I$, $k=1,2,...,m_i$. For each $\beta\in Q$, let

$$U_\beta=\{x\in U|q^h x q^{-h}=q^{\beta(h)}x, \quad \text{for all} \;\; h\in P^\vee\}. \tag{10}$$

We similarly define $U^\pm_{\pm\beta}$, $U^{\geqslant 0}_{\pm\beta}$, and $U^{\leqslant 0}_{\pm\beta}$ for $\beta\in Q^+$. We then have the following.

Proposition II.3 (Ref. 6):

(a) $U\cong U^-\otimes U^0\otimes U^+$.

(b) $U^0=\oplus_{h\in P^\vee}\mathbf{C}q^h$.

(c) $U^\pm=\oplus_{\beta\in Q^+}U^\pm_{\pm\beta}$.

(d) (R5) *and* (R7) [*respectively*, (R6) *and* (R8)] *are the fundamental relations for* U^+ (*respectively*, U^-).

We give Q^+ a partial ordering by setting $\lambda\geqslant\mu$ if and only if $\lambda-\mu\in Q^+$. We will also use the notation $K_\gamma=\Pi K_i^{n_i}$ for $\gamma=\Sigma n_i\alpha_i\in Q$.

III. REPRESENTATIONS OF $U_q(\mathfrak{g})$

For $i\in I$ define the **C**-linear functionals $\Lambda_i\in\mathfrak{h}^*$ by

$$\Lambda_i(h_j)=\delta_{i,j} \quad \Lambda_i(d_j)=0, \quad \text{for all} \;\; j\in I. \tag{11}$$

Define the lattices:

$$P=\{\lambda\in\mathfrak{h}^*|\lambda(h_i),\lambda(d_i)\in\mathbf{Z},\forall i\in I\}, \tag{12}$$

$$\overline{P}=\Big(\bigoplus_{i\in I}\mathbf{Z}\alpha_i\Big)\oplus\Big(\bigoplus_{i\in I}\mathbf{Z}\Lambda_i\Big). \tag{13}$$

P is called the *weight lattice* of $\mathfrak{g}$. An element $\lambda\in P$ is said to be a *dominant integral weight* if

$$\lambda(h_i)\in\mathbf{Z}_{\geqslant 0}, \quad \text{for all} \;\; i\in I^{\text{re}}, \tag{14}$$

$$\lambda(h_i)\in 2\mathbf{Z}_{\geqslant 0}, \quad \text{for all} \;\; i\in I^{\text{re}}\cap I^{\text{odd}}, \tag{15}$$

where I^{odd} denotes the set of $i\in I$ such that $\theta_{i,i}=-1$. Let P^+ denote the set of all dominant integral weights.

Set $\bar{\mathfrak{h}}^* = \mathbf{C}\otimes_{\mathbf{Z}}\bar{P}$. Then the nondegenerate symmetric bilinear form on $\mathfrak{h}$ gives an isomorphism between $\mathfrak{h}$ and $\bar{\mathfrak{h}}^*$ hence also induces a bilinear form on $\bar{\mathfrak{h}}^*$. We may extend this bilinear form to a symmetric bilinear form on $\mathfrak{h}^*$. We extend it so that it satisfies $(\lambda|\alpha_i)=\lambda(s_i h_i)$ and $(\lambda|\Lambda_i)=\lambda(s_i d_i)$ for every $\lambda\in\mathfrak{h}^*$. Write $\lambda\perp\mu$ if $(\lambda|\mu)=0$.

For each $i\in I$ such that $a_{i,i}\neq 0$, we define the *simple reflection* $r_i\in \mathrm{GL}(\mathfrak{h}^*)$ on $\mathfrak{h}^*$ by

$$r_i(\lambda)=\lambda-\frac{2}{a_{i,i}}\lambda(h_i)\alpha_i. \tag{16}$$

The subgroup W of $\mathrm{GL}(\mathfrak{h}^*)$ generated by $r_i (i\in I^{\mathrm{re}})$ is called the *Weyl group* of $\mathfrak{g}(A,\underline{m},C)$. We denote by $l: W\to\mathbf{Z}_{\geqslant 0}$ the natural length function.

Let R be the family of all imaginary simple roots, each root occurring as many times as its multiplicity, i.e., m_i times for α_i. For $\lambda\in P^+$, define $R(\lambda)$ to be the set of all $\mu=\sum_{j=1}^{r}\alpha_{i_j}+\sum_{k=1}^{s} l_{i_k}\beta_{i_k}\in Q^+$, where α_{i_j} (resp., β_{i_k}) are distinct even (resp., odd) roots in R, satisfying (i) $\alpha_{i_j}\perp\lambda, \beta_{i_k}\perp\lambda$, for all j,k; (ii) $\alpha_{i_j}\perp\beta_{i_k}$, for all j,k; (iii) $\alpha_{i_j}\perp\alpha_{i_k}, \beta_{i_j}\perp\beta_{i_k}$, for $j\neq k$; (iv) $\beta_{i_k}\perp\beta_{i_k}$, if $l_{i_k}\geqslant 2$. In particular, $0\in R(\lambda)$. For μ as above, we define

$$ht(\mu)=r+\sum_{k=1}^{s} d_{ik}.$$

Suppose $\rho\in\mathfrak{h}^*$ satisfies $\rho(h_i)=\frac{1}{2}a_{i,i}$ for all $i\in I$.

Proposition III.1 (Refs. 18 and 19): Let $\lambda\in P^+$. *Denote by $M^q(\lambda)$ the Verma module for $U_q(\mathfrak{g})$ with highest weight λ and let $V^q(\lambda)$ be the irreducible highest weight module over $U_q(\mathfrak{g})$ with highest weight λ. Then,*

$$\mathrm{ch}\, M^q(\lambda)=\frac{e^{\lambda}}{\prod_{\alpha\in\Phi^-}(1-\theta(\alpha,\alpha)e^{\alpha})^{\theta(\alpha,\alpha)\dim\mathfrak{g}_\alpha}}=e^{\lambda}\sum_{\beta\in Q^+}(\dim U^-_{-\beta})e^{-\beta}, \tag{17}$$

$$\mathrm{ch}\, V^q(\lambda)=\frac{\sum_{w\in W,\mu\in R(\lambda)}(-1)^{l(w)+ht(\mu)}e^{w(\lambda+\rho-\mu)-\rho}}{\prod_{\alpha\in\Phi^-}(1-\theta(\alpha,\alpha)e^{\alpha})^{\theta(\alpha,\alpha)\dim\mathfrak{g}_\alpha}}. \tag{18}$$

In this formula, Φ^- is the *set of all negative roots.*

The following is a corollary to this proposition.

Lemma III.2: Let $\gamma=\sum_{i\in I}n_i\alpha_i\in Q^+$. Suppose $\lambda\in P^+$, $\lambda(h_i)>0$ for all $i\in I^{\mathrm{im}}$ and $\lambda(h_i)\geqslant n_i$ for all $i\in I^{\mathrm{re}}$. Then we have a linear isomorphism $U^-_{-\gamma}\xrightarrow{\sim}V^q(\lambda)_{\lambda-\gamma}$ given by $u\mapsto uv_\lambda$.

Proof: $U^-_{-\gamma}\to M^q(\lambda)_{\lambda-\gamma}$ is surjective, so $U^-_{-\gamma}\to V^q(\lambda)_{\lambda-\gamma}$ is also surjective. Hence, it suffices to show $\dim U^-_{-\gamma}=\dim V^q(\lambda)_{\lambda-\gamma}$. Since $(\alpha_i|\lambda)=\lambda(s_i h_i)=s_i\lambda(h_i)>0$ for all $i\in I^{\mathrm{im}}$, no nonempty subset F of R satisfies $F\perp\lambda$, and so

$$\mathrm{ch}\, V^q(\lambda)=\frac{\sum_{w\in W}(-1)^{l(w)}e^{w(\lambda+\rho)-\rho}}{\prod_{\alpha\in\Phi^-}(1-\theta(\alpha,\alpha)e^{\alpha})^{\theta(\alpha,\alpha)\dim\mathfrak{g}_\alpha}}, \tag{19}$$

$$=\left(\sum_{w\in W}(-1)^{l(w)}e^{w(\lambda+\rho)-\rho}\right)\left(\sum_{\beta\in Q^+}(\dim U^-_{-\beta})e^{-\beta}\right). \tag{20}$$

Therefore, it suffices to show that if $w(\lambda+\rho)-\rho-\beta=\lambda-\gamma$ for some $w\in W$, $\beta\in Q^+$, then $w=1$.

We will show that if $w\neq 1$, then $\gamma+w(\lambda+\rho)-(\lambda+\rho)\notin Q^+$ by using induction on the length of w.

If $w=r_i(i\in I^{\mathrm{re}})$, then

$$\gamma+r_i(\lambda+\rho)-(\lambda+\rho)=\gamma+\lambda+\rho-(\lambda(h_i)+\rho(h_i))\alpha_i-(\lambda+\rho)=\gamma-(\lambda(h_i)+1)\alpha_i\notin Q^+.$$

If $w=w' r_i(i\in I^{\mathrm{re}})$ with $l(w)=l(w')+1$, then

$$\gamma+w(\lambda+\rho)-(\lambda+\rho)=\gamma+w'r_i(\lambda+\rho)-(\lambda+\rho)$$
$$=\gamma+w'(\lambda+\rho-(\lambda(h_i)+\rho(h_i))\alpha_i)-(\lambda+\rho)$$
$$=(\gamma+w'(\lambda+\rho)-(\lambda+\rho))-(\lambda(h_i)+1)w'\alpha_i\notin Q^+.$$

This completes the proof. □

IV. THE BILINEAR FORM ON $U_q(\mathfrak{g})$

IV.1. *The bilinear form on* $U^{\geqslant 0}\times U^{\leqslant 0}$. In this section, we define a bilinear form on $U^{\geqslant 0}\times U^{\leqslant 0}$, which is nondegenerate when restricted to $U^+_\beta\times U^-_{-\beta}$, $\beta\in Q^+$.

For $\phi\in(U_\beta)^*$, $\psi\in(U_\gamma)^*$, $x\in U_\beta$, and $y\in U_\gamma$, we define $(\phi\otimes\psi)(x\otimes y)=\theta(-\gamma,\beta)\phi(x)\psi(y)$. With this, and the Hopf algebra structure on $U_q(\mathfrak{g})$, we can give an algebra structure to $\oplus_{\alpha\in Q^+}(U^{\geqslant 0}_\alpha)^*$ by setting $(\phi_1\phi_2)(x)=(\phi_1\otimes\phi_2)(\Delta(x))$ for $\phi_1,\phi_2\in\oplus_{\alpha\in Q^+}(U^{\geqslant 0}_\alpha)^*$ and $x\in U^{\geqslant 0}$. For $h\in P^\vee$ and $i\in I$, $k=1,2,...,m_i$, we define the linear functionals $\phi_h,\psi_{i,k}\in\oplus_{\alpha\in Q^+}(U^{\geqslant 0}_\alpha)^*$ by

$$\phi_h(xq^{h'})=\epsilon(x)q^{-(h|h')}\quad(x\in U^+,h'\in P^\vee),\tag{21}$$

$$\psi_{i,k}(xq^h)=0\quad(x\in U^+_\beta,\beta\in Q^+\backslash\{\alpha_i\}),\tag{22}$$

$$\psi_{i,k}(e_{i,l}q^h)=\delta_{k,l}.\tag{23}$$

Proposition IV.1: There exists an algebra homomorphism,

$$\zeta:U^{\leqslant 0}\rightarrow\bigoplus_{\alpha\in Q^+}(U^{\geqslant 0}_\alpha)^*,\tag{24}$$

given by

$$\zeta(q^h)=\phi_h,\quad(h\in P^\vee),\tag{25}$$

$$\zeta(f_{i,k})=-\frac{1}{\xi_i}\psi_{i,k}\quad(i\in I,k=1,2,...,m_i).\tag{26}$$

Proof: By Proposition II.3, we have only to check that the relations (R1), (R3), (R6), and (R8) are preserved under the map ζ. Other cases being easy, we just sketch the (R6) part.

Define $e^{(n)}_{i,k}=e^n_{i,k}/\{n\}_{q_i}!$. We may check by induction on n that

$$\Delta(e^{(n)}_{i,k})=\sum_{s+t=n}q_i^{st}e^{(s)}_{i,k}K_i^t\otimes e^{(t)}_{i,k}.$$

This shows

$$((\Delta\otimes 1)\circ\Delta)(e^{(n)}_{i,k})=\sum_{r+s+t=n}q_i^{rs+st+tr}e^{(r)}_{i,k}K_i^{s+t}\otimes e^{(s)}_{i,k}K_i^t\otimes e^{(t)}_{i,k}.$$

We again use induction to prove

$$\psi^n_{i,k}(e^{(n)}_{i,k})=(\theta_{i,i}q_i)^{n(n-1)/2}.$$

With this, it is possible to show

$$\psi_{i,k}^{N-n}\psi_{j,l}\psi_{i,k}^{n}(e_{i,k}^{(N-m)}e_{j,l}e_{i,k}^{(m)})=\sum \theta_{i,i}^{A'}\theta_{i,j}^{B'}q_i^{C'}\begin{Bmatrix}N-n\\ \alpha\end{Bmatrix}_{q_i}\begin{Bmatrix}n\\ \beta\end{Bmatrix}_{q_i},$$

with the summation over non-negative integers α, β, γ, δ such that $\alpha+\beta=N-m$, $\gamma+\delta=m$, $\alpha+\gamma=N-n$, and $\beta+\delta=n$, and where

$$A'=\beta\gamma+(N-n)n+\tfrac{1}{2}(N-n)(N-n-1)+\tfrac{1}{2}n(n-1),$$

$$B'=\beta-\gamma+(N-n)-n,$$

$$C'=\alpha\beta+\gamma\delta+2\beta\gamma+(\beta+\gamma)a_{i,j}+\tfrac{1}{2}(N-n)(N-n-1)+\tfrac{1}{2}n(n-1).$$

Noting

$$\begin{Bmatrix}N\\ n\end{Bmatrix}_{q_i}\begin{Bmatrix}N-n\\ \alpha\end{Bmatrix}_{q_i}\begin{Bmatrix}n\\ N-m-\alpha\end{Bmatrix}_{q_i}=\begin{Bmatrix}N\\ m\end{Bmatrix}_{q_i}\begin{Bmatrix}N-m\\ \beta\end{Bmatrix}_{q_i}\begin{Bmatrix}m\\ \delta\end{Bmatrix}_{q_i},$$

we can calculate

$$\left(\sum_{n=0}^{1-a_{i,j}}(-1)^n\theta_{i,j}^{n}\theta_{i,i}^{n(n-1)/2}\begin{Bmatrix}1-a_{i,j}\\ n\end{Bmatrix}_{q_i}\psi_{i,k}^{1-a_{i,j}-n}\psi_{j,l}\psi_{i,k}^{n}\right)(e_{i,k}^{(N-m)}e_{j,l}e_{i,k}^{(m)})$$
$$=\sum_{n=0}^{N}\sum \theta_{i,i}^{A}\theta_{i,j}^{B}q_i^{C}\begin{Bmatrix}N\\ m\end{Bmatrix}_{q_i}\begin{Bmatrix}N-m\\ \beta\end{Bmatrix}_{q_i}\begin{Bmatrix}m\\ \delta\end{Bmatrix}_{q_i},$$

with the second summation over non-negative integers satisfying the same conditions as before, and where

$$A=\tfrac{1}{2}N(N-1)+m\beta+\tfrac{1}{2}\beta(\beta-1)+\tfrac{1}{2}\delta(\delta-1),$$

$$B=N-m,$$

$$C=(m-mN+\tfrac{1}{2}N(N-1))+(m+1-N)\beta+(m-1)\delta.$$

This can be written as a product of two sums that simplifies to zero. □

Define a bilinear form $(\,|\,):U^{\geqslant 0}\times U^{\leqslant 0}\rightarrow \mathbf{C}(q)$ by

$$(x|y)=\zeta(y)(x)\quad (x\in U^{\geqslant 0},y\in U^{\leqslant 0}). \tag{27}$$

For $n\in\mathbf{Z}_{>0}$, we denote by $\Delta_n:U\rightarrow U^{\otimes(n+1)}$, the algebra homomorphism defined by $\Delta_1=\Delta$, $\Delta_n=(\Delta\otimes 1)\circ\Delta_{n-1}$, and we write

$$\Delta_n(x)=\sum_{(x)_n}x_{(0)}\otimes x_{(1)}\otimes\cdots\otimes x_{(n)}. \tag{28}$$

For homogeneous elements $x_i\in U_{\beta_i}^{\geqslant 0}$, $y_i\in U_{-\gamma_i}^{\leqslant 0}$ $(i=1,2)$, we define $(x_1\otimes x_2|y_1\otimes y_2)=\theta(\beta_2,-\gamma_1)(x_1|y_1)(x_2|y_2)$ and extend it by linearity. For $x\in U_\beta$, $y\in U_\gamma$, we will write $\theta(x,y)$ to mean $\theta(\beta,\gamma)$ and define $\sigma:U\otimes U\rightarrow U\otimes U$ by $\sigma(x\otimes y)=\theta(x,y)y\otimes x$ on homogeneous elements and extend it by linearity.

Proposition IV.2: The bilinear form $(\,|\,)$ *on* $U^{\geqslant 0}\times U^{\leqslant 0}$ *defined by (27) satisfies*

$$(x|y_1y_2)=(\Delta(x)|y_1\otimes y_2)\quad (x\in U^{\geqslant 0},y_1,y_2\in U^{\leqslant 0}), \tag{29}$$

$$(x_1x_2|y)=(\sigma(x_1\otimes x_2)|\Delta(y))\quad (x_1,x_2\in U^{\geqslant 0},y\in U^{\leqslant 0}), \tag{30}$$

$$(q^h|q^{h'})=q^{-(h|h')} \quad (h,h'\in P^\vee), \tag{31}$$

$$(q^h|f_{i,k})=0, \tag{32}$$

$$(e_{i,k}|q^h)=0, \tag{33}$$

$$(e_{i,k}|f_{j,l})=-\frac{1}{\xi_i}\delta_{i,j}\delta_{k,l}, \tag{34}$$

for $i,j\in I$, $k=1,2,...,m_j$.

Moreover, the bilinear form on $U^{\geqslant 0}\times U^{\leqslant 0}$ *satisfying the above equations is unique.*

Proof: Everything including uniqueness is straightforward, except for (30). It is proved by induction. Here we just show the induction part. We suppress the summation signs for simplicity. Assume $(x^1x^2|y^i)=\theta(x^1,x^2)(x^2\otimes x^1|\Delta(y^i))$ for $i=1,2$. Then

$$\begin{aligned}
(x^1x^2|y^1y^2)&=\zeta(y^1)\zeta(y^2)(x^1x^2)\\
&=\zeta(y^1)\otimes\zeta(y^2)((x^1_{(0)}\otimes x^1_{(1)})\cdot(x^2_{(0)}\otimes x^2_{(1)}))\\
&=\zeta(y^1)\otimes\zeta(y^2)(\theta(x^1_{(1)},x^2_{(0)})x^1_{(0)}x^2_{(0)}\otimes x^1_{(1)}x^2_{(1)})\\
&=\theta(x^1_{(1)},x^2_{(0)})\theta(y^2,x^1_{(0)})\theta(y^2,x^2_{(0)})(x^1_{(0)}x^2_{(0)}|y^1)(x^1_{(1)}x^2_{(1)}|y^2)\\
&=\theta(x^1_{(0)},x^1_{(1)})\theta(x^1_{(0)},x^2_{(1)})\theta(x^2_{(0)},x^2_{(1)})\theta(x^1_{(0)},x^2_{(0)})\theta(x^1_{(1)},x^2_{(1)})\\
&\quad\times(x^2_{(0)}\otimes x^1_{(0)}|y^1_{(0)}\otimes y^1_{(1)})(x^2_{(1)}\otimes x^1_{(1)}|y^2_{(0)}\otimes y^2_{(1)})\\
&=\theta(x^1_{(0)},x^1_{(1)})\theta(x^1_{(0)},x^2_{(1)})\theta(x^2_{(0)},x^2_{(1)})(x^2_{(0)}|y^1_{(0)})(x^1_{(0)}|y^1_{(1)})(x^2_{(1)}|y^2_{(0)})(x^1_{(1)}|y^2_{(1)})\\
&=\theta(y^1_{(1)},y^2_{(0)})(\Delta(x^2)|y^1_{(0)}\otimes y^2_{(0)})(\Delta(x^1)|y^1_{(1)}\otimes y^2_{(1)})\\
&=\theta(y^1_{(1)},y^2_{(0)})\theta(y^1_{(0)}y^2_{(0)},x^1)(x^2\otimes x^1|y^1_{(0)}y^2_{(0)}\otimes y^1_{(1)}y^2_{(1)})\\
&=\theta(x^1,x^2)(x^2\otimes x^1|(y^1_{(0)}\otimes y^1_{(1)})\cdot(y^2_{(0)}\otimes y^2_{(1)}))=\theta(x^1,x^2)(x^2\otimes x^1|\Delta(y^1y^2)).
\end{aligned}$$

This completes the proof. □

Lemma IV.3: (a) $(S(x)|S(y))=(x|y)$; (b) $(xq^h|yq^{h'})=q^{-(h|h')}(x|y)$ $(h,h'\in P^\vee, x\in U^+, y\in U^-)$; (c) $(U^+_\beta|U^-_{-\gamma})=0$, *if* $\gamma\neq\beta$.

Proof: To prove (a), we set $(\ |\ )'=(S()|S())$ and show $(\ |\ )'$ satisfies conditions of Proposition IV.2. The remaining two are easy. □

Lemma IV.4: For $x\in U^{\geqslant 0}$, $y\in U^{\leqslant 0}$, *which are homogeneous, we have*

$$\theta(x,y)yx=\sum_{(x)_2,(y)_2}\Theta_{xy}(x_{(0)}|S(y_{(0)}))(x_{(2)}|y_{(2)})x_{(1)}y_{(1)} \tag{35}$$

and

$$xy=\sum_{(x)_2,(y)_2}\theta(x_{(1)},y_{(1)})\Theta_{xy}(x_{(0)}|y_{(0)})(x_{(2)}|S(y_{(2)}))y_{(1)}x_{(1)}, \tag{36}$$

with $\Theta_{xy}=\theta(x_{(1)},y_{(0)})\theta(x_{(2)},y_{(0)})\theta(x_{(2)},y_{(1)})$.

Proof: By substituting (35) onto the right-hand side of (36), we can show that (35) implies (36).

To prove (35), we use induction on y and reduce the problem to showing this true for $y=q^h$ and $y=f_{i,k}$. The case $y=q^h$ is easy. The case $y=f_{i,k}$ turns out to be equivalent to showing

$$\theta(x,f_{i,k})f_{i,k}x=\sum_{(x)_1}\{(x_{(1)}|f_{i,k})x_{(0)}+\theta(x_{(1)},f_{i,k})(x_{(1)}|K_i^{-1})x_{(0)}f_{i,k}$$
$$-\theta(x_{(1)},f_{i,k})(x_{(0)}|f_{i,k})K_i^{-1}x_{(1)}\}, \tag{37}$$

which is proved by induction on the length of x. □

Lemma IV.5: Let $\beta\in Q^+\setminus\{0\}$ *and* $y\in U^-_{-\beta}$. *If* $e_{i,k}y=\theta(\alpha_i,-\beta)ye_{i,k}$, *for all* $i\in I$, $k=1,2,...,m_i$, *then* $y=0$.

Proof: Choose $\lambda\in P^+$ satisfying the assumptions of Lemma III.2. Since $e_{i,k}(y\cdot v_\lambda)=\theta(\alpha_i,-\beta)y(e_{i,k}\cdot v_\lambda)=0$ for all $i\in I$, $k=1,2,...,m_i$, and $\mathrm{wt}(y\cdot v_\lambda)=\lambda-\beta\leqslant\lambda$, $y\cdot v_\lambda$ generates a proper submodule of $V^q(\lambda)$. Hence $y\cdot v_\lambda=0$. Lemma III. 2 now says $y=0$. □

Theorem IV.6: *For* $\beta\in Q^+$, *the bilinear form* $(\,|\,):U^{\geqslant 0}\times U^{\leqslant 0}\to\mathbf{C}(q)$ *defined by* (27) *is nondegenerate when restricted to* $U^+_\beta\times U^-_{-\beta}$.

Proof: Since $\dim U^+_\beta=\dim U^-_{-\beta}$, nondegeneracy on one side implies the nondegeneracy on the other side. So we will just prove the following statement:

$$\text{if}\ \ y\in U^-_{-\beta}\quad\text{and}\quad(U^+_\beta|y)=0,\quad\text{then}\ \ y=0. \tag{38}$$

We use induction on β.

The case $\beta=0$ or α_i is easy.

Assume (38) is true for all $\gamma<\beta$ with $\beta\in Q^+\setminus(\{0\}\cup\{\alpha_i\}_{i\in I})$. Recall the notation $K_\gamma=q\Sigma_{i\in I}n_is_ih_i$ for $\gamma=\Sigma_{i\in I}n_i\alpha_i\in Q$. By definition of Δ, we see that

$$\Delta(y)=\sum_{0\leqslant\gamma\leqslant\beta}y_\gamma(1\otimes K_{-\gamma}),\quad y_\gamma\in U^-_{-\gamma}\otimes U^-_{-(\beta-\gamma)}, \tag{39}$$

with $y_0=1\otimes y$ and $y_\beta=y\otimes 1$. Fix $0<\gamma<\beta$. For any $u\in U^+_{\beta-\gamma}$ and $v\in U^+_\gamma$, we have

$$(v\otimes u|y_\gamma)=(v\otimes u|y_\gamma(1\otimes K_{-\gamma})), \tag{40}$$

$$=(v\otimes u|\Delta(y)), \tag{41}$$

$$=\theta(\gamma,\beta-\gamma)(uv|y),\quad\text{by (30)}, \tag{42}$$

$$=0. \tag{43}$$

Hence $(U^+_\gamma\otimes U^+_{\beta-\gamma}|y_\gamma)=0$. This implies $y_\gamma=0$ by our induction hypothesis. Therefore $\Delta(y)=y\otimes K_{-\beta}+1\otimes y$. We apply Lemma IV.4 to

$$\Delta_2(e_{i,k})=e_{i,k}\otimes 1\otimes 1+K_i\otimes e_{i,k}\otimes 1+K_i\otimes K_i\otimes e_{i,k}, \tag{44}$$

$$\Delta_2(y)=y\otimes K_{-\beta}\otimes K_{-\beta}+1\otimes y\otimes K_{-\beta}+1\otimes 1\otimes y, \tag{45}$$

and get

$$\theta(\alpha_i,-\beta)ye_{i,k}=e_{i,k}y,\quad\text{for all}\ \ i\in I. \tag{46}$$

Hence, $y=0$ by Lemma IV.5. □

IV.2. *The Killing form.* Recall from Proposition II.3 that $U\cong U^+\otimes U^0\otimes S(U^-)\cong U^-\otimes U^0\otimes S(U^+)$. Using the bilinear form defined in the previous section, we define a new bilinear form,

$$\langle\,|\,\rangle:U\times U\to\mathbf{C}(q^{1/2}),$$

by setting

$$\langle x_1 q^{h_1} S(y_1) | y_2 q^{h_2} S(x_2) \rangle = (x_1|y_2)(x_2|y_1) q^{-(h_1|h_2)/2} \theta(y_1,y_2)\,\theta(y_1,x_2), \tag{47}$$

for homogeneous $x_i \in U^+$, $y_i \in U^-$, $h_i \in P^\vee$ and extending by linearity.

For homogeneous $u,v \in U$, we define

$$\mathrm{ad}(u)\cdot v = \sum_{(u)_1} \theta(u_{(1)},v) u_{(0)} v S(u_{(1)}), \tag{48}$$

$$v\cdot \widetilde{\mathrm{ad}}(u) = \sum_{(u)_1} \theta(v,u_{(0)}) S(u_0) v u_{(1)}. \tag{49}$$

It is easy to check that these define left and right actions of U on U.

The bilinear form on U defined above is *invariant*, in that we have the following.

Proposition IV.7: For $u,v,v' \in U$,

$$\langle \mathrm{ad}(u)\cdot v | v' \rangle = \langle v | v' \cdot \widetilde{\mathrm{ad}}(u) \rangle \theta(u,v)\,\theta(u,v'). \tag{50}$$

Proof: It suffices to check the formula for $u = q^{h''} (h'' \in P^\vee)$, $e_{i,k}$, $f_{i,k}$ ($i \in I$, $k = 1,2,\ldots,m_i$) and for $v = xq^h S(y)$ and $v' = y' q^{h'} S(x')$ with $x \in U^+_\beta$, $x' \in U^+_{\beta'}$, $y \in U^-_{-\gamma}$, $y' \in U^-_{-\gamma'}$ $(\beta,\beta',\gamma,\gamma' \in Q^+)$. Since the case $u = f_{i,k}$ is similar to the case $u = e_{i,k}$, we will omit the case $u = f_{i,k}$.

(i) $u = q^{h''}$. The left-hand side is

$$\langle \mathrm{ad}(u)\cdot v | v' \rangle = \langle q^{h''} v q^{-h''} | v' \rangle = q^{(\beta-\gamma)(h'')} \langle v | v' \rangle,$$

and the right-hand side is

$$\langle v | v' \cdot \widetilde{\mathrm{ad}}(u) \rangle = \langle v | q^{-h''} v' q^{h''} \rangle = q^{(\gamma'-\beta')(h'')} \langle v | v' \rangle.$$

Since $\langle v | v' \rangle \neq 0$ only when $\beta = \gamma'$ and $\beta' = \gamma$, we are done.

(ii) $u = e_{i,k}$. Applying Lemma IV.4, we obtain

$$\begin{aligned} \mathrm{ad}(u)\cdot v &= e_{i,k} x q^h S(y) + \theta(\alpha_i,\beta) q^{(\alpha_i|\beta)} x K_i q^h S(e_{i,k} y) \\ &= e_{i,k} x q^h S(y) + \theta(\alpha_i,\beta) q^{(\alpha_i|\beta)} \sum_{(y)_2} \{A - B + C\}, \end{aligned}$$

where

$$A = (e_{i,k}|y_{(0)})(1|S(y_{(2)})) x K_i q^h S(y_{(1)}),$$

$$B = q^{\alpha_i(h)} (K_i|y_{(0)})(1|S(y_{(2)})) x e_{i,k} q^h S(y_{(1)}),$$

$$C = \theta(e_{i,k},y_{(1)})(K_i|y_{(0)})(e_{i,k}|S(y_{(2)})) x q^h S(y_{(1)}),$$

and

$$\begin{aligned} v'\cdot \widetilde{\mathrm{ad}}(u) &= -\theta(\beta'-\gamma',\alpha_i) q^{(\gamma'-\alpha_i|\alpha_i)} e_{i,k} y' K_i^{-1} q^{h'} S(x') - \theta(\beta',\alpha_i) q^{(\gamma'-\beta'|\alpha_i)} y' q^{h'} S(e_{i,k} x') \\ &= -\theta(\beta'-\gamma',\alpha_i) q^{(\gamma'-\alpha_i|\alpha_i)} \sum_{(y')_2} \{A' - B' + C'\} - \theta(\beta',\alpha_i) q^{(\gamma'-\beta'|\alpha_i)} y' q^{h'} S(e_{i,k} x'), \end{aligned}$$

where

$$A' = (e_{i,k}|y'_{(0)})(1|S(y'_{(2)}))y'_{(1)}K_i^{-1}q^{h'}S(x'),$$

$$B' = \theta(\alpha_i,\beta'-\gamma')q^{(\alpha_i|\alpha_i)}q^{-\alpha_i(h')}(K_i|y'_{(0)})(1|S(y'_{(2)}))y'_{(1)}q^{h'}S(x'e_{i,k}),$$

$$C' = \theta(e_{i,k},y'_{(1)})(K_i|y'_{(0)})(e_{i,k}|S(y'_{(2)}))y'_{(1)}q^{h'}S(x').$$

There are only two cases to consider: (i) $\gamma'=\beta+\alpha_i$ and $\gamma=\beta'$; (ii) $\gamma'=\beta$ and $\gamma=\beta'+\alpha_i$.

Since the latter case is similar to the former, we will only check the first case. Assume $\gamma'=\beta+\alpha_i$ and $\gamma=\beta'$. Then, in order to have $B\neq 0$, we must have $y_{(0)},y_{(2)}\in U^0$ and $y_{(1)}\in U^-_{-\gamma'}$. Similarly, $A'\neq 0$ implies $y'_{(0)}\in U^{\leqslant 0}_{-\alpha_i}$, $y'_{(1)}\in U^{\leqslant 0}_{-\beta}$, and $y'_{(2)}\in U^0$. In this case, we get $y'_{(1)}=\tilde{y}'_1K_i^{-1}$ for some $\tilde{y}'_1\in U^-_{-\beta}$. Also, $C'\neq 0$ implies $y'_{(0)}\in U^0$, $y'_{(1)}\in U^{\leqslant 0}_{-\beta}$, and $y'_{(2)}\in U^{\leqslant 0}_{-\alpha_i}$. In this case, we have $y'_{(2)}=\tilde{y}'_2K^{-1}_{\gamma'-\alpha_i}$ for some $\tilde{y}'_2\in U^-_{-\alpha_i}$. We need to prepare one more fact. Using Proposition IV.2, we obtain the following formula:

$$(x_1x_2x_3|y)=\sum_{(y)_2}\theta(x_1x_2,x_3)\,\theta(x_1x_2,y_{(0)})\,\theta(x_1,x_2)\,\theta(x_1,y_{(1)})(x_3|y_{(0)})(x_2|y_{(1)})(x_1|y_{(2)}),$$

for any $x_i\in U^+$ $(i=1,2,3)$ and $y\in U^-$. From this formula, we get

$$(x'|y)=(x'K_i|y)=\sum_{(y)_2}(K_i|y_{(0)})(x'|y_{(1)})(1|y_{(2)}),$$

$$(xe_{i,k}|y')=\sum_{(y')_2}\theta(x,e_{i,k})\,\theta(x,y'_{(0)})(e_{i,k}|y'_{(0)})(x|y'_{(1)})(1|y'_{(2)}),$$

$$(e_{i,k}x|y')=(e_{i,k}xK_i|y')=\sum_{(y')_2}\theta(e_{i,k},x)\,\theta(e_{i,k},y'_{(1)})(K_i|y'_{(0)})(x|y'_{(1)})(e_{i,k}|y'_{(2)}).$$

Now, we obtain

$$\begin{aligned}\langle \mathrm{ad}(u)\cdot v|v'\rangle &= \langle e_{i,k}xq^hS(y)|y'q^{h'}S(x')\rangle\\ &\quad -\theta(\alpha_i,\beta)q^{(\alpha_i|\beta)}q^{\alpha_i(h)}\sum_{(y)_2}(K_i|y_{(0)})\langle xe_{i,k}q^hS(y_{(1)})|y'q^{h'}S(x')\rangle\\ &=\theta(\gamma,\gamma'-\beta')q^{-(h|h')/2}\\ &\quad\times\left\{(e_{i,k}x|y')(x'|y)-\theta(\alpha_i,\beta)q^{(\alpha_i|\beta)}q^{\alpha_i(h)}\sum_{(y)_2}(K_i|y_{(0)})(xe_{i,k}|y')(x'|y_{(1)})\right\}\\ &=\theta(\gamma,\gamma'-\beta')q^{-(h|h')/2}\sum_{(y)_2,(y')_2}(K_i|y_{(0)})(x'|y_{(1)})(x|y'_{(1)})\\ &\quad\times\{(K_i|y'_{(0)})(e_{i,k}|y'_{(2)})-\theta(x,y'_{(0)})q^{(\alpha_i|\beta)}q^{\alpha_i(h)}(e_{i,k}|y'_{(0)})\}\end{aligned}$$

and

$$\langle v|v'\cdot\widetilde{\mathrm{ad}}(u)\rangle = -\theta(\beta'-\gamma',\alpha_i)q^{(\gamma'-\alpha_i|\alpha_i)}\sum_{(y')_2}\Big\{(e_{i,k}|y'_{(0)})\langle xq^hS(y)|y'_{(1)}K_i^{-1}q^{h'}S(x')\rangle$$

$$+\theta(e_{i,k},y'_{(1)})(K_i|y'_{(0)})(e_{i,k}|S(y'_{(2)}))\langle xq^hS(y)|y'_{(1)}q^{h'}S(x')\rangle\Big\}$$

$$=\theta(\beta'-\gamma',\alpha_i)\,\theta(\gamma,\beta-\beta')q^{-(h|h')/2}q^{(\beta|\alpha_i)}\sum_{(y')_2}\Big\{\theta(e_{i,k},y'_{(1)})(K_i|y'_{(0)})$$

$$\times(e_{i,k}K_i^{-1}|\tilde{y}'_2K^{-1}_{\gamma'-\alpha_i})(x|y'_{(1)})(x'|y)-(e_{i,k}|y'_{(0)})(x|\tilde{y}'_1)(x'|y)q^{\alpha_i(h)}\Big\},$$

$$=\theta(\beta'-\gamma',\alpha_i)\,\theta(\gamma,\beta-\beta')q^{-(h|h')/2}\sum_{(y)_2,(y')_2}(K_i|y_{(0)})(x'|y_{(1)})(x|y'_{(1)})$$

$$\times\Big\{\theta(e_{i,k},y'_{(1)})(K_i|y'_{(0)})(e_{i,k}|y'_{(2)})-q^{(\alpha_i|\beta)}q^{\alpha_i(h)}(e_{i,k}|y'_{(0)})\Big\}.$$

Comparing these two, we get the desired formula. □

This proposition allows us to define a right U-module structure on some subalgebra of U^*. Define $\zeta:U\to U^*$ by setting

$$[\zeta(u)](v)=\langle v|u\rangle, \tag{51}$$

for $u,v\in U$. Here, the dual space on the right should be viewed as the set of linear maps from U to $\mathbf{C}(q^{1/2})$. For $\zeta(u)\in\zeta(U)$, $x\in U$, define $\zeta(u)\cdot x$ by,

$$[\zeta(u)\cdot x](v)=\theta(u,x)\,\theta(v,x)[\zeta(u)](\mathrm{ad}(x)\cdot v).$$

Proposition IV.7 allows us to check $\zeta(u)\cdot x=\zeta(u\cdot\widetilde{\mathrm{ad}}(x))$. So this gives a right U-module structure on $\zeta(U)$ and $\zeta:U\to\zeta(U)$ becomes a U-module homomorphism.

Proposition IV.8: The bilinear form $\langle\,|\,\rangle$ is nondegenerate. Hence, the map ζ is injective.

Proof: Let $u\in U^-_{-\alpha}U^0S(U^+_\beta)$ with $\langle v|u\rangle=0$ for all $v\in U^+_\alpha U^0S(U^-_{-\beta})$. It suffices to show $u=0$. For each $\gamma\in Q^+-\{0\}$, choose a basis $\{u_i^\gamma\}_i$ of U^+_γ. And let $\{v_i^\gamma\}_i$ be a basis of $U^-_{-\gamma}$ dual to $\{u_i^\gamma\}_i$ with respect to the nondegenerate bilinear form $(\,|\,)$. Notice that the elements $u_i^\alpha q^hS(v_j^\beta)$ with $h\in P^\vee$ and i, j going over appropriate indices, form a basis for $U^+_\alpha U^0U^-_{-\beta}$. Similarly, the elements $v_k^\alpha q^{h'}S(u_l^\beta)$ form a basis for $U^-_{-\alpha}U^0U^+_\beta$. Writing $u=\Sigma_{k,h,l}a_{k,h,l}v_k^\alpha q^{h'}S(u_l^\beta)$ with $a_{k,h,l}\in\mathbf{C}(q)$, and using

$$\langle u_i^\alpha q^hS(v_j^\beta)|v_k^\alpha q^{h'}S(u_l^\beta)\rangle=\delta_{i,k}\delta_{j,l}q^{-(h|h')/2}\theta(\beta,\alpha)\,\theta(\beta,\beta),$$

we arrive at

$$\sum_{h'\in P^\vee}a_{k,h',l}q^{-(h|h')/2}=0,$$

for each k, l, and $h\in P^\vee$. Now, each map $h\mapsto q^{-(h|h')/2}$ is a group homomorphism from $P^\vee$ to the multiplicative group $\mathbf{C}(q^{1/2})^\times$. Since $q^{1/2}$ is not a root of unity, distinct h' produces distinct homomorphisms. So, by Artin's Theorem on linear independence of characters, every $a_{k,h',l}=0$. We have $u=0$, as claimed. □

V. HARISH–CHANDRA HOMOMORPHISM

We denote the center of U by $\mathfrak{z}$. For each $i\in I$ with $a_{i,i}\neq 0$, define the *simple reflection* $r_i\in\mathrm{GL}(\mathfrak{h})$ by

$$r_i(h)=h-\frac{2}{a_{i,i}}\alpha_i(h)h_i,$$

and let $\widetilde{W}=\langle r_i|i\in I,a_{i,i}\neq 0\rangle\subset \mathrm{GL}(\mathfrak{h})$. Let $(U^0)^{\widetilde{W}}$ be the subspace of U^0 consisting of the elements $\Sigma_{h\in P^\vee}c_hq^h(c_h\in\mathbf{C}(q))$ such that $c_h\neq 0$ implies $w(h)\in P^\vee$ and $c_{w(h)}=c_h$ for any $w\in\widetilde{W}$.

We define an algebra automorphism $\phi:U^0\to U^0$ by setting $\phi(q^h)=q^{-\rho(h)}q^h$ for $h\in P^\vee$. The *Harish–Chandra homomorphism* $\xi:\mathfrak{z}\to U^0$ is the restriction to $\mathfrak{z}$ of the map

$$U\xrightarrow{\sim}U^-\otimes U^0\otimes U^+\xrightarrow{\epsilon\otimes 1\otimes\epsilon}U^0\xrightarrow{\phi}U^0.$$

For later use, we define the algebra homomorphism $\chi_\lambda:U^0\to\mathbf{C}(q)$ for each $\lambda\in P^+$ by $\chi_\lambda(q^h)=q^{\lambda(h)}$.

Proposition V.1: (a) *ξ is an algebra homomorphism*. (b) *ξ is injective*.

Proof: We will just prove (b). Let $z\in\mathfrak{z}$ be such that $\xi(z)=0$. Writing $z=\Sigma_{\beta\in Q^+}z_\beta$ with $z_\beta\in U^-_{-\beta}U^0U^+_\beta$, we see that $z_0=0$. Fix any $\beta\in Q^+$ minimal with the property that $z_\beta\neq 0$. Also choose basis $\{y_r\}_r$ and $\{x_s\}_s$ of $U^-_{-\beta}$ and U^+_β, respectively. We may write $z_\beta=\Sigma_{r,s}y_ru_{r,s}x_s$ for some $u_{r,s}\in U^0$. Then

$$0=e_{i,k}z-ze_{i,k}=\sum_{\gamma\neq\beta}(e_{i,k}z_\gamma-z_\gamma e_{i,k})+\sum_{r,s}(e_{i,k}y_r-\theta(\alpha_i,-\beta)y_re_{i,k})u_{r,s}x_s$$
$$+\sum_{r,s}y_r(\theta(\alpha_i,-\beta)e_{i,k}u_{r,s}x_s-u_{r,s}x_se_{i,k}).$$

Recalling the minimality of β, we see that only the second term on the right belongs to $U^-_{-(\gamma-\alpha_i)}U^0U^+_\gamma$. So we have $\Sigma_{r,s}(e_{i,k}y_r-\theta(\alpha_i,-\beta)y_re_{i,k})u_{r,s}x_s=0$. $\{x_s\}_s$ was chosen to be a basis, so $e_{i,k}\Sigma_ry_ru_{r,s}=\theta(\alpha_i,-\beta)\Sigma_ry_re_{i,k}u_{r,s}$ for all $i\in I$ and s.

Let $v_\lambda\in V^q(\lambda)$ denote the highest weight vector. Set $v=\Sigma_r\chi_\lambda(u_{r,s})y_rv_\lambda$. Then $e_{i,k}v=\theta(\alpha_i,-\beta)\Sigma_ry_re_{i,k}u_{r,s}v_\lambda=0$ for all $i\in I$, so the irreducibility of $V^q(\lambda)$ says $v=0$. Choosing an appropriate $\lambda\in P^+$, we may use Lemma III.2 and say $\Sigma_r\chi_\lambda(u_{r,s})y_r=0$. Again, $\{y_r\}_r$ was a basis, so $\chi_\lambda(u_{r,s})=0$ for all *r,s*. By choosing a suitable set of λ, we may show $u_{r,s}=0$ for all *r,s* and we have $z_\beta=0$. This contradicts the choice of z_β. □

We now try to close in the image of ξ. For each $J\subset\{(i,k)|i\in I,k=1,2,...,m_i\}$, let $U_J=\langle e_{i,k},f_{i,k},U^0|(i,k)\in J\rangle$. We denote by $\mathfrak{z}_J$ the center of the algebra U_J and by $\xi_J:\mathfrak{z}_J\to U^0$ the Harish–Chandra homomorphism for U_J. Let U_J^+ (respectively, U_J^-) be the subalgebra of U_J generated by $e_{i,k}$ (respectively, $f_{i,k}$) with $(i,k)\in J$, and set

$$R_J^+=\{x\in U^+|(x|U_J^-)=0\}=\{x\in U^+|(x|U_J^-U^0)=0\},\tag{52}$$

$$R_J^-=\{y\in U^-|(U_J^+|y)=0\}=\{y\in U^-|(U^0U_J^+|y)=0\},\tag{53}$$

$$R_J=R_J^-U^0U^++U^-U^0R_J^+.\tag{54}$$

The following may be proved as in Ref. 15.

Lemma V.2: (a) $U=U_J\oplus R_J$; (b) $U_JR_JU_J\subset R_J$; (c) $(\epsilon\otimes 1\otimes\epsilon)(R_J)=0$.

Define $U_r^0=\oplus_h\mathbf{C}(q)q^h$, where the direct sum is over all $h\in P^\vee$, satisfying (i) $\alpha_i(h)\in s_ia_{i,i}\mathbf{Z}$, if $i\in I^{\mathrm{ev}}$; (ii) $\alpha_i(h)\in 2s_ia_{i,i}\mathbf{Z}$, if $i\in I^{\mathrm{odd}}$ and $a_{i,i}\neq 0$.

Proposition V.3: (a) $\mathrm{Im}(\xi)\subset(U^0)^{\widetilde{W}}$. (b) $\mathrm{Im}(\xi)\subset U_r^0$. (c) $\mathrm{Im}(\xi)\subset\mathrm{Im}(\xi_J)$.

Proof: (a) Let $z\in\mathfrak{z}$. Let $v_\lambda\in M^q(\lambda)$ be the highest weight vector. Then, $zv_\lambda=\chi_{\lambda+\rho}(\xi(z))v_\lambda$. Since z commutes with every element of U, z acts as $\chi_{\lambda+\rho}(\xi(z))$ on every element of $M^q(\lambda)$. Now, fix $i\in I$ such that $a_{i,i}\neq 0$. We may calculate

$$e_{i,k}f_{i,k}^n=\theta_{i,i}^n f_{i,k}^n e_{i,k}+\theta_{i,i}^{n-1}f_{i,k}^{n-1}\frac{1}{\xi_i}\left(\frac{1-\theta_{i,i}^{-n}q_i^{-na_{i,i}}}{1-\theta_{i,i}^{-1}q_i^{-a_{i,i}}}K_i-\frac{1-\theta_{i,i}^{n}q_i^{na_{i,i}}}{1-\theta_{i,i}q_i^{a_{i,i}}}K_i^{-1}\right).$$

So that, for each $\lambda\in P$ satisfying $n(\lambda):=(2/a_{i,i})\lambda(h_i)\in\mathbf{Z}_{\geqslant 0}$, we can check that $f_{i,k}^{n(\lambda)+1}v_\lambda$ is a highest weight vector. Its weight is

$$\lambda-\left(\frac{2}{a_{i,i}}\lambda(h_i)+1\right)\alpha_i=\lambda-\frac{2}{a_{i,i}}(\lambda+\rho)(h_i)\alpha_i=r_i(\lambda+\rho)-\rho.$$

The argument at the beginning of this proof applies to any highest weight vector, and we have

$$\chi_{\lambda+\rho}(\xi(z))=\chi_{r_{i(\lambda+\rho)}}(\xi(z)),$$

under the condition $(2/a_{i,i})\lambda(h_i)\in\mathbf{Z}$. Checking $\chi_{r_i\mu}(q^h)=\chi_\mu(r_iq^h)$, for any $\mu\in\mathfrak{h}^*$, the above may now be written as

$$\chi_{\lambda+\rho}(\xi(z)-r_i\xi(z))=0,$$

for every $\lambda\in P$ satisfying $\lambda(h_i)\in(a_{i,i}/2)\mathbf{Z}_{\geqslant 0}$. By choosing a suitable set of λ, we may show $\xi(z)=r_i\xi(z)$.

(b) Let $z=\Sigma_{\beta\in Q^+}z_\beta\in\mathfrak{z}$ with $z_\beta\in U_{-\beta}^-U^0U_\beta^+$. Set $x=\Sigma_{n=0}^\infty z_{n\alpha_i}$ and $y=z-x$. Then $z=x+y$ with $x\in U_{\{(i,1)\}}$ and $y\in R_{\{(i,1)\}}$. Looking at

$$0=e_{i,k}z-ze_{i,k}=(e_{i,k}x-xe_{i,k})+(e_{i,k}y-ye_{i,k}),$$

with Lemma V.2 in mind, we see that $x\in\mathfrak{z}_{\{(i,1)\}}$. By the results of Section VI.1, all of which may be obtained by direct calculation, we have (i) $z_0\in\langle K_i,q^h|\alpha_i(h)=0\rangle$ if $i\in I^{\mathrm{ev}}$; (ii) $z_0\in\langle K_i^2,q^h|\alpha_i(h)=0\rangle$ if $i\in I^{\mathrm{odd}}$ and $a_{i,i}\neq 0$.
The result follows.

(c) For $z\in\mathfrak{z}$, write $z=x+y$ with $x\in U_J$ and $y\in R_J$. As in the proof for (b), we may show $x\in\mathfrak{z}_J$. So we have $\xi(z)=\xi(x)+\xi(y)=\xi(x)=\xi_J(x)\in\mathrm{Im}(\xi_J)$. □

VI. THE CENTER OF $U_q(\mathfrak{g})$

VI.1. *Rank 1.* In this section, we list the center for the case when the index set is of size 1. All results may be obtained by direct calculation using induction after choosing a suitable basis of $U\cong U^-\otimes U^0\otimes U^+$.
If $a_{i,i}\neq 0$, $\theta_{i,i}=1$, define

$$C_i=f_{i,1}e_{i,1}+\frac{1}{\xi_i}\left(\frac{1}{1-q^{-s_ia_{i,i}}}K_i-\frac{1}{1-q^{s_ia_{i,i}}}K_i^{-1}\right).$$

If $a_{i,i}\neq 0$, $\theta_{i,i}=-1$, define

$$C_i=f_{i,1}^2e_{i,1}^2+\frac{1}{\xi_i}f_{i,1}\left(\frac{1-q^{s_ia_{i,i}}}{1+q^{s_ia_{i,i}}}K_i-\frac{1-q^{-s_ia_{i,i}}}{1+q^{-s_ia_{i,i}}}K_i^{-1}\right)e_{i,1}-\frac{1}{\xi_i^2}\left\{\frac{1}{(1+q^{-s_ia_{i,i}})^2}K_i^2+\frac{1}{(1+q^{s_ia_{i,i}})^2}K_i^{-2}\right\}.$$

If $h\in P^\vee$ satisfy $\alpha_i(h)\neq 0$, define

$$C_{ih}=f_{i,1}q^he_{i,1}+\frac{1}{\xi_i}\frac{1}{1-q^{-\alpha_i(h)}}q^h(K_i-K_i^{-1}).$$

Proposition VI.1: (a) If $J=\{(i,1)\}$ and $a_{i,i}\neq 0$, then $\mathfrak{z}_J=\langle C_i, q^h | h\in P^\vee, \alpha_i(h)=0\rangle$.
(b) *If* $J=\{(i,1)\}$, $a_{i,i}=0$, *and* $\theta_{i,i}=1$, *then* $\mathfrak{z}_J=\langle q^h | h\in P^\vee, \alpha_i(h)=0\rangle\subset U^0$.
(c) *If* $J=\{(i,1)\}$, $a_{i,i}=0$, *and* $\theta_{i,i}=-1$, *then* $\mathfrak{z}_J=\langle C_{ih}, q^{h'} | h,h'\in P^\vee, \alpha_i(h)\neq 0, \alpha_i(h')=0\rangle$,

VI.2. *Finite type*. In this section, we give a structure theorem for the center of $U_q(\mathfrak{g})$ when the Borcherds–Cartan matrix is of finite type. We take the Borcherds–Cartan matrix to be of finite type throughout this section. To simplify arguments, we redefine

$$P^\vee=\bigoplus_{i\in I}\mathbf{Z}h_i,$$

$$\mathfrak{h}=\bigoplus_{i\in I}\mathbf{C}h_i,$$

for this section. Notice that the bilinear form $(\,|\,)$ is still nondegenerate on the redefined $\mathfrak{h}$.

The irreducible highest weight module has a natural grading,

$$V^q(\lambda)=\bigoplus_{\alpha\in Q^+}V^q(\lambda)_{\lambda-\alpha}.$$

Define a map $\eta\in\mathrm{End}(V^q(\lambda))$ by setting $\eta(v)=\theta(\alpha,\alpha)v$ for $v\in V^q(\lambda)_{\lambda-\alpha}$. When the Borcherds–Cartan matrix A is of finite type, it is known (Ref. 20) that the irreducible highest weight module $V(\lambda)$ over $\mathfrak{g}(A)$ is finite dimensional for $\lambda\in P^+$. Since the classical limit (Ref. 6) of $V^q(\lambda)$ is $V(\lambda)$, $V^q(\lambda)$ is also of finite dimension when $\lambda\in P^+$. So we may define the *supertrace* for $x\in U_q(\mathfrak{g})$ acting on $V^q(\lambda)$ by

$$\mathrm{str}(x;V^q(\lambda))=\mathrm{tr}(\eta\circ x;V^q(\lambda)). \tag{55}$$

For homogeneous elements $x,y\in U$, we can easily check

$$\mathrm{str}(xy)=\theta(x,y)\mathrm{str}(yx). \tag{56}$$

Lemma VI.2: $u\in\mathfrak{z}$ *if and only if* $u\cdot\widetilde{ad}(x)=\epsilon(x)u$ *for all* $x\in U$.

Proof: Let $u\in\mathfrak{z}$. Then, $u\in U_0$ and

$$u\cdot\widetilde{\mathrm{ad}}(x)=\sum_{(x)_1}S(x_{(0)})ux_{(1)}=u\sum_{(x)_1}S(x_{(0)})x_{(1)}=\epsilon(x)u.$$

Conversely, if $u\cdot\widetilde{\mathrm{ad}}(x)=\epsilon(x)u$ for all $x\in U$,

$$q^{-h}uq^h=u\cdot\widetilde{\mathrm{ad}}(x)=\epsilon(q^h)u=u.$$

So $u\in U_0$, and we have

$$0=\epsilon(e_{i,k})u=u\cdot\widetilde{\mathrm{ad}}(e_{i,k})=-K_i^{-1}e_{i,k}u+K_i^{-1}ue_{i,k}.$$

This shows $e_{i,k}u=ue_{i,k}$. We may similarly show $f_{i,k}u=uf_{i,k}$, and hence $u\in\mathfrak{z}$. □

For each $\lambda\in P^+$, define $f_\lambda\in U^*$ by

$$f_\lambda(u)=\mathrm{str}(uK_{2\rho}^{-1};V^q(\lambda)). \tag{57}$$

Let $\nu:\mathfrak{h}\to\mathfrak{h}^*$ denote the isomorphism given by the nondegenerate symmetric bilinear form $(\,|\,)$. Define

$$\hat{Q}=\nu(P^\vee)=\bigoplus_{i\in I}\mathbf{Z}\frac{1}{s_i}\alpha_i.$$

Recall the map $\zeta:U\to U^*$ defined in (51).

Lemma VI.3: For $\lambda\in P^+$, $f_\lambda\in\mathrm{Im}(\zeta)$ *if and only if* $\lambda\in\frac{1}{2}\hat{Q}$.

Proof: From Proposition IV.8, we see that the image of ζ is the restricted dual of $U_q(\mathfrak{g})$. So

$$\mathrm{Im}(\zeta)=\Big(\bigoplus_{\beta\in Q^+}(U^-_{-\beta})^*\Big)\otimes\Big(\bigoplus_{\mu\in(1/2)\hat{Q}}\mathbf{C}(q)\chi_\mu\Big)\otimes\Big(\bigoplus_{\beta\in Q^+}(U^+_\beta)^*\Big),$$

under the identification $U\cong U^-\otimes U^0\otimes U^+$. The finite dimensionality of $V^q(\lambda)$ allows us to show $f_\lambda\in\mathrm{Im}(\zeta)$ if and only if $\lambda\in\frac{1}{2}\hat{Q}$. □

The next proposition gives elements of the center.

Proposition VI.4: For each $\lambda\in P^+\cap\frac{1}{2}\hat{Q}$, *we have* $z_\lambda:=\zeta^{-1}(f_\lambda)\in\mathfrak{z}$.

Proof: Recall from the theory of finite-dimensional simple Lie algebras that ρ may be written as a half sum of positive roots. Since the simple roots for the super case is identical to the nonsuper case, we have $2\rho\in Q^+$ in either case. Hence, in the notation previously given, $K_{2\rho}$ is a well-defined element of U^0. Using the fact that $K_{2\rho}^{-1}xK_{2\rho}=S^2(x)$ for any $x\in U$ and using the property of supertrace given by (56), we have, for any $u\in U$,

$$\begin{aligned}(f_\lambda\cdot x)(u)&=f_\lambda(\mathrm{ad}(x)\cdot u)\theta(u,x)\\&=\sum_{(x)_1}\mathrm{str}(x_{(0)}uS(x_{(1)})K_{2\rho}^{-1};V^q(\lambda))\theta(x_{(1)},u)\theta(u,x)\\&=\sum_{(x)_1}\mathrm{str}(uS(x_{(1)})K_{2\rho}^{-1}x_{(0)};V^q(\lambda))\theta(x_{(0)},x_{(1)})\\&=\mathrm{str}\left(uS\left(\sum_{(x)_1}S(x_{(0)})x_{(1)}\right)K_{2\rho}^{-1};V^q(\lambda)\right)=\epsilon(x)\mathrm{str}(uK_{2\rho}^{-1};V^q(\lambda))=\epsilon(x)f_\lambda(u).\end{aligned}$$

Thus $f_\lambda\cdot x=\epsilon(x)f_\lambda$. Recall from Proposition IV.8 that ζ is injective, and notice

$$f_\lambda\cdot x=\zeta(\zeta^{-1}(f_\lambda))\cdot x=\zeta(\zeta^{-1}(f_\lambda)\cdot\widetilde{\mathrm{ad}}(x)).$$

This shows $\zeta^{-1}(f_\lambda)\cdot\widetilde{\mathrm{ad}}(x)=\epsilon(x)\zeta^{-1}(f_\lambda)$. From Lemma VI.2, we get $\zeta^{-1}(f_\lambda)\in\mathfrak{z}$. □

We finally show that the above elements generate the whole center.

Theorem VI.5: *Suppose that the Borcherds–Cartan matrix* $A=(a_{i,j})_{i,j\in I}$ *is indecomposable and of a finite type. Then,* $\xi:\mathfrak{z}\to(U_r^0)^{\tilde{W}}$ *is an isomorphism.*

Proof: Let us calculate $\xi(z_\lambda)$. We extend the notation K_β previously introduced to $\beta\in\hat{Q}$ by setting $K_{(1/s_i)\alpha_i}=q^{h_i}$. We have the commutative diagram,

$$\begin{array}{ccc} & \overset{\zeta}{} & \\ U & \to & U^*\\ \epsilon\otimes\mathrm{id}\otimes\epsilon\downarrow & & \downarrow\\ U^0 & \to & (U^0)^*,\end{array}$$

where the right vertical arrow is the restriction map and the lower horizontal arrow is given by $K_\mu\mapsto\chi_{-\mu/2}$. Now, as maps on U^0,

$$f_\lambda=\sum_{\mu\leqslant\lambda}\theta(\lambda-\mu,\lambda-\mu)\dim(V(\lambda)_\mu)q^{-2(\rho|\mu)}\chi_\mu.$$

This shows

$$\xi(z_\lambda)=\sum_{\mu\leqslant\lambda}\theta(\lambda-\mu,\lambda-\mu)\dim(V(\lambda)_\mu)K_{-2\mu},\tag{58}$$

for $\lambda\in P^+\cap\frac{1}{2}\hat{Q}$.

Define $\hat{P}$ to be the set of elements $\mu\in\mathfrak{h}^*$ such that $\mu(h_i)\in\mathbf{Z}$ if $i\in I$ is even and $\mu(h_i)\in 2\mathbf{Z}$ if $i\in I$ is odd. Notice $P^+\subset\hat{P}$. We can now write

$$U_r^0=\bigoplus_{\mu\in 2\hat{P}\cap\hat{Q}}\mathbf{C}(q)K_\mu.$$

Action of the Weyl groups $\tilde{W}$ and W defined on $\mathfrak{h}$ and $\mathfrak{h}^*$ are compatible with the isomorphism ν. By Proposition VI.4, it suffices to show that the elements $\xi(z_\lambda)$ with $\lambda\in P^+\cap\frac{1}{2}\hat{Q}$ generate $(U_r^0)^W$.

Set $\overline{\mu}=\Sigma_{w\in W}K_{-w\mu}$ for any $\mu\in\hat{Q}$. We know the elements $\overline{\mu}$ with $\mu\in 2P^+\cap\hat{Q}$ generate $(U_r^0)^W$. Let us use induction to show that each of them belong to Im(ξ). The element $\overline{0}\in U_r^0$ is given by $\xi(z_\lambda)$, with $\lambda=0$. Choose any $\lambda\in 2P^+\cap\hat{Q}$. Then, $\frac{1}{2}\lambda\in P^+\cap\frac{1}{2}\hat{Q}$ so that $z_{(1/2)\lambda}$ is an element of the center. Recall that Im(ξ) is invariant under the action of W (Proposition V.3). Using $\dim V^q(\frac{1}{2}\lambda)_{(1/2)\lambda}=1$, we may rewrite (58) as

$$\xi(z_{(1/2)\lambda})=\overline{\lambda}+\sum_\mu n_\mu\overline{\mu},$$

with $n_\mu\in\mathbf{Z}$ and $\frac{1}{2}\mu$ running over some set of weights of $V^q(\frac{1}{2}\lambda)$. Since all $\mu<\lambda$, induction hypothesis show that each $\overline{\mu}$ belong to Im(ξ). Hence $\overline{\lambda}\in$ Im(ξ) and the induction step is complete. □

VI.3. *Other cases*. Let 2_i, 0_i, and Θ_i denote the fact that $a_{i,i}$ is, respectively, 2, 0, and negative. We will sometimes add a $\pm$ to these to reflect the sign of $\theta_{i,i}$. So, for example, 2_i^- implies that i is an odd real index. For $i,j\in I$, let us say $\odot_i$ is *connected directly* to $\odot_j$ if $a_{i,j}\neq 0$, where $\odot$ can be any one of 2, 0, or Θ. Here are some results for the case when $|J|=2$.

Lemma VI.6: Assume one of the following.

(a) $J=\{(i,1),(i,2)\}$, *with* Θ_i
(b) $J=\{(i,1),(j,1)\}$, *with* 0_i^+ *connected directly to* 0_j^-.
(c) $J=\{(i,1),(j,1)\}$, *with* 0_i^- *connected directly to* 0_j^-.
(d) $J=\{(i,1),(j,1)\}$, *with* Θ_i *connected directly to* 0_j^-.
(e) $J=\{(i,1),(j,1)\}$, *with* 2_i *connected directly to* Θ_j.
(f) $J=\{(i,1),(j,1)\}$, *with* Θ_i *connected directly to* Θ_j.

Then, $\mathfrak{z}_J\subset U^0$.

Proof: (a) and (f) may be proved as in Ref. 15, Proposition 4.5. And (e) may be proved as in Ref. 15, Proposition 4.6. (c) is proved by explicit calculation.

Let us prove (b) and (d) simultaneously. Let $z\in\mathfrak{z}_J$. Since it commutes with q^h for all $h\in P^\vee$, $z=\Sigma z_\beta$ with $z_\beta\in U_\beta^-\otimes U^0\otimes U_\beta^+$, where the sum is over all $\beta\in\mathbf{Z}_{\geqslant 0}\alpha_i\oplus\mathbf{Z}_{\geqslant 0}\alpha_j$. Let α be maximal among those $\beta\in\mathbf{Z}_{\geqslant 0}\alpha_i\oplus\mathbf{Z}_{\geqslant 0}\alpha_j$ for which z_β is nonzero and suppose $\alpha\neq 0$. Let $\{x_\mu\}$ and $\{y_\lambda\}$ be any bases of $(U_J^+)_\alpha$ and $(U_J^-)_{-\alpha}$, respectively. We can now write

$$z=\left(\sum_{\lambda,\mu,h}c_h^{\lambda,\mu}y_\lambda q^h x_\mu\right)+z'.$$

Recall Lemma IV.4 and notice

$$\Delta_2(e_{i,k})=e_{i,k}\otimes 1\otimes 1+K_i\otimes e_{i,k}\otimes 1+K_i\otimes K_i\otimes e_{i,k},$$

$$\Delta_2(y_\lambda)=1\otimes y_\lambda\otimes K_{-\alpha}+\text{``other terms.''}$$

This shows that the only part of $e_{i,k}z-ze_{i,k}$ belonging to the direct sum component $U^-_{-\alpha}\otimes U^0\otimes U^+_{\alpha+\alpha_i}$ is

$$\left(\sum_{\lambda,\mu,h} c_h^{\lambda,\mu}\theta(\alpha_i,-\alpha)y_\lambda e_{i,k}q^h x_\mu\right)-\left(\sum_{\lambda,\mu,h} c_h^{\lambda,\mu}y_\lambda q^h x_\mu e_{i,k}\right)$$

$$=\sum_{\lambda,h} y_\lambda q^h \sum_{\mu} c_h^{\lambda,\mu}(q^{-\alpha_i(h)}\theta(\alpha,\alpha_i)e_{i,k}x_\mu - x_\mu e_{i,k}).$$

Hence, for each $h\in P^\vee$ and λ,

$$\sum_{\mu} c_h^{\lambda,\mu}(q^{-\alpha_i(h)}\theta(\alpha,\alpha_i)e_{i,k}x_\mu - x_\mu e_{i,k})=0,$$

and the same statement with i replaced by j also holds. Now, $e_{j,1}^2=0$ is the only relation in U_J^+ for the case we are considering, so we may take an explicit set of monomials in $e_{i,1}$ and $e_{j,1}$ for the basis of U_α^+ and, using these, we can show that the two equations cannot be simultaneously true.□

Proposition VI.7: Assume that A is indecomposable. Suppose that energy 0_j^- *is connected directly to a* 0_i *or a* Θ_i. *If there is a nonempty subset J of* $\{(i,k)|i\in I,k=1,...,m_i\}$ *such that* $\mathfrak{z}_J\subset U^0$, *then* $\mathfrak{z}$ *is contained in* U^0.

Proof: Let $\bar{J}=\{i\in I|(i,k)\in J \text{ for some } k\}$. For $i\in I$, set

$$T_i=\bigoplus_{h\in P^\vee,\alpha_i(h)=0}\mathbf{C}(q)q^h.$$

We then have $\mathfrak{z}\cap U^0=\cap_{i\in I}T_i$ and similarly, $\mathfrak{z}_J\cap U^0=\cap_{i\in\bar{J}}T_i$. It suffices to show $\mathrm{Im}(\xi)\subset\cap_{i\in I}T_i$.

We already have $\mathrm{Im}(\xi)\subset\mathrm{Im}(\xi_J)\subset T_i$ for every $i\in\bar{J}$. Also, if 0_i^+, we have $\mathrm{Im}(\xi)\subset\mathrm{Im}(\xi_{\{(i,1)\}})\subset T_i$ by Proposition VI.1 (b). If 0_i^-, the conditions on the matrix show we may use Lemma VI.6 to write $\mathrm{Im}(\xi)\subset T_i$.

We now show that if $a_{j,j}\neq 0$ and $a_{i,j}\neq 0$, then $T_i\cap(U^0)^{\tilde{W}}\subset T_j$. Let $c=\Sigma c_h q^h\in T_i\cap(U^0)^{\tilde{W}}$. We must have $r_j c=c\in T_i$, so if $c_h\neq 0$, then $\alpha_i(h)=0$ and $\alpha_i(r_j h)=0$. But $\alpha_i(r_j h)=-(2/a_{j,j})a_{j,i}\alpha_j(h)$ so $\alpha_j(h)=0$. We have $c\in T_j$, as wanted.

Fix any $j\in I-(\bar{J}\cup\{i|a_{i,i}=0\})$. By the indecomposability of A, there exists a finite sequence $i=i_0,i_1,\ldots,i_n=j$ such that $i\in\bar{J}\cup\{i|a_{i,i}=0\}$, $i_k\notin\bar{J}\cup\{i|a_{i,i}=0\}$ for $k\geqslant 1$, and $a_{i_k,i_{k+1}}\neq 0$ for all k. What we have found above allows us to recursively show $\mathrm{Im}(\xi)\subset T_{i_k}$ and, in particular, $\mathrm{Im}(\xi)\subset T_j$. □

Proposition VI.8: Suppose three exists some finite $J\subset I$ *such that for every* $j\in J$, $a_{j,j}=2$ *and for which the corresponding submatrix* $A_J=(a_{i,j})_{i,j\in J}$ *is indecomposable and not of a finite type. Then,* $\mathfrak{z}_J\subset U^0$.

Proof: By Ref. 21, Proposition 4.9, we have $|W_J|=\infty$. Let $J'\subset J$ be such that $|W_{J'}|=\infty$ and $|W_{J''}|<\infty$ for all $J''\subsetneq J'$. We may use Proposition VI.7 if we can show $\mathfrak{z}_{J'}\subset U^0$. Hence, it suffices to show that if $h\in P^\vee$, $|W_{J'}(h)|<\infty$, then $\alpha_i(h)=0$, for all $i\in J'$. Give partial order to $\mathfrak{h}$ by setting $h_1\geqslant h_2$ if and only if $h_1-h_2\in(\Sigma_i\mathbf{Z}_{\geqslant 0}h_i)+(\Sigma_i\mathbf{Z}_{\geqslant 0}d_i)$. Let $h'\in W_{J'}(h)$ be maximal with respect to this order. Then, for each $i\in J'$, if $\alpha_i(h')<0$, then $h'<r_i h'$, so $\alpha_i(h')\geqslant 0$ for all $i\in J'$. Set $W_{h'}=\{w\in W|w(h')=h'\}$. By Ref. 21, Proposition 3.12(a), $W_{h'}=W_{J''}$ with $J''=\{i\in J|\alpha_i(h')=0\}$. If $J'',\subsetneq J'$, then $|W_{J'}(h')|=|W_{J'}/W_{h'}|=\infty$. Hence, we must have $J''=J'$ and $\alpha_i(h')=0$ for all $i\in J'$. $\{h'\}=W_{J'}(h')=W_{J'}(h)$. So $h=h'$ and $\alpha_i(h)=0$ for all $i\in J'$. □

Proposition VI.9: Let A be indecomposable, not of finite type, and $a_{i,i}=2$ *for all* $i\in I$. *Then,* $\mathfrak{z}\subset U^0$.

Proof: Suppose there exists some finite indecomposable submatrix that is not of finite type. Then we may use Proposition VI.7 and Proposition VI.8 to obtain the result.

If, to the contrary, every finite submatrix of A is of finite type, it must be one of the following types:

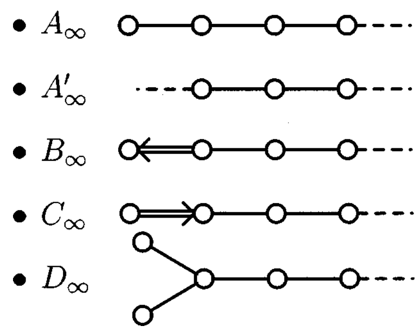

In all cases, with I naturally ordered, the matrix satisfies the following condition.

For each $i \in I$, there exists some $j > i$ such that $a_{i,j} \neq 0$, and $a_{i,k} = 0$ for $k > j$. $\qquad (*)$

Let $c = \Sigma_h c_h q^h \in \mathrm{Im}(\xi) \subset (U^0)^{\tilde{W}}$. Fix $h \in P^\vee$ for which $c_h \neq 0$. We aim to show $|\tilde{W}(h)| = \infty$ if $\alpha_j(h) \neq 0$ for some $j \in I$. We may assume that only finitely many $j \in I$ satisfy $\alpha_j(h) \neq 0$. Let $k \in I$ be the maximal of those so that $\alpha_j(h) = 0$ for all $j > k$ and $\alpha_k(h) \neq 0$. Set $i_0 = k$, and using property $(*)$, recursively choose i_n so that $i_{n+1} > i_n$ and $a_{i_n, i_{n+1}} \neq 0$. Put $h_0 = h$ and $h_{n+1} = r_{i_n} h_n$. Then, h_n cannot form a closed orbit and $|\tilde{W}(h)| = \infty$. Hence $\mathrm{Im}(\xi) \subset \oplus_{h \in P^\vee, \alpha_i(h) = 0} \mathbf{C} q^h$ and $\mathfrak{z} \subset U^0$. □

We can now collect all results and state the following.

Theorem VI.10: *Assume that the Borcherds–Cartan matrix* $A = (a_{i,j})_{i,j \in I}$ *is indecomposable and not of finite type. Suppose that every* 0_j^- *is connected directly to a* 0_i *or a* Θ_i. *Except for the case when* $|I| = 1$ *with* $m_i = 1$, *the center* $\mathfrak{z}$ *belongs to* U^0.

Proof: We apply Proposition VI.7 to each possible case.

If $|I| = 1$, the conditions imply either a 0_i^+ or a Θ_i with $m_i \geqslant 2$. These cases may be handled by Proposition VI.1 (b) and Lemma VI.6 (a), respectively.

Now suppose $|I| \geqslant 2$. Proposition VI.9 does away with the case when all $a_{i,i} = 2$. If it contains a 0_i^+, we may again use Proposition VI.1 (b). If it contains a 0_i^- but no 0_i^+, we use Lemma VI.6 (c), (d). The only other case is covered by Lemma VI.6 (e), (f). □

VII. THE UNIVERSAL *R* MATRIX

In this section, we find the universal R matrix for the quantum group $U_q(\mathfrak{g})$.

A Hopf superalgebra (or a colored Hopf algebra) H together with an element $\mathbf{R} \in H \otimes H$ is called a *quasitriangular Hopf superalgebra* if it satisfies the following: (a) $\mathbf{R}$ is invertible; (b) $\mathbf{R} \cdot \Delta(a) = \Delta'(a) \cdot \mathbf{R}$, for all $a \in H$; (c) $(\Delta \otimes 1)(\mathbf{R}) = \mathbf{R}_{13}\mathbf{R}_{23}$; (d) $(1 \otimes \Delta)(\mathbf{R}) = \mathbf{R}_{13}\mathbf{R}_{12}$, where $\Delta' = \sigma \circ \Delta$ with σ a colored transposition map, and where $\mathbf{R}_{ij}$ is an element of $H \otimes H \otimes H$, such that the ith and jth components are given by $\mathbf{R}$, and the remaining component is 1. The element $\mathbf{R}$ is called the *universal R matrix*. It satisfies the Yang–Baxter equation,

$$\mathbf{R}_{12}\mathbf{R}_{13}\mathbf{R}_{23} = \mathbf{R}_{23}\mathbf{R}_{13}\mathbf{R}_{12}. \qquad (59)$$

A Hopf superalgebra H together with an element $C \in H \otimes H$ and an algebra homomorphism $\Phi: H \otimes H \to H \otimes H$ is called a *pretriangular Hopf superalgebra* if it satisfies the following: (P1) C is invertible; (P2) $C \cdot \Delta(a) = \Phi(\Delta'(a)) \cdot C$, for all $a \in H$; (P3) $\Phi_{23} \circ \Phi_{13}(C_{12}) = C_{12}$; (P4) $\Phi_{12} \circ \Phi_{13}(C_{23}) = C_{23}$; (P5) $\Phi_{23}(C_{13}) \cdot C_{23} = (\Delta \otimes 1)(C)$; (P6) $\Phi_{12}(C_{13}) \cdot C_{12} = (1 \otimes \Delta)(C)$.

Under some conditions, it is possible to show that a pretriangular Hopf superalgebra becomes a quasitriangular Hopf superalgebra.

We set $U^{+,\beta} = \oplus_{\gamma \in Q^+, \gamma \not\leq \beta} U_\gamma^+$ for each $\beta \in Q^+$ and define the completion $\hat{U}$ of U by

$$\hat{U}=\lim_{\overleftarrow{\beta}} U/UU^{+,\beta}. \tag{60}$$

There is a natural embedding of U in $\hat{U}$ and there is a natural algebra structure on $\hat{U}$ that extends that of U under this embedding.

The completion of $U^{\otimes n}$ is similarly defined. We will write $\hat{U}\hat{\otimes}\hat{U}$ for the completion of $U\otimes U$.

Define an algebra automorphism $\Phi: U\otimes U\to U\otimes U$ by

$$\Phi(q^h\otimes q^{h'})=q^h\otimes q^{h'}, \tag{61}$$

$$\Phi(e_{i,k}\otimes 1)=e_{i,k}\otimes K_i, \quad \Phi(1\otimes e_{i,k})=K_i\otimes e_{i,k}, \tag{62}$$

$$\Phi(f_{i,k}\otimes 1)=f_{i,k}\otimes K_i^{-1}, \quad \Phi(1\otimes f_{i,k})=K_i^{-1}\otimes f_{i,k}. \tag{63}$$

It can be shown that Φ naturally extends to an algebra automorphism of $\hat{U}\hat{\otimes}\hat{U}$.

We denote by $C_\beta\in U_\beta^+\otimes U_{-\beta}^-$ the canonical element of the bilinear form $(\,|\,):U_\beta^+\times U_{-\beta}^-\to\mathbf{C}$. Define

$$C=\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_\beta|h_\beta)}(K_\beta^{-1}\otimes K_\beta)C_\beta\in\hat{U}\hat{\otimes}\hat{U}. \tag{64}$$

Lemma VII.1: (a) $C\cdot\Delta(q^h)=\Phi(\Delta'(q^h))\cdot C$ $(h\in P^\vee)$; (b) $(\Phi_{23}\circ\Phi_{13})(C_{12})=C_{12}$; (c) $(\Phi_{12}\circ\Phi_{13})(C_{23})=C_{23}$.

Proof: This is just a straightforward calculation. □

Lemma VII.2: Let $\beta\in Q^+$.

(a) $\sum_{\substack{\gamma,\delta\in Q^+\\ \gamma+\delta=\beta}}C_\gamma(K_\delta\otimes 1)(S\otimes 1)(C_\delta)=\delta_{\beta,0}$.

(b) $\sum_{\substack{\gamma,\delta\in Q^+\\ \gamma+\delta=\beta}}(K_\gamma\otimes 1)(S\otimes 1)(C_\gamma)C_\delta=\delta_{\beta,0}$.

(c) $\theta_{i,i}[1\otimes e_{i,k},C_{\beta+\alpha_i}]=C_\beta(e_{i,k}\otimes K_i^{-1})-(e_{i,k}\otimes K_i)C_\beta$.

(d) $\theta_{i,i}[f_{i,k}\otimes 1,C_{\beta+\alpha_i}]=C_\beta(K_i\otimes f_{i,k})-(K_i^{-1}\otimes f_{i,k})C_\beta$.

(e) $(\Delta\otimes 1)(C_\beta)=\sum_{\substack{\gamma,\delta\in Q^+\\ \gamma+\delta=\beta}}q^{-(h_\gamma|h_\delta)}(K_\delta\otimes 1\otimes 1)(C_\gamma)_{13}(C_\delta)_{23}$.

(f) $(1\otimes\Delta)(C_\beta)=\sum_{\substack{\gamma,\delta\in Q^+\\ \gamma+\delta=\beta}}q^{-(h_\gamma|h_\delta)}(1\otimes 1\otimes K_{-\delta})(C_\gamma)_{13}(C_\delta)_{12}$.

Proof: Here we show the proof for (a) only. Other cases may be proved in a similar spirit.

The case $\beta=0$ is trivial. So assume $\beta\in Q^+\setminus\{0\}$. The left-hand side is contained in $U_\beta^+\otimes U$, so by Theorem IV.6 it suffices to show that the application of $(\cdot|w)\otimes 1$ is zero for all $w\in U_{-\beta}^-$. We may write

$$\Delta(w)=\sum_{\substack{\gamma,\delta\in Q^+\\ \gamma+\delta=\beta}}w^{\delta,\gamma}(1\otimes K_\delta), \quad \text{with } w^{\delta,\gamma}\in U_{-\delta}^-\otimes U_{-\gamma}^-$$

and

$$w^{\delta,\gamma}=\sum_m w_{\delta,m}^{\delta,\gamma}\otimes w_{\gamma,m}^{\delta,\gamma}, \quad \text{with } w_{\delta,m}^{\delta,\gamma}\in U_{-\delta}^-, \quad w_{\gamma,m}^{\delta,\gamma}\in U_{-\gamma}^-.$$

We may also fix basis $\{x_r^\gamma\}_r$ and $\{y_r^\gamma\}_r$ of U_γ^+ and $U_{-\gamma}^-$, respectively, which are dual with respect to the bilinear form. Now

$$((\,|w)\otimes 1)(\mathrm{LHS}) = ((\,|w)\otimes 1)\left(\sum_{\gamma,\delta,r,s} \theta(y_r^\gamma, x_s^\delta) x_r^\gamma K_\delta S(x_s^\delta)\otimes y_r^\gamma y_s^\delta\right)$$

$$= \sum_{\gamma,\delta,r,s} \theta(x_r^\gamma, K_\delta S(x_s^\delta))\theta(-\gamma,\delta)(K_\delta S(x_s^\delta)\otimes x_r^\gamma|\Delta(w))y_r^\gamma y_s^\delta$$

$$= \sum_{\gamma,\delta,r,s,m} \theta(\delta,\gamma)(k_\delta S(x_s^\delta)|w_{\delta,m}^{\delta,\gamma})(x_r^\gamma|w_{\gamma,m}^{\delta,\gamma}K_{-\delta})y_r^\gamma y_s^\delta$$

$$= \sum_{\gamma,\delta,m} \theta(\delta,\gamma)\left(\sum_r (x_r^\gamma|w_{\gamma,m}^{\delta,\gamma}K_{-\delta})y_r^\gamma\right)\left(\sum_s (K_\delta S(x_s^\delta)\middle| w_{\delta,m}^{\delta,\gamma})y_s^\delta\right)$$

$$= \sum_{\gamma,\delta,m} \theta(\delta,\gamma)\left(\sum_r (x_r^\gamma|w_{\gamma,m}^{\delta,\gamma})y_r^\gamma\right)\left(\sum_s ((x_s^\delta|K_\delta^{-1}S^{-1}(w_{\delta,m}^{\delta,\gamma}))y_s^\delta)\right)$$

$$= \sum_{\gamma,\delta,m} \theta(\delta,\gamma)w_{\gamma,m}^{\delta,\gamma}K_\delta^{-1}S^{-1}(w_{\delta,m}^{\delta,\gamma})$$

$$= S^{-1}\left(\sum_{\gamma,\delta,m} w_{\delta,m}^{\delta,\gamma}S(w_{\gamma,m}^{\delta,\gamma}K_\delta^{-1})\right)$$

$$= (S^{-1}\circ m\circ(1\otimes S)\circ\Delta)(w) = \epsilon(w) = 0.$$

Hence the left-hand side is zero when $\beta \neq 0$. □

Proposition VII.3: Let

$$C' = \sum_{\beta\in Q^+} \theta(\beta,\beta)q^{(h_\beta|h_\beta)}(1\otimes K_\beta)(S\otimes 1)(C_\beta) \in \hat{U}\hat{\otimes}\hat{U}. \tag{65}$$

Then $CC' = C'C = 1$.

Proof:

$$CC' = \left(\sum_{\gamma\in Q^+} \theta(\gamma,\gamma)q^{(h_\gamma|h_\gamma)}(K_\gamma^{-1}\otimes K_\gamma)C_\gamma\right)\left(\sum_{\delta\in Q^+} \theta(\delta,\delta)q^{(h_\delta|h_\delta)}(1\otimes K_\delta)(S\otimes 1)(C_\delta)\right)$$

$$= \sum_{\beta\in Q^+} \sum_{\substack{\gamma+\delta=\beta\\ \gamma,\beta\in Q^+}} \theta(\beta,\beta)q^{(h_\gamma|h_\gamma)+(h_\delta|h_\delta)}(K_\gamma^{-1}\otimes K_\gamma)C_\gamma(1\otimes K_\delta)(S\otimes 1)(C_\delta)$$

$$= \sum_{\beta\in Q^+} \theta(\beta,\beta)q^{(h_\beta|h_\beta)}(K_\beta^{-1}\otimes K_\beta)\sum_{\substack{\gamma+\delta=\beta\\ \gamma,\beta\in Q^+}} C_\gamma(K_\delta\otimes 1)(S\otimes 1)(C_\delta).$$

We may now apply Lemma VII.2. The other part is done similary. □

Proposition VII.4: We have

$$C\cdot\Delta(e_{i,k}) = \Phi(\Delta'(e_{i,k}))\cdot C, \tag{66}$$

$$C\cdot\Delta(f_{i,k}) = \Phi(\Delta'(f_{i,k}))\cdot C. \tag{67}$$

Proof:

$$C\cdot\Delta(e_{i,k})=\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_{\beta+\alpha_i}|h_\beta)}(K_\beta^{-1}\otimes K_{\beta+\alpha_i})C_\beta(e_{i,k}\otimes K_i^{-1})$$
$$+\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_{\beta-\alpha_i}|h_\beta)}(K_{\beta-\alpha_i}^{-1}\otimes K_\beta)C_\beta(1\otimes e_{i,k}),$$

$$\Phi(\Delta'(e_{i,k}))\cdot C=\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_\beta|h_{\beta-\alpha_i})}(K_{\beta-\alpha_i}^{-1}\otimes K_\beta)(1\otimes e_{i,k})C_\beta$$
$$+\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_{\beta+\alpha_i}|h_\beta)}(K_\beta^{-1}\otimes K_{\beta+\alpha_i})(e_{i,k}\otimes K_i)C_\beta,$$

$$C\cdot\Delta(e_{i,k})-\Phi(\Delta'(e_{i,k}))\cdot C=\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_\beta|h_{\beta+\alpha_i})}(K_\beta^{-1}\otimes K_{\beta+\alpha_i})\{C_\beta(e_{i,k}\otimes K_i^{-1})$$
$$-(e_{i,k}\otimes K_i)C_\beta-\theta(\alpha_i,\alpha_i)[1\otimes e_{i,k},C_{\beta+\alpha_i}]\}.$$

We apply Lemma VII.2 to obtain the result. The other case is similar. □

Proposition VII.5: We have

$$\Phi_{23}(C_{13})\cdot C_{23}=(\Delta\otimes 1)(C),\tag{68}$$

$$\Phi_{12}(C_{13})\cdot C_{12}=(1\otimes\Delta)(C).\tag{69}$$

Proof:

$$\Phi_{23}(C_{13})=\sum_{\gamma\in Q^+}\theta(\gamma,\gamma)q^{(h_\gamma|h_\gamma)}(K_\gamma^{-1}\otimes K_\gamma^{-1}\otimes K_\gamma)(C_\gamma)_{13},$$

$$(\Phi_{23}(C_{13}))C_{23}=\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_\beta|h_\beta)}(K_\beta^{-1}\otimes K_\beta^{-1}\otimes K_\beta)$$
$$\times\sum_{\substack{\gamma,\delta\in Q^+\\ \gamma+\delta=\beta}}q^{-(h_\delta|h_\gamma)}(K_\delta\otimes 1\otimes 1)(C_\gamma)_{13}(C_\delta)_{23},$$

$$(\Delta\otimes 1)(C)=\sum_{\beta\in Q^+}\theta(\beta,\beta)q^{(h_\beta|h_\beta)}(K_\beta^{-1}\otimes K_\beta^{-1}\otimes K_\beta)(\Delta\otimes 1)(C_\beta).$$

The second case is done similarly. □

The propositions tell us that U is almost a pretriangular Hopf superalgebra.

Theorem VII.6: *The statements* $(P1)$ *and* $(P2)$ *hold in* $\hat{U}\hat{\otimes}\hat{U}$ *and the relations* $(P3)$–$(P6)$ *hold in* $\hat{U}\hat{\otimes}\hat{U}\hat{\otimes}\hat{U}$.

A weight module is *P-weighted* if all its weights belong to P. Notice $(P|P)\subset\mathbf{Z}$. This allows us to define $\mathfrak{Z}\in\mathrm{End}(V\otimes W)$ for any P-weighted $U_q(\mathfrak{g})$-modules V and W by setting,

$$\mathfrak{Z}(v\otimes w)=q^{(\mathrm{wt}(v)|\mathrm{wt}(w))}v\otimes w,\tag{70}$$

on homogeneous elements and extending by linearity. The map $\mathfrak{Z}$ is certainly invertible. There is a natural action of $U\otimes U$ on $V\otimes W$ and as endomorphisms on $V\otimes W$,

$$\Phi(a\otimes b)=\mathfrak{Z}\circ(a\otimes b)\circ\mathfrak{Z}^{-1},\tag{71}$$

for every $a\otimes b\in U\otimes U$.

Set $\mathbf{R}=\mathfrak{z}^{-1}C$. Then we finally have the following theorem.

Theorem VII.7: *Let $V_i(i=1,2,3)$ be P-weighted $U_q(\mathfrak{g})$ modules. As endomorphisms on $V_1\otimes V_2\otimes V_3$, when it can be defined,* $\mathbf{R}$ *satisfies the Yang–Baxter equation* (59).

Proof: From (P5) and Eq. (71), we have

$$\mathfrak{z}_{23}C_{13}\mathfrak{z}_{23}^{-1}C_{23}=(\Delta\otimes 1)(C), \tag{72}$$

$$\mathbf{R}_{13}\mathbf{R}_{23}=\mathfrak{z}_{13}^{-1}\mathfrak{z}_{23}^{-1}(\Delta\otimes 1)(C). \tag{73}$$

Applying $\sigma\otimes 1$ to both sides of (P5) and working as above, we get

$$\mathbf{R}_{23}\mathbf{R}_{13}=\mathfrak{z}_{23}^{-1}\mathfrak{z}_{13}^{-1}(\Delta'\otimes 1)(C). \tag{74}$$

The use of (P2) shows

$$\mathbf{R}_{23}\mathbf{R}_{13}\mathbf{R}_{12}=\mathfrak{z}_{23}^{-1}\mathfrak{z}_{13}^{-1}((\Delta'\otimes 1)(C))\mathbf{R}_{12} \tag{75}$$

$$=\mathfrak{z}_{23}^{-1}\mathfrak{z}_{13}^{-1}((\Delta'\otimes 1)(C))(\mathfrak{z}^{-1}C\otimes 1) \tag{76}$$

$$=\mathfrak{z}_{23}^{-1}\mathfrak{z}_{13}^{-1}(\mathfrak{z}^{-1}C\otimes 1)(\Delta\otimes 1)(C). \tag{77}$$

Now, the $\mathfrak{z}_{ij}$ commute with each other and (P3) with (71) says C_{12} commutes with $\mathfrak{z}_{13}^{-1}\mathfrak{z}_{23}^{-1}$, so we may use (73) to write

$$\mathfrak{z}_{23}^{-1}\mathfrak{z}_{13}^{-1}(\mathfrak{z}^{-1}C\otimes 1)(\Delta\otimes 1)(C)=\mathbf{R}_{12}\mathbf{R}_{13}\mathbf{R}_{23}. \tag{78}$$

Putting things together, we have the result. □

ACKNOWLEDGMENTS

I would like to thank Professor Seok-Jin Kang and one of my colleagues, Kyu-Hwan Lee, for help throughout this work. Lee has contributed much in finding the structure theorem for the center of finite type cases. This work was supported in part by GARC-KOSEF at Seoul National University.

[1] V. G. Drinfel'd, ''Hopf algebras and the quantum Yang–Baxter equation,'' Sov. Math. Dokl. **32**, 254–258 (1985).
[2] M. Jimbo, ''A q-difference analogue of $U(\mathfrak{g})$ and the Yang-Baxter equation,'' Lett. Math. Phys. **10**, 63–69 (1985).
[3] R. E. Borcherds, ''Generalized Kac–Moody algebras,'' J. Alg. **115**, 501–512 (1988).
[4] R. E. Borcherds, ''Monstrous moonshine and monstrous Lie superalgebras,'' Invent. Math. **109**, 405–444 (1992).
[5] S.-J. Kang, ''Quantum deformations of generalized Kac–Moody algebras and their modules,'' J. Alg. **175**, 1041–1066 (1995).
[6] G. Benkart, S.-J. Kang, and D. J. Melville, ''Quantized enveloping algebras for Borcherds superalgebras,'' Trans. Am. Math. Soc. **350**, 3297–3319 (1998).
[7] P. Baumann, ''On the center of quantized enveloping algebras,'' J. Alg. **203**, 244–206 (1998).
[8] V. G. Drinfel'd, ''Almost cocommutative Hopf algebras,'' Leningrad Math. J. **1**, 321–342 (1990).
[9] A. Joseph and G. Letzter, ''Separation of variables for quantized enveloping algebras,'' Am. J. Math. **116**, 127–177 (1994).
[10] N. Yu. Reshetikhin, ''Quasitriangle Hopf algebras and invariants of links,'' Leningrad Math. J. **1**, 291–513 (1990).
[11] N. Yu. Reshetikhin, L. A. Takhtadzhyan, and L. D. Faddeev, ''Quantization of Lie groups and Lie algebras,'' Leningrad Math. J. **1**, 193–225 (1990).
[12] M. Rosso, ''Analogues de la forme Killing et du théorème d'Harish–Chandra pour les groupes quantiques,'' Ann. Sci. Ecole. Norm. Sup. **23**, 445–467 (1990).
[13] T. Tanisaki, ''Killing forms, Harish–Chandra isomorphisms, and universal R-matrices for quantum algebras,'' Infinite Analysis, Part A, B (Kyoto, 1991), pp. 941–961, Adv. Ser. Math. Phys. **16** (World Publishing, River Edge, NJ, 1992).
[14] P. I. Etingof, ''Central elements for quantum affine algebras and affine Macdonald's operators,'' Math. Res. Lett. **2**, 611–628 (1995).
[15] S.-J. Kang and T. Tanisaki, ''Universal R-matrices and the center of the quantum generalized Kac–Moody algebras,'' Hiroshima Math. J. **27**, 347–360 (1997).
[16] V. G. Drinfel'd, ''Quantum groups,'' in *Proceedings of the International Congress of Mathematicians*, Berkeley, CA, 1986 (American Mathematical Society, Providence, RI, 1987), pp. 798–820.

[17] M. D. Gould, R. B. Zhang, and A. J. Bracken, ''Quantum double construction for graded Hopf algebras,'' Bull. Austral. Math. Soc. **47**, 353–375 (1993).
[18] M. Miyamoto, ''A generalization of Borcherds algebra and denominator formula,'' J. Alg. **180**, 631–651 (1996).
[19] U. Ray, ''A character formula for generalized Kac–Moody superalgebras,'' J. Alg. **177**, 154–163 (1995).
[20] V. G. Kac, ''Lie superalgebras,'' Adv. Math. **26**, 8–96 (1977).
[21] V. G. Kac, *Infinite dimensional Lie Algebras*, 3rd ed. (Cambridge University Press, Cambridge, 1990).

Unitary representations of the quantum algebra $su_q(2)$ on a real two-dimensional sphere for $q \in \mathbb{R}^+$ or generic $q \in S^1$

M. Irac-Astaud[a)]
Laboratoire de Physique Théorique de la Matière Condensée, Université Paris VII, 2, place Jussieu, F-75251 Paris Cedex 05, France

C. Quesne[b)]
Physique Nucléaire Théorique et Physique Mathématique, Université Libre de Bruxelles, Campus de la Plaine CP229, Boulevard du Triomphe, B-1050 Brussels, Belgium

(Received 25 March 1998; accepted for publication 22 February 1999)

Some time ago, Rideau and Winternitz introduced a realization of the quantum algebra $su_q(2)$ on a real two-dimensional sphere, or a real plane, and constructed a basis for its representations in terms of q-special functions, which can be expressed in terms of q-Vilenkin functions, and are related to little q-Jacobi functions, q-spherical functions, and q-Legendre polynomials. In their study, the values of q were implicitly restricted to $q \in \mathbb{R}^+$. In the present paper, we extend their work to the case of generic values of $q \in S^1$ (i.e., q values different from a root of unity). In addition, we unitarize the representations for both types of q values, $q \in \mathbb{R}^+$ and generic $q \in S^1$, by determining some appropriate scalar products. From the latter, we deduce the orthonormality relations satisfied by the q-Vilenkin functions. © *1999 American Institute of Physics.* [S0022-2488(99)01806-X]

I. INTRODUCTION

As is well known, most special functions of mathematical physics admit extensions to a base q, which are called q-special functions.[1–3] In the same way as Lie algebras and their representations provide a unifying framework for the former, quantum algebras[4] are relevant to the study of the latter (see, e.g., Ref. 5 and references quoted therein).

Some time ago, Rideau and Winternitz[6] introduced a realization of the quantum algebra $su_q(2)$ on a real sphere S^2 (or, via a stereographic projection, on a real plane), and constructed a basis for its irreducible representations (irreps) in terms of some functions $\Psi^J_{MNq}(\theta,\phi) \propto P^J_{MNq}(\cos\theta)\exp(-i(M+N)\phi)$. The functions $P^J_{MNq}(\cos\theta)$ were called q-Vilenkin functions because, for $q=1$, they reduce to functions $P^J_{MN}(\cos\theta)$ introduced by Vilenkin,[7,8] and related to Jacobi polynomials.

Rideau and Winternitz did establish various interesting results for the q-Vilenkin functions, including their recursion relations, explicit expression, generating function, and symmetry relations. They also compared them with other q-special functions, such as q-hypergeometric series, little q-Jacobi functions, q-spherical functions, and q-Legendre polynomials. Recently, the latter polynomials were further studied by Schmidt along similar lines.[9]

The realization of $su_q(2)$ on S^2, introduced by Rideau and Winternitz, was used by one of the present authors (MIA) to set up $su_q(2)$-invariant Schrödinger equations in the usual framework of quantum mechanics.[10] The corresponding radial equations can be easily solved for the ''free'' $su_q(2)$-invariant particle,[10] as well as for the Coulomb[10] and oscillator[11] potentials.

Although not explicitly stated in Ref. 6, the values of the deformation parameter q, considered there, are restricted to $q \in \mathbb{R}^+$. Close examination indeed shows that the explicit form of the

[a)]Electronic mail: mici@ccr.jussieu.fr
[b)]Directeur de recherches FNRS; electronic mail: cquesne@ulb.ac.be

function $Q_{Jq}(\eta)$, $\eta\equiv\cot^2(\theta/2)$, entering the definition of the q-Vilenkin functions,[6] is not valid for half-integer J values, whenever q runs over the unit circle.

Though important both from the q-special function viewpoint, and from that of their applications in quantum mechanics, the question of the $\mathrm{su}_q(2)$ irrep unitarity was also left unsolved by Rideau and Winternitz. They only noticed[6] that their realization of $\mathrm{su}_q(2)$ on S^2 is not unitary with respect to the scalar product used to unitarize the corresponding realization of su(2), and that a new scalar product should therefore be determined to cope with this drawback.

The purpose of the present paper is twofold: first, to find a solution for $Q_{Jq}(\eta)$ for generic $q\in S^1$ (i.e., for q different from a root of unity), and second, to unitarize the representations for both $q\in\mathbb{R}^+$, and generic $q\in S^1$. As a consequence, the explicit orthonormality relations of the q-Vilenkin and related functions will be established.

In Sec. II, the representations of $\mathrm{su}_q(2)$ on S^2, derived by Rideau and Winternitz, are briefly reviewed. The function $Q_{Jq}(\eta)$ is determined in Sec. III. The unitarization of the representations is dealt with in Sec. IV. Section V contains the conclusion.

II. REPRESENTATIONS OF $\mathrm{su}_q(2)$ ON S^2

Let us consider functions $f(\theta,\phi)$ on a sphere S^2, defined by $x_0^2+y_0^2+z_0^2=\frac{1}{4}$. These functions can also be viewed as functions on a real plane, via the stereographic projection $x=x_0/(\frac{1}{2}-z_0)$, $y=y_0/(\frac{1}{2}-z_0)$. In terms of spherical coordinates on S^2 and polar ones on the plane, we have

$$x_0=\tfrac{1}{2}\sin\theta\cos\phi,\quad y_0=\tfrac{1}{2}\sin\theta\sin\phi,\quad z_0=\tfrac{1}{2}\cos\theta,$$

$$x=\rho\cos\phi,\quad y=\rho\sin\phi,\quad \rho=\cot\frac{\theta}{2}, \tag{2.1}$$

$$0\leqslant\theta\leqslant\pi,\quad 0\leqslant\phi<2\pi,\quad 0\leqslant\rho<\infty.$$

Instead of the real variables x and y, one can use complex ones

$$z=x+iy=\rho e^{i\phi},\quad \overline{z}=x-iy=\rho e^{-i\phi}. \tag{2.2}$$

Functions $f(\theta,\phi)$ on S^2 can thus be projected onto functions $f(\rho,\phi)$ on the real plane, or functions $f(z,\overline{z})$ of a complex variable and its conjugate.

The $\mathrm{su}_q(2)$ generators H_3, H_+, H_- satisfy the commutation relations[4]

$$[H_3,H_\pm]=\pm H_\pm,\quad [H_+,H_-]=[2H_3]_q\equiv\frac{q^{2H_3}-q^{-2H_3}}{q-q^{-1}}, \tag{2.3}$$

and the Hermiticity properties

$$H_3^\dagger=H_3,\quad H_\pm^\dagger=H_\mp, \tag{2.4}$$

where in Eq. (2.3), we assume $q=e^\tau\in\mathbb{R}^+$, or $q=e^{i\tau}\in S^1$ (but different from a root of unity). From H_3 and $H_\pm$, one can construct a Casimir operator

$$\mathcal{C}=H_+H_-+[H_3]_q[H_3-1]_q=H_-H_++[H_3]_q[H_3+1]_q, \tag{2.5}$$

such that $[\mathcal{C},H_3]=[\mathcal{C},H_\pm]=0$.

The generators H_3, H_+, H_- can be realized[6] by the following operators, acting on functions $f(z,\overline{z})$ or $f(\theta,\phi)$,

$$H_3=-z\partial_z+\overline{z}\partial_{\overline{z}}-N=i\partial_\phi-N,$$

$$H_+=-z^{-1}[T]_q q^{\overline{T}-(N/2)}-q^{T+(N/2)}\overline{z}[\overline{T}-N]_q, \tag{2.6}$$

$$H_- = z[T+N]_q q^{\overline{T}-N/2} + q^{T+(N/2)} \overline{z}^{-1} [\overline{T}]_q ,$$

where

$$T = z\partial_z = -\tfrac{1}{2}(\sin\theta\partial_\theta + i\partial_\phi), \quad \overline{T} = \overline{z}\partial_{\overline{z}} = -\tfrac{1}{2}(\sin\theta\partial_\theta - i\partial_\phi). \tag{2.7}$$

For future use, it is also convenient to write $H_\pm$ in terms of polar coordinates on the real plane as

$$H_\pm = \mp \frac{e^{\mp i\phi}}{q-q^{-1}} \left\{ \left(\rho + \frac{1}{\rho}\right) q^{\rho\partial_\rho \mp (N/2)} - \rho q^{\mp i\partial_\phi \pm (3N/2)} - \frac{1}{\rho} q^{\pm i\partial_\phi \mp (N/2)} \right\}. \tag{2.8}$$

Basis functions $\Psi^J_{MNq}(z,\overline{z})$ for the $(2J+1)$-dimensional irrep of $\mathrm{su}_q(2)$ satisfy the relations[4]

$$\begin{gathered} H_3 \Psi^J_{MNq} = M \Psi^J_{MNq}, \quad H_\pm \Psi^J_{MNq} = ([J\mp M]_q [J\pm M+1]_q)^{1/2} \Psi^J_{M\pm 1,Nq}, \\ \mathcal{C}\Psi^J_{MNq} = [J]_q [J+1]_q \Psi^J_{MNq}, \quad M = \{-J, -J+1, ..., J\}, \quad |N| \le J, \end{gathered} \tag{2.9}$$

where J, M and N are simultaneously integers or half-integers. Let us remark that, when $q \in S^1$, the existence of such a representation implies that the factorials do not vanish, hence that q is not a root of unity.

Following Rideau and Winternitz,[6] let us write $\Psi^J_{MNq}(z,\overline{z})$ as

$$\Psi^J_{MNq}(z,\overline{z}) = N^J_{MNq} Q_{Jq}(\eta) q^{-NM/2} R^J_{MNq}(\eta) \overline{z}^{M+N}, \quad \eta = z\overline{z}. \tag{2.10}$$

Here, N^J_{MNq} is a constant, which can be expressed as

$$\begin{gathered} N^J_{MNq} = C_{JNq} \left(\frac{[J+M]_q!}{[J-M]_q! [2J]_q!} \right)^{1/2}, \\ C_{JNq} = \frac{1}{\sqrt{2\pi}} \left(\frac{[J+N]_q! [2J+1]_q!}{[J-N]_q!} \right)^{1/2} \gamma(J,N,q), \end{gathered} \tag{2.11}$$

in terms of some yet undetermined normalization constant $\gamma(J,N,q)$, and q-factorials, defined by $[x]_q! \equiv [x]_q [x-1]_q \cdots [1]_q$ if $x \in \mathbb{N}^+$, $[0]_q! \equiv 1$, and $([x]_q!)^{-1} \equiv 0$ if $x \in \mathbb{N}^-$. Equation (2.10) also contains two functions of η, $Q_{Jq}(\eta)$ and $R^J_{MNq}(\eta)$. The latter is a polynomial, whose explicit form is given by

$$R^J_{MNq}(\eta) = [J-N]_q! [J-M]_q! \sum_k \frac{(-\eta)^k}{[k]_q! [J-M-k]_q! [J-N-k]_q! [M+N+k]_q!}, \tag{2.12}$$

the summation over k being restricted by the condition that all the factorials in the denominator be positive. The former is defined by the functional equation

$$Q_{Jq}(q^2\eta)(1+\eta) = Q_{Jq}(\eta)(1+q^{-2J}\eta), \tag{2.13}$$

whose solution, only determined up to an arbitrary multiplicative factor $f_{Jq}(\eta)$ such that

$$f_{Jq}(q^2\eta) = f_{Jq}(\eta), \tag{2.14}$$

will be discussed in detail for both $q \in \mathbb{R}^+$, and generic $q \in S^+$, in the next section.

In terms of spherical coordinates, Eq. (2.10) becomes[6]

$$\Psi^J_{MNq}(\theta,\phi) = C_{JNq} \left(\frac{[J-N]_q!}{[J+N]_q! [2J]_q!} \right)^{1/2} i^{-2J+M+N} q^{-NM/2} P^J_{MNq}(\cos\theta) e^{-i(M+N)\phi}, \tag{2.15}$$

where

$$P^J_{MNq}(\xi)=i^{2J-M-N}\left(\frac{[J+M]_q![J+N]_q!}{[J-M]_q![J-N]_q!}\right)^{1/2}\eta^{(M+N)/2}Q_{Jq}(\eta)R^J_{MNq}(\eta),$$
$$\xi=\cos\theta,\quad \eta=\frac{1+\xi}{1-\xi}=\cot^2\frac{\theta}{2}, \tag{2.16}$$

are q-Vilenkin functions. For integer J values, the functions $\Psi^J_{M0q}(\theta,\phi)$ are proportional to q-spherical harmonics, while $P_{Jq}(\xi)\equiv P^J_{M0q}(\xi)$ are q-analogs of Legendre polynomials.

In the $q\to1$ limit, the $\mathrm{su}_q(2)$ realization (2.6) goes over into the su(2) realization

$$H_3=-z\partial_z+\bar{z}\partial_{\bar{z}}-N,\quad H_+=-\partial_z-\bar{z}^2\partial_{\bar{z}}+N\bar{z},\quad H_-=z^2\partial_z+\partial_{\bar{z}}+Nz, \tag{2.17}$$

the constant $\gamma(J,N,q)$ into $\gamma(J,N,1)=1$, and the q-Vilenkin functions into ordinary ones $P^J_{MN}(\xi)$. The latter are given by Eq. (2.16), where $[x]_q\to x$, and $Q_{Jq}(\eta)\to Q_J(\eta)=(1+\eta)^{-J}$. The operators (2.17) satisfy Eq. (2.4), and the functions Ψ^J_{MN}, $J=|N|,|N|+1,\ldots$, $M=-J,-J+1,\ldots,J$, form an orthonormal set with respect to the scalar product

$$\langle\psi_1|\psi_2\rangle=2\int\frac{dz\,d\bar{z}}{(1+z\bar{z})^2}\overline{\psi_1(z,\bar{z})}\psi_2(z,\bar{z})=\frac{1}{2}\int_0^\pi d\theta\sin\theta\int_0^{2\pi}d\phi\,\overline{\psi_1(\theta,\phi)}\psi_2(\theta,\phi), \tag{2.18}$$

where the integral over z, $\bar{z}$ extends over the whole complex plane.

III. DETERMINATION OF $Q_{Jq}(\eta)$

Following Rideau and Winternitz,[6] as a solution of Eq. (2.13), we may consider the function

$$Q_{Jq}(\eta)={}_1\Phi_0(q^{2J};-;q^2,-q^{-2J}\eta)={}_1\Phi_0(q^{-2J};-;q^{-2},-q^{-2}\eta), \tag{3.1}$$

where ${}_1\Phi_0$ is a basic hypergeometric series in the notations of Ref. 3.

For $q\in\mathbb{R}^+$, use of the q-binomial theorem[3] leads to the expressions

$$Q_{Jq}(\eta)=\prod_{k=0}^{\infty}\frac{(1+q^{2k}\eta)}{(1+q^{-2J+2k}\eta)} \tag{3.2}$$

if $0<q<1$, and

$$Q_{Jq}(\eta)=\prod_{k=0}^{\infty}\frac{(1+q^{-2J-2k-2}\eta)}{(1+q^{-2k-2}\eta)} \tag{3.3}$$

if $q>1$. For integer J values, both expressions reduce to the inverse of a polynomial,

$$Q_{Jq}(\eta)=\prod_{k=0}^{J-1}\left(\frac{1}{1+\eta q^{-2J+2k}}\right), \tag{3.4}$$

whereas for half-integer J values, we are left with convergent infinite products.

For generic $q\in S^1$ and integer J values, Eq. (3.4) still remains a valid solution of Eq. (2.13). However, for half-integer J values, the infinite products contained in Eqs. (3.2) and (3.3), as well as other expressions of ${}_1\Phi_0$ in terms of infinite series or products, found in Refs. 1 and 3, are divergent. We therefore have to look for another solution to Eq. (2.13).

For such a purpose, let us linearize Eq. (2.13) into

$$K_{Jq}(q^2\eta)-K_{Jq}(\eta)=\ln\frac{1+q^{-2J}\eta}{1+\eta}, \tag{3.5}$$

by setting

$$K_{Jq}(\eta)=\ln Q_{Jq}(\eta). \tag{3.6}$$

In terms of the operator $X\equiv\eta\partial_\eta$, Eq. (3.5) can be rewritten as

$$(q^{2X}-1)K_{Jq}(\eta)=(q^{-2JX}-1)\ln(1+\eta). \tag{3.7}$$

Let us consider the difference equation

$$(q^X-q^{-X})L_q(\eta)\equiv L_q(q\eta)-L_q(q^{-1}\eta)=\ln(1+\eta). \tag{3.8}$$

If we are able to find a solution to the latter, then

$$K_{Jq}(\eta)\equiv q^{-X}(q^{-2JX}-1)L_q(\eta)=L_q(q^{-2J-1}\eta)-L_q(q^{-1}\eta) \tag{3.9}$$

will be a solution of Eq. (3.7).

We will now proceed to demonstrate the following.

Lemma III.1: For $0<\eta<\infty$, and $q=e^{i\tau}$ different from a root of unity, the function

$$L_q(\eta)=\frac{1}{2\pi i}\int_0^\infty\frac{dt}{t(1+t)}\ln(1+\eta t^{\tau/\pi}),\quad if\ \ 0<\tau<\pi, \tag{3.10}$$

$$L_q(\eta)=-\frac{1}{2\pi i}\int_0^\infty\frac{dt}{t(1+t)}\ln(1+\eta t^{-\tau/\pi}),\quad if -\pi<\tau<0, \tag{3.11}$$

is a solution of Eq. (3.8).

Proof: We note that if some function $L_q(\eta)$ is a solution of Eq. (3.8) for $q=e^{i\tau}$, $0<\tau<\pi$, then $-L_{q^{-1}}(\eta)$ is also a solution of the same. Hence, Eq. (3.11) directly results from Eq. (3.10). It is also a simple matter to show that the integral on the right-hand side of Eq. (3.10) is convergent. It therefore only remains to prove that the latter satisfies Eq. (3.8). For such a purpose, we have to separately consider the integral when η is replaced by $\eta e^{i\tau}$, or by $\eta e^{-i\tau}$.

Let us introduce a function $M(v)$ of a complex variable v, defined by

$$M(v)=[v(1+v)]^{-1}\ln(1+\eta e^{-i\tau}v^{\tau/\pi}), \tag{3.12}$$

where on the right-hand side, there appear two multivalued functions $v^{\tau/\pi}$, and $\ln w$, where $w=1+\eta e^{-i\tau}v^{\tau/\pi}$.

For the function $v^{\tau/\pi}$, let us choose a branch cut along the positive real axis, so that $v^{\tau/\pi}=|v^{\tau/\pi}|\exp(i\tau\alpha/\pi)$, where $v=|v|\exp(i\alpha)$, and $0<\alpha<2\pi$. On the two sides of such a cut, the argument of the logarithm on the right-hand side of Eq. (3.12) takes the values

$$w_+=1+\eta e^{-i\tau}|v|^{\tau/\pi}\quad \text{if } \operatorname{Re}v>0, \operatorname{Im}v=+0, \tag{3.13}$$

and

$$w_-=1+\eta e^{i\tau}|v|^{\tau/\pi}\quad \text{if } \operatorname{Re}v>0, \operatorname{Im}v=-0, \tag{3.14}$$

respectively.

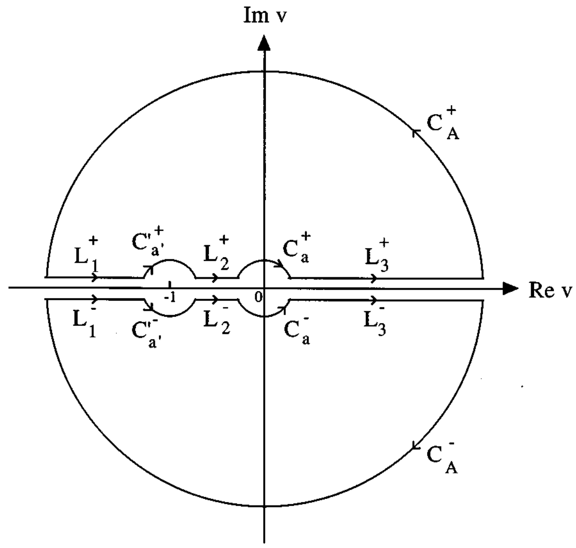

FIG. 1. Contours in the complex v plane used in the proof of Lemma III.1.

Considering next the function $\ln w$, it is easy to show that its branch point at $w=0$ and its branch cut along the negative real axis in the complex w plane cannot be reached within the truncated v plane, since the condition $\exp[i\tau(\alpha/\pi-1)]=-1$ cannot be fulfilled for $0<\tau<\pi$, and $0<\alpha<2\pi$.

Hence, when integrating the function $M(v)$ in the complex v plane, one should consider contours avoiding the branch point $v=0$, the branch cut $\mathrm{Re}\, v>0$, $\mathrm{Im}\, v=0$, and the simple pole at $v=-1$. Let us consider the two vanishing coutour integrals

$$\int_{\Gamma^+} M(v)dv = \int_{\Gamma^-} M(v)dv = 0, \tag{3.15}$$

where Γ^+ and Γ^- are the paths in the upper and lower halves of the v plane, displayed on Fig. 1. The former consists of the upper half C_A^+ of a large circle of radius A centered at the origin, and described in the counterclockwise sense, the upper halves C_a^+, $C'^+_{a'}$ of two small circles of radius a, a', centered at $v=0$ and $v=-1$, respectively, both described in the clockwise sense, and three straight lines L_1^+, L_2^+, L_3^+ lying just above the real axis, and going from $-A$ to $-1-a'$, from $-1+a'$ to $-a$, and from a to A, respectively. The latter path Γ^- is defined in a similar way.

Taking now Eqs. (3.13) and (3.14) into account, we obtain

$$2\pi i(L_q(q\eta)-L_q(q^{-1}\eta)) = \lim_{\substack{a\to 0\\ A\to\infty}} \left\{ \int_{L_3^-} M(v)dv - \int_{L_3^+} M(v)dv \right\}. \tag{3.16}$$

Owing to Eq. (3.15), each of the integrals on the right-hand side of Eq. (3.16) can be rewritten in terms of integrals along the other parts of the path Γ^- or Γ^+. Those along L_1^+ (resp. L_2^+) and L_1^- (resp. L_2^-) obviously cancel. Furthermore,

$$\lim_{A\to\infty} \left| \int_{C_A^+} M(v)dv - \int_{C_A^-} M(v)dv \right| \sim \lim_{A\to\infty} \frac{\ln A}{A} = 0, \tag{3.17}$$

and

$$\lim_{a\to 0}\left|\int_{C_a^+} M(v)dv - \int_{C_a^-} M(v)dv\right| \sim \lim_{a\to 0} a^{\tau/\pi} = 0, \tag{3.18}$$

so that

$$2\pi i(L_q(q\eta) - L_q(q^{-1}\eta)) = -\int_{C'_{a'}} M(v)dv = -2\pi i \operatorname{Res} M(-1) = 2\pi i \ln(1+\eta), \tag{3.19}$$

where $C'_{a'}$ denotes the circle of radius a' centered at $v=-1$, and described in the counterclockwise sense. Equation (3.19) completes the proof. ■

The results of the present section can be collected into the following.

Proposition III.2: The function $Q_{Jq}(\eta)$, appearing on the right-hand side of Eq. (2.10), is given by Eq. (3.4) for integer J values, and either $q \in \mathbb{R}^+$ or generic $q \in S^1$, and by Eqs. (3.2) and (3.3) for half-integer J values, and $q \in \mathbb{R}^+$. For half-integer J values, and generic $q \in S^1$, it can be expressed as

$$Q_{Jq}(\eta) = \exp\{L_q(q^{-2J-1}\eta) - L_q(q^{-1}\eta)\}, \tag{3.20}$$

where $L_q(\eta)$ admits the integral representation given in Lemma III.1.

IV. UNITARIZATION OF THE REPRESENTATIONS OF $su_q(2)$ ON S^2

In the present section, we will determine a new scalar product $\langle \psi_1 | \psi_2 \rangle_q$ that unitarizes the realization (2.6) of $su_q(2)$, and goes over into the old one $\langle \psi_1 | \psi_2 \rangle$, defined in Eq. (2.18), whenever $q \to 1$. For such a purpose, we shall first impose that Eq. (2.4) is satisfied by the realization (2.6) with respect to $\langle \psi_1 | \psi_2 \rangle_q$. The residual arbitrariness in the measure will then be lifted by demanding that $\langle \psi_1 | \psi_2 \rangle_q$ satisfies the usual properties of a scalar product.

We shall successively consider hereunder the cases where $q \in \mathbb{R}^+$, and generic $q \in S^1$.

A. The case where $q \in \mathbb{R}^+$

Let us make the following ansatz for $\langle \psi_1 | \psi_2 \rangle_q$,

$$\langle \psi_1 | \psi_2 \rangle_q = \int_0^\infty d\rho \int_0^{2\pi} d\phi \Big(\overline{A_q \psi_1(\rho,\phi,q)} f_1(\rho,q) q^{a_1 \rho \partial_\rho} \psi_2(\rho,\phi,q) + \overline{\psi_1(\rho,\phi,q)} f_2(\rho,q) q^{a_2 \rho \partial_\rho} A_q \psi_2(\rho,\phi,q) \Big), \tag{4.1}$$

in terms of the polar coordinates ρ, ϕ on the real plane, defined in Eq. (2.1). Here a_1, a_2, and $f_1(\rho,q)$, $f_2(\rho,q)$ are some yet undetermined constants and functions of the indicated arguments, respectively, and $A_q \equiv q^{-2q\partial_q}$ is the operator that changes q into q^{-1}, when acting on any function of q,

$$A_q \psi(\rho,\phi,q) = \psi(\rho,\phi,q^{-1}). \tag{4.2}$$

It is easy to check that

$$\langle \psi_1 | H_3 \psi_2 \rangle_q = \langle H_3 \psi_1 | \psi_2 \rangle_q \tag{4.3}$$

with respect to (4.1). Let us now impose the condition

$$\langle \psi_1 | H_+ \psi_2 \rangle_q = \langle H_- \psi_1 | \psi_2 \rangle_q . \tag{4.4}$$

By combining Eqs. (2.6) and (4.1), the left-hand side of this condition can be written as

$$\langle \psi_1 | H_+ \psi_2 \rangle_q = (q-q^{-1})^{-1} \int_0^\infty d\rho \int_0^{2\pi} d\phi \Big\{ \overline{\psi_1(\rho,\phi,q^{-1})} f_1(\rho,q) e^{-i\phi}$$
$$\times \left(-\left(q^{a_1}\rho + \frac{1}{q^{a_1}\rho} \right) q^{\rho\partial_\rho - (N/2)} + q^{a_1}\rho q^{-i\partial_\phi + (3N/2)} + \frac{1}{q^{a_1}\rho} q^{i\partial_\phi - (N/2)} \right)$$
$$\times \psi_2(q^{a_1}\rho,\phi,q) - \overline{\psi_1(\rho,\phi,q)} f_2(\rho,q) e^{-i\phi} \left(-\left(q^{a_2}\rho + \frac{1}{q^{a_2}\rho} \right) q^{-\rho\partial_\rho + (N/2)} \right.$$
$$\left. + q^{a_2}\rho q^{i\partial_\phi - (3N/2)} + \frac{1}{q^{a_2}\rho} q^{-i\partial_\phi + (N/2)} \right) \psi_2(q^{a_2}\rho,\phi,q^{-1}) \Big\}. \tag{4.5}$$

After integrating by parts and making some straightforward transformations, it becomes

$$\langle \psi_1 | H_+ \psi_2 \rangle_q = (q-q^{-1})^{-1} \int_0^\infty d\rho \int_0^{2\pi} d\phi e^{-i\phi} \Big\{ -\Big(\Big(q^{a_1-1}\rho + \frac{1}{q^{a_1-1}\rho} \Big)$$
$$\times f_1(q^{-1}\rho,q) q^{-\rho\partial_\rho - 1 - (N/2)} \overline{\psi_1(\rho,\phi,q^{-1})} \Big) \psi_2(q^{a_1}\rho,\phi,q)$$
$$+ \Big(\Big(\rho q^{i\partial_\phi + a_1 + 1 + (3N/2)} + \frac{1}{\rho} q^{-i\partial_\phi - a_1 - 1 - (N/2)} \Big) \overline{\psi_1(\rho,\phi,q^{-1})} \Big)$$
$$\times f_1(\rho,q) \psi_2(q^{a_1}\rho,\phi,q) + \Big(\Big(q^{a_2+1}\rho + \frac{1}{q^{a_2+1}\rho} \Big)$$
$$\times f_2(q\rho,q) q^{\rho\partial_\rho + 1 + (N/2)} \overline{\psi_1(\rho,\phi,q)} \Big) \psi_2(q^{a_2}\rho,\phi,q^{-1})$$
$$- \Big(\Big(\rho q^{-i\partial_\phi + a_2 - 1 - (3N/2)} + \frac{1}{\rho} q^{i\partial_\phi - a_2 + 1 + (N/2)} \Big) \overline{\psi_1(\rho,\phi,q)} \Big)$$
$$\times f_2(\rho,q) \psi_2(q^{a_2}\rho,\phi,q^{-1}) \Big\}. \tag{4.6}$$

On the other hand, for real q values the right-hand side of Eq. (4.4) can be written as

$$\langle H_- \psi_1 | \psi_2 \rangle_q = (q-q^{-1})^{-1} \int_0^\infty d\rho \int_0^{2\pi} d\phi e^{-i\phi}$$
$$\times \Big\{ \Big(\Big\{ -\Big(\rho + \frac{1}{\rho} \Big) q^{-\rho\partial_\rho - (N/2)} + \rho q^{i\partial_\phi + (3N/2)} + \frac{1}{\rho} q^{-i\partial_\phi - (N/2)} \Big\} \overline{\psi_1(\rho,\phi,q^{-1})} \Big)$$
$$\times f_1(\rho,q) \psi_2(q^{a_2}\rho,\phi,q) + \Big(\Big\{ \Big(\rho + \frac{1}{\rho} \Big) q^{\rho\partial_\rho + (N/2)} - \rho q^{-i\partial_\phi - (3N/2)}$$
$$- \frac{1}{\rho} q^{i\partial_\phi + (N/2)} \Big\} \overline{\psi_1(\rho,\phi,q)} \Big) f_2(\rho,q) \psi_2(q^{a_2}\rho,\phi,q^{-1}) \Big\}. \tag{4.7}$$

It now remains to equate the right-hand side of Eq. (4.6) with that of Eq. (4.7). Both of them being some linear combinations of four different types of terms, containing one of the operators $q^{-i\partial_\phi}$, $q^{i\partial_\phi}$, $q^{-\rho\partial_\rho}$, or $q^{\rho\partial_\rho}$, acting on some function, respectively, it is sufficient to separately equate such terms. The conditions on the first two classes of terms impose that

$$a_1 = -1, \quad a_2 = 1, \tag{4.8}$$

while those on the last two lead to the equations

$$q^{-1}\left(q^{-2}\rho+\frac{1}{q^{-2}\rho}\right)f_1(q^{-1}\rho,q)=\left(\rho+\frac{1}{\rho}\right)f_1(\rho,q),$$
$$q\left(q^{2}\rho+\frac{1}{q^{2}\rho}\right)f_2(q\rho,q)=\left(\rho+\frac{1}{\rho}\right)f_2(\rho,q), \tag{4.9}$$

whose solutions are given by

$$f_1(\rho,q)=\frac{B_1(q)q^{-1}\rho}{(1+\rho^2)(1+q^{-2}\rho^2)},\quad f_2(\rho,q)=\frac{B_2(q)q\rho}{(1+\rho^2)(1+q^{2}\rho^2)}, \tag{4.10}$$

in terms of two undetermined constants $B_1(q)$, and $B_2(q)$.

Let us now further restrict the sesquilinear form (4.1), where substitutions (4.8) and (4.10) have been made, by imposing that it is Hermitian, i.e.,

$$\overline{\langle\psi_1|\psi_2\rangle_q}=\langle\psi_2|\psi_1\rangle_q. \tag{4.11}$$

By a straightforward calculation, similar to that carried out for condition (4.4), it can be shown that Eq. (4.11) leads to the relation

$$B_2(q)=\overline{B_1(q)}. \tag{4.12}$$

As a consequence, there only remains a single undetermined constant $B(q)\equiv B_1(q)$ in Eq. (4.1). At this stage, it is important to notice that had we only considered a single term, instead of two, in Eq. (4.1), it would have been impossible to fulfill condition (4.11).

In addition, we remark that Eqs. (4.4) and (4.11) imply that

$$\langle\psi_1|H_-\psi_2\rangle_q=\langle H_+\psi_1|\psi_2\rangle_q. \tag{4.13}$$

Hence, all the Hermiticity conditions (2.4) on the $\mathrm{su}_q(2)$ generators are satisfied by the form defined in Eqs. (4.1), (4.8), (4.10), and (4.12). The functions $\Psi^J_{MNq}(z,\overline{z})$, defined in Eq. (2.10), and corresponding to a fixed N value, but different J and/or M values, are therefore orthogonal with respect to such a form.

To make $\langle\psi_1|\psi_2\rangle_q$ into a scalar product, it only remains to impose that it is a positive definite form. Since we also want that in the resulting Hilbert space, the functions Ψ^J_{MNq} with given J and N values, and $M=-J,\ -J+1,...,J,$ form an orthonormal basis for the $\mathrm{su}_q(2)$ irrep characterized by J, a condition that combines both requirements is

$$\langle\Psi^J_{MNq}|\Psi^J_{MNq}\rangle_q=1,\quad M=-J,-J+1,...,J. \tag{4.14}$$

By using Eqs. (2.5) and (2.9) for $M\neq J$, Eq. (4.14) can be transformed into the condition

$$\langle\Psi^J_{JNq}|\Psi^J_{JNq}\rangle_q=1. \tag{4.15}$$

In Appendix A, the squared norm of Ψ^J_{JNq} is calculated by using Eqs. (2.10), (2.11), (2.12), (3.2), and (3.3), and by taking Eqs. (4.1), (4.8), (4.10), and (4.12) into account. The resulting condition (4.15) reads

$$\frac{\ln q}{q-q^{-1}}\Big(B(q)\overline{\gamma(J,N,q^{-1})}\gamma(J,N,q)+\overline{B(q)\ \gamma(J,N,q)}\gamma(J,N,q^{-1})\Big)=1. \tag{4.16}$$

Since in the limit $q\rightarrow 1$, $\gamma(J,N,q)\rightarrow 1$, we may choose

$$\gamma(J,N,q)=1, \quad B(q)=\overline{B(q)}=\frac{q-q^{-1}}{2\ln q}=\frac{\sinh\tau}{\tau}. \tag{4.17}$$

For $q\to 1$ or $\tau\to 0$, we find that $B(q)\to 1$, so that $\langle\psi_1|\psi_2\rangle_q\to\langle\psi_1|\psi_2\rangle$, where the latter is given by Eq. (2.18), as it should be.

The results obtained can be summarized as follows:

Proposition IV.1: For $q\in\mathbb{R}^+$, the scalar product

$$\langle\psi_1|\psi_2\rangle_q=\frac{q-q^{-1}}{2\ln q}\int dz d\bar{z}\Bigg(\overline{\psi_1(z,\bar{z},q^{-1})}\frac{1}{(1+z\bar{z})(1+q^{-2}z\bar{z})}q^{-z\partial_z-\bar{z}\partial_{\bar{z}}-1}\psi_2(z,\bar{z},q)$$
$$+\overline{\psi_1(z,\bar{z},q)}\frac{1}{(1+z\bar{z})(1+q^2 z\bar{z})}q^{z\partial_z+\bar{z}\partial_{\bar{z}}+1}\psi_2(z,\bar{z},q^{-1})\Bigg), \tag{4.18}$$

or

$$\langle\psi_1|\psi_2\rangle_q=\frac{q-q^{-1}}{8\ln q}\int_0^{\pi}d\theta\sin\theta\int_0^{2\pi}d\phi$$
$$\times\Bigg(\overline{\psi_1(\theta,\phi,q^{-1})}\frac{1}{\sin^2(\theta/2)+q^{-2}\cos^2(\theta/2)}q^{\sin\theta\partial_\theta-1}\psi_2(\theta,\phi,q)$$
$$+\overline{\psi_1(\theta,\phi,q)}\frac{1}{\sin^2(\theta/2)+q^{2}\cos^2(\theta/2)}q^{-\sin\theta\partial_\theta+1}\psi_2(\theta,\phi,q^{-1})\Bigg), \tag{4.19}$$

unitarizes the $\mathrm{su}_q(2)$ realization (2.6), where N may take any integer or half-integer value. The functions $\Psi^J_{MNq}(z,\bar{z})$, or $\Psi^J_{MNq}(\theta,\phi)$, defined in Eqs. (2.10) and (2.15), where $J=|N|, |N|+1,\ldots, M=-J, -J+1,\ldots,J$, and $\gamma(J,N,q)=1$, form an orthonormal set with respect to such a scalar product.

From Proposition IV.1, we easily obtain the following corollary.

Corollary IV.2: For $q\in\mathbb{R}^+$, the q-Vilenkin functions $P^J_{MNq}(\xi)$, defined in Eq. (2.16), satisfy the orthonormality relation

$$\frac{q-q^{-1}}{4\ln q}\int_{-1}^{+1}d\xi\Bigg(\overline{P^{J'}_{MNq^{-1}}(\xi)}\frac{1}{q+q^{-1}-(q-q^{-1})\xi}q^{(\xi^2-1)\partial_\xi}P^J_{MNq}(\xi)$$
$$+\overline{P^{J'}_{MNq}(\xi)}\frac{1}{q+q^{-1}+(q-q^{-1})\xi}q^{-(\xi^2-1)\partial_\xi}P^J_{MNq^{-1}}(\xi)\Bigg)=\frac{\delta_{J',J}}{[2J+1]_q}. \tag{4.20}$$

B. The case where $q\in S^1$

Whenever $q\in S^1$, the ansatz (4.1) does not work, because though Eq. (4.6) remains valid, Eq. (4.7) is changed in such a way that both cannot be matched. Let us therefore change Eq. (4.1) into the following ansatz

$$\Big\langle\psi_1\Big|\psi_2\Big\rangle_q=\int_0^\infty d\rho\int_0^{2\pi}d\phi\Big(\overline{\psi_1(\rho,\phi,q)}f_1(\rho,q)q^{a_1\rho\partial_\rho}\psi_2(\rho,\phi,q)$$
$$+\overline{A_q\psi_1(\rho,\phi,q)}f_2(\rho,q)q^{a_2\rho\partial_\rho}A_q\psi_2(\rho,\phi,q)\Big), \tag{4.21}$$

where a_1, a_2, $f_1(\rho,q)$, $f_2(\rho,q)$, and A_q keep the same meaning as before.

Condition (4.3) is again automatically satisfied. Turning now to condition (4.4), it is easy to see that Eqs. (4.6) and (4.7) remain valid, except for the interchange of $\overline{\psi_1(\rho,\phi,q)}$ with $\overline{\psi_1(\rho,\phi,q^{-1})}$. Hence, Eq. (4.4) is also fulfilled by choosing a_1, a_2, $f_1(\rho,q)$, and $f_2(\rho,q)$ as given in Eqs. (4.8), and (4.10).

A difference with the case where $q\in\mathbb{R}^+$ appears when imposing the Hermiticity condition (4.11). The latter is now equivalent to the relations

$$\overline{B_1(q)}=B_1(q), \quad \overline{B_2(q)}=B_2(q), \tag{4.22}$$

showing that the real constants $B_1(q)$, and $B_2(q)$ remain independent. In the present case, keeping only one of the two terms on the right-hand side of Eq. (4.21) would therefore lead to a well-behaved scalar product.

As shown in Appendix B, condition (4.15) now reads

$$\frac{\ln q}{q-q^{-1}}(B_1(q)|\gamma(J,N,q)|^2+B_2(q)|\gamma(J,N,q^{-1})|^2)=1. \tag{4.23}$$

Among the infinitely many solutions of this equation, we may select the most symmetrical one,

$$\gamma(J,N,q)=1, \quad B_1(q)=B_2(q)=\frac{q-q^{-1}}{2\ln q}=\frac{\sin\tau}{\tau}. \tag{4.24}$$

Hence, whenever $q\to1$ or $\tau\to0$, the limit of $\langle\psi_1|\psi_2\rangle_q$ is again $\langle\psi_1|\psi_2\rangle$, as it should be.

In conclusion, we obtain the following proposition.

Proposition IV.3: For generic $q\in S^1$, the scalar product

$$\langle\psi_1|\psi_2\rangle_q=\frac{q-q^{-1}}{2\ln q}\int dz d\overline{z}\left(\overline{\psi_1(z,\overline{z},q)}\frac{1}{(1+z\overline{z})(1+q^{-2}z\overline{z})}q^{-z\partial_z-\overline{z}\partial_{\overline{z}}-1}\psi_2(z,\overline{z},q)\right.$$
$$\left.+\overline{\psi_1(z,\overline{z},q^{-1})}\frac{1}{(1+z\overline{z})(1+q^2z\overline{z})}q^{z\partial_x+\overline{z}\partial_{\overline{z}}+1}\psi_2(z,\overline{z},q^{-1})\right), \tag{4.25}$$

or

$$\langle\psi_1|\psi_2\rangle_q=\frac{q-q^{-1}}{8\ln q}\int_0^\pi d\theta\sin\theta\int_0^{2\pi}d\phi$$
$$\times\left(\overline{\psi_1(\theta,\phi,q)}\frac{1}{\sin^2(\theta/2)+q^{-2}\cos^2(\theta/2)}q^{\sin\theta\partial_\theta-1}\psi_2(\theta,\phi,q)\right.$$
$$\left.+\overline{\psi_1(\theta,\phi,q^{-1})}\frac{1}{\sin^2(\theta/2)+q^2\cos^2(\theta/2)}q^{-\sin\theta\partial_\theta+1}\psi_2(\theta,\phi,q^{-1})\right), \tag{4.26}$$

unitarizes the $\mathrm{su}_q(2)$ realization (2.6), where N may take any integer or half-integer value. The functions $\Psi^J_{MNq}(z,\overline{z})$, or $\Psi^J_{MNq}(\theta,\phi)$, defined in Eqs. (2.10) and (2.15), where $J=|N|,|N|+1,\ldots$, $M=-J,-J+1,\ldots,J$, and $\gamma(J,N,q)=1$, form an orthonormal set with respect to such a scalar product.

Corollary IV.4: For generic $q\in S^1$, the q-Vilenkin functions $P^J_{MNq}(\xi)$, defined in Eq. (2.16), satisfy the orthonormality relation

$$\frac{q-q^{-1}}{4\ln q}\int_{-1}^{+1}d\xi\left(\overline{P_{MNq}^{J'}(\xi)}\frac{1}{q+q^{-1}-(q-q^{-1})\xi}q^{(\xi^2-1)\partial_\xi}P_{MNq}^{J}(\xi)\right.$$
$$\left.+\overline{P_{MNq^{-1}}^{J'}(\xi)}\frac{1}{q+q^{-1}+(q-q^{-1})\xi}q^{-(\xi^2-1)\partial_\xi}P_{MNq^{-1}}^{J}(\xi)\right)=\frac{\delta_{J',J}}{[2J+1]_q}. \tag{4.27}$$

V. CONCLUSION

In the present paper, we did extend the study of the $\mathrm{su}_q(2)$ representations on a real two-dimensional sphere, carried out by Rideau and Winternitz,[6] in two ways.

First, we did prove that such representations exist not only for $q\in\mathbb{R}^+$, but also for generic $q\in S^1$. For such a purpose, we did provide an integral representation for the functions $Q_{Jq}(\eta)$, entering the definition of the q-Vilenkin functions, whenever J takes any half-integer value.

Second, we did unitarize the representations by determining appropriate scalar products for both ranges of q values. Such scalar products are expressed in terms of ordinary integrals, instead of q-integrals, as is usually the case.[5]

The resulting orthonormality relations for the q-Vilenkin and related functions should play an important role in applications to quantum mechanics, such as those considered in Refs. 10 and 11.

APPENDIX A: PROOF OF EQ. (4.16)

The purpose of this appendix is to evaluate the squared norm of the function $\Psi^J_{JNq}(z,\overline{z})$ when the scalar product (4.1) is used, and Eqs. (4.8), (4.10), and (4.12) are taken into account.

From Eqs. (2.2), (2.10), (2.11), and (2.12), Ψ^J_{JNq} can be written in polar coordinates as

$$\Psi^J_{JNq}=\frac{C_{JNq}}{[J+N]_q!}q^{-JN/2}Q_{Jq}(\rho^2)\rho^{J+N}e^{-i(J+N)\phi}. \tag{A1}$$

Its squared norm can therefore be expressed as

$$\langle\Psi^J_{JNq}|\Psi^J_{JNq}\rangle_q=\frac{\pi}{([J+N]_q!)^2}\left(B(q)\overline{C_{JNq^{-1}}}C_{JNq}q^{-J-N-1}\mathcal{I}_q+\overline{B(q)C_{JNq}}C_{JNq^{-1}}q^{J+N+1}\mathcal{I}_{q^{-1}}\right), \tag{A2}$$

in terms of the integral

$$\mathcal{I}_q=\int_0^\infty d\eta Q_{Jq^{-1}}(\eta)\frac{\eta^{J+N}}{(1+\eta)(1+q^{-2}\eta)}Q_{Jq}(q^{-2}\eta), \tag{A3}$$

and the same with q replaced by q^{-1}.

By introducing Eqs. (3.2) and (3.3) into Eq. (A3), we obtain

$$\mathcal{I}_q=\int_0^\infty d\eta\eta^{J+N}\prod_{k=0}^{\infty}\frac{(1+q^{2J+2k+2}\eta)}{(1+q^{-2J+2k-2}\eta)}=q^{2(J+1)(J+N+1)}\tilde{B}_{q^2}(J+N+1,J-N+1) \tag{A4}$$

if $0<q<1$, and

$$\mathcal{I}_q=\int_0^\infty d\eta\eta^{J+N}\prod_{k=0}^{\infty}\frac{(1+q^{-2J-2k-4}\eta)}{(1+q^{2J-2k}\eta)}=q^{-2J(J+N+1)}\tilde{B}_{q^{-2}}(J+N+1,J-N+1) \tag{A5}$$

if $q>1$. In Eqs. (A4) and (A5), we denote by $\tilde{B}_q(x,y)$ Ramanujan's continuous q-analog of the beta integral[2]

$$\tilde{B}_q(x,y)=\int_0^\infty dt\, t^{x-1}\prod_{k=0}^{\infty}\frac{(1+q^{x+y+k}t)}{(1+q^k t)},\quad 0<q<1, \tag{A6}$$

to distinguish it from the discrete q-analog of the same, known as $B_q(x,y)$ [see e.g. Eq. (1.11.7) of Ref. 3].

From Eq. (5.8) of Ref. 2, $\tilde{B}_q(x,y)$ is given for generic x values by

$$\tilde{B}_q(x,y)=\frac{\pi}{\sin \pi x}\prod_{k=1}^{\infty}\frac{(1-q^{k-x})(1-q^{x+y+k-1})}{(1-q^k)(1-q^{y+k-1})}. \tag{A7}$$

The values of x, which appear in Eqs. (A4) and (A5), being $x=J+N+1\in\mathbb{N}^+$, we have to calculate the limit of the right-hand side of Eq. (A7) when $x\to m\in\mathbb{N}^+$. Using L'Hospital's rule, we find

$$\lim_{x\to m}\frac{1-q^{m-x}}{\sin \pi x}=(-1)^m\frac{\ln q}{\pi},\quad m\in\mathbb{N}^+. \tag{A8}$$

Hence, for $x=m,\ y=n,\ m,n\in\mathbb{N}^+$, Eq. (A7) becomes

$$\tilde{B}_q(m,n)=(-1)^m(\ln q)\frac{\Pi_{k=1}^{m-1}(1-q^{k-m})}{\Pi_{k=1}^{m}(1-q^{n+k-1})}=\frac{(\ln q)q^{-m(n+m-1)/2}[m-1]_{q^{1/2}}![n-1]_{q^{1/2}}!}{(q^{1/2}-q^{-1/2})[n+m-1]_{q^{1/2}}!}, \tag{A9}$$

where in the last step, we introduced q-factorials, defined as in Sec. II.

From Eqs. (A4), (A5), and (A9), it follows that for any $q\in\mathbb{R}^+$

$$\mathcal{I}_q=\frac{2(\ln q)q^{J+N+1}[J+N]_q![J-N]_q!}{(q-q^{-1})[2J+1]_q!}. \tag{A10}$$

By taking Eq. (2.11) into account, the squared norm of Ψ^J_{JNq}, defined in Eq. (A2), therefore becomes

$$\langle\Psi^J_{JNq}|\Psi^J_{JNq}\rangle_q=\frac{\ln q}{q-q^{-1}}\left(B(q)\overline{\gamma(J,N,q^{-1})}\gamma(J,N,q)+\overline{B(q)\gamma(J,N,q)}\gamma(J,N,q^{-1})\right), \tag{A11}$$

which proves Eq. (4.16).

APPENDIX B: PROOF OF EQ. (4.23)

The purpose of this appendix is to evaluate the squared norm of the function $\Psi^J_{JNq}(z,\overline{z})$ when the scalar product (4.21) is used, and Eqs. (4.8), (4.10), and (4.22) are taken into account.

Since for $q\in S^1$, Ψ^J_{JNq} is still given by Eq. (A1), its squared norm reads

$$\langle\Psi^J_{JNq}|\Psi^J_{JNq}\rangle_q=\frac{\pi}{([J+N]_q!)^2}\left(B_1(q)|C_{JNq}|^2q^{-J-N-1}\mathcal{I}'_q+B_2(q)|C_{JNq^{-1}}|^2q^{J+N+1}\mathcal{I}'_{q^{-1}}\right). \tag{B1}$$

Here $\mathcal{I}'_q$ denotes the integral

$$\mathcal{I}'_q=\int_0^\infty d\eta F_{Jq}(\eta)\eta^{J+N} \tag{B2}$$

with

$$F_{Jq}(\eta)=\overline{Q_{Jq}(\eta)}\frac{1}{(1+\eta)(1+q^{-2}\eta)}Q_{Jq}(q^{-2}\eta). \tag{B3}$$

According to whether J is integer or half-integer, we have to insert Eq. (3.4) or Eq. (3.20) into Eq. (B3). In both cases, the result reads

$$F_{Jq}(\eta)=\prod_{p=0}^{2J+1}\frac{1}{1+q^{2J-2p}\eta}. \tag{B4}$$

This is obvious in the former case. In the latter, by using the property $\overline{L_q(\eta)}=-L_q(\eta)$, Eq. (B3) can be transformed into

$$\begin{aligned}F_{Jq}(\eta)&=\exp\{-L_q(q^{2J+1}\eta)+L_q(q\eta)\}\frac{1}{(1+\eta)(1+q^{-2}\eta)}\exp\{L_q(q^{-2J-3}\eta)-L_q(q^{-3}\eta)\}\\&=\frac{1}{(1+\eta)(1+q^{-2}\eta)}\exp\Bigg\{-\sum_{p=0}^{2J+1}[L_q(q^{2J+1-2p}\eta)-L_q(q^{2J-1-2p}\eta)]\\&\quad+[L_q(q\eta)-L_q(q^{-1}\eta)]+[L_q(q^{-1}\eta)-L_q(q^{-3}\eta)]\Bigg\}.\end{aligned} \tag{B5}$$

Repeated use of Eq. (3.8) for various arguments then directly leads to the searched for result (B4).

To evaluate $\mathcal{I}_q'$ for $F_{Jq}(\eta)$ given by Eq. (B4), we cannot use the same method as that employed in Appendix A to calculate $\mathcal{I}_q$, because in the q-analog of the beta integral, given in Eq. (A6), q is assumed real. Let us therefore rewrite the integrand of $\mathcal{I}_q'$ in the form

$$\eta^{J+N}F_{Jq}(\eta)=\sum_{p=0}^{2J+1}\frac{a_p^{(J)}}{\eta+q^{2p-2J}}, \tag{B6}$$

where the coefficient $a_p^{(J)}$ is the residue of $\eta^{J+N}F_{Jq}(\eta)$ at the pole $\eta=-q^{2p-2N}$, i.e.,

$$a_p^{(J)}=(-1)^{J+N}\frac{q^{J+1}}{(q-q^{-1})^{2J+1}}\times\frac{(-1)^p q^{N(2p-2J)}}{[p]_q![2J-p+1]_q!}. \tag{B7}$$

Then

$$G_{Jq}(\eta)\equiv\int d\eta F_{Jq}(\eta)\,\eta^{J+N}=(-1)^{J+N}\frac{q^{J+1}}{(q-q^{-1})^{2J+1}}\sum_{p=0}^{2J+1}\frac{(-1)^p q^{N(2p-2J)}}{[p]_q![2J-p+1]_q!}\ln(\eta+q^{2p-2J}). \tag{B8}$$

To calculate the values of $G_{Jq}(\eta)$ for $\eta\to\infty$ and $\eta=0$, the following identities[1] will be useful:

$$(1;\eta)_q^{2J+1}\equiv\sum_{p=0}^{2J+1}\begin{bmatrix}2J+1\\p\end{bmatrix}_q\eta^p=\prod_{p=0}^{2J}(1+q^{2p-2J}\eta), \tag{B9}$$

$$\frac{d}{d\eta}(1;\eta)_q^{2J+1}=\sum_{p=0}^{2J+1}\begin{bmatrix}2J+1\\p\end{bmatrix}_q p\,\eta^{p-1}=\sum_{p=0}^{2J}q^{2p-2J}\prod_{\substack{r=0\\r\neq p}}^{2J}(1+q^{2r-2J}\eta), \tag{B10}$$

where

$$\begin{bmatrix} n \\ p \end{bmatrix}_q \equiv \frac{[n]_q!}{[p]_q![n-p]_q!} \tag{B11}$$

is a q-binomial coefficient. From Eq. (B9), we obtain

$$(1;-q^{2N})_q^{2J+1} = \sum_{p=0}^{2J+1} (-1)^p \begin{bmatrix} 2J+1 \\ p \end{bmatrix}_q q^{2Np} = 0, \tag{B12}$$

because on the right-hand side, the factor $(1-q^{2p-2J+2N})$ vanishes for $p=J-N$. Similarly, from Eq. (B10), we get

$$\begin{aligned} \frac{d}{d\eta}(1;\eta)_q^{2J+1}\bigg|_{\eta=-q^{2N}} &= \sum_{p=0}^{2J+1} (-1)^{p-1} p \begin{bmatrix} 2J+1 \\ p \end{bmatrix}_q q^{2N(p-1)} \\ &= (-1)^{J+N}(q-q^{-1})^{2J}[J-N]_q![J+N]_q!q^{N(2J-1)}, \end{aligned} \tag{B13}$$

since on the right-hand side only the term corresponding to $p=J-N$ leads to a nonvanishing result.

By noting that for $\eta \gg 1$,

$$\ln(\eta + q^{2p-2J}) \simeq \ln(\eta) + \frac{q^{2p-2J}}{\eta} + O\left(\frac{1}{\eta^2}\right), \tag{B14}$$

it directly results from Eq. (B12) that

$$\lim_{\eta\to\infty} G_{Jq}(\eta) = 0. \tag{B15}$$

Furthermore, from Eqs. (B12) and (B13), we obtain

$$\begin{aligned} G_{Jq}(0) &= \frac{(-1)^{J+N} q^{J+1} \ln q}{(q-q^{-1})^{2J+1}} \sum_{p=0}^{2J+1} \frac{(-1)^p (2p-2J) q^{N(2p-2J)}}{[p]_q![2J-p+1]_q!} \\ &= \frac{2(-1)^{J+N} q^{J+1} \ln q}{[2J+1]_q!(q-q^{-1})^{2J+1}} \sum_{p=0}^{2J+1} (-1)^p p \begin{bmatrix} 2J+1 \\ p \end{bmatrix}_q q^{N(2p-2J)} \\ &= -\frac{2[J+N]_q![J-N]_q! q^{J+N+1} \ln q}{[2J+1]_q!(q-q^{-1})}. \end{aligned} \tag{B16}$$

By taking Eqs. (B8), (B15), and (B16) into account, we conclude that for generic $q \in S^1$, $\mathcal{I}_q'$, defined in Eq. (B2), is given by

$$\mathcal{I}_q' = \frac{2(\ln q) q^{J+N+1} [J+N]_q! [J-N]_q!}{(q-q^{-1})[2J+1]_q!}. \tag{B17}$$

By combining this result with Eqs. (2.11), (4.15), and (B1), Eq. (4.23) directly follows.

[1] H. Exton, *q-Hypergeometric Functions and Applications* (Ellis Horwood, New York, 1983).
[2] G. E. Andrews, *q-Series: Their Development and Application in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra* (American Mathematical Society, Providence, RI, 1986).
[3] G. Gasper and M. Rahman, *Basic Hypergeometric Series* (Cambridge U.P., Cambridge, 1990).
[4] V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge U.P., Cambridge, 1994).
[5] N. Ja. Vilenkin and A. U. Klimyk, *Representation of Lie Groups and Special Functions* (Kluwer, Dordrecht, 1991), 3 volumes.
[6] G. Rideau and P. Winternitz, J. Math. Phys. **34**, 6030 (1993).

[7] N. Ja. Vilenkin, *Special Functions and the Theory of Group Representations* (American Mathematical Society, Providence, RI, 1968).
[8] Note that the functions $P^J_{MN}(\cos\theta)$, defined in Ref. 5, Vol. 1, p. 85, differ by a phase factor from those considered in Refs. 6 and 7.
[9] J. R. Schmidt, J. Math. Phys. **37**, 3062 (1996).
[10] M. Irac-Astaud, Lett. Math. Phys. **36**, 169 (1996); Czech. J. Phys. **46**, 179 (1996).
[11] M. Irac-Astaud and C. Quesne, unpublished.

On an application of Crum–Krein transform to expansions in products of solutions of two Sturm–Liouville equations

E. Kh. Khristov
Department of Mathematics and Informatics, Sofia University, ''St. Kl. Ohridski,'' J. Bourchier 5, 1164 Sofia, Bulgaria

(Received 10 July 1998; accepted for publication 1 December 1998)

By using the Crum–Krein transform we construct explicit formulas allowing us to reduce the problem of expansions in products of solutions for two Sturm–Liouville problems on the semiaxis with general boundary conditions to the same problem for two Sturm–Liouville problems with zero boundary conditions. We also obtain the transformation formulas for the corresponding $\mathbf{\Lambda}$ operators. © *1999 American Institute of Physics.* [S0022-2488(99)01306-7]

I. INTRODUCTION

Let us consider two self-adjoint Sturm–Liouville problems with a spectral parameter $\lambda = k^2$ for the equations

$$y_n'' + (k^2 - v_n(x))y_n = 0, \quad 0 < x < \infty, \quad n = 1,2 \tag{1.1}$$

with boundary conditions

$$y_n'(0) - \gamma_n y_n(0) = 0 \quad ('=d/dx). \tag{1.2}$$

Here and below γ_n are finite real numbers and the real-valued potentials $v_n(x)$ belong to the space X_1, where

$$X_1 = \left\{ f(x): \int_0^\infty (1+x)|f(x)|dx < \infty \right\}.$$

The main purpose of this paper is the construction of transforms which allow one to obtain the expansions in products $Y(x,k) = y_1(x,k)y_2(x,k)$ of solutions of the problems (1.1), (1.2) and the operators associated with them, usually called $\mathbf{\Lambda}$ operators, starting from the problems

$$y_n'' + (k^2 - r_n(x))y_n = 0, \quad 0 < x < \infty, \quad n = 1,2, \tag{1.3}$$

$$y_n(0) = 0, \tag{1.4}$$

where $r_n(x) \in X_1$. For the problem (1.3), (1.4) the corresponding expansion formulas are well known and their proofs are simpler, see, e.g., Ref. 1. Combining the results in this paper with the previous ones[1–3] we obtain an algebraic scheme allowing us to reduce the problem of expansion in products of solutions for two radial Schrödinger equations (1.3) with $r_n^{(l)}(x) = l(l+1))x^{-2} + r_n(x)$, $l = 1,2,...$, as well as those connected with (1.1) and (1.2) to the simplest case $l = 0$ with a boundary condition $y_n(0) = 0$ and purely continuous spectrum. Let us note that each of the problems (1.1), (1.2) and (1.3), (1.4) can be treated separately for example using the method of contour integration.[1,4] In contrast with the case (1.3) and (1.4) the spectral problem for the Λ operator in the case (1.1) and (1.2) leads to nonlocal k-dependent boundary conditions. The method we present here avoids the above difficulties.

The interest in these topics is motivated mainly by the applications to the theory of the soliton equations, see, e.g., Refs. 5–8. It is also known[9] that the completeness of the products of solutions of two Sturm–Liouville problems plays an important role in various theorems concerning the associated direct and inverse problems. Articles 10–12 are among the first where expansion formulas in terms of the squared Sturm–Liouville or Zakharov–Shabat eigenstates are derived. Here we specifically mention paper 13, where expansions in products of solutions of two Sturm–Liouville problems with $v_n(x) \in L_{\mathrm{loc}}$ are considered (see also Ref. 1, where one can find a detailed bibliography and additional remarks about the spectral theory of Λ operators). As pointed out in Refs. 14 and 15, the solution of the inverse scattering problem by using the Gelfand–Levitan–Marchenko equation for the operator (1.1) and (1.2) is essentially different from that with a boundary condition $y(0)=0$, which is confirmed in the present work as well. In Refs. 16 and 17 it has been shown that by using Crum–Krein transform[18,19] the inverse problem for the operator (1.1) and (1.2) can be reduced to the inverse problem for the operator (1.3) and (1.4). Similar transforms on a finite interval have been systematically developed in Refs. 20 and 21 on the semiaxis such transforms have been well known since the pioneer works 17,19 and 22; see also Ref. 23 where, in particular, these results are stated in an operator language.

Let us introduce some notations which we shall use later on, and outline our main results. We denote by $\varphi_n(x,k)$ and $h_n(x,k)$ the regular and the Jost solutions of Eq. (1.3), which are defined by the conditions

$$\varphi_n(0,k)=0, \quad \varphi_n'(0,k)=1, \quad \lim_{x\to\infty} h_n(x,k)\exp(-ikx)=1 \quad (\operatorname{Im} k \geqslant 0),$$

and let $h_n(k)=h_n(0,k)=W(h_n,\varphi_n)(W(f,g)=fg'-f'g)$ be the characteristic function of the problem (1.3), (1.4). We also introduce the products $\Phi(x,k)=\varphi_1(x,k)\varphi_2(x,k)$, $H(x,k)=h_1(x,k)h_2(x,k)$, and $H(k)=H(0,k)=h_1(k)h_2(k)$. Furthermore, we define

$$\tilde{H}(x,k)=\frac{4}{\pi}k\,\operatorname{Im}\{H(x,k)H^{-1}(k)\}. \tag{1.5}$$

As a result we obtain

$$\tilde{H}(0,k)=0, \quad \tilde{H}'(0,k)=\frac{4}{\pi}\left\{\frac{k^2}{|h_1(k)|^2}+\frac{k^2}{|h_2(k)|^2}\right\}. \tag{1.6}$$

The following theorem deals with expansion formulas associated with the products $\Phi(x,k)$ and $H(x,k)$.[1]

Theorem 1.1: *Let the problems (1.3), (1.4) with purely continuous spectrum, i.e.,* $h_n(k)\neq 0$, $\operatorname{Im} k\geqslant 0$, $n=1,2$, *be given, and the functions* $\Phi(x,k)$ *and* $\tilde{H}(x,k)$ *be constructed as above. Then for any absolutely continuous function* $f\in X_1$ *we have*

$$f(x)=\int_0^\infty \tilde{H}(x,k)(f,\Phi'(k))dk, \quad f(x)=\int_0^\infty \Phi'(x,k)(f,\tilde{H}(k))dk. \tag{1.7}$$

If, in addition, $f\in L_1^0=\{f\in L_1=L_1(0,\infty):\int_0^\infty f(x)dx=0\}$ *then*

$$f(x)=-\int_0^\infty \tilde{H}'(x,k)(f,\Phi(k))dk, \tag{1.8}$$

where $(f,\Phi'(k))=\int_0^\infty f(x)\Phi'(x,k)dx$, *etc.*

Now in connection with the problem (1.1), (1.2) we denote by $\psi_n(x,k)$ and $f_n(x,k)$ the solutions of (1.1) for which

$$\psi_n(0,k)=1,\quad \psi_n'(0,k)=\gamma_n,\quad \lim_{x\to\infty} f_n(x,k)\exp(-ikx)=1(\operatorname{Im} k\geqslant 0),$$

and let $e_n(k)=f_n'(0,k)-\gamma_n f_n(0,k)=W(\psi_n,f_n)$ be the characteristic function of the problem (1.1), (1.2). Here we introduce the products $\Psi(x,k)=\psi_1(x,k)\psi_2(x,k)$, $F(x,k)=f_1(x,k)f_2(x,k)$, and $E(k)=e_1(k)e_2(k)$.

Let us also define the real space $\mathcal{N}=L_2(0,\infty)\oplus\mathbb{R}\ni\hat{f}=(f(x),\alpha)$ with a scalar product

$$(\hat{f},\hat{g})_1=(f,g)+\alpha\beta,\quad (f,g)=\int_0^\infty f(x)g(x)dx.$$

We will use that notation also in the case when $f\in X_1$ (or $f\in L_1$) and $g(x)\in L_\infty=L_\infty(0,\infty)$. We denote the corresponding spaces by $\mathcal{N}_1$ and $\mathcal{N}_\infty$. Let $C^k=C^k(0,\infty)$ be, as usual, the space of functions $f(x)$ with continuous kth derivatives and $C_0^k=C_0^k(0,\infty)$ be the space of functions $f(x)\in C^k$ such that $\lim_{x\to\infty} f^{(l)}(x)=0$, $l=0,1,\dots,k$. Let also $\mathcal{N}^k=C^k\oplus\mathbf{R}$, $\mathcal{N}_0^k=C_0^k\oplus\mathbf{R}$. Next, we introduce the functions from $\mathcal{N}_\infty$,

$$\hat{\tilde{F}}(x,k)=(\tilde{F}(x,k),\tilde{F}(0,k)),\quad \tilde{F}(x,k)=\frac{4}{\pi}k\operatorname{Im}\{F(x,k)E^{-1}(k)\} \tag{1.9}$$

and

$$\hat{\Psi}'(x,k)=(\Psi'(x,k),1/2). \tag{1.10}$$

Here the analog of Theorem 1.1 is the following theorem.[1,4]

Theorem 1.2: *Let the problems (1.1), (1.2) with purely continuous spectrum, i.e.,* $e_n(k)\neq 0$, $\operatorname{Im} k\geqslant 0$, $n=1,2$, *be given, and the functions* $\hat{\Psi}'(x,k)$ *and* $\hat{\tilde{F}}(x,k)$ *be constructed as above. Then for any function* $\hat{f}=(f(x),\alpha)\in\mathcal{N}_1$ *where* $f(x)$ *is absolutely continuous the following expansion formulas take place*:

$$\hat{f}=\int_0^\infty \hat{\tilde{F}}(x,k)(\hat{f},\hat{\Psi}'(k))_1\,dk,\quad \hat{f}=\int_0^\infty \hat{\Psi}'(x,k)(\hat{f},\hat{\tilde{F}}(k))_1\,dk, \tag{1.11}$$

and

$$f(x)=-\int_0^\infty \tilde{F}'(x,k)(f,\tilde{\Psi}(k))dk, \tag{1.12}$$

where $\tilde{\Psi}(x,k)=\Psi(x,k)-1/2$.

The article is organized in the following way. In Sec. I we show that the Crum operator

$$\mathcal{F}(\hat{v}_n;x)\overset{\text{def}}{=}v_n(x)-2\frac{d^2}{dx^2}\ln z_n(x)=r_n(x),\quad (\hat{v}_n=(v_n(x),\gamma_n)), \tag{1.13}$$

where $z_n(x)=\psi(x,k_0)$, $k_0=i\tau_0$, $\tau_0>0$ generates transforms which reduce the expansions in Theorem 1.2 to those in Theorem 1.1 and vice versa. In Sec. III, starting from some well-known results concerning $\mathbf{\Lambda}$ operators associated with the problems (1.3), (1.4), we construct a one-parameter family of $\mathbf{\Lambda}$ operators (τ_0 being the parameter) associated with (1.1), (1.2). The $\mathbf{\Lambda}$ operators obtained here do not coincide with the ones given in Ref. 4 even though they have the same decompositions of unity generated by the expansion formulas given in Theorem 1.2. In Sec. IV, some applications of our results to the scattering theory for the problem (1.1), (1.2) are given. In particular, here we show that if $\hat{v}_1=\hat{v}_2$ then the transform given in Sec. II can be obtained as a derivative of $\mathcal{F}(\hat{v};x)$ with respect to $\hat{v}$. Let us note that the restriction for a purely continuous spectrum of the problems considered is not essential in view of the results 1, 2 and is used only to keep the size of the paper reasonable. In Sec. V, the general case when a discrete spectrum exists is considered in brief. Some well-known results concerning the Sturm–Liouville problems considered, which we use below without reference, can be found in Refs. 15, 17, and 24.

II. TRANSFORM OF THE EXPANSION FORMULAS

The basis for our constructions as well as for those in Refs. 1–3 is the following lemma.

Lemma 2.1: Consider the equations

$$y_n''+(\lambda-v_n(x))y_n=0,\quad a<x<b,\quad n=1,2 \tag{2.1}$$

and construct the equations

$$y_n^{(1)''}+(\lambda-r_n(x))y_n^{(1)}=0,\quad r_n(x)=v_n(x)-2\,\frac{d^2}{dx^2}\ln z_n(x),$$

where $z_n(x)$ *are solutions of (2.1) for* $\lambda=\lambda_0$, $z_n(x)\neq 0$, $a<x<b$. *Then*

$$\begin{aligned} W(Z(x),Y(x,\lambda))&=(\lambda_0-\lambda)\frac{d}{dx}(Z(x)Y_1(x,\lambda)),\\ W(Z_1(x),Y_1(x,\lambda))&=(\lambda_0-\lambda)^{-1}\frac{d}{dx}(Z_1(x)Y(x,\lambda)), \end{aligned} \tag{2.2}$$

where $Y(x,\lambda)=y_1(x,\lambda)y_2(x,\lambda)$, $Y_1(x,\lambda)=y_1^{(1)}(x,\lambda)y_2^{(1)}(x,\lambda)$ *and* $Z(x)=z_1(x)z_2(x)$, $Z_1(x)=Z^{-1}(x)$.

Lemma 2.1 is a direct consequence of the well-known Crum's lemma[18] (see also Lemma 2.2 below) in view of the identity[8]

$$W(Y(x,\lambda),Z(x,\lambda_0))=\frac{1}{\lambda-\lambda_0}\frac{d}{dx}\prod_{n=1,2}W(y_n(x,\lambda),z_n(x,\lambda_0)).$$

Let us now consider, along with the problems (1.1), (1.2), the problems (1.3), (1.4) where $r_n(x)$ is defined as in (1.13). The following lemma has been obtained in Refs. 16 and 17.

Lemma 2.2. Let $z(x)=\psi(x,k_0)$ *be a solution of the Eq. (1.1) for* $k_0=i\tau_0$, $\tau_0>0$, $z(x)\neq 0$, $0\leq x<\infty$. *Then the potential* $r(x)$ *defined by (1.13) belongs to* X_1 *and the solutions* $\varphi(x,k)$ *and* $h(x,k)$ *of* (1.3) *are expressed via the solutions* $\psi(x,k)$ *and* $f(x,k)$ *as follows:*

$$\varphi(x,k)=\frac{W(z(x),\psi(x,k))}{(k_0^2-k^2)z(x)},\quad h(x,k)=\frac{W(z(x),f(x,k))}{i(k_0+k)z(x)} \tag{2.3}$$

and

$$h(k)=-i(k_0+k)^{-1}e(k). \tag{2.4}$$

Here $y(x)=z^{-1}(x)$ *is the solution of Eq. (1.3) for* $k=k_0$,

$$y(x)\sim Ce^{ik_0x},\quad x\to\infty, \tag{2.5}$$

where C is a constant.

Let us denote as usual by $\mathcal{L}(X,Y)$ the space of bounded linear operators defined in X with values in Y. Now we introduce in the space $X=X_1$ (or $X=L_\infty$) the operators

$$\mathbf{A}f=f(x)=2Z(x)\int_x^\infty Y'(t)f(t)dt,\quad \tilde{\mathbf{A}}f=f(x)-2Z'(x)\int_x^\infty Y(t)f(t)dt,$$

where

$$Y(x)=Z^{-1}(x),\quad Z(x)=\Psi(x,k_0). \tag{2.6}$$

One can directly check that

$$\mathbf{A}Y=0,\quad \tilde{\mathbf{A}}Y'=0, \tag{2.7}$$

where the second equation follows from the first one due to $d\mathbf{A}f(x)/dx=\tilde{\mathbf{A}}f'(x)$. Also, (2.5) leads to $\mathbf{A},\tilde{\mathbf{A}}\in\mathcal{L}(X_1,X_1)$ (respectively, $\mathcal{L}(L_\infty,L_\infty)$). By using the operators $\mathbf{A}$ and $\tilde{\mathbf{A}}$ we construct operators $\hat{\mathbf{A}},\hat{\tilde{\mathbf{A}}}:X_1\to\mathcal{N}_1(L_\infty\to\mathcal{N}_\infty)$ in the following way:

$$\hat{\mathbf{A}}f=(\mathbf{A}f(x),2(Y',f)),\quad \hat{\tilde{\mathbf{A}}}f=(\tilde{\mathbf{A}}f(x),-(Y,f)). \tag{2.8}$$

Next, we introduce the operators

$$\mathbf{B}f=f(x)-2Y(x)\int_0^x Z'(t)f(t)dt,\quad \tilde{\mathbf{B}}f=f(x)+2Y'(x)\int_0^x Z(t)f(t)dt.$$

In view of (2.6) for the adjoint operators $\mathbf{B}^*$ and $\tilde{\mathbf{B}}^*$ we have

$$\mathbf{B}^*=\tilde{\mathbf{A}},\quad \tilde{\mathbf{B}}^*=\mathbf{A}. \tag{2.9}$$

Further on, we construct the operators $\hat{\mathbf{B}}$ and $\hat{\tilde{\mathbf{B}}}:\mathcal{N}_1\to X_1(\mathcal{N}_\infty\to L_\infty)$ via

$$\hat{\mathbf{B}}\hat{f}=\mathbf{B}f(x)-\alpha Y(x),\quad \hat{\tilde{\mathbf{B}}}\hat{f}=\tilde{\mathbf{B}}f(x)+2\alpha Y'(x). \tag{2.10}$$

From (2.7) and (2.9) it follows that

$$\hat{\tilde{\mathbf{A}}}=\hat{\tilde{\mathbf{B}}}^{-1},\quad \hat{\mathbf{A}}=\hat{\mathbf{B}}^{-1}. \tag{2.11}$$

The next theorem gives the general transformation formulas for the functions $\hat{\Psi}(x,k)$, $\hat{\tilde{F}}(x,k)$ and $\Phi(x,k)$, $\tilde{H}(x,k)$.

Theorem 2.1: *Let the problems (1.1), (1.2) and (1.3), (1.4) be connected by the transform (1.13). Then the following representations hold:*

$$\hat{\Psi}'(x,k)=(k_0^2-k^2)\hat{\tilde{\mathbf{A}}}\Phi'(x,k),\quad \hat{\tilde{F}}(x,k)=(k_0^2-k^2)^{-1}\hat{\mathbf{A}}\tilde{H}(x,k), \tag{2.12}$$

$$\Phi'(x,k)=(k_0^2-k^2)^{-1}\hat{\tilde{\mathbf{B}}}\hat{\Psi}'(x,k),\quad \tilde{H}(x,k)=(k_0^2-k^2)\hat{\mathbf{B}}\hat{\tilde{F}}(x,k). \tag{2.13}$$

Proof: From (2.11) it follows that the equalities in (2.13) are a direct consequence of (2.12). Let us prove (2.12). By using (2.2) with $Z_1=Y(x)$, $[Y(x)$ is defined by (2.6)], $Y_1(x,\lambda)=\Phi(x,k)$, $Y(x,\lambda)=\Psi(x,k)$ and the first of the formulas in (2.3) we get

$$W(Y(x),\Phi(x,k))=(k_0^2-k^2)^{-1}\frac{d}{dx}(Y(x)\Psi(x,k)).$$

After integrating from x to ∞ and having in mind (2.5) we obtain the first formula in (2.12). The second formula in (2.12) can be obtained in a similar way taking into account (2.4). □

Now we shall show how starting from the expansion formulas given in the Theorem 1.1 one can derive the corresponding expansions in Theorem 1.2 and vice versa. We explain in more detail only the transition (1.12)$\leftrightarrow$(1.8). The transitions (1.11)$\leftrightarrow$(1.7) could be obtained analogously and so we omit its proof here.

Proof of the transition (1.12)$\leftrightarrow$(1.8)*:* First we show how starting from the expansion in (1.8) one can obtain the one in (1.12). In order to do that we put $f(x)=\tilde{\mathbf{B}}g(x)-Y'(x)\int_0^\infty g(x)dx$ in (1.8). Evidently $f\in L_1^0$. Let us mention here that if $f(x)=-\int_x^\infty f'(t)dt,\ f'(x)\in X_1$, then

$$(\hat{f},\hat{\Psi}'(k))_1=\tfrac{1}{2}(\alpha-f(0))-(f',\tilde{\Psi}(k)). \tag{2.14}$$

Now we construct the function $\hat{G}_0(x)=(G(x)=-\int_x^\infty g(t)dt,G(0))$. From (2.13) and (2.14) we obtain

$$-(\hat{\mathbf{B}}\hat{G}_0,\Phi'(k))=((\hat{\mathbf{B}}\hat{G}_0)',\Phi(k))=(k_0^2-k^2)^{-1}(g,\tilde{\Psi}(k)).$$

Therefore, (1.8) leads to

$$\tilde{\mathbf{B}}g(x)-Y'(x)\int_0^\infty g(t)dt=-\int_0^\infty \tilde{H}'(x,k)(k_0^2-k^2)^{-1}(g,\tilde{\Psi}(k))dk. \tag{2.15}$$

By applying the operator $\tilde{\mathbf{A}}$ to both sides of (2.15) we obtain the expansion (1.12) in view of (2.7), and the representation $\tilde{F}'(x,k)=(k_0^2-k^2)^{-1}\tilde{\mathbf{A}}\tilde{H}'(x,k)$ which is a direct consequence of (2.12). In connection with deriving the expansion (1.8) from (1.12) it is sufficient to mention that from (2.13) it follows $(g,\Phi(k))=(k_0^2-k^2)^{-1}(\tilde{\mathbf{A}}g,\tilde{\Psi}(k))$ where $g(x)\in L_1^0$. □

Remark 2.1: It can be proved (see, e.g., Refs. 1 and 2) that the system $\{\Phi'(x,k),0<k<\infty\}$ is biorthogonal to the system $\{\tilde{H}(x,\mu),0<\mu<\infty\}$, i.e., if $k,\mu>0$ then there exists (in a distributional sense) the limit

$$(\tilde{H}(\mu),\Phi'(k))=\lim_{N\to\infty}\int_0^N \tilde{H}(x,\mu)\Phi'(x,k)dx=\frac{\sin 2N(\mu-k)}{\pi(\mu-k)}=\delta(\mu-k). \tag{2.16}$$

From the transformation formulas given in Theorem 2.1 and using the identity $(\hat{\mathbf{B}}\hat{f},\hat{\tilde{\mathbf{B}}}\hat{g})_1=(\hat{f},\hat{g})_1$, we have that

$$(\tilde{H}(k),\Phi'(\mu))=\frac{k_0^2-k^2}{k_0^2-\mu^2}(\hat{\tilde{F}}(k),\hat{\Psi}'(\mu))_1.$$

The last relation together with (2.16) shows that the system $\{\hat{\Psi}'(x,k),0<k<\infty\}$ is biorthogonal to the system $\{\hat{\tilde{F}}(x,\mu),0<\mu<\infty\}$ with respect to the scalar product $(\cdot,\cdot)_1$.

III. Λ OPERATORS FOR THE PROBLEMS (1.1), (1.2)

In the present section we construct the $\mathbf{\Lambda}$ operators for which the expansions given in Theorem 1.2 are decomposition of unity. We shall use some of the well-known results from the Λ operator theory related with the problem (1.3), (1.4) which we introduce in the beginning. The main result of this section is Theorem 3.2 below.

In Ref. 25 it has been shown that if one introduces the operator

$$\mathbf{\Lambda}_{x_0}=\frac{1}{4}\left[-\frac{d^2}{dx^2}+2s(r;x)-\int_{x_0}^x dt\, s_t(r;t)-\int_{x_0}^x dt\,\Delta(r;t)\int_{x_0}^t dz\,\Delta(r;z)\right],$$

where $s(r;x)=r_1(x)+r_2(x)$, $\Delta(r;x)=r_1(x)-r_2(x)$ then the product $Y(x,k)=y_1(x,k)y_2(x,k)$ of any two solutions y_n of (1.3) satisfies the equation

$$\mathbf{\Lambda}_{x_0} Y(x,k) = \lambda Y(x,k) + B(x_0;x,k), \quad \lambda = k^2, \tag{3.1}$$

where

$$B(x_0;x,k) = -\frac{1}{4}\Bigg[W(y_1(x),y_2(x))\Big|_{x=x_0} \int_{x_0}^{x} \Delta(r;t)\,dt + (2\lambda - s(r;x_0))Y(x_0,k) + 2y_1'(x_0,k)y_2'(x_0,k)\Bigg].$$

Now let us introduce the operators $\mathbf{\Lambda} = \mathbf{\Lambda}_{x_0=\infty}$, $\mathbf{\Lambda}^* = (d/dx)\mathbf{\Lambda}_{x_0=0}\int_0^x$, i.e.,

$$\mathbf{\Lambda} = \frac{1}{4}\left[-\frac{d^2}{dx^2} + 2s(r;x) + \int_x^\infty dt\, s_t(r;t) - \int_x^\infty dt\, \Delta(r;t) \int_t^\infty dz\, \Delta(r;z) \right],$$

$$\mathbf{\Lambda}^* = \frac{1}{4}\left[-\frac{d^2}{dx^2} + 2s(r;x) + s_x(r;x)\int_0^x dt - \Delta(r;x) \int_0^x dt\, \Delta(r;t) \int_0^t dz \right]$$

with a definition domain

$$\mathcal{D}(\mathbf{\Lambda}) = \mathcal{D}(\mathbf{\Lambda}^*) = (f(x): f(0)=0, f(x) \in C_0^2 \cap L^1).$$

Then the operator $\mathbf{\Lambda}^*$ is adjoint to $\mathbf{\Lambda}$, i.e.,

$$(\mathbf{\Lambda} f, g) = (f, \mathbf{\Lambda}^* g) \quad \text{for } f,g \in \mathcal{D}(\mathbf{\Lambda}). \tag{3.2}$$

It is well known[7,25] that the function $H(x,k)$ [as well as $\tilde{H}(x,k)$] is a solution of the equation

$$\mathbf{\Lambda} H(x,k) = k^2 H(x,k), \quad \operatorname{Im} k \geqslant 0, \tag{3.3}$$

where $\tilde{H}(x,k)$ satisfies the boundary conditions (1.6). Similarly, the function $\Phi(x,k)$ solves the equation

$$\mathbf{\Lambda}^* \Phi'(x,k) = k^2 \Phi'(x,k), \tag{3.4}$$

with boundary conditions $\Phi(0,k) = \Phi'(0,k) = 0$, $\Phi''(0,k) = 2$.

These equations together with Theorem 1.1 show[1] that the expansions (1.7) are decompositions of unity for the operators $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^*$, i.e.,

$$\mathbf{\Lambda} f(x) = -\int_0^\infty k^2 \tilde{H}(x,k)(f, \Phi'(k))\,dk \quad \text{for } f(x) \in \mathcal{D}(\mathbf{\Lambda}), \tag{3.5}$$

$$\mathbf{\Lambda}^* f(x) = -\int_0^\infty k^2 \Phi'(x,k)(f, \tilde{H}(k))\,dk \quad \text{for } f(x) \in \mathcal{D}(\mathbf{\Lambda}). \tag{3.6}$$

Let us now introduce the operators

$$\hat{\mathbf{\Lambda}}^* = \hat{\tilde{\mathbf{A}}}\mathbf{\Lambda}^*\hat{\tilde{\mathbf{B}}} \equiv \hat{\tilde{\mathbf{B}}}^{-1}\mathbf{\Lambda}^*\hat{\tilde{\mathbf{B}}}, \quad \hat{\mathbf{\Lambda}} = \hat{\mathbf{A}}\mathbf{\Lambda}\hat{\mathbf{B}} \equiv \hat{\mathbf{B}}^{-1}\mathbf{\Lambda}\hat{\mathbf{B}}, \tag{3.7}$$

acting in the space $\mathcal{N}_0^2(\mathcal{N}_\infty^2)$ according to

$$\hat{\mathbf{\Lambda}}^* \hat{f} = (\tilde{\mathbf{A}}\mathbf{\Lambda}^*\tilde{\mathbf{B}} f(x) + 2\alpha \tilde{\mathbf{A}}\mathbf{\Lambda}^* Y'(x), -(Y, \mathbf{\Lambda}^*\tilde{\mathbf{B}} f) - 2\alpha(Y, \mathbf{\Lambda}^* Y')), \tag{3.8}$$

$$\hat{\mathbf{\Lambda}} \hat{f} = (\mathbf{A}\mathbf{\Lambda}\mathbf{B} f(x) - \alpha \mathbf{A}\mathbf{\Lambda} Y(x), 2(Y', \mathbf{\Lambda}\mathbf{B} f) - 2\alpha(Y', \mathbf{\Lambda} Y)). \tag{3.9}$$

From (3.2) it follows that

$$(\hat{\mathbf{\Lambda}}\hat{f},\hat{g})_1=(\hat{f},\hat{\mathbf{\Lambda}}^*\hat{g})_1 \quad \text{for } \hat{f}\in\mathcal{D}(\hat{\mathbf{\Lambda}}), \quad \hat{g}\in\mathcal{D}(\mathbf{\Lambda}^*),$$

where $\mathcal{D}(\hat{\mathbf{\Lambda}})=\hat{\mathbf{B}}\mathcal{D}(\mathbf{\Lambda})$ and $\mathcal{D}(\hat{\mathbf{\Lambda}}^*)=\hat{\hat{\mathbf{B}}}\mathcal{D}(\mathbf{\Lambda})$, i.e.,

$$\mathcal{D}(\hat{\mathbf{\Lambda}})=(\hat{f}\in\mathcal{N}_0^2 : f(0)=\alpha, f(x)\in C_0^2\cap L^1),$$

$$\mathcal{D}(\hat{\mathbf{\Lambda}}^*)=(\hat{f}\in\mathcal{N}_0^2 : f(0)=2(\gamma_1+\gamma_2)\alpha, f(x)\in C_0^2\cap L^1).$$

Now, having in mind Theorem 2.1, it is easy to obtain from Eqs. (3.3), (3.4) and decompositions (3.5), (3.6) the following theorem.

Theorem 3.1: *Let the operators* $\hat{\mathbf{\Lambda}}^*,\hat{\mathbf{\Lambda}}$ *be constructed as in (3.7). Then:* (i) *The function* $\hat{\hat{F}}(x,k)$ *defined by (1.9) is a solution of the equation*

$$\hat{\mathbf{\Lambda}}\hat{\hat{F}}(x,k)=k^2\hat{\hat{F}}(x,k)$$

and $\hat{\Psi}'(x,k)$ *defined by (1.10) is a solution of the equation*

$$\hat{\mathbf{\Lambda}}^*\hat{\Psi}'(x,k)=k^2\hat{\Psi}'(x,k)$$

with a boundary condition $\Psi'(0,k)=(\gamma_1+\gamma_2)\Psi(0,k)$, $(\Psi(0,k)=1)$.

(ii) *The expansion formulas (1.11) are decompositions of unity for the operators* $\hat{\mathbf{\Lambda}}^*$ *and* $\hat{\mathbf{\Lambda}}$, *i.e.,*

$$\hat{\mathbf{\Lambda}}^*\hat{f}=\int_0^\infty k^2\hat{\Psi}'(x,k)(\hat{f},\hat{\hat{F}}(k))_1 dk \quad \text{for } \hat{f}\in\mathcal{D}(\hat{\mathbf{\Lambda}}^*), \tag{3.10}$$

$$\hat{\mathbf{\Lambda}}\hat{f}=\int_0^\infty k^2\hat{\hat{F}}(k,x)(\hat{f},\hat{\Psi}'(k))_1 dk \quad \text{for } \hat{f}\in\mathcal{D}(\hat{\mathbf{\Lambda}}). \tag{3.11}$$

Next let us consider following Refs. 1 and 4 the operators

$$\hat{\mathbf{\Lambda}}_{\hat{v}}^*=\begin{pmatrix} \mathbf{\Lambda}_v^* & S^*(x) \\ -\frac{1}{4}\left.\frac{d}{dx}\right|_{x=0} & c^* \end{pmatrix}, \tag{3.12}$$

$$\hat{\mathbf{\Lambda}}_{\hat{v}}=\begin{pmatrix} \mathbf{\Lambda}_v & 0 \\ (\cdot,S^*)-\frac{1}{2}(\gamma_1+\gamma_2)\left.\frac{d}{dx}\right|_{x=0} & c^* \end{pmatrix}, \tag{3.13}$$

where $\mathbf{\Lambda}_v=\mathbf{\Lambda}_{x_0=\infty}$, $\mathbf{\Lambda}_v^*=(d/dx)\mathbf{\Lambda}_{x_0=0}\int_0^x$, $(r_n\to v_n)$ and $S^*(x)=\frac{1}{2}w(x)-\frac{1}{2}(\gamma_1-\gamma_2)\Delta(v;x)$, $c^*=\frac{1}{2}s(v;0)+\gamma_1\gamma_2$, $w(x)=s_x(v;x)-\Delta(v;x)\int_0^x\Delta(v;t)dt$. That is, for any function $\hat{f}\in\mathcal{N}_0^2(\mathcal{N}_\infty^2)$ we have

$$\hat{\mathbf{\Lambda}}_{\hat{v}}^*\hat{f}=(\mathbf{\Lambda}_v^*f(x)+\alpha S^*(x),\tfrac{1}{4}f'(0)+\alpha c^*)$$

and

$$\hat{\mathbf{\Lambda}}_{\hat{v}}\hat{f}=(\mathbf{\Lambda}_v f(x),(f,S^*)-\tfrac{1}{2}(\gamma_1+\gamma_2)f'(0)+\alpha c^*).$$

Then

$$\hat{\mathbf{\Lambda}}_{\hat{v}}^{*}\hat{f}=\int_0^\infty k^2\hat{\Psi}'(x,k)(\hat{f},\hat{\tilde{F}}(k))_1 dk \quad \text{for } \hat{f}\in\mathcal{D}(\hat{\mathbf{\Lambda}}^*), \tag{3.14}$$

$$\hat{\mathbf{\Lambda}}_{\hat{v}}\hat{f}=\int_0^\infty k^2\hat{\tilde{F}}(k,x)(\hat{f},\hat{\Psi}'(k))_1 dk \quad \text{for } \hat{f}\in\mathcal{D}(\hat{\mathbf{\Lambda}}). \tag{3.15}$$

Theorem 3.2: *Let the operators* $\hat{\mathbf{\Lambda}}^*,\hat{\mathbf{\Lambda}}$ *be constructed as in (3.7) and let the operators* $\hat{\mathbf{\Lambda}}_{\hat{v}}^*,\hat{\mathbf{\Lambda}}_{\hat{v}}$ *be defined as in (3.12), (3.13). Then:*

$$\hat{\mathbf{\Lambda}}^*\hat{f}=\hat{\mathbf{\Lambda}}_{\hat{v}}^*\hat{f} \quad \text{for } \hat{f}\in\mathcal{D}(\hat{\mathbf{\Lambda}}^*), \quad \hat{\mathbf{\Lambda}}\hat{f}=\hat{\mathbf{\Lambda}}_{\hat{v}}\hat{f} \quad \text{for } \hat{f}\in\mathcal{D}(\hat{\mathbf{\Lambda}}), \tag{3.16}$$

where the explicit form of the operators $\hat{\mathbf{\Lambda}}$ *and* $\hat{\mathbf{\Lambda}}^*$ *in terms of* $v_n(x),\gamma_n$ *is given by the expressions*

$$\hat{\mathbf{\Lambda}}^*=\begin{pmatrix} \mathbf{\Lambda}_v^* & S^*(x) \\ -\frac{1}{4}\frac{d}{dx}\Big|_{x=0}+\frac{1}{4}(\gamma_1+\gamma_2)\Big|_{x=0} & c^*-\frac{1}{2}(\gamma_1+\gamma_2)^2 \end{pmatrix}, \tag{3.17}$$

$$\hat{\mathbf{\Lambda}}=\begin{pmatrix} \mathbf{\Lambda}_v & 0 \\ (\cdot,S^*)-\frac{1}{2}(\gamma_1+\gamma_2)\frac{d}{dx}\Big|_{x=0}+(c^*+\tau_0^2)|_{x=0} & -\tau_0^2 \end{pmatrix}. \tag{3.18}$$

Proof: By comparing (3.14), (3.15) with (3.10), (3.11) we obtain straightforwardly (3.16). In view of (2.5), Eq. (3.1) yields the equation

$$\hat{\mathbf{\Lambda}}Y(x)=-\tau_0^2 Y(x).$$

As a result, we find that

$$-\mathbf{A\Lambda}Y(x)=0, \quad -2(\mathbf{\Lambda}Y,Y')=-\tau_0^2, \quad -2(Y,\mathbf{\Lambda}^*Y')=\tfrac{1}{2}s(v;0)-\tfrac{1}{2}(\gamma_1^2+\gamma_2^2),$$

$$-(Y,\mathbf{\Lambda}^*\tilde{\mathbf{B}}g)=-\tfrac{1}{4}g'(0)+\tfrac{1}{4}(\gamma_1+\gamma_2)g(0),$$

and

$$2(Y',\mathbf{AB}f)=2(f,\tilde{\mathbf{A}}\mathbf{\Lambda}^*Y')-\tfrac{1}{2}(\gamma_1+\gamma_2)f'(0)+(\tau_0^2+\tfrac{1}{2}s(v;0)+\gamma_1\gamma_2)f(0).$$

By inserting these equalities in (3.8), (3.9) we obtain that if $g(0)=2(\gamma_1+\gamma_2)\beta$ then

$$-(Y,\mathbf{\Lambda}^*\tilde{\mathbf{B}}g)-2\beta(Y,\mathbf{\Lambda}^*Y'))=-\tfrac{1}{4}g'(0)+c^*\beta$$

and if $f(0)=\alpha$ then

$$2(Y',\mathbf{\Lambda B}f)-2\alpha(Y',\mathbf{\Lambda}Y))=(f,S^*)-\tfrac{1}{2}(\gamma_1+\gamma_2)f'(0)+c^*f(0).$$

Thus we obtain $S^*(x)=2\tilde{\mathbf{A}}\mathbf{\Lambda}^*Y'(x)$ since $f(x)$ is an arbitrary function in C_0^2. From here, in view of (3.16), we conclude due to the arbitrariness of $f(x)$ that $\Lambda_v^*=\tilde{\mathbf{A}}\mathbf{\Lambda}^*\tilde{\mathbf{B}}$, $\mathbf{\Lambda}_v=\mathbf{A\Lambda B}$. □

Remark 3.1: From the above Theorem one can see that the relations $\hat{\Lambda}\hat{f}=\hat{\Lambda}_{\hat{v}}\hat{f}$ and $\hat{\Lambda}^*\hat{f}=\hat{\Lambda}_{\hat{v}}^*\hat{f}$ do not hold in general (for an arbitrary function $\hat{f}$ in $\mathcal{N}_0^2$) since from (3.17), (3.18) it follows that the necessary and sufficient conditions for that are $-\tau_0^2=\frac{1}{2}s(v;0)+\gamma_1\gamma_2$, $\gamma_1+\gamma_2=0$.

IV. APPLICATIONS

In this section first we show that the operators generating the transforms in Sec. II are functional derivative of the operator (1.13),

$$\mathcal{F}(\hat{v};x) \overset{\text{def}}{=} v(x) - 2\frac{d^2}{dx^2}\ln z(x) = r(x), \tag{4.1}$$

where $z(x) = \psi(\hat{v};x,k_0) = \psi(x,k_0)$. The potential $\hat{v} = (v(x),\gamma)$, which determines the problem (1.1), (1.2) is considered as an element in the space $\mathcal{N}_1$, the potential $r(x)$ in (1.3) as an element in X_1 and $\mathcal{F}$ as an operator $\mathcal{F}:\mathcal{N}_1 \to X_1$. According to Lemma 2.2, the inverse to $\mathcal{F}(\hat{v};x)$ operator $\mathcal{G}(r;x):X_1 \to \mathcal{N}_1$, is given by the expression

$$\mathcal{G}(r;x) = \left(r(x) - 2\frac{d^2}{dx^2}\ln y(r;x), -\frac{d}{dx}\ln y(r;x)\bigg|_{x=0} \right), \tag{4.2}$$

where $y(r;x) = \psi^{-1}(\hat{v};x,k_0)$. The main result is the following.

Theorem 4.1: *At any point $\hat{v} \in \mathcal{N}_1$ there exists the derivative:*

$$\frac{d}{d\epsilon}\mathcal{F}(\hat{v}+\epsilon\hat{g};x)|_{\epsilon=0} = \frac{\partial \mathcal{F}}{\partial \hat{v}}\hat{g} = -\hat{\hat{\mathbf{B}}}\hat{g} \quad (\hat{g} = (g(x),\beta)), \tag{4.3}$$

where $\partial\mathcal{F}/\partial\hat{v} = (\partial\mathcal{F}/\partial v, \partial\mathcal{F}/\partial\gamma)$ and $\hat{\hat{\mathbf{B}}}$ is defined by (2.10) for $\hat{v}_1 = \hat{v}_2 = \hat{v}$. The operator $\mathcal{G}$ is differentiable at any point $r \in X_1$ and its derivative is

$$\frac{d}{d\epsilon}\mathcal{G}(r+\epsilon f;x)|_{\epsilon=0} = -\hat{\hat{\mathbf{A}}}f, \tag{4.4}$$

where $\hat{\hat{\mathbf{A}}}$ is defined by (2.8) with $r_1 = r_2 = r$.

Proof: Equation (4.1) yields

$$\frac{d}{d\epsilon}\mathcal{F}(v+\epsilon g,\gamma+\epsilon\beta;x)|_{\epsilon=0} = g(x) - 2\frac{d^2}{dx^2}\left(\psi^{-1}(\hat{v};x)\frac{\partial\psi(\hat{v};x)}{\partial v}g(x) \right) - \frac{d^2}{dx^2}\left(\psi^{-1}(\hat{v};x)\frac{\partial\psi(\hat{v};x)}{\partial\gamma}\beta \right). \tag{4.5}$$

In a standard way one obtains for any fixed k_0 and γ the following expression for the derivative:

$$\frac{\partial\psi(\hat{v};x,k_0)}{\partial v}g(x) = \psi(x,k_0)\int_0^x \psi^{-2}(t,k_0)\left(\int_0^t \psi^2(z,k_0)g(z)dz \right)dt.$$

Therefore,

$$-2\frac{d^2}{dx^2}\left(\psi^{-1}(x,k_0)\frac{\partial\psi(x,k_0)}{\partial v}g(x) \right) = -2Y'(x)\int_0^x Y^{-1}(t)g(t)dt - 2g(x). \tag{4.6}$$

In order to obtain

$$-2\frac{d^2}{dx^2}\left(\psi^{-1}(\hat{v};x,k_0)\frac{\partial\psi(\hat{v};x,k_0)}{\partial\gamma} \right) = -2Y'(x) \tag{4.7}$$

it remains to mention that $\partial\psi(\hat{v};x,k_0)/\partial\gamma=\psi(x,k_0)\int_0^x\psi^{-2}(t,k_0)dt$. Now we derive (4.3) by inserting (4.6) and (4.7) into the right-hand side of (4.5). The representation (4.4) is obtained in a similar manner. □

Let us denote by $\eta(\hat{v};k)$ the phase shift of the problem (1.1), (1.2) where $e(\hat{v};k)=|e(\hat{v};k)|\exp(i\eta(\hat{v};k))$ and by $\delta(r;k)$ the phase shift of the problem (1.3), (1.4) where $h(r;k)=|h(r;k)|\exp(i\delta(r;k))$. Then from (2.4) it follows that

$$\delta(\mathcal{F}(\hat{v});k)=\arctan\frac{k}{\tau_0}+\eta(\hat{v};k),\quad 0<k<\infty. \tag{4.8}$$

and

$$\rho(\mathcal{F}(\hat{v});k)=(k^2-k_0^2)\sigma(\hat{v};k), \tag{4.9}$$

where $\sigma(\hat{v};k)=(2/\pi)k^2|e(\hat{v};k)|^{-2}$ and $\rho(r;k)=(2/\pi)k^2|h(r;k)|^{-2}$ are the spectral density of the problems (1.1), (1.2) and (1.3), (1.4), respectively.

Now we recall (see, e.g., Ref. 24) that at any point $r\in X_1$ where $h(r;0)\neq 0$ there exists the derivative

$$\frac{\partial\delta(r;k)}{\partial r(x)}=\frac{k}{|H(r;k)|}\Phi(r;x,k). \tag{4.10}$$

Let us differentiate (4.8) with respect to $\hat{v}$. According to (4.3) we have

$$\left(\frac{\partial\delta(r;k)}{\partial r},\frac{\partial\mathcal{F}}{\partial\hat{v}}\hat{g}\right)=\left(\frac{\partial\eta(\hat{v};k)}{\partial\hat{v}},\hat{g}\right)_1, \tag{4.11}$$

where $\partial\eta(\hat{v};k)/\partial\hat{v}=(\partial\eta(\hat{v};k)/\partial v(x),\partial\eta(\hat{v};k)/\partial\gamma)$. By using the representations (2.12) and (4.9) we find that $\partial\eta(\hat{v};k)/\partial\hat{v}=k|E(\hat{v};k)|^{-1}(\Psi(\hat{v};x,k),1)$. Now let us replace $f(x)$ in the expansion (1.8) by $f(x)=\hat{\tilde{\mathbf{B}}}\hat{v}_t(x,t)$ where $\hat{v}(t)=(v(x,t),\gamma(t))$ is the family of differentiable in t potentials such that $\hat{v}_t=(v_t(x,t),\gamma_t(t))\in\mathcal{N}_1$. Having in mind (4.11) we obtain

$$\hat{\tilde{\mathbf{B}}}\hat{v}_t(x,t)=\int_0^\infty\chi'(x,k)\eta_t(\hat{v};k)dk, \tag{4.12}$$

where $\chi(x,k)=k^{-1}|H(k)|\tilde{H}(x,k)$, $\eta_t(\hat{v};k)=(\partial\eta(\hat{v};k)/\partial\hat{v},\hat{v}_t)_1$. The condition $r_t(x,t)=-\hat{\tilde{\mathbf{B}}}\hat{v}_t(x,t)\in L_1^0$ yields $I_{1,t}(\hat{v}(t))=0$, $I_1(\hat{v})=\gamma+\frac{1}{2}\int_0^\infty v(x)dx$. Note that I_1 determines[14] the asymptotics $\eta(\hat{v};k)=k^{-1}I_1(\hat{v})+o(k^{-1})$ as $k\to\infty$. Now we put $x=0$ in (4.12) and use (1.6) to obtain the trace formula

$$\frac{d}{dt}\{v(0,t)-2\gamma^2(t)\}=\frac{8}{\pi}\int_0^\infty k\eta_t(\hat{v};k)dk. \tag{4.13}$$

Note that (4.13) could be obtained directly from the well-known Faddeev–Newton[26,27] trace formula

$$r_t(0,t)=\frac{8}{\pi}\int_0^\infty k\delta_t(r;k)dk \tag{4.14}$$

by implementing (4.1). Let us here also make the remark that (4.14) follows from the expansion (1.8) with $f(x)=r_t(x,t)$, boundary conditions (1.6) and the expression (4.10) for $\partial\delta(r;k)/\partial r(x)$ in contrast to the formula (4.13) which is not a direct consequence of (1.12).

V. TRANSFORMS IN THE CASE OF A DISCRETE SPECTRUM

In this section is given a generalization of the scheme for deriving transforms proposed in Sec. II to the general case when we have a discrete spectrum. For the problem (1.3), (1.4) it is defined by the zeros of the characteristic function $h_n(k)$ at $\operatorname{Im} k>0$. We put

$$\sigma_n=\{k_{n,j}:h_n(k_{n,j})=0,k_{n,j}=i\tau_{n,j},\tau_{n,j}>0,\quad j=1,2,...,N_n\}$$

and $\sigma=\sigma_1\cup\sigma_2$, $\sigma''=\sigma_1\cap\sigma_2$, $\sigma'=\sigma\backslash\sigma''$. If $k_{n,j}\in\sigma'$ then we put

$$\tilde{H}_{n,j}(x)=-m_{n,j}H(x,k_{n,j}),\quad m_{n,j}=4k_{n,j}\dot{H}^{-1}(k_{n,j}),\quad (\dot{}=\partial/\partial k)$$

and with each $k_j\in\sigma''$ we associate a pair of functions

$$\tilde{H}_{j,1}(x)=-p_j(\dot{H}(x,k_j)+g_jH(x,k_j)),\tilde{H}_{j,2}(x)=-p_jH(x,k_j),$$

where $p_j=8k_j\ddot{H}^{-1}(k_j)$, $q_j=k_j^{-1}-\dddot{H}(k_j)(3\ddot{H}(k_j))^{-1}$. The analog of (1.7) here is the following expansion formula:[1]

$$f(x)=\int_0^\infty\tilde{H}(x,k)(f,\Phi'(k))dk+\Sigma_{k_{n,j}\in\sigma'}\tilde{H}_{n,j}(x)(f,\Phi'(k_{n,j}))$$
$$+\Sigma_{k_j\in\sigma''}\{\tilde{H}_{j,1}(x)(f,\Phi'(k_j))+\tilde{H}_{j,2}(x)(f,\dot{\Phi}'(k_j))\},\qquad(5.1)$$

where $f(x)$ is an absolutely continuous function in X_1.

In a similar way, for the problems (1.1), (1.2) we denote, as above; $\sigma_n=\{k_{n,j}:e_n(k_{n,j})=0,k_{n,j}=i\tau_{n,j},\tau_{n,j}>0,\quad j=1,2,...,N_n\}$ and $\sigma=\sigma_1\cup\sigma_2$, $\sigma''=\sigma_1\cap\sigma_2$, $\sigma'=\sigma\backslash\sigma''$. Note that under the condition (1.13) the spectrum σ_n of (1.1), (1.2) coincides with the spectrum σ_n of (1.3), (1.4) which can be seen from (2.4). Here we put

$$\hat{\tilde{F}}_{n,j}(x)=-a_{n,j}\hat{F}(x,k_{n,j}),\quad a_{n,j}=4k_{n,j}\dot{E}^{-1}(k_{n,j})\quad(k_{n,j}\in\sigma')$$

and

$$\hat{\tilde{F}}_{j,1}(x)=-b_j(\dot{\hat{F}}(x,k_j)+d_j\hat{F}(x,k_j)),\quad \hat{\tilde{F}}_{j,2}(x)=-b_j\hat{F}(x,k_j),$$

where $b_j=8k_j\ddot{E}^{-1}(k_j)$, $d_j=k_j^{-1}-\dddot{E}(k_j)(3\ddot{E}(k_j))^{-1}(k_j\in\sigma'')$.

The analog of (1.12) is the following expansion formula:[4]

$$\hat{f}=\int_0^\infty\hat{\tilde{F}}(x,k)(\hat{f},\hat{\Psi}'(k))_1dk+\Sigma_{k_{n,j}\in\sigma'}\hat{\tilde{F}}_{n,j}(x)(\hat{f},\hat{\Psi}'_{n,j})_1$$
$$+\Sigma_{k_j\in\sigma''}\{\hat{\tilde{F}}_{j,1}(x)(\hat{f},\hat{\Psi}'(k_j))_1+\hat{\tilde{F}}_{j,2}(x)(\hat{f},\dot{\hat{\Psi}}'(k_j))_1\},\qquad(5.2)$$

where $\hat{f}=(f(x),\alpha)\in\mathcal{N}_1$.

Here we sketch only the proof of the transition $(5.1)\rightarrow(5.2)$. Since the transform related to the continuous spectrum has already been obtained in Sec. II, here it is sufficient to consider only the transform of the expansion

$$f(x)=\tilde{H}_{j,1}(x)(f,\Phi'(k_j))+\tilde{H}_{j,2}(x)(f,\dot{\Phi}'(k_j))+\tilde{H}_{n,l}(x)(f,\Phi'(k_{n,l}))\qquad(5.3)$$

in order to derive (5.2) starting from (5.1). By using equality (2.4) where $\tau_0>\tau_{n,j}\tau_{n,j}\in\sigma$ we obtain that the coefficients $a_{n,j}=-(k_0+k_{n,j})^{-2}m_{n,j}$ $(k_{n,j}\in\sigma')$ and $b_j=-(k_0+k_j)^{-2}p_j$, $d_j=q_j-2(k_0+k_j)^{-1}(k_j\in\sigma'')$. Let us put in (5.3) $f=\hat{\mathbf{B}}\hat{g}$, $\hat{g}=(g(x),\beta)\in\mathcal{N}_1$. It follows from (2.13) that

$$(\hat{\mathbf{B}}\hat{g},\Phi'(k_{n,j}))=(k_0^2-k_{n,j}^2)^{-1}(\hat{g},\hat{\Psi}'(k_{n,j}))_1 \quad (k_{n,j}\in\sigma)$$

and

$$(\hat{\mathbf{B}}\hat{g},\dot{\Phi}'(k_j))=(k_0^2-k_j^2)^{-1}(\hat{g},\dot{\hat{\Psi}}'(k_j))_1+2k_j(k_0^2-k_j^2)^{-2}(\hat{g},\hat{\Psi}'(k_j))_1, \quad (k_j\in\sigma'').$$

Therefore, we obtain

$$\hat{\mathbf{B}}\hat{g}=\frac{k_0+k_j}{k_0-k_j}b_j\left(\dot{H}(x,k_j)+\left(d_j+\frac{2k_j}{k_0^2-k_j^2}\right)H(x,k_j)\right)(\hat{g},\hat{\Psi}'(k_j))_1$$
$$+\frac{k_0+k_j}{k_0-k_j}b_jH(x,k_j)(\hat{g},\dot{\hat{\Psi}}'(k_j))_1+\frac{k_0+k_{n,j}}{k_0-k_{n,l}}a_{n,l}H(x,k_{n,l})(\hat{g},\hat{\Psi}'(k_{n,l}))_1.$$

Finally, by applying the operator $\hat{\mathbf{A}}$, in view of (2.11) and (2.12) we come to

$$\hat{g}=\hat{\tilde{F}}_{j,1}(x)(\hat{g},\hat{\Psi}'(k_j))_1+\hat{\tilde{F}}_{j,2}(x)(\hat{g},\dot{\hat{\Psi}}'(k_j))_1+\hat{\tilde{F}}_{n,l}(x)(\hat{g},\hat{\Psi}'_{n,l})_1.$$

The case (5.2)→(5.1) can be treated just as above and so we omit the proof here.

ACKNOWLEDGMENT

This work is partially supported by the Bulgarian Ministry of Education and Science under Contract No. MM 608/96.

[1] I. D. Iliev, E. Kh. Khristov, and K. P. Kirchev, *Spectral Methods in Soliton Equations* (Pitman Monographs and Surveys in Pure and Applied Mathematics, London, 1994), Vol. 73.
[2] E. Kh. Khristov and Y. P. Mishev, Ann. de L'Univ. Sofia, Fac. Math. and Inf., Livre 2-Mechanique **82**, 219 (1988).
[3] Y. P. Mishev, Inverse Probl. **7**, 379 (1991).
[4] E. Kh. Khristov, Inverse Probl. **14**, 647 (1998).
[5] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[6] V. A. Arkad'ev, A. K. Pogrebkov, and M. K. Polivanov, Teor. Mat. Fiz. **72**, 323 (1987) (Theor. Math. Phys.).
[7] F. Calogero and A. Degasperis, *Spectral Transform and Solitons* (North-Holland, Amsterdam, 1982), Vol. I.
[8] V. E. Zakharov and L. D. Faddeev, Funct. Anal. Appl. **5**, 332 (1971).
[9] G. Borg, Acta Math. **78**, 1 (1946).
[10] V. Barcelon, J. Math. Phys. **15**, 429 (1974).
[11] D. J. Kaup, J. Math. Anal. Appl. **54**, 849 (1976).
[12] R. L. Sachs, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **14**, 674 (1983).
[13] M. G. Gasymov and B. M. Levitan, DAN USSR **310**, 781 (1990).
[14] B. M. Levitan, Mat. Zametki **17**, 611 (1975).
[15] B. M. Levitan, *Inverse Sturm–Liouville Problems* (Nauka, Moscow, 1984).
[16] V. F. Korop, Sibirskii Math. J. **11**, 672 (1981).
[17] V. A. Marchenko, *Sturm–Liouville Operators and Applications* (Naukova Dumka, Kiev, 1977).
[18] M. M. Crum, Q. J. Math. **6**, 121 (1955).
[19] M. G. Krein, DAN USSR **113**, 970 (1957).
[20] E. L. Isaacson, H. P. McKean, and E. Trubowitz, Commun. Pure Appl. Math. **37**, 1 (1984).
[21] J. Pöschel and E. Trubowitz, *Inverse Spectral Theory*, Pure and Applied Mathematics Vol. 130 (Academic, New York, 1987).
[22] L. D. Faddeev, UMN **14**, 57 (1959).
[23] P. A. Deift, Duke Math. J. **45**, 267 (1978).
[24] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. (Springer, Berlin, 1989).
[25] F. Calogero, *Stud. Math. Phys. (Essays in Honor of V. Bargman)*, edited by E. H. Lieb, B. Simon, and A. S. Wightman (Princeton University Press, Princeton, NJ, 1976), p. 23.
[26] L. D. Faddeev, DAN USSR **115**, 878 (1957).
[27] R. G. Newton, Phys. Rev. **101**, 1588 (1956).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 6 JUNE 1999

Geometry and representations of the quantum supergroup $\mathrm{OSP}_q(1|2n)$

H. C. Lee
Department of Physics and Center for Complex Systems, National Central University, Chungli, Taiwan, Republic of China

R. B. Zhang
Department of Pure Mathematics, University of Adelaide, Adelaide, Australia

(Received 8 June 1998; accepted for publication 4 November 1998)

The quantum supergroup $\mathrm{OSP}_q(1|2n)$ is studied systematically. A Haar functional is constructed, and an algebraic version of the Peter–Weyl theory is extended to this quantum supergroup. Quantum homogeneous superspaces and quantum homogeneous supervector bundles are defined following the strategy of Connes' theory. Parabolic induction is developed by employing the quantum homogeneous supervector bundles. Quantum Frobenius reciprocity and a generalized Borel–Weil theorem are established for the induced representations. © *1999 American Institute of Physics.* [S0022-2488(99)00205-4]

I. INTRODUCTION

Quantized universal enveloping algebras of Lie superalgebras were introduced in the late 1980s[1,2] to describe the type of supersymmetries exhibited by some two-dimensional statistical mechanics models.[3] Since then these quantum superalgebras have been intensively studied, leading to the development of an extensive theory on both the structure and representations. We mention in particular that the quasi-triangular Hopf superalgebraic structure of the quantum superalgebras was investigated in Ref. 4; the representation theory of the type I quantum superalgebras, the $\mathrm{gl}(m|n)$ super Yangians and the quantum affine superalgebras with symmetrizable Cartan matrices were developed in Ref. 5. The theory of quantum superalgebras had significant impact on a range of areas of physics and mathematics. Its applications to two-dimensional integrable models in statistical mechanics and quantum field theory were extensively explored in Refs. 1 and 6 and many other publications. The application to knot theory and three-manifolds[7,8] has yielded many new topological invariants, notably, the multi-parameter generalizations of Alexander–Conway polynomials.

The associated quantum supergroups are in contrast less studied in the literature. So far only the quantum supergroup $\mathrm{GL}_q(m|n)$ has been systematically investigated.[9] In Ref. 9, the structure and representation theories of $\mathrm{GL}_q(m|n)$ were developed. The irreducible covariant and contravariant tensorial representations were studied in detail within the framework of parabolic induction, resulting in a quantum Borel–Weil theorem for these representations. The aim of this paper is to treat the $\mathrm{osp}(1|2n)$ series of quantum supergroups at generic q.

The $\mathrm{osp}(1|2n)$ series of Lie superalgebras played an important role in the study of supersymmetry on de Sitter space.[10] These Lie superalgebras, especially $\mathrm{osp}(1|32)$, also featured prominently in recent developments of string theory. An Inonu–Wigner contraction of $\mathrm{osp}(1|32)$ yields the 11-dimensional Poincaré superalgebra with two and five form central charges, which is the underlying symmetry of M theory; the superalgebra $\mathrm{osp}(1|32)$ itself also plays an important role in the theory of supermembranes.[11] From a mathematical point of view, $\mathrm{osp}(1|2n)$ is also rather exceptional amongst all the finite-dimensional simple Lie superalgebras in that its Cartan matrix is symmetrizable, and the structure of its finite-dimensional representations is completely understood. In particular, it is known that all finite-dimensional representations are completely reducible.

Many properties of $\mathrm{osp}(1|2n)$ carry over to the quantum case when q is generic. It is particularly useful to recall that the Drinfeld version of $\mathrm{U}_q(\mathrm{osp}(1|2n))$ is, algebraically, a trivial deformation of $\mathrm{U}(\mathrm{osp}(1|2n))$ in the sense of Gerstenhaber. (This fact is known to experts, and may be easily inferred from results of Ref. 12.) Therefore, *finite-dimensional representations* of $\mathrm{U}_q(\mathrm{osp}(1|2n))$ *are also completely reducible*. This remains true for the Jimbo version of $\mathrm{U}_q(\mathrm{osp}(1|2n))$ at generic q. One way to see this is through the specialization of the indeterminate of the Drinfeld algebra to a generic complex parameter; the other is through the isomorphism between $\mathrm{U}_q(\mathrm{osp}(1|2n))$ and $U_{-q}(\mathrm{so}(2n+1))$ established by a kind of Bose–Fermi transmutation.[13] There is also an interesting connection between the representation theory of $\mathrm{U}_q(\mathrm{osp}(1|2n))$ and quantum para-statistics, details on which can be found in Ref. 14.

This paper will study structural and representation theoretical properties of the quantum supergroup $\mathrm{OSP}_q(1|2n)$, and also investigate its underlying geometries. This quantum supergroup will be defined by its superalgebra of functions, which is the $\mathbf{Z}_2$-graded Hopf algebra generated by the matrix elements of the vector representation of $\mathrm{U}_q(\mathrm{osp}(1|2n))$. Two major results in the structure theory are presented, namely, the existence of a left and right integral, which will be called a quantum Haar functional, and a quantum Peter–Weyl theorem.

Corresponding to each reductive subalgebra $\mathrm{U}_q(\mathbf{k})$ of $\mathrm{U}_q(\mathrm{osp}(1|2n))$, we introduce a quantum homogeneous superspace, which is defined by specifying its superalgebra of functions $\mathcal{A}_q^{\mathbf{k}}$. A quantum homogeneous supervector bundle over the quantum homogeneous superspace is induced from any given finite-dimensional $\mathrm{U}_q(\mathbf{k})$ module. We shall show that the space of sections $\Gamma_q^{\mathbf{k}}(V)$ of this bundle is projective and is of finite type both as a left and a right module over $\mathcal{A}_q^{\mathbf{k}}$. Therefore our definition of quantum homogeneous supervector bundles is consistent with the general definition of noncommutative vector bundles in Connes' theory.[15]

Quantum homogeneous supervector bundles will be applied to develop a theory of induced representations for $\mathrm{OSP}_q(1|2n)$. Amongst the results obtained are quantum versions of Frobenius reciprocity and the Borel–Weil theorem. The latter provides a concrete realization of finite-dimensional irreducible $\mathrm{OSP}_q(1|2n)$ representations in terms of quantum analogs of ''holomorphic'' sections of quantum homogeneous supervector bundles.

We wish to point out that in the context of Lie supergroups at the classical level, the mathematical theories of homogeneous superspaces and homogeneous supervector bundles were studied in Refs. 16 and 17. The development of a Bott–Borel–Weil theory was also initiated and extensively investigated by Penkov and co-workers.[17] However, complications arising from supermanifold geometry render these subjects very difficult to study. So far as we are aware, many aspects of the subjects remain to be fully developed. It seems that the Hopf algebraic approach developed here and in Ref. 9 is also worth exploring at the classical level, and is likely to provide a new method complementary to the geometric approach of Refs. 16 and 17.

The organization of the paper is as follows. In Sec. II we review some known facts about $\mathrm{U}_q(\mathrm{osp}(1|2n))$, which will be needed later. In Sec. III we study the quantum supergroup $\mathrm{OSP}_q(1|2n)$. In Sec. IV we investigate the quantum homogeneous superspaces and quantum homogeneous supervector bundles determined by this quantum supergroup, while the last section applies results of Sec. IV to study the representation theory of $\mathrm{OSP}_q(1|2n)$.

II. $\mathrm{U}_q(\mathrm{osp}(1|2n))$

This section reviews some known results on the quantized universal enveloping algebra $\mathrm{U}_q(\mathrm{osp}(1|2n))$. Let E be the n-dimensional Euclidean space spanned by the vectors ϵ_i, with the inner product (,) defined by $(\epsilon_i,\epsilon_j)=\delta_{ij}$. We can express the simple roots of the Lie superalgebra $\mathrm{osp}(1|2n)$ in terms of the ϵ's as

$$\alpha_i=\epsilon_i-\epsilon_{i+1},\quad i=1,2,\dots,n-1,\quad \alpha_n=\epsilon_n,$$

where α_n is the odd simple root. The Cartan matrix $A=(a_{ij})_{i,j=1}^n$ of $\mathrm{osp}(1|2n)$ is then given by $a_{ij}=2(\alpha_i,\alpha_j)/(\alpha_i,\alpha_i)$. An element $\mu\in E$ will be called integral if

$$l_i = \frac{2(\mu,\alpha_i)}{(\alpha_i,\alpha_i)} \in \mathbf{Z}, \quad \forall i<n, \quad l_n = \frac{(\mu,\alpha_n)}{(\alpha_n,\alpha_n)} \in \mathbf{Z},$$

and the set of all integral elements will be denoted by $\mathcal{P}$. (Note the unusual form of l_n.) Set $\mathcal{P}_+ = \{\mu \in \mathcal{P} | l_i, l_n \in \mathbf{Z}_+\}$. Elements of $\mathcal{P}_+$ will be called integral dominant.

The Jimbo version of the quantum superalgebra $\mathrm{U}_q(\mathrm{osp}(1|2n))$ is a $\mathbf{Z}_2$-graded complex associative algebra generated by $\{k_i^{\pm 1}, e_i, f_i, i \in \mathbf{N}_n\}$, $\mathbf{N}_n = \{1,2,...,n\}$, subject to the relations

$$\begin{gathered} k_i k_i^{-1} = 1, \quad k_i k_j = k_j k_i, \\ k_i e_j = q^{(\alpha_i,\alpha_j)} e_j k_i, \quad k_i f_j = q^{-(\alpha_i,\alpha_j)} f_j k_i, \\ [e_i, f_j\} = \delta_{ij} \frac{k_i - k_i}{q-q}, \quad \forall i,j \in I, \\ (\mathrm{Ad}\, e_i)^{1-a_{ij}}(e_j) = 0, \quad (\mathrm{Ad}\, f_i)^{1-a_{ij}}(f_j) = 0, \quad \forall i \neq j. \end{gathered} \tag{1}$$

All the generators are chosen to be homogeneous, with $k_i^{\pm 1}$, $\forall i$, and e_j, f_j, $j<n$, being even, and e_n, f_n being odd. For a homogeneous element x, we define $[x]=0$ if x is even, and $[x]=1$ when odd. The graded commutator $[.,.\}$ represents the usual commutator when any one of the two arguments is even, and the anticommutator when both arguments are odd. The adjoint operation Ad is defined by

$$\begin{gathered} \mathrm{Ad}\, e_i(x) = e_i x - (-1)^{[e_i][x]} k_i x k_i^{-1} e_i, \\ \mathrm{Ad}\, f_i(x) = f_i x - (-1)^{[f_i][x]} k_i x k_i^{-1} f_i. \end{gathered}$$

For x being a monomial in e_j's or f_j's it carries a definite weight $\omega(x) \in H^*$. Then $\mathrm{Ad}\, e_i(x) = e_i x - (-1)^{[e_i[x]} q^{(\alpha_i,\omega(x))} x e_i$, and similarly for $\mathrm{Ad}\, f_i(x)$. For convenience, we will use the notation $\mathbf{g}$ to denote $\mathrm{osp}(1|2n)$, and $\mathrm{U}_q(\mathbf{g})$ to denote $\mathrm{U}_q(\mathrm{osp}(1|2n))$. As is well known, this algebra has the structures of a $\mathbf{Z}_2$-graded Hopf algebra. We will denote the comultiplication by Δ, the counit by ϵ, and the antipode by S.

The representation theory of $\mathrm{U}_q(\mathbf{g})$ was developed in Ref. 13. For any finite-dimensional $\mathrm{U}_q(\mathbf{g})$ module, there exists a homogeneous basis relative to which the k_i are represented by diagonal matrices. Here we will only consider such finite-dimensional $\mathrm{U}_q(\mathbf{g})$ modules that the eigenvalues of the k_i tend to 1 as q approaches 1. We will denote the set of all such $\mathrm{U}_q(\mathbf{g})$ modules by $\mathbf{Mod}_q(\mathbf{g})$. Recall that all objects of $\mathbf{Mod}_q(\mathbf{g})$ are semi-simple.

If $W(\lambda)$ is a simple object of $\mathbf{Mod}_q(\mathbf{g})$, then there exists the unique (up to scalar multiples) highest weight vector v_+, such that

$$e_i v_+ = 0, \quad k_i v_+ = q^{(\lambda,\alpha_i)} v_+, \quad \lambda \in \mathcal{P}_+,$$

and the module $W(\lambda)$ is uniquely determined by the highest weight λ. We will denote the lowest weight of $W(\lambda)$ by $\overline{\lambda}$, and define $\lambda^\dagger = -\overline{\lambda}$. The dual module of $W(\lambda)$ has highest weight $\lambda^\dagger$.

The irreducible $\mathrm{U}_q(\mathbf{g})$ module with highest weight ϵ_1 plays a special role in the representation theory of $\mathrm{U}_q(\mathbf{g})$. We denote this module by $\mathbf{E}$, and refer to it as the vector module. Let us now examine this module in some detail. Denote by w_1 the highest weight vector of $\mathbf{E}$, which is assumed to be even. Define

$$\begin{gathered} w_i = f_{i-1} w_{i-1}, \quad 1 < i \leq n, \\ w_0 = f_n w_n, \quad w_{-n} = f_n w_n, \\ w_{-j} = f_j w_{-j-1}, \quad n > j \geq 1. \end{gathered}$$

Then $\{w_\mu|\mu=0,\pm1,\pm2,...,\pm n\}$ forms a weight basis of $\mathbf{E}$. We will denote by t the irreducible representation relative to this basis. The matrix elements of the e_i, f_i and k_i can be immediately written down. We have

$$t(e_i)_{\mu\nu}=\delta_{\mu i}\delta_{\nu,i+1}+\delta_{\mu,-i-1}\delta_{\nu,-i},$$

$$t(f_i)_{\mu\nu}=\delta_{\mu,i+1}\delta_{\nu i}+\delta_{\mu,-i}\delta_{\nu,-i-1},\quad i<n,$$

$$t(e_n)_{\mu\nu}=\delta_{\mu n}\delta_{\nu 0}-\delta_{\mu 0}\delta_{\nu,-n},$$

$$t(f_n)_{\mu\nu}=\delta_{\mu 0}\delta_{\nu n}+\delta_{\mu,-n}\delta_{\nu 0},$$

$$t(k_j)_{\mu\nu}=\delta_{\mu\nu}q^{(\alpha_j,\epsilon_\mu)},\quad 1\leqslant j\leqslant n,$$

where $\epsilon_0=0$, and $\epsilon_{-i}=-\epsilon_i$.

Let $\{w_\mu^*\}$ be the basis of $\mathbf{E}^*$ defined by $w_\mu^*(w_\nu)=\delta_{\mu\nu}$. Here $\mathbf{E}^*$ has a natural $\mathrm{U}_q(\mathbf{g})$-module structure with the $\mathrm{U}_q(\mathbf{g})$ action given by

$$xw_\mu^*=\sum_\nu (-1)^{[x]\delta_{\mu 0}}t(S(x))_{\mu\nu}w_\nu^*. \tag{2}$$

The lowest weight of $\mathbf{E}$ is $-\epsilon_1$. Thus the module $\mathbf{E}$ is self-dual. This implies that there exists a $\mathrm{U}_q(\mathbf{g})$-module isomorphism $M:\mathbf{E}\rightarrow\mathbf{E}^*$, which is unique up to scaler multiples. The w_{-1}^*, being the highest weight vector of $\mathbf{E}^*$, will be identified with w_1 so that this arbitrariness in M can be removed. Now let

$$w_\mu^*=\sum_\nu w_\nu M_{\nu\mu}.$$

Then

$$M_{\mu\nu}=m_\mu\delta_{\mu+\nu,0},\quad m_\mu=\begin{cases}(-q)^{\mu-1}, & \mu>0,\\ (-q)^n, & \mu=0,\\ (-q)^{2n+\mu}, & \mu<0.\end{cases} \tag{3}$$

It follows from earlier discussions that repeated tensor products of $\mathbf{E}$ are completely reducible. Furthermore, every finite-dimensional irreducible $\mathrm{U}_q(\mathbf{g})$ module is embedded in some $\mathbf{E}^{\otimes k}$ for at least one $k\geqslant 0$.

For later use, we consider two classes of $\mathbf{Z}_2$-graded Hopf subalgebras of $\mathrm{U}_q(\mathbf{g})$. Corresponding to any subset Θ of $\mathbf{N}_n$, we introduce

$$\mathcal{S}_k=\{k_i^{\pm1}, i\in\mathbf{N}_n;\quad e_j, f_j, j\in\boldsymbol{\Theta}\};$$

$$\mathcal{S}_p=\mathcal{S}_k\cup\{e_j, j\in\mathbf{N}_n\backslash\boldsymbol{\Theta}\}.$$

The elements of each set generate a $\mathbf{Z}_2$-graded Hopf subalgebra of $\mathrm{U}_q(\mathbf{g})$. The subalgebra generated by the elements of $\mathcal{S}_k$ will be denoted by $\mathrm{U}_q(\mathbf{k})$, and called a reductive subalgebra of $\mathrm{U}_q(\mathbf{g})$, while that generated by the elements of $\mathcal{S}_p$ will be denoted by $\mathrm{U}_q(\mathbf{p})$ and called a parabolic subalgebra. Note that $\mathrm{U}_q(\mathbf{k})$ is a $\mathbf{Z}_2$-graded Hopf subalgebra of $\mathrm{U}_q(\mathbf{p})$. If we replace e_i by f_i and vice versa in $\mathcal{S}_p$, we obtain another set, which will generate a $\mathbf{Z}_2$-graded Hopf subalgebra of $\mathrm{U}_q(\mathbf{g})$ having similar properties as $\mathrm{U}_q(\mathbf{p})$. Results presented in the remainder of the paper can also be formulated using such algebras.

Observe that there are two types of reductive subalgebras, depending on whether $\boldsymbol{\Theta}$ contains n. The first type arises when $n\notin\boldsymbol{\Theta}$, and in this case, $\mathrm{U}_q(\mathbf{k})$ is the direct product of quantized universal enveloping algebras associated with a series of ordinary (i.e., nongraded) Lie algebras of type A supplemented by the algebra generated by some $k_i^{\pm 1}$. The second type arises when $n \in \Theta$. This time, $\mathrm{U}_q(\mathbf{k})$ is the direct product of the first type with a $\mathrm{U}_q(\mathrm{osp}(1|2m))$ for some $m < n$. In both cases, the finite-dimensional representations of $\mathrm{U}_q(\mathbf{k})$ are completely reducible. This fact will be of great importance to the main subject of the paper.

Let V_μ be a finite-dimensional irreducible $\mathrm{U}_q(\mathbf{k})$ module. Then V_μ is of highest weight type. Let μ be the highest weight and $\tilde{\mu}$ the lowest weight of V_μ respectively. We can extend V_μ in a unique fashion to a $\mathrm{U}_q(\mathbf{p})$ module, which is still denoted by V_μ, such that the elements of $\mathcal{S}_p \backslash \mathcal{S}_k$ act by zero. It is not difficult to see that all finite dimensional irreducible $\mathrm{U}_q(\mathbf{p})$ modules are of this kind.

Consider a finite-dimensional irreducible $\mathrm{U}_q(\mathbf{g})$ module $W(\lambda)$, with highest weight λ and lowest weight $\bar{\lambda}$. $W(\lambda)$ can be restricted in a natural way to a $\mathrm{U}_q(\mathbf{p})$ module, which is always indecomposable, but not irreducible in general. It can be readily shown that

$$\dim_{\mathbf{C}} \mathrm{Hom}_{\mathrm{U}_q(\mathbf{p})}(W(\lambda), V_\mu) = \begin{cases} 1, & \bar{\lambda} = \tilde{\mu}, \\ 0, & \bar{\lambda} \neq \tilde{\mu}. \end{cases}$$

III. THE QUANTUM SUPERGROUP $\mathrm{OSP}_q(1|2n)$

There exist well-established methods for quantizing ordinary Lie groups in the nonsupersymmetric setting. (See Ref. 18 and references therein.) These methods can also be extended to construct $\mathrm{OSP}_q(1|2n)$, and this will be done here. However, we should point out that it is, in general, much more difficult to study quantum supergroups. See Ref. 9 for details on $\mathrm{GL}_q(m|n)$.

We will show that the quantum supergroup $\mathrm{OSP}_q(1|2n)$ admits a quantum Haar functional, and also a Peter–Weyl basis. This, however, is an exception rather than the rule. It is known that the finite-dimensional representations of all the quantum superalgebras but $\mathrm{U}_q(\mathrm{osp}(1|2n))$ are not completely reducible. This fact renders it impossible to construct Peter–Weyl bases for the corresponding quantum supergroups [which are yet to be defined except $\mathrm{GL}_q(m|n)$].

Let us recall some general results about $\mathbf{Z}_2$-graded Hopf algebras. Let A be a $\mathbf{Z}_2$-graded Hopf algebra with comultiplication Δ, counit ϵ, and antipode S. We define the finite dual A^0 of A to be a subspace of A^* such that for any $f \in A^0$, $\mathrm{Ker}\, f$ contains a two-sided ideal $\mathcal{I}$ of A which is of finite codimension, i.e., $\dim A/\mathcal{I} < \infty$. Of course in the most general situation, there is no guarantee that A^0 will not be zero. But when A^0 is nontrivial, then it is also a $\mathbf{Z}_2$-graded Hopf algebra with a structure dualizing that of A. More explicitly, the multiplication is defined, for f, $g \in A^0$, a, $b \in A$, by

$$\langle fg, a\rangle = \langle f \otimes g, \Delta(a)\rangle = \sum_{(a)} (-1)^{[g][a_{(1)}]} \langle f, a_{(1)}\rangle \langle g, a_{(2)}\rangle.$$

It is easy to see that the unit of A^0 is ϵ. Denote the comultiplication, the counit, and the antipode of A^0 respectively by Δ_0, ϵ^0 and S_0. Then

$$\langle \Delta_0(f), a \otimes b\rangle = \sum_{(f)} (-1)^{[f_{(1)}][f_{(2)}]} \langle f_{(1)}, a\rangle \langle f_{(2)}, b\rangle = \langle f, ab\rangle,$$

$$\langle S_0(f), a\rangle = \langle f, S(a)\rangle, \quad \epsilon^0(f) = \langle f, \mathbb{1}_A\rangle.$$

Now we come back to the quantum supergroup $\mathrm{OSP}_q(1|2n)$. As is well known, we cannot define the quantum supergroup directly. Instead, we need to find the algebra of functions on it. Introduce $t_{\mu\nu} \in (\mathrm{U}_q(\mathbf{g}))^*$, $\mu, \nu = 0, \pm 1, \pm 2, \ldots, \pm n$, defined by

$$t_{\mu\nu}(x) = t(x)_{\mu\nu}, \quad \forall x \in \mathrm{U}_q(\mathbf{q}),$$

where t is the vector representation of $\mathrm{U}_q(\mathbf{g})$. We call the $t_{\mu\nu}$ the matrix elements of t. Finite dimensionality of $\mathbf{E}$ implies that $t_{\mu\nu}\in(\mathrm{U}_q(\mathbf{g}))^0$, $\forall\mu,\ \nu$.

We define the superalgebra $\mathcal{T}_q(\mathbf{g})$ of functions on $\mathrm{OSP}_q(1|2n)$ to be the $\mathbf{Z}_2$-graded subalgebra of $(\mathrm{U}_q(\mathbf{g}))^0$ generated by the matrix elements of the vector representation of $\mathrm{U}_q(\mathbf{g})$, i.e., $t_{\mu\nu}$, $\mu,\nu=0,\pm 1,\pm 2,...,\pm n$. Then we have the following theorem.

Theorem 1: (1) *$\mathcal{T}_q(\mathbf{g})$ is a $\mathbf{Z}_2$-graded Hopf algebra.*

(2) *Let $t^{(\lambda)}$ be the irreducible representation of $\mathrm{U}_q(\mathbf{g})$ with highest weight $\lambda\in\mathcal{P}_+$, and let $t_{ij}^{(\lambda)}$, $i,j=1,2,...,d_\lambda$ $(d_\lambda=\dim t^{(\lambda)})$, be the matrix elements of $t^{(\lambda)}$. Then*

$$\mathcal{T}_q(\mathbf{g})=\bigoplus_{\lambda\in\mathcal{P}_+}\bigoplus_{i,j=1}^{d_\lambda}\mathbf{C}t_{ij}^{(\lambda)}. \tag{4}$$

Proof: The $\mathbf{Z}_2$-graded bialgebra structure of $\mathcal{T}_q(\mathbf{g})$ is obvious, and the existence of the antipode follows from the self-duality of the vector module $\mathbf{E}$ over $\mathrm{U}_q(\mathbf{g})$. Part (2) immediately follows from the complete reducibility of finite-dimensional representations of $\mathrm{U}_q(\mathbf{g})$. □

Let us now work out the explicit forms of the comultiplication and the antipode. The comultiplication is given by

$$\Delta_0(t_{\mu\nu})=\sum_\sigma\ (-1)^{(\delta_{\mu 0}+\delta_{\sigma 0})(\delta_{\nu 0}+\delta_{\sigma 0})}t_{\mu\sigma}\otimes t_{\sigma\nu}.$$

The antipode can be constructed from (2) by using the $\mathrm{U}_q(\mathbf{g})$-module isomorphism M. We have

$$\begin{aligned}S_0(t_{\mu\nu})&=(-1)^{(\delta_{\mu 0}+\delta_{\nu 0})\delta_{\mu 0}}(M^{-1}tM)_{\nu\mu}\\&=(-1)^{(\delta_{\mu 0}+\delta_{\nu 0})\delta_{\mu 0}}\frac{m_{-\mu}t_{-\nu,-\mu}}{m_{-\nu}},\end{aligned}$$

where m_μ is given by (3).

Here we introduce more notations for later use. Let $\{w_i^{(\lambda)}|i=1,2,...,d_\lambda\}$ be the homogeneous basis of $W(\lambda)$ with respect to which the representation $t^{(\lambda)}$ is defined. We denote by $\{\tilde{w}_i^{(\lambda)}|i=1,2,...,d_\lambda\}$ the basis of $W(\lambda)^*=W(\lambda^\dagger)$ such that $\tilde{w}_i^{(\lambda)}(w_j^{(\lambda)})=\delta_{ij}$. The $\mathrm{U}_q(\mathbf{g})$-module structure of $W(\lambda)^*$ enables us to define $\tilde{t}_{ij}^{(\lambda)}\in\mathcal{T}_q(\mathbf{g})$ by

$$x\tilde{w}_i^{(\lambda)}=\sum_j\ \tilde{t}_{ji}^{(\lambda)}(x)\tilde{w}_j^{(\lambda)},\quad \forall x\in\mathrm{U}_q(\mathbf{g}).$$

Then

$$\tilde{t}_{ji}^{(\lambda)}=(-1)^{[i]([i]+[j])}S_0(t_{ij}^{(\lambda)}),$$

where $[i]=0$ or 1 depending on whether w_i is even or odd. Clearly the $\tilde{t}_{ji}^{(\lambda)}$ are linear combinations of $t_{ij}^{(\lambda^\dagger)}$. Furthermore, the $\tilde{t}_{ji}^{(\lambda)}$, $\forall\lambda\in\mathcal{P}_+$, also form a basis of $\mathcal{T}_q(\mathbf{g})$.

From here on, we will omit the subscript 0 from Δ_0 and S_0.

Let us now turn to the discussion of a Haar functional on the quantum supergroup $\mathcal{T}_q(\mathbf{g})$. But before embarking on this task, we first consider the notion of an integral on an arbitrary $\mathbf{Z}_2$-graded Hopf algebra A. Let A^* be its dual, which has a natural $\mathbf{Z}_2$-graded algebraic structure induced by the co-algebraic structure of A. An even homogeneous element $\int^l\in A^*$ is called a left integral on A if

$$f\cdot\int^l=\langle f,\mathbb{1}_A\rangle\int^l,\quad \forall f\in A^*.$$

Similarly, an even homogeneous element $\int^r\in A^*$ is called a right integral on A if

$$\int^r \cdot f = \langle f, 1_A \rangle \int^r, \quad \forall f \in A^*.$$

A straightforward calculation shows that the defining properties of the integrals are equivalent to the following requirements

$$\left(\mathrm{id} \otimes \int^l\right) \Delta(x) = \int^l x, \quad \left(\int^r \otimes \mathrm{id}\right) \Delta(x) = \int^r x, \quad \forall x \in A. \tag{5}$$

where id is the identity map on A.

A Haar functional $\int \in A^*$ on A is an integral on A which is both left and right, and sends 1_A to 1, i.e.,

$$\begin{gathered} (i) \quad \left(\int \otimes \mathrm{id}\right) \Delta(x) = \left(\mathrm{id} \otimes \int\right) \Delta(x) = \int x, \quad \forall x \in A, \\ (ii) \quad \int 1_A = 1. \end{gathered} \tag{6}$$

In the case of $\mathcal{T}_q(\mathbf{g})$, it is an entirely straightforward matter to show the following.

Theorem 2: *The element* $\int \in (\mathcal{T}_q(\mathbf{g}))^*$ *defined by*

$$\int 1_{\mathcal{T}_q(\mathbf{g})} = 1; \quad \int t_{ij}^{(\lambda)} = 0, \quad 0 \neq \lambda \in \mathcal{P}_+,$$

gives rise to a Haar functional on $\mathcal{T}_q(\mathbf{g})$.

Denote by 2ρ the sum of the positive roots of $\mathbf{g}$. Let $K_{2\rho}$ be the product of powers of $k_i^{\pm 1}$'s such that

$$K_{2\rho} e_i K_{2\rho}^{-1} = q^{(2\rho, \alpha_i)} e_i, \quad \forall i.$$

Then it can be easily shown that

$$S^2(x) = K_{2\rho} x K_{2\rho}^{-1}, \quad \forall x \in \mathrm{U}_q(\mathbf{q}).$$

We define the quantum superdimension of the irreducible $\mathrm{U}_q(\mathbf{g})$ module $W(\lambda)$ by

$$\mathrm{SD}_q(\lambda) := \mathrm{Str}\{t^{(\lambda)}(K_{2\rho})\}.$$

For quantum superalgebras other than the $\mathrm{osp}(1|2n)$ series, there exists a class of finite-dimensional irreducible representations, the typicals, of which the super-dimensions vanish identically. Again, $\mathrm{U}_q(\mathrm{osp}(1|2n))$ is an exception, and we have the following important property: for any irreducible $\mathrm{U}_q(\mathrm{osp}(1|2n))$ module $W(\lambda)$ with highest weight $\lambda \in \mathcal{P}_+$,

$$\mathrm{SD}_q(\lambda) \neq 0.$$

Now the Haar functional $\int$ satisfies the following properties.

Lemma 1:

$$\begin{gathered} \int t_{ij}^{(\lambda)} \tilde{t}_{rs}^{(\mu)} (-1)^{[j][r]+[i]+[j]} = \delta_{ir} \delta_{\lambda\mu} \frac{t_{sj}^{(\lambda)}(K_{2\rho})}{SD_q(\lambda)}, \\ \int \tilde{t}_{ij}^{(\lambda)} t_{rs}^{(\mu)} (-1)^{[j][r]} = \delta_{js} \delta_{\lambda\mu} \frac{\tilde{t}_{ir}^{(\lambda)}(K_{2\rho})}{SD_q(\lambda)}. \end{gathered} \tag{7}$$

Proof: Consider the first equation. The $\lambda \neq \mu$ case is easy to prove: the integral vanishes because the tensor product $W(\lambda)\otimes W(\mu^{\dagger})$ does not contain the trivial $\mathrm{U}_q(\mathbf{g})$ module. When $\lambda = \mu$, we introduce the notations

$$\phi_{ir;sj}=\int t_{ij}^{(\lambda)}\tilde{t}_{rs}^{(\lambda)}(-1)^{[j][r]+[i]+[j]};\quad \Phi[s,j]=(\phi_{ir;sj})_{i,r=1}^{d_\lambda};\quad \Psi[i,r]=(\phi_{ir;sj})_{s,j=1}^{d_\lambda}.$$

It is clearly true that $\mathrm{Str}(\Psi[i,r])=\delta_{ir}$.

Note that corresponding to each $x\in \mathrm{U}_q(\mathbf{g})$, there exists an $\tilde{x}\in(\mathcal{T}_q(\mathbf{g}))^*$ defined by $\tilde{x}(a)=\langle a,x\rangle$, $\forall a\in\mathcal{T}_q(\mathbf{g})$. The left integral property of $\int$ leads to

$$\begin{aligned}\epsilon(x)\phi_{ir;sj}&=\left(\tilde{x}.\int\right)t_{ij}^{(\lambda)}\tilde{t}_{rs}^{(\lambda)}(-1)^{[j][r]+[i]+[j]}\\&=\sum_{(x)}\sum_{i',r'}t_{ii'}^{(\lambda)}(x_{(1)})t_{r'r}^{(\lambda)}(S(x_{(2)}))\phi_{i'r';sj}(-1)^{[x]([i]+[j])+[x_{(2)}]([j]+[s])},\end{aligned}$$

i.e.,

$$\epsilon(x)\Phi[s,j]=\sum_{(x)}t^{(\lambda)}(x_{(1)})\Phi[s,j]t^{(\lambda)}(S(x_{(2)}))(-1)^{[x_{(2)}]([j]+[s])},\quad \forall x\in \mathrm{U}_q(\mathbf{g}).$$

Schur's lemma forces $\Phi[s,j]$ to be proportional to the identity matrix, and we have

$$\Psi[i,r]=\delta_{ir}\psi,$$

for some $d_\lambda\times d_\lambda$ matrix ψ. The right integral property of $\int$ leads to

$$\epsilon(y)\psi=\sum_{(y)}t^{(\lambda)}(K_{2\rho})t^{(\lambda)}(y_{(1)})t^{(\lambda)}(K_{2\rho}^{-1})\psi t^{(\lambda)}(S(y_{(2)})).$$

Again by using Schur's lemma we conclude that ψ is proportional to $t^{(\lambda)}(K_{2\rho})$. Since its supertrace is 1, we have

$$\psi=\frac{t^{(\lambda)}(K_{2\rho})}{\mathrm{SD}_q(\lambda)}.$$

This completes the proof of the first equation of the lemma. The second equation can be shown in exactly the same way. □

It is worth observing that this Lemma and part (2) of Theorem 1 provide a quantum analog of the Peter–Weyl theorem for $\mathrm{OSP}_q(1|2n)$.

IV. QUANTUM HOMOGENEOUS SUPERVECTOR BUNDLES

In this section we will investigate the quantum homogeneous superspaces and quantum homogeneous supervector bundles arising from the quantum supergroup $\mathrm{OSP}_q(1|2n)$ by adapting the methods and techniques of Refs. 9 and 19 to the present context. Let us start by introducing two types of actions of $\mathrm{U}_q(\mathbf{g})$ on $\mathcal{T}_q(\mathbf{g})$. The first action will be denoted by $\circ$, which corresponds to the right translation in the classical theory of Lie groups. It is defined by

$$x\circ f=\sum_{(f)}(-1)^{[f_{(1)}][f_{(2)}]}f_{(1)}\langle f_{(2)},x\rangle,\quad x\in \mathrm{U}_q(\mathbf{g}),\quad f\in\mathcal{T}_q(\mathbf{g}).\tag{8}$$

Straightforward calculations show that

$$x\circ(y\circ f)=(xy)\circ f,\quad (x\circ f)(y)=f(yx),\quad (\mathrm{id}_{\mathcal{T}_q(\mathbf{g})}\otimes x\circ)\Delta(f)=\Delta(x\circ f).$$

The other action, which corresponds to the left translation in the classical Lie group theory, will be denoted by $\cdot$. It is defined by

$$x\cdot f=\sum_{(f)}\langle f_{(1)},S^{-1}(x)\rangle f_{(2)}.\tag{9}$$

It can be easily shown that

$$(x\cdot f)(y)=(-1)^{[x][y]}f(S^{-1}(x)y),$$

$$x\cdot(y\cdot f)=(xy)\cdot f,\quad x,y\in \mathrm{U}_q(\mathbf{g}),\quad f\in\mathcal{T}_q(\mathbf{g}).$$

Furthermore, the two actions graded commute in the following sense

$$x\circ(y\cdot f)=(-1)^{[x][y]}y\cdot(x\circ f),\quad x,y\in \mathrm{U}_q(\mathbf{g}),\quad f\in\mathcal{T}_q(\mathbf{g}).$$

Let V be a finite-dimensional module over $\mathrm{U}_q(\mathbf{k})$. We extend the actions $\circ$ and $\cdot$ trivially to $V\otimes\mathcal{T}_q(\mathbf{g})$: for any $\zeta=\Sigma v_i\otimes f_i\in V\otimes\mathcal{T}_q(\mathbf{g})$,

$$x\cdot\zeta=\sum(-1)^{[x][v_i]}v_i\otimes x\cdot f_i,$$

$$x\circ\zeta=\sum(-1)^{[x][v_i]}v_i\otimes x\circ f_i,\quad x\in \mathrm{U}_q(\mathbf{g}).$$

We now introduce two important definitions:

$$\mathcal{A}_q^{\mathbf{k}}:=\{f\in\mathcal{T}_q(\mathbf{g})|x\circ f=\epsilon(x)f,\quad \forall x\in \mathrm{U}_q(\mathbf{k})\};\tag{10}$$

$$\Gamma_q^{\mathbf{k}}(V):=\{\zeta\in V\otimes\mathcal{T}_q(\mathbf{g})|x\circ\zeta=(S(x)\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})\zeta,\quad\forall x\in \mathrm{U}_q(\mathbf{k})\}.\tag{11}$$

The remainder of this section is devoted to studying the properties of these objects. Let us first prove the following.

Proposition 1: (1) $A_q^{\mathbf{k}}$ *is an infinite-dimensional subalgebra of* $\mathcal{T}_q(\mathbf{g})$.

(2) $\Gamma_q^{\mathbf{k}}(V)$ *is an infinite-dimensional supervector space if the weight of any vector of V is* $\mathrm{U}_q(\mathbf{g})$ *integral, and is zero otherwise.*

Proof: We first show that $\mathcal{A}_q^{\mathbf{k}}$ is a subalgebra of $\mathcal{T}_q(\mathbf{g})$. Since $\mathrm{U}_q(\mathbf{k})$ is a Hopf subalgebra of $\mathrm{U}_q(\mathbf{g})$, for any $x\in \mathrm{U}_q(\mathbf{k})$, $\Delta(x)=\Sigma_{(x)}x_{(1)}\otimes x_{(2)}\in \mathrm{U}_q(\mathbf{k})\otimes \mathrm{U}_q(\mathbf{k})$. Hence

$$x\circ(ab)=\sum_{(x)}(-1)^{[x_{(2)}][a]}\{x_{(1)}\circ a\}\{x_{(2)}\circ b\}=\epsilon(x)ab,$$

that is, $ab\in\mathcal{A}_q^{\mathbf{k}}$.

Since the finite-dimensional representations of $\mathrm{U}_q(\mathbf{k})$ are completely reducible, the study of properties of $\Gamma_q^{\mathbf{k}}(V)$ reduces to the case when V is irreducible. Let V_μ be a finite-dimensional irreducible $\mathrm{U}_q(\mathbf{k})$ module with highest weight μ and lowest weight $\tilde{\mu}$. Any element $\zeta\in\Gamma_q^{\mathbf{k}}(V_\mu)$ can be expressed in the form

$$\zeta=\sum_{\lambda\in\mathcal{P}_+}\sum_{i,j}v_{ij}^{(\lambda)}\otimes\tilde{t}_{ij}^{(\lambda)},$$

for some $v_{ij}^{(\lambda)} \in V_\mu$. Fix an arbitrary $\lambda \in \mathcal{P}_+$. For any nonvanishing $w \in W(\lambda)$, the following linear map is clearly surjective:

$$\mathrm{Hom}_{\mathbf{C}}(W(\lambda), V_\mu) \otimes w \to V_\mu,$$

$$\phi \otimes w \mapsto \phi(w).$$

Thus there exist $\phi_i^{(\lambda)} \in \mathrm{Hom}_{\mathbf{C}}(W(\lambda), V_\mu)$ such that $v_{ij}^{(\lambda)} = \phi_i^{(\lambda)}(w_j^{(\lambda)})$, where $\{w_i^{(\lambda)}\}$ is the basis of $W(\lambda)$ discussed before. Therefore, we can rewrite ζ as

$$\zeta = \sum_{\lambda \in \mathcal{P}_+} \sum_{i,j} \phi_i^{(\lambda)}(w_j^{(\lambda)}) \otimes \tilde{t}_{ij}^{(\lambda)}.$$

The defining property of $\Gamma_q^{\mathbf{k}}(V_\mu)$ states that

$$\ell \circ \zeta = (\mathrm{id}_{\mathcal{T}_q(\mathbf{g})} \otimes S(\ell))\zeta, \quad \forall \ell \in \mathrm{U}_q(\mathbf{k}).$$

Thus we have

$$\sum_{\lambda \in \mathcal{P}_+} \sum_{i,j,k} t_{jk}^{(\lambda)}(S(\ell))\phi_i^{(\lambda)}(w_j^{(\lambda)}) \otimes (-1)^{[\ell][\phi_i^{(\lambda)}]}\tilde{t}_{ik}^{(\lambda)} = \sum_{\lambda \in \mathcal{P}_+} \sum_{i,j} S(\ell)\phi_i^{(\lambda)}(w_j^{(\lambda)}) \otimes \tilde{t}_{ij}^{(\lambda)}.$$

Recalling that the $\tilde{t}_{ki}^{(\lambda)}$ are linearly independent, the above is equivalent to

$$\ell \phi_i^{(\lambda)}(w_j^{\lambda}) = (-1)^{[\ell][\phi_i^{(\lambda)}]}\phi_i^{(\lambda)}(\ell w_j^{(\lambda)}), \quad \forall \ell \in \mathrm{U}_q(\mathbf{k}).$$

This equation is precisely the statement that the $\phi_i^{(\lambda)}$ be $\mathrm{U}_q(\mathbf{k})$-module homomorphisms of degrees $[\phi_i^{(\lambda)}]$,

$$\phi_i^{(\lambda)} \in \mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W(\lambda), V_\mu) \subset \mathrm{Hom}_{\mathbf{C}}(W(\lambda), V_\mu), \quad \forall i.$$

Thus finding sections in $\Gamma_q^{\mathbf{k}}(V_\mu)$ is equivalent to finding, for all $\lambda \in \mathcal{P}_+$, the homomorphisms $\phi^{(\lambda)} \in \mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W(\lambda), V_\mu)$. Note that each such homomorphism $\phi^{(\lambda)}$ determines d_λ linearly independent sections:

$$\zeta_i^{(\lambda)} = \sum_j \phi^{(\lambda)}(w_j^{(\lambda)}) \otimes \tilde{t}_{ij}^{(\lambda)}.$$

However, when μ is not integral with respect to $\mathrm{U}_q(\mathbf{g})$, $\mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W(\lambda), V_\mu) = 0$, and hence $\Gamma_q^{\mathbf{k}}(V_\mu)$ vanishes in this case.

Now consider the case with $\mu = 0$; we have $\Gamma_q^{\mathbf{k}}(V_{\mu=0}) = \mathcal{A}_q^{\mathbf{k}}$ as supervector spaces. There is a homomorphism from the trivial representation of $\mathrm{U}_q(\mathbf{g})$, $W(0) = \mathbf{C}$, onto $V_0 = \mathbf{C}$. This gives the constant sections of $\mathcal{A}_q^{\mathbf{k}}$. Let γ be the highest root of $\mathbf{g}$. Recall that in the classical situation, $\mathbf{k}$ is reductive with $N = r - |\Theta|$ independent central elements. This, transcribed to the quantum case, implies the existence of N linearly independent $\mathrm{U}_q(\mathbf{k})$ homomorphisms $W(\gamma) \to \mathbf{C}$. As mentioned above, each of these corresponds to $d = \dim(\mathbf{g})$ linearly independent sections. So the representation $W(\gamma)$ determines Nd linearly independent sections. Further linearly independent sections can be obtained using the following lemma.

Lemma 2: Suppose there are nontrivial $\mathrm{U}_q(\mathbf{k})$ *homomorphisms* $W(\lambda_1) \to V_{\mu_1}$ *and* $W(\lambda_2) \to V_{\mu_2}$. *Then there is an induced nontrivial* $\mathrm{U}_q(\mathbf{k})$ *homomorphism*

$$W(\lambda_1 + \lambda_2) \to V_{\mu_1 + \mu_2}.$$

For example, for any positive integer m, there exist $(m|N)$ (partition of m into $\leqslant N$ parts) linearly independent homomorphisms $W(m\gamma)\rightarrow\mathbf{C}$. Thus we have proved that the algebra $\mathcal{A}_q^{\mathbf{k}}$ is infinite dimensional.

Now let us consider the case with $0\neq\mu\in\mathcal{P}$. It is an elementary exercise to verify that V_μ is $\mathrm{U}_q(\mathbf{k})$-isomorphic to a $\mathrm{U}_q(\mathbf{k})$-irreducible part of $W(\lambda')$, where λ' is the dominant weight in the Weyl group orbit of μ. Thus there is a nontrivial $\mathrm{U}_q(\mathbf{k})$ homomorphism

$$W(\lambda')\rightarrow V_\mu,$$

and this determines at least $d_{\lambda'}$ linearly independent sections in $\Gamma_q^{\mathbf{k}}(V_\mu)$. Further linearly independent sections can be constructed explicitly using Lemma 2 which promises a family of homomorphisms

$$W(\lambda'+m\gamma)\rightarrow V_\mu, \quad m\in\mathbf{N}_+ .$$

This establishes that $\Gamma_q^{\mathbf{k}}(V_\mu)$ is infinite dimensional. □

$\mathcal{A}_q^{\mathbf{k}}$ may be regarded as the quantum analog of the algebra of functions over the superspace $\mathrm{OSP}(1|2n)/K$, where K is the subgroup of $\mathrm{OSP}(1|2n)$ with Lie superalgebra $\mathbf{k}$. Such homogeneous superspaces were studied in the work of Manin,[16] Penkov,[17] and others. Here we wish to make some investigations into their quantum analogs.

As is well known, one cannot define a noncommutative (in the $\mathbf{Z}_2$-graded sense) space directly in geometrical terms. Instead, such a space has to be defined by specifying its algebra of functions. We will take $\mathcal{A}_q^{\mathbf{k}}$ as the algebra of functions over the quantum homogeneous superspace which corresponds to $\mathrm{OSP}(1|2n)/K$ in the classical situation. Let us now study properties of $\Gamma_q^{\mathbf{k}}(V)$. First observe the following.

Theorem 3: *$\Gamma_q^{\mathbf{k}}(V)$ furnishes a two-sided $\mathcal{A}_q^{\mathbf{k}}$ module under the multiplication of $\mathcal{T}_q(\mathbf{g})$.*

Proof: The left and right actions of $\mathcal{A}_q^{\mathbf{k}}$ on $\Gamma_q^{\mathbf{k}}(V)$ are respectively defined by

$$a\zeta=\sum_r (-1)^{[a][v_i]}v_i\otimes af_i, \quad \zeta a=\sum_r v_i\otimes f_i a,$$

where $a\in\mathcal{A}_q^{\mathbf{k}}$ and $\zeta=\Sigma_i v_i\otimes f_i\in\Gamma_q^{\mathbf{k}}(V)$. Now for $p\in\mathrm{U}_q(\mathbf{k})$,

$$p\circ(a\zeta)=\sum_{(p)} (-1)^{[p_{(2)}][a]}\{p_{(1)}\circ a\}\{p_{(2)}\circ\zeta\}=(-1)^{[p][a]}a\{p\circ\zeta\}=(S(p)\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})a\zeta;$$

$$p\circ(\zeta a)=\sum_{(p)} (-1)^{[p_{(2)}][\zeta]}\{p_{(1)}\circ\zeta\}\{p_{(2)}\circ a\}=\{p\circ\zeta\}a=(S(p)\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})\zeta a.$$

This completes the proof. □

When V is actually a $\mathrm{U}_q(\mathbf{g})$ module, the $\mathcal{A}_q^{\mathbf{k}}$ module $\Gamma_q^{\mathbf{k}}(V)$ has a particularly simple structure.

Proposition 2: Let W be a finite-dimensional left $\mathrm{U}_q(\mathbf{g})$ module, which we regard as a left $\mathrm{U}_q(\mathbf{k})$ module by restriction. Then $\Gamma_q^{\mathbf{k}}(W)$ is isomorphic to $W\otimes\mathcal{A}_q^{\mathbf{k}}$ either as a left or right $\mathcal{A}_q^{\mathbf{k}}$ module.

Proof: We first construct the right $\mathcal{A}_q^{\mathbf{k}}$ module isomorphism. Being a left $\mathrm{U}_q(\mathbf{g})$ module, W carries a natural right $\mathcal{T}_q(\mathbf{g})$ comodule structure with the comodule action $\delta: W\rightarrow W\otimes\mathcal{T}_q(\mathbf{g})$ defined by

$$\delta(w)(x)=xw, \quad x\in\mathrm{U}_q(\mathbf{g}), \quad w\in W. \tag{12}$$

[Here the notation requires some clarification. If we express $\delta(w)=\Sigma_{(w)}w_{(1)}\otimes w_{(2)}$, then $\delta(w)(x)=\Sigma_{(w)}(-1)^{[x][w_{(1)}]}w_{(1)}\langle w_{(2)},x\rangle$.] Define $\eta: W\otimes\mathcal{T}_q(\mathbf{g})\rightarrow W\otimes\mathcal{T}_q(\mathbf{g})$ by the composition of maps

$$W\otimes\mathcal{T}_q(\mathbf{g})\xrightarrow{\delta\otimes\mathrm{id}}W\otimes\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\to W\otimes\mathcal{T}_q(\mathbf{g}),$$

where the last map is the multiplication of $\mathcal{T}_q(\mathbf{g})$. Then η defines a right $\mathcal{A}_q^{\mathbf{k}}$ module isomorphism, with the inverse map given by the composition

$$W\otimes\mathcal{T}_q(\mathbf{g})\xrightarrow{\delta\otimes\mathrm{id}}W\otimes\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\xrightarrow{(\mathrm{id}\otimes S\otimes\mathrm{id})}W\otimes\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\to W\otimes\mathcal{T}_q(\mathbf{g}),$$

where the last map is again the multiplication of $\mathcal{T}_q(\mathbf{g})$. It is not difficult to show that

$$x\circ\eta(\zeta)=\eta\left(\sum_{(x)}(x_{(1)}\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})x_{(2)}\circ\zeta\right),$$

$$x\circ\eta^{-1}(\zeta)=\eta^{-1}\left(\sum_{(x)}(S(x_{(1)})\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})x_{(2)}\circ\zeta\right),\quad\forall\zeta\in\mathcal{T}_q(\mathbf{g})\otimes W,\quad x\in\mathrm{U}_q(\mathbf{g}).$$

Consider $\zeta\in\Gamma_q^{\mathbf{k}}(W)$. We have

$$p\circ\eta(\zeta)=\eta\left(\sum_{(p)}(p_{(1)}\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})p_{(2)}\circ\zeta\right)=\eta\left(\sum_{(p)}(p_{(1)}S(p_{(2)})\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})\zeta\right)$$
$$=\epsilon(p)\,\eta(\zeta),\quad\forall p\in\mathrm{U}_q(\mathbf{k}).$$

Hence $\eta(\Gamma_q^{\mathbf{k}}(W))\subset W\otimes\mathcal{A}_q^{\mathbf{k}}$. Conversely, given any $\xi\in W\otimes\mathcal{A}_q^{\mathbf{k}}$, we have

$$p\circ\eta^{-1}(\xi)=\eta^{-1}\left(\sum_{(p)}(S(p_{(1)})\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})p_{(2)}\circ\xi\right)$$
$$=\eta^{-1}\left(\sum_{(p)}(S(p_{(1)})\epsilon(p_{(2)})\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})\xi\right)=(S(p)\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})\,\eta^{-1}(\xi),\quad\forall p\in\mathrm{U}_q(\mathbf{k}).$$

Thus $\eta^{-1}(W\otimes\mathcal{A}_q^{\mathbf{k}})\subset\Gamma_q^{\mathbf{k}}(W)$. Therefore the restriction of η to $\Gamma_q^{\mathbf{k}}(W)$ provides the desired right $\mathcal{A}_q^{\mathbf{k}}$ module isomorphism.

The left module isomorphism is given by the restriction to $\Gamma_q^{\mathbf{k}}(W)$ of the linear map $\kappa:W\otimes\mathcal{T}_q(\mathbf{g})\to W\otimes\mathcal{T}_q(\mathbf{g})$, which is defined by the following composition of maps

$$W\otimes\mathcal{T}_q(\mathbf{g})\xrightarrow{\delta\otimes\mathrm{id}}W\otimes\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\xrightarrow{\mathrm{id}\otimes P(S^2\otimes\mathrm{id})}W\otimes\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\to W\otimes\mathcal{T}_q(\mathbf{g}),$$

where

$$P:\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\to\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g}),$$
$$a\otimes b\mapsto(-1)^{[a][b]}b\otimes a.\tag{13}$$

The inverse map κ^{-1} is given by

$$W\otimes\mathcal{T}_q(\mathbf{g})\xrightarrow{\delta\otimes\mathrm{id}}W\otimes\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\xrightarrow{\mathrm{id}\otimes P(S\otimes\mathrm{id})}W\otimes\mathcal{T}_q(\mathbf{g})\otimes\mathcal{T}_q(\mathbf{g})\to W\otimes\mathcal{T}_q(\mathbf{g}).$$

□

With the help of this Proposition, we can now prove the following important result.

Theorem 4: $\Gamma_q^{\mathbf{k}}(V)$ *is projective and of finite type both as a left and right module over the superalgebra* $\mathcal{A}_q^{\mathbf{k}}$ *of functions on the quantum homogeneous superspace.*

Proof: Since $\mathrm{U}_q(\mathbf{k})$ is a reductive subalgebra of $\mathrm{U}_q(\mathbf{g})$, all finite-dimensional representations of $\mathrm{U}_q(\mathbf{k})$ are completely reducible. Let V_s, $s=1,2,...,K<\infty$, be the irreducible direct summands of

V such that their weights are all integral with respect to $\mathrm{U}_q(\mathbf{g})$. Then $\Gamma_q^{\mathbf{k}}(V)=\oplus_s\Gamma_q^{\mathbf{k}}(V_s)$. Consider any V_s, and denote its highest weight by μ_s. There exists such a $\hat{\mu}_s$ in the Weyl group orbit of $\mathbf{g}$ that is integral dominant with respect to $\mathbf{g}$. Let $W(\hat{\mu}_s)$ be the irreducible $\mathrm{U}_q(\mathbf{g})$ module with highest weight $\hat{\mu}_s$, which can be regarded as a $\mathrm{U}_q(\mathbf{k})$ module in the natural way. There always exists a $\mathrm{U}_q(\mathbf{k})$ module $V_s^{\perp}$ such that $W(\hat{\mu}_s)=V_s\oplus V_s^{\perp}$. Write $V^{\perp}=\oplus_s V_s^{\perp}$, and $W=\oplus_s W(\hat{\mu}_s)$. We have

$$\Gamma_q^{\mathbf{k}}(V)\oplus\Gamma_q^{\mathbf{k}}(V^{\perp})=\Gamma_q^{\mathbf{k}}(W)\cong W\otimes\mathcal{A}_q^{\mathbf{k}},$$

where the last step follows from Proposition 2. □

Recall that in classical differential geometry, the space $\mathcal{H}$ of sections of a vector bundle over a compact manifold M furnishes a module over the algebra $\mathcal{A}(M)$ of functions on M. It then follows from Swan's theorem that this module must be projective and is of finite type. Conversely, any projective module of finite type over $\mathcal{A}(M)$ is isomorphic to the space of sections of some vector bundle over M. This result is taken as the starting point for studying vector bundles in noncommutative geometry: one defines a vector bundle over a noncommutative space in terms of the space of sections which is required to be a finite-type project module over the noncommutative algebra of functions on the virtual noncommutative space. Therefore, $\Gamma_q^{\mathbf{k}}(V)$ will be called the space of sections of a quantum supervector bundle over the quantum homogeneous superspace associated with $\mathcal{A}_q^{\mathbf{k}}$.

Homogeneous supervector bundles at the classical level were studied in Refs. 16 and 17. We will not enter the discussion of the subject, but merely mention that the subject proves to be extremely rich and many aspects of it remain to be developed.

Following the classical terminology, we will call a quantum supervector bundle trivial if the sections form a free module over the superalgebra of functions on the quantum superspace. The following proposition is an immediate consequence of Proposition 2.

Proposition 3: If the $\mathrm{U}_q(\mathbf{k})$ *module V is in fact a finite-dimensional left* $\mathrm{U}_q(\mathbf{g})$ *module, then the quantum homogeneous supervector bundle with the space of sections* $\Gamma_q^{\mathbf{k}}(V)$ *is trivial.*

V. INDUCED REPRESENTATIONS

In this section we will investigate induced representations of the quantum supergroup $\mathrm{OSP}_q(1|2n)$ by using results of the last section. The following proposition explains how quantum homogeneous supervector bundles enter representation theory.

Proposition 4: $\Gamma_q^{\mathbf{k}}(V)$ *furnishes a left* $\mathrm{U}_q(\mathbf{g})$ *module under the* $\cdot$ *action, and also a right* $\mathcal{T}_q(\mathbf{g})$ *comodule under the action* $\omega=\mathrm{id}_V\otimes(\mathrm{id}_{\mathcal{T}_q(\mathbf{g})}\otimes S^{-1})\Delta$.

Proof: For $p\in\mathrm{U}_q(\mathbf{k})$, $x\in\mathrm{U}_q(\mathbf{g})$, and $\zeta\in\Gamma_q^{\mathbf{k}}(V)$, we have

$$p\circ(x\cdot\zeta)=(-1)^{[p][x]}x\cdot(p\circ\zeta)=(S(p)\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})(x\cdot\zeta).$$

Thus $\Gamma_q^{\mathbf{k}}(V)$ indeed furnishes a left $\mathrm{U}_q(\mathbf{g})$ module under the $\cdot$ action. The $\mathcal{T}_q(\mathbf{g})$ coaction ω is just the dual of this left $\mathrm{U}_q(\mathbf{g})$ action.

We call $\Gamma_q^{\mathbf{k}}(V)$ an induced $\mathrm{U}_q(\mathbf{g})$ module, and also an induced $\mathcal{T}_q(\mathbf{g})$ comodule. For such induced modules, we have the following quantum analog of Frobenius reciprocity.

Theorem 5: *Let W be a* $\mathrm{U}_q(\mathbf{g})$ *module, the restriction of which furnishes a* $\mathrm{U}_q(\mathbf{k})$ *module in a natural way. Then there exists a canonical isomorphism*

$$\mathrm{Hom}_{\mathrm{U}_q(\mathbf{g})}(W,\Gamma_q^{\mathbf{k}}(V))\cong\mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W,V), \tag{14}$$

where $\mathrm{U}_q(\mathbf{g})$ *acts on the left module* $\Gamma_q^{\mathbf{k}}(V)$ *via the* $\cdot$ *action.*

Proof: We prove the proposition by explicitly constructing the isomorphism, which we claim to be the linear map

$$F:\mathrm{Hom}_{\mathrm{U}_q(\mathbf{g})}(W,\Gamma_q^{\mathbf{k}}(V))\rightarrow \mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W,V), \quad \psi\mapsto\psi(1_{\mathrm{U}_q(\mathbf{g})}),$$

with the inverse map

$$\overline{F}:\mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W,V)\rightarrow \mathrm{Hom}_{\mathrm{U}_q(\mathbf{g})}(W,\Gamma_q^{\mathbf{k}}(V)), \quad \phi\mapsto\overline{\phi}=(\phi\otimes S)\delta,$$

where $\delta: W\rightarrow W\otimes\mathcal{T}_q(\mathbf{g})$ is the right $\mathcal{T}_q(\mathbf{g})$ comodule action defined by (12).

To verify our claim, we first need to demonstrate that the image of F is contained in $\mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W,V)$. Consider $\psi\in\mathrm{Hom}_{\mathrm{U}_q(\mathbf{g})}(W,\Gamma_q^{\mathbf{k}}(V))$. For any $p\in\mathrm{U}_q(\mathbf{k})$ and $w\in W$, we have

$$p(F\psi(w))=(S^{-1}(p)\circ\psi(w))(\mathbb{1}_{\mathrm{U}_q(\mathbf{g})}),$$

where we have used the defining property of $\Gamma_q^{\mathbf{k}}(V)$. Note that

$$(S^{-1}(p)\circ\psi(w))(\mathbb{1}_{\mathrm{U}_q(\mathbf{g})})=(p\cdot\psi(w))(\mathbb{1}_{\mathrm{U}_q(\mathbf{g})}).$$

The $\mathrm{U}_q(\mathbf{g})$-module structure of $\Gamma_q^{\mathbf{k}}(V)$ and the given condition that ψ is a $\mathrm{U}_q(\mathbf{g})$-module homomorphism immediately leads to

$$p(F\psi(w))=(-1)^{[\psi][p]}\psi(pw)(\mathbb{1}_{\mathrm{U}_q(\mathbf{g})})=(-1)^{[\psi][p]}F\psi(pw), \quad p\in\mathrm{U}_q(\mathbf{k}); \quad w\in W.$$

Now consider $\overline{F}$. We first show that the image $\mathrm{Im}(\overline{F})$ of $\overline{F}$ is contained in $\mathrm{Hom}_{\mathrm{U}_q(\mathbf{g})}(W,\Gamma_q^{\mathbf{k}}(V))$. Note that $\mathrm{Im}(\overline{F})\subset\mathrm{Hom}_{\mathbf{C}}(W,V\otimes\mathcal{T}_q(\mathbf{g}))$. Some relatively simple manipulations lead to

$$(x\cdot\overline{\phi}(w))=\overline{\phi}(xw),$$

$$(p\circ\overline{\phi}(w))=(S(p)\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})\overline{\phi}(w), \quad x\in\mathrm{U}_q(\mathbf{g}), \quad p\in\mathrm{U}_q(\mathbf{k}), \quad w\in W.$$

Therefore, $\mathrm{Im}(\overline{F})\subset\mathrm{Hom}_{\mathrm{U}_q(\mathbf{g})}(W,\Gamma_q^{\mathbf{k}}(V))$. Now we show that F and $\overline{F}$ are inverse to each other. For $\psi\in\mathrm{Hom}_{\mathrm{U}_q(\mathbf{g})}(W,\Gamma_q^{\mathbf{k}}(V))$, and $\phi\in\mathrm{Hom}_{\mathrm{U}_q(\mathbf{k})}(W,V)$, we have

$$(F\overline{F}\phi)(w)=(\overline{F}\phi)(w)(1_{\mathrm{U}_q(\mathbf{g})})=\phi(w),$$

$$\begin{aligned}(\overline{F}F\psi)(w)(x)&=(-1)^{[x]([w]+1)}(F\psi)(S(x)w)\\&=(-1)^{[x]([w]+1)}\psi(S(x)w)(1_{\mathrm{U}_q(\mathbf{g})})\\&=(-1)^{[x]([w]+[\psi]+1)}(S(x)\cdot\psi(w))(1_{\mathrm{U}_q(\mathbf{g})})=\psi(w)(x), \quad x\in\mathrm{U}_q(\mathbf{g}), \quad w\in W.\end{aligned}$$

This completes the proof of the Proposition. □

Let V_μ be a finite-dimensional irreducible $\mathrm{U}_q(\mathbf{p})$ module with highest weight μ and lowest weight $\tilde{\mu}$. Since V_μ is a $\mathrm{U}_q(\mathbf{p})$ module the following is a well-defined subspace of $\Gamma_q^{\mathbf{k}}(V_\mu)$,

$$\mathcal{O}_q(V_\mu):=\{\zeta\in\Gamma_q^{\mathbf{k}}(V_\mu)|p\circ\zeta=(S(p)\otimes\mathrm{id}_{\mathcal{T}_q(\mathbf{g})})\zeta, \quad \forall p\in\mathrm{U}_q(\mathbf{p})\}.$$

We may regard $\mathcal{O}_q(V_\mu)$ as the quantum analog of the space of ''holomorphic sections.'' Recall that the notation $W(\lambda)$ denotes the irreducible $\mathrm{U}_q(\mathbf{g})$ module with highest weight λ. We have the following result.

Theorem 6: *There exists the following* $\mathrm{U}_q(\mathbf{g})$ *module isomorphism*

$$\mathcal{O}_q(V_\mu)\cong\begin{cases}W((-\tilde{\mu})^\dagger), & -\tilde{\mu}\in\mathcal{P}_+,\\ 0, & \text{otherwise}.\end{cases} \tag{15}$$

Proof: Each $\zeta\in\mathcal{O}_q(V_\mu)$ can be expressed in the form

$$\zeta=\sum_{\lambda\in\mathcal{P}_+}\sum_{i,j} v_{ij}^{(\lambda)}\otimes\tilde{t}_{ij}^{(\lambda)},$$

for some $v_{ij}^{(\lambda)}\in V_\mu$ $(i,j=1,...,d_\lambda)$. Arguing as in the proof of Proposition 1 one concludes, for each $\lambda\in\mathcal{P}_+$, that there exist $\phi_i^{(\lambda)}\in\mathrm{Hom}_{\mathbf{C}}(W(\lambda),V_\mu)$ such that $v_{ij}^{(\lambda)}=\phi_i^{(\lambda)}(w_j^{(\lambda)})$, where $\{w_i^{(\lambda)}\}$ is the basis of $W(\lambda)$, relative to which the irreducible representation $t^{(\lambda)}$ of $\mathrm{U}_q(\mathbf{g})$ is defined. Thus we can rewrite ζ as

$$\zeta=\sum_{\lambda\in\mathcal{P}_+}\sum_{i,j} \phi_i^{(\lambda)}(w_j^{(\lambda)})\otimes\tilde{t}_{ij}^{(\lambda)}.$$

Similar reasoning as in the proof of Proposition 1 shows that the $\phi_i^{(\lambda)}$ must be $\mathrm{U}_q(\mathbf{p})$-module homomorphisms of degree $[\phi_i^{(\lambda)}]$. It immediately follows from (4) that

$$\phi_i^{(\lambda)}=c_i\phi^{(\lambda)},\quad c_i\in\mathbf{C},$$

and $\phi^{(\lambda)}$ may be nonzero only when

$$\overline{\lambda}=\tilde{\mu}.$$

Hence, if $-\tilde{\mu}\notin\mathcal{P}_+$, we have $\mathcal{O}_q(V_\mu)=0$. When $-\tilde{\mu}\in\mathcal{P}_+$, we set

$$\nu=(-\tilde{\mu})^{\dagger}.$$

Then, we may conclude that $\mathcal{O}_q(V_\mu)$ is spanned by

$$\zeta_i=\sum_j \phi^{(\nu)}(w_j^{(\nu)})\otimes\tilde{t}_{ij}^{(\lambda)},\tag{16}$$

which are obviously linearly independent. Furthermore,

$$x\cdot\zeta_i=(-1)^{[x][\phi^{(\nu)}]}\sum_j t_{ji}^{(\nu)}(x)\zeta_j,\quad x\in\mathrm{U}_q(\mathbf{g}).$$

Thus $\mathcal{O}_q(V_\mu)\cong W(\nu)$. More explicitly, the isomorphism is given by

$$W(\nu)\xrightarrow{(\mathrm{id}\otimes S)\delta}\mathcal{O}_q(W(\nu))\xrightarrow{\phi^{(\nu)}\otimes\mathrm{id}}\mathcal{O}_q(V_\mu).\tag{17}$$

This completes the proof of the theorem. □

This result provides an analog of the celebrated Borel–Weil theorem for the quantum supergroup $\mathrm{OSP}_q(1|2n)$. For the classical Lie supergroups, the program of developing a Bott–Borel–Weil theory was extensively investigated by Penkov and co-workers.[17] Also, a quantum Borel–Weil theorem for the covariant and contravariant tensor representations of quantum $\mathrm{GL}(m|n)$ was obtained in Ref. 9.

When $\mu=0$, the theorem implies that

$$\{f\in\mathcal{T}_q(\mathbf{g})|p\circ f=\epsilon(p)f,\quad\forall p\in\mathrm{U}_q(\mathbf{p})\}=\mathbf{C}\epsilon.$$

Combining this result with with Proposition 2, we obtain the following

Corollary: Let W be any finite-dimensional $\mathrm{U}_q(\mathbf{g})$ *module. Then, as* $\mathrm{U}_q(\mathbf{g})$*-modules,*

$$\mathcal{O}_q(W)\cong\epsilon\otimes W.$$

ACKNOWLEDGMENTS

This work is supported by the National Science Council (ROC) Grant No. NSC-87-2112-M008-002. Zhang wishes to thank the Department of Physics and the Center for Complex Systems at the National Central University for the hospitality extended to him during a visit from late 1997 to early 1998, when this work was largely completed.

[1] A. J. Bracken, M. D. Gould, and R. B. Zhang, Mod. Phys. Lett. A **5**, 831 (1990).
[2] M. Chaichian and P. P. Kulish, Phys. Lett. B **234**, 72 (1990); R. Floreanini, V. P. Spiridonov, and L. Vinet, Commun. Math. Phys. **137**, 149 (1991); M. Scheunert, Lett. Math. Phys. **34**, 320 (1993); H. Yamane, Proc. Jpn. Acad., Ser. A: Math. Sci. **70**, 31 (1994).
[3] J. H. H. Perk and C. L. Schultz, Phys. Lett. **84A**, 407 (1981); V. V. Bazhanov and A. G. Shadrikov, Theor. Math. Phys. **73**, 1302 (1987).
[4] M. D. Gould, R. B. Zhang, and A. J. Bracken, Bull Austral. Math. Soc. **47**, 353 (1993); H. Yamane, Proc. Jpn. Acad., Ser. A: Math. Sci. **67**, 108 (1991); S. M. Khoroshkin and V. N. Tolstoy, Commun. Math. Phys. **141**, 599 (1991).
[5] T. D. Palev and V. N. Tolstoy, Commun. Math. Phys. **141**, 549 (1991); R. B. Zhang, J. Math. Phys. **34**, 1236 (1993); R. B. Zhang, J. Phys. A **26**, 7041 (1993); T. D. Palev, N. I. Stoilova, and J. Van der Jeugt, Commun. Math. Phys. **166**, 367 (1994); R. B. Zhang, Lett. Math. Phys. **37**, 419 (1996); R. B. Zhang, J. Math. Phys. **38**, 535 (1997).
[6] R. B. Zhang, A. J. Bracken, and M. D. Gould, Phys. Lett. B **257**, 133 (1991); A. J. Bracken, M. D. Gould, J. R. Links, and Y. Z. Zhang, Phys. Rev. Lett. **74**, 2768 (1995); M. J. Martins, Nucl. Phys. B **450**, 768 (1995).
[7] H. C. Lee, NATO ASI Ser., Ser. B **245**, 359 (1990); R. B. Zhang, J. Math. Phys. **33**, 3918 (1992); M. D. Gould, I. Tsohantjis, and A. J. Bracken, Rev. Math. Phys. **5**, 533 (1993); J. R. Links, M. D. Gould, and R. B. Zhang, Rev. Math. Phys. **5**, 345 (1993).
[8] R. B. Zhang, Rev. Math. Phys. **7**, 809 (1995); R. B. Zhang and H. C. Lee, Mod. Phys. Lett. A **11**, 2397 (1996).
[9] R. B. Zhang, Commun. Math. Phys. **195**, 525 (1998).
[10] C. Fronsdal, *Essays on supersymmetry*, Mathematical Physics Studies Vol. 8 (Reidel, Dordrecht, 1986).
[11] P. K. Townsend, ''*M*-theory from its superalgebras,'' hep-th/9712004; M. Gunaydin, ''Unitary supermultiplets of Osp (1/32, *R*) and *M* theory,'' hep-th/9803138.
[12] M. Scheunert and R. B. Zhang, J. Math. Phys. **39**, 5024 (1998).
[13] R. B. Zhang, Lett. Math. Phys. **25**, 317 (1992).
[14] T. D. Palev, J. Phys. A **26**, L1111 (1993); T. D. Palev and J. Van der Jeugt, J. Phys. A **28**, 2605 (1995).
[15] A. Connes, *Noncommutative geometry* (Academic, New York, 1994).
[16] Y. I. Manin, *Gauge field theory and complex geometry* (Springer-Verlag, Berlin, 1988).
[17] I. B. Penkov, Sov. Probl. Math. **32**, (1988) I. B. Penkov and V. Serganova, Annales de L'institut Fourier **39**, 846 (1989).
[18] V. Chari and A. Pressley, *A guide to quantum groups* (Cambridge U.P., Cambridge, 1994).
[19] A. R. Gover and R. B. Zhang, ''Geometry of quantum homogeneous vector bundles and representation theory of quantum groups. I,'' Rev. Math. Phys. (in press).

Toroidal and level 0 $U'_q(\widehat{sl_{n+1}})$ actions on $U_q(\widehat{gl_{n+1}})$ modules

Kei Miki
Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka 560, Japan

(Received 14 August 1998; accepted for publication 17 March 1999)

(1) Utilizing a braid group action on a completion of $U_q(\widehat{sl_{n+1}})$, an algebra homomorphism from the toroidal algebra $U_q(sl_{n+1,\mathrm{tor}})$ $(n\geqslant 2)$ to a completion of $U_q(\widehat{gl_{n+1}})$ is obtained. (2) The toroidal actions by Saito induces a level 0 $U'_q(\widehat{sl_{n+1}})$ action on level 1 integrable highest weight modules of $U_q(\widehat{sl_{n+1}})$. Another level 0 $U'_q(\widehat{sl_{n+1}})$ action was defined by Jimbo *et al.*, in the case $n=1$. Using the fact that the intertwiners of $U_q(\widehat{sl_{n+1}})$ modules are intertwiners of toroidal modules for an appropriate comultiplication, the relation between these two level 0 $U'_q(\widehat{sl_{n+1}})$ actions is clarified.
[S0022-2488(99)04806-9]

I. INTRODUCTION

In Refs. 1 and 2, a quantum toroidal algebra $U_q(sl_{n+1,\mathrm{tor}})$ was introduced. Up to now several results have been obtained on this algebra. In Ref. 2 the connection between toroidal modules and an extension of the double affine Hecke algebra[3] was noticed and the Schur-type duality was obtained. The vertex representations were constructed on level 1 $U_q(\widehat{gl_{n+1}})$ modules by Saito.[4] In Ref. 5, toroidal actions were shown to be defined on any integrable highest weight module of $U_q(\widehat{gl_{n+1}})$, using the level—rank duality. Since the toroidal algebra has homomorphic images of $U_q(\widehat{sl_{n+1}})$ and $U'_q(\widehat{sl_{n+1}})$, $U_q(\widehat{sl_{n+1}})$ and $U'_q(\widehat{sl_{n+1}})$ actions are defined on toroidal modules. The known $U'_q(\widehat{sl_{n+1}})$ actions obtained in this way have level 0 (see Ref. 6). Therefore level 0 $U'_q(\widehat{sl_{n+1}})$ actions on $U_q(\widehat{sl_{n+1}})$ modules are closely related to toroidal modules. In Refs. 7 and 8, the level 1 $U_q(\widehat{sl_{n+1}})$ action on the fermionic Fock space[9] and the level 0 $U'_q(\widehat{sl_{n+1}})$ action via the affine Hecke algebra were shown to be combined into a toroidal action. In Ref. 10, motivated by Ref. 11, a level 0 $U'_q(\widehat{sl_2})$ action was defined on level 1 integrable $U_q(\widehat{sl_2})$ modules, utilizing the intertwiners and the representation of the affine Hecke algebra.

In this paper we obtain two results on these problems. In Sec. III, utilizing a braid group action on a completion of $U_q(\widehat{sl_{n+1}})$,[12] an algebra homomorphism from the toroidal algebra $U_q(sl_{n+1,\mathrm{tor}})$ $(n\geqslant 2)$ to a completion of $U_q(\widehat{gl_{n+1}})$ is constructed. This implies that any highest weight module of $U_q(\widehat{gl_{n+1}})$ is a toroidal module. This result corresponds to the fact that the algebra homomorphism from $U_q(\widehat{sl_{n+1}})$ to $U_q(gl_{n+1})$ by Jimbo[13] is neatly expressed in terms of the braid group action by Lusztig.[14] In Secs. IV and V, assuming a triangular decomposition of the toroidal algebra, we consider the level 1 toroidal modules by Saito. Utilizing the fact that the intertwiners of $U_q(\widehat{sl_{n+1}})$ modules are intertwiners of toroidal modules for an appropriate comultiplication, we clarify the relation between the level 0 $U'_q(\widehat{sl_{n+1}})$ action induced by the toroidal action and that in Ref. 10. Note that in Ref. 11, first, a Yangian action on level 1 integrable highest weight modules of $\widehat{sl}_2$ was constructed in terms of the currents and then the intertwining property of the vertex operators was used. Therefore our approach is closer to the original one. Clarifying the connection between our results and Refs. 5, 7, and 8 would be interesting.

II. DEFINITION OF ALGEBRAS

Let q be an indeterminate and set $F=\mathbf{Q}(q)$. Fix $n\geqslant 2$.

Let $(a_{ij})_{1\leqslant i,j\leqslant n}$ be the Cartan matrix of type A_n and set $\kappa_{ij}=1$. Let $\mathrm{U}_q(\widehat{sl_{n+1}})$ (Ref. 15) be the F algebra defined by generators $E_{i,m}$, $F_{i,m}$, $h_{i,r}$, $k_i^{\pm 1}$, $C^{\pm 1}$, $D^{\pm 1}$ $(1\leqslant i\leqslant n,\ m\in\mathbf{Z},\ r\in\mathbf{Z}\setminus\{0\})$ and relations,

$$C^{\pm 1}\text{central},\quad C^{\pm 1}C^{\mp 1}=D^{\pm 1}D^{\mp 1}=1,\quad DX_i(z)D^{-1}=X_i(z/q)\ \ (X=E,F), \tag{2.1}$$

$$k_i^{\pm 1}k_i^{\mp 1}=1,\quad D\phi_i^{\pm}(z)D^{-1}=\phi_i^{\pm}(z/q), \tag{2.2}$$

$$\phi_i^{\pm}(z)\phi_j^{\pm}(w)=\phi_j^{\pm}(w)\phi_i^{\pm}(z), \tag{2.3}$$

$$\frac{1-q^{-a_{ij}}C^{-1}\kappa_{ij}z/w}{1-q^{a_{ij}}C^{-1}\kappa_{ij}z/w}\,\phi_i^{+}(z)\phi_j^{-}(w)=\frac{1-q^{-a_{ij}}C\kappa_{ij}z/w}{1-q^{a_{ij}}C\kappa_{ij}z/w}\,\phi_j^{-}(w)\phi_i^{+}(z), \tag{2.4}$$

$$\phi_i^{\pm}(z)E_j(w)\phi_i^{\pm}(z)^{-1}=q^{\mp a_{ij}}\frac{1-q^{\pm a_{ij}}C^{-(1/2)\mp(1/2)}(\kappa_{ij}z/w)^{\pm 1}}{1-q^{\mp a_{ij}}C^{-(1/2)\mp(1/2)}(\kappa_{ij}z/w)^{\pm 1}}E_j(w), \tag{2.5}$$

$$\phi_i^{\pm}(z)F_j(w)\phi_i^{\pm}(z)^{-1}=q^{\pm a_{ij}}\frac{1-q^{\mp a_{ij}}C^{(1/2)\mp(1/2)}(\kappa_{ij}z/w)^{\pm 1}}{1-q^{\pm a_{ij}}C^{(1/2)\mp(1/2)}(\kappa_{ij}z/w)^{\pm 1}}F_j(w), \tag{2.6}$$

$$[E_i(z),F_j(w)]=\frac{\delta_{ij}}{q-q^{-1}}(\delta(Cw/z)\phi_i^{-}(z)-\delta(Cz/w)\phi_i^{+}(w)), \tag{2.7}$$

$$q^{a_{ij}}(1-q^{-a_{ij}}\kappa_{ij}z/w)E_i(z)E_j(w)=(1-q^{a_{ij}}\kappa_{ij}z/w)E_j(w)E_i(z), \tag{2.8}$$

$$q^{-a_{ij}}(1-q^{a_{ij}}\kappa_{ij}z/w)F_i(z)F_j(w)=(1-q^{-a_{ij}}\kappa_{ij}z/w)F_j(w)F_i(z); \tag{2.9}$$

for i,j such that $a_{ij}=-1$,

$$E_i(z_1)E_i(z_2)E_j(w)-(q+q^{-1})E_i(z_1)E_j(w)E_i(z_2)+E_j(w)E_i(z_1)E_i(z_2)+(z_1\leftrightarrow z_2)=0, \tag{2.10}$$

$$F_i(z_1)F_i(z_2)F_j(w)-(q+q^{-1})F_i(z_1)F_j(w)F_i(z_2)+F_j(w)F_i(z_1)F_i(z_2)+(z_1\leftrightarrow z_2)=0; \tag{2.11}$$

for i,j such that $a_{ij}=0$,

$$[E_i(z),E_j(w)]=0,\quad [F_i(z),F_j(w)]=0, \tag{2.12}$$

where

$$E_i(z)=\sum_{m\in\mathbf{Z}}E_{i,m}/z^m,\quad F_i(z)=\sum_{m\in\mathbf{Z}}F_{i,m}/z^m,$$

$$\phi_i^{\pm}(z)=k_i^{\mp 1}\exp\left(\mp(q-q^{-1})\sum_{r>0}h_{i,\mp r}z^{\pm r}\right). \tag{2.13}$$

As is well known, this algebra is also described by the Chevally generators e_i, f_i, $k_i^{\pm 1}(0\leqslant i\leqslant n)$, and $D^{\pm 1}$. Later we need its comultiplication Δ_0 determined by

$$\Delta_0(e_i)=e_i\otimes 1+k_i\otimes e_i,\quad \Delta_0(f_i)=f_i\otimes k_i^{-1}+1\otimes f_i,$$

$$\Delta_0(k_i)=k_i\otimes k_i,\quad \Delta_0(D)=D\otimes D. \tag{2.14}$$

$\mathrm{U}_q(\widehat{gl_{n+1}})$ (Ref. 16) is defined to be the F algebra generated by $E_{i,m}$, $F_{i,m}$, $a_{k,r}$, $t_k^{\pm 1}$, $C^{\pm 1}$, $D^{\pm 1}$ ($1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant n+1$, $m \in \mathbf{Z}$, $r \in \mathbf{Z} \backslash \{0\}$) with defining relations (2.1), (2.7)–(2.12) and the following:

$$t_k^{\pm 1} t_k^{\mp 1} = 1, \quad D \psi_k^{\pm}(z) D^{-1} = \psi_k^{\pm}(z/q), \tag{2.15}$$

$$\psi_k^{\pm}(z) \psi_l^{\pm}(w) = \psi_l^{\pm}(w) \psi_k^{\pm}(z), \tag{2.16}$$

$$\begin{aligned} &\frac{1-q^{-2}C^{-1}z/w}{1-C^{-1}z/w} \frac{1-q^{2\theta(k>l)}C^{-1}z/w}{1-q^{-2\theta(k<l)}C^{-1}z/w} \psi_k^{+}(z) \psi_l^{-}(w) \\ &= \frac{1-Cz/w}{1-q^2 Cz/w} \frac{1-q^{2\theta(k>l)}Cz/w}{1-q^{-2\theta(k<l)}Cz/w} \psi_l^{-}(w) \psi_k^{+}(z), \end{aligned} \tag{2.17}$$

$$\psi_k^{\pm}(z) E_j(w) \psi_k^{\pm}(z)^{-1} = q^{\mp b_{kj}} \frac{1-q^{\pm(k-(1/2)+(1/2)b_{kj})} C^{-(1/2)\mp(1/2)} (z/w)^{\pm 1}}{1-q^{\pm(k-(1/2)-(3/2)b_{kj})} C^{-(1/2)\mp(1/2)} (z/w)^{\pm 1}} E_j(w), \tag{2.18}$$

$$\psi_k^{\pm}(z) F_j(w) \psi_k^{\pm}(z)^{-1} = q^{\pm b_{kj}} \frac{1-q^{\pm(k-(1/2)-(3/2)b_{kj})} C^{(1/2)\mp(1/2)} (z/w)^{\pm 1}}{1-q^{\pm(k-(1/2)+(1/2)b_{kj})} C^{(1/2)\mp(1/2)} (z/w)^{\pm 1}} F_j(w). \tag{2.19}$$

Here $b_{kj} = \delta_{kj} - \delta_{kj+1}$, $\theta(\cdot)$ is a step function, and

$$\psi_k^{\pm}(z) = t_k^{\mp 1} \exp\left(\mp (q-q^{-1}) \sum_{r>0} a_{k,\mp r} z^{\pm r} \right). \tag{2.20}$$

In Eq. (2.7), $\phi_i^{\pm}(z)$ should be understood as follows:

$$\phi_i^{\pm}(q^i z) = \psi_i^{\pm}(z) / \psi_{i+1}^{\pm}(z). \tag{2.21}$$

Let $(a_{ij})_{0 \leqslant i,j \leqslant n}$ be the Cartan matrix of type $A_n^{(1)}$ and set $\kappa_{ij} = 1$ $((i,j) \neq (n,0),(0,n))$, $\kappa_{n0} = \kappa_{0n}^{-1} = \kappa$. Let $\mathrm{U}_q(sl_{n+1,\mathrm{tor}})$ (Refs. 1 and 2) be the F algebra defined by generators $E_{i,m}$, $F_{i,m}$, $h_{i,r}$, $k_i^{\pm 1}$, $C^{\pm 1}$, $D^{\pm 1}$ ($0 \leqslant i \leqslant n$, $m \in \mathbf{Z}$, $r \in \mathbf{Z} \backslash \{0\}$), $\kappa^{\pm 1}$ and relations (2.1)–(2.12) and

$$\kappa^{\pm 1} \text{ central}, \quad \kappa^{\pm 1} \kappa^{\mp 1} = 1. \tag{2.22}$$

Note that we include only one scaling element D and its inverse among the generators.

Hereafter we shall write U for $\mathrm{U}_q(\widehat{sl_{n+1}})$ and let U^+, U^-, and U^0 denote the subalgebras of U generated by e_i ($0 \leqslant i \leqslant n$), f_i ($0 \leqslant i \leqslant n$), and $k_i^{\pm 1}$ ($0 \leqslant i \leqslant n$) and $D^{\pm 1}$, respectively.

At the end of the section, for completeness, we shall prove the following lemma, though it would be well known. The first claim enables us to identify U with the subalgebra of $\mathrm{U}_q(\widehat{gl_{n+1}})$ generated by $E_{i,m}$, $F_{i,m}$, $h_{i,r}$, $k_i^{\pm 1}$ ($1 \leqslant i \leqslant n$), $C^{\pm 1}$, and $D^{\pm 1}$.

Lemma 1: (1) The algebra homomorphism i: $U \to U_q(\widehat{gl_{n+1}})$ *such that*

$$\begin{aligned} &E_i(z) \mapsto E_i(z), \quad F_i(z) \mapsto F_i(z), \\ &\phi_i^{\pm}(q^i z) \mapsto \psi_i^{\pm}(z) \psi_{i+1}^{\pm}(z)^{-1}, \quad C \mapsto C, \quad D \mapsto D, \end{aligned} \tag{2.23}$$

is injective.

(2) Set $\beta_r = \sum_{k=1}^{n+1} q^{(2k-n-2)r} a_{k,r} / [(n+1)r]$ *where* $[m] = (q^m - q^{-m})/(q-q^{-1})$. *Let* $A^{\pm}$ *(resp.* A^0*) be the subalgebras of* $U_q(\widehat{gl_{n+1}})$ *generated by* $i(U^{\pm})$ *and* $\beta_{\pm r}$ $(r>0)$ *(resp.* $t_k^{\pm 1}$, $C^{\pm 1}$, *and* $D^{\pm 1}$*). Then the multiplication map* $A^- \otimes A^0 \otimes A^+ \to U_q(\widehat{gl_{n+1}})$ *is an isomorphism of vector spaces.*

Before proving the lemma, we shall introduce some notations, which we shall also need in Sec. IV A. Let $\mathcal{A}$ be the F algebra defined by generators $E_{i,m}$, $F_{i,m}$, $h_{i,r}$, $t_k^{\pm 1}$, $C^{\pm 1}$, $D^{\pm 1}$ ($1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant n+1$, $m \in \mathbf{Z}$, $r \in \mathbf{Z} \backslash \{0\}$) and the relations

$$t_k^{\pm 1} t_k^{\mp 1} = 1, \quad t_k t_l = t_l t_k, \quad D t_k = t_k D, \quad t_k \phi_i^{\pm}(z) t_k^{-1} = \phi_i^{\pm}(z), \tag{2.24}$$

$$t_k E_j(z) t_k^{-1} = q^{b_{kj}} E_j(z), \quad t_k F_j(z) t_k^{-1} = q^{-b_{kj}} F_j(z), \tag{2.25}$$

and (2.1)–(2.12), where $(a_{ij})_{1 \leqslant i,j \leqslant n}$ is the Cartan matrix of type A_n, $\kappa_{ij}=1$ and $k_i = t_i t_{i+1}^{-1}$. Let further $\mathcal{B}$ be the F algebra generated by b_r ($r \in \mathbf{Z} \backslash \{0\}$), $C^{\pm 1}$, and $D^{\pm 1}$ with relations

$$C^{\pm 1} \text{ central}, \quad C^{\pm 1} C^{\mp 1} = D^{\pm 1} D^{\mp 1} = 1, \quad D b_r D^{-1} = q^r b_r,$$

$$[b_r, b_s] = \delta_{r+s,0} \frac{(q^{n+1}C)^r - (q^{n+1}C)^{-r}}{q - q^{-1}} \frac{[r]^2}{r[(n+1)r]}. \tag{2.26}$$

Finally let $j: \mathrm{U}_q(\widehat{gl_{n+1}}) \to \mathcal{A} \otimes \mathcal{B} / \langle C \otimes 1 - 1 \otimes C \rangle$ be the algebra homomorphism determined by

$$E_i(z) \mapsto E_i(z) \dot{\otimes} 1, \quad F_i(z) \mapsto F_i(z) \dot{\otimes} 1,$$

$$\psi_k^{\pm}(z) \mapsto (t_k^{\mp 1} \dot{\otimes} 1) \exp\left(\mp (q - q^{-1}) \sum_{r>0} \hat{a}_{k,\mp r} z^{\pm r} \right),$$

$$C \mapsto C \dot{\otimes} 1, \quad D \mapsto D \dot{\otimes} D, \tag{2.27}$$

where $x \dot{\otimes} y$ denotes the image of $x \otimes y$ under the quotient map $\mathcal{A} \otimes \mathcal{B} \to \mathcal{A} \otimes \mathcal{B} / \langle C \otimes 1 - 1 \otimes C \rangle$ and $\hat{a}_{k,r} = \overline{a}_{k,r} \dot{\otimes} 1 + 1 \dot{\otimes} b_r$ with

$$\overline{a}_{k,r} = \frac{1}{[(n+1)r]} \left(\sum_{i=k}^{n} [(n+1-i)r] h_{i,r} - q^{-(n+1)r} \sum_{i=1}^{k-1} [ir] h_{i,r} \right). \tag{2.28}$$

Proof of Lemma 1: Letting ζ be the algebra homomorphism from U to $\mathcal{A}$ such that

$$X_i(z) \mapsto X_i(z) \quad (X = E, F, \phi^{\pm}), \quad C \mapsto C, \quad D \mapsto D, \tag{2.29}$$

set $\mathcal{A}^{\pm} = \zeta(\mathrm{U}^{\pm})$. Let further $\mathcal{A}^0$ be the subalgebra of $\mathcal{A}$ generated by $t_k^{\pm 1}$, $C^{\pm 1}$, and $D^{\pm 1}$. Then Refs. 17 and 18 imply that $\mathcal{A}^- \otimes \mathcal{A}^0 \otimes \mathcal{A}^+ \simeq \mathcal{A}$, $\mathcal{A}^{\pm} \simeq \mathrm{U}^{\pm}$ and the vectors $\Pi_k t_k^{m_k} C^a D^b$ ($m_k, a, b \in \mathbf{Z}$) form a basis of $\mathcal{A}^0$. Note that this triangular decomposition of $\mathcal{A}$ proves the injectivity of ζ. Let $\mathcal{B}^{\pm}$ (resp. $\mathcal{B}^0$) be the subalgebras of $\mathcal{B}$ generated by $b_{\pm r}$ ($r>0$) (resp. $C^{\pm 1}$ and $D^{\pm 1}$). Then $\mathcal{B}^- \otimes \mathcal{B}^0 \otimes \mathcal{B}^+ \simeq \mathcal{B}$ and the vectors $\Pi_{r>0} b_{\pm r}^{m_r}$ ($m_r \in \mathbf{Z}_{\geqslant 0}$, $m_r = 0$ if $r \gg 0$) and $C^a D^b$ ($a, b \in \mathbf{Z}$) form bases of $\mathcal{B}^{\pm}$ and $\mathcal{B}^0$, respectively. Set $\mathcal{C}^0 = \mathcal{A}^0 \otimes \mathcal{B}^0 / \langle C \otimes 1 - 1 \otimes C \rangle$. Then from the above properties of $\mathcal{A}$ and $\mathcal{B}$ we find that $\mathcal{A} \otimes \mathcal{B} / \langle C \otimes 1 - 1 \otimes C \rangle$ has the triangular decomposition

$$(\mathcal{A}^- \otimes \mathcal{B}^-) \otimes \mathcal{C}^0 \otimes (\mathcal{A}^+ \otimes \mathcal{B}^+) \simeq \mathcal{A} \otimes \mathcal{B} / \langle C \otimes 1 - 1 \otimes C \rangle, \tag{2.30}$$

and a basis of $\mathcal{C}^0$ is given by the elements $\Pi_k t_k^{m_k} C^a D^b \otimes D^c$ ($m_k, a, b, c \in \mathbf{Z}$).

Let $\xi: \mathcal{A} \to \mathrm{U}_q(\widehat{gl_{n+1}})$ be the algebra homomorphism such that

$$X_i(z) \mapsto X_i(z) \quad (X = E, F), \quad \phi_i^{\pm}(q^i z) \mapsto \psi_i^{\pm}(z) \psi_{i+1}^{\pm}(z)^{-1},$$

$$t_k \to t_k, \quad C \mapsto C, \quad D \mapsto D. \tag{2.31}$$

From the above triangular decompositions of $\mathcal{A}$ and $\mathcal{A}\otimes\mathcal{B}/\langle C\otimes 1-1\otimes C\rangle$, we find that $j\circ\xi$ and, hence, $i=\xi\circ\zeta$ are injective. Since β_r's commute with $i(\mathrm{U}^{\pm})$, we have $A^{\pm}=\Sigma_{m_r}i(\mathrm{U}^{\pm})\Pi_{r>0}\beta_{\pm r}^{m_r}$. Noting this, $j(i(y))=\zeta(y)\dot{\otimes}1(y\in\mathrm{U}^{\pm})$ and $j(\beta_r)=1\dot{\otimes}b_r$, we find that the map j induces the isomorphisms $A^{\pm}\simeq\mathcal{A}^{\pm}\otimes\mathcal{B}^{\pm}$. This map j also induces the injective map $\mathcal{A}^0\rightarrow\mathcal{C}^0$. Therefore since $\mathrm{U}_q(\widehat{gl_{n+1}})=A^-A^0A^+$, we obtain (2). ∎

III. HOMOMORPHISM FROM $\mathrm{U}_q(sl_{n+1,\mathrm{tor}})$ TO A COMPLETION OF $\mathrm{U}_q(\widehat{gl_{n+1}})$

A. Braid group action on a completion of $\mathrm{U}_q(\widehat{sl_{n+1}})$

Setting

$$\mathrm{U}_r=\{y\in\mathrm{U}|DyD^{-1}=q^r y\},\quad (r\in\mathbf{Z}),\tag{3.1}$$

put

$$\mathrm{U}_{kl}=\sum_{\substack{r\geqslant k\\ s\geqslant l}}\mathrm{U}_{-r}\mathrm{U}\mathrm{U}_s,\quad (k,l\in\mathbf{Z}_{\geqslant 0}).\tag{3.2}$$

We introduce a linear topology on U by letting (U_{kl}) be a fundamental system of neighborhoods of the origin. Then U is a topological algebra. Define $\mathrm{U}_r^{\pm}$ similarly to U_r. Then the following holds:

$$\mathrm{U}_{kl}=\sum_{r\geqslant k}\mathrm{U}_{-r}^{-}\mathrm{U}^0\sum_{s\geqslant l}\mathrm{U}_s^{+}.\tag{3.3}$$

Therefore U is separated thanks to the triangular decomposition of U.[18] We shall denote the completion of U by $\hat{\mathrm{U}}$.

Let T_i $(1\leqslant i\leqslant n)$ (Ref. 12) be the continuous algebra automorphisms of $\hat{\mathrm{U}}$ determined by

$$E_i(z)\mapsto -F_i(z/Cq^2)\phi_i^-(z/q^2)^{-1},\quad F_i(z)\mapsto -\phi_i^+(z/q^2)^{-1}E_i(z/Cq^2),$$
$$\phi_i^{\pm}(z)\mapsto\phi_i^{\pm}(z/q^2)^{-1},\quad C^{\pm 1}\mapsto C^{\pm 1},\quad D^{\pm}\mapsto D^{\pm 1},\tag{3.4}$$

$$E_j(z)\mapsto\oint\frac{du}{u}\left(E_i(u)E_j(z)-q\frac{1-q^{-1}u/z}{1-qu/z}E_j(z)E_i(u)\right),$$

$$F_j(z)\mapsto\oint\frac{du}{u}\left(F_j(z)F_i(u)-q^{-1}\frac{1-qz/u}{1-q^{-1}z/u}F_i(u)F_j(z)\right),$$

$$\phi_j^{\pm}(z)\mapsto\phi_j^{\pm}(z)\phi_i^{\pm}(z/q),\ \text{when}\ |i-j|=1,\tag{3.5}$$

$$E_j(z)\mapsto E_j(z),\quad F_j(z)\mapsto F_j(z),\quad \phi_j^{\pm}(z)\mapsto\phi_j^{\pm}(z),\ \text{when}\ |i-j|>1.\tag{3.6}$$

Here $\oint(du/u)$ denotes the operation which picks out the coefficient of u^0. Then they satisfy the Coxeter relations,

$$T_iT_jT_i=T_jT_iT_j\ \text{when}\ |i-j|=1,$$
$$T_iT_j=T_jT_i\ \text{when}\ |i-j|>1,\tag{3.7}$$

and have the following property:

$$T_iT_{i\pm 1}X_i(z)=X_{i\pm 1}(z/q),\quad (X=E,F,\phi^{\pm}).\tag{3.8}$$

Note that the inverse T_i^{-1} is given by $\eta\circ T_i\circ\eta$ where η is the continuous algebra antiautomorphism of $\hat{\mathrm{U}}$ determined by

$$E_i(z)\mapsto E_i(z^{-1}),\quad F_i(z)\mapsto F_i(z^{-1}),\quad \phi_i^{\pm}(z)\mapsto\phi_i^{\mp}(Cz^{-1})\quad C^{\pm1}\mapsto C^{\pm1},\quad D^{\pm1}\mapsto D^{\pm1}. \tag{3.9}$$

B. Homomorphism from $\mathrm{U}_q(sl_{n+1,\mathrm{tor}})$ to a completion of $\mathrm{U}_q(\widehat{gl_{n+1}})$

As in the case U, we introduce a separated linear topology on $\mathrm{U}_q(\widehat{gl_{n+1}})$ and denote its completion by $\hat{\mathrm{U}}_q(\widehat{gl_{n+1}})$. Equation (3.3), the corresponding one for $\mathrm{U}_q(\widehat{gl_{n+1}})_{kl}$ and Lemma 1 give $\mathrm{U}_{kl}=\mathrm{U}\cap\mathrm{U}_q(\widehat{gl_{n+1}})_{kl}$. Therefore we can identify $\hat{\mathrm{U}}$ with the subalgebra of $\hat{\mathrm{U}}_q(\widehat{gl_{n+1}})$. Utilizing the above braid group action on $\hat{\mathrm{U}}$, we obtain the following theorem:

Theorem 1: *Set*

$$\tilde{E}_0^{(\epsilon)}(z)=-T_n^{\epsilon}\cdots T_1^{\epsilon}E_1(zq^{\epsilon}),\quad \tilde{F}_0^{(\epsilon)}(z)=-T_n^{\epsilon}\cdots T_1^{\epsilon}F_1(zq^{\epsilon}),$$

$$\tilde{\phi}_0^{\pm,(\epsilon)}(z)=T_n^{\epsilon}\cdots T_1^{\epsilon}\phi_1^{\pm}(zq^{\epsilon}),\quad (\epsilon=\pm1). \tag{3.10}$$

For $\epsilon=\pm1$, there exists an algebra homomorphism $f_\epsilon:U_q(sl_{n+1,tor})\to\hat{U}_q(\widehat{gl_{n+1}})$ determined by

$$E_i(z)\mapsto E_i(z),\quad F_i(z)\mapsto F_i(z),\quad \phi_i^{\pm}(z)\mapsto\phi_i^{\pm}(z),\quad (1\leqslant i\leqslant n),$$

$$C\mapsto C,\quad D\mapsto D,$$

$$E_0(z)\mapsto\psi_{n+1}^{+}(z/\mu_\epsilon)^{\epsilon}\tilde{E}_0^{(\epsilon)}(z)\psi_{n+1}^{-}(Cz/\mu_\epsilon)^{-\epsilon},$$

$$F_0(z)\mapsto\psi_{n+1}^{+}(Cz/\mu_\epsilon)^{-\epsilon}\tilde{F}_0^{(\epsilon)}(z)\psi_{n+1}^{-}(z/\mu_\epsilon)^{\epsilon},$$

$$\phi_0^{\pm}(z)\mapsto\tilde{\phi}_0^{\pm,(\epsilon)}(z)\psi_{n+1}^{\pm}(z/C\mu_\epsilon)^{\epsilon}\psi_{n+1}^{\pm}(Cz/\mu_\epsilon)^{-\epsilon},$$

$$\kappa\mapsto(\mu_0C^2)^{\epsilon}, \tag{3.11}$$

where $\mu_0=q^{n+1}$ and $\mu_\epsilon=\mu_0(\mu_0C)^{\epsilon}$.

Proof: As an example, we shall check the Serre relation (2.10) with $(i,j)=(0,n)$ for the case $\epsilon=1$. Set $T_w=T_n\cdots T_1$. Thanks to (3.8), we get

$$T_w^{-1}E_{n-1}(z/q)=E_n(z),\quad T_w^{-1}E_n(z/q)=-\tilde{\phi}_0^{+,(+)}(\kappa z/C)\tilde{E}_0^{(+)}(\kappa z)\tilde{\phi}_0^{-,(+)}(\kappa z)^{-1}, \tag{3.12}$$

where $\kappa=\mu_0C^2$. Therefore applying T_w^{-1} to the Serre relation (2.10) with $(i,j)=(n,n-1)$ of $\mathrm{U}_q(\widehat{gl_{n+1}})$ and using the relations among $\psi_k^{\pm}(z)$, $E_i(z)$, and $F_i(z)$, we obtain the desired relation. ■

Corollary 1: Let V be a $U_q(\widehat{gl_{n+1}})$ module such that

$$V=\oplus_{m\in\mathbf{Z}}V_m,\quad V_m=\{v\in V|Dv=q^m v\},\quad V_m=\{0\}\quad (m\gg0).$$

If C acts as $C_0\in F^{\times}$ on V, then V has two $U_q(sl_{n+1,tor})$ module structures where κ acts as $\mu_0C_0^2$ and $(\mu_0C_0^2)^{-1}$, *respectively.*

Corollary 2: The algebra homomorphism $g:U\to U_q(sl_{n+1,tor})$ such that

$$E_i(z)\mapsto E_i(z),\quad F_i(z)\mapsto F_i(z),\quad \phi_i^{\pm}(z)\mapsto\phi_i^{\pm}(z),\quad C\mapsto C,\quad D\mapsto D, \tag{3.13}$$

is injective.

Proof: The claim follows from the fact that the composite $f_\epsilon \circ g$ is the identity map from U to the subalgebra U of $\hat{\mathrm{U}}_q(\widehat{gl_{n+1}})$. ■

IV. INTERTWINERS OF TOROIDAL MODULES

A. Level 1 toroidal modules by Saito

For $\mu \in F^\times$, set $\mathcal{U}_\mu = \mathrm{U}_q(sl_{n+1,\mathrm{tor}})/\langle \kappa - \mu \rangle$. Hereafter we consider this algebra. We shall often use $p = \mu_0 \mu$ instead of μ.

Let Λ_i $(0 \leqslant i \leqslant n)$ be the fundamental weights of $\widehat{sl_{n+1}}$. Let further $V(\Lambda_i)$ be the irreducible highest weight module of U with highest weight Λ_i and v_{Λ_i} be its highest weight vector. Set $\mathcal{B}_1 = \mathcal{B}/I$, where I is the left ideal generated by b_r $(r>0)$, $C-q$ and $D-1$. For $m \in \mathbf{Z}$, set $W_m = V(\Lambda_j) \otimes \mathcal{B}_1$ where $j \equiv m \bmod n+1$. Let α_i's be the simple roots of $\widehat{sl_{n+1}}$ and $(\ |\ )$ be the standard symmetric bilinear form on the weight space of $\widehat{sl_{n+1}}$ normalized by $(\alpha_i | \alpha_i) = 2$. Set

$$\epsilon_k = \left(\sum_{i=k}^{n} (n+1-i)\alpha_i - \sum_{i=1}^{k-1} i\alpha_i \right)/(n+1), \quad (1 \leqslant k \leqslant n+1).$$

For $1 \leqslant k \leqslant n+1$, let ∂_{ϵ_k} be a linear operator on $V(\Lambda_j)$ such that $\partial_{\epsilon_k} v = (\epsilon_k | \nu) v$, where v is a weight vector with weight ν. Then W_m is a $\mathrm{U}_q(\widehat{gl_{n+1}})$ module via the map j if the U module $V(\Lambda_j)$ is extended to a $\mathcal{A}$ module by letting t_k acts as $q^{\partial_{\epsilon_k} + [m/(n+1)]}$.

The results by Saito[4] on level 1 toroidal modules can be stated as follows.

Proposition 1: Set $a = (-1)^{n-1} q^n$.

(1) For $\epsilon = \pm 1$, *the U module structure on* $V(\Lambda_j)$ $(0 \leqslant j \leqslant n)$ *is extended to a* $\mathcal{U}_{\mu_0^\epsilon}$ *module structure by*

$$E_0(z) = a \tilde{E}_0^{(\epsilon)}(z), \quad F_0(z) = a^{-1} \tilde{F}_0^{(\epsilon)}(z), \quad \phi_0^{\pm}(z) = \tilde{\phi}_0^{\pm,(\epsilon)}(z). \tag{4.1}$$

(2) Letting α_r $(r \in \mathbf{Z} \backslash \{0\})$ *be the elements of* $F^\times$ *such that*

$$\alpha_r \alpha_{-r} = \frac{\mu^r + \mu^{-r} - \mu_0^r - \mu_0^{-r}}{(q^r - q^{-r})(q^r \mu_0^r - q^{-r} \mu_0^{-r})}, \tag{4.2}$$

set

$$X^{\pm}(z) = \exp\left(\mp (q - q^{-1}) \sum_{r>0} x_{\mp r} z^{\pm r} \right),$$

$$x_r = (q\mu_0^2)^r \frac{(\mu/\mu_0)^r - 1}{q^{2r} - 1} \overline{a}_{n+1,r} \otimes 1 + \alpha_r 1 \otimes b_r . \tag{4.3}$$

Then the U module structure on W_m *is extended to a* $\mathcal{U}_\mu$ *module structure by*

$$E_0(z) = a X^+(z) (\tilde{E}_0^{(+)}(z) \otimes 1) X^{-1}(qz)^{-1} ((\mu_0/\mu)^{\partial_{\epsilon_{n+1}} + m/(n+1)} \otimes 1),$$

$$F_0(z) = a^{-1} (((\mu/\mu_0)^{\partial_{\epsilon_{n+1}} + m/(n+1)} \otimes 1) X^+(qz)^{-1} (\tilde{F}_0^{(+)}(z) \otimes 1) X^-(z),$$

$$\phi_0^{\pm}(z) = \frac{X^{\pm}(q^{-1}z)}{X^{\pm}(qz)} (\tilde{\phi}_0^{\pm,(+)}(z) \otimes 1). \tag{4.4}$$

We have chosen the above normalization of $E_0(z)$ and $F_0(z)$ for a later convenience. Note that in the case $\mu = \mu_0 q^2$ the above result coincides with Theorem 1.

B. Level 0 toroidal modules

Set $V=F^{n+1}$ and let $v_1,\ldots,v_{n+1}$ be its standard basis. $V[x,x^{-1}]$ is a U module by the following:

$$E_i(q^iz)=\delta(z/x)E_{ii+1},\quad F_i(q^iz)=\delta(z/x)E_{i+1i},$$

$$\phi_i^+(q^iz)=\sum_{k\neq i,i+1}E_{kk}+q^{-1}\frac{1-q^2z/x}{1-z/x}E_{ii}+q\frac{1-q^{-2}z/x}{1-z/x}E_{i+1i+1},$$

$$\phi_i^-(q^iz)=\sum_{k\neq i,i+1}E_{kk}+q\frac{1-q^{-2}x/z}{1-x/z}E_{ii}+q^{-1}\frac{1-q^2x/z}{1-x/z}E_{i+1i+1},\quad (1\leqslant i\leqslant n),$$

$$C^{\pm 1}=1,\quad D^{\pm 1}=q^{\pm\vartheta}. \tag{4.5}$$

Here E_{ij}'s are matrix units, $\delta(z)=\Sigma_{m\in\mathbf{Z}}z^m$, x acts by multiplication, and a^{ϑ} $(a\in F^{\times})$ is a linear operator on $V[x,x^{-1}]$ defined by $a^{\vartheta}v_{\epsilon}x^m=a^mv_{\epsilon}x^m$. We shall denote this representation by (π,V_x).

Set $p=\mu_0\mu$. The U module structure on V_x is extended to a $\mathcal{U}_{\mu}$ structure[2] by

$$E_0(z)=ap^{\vartheta}\delta(z/x)E_{n+11},\quad F_0(z)=a^{-1}\delta(z/x)E_{1n+1}p^{-\vartheta},$$

$$\phi_0^+(z)=\sum_{k\neq 1,n+1}E_{kk}+q^{-1}\frac{1-q^2z/px}{1-z/px}E_{n+1n+1}+q\frac{1-q^{-2}z/x}{1-z/x}E_{11},$$

$$\phi_0^-(z)=\sum_{k\neq 1,n+1}E_{kk}+q\frac{1-q^{-2}px/z}{1-px/z}E_{n+1n+1}+q^{-1}\frac{1-q^2x/z}{1-x/z}E_{11}, \tag{4.6}$$

where $a\in F^{\times}$. We shall denote this representation by (π_a,V_a).

C. Bialgebra structure of $\mathcal{U}_{\mu}$

In this subsection, fixing $\mu\in F^{\times}$, we omit the subscript μ of $\mathcal{U}_{\mu}$.

1. Completion $\mathcal{U}_{\mu}^{\hat{\otimes}N}$ and bialgebra structure of $\mathcal{U}_{\mu}$

Set $c_i=i(n+1-i)$ $(1\leqslant i\leqslant n)$ and put $K=D^{2(n+1)}\Pi_{1\leqslant i\leqslant n}k_i^{c_i}$. Note that $(\Sigma c_i\alpha_i|\alpha_j)=2$ for $1\leqslant j\leqslant n$. For $N\geqslant 1$ and $r\in\mathbf{Z}$, set

$$\mathcal{U}^{\otimes N}{}_r=\{y\in\mathcal{U}^{\otimes N}|K^{\otimes N}y(K^{\otimes N})^{-1}=q^ry\}. \tag{4.7}$$

We assume the following:

Assumption 1: (1) U is identified with the subalgebra of $\mathcal{U}$ generated by $E_{i,m}$, $F_{i,m}$, $h_{i,r}$, $k_1^{\pm 1}$ $(1\leqslant i\leqslant n)$, $C^{\pm 1}$ and $D^{\pm 1}$ *by the algebra homomorphism from U* to $\mathcal{U}$ *determined by*

$$E_i(z)\mapsto E_i(z),\quad F_i(z)\mapsto F_i(z),\quad \phi_i^{\pm}(z)\mapsto\phi_i^{\pm}(z)\quad C\mapsto C,\quad D\mapsto D. \tag{4.8}$$

(2) $\mathcal{U}$ has subalgebras $\mathcal{U}^-$, $\mathcal{U}^0$, and *$\mathcal{U}^+$ such that the multiplication map $\mathcal{U}^-\otimes\mathcal{U}^0\otimes\mathcal{U}^+\to\mathcal{U}$ is an isomorphism of vector spaces and*

$$U^{\pm}\subset\mathcal{U}^{\pm}=\oplus_{r\geqslant 0}\mathcal{U}^{\pm}\cap\mathcal{U}_{\pm r},\quad U^0\subset\mathcal{U}^0\subset\mathcal{U}_0. \tag{4.9}$$

Set

$$\mathcal{U}^{\otimes N}{}_{0l}=\sum_{s\geqslant l}\mathcal{U}^{\otimes N}\mathcal{U}^{\otimes N}{}_s,\quad (l\in\mathbf{Z}_{\geqslant 0}). \tag{4.10}$$

We introduce a linear topology on $\mathcal{U}^{\otimes N}$ by letting $(\mathcal{U}^{\otimes N}{}_{0l})_{l\geqslant 0}$ be a fundamental system of neighborhoods of the origin. [Hereafter we shall simply say ''introduce a linear topology by $(\mathcal{U}^{\otimes N}{}_{0l})$.''] The $\mathcal{U}^{\otimes N}$ is a separated topological algebra. We denote its completion by $\hat{\mathcal{U}}$ $(N=1)$ and $\mathcal{U}^{\hat{\otimes} N}$ $(N\geqslant 2)$. Note that for $N\geqslant 2$, the topology on $\mathcal{U}^{\otimes N}$ is equivalent to the tensor product topology.

Let $\overline{\Delta}_{\mathrm{I}}$ and $\overline{\Delta}_{\mathrm{II}}$ be the continuous algebra homomorphisms from $\hat{\mathcal{U}}$ to $\mathcal{U}\hat{\otimes}\mathcal{U}$ determined by

$$\overline{\Delta}_{\mathrm{I}}(C)=\overline{\Delta}_{\mathrm{II}}(C)=C\otimes C,\quad \overline{\Delta}_{\mathrm{I}}(D)=\overline{\Delta}_{\mathrm{II}}(D)=D\otimes D,$$

$$\overline{\Delta}_{\mathrm{I}}(\phi_i^+(z))=\phi_i^+(z/C_2)\otimes\phi_i^+(z),\quad \overline{\Delta}_{\mathrm{I}}(\phi_i^-(z))=\phi_i^-(z)\otimes\phi_i^-(z/C_1),$$

$$\overline{\Delta}_{\mathrm{I}}(E_i(z))=E_i(z)\otimes 1+\phi_i^-(z)\otimes E_i(z/C_1),\quad \overline{\Delta}_{\mathrm{I}}(F_i(z))=1\otimes F_i(z)+F_i(z/C_2)\otimes\phi_i^+(z),$$

$$\overline{\Delta}_{\mathrm{II}}(\phi_i^+(z))=\phi_i^+(z/C_2)\otimes\phi_i^+(z),\quad \overline{\Delta}_{\mathrm{II}}(\phi_i^-(z))=\phi_i^-(z)\otimes\phi_i^-(z/C_1),$$

$$\overline{\Delta}_{\mathrm{II}}(E_i(z))=E_i(z)\otimes 1+\phi_i^+(Cz)\otimes E_i(C_1z),\quad \overline{\Delta}_{\mathrm{II}}(F_i(z))=1\otimes F_i(z)+F_i(C_2z)\otimes\phi_i^-(Cz), \tag{4.11}$$

where $C_1=C\otimes 1$ and $C_2=1\otimes C$. Let further $\epsilon:\hat{\mathcal{U}}\to F$ denote the continuous algebra homomorphism determined by

$$\epsilon(E_i(z))=\epsilon(F_i(z))=0,\quad \epsilon(\phi_i^{\pm}(z))=\epsilon(C)=\epsilon(D)=1. \tag{4.12}$$

Here F is given a discrete topology. Then $(\hat{\mathcal{U}},\overline{\Delta}_i,\epsilon)$ $(i=\mathrm{I,II})$ are bialgebras. These are straightforward generalizations of the bialgebra structures found by Drinfeld for $\hat{U}$. Let $\mathcal{R}=\mathcal{R}^+\mathcal{R}^0\mathcal{R}^-$ be the Gauss decomposition of the universal R matrix of U with Δ_0 as a comultiplication.[19] Letting

$$\mathcal{F}_{\mathrm{I}}=\mathcal{R}^-,\quad \mathcal{F}_{\mathrm{II}}=\sigma(\mathcal{R}^+)^{-1},\quad (\sigma(a\otimes b)=b\otimes a), \tag{4.13}$$

set

$$\Delta_i(\cdot)=\mathcal{F}_i^{-1}\overline{\Delta}_i(\cdot)\mathcal{F}_i,\quad (i=\mathrm{I,II}). \tag{4.14}$$

It is known[19] that

$$(\epsilon\hat{\otimes}1)\mathcal{F}_i=(1\hat{\otimes}\epsilon)\mathcal{F}_i=1,\quad (\mathcal{F}_i)_{12}(\Delta_0\hat{\otimes}1)\mathcal{F}_i=(\mathcal{F}_i)_{23}(1\hat{\otimes}\Delta_0)\mathcal{F}_i, \tag{4.15}$$

$$\Delta_i(y)=\Delta_0(y),\quad (i=\mathrm{I,II},\ \ y\in U), \tag{4.16}$$

and that $(\hat{\mathcal{U}},\Delta_i,\epsilon)$ $(i=\mathrm{I,II})$ are bialgebras. Set $\Delta_{\mathrm{I}}^{\mathrm{op}}=\sigma\circ\Delta_{\mathrm{I}}$ and $\mathcal{F}_{\mathrm{I}}^{\mathrm{op}}=\sigma(\mathcal{F}_{\mathrm{I}})$. For $(\Delta,\mathcal{F})=(\Delta_{\mathrm{I}}^{\mathrm{op}},\mathcal{F}_{\mathrm{I}}^{\mathrm{op}})$, $(\Delta_{\mathrm{II}},\mathcal{F}_{\mathrm{II}})$, etc., we define $\Delta^{(N)}:\hat{\mathcal{U}}\to\mathcal{U}^{\hat{\otimes}N+1}$ $(N\geqslant 1)$ by

$$\Delta^{(1)}=\Delta,\quad \Delta^{(N)}=(\Delta\hat{\otimes}1^{\hat{\otimes}N-1})\circ\Delta^{(N-1)}\quad (N\geqslant 2) \tag{4.17}$$

and $\mathcal{F}^{(N)}\in\mathcal{U}^{\hat{\otimes}N}$ $(N\geqslant 2)$ by

$$\mathcal{F}^{(2)}=\mathcal{F},\quad \mathcal{F}^{(N)}=(\mathcal{F}^{(2)})_{12}(\Delta\hat{\otimes}1^{\hat{\otimes}N-2})\mathcal{F}^{(N-1)},\quad (N\geqslant 3). \tag{4.18}$$

Then the following holds:

$$\Delta^{(N-1)}(y)=\mathcal{F}^{(N)-1}\overline{\Delta}^{(N-1)}(y)\mathcal{F}^{(N)},\quad (y\in\mathcal{U}). \tag{4.19}$$

Later we need the following property of $\mathcal{F}_{\rm I}^{\rm op}$ and $\mathcal{F}_{\rm II}$.[19] Set $U^{\geqslant 0}=U^0U^+$ and $U^{\leqslant 0}=U^-U^0$. Set further $\overline{Q}=\oplus_{i=1}^n\mathbf{Z}\alpha_i$, $\overline{Q}_+=-\overline{Q}_-=\oplus_{i=1}^n\mathbf{Z}_{\geqslant 0}\alpha_i$, and $\delta=\Sigma_{i=0}^n\alpha_i$. For $a=\geqslant 0, \leqslant 0$, $m\in\mathbf{Z}$ and $\lambda\in\overline{Q}$, put

$$U^a_{m\delta+\lambda}=\{y\in U^a|DyD^{-1}=q^my,\quad k_iyk_i^{-1}=q^{(\alpha_i|\lambda)}y,\quad (1\leqslant i\leqslant n)\}.$$

Let $M_\pm$ be the closure of $\Sigma_{m\geqslant 0}\Sigma_{\lambda\in\overline{Q}_\pm\backslash\{0\}}U^{\leqslant 0}_{-(m\delta+\lambda)}\otimes U^{\geqslant 0}_{m\delta+\lambda}$. Then

$$\mathcal{F}_{\rm I}^{\rm op}-1\in M_-,\quad \mathcal{F}_{\rm II}-1\in M_+. \tag{4.20}$$

2. Completion $\widehat{\widetilde{\mathcal{U}_\mu^{\otimes N}}}$ and its representation $\hat{\mathcal{W}}$

For $N\geqslant 2$ and $m\in\mathbf{Z}$, set

$$\mathcal{U}^{\otimes N}(m)=\sum_{\Sigma_{i=2}^N(i-1)m_i\geqslant m}\mathcal{U}\otimes\mathcal{U}_{m_2}\otimes\cdots\otimes\mathcal{U}_{m_N}. \tag{4.21}$$

Let $\widetilde{\mathcal{U}^{\otimes N}}$ $(N\geqslant 2)$ be the F algebra $\mathcal{U}^{\otimes N}$ on which a linear topology is introduced by $(\mathcal{U}^{\otimes N}(m))$. Then $\widetilde{\mathcal{U}^{\otimes N}}$ is a separated topological algebra. We denote its completion by $\widehat{\widetilde{\mathcal{U}^{\otimes N}}}$.

Let $\mathcal{W}_i$ $(1\leqslant i\leqslant N)$ be $\mathcal{U}$ modules such that

$$\mathcal{W}_i=\oplus_{m\in\mathbf{Z}}\mathcal{W}_{i,m},\quad \mathcal{W}_{i,m}=\{w\in\mathcal{W}_i|Kw=q^mw\}, \tag{4.22}$$

and set $\mathcal{W}=\mathcal{W}_1\otimes\cdots\otimes\mathcal{W}_N$. For $m\in\mathbf{Z}$ setting

$$\mathcal{W}(m)=\sum_{\Sigma_{i=2}^N(i-1)m_i\geqslant m}\mathcal{W}_1\otimes\mathcal{W}_{2,m_2}\otimes\cdots\otimes\mathcal{W}_{N,m_N}, \tag{4.23}$$

introduce a separated linear topology on $\mathcal{W}$ by $(\mathcal{W}(m))$. Then $\hat{\mathcal{W}}$, the completion of $\mathcal{W}$, is a topological $\widehat{\widetilde{\mathcal{U}^{\otimes N}}}$ module.

Remark 1: $\hat{\mathcal{W}}=\mathcal{W}$ in the case $N=2$ and $\mathcal{W}_{N,m}=\{0\}$ $(m\geqslant 0)$.

3. U_μ module structure on $\hat{\mathcal{W}}$

Lemma 2: (1) Let τ be the identity map from $\widetilde{\mathcal{U}^{\otimes N}}$ to $\mathcal{U}^{\otimes N}$. Let further $\hat{\tau}:\widehat{\widetilde{\mathcal{U}^{\otimes N}}}\to\mathcal{U}^{\hat{\otimes}N}$ be the continuous extension of the continuous algebra homomorphism τ. Then $\hat{\tau}$ is injective.

(2) If we identify $\widehat{\widetilde{\mathcal{U}^{\otimes N}}}$ with a subalgebra of $\mathcal{U}^{\hat{\otimes}N}$ via the map $\hat{\tau}$, then the following holds:

$$\Delta_{\rm I}^{{\rm op}(N-1)}(\mathcal{U})\subset\widehat{\widetilde{\mathcal{U}^{\otimes N}}},\ \ \Delta_{\rm II}^{(N-1)}(\mathcal{U})\subset\widehat{\widetilde{\mathcal{U}^{\otimes N}}}. \tag{4.24}$$

Proof: (1) For $l\in\mathbf{Z}_{\geqslant 0}$, set

$$M_l=\sum_{\Sigma l_i=l}\mathcal{U}^{-\otimes N}\otimes\mathcal{U}^{0\otimes N}\otimes\mathcal{U}^+_{l_1}\otimes\cdots\otimes\mathcal{U}^+_{l_N}. \tag{4.25}$$

Let $m\geqslant 0$. Under the identification of $\mathcal{U}^{\otimes N}\simeq(\mathcal{U}^-\otimes\mathcal{U}^0\otimes\mathcal{U}^+)^{\otimes N}$ with $\mathcal{U}^{-\otimes N}\otimes\mathcal{U}^{0\otimes N}\otimes\mathcal{U}^{+\otimes N}$, the following holds:

$$\mathcal{U}^{\otimes N}(m)=\oplus_l(\mathcal{U}^{\otimes N}(m)\cap M_l),\quad \mathcal{U}^{\otimes N}_{0m}=\oplus_{l\geqslant m}M_l. \tag{4.26}$$

From this, we obtain

$$\cap_l(\mathcal{U}^{\otimes N}(m)+\mathcal{U}_{0l}^{\otimes N})=\mathcal{U}^{\otimes N}(m). \tag{4.27}$$

Utilizing the last equality, it is easy to show that any Cauchy sequence in $\widetilde{\mathcal{U}^{\otimes N}}$ whose image under τ converges to 0 in $\mathcal{U}^{\hat{\otimes} N}$ converges to 0 in $\widetilde{\widetilde{\mathcal{U}^{\otimes N}}}$.

(2) The claim can be easily shown for $\overline{\Delta}_{\rm I}^{\rm op}$ and $\overline{\Delta}_{\rm II}$ instead of $\Delta_{\rm I}^{\rm op}$ and $\Delta_{\rm II}$. Utilizing (4.18) and (4.20), we can check $\mathcal{F}_{\rm I}^{{\rm op}(N)}$, $\mathcal{F}_{\rm II}^{(N)}\in\widetilde{\widetilde{\mathcal{U}^{\otimes N}}}$. Therefore thanks to (4.19) we obtain the claim. ■

The above lemma and (4.16) give the following.

Proposition 2: The completion $\hat{\mathcal{W}}$ defined in Sec. IV C 2 is endowed with $\mathcal{U}$ module structures via the comultiplications $\Delta_1^{\rm op}$ and $\Delta_{\rm II}$. Their U module structures are the ones via the comultiplications $\Delta_0^{\rm op}$ and Δ_0, respectively.

Remark 2: For $(\Delta_{\rm I}^{\rm op},\mathcal{F}_{\rm I}^{\rm op})$, we can replace K by D in (4.7) and (4.22).

D. Intertwiners of toroidal modules

Let $\Phi^{(j)}$ and $\Psi^{(j)}$ $(0\leqslant j\leqslant n)$ denote the intertwiners of U modules determined by

$$\Phi^{(j)}:V(\Lambda_j)\otimes V_x\to V(\Lambda_{j+1}),\quad \Psi^{(j)}:V_x\otimes V(\Lambda_j)\to V(\Lambda_{j+1}),$$
$$\Phi^{(j)}(v_{\Lambda_j}\otimes v_{j+1})=v_{\Lambda_{j+1}},\quad \Psi^{(j)}(v_{j+1}\otimes v_{\Lambda_j})=v_{\Lambda_{j+1}}, \tag{4.28}$$

where $V(\Lambda_j)\otimes V_x$ and $V_x\otimes V(\Lambda_j)$ are U modules via the comultiplication Δ_0, and Λ_{n+1} should be understood as Λ_0. As usual we define their components by

$$\Phi_{k,m}^{(j)}u=\Phi^{(j)}(u\otimes v_kx^m),\quad \Psi_{k,m}^{(j)}u=\Psi^{(j)}(v_kx^m\otimes u), \tag{4.29}$$

and set

$$\Phi_k^{(j)}(z)=\sum_{m\in\mathbf{Z}}\Phi_{k,m}^{(j)}z^{-m},\quad \Psi_k^{(j)}(z)=\sum_{m\in\mathbf{Z}}\Psi_{k,m}^{(j)}z^{-m}. \tag{4.30}$$

Proposition 3: For $m\in\mathbf{Z}$, set $a_m=q^{2m}$. In the following, give $V(\Lambda_j)\otimes V_a$ and $W_m\otimes V_a$ (resp. $V_a\otimes V(\Lambda_j)$ and $V_a\otimes W_m$) toroidal module structures via the comultiplication $\Delta_{\rm I}$ (resp. $\Delta_{\rm II}$) (see Remark 1).

(1) For $0\leqslant j\leqslant n$, $\Phi^{(j)}:V(\Lambda_j)\otimes V_{a_j}\to V(\Lambda_{j+1})$ and $\Psi^{(j)}:V_{a_j}\otimes V(\Lambda_j)\to V(\Lambda_{j+1})$ are intertwiners of $\mathcal{U}_{\mu_0}$ modules.

(2) Let $p=\mu_0\mu\neq1$ and α_r's be the ones defined in Proposition 1. Setting

$$Y^{\pm}(z)=\exp\left(\mp(q-q^{-1})\sum_{r>0}\frac{p^{(r\pm r)/2}}{1-p^r}\alpha_{\mp r}b_{\mp r}z^{\pm r}\right), \tag{4.31}$$

put

$$\Xi(z)=Y^+(qz)Y^-(z),\quad \Sigma(z)=Y^+(z)Y^-(qz). \tag{4.32}$$

For $m=(n+1)s+j$ $(s\in\mathbf{Z},\ 0\leqslant j\leqslant n)$, define $\tilde{\Phi}^{(m)}:W_m\otimes V_{a_m/p^s}\to W_{m+1}$ and $\tilde{\Psi}^{(m)}:V_{a_m/p^s}\otimes W_m\to W_{m+1}$ by the generating series

$$\tilde{\Phi}_k^{(m)}(z)=\Phi_k^{(j)}(z)\otimes\Xi(z)\ \ and\ \ \tilde{\Psi}_k^{(m)}(z)=\Psi_k^{(j)}(z)\otimes\Sigma(z), \tag{4.33}$$

respectively. Here $\tilde{\Phi}_k^{(m)}(z)$ and $\tilde{\Psi}_k^{(m)}(z)$ are defined similarly to $\Phi_k^{(j)}(z)$ and $\Psi_k^{(j)}(z)$. Then $\tilde{\Phi}^{(m)}$ and $\tilde{\Psi}^{(m)}$ are intertwiners of $\mathcal{U}_\mu$ modules.

Proof: We consider the case $\tilde{\Phi}^{(m)}$. In this proof, we give $W_m\otimes V_a$ (resp. $V(\Lambda_j)\otimes V_x$) a $\mathcal{U}_\mu$ (resp. U) module structure via the comultiplication $\overline{\Delta}_{\rm I}$. Set $\check{\Phi}^{(m)}=\tilde{\Phi}^{(m)}\mathcal{F}_I^{-1}$ and $\overline{\Phi}^{(j)}=\Phi^{(j)}\mathcal{F}_I^{-1}$

so that $\check{\Phi}_k^{(m)}(z)=\overline{\Phi}_k^{(j)}(z)\otimes\Xi(z)$. Then $\overline{\Phi}^{(j)}:V(\Lambda_j)\otimes V_x\to V(\Lambda_{j+1})$ is an intertwiner of U modules and its explicit expression in terms of bosons is known.[20] Utilizing the expression, it is straightforward to show $\check{\Phi}^{(m)}:W_m\otimes V_{a_m/p^s}\to W_{m+1}$ is an intertwiner of $\mathcal{U}_\mu$ modules. The claim follows from this. ■

V. LEVEL 0 $U_q'(\widehat{sl_{n+1}})$ action on level 1 $U_q(\widehat{gl_{n+1}})$ and $U_q(\widehat{sl_{n+1}})$ modules

A. Completions of $V_x^{\otimes N}$

Following Ref. 10, we introduce two completions of $V_x^{\otimes N}$. For $m\in\mathbf{Z}$, set

$$\mathcal{V}_N[m]=\text{span}\left\{ v_{\epsilon_1}x^{m_1}\otimes\cdots\otimes v_{\epsilon_N}x^{m_N}\,\middle|\,\sum_{i=2}^{N}(i-1)m_i\geqslant m\right\}, \tag{5.1}$$

$$\mathcal{V}_N[[m]]=\text{span}\{v_{\epsilon_1}x^{m_1}\otimes\cdots\otimes v_{\epsilon_N}x^{m_N}|\max\{m_1,\ldots,m_N\}\geqslant m\}. \tag{5.2}$$

Introduce two linear topologies on the vector space $V_x^{\otimes N}$ by $(\mathcal{V}_N[m])$ and $(\mathcal{V}_N[[m]])$. We denote the thus obtained separated topological vector spaces by $\mathcal{V}_N'$ and $\mathcal{V}_N$, respectively, and let $\hat{\mathcal{V}}_N'$ and $\hat{\mathcal{V}}_N$ signify their completions. Let ι_N be the identity map from $\mathcal{V}_N'$ to $\mathcal{V}_N$. Let further $\hat{\iota}_N:\hat{\mathcal{V}}_N'\to\hat{\mathcal{V}}_N$ denote the continuous extension of the continuous linear map ι_N. (This map is shown to be injective as in Lemma 2.) Note that our completions are a little bit different from those in Ref. 10.

Lemma 3: The topology on $\mathcal{V}_N'$ is equivalent to the one by a family of subspaces $(\mathcal{V}_x^{\otimes N}(m))$ *defined in Sec. IV C 2.*

Proof: The claim is easily checked using the equality $Kv_\epsilon x^m=q^{2(n+1)m+n-2\epsilon+2}v_\epsilon x^m$. ■

Set

$$v_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N)=\sum_{m_i\in\mathbf{Z}}v_{\epsilon_1}x^{m_1}\otimes\cdots\otimes v_{\epsilon_N}x^{m_N}z_1^{-m_1}\cdots z_N^{-m_N}. \tag{5.3}$$

Let $\mathcal{N}_N'$ (resp. $\mathcal{N}_N$) be the closure in $\hat{\mathcal{V}}_N'$ (resp. $\hat{\mathcal{V}}_N$) of the span of the coefficients of the following generating series:

$$\begin{aligned}&v_{\ldots,\epsilon_i,\epsilon_{i+1},\ldots}(\ldots,z_i,z_{i+1},\ldots)-(1-q^2)\frac{(z_i/z_{i+1})^{\theta(\epsilon_i<\epsilon_i+1)}}{1-q^2z_i/z_{i+1}}v_{\ldots,\epsilon_i,\epsilon_{i+1},\ldots}(\ldots,z_{i+1},z_i,\ldots)\\&\quad+q\frac{1-z_i/z_{i+1}}{1-q^2z_i/z_{i+1}}v_{\ldots,\epsilon_{i+1},\epsilon_i,\ldots}(\ldots,z_{i+1},z_i,\ldots),\quad(\epsilon_i\neq\epsilon_{i+1}),\end{aligned}$$

$$v_{\ldots,\epsilon_i,\epsilon_{i+1},\ldots}(\ldots,z_i,z_{i+1},\ldots)+q^2\frac{1-q^{-2}z_i/z_{i+1}}{1-q^2z_i/z_{i+1}}v_{\ldots,\epsilon_{i+1},\epsilon_i,\ldots}(\ldots,z_{i+1},z_i,\ldots),\quad(\epsilon_i=\epsilon_{i+1}), \tag{5.4}$$

where $1 \leqslant i \leqslant N-1$. Then $\mathcal{N}_N = \overline{\hat{\iota}_N(\mathcal{N}'_N)}$. Note that in terms of the R matrix $R(z) \in \mathrm{End}(V \otimes V) \times [[z]]$ defined by

$$R(z) v_{\epsilon_1} \otimes v_{\epsilon_2} = \sum_{\epsilon'_1, \epsilon'_2} R(z)^{\epsilon'_1, \epsilon'_2}_{\epsilon_1, \epsilon_2} v_{\epsilon'_1} \otimes v_{\epsilon'_2},$$

$$= (1-q^2) \frac{z^{\theta(\epsilon_1 < \epsilon_2)}}{1-q^2 z} v_{\epsilon_1} \otimes v_{\epsilon_2} - q \frac{1-z}{1-q^2 z} v_{\epsilon_2} \otimes v_{\epsilon_1}, \quad (\epsilon_1 \neq \epsilon_2),$$

$$= -q^2 \frac{1-q^{-2} z}{1-q^2 z} v_{\epsilon_1} \otimes v_{\epsilon_2}, \quad (\epsilon_1 = \epsilon_2), \tag{5.5}$$

the generating series (5.4) are written as follows:

$$v_{\ldots, \epsilon_i, \epsilon_{i+1}, \ldots}(\ldots, z_i, z_{i+1}, \ldots) - \sum_{\epsilon'_i, \epsilon'_{i+1}} R(z_i / z_{i+1})^{\epsilon'_{i+1}, \epsilon'_i}_{\epsilon_i, \epsilon_{i+1}} v_{\ldots, \epsilon'_{i+1}, \epsilon'_i, \ldots}(\ldots, z_{i+1}, z_i, \ldots). \tag{5.6}$$

B. Level 0 $U'_q(\widehat{sl_{n+1}})$ actions on $\hat{\mathcal{V}}'_{\mathcal{N}} / \mathcal{N}'_N$ and $\hat{\mathcal{V}}_N / \mathcal{N}_N$

Let $U'_q(\widehat{sl_{n+1}})$ be the subalgebra of $U_q(\widehat{sl_{n+1}})$ generated by e_i, f_i, and $k_i^{\pm 1}$ $(0 \leqslant i \leqslant n)$. Set $U'_{c=0} = U'_q(\widehat{sl_{n+1}}) / \langle k_0 \cdots k_n - 1 \rangle$ and give this algebra a discrete topology.

Proposition 4: Let $S_N : \hat{\mathcal{V}}'_N \to \hat{\mathcal{V}}'_N$ be the homeomorphic linear map defined by

$$v_{\epsilon_1, \ldots, \epsilon_N}(z_1, \ldots, z_N) \mapsto v_{\epsilon_1, \ldots, \epsilon_N}(z_1, \ldots, z_N) \Big/ \prod_{i<j} \eta(z_j / z_i), \tag{5.7}$$

where

$$\eta(z) = \frac{(q^2 p z; p)_\infty}{(p z; p)_\infty}, \quad (z;p)_\infty = \prod_{m=0}^{\infty} (1 - p^m z). \tag{5.8}$$

Let further $\sigma_N : \mathcal{U}_\mu \to \mathrm{End}\,(\hat{\mathcal{V}}'_N)$ be the map defined by

$$\sigma_N(y) = S_N^{-1} \circ (\pi_{a_{N-1}} \otimes \cdots \otimes \pi_{a_0}) \Delta_{\mathrm{I}}^{\mathrm{op}(N-1)}(y) \circ S_N, \quad y \in \mathcal{U}_\mu. \tag{5.9}$$

Then the following holds:

(1) $\hat{\mathcal{V}}'_N$ is a $\mathcal{U}_\mu$ module via the map σ_N.

(2) $\mathcal{N}'_N$ is $\mathcal{U}_\mu$ invariant and the $\mathcal{U}_\mu$ action on $\hat{\mathcal{V}}'_N$ induces a $\mathcal{U}_\mu$ action on $\hat{\mathcal{V}}'_N / \mathcal{N}'_N$.

(3) The $\mathcal{U}_\mu$ modules $\hat{\mathcal{V}}'_N$ and $\hat{\mathcal{V}}'_N / \mathcal{N}'_N$ are considered as $U'_{c=0}$ modules via the algebra homomorphism $\varphi : U'_q(\widehat{sl_{n+1}}) \to \mathcal{U}_\mu$ determined by $e_i \mapsto E_{i,0}$, $f_i \mapsto F_{i,0}$, $k_i \mapsto k_i$.

Proof: (1) Follows from Proposition 2 and Lemma 3. Thanks to Lemma A.3, (2) is proven if we show

$$S_N^{-1} \circ (\pi_{a_{N-1}} \otimes \cdots \otimes \pi_{a_0}) \overline{\Delta}_{\mathrm{I}}^{\mathrm{op}(N-1)}(y) \circ S_N(\mathcal{N}'^{\mathcal{F}}_N) \subset \mathcal{N}'^{\mathcal{F}}_N, \quad y \in \mathcal{U}_\mu. \tag{5.10}$$

(See Lemma A.3 for the definition of $\mathcal{N}'^{\mathcal{F}}_N$.) This is easily checked using (4.5), (4.6), and (4.11). (3) Follows simply from the fact that $k_0 \cdots k_n \in \mathcal{U}_\mu$ acts as 1 on $\hat{\mathcal{V}}'_N$. ■

Proposition 5 (Ref. 10): For $i \neq j$, letting K_{ij} be the operator which interchanges z_i and z_j, set $\xi_{ij} = q^{-1} + (q - q^{-1})[z_j / (z_i - z_j)](K_{ij} - 1) \in \mathrm{End}\, F[z_1^{\pm 1}, \ldots, z_N^{\pm 1}]$. Set further

$$Y_j = \xi_{j j+1} \cdots \xi_{jN} p^{\vartheta_j} \xi_{1j}^{-1} \cdots \xi_{j-1 j}^{-1}, \tag{5.11}$$

where $p=\mu_0\mu$ and p^{ϑ_j} signifies the scale operator p^{ϑ} acting on the variable z_j. Then

$$\sum_{m_i}(\hat{Y}_j v_{\epsilon_1}x^{m_1}\otimes\cdots\otimes v_{\epsilon_N}x^{m_N})z_1^{-m_1}\cdots z_N^{-m_N}=Y_j v_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N), \tag{5.12}$$

define continuous linear operators $\hat{Y}_j$'s on $\hat{\mathcal{V}}_N$ and the following holds.

(1) $\hat{\mathcal{V}}_N$ is given a topological $U'_{c=0}$ module structure by the map

$$e_0\mapsto q^{N-1}\sum_{j=1}^{N}\hat{Y}_j^{-1}1^{\otimes j-1}\otimes E_{n+1,1}\otimes(q^{E_{n+1,n+1}-E_{1,1}})^{\otimes N-j},$$

$$f_0\mapsto q^{-(N-1)}\sum_{j=1}^{N}\hat{Y}_j(q^{E_{1,1}-E_{n+1,n+1}})^{\otimes j-1}\otimes E_{1,n+1}\otimes 1^{\otimes N-j},$$

$$k_0\mapsto(q^{E_{n+1,n+1}-E_{1,1}})^{\otimes N},$$

$$y\mapsto(\pi\otimes\cdots\otimes\pi)\Delta_0^{\mathrm{op}(N-1)}(y),\quad y=e_i,f_i,k_i,\quad(1\leqslant i\leqslant n). \tag{5.13}$$

(2) $\mathcal{N}_N$ is $U'_{c=0}$ invariant and the $U'_{c=0}$ action on $\hat{\mathcal{V}}_N$ induces a $U'_{c=0}$ action on $\hat{\mathcal{V}}_N/\mathcal{N}_N$.

For the two $U'_{c=0}$ modules $\hat{\mathcal{V}}'_N/\mathcal{N}'_N$ and $\hat{\mathcal{V}}_N/\mathcal{N}_N$, we can show the following:

Proposition 6: The map $\tilde{\iota}_N:\hat{\mathcal{V}}'_N/\mathcal{N}'_N\to\hat{\mathcal{V}}_N/\mathcal{N}_N$ induced by $\hat{\iota}_N$ is a homomorphism of $U'_{c=0}$ modules.

Proof: $\mathcal{V}'_N/\mathcal{N}'_N$ is a topological $U'_{c=0}$ module; $\tilde{\iota}_N$ is continuous; $\hat{\mathcal{V}}_N/\mathcal{N}_N$ is separated. Therefore the claim follows from the following two lemmas. ∎

Lemma 4: Let X_N denote the span of the vectors $v_{\epsilon_1}x^{m_1}\otimes\cdots\otimes v_{\epsilon_N}x^{m_N}$ ($\epsilon_1\leqslant\cdots\leqslant\epsilon_N$, $m_i\in\mathbf{Z}$). Then $(X_N+\mathcal{N}'_N)/\mathcal{N}'_N$ is dense in $\hat{\mathcal{V}}'_N/\mathcal{N}'_N$.

Proof: It is sufficient to show $\hat{\mathcal{V}}'_N=\overline{X_N+\mathcal{N}'_N}$. Therefore the claim follows if we show the following:

$$v_{\epsilon_1}x^{m_1}\otimes\cdots\otimes v_{\epsilon_N}x^{m_N}\in\overline{X_N+\mathcal{N}'_N}. \tag{5.14}$$

Let $\epsilon_i<\epsilon_{i+1}$ and Y_N be the the span of the coefficients of $v_{\ldots,\epsilon_i,\epsilon_{i+1},\ldots}(\ldots,z_i,z_{i+1},\ldots)$. Equation (5.4) implies that the coefficients of

$$(1-(z_i/z_{i+1})^m)v_{\ldots,\epsilon_{i+1},\epsilon_i,\ldots}(\ldots,z_{i+1},z_i,\ldots)$$

belong to $Y_N+\mathcal{N}'_N$ for any integer m. Hence the coefficients of $v_{\ldots,\epsilon_{i+1},\epsilon_i\ldots}(\ldots,z_{i+1},z_i,\ldots)$ belong to $\overline{Y_N+\mathcal{N}'_N}$. Using the last argument repeatedly, we can show (5.14). ∎

Lemma 5: For $\epsilon_1\leqslant\cdots\leqslant\epsilon_N$ and $y=e_i$, f_i, $k_i^{\pm1}$ $(0\leqslant i\leqslant n)$, the following holds:

$$\tilde{\iota}_N(y\dot{v}_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N))=y\tilde{\iota}_N(\dot{v}_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N)), \tag{5.15}$$

where $\dot{v}_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N)$ signifies the image of $v_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N)$ under the quotient map $\hat{\mathcal{V}}'_N\to\hat{\mathcal{V}}'_N/\mathcal{N}'_N$.

Proof: The above equality clearly holds except for e_0 and f_0. The case e_0 is shown in the Appendix and the case f_0 is similarly proven. ∎

C. Level 0 $U'_q(\widehat{sl_{n+1}})$ actions on $\hat{\mathcal{V}}_N$ and W_N

Give W_N a discrete topology. Set $|0\rangle=v_{\Lambda_0}\otimes1\in W_0$. Let $\rho_N:\mathcal{V}_N\to W_N$ be the linear map defined by

$$v_{\epsilon_1,\dots,\epsilon_N}(z_1,\dots,z_N)\mapsto\tilde{\Phi}^{(N-1)}_{\epsilon_1}(z_1)\cdots\tilde{\Phi}^{(0)}_{\epsilon_N}(z_N)|0\rangle\prod_{0\le k\le N-1} z_{N-k}^{s_k}\Big/\prod_{i<j}\eta(z_j/z_i),\tag{5.16}$$

where s_k is the integer part of $k/(n+1)$. Then, as in Ref. 10, thanks to the commutation relations among the intertwiners $\tilde{\Phi}^{(m)}$,

$$\begin{aligned}&\tilde{\Phi}^{(m+1)}_{\epsilon_1}(z_1)\tilde{\Phi}^{(m)}_{\epsilon_2}(z_2)/\eta(z_2/z_1)\\&=\sum_{\epsilon_1',\epsilon_2'}(z_1/z_2)^{s_m-s_{m+1}}R(z_1/z_2)^{\epsilon_2',\epsilon_1'}_{\epsilon_1,\epsilon_2}\tilde{\Phi}^{(m+1)}_{\epsilon_2'}(z_2)\tilde{\Phi}^{(m)}_{\epsilon_1'}(z_1)/\eta(z_1/z_2),\end{aligned}\tag{5.17}$$

the map ρ_N is continuous and extends to a continuous linear map $\hat{\rho}_N:\hat{\mathcal{V}}_N\to W_N$. Moreover the latter map induces $\tilde{\rho}_N:\hat{\mathcal{V}}_N/\mathcal{N}_N\to W_N$.

Proposition 7: Let us regard the $\mathcal{U}_\mu$ module W_N as a $U'_{c=0}$ module via the map φ. Then $\tilde{\rho}'_N:=\tilde{\rho}_N\circ\tilde{\iota}_N:\hat{\mathcal{V}}'_N/\mathcal{N}'_N\to W_N$ is a homomorphism of $U'_{c=0}$ modules.

Proof: In this proof, we complete tensor products $V_{b_1}\otimes\cdots\otimes V_{b_k}$ and $V_{b_1}\otimes\cdots\otimes V_{b_k}\otimes W_m$ ($b_i\in F^\times$) as in Sec. IV C 2. We give the completions $\mathcal{U}_\mu$ module structures via the comultiplication $\Delta_1^{\rm op}$ as in Proposition 2 and consider them as $U'_{c=0}$ modules via the map φ. Note that we treat the completion $\hat{\mathcal{V}}'_N$ in the same way (not as in Proposition 4). To prove the proposition, it is sufficient to show that the continuous linear map $\hat{\rho}_N\circ\hat{\iota}_N\circ S_N^{-1}:\hat{\mathcal{V}}'_N\to W_N$ is $U'_{c=0}$ linear. Note that this map is determined by

$$v_{\epsilon_1,\dots,\epsilon_N}(z_1,\dots,z_N)\mapsto\tilde{\Phi}^{(N-1)}_{\epsilon_1}(z_1)\cdots\tilde{\Phi}^{(0)}_{\epsilon_N}(z_N)|0\rangle\prod_{0\le k\le N-1} z_{N-k}^{s_k}.\tag{5.18}$$

Let $\mathcal{F}$ and $\bar{\Delta}$ stand for $\mathcal{F}_1^{\rm op}$ and $\bar{\Delta}_1^{\rm op}$, respectively. From $(1^{\hat{\otimes}N}\hat{\otimes}\epsilon)\mathcal{F}^{(N+1)}=\mathcal{F}^{(N)}$ and (4.20), we get $\mathcal{F}^{(N+1)\pm1}(u\otimes|0\rangle)=(\mathcal{F}^{(N)\pm1}u)\otimes|0\rangle$ for $u\in\hat{\mathcal{V}}'_N$. Moreover we have $\bar{\Delta}^{(N)}(y)(u\otimes|0\rangle)=(\bar{\Delta}^{(N-1)}(y)u)\otimes|0\rangle$ for $y\in U'_{c=0}$ and $u\in\hat{\mathcal{V}}'_N$ since $E_{0,m}|0\rangle=F_{0,m}|0\rangle=0$ $(m\geqslant0)$ and $\phi_0^-(z)|0\rangle=|0\rangle$. Hence the continuous map from the completion of $V_{a_{N-1}}\otimes\cdots\otimes V_{a_0}$ to the completion of $V_{a_{N-1}}\otimes\cdots\otimes V_{a_0}\otimes W_0$ defined by $u\mapsto u\otimes|0\rangle$ is $U'_{c=0}$ linear. Set $\tilde{\Phi}^{{\rm op}(m)}=\tilde{\Phi}^{(m)}\circ(1\otimes x^{s_m})\circ\sigma$, $(\sigma(v\otimes w)=w\otimes v)$. The linear map $1^{\otimes N-m-1}\otimes\tilde{\Phi}^{{\rm op}(m)}:V_{a_{N-1}}\otimes\cdots\otimes V_{a_m}\otimes W_m\to V_{a_{N-1}}\otimes\cdots\otimes V_{a_{m+1}}\otimes W_{m+1}$ extends continuously to a map between the completions and the latter map is shown to be an intertwiner of $U'_{c=0}$ modules. From the above we can show the claim stated in the first part of the proof. ■

From Propositions 6 and 7, we can show the following:

Theorem 2: Regard $\hat{\mathcal{V}}_N$ and W_N as $U'_{c=0}$ modules as in Propositions 5 and 7. Then $\hat{\rho}_N:\hat{\mathcal{V}}_N\to W_N$ is a homomorphism of $U'_{c=0}$ modules.

Proof: Since $\tilde{\rho}_N\circ\tilde{\iota}_N=\tilde{\rho}'_N$ and the maps $\tilde{\iota}_N$ and $\tilde{\rho}'_N$ are $U'_{c=0}$ linear, we get $\tilde{\rho}_N(yw)=y\tilde{\rho}_N(w)$ for $y\in U'_{c=0}$ and $w\in{\rm Im}\,\tilde{\iota}_N$. ${\rm Im}\,\tilde{\iota}_N$ is dense in $\hat{\mathcal{V}}_N/\mathcal{N}_N$; $\hat{\mathcal{V}}_N/\mathcal{N}_N$ and W_N are separated topological $U'_{c=0}$ modules; $\tilde{\rho}_N$ is continuous. Therefore $\tilde{\rho}_N$ is a homomorphism of $U'_{c=0}$ modules. The claim follows from this. ■

Finally we consider the case $\mu=\mu_0$. Let $\hat{\gamma}_N:\hat{\mathcal{V}}_N\to V(\Lambda_j)$ $(j\equiv N\bmod n+1)$ denote the continuous linear map defined by (5.16) with $\tilde{\Phi}^{(m)}$ and $|0\rangle$ replaced by $\Phi^{(i)}(i\equiv m\bmod n+1)$ and v_{Λ_0}, respectively. Let further $\hat{\gamma}:\hat{\mathcal{V}}:=\oplus_{N\geqslant1}\hat{\mathcal{V}}_N\to\mathcal{H}:=\oplus_{j=0}^n V(\Lambda_j)$ signify the linear map obtained from $\hat{\gamma}_N$'s. We consider the $\mathcal{U}_{\mu_0}$ modules $\hat{\mathcal{V}}$ and $\mathcal{H}$ as $U'_{c=0}$ modules via the map φ.

Theorem 3: *Consider the case $\mu=\mu_0$ $(p=\mu_0^2)$ and let $\hat{\mathcal{V}}$, $\mathcal{H}$, and $\hat{\gamma}$ be as above. Then $\hat{\gamma}:\hat{\mathcal{V}}\to\mathcal{H}$ is a surjective homomorphism of $U'_{c=0}$ modules.*

Proof: The surjectivity is shown as in Ref. 10. Thanks to Proposition 3 (1), Proposition 7 holds also for the map $\hat{\mathcal{V}}'_N/\mathcal{N}'_N \to V(\Lambda_j)$ $(j \equiv N \bmod n+1)$ induced by $\hat{\gamma}_N \circ \hat{\iota}_N$. Therefore the remaining claim is proven as in Theorem 2. ■

Remark 3: In Ref. 10, in the case $n=1$, by showing that $Ker\ \hat{\gamma}$ is $U'_{c=0}$ invariant, a $U'_{c=0}$ action was defined on $\mathcal{H}$ so that $\hat{\gamma}$ is a homomorphism of $U'_{c=0}$ modules. Therefore Theorem 3 clarifies the relation between the $U'_{c=0}$ module structure on $\mathcal{H}$ induced by the toroidal action[4] and the one by the approach of Ref. 10.

APPENDIX: PROOF OF LEMMA 5

The following is immediate. See Ref. 10 for (1) and (4).

Lemma A.1: For $I \subset \{1,...,N\}$, let K_I be the set of generating series $w(z_1,\ldots,z_N) = \Sigma w_{m_1,\ldots,m_N} z_1^{-m_1} \cdots z_N^{-m_N}$ such that $w_{m_1,\ldots,m_N} \in \overline{\mathcal{V}_N[[M]]}$, where $M = \max\{m_i\}_{i\in I}$. For $I \subset \{1,...,N\}$ and $k \in \{1,...,N\}$, let $A_{I,k}$ denote the algebra of formal power series in the variables z_i/z_k $(i \in I \backslash \{k\})$.

(1) If f is a formal series of the form $f = \Sigma_{\Sigma m_i = m} f_{m_1,\ldots,m_r} z_{i_1}^{-m_1} \cdots z_{i_r}^{-m_r}$ $(1 \leqslant i_1 < \cdots < i_r \leqslant N$, $f_{m_1,\ldots,m_r} \in F$, $m \in \mathbf{Z})$ and $w \in K_{\{i_1,\ldots,i_r\}}$, then fw is a well defined formal series with $\hat{\mathcal{V}}_N$ as coefficients.

(2) $A_{I,k} K_I \subset K_{I\backslash\{k\}}$. Moreover $A_{I,k}$ acts on K_I when $k \notin I$.

(3) Let i, $j \in I$, $i \neq j$, k, $l \in \{1,...,N\} \backslash I$, $k \neq l$ and $w \in K_I$. Let further $f(z)$ and $g(z)$ be formal power series in z. Then for $(x,y) = (z_i/z_k, z_i/z_l)$ and $(z_i/z_j, z_j/z_k)$, the following holds:

$$f(x)(g(y)w) = g(y)(f(x)w) = (f(x)g(y))w.$$

(4) $\xi_{ij}^{\pm 1} K_I \subset K_I$ if $i,j \in I$.

(5) Let i, $j \in I$, $i \neq j$ and $k \in \{1,...,N\} \backslash I$. If $f \in A_{I,k}$ satisfies $f\xi_{ij} = \xi_{ij} f$ when acted on Laurent polynomials in z_l $(1 \leqslant l \leqslant N)$, then $f\xi_{ij}|_{K_I} = \xi_{ij} f|_{K_I}$.

Lemma A.2: Let L be the set of generating series $w(z_1,\ldots,z_N) = \Sigma w_{m_1,\ldots,m_N} z_1^{-m_N} \cdots z_N^{-m_N}$ such that $w_{m_1,\ldots,m_N} \in \overline{\mathcal{V}_N[M]}$, where $M = \Sigma_{i=2}^N (i-1) m_i$.

(1) Let B be the algebra of formal power series in the variables z_{i+1}/z_i $(1 \leqslant i < N)$. Then B acts on L.

(2) Set $u_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N) = \mathcal{F}_{\mathrm{I}}^{op(N)-1} v_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N)$. Then $u_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N) \in L$ for any ϵ_i, and $u_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N) = v_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N)$ when $\epsilon_1 \leqslant \cdots \leqslant \epsilon_N$.

Proof: (1) is immediate. (2) follows from (4.20). ■

Lemma A.3: Let $\mathcal{N}'^{\mathcal{F}}_N$ be the closure in $\hat{\mathcal{V}}'_N$ of the span of the coefficients of the following generating series:

$$v_{\ldots,\epsilon_i,\epsilon_{i+1},\ldots}(\ldots,z_i,z_{i+1},\ldots) + q\,\frac{1 - z_i/z_{i+1}}{1 - q^2 z_i/z_{i+1}}\, v_{\ldots,\epsilon_{i+1},\epsilon_i,\ldots}(\ldots,z_{i+1},z_i,\ldots), \quad (\epsilon_i < \epsilon_{i+1}),$$

$$v_{\ldots,\epsilon_i,\epsilon_{i+1},\ldots}(\ldots,z_i,z_{i+1},\ldots) + q^2\,\frac{1 - q^{-2} z_i/z_{i+1}}{1 - q^2 z_i/z_{i+1}}\, v_{\ldots,\epsilon_{i+1},\epsilon_i,\ldots}(\ldots,z_{i+1},z_i,\ldots), \quad (\epsilon_i = \epsilon_{i+1}), \tag{A1}$$

where $1 \leqslant i \leqslant N-1$. Then $\mathcal{F}_{\mathrm{I}}^{\mathrm{op}(N)} \mathcal{N}'_N = \mathcal{N}'^{\mathcal{F}}_N$.

Proof: Let $\mathcal{F}^{(N)}$ signify $\mathcal{F}_{\mathrm{I}}^{\mathrm{op}(N)}$. Firstly we consider the case $N=2$. Let $\epsilon_1 < \epsilon_2$. From the definition we get

$$v_{\epsilon_1,\epsilon_2}(z_1,z_2)+q\frac{1-z_2/z_1}{1-q^2z_2/z_1}v_{\epsilon_2,\epsilon_1}(z_1,z_2)+(z_1\leftrightarrow z_2)\in\mathcal{N}_2'^{\mathcal{F}}. \tag{A2}$$

This implies

$$\frac{z_2}{1-z_2/z_1}v_{\epsilon_1,\epsilon_2}(z_1,z_2)+\frac{qz_2}{1-q^2z_2/z_1}v_{\epsilon_2,\epsilon_1}(z_1,z_2)-(z_1\leftrightarrow z_2)\in\mathcal{N}_2'^{\mathcal{F}}. \tag{A3}$$

Let N_2' (resp. $N_2'^{\mathcal{F}}$) be the span of the coefficients of (5.4) (resp. (A1) and (A3)). From

$$\mathcal{F}^{(2)}v_{\epsilon_1,\epsilon_2}(z_1,z_2)=v_{\epsilon_1,\epsilon_2}(z_1,z_2),\quad(\epsilon_1\leqslant\epsilon_2),$$

$$=v_{\epsilon_1,\epsilon_2}(z_1,z_2)-(q-q^{-1})\frac{z_2/z_1}{1-z_2/z_1}v_{\epsilon_2,\epsilon_1}(z_1,z_2),\quad(\epsilon_1>\epsilon_2), \tag{A4}$$

we obtain $\mathcal{F}^{(2)}N_2'=N_2'^{\mathcal{F}}$ and, hence, the claim in the case $N=2$. Next we consider the case $N>2$. Equation (4.15) gives

$$\mathcal{F}^{(N)}=\mathcal{F}^{(2)}_{k,k+1}(1^{\hat{\otimes}k-1}\hat{\otimes}\Delta_{\mathrm{I}}^{\mathrm{op}}\hat{\otimes}1^{\hat{\otimes}N-k-1})\mathcal{F}^{(N-1)}\quad\text{for}\quad 1\leqslant k\leqslant N-1. \tag{A5}$$

It is easy to check that N_2' is U invariant. Hence we obtain

$$\mathcal{F}^{(N)}V_x^{\otimes k-1}\otimes N_2'\otimes V_x^{\otimes N-k-1}\subset\overline{V_x^{\otimes k-1}\otimes\mathcal{F}^{(2)}N_2'\otimes V_x^{\otimes N-k-1}}. \tag{A6}$$

From this, we get $\mathcal{F}^{(N)}\mathcal{N}_N'\subset\mathcal{N}_N'^{\mathcal{F}}$. The reverse inclusion is shown similarly. ∎

Hereafter we shall let $\equiv$ denote the equality in $\hat{\mathcal{V}}_N \bmod \mathcal{N}_N$.

Lemma A.4: The action of e_0 on the $U'_{c=0}$ module $\hat{\mathcal{V}}_N$ and that of $E_{0,0}$ on the $\mathcal{U}_\mu$ module $\hat{\mathcal{V}}'_N$ satisfy the following relations:

(1) $e_0v_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N)$

$$\equiv(-q)^{N-1}\sum_{j=1}^{N}\delta_{\epsilon_j,1}(-1)^{j-1}q^{-\Sigma_{j<i\leqslant N}\delta_{\epsilon_i,1}}\xi_{j-1j}\cdots\xi_{1j}$$
$$\times v_{\epsilon_1,\ldots,\hat{\epsilon}_j,\ldots,\epsilon_N,n+1}(z_1,\ldots,\hat{z}_j,\ldots,z_N,z_j/p). \tag{A7}$$

(2) *For* $\epsilon_1\leqslant\cdots\leqslant\epsilon_N$,

$$\hat{\iota}_N(E_{0,0}v_{\epsilon_1,\ldots,\epsilon_N}(z_1,\ldots,z_N))$$
$$\equiv(-q)^{N-1}\sum_{j=1}^{N}\delta_{\epsilon_j,1}(-q)^{-(j-1)}q^{-\Sigma_{j<i\leqslant N}\delta_{\epsilon_i,1}}\prod_{i=1}^{j-1}\frac{1-q^2z_j/z_i}{1-z_j/z_i}\prod_{i=j+1}^{N}\left(\frac{1-q^2z_i/z_j}{1-z_i/z_j}\right)^{\delta_{\epsilon_i,1}}$$
$$\times v_{\epsilon_1,\ldots,\hat{\epsilon}_j,\ldots,\epsilon_N,n+1}(z_1,\ldots,\hat{z}_j,\ldots,z_N,z_j/p). \tag{A8}$$

Here $\hat{}$ denotes the deletion of variables.

Proof: (1) The definition of ξ_{ij} and (5.4) give

$$\xi_{ii+1}^{-1}v_{\ldots,\epsilon_i,\epsilon_{i+1},\ldots}(\ldots,z_i,z_{i+1},\ldots)\equiv-q^{-\delta_{\epsilon_i,\epsilon_{i+1}}}v_{\ldots,\epsilon_{i+1},\epsilon_i,\ldots}(\ldots,z_{i+1},z_i,\ldots),\quad(\epsilon_i\geqslant\epsilon_{i+1}). \tag{A9}$$

Using this equality, we get the claim.

(2) Set $f(z)=q^{-1}(1-q^2z)/(1-z)$. Set further $g_\epsilon(z)=q(1-z)/(1-q^2z)$ for $1\leqslant\epsilon\leqslant n$ and $=q^2(1-q^{-2}z)/(1-q^2z)$ for $\epsilon=n+1$. Then thanks to (4.19) we get

$$E_{0,0}u_{\epsilon_1,\dots\epsilon_N}(z_1,\dots,z_N)=q^{N-1}\sum_{j=1}^{N}\delta_{\epsilon_j,1}\prod_{i=1}^{j-1}f(z_j/z_i)\prod_{i=j+1}^{N}f(z_i/z_j)^{\delta_{\epsilon_i,1}}\times\prod_{i=j+1}^{N}g_{\epsilon_i}(pz_i/z_j)u_{\epsilon_1,\dots,\epsilon_{j-1},n+1,\epsilon_{j+1},\dots,\epsilon_N}(z_1,\dots,z_j/p,\dots,z_N), \tag{A10}$$

for any ϵ_i. Lemmas A.2 and A.3 imply

$$\hat{\iota}_N\left(\prod_{i=j+1}^{N}g_{\epsilon_i}(pz_i/z_j)u_{\epsilon_1,\dots,\epsilon_{j-1},n+1,\epsilon_{j+1},\dots,\epsilon_N}(z_1,\dots,z_j/p,\dots z_N)\right)\equiv(-1)^{N-j}\hat{\iota}_N(u_{\epsilon_1,\dots,\hat{\epsilon}_j,\dots,\epsilon_N,n+1}(z_1,\dots,\hat{z}_j,\dots,z_N,z_j/p)). \tag{A11}$$

In the case $\epsilon_1\leqslant\cdots\leqslant\epsilon_N$, thanks to Lemma A.1 (1) and Lemma A.2, multiplying the above equation by $\Pi_{i=1}^{j-1}f(z_j/z_i)\Pi_{i=j+1}^{N}f(z_i/z_j)^{\delta_{\epsilon_i,1}}$ is meaningful. Hence we obtain the claim. ■

Thanks to the above lemma, (5.15) with $y=e_0$ follows from the following lemma with $w(z_1,\dots,z_N)=v_{\underbrace{1,\dots,1}_{s-1},\epsilon_{s+1},\dots,\epsilon_N,n+1}(z_1,\dots,z_N)$ $(1<\epsilon_{s+1}\leqslant\cdots\leqslant\epsilon_N)$.

Lemma A.5: Let $1\leqslant t\leqslant s\leqslant N$ and $w(z_1,\dots,z_N)\in K_{\{t,\dots,N\}}$. If the coefficients of the generating series

$$w(z_1,\dots,z_i,z_{i+1},\dots,z_N)+q^2\frac{1-q^{-2}z_i/z_{i+1}}{1-q^2z_i/z_{i+1}}w(z_1,\dots,z_{i+1},z_i,\dots,z_N),\quad(t\leqslant i\leqslant s-2)$$

belong to $\mathcal{N}_N$, then the following equality holds in $\hat{\mathcal{V}}_N$:

$$\sum_{j=t}^{s}(-q)^{j-t}\xi_{j-1j}\cdots\xi_{tj}w(z_1,\dots,\hat{z}_j,\dots,z_N,z_j/p)\equiv\sum_{j=t}^{s}(-1)^{j-t}\prod_{i=t}^{j-1}\frac{1-q^2z_j/z_i}{1-z_j/z_i}\prod_{i=j+1}^{s}\frac{1-q^2z_i/z_j}{1-z_i/z_j}w(z_1,\dots,\hat{z}_j,\dots,z_N,z_j/p). \tag{A12}$$

Proof: This can be shown by induction on $s-t$. Thanks to Lemma A.1, the left hand side is rewritten as follows:

$$w(z_1,\dots,\hat{z}_t,\dots,z_N,z_t/p)+(1-q^2)\sum_{j=t+1}^{s}(-q)^{j-t-1}\xi_{j-1j}\cdots\xi_{t+1j}\times\frac{z_j/z_t}{1-z_j/z_t}w(z_1,\dots,z_{t-1},z_j,z_{t+1},\dots,\hat{z}_j,\dots,z_N,z_t/p)-\frac{1}{\Pi_{i>t}\eta(z_i/z_t)}\sum_{j=t+1}^{s}(-q)^{j-t-1}\xi_{j-1j}\cdots\xi_{t+1j}\overline{w}(z_1,\dots,\hat{z}_j,\dots,z_N,z_j/p), \tag{A13}$$

where

$$\overline{w}(z_1,\ldots,z_N)=\prod_{i>t}\eta(z_i/z_t)\times w(z_1,\ldots,z_N)\in K_{\{t+1,\ldots,N\}}. \tag{A14}$$

For $t+1\leqslant i<j\leqslant s$, thanks to Lemma A.1, we obtain,

$$\begin{aligned}&\xi_{ij}\frac{z_j/z_t}{1-z_j/z_t}w(z_1,\ldots,\hat{z}_t,\ldots,z_{i-1},z_j,z_i,\ldots,\hat{z}_j,\ldots,z_N,z_t/p)\\ &\quad\equiv -q^{-1}\frac{1-q^2z_i/z_t}{1-z_i/z_t}\,\frac{z_j/z_t}{1-z_j/z_t}w(z_1,\ldots,\hat{z}_t,\ldots,z_i,z_j,z_{i+1},\ldots,\hat{z}_j,\ldots,z_N,z_t/p).\end{aligned} \tag{A15}$$

Since the action of $\xi_{j-1j}\cdots\xi_{i+1j}$ on both sides of the above is well defined, we get the following equality:

$$\begin{aligned}&\xi_{j-1j}\cdots\xi_{ij}\frac{z_j/z_t}{1-z_j/z_t}w(z_1,\ldots,\hat{z}_t,\ldots,z_{i-1},z_j,z_i,\ldots,\hat{z}_j,\ldots,z_N,z_t/p)\\ &\quad\equiv -q^{-1}\frac{1-q^2z_i/z_t}{1-z_i/z_t}\,\xi_{j-1j}\cdots\xi_{i+1j}\frac{z_j/z_t}{1-z_j/z_t}w(z_1,\ldots,\hat{z}_t,\ldots,z_i,z_j,z_{i+1},\ldots,\hat{z}_j,\ldots,z_N,z_t/p).\end{aligned} \tag{A16}$$

Repeating this argument, the sum of the first two terms is found to be

$$\prod_{i=t+1}^{s}\frac{1-q^2z_i/z_t}{1-z_i/z_t}w(z_1,\ldots,\hat{z}_t,\ldots,z_N,z_t/p).$$

Applying the assumption of the induction to the last sum, we obtain the claim. ■

[1] V. Ginzburg, M. Kapranov, and E. Vasserot, ''Langlands reciprocity for algebraic surfaces,'' Math. Res. Lett. **2**, 147–160 (1995).
[2] M. Varagnolo and E. Vasserot, ''Schur duality in the toroidal setting,'' Commun. Math. Phys. **182**, 469–484 (1996).
[3] I. V. Cherednik, ''Double affine Hecke algebras and Macdonald's conjectures,'' Ann. Math. **141**, 191–216 (1995).
[4] Y. Saito, ''Quantum toroidal algebras and their vertex representations,'' Publ. RIMS. Kyoto Univ. **34**, 155–177 (1998).
[5] K. Takemura and D. Uglov, ''Representations of the quantum toroidal algebra on highest weight modules of the quantum affine algebra of type $gl(N)$'' (preprint math/9806134).
[6] V. Chari and A. Pressley, ''Quantum affine algebras,'' Commun. Math. Phys. **142**, 261–283 (1991).
[7] M. Varagnolo and E. Vasserot, ''Double-loop algebras and the Fock space,'' Invent. Math. **133**, 133–159 (1998).
[8] Y. Saito, K. Takemura, and D. Uglov, ''Toroidal actions on level 1 modules of $U_q(\widehat{sl}_n)$,'' Trans. Groups **1**, 75–102 (1998).
[9] M. Kashiwara, T. Miwa, and E. Stern, ''Decomposition of q-deformed Fock spaces,'' Selecta Math., New Series **1**, 787–805 (1995).
[10] M. Jimbo, R. Kedem, H. Konno, T. Miwa, and J.-U. H. Petersen, ''New level-0 action of $U_q(\widehat{sl}_2)$ on level-1 modules,'' in Proceedings of the Statistical Mechanics and Quantum Field Theory, University of Southern California, 16–21 May 1994 (unpublished).
[11] D. Bernard, V. Pasquier, and D. Serban, ''Spinons in conformal field theory,'' Nucl. Phys. B **428**, 612–628 (1994).
[12] J. Ding and S. Khoroshkin, ''Weyl group extensions of quantum current algebras'' (preprint math/9804139).
[13] M. Jimbo, ''A q-analog of $U(gl(N+1))$, Hecke algebra, and the Yang–Baxter equation,'' Lett. Math. Phys. **11**, 247–252 (1986).
[14] G. Lusztig, ''Quantum deformations of certain simple modules over enveloping algebras,'' Adv. Math. **70**, 237–249 (1988).
[15] V. G. Drinfeld, ''New realizations of Yangian and quantum affine algebra,'' Sov. Math. Dokl. **36**, 212–216 (1988).
[16] J. Ding and I. Frenkel, ''Isomorphism of two realizations of quantum affine algebra $U_q(\widehat{gl}_n)$,'' Commun. Math. Phys. **156**, 277–300 (1993).
[17] J. Beck, ''Braid group actions and quantum affine algebras,'' Commun. Math. Phys. **165**, 555–568 (1994).
[18] M. Rosso, ''Finite dimensional representations of the quantum analog of the enveloping algebra of a complex simple Lie

algebra,'' Commun. Math. Phys. **117**, 581–593 (1988).
[19] S. M. Khoroshkin and V. N. Tolstoy, ''Twisting of quantum (super) algebras. Connection of Drinfeld's and Cartan–Weyl realizations for quantum affine algebras'' (preprint hep-th/9404036).
[20] J. Ding and K. Iohara, ''Drinfeld comultiplication and vertex operators,'' J. Geom. Phys. **23**, 1–13 (1997).

Picard–Fuchs ordinary differential systems in $N=2$ supersymmetric Yang–Mills theories

Yűji Ohta
Research Institute for Mathematical Sciences, Kyoto University, Sakyoku, Kyoto 606, Japan

(Received 21 December 1998; accepted for publication 26 January 1999)

In general, Picard–Fuchs systems in $N=2$ supersymmetric Yang–Mills theories are realized as a set of simultaneous partial differential equations. However, if the quantum chromodynamics (QCD) scale parameter is used as a unique independent variable instead of moduli, the resulting Picard–Fuchs systems are represented by a single ordinary differential equation (ODE) whose order coincides with the total number of independent periods. This paper discusses some properties of these Picard–Fuchs ODEs. In contrast with the usual Picard–Fuchs systems written in terms of moduli derivatives, there exists a Wronskian for this ordinary differential system and this Wronskian produces a new relation among periods, moduli, and QCD scale parameter, which in the case of SU(2) is reminiscent of the scaling relation of prepotential. On the other hand, in the case of the SU(3) theory, there are two kinds of ordinary differential equations, one of which is the equation directly constructed from periods and the other is derived from the SU(3) Picard–Fuchs equations in moduli derivatives identified with Appell's F_4 hypergeometric system, i.e., Burchnall's fifth-order ordinary differential equation published in 1942. It is shown that four of the five independent solutions to the latter equation actually correspond to the four periods in the SU(3) gauge theory and the closed form of the remaining one is established by the SU(3) Picard–Fuchs ODE. The formula for this fifth solution is a new one.
[S0022-2488(99)01106-8]

I. INTRODUCTION

It has been recognized that the low energy effective action of $N=2$ supersymmetric Yang–Mills theory for any Lie gauge group including at most two derivatives and four fermions is dominated by a holomorphic function called prepotential $\mathcal{F}$.[1] Perturbatively, this prepotential is a sum of the classical part and one-loop contribution, and further contributions from higher loop diagrams are excluded by the nonrenormalization theorem in $N=2$ theory. However, $\mathcal{F}$ was expected to be affected by instantons and hence its nonperturbative determination was a long-standing problem.

In the case of SU(2) gauge theory, without any quark hypermultiplet, Seiberg and Witten[2] showed that the vacuum configuration of the $N=2$ action was parametrized by the moduli $u=\langle \mathrm{tr}\,\phi^2\rangle$ (ϕ is a complex scalar field in the adjoint representation of the gauge group) and singularities on this parameter space (moduli space) could split into pieces by instanton effect. This instanton corrected moduli space is often called quantum moduli space and Seiberg and Witten[2,3] identified the quantum moduli space with the moduli space of a certain elliptic curve of genus one. According to their ansatz, since the vacuum expectation value $a=\langle\phi\rangle$ and its magnetic dual can also be regarded as periods of a meromorphic one-form on the elliptic Riemann surface, these periods can be calculated as a linear combination of solutions to Picard–Fuchs equations (for a historical review and introduction of the Picard–Fuchs equation in mathematics, see Gray[4]). Once the periods are calculated, it is immediate to obtain the prepotential because of rigid special geometry. In this way, Klemm *et al.*[5] determined the SU(2) prepotential.

Seiberg and Witten's approach to $N=2$ supersymmetric SU(2) Yang–Mills theory[2,3] was extended to other higher rank gauge group cases coupled with or without quark hypermultiplets,[5–16] and it was found that the quantum moduli spaces of those gauge theories could be identified with those of Riemann surfaces with certain genus. In these studies, general algorithms to get Picard–Fuchs equations were developed,[17–21] but these equations are in general realized as a set of simultaneous partial differential equations (PDEs) in terms of moduli derivatives. For this reason, it is not easy to solve Picard–Fuchs equations, especially, in higher rank gauge goup cases.

However, if the quantum chromodynamics (QCD) scale parameter instead of moduli is used as the unique independent variable, the resulting Picard–Fuchs systems will be represented by a single ordinary differential equation (ODE) whose order coincides with the total number of independent periods. Then the problem solving Picard–Fuchs equations can be encoded into the language of ODE and therefore the study of periods are simplified. As another feature of this formalism, we remark that in contrast with the usual Picard–Fuchs systems written in terms of moduli derivatives there exists a Wronskian for these ordinary differential systems and the Wronskian produces a new relation among periods, moduli, and QCD scale parameter. Especially, in the case of SU(2) it is quite reminiscent of the scaling relation of the prepotential.[22–24] This relation is highly nonlinear in the case of higher rank gauge group, but it reflects the structure of the Picard–Fuchs ODE. Section II discusses these Picard–Fuchs ordinary differential systems. This realization of Picard–Fuchs systems via ODE becomes interesting when we consider a relation to hypergeometric differential equations in multiple variables. For example, in the case of SU(3) gauge theory, we can find another ODE which gives equivalent periods. That is Burchnall's fifth-order equation directly constructed from Appell's F_4 hypergeometric differential equations.[25] In general, the dimension of the solution space of a single ODE constructed from simultaneous PDEs can exceed that of the original PDEs[25] (see also Srivastava and Karlsson[26] and references therein), and Burchnall's equation is the case. In addition, the extra solution which is not a solution to the original PDEs is known to have a very characteristic form. In the case of Burchnall's equation for the SU(3) gauge theory, since four of the five independent solutions are found to correspond to the four periods of the SU(3) gauge theory (this identification is explicitly checked at the semiclassical regime) and these four periods are also solutions to the SU(3) Picard–Fuchs ODE, it is possible to extract a differential equation only for the fifth solution (although the fifth solution is irrelevant to the underlying physics). The formula for this fifth solution obtained in this way is a new one and takes a very different form compared with other fifth-order ODEs constructed from a set of PDEs, e.g., a product of two Bessel functions or Whittaker functions (the ''fifth solution'' to these two cases are quite reminiscent of each other). In our derivation, it is crucial to notice that the fifth solution of the fifth-order ODE (that is, Burchnall's equation) associated with Appell's F_4 is constructed by ''subtracting'' the fourth-order equation satisfied only by periods of the SU(3) Seiberg–Witten curve which are part of the solutions to the fifth-order ODE. In Sec. III, we discuss these aspects of Burchnall's equation as an application of the SU(3) Picard–Fuchs ODE to a theory of hypergeometric equations. Section IV is a brief summary.

Remark: When we simply say ''Picard–Fuchs ODE,'' it always means a single ordinary differential equation in terms of QCD scale parameter derivatives.

II. PICARD–FUCHS ORDINARY DIFFERENTIAL SYSTEMS

A. The hyperelliptic curve

First, let us recall the exact solution to the $SU(n+1)(n\in\mathbf{N})$ gauge theory as an example. On the affine local coordinates $x,y\in\mathbf{C}$, the hyperelliptic curve and the Seiberg–Witten differential are given by[5–7,9,16]

$$y^2=\widetilde{W}^2_{SU(n+1)}-z,\quad \lambda_{SW}=\frac{x\,\partial_x\widetilde{W}_{SU(n+1)}}{y}\,dx, \tag{2.1}$$

where $z=\Lambda_{\mathrm{SU}(n+1)}^{2(n+1)}$ and

$$\widetilde{W}_{\mathrm{SU}(n+1)}=x^{n+1}-\sum_{i=2}^{n+1} s_i x^{n+1-i}. \tag{2.2}$$

Equation (2.2) shows the simple singularity with moduli s_i. This hyperelliptic curve can be compactified to a Riemann surface of genus n after addition of infinity, hence there must be noncontractible $2n$ cycles on this surface and these cycles can be chosen as the canonical bases, i.e., $\alpha_i \cap \alpha_j=\beta_i \cap \beta_j=0$, $\alpha_i \cap \beta_j=-\beta_j \cap \alpha_i=+\delta_{i,j}$. Then we introduce the period vector

$$\Pi=\begin{pmatrix} a_{D_i} \\ a_i \end{pmatrix}, \tag{2.3}$$

where

$$a_i=\oint_{\alpha_i}\lambda_{\mathrm{SW}}, \quad a_{D_i}=\oint_{\beta_i}\lambda_{\mathrm{SW}}. \tag{2.4}$$

Below, we often denote moduli as $u\equiv s_2$ and $v\equiv s_3$.

B. Derivation of Picard–Fuchs ODE

A physically interesting behavior of periods is in the weak coupling region. Of course, the study of periods can be proceeded by using Picard–Fuchs equations, and in general periods are represented by series in moduli. However, by some rearrangement they are found to have a striking feature in their form. Namely, they can be summarized as

$$\text{periods}=\text{classical part}+\text{quantum corrections}. \tag{2.5}$$

Equation (2.5) suggests that it is more convenient to construct periods as a series in QCD scale parameter Λ rather than moduli. In a sense, (2.5) might be implied by Ito and Sasakura[27] in their observation of a general form of (a certain type of) Picard–Fuchs operators, which is summarized schematically as

$$L=L_{\mathrm{cl}}+\Lambda L_{\Lambda}. \tag{2.6}$$

This equality means that the Picard–Fuchs operator L is a sum of the operator L_{cl} whose kernel is classical periods and some operator L_{Λ}. Once the classical periods are known, we can calculate instanton corrected periods, that is, the kernel of L, by assuming L_{Λ} as a perturbation term for small Λ (at semiclassical regime). However, since in the case of other gauge theories with higher rank gauge groups the Picard–Fuchs system is represented by a set of PDEs, such perturbative calculation involves technical problems, therefore another method to obtain instanton corrected periods should be developed. One of the candidates in view of the differential equation is to construct Picard–Fuchs equations by regarding Λ as a unique independent variable instead of moduli. Then the resulting Picard–Fuchs equations will be expressed by an ordinary differential equation. Since the QCD scale parameter always appears in any gauge theory with or without (massive) hypermultiplets, this formulation is convenient when we generalize the method to various gauge theories with any rank gauge group.

Now, let us consider the derivation of this ordinary differential equation. In general, k-times differentiation of λ_{SW} over z gives

$$\frac{d^k\lambda_{\mathrm{SW}}}{dz^k}=\frac{\text{polynomial in } x}{y^{2k+1}}dx, \tag{2.7}$$

but the right-hand side can be decomposed into a sum of Abelian differentials and a total derivative term, if the well-known reduction algorithm is used.[17–21] Accordingly, collecting (2.7) for various k can generate a differential equation. In addition, since all independent periods should be solutions to this equation, the order of the equation must coincide with the total number of them, and in fact it is determined as $2n$. This reduction method is easily confirmed, if the Seiberg–Witten curves are hyperelliptic type. In this way, we get the Picard–Fuchs ODE in the form

$$\left[\frac{d^{2n}}{dz^{2n}}+c_{2n-1}\frac{d^{2n-1}}{dz^{2n-1}}+\cdots+c_0\right]\Pi=0, \tag{2.8}$$

where c_i are functions in moduli.

Also when massive quarks are included, we can obtain a similar ordinary differential equation, but in this case a mass dependent polynomial appears in the denominator of the right-hand side of (2.7). Reduction of such massive differential was also recognized by Marshakov *et al.*[28] in their construction of massive WDVV equations.

C. Examples of Picard–Fuchs ODE

Let us see examples of Picard–Fuchs ODE. The first one is the SU(2) case and then the coefficients in (2.8) are given by

$$c_1=\frac{1}{z}, \quad c_0=\frac{1}{16z(u^2-z)}. \tag{2.9}$$

Next, let us consider the SU(3) case. In this case, the coefficients are

$$\begin{aligned}
c_0&=\frac{-45(3z-4u^3+27v^2)}{2z^2\tilde{\Delta}_{\mathrm{SU}(3)}},\\
c_1&=\frac{45(1053z^2-538zu^3+40u^6+3267zv^2-54u^3v^2-1458v^4)}{2z^2\tilde{\Delta}_{\mathrm{SU}(3)}},\\
c_2&=\frac{1}{4z^2\tilde{\Delta}_{\mathrm{SU}(3)}}[445\,905z^3-8(4u^3-27v^2)^3+z^2(-217\,368u^3+734\,589v^2)\\
&\quad+36z(676u^6-135u^3v^2-29\,889v^4)],\\
c_3&=\frac{1}{z\tilde{\Delta}_{\mathrm{SU}(3)}}[76\,545z^3-162z^2(244u^3-297v^2)-4(4u^3-27v^2)^3\\
&\quad+9z(656u^6-1080u^3v^2-22\,599v^4)],
\end{aligned} \tag{2.10}$$

where $\tilde{\Delta}_{\mathrm{SU}(3)}$ is the product

$$\tilde{\Delta}_{\mathrm{SU}(3)}=(15z-4u^3+27v^2)\Delta_{\mathrm{SU}(3)} \tag{2.11}$$

with the discriminant

$$\Delta_{\mathrm{SU}(3)}=[729z^2+(4u^3-27v^2)^2-54z(4u^3+27v^2)] \tag{2.12}$$

of the SU(3) hyperelliptic curve.

We can easily obtain Picard–Fuchs ODE for SO(5) or Sp(4) gauge group in a similar manner. Though the Sp(4) hyperelliptic curve may be seen to be different from that of the SO(5) theory, the isomorphism of Picard–Fuchs equations between these two theories could be observed by Ito and Sasakura.[27] Also in the case of Picard–Fuchs ODE, this isomorphism can be easily established by the same transformation.[27]

D. Wronskian

As is well known, the scaling relation of the SU(2) prepotential can be generated from Wronskian of the Picard–Fuchs equation, but this is valid only for this SU(2) theory because the Picard–Fuchs equations in other gauge theories consist of partial differential equations. In such theories, ''Wronskian'' does not generally exist. However, our Picard–Fuchs ODE (2.8) admits a Wronskian given by

$$W_{\mathrm{SU}(n+1)}=\begin{vmatrix} a_1 & \cdots & a_n & a_{D_1} & \cdots & a_{D_n} \\ a_1' & \cdots & a_n' & a_{D_1}' & \cdots & a_{D_n}' \\ \vdots & & \vdots & \vdots & & \vdots \\ a_1^{(2n-1)} & \cdots & a_n^{(2n-1)} & a_{D_1}^{(2n-1)} & \cdots & a_{D_n}^{(2n-1)} \end{vmatrix}, \tag{2.13}$$

where $'=d/dz$. Substituting (2.13) into (2.8) shows that $W_{\mathrm{SU}(n+1)}$ satisfies

$$W_{\mathrm{SU}(n+1)}'+c_{2n-1}W_{\mathrm{SU}(n+1)}=0, \tag{2.14}$$

which is integrated to give

$$W_{\mathrm{SU}(n+1)}=\text{const.}\exp\left(-\int c_{2n-1}dz\right), \tag{2.15}$$

where const. is the integration constant to be determined from the comparison of the left- and right-hand sides of (2.15) by asymptotic behavior of periods, but may depend on moduli because in our formulation moduli are regarded as constant.

Note that (2.15) produces a new nonlinear relation between periods and other parameters. For example, in the case of the SU(2) theory with the normalization used by Klemm *et al.*,[5] (2.15) gives

$$W_{\mathrm{SU}(2)}=-\frac{iu}{2\pi z}, \tag{2.16}$$

which is a quite reminiscent expression with the homogeneity relation of prepotential.[22–24] Similarly, for the SU(3) theory, we have

$$W_{\mathrm{SU}(3)}=\frac{15z-(4u^3-27v^2)}{z^4\Delta_{\mathrm{SU}(3)}^2}, \tag{2.17}$$

where the integration constant is normalized to 1 for convenience in Sec. III. Note that the regular singularities of (2.17) are the same as those of the SU(3) Picard–Fuchs ODE.

III. BURCHNALL'S EQUATION

A. Multiterm differential equation

Picard–Fuchs equations obtained in Sec. II can be shown to be classified in terms of a multiterm ordinary differential equation discussed by Burchnall in his study of the relationship among hypergeometric differential equations in multiple variables and certain types of ordinary differential equations.[25] Here, multiterm ordinary differential equation is defined by:

Definition: The k-term ordinary differential equation is the differential equation taking the form

$$\left[f(\theta_z)+\sum_{i=1}^{l} z^i g_i(\theta_z)\right]\Pi=0 \tag{3.1}$$

for some l, where f and g_i are polynomial differential operators in the Euler derivative $\theta_z = z d/dz$. k is the total number of f and nonzero g_i.

For example, in terms of the Euler derivative, our SU(3) Picard–Fuchs ODE can be rewritten as the four-term ODE with

$$\begin{aligned}
f&=2916(x-y)^3(-1+\theta_z)^2\theta_z{}^2,\\
g_1&=-81(x-y)\theta_z{}^2[43x-52y-60(x-y)\theta_z+4(23x+13y)\theta_z{}^2],\\
g_2&=9[(386x-251y)\theta_z{}^2+144(2x-7y)\theta_z{}^3+36(19x+y)\theta_z{}^4-(x-y)(10+13\theta_z)],\\
g_3&=-5(1+3\theta_z)(2+3\theta_z)(-1+6\theta_z)(1+6\theta_z),
\end{aligned} \tag{3.2}$$

where $x=4u^3/27$ and $y=v^2$.

According to Srivastava and Saran,[29] who extended the work of Burchnall to four-term ODE, our SU(3) Picard–Fuchs ODE seems to be representable by a hypergeometric function in the homogeneity form $F(pz,qz,rz)$, where p, q, and r are parameters. In this paper, we could not specify this function, but since the kernel of the SU(3) Picard–Fuchs ODE is essentially written by Appell's F_4 function, some property of F_4 may appear in our SU(3) Picard–Fuchs ODE as a four-term equation. Furthermore, more detailed study indicates that Picard–Fuchs ODE in any rank gauge group can be classified as a k-term equation, but k seems to correspond to $2\times$ (rank of the gauge group).

Finally, note that the Picard–Fuchs ODE in SU(3) gauge theory has a factor $(-1+\theta_z)^2\theta_z^2$ in f polynomial. Therefore, the indicial indices at semiclassical regime are degenerated to -1 and 0. This indicates that there are logarithmic solutions at this regime. Of course, similar observation also holds for SO(5) and Sp(4) Picard–Fuchs ODEs.

B. Appell's equations and Burchnall's equation

As is well known, a hypergeometric function admits a lot of transformations and reducibilities. For example, a Gaussian ${}_2F_1$ system has 24 solutions and Appell's F_1 has 60 solutions. However, these solutions can be more systematically constructed, if we consider an equivalent ODE. In fact, Srivastava and Saran succeeded in finding 120 solutions to the ODE for Lauricella's $F_D^{(3)}$ function.[29] Also in this sense, the study of ODE for a hypergeometric partial differential system is interesting. The method used in these studies followed Burchnall's work.[25] In this paper, we do not attempt to obtain all solutions to Burchnall's equation (see below) like Kummer's 24 solutions, but we can show that the five basic solutions to Burchnall's equation, especially, the extra solution which is not a solution to the F_4 system, can be derived by using SU(3) Picard–Fuchs ODE. Of course, it may also be interesting if this extra solution can be represented by Appell's F_4, but we do not know whether it is possible or not. Nevertheless, we can establish the fifth solution as a simple formula. The reader should notice that the method presented in this paper is to use a fourth-order ODE [SU(3) Picard–Fuchs ODE] satisfied by periods of Riemann surface in genus two [SU(3) Seiberg–Witten curve] and therefore our method is quite different from those mentioned above.

First, let us recall that the SU(3) Picard–Fuchs system[5]

$$\left[\theta_{\tilde{x}}\left(\theta_{\tilde{x}}-\frac{1}{3}\right)-\tilde{x}\left(\theta_{\tilde{x}}+\theta_{\tilde{y}}-\frac{1}{6}\right)\left(\theta_{\tilde{x}}+\theta_{\tilde{y}}-\frac{1}{6}\right)\right]\Pi=0,$$
$$\left[\theta_{\tilde{y}}\left(\theta_{\tilde{y}}-\frac{1}{2}\right)-\tilde{y}\left(\theta_{\tilde{x}}+\theta_{\tilde{y}}-\frac{1}{6}\right)\left(\theta_{\tilde{x}}+\theta_{\tilde{y}}-\frac{1}{6}\right)\right]\Pi=0, \tag{3.3}$$

where we have introduced $\tilde{x}=4u^3/(27\Lambda^6_{\mathrm{SU}(3)})$ and $\tilde{y}=v^2/\Lambda^6_{\mathrm{SU}(3)}$, and $\theta_x=x\partial/\partial x$ and $\theta_y=y\partial/\partial y$ are Euler partial derivatives. In the terminology of Sec. III A, (3.3) consists of two two-term equations and is nothing but the Appell's differential system for the type F_4 hypergeometric double series

$$F_4(\alpha,\beta;\gamma,\gamma';x,y)=\sum_{m,n=0}^{\infty}\frac{(\alpha)_{m+n}(\beta)_{m+n}}{(\gamma)_m(\gamma')_n}\frac{x^m}{m!}\frac{y^n}{n!}, \tag{3.4}$$

where $\alpha=\beta=-1/6$, $\gamma=2/3$, and $\gamma'=1/2$. However, by the scaling transformation

$$\tilde{x}=x\tilde{z},\quad \tilde{y}=y\tilde{z},\quad x=\frac{4u^3}{27},\quad y=v^2,\quad \tilde{z}=\frac{1}{\Lambda^6_{A_2}}, \tag{3.5}$$

we see that (3.3) turns to

$$[\theta_x(\theta_x+\gamma-1)-x\tilde{z}(\theta_x+\theta_y+\alpha)(\theta_x+\theta_y+\beta)]F=0,$$
$$[\theta_y(\theta_y+\gamma'-1)-y\tilde{z}(\theta_x+\theta_y+\alpha)(\theta_x+\theta_y+\beta)]F=0, \tag{3.6}$$

whose analytic solution near $(x,y)=(0,0)$ is given by

$$F(x\tilde{z},y\tilde{z})=\sum_{m,n=0}^{\infty}\frac{(\alpha)_{m+n}(\beta)_{m+n}}{(\gamma)_m(\gamma')_n}\frac{(x\tilde{z})^m}{m!}\frac{(y\tilde{z})^n}{n!}. \tag{3.7}$$

At first sight, since $F(x\tilde{z},y\tilde{z})$ reduces to F_4 for $\tilde{z}\to 1$, this scale transformation may be trivial, but (3.6) was used as a starting point in Burchnall's work on a set of partial differential equations.[25]

In fact, Burchnall noticed on the homogeneity relation of F

$$(\theta_x+\theta_y-\theta_{\tilde{z}})F=0, \tag{3.8}$$

where $\theta_{\tilde{z}}$ is the ordinary differential operator $\theta_{\tilde{z}}=\tilde{z}d/d\tilde{z}$, and finally arrived at the ordinary differential equation of fifth order (see also Appendix A)

$$\Big[f_0-2(x+y)\tilde{z}f_1(\theta_{\tilde{z}}+\alpha)(\theta_{\tilde{z}}+\beta)+\frac{1}{2}(x-y)\tilde{z}f_2(\theta_{\tilde{z}}+\alpha)(\theta_{\tilde{z}}+\beta)$$
$$+(x-y)^2\tilde{z}^2f_3(\theta_{\tilde{z}}+\alpha)(\theta_{\tilde{z}}+\alpha+1)(\theta_{\tilde{z}}+\beta)(\theta_{\tilde{z}}+\beta+1)\Big]F=0, \tag{3.9}$$

where

$$f_0=\theta_{\tilde{z}}(\theta_{\tilde{z}}+\gamma-1)(\theta_{\tilde{z}}+\gamma'-1)(\theta_{\tilde{z}}+\gamma+\gamma'-2)\left(\theta_{\tilde{z}}+\frac{\gamma}{2}+\frac{\gamma'}{2}-2\right),$$
$$f_1=\left(\theta_{\tilde{z}}+\frac{\gamma}{2}+\frac{\gamma'}{2}\right)\left(\theta_{\tilde{z}}+\frac{\gamma}{2}+\frac{\gamma'}{2}-\frac{1}{2}\right)\left(\theta_{\tilde{z}}+\frac{\gamma}{2}+\frac{\gamma'}{2}-1\right),$$
$$f_2=(\gamma-\gamma')(\gamma+\gamma'-2)\left(\theta_{\tilde{z}}+\frac{\gamma}{2}+\frac{\gamma'}{2}-\frac{1}{2}\right),$$
$$f_3=\left(\theta_{\tilde{z}}+\frac{\gamma}{2}+\frac{\gamma'}{2}+1\right). \tag{3.10}$$

In contrast with the SU(3) Picard–Fuchs ODE, since Burchnall's equation (3.9) is classified as a three-term ODE, it reflects a consequence of the nature of two variables hypergeometric function in the form $F(pz,qz)$, where p and q are parameters.[25]

As a direct calculation shows, (3.9) cannot be expressed in the form $[\theta_{\tilde{z}}+K(x,y,\tilde{z})]LF=0$, where $K(x,y,\tilde{z})$ is some function of x, y, and $\tilde{z}$, and L is some differential operator of fourth order. Accordingly, such L does not exist and therefore the relation between our SU(3) Picard–Fuchs ODE and Burchnall's equation is nontrivial. However, note that the coefficient of highest power in the Euler derivative corresponds to the discriminant of the SU(3) curve

$$\Delta\equiv[1-2(x+y)\tilde{z}+(x-y)^2\tilde{z}^2]=\frac{1}{729\Lambda_{\mathrm{SU}(3)}^{12}}[4u^3-27(v+\Lambda_{\mathrm{SU}(3)}^3)^2][4u^3-27(v-\Lambda_{A_2}^3)^2]. \tag{3.11}$$

Therefore, in a sense Burchnall's equation is reminiscent of the SU(3) Picard–Fuchs ODE, but these two are not completely equivalent. Clarifying the relation between these two equations is the subject of the rest of the paper.

C. Four solutions at semiclassical regime

It would be instructive to get solutions explicitly around $\Lambda_{\mathrm{SU}(3)}=0(\tilde{z}=\infty)$, which is a regular singular point of the equation. In the work of Burchnall, solutions of (3.9) were not calculated at any singularities, but since (3.9) is a linear ordinary differential equation, it is easy to solve it by traditional Frobenius's method under the assumption $F=\tilde{z}^{-\nu}\Sigma_{n=0}^{\infty}A_n\tilde{z}^{-n}$ for some ν and A_n. Then the indicial indices are determined as

$$\nu=\alpha,\beta,\alpha+1,\beta+1,(\gamma+\gamma'+2)/2 \tag{3.12}$$

or equivalently,

$$\nu_1=-1/6,\quad \nu_2=5/6,\quad \nu_3=19/12, \tag{3.13}$$

where ν_1 and ν_2 are actually double roots. The solution for ν_3 is the subject of Sec. III D.

Equation (3.9) produces the recursion relations

$$\begin{gathered}(x-y)^2\rho_1A_1-(\nu_i-\alpha)(\nu_i-\beta)\phi_1A_0=0,\\ (x-y)^2\rho_nA_n-(\nu_i+n-\alpha-1)(\nu_i+n-\beta-1)\sigma_nA_{n-1}+\chi_nA_{n-2}=0,\quad n>1,\end{gathered} \tag{3.14}$$

where

$$\begin{aligned}\rho_n&=(2\nu_i+2n-\gamma-\gamma'-2)(\nu_i+n-\alpha)(\nu_i+n-\beta)(\nu_i+n-\alpha-1)(\nu_i+n-\beta-1),\\ \sigma_n&=(\nu_i+n-\gamma-\gamma')(2\nu_i+2n-\gamma-\gamma')[x(\nu_i+n-\gamma'-1)+y(\nu_i+n-\gamma-1)]\\ &\quad+(\nu_i+n-1)(2\nu_i+2n-\gamma-\gamma'-2)[x(\nu_i+n-\gamma)+y(\nu_i+n-\gamma')],\\ \chi_n&=(\nu_i+n-2)(\nu_i+n-\gamma-1)(\nu_i+n-\gamma'-1)(\nu_i+n-\gamma-\gamma')(2\nu_i+2n-\gamma-\gamma')\end{aligned} \tag{3.15}$$

with $A_0=1$. Here, repeated indices are assumed *not* to be summed. If these recursion relations are used, the solutions corresponding to respective indicial indices will be obtained, but we must be careful, because there are indicial indices which differ by unit among them, i.e., α and $\alpha+1$, and β and $\beta+1$. For example, let $\nu=\alpha$. Then the recursion relations produce the nth coefficient as a linear combination of A_0 and A_1. Thus the solution is given by a linear combination of $\tilde{z}^{-\alpha}(A_0+\cdots)$ and $\tilde{z}^{-\alpha-1}(A_1+\cdots)$, but the ''indicial index'' of the last series can be seen as $\alpha+1$. Therefore, the last series can be also regarded as a solution corresponding to this index. In fact, it is easy to see that explicit construction of the solution supports this observation. Of course, in this case, since we would like to get a solution corresponding to the index α, A_1 can be set to zero without loss of generality. For this reason, A_1 is chosen as zero for the indices α and β, while that for $\alpha+1$, $\beta+1$, and $(\gamma+\gamma'-2)/2$ should be determined from the first equation in (3.14).

In this way, we get the regular series solutions ($i=1,2$)

$$\varphi_i=\tilde{z}^{-\nu_i}\sum_{n=1}^{\infty} A_{i,n}\tilde{z}^{-n}, \tag{3.16}$$

where the first few coefficients are given by

$$\begin{aligned} A_{i,0}&=1,\\ A_{1,1}&=0,\\ A_{1,2}&=\frac{5}{648(x-y)^2},\\ A_{1,3}&=\frac{35(41x+40y)}{209\,952(x-y)^4},\\ A_{2,1}&=\frac{5(5x+4y)}{48(x-y)^2},\\ A_{2,2}&=\frac{35(157x^2+460xy+112y^2)}{15\,552(x-y)^4},\\ A_{2,3}&=\frac{385(18\,671x^3+119\,352x^2y+105\,504xy^2+12\,352y^3)}{26\,873\,856(x-y)^6}. \end{aligned} \tag{3.17}$$

On the other hand, the degeneracy of ν_1 and ν_2 produce the logarithmic solutions ($j=1,2$)

$$\tilde{\varphi}_j=\varphi_j \ln\frac{1}{\tilde{z}}+\tilde{z}^{-\nu_j}\sum_{n=1}^{\infty} B_{j,n}\tilde{z}^{-n}, \tag{3.18}$$

where some of $B_{j,n}$ are

$$\begin{aligned} B_{1,1}&=0,\\ B_{1,2}&=-\frac{17}{1296(x-y)^2},\\ B_{1,3}&=-\frac{(11\,761x+11\,216y)}{1\,259\,712(x-y)^4},\\ B_{2,1}&=\frac{49x+104y}{144(x-y)^2},\\ B_{2,2}&=\frac{14\,273x^2+70\,940xy+28\,268y^2}{46\,656(x-y)^4},\\ B_{2,3}&=\frac{41\,936\,917x^3+383\,568\,144x^2y+472\,854\,144xy^2+79\,210\,112y^3}{161\,243\,136(x-y)^6}. \end{aligned} \tag{3.19}$$

It is interesting to notice that these series are composed by a series in powers of $1/(x-y)=27/(4u^3-27v^2)$ which detects the discriminant of the semiclassical SU(3) curve. This feature is useful when we consider the structure of the quantum moduli space of the SU(3) gauge theory.

For this purpose, let us recall the work of Klemm *et al.*[5] In the course of the analysis of the quantum moduli space of the SU(3) gauge theory, Klemm *et al.*[5] found that the quantum moduli space could be better understood as the complex projective space $\mathbf{CP}^2$ with singularities which correspond to the strong coupling regime. Then this space can be covered by the three local (inhomogeneous) coordinates

$$P_1:\left(\frac{4u^3}{27\Lambda^6}:\frac{v^2}{\Lambda^6}:1\right),\quad P_2:\left(\frac{4u^3}{27v^2}:1:\frac{\Lambda^6}{v^2}\right),\quad P_3:\left(1:\frac{27v^2}{4u^3}:\frac{27\Lambda^6}{4u^3}\right), \tag{3.20}$$

where $\Lambda\equiv\Lambda SU(3)$. For this reason, periods should be obtained at each coordinate patch, hence the periods derived in this way are locally valid. However, our solutions have a (slightly) nice property, because the basis of solution space are common both on P_2 and P_3. That is, to get periods on P_2, it is enough to further expand φ_i by v, while on P_3 by large u.

In fact, we can see that the four periods are expressed by linear combinations of (3.16) and (3.18). For example, to match periods on the patch P_3, let us define ω_i and Ω_j by linear combinations of φ_i

$$\tilde{\omega}_1=c_1\varphi_1+c_2\varphi_2,\quad \tilde{\omega}_2=c_3\varphi_1+c_4\varphi_2,$$

$$\tilde{\Omega}_1=\tilde{\omega}_1\ln\frac{27}{4u^3\tilde{z}}+\sum_{i=1}^{2}\sum_{n=1}^{\infty}c_iB_{i,n}\tilde{z}^{-\nu_i-n}+c_5\varphi_1+c_6\varphi_2, \tag{3.21}$$

$$\tilde{\Omega}_2=\tilde{\omega}_2\ln\frac{27}{4u^3\tilde{z}}+\sum_{i=1}^{2}\sum_{n=1}^{\infty}c_{i+2}B_{i,n}\tilde{z}^{-\nu_i-n}+c_7\varphi_1+c_8\varphi_2,$$

where

$$c_1={}_2F_1\left(-\frac{1}{6},\frac{1}{6};\frac{1}{2};\frac{27v^2}{4u^3}\right),\quad c_2=-\frac{3}{16u^3}\,{}_2F_1\left(\frac{5}{6},\frac{7}{6};\frac{1}{2};\frac{27v^2}{4u^3}\right),$$

$$c_3={}_2F_1\left(\frac{1}{3},\frac{2}{3};\frac{3}{2};\frac{27v^2}{4u^3}\right),\quad c_4=\frac{3}{2u^3}\,{}_2F_1\left(\frac{4}{3},\frac{5}{3};\frac{3}{2};\frac{27v^2}{4u^3}\right),$$

$$c_5=c_1+\sum_{n=1}^{\infty}\frac{(-1/6)_n(1/6)_n}{(1/2)_n n!}\left[\psi\left(n-\frac{1}{6}\right)-\psi\left(-\frac{1}{6}\right)+\psi\left(n+\frac{1}{6}\right)-\psi\left(\frac{1}{6}\right)\right]\left(\frac{27v^2}{4u^3}\right)^n, \tag{3.22}$$

$$c_6=c_2-\frac{3}{16u^3}\sum_{n=0}^{\infty}\frac{(5/6)_n(7/6)_n}{(1/2)_n n!}\left[2\psi(1)-2\psi(2)+\psi\left(n+\frac{5}{6}\right)-\psi\left(-\frac{1}{6}\right)+\psi\left(n+\frac{7}{6}\right)-\psi\left(\frac{1}{6}\right)\right]\times\left(\frac{27v^2}{4u^3}\right)^n,$$

$$c_7=c_3+\sum_{n=1}^{\infty}\frac{(1/3)_n(2/3)_n}{(3/2)_n n!}\left[\psi\left(n+\frac{1}{3}\right)-\psi\left(\frac{1}{3}\right)+\psi\left(n+\frac{2}{3}\right)-\psi\left(\frac{2}{3}\right)\right]\left(\frac{27v^2}{4u^3}\right)^n,$$

$$c_8 = c_4 + \frac{3}{2u^3}\sum_{n=0}^{\infty}\frac{(4/3)_n(5/3)_n}{(3/2)_n n!}\left[2\psi(1) - 2\psi(2) + \psi\left(n+\frac{4}{3}\right) - \psi\left(\frac{1}{3}\right) + \psi\left(n+\frac{5}{3}\right) - \psi\left(\frac{2}{3}\right)\right]$$
$$\times\left(\frac{27v^2}{4u^3}\right)^n.$$

Here, $\psi(x) = d\ln\Gamma(x)/dx$ is the digamma function and ${}_2F_1$ is the hypergeometric function whose series representation is given by

$$ {}_2F_1(a,b;c;x) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\frac{x^n}{n!}, \tag{3.23}$$

where $(*)_n = \Gamma(*+n)/\Gamma(*)$ is the Pochhammer symbol. In order to make a contact with the normalization used by Klemm *et al.*,[5] we rescale as

$$\omega_1 = 2\sqrt{u}\Lambda\tilde{\omega}_1, \quad \omega_2 = \frac{v\Lambda}{u}\tilde{\omega}_2,$$
$$\Omega_1 = 2\sqrt{u}\Lambda\tilde{\Omega}_1, \quad \Omega_2 = \frac{v\Lambda}{u}\tilde{\Omega}_2. \tag{3.24}$$

Then the periods a_j and a_{D_j} can be given by

$$a_1 = \frac{1}{2}(\omega_1 + \omega_2), \quad a_2 = \frac{1}{2}(\omega_1 - \omega_2),$$
$$a_{D_1} = -\frac{i}{4}(\Omega_1 + 3\Omega_2) - \frac{i}{\pi}(\delta_1\omega_1 - \delta_2\omega_2), \tag{3.25}$$
$$a_{D_2} = -\frac{i}{4}(\Omega_1 - 3\Omega_2) - \frac{i}{\pi}(\delta_1\omega_1 + \delta_2\omega_2),$$

where $\delta_1 = i(5 - 3\ln 3 - 4\ln 2)/4$ and $\delta_2 = 3i(1 + 3\ln 3)/4$ are constants determined from asymptotic expansion of periods.[5] The identification of periods by our solutions can be easily established by expanding (3.25) at $u = \infty$. On the other hand, for large v, i.e., on the patch P_2, φ_i are expanded at $v = \infty$ and then consider similar linear combinations.

In this way, we can check that the series solutions for $v = v_1$ and v_2 comprise in fact the four periods.

D. The fifth solution

We have seen that the four solutions with indicial indices ν_1 and ν_2 of Burchnall's fifth-order equation in fact yield the four periods of the SU(3) gauge theory. However, there exists an extra solution in (3.9). Therefore, the appearance can be regarded as characteristic in the ordinary differential form of the partial differential system.

In this section, we show that it is possible to derive a fourth-order equation satisfied by this fifth solution with aid of the SU(3) Picard–Fuchs ODE and we derive the closed formula for this fifth solution as an application of the SU(3) Picard–Fuchs ODE.

First, let us rewrite (3.9) in the form

$$\left[\frac{d^5}{dz^5} + c_{B,4}\frac{d^4}{dz^4} + \cdots + c_{B,0}\right]F = 0, \tag{3.26}$$

where $z=1/\tilde{z}$ and $c_{B,i}$ are some functions in x, y, and z. As we have already seen in Sec. III C, since four of the five independent solutions to (3.9) are regarded as the four periods of the SU(3) gauge theory, we can write the Wronskian for (3.26) as

$$W_B=\begin{vmatrix} a_1 & a_2 & a_{D_1} & a_{D_2} & h \\ a_1' & a_2' & a_{D_1}' & a_{D_2}' & h' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_1^{(4)} & a_2^{(4)} & a_{D_1}^{(4)} & a_{D_2}^{(4)} & h^{(4)} \end{vmatrix}, \tag{3.27}$$

where $'=d/dz$ and the fifth solution is denoted by h. Again from basic differential calculation, W_B is generated from $c_{B,4}$ and is found to be [cf. (2.15)]

$$W_B=\frac{1}{z^{85/12}[x^2+(y-z)^2-2x(y+z)]^3} \tag{3.28}$$

up to the normalization of integration constant, which is irrelevant to the following discussion.

Next, recall that the fourth-order derivatives of periods can be reduced to a linear combination of lower order derivatives by using the SU(3) Picard–Fuchs ODE. Therefore, from (3.27) with the SU(3) Picard–Fuchs ODE, we see that

$$W_B=W_{\mathrm{SU}(3)}(h^{(4)}+c_3h'''+c_2h''+c_1h'+c_0h), \tag{3.29}$$

where c_i are given by (2.10). From (3.28) and (2.17), it is immediate to obtain the differential equation for h,

$$h^{(4)}+c_3h'''+c_2h''+c_1h'+c_0h=R, \tag{3.30}$$

where

$$R=\frac{1}{z^{37/12}[5z-9(x-y)][x^2+(y-z)^2-2x(y+z)]}. \tag{3.31}$$

It is interesting to note that h satisfies the SU(3) Picard–Fuchs ODE with a source term. It is now easy to get a general solution to (3.30) (see also Appendix B)

$$h=\sum_{i=1}^{2}\rho_i a_i+\sum_{i=1}^{2}\epsilon_i a_{D_i}-a_1\int_0\frac{Rw_1}{W_{\mathrm{SU}(3)}}dx+a_2\int\frac{Rw_2}{W_{\mathrm{SU}(3)}}dx-a_{D_1}\int_0\frac{Rw_3}{W_{\mathrm{SU}(3)}}dx+a_{D_2}\int_0\frac{Rw_4}{W_{\mathrm{SU}(3)}}dx, \tag{3.32}$$

where ρ_i and ϵ_i are integration constants, the integration symbol is the integration constant free integral, and w_i are determinants defined by

$$w_1=\begin{vmatrix} a_2 & a_{D_1} & a_{D_2} \\ a_2' & a_{D_1}' & a_{D_2}' \\ a_2'' & a_{D_1}'' & a_{D_2}'' \end{vmatrix}, \quad w_2=\begin{vmatrix} a_1 & a_{D_1} & a_{D_2} \\ a_1' & a_{D_1}' & a_{D_2}' \\ a_1'' & a_{D_1}'' & a_{D_2}'' \end{vmatrix},$$

$$w_3=\begin{vmatrix} a_1 & a_2 & a_{D_2} \\ a_1' & a_2' & a_{D_2}' \\ a_1'' & a_2'' & a_{D_2}'' \end{vmatrix}, \quad w_4=\begin{vmatrix} a_1 & a_2 & a_{D_1} \\ a_1' & a_2' & a_{D_1}' \\ a_1'' & a_2'' & a_{D_1}'' \end{vmatrix}. \tag{3.33}$$

$W_{SU(3)}$ in (3.32) can be identified with (2.17).

To summarize, we have succeeded in finding a closed representation of the fifth solution by using the SU(3) Picard–Fuchs ODE. It is interesting to compare (3.32) with the fifth solution of other three-term differential equation discussed by Burchnall.[25] In the case of a product of two Bessel functions, for instance, the fifth solution is expressed by an integral of a product of two ''Wronskians'' each of which is written by two other independent solutions.[25] However, our fifth solution does not admit such factorization, so (3.32) seems to imply the fact that the Appell function F_4 with parameters $\alpha=\beta=-1/6$, $\gamma=2/3$, and $\gamma'=1/2$ [or equivalently, the SU(3) periods] cannot be factored into a product of two (nontrivial) functions.

Remark: In the semiclassical regime, series representation of this fifth solution corresponds to $\nu_3=19/12$ *in (3.13), and this can be also seen from the terms not including* ρ_i *and* ϵ_i *in (3.32).*

IV. SUMMARY

In this paper, we have discussed the Picard–Fuchs equations appearing in $N=2$ supersymmetric Yang–Mills theories in view of ordinary differential equation and realized Picard–Fuchs equations as a system of ODEs. This construction has given a new relation among periods, moduli, and QCD parameter by using the Wronskian and this is a systematic way to get such nonlinear relation among periods in higher rank gauge group cases.

In the case of SU(3), we have also found that Burchnall's ordinary differential equation for Appell's F_4 is a candidate of Picard–Fuchs ODE by identifying the SU(3) QCD mass scale parameter with the scaling variable used in Burchnall's observation and confirmed that four of the five solutions of Burchnall's equation in fact coincide with SU(3) periods. As for the fifth solution, it has been shown that it has a simple and closed form by using the SU(3) Picard–Fuchs ODE. Of course, even if we consider arbitrary Riemann surface of genus two and try to get a similar result for the fifth solution to Burchnall's equation, the derivation will be failed because we cannot always have an equation in fourth order like SU(3) Picard–Fuchs ODE. Note that the SU(3) Seiberg–Witten curve is a specific choice and the appearance of $\Lambda_{SU(3)}$ plays the central role in the discussion.

Generalization of Burchnall's construction of ordinary differential equation from a set of partial differential equations to Picard–Fuchs equations in other gauge theories is straightforward, but we do not know whether Burchnall type equations exist for Picard–Fuchs equations in those gauge theories. Studying these cases will open further aspects of Picard–Fuchs equations and hypergeometric nature of the equations constructed from the homogeneous hypergeometric equations in multiple variables.

Finally, as another direction, since our equation is ODE in contrast with the usual Picard–Fuchs systems, it may be possible to consider a relation to classical W-algebras in view of Picard–Fuchs ODE.[30]

ACKNOWLEDGMENT

I acknowledge Professor P. W. Karlsson, who kindly gave me a comment on my lengthy questions about Burchnall's work.

APPENDIX A: DERIVATION OF THE BURCHNALL'S EQUATION

In this appendix, we briefly review the derivation of Burchnall's equation. The reader is also recommended to refer to the original paper.[25]

First, notice that from (3.6) it is easy to obtain

$$(\theta_{\tilde{z}}+\gamma+\gamma'-2)\theta_x\theta_y F=(x\tilde{z}\theta_y+y\tilde{z}\theta_x)XF, \tag{A1}$$

where $X=(\theta_{\tilde{z}}+\alpha)(\theta_{\tilde{z}}+\beta)$. We can also obtain

$$UF \equiv [\theta_{\tilde z}(\theta_{\tilde z}+\gamma-1)-(x+y)\tilde zX]F=(2\theta_x+\gamma-\gamma')\theta_y F,$$
$$U'F \equiv [\theta_{\tilde z}(\theta_{\tilde z}+\gamma'-1)-(x+y)\tilde zX]F=(2\theta_y+\gamma'-\gamma)\theta_x F \tag{A2}$$

and

$$\theta_x(\theta_{\tilde z}+\gamma-1)(\theta_{\tilde z}+\gamma'-1)F=[(\theta_{\tilde z}+\gamma'-1)\theta_x\theta_y+x\tilde z(\theta_{\tilde z}+\gamma')X]F,$$
$$\theta_y(\theta_{\tilde z}+\gamma'-1)(\theta_{\tilde z}+\gamma-1)F=[(\theta_{\tilde z}+\gamma-1)\theta_x\theta_y+y\tilde z(\theta_{\tilde z}+\gamma)X]F. \tag{A3}$$

Addition of the two equations in (A3) provides

$$VF \equiv [\theta_{\tilde z}(\theta_{\tilde z}+\gamma-1)(\theta_{\tilde z}+\gamma'-1)-x\tilde z(\theta_{\tilde z}+\gamma')X-y\tilde z(\theta_{\tilde z}+\gamma)X]F=(2\theta_{\tilde z}+\gamma+\gamma'-2)\theta_x\theta_y F, \tag{A4}$$

thus

$$(\theta_{\tilde z}+\gamma+\gamma'-2)VF=[4xy\tilde z^2XX_{+1}+x\tilde zXU+y\tilde zXU'+2(\theta_{\tilde z}+\gamma+\gamma'-2)\theta_x\theta_y]F, \tag{A5}$$

where $X_{+1}=(\theta_{\tilde z}+\alpha+1)(\theta_{\tilde z}+\beta+1)$, and (A1) and (A2) have been used. Moreover, from (A5) we have

$$\begin{aligned} WF \equiv [&\theta_{\tilde z}(\theta_{\tilde z}+\gamma+\gamma'-2)(\theta_{\tilde z}+\gamma-1)(\theta_{\tilde z}+\gamma'-1)-x\tilde z[(\theta_{\tilde z}+\gamma+\gamma'-1)(\theta_{\tilde z}+\gamma') \\ &+\theta_{\tilde z}(\theta_{\tilde z}+\gamma-1)]X-y\tilde z[(\theta_{\tilde z}+\gamma+\gamma'-1)(\theta_{\tilde z}+\gamma)+\theta_{\tilde z}(\theta_{\tilde z}+\gamma'-1)]X \\ &+(x-y)^2\tilde z^2XX_{+1}]F=2(\theta_{\tilde z}+\gamma+\gamma'-2)\theta_x\theta_y F. \end{aligned} \tag{A6}$$

Therefore, the expected equation is given by rearrangement of

$$(2\theta_{\tilde z}+\gamma+\gamma'-2)WF=2(\theta_{\tilde z}+\gamma+\gamma'-2)VF, \tag{A7}$$

i.e.,

$$[Y_0-x\tilde zY_1X-y\tilde zY_2X+(x-y)^2\tilde z^2(2\theta_{\tilde z}+\gamma+\gamma'+2)XX_{+1}]F=0, \tag{A8}$$

where

$$\begin{aligned} Y_0&=\theta_{\tilde z}(\theta_{\tilde z}+\gamma-1)(\theta_{\tilde z}+\gamma'-1)(\theta_{\tilde z}+\gamma+\gamma'-2)(2\theta_{\tilde z}+\gamma+\gamma'-4), \\ Y_1&=(\theta_{\tilde z}+\gamma')(\theta_{\tilde z}+\gamma+\gamma'-1)(2\theta_{\tilde z}+\gamma+\gamma'-2)+\theta_{\tilde z}(\theta_{\tilde z}+\gamma-1)(2\theta_{\tilde z}+\gamma+\gamma'), \\ Y_2&=(\theta_{\tilde z}+\gamma)(\theta_{\tilde z}+\gamma+\gamma'-1)(2\theta_{\tilde z}+\gamma+\gamma'-2)+\theta_{\tilde z}(\theta_{\tilde z}+\gamma'-1)(2\theta_{\tilde z}+\gamma+\gamma'). \end{aligned} \tag{A9}$$

Of course (A8) is equivalent to (3.9).

APPENDIX B: GENERAL SOLUTION TO FOURTH-ORDER ODE

This Appendix reviews a construction of a general solution to the fourth-order linear ordinary differential equation

$$y^{(4)}+P(x)y'''+Q(x)y''+R(x)y'+S(x)y=T(x), \tag{B1}$$

where $'=d/dx$ and P, Q, R, S, and T are some functions in x.

First, let $T=0$ and let y_i $(i=1,...,4)$ be the fundamental solutions to

$$y^{(4)}+P(x)y'''+Q(x)y''+R(x)y'+S(x)y=0. \tag{B2}$$

Then we assume that the general solution to (B1) is represented in the form

$$y=\sum_{i=1}^{4} C_i(x)y_i \tag{B3}$$

by using unknown coefficients C_i. Differentiating (B3), we can obtain

$$y'=\sum_{i=1}^{4} C_i'y_i+\sum_{i=1}^{4} C_i y_i', \tag{B4}$$

but we further assume that the first term in the right-hand side vanishes. Namely, we have

$$\sum_{i=1}^{4} C_i'y_i=0 \tag{B5}$$

and

$$y'=\sum_{i=1}^{4} C_i y_i'. \tag{B6}$$

Repeating differentiation and imposing the vanishing of terms including C_i', we get

$$\sum_{i=1}^{4} C_i'y_i'=0, \quad \sum_{i=1}^{4} C_i'y_i''=0 \tag{B7}$$

and

$$y''=\sum_{i=1}^{4} C_i y_i'', \quad y'''=\sum_{i=1}^{4} C_i y_i'''. \tag{B8}$$

As for $y^{(4)}$, we assume

$$y^{(4)}=\sum_{i=1}^{4} C_i'y_i'''+\sum_{i=1}^{4} C_i y_i^{(4)}. \tag{B9}$$

Then from (B1), (B9), and (B8), we get

$$\sum_{i=1}^{4} C_i'y_i'''=T. \tag{B10}$$

In this way, we can arrive at the matrix equation determining all C_i

$$YC={}^{T}(0,0,0,T), \tag{B11}$$

where

$$Y=\begin{pmatrix} y_1 & \cdots & y_4 \\ y_1' & \cdots & y_4' \\ y_1'' & \cdots & y_4'' \\ y_1''' & \cdots & y_4''' \end{pmatrix}, \quad C=\begin{pmatrix} C_1' \\ \vdots \\ C_4' \end{pmatrix}. \tag{B12}$$

Consequently, C_i are given by

$$C_1=c_1-\int_0 \frac{Tw_1}{\det Y}dx, \quad C_2=c_2+\int_0 \frac{Tw_2}{\det Y}dx, \quad C_3=c_3-\int_0 \frac{Tw_3}{\det Y}dx, \quad C_4=c_4+\int_0 \frac{Tw_4}{\det Y}dx, \tag{B13}$$

where c_i are integration constants, the integration symbol is the integration constant free integral and

$$w_1=\begin{vmatrix} y_2 & y_3 & y_4 \\ y_2' & y_3' & y_4' \\ y_2'' & y_3'' & y_4'' \end{vmatrix}, \quad w_2=\begin{vmatrix} y_1 & y_3 & y_4 \\ y_1' & y_3' & y_4' \\ y_1'' & y_3'' & y_4'' \end{vmatrix}, \quad w_3=\begin{vmatrix} y_1 & y_2 & y_4 \\ y_1' & y_2' & y_4' \\ y_1'' & y_2'' & y_4'' \end{vmatrix}, \quad w_1=\begin{vmatrix} y_1 & y_2 & y_3 \\ y_1' & y_2' & y_3' \\ y_1'' & y_2'' & y_3'' \end{vmatrix}. \tag{B14}$$

[1] N. Seiberg, Phys. Lett. B **206**, 75 (1988).
[2] N. Seiberg and E. Witten, Nucl. Phys. B **431**, 484 (1994).
[3] N. Seiberg and E. Witten, Nucl. Phys. B **435**, 129 (1994).
[4] J. J. Gray, Bull. Am. Math. Soc. **10** (1), 1 (1984).
[5] A. Klemm, W. Lerche, and S. Theisen, Int. J. Mod. Phys. A **11**, 1929 (1996).
[6] A. Klemm, W. Lerche, S. Yankielowicz, and S. Theisen, Phys. Lett. B **344**, 169 (1995).
[7] P. C. Argyres and A. Faraggi, Phys. Rev. Lett. **74**, 3931 (1995).
[8] P. C. Argyres, M. R. Plesser, and A. D. Shapere, Phys. Rev. Lett. **75**, 1699 (1995).
[9] A. Hanany and Y. Oz, Nucl. Phys. B **452**, 283 (1995).
[10] U. H. Danielsson and B. Sundborg, Phys. Lett. B **358**, 273 (1995).
[11] U. H. Danielsson and B. Sundborg, Phys. Lett. B **370**, 83 (1996).
[12] A. Brandhuber and K. Landsteiner, Phys. Lett. B **358**, 73 (1995).
[13] A. Hanany, Nucl. Phys. B **466**, 85 (1996).
[14] M. Alishahiha, F. Ardalan, and F. Mansouri, Phys. Lett. B **381**, 446 (1996).
[15] K. Landsteiner, J. M. Pierre, and S. B. Giddings, Phys. Rev. D **55**, 2367 (1997).
[16] M. R. Abolhasani, M. Alishahiha, and A. M. Ghezelbash, Nucl. Phys. B **480**, 279 (1996).
[17] J. M. Isidro, A. Mukherjee, J. P. Nunes, and H. J. Schnitzer, Nucl. Phys. B **492**, 647 (1997).
[18] J. M. Isidro, A. Mukherjee, J. P. Nunes, and H. J. Schnitzer, Int. J. Mod. Phys. A **13**, 233 (1998).
[19] J. M. Isidro, A. Mukherjee, J. P. Nunes, and H. J. Schnitzer, Nucl. Phys. B **502**, 363 (1997).
[20] M. Alishahiha, Phys. Lett. B **398**, 100 (1997).
[21] M. Alishahiha, Phys. Lett. B **418**, 317 (1998).
[22] M. Matone, Phys. Lett. B **357**, 342 (1995).
[23] J. Sonnenschein, S. Theisen, and S. Yankielowicz, Phys. Lett. B **367**, 145 (1996).
[24] T. Eguchi and S.-K. Yang, Mod. Phys. Lett. A **11**, 131 (1996).
[25] J. L. Burchnall, Q. J. Math. **13**, 90 (1942).
[26] H. M. Srivastava and P. W. Karlsson, *Multiple Gaussian Hypergeometric Series* (Horwood, West Sussex, 1985).
[27] K. Ito and N. Sasakura, Nucl. Phys. B **484**, 141 (1997).
[28] A. Marshakov, A. Mironov, and A. Morozov, ''More evidence for the WDVV equations in $N=2$ SUSY Yang–Mills theories,'' hep-th/9701123.
[29] H. M. Srivastava and S. Saran, Ganita **16**, 139 (1965).
[30] P. Di Francesco, C. Itzykson, and J.-B. Zuber, Commun. Math. Phys. **140**, 543 (1991).

On the signature of certain intersection forms

Feng Xu[a)]
Department of Mathematics, University of Oklahoma,
601 Elm Avenve, Room 423 Norman, Oklahoma 73019

(Received 9 December 1998; accepted for publication 7 January 1999)

We prove a conjecture of Zuber on the signature of intersection forms associated with affine algebras of type A, which is based on connections between $N=2$ integrable models in two dimensions and certain class of graphs. © *1999 American Institute of Physics.* [S0022-2488(99)03006-6]

I. INTRODUCTION

Let $N\geqslant 2$ be a positive integer and $\lambda'=(\lambda_1',\ldots,\lambda_{N-1}')$, $0<\lambda_i'<1$, $i=1,2,...,N-1$ with $\Sigma_{0<i<N}\lambda_i'<1$. Define

$$p_i(\lambda'):=\lambda_i'+\cdots+\lambda_{N-1}'-\frac{1}{N}\sum_{0<j<N} j\lambda_j', \quad i=1,2,...,N-1$$

and $p_N(\lambda'):=-(1/N)\Sigma_{0<j<N} j\lambda_j'$. Define

$$q_i(\lambda'):=-Np_i(\lambda')+\frac{N+1}{2}-i$$

and

$$g_i(\lambda'):=(-1)^i\prod_{r=1}^{N} 2\cos\left(\pi\left(p_r(\lambda')-\frac{i}{N}\right)\right), \quad i=1,2,...,N.$$

If S is a finite sequence of real numbers, we define $b_+(S)$ [respectively $b_-(S)$, $b_0(S)$] to be the number of positive (respectively, negative, zero) elements in S. Let $a(S):=b_+(S)-b_-(S)$ and denote by $Q_{\lambda'}$, $G_{\lambda'}$ the following two sets:

$$Q_{\lambda'}:=\{\cos(\pi q_1(\lambda')),\ldots,\cos(\pi q_N(\lambda'))\},$$

$$G_{\lambda'}:=\{g_1(\lambda'),\ldots,g_N(\lambda')\}.$$

Notice that since $\cos(\pi q_i(\lambda'))>0$ iff $q_i(\lambda')\in]2p-\frac{1}{2}, 2p+\frac{1}{2}[$for some integer p, $b_+(Q_{\lambda'})$ is much easier to calculate than $b_+(G_{\lambda'})$ and the same is true for b_-'s. The main theorem in this paper is the following:

Theorem 1: *Let $\lambda'=(\lambda_1',\ldots,\lambda_{N-1}')$ be as above. Then:*

$$b_+(Q_{\lambda'})=b_+(G_{\lambda'}),b_-(Q_{\lambda'})=b_-(G_{\lambda'}),b_0(Q_{\lambda'})=b_0(G_{\lambda'}).$$

This theorem implies Zuber's conjecture about the signature of intersection forms associated with affine algebras of type A (cf. Ref. 1) which is the motivation of this paper. Note that $b_0(Q_{\lambda'})=b_0(G_{\lambda'})$ is already noticed in a slightly different form in Ref. 1.

Zuber's conjecture appeared as Conjecture 2.5 of Ref. 1. It is based on the mysterious connections between integrable models with two supersymmetries $(N=2)$ in two dimensions (cf. Ref.

[a)]Electronic mail: xufeng@math.ou.edu

2) and the class of graphs constructed in Ref. 1 (see also Ref. 3). In the special case when the graphs are regular (cf. Sec. II C), the conjecture can be proved (cf. p. 14 of Ref. 4) by combining the results of Ref. 4 and Ref. 5. In fact, in Ref. 4, a connection between regular graphs and singularity theory is established, and combined with Ref. 5 which is based on mixed Hodge structures, gives a rather indirect proof of Zuber's conjecture in the special case when the graphs are regular. However, for other graphs in Ref. 1 (see also Ref. 6), the connection with singularity theory, if any, is not clear at all. We also failed to prove Zuber's conjecture by using the connection with Ref. 2 as mentioned in Ref. 1.

We came to the realization that a statement as in theorem 1 may be true by first observing lemma 1 (cf. Sec. II A) which was already noticed in a slightly different form in Ref. 1. We then checked that theorem 1 is true explicitly in the case when $N=3$, 4 and some other cases which motivated us to give a general proof.

The idea of the proof of theorem 1 is as follows. When λ' changes, $b_+(G_{\lambda'})$ [respectively, $b_-(G_{\lambda'})$] may change only if some of $g_i(\lambda')$'s become 0 or change its sign, i.e., $g_i(\lambda')$ intersect the hyperplanes on which $g_i(\lambda')=0$. By lemma 1 of Sec. II A, these hyperplanes are the same as the hyperplanes on which some $q_j(\lambda')$'s lie in $\mathbb{Z}+\frac{1}{2}$. Consider the domain $D:=\{\lambda'=(\lambda'_1,\ldots,\lambda'_{N-1})|0<\lambda'_i<1,\Sigma_{0<i<N}\lambda'_i<1\}$. D is separated by the above hyperplanes into disjoint open regions. In each open region, the set of numbers compared in theorem 1 should be completely determined. In Sec. II A we determine these numbers in a given open region and find that they miraculously satisfy theorem 1. In Sec. II B, we show that theorem 1 also holds for any $\lambda'\in D$ which is on the boundary of the open region: this follows from Sec. II A and lemma 1. In Sec. II C, after introducing Zuber's conjecture, we show how theorem 1 implies that the conjecture is true. In Sec. III, we present our conclusions and questions.

II. THE PROOF

A. The interior case

We shall use the notations of Sec. I. Recall

$$D=\left\{\lambda'=(\lambda'_1,\ldots,\lambda'_{N-1})|\lambda'_i>0,\quad \sum_{0<i<N}\lambda'_i<1\right\}.$$

For $\lambda'\in D$, recall

$$p_i(\lambda')=\lambda'_i+\cdots+\lambda'_{N-1}-\frac{1}{N}\sum_{0<j<N}j\lambda'_j$$

$$=\frac{1}{N}[(-\lambda'_1-2\lambda'_2-\cdots-(i-1)\lambda'_{i-1})+(N-i)\lambda'_i+\cdots+\lambda'_{N-1}]$$

for $i=1,2,\ldots,N-1$. We have:

$$p_i(\lambda')>\frac{1}{N}[(-\lambda'_1-2\lambda'_2-\cdots-(i-1)\lambda'_{i-1})]>-\frac{1}{N}(i-1)\sum_{0<j<N}\lambda'_j>-\frac{1}{N}(i-1)$$

and

$$p_i(\lambda')<\frac{1}{N}[(N-i)\lambda'_i+\cdots+\lambda'_{N-1}]<\frac{1}{N}(N-i)\sum_{0<j<N}\lambda'_j<\frac{1}{N}(N-i).$$

Similarly one can show $(1/N)(1-N)<p_N(\lambda')<0$. So we have:

$$\frac{1}{N}(1-i)<p_i(\lambda')<\frac{1}{N}(N-i),$$

$i=1,2,...,N$. If for some i, $q_i(\lambda')=-Np_i(\lambda')+\frac{1}{2}(N+1)-i=j+\frac{1}{2}$ with $j\in\mathbb{Z}$, then

$$i-\tfrac{1}{2}N<i+j<i+\tfrac{1}{2}N-1.$$

Define $0<r_{\lambda'}(i)<N+1$ to be the unique integer such that $(1/N)(r_{\lambda'}(i)+j+i)\in\mathbb{Z}$. In fact, if $j+i<0$, $r_{\lambda'}(i)=-(j+i)$, if $0\leqslant j+i<N$, $r_{\lambda'}(i)=N-(j+i)$, and if $N\leqslant(j+i)<\frac{3}{2}N$, $r_{\lambda'}(i)=2N-(j+i)$. It follows that

$$g_{r_{\lambda'}}(i)(\lambda')=0.$$

We have the following lemma which, in a slightly different form also appeared on p. 17 of Ref. 1.

Lemma 1: For any $\lambda'\in D$, *the map* $i\to r_{\lambda'}(i)$ *defined above from* $\{i|q_i(\lambda')\in\mathbb{Z}+\frac{1}{2}\}$ *to* $\{r|g_r(\lambda')=0\}$ *is a one to one and onto map. Moreover* $q_i(\lambda')\in\mathbb{Z}+\frac{1}{2}$ *iff* $g_{r_{\lambda'}}(i)(\lambda')=0$ *and* $r_{\lambda'}(i)$ *depends only on i and* $q_i(\lambda')$.

Proof: By using definitions we have

$$g_r(\lambda')=(-1)^r\prod_{i=1}^{N}2\sin\left(\frac{\pi}{N}\left(q_i(\lambda')+i+r-\frac{1}{2}\right)\right).$$

If $g_r(\lambda')=0$, then there exists $i:=i_{\lambda'}(r)$ such that

$$\frac{1}{N}\left(q_i(\lambda')+i+r-\frac{1}{2}\right)\in\mathbb{Z}.$$

Let $j\in\mathbb{Z}$ with $q_i(\lambda')+i+r-\frac{1}{2}=j+i+r$, then $q_i(\lambda')=-Np_i(\lambda')+\frac{1}{2}(N+1)-i\in\mathbb{Z}+\frac{1}{2}$. Using the fact that $0<|p_a-p_b|<1$ for any $1\leqslant a\neq b\leqslant N$, it is easy to see that such an $i:=i_{\lambda'}(r)$ is also unique. It is then easy to check that the map $i\to r_{\lambda'}(i)$ and $r\to i_{\lambda'}(r)$ are inverse to each other. The rest of the lemma follows from the definitions of $r_{\lambda'}(i)$.

Let $\beta_i:=\frac{1}{2}N-i+\gamma_i$ with $\gamma_i\in\mathbb{Z}$, $i=1,2,...,N$. Let $D_\gamma:=\{\lambda'\in D|\beta_i<Np_i(\lambda')<\beta_i+1, i=1,2,...,N\}$. D_γ will be called open regions. It is clear that if $\gamma\neq\gamma'$, then $D_\gamma\cap D_{\gamma'}=0$. Notice that $q_i(\lambda')\in\mathbb{Z}+\frac{1}{2}$ iff λ' lies on the boundary of some D_γ, and by lemma 1, $g_{r_{\lambda'}(i)}(\lambda')=0$ iff λ' lies on the boundary of some D_γ. Suppose $D_\gamma\neq\varnothing$ and $\lambda'\in D_\gamma$. Then we have:

(1) If $i<j$, then $\beta_i\geqslant\beta_j$ which follows from the fact that $i<j$, then $p_i(\lambda')>p_j(\lambda')$ and $\beta_i-\beta_j\in\mathbb{Z}$;

(2) $\beta_1-\beta_N\leqslant N$ which follows from $p_1(\lambda')-p_N(\lambda')<1$.

By (1) we can assume that

$$\beta_1=\cdots=\beta_{i_1}>\beta_{i_1+1}=\cdots=\beta_{i_2}>\cdots>\beta_{i_{t-1}+1}=\ldots=\beta_{i_t},$$

where $1\leqslant i_1<i_2<\cdots<i_t=N$. We determine the sign of $g_r(\lambda')$ for a fixed $1\leqslant r\leqslant N$. Since $0<p_1(\lambda')-p_N(\lambda')<1$, there is at most one $k+\frac{1}{2}$ with $k\in\mathbb{Z}$ such that

$$p_N(\lambda')-\frac{r}{N}<k+\frac{1}{2}<p_1(\lambda')-\frac{r}{N}.$$

Also notice that if

$$\frac{\beta_i-r}{N}<k+\frac{1}{2}<\frac{\beta_i-r+1}{N},$$

then we have

$$\gamma_i-i-Nk<r<\gamma_i-i-Nk+1,$$

which is impossible since r, γ_i are integers. So if there is a $k\in\mathbb{Z}$ such that

$$p_N(\lambda')-\frac{r}{N}<k+\frac{1}{2}<p_1(\lambda')-\frac{r}{N},$$

then there is a unique integer, denoted by $1\leqslant f(r)\leqslant t-1$, such that:

$$\frac{\beta_{i_{f(r)+1}}+1-r}{N}\leqslant k+\frac{1}{2}\leqslant\frac{\beta_{i_{f(r)}}-r}{N},$$

and the sign of $g_r(\lambda')$ is:

$$(-1)^r(-1)^{kk'}(-1)^{(k-1)(N-k')}=(-1)^{r+Nk-i_{f(r)}},$$

where $k':=\{\beta_j:\beta_j<\beta_{f(r)}\}^{\#}=N-i_{f(r)}$. If there is no $k_1\in\mathbb{Z}$ such that

$$p_N(\lambda')-\frac{r}{N}<k_1+\frac{1}{2}<p_1(\lambda')-\frac{r}{N},$$

then there is a $k\in\mathbb{Z}$ such that:

$$k-\frac{1}{2}\leqslant\frac{\beta_N-r}{N}<\frac{\beta_1+1-r}{N}\leqslant k+\frac{1}{2},$$

and the sign of $g_r(\lambda')$ is:

$$(-1)^{r+kN}.$$

We define $f(r)=t$ in this case. Let $s:=r+kN$, then the signs of the set $\{g_r(\lambda')\}$ with $1\leqslant f(r)\leqslant t-1$ are given by

$$(-1)^{s-i_{f(r)}}$$

with $\gamma_{i_{f(r)+1}}+1-i_{f(r)+1}\leqslant s\leqslant\gamma_{i_{f(r)}}-i_{f(r)}$, and the sign of the set $\{g_r(\lambda')\}$ with $f(r)=t$ is given by

$$(-1)^{s-i_t}$$

with $\gamma_{i_1}+1-i_1-N\leqslant s\leqslant\gamma_{i_t}-i_t$. Now we determine the sign of $\cos(\pi q_i(\lambda'))$. Recall $\beta_i=(N/2)-i+\gamma_i$, $q_i(\lambda')=-Np_i(\lambda')+(N+1)/2-i$, and $\beta_i<Np_i(\lambda')<\beta_i+1$, we have

$$-\gamma_i-\tfrac{1}{2}<q_i(\lambda')<-\gamma_i+\tfrac{1}{2}.$$

So $\cos(\pi q_i(\lambda'))>0$ [respectively, $\cos(\pi q_i(\lambda'))<0$] iff $\gamma_i\in 2\mathbb{Z}$ (respectively, $\gamma_i\in 2\mathbb{Z}+1$). Recall from the introduction we have that for a finite sequence S of real numbers $a(S)=b_+(S)-b_-(S)$. To save some writing for any integer x we define $\{x\}:=[1-(-1)^x]/2$. Then the a of the following sequence $\{\cos(\pi q_i(\lambda')),i_{u-1}+1\leqslant i\leqslant i_u\}$ is

$$(-1)^{\gamma_{i_u}}\{i_u-i_{u-1}\}$$

and the a of the following sequence $\{(-1)^{s-i_u},\gamma_{i_{u+1}}+1-i_{u+1}\leqslant s\leqslant\gamma_{i_u}-i_u\}$ is

$$(-1)^{\gamma_{i_u}}\{\gamma_{i_u}-\gamma_{i_{u+1}}+i_{u+1}-i_u\},$$

where we define $i_{u-1}=0$ if $u=1$, and $\gamma_{i_{u+1}}-i_{u+1}=\gamma_{i_1}-i_1-N$ if $u=t$. It follows that $a(G_{\lambda'})-a(Q_{\lambda'})$ is given by

$$\sum_{u=1}^{t}(-1)^{\gamma_{i_u}}(\{\gamma_{i_u}-\gamma_{i_{u+1}}+i_{u+1}-i_u\}-\{i_u-i_{u-1}\})$$
$$=(-1)^{\gamma_{i_1}}(\{\gamma_{i_1}-\gamma_{i_2}+i_2-i_1\}-\{i_1\})+(-1)^{\gamma_{i_2}}(\{\gamma_{i_2}-\gamma_{i_3}+i_3-i_2\}-\{i_2-i_1\})$$
$$+\ldots+(-1)^{\gamma_{i_{t-1}}}(\{\gamma_{i_{t-1}}-\gamma_{i_t}+i_t-i_{t-1}\}-\{i_{t-1}-i_{t-2}\})$$
$$+(-1)^{\gamma_{i_t}}(\{\gamma_{i_t}-\gamma_{i_t}+i_1\}-\{i_t-i_{t-1}\}).$$

By using

$$\{\gamma_{i_u}-\gamma_{i_{u+1}}+i_{u+1}-i_u\}=\{\gamma_{i_u}-\gamma_{i_{u+1}}\}+(-1)^{\gamma_{i_u}-\gamma_{i_{u+1}}}\{i_{u+1}-i_u\},$$

which follows easily from the definition of $\{.\}$, we see that $\pm\{i_{u+1}-i_u\}$ terms canceled each other in the above summation and the remaining terms are:

$$(-1)^{\gamma_{i_1}}\{\gamma_{i_1}-\gamma_{i_2}\}+(-1)^{\gamma_{i_2}}\{\gamma_{i_2}-\gamma_{i_3}\}+\cdots+(-1)^{\gamma_{i_t}}\{\gamma_{i_t}-\gamma_{i_1}\}$$

which is also 0 since $\{x\}=[1-(-1)^x]/2$. So we have shown that

$$a(G_{\lambda'})-a(Q_{\lambda'})=0,$$

i.e.,

$$b_+(G_{\lambda'})-b_-(G_{\lambda'})=b_+(Q_{\lambda'})-b_-(Q_{\lambda'}).$$

Since

$$b_+(G_{\lambda'})+b_-(G_{\lambda'})=N=b_+(Q_{\lambda'})+b_-(Q_{\lambda'}),$$

it follows that theorem 1 is true for $\lambda'\in D_\gamma$.

B. The boundary case

Assume $\lambda'\in D$ and λ' is on the boundary of some D_γ. Assume $\{q_i(\lambda')|q_i(\lambda')\in\mathbb{Z}+\frac{1}{2}\}=\{q_{k_1}(\lambda'),\ldots,q_{k_s}(\lambda'),s\geqslant 1\}$. Let $k_i\rightarrow r_{\lambda'}(k_i)$ be as in lemma 1. We can choose a small neighborhood W of λ' such that for any $\mu\in W$ and $l\neq k_i$, $i=1,\ldots,s$ [respectively, $m\neq r(k_i)$, $i=1,\ldots,s$], $\cos(\pi q_l(\mu))$ [respectively $g_m(\mu)$] has the same sign as $\cos(\pi q_l(\lambda'))$ [respectively $g_m(\lambda')$] since $\cos(\pi q_l(\lambda'))$ [respectively $g_m(\lambda')$] is not zero. Let $\mu_1\in D_\gamma\cap W$. We compare $b_+(G_{\lambda'})$ [respectively $b_+(Q_{\lambda'})$] with $b_+(G_{\mu_1})$ [respectively, $b_+(Q_{\mu_1})$]. Since $b_+(G_{\mu_1})=b_+(Q_{\mu_1})$ by Sec. II A, to prove $b_+(G_{\lambda'})=b_+(Q_{\lambda'})$ we just have to show that if $\cos(\pi q_{k_i}(\mu_1))>0$ for some k_i, $1\leqslant i\leqslant s$, then $g_{r_{\lambda'}(k_i)}(\mu_1)>0$ and vice versa. Let us consider a small line segment with end points μ_1, μ_2 which passes from D_γ to its neighbor $D_{\gamma'}$, intersects the hyperplane $q_{k_i}(\mu)=q_{k_i}(\lambda')$ at μ_0, and does not intersect any other hyperplanes. Then we have $\cos(\pi q_{k_i}(\mu_0))=0$, so by lemma 1, $g_{r_{\mu_0}(k_i)}(\mu_0)=0$. Again by lemma 1, $r_{\mu_0}(k_i)$ depends only on k_i and $q_{k_i}(\mu_0)=q_{k_i}(\lambda')$, so $r_{\mu_0}(k_i)=r_{\lambda'}(k_i)$. As μ goes from μ_1 to μ_2 on the above line segment, $\cos(\pi q_{k_i}(\mu))$, $g_{r_{\lambda'}(k_i)}(\mu)$ change their signs while the signs of all other $\cos(\pi q_i(\mu))$, $g_j(\mu)$'s do not change. By Sec. II A, $b_+(Q_{\mu_l})=b_+(Q_{\mu_l})$, $l=1, 2$, it follows that if $\cos(\pi q_{k_i}(\mu_1))>0$ for some k_i, $1\leqslant i\leqslant s$, then $g_{r_{\lambda'}(k_i)}(\mu_1)>0$ and vice versa. So we have proved that $b_+(G_{\lambda'})=b_+(Q_{\lambda'})$, and since $b_0(G_{\lambda'})=b_0(Q_{\lambda'})$ by lemma 1, and both $G_{\lambda'}$ and $Q_{\lambda'}$ have N elements, theorem 1 is proved for $\lambda'\in D$ which lies on the boundary of some D_γ. By Sec.II A and II B theorem 1 is proved.

C. Zuber's conjecture

To describe Zuber's conjecture, we have to introduce some notations from Ref. 1 to which the reader is referred for more details.

Let $\Lambda_1,\dots,\Lambda_{N-1}$ be the fundamental weights of $\mathrm{SL}(N)$. Let $k\in\mathbb{N}$. Recall that the set of integrable weights of the affine algebra $\mathrm{SL}(N)$ at level k is the following subset of the weight lattice of $\mathrm{SL}(N)$:

$$P_{++}^{(h)}=\{\lambda=\lambda_1\Lambda_1+\cdots+\lambda_{N-1}\Lambda_{N-1}|\lambda_i\in\mathbb{N},\lambda_1+\cdots+\lambda_{N-1}<h\},$$

where $h=k+N$. This set admits a $\mathbb{Z}_N$ automorphism generated by

$$\sigma:\lambda=(\lambda_1,\lambda_2,\dots,\lambda_{N-1})\rightarrow\sigma(\lambda)=\left(h-\sum_{j=1}^{N-1}\lambda_j,\lambda_1,\dots,\lambda_{N-2}\right).$$

We then introduce the weights e_i of the standard N-dimensional representation of $\mathrm{SL}(N)$

$$e_1=\Lambda_1,\quad e_i=\Lambda_i-\Lambda_{i-1},\quad i=2,\dots,N-1,\quad e_N=-\lambda_{N-1}$$

endowed with the scalar product $(e_i,e_j)=\delta_{ij}-1/N$. We shall be concerned with type II class of graphs introduced in Sec. 1 of Ref 1. These graphs generalize the classical *A, D, E* Dynkin diagrams which may be regarded as related to the SL(2) algebra. The axioms on these graphs are given in Sec. 1.2 of Ref. 1 as follows.

(1) A set ν of $|\nu|=n$ vertices is given. These vertices are denoted by Latin letters a, b,... . There exists an involution $a\rightarrow a$ and the set ν admits a $\mathbb{Z}_n$ grading denoted by $\tau(a)$ such that $\tau(\overline{a})=-\tau(a)\bmod N$.

(2) A set of $N-1$ commuting $n\times n$ matrices G_p, $p=1,2,\dots,N-1$ is given. Their matrix elements are assumed to be non-negative integers, so they may be regarded as adjacency matrices of $N-1$ graphs g_p. g_1 is also assumed to be connected.

(3) The edges of the graphs g_p are compatible with the grading τ in the sense that $(G_p)_{ab}=0$ if $\tau(b)\neq\tau(a)+p\bmod N$.

(4) The matrices are transposed of one another $G_p^t=G_{N-p}$ and $(G_p)_{ab}=(G_p)_{\overline{b}\overline{a}}$.

(5) As a consequence of axions (2) and (4), the matrices G_p are commuting normal matrices and may thus be simultaneously diagonalized in a common orthonormal basis. This basis, denoted by $\psi^{(\lambda,i)}$, is assumed to be labeled by the weights λ of $\mathrm{SL}(N)$, that are restricted to $P_{++}^{(h)}$, for some integer $h>N$, in a way that the eigenvalues $\gamma_p^{(\lambda)}$ have the form $\gamma_p^{(\lambda)}=\chi_p(M(\lambda))$, where χ_p is the ordinary character for the pth fundamental representation of the group $\mathrm{SU}(N)$, and $M(\lambda)$ denotes the diagonal matrix $M(\lambda)=\mathrm{diag}(\epsilon_j(\lambda))_{j=1,\dots,N}$. Here $\epsilon_j(\lambda):=\exp(-(2\pi i/h)(e_j,\lambda)$, and i in (λ,i) is an index integer, $1\leqslant i\leqslant m_\lambda$ with m_λ being the multiplicity of eigenvalue $\gamma_p^{(\lambda)}$. The set of (λ,i)'s will be denoted by Exp.

There exists a special class of solutions known for all N and $h>N$, namely the fusion graphs of the affine algebra $\mathrm{SL}(N)$ at level $k=h-N$. The vertices are the integrable weights described above, i.e., $\nu=P_{++}^{(h)}$. The matrices G_p are the Verlinde matrices, which describe the fusion by the pth fundamental representation. The fusion rules are given on p. 288 of Ref. 7. Their diagonalization is known, thanks to the Verlinde formula (cf. p. 288 of Ref. 7), and the eigenvalues are the $\gamma_p^{(\lambda)}$, where λ takes all the values in$=P_{++}^{(h)}$. We will call these graphs *regular graphs* in this paper. In the case of $N=2$, these regular graphs reduce to the A_{h-1} Dynkin diagrams.

More solutions are known (cf. Ref. 1). In Ref. 6 (in particular Theorem 3.10 and (5) of Theorem 3.8), infinite series of such graphs are constructed from the maximal conformal inclusions of the form $\mathrm{SU}(N)\subset G$ with G being a simple and simply connected compact Lie group.

Given graphs of the previous type, let V be a complex vector space with a basis α_a labeled by the vertices of the set ν. A bilinear form g is defined by:

$$g_{ab}=\langle\alpha_a,\alpha_b\rangle=2\delta_{ab}+G_{ab}$$

in terms of the matrix $G_{ab}=\Sigma_{p=1}^{N-1}(G_p)_{ab}$. g will be called the intersection form. This is the intersection form in the title of this paper.

The eigenvalue of the matrix (g_{ab}) with eigenvector $\psi_p^{(\lambda,i)}$ is (cf. (34) of Ref. 1):

$$g^{(\lambda)}=\prod_{i=1}^{N}\left(1+\exp\left(-\frac{2\pi i}{h}(e_i,\lambda)\right)\right).$$

For $(\lambda,i)\in\mathrm{Exp}$ define real numbers which depend only on λ (cf. (46) of Ref. 1) by:

$$q_\lambda^{(R)}:=\frac{1}{h}\sum_{j=1}^{N-1} j(\lambda_j-1)+\frac{(N-h)(N-1)}{2h}.$$

We can now state Zuber's conjecture on the signature of g (cf. Conjecture 2.5 of Ref. 1):

Zuber's Conjecture. The signature of the bilinear form g for class II graphs is $(x+,y-,z0)$ *where x is the number of* $q_\lambda^{(R)}$ *which fall in an interval* $]2p-\frac{1}{2},2p+\frac{1}{2}[$ *for some* $p\in\mathbb{Z}$ *(p may depend on* $q_\lambda^{(R)}$*,) y is the number of those in an interval* $]2p'+\frac{1}{2},2p'+\frac{3}{2}[$ *for some* $p'\in\mathbb{Z}$ *(p′* may depend on $q_\lambda^{(R)}$*, and* $t=n-r-s$ *is the number of those* $q_\lambda^{(R)}$ *which are half-integers.*

We now prove this conjecture.

Let us first notice a simple consequence of the axioms on the graphs. It follows from Proposition 1.2 of Ref. 1 that Exp is invariant under the action of σ. In fact, if $\Sigma_a\psi_a^{(\lambda,i)}a$ is an eigenvector of G_p with eigenvalue $\gamma_p^{(\lambda)}$, then Proposition 1.2 of Ref. 1 implies that $\Sigma_a\psi_a^{(\lambda,i)}\exp(2\pi i\tau(a)/N)a$ is an eigenvector of G_p with eigenvalue $\gamma_p^{(\sigma(\lambda))}$. Since $a\to\exp(2\pi i\tau(a)/N)a$ is an invertible map, it follows that the multiplicity of eigenvalue $\gamma_p^{(\sigma(\lambda))}$ is the same as that of eigenvalue $\gamma_p^{(\lambda)}$. We can therefore define $\sigma(\lambda,i)=(\sigma(\lambda),i)$. It follows that Exp can be written as a disjoint union of the orbits under the action of σ. To prove Zuber's conjecture, we just have to show it is true on each orbit.

Let $(\lambda,i)\in\mathrm{Exp}$ and let d be the smallest positive integer such that $\sigma^d(\lambda)=\lambda$. Then d/N and let $N=dd_1$. Let $G'_\lambda:=\{g^{(\sigma^i(\mu))},i=1,2,...,d\}$, $Q'_\lambda:=\{\cos(\pi q^{(R)}_{\sigma^i(\mu)}),i=1,2,...,d\}$, we need to show

$$b_+(G'_\lambda)=b_+(Q'_\lambda),b_0(G'_\lambda)=b_0(Q'_\lambda),b_-(G'_\lambda)=b_-(Q'_\lambda).$$

Note that $b_0(G'_\lambda)=b_0(Q'_\lambda)$ was already noticed on p. 17 of Ref. 1. Let $\lambda'=(\lambda_1/h,\ldots,\lambda_{N-1}/h)$, then $p_i(\lambda')=(e_i,\lambda)/h$, $i=1,2,...,N$. To use theorem 1, we make use of the following identities which follow from the definitions:

$$q^{(R)}_{\sigma^{-j}(\lambda)}=q_j(\lambda')$$

and

$$\begin{aligned}g^{(\sigma^j(\lambda))}&=\prod_{l=1}^{N}\left(1+\exp\left(+\frac{2\pi ij}{N}\epsilon_l(\lambda)\right)\right)\\&=\prod_{l=1}^{N}2\cos\left(\pi\left(p_l(\lambda')-\frac{j}{N}\right)\right)\times\exp\left(\pi ij-\sum_{l=1}^{N}\frac{1}{h}(\lambda,e_l)\right)\\&=(-1)^j\prod_{l=1}^{N}2\cos\left(\pi\left(p_l(\lambda')-\frac{j}{N}\right)\right)\\&=g_j(\lambda').\end{aligned}$$

Now it is clear that the size of $G_{\lambda'}$ (respectively, $Q_{\lambda'}$) is d_1 times the size of G'_λ (respectively, Q'_λ) and we have:

$$b_+(G_{\lambda'}) = d_1 b_+(G'_\lambda), \quad b_0(G_{\lambda'}) = d_1 b_0(G'_\lambda), \quad b_-(G_{\lambda'}) = d_1 b_-(G'_\lambda)$$

and

$$b_+(Q_{\lambda'}) = d_1 b_+(Q'_\lambda), \quad b_0(Q_{\lambda'}) = d_1 b_0(Q'_\lambda), \quad b_-(Q_{\lambda'}) = d_1 b_-(Q'_\lambda).$$

By theorem 1, we have proved:

$$b_+(G'_\lambda) = b_+(Q'_\lambda), \quad b_0(G'_\lambda) = b_0(Q'_\lambda), \quad b_-(G'_\lambda) = b_-(Q'_\lambda).$$

Let us summarize the result in the following:

Corollary 1: Zuber's Conjecture as stated above is true.

III. CONCLUSIONS AND QUESTIONS

In this paper we proved Zuber's conjecture on the signature of certain intersection forms by using theorem 1.

Our results imply that the infinite series of graphs which are constructed in Ref. 6 by using subfactors associated with conformal inclusions satisfy Zuber's conjecture. This lends further support to the idea that these graphs may be associated with the integrable models in Ref. 2 which is the basis of Zuber's conjecture. Such a relation is not very clear and should be very interesting.

ACKNOWLEDGMENT

This work is partially supported by NSF Grant No. DMS-9500882

[1] J.-B. Zuber, *Generalized Dynkin diagrams and root systems and their folding*, hep-th 9707046, to appear in the proceedings of the Taniguchi Symposium *Topological Field theory, Primitive forms and Related Topics*, Kyoto, December 1996, edited by M. Kashiwara, A. Matsuo, K. Saito, and I. Satake.
[2] S. Cecotti and C. Vafa, Nucl. Phys. **A367**, 359–461 (1991); Commun. Math. Phys. **158**, 569–644 (1993).
[3] J.-B. Zuber, Commun. Math. Phys. **179**, 265–294 (1996).
[4] S. M. Gusein-Zade and A. Varchenko, Selecta Math. (N.S.) **3**, 79–97 (1997).
[5] J. Steenbrink, Compositio Mathematica **34**, 211–223 (1977).
[6] F. Xu, Commun. Math. Phys. **192**, 349–403 (1998).
[7] V. Kac, *Infinite Dimensional Lie Algebras* (Cambridge University Press, Cambridge).

Erratum: ''Augmented temporal logic formalism for histories-based generalized quantum mechanics'' [J. Math. Phys. 39, 704 (1998)]

Tulsi Dass and Yogesh Joglekar
Department of Physics, Indian Institute of Technology, Kanpur 208016, India

(Received 3 December 1998; accepted for publication 14 December 1998)

On p. 709, line no. 19, replace ''all the psg's introduced above'' by ''$\mathcal{K}_1$ and $\mathcal{K}_3$.''

Quantum field theory of partitions

Carl M. Bender
Department of Physics, Washington University, St. Louis, Missouri 63130

Dorje C. Brody
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge CB3 9EW, United Kingdom

Bernhard K. Meister
Goldman Sachs, ARK Mori Building, 12-32-1 Akasaka. Minato-ku, Tokyo 107, Japan

(Received 18 January 1999; accepted for publication 17 February 1999)

Given a sequence of numbers $\{a_n\}$, it is always possible to find a set of Feynman rules that reproduce that sequence. For the special case of the partitions of the integers, the appropriate Feynman rules give rise to graphs that represent the partitions in a clear pictorial fashion. These Feynman rules can be used to generate the Bell numbers $B(n)$ and the Stirling numbers $S(n,k)$ that are associated with the partitions of the integers. © *1999 American Institute of Physics.*
[S0022-2488(99)03207-7]

I. INTRODUCTION

The *partition* of an integer n is the set of all distinct ways to represent n as a sum of positive integers smaller than or equal to n. The number of elements in the partition of n is designated $p(n)$. For example, since 5 can be represented as 1+1+1+1+1, 1+1+1+2, 1+1+3, 1+2+2, 1+4, 2+3, or simply 5, we have $p(5)=7$. Defining $p(0)=1$, we obtain a well-known formula for the generating function of the numbers $p(n)$.[1]

$$\sum_{n=0}^{\infty} p(n)x^n=\prod_{k=1}^{\infty} \frac{1}{1-x^k}. \tag{1}$$

There is a simple graphical way to represent the partitions because there is a one-to-one correspondence between the nth partition and the set of transitive nondirected n-vertex graphs. A graph is said to be *transitive* if for every path connecting two vertices there is an edge joining the two vertices. In Fig. 1 we display several transitive graphs and several nontransitive graphs.

In Fig. 2 we display the transitive graphs having from one through five vertices; note that the number of such n-vertex graphs equals the number of partitions $p(n)$: $p(1)=1$, $p(2)=2$, $p(3)=3$, $p(4)=5$, $p(5)=7$.

In the study of quantum field theory graphs are used to represent terms in a diagrammatic perturbation expansion. Here, one associates with each graph a numerical amplitude. This amplitude is the product of the symmetry number, the vertex numbers, and a Feynman integral for the graph.[2] In this paper we will work in zero-dimensional space only so that Feynman integrals become trivial and are merely the product of the line amplitudes.

To illustrate the procedure of assigning amplitudes to graphs, let us consider the graphs in Fig. 2. We associate symmetry numbers with each of the graphs shown in Fig. 2, take the vertex amplitudes and line amplitudes to be unity, and then sum over all graphs having n vertices. We obtain the result $B(n)/n!$, where $B(n)$ is the nth Bell number.[3] The *Bell numbers* $B(0)$, $B(1)$, $B(2)$, $B(3)$,... are a sequence of positive integers that begins 1, 1, 2, 5, 15, 52, 203, 877,... . The Bell number $B(n)$ is the number of *labeled* partitions of the positive integer n, which can be seen by labeling all the vertices in Fig. 2 and counting the number of transitive nondirected graphs. There is a simple generating function $G(x)$ for the Bell numbers:

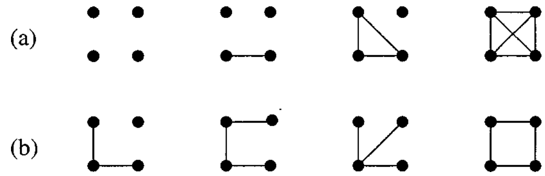

FIG. 1. Examples of (a) four transitive graphs and (b) four nontransitive graphs, each having four vertices.

$$G(x)=\exp(e^{x}-1)=\sum_{n=0}^{\infty}\frac{B(n)}{n!}x^{n}. \tag{2}$$

From this representation one can derive a representation for $B(n)$ in terms of a sum:

$$B(n)=\frac{1}{e}\sum_{k=0}^{\infty}\frac{k^{n}}{k!}. \tag{3}$$

In quantum field theory there is a simple and general formal construction for the set of all graphs in terms of the exponential of a derivative operator.[4] For example,

$$\exp\left(\frac{a}{2!}\frac{d^{2}}{dx^{2}}\right)\exp\left(\frac{g}{4!}x^{4}\right)\Bigg|_{x=0} \tag{4}$$

is the generating function for the set of all four-vertex vacuum diagrams, connected and disconnected, in which the line amplitude is a and the vertex amplitude is g. If we expand this expression as a series in powers of g, the coefficient of g^{n} is the sum over all four-vertex vacuum graphs containing n vertices with each graph weighted by its symmetry number.

We can generalize Eq. (4) to include n-point vertices and we can even extend these ideas to include generalized lines having m legs.[5] The formula

$$\exp\left(a\sum_{m=1}^{\infty}\frac{L_{m}}{m!}\frac{d^{m}}{dx^{m}}\right)\exp\left(g\sum_{n=1}^{\infty}\frac{V_{n}}{n!}x^{n}\right)\Bigg|_{x=0} \tag{5}$$

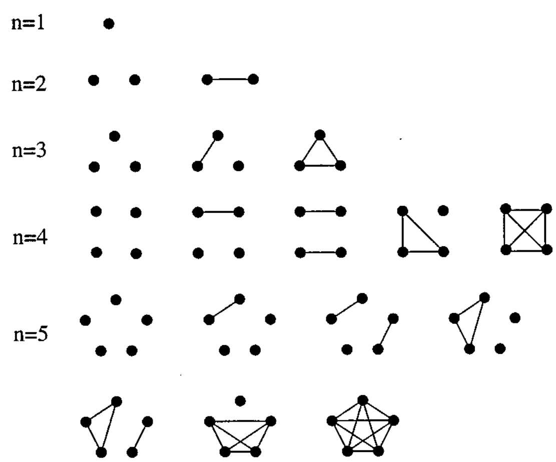

FIG. 2. Transitive nondirected graphs having from $n=1$ to $n=5$ vertices. Note that for each value of n, the number of such graphs is $p(n)$, the number of elements in the partition of the integer n.

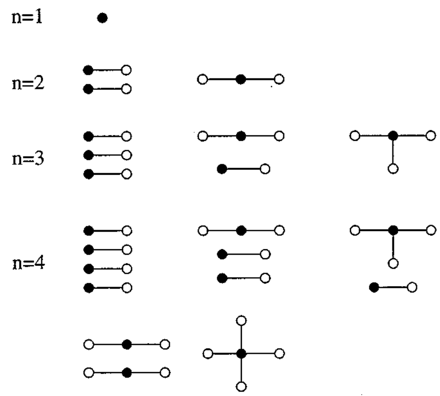

FIG. 3. The graphs in a field theory whose n-point vertex amplitudes are all unity and whose line amplitudes all vanish except for lines having just one end. If the amplitude of such lines is z, then the coefficient of z^n gives all n-line graphs contributing to the partition of the integer n. These graphs correspond exactly to the partition graphs shown in Fig. 2. This correspondence is seen by exchanging lines and vertices.

represents the set of all vacuum diagrams, connected and disconnected, constructed from n-point vertices whose amplitudes are V_n and generalized lines having m legs whose amplitudes are L_m. If we expand this expression as a formal power series in powers of a, then the coefficient of a^n is the sum of the symmetry numbers of all graphs having n lines, and if we expand this expression as a series in powers of g, then the coefficient of g^n is the sum of the symmetry numbers of all graphs having n vertices. If we first take the natural logarithm of the expression in (5) and then expand in formal power series in a or g, then the coefficients represent the sum of the symmetry numbers of just the *connected* graphs.[4] Note that, in general, these formal power series are divergent series because the number of graphs grows like a factorial.

II. PARTITIONS

Given the general graph construction procedure in Eq. (5) one may ask whether it is possible to find a quantum field theoretic set of Feynman rules for which the graphs in its perturbation expansion are the partition graphs in Fig. 2. Let us consider a field theory for which the n-vertex amplitudes and the coupling constant g are all unity. Note that this gives the potential

$$e^x - 1 = \sum_{n=1}^{\infty} \frac{V_n}{n!} x^n, \tag{6}$$

which implies that $V_n = 1$ for all n. Next, let us impose the requirement that there be only one kind of line, a line having just *one* end. We do so by taking $L_m = 0 (m > 1)$. Choosing $aL_1 = z$, we construct from Eq. (5) the n-line vacuum graphs in this theory; these graphs are shown, for the number of lines ranging from one through five, in Fig. 3. Observe that the graphs in Fig. 3 correspond, graph by graph, to the partition graphs in Fig. 2. From Eq. (5) we see that the generating function for these graphs is

$$G(z) = \exp\left(z \frac{d}{dx} \right) \exp(e^x - 1) \bigg|_{x=0}. \tag{7}$$

However, $\exp(zd/dx)$ is just the translation operator. Thus, (7) reduces immediately to

$$G(z) = \exp(e^z - 1). \tag{8}$$

This argument reproduces the generating function $G(x)$ in (2) for the Bell numbers $B(n)$.

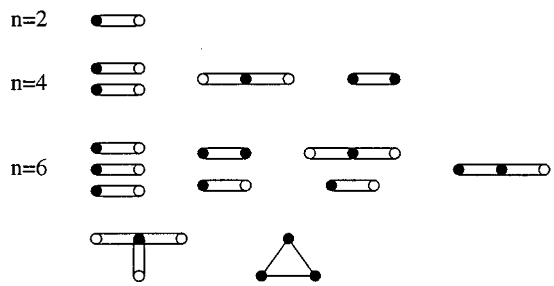

FIG. 4. The graphs in a theory whose Feynman rules allow for any number of even vertices but lines having only two ends. Summing the amplitudes of these graphs gives the generating function in (9).

An alternative way to derive Eq. (8) is to argue that in the field theory described by (7) there is only one connected graph having n lines. This graph consists of the n-point vertex to which n one-ended lines are attached. The symmetry number of this graph is just $1/n!$. Thus, the generating function $C(z)$ of the connected graphs is just $C(z)=\Sigma_{n=1}^{\infty}z^n/n!=e^z-1$. However, the generating function $G(z)$ for the set of all connected and disconnected graphs is just the exponential of the generating function $C(z)$ for the connected graphs. Thus, we have reproduced the result in (8).

Apparently, one can always translate the general problem of finding the nth term in a sequence of numbers to the problem of constructing a set of Feynman rules that produces this sequence. For example, let us ask whether there is a field theory (a set of Feynman rules) that reproduces directly the amplitudes in $C(z)$. This means that we want to find a quantum field theory whose graphs correspond to the connected graphs in Fig. 2. Let us return to Eq. (5) and consider a field theory whose lines have two ends ($L_m=0$ for $m\neq 2$) and which may have all even vertices but no odd vertices. Choosing $L_2=2$ and $g=1$ we find that the expansion of Eq. (5) as a series in powers of a has the form

$$1+aV_2+\tfrac{1}{2}a^2(V_4+3V_2^2)+\tfrac{1}{6}a^3(15V_2^2+15V_2V_4+V_6)$$
$$+\tfrac{1}{24}a^4(105V_2^4+210V_2^2V_4+35V_4^2+28V_2V_6+V_8)+\cdots. \tag{9}$$

The graphs that give rise to this sequence of coefficients are shown in Fig. 4.

If we demand that the coefficient of a^n agree with the coefficient of z^n in $C(z)$, then we obtain a sequence of algebraic equations for the vertex amplitudes V_2: $V_2=1$. $V_4+3V_2^2=1$, $15V_2^3+15V_2V_4+V_6=1$, and so on. The solution to this system of equations is $V_2=1$, $V_4=-2$, $V_6=16$, $V_8=-272$, and so on. These are the tangent numbers t_n:

$$\tanh x=\sum_{n=1}^{\infty}\frac{t_n}{(2n-1)!}x^{2n-1}. \tag{10}$$

To understand why the tangent numbers have appeared, note that upon integrating (10) with respect to x we get

$$\ln(\cosh x)=\sum_{n=1}^{\infty}\frac{t_n}{(2n)!}x^{2n}. \tag{11}$$

Thus, from Eq. (5) we have

$$\exp\left(a\frac{d^2}{dx^2}\right)\exp[\ln(\cosh x)]\bigg|_{x=0}=\sum_{n=0}^{\infty}\frac{a^n}{n!}\frac{d^{2n}}{dx^{2n}}\cosh x\bigg|_{x=0}=\sum_{n=0}^{\infty}\frac{a^n}{n!}=e^a, \tag{12}$$

which is precisely the generating function that we tried to reproduce.

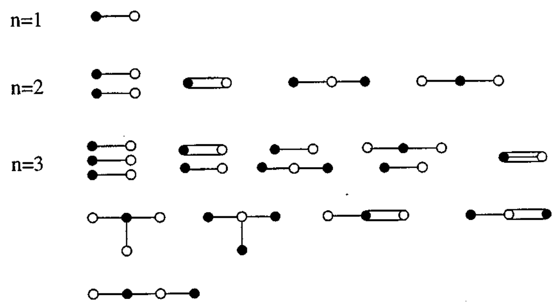

FIG. 5. The graphs in a theory whose Feynman rules allow for n-point vertices ($n=1,2,3,...$) and m-legged lines ($m=1,2,3,...$). If the vertex amplitudes are all unity and the m-legged line amplitude is z^m, then the generating function $G(z)$ for the graphs, as given in (17), has a Taylor expansion for which the coefficient of z^n is $[B(n)]^2/n!$. That is, the number of labeled graphs of order n is the square of the nth Bell number.

III. BELL NUMBERS

We can also examine alternative theories to the theory described in (7). Suppose, for example, we construct a field theory with the same vertices but with a two-ended line instead of a one-ended line,

$$G(z)=\exp\left(\frac{z}{2!}\frac{d^2}{dx^2}\right)\exp(e^x-1)\bigg|_{x=0}, \tag{13}$$

or even a generalized line having M ends,

$$G(z)=\exp\left(\frac{z}{M!}\frac{d^M}{dx^M}\right)\exp(e^x-1)\bigg|_{x=0}. \tag{14}$$

It is easy to show that the coefficient of z^n in the power series expansion of (14) is proportional to the nMth Bell number:

$$G(z)=\sum_{n=0}^{\infty}\frac{1}{n!}z^n(M!)^{-n}B(nM). \tag{15}$$

This line of thinking suggests that one can even construct a field theory whose lines have any number of ends:

$$G(z)=\exp(e^{zd/dx}-1)\exp(e^x-1)|_{x=0} \tag{16}$$

In this case we obtain the beautiful result that the coefficient of z^n is the *square* of the nth Bell number:

$$G(z)=\sum_{n=0}^{\infty}\frac{1}{n!}z^n[B(n)]^2. \tag{17}$$

The graphs contributing to the generating function in (17) are shown in Fig. 5. The numbers of these graphs are 1, 1, 4, 10, 33,..., which apparently is an unknown sequence. However, if we label the vertices for these graphs, then the number of corresponding graphs are given by the squares of the Bell numbers.

IV. STIRLING NUMBERS

The labeled partitions of the integers n can be grouped into classes described by the Stirling numbers $S(n,k)$.[6] Let the n be the number of objects and k be the number of groups. For example, consider the case $n=3$. Given three numbers, x, y, and z, we can group these objects *one* way as a sum of one group of three numbers $(x+y+z)$, *three* ways as a sum of two groups of numbers, $(x+y)+(z)$, $(x+z)+(y)$, and $(y+z)+(x)$, and *one* way as a sum of three numbers $(x)+(y)+(z)$. Thus, we say that $S(3,1)=1$, $S(3,2)=3$, and $S(3,3)=1$. The Stirling numbers are all integers: $S(1,1)=1$, $S(2,1)=S(2,2)=1$, $S(4,1)=1$, $S(4,2)=7$, $S(4,3)=6$, $S(4,4)=1$, $S(5,1)=1$, $S(5,2)=15$, $S(5,3)=25$, $S(5,4)=10$, $S(5,5)=1$, and so on. There is an elementary representation for $S(n,k)$ as a sum:

$$S(n,k) \equiv \sum_{j=1}^{k} \frac{(-1)^{k-j} j^n}{j!(k-j)!}. \tag{18}$$

Recall that the Bell number $B(n)$ represents the total number of labeled partitions of the positive integer n. Hence, summing with respect to k over the Stirling numbers $S(n,k)$ gives $B(n)$:

$$B(n) = \sum_{k=1}^{n} S(n,k). \tag{19}$$

The Stirling numbers emerge nicely in terms of Feynman rules. Let us generalize Eq. (7) slightly so that each vertex has amplitude v instead of 1:

$$G(z,v) = \exp\left(z \frac{d}{dx}\right) \exp[v(e^x - 1)]\bigg|_{x=0}. \tag{20}$$

If we now expand the generating $G(z,v)$ as a series in powers of z and v, then the coefficient of $z^n/n!$ and v^k is precisely the Stirling number $S(n,k)$:

$$G(z,v) = \sum_{n=0}^{\infty} \frac{z^n}{n!} \left(\sum_{k=1}^{n} S(n,k) v^k \right). \tag{21}$$

If we set $v=1$ in this equation and compare with Eq. (2), then we recover Eq. (19).

V. CONCLUSION

In this paper we have illustrated how one can impose purely combinatorial or topological constraints upon the dynamics of a quantum field theory. In particular, we have demonstrated examples of field theories that give rise to the various partitions of integers. If we introduce further structures in the underlying dynamics, such as Fermions (which give rise to directed graphs), then the method we have sketched here would shed insights to some very difficult problems in combinatorics, such as determining the number of partially ordered sets (posets). We hope to study these applications elsewhere.

ACKNOWLEDGMENTS

CMB thanks the Theoretical Physics Group at Imperial College, London, for its hospitality and the PPARC and the Fulbright Association for financial support. DCB thanks PPARC for financial support. DCB and BKM thank the Physics Department at Washington University for its hospitality. This work was supported in part by the U.S. Department of Energy.

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, DC, 1964), p. 825.
[2] J. D. Bjorken and S. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).
[3] G.-C. Rota, Am. Math. Monthly **71**, 498 (1964).
[4] C. M. Bender and W. E. Caswell, J. Math. Phys. **19**, 2579 (1978).
[5] C. M. Bender, F. Cooper, G. S. Guralnik, D. H. Sharp, and M. L. Silverstein, Phys. Rev. D **20**, 1374 (1979).
[6] L. Comtet, *Advanced Combinatorics* (Reidel, Dordrecht, 1974).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Connections and metrics respecting purification of quantum states

J. Dittmann[a)]
Mathematisches Institut, Universität Leipzig, Augustusplatz 10/11, 04109 Leipzig, Germany

A. Uhlmann
Institut fur Theoretische Physik, Universität Leipzig, Leipzig, Germany

(Received 5 January 1999; accepted for publication 5 April 1999)

Standard purification interlaces Hermitian and Riemannian metrics on the space of density operators with metrics and connections on the purifying Hilbert–Schmidt space. We discuss connections and metrics which are well adopted to purification, and present a selected set of relations between them. A connection, as well as a metric on state space, can be obtained from a metric on the purification space. We include a condition, with which this correspondence becomes one to one. Our methods are borrowed from elementary *-representation and fiber space theory. We lift, as an example, solutions of a von Neumann equation, write down holonomy invariants for cyclic ones, and ''add noise'' to a curve of pure states. © *1999 American Institute of Physics.* [S0022-2488(99)02107-6]

I. INTRODUCTION

In Ref. 1, see also Ref. 2, the monotone Hermitian and Riemannian metrics in the (finite dimensional) spaces of all density operators are classified. Based on the theory of operator means[3] they are indexed by a real function, f, operator monotone[4] on $(0,\infty)$. These metrics play an important role in domains like quantum information geometry, quantum versions of statistical estimation, and decision rules.[5–7]

D. Petz communicated his main results to us prior to publication, and about that time we started to ask for the effect of a purifying lift to these metrics. There are clear reasons for this. One of the present authors (A.U.) had defined 1986 in Ref. 8 an extension of the geometric phase,[9,10] see also Refs. 11 and 12, to curves of density operators by the help of a ''parallelity condition.'' The condition singles out, up to a global gauge (or a global partial isometry), a distinguished ''parallel lift'' within all purifying lifts of a curve of density operators. It turns out[13] that a connection form (a gauge potential), here called $\mathbf{a}^{\mathrm{geo}}$, is governing the transport of the purifying vectors, such that the parallelity condition results from the request for horizontality. In 1992 G. Rudolph and one of the authors (J.D.) considered a large class of gauge potentials, including $\mathbf{a}^{\mathrm{geo}}$, which rests on a purification scheme and which enables variants of the geometric phase along curves of density operators. It seems natural to ask for a link between these objects: (a) the connection forms just mentioned, (b) certain Hermitian (Riemannian) metrics on the purification space, and, if respecting the symmetry of the scheme, (c) metrics induced from (b) on the space of density operators.

Purification is essentially representation theory of observables and of the algebra in which they are contained. Principally one may use any unital *-representation of the ''algebra of observables'' over which the states can be defined. Its Hilbert representation space should only be large enough to allow for a representation of the states by vectors. If this condition is fulfilled, transport mechanism, its noncommutative phases, metrics, and other geometric objects can be constructed by relying on their form and appearance in the pure state case.

[a)]Electronic mail: dittmann@mathematik.uni-leipzig.de

In our paper we remain within an elementary setting: Our density operators live on a Hilbert space $\mathcal{H}$ of finite dimension n. In our convention, a density operator should not necessarily be normalized. We speak of ''density operators'' whether their trace is one or not. The algebra of observables is the algebra $\mathcal{B}(\mathcal{H})$ of all operators acting on $\mathcal{H}$. The representation or purification space, $\mathcal{W}$, is identified with the algebra of operators and equipped with the Hilbert–Schmidt scalar product. (In infinite dimensions $\mathcal{W}$ will be the space of Hilbert–Schmidt operators.) We try to emphasize the different meaning of operators by different notations: Operators acting on $\mathcal{H}$ are denoted by lower case letters, those acting on $\mathcal{W}$ often by capital letters. [Some authors call the operators of $\mathcal{B}(\mathcal{W})$ ''superoperators.''] Section II is devoted to explaining our notation in more details. In our paper purification takes place in the standard representation of $\mathcal{B}(\mathcal{H})$, i.e., in the GNS-representation based on the trace. For that reason we called it *standard purification*. In Sec. III the formalism is extended to velocity vectors, i.e., to tangents, at density operators and at their purifications. Purification defines vertical tangents in a canonical way. A tangent, orthogonal to the space of vertical tangents, is called horizontal, provided the tangent spaces carry a real Hilbert space structure, i.e., a Riemannian metric. Equivalently, within all purifying lifts of a given curve of density operators, those with the least length are horizontal.

Section IV exemplifies our task in defining horizontality by the real part of the Hilbert–Schmidt metric. As one knows, the Bures length of a curve of density operators and the Hilbert–Schmidt length of an horizontal lift are equal one to another. In deriving the parallelity condition we meet some peculiarities with tangents of purifying vectors if they belong to density operators with some vanishing eigenvalues. The reader will find a short account of the relation between the connection form $\mathbf{a}^{\text{geo}}$ (Ref. 13) governing the geometric phase, and the Riemannian Bures metric.

Indeed, we devote some time to asking, and giving an affirmative answer to the following question: Is the topological metric of Bures Riemannian?[14–16] Essential differential geometric properties are in Ref. 17, see also Ref. 18 for $\dim \mathcal{H}=3$. Relations to quantum information theory can be seen in Refs. 19 and 20. However, a parameterization in terms of the operators' matrix elements remains cumbersome, except for $\dim \mathcal{H}=2$.

Concerning $\mathbf{a}^{\text{geo}}$, which extends the geometric phase to (closed) curves of density operators, examples are shown in Sec. VIII. There is a further issue to be mentioned: The gauge potential for the two-dimensional density operators[21] living on a four-dimensional purification space, satisfies the Yang–Mills equations. With a certain cosmological constant, it even is a solution of the combined Yang–Mills–Einstein equations.[22] Meanwhile we know[23] $\mathbf{a}^{\text{geo}}$ satisfies the Yang–Mills equations for every finite dimension of the supporting Hilbert space $\mathcal{H}$. These findings may be seen as extensions to mixed states of numerous examples relating the original Berry phase to Dirac monopoles, and the Wilczek and Zee phase[24] to instantons.

Section VI is devoted to the class of connections introduced in Ref. 25, which are, so to say, ''relatives'' of $\mathbf{a}^{\text{geo}}$, compatible with the purification scheme. They are characterized by a function F, defined on $(0,\infty)$, and fulfilling $\overline{F}(1/t)=-F(t)$. Some equations become more appealing by using the function r, the arithmetic mean of $\overline{F}$ and 1. The connection forms $\mathbf{a}$ assign to every tangent x at the lift $w\in\mathcal{W}$ of $\varrho=ww^*$ a value in the Lie algebra of $U(n)$. The action of the gauge group induces the ''canonical'' connection $\mathbf{a}^{\text{can}}$. The canonical connection is gained with the choice $F=0$. The connection $\mathbf{a}^{\text{geo}}$ is constructed with $F(t)=(t-1)/(t+1)$. As we shall see, only connections with real F can be obtained from an appropriate Hermitian metric. We believe the complete class is a more natural object at the complexified tangents. They all decompose as $\theta-\theta^*$ with θ of type (1,0).

We specify the class of Hermitian metrics by another positive and real valued function, k, on the positive half-axis. The metrical form for the tangents at a purifying vector, w, will be given by the inverse of the (''super'')operator $k(\Delta_w)$, where Δ is the field of modular operators. There is an antilinear operator, a modification of Tomita–Takasaki's S_w operator, which admits just the horizontal tangents as fix points. The connection adjusted to the metric is characterized by various relations between the functions k, F, and r. Moreover, every one of the Hermitian metrics considered on the tangent space of $\mathcal{W}$ is a lift of exactly one Hermitian form on the space of density operators. The latter depends on a function f which is related to k. The Riemannian metric on the density operators is gained as the real part of the Hermitian one, and it corresponds to the harmonic mean of $f(t)$ and $tf(1/t)$. Further we discuss an additional condition, which enables us to assign a unique connection form to a given monotone Riemannian state space metric. These metrics are induced from the Hilbert–Schmidt metric by some constraints on the purifying vectors replacing the orthogonality condition of the Bures case.

The starting point has been a set of connections, compatible with the purification procedure, to define reasonable parallel transports along curves of density operators. We return to this issue in purifying horizontally solutions of von Neumann equations. Cyclic solutions give rise to some holonomy invariants. There are constraints on F for extending the parallelity conditions to the boundary, in particular to pure states. If they are fulfilled, the holonomy invariants reduce to the well-known geometric phase of Berry for pure states. At the end we ask what happened if ''noise'' is added to a closed path of pure states.

II. STANDARD PURIFICATION

We start by reviewing some basic ideas of the purification procedure. Let $\mathcal{H}$ be a complex Hilbert space of finite dimension n with scalar product $\langle .,. \rangle$ antilinear in its left argument. $\mathcal{B}(\mathcal{H})$ denotes the $*$-algebra of linear operators acting on $\mathcal{H}$. A *state* is a positive linear form over the algebra which takes the value 1 at the identity of $\mathcal{B}(\mathcal{H})$. Generally, a linear form l over our algebra is uniquely represented by

$$l(b) = \mathrm{Tr}\, b\omega, \quad \forall b \in \mathcal{B}(\mathcal{H}).$$

The linear form is positive if and only if ω is a positive element of $\mathcal{B}(\mathcal{H})$. We then call ω a *density operator* in accordance with its usage in physics. A density operator represents a state iff its trace is one.

A *purification* of a positive linear form over $\mathcal{B}(\mathcal{H})$ is a lift to a pure linear form of a larger algebra. A way to do so is as follows: With another auxiliary Hilbert space $\mathcal{H}^{\mathrm{aux}}$, with at least the same dimension, we consider

$$\mathcal{H} \otimes \mathcal{H}^{\mathrm{aux}}, \quad n = \dim \mathcal{H} \leqslant \mathcal{H}^{\mathrm{aux}}$$

and the inclusion (which, indeed, is a $*$-representation)

$$\mathcal{B}(\mathcal{H}) \hookrightarrow \mathcal{B}(\mathcal{H}) \otimes 1^{\mathrm{aux}} \subset \mathcal{B}(\mathcal{H} \otimes \mathcal{H}^{\mathrm{aux}}) \tag{1}$$

into the operator algebra of the extended Hilbert space. Let ϱ be the density operator of a positive linear form l over $\mathcal{B}(\mathcal{H})$. A vector ψ of $\mathcal{H} \otimes \mathcal{H}^{\mathrm{aux}}$ is said to *purify* l, and hence ϱ, iff

$$l(b) \equiv \mathrm{Tr}\, b\varrho = \langle \psi, b \otimes 1^{\mathrm{aux}} \psi \rangle \quad \forall b \in \mathcal{B}(\mathcal{H}). \tag{2}$$

A distinguished way to choose the auxiliary Hilbert space is to require

$$\mathcal{H}^{\mathrm{aux}} = \mathcal{H}^*, \quad \mathcal{W} := \mathcal{H} \otimes \mathcal{H}^*, \tag{3}$$

which results in the *standard purification,* based on the standard representation of $\mathcal{B}(\mathcal{H})$. In what follows this choice is assumed, and we have to fix some notations and conventions at the beginning.

Let $\phi \in \mathcal{H}$. The element $\phi^* \in \mathcal{H}^*$, is defined by $\phi^*(\phi') = \langle \phi, \phi' \rangle$. In Dirac's notation:

$$\phi \leftrightarrow |\phi\rangle, \quad \phi^* \leftrightarrow \langle \phi |.$$

Being in finite dimensions, every operator is Hilbert–Schmidt, and $\mathcal{W}$ is canonically isomorphic to $\mathcal{B}(\mathcal{H})$. This can be made explicit with two arbitrarily chosen orthonormal bases $\phi_1, \phi_2, \ldots$ and $\phi'_1, \phi'_2, \ldots$ of $\mathcal{H}$ in writing

$$w=\sum |\phi_j\rangle\langle\phi_j, w\phi_k'\rangle\langle\phi_k'|, \quad w\in\mathcal{W}. \tag{4}$$

The Hilbert–Schmidt scalar product on $\mathcal{W}$ is

$$(w_2, w_1) := \operatorname{Tr} w_2^* w_1 = \sum \langle w_2\phi_k', \phi_j\rangle\langle\phi_j, w_1\phi_k'\rangle. \tag{5}$$

The star operation in $\mathcal{B}(\mathcal{H})$ is equivalent with a conjugation in $\mathcal{W}$,

$$w \mapsto w^* \quad \text{or} \quad (\phi\otimes\tilde{\phi}^*)^* = \tilde{\phi}\otimes\phi^*.$$

We need some operators acting on $\mathcal{W}$. The standard representation of $\mathcal{B}(\mathcal{H})$ is the inclusion (1), specified by (3), and acting as follows:

$$b\mapsto L_b, \quad L_b w := bw, \quad b\in\mathcal{B}(\mathcal{H}).$$

We also need the right multiplication R_b, i.e., $R_b w = wb$. The right multiplication can be used to implement the standard representation of $\mathcal{B}(\mathcal{H}^*)$. Notice the different meaning of the *-operations on $\mathcal{W}=\mathcal{B}(\mathcal{H})$ and on $\mathcal{B}(\mathcal{W})$ seen in

$$(L_b)^* = L_{b^*}, \quad (L_b w)^* = (R_b)^* w^*$$

and in similar relations after exchanging L_b and R_b. Now, let $\hat{l}$ be a linear form on $\mathcal{B}(\mathcal{W})$ and l its *restriction* or reduction onto $\mathcal{B}(\mathcal{H})$. The relation

$$\hat{l}\mapsto l, \quad l(b) := \hat{l}(L_b), \quad b\in\mathcal{B}(\mathcal{H}) \tag{6}$$

encodes the partial trace over $\mathcal{H}^*$ on $\mathcal{W}$. Focusing our attention on the purification procedure, we shall apply this well-known mapping mainly to linear functionals of rank one. In that case the essence of the reduction mapping to the factors of $\mathcal{W}$ is contained in

$$(w_2, L_b R_c w_1) = \operatorname{Tr} w_2^* b w_1 c. \tag{7}$$

Its left-hand-side defines a linear form $B\mapsto(w_2, Bw_1)$ over $\mathcal{B}(\mathcal{W})$, and, varying w_1 and w_2 within $\mathcal{W}$, one can get every linear functional of rank one. Presently we need to consider (7) with $w_1 = w_2 = w$ and with either c or b the identity operator. Then, for $B\in\mathcal{B}(\mathcal{W})$ and $b,c\in\mathcal{B}(\mathcal{H})$, the left- and the right-hand sides of (7) may be rewritten

$$\hat{l}(B) = (w, Bw), \quad l(b) = \operatorname{Tr} ww^* b, \quad l'(c) = \operatorname{Tr} w^* w c.$$

$\varrho = \varrho_l := ww^*$ is called the *density* or the *density operator* of l, while w is said to *purify* l. In the same spirit, a positive linear functional $\hat{l}$ of rank one, which reduces to l, is a *purification* of l.

From now on, instead of switching forth and back between linear forms and their densities, we remain mainly with the latter. Accordingly we define the mappings

$$\Pi w = ww^*, \quad \Pi' w = w^* w.$$

The mapping Π (and similarly the mapping Π'), is slightly more subtle than the reduction mapping (6). Its domain of definition is $\mathcal{W}$. Thus Π is composed of a Hopf bifurcation from w to the rank one density operator $|w)(w|$, representing the linear form $B\to(w,Bw)$, followed by the reduction (6):

$$w\mapsto|w)(w|\mapsto ww^*.$$

Here we used Dirac's notation relative to the scalar product (5) in $\mathcal{W}$. Π is a bundle projection, where the bundle space is $\mathcal{W}$ and the base space is the cone of (not necessarily normalized) density operators (i.e., positive trace class operators). Being in finite dimension, the base space is the positive cone of $\mathcal{B}(\mathcal{H})$. The bundle fibers are manifolds. However, the dimension of the fibers vary with the rank n_w of $w \in \mathcal{W}$. Therefore certain discontinuities occur if the rank is changing.

All this can be seen by the ''diagonal'' form of (4), which is the Gram–Schmidt decomposition of w. Let $\lambda_1, \lambda_2, \ldots$ be the n_w nonzero eigenvalues of ww^* and $\phi_1, \phi_2, \ldots$ their orthonormal eigenvectors,

$$ww^* = \sum \lambda_j |\phi_j\rangle\langle\phi_j|, \quad \lambda_k > 0.$$

There exists exactly one other orthonormal basis of vectors, $\phi'_1, \phi'_2, \ldots$ of the same length n_w, fulfilling

$$w = \sum \sqrt{\lambda_k} |\phi_k\rangle\langle\phi'_k|, \quad w^* w = \sum \lambda_j |\phi'_j\rangle\langle\phi'_j| \tag{8}$$

and the positive numbers λ_j sum up to (w,w). From (8) one can read off the polar decompositions

$$w = \sqrt{ww^*}\, v = v \sqrt{w^* w}, \quad v = \sum |\phi_k\rangle\langle\phi'_k|. \tag{9}$$

The index k runs from 1 to n_w. One may call v the *phase of* w *relative to* $\varrho = ww^*$. The projection operators $v^* v$ and vv^*, attached to the partial isometry v, map $\mathcal{H}$ onto the support spaces of $w^* w$ and ww^*, respectively. Later on we need the operator $J = J_w$,

$$J_w x = v x^* v = \sum |\phi_j\rangle\langle\phi'_j, x^* \phi_k\rangle\langle\phi'_k|, \tag{10}$$

which, for completely entangled w, is the well-known *modular conjugation*. One easily establishes

$$(J_w)^2 x = (vv^*) x (v^* v), \quad (Jx, y) = (Jy, x). \tag{11}$$

If $\varrho \geqslant 0$ is a density operator, the set $\Pi^{-1}\varrho$ consists of all w satisfying $\varrho = ww^*$. Along this fiber the orthoframe $\phi'_1, \phi'_2 \ldots$ in (8) and (9) varies arbitrarily. Thus the fiber at ϱ is isomorphic, though not canonically, to a complex Stiefel manifold.[26] These isomorphisms are parametrized by the different possibilities to choose an orthoframe for the nonzero eigenvalues of ϱ. The *structure* or *gauge group* of $\Pi^{-1}\varrho$ consists of all unitary $u \in \mathcal{B}(\mathcal{H})$ acting by R_u.

Iff ϱ is already pure, $\varrho = |\phi\rangle\langle\phi|$, its purifications reads $w = |\phi\rangle\langle\phi'|$. That is, the purifying vectors are necessarily product vectors (''unentangled'' vectors).

In case the rank of ϱ is larger than one, w is called *entangled* in the domain of quantum information theory. Accordingly, *complete entanglement* of w is reached if the density operator ϱ is of maximal rank $n_w = \dim \mathcal{H}$. In this case, in traditional *-representation theory, ϱ is called *faithful* and w *separating*. $\varrho = ww^*$ is faithful iff w is invertible.

The set of all faithful ϱ is the base space of a principal fiber bundle with free action of the unitaries R_u. The fiber space consists of all invertible w, the projection is Π.

III. PURIFICATION AND TANGENTS

A smooth, oriented curve in $\mathcal{W}$, passing through w, defines at w a *tangent* or *velocity vector* x. Hence the tangent space, $\mathcal{T}_w$ at w, may be identified with $\mathcal{W}$ if considered as a real linear space.

Assume that w and the unitaries u depend smoothly on a parameter, and let us use a dot to show parameter differentiation. The gauge transformation $w \mapsto w' := wu$ induces the relation

$$x \mapsto x' = xu + w\dot{u}, \quad x = \dot{w}, \quad x' = \dot{w}'. \tag{12}$$

Let us now consider Π, and assume $\Pi w = \varrho$. Π induces a mapping Π_* from the tangent space of $\mathcal{W}$ into the density operator's tangents.

Being a first-order problem, it is sufficient for the following to assume a curve as simple as possible, say $w(\lambda) = w + \lambda x$. The curve is projected by Π to a curve of density operators $\varrho_\lambda = w(\lambda) w^*(\lambda)$ of $\mathcal{B}(\mathcal{H})$. Differentiating at $\lambda = 0$ results in a tangent $\Pi_* x = \xi$ at ϱ,

$$\xi = \dot{\varrho}, \quad \xi = (ww^*)^{\cdot} = xw^* + wx^*. \tag{13}$$

A tangent vector x at w is called *vertical* iff $\Pi_* x = 0$. The real vector space of the vertical tangents at w is denoted by $\mathcal{T}_w^{\mathrm{ver}}$. It is a straightforward and well-known exercise to show: The gauge transformation $x \mapsto x'$ of (12) maps vertical tangents at w to vertical tangents at w'.

We look at vertical tangents as labels for the physical phase. The phase of a single state or of its density operator is not an observable. Which purifying vector w we choose is physically irrelevant. What can be observed are relative phases, for example in interference experiments. The relative phases should depend on the way a density operator is changed to become another one. There should be a protocol according to which the tangents, and hence the phases, are transported along a curve within the space of density operators. This can be achieved by the help of a parallel transport.

The standard procedure is to split the tangent space at every w into a direct sum of the vertical and of an horizontal part. Respecting the complex linear structures, we restrict ourselves to decompositions defined by the real part of an Hermitian metric: We assume at every w a distinguished positive Hermitian sesquilinear form

$$w \mapsto (x_2, x_1)_w, \quad x_1, x_2 \in \mathcal{T}_w. \tag{14}$$

For completely entangled w it should be positive definite. Now $\mathrm{Re}\,(.,.)_w$, the real part of (14), converts the tangent space at w into a *real* Hilbert space. The *velocity* with which a curve goes through w is the square root of $(x,x)_w$ with x the tangent at that point. In this setting, parallel transport is asking for a minimal velocity lift of a given tangent at the base space. This, in turn, induces a metrical structure at the base space: One calls the *velocity of a base space tangent* the minimum of the velocities of all possible lifts.

Thus, the *horizontal part*, x^{hor}, of a tangent x at w is the unique element of the set $x + \mathcal{T}_w^{\mathrm{ver}}$ with the smallest velocity. This is in accordance with the definition of $\mathcal{T}_w^{\mathrm{hor}}$ as the orthogonal complement of $\mathcal{T}_w^{\mathrm{ver}}$ in the real Hilbert space $\mathcal{T}_w$, the latter equipped with the scalar product $\mathrm{Re}\,(.,.)_w$.

There is a distinguished real subspace, $\mathcal{T}_w^{\mathrm{Ver}} \subset \mathcal{T}_w^{\mathrm{ver}}$, containing all tangents

$$x = wa, \quad a = -a^* \in \mathcal{W},$$

which are obviously vertical. If w is invertible (completely entangled), every vertical tangent can be uniquely expressed in that way. But generally, $\mathcal{T}_w^{\mathrm{Ver}}$ is a proper subspace of $\mathcal{T}_w^{\mathrm{ver}}$. We call a vertical tangent *neutral* iff it is orthogonal to $\mathcal{T}_w^{\mathrm{Ver}}$ with respect to $\mathrm{Re}\,(.,.)_w$. Hence, every tangent x allows for an orthogonal decomposition

$$x = x^{\mathrm{hor}} + x^{\mathrm{ver}}, \quad x^{\mathrm{ver}} = x^{\mathrm{neutral}} + x^{\mathrm{Ver}}. \tag{15}$$

IV. PHASE TRANSPORT AND BURES METRIC

The most natural and simple choice for the Hermitian metric $(x_2, x_1)_w$ of (14) is certainly the Hilbert–Schmidt scalar product (5). This choice is particularly interesting for several reasons.

At first it gives a straightforward generalization of the geometric phase by the parallel transport evolving from this choice. Indeed, one obtains a natural extension of the Fock,[27], Berry,[9] Simon,[10] Wilczek and Zee[24] parallel transport to density operators.

Transport of state vectors along closed curves generates a holonomy problem. In the period between V. Fock and M. Berry this has become clear. B. Simons explained how to calculate the holonomy by the second Chern class of the Hilbert space if considered as a line bundle. There is an extensive literature on the transport of phases along curves and loops of pure states, see Ref. 28 for a selection of important results, applications, and references. Particular examples in using and calculating the geometric phase can already be found in papers of decades past.

Second, one gets a Riemannian metric[14] on the (not necessarily normalized) density operators of $\mathcal{B}(\mathcal{H})$. Its distance function is the distance introduced by Bures[29] in following a similar construction of Kakutani[30] in probability spaces. Being the infinitesimal version of *Bures' distance*, we call this Riemannian metric *Bures' metric*.

And, finally, already the choice

$$(x_2,x_1)_w=(x_2,x_1),\quad \forall w,$$

shows essential problems in deviating from a genuine fiber bundle.

We start by enumerating the tangents y orthogonal to $\mathcal{T}_w^{\mathrm{Ver}}$

$$(y,wa)+(wa,y)=0,\quad \forall a+a^*=0.$$

That condition straightforwardly comes down to

$$y^*w=w^*y \tag{16}$$

and y is orthogonal to all Ver-tangents iff w^*y is Hermitian. (16) is the *parallelity condition*,[8] which extends the transport condition for the geometric phase from pure to mixed states.

To decompose y in its neutral and horizontal part, we start by completing the two orthonormal systems of the Schmidt decomposition (8) arbitrarily and set $\lambda_j=0$ if $j>n_w$. By sandwiching (16) between the orthobase $\{\phi_i'\}$ we get

$$\sqrt{\lambda_k}\langle\phi_j',y^*\phi_k\rangle=\sqrt{\lambda_j}\langle\phi_j,y\phi_k'\rangle.$$

There evolve two conditions on the matrix elements:

$$j\leq n_w,\quad k>n_w\Rightarrow\langle\phi_j',y\phi_k\rangle=0.$$

$$k,j\leq n_w\Rightarrow\frac{\langle\phi_j',y^*\phi_k\rangle}{\sqrt{\lambda_j}}=\frac{\langle\phi_j,y\phi_k'\rangle}{\sqrt{\lambda_k}}.$$

No restriction occurs for $j>n_w$, $k\leq n_w$. There is an Hermitian g such that

$$\langle\phi_j,g\phi_k\rangle=\frac{\langle\phi_j,y\phi_k'\rangle}{\sqrt{\lambda_k}},\quad k\leq n_w.$$

One may choose the matrix elements of g with indices both larger than n_w arbitrarily but consistent with $g=g^*$.

The tangent $y_1=gw$ is horizontal[31,32] because it is orthogonal to all ver-tangents x. Indeed, $xw^*+wx^*=0$ implies $(gw,x)+(x,gw)=(g,xw^*+wx^*)=0$. What remains to check is the case of a tangent y_0, real orthogonal to all gw, $g=g^*$, and to all Ver-tangents. From the first condition it follows $wy_0^*+y_0w^*=0$, hence verticality, and from the second we obtain $w^*y_0=y_0^*w$. This is equivalent with

$$\langle\phi_j,y_0\phi_k'\rangle=0,\quad \forall j,k\leq n_w$$

or

$$y \quad \text{neutral} \quad \Leftrightarrow w^* y = y w^* = 0. \tag{17}$$

We conclude that every tangent x allows for a unique decomposition

$$x = gw + x_0 + wa \tag{18}$$

in a horizontal, a neutral, and a vertical part where g is Hermitian, a anti-Hermitian, and x_0 satisfies (17). With the extra conditions

$$\langle \phi_j , g\phi_k' \rangle = \langle \phi_j , a\phi_k' \rangle = 0, \quad k,j \geqslant n_w ,$$

both g and a are unique. The last conditions are equivalent to the choice of maximal null-spaces, i.e., *minimal supports* for g and a. They allow one to define g and a uniquely.

The transformation property (12) implies

$$w \mapsto w' = wu \Rightarrow a \mapsto a' = u^* a u + u^* \dot{u} , \tag{19}$$

so that $x \mapsto a$ is a connection form (gauge potential) **a** for the gauge group $u \mapsto R_u$. However, support properties may not change continuously. For parameter values at which the rank of w is changing, one has to understand g or a as equivalence class with respect to the kernel of $g \mapsto gw$ or $a \mapsto wa$, respectively. Then (19) remains meaningful even in those cases.

In our next step we look at g and a. g, which describes the horizontal part of a tangent vector x, can be expressed by $\xi := \Pi_* x$ and $\varrho := ww^* = \Pi w$. We need the pair x and $\tilde{\varrho} := w^* w$ to gain a. We get

$$\varrho g + g \varrho = \xi, \quad \tilde{\varrho} a + a \tilde{\varrho} = w^* x - x^* w. \tag{20}$$

The first equation[32,31] is obtained from (13). To see the second one,[13] insert (18) into its right-hand side.

Apart from an obvious restriction on ξ, (20) can be solved to get g or a, and several ways to do so are well known. A review is in Ref. 33. The restriction in question reads $\langle \phi, \xi \phi \rangle = 0$ whenever ϕ is in the null space of ϱ for the first equation, and $\langle \phi', \xi \phi' \rangle = 0$ whenever ϕ' is in the null space of $\tilde{\varrho}$. Below we assume they are satisfied.

With the solvability conditions in mind we rewrite (20) as equations between operators in $\mathcal{B}(\mathcal{W})$. In order not to overload notations we abbreviate

$$\mathrm{L} \equiv L_\varrho , \quad \mathrm{R} \equiv R_\varrho , \quad \tilde{\mathrm{L}} \equiv L_{\tilde{\varrho}} , \quad \tilde{\mathrm{R}} \equiv R_{\tilde{\varrho}} .$$

These are families of operators indexed by ϱ or $\tilde{\varrho}$.

Let us start now from (20). The equations can be solved by

$$g = (\mathrm{L} + \mathrm{R})^{-1} \xi, \quad a = (\tilde{\mathrm{L}} + \tilde{\mathrm{R}})^{-1} (w^* x - x^* w). \tag{21}$$

The operationally defined inverse exists by the solvability condition above. With two tangents ξ_j at ϱ and their horizontal lifts x_j^{hor} we get the Riemannian metric[16,14] belonging to the Bures distance

$$(\xi_2 , \xi_1)^{\mathrm{Bures}} := \mathrm{Re}\,(x_1^{\mathrm{hor}}, x_2^{\mathrm{hor}}) = \tfrac{1}{2} \mathrm{Tr}\, \varrho (g_1 g_2 + g_2 g_1) \tag{22a}$$

or, equivalently,

$$(\xi_2 , \xi_1)^{\mathrm{Bures}} = \tfrac{1}{2} \mathrm{Tr}\, \xi_2 g_1 = \tfrac{1}{2} \mathrm{Tr}\, \xi_2 (\mathrm{L} + \mathrm{R})^{-1} \xi_1 . \tag{22b}$$

There is a similar procedure with the second equation of (21) resulting in the connection $\mathbf{a}^{\text{geo}}$ with $\mathbf{a}^{\text{geo}}(x):=wa$. The superscript ''geo,'' if used, is a reminder for the physical important geometric phase. From (21) we get

$$\mathbf{a}^{\text{geo}}=\frac{\tilde{\mathrm{L}}}{\tilde{\mathrm{L}}+\tilde{\mathrm{R}}}(w^{-1}dw)-\frac{\tilde{\mathrm{R}}}{\tilde{\mathrm{L}}+\tilde{\mathrm{R}}}(w^{-1}dw)^{*}, \tag{23a}$$

where $w^{-1}\,dw$ is the left canonical 1-form with values in the Lie algebra of $\mathrm{GL}(\mathcal{H})$. $\mathbf{a}^{\text{geo}}$ takes values in the Lie algebra of the gauge group $U(\mathcal{H})$ acting from the right via $u\mapsto R_u$.

Formula (23a) represents $\mathbf{a}^{\text{geo}}$ as the difference of two Hermitian conjugated parts of type (1,0) and (0,1), respectively:

$$\mathbf{a}^{\text{geo}}=\mathbf{a}_{1,0}-\mathbf{a}_{0,1}, \quad \mathbf{a}_{0,1}=\mathbf{a}_{1,0}^{*}.$$

Another interesting equation expresses $\mathbf{a}^{\text{geo}}$ as sum of the canonical 1-form $\mathbf{a}^{\text{can}}$ of the bundle $\mathrm{GL}(\mathcal{H})/U(\mathcal{H})$ and a horizontal Ad-1-form[25]

$$\mathbf{a}^{\text{geo}}=\frac{w^{-1}\,dw-(w^{-1}\,dw)^{*}}{2}+\frac{\tilde{\mathrm{L}}-\tilde{\mathrm{R}}}{\tilde{\mathrm{L}}+\tilde{\mathrm{R}}}\,\frac{w^{-1}dw+(w^{-1}dw)^{*}}{2}. \tag{23b}$$

Since the second form is horizontal, it can be rewritten in terms of $d\varrho$ and we get

$$\mathbf{a}^{\text{geo}}=\mathbf{a}^{\text{can}}+w^{-1}\left(\frac{\mathrm{L}-\mathrm{R}}{2(\mathrm{L}+\mathrm{R})}\,d\varrho\right)(w^{-1})^{*} \tag{23b$'$}$$

$$=w^{-1}dw-w^{-1}\left(\frac{\mathrm{R}}{\mathrm{L}+\mathrm{R}}d\varrho\right)(w^{-1})^{*}. \tag{23c}$$

It becomes immediately clear that $\mathbf{a}^{\text{geo}}(x)=\mathbf{a}^{\text{can}}(x)$ iff $\mathrm{L}\xi=\mathrm{R}\xi$, where $\xi:=wx^{*}+xw^{*}$, i.e., iff ϱ commutes with $\dot{\varrho}$.

This observation motivates the decomposition

$$\mathcal{T}_{\varrho}=\mathcal{T}_{\varrho}^{\parallel}+\mathcal{T}_{\varrho}^{\perp} \tag{24}$$

of the tangent space $\mathcal{T}_{\varrho}$ into a direct sum, where $\xi\in\mathcal{T}_{\varrho}^{\parallel}$ iff ξ commutes with $\varrho=ww^{*}$ or, equivalently, iff $\langle\phi_j,\xi\phi_k\rangle=0$ for any two eigenvectors ϕ_j, ϕ_k, of ϱ with different eigenvalues. On the other hand, $\xi\in\mathcal{T}_{\varrho}^{\perp}$ iff it can be written as a commutator $i[b,\varrho]$ with a suitable Hermitian b. (24) is a well-known matrix decomposition: Assume ϱ represented as block diagonal matrix, every block belongs to just one eigenvalue. This induces a block representation of any matrix ξ. One gets $\xi^{\parallel}$ by setting zero every off-diagonal block of ξ. If the entries in the diagonal blocks are set to zero, one obtains $\xi^{\perp}$. In our present field of interest Hübner[18] obtained a decomposition (24) of the Bures Riemannian metric. For larger classes of metrics this has been done by Hasegawa and Petz (Refs. 34 and 35).

This brings us back to the metric (22). There is a solution g_1 commuting with ϱ iff ξ_1 does so: The support ϱ cannot be smaller than the support of ξ. Hence $2g_1=\varrho^{-1}\xi_1=\xi_1\varrho^{-1}$ is operational well defined. Inserting in (22b) results in

$$(\xi_2,\xi_1)^{\text{Bures}}=\tfrac{1}{4}\mathrm{Tr}\,\xi_2\xi_1\varrho^{-1}, \quad \xi_1\in\mathcal{T}_{\varrho}^{\parallel}. \tag{25}$$

Comparing this with the Riemannian metric

$$(\xi_2,\xi_1)^{\text{can}}:=\tfrac{1}{8}\mathrm{Tr}\,(\xi_2\xi_1+\xi_1\xi_2)\varrho^{-1}=\mathrm{Tr}\,\xi_2(\mathrm{L}^{-1}+\mathrm{R}^{-1})\xi_1,$$

the inequality $4/(\mathrm{L}+\mathrm{R})\leqslant(1/\mathrm{L})+(1/\mathrm{R})$ gives[36]

$$(\xi,\xi)^{\text{Bures}} \leqslant (\xi,\xi)^{\text{can}}$$

and equality holds if and only if $\xi \in \mathcal{T}_\varrho^\parallel$, or, what is the same, if ξ commutes with ϱ.

Let $\phi_1, \ldots$ be a complete orthonormal eigenvector basis of $\varrho = ww^*$ and ξ with eigenvalues λ_j and $\dot{\lambda}_j$, respectively. Then we get from (25) the following quadratic form:

$$\frac{1}{4}\sum d\lambda_j^2 \lambda_j^{-1} = \sum d\mu_j^2, \quad \mu_j := \sqrt{\lambda_j}.$$

This is an Euclidean metric. However, restricted to the state space, where $\lambda_1, \ldots$ becomes a probability vector, we get Fisher's metric (''Fisher–Rao metric'').[37]

If the Bures metric is restricted to a submanifold of mutual commuting states, the Fisher metric is obtained. Moreover, *on any submanifold of commuting density operators*, whether normalized or not, *the phase transport is holonomically trivial.*

Indeed, we can form the lift $\varrho \to w = \sqrt{\varrho}$. The assumed commutativity provides us with Hermitian and commutative w and $x = \dot{w}$, and with $\varrho = ww^* = w^*w = \tilde{\varrho}$. Hence (21) comes down to $\mathbf{a}(x) = 0$, and the lift is horizontal. There is no room for a nontrivial phase.

We see a nontrivial geometric phase is definitely an effect of noncommutativity. We need for them curves with mutually not commuting density operators.

V. AUXILIARY TOOLS

In order to extend our previous considerations to a larger class of connections[25] we need some auxiliary tools.

Looking at Eqs. (23) one can identify functions of L/R and $\tilde{\text{L}}/\tilde{\text{R}}$. These operators are relatives of $\text{L}/\tilde{\text{R}} = \Delta_w$, the Tomita–Takesaki modular operator of the representation $b \mapsto L_b$ with GNS-vector w. The operators are defined if w^{-1} exists, that is for completely entangled w. But, as (23) shows, certain functions of these operators can be defined for every w.

Let $t \mapsto f(t)$ be a function defined for $0 < t < \infty$. We assume the existence of

$$f(0) := \lim_{t\to 0} f(t), \quad f(\infty) := \lim_{t\to\infty} f(t). \tag{26}$$

The assumption is necessary if we like to extend the formalism to density operators which are not invertible. Without it, we have to restrict ourselves to completely entangled w, i.e., to faithful density operators.

To treat an example with the assumption (26), we define $f(\text{L}/\tilde{\text{R}}) =: f(\Delta)$. The positive operators L and $\tilde{\text{R}}$ commute. Let λ_j be the eigenvalue of ww^* and of w^*w with the eigenvectors ϕ_j and ϕ_j'. The eigenvectors, suitably chosen, collect in a complete orthonormal basis satisfying the Gram–Schmidt decomposition (8). λ_j is zero if $j > n_w$ and positive otherwise. Now

$$\text{L}v_{jk} = \lambda_j v_{jk}, \quad \tilde{\text{R}}v_{jk} = \lambda_k v_{jk}, \quad v_{jk} := |\phi_j\rangle\langle\phi_k'|.$$

The elements v_{jk} constitute a complete orthonormal basis of the Hilbert–Schmidt space $\mathcal{W}$. We like $f(\Delta)$ to be diagonalizable with eigenvectors v_{jk}. Remembering $\Delta = \text{L}/\tilde{\text{R}}$ we start with

$$f(\Delta)v_{jk} = f(\lambda_j/\lambda_k)v_{jk}, \quad \text{if } \lambda_k > 0.$$

The remaining possibility is done ''by hand'' in requiring

$$f(\Delta)v_{jk} = f(\infty)v_{jk}, \quad \text{if } \lambda_j > 0, \quad \lambda_k = 0,$$

$$f(\Delta)v_{jk} = f(1)v_{jk}, \quad \text{if } \lambda_j = \lambda_k = 0.$$

With this convention v_{jj} is an eigenvector of $f(\Delta)$ with eigenvalue $f(1)$ for all j.

The same game is to play with $f(\mathrm{L/R})$ and $f(\tilde{\mathrm{L}}/\tilde{\mathrm{R}})$. While the spectra of $f(\mathrm{L/R})$ and $f(\tilde{\mathrm{L}}/\tilde{\mathrm{R}})$ coincide with that of $f(\Delta)$, their eigenvectors are, respectively,

$$|\phi_j\rangle\langle\phi_k| = v_{ji} v_{ik}^*, \quad |\phi'_j\rangle\langle\phi'_k| = v_{ji}^* v_{ik}.$$

VI. A CLASS OF CONNECTIONS

Our aim is to describe a class of connections, essentially that of Dittmann and Rudolph[25]. These objects, as will be seen, are particularly well adapted to the purification of the $\mathcal{H}$-system by that of $\mathcal{W}=\mathcal{H}\otimes\mathcal{H}^*$. We assume w to be completely entangled, so that $\varrho=\Pi w$ is faithful (invertible). Whether it is possible to skip this assumption, either by calculating modulo neutral tangents or by continuity arguments, depends on the asymptotic behavior of certain functions to be introduced below.

Let $[0,\infty]\ni s\mapsto r(s)\in\mathbb{C}$ be a smooth function and $r(1)=1/2$. Then

$$(r(\tilde{\mathrm{L}}/\tilde{\mathrm{R}})y)^* = \overline{r}(\tilde{\mathrm{R}}/\tilde{\mathrm{L}})y^*.$$

Mimicking Eq. (23a) we define the form

$$\mathbf{a} := \overline{r}(\tilde{\mathrm{L}}/\tilde{\mathrm{R}})(w^{-1}\,dw) - r(\tilde{\mathrm{R}}/\tilde{\mathrm{L}})(w^{-1}\,dw)^*. \tag{27a}$$

It transforms like a connection and takes anti-Hermitian values. To be a connection it must take the correct values at vertical vectors, i.e., $\mathbf{a}(wa)=a$, for all anti-Hermitian a. Thus we need to have

$$\overline{r}(t)+r(1/t)=1, \quad F(t):=\overline{r}(t)-r(1/t)=-\overline{F}(1/t), \tag{28}$$

to get a genuine connection with respect to the gauge group $\mathrm{U}(\mathcal{H})$ acting by $u\mapsto R_u$. Furthermore, as a consequence of $r(1)=1/2$, one observes rescaling invariance of this connection form. Indeed, $\mathbf{a}$ is invariant under $w\mapsto\lambda(w)w$, where $\lambda:\mathcal{W}\to\mathbb{R}$:

$$\mathbf{a}_w(x) = \mathbf{a}_{\lambda w}(d\lambda(x)w+\lambda x),$$

so that there is no need to normalize w in calculating $\mathbf{a}$. The second equation in (28) introduces the function F used in Ref. 25 to label their gauge potentials, and we are allowed now to rewrite (27a) in a manner known already from (23):

$$\mathbf{a} = \mathbf{a}^{\mathrm{can}} + F(\tilde{\mathrm{L}}/\tilde{\mathrm{R}})\frac{(w^{-1}\,dw)+(w^{-1}\,dw)^*}{2} \tag{27b$'$}$$

$$= \mathbf{a}^{\mathrm{can}} + w^{-1}(F(\mathrm{L/R})\,d\varrho)(w^{-1})^* \tag{27b}$$

$$= w^{-1}\,dw - w^{-1}(r(\mathrm{R/L})\,d\varrho)(w^{-1})^*. \tag{27c}$$

One returns to the Bures case by

$$\mathbf{a}=\mathbf{a}^{\mathrm{geo}} \Leftrightarrow r(t)=\frac{t}{1+t} \Leftrightarrow F(t)=(t-1)/(t+1).$$

Before deriving expressions for the vertical and horizontal part of a given tangent x, we draw an important conclusion:

The value of a connection at the lift of a $\|$-tangent is independent of F, respectively, r. Indeed, $F(1)=0$ and $\mathrm{L}x=\mathrm{R}x$ for these tangents, and we get from (27b') immediately

$$\Pi_*(x) \in \mathcal{T}^{\|} \Rightarrow \mathbf{a}(x) = \mathbf{a}^{\text{can}}(x), \quad \forall F$$

allowing to extend a conclusion of Sec. IV:

On submanifolds with mutually commuting density operators the holonomy of every loop is trivial for the whole class of connections considered here.

Indeed, the lift $\varrho \to \sqrt{\varrho}$ is horizontal along every curve of commuting densities.

To obtain the vertical and horizontal part of a tangent x let us apply Eq. (27c) to x multiplied by w from the left. We assumed w to be separating so that there are no nonvanishing neutral tangents. Therefore

$$x^{\text{ver}} = x^{\text{Ver}} = w\mathbf{a}(x) = x - (r(\text{R}/\text{L})\xi)(w^*)^{-1} \tag{29a}$$

$$= x - r(\Delta^{-1})(x + wx^* w^{*-1}) \tag{29b}$$

$$= x - r(\Delta^{-1})[x + \Delta^{1/2}Jx], \tag{29c}$$

reminding $wx^*(w^*)^{-1} = \Delta^{1/2}Jx = J\Delta^{-1/2}x$. (29) reflects the decomposition of a general tangent into a vertical and a horizontal part, see (15). We conclude

$$x^{\text{hor}} = (r(\text{R}/\text{L})\xi)(w^*)^{-1} = r(\Delta^{-1})[x + \Delta^{1/2}Jx]. \tag{30}$$

A connection form $\mathbf{a}$ regulates the change of the phase v along a horizontal lift, $w_t = \sqrt{\varrho_t}v_t$, of a curve ϱ_t. We express $\mathbf{a}$ by

$$\begin{aligned}
\mathbf{a}(\dot{w}) = \mathbf{a}(\sqrt{\varrho}\dot{v} + (\sqrt{\varrho})^{\cdot} v) &= \mathbf{a}(\sqrt{\varrho} v v^* \dot{v} + (\sqrt{\varrho})^{\cdot} v) \\
&= v^*\dot{v} + v^*\mathbf{a}(\sqrt{\varrho}^{\cdot})v \\
&= v^*\dot{v} + v^*\mathbf{a}\left(\frac{1}{\sqrt{\text{L}} + \sqrt{\text{R}}}\dot{\varrho}\right)v \\
&= v^*\dot{v} + v^*\frac{1}{2}\frac{1}{\sqrt{\text{LR}}}\left(F(\text{L}/\text{R}) + \frac{\sqrt{\text{R}} - \sqrt{\text{L}}}{\sqrt{\text{R}} + \sqrt{\text{L}}}\right)(\dot{\varrho})v.
\end{aligned}$$

and see that the horizontality of w_t is equivalent with

$$0 = \dot{v}v^* + \frac{1}{2}\frac{1}{\sqrt{\text{LR}}}\left(F(\text{L}/\text{R}) + \frac{\sqrt{\text{R}} - \sqrt{\text{L}}}{\sqrt{\text{R}} + \sqrt{\text{L}}}\right)(\dot{\varrho}). \tag{31}$$

One observes, that there is one and only one connection in our setting with a global horizontal section, $\varrho \mapsto \sqrt{\varrho}$. That connection is given by

$$F(t) = -\frac{1 - \sqrt{t}}{1 + \sqrt{t}}, \quad r(t) = \frac{\sqrt{t}}{1 + \sqrt{t}}.$$

VII. CONNECTION AND METRIC

In this section we specify a class of Hermitian metrics (14) on $\mathcal{W}$, which respects the purification scheme. Our first task is to ask for Hermitian metrics on the complex manifold $\mathcal{W}$, the real part of which is compatible with a given connection form of Sec. VI. We demand: At every completely entangled $w \in \mathcal{W}$, the vertical tangents are real orthogonal to the horizontal ones. In the case where there exists a Hermitian metric doing this task, the functions F and r characterizing the connection have to be real. In the next step we describe the Hermitian and Riemannian metric one obtains by reduction from the purification space to that of (unnormalized) density operators.

Starting with a connection (27a), (28), there is some freedom in the choice of the Hermitian metric. It is an interesting question in its own right, whether, by a reasonable condition, the Hermitian metric becomes unique. We explain in the last part of this section how this can be done. If we start from a Riemannian metric on the density operators, the uniqueness problem is more involved. Nevertheless, our additional condition solves it also, at least for the monotone Riemannian metrics.

To start our little program we construct Hermitian metrics (14) by modifying the Hilbert Schmidt scalar product on $\mathcal{W}$ by a function $k(\Delta)$ of the modular operator. Like R and L the modular operator Δ depends on w. Our ansatz for the Hermitian product in $T_w\mathcal{W}$ reads

$$(x_2,x_1)_w := (x_2, k(\Delta_w)^{-1} x_1), \tag{32}$$

where k is a real positive smooth function defined either only on $0<t<\infty$ or on the closed interval $0\leqslant t\leqslant\infty$. We use the rules explained in Sec. V. There are two main merits with such a choice of the modified Hermitian metric: The symmetry group of the metric contains the unitary group $\mathrm{U}(\mathcal{H})\times\mathrm{U}(\mathcal{H}^*)$. The second is the rescaling invariance of Δ under $w\mapsto\lambda(w)w$, where $\lambda(w)$ denotes (a sufficiently smooth) real function on $\mathcal{W}$. Rescaling invariance is a further reason not to insist on normalized density operators.

In determining the connection form compatible with (32), we follow the recipe of Sec. III. We need the real-orthogonal complement of the vertical directions. They are to gain by the metrical independence of verticality. Namely, if a tangent x is real orthogonal to all vertical ones, $k(\Delta)^{-1}x$ is horizontal with respect to the Hilbert–Schmidt metric. Therefore, as shown in Sec. IV, we are allowed to write $x=gw$ with a Hermitian g. Conclusion:

A tangent x is horizontal with respect to (32), if it can be represented as

$$x = k(\Delta)(gw) = k(\mathrm{L}/\mathrm{R})(g)w, \quad g = g^*. \tag{33}$$

The real space of horizontal tangents is the fix point set of an antilinear operator, S_w^k, acting on $\mathcal{W}$. Our notation is borrowed from that of the Tomita–Takesaki operator $S_w = J\sqrt{\Delta}$, which will be returned if $k\equiv 1$. Our definition is

$$S_w^k := Jk(\Delta^{-1})k(\Delta)^{-1}\sqrt{\Delta} = k(\Delta)k(\Delta^{-1})^{-1}S_w. \tag{34a}$$

If this operator acts on $x=k(\Delta)(gw)$ the result is $k(\Delta)(g^*w)$. Comparison with (33) establishes: x is a fix point of S_w^k if and only if x is horizontal.

The square of the operator (34a) is J^2; compare (11). J^2 is the identity of $\mathcal{W}$ iff w is invertible. Further, the adjoint of S_w^k with respect to (32) is $\sqrt{\Delta}J$ and, as it should be, independent of k. (Tomita–Takesaki theory calls it "F_w.") Finally we polar decompose (34a) to get the appropriate modifications of the modular operator, $\Delta=\Delta_w$, and of the modular conjugation, $J=J_w$,

$$S_w^k = J_w^k |S_w^k|, \quad \Delta_w^k := |S_w^k|^2, \tag{34b}$$

$$\Delta_w^k = k(\Delta^{-1})k(\Delta)^{-1}\Delta, \quad J_w^k = J\sqrt{k(\Delta^{-1})k(\Delta)^{-1}}. \tag{34c}$$

We now ask for the connection coming with the metric. The connection form belonging to (32) annihilates all the horizontal vectors (33). This reasoning, applied to (27a) or (27b), determines the function r or F. The calculation shows, in accordance with (28),

$$r(t) = \frac{tk(1/t)}{k(t)+tk(1/t)}, \quad \text{respectively}, \quad F(t) = \frac{tk(1/t)-k(t)}{tk(1/t)+k(t)}. \tag{35}$$

Obviously, *the functions r and F are real valued* if the connection is gained from a Hermitian metric (32). A cross check of (35) is in setting $k \equiv 1$. We get $r(t) = t/(1+t)$ and $F(t) = (t-1)/(t+1)$ as it should be for the Bures case.

On the other hand, given r or F, there is some freedom for k since the induced connection depends on $k(t)/k(1/t)$ only.

$$\frac{k(t)}{k(1/t)} = 1 \Leftrightarrow r(t) = \frac{t}{1+t}, \quad F(t) = \frac{(t-1)}{(t+1)}, \quad \mathbf{a} = \mathbf{a}^{\text{geo}},$$

$$\frac{k(t)}{k(1/t)} = t \Leftrightarrow r(t) = \frac{1}{2}, \quad F(t) = 0, \quad \mathbf{a} = \mathbf{a}^{\text{can}}.$$

In particular, there is no modification of the Tomita–Takesaki operators by (34) if the connection is $\mathbf{a}^{\text{geo}}$. More generally, from (35) we get

$$\frac{k(t)}{k(1/t)} = t\frac{r(1/t)}{r(t)} = t\frac{1-F(t)}{1+F(t)} \tag{36}$$

and find, remarkably enough, the modified Tomita–Takesaki operators (34) depending on F only. Further, by (36), the positivity of k enforces the inequality

$$-1 < F(t) < 1 \tag{37}$$

for F to be obtained from a k. In order to invert (36), the inequality is also sufficient. According to (28) one needs only to check $F<1$ for real F. Then, given F, the general solution of the problem is

$$k(t) := \sqrt{t}(1-F(t))q(t),$$

q being an arbitrary positive function fulfilling $q(t) = q(1/t)$.

We started from a Hermitian metric on $\mathcal{W}$, derived conditions for horizontality, and determined the connection. Now we go back to $\mathcal{H}$ and to its density operators: We ask for the Hermitian and Riemannian metric induced on the space of density operators. That is, with two tangents ξ and η at $\Pi w = \varrho$, we are concerned with

$$(\eta,\xi)_\varrho := (y^{\text{hor}}, x^{\text{hor}})_w, \quad \text{Re}\,(\eta,\xi)_\varrho = \frac{(\eta,\xi)_\varrho + (\xi,\eta)_\varrho}{2}.$$

x^{hor} and y^{hor} are the horizontal lifts of ξ and η. In the present paper the $\mathbb{C}$-valued $\mathbb{R}$-linear form $\xi,\eta \mapsto (\eta,\xi)_\varrho$ is defined on the real tangents. Nevertheless, for obvious reasons, we call it ''Hermitian.'' Relying on (30) we conclude

$$(y^{\text{hor}}, x^{\text{hor}})_w = \text{Tr}\, r(\text{L/R})(\eta)\frac{r(\text{R/L})}{\text{R}k(\text{L/R})}(\xi) = \text{Tr}\,\eta\frac{r(\text{R/L})^2}{\text{R}k(\text{L/R})}\xi,$$

so that

$$(\eta,\xi)_\varrho = \text{Tr}\,\eta\frac{\text{R}k(\text{L/R})}{[\text{R}k(\text{L/R}) + \text{L}k(\text{R/L})]^2}\xi, \tag{38a}$$

where r has been substituted by k by the aid of (35). The real part is a Riemannian metric. By standard rules we get

$$\mathrm{Re}\,(\eta,\xi)_\varrho=\frac{1}{2}\mathrm{Tr}\,\eta\frac{1}{\mathrm{R}k(\mathrm{L}/\mathrm{R})+\mathrm{L}k(\mathrm{R}/\mathrm{L})}\xi. \tag{38b}$$

Petz[36,1,2] was able to classify all monotone Hermitian metrics on the state space, i.e., those for which $(\cdot,\cdot)_\varrho$ does not increase under the action of completely positive and unital mappings. At the heart of his result is the characterization of a monotone metric by an operator monotone function, *f*, defined on $0<t<\infty$, such that

$$(\eta,\xi)_\varrho=\frac{1}{4}\mathrm{Tr}\,\eta\frac{\mathrm{R}^{-1}}{f(\mathrm{L}/\mathrm{R})}\xi. \tag{39}$$

(The factor 1/4 is a normalization convention.) Note that this Hermitian metric becomes symmetric, and hence a Riemannian one, if and only if the function *f* satisfies $f(t)=tf(1/t)$. A function with this algebraic property we call self-transposed, following the terminology for operator means introduced in Ref. 3. Presently, however, the monotonicity of the metric (39) or of its real part is *not* assumed. We need a more general frame. Having this in mind, we compare (39) with (38a) and obtain

$$f(t)=\frac{(k(t)+tk(1/t))^2}{4k(t)}. \tag{40}$$

This equation has a unique solution for *k* depending on *f*, therefore, every Hermitian metric (39) can be reached by exactly one Hermitian metric (32) on the purification space. Indeed, the harmonic mean of $f(t)$ and its transpose, $tf(1/t)$, yields

$$\frac{1}{f(t)}+\frac{1}{tf(1/t)}=\frac{4}{k(t)+tk(1/t)}$$

so that one can insert this into the right-hand side of (40) to express *k* by *f*:

$$k(t)=f(t)\frac{4t^2f(1/t)^2}{[f(t)+tf(1/t)]^2}. \tag{41}$$

Moreover, using (35) we get

$$r(t)=\frac{f(t)}{f(t)+tf(1/t)},\quad F(t)=\frac{f(t)-t(1/t)}{f(t)+tf(1/t)}. \tag{42}$$

These equations describe the relation between the connection on $\mathcal{W}$ and the Hermitian metric living on the density operators. It is Riemannian iff *f* is self-transposed. (41) yields $f=k$ in this case, and (42) degenerates to $r\equiv 1/2$. Hence, *if the induced Hermitian form is Riemannian, the induced connection is necessarily the canonical one.* This way we do not get an interesting mapping from the class of Riemannian metrics to the class of connections. Especially, the function $f(t)=(1+t)/2$ belonging to the Bures metric cannot be gained from $\mathbf{a}^{\mathrm{geo}}$ as one might expect.

Moreover, if we like to gain the connection form $\mathbf{a}^{\mathrm{geo}}$, $r(t)=t/(t+1)$, belonging to the geometric phase, we need, according to (42), $t^2f(1/t)=f(t)$ or, equivalently, $k(t)=k(1/t)$. If *f* is operator monotone, so is $tf(1/t)$. Therefore, $t^2f(1/t)$ is convex (lemma 5.2 of Ref. 3). Thus, *f* is convex and, as an operator monotone function, concave. Being convex and concave, *f* has to be affine. An affine function on the positive real axis, fulfilling $t^2f(1/t)=f(t)$, is a multiple of *t*.

If $\mathbf{a}=\mathbf{a}^{\mathrm{geo}}$ *and f is operator monotone with* $f(1)=1$, *then* $f(t)=t$.

However, for $k(t)=1$ [respectively, $k(t)=2t/(t+1)$] we get $\mathbf{a}=\mathbf{a}^{\mathrm{geo}}$ (respectively, $\mathbf{a}=\mathbf{a}^{\mathrm{can}}$) and obtain from (38b) for the real part

$$\mathrm{Re}\,(\eta,\xi)_\varrho=\frac{1}{4}\mathrm{Tr}\,\eta\frac{R^{-1}}{f_s(L/R)}\xi \tag{43}$$

with $f_s(t)=(1+t)/2$ (respectively, $f_s(t)=2t/(t+1)$). These f_s are distinguished (self-transposed) operator monotone functions. Moreover, in these cases (38b) restricted to the horizontal vectors coincides with the real part of the Hilbert–Schmidt metric. This is the motivation to deal in the following with the real part of the Hermitian metric induced on the state space.

First of all, this Riemannian metric is of the form (43) with a certain self-transposed function f_s depending on k. From (38b) we get

$$f_s(t)=\frac{k(t)+tk(1/t)}{2}. \tag{44}$$

$f_s(t)$ is the harmonic mean of $f(t)$ and $tf(1/t)$, with f given by (40).

Clearly, in starting with a self-transposed f_s there is some arbitrariness in choosing k respecting (44). Moreover, given a self-transposed f_s, the only restriction for F is $-F(1/t)=F(t)<1$. Indeed, Eqs. (35) and (44) then have the unique solution

$$k(t)=f_s(t)(1-F(t)). \tag{45}$$

In order to remove the arbitrariness in going from f_s to F and vice versa or from f_s to k, we impose an additional requirement on the class (32) of Hermitian metrics $(x,y)_w$. The aim is to ensure that, given f_s, there is only one k and one F fulfilling (35) and (44). We shall prove that we meet our goal for operator monotone f_s by the following natural demand:

Condition HS: For x and y belonging to the horizontal spaces defined by the Hermitian metric (32), the real part, $\mathrm{Re}\,(x,y)_w$, *of* $(x,y)_w$ *coincides with the real part,* $\mathrm{Re}\,(x,y)$, *of the Hilbert–Schmidt product of x and y.*

At first, by the aid of (33), the condition HS becomes

$$\mathrm{Re}\,(k(\Delta)(gw),g'w)=\mathrm{Re}\,(k(\Delta)(gw),k(\Delta)(g'w))$$

with arbitrary Hermitian g and g'. It yields the constraint

$$k(t)+tk(1/t)=k(t)^2+tk(1/t)^2. \tag{46}$$

Next, we have the following crucial observation, which one verifies straightforwardly:

There is a one-to-one correspondence between positive functions k fulfilling the constraint (46) and functions F with $-F(1/t)=F(t)<1$. *The correspondence is given by (35) and*

$$k(t)=\frac{2t(1-F(t)}{(1+F(t))^2+t(1-F(t))^2}. \tag{47}$$

By (44) or, equally well, by (45) we get the relation between F and f_s,

$$f_s(t)=\frac{2t}{(1+F(t))^2+t(1-F(t))^2}. \tag{48}$$

Hence, under condition HS, a function f_s can be gained from a k iff f_s has a representation (48) with a suitable F, $F(t)<1$. To explain which functions f_s can be reached, we rewrite relation (48) into the equivalent form

$$\frac{1+t}{2}-f_s(t)=\frac{f_s(1/t)(1+t)^2}{4}\left(\frac{t-1}{t+1}-F(t)\right)^2.$$

Therefore, necessary conditions for f_s are $f_s(1)=1$, $f_s \leqslant (1+t)/2$ and, moreover, $t \mapsto (1+t)/2 - f_s(t)$ must be the square of a smooth function.

Now suppose we have such a pair f_s, F. We define an auxiliary smooth function

$$\delta(t) := \frac{\sqrt{f_s(1/t)}(1+t)}{2} \left(\frac{t-1}{t+1} - F(t) \right).$$

It fulfills

$$\delta(t)^2 = \frac{1+t}{2} - f_s(t), \quad \sqrt{t}\,\delta(1/t) + \delta(t) = 0. \tag{49}$$

The second equation is a consequence of $F(1/t) = -F(t)$ and $f_s(t) = t f_s(1/t)$. F can be expressed in terms of δ and f_s by

$$F(t) = \frac{t-1}{t+1} - \frac{2}{(1+t)\sqrt{f_s(1/t)}}\,\delta(t). \tag{50}$$

Conversely, for a given self-transposed f_s, $f_s(1)=1$, the possibilities in choosing δ with the properties (49) enumerate via (50) the solutions F of (48) and $-F(1/t) = F(t)$. But such an F may not fulfill $F(t) < 1$ if we did not choose appropriately the signs for δ in (49). The desired choice may be neither unique nor possible. But if so, the function k defined by

$$k(t) := \frac{2}{t+1} \left(f_s(t) + \sqrt{t f_s(t)}\,\delta(t) \right) \tag{51}$$

satisfies (44) and (35).

The question, which functions f_s, $f(1)=1$, bounded by $0 < f(t) \leqslant (1+t)/2$, can arise from F or, equivalently, from a Hermitian metric (32), depends also on regularity requirements on F and k. We do not discuss this in detail. Instead we have the following uniqueness result:

Lemma: For every self-transposed operator monotone function $f_s : (0,\infty) \to \mathbb{R}$ with $f(1)=1$ there exists exactly one positive real analytic function $k : (0,\infty) \to \mathbb{R}$ fulfilling (44) and (46). k and its corresponding function F are given by

$$k(t) = \frac{2 f_s(t)}{t+1} \left(1 + \frac{t-1}{|t-1|} \sqrt{t} \sqrt{\frac{1+t}{2 f_s(t)} - 1} \right), \tag{52}$$

$$F(t) = \frac{t-1}{t+1} \left(1 - \frac{2\sqrt{t}}{|t-1|} \sqrt{\frac{1+t}{2 f_s(t)} - 1} \right) \tag{53}$$

for $t \neq 1$ and $k(1)=1$, $F(1)=0$.

We prove this assertion in the Appendix. (It should be emphasized that k and F are real analytic although the last formulas involve $1/|t-1|$, see the Appendix.) From this lemma we get:

For every monotone Riemannian metric (43), $f_s(1)=1$, on the manifold of completely entangled states there exists exactly one Hermitian metric (32) satisfying the condition HS such that the real part of the induced Hermitian metric is just the given monotone metric. For a given f_s the Hermitian metric and the corresponding connection form are obtained from (52) and (53).

The obtained connection we call the connection associated to the monotone Riemannian metric. For the Bures metric we return to the Hilbert–Schmidt metric and the connection above called $\mathbf{a}^{\text{geo}}$.

Since we used only certain properties of operator monotone functions this assertion would be true for a larger class of metrics, but we will not deal with this problem.

Although the condition HS seems to be natural, perhaps a short comment would be worthwhile. The induced Riemannian metrics are obtained, essentially, by taking the real part of the Hermitian metric of horizontally lifted vectors. But, because of HS, this is the same as the real part of the Hilbert–Schmidt metric. Forgetting for a moment about the underlying Hermitian metric, which forced horizontality, we can take the following point of view: The monotone metrics are obtained from the originally given Hilbert–Schmidt metric similarly to the Bures metric (Sec. IV) whereas the deviation from the Bures metric is caused by some constraints on the purifying lifts.

VIII. EXAMPLES

At first we look at curves of density operators satisfying a von Neumann equation

$$i\dot{\varrho}=[h,\varrho],\quad h=h^*,\quad \dot{h}=0 \tag{54}$$

and their lifts. We may think of $h\in\mathcal{B}(\mathcal{H})$ as of a given Hamiltonian and of the curve parameter, t, as time. This interpretation is not obligatory: h may be the generator of any one-parameter group. (The parameter t should not be confused with the use of the same letter as a dummy variable in several functions like f, k, r, F.) To fix a solution of (54), we start at an initial time, $t_{\rm in}$, with an initial density operator $\varrho_{\rm in}$. The solution may be written

$$\varrho_t=u_t^*\varrho_{\rm in}u_t,\quad u_t:=\exp i(t-t_{\rm in})h. \tag{55}$$

Now a general lift w_t is polar decomposed, $w_t=\sqrt{\varrho_t}v_t$, according to (9).

Our aim is to prove the following: *Given a connection form and an initial* $\varrho_{\rm in}$ *at* $t_{\rm in}$ *there is a t-independent Hermitian* $\tilde{h}$ *such that*

$$u_tv_t=\exp i(t-t_{\rm in})\tilde{h} \tag{56}$$

implies horizontality of w_t. At first we see from (55) and (56) the validity of a Schrödinger equation in $\mathcal{W}$,

$$i\dot{w}=Hw,\quad Hw:=hw-w\tilde{h}. \tag{57}$$

By the help of our menagerie of equations it is not particularly difficult to prove the statement above and to obtain an expression for $\tilde{h}$. At first let us multiply (57) by w^* from the right. By (30) the condition for horizontality is in equating $i\dot{w}w^*$ with $r(R/L)i\dot{\varrho}$. Now (54) yields

$$r(\mathrm{R/L})(h\varrho-\varrho h)=h\varrho-w\tilde{h}w^*.$$

This equation is sufficient to guarantee horizontality. Now $w\tilde{h}w^*$ can be computed by (56) to $u_t^*\sqrt{\varrho_{\rm in}}\tilde{h}\sqrt{\varrho_{\rm in}}u_t$. Therefore, our horizontality condition is the Ad-transform with u_t^* of the equation

$$r(\mathrm{R_{in}/L_{in}})(h\varrho_{\rm in}-\varrho_{\rm in}h)=h\varrho_{\rm in}-\sqrt{\varrho_{\rm in}}\tilde{h}\sqrt{\varrho_{\rm in}},$$

where R and L at $t=t_{\rm in}$ is indexed by in. In other words, if we choose $\tilde{h}$ t-independent and v according to (56), we can satisfy the horizontality condition.

To get a unique $\tilde{h}$, we require the support of $\tilde{h}$ to be smaller than that of $\varrho_{\rm in}$. Finally, with the help of (28), we get the expression

$$\tilde{h}=(\sqrt{\mathrm{R/L}}\,r(\mathrm{L/R})+\sqrt{\mathrm{L/R}}\,r(\mathrm{R/L}))h,\quad t=t_{\rm in}. \tag{58}$$

Let us consider a solution (55) of (54) from $t_{\rm in}$ to $t_{\rm out}$. *Then* $w_{\rm out}w_{\rm in}^*$ *is a gauge invariant.* Its trace in $\mathcal{H}$,

$$(w_{\rm in},w_{\rm out})=(w_{\rm in},[\exp i(t_{\rm out}-t_{\rm in})H]w_{\rm in})=\operatorname{Tr}\sqrt{\varrho_{\rm in}}\sqrt{\varrho_{\rm out}}\exp(i(t_{\rm in}-t_{\rm out})h)\exp(i(t_{\rm out}-t_{\rm in})\tilde{h}), \tag{59}$$

may be called a *relative geometric phase*. For pure states that object has been introduced in Ref. 38. These authors called it the ''non-cyclic geometric phase.'' One may think of shortcutting the in- and the out-state to a closed curve by a Fubini Study geodesic arc. Whether one has a similar interpretation in our much more general case remains an open question.

For a cyclic solution of (54), i.e., $\varrho_{\rm in}=\varrho_{\rm out}$, $t_{\rm cycle}=t_{\rm out}-t_{\rm in}$, the expression $w_{\rm out}w_{\rm in}^*$ is a (pointed) holonomy invariant, i.e., it depends on the choice of $\varrho_{\rm in}$. To change the in-state of our cyclic curve one has to perform a u_t-transformation. Consequently, *all eigenvalues of* $w_{\rm out}w_{\rm in}^*$ *are (absolute) holonomy invariants*. of our cyclic curve. They are encoded in the traces

$$\operatorname{Tr}(w_{\rm out}w_{\rm in}^*)^m=\operatorname{Tr}[\varrho_{\rm in}\exp(-it_{\rm cycle}h)\exp(it_{\rm cycle}\tilde{h})]^m, \tag{60}$$

where $\exp(-it_{\rm cycle}h)$ commutes with $\varrho_{\rm in}$.

There are a few examples where one can become more explicit. One of them is in *adding noise to a curve of pure states* p_t. In this important example one can study the influence of ''noise'' on the geometric phase, and the behavior of gauge and holonomy invariants in coming from the interior to the extreme boundary of the cone of unnormalized density operators. For this purpose we fix two positive real numbers, α and β, and consider the curve of density operators ϱ,

$$\varrho=\alpha p+\beta\mathbf{1},\quad p=|\psi\rangle\langle\psi|,\quad \langle\psi,\psi\rangle=1. \tag{61}$$

$\alpha+\beta$ is a simple and β, if n denotes the dimension of $\mathcal{H}$, a $(n-1)$-fold eigenvalue of ϱ. ψ, p and ϱ depend on a parameter t, but we will not suppose a von Neumann equation.

Remark: The line element of this curve with respect to the metric induced from (32) is

$$ds^2=\frac{2\alpha(1-\tau)}{\tau k(1/\tau)+k(\tau)}ds^2_{\rm Bures},\quad \tau:=\frac{\beta}{\alpha+\beta},$$

where $ds^2_{\rm Bures}$ denotes the Bures line element of the curve of pure states p_t. □

All t-derivations will be indicated by a dot, in particular

$$\dot{\varrho}=\alpha\dot{p},\quad \dot{p}=\dot{p}p+p\dot{p},\quad p\dot{p}p=0.$$

$\dot{\varrho}$ belongs to $\mathcal{T}^\perp$. As an application one calculates

$$R_\varrho\dot{p}=\dot{p}(\alpha p+\beta\mathbf{1})=(\alpha+\beta)\dot{p}p+\beta p\dot{p}.$$

In this manner one gets

$$R_\varrho(p\dot{p})=\beta p\dot{p},\quad R_\varrho(\dot{p}p)=(\alpha+\beta)\dot{p}p,$$

$$L_\varrho(p\dot{p})=(\alpha+\beta)p\dot{p},\quad L_\varrho(\dot{p}p)=\beta\dot{p}p$$

and, finally, skipping the index of L_ϱ and R_ϱ,

$$({\rm L/R})(p\dot{p})=\left(\frac{\alpha+\beta}{\beta}\right)p\dot{p},\quad ({\rm L/R})(\dot{p}p)=\left(\frac{\beta}{\alpha+\beta}\right)\dot{p}p.$$

For instance, $\dot{p}p$ and $\dot{p}p$ are eigenvectors of LR with the eigenvalue $(\alpha+\beta)\beta$. At this stage we do not suppose a von Neumann equation (54) but rely on (31). From the last equation and $F(t)=-F(1/t)$, we get

$$F(\mathrm{L/R})\dot{p}=F\left(\frac{\beta}{\alpha+\beta}\right)(\dot{p}p-p\dot{p}).$$

Hence, in solving (31) with (61) we are faced with an equation

$$\dot{v}v^*=\frac{1}{2}\frac{\alpha}{\sqrt{(\alpha+\beta)\beta}}\left[F\left(\frac{\beta}{\alpha+\beta}\right)+\frac{\sqrt{\alpha+\beta}-\sqrt{\beta}}{\sqrt{\alpha+\beta}+\sqrt{\beta}}\right](p\dot{p}-\dot{p}p), \tag{62}$$

which may be rewritten as

$$\dot{v}^*=v^*(1-\mu)(p\dot{p}-\dot{p}p), \quad \mu=\frac{1}{2}\frac{\alpha}{\sqrt{(\alpha+\beta)\beta}}\left[F\left(\frac{\beta}{\alpha+\beta}\right)+\frac{\alpha+2\beta}{\alpha}\right]. \tag{63}$$

Can we go by $\beta\rightarrow 0$ to the pure states? A necessary condition is

$$F(0)=-1$$

or, equivalently, $r(0)=0$. To be sufficient we additionally need the existence of

$$\kappa:=\lim_{\beta\rightarrow 0}\mu=\lim_{\lambda\rightarrow 0}\frac{1+F(\lambda)}{2\sqrt{\lambda}}=\lim_{\lambda\rightarrow 0}\lambda^{-1/2}r(\lambda). \tag{64}$$

Then the limit $\beta\rightarrow 0$ can be performed in (62):

$$(v\dot{v}^*)^{\mathrm{pure}}=(1-\kappa)(p\dot{p}-\dot{p}p). \tag{65}$$

With $\mathbf{a}^{\mathrm{geo}}$, or, more generally, with $s>1/2$ in $r(\lambda)=\lambda^s/(1+\lambda^s)$, we get $\kappa=0$. *With* $\kappa=0$ *we obtain the Berry phase for pure states.*

Indeed, imposing $\langle\psi,\dot{\psi}\rangle=0$ a la Berry[9] and Fock[27], we find $\dot{v}^*\psi+v^*\dot{\psi}=0$ from (63). Hence, with $\kappa=0$, the vector $v^*\psi$ is t-independent. This yields $w=|\psi\rangle\langle\varphi|$, $\dot{\varphi}=0$. It then follows

$$\mathrm{Tr}\,(w_{\mathrm{out}}w_{\mathrm{in}}^*)^m=\langle\psi_{\mathrm{in}},\psi_{\mathrm{out}}\rangle^m.$$

This is the mth power of the Berry phase, because we had supposed the validity of Berry's transport condition. Remark that this goes not through if $\kappa\neq 0$ or if, as for $\mathbf{a}^{\mathrm{can}}$, (64) does not exist.

Something more can be said if (61) satisfies a von Neumann equation (54). Computing $\tilde{h}$ with this assumption by the help of (58) ends up with

$$\tilde{h}=h+\mu[(\mathbf{1}-p_{\mathrm{in}})hp_{\mathrm{in}}+p_{\mathrm{in}}h(\mathbf{1}-p_{\mathrm{in}})]. \tag{66}$$

Looking at $\tilde{h}$ as a block matrix with respect to p_{in} and $\mathbf{1}-p_{\mathrm{in}}$, the deviation from h is in multiplying the off-diagonal blocks by μ. If (64) exists and $\kappa=0$ then the off-diagonal blocks become zero at the pure state limit.

ACKNOWLEDGMENTS

The authors would like to thank P. M. Alberti, C. Crell, and M. B. Ruskai for valuable remarks. Our particular thanks go to Dènes Petz and Gerd Rudolph for enlightening and helpful discussions. J.D. is grateful to H. B. Rademacher for valuable advice. The authors gratefully acknowledge a stay at the ''Stefan Banach International Mathematical Center,'' Warszawa, and one of us (A.U.) would like to thank the ''Erwin Schrödinger International Institute for Mathematical Physics,'' Vienna, and the ''1997 Elsag-Bailey – I.S.I. Foundation research meeting on quantum computation,'' Torino, where part of the research was done.

APPENDIX: PROOF OF THE LEMMA OF SECTION VII

Every self-transposed operator monotone function f_s with $f_s(1)=1$ has a unique integral representation

$$\begin{aligned} f_s(t) &= m(\{0\})\,\frac{1+t}{2} + \int_{(0,1]} \frac{1+x}{2}\left(\frac{t}{t+x}+\frac{t}{tx+1}\right) dm(x) \\ &= \frac{1+t}{2} + \int_{(0,1]} \left\{ -\frac{1+t}{2} + \frac{1+x}{2}\left(\frac{t}{t+x}+\frac{t}{tx+1}\right)\right\} dm(x) \\ &= \frac{1+t}{2} - (1-t)^2 \int_{(0,1]} \frac{x(t+1)}{2(t+x)(tx+1)}\, dm(x), \end{aligned} \tag{A1}$$

where m is a normalized positive Radon measure on $[0,1]$, see Ref. 3. If the measure is not concentrated at 0, the last integral is strictly positive for all $t\in\mathbb{R}_+$. Its positive root, for the time being denoted by τ, is a real analytic function. Hence, every such function f_s can be represented as

$$f_s(t) = \frac{1+t}{2} - (t-1)^2 \tau(t)^2 \tag{A2}$$

with a certain τ, positive or trivial. Therefore, $(1+t)/2 - f_s(t)$ has exactly two real analytic roots,

$$\delta_+(t) = (t-1)\,\tau(t), \quad \delta_-(t) = -(t-1)\,\tau(t),$$

or is vanishing. The self-transposeness of f_s implies $\tau(1/t)=\sqrt{t}\,\tau(t)$ and both roots fulfill the condition (49). As explained in Sec. VII, a solution for k of our problem corresponds to such a root δ, which leads via (50) to $F(t)<1$. We infer: If selecting the root δ_+, the condition $F(t)<1$, $t>0$, is equivalent to $f_s(t)>1/2$ for all $t>1$. Because f_s is monotone increasing and $f_s(1)=1$ the latter inequality is true. On the other hand, F cannot fulfill $F(t)<1$ for all $t>1$ if the root δ_- is chosen, except $\delta_-=0$. Otherwise we could conclude $f_s(t)>t/2$ for all $t>1$. But the self-transposeness effects $f_s'(1)=1/2$ and f_s must be concave. Therefore, $\delta := \delta_+$ is the only real analytic root leading to an appropriate F. Inserting

$$\delta(t) = (t-1)\,\tau(t) = \frac{t-1}{|t-1|}\sqrt{\frac{1+t}{2} - f_s(t)}, \quad \delta(1)=0, \tag{A3}$$

into formulas (50), (51) yields (53) and (52).

[1] D. Petz, Linear Algebr. Appl. **244**, 81 (1996).
[2] D. Petz and C. Sudar, J. Math. Phys. **37**, 2662 (1996).
[3] F. Kubo and T. Ando, Math. Ann. **246**, 205 (1980).
[4] W. Donoghue, *Monotone Matrix Functions and Analytic Continuation* (Springer, Berlin, 1974).
[5] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).
[6] R. S. Ingarden, H. Janyszek, A. Kossakowski, and T. Kawaguchi, Tensor **37**, 105 (1982).
[7] M. Ohya and D. Petz, *Quantum Entropy and Its Use*, Texts and Monographs in Physics (Springer, Berlin, 1993).
[8] A. Uhlmann, Rep. Math. Phys. **24**, 229 (1986).
[9] M. V. Berry, Proc. R. Soc. London, Ser. A **392**, 45 (1984).
[10] B. Simon, Phys. Rev. Lett. **51**, 2167 (1983).
[11] Y. Aharonow and J. Anandan, Phys. Rev. Lett. **58**, 1593 (1987).
[12] J. Anandan and L. Stodolsky, Phys. Rev. D **35**, 2597 (1987).
[13] A. Uhlmann, Lett. Math. Phys. **21**, 229 (1991).
[14] A. Uhlmann, in *Quantum Groups and Related Topics*, edited by R. Gielerak *et al.* (Kluwer, Dordrecht, 1992), pp. 267–274.
[15] M. Hübner, Phys. Lett. A **163**, 239 (1992).
[16] A. Uhlmann, Phys. Lett. A **161**, 329 (1992).
[17] J. Dittmann, Sem. S. Lie **3**, 73 (1993).

[18] M. Hübner, Phys. Lett. A **179**, 226 (1993).
[19] S. L. Braunstein and C. M. Caves, Phys. Rev. Lett. **72**, 3439 (1994).
[20] C. A. Fuchs and C. M. Caves, Open Sys. Inf. Dyn. **3**, 1 (1995).
[21] J. Dittmann and G. Rudolph, J. Geom. Phys. **10**, 93 (1992).
[22] G. Rudolph and T. Tok, Rep. Math. Phys. **39**, 433 (1997).
[23] J. Dittmann, Lett. Math. Phys. **46**, 281 (1998).
[24] F. Wilczek and A. Zee, Phys. Rev. Lett. **52**, 2111 (1984).
[25] J. Dittmann and G. Rudolph, J. Math. Phys. **33**, 4148 (1992).
[26] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1969), Vol II.
[27] V. Fock, Z. Phys. **49**, 323 (1928).
[28] *Geometric Phases in Physics*, edited by A. Shapere and F. Wilczek (World Scientific, Singapore, 1990); *Anomalies, Phases, Defects*, edited by M. Bregola, G. Marmo, and G. Morandi (Bibliopolis, Naples, 1990); *Topological Phases in Quantum Theory*, edited by B. Markowski and S. I. Vinitski (World Scientific, Singapore, 1989).
[29] D. J. C. Bures, Trans. Am. Math. Soc. **135**, 199 (1969).
[30] S. Kakutani, Ann. Math. **49**, 214 (1948).
[31] L. Dabrowski and H. Grosse, Lett. Math. Phys. **19**, 205 (1990).
[32] A. Uhlmann, Ann. Phys. (Leipzig) **46**, 63 (1989).
[33] R. Bhatia and P. Rosenthal, Bull. London Math. Soc. **29**, 1 (1997).
[34] H. Hasegawa, Rep. Math. Phys. **33**, 87 (1993).
[35] D. Petz and C. Hasegawa, Lett. Math. Phys. **38**, 221 (1996).
[36] D. Petz, Linear Algebr. Appl. **244**, 81 (1996).
[37] R. A. Fisher, Proc. Cambridge Philos. Soc. **22**, 700 (1925).
[38] J. Samuel and R. Bhandari, Phys. Rev. Lett. **60**, 2339 (1988).

On form-preserving transformations for the time-dependent Schrödinger equation

Federico Finkel and Artemio González-López
Departamento de Física Teórica II, Universidad Complutense de Madrid, E-28040 Madrid, Spain

Niky Kamran
Department of Mathematics and Statistics, McGill University, Montréal, Québec H3A 2K6, Canada

Miguel A. Rodríguez
Departamento de Física Teórica II, Universidad Complutense de Madrid, E-28040 Madrid, Spain

(Received 22 December 1998; accepted for publication 18 March 1999)

In this paper we point out a close connection between the Darboux transformation and the group of point transformations which preserve the form of the time-dependent Schrödinger equation (TDSE). In our main result, we prove that any pair of time-dependent real potentials related by a Darboux transformation for the TDSE may be transformed by a suitable point transformation into a pair of time-independent potentials related by a usual Darboux transformation for the stationary Schrödinger equation. Thus, any (real) potential solvable via a time-dependent Darboux transformation can alternatively be solved by applying an appropriate form-preserving point transformation of the TDSE to a time-independent potential. The pre-eminent role of the latter type of transformations in the solution of the TDSE is illustrated with a family of quasi-exactly solvable time-dependent anharmonic potentials. © *1999 American Institute of Physics.*
[S0022-2488(99)00207-8]

I. INTRODUCTION

A considerable amount of research has been devoted over the past few years to the exact solution of the time-dependent Schrödinger equation (TDSE) in $1+1$ dimensions. Several modifications of the celebrated Darboux transformation for the stationary Schrödinger equation,[1,2] have been proposed in this respect in the literature. Matveev and Salle showed that the usual Darboux transformation for the stationary Schrödinger equation could also be applied to the TDSE with a time-dependent potential.[3] An equivalent approach was followed by Bluman and Shtelen in Ref. 4, who considered a nonlocal transformation which is precisely the inverse map of the usual Darboux transformation. The Darboux transform of a time-dependent potential is in general a complex-valued function. (The explicit conditions for the resulting potential to be real-valued appear in a recent paper by Bagrov and Samsonov.[5]) For this reason, several generalizations of the Darboux transformation mapping real potentials to real potentials have been proposed in the literature. The best known of these generalizations is the binary Darboux transformation described in Ref. 3, which is in fact one of the main tools for finding exact solutions of integrable equations.

A seemingly unrelated method of constructing exact solutions of the TDSE which has proved remarkably successful is based on the use of point transformations which preserve the form of the TDSE. The idea goes back to the work of Leach on the time-dependent harmonic oscillator,[6] arising, e.g., in the study of the motion of charged particles in a Paul trap.[7] The method was subsequently extended by Bluman[8,9] and Ray[10] to obtain exact solutions of the TDSE for a quadratic potential with arbitrary time-dependent coefficients. The technique has also been applied to time-dependent harmonic oscillators with a repulsive barrier,[11] and to anisotropic time-dependent harmonic potentials in $2+1$ dimensions.[12]

The main purpose of this paper is to characterize the most general time-dependent real potential whose Darboux transform is real. A partial result along these lines for potentials rapidly decreasing at spatial infinity was mentioned in Ref. 13. As noted in this reference, the latter potentials are of limited interest regarded as solutions of the KP equation. However, potentials of this type (and, more generally, of the type considered in this paper) are interesting from the point of view of exactly solving the time-dependent Schrödinger equation, as underscored by the recent work of Bagrov and Samsonov.[5,14]

In this paper we show that the Darboux transformation for the TDSE is in fact closely related to the point transformation method. To this end, in Sec. II we briefly review the Darboux transformation and the point transformations preserving the form of the TDSE, applying the latter to construct a time-dependent anharmonic oscillator potential admitting a certain number of algebraically computable wave functions. In Sec. III we derive the main results of our paper, proving that any time-dependent real-valued potential for which the Darboux transformation yields a real potential may always be mapped to a time-independent potential by a form-preserving point transformation of the TDSE. Moreover, the Darboux transformation for any such potential is equivalent to a Darboux transformation for its associated time-independent potential, followed by the inverse of the corresponding point transformation. Finally, Sec. IV is devoted to our concluding remarks and related open questions.

II. GENERAL BACKGROUND

In this section we summarize the basics of the Darboux and the form-preserving point transformations for the TDSE. Following Bagrov and Samsonov,[5] we take as the starting point of the Darboux transformation for the TDSE the intertwining relation

$$L(i\partial_t - H_0) = (i\partial_t - H_1)L, \tag{1}$$

where

$$H_i = -\partial_x^2 + V_i(x,t), \quad i = 0,1, \tag{2}$$

and L is a first-order differential operator of the form

$$L = L_1(x,t)\partial_x + L_0(x,t).$$

It follows immediately from the intertwining relation (1) that if ψ_0 solves the TDSE with Hamiltonian H_0, then $\psi_1 = L\psi_0$ will solve the TDSE with Hamiltonian H_1. It is also easily verified that the intertwining relation (1) will be satisfied if and only if

$$L = L_1 \cdot (\partial_x + \chi_x), \quad V_1 = V_0 + 2\chi_{xx} + i(\log L_1)_t, \tag{3}$$

where $e^{-\chi}$ is a solution of the TDSE with the potential V_0, and $L_1 = L_1(t)$ is an arbitrary function. The transformed potential $V_1(x,t)$ is a real-valued function if and only if

$$\operatorname{Im} \chi_{xxx} = 0 \tag{4}$$

and

$$|L_1| = \exp\left[-2\int_{t_0}^{t} \operatorname{Im} \chi_{xx}(x,s)\,ds\right]. \tag{5}$$

Without loss of generality, we shall assume from now on that L_1 is real and positive, and is therefore given by the right-hand side of (5).

Just as in the time-independent case, the Darboux transformation for the TDSE can be inverted. Indeed, if ψ_1 is a solution of the TDSE with potential V_1 given by (3), the function

$$\psi_0(x,t)=\frac{e^{-\chi(x,t)}}{L_1(t)}\left[\int_{x_0}^{x}e^{\chi(y,t)}\psi_1(y,t)dy+c_0(t)\right] \tag{6}$$

with $c_0(t)$ given by

$$c_0(t)=iL_1(t)\int_{t_0}^{t}\frac{e^{\chi(x_0,s)}}{L_1(s)}(\psi_{1,x}(x_0,s)-\chi_x(x_0,s)\psi_1(x_0,s))ds$$

solves the TDSE with potential V_0. If the factor L_1 is taken as unity, the mapping $\psi_1\mapsto\psi_0$ given by (6) reduces to the nonlocal transformation considered by Bluman and Shtelen in Ref. 4.

The most general point transformation mapping the TDSE

$$(i\partial_t+\partial_x^2-V_0(x,t))\psi_0(x,t)=0 \tag{7}$$

for any given potential V_0 into another TDSE with potential $\overline{V}_0$ for the transformed wave function $\overline{\psi}_0$ is defined by

$$\overline{x}=\frac{x}{C(t)}+B(t),\quad \overline{t}=\int_{t_0}^{t}\frac{ds}{C^2(s)},$$

$$\psi_0(x,t)=|C|^{-1/2}\exp\left[\frac{i}{4}\left(\frac{\dot{C}}{C}x^2-2\dot{B}Cx+A(t)\right)\right]\overline{\psi}_0(\overline{x},\overline{t}), \tag{8}$$

$$V_0(x,t)=\frac{1}{C^2}\overline{V}_0(\overline{x},\overline{t})-\frac{\ddot{C}}{4C}x^2+\left(\frac{C\ddot{B}}{2}+\dot{B}\dot{C}\right)x-\frac{1}{4}(C^2\dot{B}^2+\dot{A}),$$

where *A*, *B*, and $C\neq0$ are real-valued functions of *t*. Note that square-integrability is preserved under the transformation (8). As remarked before, the point transformation (8) has been employed to construct exact solutions of the TDSE for quadratic potentials with time-dependent coefficients. The interest of the transformation (8) is not limited, however, to quadratic (or exactly solvable) time-dependent potentials, as evidenced by the following example:

Example: Consider the two-parameter family of anharmonic oscillator potentials given by[15–17]

$$\overline{V}(\overline{x})=\overline{x}^6+2\alpha\overline{x}^4+(\alpha^2-4n-3)\overline{x}^2, \tag{9}$$

where $\alpha\in\mathbb{R}$ and $n\in\mathbb{N}$. The sextic potential (9) is a well-known example of the class of *quasi-exactly solvable* potentials, for which a certain subset of the spectrum can be computed by purely algebraic means; see Ref. 18 for an extensive review of the field. The first $n+1$ even bound states of the potential (9) are of the form

$$\phi(\overline{x})=\exp\left[-\frac{1}{4}\overline{x}^4-\frac{\alpha}{2}\overline{x}^2\right]p(\overline{x}^2), \tag{10}$$

where $p(s)$ is a polynomial in s of degree less than or equal to n which can be computed algebraically. The point transformation (8) with $A=B=0$ and $C=\omega^{-1/2}$, $\omega=\omega(t)$ being a positive function, leads directly to the potential $V(x,t)$ given by

$$V(x,t)=\omega^4x^6+2\alpha\omega^3x^4+\left(\alpha^2-4n-3-\frac{3\dot{\omega}^2-2\omega\ddot{\omega}}{16\omega^4}\right)\omega^2x^2. \tag{11}$$

The TDSE with potential (11) possesses $n+1$ square-integrable solutions of the form

$$\psi(x,t)=\omega^{1/4}\exp\left[-i\left(\frac{\dot{\omega}}{8\omega}x^2+E\int_{t_0}^{t}\omega(s)ds\right)\right]\phi(\sqrt{\omega}x), \tag{12}$$

where $\phi(\overline{x})$ is an algebraic eigenfunction of the form (10) with eigenvalue E of the Hamiltonian

$$\overline{H}=-\partial_x^2+\overline{V}(\overline{x}).$$

The potential (11) thus provides a natural extension of the notion of quasi-exact solvability to the time-dependent case, in the sense that the associated TDSE admits a certain number of solutions which can be determined algebraically. In particular, note that if $\omega(t)$ is of the form

$$\omega_0(t)=\beta[\gamma+(\gamma^2-(\alpha^2-4n-3)\beta)^{1/2}\sin(4\sqrt{\beta}t+\delta)]^{-1},$$

with $\delta\in\mathbb{R}$, $\beta>0$ and $\gamma^2>(\alpha^2-4n-3)\beta>0$, the potential (11) reduces to a harmonic oscillator with a periodic-in-time anharmonic perturbation, namely,

$$V_0(x,t)=\omega_0^4(t)x^6+2\alpha\omega_0^3(t)x^4+\beta x^2.$$

III. THE REALITY CONDITION AND THE FORM-PRESERVING POINT TRANSFORMATIONS

In this section we prove the main results of our paper, starting with the following theorem:

Theorem: Let $e^{-\chi}$ be a solution of the TDSE with potential $V_0(x,t)$. If χ satisfies the reality condition (4), then $V_0(x,t)$ may be mapped to a time-independent potential $\overline{V_0}(\overline{x})$ by a point transformation (8).

Proof: Let $\chi_0=\mathrm{Re}\,\chi$, $\chi_1=\mathrm{Im}\,\chi$. The TDSE for $e^{-\chi}$ is then equivalent to the pair of real equations given by

$$\chi_{0,t}+\chi_{1,xx}-2\chi_{0,x}\chi_{1,x}=0, \tag{13}$$

$$V_0(x,t)=\chi_{1,t}-\chi_{0,xx}-\chi_{1,x}^2+\chi_{0,x}^2. \tag{14}$$

If the reality condition (4) holds, i.e., if χ_1 is of the form

$$\chi_1=a(t)x^2+b(t)x+c(t), \tag{15}$$

Eqs. (13) and (14) reduce to

$$\chi_{0,t}-2(2ax+b)\chi_{0,x}+2a=0, \tag{16}$$

$$V_0(x,t)=(\dot{a}-4a^2)x^2+(\dot{b}-4ab)x+\dot{c}-b^2+\chi_{0,x}^2-\chi_{0,xx}. \tag{17}$$

The general solution of Eq. (16) is of the form[19]

$$\chi_0=-2\int_{t_0}^{t}a(s)ds+F\left(e^{4\int_{t_0}^{t}a(s)ds}x+2\int_{t_0}^{t}b(s)e^{4\int_{t_0}^{s}a(r)dr}ds\right), \tag{18}$$

where F is an arbitrary real-valued function. Substituting this expression into (17), we immediately conclude that the transformation (8) determined by

$$C(t)=e^{-4\int_{t_0}^{t}a(s)ds},\quad B(t)=2\int_{t_0}^{t}\frac{b(s)}{C(s)}ds,\quad A(t)=-4c(t), \tag{19}$$

maps the potential $V_0(x,t)$ into the time-independent potential

$$\overline{V}_0(\overline{x}) = F'^2(\overline{x}) - F''(\overline{x}). \qquad \text{Q.E.D.}$$

Thus, any potential for which the Darboux transformation yields another real-valued potential may be mapped into a time-independent potential by a form-preserving point transformation. It is easy to check that the transform of $e^{-\chi}$ under the point transformation defined by (8), (15), (18), (19) is e^{-F}, which is therefore an eigenfunction of the time-independent potential $\overline{V}_0(\overline{x})$. The next Corollary shows that the usual Darboux transform of the associated time-independent potential generated by F is related to the Darboux transform of the original potential by the same point transformation.

Corollary: Let $e^{-\chi} = e^{-\chi_0 - i\chi_1}$ be a solution of the TDSE with potential $V_0(x,t)$, with χ satisfying the reality condition (4). Let $\mathcal{D}$ and $\mathcal{T}$ denote, respectively, the Darboux transformation (3) and the point transformation (8) defined by χ via Eqs. (15), (18), (19). Let $\overline{\mathcal{D}}$ denote the Darboux transformation generated by F. Then

$$\mathcal{T} \circ \mathcal{D} = \overline{\mathcal{D}} \circ \mathcal{T}. \tag{20}$$

Remark: This result may be easily visualized with the help of the following commutative diagram:

$$\begin{array}{ccc} V_0(x,t) & \xrightarrow{\mathcal{T}} & \overline{V}_0(\overline{x}) = F'^2 - F'' \\ \mathcal{D}\downarrow & & \downarrow\overline{\mathcal{D}} \\ V_1(x,t) & \xrightarrow{\mathcal{T}} & \overline{V}_1(\overline{x}) = F'^2 + F'' \end{array}.$$

Proof: The proof follows from a straightforward application of the appropriate formulas for the transformed potentials and wave functions. Q.E.D.

The above Corollary shows, in particular, that the potential $V_1(x,t)$ is the image under the inverse of the form-preserving point transformation $\mathcal{T}$ determined by (8), (15), (18), (19) of a time-independent potential $\overline{V}_1(\overline{x})$. Exact solutions of the TDSE with potential V_1 can therefore be obtained simply by applying the point transformation $\mathcal{T}^{-1}$ to solutions of the TDSE for the *time-independent* potential $\overline{V}_1$.

Another important consequence of the above Corollary is that, if the potential $\overline{V}_0$ satisfies (for instance) the condition

$$\int_{-\infty}^{\infty} |\overline{V}_0(\overline{x})|(1 + |\overline{x}|)\, d\overline{x} < \infty, \tag{21}$$

then the time-dependent Darboux transformation (3) preserves the square-integrability of eigenfunctions. Indeed, if $\overline{V}_0$ verifies (21) then the time-independent Darboux transformation $\overline{\mathcal{D}}$ determined by a nonvanishing eigenfunction of $\overline{V}_0$ preserves square integrability.[20,21] The result stated above then follows easily from (20), the invertibility of $\mathcal{T}$, and the fact that the form-preserving point transformation $\mathcal{T}$ always preserves square integrability.

Example: It is straightforward to verify that all the examples of time-dependent potentials appearing in Refs. 4, 5, 14 which are solvable by means of a Darboux transformation are indeed the images of certain exactly solvable time-independent potentials under suitable form-preserving point transformations.

For instance, the free-particle potential $V_0(x,t) = 0$ admits a one-parameter family of solutions

$$\psi_\lambda(x,t) = (1+t^2)^{-1/4} \exp\left[\frac{i}{4}\left(\frac{tx^2}{1+t^2} + 4\lambda \arctan t\right)\right] Q_\lambda(x/\sqrt{1+t^2})$$

satisfying the reality condition (4), where Q_λ is a (real-valued) solution of Weber's equation

$$Q''_\lambda(y)-\left(\frac{y^2}{4}+\lambda\right)Q_\lambda(y)=0 \tag{22}$$

(see Ref. 14, Eq. (18)). By the Theorem at the beginning of this section, it follows that V_0 is related to a certain time-independent potential $\overline{V}_0$ by a point transformation (8). Indeed, in this case we have

$$\chi_1=-\operatorname{Im}\log\psi_\lambda=-\frac{tx^2}{4(1+t^2)}-\lambda\arctan t,$$

so from (19) it follows that

$$C=\sqrt{1+t^2},\quad B=0,\quad A=4\lambda\arctan t. \tag{23}$$

Substituting these formulas into (8) we find that

$$\overline{x}=\frac{x}{\sqrt{1+t^2}} \tag{24}$$

and

$$\overline{V}_0(\overline{x})=\frac{\overline{x}^2}{4}+\lambda \tag{25}$$

is a harmonic oscillator potential.

For all $n\in\mathbb{N}$, let H_n denote the nth Hermite polynomial. The functions[22]

$$Q_{n+1/2}(y)=i^n e^{y^2/4}H_n(iy/\sqrt{2})$$

are real-valued solutions of Weber's equation (22) with $\lambda=n+\frac{1}{2}$ without zeros on the positive real semiaxis.[23] Hence, for all $n\in\mathbb{N}$ the Darboux transformation determined by the eigenfunction $\psi_{n+1/2}$ of V_0 is well-defined on the positive real semiaxis. From Eqs. (3), (5) and the definition of ψ_λ, it follows that the transformed potential V_1 is given by

$$V_1(x,t)=2\chi_{0,xx}=-2[\log Q_{n+1/2}(\overline{x})]_{xx}=\frac{2}{1+t^2}\left(\frac{Q'^2_{n+1/2}(\overline{x})}{Q^2_{n+1/2}(\overline{x})}-\frac{\overline{x}^2}{4}-n-\frac{1}{2}\right),$$

where $\overline{x}$ is given by (24). Using standard identities for the derivatives of the Hermite polynomials, the reader can easily verify that this formula for $V_1(x,t)$ agrees with the corresponding expression given in Ref. 14. As stated in the Corollary, the potential $V_1(x,t)$ is related by the point transformation (8) defined by (23) to a time-independent potential $\overline{V}_1(\overline{x})$ obtained from $\overline{V}_0(\overline{x})$ by applying a time-independent Darboux transformation $\overline{\mathcal{D}}$. From (18) and the definition of ψ_λ, it easily follows that the function $F(\overline{x})$ generating the Darboux transformation $\overline{\mathcal{D}}$ is given by

$$F(\overline{x})=-\log Q_{n+1/2}(\overline{x}),$$

and therefore

$$\overline{V}_1(\overline{x})=F'^2(\overline{x})+F''(\overline{x})=2\frac{Q'^2_{n+1/2}(\overline{x})}{Q^2_{n+1/2}(\overline{x})}-\frac{\overline{x}^2}{4}-n-\frac{1}{2}.$$

It is straightforward to check that the potentials $V_1(x,t)$ and $\overline{V}_1(\overline{x})$ are indeed related by the point transformation (8) determined by (23).

IV. CONCLUSIONS

In this paper, we have shown that the Darboux transformation for the time-dependent Schrödinger equation is essentially equivalent to the usual Darboux transformation for the stationary Schrödinger equation. Any (real) potential $V_1(x,t)$ solvable via a time-dependent Darboux transformation starting from a real potential $V_0(x,t)$ can alternatively be solved by applying a form-preserving point transformation to a time-independent potential $\overline{V_1}(\overline{x})$. As a matter of fact, although a large number of methods and papers have been devoted in recent times to the exact solution of the TDSE,[24] most of the associated time-dependent Hamiltonians either are not of the standard form (2), or are also obtainable from a time-independent Hamiltonian by a form-preserving point transformation.

The interest of the Darboux transformation for the TDSE as a useful method to obtain new (quasi-)exactly solvable time-dependent potentials is therefore very limited. It should be noted, however, that the Darboux transformation for the TDSE could still render helpful results if the starting potential is not a real-valued function but only the transformed potential is real.[25] As a final remark, we would like to stress that the Darboux transformation may still be useful to construct exact solutions to real-valued diffusion equations of the Fokker–Planck-type, for which no reality condition as Eq. (4) must be considered.

ACKNOWLEDGMENTS

F.F., A.G.-L., and M.A.R. would like to acknowledge the partial financial support of the DGICYT under Grant No. PB95-0401. N.K. was supported in part by NSERC Grant No. 0GP0105490.

[1] G. Darboux, C. R. Acad. Sci. Paris **94**, 1456 (1882).
[2] F. Cooper, A. Khare, and U. Sukhatme, Phys. Rep. **251**, 267 (1995).
[3] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons* (Springer, Berlin, 1991).
[4] G. Bluman and V. Shtelen, J. Phys. A **29**, 4473 (1996).
[5] V. G. Bagrov and B. F. Samsonov, Phys. Lett. A **210**, 60 (1996).
[6] P. G. L. Leach, J. Math. Phys. **18**, 1902 (1977).
[7] L. S. Brown, Phys. Rev. Lett. **66**, 527 (1991).
[8] G. Bluman, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **39**, 238 (1980).
[9] G. Bluman, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **43**, 1259 (1983).
[10] J. R. Ray, Phys. Rev. A **26**, 729 (1982).
[11] R. S. Kaushal and D. Parashar, Phys. Rev. A **55**, 2610 (1997).
[12] J. R. Burgan, M. R. Feix, E. Fijalkow, and A. Munier, Phys. Lett. A **74**, 11 (1979).
[13] M. Boiti, F. Pempinelli, A. K. Pogrebkov, and M. C. Polivanov, Inverse Probl. **7**, 43 (1991).
[14] V. G. Bagrov and B. F. Samsonov (preprint, quant-ph/9709040).
[15] V. Singh, S. N. Biswas, and K. Datta, Phys. Rev. D **18**, 1901 (1978).
[16] A. V. Turbiner and A. G. Ushveridze, Phys. Lett. A **126**, 181 (1987).
[17] A. González-López, N. Kamran, and P. J. Olver, Commun. Math. Phys. **153**, 117 (1993).
[18] A. G. Ushveridze, *Quasi-Exactly Solvable Models in Quantum Mechanics* (IOP, Bristol, 1994).
[19] Note that Eq. (13) cannot in general be solved in closed form if χ_1 is a polynomial in x of degree $k>2$.
[20] P. Deift and E. Trubowitz, Commun. Pure Appl. Math. **32**, 121 (1979).
[21] J. M. Sparenberg and D. Baye, J. Phys. A **28**, 5079 (1995).
[22] Our choice of $Q_{n+1/2}$ differs from the one in Ref. 14 by the irrelevant constant factor i^n.
[23] If n is even $Q_{n+1/2}$ is nonzero on the whole real axis.
[24] A comprehensive review of these methods is beyond the scope of this paper; see Refs. 26–30, and references therein for a detailed treatment.
[25] Some families of time-independent *complex* potentials which are quasi-exactly solvable have been recently proposed in the literature (Ref. 31).
[26] W. Magnus, Commun. Pure Appl. Math. **7**, 649 (1954).
[27] J. Wei and E. Norman, J. Math. Phys. **4**, 575 (1963).
[28] D. R. Truax, J. Math. Phys. **22**, 1959 (1981).
[29] C. M. Cheng and P. C. W. Fung, J. Phys. A **21**, 4115 (1988).
[30] S. Zhang and F. Li, J. Phys. A **24**, 6143 (1996).
[31] C. M. Bender and S. Boettcher (preprint, physics/9801007).

Lifetimes of impurity states in crossed magnetic and electric fields

Sébastien Gyger and Philippe A. Martin
Institut de Physique Théorique, Ecole Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland

(Received 12 January 1998; accepted for publication 9 February 1999)

We study the quantum dynamics of localized impurity states created by a point interaction for an electron moving in two dimensions under the influence of a perpendicular magnetic field and an in-plane weak electric field. All impurity states are unstable in presence of the electric field. Their lifetimes are computed and shown to grow in a Gaussian way as the electric field tends to zero. © *1999 American Institute of Physics.* [S0022-2488(99)01807-1]

I. INTRODUCTION

A detailed understanding of the dynamics of electrons in two dimensions (2D) in crossed magnetic and electric fields and in the presence of impurity scattering plays an important role in the context of the quantum Hall effect and transport theory.

We consider an electron of charge e and mass m confined to the two-dimensional plane (x,y) (without boundaries). A uniform magnetic field of magnitude B acts perpendicular to the plane and an electric field of magnitude E acts in the x direction. Moreover, an impurity located at the origin scatters the electron with a short-range potential $V(x,y)$. When $E=0$, the electron remains localized in the course of the time, both for the classical and the quantum dynamics. Classically, modeling for instance the interaction with the impurity by a hard disk of radius a, the electron accomplishes the usual circular cyclotronic motion outside of the disk or bounces around the surface of the disk by a succession of segments of cyclotronic motion. Quantum mechanically, the impurity potential, considered as a short-range perturbation of the pure magnetic problem (the Landau Hamiltonian), will preserve the essential spectrum (i.e., the infinitely degenerated Landau levels) and create at most point spectrum in-between these levels (and possibly below them if V is attractive).

When the electric field is applied, the situation changes. In the classical case, the center of the cyclotronic orbit acquires a constant drift of velocity E/B in the y direction: the trajectories around the disk get distorted by the acceleration imposed by the electric field and may eventually leave the disk. In the quantum case, the Hamiltonian of the homogeneous system with the crossed magnetic and electric fields has an absolutely continuous spectrum on $\mathbb{R}$ and this spectrum will remain present after the introduction of the short-range impurity potential, in view of general theorems on the stability of the absolutely continuous spectrum under such perturbations. An interesting result has recently been obtained for the classical dynamics: for nonzero but sufficiently small electric field, there exists a set of positive Lebesgue measure of trajectories that remain trapped near the disk.[1] The corresponding quantum mechanical question is whether point spectrum can survive the switching on of the electric field and remain embedded into the continuous spectrum, provided that this field is sufficiently weak. To our knowledge there is no definite answer to this question at the moment. The existence of an appreciable domain of the classical phase space supporting localized trajectories may be an argument in favor of a corresponding localized quantum state, but quantum interferences and tunneling phenomena may invalidate this anticipation.

In this article, we bring a partial contribution to this problem by considering a point impurity acting as a δ-potential. This model has been introduced by Prange[2] in relation with the quantum Hall effect (with the electron in a finite strip). Our contribution consists of a proof that all localized states created by the impurity are turned into resonances and of an exact determination of their lifetimes, shown to be of the order of $\exp((|B|/|e|\hbar)(\Delta/E)^2)$ as $E\to 0$ where Δ is the distance of the resonance energy to the closest Landau level (for a more precise formula, see Sec. IV). However, if one allows for an impurity interaction which is not of potential form (i.e., an integral kernel in configuration space), one can have a point spectrum embedded into the continuum. We give an example with a rank one perturbation (Sec. V). A closely related analysis by E. H. Hauge and J. M. J. van Leeuwen[3] will be discussed in our concluding remarks.

II. THE MODEL

We denote H^0 the 2D Hamiltonian of the electron in the crossed fields with potential vector $\mathbf{A}=(0,Bx)$ in the Landau gauge

$$H^0=\frac{1}{2}p_x^2+\frac{1}{2}(p_y-x)^2-\mu x,\quad p_x=-i\frac{\partial}{\partial x},\quad p_y=-i\frac{\partial}{\partial y}. \tag{1}$$

Here H^0 is written in dimensionless variables by choosing $\sqrt{\hbar/|eB|}$ as the unit of length and $\hbar|eB|/m$ as the unit of energy, with $\mu=E\sqrt{m^2/|eB^3|\hbar}$. Since p_y commutes with H^0, there is a direct integral decomposition of $H^0=\int^{\oplus}dk\,H_k^0$, with

$$H_k^0=-\frac{1}{2}\frac{d^2}{dx^2}+\frac{1}{2}(k-x)^2-\mu x,\quad k\in\mathbb{R}. \tag{2}$$

The spectral representation of H^0 is explicitly given in terms of the generalized eigenfunctions

$$(xy|nk\rangle=\frac{1}{\sqrt{2\pi}}e^{iky}u_n(x-k-\mu),\quad n=0,1,\ldots,\quad k\in\mathbb{R}, \tag{3}$$

and spectral branches

$$\epsilon_n(k)=\left(n+\frac{1}{2}\right)-\mu k-\frac{\mu^2}{2}. \tag{4}$$

In (3), u_n are the usual normalized eigenfunctions of a harmonic oscillator of frequency equal to one,

$$u_n(x)=\left(\frac{1}{\sqrt{\pi}2^n n!}\right)^{1/2}e^{-(1/2)x^2}H_n(x), \tag{5}$$

with H_n the nth Hermite polynomial. For $\mu\neq 0$, the branches $\epsilon_n(k)$ are linear functions of k, thus without points of constancy, implying that the spectrum of H^0 is absolutely continuous on $\mathbb{R}$.[4] If $\mu=0$, $\epsilon_n(k)=n+\frac{1}{2}$ are constant, the spectrum reduces to the infinitely degenerated Landau levels, and k labels the degeneracy in the corresponding subspaces.

The total Hamiltonian H is obtained by formally adding to H^0 the singular potential $V(x,y)=\lambda\delta(x,y)$, where $\delta(x,y)$ is the two-dimensional Dirac function. It is well known that this singularity is too strong in 2D and a renormalization of the coupling constant is needed.[5–7] Introducing the resolvents $R_z=(H-z)^{-1}$ and $R_z^0=(H^0-z)^{-1}$, $z\in C$, and solving the resolvent equation $R_z=R_z^0-R_z^0VR_z$ leads to

$$R_z=R_z^0-\frac{R_z^0|00)(00|R_z^0}{g(z)}, \tag{6}$$

$$g(z)=\lambda^{-1}+(00|R_z^0|00), \tag{7}$$

and from the spectral decomposition (3) of H^0

$$(00|R_z^0|00)=\sum_{n=0}^{\infty}\int dk\frac{|(00|nk)\rangle|^2}{\epsilon_n(k)-z}. \tag{8}$$

The n-summation in (8) diverges logarithmically since $\epsilon_n(k)\sim n,\ n\to\infty$. Following the same procedure as in Ref. 7, it can be made finite by substracting out the diverging part and combining it with the coupling constant [using also the normalization of the functions (3)]

$$g(z)=\lim_{N\to\infty}\left(\lambda^{-1}+\frac{1}{2\pi}\sum_{n=0}^{N}\frac{1}{n+\frac{1}{2}}\right)+\lim_{N\to\infty}\sum_{n=0}^{N}\int dk\,|(00|nk)|^2\left(\frac{1}{\epsilon_n(k)-z}-\frac{1}{n+\frac{1}{2}}\right). \tag{9}$$

In the first term of (9) one chooses $\lambda=\lambda_N<0$ be N-dependent and negative, and requires that $\lambda_N\to 0$ in such a way that $\lim_{N\to\infty}(\lambda_N^{-1}+(1/2\pi)\Sigma_{n=0}^{N}[1/(n+1/2)])=\lambda_r$, where λ_r is a finite renormalized coupling constant. The limit of the second term in (9) exists (see Appendix A) so that the model is defined by its resolvent (6) setting

$$g(z)=\lambda_r+\sum_{n=0}^{\infty}\int dk|(00|nk\rangle|^2\left(\frac{1}{\epsilon_n(k)-z}-\frac{1}{n+\frac{1}{2}}\right). \tag{10}$$

Its spectrum will be determined by the nature of the singularities of R_z as z approaches the real axis. When $\mu=0$, the model reduces to that studied in Ref. 7:

$$g(z)|_{\mu=0}=\lambda_r+\frac{1}{2\pi}\sum_{n=0}^{\infty}\left(\frac{1}{(n+\frac{1}{2})-z}-\frac{1}{n+\frac{1}{2}}\right) \tag{11}$$

has one zero ϵ_j in between each of the Landau levels, $j+1/2<\epsilon_j<j+3/2$, $j=0,1,\ldots$, and a zero $\epsilon_g<1/2$ lying below all Landau levels. These zeros give poles in $R_z|_{\mu=0}$ that correspond to nondegenerate eigenvalues ϵ_j of $H|_{\mu=0}$ with normalized eigenvectors ψ_j (impurity states). As $z\to\epsilon_j$, $g(z)|_{\mu=0}\sim a_j(z-\epsilon_j)$ and $(\epsilon_j-z)\langle\psi_j|R_z|_{\mu=0}|\psi_j\rangle\to 1$. From (6) this determines the coefficient a_j to be

$$a_j=|(00|R^0_{\epsilon_j}|_{\mu=0}|\psi_j\rangle|^2. \tag{12}$$

The poles coming from $R_z^0|_{\mu=0}$ correspond to the Landau levels that remain unaffected by the presence of the impurity (stability of the essential spectrum).

III. WEAK ELECTRIC FIELD ASYMPTOTICS

In presence of the electric field, the boundary values of $g(z)$ as $z=\zeta\pm i\eta$ approaches the real axis are obtained by an application of the Cauchy principal value formula (writing now explicitly the μ dependence in the arguments of the functions)

$$\lim_{\eta\to 0^+} g(\zeta\pm i\eta,\mu)=\alpha(\zeta,\mu)\pm i\beta(\zeta,\mu) \tag{13}$$

with

$$\alpha(\zeta,\mu)=\lambda_r+\sum_{n=0}^{\infty}\mathcal{P}\!\!\int dk|(00|nk\rangle|^2\left(\frac{1}{\epsilon_n(k)-\zeta}-\frac{1}{n+\frac{1}{2}}\right), \tag{14}$$

$$\beta(\zeta,\mu)=\frac{\pi}{\mu}\sum_{n=0}^{\infty}|(00|nk_n(\zeta,\mu)\rangle|^2, \tag{15}$$

where $k_n(\zeta,\mu)$ solves $\epsilon_n(k)=\zeta$, i.e.,

$$k_n(\zeta,\mu)=\frac{1}{\mu}\left(n+\frac{1}{2}-\zeta-\frac{\mu^2}{2}\right). \tag{16}$$

The functions $\alpha(\zeta,\mu)$ and $\beta(\zeta,\mu)$ will determine respectively the spectral shifts and the lifetimes of the impurity states as $\mu\to 0$. From now on we focus attention on the jth impurity state by studying these functions in a neighborhood of the unperturbed energy ϵ_j that does not contain the nearest Landau levels. At $\mu=0$, $\alpha(\zeta,0)$ reduces to the expression (11) and vanishes at ϵ_j with $\partial\alpha(\zeta,0)/\partial\zeta|_{\zeta=\epsilon_j}>0$: the implicit function theorem ensures then that for μ small, $\alpha(\zeta,\mu)$ has a nearby zero at $\zeta=\epsilon_j(\mu)$ and

$$\alpha(\zeta,\mu)=a_j(\mu)(\zeta-\epsilon_j(\mu))+O((\zeta-\epsilon_j(\mu))^2), \quad a_j(\mu)>0 \tag{17}$$

The spectral shift $\epsilon_j(\mu)-\epsilon_j=\epsilon_j^{(1)}\mu+O(\mu^2)$ can itself be expanded in μ and $a_j(\mu)=a_j+O(\mu)$ where a_j has the value (12).(One can verify that the linear correction $\epsilon_j^{(1)}\mu$ is also obtained by formally applying the regular perturbation theory to the eigenvalue ϵ_j when the electric field is switched on).

We study now asympotic behavior of $\beta(\zeta,\mu)$ as $\mu\to 0$ with ζ in a neighborhood of ϵ_j. We denote $\Delta_j=\min(\epsilon_j-(j+1/2),j+3/2-\epsilon_j)<1/2$ the gap between ϵ_j and the nearest Landau level: Δ_j can be equal to either quantity depending on the value of λ_r. For sake of definiteness assume in the sequel that $\Delta_j=\epsilon_j-(j+1/2)$.

Proposition: Set $\zeta-\mu^2/2=j+1/2+\delta$, $0<\delta<1/2$. Then

$$\beta(\zeta,\mu)=\frac{1}{2\sqrt{\pi}}\frac{2^j}{j!}\frac{1}{\mu}\left(\frac{\delta}{\mu}\right)^{2j}e^{-(\delta/\mu)^2}(1+O(\mu^2)). \tag{18}$$

Proof: From (3), (5), (15), and (16) $\beta(\zeta,\mu)$ reads

$$\beta(\zeta,\mu)=\frac{1}{2\sqrt{\pi}}\frac{1}{\mu}\sum_{n=0}^{\infty}\frac{1}{2^n n!}e^{-((n-j-\delta)/\mu)^2}H_n^2\left(\frac{n-j-\delta}{\mu}\right). \tag{19}$$

The control of these series necessitates an estimate of the Hermite polynomials when their argument is of the same magnitude as their order. This is provided by the next lemma (proof in Appendix B).

Lemma: Let x be any real number not equal to a positive integer. Then there exists $\mu_0>0$ such that

$$H_n\left(\frac{n-x}{\mu}\right)=2^n\left(\frac{n-x}{\mu}\right)^n(1-r_n(\mu)), \quad 0<\mu\leqslant\mu_0 \tag{20}$$

where $r_n(\mu)\geqslant 0$ and $r_n(\mu)=O(\mu^2)$ as $\mu\to 0$.

The lemma gives the upper bound

$$\beta(\zeta,\mu)\leqslant\sum_{n=0}^{\infty}b_n(\mu)=b_j(\mu)\left(1+\sum_{n\neq j}^{\infty}\frac{b_n(\mu)}{b_j(\mu)}\right) \tag{21}$$

with

$$b_n(\mu)=\frac{1}{2\sqrt{\pi}}\frac{2^n}{n!}\frac{1}{\mu}\left(\frac{n-j-\delta}{\mu}\right)^{2n}e^{-((n-j-\delta)/\mu)^2}. \tag{22}$$

For $n\neq j$, the largest Gaussian factor in the ratios $b_n(\mu)/b_j(\mu)$ occurs when $n=j+1$. Factorizing it out, we write these ratios in the form

$$\frac{b_n(\mu)}{b_j(\mu)}=\frac{1}{\mu^2}e^{-[(1-\delta)^2-\delta^2]/\mu^2}c_n(\mu). \tag{23}$$

One checks from (22) and the above definition that the $c_n(\mu)$ are bounded as $\mu\to 0$ and $c_n(\mu)\leqslant c_n(1)\sim n^n e^{-(n^2+O(n))}$ for n large. Hence the series $\Sigma_{n\neq j}^{\infty}c_n(\mu)$ converges and is uniformly bounded with respect to μ, implying with (21) and (23) that

$$\beta(\zeta,\mu)\leqslant b_j(\mu)(1+O(e^{-c/\mu^2})) \tag{24}$$

for some $C>0$. On the other hand, one concludes from (19) and the lemma that

$$\beta(\zeta,\mu)\geqslant b_j(\mu)(1-r_j(\mu))^2,\quad r_j(\mu)=O(\mu^2). \tag{25}$$

Combining (24) and (25) gives the result of the proposition.

IV. SHAPE OF RESONANCES AND LIFETIMES

The time-dependent decay amplitude $\langle\psi_j|\exp(-iHt)|\psi_j\rangle$ of the jth impurity state under a weak electric field is given by the Fourier transform of the density of states

$$\rho_j(\zeta,\mu)=\frac{1}{2i\pi}\lim_{\eta\to 0^+}\langle\psi_j|(R_{\zeta+i\eta}-R_{\zeta-i\eta})|\psi_j\rangle \tag{26}$$

as $\mu\to 0$. One finds from (6)

$$\rho_j=\frac{1}{2i\pi}\left(\frac{f_+f_-^*}{\alpha-i\beta}-\frac{f_+^*f_-}{\alpha+i\beta}\right)+\rho_j^0=\frac{\beta}{\pi(\alpha^2+\beta^2)}\mathcal{R}(f_+f_-^*)+\frac{\alpha}{\pi(\alpha^2+\beta^2)}\mathcal{J}(f_+f_-^*)+\rho_j^0, \tag{27}$$

where all the functions depend on the energy ζ; $\rho_j^0(\zeta,\mu)$ is the corresponding density of states of the crossed fields Hamiltonian H^0, $\alpha(\zeta,\mu)$ and $\beta(\zeta,\mu)$ are the functions previously discussed and

$$f_\pm(\zeta)=\lim_{\eta\to 0^+}(00|R^0_{\zeta\pm i\eta}|\psi_j\rangle. \tag{28}$$

In view of (17) and the fact that $\beta(\zeta,\mu)$ tends to zero, the first term in the right-hand side of (27) behaves as a Lorentzian in a neighborhood of ϵ_j for μ small,

$$\frac{\beta_j(\zeta,\mu)}{\pi[(a_j(\mu)(\zeta-\epsilon_j(\mu)))^2+\beta_j(\zeta,\mu)^2]}\mathcal{R}(f_+(\zeta)f_-^*(\zeta))\sim\frac{\frac{1}{2}\Gamma(\mu)}{\pi[(\zeta-\epsilon_j)^2+(\frac{1}{2}\Gamma(\mu))^2]}, \tag{29}$$

with

$$\Gamma_j(\mu)=2\frac{\beta_j(\epsilon_j,\mu)}{a_j}. \tag{30}$$

In (29), we have kept the dominant behavior as $\mu\to 0$ by evaluating $a_j(\mu)$ and $\epsilon_j(\mu)$ at $\mu=0$ and $\beta(\zeta,\ \mu)$, $f_\pm(\zeta)$ at $\zeta=\epsilon_j$. The Lorentzian is properly normalized because $\lim_{\mu\to 0}\mathcal{R}(f_+(\epsilon_j)f_-^*(\epsilon_j))=|(00|R^0_{\epsilon_j}|_{\mu=0}|\psi_j\rangle|^2=a_j$ by (12). The last two terms in (27) remain bounded for ζ in a neighborhood of ϵ_j. Moreover, these two terms vanish as $\mu\to 0$ since $\lim_{\mu\to 0}f_+=\lim_{\mu\to 0}f_-$ is a real quantity and $\lim_{\mu\to 0}\rho_j^0(\zeta,\mu)=0$ when ζ is in between Landau levels.

In (29), $\Gamma(\mu)=((\tau(\mu))^{-1}$ is the inverse lifetime of the jth resonance so that by (30) and (18) $\tau_j(\mu)$ has the form

$$\tau_j(\mu)=C_j\mu\left(\frac{\mu}{\Delta_j}\right)^{2j}\exp\left(\frac{\Delta_j}{\mu}\right)^2, \tag{31}$$

where $\Delta_j=\epsilon_j-(j+1/2)$. The analysis and the results are similar if $\Delta_j=(j+\frac{3}{2})-\epsilon_j$. The lifetime of the resonance corresponding to the lowest energy state ϵ_g is found to be $\tau_g(\mu)=C_g\mu\exp((1/2-\epsilon_g)/\mu)^2$.

V. CONCLUDING REMARKS

We have shown that all impurity states are delocalized under the influence of an electric field, how weak it may be, but with Gaussian long lifetimes. This has to be compared with pure Stark resonances that have exponentially long lifetimes.[8]

Here the calculation has been performed with an attractive δ-potential. (The renormalization procedure given in Section II requires a negative bare coupling constant). The repulsive case is studied in Ref. 3, with a potential with narrow support of diameter d. The infinite n-summations are convergent, and the authors approximate them by finite sums up to a natural cutoff value N determined by the ratio of d to the magnetic length. The result for lifetimes are the same as ours. In addition, they investigate current carrying states as well as the semi-classical regime.

The conclude we remark that our results are sensitive to the potential nature of the impurity interaction. Consider, for instance, the model $H=H^0+\lambda|\phi\rangle\langle\phi|$ obtained by adding a rank-one interaction to the crossed fields Hamiltonian, with $\phi(x,y)$ a square integrable function on $\mathbb{R}^2$, namely the impurity interaction is represented by the nonlocal separable kernel $\lambda\phi(x,y)\phi^*(x',y')$. Energies and lifetimes of resonances are now found from the function

$$g_\phi(z)=\lambda^{-1}+\langle\phi|R_z^0|\phi\rangle \tag{32}$$

in the place of (7) and

$$\beta_\phi(\zeta,\mu)=\frac{\pi}{\mu}\sum_{n=0}^{\infty}|\langle\phi|nk_n(\zeta,\mu)\rangle|^2, \tag{33}$$

$$\langle nk|\phi\rangle=\sum_{n=0}^{\infty}\int dx\, u_n(x)\tilde{\phi}(x,k), \tag{34}$$

where $\tilde{\phi}(x,k)$ is the Fourier transform of $\phi(x,y)$ with respect to the y coordinate. (Here renormalization of the coupling constant is not needed.) Suppose now that

$$\tilde{\phi}(x,k)=0, \quad k\geqslant k_0>0, \tag{35}$$

Then when μ is small enough (say $\mu\ll\Delta_j/k_0$), there exists a neighborhood of ϵ_j such that $k_n(\zeta,\mu)\geqslant k_0$ for all n and thus $\beta_\phi(\zeta,\mu)$ vanishes when ζ is in this neighborhood. This implies that the jth impurity state remains an eigenvector of H with an eigenvalue close to ϵ_j embedded in the continuum. One sees on this example that the interaction has to be sufficiently nonlocal since by the condition (35) the support of $\phi(x,y)$ must extend on the whole y-axis.

APPENDIX A: EXISTENCE OF THE RENORMALIZED MODEL

Introducing (3) in (10) (using also parity of the function u_n), we split the sum in two terms

$$\sum_{n=0}^{\infty}\int dk\,|u_n(-k)|^2\left(\frac{1}{\epsilon_n(k-\mu)-z}-\frac{1}{n+\frac{1}{2}}\right)=I_1+I_2, \tag{A1}$$

$$I_1=\sum_{n=0}^{\infty}\frac{1}{(n+\frac{1}{2})n^{\zeta}}\int_{|k|\leqslant n^{\gamma}}dk\,|u_n(k)|^2 n^{\zeta}\left(\frac{k\mu-(\mu^2/2)+z}{\epsilon_n(k-\mu)-z}\right),$$

$$I_2=\sum_{n=0}^{\infty}\int_{|k|\geqslant n^{\gamma}}dk\,\frac{|u_n(k)|^2}{\epsilon_n(k-\mu)-z}-\sum_{n=0}^{\infty}\int_{|k|\geqslant n^{\gamma}}dk\,\frac{|u_n(k)|^2}{n+\frac{1}{2}},$$

with

$$0<\zeta<\tfrac{1}{4},\quad \tfrac{1}{2}<\gamma=1-2\zeta<1. \tag{A2}$$

One sees in the integrand of I_1 that the fraction

$$\left|n^{\zeta}\left(\frac{z+k\mu-\mu^2/2}{-z+(n+1/2)-\mu k+\mu^2/2}\right)\right|<C<\infty$$

is bounded uniformly with respect to k and n for $|k|\leqslant n^{\gamma}$, $\mathcal{J}z\neq 0$. Hence, since the u_n are normalized, $I_1\leqslant C\sum_{n=0}^{\infty}1/(n+1/2)n^{\zeta}<\infty$. For I_2, since $1/|\epsilon_n(k-\mu)-z|\leqslant 1/|\mathcal{J}z|$, it is sufficient to show that $\sum_{n=0}^{\infty}\int_{|k|\geqslant n^{\gamma}}dk|u_n(k)|^2<\infty$. From the integral representation of the Hermite polynomials[9]

$$H_n(y)=2^n\frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}dt(y+it)^n e^{-t^2}, \tag{A3}$$

one deduces [using $\log(1+y)\leqslant y$]

$$H_n(k)\leqslant\frac{2^n k^n}{\sqrt{\pi}}\int dt\left(1+\left|\frac{t}{k}\right|\right)^n e^{-t^2}\leqslant\frac{2^n k^n}{\sqrt{\pi}}\int dt\,e^{n|t/k|}e^{-t^2}\leqslant 2^{n+1}k^n e^{n^2/4k^2}. \tag{A4}$$

Noting that n sufficiently large ($n\geqslant n_0$) and $|k|>n^{\gamma}$, $\gamma>1/2$ implies $e^{-(k^2/2)(1-n^2/2k^4)}\leqslant e^{-k^2/4}$, one finds from (A4)

$$\begin{aligned}\sum_{n=n_0}^{\infty}\int_{|k|\geqslant n^{\gamma}}dk|u_n(k)|^2&=\frac{1}{2\pi\sqrt{\pi}}\sum_{n=n_0}^{\infty}\frac{1}{2^n n!}\int_{|k|\geqslant n^{\gamma}}dk\,e^{-k^2}H_n^2(k)\\&\leqslant\frac{1}{2\pi\sqrt{\pi}}\sum_{n=n_0}^{\infty}\frac{2^n}{n!}e^{-(1/2)n^{2\gamma}}\int_{|k|\geqslant n^{\gamma}}dk\,e^{-k^2/4}k^{2n}\\&\leqslant\frac{8e^4}{\pi}\sum_{n=n_0}^{\infty}2^n n!\,e^{-(1/2)n^{2\gamma}}<\infty,\end{aligned} \tag{A5}$$

where the last inequality follows from $|k|^n\leqslant n!\,e^{|k|}$. The series (A5) converges for $\gamma>1/2$.

APPENDIX B: PROOF OF THE LEMMA

The lemma is true by inspection for the cases $n=1,2,3$. If $n\geqslant 4$, one uses the formula (A3),

$$H_n\left(\frac{n-x}{\mu}\right)=2^n\left(\frac{n-x}{\mu}\right)^n f_n(\mu),\quad f_n(\mu)=\frac{1}{\sqrt{\pi}}\int dt\left(1+\frac{i\mu t}{n-x}\right)^n e^{-t^2}. \tag{B1}$$

The limited Taylor expansion of $f_n(\mu)$ around $\mu=0$ gives

$$f_n(\mu)=1+\frac{1}{2!}\mu^2 f_n''(0)+\frac{1}{4!}\mu^4 f_n''''(\bar{\mu}),\quad 0\leqslant\bar{\mu}\leqslant\mu, \tag{B2}$$

with

$$f_n''(0)=-\frac{n(n-1)}{(n-x)^2},$$

$$f_n''''(\bar{\mu})=\frac{n!}{(n-4)!(n-x)^4}\frac{1}{\sqrt{\pi}}\int dt\, t^4\left(1+i\frac{\bar{\mu}t}{n-x}\right)^{n-4}e^{-t^2}.$$

If x is not a positive integer, there exists C_1 independent of n, $0<C_1<\infty$, such that $f_n''(0)\geqslant -C_1$. Moreover, one has for $n\geqslant 4$ and using $\log(1+y)\leqslant y$

$$|f_n''''(\bar{\mu})|\leqslant\frac{n!}{(n-4)!(n-x)^4}\frac{1}{\sqrt{\pi}}\int dt\, t^4\left(1+\bar{\mu}\left|\frac{t}{n-x}\right|\right)^{n-4}e^{-t^2}$$

$$\leqslant\frac{n!}{(n-4)(n-x)^4}\frac{1}{\sqrt{\pi}}\int dt\, t^4 e^{(n-4)\bar{\mu}|t|/|n-x|}e^{-t^2}$$

$$\leqslant C_2<\infty,$$

with C_2 independent of n and of $\bar{\mu}$ in compact sets. This leads to the conclusion of the lemma.

[1] N. Berglund, A. Hansen, E. H. Hauge, and J. Piasecki, Phys. Rev. Lett. **77**, 2149 (1996).
[2] R. E. Prange, Phys. Rev. B **23**, 4802 (1981).
[3] E. H. Hauge and J. M. J. van Leeuwen, Bound states and metastability near a scatterer in crossed electromagnetic fields, to be published in Physica A.
[4] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, IV. Analysis of operators*, (Academic, London, 1978) Theorem XIII.86.
[5] F. Gesztesy, H. Holden, and P. Seba, in *Schrödinger operators, standard and non-standard*, edited by P. Exner and P. Seba (World Scientific, Singapore, 1989), p. 147.
[6] F. Perez and F. A. B. Coutinho, Am. J. Phys. **59**, 52 (1991).
[7] R. M. Cavalcanti and C. A. A. de Carvalho, J. Phys. A **31**, 2391 (1998).
[8] L. Landau and E. Lifchitz, *Quantum Mechanics* (Pergamon, New York, 1965), Chap. X.
[9] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953), Chap. VI.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Optimal cloning of pure states, testing single clones

M. Keyl[a)] and R. F. Werner[b)]
Institut für Mathematische Physik, TU Braunschweig, Mendelssohnstraße 3, 38106 Braunschweig, Germany

(Received 4 August 1998; accepted for publication 7 April 1999)

We consider quantum devices for turning a finite number N of d-level quantum systems in the same unknown pure state σ into $M>N$ systems of the same kind, in an approximation of the M-fold tensor product of the state σ. In a previous paper it was shown that this problem has a unique optimal solution, when the quality of the output is judged by arbitrary measurements, involving also the correlations between the clones. We show in this paper, that if the quality judgment is based solely on measurements of single output clones, there is again a unique optimal cloning device, which coincides with the one found previously. © *1999 American Institute of Physics.* [S0022-2488(99)03707-X]

I. INTRODUCTION

According to the well-known ''no-cloning theorem''[1] perfect copying of quantum information is impossible, i.e., there is no machine which takes a quantum system as input and produces two systems of the same kind, both of them indistinguishable from the input. However, from the point of view of practical applications in Quantum Information Theory this Theorem by itself is not very useful, because it only asserts that the cloning task cannot be performed *exactly*—but then no task can be performed exactly by real devices. The fundamental importance of the No-Cloning Theorem is expressed much better in stronger versions of the Theorem, which also give explicit lower bounds on the error made in any attempt to build a cloning device. Some such bounds have been established, as well.[2,3] Even more insight into the cloning problem is given by results showing how to minimize the error, i.e., how to construct *optimal* cloning devices.[4–8] Other recent related work can be found in Refs. 9–15.

In this paper we consider cloning devices, which take as input a certain number N of identically prepared systems, and produce a larger number M of systems as output. Again, the cloning task is to make the output state resemble as much as possible a state of M systems all prepared in the same state as the inputs. This variant of the problem is of interest as a ''quantum amplifier.'' It also has a better chance of reasonable success than a cloning device operating on single input systems: In the limit of many input systems the device can make a good statistical estimate of the input density matrix and hence produce arbitrarily good clones.

Different variants of this problem arise by different choices of the type of systems and the set of states which should be copied, e.g., pure versus mixed states, or a finite number of states arising in a cryptographic protocol. In the present paper we are exclusively concerned with the cloning of arbitrary unknown pure states.

A second choice to be made is the precise notion of approximation between the output states of the cloning device and the (inattainable) target state. Apart from technicalities the basic issue here is whether the full states are compared, or only the one-clone marginals. Approximation in the first sense means that the expectations of all observables, including those testing correlations and entanglement between different clones, are close in the two states being compared. On the other hand, approximation in the second sense means closeness of expectations of single clone observables only. Perhaps this second condition has more of the flavor of the No-Cloning Theo-

[a)]Electronic mail: M.Keyl@tu-bs.de
[b)]Electronic mail: R.Werner@tu-bs.de

rem, since in that theorem, too, the requirement is that each single (!) clone be indistinguishable from the input.

In Ref. 16 we showed that the pure state cloning problem with all-particle test criterion has a unique optimal solution. In this paper we show the same for the single-particle test criterion, and that the two optimal cloning devices are actually the same. The difference between the two results may not seem great. However, the result in the present paper required much heavier mathematical machinery, and we believe it to be considerably deeper. The reason is that one-particle tests by far do not exhaust the linear space of M-particle observables. In particular, all correlations of the cloner's output are ignored by the test, which would make a *unigue* optimal solution appear rather unlikely. Nevertheless, this is what we prove.

II. STATEMENT OF THE PROBLEM AND MAIN RESULT

Let us start with a precise formulation of the question we are going to consider. First of all, we will study throughout this paper only d-level systems with arbitrary but finite d. Hence the one-particle Hilbert space $\mathcal{H}$ we are using is $\mathcal{H}=\mathbb{C}^d$. The Hilbert space for the input to the cloning device is therefore the N-fold tensor product $\mathcal{H}^{\otimes N}$ of $\mathcal{H}$ with itself. In fact, because we only consider tensor powers of pure states as inputs, it suffices to take the subspace of $\mathcal{H}^{\otimes N}$ spanned by vectors of the form $\varphi^{\otimes N}$ with $\varphi\in\mathcal{H}$. This is precisely the ''Bose'' subspace $\mathcal{H}_+^{\otimes N}\subset\mathcal{H}^{\otimes N}$, i.e., the space of vectors invariant under all permutations. The output Hilbert space will be $\mathcal{H}^{\otimes M}$ with $M>N$. On this space we cannot impose an *a priori* symmetry restriction, although such a restriction will come out *a posteriori*, as a special property of optimal cloning devices.

A *cloning map* is a completely positive, unital map $T:\mathcal{B}(\mathcal{H}^{\otimes M})\rightarrow\mathcal{B}(\mathcal{H}_+^{\otimes N})$. This describes the action of the device on observables. Its (pre-)dual, describing the same operation in terms of states, will be denoted[17] by $T_*:\mathcal{B}_*(\mathcal{H}_+^{\otimes N})\rightarrow\mathcal{B}_*(\mathcal{H}^{\otimes M})$. If we identify states with density operators, this means that $\mathrm{tr}(\rho T(A))=\mathrm{tr}(T_*(\rho)A)$ for arbitrary density operators ρ and observables A. The input of the cloning device are N systems, prepared independently according to the same state σ. Thus the overall input state is $\sigma^{\otimes N}$. We will assume σ to be pure, i.e., the density matrix of σ is a one-dimensional projection onto a wave vector $\psi\in\mathcal{H}$, say. Then $\sigma^{\otimes N}$ is the projection onto the vector $\psi^{\otimes N}$. The output of the cloning device is the state $T_*(\sigma^{\otimes N})$, which is a (generally entangled) state of $M>N$ systems. Our aim is to design T so that the output states $T_*(\sigma^{\otimes N})$ approximate the product states $\sigma^{\otimes M}$.

The one particle observables, on which the comparison will be based, will be written as $a_{(k)}=\mathbb{1}^{\otimes(k-1)}\otimes a\otimes\mathbb{1}^{\otimes(M-k)}\in\mathcal{B}(\mathcal{H}^{\otimes M})$, for all $a\in\mathcal{B}(\mathcal{H})$. Thus the optimal cloning problem for arbitrary pure input states T is to make the expectations

$$\mathrm{tr}(a_{(k)}T_*(\sigma^{\otimes N}))=\mathrm{tr}(T(a_{(k)})\sigma^{\otimes N})=\langle\psi^{\otimes N},T(a_{(k)})\psi^{\otimes N}\rangle,$$

$$\mathrm{tr}(a_{(k)}\sigma^{\otimes M})=\mathrm{tr}(a\sigma)=\langle\psi,a\psi\rangle$$

as similar as possible for arbitrary one-particle observables a and one-particle vectors ψ. Of course, when taking a supremum over such differences, the size of a has to be constrained somehow. We will choose the constraint $0\leq a\leq\mathbb{1}$, which is to say that the above two expressions have an immediate interpretation as probabilities. The largest difference of such probabilities is now the error functional for cloning maps, which we will seek to minimize:

$$\Delta_{\mathrm{one}}(T)=\sup_{a,\psi,k}|\langle\psi^{\otimes N},T(a_{(k)})\psi^{\otimes N}\rangle-\langle\psi,a\psi\rangle|, \tag{1}$$

where the supremum is taken over all $\psi\in\mathcal{H}$ with $\|\psi\|=1$, all operators $a\in\mathcal{B}(\mathcal{H})$ with $0\leq a\leq\mathbb{1}$, and all integers $1\leq k\leq M$.

The corresponding quantity based on tests of the full output state (including correlations) is

$$\Delta_{\mathrm{all}}(T)=\sup_A\sup_{\sigma,\mathrm{pure}}|\mathrm{tr}(T(A)\sigma^{\otimes N})-\mathrm{tr}(A\sigma^{\otimes M})|,$$

where the supremum is taken over all $A \in \mathcal{B}(\mathcal{H}^{\otimes M})$ with $0 \leq A \leq \mathbb{1}$ and over all pure states $\sigma \in \mathcal{B}_*(\mathcal{H})$. Due to the properties of the trace norm $\|\cdot\|_1$ this functional can be expressed by

$$\Delta_{\mathrm{all}}(T) = \sup_{\sigma,\mathrm{pure}} \|T_*(\sigma^{\otimes N}) - \sigma^{\otimes M}\|_1 . \tag{2}$$

It turns out that there is exactly one cloning map $\hat{T}$ which minimizes this error functional. This can be proven with a minor adaptation of the arguments in Ref. 16, which start from a slightly different criterion, namely the maximization of the ''fidelity'' $\mathcal{F}(T_*) = \sup_{\sigma,\mathrm{pure}}(\sigma^{\otimes M} T_*(\sigma^{\otimes N}))$. The unique solution $T = \hat{T}$ minimizing (2), or maximizing $\mathcal{F}(T_*)$, is best expressed in terms of its action on states, i.e.,

$$\hat{T}_*(\rho) = \frac{d[N]}{d[M]} S_M(\rho \otimes \mathbb{1}^{M-N}) S_M . \tag{3}$$

Here $d[N] = \binom{d+N-1}{N}$ denotes the dimension of the symmetric subspace $\mathcal{H}_+^{\otimes N}$, S_M is the projection from $\mathcal{H}^{\otimes M}$ to $\mathcal{H}_+^{\otimes M}$, and ρ is an arbitrary density operator on $\mathcal{H}_+^{\otimes N}$. In Ref. 16 we also computed the one-site restriction of the output states of this cloner:

$$\mathrm{tr}(\hat{T}(a_k)\sigma^{\otimes N}) = \gamma(\hat{T})\sigma(a) + (1 - \gamma(\hat{T}))\mathrm{tr}(a)/d,$$

where

$$\gamma(\hat{T}) = \frac{N}{N+d} \frac{M+d}{M}$$

is the so-called Black Cow factor of $\hat{T}$, interpreted as a ''shrinking factor of the Poincaré sphere'' in the discussions of the qubit $(d=2)$ case. This makes it easy to verify the case of equality in the following Theorem, which is our main result.

Theorem 1: *For any cloning map* $T: \mathcal{B}(\mathcal{H}^{\otimes M}) \to \mathcal{B}(\mathcal{H}_+^{\otimes N})$ *we have*

$$\Delta_{\mathrm{one}}(T) \geq \frac{d-1}{d} \left| 1 - \frac{N}{N+d} \frac{M+d}{M} \right|$$

with equality iff $T = \hat{T}$ *with* $\hat{T}$ *from Eq. (3).*

III. FINDING THE OPTIMAL CLONING MAP

A. Reduction to the covariant case

In this section we will give the proof of our main theorem, apart from some group theoretical Lemmas, which will be proved in Appendix A. Throughout, the symmetry of sitewise unitary rotation of clones and input states will play a crucial role. The necessary background information on unitary representations of SU(d) will also be supplied in Appendix A.

We establish some notation first. By U(d) we will denote the group of unitary $d \times d$-matrices, i.e., the unitary group on our underlying one-particle space $\mathcal{H} \equiv \mathbb{C}^d$. Unitary representations of this group will be denoted by the letter π with suitable indices. $\pi_\square$ is the defining representation on $\mathbb{C}^d$, and its nth tensor power, acting on $\mathcal{H}^{\otimes N}$ by the operators $\pi_\square^{\otimes N}(u) = u^{\otimes N}$ is $\pi_\square^{\otimes N}$. The restriction of this representation to the symmetric subspace $\mathcal{H}_+^{\otimes N}$ will be denoted by π_N^+. Thus a cloning map $T: \mathcal{B}(\mathcal{H}^{\otimes M}) \to \mathcal{B}(\mathcal{H}_+^{\otimes N})$ is called U(d)-*covariant*, if

$$T(\pi_\square^{\otimes M}(u) A \pi_\square^{\otimes M}(u)^*) = \pi_N^+(u) T(A) \pi_N^+(u)^* . \tag{4}$$

This equation merely expresses that T does not prefer any direction in $\mathcal{H}$. It would be a natural initial assumption for good cloning devices but, of course, in our case it will come out as a result of the minimization: $\hat{T}$ from Eq. (3) is obviously covariant, because S_M commutes with all $\pi_\square^{\otimes M}(u)$. It is convenient to state the covariance condition as a fixed point property: We define the action τ of unitary rotations on cloning maps by

$$(\tau_u T)(A) = \pi_N^+(u)^* T(\pi_\square^{\otimes M}(u) A \pi_\square^{\otimes M}(u)^*) \pi_N^+(u), \tag{5}$$

so that T is covariant iff $\tau_u(T) = T$ for all $u \in U(d)$. We denote by $\overline{T}$ the average of $\tau_u T$ with respect to u, i.e.,

$$\overline{T} = \int du\ \tau_u(T), \tag{6}$$

where ''du'' denotes the normalized Haar measure on $\mathrm{U}(d)$.

The fact that the cloning error Δ_{one} does not single out a direction on $\mathcal{H}$ either is expressed by the—easily verified—equation

$$\Delta_{\mathrm{one}}(\tau_u T) = \Delta_{\mathrm{one}}(T). \tag{7}$$

Similarly, we can get an estimate of $\Delta_{\mathrm{one}}(\overline{T})$: The functional Δ_{one} is defined as the supremum of a set of convex expressions in T. Therefore, it is convex, and $\Delta_{\mathrm{one}}(\overline{T}) \leqslant \Delta_{\mathrm{one}}(T)$. So as long as we are only interested in finding *some* cloning map with minimal Δ_{one}, we may restrict attention to covariant ones.

There is a similar simplification, which we can make ''without loss of cloning quality'': Δ_{one} is invariant under a change of the ordering of the clones. That is to say, if $V:\mathcal{H}^{\otimes M} \to \mathcal{H}^{\otimes M}$ is a permutation operator, and if we define $\tau_V T$ by $(\tau_V T)(A) = T(VAV^*)$, we may replace T by its average over permutations without loss of cloning quality. That is, we may assume that $\tau_V T = T$ for all permutations V. We will refer to this property as *permutation invariance*.

Our strategy is now to assume $\mathrm{U}(d)$-covariance and permutation invariance of T, and to show that there is a unique solution to the variational problem with these additional properties. The above convexity argument then implies that no other cloning map can do better. But since the functional Δ_{one} is not strictly convex, we will need an extra step to establish uniqueness. This we will do in Sec. III F by showing that any cloning map whose mean is the optimal covariant cloner has to be covariant itself.

B. Reduction to the extremal covariant case

The functional Δ_{one} involves only operators $T(A)$ with A of the special form $A = a_{(k)} = \mathbb{1}^{\otimes(k-1)} \otimes a \otimes \mathbb{1}^{\otimes(M-k)} \in \mathcal{B}(\mathcal{H}^{\otimes M})$. Now due to permutation invariance $T(a_{(k)})$ does not depend on k, and we have

$$T(a_{(k)}) = \frac{1}{M} T\left(\sum_k a_{(k)} \right). \tag{8}$$

What makes Eq. (8) useful is that on the right-hand side T is now applied to one of the generators of the representation $\pi_\square^{\otimes N}$: we have $\exp(i\Sigma_{k=1}^M a_{(k)}) = (\exp(ia))^{\otimes M}$. Because T is covariant, we can determine how the operators in Eq. (8) transform under $\mathrm{U}(d)$-rotations:

$$\pi_N^+(u) T(a_{(k)}) \pi_N^+(u)^* = T((uau^*)_{(k)}), \tag{9}$$

where the multiplication of a and u on the right-hand side is in the $d \times d$-matrices. This property fixes the ''transformation behavior'' of the operators $T(a_{(k)})$, and as we will see, this essentially fixes the tuple of operators $T(a_{(k)})$. Of course, $a = \mathbb{1}$ in (9) simply leads to $T(\mathbb{1}_{(k)}) = T(\mathbb{1}) = \mathbb{1}$. The operator $i\mathbb{1}$ is the (anti-Hermitian) generator of the subgroup of unitaries multiplying each vector with the same phase. More interesting are the generators of $\mathrm{SU}(d)$, in which such trivial phases have been eliminated. These generators, in other words the Lie algebra $\mathfrak{su}(d)$, are exactly the traceless anti-Hermitian $d\times d$-matrices. In the qubit case ($d=2$) $\mathfrak{su}(2)$ is spanned by the Pauli matrices (multiplied by i), and 3-tuples of operators transforming like the generators are known in physics literature as ''vector operators.'' It is well-known that, due to the simple reducibility of SU(2), each irreducible representation of SU(2) contains exactly one vector operator (up to a factor), namely the generators (angular momentum operators) of the representation themselves. So all operators $T(a_{(k)})$ are determined by the single numerical factor relating the operators $T(a_{(k)})$ to the generators of the irreducible representation π_N^+.

It turns out that the same idea works in the SU(d)-case for arbitrary d. In order to state it precisely, we need a notation for the Lie algebra representation associated with a unitary representation of a Lie group. We define $\partial\pi(X)$ to be the anti-Hermitian generator of the one-parameter subgroup generated by X, i.e.,

$$\partial\pi(X)=\frac{d}{dt}\,\pi(e^{tX})\bigg|_{t=0}. \tag{10}$$

Then the desired property of a representation is stated in the following definition:

Definition 2: Let $\pi:G\to\mathcal{B}(\mathcal{H}_\pi)$ *be a finite dimensional unitary representation of a Lie group* G *with Lie algebra* $\mathfrak{g}$. *Then* $\mathfrak{g}$ *is said to be nondegenerate in* $\mathcal{B}(\mathcal{H}_\pi)$ *with respect to* π, *if any linear operator* $L:\mathfrak{g}\to\mathcal{B}(\mathcal{H}_\pi)$ *with the covariance property* $\pi(g)L(X)\pi(g)^*=L(gXg^{-1})$ *is of the form* $L(X)=\lambda\,\partial\pi(X)$, *for some factor* $\lambda\in\mathbb{C}$.

As we argued above, $\mathfrak{su}(2)$ is nondegenerate in *every* irreducible representation of SU(2). However, for $d\geq 3$ we can find representations containing degenerate copies of the generators, and we have to make sure that the special representations occurring in the present problem are of the ''good'' kind. This is the content of the following Lemma, proved in the Appendix.

Lemma 3: $\mathfrak{su}(d)$ *is nondegenerate in* $\mathcal{B}(\mathcal{H}_+^{\otimes N})$ *with respect to* π_N^+.

Corollary 4: Let $\pi:U(d)\to\mathcal{B}(\mathcal{H}_\pi)$ *be a unitary representation, and let* $T:\mathcal{B}(\mathcal{H}_\pi)\to\mathcal{B}(\mathcal{H}_+^{\otimes N})$ *be a completely positive normalized and* U(d)*-covariant map*, i.e., $T(\pi(u)A\pi(u)^*)=\pi_N^+(u)T(A)\pi_N^+(u)^*$. *Then there is a number ω(T) such that*

$$T(\partial\pi(a))=\omega(T)\sum_{k=1}^{N}a_{(k)},$$

for every $a\in\mathcal{B}(\mathcal{H})$ *with* $\mathrm{tr}(a)=0$.

Given $\omega(T)$ for $\pi=\pi_\square^{\otimes M}$, we can compute the cloning error $\Delta_{\mathrm{one}}(T)$ as follows: Given $a\in\mathcal{B}(\mathcal{H})$ with $0\leq a\leq\mathbb{1}$, we can write $a=\alpha\mathbb{1}+a'$ with $\mathrm{tr}\,a'=0$. Then

$$T(a_{(k)})=\alpha\mathbb{1}+\frac{1}{M}T\left(\sum_{l=1}^{M}a'_{(l)}\right)=\alpha\mathbb{1}+\frac{\omega(T)}{M}\left(\sum_{l=1}^{N}a'_{(l)}\right)$$

and with $a'=a-\alpha\mathbb{1}$ and $\alpha=(\mathrm{tr}\,a)/d$

$$T(a_{(k)})=\frac{\mathrm{tr}\,a}{d}\left(1-\frac{N\omega(T)}{M}\right)\mathbb{1}+\frac{\omega(T)}{M}\left(\sum_{l=1}^{N}a_{(l)}\right).$$

In any state $\psi\in\mathcal{H}$ we get

$$\langle\psi,a\psi\rangle-\langle\psi^{\otimes N},T(a_{(k)})\psi^{\otimes N}\rangle=(1-\gamma(T))\left(\langle\psi,a\psi\rangle-\frac{\mathrm{tr}\,a}{d}\right),$$

where $\gamma(T)=(N/M)\omega(T)$ is the Black-Cow factor already mentioned in Sec. II. With

$$\sup_{\psi,a}\left(\langle\psi,a\psi\rangle-\frac{\operatorname{tr}a}{d}\right)=\frac{d-1}{d}$$

we get

$$\Delta_{\rm one}(T)=\frac{d-1}{d}\left|1-\frac{N}{M}\,\omega(T)\right|=\frac{d-1}{d}\left|1-\gamma(T)\right|. \tag{11}$$

We remark that the largest possible $\omega(T)$, to be determined below, still makes the second term in the absolute value less than 1, so we could omit the absolute value signs. In any case, we will only seek to maximize $\omega(T)$ from now on, ignoring the possibility of $\omega(T)>M/N$, anticipating that it will be ruled out by the result of the maximization anyway.

An important observation about the Corollary and formula (11) is that ω is clearly an *affine* functional on the convex set of covariant cloning maps (i.e., ω respects convex combinations). Whereas we previously used the convexity of $\Delta_{\rm one}$ to conclude that averaging over rotations and permutations (and hence a move toward the interior of the convex set of cloning maps) generally improves the cloning quality, we now see that the optimum can be sought, as for any affine functional, on the extreme boundary of the subset of covariant cloning maps. Therefore our next steps will be aimed at the determination of the extremal $\mathrm{U}(d)$-covariant and permutation invariant cloning maps, and, subsequently the solution of the variational problem for these extremal cases.

C. Convex decomposition of covariant cloning maps

For the first reduction step we use the close connection between the permutation operators on $\mathcal{H}^{\otimes M}$ and the representation $\pi_{\square}^{\otimes M}$. Let $(\pi_{\square}^{\otimes M})'$ denote the algebra of all operators on $\mathcal{H}^{\otimes M}$ commuting with all $\pi_{\square}^{\otimes M}(u)\equiv u^{\otimes M}$. This algebra consists precisely of the linear combinations of permutation unitaries (see Theorem IX.11.5 in Ref. 18). So consider a reduction of $\pi_{\square}^{\otimes M}$ into irreducibles, i.e., an orthogonal decomposition of the identity into minimal projections $E_\alpha \in(\pi_{\square}^{\otimes M})'$. Then due to covariance the operators $T(E_\alpha)$ commute with all $\pi_N^+(u)$, and because the latter representation is irreducible, they must be multiples of the identity, $T(E_\alpha)=r_\alpha\mathbb{1}$, say. Because $T(VA)=T(AV)$ for permutation operators V, we also have $T(AE_\alpha)=T(E_\alpha AE_\alpha)$. Hence

$$T_\alpha(A)=r_\alpha^{-1}T(E_\alpha AE_\alpha) \tag{12}$$

is a legitimate cloning map in its own right (provided $r_\alpha\neq 0$). Moreover,

$$T(A)=\sum_\alpha T(AE_\alpha)=\sum_\alpha T(E_\alpha AE_\alpha)=\sum_\alpha r_\alpha T_\alpha(A) \tag{13}$$

is a convex decomposition of the given T into such summands. Maximizing $\omega(T) =\Sigma_\alpha r_\alpha\omega(T_\alpha)$ thus means concentrating the coefficients r_α on those α, for which $\omega(T_\alpha)$ is maximal. At this stage it is perhaps already plausible that only the summand T_α, for which $E_\alpha =S_M$ is the projection onto the symmetric subspace, will give the best $\omega(T_\alpha)$, because this is the space supporting the pure states $\sigma^{\otimes M}$ the cloner is supposed to approximate. In fact, for the optimization of $\Delta_{\rm all}$ in Ref. 16 this idea leads directly to a simple solution. In the present case we found no direct proof of this plausible statement.

We therefore have to enter into the further convex decomposition of each T_α. The output states of this cloning map are supported by $\mathcal{H}_\alpha\equiv E_\alpha\mathcal{H}^{\otimes M}$, and we will restrict T_α accordingly, i.e., we consider it as a covariant map $T_\alpha:\mathcal{B}(\mathcal{H}_\alpha)\rightarrow\mathcal{B}(\mathcal{H}_+^{\otimes N})$, which is covariant with respect to the restricted representation $\pi_\alpha=\pi^{\otimes M}\upharpoonright\mathcal{H}_\alpha$ and π_N^+.

As for any completely positive map, the convex decompositions of T_α are governed by the Stinespring dilation.[19] Since we are looking, more specifically, for decompositions into covariant completely positive maps, we have to invoke a ''covariant'' version of the Stinespring dilation Theorem,[20] which is stated in Appendix B for the convenience of the reader. According to this

Theorem we can write a covariant completely positive $T_\alpha:\mathcal{B}(\mathcal{H}_\alpha)\to\mathcal{B}(\mathcal{H}_+^{\otimes N})$ as $T_\alpha(A)=V^*(A\otimes\mathbb{1}_\mathcal{K})V$, where $\mathcal{K}$ is some auxiliary Hilbert space carrying a unitary representation $\tilde{\pi}:\mathrm{U}(d)\to\mathcal{B}(\mathcal{K})$, and $V:\mathcal{H}_+^{\otimes N}\to\mathcal{H}_\alpha\otimes\mathcal{K}$ is an isometry intertwining the respective representations, i.e.,

$$V\pi_N^+(u)=(\pi_\alpha(u)\otimes\tilde{\pi}(u))V. \tag{14}$$

The convex reduction theory of T_α is now the same as the reduction theory of $\tilde{\pi}$ into irreducibles: if F_β is a minimal projection in the algebra $\tilde{\pi}'$, and hence $\tilde{\pi}\upharpoonright F_\beta\mathcal{K}$ is irreducible, then $A\mapsto V^*(A\otimes F_\beta)V$ is a covariant map, which cannot be further decomposed into a sum of covariant completely positive maps (see Appendix B). Note that $V^*(\mathbb{1}\otimes F_\beta)V$ commutes with the irreducible representation π_N^+, so that once again this summand is normalized up to a factor: $V^*(\mathbb{1}\otimes F_\beta)V=r_\beta\mathbb{1}$. Therefore $T_\alpha=\Sigma_\beta r_\beta T_{\alpha\beta}$, where each $T_{\alpha\beta}(A)=r_\beta^{-1}V^*(A\otimes F_\beta)V$ is again an admissible cloning map. The following statement summarizes the result of the decomposition theory of T.

Proposition 5: Let $T:\mathcal{B}(\mathcal{H}^{\otimes M})\to\mathcal{B}(\mathcal{H}_+^{\otimes N})$ *be a* $\mathrm{U}(d)$*-covariant and permutation invariant cloning map. Then* T *is a convex combination* $T=\Sigma_{\alpha\beta}r_{\alpha\beta}T_{\alpha\beta}$ *such that each* $T_{\alpha\beta}$ *is of the following special form:* $T_{\alpha\beta}(A)=V^*(A\otimes\mathbb{1}_\beta)V$*, where* V *is an intertwining isometry between* π_N^+ *and* $\pi_\alpha\otimes\pi_\beta$*, such that* $\pi_\alpha:\mathrm{U}(d)\to\mathcal{B}(\mathcal{H}_\alpha)$ *is an irreducible subrepresentation of* $\pi_\square^{\otimes M}$*, and* $\pi_\beta:\mathrm{U}(d)\to\mathcal{B}(\mathcal{H}_\beta)$ *is also an irreducible unitary representation.*

This Proposition summarizes all that is needed for the further treatment of the variational problem. However, we could have made a slightly stronger statement by eliminating the nonuniqueness introduced by the choice of the minimal projections E_α. If the subrepresentations π_α and $\pi_{\alpha'}$ are unitarily equivalent, then they can be connected by a unitary, which is again a linear combination of permutations. Hence the contribution of the term $r_\alpha T_\alpha=\Sigma_\beta r_{\alpha\beta}T_{\alpha\beta}$ to $\omega(T)$ depends only on the isomorphism type of π_α.

What we cannot assert in general, however, is that V is determined by the isomorphism types of π_α and π_β: Among the groups $\mathrm{SU}(d)$ only $d=2$ is ''simply reducible,'' which means that the space of intertwiners between π_γ and $\pi_\alpha\otimes\pi_\beta$ is at most one dimensional for arbitrary irreducible representations π_α, π_β, π_γ. In Sec. III D we will therefore focus on the qubit case, and show how to determine $\omega(T_{\alpha\beta})$ from the representations involved. This procedure will then be generalized to arbitrary d, and it will turn out that, perhaps surprisingly, in the general case $\omega(T_{\alpha\beta})$ also depends on π_α, π_β only up to unitary equivalence.

D. Maximizing ω in the case $d=2$

For $d=2$ the representations of SU(2) are conventionally labeled by their ''total angular momentum'' $j=0,1/2,1,\ldots$. The irreducible representation π_j has dimension $2j+1$, and is isomorphic to π_N^+ with $N=2j$ in the notation used above. For $j=1$ we get the three-dimensional representation isomorphic to the rotation group, which is responsible for the importance of this group in physics. In a suitable basis X_1, X_2, X_3 of the Lie algebra $\mathfrak{su}(2)$ we get the commutation relations $[X_1,X_2]=X_3$, and cyclic permutations of the indices thereof. In the $j=1$ representation $\partial\pi_1(X_k)$ generates the rotations around the k axis in 3-space. The Casimir operator of SU(2) is the square of this vector operator, i.e., $\tilde{\mathbf{C}}_2=\Sigma_{k=1}^3X_k^2$. In the representation π_j it is the scalar $j(j+1)$, i.e., if we extend the representation $\partial\pi$ of the Lie algebra to the universal enveloping algebra (which also contains polynomials in the generators), we get $\partial\pi_j(\tilde{\mathbf{C}}_2)=j(j+1)\mathbb{1}$. We can use this to determine $\omega(T_{\alpha\beta})$ for arbitrary irreducible representations. This computation can be seen as an elementary computation of a so-called $6j$-symbol (see also Ref. 21 for a context in which the same computation arises), but we will not need to invoke any of the $6j$-machinery.

So let V be an intertwining isometry between π_γ and $\pi_\alpha\otimes\pi_\beta$, where $\alpha,\beta,\gamma\in\{0,1/2,\ldots\}$ label irreducible representations. Then ω is defined by

$$\omega\cdot\partial\pi_\gamma(X_k)=V^*(\partial\pi_\alpha(X_k)\otimes\mathbb{1}_\beta)V. \tag{15}$$

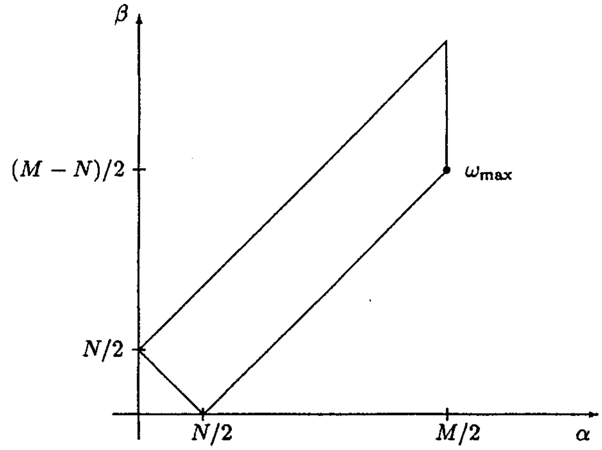

FIG. 1. Area of admissible pairs (α,β).

We multiply this equation by $\partial\pi_\gamma(X_k)$, use the intertwining property of V in the form $V\partial\pi_\gamma(X)=(\partial\pi_\alpha(X)\otimes\mathbb{1}_\beta+\mathbb{1}_\alpha\otimes\partial\pi_\beta(X))V$, and sum over k to get

$$\omega\cdot\partial\pi_\gamma(\tilde{\mathbf{C}}_2)=V^*(\partial\pi_\alpha(\tilde{\mathbf{C}}_2)\otimes\mathbb{1}_\beta)V+\sum_k V^*(\partial\pi_\alpha(X_k)\otimes\partial\pi_\beta(X_k))V.$$

The tensor product in the second summand can be re-expressed in terms of Casimir operators as

$$\sum_k (\partial\pi_\alpha(X_k)\otimes\partial\pi_\beta(X_k))=\frac{1}{2}\sum_k (\partial\pi_\alpha(X_k)\otimes\mathbb{1}_\beta+\mathbb{1}_\alpha\otimes\partial\pi_\beta(X_k))^2-\tfrac{1}{2}\partial\pi_\alpha(\tilde{\mathbf{C}}_2)\otimes\mathbb{1}_\beta-\tfrac{1}{2}\mathbb{1}_\alpha\otimes\partial\pi_\alpha(\tilde{\mathbf{C}}_2).$$

Inserting this into the previous equation, using the intertwining property once again, and inserting the appropriate scalars for $\partial\pi(\tilde{\mathbf{C}}_2)\equiv\tilde{C}_2(\pi)\mathbb{1}$, we find that $\omega\cdot\tilde{C}_2(\pi_\gamma)=\tilde{C}_2(\pi_\alpha)+\frac{1}{2}(\tilde{C}_2(\pi_\gamma)-\tilde{C}_2(\pi_\alpha)-\tilde{C}_2(\pi_\beta))$, and hence

$$\omega=\frac{1}{2}+\frac{\tilde{C}_2(\pi_\alpha)-\tilde{C}_2(\pi_\beta)}{2\tilde{C}_2(\pi_\gamma)}. \tag{16}$$

Note that we have only used the fact that the Casimir operator $\tilde{\mathbf{C}}_2$ is some fixed quadratic expression in the generators. This is also true for SU(d). Hence Eq. (16) also holds in the general case. In particular, we have shown that for the purpose of optimizing $\omega(T_{\alpha\beta})$ only the isomorphism types of π_α and π_β are relevant, but not the particular intertwiner V.

Specializing again to the case $d=2$, we find

$$\omega=\frac{1}{2}+\frac{\alpha(\alpha+1)-\beta(\beta+1)}{2\gamma(\gamma+1)}. \tag{17}$$

Here $\gamma=N/2$ is fixed by the number N of input systems. α is constrained by the condition that π_α must be a subrepresentation of $\pi_{j=1/2}^{\otimes M}$, which is equivalent to $\alpha\leq M/2$. Finally, β is constrained by the condition that there must be a nonzero intertwiner between π_γ and $\pi_\alpha\otimes\pi_\beta$. It is well known that this condition is equivalent to the inequality $|\alpha-\beta|\leq\gamma\leq\alpha+\beta$. This is the same as the ''triangle inequality'': the sum of any two of α,β,γ is larger than the third. The area of admissible pairs (α,β) is represented in Fig. 1.

Since $x \mapsto x(x+1)$ is increasing for $x \geqslant 0$, we maximize ω with respect to β in Eq. (17) if we choose β as small as possible, i.e., $\beta = |\alpha - \gamma|$. Then the numerator in Eq. (17) becomes

$$\alpha(\alpha+1) - \beta(\beta+1) = 2\alpha\gamma - \gamma^2 + \max\{\gamma, 2\alpha - \gamma\},$$

which is strictly increasing in α. Hence the maximum

$$\omega_{\max} = \frac{M+2}{N+2} \tag{18}$$

is attained for and only for $\alpha = M/2$ and $\beta = (M-N)/2$.

Note that the seemingly simpler procedure of first maximizing α and then minimizing β to the smallest value consistent with $\alpha = M/2$ leads to the same result, but is fallacious because it fails to rule out possibly larger values of ω in the lower triangle of the admissible region in Fig. 1. The same problem arises for higher d, and one has to be careful to find a maximization procedure which takes into account all constraints.

E. Maximizing ω in the general case

Let us generalize now the previous discussion to arbitrary but finite d. In this case irreducible representations of U(d) are labeled, according to Appendix A 2 by their highest weight $\mathbf{m} = (m_1, \ldots, m_d)$. Hence we can decompose $T: \mathcal{B}(\mathcal{H}^{\otimes M}) \to \mathcal{B}(\mathcal{H}_+^{\otimes N})$ as described in the Proposition into the sum $T = \Sigma_{(\mathbf{m},\mathbf{n}) \in W} r_{\mathbf{m},\mathbf{n}} T_{\mathbf{m},\mathbf{n}}$, taken over the set

$$W = \{(\mathbf{m},\mathbf{n}) \in \mathbb{Z}_+^d \times \mathbb{Z}_+^d \,|\, \pi_{\mathbf{m}} \subset \pi_{\square}^{\otimes M} \quad \text{and} \quad \pi_N^+ \subset \pi_{\mathbf{m}} \otimes \pi_{\mathbf{n}}\}.$$

Here $\mathbb{Z}_+^d$ is an abbreviation for the set of all possible highest weights of irreducible U(d) representations, i.e., $\mathbb{Z}_+^d = \{(m_1, \ldots, m_d) | m_1 \geqslant m_2 \geqslant \ldots \geqslant m_d\}$.

Our task is now to determine $(\mathbf{m},\mathbf{n}) \in W$ such that $\omega = \omega(T_{\mathbf{m},\mathbf{n}})$ becomes maximal. To this end we consider in analogy to (15) the equation

$$\omega \cdot \partial \pi_N^+(X) = V^*(\partial \pi_{\mathbf{m}}(X) \otimes \mathbb{1}_n) V, \qquad \forall X \in \mathfrak{su}(d), \tag{19}$$

where V is an intertwining isometry between π_N^+ and $\pi_{\mathbf{m}} \otimes \pi_{\mathbf{n}}$. Note that Eq. (19) is valid only for $X \in \mathfrak{su}(d)$ [and not for $X \in \mathfrak{u}(d)$ in general]. Hence we have to consider the second-order Casimir operator $\tilde{\mathbf{C}}_2$ of SU(d) which is given, according to Appendix A 5, by an expression of the form $\tilde{\mathbf{C}}_2 = \Sigma_{jk} g^{jk} X_j X_k$. This is all we needed in the derivation of Eq. (16) in the SU(2)-case. The generalization to arbitrary d hence reads

$$\omega = \frac{1}{2} + \frac{\tilde{C}_2(\pi_{\mathbf{m}}) - \tilde{C}_2(\pi_{\mathbf{n}})}{2\tilde{C}_2(\pi_N^+)}. \tag{20}$$

The concrete form of $\tilde{C}_2(\pi_{\mathbf{m}})$ as a function of the weights $\mathbf{m}$ is given in Eq. (A8), and will be needed only later. Since $\tilde{C}_2(\pi_N^+)$ is a positive constant we have to maximize the function

$$W \ni (\mathbf{m},\mathbf{n}) \mapsto F(\mathbf{m},\mathbf{n}) = \tilde{C}_2(\pi_{\mathbf{m}}) - \tilde{C}_2(\pi_n) \in \mathbb{Z} \tag{21}$$

on its domain W.

The first step in this direction is to reexpress $F(\mathbf{m},\mathbf{n})$ in terms of the U(d) Casimir operators $\mathbf{C}_2$ and $\mathbf{C}_1^2$. Note in this context that although Eq. (19) is, as already stated, valid only for $X \in \mathfrak{su}(d)$ the representations $\pi_{\mathbf{m}}$ and $\pi_{\mathbf{n}}$ are still U(d) representations. Hence we can apply the equation $\tilde{\mathbf{C}}_2 = \mathbf{C}_2 - (1/d)\mathbf{C}_1^2$ given in Appendix A 5:

$$F(\mathbf{m},\mathbf{n}) = C_2(\pi_{\mathbf{m}}) - C_2(\pi_{\mathbf{n}}) - \frac{1}{d}(C_1^2(\pi_{\mathbf{m}}) - C_1^2(\pi_{\mathbf{n}})). \tag{22}$$

This rewriting is helpful, because the invariants C_1 turn out to be independent of the variational parameters: Since $\pi_{\mathbf{m}} \subset \pi_{\square}^{\otimes M}$, and $\partial\pi_{\square}^{\otimes M}(\mathbb{1}_d) = M\mathbb{1}$, we also have $C_1(\pi_{\mathbf{m}}) = M$. On the other hand, the existence of an intertwining isometry V with $V\pi_N^+ = \pi_{\mathbf{m}} \otimes \pi_{\mathbf{n}} V$ implies

$$VC_1(\pi_N^+)\mathbb{1} = V\partial\pi_N^+(\mathbf{C}_1) = (\partial\pi_{\mathbf{m}}(\mathbf{C}_1)\otimes \mathbb{1}_{\mathbf{n}} + \mathbb{1}_{\mathbf{m}} \otimes \partial\pi_{\mathbf{n}}(\mathbf{C}_1))V = (C_1(\pi_{\mathbf{m}})\mathbb{1} + C_1(\pi_{\mathbf{n}})\mathbb{1})V$$

and therefore $C_1(\pi_N^+) = C_1(\pi_{\mathbf{m}}) + C_1(\pi_{\mathbf{n}})$. Since $C_1(\pi_N^+) = N$ and $C_1(\pi_{\mathbf{m}}) = M$ we get $C_1(\pi_{\mathbf{n}}) = N - M$. Inserting this into Eq. (22) we find the functional

$$F(\mathbf{m},\mathbf{n}) = F_1(\mathbf{m},\mathbf{n}) - \frac{2MN - N^2}{d}, \tag{23}$$

where only F_1 depends on the variational parameters, and is expressed explicitly [see Eq. (A7)] as

$$W \ni (\mathbf{m},\mathbf{n}) \mapsto F_1(\mathbf{m},\mathbf{n}) = C_2(\pi_{\mathbf{m}}) - C_2(\pi_{\mathbf{n}}) = \sum_{j=1}^{d} (m_j^2 - n_j^2) + \sum_{k=1}^{d} (d - 2k + 1)(m_k - n_k) \in \mathbb{Z}, \tag{24}$$

which remains to be maximized over W.

To do this we have to express the constraints defining the domain W more explicitly. We have already seen that $\mathbf{m} \in \mathbb{Z}_+^d$ has to satisfy the constraint $\Sigma_{j=1}^d m_j = M$. In addition we get, due to Eq. (A1), $m_d > 0$. To fix the constraints for $\mathbf{n}$ note that according to Eq. (A4) $\pi_N^+ \subset \pi_{\mathbf{m}} \otimes \pi_{\mathbf{n}}$ is equivalent to $\pi_{\mathbf{m}} \subset \pi_N^+ \otimes \pi_{\tilde{\mathbf{n}}}$. Here we have introduced $\tilde{\mathbf{n}} = (\tilde{n}_1, \ldots, \tilde{n}_d) = (-n_d, \ldots, -n_1)$ as a notation for the highest weight of the representation $\overline{\pi_{\mathbf{n}}}$ conjugate to $\pi_{\mathbf{n}}$ (i.e., $\overline{\pi_{\mathbf{n}}} = \pi_{\tilde{\mathbf{n}}}$). Now we can apply Eq. (A3) to get

$$\pi_N^+ \subset \pi_{\mathbf{m}} \otimes \pi_{\mathbf{n}} \Leftrightarrow \tilde{n}_k = m_k - \mu_k$$

with

$$0 \leq \mu_k \leq m_k - m_{k+1} \forall k = 1,\ldots,d-1, \quad \sum_{k=1}^{d} \mu_k = N.$$

In other words

$$W = \{(\mathbf{m},\mathbf{n}) | \tilde{\mathbf{n}} = \mathbf{m} - \mu, \quad \text{and } (\mathbf{m},\mu) \in W_1\}$$

with

$$W_1 = \left\{ (\mathbf{m},\mu) \in \mathbb{Z}_+^d \times \mathbb{Z}^d \,\middle|\, \sum_{k=1}^{d} m_k = M, \quad \sum_{k=1}^{d} \mu_k = N \quad \text{and } 0 \leq \mu_k \leq m_k - m_{k+1} \forall k = 1,\ldots,d-1 \right\}.$$

The function F_1 can now be re-expressed in terms the new variables $(\mathbf{m},\mu)$. To this end note that $C_2(\pi_{\mathbf{n}}) = C_2(\overline{\pi_{\mathbf{n}}}) = C_2(\pi_{\tilde{\mathbf{n}}})$. Hence we have

$$F_1(\mathbf{m},\mathbf{n}) = F_1(\mathbf{m},\tilde{\mathbf{n}}) = F_1(\mathbf{m},\mathbf{m} - \mu)$$

and therefore with Eq. (24):

$$F_1(\mathbf{m},\tilde{\mathbf{n}}) = \sum_{k=1}^{d} \mu_k(2m_k - 2k - \mu_k) + (d+1)\sum_{k=1}^{d} \mu_k = F_2(\mathbf{m},\mu) + (d+1)N \tag{25}$$

with the new function

$$W_1 \ni (\mathbf{m},\mu) \mapsto F_2(\mathbf{m},\mu) = \sum_{k=1}^{d} \mu_k(2m_k - 2k - \mu_k) \in \mathbb{Z}. \tag{26}$$

Hence we have reduced our problem to the following Lemma:

Lemma 6: The function $F_2 : W_1 \to \mathbb{Z}$ defined in Eq. (26) attains its maximum for and only for

$$\mathbf{m}_{\max} = (M,0,...,0) \quad \textit{and} \quad \mu_{\max} = \begin{cases} (N,0,...,0) & \textit{for } N \leqslant M \\ (M,0,...,0,N-M) & \textit{for } N \geqslant M. \end{cases}$$

Proof: We consider a number of cases in each of which we apply a different strategy for increasing F_2. In these procedures we consider d to be a variable parameter, too, because if $\mu_d = m_d = 0$, the further optimization will be treated as a special case of the same problem with d reduced by one.

Case A: $\mu_d > 0$, $\mu_i < m_i - m_{i+1}$ for some $i < d$.

In this case we apply the substitution $\mu_i \mapsto (\mu_i + 1)$, $\mu_d \mapsto (\mu_d - 1)$, which leads to the change

$$\delta F_2 = 2(-\mu_i + \mu_d + (d-i-1) + m_{i+1} - m_d) \geqslant 2(\mu_d + (d-i-1)) > 0$$

in the target functional. In this way we proceed until either all μ_i with $i < d$ satisfy the upper bound with equality (Case B below) or $\mu_d = 0$, i.e., Case C or Case D applies.

Case B: $\mu_d > 0$, $\mu_i = m_i - m_{i+1}$ for all $i < d$. In this case all μ_k, including μ_d, are determined by the m_k and by the normalization ($\mu_d = N - m_1 + m_d$). Inserting these values into F_2, and using the normalization conditions, we get $F_2(\mathbf{m},\mathbf{n}) = F_3(\mathbf{m}) - 2(M + dN) - N^2$ with

$$F_3(\mathbf{m}) = 2(N+d)m_1$$

constrained by

$$m_1 \geqslant \ldots \geqslant m_d \geqslant 0, \quad \sum_k m_k = M, \quad \text{and} \quad m_1 - m_d \leqslant N.$$

This defines a variational problem in its own right. Any step increasing m_1 at the expense of some other m_k increases F_2. This process terminates either when $M = m_1$, and all other $m_k = 0$. This is surely the case for $M < N$, because then $\mu_d = N - m_1 + m_d \geqslant N - M > 0$. This is already the final result claimed in the Lemma. On the other hand, the process may terminate because μ_d reaches 0 or would become negative. In the former case we get $\mu_d = 0$, and hence Case C or Case D. The latter case (termination at $\mu_d = 1$) may occur because the transformation $m_1 \mapsto (m_1 + 1)$, $m_d \mapsto (m_d - 1)$ changes $\mu_d = N - m_1 + m_d$ by -2. There are two basic situations in which changing both m_1 and m_d is the only option for maximizing F_3, namely $d = 2$ and $m_1 = m_2 = \cdots = m_d$. The first case is treated below as Case E. In the latter case we have $1 = N - m_1 + m_d = N$. Then the overall variational problem in the Lemma is trivial, because only one term remains, and one only has to maximize the quantity $2m_k - 2k - 1$, with trivial maximum at $k = 1$, $m_1 = M$.

Case C: $\mu_d = 0$, $m_d > 0$. For $\mu_d = 0$, the number m_d does not enter in the function F_2. Therefore, the move $m_d \mapsto 0$ and $m_1 \mapsto m_1 + m_d$, increases F_2 by $\mu_1 m_d \geqslant 0$. Note that this is always compatible with the constraints, and we end up in Case D.

Case D: $\mu_d = 0$, $m_d = 0$, $d > 2$. Set $d \mapsto (d-1)$. Note that we could now use the extra constraint $\mu_{d'} \leqslant m_{d'}$, where $d' = d - 1$. We will *not* use it, so in principle we might get a larger maximum. However, since we do find a maximizer satisfying all constraints, we still get a valid maximum.

Case E: $d = 2$, $\mu_1 = m_1 - m_2$, $\mu_2 = 1$. In this case $\mathbf{m} = (m_1, m_2)$ is completely fixed by the constraints. We have: $m_1 + m_2 = M$ and $\mu_1 + \mu_2 = m_1 - m_2 + 1 = N$, hence $m_1 - m_2 = N - 1$. This implies $2m_1 = M + N - 1$, $2m_2 = M - N + 1$ and since $m_2 \geqslant 0$ we get $M \geqslant N - 1$. If $M = N - 1$ holds we get $m_1 = N - 1 = M$, $m_2 = 0$ and consequently $\mu_1 = N - 1$. Together with $\mu_2 = 1 = N - M$ these are exactly the parameters where F_2 should take its maximum according to the Lemma. Hence assume $M \geqslant N$. In this case $\mu_2 = 1$ implies that F_2 becomes $NM - 3N - 4$, which is, due to $M \geqslant N$, strictly smaller than $F_2(M,0;N,0) = 2MN - N^2 - 2N$.

Uniqueness: In all cases just discussed the manipulations described lead to a strict increase of $F_2(\mathbf{m},\mu)$ as long as $(\mathbf{m},\mu) \neq (\mathbf{m}_{\max},\mu_{\max})$ holds. The only exception is Case C with $\mu_1 = 0$. In this situation there is a $1 < k < d$ with $\mu_k > 0$. Hence we can apply the maps $d \mapsto d-1$ (Case D) and $m_d \mapsto 0$ and $m_1 \mapsto m_1 + m_d$ (Case C) until we get $\mu_d \neq 0$ (i.e., d reaches k). Since $\mu_1 = 0$ the corresponding $(\mathbf{m},\mu)$ is not equal to $(\mathbf{m}_{\max},\mu_{\max})$. Therefore we can apply one of the manipulations described in Case A, Case B, or Case E which leads to a strict increase of $F_2(\mathbf{m},\mu)$. This shows that $F_2(\mathbf{m},\mu) < F_2(\mathbf{m}_{\max},\mu_{\max})$ as long as $(\mathbf{m},\mu) \neq (\mathbf{m}_{\max},\mu_{\max})$ holds. Consequently the maximum is unique. □

With this result and Eqs. (20), (21), (23), (25), and (26) we can easily calculate $\omega_{\max}$:

$$\omega_{\max} = \omega(\hat{T}) = \frac{M+d}{N+d}$$

and with (11) we get $\Delta(T) \geqslant \Delta(\hat{T})$ with $\Delta(\hat{T})$ from Theorem 1.

F. Proving uniqueness

One part of the uniqueness proof is already given above: There is only one optimal *covariant* cloning map, namely $\hat{T}$. This follows easily from the uniqueness of the maximum found in Lemma 6 and from the fact that the representation π_N^+ is contained exactly once in the tensor product $\pi_M^+ \otimes \overline{\pi_{M-n}^+}$ [see Eq. (A3) and the discussion in Sec. III C].

Suppose now that T is a noncovariant cloning map, which also attains the best value: $\Delta_{\text{one}}(T) = \Delta_{\text{one}}(\hat{T})$. Then we may consider the average $\overline{T}$ of T [see Eq. (6)], which is also optimal and, in addition, covariant. Therefore $\overline{T} = \hat{T}$. The uniqueness part of the proof thus follows immediately from the following proposition:

Proposition 7: Each completely positive, unital map $T: \mathcal{B}(\mathcal{H}^{\otimes M}) \rightarrow \mathcal{B}(\mathcal{H}_+^{\otimes N})$ satisfying the equation $\overline{T} = \hat{T}$ equals $\hat{T}$.

Proof: We trace back this statement to the main theorem of Ref. 16. To this end note that $\overline{T} = \hat{T}$ implies the equivalent equation for the preduals:

$$\overline{T_*} = \int \tau_u T_* \, du = \hat{T}_*,$$

where τ_u acts on T_* by

$$\tau_u T_*(\sigma) = \pi_{\square}^{\otimes M}(u)^* T_*(\pi_N^+(u)\sigma\pi_N^+(u)^*)\pi_{\square}^{\otimes M}(u).$$

Furthermore we know from the main theorem of Ref. 16 that $\text{tr}(\sigma^{\otimes M} T_*(\sigma^{\otimes N})) \leqslant d[N]/d[M]$ is true for all pure states $\sigma \in \mathcal{B}_*(\mathcal{H})$ and that equality holds iff $T = \hat{T}$. Consequently we have

$$\int \left(\frac{d[N]}{d[M]} - \text{tr}(\sigma^{\otimes M} \tau_u T(\sigma^{\otimes N})) \right) du = \frac{d[N]}{d[M]} - \text{tr}(\sigma^{\otimes M} \overline{T}_*(\sigma^{\otimes N})) = \frac{d[N]}{d[M]} - \text{tr}(\sigma^{\otimes M} \hat{T}_*(\sigma^{\otimes N})) = 0.$$

Since the integral on the left-hand site of this equation is taken over positive quantities the integrand has to vanish for all values of $u \in \text{U}(d)$. This implies $\text{tr}(\sigma^{\otimes M} T(\sigma^{\otimes N})) = d[N]/d[M]$ for all pure states $\sigma \in \mathcal{B}_*(\mathcal{H})$. However this is, according to Ref. 16, only possible if $T = \hat{T}$. □

ACKNOWLEDGMENTS

This paper is a response to the many discussions about the cloning problem one of us (R.F.W.) had with participants of the ISI Workshop on Quantum Computing in Torino in July 1997. We would like to thank D. Sondermann for a critical reading of the manuscript.

APPENDIX A: REPRESENTATIONS OF UNITARY GROUPS

Throughout this paper many arguments from representation theory of unitary groups are used. In order to fix the notation and to state the most relevant theorems we will recall in this appendix some well-known facts from representation theory of Lie groups. General references are the books of Barut and Raczka,[22] Zhelobenko,[23] and Simon.[18]

1. The groups and their Lie algebras

Let us consider first the group $\mathrm{U}(d)$ of all complex $d\times d$ unitary matrices. Its Lie algebra $\mathfrak{u}(d)$ can be identified with the Lie algebra of all anti-Hermitian $d\times d$ matrices. The exponential function is then given by the usual matrix exponential $X\mapsto\exp(X)$. $\mathfrak{u}(d)$ is a real Lie algebra. Hence we can consider its complexification $\mathfrak{u}(d)\otimes\mathbb{C}$ which coincides with the set of all $d\times d$ matrices and at the same time with the Lie algebra $\mathfrak{gl}(d,\mathbb{C})$ of the general linear group $\mathrm{GL}(d,\mathbb{C})$. In other words $\mathfrak{u}(d)$ is a real form of $\mathfrak{gl}(d,\mathbb{C})$. A basis of $\mathfrak{gl}(d,\mathbb{C})$ is given by the matrices $E_{jk}=|j\rangle\langle k|$.

The set of elements of $\mathrm{U}(d)$ with determinant one forms the subgroup $\mathrm{SU}(d)$ of $\mathrm{U}(d)$. Its Lie algebra $\mathfrak{su}(d)$ is the subalgebra of $\mathfrak{u}(d)$ consisting of the elements with zero trace. Hence the complexification $\mathfrak{su}(d)\otimes\mathbb{C}$ of $\mathfrak{su}(d)$ is the Lie algebra of trace-free matrices and coincides therefore with the Lie algebra $\mathfrak{sl}(d,\mathbb{C})$ of the special linear group $\mathrm{SL}(d,\mathbb{C})$. As well as in the $\mathrm{U}(d)$ case this means that $\mathfrak{su}(d)$ is a real from of $\mathfrak{sl}(d,\mathbb{C})$. The matrices E_{jk} are no longer a basis for $\mathfrak{sl}(d,\mathbb{C})$ since the E_{jj} are not trace free. Instead we have to consider E_{jk}, $j\neq k$ and $H_j=E_{jj}-E_{j+1,j+1}$, $j=1,...,d-1$. The difference between $\mathfrak{sl}(d,\mathbb{C})$ and $\mathfrak{gl}(d,\mathbb{C})$ is exactly the center of $\mathfrak{gl}(d,\mathbb{C})$, i.e., all complex multiples of the identity matrix. In other words we have $\mathfrak{gl}(d,\mathbb{C})=\mathfrak{sl}(d,\mathbb{C})\oplus\mathbb{C}\mathbb{1}$. A similar result holds for the real forms: $\mathfrak{u}(d)=\mathfrak{su}(d)\oplus\mathbb{R}\mathbb{1}$.

The (real) span of all iE_{jj}, $j=1,...,d$ is a subalgebra of $\mathfrak{u}(d)$ which is maximal Abelian, i.e., a Cartan subalgebra of $\mathfrak{u}(d)$. We will denote it in the following by $\mathfrak{t}(d)$ and its complexification by $\mathfrak{t}_{\mathbb{C}}(d)\subset\mathfrak{gl}(d,\mathbb{C})$. The intersection of $\mathfrak{t}(d)$ with $\mathfrak{su}(d)$ results in a Cartan subalgebra $\mathfrak{st}(d)$ of $\mathfrak{su}(d)$. We will denote the complexification by $\mathfrak{st}_{\mathbb{C}}(d)$. Again the two algebras $\mathfrak{t}(d)$ and $\mathfrak{st}(d)$ differ by the center of $\mathfrak{u}(d)$, i.e., $\mathfrak{t}(d)=\mathfrak{st}(d)\oplus\mathbb{R}\mathbb{1}$ and $\mathfrak{t}_{\mathbb{C}}(d)=\mathfrak{st}_{\mathbb{C}}(d)\oplus\mathbb{C}\mathbb{1}$ in the complexified case.

2. Representations

Consider now a finite-dimensional[24] representation $\pi:\mathrm{U}(d)\to\mathrm{GL}(N,\mathbb{C})$ of $\mathrm{U}(d)$. It is characterized uniquely by the corresponding representation $\partial\pi:\mathfrak{u}(d)\to\mathfrak{gl}(N,\mathbb{C})$ of its Lie algebra, i.e., we have $\pi(\exp(X))=\exp(\partial\pi(X))$. The representation $\partial\pi$ can be extended by complex linearity to a representation of $\mathfrak{gl}(d,\mathbb{C})$ which we will denote by $\partial\pi$ as well. Hence $\partial\pi$ leads to a representation π of the group $\mathrm{GL}(d,\mathbb{C})$. Similar notations we will adopt for representations of $\mathrm{SU}(d)$ and $\mathrm{SL}(d,\mathbb{C})$.

Assume now that π is an irreducible representation of $\mathrm{GL}(d,\mathbb{C})$. An infinitesimal weight of π (or simply a weight in the following) is an element λ of the dual of $\mathfrak{t}_{\mathbb{C}}^*(d)$ of $\mathfrak{t}_{\mathbb{C}}(d)$ such that $\partial\pi(X)x=\lambda(X)x$ holds for all $X\in\mathfrak{t}_{\mathbb{C}}(d)$ and for a nonvanishing $x\in\mathbb{C}^N$. The linear subspace $V_\lambda\subset\mathbb{C}^N$ of all such x is called the weight subspace of the weight λ. The set of weights of π is not empty and, due to irreducibility, there is exactly one weight $\mathbf{m}$, called the highest weight, such that $\partial\pi(E_{jk})x=0$ for all x in the weight subspace of $\mathbf{m}$ and for all $j,k=1,...,d$ with $j<k$. The representation π is (up to unitary equivalence) uniquely determined by its highest weight. On the other hand the weight $\mathbf{m}$ is uniquely determined by its values $\mathbf{m}(E_{jj})=m_j$ on the basis E_{jj} of $\mathfrak{t}_{\mathbb{C}}(d)$. We will express this fact in the following as ''$\mathbf{m}=(m_1,\ldots,m_d)$ is the highest weight of the representation π.'' For each analytic representation of $\mathrm{GL}(d,\mathbb{C})$ the m_j are integers satisfying the inequalities $m_1\geqslant m_2\geqslant\cdots\geqslant m_d$ and the converse is also true: each family of integers with this property defines the highest weight of an analytic, irreducible representation of $\mathrm{GL}(d,\mathbb{C})$.

In a similar way we can define weights and highest weights for representations of the group $SL(d,\mathbb{C})$ as linear forms on the Cartan subalgebra $\mathfrak{st}_{\mathbb{C}}(d)$. As in the $GL(d,\mathbb{C})$-case an irreducible representation π of $SL(d,\mathbb{C})$ is characterized uniquely by its highest weight $\mathbf{m}$. However we cannot evaluate $\mathbf{m}$ on the basis E_{jj} since these matrices are not trace free. One possibility is to consider an arbitrary extension of $\mathbf{m}$ to the algebra $\mathfrak{t}_{\mathbb{C}}(d)=\mathfrak{st}_{\mathbb{C}}(d)\oplus\mathbb{C}\mathbb{1}$. Obviously this extension is not unique. Therefore the values $\mathbf{m}(E_{jj})=m_j$ are unique only up to an additive constant. To circumvent this problem we will use usually the normalization condition $m_d=0$. In this case the integer m_j corresponds to the number of boxes in the jth row of the Young tableau usually used to characterize the irreducible representation π. Another possibility to describe the weight $\mathbf{m}$ is to use the basis H_j of $\mathfrak{st}_{\mathbb{C}}(d)$. We get a sequence of integers $l_j=\mathbf{m}(H_j)$, $j=1,...,d-1$. They are related to the m_j by $l_j=m_j-m_{j+1}$. Each sequence $l_1,\ldots,l_{d-1}$ defines the highest weight of an irreducible representation of $SL(d,\mathbb{C})$ iff the l_j are positive integers.

Finally consider the representation $\overline{\pi}$ conjugate to π, i.e., $\overline{\pi}(u)=\overline{\pi(u)}$. If π is irreducible the same is true for $\overline{\pi}$. Hence $\overline{\pi}$ admits a highest weight which is given by $(-m_d,-m_{d-1},\ldots,-m_1)$. If π is a $SU(d)$ representation we can apply the normalization $m_d=0$. Doing this as well for the conjugate representation we get $(m_1,m_1-m_{d-1},\ldots,m_1-m_2,0)$. In terms of Young tableaus this corresponds to the usual rule to construct the tableau of the conjugate representation: Complete the Young tableau of π to form a $d\times m_1$ rectangle. The complementary tableau rotated by 180° is the Young tableau of $\overline{\pi}$.

3. Tensor products of representations

Consider now two finite dimensional irreducible representations $\pi_{\mathbf{m}}$, $\pi_{\mathbf{n}}$ of $U(d)$ with highest weights $\mathbf{m},\mathbf{n}$. Their tensor product $\pi_{\mathbf{m}}\otimes\pi_{\mathbf{n}}$ is completely reducible. If r_π denotes the multiplicity of the irreducible representation π in $\pi_{\mathbf{m}}\otimes\pi_{\mathbf{n}}$ then this means that $\pi_{\mathbf{m}}\otimes\pi_{\mathbf{n}}=\oplus_\pi r_\pi\pi$. Hence to decompose the representation $\pi_{\mathbf{m}}\otimes\pi_{\mathbf{n}}$ we have to compute the integer valued functions $(\mathbf{m},\mathbf{n})\mapsto r_\pi(\mathbf{m},\mathbf{n})$. There are several general schemes to do this (see, e.g., Chap. XII of Ref. 23). However we are only interested in the following special cases. The highest weight of the representation $\pi_{\mathbf{1}}: U(d)\ni U\mapsto U\in GL(d,\mathbb{C})$ (denoted $\pi_{\square}$ in Sec. III A) is $\mathbf{1}=(1,0,...,0)$. Consider the N-fold tensor product of this representation. It can be decomposed as follows:

$$\pi_{\mathbf{1}}^{\otimes N}=\sum_{\substack{m_1+\cdots+m_d=N\\ m_d\geqslant 0}} r(m_1,\ldots,m_d)\,\pi_{m_1,\ldots,m_d}, \tag{A1}$$

where $\pi_{m_1,\ldots,m_d}$ denotes the irreducible representation with highest weight $(m_1,\ldots,m_d)$. The coefficients $r(m_1,\ldots,m_d)$ are determined by the following recurrence relation:

$$r(m_1,\ldots,m_d)=r(m_1-1,...,m_d)+r(m_1,m_2-1,...,m_d)+\cdots+r(m_1,\ldots,m_d-1). \tag{A2}$$

Consider now the N-fold symmetric tensor product of $\pi_{\mathbf{1}}$ (denoted π_N^+ in Sec. III A). It is irreducible with highest weight $N\mathbf{1}=(N,0,...,0)$ (hence $\pi_N^+=\pi_{N\mathbf{1}}$). The tensor product of this representation with an arbitrary irreducible representation $\pi_{\mathbf{m}}$ [with highest weight $\mathbf{m}=(m_1,\ldots,m_d)$] is

$$\pi_{N\mathbf{1}}\otimes\pi_{\mathbf{m}}=\sum_{\substack{0\leqslant\mu_{k+1}\leqslant m_k-m_{k+1}\\ \mu_1+\cdots+\mu_d=N}} \pi_{m_1+\mu_1,\ldots,m_d+\mu_d}. \tag{A3}$$

From Eq. (A3) we also get a condition for $\pi_{N\mathbf{1}}$ to be contained in an arbitrary tensor product $\pi_{\mathbf{m}}\otimes\pi_{\mathbf{n}}$ which we need in Sec. III E: For arbitrary weights $\mathbf{m},\mathbf{n},\mathbf{p}$ we have

$$\pi_{\mathbf{m}}\subset\pi_{\mathbf{n}}\otimes\pi_{\mathbf{p}}\Leftrightarrow\pi_{\mathbf{n}}\subset\overline{\pi_{\mathbf{p}}}\otimes\pi_{\mathbf{m}}. \tag{A4}$$

If two irreducible representations $\pi_{\mathbf{m}}$, $\pi_{\mathbf{n}}$ of SU(d) are given we can characterize them, as described above, by their highest weights $\mathbf{m}=(m_1,\ldots,m_d)$ and $\mathbf{n}=(n_1,\ldots,n_d)$ using the normalizations $m_d=0$ and $n_d=0$. After applying the stated theorems to the tensor product of the corresponding U(d) representations we can restrict the summands in the resulting spectral decomposition back to SU(d), i.e., we renormalize the highest weights $(m_1,\ldots,m_d)$ to the $m_d=0$ case.

4. Nondegeneracy of $\mathfrak{su}(d)$

We are now ready to discuss the group theoretic part of the proof of our main theorem, i.e., Lemma 3 which we have only stated in Sec. III. According to Definition 2 we have to show that each linear operator $L:\mathfrak{su}(d)\to\mathcal{H}_+^{\otimes N}$ with the covariance property

$$\pi_N^+(g)L(X)\pi_N^+(g^{-1})=L(gXg^{-1}) \tag{A5}$$

is of the form $L(X)=\lambda\partial\pi_+^N(X)$ with a constant factor λ. Here π_N^+ is the irreducible representation of SU(d) introduced in Sec. III A. (Hence we have $\pi_N^+=\pi_{N\mathbf{1}}$ using the notation introduced in Appendix A 3.)

To reformulate this statement note first that the map $g\mapsto\pi_N^+(g)\cdot\pi_N^+(g^{-1})$ can be interpreted as a unitary representation of SU(d) on the representation space $\mathcal{H}_+^{\otimes N}\otimes\mathcal{H}_+^{\otimes N}$. In fact it is (unitarily equivalent to) the tensor product $\pi_N^+\otimes\overline{\pi_N^+}$. Since $\mathrm{SU}(d)\ni g\mapsto g\cdot g^{-1}\in\mathcal{B}(\mathfrak{su}(d))$ is the adjoint representation of SU(d) this implies that each map X satisfying (A5) intertwines $\pi_N^+\otimes\overline{\pi_N^+}$ and the adjoint representation Ad. Note second that the representation $\partial\pi_N^+$ of the Lie algebra $\mathfrak{su}(d)$ satisfies Eq. (A5) in an obvious way (with $\lambda=1$) hence we have to show that all such intertwiners are proportional, or in other words that Ad is contained in $\pi_N^+\otimes\overline{\pi_N^+}$ exactly once.

Let us discuss now the tensor product $\pi_N^+\otimes\overline{\pi_N^+}$. The irreducible representation π_N^+ has highest weight $(N,0,...,0)$ (see Appendix A 2) and consequently the highest weight of its conjugate is $(N,\ldots,N,0)$. We can apply now Eq. (A3), which shows that the adjoint representation whose highest weight is $(2,1,...,1,0)$ is contained in $\pi_N^+\otimes\overline{\pi_N^+}$ exactly once. This shows together with our previous discussion that $\mathfrak{su}(d)$ is nondegenerate in $\mathcal{H}_+^{\otimes N}$ with respect to π_N^+.

5. The Casimir invariants

To each Lie algebra $\mathfrak{g}$ we can associate its universal enveloping algebra $\mathfrak{G}$. It is defined as the quotient of the full tensor algebra $\oplus_{n\in\mathbb{N}_0}\mathfrak{g}^{\otimes n}$ with the two sided ideal $\mathfrak{I}$ generated by $X\otimes Y-Y\otimes X-[X,Y]$, i.e., $\mathfrak{G}$ is an associative algebra. The original Lie algebra $\mathfrak{g}$ can be embedded in its enveloping algebra $\mathfrak{G}$ by $\mathfrak{g}\ni X\mapsto X+\mathfrak{I}\in\mathfrak{G}$. The Lie bracket is then simply given by $[X,Y]=XY-YX$. Moreover $\mathfrak{G}$ is algebraically generated by $\mathfrak{g}$ and $\mathbb{1}$. Hence each representation $\partial\pi$ of $\mathfrak{g}$ generates a unique representation $\partial\pi$ of $\mathfrak{G}$ simply by $\partial\pi(X_1\cdots X_k)=\partial\pi(X_1)\cdots\partial\pi(X_k)$. If $\partial\pi$ is irreducible the induced representation $\partial\pi$ is irreducible as well.

We are interested not in the whole algebra but only in its center $\mathfrak{Z}(\mathfrak{G})$, i.e., the subalgebra consisting of all $Z\in\mathfrak{G}$ commuting with all elements of $\mathfrak{G}$. The elements of $\mathfrak{Z}(\mathfrak{G})$ are called central elements or Casimir elements. If $\partial\pi$ is a representation of $\mathfrak{G}$ the representatives $\partial\pi(Z)$ of Casimir elements commute with all other representatives $\partial\pi(X)$. This implies for irreducible representations that all $\partial\pi(Z)$ are multiples of the identity.

Consider now the case $\mathfrak{g}=\mathfrak{gl}(d,\mathbb{C})$. In this case we can identify the enveloping algebra $\mathfrak{G}$ with the set of all left invariant differential operators on GL($d,\mathbb{C}$) (a similar statement is true for any Lie group). Of special interest for us are the Casimir elements belonging to operators of first and second order. Using the standard basis E_{ij} of $\mathfrak{gl}(d,\mathbb{C})$ introduced in Appendix A 1 they are given by

$$\mathbf{C}_1=\sum_{j=1}^{d}E_{jj}\,,\quad \mathbf{C}_2=\sum_{j,k=1}^{d}E_{jk}E_{kj}\,.$$

Of course $\mathbf{C}_1^2$ is as well of second order and it is linearly independent of $\mathbf{C}_2$. Hence each second-order Casimir element of $\mathfrak{G}$ is a linear combination of $\mathbf{C}_2$ and $\mathbf{C}_1^2$.

If $\partial\pi$ is an irreducible representation of $\mathfrak{gl}(d,\mathbb{C})$ with highest weight $(m_1,\dots,m_d)$ it induces, as described above, an irreducible representation $\partial\pi$ of $\mathfrak{G}$ and the images of $\partial\pi(\mathbf{C}_1)$ and $\partial\pi(\mathbf{C}_2)$ are multiples of the identity, i.e., $\partial\pi(\mathbf{C}_1)=C_1(\pi)\mathbb{1}$ and $\partial\pi(\mathbf{C}_2)=C_2(\pi)\mathbb{1}$ with

$$C_1(\pi)=\sum_{j=1}^{d} m_j, \quad C_2(\pi)=\sum_{j=1}^{d} m_j^2+\sum_{j<k}(m_j-m_k). \tag{A6}$$

Let us discuss now the Casimir elements of $\mathrm{SL}(d,\mathbb{C})$. Since $\mathrm{SL}(d,\mathbb{C})$ is a subgroup of $\mathrm{GL}(d,\mathbb{C})$ its enveloping algebra $\mathfrak{S}$ is a subalgebra of $\mathfrak{G}$. However the corresponding Lie algebras differ only by the center of $\mathfrak{gl}(d,\mathbb{C})$. Hence the center $\mathfrak{Z}(\mathfrak{S})$ of $\mathfrak{G}$ is a subalgebra of $\mathfrak{Z}(\mathfrak{G})$. Since $\mathfrak{sl}(d,\mathbb{C})$ is simple there is no first-order Casimir element and there is only one second-order Casimir element $\tilde{\mathbf{C}}_2$ which is therefore a linear combination $\tilde{\mathbf{C}}_2=\mathbf{C}_2+\alpha\mathbf{C}_1^2$ of $\mathbf{C}_1^2$ and $\mathbf{C}_2$. Obviously the factor α is uniquely determined by the condition that the expression

$$\tilde{C}_2(\pi)=C_1(\pi)+\alpha C_1^2(\pi)=\sum_{j=1}^{d} m_j^2+\sum_{j<k}(m_j-m_k)+\alpha\left(\sum_{j=1}^{d} m_j\right)^2 \tag{A7}$$

with $\partial\pi(\tilde{\mathbf{C}}_2)=\tilde{C}_2(\pi)\mathbb{1}$ is invariant under the renormalization $(m_1,\dots,m_d)\mapsto(m_1+\mu,\dots,m_d+\mu)$ with an arbitrary constant μ. Straightforward calculations show that $\alpha=-1/d$. Hence we get $\tilde{\mathbf{C}}_2=\mathbf{C}_2-(1/d)\mathbf{C}_1^2$ and

$$\tilde{C}_2(\pi)=\frac{1}{d}\left((d-1)\sum_{j=1}^{d} m_j^2-\sum_{j\neq k}^{d} m_j m_k+d\sum_{j<k}(m_j-m_k)\right). \tag{A8}$$

Alternatively $\tilde{\mathbf{C}}_2$ can be expressed in terms of a basis $(X_j)_j$ of $\mathfrak{sl}(d,\mathbb{C})$. In fact there is a symmetric second rank tensor $g^{jk}X_j\otimes X_k\in\mathfrak{sl}(d,\mathbb{C})\otimes\mathfrak{sl}(d,\mathbb{C})$ such that $\tilde{\mathbf{C}}_2$ coincides with the equivalence class of g^{jk} in $\mathfrak{S}$. In other words $\tilde{\mathbf{C}}_2=\Sigma_{jk}g^{jk}X_jX_k$ holds which leads to

$$\tilde{C}_2(\pi)\mathbb{1}=\sum_{jk} g^{jk}\partial\pi(X_i)\partial\pi(X_j)$$

for an irreducible representation π of $\mathrm{SU}(d)$.

APPENDIX B: STINESPRING THEOREM FOR COVARIANT CP-MAPS

In this appendix we will state the covariant version of Stinespring's theorem[20] which we have used in the proof of Theorem 1. However, as in the rest of the paper, we will restrict the discussion to finite dimensional Hilbert spaces (i.e., only cp-maps between finite von Neumann factors are considered).

Theorem 8: *Let G be a group with finite dimensional unitary representations* $\pi_i:G\to\mathcal{B}(\mathcal{H}_i)(i=1,2)$, *and* $T:\mathcal{B}(\mathcal{H}_2)\to\mathcal{B}(\mathcal{H}_1)$ *a completely positive map with the covariance property* $\pi_1(g)T(X)\pi_1(g)^*=T(\pi_2(g)X\pi_2(g)^*)$.

(1) Then there is another finite dimensional unitary representation $\tilde{\pi}:G\to\mathcal{B}(\tilde{\mathcal{H}})$ and an intertwiner $V:\mathcal{H}_1\to\mathcal{H}_2\otimes\tilde{\mathcal{H}}$ with $V\pi_1(g)=\pi_2\otimes\tilde{\pi}V$ such that $T(X)=V^*(X\otimes\mathbb{1})V$ holds.
(2) If $T=\Sigma_\alpha T^\alpha$ is a decomposition of T in completely positive terms, there is a decomposition $\mathbb{1}=\Sigma_\alpha F^\alpha$ of the identity operator on $\tilde{\mathcal{H}}$ into positive operators $F^\alpha\in\mathcal{B}(\tilde{\mathcal{H}})$ with $[F^\alpha,\tilde{\pi}(g)]=0$ such that $T^\alpha(X)=V^*(X\otimes F^\alpha)V$.

We only sketch the main ideas of the proof. The first step is Stinespring's theorem in its general form:[19] There exists a representation $\eta:\mathcal{B}(\mathcal{H}_2)\to\mathcal{B}(\mathcal{K})$ of the C^*-algebra $\mathcal{B}(\mathcal{H}_2)$ on a

Hilbert space $\mathcal{K}$ and a bounded operator $V:\mathcal{H}_1\rightarrow\mathcal{K}$ such that $T(X)=V^*\eta(X)V$ holds. Up to unitary equivalence there is exactly one such triple $(\mathcal{K},V,\pi)$ such that the vectors $\pi(A)V\psi\in\mathcal{K}$ with $\psi\in\mathcal{H}_1$ and $A\in\mathcal{B}(\mathcal{H}_2)$ span $\mathcal{K}$.

It is this uniqueness, from which the representation $\tilde{\pi}$ of G is constructed. Indeed, the objects $V_g=V\pi_1(g)$, and $\eta_g(X)=\eta(\pi_2(g)X\pi_2(g)^*)$ form a Stinespring dilation of the completely positive map $T_g(X)=\pi_1(g)^*T(\pi_2(g)X\pi_2(g)^*)\pi_1(g)$, which by covariance is equal to T. Hence by ''uniqueness up to unitary equivalence'' there is a unique unitary operator $U_g\in\mathcal{B}(\mathcal{K})$ such that $V_g=V\pi_1(g)=U_gV$, and $\eta_g(X)=\eta(\pi_2(g)X\pi_2(g)^*)=U_g\eta(X)U_g^*$. This can be simplified a bit further by the observation that according to the second equation the operators $\tilde{U}_g=\eta(\pi_2(g))^*U_g$ commute with all $\eta(X)$. It is easy to see that the U_g are a representation, and hence so is $\tilde{U}$: we have $\tilde{U}_g\tilde{U}_h=\eta(\pi_2(g))^*U_g\eta(\pi_2(h)^*)U_h=\eta(\pi_2(g))^*\eta(\pi_2(g)\pi_2(h)^*\times\pi_2(g)^*)U_gU_h=\eta(\pi_2(g)^*\pi_2(g)\pi_2(h)^*\pi_2(g)^*)U_{gh}=\eta(\pi_2(gh)^*)U_{gh}=\tilde{U}_{gh}$.

For a proof of part (1) we now only need to invoke the observation that all representations of $\mathcal{B}(\mathcal{H}_2)$ are of the form $\eta\simeq\mathrm{id}\otimes\mathbb{1}$ with $\mathcal{K}=\mathcal{H}_2\otimes\tilde{\mathcal{H}}$. (Here ''$\simeq$'' denotes a unitary equivalence, which we will include as a factor in V.) Since $\tilde{U}_g$ commutes with all $\eta(X)=X\otimes\mathbb{1}$, it is of the form $\tilde{U}_g=\mathbb{1}\otimes\tilde{\pi}(g)$, which proves the assertion.

The second part of the theorem stated for a trivial group $G=\{e\}$ is also known as the Radon–Nikodyn theorem coming with the Stinespring theorem. In general it asserts the existence of a partition of the identity operator on $\mathcal{K}$ into operators $\tilde{F}^\alpha$ commuting with all $\eta(X)$, giving the decomposition of T as $T^\alpha=V^*\eta(X)F^\alpha V$. Again, we can write these as $\tilde{F}^\alpha=\mathbb{1}\otimes F^\alpha$. Since the $F^\alpha V$ are uniquely determined by the T^α, it is easy to see that covariance of T^α is equivalent to $F^\alpha=\tilde{\pi}_gF^\alpha\tilde{\pi}_g^*$.

[1] W. K. Wootters and W. H. Zurek, ''A single quantum cannot be cloned'' Nature (London) **299**, 802–803 (1982).
[2] V. Bužek and M. Hillery, ''Quantum copying: Beyond the no-cloning theorem,'' Phys. Rev. A **54**, 1844–1852 (1996), quant-ph/9607018.
[3] M. Hillery and V. Bužek, ''Quantum copying: Fundamental inequalities,'' Phys. Rev. A **56**, 1212–1216 (1997).
[4] N. Gisin and S. Massar, ''Optimal quantum cloning machines,'' Phys. Rev. Lett. **79**, 2153–2156 (1997), quant-ph/9705046.
[5] D. Bruß, D. P. DiVincenzo, A. Ekert, C. A. Fuchs, C. Machiavello, and J. A. Smolin, ''Optimal universal and state-dependent cloning,'' Phys. Rev. A **57**, 2368–2378 (1998), Los Alamos Report, quant-ph/9705038.
[6] D. Bruß, A. Ekert, and C. Machiavello, ''Optimal universal cloning and state estimation,'' Los Alamos Report, quant-ph/9712019.
[7] V. Bužek and M. Hillery, ''Universal optimal cloning of qubits and quantum registers,'' Los Alamos Report, quant-ph/9801009.
[8] P. Zanardi, ''A note on quantum cloning in *d* dimensions,'' Los Alamos Report, quant-ph/9804011.
[9] N. Gisin and B. Huttner, ''Quantum cloning, eavesdropping and Bell's inequality,'' Phys. Lett. A **228**, 13–21 (1997).
[10] N. Gisin, ''Quantum cloning without signaling,'' Los Alamos Report, quant-ph/9801005.
[11] T. Mor, ''No-cloning of orthogonal states in composite systems,'' Los Alamos Report, quant-ph/9802036.
[12] N. J. Cerf, ''Quantum cloning and the capacity of the Pauli channel,'' Los Alamos Report, quant-ph/9803058.
[13] N. J. Cerf, ''Asymmetric quantum cloning machines,'' Los Alamos Report, quant-ph/9805024.
[14] C. S. Niu and R. B. Griffiths, ''Optimal copying of one quantum bit,'' Los Alamos Report, quant-ph/9805073.
[15] M. Murao, D. Jonathan, M. B. Plenio, and V. Vedral, ''Quantum telecloning and multiparticle entanglement,'' Los Alamos Report, quant-ph/9806082.
[16] R. F. Werner, ''Optimal cloning of pure states,'' Phys. Rev. A **58**, 1827–1832 (1998); Los Alamos Report, quant-ph/9804001.
[17] In Ref. 16 the symbol T was used for T_*. In contrast to Ref. 16 the key arguments in the present paper are phrased more readily in terms of the map on observables than in terms of the map on states. Therefore, we decided to change this notation, which then also agrees better with the usage for completely positive maps on operator algebras.
[18] B. Simon, ''Representations of finite and compact groups,'' Am. Math. Soc. (1996).
[19] W. F. Stinespring, ''Positive functions on C^*-algebras,'' Proc. Am. Math. Soc. **6**, 211–216 (1955).
[20] H. Scutaru, ''Some remarks on covariant completely positive linear maps on C*-algebras,'' Rep. Math. Phys. **16**, 79–87 (1979).
[21] M. Fannes, B. Nachtergaele, and R. F. Werner, ''Exact antiferromagnetic ground states of quantum spin chains,'' Europhys. Lett. **10**, 633–637 (1989).
[22] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (World Scientific, Singapore, 1986).
[23] D. P. Zhelobenko, ''Compact Lie groups and their representations,'' Am. Math. Soc. (1978).
[24] All representations in this paper are finite dimensional.

Tunneling of a massless field through a 3D Gaussian barrier

Giovanni Modanese[a)]
European Centre for Theoretical Studies in Nuclear Physics and Related Areas, Villa Tambosi, Strada delle Tabarelle 286, I-38050 Villazzano (TN), Italy

(Received 29 July 1998; accepted for publication 26 January 1999)

We propose a method for the approximate computation of the Green function of a scalar massless field ϕ subjected to potential barriers of given size and shape in space–time. This technique is applied to the case of a 3D Gaussian ellipsoidlike barrier, placed on the axis between two pointlike sources of the field. Instead of the Green function we compute its temporal integral, that gives the static potential energy of the interaction of the two sources. Such interaction takes place in part by tunneling of the quanta of ϕ across the barrier. We evaluate numerically the correction to the potential in dependence on the barrier size and on the barrier-sources distance. © *1999 American Institute of Physics.* [S0022-2488(99)04906-3]

I. INTRODUCTION

In Quantum Field Theory it is useful in several occasions to have a general expression for the Euclidean two-point correlation function of a massless scalar field ϕ in the presence of potential ''barriers'' in space–time of the form

$$V(\phi(x)) = \xi J_\Omega(x)[\phi^2(x) - \phi_0^2]^2, \tag{1}$$

$J_\Omega(x)$ being the characteristic function of the 4-region Ω where the potential has support ($J_\Omega = 1$ for $x \epsilon \Omega$, $J_\Omega = 0$ elsewhere). The region Ω can be multiple connected, thus representing several barriers placed at different points in space–time.

Possible applications are connected for instance to the fact that a potential of the form (1) represents a localized imaginary mass term ($m^2 < 0$) in the action of the scalar field ϕ. Terms of this kind can be present in cosmological models with inflationary fields. It is also known that every quantum field with nonvanishing vacuum expectation value (VEV) has a global imaginary mass term in its Lagrangian,[1] which couples to the gravitational field as a cosmological term; one can show[2] that if the VEV is not constant but depends on x, it becomes a *local* cosmological term for the gravitational field.

More generally, suppose we have a system of two interacting fields and regard one of them (or its VEV) as a fixed external source. The coupling term of the two fields becomes a local constraint for the dynamical field, a sort of external potential localized in the regions where the external field has support. It is therefore important to study the tunneling of the dynamical field through these regions, that is, its Green functions. Note that in systems like this translational invariance is generally lost.

It is easy to check that the potential $V(\phi)$ in Eq. (1) implements in fact a constraint in the functional integral of the field; writing this integral as

$$z = \int d[\phi] \exp\left[-\int d^4x (\partial\phi)^2 - \int d^4x V(\phi) \right],$$

one sees that for large ξ the square of the field is forced to take the value ϕ_0^2 within the region Ω.

[a)]Electronic mail: modanese@science.unitn.it

For a characteristic function $J_\Omega(x)$ like the one specified above we say that the constraint is imposed in a ''sharp'' way; in space–time the potential barrier looks like a step at the boundary of Ω. Smoothing J_Ω we can obtain a smooth potential barrier. In the following we shall be more interested in this second case.

Note that the potential (1) has the shape of a double *well*, as long as considered only a function of the field ϕ, but regarded as a function of x it is positive and reminds much more a *barrier*.

Let us focus on the case of weak fields, such that ϕ (Ref. 4) can be disregarded with respect to ϕ^2. If the product $\gamma\equiv\xi\phi_0^2$ is small, then the effect of the barriers on field correlations is small, too, and can be treated as a perturbation. One can solve the equation for the modified propagator $G'(x_1,x_2)=\langle\phi(x_1)\phi(x_2)\rangle_V$ in closed form (see the Appendix), finding that G' is given by a double inverse Fourier transform, with the direct transform of J_Ω evaluated at $(p+k)$,

$$G(x_1,x_2)=G^0(x_1,x_2)+\gamma G'(x_1,x_2),$$

$$G^0(x_1,x_2)=\int d^4k e^{-ik(x_1-x_2)},$$

$$G'(x_1,x_2)=\int d^4p\int d^4k e^{ipx_1}e^{ikx_2}\frac{\tilde{J}_\Omega(p+k)}{k^2p^2}. \tag{2}$$

In finite-dimensional quantum mechanics computing $G'(x_1,x_2)$ corresponds to compute the Feynman transition amplitude, related in turn to the system's wave function in the presence of barriers. In field theory the intuitive meaning of $G'(x_1,x_2)$ is less immediate. However, we can derive from $G'(x_1,x_2)$ a quantity with a direct physical interpretation; the static potential $U(\mathbf{x}_1,\mathbf{x}_2)$ of the interaction of two pointlike sources q_1 and q_2 of the field ϕ at rest. This interaction is mediated by the exchange of quanta of ϕ. If the barriers are placed somewhere between the sources, the interaction is clearly affected, but it still takes place—provided the product γ is small—with the quanta of ϕ ''tunneling'' through the barriers (or passing over the wells, depending on the interpretation).

The leading contribution to the static potential $U(\mathbf{x}_1,\mathbf{x}_2)$ is obtained from (2) as follows.[3] First one defines $J_\Omega(x)$ as the product of a 3D function $j_\Omega(\mathbf{x})$ and a function constant in time, then one integrates over t_1 and t_2, multiplies by q_1q_2 and divides by $-T$, taking the limit for $T\to\infty$. The result is

$$\begin{aligned}U(\mathbf{x}_1,\mathbf{x}_2)&=U^0(\mathbf{x}_1,\mathbf{x}_2)+\gamma U'(\mathbf{x}_1,\mathbf{x}_2)\\&=\frac{q_1q_2}{|\mathbf{x}_1-\mathbf{x}_2|}-\gamma(2\pi)^8q_1q_2\int d\mathbf{p}\int d\mathbf{k}e^{i\mathbf{p}\mathbf{x}_1}e^{i\mathbf{k}\mathbf{x}_2}\frac{\tilde{j}_\Omega(\mathbf{p}+\mathbf{k})}{\mathbf{k}^2\mathbf{p}^2}.\end{aligned} \tag{3}$$

This formula is easily generalized to the case of N charges $q_1,\ldots,q_N$, placed, respectively, at $\mathbf{x}_1,\ldots,\mathbf{x}_N$.

A limit case of the physical situation we are considering is represented by the electrostatic potential of pointlike charges in the presence of perfect conductors. In this case the field is exactly zero within the region Ω, and Ω has sharp boundaries—thus $j_\Omega(\mathbf{x})$ is a step function and $\tilde{j}_\Omega(\mathbf{p})$ a strongly oscillating function. Equation (3) could be applied to this case only if the parameters ϕ_0 and ξ could be chosen in such a way that $\phi_0\to0$ and $\xi\to\infty$, the product $\gamma=\xi\phi_0^2$ still being finite and small. We know, however, that usually in an electrostatic system the change in potential energy due to the presence of perfect conductors is not just a small correction. (It can be computed exactly, in principle, solving a classical field equation with suitable boundary conditions.)

The case of interest here is actually more subtle. In the following $j_\Omega(\mathbf{x})$ is supposed to be a smooth function and both ϕ_0 and ξ are taken to be finite. The field square has only a certain probability to be equal to ϕ_0^2 within Ω. This probability is maximum at the center of Ω and decreases towards the boundary of Ω. Since $j_\Omega(\mathbf{x})$ is smooth (a Gaussian function), its Fourier transform $\tilde{j}_\Omega(\mathbf{p})$ is smooth too, and the integral (3) can be computed numerically.

It is interesting to study U' in dependence on the geometrical features of the barrier Ω and on the position of q_1 and q_2 with respect to it. Take, for instance, a finite size barrier (Gaussian ellipsoid, see Sec. II) lying on the axis joining $\mathbf{x}_1$ to $\mathbf{x}_2$. We may expect that if one of the two charges is close to Ω, then $|U'/U^0|$ is larger, decreasing if both charges are far away from Ω—or if Ω is not on their axis. This behavior is confirmed and specified by our numerical results.

The paper is organized as follows. In Sec. II we compute the leading correction to the static potential for a barrier with the shape of an ellipsoid. Due to the peculiar behavior of the integrand, the procedure for numerical integration is not trivial and requires some care. We describe it in detail. Results are given in Sec. III. They concern in particular the dependence of the correction to $U(\mathbf{x}_1,\mathbf{x}_2)$ on the geometrical setting (size of the barrier and its position with respect to the pointlike sources). Far from exploring all the conceivable variations and related phenomenology, the main aim of this work is to show that the general technique can be successfully applied to real cases.

II. THE CASE OF TWO STATIC SOURCES

Let us focus now on a configuration with two static sources and one barrier only. We choose our reference frame in such a way that the sources lie on the z-axis,

$$\mathbf{x}_1=(0,0,L_1);\quad \mathbf{x}_2=(0,0,-L_2).$$

The spatial shape and size of the barrier are defined by the function

$$j_\Omega(\mathbf{x})=\exp\left(-\frac{x^2+y^2}{a^2}-\frac{z^2}{b^2}\right). \tag{4}$$

This means that the region Ω is like an ellipsoid centered at the origin, with simmetry axis along Oz, radius of the order of a and thickness of the order of b. We suppose that $a>b$, thus the ellipsoid is ''squeezed'' on the xy-plane. More precisely, the region Ω itself is not sharply defined, but the surfaces where $j_\Omega(\mathbf{x})$ is constant are ellipsoids. For instance, on the surface defined by

$$\frac{x^2+y^2}{a^2}+\frac{z^2}{b^2}=1,$$

the function $j_\Omega(\mathbf{x})$ is constant and equal to e^{-1}.

The Fourier transform of (4) is

$$\tilde{j}_\Omega(\mathbf{p})=\pi^{3/2}a^2b\,\exp\{\tfrac{1}{4}[-a^2(p_x^2+p_y^2)-b^2p_z^2]\}. \tag{5}$$

The charges q_1 and q_2 can be taken to be unitary and the distances L_1 and L_2 expressed as multiples of the ellipsoid radius a: $L_1\equiv n_1a$, $L_2\equiv n_2a$. From (3), (5) we obtain

$$U'(\mathbf{x}_1,\mathbf{x}_2)=-(2\pi)^8\pi^{3/2}a^2b\int d\mathbf{p}\int d\mathbf{k}\,\frac{e^{ik_zn_1a-ip_zn_2a}}{\mathbf{k}^2\mathbf{p}^2}\exp\left[-\frac{1}{4}a^2(\mathbf{p}+\mathbf{k})_{xy}^2-\frac{1}{4}b^2(p_z+k_z)^2\right],$$

where $\mathbf{V}_{xy}$ denotes the component of a vector $\mathbf{V}$ in the plane xy. In the following we shall be most interested in the case with one charge far from the barrier $(n_1\gg 1)$, while the other charge is close to it (typically in our numerical calculations n_2 ranges between 1 and 15). Accordingly we set $n_1^{-1}=\varepsilon$, $n_2=n$. After rescaling $k_z\to\varepsilon k_z$ we obtain

$$U'(\mathbf{x}_1,\mathbf{x}_2)=-(2\pi)^8\pi^{3/2}\frac{a^3b}{L_1}\int d\mathbf{p}\int d\mathbf{k}\frac{e^{ik_za-ip_zna}}{\mathbf{p}^2(k_x^2+k_y^2+\varepsilon k_z^2)}$$
$$\times\exp\left[-\frac{1}{4}a^2(\mathbf{p}+\mathbf{k})_{xy}^2-\frac{1}{4}b^2(p_z+\varepsilon k_z)^2\right].$$

Then we eliminate any further dimensional parameters by rescaling $k\rightarrow 2k/a$ and $p\rightarrow 2p/a$, obtaining

$$U'(\mathbf{x}_1,\mathbf{x}_2)=-(2\pi)^8\pi^{3/2}\frac{4ab}{L_1}\int d\mathbf{p}\int d\mathbf{k}\frac{e^{ik_z-ip_zn}}{\mathbf{p}^2(k_x^2+k_y^2+\varepsilon k_z^2)}\exp[-(\mathbf{p}+\mathbf{k})_{xy}^2-\rho^2(p_z+\varepsilon k_z)^2],$$

where $\rho=b/a$ is the ratio between the thickness b and the radius a of the ellipsoid.

Next we introduce the polar variables θ_k, θ_p, θ_k, and ϕ_p. In the following k and p will not denote four-vectors anymore, but $|\mathbf{k}|$ and $|\mathbf{p}|$, respectively. The square of the component of the vector $(\mathbf{p}+\mathbf{k})$ in the xy plane is

$$(\mathbf{p}+\mathbf{k})_{xy}^2=p^2\sin^2\theta_p+k^2\sin^2\theta_k+2pk\sin\theta_p\sin\theta_k\cos(\phi_k-\phi_p).$$

The other components are

$$k_z=k\cos\theta_k;\quad p_z=p\cos\theta_p;$$

$$k_x^2+k_y^2=\mathbf{k}_{xy}^2=k^2\sin^2\theta_k.$$

Finally, introducing the variables

$$s=\cos\theta_k,\quad t=\cos\theta_p,\quad \phi=(\phi_k-\phi_p),$$

one obtains, remembering that the integrand is even in s,t, the following basic formula:

$$U'(\mathbf{x}_1,\mathbf{x}_2)=-\frac{(2\pi)^{10}}{\sqrt{\pi}}\frac{ab}{L_1}2\pi\int_0^{2\pi}d\phi\int_{-1}^1 ds\int_{-1}^1 dt\int_0^\infty dk\int_0^\infty dp\frac{\cos(ks-npt)}{1-s^2(1-\varepsilon^2)}$$
$$\times\exp[-\rho^2(pt+\varepsilon ks)^2-p^2(1-t^2)-k^2(1-s^2)-2pk\cos\phi\sqrt{(1-t^2)(1-s^2)}]$$
$$\equiv-\frac{(2\pi)^{11}}{\sqrt{\pi}}\frac{ab}{L_1}\int_0^{2\pi}d\phi\int_{-1}^1 ds\int_{-1}^1 dt\int_0^\infty dk\int_0^\infty dp f(\phi,s,t,k,p;\varepsilon,\rho,n)$$
$$\equiv-\frac{(2\pi)^{11}}{\sqrt{\pi}}\frac{ab}{L_1}F(\varepsilon,\rho,n). \tag{6}$$

A. Preliminary study of the integrand

It is important to discuss in advance the case in which ρ and ε take values much smaller than 1, that is, Ω is very thin and the distance of the first charge from Ω is much larger than a. When t and s approach $+1$ or -1, for small values of ρ the integral over k and p converges very slowly at infinity and the factor $\cos(ks-npt)$ performs a large number of oscillations. For very small ε there are many more oscillations in k than in p. (In the limit $\rho\rightarrow 0$ the integral makes sense only as a distribution. We shall never approach this limit, however.)

Let us set, for instance, $s=1$, $t=1$ and $\phi=\pi/2$ in the argument of the exponential in (6). We obtain the exponential factors

$$\exp[-\rho^2(pt+\varepsilon ks)^2]=\exp[-(\rho p)^2]\exp[-(\rho\varepsilon k)^2]\exp[-2\rho^2\varepsilon kp]. \tag{7}$$

TABLE I. Integration ranges in the four domains, for some values of ε,ρ.

ϵ	ρ	$\mathbf{K}_1$	$\mathbf{P}_1$	$\mathbf{K}_2$	$\mathbf{P}_2$	$\mathbf{K}_3$	$\mathbf{P}_3$	$\mathbf{K}_4$	$\mathbf{P}_4$
0.1	0.3	12	10	80	10	12	10	80	10
0.1	0.1	10	10	100	25	20	20	120	40
0.1	0.032	10	10	600	20	20	60	600	60
0.1	0.01	20	20	1500	30	20	200	1500	200
0.032	0.032	10	10	2000	15	20	70	1800	80
0.032	0.01	12	12	6000	15	15	180	6000	250
0.01	0.032	10	10	7500	15	10	70	7500	90
0.032	0.0032	15	15	18000	15	15	650	18000	700

The first factor on the r.h.s. of (7) has a range in p of the order of ρ^{-1} and the second factor has a range in k of the order of $(\rho\varepsilon)^{-1}$. The third factor has a range in p, for fixed k, of the order of $(\rho\sqrt{\varepsilon k})^{-1}$ and a range in k, for fixed p, of the order of $(\rho\sqrt{\varepsilon p})^{-1}$. Fortunately this latter factor is not relevant; if its range is larger than the other two ranges then it does not play any role; if it is

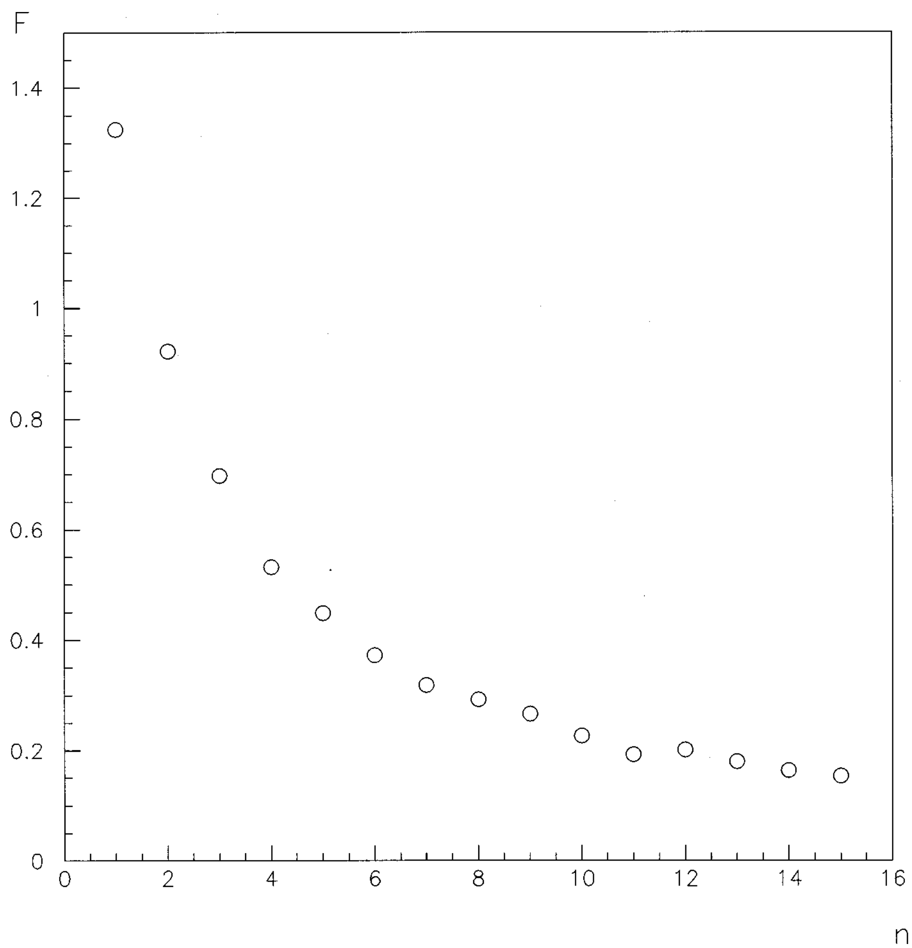

FIG. 1. Dependence of $F(\varepsilon,\rho,n)$ on n, in the range $n=1-15$, for $\varepsilon=0.1$ and $\rho=0.3$. Errors are ~ 0.01.

smaller then it is sufficient to refer to the other ranges.

As soon as s^2 and t^2 go away from 1, the number of oscillations of the integrand decreases. For instance, setting $s=t=0.98$ we obtain the exponential factors

$$\sim\exp[-(\rho p)^2-(\rho\varepsilon k)^2-2\rho^2\varepsilon kp]\exp[-0.04p^2-0.04k^2].$$

When ρ is much smaller than 1 the range of this product is determined by the second exponential and does not depend on ρ.

It is also easy to take into account the term proportional to cos ϕ. After setting $\phi=\pi$ that term gives a positive contribution to the argument of the exponential; thus studying the range of the resulting expression we obtain an upper limit valid for any ϕ.

B. Integration domains

Independently of the considerations above, it is possible to plot the integrand $f(\phi,s,t,k,p;\varepsilon,\rho,n)$ for several different values of ρ and ε and check the ranges of the exponentials. In order to better control the oscillations of f, we study it in 4 different domains of the variables s,t,

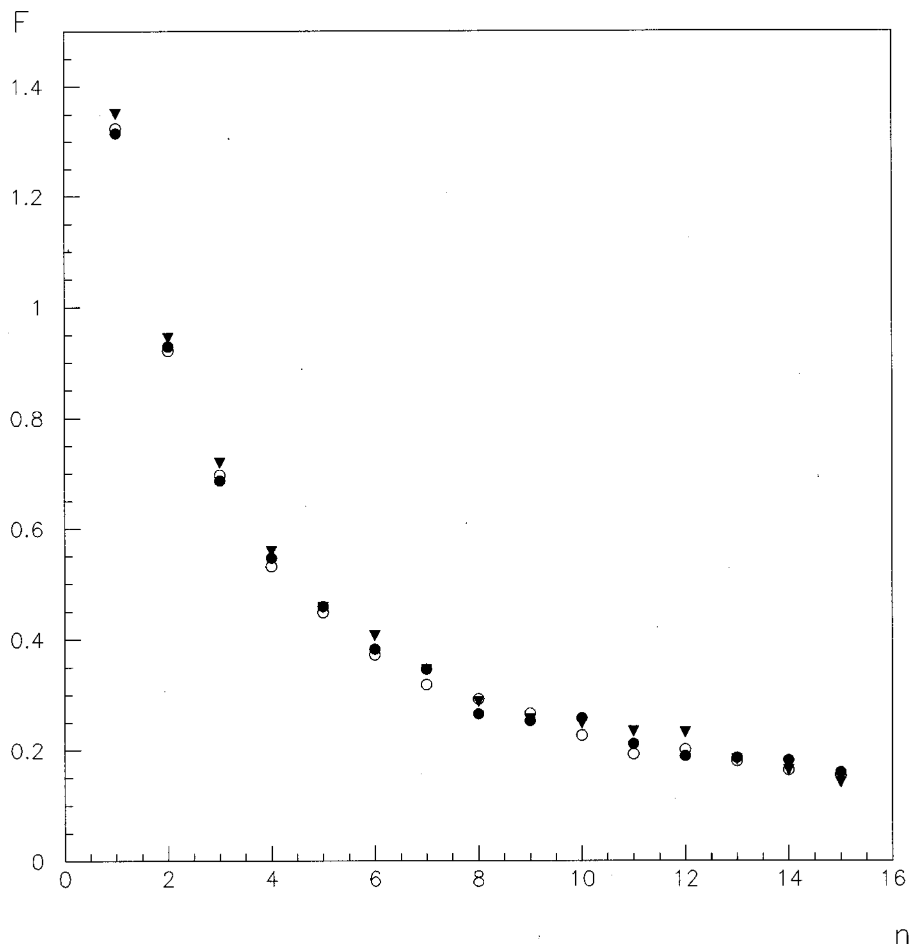

FIG. 2. Comparison of the values of $F(\varepsilon,\rho,n)$ for $\varepsilon=0.1$ and $\rho=0.3$ (white circles), $\rho=0.1$ (black triangles), and $\rho=0.032$ (black circles).

$$\text{Domain 1: } t,s\epsilon[0,1-\alpha];$$
$$\text{Domain 2: } t\epsilon[0,1-\alpha];s\epsilon[1-\alpha,1];$$
$$\text{Domain 3: } t\epsilon[1-\alpha,1];s\epsilon[0,1-\alpha];$$
$$\text{Domain 4: } t,s\epsilon[1-\alpha,1-\alpha].$$

A typical value of α employed in the program is $\alpha=0.02$. The total integration domain in s,t is obtained by ''reflecting'' each of the domains above with respect to one axis and then reflecting again the result with respect to the origin $(s\rightarrow -s,\ t\rightarrow -t,\ s,t\rightarrow -s,\ -t)$. In each domain i there is a maximum value for the variables k and p, beyond which f is equal to zero for any practical purpose. Denoting by K_i and P_i these ranges, Table I shows the results found for some considered values of ε and ρ.

Since the integration over k and p is extended to wide ranges, the most reasonable technique for the numerical computation of the integral (6) appears to be a Monte Carlo sampling of the integrand. The sampling algorithm evaluates the average value of f in each domain, extending the values of k and p up to the maximum range necessary for that domain. At the end the global average is computed, weighing each single average with the ratio between the domain volume and the total volume. Denoting by f_i the average of f in the domain i and by V_i the domain volume we have

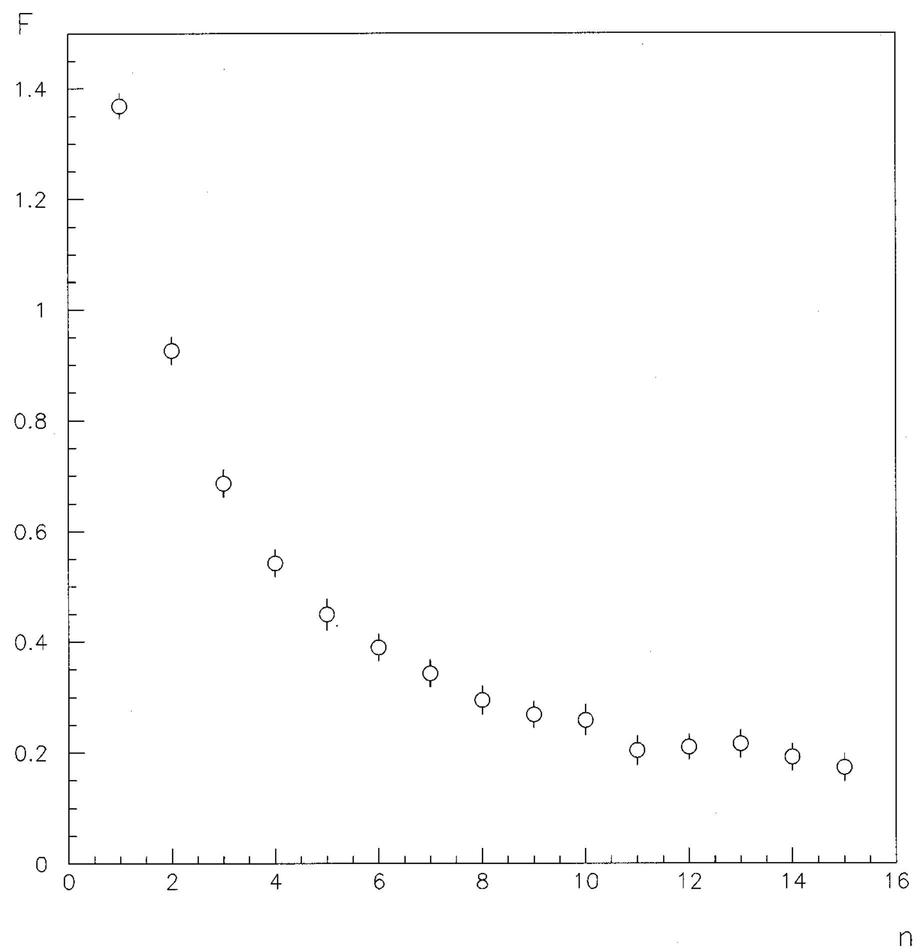

FIG. 3. Same as in Fig. 1, for $\varepsilon=0.05$. Errors are larger, as shown.

TABLE II. Results of the best fit $F(n)=\exp(-mn+q)+b$.

	b	m	q
$\varepsilon=0.1$ (Fig. 1)	0.14	0.29	0.3
with $n>2$	0.12	0.24	0.1
$\varepsilon=0.05$ (Fig. 3)	0.17	0.32	0.4
with $n>2$	0.16	0.26	0.1

$$F=\sum_{i=1}^{4} f_i V_i = 8\pi[(1-\alpha)^2 K_1 P_1 f_1 + \alpha(1-\alpha) K_2 P_2 f_2 + \alpha(1-\alpha) K_3 P_3 f_3 + \alpha^2 K_4 P_4 f_4]. \quad (8)$$

III. RESULTS OF THE NUMERICAL INTEGRATION

The contributions of the domains 2, 3, and 4 to the integral F (compare (8)) are found to be small with respect to the contribution of domain 1. The fluctuations of the average of f in domains 2 and 4 (where s^2 approaches 1) may be very large. In order to achieve a sufficient precision these regions have been sampled with a large number of points (up to $\sim 10^{10}$). The standard routine ''ran2'' (Ref. 4) was used for random numbers generation.

The dependence of the integral F on the parameters ε and ρ is very weak, thus U' depends on a, b and L_1 mainly as ab/L_1 (see Eq. (6)). The study of the dependence of U' on n is more difficult, because this dependence is entirely contained in the integral F and can be evaluated only numerically. One needs to insert in the program a cycle which samples the integrand for different values of n, typically between 1 and 15. This is possible because the ranges P_i, K_i do not depend on n.

The numerical evaluation of F as a function of n in the range $n=1-15$, with $\varepsilon=0.1$ and $\rho=0.3$, gives the results shown in Fig. 1. With $\rho=0.1$ and $\rho=0.032$ one obtains very similar results, thus confirming the weak dependence on ρ (Fig. 2). As expected varying ε does not affect much the value of F either, since the dependence on the distance L_1 is already factorized out of the integral (compare Fig. 3).

Figures 1, 2, 3 reveal an exponential behavior of $F(n)$ of the form

$$F(n)\sim\exp(-mn+q)+b.$$

It is also clear just from the graphs that the exponential decrease of F for large n leaves an asymptotic value $F=b$, with b in the interval 0.1–0.2. This is an interesting behavior, as it means that the ''shadow'' produced by the barrier in the static field of the two sources has a long, constant tail.

A least-squares fit of the data gives the results of Table II. Excluding from the fit the first two points ($n=1,2$) we obtain better estimates for the distribution tail and for b. The errors on the parameters of the fit, in particular those on b, are small. They can be estimated knowing that the least-squares sum of the perceptual errors $S=\Sigma_n\{1-[\exp(-mn+q)+b]/F(n)\}^2$ has a minimum value $S_{\min}\sim 0.05$ and that its second partial derivatives at the minimum are of the order of $(\partial^2 S/\partial b^2)\sim 4\cdot 10^2$, $(\partial^2 S/\partial m^2)\sim 10^2$, $(\partial^2 S/\partial q^2)\sim 10$.

IV. CONCLUSIONS

Our technique for the computation of the Green function and the static potential of two pointlike sources appears to work well for weak fields, yielding reasonable results. The method is based upon a double 3D Fourier transform of the function which represents size and position in space of the potential well or barrier. This double transform is necessary, due to the lack of translational invariance of the system. Its numerical evaluation requires a preliminary analytical study and a subdivision of the integration volume in a few domains, because the range of the real exponential factors in the integrand varies considerably.

We studied the case of a smooth barrier with the form of a Gaussian ellipsoid in coordinate and momentum space. For values of ρ and ε not much smaller than 1 a good precision was obtained. (ρ is the ratio between the lengths a and b of the ellipsoids axes and ε is the ratio between the length of the major axis a and the distance L_1 of the first source from the ellipsoid.)

Denoting by n the distance of the second source in units of the major axis, we found that the correction to the interaction potential along the line joining the two sources and the barrier has the following form (compare Eqs. (3), (6)):

$$U=U^0+\gamma U'=\frac{1}{|\mathbf{x}_1-\mathbf{x}_2|}-\frac{(2\pi)^{11}}{\sqrt{\pi}}\frac{\gamma ab}{L_1}F(\varepsilon,\rho,n)$$
$$=\frac{1}{|\mathbf{x}_1-\mathbf{x}_2|}\left[1-\frac{(2\pi)^{11}}{\sqrt{\pi}}\gamma ab(1+n\varepsilon)F(\varepsilon,\rho,n)\right].$$

The function F depends very weakly on ρ and ε. Its dependence on n is displayed in Figs. 1–3 and shows an exponential decay followed by a constant tail.

The behavior summarized above is interesting in itself, being the result of a sort of ''tunneling'' of the scalar field through a region where it is constrained or has imaginary mass. We have seen that the local imaginary mass term affects the propagation of the field also outside the region Ω where it has support. This feature is easily understood from the physical point of view; we gave here a method for its quantitative evaluation.

ACKNOWLEDGMENT

This work has been partially supported by the A.S.P., Associazione per lo Sviluppo Scientifico e Tecnologico del Piemonte, Turin, Italy.

APPENDIX: PROOF OF THE EXPRESSIONS FOR G', U'

We give here the proof of Eqs. (2) and (3) of the main text. Expanding the square in (1) we obtain for $W[J]$,

$$W[J]=\int d[\phi]\exp\left\{-\int d^4x[(\partial\phi)^2-2\xi\phi_0^2J_\Omega(x)\phi^2(x)+\xi J_\Omega(x)\phi^4(x)+\xi J_\Omega(x)\phi_0^4]\right\}.$$

The last term in the square bracket is constant with respect to $\phi(x)$ and its exponential can be factorized out of the functional integral. In a first instance—for weak fields—we can disregard the $\phi^4(x)$ term. We are then led to consider a quadratic functional integral, and the ''modified propagator'' $G(x,y)=\langle\phi(x)\phi(y)\rangle_J$, which by definition satisfies the equation

$$[\partial_x^2+\gamma J_\Omega(x)]G(x,y)=-(2\pi)^4\delta^4(x-y), \tag{9}$$

where $\gamma=2\xi\phi_0^2>0$. Let us focus on the case when $\phi_0^2=0$ inside the regions Ω_i and let us take the limit $\phi_0\to 0$ and $\xi\to\infty$ in such a way that γ is finite and very small, so that the term $\gamma J_\Omega(x)$ in Eq. (9) constitutes only a small perturbation, compared to the kinetic term. Then we can set

$$G(x,y)=G^0(x,y)+\gamma G'(x,y),$$

where $G^0(x,y)$ is the propagator of the free scalar field, and we find immediately that $G'(x,y)$ satisfies the equation

$$\partial_x^2G'(x,y)=-J_\Omega(x)G^0(x,y). \tag{10}$$

Unlike $G^0(x,y)$, in general $G'(x,y)$ will not depend only on $(x-y)$, because the source breaks the translation invariance of the system. In order to go to momentum space it will therefore be necessary to consider the Fourier transform of $G'(x,y)$ with respect to both arguments. We define $\tilde{G}'(p,k)$ and $\tilde{J}_\Omega(p)$ as follows:

$$G'(x,y)=\int d^4p\int d^4k e^{ipx}e^{iky}\tilde{G}'(p,k)$$

and

$$J_\Omega(x)=\int d^4p e^{ipx}\tilde{J}_\Omega(p),\quad G^0(x,y)=\int d^4k\frac{e^{-ik(x-y)}}{k^2}.$$

The right-hand side of (10) can be rewritten as

$$J_\Omega(x)G^0(x,y)=\int d^4p\int d^4k e^{ipx}\tilde{J}_\Omega(p)\frac{e^{-ik(x-y)}}{k^2}=\int d^4k\int d^4p e^{iky}e^{ipx}\frac{\tilde{J}_\Omega(p+k)}{k^2},$$

and we obtain the following algebraic equation for the double Fourier transform of the first order correction to the propagator:

$$p^2\tilde{G}'(p,k)=\frac{\tilde{J}_\Omega(p+k)}{k^2}.$$

Transforming back, in conclusion we find Eq. (2) of the main text, namely,

$$G'(x,y)=\int d^4p\int d^4k e^{ipx}e^{iky}\frac{\tilde{J}_\Omega(p+k)}{k^2p^2}. \tag{11}$$

Therefore, if we know the Fourier transform of the characteristic function J_Ω of the space–time region where the constraint is imposed, we can in principle compute the leading order correction to the field propagator and thus to $W[J]$.

It is known[5] that the vacuum-to-vacuum amplitude $W[J]=\langle 0^+|0^-\rangle_J$ of a field system in the presence of an external source J is related to the logarithm of the systems' ground state energy,

$$E_0[J]=-T^{-1}\ln W[J],$$

where the functional integral is supposed to be suitably normalized and the source vanishes outside the temporal interval $[-T/2,+T/2]$, with T eventually approaching infinity. (We use units in which $\hbar=c=1$.)

An interesting application of (11) occurs in the case when the field $\phi(x)$ also interacts with N static pointlike sources placed at $\mathbf{x}_1,\mathbf{x}_2\ldots\mathbf{x}_N$. Namely, let us add a further, linear coupling term S_Q to the action of the system,

$$S_Q=\int d^4xQ(x)\phi(x),\ \ \text{with}\ \ Q(x)=\sum_{j=1}^{N}q_j\delta^3(\mathbf{x}-\mathbf{x}_j).$$

The ground state energy of the system corresponds, up to a constant, to the static potential energy of the interaction of the sources through the field ϕ. As before, it is obtained from the functional average of the interaction term, computed keeping the constraint into account,

$$E_0[J,Q]=U(\mathbf{x}_1,\ldots,\mathbf{x}_N)=-T^{-1}\ln\langle\exp\{-S_Q\}\rangle_J. \tag{12}$$

Expanding (12) one finds that to leading order in the $q_j s$, $U(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ is given by a sum of propagators integrated on time,

$$U(\mathbf{x}_1,\ldots,\mathbf{x}_N) = -T^{-1}\sum_{j,l=1}^{N} q_j q_l \int dt_j \int dt_l \langle \phi(t_j,\mathbf{x}_j)\phi(t_l,\mathbf{x}_l)\rangle_J ,$$

where $t_j, t_l \epsilon [-T/2,+T/2]$. Since the regions Ω_i are infinitely elongated in the temporal direction the function $\tilde{J}_\Omega(p+k)$ gets factorized as

$$\tilde{J}_\Omega(p+k) = (2\pi)^4 \delta(p_0+k_0)\tilde{j}_\Omega(\mathbf{p}+\mathbf{k}). \tag{13}$$

Clearly the potential is disturbed by the presence of the ''barriers'' $j_\Omega(\mathbf{x})$. To first order in γ we can write

$$U(\mathbf{x}_1,\ldots,\mathbf{x}_N) = U^0(\mathbf{x}_1,\ldots,\mathbf{x}_N) + \gamma U'(\mathbf{x}_1,\ldots,\mathbf{x}_N)$$

and taking into account Eqs. (11), (13) we find

$$\begin{aligned} U'(\mathbf{x}_1,\ldots,\mathbf{x}_N) &= -T^{-1}\sum_{j,l=1}^{N} q_j q_l \int dt_j \int dt_l G'(x_j,x_l) \\ &= -(2\pi)^4 T^{-1}\sum_{j,l=1}^{N} q_j q_l \int dt_j \int dt_l \int d^4p \int d^4k \, \frac{e^{ipx_j+ikx_l}\tilde{J}_\Omega(p+k)}{k^2p^2} \\ &= -(2\pi)^8 T^{-1}\sum_{j,l=1}^{N} q_j q_l \int dt_j \int dt_l \int d^4p \int d\mathbf{k} \, \frac{e^{ip_0(t_j-t_l)+i\mathbf{p}\mathbf{x}_j+i\mathbf{k}\mathbf{x}_l}\tilde{j}_\Omega(\mathbf{p}+\mathbf{k})}{(p_0^2+\mathbf{k}^2)(p_0^2+\mathbf{p}^2)}. \end{aligned}$$

Changing to variables $t=t_j-t_l$ and $s=t_j+t_l$ and integrating we finally obtain the contribution of the perturbation to the static potential energy [Eq. (3) of the main text],

$$U'(\mathbf{x}_1,\ldots,\mathbf{x}_N) = -(2\pi)^8 \sum_{j,l=1}^{N} q_j q_l \int d\mathbf{p} \int d\mathbf{k} e^{i\mathbf{p}\mathbf{x}_j+i\mathbf{k}\mathbf{x}_l} \frac{\tilde{j}_\Omega(\mathbf{p}+\mathbf{k})}{\mathbf{k}^2\mathbf{p}^2}.$$

[1] S. Weinberg, Rev. Mod. Phys. **61**, 1 (1989).
[2] G. Modanese, Phys. Rev. D **54**, 5002 (1996).
[3] M. Bander, Phys. Rep. **75**, 205 (1981).
[4] B. P. Flannery, W. H. Press, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in* FORTRAN 77, 2nd ed. (Cambridge University Press, Cambridge, 1992), p. 272.
[5] K. Symanzik, Commun. Math. Phys. **16**, 48 (1970); see also G. Modanese, Nucl. Phys. B **434**, 697 (1995).

Generalized adiabatic product expansion: A nonperturbative method of solving the time-dependent Schrödinger equation

Ali Mostafazadeh[a)]
Department of Mathematics, Koç University, Istinye 80860 Istanbul, Turkey

(Received 6 January 1999; accepted for publication 17 March 1999)

We outline a method based on successive canonical transformations which yields a product expansion for the evolution operator of a general (possibly non-Hermitian) Hamiltonian. For a class of such Hamiltonians this expansion involves a finite number of terms, and our method gives the exact solution of the corresponding time-dependent Schrödinger equation. We apply this method to study the dynamics of a general nondegenerate two-level quantum system, a time-dependent classical harmonic oscillator, and a degenerate system consisting of a spin 1 particle interacting with a time-dependent electric field $\vec{\mathcal{E}}(t)$ through the Stark Hamiltonian $H=\lambda(\vec{J}\cdot\vec{\mathcal{E}})^2$. © *1999 American Institute of Physics.* [S0022-2488(99)02207-0]

I. INTRODUCTION

Recently, a method based on successive canonical transformations has been used to obtain exact solution of the Schrödinger equation

$$i\frac{d}{dt}|\psi(t)\rangle=H(t)|\psi(t)\rangle \tag{1}$$

for a class of dipole Hamiltonians[1–3] and time-dependent harmonic oscillators.[4] For these systems the Hamiltonian is a nondegenerate Hermitian operator. The purpose of the present article is to extend the application of this method to the cases where the Hamiltonian is non-Hermitian and involves degenerate eigenvalues.

Non-Hermitian Hamiltonians have been used to model a variety of physical systems involving decaying states.[5] The solution of the Schrödinger equation for a time-dependent two-level non-Hermitian Hamiltonian has been considered in Refs. 6 and 7. Another motivation for the study of the Schrödinger equation for a time-dependent non-Hermitian Hamiltonian is the fact that the solution of every linear ordinary differential equation (ODE) may be reduced to the solution of a system of first-order linear ODEs which can be written in the form of the time-dependent Schrödinger equation (1) or alternatively

$$|\psi(t)\rangle=U(t)|\psi(0)\rangle, \tag{2}$$

$$i\frac{d}{dt}U(t)=H(t)U(t), \tag{3}$$

$$U(t)=1, \tag{4}$$

where $U(t)$ is the evolution operator. For a general linear ODE the corresponding Hamiltonian $H(t)$ may be a non-Hermitian matrix with degenerate eigenvalues.

[a)]Electronic mail: amostafazadeh@ku.edu.tr

The method of *adiabatic product expansion* developed in Refs. 1 and 2 does not directly apply to quantum systems with non-Hermitian Hamiltonians. In this article we shall present a generalization of this method which applies to arbitrary (possibly) non-Hermitian Hamiltonians with degenerate as well as nondegenerate eigenvalues.

The organization of the article is as follows. In Sec. II we review the basic results concerning the adiabatic approximation for degenerate and non-Hermitian Hamiltonians. In Sec. III we discuss the generalization of the method of adiabatic product expansion to these Hamiltonians. In Sec. IV we use the results of Sec. III to study the solution of the Schrödinger equation for a general nondegenerate non-Hermitian two-level Hamiltonian. In Sec. V, we apply the general results obtained in Sec. IV to treat the classical equation of motion for a harmonic oscillator with a time-dependent frequency. In Sec. VI, we discuss the application of the adiabatic product expansion to study the quadrupole interaction of a spin 1 particle with a time-dependent electric field $\vec{\mathcal{E}}=(\mathcal{E}_1(t),\mathcal{E}_2(t),0)$. We show that the corresponding Hamiltonian which has a degenerate and a nondegenerate eigenvalue is canonically equivalent to a Hamiltonian which has only nondegenerate eigenvalues. Furthermore, we show that if the direction of the electric field depends in a particular way on its magnitude, then our method yields the exact solution of the Schrödinger equation. Finally we present our conclusions in Sec. VII.

II. ADIABATIC APPROXIMATION FOR NON-HERMITIAN HAMILTONIANS

Let $H=H[R]$ be a parametric Hamiltonian which depends on a set of real parameters $R=(R^1,R^2,\ldots,R^d)$ labeling the points of a smooth manifold M. Let $E_n[R]$ denote the eigenvalues of $H[R]$ and $\mathcal{H}_n[R]$ be the degeneracy subspace associated with $E_n[R]$. Let $\mathcal{N}$ denote the degree of degeneracy of $E_n[R]$, i.e., the complex dimension of $\mathcal{H}_n[R]$. We shall assume that the spectrum of $H[R]$ is discrete and $\mathcal{N}$ does not depend on R.

Now let $|\psi_n,a;R\rangle$ and $|\phi_n,a;R\rangle$ form a complete biorthonormal basis of the Hilbert space.[8,9] This means that $|\psi_n,a;R\rangle$ with $a\in\{1,2,...,\mathcal{N}\}$ form a basis of $\mathcal{H}_n[R]$, in particular

$$H[R]|\psi_n,a;R\rangle=E_n[R]|\psi_n,a;R\rangle, \tag{5}$$

and $|\phi_n,a;R\rangle$ satisfy

$$H[R]^\dagger|\phi_n,a;R\rangle=E_n^*[R]|\phi_n,a;R\rangle, \tag{6}$$

$$\langle\phi_m,b;R|\psi_n,a;R\rangle=\delta_{mn}\delta_{ab}, \tag{7}$$

$$\sum_n\sum_{a=1}^{\mathcal{N}}|\psi_n,a;R\rangle\langle\phi_n,a;R|=1. \tag{8}$$

Next suppose that the parameters R^i depend on time t, then $R(t)$ defines a curve $\mathcal{C}$ in the parameter space M, and the Hamiltonian, its eigenvalues, and eigenvectors become time dependent. In this case we use the notation $H(t):=H[R(t)]$, $E_n(t):=E_n[R(t)]$, $|\psi_n,a;t\rangle:=|\psi_n,a;R(t)\rangle$, and $|\phi_n,a;t\rangle:=|\phi_n,a;R(t)\rangle$. We shall assume that $E_n(t)$, $|\psi_n,a;t\rangle$ and $|\phi_n,a;t\rangle$ are smooth functions of t and that during the evolution of the system the eigenvalues of the Hamiltonian do not cross, i.e., if $E_m(0)<E_n(0)$, then for all $t\in[0,\tau]$, $E_m(t)<E_n(t)$, where τ denotes the duration of the evolution of the system.

Differentiating both sides of Eq. (5) with respect to t, taking the inner product of both sides of the resulting equation with $|\phi_m,b;t\rangle$, for arbitrary m and b, and using Eqs. (5)–(7), we have

$$[E_m(t)-E_n(t)]\langle\phi_m,b;t|\frac{d}{dt}|\psi_n,a;t\rangle+\langle\phi_m,b;t|\dot{H}(t)|\psi_n,a;t\rangle-\delta_{mn}\delta_{ab}\dot{E}(t)=0. \tag{9}$$

Here a dot denotes differentiation with respect to t. For $m\neq n$, Eq. (9) reads

$$\langle \phi_m, b; t | \frac{d}{dt} | \psi_n, a; t \rangle = \frac{\langle \phi_m, b; t | \dot{H}(t) | \psi_n, a; t \rangle}{E_n(t) - E_m(t)} \quad \text{for } m \neq n. \tag{10}$$

Now let us express the solution of the Schrödinger equation (1) in the basis $\{|\psi_n, a; t\rangle\}$. Then

$$|\psi(t)\rangle = \sum_n \sum_{a=1}^{\mathcal{N}} C_a^n(t) |\psi_n, a; t\rangle, \tag{11}$$

where $C_a^n(t)$ are complex coefficients. Substituting Eq. (11) in the Schrödinger equation (1), taking the inner product of both sides of the resulting equation with $|\phi_m, b; t\rangle$, and making use of Eqs. (5), (6), (7), and (10), we find

$$i\dot{C}_b^m - E_m C_b^m + \sum_{a=1}^{\mathcal{N}} i \left\langle \phi_m, b; t \left| \frac{d}{dt} \psi_m, a; t \right. \right\rangle C_a^m = -i \sum_{n \neq m} \sum_{a=1}^{\mathcal{N}} \frac{\langle \phi_m, b; t | \dot{H}(t) | \psi_n, a; t \rangle}{E_n(t) - E_m(t)}. \tag{12}$$

The special case of this equation with $\mathcal{N} = 1$, i.e., the nondegenerate case, has been originally derived by Garrison and Wright[9] in their investigation of the adiabatic geometric phase[10] for non-Hermitian Hamiltonians.[9,11–14]

If the right-hand side of Eq. (12) is negligible, then one says that the system undergoes an adiabatic evolution.[15,16,2,9,17] In this case, the equations for C_a^n decouple and their solution is given by

$$C_a^n(t) = \sum_{b=1}^{\mathcal{N}} K_{ab}^n(t) C_b^n(0), \tag{13}$$

where $K_{ab}^n(t)$ are entries of the invertible matrix

$$K^n(t) := e^{-i\int_0^t E_n(s)ds} \mathcal{P} \exp\left[i \int_{R(0)}^{R(t)} \mathcal{A}^n[R] \right], \tag{14}$$

$\mathcal{P}$ denotes the path-ordering operator, $\mathcal{A}^n$ is the matrix of one-forms with entries

$$\mathcal{A}_{ab}^n[R] := i \langle \phi_n, a; R | d | \psi_n, b; R \rangle, \tag{15}$$

d stands for the exterior derivative with respect to R^i, and the line integral in Eq. (14) is evaluated along the curve $\mathcal{C}$ defined by $R(t)$. If $\mathcal{C}$ is a closed curve in M, the Hamiltonian has a periodic time dependence and the path-ordered exponential in Eq. (14), which takes the form

$$\mathcal{P} \exp\left[i \oint_{\mathcal{C}} \mathcal{A}^n[R] \right], \tag{16}$$

is the *non-Hermitian* analog of the *non-Abelian adiabatic geometric phase*.[18]

Note that if the initial vector $|\psi(0)\rangle$ is an eigenvector of the initial Hamiltonian $H(0)$, then the adiabaticity of the evolution implies that $|\psi(t)\rangle$ is an eigenvector of $H(t)$ for all $t \in [0, \tau]$. In terms of the time-evolution operator $U(t)$ of Eq. (3) this is expressed by

$$U(t) \approx U^{(0)}(t), \tag{17}$$

where

$$U^{(0)}(t) := \sum_n \sum_{a,b=1}^{\mathcal{N}} K_{ab}^n(t) |\psi_n, a; t\rangle \langle \phi_n, b; 0|. \tag{18}$$

One can easily show that $U^{(0)}(t)$ is invertible, and its inverse is given by

$$U^{(0)^{-1}}(t)=\sum_n\sum_{a,b=1}^{\mathcal{N}}K_{ab}^{n^{-1}}(t)|\psi_n,a;0\rangle\langle\phi_n,b;t|,\tag{19}$$

where $K^{n^{-1}}(t)$ is the inverse of $K^n(t)$.

III. ADIABATIC CANONICAL TRANSFORMATIONS AND THE GENERALIZED ADIABATIC PRODUCT EXPANSION

Let $g(t)$ be an invertible linear operator acting on the Hilbert space. Then the transformations:

$$|\psi(t)\rangle\rightarrow|\psi'(t)\rangle:=g(t)|\psi(t)\rangle,\tag{20}$$

$$H(t)\rightarrow H'(t):=g(t)H(t)g(t)^{-1}-ig(t)\frac{d}{dt}g(t)^{-1},\tag{21}$$

$$U(t)\rightarrow U'(t):=g(t)U(t)g(0)^{-1},\tag{22}$$

leave the form of the Schrödinger equation invariant. We shall call such a transformation a *canonical transformation*.

Now let us investigate the consequences of the canonical transformation defined by $g(t)=U^{(0)}(t)^{-1}$. We shall call this transformation the *adiabatic canonical transformation*. Denoting the transformed Hamiltonian H' by $H^{(1)}$, we have

$$H^{(1)}(t)=\sum_{n,m\neq n}\sum_{a=1}^{\mathcal{N}}\sum_{b=1}^{\mathcal{M}}H_{ab}^{(1)^{nm}}(t)|\psi_n,a;0\rangle\langle\phi_m,b;0|,\tag{23}$$

where

$$H_{ab}^{(1)nm}(t):=-K_{ac}^{(n)}(t)^{-1}A_{cd}^{nm}(t)K_{db}^{m}(t),\quad A_{cd}^{nm}(t):=i\langle\phi_nc;t|\frac{d}{dt}|\psi_m,d;t\rangle.\tag{24}$$

Because $g(0)=U^{(0)}(0)^{-1}=1$, the transformed evolution operator is given by

$$U'(t)=U^{(0)}(t)^{-1}U(t).\tag{25}$$

Clearly if the adiabatic approximation is valid, $H^{(1)}(t)\approx 0$ and $U'(t)\approx 1$.

Let us suppose that $H^{(1)}(t)$ has a discrete spectrum and denote by $E_{n_1}^{(1)}(t)$ and $\mathcal{N}_1$ the eigenvalues of $H^{(1)}(t)$ and their degree of degeneracy. Furthermore, let $\{|\psi_{n_1}^{(1)},a_1;t\rangle,|\phi_n^{(1)},a_1;t\rangle\}$ be a biorthonormal eigenbasis of the Hilbert space, i.e.,

$$H^{(1)}(t)|\psi_{n_1}^{(1)},a_1;t\rangle=E_{n_1}^{(1)}(t)|\psi_{n_1}^{(1)},a_1;t\rangle,$$

$$H^{(1)}(t)^{\dagger}|\phi_{n_1}^{(1)},a_1;t\rangle=E_{n_1}^{(1)*}(t)|\phi_{n_1}^{(1)},a_1;t\rangle,$$

$$\langle\phi_{m_1}^{(1)},b_1;t|\psi_{n_1}^{(1)},a_1;t\rangle=\delta_{m_1n_1}\delta_{a_1b_1},\quad\sum_{n_1}\sum_{a_1=1}^{\mathcal{N}_1}|\psi_{n_1}^{(1)},a_1;t\rangle\langle\phi_{n_1}^{(1)},a_1;t|=1.$$

Then $H^{(1)}(t)$ shares the properties of the original Hamiltonian $H(t)$, and we can repeat the above analysis using $H^{(1)}(t)$ in place of $H(t)$. In this way the adiabatic approximation yields the approximate evolution operator

$$U^{(1)}(t)=\sum_{n_1}\sum_{a_1,b_1=1}^{\mathcal{N}_1} K^{(1)n_1}_{a_1b_1}(t)|\psi^{(1)}_{n_1},a_1;t\rangle\langle\phi^{(1)}_{n_1},b_1;0| \tag{26}$$

for $H^{(1)}(t)$, where $K^{(1)n_1}_{a_1b_1}(t)$ are the entries of the matrix $K^{(1)n_1}$ obtained by replacing $E_n(t)$, $|\psi_n,a;t\rangle$ and $|\phi_n,a;t\rangle$ in Eqs. (14) and (15) by $E^{(1)}_{n_1}(t)$, $|\psi^{(1)}_{n_1},a_1;t\rangle$, and $|\phi^{(1)}_{n_1},a_1;t\rangle$, respectively.

Next we perform the adiabatic canonical transformation defined by $g(t)=U^{(1)}(t)^{-1}$. This leads to a transformed Hamiltonian $H^{(2)}(t)$ which is related to $H^{(1)}(t)$ according to Eqs. (23) and (24) with K^n, $|\psi_n,a;t\rangle$, and $|\phi_n,b;t\rangle$ replaced by $K^{(1)n_1}$, $|\psi^{(1)}_{n_1},a_1;t\rangle$, and $|\phi^{(1)}_{n_1},a_1;t\rangle$. The transformed evolution operator is given by

$$U^{(1)}(t)^{-1}U^{(0)}(t)^{-1}U(t).$$

Repeating this procedure we obtain, after N successive adiabatic canonical transformations, a transformed Hamiltonian $H^{(N)}(t)$ and a transformed evolution operator which is given by

$$U^{(N-1)}(t)^{-1}U^{(N-2)}(t)^{-1}\cdots U^{(0)}(t)^{-1}U(t).$$

Here $U^{(\ell)}(t)$, with $\ell\in\{1,2,...,N-1\}$, denotes the approximate evolution operator obtained by performing adiabatic approximation on the Hamiltonian $H^{(\ell)}(t)$.

If for some N the adiabatic approximation yields the exact solution of the Schrödinger equation for the Hamiltonian $H^{(N)}(t)$, then by construction $H^{(N+1)}(t)=0$ and $U^{(N+1)}(t)=1$. In this case, the original evolution operator is given by

$$U(t)=U^{(0)}(t)U^{(1)}(t)\cdots U^{(N)}(t). \tag{27}$$

If the adiabatic approximation fails for all $H^{(N)}(t)$, then there are two possibilities

(i) One obtains an infinite product expansion for the evolution operator

$$U(t)=\prod_{\ell=0}^{\infty} U^{(\ell)}(t):=U^{(0)}(t)U^{(1)}(t)\cdots U^{(\ell)}(t)\cdots. \tag{28}$$

In this case, one may view Eq. (27) as a *generalization* of the adiabatic approximation.

(ii) One obtains $H^{(i)}(t)=H^{(j)}(t)$ for some i and j with $i\neq j$. In this case a direct application of the method of adiabatic product expansion does not produce a solution. However, as we shall see in Sec. IV, sometimes it is possible to modify this method by combining the adiabatic canonical transformation with other canonical transformations, so that one obtains a finite or an infinite product expansion with distinct terms.

IV. APPLICATION TO TWO-LEVEL HAMILTONIANS

Two-level nondegenerate Hamiltonians provide the simplest nontrivial quantum systems. This has been one of the main reasons for the study of these Hamiltonians since the early days of quantum mechanics. In this section we shall consider the most general nondegenerate two-level Hamiltonian which may or may not be Hermitian.

In an arbitrary basis of the Hilbert space ($\mathbb{C}^2$), the Hamiltonian is given by a two-by-two complex matrix $\overline{H}$. One can perform a quantum canonical transformation (21) defined by $g(t)=\exp\{i\int_0^t[\mathrm{tr}\,\overline{H}(s)]ds/2\}$ to map the Hamiltonian $\overline{H}$ to a traceless Hamiltonian of the form

$$H:=\begin{pmatrix} a & b\\ c & -a\end{pmatrix}, \tag{29}$$

where $\mathrm{tr}\,\overline{H}$ denotes the trace of $\overline{H}$, and $a=a(t)$, $b=b(t)$, $c=c(t)$ are complex-valued smooth functions of t.

We can easily solve the eigenvalue problem for the Hamiltonian (29). The eigenvalues are given by

$$E_1(t):=-E(t), \quad E_2(t):=E(t),$$

where

$$E:=\sqrt{a^2+bc}. \tag{30}$$

We shall demand that during the time interval $[0,\tau]$ of interest $E\neq 0$, so that the eigenvalues are nondegenerate. In particular, no level crossings occur. Then a possible choice for a biorthonormal eigenbasis is

$$|\psi_1;R\rangle=\begin{pmatrix} -b \\ a+E \end{pmatrix}, \quad |\psi_2;R\rangle=\begin{pmatrix} a+E \\ c \end{pmatrix}, \tag{31}$$

$$|\phi_1;R\rangle=\frac{1}{N^*}\begin{pmatrix} -c^* \\ a^*+E^* \end{pmatrix}, \quad |\phi_2;R\rangle=\frac{1}{N^*}\begin{pmatrix} a^*+E^* \\ b^* \end{pmatrix}, \tag{32}$$

where $R=(a,b,c)$ and $N:=2E(a+E)$.

Next we compute $U^{(0)}$ and $H^{(1)}$. Using Eqs. (18), (14), (15), and (23), we find

$$U^{(0)}(t)=K^1(t)|\psi_1;t\rangle\langle\phi_1;0|+K^2(t)|\psi_2;t\rangle\langle\phi_2;0|, \tag{33}$$

$$H^{(1)}(t)=\xi(t)|\psi_1;0\rangle\langle\phi_2;0|+\zeta(t)|\psi_2;0\rangle\langle\phi_1;0|, \tag{34}$$

where

$$K^1(t):=K^1_{11}(t)=\exp\left(\frac{i\eta(t)}{2}-\int_{R(0)}^{R(t)}\left[(2E)^{-1}\left(da+dE+\frac{cdb}{a+E}\right)\right]\right),$$

$$K^2(t):=K^2_{11}(t)=\exp\left(\frac{-i\eta(t)}{2}-\int_{R(0)}^{R(t)}\left[(2E)^{-1}\left(da+dE+\frac{bdc}{a+E}\right)\right]\right), \tag{35}$$

$$\eta(t):=2\int_0^t E(s)ds,$$

$$\xi(t):=H^{12}_{11}(t)=\left(-\frac{ie^{-2i\alpha(t)}}{2}\right)\left[1+\frac{a(t)}{E(t)}\right]\frac{d}{dt}\left[\frac{c(t)}{a(t)+E(t)}\right], \tag{36}$$

$$\zeta(t):=H^{21}_{11}(t)=\left(\frac{ie^{2i\alpha(t)}}{2}\right)\left[1+\frac{a(t)}{E(t)}\right]\frac{d}{dt}\left[\frac{b(t)}{a(t)+E(t)}\right], \tag{37}$$

$$\alpha(t):=\frac{\eta(t)}{2}+\frac{i}{4}\int_{R(0)}^{R(t)}\frac{cdb-bdc}{E(E+a)}. \tag{38}$$

The transformed Hamiltonian has the following matrix expression:

$$H^{(1)}(t)=\begin{pmatrix} a^{(1)}(t) & b^{(1)}(t) \\ c^{(1)}(t) & -a^{(1)}(t) \end{pmatrix}, \tag{39}$$

where

$$a^{(1)}(t):=-\frac{b_0\xi(t)+c_0\zeta(t)}{2E_0}, \tag{40}$$

$$b^{(1)}(t):=-\frac{b_0^2\xi(t)-(a_0+E_0)^2\zeta(t)}{2E_0(E_0+a_0)}, \tag{41}$$

$$c^{(1)}(t):=-\frac{-(a_0+E_0)^2\xi(t)+c_0^2\zeta(t)}{2E_0(E_0+a_0)}, \tag{42}$$

$a_0:=a(0)$, $b_0:=b(0)$, $c_0:=c(0)$, $E_0:=E(0)$, and we have used Eqs. (31), (32), and (34).

Note that the transformed Hamiltonian $H^{(1)}(t)$ is traceless, and one can obtain $H^{(2)}(t)$ by substituting $a^{(1)}$ for $a, b^{(1)}$ for $b, c^{(1)}$ for c, and $E^{(1)}:=\sqrt{(a^{(1)})^2+b^{(1)}c^{(1)}}$ for E in Eqs. (34), and (36)–(38). Clearly this can be repeated indefinitely, and one can compute $H^{(\ell)}$ for arbitrary ℓ.

The adiabatic approximation corresponds to the cases where the matrix elements of $H^{(1)}(t)$ can be neglected. As seen from Eqs. (39) to (42) this happens whenever both ξ and ζ are negligible. One can also check that if only one of these quantities is negligible, then $H^{(1)}(t)$ is equal to the other times a constant matrix. This means that $H^{(1)}(t)$ has essentially stationary eigenvectors and the adiabatic approximation would yield the solution of the Schrödinger equation for $H^{(1)}$. In fact, it is not difficult to check that for the cases that either ξ or ζ is negligible, $H^{(2)}(t)\approx 0$. In particular, setting $\xi=0$ or $\zeta=0$ implies $H^{(2)}(t)=0$ and the evolution operator is given by

$$U(t)=U^{(0)}(t)U^{(1)}(t). \tag{43}$$

Therefore, the conditions $\xi=0$ and $\zeta=0$ each define a class of exactly solvable two-level systems. In view of Eqs. (36) and (37), these are as follows.

Class 1: The two-level systems for which $c/(a+E)=\mu=\text{constant}$, or alternatively $c=\mu(\mu b+\sqrt{4a^2+\mu^2b^2})/2$.

Class 2: The two-level systems for which $b/(a+E)=\nu=\text{constant}$, or alternatively $c=b/\nu^2-a^2/b$.

In general ξ and ζ do not vanish and the adiabatic product expansion does not terminate. There is also a special class of two-level systems for which the product expansion has a periodic structure in the sense of case (ii) of Sec. III. This is

Class 3: The two-level systems for which $a=0$.

Setting $a=0$ in Eqs. (38), (35), (36), and (37) and defining $f(t):=i\sqrt{c(t)/b(t)}$, we have

$$\alpha(t)=\frac{\eta(t)}{2}+\frac{i}{4}\ln\left(\frac{c_0b(t)}{b_0c(t)}\right), \quad \eta(t)=2\int_0^t\sqrt{b(s)c(s)}\,ds,$$

$$\xi(t)=-\frac{f_0\dot{f}(t)e^{-i\eta(t)}}{2\,f(t)}, \quad \zeta(t)=\frac{\dot{f}(t)e^{i\eta(t)}}{2\,f_0f(t)},$$

where $f_0:=f(0)$. Substituting these equations in Eq. (39), we obtain

$$H^{(1)}(t)=E^{(1)}(t)\begin{pmatrix}\cos\eta(t) & f_0^{-1}\sin\eta(t)\\ f_0\sin\eta(t) & -\cos\eta(t)\end{pmatrix}, \tag{44}$$

where

$$E^{(1)}(t)=\frac{i\dot{f}(t)}{2\,f(t)}.$$

This Hamiltonian has two interesting properties.

(1) If $b_0=c_0$, then $f_0=i$ and

$$H^{(1)}(t)=E^{(1)}(t)[\sin\eta(t)\sigma_2+\cos\eta(t)\sigma_3]=E^{(1)}(t)e^{i\eta(t)\sigma_1/2}\sigma_3e^{-i\eta(t)\sigma_1/2}, \tag{45}$$

where σ_i are Pauli matrices, and we have used the identity

$$e^{-i\varphi\sigma_i}\sigma_j e^{i\varphi\sigma_i}=\cos(2\varphi)\sigma_j+\sin(2\varphi)\sum_{k=1}^{3}\epsilon_{ijk}\sigma_k \quad \text{for } i\neq k. \tag{46}$$

In Eq. (46), φ is an arbitrary complex variable and ϵ_{ijk} is the totally antisymmetric Levi–Civita symbol with $\epsilon_{123}=1$. For the time periods during which $\sqrt{b(t)c(t)}$ is real, $\eta(t)$, is real, and the Hamiltonian (45) is anti-Hermitian. In particular, its eigenvectors are orthogonal. Up to a factor of i this Hamiltonian describes the interaction of a spin 1/2 magnetic dipole with a changing magnetic field. This system has an SU(2) dynamical group.[19,20,1,3] For the time periods during which $\sqrt{b(t)c(t)}$ is imaginary, $\eta(t)$ is imaginary, and up to a factor of i the Hamiltonian (45) describes a quantum system with a SU(1,1) dynamical group. A Hermitian analog of such a system is the time-dependent generalized harmonic oscillator.[21,20,22]

(2) Performing the adiabatic canonical transformation on (45), we arrive at the unexpected result

$$H^{(2)}(t)=H(t). \tag{47}$$

Therefore, direct application of the method of adiabatic product expansion does not lead to a solution.

Next we shall describe a modification of the method of adiabatic product expansion which yields an infinite product expansion for the evolution operator of the Class 3 systems which involve distinct terms.

Consider the transformed Hamiltonian (44). We can express this Hamiltonian using Eq. (39) with

$$a^{(1)}(t)=E^{(1)}(t)\cos\eta(t),\quad b^{(1)}(t)=f_0^{-1}E^{(1)}(t)\sin\eta(t),\quad c^{(1)}(t)=f_0E^{(1)}(t)\sin\eta(t). \tag{48}$$

Although this Hamiltonian does not belong to Class 3, it can be canonically transformed to a Hamiltonian which belongs to Class 3, i.e., its diagonal matrix elements vanish. This transformation is defined by $g(t)=\exp\{i\int_0^t a^{(1)}(s)ds\sigma_3\}$. The corresponding transformed Hamiltonian is given by

$$H_1(t)=\begin{pmatrix} 0 & b_1(t)\\ c_1(t) & 0\end{pmatrix}, \tag{49}$$

where

$$b_1(t):=b^{(1)}(t)e^{i\gamma_1(t)},\quad c_1(t):=c^{(1)}(t)e^{-i\gamma_1(t)},\quad \gamma_1(t):=2\int_0^t a^{(1)}(s)ds. \tag{50}$$

The evolution operator U_1 of H_1 is related to the evolution operator of the original Hamiltonian H according to

$$U_1(t)=\exp\left(i\int_0^t a^{(1)}(s)ds\sigma_3\right)U^{(0)}(t)^\dagger U(t), \tag{51}$$

where we have used Eq. (22).

Now since H_1 has the same form as H, we can repeat the above analysis using H_1 in place of H. Performing an adiabatic canonical transformation on H_1 we obtain the transformed Hamiltonian

$$H_1^{(1)}(t)=\begin{pmatrix} a_1^{(1)}(t) & b_1^{(1)}(t) \\ c_1^{(1)} & -a_1^{(1)}(t) \end{pmatrix}, \tag{52}$$

where

$$\begin{gathered} a_1^{(1)}(t):=E_1^{(1)}(t)\cos\eta_1(t), \quad b_1^{(1)}(t):=f_{1,0}^{-1}E_1^{(1)}\sin\eta_1(t), \\ c_1^{(1)}(t):=f_{1,0}E_1^{(1)}(t)\sin\eta_1(t), \quad E_1^{(1)}(t):=E^{(1)}(t)\cos\eta(t), \\ f_1(t):=i\sqrt{\frac{c_1(t)}{b_1(t)}}=if_0e^{-i\gamma_1(t)}, \quad f_{1,0}:=f_1(0)=if_0, \\ \eta_1(t):=2\int_0^t E^{(1)}(s)\sin\eta(s)ds. \end{gathered} \tag{53}$$

Clearly we can repeat this procedure indefinitely and construct an infinite product expansion for the evolution operator. Again if we compute only a finite number of terms in this expansion, then we obtain a generalization of the adiabatic approximation. The validity of this approximation may be checked by computing the transformed Hamiltonians. It is not difficult to show that the transformed Hamiltonian obtained after ℓ adiabatic canonical transformations is of the form

$$H_\ell^{(1)}(t)=h_\ell(t)S(t),$$

where

$$H_\ell(t)=E^{(1)}(t)\cos\eta(t)\cos\eta_1(t)\cos\eta_2(t)\cdots\cos\eta_{\ell-1}(t),$$

$$\eta_j(t):=2\int_0^t E^{(1)}(s)\cos\eta(s)\cos\eta_1(s)\cdots\cos\eta_{j-1}(s)\sin\eta_{j-1}(s)ds,$$

where $j\in\{2,3,...,\ell-1\}$ and $S(t)$ is a two-by-two matrix of unit determinant. Clearly if for some ℓ, $h_\ell(t)$ is negligible, then the above-mentioned generalization of the adiabatic approximation is valid.

Finally let us note that in general the initial Hamiltonian (29) can be written in the form

$$H(t)=\alpha_1(t)\sigma_1+\alpha_2(t)\sigma_2+a(t)\sigma_3, \tag{54}$$

with $\alpha_1=(b+c)/2$ and $\alpha_2=i(b-c)/2$. Performing the canonical transformation (21) defined by $g(t)=\exp\{i\int_0^t\alpha_2(s)ds\,\sigma_2\}$, we transform the Hamiltonian (54) into

$$H'(t)=\alpha'(t)\sigma_1+a'(t)\sigma_3=\begin{pmatrix} a'(t) & \alpha'(t) \\ \alpha'(t) & -a'(t) \end{pmatrix},$$

where

$$\alpha'(t):=\alpha_1(t)\cos\xi(t)-a(t)\sin\xi(t), \quad a'(t):=\alpha_1(t)\sin\xi(t)+a(t)\cos\xi(t),$$

$$\xi(t):=\int_0^t\alpha_2(s)ds. \tag{55}$$

Here we have used Eqs. (21) and (46). Next we perform another canonical transformation, namely the one defined by $g(t)=\exp\{i\int_0^t a'(s)ds\,\sigma_3\}$. This transformation maps the Hamiltonian (55) into

$$H''(t)=\alpha'(t)e^{i\eta'(t)\sigma_3/2}\sigma_1 e^{-i\eta'(t)\sigma_3/2}$$
$$=\alpha'(t)[\cos\eta'(t)\sigma_1-\sin\eta'(t)\sigma_2]=\alpha'(t)\begin{pmatrix}0 & e^{i\eta'(t)}\\ e^{-i\eta'(t)} & 0\end{pmatrix}, \tag{56}$$

where $\eta'(t):=2\int_0^t a'(s)ds$. This Hamiltonian is not only a member of Class 3 Hamiltonians, but initially (at $t=0$) its off-diagonal matrix elements are equal. In particular, it has the properties 1 in the above list. Note that we can carry out these canonical transformations on any two-level Hamiltonian. Therefore, every two-level Hamiltonian is canonically equivalent to a Class 3 Hamiltonian of the form (56). This means that the results obtained for Class 3 Hamiltonians apply to arbitrary two-level Hamiltonians.

V. TIME-DEPENDENT SIMPLE HARMONIC OSCILLATOR

It is well-known that the solution of every second-order linear ODE[23] can be reduced to the classical equation of motion for a simple harmonic oscillator with a time-dependent frequency $\omega=\omega(t)$,

$$\ddot{x}(t)+\omega^2(t)x(t)=0. \tag{57}$$

It is also well-known that one can reduce both the classical and quantum equations of motion for a generalized harmonic oscillator to Eq. (57).[24,25,22,26] This equation has, therefore, many physical applications.[25,27] Yet an exact analytic expression for the general solution of this equation is not known even for the case of real frequency.[28] The lack of an exact analytic solution of Eq. (57) is not surprising. One way to see this is to recall that the time-independent Schrödinger equation for an arbitrary potential $V(x)$ in one dimension is given by

$$\frac{d^2\psi_n}{dx^2}+\left(\frac{\hbar^2[E_n-V(x)]}{2m}\right)\psi_n=0, \tag{58}$$

where E_n and ψ_n are the energy eigenvalues and eigenfunctions, respectively. Equation (58) can be easily identified with Eq. (57) provided that one makes the change of variables: $x\to t$, $\psi_n\to x$, and $\{\hbar^2[E-V(x)]\}/(2m)\to\omega^2(t)$. This shows that if one was able to find the exact analytic solution of Eq. (57) for arbitrary frequency ω, then one would have been able to find the general solution of the time-independent Schrödinger equation for any potential V.

In the following we shall consider the case of an ordinary time-dependent harmonic oscillator (57) with real frequency. In order to apply the results of Sec. IV to Eq. (57), we first express it in the form of a system of first-order ODEs. Defining,

$$|\psi(t)\rangle:=\begin{pmatrix}x(t)\\ v(t)\end{pmatrix},\quad v(t):=\dot{x}(t),$$

we can write Eq. (57) in the form of the Schrödinger equation (1) with a two-level Hamiltonian of the form (29) with

$$a=0,\quad b=i,\quad c=-i\omega(t)^2,\quad E=\omega(t). \tag{59}$$

Since $a=0$, this system belongs to the Class 3 of Sec. IV with

$$f(t)=\omega(t), \quad \eta(t)=2\int_0^t \omega(s)ds,$$
$$H^{(1)}=E^{(1)}(t)\begin{pmatrix} \cos\eta(t) & \dfrac{\sin\eta(t)}{\omega_0} \\ \omega_0\sin\eta(t) & \cos\eta(t)\end{pmatrix}, \quad E^{(1)}(t):=\frac{i\dot{\omega}(t)}{2\omega(t)}. \tag{60}$$

Clearly, for real frequency $\omega(t)$ we can scale the time variable t so that $\omega_0=1$. Then the transformed Hamiltonian (60) takes the form

$$H^{(1)}(t)=E^{(1)}(t)[\sin\eta(t)\sigma_1+\cos\eta(t)\sigma_3]=E^{(1)}(t)e^{-i\eta(t)\sigma_2/2}\sigma_3 e^{i\eta(t)\sigma_2/2}. \tag{61}$$

Note that since $\eta(t)$ is real, the Hamiltonian (61) is an anti-Hermitian matrix with orthogonal eigenvectors. Therefore, up to a factor of i it describes a two-level spin system with a SU(2) dynamical group.[19,3] [Note that one can absorb the factor of i in the definition of the time variable t, i.e., by defining the imaginary time variable $\tau:=-it$. Therefore, the dynamics given by the Hamiltonian (61) may be viewed as the dynamics of a spin system with imaginary time.] This is rather surprising, for it is well-known that the quantum harmonic oscillator has SU(1,1) dynamical group and that its Schrödinger equation may be reduced to Eq. (57) by means of a quantum canonical transformation corresponding to a time-dependent dilatation.[26]

In view of the fact that $E^{(1)}$ is proportional to the derivative of $\ln\omega$, we can make a change of independent variable, namely $t\to\eta$. Note that η is the integral of a positive real function of t. Hence, it is a monotonically increasing function of t. Making this change of variable the Schrödinger equation for the Hamiltonian (61) becomes

$$i\frac{d}{d\eta}\tilde{U}(\eta)=\tilde{H}(\eta)\tilde{U}(\eta), \quad \tilde{U}(\eta)=1,$$

where $\tilde{U}(\eta):=U'(t(\eta))$, $U'(t)$ is the evolution operator for the the Hamiltonian (61),

$$\tilde{H}(\eta):=\tilde{E}(\eta)e^{-i\eta\sigma_2/2}\sigma_3 e^{i\eta\sigma_2/2}=\tilde{E}(\eta)(\sin\eta\sigma_1+\cos\eta\sigma_3), \tag{62}$$

$$\tilde{E}(\eta):=\frac{i\omega'(\eta)}{2\omega(\eta)}, \quad \omega':=\frac{d\omega}{d\eta}. \tag{63}$$

Up to a factor of i, the Hamiltonian (62) describes the interaction of a spin 1/2 magnetic dipole with a changing magnetic field whose direction rotates uniformly in the $x-z$ plane.

As we mentioned in Sec. IV for the Class 3 systems $H^{(2)}(t)=H(t)$. Hence direct application of the method of the adiabatic product expansion does not lead to a solution of the Schrödinger equation for the Hamiltonian (61) or (62). In this case, either one constructs the modified adiabatic product expansion of Sec. IV or examines the adiabatic series expansion of Ref. 2. The latter yields a series expansion for the evolution operator $\tilde{U}(\eta)$ of the Hamiltonian $\tilde{H}(\eta)$, namely

$$\begin{aligned}\tilde{U}(\eta)&=\mathcal{T}\exp\left(-i\int_0^\eta \tilde{H}(s)ds\right)\\ &=1-i\int_0^\eta \tilde{H}(s)ds+\frac{(-i)^2}{2}\int_0^\eta\int_0^\eta \mathcal{T}[\tilde{H}(s_1)\tilde{H}(s_2)]ds_1\,ds_2+\cdots+\frac{(-i)^n}{n!}\\ &\quad\times\int_0^\eta\cdots\int_0^\eta \mathcal{T}[\tilde{H}(s_1)\cdots\tilde{H}(s_n)]ds_1\cdots ds_n+\cdots,\end{aligned} \tag{64}$$

where $\mathcal{T}$ stands for the time-ordering operator. Since $\tilde{H}(\eta)$ is proportional to $\omega'(\eta)$, for slowly varying ω one obtains an approximate expression for $\tilde{U}(\eta)$ by computing a finite number of terms in this series. This is in fact another generalization of the adiabatic approximation, because if one keeps only the first term in this series and neglects the other terms one is essentially neglecting $\tilde{H}$ or alternatively $H^{(1)}$. As we explained above, this is just the adiabatic approximation. If one keeps more terms in this series, then one obtains a better approximation than the adiabatic approximation.

VI. QUADRUPOLE INTERACTION OF A SPIN 1 PARTICLE WITH A CHANGING ELECTRIC FIELD

Consider a spin 1 particle interacting with a changing electric field $\vec{\mathcal{E}}(t) = (\mathcal{E}_1(t),\mathcal{E}_2(t),\mathcal{E}_3(t))$ according to the Stark Hamiltonian

$$H(t)=\lambda[\vec{J}\cdot\vec{\mathcal{E}}(t)]^2, \tag{65}$$

where λ is a real coupling constant and $\vec{J}$ is the angular momentum of the particle. The quadrupole interactions of the form (65) have been extensively studied for fermionic systems in relation with the non-Abelian geometric phases[29–31] (See also Ref. 32.) The occurrence of non-Abelian geometric phases for the degenerate spin 1 systems has been pointed out in Ref. 33. For these systems, the particle has a definite angular momentum $j=1$ and the Hamiltonian is a 3×3 matrix. Using the spin $j=1$ representation of J_i, we can express the Stark Hamiltonin (65) in the form

$$H=\left(\frac{\lambda r^2}{2}\right)\begin{pmatrix} 1+2z^2 & \sqrt{2}ze^{-i\theta} & e^{-2i\theta} \\ \sqrt{2}ze^{i\theta} & 2 & -\sqrt{2}ze^{-i\theta} \\ e^{2i\theta} & -\sqrt{2}ze^{i\theta} & 1+2z^2 \end{pmatrix}, \tag{66}$$

where r, θ, and z are defined by

$$r:=\sqrt{\mathcal{E}_1^2+\mathcal{E}_2^2},\quad e^{i\theta}:=\frac{\mathcal{E}_1+i\mathcal{E}_2}{r},\quad z:=\frac{\mathcal{E}_3}{r}.$$

In view of the general results of Ref. 33, if $r\neq 0$ then the Hamiltonian (66) has a degenerate and a nondegenerate eigenvalue. In the following we shall consider the case where $\mathcal{E}_3=0$. The general case $\mathcal{E}_3\neq 0$ can be similarly treated.

If $\mathcal{E}_3=0$, then $z=0$ and

$$H=\left(\frac{\lambda r^2}{2}\right)\begin{pmatrix} 1 & 0 & e^{-2i\theta} \\ 0 & 2 & 0 \\ e^{2i\theta} & 0 & 1 \end{pmatrix}. \tag{67}$$

The eigenvalues of this Hamiltonian are given by

$$E_1=0,\quad E_2=\lambda r^2. \tag{68}$$

For $r\neq 0$, E_1 is nondegenerate and E_2 is doubly degenerate. A set of orthonormal eigenvectors of this Hamiltonian is given by

$$|\psi_1;R\rangle:=\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 0 \\ e^{2i\theta} \end{pmatrix},\quad |\psi_2,1;R\rangle:=\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ e^{2i\theta} \end{pmatrix},\quad |\psi_2,2;R\rangle:=\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{69}$$

where $R=(r,\theta)$.

Next we compute $U^{(0)}(t)$ for this system. In order to do this we first use Eq. (15) to calculate $\mathcal{A}^n$. In view of the fact that the Hamiltonian (67) is Hermitian, $|\phi_n,a;R\rangle=|\psi_n,a;R\rangle$ and Eq. (15) leads to

$$\mathcal{A}^1=-d\theta, \quad \mathcal{A}^2=\begin{pmatrix} -d\theta & 0 \\ 0 & 0 \end{pmatrix}. \tag{70}$$

Substituting Eqs. (70) into Eq. (14) and making use of Eq. (68), we find

$$K^1(t)=e^{-i[\theta(t)-\theta_0]}, \quad K^2(t)=e^{-i\rho(t)}\begin{pmatrix} e^{-i[\theta(t)-\theta_0]} & 0 \\ 0 & 1 \end{pmatrix}, \tag{71}$$

where

$$\theta_0:=\theta(0), \quad \rho(t):=\lambda\int_0^t r(s)^2\,ds. \tag{72}$$

Using Eqs. (18) and (69), we have

$$U^{(0)}(t)=\frac{1}{2}\begin{pmatrix} (1+e^{-i\rho(t)})e^{i\theta_-(t)} & 0 & (-1+e^{-i\rho(t)})e^{-i\theta_+(t)} \\ 0 & 2e^{-i\rho(t)} & 0 \\ (-1+e^{-i\rho(t)})e^{i\theta_+(t)} & 0 & (1+e^{-i\rho(t)})e^{i\theta_-(t)} \end{pmatrix}, \tag{73}$$

where $\theta_\pm(t):=\theta(t)\pm\theta_0$.

Next we compute the Hamiltonian $H^{(1)}(t)$. This involves the calculation of $A_{ab}^{nm}(t)$ and $H_{ab}^{mn}(t)$ for $m\neq n$. Using Eqs. (24) and (69) we have

$$A_{11}^{12}=A_{11}^{21}=-\dot{\theta}, \quad A_{12}^{12}=A_{21}^{21}=0,$$

$$H_{11}^{21}=H_{11}^{12}=\dot{\theta}e^{i\rho(t)}, \quad H_{12}^{12}=H_{21}^{21}=0.$$

Substituting these equations in Eq. (23) and using Eq. (69), we obtain

$$H^{(1)}(t)=-\dot{\theta}(t)\begin{pmatrix} \cos\rho(t) & 0 & -i\sin\rho(t) \\ 0 & 0 & 0 \\ i\sin\rho(t) & 0 & -\cos\rho(t) \end{pmatrix}=-\dot{\theta}(t)[\sin\rho(t)\Sigma_2+\cos\rho(t)\Sigma_3], \tag{74}$$

where

$$\Sigma_2:=\begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \Sigma_3:=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{75}$$

It is not difficult to recognize Σ_2 and Σ_3 as the Pauli matrices σ_2 and σ_3 represented in a $(0+1/2)$ representation of SU(2). In view of this identification we can express $H^{(1)}(t)$ in the form

$$H^{(1)}(t)=-\dot{\theta}(t)e^{i\rho(t)\Sigma_1/2}\Sigma_3e^{-i\rho(t)\Sigma_1/2}, \tag{76}$$

where

$$\Sigma_1 := \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

and we have used Eq. (46).

The Hamiltonian $H^{(1)}(t)$ has the following interesting properties.

(a) For $\theta=0$, i.e., $\theta=$constant, the adiabatic approximation is exact and $U(t)=U^{(0)}(t)$.

(b) In view of Eq. (76) for $\dot\theta\neq 0$, $H^{(1)}(t)$ has three nondegenerate eigenvalues, namely $-\dot\theta$, 0, and $\dot\theta$. This is quite remarkable because it shows that the adiabatic canonical transformation maps the degenerate Hamiltonian (67) into the nondegenerate Hamiltonian (76).

(c) Since Σ_i is a representation of the Pauli matrix σ_i, the Hamiltonian (76) belongs to a representation of the Lie algebra of SU(2). This means that one can reduce the Schrödinger equation for this Hamiltonian to that of the dipole Hamiltonian[2,3]

$$H_{\rm dp} = -2\,\dot\theta(t) e^{i\rho(t)J_1} J_3 e^{-i\rho(t)J_1}.$$

(d) We can perform another canonical transformation, namely the one defined by $g(t) = \exp[-i\rho(t)\Sigma_1/2]$ to transform the Hamiltonian (76) into

$$H^{(1)'}(t) = \frac{r(t)^2}{2}\Sigma_1 - \dot\theta(t)\Sigma_3, \tag{77}$$

where we have used Eqs. (76), (21), and (72). In particular if $\dot\theta$ and r^2 happen to be proportional, i.e., for some $c\in\mathbb{R}$

$$\dot\theta(t) = c\,r(t)^2, \tag{78}$$

then $H^{(1)'}(t) = r(t)^2(\frac{1}{2}\Sigma_1 - c\Sigma_3)$. In this case the eigenvectors of $H^{(1)'}(t)$ are constant and the adiabatic approximation yields the exact solution of the Schrödinger equation for $H^{(1)'}(t)$. The corresponding evolution operator is then given by

$$U^{(1)'}(t) = \exp\left(i\left(\frac{1}{2}\Sigma_1 - c\Sigma_3\right)\int_0^t r(s)^2\, ds\right). \tag{79}$$

Having obtained the evolution operator for $H^{(1)'}(t)$ we can use Eq. (22) to obtain the evolution operator for $H^{(1)}(t)$ and $H(t)$. This yields the following expression for the evolution operator for $H(t)$:

$$U(t) = U^{(0)}(t) e^{i\rho(t)\Sigma_1/2} U^{(1)'}(t), \tag{80}$$

where $U^{(0)}(t)$ and $U^{(1)'}(t)$ are given by Eqs. (73) and (79).

The above analysis shows that the condition (78) defines a class of exactly solvable time-dependent Stark Hamiltonians. If $\theta=\omega t$, for some constant frequency ω, this condition corresponds to the case of the rotating electric field $\vec E = r(\sin\omega t, \cos\omega t, 0)$ with magnitude r.

VII. CONCLUSION

In this article we have extended the method of the adiabatic product expansion to non-Hermitian and degenerate Hamiltonians. We showed that in general there were three possibilities for the adiabatic product expansion.

(1) The expansion terminates after a finite number of iteractions. This happens when one of the transformed Hamiltonians vanishes. In this case the method yields the exact solution for the Schrödinger equation.

(2) The expansion consists of an infinite number of distinct terms. In this case, the method does not lead to an exact solution, but it gives rise to a generalization of the adiabatic approximation. This approximation is performed by keeping a finite number of terms in the product expansion. The general asymptotic behavior of the adiabatic product expansion has not been studied. However, one can interpret this approximation by recalling that the condition for the termination of the product expansion corresponds to the validity of the conventional adiabatic approximation for one of the transformed Hamiltonians.

(3) The expansion involves terms which are not distinct. In this case the expansion does not lead to a solution. However, usually one can make another time-dependent canonical transformation after each adiabatic transformation and obtain an infinite product expansion with the properties of case (2) above.

We have considered some specific problems that one can attempt to solve using this method. We treated the case of a general nondegenerate two-level system and applied our general results to the more specific case of the classical equation of motion for a harmonic oscillator with a time-dependent frequency. In this case, we showed that the adiabatic canonical transformation mapped the corresponding two-level quantum system to a quantum system with an anti-Hermitian Hamiltonian. Although the direct application of the method of adiabatic product expansion did not yield a solution, we could construct the modified adiabatic product expansion. We have also outlined an adiabatic series expansion for the time-evolution operator of this system which led to another generalization of the adiabatic approximation. Finally, we considered the application of our method to treat the quadrupole interaction of a spin 1 particle with a changing electric field. The corresponding (Stark) Hamiltonian had a nondegenerate as well as a degenerate eigenvalue. We showed that the adiabatic canonical transformation mapped this Hamiltonian to a Hamiltonian which had nondegenerate eigenvalues and belonged to a reducible $(0+1/2)$ representation of the Lie algebra of SU(2). This means that we can directly use the results of Refs. 2 and 3 which treat the Schrödinger equation for a nondegenerate Hamiltonian belonging to (an irreducible representation of) the Lie algebra of SU(2). Furthermore, we identified a class of exactly solvable spin 1 quadruple Hamiltonians.

[1] A. Mostafazadeh, Phys. Lett. A **228**, 7 (1997).
[2] A. Mostafazadeh, Phys. Rev. A **55**, 1653 (1997).
[3] A. Mostafazadeh, J. Math. Phys. **38**, 3489 (1997).
[4] A. Mostafazadeh, Phys. Rev. A **55**, 4084 (1997).
[5] H. C. Baker, Phys. Rev. Lett. **50**, 1579 (1983); A. Siegma, Opt. Commun. **31**, 369 (1979); G. Dattoli, T. Hermsen, L. Mezi, and A. Torrr, Phys. Rev. A **37**, 4334 (1988).
[6] G. Dattoli, A. Torrr, and R. Mignani, Phys. Rev. A **42**, 1467 (1990).
[7] A. Kvitsinsky and S. Putterman, J. Math. Phys. **32**, 1403 (1991).
[8] J. Wong, J. Math. Phys. **8**, 2039 (1967).
[9] J. C. Garrison and E. M. Wright, Phys. Lett. A **128**, 177 (1988).
[10] M. V. Berry, Proc. R. Soc. London, Ser. A **392**, 45 (1984).
[11] G. Dattoli, R. Mignani, and A. Torre, J. Phys. A **23**, 5795 (1990).
[12] Ch. Miniature, C. Sire, J. Baudon, and J. Bellissard, Europhys. Lett. **13**, 199 (1990).
[13] A. Mondragón and E. Hernández, J. Phys. A **29**, 2567 (1996).
[14] A. Mostafazdeh, Koç University, preprint.
[15] M. Born and V. Fock, Z. Phys. **51**, 165 (1928).
[16] T. Kato, J. Phys. Soc. Jpn. **5**, 435 (1950).
[17] G. Nenciu and G. Rasche, J. Phys. A **25**, 5741 (1992).
[18] F. Wilczek and A. Zee, Phys. Rev. Lett. **52**, 2111 (1984).
[19] S. J. Wang and W. Zuo, Phys. Lett. A **196**, 13 (1994).
[20] S. V. Prants, J. Phys. A **19**, 3657 (1986); Y.-Z. Lai, J.-Q. Liang, H. J. W. Müller-Kirsten, and J.-G. Zhuo, *ibid.* **29**, 1773 (1996); S. Zhang and F. Li, *ibid.* **29**, 6143 (1996).
[21] S. J. Wang, W. Zuo, A. Weiguny, and F. L. Li, Phys. Lett. A **196**, 7 (1994); V. Penna, Ann. Phys. (N.Y.) **245**, 389 (1996).
[22] R. Jackiw, Int. J. Mod. Phys. A **3**, 285 (1988).
[23] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956).
[24] K. Husimi, Prog. Theor. Phys. **9**, 381 (1953).
[25] V. V. Dodonov and M. A. Mańko, in *Invariants and the Evolution of Nonstationary Quantum Systems*, edited by M. A. Markov (Nova, Commack, NY, 1989).
[26] A. Mostafazadeh, J. Phys. A **31**, 6495 (1998).

[27] W. Paul, Rev. Mod. Phys. **62**, 531 (1990); F. Diedrich, J. C. Bergquist, W. M. Itano, and D. J. Wineland, Phys. Rev. Lett. **62**, 403 (1989); D. J. Heinzen and D. J. Wineland, Phys. Rev. A **42**, 2977 (1990); G. S. Agrawal and S. Arun Kumar, Phys. Rev. Lett. **67**, 3665 (1991).
[28] R. Bellman, *Perturbation Techniques in Mathematics, Physics, and Engineering* (Holt, Rinehart and Winston, New York, 1964).
[29] C. A. Mead, Phys. Rev. Lett. **59**, 161 (1987).
[30] J. E. Avron, L. Sadun, J. Segret, and B. Simon, Phys. Rev. Lett. **61**, 1329 (1988).
[31] J. E. Avron, L. Sadun, J. Segret, and B. Simon, Commun. Math. Phys. **124**, 595 (1989).
[32] M. T. Johnsson and I. J. Aitchison, J. Phys. A **30**, 2085 (1997).
[33] A. Mostafazadeh, J. Phys. A **30**, 7525 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Geometric algebra and the causal approach to multiparticle quantum mechanics

Shyamal Somaroo, Anthony Lasenby, and Chris Doran[a)]
Astrophysics Group, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom

(Received 28 January 1998; accepted 14 February 1999)

It is argued that geometric algebra, in the form of the multiparticle spacetime algebra, is well suited to the study of multiparticle quantum theory, with advantages over conventional techniques both in ease of calculation and in providing an intuitive geometric understanding of the results. This is illustrated by comparing the geometric algebra approach for a system of two spin-1/2 particles with the nonrelativistic approach of Holland [Phys. Rep. **169**, 294 (1988)]. © *1999 American Institute of Physics.* [S0022-2488(99)00907-X]

I. INTRODUCTION

Geometric (Clifford) algebra is a powerful algebraic tool with applications throughout the fields of physics and engineering. The geometric algebra of space–time—the spacetime algebra or STA—is well suited to describing many aspects of classical and quantum relativistic physics[1–4] including gravitation.[5] In Refs. 2 and 6 the multiparticle spacetime algebra (MSTA) was introduced and applied to relativistic multiparticle quantum theory. In the present paper the algebraic advantages of the MSTA approach are demonstrated through a comparison with work on a causal approach to nonrelativistic multiparticle quantum theory based on the Pauli equation.[7,8] We show that the MSTA elucidates a number of features of the multiparticle causal theory and, in particular, clarifies its geometric content.

The causal, or Bohmian, approach to quantum mechanics is an interpretation in which the statistical results of quantum theory are recovered from an ensemble of deterministically evolving systems. The approach is based on establishing a connection between the wave equation and a deterministic model that is supposed to underlie the quantum process. In the case of one spin-1/2 particle, this model consists of a classical spinning rigid body under the additional influence of a quantum potential.[7] In this way physical properties can be associated with the quantum particle, and equations for their evolution obtained from the conventional wave equation. Furthermore, the variables (including spin) on which the wave function depends are consistently interpreted as the spatial position and orientation of the particle through this model.

In n-particle nonrelativistic quantum theory the wave function depends on a dynamical configuration space of dimension $3n$, as well as on a temporal parameter t. To apply a causal approach to this system one must first associate the wave function in configuration space with a set of physical properties. These are then interpreted as the properties of the individual particles in the ensemble making up the system under consideration. Equations describing the evolution of these properties are then derived from the conventional n-particle Pauli wave equation.

Holland[7,8] has addressed the problem of extracting a set of physical properties from the n-particle wave function. His method is to construct a set of tensor variables from quadratic combinations of the spinorial wave function. These tensor variables are more easily associated with a set physical properties than the underlying spinorial degrees of freedom.

Here we show that the MSTA formulation of multiparticle quantum theory considerably simplifies the task of extracting these physical variables. Its lack of redundant mathematical

[a)]Electronic mail: C.Doran@mrao.cam.ac.uk

complications considerably simplifies calculations, as well as clarifying the relation between the spinorial and tensorial degrees of freedom. Furthermore, the method of construction readily lends itself to an interpretation in terms of an underlying physical model, though such considerations are not pursued here. In fact, since we are only concerned with nonrelativistic quantum theory, only a fraction of the full power of the MSTA is brought into play here. An introduction to the MSTA approach to relativistic multiparticle quantum theory is contained in Ref. 2 (see also Ref. 6).

We start with an outline of the MSTA and introduce our conventions and notations. We then give a schematic overview of Holland's method of relating spinorial and tensorial variables. In Sec. IV we study the one-particle case, giving a detailed account of the correspondence scheme between the MSTA and Holland's approach. In Sec. V we turn to the two-particle case, as discussed first in Ref. 8. We show how the various physical variables are found more easily in the MSTA approach, and reveal their simple geometric origins. We end with a brief look at how the MSTA approach generalizes to the *n*-particle case.

II. MULTIPARTICLE SPACE–TIME ALGEBRA

Spacetime algebra is the geometric, or Clifford, algebra of Minkowski space–time. Both geometric and spacetime algebra have been widely discussed by many authors (see Refs. 1, 2, 9, and 10 for further material). The multiparticle spacetime algebra (MSTA) was introduced to tackle the problem of formulating relativistic multiparticle mechanics within geometric algebra.[2,6] It is the geometric algebra of *n*-particle configuration space which, for relativistic systems, consists of *n* copies of Minkowski space–time. We usually refer to each copy as a "one-particle space."

An appropriate orthonormal basis for the MSTA is provided by the set $\{\gamma_\mu^a\}$, where $\mu=0,...,3$ labels the space–time vector, and $a=1,...,n$ labels the particle space. For cases where only one particle is present this index is often omitted. These vectors have an associative (geometric) product denoted by juxtaposition. The symmetrized product is denoted by a dot and satisfies

$$\gamma_\mu^a \cdot \gamma_\nu^b = \tfrac{1}{2}(\gamma_\mu^a \gamma_\nu^b + \gamma_\nu^a \gamma_\mu^b) = \eta_{\mu\nu}\delta^{ab}, \tag{2.1}$$

where $\eta_{\mu\nu}=\mathrm{diag}(+,-,-,-)$. Vectors from different particle spaces anticommute as a consequence of their orthogonality. The remaining, antisymmetrized product is denoted by a wedge and generates a bivector,

$$\gamma_\mu^i \wedge \gamma_\nu^j \equiv \tfrac{1}{2}(\gamma_\mu^i \gamma_\nu^j - \gamma_\nu^i \gamma_\mu^j). \tag{2.2}$$

In this manner a basis for the entire MSTA can be constructed. This has 2^{4n} degrees of freedom. A general element of the MSTA is termed a *multivector*.

In this paper we deal only with nonrelativistic quantum mechanics. We therefore need to pick out a preferred timelike vector for each of the particle spaces. We take this vector to be γ_0^a for each *a*. Spatial vectors relative to these timelike vectors are modeled as bivectors through a "space–time split."[3,4] For these we introduce the notation (with no sum over *a*)

$$\sigma_j^a \equiv \gamma_j^a \gamma_0^a, \quad j=1,...,3, \quad a=1,...,n. \tag{2.3}$$

For each particle space the set $\{\sigma_j^a\}$ generates the geometric algebra of relative space, which we denote $\mathcal{G}_3$. Each has a basis of the form

$$1, \quad \{\sigma_j\}, \quad \{i\sigma_j\}, \quad i, \tag{2.4}$$

where the volume element i is defined by

$$i \equiv \sigma_1 \sigma_2 \sigma_3 \tag{2.5}$$

and we have suppressed the particle-space indices. The notation reflects the fact that the geometric algebra of (relative) space is isomorphic to the Pauli algebra, though we stress that the $\{\sigma_j^a\}$ are a basis set of vectors for three-space, and not abstract operators in spin-space. The even subalgebra of the basis (2.4) is spanned by $\{1,\ i\sigma_j\}$ and is isomorphic to the quaternion algebra.

An important property of the $\{\sigma_j^a\}$ is that, unlike space–time basis vectors, relative vectors from separate particle spaces commute. This follows immediately from their definition:

$$\sigma_i^a\sigma_j^b=\gamma_i^a\gamma_0^a\gamma_j^b\gamma_0^b=\gamma_i^a\gamma_j^b\gamma_0^b\gamma_0^a=\gamma_j^b\gamma_0^b\gamma_i^a\gamma_0^a=\sigma_j^b\sigma_i^a \quad (a\neq b). \tag{2.6}$$

It follows that the $\{\sigma_j^a\}$ generate the direct product space $\mathcal{G}_3^n\equiv\mathcal{G}_3\otimes\cdots\otimes\mathcal{G}_3$ of n copies of the geometric algebra of three-dimensional space $\mathcal{G}_3$. All properties of this space follow from the properties of the fully relativistic MSTA.

Within the Pauli algebra of space, an important role is played by *rotors*. These are elements of the even subalgebra of the Pauli algebra satisfying the relation

$$R\tilde{R}=1, \tag{2.7}$$

where the tilde denotes the operation of reversing the order of the vectors in any geometric product in the MSTA. The operation of rotating a multivector is performed by

$$A\mapsto A'=RA\tilde{R}, \tag{2.8}$$

which is easily shown to keep lengths and angles unchanged.

A spinor transforms single sidedly under the action of a rotor, and can be defined as an element of a linear space that is closed under left multiplication by the rotor group. Traditionally, Pauli spinors are either taken as complex column vectors acted on by the 2×2 Pauli matrices, or as elements of a minimal left ideal of the Pauli algebra.[11] A third approach, which turns out to be very powerful in applications, is to represent spinors as elements of the even subalgebra of the Pauli algebra. This space has four real dimensions, and is closed under the action of the rotor group. It is a straightforward matter to establish a $1\leftrightarrow 1$ map between Pauli column spinors and elements of the even subalgebra.[2,12] We start with the Pauli spin matrices in the form

$$\hat{\sigma}_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix},\quad \hat{\sigma}_2=\begin{pmatrix}0 & -j\\ j & 0\end{pmatrix},\quad \hat{\sigma}_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \tag{2.9}$$

where the carets denote that the $\{\hat{\sigma}_i\}$ are explicitly matrices, and j is used for the scalar unit imaginary of quantum mechanics since the symbol i is already employed for the spatial volume element. A column spinor ψ^a is then placed into a $1\leftrightarrow 1$ correspondence with an element of the even subalgebra as follows:

$$\psi^a=\begin{pmatrix}a^0+ja^3\\ -a^2+ja^1\end{pmatrix}\leftrightarrow\psi=a^0+a^k i\sigma_k. \tag{2.10}$$

The action of the quantum operators $\{\hat{\sigma}_k\}$ and j is now replaced by the operations

$$\hat{\sigma}_k|\psi\rangle\leftrightarrow\sigma_k\psi\sigma_3 \quad (k=1,...,3), \tag{2.11}$$

$$j|\psi\rangle\leftrightarrow\psi i\sigma_3. \tag{2.12}$$

Every calculation that can be performed with the column spinor ψ^a can also be performed with the even element ψ, and in practice the latter approach is usually easier. One reason for this is the natural decomposition of ψ into a density term and a rotor:

$$\psi=\rho^{1/2}R, \tag{2.13}$$

where

$$\rho \equiv \psi\tilde{\psi}. \tag{2.14}$$

The rotor R is an instruction to rotate the fixed $\{\sigma_i\}$ frame onto the frame of observables. This establishes a natural link with the description of a rotating rigid body.[9,12]

Nonrelativistic multiparticle spinors are formed from direct products of single particle spinors. If we denote the even subalgebra of $\mathcal{G}_3$ by $\mathcal{G}_3^+$, we see that nonrelativistic MSTA spinors belong to $(\mathcal{G}_3^+)^n = \mathcal{G}_3^+ \otimes \cdots \otimes \mathcal{G}_3^+$. An advantage of the MSTA approach is that this direct product coincides with the geometric product already defined. At this point it is useful to introduce the notation

$$i\sigma_j^a = i^a \sigma_j^a, \tag{2.15}$$

which removes some superscripts without introducing any ambiguity. Multiparticle spinors therefore belong to the space generated by the elements $\{1,\ i\sigma_j^a\}$. This space is closed under the left-sided action of the group of rotors of the form $R^1R^2,\ldots,R^n$, where each R^a denotes a copy of the same rotor for each particle space.

In Eq. (2.12) we saw that the role of the unit imaginary of traditional quantum theory is played by right multiplication by $i\sigma_3$. For the n-particle case there will be n copies of $i\sigma_3$, and right-multiplication by all of these must yield the same result. This is achieved by introducing the n-particle ''correlator''[2]

$$E_n \equiv \tfrac{1}{2}(1 - i\sigma_3^1 i\sigma_3^2)\cdots\tfrac{1}{2}(1 - i\sigma_3^1 i\sigma_3^n), \tag{2.16}$$

which locks the various one-particle complex structures;

$$E_n i\sigma_3^1 = E_n i\sigma_3^2 = \cdots = E_n i\sigma_3^n \equiv J_n. \tag{2.17}$$

The E_n and J_n satisfy

$$E_nE_n = E_n, \quad J_nJ_n = -E_n. \tag{2.18}$$

Correlating all n-particle states $\psi \in (\mathcal{G}_3^+)^n$ by right-multiplying by the idempotent E_n ensures that the conventional complex structure is reproduced by the operation of right multiplication by any of the $i\sigma_3^a$ or J_n. The correlator also reduces the degrees of freedom in an n-particle spinor from 4^n to the expected 2^{n+1}. It is worth noting that one effect of the correlator is to ''phase lock'' all one-particle phase-factors:

$$e^{\alpha i\sigma_3^1}E_n = e^{\alpha i\sigma_3^2}E_n = \cdots = e^{\alpha J_n}E_n. \tag{2.19}$$

This suggests an interesting substructure to the theory, which could prove useful in constructing a suitable particle model for a causal interpretation.

As an example of the above scheme, consider the MSTA analog of the spin singlet state

$$|\epsilon\rangle = \frac{1}{\sqrt{2}}\left\{\begin{pmatrix}1\\0\end{pmatrix}\otimes\begin{pmatrix}0\\1\end{pmatrix} - \begin{pmatrix}0\\1\end{pmatrix}\otimes\begin{pmatrix}1\\0\end{pmatrix}\right\}. \tag{2.20}$$

This is represented in the two-particle MSTA by the multivector

$$\epsilon = \frac{1}{\sqrt{2}}(i\sigma_2^1 - i\sigma_2^2)\,\frac{1}{2}(1 - i\sigma_3^1 i\sigma_3^2). \tag{2.21}$$

It can be shown that ϵ satisfies[2]

$$M^1\epsilon = \tilde{M}^2\epsilon \tag{2.22}$$

for an arbitrary Pauli-even multivector M. This result quickly establishes the rotation invariance of ϵ, since under a rotation in two-particle space, ϵ transforms as

$$\epsilon \mapsto R^1 R^2 \epsilon = R^1 \tilde{R}^1 \epsilon = \epsilon. \tag{2.23}$$

Throughout this paper superscripts on one-particle multivectors denote the one-particle space being referred to. For example, ϕ^k denotes a copy of the one-particle spinor ϕ in the particle-k space. Information regarding the kth particle is extracted from an arbitrary MSTA multivector by projecting it onto the particle-k space. Following Holland, Pauli spin indices are denoted by lower case letters $a-h$, and matrices are denoted by ''caretted'' versions of symbols denoting their algebraic analogs. When considered as a matrix algebra generated by (2.9) over the complex field, the Pauli algebra is denoted as $\mathcal{P}$, with $\mathcal{P}^n$ being the corresponding n-copy direct product. The range $i-l$ is used to denote spatial tensor indices, and the Einstein summation convention also applies throughout unless stated otherwise.

III. THE CAUSAL APPROACH TO MULTIPARTICLE STATES

Conventionally, a nonrelativistic spin-1/2 particle is described by a Pauli spinor ψ^a, usually viewed as a complex linear combination of the spin basis states $|\uparrow\rangle$ and $|\downarrow\rangle$, or more explicitly as a 2×1 complex column matrix. Two spin-1/2 particles are described by a rank-2 spinor, or spin tensor, ψ^{ab}. A basis for this is taken to be a complex linear combination of the direct product of two copies of the spin basis $\{|\uparrow\rangle, |\downarrow\rangle\}$. We may also view ψ^{ab} as a 4×1 complex column matrix by considering the two separate spin indices a and b in ψ^{ab} as one compounded index $\Delta=[ab]$, $\Delta=1,...,4$. This extends to a rank-n spinor $\psi^{abc\cdots}$, which describes a system of n spin-1/2 particles. $\psi^{abc\ldots}$ has 2^n complex degrees of freedom and can be viewed as a $2^n\times 1$ complex column matrix ψ^Δ, where Δ is the compounded index $\Delta=[abc\ldots]$.

Following Holland,[8] a tensor can be constructed from a pair of spinors $\psi^\Delta \equiv \psi^{abc\ldots}$ and $\xi^{efg\cdots}\equiv \xi^\Theta$ of the same rank (2^n) by first constructing the $2^n\times 2^n$ complex matrix

$$A^\Delta{}_\Theta = A^{abc\cdots}{}_{efg\ldots} \equiv \psi^{abc\cdots}(\xi_{efg\ldots})^T = \psi^\Delta(\xi_\Theta)^T, \tag{3.1}$$

where the superscript T denotes matrix transposition. Spinor indices are raised and lowered by the spin metric[13]

$$\epsilon_{ab} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = j\hat{\sigma}_2 . \tag{3.2}$$

(Conventionally, raised or lowered indices on an object indicate different coordinate representations of the same object. Here however, its significance is to indicate a distinct object which carries the same information as the original but in a different form.)

Any $2^n\times 2^n$ (complex) matrix may be expanded in terms of a set of independent basis matrices $\{(\hat{e}_{klm\ldots})^\Delta{}_\Theta\}$ for the direct product of n Pauli algebras $\mathcal{P}^n$,

$$A^\Delta{}_\Theta = \sum_{k,l,m,\ldots} c_{klm\ldots}(\hat{e}_{klm\ldots})^\Delta{}_\Theta . \tag{3.3}$$

The basis matrices $\{(\hat{e}_{klm\ldots})^\Delta{}_\Theta\}$ carry both spinor indices (Δ and Θ) and tensor indices ($\{k,l,m,\ldots\}$) which they inherit from the Pauli matrices. The complex expansion coefficients $c_{klm\ldots}$ are spatial tensors and may be determined from A via

$$c_{klm\ldots} = \frac{1}{2^n}\mathrm{Tr}[\psi^\Delta \xi_\Theta(\hat{e}_{klm\ldots})^\Theta{}_\Omega] = \frac{1}{2^n}\xi_\Theta(\hat{e}_{klm\ldots})^\Theta{}_\Delta \psi^\Delta, \tag{3.4}$$

where Tr denotes the matrix trace.

If ξ_Θ is obtained directly from ψ^Δ then $c_{klm\ldots}$ will characterize some of the information in ψ^Δ. This is the basis of Holland's approach. Note that, since A satisfies the relation

$$A^2 = A^\Delta{}_\Theta A^\Theta{}_\Omega = \psi^\Delta(\xi_\Theta)^T \psi^\Theta(\xi_\Omega)^T = A^\Theta{}_\Theta A^\Delta{}_\Omega = \mathrm{Tr}(A)A, \tag{3.5}$$

A therefore belongs to an ideal of the algebra $\mathcal{P}^n$. It is via this ''ideal'' characterization of ψ^Δ that Holland associates the tensors $c_{klm\ldots}$ with spinor degrees of freedom. We now study how geometric algebra both simplifies this scheme, and reveals much of the hidden geometry relating spinors to tensors. We start with an analysis of the one-particle setup.

IV. THE ONE-PARTICLE CASE

In the one-particle setup we expand the 2×2 complex matrix $A^a{}_b$ in terms of the Pauli matrices:

$$A^a{}_b \equiv \psi^a(\xi_b)^T \equiv s + u_k\hat{\sigma}_k = \Re(s) + j\Im(s) + [\Re(u_k) + j\Im(u_k)]\hat{\sigma}_k. \tag{4.1}$$

We can construct a MSTA version of this by first writing out $A^a{}_b$ explicitly in terms of the spinor components

$$\psi^a = \begin{pmatrix} \psi^1 \\ \psi^2 \end{pmatrix}, \quad \xi^a = \begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix}. \tag{4.2}$$

The matrix $A^a{}_b$ then has components

$$A^a{}_b = \begin{pmatrix} \psi^1\xi^2 & -\psi^1\xi^1 \\ \psi^2\xi^2 & -\psi^2\xi^1 \end{pmatrix} = \begin{pmatrix} \psi^1 & 0 \\ \psi^2 & 0 \end{pmatrix} \begin{pmatrix} \xi^2 & 0 \\ -\xi^1 & 0 \end{pmatrix}^T. \tag{4.3}$$

By writing this as the product of two matrices, we can easily establish an equivalent expression within the one-particle geometric algebra of space. First, we note the following equivalence:

$$\begin{pmatrix} \psi^1 & 0 \\ \psi^2 & 0 \end{pmatrix} \leftrightarrow \psi\tfrac{1}{2}(1+\sigma_3), \tag{4.4}$$

where ψ is the Pauli-even multivector formed from ψ^a according to Eq. (2.10). Second, we need the analog of matrix transposition for multivectors. It is easily confirmed that this is performed by

$$M^T = \sigma_2 \tilde{M} \sigma_2. \tag{4.5}$$

If we now denote the multivector equivalent of $A^a{}_b$ by A, we find that

$$A = \psi\tfrac{1}{2}(1+\sigma_3)[i\sigma_2\xi\tfrac{1}{2}(1+\sigma_3)]^T = \psi\tfrac{1}{2}(1+\sigma_3)\sigma_2\tfrac{1}{2}(1-\sigma_3)\tilde{\xi}(-i\sigma_2)\sigma_2 = -\psi\tfrac{1}{2}(\sigma_1 + i\sigma_2)\tilde{\xi}, \tag{4.6}$$

which immediately gives all of the components s and $\{u_k\}$ by simply reading off the terms of different grades. Each grade returns a genuine geometric object, since under the rotation $\psi \mapsto R\psi$, $\xi \mapsto R\xi$, we find that A transforms as

$$A \mapsto RA\tilde{R}, \tag{4.7}$$

which is the correct transformation law for geometric objects. The same approach extends easily to the case where the rotor R includes Lorentz transformations. This is not such a surprise when one considers that the algebraic manipulations described here closely resemble those of the 2-spinor calculus,[13,14] which are designed to be fully relativistic. We can now give the following explicit formulas for the s and $\{u_k\}$:

$$\Re(s) = -\tfrac{1}{2}\langle \psi i \sigma_2 \tilde{\xi}\rangle, \quad \Im(s) = \tfrac{1}{2}\langle \psi i \sigma_1 \tilde{\xi}\rangle,$$
$$\Re(u_k) = -\tfrac{1}{2}\langle \psi \sigma_1 \tilde{\xi} \sigma_k\rangle, \quad \Im(u_k) = -\tfrac{1}{2}\langle \psi \sigma_2 \tilde{\xi}\sigma_k\rangle, \tag{4.8}$$

where $\langle M\rangle$ denotes the result of projecting out the scalar part of the multivector M.

Having established the correspondence between the matrix formalism and geometric algebra, it is now straightforward to consider the choices Holland makes for ξ_b and the resulting tensors obtained.[8] The first choice is

$$\xi_b = \psi^{a*} \tag{4.9}$$

$$\Rightarrow \xi^a = \epsilon^{ab}\xi_b = -\epsilon_{ab}\psi^{a*}. \tag{4.10}$$

The multivector analog of complex conjugation is defined by

$$\psi^* = \sigma_2 \psi \sigma_2, \tag{4.11}$$

so this choice corresponds to setting

$$\xi = (-i\sigma_2)\sigma_2\psi\sigma_2 = -\psi i \sigma_2. \tag{4.12}$$

It follows that A is given by

$$A = \psi \tfrac{1}{2}(1+\sigma_3)\tilde{\psi} = \tfrac{1}{2}\psi\tilde{\psi} + \tfrac{1}{2}\psi\sigma_3\tilde{\psi}, \tag{4.13}$$

and from (4.8) we find

$$\rho \equiv 2s = \psi\tilde{\psi}, \tag{4.14}$$

$$S \equiv 2u_k\sigma_k = \psi\sigma_3\tilde{\psi}, \tag{4.15}$$

where ρ and S are the symbols used by Holland. As expected, we have isolated the scalar density $\psi\tilde{\psi}$ and the spin vector $\psi\sigma_3\tilde{\psi}$. On decomposing ψ in the form

$$\psi = \rho^{1/2} R = \rho^{1/2} e^{i\sigma_3\theta/2} e^{i\sigma_1\varphi/2} e^{i\sigma_3\chi/2}, \tag{4.16}$$

where we have written the rotor R in terms of the Euler angles, we see that the scalar ρ and the vector

$$S = \rho[\sin(\theta)\cos(\varphi)\sigma_1 + \sin(\theta)\sin(\varphi)\sigma_2 + \cos(\theta)\sigma_3] \tag{4.17}$$

are independent of the phase χ. This pair of tensors therefore only embodies three of the four degrees of freedom in ψ and consequently an additional characterization of ψ is required. From our geometric viewpoint it is clear that the only other tensors that can be obtained from ψ are the vectors $\psi\sigma_1\tilde{\psi}$ and $\psi\sigma_2\tilde{\psi}$, or their corresponding duals. To verify this, consider Holland's other choice:

$$\xi_b = \psi_b \Rightarrow \xi = \psi. \tag{4.18}$$

On substituting this into (4.13) we find that

$$A = -\psi\tfrac{1}{2}(1+\sigma_3) i \sigma_2 \tilde{\psi} = -\tfrac{1}{2}(\psi i \sigma_2 \tilde{\psi} + \psi \sigma_1 \tilde{\psi}). \tag{4.19}$$

Clearly for this choice of ξ, $s=0$ and from (4.8)

$$M \equiv 2\Re(\mathbf{u}) = -\psi\sigma_1\tilde{\psi}, \quad N \equiv 2\Im(\mathbf{u}) = -\psi\sigma_2\tilde{\psi}, \tag{4.20}$$

where M and N are Holland's Cartan–Kramers vectors.

The mutual geometric relations of M, N, and S are transparent from the geometric point of view, since the set is obtained by a rotation and dilation of the orthonormal basis $\{\sigma_1, \sigma_2, \sigma_3\}$. It follows immediately that

$$\rho = |M| = |N| = |S| \tag{4.21}$$

and that M, N, and S are mutually orthogonal. Since the phase factor $\exp\{i\sigma_3\chi/2\}$ does not commute with either of σ_1 or σ_2, the pair $\{M,N\}$ contain information about all four degrees of freedom in ψ. Indeed, any pair from the set $\{S,M,N\}$ contains information about all four degrees of freedom in ψ, and can be used to reconstruct ψ (and hence ψ^a, if required) up to an arbitrary sign. The simple manner in which the triad $\{M,N,S\}$ is formed and understood in the geometric algebra approach fully demonstrates its advantages.

V. THE TWO-PARTICLE CASE

Two-particle states are conventionally represented by rank-2 spinors ψ^{ab}. From these we construct the matrix

$$A^{ab}{}_{cd} = \psi^{ab}(\xi_{cd})^T, \tag{5.1}$$

which is a 4×4 complex matrix, where the transpose is understood to be with respect to the compounded index $[cd]$. This matrix can be expanded in the 16-dimensional basis

$$\begin{aligned} \mathbf{1}^{ab}_{cd} &= \delta^a{}_c \delta^b{}_d, \quad (\hat{e}_{1k})^{ab}{}_{cd} = \hat{\sigma}_k{}^a{}_c \delta^b{}_d \\ (\hat{e}_{2l})^{ab}{}_{cd} &= \delta^a{}_c \hat{\sigma}_l{}^b{}_d, \quad (\hat{e}_{1k}\hat{e}_{2l})^{ab}{}_{cd} = \hat{\sigma}_k{}^a{}_c \hat{\sigma}_l{}^b{}_d \end{aligned} \tag{5.2}$$

belonging to $\mathcal{P}^2$. With a suppression of spinor indices, we can write (following Ref. 8)

$$A = s\mathbf{1} + c_k\hat{e}_{1k} + d_k\hat{e}_{2k} + f_{kl}\hat{e}_{1k}\hat{e}_{2l}. \tag{5.3}$$

These coefficients are not independent because of the relation $A^2 = \mathrm{Tr}(A)A$.

A complete basis for two-particle spin states, together with their MSTA analogs, is provided by

$$\begin{aligned} \begin{pmatrix}1\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\end{pmatrix} &\leftrightarrow E, \\ \begin{pmatrix}0\\1\end{pmatrix} \otimes \begin{pmatrix}1\\0\end{pmatrix} &\leftrightarrow -i\sigma_2^1 E, \\ \begin{pmatrix}1\\0\end{pmatrix} \otimes \begin{pmatrix}0\\1\end{pmatrix} &\leftrightarrow -i\sigma_2^2 E, \\ \begin{pmatrix}0\\1\end{pmatrix} \otimes \begin{pmatrix}0\\1\end{pmatrix} &\leftrightarrow i\sigma_2^1 i\sigma_2^2 E, \end{aligned} \tag{5.4}$$

where $E = \frac{1}{2}(1 - i\sigma_3^1 i\sigma_3^2)$ is the two-particle correlator. A rank-2 spinor ψ^{ab} can be expanded in terms of this basis and hence mapped directly to the MSTA element $\psi = \psi E \in (\mathcal{G}_3^+)^2$, where

$$\psi = (\phi_1 - i\sigma_2^1\phi_2 - i\sigma_2^2\phi_3 + i\sigma_2^1 i\sigma_2^2\phi_4)E \tag{5.5}$$

and ϕ_n, $n=1,...,4$ are the combinations of 1 and J appropriate to the expansion of ψ^{ab}. All MSTA spinors $\psi \in (\mathcal{G}_3^+)^2$ contain an implicit factor of E. This is only shown explicitly in cases where its presence increases clarity.

We can again construct the matrix $A^{ab}{}_{cd}$ by first introducing the MSTA analog of a 4×4 complex matrix in which ψ^{ab} is the first and only nonzero column. The multivector equivalent of this is

$$\psi\tfrac{1}{2}(1+\sigma_3^1)\tfrac{1}{2}(1+\sigma_3^2)E=\psi\tfrac{1}{2}(1+\sigma_3^1)\tfrac{1}{2}(1+\sigma_3^1)\tfrac{1}{2}(1-i^1i^2). \tag{5.6}$$

The analog of the matrix transpose operation is now

$$M^T=\sigma_2^1\sigma_2^2\tilde{M}\sigma_2^1\sigma_2^2 \tag{5.7}$$

and we find that the information contained in $A^{ab}{}_{cd}$ can be encoded in the multivector

$$A=\psi\tfrac{1}{2}(\sigma_1^1+i\sigma_2^1)\tfrac{1}{2}(\sigma_1^2+i\sigma_2^2)E\tilde{\xi}=\psi\tfrac{1}{2}(\sigma_1^1+i\sigma_2^1)\tfrac{1}{2}(\sigma_1^2+i\sigma_2^2)\tilde{\xi}(1-i^1i^2). \tag{5.8}$$

The fact that A satisfies $A=A\tfrac{1}{2}(1-i^1i^2)$ is to be expected. The tensor product $\mathcal{G}_3\otimes\mathcal{G}_3$ defines a space of 64 real dimensions, whereas the product $\mathcal{P}\otimes\mathcal{P}$ defines a 16-dimensional complex space, with 32 real dimensions. The space $\mathcal{P}\otimes\mathcal{P}$ therefore fails to provide a matrix representation of the full algebra $\mathcal{G}_3\otimes\mathcal{G}_3$, the reason being that the two pseudoscalars i^1 and i^2 are given the same representation in terms of the unit imaginary j. The only multivectors in $\mathcal{G}_3\otimes\mathcal{G}_3$ which do correspond directly to matrices in $\mathcal{P}\otimes\mathcal{P}$ are therefore those which contain a factor of the idempotent $\tfrac{1}{2}(1-i^1i^2)$, which links the pseudoscalars together.

The quantities of interest are the complex tensor coefficients s, c_k, d_l, and f_{kl} in (5.3). These can all be recovered by taking appropriate traces of the form

$$\tfrac{1}{4}\mathrm{Tr}(A^{ab}{}_{cd}\Gamma^{cd}{}_{ef})=\tfrac{1}{4}\xi_{cd}\Gamma^{cd}{}_{ab}\psi^{ab}, \tag{5.9}$$

where Γ is some combination of the basis elements (5.2). From the above scheme, a general matrix of the form $A^{ab}{}_{cd}\Gamma^{cd}{}_{ef}$ will have a multivector equivalent $M=M\tfrac{1}{2}(1-i^1i^2)$. With this multivector we have the explicit relations

$$\begin{aligned}\Re(\tfrac{1}{4}\xi_{cd}\Gamma^{cd}{}_{ab}\psi^{ab})&=2\langle M\rangle,\\ \Im(\tfrac{1}{4}\xi_{cd}\Gamma^{cd}{}_{ab}\psi^{ab})&=-2\langle Mi^1\rangle.\end{aligned} \tag{5.10}$$

The extra factor of 2 is required because of the presence of the factor of $\tfrac{1}{2}(1-i^1i^2)$ in M.

We now have analogs for most of the operations performed on rank-2 spinors. The one remaining operation is that of particle interchange

$$\psi^{ab}\mapsto\psi^{ba}. \tag{5.11}$$

The algebraic effect of this on the MSTA spinor ψ is

$$\psi\mapsto\psi^I=(\phi_1-i\sigma_2^2\phi_2-i\sigma_2^1\phi_3+i\sigma_2^1i\sigma_2^2\phi_4)E. \tag{5.12}$$

This operation has no one-particle analog. As an algebraic operation it can be expressed as

$$\psi\to\psi^I=E\psi E-i\sigma_2^1i\sigma_2^2\bar{E}\psi E=\tfrac{1}{2}(1-i\sigma_k^1i\sigma_k^2)\psi, \tag{5.13}$$

where $\bar{E}\equiv\tfrac{1}{2}(1+i\sigma_3^1i\sigma_3^2)$. If we recall the definition of the MSTA rotation singlet state from Sec. II,

$$\epsilon \equiv \frac{1}{\sqrt{2}}(i\sigma_2^1 - i\sigma_2^2)E, \tag{5.14}$$

we find that

$$\epsilon\tilde{\epsilon} = \tfrac{1}{2}(1 + i\sigma_k^1 i\sigma_k^2). \tag{5.15}$$

It follows that

$$\psi^I = (1 - \epsilon\tilde{\epsilon})\psi, \tag{5.16}$$

confirming that the antisymmetrized state $\psi - \psi^I = \epsilon\tilde{\epsilon}\psi$ is a rotation singlet. Equation (2.23) ensures that the interchange operation is rotationally covariant,

$$R^1R^2(1-\epsilon\tilde{\epsilon})\psi = R^1R^2\psi - \epsilon\tilde{\epsilon}\psi = (1-\epsilon\tilde{\epsilon})R^1R^2\psi, \tag{5.17}$$

as expected.

We are now in a position to give explicit MSTA formulas for the two-particle tensors constructed in Ref. 8. With ψ the direct map of ψ^{ab}, according to Eq. (5.5), we find that

$$\begin{aligned} \rho &\equiv \psi^{ab}\psi^{*ab} = 2\langle \psi\tilde{\psi}\rangle, \\ S_{1k} &\equiv \psi^{*ab} e_{1k}{}^{ab}{}_{cd}\psi^{cd} = -2(\psi J\tilde{\psi})\cdot(i\sigma_k^1), \\ S_{2k} &\equiv \psi^{*ab} e_{2k}{}^{ab}{}_{cd}\psi^{cd} = -2(\psi J\tilde{\psi})\cdot(i\sigma_k^2), \\ S_{kl} &\equiv \psi^{*ab} e_{1k}{}^{ab}{}_{cd} e_{2k}{}^{cd}{}_{ef}\psi^{ef} = -2(\psi\tilde{\psi})\cdot(i\sigma_k^1 i\sigma_l^2). \end{aligned} \tag{5.18}$$

The only terms involved in the MSTA approach are the scalar + four-vector quantities $\psi\tilde{\psi} = \psi E\tilde{\psi}$ and the bivector $\psi J\tilde{\psi}$. This information is summarized in the single multivector

$$A = \psi\tfrac{1}{2}(1+\sigma_3^1)\tfrac{1}{2}(1+\sigma_3^2)\tilde{\psi}, \tag{5.19}$$

as expected.

Holland interprets S_{kl} as a spin correlation tensor, an observation justified by that fact that S_{kl} encodes the four-vector component of the ''expectation,'' $\psi E\tilde{\psi}$, of the correlator E in the MSTA formulation. The quantity $\psi J\tilde{\psi}$ is a bivector on account of its even grade and reversion asymmetry. Since we work in the closed algebra $(\mathcal{G}_3^+)^2$, it can only have $i\sigma_k^1$ and $i\sigma_k^2$ parts. We can therefore write

$$\psi J\tilde{\psi} = \tfrac{1}{2}(S_1^1 + S_2^2), \tag{5.20}$$

where S_1^1 and S_2^2 are one-particle bivectors. The subscripts indicate that S_1 and S_2 are separate variables, while the superscripts denote the particle spaces these quantities are expressed in. The quantity $\psi J\tilde{\psi}$ is the two-particle spin bivector whose one-particle projections S_1 and S_2 are the spin bivectors of particle-1 and particle-2, respectively.[2]

One surprising result proved in Ref. 8 is that the spin bivectors S_1 and S_2 have the same magnitude, $|S_1| = |S_2|$. This result is obviously true for direct-product states, but it is not intuitively obvious why it should hold for general superpositions. It is therefore instructive to see how to prove the result in the MSTA. We start by noting that the components of S_1 are given by

$$-\tfrac{1}{2}S_{1k} = \tfrac{1}{2}S_1^1\cdot(i\sigma_k^1) = (\psi J\tilde{\psi})\cdot(i\sigma_k^1) = (\tilde{\psi} i\sigma_k^1\psi)\cdot(i\sigma_3^1). \tag{5.21}$$

The implicit idempotent E on either side of the bivector $\tilde{\psi} i\sigma_k^1 \psi = E\tilde{\psi} i\sigma_k^1 \psi E$, acts as a projection and implies that $\tilde{\psi} i\sigma_k^1 \psi$ can only contain an equal linear combination of $i\sigma_3^1$ and $i\sigma_3^2$. We can therefore write

$$\tilde{\psi} i\sigma_k^1 \psi = S_{1k} J, \quad \tilde{\psi} i\sigma_k^2 \psi = S_{2k} J. \tag{5.22}$$

From this we deduce that the magnitudes of S_1 and S_2 are given by

$$|S_a|^2 = S_{ak} S_{ak} = -2\langle S_{ak} J S_{ak} J\rangle = -2\langle \tilde{\psi} i\sigma_k^a \psi \tilde{\psi} i\sigma_k^a \psi\rangle, \tag{5.23}$$

where $a = 1,2$ and no summation over a is implied. Since $\psi\tilde{\psi}$ is even and reversion symmetric, it contains only a scalar and four-vector part. The four-vector part necessarily has the form $B^1 C^2$, where B^1 and C^2 are bivectors in the separate one-particle spaces. Employing the one-particle identity

$$i\sigma_k^1 B^1 i\sigma_k^1 = B^1, \tag{5.24}$$

we see that (with no sum over a)

$$i\sigma_k^a \psi\tilde{\psi} i\sigma_k^a = i\sigma_k^a (\langle \psi\tilde{\psi}\rangle + \langle \psi\tilde{\psi}\rangle_4) i\sigma_k^a = -3\langle \psi\tilde{\psi}\rangle + \langle \psi\tilde{\psi}\rangle_4 = \psi\tilde{\psi} - 4\langle \psi\tilde{\psi}\rangle. \tag{5.25}$$

This result is independent of the label a, and upon substitution into (5.23) implies that $|S_1|^2 = |S_2|^2$. Specifically,

$$|S_a|^2 = -2\langle \tilde{\psi}(\psi\tilde{\psi} - 4\langle \psi\tilde{\psi}\rangle)\psi\rangle = -2\langle \tilde{\psi}\psi\tilde{\psi}\psi\rangle + 2\rho^2. \tag{5.26}$$

This result does not generalize to higher particle numbers.

In Ref. 8 the quantity Ω is defined as

$$|S_1|^2 = |S_2|^2 = 2\Omega - \rho^2, \tag{5.27}$$

and it follows from Eq. (5.26) that we can write

$$\Omega = \tfrac{3}{2}\rho^2 - \langle \tilde{\psi}\psi\tilde{\psi}\psi\rangle. \tag{5.28}$$

Since the elements S_{kl} are the components of $-2\langle \psi\tilde{\psi}\rangle_4$, we can substitute

$$\psi\tilde{\psi} = \tfrac{1}{2}(\rho - S_{kl} i\sigma_k^1 i\sigma_l^2) \tag{5.29}$$

into the preceding expression for Ω to obtain

$$\Omega = \tfrac{1}{4}(5\rho^2 - S_{kl} S_{kl}), \tag{5.30}$$

recovering Ω in terms of previously defined quantities. This result is not so easily deduced in the spin-tensor approach.

A second choice of ξ_{cd} considered in Ref. 8 is

$$\xi_{cd} = \psi^{*dc}. \tag{5.31}$$

The equivalent MSTA spinor is

$$\xi = (1 - \epsilon\tilde{\epsilon})\psi i\sigma_2^1 i\sigma_2^2, \tag{5.32}$$

leading to the new multivector

$$A' = \psi \tfrac{1}{2}(1 + \sigma_3^1)\tfrac{1}{2}(1 + \sigma_3^2)\tilde{\psi}(1 - \epsilon\tilde{\epsilon}) = A(1 - \epsilon\tilde{\epsilon}), \tag{5.33}$$

where A is as defined in Eq. (5.19). This expression makes it immediately clear that A' contains precisely the same information as A. This conclusion is much harder to reach in the tensor approach, where one is forced to consider each of the terms in the matrix $\psi^{ab}\psi^{*dc}$ and show that they can be written in terms of the $\psi^{ab}\psi^{cd}$.

It is clear from the expressions $\psi\tilde{\psi}$ and $\psi J\tilde{\psi}$ that $\{\rho, S_{1k}, S_{2l}, S_{kl}\}$ are invariant under an overall phase change of ψ. Consequently, this set only encodes seven of the eight degrees of freedom in ψ and some other form for ξ must be used to recover all of the information in ψ. The MSTA approach enables us to immediately write down two further geometric entities which pick up the phase information. These are

$$\tfrac{1}{2}W \equiv \psi i\sigma_2^1 i\sigma_2^2 \tilde{\psi}, \quad \tfrac{1}{2}V \equiv \psi J i\sigma_2^1 i\sigma_2^2 \tilde{\psi}. \tag{5.34}$$

With these we can proceed to give MSTA equivalents of the remaining tensors defined in Ref. 8. The first set are formed by taking $\xi_{ab} = \psi_{ab}$, yielding the two new quantities

$$\overline{\rho} \equiv \psi_{ab}\psi^{ab} = 2\langle i\sigma_2^1 i\sigma_2^2 \tilde{\psi}\psi\rangle - 2\langle i\sigma_2^1 i\sigma_2^2 \tilde{\psi}\psi J\rangle j = \langle W\rangle - j\langle V\rangle \tag{5.35}$$

and

$$\begin{aligned} T_{kl} \equiv \psi_{ab} e_{1k}{}^{ab}{}_{cd} e_{2k}{}^{cd}{}_{ef} \psi^{ef} &= -2\langle i\sigma_2^1 i\sigma_2^2 \tilde{\psi} i\sigma_k^1 i\sigma_l^2 \psi\rangle + 2j\langle i\sigma_2^1 i\sigma_2^2 \tilde{\psi} i\sigma_k^1 i\sigma_l^2 \psi J\rangle \\ &= -W\cdot(i\sigma_k^1 i\sigma_l^2) + jV\cdot(i\sigma_k^1 i\sigma_l^2). \end{aligned} \tag{5.36}$$

As expected, V and W are the only quantities necessary for the evaluation of these coefficients. Since $i\sigma_2^1 i\sigma_2^2$ and $i\sigma_1^1 i\sigma_2^2$ anticommute with J, the phase transformation $\psi \to \psi e^{\theta J}$ implies that

$$W \to \cos(2\theta)W + \sin(2\theta)V, \quad V \to \cos(2\theta)V - \sin(2\theta)W. \tag{5.37}$$

By taking scalar parts of these transformations we can deduce the behavior of $\overline{\rho}$ under phase changes. Similarly, taking the four-vector parts gives us the phase transformation properties for the real and imaginary parts of T_{kl}.

The remaining quantities defined in Ref. 8 are obtained from the choice $\xi_{ab} = \psi_{ba}$. Again, the MSTA equivalent for this choice,

$$\xi = (1 - \epsilon\tilde{\epsilon})\psi \tag{5.38}$$

makes it clear that this choice yields nothing new, and that all of the tensor coefficients derived by setting $\xi_{ab} = \psi_{ba}$ can be recovered from V and W.

One remaining MSTA construct is the scalar + four-vector quantity $\tilde{\psi}\psi$. By its construction this quantity is automatically invariant under rotations. The scalar term is just the density ρ already defined. The four-vector invariant is more interesting since it picks up phase information. In terms of the decomposition (5.5) we have explicitly

$$\tilde{\psi}\psi = [\rho + 2i\sigma_2^1 i\sigma_2^2(\phi_1\phi_4 - \phi_2\phi_3)]E, \tag{5.39}$$

which demonstrates that it is the complex quantity $\phi_1\phi_4 - \phi_2\phi_3$ which picks up the phase information. The same information is encoded in the trace T_{kk}, so $\tilde{\psi}\psi$ yields no new information.

In summary, we see that all of the information required to completely encode ψ in tensor form is contained in the set of multivectors

$$\{\psi\tilde{\psi},\ \psi J\tilde{\psi},\ \psi i\sigma_2^1 i\sigma_2^2 \tilde{\psi},\ \psi i\sigma_2^1 i\sigma_2^2 J\tilde{\psi}\}. \tag{5.40}$$

This is the complete set of distinct objects obtainable by taking any basis element Γ of the direct product of two Pauli algebras and forming the bilinear construct $\psi\Gamma\tilde{\psi}$. The other objects one might try to construct would be of the form $\psi i\sigma_1^1\tilde{\psi}$, but the presence of the idempotent E ensures that all quantities of this form vanish. (If one looks to construct models beyond those suggested by quantum theory, however, one can contemplate not correlating the phases of the particles, in which case such quantities would come into consideration.)

The set (5.40) is directly analogous to the one-particle set $\{\rho, iM, iN, iS\}$ which can be written as

$$\{\psi\tilde{\psi},\ \ \psi J_1\tilde{\psi},\ \ \psi i\sigma_2\tilde{\psi},\ \ \psi i\sigma_2 J_1\tilde{\psi}\}. \tag{5.41}$$

We are now in a position to appreciate just how systematic and simple the MSTA approach is. The full set of distinct objects (5.40) could have been written down easily at the start of the analysis, and all terms calculated simply with the MSTA, without requiring laborious matrix and tensor manipulations. Furthermore, the MSTA approach is very amenable to generalization to higher dimensions, as discussed in Sec. VI.

VI. EXTENSIONS AND FURTHER WORK

As in the one-particle and two-particle cases, the n-particle spinor $\psi^{abc\cdots}$ can be mapped directly to an element in the direct product of n Pauli-even algebras $(\mathcal{G}_3^+)^n$. The ambiguity in the complex structure for each of the factors in $(\mathcal{G}_3^+)^n$ requires the introduction of the n-particle correlator E_n defined at Eq. (2.16). We find that the spinor ψ can be written as a combination of 2^n terms:

$$\psi = \left(\phi + \sum_{a=1}^{n} i\sigma_2^a \phi_a + \sum_{a<b} i\sigma_2^a i\sigma_2^b \phi_{ab} + \cdots + i\sigma_2^1 i\sigma_2^2 \cdots i\sigma_2^n \phi_{12\ldots n} \right) E_n, \tag{6.1}$$

where the ϕ are complex combinations of 1 and J_n. The tensor observables formed from two n-particle spinors are summarized in the multivector

$$A = \psi\left(\tfrac{1}{2}(\sigma_1^1 + i\sigma_2^1)\cdots\tfrac{1}{2}(\sigma_1^n + i\sigma_2^n)\right)\tilde{\xi}, \tag{6.2}$$

and it is clear that the various tensors one might construct correspond to the various multivector parts of bilinear constructs of the form $\psi\Gamma\tilde{\psi}$, where Γ is some fixed set of MSTA basis elements.

For example, one of the key objects to analyze is the multivector $\psi J\tilde{\psi}$. This has grade-2, grade-6,..., components, of which the grade-2 component is the multiparticle spin bivector[2]

$$S \equiv 2^{n-1}\langle \psi J_n \tilde{\psi}\rangle_2 . \tag{6.3}$$

Interpretations for the other components of $\psi J\tilde{\psi}$ can be made in terms of spin correlations between particles. This approach to constructing tensors from spinors is clearly more economic than the matrix/tensor approach, which gets progressively worse with increasing particle number due to the large degree of redundancy in the tensor coefficients contained in the various bilinear constructs.

A further advantage of the MSTA approach is that it is easily generalized to the relativistic domain. A discussion of this is contained in Ref. 2 and further details are contained in Ref. 6. Such an extension is essential if this approach is to shed light on questions of nonlocality in Einstein–Podolski–Rosen- (EPR)-type experiments—in particular a full analysis of these must incorporate relativity, as this lies at the heart of the paradox. A simple model for two-particle relativistic spin correlations is contained in Ref. 2, though more work is needed to extend this work to model an EPR-type setup. Finally, it should be borne in mind that the MSTA is equally applicable to classical as well as quantum physics, and many of the techniques described here are useful for studying multiparticle classical relativistic dynamics, a notoriously difficult subject.

ACKNOWLEDGMENTS

S.S. is grateful to the Cambridge Livingstone Trust for funding during the course of this work. C.D. is funded by a grant from the Lloyd's of London Tercentenary Foundation.

[1] S. F. Gull, A. N. Lasenby, and C. J. L. Doran, ''Imaginary numbers are not real—the geometric algebra of spacetime,'' Found. Phys. **23**, 1175 (1993).
[2] C. J. L. Doran, A. N. Lasenby, S. F. Gull, S. S. Somaroo, and A. D. Challinor, ''Spacetime algebra and electron physics,'' Adv. Imaging Electron Phys. **95**, 271 (1996).
[3] D. Hestenes, *Space–Time Algebra* (Gordon and Breach, New York, 1966).
[4] D. Hestenes, ''Proper dynamics of a rigid point particle,'' J. Math. Phys. **15**, 1778 (1974).
[5] A. N. Lasenby, C. J. L. Doran, and S. F. Gull, ''Gravity, gauge theories and geometric algebra,'' Philos. Trans. R. Soc. London, Ser. A **356**, 487 (1998).
[6] C. J. L. Doran, ''Geometric algebra and its application to mathematical physics,'' Ph.D. thesis, Cambridge University, 1994.
[7] P. R. Holland, *The Quantum Theory of Motion* (Cambridge University Press, Cambridge, 1993).
[8] P. R. Holland, ''Causal interpretation of a system of two spin-1/2 particles,'' Phys. Rep. **169**, 294 (1988).
[9] C. J. L. Doran, A. N. Lasenby, S. F. Gull, and J. Lasenby, ''Lectures in geometric algebra,'' in *Clifford (Geometric) Algebras*, edited by W. E. Baylis (Birkhauser, Boston, 1996), pp. 65–236.
[10] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus* (Reidel, Dordrecht, 1984).
[11] I. W. Benn and R. W. Tucker, *An Introduction to Spinors and Geometry* (Hilger, Bristol, 1988).
[12] C. J. L. Doran, A. N. Lasenby, and S. F. Gull, ''States and operators in the spacetime algebra,'' Found. Phys. **23**, 1239 (1993).
[13] R. Penrose and W. Rindler, *Spinors and Space–time Volume 1: Two-spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge, 1984).
[14] A. N. Lasenby, C. J. L. Doran, and S. F. Gull, ''2-spinors, twistors and supersymmetry in the spacetime algebra,'' in *Spinors, Twistors, Clifford Algebras and Quantum Deformations*, edited by Z. Oziewicz, B. Jancewicz, and A. Borowiec (Kluwer, Dordrecht, 1993), p. 233.

A differential-geometric interpretation of Kirchhoff's elastic rods

Kai Hu[a)]
Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095-1555

(Received 28 July 1998; accepted for publication 2 April 1999)

In this article, Kirchhoff's elasticity theory of rods is revisited but from a viewpoint of Riemannian geometry. By means of the Cayley–Klein parameter, the theory under clamped-end conditions can be regarded as a geometry of paths on some geometrically distorted three-sphere produced by a constrained elastic energy functional. Using this geometric formulation, the uniqueness of elasticae with prescribed initial values of the strains can be easily shown. On the other hand, a family of elasticae with prescribed values of orthonormal frames at two endpoints is demonstrated to be parametrized by an open set in $\mathbb{R}^3$. In particular, a criterion of the nonuniqueness of elasticae satisfying clamped-end conditions is given in terms of a geometric concept—conjugate points. © *1999 American Institute of Physics.* [S0022-2488(99)01307-9]

I. INTRODUCTION

Over the last two decades, Kirchhoff's elasticity theory has been extensively applied to study shapes of two-stranded polymers, like DNA (suggested by Fuller[1]). Also, the theory provides energetic explanations of supercoiling of DNA which is important for DNA functions like recombination, DNA being a torus knot observed under electron microscopy.[2] Recently, an interesting area from molecular biology deals with the study of a special case of Kirchhoff's theory of which the stress-free state is not a (twisted) straight line anymore.[3–8] Some relaxed DNA observed in experiments form a closed circle, such as SV40.[4,5] Thus, intrinsically curved elasticity is needed to model circular DNA. Indeed, theoretically, one cannot simply ignore, for instance, base stacking and hydrogen bonding energies existing in DNA. Therefore, in this article we turn our attention to general elasticity theory, particularly with an emphasis on geometric methods.

An elastic rod is *inextensible* (resp. *unshearable*) if the length of its axis (resp. the normal cross sections being normal to its axis) is unaltered while undergoing deformations. These are reasonable assumptions when we study conformations of DNA. Mathematically, an elastic rod can be simplified to a *framed curve* (C,ν) consisting of an immersed, unit-speed space curve C, called axis, and a preferred unit normal vector field ν defined along C, called material direction. Therefore, it is easy to see that an unshearable elastic rod under deformation can be thought of as a framed curve being perturbed within the family of framed curves. How we take inextensibility into account will be demonstrated in Sec. III.

According to Kirchhoff's theory, an *elastica* is a critical point of the so-called *elastic energy* functional,[7]

$$\mathcal{E}=\frac{1}{2}\int_0^l \rho_1(\omega_1-\kappa_1)^2+\rho_2(\omega_2-\kappa_2)^2+\rho_3(\omega_3-\kappa_3)^2\,ds, \tag{1}$$

where ρ_1 and ρ_2 are both called *bending stiffnesses* and ρ_3 is called *twisting stiffness*. All ρ_i's are assumed positive constants; ω_i's are called *strains* and ω_3 is specially called *twisting density*

[a)]Electronic mail: khu@math.ucla.edu

(I will explain how the strains of an elastic rod can be computed). Here κ_1, κ_2 are called *intrinsic curvatures*, and κ_3 is called *intrinsic twisting density*. In fact, κ_i's are the ω_i's of the stress-free state (and this state is also called relaxed state in molecular biology). Also, s is the arc-length parameter of the axis of an elastic rod and l is its length. The sum of the first two terms of (1) is called *bending energy* and the remaining term alone is called *twisting energy*. An elastic rod is called *isotropic* if $\rho_1 = \rho_2$. It is obvious that, for example, if the largest stiffness is ρ_3, then torsional fluctuations are less energetically favorable.

In the application to molecular biology, unless DNA is denatured, it is always possible to determine the axis of DNA as a curve traced out by the centers of base pairs. In fact, what we are concerned with here is the so-called B-form DNA which occurs when DNA is fully hydrated as it is *in vivo*. While a framed curve is employed to model such DNA, one may assume that C is exactly the axis and oriented by the $3'-5'$ direction of the leading backbone of the double helix; ν is chosen to be parallel to the long axis of the plane containing a base pair, pointing in the direction of the leading backbone.[7] Thus, κ_3 represents the naturally helical structure of DNA. Especially, κ_3 may be regarded as a constant because the structure is uniform. The bending and twisting stiffnesses usually vary according to different environmental factors, such as temperature, pH, and ion concentration of solutions, because these factors have influence on the elastic stiffnesses.[9]

In this article, I continue the use of the Cayley–Klein parameter[10] inspired by Li and Maddocks' paper.[6] By means of this parameter, the whole elasticity theory of rods can be transformed to a geometry of paths on a geometrically distorted S^3 governed by the elastic energy functional. In particular, in the case of *intrinisically straight* elasticity defined by each $\kappa_i = 0$, the theory is transformed exactly into the geometry of paths determined by the Dirichlet functional, which is well understood in Riemannian geometry.

This geometric method is heavily applied to derive Euler–Lagrange equations in Sec. III (all geometric calculations involved there can be found in Chern's lecture notes).[11] We set up a clamped-end variational principle such that the theory has no symmetry of rigid motions of $\mathbb{R}^3$ generically also a correspondence between deformations of elastic rods (satisfying three out of four clamped-end conditions and length preserving) and variations of paths (with endpoints fixed), see Lemma 1. Additionally, regularity of elasticae is established and basic properties of elasticae in the general case are developed. In Sec. IV, we formulate two questions. Physically, Question 1 asks if one can find an elastica whenever the initial values of the strains are known and when such an elastica is unique; Question 2 asks the same thing if the boundary values of orthonormal frames [i.e., conditions (13) and (14)] are given instead. Both answers to Question 1 are affirmative, but they are uncertain for Question 2—Theorems 4 and 5 are partial answers. Moreover, we give a geometric criterion of the nonuniqueness of clamped-end elasticae (Theorem 6). All geometry terminologies used in Sec. IV can be found in Cheeger and Ebin's book.[12]

Author's apology. There exists a vast collection of papers on both elasticity theory and elastic models on DNA. Therefore, I most sincerely apologize to those esteemed scientists whose work is not mentioned in this article. I only list those papers having direct influence on this article.

II. THE GEOMETRIC FORMULATION OF ELASTIC ENERGY

Given a framed curve (C,ν), one may assume e_1, the unit tangent vector of C, $e_2 = \nu$, and $e_3 = e_1 \times e_2$, which points in the direction of the major groove of DNA and is parallel to the short axis of the base plane when modeling DNA is the concern.[7] Arising from (i,j,k) the standard basis of $\mathbb{R}^3$, such an orthonormal frame (e_1, e_2, e_3) gives a curve in SO(3). Recall the double cover π:SU(2)$\rightarrow$SO(3) given by $\pi(q)$ acting on $\mathbb{R}^3$ via $q^{-1}xq$, where $q^{-1}xq$ should be realized as a product of quaternions and $x \in \mathbb{R}^3$ identified with the purely imaginary quaternions. Therefore, the aforementioned curve in SO(3) can be lifted to SU(2), called the *Cayley–Klein* parameter of (C,ν).

Using the facts $q = q_1 + iq_2 + jq_3 + kq_4$, $q^{-1} = \overline{q}$ (because $q \in S^3$), and $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $jk = -kj = i$, $ki = -ik = j$, one can derive some useful equations relating e_i's of a framed curve to its Cayley–Klein parameter q as follows:

$$e_1 = q^{-1} i q = \begin{pmatrix} q_1^2 + q_2^2 - q_3^2 - q_4^2 \\ 2(q_2 q_3 - q_1 q_4) \\ 2(q_1 q_3 + q_2 q_4) \end{pmatrix}, \tag{2}$$

$$e_2 = q^{-1} j q = \begin{pmatrix} 2(q_1 q_4 + q_2 q_3) \\ q_1^2 - q_2^2 + q_3^2 - q_4^2 \\ 2(q_3 q_4 - q_1 q_2) \end{pmatrix}, \tag{3}$$

$$e_3 = q^{-1} k q = \begin{pmatrix} 2(q_2 q_4 - q_1 q_3) \\ 2(q_1 q_2 + q_3 q_4) \\ q_1^2 - q_2^2 - q_3^2 + q_4^2 \end{pmatrix}. \tag{4}$$

The structure equations[11] of (C,ν) are given by $de_i = \omega_{ij} e_j$, where ω_{ij} is the connection one-form satisfying $\omega_{ij} + \omega_{ji} = 0$, $1 \leqslant i,j \leqslant 3$. Then the strains are given by $\omega_1 = \omega_{12}(e_1)$, $\omega_2 = \omega_{13}(e_1)$, and $\omega_3 = \omega_{23}(e_1)$, where $\omega_{..}(e_.)$ denotes the natural pairing between one-forms and vector fields. So $k^2 = \omega_1^2 + \omega_2^2$, where k is the (principal normal) curvature of C. On the other hand, by means of (2)–(4), one has

$$\omega_1 = 2(\dot q_1 q_4 + \dot q_2 q_3 - \dot q_3 q_2 - \dot q_4 q_1) = 2\dot q \cdot \mathrm{B}_1 q, \tag{5}$$

$$\omega_2 = 2(-\dot q_1 q_3 + \dot q_2 q_4 + \dot q_3 q_1 - \dot q_4 q_2) = 2\dot q \cdot \mathrm{B}_2 q, \tag{6}$$

$$\omega_3 = 2(\dot q_1 q_2 - \dot q_2 q_1 + \dot q_3 q_4 - \dot q_4 q_3) = 2\dot q \cdot \mathrm{B}_3 q, \tag{7}$$

where $\dot q$ means the tangent vector of Cayley–Klein parameter q, the dot between $\dot q$ and $\mathrm{B}_i q$ denotes the standard inner product of $\mathbb{R}^4$, and

$$\mathrm{B}_1 q = q_4 \frac{\partial}{\partial q_1} + q_3 \frac{\partial}{\partial q_2} - q_2 \frac{\partial}{\partial q_3} - q_1 \frac{\partial}{\partial q_4},$$

$$\mathrm{B}_2 q = -q_3 \frac{\partial}{\partial q_1} + q_4 \frac{\partial}{\partial q_2} + q_1 \frac{\partial}{\partial q_3} - q_2 \frac{\partial}{\partial q_4},$$

$$\mathrm{B}_3 q = q_2 \frac{\partial}{\partial q_1} - q_1 \frac{\partial}{\partial q_2} + q_4 \frac{\partial}{\partial q_3} - q_3 \frac{\partial}{\partial q_4}.$$

One thus can obtain the Lie bracket of $\mathrm{B}_i q$ and $\mathrm{B}_j q$ as follows:

$$[\mathrm{B}_i q, \mathrm{B}_j q] = 2\epsilon_{ijk} \mathrm{B}_k q, \quad \text{for all } i,j,k, \tag{8}$$

where ϵ_{ijk} is the permutation sign of (ijk). Equations (5)–(7) also suggest that we write $\dot q$ as

$$\dot q = \sum_{i=1}^{3} \frac{\omega_i}{2} \mathrm{B}_i q. \tag{9}$$

Notice that with respect to the standard orientation on $S^3, \mathrm{B}_1 q, \mathrm{B}_2 q, \mathrm{B}_3 q$ forms a negatively oriented basis of $T_q S^3$. It is also worth emphasizing that if the Frenét frame of C exists, then $\omega_1 = k \cos\theta$, $\omega_2 = k \sin\theta$, and $\omega_3 = \tau - \dot\theta$, where τ is the geometric torsion of C and θ is the angle measured from the principal normal of C to e_2. Thus, the relations are very useful in uncovering complete information of an elastica once its axis is *nondegenerate*, i.e., $k > 0$.

Equipped with (5)–(7), one can rewrite (1) as

$$\mathcal{E}(q)=\frac{1}{2}\int_0^l \rho_1(2\dot{q}\cdot \mathrm{B}_1 q-\kappa_1)^2+\rho_2(2\dot{q}\cdot \mathrm{B}_2 q-\kappa_2)^2+\rho_3(2\dot{q}\cdot \mathrm{B}_3 q-\kappa_3)^2\,ds.$$

Because $\mathrm{B}_i q$'s, for $1\leqslant i\leqslant 3$, also form an orthonormal frame on S^3 with respect to the standard metric tensor written as

$$g_{\mathrm{std}}=(\sigma_1)^2+(\sigma_2)^2+(\sigma_3)^2,$$

where σ_i is the one-form dual to $\mathrm{B}_i q$, one can introduce a change of g_{std} immediately shown below:

$$g=\rho_1(\sigma_1)^2+\rho_2(\sigma_2)^2+\rho_3(\sigma_3)^2.$$

Therefore, the elastic energy functional can be formulated as

$$\mathcal{E}(q)=2\int_0^l \|\dot{q}\|^2\,ds-2\sum_{i=1}^{3}\int_0^l \kappa_i\langle \dot{q},\mathrm{B}_i q\rangle\,ds+\frac{1}{2}\sum_{i=1}^{3}\int_0^l \rho_i\kappa_i^2\,ds. \tag{10}$$

Here, the first term of (10) is four times the Dirichlet functional and the $\langle\cdot,\cdot\rangle$ in the second term is the inner product on (S^3,g).

III. EULER–LAGRANGE EQUATIONS OF ELASTIC ENERGY FUNCTIONAL

A. Clamped-end variational principle

In this article, we are interested in the elasticity of *clamped ends:*

$$C(0)=\mathbf{0}, \tag{11}$$

$$C(l)=(a_1,a_2,a_3), \tag{12}$$

$$(e_1(0),e_2(0),e_3(0))=(i,j,k), \tag{13}$$

$$(e_1(l),e_2(l),e_3(l))=(v_1,v_2,v_3), \tag{14}$$

where $\mathbf{0}$ is the origin of $\mathbb{R}^3$,a_i's are real numbers satisfying $\Sigma_{i=1}^3 a_i^2\leqslant l^2$, and (v_1,v_2,v_3) is a set of orthonormal frame at $C(l)$.

Because of (13), we are interested in those Cayley–Klein parameters starting at $1=(1,0,0,0)\in S^3\subset\mathbb{R}^4$. The other endpoint of such a Cayley–Klein parameter corresponding to condition (14) is usually denoted by x.

Lemma 1: The relation between deformations of framed curves and variations of their Cayley–Klein parameters is that a variation of curves on S^3 starting at 1 and ending at a given x can be realized as a deformation of framed curves satisfying (11), (13), and (14) and a property that each framed curve in the deformation has the same length, and vice versa.

Proof: A variation of curves on S^3 starting at 1 and ending at x is a mapping $h:[0,l]\times[0,1]\rightarrow S^3$ satisfying $h(0,\epsilon)=1$ and $h(l,\epsilon)=x$. By means of (2)–(4), one may construct one-parameter families of e_1 and e_2 corresponding to h_ϵ where $h_\epsilon=h(\cdot,\epsilon)$. Thus, one obtains a deformation of framed curves with the required properties.

Conversely, suppose that a deformation of framed curves of equal lengths is given by $H:[0,l]\times[0,1]\rightarrow T\mathbb{R}^3$ where $H(\cdot,\epsilon)=(C_\epsilon,\nu_\epsilon)$ satisfies conditions (11), (13), and (14). With the corresponding frames $(e_1^\epsilon,e_2^\epsilon,e_3^\epsilon)$ where $(e_1^\epsilon,e_2^\epsilon,e_3^\epsilon)$ is the orthonormal frame constructed by $(C_\epsilon,\nu_\epsilon)$, one can produce the associated one-parameter family of curves in SO(3) when compared with (i,j,k). Notice that all curves of the family start and end at the same points, respectively. By means of the homotopy lifting property, these curves can be lifted to SU(2) which all start at 1 and end at x. □

Because the Cayley–Klein parameter only describes motion of the orthonormal frame, condition (12) should be captured by incorporating a Lagrange multiplier with elastic energy as follows:

$$\mathcal{E}_c(q)=\mathcal{E}(q)-\int_0^l \lambda\cdot\left(e_1-\frac{a}{l}\right)ds,$$

where λ is a vector-valued Lagrange multiplier, and the dot appearing in the integrand of the second term of $\mathcal{E}_c$ is the standard inner product of $\mathbb{R}^3$.

B. The Euler–Lagrange equations

The structure equations of the standard S^3 are $d\sigma_i=\sigma_j\wedge\sigma_{ij}$, where the connection one-form $\sigma_{ij}=\epsilon_{ijk}\sigma_k$ can be deduced from (8). Let $f_i=\mathrm{B}_i q/\sqrt{\rho_i}$ and the dual one-form $\eta_i=\sqrt{\rho_i}\sigma_i$ and notice that $g=(\eta_1)^2+(\eta_2)^2+(\eta_3)^2$.

Lemma 2: Let the structure equations of (S^3,g) *be written as* $d\eta_i=\eta_j\wedge\eta_{ij}$, *where* η_{ij} *is the connection one-form. Then*

$$\eta_{ij}=\epsilon_{ijk}\frac{\rho_i+\rho_j-\rho_k}{\sqrt{\rho_i\rho_j\rho_k}}\,\eta_k\,.$$

Proof: Since $\eta_i=\sqrt{\rho_i}\sigma_i$ and

$$d\eta_i=\epsilon_{ijk}\frac{\rho_i}{\sqrt{\rho_i\rho_j\rho_k}}\,\eta_j\wedge\eta_k\,,$$

by means of $d\eta_i=\eta_j\wedge\eta_{ij}$, one has

$$\eta_j\wedge\left(\eta_{ij}-\epsilon_{ijk}\frac{\rho_i}{\sqrt{\rho_i\rho_j\rho_k}}\,\eta_k\right)=0.$$

By Cartan's lemma,[13] there exists (locally defined) functions a_{ijk}'s satisfying $a_{ijk}=a_{ikj}$ such that

$$\eta_{ij}-\epsilon_{ijk}\frac{\rho_i}{\sqrt{\rho_i\rho_j\rho_k}}\,\eta_k=a_{ijk}\eta_k\,.$$

After rearrangement, due to $\eta_{ij}+\eta_{ji}=0$, an identity is obtained

$$a_{ijk}+a_{jik}=\epsilon_{ijk}\frac{\rho_j-\rho_i}{\sqrt{\rho_i\rho_j\rho_k}}.$$

Consider the following three equations:

$$a_{ijk}+a_{jik}=\epsilon_{ijk}\frac{\rho_j-\rho_i}{\sqrt{\rho_i\rho_j\rho_k}},$$

$$a_{jki}+a_{kji}=\epsilon_{jki}\frac{\rho_k-\rho_j}{\sqrt{\rho_i\rho_j\rho_k}},$$

$$a_{kij}+a_{ikj}=\epsilon_{kij}\frac{\rho_i-\rho_k}{\sqrt{\rho_i\rho_j\rho_k}}.$$

Now, we subtract the second equation from the sum of the first and third ones to get

$$a_{ijk}=\epsilon_{ijk}\frac{\rho_j-\rho_k}{\sqrt{\rho_i\rho_j\rho_k}}.$$

Thus, the expression of η_{ij} is obtained. □

Now, one can derive the differential-equation part of the Euler–Lagrange equations of $\mathcal{E}_c$.

Lemma 3: An elastica must satisfy the following equation, where ∇ *is the covariant differentiation on* (S^3,g) *given by* $\nabla_{f_j}f_i=\eta_{ik}(f_j)f_k$,

$$\begin{aligned}2\nabla_{\dot q}\dot q=&(-\lambda\cdot e_2+\rho_1\dot\kappa_1+\rho_3\kappa_3\omega_2-\rho_2\kappa_2\omega_3)\frac{f_1}{\sqrt{\rho_1}}\\&+(-\lambda\cdot e_3+\rho_2\dot\kappa_2-\rho_3\kappa_3\omega_1+\rho_1\kappa_1\omega_3)\frac{f_2}{\sqrt{\rho_2}}\\&+(\rho_3\dot\kappa_3+\rho_2\kappa_2\omega_1-\rho_1\kappa_1\omega_2)\frac{f_3}{\sqrt{\rho_3}}.\end{aligned}\tag{15}$$

Notice that the other part of the Euler–Lagrange equations is simply condition (12).

Sketch of Proof: We will use physicists' notation to help derive the equation:

$$\delta\left(2\int_0^l\|\dot q\|^2\,ds\right)=\int_0^l\langle\delta q,-4\nabla_{\dot q}\dot q\rangle ds,$$

$$\delta\left(-2\sum_{i=1}^3\int_0^l\kappa_i\langle\dot q,\mathrm{B}_iq\rangle ds\right)=\int_0^l\left\langle\delta q,2\sum_{i=1}^3\left(\dot\kappa_i\sqrt{\rho_i}f_i+\rho_i\kappa_i\epsilon_{ikj}\frac{\omega_j}{\sqrt{\rho_k}}f_k\right)\right\rangle ds$$

and

$$\delta\left(-\int_0^l\lambda\cdot\left(e_1-\frac{a}{l}\right)ds\right)=\int_0^l\left\langle\delta q,\frac{-2\lambda\cdot e_2}{\sqrt{\rho_1}}f_1+\frac{-2\lambda\cdot e_3}{\sqrt{\rho_2}}f_2\right\rangle ds.$$

Because $\delta\mathcal{E}_c=0$, we obtain (15). □

Componentwise, one has

$$\rho_1\dot\omega_1+(\rho_3-\rho_2)\omega_2\omega_3=-\lambda\cdot e_2+\rho_1\dot\kappa_1+\rho_3\kappa_3\omega_2-\rho_2\kappa_2\omega_3,\tag{16}$$

$$\rho_2\dot\omega_2+(\rho_1-\rho_3)\omega_1\omega_3=-\lambda\cdot e_3+\rho_2\dot\kappa_2-\rho_3\kappa_3\omega_1+\rho_1\kappa_1\omega_3,\tag{17}$$

$$\rho_3\dot\omega_3+(\rho_2-\rho_1)\omega_1\omega_2=\rho_3\dot\kappa_3+\rho_2\kappa_2\omega_1-\rho_1\kappa_1\omega_2.\tag{18}$$

An immediate consequence is the regularity of solutions to the above Euler–Lagrange equations when all κ_i's are assumed smooth.

Corollary 1: Each elastica is smooth, i.e., C and ν *are both* C^∞.

Proof: Considering (15), since it is elliptic of second order with respect to q, any solution to (15) is smooth [notice that (15) differs from the equation of harmonic mappings[14] by lower-order terms with respect to q]. Then by (2) and (3), the result is obtained. □

In fact, a stronger regularity of elasticae can be also obtained—each elastica is (real) analytic if each κ_i is assumed analytic. This is because SU(2), as a Lie group, is an analytic manifold; also the metric tensor g is analytic. Therefore, q as a solution to (15) is an analytic mapping from $[0,l]$ to S^3 (by the same reason used in the proof of the last corollary). Although the analyticity of solutions is obtained, it is hardly useful in helping us to understand elasticae in this article. Perhaps, the only usefulness is $k>0$ a.e. over $[0,l]$ unless $k=0$ identically.[15]

No matter which regularity one is going to make use of, it is important to notice that although each elastica is smooth (or analytic), the axis might yet have a discontinuity between $C^{(k)}(0)$ and $C^{(k)}(l)$ for some k if $C(0)=C(l)$, where $C^{(k)}$ denotes the kth-order derivative of C. This is because C has been parametrized over $[0,l]$, not S^1.

C. Basic properties of elasticae

Now consider the inner product of $2\nabla_{\dot q}\dot q$ with $\dot q$. By means of (15), one obtains

$$\langle 2\nabla_{\dot q}\dot q,\dot q\rangle=-\lambda\cdot\left(\frac{\omega_1}{2}e_2+\frac{\omega_2}{2}e_3\right)+\frac{1}{2}\sum_{i=1}^{3}\rho_i\dot\kappa_i\omega_i\,. \tag{19}$$

The lhs is equal to the derivative of $\|\dot q\|^2$ with respect to s and the first term of the rhs is exactly the derivative of $-\frac{1}{2}\lambda\cdot e_1$. Therefore,

$$\|\dot q\|^2+\frac{1}{2}\lambda\cdot e_1=\frac{c}{2}+\frac{1}{2}\sum_{i=1}^{3}\int_0^s\rho_i\dot\kappa_i\omega_i\,,$$

where $c=2\|\dot q(0)\|^2+\lambda\cdot i$.

Let *total energy* of (C,ν) be defined by

$$H=2\|\dot q\|^2+\lambda\cdot e_1\,.$$

In the expression of H, the term $2\|\dot q\|^2$ can be regarded as *kinetic energy* and the other term $\lambda\cdot e_1$ as *potential energy* since it is the price one pays for fixing endpoint $C(l)$.

An elastic rod is *uniform* if each κ_i is a constant. Under such an assumption, total energy is a conserved quantity for each elastica, known to G. Kirchhoff and A. Clebsch.[16]

Theorem 1: *Total energy is a conserved quantity for a uniform elastica.*

By (16) and (17), one can write λ as

$$\lambda=\lambda_1e_1-\lambda_2e_2-\lambda_3e_3\,,$$

where λ_1 is potential energy, and

$$\lambda_2=\rho_1\dot\omega_1+(\rho_3-\rho_2)\omega_2\omega_3-\rho_1\dot\kappa_1-\rho_3\kappa_3\omega_2+\rho_2\kappa_2\omega_3\,,$$

$$\lambda_3=\rho_2\dot\omega_2+(\rho_1-\rho_3)\omega_1\omega_3-\rho_2\dot\kappa_2+\rho_3\kappa_3\omega_1-\rho_1\kappa_1\omega_3\,.$$

Since $d\lambda=0$, one obtains

$$\dot\lambda_1+\lambda_2\omega_1+\lambda_3\omega_2=0, \tag{20}$$

$$\lambda_1\omega_1-\dot\lambda_2+\lambda_3\omega_3=0, \tag{21}$$

$$\lambda_1\omega_2-\lambda_2\omega_3-\dot\lambda_3=0. \tag{22}$$

Note (20) is exactly (19) [we need (18) to verify the claim]. The importance of (21) and (22) is that they are bases from which one can solve for elasticae. [Observe that (16) and (17) still involve e_2 and e_3, respectively, and they make solving for elasticae more difficult.] Also, one may make use of (21), (22) and (18) to write a system of differential equations which is equivalent to the (differential equation part of) Euler–Lagrange equations, and then the regularity proved in Corollary 1 is merely a direct application of Cauchy's theorem of differential equations.[17]

Recall that a relaxed state is a state of zero elastic energy. Thus, $\omega_i = \kappa_i$ for each i. Using the above expression of λ, we have $\lambda = \lambda_1 e_1$ for a relaxed state.

Corollary 2: A relaxed state has $\lambda = \lambda_1 e_1$ *where* λ_1 *is a constant. Moreover, if* $\kappa_1^2 + \kappa_2^2 \neq 0$ *at some point in* $[0,l]$, *then* $\lambda_1 = 0$.

Proof: Equations (20)–(22) become

$$\dot{\lambda}_1 = 0 \quad \text{and} \quad \lambda_1 \kappa_1 = \lambda_1 \kappa_2 = 0,$$

respectively. Therefore, one may conclude the corollary. □

To see the existence and uniqueness of a relaxed state, one has to know the differential equation of it:

$$\dot{q} = \sum_{i=1}^{3} \frac{\sqrt{\rho_i}}{2} \kappa_i f_i .$$

The above expression can be derived from the facts $\omega_i = \kappa_i$, $\mathrm{B}_i q = \sqrt{\rho_i} f_i$, and (9).

Theorem 2: *A relaxed state exists uniquely.*

Proof: The result may be obtained by the fundamental theorem of differential equations.[17] □

IV. MAIN RESULTS

Let us study the following two fundamental questions:

Question 1: Given a point $1 \in S^3$ *and a vector* $v \in T_1 S^3$, *is there an elastica whose Cayley–Klein parameter q satisfies* $q(0) = 1$ *and* $\dot{q}(0) = v$? *Such an elastica (if it exists) should be unique with respect to the initial conditions.*

Question 2: Given two points 1, $x \in S^3$, *is there an elastica whose Cayley–Klein parameter connects these two points? Under what condition is such an elastica unique with respect to the boundary conditions?*

The answer to Question 1 is the following.

Theorem 3: *There exists uniquely a solution curve to (15) satisfying the prescribed initial conditions* $q(0) = 1$ *and* $\dot{q}(0) = v$.

Sketch of Proof: On TS^3, let (q,p) be the canonical coordinates. A solution curve one seeks thus corresponds to an integral curve of the following differential equations defined on TS^3.

$$\dot{q} = p,$$

$$\dot{p} = \text{one-half of the r.h.s. of (15)}.$$

By means of the fundamental theorem of differential equations, one obtains the result. □

A partial answer to Question 2 may be given as follows. Let Ω be the path space of all paths on S^3 from 1 to x; here paths are assumed sufficient differentiability and parametrized over $[0,l]$. While proving the next theorem, we do not need that each κ_i is smooth, only C^2. Therefore, Corollary 1 can be restated as: The strains of each elastica are C^{k+1} if all κ_i's are C^k. This implies that the axis of an elastica is C^{k+3} and the material direction is C^{k+2}.

Once κ_1, κ_2, and κ_3 are written collectively as $\kappa = (\kappa_1, \kappa_2, \kappa_3)$, we define $\mathcal{P} = C^2[0,l] \times C^2[0,l] \times C^2[0,l]$ with product topology where $C^2[0,l]$ is topologized by $\mathcal{C}^2$-norm

$$\|u\|_{\mathcal{C}^2} = \max_{x \in [0,l]} |u(x)| + |\dot{u}(x)| + |\ddot{u}(x)|, \quad \text{for } u \in C^2[0,l].$$

Let $\mathcal{V}_q$ be the space of all vector fields defined along $q \in \Omega$ and $\mathcal{V} = \amalg_{q \in \Omega} \mathcal{V}_q$, where ⨿ denotes the disjoint union.

Lemma 4: $\mathcal{V}$ *is a vector bundle over* Ω.

Proof: Only the local triviality of $\mathcal{V}$ is proved here. First, we pick an open neighborhood $\mathcal{O}$ of q in Ω (equipped with compact-open topology) such that for any $\tilde{q}$ in the neighborhood, there exists a unique minimal geodesic connecting $\tilde{q}(s)$ and $q(s)$ for all $s \in [0,l]$. Therefore, using parallel transport along these minimal geodesics, we can establish a one-to-one correspondence between vector fields defined along q and $\tilde{q}$. Such a correspondence gives local triviality of $\mathcal{V}$ as a product of $\mathcal{O}$ and $\mathcal{V}_q$. □

Let F be a functional

$$F:\Omega \times \mathcal{P} \to \mathcal{V}$$

defined by

$$(q,\kappa) \mapsto \left(q, \nabla_{\dot{q}}\dot{q} - \sum_{i=1}^{3} \frac{\alpha_i}{2\sqrt{\rho_i}} f_i \right),$$

where α_i is the coefficient of $f_i/\sqrt{\rho_i}$ of the rhs of (15).

Theorem 4: *Let q_0, from 1 to x, be a Cayley–Klein parameter of some elastica whose intrinsic curvatures and twisting density are κ_1^0, κ_2^0, and κ_3^0, simply written as $\kappa^0 = (\kappa_1^0, \kappa_2^0, \kappa_3^0) \in \mathcal{P}$. Then there is an open neighborhood of κ^0 in $\mathcal{P}$ such that a family of solution curves to* (15) *exists and is parametrized by the neighborhood.*

Note, elasticae of the family have the same λ.

Proof: Let $\tilde{q}(\epsilon), -1 < \epsilon < 1$, be a curve in Ω such that $\tilde{q}(0) = q_0$. The tangent vector to $\tilde{q}$ at $\epsilon = 0$ is a vector field W defined along q_0 while it is realized on S^3. Thus,

$$\left(\frac{\partial}{\partial q} \nabla_{\dot{q}}\dot{q} \right)(q_0,\kappa^0) = \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} \nabla_{\dot{\tilde{q}}}\dot{\tilde{q}} = \nabla_W \nabla_{\dot{\tilde{q}}}\dot{\tilde{q}}|_{\epsilon=0}$$

and the rhs of the last equality is equal to $\nabla_{\dot{q}_0}\nabla_{\dot{q}_0}W + R(W,\dot{q}_0)\dot{q}_0$, where $R(\cdot,\cdot)\cdot$ is the curvature tensor of (S^3,g), by means of the following identities:[17]

$$\nabla_{\tilde{W}}\nabla_{\dot{\tilde{q}}}\dot{\tilde{q}} - \nabla_{\dot{\tilde{q}}}\nabla_{\tilde{W}}\dot{\tilde{q}} - \nabla_{[\tilde{W},\dot{\tilde{q}}]}\dot{\tilde{q}} = R(\tilde{W},\dot{\tilde{q}})\dot{\tilde{q}} \quad \text{and} \quad \nabla_{\tilde{W}}\dot{\tilde{q}} - \nabla_{\dot{\tilde{q}}}\tilde{W} = [\tilde{W},\dot{\tilde{q}}] = 0,$$

where

$$\tilde{W} = V_*\left(\frac{\partial}{\partial \epsilon} \right) \quad \text{and} \quad \dot{\tilde{q}} = V_*\left(\frac{\partial}{\partial s} \right),$$

and V is a one-parameter variation of q_0 which can be regarded as a realization of a curve in Ω on S^3. More precisely, if

$$\left. \frac{d}{d\epsilon} \right|_{\epsilon=0} \tilde{q} = W \in T_{q_0}\Omega,$$

then V is a mapping $V:]-1,1[\times[0,l] \to S^3$ such that $V(\epsilon,0) = 1$, $V(\epsilon,l) = x$ for $\epsilon \in]-1,1[$, $V(0,\cdot) = q_0$ and $W = \tilde{W}(0,\cdot)$ [of course, $\dot{q}_0 = \dot{\tilde{q}}(0,\cdot)$]. Let us write $W = \Sigma W_i f_i$.

Similarly, one can compute

$$\left(\frac{\partial}{\partial q} f_i \right)(q_0,\kappa^0) = \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} f_i = \nabla_{\tilde{W}} f_i|_{\epsilon=0} = \sum_{j,k} \epsilon_{ikj} W_j \frac{\rho_i + \rho_k - \rho_j}{\sqrt{\rho_i \rho_j \rho_k}} f_k,$$

$$\left(\frac{\partial}{\partial q} (-\lambda \cdot e_2) \right)(q_0,\kappa^0) = \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} (-\lambda \cdot e_2) = 2(W_1 \lambda \cdot e_1 - W_3 \lambda \cdot e_3),$$

$$\left(\frac{\partial}{\partial q}(-\lambda\cdot e_3)\right)(q_0,\kappa^0)=\frac{\partial}{\partial\epsilon}\bigg|_{\epsilon=0}(-\lambda\cdot e_3)=2(W_2\lambda\cdot e_1+W_3\lambda\cdot e_2),$$

$$\left(\frac{\partial}{\partial q}\omega_i\right)(q_0,\kappa^0)=\frac{\partial}{\partial\epsilon}\bigg|_{\epsilon=0}\tilde{\omega}_i=\frac{2}{\sqrt{\rho_i}}\left(\dot{W}_i+\epsilon_{ikj}W_j\omega_k\sqrt{\frac{\rho_i}{\rho_j}}\right),$$

where $\tilde{\omega}_i=2\langle\dot{\tilde{q}},f_i\rangle/\sqrt{\rho_i}$. Therefore, we can write

$$\frac{\partial F}{\partial q}(q_0,\kappa^0)(W)=(\nabla_{\dot{q}_0})^2W+R(W,\dot{q}_0)\dot{q}_0-\sum_{i=1}^{3}\beta_i f_i, \tag{23}$$

where β_i's are functions involved with W_i's and their first-order derivatives (and of course also with λ, κ_i's, and ω_i's).

Now, let $K\subset T_{q_0}\Omega$ and R denote the kernel and range of $(\partial F/\partial q)(q_0,\kappa^0)$, respectively. Then K is finite dimensional and R has finite codimension since (23) is an elliptic differential equation in W. Because there is a natural inner product on $T_{q_0}\Omega$ defined by

$$\langle U,V\rangle_{q_0}=\int_0^l\langle U(s),V(s)\rangle ds,$$

one can identify the quotient space $T_{q_0}\Omega/K$ with the orthogonal complement of K in $T_{q_0}\Omega$, written as $K^\perp$. Let Ω^* be the image of $K^\perp$ under the exponential map at q_0,

$$\mathrm{Exp}_{q_0}(U)\in\Omega \quad \text{defined by} \quad \mathrm{Exp}_{q_0}(U)(s)=\exp_{q_0(s)}U(s).$$

Then the restriction of F to $\Omega^*\times\mathcal{P}$ (if necessary, we also need to replace $\mathcal{V}$ by the image of R under the corresponding exponential map when $\mathcal{V}$ is regarded as a manifold) has $\partial F/\partial q$ as an isomorphism at (q_0,κ^0). By the implicit function theorem, one obtains the theorem. □

The geometric picture of the proof of the above theorem is clearer if we restrict our attention to the so-called *geodesic elastica*,[18] especially in the case of intrinsically straight elasticity. An elastica is called geodesic if its Cayley–Klein parameter is a geodesic on (S^3,g). The reason for considering geodesic elasticae is that given any two points on S^3, there is always a geodesic connecting them. However, not every geodesic can be a Cayley–Klein parameter of some elastica. Now if we consider the intrinsically straight elasticity, all β_i's in (23) are zero for a geodesic elastica whose axis is not a straight line. This implies that the kernel is the space of Jacobi fields vanishing at endpoints. Therefore, x is a conjugate point of 1 along q_0 if the kernel is nontrivial.

Motivated by the last explanation of the proof of Theorem 4, let us call an elasticity *geometric* if every geodesic is a Cayley–Klein parameter of some elastica. Such a special category includes intrinsically straight elasticity and isotropic elasticity of zero intrinsic curvatures, but not isotropic O-ring elasticity[3] defined by κ_1 being a constant and $\kappa_2=\kappa_3=0$ because a geodesic being a Cayley–Klein parameter of some O-ring elastica must have $\omega_2=0$ and ω_3 as a constant. Now employing Theorem 4 alone, one can show that Question 2 has an affirmative answer to the existence under some condition on intrinsic curvatures and twisting density. For this we first assume a geodesic elastica of some geometric elasticity whose Cayley–Klein parameter solves Question 2, and then applying Theorem 4 to this elastica we are able to get an elastica whose intrinsic curvatures and twisting density are close (with respect to the product topology of $\mathcal{P}$) to those of the assumed geodesic elastica.

Using the proof of Theorem 4, one can also show the following.

Theorem 5: *Suppose that there is an elastica whose Cayley–Klein parameter is from 1 to x, say* (q_0,λ_0). *Then there exists a family of solution curves to (15) also connecting 1 and x. Moreover, along with the presumed elastica, such a family can be parametrized by an open set in* $\mathbb{R}^3$.

Notice that in order to prove Theorem 5, we replace F by

$$G:\Omega\times\mathbb{R}^3\rightarrow\mathcal{V}$$

defined by

$$(q,\lambda)\mapsto\left(q,\nabla_{\dot q}\dot q-\sum_{i=1}^{3}\frac{\alpha_i}{2\sqrt{\rho_i}}f_i\right).$$

So, elasticae of the family have the same intrinsic curvatures and twisting density.

An interesting application of Theorem 5 is to demonstrate the following.

Theorem 6: *Let q_0, from 1 to x, be a Cayley-Klein parameter of some geodesic elastica of a geometric elasticity such that x is a conjugate point of 1 along q_0. Then there are two distinct solution curves to (15) such that the corresponding elasticae satisfy the same conditions (11)–(14).*

Proof: Consider the given geodesic elastica and all perturbed solutions obtained by Theorem 5, written as q_λ. Since the integral $\int_0^l e_1(q_\lambda)ds$ is continuous in q_λ (or more accurately λ), it is true that either (i) there are two different values of λ, say λ_α and λ_β, such that the elasticae whose Cayley–Klein parameters are q_{λ_α}, and q_{λ_β}, respectively, share the same endpoint [condition (12)], or (ii) the total integral of e_1 regarded as a mapping of λ is one-to-one. In the latter, since q_0 is a geodesic whose endpoints are conjugate to each other, there must be a variation of q_0 consisting of geodesics from 1 to x.[12] Since elasticae whose Cayley–Klein parameters are q_λ exhaust neighboring points of $\int_0^l e_1(q_0)ds$, there must be one elastica whose Cayley–Klein parameter is in the aforementioned geodesic variation such that its endpoint at $s=l$ lies in the image resulting from the elasticae associated to q_λ. Thus, the proof is completed. □

ACKNOWLEDGMENTS

I want to thank my advisor, Professor James White, for his constant encouragement, very patient guidance, and support. Gratitude also goes to Professor John Maddocks, Professor Robert Manning, and Professor Hong Qian for many helpful discussions, and to Professor Joseph Rudnick for his comments. I very much appreciate the referee for the comments and suggestions used to revise this paper.

[1] F. B. Fuller, Proc. Natl. Acad. Sci. USA **68**, 815 (1971).

[2] The corresponding mathematical theory was done in the case of Bernoulli's elasticity by J. Langer and D. A. Singer, J. London Math. Soc. **30**, 512 (1984), and it can be generalized to Kirchhoff's elasticity of isotropic rods with zero intrinsic curvatures.

[3] W. R. Bauer, R. A. Lund, and J. H. White, Proc. Natl. Acad. Sci. USA **90**, 833 (1993).

[4] A. Bolshoy, P. McNamara, R. E. Harrington, and E. N. Trifonov, Proc. Natl. Acad. Sci. USA **88**, 2312 (1991).

[5] P. De Santis, A. Palleschi, M. Savino, and A. Scipioni, Biophys. Chem. **42**, 147 (1992).

[6] Y. Li and J. H. Maddocks, to appear in J. Comput. Phys. (the Gzipped postscript is located at http://lcvmwww.epfl.ch/~jhm/research.html).

[7] T. P. Westcott, I. Tobias and W. K. Olson, J. Phys. Chem. **99**, 17926 (1995).

[8] J. H. White, R. A. Lund, and W. R. Bauer, Biopolymers **38**, 235 (1996).

[9] P. J. Hagerman, Annu. Rev. Biophys. Biophys. Chem. **17**, 265 (1988).

[10] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison–Wesley, Reading, MA, 1980).

[11] S.-s. Chern, *Topics in Differential Geometry* (1951), also found in *Selected Papers Vol. IV* (Springer-Verlag, New York, 1989).

[12] J. Cheeger and D. Ebin, *Comparison Theorems in Riemannian Geometry* (North-Holland, Amsterdam, 1975).

[13] S. Sternberg, *Lectures on Differential Geometry* (Prentice–Hall, Englewood Cliffs, NJ, 1964).

[14] J. Eells and L. Lemaire, Bull. London Math. Soc. **10**, 1 (1978).

[15] The result can be proved by the fact: if u defined on $[0,l]$ is analytic, then $u^{-1}(c)$ either is the whole interval or contains no intervals. For the latter, such $u^{-1}(c)$ is called zero-dimensional. A characterization of some zero-dimensional spaces is that such a space is a union of the Cantor set and an open countable subset, which is applicable to what we are interested in. For details, see a preprint on Kirchhoff's elasticity of isotropic *O*-ring rods (a postscript is located at http://www.math.ucla.edu/~khu/works.html).

[16] A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, 4th ed. (Dover, New York, 1944).

[17] G. Birkhoff and G.-C. Rota, *Ordinary Differential Equations*, 2nd ed. (Wiley, New York, 1969).

[18] General properties of geodesic elasticae will not be presented in this paper because there are few geodesic elasticae when nontrivial intrinsic curvatures exist. This situation is very conceivable, because the geodesic equation $\nabla_{\dot{q}}\dot{q}=0$ is independent of intrinsic curvatures, yet the existence of the latter puts restrictions on a geodesic [cf. the both sides of (16)–(18)]. In fact, in my previous study on this topic in the case of the isotropic elasticity in which $\kappa_2=\kappa_3=0$, it was shown that if κ_1 is neither a linear function nor a harmonic oscillation, then there is no geodesic elastica.

Symplectic-energy-momentum preserving variational integrators

C. Kane
Graduate Aeronautical Laboratories and CDS, California Institute of Technology, Pasadena, California 91125

J. E. Marsden
CDS 107-81, California Institute of Technology, Pasadena, California 91125

M. Ortiz
Graduate Aeronautical Laboratories, California Institute of Technology, Pasadena, California 91125

(Received 20 January 1999; accepted for publication 3 February 1999)

The purpose of this paper is to develop variational integrators for conservative mechanical systems that are symplectic and energy and momentum conserving. To do this, a space–time view of variational integrators is employed and time step adaptation is used to impose the constraint of conservation of energy. Criteria for the solvability of the time steps and some numerical examples are given. [S0022-2488(99)01707-7]

I. INTRODUCTION

The purpose of this paper is to develop variational integrators for conservative mechanical systems with adaptive time steps. The resulting algorithms are symplectic, energy preserving, and they also preserve the momentum maps (Noether quantities) associated with symmetry groups.

An important idea for how to develop such integrators comes from the paper of Marsden, Patrick, and Shkoller[1] in which the space–time view is stressed. This viewpoint is very important for two main reasons:

(1) to avoid conflicts with well-known theorems (Ge and Marsden[2]), which limit the possibility that *constant time stepping* algorithms be symplectic and energy and momentum preserving; and
(2) to give meaning to the term *symplectic* in the context of adaptive time stepping algorithms since the algorithm is not given by a single mapping associated with a constant time step.

The basic algorithm itself consists of two parts. First, to update positions, the variational approach of Veselov,[3,4] and Moser and Veselov[5] is adopted; these ideas were implemented numerically by Wendlandt and Marsden.[6] Second, to compute the time steps themselves, energy preservation is imposed. Roughly speaking, we make use of the fact that time and energy are conjugate variables.

In this paper, our main purpose is the following:

(1) to set up the basic algorithm that implements these symplectic-energy-momentum (SEM) integrators;
(2) to make precise the sense in which the algorithms are symplectic;
(3) to investigate the solvability conditions for the time step; and
(4) to give some simple numerical examples.

With regard to solving for the time step, we shall indicate how the solvability conditions are closely related to the positivity of the numerically computed kinetic energy. In particular, when one is close to a location with zero velocity (roughly speaking, ‘‘turning points’’), our solvability criterion and examples indicate that one should move through such points using a criterion other than energy preservation. Numerically, this causes a slight adjustment to the energy. How one should best treat these points requires further investigation. *It is not the purpose of this paper to give any extensive numerical tests or implementations of these methods*. This is a nontrivial task, but clearly needs to be done, and is planned for other publications.

In future papers we will also investigate the use of these ideas in our collision algorithms[7] and how one can incorporate dissipative effects.

II. BRIEF REVIEW OF VARIATIONAL INTEGRATORS

In this section, we recall some of the essential features of variational integrators that are needed in this paper. For additional details, see Refs. 6 and 1.

A. Limitations on mechanical integrators

There has been a large literature developing on the use of energy-momentum and symplectic-momentum integrators. We shall not attempt to survey this all here, but rather refer the reader to some of the recent literature, such as the collection of papers in Ref. 8 and Ref. 9. We do mention that for time stepping algorithms with fixed time steps, the theorem of Ge and Marsden[2] has led to a general division of algorithms into those that are energy-momentum preserving and those that are symplectic-momentum preserving. One of our main points is that if one takes a space–time view of variational integrators, as is advocated in Ref. 1, then one can have all three of these properties. Papers typified by Simo and Tarnow,[10] Simo, Tarnow, and Wong,[11] and Gonzalez[12] have focused on energy preserving algorithms, but they presumably fail (except, perhaps, in special cases, such as integrable systems) to be symplectic. Other approaches based on Hamilton's principle are those of Shibberu[13] and Lewis.[14] See also the work by Lee.[15]

B. Accuracy of solutions

We should, at the outset, make another point clear. We are not claiming anything about the accuracy of individual trajectories. Indeed, it is well known that structure preservation alone does not guarantee this. (See, e.g., Refs. 16, 17.) For systems with complicated, unstable, or chaotic trajectories, it is not clear that accuracy of individual trajectories is the correct question to ask. Rather, one should probably concentrate on statistical properties of solutions. These are deep questions that we do not attempt to address here, but one hopes that by preserving as much of the structure as possible, one is closer to addressing such issues. In some cases the advantage of being symplectic is clear; for instance, the condition of being symplectic guarantees that the phase space volume is preserved and this can be an obvious limitation on many integrators; even after relatively short times, one can see phase space volume not preserved in many integrators.

C. Some common integrators

In structural mechanics, the $\beta=0$, $\gamma=\frac{1}{2}$ member of the widely used Newmark family is a variational integrator (see, for example, Ref. 11) and therefore is symplectic and momentum preserving. In fact, the whole Newmark family of algorithms is variational.[18] Our methods can be used to make these integrators also preserve energy by using time-adaptive stepping. We also mention that the popular Verlet methods and shake algorithms are variational integrators (see Refs. 6 and 18 for further discussion and references).

D. Dissipation and constraints

While dissipation and forcing are of course very important, as we have mentioned, we leave their discussion for future publications. One possibility is that dissipative effects can be dealt with by means of product formulas, as in Refs. 19–21, for example. Another is to incorporate the dissipative effects into the variational principle, as in Refs. 18 and 22.

Constraints are also very important for integrators. We also do not discuss these in any detail in this paper. However, we do mention that variational integrators handle constraints in a simple and efficient way.

Variational methods also generalize to partial differential equations (PDE's) using multisymplectic geometry with the result being a class of multisymplectic momentum integrators. See Ref. 1 for details and numerical examples. This type of approach should ultimately be of use in elastodynamics as well as ocean dynamics, for example.

E. Symmetry and reduction

We should also mention that for mechanical systems with symmetry, the investigation of discrete versions of reduction theory, such as Euler–Poincaré reduction,[23] are of current interest.[24,25] We will not be making use of this reduction theory in this paper, but it is related since our integrators are intended to preserve symmetry. It would be of interest to develop time-adapted integrators in the sense of the present paper in the general context of discrete reduction.

F. The discrete variational principle

Given a *configuration space* Q, a **discrete Lagrangian** is a map

$$L_d : Q \times Q \to \mathbb{R}.$$

In practice, L_d is obtained by approximating a given Lagrangian as we shall discuss later, but regard L_d as given for the moment. The time step information will be contained in L_d and we regard L_d as a function of 2 nearby points (q_k, q_{k+1}).

For a positive integer N, the **action sum** is the map $S_d : Q^{N+1} \to \mathbb{R}$ defined by

$$S_d = \sum_{k=0}^{N-1} L_d(q_k, q_{k+1}),$$

where $q_k \in Q$ and k is a nonnegative integer. The action sum is the discrete analog of the action integral in mechanics.

The **discrete variational principle** states that the evolution equations extremize the action sum given fixed end points, q_0 and q_N. Extremizing S_d over $q_1, \ldots, q_{N-1}$ leads to the **discrete Euler–Lagrange (DEL) equations:**

$$D_1 L_d(q_k, q_{k+1}) + D_2 L_d(q_{k-1}, q_k) = 0$$

for all $k = 1,...,N-1$. We can write this equation in terms of a discrete **algorithm**

$$\Phi : Q \times Q \to Q \times Q$$

defined implicitly by

$$D_1 L_d \circ \Phi + D_2 L_d = 0,$$

i.e.,

$$\Phi(q_{k-1}, q_k) = (q_k, q_{k+1}).$$

If, for each $q \in Q$, $D_1 L_d(q,q) : T_q Q \to T_q^* Q$ is invertible, then $D_1 L_d : Q \times Q \to T^*Q$ is locally invertible and so the algorithm Φ, which flows the system forward in discrete time, is well defined for small time steps.

G. Variational algorithms are symplectic

To explain the sense in which the algorithm is symplectic, first define the **fiber derivative** (or the discrete Legendre transform) by

$$\mathbb{F}L_d : Q \times Q \to T^*Q; \quad (q_0, q_1) \mapsto (q_0, D_1 L_d(q_0, q_1)),$$

and define the two-form ω on $Q\times Q$ by pulling back the canonical two-form $\Omega_{\text{CAN}}=dq^i\wedge dp_i$ from T^*Q to $Q\times Q$:

$$\omega=\mathbb{F}L_d^*(\Omega_{\text{CAN}}).$$

The fiber derivative is analogous to the standard Legendre transform.

The coordinate expression for ω is

$$\omega=\frac{\partial^2 L_d}{\partial q_k^i \partial q_{k+1}^j}(q_k,q_{k+1})dq_k^i\wedge dq_{k+1}^j.$$

A fundamental fact is that

the algorithm Φ exactly preserves the symplectic form ω.

One proof of this is to simply verify it with a straightforward calculation—see Ref. 6 for the details. Another is to derive the same conclusion directly from the variational structure, as is done in Ref. 1.

H. The algorithm preserves momentum

Recall that *Noether's theorem* states that a continuous symmetry of the Lagrangian leads to conserved quantities, as with linear and angular momentum. A nice way to derive these conservation laws (the way Noether did it) is to use the *invariance of the variational principle*.

Assume that the discrete Lagrangian is invariant under the action of a Lie group G on Q, and let $\xi\in\mathfrak{g}$, the Lie algebra of G. By analogy with the continuous case, define the **discrete momentum map**, $\mathbf{J}_d:Q\times Q\to\mathfrak{g}^*$ by

$$\langle\mathbf{J}_d(q_k,q_{k+1}),\xi\rangle:=\langle D_1L_d(q_k,q_{k+1}),\xi_Q(q_k)\rangle.$$

A second fundamental fact is that

the algorithm Φ exactly preserves the momentum map.

I. Construction of mechanical integrators

Assume we have a mechanical system with a constraint manifold, $Q\subset V$, where V is a real finite-dimensional vector space, and that we have an unconstrained Lagrangian, $L:TV\to\mathbb{R}$ which, by restriction of L to TQ, defines a **constrained Lagrangian**, $L^c:TQ\to\mathbb{R}$. Roughly speaking, V is a containing vector space in which the computer arithmetic will take place. In particular, coordinate charts on Q are *not* chosen for this purpose. In fact, apart from the use of the containing vector space V, the algorithms developed here are independent of the use of coordinates on Q.

We also assume that we have a vector-valued **constraint function**, $g:V\to\mathbb{R}^k$, such that our constraint manifold is given by $g^{-1}(0)=Q\subset V$, with 0 a regular value of g. The dimension of V is denoted n and therefore, the dimension of Q is $m=n-k$.

Define a **discrete, unconstrained Lagrangian**, $L_d:V\times V\to\mathbb{R}$ in some consistent manner, such as

$$L_d(x,y)=L\left(\gamma x+(1-\gamma)y,\frac{y-x}{h}\right), \tag{II.1}$$

where $h\in\mathbb{R}_+$ is the time step and $0\leqslant\gamma\leqslant1$ is an interpolation parameter.

The corresponding discrete Euler–Lagrange equations give an algorithm closely related to (in a sense made precise in Ref. 18) the **Newmark algorithm** for the standard choice of Lagrangian given by kinetic minus potential energy. We get the *central difference method* for $\gamma=\frac{1}{2}$ and we get the *shake* algorithm with $\gamma=1$ (the *Verlet* algorithm is the unconstrained version of the shake algorithm). We also note that the *Moser–Veselov discrete Lagrangian* for the rigid body is constructed using either $\gamma=1$ or $\gamma=0$ (see Ref. 24 for details).

We remark in passing that other choices for discrete Lagrangians are also possible, such as

$$L_d(x,y)=\sigma L\left(\gamma_1 x+(1-\gamma_1)y,\frac{y-x}{h}\right)+(1-\sigma)L\left(\gamma_2 x+(1-\gamma_2)y,\frac{y-x}{h}\right),$$

where σ, λ_1, and λ_2 are between 0 and 1. These other choices, which give algorithms such as the midpoint rule, are not investigated here (see Ref. 18). Alternative choices, of which this is an example, as well as issues of local truncation error and accuracy are investigated in Ref. 26.

The **unconstrained action sum** is defined by

$$S_d=\sum_{k=0}^{N-1} L_d(x_k,x_{k+1}).$$

Extremize $S_d:V^{N+1}\to\mathbb{R}$ *subject to the constraint* that $x_k\in Q\subset V$ for $k=1,...,N-1$, i.e., solve

$$D_1L_d(x_k,x_{k+1})+D_2L_d(x_{k-1},x_k)+\lambda_k^T Dg(x_k)=0$$

(no sum on k) with $g(x_k)=0$ for $k=1,...,N-1$. Here, the λ_k are *Lagrange multipliers*, chosen to enforce the constraints.

Thus, the algorithm is defined by starting with x_k and x_{k-1} in $Q\subset V$, i.e., $g(x_k)=0$ and $g(x_{k-1})=0$, and solving

$$D_1L_d(x_k,x_{k+1})+D_2L_d(x_{k-1},x_k)+\lambda_k^T Dg(x_k)=0$$

subject to $g(x_{k+1})=0$, for x_{k+1} and λ_k. In terms of the *unconstrained Lagrangian*, the algorithm reads as follows:

$$\begin{aligned}&\frac{1}{h}\left[\frac{\partial L}{\partial \dot{x}}\left(\gamma x_{k-1}+(1-\gamma)x_k,\frac{x_k-x_{k-1}}{h}\right)-\frac{\partial L}{\partial \dot{x}}\left(\gamma x_k+(1-\gamma)x_{k+1},\frac{x_{k+1}-x_k}{h}\right)\right]\\&\quad\times(1-\gamma)\frac{\partial L}{\partial x}\left(\gamma x_{k-1}+(1-\gamma)x_k,\frac{x_k-x_{k-1}}{h}\right)\\&\quad+\gamma\frac{\partial L}{\partial x}\left(\gamma x_k+(1-\gamma)x_{k+1},\frac{x_{k+1}-x_k}{h}\right)+D^Tg(x_k)\lambda_k=0\end{aligned}$$

together with $g(x_{k+1})=0$.

Example: If the continuous Lagrangian is

$$L(q,\dot{q})=\tfrac{1}{2}\dot{q}^T M\dot{q}-V(q)$$

with constraint $g(q)=0$, where M is a constant mass matrix, and V is the potential energy, then the DEL equations are

$$\begin{aligned}&M\left(\frac{x_{k+1}-2x_k+x_{k-1}}{h^2}\right)+(1-\gamma)\frac{\partial V}{\partial q}(\gamma x_{k-1}+(1-\gamma)x_k)\\&\quad+\gamma\frac{\partial V}{\partial q}(\gamma x_k+(1-\gamma)x_{k+1})-D^Tg(x_k)\lambda_k=0\end{aligned}$$

with $g(x_{k+1})=0$.

Wendlandt and Marsden[6] show that *the algorithm defined using Lagrange multipliers coincides with that defined intrinsically using the constrained discrete Lagrangian on* $Q\times Q$, *so it is symplectic and momentum preserving.*

J. An intrinsic variational viewpoint

Recall that given a Lagrangian function $L:TQ\to\mathbb{R}$, we construct the corresponding **action functional** $\mathfrak{S}$ on C^2 curves $q(t)$ by (using coordinate notation)

$$\mathfrak{S}(q(\cdot))\equiv\int_a^b L\left(q^i(t),\frac{dq^i}{dt}(t)\right)dt. \tag{II.2}$$

The action functional depends on a and b, but this is not explicit in the notation. Hamilton's principle seeks the curves $q(t)$ for which the functional $\mathfrak{S}$ is stationary under variations of $q^i(t)$ with fixed endpoints. It will be useful to recall this calculation; namely, we seek curves $q(t)$ which satisfy

$$\mathbf{d}\mathfrak{S}(q(t))\cdot\delta q(t)\equiv\left.\frac{d}{d\epsilon}\right|_{\epsilon=0}\mathfrak{S}(q(t)+\epsilon\delta q(t))=0 \tag{II.3}$$

for all $\delta q(t)$ with $\delta q(a)=\delta q(b)=0$. Abbreviating $q_\epsilon\equiv q+\epsilon\delta q$, and using integration by parts, the calculation is

$$\mathbf{d}\mathfrak{S}(q(t))\cdot\delta q(t)=\left.\frac{d}{d\epsilon}\right|_{\epsilon=0}\int_a^b L\left(q_\epsilon^i(t),\frac{dq_\epsilon^i}{dt}(t)\right)dt=\int_a^b\delta q^i\left(\frac{\partial L}{\partial q^i}-\frac{d}{dt}\frac{\partial L}{\partial\dot{q}^i}\right)dt+\left.\frac{\partial L}{\partial\dot{q}^i}\delta q^i\right|_a^b. \tag{II.4}$$

The last term in (II.4) vanishes since $\delta q(a)=\delta q(b)=0$, so that the requirement (II.3) for $\mathfrak{S}$ to be stationary yields the **Euler–Lagrange equations**

$$\frac{\partial L}{\partial q^i}-\frac{d}{dt}\frac{\partial L}{\partial\dot{q}^i}=0. \tag{II.5}$$

Notice that the boundary term in the first variation of the action is the canonical one form $p_i\dot{q}^i$. This is the starting point for the variational proof that the algorithm is symplectic. The idea is to restrict $\mathfrak{S}$ to the space of solutions and to use the general identity $d^2\mathfrak{S}=0$ to derive the symplectic nature of flow of the Euler–Lagrange equations. The point is that *this type of derivation also is valid for the discrete case*, as is shown in Ref. 1.

We also mention that one can similarly give a derivation of the conservation of momentum maps entirely based on the variational principle as well.

We end this section with one further important remark. Namely, one may think that the discrete symplectic form and momentum map that are conserved by the variational algorithm are somehow ''concocted'' to be conserved. This is not the case. Indeed, one can, via the discrete Legendre transform, transfer the algorithm to position-momentum space. Transferred to these variables, the algorithm will preserve the *standard* symplectic form $dq^i\wedge dp_i$ and the *standard* momentum map. As M. West pointed out to us, to get a corresponding algorithm that is consistent with the corresponding continuous Hamiltonian system on T^*Q, and one that is also in line with our discrete energy developed below, one should really use the map

$$(q_0,q_1)\mapsto(q_0,-hD_1L_d(q_0,q_1)),$$

where h is the time step, but this does not affect the results here.

III. REVIEW OF ENERGY AND SYMPLECTICITY CONSERVATION IN THE CONTINUOUS CASE

The main issue addressed in the following section is how we can achieve conservation of energy using adaptive time steps. We shall see that apart from some exceptional circumstances, which we can algorithmically identify, one can achieve this.

To address these issues, we first consider the continuous time case.

A. Conservation of energy

We will first recall how conservation of energy is derived directly from Hamilton's principle in the case where $L(q,\dot{q})$ is time independent. This provides a clue about how one should proceed with the time-adaptive steps.

Assume that $q(t)$ is a solution of the Euler–Lagrange equations. Let $s_\varepsilon(t)$ be a family of functions of t depending on the parameter ε and with $s_0(t)=t$ and with $s_\varepsilon(a)=a$, $s_\varepsilon(b)=b$. Let

$$\delta s(t)=\frac{d}{d\varepsilon}s_\varepsilon\bigg|_{\varepsilon=0}.$$

We consider the associated family of curves $q(s_\varepsilon(t))$ which has the variation

$$\delta q(t)=\delta s(t)\dot{q}(t).$$

Hamilton's principle $(\delta\int L\,dt=0)$ in this case gives us

$$\int \frac{\partial L}{\partial q}\delta q+\frac{\partial L}{\partial \dot{q}}(\dot{\delta q})=\int\left(\frac{\partial L}{\partial q}-\frac{d}{dt}\frac{\partial L}{\partial \dot{q}}\right)\delta q. \tag{III.1}$$

Using the special form of the variation, this becomes

$$\int \frac{\partial L}{\partial q}\dot{q}\,\delta s-\left(\frac{d}{dt}\frac{\partial L}{\partial \dot{q}}\right)\dot{q}\,\delta s\,dt=0. \tag{III.2}$$

Equation (III.2) gives

$$\frac{\partial L}{\partial q}\dot{q}-\frac{d}{dt}\frac{\partial L}{\partial \dot{q}}\dot{q}=0=\frac{dE}{dt}, \tag{III.3}$$

where, as usual,

$$E=\frac{\partial L}{\partial \dot{q}}\dot{q}-L(q,\dot{q}). \tag{III.4}$$

The point here is that we see a sense in which the Hamiltonian H arises naturally when one considers variations of the curve $q(t)$ that are given by time reparametrizations.

B. Symplecticity in the space–time sense

In Ref. 1 it is shown how the variational principle naturally leads to boundary terms in both the continuous and the discrete case that leads to a deeper understanding of why the Euler–Lagrange and Hamilton equations themselves preserve the symplectic structure, as well as their discrete counterparts. We shall make use of this type of argument below in the discrete case. To help motivate the result, we make some relevant remarks here.

Consider a (possibly time dependent) Hamiltonian $H(q,p,t)$ in the canonically conjugate variables q^i, p_i and introduce an extended Hamiltonian $\bar{H}$, a function of q, p and two new real variables q_0 and p_0 by

$$\bar{H}(q,p,q_0,p_0)=H(q,p,q_0)+p_0.$$

Hamilton's equations for this new autonomous Hamiltonian agree with the time-dependent equations for the original H if we identify q_0 with the time and p_0 with $-H$. In addition, this leads one to the conservation of the canonical symplectic structure in the space–time sense, namely,

$$\Omega_H=\omega+dH\wedge dt,$$

where $\omega = \Sigma_i dq_i \wedge dp_i$ is the canonical symplectic form. What is interesting for us later is the following remark in case H is time independent (so H is conserved). Consider the flow $\bar{F}_s$ of the extended Hamiltonian $\bar{H}$, which is given by advancing the time by an amount s:

$$(q(t),p(t),t,H) \mapsto (q(t+s),p(t+s),t+s,H),$$

where q and p advance by the flow F_s of the original Hamiltonian system. Imagine, for later purposes, that s is a function of the initial point and t, so that ds is a nontrivial differential. Then the statement that the flow $\bar{F}_s$ preserves the symplectic form above reads: $\bar{F}_s^* \Omega_H = \Omega_H$ or, equivalently,

$$F_s^* \omega + dH \wedge d(t+s) = \omega + dH \wedge dt.$$

Cancelling $dH \wedge dt$, this reads

$$F_s^* \omega + dH \wedge ds = \Omega_0. \tag{III.5}$$

If we think of s as the time advancement, or the time step, then below we will prove a discrete analog of this identity, which is how we will interpret symplecticity in the time-dependent sense.

Following the arguments in the discrete case later, or the methods of Marsden *et al.*,[1] one can also derive a Lagrangian version of this identity from the variational principle.

IV. THE VARIATIONAL ENERGY ALGORITHM

We consider a discrete Lagrangian that is (possibly) time dependent and has an associated time step h_1 that may be coupled to the current choice of points (q_1,q_2); we denote this discrete Lagrangian by $L_{h_1}(q_1,q_2) := L_d(q_1,q_2,h_1)$.

Given (q_0,q_1,h_0), we seek to find (q_1,q_2,h_1). In general, this will give us a way to pass from data (q_{k-1},q_k,h_{k-1}) to (q_k,q_{k+1},h_k).

This setup differs from the usual discrete Lagrangian procedures in the inclusion of time step information h_k that is coupled to the current configuration data (q_k,q_{k+1}).

We will find q_2 and h_1 together, by solving an equation similar to the discrete Euler–Lagrange equation for q_2, while we solve for h_1 using the equation $E_d(q_0,q_1,h_0) = E_d(q_1,q_2,h_1)$ where E_d is the discrete energy function, defined below.

(a) The discrete action. One choice of discrete action is obtained by just using the following approximation to the action integral for the first two sets of points, (q_0,q_1,h_0) and (q_1,q_2,h_1):

$$[h_0 L_d(q_0,q_1,h_0) + h_1 L_d(q_1,q_2,h_1)]. \tag{IV.1}$$

One could also use other, more accurate methods to approximate the action integral. This might lead to some interesting new, more accurate algorithms, but we shall not explore them in this paper.

(b) The discrete algorithm. To derive the algorithm, we consider the same discrete variational principle as before, but now parametrized by the time step information:

$$\frac{\partial}{\partial q_1}[h_0 L_d(q_0,q_1,h_0) + h_1 L_d(q_1,q_2,h_1)] = 0. \tag{IV.2}$$

We also write this as

$$h_0 D_2 L_d(q_0,q_1,h_0) + h_1 D_1 L_d(q_1,q_2,h_1) = 0.$$

In this relation, the time steps h_0,h_1 are held fixed. We will later derive the symplectic relation by considering variations of solutions, just as in the continuous case.

A. The discrete energy

The above variational equation (IV.2) will be coupled with an energy equation that will enable us to solve for both q_2 and h_2.

We define the **discrete energy** as follows:

$$E_d(q_0,q_1,h_0) = -h_0 D_3 L_d(q_0,q_1,h_0) - L_d(q_0,q_1,h_0) = -\frac{\partial}{\partial h_0}[h_0 L_d(q_0,q_1,h_0)]. \quad \text{(IV.3)}$$

This intrinsic definition is motivated in part by the fact that for Lagrangians of the form of kinetic minus potential energy, and with the choice of discrete Lagrangian given by (II.1), the discrete energy is given by the expression one would naturally think of, namely it is easily verified that, in this case, we have

$$E_d(q_0,q_1,h_0) = \frac{1}{2}\left(\frac{q_1-q_0}{h_0}\right)^T M \left(\frac{q_1-q_0}{h_0}\right) + V(\gamma q_0 + (1-\gamma) q_1). \quad \text{(IV.4)}$$

In this case, this can also be written as

$$E_d(q_0,q_1,h_0) = E\left(\gamma q_0 + (1-\gamma) q_1, \left(\frac{q_1-q_0}{h_0}\right)\right), \quad \text{(IV.5)}$$

where $E(q,\dot{q})$ is the energy associated with the original Lagrangian $L(q,\dot{q})$. As pointed out to us by Matt West, one can also motivate this definition of the energy using the variational principle and a discrete version of the Hamilton-Jacobi equation.

The main second equation defining the algorithm is

$$E_d(q_0,q_1,h_0) = E_d(q_1,q_2,h_1), \quad \text{(IV.6)}$$

or, with algorithmic notation,

$$E_{k-1} := E_d(q_{k-1},q_k,h_{k-1}) = E_d(q_k,q_{k+1},h_k) := E_k. \quad \text{(IV.7)}$$

For example, in the case the Lagrangian equals kinetic plus potential energy, the condition $E_0 = E_1$ is equivalent to

$$h_1^2 = \frac{(q_2-q_1)^T M (q_2-q_1)}{2[E_0 - V(\gamma q_1 + (1-\gamma) q_2)]}. \quad \text{(IV.8)}$$

For this to remain meaningful as we compute, we need to make sure that the *computed kinetic energy*

$$E_0 - V(\gamma q_1 + (1-\gamma) q_2) \quad \text{(IV.9)}$$

remains positive.

To realize this condition, one can consider verifying it *a posteriori* as follows.

(1) First compute the square of the new time step h_1.

(2) Substitute h_1^2 in the relation

$$h_0 D_2 L_{h_0}(q_0,q_1) + h_1 D_1 L_{h_1}(q_1,q_2) = 0, \quad \text{(IV.10)}$$

which gives an implicit equation for q_2.

(3) Compute q_2 implicitly.

(4) Verify from the formula for h_1^2 that one gets a positive answer.

(5) If so, we proceed. If not, we keep the time step from the last iterate and proceed.

This approach will of course induce an energy variation in such cases. As we shall see below in greater detail, *this will only happen near ''turning points;'' that is, near points where the velocity is nearly zero.*

In the specific example, the equation for q_2 reads

$$m\left[\frac{(q_1-q_0)}{h_0^2}-(1-\gamma)V'(\gamma q_0+(1-\gamma)q_1)\right]h_0+\left[-\frac{m}{h_1^2}(q_2-q_1)-\gamma V'(\gamma q_1+(1-\gamma)q_2)\right]h_1=0.$$

We define $B(q_0,q_1,h_0)$ to be

$$B(q_0,q_1,h_0):=\left[m\,\frac{(q_1-q_0)}{h_0^2}-(1-\gamma)V'(\gamma q_0+(1-\gamma)q_1)\right]h_0,$$

and we let, for notational convenience, $u=\gamma q_1+(1-\gamma)q_2$. We then have

$$B=\left[2(1-\gamma)\frac{M(E_0-V(u))}{(u-q_1)^TM(u-q_1)}(u-q_1)+\gamma V'(u)\right]\sqrt{2\,\frac{(u-q_1)^TM(u-q_1)}{E_0-V(u)}}. \tag{IV.11}$$

In the particular case where $\gamma=\frac{1}{2}$ and q_1,q_2 are scalars, this expression simplifies to

$$B=\left[\frac{E_0-V(u)}{u-q_1}+\frac{1}{2}V'(u)\right]\sqrt{\frac{2(u-q_1)^2M}{E_0-V(u)}}. \tag{IV.12}$$

In other words,

$$E_0-V(u)+\frac{1}{2}V'(u)(u-q_1)-\frac{\sqrt{E_0-V(u)}}{\sqrt{2M}}B=0. \tag{IV.13}$$

We solve the previous equation for u. Then q_2 follows in a straightforward way.

V. TIME STEP SOLVABILITY AND AN OPTIMIZATION METHOD

We have seen previously that the condition $E_0-V(u)>0$ should be verified in order to compute the next time step h_1, given h_0. To clarify the exposition, we write $E_0=K_0+V_0$ and we take $V_1=V(u)$. It follows that

$$E_0-V(u)=K_0+V_0-V_1.$$

Our condition for solvability is therefore

$$K_0+V_0-V_1>0. \tag{V.1}$$

If K_0 is large and the time step small, then in this case (V.1) is automatically verified. Indeed, K_0 large and $V_0\simeq V_1$ implies that the computed kinetic energy is positive. If $q_1\simeq q_2$ and h_0 is not so small, this is a rather delicate situation but can be explored by writing

$$\frac{1}{2}\,\frac{(q_1-q_0)^Tm(q_1-q_0)}{h_0^2}+V(\gamma q_0+(1-\gamma)q_1)-V(\gamma q_1+(1-\gamma)q_2). \tag{V.2}$$

Taylor expanding the potential terms when $\gamma=\frac{1}{2}$ gives

$$V\left(\frac{q_0+q_1}{2}\right)-V\left(\frac{q_1+q_2}{2}\right)=-\frac{1}{2}V'\left(\frac{q_0+q_1}{2}\right)\frac{q_2-q_0}{2}. \tag{V.3}$$

When the kinetic term is small, the condition should reduce to the condition that q_2-q_0 has the same sign as $-V'((q_0+q_1)/2)$. This tells one in which direction q_2 should move.

(a) An optimization method. An alternative strategy to deal with this issue of how to compute the time steps near turning points, which is the one we adapt in this paper, is to adopt the following optimization technique. Given $h_0,\ q_0,\ q_1$ we have to find $h_1,\ q_2$ such that q_2 is determined by the DEL equations

$$g(q_0,q_1,q_2,h_0,h_1) := h_0 D_2 L(q_0,q_1,h_0)+h_1 D_1 L(q_1,q_2,h_1)=0 \tag{V.4}$$

and the energy condition

$$f(q_0,q_1,q_2,h_0,h_1) := E(q_1,q_2,h_1)-E(q_0,q_1,h_0)=0. \tag{V.5}$$

The basic equations we want to solve are thus

$$f(q_0,q_1,q_2,h_0,h_1)=0, \tag{V.6}$$

$$g(q_0,q_1,q_2,h_0,h_1)=0, \tag{V.7}$$

to be solved for the variables q_2 and h_1 as a function of q_0, q_1, and h_0. The technique we use is to minimize the quantity

$$\mathcal{B}=[f(q_0,q_1,q_2,h_0,h_1)]^2+[g(q_0,q_1,q_2,h_0,h_1)]^2 \tag{V.8}$$

over the variables $h_1,\ q_2$, with the other variables given, and *subject to the constraint* $h_1>0$. As above, this constraint means, in practice, that the computed kinetic energy is positive. Of course, this is then iterated and defines our algorithm as a map

$$(q_{k-1},q_k,h_{k-1})\mapsto(q_k,q_{k+1},h_k).$$

This method may be implemented in a standard way using a quasi-Newton algorithm, as in Ref. 27 (see also Ref. 28). Of course, other methods for efficiently solving the system of equations (VIII.9) can be considered as well, but as we have mentioned, we do not carry out any extensive comparitive or implementation tests in this paper. In the simple examples we do give in the section below, we use this optimization method.

VI. SYMPLECTIC NATURE OF THE ALGORITHM

We now show the sense in which the algorithm above is symplectic. This will be in the form of an identity that the mapping

$$\Phi: Q\times Q\times\mathbb{R}\to Q\times Q\times\mathbb{R}$$

defined by

$$(q_0,q_1,h_0)\mapsto(q_1,q_2,h_1),$$

where, as we have seen, q_2 and h_1 are defined by

$$h_0 D_2 L_d(q_0,q_1,h_0)+h_1 D_1 L_d(q_1,q_2,h_1)=0$$

and

$$E_d(q_0,q_1,h_0)=E_d(q_1,q_2,h_1),$$

where we recall that the discrete energy is defined by

$$E_d(q_0,q_1,h_0) = -D_{h_0}[h_0 L_d(q_0,q_1,h_0)].$$

To determine the symplectic nature of the mapping $\bar{\Phi}$, we follow the general line of reasoning in Ref. 1. Namely, we consider the action sum

$$S = [h_0 L_d(q_0,q_1,h_0) + h_1 L_d(q_1,q_2,h_1)]$$

and take its full differential as a function of all the variables, keeping in mind that h_1 and q_2 are functions of (q_0,q_1,h_0). Using the definition of the discrete energy, we get

$$dS = h_0 D_1 L_d(q_0,q_1,h_0)dq_0 + h_0 D_2 L_d(q_0,q_1,h_0)dq_1 + h_1 D_1 L_d(q_1,q_2,h_1)dq_1 + h_1 D_2 L_d(q_1,q_2,h_1)dq_2 - E_d(q_0,q_1,h_0)dh_0 - E_d(q_1,q_2,h_1)dh_1.$$

Because of the discrete Euler–Lagrange equations, this simplifies to

$$dS = h_0 D_1 L_d(q_0,q_1,h_0)dq_0 + h_1 D_2 L_d(q_1,q_2,h_1)dq_2 - E_d(q_0,q_1,h_0)dh_0 - E_d(q_1,q_2,h_1)dh_1.$$

In view of the equations defining the algorithm, we can write this as

$$dS = \Theta_L^- + \bar{\Phi}^* \Theta_L^+, \tag{VI.1}$$

where the one-forms Θ_L^- and Θ_L^+ are defined by

$$\Theta_L^-(q_0,q_1,h_0) = h_0 D_1 L_d(q_0,q_1,h_0)dq_0 - E_d dh_0$$

and

$$\Theta_L^+(q_0,q_1,h_0) = h_0 D_2 L_d(q_0,q_1,h_0)dq_1 - E_d dh_0.$$

Now notice that because of the definition of E_d, we have

$$\Theta_L^-(q_0,q_1,h_0) + \Theta_L^+(q_0,q_1,h_0) = d[h_0 L_d] - E_d dh_0. \tag{VI.2}$$

Substituting (VI.2) into (VI.1) gives

$$dS = d[h_0 L_d] - \theta_d^+ + \bar{\Phi}^* \Theta_L^+, \tag{VI.3}$$

where θ_d^+ is the discrete analog of the canonical one form, $p_i dq^i$, namely

$$\theta_d^+ = \Theta_d^+ + E_d dh_0 = h_0 D_2 L_d(q_0,q_1,h_0)dq_1.$$

Taking the differential of (VI.3), using $d^2=0$ and the fact that pull back commutes with the differential gives our final identity, namely

$$\bar{\Phi}^* \Omega_d = \omega_d, \tag{VI.4}$$

where

$$\Omega_d = -d\Theta_d^+$$

is the discrete analog of the space–time symplectic form and where

$$\omega_d = -d\theta_d^+$$

is the discrete analog of the phase space symplectic form. Notice that the identity (VI.4) is the discrete analog of the identity (III.5) in the continuous case. Thus, we may interpret the identity (VI.4) as the symplectic nature of the algorithm.

(a) Momentum conservation. We note that one proves conservation of momentum for algorithms invariant under a symmetry group in the same way as usual, following, Ref. 1; we need not repeat the argument.

VII. SUMMARY OF THE FEATURES OF THE ALGORITHM

In this section we summarize the three main features of the algorithm.

(1) The algorithm conserves energy.
(2) The algorithm is symplectic in the sense spelled out in the previous section.
(3) The algorithm conserves momentum.

We have designed it to preserve energy. The discrete version of the arguments given in the continuous case shows the ''space–time'' sense in which the algorithm is symplectic, as we have explained.

VIII. NUMERICAL EXAMPLES

The first example is one-dimensional and integrable and the second example consists of the first one coupled to an oscillator.

In each case, we will compare the constant time step method, which will show the orbits in phase space and variations in the energy as a function of time, with the corresponding results for the adaptive time step algorithm.

A. One degree of freedom example

We will use a Lagrangian that is of the standard form kinetic minus potential, namely

$$L(q,\dot{q})=\frac{m}{2}\dot{q}^2-V(q),$$

where q and $\dot{q}$ are real numbers, with the corresponding discrete Lagrangian (with $\gamma=\frac{1}{2}$) given by

$$L_d(q_0,q_1,h)=\frac{1}{2}m\left(\frac{q_1-q_0}{h}\right)^2-V\left(\frac{q_0+q_1}{2}\right),$$

where q_0, q_1, and $h>0$ are also real numbers. The corresponding energy, according to formula (IV.4) is given by

$$E_h(q_0,q_1)=\frac{1}{2}m\left(\frac{q_1-q_0}{h}\right)^2+V\left(\frac{q_0+q_1}{2}\right).$$

(a) Constant time step algorithm. We find q_2 using the DEL equations:

$$h[D_2L(q_0,q_1)+D_1L(q_1,q_2)]=0. \tag{VIII.1}$$

As $h\neq 0$, Eq. (VIII.1) leads to

$$\frac{m}{h^2}(q_1-q_0)-\frac{1}{2}V'\left(\frac{q_0+q_1}{2}\right)-\frac{m}{h^2}(q_2-q_1)-V'\left(\frac{q_1+q_2}{2}\right)=0 \tag{VIII.2}$$

to be solved for q_2. Keeping h fixed, this is the variational integrator that we use for the constant time step algorithm.

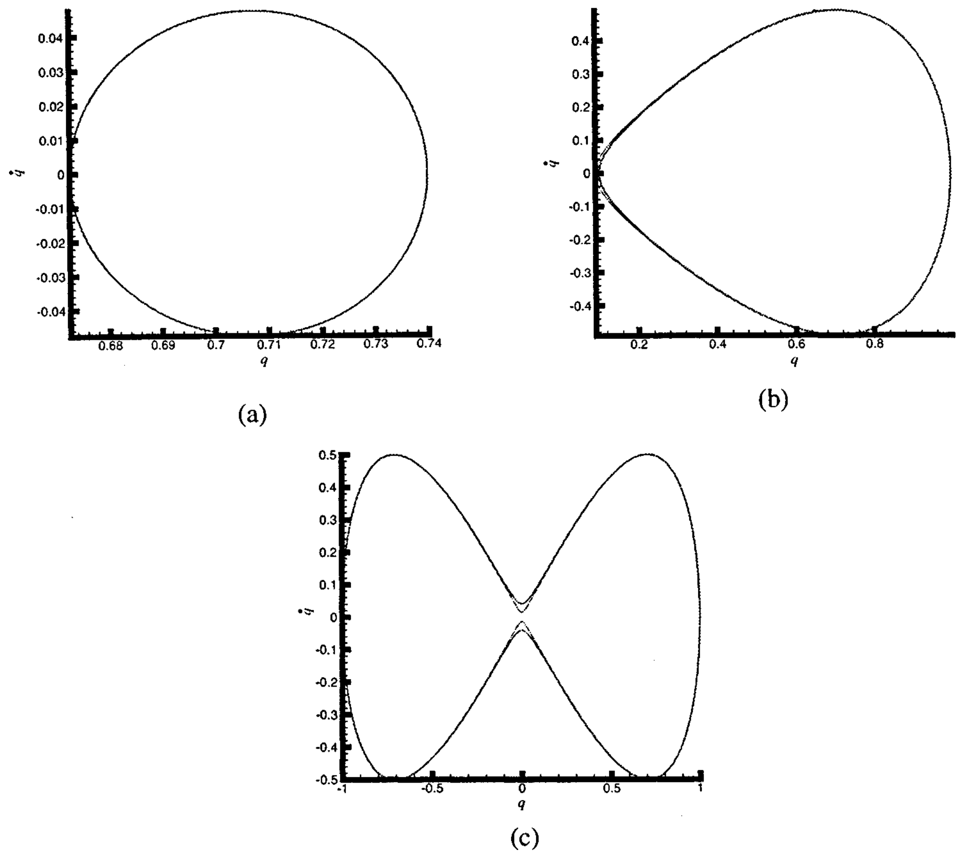

FIG. 1. Three initial conditions are studied for the particle in the double well potential. The orbits for both the fixed time step and the adaptive time step algorithms are plotted. The initial time step used in all cases is $h_0=0.1$. The initial data is (a) $q_0=q_1=0.74$, (b) $q_0=q_1=0.995$, and (c) $q_0=1.0$, $q_1=1.0$. The two orbits in each figure are nearly indistinguishable to the eye, but the adaptive time step computation is somewhat more accurate.

(b) Adaptive time step algorithm. Given $h_0,\ q_0,\ q_1$ we have to find $h_1,\ q_2$ such that q_2 is determined by the DEL equations

$$h_0 D_2 L(q_0,q_1,h_0)+h_1 D_1 L(q_1,q_2,h_1)=0 \tag{VIII.3}$$

and the energy condition

$$E(q_0,q_1,h_0)=E(q_1,q_2,h_1). \tag{VIII.4}$$

We write the energy condition as follows. Define

$$\begin{aligned} f(q_0,q_1,q_2,h_0,h_1)&=E(q_1,q_2,h_1)-E(q_0,q_1,h_0)\\ &=\frac{1}{2}m\,\frac{(q_2-q_1)^2}{h_1^2}+V\left(\frac{q_1+q_2}{2}\right)-E(q_0,q_1,h_0). \end{aligned} \tag{VIII.5}$$

The energy equation is written this way because $E_0 \coloneqq E(q_0,q_1,h_0)$ will have been computed and stored at the previous step.

The DEL equation (VIII.3) is written as follows:

$$g(q_0,q_1,q_2,h_0,h_1)=h_0 B_0+h_1\left[-m\,\frac{q_2-q_1}{h_1^2}-\frac{1}{2}V'\left(\frac{q_1+q_2}{2}\right)\right], \tag{VIII.6}$$

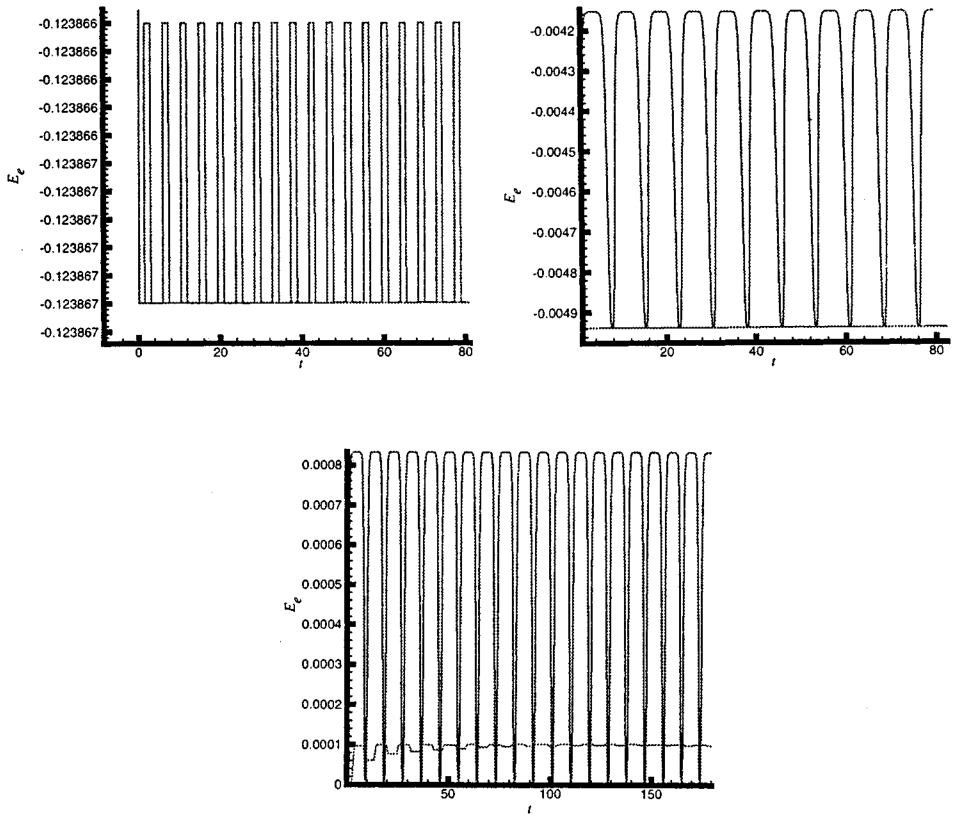

FIG. 2. The relatively large amplitude curve shows the (small) energy error for the constant time step algorithm as a function of time, while the lower curve shows the energy error for the adaptive time step algorithm. The initial data correspond to the three regions shown in the preceding figure.

where

$$B_0 = \left[m \frac{q_1 - q_0}{h_0^2} - \frac{1}{2} V' \left(\frac{q_0 + q_1}{2} \right) \right], \tag{VIII.7}$$

again, a quantity that will have been computed at the previous step.

(c) The numerical technique. The basic equations we want to solve are the following:

$$f(q_0, q_1, q_2, h_0, h_1) = 0, \tag{VIII.8}$$

$$g(q_0, q_1, q_2, h_0, h_1) = 0, \tag{VIII.9}$$

to be solved for the variables q_2 and h_1 as a function of q_0, q_1, and h_0. The technique we use is to minimize the quantity

$$\mathcal{B} = [f(q_0, q_1, q_2, h_0, h_1)]^2 + [g(q_0, q_1, q_2, h_0, h_1)]^2 \tag{VIII.10}$$

over the variables h_1, q_2, with the other variables given, and subject to the constraint $h_1 > 0$. As mentioned in the general theory, this method is implemented using descent methods, following Byrd *et al.*[27] and Zhu *et al.*[28]

(d) The double well potential. To illustrate the procedures, we choose $m = 1$ and take

$$V(q) = \tfrac{1}{2}(q^4 - q^2). \tag{VIII.11}$$

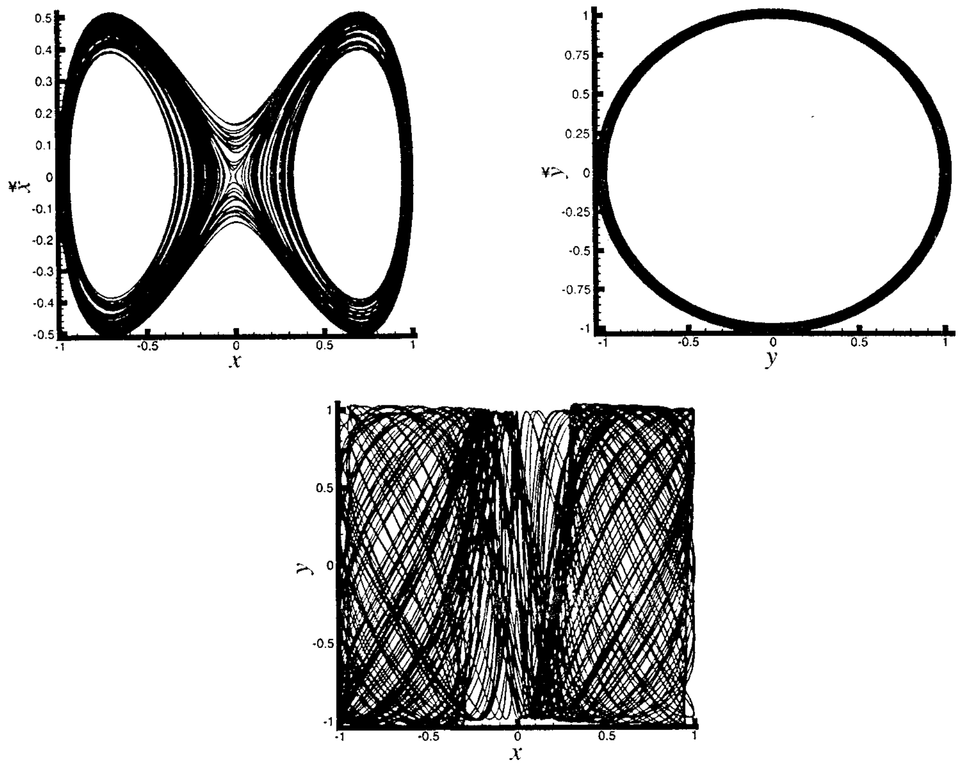

FIG. 3. An orbit in the coupled double well–oscillator system. In this plot, the initial conditions used were $x_0=y_0=x_1=y_1=1.00$, $h_0=0.1$, and $\epsilon=0.01$.

We study three regions of phase space as shown in Fig. 1. The axes show the computed position $q=(q_k+q_{k+1})/2$ and the computed velocity $\dot{q}=(q_{k+1}-q_k)/h_k$ as functions of time. The orbit in Fig. 1(a) is a periodic orbit that oscillates around the stable equilibrium position $q=1/\sqrt{2}$, $\dot{q}=0$. That in Fig. 1(b) is a periodic orbit with high period just inside the homoclinic orbit in the positive q half-space, while that in Fig. 1(c) is a periodic orbit just outside the homoclinic orbit.

The energy errors for both the constant time step and the adaptive time step algorithms are shown in Fig. 2. The amplitude in the variation of the energy depends on the time step. The smaller the time step, the smaller the amplitude, but we note that there is not a big difference in the periods. The same sort of behavior can also be seen in the corresponding plots in Ref. 6.

The small changes in the energy are, we believe, due to the effect of the turning points as we explained earlier. Of course, one can contemplate methods whereby these can be compensated or reduced further, but we do not explore these issues in this paper.

B. A two degree of freedom example

Now we consider an oscillator coupled with our previous double well potential example. The system now has chaotic orbits, so it is somewhat more interesting.

The continuous Lagrangian we choose is given by

$$L(x,y,\dot{x},\dot{y})=\tfrac{1}{2}\dot{x}^2-V(x)+\tfrac{1}{2}\dot{y}^2-\tfrac{1}{2}y^2+\varepsilon xy, \tag{VIII.12}$$

where ε introduces a small perturbation. This is a very simple example of a chaotic system arising as a perturbation of an integrable one. Shortly we will choose V to be the potential used in the preceding subsection.

Using the notation

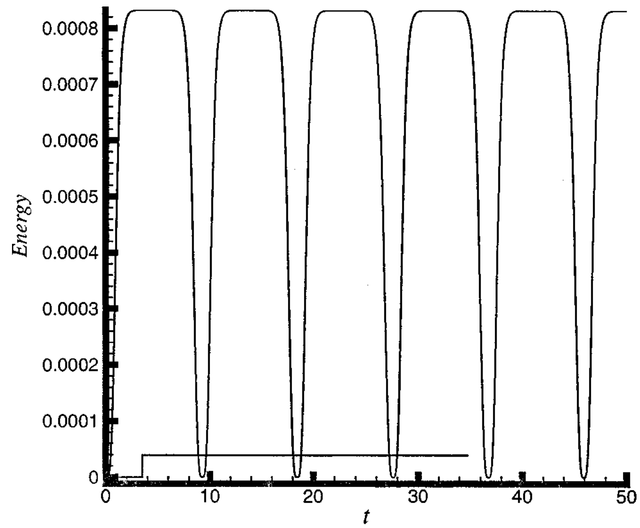

FIG. 4. The energy behavior for the variational versus the symplectic-energy method in the coupled double well–oscillator system.

$$q_0=(x_0,y_0),q_1=(x_1,y_1),$$

the corresponding discrete Lagrangian is

$$L_d(x_0,y_0,x_1,y_1,h_0)=\frac{1}{2}\frac{(x_1-x_0)^2}{h_0^2}-V\left(\frac{x_0+x_1}{2}\right)+\frac{1}{2}\frac{(y_1-y_0)^2}{h_0^2}-\frac{1}{2}\left(\frac{y_0+y_1}{2}\right)+\epsilon\left(\frac{x_0+x_1}{2}\right)\left(\frac{y_0+y_1}{2}\right). \tag{VIII.13}$$

Given $h_0,\ x_0,\ y_0,\ x_1,\ y_1$ we have to find $h_1,\ x_2,\ y_2$ such that the DEL equations

$$h_0D_2L(x_0,y_0,x_1,y_1,h_0)+h_1D_1L(x_1,y_1,x_2,y_2,h_1)=0 \tag{VIII.14}$$

and the energy condition

$$E(x_0,y_0,x_1,y_1,h_0)=E(x_1,y_1,x_2,y_2,h_1). \tag{VIII.15}$$

We write the energy condition in the form $f=0$ as follows. Define

$$\begin{aligned}f(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)&=E(x_1,y_1,x_2,y_2,h_1)-E(x_0,y_0,x_1,y_1,h_0)\\&=\frac{1}{2}\frac{(x_2-x_1)^2}{h_1^2}+V\left(\frac{x_1+x_2}{2}\right)+\frac{1}{2}\frac{(y_2-y_1)^2}{h_1^2}+\frac{1}{2}\left(\frac{y_1+y_2}{2}\right)^2\\&\quad-\epsilon\left(\frac{x_1+x_2}{2}\right)\left(\frac{y_1+y_2}{2}\right)-E(x_0,y_0,x_1,y_1,h_0).\end{aligned} \tag{VIII.16}$$

The DEL equation (VIII.14) is written in the form of a system $g=0$, $k=0$ as follows. Define g by

$$g(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)=h_0B_0+h_1\left[-\frac{x_2-x_1}{h_1^2}-\frac{1}{2}V'\left(\frac{x_1+x_2}{2}\right)+\frac{\epsilon}{2}\left(\frac{y_1+y_2}{2}\right)\right], \tag{VIII.17}$$

where

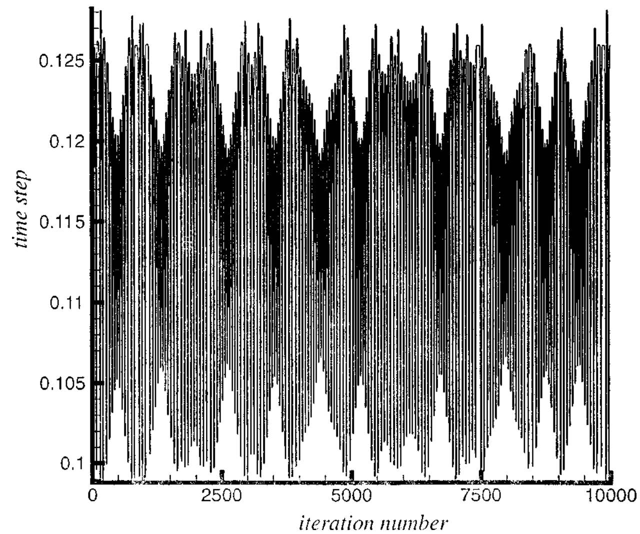

FIG. 5. The time step versus the iteration number for the symplectic-energy method.

$$B_0=\left[\frac{x_1-x_0}{h_0^2}-\frac{1}{2}V'\left(\frac{x_0+x_1}{2}\right)+\frac{\epsilon}{2}\left(\frac{y_0+y_1}{2}\right)\right]. \tag{VIII.18}$$

Define k by

$$k(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)=h_0C_0+h_1\left[-\frac{y_2-y_1}{h_1^2}-\frac{1}{2}\left(\frac{y_1+y_2}{2}\right)+\frac{\epsilon}{2}\left(\frac{x_1+x_2}{2}\right)\right], \tag{VIII.19}$$

where

$$C_0=\left[\frac{y_1-y_0}{h_0^2}-\frac{1}{2}\left(\frac{y_0+y_1}{2}\right)+\frac{\epsilon}{2}\left(\frac{x_0+x_1}{2}\right)\right]. \tag{VIII.20}$$

(e) The numerical technique. The basic equations we want to solve are the following:

$$f(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)=0, \tag{VIII.21}$$

$$g(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)=0,$$

$$k(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)=0, \tag{VIII.22}$$

to be solved for the variables x_2, y_2 and h_1 as a function of x_0, y_0, x_1, y_1 and h_0.

The technique used is to minimize the quantity

$$\mathcal{B}=[f(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)]^2+[g(x_0,y_0,x_1,y_1,x_2,y_2,h_0,h_1)]^2 \tag{VIII.23}$$

over the variables h_1, x_2, y_2, with the other variables given, and subject to the constraint $h_1>0$. This method is then implemented by the same method as in the preceding example.

(f) The double well potential coupled with an oscillator. To illustrate the procedures, we choose, as before,

$$V(x)=\tfrac{1}{2}(x^4-x^2). \tag{VIII.24}$$

We study a single chaotic orbit shown in Fig. 3. The different projections of the orbit, to the $x, \dot{x}$ and the $y, \dot{y}$ spaces and to the configuration space x, y, are shown.

Figure 4 shows the energy behavior, as before, for the standard variational integrator versus our symplectic-energy algorithm.

Finally, Fig. 5 shows how the time step varies with the iteration.

ACKNOWLEDGMENTS

We thank Debra Lewis, Robert McLachlan, George Patrick, Steve Shkoller, and Matt West for useful remarks.

[1] J. E. Marsden, G. W. Patrick, and S. Shkoller, ''Multisymplectic Geometry, Variational Integrators, and Nonlinear PDEs,'' Commun. Math. Phys. **199**, 351–395 (1998).

[2] Z. Ge and J. E. Marsden, ''Lie-Poisson integrators and Lie-Poisson Hamilton-Jacobi theory,'' Phys. Lett. A **133**, 134–139 (1988).

[3] A. P. Veselov, ''Integrable discrete-time systems and difference operators.'' Funct. Anal. Appl. **22**, 83–94 (1988).

[4] A. P. Veselov, ''Integrable Lagrangian correspondences and the factorization of matrix polynomials,'' Funct. Anal. Appl. **25**, 112–123 (1991).

[5] J. Moser and A. P. Veselov, ''Discrete versions of some classical integrable systems and factorization of matrix polynomials,'' Commun. Math. Phys. **139**, 217–243 (1991).

[6] J. M. Wendlandt and J. E. Marsden, ''Mechanical integrators derived from a discrete variational principle,'' Physica D **106**, 223–246 (1997).

[7] C. Kane, E. A. Repetto, M. Ortiz, and J. E. Marsden, ''Finite element analysis of nonsmooth contact,'' Comput. Meth. Appl. Mech. Eng. (to appear).

[8] J. E. Marsden, G. W. Patrick, and W. F. Shadwick, *Integration Algorithms and Classical Mechanics*, Fields Institute Communications, Vol. 10 (American Mathematical Society, Providence, RI, 1996).

[9] J. M. Sanz-Serna and M. Calvo, *Numerical Hamiltonian Problems* (Chapman and Hall, London, 1994).

[10] J. C. Simo and N. Tarnow, ''The discrete energy momentum method. Conserving algorithms for nonlinear elastodynamics,'' ZAMP **43**, 757–792 (1992).

[11] J. C. Simo, N. Tarnow, and K. K. Wong, ''Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics,'' Comput. Methods Appl. Mech. Eng. **100**, 63–116 (1992).

[12] O. Gonzalez, ''Time integration and discrete Hamiltonian systems,'' J. Nonlinear Sci. **6**, 449–468 (1996).

[13] Y. Shibberu, ''Time-discretization of Hamiltonian systems,'' Comput. Math. Appl. **28**, 123–145 (1994).

[14] H. R. Lewis and P. Kostelec, ''The use of Hamilton's principle to derive time-advance algorithms for ordinary differential equations,'' Comput. Phys. Commun. **96**, 129–151 (1996).

[15] T. D. Lee, ''Difference equations and conservation laws,'' J. Stat. Phys. **46**, 843–860 (1987).

[16] M. Ortiz, ''A note on energy conservation and stability of nonlinear time-stepping algorithms,'' Comput. Struct. **24**, 167–168 (1986).

[17] J. C. Simo and O. Gonzalez, ''Assessment of energy-momentum and symplectic schemes for stiff dynamical systems,'' in *Proceedings ASME Winter Annual Meeting*, New Orleans, December 1993.

[18] C. Kane, J. E. Marsden, M. Ortiz, and M. West, ''Variational and Newmark algorithms for conservative and dissipative systems'' (preprint).

[19] F. Armero and J. C. Simo, ''A New Unconditionally Stable Fractional Step Method for Non-Linear Coupled Thermomechanical Problems,'' Int. J. Numer. Methods Eng. **35**, 737–766 (1992).

[20] F. Armero and J. C. Simo, ''A-Priori Stability Estimates and Unconditionally Stable Product Formula Algorithms for Non-Linear Coupled Thermoplasticity,'' Int. J. Plast. **9**, 149–182 (1993).

[21] F. Armero and J. C. Simo, ''Long-Term Dissipativity of Time-Stepping Algorithms for an Abstract Evolution Equation with Applications to the Incompressible MHD and Navier-Stokes Equations,'' Comput. Methods Appl. Mech. Eng. **131**, 41–90 (1996).

[22] M. Ortiz and L. Stainier, ''The variational formulation of viscoplastic constitutive updates,'' Comput. Methods Appl. Mech. Eng. (to appear).

[23] J. E. Marsden and T. S. Ratiu, *Mechanics and Symmetry*, Texts in Applied Mathematics, Vol. 17, 2nd ed. (Springer-Verlag, New York, 1999).

[24] J. E. Marsden, S. Pekarsky, and S. Shkoller, ''The discrete Euler–Poincaré equation and reduction,'' preprint (1998).

[25] A. I. Bobenko and Y. B. Suris, ''Discrete time Lagrangian mechanics on Lie groups, with an application to the Lagrange top,'' preprint (1998).

[26] M. West, Accuracy of variational integrators (in preparation).

[27] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, ''A limited memory algorithm for bound constrained optimization,'' J. Sci. Comput. **16**, 1190–1208 (1995).

[28] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, ''L-BFGS-B: FORTRAN Subroutines for Large Scale Bound Constrained Optimization,'' Tech. Report, NAM-11, EECS Department, Northwestern University (1995).

Magnetohydrodynamic boundary layer on a flat plate: Further analytic results

Bhimsen K. Shivamoggi and David K. Rollins[a)]
Department of Mathematics, University of Central Florida, Orlando, Florida 32816

(Received 2 June 1998; accepted for publication 4 March 1999)

Further analytic results are deduced with the magnetohydrodynamic boundary layer equations for a flat plate. The asymptotic behavior of the solutions is deduced using the scaling group method. Then, an analytic perturbative procedure is used to determine an approximate solution that exhibits this asymptotic behavior. © *1999 American Institute of Physics.* [S0022-2488(99)02906-0]

I. INTRODUCTION

Greenspan and Carrier[1] considered the flow of a viscous, electrically conducting, incompressible fluid past a semi-infinite flat plate in the presence of a magnetic field which is uniform at infinity and parallel to the stream. They reduced the boundary-layer equations then to a pair of coupled nonlinear ordinary differential equations—

$$g'' + \mu(fg' - f'g) = 0, \tag{1}$$

$$f''' + ff'' - \frac{1}{A^2} gg'' = 0, \tag{2}$$

subject to the boundary conditions

$$\eta = 0: f = 0, \quad f' = 0, \quad g = 0, \tag{3a}$$

$$\eta \Rightarrow \infty: f \approx 2\eta, \quad g \approx 2\eta. \tag{3b}$$

Here, primes denote differentiation with respect to the independent variable $\eta = \frac{1}{2} y \sqrt{U/\nu x}$; y measures the distance from the plate, x is the distance along the plate from the leading edge, U is the undisturbed velocity, and ν is the kinematic viscosity. Further, if u and B_x are the x components of the velocity and magnetic fields, then

$$u = \tfrac{1}{2} U f'(\eta), \quad B_x = \tfrac{1}{2} B_0 g'(\eta),$$

B_0 being the ambient magnetic field intensity. Finally, $A \equiv U/V_{\mathrm{A}}$, where V_{A} is the Alfvén velocity $V_{\mathrm{A}} \equiv B_0/\sqrt{4\pi\rho}$, and $\mu \equiv 4\pi\sigma\nu$, ρ being the density and σ the electrical conductivity of the fluid.

For sub-Alfvénic flows $(A<1)$, disturbances travel upstream of the plate, invalidating the notion of a boundary layer originating at the leading edge of the plate. As Greenspan and Carrier[1] pointed out, this can be clearly seen by considering the case with infinite electrical conductivity, $\mu \Rightarrow \infty$. For this case, Eqs. (1) and (2) become $g = f$, $f''' + (1 + (1/A^2)) ff'' = 0$, so that one needs $A^2 > 1$ in order to preserve the usual boundary-layer situation. Reuter and Stewartson[2] showed that, for this case, the problem is mathematically ill posed, in the sense that it does not admit any solutions such that $f''(0) > 0$ and $g'(0) > 0$. Stewartson and Wilson[3] showed further that, even for certain values of $A > 1$, the solutions turn out to be nonunique whenever $\mu < 1$.

[a)]Electronic mail: drollins@pegasus.cc.ucf.edu

Equations (1) and (2) are, on the other hand, highly nonlinear and, therefore, one may not anticipate explicit analytical solutions for them. In this paper, we first deduce the asymptotic behavior of the solutions of Eqs. (1) and (2) using the scaling group method (Bluman and Kumei[4]). We then use an analytical perturbative procedure due to Bender *et al.*[5] to determine an approximate solution that has the above asymptotic behavior.

II. ASYMPTOTIC BEHAVIOR OF THE SOLUTION

In order to find the asymptotic behavior of the solutions of (1)–(3), note first that Eqs. (1) and (2) admit solutions of the form

$$f \sim \frac{1}{\eta}, \quad g \sim \frac{1}{\eta}; \tag{4}$$

(4) implies that Eqs. (1) and (2) have the scaling group

$$\overline{f} = \alpha^{-1} f, \quad \overline{g} = \alpha^{-1} g, \quad \overline{\eta} = \alpha \eta. \tag{5}$$

We may therefore introduce the following canonical coordinates

$$s = f\eta, \quad t = \eta^2 \frac{df}{d\eta}, \quad q = \eta g. \tag{6}$$

The transformation from (s,t) to (f,η) is given differentially by

$$\frac{ds}{t+s} = \frac{d\eta}{\eta}. \tag{7}$$

The transformation rules of the various derivatives are

$$\begin{gathered}
\frac{d^2 f}{d\eta^2} = -\frac{2t}{\eta^3} + \frac{1}{\eta^3}(t+s)\frac{dt}{ds}, \\
\frac{d^3 f}{d\eta^3} = \frac{6t}{\eta^4} - \frac{5}{\eta^4}(t+s)\frac{dt}{ds} + \frac{1}{\eta^4}(t+s)^2\frac{d^2 t}{ds^2} + \frac{1}{\eta^4}(t+s)\frac{dt}{ds}\left(\frac{dt}{ds}+1\right), \\
\frac{dg}{d\eta} = -\frac{q}{\eta^2} + \frac{1}{\eta^2}(t+s)\frac{dq}{ds}, \\
\frac{d^2 g}{d\eta^2} = \frac{2}{\eta^3}q - \frac{3}{\eta^3}(t+s)\frac{dq}{ds} + \frac{1}{\eta^3}(t+s)^2\frac{d^2 q}{ds^2} + \frac{1}{\eta^3}(t+s)\frac{dq}{ds}\left(\frac{dt}{ds}+1\right).
\end{gathered} \tag{8}$$

In terms of the new variables (s,t,q), the boundary-value problem (1)–(3) becomes

$$\begin{aligned}
&6t - 5(t+s)\frac{dt}{ds} + (t+s)^2\frac{d^2 t}{ds^2} + (t+s)\frac{dt}{ds}\left(\frac{dt}{ds}+1\right) - 2st + s(t+s)\frac{dt}{ds} \\
&\qquad - \frac{1}{A^2} q[2q - 3(t+s)]\frac{dq}{ds} + (t+s)^2\frac{d^2 q}{ds^2} + (t+s)\frac{dq}{ds}\left(\frac{dt}{ds}+1\right) = 0,
\end{aligned} \tag{9}$$

$$2q - 3(t+s)\frac{dq}{ds} + (t+s)^2\frac{d^2 q}{ds^2} + (t+s)\frac{dq}{ds}\left(\frac{dt}{ds}+1\right) + \mu\left[s(t+s)\frac{dq}{ds} - q(t+s)\right] = 0, \tag{10}$$

$$s = 0: t = 0, \quad q = 0, \tag{11}$$

$$s \Rightarrow \infty : t \Rightarrow \infty, \quad q \Rightarrow \infty. \tag{12}$$

Near $s=0$, Eqs. (9) and (10) show that

$$t \approx \lambda_1 s, \quad q \approx \lambda_2 s^p, \tag{13}$$

with

$$6\lambda_1 - 5\lambda_1(\lambda_1+1) + \lambda_1(\lambda_1+1)^2 \approx 0, \tag{14a}$$

$$2\lambda_2 - 3p\lambda_2(\lambda_1+1) + p(p-1)\lambda_2(\lambda_1+1)^2\lambda_2 p(\lambda_1+1)^2 \approx 0, \tag{15a}$$

or

$$\lambda_1 = 1,2, \tag{14b}$$

$$\lambda_1 = 2, \lambda_2 \ \text{ arbitrary and } \ p = \tfrac{1}{3}, \tfrac{2}{3}. \tag{15b}$$

$\lambda_1 = 1$ turns out to be the spurious root. The root $p=\frac{1}{3}$ is to be discarded because we require from (6) and (13) that $2p > 1$.

For $\lambda_1 = 2$, we obtain from (7) and (13),

$$s \sim \eta^3. \tag{16}$$

Using (16), we have from (6),

$$\eta \Rightarrow 0 : f \sim \eta^2, \quad g \sim \eta \tag{17}$$

Near $s \Rightarrow \infty$, equations (9) and (10) show that

$$t \approx \tilde{\lambda}_1 s, \quad q \approx \tilde{\lambda}_2 s, \tag{18}$$

with

$$-2\tilde{\lambda}_1 + \tilde{\lambda}_1(\tilde{\lambda}_1+1) - \frac{1}{A^2}\tilde{\lambda}_2[2\tilde{\lambda}_2 - 3\tilde{\lambda}_2(\tilde{\lambda}_1+1) + \tilde{\lambda}_2(\tilde{\lambda}_1+1)^2] \approx 0, \tag{19a}$$

$$2\tilde{\lambda}_2 - 3\tilde{\lambda}_2(\tilde{\lambda}_1+1) + \tilde{\lambda}_2(\tilde{\lambda}_1+1)^2 \approx 0, \tag{20a}$$

or

$$\tilde{\lambda}_1 = 1, \tag{19b}$$

$$\tilde{\lambda}_2 \ \text{ arbitrary.} \tag{20b}$$

Using (19) and (20), we obtain from (7) and (18),

$$s \sim \eta^2. \tag{21}$$

Using (21), we have from (6),

$$\eta \Rightarrow \infty : f \sim \eta, \quad g \sim \eta. \tag{22}$$

Observe that the asymptotic behavior of the solutions, as exhibited by (17) and (22), is independent of the Alfvén number A.

III. AN ANALYTIC PERTURBATIVE SOLUTION

We now use a perturbative procedure due to Bender *et al.*[5] to solve Eqs. (1) and (2) analytically. This method has been used recently (Shivamoggi and Rollins[6]) to solve the Kadomtsev equation for a heavy atom in a very strong magnetic field with very good results. We first replace Eqs. (1) and (2) by ones that contain a parameter δ, i.e.,

$$g'' + \mu[g' f^{\delta} - f' g^{\delta}] = 0, \tag{23}$$

$$f''' + f'' f^{\delta} - \frac{1}{A^2} g'' g^{\delta} = 0. \tag{24}$$

Note that Eqs. (1) and (2) are recovered when $\delta = 1$, and $\delta = 0$ corresponds to the linear zeroth-order approximation. By identifying δ as the perturbation parameter, the solution (f,g) is then expanded in a power series in δ,

$$\begin{aligned} f &= f_0 + \delta f_1 + \delta^2 f_2 + \cdots, \\ g &= g_0 + \delta g_1 + \delta^2 g_2 + \cdots. \end{aligned} \tag{25}$$

This then leads to a set of linear equations for (f_n, g_n): $O(1)$,

$$g_0'' + \mu(g_0' - f_0') = 0, \tag{26}$$

$$f_0''' + f_0'' - \frac{1}{A^2} g_0'' = 0; \tag{27}$$

$O(\delta)$,

$$g_1'' + \mu(g_1' - f_1') = -\mu(g_0' \cdot \ln f_0 - f_0' \cdot \ln g_0), \tag{28}$$

$$f_1''' + f_1'' - \frac{1}{A^2} g_1'' = -f_0'' \cdot \ln f_0 + \frac{1}{A^2} g_0'' \cdot \ln g_0, \tag{29}$$

etc.

Successive integrations of Eqs. (26) and (27), along with the use of (3), lead to

$$g_0' + \mu(g_0 - f_0) = \alpha, \tag{30}$$

$$f_0' + f_0 - \frac{1}{A^2} g_0 = c\eta, \tag{31}$$

where c is an arbitrary constant and

$$\alpha \equiv g_0'(0) > 0. \tag{32}$$

Using the boundary condition (3) at $\eta \Rightarrow \infty$, we obtain, from (31),

$$c = 2\epsilon, \tag{33}$$

where

$$\epsilon \equiv 1 - \frac{1}{A^2} > 0.$$

We have, from Eqs. (30) and (31),

$$f_0''+(1+\mu)f_0'+\epsilon\mu f_0=2\epsilon\mu\eta+2\epsilon+\alpha(1-\epsilon), \tag{34}$$

from which

$$f_0=(D_1e^{\sigma_1\eta}+D_2e^{\sigma_2\eta})-\frac{2\mu+(2-\alpha)(1-\epsilon)}{\mu\epsilon}+2\eta, \tag{35}$$

where

$$\sigma_{1,2}=\tfrac{1}{2}[-(1+\mu)\pm\sqrt{(1+\mu)^2-4\mu\epsilon}].$$

Using the boundary condition (3) at $\eta=0$, we obtain

$$D_1=-\frac{1}{\sigma_1-\sigma_2}\left[\sigma_2\frac{2\mu+(2-\alpha)(1-\epsilon)}{\mu\epsilon}+2\right],$$

$$D_2=\frac{1}{\sigma_1-\sigma_2}\left[\sigma_1\frac{2\mu+(2-\alpha)(1-\epsilon)}{\mu\epsilon}+2\right]. \tag{36}$$

We have from (32) and (35) and (36), for small η,

$$f_0\approx\tfrac{1}{2}[\alpha(1-\epsilon)+2\epsilon]\eta^2, \tag{37a}$$

$$g_0\approx\alpha\eta. \tag{37b}$$

One may also obtain (37) directly from Eqs. (30) and (31).

Observe from (37a) that, for super-Alfvénic flows ($\epsilon>0$), $f_0''(0)>0$, because

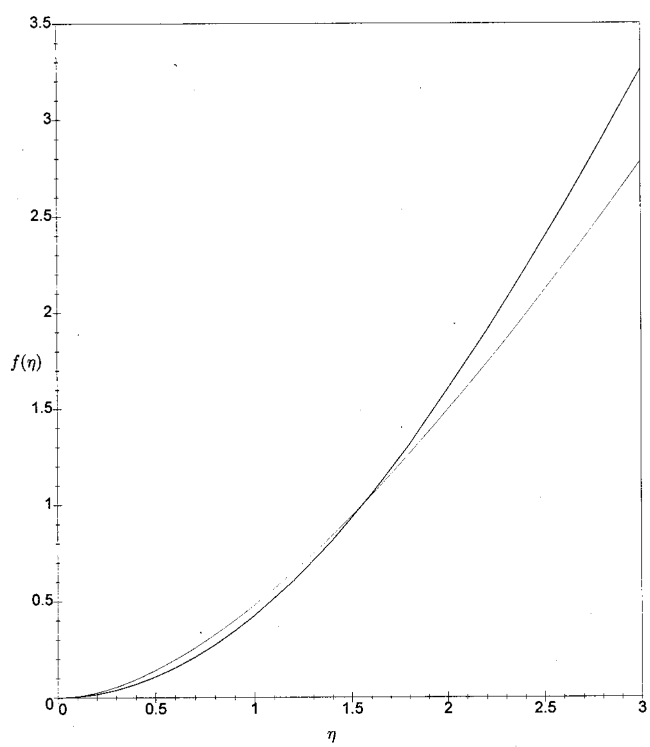

FIG. 1. Comparison of zeroth order approximate solution and numerical solution (bold) for nonlinear boundary value problem for magnetohydrodynamic case.

$$\alpha(1-\epsilon)+2\epsilon>0,$$

on noting (32) and that $(1-\epsilon)>0$.

Further, we have from (31), (33), and (35), for large η,

$$f_0 \approx 2\eta, \tag{38a}$$

$$g_0 \approx 2\eta. \tag{38b}$$

One may also obtain (38) directly from Eqs. (30) and (31).

The agreement of (37) with (17) on the one hand, and (38) with (3b) on the other hand, indicates that the asymptotic behavior (for *both* small and large η) of the solution of Eqs. (1) and (2) can be accurately provided by the linearized versions of the latter. Indeed, the linearized (or the zeroth-order) solution turns out to provide a reasonably accurate representation of the exact numerical solution of Eqs. (1) and (2) elsewhere as well.

In Fig. 1, the zeroth-order approximate analytic solution f_0 given in (35) is compared with the exact numerical solution of Eqs. (1) and (2). The agreement seems to be very good, even though f_0 is meant to be only a crude approximation to the exact solution. In fact, this feature is a carryover from the hydrodynamic case. In the latter case, (35) reduces to

$$f_0(\eta)=2e^{-\eta}-2+2\eta, \tag{39}$$

in agreement with the one given by Bender *et al.*[5] In Fig. 2, the zeroth-order approximate analytic solution given by (39) is compared with the exact numerical solution of the Blasius equation,

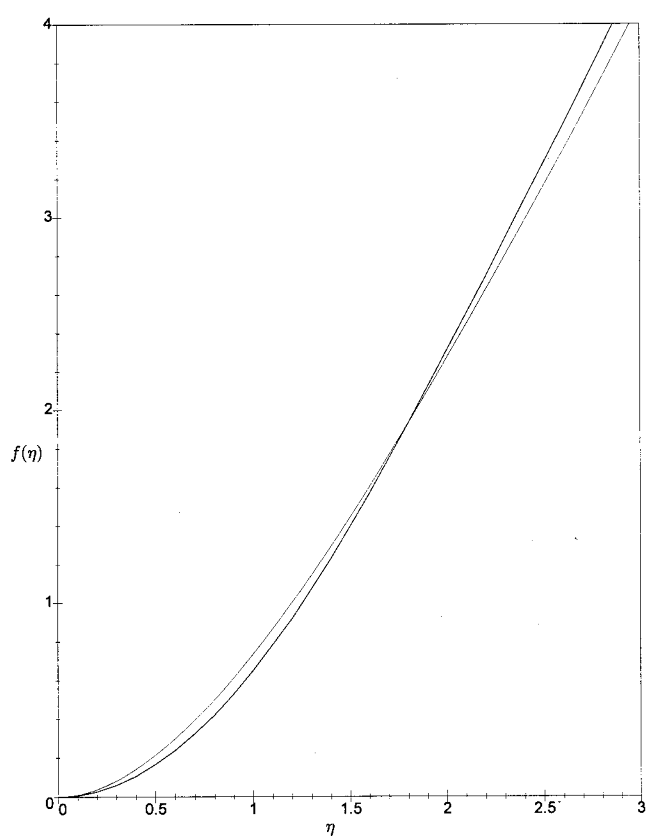

FIG. 2. Comparison of zeroth order approximate solution and numerical solution (bold) for nonlinear boundary value problem for hydrodynamic case.

$$f''' + ff'' = 0. \tag{40}$$

Again, the agreement seems to be very good.

Next, using (31) and (35) in the $O(\delta)$ equations (28) and (29), we see that a closed-form analytic solution of these equations becomes very difficult to find.

IV. DISCUSSION

In this paper, we have deduced further analytic results with the magnetohydrodynamic boundary layer equations (1) and (2) for a flat plate. We first derived the asymptotic behavior of the solutions using the scaling group method (Bluman and Kumei[4]). We then sought to use an analytic perturbative procedure due to Bender *et al.*[5] to determine an approximate solution. However, the linearized (or the zeroth-order) solution of the boundary-layer equations (1) and (2) turned out to provide not only the required asymptotic behavior (for both small and large η) of the exact numerical solution of Eqs. (1) and (2), but also a reasonably accurate representation of the exact numerical solution elsewhere.

ACKNOWLEDGMENTS

The authors are indebted to the referee for his valuable remarks that contributed to making the presentation of the material more accurate.

[1] H. P. Greenspan and G. F. Carrier, J. Fluid Mech. **6**, 77 (1959).
[2] G. E. H. Reuter and K. Stewartson, Phys. Fluids **4**, 276 (1961).
[3] K. Stewartson and D. H. Wilson, J. Fluid Mech. **18**, 337 (1964).
[4] G. W. Bluman and S. Kumei: *Group Symmetries and Differential Equations* (Springer-Verlag, Berlin, 1989).
[5] C. M. Bender, K. A. Milton, S. S. Pensky, and L. M. Simmons, Jr., J. Math. Phys. **30**, 1447 (1989).
[6] B. K. Shivamoggi and D. K. Rollins, J. Phys. A **30**, 3681 (1997).

The Weierstrass–Enneper system for constant mean curvature surfaces and the completely integrable sigma model

P. Bracken[a)] and A. M. Grundland[b)]
Centre de Recherches Mathématique, Université de Montréal, Montréal, Québec H3C 3J7, Canada

L. Martina[c)]
Department of Physics, University of Lecce, 73100 Lecce, Italy

(Received 27 July 1998; accepted for publication 5 March 1999)

The integrability of a system which describes constant mean curvature surfaces by means of the adapted Weierstrass–Enneper inducing formula is studied. This is carried out by using a specific transformation which reduces the initial system to the completely integrable two-dimensional Euclidean nonlinear sigma model. Through the use of the apparatus of differential forms and Cartan theory of systems in involution, it is demonstrated that the general analytic solutions of both systems possess the same degree of freedom. Furthermore, a new linear spectral problem equivalent to the initial Weierstrass–Enneper system is derived via the method of differential constraints. A new procedure for constructing solutions to this system is proposed and illustrated by several elementary examples, including a multi-soliton solution. © *1999 American Institute of Physics.* [S0022-2488(99)03307-1]

I. INTRODUCTION

Since the last century, the problems of surfaces and their deformations under various types of dynamics have generated a great deal of interest and activity in several mathematical as well as physical fields of research.[1–9] In particular, surfaces with constant mean curvature have been shown to play an essential role in several applications to nonlinear phenomena in such areas of physics as two-dimensional gravity,[4,10] quantum field theory,[4,11] statistical physics,[3,12] and fluid dynamics.[13,14] The Weierstrass–Enneper formula for inducing minimal surfaces has been studied for several years,[15–17] most recently by B. Konopelchenko and I. Taimanov.[18,19] They established a direct connection between certain classes of constant curvature surfaces and an integrable finite-dimensional Hamiltonian system (for a summary of their results, see Ref. 19). In general, it was shown by B. Konopelchenko[18] that the following infinite-dimensional Hamiltonian system describes constant mean curvature surfaces,

$$\partial\psi_1 = 2H(|\psi_1|^2 + |\psi_2|^2)\psi_2, \quad \bar{\partial}\psi_2 = -2H(|\psi_1|^2 + |\psi_2|^2)\psi_1, \tag{1.1}$$

where ψ_1 and ψ_2 are complex functions of the complex variables $(z,\bar{z})$. The bar denotes the complex conjugate, $\partial = \partial/\partial z$ and $\bar{\partial} = \partial/\partial\bar{z}$, and H denotes the constant mean curvature of the surface. One can assume, without loss of generality, $H = \frac{1}{2}$. Then, system (1.1) takes the form

$$\partial\psi_1 = p\psi_2, \quad \bar{\partial}\psi_2 = -p\psi_1, \quad p = |\psi_1|^2 + |\psi_2|^2, \tag{1.2a}$$

and its respective complex conjugate is

[a)]Electronic mail: bracken@CRM.UMontreal.ca
[b)]Electronic mail: grundlan@CRM.UMontreal.ca
[c)]Electronic mail: martina@le.infn.it

$$\overline{\partial \psi_1} = p \overline{\psi_2}, \quad \overline{\partial \psi_2} = -p \overline{\psi_1}. \tag{1.2b}$$

The above system can be considered a variant of the original Weierstrass–Enneper (WE) system and we will refer to it as such.

The system (1.2) determines a set of constant mean curvature surfaces obtained by the following parametrization $(z,\bar{z}) \to (X_1(z,\bar{z}), X_2(z,\bar{z}), X_3(z,\bar{z}))$

$$X_1 + iX_2 = 2i \int_{z_0}^{z} (\overline{\psi_1^2}\, dz' - \overline{\psi_2^2} d\overline{z'}),$$

$$X_1 - iX_2 = 2i \int_{z_0}^{z} (\psi_2^2\, dz' - \psi_1^2 d\overline{z'}), \tag{1.3}$$

$$X_3 = -2 \int_{z_0}^{z} (\psi_2 \overline{\psi_1}\, dz' + \psi_1 \overline{\psi_2}\, d\overline{z'}).$$

Using the standard formulas, we find that the first fundamental form on the surface is given by

$$\Omega = 4p^2\, dz d\bar{z}, \tag{1.4}$$

and the Gaussian curvature is[18]

$$K = -\frac{\partial \bar{\partial} (\ln p)}{p^2} \tag{1.5}$$

in isothermic coordinates.

The results obtained in Ref. 19 give a certain indication suggesting complete integrability of the WE system (1.2). However, a systematic approach to its integrability still remains an open problem. It will be shown here that the WE system (1.2) passes the Painlevé test, which means that it satisfies the necessary condition for complete integrability. This fact will be deduced from the existence of a linear spectral problem for the WE system (1.2). Moreover, the Lie algebra of infinitesimal symmetries of the WE system (1.2) is spanned by the vector fields

$$\alpha_1 = \partial, \quad \alpha_2 = \bar{\partial}, \quad \beta = \psi_1 \partial_{\psi_1} - \overline{\psi_1} \partial_{\overline{\psi_1}} + \psi_2 \partial_{\psi_2} - \overline{\psi_2} \partial_{\overline{\psi_2}} \tag{1.6}$$

and

$$\alpha_\xi = \xi \partial + \overline{\xi \partial} - \tfrac{1}{2} [\psi_1 (\overline{\partial \xi}) \partial_{\psi_1} + \overline{\psi_1} (\partial \xi) \partial_{\overline{\psi_1}} + \psi_2 (\partial \xi) \partial_{\psi_2} + \overline{\psi_2} (\overline{\partial \xi}) \partial_{\overline{\psi_2}}], \tag{1.7}$$

where $\xi(z)$ is an arbitrary analytic function and $\bar{\xi}(\bar{z})$ denotes its complex conjugate. Then, the set $\{\alpha_\xi\}$ generates an infinite-dimensional Lie algebra, realizing the conformal symmetry property of the system (1.2). This algebra contains a Virasoro subalgebra,[20] which has translations and dilations as special elements and sl(2) as unique simple subalgebras. The vector field β commutes with the α_ξ and it describes a scaling transformation involving only the dependent variables ψ_i and $\bar{\psi}_i$.

To obtain a one-parameter subgroup, we integrate α_ξ. Thus, we obtain the transformations

$$z'(\lambda, z) = F^{-1}(\lambda + F(z)) = \int_{z_0}^{z} \frac{\xi[z'(\lambda, w)]}{\xi(w)} dw,$$

$$\psi_1'(\lambda, \psi_1) = \psi_1 \left(\frac{\bar{\xi}(\bar{z})}{\xi(\overline{z'}(\lambda, z))} \right)^{1/2}, \quad \psi_2'(\lambda, \psi_2) = \psi_2 \left(\frac{\xi(z'(\lambda, z))}{\xi(z)} \right)^{1/2},$$

where

$$F(z)=\int_c^z \frac{ds}{\xi(s)},$$

with z_0 and c suitable complex numbers. Although the existence of a conformal symmetry does not imply in general complete integrability, it is, however, a strong indication in this direction.

Moreover, the WE system (1.2) possesses several conserved quantities.[10] The conservation of current which is defined by

$$J=\bar{\psi}_1\partial\psi_2-\psi_2\partial\bar{\psi}_1 \tag{1.8}$$

leads to interesting consequences. Differentiation of J gives

$$\begin{aligned}\bar{\partial}J=\bar{\partial}(\bar{\psi}_1\partial\psi_2-\psi_2\partial\bar{\psi}_1)&=(\overline{\partial\psi_1})\partial\psi_2+\bar{\psi}_1(\bar{\partial}\partial\psi_2)-(\bar{\partial}\psi_2)(\partial\bar{\psi}_1)-\psi_2(\bar{\partial}\partial\bar{\psi}_1)\\&=p\bar{\psi}_2\partial\psi_2+\bar{\psi}_1(\bar{\partial}\partial\psi_2)+p\psi_1\partial\bar{\psi}_1-\psi_2(\bar{\partial}\partial\bar{\psi}_1).\end{aligned} \tag{1.9}$$

The mixed derivatives obtained from (1.2) are

$$\begin{aligned}\partial\overline{\partial\psi_1}&=\partial(p\bar{\psi}_2)=(\partial p)\bar{\psi}_2+p\partial\bar{\psi}_2=(\partial p)\bar{\psi}_2-p^2\bar{\psi}_1,\\ \partial\bar{\partial}\psi_2&=-\partial(p\psi_1)=-(\partial p)\psi_1-p\partial\psi_1=-(\partial p)\psi_1-p^2\psi_2.\end{aligned} \tag{1.10}$$

Consequently, the derivative of J vanishes,

$$\begin{aligned}\bar{\partial}J&=p\bar{\psi}_2\partial\psi_2-|\psi_1|^2(\partial p)-p^2\bar{\psi}_1\psi_2+p\psi_1\partial\bar{\psi}_1-|\psi_2|^2(\partial p)+p^2\bar{\psi}_1\psi_2\\&=-p(\partial p)+p(\bar{\psi}_2\partial\psi_2+\psi_1\partial\bar{\psi}_1)=-p(\partial p)+p(\partial p)=0.\end{aligned} \tag{1.11}$$

Note that $\bar{\partial}J=0$ holds even when no restriction has been placed on ∂p. Exactly the same situation occurs for the conjugate equation, $\partial\bar{J}=0$. This means that the current J is an entire function.

In this paper, we examine certain aspects of complete integrability of the WE system (1.2) in the context of a two-dimensional Euclidean sigma model. In particular, we focus on constructing a linear spectral problem for this system where the explicit form has not been known up to now.

This paper is organized as follows. In Sec. II, we perform the reduction of the original system to a certain second-order system of partial differential equations (PDEs). In Sec. III, we present an estimation of the degree of freedom of the general analytic solutions of both systems. This analysis is carried out by means of the Cartan theory of systems in involution. In Sec. IV, a linear spectral problem is derived for the WE system via a two-dimensional nonlinear sigma model based on the related second-order system. This procedure amounts to a new technique for generating certain classes of solutions of the WE system which is illustrated with several examples in Sec V. Section VI contains final remarks and possible future developments.

II. THE SECOND-ORDER SYSTEM ASSOCIATED WITH THE WEIERSTRASS–ENNEPER SYSTEM

In our investigation of the integrability of the WE system (1.2), we subject it to several transformations in order to simplify its structure.

We start by introducing the new complex variable

$$\rho=\frac{\psi_1}{\psi_2}. \tag{2.1}$$

Using Eqs. (1.2) and the relation $p=|\psi_2|^2(1+|\rho|^2)$, one obtains

$$\partial\rho=\frac{\partial\psi_1}{\psi_2}-\frac{\psi_1}{\psi_2^2}\partial\bar{\psi}_2=\frac{p\psi_2}{\psi_2}-\frac{\psi_1}{\psi_2^2}(-p\bar{\psi}_1)=\frac{p^2}{\psi_2^2}=(1+|\rho|^2)^2\psi_2^2. \tag{2.2}$$

Note that $\partial\rho$ and ψ_2^2 are related by a real function $(1+|\rho|^2)^2$. Consequently, they have the same polar angle in the complex plane. Dividing (2.2) by $(1+|\rho|^2)^2$ and taking the principal square root, one obtains ψ_2. The complex conjugate of ψ_2 is found in the usual way by reflecting through the real axis. Using (2.1), ψ_1 can be obtained from the product of ρ and $\bar{\psi}_2$. This generates the following transformation from the variable ρ into the set of variables ψ_i,

$$\psi_1=\epsilon\rho\frac{(\overline{\partial\rho})^{1/2}}{1+|\rho|^2},\quad \psi_2=\epsilon\frac{(\partial\rho)^{1/2}}{1+|\rho|^2}\quad (\epsilon^2=1). \tag{2.3}$$

Let us now state the following proposition.

Proposition 1: If ψ_1 and ψ_2 are solutions of the system (1.2), then the function ρ defined by (2.1) is a solution of the following second-order system,

$$\partial\bar{\partial}\rho-\frac{2\bar{\rho}}{1+|\rho|^2}\partial\rho\bar{\partial}\rho=0, \tag{2.4a}$$

$$\partial\bar{\partial}\bar{\rho}-\frac{2\rho}{1+|\rho|^2}\partial\bar{\rho}\bar{\partial}\bar{\rho}=0. \tag{2.4b}$$

Proof: Differentiation of Eq. (2.1) with respect to $\bar{z}$ yields

$$\bar{\partial}\rho=\frac{\bar{\partial}\psi_1}{\bar{\psi}_2}-\frac{\psi_1}{\bar{\psi}_2^2}\overline{\partial\psi_2}=(\bar{\psi}_2)^{-2}[\bar{\psi}_2\bar{\partial}\psi_1-\psi_1\overline{\partial\psi_2}]. \tag{2.5}$$

By an easy computation, one obtains from (2.2) and (2.5)

$$\begin{aligned}\partial\bar{\partial}\rho&=\frac{\partial\bar{\partial}\psi_1}{\bar{\psi}_2}-\frac{\bar{\partial}\psi_1}{(\bar{\psi}_2)^2}\partial\bar{\psi}_2-\frac{\partial\psi_1}{(\bar{\psi}_2)^2}\overline{\partial\psi_2}+2\frac{\psi_1}{(\bar{\psi}_2)^3}\partial\bar{\psi}_2\overline{\partial\psi_2}-\frac{\psi_1}{(\bar{\psi}_2)^2}\partial\overline{\partial\psi_2}\\&=(\bar{\psi}_2)^{-3}[\bar{\psi}_2^2(\partial\bar{\partial}\psi_1)-\bar{\psi}_2\bar{\partial}\psi_1\partial\bar{\psi}_2-\bar{\psi}_2\partial\psi_1\overline{\partial\psi_2}+2\psi_1(\partial\bar{\psi}_2)(\overline{\partial\psi_2})-\psi_1\bar{\psi}_2(\partial\overline{\partial\psi_2})]\\&=(\bar{\psi}_2)^{-3}[\bar{\psi}_2^2(\partial\bar{\partial}\psi_1)+p\bar{\psi}_2\bar{\psi}_1\bar{\partial}\psi_1-p|\psi_2|^2\overline{\partial\psi_2}-2p|\psi_1|^2\overline{\partial\psi_2}-\psi_1\bar{\psi}_2(\partial\overline{\partial\psi_2})],\end{aligned} \tag{2.6}$$

and its respective complex conjugate equation is

$$\partial\bar{\partial}\bar{\rho}=(\psi_2)^{-3}[\psi_2^2(\bar{\partial}\partial\psi_1)+p\psi_2\psi_1\partial\bar{\psi}_1-p|\psi_2|^2\partial\psi_2-2p|\psi_1|^2\partial\psi_2-\bar{\psi}_1\psi_2(\bar{\partial}\partial\psi_2)]. \tag{2.7}$$

Using (1.2) the second derivatives (1.9) become

$$\begin{aligned}\partial\overline{\partial\psi_1}&=(\psi_1\partial\bar{\psi}_1+\bar{\psi}_2\partial\psi_2)\bar{\psi}_2-p^2\bar{\psi}_1,\\ \bar{\partial}\partial\psi_1&=(\bar{\psi}_1\bar{\partial}\psi_1+\psi_2\overline{\partial\psi_2})\psi_2-p^2\psi_1,\\ \partial\bar{\partial}\psi_2&=-(\psi_1\partial\bar{\psi}_1+\bar{\psi}_2\partial\psi_2)\psi_1-p^2\psi_2,\\ \bar{\partial}\partial\bar{\psi}_2&=-(\bar{\psi}_1\bar{\partial}\psi_1+\psi_2\overline{\partial\psi_2})\bar{\psi}_1-p^2\bar{\psi}_2.\end{aligned} \tag{2.8}$$

Substituting (2.8) into (2.6) and (2.7), the following compact formulas for the mixed ρ derivatives can be obtained:

$$\partial\bar{\partial}\rho = \frac{2\bar{\psi}_1 p}{\psi_2^3}\left(\bar{\psi}_2\bar{\partial}\psi_1 - \psi_1\overline{\partial\psi_2}\right),$$
$$\bar{\partial}\partial\bar{\rho} = \frac{2\psi_1 p}{\psi_2^3}\left(\psi_2\partial\bar{\psi}_1 - \bar{\psi}_1\partial\psi_2\right). \tag{2.9}$$

Substituting (2.1), (2.2), (2.5), and (2.9) into the left-hand side of (2.4a), one obtains

$$\begin{aligned}\partial\bar{\partial}\rho - \frac{2\bar{\rho}}{1+|\rho|^2}\partial\rho\bar{\partial}\rho &= \frac{2\bar{\psi}_1 p}{\psi_2^3}\left(\bar{\psi}_2\bar{\partial}\psi_1 - \psi_1\overline{\partial\psi_2}\right) - \frac{2\bar{\psi}_1\psi_2}{\psi_2^2}(1+|\rho|^2)\left(\bar{\psi}_2\bar{\partial}\psi_1 - \psi_1\overline{\partial\psi_2}\right)\\ &= \frac{2\bar{\psi}_1 p}{\psi_2^3}\left(\bar{\psi}_2\bar{\partial}\psi_1 - \psi_1\overline{\partial\psi_2}\right) - \frac{2\bar{\psi}_1 p}{\psi_2^3}\left(\bar{\psi}_2\bar{\partial}\psi_1 - \psi_1\overline{\partial\psi_2}\right) = 0.\end{aligned}$$

An analogous result holds for the conjugate equation (2.4b). Q.E.D.

Similar formulas to those of (2.4) can be found in the literature[10,11] in the context involving conformal immersions of a Riemann surface in $\mathbb{R}^n$.

The converse of Proposition 1 can be formulated as follows.

Proposition 2: If ρ is a solution to the system (2.4), then the functions ψ_1 and ψ_2 defined by (2.3) in terms of ρ satisfy the WE system (1.2).

Proof: Differentiating (2.3) with respect to z, we obtain

$$\partial\psi_1 = \epsilon\left\{\partial\rho\frac{(\bar{\partial}\bar{\rho})^{1/2}}{1+|\rho|^2} + \rho\frac{(\bar{\partial}\bar{\rho})^{\sqrt{-1/2}}}{2(1+|\rho|^2)}\partial\bar{\partial}\bar{\rho} - \rho\frac{(\bar{\partial}\bar{\rho})^{1/2}}{(1+|\rho|^2)^2}(\rho\partial\bar{\rho} + \bar{\rho}\partial\rho)\right\}.$$

Substituting Eq. (2.4) into this expression, we get

$$\partial\psi_1 = \epsilon\left\{\partial\rho\frac{(\bar{\partial}\bar{\rho})^{1/2}}{1+|\rho|^2} - |\rho|^2\frac{(\bar{\partial}\bar{\rho})^{1/2}}{(1+|\rho|^2)^2}\partial\rho\right\} = \frac{(\partial\rho\bar{\partial}\bar{\rho})^{1/2}}{(1+|\rho|^2)}\psi_2. \tag{2.10}$$

Multiplying both equations in (2.3) together, the following expression for p results:

$$p = \frac{(\partial\rho\bar{\partial}\bar{\rho})^{1/2}}{(1+|\rho|^2)}. \tag{2.11}$$

Therefore,

$$\partial\psi_1 = \frac{(\partial\rho\bar{\partial}\bar{\rho})^{1/2}}{(1+|\rho|^2)}\psi_2 = p\psi_2.$$

Similarly, differentiation of (2.3) with respect to $\bar{z}$ gives

$$\bar{\partial}\psi_2 = \epsilon\left\{\frac{(\partial\rho)^{-1/2}}{2(1+|\rho|^2)}\bar{\partial}\partial\rho - \frac{(\partial\rho)^{1/2}}{(1+|\rho|^2)^2}(\rho\bar{\partial}\bar{\rho} + \bar{\rho}\bar{\partial}\rho)\right\}. \tag{2.12}$$

Substituting (2.4) and (2.11) into (2.12), we obtain

$$\bar{\partial}\psi_2 = -\frac{(\bar{\partial}\bar{\rho}\partial\rho)^{1/2}}{(1+|\rho|^2)}\psi_1 = -p\psi_1,$$

which completes the proof. Q.E.D.

In some cases, it is more convenient to deal with (2.4) than the original system (1.2), since it consists of only two equations for two dependent variables ρ and $\bar{\rho}$. For example, a very large class of solutions of (2.4) can be found simply by requiring the holomorphicity ($\bar{\partial}\rho=0$) or antiholomorphicity ($\partial\rho=0$) of the function ρ. We will show later in Sec. V some examples of this type of solution. In the context of differential geometry, the system (2.4) was introduced by Kenmotsu in his seminal paper,[15] and then often used by subsequent authors.[10,18,19]

It is worth noting that, as in the case of system (1.2), the classical symmetry groups of (2.4) are conformal and scaling transformations. The corresponding symmetry algebra is spanned by

$$\alpha_1=\xi(z)\partial, \quad \alpha_2=\eta(\bar{z})\bar{\partial}, \quad \alpha_3=\rho\partial_p-\bar{\rho}\partial_{\bar{\rho}}, \tag{2.13}$$

where ξ and η are arbitrary functions of their arguments. This algebra can be decomposed as a direct sum of two infinite-dimensional simple Lie subalgebras with direct sum a one-dimensional algebra generated by α_3. Assuming that the functions ξ and η are analytic in a proper open subset Ω of $\mathbb{C}$, they can be developed in a Laurent series so we can provide a base for two centerless Virasoro algebras. Finite-dimensional subalgebras are spanned by $\{\partial\}$, $\{\partial,z\partial\}$, $\{\partial,z\partial,z^2\partial\}$,... and $\{\bar{\partial}\}$, $\{\bar{\partial},\bar{z}\bar{\partial}\}$, $\{\bar{\partial},\bar{z}\bar{\partial},\bar{z}^2\bar{\partial}\}$,..., respectively. In particular, the invariants of the one-dimensional subalgebra $\{\partial\}$ are given by $\{\bar{z},\rho\}$. Then, the invariant solutions are any antiholomorphic functions ρ of $\bar{z}$. A detailed study of solutions invariant under vector fields (2.13) is beyond the scope of the present work, but there is no difficulty in treating them.

Finally, an interesting feature of the WE system (1.2) can be derived from the Gaussian curvature (1.5). It can be expanded in the following way:

$$p^2K=-\bar{\partial}\partial\ln(p)=-\bar{\partial}\left(\frac{1}{p}\partial p\right)=\frac{1}{p^2}(\partial p)(\bar{\partial}p)-\frac{1}{p}\bar{\partial}\partial p.$$

Using the system of equations (1.2), the differentiation of the function p with respect to z and $\bar{z}$, respectively, yields

$$\partial p=\psi_1(\partial\bar{\psi}_1)+\bar{\psi}_2(\partial\psi_2), \quad \bar{\partial}p=\bar{\psi}_1(\bar{\partial}\psi_1)+\psi_2(\overline{\partial\psi_2}). \tag{2.14}$$

The mixed derivative of p becomes

$$\bar{\partial}\partial p=\bar{\partial}\psi_1\partial\bar{\psi}_1+\psi_1\bar{\partial}\partial\bar{\psi}_1+\bar{\psi}_2\bar{\partial}\partial\psi_2+\partial\psi_2\overline{\partial\psi_2}=\bar{\partial}\psi_1\partial\bar{\psi}_1+\partial\psi_2\overline{\partial\psi_2}-p^3. \tag{2.15}$$

The product of the derivatives (2.14) is given by

$$\bar{\partial}p\,\partial p=|\psi_1|^2(\bar{\partial}\psi_1)(\partial\bar{\psi}_1)+\psi_1\psi_2(\partial\bar{\psi}_1)(\overline{\partial\psi_2})+\bar{\psi}_1\bar{\psi}_2(\bar{\partial}\psi_1)(\partial\psi_2)+|\psi_2|^2(\partial\psi_2)(\overline{\partial\psi_2}).$$

Substituting these derivatives into the expression for p^2K, we obtain the following result:

$$p^4K=\psi_1\psi_2(\partial\bar{\psi}_1)(\overline{\partial\psi_2})+\bar{\psi}_1\bar{\psi}_2(\bar{\partial}\psi_1)(\partial\psi_2)-|\psi_1|^2(\partial\psi_2)(\overline{\partial\psi_2})-|\psi_2|^2(\bar{\partial}\psi_1)(\partial\bar{\psi}_1)+p^4. \tag{2.16}$$

This gives an explicit form for the Gaussian curvature K in terms of the functions ψ_1 and ψ_2.

III. THE ESTIMATION OF DEGREE OF INDETERMINANCY OF GENERAL SOLUTIONS

Now, let us demonstrate that the general analytic solutions of the Weierstrass–Enneper system (1.2) and system (2.4) possess the same degree of freedom. To this end, we employ Cartan's theory of systems in involution.[21] For more information on this subject, see Refs. 22–24.

For computational purposes, it is useful to examine the systems of Pfaffian forms equivalent to the considered systems of equations (1.2) and (2.4). We determine the Cartan numbers of these systems and the numbers of arbitrary parameters admitted by the general solutions of their polar equations.[21]

A. The Weierstrass–Enneper system

If one introduces the following notation,

$$x^1=\overline{z}, \quad x^2=z, \quad u_1=\psi_1, \quad u_2=\psi_2, \quad u_3=\overline{\psi_1}, \quad u_4=\overline{\psi_2},$$
$$u_5=\partial\psi_1, \quad u_6=\overline{\partial}\psi_2, \quad u_7=\overline{\partial\psi_1}, \quad u_8=\partial\overline{\psi_2}, \tag{3.1}$$

and

$$\xi^1=\overline{\partial}\psi_1, \quad \xi^2=\partial\psi_2, \quad \xi^3=\partial\overline{\psi_1}, \quad \xi^4=\overline{\partial\psi_2}, \tag{3.2}$$

then system (1.2) takes the form

$$u_5=(u_1u_3+u_2u_4)u_2, \quad u_7=(u_1u_3+u_2u_4)u_4,$$
$$u_6=-(u_1u_3+u_2u_4)u_1, \quad u_8=-(u_1u_3+u_2u_4)u_3. \tag{3.3}$$

If one chooses (3.2) as parameters, then, in terms of (3.1), Eq. (3.3) can be written as a system of differential one-forms,

$$\omega_1=du_1-(\xi^1dx^1+u_5dx^2)=0,$$
$$\omega_2=du_2-(u_6dx^1+\xi^2dx^2)=0,$$
$$\omega_3=du_3-(u_7dx^1+\xi^3dx^2)=0,$$
$$\omega_4=du_4-(\xi^4dx^1+u_8dx^2)=0,$$
$$\omega_5=du_5-\{[(\xi^1u_3+u_1u_7+u_6u_4+u_2\xi^4)u_2+(u_1u_3+u_2u_4)u_6]dx^1$$
$$+[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_2+(u_1u_3+u_2u_4)\xi^2]dx^2\}=0,$$
$$\omega_6=du_6+\{[(\xi^1u_3+u_1u_7+u_6u_4+u_2\xi^4)u_1+(u_1u_3+u_2u_4)\xi^1]dx^1$$
$$+[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_1+(u_1u_3+u_2u_4)u_5]dx^2\}=0,$$
$$\omega_7=du_7-\{[(\xi^1u_3+u_1u_7+u_6u_4+u_2\xi^4)u_4+(u_1u_3+u_2u_4)\xi^4]dx^1$$
$$+[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_4+(u_1u_3+u_2u_4)u_8]dx^2\}=0,$$
$$\omega_8=du_8+\{[(\xi^1u_3+u_1u_7+u_6u_4+u_2\xi^4)u_3+(u_1u_3+u_2u_4)u_7]dx^1$$
$$+[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_3+(u_1u_3+u_2u_4)\xi^3]dx^2\}=0. \tag{3.4}$$

Here, we interpret $\overline{z}$ and z as independent coordinates x^1 and x^2, respectively, in $\mathbb{R}^2$ space. The variables $u=(u_1,\ldots,u_8)$ are considered as coordinates in $\mathbb{R}^8$ space. The quantity $\xi=(\xi^1,\ldots,\xi^4)$ represents a vector of all first derivatives of ψ_i which do not appear in WE system (1.2). If we consider the variables u and ξ as unknown functions of $x=(x^1,x^2)$, then, in terms of (3.1) and (3.2), the WE system (1.2) is equivalent to the system of differential one-forms (3.4). After exterior differentiation of (3.4) we obtain the following system of two-forms, modulo (3.4),

$$\Omega_1\equiv d\omega_1=dx^1\wedge d\xi^1-[(\xi^1u_3+u_1u_7+u_6u_4+u_2\xi^4)u_2+(u_1u_3+u_2u_4)u_6]dx^1\wedge dx^2,$$
$$\Omega_2\equiv d\omega_2=-[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_1+(u_1u_3+u_2u_4)u_5]dx^1\wedge dx^2+dx^2\wedge d\xi^2,$$
$$\Omega_3\equiv d\omega_3=[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_4+(u_1u_3+u_2u_4)u_8]dx^1\wedge dx^2+dx^2\wedge d\xi^3,$$

$$\Omega_4 \equiv d\omega_4 = dx^1 \wedge d\xi^4 + [(\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_3 + (u_1 u_3 + u_2 u_4) u_7] dx^1 \wedge dx^2,$$

$$\begin{aligned}\Omega_5 \equiv d\omega_5 = & -u_2 u_3 d\xi^1 \wedge dx^1 + u_2^2 dx^1 \wedge d\xi^4 - u_1 u_2 d\xi^3 \wedge dx^2 - (u_1 u_3 + 2u_2 u_4) d\xi^2 \wedge dx^2 \\ & + [u_1 u_2((u_3 u_5 + u_1 \xi^3 + \xi^2 u_4 + u_2 u_8) u_4 + (u_1 u_3 + u_2 u_4) u_8) - (u_1 u_3 + 2u_2 u_4)((u_3 u_5 \\ & + u_1 \xi^3 + \xi^2 u_4 + u_2 u_8) u_1 + (u_1 u_3 + u_2 u_4) u_5) - u_2 u_3((\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_2 \\ & + (u_1 u_3 + u_2 u_4) u_6) + u_2^2((\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_3 \\ & + (u_1 u_3 + u_2 u_4) u_7)] dx^1 \wedge dx^2,\end{aligned} \tag{3.5}$$

$$\begin{aligned}\Omega_6 \equiv d\omega_6 = & -(2u_1 u_3 + u_2 u_4) dx^1 \wedge d\xi^1 - u_1 u_2 dx^1 \wedge d\xi^4 - u_1 u_4 dx^2 \wedge d\xi^2 - u_1^2 dx^2 \wedge d\xi^3 \\ & + [-u_1^2((u_3 u_5 + u_1 \xi^3 + \xi^2 u_4 + u_2 u_8) u_4 + (u_1 u_3 + u_2 u_4) u_8) + u_1 u_4((u_3 u_5 + u_1 \xi^3 \\ & + \xi^2 u_4 + u_2 u_8) u_1 + (u_1 u_3 + u_2 u_4) u_5) - u_1 u_2((\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_3 \\ & + (u_1 u_3 + u_2 u_4) u_7) + ((\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_2 \\ & + (u_1 u_3 + u_2 u_4) u_6)(2u_1 u_3 + u_2 u_4)] dx^1 \wedge dx^2,\end{aligned}$$

$$\begin{aligned}\Omega_7 \equiv d\omega_7 = & \, u_3 u_4 dx^1 \wedge d\xi^1 + (u_1 u_3 + 2u_2 u_4) dx^1 \wedge d\xi^4 + u_4^2 dx^2 \wedge d\xi^2 + u_1 u_4 dx^2 \wedge d\xi^3 \\ & + [u_1 u_4((u_3 u_5 + u_1 \xi^3 + u_4 \xi^2 + u_2 u_8) u_4 + (u_1 u_3 + u_2 u_4) u_8) - u_4^2 \\ & \times ((u_3 u_5 + u_1 \xi^3 + u_4 \xi^2 + u_2 u_8) u_1 + (u_1 u_3 + u_2 u_4) u_5) - u_3 u_4((\xi^1 u_3 + u_1 u_7 + u_4 u_6 \\ & + u_2 \xi^4) u_2 + (u_1 u_3 + u_2 u_4) u_6) + (u_1 u_3 + 2u_2 u_4)((\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_3 \\ & + (u_1 u_3 + u_2 u_4) u_7)] dx^1 \wedge dx^2,\end{aligned}$$

$$\begin{aligned}\Omega_8 \equiv d\omega_8 = & -u_3^2 dx^1 \wedge d\xi^1 - u_2 u_3 dx^1 \wedge d\xi^4 - u_3 u_4 dx^2 \wedge d\xi^2 - (2u_1 u_3 + u_2 u_4) dx^2 \wedge d\xi^3 \\ & + [u_3 u_4((u_3 u_5 + u_1 \xi^3 + \xi^2 u_4 + u_2 u_8) u_1 + (u_1 u_3 + u_2 u_4) u_5) - (2u_1 u_3 + u_2 u_4) \\ & \times ((u_3 u_5 + u_1 \xi^3 + \xi^2 u_4 + u_2 u_8) u_4 + (u_1 u_3 + u_2 u_4) u_8) \\ & + u_3^2((\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_2 + (u_1 u_3 + u_2 u_4) u_6) \\ & - u_2 u_3((\xi^1 u_3 + u_1 u_7 + u_4 u_6 + u_2 \xi^4) u_3 + (u_1 u_3 + u_2 u_4) u_7)] dx^1 \wedge dx^2.\end{aligned}$$

Note that the quantities ξ^i enter linearly into the expressions (3.4) and (3.5).

The vector fields Y_j, $j=1,2$ which annihilate the one-forms ω_s and the two-forms Ω_s, satisfy the polar equations,

$$\langle \omega_s \lrcorner Y_j \rangle = 0, \quad \langle \Omega_s \lrcorner Y_1, Y_2 \rangle = 0, \quad s = 1,...,8, \quad j = 1,2. \tag{3.6}$$

From conditions (3.6) one finds

$$\begin{aligned}Y_1 = & \, \partial_{x^1} + \sum_{r=1}^{4} a^r \partial_{\xi^r} + \xi^1 \partial_{u_1} + u_6 \partial_{u_2} + u_7 \partial_{u_3} + \xi^4 \partial_{u_4} - [(\xi^1 u_3 + u_1 u_7 + u_6 u_4 + u_2 \xi^4) u_2 \\ & + (u_1 u_3 + u_2 u_4) u_6] \partial_{u_5} + [(\xi^1 u_3 + u_1 u_7 + u_6 u_4 + u_2 \xi^4) u_1 + (u_1 u_3 + u_2 u_4) \xi^1] \partial_{u_6} \\ & - [(\xi^1 u_3 + u_1 u_7 + u_6 u_4 + u_2 \xi^4) u_4 + (u_1 u_3 + u_2 u_4) \xi^4] \partial_{u_7} \\ & + [(\xi^1 u_3 + u_1 u_7 + u_6 u_4 + u_2 \xi^4) u_3 + (u_1 u_3 + u_2 u_4) u_7] \partial_{u_8},\end{aligned}$$

and

$$Y_2=\partial_{x^2}+\sum_{r=1}^{4} b^r\partial_{\xi^r}+u_5\partial_{u_1}+\xi^2\partial_{u_2}+\xi^3\partial_{u_3}+u_8\partial_{u_4}-[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_2$$
$$+(u_1u_3+u_2u_4)\xi^2]\partial_{u_5}+[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_1+(u_1u_3+u_2u_4)u_5]\partial_{u_6}$$
$$-[(u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_4+(u_1u_3+u_2u_4)u_8]\partial_{u_7}$$
$$+[(u_5u_3+u_1\xi_3+\xi^2u_4+u_2u_8)u_3+(u_1u_3+u_2u_4)\xi^3]\partial_{u_8},$$

where

$$a^2=-((u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_1+(u_1u_3+u_2u_4)u_5),$$
$$a^3=((u_5u_3+u_1\xi^3+\xi^2u_4+u_2u_8)u_4+(u_1u_3+u_2u_4)u_8),$$
$$b^1=((\xi^1u_3+u_1u_7+u_6u_4+u_2\xi^4)u_2+(u_1u_3+u_2u_4)u_6),$$
$$b^4=-((\xi^1u_3+u_1u_7+u_6u_4+u_2\xi^4)u_3+(u_1u_3+u_2u_4)u_7),$$

and the quantities a^1,a^4,b^2,b^3 are arbitrary. Thus the number of free parameters in (3.6) is

$$N=4. \tag{3.7}$$

Under the chosen notation (3.1) and (3.2) the Pfaffian system (3.4) takes the abbreviated form

$$\omega_s=du_s-G_{s\mu}(x,\xi,u)dx^\mu,\quad s=1,...,8,\quad \mu=1,2, \tag{3.8}$$

where

$$x=(x^1,x^2),\quad \xi=(\xi^1,\ldots,\xi^4),\quad u=(u^1,\ldots,u^8),$$

and $G_{s\mu}$ depends linearly on ξ. The elements of the 8×4 matrix,

$$a_{sr}=\left(\frac{\partial G_{s\mu}}{\partial\xi^r}(x,\xi,u)X^\mu\right),\quad X=(X^1,X^2)\in\mathbb{R}^2, \tag{3.9}$$

determine the values of the Cartan quasicharacters s_i, $i=1,2$. The nonzero elements of the matrix (a_{sr}) are

$$a_{11}=X^1\quad a_{51}=u_3u_2X^1\quad a_{61}=-(2u_1u_3+u_2u_4)X^1$$
$$a_{22}=X^2\quad a_{52}=(2u_2u_4+u_1u_3)X^2\quad a_{62}=-u_1u_4X^2$$
$$a_{33}=X^2\quad a_{53}=u_1u_2X^2\quad a_{63}=-u_1^2X^2$$
$$a_{44}=X^1\quad a_{54}=u_2^2X^1\quad a_{64}=-u_1u_2X^1$$
$$a_{71}=u_3u_4X^1\quad a_{81}=-u_3^2X^1$$
$$a_{72}=u_4^2X^2\quad a_{82}=-u_2u_3X^1$$
$$a_{73}=u_1u_4X^2\quad a_{83}=-(2u_1u_3+u_2u_4)X^2$$
$$a_{74}=(u_1u_3+2u_2u_4)X^1\quad a_{84}=-u_3u_4X^2.$$

Thus, the Cartan quasicharacters are given by

$$s_1=\max_{X\in\mathbb{R}^2}\operatorname{rank}(a_{sr})=4,\quad s_2=p-s_1=0,$$

where p is the number of coordinates ξ, that is, $p=4$. From the definition of the Cartan number Q one has

$$Q=s_1+2s_2=4. \tag{3.10}$$

Since the number N of free parameters appearing in (3.6) equals 4, one has $Q=N$. Thus, according to Cartan's theorem,[21] system (3.4) [as well as (1.2)] is in involution. Its general analytic solution exists in some neighborhood of a regular point (x_0,ξ_0,u_0) and depends on four arbitrary real analytic functions of one real variable.

B. The second–order system of PDEs

Now, for the system (2.4), a similar analysis is performed. We demonstrate that locally the solution space of (2.4) has the same dimension as the system (1.2). For computational purposes, it is useful to write Eqs. (2.4) in the form

$$\begin{aligned} &u_{1,x^1x^2}-\frac{2u_2}{1+u_1u_2}u_{1,x^1}u_{1,x^2}=0,\\ &u_{2,x^1x^2}-\frac{2u_1}{1+u_1u_2}u_{2,x^1}u_{2,x^2}=0, \end{aligned} \tag{3.11}$$

where the following notation has been used:

$$x^1=z,\quad x^2=\overline{z},\quad u_1=\rho,\quad u_2=\overline{\rho}. \tag{3.12}$$

The system of differential one-forms corresponding to (3.11) can be written as follows:

$$\begin{aligned} &\omega_1=du_1-u_3dx^1-u_5dx^2=0,\\ &\omega_2=du_2-u_4dx^1-u_6dx^2=0,\\ &\omega_3=du_3-\xi^1dx^1-\frac{2u_2}{1+u_1u_2}u_3u_5dx^2=0,\\ &\omega_4=du_4-\xi^2dx^1-\frac{2u_1}{1+u_1u_2}u_4u_6dx^2=0,\\ &\omega_5=du_5-\frac{2u_2}{1+u_1u_2}u_3u_5dx^1-\xi^3dx^2=0,\\ &\omega_6=du_6-\frac{2u_1}{1+u_1u_2}u_4u_6dx^1-\xi^4dx^2=0,\\ &\omega_7=du_7-u_{11}dx^1-u_{12}dx^2,\\ &\omega_8=du_8-u_{21}dx^1-u_{22}dx^2, \end{aligned} \tag{3.13}$$

where we use the standard notation

$$u_3=u_{1,x^1},\quad u_4=u_{2,x^1},\quad u_5=u_{1,x^2},\quad u_6=u_{2,x^2},\quad u_7=u_{1,x^1x^2},\quad u_8=u_{2,x^1x^2}, \tag{3.14}$$

and, for the sake of simplicity, introduce the additional notation

$$u_{11}=2\left(\frac{u_2u_3u_5}{1+u_1u_2}\right)_{x^1}=2\left[\frac{u_3u_4u_5+u_2u_5\xi^1+u_2u_3u_7}{1+u_1u_2}-\frac{(u_2u_3u_5)(u_2u_3+u_1u_4)}{(1+u_1u_2)^2}\right],$$

$$u_{12}=2\left(\frac{u_2u_3u_5}{1+u_1u_2}\right)_{x^2}=2\left[\frac{u_3u_5u_6+u_2u_5u_7+u_2u_3\xi^3}{1+u_1u_2}-\frac{(u_2u_3u_5)(u_2u_5+u_1u_6)}{(1+u_1u_2)^2}\right],$$

$$u_{21}=2\left(\frac{u_1u_4u_6}{1+u_1u_2}\right)_{x^1}=2\left[\frac{u_3u_4u_6+u_1u_6\xi^2+u_1u_4u_8}{1+u_1u_2}-\frac{(u_1u_4u_6)(u_2u_3+u_1u_4)}{(1+u_1u_2)^2}\right],$$

$$u_{22}=2\left(\frac{u_1u_4u_6}{1+u_1u_2}\right)_{x^2}=2\left[\frac{u_4u_5u_6+u_1u_6u_8+u_1u_4\xi^4}{1+u_1u_2}-\frac{(u_1u_4u_6)(u_2u_5+u_1u_6)}{(1+u_1u_2)}\right].$$

We choose as parameters

$$\xi^1=u_{1,x^1x^1},\quad \xi^2=u_{1,x^2x^2},\quad \xi^3=u_{2,x^1x^1},\quad \xi^4=u_{2,x^2x^2}. \tag{3.15}$$

As in the previous case, given the chosen notation (3.12), the Pfaffian system (3.13) takes the form (3.8). Note that, in this case, the matrix (3.9) has the same dimension 8×4 as in the case of system (1.2). The nonzero elements of the matrix (a_{sr}) are

$$a_1^3=X^1,\quad a_2^4=-X^1,\quad a_3^5=-X^2,\quad a_4^6=-X^2.$$

Thus, the Cartan quasi-characters are given by

$$s_1=\max_{X\in\mathbb{R}^2}\operatorname{rank}(a_{sr})=4,\quad s_2=p-s_1=0,$$

where p is the number of coordinates ξ, that is, $p=4$. Consequently, the Cartan number Q equals

$$Q=s_1+2s_2=4. \tag{3.16}$$

After exterior differentiation system (3.13) with the chosen notation (3.12), (3.14), and (3.15) takes the form

$$\Omega_l\equiv d\omega_l\equiv 0,\quad l=1,2,7,8,$$

$$\Omega_3\equiv d\omega_3=-dx^1\wedge d\xi^1-u_{11}dx^1\wedge dx^2,$$

$$\Omega_4\equiv d\omega_4=-dx^1\wedge d\xi^2-u_{21}dx^1\wedge dx^2,$$

$$\Omega_5\equiv d\omega_5=u_{12}dx^1\wedge dx^2+dx^2\wedge d\xi^3,$$

$$\Omega_6\equiv d\omega_6=u_{22}dx^1\wedge dx^2+dx^2\wedge d\xi^4,$$

which is satisfied modulo (3.13).

The vector fields Y_j, $j=1,2$, which annihilate the one-forms ω_s and the two-forms Ω_s satisfy the polar equations (3.6). From these equations we have

$$Y_1=\partial_{x^1}+\sum_{r=1}^{4} a^r\partial_{\xi^r}+u_3\partial_{u^1}+u_4\partial_{u^2}+\xi^1\partial_{u^3}+\xi^2\partial_{u^4}+2\,\frac{u_2u_3u_5}{1+u_1u_2}\,\partial_{u_5}$$
$$+2\,\frac{u_1u_4u_6}{1+u_1u_2}\,\partial_{u_6}+u_{11}\partial_{u_7}+u_{21}\partial_{u_8},$$
$$Y_2=\partial_{x^2}+\sum_{r=1}^{4} b^r\partial_{\xi^r}+u_5\partial_{u_1}+u_6\partial_{u_2}+2\,\frac{u_2u_3u_5}{1+u_1u_2}\,\partial_{u_3}+2\,\frac{u_1u_4u_6}{1+u_1u_2}\,\partial_{u_4}+\xi^3\partial_{u_5}$$
$$+\xi^4\partial_{u_6}+u_{12}\partial_{u_7}+u_{22}\partial_{u_8}, \tag{3.17}$$

where

$$b^1=-u_{11},\quad b^2=-u_{21},\quad a^3=u_{12},\quad a^4=u_{22}.$$

As in the previous case, we have four free parameters, a^1,a^2,b^3,b^4. Thus, we get $Q=N$ and, according to Cartan's theorem, system (3.11) [as well as (2.4)] is in involution. Its general analytic solution depends on four arbitrary real analytic functions of one real variable.

Note that, since the systems of one-forms (3.4) and (3.13) are equivalent to systems (1.2) and (2.4), respectively, the Cartan theorem implies the existence of the general analytic solutions of (1.2) and (2.4). Now, if we interpret $\bar{z}$ and z as coordinates in the complex plane $\mathbb{C}$ and ψ_i and $\bar{\psi}_i$ as complex and complex conjugate functions on $\mathbb{C}$, then the general solutions of both systems (1.2) and (2.4) depend on two arbitrary complex analytic functions of one complex variable and their complex conjugate functions.

We have shown that systems (1.2) and (2.4) possess the same degree of freedom in terms of their general analytic solutions. Using Cartan's theorem, we can formulate the following conclusion.

Proposition 3: Suppose the systems (1.2) and (2.4) are both in involution at regular points (z_0,ξ_0,ψ_0) and (z_0,ξ_0,ρ_0), respectively. Then their general analytic solutions exist in some neighborhood of these regular points and both depend on two arbitrary complex analytic functions of one complex variable, and their complex conjugate functions.

Note that the mapping given by (2.1), from the solution of (2.4) to the solution of (1.2), does not restrict the type of boundary value conditions imposed on (2.4) and (1.2).

IV. COMPLETE INTEGRABILITY OF THE WEIERSTRASS–ENNEPER SYSTEM IN THE CONTEXT OF THE SIGMA MODEL

A. The linear spectral problem associated with the Weierstrass–Enneper system

The objective of this section is to demonstrate a connection between the Weierstrass–Enneper system (1.2) and the completely integrable Euclidean sigma model in two dimensions and, next, to derive through this link the linear spectral problem for the Weierstrass–Enneper system.

Let us identify (2.3) with the stereographic coordinate representation[25] of the two-dimensional Euclidean nonlinear sigma model

$$[S,\partial\bar{\partial}S]=0, \tag{4.1}$$

where the spin matrix

$$S=\begin{pmatrix} s_3 & \bar{s}_+ \\ s_+ & -s_3 \end{pmatrix},\quad \det S=-1,$$

belongs to the Hermitian space SU(2)/U(1). In the stereographic coordinate representation, the matrix S is given by

$$S=\frac{1}{1+|\rho|^2}\begin{pmatrix}1-|\rho|^2 & 2\bar{\rho}\\ 2\rho & -1+|\rho|^2\end{pmatrix}, \tag{4.2a}$$

where

$$s_+=\frac{2\rho}{1+|\rho|^2},\quad s_3=\frac{1-|\rho|^2}{1+|\rho|^2}. \tag{4.2b}$$

By substituting the matrix S given by (4.2) into (4.1), we obtain the following condition,

$$\left[\bar{\rho}\left(\bar{\partial}\partial\rho-\frac{2\bar{\rho}}{1+|\rho|^2}\partial\rho\bar{\partial}\rho\right)-\rho\left(\partial\bar{\partial}\bar{\rho}-\frac{2\rho}{1+|\rho|^2}\bar{\partial}\bar{\rho}\partial\bar{\rho}\right)\right]I=0,$$

where I is the unit matrix in this equation. This is identically satisfied whenever Eqs. (2.4) hold.

In terms of the complex functions ψ_i and $\bar{\psi}_i$, $i=1,2$, which appear in (1.2), the spin matrix S takes the form

$$S=\frac{1}{p}\begin{pmatrix}-|\psi_1|^2+|\psi_2|^2 & 2\bar{\psi}_1\bar{\psi}_2\\ 2\psi_1\psi_2 & |\psi_1|^2-|\psi_2|^2\end{pmatrix}. \tag{4.3}$$

From (2.3), we obtain that the inverse mapping of (4.3) is double valued and is provided by

$$\psi_1=\frac{\epsilon}{2}s_+\left[\bar{\partial}\left(\frac{\bar{s}_+}{1+s_3}\right)\right]^{1/2},\quad \psi_2=\frac{\epsilon}{2}(1+s_3)\left[\partial\left(\frac{s_+}{1+s_3}\right)\right]^{1/2},\quad \epsilon^2=1, \tag{4.4a}$$

where

$$\rho=\frac{s_+}{1+s_3},\quad \bar{\rho}=\frac{\bar{s}_+}{1+s_3}. \tag{4.4b}$$

Proposition 4: If ψ_1 and ψ_2 are solutions of the WE system (1.2), then the spin matrix S given by (4.3) is a solution of the sigma model equation (4.1).

Proof: The results are directly obtained by substituting the spin matrix S given by (4.3) into the commutator (4.1) and assuming that the functions ψ_i satisfy (1.2). This computation leads to a vanishing commutator (4.1). Q.E.D.

The procedure for constructing solutions to (1.2) can be reduced to the following. Take any solution of the sigma model (4.1) and substitute it into Eqs. (4.4b). The function ρ thus obtained provides us, by means of transformation (2.3), with solutions ψ_1 and ψ_2 of the WE system (1.2). The possibility of constructing such solutions is demonstrated in the next section.

We now consider the possibility of constructing a linear spectral problem for the WE system (1.2). Let us introduce a new set of complex functions φ_1 and $\varphi_2:\mathbb{C}\to\mathbb{C}$ which are related to the complex functions ψ_1 and ψ_2 in the following way:

$$\begin{aligned}\psi_1&=f(z,\bar{z})\varphi_1,\quad \bar{\psi}_2=f(z,\bar{z})\bar{\varphi}_2,\\ \bar{\psi}_1&=\bar{f}(\bar{z},z)\bar{\varphi}_1,\quad \psi_2=\bar{f}(\bar{z},z)\varphi_2,\end{aligned} \tag{4.5}$$

for any complex function $f:\mathbb{C}\to\mathbb{C}$. From the definition (2.1), it is evident that the transformation (4.5) leaves the functions ρ and $\bar{\rho}$ invariant,

$$\rho=\frac{\varphi_1}{\bar{\varphi}_2},\quad \bar{\rho}=\frac{\bar{\varphi}_1}{\varphi_2}, \tag{4.6}$$

and the structure of the spin matrix S given by (4.3) is also preserved. This means that there exists a freedom which resembles a type of gauge freedom in the definition of the ρ variable, since the numerator and denominator of (4.6) can be multiplied by any complex function. The crux of the matter is that it is not required that the set of functions φ_i satisfy the original system (1.2), but that the ratio of φ_1 over $\overline{\varphi_2}$ satisfy (2.4). Let us express Eqs. (1.2) in terms of φ_i and f. The derivatives of ψ_1 and ψ_2 take the form

$$\partial\psi_1 = (\partial f)\varphi_1 + f(\partial\varphi_1), \quad \bar{\partial}\psi_2 = (\overline{\partial f})\varphi_2 + f(\bar{\partial}\varphi_2).$$

We define the variable $q = |\varphi_1|^2 + |\varphi_2|^2$, and from (1.2a) it follows that $p = |f|^2 q$. Taking the above into account we can write (1.2) as

$$(\partial f)\varphi_1 + f(\partial\varphi_1) = p\bar{f}\varphi_2, \quad (\overline{\partial f})\varphi_2 + \overline{f}(\bar{\partial}\varphi_2) = -pf\varphi_1.$$

Solving the above equations for $\partial\varphi_1$ and $\bar{\partial}\varphi_2$, respectively, we obtain the equations of motion,

$$\begin{aligned} \partial\varphi_1 &= q\overline{f^2}\varphi_2 - (\partial \ln f)\varphi_1, \quad \bar{\partial}\varphi_2 = -qf^2\varphi_1 - (\bar{\partial}\ln\bar{f})\varphi_2, \\ \bar{\partial}\overline{\varphi_1} &= qf^2\overline{\varphi_2} - (\bar{\partial}\ln\bar{f})\overline{\varphi_1}, \quad \partial\overline{\varphi_2} = -q\overline{f^2}\,\overline{\varphi_1} - (\partial\ln f)\overline{\varphi_2}. \end{aligned} \tag{4.7}$$

Using (4.7), the differentiation of Eqs. (4.6) with respect to z and $\bar{z}$, respectively, yields a pair of relations similar to (2.3) which relate the functions φ_i to ρ and a nonzero f,

$$\varphi_1 = \epsilon\rho\frac{(\bar{\partial}\bar{\rho})^{1/2}}{f(1+|\rho|^2)}, \quad \varphi_2 = \epsilon\frac{(\partial\rho)^{1/2}}{f(1+|\rho|^2)}. \tag{4.8}$$

Relations (4.8) can also be obtained in a more straightforward way by substituting (4.5) into Eqs. (2.3). Note that as well as (2.3), the transformations (4.8) are doubled valued. Now we can formulate the following.

Proposition 5: If the function ρ defined by (4.6) is a solution of the system (2.4), then the functions φ_1 and φ_2 satisfy the following system of equations:

$$\begin{aligned} \partial\varphi_1 &= q\overline{f^2}\varphi_2, \quad \bar{\partial}\overline{\varphi_1} = qf^2\overline{\varphi_2}, \\ \bar{\partial}\varphi_2 &= -qf^2\varphi_1, \quad \partial\overline{\varphi_2} = -q\overline{f^2}\,\overline{\varphi_1}, \end{aligned} \tag{4.9}$$

for any function $f:\mathbb{C}\to\mathbb{C}$ satisfying

$$\partial f = 0. \tag{4.10}$$

Proof: Indeed, by an easy computation, one obtains from (4.6) and (4.7) the first derivatives of ρ and $\bar{\rho}$,

$$\begin{aligned} \partial\rho &= q^2|f|^2(\overline{\varphi_2})^{-2}, \quad \partial\bar{\rho} = (\varphi_2)^{-2}(\varphi_2\partial\overline{\varphi_1} - \overline{\varphi_1}\partial\varphi_2), \\ \bar{\partial}\rho &= (\overline{\varphi_2})^{-2}(\overline{\varphi_2}\bar{\partial}\varphi_1 - \varphi_1\bar{\partial}\overline{\varphi_2}), \quad \bar{\partial}\bar{\rho} = q^2|f|^2(\varphi_2)^{-2}, \end{aligned} \tag{4.11a}$$

and the second derivatives of ρ and $\bar{\rho}$,

$$\begin{aligned} \partial\bar{\partial}\rho &= (\overline{\varphi_2})^{-3}[2q\overline{f^2}\,\overline{\varphi_1}\,\overline{\varphi_2}\bar{\partial}\varphi_1 - q\overline{f^2}(q+|\varphi_1|^2-|\varphi_2|^2)\bar{\partial}\overline{\varphi_2}], \\ \partial\bar{\partial}\bar{\rho} &= (\varphi_2)^{-3}[2qf^2\varphi_1\varphi_2\partial\overline{\varphi_1} - qf^2(q+|\varphi_1|^2-|\varphi_2|^2)\partial\varphi_2]. \end{aligned} \tag{4.11b}$$

Substituting expressions (4.11) into (2.4), we get a differential constraint for the function f and its respective complex conjugate

$$(\overline{f}^2-1)\overline{\partial f}=0, \quad (f^2-1)\partial f=0. \tag{4.12}$$

Thus, the general solution of system (4.12) is given by any antiholomorphic function f such that relation (4.10) holds. Consequently, the equations of motion (4.7) become those in (4.9). Q.E.D.

Proposition 6: If the boundary value problem for the WE system (1.2) is given by two arbitrary complex analytic functions of one complex variable (and their complex conjugate functions), then the solution of (1.2) is unique up to a gauge transformation (4.5).

Proof: By virtue of Propositions 1 and 2, the map from Eqs. (1.2) to (2.4) is one-to-one. The map from Eqs. (2.4) to the sigma model (4.1) is also one-to-one because of the transformation (4.2b). The solution of the boundary value problem for (4.1) possesses a unique solution.[25] Hence, from Proposition 3 and Eqs. (4.5) and (4.8), it follows that the solution of the boundary value problem for WE system (1.2) is unique up to multiplication by any function $f(\overline{z})$ satisfying (4.10). This means that the freedom of solutions to (1.2) and (2.4) is the same, up to a gauge function f. Q.E.D.

Now we examine certain aspects of complete integrability of the equations of motion (4.9) in the context of a two-dimensional Euclidean sigma model (4.1).

As it was shown by A. V. Mikhailov in Ref. 26, Eq. (4.1) is a compatibility condition for the two linear spectral problems

$$\partial\Phi=\frac{1}{\lambda+1}\mathcal{U}\Phi, \quad \overline{\partial}\Phi=\frac{1}{\lambda-1}\mathcal{U}^{\dagger}\Phi. \tag{4.13}$$

Here, $\mathcal{U}=\partial SS$, $\mathcal{U}^{\dagger}=S\overline{\partial}S$ with S given by (4.3), and $\Phi(z,\overline{z},\lambda)$ is a matrix of fundamental solutions, while λ represents the spectral parameter. The denominators which contain the spectral parameter cannot be absorbed in the derivatives $\partial\Phi$ and $\overline{\partial}\Phi$. Note that there is a direct connection between the matrix eigenfunction $\Phi(z,\overline{z},\lambda)$ in expression (4.13) and the fields ψ_i, through the mapping (4.4) since there exists the relation[26]

$$S=\Phi(z,\overline{z},0). \tag{4.14}$$

Then we could say that the WE system (1.2) is completely integrable, because of the mappings (4.3) and (4.4). Indeed, by expressing the spin matrix S in terms of the functions φ_i and $\overline{\varphi}_i$, one obtains the explicit form of the linear spectral problem (4.13) for the equation of motion (4.9)

$$\partial\Phi=\frac{2}{\lambda+1}M\Phi, \quad \overline{\partial}\Phi=\frac{2}{\lambda-1}M^{\dagger}\Phi, \tag{4.15}$$

where

$$M=\begin{pmatrix} b/2 & a \\ -c & -b/2 \end{pmatrix}. \tag{4.16}$$

We introduce the following notation:

$$\begin{aligned} a&=-\overline{f}^2\overline{\varphi}_1^2+\frac{1}{fq^2}[\overline{\varphi}_1\overline{\varphi}_2^2\partial(\overline{f}\varphi_2)-\overline{\varphi}_2|\varphi_2|^2\partial(\overline{f}\overline{\varphi}_1)],\\ b&=2\left[-\overline{f}^2\overline{\varphi}_1\varphi_2+\frac{1}{fq^2}(\varphi_1|\varphi_2|^2\partial(\overline{f}\overline{\varphi}_1)-\overline{\varphi}_2|\varphi_1|^2\partial(\overline{f}\varphi_2))\right],\\ c&=-\overline{f}^2\varphi_2^2+\frac{1}{fq^2}[\varphi_1|\varphi_1|^2\partial(\overline{f}\varphi_2)-\varphi_1^2\varphi_2\partial(\overline{f}\overline{\varphi}_1)]. \end{aligned} \tag{4.17}$$

Making use of (4.5), one can find an explicit form for the coefficients (4.17) in terms of the functions ψ_i. The matrix M can be written in the form

$$M=A+\frac{J}{p^2}A^\dagger, \quad \det M=-\frac{2J}{p^2}, \tag{4.18}$$

where J is the current (1.8) and A is a degenerate nilpotent matrix which can be decomposed as follows,

$$A=-\overline{\psi_1}\psi_2\sigma_3-\overline{\psi_1^2}\sigma_+ +\psi_2^2\sigma_-, \quad \sigma_\pm=\tfrac{1}{2}(\sigma_1\pm i\sigma_2), \tag{4.19}$$

where σ_1, σ_2, σ_3 are Pauli matrices

$$\sigma_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \sigma_2=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}, \quad \sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

However, in order to be able to use results from the inverse scattering method to construct soliton solutions of the WE system (1.2), it is convenient to simplify the form of the linear spectral problem (4.15).

Proposition 7: For any bounded entire function J, the linear spectral problem for the WE system (1.2) has the form

$$\partial\Phi=\frac{2}{\lambda+1}A\Phi, \quad \bar{\partial}\Phi=\frac{2}{\lambda-1}A^\dagger\Phi. \tag{4.20}$$

Proof: Indeed, if we substitute (4.18) into the system (4.15), then the system takes the form

$$\partial\Phi=\frac{2}{\lambda+1}\left(A+\frac{J}{p^2}A^\dagger\right)\Phi, \tag{4.21}$$

$$\bar{\partial}\Phi=\frac{2}{\lambda-1}\left(A^\dagger+\frac{\bar{J}}{p^2}A\right)\Phi. \tag{4.22}$$

From the conservation of the current (1.11), we obtain that the current J is a holomorphic function. According to Liouville's theorem if $J(z)$ is an entire function, and if $|J(z)|\leqslant M$ for all $z\in\mathbb{C}$, then $J(z)\equiv$ constant. Consequently, one can take the current J to be equal to zero, hence Eqs. (4.21) and (4.22) become (4.20). The compatibility condition for the two equations in (4.20), namely,

$$\bar{\partial}A-\partial A^\dagger+[A,A^\dagger]=0,$$

is satisfied, whenever the WE system (1.2) holds. Under these circumstances, the linear spectral problem (4.20) holds for the WE system (1.2). So, matrices A and $A^\dagger$ can be identified as the Lax pair for the WE system. Q.E.D.

Moreover, an interesting feature of the WE system (1.2) has been observed. Namely, the system (4.20) has the WE system of equations as compatibility conditions for any function $J(z)$, not necessarily bounded. This fact can be easily verified by direct calculation.

Note that the system of Riccati equations corresponding to (4.20) is

$$\partial y=-\frac{2}{\lambda+1}(\overline{\psi_1}+\psi_2 y)^2, \quad \bar{\partial}y=\frac{2}{\lambda-1}(\overline{\psi_2}-\psi_1 y)^2,$$

where y is a complex function (called the pseudopotential[27] given by the ratio of the components of the vector Φ, that is, $y=\phi_1/\phi_2$). We conclude that the existence of the linear spectral problem (4.20) for the WE system implies that this system is completely integrable.

Finally, a property of the WE system (1.2) in the context of the sigma model is the existence of a topological charge. Indeed, it is well known that the sigma model (4.1) possesses a topological charge,[27–29] which we denote by I. Making use of the current J, the transformations (4.3), and the equations of motion (1.2), one finds that if the integral

$$I=\frac{i}{8\pi}\int_C \mathrm{Tr}\,(\mathbf{S}\cdot[\partial\mathbf{S},\bar{\partial}\mathbf{S}])\,dz d\bar{z}=-\frac{i}{2\pi}\int_C \frac{1}{p^2}[|J|^2-p^4]\,dz d\bar{z} \tag{4.23}$$

exists, it is an integer, where J is given by Eq. (1.8).

B. Reduction of the Weierstrass–Enneper system to a decoupled linear system

Now we discuss a set of conditions which allow the system (1.2) to become a linear decoupled system of equations.

Proposition 8: If the functions ψ_1 and ψ_2 satisfy the overdetermined system composed of the equations of motion (1.2) and differential conditions

$$\bar{\psi}_1\bar{\partial}\psi_1+\psi_2\overline{\partial\psi_2}=0,\quad \bar{\psi}_2\partial\psi_2+\psi_1\partial\bar{\psi}_1=0, \tag{4.24}$$

then the overdetermined system is equivalent to a linear decoupled system of the form

$$\begin{gathered}\partial\bar{\partial}\psi_i+p_0^2\psi_i=0,\quad \partial\overline{\partial\psi_i}+p_0^2\bar{\psi}_i=0,\quad i=1,2,\\ |\psi_1|^2+|\psi_2|^2=p_0\in\mathbb{R}.\end{gathered} \tag{4.25}$$

Proof: Making use of (1.2) and conditions (4.24) we obtain that the derivatives of p given by (2.14) vanish:

$$\partial p=\psi_1(\partial\bar{\psi}_1)+\bar{\psi}_2(\partial\psi_2)=0,\quad \bar{\partial}p=\bar{\psi}_1(\bar{\partial}\psi_1)+\psi_2(\overline{\partial\psi_2})=0. \tag{4.26}$$

This means that if (4.24) holds, then p is a real constant, say p_0. Thus,

$$|\psi_1|^2+|\psi_2|^2=p_0 \tag{4.27}$$

is a conserved quantity. Hence, the Weierstrass–Enneper system (1.2) becomes a linear system which can be decoupled in terms of the functions ψ_i such that (4.25) holds. Q.E.D.

Let us now investigate the case in which all the derivatives of the functions ψ_i and $\bar{\psi}_i$ are specified. This means that we supplement the WE system (1.2) with some additional differential constraints, so we can formulate the following.

If the conditions (4.24) hold, then we show that the system (1.2) can be extended to the system of the form

$$\partial\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=\begin{pmatrix}p\psi_2\\ \alpha\end{pmatrix},\quad \bar{\partial}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=\begin{pmatrix}\beta\\ -p\psi_1\end{pmatrix}, \tag{4.28a}$$

and the respective conjugate system,

$$\partial\begin{pmatrix}\bar{\psi}_1\\ \bar{\psi}_2\end{pmatrix}=\begin{pmatrix}\bar{\beta}\\ -p\bar{\psi}_1\end{pmatrix},\quad \bar{\partial}\begin{pmatrix}\bar{\psi}_1\\ \bar{\psi}_2\end{pmatrix}=\begin{pmatrix}p\bar{\psi}_2\\ \bar{\alpha}\end{pmatrix}, \tag{4.28b}$$

where the quantities α and β are assumed to be some polynomial functions expressible in terms of ψ_i and $\bar{\psi}_i$, with constant coefficients. The system (4.28) will be called the augmented system. Our aim is to find an explicit form for α and β in such a way that they do not provide any additional differential constraints on ψ_i and $\bar{\psi}_i$ other than (1.2) and (4.24) when the compatibility conditions for (4.28) are added, namely, (2.8) and

$$\begin{aligned}\partial\partial\psi_1&=\partial(p\psi_2)=(\psi_1\partial\bar{\psi}_1+\bar{\psi}_2\partial\psi_2)\psi_2+p\partial\psi_2,\\ \overline{\partial\partial\psi_1}&=\bar{\partial}(p\bar{\psi}_2)=(\bar{\psi}_1\bar{\partial}\psi_1+\psi_2\overline{\partial\psi_2})\bar{\psi}_2+p\overline{\partial\psi_2},\\ \overline{\partial}\partial\psi_2&=-\bar{\partial}(p\psi_1)=-(\bar{\psi}_1\bar{\partial}\psi_1+\psi_2\overline{\partial\psi_2})\psi_1-p\bar{\partial}\psi_1,\\ \partial\partial\bar{\psi}_2&=-\partial(p\bar{\psi}_1)=-(\psi_1\partial\bar{\psi}_1+\bar{\psi}_2\partial\psi_2)\bar{\psi}_1-p\partial\bar{\psi}_1.\end{aligned}\tag{4.29}$$

Indeed, from the compatibility conditions (2.8) and (4.29), the analysis of the dominant terms in the ψ_i and $\bar{\psi}_i$ functions leads to the requirement that all unknown derivatives, $(\partial\psi_1,\partial\bar{\psi}_1,\partial\psi_2,\partial\bar{\psi}_2)$, other than those appearing in (1.2), have to be cubic in terms of the fields ψ_i and $\bar{\psi}_i$. Moreover, if one assumes that the discrete symmetry of the system (1.2), invariant under the reflection symmetry in the space of dependent and independent variables, namely,

$$\psi_i\to-\psi_j,\quad \bar{\psi}_i\to-\bar{\psi}_j,\quad i\neq j=1,2,\quad \partial\to-\bar{\partial},\quad \bar{\partial}\to-\partial,\tag{4.30}$$

can be extended to the augmented system (4.28), then one obtains the following relations,

$$\begin{aligned}\partial\bar{\psi}_1&=-p[\bar{c}_2\bar{\psi}_1+\bar{c}_1\bar{\psi}_2+\bar{c}_4\psi_1+\bar{c}_3\psi_2],\\ \bar{\partial}\psi_1&=-p[c_2\psi_1+c_1\psi_2+c_4\bar{\psi}_1+c_3\bar{\psi}_2],\\ \partial\psi_2&=p[c_1\psi_1+c_2\psi_2+c_3\bar{\psi}_1+c_4\bar{\psi}_2],\\ \overline{\partial\psi_2}&=p[\bar{c}_1\bar{\psi}_1+\bar{c}_2\bar{\psi}_2+\bar{c}_3\psi_1+\bar{c}_4\psi_2].\end{aligned}\tag{4.31}$$

It is assumed that the c_i, $i=1,...,4$, are constants to be determined from the compatibility conditions for (2.8) and (4.29). Substituting (4.31) into (4.29) leads us to a system of equations which are polynomial in ψ_i and $\bar{\psi}_i$. The unique solution of this system has the form

$$\partial\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=p\begin{pmatrix}\psi_2\\ -\psi_1\end{pmatrix}=\bar{\partial}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}\tag{4.32a}$$

and its respective conjugate system

$$\partial\begin{pmatrix}\bar{\psi}_1\\ \bar{\psi}_2\end{pmatrix}=p\begin{pmatrix}\bar{\psi}_2\\ -\bar{\psi}_1\end{pmatrix}=\bar{\partial}\begin{pmatrix}\bar{\psi}_1\\ \bar{\psi}_2\end{pmatrix}.\tag{4.32b}$$

Note that the same formula (4.32) can be found in Ref. 10. Under the conditions (4.24), we show that the system (1.2) admits a conserved quantity (4.27) with p a real constant. This means that in this case by virtue of (1.5), the Gaussian curvature $K=0$, which implies the space is flat.

Proposition 9: If the functions ψ_1 and ψ_2 satisfy the overdetermined system composed of the WE system (1.2) and the following differential constraint,

$$-(\bar{\psi}_1\bar{\partial}\psi_1+\psi_2\overline{\partial\psi_2})+(\psi_1\partial\bar{\psi}_1+\bar{\psi}_2\partial\psi_2)=0,\tag{4.33}$$

then the conserved quantity p is a real-valued function of a real argument $(z+\bar{z})/2$ and

$$|\psi_1|^2+|\psi_2|^2=p((z+\bar{z})/2).\tag{4.34}$$

Proof: Indeed, using the derivatives of p, and taking into account (4.33), we obtain

$$(\partial-\bar{\partial})p=\psi_1(\partial\bar{\psi}_1)+\bar{\psi}_2(\partial\psi_2)-\bar{\psi}_1(\bar{\partial}\psi_1)-\psi_2(\overline{\partial\psi_2})=0.$$

This completes the proof. Q.E.D.

Consequently, in the case when (4.33) holds, the first fundamental form (1.4) and the Gaussian curvature (1.5) take the following form:

$$\Omega = 4p^2(x)dzd\bar{z}, \quad K = -\ddot{p}(x)/p^2(x),$$

respectively. Here, we introduce the notation $\ddot{p} = d^2p/dx^2$.

V. EXAMPLES AND APPLICATIONS

At this point, we would like to illustrate the proposed procedure for constructing solutions of the WE system (1.2) with several elementary examples.

Now, let us discuss some classes of solutions to the WE system (1.2), which can be obtained directly by applying the transformation (2.3). First, we consider the class of solutions which correspond to analytic choices of the function ρ. It is easy to check that for this class of solutions the conserved density J in (1.11) is identically equal to zero.

(1) The simplest solutions of this type are given by

$$\rho = \left[\frac{(z-z_0)}{\lambda}\right]^n, \tag{5.1}$$

where λ and z_0 are arbitrary real and complex numbers, respectively. From the point of view of the sigma model, this form of ρ corresponds to the instanton of charge $I = -n$ located at z_0 and of size λ. By virtue of the invariance of the system (1.2) under conformal transformations, we can set without loss of generality, $z_0 = 0$ and $\lambda = 1$. Then, using (2.3) we find that the solutions of (1.2) are given by

$$\psi_1 = \epsilon n^{1/2} \frac{z^n \bar{z}^{(n-1)/2}}{1+|z|^{2n}}, \quad \psi_2 = \epsilon n^{1/2} \frac{z^{(n-1)/2}}{1+|z|^{2n}}. \tag{5.2}$$

Each of these solutions belongs to a different topological sector of index n. Furthermore, notice that for all even n, the solutions are double valued. Nevertheless, this fact has no influence on the surfaces parametrized by the relations (1.3). Actually, the solutions (5.2) correspond to only one constant mean curvature surface, which is covered n times as z runs over the complex plane. This surface is obtained (modulo translations) by revolving the curve

$$X_2 = (X_3 - 2)\left(\frac{X_3}{4-X_3}\right)^{1/2} \tag{5.3}$$

around the axis X_3. It possesses a conic point in (0,0,2).

(2) Another class of solutions is provided by the analytic function

$$\rho = e^{\lambda z}, \tag{5.4}$$

corresponding to a static domain wall in the isotropic $O(3)$ magnet. The associated solution of the system (1.2) is

$$\psi_1 = \epsilon \bar{\lambda}^{1/2} \frac{e^{\bar{\lambda}\bar{z}/2}}{e^{-\lambda z} + e^{\bar{\lambda}\bar{z}}}, \quad \psi_2 = \epsilon \lambda^{1/2} \frac{e^{-\lambda z/2}}{e^{-\lambda z} + e^{\bar{\lambda}\bar{z}}}. \tag{5.5}$$

Also in this case, the whole class of solutions parametrized by λ represents a unique constant mean curvature surface (modulo translations), obtained from (1.3) by revolving the following curve around the X_3 axis:

$$X_3 = 2\,\frac{X_2^2}{1 \pm \sqrt{1 - X_2^2}}. \tag{5.6}$$

Many other solutions of (1.2) admitting ρ to be a meromorphic function can be found. For the present, we do not discuss them.

(3) Let us assume now, as opposed to the previous cases, that the conserved current density in (1.8) is a nonvanishing holomorphic function. In such a case, one can check that the matrix $\mathcal{U}$ in the spectral problem (4.13) has a nonvanishing $\mathrm{Tr}\,\mathcal{U}^2$. The simplest choice is to put

$$\mathcal{U} = g(z)\sigma_3. \tag{5.7}$$

The solution of the corresponding spectral problem can be readily obtained:[26]

$$\Phi = \exp[2i\chi\sigma_3]\sigma_1, \tag{5.8}$$

where σ_i are Pauli matrices, $\chi = \mathrm{Im}\int_\Gamma g(z)\,dz$ and Γ is an arbitrary curve in the domain in which g is analytic. Then, resorting to the relations (4.14) and (4.4), one obtains the following solutions to Eq. (1.2):

$$\psi_1 = -\frac{\epsilon}{2} i e^{-i\chi} \bar{g}^{1/2}, \quad \psi_2 = \frac{\epsilon}{2} i e^{-i\chi} g^{1/2}. \tag{5.9}$$

The associated surface is given by the parametric equations

$$\begin{gathered} X_1 = \sin 2\chi + X_{10}, \quad X_2 = -\cos 2\chi + X_{20}, \\ X_3 = \omega + X_{30} \quad \left(\omega = \mathrm{Re}\int_\Gamma f(z)\,dz\right), \end{gathered} \tag{5.10}$$

which describe a cylinder having X_3 as a symmetry axis. Nontrivial deformations of this type of solution can be found by using the recurrence N-soliton wave function formula[26] in expression (4.14).

Now, let us discuss a simple example to illustrate the construction introduced in Sec. IV.

(4) Consider the possibility where f is real and $f = q^{-1/2}$ with φ_i chosen to make $\partial f = 0$. In this case, from (4.9) one has

$$\partial\varphi_1 = q\bar{f}^2\varphi_2 = \varphi_2, \quad \bar{\partial}\varphi_2 = -qf^2\varphi_1 = -\varphi_1. \tag{5.11}$$

This system reduces to two second-order linear equations of the form (4.25)

$$\partial\bar{\partial}\varphi_i + \varphi_i = 0, \quad i = 1,2.$$

For example, let us write a simple set of solutions to this equation

$$\begin{gathered} \varphi_1 = -ie^{i(z+\bar{z})}, \quad \bar{\varphi}_1 = ie^{-i(z+\bar{z})}, \\ \varphi_2 = e^{i(z+\bar{z})}, \quad \bar{\varphi}_2 = e^{-i(z+\bar{z})}. \end{gathered}$$

Therefore, one has

$$q = |\varphi_1|^2 + |\varphi_2|^2 = 2, \quad f = \frac{1}{\sqrt{2}}.$$

Note that for these functions φ_i, conditions (4.24) are identically satisfied. From (4.6), we obtain

$$\rho = -ie^{2i(z+\bar{z})}.$$

It is also easy to show that for this class of solutions ρ, Eqs. (2.4) are identically satisfied. Substituting the functions ρ and f into Eqs. (4.8), one obtains an explicit solution of the WE system (1.2):

$$\psi_1 = -i\frac{\epsilon}{\sqrt{2}}e^{i(z+\bar{z})}, \quad \psi_2 = \frac{\epsilon}{\sqrt{2}}e^{i(z+\bar{z})}.$$

This solution represents a phase plane wave since the argument is one-dimensional $x=z+\bar{z}$, its absolute value is constant, and the solution has exponential form.

(5) A special class of exponential solutions to (1.2) can be found to hold when p is constant. According to Proposition 8, we have to solve (4.25). Thus, a vacuum solution takes the form

$$\psi_1 = c_1 e^{i(hz+k\bar{z})}, \qquad \psi_2 = ic_1\frac{h}{p}e^{i(hz+k\bar{z})}, \tag{5.12}$$

where c_1 is a complex constant and h, k are real constants such that

$$|c_1|^2 = \frac{p^3}{(p^2+h^2)}, \quad p = hk.$$

Due to the linearity of Eqs. (4.25), we can look for a more general class of solutions which represent a superposition of exponential functions. The one-soliton solution of (4.25) is given by

$$\psi_1 = c_1 e^{i(h_1 z + k_1\bar{z})} + c_2 e^{i(h_2 z + k_2\bar{z})}, \qquad \psi_2 = \frac{i}{p}(h_1 c_1 e^{i(h_1 z + k_1\bar{z})} + h_2 c_2 e^{i(h_2 z + k_2\bar{z})}), \tag{5.13}$$

where the c_i are complex constants and the h_i, k_i are real constants which satisfy

$$h_1 k_1 = p^2, \quad h_2 k_2 = p^2, \quad h_1 h_2 = -p^2,$$

and

$$|c_2|^2 = \frac{p - |c_1|^2(1+h_1^2/p^2)}{1+p^2/h_1^2}.$$

From (2.1), we obtain the expression for ρ corresponding to the one-soliton solutions,

$$\rho = -i\frac{c_1 e^{ih_1(z+\bar{z})} + c_2 e^{-ih_1(z+\bar{z})}}{\bar{c}_1 e^{ih_1(z+z)} - \bar{c}_2 e^{ih_1(z+z)}},$$

for which condition (2.4) is identically satisfied.

(6) Now, let us discuss the construction of multi-soliton solutions to the WE system (1.2), which can be obtained by exploiting the linear spectral problem (4.15). According to the first step of the procedure, we choose an antiholomorphic function of the form

$$f = \epsilon(a-b)\frac{(2\bar{z}-a-b)z - a(\bar{z}-a) - b(\bar{z}-b)}{|z-a|^2+|z-b|^2}, \quad \epsilon = \pm 1, \quad a,b \in \mathbb{R} \tag{5.14}$$

and look for a nontrivial solution ρ of the system (2.4),

$$\rho=\frac{z-a}{z-b}. \tag{5.15}$$

The substitution of (5.14) and (5.15) into the relations (4.8) gives

$$\varphi_1=\frac{z-a}{(2\bar{z}-a-b)z-a(\bar{z}-a)-b(\bar{z}-b)},\qquad \varphi_2=\frac{\bar{z}-b}{(2z-a-b)\bar{z}-a(z-a)-b(z-b)}. \tag{5.16}$$

Finally, from the solution of the linear spectral problem (4.15) and relations (4.14) and (4.4), we obtain an explicit one-soliton solution of the WE system (1.2):

$$\psi_1=\epsilon(a-b)\frac{z-a}{|z-a|^2+|z-b|^2},\qquad \psi_2=\epsilon(a-b)\frac{\bar{z}-b}{|z-a|^2+|z-b|^2}. \tag{5.17}$$

A similar computation can be performed for the case in which ρ satisfies (2.4) and has a more general form than (5.15):

$$\rho=\prod_{j=1}^{N}\frac{z-a_j}{z-b_j},\qquad a_j,b_j\in\mathbb{R}, \tag{5.18}$$

with distinct parameters such that a and b are replaced by a_j and b_j, respectively. The same process is done with the function f in (5.14). Thus, we can determine explicitly the corresponding form of a multi-soliton solution by applying the recurrence N-soliton wave function formula[26] in the expression (4.14) to obtain

$$\psi_1=\epsilon\frac{\Pi_{j=1}^{N}(z-a_j)/(z-b_j)}{1+\Pi_{j=1}^{N}|(z-a_j)/(z-b_j)|^2}\left(\sum_{s=1}^{N}\frac{1}{(\bar{z}-b_s)}\left(\prod_{\substack{j=1\\ j\neq s}}^{N}\frac{(\bar{z}-a_j)}{(\bar{z}-b_j)}-\prod_{j=1}^{N}\frac{(\bar{z}-a_j)}{(\bar{z}-b_j)}\right)\right)^{1/2},$$
$$\psi_2=\frac{\epsilon}{1+\Pi_{j=1}^{N}|(z-a_j)/(z-b_j)|^2}\left(\sum_{s=1}^{N}\frac{1}{(z-b_s)}\left(\prod_{\substack{j=1\\ j\neq s}}^{N}\frac{(z-a_j)}{(z-b_j)}-\prod_{j=1}^{N}\frac{(z-a_j)}{(z-b_j)}\right)\right)^{1/2}. \tag{5.19}$$

Note that this solution admits simple poles. The topological charge (4.23) for each of the instanton solutions (5.19) corresponds to $I=\epsilon N$.

VI. FUTURE OUTLOOK

In this paper, we have shown that the adapted WE system (1.2), proposed by B. Konopelchenko and I. A. Taimanov as a tool to induce constant mean curvature surfaces, is closely related to the nonlinear Euclidean sigma-model SU(2). This link enabled us to propose a new approach to the construction of solutions, based on the intermediate system of equations (4.9) with which the sigma model (4.1) is associated.

Let us now consider a system of the form

$$\partial\psi_1=h(z,\bar{z})p\,\psi_2,\qquad \overline{\partial\psi_1}=h(z,\bar{z})p\,\bar{\psi}_2,$$
$$\bar{\partial}\psi_2=-h(z,\bar{z})p\,\psi_1,\qquad \partial\bar{\psi}_2=-h(z,\bar{z})p\,\bar{\psi}_1, \tag{6.1}$$

where h is assumed to be a real function of z and $\bar{z}$ not equal to one. Otherwise, we have the previous case. We are interested in conditions under which system (6.1) becomes a completely integrable one.

Making use of the transformation (2.1), by calculations similar to those done in Sec. II, we find

$$\psi_1=\epsilon\rho\frac{(\overline{\partial\rho})^{1/2}}{h^{1/2}(1+|\rho|^2)},\quad \psi_2=\epsilon\frac{(\partial\rho)^{1/2}}{h^{1/2}(1+|\rho|^2)},\quad \epsilon^2=1, \tag{6.2}$$

and system (6.1) becomes

$$\bar{\partial}\partial\rho=\frac{2\bar{\rho}}{1+|\rho|^2}\partial\rho\bar{\partial}\rho+(\partial(\ln h))(\partial\rho), \tag{6.3a}$$

$$\partial\bar{\partial}\rho=\frac{2\rho}{1+|\rho|^2}\overline{\partial\rho}\,\overline{\partial\rho}+(\bar{\partial}(\ln h))(\overline{\partial\rho}). \tag{6.3b}$$

The above form is more convenient to analyze than the equations (6.1). Employing the conditional symmetry method,[30,31] we look for conditions necessary for solvability of a class of equations (6.3) which admit compatible first-order differential constraints. We consider here the simplest case where the differential constraints are based on an sl(2,$\mathbb{C}$) representation. So, we assume that they take the form of coupled Riccati equations (and their complex conjugate equations) with nonconstant coefficients,

$$\begin{aligned}\partial\rho&=A_1^0(z,\bar{z})+A_1^1(z,\bar{z})\rho+A_1^2(z,\bar{z})\rho^2,\\ \bar{\partial}\rho&=A_2^0(z,\bar{z})+A_2^1(z,\bar{z})\rho+A_2^2(z,\bar{z})\rho^2.\end{aligned} \tag{6.4}$$

The compatibility condition for the system (6.4) requires appending to it the zero curvature conditions,

$$A^l_{[\mu,\nu]}+\tfrac{1}{2}C^l_{ab}A^a_\mu A^b_\nu=0,\quad a,b,l=0,1,2\quad \mu,\nu=1,2, \tag{6.5}$$

where $(z^\mu)=(z,\bar{z})$ and C^l_{ab} are structure constants of sl(2,$\mathbb{C}$). The brackets $[\mu,\nu]$ denote here the alternation with respect to the indices μ and ν. We look for conditions on the function h which ensures that the overdetermined system composed of the equations (6.3), differential constraints (6.4), and conditions (6.5) are in involution. These involutivity conditions give us the specific differential restrictions on the class of function *h:*

$$\bar{\partial}\partial\left(\frac{1}{h}\right)=0. \tag{6.6}$$

The general solution of (6.6) is given by

$$h(z,\bar{z})=\frac{1}{r(z)+r(\bar{z})}. \tag{6.7}$$

Here, r is an arbitrary real function. Then, system (6.1) becomes

$$\begin{aligned}\partial\psi_1&=\frac{p}{r(z)+r(\bar{z})}\psi_2,\quad \overline{\partial\psi_1}=\frac{p}{r(z)+r(\bar{z})}\bar{\psi}_2,\\ \bar{\partial}\psi_2&=-\frac{p}{r(z)+r(\bar{z})}\psi_1,\quad \partial\bar{\psi}_2=-\frac{p}{r(z)+r(\bar{z})}\bar{\psi}_1,\end{aligned} \tag{6.8}$$

and, consequently, system (6.3) takes the form

$$
\begin{aligned}
\bar\partial\partial\rho &= \frac{2\bar\rho}{1+|\rho|^2}\,\partial\rho\,\bar\partial\rho - \frac{\partial r}{r(z)+r(\bar z)}\,(\partial\rho),\\
\bar\partial\partial\bar\rho &= \frac{2\rho}{1+|\rho|^2}\,\bar\partial\bar\rho\,\partial\bar\rho - \frac{\bar\partial r}{r(z)+r(\bar z)}\,(\bar\partial\bar\rho).
\end{aligned}
\tag{6.9}
$$

It is easy to show that system (6.8) cannot be transformed into the original WE system (1.1), corresponding to constant mean curvature surfaces, by any change of independent variables. Particular case of system (6.8) has been recently discussed.[32]

An analysis of system (6.9) similar to the one carried out in Sec. IV, can provide us with an explicit form of the spectral problem for (6.9). Since system (1.1) constitutes a special case of system (6.1), it is evident that our approach can be applied to systems which describe much more diverse types of surfaces. This task will be undertaken in future work.

ACKNOWLEDGMENTS

The authors thank J. Harnad (Concordia University) and B. Konopelchenko (Universita di Lecce) for helpful and interesting discussions on the topic of this paper. We also thank W. Hereman (Colorado School of Mines) for providing computer generated symmetry calculations of the WE system (1.2).

This work was supported in part by research grants from NSERC of Canada and the Fonds FCAR du Gouvernement du Québec, as well by the INFN-Sezione di Lecce and by MURST of Italy, and partially by the NATO Collaborative Program Research Grant No CRG 960717.

[1] A. R. Bishop, L. G. Campbell, and P. G. Channel (editors), *Fronts, Interfaces and Patterns* (North-Holland, Amsterdam, 1984).
[2] P. Pelce (editor), *Dynamics of Curved Fronts* (Academic, New York, 1986).
[3] D. Nelson, T. Piran, and S. Weinberg, *Statistical Mechanics of Membranes and Surfaces* (World Scientific, Singapore, 1989).
[4] D. G. Gross, C. N. Pope, and S. Weinberg (editors), *Two-dimensional Quantum Gravity and Random Surfaces* (World Scientific, Singapore, 1992).
[5] K. A. Brakke, *The Motion of a Surface by its Mean Curvature* (Princeton U.P., Princeton, NJ, 1978).
[6] S.-T. Yau (editor), *Seminar on Differential Geometry* (Princeton U.P., Princeton, NJ, 1982).
[7] H. C. Wente, Counterexample of a conjecture of H. Hopf, Pac. J. Math. **121**, 193–243 (1986).
[8] U. Pinkall and I. Sterling, ''On the classification of constant mean curvature tori,'' Ann. Math. **130**, 407–451 (1989).
[9] N. Kapouleas, ''Complete constant mean curvature surfaces in Euclidean three-space,'' Ann. Math. **131**, 239–330 (1990).
[10] R. Carroll and B. Konopelchenko, ''Generalized Weierstrass-Enneper inducing conformal immersions and gravity,'' Int. J. Mod. Phys. A **11(7)**, 1183–1216 (1996).
[11] K. Viswanathan and R. Parthasarathy, Ann. Phys. **206**, 237 (1991), Phys. Rev. D **51**, 5830 (1995).
[12] D. Amit, *Field Theory, the Renormalization Group and Critical Phenomena* (McGraw-Hill, New York, 1978).
[13] Z. Peradzynski, in *Geometry and Nonlinear Partial Differential Equations in the Cauchy Problem for Dissipative Media*, edited by W. Kosinski (Ossolineum, Wroclaw, 1979).
[14] B. L. Rozdestvenskii and N. N. Janenko, *Systems of Quasilinear Equations and their Applications to Gas Dynamics*, Vol. 55 (Ams, Providence, RI, 1983).
[15] K. Kenmotsu, ''Weierstrass Formula for surfaces of prescribed mean curvature,'' Math. Ann. **245**, 89–99 (1979).
[16] D. A. Hoffman and R. Osserman, ''The Gauss map of surfaces in $\mathbb{R}^3$ and $\mathbb{R}^4$,'' Proc. London Math. Soc. **50**, 27–56 (1985).
[17] R. Osserman, *A Survey of Minimal Surfaces* (Dover, New York, 1996).
[18] B. Konopelchenko, ''Induced surfaces and their integrable dynamics,'' Stud. Appl. Math. **96**, 9–51 (1996).
[19] B. G. Konopelchenko and I. A. Taimanov, ''Constant mean curvature surfaces via an integrable dynamical system,'' J. Phys. A **29**, 1261–1265 (1996).
[20] V. G. Kac, *Infinite Dimensional Lie Algebras* (Birkhauser, Boston, 1983).
[21] E. Cartan, *Sur la Structure des Groupes Infinis de Transformation*, Chapitre I, Les Systèmes Differential en Involution. (Gauthier-Villars, Paris, 1953).
[22] W. Slebodzinski, *Exterior Differential Forms and Their Applications* (Polish Scientific PWN, Warsaw, 1957).
[23] M. Kuranischi, *Lectures on Involutive Systems of Partial Differential Equations* (Brasil Math. Soc., Sao Paulo, 1967).
[24] R. L. Brayant, S. S. Chern, R. B. Gardner, H. L. Goldschmidt, and P. A. Griffiths, *Exterior Differential Systems*,

Mathematical Sciences Research Institute Publications, Vol. 18 (Springer-Verlag, New York, 1991).
[25] V. G. Makhankov and O. K. Pashaev, ''Integrable Pseudospin Models in Condensed Matter,'' Sov. Sci. Rev. Math. Phys. **9**, 1 (1992).
[26] A. V. Mikhailov, *Integrable Magnetic Models*, in *Solitons*, edited by S. E. Trullinger, V. E. Zakharov, and V. L. Pokrovsky (North-Holland, Amsterdam, 1986), pp. 623–690.
[27] M. Ablowitz and P. Clarkson, *Solitons Nonlinear Evolution Equations and Inverse Scattering* (Cambridge U.P., London, 1991).
[28] R. Rajamaran, *Solitons and Instantons* (North-Holland, Amsterdam, 1982).
[29] W. Zakrzewski, *Low Dimensional Sigma-Models* (Hilger, New York, 1989).
[30] A. M. Grundland, L. Martina, and G. Rideau, *Partial differential equations with differential constraints*, in CRM Proceedings and Lecture Notes (American Mathematical Society, Providence, RI, 1997), Vol. 11, pp. 135–154.
[31] I. Ayari and A. M. Grundland, ''Conditional symmetries for *k*-th order partial differential equations,'' preprint, CRM-2488, Université de Montréal (1997).
[32] A. I. Bobenko, ''Surfaces in terms of 2 by 2 matrices. Old and new integrable cases,'' in *Aspects of Mathematics*, A. P. Fordy and J. C. Wood (editors) (Vieweg, Wiesbaden, 1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Unitary deformations and complex soliton equations

Péter Varga[a)]
Institute of Mathematics and Informatics, Lajos Kossuth University, H-4010 Debrecen, Hungary

(Received 18 January 1999; accepted for publication 8 February 1999)

The generalized Lax-equation $\dot{L}=i(AL-LB)$ leaves the spectrum of L^*L invariant if A and B are self-adjoint operators. Consequently such equations possess many conversed quantities. With the help of this scheme, complex equations of Korteweg–de Vries type are derived. [S0022-2488(99)03607-5]

Since the Lax-equation

$$\dot{L}=PL-LP$$

is the infinitesimal form of the similarity transformation $L\to e^{tP}Le^{-tP}$, such equations leave invariant the ''spectrum'' of L,[1] ensuring the existence of nontrivial conserved quantities. In many cases this scheme leads to the complete integrability of the corresponding equation (see Ref. 2 for a textbook account). The following modification

$$\dot{L}=PL-LQ, \tag{1}$$

of the Lax-equation arises in several situation. It was noted by Drinfeld and Sokolov,[3] that the spectrum of L_1L_2 is preserved by the following transformations: $L_1\to e^{tP}L_1e^{-tQ}$, $L_2\to e^{tQ}L_2e^{-tP}$. The infinitesimal form of these transformations are

$$\dot{L}_1=PL_1-L_1Q, \quad \dot{L}_2=QL_2-L_2P.$$

With the choice of $L_1=\partial_x^2+u+\phi$, $L_2=\partial_x^2+u-\phi$ one obtains the Hirota–Satsuma equations.[4,5] Another example of this scheme is the constrained KP hierarchy (cKP).[6–10] Here the spectrum of $L_1^{-1}L_2$ is preserved by the $L_1\to e^{tP}L_1e^{-tQ}$, $L_2\to e^{tP}L_2e^{-tQ}$ transformations. A somewhat different type of Eq. (1) occurs in the works of Semenov–Tian–Shansky on integrable lattice systems. There P and Q are related by $Q=\tau(P)$, where τ is an automorphism which commutes with the ''R-matrix'' of the problem (see, Ref. 11 for details).

The generalized Lax-equation also arises if L can be regarded as a linear operator on a Hilbert-space $\mathfrak{h}$. In this case one can try to preserve the spectrums of L^*L and LL^* instead of L's one. In fact, it is fairly natural to associate the operator L^*L to L, since every $L\in\mathbf{B}(\mathfrak{h})$ can be uniquely written as $L=U|L|$, where L's absolute value $|L|=(L^*L)^{1/2}$ is a positive operator, while U is a partial isometry.[12]

The spectrum of L^*L is not preserved by general similarity transformations of L. However, the transformation

$$L\to e^{itP}Le^{-itQ}$$

does leave it invariant if e^{itP} and e^{-itQ} are unitary, i.e., if P and Q are self-adjoint operators. Indeed

[a)]Electronic mail: varga@math.klte.hu

$$LL^* \to (e^{itP}Le^{-itQ})((e^{-itQ})^* L^* (e^{itP})^*) = e^{itP}LL^* e^{-itP}.$$

These considerations suggest that it might be possible to obtain integrable equations in the following form:

$$\dot{L} = i(PL - LQ), \quad P = P^*, \quad Q = Q^*. \tag{2}$$

As an illustration, we apply this method to the $L = \partial_x^2 + v(x)\partial_x + u(x)$ operator, where $u(x)$ and $v(x)$ are complex functions. If $v(x) = 0$, then the obtained equations are simple reductions of the (complex) Hirota–Satsamura equations. The $v(x) = 0$ constraint is not preserved by the even flows of the hierarchy.

First, let us recall the Gelfand–Dickey[13,2] construction of the Korteweg–de Vries (KdV) hierarchy. Their starting point is the Schrödinger operator

$$L = \partial^2 + u(x),$$

$(\partial = \partial_x)$. Its square root $L^{1/2}$ is a pseudodifferential operator

$$L^{1/2} = \partial + l_{-1}^{[1/2]}\partial^{-1} + l_{-2}^{[1/2]}\partial^{-2} + l_{-3}^{[1/2]}\partial^{-3} + \cdots,$$

where the $l_i^{[1/2]}$'s are polynomials of u and its derivatives. They are recursively determined by the condition $(L^{1/2})^2 = L$. The crucial property of $L^{1/2}$ is that

$$[L, L^{1/2}] = 0.$$

Then

$$0 = [L, L^{k/2}] = [L, (L^{k/2})_+ + (L^{k/2})_-] \Rightarrow [L, (L^{k/2})_+] = -[L, (L^{k/2})_-],$$

where $(L^{k/2})_+$ is a differential operator containing the non-negative powers of ∂, while $(L^{k/2})_-$ consists of terms of negative powers of ∂. Since $[L, (L^{k/2})_+]$ is a differential operator, and $[L, (L^{k/2})_-]$ cannot contain positive powers of ∂, both expressions must be polynomials of u and its derivatives, so

$$\partial_{t_k} L = \partial_{t_k} u = [L, (L^{k/2})_+] = -[L, (L^{k/2})_-]$$

is a partial differential equation for $u(x,t)$.

With some minor modification, this scheme works for the generalized Lax-equation, too. As the self-adjointness of operators has an important role, we use the self-adjoint derivation $D = i\partial$ instead of ∂. Let

$$L = D^2 + v(x)D + u(x),$$

where u and v are complex functions of x. Instead of $L^{1/2}$ (which is not self-adjoint), we would like to obtain self-adjoint pseudodifferential operators

$$A = D + a_0 + a_{-1}D^{-1} + a_{-2}D^{-2} + \cdots,$$

$$B = D + b_0 + b_{-1}D^{-1} + b_{-2}D^{-2} + \cdots,$$

satisfying

$$i(AL - LB) = 0.$$

So

$$AL=LB\Rightarrow L^{-1}AL=B=B^*=L^*AL^{-1*}\Rightarrow A(LL^*)=(LL^*)A,$$

which implies that

$$A=(LL^*)^{1/4},\quad B=(L^*L)^{1/4}.$$

The operator equations

$$\partial_{t_k}L=\partial_{t_k}vD+\partial_{t_k}u=i\{(A^k)_+L-L(B^k)_+\}=-i\{(A^k)_-L-L(B^k)_-\} \tag{3}$$

generate integrable equations for $u(x,t)$ and $v(x,t)$. Note that the splitting $X=X_++X_-$ respects self-adjointness, i.e., $X=X^*\Rightarrow(X_+=X_+^*\wedge X_-=X_-^*)$. This property would not be true, if the adjoint was computed with respect to a more general $(\psi_1,\psi_2)=\int\overline{\psi_1}F\psi_2\,dx$ inner product, where F is some pseudodifferential operator. In the KdV hierarchy v can be set to zero, since the term containing ∂ drops out from

$$\partial_{t_k}L=\partial_{t_k}(\partial^2+u)=-[(\partial^2+u),(L^{k/2})_-]=-[\partial^2+u,l_{-1}^{[1/2]}\partial_{-1}+l_{-2}^{[1/2]}\partial_{-2}+\cdots].$$

This reduction does not necessarily work in our case, since

$$\begin{aligned}-i((A^k)_-(D^2+u)-(D^2+u)(B^k)_-)&=-i((a_{-1}^{[k]}D^{-1}+\cdots)(D^2+u)-(D^2+u)(b_{-1}^{[k]}D^{-1}+\cdots))\\&=-i(a_{-1}^{[k]}-b_{-1}^{[k]})D,\end{aligned}$$

so the constraints $v(x)=0$ can be violated by the evolution equation. Since the evolution of LL^* and L^*L have the usual Lax form

$$\partial_{t_k}(LL^*)=i[(A^k)_+,LL^*],\quad \partial_{t_k}(L^*L)=i[(B^k)_+,L^*L],$$

the standard machinery of the KdV equations[2] can be applied. For example, the integrals of motion are

$$\int \operatorname{res}A^k\,dx,\quad \int\operatorname{res}B^k\,dx,\quad k=1,2,3,\ldots,$$

where $\operatorname{res}X$ is the coefficient of D^{-1} in the pseudodifferential operator X. The fact that Eq. (3) can be embedded into the KdV hierarchy generated by the Lax-operator

$$L^{(4)}=D^4+u_3D^3+u_2D^2+u_1D1+u_0$$

ensures their integrability. The embedding is determined by the constraints

$$L^{(4)}=LL^*=(D^2+vD+u)(D^2+D\overline{v}+\overline{u})=D^4+(v+\overline{v})D^3+\cdots.$$

Since in the KdV hierarchy of $L^{(4)}$ the constraint $u_3=0$ is preserved by the evolution equations, the constraint $v+\overline{v}=0\to\Re v=0$ can be imposed on Eq. (3), too. However, even $v=0$ is compatible with (3) for odd k, as these equations are basically equivalent to the Hirota–Satsamura hierarchy.

Now we (me and a symbolic algebra package) compute the explicit forms of the first few of the hierarchy (3). For the odd flows, we present only their constrained, $v=0$ form, while for the even flows, v's value is constrained to be pure imaginary ($v=iw$).

(1) Flow:

$$\partial_{t_1}u=u'.$$

(2) Flow:

$$\partial_{t_2} w = 2\mathfrak{F} u',$$

$$\partial_{t_2} u = \frac{i}{2}(w''' - w'(2w' + 2w^2 - 4u) - w(u' - \overline{u}' - w'')).$$

(3) Flow:

$$\partial_{t_3} u = \tfrac{1}{8}(u''' - 3\overline{u}''' - 6uu' + 6u'\overline{u} + 12u\overline{u}').$$

(4) Flow:

$$\partial_{t_4} u = \partial_{t_4} w = 0.$$

(5) Flow:

$$\partial_{t_5} u = \tfrac{1}{32}\{-3u^{(5)} + 5\overline{u}^{(5)} + 5u'''(3u - \overline{u}) + 5\overline{u}'''(-5u + \overline{u}) + 5u'(-3u^2 + 6u\overline{u} + \overline{u}^2 - 3u'' - 5\overline{u}'') + 5\overline{u}'(4u^2 + 4u\overline{u} - 3u'' - 3\overline{u}'')\}.$$

The presented construction can be applied to other integrable systems, too. For example, a Kadomtsev–Petviashvilli type hierarchy[2] can be derived without any difficulty. Let

$$L = D + u_0 + u_{-1}D^{-1} + u_{-2}D^{-2} + \cdots,$$

$$\partial_m L = i(B_m L - L\overline{B_m}), \quad B_m = (LL^*)^m_+ \quad \overline{B_m} = (L^*L)^m_+ .$$

Then the zero-curvature condition

$$\partial_m B_n - \partial_n B_m - i[B_m, B_n] = 0$$

provides the equations of a KP type hierarchy. The proof of the zero-curvature condition is almost the same as for the standard KP equations:

$$\partial_m B_n - \partial_n B_m - i[B_m, B_n] = i\left(\sum_{j=0}^{n-1} (LL^*)^{n-1-j}([B_m L - L\overline{B_m}]L^* + L[\overline{B_m}L^* - L^* B_m])(LL^*)^j - (n \leftrightarrow m)\right)_+ - [B_m, B_n]$$

$$= i([B_m, (LL^*)^n]_+ - [B_n, (LL^*)^m]_+ - [B_m, B_n])$$

$$= i([B_n - (LL^*)^n, B_m - (LL^*)^m]_+ = [(LL^*)^m_-, (LL^*)^n_-]) = 0.$$

These computations are very similar to the ones which appear in the theory of the constrained KP hierarchy.[10] This is not surprising, since here we isospectrally deform LL^*, while in the case of the cKP hierarchy $L_1^{-1}L_2$'s deformation is isospectral. The verification of the commutativity of the flows ∂_m is standard.

Finally we present another variation on the theme of this paper. We determine those $L \to e^{tA}Le^{-tB}$ transformations which generate isospectral deformations of $L\overline{L}$. Since $L\overline{L} \to e^{tA}Le^{-tB}\overline{e^{tA}}\overline{L}\overline{e^{-tB}}$, we obtain that $B = \overline{A}$. With the $A = (L\overline{L})^{k/n}_+$, $B = (\overline{L}L)^{k/n}_+$ choices, it is possible to repeat the steps of the derivation of the Hirota–Satsamura type coupled KdV hierarchy. Note, however, that this sort of scheme does not work for the LL^T product. In that case A and B should satisfy $B = -A^T$, but this condition does not hold for $A = (LL^T)^\alpha$ and $B = (L^T L)^\alpha$. So it seems that it is possible to modify only the complex KdV equations.

ACKNOWLEDGMENT

We are gratefully acknowledge the suggestions and encouragement of László Fehér.

[1] P. Lax, Commun. Pure Appl. Math. **21**, 467 (1968).
[2] L. A. Dickey, *Soliton Equations and Hamiltonian Systems* (World Scientific, Singapore, 1991).
[3] V. G. Drinfeld and V. V. Sokolov, Sov. Math. Dokl. **23**, 457 (1981).
[4] R. Hirota and J. Satsuma, Phys. Lett. **89A**, 168 (1982).
[5] R. Dodd and A. Fordy, Phys. Lett. **89A**, 168 (1982).
[6] W. Oevel and W. Schief, Rev. Math. Phys. **6**, 1301 (1994).
[7] A. Kundu, W. Oevel, and W. Strampp, J. Math. Phys. **36**, 2972 (1995).
[8] A. Krichever, Physica D **87**, 14 (1995).
[9] L. A. Dickey, Lett. Math. Phys. **35**, 229 (1995).
[10] H. Aratyn, E. Nissimov, and S. Pacheva, Int. J. Mod. Phys. A **12**, 1265 (1977).
[11] M. A. Semenov-Tian-Shansky, Publ. Res. Inst. Math. Sci. **21**, 1237 (1985).
[12] G. K. Pedersen, *Analysis Now*, Graduate Text in Mathematics (Springer, New York, 1989).
[13] I. M. Gelfand and L. Dickey, Funkz. Anal. Priloz. **10**, 13–29 (1976).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

A finite-dimensional integrable system associated with the three-wave interaction equations

Yongtang Wu
Department of Computer Science, Hong Kong Baptist University, 224 Waterloo Road, Kowloon, Hong Kong, People's Republic of China

Xianguo Geng
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China and Department of Mathematics, Zhengzhou University, Zhengzhou, Henan 450052, People's Republic of China

(Received 2 November 1998; accepted for publication 17 March 1999)

Under a constraint between the potentials and the eigenfunctions, the 3×3 AKNS matrix spectral problem and its adjoint spectral problem associated with the three-wave interaction equations are nonlinearized so as to be a new finite-dimensional Hamiltonian system. A general scheme for generating involutive systems of conserved integrals and their two new generators are proposed, by which the finite-dimensional Hamiltonian system is further proved to be completely integrable in the Liouville sense. Moreover, the involutive solutions of the three-wave interaction equations are given.
[S0022-2488(99)00407-7]

I. INTRODUCTION

It has been known that there are several systematic approaches to obtain explicit solutions of soliton equations, such as the inverse scattering transformation, the Hirota technique, the algebra–geometric method, the polar expansion solution method, etc.[1–6] Some interesting explicit solutions have been found, the most important among which are pure-soliton solutions, finite-band solutions and polar expansion solutions.

The observation that all the explicit solutions mentioned above have a finite number of parameters, which means that they satisfy some kind of ordinary differential equations, suggests an important approach to get new finite-dimensional integrable systems from soliton equations. Recently an effective method, the so-called nonlinearization of eigenvalue problems or Lax pairs,[7–9] has been developed and applied to various soliton hierarchies associated with 2×2 zero-trace matrix spectral problems, from which a considerable number of new finite-dimensional systems are obtained that are completely integrable in the Liouville sense. Another important application of the nonlinearization method is that it provides a way of solving soliton equations, integrable nonlinear partial differential equations, by separation of spatial and temporal variables. At the same time the inter-relation between soliton equations and finite-dimensional integrable systems is revealed. The method is sometimes called the method of separation of variables for nonlinear partial differential equations, which generalizes the corresponding method for linear ones.

A similar method, the restricted flow technique, for bi-Hamiltonian soliton hierarchies is proposed in Refs. 10, 11 and bi-Hamiltonian structures for the resulting finite-dimensional integrable systems can also be worked out through a Miura map.[11,12] There are attempts to apply the nonlinearization method or the restricted flow technique to discrete systems in order to get integrable symplectic maps.[13–17]

Very recently, the nonlinearization method has been generalized to discuss Lax pairs and adjoint Lax pairs of soliton equations[18–21] so that it may also be suitable for the cases of 2×2 nonzero-trace matrix spectral problems. It has been applied successfully to the hierarchy of Harry

Dym type equations and the Blaszak discrete soliton hierarchy,[20,21] which correspond to 3×3 matrix spectral problems.

The key to the complete integrability of a finite-dimensional Hamiltonian system is the existence of an involutive system of conserved integrals. However, it is difficult for us to search for an involutive system of conserved integrals of a given finite-dimensional Hamiltonian system. In this paper, based on the above works we are going to discuss the nonlinearization for a 3×3 AKNS matrix spectral problem and its adjoint spectral problem associated with the three-wave interaction equations,[1,22] from which a new finite-dimensional Hamiltonian system is obtained. Resorting to the characteristic polynomial of solution matrix of the stationary zero-curvature equation, we propose a general scheme for generating involutive systems of enough conserved integrals of the resulting finite-dimensional Hamiltonian system. To prove the functional independence of conserved integrals, two new generators of involutive systems of conserved integrals are introduced, which are two natural generalizations of the 2×2 case.[23] This shows that the finite-dimensional Hamiltonian system is completely integrable in the Liouville sense.

Consider the $n\times n$ matrix spectral problem

$$\psi_x=U(u,\lambda)\psi,\quad \psi=(\psi^1,\ldots,\psi^n)^T. \tag{1.1}$$

In order to derive the isospectral hierarchy associated with Eq. (1.1), we proceed first to solve the stationary zero-curvature equation,

$$V_x-[U,V]=0,\quad V=\sum_{j\geqslant0}V^{(j)}\lambda^{-j}, \tag{1.2}$$

which usually is equivalent to Lenard recursive equation

$$KG_{j-1}=JG_j,\quad JG_{-1}=0,\quad j\geqslant0. \tag{1.3}$$

Here K and J are two skew-symmetric operators. The soliton hierarchy $u_t=JG_m$ has a Lax pair, the spectral problem (1.1) and the auxiliary problem

$$\psi_{t_m}=V_m\psi,\quad V_m=(\lambda^mV)_+, \tag{1.4}$$

where the symbol $+$ stands for the choice of non-negative power of λ. The introduction of the adjoint problem of Eq. (1.1),

$$\phi_x=-U(u,\lambda)^T\phi,\quad \phi=(\phi^1,\ldots,\phi^n)^T \tag{1.5}$$

allows the calculation of the functional gradient of the eigenvalues with regard to the potential u (see, e.g., Sec. III). Usually such a functional gradient $\nabla\lambda_j$ satisfies the following equation:

$$K\nabla\lambda_j=\rho(\lambda_j)J\nabla\lambda_j,\quad \rho(\lambda_j)=c_1\lambda_j+c_2\lambda_j^2,\quad 1\leqslant j\leqslant N, \tag{1.6}$$

where $\lambda_1,\ldots,\lambda_N$ are the eigenvalues of Eqs. (1.1) and (1.5), c_1 and c_2 are constants. The following two kinds of constraints:

$$G_0=\sum_{j=1}^N\nabla\lambda_j,\quad G_{-1}=\sum_{j=1}^N\nabla\lambda_j, \tag{1.7}$$

which are called the Bargmann and Neumann constraints, respectively, play a central role in the process of nonlinearization of the eigenvalue problems (1.1) and (1.5). From (1.7) we can obtain the relations

$$u=f(q,p)\quad\text{and}\quad g(q,p)=0,\quad u=f(q,p), \tag{1.8}$$

where $q=(q_1^1,\ldots,q_N^1,\ldots,q_1^n,\ldots,q_N^n)^T$, $p=(p_1^1,\ldots,p_N^1,\ldots,p_1^n,\ldots,p_N^n)^T$, $q_j^i=\psi^i(\lambda_j)$, $p_j^i=\phi^i(\lambda_j)$, $1\leqslant i\leqslant n$, $1\leqslant j\leqslant N$. Under the two constraints, N replicas of the spectral problems (1.1) and (1.5) associated with $\lambda_1,\ldots,\lambda_N$ are nonlinearized into two finite-dimensional Hamiltonian systems

$$q_x=U(f(q,p),\Lambda)q,\quad p_x=-U(f(q,p),\Lambda)^Tp,\quad \Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_N), \tag{1.9}$$

$$q_x=U(f(q,p),\Lambda)q,\quad p_x=-U(f(q,p),\Lambda)^Tp,\quad g(q,p)=0, \tag{1.10}$$

which are called the Bargmann and Neumann systems, respectively. In the following, we propose a general scheme for generating conserved integrals of Eq. (1.9) or (1.10). Noticing the matrix $\mu I-V$ is also a solution of the stationary zero-curvature equation (1.2), which implies that $\det(\mu I-V)$ is a constant with respect to the variable x, for each value of the spectral parameter λ. Here μ is a parameter, I is an $n\times n$ matrix. Let us consider the characteristic polynomial of solution matrix V of Eq. (1.2),

$$\det(\mu I-V)=\mu^n-\mathcal{F}_\lambda^{(0)}\mu^{n-1}+\mathcal{F}_\lambda^{(1)}\mu^{n-2}+\cdots+(-1)^n\mathcal{F}_\lambda^{(n-1)}, \tag{1.11}$$

where

$$\mathcal{F}_\lambda^{(0)}=\mathrm{tr}\,V,\quad \mathcal{F}_\lambda^{(1)}=\sum_{1\leqslant i<j\leqslant n}\begin{vmatrix}V_{ii} & V_{ij}\\ V_{ji} & V_{jj}\end{vmatrix},$$

$$\mathcal{F}_\lambda^{(2)}=\sum_{1\leqslant i<j<k\leqslant n}\begin{vmatrix}V_{ii} & V_{ij} & V_{ik}\\ V_{ji} & V_{jj} & V_{jk}\\ V_{ki} & V_{kj} & V_{kk}\end{vmatrix},\ldots,\quad \mathcal{F}_\lambda^{(n-1)}=\det V. \tag{1.12}$$

By using Eqs. (1.3), (1.6) and the corresponding constraint, Eqs. (1.12) are reduced to generating functions of the conserved integrals of the Bargmann system (1.9), or the Neumann system (1.10). Sometimes some modifications are made, especially for the Neumann system. Thus we easily get the conserved integrals of the system (1.9) or (1.10) from their generating functions.

The outline of the paper is as follows: In Sec. II, we shall reconstruct the soliton hierarchy associated with the 3×3 AKNS spectral problem and establish their Hamiltonian structures. In Sec. III, we shall introduce the Bargmann constraint between the potentials and eigenfunctions. Under the constraint, a new finite-dimensional Hamiltonian system is obtained by nonlinearization of the 3×3 AKNS spectral problem and its adjoint one. In Sec. IV, we shall show how the scheme is applied to generate involutive systems of conserved integrals of the finite-dimensional Hamiltonian system. Further we prove that the finite-dimensional Hamiltonian system is completely integrable in the Liouville sense. In Sec. V, the representation of involutive solutions of the three-wave interaction equations and the soliton hierarchy is given. Finally in the Appendix, by means of the generating functions of conserved integrals, we give the involutivity of conserved integrals. Then we introduce two new generators of involutive systems of conserved integrals and prove the functional independence of conserved integrals.

II. THE SOLITION HIERARCHY AND HAMILTONIAN STRUCTURES

Let us consider the 3×3 AKNS matrix spectral problem

$$\psi_x=U(u,\lambda)\psi,\quad \psi=\begin{pmatrix}\psi^1\\ \psi^2\\ \psi^3\end{pmatrix},\quad U=\begin{pmatrix}\alpha_1\lambda & u_{12} & u_{13}\\ u_{21} & \alpha_2\lambda & u_{23}\\ u_{31} & u_{32} & \alpha_3\lambda\end{pmatrix}, \tag{2.1}$$

where the potential $u=(u_{12},u_{21},u_{13},u_{31},u_{23},u_{32})^T$, λ is a constant spectral parameter, $\alpha_i' s(1\leqslant i\leqslant 3)$ are three distinct constants. Our aim is to engender the soliton hierarchy from the spectral problem (2.1). To this end, we first solve the stationary zero-curvature equation,

$$V_x-[U,V]=0,\quad V=(V_{ij})_{3\times 3}, \tag{2.2}$$

which is equivalent to

$$V_{ijx}+u_{ij}(V_{ii}-V_{jj})+\sum_{\substack{k=1\\k\neq i,j}}^{3}(u_{kj}V_{ik}-u_{ik}V_{kj})-\lambda(\alpha_i-\alpha_j)V_{ij}=0,\quad i\neq j, \tag{2.3a}$$

$$V_{iix}=\sum_{\substack{k=1\\k\neq i}}^{3}(u_{ik}V_{ki}-u_{ki}V_{ik}),\quad 1\leqslant i,j\leqslant 3. \tag{2.3b}$$

Substitution the expansion

$$V_{ij}=\sum_{n\geqslant 0}V_{ij}^{(n)}\lambda^{-n} \tag{2.4}$$

into Eq. (2.3), we obtain the recurrence relations

$$V_{iix}^{(0)}=0,\quad V_{ij}^{(0)}=0,\quad (i\neq j),$$

$$V_{ijx}^{(n)}+u_{ij}(V_{ii}^{(n)}-V_{jj}^{(n)})+\sum_{\substack{k=1\\k\neq i,j}}^{3}(u_{kj}V_{ik}^{(n)}-u_{ik}V_{kj}^{(n)})-(\alpha_i-\alpha_j)V_{ij}^{(n+1)}=0,\quad i\neq j,$$

$$V_{iix}^{(n)}=\sum_{\substack{k=1\\k\neq i}}^{3}(u_{ik}V_{ki}^{(n)}-u_{ki}V_{ik}^{(n)}),\quad 1\leqslant i,j\leqslant 3,n\geqslant 0. \tag{2.5}$$

By Eq. (2.5) we have

$$V_{ii}^{(0)}=\beta_i(\text{constant}),\qquad V_{ij}^{(0)}=0,\quad i\neq j,$$

$$V_{ii}^{(1)}=0,\quad V_{ij}^{(1)}=\frac{\beta_i-\beta_j}{\alpha_i-\alpha_j}u_{ij},\quad i\neq j, \tag{2.6}$$

and require that

$$\beta_i\neq\beta_j(i\neq j),\qquad V_{ij}^{(n)}|_{u=0}=0,\quad n\geqslant 1, \tag{2.7}$$

where the condition (2.7) means to identify constants of the integration to be zero. Hence $V_{ij}^{(n)}$ is uniquely determined. It is easy to calculate that

$$V_{ij}^{(2)}=\frac{\beta_i-\beta_j}{(\alpha_i-\alpha_j)^2}u_{ijx}+\frac{1}{\alpha_i-\alpha_j}\sum_{\substack{k=1\\k\neq i,j}}^{3}\left(\frac{\beta_k-\beta_i}{\alpha_k-\alpha_i}-\frac{\beta_k-\beta_j}{\alpha_k-\alpha_j}\right)u_{ik}u_{kj},\qquad i\neq j, \tag{2.8a}$$

$$V_{ii}^{(2)}=\sum_{\substack{k=1\\k\neq i}}^{3}\frac{\beta_k-\beta_i}{(\alpha_k-\alpha_i)^2}u_{ik}u_{ki}. \tag{2.8b}$$

Equations (2.5)–(2.7) can be equivalently written as the Lenard form

$$KG_{n-1}=JG_n\,,\quad G_{n-1}=(V_{21}^{(n)},V_{12}^{(n)},V_{31}^{(n)},V_{13}^{(n)},V_{32}^{(n)},V_{23}^{(n)})^T,\qquad n\geqslant 1, \tag{2.9a}$$

$$G_0=\left(\frac{\beta_1-\beta_2}{\alpha_1-\alpha_2}u_{21},\frac{\beta_1-\beta_2}{\alpha_1-\alpha_2}u_{12},\frac{\beta_1-\beta_3}{\alpha_1-\alpha_3}u_{31},\frac{\beta_1-\beta_3}{\alpha_1-\alpha_3}u_{13},\frac{\beta_2-\beta_3}{\alpha_2-\alpha_3}u_{32},\frac{\beta_2-\beta_3}{\alpha_2-\alpha_3}u_{23}\right)^T \tag{2.9b}$$

with $G_n|_{u=0}=0$. Here J and K are two skew-symmetric operators,

$$J=\begin{pmatrix} 0 & \alpha_1-\alpha_2 & 0 & 0 & 0 & 0\\ \alpha_2-\alpha_1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & \alpha_1-\alpha_3 & 0 & 0\\ 0 & 0 & \alpha_3-\alpha_1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & \alpha_2-\alpha_3\\ 0 & 0 & 0 & 0 & \alpha_3-\alpha_2 & 0 \end{pmatrix},$$

$$K=(K_{lm})_{6\times 6},\quad K_{ml}^*=-K_{lm}$$

with

$$K_{11}=2u_{12}\partial^{-1}u_{12},\quad K_{12}=\partial-2u_{12}\partial^{-1}u_{21},\quad K_{13}=u_{12}\partial^{-1}u_{13},\quad K_{14}=u_{32}-u_{12}\partial^{-1}u_{31},$$

$$K_{15}=-u_{13}-u_{12}\partial^{-1}u_{23},\quad K_{16}=u_{12}\partial^{-1}u_{32},\quad K_{22}=2u_{21}\partial^{-1}u_{21},\quad K_{23}=-u_{23}-u_{21}\partial^{-1}u_{13},$$

$$K_{24}=u_{21}\partial^{-1}u_{31},\quad K_{25}=u_{21}\partial^{-1}u_{23},\quad K_{26}=u_{31}-u_{21}\partial^{-1}u_{32},\quad K_{33}=2u_{13}\partial^{-1}u_{13},$$

$$K_{34}=\partial-2u_{13}\partial^{-1}u_{31},\quad K_{35}=u_{13}\partial^{-1}u_{23},\quad K_{36}=-u_{12}-u_{13}\partial^{-1}u_{32},\quad K_{44}=2u_{31}\partial^{-1}u_{31},$$

$$K_{45}=u_{21}-u_{31}\partial^{-1}u_{23},\quad K_{46}=u_{31}\partial^{-1}u_{32},\quad K_{55}=2u_{23}\partial^{-1}u_{23},\quad K_{56}=\partial-2u_{23}\partial^{-1}u_{32},$$

$$K_{66}=2u_{32}\partial^{-1}u_{32},\quad \partial=\partial/\partial x,\quad \partial\partial^{-1}=\partial^{-1}\partial=1.$$

Using the trace identity technique,[24] we have

$$\frac{\delta H_n}{\delta u_{ij}}=V_{ji}^{(n)},\quad H_n=-\frac{1}{n}(\alpha_1V_{11}^{(n+1)}+\alpha_2V_{22}^{(n+1)}+\alpha_3V_{33}^{(n+1)}), \tag{2.10}$$

where the potentials $u_{ij},i\neq j$, are assumed to belong to the Schwartz space $\mathcal{S}(\Omega)$, $\Omega=(-\infty,\infty)$.

Now we introduce the auxiliary problem of the spectral problem (2.1),

$$\psi_{t_m}=V^{(m)}\psi,\quad V^{(m)}=V^{(m)}(u,\lambda)=(\lambda^mV)_+\,,\quad m\geqslant 1. \tag{2.11}$$

The compatibility condition between Eqs. (2.1) and (2.11) leads to the zero-curvature equation, $U_t-V_x+[U,V]=0$, that is the hierarchy of soliton equations with bi-Hamiltonian forms

$$u_{t_m}=X_m=K\frac{\delta H_m}{\delta u}=J\frac{\delta H_{m+1}}{\delta u},\quad m\geqslant 1, \tag{2.12}$$

where the vector field $X_m=KG_{m-1}=JG_m$ and Eq. (2.10) is used. The typical nonlinear system in the hierarchy is the famous three-wave interaction equations,

$$u_{ijt_1}=\frac{\beta_i-\beta_j}{\alpha_i-\alpha_j}u_{ijx}+\sum_{\substack{k=1\\k\neq i,j}}^{3}\left(\frac{\beta_k-\beta_i}{\alpha_k-\alpha_i}-\frac{\beta_k-\beta_j}{\alpha_k-\alpha_j}\right)u_{ik}u_{kj},\quad i\neq j,\quad 1\leqslant i,\quad j\leqslant 3, \tag{2.13}$$

which has important applications in physics.[1,22,25]

III. A FINITE-DIMENSIONAL HAMILTONIAN SYSTEM

In order to give the constraint between the potentials and the eigenfunctions, it is necessary to calculate the functional gradient of the eigenvalue with respect to the potential. We introduce the adjoint spectral problem of (2.1),

$$\phi_x=-U(u,\lambda)^T\phi,\quad \phi=(\phi^1,\phi^2,\phi^3)^T. \tag{3.1}$$

Suppose that $u_{ij}\to u_{ij}+\epsilon\delta u_{ij}$, $1\leqslant i\neq j\leqslant 3$, denote $\partial/\partial\epsilon|_{\epsilon=0}$ by a dot. The underlying interval Ω is $(-\infty,\infty)$ under the decaying condition at infinity. A direct calculation shows by Eqs. (2.1) and (3.1) that

$$(\phi^T\dot{\psi})_x=\phi^T\dot{U}\psi. \tag{3.2}$$

If λ is an eigenvalue of the spectral problems (2.1) and (3.1), the integration of the left-hand side of equality (3.2) vanishes because of the boundary conditions. Then we have

$$\int_\Omega \phi^T\dot{U}\psi dx=0. \tag{3.3}$$

Let $\lambda_1,\ldots,\lambda_N$ be N distinct eigenvalues. Then the systems associated with spectral problems (2.1) and (3.1) can be written in the form

$$(q_l^1,q_l^2,q_l^3)_x=(q_l^1,q_l^2,q_l^3)U(u,\lambda_l)^T,\quad (p_l^1,p_l^2,p_l^3)_x=-(p_l^1,p_l^2,p_l^3)U(u,\lambda_l), \tag{3.4}$$

where $q_l^i=\psi^i(\lambda_l)$, $p_l^i=\phi^i(\lambda_l)$, $1\leqslant i\leqslant 3$, $1\leqslant l\leqslant N$, are eigenfunctions. Noticing Eq. (3.4) and $\dot{U}(u,\lambda_l)=U(\delta u,\delta\lambda_l)$, we obtain by Eq. (3.3) that

$$\int_\Omega(p_l^1,p_l^2,p_l^3)U(\delta u,\delta\lambda_l)(q_l^1,q_l^2,q_l^3)^Tdx=0,$$

which implies that the functional gradient of the eigenvalue λ_l with regard to the potential u is

$$\nabla\lambda_l=\frac{\delta\lambda_l}{\delta u}=\left(\frac{\delta\lambda_l}{\delta u_{12}},\frac{\delta\lambda_l}{\delta u_{21}},\frac{\delta\lambda_l}{\delta u_{13}},\frac{\delta\lambda_l}{\delta u_{31}},\frac{\delta\lambda_l}{\delta u_{23}},\frac{\delta\lambda_l}{\delta u_{32}}\right)^T=(q_l^2p_l^1,q_l^1p_l^2,q_l^3p_l^1,q_l^1p_l^3,q_l^3p_l^2,q_l^2p_l^3)^T, \tag{3.5}$$

where we assume that

$$\int_\Omega(\alpha_1q_l^1p_l^1+\alpha_2q_l^2p_l^2+\alpha_3q_l^3p_l^3)dx=-1.$$

A direct calculation shows that the gradient $\nabla\lambda_l$ satisfies the following equation:

$$K\nabla\lambda_l=\lambda_lJ\nabla\lambda_l. \tag{3.6}$$

As a matter of fact, we obtain from Eq. (3.4) that

$$(q_l^1p_l^1)_x-(q_l^2p_l^2)_x=2u_{12}q_l^2p_l^1-2u_{21}q_l^1p_l^2+u_{13}q_l^3p_l^1-u_{31}q_l^1p_l^3-u_{23}q_l^3p_l^2+u_{32}q_l^2p_l^3, \tag{3.7}$$

$$(q_l^1 p_l^2)_x + u_{32} q_l^1 p_l^3 - u_{13} q_l^3 p_l^2 = u_{12}(q_l^2 p_l^2 - q_l^1 p_l^1) + \lambda_l(\alpha_1 - \alpha_2) q_l^1 p_l^2. \tag{3.8}$$

Noticing Eqs. (3.7) and (3.8), we have

$$\begin{aligned} &K_{11} q_l^2 p_l^1 + K_{12} q_l^1 p_l^2 + K_{13} q_l^3 p_l^1 + K_{14} q_l^1 p_l^3 + K_{15} q_l^3 p_l^2 + K_{16} q_l^2 p_l^3 \\ &\quad = u_{12} \partial^{-1}(2u_{12} q_l^2 p_l^1 - 2u_{21} q_l^1 p_l^2 + u_{13} q_l^3 p_l^1 - u_{31} q_l^1 p_l^3 - u_{23} q_l^3 p_l^2 + u_{32} q_l^2 p_l^3) \\ &\qquad + (q_l^1 p_l^2)_x + u_{32} q_l^1 p_l^3 - u_{13} q_l^3 p_l^2 = \lambda_l(\alpha_1 - \alpha_2) q_l^1 p_l^2, \end{aligned} \tag{3.9}$$

which implies that the sign of equality in the first row of Eq. (3.6) holds. In a similar way, we may prove that other rows of Eq. (3.6) are also identical equations.

Now we consider the Bargmann constraint

$$G_0 = \sum_{l=1}^{N} \nabla \lambda_l , \tag{3.10}$$

which implies

$$u_{ij} = \frac{\alpha_i - \alpha_j}{\beta_i - \beta_j} \langle q^i, p^j \rangle, \quad i \neq j, \ \ 1 \leqslant i,j \leqslant 3, \tag{3.11}$$

where $\langle \cdot , \cdot \rangle$ is the standard inner-product in $\mathcal{R}^N$, $q^i = (q_1^i, \ldots, q_N^i)^T$, $p^i = (p_1^i, \ldots, p_N^i)^T$. Substituting Eq. (3.11) into Eq. (3.4), we obtain a finite-dimensional Hamiltonian system

$$q_x^i = \frac{\partial H}{\partial p^i}, \quad p_x^i = -\frac{\partial H}{\partial q^i}, \quad 1 \leqslant i \leqslant 3, \tag{3.12}$$

with the Hamiltonian

$$\begin{aligned} H = {} & \alpha_1 \langle \Lambda q^1, p^1 \rangle + \alpha_2 \langle \Lambda q^2, p^2 \rangle + \alpha_3 \langle \Lambda q^3, p^3 \rangle + \frac{\alpha_1 - \alpha_2}{\beta_1 - \beta_2} \langle q^1, p^2 \rangle \langle q^2, p^1 \rangle \\ & + \frac{\alpha_1 - \alpha_3}{\beta_1 - \beta_3} \langle q^1, p^3 \rangle \langle q^3, p^1 \rangle + \frac{\alpha_2 - \alpha_3}{\beta_2 - \beta_3} \langle q^2, p^3 \rangle \langle q^3, p^2 \rangle, \end{aligned}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$. Moreover, we have from Eq. (3.4) that

$$[(q_l^1, q_l^2, q_l^3)(p_l^1, p_l^2, p_l^3)^T]_x = 0,$$

which implies

$$q_l^1 p_l^1 + q_l^2 p_l^2 + q_l^3 p_l^3 = \text{constant}. \tag{3.13}$$

It is easy to see that Eq. (3.13) are the conserved integrals of the system (3.12).

IV. THE INTEGRABILITY

In this section, we shall show how the characteristic polynomial of the solution matrix of Eq. (2.2) is used to generate the involutive systems of conserved integrals of the finite-dimensional Hamiltonian system (3.12). To this end, we first consider the characteristic polynomial

$$\det(\mu I - V) = \mu^3 - \mathcal{F}_\lambda^{(0)} \mu^2 + \mathcal{F}_\lambda^{(1)} \mu - \mathcal{F}_\lambda^{(2)}, \tag{4.1}$$

where

$$\mathcal{F}_\lambda^{(0)}=\operatorname{tr} V,\quad \mathcal{F}_\lambda^{(1)}=\sum_{1\leqslant i<j\leqslant 3}\begin{vmatrix} V_{ii} & V_{ij} \\ V_{ji} & V_{jj}\end{vmatrix},\quad \mathcal{F}_\lambda^{(2)}=\begin{vmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33}\end{vmatrix}. \tag{4.2}$$

It is easy to see that $\mathcal{F}_\lambda^{(0)},\mathcal{F}_\lambda^{(1)},\mathcal{F}_\lambda^{(2)}$ are constants with respect to the variable x. For the sake of convenience, we introduce a bilinear function $\overset{ij}{Q}_\lambda$ on $\mathcal{R}^N$ and its partial-fraction expansion and Laurent expansion,

$$\overset{ij}{Q}_\lambda=\langle(\lambda-\Lambda)^{-1}q^i,p^j\rangle=\sum_{l=1}^{N}\frac{q_l^i p_l^j}{\lambda-\lambda_l}=\sum_{n\geqslant 0}\lambda^{-n-1}\langle\Lambda^n q^i,p^j\rangle.$$

By using Eqs. (2.9), (3.6) and the constraint (3.10), we take the following restriction:

$$G_n=\sum_{l=1}^{N}\lambda_l^n\nabla\lambda_l, \tag{4.3}$$

which is a special solution of Eq. (2.9) and can be written as follows:

$$V_{ij}^{(n)}=\langle\Lambda^{n-1}q^i,p^j\rangle,\quad i\neq j,\quad 1\leqslant i,\quad j\leqslant 3,\quad n\geqslant 1. \tag{4.4a}$$

From Eq. (4.4a) and the third expression of Eq. (2.5), we have

$$V_{ii}^{(n)}=\langle\Lambda^{n-1}q^i,p^i\rangle,\quad 1\leqslant i\leqslant 3, n\geqslant 1. \tag{4.4b}$$

By utilizing Eqs. (2.4) and (4.4), we get

$$V_{ij}=\sum_{n\geqslant 1}\langle\Lambda^{n-1}q^i,p^j\rangle\lambda^{-n}=\overset{ij}{Q}_\lambda,\quad i\neq j,\quad 1\leqslant i,j\leqslant 3, \tag{4.5a}$$

$$V_{ii}=\beta_i+\sum_{n\geqslant 1}\langle\Lambda^{n-1}q^i,p^i\rangle\lambda^{-n}=\beta_i+\overset{ii}{Q}_\lambda,\quad 1\leqslant i\leqslant 3. \tag{4.5b}$$

Substituting Eq. (4.5) into Eq. (4.2) yields generating functions of integrals of motion for Eq. (3.12),

$$\hat{\mathcal{F}}_\lambda^{(0)}=\overset{11}{Q}_\lambda+\overset{22}{Q}_\lambda+\overset{33}{Q}_\lambda, \tag{4.6}$$

$$\hat{\mathcal{F}}_\lambda^{(1)}=(\beta_2+\beta_3)\overset{11}{Q}_\lambda+(\beta_1+\beta_3)\overset{22}{Q}_\lambda+(\beta_1+\beta_2)\overset{33}{Q}_\lambda+\sum_{1\leqslant i<j\leqslant 3}(\overset{ii}{Q}_\lambda\overset{jj}{Q}_\lambda-\overset{ij}{Q}_\lambda\overset{ji}{Q}_\lambda), \tag{4.7}$$

$$\hat{\mathcal{F}}_\lambda^{(2)}=\beta_2\beta_3\overset{11}{Q}_\lambda+\beta_1\beta_3\overset{22}{Q}_\lambda+\beta_1\beta_2\overset{33}{Q}_\lambda+\sum_{1\leqslant i<j\leqslant 3}\beta_{6-i-j}(\overset{ii}{Q}_\lambda\overset{jj}{Q}_\lambda-\overset{ij}{Q}_\lambda\overset{ji}{Q}_\lambda)+\begin{vmatrix} \overset{11}{Q}_\lambda & \overset{12}{Q}_\lambda & \overset{13}{Q}_\lambda \\ \overset{21}{Q}_\lambda & \overset{22}{Q}_\lambda & \overset{23}{Q}_\lambda \\ \overset{31}{Q}_\lambda & \overset{32}{Q}_\lambda & \overset{33}{Q}_\lambda\end{vmatrix}, \tag{4.8}$$

where $\hat{\mathcal{F}}_\lambda^{(0)}$, $\hat{\mathcal{F}}_\lambda^{(1)}$, and $\hat{\mathcal{F}}_\lambda^{(2)}$ are defined by

$$\hat{\mathcal{F}}_\lambda^{(0)}=\hat{\mathcal{F}}_\lambda^{(0)}-\beta_1-\beta_2-\beta_3,\quad \hat{\mathcal{F}}_\lambda^{(1)}=\mathcal{F}_\lambda^{(1)}-\beta_1\beta_2-\beta_1\beta_3-\beta_2\beta_3,$$

$$\hat{\mathcal{F}}_\lambda^{(2)}=\mathcal{F}_\lambda^{(2)}-\beta_1\beta_2\beta_3.$$

Substituting the Laurent expansion of $\overset{ik}{Q_\lambda}$ into Eqs. (4.6)–(4.8), respectively, we have

$$\hat{\mathcal{F}}_\lambda^{(0)}=\sum_{m\geqslant 0}\lambda^{-m-1}F_m^{(0)},\quad \hat{\mathcal{F}}_\lambda^{(1)}=\sum_{m\geqslant 0}\lambda^{-m-1}F_m^{(1)},\quad \hat{\mathcal{F}}_\lambda^{(2)}=\sum_{m\geqslant 0}\lambda^{-m-1}F_m^{(2)},\tag{4.9}$$

where

$$F_m^{(0)}=\langle\Lambda^m q^1,p^1\rangle+\langle\Lambda^m q^2,p^2\rangle+\langle\Lambda^m q^3,p^3\rangle,\quad m\geqslant 0,\tag{4.10}$$

$$F_0^{(1)}=(\beta_2+\beta_3)\langle q^1,p^1\rangle+(\beta_1+\beta_3)\langle q^2,p^2\rangle+(\beta_1+\beta_2)\langle q^3,p^3\rangle,\tag{4.11a}$$

$$F_m^{(1)}=(\beta_2+\beta_3)\langle\Lambda^m q^1,p^1\rangle+(\beta_1+\beta_3)\langle\Lambda^m q^2,p^2\rangle+(\beta_1+\beta_2)\langle\Lambda^m q^3,p^3\rangle$$
$$+\sum_{1\leqslant i<j\leqslant 3}\sum_{l=1}^{m}\begin{vmatrix}\langle\Lambda^{l-1}q^i,p^i\rangle & \langle\Lambda^{m-l}q^j,p^i\rangle\\ \langle\Lambda^{l-1}q^i,p^j\rangle & \langle\Lambda^{m-l}q^j,p^j\rangle\end{vmatrix},\quad m\geqslant 1,\tag{4.11b}$$

$$F_0^{(2)}=\beta_2\beta_3\langle q^1,p^1\rangle+\beta_1\beta_3\langle q^2,p^2\rangle+\beta_1\beta_2\langle q^3,p^3\rangle,\tag{4.12a}$$

$$F_1^{(2)}=\beta_2\beta_3\langle\Lambda q^1,p^1\rangle+\beta_1\beta_3\langle\Lambda q^2,p^2\rangle+\beta_1\beta_2\langle\Lambda q^3,p^3\rangle$$
$$+\sum_{1\leqslant i<j\leqslant 3}\beta_{6-i-j}\begin{vmatrix}\langle q^i,p^i\rangle & \langle q^j,p^i\rangle\\ \langle q^i,p^j\rangle & \langle q^j,p^j\rangle\end{vmatrix},\tag{4.12b}$$

$$F_m^{(2)}=\beta_2\beta_3\langle\Lambda^m q^1,p^1\rangle+\beta_1\beta_3\langle\Lambda^m q^2,p^2\rangle+\beta_1\beta_2\langle\Lambda^m q^3,p^3\rangle$$
$$+\sum_{1\leqslant i<j\leqslant 3}\sum_{\substack{l+n=m-1\\ l,n\geqslant 0}}\beta_{6-i-j}\begin{vmatrix}\langle\Lambda^l q^i,p^i\rangle & \langle\Lambda^n q^j,p^i\rangle\\ \langle\Lambda^l q^i,p^j\rangle & \langle\Lambda^n q^j,p^j\rangle\end{vmatrix}$$
$$+\sum_{\substack{l+n+s=m-2\\ l,n,s\geqslant 0}}\begin{vmatrix}\langle\Lambda^l q^1,p^1\rangle & \langle\Lambda^n q^1,p^2\rangle & \langle\Lambda^s q^1,p^3\rangle\\ \langle\Lambda^l q^2,p^1\rangle & \langle\Lambda^n q^2,p^2\rangle & \langle\Lambda^s q^2,p^3\rangle\\ \langle\Lambda^l q^3,p^1\rangle & \langle\Lambda^n q^3,p^2\rangle & \langle\Lambda^s q^3,p^3\rangle\end{vmatrix},\quad m\geqslant 2.\tag{4.12c}$$

In this way, we obtain the conserved integrals $\{F_m^{(i)}\}$, $0\leqslant i\leqslant 2$, of the Hamiltonian system (3.12). The Poisson bracket of two functions in the symplectic space $(\mathcal{R}^{6N},\Sigma_{i=1}^3 dp^i\wedge dq^i)$ is defined as

$$\{f,g\}=\sum_{j=1}^{N}\sum_{i=1}^{3}\left(\frac{\partial f}{\partial q_j^i}\frac{\partial g}{\partial p_j^i}-\frac{\partial f}{\partial p_j^i}\frac{\partial g}{\partial q_j^i}\right)=\sum_{i=1}^{3}\left(\left\langle\frac{\partial f}{\partial q^i},\frac{\partial g}{\partial p^i}\right\rangle-\left\langle\frac{\partial f}{\partial p^i},\frac{\partial g}{\partial q^i}\right\rangle\right).$$

We can prove the following assertions:

Theorem 4.1: The functions $\{F_m^{(i)}\}$, $0\leqslant i\leqslant 2$, $m\geqslant 0$, are in involution in pairs, $\{F_m^{(i)},F_l^{(j)}\}=0$, $0\leqslant i,\ j\leqslant 2$, for any $m,l\geqslant 0$.

Theorem 4.2: The $3N$ 1-forms $dF_l^{(i)}$, $1\leqslant l\leqslant N$, $0\leqslant i\leqslant 2$, are linearly independent. The proof of the above two theorems is given in the Appendix.

A direct calculation shows that the Hamiltonian function H of the system (3.12) can be rewritten as follows:

$$H=\gamma_0F_1^{(0)}+\gamma_1F_1^{(1)}+\gamma_2F_1^{(2)}+\gamma_3(\beta_1^2F_0^{(0)}-\beta_1F_0^{(1)}+F_0^{(2)})(\beta_2^2F_0^{(0)}\beta_2F_0^{(1)}+F_0^{(2)})$$
$$+\gamma_4(\beta_1^2F_0^{(0)}-\beta_1F_0^{(1)}+F_0^{(2)})(\beta_3^2F_0^{(0)}-\beta_3F_0^{(1)}$$
$$+F_0^{(2)})+\gamma_5(\beta_2^2F_0^{(0)}-\beta_2F_0^{(1)}+F_0^{(2)})(\beta_3^2F_0^{(0)}-\beta_3F_0^{(1)}+F_0^{(2)}), \tag{4.13}$$

where

$$\gamma_0=\frac{\alpha_1(\beta_2-\beta_3)\beta_1^2+\alpha_2(\beta_3-\beta_1)\beta_2^2+\alpha_3(\beta_1-\beta_2)\beta_3^2}{(\beta_1-\beta_2)(\beta_2-\beta_3)(\beta_1-\beta_3)},$$

$$\gamma_1=-\frac{\alpha_1(\beta_2-\beta_3)\beta_1+\alpha_2(\beta_3-\beta_1)\beta_2+\alpha_3(\beta_1-\beta_2)\beta_3}{(\beta_1-\beta_2)(\beta_2-\beta_3)(\beta_1-\beta_3)},$$

$$\gamma_2=\frac{\alpha_1(\beta_2-\beta_3)+\alpha_2(\beta_3-\beta_1)+\alpha_3(\beta_1-\beta_2)}{(\beta_1-\beta_2)(\beta_2-\beta_3)(\beta_1-\beta_3)},\quad \gamma_3=\frac{\alpha_2-\alpha_1}{(\beta_1-\beta_2)^3(\beta_1-\beta_3)(\beta_2-\beta_3)},$$

$$\gamma_4=\frac{\alpha_1-\alpha_3}{(\beta_1-\beta_3)^3(\beta_1-\beta_2)(\beta_2-\beta_3)},\quad \gamma_5=\frac{\alpha_3-\alpha_2}{(\beta_2-\beta_3)^3(\beta_1-\beta_2)(\beta_1-\beta_3)}.$$

Hence the integrability of Eq. (3.12) is established resorting to Theorems 4.1 and 4.2.

Theorem 4.3: The finite-dimensional Hamiltonian system (3.12) is completely integrable in the Liouville sense.

V. THE INVOLUTIVE REPRESENTATION OF SOLUTIONS

In this section, we shall give the involutive representation of solutions of the AKNS soliton hierarchy. We first introduce the Lenard gradients $g_l^{(i)}$ defined recursively by

$$Kg_{l-1}^{(i)}=Jg_l^{(i)},\quad g_l^{(i)}|_{u=0}=0,\quad 1\leqslant i\leqslant 3,\quad l\geqslant 1, \tag{5.1}$$

with

$$g_0^{(1)}=\left(\frac{u_{21}}{\alpha_1-\alpha_2},\frac{u_{12}}{\alpha_1-\alpha_2},\frac{u_{31}}{\alpha_1-\alpha_3},\frac{u_{13}}{\alpha_1-\alpha_3},0,0\right)^T, \tag{5.2}$$

$$g_0^{(2)}=\left(\frac{u_{21}}{\alpha_2-\alpha_1},\frac{u_{12}}{\alpha_2-\alpha_1},0,0,\frac{u_{32}}{\alpha_2-\alpha_3},\frac{u_{23}}{\alpha_2-\alpha_3}\right)^T, \tag{5.3}$$

$$g_0^{(3)}=\left(0,0,\frac{u_{31}}{\alpha_3-\alpha_1},\frac{u_{13}}{\alpha_3-\alpha_1},\frac{u_{32}}{\alpha_3-\alpha_2},\frac{u_{23}}{\alpha_3-\alpha_2}\right)^T. \tag{5.4}$$

The corresponding mth order vector is represented by

$$X_m(u,\omega)=J(\omega_1g_m^{(1)}+\omega_2g_m^{(2)}+\omega_3g_m^{(3)}),\quad \omega=(\omega_1,\omega_2,\omega_3), \tag{5.5}$$

from which it is easy to see that $X_m=X_m(u,\beta)$, $\beta=(\beta_1,\beta_2,\beta_3)$. Now we consider the canonical system of the $\overline{F}_m^{(1)}$-flow with $\overline{F}_m^{(1)}=-F_m^{(1)}$,

$$\frac{\partial q^i}{\partial t_m}=\frac{\partial\overline{F}_m^{(1)}}{\partial p^i},\quad \frac{\partial p^i}{\partial t_m}=-\frac{\partial\overline{F}_m^{(1)}}{\partial q^i},\quad 1\leqslant i\leqslant 3,\quad m\geqslant 1. \tag{5.6}$$

Then the systems (3.12) and (5.6) are compatible and their Hamiltonian phase flows $g_H^x, g_{F_m}^{t_m}$ commute,[26] which imply that there exists the involutive solution[26,9] of the consistent system of Eqs. (3.12) and (5.6), represented by

$$q^i(x,t_m)=q_H^x g_{F_m}^{t_m} q^i(0,0),\ \ p^i(x,t_m)=g_H^x g_{F_m}^{t_m} p^i(0,0),\ \ 1\leqslant i\leqslant 3.$$

Here $q^i(0,0), p^i(0,0),\ 1\leqslant i\leqslant 3$ are the given initial values.

Theorem 5.1: Let $\lambda_1,\ldots,\lambda_N$ be N distinct parameters. If $(q^i(x,t_1),p^i(x,t_1)),\ 1\leqslant i\leqslant 3$, is an involutive solution of the system of Eqs. (3.12) and (5.6) with $m=1$, then

$$u_{ij}(x,t_1)=\frac{\alpha_i-\alpha_j}{\beta_i-\beta_j}\langle q^i(x,t_1),p^j(x,t_1)\rangle,\ \ 1\leqslant i,\ \ j\leqslant 3,\ \ i\neq j, \tag{5.7}$$

solve the three-wave interaction equations

$$u_{ijt_1}=\frac{\beta_i-\beta_j}{\alpha_i-\alpha_j}u_{ijx}+\sum_{\substack{k=1\\k\neq i,j}}^{3}\left(\frac{\beta_k-\beta_i}{\alpha_k-\alpha_i}-\frac{\beta_k-\beta_j}{\alpha_k-\alpha_j}\right)u_{ik}u_{kj}+a_{ij}u_{ij},\ \ i\neq j,\ \ 1\leqslant i,\ \ j\leqslant 3, \tag{5.8}$$

where $a_{ij}=\langle q^i(0,0),p^i(0,0)\rangle-\langle q^j(0,0),p^j(0,0)\rangle,\ i\neq j,\ 1\leqslant i,\ j\leqslant 3$, are constants independent of x,t.

Proof: Using Eqs. (4.10), (4.11a), and (4.12a), we have

$$\langle q^1,p^1\rangle=\frac{F_0^{(2)}-\beta_1F_0^{(1)}+\beta_1^2F_0^{(0)}}{(\beta_1-\beta_2)(\beta_1-\beta_3)},\quad \langle q^2,p^2\rangle=\frac{-F_0^{(2)}+\beta_2F_0^{(1)}-\beta_2^2F_0^{(0)}}{(\beta_1-\beta_2)(\beta_2-\beta_3)},$$

$$\langle q^3,p^3\rangle=\frac{F_0^{(2)}-\beta_3F_0^{(1)}+\beta_3^2F_0^{(0)}}{(\beta_1-\beta_3)(\beta_2-\beta_3)},$$

which imply that $\langle q^i,p^i\rangle=\langle q^i(0,0),p^i(0,0)\rangle,\ 1\leqslant i\leqslant 3$, are constants. Noticing Eqs. (3.12) and (5.6) with $m=1$, through direct calculations we get

$$\langle q^i,p^j\rangle_x=(\alpha_i-\alpha_j)\langle\Lambda q^i,p^j\rangle-\frac{\alpha_i-\alpha_j}{\beta_i-\beta_j}a_{ij}\langle q^i,p^j\rangle+\sum_{\substack{k=1\\k\neq i,j}}^{3}\left(\frac{\alpha_k-\alpha_i}{\beta_k-\beta_i}-\frac{\alpha_k-\alpha_j}{\beta_k-\beta_j}\right)\langle q^i,p^k\rangle\langle q^k,p^j\rangle,$$

$$\langle q^i,p^j\rangle_{t_1}=(\beta_i-\beta_j)\langle\Lambda q^i,p^j\rangle,\ \ i\neq j,1\leqslant i,j\leqslant 3.$$

Combining these results together yield the desired three-wave interaction equations (5.8).

Generally, we have the following fact:

Theorem 5.2: Let $\lambda_1,\ldots,\lambda_N$ be N distinct parameters. If $(q^i(x,t_m),p^i(x,t_m)),\ 1\leqslant i\leqslant 3$, is an involutive solution of the system of Eqs. (3.12) and (5.6), then

$$u_{ij}(x,t_m)=\frac{\alpha_i-\alpha_j}{\beta_i-\beta_j}\langle q^i(x,t_m),p^j(x,t_m)\rangle,\ \ 1\leqslant i,j\leqslant 3,i\neq j, \tag{5.9}$$

satisfy the soliton equations

$$u_{t_m}=X_m(u,\beta)+X_{m-1}(u,\sigma^{(0)})+\cdots+X_0(u,\sigma^{(m-1)}),\ \ m\geqslant 0, \tag{5.10}$$

with suitably chosen constant vectors $\sigma^{(l)}=(\sigma_1^{(l)},\sigma_2^{(l)},\sigma_3^{(l)}),\ 0\leqslant l\leqslant m-1$.

Proof: By using Eqs. (5.9) and (5.6), a direct calculation gives

$$u_{t_m}=J\sum_{l=1}^{N}\lambda_l^m\nabla\lambda_l. \tag{5.11}$$

Operating with the operator $(J^{-1}K)^m$ upon the constraint $G_0=\Sigma_{l=1}^{N}\nabla\lambda_l$, we have

$$G_m+\sum_{l=0}^{m-1}(\sigma_1^{(l)}g_{m-l-1}^{(1)}+\sigma_2^{(l)}g_{m-l-1}^{(2)}+\sigma_3^{(l)}g_{m-l-1}^{(3)})=\sum_{l=1}^{N}\lambda_l^m\nabla\lambda_l, \tag{5.12}$$

where $\sigma_i^{(l)}$, $0\leqslant l\leqslant m-1$, $1\leqslant i\leqslant 3$, are integral constants. Substituting Eq. (5.12) into Eq. (5.11) and noticing Eq. (5.5), we obtain Eq. (5.10). The proof is finished.

VI. SUMMARY AND CONCLUSIONS

The procedure for the nonlinearization of the $n\times n$ matrix spectral problem and its adjoint spectral problem has been described briefly. To illustrate the general principles, the nonlinearization of the 3×3 AKNS matrix spectral problem and its adjoint spectral problem associated with the three-wave interaction equations is discussed in detail. The characteristic polynomial of solution matrix of the stationary zero-curvature equation is used to generate involutive system of enough conserved integrals of the resulting finite-dimensional Hamiltonian system. This scheme is general, which is suitable for the other systems. It is interesting that the canonical equations of the $F_m^{(1)}$-flow (up to a constant factor) given by $\mathcal{F}_\lambda^{(1)}$ of Eq. (4.1) are exactly the nonlinearized temporal parts of the Lax pairs and adjoint Lax pairs for soliton hierarchy related to the 3×3 AKNS matrix spectral problem. The solutions of the three-wave interaction equations are reduced to solving the two compatible systems of ordinary differential equations. Two generators of involutive systems of conserved integrals are introduced, from which the functional independence of conserved integrals is rigorously proved (see Appendix). We point out that the method used here is general, which is suitable for the cases of $n\times n$ matrix spectral problems. Similar results will be left to a future publication. Moreover, we may also consider construction of action-angle variables for the finite dimensional integrable system and further give the finite-band solutions for the three-wave interaction equations, which will be discussed in other papers.

APPENDIX:

1. The proof of theorem 4.1

In order to prove theorem 4.1, We first introduce the notations

$$I_\lambda^{(1)}=(\beta_2+\beta_3)\overset{11}{Q}_\lambda+(\beta_1+\beta_3)\overset{22}{Q}_\lambda+(\beta_1+\beta_2)\overset{33}{Q}_\lambda, \tag{A1}$$

$$T_\lambda^{(1)}=\sum_{1\leqslant i<j\leqslant 3}(\overset{ii}{Q}_\lambda\overset{jj}{Q}_\lambda-\overset{ij}{Q}_\lambda\overset{ji}{Q}_\lambda), \tag{A2}$$

$$I_\lambda^{(2)}=\beta_2\beta_3\overset{11}{Q}_\lambda+\beta_1\beta_3\overset{22}{Q}_\lambda+\beta_1\beta_2\overset{33}{Q}_\lambda, \tag{A3}$$

$$T_\lambda^{(2)}=\sum_{1\leqslant i<j\leqslant 3}\beta_{6-i-j}(\overset{ii}{Q}_\lambda\overset{jj}{Q}_\lambda-\overset{ij}{Q}_\lambda\overset{ji}{Q}_\lambda), \tag{A4}$$

$$R_\lambda = \begin{vmatrix} \overset{11}{Q}_\lambda & \overset{12}{Q}_\lambda & \overset{13}{Q}_\lambda \\ \overset{21}{Q}_\lambda & \overset{22}{Q}_\lambda & \overset{23}{Q}_\lambda \\ \overset{31}{Q}_\lambda & \overset{32}{Q}_\lambda & \overset{33}{Q}_\lambda \end{vmatrix}, \tag{A5}$$

and prove the following several assertions:

Lemma A.1: Let

$$I_\lambda(\sigma) = \sigma_1 \overset{11}{Q}_\lambda + \sigma_2 \overset{22}{Q}_\lambda + \sigma_3 \overset{33}{Q}_\lambda, \quad \sigma = (\sigma_1, \sigma_2, \sigma_3) \in \mathcal{C}^3. \tag{A6}$$

Then we have

$$\{I_\mu(\sigma), I_\lambda(\tau)\} = 0, \quad \tau = (\tau_1, \tau_2, \tau_3) \in \mathcal{C}^3, \tag{A7}$$

$$\begin{aligned}(\mu-\lambda)\{I_\mu(\sigma), R_\lambda\} = {} & (\sigma_1-\sigma_2)[\overset{12}{Q}_\mu(\overset{23}{Q}_\lambda \overset{31}{Q}_\lambda - \overset{21}{Q}_\lambda \overset{33}{Q}_\lambda) - \overset{21}{Q}_\mu(\overset{13}{Q}_\lambda \overset{32}{Q}_\lambda - \overset{12}{Q}_\lambda \overset{33}{Q}_\lambda)] + (\sigma_1-\sigma_3) \\ & \times[\overset{13}{Q}_\mu(\overset{21}{Q}_\lambda \overset{32}{Q}_\lambda - \overset{22}{Q}_\lambda \overset{31}{Q}_\lambda) - \overset{31}{Q}_\mu(\overset{12}{Q}_\lambda \overset{23}{Q}_\lambda - \overset{13}{Q}_\lambda \overset{22}{Q}_\lambda)] + (\sigma_2-\sigma_3) \\ & \times[\overset{23}{Q}_\mu(\overset{12}{Q}_\lambda \overset{31}{Q}_\lambda - \overset{11}{Q}_\lambda \overset{32}{Q}_\lambda) - \overset{32}{Q}_\mu(\overset{13}{Q}_\lambda \overset{21}{Q}_\lambda - \overset{11}{Q}_\lambda \overset{23}{Q}_\lambda)], \quad \forall \lambda, \mu \in \mathcal{C}. \end{aligned} \tag{A8}$$

Proof: By using the definition of the Poisson bracket, we have

$$\{\overset{ii}{Q}_\mu, \overset{ii}{Q}_\lambda\} = \left\langle \frac{\partial \overset{ii}{Q}_\mu}{\partial q^i}, \frac{\partial \overset{ii}{Q}_\lambda}{\partial p^i} \right\rangle - \left\langle \frac{\partial \overset{ii}{Q}_\mu}{\partial p^i}, \frac{\partial \overset{ii}{Q}_\lambda}{\partial q^i} \right\rangle \quad \{\overset{ii}{Q}_\mu, \overset{jj}{Q}_\lambda\} = 0, \quad (i \neq j),$$

which together with the equalities

$$\frac{\partial \overset{ij}{Q}_\lambda}{\partial q^k} = \delta_{ik}(\lambda - \Lambda)^{-1} p^j, \quad \frac{\partial \overset{ij}{Q}_\lambda}{\partial p^k} = \delta_{jk}(\lambda - \Lambda)^{-1} q^i, \quad 1 \leqslant i,j,k \leqslant 3 \tag{A9}$$

leads to

$$\{\overset{ii}{Q}_\mu, \overset{jj}{Q}_\lambda\} = 0, \quad 1 \leqslant i,j \leqslant 3. \tag{A10}$$

Resorting to Eq. (A10) and the bilinear property of the Poisson bracket, a direct calculation shows that Eq. (A7) holds. Let

$$f_1(\mu,\lambda) = (\mu-\lambda) \sum_{k=1}^{3} \sigma_k \left(\left\langle \frac{\partial \overset{kk}{Q}_\mu}{\partial q^k}, \frac{\partial R_\lambda}{\partial p^k} \right\rangle - \left\langle \frac{\partial \overset{kk}{Q}_\mu}{\partial p^k}, \frac{\partial R_\lambda}{\partial q^k} \right\rangle \right). \tag{A11}$$

It is easy to see that

$$(\mu-\lambda)\{I_\mu(\sigma), R_\lambda\} = f_1(\mu,\lambda). \tag{A12}$$

In the following calculations of the *Mathematica*, we usually write $\overset{ij}{Q}_\lambda = Q_{\lambda,ij}$ for the sake of convenience. By using the *Mathematica*, we can verify Eq. (A8),

$$a_{11}[\lambda\_] := Q_{\lambda,22}Q_{\lambda,33} - Q_{\lambda,23}Q_{\lambda,32}; \quad a_{12}[\lambda\_] := Q_{\lambda,23}Q_{\lambda,31} - Q_{\lambda,21}Q_{\lambda,33};$$

$$a_{13}[\lambda\_] := Q_{\lambda,21}Q_{\lambda,32} - Q_{\lambda,22}Q_{\lambda,31}; \quad a_{21}[\lambda\_] := Q_{\lambda,13}Q_{\lambda,32} - Q_{\lambda,12}Q_{\lambda,33};$$

$$a_{22}[\lambda\_] := Q_{\lambda,11}Q_{\lambda,33} - Q_{\lambda,13}Q_{\lambda,31}; \quad a_{23}[\lambda\_] := Q_{\lambda,12}Q_{\lambda,31} - Q_{\lambda,11}Q_{\lambda,32};$$

$$a_{31}[\lambda\_] := Q_{\lambda,12}Q_{\lambda,23} - Q_{\lambda,13}Q_{\lambda,22}; \quad a_{32}[\lambda\_] := Q_{\lambda,13}Q_{\lambda,21} - Q_{\lambda,11}Q_{\lambda,23};$$

$$a_{33}[\lambda\_] := Q_{\lambda,11}Q_{\lambda,22} - Q_{\lambda,12}Q_{\lambda,21};$$

$$\begin{aligned} f_1[\mu\_,\lambda\_] :=& \sigma_1(a_{21}[\lambda](Q_{\lambda,21}-Q_{\mu,21}) + a_{31}[\lambda](Q_{\lambda,31}-Q_{\mu,31}) - a_{12}[\lambda](Q_{\lambda,12}-Q_{\mu,12}) - a_{13}[\lambda] \\ &\times(Q_{\lambda,13}-Q_{\mu,12})) + \sigma_2(a_{12}[\lambda](Q_{\lambda,12}-Q_{\mu,12}) + a_{32}[\lambda](Q_{\lambda,32}-Q_{\mu,32}) - a_{21}[\lambda] \\ &\times(Q_{\lambda,21}-Q_{\mu,21}) - a_{23}[\lambda](Q_{\lambda,23}-Q_{\mu,23})) + \sigma_3(a_{13}[\lambda](Q_{\lambda,13}-Q_{\mu,13}) + a_{23}[\lambda] \\ &\times(Q_{\lambda,23}-Q_{\mu,23}) - a_{31}[\lambda]Q_{\lambda,31}-Q_{\mu,31}) - a_{32}[\lambda](Q_{\lambda,32}-Q_{\mu,32})); \end{aligned}$$

$$\begin{aligned} g_1 =& (\sigma_1-\sigma_2)(a_{12}[\lambda]Q_{\mu,12} - a_{21}[\lambda]Q_{\mu,21}) + (\sigma_1-\sigma_3)(a_{13}[\lambda]Q_{\mu,13} - a_{31}[\lambda]Q_{\mu,31}) \\ &+ (\sigma_2-\sigma_3)(a_{23}[\lambda]Q_{\mu,23} - a_{32}[\lambda]Q_{\mu,32}); \end{aligned}$$

$$\mathrm{Simplify}[f_1[\mu,\lambda] - g_1]$$

$$\mathrm{Out}[1] = 0.$$

Here the expression of $f_1[\mu\_,\lambda\_]$ can be obtained by substituting Eq. (A5) into Eq. (A11) and using Eq. (A9) and the equality

$$\langle(\mu-\Lambda)^{-1}(\lambda-\Lambda)^{-1}q^i,p^j\rangle = (\mu-\lambda)^{-1}(\overset{ij}{Q}_\lambda - \overset{ij}{Q}_\mu). \tag{A13}$$

Lemma A.2: Let

$$T_\lambda(\sigma) = \sum_{1\le i<j\le 3} \sigma_{6-i-j}(\overset{ii}{Q}_\lambda \overset{jj}{Q}_\lambda - \overset{ij}{Q}_\lambda \overset{ji}{Q}_\lambda), \quad \sigma = (\sigma_1,\sigma_2,\sigma_3) \in \mathcal{C}^3. \tag{A14}$$

Then we have

$$\begin{aligned} (\mu-\lambda)\{I_\mu(\sigma),T_\lambda(\tau)\} =& \tau_3(\sigma_1-\sigma_2)(\overset{21}{Q}_\mu \overset{12}{Q}_\lambda - \overset{12}{Q}_\mu \overset{21}{Q}_\lambda) + \tau_2(\sigma_1-\sigma_3) \\ &\times(\overset{31}{Q}_\mu \overset{13}{Q}_\lambda - \overset{13}{Q}_\mu \overset{31}{Q}_\lambda) + \tau_1(\sigma_2-\sigma_3)(\overset{32}{Q}_\mu \overset{23}{Q}_\lambda - \overset{23}{Q}_\mu \overset{32}{Q}_\lambda), \end{aligned} \tag{A15}$$

$$\begin{aligned} (\mu-\lambda)\{T_\mu(\tau),T_\lambda(\sigma)\} =& \tau_1\sigma_2W_1(\mu,\lambda) + \tau_2\sigma_1W_1(\lambda,\mu) + \tau_2\sigma_3W_2(\mu,\lambda) \\ &+ \tau_3\sigma_2W_2(\lambda,\mu) + \tau_1\sigma_3W_3(\mu,\lambda) + \tau_3\sigma_1W_3(\lambda,\mu), \\ &\forall\lambda,\mu\in\mathcal{C}, \tau=(\tau_1,\tau_2,\tau_3)\in\mathcal{C}^3, \end{aligned} \tag{A16}$$

where

$$\begin{aligned} W_1(\mu,\lambda) = {} & Q^{31}_\lambda(Q^{13}_\mu Q^{22}_\mu - Q^{12}_\mu Q^{23}_\mu) + Q^{13}_\lambda(Q^{32}_\mu Q^{21}_\mu - Q^{22}_\mu Q^{31}_\mu) \\ & + Q^{32}_\mu(Q^{11}_\lambda Q^{23}_\lambda - Q^{13}_\lambda Q^{21}_\lambda) + Q^{23}_\mu(Q^{12}_\lambda Q^{31}_\lambda - Q^{11}_\lambda Q^{32}_\lambda), \end{aligned}$$

$$\begin{aligned} W_2(\mu,\lambda) = {} & Q^{12}_\lambda(Q^{21}_\mu Q^{33}_\mu - Q^{23}_\mu Q^{31}_\mu) + Q^{21}_\lambda(Q^{13}_\mu Q^{32}_\mu - Q^{12}_\mu Q^{33}_\mu) \\ & + Q^{13}_\mu(Q^{22}_\lambda Q^{31}_\lambda - Q^{21}_\lambda Q^{32}_\lambda) + Q^{31}_\mu(Q^{12}_\lambda Q^{23}_\lambda - Q^{13}_\lambda Q^{22}_\lambda), \end{aligned}$$

$$\begin{aligned} W_3(\mu,\lambda) = {} & Q^{21}_\lambda(Q^{12}_\mu Q^{33}_\mu - Q^{13}_\mu Q^{32}_\mu) + Q^{12}_\lambda(Q^{23}_\mu Q^{31}_\mu - Q^{21}_\mu Q^{33}_\mu) \\ & + Q^{23}_\mu(Q^{11}_\lambda Q^{32}_\lambda - Q^{12}_\lambda Q^{31}_\lambda) + Q^{32}_\mu(Q^{13}_\lambda Q^{21}_\lambda - Q^{11}_\lambda Q^{23}_\lambda). \end{aligned}$$

Proof: In a way similar to the proof of Eq. (A8), we can verify Eq. (A15). Now we prove Eq. (A16). Let

$$f_2(\mu,\lambda) = (\mu-\lambda)\sum_{k=1}^{3}\left(\left\langle \frac{\partial T_\mu(\tau)}{\partial q^k}, \frac{\partial T_\lambda(\sigma)}{\partial p^k}\right\rangle - \left\langle \frac{\partial T_\mu(\tau)}{\partial p^k}, \frac{\partial T_\lambda(\sigma)}{\partial q^k}\right\rangle\right). \tag{A17}$$

Then we have

$$(\mu-\lambda)\{T_\mu(\tau), T_\mu(\sigma)\} = f_2(\mu,\lambda). \tag{A18}$$

In the following, we shall prove by using the *Mathematica* that Eq. (A16) holds

$$b_1[\lambda\_] = \sigma_3 Q_{\lambda,22} + \sigma_2 Q_{\lambda,33}; \quad b_2[\lambda\_] = \sigma_3 Q_{\lambda,11} + \sigma_1 Q_{\lambda,33};$$

$$b_3[\lambda\_] = \sigma_2 Q_{\lambda,11} + \sigma_1 Q_{\lambda,22};$$

$$c_1[\mu\_] = \tau_3 Q_{\mu,22} + \tau_2 Q_{\mu,33}; \quad c_2[\mu\_] = \tau_3 Q_{\mu,11} + \tau_1 Q_{\mu,33};$$

$$c_3[\mu\_] = \tau_2 Q_{\mu,11} + \tau_1 Q_{\mu,22};$$

$$\begin{aligned} f_2[\mu\_,\lambda\_] := {} & c_1[\mu](b_1[\lambda](Q_{\lambda,11} - Q_{\mu,11}) - \sigma_3 Q_{\lambda,12}(Q_{\lambda,21} - Q_{\mu,21}) - \sigma_2 Q_{\lambda,13}(Q_{\lambda,31} - Q_{\mu,31})) \\ & - \tau_3 Q_{\mu,21}(b_1[\lambda](Q_{\lambda,12} - Q_{\mu,12}) - \sigma_3 Q_{\lambda,12}(Q_{\lambda,22} - Q_{\mu,22}) - \sigma_2 Q_{\lambda,13}(Q_{\lambda,32} - Q_{\mu,32})) \\ & - \tau_2 Q_{\mu,31}(b_1[\lambda](Q_{\lambda,13} - Q_{\mu,13}) - \sigma_3 Q_{\lambda,12}(Q_{\lambda,23} - Q_{\mu,23}) - \sigma_2 Q_{\lambda,13}(Q_{\lambda,33} - Q_{\mu,33})) \\ & + c_2[\mu](b_2[\lambda](Q_{\lambda,22} - Q_{\mu,22}) - \sigma_3 Q_{\lambda,21}(Q_{\lambda,12} - Q_{\mu,12}) - \sigma_1 Q_{\lambda,23}(Q_{\lambda,32} - Q_{\mu,32})) \\ & - \tau_3 Q_{\mu,12}(b_2[\lambda](Q_{\lambda,21} - Q_{\mu,21}) - \sigma_3 Q_{\lambda,21}(Q_{\lambda,11} - Q_{\mu,11}) - \sigma_1 Q_{\lambda,23}(Q_{\lambda,31} - Q_{\mu,31})) \\ & - \tau_1 Q_{\mu,32}(b_2[\lambda](Q_{\lambda,23} - Q_{\mu,23}) - \sigma_3 Q_{\lambda,21}(Q_{\lambda,13} - Q_{\mu,13}) - \sigma_1 Q_{\lambda,23}(Q_{\lambda,33} - Q_{\mu,33})) \\ & + c_3[\mu](b_3[\lambda](Q_{\lambda,33} - Q_{\mu,33}) - \sigma_2 Q_{\lambda,31}(Q_{\lambda,13} - Q_{\mu,13}) - \sigma_1 Q_{\lambda,32}(Q_{\lambda,23} - Q_{\mu,23})) \\ & - \tau_2 Q_{\mu,13}(b_3[\lambda](Q_{\lambda,31} - Q_{\mu,31}) - \sigma_2 Q_{\lambda,31}(Q_{\lambda,11} - Q_{\mu,11}) - \sigma_1 Q_{\lambda,32}(Q_{\lambda,21} - Q_{\mu,21})) \\ & - \tau_1 Q_{\mu,23}(b_3[\lambda](Q_{\lambda,32} - Q_{\mu,32}) - \sigma_2 Q_{\lambda,31}(Q_{\lambda,12} - Q_{\mu,12}) - \sigma_1 Q_{\lambda,32}(Q_{\lambda,22} - Q_{\mu,22})) \\ & + b_1[\lambda](c_1[\mu](Q_{\mu,11} - Q_{\lambda,11}) - \tau_3 Q_{\mu,12}(Q_{\mu,21} - Q_{\lambda,21}) - \tau_2 Q_{\mu,13}(Q_{\mu,31} - Q_{\lambda,31})) \\ & - \sigma_3 Q_{\lambda,21}(c_1[\mu](Q_{\mu,12} - Q_{\lambda,12}) - \tau_3 Q_{\mu,12}(Q_{\mu,22} - Q_{\lambda,22}) - \tau_2 Q_{\mu,13}(Q_{\mu,32} - Q_{\lambda,32})) \end{aligned}$$

$$-\sigma_2 Q_{\lambda,31}(c_1[\mu](Q_{\mu,13}-Q_{\lambda,13})-\tau_3 Q_{\mu,12}(Q_{\mu,23}-Q_{\lambda,23})-\tau_2 Q_{\mu,13}(Q_{\mu,33}-Q_{\lambda,33}))$$
$$+b_2[\lambda](c_2[\mu](Q_{\mu,22}-Q_{\lambda,22})-\tau_3 Q_{\mu,21}(Q_{\mu,12}-Q_{\lambda,12})-\tau_1 Q_{\mu,23}(Q_{\mu,32}-Q_{\lambda,32}))$$
$$-\sigma_3 Q_{\lambda,12}(c_2[\mu](Q_{\mu,21}-Q_{\lambda,21})-\tau_3 Q_{\mu,21}(Q_{\mu,11}-Q_{\lambda,11})-\tau_1 Q_{\mu,23}(Q_{\mu,31}-Q_{\lambda,31}))$$
$$-\sigma_1 Q_{\lambda,32}(c_2[\mu](Q_{\mu,23}-Q_{\lambda,23})-\tau_3 Q_{\mu,21}(Q_{\mu,13}-Q_{\lambda,13})-\tau_1 Q_{\mu,23}(Q_{\mu,33}-Q_{\lambda,33}))$$
$$+b_3[\lambda](c_3[\mu](Q_{\mu,33}-Q_{\lambda,33})-\tau_2 Q_{\mu,31}(Q_{\mu,13}-Q_{\lambda,13})-\tau_1 Q_{\mu,32}(Q_{\mu,23}-Q_{\lambda,23}))$$
$$-\sigma_2 Q_{\lambda,13}(c_3[\mu](Q_{\mu,31}-Q_{\lambda,31})-\tau_2 Q_{\mu,31}(Q_{\mu,11}-Q_{\lambda,11})-\tau_1 Q_{\mu,32}(Q_{\mu,21}-Q_{\lambda,21}))$$
$$-\sigma_1 Q_{\lambda,23}(c_3[\mu](Q_{\mu,32}-Q_{\lambda,32})-\tau_2 Q_{\mu,31}(Q_{\mu,12}-Q_{\lambda,12})-\tau_1 Q_{\mu,32}(Q_{\mu,22}$$
$$-Q_{\lambda,22}));$$

$$W_1[\mu\_,\lambda\_]:=a_{13}[\mu]Q_{\lambda,13}-a_{31}[\mu]Q_{\lambda,31}-a_{32}[\lambda]Q_{\mu,32}+a_{23}[\lambda]Q_{\mu,23};$$

$$W_2[\mu\_,\lambda\_]:=a_{21}[\mu]Q_{\lambda,21}-a_{12}[\mu]Q_{\lambda,12}-a_{13}[\lambda]Q_{\mu,13}+a_{31}[\lambda]Q_{\mu,31};$$

$$W_3[\mu\_,\lambda\_]:=a_{12}[\mu]Q_{\lambda,12}-a_{21}[\mu]Q_{\lambda,21}-a_{23}[\lambda]Q_{\mu,23}+a_{32}[\lambda]Q_{\mu,32};$$

$$g_2=\tau_1\sigma_2 W_1[\mu,\lambda]+\tau_2\sigma_1 W_1[\lambda,\mu]+\tau_2\sigma_3 W_2[\mu,\lambda]+\tau_3\sigma_2 W_2[\lambda,\mu]$$
$$+\tau_1\sigma_3 W_3[\mu,\lambda]+\tau_3\sigma_1 W_3[\lambda,\mu];$$

$$\text{Simplify}[f_2[\mu,\lambda]-g_2]$$

$$\text{Out}[2]=0.$$

Lemma A.3: Under the same assumption as the Lemma A.2, we have

$$(\mu-\lambda)\{T_\mu(\sigma),R_\lambda\}=\sigma_1[Y_1(\mu,\lambda)-Y_2(\mu,\lambda)]+\sigma_2[Y_3(\mu,\lambda)-Y_1(\mu,\lambda)]$$
$$+\sigma_3[Y_2(\mu,\lambda)-Y_3(\mu,\lambda)],$$
$$\forall\lambda,\mu\in\mathcal{C}. \tag{A19}$$

where

$$Y_1(\mu,\lambda)=(\overset{21}{Q}_\mu\overset{33}{Q}_\mu-\overset{23}{Q}_\mu\overset{31}{Q}_\mu)(\overset{13}{Q}_\lambda\overset{32}{Q}_\lambda-\overset{12}{Q}_\lambda\overset{33}{Q}_\lambda)+(\overset{13}{Q}_\mu\overset{32}{Q}_\mu-\overset{12}{Q}_\mu\overset{33}{Q}_\mu)(\overset{23}{Q}_\lambda\overset{31}{Q}_\lambda-\overset{21}{Q}_\lambda\overset{33}{Q}_\lambda),$$

$$Y_2(\mu,\lambda)=(\overset{22}{Q}_\mu\overset{13}{Q}_\mu-\overset{12}{Q}_\mu\overset{23}{Q}_\mu)(\overset{21}{Q}_\lambda\overset{32}{Q}_\lambda-\overset{22}{Q}_\lambda\overset{31}{Q}_\lambda)+(\overset{21}{Q}_\mu\overset{32}{Q}_\mu-\overset{22}{Q}_\mu\overset{31}{Q}_\mu)(\overset{12}{Q}_\lambda\overset{23}{Q}_\lambda-\overset{13}{Q}_\lambda\overset{22}{Q}_\lambda),$$

$$Y_3(\mu,\lambda)=(\overset{11}{Q}_\mu\overset{32}{Q}_\mu-\overset{12}{Q}_\mu\overset{31}{Q}_\mu)(\overset{13}{Q}_\lambda\overset{21}{Q}_\lambda-\overset{11}{Q}_\lambda\overset{23}{Q}_\lambda)+(\overset{13}{Q}_\mu\overset{21}{Q}_\mu-\overset{11}{Q}_\mu\overset{23}{Q}_\mu)(\overset{12}{Q}_\lambda\overset{31}{Q}_\lambda-\overset{11}{Q}_\lambda\overset{32}{Q}_\lambda).$$

Proof: Let

$$f_3(\mu,\lambda)=(\mu-\lambda)\sum_{k=1}^{3}\left(\left\langle\frac{\partial T(\sigma)_\mu}{\partial q^k},\frac{\partial R_\lambda}{\partial p^k}\right\rangle-\left\langle\frac{\partial T(\sigma)_\mu}{\partial p^k},\frac{\partial R_\lambda}{\partial q^k}\right\rangle\right). \tag{A20}$$

Then we have

$$(\mu-\lambda)\{T(\sigma)_\mu,R_\lambda\}=f_3(\mu,\lambda). \tag{A21}$$

With the help of the *Mathematica*, we can verify Eq. (A19)

$$
\begin{aligned}
f_3[\mu\_,\lambda\_] := & b_1[\mu](a_{11}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{21}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{31}[\lambda](Q_{\lambda,31}-Q_{\mu,31}))\\
&-\sigma_3 Q_{\mu,21}(a_{11}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{21}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{31}[\lambda](Q_{\lambda,32}-Q_{\mu,32}))\\
&-\sigma_2 Q_{\mu,31}(a_{11}[\lambda](Q_{\lambda,13}-Q_{\mu,13})+a_{21}[\lambda](Q_{\lambda,23}-Q_{\mu,23})+a_{31}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&-b_1[\mu](a_{11}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{12}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{13}[\lambda](Q_{\lambda,13}-Q_{\mu,13}))\\
&+\sigma_3 Q_{\mu,12}(a_{11}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{12}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{13}[\lambda](Q_{\lambda,23}-Q_{\mu,23}))\\
&+\sigma_2 Q_{\mu,13}(a_{11}[\lambda](Q_{\lambda,31}-Q_{\mu,31})+a_{12}[\lambda](Q_{\lambda,32}-Q_{\mu,32})+a_{13}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&+b_2[\mu](a_{12}[\lambda](Q_{\lambda,2}-Q_{\mu,12})+a_{22}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{32}[\lambda](Q_{\lambda,32}-Q_{\mu,32}))\\
&-\sigma_3 Q_{\mu,12}(a_{12}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{22}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{32}[\lambda](Q_{\lambda,31}-Q_{\mu,31}))\\
&-\sigma_1 Q_{\mu,32}(a_{12}[\lambda](Q_{\lambda,13}-Q_{\mu,13})+a_{22}[\lambda](Q_{\lambda,23}-Q_{\mu,23})+a_{32}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&-b_2[\mu](a_{21}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{22}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{23}[\lambda](Q_{\lambda,23}-Q_{\mu,23}))\\
&+\sigma_3 Q_{\mu,21}(a_{21}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{22}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{23}[\lambda](Q_{\lambda,13}-Q_{\mu,13}))\\
&+\sigma_1 Q_{\mu,23}(a_{21}[\lambda](Q_{\lambda,31}-Q_{\mu,31})+a_{22}[\lambda](Q_{\lambda,32}-Q_{\mu,32})+a_{23}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&+b_3[\mu](a_{13}[\lambda](Q_{\lambda,13}-Q_{\mu,13})+a_{23}[\lambda](Q_{\lambda,23}-Q_{\mu,23})+a_{33}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&-\sigma_2 Q_{\mu,13}(a_{13}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{23}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{33}[\lambda](Q_{\lambda,31}-Q_{\mu,31}))\\
&-\sigma_1 Q_{\mu,23}(a_{13}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{23}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{33}[\lambda](Q_{\lambda,32}-Q_{\mu,32}))\\
&-b_3[\mu](a_{31}[\lambda](Q_{\lambda,31}-Q_{\mu,31})+a_{32}[\lambda](Q_{\lambda,32}-Q_{\mu,32})+a_{33}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&+\sigma_2 Q_{\mu,31}(a_{31}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{32}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{33}[\lambda](Q_{\lambda,13}-Q_{\mu,13}))\\
&+\sigma_1 Q_{\mu,32}(a_{31}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{32}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{33}[\lambda](Q_{\lambda,23}-Q_{\mu,23}));
\end{aligned}
$$

$$Y_1=a_{21}[\mu]a_{12}[\lambda]-a_{12}[\mu]a_{21}[\lambda];\quad Y_2=a_{13}[\mu]a_{31}[\lambda]-a_{31}[\mu]a_{13}[\lambda];$$

$$Y_3=a_{32}[\mu]a_{23}[\lambda]-a_{23}[\mu]a_{32}[\lambda];\quad g_3=\sigma_1(Y_1-Y_2)+\sigma_2(Y_3-Y_1)+\sigma_3(Y_2-Y_3);$$

$$\text{Simplify}[f_3[\mu,\lambda]-g_3]$$

$$\text{Out}[3]=0.$$

Lemma A.4:

$$\{T_\mu^{(1)},T_\lambda^{(1)}\}=0,\quad \{R_\mu,R_\lambda\}=0,\quad \{T_\mu^{(1)},R_\lambda\}=0,\quad \forall\lambda,\mu\in\mathcal{C}. \tag{A22}$$

Proof: By Eqs. (A16) and (A19), we get that

$$\{T_\mu^{(1)},T_\lambda^{(1)}\}=\{T_\mu(\tau),T_\lambda(\sigma)\}|_{\tau=\sigma=(1,1,1)}=0,\quad \{T_\mu^{(1)},R_\lambda\}=\{T_\mu(\sigma),R_\lambda\}|_{\sigma=(1,1,1)}=0.$$

Let

$$f_4(\mu,\lambda)=(\mu-\lambda)\sum_{k=1}^{3}\left(\left\langle\frac{\partial R_\mu}{\partial q^k},\frac{\partial R_\lambda}{\partial p^k}\right\rangle\right). \tag{A23}$$

It is easy to see that

$$(\mu-\lambda)\{R_\mu,R_\lambda\}=f_4(\mu,\lambda)+f_4(\lambda,\mu). \tag{A24}$$

The second expression of Eq. (A22) can be verified by the *Mathematica*,

$$
\begin{aligned}
f_4[\mu\_,\lambda\_]:=a_{11}[\mu](&a_{11}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{11}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{31}[\lambda](Q_{\lambda,31}-Q_{\mu,31}))\\
&+a_{12}[\mu](a_{11}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{21}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{31}[\lambda](Q_{\lambda,32}-Q_{\mu,32}))\\
&+a_{13}[\mu](a_{11}[\lambda](Q_{\lambda,13}-Q_{\mu,13})+a_{21}[\lambda](Q_{\lambda,23}-Q_{\mu,23})+a_{31}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&+a_{21}[\mu](a_{12}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{22}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{32}[\lambda](Q_{\lambda,31}-Q_{\mu,31}))\\
&+a_{22}[\mu](a_{12}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{22}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{32}[\lambda](Q_{\lambda,32}-Q_{\mu,32}))\\
&+a_{23}[\mu](a_{12}[\lambda](Q_{\lambda,13}-Q_{\mu,13})+a_{22}[\lambda](Q_{\lambda,23}-Q_{\mu,23})+a_{32}[\lambda](Q_{\lambda,33}-Q_{\mu,33}))\\
&+a_{31}[\mu](a_{13}[\lambda](Q_{\lambda,11}-Q_{\mu,11})+a_{23}[\lambda](Q_{\lambda,21}-Q_{\mu,21})+a_{33}[\lambda](Q_{\lambda,31}-Q_{\mu,31}))\\
&+a_{32}[\mu](a_{13}[\lambda](Q_{\lambda,12}-Q_{\mu,12})+a_{23}[\lambda](Q_{\lambda,22}-Q_{\mu,22})+a_{33}[\lambda](Q_{\lambda,32}-Q_{\mu,32}))\\
&+a_{33}[\mu](a_{13}[\lambda](Q_{\lambda,13}-Q_{\mu,13})+a_{23}[\lambda](Q_{\lambda,23}-Q_{\mu,23})+a_{33}[\lambda](Q_{\lambda,33}-Q_{\mu,33}));
\end{aligned}
$$

$$\mathrm{Simplify}[f_4[\mu,\lambda]+f_4[\lambda,\mu]]$$

$$\mathrm{Out}[4]=0.$$

Proposition A.5:

$$\{\hat{\mathcal{F}}_\mu^{(i)},\hat{\mathcal{F}}_\lambda^{(j)}\}=0,\quad 1\leqslant i,j\leqslant 3,\forall\lambda,\mu\in\mathcal{C}, \tag{A25}$$

with $\hat{\mathcal{F}}_\lambda^{(1)}=I_\lambda^{(1)}+T_\lambda^{(1)}$, $\hat{\mathcal{F}}_\lambda^{(2)}=I_\lambda^{(2)}+T_\lambda^{(2)}+R_\lambda$.

Proof: It is easy to see that the equalities

$$\hat{\mathcal{F}}_\lambda^{(0)}=I_\lambda(\sigma)|_{\sigma=(1,1,1)},\quad I_\lambda^{(1)}=I_\lambda(\sigma)|_{\sigma=(\beta_2+\beta_3,\beta_1+\beta_3,\beta_1+\beta_2)},$$

$$I_\lambda^{(2)}=I_\lambda(\sigma)|_{\sigma=(\beta_2\beta_3,\beta_1\beta_3,\beta_1\beta_2)},\quad T_\lambda^{(1)}=T_\lambda(\sigma)|_{\sigma=(1,1,1)},\quad T_\lambda^{(2)}=T_\lambda(\sigma)|_{\sigma=(\beta_1,\beta_2,\beta_3)}.$$

Obviously $\{\hat{\mathcal{F}}_\mu^{(0)},\hat{\mathcal{F}}_\lambda^{(0)}\}=0$ from Eq. (A7). Using Eq. (A7) and the symmetry of Eq. (A15), we obtain

$$\{I_\mu^{(i)},I_\lambda^{(i)}\}=0,\quad \{I_\mu^{(i)},T_\lambda^{(i)}\}+\{T_\mu^{(i)},I_\lambda^{(i)}\}=0,\quad i=1,2, \tag{A26}$$

which together with the first expression of Eq. (A22) imply $\{\hat{\mathcal{F}}_\mu^{(1)},\hat{\mathcal{F}}_\lambda^{(1)}\}=0$ resorting to the bilinear property of the Poisson bracket. Noticing Eqs. (A8), (A16), and (A19), we get that

$$\{I_\mu^{(2)},R_\lambda\}+\{R_\mu,I_\lambda^{(2)}\}+\{T_\mu^{(2)},T_\lambda^{(2)}\}=0,\quad \{T_\mu^{(2)},R_\lambda\}+\{R_\mu,T_\lambda^{(2)}\}=0, \tag{A27}$$

which together with Eqs. (A26) and (A22) lead up to $\{\hat{\mathcal{F}}_\mu^{(2)},\hat{\mathcal{F}}_\lambda^{(2)}\}=0$. In view of Eqs. (A7), (A8), and (A15), we obtain

$$\{\hat{\mathcal{F}}_\mu^{(0)},I_\lambda^{(i)}\}=0,\quad \{\hat{\mathcal{F}}_\mu^{(0)},R_\lambda\}=0,\quad \{\hat{\mathcal{F}}_\mu^{(0)},T^{(i)}\}=0,\quad i=1,2,$$

$$\{I_\mu^{(1)},I_\lambda^{(2)}\}=0,\quad \{I_\mu^{(1)},R_\lambda\}+\{T_\mu^{(1)},T_\lambda^{(2)}\}=0,\quad \{I_\mu^{(1)},T_\lambda^{(2)}\}+\{T_\mu^{(1)},I_\lambda^{(2)}\}=0. \tag{A28}$$

By using Eq. (A28) and the third expression of Eq. (A22), we have

$$\{\hat{\mathcal{F}}_\mu^{(0)},\hat{\mathcal{F}}_\lambda^{(i)}\}=0,\quad \{\hat{\mathcal{F}}_\mu^{(1)},\hat{\mathcal{F}}_\lambda^{(2)}\}=0,\quad i=1,2.$$

The proof is completed.

Equation (4.9) and Proposition A.5 imply Theorem 4.1.

2. Two generators and the proof of Theorem 4.2

To prove Theorem 4.2, we introduce two generators of involutive systems of conserved integrals

$$\Gamma_l^{(ij)}=\sum_{\substack{k=1\\k\neq l}}^{N}\frac{B_{lk}^{(ij)}}{\lambda_l-\lambda_k},\quad B_{lk}^{(ij)}=(q_l^iq_k^j-q_k^iq_l^j)(p_l^ip_k^j-p_k^ip_l^j),\tag{A29}$$

$$\Upsilon_l=\sum_{\substack{s=1\\s\neq k}}^{N}\sum_{\substack{k=1\\k\neq l}}^{N}\frac{A_{lks}}{(\lambda_l-\lambda_k)(\lambda_k-\lambda_s)},\quad A_{lks}=\begin{vmatrix}q_l^1 & q_l^2 & q_l^3\\ q_k^1 & q_k^2 & q_k^3\\ q_s^1 & q_s^2 & q_s^3\end{vmatrix}\begin{vmatrix}p_l^1 & p_l^2 & p_l^3\\ p_k^1 & p_k^2 & p_k^3\\ p_s^1 & p_s^2 & p_s^3\end{vmatrix},\tag{A30}$$

which are two natural generalizations of the 2×2 case.[22]

Proposition A.6:

$$\overset{ii}{Q}_\lambda\overset{jj}{Q}_\lambda-\overset{ij}{Q}_\lambda\overset{ji}{Q}_\lambda=\sum_{l=1}^{N}\frac{\Gamma_l^{(ij)}}{\lambda-\lambda_l},\quad 1\leqslant i,j\leqslant 3,\tag{A31}$$

$$R_\lambda=\sum_{l=1}^{N}\frac{\Upsilon_l}{\lambda-\lambda_l}.\tag{A32}$$

Proof: Put the partial-fraction expansion of $\overset{ij}{Q}_\lambda$, i.e.,

$$\overset{ij}{Q}_\lambda=\sum_{l=1}^{N}\frac{q_l^ip_l^j}{\lambda-\lambda_l},\tag{A33}$$

into the right-hand side of Eq. (A31). The left-hand side of Eq. (A31) is obtained through direct calculations on account of

$$\frac{1}{(\lambda-\lambda_l)(\lambda-\lambda_k)}=\frac{1}{\lambda_l-\lambda_k}\left(\frac{1}{\lambda-\lambda_l}-\frac{1}{\lambda-\lambda_k}\right).\tag{A34}$$

Substituting Eq. (A33) into Eq. (A5) yields

$$\begin{aligned}R_\lambda&=\sum_{l=1}^{N}\sum_{k=1}^{N}\sum_{s=1}^{N}\frac{q_l^1q_k^2q_s^3}{(\lambda-\lambda_l)(\lambda-\lambda_k)(\lambda-\lambda_s)}\begin{vmatrix}p_l^1 & p_l^2 & p_l^3\\ p_k^1 & p_k^2 & p_k^3\\ p_s^1 & p_s^2 & p_s^3\end{vmatrix}\\&=\frac{1}{6}\sum_{l=1}^{N}\sum_{k=1}^{N}\sum_{s=1}^{N}\frac{A_{lks}}{(\lambda-\lambda_l)(\lambda-\lambda_k)(\lambda-\lambda_s)}.\end{aligned}$$

Resorting to Eq. (A34), we get

$$
\begin{aligned}
R_\lambda &= \frac{1}{6}\sum_l \sum_k \sum_{s\neq k} \frac{A_{lks}}{(\lambda_k-\lambda_s)(\lambda-\lambda_l)}\left(\frac{1}{\lambda-\lambda_k}-\frac{1}{\lambda-\lambda_s}\right)\\
&= \frac{1}{6}\sum_l \sum_{k\neq l} \sum_{s\neq k,l} \frac{A_{lks}}{\lambda_k-\lambda_s}\left[\frac{1}{\lambda_l-\lambda_k}\left(\frac{1}{\lambda-\lambda_l}-\frac{1}{\lambda-\lambda_k}\right)-\frac{1}{\lambda_l-\lambda_s}\left(\frac{1}{\lambda-\lambda_l}-\frac{1}{\lambda-\lambda_s}\right)\right]\\
&= \frac{2}{6}\sum_{l=1}^N \frac{\Upsilon_l}{\lambda-\lambda_l}+\frac{2}{6}\sum_l \sum_k \sum_{s\neq k,l} \frac{A_{lks}}{(\lambda_k-\lambda_s)(\lambda_l-\lambda_s)(\lambda-\lambda_s)}\\
&= \frac{2}{6}\sum_{l=1}^N \frac{\Upsilon_l}{\lambda-\lambda_l}+\frac{2}{6}\sum_l \sum_{k\neq l} \sum_{s\neq k,l} \frac{A_{lks}}{(\lambda-\lambda_s)(\lambda_l-\lambda_k)}\left(\frac{1}{\lambda_k-\lambda_s}-\frac{1}{\lambda_l-\lambda_s}\right)\\
&= \sum_{l=1}^N \frac{\Upsilon_l}{\lambda-\lambda_l},
\end{aligned}
$$

where it is used that $A_{llk}=A_{lkl}=A_{lkk}=0$, $A_{lks}=A_{lsk}=A_{skl}$. The proof is completed.

Proposition A.7: Let $E_1^{(i)},\ldots,E_N^{(i)}$ $(0\leqslant i\leqslant 2)$ defined as follows:

$$E_l^{(0)}=q_l^1p_l^1+q_l^2p_l^2+q_l^3p_l^3, \tag{A35}$$

$$E_l^{(1)}=(\beta_2+\beta_3)q_l^1p_l^1+(\beta_1+\beta_3)q_l^2p_l^2+(\beta_1+\beta_2)q_l^3p_l^3+\Gamma_l^{(12)}+\Gamma_l^{(13)}+\Gamma_l^{(23)}, \tag{A36}$$

$$E_l^{(2)}=\beta_2\beta_3q_l^1p_l^1+\beta_1\beta_3q_l^2p_l^2+\beta_1\beta_2q_l^3p_l^3+\beta_3\Gamma_l^{(12)}+\beta_2\Gamma_l^{(13)}+\beta_1\Gamma_l^{(23)}+\Upsilon_l. \tag{A37}$$

Then

$$F_m^{(i)}=\sum_{l=1}^N \lambda_l^m E_l^{(i)},\quad 0\leqslant i\leqslant 2. \tag{A38}$$

Proof: By Proposition A.6 and Eq. (A33), it is easy to see that

$$\mathcal{F}_\lambda^{(i)}=\sum_{l=1}^N \frac{E_l^{(i)}}{\lambda-\lambda_l}. \tag{A39}$$

Expanding $(\lambda-\lambda_l)^{-1}$ as a power series in λ^{-1} and substituting into Eq. (A39), we obtain

$$\mathcal{F}_\lambda^{(i)}=\sum_{l=0}^\infty \lambda^{-m-1}\sum_{l=1}^N \lambda_l^m E_l^{(i)},$$

which together with Eq. (4.9) gives $F_m^{(i)}=\sum_{l=1}^N\lambda_l^mE_l^{(i)}$.

Proposition A.8: The $3N$ 1-forms $dE_l^{(i)}$, $1\leqslant l\leqslant N$, $0\leqslant i\leqslant 2$, are linearly independent.

Proof: Suppose that there exist $3N$ constants $c_l^{(i)}$, $1\leqslant l\leqslant N$, $0\leqslant i\leqslant 2$, satisfying

$$\sum_{l=1}^N (c_l^{(0)}dE_l^{(0)}+c_l^{(1)}dE_l^{(1)}+c_l^{(2)}dE_l^{(2)})=0. \tag{A40}$$

Then Eq. (A40) implies

$$\sum_{l=1}^N \left(c_l^{(0)}\frac{\partial E_l^{(0)}}{\partial p^i}+c_l^{(1)}\frac{\partial E_l^{(1)}}{\partial p^i}+c_l^{(2)}\frac{\partial E_l^{(2)}}{\partial p^i}\right)=0,\quad 1\leqslant i\leqslant 3. \tag{A41}$$

In order to deduce all constants $c_l^{(i)}=0$, $1\leqslant l\leqslant N$, $0\leqslant i\leqslant 2$, we demand the following equalities, which can be calculated directly:

$$\frac{\partial^2 E_l^{(0)}}{\partial q^i \partial p^i}=(\underbrace{0,\ldots,0}_{l-1},1,\underbrace{0,\ldots,0}_{N-l})^T,\quad 1\leqslant i\leqslant 3, \tag{A42}$$

$$\frac{\partial^2 \Gamma_l^{(ij)}}{\partial q^i \partial p^i}=\left(\frac{q_l^j p_l^j}{\lambda_l-\lambda_1},\ldots,\frac{q_l^j p_l^j}{\lambda_l-\lambda_{l-1}},\sum_{\substack{k=1\\k\neq l}}^N \frac{q_k^j p_k^j}{\lambda_l-\lambda_k},\frac{q_l^j p_l^j}{\lambda_l-\lambda_{l+1}},\ldots,\frac{q_l^j p_l^j}{\lambda_l-\lambda_N}\right)^T, \tag{A43a}$$

$$\frac{\partial^2 \Gamma_l^{(ij)}}{\partial q^j \partial p^j}=\left(\frac{q_l^i p_l^i}{\lambda_l-\lambda_1},\ldots,\frac{q_l^i p_l^i}{\lambda_l-\lambda_{l-1}},\sum_{\substack{k=1\\k\neq l}}^N \frac{q_k^i p_k^i}{\lambda_l-\lambda_k},\frac{q_l^i p_l^i}{\lambda_l-\lambda_{l+1}},\ldots,\frac{q_l^i p_l^i}{\lambda_l-\lambda_N}\right)^T,\quad 1\leqslant i<j\leqslant 3, \tag{A43b}$$

$$\partial^2 \Gamma_l^{(ij)}/(\partial q^k \partial p^k)=0,\quad k\neq i,j. \tag{A43c}$$

Now we introduce the operator

$$D_s^{(j)}=\mathrm{diag}(\underbrace{\partial^2/(\partial q_s^j \partial p_s^j),\ldots,\partial^2/(\partial q_s^j \partial p_s^j)}_{N}),\quad 1\leqslant j\leqslant 3, 1\leqslant s\leqslant N.$$

Then we have from Eq. (A43) that

$$D_s^{(j)}\frac{\partial^2 \Gamma_l^{(ij)}}{\partial q^i \partial p^i}=\begin{cases}\left(\frac{1}{\lambda_l-\lambda_1},\ldots,\frac{1}{\lambda_l-\lambda_{l-1}},0,\frac{1}{\lambda_l-\lambda_{l+1}},\ldots,\frac{1}{\lambda_l-\lambda_N}\right)^T, & l=s\\ 0,\quad l\neq s, 1\leqslant i<j\leqslant 3,\end{cases} \tag{A44a}$$

$$D_s^{(i)}\frac{\partial^2 \Gamma_l^{(1j)}}{\partial q^1 \partial p^1}=0,\quad D_s^{(j)}\frac{\partial^2 \Gamma_l^{(23)}}{\partial q^1 \partial p^1}=0,\quad D_s^{(3)}\frac{\partial^2 \Gamma_l^{(1j)}}{\partial q^2 \partial p^2}=0,\quad 2\leqslant i,j\leqslant 3, i\neq j. \tag{A44b}$$

Acting with the operator $D_s^{(j)}\partial/(\partial q^i)|_{q=p=0}$, $(1\leqslant i<j\leqslant 3, 1\leqslant s\leqslant N)$, upon Eq. (A41), we get

$$c_s^{(1)}+\beta_1 c_s^{(2)}=0,\quad c_s^{(1)}+\beta_2 c_s^{(2)}=0,\quad c_s^{(1)}+\beta_3 c_s^{(2)}=0,\quad 1\leqslant s\leqslant N, \tag{A45}$$

in view of Eq. (A44) and the following equality:

$$D_s^{(j)}\frac{\partial^2 \Upsilon_l}{\partial q^i \partial p^i}\bigg|_{q=p=0}=0,\quad q=(q^1,q^2,q^3),\quad p=(p^1,p^2,p^3),\quad 1\leqslant i,j\leqslant 3, 1\leqslant l,s\leqslant N.$$

Equations (A45) imply $c_s^{(1)}=c_s^{(2)}=0$, $1\leqslant s\leqslant N$. Operating with $\partial/(\partial q^i)$ on Eq. (A41) and noticing $c_l^{(1)}=c_l^{(2)}=0$, we derive $c_l^{(0)}=0$, $1\leqslant l\leqslant N$. Hence the $3N$ 1-forms $dE_l^{(i)}$, $0\leqslant i\leqslant 2$, $1\leqslant l\leqslant N$, are linearly independent.

The proof of Theorem 4.2: Assume that there exist $3N$ constants $b_m^{(i)}$, $1\leqslant m\leqslant N$, $0\leqslant i\leqslant 2$, so that

$$\sum_{m=1}^N (b_m^{(0)}dF_m^{(0)}+b_m^{(1)}dF_m^{(1)}+b_m^{(2)}dF_m^{(2)})=0. \tag{A46}$$

Substituting Eq. (A38) into Eq. (A46) and noting the independence of the $dE_l^{(i)}$, $1\leqslant l\leqslant N$, $0\leqslant i\leqslant 2$, we get

$$\sum_{m=1}^{N} b_m^{(i)}\lambda_l^{m-1}=0, \quad 1\leqslant l\leqslant N, 0\leqslant i\leqslant 2,$$

which implies $b_m^{(i)}=0$, $1\leqslant m\leqslant N$, $0\leqslant i\leqslant 2$, by utilizing that Vandermonde determinant is not zero. The proof is finished.

ACKNOWLEDGMENTS

Project 19671074 was supported by National Natural Science Foundation of China. One of the authors (X. Geng) would like to thank the Henan Science Foundation Committee of China for financial support. This work was partially supported by Hong Kong RGC/97-98/21.

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[2] A. C. Newell, *Solitons in Mathematics and Physics* (SIAM, Philadelphia, 1985).
[3] S. P. Novikov, S. V. Manakov, L. P. Pitaevskii, and V. E. Zakharov, *Theory of Solitons, The Inverse Scattering Methods* (Nauka, Moskva, 1980) (in Russian) [English translation (Plenum, New York, 1984)].
[4] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer, Berlin, 1987).
[5] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[6] V. I. Arnold and S. P. Novikov, *Dynamical Systems IV* (Springer, Berlin, 1990).
[7] C. W. Cao and X. G. Geng, in *Nonlinear Physics, Research Reports in Physics*, edited by C. H. Gu, Y. S. Li, and G. Z. Tu (Springer, Berlin, 1990), pp. 68–78.
[8] C. W. Cao, Sci. China, Ser. A **33**, 528 (1990).
[9] C. W. Cao and X. G. Geng, J. Phys. A **23**, 4117 (1990).
[10] M. Antonowicz and S. Rauch-Wojciechowski, J. Phys. A **24**, 5043 (1991).
[11] M. Antonowicz and S. Rauch-Wojciechowski, J. Math. Phys. **33**, 2115 (1992).
[12] M. Blaszak, J. Phys. A **26**, 5985 (1993).
[13] O. Ragnisco, in *Solitons and Chaos*, edited by I. Antoniou and F. J. Lambert (Springer, Berlin, 1991), pp. 227–231.
[14] O. Ragnisco, Phys. Lett. A **167**, 165 (1992).
[15] X. G. Geng, Acta Math. Sin., New Ser. **8**, 78 (1992).
[16] X. G. Geng, J. Math. Phys. **34**, 805 (1993).
[17] O. Ragnisco, C. W. Cao, and Y. T. Wu, J. Phys. A **28**, 573 (1995).
[18] W. X. Ma and W. Strampp, Phys. Lett. A **185**, 277 (1994).
[19] W. X. Ma, Q. Ding, W. G. Zhang, and B. Q. Lu, Nuovo Cimento B **111**, 1135 (1996).
[20] X. G. Geng, Phys. Lett. A **194**, 44 (1994).
[21] Y. T. Wu and X. G. Geng, J. Math. Phys. **37**, 2338 (1996).
[22] C. H. Guo, in *Theory of Solitons and Its Application*, edited by C. H. Gu *et al.* (Zhejiang Publishing House of Science and Technology, China, 1990), pp. 141–174.
[23] J. Moser, *Integrable Hamiltonian System and Spectral Theory*, in *Proceedings of the Symposium on Differential Geometry and Differential Equations* (Science, Beijing, 1986), pp. 157–229.
[24] G. Z. Tu, J. Math. Phys. **30**, 330 (1989).
[25] D. J. Kaup, Stud. Appl. Math. **55**, 9 (1976).
[26] V. I. Arnold, *Mathematical Methods in Classical Mechanics* (Springer, Berlin, 1978).

Phase space reduction and Poisson structure

Nadhem Zaalani
Institut de Mathématiques, Université Pierre et Marie Curie 4, place Jussieu 75252, Paris Cedex 05, France

(Received 13 July 1998; accepted for publication 11 December 1998)

Let (P,π,B,G) be a G-principal fiber bundle. The action of G on the cotangent bundle T^*P is free and Hamiltonian. By Liberman and Marle [*Symplectic Geometry and Analytical Mechanics* (Reidel, Dortrecht, 1987)] and Marsden and Ratiu [Lett. Math. Phys. **11**, 161 (1981)] the quotient space T^*P/G is a Poisson manifold. We will determine the Poisson bracket on the reduced Poisson manifold T^*P/G, and its symplectic leaves. © *1999 American Institute of Physics.* [S0022-2488(99)02606-7]

I. INTRODUCTION

When a Lie group G acts on a manifold Q, it may foliate it into the orbits under the action. The quotient with respect to this foliation is Q/G. Under certain assumptions Q/G is a manifold with smooth projection from Q to Q/G. If Q is the phase space of a mechanical system, Q/G will often play the role of a reduced phase manifold, the carrier manifold of a reduced system. For this reason there is some interest in how structures, and more generally properties of Q, carry down to Q/G. Assume that Q is a Poisson manifold and that G acts on Q, with a Hamiltonian and free action. A result of Marsden and Ratiu[2] shows that the Poisson bracket on Q, projects onto a Poisson bracket on the quotient space Q/G. In particular, when Q is the phase space of a mechanical system, endowed with a free and Hamiltonian action of a Lie group G, the quotient space is a Poisson manifold.

Let G be a Lie group, $\mathcal{G}$ its Lie algebra, and (P,π,B,G) a G-principal bundle with total space P, base space B, and a principal action on the left

$$\Phi: G\times P\to P,$$

$$(g,p)\mapsto\Phi(g,p).$$

The canonical lifting $\hat{\Phi}$ of Φ to the cotangent bundle T^*P is a free and Hamiltonian action. By Refs. 1 and 2, the quotient space T^*P/G has a unique Poisson structure for which the canonical projection $\tau: T^*P\to T^*P/G$ is a Poisson map.

When a connection is given on P, we show that T^*P/G can be considered as a vector bundle over T^*B with $\mathcal{G}^*$ as typical fiber, and that its Poisson structure is the sum of three terms: the first one comes from the canonical symplectic structure of T^*B, the second one from the Lie–Poisson structure on $\mathcal{G}^*$, and the third one involves the curvature of the connection on P. Finally we show that the symplectic leaves of T^*P/G are bundles over T^*B with coadjoint orbit in $\mathcal{G}^*$ as a typical fiber. These fiber bundles are introduced by Sternberg in Ref. 3, who has given their symplectic structure, thus giving a symplectic formulation of the motion of a particule in a Yang–Mills field. Weinstein[4] has shown that these manifolds are reduced manifolds at $0\in\mathcal{G}^*$ of the diagonal action of G on the product $T^*P\times\mathcal{O}$, where $\mathcal{O}$ is a coadjoint orbit in $\mathcal{G}^*$. Note that Kummer[5] has shown that there is a symplectomorphism between the reduced symplectic manifold of T^*P at a G-invariant point $\mu\in\mathcal{G}^*$, and the symplectic space T^*B endowed with the symplectic form, sum of its canonical symplectic form and of a magnetique term depending on the curvature of the connection on P.

Those authors (Sternberg, Weinstein, and Kummer), were interested in the symplectic structures of the leaves of T^*P/G, by considering each one separately. We are interested in the Poisson structure of T^*P/G, which gathers together all of the symplectic structures of the leaves of T^*P/G.

Note also that Montgomery[6] has given a global formula of the Poisson bracket on T^*P/G, without proof. One of the principal objects of this paper is to give a proof of this formula.

It is important to remark that the Poisson structure on T^*P/G is useful especially for the study of the stability of some motions of a mechanical system, whose phase space is T^*P, when a Lie group G of symmetry of the system acts freely on the configuration space P, which is the set of all possible configurations of the mechanical system.

II. CONNECTION AND CANONICAL SYMPLECTIC FORM

Let (P,π,B,G) be a G-principal fiber bundle. This means we are given a free action, that we suppose left action

$$\Phi: G\times P\to P, \quad (g,p)\mapsto \Phi(g,p).$$

The canonical lift of Φ is a free and Hamiltonian action $\hat{\Phi}$ of G on T^*P, with moment mapping (see the definition in Refs. 1 or 7–9) $J: T^*P\to \mathcal{G}^*$, given for all $z\in T^*P$ and $\xi\in\mathcal{G}$ by

$$\langle J(z),\xi\rangle = \langle z,\xi_P(q_P(z))\rangle,$$

where $q_P: T^*P\to P$ is the canonical projection and ξ_P is the fundamental vector field on P associated with ξ, given by

$$\xi_P(p) = \frac{d}{dt}\Phi(\exp(-t\xi),p)\big|_{t=0}.$$

Suppose we are given a connection on P (see Refs. 10 and 11). It is a G-equivariant one-form ω on P with values in $\mathcal{G}$, $\Phi_g^*\omega = \mathrm{Ad}_g\omega$, and such that for all $p\in P$ and $\xi\in\mathcal{G}$, $\omega(\xi_P(p)) = \xi$. The tangent space at p splits into a direct sum $T_pP = V_pP\oplus H_pP$, where $V_pP = \ker T_p\pi$ is the vertical subspace and $H_pP = \ker\omega$ is the horizontal subspace defined by the connection. Then ω restricted to V_pP is the inverse of the mapping

$$u_p: \mathcal{G}\to V_p,$$

$$\xi\mapsto \xi_P(p).$$

The horizontal subspace H_pP is isomorphic to $T_{\pi(p)}B$, therefore, the cotangent space at p splits into the direct sum $T_p^*P = V_p^*P\oplus T_{\pi(p)}^*B$. Thus we obtain a projection

$$k_p: T_p^*P\to T_{\pi(p)}^*B,$$

which is the transpose of the horizontal lifting of vectors in $T_{\pi(p)}B$ to horizontal vectors in T_pP.

Let $z\in T_p^*P$, $Z\in T_z(T^*P)$ and $X = Tq_P(Z)\in T_pP$. Let $z = z_{\mathrm{ver}} + z_{\mathrm{hor}}$, $X = X_{\mathrm{ver}} + X_{\mathrm{hor}}$, be the decompositions of z and X relative to the direct sum decompositions given above, then the Liouville one-form evaluated on Z is

$$(\alpha_P)_z(Z) = \langle z,X\rangle = \langle z, X_{\mathrm{ver}} + X_{\mathrm{hor}}\rangle = \langle z,\xi_P(p)\rangle + \langle z, X_{\mathrm{hor}}\rangle.$$

Now, by definition of J,k and the connection one form ω, we have

$$\langle z,\xi_P(p)\rangle = \langle J(z),\xi\rangle, \quad \langle z,X_{\mathrm{hor}}\rangle = \langle k_p(z), T_p\pi(X)\rangle, \quad \omega(\xi_P(p)) = \xi,$$

therefore,

$$
\begin{aligned}
(\alpha_P)_z(Z) = \langle J(z),\xi\rangle + \langle k_p(z), T_p\pi(X)\rangle &= \langle J(z),\omega(\xi_P)\rangle + \langle k_p(z), T_p\pi(X)\rangle \\
&= \langle J(z),\omega(X)\rangle + \langle k_p(z), T_p\pi(X)\rangle \\
&= \langle J(z),(q_P^*\omega)(Z)\rangle + \langle k_p(z), Tk(Z)\rangle,
\end{aligned}
$$

and then we have the following proposition.

Proposition II.1: The Liouville one-form α_P and the canonical symplectic two-form $d\alpha_P$ on T^*P, are expressed by the following formulas:

$$\alpha_P = \langle J, q_P^*\omega\rangle + k^*\alpha_B, \quad d\alpha_P = dJ\wedge q_P^*\omega + \langle J, q_P^* d\omega\rangle + k^* d\alpha_B.$$

III. DECOMPOSITION OF THE COTANGENT BUNDLE

We denote by $\tilde{P} = \{(\beta,p)\in T^*B\times P / q_P(\beta) = \pi(p)\}$, the fiber product of T^*B and P over B. Let $\hat{\pi}:\tilde{P}\to P$ and $\tilde{\pi}:\tilde{P}\to T^*B$ be the first and the second projections. Then the following diagram is commutative:

$$
\begin{array}{ccc}
\tilde{P} & \xrightarrow{\hat{\pi}} & P \\
\tilde{\pi}\downarrow & & \downarrow\pi \\
T^*B & \xrightarrow{q_B} & B
\end{array}
$$

and $(\tilde{P},\tilde{\pi},T^*B,G)$ is a G-principal fiber bundle. The mapping $\hat{\pi}$ is a G-principal fiber bundles homomorphism. The mapping $\tilde{\omega} = \hat{\pi}^*\omega$ is then a connection on $\tilde{P}$. Let $\tilde{\Omega}$ be its curvature form. We have $\tilde{\Omega} = \tilde{\pi}^*\Omega$, where Ω is the curvature two-form of ω. It is a basic two-form on $\tilde{P}$ with values in $\mathcal{G}$, it can be thought as a two-form on T^*B with values in the associated fiber bundle to $\tilde{P}$ with typical fiber $\mathcal{G}$:$\tilde{P}\times_G\mathcal{G}$ (see Ref. 10).

Proposition III.1: (Maurer–Cartan equation). The curvature form Ω, satisfies the structure equation of Maurer–Cartan:

$$\text{for all } p\in P, X \text{ and } Y\in T_pP, \quad \Omega(X,Y) = d\omega(X,Y) + [\omega(X),\omega(Y)].$$

Proof: Since $TP = VP\oplus HP$, it is sufficient to consider the three following cases:
First case: X,Y are horizontal.

In this case, $\omega(X)=\omega(Y)=0$, $X_{\rm hor}=X$, and $Y_{\rm hor}=Y$, the equality reduces to the definition of Ω, $\Omega(X,Y) = d\omega(X_{\rm hor},Y_{\rm hor})$.
Second case: X,Y are vertical.

In this case, there exist ξ and ξ' in $\mathcal{G}$ such that $X=\xi_P(p)$ and $Y=\xi'_P(p)$

$$\omega(X)=\xi \text{ and } \omega(Y)=\xi'$$

$$X_{\rm hor}=Y_{\rm hor}=0 \text{ then } \Omega(X,Y)=0.$$

On the other hand,

$$d\omega(X,Y)+[\omega(X),\omega(Y)] = \xi_P(\omega(\xi'_P)) - \xi'_P(\omega(\xi_P)) - \omega([\xi_P,\xi'_P]) + [\omega(\xi_P),\omega(\xi'_P)].$$

Now, $\omega(\xi_P)$ and $\omega(\xi'_P)$ are two constant functions on P, then the two first right terms cancel. We have $[\xi_P,\xi'_P]=[\xi,\xi']_P$ then, $d\omega(\xi_P,\xi'_P)+[\xi,\xi'] = -[\xi,\xi']+[\xi,\xi']=0$.
Third case: X horizontal and Y vertical.
In this case, $X_{\rm hor}=X$, $Y_{\rm hor}=0$, $\omega(X)=0$, and $\omega(Y)=\xi$, where ξ is the unique element of $\mathcal{G}$, such that $\xi_P(p)=Y$.

We extend X to a horizontal vector field on P, which will be also denoted by X

$$\Omega(X,Y)=d\omega(X_{\text{hor}},Y_{\text{hor}})=0.$$

On the other hand,

$$d\omega(X,Y)+[\omega(X),\omega(Y)]=X(\omega(\xi_P))-\xi_P(\omega(X))-\omega([X,\xi_P])+[\omega(X),\omega(\xi_P)]=\omega([X,\xi])$$

Since X is horizontal, then $[X,\xi_P]$ also, therefore $\omega([X,\xi_P])=0$. ■

We assume that $\mathcal{G}^*$ is equipped with its canonical Lie–Poisson structure defined by the following bracket:

$$\text{for} \quad f_1 \ \text{and} \ f_2 \ \text{in} \ C^\infty(\mathcal{G}^*), \quad \{f_1,f_2\}(\mu)=-\langle \mu,[d_\mu f_1,d_\mu f_2]\rangle.$$

Here $\mathcal{G}^{**}$ is identified with $\mathcal{G}$ and [,] is the Lie bracket on $\mathcal{G}$.
We consider the following action on the product $\tilde{P}\times\mathcal{G}^*$, for $g\in G,(\beta,p)\in\tilde{P}$ and $\mu\in\mathcal{G}^*,g((\beta,p),\mu)=((\beta,\Phi(g,p)),\text{Ad}_g^*\mu)$.

In this case, the mapping

$$\psi:\tilde{P}\times\mathcal{G}^*\rightarrow T^*P, \quad ((\beta,p),\mu)\mapsto\beta+{}^t\omega_p(\mu),$$

is a G-equivariant isomorphism with inverse

$$\psi^{-1}:T^*P\rightarrow\tilde{P}\times\mathcal{G}^*, \quad z=z_{\text{ver}}+z_{\text{hor}}\mapsto(z_{\text{hor}},{}^tu(z)).$$

Thus, T^*P is isomorphic to $\tilde{P}\times\mathcal{G}^*$ and T^*P/G to the associated fiber bundle $\tilde{P}\times_G\mathcal{G}^*$, with base space T^*B and fiber $\mathcal{G}^*$. We denote $\pi_{\mathcal{G}^*}:\tilde{P}\times_G\mathcal{G}^*\rightarrow T^*B$ the canonical projection on the base space.

A. Moment mapping and canonical symplectic form

Let $z\in T^*P$ and $\xi\in\mathcal{G}$,

$$\langle {}^tu(z),\xi\rangle=\langle z,u(\xi)\rangle=\langle z,\xi_P(p)\rangle=\langle J(z),\xi\rangle.$$

The moment mapping J is then identified with the second projection from $\tilde{P}\times\mathcal{G}^*$ to $\mathcal{G}^*$, the mapping k with the projection $\tilde{\pi}\circ pr_1$, where $pr_1:\tilde{P}\times\mathcal{G}^*\rightarrow\tilde{P}$ is the first projection, and the mapping q_P is identified with $\hat{\pi}\circ pr_1$.

By proposition (II.1), $\alpha_P=\langle J,q_P^*\omega\rangle+k^*\alpha_B$, then after identification of T^*P with $\tilde{P}\times\mathcal{G}^*$, for all $z=(\beta,p,\mu)\in\tilde{P}\times\mathcal{G}^*$

$$\alpha_P(z)=\langle J,\tilde{\omega}\rangle+k^*\alpha_B(\beta).$$

A vector $Z\in T_z(\tilde{P}\times\mathcal{G}^*)$ can be written as $Z=(Z',\eta)$, where $Z'\in T_{(\beta,p)}\tilde{P}$ and $\eta\in T_\mu\mathcal{G}^*\equiv\mathcal{G}^*$, $\eta=T_zJ(Z)$.

Let $\xi=\tilde{\omega}(Z')$. In this case,

$$\alpha_P(Z)=\langle J,\tilde{\omega}\rangle(Z)+k^*\alpha_B(\beta)(Z)=\langle\mu,\xi\rangle+\alpha_B(\beta)(Tk(Z')).$$

Then, for all $Z_1=(Z_1',\eta_1)$ and $Z_2=(Z_2',\eta_2)$ in $T_z(\tilde{P}\times\mathcal{G}^*)$,

$$d\alpha_P(Z_1,Z_2)=dJ\wedge\tilde{\omega}(Z_1,Z_2)+\langle\mu,d\tilde{\omega}(Z_1',Z_2')\rangle+d\alpha_B(\beta)(Tk(Z_1'),Tk(Z_2')).$$

Now by the Maurer–Cartan equation (Proposition III.1), applied to $\tilde{\Omega}$, we have

$$d\tilde{\omega}(Z_1',Z_2')=\tilde{\Omega}(Z_1',Z_2')-[\tilde{\omega}(Z_1'),\tilde{\omega}(Z_2')],$$

then, if $\tilde{\omega}(Z_1')=\xi_1$ and $\tilde{\omega}(Z_2')=\xi_2$, we obtain

$$d\alpha_P(Z_1,Z_2)=dJ\wedge\tilde{\omega}(Z_1,Z_2)+\langle\mu,\tilde{\Omega}(Z_1',Z_2')\rangle-\langle\mu,[\xi_1,\xi_2]\rangle+d\alpha_B(\beta)(Tk(Z_1'),Tk(Z_2')). \tag{1}$$

B. Covariant differential of a function

The connection $\tilde{\omega}$ on $\tilde{P}$ generates a connection on the associated bundle $\tilde{P}\times_G\mathcal{G}^*$, which is given by a horizontal subbundle $H(\tilde{P}\times_G\mathcal{G}^*)$, complementary in $T(\tilde{P}\times_G\mathcal{G}^*)$ to the vertical subbundle $V(\tilde{P}\times_G\mathcal{G}^*)=\ker\pi_{\mathcal{G}^*}$. Let h be the horizontal lifting of the tangent vectors on T^*B to the horizontal vectors on $\tilde{P}\times_G\mathcal{G}^*$.

Definition: We define the covariant differential of a function $f\in C^\infty(\tilde{P}\times_G\mathcal{G}^*)$ on $\hat{z}\in\tilde{P}\times_G\mathcal{G}^*$ to be that covector $d_{\tilde{\omega}}f(\hat{z})$ on $\beta=\pi_{\mathcal{G}^*}(\hat{z})\in T^*B$, given for all $V_\beta\in T_\beta(T^*B)$ by

$$\langle d_{\tilde{\omega}}f(\hat{z}),V_\beta\rangle=\langle df(\hat{z}),hV_\beta\rangle.$$

This may be thought of as the horizontal part of $df(\hat{z})$. The vertical part may be thought of as an element in the dual bundle to $\tilde{P}\times_G\mathcal{G}^*$ which is the adjoint bundle $\tilde{P}\times_G\mathcal{G}$, with base space T^*B and fiber $\mathcal{G}$. This vertical part is the differential of f with respect to the variable $\hat{\mu}$ in the fiber $(\mathcal{G}^*)_{\hat{z}}$ through $\hat{z}$. Thus,

$$df(\hat{z})=d_{\tilde{\omega}}f(\hat{z})+d_{\hat{\mu}}f(\hat{z}).$$

The associated fiber bundle $\tilde{P}\times_G\mathcal{G}$ is a bundle with base space T^*B and typical fibre $\mathcal{G}$. Then every fiber over an element $\beta\in T^*B$ has a Lie algebra structure. We denote by $[,]$ its Lie bracket.

IV. REDUCED POISSON BRACKET

Theorem IV.1: Let (P,π,B,G) be a G-principal fiber bundle, and ω a connection on P. The quotient space T^*P/G has a unique Poisson structure for which the canonical projection is a Poisson map, and after the identification of T^*P/G with $\tilde{P}\times_G\mathcal{G}^*$, the Poisson bracket of two functions f_1 and f_2 in $C^\infty(\tilde{P}\times_G\mathcal{G}^*)$, evaluated at a point $\hat{z}$ in $\tilde{P}\times_G\mathcal{G}^*$, is

$$\{f_1,f_2\}(\hat{z})=d\alpha_B(\beta)(d_{\tilde{\omega}}f_1^\sharp(\hat{z}),d_{\tilde{\omega}}f_2^\sharp(\hat{z}))+\langle\hat{z},\tilde{\Omega}(\beta)(d_{\tilde{\omega}}f_1^\sharp(\hat{z}),d_{\tilde{\omega}}f_2^\sharp(\hat{z}))\rangle$$
$$-\langle\hat{z},[d_{\hat{\mu}}f_1(\hat{z}),d_{\hat{\mu}}f_2(\hat{z})]\rangle.$$

In this formula, the vector $d_{\tilde{\omega}}f^\sharp(\hat{z})$ is the unique vector in $T_\beta(T^*B)$, associated to $d_{\tilde{\omega}}f(\hat{z})$ by the isomorphism generated by the symplectic form $d\alpha_B$, between $T^*(T^*B)$ and $T(T^*B)$. The two-form $\tilde{\Omega}$ is considered as a two-form on T^*B with values in $\tilde{P}\times_G\mathcal{G}$, by the mapping

$$(V_\beta,W_\beta)\mapsto(\beta,\tilde{\Omega}(hV_\beta,hW_\beta)),$$

where $\beta\in T^*B$, V_β and $W_\beta\in T^*B$, and hV_β,hW_β are their corresponding horizontal liftings, with respect to the connection $\tilde{\omega}$.

Proof: Let $\tau_{\mathcal{G}^*}:\tilde{P}\times\mathcal{G}^*\to\tilde{P}\times_G\mathcal{G}^*$, be the canonical projection, which maps a point $z=(\beta,p,\mu)\in\tilde{P}\times\mathcal{G}^*$, to its G orbit, $\hat{z}=\tau_{\mathcal{G}^*}(z)\in\tilde{P}\times_G\mathcal{G}^*$.
As show by Marsden and Ratiu in Ref. 2, the quotient space of a Poisson manifold by a Hamiltonian and free action of a Lie group, has a unique Poisson structure, for which the projection, which sends a point to its orbit, is a Poisson map.

The associated fiber bundle $\tilde{P}\times_G\mathcal{G}^*$ is endowed with the unique Poisson structure for which the projection $\tau_{\mathcal{G}^*}$ is a Poisson map. Then, if f_1,f_2 are in $C^\infty(\tilde{P}\times_G\mathcal{G}^*)$, we have

$$\{f_1,f_2\}(\hat{z})=d\alpha_P(z)(d(f_1\circ\tau_{\mathcal{G}^*})^\sharp,d(f_2\circ\tau_{\mathcal{G}^*})^\sharp).$$

Now,

$$df_i(\hat{z}) = d_{\tilde{\omega}} f_i(\hat{z}) + d_{\hat{n}} f_i(\hat{z}) \quad \text{for } i=1,2,$$

and by formula (1), we have

$$d\alpha_P(z)(d(f_1\circ\tau_{\mathcal{G}^*})^\sharp, d(f_1\circ\tau_{\mathcal{G}^*})^\sharp) = dJ\wedge\tilde{\omega}(d(f_1\circ\tau_{\mathcal{G}^*})^\sharp, d(f_2\circ\tau_{\mathcal{G}^*})^\sharp) + \langle J(z), \tilde{\Omega}(z) \times (d_{\tilde{\omega}}(f_1\circ\tau_{\mathcal{G}^*})^\sharp, d_{\tilde{\omega}}(f_2\circ\tau_{\mathcal{G}^*})^\sharp)\rangle - \langle J(z), [d_\mu(f_1\circ\tau_{\mathcal{G}^*})^\sharp, d_\mu \times (f_2\circ\tau_{\mathcal{G}^*})^\sharp]\rangle + d\alpha_B(\beta)(d_{\tilde{\omega}}(f_1\circ\tau_{\mathcal{G}^*})^\sharp, d_{\tilde{\omega}}(f_2\circ\tau_{\mathcal{G}^*})^\sharp).$$

The quadruple $(\tilde{P}\times\mathcal{G}^*, \tau_{\mathcal{G}^*}, \tilde{P}\times_G\mathcal{G}^*, G)$ is a G-principal fiber bundle (see Kobayashi and Nomizu[11]). The connection one-form $\tilde{\omega}$ on $\tilde{P}$ can be thought of as a one-form on $\tilde{P}\times\mathcal{G}^*$ (depending only on coordinates in $\tilde{P}$). For $i=1,2$ $f_i\circ\tau_{\mathcal{G}^*}$ is a G-invariant function on $\tilde{P}\times\mathcal{G}^*$. There is an isomorphism between the set of functions on $\tilde{P}\times_G\mathcal{G}^*$ and the set of G-invariant functions on $\tilde{P}\times\mathcal{G}^*$. Then, $d(f_i\circ\tau_{\mathcal{G}^*})(z)^\sharp$ is the horizontal lifting of $df_i^\sharp(\hat{z})$ to $\tilde{P}\times\mathcal{G}^*$. It is then a horizontal vector. We have then

$$\tilde{\omega}(d(f_i\circ\tau_{\mathcal{G}^*})^\sharp) = 0 \quad dJ\wedge\tilde{\omega}(d(f_1\circ\tau_{\mathcal{G}^*})^\sharp, d(f_2\circ\tau_{\mathcal{G}^*})^\sharp) = 0.$$

On the other hand the moment mapping J is the second projection from $\tilde{P}\times\mathcal{G}^*$ to $\mathcal{G}^*$. It generates a mapping $\tilde{J}$ from $\tilde{P}\times_G\mathcal{G}^*$ on $\tilde{P}\times_G\mathcal{G}^*$, which is the identity mapping. Then,

$$\begin{aligned}\{f_1,f_2\}(\hat{z}) &= d\alpha_B(\beta)(d_{\tilde{\omega}}f_1^\sharp(\tau_{\mathcal{G}^*}(z), d_{\tilde{\omega}}f_2^\sharp(\tau_{\mathcal{G}^*}(z))) + \langle\tilde{J}(\hat{z}), \tilde{\Omega}(\beta)(d_{\tilde{\omega}}f_1^\sharp(\tau_{\mathcal{G}^*}(z)), d_{\tilde{\omega}}f_2^\sharp(\tau_{\mathcal{G}^*}(z)))\rangle \\ &\quad - \langle\tilde{J}(\hat{z}), [d_{\hat{\mu}}f_1(\tau_{\mathcal{G}^*}(z)), d_{\hat{\mu}}f_2(\tau_{\mathcal{G}^*}(z))]\rangle \\ &= d\alpha_B(\beta)(d_{\tilde{\omega}}f_1^\sharp(\hat{z}), d_{\tilde{\omega}}f_2^\sharp(\hat{z})) + \langle\hat{z}, \tilde{\Omega}(\beta)(d_{\tilde{\omega}}f_1^\sharp(\hat{z}), d_{\tilde{\omega}}f_2^\sharp(\hat{z})\rangle - \langle\hat{z}, [d_{\hat{\mu}}f_1(\hat{z}), d_{\hat{\mu}}f_2(\hat{z})]\rangle.\end{aligned}$$

■

V. SYMPLECTIC LEAVES

Let us recall a result of Sternberg.[3] Let (F,Ω_F) be a symplectic manifold with symplectic form Ω_F, on which a Lie group G acts by a Hamiltonian action on the right, with a G-equivariant moment $J_F: F\to\mathcal{G}^*$. Then the choice of a connection on P determines a symplectic structure on the associated bundle $\tilde{P}\times_G F$. The symplectic form on $\tilde{P}\times_G F$ is

$$\sigma = \pi_F^* d\alpha_B - \langle\tilde{J}_F, \pi_F^*\tilde{\Omega}\rangle + \tilde{\Omega}_F,$$

where $\tilde{J}$ is the mapping on $\tilde{P}\times_G F$ with values in $\tilde{P}\times_G\mathcal{G}^*$ generated by the G-equivariant moment map J_F, which makes the following diagram commutative:

$$\begin{array}{ccc} \tilde{P}\times F & \xrightarrow{id_{\tilde{P}}\times J_F} & \tilde{P}\times\mathcal{G}^* \\ \downarrow & & \downarrow \\ \tilde{P}\times_G F & \xrightarrow{\tilde{J}_F} & \tilde{P}\times_G\mathcal{G}^* \end{array}$$

The curvature two-form $\tilde{\Omega}$ on $\tilde{P}$ is equivariant and cancels on vertical vector fields. It can be considered as a two-form on T^*B with values in $\tilde{P}\times_G\mathcal{G}$ (see Ref. 10). The form $\tilde{\Omega}_F$ is the two-form on $\tilde{P}\times_G F$ generated by the G-invariant symplectic two-form Ω_F. The mapping $\pi_F: \tilde{P}\times_G F\to T^*B$ is the canonical projection on the base space.

Let us take $F=\mathcal{O}$ a coadjoint orbit in $\mathcal{G}^*$, with its canonical symplectic two-form $\Omega_{\mathcal{O}}$, given for all $\mu\in\mathcal{O}$ and $\xi,\zeta\in\mathcal{G}$, by

$$\Omega_{\mathcal{O}}(\mu)(\xi_{\mathcal{O}},\zeta_{\mathcal{O}})=-\langle\mu,[\xi,\zeta]\rangle.$$

Then the moment map $J_{\mathcal{O}}$ is equal to $-i_{\mathcal{O}}$, where $i_{\mathcal{O}}$ is the injection of $\mathcal{O}$ in $\mathcal{G}^*$. Consequently the mapping $-\tilde{J}_{\mathcal{O}}$ is the injection of $\tilde{P}\times_G\mathcal{O}$ in $\tilde{P}\times_G\mathcal{G}^*$: we have $\tilde{J}_{\mathcal{O}}(\hat{z})=-\hat{z}$. The symplectic two-form on the symplectic space $\tilde{P}\times_G\mathcal{O}$, evaluated at a point $\hat{z}\in\tilde{P}\times_G\mathcal{O}$, $\beta=\pi_{\mathcal{O}}(\hat{z})\in T^*B$, and $\hat{Z}_1,\hat{Z}_2\in T_{\hat{z}}(\tilde{P}\times_G\mathcal{O})$ is

$$\sigma(\hat{Z}_1,\hat{Z}_2)=d\alpha_B(T\pi_{\mathcal{O}}(\hat{Z}_1),T\pi_{\mathcal{O}}(\hat{Z}_2))+\langle\hat{z},\tilde{\Omega}(T\pi_{\mathcal{O}}(\hat{Z}_1),T\pi_{\mathcal{O}}(\hat{Z}_2))\rangle-\langle\hat{z},[\hat{Z}_{1\mathrm{ver}},\hat{Z}_{2\mathrm{ver}}]\rangle.$$

Theorem V.1: The symplectic leaves of $\tilde{P}\times_G\mathcal{G}^*$ are the symplectic fiber bundles $\tilde{P}\times_G\mathcal{O}$, associated to the principal fiber bundle $\tilde{P}$, with base space T^*B and typical fiber a coadjoint orbit $\mathcal{O}$ in $\mathcal{G}^*$.

Proof: The symplectic leaves of T^*P/G are the Marsden–Weinstein reduced spaces (see Ref. 12) $J^{-1}(\mu)/G_\mu$, where $\mu\in\mathcal{G}^*$ and G_μ its isotropy subgroup with respect to the coadjoint action of G on $\mathcal{G}^*$. By Marle[13] the reduced space $J^{-1}(\mu)/G_\mu$ is symplectomorphic to the symplectic space $J^{-1}(\mathcal{O})/G$, where $\mathcal{O}$ is the coadjoint orbit containing μ. After identification of T^*P with $\tilde{P}\times\mathcal{G}^*$, the moment mapping J, is the second projection from $\tilde{P}\times\mathcal{G}^*$ to $\mathcal{G}^*$. Therefore,

$$J^{-1}(\mathcal{O})=\tilde{P}\times\mathcal{O}\quad\text{and}\quad J^{-1}(\mathcal{O})/G=\tilde{P}\times_G\mathcal{O}.$$

Let us take $\hat{z}\in\tilde{P}\times_G\mathcal{O}$, $\beta=\pi_{\mathcal{O}}(\hat{z})\in T^*B$, f_1 and f_2 in $C^\infty(\tilde{P}\times_G\mathcal{O})$, and let us denote by, $\hat{Z}_{f_i}$ the value on $\hat{z}$, of the Hamiltonian vector field associated to f_i, for $i=1,2$, then

$$\begin{aligned}\sigma_{\hat{z}}(\hat{Z}_{f_1},\hat{Z}_{f_2})&=d\alpha_B(\beta)(T\pi_{\mathcal{O}}(\hat{Z}_{f_1}),T\pi_{\mathcal{O}}(\hat{Z}_{f_2})+\langle\hat{z},\tilde{\Omega}(\beta)(T\pi_{\mathcal{O}}(\hat{Z}_{f_1}),T\pi_{\mathcal{O}}(\hat{Z}_{f_2}))\rangle\\&\quad-\langle\hat{z},[(\hat{Z}_{f_1})_{\mathrm{ver}},(\hat{Z}_{f_2})_{\mathrm{ver}}]\rangle\\&=d\alpha_B(\beta)(d_{\tilde{\omega}}f_1^\sharp(\tau_{\mathcal{O}}(z)),d_{\tilde{\omega}}f_2^\sharp(\tau_{\mathcal{O}}(z)))+\langle\hat{z},\tilde{\Omega}(\beta)(d_{\tilde{\omega}}f_1^\sharp(\tau_{\mathcal{O}}(z)),d_{\tilde{\omega}}f_2^\sharp(\tau_{\mathcal{O}}(z)))\rangle\\&\quad-\langle\hat{z},[d_{\hat{\mu}}f_1^\sharp(\tau_{\mathcal{O}}(z)),d_{\hat{\mu}}f_2^\sharp(\tau_{\mathcal{O}}(z))]\rangle\\&=\{f_1,f_2\}_{T^*P/G}.\end{aligned}$$

We see that the restriction of the Poisson bracket of $\tilde{P}\times_G\mathcal{G}^*$ to $\tilde{P}\times_G\mathcal{O}$, is equal to the Poisson bracket generated by the symplectic form on $\tilde{P}\times_G\mathcal{O}$. This proves the theorem V.1. ■

VI. EXAMPLE

Let $G=\mathrm{SO}(3)$, be the rotation group of the Euclidean three space $\mathbb{R}^3$ about its origin, and let $K=S^1$ be the rotation group of the plane $\mathbb{R}^2$ considered as a subgroup of G, by identification with elements of the form

$$\begin{pmatrix}a&-b&0\\b&a&0\\0&0&1\end{pmatrix},\quad a^2+b^2=1.$$

Let us consider the action Φ of K on G given by the restriction of the left translations of G to $K\times G$

$$\Phi:K\times G\to G,\quad (k,g)\mapsto kg.$$

The set of orbits $\mathrm{SO}(3)/S^1$ can be identified with S^2. Let $\pi_K:\mathrm{SO}(3)\rightarrow S^2$, be the canonical projection which sends an element $A\in \mathrm{SO}(3)$ to its K orbit. In this case (G,π_K,S^2,S^1) is a S^1-principal fiber bundle with total space SO(3) and base space S^2. The Lie algebra so(3) of SO(3), endowed with its Lie bracket

$$[A,B]=AB-BA,$$

is identified with $\mathbb{R}^3$ endowed with the vector product $\times$ (see Ref. 7). The tangent fiber bundle to SO(3) is identified with $\mathrm{SO}(3)\times\mathbb{R}^3$ (see Ref. 1). The Lie algebra of S^1 is identified with $\mathbb{R}$, with trivial Lie bracket. The mapping which sends $(A,a), A\in \mathrm{SO}(3)$ and $a=(a_1,a_2,a_3)\in\mathbb{R}^3$, to a_3, is a connection on $(\mathrm{SO}(3),\pi_K,S^2,S^1)$. Its curvature two-form is the mapping Ω, evaluated at a pair $((A,a),(A',a'))\in T\mathrm{SO}(3)\times T\mathrm{SO}(3)$, is (by Ref. 10)

$$\Omega((A,a),(A',a'))=-(a\times a')_3=a_1a_2'-a_2a_1'.$$

The orbit set of the canonical lifting $\hat{\Phi}$ of Φ to the cotangent bundle $T^*\mathrm{SO}(3)$, is a fiber bundle with base space T^*S^2 and typical fiber $\mathbb{R}$. The covariant differential of a function $f\in C^\infty(T^*\mathrm{SO}(3)/S^1)$ is the differential with respect to $(s,(a_1,a_2))$, where $s\in S^2$ and $(a_1,a_2)\in T_s^*S^2$.

The reduced Poisson bracket of two functions f_1 and $f_2\in C^\infty(T^*\mathrm{SO}(3)/S^1)$, evaluated at a point $(s,(a_1,a_2),a_3)\in T^*\mathrm{SO}(3)/S^1$, is

$$\begin{aligned}\{f_1,f_2\}&=d\alpha_{S^2}(d_{(s,(a_1,a_2))}f_1^\#,d_{(s,(a_1,a_2))}f_2^\#)+\langle a_3,-(df_1\times df_2)_3\rangle\\&=\langle d_sf_1,d_{(a_1,a_2)}f_2(s)\rangle-\langle d_sf_2,d_{(a_1,a_2)}f_1(s)\rangle-\langle a_3,(df_1\times df_2)_3\rangle\\&=\langle d_sf_1,d_{(a_1,a_2)}f_2(s)\rangle-\langle d_sf_2,d_{(a_1,a_2)}f_1(s)\rangle-a_3\left(\frac{\partial f_1}{\partial a_1}\frac{\partial f_2}{\partial a_2}-\frac{\partial f_1}{\partial a_2}\frac{\partial f_2}{\partial a_1}\right).\end{aligned}$$

The Lie bracket on the Lie algebra of S^1 is trivial. Then the symplectic leaves of the reduced cotangent bundle to SO(3) are all diffeomorphic to T^*S^2. ■

[1] P. Liberman and C. M. Marle, *Symplectic Geometry and Analytical Mechanics* (Reidel, Dortrecht, 1987).
[2] J. E. Marsden and T. S. Ratiu, ''Reduction of Poisson manifolds,'' Lett. Math. Phys. **11**, 161–170 (1986).
[3] S. Sternberg, ''Minimal coupling and the symplectic mechanics of a classical particle in the presence of a Yang-Mills field,'' Proc. Natl. Acad. Sci. USA **74**, 5253–5254 (1977).
[4] A. Weinstein, ''A universal phase space for particules in Yang–Mills fields,'' Lett. Math. Phys. **2**, 417–420 (1978).
[5] M. Kummer, ''On the construction of reduced phase space of a Hamiltonian system with symmetry,'' Indiana. Univ. Math. J. **30**, 281–291 (1981).
[6] R. Montgomery, J. E. Marsden, and T. S. Ratiu, ''Gauged Lie Poisson structures,'' Contemp. Math. **28**, 101–114 (1984).
[7] R. Abraham and J. E. Marsden, *Foundations of Mechanics*. 2nd ed. (Addison–Wesley, Reading, MA, 1978).
[8] J. M. Souriau, *Structure des Systèmes Dynamiques* (Dunod, Paris, 1970).
[9] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics* (Cambridge University Press, New York, 1984).
[10] W. Greub, S. Halperin, and R. Vanstone, *Connections, Curvatures and Cohomology* (Academic, New York, 1973).
[11] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1963, 1969).
[12] J. E. Marsden and A. Weinstein, ''Reduction of symplectic manifolds with symmetry,'' Rep. Math. Phys. **5**, 121–130 (1974).
[13] C. M. Marle, *Symplectic Manifolds, Dynamical Groups and Hamiltonian Mechanics* (Reidel, Dortrecht, 1976).

Cauchy analysis of the linearized skew sector of the massive nonsymmetric gravitational theory

P. Baki[a] and J. O. Malo[a]
Department of Physics, University of Nairobi, Box 30197, Nairobi, Kenya

(Received 4 November 1998; accepted for publication 14 December 1998)

Cauchy analysis of the linearized field equations of the skew sector of the massive nonsymmetric gravitational theory shows that small perturbations give rise to bounded accelerations thereby ensuring good asymptotic behavior for the skew part of the fundamental tensor.
[S0022-2488(99)04406-0]

I. INTRODUCTION

The nonsymmetric gravitational theory (NGT) arose out of a reinterpretation of Einstein's unified field theory[1,2] as purely a theory of gravitation.[3–5] However, the original versions of NGT were confronted with perturbative consistency problems due to the absence of a massless gauge invariance in the skew sector of the theory. The only gauge invariance was that associated with general covariance of the theory[6–8] which was clarified by Damour *et al.*, who showed that the wave solutions of the weak-field equations did not decrease at large distances from the source along the forward light cone.[9,10] They proposed that the theory should reduce to the massive Kalb–Ramond theory, which does not require gauge invariance for well behaved positive energy solutions.

The theory was subsequently altered to massive nonsymmetric gravitational theory (MNGT)[11–13] by requiring that the linearized field equations reduce to those of a massive Kalb–Ramond field,[14] guaranteeing that the linearized fields are well behaved asymptotically far from the source.[15] A study of the instabilities of the linearized MNGT showed that the instabilities arise from the fact that small skew sector perturbations lead to unboundedly large accelerations.[16] However, it will be shown that such instabilities may not in fact exist.

The organization of the paper is as follows: Section II presents an overview of the skew sector of MNGT on a general relativity (GR) background, while in Sec. III the linearized field equations are expressed in $3+1$ decomposed form. The analysis of these equations is carried out in Sec. IV, while in the Appendix we give the mathematical details of Sec. III.

II. THE LINEARIZED SKEW SECTOR OF MNGT

The linearized MNGT presented here is that due to Clayton,[15] in which the dynamics of the theory are determined from the first-order action ($G=c=1$):

$$S=\int d^4x\left\{-\hat{g}^{\mu\nu}R^{NS}_{\mu\nu}(\Gamma)-\hat{g}^{\mu\nu}\partial_{[\mu W\nu]}+\hat{l}^{\mu}\Gamma_{\mu}+\frac{1}{2}\lambda\hat{g}^{(\mu\nu)}W_{\mu}W_{\nu}+\frac{1}{4}m^2\hat{g}^{[\mu\nu]}g_{[\mu\nu]}\right\}+S_M \tag{1}$$

where the quantities with a caret are densitized, and

$$\frac{\delta S_M}{\delta g^{\mu\nu}}=\hat{T}_{\mu\nu}$$

[a]Electronic mail: physics@ken.healthnet.org

is the matter stress energy tensor derived from the variation of the matter action S_M, that acts as a source in the gravitational field equations. The analysis of the theory is carried out in the context of the Palatini formalism and takes place at the level of the field equations. Thus performing a variation of the action with respect to $g_{\mu\nu}$ yields the field equations:

$$\frac{\delta S}{\delta \hat{g}^{\mu\nu}}=R^{NS}_{\mu\nu}+\partial_{[\mu W_{\nu}]}-\frac{1}{2}\lambda W_\mu W_\nu-\frac{1}{4}m^2\left(g_{[\mu\nu]}-g_{\alpha\mu}g_{\nu\beta}g^{[\alpha\beta]}+\frac{1}{2}g_{\nu\mu}g^{[\alpha\beta]}g_{[\alpha\beta]}\right)=\hat{T}_{\mu\nu}-\frac{1}{2}g_{\mu\nu}\hat{T}. \tag{2}$$

If one considers all the antisymmetric parts of the fundamental tensor as perturbations about a symmetric GR background,[15,7,13] i.e.,

$$g_{\mu\nu}\rightarrow g_{(\mu\nu)}+h_{\mu\nu} \tag{3}$$

and also the affine connections as:

$$\Gamma^{\alpha}_{\mu\nu}\rightarrow\left\{\begin{matrix}\alpha\\ \mu\nu\end{matrix}\right\}+\gamma^{\alpha}_{\mu\nu}, \tag{4}$$

where $h_{\mu\nu}$ and $\gamma^{\alpha}_{\mu\nu}$ are the perturbed quantities, and $g_{(\mu\nu)}$ and $\{{}^{\alpha}_{\mu\nu}\}$ are the symmetric background metric and christoffel symbols, respectively; with the fundamental tensor defined by

$$g_{\mu\nu}g^{\alpha\nu}=\delta^{\nu}_{\mu}. \tag{5}$$

Following Eq. (2), the field equations of MNGT expanded to first order about a GR background yields

$${}^1R_{\mu\nu}+\frac{1}{\lambda}\nabla_{[\mu}\nabla^{\alpha}h_{[\mu\nu]]}-\frac{1}{2}m^2h_{[\mu\nu]}=\hat{T}_{\mu\nu}-\frac{1}{2}g_{\mu\nu}\hat{T}, \tag{6}$$

where the first-order correction to Ricci-like tensor in NGT is given by

$${}^1R_{\mu\nu}=\nabla_{\alpha}\gamma^{\alpha}_{\mu\nu}-\nabla_{(\nu}\gamma^{\alpha}_{\alpha\mu)}. \tag{7}$$

By ignoring symmetric GR perturbations, the antisymmetric part of (6) is obtained as:

$$\begin{aligned}\nabla_{\alpha}\gamma^{\alpha}_{[\mu\nu]}+\frac{1}{\lambda}\nabla_{[\mu}\nabla^{\alpha}h_{[\mu\nu]]}-\frac{1}{2}m^2h_{[\mu\nu]}&=\frac{1}{2}(\nabla_{\alpha}F_{\mu\nu\alpha}+m^2h_{[\mu\nu]})-2\nabla^{\alpha}\nabla_{[\mu}h_{[\alpha\nu]]}\\&\quad+\frac{1}{\lambda}[1+\tfrac{2}{3}\lambda]\nabla_{[\mu}\nabla^{\alpha}h_{[\alpha\nu]]}\\&=\hat{T}_{[\mu\nu]},\end{aligned} \tag{8}$$

where

$$F_{\mu\nu\alpha}=(\partial_{\mu}h_{\nu\alpha}+\partial_{\alpha}h_{\mu\nu}+\partial_{\nu}h_{\alpha\mu}). \tag{9}$$

If one expands the action (1) to second order, ignores the surface terms, and imposes the compatibility conditions following the removal of the torsion vector $W_{\mu\nu}$ from the dynamics,[15] then one obtains the action as:

$$S=\tfrac{1}{12}F^{\mu\nu\alpha}F_{\mu\nu\alpha}-\tfrac{1}{4}m^2h^{[\mu\nu]}h_{[\mu\nu]}-\nabla^{\alpha}h^{[\mu\nu]}\nabla_{\mu}h_{[\alpha\mu]}(-\tfrac{1}{2}\lambda+\tfrac{1}{3})\nabla_{\nu}h^{[\mu\nu]}\nabla^{\alpha}h_{[\mu\alpha]}. \tag{10}$$

And by choosing $\lambda=\frac{3}{4}$ one obtains kinetic terms identical to those of Kalb–Ramond theory on a GR background, giving the skew sector action

$$S_s = \tfrac{1}{12} F^{\mu\nu\alpha} F_{\mu\nu\alpha} - \tfrac{1}{4} m^2 h^{[\mu\nu]} h_{[\mu\nu]} - h^{[\mu\nu]} h^{[\alpha\beta]} R_{\alpha\mu\beta\nu}, \tag{11}$$

where ${}^0R_{\alpha\mu\beta\nu}$ designates the background curvature. By taking a variation of the action (11) with respect to $h^{[\mu\nu]}$ one obtains the linearized field equations

$$\nabla_\alpha F_{\mu\nu\alpha} + m^2 h_{[\mu\nu]} - 4 h^{[\alpha\beta]} R_{\alpha\mu\beta\nu} = 2\hat{T}_{[\mu\nu]}. \tag{12}$$

III. 3+1 DECOMPOSITION OF THE LINEARIZED FIELD EQUATIONS

One considers space–time as composed of three-dimensional spacelike surface and time, embedded in four-dimensional Riemann space. If one treats these spacelike surfaces as the element of a family of surfaces to which one could define non-null normal vectors n^μ such that

$$n_\nu n^\mu = \epsilon = \pm 1, \tag{13}$$

where the components of the normal vectors are denoted by

$$n_\mu = (0,0,0,\epsilon N), \quad n^\mu = \left(-\frac{N^a}{N}, \frac{1}{N} \right), \tag{14}$$

$\mu,\nu,\ldots = 0,1,2,3$ and $a,b,\ldots = 1,2,3$, then the metric tensor of the hypersurface, written in terms of the skew part of the 3-metric h_{ab} takes the form

$$dS^2 = h_{ab}\, dx^a\, dx^b \tag{15}$$

and the skew metric tensor $h_{\mu\nu}$ of space–time is given by

$$dS^2 = h_{\mu\nu}\, dx^\mu\, dx^\nu, \tag{16}$$

the two being related as

$$dS^2 = h_{\mu\nu}\, dx^\mu\, dx^\nu = h_{ab}(dx^a + N^a dx^0)(dx^b + N^b\, dx^0) + \epsilon (N\, dx^0)^2, \tag{17}$$

where N and N^a are, respectively, the lapse function and shift vector, and the quantity x^0 is timelike. In the same way the $3+1$ split of metric tensor leads to:

$$\begin{aligned} h_{\mu\nu} &= {}^3h_{\mu\nu}, \quad h_{0\delta} = N_\delta, \\ h_{\gamma 0} &= N_\gamma, \quad h_{00} = N_\eta N^\eta, \end{aligned} \tag{18}$$

with its inverse as

$$\begin{aligned} h^{\mu\nu} &= h^{ab} + \frac{\epsilon}{N^2} N^a N^b, \quad h^{0\delta} = -\epsilon \frac{N^a}{N^2}, \\ h^{\gamma 0} &= -\epsilon \frac{N^b}{N^2}, \quad h^{00} = -\frac{\epsilon}{N^2}. \end{aligned} \tag{19}$$

Since our primary goal is to express the field equations (12) in $3+1$ decomposed form, we perform the split of the background curvature into components parallel or perpendicular to the normal vector to the hypersurface by making use of the projection tensor as in Appendix A. Thus one obtains the components of the background curvature tensor as

$$R_{cdba} = {}^3R_{cdba} + \epsilon (K_{da} K_{cb} - K_{db} K_{ca}), \tag{20}$$

$$R^\mu_{dab} n_\mu = (K_{db|a} - K_{da|b}), \tag{21}$$

$$R^{\mu}_{d\rho b} n_{\mu} n^{\rho} = \dot{n}_{(d;b)} + K_{da} K^{a}_{b} - \epsilon \dot{n}_{d} \dot{n}_{b} + \& K_{ab} . \tag{22}$$

Equations (20) and (21) are analogous to the Gauss–Codazzi equations. The quantities K_{da}, K_{cb}, etc. denote extrinsic curvatures while & denotes the Lie derivative. By choosing the coordinates such that the $X^a = constant$ lines are normal to the hypersurface, one obtains $N^a = 0$ and Eqs. (20)–(22) simplify to

$$R_{cdab} = {}^{3}R_{cdab} + \epsilon (K_{da} K_{cb} - K_{db} K_{ca}), \tag{23a}$$

$$R^{0}_{dab} = \frac{\epsilon}{N} (K_{db|a} - K_{da|b}), \tag{23b}$$

$$R^{0}_{do\nu} = \frac{\epsilon}{N} (K_{db,o} - N_{,d|b}) + \epsilon K_{cd} K^{c}_{b} . \tag{23c}$$

Equations (23a)–(23c) are then used to rewrite the linearized field equations (12) in $3+1$ decomposed form as in Appendix B, resulting in two sets of field equations:

$${}^{3}R_{cdab} - K_{da} K_{cb} + K_{db} K_{ca} = 0, \tag{24a}$$

$$K_{db,0} - N_{,d|b} + N K_{cd} K^{c}_{b} = 0. \tag{24b}$$

IV. ANALYSIS OF THE DECOMPOSED FIELD EQUATIONS

An examination of the field equations shows the following.

(a) The linearized field equations (24a) contain only the spiral derivatives of the metric tensors and their corresponding extrinsic curvatures and consequently the initial values of, say, h_{cb} and K_{cb} cannot be chosen freely. Thus the equations play the role of constraint equations. These 14 equations are also connected by 4 Bianchi identities, thus reducing the actual number of constraints to 10.

(b) Since the highest time derivative occurs in Eq. (24b) namely $K_{db,0}$ and given that the equation has no curvature terms the antisymmetric components of the fundamental tensor ensure good asymptotic behavior. This contrasts with what had been found by Clayton,[16] who showed by an examination of the exact field equations of MNGT near initially GR field configuration, that arbitrarily small antisymmetric sector fields lead to large accelerations, thereby giving the possibility of linearization instability of MNGT. Our results suggest that the existence of such an instability is suspect on these grounds.

Now since it had been shown[16] that small skew sector data lead to nonsingular evolution for the majority of the possible choices of perturbatively small skew sector initial data, we are of the opinion that our results would be a good starting point for probing the phenomenological results on galaxy dynamics, i.e., whether the collapse of a spherically symmetric matter would be nonsingular.

V. CONCLUSION

It has been shown that the linearized field equations of MNGT have a dynamical part which leads to bounded accelerations, thus giving promising results to the effect that the skew part of the fundamental tensor exhibits good fall-off.

Thus our results effectively strengthen the view that the skew part of the fundamental tensor could be considered just as a perturbation of the symmetric GR background in the linear approximation.

ACKNOWLEDGMENTS

We are grateful to the department of Physics, University of Nairobi and the German Academic Exchange Service, DAAD, for providing research grant in support of this work.

APPENDIX A: 3+1 DECOMPOSITION OF BACKGROUND CURVATURE TENSOR

In order to decompose tensors into components parallel or perpendicular to the normal vector to the hypersurface, one defines a projection tensor $P_{\mu\nu}=h_{\mu\nu}-\epsilon n_\mu P_\nu$ which has the properties[17]

$$P_{\mu\nu}P^{\nu}_{\mu}=P_{\mu\alpha}, \quad P_{\mu\nu}n^{\mu}=0,$$
$$P_{cd}=h_{cd}, \quad P^{cd}=h^{cd}, \tag{A1}$$
$$h^{0}_{\nu}=0, \quad P^{d}_{m}=\delta^{d}_{m}.$$

Since curvature tensors are obtained from covariant derivatives of normal vectors, we may obtain the $3+1$ dimensional split of the background curvature tensor $R_{\alpha\mu\beta\nu}$ by making use of projection tensors as follows:

$$(T_{d|a|b}-T_{c|b|a})P^{d}_{\nu}P^{a}_{\rho}P^{b}_{\lambda}=(T_{\sigma;\eta}P^{\sigma}_{\gamma}P^{\eta}_{\delta});\xi P^{\gamma}_{\nu}(P^{\delta}_{\rho}P^{\xi}_{\lambda}-P^{\delta}_{\lambda}P^{\xi}_{\delta})$$
$$=(T_{\sigma;\eta;\xi}-T_{\sigma;\xi;\eta})P^{\sigma}_{\nu}P^{\eta}_{\rho}P^{\xi}_{\lambda}+T_{\sigma;\eta}(P^{\eta}_{\gamma}P^{\eta}_{\delta});P^{\gamma}_{\nu}(P^{\delta}_{\rho}P^{\xi}_{\lambda}-P^{\delta}_{\lambda}P^{\xi}_{\rho}), \tag{A2}$$

where, say, T_d, etc., are three vectors and $T_{d|a}$ etc refer to 3-dim covariant derivative of the three vectors. From this equation one obtains the relation

$${}^{3}R_{cdab}T^{c}P^{d}_{\nu}P^{a}_{\rho}P^{b}_{\lambda}=R_{c\sigma\eta\xi}T^{c}P^{\sigma}_{\nu}P^{\eta}_{\rho}P^{\xi}_{\lambda}+\epsilon(K_{\lambda\nu}K_{\rho c}-K_{\nu\rho}K_{\lambda c})T^{c}. \tag{A3}$$

Since the equation holds for every vector T^c, one obtains the relation

$$R_{cdab}={}^{3}R_{cdba}+\epsilon(K_{da}K_{cb}-K_{db}K_{ac}), \tag{A4}$$

analogous to the Gauss equation. Now since the covariant derivative of the projection tensor $P_{\mu\nu;\gamma}$ is given by

$$P_{\mu\nu;\gamma}=-(\dot{n}_\mu n_\nu+n_\mu\dot{n}_\nu)n_\gamma+\epsilon(K_{\mu\gamma}n_\nu+K_{\mu\gamma}n_\mu) \tag{A5}$$

and the covariant derivatives of the vector T is given by

$$P^{\gamma}_{\mu}P^{\delta}_{\nu}T_{\gamma;\delta}=P^{c}_{d}P^{d}_{\nu}T_{c|d}, \tag{A6a}$$

$$T_{c|d}=T_{c,d}-{}^{3}\Gamma^{m}_{cd}T_{m} \tag{A6b}$$

we have that

$$(n_{\xi;\sigma;\eta}-n_{\xi;\eta;\sigma})P^{\xi}_{\nu}P^{\sigma}_{\rho}P^{\eta}_{\lambda}=[(n_{\xi;\gamma}P^{\gamma}_{\sigma})_{;\eta}-(n_{\xi;\gamma}P^{\gamma}_{\eta})_{;\sigma}]P^{\xi}_{\nu}P^{\sigma}_{\rho}P^{\eta}_{\lambda}$$
$$=(K_{\xi\eta;\sigma}-K_{\xi\sigma;\eta})P^{\xi}_{\nu}P^{\sigma}_{\rho}P^{\eta}_{\lambda}=(K_{db|a}-K_{da|b})P^{c}_{\nu}P^{a}_{\rho}P^{b}_{\lambda},$$

from which we obtain the expression

$$R^{\mu}_{dab}n_\mu=(K_{db|a}-K_{da|b}). \tag{A7}$$

Analogously from

$$(n_{\xi;\sigma;\eta}-n_{\xi;\eta;\sigma})P^{\xi}_{\nu}n^{\sigma}P^{\eta}_{\lambda}=[(\epsilon\dot{n}_{\xi}n_{\sigma}-K_{\xi\sigma})_{;\eta}-\epsilon(\dot{n}_{\xi}n_{\eta}-K_{\xi\eta})_{;\sigma}]n^{\sigma}P^{\xi}_{\nu}P^{\eta}_{\lambda}$$
$$=\dot{n}_{\xi;\eta}P^{\eta}_{\nu}P^{\eta}_{\lambda}+K_{\xi\sigma}n^{\sigma}_{;\eta}P^{\xi}_{\nu}P^{\eta}_{\lambda}-\epsilon\dot{n}_{\mu}\dot{n}_{\lambda}+K_{\nu\lambda;\sigma}n^{\sigma}-K_{\xi\eta}(P^{\xi}_{\nu}P^{\eta}_{\lambda})_{;\sigma}n^{\sigma},$$

from which it follows that

$$R^{\mu}_{d\rho b}n_{\mu}n^{\rho}=\dot{n}_{(d;b)}+K_{da}K^{a}_{b}-\epsilon\dot{n}_{d}\dot{n}_{b}+\&K_{db}\,, \tag{A8}$$

where & denotes the Lie derivative and K_{db}, etc., are the extrinsic curvature tensors

Now for the special case when the shift vector $N^{\alpha}=0$, Eqs. (A4)–(A8) simplify to

$$R_{cdab}={}^{3}R_{cdab}=\epsilon(K_{da}K_{cb}-K_{db}K_{ca}), \tag{A9a}$$

$$R^{0}_{dab}=\frac{\epsilon}{N}(K_{db|a}-K_{da|b}), \tag{A9b}$$

$$R^{0}_{d0\nu}=\frac{\epsilon}{N}(K_{db,0}-N_{,d|b})+\epsilon K_{cd}K^{c}_{b}\,, \tag{A9c}$$

which are the decomposed components of the background curvature tensor ${}^{0}R_{\alpha\mu\beta\nu}$.

APPENDIX B: DECOMPOSITION OF FIELD EQUATIONS

The decomposed background curvature tensor (A9) may be used to rewrite the linearized field equations (12) in $3+1$ decomposed form as follows:

$$\nabla^{\alpha}F_{\mu\nu\alpha}+m^{2}h_{[\mu\nu]}-4h^{[\alpha\beta]}\,{}^{3}R_{cdab}+4h^{[\alpha\beta]}\epsilon(K_{da}K_{cb}-K_{db}K_{ca})=-2\hat{T}_{[\mu\nu]}\,, \tag{B1}$$

$$\nabla^{\alpha}F_{\mu\nu\alpha}+m^{2}h_{[\mu\nu]}-4h^{[\alpha\beta]}\frac{\epsilon}{N}(K_{db|a}-K_{da|b})=-2\hat{T}_{[\mu\nu]}\,, \tag{B2}$$

$$\nabla^{\alpha}F_{\mu\nu\alpha}+m^{2}h_{[\mu\nu]}-4h^{[\alpha\beta]}\frac{\epsilon}{N}\{(K_{db,0}-N_{,d|b})+K_{cd}K^{c}_{b}\}=-2\hat{T}_{[\mu\nu]}\,. \tag{B3}$$

Now since each of these equations are equal, we may subtract Eq. (B2) from the remaining two and after some rearrangement, we end up with two sets of field equations:

$${}^{3}R_{cdab}-K_{da}K_{cb}+K_{db}K_{ca}=0, \tag{B4}$$

and

$$K_{db,0}-N_{,d|b}+NK_{cd}K^{c}_{b}=0, \tag{B5}$$

which are, respectively, the constraint equations and the evolution equations.

[1] A. Einstein, Ann. Math. **46**, 578 (1945).
[2] A. Einstein and E. G. Strauss, Ann. Math. **47**, 731 (1946).
[3] J. W. Moffat, Phys. Rev. D **19**, 3554 (1959).
[4] J. W. Moffat, Found. Phys. **14**, 1217 (1984).
[5] J. W. Moffat, *Gravitation: A Banff Summer Institute*, edited by R. Mann and P. Wesson (World Scientiffic, Singapore, 1991), p. 523.
[6] G. Kunstatter, H. P. Levio, and P. Savaria, Class. Quantum Grav. **1**, 7 (1984).
[7] P. F. Kelly, Class. Quantum Grav. **8**, 1217 (1991).
[8] P. F. Kelly, Class. Quantum Grav. **9**, 1423 (1992).
[9] T. Damour, S. Deser, and J. McCarthy, Phys. Rev. D **45**, R3289 (1992).
[10] T. Damour, S. Deser, and J. McCarthy, Phys. Rev. D **47**, 1541 (1993).
[11] J. Legare' and J. W. Moffat, Gen. Relativ. Gravit. **27**, 761 (1995).
[12] J. W. Moffat, Phys. Lett. B **335**, 447 (1995).

[13] J. W. Moffat, Phys. Rev. D **36**, 3722 (1995).
[14] M. Kalb and P. Ramond, Phys. Rev. D **9**, 2273 (1974).
[15] M. A. Clayton, J. Math. Phys. **37**, 395 (1996).
[16] M. A. Clayton, Class. Quantum Grav. **13**, 2851 (1996).
[17] H. Stephani, *General Relativity*, 2nd ed. (Cambridge University Press, Cambridge, 1990), p. 160.

Almost-complex and almost-product Einstein manifolds from a variational principle

Andrzej Borowiec,[a] Marco Ferraris, Mauro Francaviglia, and Igor Volovich[b]
Dipartimento di Matematica, Università di Torino, Via C. Alberto 10, 10123 Torino, Italy

(Received 7 July 1997; accepted for publication 2 February 1999)

It is shown that the first-order (Palatini) variational principle for a generic nonlinear metric-affine Lagrangian depending on the (symmetrized) Ricci square invariant leads to an almost-product Einstein structure or to an almost-complex anti-Hermitian Einstein structure on a manifold. It is proved that a real anti-Hermitian metric on a complex manifold satisfies the Kähler condition on the same manifold treated as a real manifold if and only if the metric is the real part of a holomorphic metric. A characterization of anti-Kähler Einstein manifolds and almost-product Einstein manifolds is obtained. Examples of such manifolds are considered. © *1999 American Institute of Physics.* [S0022-2488(99)03107-2]

I. INTRODUCTION

Almost-complex and almost-product structures are among the most important geometrical structures which can be considered on a manifold.[1–4] The aim of this paper is to show that structures of this kind appear in a natural way from a variational principle based on a general class of Lagrangians depending on the Ricci square invariant constructed out of a metric and a symmetric connection; in particular we will show that an anti-Hermitian metric and its special case, the so-called ''anti-Kählerian'' metric, appear naturally from our variational principle. Manifolds with such metrics are much less studied than the familiar Hermitian and Kählerian cases. We hope that our variational principle and ''the universality of the Einstein equations'' (see below) will provide an additional motivation for investigating these manifolds.

Let M be a differentiable manifold of dimension n and $L(M)$ be its frame bundle, a principal fiber bundle over M with group $\mathrm{GL}(n;\mathbb{R})$. Let G be a Lie subgroup of $\mathrm{GL}(n;\mathbb{R})$. A differentiable subbundle Q of $L(M)$ with structure group G is called a *G-structure on M*.[2,4] The classification and integrability of G-structures have been studied in differential geometry; algebraic-topological conditions on M which are necessary for the existence of a G-structure on M can be given in terms of characteristic classes (see, for example, Ref. 5). We also recall that there is a natural one-to-one correspondence between pseudo-Riemannian metrics of signature q on M and $\mathrm{O}(p,q;\mathbb{R})$-structures on M, with $p+q=n$. An $\mathrm{O}(p,q;\mathbb{R})$-structure is integrable if and only if the corresponding pseudo-Riemannian metric has vanishing Riemann curvature. If $G=\mathrm{GL}(p;\mathbb{R})\times\mathrm{GL}(q;\mathbb{R})$, then the G-structure is called an *almost-product structure*;[3,6,7] if $G=\mathrm{O}(r,s;\mathbb{R})\times\mathrm{O}(k,l;\mathbb{R})$, then the G-structure is called a *(pseudo-) Riemannian almost-product structure*.[3,8,9] If n is even, $n=2m$, and one considers $\mathrm{GL}(m;\mathbb{C})$ as a subgroup of $\mathrm{GL}(2m;\mathbb{R})$, then a $\mathrm{GL}(m;\mathbb{C})$-structure is called an *almost-complex structure*;[3,4] if, moreover, one considers $\mathrm{O}(m;\mathbb{C})$ as a subgroup of $\mathrm{GL}(m;\mathbb{C})$, then an $\mathrm{O}(m;\mathbb{C})$-structure defines an *almost-complex anti-Hermitian structure*.[10–14]

We will here use an equivalent description of these G-structures. Let M be a manifold and P be an endomorphism of the tangent bundle TM satisfying $P^2=I$, where I=*identity*. Then P

[a]On leave from the Institute of Theoretical Physics, University of Wrocław, pl. Maksa Borna 9, 50-204 Wroclaw, Poland. Electronic mail: borow@ift.uni.wroc.pl
[b]Permanent address: Steklov Mathematical Institute, Russian Academy of Sciences, Vavilov St. 42, GSP-1, 117966 Moscow, Russia.

defines an almost-product structure on M. If g is a metric on M such that $g(PX,PY)=g(X,Y)$ for arbitrary vectorfields X and Y on M, then the triple (M,g,P) defines a (pseudo-) Riemannian almost-product structure. Geometric properties of (pseudo-) Riemannian almost-product structures have been studied in Refs. 1, 3, 6, 8, 9, and 15–19. If, moreover, g is an Einstein metric [i.e., $\mathrm{Ric}(g)=\gamma g$ holds, where $\mathrm{Ric}(g)$ is the Ricci tensor and γ is a constant], then the triple (M,g,P) shall be called an *almost-product Einstein manifold*.

Analogously, if J is an endomorphism of the tangent bundle TM satisfying $J^2=-I$, then J defines an almost-complex structure on M. An almost-complex structure is integrable if and only if it comes from a complex structure (see Ref. 20). If g is a metric on M such that $g(JX,JY)=-g(X,Y)$ for arbitrary vectorfields X and Y on M, then the triple (M,g,J) defines an almost-complex anti-Hermitian structure. The metric g in this case is called a *Norden metric* and in complex coordinates it has the form

$$ds^2=g_{ab}dz^a dz^b+g_{\bar{a}\bar{b}}d\bar{z}^{\bar{a}}d\bar{z}^{\bar{b}},$$

where $g_{\bar{a}\bar{b}}=\overline{g_{ab}}$. This canonical form differs from the well-known form of a Hermitian metric $ds^2=2g_{a\bar{b}}dz^a d\bar{z}^b$. We will show (Theorem 4.2) that the condition $\nabla J=0$, where ∇ is the Levi–Civita connection, is equivalent in this case to analyticity of the metric: $\partial_{\bar{c}}g_{ab}=0$. Such anti-Hermitian metrics shall be called *anti-Kählerian metrics* since for a Hermitian metric the condition $\nabla J=0$ defines a Kählerian metric.

If g is an Einstein metric, i.e., $\mathrm{Ric}(g)=\gamma g$ holds, then the almost-complex anti-Hermitian manifold (M,g,J) is called an *anti-Hermitian Einstein manifold*. We will consider an important particular class of such manifolds, namely those characterized by *anti-Kählerian Einstein metrics*. Let us stress that we treat *the whole* complex manifold as a real manifold and in this way we get a real Einstein metric with signature (m,m). Another approach to complex Einstein equations, dealing with a *real section* of a complex manifold and aiming to get the Lorentz signature, has been considered in Refs. 21–26.

These G-structures can be conveniently defined as a triple (M,g,K), where g is a metric on M and K is a (1,1) tensor field on M such that $K^2=\epsilon I$ and $g(KX,KY)=\epsilon g(X,Y)$ for arbitrary vectorfields X and Y on M ($\epsilon\neq 0$ is a real constant). We shall also call them *K-structures*. If $\epsilon=1$, then K defines an almost-product structure on M; if $\epsilon=-1$, then K defines an almost-complex structure on M. The more general case $\epsilon>0$ can be reduced to $\epsilon=1$ by a suitable rescaling, while the case $\epsilon<0$ is reduced to $\epsilon=-1$. In any coordinate system one has $K^\mu_\alpha K^\alpha_\nu=\epsilon\delta^\mu_\nu$ and $K^t gK=\epsilon g$, where K^t is the transpose matrix, $\mu,\nu,\alpha=1,2,...,n=\dim(M)$, and δ^μ_ν is the Kronecker symbol.

One can then define a new metric h by the relation $h(X,Y)=g(KX,Y)$, or equivalently $h=gK$, i.e., in local coordinates $h_{\mu\nu}=g_{\mu\alpha}K^\alpha_\nu$. Then the following holds:

$$(g^{-1}h)^2=\epsilon I. \tag{1.1}$$

The relation (1.1) for $\epsilon=+1$ or -1 is equivalent to $(h^{-1}g)^2=\epsilon I$ and there is a one-to-one correspondence between the G-structure (M,g,K) and the G-structure (M,h,K^{-1}). Hence the G-structure given by the triple (M,g,K) can be equivalently described by the triple (M,g,h), where g and h are metrics on M satisfying (1.1). We call such metrics *twin metrics* or *dual metrics*.

In this paper, starting from a manifold M endowed with a metric $h=(h_{\mu\nu})$ and a symmetric linear connection $\Gamma=(\Gamma^\alpha_{\mu\nu})$, we obtain a K-structure as

$$K^\alpha_\mu=h^{\alpha\nu}S_{\mu\nu},$$

where $S_{\mu\nu}\equiv R_{(\mu\nu)}(\Gamma)$ is the symmetric part of the Ricci tensor of the given connection Γ. The general idea is the following. Let us first set $g_{\mu\nu}=R_{(\mu\nu)}(\Gamma)$. According to the results of our earlier paper,[27] one can show that g is, in fact, a new metric and that h and g are ‘‘twin metrics’’ if one assumes a suitable variational principle based on the action

$$A(\Gamma,h)=\int_M f(S)\sqrt{h}\,dx \tag{1.2}$$

and imposes independent variations over the metric and the connection. Here $f(S)$ is a given function of one real variable, which we assume to be analytic, while the scalar $S=S(\Gamma,h)$ is the Ricci square invariant

$$S=h^{\mu\alpha}h^{\nu\beta}S_{\alpha\beta}S_{\mu\nu}.$$

If f is a generic analytic function and $n>2$, one gets either a (pseudo-) Riemannian almost-product structure or an almost-complex structure. In fact, as we shall see below (Theorem 2.4), the Euler–Lagrange equations for (1.2) are generically equivalent to the following system of equations for two metrics $h_{\mu\nu}$ and $g_{\mu\nu}$:

$$(h^{-1}g)^2=\frac{c}{n}I, \tag{1.3}$$

$$\mathrm{Ric}\,(g)=g, \tag{1.4}$$

where the real number c is a root of the equation

$$f'(S)S-\frac{n}{4}f(S)=0. \tag{1.5}$$

As an example, if one takes $f(S)=(nS+c(8-n))^2$, then $S=c$ is a solution of (1.5) if $n\neq 8$ and another (degenerate) solution is $S=c(n-8)/n$. Turning to the general discussion, if $c>0$, then, as it was explained above, solutions of (1.3) and (1.4) are in one-to-one correspondence with almost-product Einstein manifolds (M,g,P), while for $c<0$ one gets anti-Hermitian Einstein manifolds (M,g,J).

Before proceeding further let us explain why the action (1.2) is interesting and important in mathematical physics and especially in the theory of gravity.

As is well known, gravitational Lagrangians which are nonlinear in the scalar curvature of a metric give rise to equations with higher (more than second) derivatives or to the appearance of additional matter fields.[28,29] This strongly depends on having taken a metric as the only basic variable and the equations ensuing from such Lagrangians show an explicit dependence on the Lagrangian itself. An important example of a nonlinear Lagrangian leading to equations with higher derivatives is given by Calabi's variational principle,[30] which shall be discussed in a forthcoming paper.

It was shown in Ref. 31 that, in contrast, working in the first-order (Palatini) formalism, i.e., assuming independent variations with respect to a metric and a symmetric connection, then, for a large class of Lagrangians of the form $f(R)$, where R is the scalar curvature, the equations obtained are almost independent on the Lagrangian, the only such dependence being, in fact, encoded into constants (cosmological and Newton's ones). In this sense the equations obtained are ''universal'' and turn out to be Einstein equations in generic cases. Considering nonlinear gravitational Lagrangians which still generate Einstein equations is particularly important since they provide a simple but general approach to governing topology in dimension two[32] and in view of applications to string theory.[33]

In a previous paper of ours[27] this discussion was extended to the case of Lagrangians with an arbitrary dependence on the square of the symmetrized Ricci tensor of a metric and a (torsionless) connection, finding roughly that the universality of Einstein equations also extends to this class of space–times. In this case, however, new important properties appear: as we have already mentioned above, depending, in fact, on the form of the Lagrangian and on the signature of the metric, one gets an almost-product Einsteinian structure or an almost-complex Einsteinian structure. Topological and geometrical obstructions for the global existence of a solution of the variational problem for this class of Lagrangians will be here considered in Secs. V and VI.

Recently there has been some interest in the problem of signature change in general relativity;[34–36] the nonstandard signature $(10+2)$ has been considered also in superstring theory (F-theory[37]) and extra timelike dimensions in Kaluza–Klein theory are considered in Refs. 38–40. Our results seem, therefore, to show new aspects of this problem which can be relevant also for quantum gravity. For a mathematical consideration of metrics with arbitrary signature which can be relevant to mathematical physics, see, for examples Refs. 41–43.

This paper is organized as follows. In the next section it will be shortly recalled how to get Eqs. (1.3) and (1.4) from the action (1.2). In Sec. III it is shown that Eq. (1.3) can be always solved locally for any given metric g, in particular satisfying (1.4). In Sec. IV we discuss the K-structures. In Sec. V we discuss the problems of the global existence as well as the classification of almost-product Einstein manifolds. In Sec. VI we prove that a real anti-Hermitian metric on a complex manifold satisfies the Kähler condition on the same manifold treated as a real manifold if and only if the metric is the real part of a holomorphic metric on this manifold. Finally, we consider also examples of almost-product Einstein manifolds and anti-Kählerian Einstein manifolds. Theorems of Sec. III are proved in the Appendix.

II. FIELD EQUATIONS

In this section we shall present in a more geometrical form the results of our earlier paper,[27] which form the basis of the further results presented hereafter. Consider, in the first-order (Palatini) formalism, the family of actions

$$A(\Gamma,h)=\int_M f(S)\sqrt{h}\,dx, \tag{2.1}$$

where M is an n-dimensional manifold $(n>2)$ endowed with a metric $h_{\mu\nu}$ and a torsionless (i.e., symmetric) connection $\Gamma^{\sigma}_{\mu\nu}$; the Lagrangian density is $L=f(S)\sqrt{h}$, where $f(S)$ is a given function of one real variable, which we assume to be analytic and $\sqrt{h}$ is a shorthand for $|\det(h_{\mu\nu})|^{1/2}$; and the scalar S is the symmetric part of Ricci square-invariant, considered as a first-order scalar concomitant of a metric and (torsionless) connection, i.e.,

$$S=S(h,\Gamma)=h^{\mu\alpha}h^{\nu\beta}S_{\alpha\beta}S_{\mu\nu} \tag{2.2}$$

being $S_{\mu\nu}=R_{(\mu\nu)}(\Gamma)$ the symmetric part of Ricci tensor, defined according to

$$R^{\lambda}_{\mu\nu\sigma}(\Gamma)=\partial_\nu\Gamma^{\lambda}_{\mu\sigma}-\partial_\sigma\Gamma^{\lambda}_{\mu\nu}+\Gamma^{\lambda}_{\alpha\nu}\Gamma^{\alpha}_{\mu\sigma}-\Gamma^{\lambda}_{\alpha\sigma}\Gamma^{\alpha}_{\mu\nu},$$

$$R_{\mu\sigma}(\Gamma)=R^{\nu}_{\mu\nu\sigma}(\Gamma)\quad(\alpha,\mu,\nu,\ldots=1,\ldots,n).$$

Following Ref. 27, the Euler–Lagrange equations of the action (2.1) with respect to independent variations of h and Γ are

$$f'(S)h^{\alpha\beta}S_{\mu\alpha}S_{\nu\beta}-\tfrac{1}{4}f(S)h_{\mu\nu}=0, \tag{2.3}$$

$$\nabla_\lambda(f'(S)\sqrt{h}h^{\mu\alpha}h^{\nu\beta}S_{\alpha\beta})=0, \tag{2.4}$$

where ∇_λ is the covariant derivative with respect to Γ. Transvecting (2.3) with $h^{\mu\nu}$ tells us that the scalar S has to obey the following real analytic equation,

$$f'(S)S-\frac{n}{4}f(S)=0, \tag{2.5}$$

which allows the description of the general features of the nonlinear system (2.3) and (2.4) and tells us, in turn, that S is generically forced to be a constant. More precisely, it was shown in Ref. 27 that whenever (2.5) admits an (isolated) simple root $S=c$, then the system is ''essentially equivalent'' to Einstein equations for a new metric g with a cosmological constant, in the precise sense which is hereafter described in greater detail. Consider, in fact, any solution

$$S=c \tag{2.6}$$

of Eq. (2.5) and assume that $f'(c)\neq 0$. Then Eq. (2.4) reduces to

$$\nabla_\lambda(\sqrt{h}h^{\mu\alpha}h^{\nu\beta}S_{\alpha\beta})=0, \tag{2.7}$$

while Eq. (2.3) reduces to

$$h^{\alpha\beta}S_{\mu\alpha}S_{\nu\beta}=\epsilon h_{\mu\nu}, \tag{2.8}$$

where a new constant ϵ depending on c arises according to the rule

$$\epsilon=f(c)/4f'(c)=c/n. \tag{2.9}$$

From (2.8) the regularity condition

$$[\det(S_{\mu\nu})]^2=\epsilon^n[\det(h_{\mu\nu})]^2 \tag{2.10}$$

follows, which entails, in particular, that $\det(S_{\mu\nu})\neq 0$ provided $\epsilon\neq 0$. Under this last hypothesis, let $S^{\mu\nu}$ be the inverse matrix of $S_{\mu\nu}$, so that from (2.8) we have

$$h^{\mu\alpha}h^{\nu\beta}S_{\alpha\beta}=\epsilon S^{\mu\nu}. \tag{2.11}$$

By using (2.10) and (2.11), we finally rewrite (2.7) as follows:

$$\nabla_\lambda[\sqrt{|\det(S_{\alpha\beta}(\Gamma))|}S^{\mu\nu}(\Gamma)]=0, \tag{2.12}$$

which will be considered as a new equation in Γ.

Let us recall now the following well-known result, essentially due to Levi-Civita: for $n>2$, any metric g, and any symmetric connection Γ, the general solution of the equation

$$\nabla_\alpha(\sqrt{g}g^{\mu\nu})=0 \tag{2.13}$$

considered as an equation for Γ is the Levi-Civita connection $\Gamma=\Gamma_{\mathrm{LC}}(g)$, i.e., locally

$$\Gamma^\sigma_{\mu\nu}(g)=\tfrac{1}{2}g^{\sigma\alpha}(\partial_\mu g_{\nu\alpha}+\partial_\nu g_{\mu\alpha}-\partial_\alpha g_{\mu\nu}). \tag{2.14}$$

Therefore, the Ricci tensor $R_{\mu\nu}(\Gamma)$ of Γ is automatically symmetric and, in fact, identical to the Ricci tensor $R_{\mu\nu}(g)$ of the metric g itself.

We can then prove the following:

Proposition 2.1: Let us assume that $\det(S_{\alpha\beta})\neq 0$. *Then a connection* Γ *satisfies Eq. (2.12) if and only if there exists a metric* $g_{\mu\nu}$ *such that*

$$R_{\mu\nu}(g)=g_{\mu\nu} \tag{2.15}$$

and $\Gamma=\Gamma_{\mathrm{LC}}(g)$ *is the Levi-Civita connection of* g.

Proof: Let Γ be a connection satisfying Eq. (2.12) and let us set

$$g_{\mu\nu}=S_{\mu\nu}(\Gamma). \tag{2.16}$$

The tensorfield g is a metric due to the condition $\det(S_{\alpha\beta})\neq 0$. Then it follows that Γ has to be the Levi-Civita connection of the metric g; moreover, one has

$$S_{\mu\nu}(\Gamma)=R_{(\mu\nu)}(\Gamma)=R_{\mu\nu}(g), \tag{2.17}$$

so that (2.15) follows from (2.16) and (2.17).

Conversely, let us give a metric $g_{\mu\nu}$ satisfying (2.15) and let us take $\Gamma=\Gamma_{\mathrm{LC}}(g)$. One has again relations (2.17). From (2.17) and (2.15) it follows then that (2.16) holds. Therefore $g^{\mu\nu}=S^{\mu\nu}(\Gamma)$ and $\det(S_{\mu\nu}(\Gamma))=\det(g_{\mu\nu})\neq 0$. Hence we see that Eq. (2.12) reduces to (2.13), which is satisfied since $\Gamma=\Gamma_{\mathrm{LC}}(g)$. Our claim is then proved. (Q.E.D.)

According to the previous discussion, we see that the Euler–Lagrange equations (2.3) and (2.4) are hence equivalent to the following equations for two metrics h and g:

$$h^{\alpha\beta}g_{\mu\alpha}g_{\nu\beta}=\epsilon h_{\mu\nu}, \tag{2.18}$$

$$R_{\mu\nu}(g)=g_{\mu\nu}, \tag{2.19}$$

which are, in fact, nothing but Eqs. (1.3) and (1.4) of the Introduction. The relation between the system (2.18) and (2.19) and the Euler–Lagrange equations in the form (2.3) and (2.4) or (2.11) and (2.12) is given by setting $g_{\mu\nu}=R_{(\mu\nu)}(\Gamma)$.

We will now use the description of (pseudo-) Riemannian almost-product structures and almost-complex structures with a Norden metric in terms of a pair of metrics (twin or dual metrics). Let us consider a triple (M,h,K), where M is a differentiable manifold, h is a metric on M, and K is a (1,1) tensorfield on M such that the following holds:

$$K^2=\epsilon I, \quad K^t hK=\epsilon h.$$

As we said in the Introduction, such a triple defines a (pseudo-)Riemannian almost-product structure if $\epsilon=1$, while it defines an almost-complex structure with a Norden metric h if $\epsilon=-1$. The triple (M,h,K) admits an equivalent description as another triple (M,h,g), where g is a metric on M satisfying the relation $(h^{-1}g)^2=\epsilon I$ or the equivalent relation $(g^{-1}h)^2=cI$, because of the following elementary proposition.

Proposition 2.2: Let K and h be real $n\times n$ matrices and $\epsilon=+1$ or -1. Then the matrices K and h satisfy the relations

$$h^t=h, \quad \det h\neq 0, \quad K^2=\epsilon I, \quad K^t hK=\epsilon h \tag{2.20}$$

if and only if there exists a real matrix g such that g and h satisfy the relations

$$h^t=h, \quad \det h\neq 0, \quad g^t=g, \quad (h^{-1}g)^2=\epsilon I. \tag{2.21}$$

Moreover, one has

$$g=hK \tag{2.22}$$

and

$$(K^{-1})^2=\epsilon l, \quad K^{-1t}gK^{-1}=\epsilon g. \tag{2.23}$$

Proof: If one has (2.20), then define g by (2.22) and check (2.21) and (2.23). Conversely, if one has (2.21), then define $K=h^{-1}g$ and check (2.20). The claim is proved. (Q.E.D.)

We can therefore state the following theorem.

Theorem 2.3: *Let M be a n-dimensional manifold, $n>2$, with a metric h and a symmetric connection Γ and let us consider the Euler–Lagrange equations (2.3) and (2.4) for the action*

(2.1). Let us assume that the analytic function $f(S)$ is such that Eq. (2.5) has an isolated root $S=c$ with $f'(c)\neq 0$. Setting then $g_{\mu\nu}=R_{(\mu\nu)}(\Gamma)$, the Euler–Lagrange equations imply the relations $(h^{-1}g)^2=\epsilon I$, $\mathrm{Ric}\,(g)=g$. Therefore,

(i) if $\epsilon>0$, after rescaling and denoting $P=g^{-1}h$ one gets an almost-product Einstein manifold (M,g,P), i.e.,

$$\mathrm{Ric}\,(g)=\gamma g,$$

$$P^2=1,\quad g(PX,PY)=g(X,Y),\quad X,Y\in\chi(M).$$

(ii) If $c<0$, after rescaling and denoting $J=g^{-1}h$ one gets instead an anti-Hermitian Einstein manifold (M,g,J), i.e.,

$$\mathrm{Ric}\,(g)=\gamma g,$$

$$J^2=-I,\quad g(JX,JY)=-g(X,Y),\quad X,Y\in\chi(M).$$

Here $\chi(M)$ is the Lie algebra of vectorfields on M.

Notice that the signatures of the metrics h and g in the almost-product case can be, in principle, arbitrary, while in the almost-complex case the signature is (m,m). In any case they will be lower-semicontinuous functions, so that without any restriction we can assume that they remain constant in the neighborhood of a generic point; in particular, they will not change in connected components of the manifold. In the next section we will therefore study the equation $(h^{-1}g)^2=\epsilon I$ for a generic point on the manifold, i.e., study it algebraically as a matrix equation.

Remark: Notice that $\epsilon=0$ corresponds to the case $S=0$, which holds iff $f(0)=0$. Then we have to distinguish two subcases $f'(0)=0$ and $f'(0)\neq 0$. In the first subcase both Eqs. (2.3) and (2.4) are automatically satisfied and no condition for g and Γ arise. Therefore, any pair (g,Γ) solves this subcase. When $f'(0)\neq 0$, then (2.3) is instead equivalent to the algebraic equation $[h^{-1}S(\Gamma)]^2=0$, which leads to an *almost-tangent* structure,[44] while (2.4) remains unchanged. We shall not discuss this case in the present paper.

III. SOLUTIONS OF THE MATRIX EQUATIONS

Let us then consider the matrix equation

$$(h^{-1}g)^2=\epsilon\mathbf{I}_n,\tag{3.1}$$

where h and g are symmetric nondegenerate real $n\times n$ matrices, $\mathbf{I}_n$ is the identity matrix in n dimensions, and ϵ is a nonvanishing real number.

In order to solve Eq. (3.1) we first notice that it is manifestly $\mathrm{Gl}(n,\mathbb{R})$-invariant under the canonical right-action $(h,g)\mapsto(A^thA,A^tgA)$, where A^t denotes the matrix transpose to A. More exactly, transforming h and g as metrics one can observe that $P=h^{-1}g$ transforms by a similarity transformation $P\to A^{-1}PA$ [i.e., as a (1,1) tensor]. Equation (3.1) is also invariant under the transformation $(h,g,\epsilon)\mapsto(g,h,\epsilon^{-1})$. Moreover, we have $(\det(g)/\det(h))^2=\epsilon^n$, so that when n is even there are no restrictions on the sign of ϵ, while ϵ has to be positive when n is odd.

It is always possible to rescale g by $\sqrt{|\epsilon|}$ and reduce (3.1) to the canonical form $(h^{-1}g)^2=\pm\mathbf{I}_n$.

When ϵ is positive, Eq. (3.1) admits always the trivial solution $g=\sqrt{\epsilon}h$; however, this does not exhaust all the possible solutions. Let us first observe, in fact, by standard minimal-polynomial arguments (see, e.g., Ref. 45), that the matrix equation

$$P^2=\mathbf{I}_n\tag{3.2}$$

admits only solutions of the form $P=M^{-1}D_kM$ for some nonsingular matrix M, where the matrices D_k (Jordan forms) are diagonal,

$$D_k=\begin{pmatrix} -\mathbf{I}_k & 0 \\ 0 & \mathbf{I}_{n-k} \end{pmatrix}, \tag{3.3}$$

and $k=0,...,n$. This result can be restated as follows: *if an automorphism P of the vector space* $\mathbb{R}^n$ *satisfies (3.2), then there exists a basis in which P is represented by one of the matrices* D_k. The non-negative integer k is an invariant of this authomorphism. In fact, such an automorphism P ($P^2=id$) represents an almost-product structure on $\mathbb{R}^n$. The set of all solutions of the equation $(h^{-1}g)^2=\mathbf{I}_n$ is then described by the following theorem (compare Ref. 45, Th. 4.5.15, case II b):

Theorem 3.1: *Let* $g=g^t$ *and* $h=h^t$ *be two real (symmetric) nondegenerate matrices (metrics). Then the following are equivalent*:

(a) $(h^{-1}g)^2=\mathbf{I}_n$.

(b) *The two metrics h and g are simultaneously diagonalizable with* ± 1 *on the diagonal, i.e., there exists a real nondegenerate matrix R such that*

$$h=R^tD_hR, \quad g=R^tD_gR,$$

and D_h *and* D_g *are diagonal matrices with* $+1$ *or* -1 *on the diagonal*.

The proof of this theorem is given in the Appendix.

Let us proceed to discuss the case $(h^{-1}g)^2=-\mathbf{I}_n$. It is known that if J is any $n\times n$ real matrix satisfying the relation

$$J^2=-\mathbf{I}_n, \tag{3.4}$$

then n must be even, $n=2m$, J can be represented as

$$J=MJ_oM^{-1}$$

where J_o is the canonical form

$$J_o=\begin{pmatrix} 0 & \mathbf{I}_m \\ -\mathbf{I}_m & 0 \end{pmatrix}, \tag{3.5}$$

and M is a nondegenerate real matrix. In fact, one deals with a complex structure and the matrix J_o gives the canonical complex structure on $\mathbb{R}^{2m}$. The following holds true:

Theorem 3.2: *Let* $h=h^t$ *and* $g=g^t$ *be two* $2m\times 2m$ *real (symmetric) nondegenerate matrices (metrics). Then the following are equivalent:*

(a) $(h^{-1}g)^2=-\mathbf{I}_{2m}$.

(b) *There exists a real nondegenerate matrix R such that*

$$h=R^t\begin{pmatrix} \mathbf{I}_m & 0 \\ 0 & -\mathbf{I}_m \end{pmatrix}R, \quad g=R^t\begin{pmatrix} 0 & \mathbf{I}_m \\ \mathbf{I}_m & 0 \end{pmatrix}R,$$

i.e., in the appropriate coordinate system the two metrics g and h take the following canonical forms:

$$K_h=\begin{pmatrix} \mathbf{I}_m & 0 \\ 0 & -\mathbf{I}_m \end{pmatrix}, \quad K_g=\begin{pmatrix} 0 & \mathbf{I}_m \\ \mathbf{I}_m & 0 \end{pmatrix}.$$

This proof is also given in the Appendix.

From Theorems 3.1 and 3.2 it follows that locally for any given metric g one can construct a twin metric h. If g satisfies $\mathrm{Ric}(g)=g$, this means that locally one produces an almost-product or an almost-complex Einsteinian structure.

IV. *K*-STRUCTURES AND KÄHLER-LIKE MANIFOLDS

We first present here a formalism which at once describes properties of various structures important in differential geometry, such as almost-complex and almost-product structures, Hermitian and anti-Hermitian metrics, Kähler manifolds and locally decomposable manifolds, etc., and then, in the next sections, consider in more detail the pseudo-Kählerian and anti-Käherian metrics.

Let M be a smooth manifold, TM its tangent bundle, and $\chi(M)$ the algebra of vectorfields on M. A (K,ϵ)-*structure* (K-*structure* in short) on M is a field of endomorphisms K on TM such that $K^2=\epsilon I$, where $\epsilon=\pm 1$. Thus $\epsilon=1$ corresponds to an almost-product structure, while $\epsilon=-1$ provides an almost-complex structure.

Let $\nabla:\chi(M)\times\chi(M)\to\chi(M)$ be a connection, denoted by $(X,Y)\to\nabla_X Y$. A K-structure is *integrable* iff there exists a linear torsionless connection on M such that $\nabla K=0$, or, equivalently, the Nijenhuis tensor N,

$$N(X,Y)=[KX,KY]-K[KX,Y]-K[X,KY]+\epsilon[X,Y],$$

vanishes. If K is integrable, then there exists an atlas of adapted coordinate charts on M in which K takes a canonical form (see e.g., Ref. 4).

Definition 1: A five-tuple (M,K,g,ϵ,σ) is called a (K,g)-*manifold* if g is a metric on M and K is a K-structure, $K^2=\epsilon I$, such that

$$g(KX,KY)=\sigma g(X,Y) \tag{4.1}$$

for all vectorfields X and Y on M. Here $\sigma=\pm 1$. In this case, we shall say that the metric g is *K-compatible* (or a *K-metric* in short).

The definition above unifies the following four cases: the case $\epsilon=1$, $\sigma=1$ corresponds to a (pseudo-) Riemannian almost-product structure; the case $\epsilon=1$, $\sigma=-1$ provides an almost para-Hermitian structure; the case $\epsilon=-1$, $\sigma=1$ is known as an almost-(pseudo)-Hermitian structure; and, finally, the case $\epsilon=-1$, $\sigma=-1$ corresponds to an almost-complex structure with a Norden metric.

Introduce a (0,2) tensorfield h, the *twin* of g, by

$$h(X,Y)=g(KX,Y). \tag{4.2}$$

Then

$$h(X,Y)=\epsilon\sigma h(Y,X), \quad h(KX,KY)=\sigma h(X,Y). \tag{4.3}$$

Notice that for $\epsilon\sigma=1$ the twin tensor is a metric (and this is, in fact, the case we have obtained from our variational principle), while for $\epsilon\sigma=-1$ the twin tensor is a two-form (and one deals with an almost-Hermitian or almost para-Hermitian structure).

Let ψ be a (0,3) tensorfield defined by the formulas

$$\psi(X,Y,Z)=g((\nabla_X K)Y,Z)\equiv(\nabla_X h)(Y,Z). \tag{4.4}$$

In a coordinate language $\psi_{\alpha\mu\nu}$ is nothing but $\nabla_\alpha h_{\mu\nu}$. It possesses the following properties:

$$\psi(X,Y,Z)=-\sigma\psi(X,KY,KZ)=\sigma\epsilon\psi(X,Z,Y). \tag{4.5}$$

Notice that the classification of almost-Hermitian structures,[46] Riemannian almost-product structures[15] as well as almost-complex structures with a Norden metric[11] is based on algebraic properties of ψ: namely, one decomposes ψ into irreducible components under the action of the appropriate group. The most restrictive class is Kähler-like, when simply $\psi=0$. If the tensorfield ψ

vanishes, then automatically $\nabla K=0$ for a torsionless (Levi-Civita) connection and the Nijenhuis tensor N is forced to vanish, too [Refs. 3 and 7, cf. also formulas (6.7)]. Therefore, the corresponding K-structure is integrable. It leads to the following definition:

Definition 2: A metric g on a (K,g)-manifold is called a *Kähler-like metric* if $\nabla K=0$, i.e.,

$$\nabla_X(KY)=K\nabla_X Y, \quad X,Y\in\chi(M), \tag{4.6}$$

where ∇ is the Levi-Civita connection of g itself.

If $\epsilon=-1$, $\sigma=1$, then a K-metric is called a Kähler metric. If $\epsilon=1$, $\sigma=1$ then a K-metric shall be called a *pseudo-Kählerian metric* [it is also called a (pseudo-) Riemannian locally decomposable metric]. The case $\epsilon=1$, $\sigma=-1$ shall be considered in a forthcoming publication[47] (see also Refs. 48 and 43). Our results on anti-Kählerian manifolds ($\epsilon=-1$, $\sigma=-1$) will be presented in Secs. VI and VII. The following Proposition extends Proposition 3.6 in Ref. 20 to an arbitrary (K,g,ϵ,σ)-structure.

Proposition 4.1: The Riemann curvature $R(X,Y)Z$ and the Ricci tensor $S(X,Y)$ of a Kähler-like manifold (M,K,g) satisfy the following properties:

$$R(X,Y)\circ K=K\circ R(X,Y), \quad R(KX,KY)=\sigma R(X,Y), \tag{4.7}$$

$$S(KX,KY)=\sigma S(X,Y), \quad (\sigma-\epsilon)S(X,Y)=\mathrm{tr}\,[V\mapsto K(R(X,KY)V)]. \tag{4.8}$$

Proof: The proof is a simple repetition of the proof of Proposition 3.6 in Ref. 20, provided one suitably takes into account formulas (4.3). (Q.E.D.)

Consider now the twin F of the Ricci tensor S:

$$F(X,Y)=S(KX,Y). \tag{4.9}$$

Then,

$$F(X,Y)=\epsilon\sigma F(Y,X), \quad F(KX,KY)=\sigma F(X,Y). \tag{4.10}$$

Notice that the symmetry property of F is exactly the same as for h. Therefore, we conclude the following:

Lemma 4.2: A Kähler-like manifold is Einstein iff the dual of S is proportional to the dual of g, i.e., $F(X,Y)\sim h(X,Y)$.

Notice that for $\epsilon\sigma=-1$ both twins F and h are two-forms. This lemma in the Kählerian case leads to a necessary condition on the first Chern class for a manifold to have an Einstein–Kähler metric.[30,49] Recall the Goldberg conjecture[50] (see also Refs. 51 and 52) saying that almost-Kähler Einsteinian manifold is a complex one. It means that an Einstein almost-Hermitian manifold with a closed Kähler form is automatically Hermitian, i.e., its almost-complex structure is integrable. An extension of the Goldberg conjecture to the other Kähler-like manifolds will be discuss elsewhere.[47]

V. ALMOST-PRODUCT EINSTEIN MANIFOLDS

In this section we consider the problems of the global existence and classification of almost-product Einstein manifolds. At the beginning we shall recall some basic facts about almost-product and (pseudo-) Riemannian almost-product structures.

The simplest examples of almost-product structures are product manifolds, i.e., manifolds which are the Cartesian product of two manifolds

$$M=M_1\times M_2. \tag{5.1}$$

In this case the tangent bundle splits as $TM=TM_1\oplus TM_2$ and $P=P_2-P_1$, where P_i are the corresponding projections on TM_i, $i=1,2$. More generally, giving an almost-product structure is equivalent to splitting the tangent bundle into two complementary subbundles (*distributions* or *almost-foliations*): $TM=V\oplus H$; in this case $P=P_V-P_H$. P is integrable iff both the distributions are integrable (i.e., they are *foliations*). Integrable almost-product structures are also called *locally product manifolds*[3] since locally they have the form (5.1). It means that locally (around any point), there exists an adapted coordinate system (x^a,y^α), $a=1,...,k$ and $\alpha=1,...,n-k$, such that the tensorfield P takes the canonical form (3.3), i.e., $P\partial_a=-\partial_a$ and $P\partial_\alpha=\partial_\alpha$.

Similarly, one can consider other structures: for example, a (pseudo-) Riemannian product manifold as a product of two (pseudo-) Riemannian manifolds,

$$(M,g)=(M_1,g_1)\times(M_2,g_2), \tag{5.2}$$

where $g=g_1\oplus g_2$. More generally, an almost-product (pseudo-) Riemannian structure (M,P,g) is integrable iff $\nabla^g P=0$ for the Levi-Civita connection ∇^g of g. In this case we speak of a *locally decomposable* (pseudo-) Riemannian manifold; in the present article we shall, however, propose to call it a *pseudo-Kähler manifold*. For locally decomposable (pseudo-) Riemannian structures both foliations are *totally geodesic*.[3,7] In an adapted coordinate system the metric g ''separates the variables'' (see Refs. 3 and 7)

$$ds^2=g_{ab}(x)dx^a dx^b+g_{\alpha\beta}(y)dy^\alpha dy^\beta. \tag{5.3}$$

The twin metric has the form $h=h_1\ominus h_2$, i.e.,

$$ds_h^2=g_{ab}(x)dx^a dx^b-g_{\alpha\beta}(y)dy^\alpha dy^\beta:$$

Since $\nabla^g_\mu h_{\alpha\beta}\equiv\psi_{\mu\alpha\beta}=0$, then we have in this case $\Gamma(g)=\Gamma(h)$. In fact, this property gives an equivalent definition of pseudo-Kähler manifolds, provided g and h are twin metrics.

Recall that an almost-product Einstein manifold is a triple (M,g,P), where g is a metric and P is a (1,1) tensorfield which satisfy

$$\mathrm{Ric}\,(g)=\gamma g, \tag{5.4}$$

$$P^2=I,\quad g(PX,PY)=g(X,Y),\quad X,Y\epsilon\chi(M). \tag{5.5}$$

Notice that $P=I$ and $P=-I$ give trivial examples of an almost-product structure. Nontrivial examples are given by the following:

Proposition 5.1: Let (M^n,g) *be an Einstein manifold satisfying (5.4) with an indefinite metric* g *of signature* q, $1\leqslant q<n$. *Then there exists on* M^n *a nontrivial almost-product structure* P *satisfying (5.5) and therefore one gets an almost-product Einstein manifold (M,g,P).*

Proof: If g is a pseudo-Riemannian metric on M, then it was proved in Ref. 17 that there exist a (strictly) Riemannian metric h and an almost-product structure P on M such that $g(X,Y)=h(PX,Y)$ and $h(PX,PY)=h(X,Y)$ for all vectorfields X and Y on M. Therefore, one gets the relations (5.5). The almost-product structure P is nontrivial since if $P=\pm I$, then $g=\pm h$; but, in fact, the metric g is pseudo-Riemannian while h is strictly Riemannian. (Q.E.D.)

It follows from the proposition above that any manifold with a strictly pseudo-Riemannian Einstein metric serves as an example of a pseudo-Riemannian almost-product Einsteinian manifold.

It should also be noted that construction of P for a given pseudo-Riemannian metric g is not a canonical one (and, in fact, depends on a choice of some ''background'' Riemannian metric on M). Therefore, a single (possibly Einstein) pseudo-Riemannian metric leads, in principle, to several almost-product (Einsteinian) manifolds. It makes a striking difference between the solutions of our variational problem corresponding to the positive roots of the fundamental equation (2.5) and those corresponding to the negative roots. In the second case, as we shall see in the next section, there are further topological obstructions for the existence of an almost-complex structure.

Let M be a pseudo-Kähler manifold, i.e., (locally) in adapted coordinate systems (x^a,y^α) the metric g splits as $g=g_1\oplus g_2$, where $g_{ab}=g_1(x)$ and $g_{\alpha\beta}=g_2(y)$. If both metrics in (5.3) are Einstein,

$$R_{ab}(g_1)=\gamma g_{ab}, \quad R_{\alpha\beta}(g_2)=\gamma g_{\alpha\beta}, \tag{5.6}$$

for the same constant γ, then it follows

$$R_{AB}(g)=\gamma g_{AB}. \tag{5.7}$$

Therefore, one has the following.

Proposition 5.2: A pseudo-Kählerian manifold is Einstein iff in any adapted coordinates (x^a,y^α) *both metrics are Einslein for the same constant* γ.

Proof: See Refs. 3 and 7. (Q.E.D.)

Interesting examples of locally product (pseudo-) Riemannian manifolds which are not locally decomposable are given by *warped product* space–times.[53–55] Given two (pseudo-) Riemannian manifolds (M_i,g_i), $i=1,2$, and a smooth function θ: $M_1\rightarrow\mathbb{R}$, on the product manifold $M=M_1\times M_2$ put the metric $g=g_1\oplus e^{2\theta}g_2$. The resulting (pseudo-) Riemannian manifold $M=M_1\times_\theta M_2$ is called a *warped product* manifold. It is, of course, an almost-product (pseudo-) Riemannian manifold and it is conformal to a locally decomposable one. It is an interesting and intriguing fact that many exact solutions of Einstein equations (including, e.g., Schwarzschild, Robertson–Walker, Reissner–Nordström, de Sitter, etc.,) and also p-brane solutions (see, e.g., Ref. 56) are, in fact, warped product space–times.[55] Therefore, these exact solutions provide beautiful examples of almost-product Einstein manifolds. Some other examples are provided by the Kaluza-Klein-type theories, $n+m$ decompositions, and more generally so-called *split* structures.[57] The explicit form of the zeta function on product spaces and of the multiplicative anomaly has been derived recently in Ref. 58.

There are topological restrictions on a (paracompact) manifold M for the existence of an almost-product structure of rank k, which are the same as for the existence of a (strictly) pseudo-Riemannian metric of signature $(k,n-k)$, which are again the same as for the existence of a k-dimensional distribution. For example, for the existence of a metric with Lorentz signature on a compact manifold M (i.e., for the existence of a nowhere-vanishing vectorfield), the necessary and sufficient condition is that the Euler characteristic number vanishes.

VI. ANTI-KÄHLERIAN MANIFOLDS

Now we consider in some detail the case of a K-metric with $\epsilon=-1$, $\sigma=-1$, which we call an anti-Kählerian metric.

Definition 3: A triple (M,g,J), where J is an almost-complex structure and the metric g is anti-Hermitian: $g(JX,JY)=-g(X,Y), X,Y\epsilon\chi(M)$ is called an *anti-Kählerian manifold* if $\nabla J=0$, where ∇ is the Levi-Civita connection.

We will prove that a real anti-Hermitian metric on a complex manifold satisfies the Kähler condition $\nabla J=0$ on the same manifold treated as a real manifold if and only if the metric is the real part of a holomorphic metric on this manifold.

Let (M,J) be a $2m$-dimensional almost-complex real manifold and let g be an anti-Hermitian metric on M. We extend J, g, and the Levi-Civita connection ∇ by $\mathbb{C}$-linearity to the complexification of the tangent bundle $T_CM=T_M\otimes\mathbb{C}$. We use the same notation for the complex extended g, J, and ∇. Then the Levi-Civita connection is the mapping $(X,Y)\rightarrow\nabla_XY$, where X and Y are now complex vectorfields (i.e., sections of T_CM). Then the (complex extended) torsion tensor T vanishes,

$$T(X,Y)=\nabla_XY-\nabla_YX-[X,Y]=0,$$

and the ordinary formulas are valid for the connection,

$$\nabla_Xg(Y,Z)=Xg(Y,Z)-g(\nabla_XY,Z)-g(Y,\nabla_XZ)=0, \tag{6.1}$$

and for the Riemann tensor,

$$R(X,Y)Z=[\nabla_X,\nabla_Y]Z-\nabla_{[X,Y]}Z, \tag{6.2}$$

where *X*, *Y*, *Z* are complex vectorfields. For the sake of clarity, we stress that for the moment we are just complexifying the tangent bundle, but we do not assume the almost-complex structure *J* is integrable. Let us now fix a (real) basis $\{X_1,\dots,X_m,JX_1,\dots,JX_m\}$ in each tangent space T_xM. Then, the set $\{Z_a,Z_{\bar a}\}$, where $Z_a=X_a-iJX_a$, $Z_{\bar a}=X_a+iJX_a$, forms a basis for each complexified tangent space $T_xM\otimes\mathbb{C}$. Unless otherwise stated, little Latin indices $a,b,c,\dots$ run from *l* to *m*, while Latin capitals $A,B,C,\dots$ run through $1,\dots,m$, $\bar 1,\dots,\bar m$; for notational convenience we shall also use bar capital indices and we shall assume $\bar{\bar A}=A$. One has $JZ_a=iZ_a$ and $JZ_{\bar a}=-iZ_{\bar a}$. We set $g_{AB}=g(Z_A,Z_B)=g_{BA}$. Then the following holds:

Proposition 6.1: Let (M,J) *be an almost-complex manifold and g be an anti-Hermitian metric on it. Then the complex extended metric g (in the complex basis constructed above) satisfies the following conditions*:

$$g_{a\bar b}=g_{\bar b a}=0, \tag{6.3}$$

$$g_{\bar A\bar B}=\overline{g_{AB}}. \tag{6.4}$$

Proof: Since the metric *g* is anti-Hermitian, we have

$$g(Z_a,Z_{\bar b})=-g(JZ_a,JZ_{\bar b})=-g(iZ_a,-iZ_{\bar b})=-g(Z_a,Z_{\bar b}).$$

Therefore, $g_{a\bar b}=0$, which proves (6.3). The proof of (6.4) is well known. In fact, we have

$$\begin{aligned} g_{\bar a\bar b}=g(Z_{\bar a},Z_{\bar b})&=g(X_a+iJX_a,X_b+iJX_b)\\ &=g(X_a,X_b)-g(JX_a,JX_b)+ig(JX_a,X_b)+ig(X_a,JX_b)\end{aligned}$$

and

$$g_{ab}=g(X_a-iJX_a,X_b-iJX_b)=g(X_a,X_b)-g(JX_a,JX_b)-ig(JX_a,X_b)-ig(X_a,JX_b).$$

Therefore, we get $g_{\bar a\bar b}=\overline{g_{ab}}$. Similarly we consider the other components of g_{AB} and thence we prove (6.4). (Q.E.D.)

It is customary to write a metric satisfying (6.3) and (6.4) as

$$ds^2=g_{ab}dz^a dz^b+g_{\bar a\bar b}dz^{\bar a}dz^{\bar b}. \tag{6.5}$$

We define now the complex Christoffel symbols Γ^C_{AB} as

$$\nabla_{Z_A}Z_B=\Gamma^C_{AB}Z_C. \tag{6.6}$$

It is known[1,3,20] that if $\nabla J=0$, then the torsion *T* and the Nijenhuis tensor *N* satisfy the identity

$$T(JX,JY)=\tfrac{1}{2}N(X,Y) \tag{6.7}$$

for any vectorfields *X* and *Y*. Since the complex extended Levi-Civita connection ∇ has no torsion, the complex Christoffel symbols are symmetric. In this case the complex structure *J* is integrable so that the real manifold *M* inherits the structure of a complex manifold. Let us now recall (see, e.g., Ref. 20) that there is a one-to-one correspondence between complex manifolds and real manifolds with an integrable complex structure. This means that there exist real, adapted (local) coordinates $(x^1,\dots,x^m,y^1,\dots,y^m)$ such that

$$J\left(\frac{\partial}{\partial x^a}\right)=\frac{\partial}{\partial y^a},\quad J\left(\frac{\partial}{\partial y^a}\right)=-\frac{\partial}{\partial x^a}.$$

Setting $z^a = x^a + iy^a$ and taking $X_a = \partial/\partial x^a$ one gets

$$Z_a = X_a - iJX_a = 2\partial/\partial z^a = 2\partial_a, \quad Z_{\bar{a}} = X_a + iJX_a = 2\partial/\partial \overline{z^a} = 2\partial_{\bar{a}},$$

where $\partial_A = \partial/\partial z^A$ and $z^{\bar{a}} = \overline{z^a}$. It appears that z^a's form a complex (analytic) coordinate chart on M. Now from (6.1) one gets

$$\Gamma^C_{AB} = \tfrac{1}{2} g^{CD}(Z_A g_{DB} + Z_B g_{DA} - Z_D g_{AB}) = g^{CD}(\partial_A g_{DB} + \partial_B g_{DA} - \partial_D g_{AB}). \tag{6.8}$$

Notice that the relation (6.1) is valid for the complex extended metric g and complex vector fields X, Y, Z if and only if it is valid for real vectorfields.

Theorem 6.2: *Let M be an m-dimensional complex manifold, thought as a real 2m-dimensional manifold with a complex structure J. Let us further assume that M is provided with an anti-Hermitian metric g. We extend J, g and the Levi-Civita connection* ∇ *by* $\mathbb{C}$*-linearity to the complexified tangent bundle* $T_{\mathbb{C}}M$. *Then the following conditions are equivalent*:

(i)

$$\nabla_X(JY) = J\nabla_X Y, \tag{6.9}$$

where X and Y are arbitrary real vectorfields.

(ii) The (complex) Christoffel symbols satisfy

$$\Gamma^C_{AB} = 0 \quad \textit{except for} \quad \Gamma^c_{ab} \quad \textit{and} \quad \Gamma^{\bar{c}}_{\bar{a}\bar{b}} = \overline{\Gamma^c_{ab}}. \tag{6.10}$$

(iii) There exists a local complex coordinate system $(z^1,\ldots,z^m)$ *on M such that the components of the complex extended metric* g_{ab} *in the canonical form (6.5) are holomorphic functions*

$$\partial_{\bar{c}} g_{ab} = 0. \tag{6.11}$$

Proof: From (6.6) we have

$$\overline{\Gamma^C_{AB}} = \Gamma^{\bar{C}}_{\bar{A}\bar{B}}.$$

The connection satisfies the conditions

$$\nabla_{Z_B}(JZ_c) = J\nabla_{Z_B} Z_c = i\nabla_{Z_B}(Z_c),$$

$$\nabla_{Z_B}(JZ_{\bar{c}}) = J\nabla_{Z_B} Z_{\bar{c}} = -i\nabla_{Z_B}(Z_{\bar{c}}),$$

if and only if

$$\Gamma^a_{B\bar{c}} = \Gamma^{\bar{a}}_{Bc} = 0. \tag{6.12}$$

This proves the equivalence between (*i*) and (*ii*). Then, for the Christoffel symbols (6.8) by taking (6.3) into account one gets

$$\Gamma^a_{b\bar{c}} = g^{aD}(\partial_b g_{D\bar{c}} + \partial_{\bar{c}} g_{Db} - \partial_D g_{D\bar{c}}) = g^{ad}\partial_{\bar{c}} g_{bd}, \tag{6.13}$$

and from (6.12) it follows that

$$\partial_{\bar{c}} g_{bd} = 0. \tag{6.14}$$

The other relations (6.12) also are reduced to (6.14) or its complex conjugated. Therefore, the relation (6.14) is equivalent to (6.11). This proves the equivalence between (*i*) and (*iii*). Our claim is hence proved. (Q.E.D.)

We have proved that a real anti-Hermitian metric on a complex manifold satisfies the Kähler condition $\nabla J=0$ on the same manifold treated as a real manifold if and only if the metric is the real part of a holomorphic metric on this manifold. Therefore, there exists a one-to-one correspondence between anti-Kähler manifolds and complex Riemannian manifolds with a holomorphic metric as they were defined in Ref. 23 (see also, Ref. 13). It should be also remarked that the theory of complex manifolds with holomorphic metric (so-called *complex Riemannian* manifolds) has become one of the cornerstones of the twistor theory.[25] This includes a *nonlinear graviton*,[21] *ambitwistor* formalism,[23] and theory of *H-spaces*.[59]

From an algebraic viewpoint, let us mention that we have been dealing with the following construction. Let V be a real vector space with a complex structure J and let G be a complex-valued bilinear form on V. Let us set

$$F(X,Y)=G(X,Y)-G(JX,JY)-iG(X,JY)-iG(JX,Y).$$

Then we have

$$F(JX,JY)=-F(X,Y).$$

Now one takes the real (or imaginary) part of F to get a real anti-Hermitian bilinear form on V.

VII. ANTI-KÄHLERIAN EINSTEIN MANIFOLDS

In this section we consider the problems of the global existence and classification of anti-Hermitian Einstein manifolds. Recall that an anti-Hermitian Einstein manifold is a triple (M,g,J) where g is a metric and J is a (1,1) tensorfield which satisfy

$$\mathrm{Ric}\,(g)=\gamma g, \tag{7.1}$$

$$J^2=-I, \quad g(JX,JY)=-g(X,Y), \quad X,Y\in\chi(M). \tag{7.2}$$

Then the metric g has necessarily the signature (m,m) (see Sec. III), being $2m=\dim M$. Let us show that by taking the real part of a holomorphic Einstein metric on a complex manifold of complex dimension m one can get a real Einstein manifold of real dimension $2m$.

From (4.8) we have for the Ricci tensor

$$\mathrm{Ric}\,(g)(JX,JY)=-\mathrm{Ric}\,(g)(X,Y), \quad X,Y\epsilon\chi(M).$$

Therefore, analogously to (6.3), we have

$$R_{a\bar{b}}=0. \tag{7.3}$$

We shall not attempt here to consider solutions of Einstein equations for a generic metric of the form (6.5), but consider only the case when g_{ab} is a holomorphic function:

$$\partial_{\bar{c}}g_{ab}=0. \tag{7.4}$$

From (7.4) and (4.7) we get for the Riemann tensor

$$R^D_{ABC}=0 \quad \text{except for } R^d_{abc} \quad \text{and} \quad R^{\bar{d}}_{\bar{a}\bar{b}\bar{c}}=\overline{R^d_{abc}}. \tag{7.5}$$

The (complex) Einstein equations

$$R_{AB}(g)=\gamma g_{AB} \tag{7.6}$$

are hence equivalent to a pair of equations

$$R_{ab}(g_{cd})=\gamma g_{ab}, \tag{7.7a}$$

$$R_{\overline{ab}}(g_{\overline{cd}})=\gamma g_{\overline{ab}}. \tag{7.7b}$$

To get a real solution of Einstein equations (7.1) from (7.7) one uses real coordinates (x^μ), $\mu=1,...,2m$ on M, i.e., $z^a=x^a+ix^{m+a}$, $a=1,...,m$, and writes the metric (6.5) as

$$ds^2=g_{ab}dz^a dz^b+g_{\overline{ab}}d\overline{z^a}d\overline{z^b}=g_{\mu\nu}dx^\mu dx^\nu, \tag{7.8}$$

where $g_{\mu\nu}$ is a real metric. We have hence proved the following theorem:

Theorem 7.1: *If* (M,g,J) *is an anti-Kählerian manifold, i.e., a complex manifold of complex dimension m with a holomorphic metric* $g_{ab}(z),a,b=1,...,m$, *and a real metric* $g_{\mu\nu}(x),\mu,\nu=1,...,2m$, *defined by (7.8), then the holomorphic metric* $g_{ab}(z)$ *satisfies (7.7a) if and only if the real metric* $g_{\mu\nu}(x)$ *is a solution of the Einstein equations (7.1):*

$$R_{\mu\nu}(g)=\gamma g_{\mu\nu}. \tag{7.9}$$

As an example, one can take a complex analytic continuation of any real analytic solution of Einstein equations. A simple example is

$$ds^2=dz^a dz^a+\frac{(z^a dz^a)^2}{1-z^a z^a}+\text{complex conj.}=g_{\mu\nu}dx^\mu dx^\nu. \tag{7.10}$$

This metric $g_{\mu\nu}$ on ‘‘the complex sphere,’’ $w_1^2+\cdots+w_{m+1}^2=1$ (which can be interpreted as a quadric $\zeta_1^2+\cdots+\zeta_{m+1}^2-\zeta_{m+2}^2=0$ in $\mathbb{C}P^{m+1}$ if one takes $w_i=\zeta_i/\zeta_{m+2},$), gives a solution of the Einstein equations (7.8) and provides an example of an anti-Hermitian Einstein manifold (M,g,J).

In particular, for $m=2$ we get a real solution of Einstein equations on the four-dimensional real manifold $(w_1^2+w_2^2+w_3^2=1, w_i\in\mathbb{C})$ with a metric of signature $(++--)$.

Notice also that any Einstein metric on a compact Riemannian manifold M^n leads to an anti-Kählerian Einstein metric on another real manifold $\mathcal{M}^{2n}$. It follows from the known fact[60] that any Einstein metric is analytic in a certain atlas on M^n. Therefore there exists a complex analytic continuation of the metric to a complex manifold of complex dimension n which is a real anti-Kählerian manifold $\mathcal{M}^{2n}$.

ACKNOWLEDGMENTS

The authors wish to thank Zbigniew Olszak for drawing their attention to works on Norden structures. Two of us (AB) and (IV) gratefully acknowledge the hospitality of the Institute of Mathematical Physics ‘‘J.-L. Lagrange’’ of the University of Torino and the support of G.N.F.M. of Italian C.N.R. This work is sponsored by G.N.F.M., M.U.R.S.T. (40% Proj. ‘‘Metodi Geometrici e Probabilistici in Fisica Matematica’’); one of us (AB) acknowledges also the support from KBN 2P302 023 07 and from CONACyT No. 27670 E.

APPENDIX: PROOF OF THEOREMS 3.1 AND 3.2

Proof of Theorem 3.1: The proof $(b)\Rightarrow(a)$ follows obviously from the fact that each two diagonal matrices with ± 1 on their diagonals do satisfy our equation, which is invariant under the appropriate transformation.

The converse $(a)\Rightarrow(b)$ is less obvious. As we already know, there exists a real nondegenerate matrix M such

$$h^{-1}g=MD_kM^{-1}. \tag{A1}$$

From this one gets

$$\tilde{g}=\tilde{h}D_k,$$

where $\tilde{g}=M^tgM$ and $\tilde{h}=M^thM$. Since $\tilde{g}$, $\tilde{h}$, and D_k are symmetric matrices one has hence

$$\tilde{h}D_k=D_k\tilde{h}. \tag{A2}$$

Let us now represent $\tilde{h}$ in block form:

$$\begin{pmatrix} \tilde{h}_{11} & \tilde{h}_{12} \\ \tilde{h}_{12}^t & \tilde{h}_{22} \end{pmatrix},$$

where $\tilde{h}_{11}^t=\tilde{h}_{11}$, $\tilde{h}_{22}^t=\tilde{h}_{22}$, and $\tilde{h}_{11}$ is a $k\times k$ matrix. Then from (A2) we obtain

$$\begin{pmatrix} \tilde{h}_{11} & -\tilde{h}_{12} \\ \tilde{h}_{12}^t & -\tilde{h}_{22} \end{pmatrix}=\begin{pmatrix} \tilde{h}_{11} & \tilde{h}_{12} \\ -\tilde{h}_{12}^t & -\tilde{h}_{22} \end{pmatrix}$$

Therefore $\tilde{h}_{12}=0$ and one gets

$$\tilde{h}=\begin{pmatrix} \tilde{h}_{11} & 0 \\ 0 & \tilde{h}_{22} \end{pmatrix}, \quad \tilde{g}=D_k\tilde{h}=\begin{pmatrix} \tilde{h}_{11} & 0 \\ 0 & -\tilde{h}_{22} \end{pmatrix}. \tag{A3}$$

Now we make use of the fact that any real nondegenerate symmetric matrix is ^{t}-congruent to a diagonal matrix whose diagonal elements are equal to $+1$ or -1 i.e., $h_{11}=S_1^tD_{h_{11}}S_1$ (and analogously $h_{22}=S_2^tD_{h_{22}}S_2$), where $D_{h_{11}}$ (resp. $D_{h_{22}}$) is a diagonal matrix with ±1 along the diagonal. Therefore one has

$$\tilde{h}=S^t\begin{pmatrix} D_{h_{11}} & 0 \\ 0 & D_{h_{22}} \end{pmatrix}S, \quad \tilde{g}=S^t\begin{pmatrix} D_{h_{11}} & 0 \\ 0 & -D_{h_{22}} \end{pmatrix}S,$$

where $S=\left(\begin{smallmatrix} S_t & 0 \\ 0 & S_2 \end{smallmatrix}\right)$. Notice that S commutes with D_k, i.e., $SD_k=D_kS$. Taking then $R=SM$ the theorem is proved. (Q.E.D.)

Proof of Theorem 3.2: To prove $(b)\Rightarrow(a)$, check that $K_h^{-1}K_g=J_o$ and then make use of the appropriate transformation properties.

In order to prove the converse $(a)\Rightarrow(b)$, in full analogy with the previous case one makes use of the fact that $h^{-1}g=MJ_oM^{-1}$. Now this leads to the condition (with the same notation)

$$\tilde{h}J_o=-J_o\tilde{h}. \tag{A4}$$

Writing then $\tilde{h}$ in block form,

$$\tilde{h}=\begin{pmatrix} a & b \\ b^t & d \end{pmatrix},$$

one has from (A4)

$$\begin{pmatrix} -b & a \\ -d & b^t \end{pmatrix}=\begin{pmatrix} b^t & d \\ -a & -b \end{pmatrix}$$

Therefore $a=-d$, $b=b^t$ and we have

$$\tilde{h}=\begin{pmatrix} a & b \\ b & -a \end{pmatrix}, \tag{A5a}$$

$$\tilde{g}=\tilde{h}J_o=\tilde{h}=\begin{pmatrix} -b & a \\ a & b \end{pmatrix}. \tag{A5b}$$

Let us further show that there exists a real matrix S such that

$$SJ_o=J_oS, \tag{A6}$$

$$\tilde{h}=S^tK_hS, \tag{A7a}$$

$$\tilde{g}=S^tK_gS. \tag{A7b}$$

To be sure that (A6) holds, take the matrix S in the form

$$S=\begin{pmatrix} s & u \\ -u & s \end{pmatrix},$$

where s and u are real $m\times m$ matrices to be determined from the conditions (A7a) and (A7b). Equation (A7a) reads as

$$\begin{pmatrix} s^ts-u^tu & s^tu+u^ts \\ u^ts+s^tu & u^tu-s^ts \end{pmatrix}=\begin{pmatrix} a & b \\ b & -a \end{pmatrix}, \tag{A8a}$$

while the equation (A7b) gives

$$\begin{pmatrix} -u^ts-s^tu & s^ts-u^tu \\ s^ts-u^tu & s^tu+u^ts \end{pmatrix}=\begin{pmatrix} -b & a \\ a & b \end{pmatrix}. \tag{A8b}$$

Notice that Eq. (A8a) is equivalent to (A8b) so that we are left with the following conditions:

$$s^ts-u^tu=a, \quad s^tu+u^ts=b.$$

This last equation can be rewritten in the complex form

$$(s+iu)^t(s+iu)=a+ib.$$

Now it is known (see, for example, Ref. 45) that any nondegenerate symmetric complex matrix $a+ib$ can be represented in the form

$$a+ib=N^tN,$$

where N is a (nondegenerate) complex matrix. Taking $s+iu=N$ and $R=SM^{-1}$ the theorem is proved. (Q.E.D.)

[1] A. Besse, *Einslein Manifolds* (Springer-Verlag, Berlin, 1987).
[2] S. S. Chern, Bull. Am. Math. Soc. **72**, 167 (1966).
[3] K. Yano, *Differential geometry on complex and almost complex spaces* (Pergamon, Oxford, 1965).
[4] S. Kobayashi, *Transformation groups in differential geometry* (Springer-Verlag, Berlin, 1972).
[5] D. Sundararaman, *Moduli, deformations and classifications of compact complex manifolds*, Research Notes in Mathematics, Vol. 45 (Pitman, Boston, 1980).
[6] B. L. Reinhart, *Differential geometry of foliations* (Springer-Verlag, Berlin, 1983).
[7] K. Yano and M. Kon, *Structures on Manifolds*, Series in Pure Mathematics, Vol. 3 (World Scientific, Singapore, 1984).
[8] A. G. Walker, Q. J. Math. Oxford **6**(2), 301 (1955); **9**, 221 (1958).
[9] A. Gray, J. Math. Mech. **16**, 715 (1967).

[10] A. P. Norden, Izv. Vyssh. Uchebn. Zaved. Ser. Mat. **4**, 145 (1960).
[11] G. T. Ganchev and A. V. Borisov, C. R. Acad. Bulg. Sci. **39**, 31 (1986).
[12] R. Castro, L. M. Hervella, and E. G. Rio, Riv. Mat. Univ. Parma **15**(4), 133 (1989).
[13] G. Ganchev and S. Ivanov, Riv. Mat. Univ. Parma **1**(5), 155 (1992).
[14] N. Blaźić and N. Bokan, ''Invariant Theory and Affine Differential Geometry,'' in *Differential Geometry and Its Application*, edited by J. Janyśka *et al.* (Masaryk University, Brno, 1996), p. 249.
[15] A. M. Naveira, Rend. Mat. **3**, 577 (1983).
[16] O. Gil-Medrano and A. M. Naveira, Can. Math. Bull. **26**(3), 358 (1983).
[17] F. J. Carreras and V. Miquel, Rend. Circ. Mat. Palermo, Ser. II **35**, 50 (1986).
[18] A. H. Rocamora, Ill. J. Math. **32**(4), 654 (1988).
[19] S. E. Stepanov, Tensor (NS) **55**, 209 (1994).
[20] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Vol. II* (Interscience, New York, 1963).
[21] R. Penrose, Gen. Relativ. Gravit. **7**, 31 (1976).
[22] Yu. I. Manin, *Gauge Field Theory and Complex Geometry* (Springer-Verlag, Berlin, 1988).
[23] C. LeBrun, Trans. Am. Math. Soc. **278**, 209 (1983).
[24] N. J. Hitchin, Lect. Notes Math. **970**, 73 (1982).
[25] E. J. Flaherty, *Hermitian and Kählerian Ceometry in Relativity*, Lecture Notes in Physics, Vol. 46 (Springer-Verlag, Berlin, 1976).
[26] S. G. Gindikin, Lect. Notes Math. **970**, 2 (1982).
[27] A. Borowiec, M. Ferraris, M. Francaviglia, and I. Volovich, Class. Quantum Grav. **15**, 43 (1998).
[28] K. A. Stelle, Gen. Relativ. Gravit. **5**, 353 (1978).
[29] G. Magnano, M. Ferraris, and M. Francaviglia, Gen. Relativ. Gravit. **19**(5), 465 (1987).
[30] E. Calabi, Ann. Math. Studies **102**, 259 (1982).
[31] M. Ferraris, M. Francaviglia, and I. Volovich, Class. Quantum Grav. **11**, 1505 (1994).
[32] M. Ferraris, M. Francaviglia, and I. Volovich, Int. J. Mod. Phys. A **12**, 5067 (1997).
[33] I. Volovich, Mod. Phys. Lett. A **8**, 1827 (1993).
[34] G. F. R. Ellis, Gen. Relativ. Gravit. **24**, 1047 (1992).
[35] S. A. Hayward, Class. Quantum Grav. **9**, 1851 (1992).
[36] T. Dereli and R. W. Tucker, Class. Quantum Grav. **10**, 365 (1993).
[37] C. Vafa, Nucl. Phys. B **469**, 403 (1996).
[38] A. D. Sakharov, Zh. Eksp. Teor. Fiz. **87**, 375 (1984).
[39] I. Ya. Aref'eva and I. V. Volovich, Phys. Lett. B **164**, 287 (1985).
[40] I. Ya. Aref'eva, B. G. Dragović, and I. V. Volovich, Phys. Lett. B **177**, 357 (1986).
[41] S. S. Chern, Ann. Acad. Brasileira Ciencias **35**, 17 (1963).
[42] V. Guillemin and S. Sternberg, Lett. Math. Phys. **12**, 1 (1986).
[43] N. Blaźić, N. Bokan, and Z. Rakić, J. Math. Phys. **39**, 6118 (1998).
[44] H. A. Eliopoulos, Can. Math. Bull. **8**(6), 721 (1965).
[45] R. A. Horn and Ch. A. Johnson, *Matrix Analysis* (Cambridge U.P., London, 1985).
[46] A. Gray and L. M. Hervella, Ann. Mat. Pura Appl. **73**, 35 (1980).
[47] A. Borowiec, M. Francaviglia, and I. Volovich, ''On Para-Kähler Manifolds,'' (in preparation).
[48] E. Reyes, A. Montesinos, and P. M. Gadea, Ann. Polonici Math. **XLVIII**, 307 (1988).
[49] S. T. Yau, Comm. Pure. Appl. Math. **31**, 339 (1978).
[50] S. I. Goldberg, Proc. Am. Math. Soc. **21**, 96 (1969).
[51] Z. Olszak, Bull. Pol. Acad. Sci. Sér. Math. Astronom. Phys. **26**, 139 (1978).
[52] K. Sekigawa and L. Vanhecke, Ann. Mat. Pura Appl. **157**, 149 (1990).
[53] B. O'Neill, *Semi-Riemannian Geometry* (Academic, New York, 1983).
[54] R. Deszcz, L. Verstralen, and L. Vrancken, Gen. Relativ. Gravit. **23**, 671 (1991).
[55] J. Carot and J. da Costa, Class. Quantum Grav. **10**, 461 (1993).
[56] I. Ya. Aref'eva, K. S. Viswanathan, and I. V. Volovich, Phys. Rev. D **55**, 4748 (1997).
[57] V. D. Gladush and R. A. Konoplya, J. Math. Phys. **40**, 955 (1999).
[58] A. A. Bytsenko and F. L. Williams, J. Math. Phys. **39**, 1075 (1998).
[59] C. P. Boyer, J. D. Finley III, and J. F. Plebański, in *General Relativity and Gravitation*, Einstein memorial volume, edited by A. Held (Plenum, New York, 1980), p. 241.
[60] J. P. Bourguignon, Lect. Notes Math. **838**, 42 (1981).

Null particle solutions in three-dimensional (anti-) de Sitter spaces

Rong-Gen Cai[a)]
Center for Theoretical Physics, Seoul National University, Seoul 151-742, Korea

J. B. Griffiths[b)]
Department of Mathematical Sciences, Loughborough University, Loughborough, Leics LE11 3TU, United Kingdom

(Received 18 May 1998; accepted for publication 16 April 1999)

We obtain a class of exact solutions representing null particles moving in three-dimensional (anti-) de Sitter spaces by boosting the corresponding static point source solutions given by Deser and Jackiw. In de Sitter space the resulting solution describes two null particles moving on the (circular) cosmological horizon, while in anti-de Sitter space it describes a single null particle propagating from one side of the universe to the other. We also boost the Bañados–Teitelboim–Zanelli black hole solution to the ultrarelativistic limit and obtain the solution for a spinning null particle moving in anti-de Sitter space. We find that the ultrarelativistic geometry of the black hole is exactly the same as that resulting from boosting the Deser–Jackiw solution when the angular momentum of the hole vanishes. A general class of solutions is also obtained which represents several null particles propagating in the Deser–Jackiw background. The differences between the three-dimensional and four-dimensional cases are also discussed. [S0022-2488(99)03907-9]

I. INTRODUCTION

Although the Einstein equations still hold in three-dimensional space–time, the nature of gravity is quite different from that in four-dimensional space–time. Because the Einstein and Riemann tensors are equivalent in three-dimensional space–time, general relativity is dynamically trivial there. That is, the vacuum space–time is flat. The localized sources have effects only on the global geometry. In 1984, Deser, Jackiw, and 't Hooft[1] investigated in detail the Einstein gravity with static point sources in three-dimensional space–time. For a single static particle, the geometry is given by cutting a sector out of the Euclidean two-plane along two straight lines, and identifying the edges to form a cone. Gravity theories with lightlike sources and spacelike source in three-dimensional flat space–time have also been analyzed in Refs. 2 and 3, respectively.

When a nonvanishing cosmological constant is introduced to the three-dimensional Einstein gravity, some significant changes occur. In this case, the space–time has constant curvature and corresponds either to de Sitter or to anti-de Sitter space. In the de Sitter space, the static two-particle solution is a sphere minus a wedge with the edges identified. This is because the two-space is a sphere in a three-dimensional covering space. To obtain the effect of a point particle one can cut the sphere from the location of the source along two great circles. On a sphere, these cuts meet again at the antipodal point. By identifying along the cuts, this procedure automatically creates a ‘‘mirror’’ source. There is no pure one-particle solution globally. For the anti-de Sitter case, the two-space is a hyperboloid. This can be cut along two lines and the cuts identified to produce single particle solutions. Deser and Jackiw[4] have obtained a metric (hereafter denoted the

[a)]Electronic mail: cai@wormhole.snu.ac.kr
[b)]Electronic mail: J. B. Griffiths@Lboro.ac.uk

DJ solution) and confirmed the above geometrical picture for static point sources by directly solving the Einstein equations with a cosmological constant.

In this paper we investigate the Einstein gravity with null particle sources in the three-dimensional de Sitter and anti-de Sitter spaces. Initially, we employ the boost method that was first used by Aichelburg and Sexl[5] to derive the gravitational field generated by a photon. By boosting the Schwarzschild solution to the ultrarelativistic limit in which the velocity of the source approaches the speed of light and the mass is scaled to zero in an appropriate manner, Aichelburg and Sexl derived a solution describing an impulsive gravitational wave propagating in a flat space–time. This method has subsequently been widely used to investigate the gravitational fields generated by various null sources moving in flat space–times (for a brief review see Ref. 6 and references cited therein). Due to the fact that the four-dimensional (anti-) de Sitter space can be represented as a four-dimensional hyperboloid embedded in a five-dimensional flat space–time, Hotta and Tanaka[7] succeeded in obtaining exact solutions for null particles moving in (anti-) de Sitter space–times by boosting the Schwarzschild–(anti-)de Sitter solutions. The impulsive wave surfaces generated have been discussed in detail by Podolský and Griffiths.[8] Further they considered more general gravitational wave solutions in (anti-) de Sitter spaces.[9] These can be interpreted as impulsive gravitational waves generated by an arbitrary distribution of null particles each with arbitrary multipole structure.

The plan of this paper is as follows. In Sec. II, we will introduce the DJ solution and boost the space–time to the ultrarelativistic limit in the three-dimensional de Sitter and anti-de Sitter spaces and then analyze the resulting geometries. We will also boost the Bañados–Teitelboim–Zanelli (BTZ) black hole[10] in the anti-de Sitter space in Sec. III. Although the BTZ black hole solution is quite different from the DJ solution globally, we find that, when the angular momentum of the BTZ black hole vanishes, the resulting geometries are equivalent to each other. In Sec. IV we will consider the null-particle solution in the DJ background, and further confirm the result derived using the boost method. A brief discussion of the main results is included in Sec. V.

II. BOOSTING THE DJ SOLUTIONS IN THE (ANTI-) DE SITTER SPACES

The three-dimensional Einstein equations with a nonvanishing cosmological constant can be written as

$$R_{\mu\nu}-\tfrac{1}{2}Rg_{\mu\nu}+\Lambda g_{\mu\nu}=8\pi T_{\mu\nu}, \tag{1}$$

where Λ denotes the cosmological constant and $T_{\mu\nu}$ the energy–momentum tensor of the sources. Here the gravitational constant G has been set to one.

The solutions which describe a static point particle at the origin in the (anti-) de Sitter spaces were found to be[4]

$$ds^2=-N^2(R)dt^2+\Phi(R)(dR^2+R^2\,d\phi^2), \tag{2}$$

where

$$\Phi(R)=\frac{4\alpha^2}{\Lambda R^2[(R/R_0)^{\alpha}+(R/R_0)^{-\alpha}]^2},$$
$$N(R)=\frac{(R/R_0)^{\alpha}-(R/R_0)^{-\alpha}}{(R/R_0)^{\alpha}+(R/R_0)^{-\alpha}}, \tag{3}$$

R_0 is an integration constant and $\alpha=1-4M$. The constant M is the mass of the point particle. Performing a simple coordinate transformation in (2) gives

$$ds^2=-(1-\Lambda r^2/\alpha^2)dt^2+\alpha^{-2}(1-\Lambda r^2/\alpha^2)^{-1}\,dr^2+r^2\,d\phi^2. \tag{4}$$

When $\Lambda>0$, the solution (4) has a cosmological event horizon at $r_c=\alpha/\sqrt{\Lambda}$ with surface gravity $\kappa=\sqrt{\Lambda}$. When $\alpha=1$, that is for the vacuum case $M=0$, the DJ solution (4) reduces to the familiar form

$$ds^2=-(1-\Lambda r^2)dt^2+(1-\Lambda r^2)^{-1}\,dr^2+r^2\,d\phi^2, \tag{5}$$

which is just the three-dimensional de Sitter ($\Lambda>0$) or anti-de Sitter ($\Lambda<0$) space in static coordinates.

Similar to the case in four dimensions, the three-dimensional (anti-) de Sitter space can also be represented as a hyperboloid embedded in a four-dimensional flat space–time. Let us first consider the case of the de Sitter space.

(i) *In de Sitter space.* In this case, the hyperboloid satisfies

$$-Z_0^2+Z_1^2+Z_2^2+Z_3^2=a^2, \tag{6}$$

where $a^2=1/\Lambda$. The de Sitter space can be expressed as the following SO(1,3) invariant line element satisfying the constraint (6):

$$ds_{\mathrm{dS}}^2=-dZ_0^2+dZ_1^2+dZ_2^2+dZ_3^2. \tag{7}$$

Obviously, when we parametrize the hypersurface (6) with the following coordinates:

$$\begin{aligned} Z_0&=\sqrt{a^2-r^2}\,\sinh(t/a), \quad Z_1=r\cos\phi,\\ Z_3&=\pm\sqrt{a^2-r^2}\,\cosh(t/a), \quad Z_2=r\sin\phi, \end{aligned} \tag{8}$$

the metric (5) can be deduced from (7). When boosting the DJ solution in the de Sitter space (7), it is appropriate first to expand the solution (4) up to the first order of the mass M (higher order contributions will vanish due to the boost). This yields

$$ds^2\approx ds_{\mathrm{dS}}^2+8M\Lambda r^2\,dt^2+\frac{8M}{(1-\Lambda r^2)^2}\,dr^2, \tag{9}$$

where ds_{dS}^2 denotes the de Sitter space (5). Using the coordinates (8), we can rewrite (9) as

$$ds^2=ds_{\mathrm{dS}}^2+\frac{8M}{(Z_3^2-Z_0^2)^2}\left[(a^2+Z_0^2-Z_3^2)(Z_3dZ_0-Z_0dZ_3)^2+\frac{a^4}{(a^2+Z_0^2+Z_3^2)}(Z_3dZ_3-Z_0dZ_0)^2\right]. \tag{10}$$

We now make a Lorentz boost in the Z_1 direction, that is, a Lorentz transformation

$$Z_0\rightarrow\frac{Z_0+vZ_1}{\sqrt{1-v^2}}, \quad Z_1\rightarrow\frac{vZ_0+Z_1}{\sqrt{1-v^2}}, \quad Z_2\rightarrow Z_2, \quad Z_3\rightarrow Z_3, \tag{11}$$

where v is the boost velocity. To obtain a result of physical interest, the mass must be reduced to zero in an appropriate way. Following Ref. 5, we scale mass as

$$M=p\sqrt{1-v^2}, \tag{12}$$

where p is a constant which can be interpreted as the energy of the null particle. Substituting (11) and (12) into (10), we obtain

$$ds^2=ds_{\mathrm{dS}}^2+\frac{8p\sqrt{1-v^2}}{(Z_3^2-z^2)^2}\left[(a^2+z^2-Z_3^2)(Z_3\,dz-z\,dZ_3)^2+\frac{a^4}{(a^2+z^2-Z_3)}(Z_3dZ_3-z\,dz)^2\right], \tag{13}$$

where $z^2=(Z_0+vZ_1)^2/(1-v^2)$. Using the identity

$$\lim_{v\to 1}\frac{1}{\sqrt{1-v^2}}f(z^2)=\delta(Z_0+Z_1)\int_{-\infty}^{\infty}f(z^2)dz, \tag{14}$$

and taking the limit $v\to 1$ in (13), we obtain

$$ds^2=ds_{\rm dS}^2-8\pi p|Z_2|\delta(Z_0+Z_1)(dZ_0+dZ_1)^2. \tag{15}$$

This looks like an impulsive wave solution in the de Sitter space, located on the surface $Z_0+Z_1=0$, $Z_2^2+Z_3^2=a^2$ which at any time is a circle of constant radius.

In order to further analyze this solution, it proves convenient to use the following coordinates:[8]

$$Z_0=\frac{1}{2\eta}[a^2-\eta^2+(x-a)^2+y^2],\quad Z_1=\frac{a}{\eta}(x-a),$$
$$Z_3=\frac{1}{2\eta}[a^2+\eta^2-(x-a)^2-y^2],\quad Z_2=\frac{a}{\eta}y. \tag{16}$$

Further we can put

$$x=\rho\cos\phi,\quad y=\rho\sin\phi, \tag{17}$$

with $\rho\in[0,\infty)$, $\phi\in[0,2\pi)$. The de Sitter space can then be described as

$$ds_{\rm dS}^2=\frac{a^2}{\eta^2}(-d\eta^2+d\rho^2+\rho^2\,d\phi^2), \tag{18}$$

which is in conformally flat form. The solution (15) can then be rewritten as

$$ds^2=ds_{\rm dS}^2-8\pi pa|\sin\phi|[\delta(\eta-\rho)(d\eta-d\rho)^2+\delta(\eta+\rho)(d\eta+d\rho)^2]. \tag{19}$$

This looks like two impulsive wave fronts. However, as pointed out in Ref. 8 in the four-dimensional case, both components are required for the conformal picture to be geodesically complete. Because $\rho\geqslant 0$, the term $\delta(\eta-\rho)$ does not vanish for $\eta\geqslant 0$ only, while $\delta(\eta+\rho)$ is required for $\eta\leqslant 0$. From (19), it is clear that the particles are located on the circle $\rho=|\eta|$ which is the cosmological horizon of the de Sitter space.

It can then be shown that the energy–momentum tensor is only nonzero at the two points $Z_0+Z_1=0$, $Z_2=0$, $Z_3=\pm a$, which thus represent two null particles. At all other points on this null surface, the impulsive component can in fact be removed by a discontinuous coordinate transformation. The solution can thus be represented as a three-dimensional de Sitter space cut along the cosmological horizon $Z_0+Z_1=0$, with the two halves reattached in such a way as to create two null particles at the points $Z_2=0$, or $y=0$, or $\phi=0,\pi$, which are at opposite points on the horizon. This situation is very like that in the four-dimensional case, in which instead of the circle the wave surface is spherical and the particles are located at opposite poles. The significant difference, however, is that in the four-dimensional case the Weyl tensor has some nonzero components on the spherical surface and these can be interpreted as describing gravitational wave components generated by the null particles. In the three-dimensional theory such free gravitational waves cannot occur.

(ii) *In anti-de Sitter space.* We now turn to the case of the three-dimensional anti-de Sitter space. This can be regarded as a hyperboloid

$$-Z_0^2+Z_1^2+Z_2^2-Z_3^2=-a^2, \tag{20}$$

embedded in an SO(2,2) invariant four-dimensional flat space–time

$$ds^2_{\mathrm{adS}} = -dZ_0^2 + dZ_1^2 + dZ_2^2 - dZ_3^2, \tag{21}$$

where $a^2 = -1/\Lambda > 0$. Obviously, the anti-de Sitter space (5) can be parametrized by the following coordinates:

$$\begin{aligned} Z_0 &= \sqrt{a^2+r^2}\,\sin(t/a), \quad Z_1 = r\cos\phi, \\ Z_3 &= \sqrt{a^2+r^2}\,\cos(t/a), \quad Z_2 = r\sin\phi. \end{aligned} \tag{22}$$

We now boost the DJ solution in the anti-de Sitter space. Again expanding the solution up to the first order in the mass M and using the coordinates (22), we arrive at

$$\begin{aligned} ds^2 = ds^2_{\mathrm{adS}} + \frac{8M}{(Z_0^2+Z_3^2)^2}\bigg[&(Z_0^2+Z_3^2-a^2)(Z_3 dZ_0 - Z_0 dZ_3)^2 \\ &+ \frac{a^4}{(Z_0^2+Z_3^2-a^2)}(Z_3 dZ_3 - Z_0 dZ_0)^2 \bigg]. \end{aligned} \tag{23}$$

Repeating the same steps as in the case of the de Sitter space, that is, using the Lorentz transformation (11), rescaling the mass as (12), and taking the limit $v \to 1$, finally we can obtain

$$ds^2 = ds^2_{\mathrm{adS}} - 8\pi p|Z_2|\,\delta(Z_0+Z_1)(dZ_0+dZ_1)^2. \tag{24}$$

Comparing with (15), it is easy to see that the expression for the apparent impulsive part is the same as that in the de Sitter space. However, the interpretation is quite different. In this case the impulsive component is given by the surface $Z_0+Z_1=0$, $Z_2^2-Z_3^2=-a^2$, which at any time is a hyperbola. Let us now analyze this solution.

In the anti-de Sitter space, introduce first the following coordinates:

$$\begin{aligned} Z_0 &= \frac{1}{2x}[a^2-\eta^2+x^2+(y-a)^2], \quad Z_1 = \frac{a}{x}(y-a), \\ Z_2 &= \frac{1}{2x}[a^2+\eta^2-x^2-(y-a)^2], \quad Z_3 = \frac{a}{x}\,\eta, \end{aligned} \tag{25}$$

and then use

$$x = \rho\cos\phi, \quad y = \rho\sin\phi. \tag{26}$$

This produces the anti-de Sitter space written in the conformally flat form

$$ds^2_{\mathrm{adS}} = \frac{a^2}{\rho^2\cos^2\phi}[-d\eta^2 + d\rho^2 + \rho^2\,d\phi^2]. \tag{27}$$

The solution (24) can be rewritten in these coordinates as

$$ds^2 = ds^2_{\mathrm{adS}} - \frac{8\pi pa|\sin\phi|}{\cos^2\phi}[\delta(\eta-\rho)(d\eta-d\rho)^2 + \delta(\eta+\rho)(d\eta+d\rho)^2]. \tag{28}$$

Here it should be stressed that the solution (28) does not mean two impulses again. As in the de Sitter case, $\delta(\eta-\rho)$ works only for $\eta>0$ while $\delta(\eta+\rho)$ for $\eta<0$. The two components are required for globally geodesic completeness. From (28), it is clear that the impulsive component is located on the line $\rho=|\eta|$, or $x^2+y^2=\eta^2$. However, this is not a circle—according to (27) it is conformal to a circle, and the coordinate ϕ is restricted to $-\pi/2<\phi<\pi/2$. In fact it is a hyperbola. It can then be shown that this solution represents a single null particle located at $Z_2=0$ on the null surface $Z_0+Z_1=0$, i.e., at $y=0$, $x=\eta$ (or $x=-\eta$). The particle thus clearly propagates from one side of the universe to the other and (since this space–time contains closed timelike lines) may then be considered to propagate back in the opposite direction.

Thus, by boosting the DJ solutions, we have obtained two kinds of exact solutions describing null particles moving in the three-dimensional de Sitter and anti-de Sitter spaces. (In Sec. IV we will further confirm these results by directly solving the Einstein equations.) Although static particles only have effect on the global geometry, which is quite different from the situation in the four-dimensional case, we still find that the boost method is sufficiently powerful to derive null particle solutions from their corresponding static particle solutions. In Sec. III we will boost the BTZ black hole solution in the anti-de Sitter space. In the static situation, this is quite different from the DJ solution from the aspect of global properties. However, the resulting ultrarelativistic geometry is found to be identical, at least in the nonrotating case.

III. BOOSTING THE BTZ BLACK HOLE SOLUTION IN THE ANTI-DE SITTER SPACE

Due to the special properties of three-dimensional gravity, it was a surprising discovery when Bañados, Teitelboim, and Zanelli[10] claimed that they found a black hole solution in the Einstein gravity with a negative cosmological constant. The solution they found is

$$ds^2=-N^2(r)dt^2+N^{-2}\,dr^2+r^2(N^\phi(r)dt+d\phi)^2, \tag{29}$$

where

$$N^2=-8M+\frac{r^2}{a^2}+\frac{16J^2}{r^2},\quad N^\phi=-\frac{4J}{r^2}. \tag{30}$$

Here $-1/a^2$ denotes the negative cosmological constant. The integration constants M and J can be interpreted as the mass and angular momentum of the black hole. This black hole has two horizons at

$$r_\pm^2=4Ma^2\left[1\pm\sqrt{1-\left(\frac{J}{Ma}\right)^2}\right], \tag{31}$$

provided $M>0$ and $J<Ma$. This solution is asymptotically an anti-de Sitter space–time and can be constructed by identifying some discrete points in the three-dimensional anti-de Sitter space. It is of interest to note, however, that when the mass and angular momentum of the hole vanish, the solution does not reduce to the anti-de Sitter space. Rather, the anti-de Sitter space–time (5) can only be obtained from (29) in the limit as $8M\to-1$ and $J\to 0$.

Before boosting the BTZ solution, it is first appropriate to expand it about the background of the anti-de Sitter space. To achieve this, we expand the BTZ solution (29) to first order in the mass term $8M+1$ and the angular momentum J. The result is

$$ds^2\approx ds^2_{\mathrm{adS}}+(8M+1)dt^2+\frac{8M+1}{(1+r^2/a^2)^2}\,dr^2-8J\,dt\,d\phi. \tag{32}$$

Using the coordinates in (22), the above metric can be rewritten as

$$\begin{aligned}ds^2=ds^2_{\mathrm{adS}}&+\frac{(8M+1)a^2}{(Z_0^2+Z_3^2)^2}\left[(Z_3dZ_0-Z_0dZ_3)^2+\frac{a^2}{Z_3^2+Z_0^2-a^2}(Z_3dZ_3+Z_0dZ_0)^2\right]\\&-\frac{8Ja}{(Z_3^2+Z_0^2)(Z_3^2+Z_0^2-a^2)}(Z_3dZ_0-Z_0dZ_3)(Z_1dZ_2-Z_2dZ_1).\end{aligned} \tag{33}$$

We now make a Lorentz boost (11) in the Z_1 direction, rescaling the mass and angular momentum as

$$8M+1=8p\sqrt{1-v^2}, \quad J=s\sqrt{1-v^2}. \tag{34}$$

We then proceed to the ultrarelativistic limit $v\to 1$. In this case, the two constants p and s can be interpreted physically as the energy and spin angular momentum of the resulting null particle, respectively. It may be observed that, in this limit, the inequality $J<Ma$ mentioned above is strictly violated. This is because the mass M and angular momentum J are rescaled in different ways. However, the limit is still an exact solution even though it is not strictly the limit of a real rotating black hole. Using this procedure in (33), we obtain

$$\begin{aligned} ds^2 = {} & ds^2_{\text{adS}} + 8\pi p(Z_3-\sqrt{Z_3-a^2})\,\delta(Z_0+Z_1)(dZ_0+dZ_1)^2 \\ & -\frac{8\pi s}{a}\left[Z_2-\frac{Z_2Z_3}{\sqrt{Z_3^2-a^2}}\right]\delta(Z_0+Z_1)(dZ_0+dZ_1)^2. \end{aligned} \tag{35}$$

The two linear terms in the solution (35) can be removed by the following discontinuous linear transformation:

$$\begin{aligned} & Z_2\to Z_2-\frac{4\pi s}{a}U\Theta(U), \\ & Z_3\to Z_3-4\pi pU\Theta(U), \\ & U\to U, \\ & V\to V-16\pi^2p^2U\Theta(U)+8\pi pZ_3\Theta(U)+\frac{16\pi^2s^2}{a^2}U\Theta(U)-\frac{8\pi s}{a}Z_2\Theta(U), \end{aligned} \tag{36}$$

where $U=Z_0+Z_1$, $V=Z_0-Z_1$, and Θ is the Heaviside step function. Therefore, the solution (35) can be reduced to

$$ds^2=ds^2_{\text{adS}}-8\pi p|Z_2|\,\delta(Z_0+Z_1)(dZ_0+dZ_1)^2+\frac{8\pi sZ_3}{a}\operatorname{sign}(Z_2)\,\delta(Z_0+Z_1)(dZ_0+dZ_1)^2. \tag{37}$$

It is now easy to see that when $s=0$, that is when the angular momentum vanishes in the original BTZ solution, the solution (37) is identical to (24). Thus, both ultrarelativistic limits of the DJ solution for $\Lambda<0$ and the spinless BTZ solution are equivalent to each other. Obviously, the third term in (37) is the spin effect of the null particle. In the coordinates (25) and (26), we can rewrite (37) as

$$\begin{aligned} ds^2 &= ds^2_{\text{adS}}+\left(-8\pi pa\frac{|\sin\phi|}{\cos^2\phi}+\frac{8\pi s\ \operatorname{sign}(\tan\phi)}{\cos\phi|\cos\phi|}\right)[\delta(\eta-\rho)(d\eta-d\rho)^2+\delta(\eta+\rho)(d\eta+d\rho)^2] \\ &= ds^2_{\text{adS}}+\left(-8\pi pa\frac{|\sin\phi|}{\cos^2\phi}+\frac{8\pi s}{\cos^2\phi}\operatorname{sign}(\sin\phi)\right)[\delta(\eta-\rho)(d\eta-d\rho)^2+\delta(\eta+\rho)(d\eta+d\rho)^2] \end{aligned} \tag{38}$$

which is clearly identical to (28) when $s=0$.

IV. NULL PARTICLES IN THE DJ BACKGROUND

In this section we will directly solve the Einstein equations with null particle sources and re-obtain some of the results of previous sections that were derived by the boost method. In Ref. 11 Dray and 't Hooft considered a particle moving with the speed of light on the Schwarzschild black hole horizon, and investigated the back reaction of the particle on the geometry. In this case, the particle produces an impulsive gravitational wave located on the Schwarzschild horizon. Loustó and Sánchez[12] and Sfetsos[13] further extended the work of Dray and 't Hooft to nonvacuum backgrounds and investigated the conditions that should be satisfied when an impulsive wave is introduced into curved space–times. The null particle solution in the BTZ background (29) has already been considered in Ref. 13.

Here we first note that the DJ solution (4) can be rewritten, after rescaling the coordinate r, as

$$ds^2 = -(1-\Lambda r^2)dt^2 + (1-\Lambda r^2)^{-1}\,dr^2 + \alpha^2 r^2\,d\phi^2. \tag{39}$$

Further defining $\phi' = \alpha\phi$ with $\phi' \in [0,2\pi\alpha)$, we have

$$ds^2 = -(1-\Lambda r^2)dt^2 + (1-\Lambda r^2)^{-1}\,dr^2 + r^2\,d\phi'^2, \tag{40}$$

which is obviously equivalent to the de Sitter space locally, but has a deficit angle $\delta = (1-\alpha)2\pi$. We will first discuss solutions with null particles located on the cosmological horizon of the de Sitter case $(\Lambda = 1/a^2 > 0)$. The metric (39) or (40) also can be regarded as a hypersurface embedded in the flat space–time (7) with

$$\begin{aligned} Z_0 &= \sqrt{a^2-r^2}\,\sinh(t/a), \quad Z_1 = r\cos(\alpha\phi), \\ Z_3 &= \pm\sqrt{a^2-r^2}\,\cosh(t/a), \quad Z_2 = r\sin(\alpha\phi). \end{aligned} \tag{41}$$

Introducing the null coordinates

$$u = e^{t/a}F(r), \quad v = e^{-t/a}F(r), \tag{42}$$

where the function $F(r)$ is defined as

$$F(r) \equiv \exp\left(-\frac{1}{a}\int \frac{dr}{(1-r^2/a^2)}\right) = \left(\frac{a-r}{a+r}\right)^{1/2}, \tag{43}$$

we can reexpress the DJ solution (40) as

$$ds^2 = 2A(u,v)du\,dv + r^2(u,v)d\phi'^2, \tag{44}$$

where

$$A(u,v) = \frac{(a^2-r^2)}{2F^2(r)}, \quad r(u,v) = \frac{a(1-uv)}{1+uv}. \tag{45}$$

We can now consider the effect of null particles located on the null surface $u=0$ which is clearly the cosmological horizon $r=a$. In this three-dimensional theory, this horizon is circular. Following Refs. 11 and 13, we can adopt the coordinate shift method on the background (40). That is, the following ansatz is employed: For $u<0$ the space–time is still the background (44) while for $u>0$ the space–time is (44) with v shifted as $v \to v + f(\phi')$. The function $f(\phi')$ which will be determined later describes the effect of the sources. Using this approach, the new solution has the form

$$ds^2 = 2A(u,v)du\,dv - 2A(u,v)f(\phi')\delta(u)du^2 + r^2(u,v)d\phi'^2, \tag{46}$$

which comes from (44) after making the coordinate shift:

$$u \to u, \quad v \to v - f(\phi')\Theta(u), \quad \phi' \to \phi'. \tag{47}$$

For this solution to be consistent with the Einstein equations, the following conditions must be satisfied at $u=0$:[13]

$$A_{,v} = r^2_{,v} = T_{vv} = 0, \tag{48}$$

$$\frac{d^2 f}{d\phi'^2} - \frac{r^2_{,uv}}{2A} f = \frac{8\pi r^2}{A} \tilde{T}_{uu}, \tag{49}$$

where T is the energy–momentum tensor of matter generating the DJ geometry, that is, the cosmological constant and possibly some static point particles, and $\tilde{T}$ is the energy–momentum tensor of any null particles located on the surface. Here it should be noticed that the only non-vanishing component of the energy–momentum tensor for null particles is $\tilde{T}_{uu}$, and that this is zero everywhere except at the points where the particles are located.

At the $u=0$ null surface—that is, on the cosmological horizon $r_c = a = 1/\sqrt{\Lambda}$ for the DJ $(\Lambda > 0)$ solution (40)—it is easy to see that the conditions (48) are satisfied and

$$A(u,v)|_{u=0} = 2a^2, \quad r^2_{,uv}|_{u=0} = -4a^2. \tag{50}$$

Then (49) reduces to

$$\frac{d^2 f}{d\phi'^2} + f = 4\pi \tilde{T}_{uu}. \tag{51}$$

It may immediately be observed that a solution with $f = 4\pi\rho$, where ρ is a constant, represents a uniform distribution of null matter (of density ρ) over the circular horizon. Since Eq. (51) is linear, this component can always be added to other components. However, we will ignore this possibility in the remainder of this section.

In those parts of the null surface on which $\tilde{T}_{uu} = 0$, Eq. (51) has the solution

$$f = c \sin(\phi' + \omega') = c \sin \alpha(\phi + \omega), \tag{52}$$

where c and $\omega' = \alpha\omega$ are arbitrary constants. This solution for f around the circular horizon can always be removed by a discontinuous coordinate transformation. However, solutions describing several discrete particles can be constructed by patching different sections of the sine wave, each with different amplitude and phase. Points at which f is C^0 but has a discontinuous first derivative can be interpreted as points at which null particles are located. The energy of each particle is then represented by the jump in the derivative of f, and the energy–momentum tensor $\tilde{T}_{uu}$ is given by a δ function. On considering (38), it may be observed that discontinuities in f may also be permitted. These represent point particles with spin and, in this case, the energy–momentum tensor $\tilde{T}_{uu}$ contains a derivative of a δ function.

For example, consider the case in which n particles each of energy p_i, $i = 1 \cdots n$, are located at points $\phi = \phi_i$ around the circular wave. The solution is then given by

$$f = c_i \sin \alpha(\phi + \omega_i) \quad \text{for} \quad \phi_{i-1} \leqslant \phi \leqslant \phi_i, \tag{53}$$

where $n+1 \to 1$. It is then possible to choose the $2n$ arbitrary constant c_i and ω_i such that

$$c_{i+1} \sin \alpha(\phi_i + \omega_{i+1}) - c_i \sin \alpha(\phi_i + \omega_i) = 0,$$

$$c_{i+1}\cos\alpha(\phi_i+\omega_{i+1})-c_i\cos\alpha(\phi_i+\omega_i)=\frac{4\pi p_i}{a}. \tag{54}$$

By choosing the constants appropriately, it is possible to construct solutions in which $n(\geqslant 2)$ null particles of arbitrary energy are distributed arbitrarily around the circular wave.

In particular, we can consider the two-particle solution in which the particles are located at opposite ends of a diameter of the circle. Since $0\leqslant\phi'<2\pi\alpha$ around the circle, we may consider the particles to be located at points given by $\phi'=0$ and $\phi'=\pi\alpha$. We may also restrict attention to the case in which the two particles have identical energy p. Such a solution can be constructed by the above method in which

$$c_1=c_2=\frac{2\pi p}{a}\operatorname{cosec}\frac{\pi\alpha}{2},\quad \omega_1=\frac{(1-\alpha)\pi}{2\alpha},\quad \omega_2=\frac{(1-3\alpha)\pi}{2\alpha}. \tag{55}$$

This solution represents two null particles propagating in the DJ $(\Lambda>0)$ background. In the case in which $\alpha=1$, the background is the de Sitter space and the solution can alternatively be written in the form

$$f=\frac{2\pi p}{a}|\sin\phi|. \tag{56}$$

This is clearly identical (after some rescaling) to the solution (19) that was obtained by boosting two static particles in the de Sitter background, and thus confirms this solution.

It may also be observed that, in the particular case in which $\alpha=1/2$, the deficit angle in ϕ' is π and a one-particle solution at $\phi=0$ can easily be constructed using

$$f=c\sin(\phi/2),$$

where

$$0\leqslant\phi\leqslant 2\pi. \tag{57}$$

It is also possible to obtain solutions for null particles propagating in the DJ background with $\Lambda<0$. However, the Dray–'t Hooft[11] method cannot be directly used in this case. Nevertheless, equivalent equations can be obtained and these will include, for $\alpha=1$, the special cases (28) of a null particle and (38) of a spinning null particle propagating in an anti-de Sitter space.

V. CONCLUSION AND DISCUSSION

We have investigated null particle solutions in the three-dimensional de Sitter and anti-de Sitter spaces by boosting the corresponding static point source solutions (DJ solutions)[4] in the (anti-) de Sitter backgrounds. For the de Sitter case, the resulting solution describes two null particles located at opposite points on the cosmological horizon which forms a circle of constant size. For the anti-de Sitter case, the solution describes a single null particle located at the point of symmetry of a propagating hyperbola. We have also boosted the BTZ black hole solution to the ultrarelativistic limit. Although the BTZ black hole is quite different from the DJ solution globally, we have found that these two ultrarelativistic limits are equivalent to each other when the angular momentum of the hole is zero. This means that the boost method may lose some memory of the original solution in the process of the boost. In addition, we believe that the angular momentum of the hole gives the spin effect of the corresponding null particle.

By using the coordinate shift method, we have also obtained null particle solutions in DJ background. When $\alpha=1$, the DJ $(\Lambda>0)$ solution reduces to the de Sitter space, and the results obtained include that derived using the boost method. It may be observed that the boost method indeed is very powerful in the derivation of null particle solutions, not only in flat space–time, but also in the (anti-) de Sitter space in three dimensions as well as four.

Due to the special properties of the geometry in three-dimensional space–time, the nature of Einstein gravity is rather different from that in four dimensions. The static point source solutions, given in Ref. 1 in flat space–time and in Ref. 4 in (anti-) de Sitter space, clearly demonstrate the differences in the local and global aspects from the four-dimensional Schwarzschild and Schwarzschild-(anti-)de Sitter solutions. By comparing the null particle solution given by Deser and Steif[2] in three-dimensional flat space–time and some results given in this paper, we may observe some similarities as well as some differences in the null particle solutions in three and four dimensions. The main difference in four-dimensional space–time is that the null particles generate impulsive gravitational waves which are forbidden in three-dimensional theories. It is of some interest to further compare spacelike source solutions in three- and four-dimensional space–times. Furthermore, it also might be interesting to discuss the geodesics and particle scattering in the null particle solutions in the (anti-) de Sitter space.

[1] S. Deser, R. Jackiw, and G. 't Hooft, Ann. Phys. (N.Y.) **152**, 220 (1984).
[2] S. Deser and A. R. Steif, Class. Quantum Grav. **9**, L153 (1992).
[3] S. Deser, R. Jackiw, and G. 't Hooft, Phys. Rev. Lett. **68**, 267 (1992).
[4] S. Deser and R. Jackiw, Ann. Phys. (N.Y.) **153**, 405 (1984).
[5] P. C. Aichelburg and R. U. Sexl, Gen. Relativ. Gravit. **2**, 303 (1971).
[6] R. G. Cai, J. Y. Ji, and K. S. Soh, Nucl. Phys. B **528**, 265 (1998).
[7] M. Hotta and M. Tanaka, Class. Quantum Grav. **10**, 307 (1993).
[8] J. Podolský and J. B. Griffiths, Phys. Rev. D **56**, 4756 (1997).
[9] J. Podolský and J. B. Griffiths, Class. Quantum Grav. **15**, 453 (1998).
[10] M. Bañados, C. Teitelboim, and J. Zanelli, Phys. Rev. Lett. **69**, 1849 (1992); M. Bañados, M. Henneaux, C. Teitelboim, and J. Zanelli, Phys. Rev. D **48**, 1506 (1993).
[11] T. Dray and G. 't Hooft, Nucl. Phys. B **253**, 173 (1985).
[12] C. O. Loustó and N. Sánchez, Phys. Lett. B **220**, 55 (1989).
[13] K. Sfetsos, Nucl. Phys. B **436**, 721 (1995).

Canonical general relativity: Matter fields in a general linear frame

M. A. Clayton[a)]
9 Hurndale Ave., Toronto, Ontario M4K 1R6, Canada

(Received 18 November 1998; accepted for publication 29 January 1999)

Building on the results of previous work [M. A. Clayton, ''Canonical general relativity: Diffeomorphism constraints and spatial frame transformations,'' J. Math. Phys. **39**, 3805–3816 (1998)], we demonstrate how matter fields are incorporated into the general linear frame approach to general relativity. When considering the Maxwell one-form field, we find that the system that leads naturally to canonical vierbein general relativity has the extrinsic curvature of the Cauchy surface represented by gravitational as well as nongravitational degrees of freedom. Nevertheless the metric compatibility conditions are undisturbed, and this apparent derivative-coupling is seen to be an effect of working with (possibly orthonormal) linear frames. The formalism is adapted to consider a Dirac Fermion, where we find that a milder form of this apparent derivative-coupling appears. © *1999 American Institute of Physics.* [S0022-2488(99)03407-6]

I. INTRODUCTION

In Ref. 1 we gave a detailed description of the surface geometry and canonical structure of general relativity (GR) in a general linear frame, using two different choices of the diffeomorphism constraints; the ''unprimed'' constraints which are compatible with a coordinate frame gauge choice, and the ''primed'' constraints which are compatible with an orthonormal (or Lorentz) frame gauge choice. These arose from considering an action that allows one to treat the metric and vierbein fields independently. Specializing to a surface-normal frame, we were led to a Hamiltonian formalism that treats the spatial frame as initial data that is independent of the spatial metric degrees of freedom, and with atlas fields $(N,N^a,N^a{}_b)$ that enforce the Hamiltonian, momentum and frame constraints respectively. The unprimed: $\underline{\mathcal{H}}$ and $\underline{\mathcal{H}}_a$, and primed: $\underline{\mathcal{H}}'$ and $\underline{\mathcal{H}}'_a$, constraints are related by nontrivial factors involving the generators of infinitesimal frame transformations: $\underline{\mathcal{J}}^a{}_b$. This work is essentially an extension of this, showing that the Hamiltonian description of matter fields minimally coupled to general relativity is compatible with the general linear frame formalism. Note that, as in Ref. 2, we have specialized somewhat by choosing a surface-normal frame and assuming the existence of a spatial coordinate frame, thereby choosing a minimal representation of the evolution of the spatial hypersurface in spacetime. Operationally this should be seen as no particular limitation since any numerical scheme would certainly begin from such a point of view. (For alternative points of view, see Refs. 3 and 4.)

We shall see that, as expected, the addition of matter fields poses no great difficulty. The somewhat surprising feature that appears is the breakdown of DeWitt's ''Riemannian Structure,''[5] since matter fields enter into the definition of the momenta conjugate to the frame fields (also occurring for any tensor field that behaves nontrivially under frame transformations). This is operationally due to the presence of derivatives of the frame field in the matter action, entering through the nonvanishing structure ''constants'' $C_{AB}{}^C$. Despite this mixing of gravitational and matter sectors, we show that the evolution of the spatial metric is identical to that required by the compatibility equations (i.e., the definition of the extrinsic curvature is unaltered) and the field equations are equivalent to those derived in a coordinate frame. Thus this form of derivative-

a)Electronic mail: Michael.A.Clayton@cern.ch

coupling is benign in the sense that the light cone structure of the theory is undisturbed[5]—trivially because it is a coordinate frame theory that has been rewritten. Despite naive expectation, this works because the matter stress-energy tensor remains independent of second-derivatives of the gravitational variables.

The ''unprimed'' system of constraints is closely related to coordinate frame approaches; the evolution equations and Lie transport act solely on the components of tensors with the frame playing a passive role. To be specific, infinitesimal spatial diffeomorphisms are defined to leave the frame unaffected and change the tensor components in the usual manner,[6] $\xi: T_a \to \mathcal{L}_{\xi}[T]_a$. Writing Einstein's equations in terms of the components of the metric γ_{ab}, frame $E^i{}_a$, and extrinsic curvature K_{ab} of a spacelike hypersurface Σ, we have

$$d_\perp[\gamma]_{ab} = 2K_{ab}, \quad d_\perp[E^i{}_a] = 0, \tag{1a}$$

$$d_\perp[K]_{ab} = -R_{ab} + \nabla_a \nabla_b[N]/N - KK_{ab} + 2K_{ac}K^c{}_b + \tfrac{1}{2}\overline{T}^{\mathrm{M}}_{ab}, \tag{1b}$$

where $d_\perp$ is the moving frame generalization (see the discussion in Sec. III B of Ref. 1 for more details) of the surface-covariant normal derivative operator.[7] We have added a matter stress-energy contribution $\overline{T}_{AB} := T_{AB} - \frac{1}{2} g_{AB} T$ to (1) for later reference, and have chosen $16\pi G = c = 1$. Lowercase Greek and Latin letters represent spacetime and spatial coordinate components and upper and lower case Roman letters indicate spacetime and spatial frame components, respectively. Symmetric and antisymmetric projection on any pair of indices is indicated by, for example, $T_{(ab)} := \frac{1}{2}(T_{ab} + T_{ba})$ and $T_{[ab]} := \frac{1}{2}(T_{ab} - T_{ba})$. Densities with respect to the spatial metric are represented as boldfaced symbols, and densities with respect to the determinant of the spatial vierbein with an underline. We also use ${}^4\nabla_A$ for the space–time covariant derivative and ∇_a for the intrinsic surface covariant derivative.

In the Hamiltonian system the action of $\mathcal{L}$ is represented on phase space by the momentum constraint $\underline{\boldsymbol{\mathcal{H}}}_a$ as $\mathcal{L}_{\vec{\xi}}[\,\cdot\,] \to \{\,\cdot\,, \underline{\boldsymbol{\mathcal{H}}}_a[\xi^a]\}$. We make use of the notation, for example, $\underline{\boldsymbol{\mathcal{H}}}_a[\xi^a] := \int_\Sigma d^3x\, \xi^a \underline{\boldsymbol{\mathcal{H}}}_a$, and $\{\cdot,\cdot\}$ is the Poisson bracket given by (32) of Ref. 1. The Hamiltonian constraint $\underline{\boldsymbol{\mathcal{H}}}$ represents the operator $d_\perp$ on phase space, and reproduces (1) weakly,

$$\{\gamma_{ab}, \underline{\boldsymbol{\mathcal{H}}}^{GR}[f]\} = 2\, f k_{ab}, \quad \{E^i{}_a, \underline{\boldsymbol{\mathcal{H}}}^{GR}[f]\} = 0, \tag{2a}$$

$$\{k_{ab}, \underline{\boldsymbol{\mathcal{H}}}^{GR}[f]\} = -fR_{ab} + \nabla_a \nabla_b[f] - fkk_{ab} + 2\, fk_{ac}k^c{}_b - \tfrac{1}{4} f \gamma_{ab} \mathcal{H}^{GR}, \tag{2b}$$

$$\{k'_{ab}, \underline{\boldsymbol{\mathcal{H}}}^{GR}[f]\} = -fR_{ab} + \nabla_a \nabla_b[f] - fkk'_{ab} + 2\, fk_{(a}{}^c k'_{cb)} - \tfrac{1}{4} f \gamma_{ab} \mathcal{H}^{GR} + fk_{(a}{}^c \mathcal{J}^{GR}_{[cb)]}. \tag{2c}$$

The extrinsic curvature K_{ab} is equivalently represented on phase space by the two tensors ($\underline{\pi}_{ab}$ and $\mathbf{p}^a{}_i$ are the momenta conjugate to $\boldsymbol{\gamma}^{ab}$ and $\underline{E}^i{}_a$, respectively),

$$k_{ab} := -(\pi_{ab} - \tfrac{1}{4}\gamma_{ab}\pi), \quad k'_{ab} := -\tfrac{1}{2}\gamma_{(ac} p^c{}_i E^i{}_{b)}, \tag{3}$$

and the weak equivalence of (2b) and (2c) guarantee that the evolution equations reproduce Einstein's equations (1). The generators of frame transformations on Σ: $\underline{\boldsymbol{\mathcal{J}}}^a{}_b$, act, for example on a covector field, as $\{T_a, \underline{\boldsymbol{\mathcal{J}}}^b{}_c[\omega^c{}_b]\} = \Delta_{\tilde{\omega}}[T]_a = -\omega^b{}_a T_b$, and satisfy the Lie algebra of $\mathfrak{gl}(3,\mathbb{R})$,

$$\{\underline{\boldsymbol{\mathcal{J}}}^a{}_b, \underline{\boldsymbol{\mathcal{J}}}^c{}_d[\omega^d{}_c]\} = \Delta_{\tilde{\omega}}[\underline{\boldsymbol{\mathcal{J}}}]^a{}_b. \tag{4}$$

Note that $\mathcal{J}^{GR}_{(ab)} = -2(k_{ab} - \gamma_{ab}k) + 2(k'_{ab} - \gamma_{ab}k')$, so that its weak vanishing implies that $k_{ab} \approx k'_{ab}$. The constraint algebra consists of (4) combined with

$$\{\underline{\boldsymbol{\mathcal{H}}}[f], \underline{\boldsymbol{\mathcal{J}}}^a{}_b[\omega^b{}_a]\} = 0, \quad \{\underline{\boldsymbol{\mathcal{H}}}_a[f^a], \underline{\boldsymbol{\mathcal{J}}}^a{}_b[\omega^b{}_a]\} = \int_\Sigma d^3x\, f^a \Delta_{\tilde{\omega}}[\underline{\boldsymbol{\mathcal{H}}}]_a, \tag{5a}$$

$$\{\underline{\boldsymbol{\mathcal{H}}}[f], \underline{\boldsymbol{\mathcal{H}}}[g]\} = (f\nabla_a[g] - g\nabla_a[f])\gamma^{ab}\underline{\boldsymbol{\mathcal{H}}}_b, \tag{5b}$$

$$\{\underline{\mathcal{H}}[f],\underline{\mathcal{H}}_a[g^a]\}=\int_\Sigma dx f\mathcal{L}_{\vec{g}}[\underline{\mathcal{H}}],\quad \{\underline{\mathcal{H}}_a[f^a],\underline{\mathcal{H}}_b[g^b]\}=\int_\Sigma d^3x f^a\mathcal{L}_{\vec{g}}[\underline{\mathcal{H}}]_a. \tag{5c}$$

This result was also derived in Ref. 2 using the geometric arguments of Teitelboim,[8,9] and is a fairly straightforward generalization of that found elsewhere (see, for example, Ref. 7).

We then noted [from the form of (1a) or (2a)] that these generators are not very convenient if one wishes to consider the restriction to orthonormal frames on Σ, since the choice $\gamma_{ab}=\delta_{ab}$ is not left invariant by the actions of $d_\perp$ and $\mathcal{L}$ (and therefore $\underline{\mathcal{H}}$ and $\underline{\mathcal{H}}_a$). There are, however, alternative representations of these operators in which the spatial frame plays a more active role. In particular, we represent an infinitesimal diffeomorphism by $\mathcal{L}'$ which acts on a frame as $\mathcal{L}'_{\vec{\xi}}[E^i{}_a]=\Delta_{\tilde{\nabla}\xi}[E^i{}_a]$, and on the components of tensors as the covariant derivative, $\mathcal{L}'_{\vec{\xi}}[T]_a=\xi^b\nabla_b[T]_a$; the action on the tensor itself is identical to that of $\mathcal{L}$: $\mathcal{L}_{\vec{\xi}}[T_a\theta^a]=\mathcal{L}'_{\vec{\xi}}[T_a\theta^a]$. The action of $\mathcal{L}'$ on the components of the spatial metric vanishes due to metric compatibility, and is therefore consistent with the limit to orthonormal spatial frames—note though that it is no longer consistent with the limit to a coordinate frame. Similarly we define the operator $d'_\perp$ to act on the frame as: $d'_\perp[E^i{}_a]=\Delta_{\tilde{K}}[E^i{}_a]=-K^b{}_aE^i{}_b$, and on the components of tensors as $d'_\perp=d_\perp+\Delta_{\tilde{K}}$. Using these, Einstein's equations (1) appear as

$$d'_\perp[\gamma]_{ab}=0,\quad d'_\perp[E^i{}_a]=\Delta_{\tilde{K}}E^i{}_b=-K^b{}_aE^i{}_b, \tag{6a}$$

$$d'_\perp[K]_{ab}=-R_{ab}+\nabla_a\nabla_b[N]/N-KK_{ab}+\tfrac{1}{2}\overline{T}^{\mathrm{M}}_{ab}. \tag{6b}$$

The action of $d'_\perp$ and $\mathcal{L}'$ are represented on phase space by the primed Hamiltonian constraint $\underline{\mathcal{H}}'$ and the primed momentum constraint $\underline{\mathcal{H}}'_a$, respectively (which correspond to a particular choice of mixing of the unprimed constraints with the frame rotation generators),

$$\{\gamma_{ab},\underline{\mathcal{H}}'^{GR}[f]\}=0,\quad \{E^i{}_a,\underline{\mathcal{H}}'^{GR}[f]\}=\Delta_{\tilde{k}'}[E^i{}_a]=-k'^b{}_aE^i{}_b, \tag{7a}$$

$$\{k_{ab},\underline{\mathcal{H}}'^{GR}[f]\}=-fR_{ab}+\nabla_a\nabla_b[f]-fk'k_{ab}-\tfrac{1}{4}f\gamma_{ab}\mathcal{H}'^{GR}-k'_{(a}{}^c\mathcal{J}^{GR}_{[b)c]}, \tag{7b}$$

$$\{k'_{ab},\underline{\mathcal{H}}'^{GR}[f]\}=-fR_{ab}+\nabla_a\nabla_b[f]-fk'k'_{ab}-\tfrac{1}{4}f\gamma_{ab}\mathcal{H}'^{GR}-fk'_{(a}{}^c\mathcal{J}^{GR}_{[b)c]}, \tag{7c}$$

and the constraint algebra for this primed system consists of (4) combined with (also derived in Ref. 2),

$$\{\underline{\mathcal{H}}'[f],\underline{\mathcal{J}}^a{}_b[\omega^b{}_a]\}=0,\quad \{\underline{\mathcal{H}}'_a[N^a],\underline{\mathcal{J}}^a{}_b[\omega^b{}_a]\}=\int_\Sigma d^3xN^a\Delta_{\tilde{\omega}}[\underline{\mathcal{H}}']_a, \tag{8a}$$

$$\{\underline{\mathcal{H}}'[f],\underline{\mathcal{H}}'[g]\}=\int_\Sigma d^3x(f\nabla_a[g]-g\nabla_a[f])(\gamma^{ab}\underline{\mathcal{H}}'_b-E\nabla_b[\underline{\mathcal{J}}]^{[ab]}), \tag{8b}$$

$$\{\underline{\mathcal{H}}'[f],\underline{\mathcal{H}}'_a[g^a]\}=\int_\Sigma d^3x(f\mathcal{L}_{\vec{g}}[\underline{\mathcal{H}}']-fg^a\Delta_{\tilde{k}'}[\underline{\mathcal{H}}']_a+2g^a\nabla_c[fk'_{ab}]\underline{\mathcal{J}}^{[bc]}), \tag{8c}$$

$$\{\underline{\mathcal{H}}'_a[f^a],\underline{\mathcal{H}}'_b[g^b]\}=-\int_\Sigma d^3xf^ag^bR^c{}_{dab}\underline{\mathcal{J}}^d{}_c. \tag{8d}$$

As we shall see, including matter presents no great difficulties. In Sec. II, we consider a scalar field as a brief demonstration of how matter fields fit into the generalized structure. In Sec. III, a one-form (gauge) field is introduced, transforming nontrivially under spatial frame rotations, and an additional constraint generating U(1) transformations appears. We find nontrivial mixing of the Maxwell canonical pair with that of the spatial vierbein, mixing that, at least superficially, resembles derivative-coupling. We will see that this is a manifestation of the gauge choice and is not ''true'' derivative-coupling since the metric compatibility conditions are unaffected. Finally, in Sec. IV, the spatial frame is constrained to be orthogonal in order to introduce a Dirac field.

In all three cases the matter action is added to that of vacuum *GR*: $S^{GR}\rightarrow S^{GR}+S^{M}$, and since both sets of vacuum *GR* constraints: $(\mathcal{H}^{GR},\mathcal{H}_a^{GR})$ and $(\mathcal{H}'^{\,GR},\mathcal{H}_a'^{\,GR})$, are equivalent to the components of the Einstein tensor: $(-2G^{\perp}{}_{\perp},-2G^{\perp}{}_{a})$, we find that the additively combined vacuum *GR* and matter constraints are weakly equivalent to adding $(T^{\perp}{}_{\perp},T^{\perp}{}_{a})$ to these. The matter stress-energy tensor is derived from the matter action S^M via

$$T_{AB}=\frac{2}{E}\left(\frac{\delta S^M}{\delta \mathbf{g}^{AB}}-\frac{1}{2}g_{AB}g^{CD}\frac{\delta S^M}{\delta \mathbf{g}^{CD}}\right)=\frac{1}{\sqrt{-g}}\left(g_{AC}E^{\mu}{}_{B}\frac{\delta S^M}{\delta \underline{E}^{\mu}{}_{C}}-g_{AB}E^{\mu}{}_{C}\frac{\delta S^M}{\delta \underline{E}^{\mu}{}_{C}}\right); \tag{9}$$

the equivalence of these two forms is a basic consistency requirement for the generalized treatment considered herein. By an argument of Floreanini and Percacci,[10] we know that once the matter action has been properly written in a general linear frame these forms will be consistent. Note though that in (9) we consider the metric and vierbein degrees of freedom *independently*, which will *not* work with the E–D system—only the second form is applicable. We will show (briefly) that the matter field equations and the stress-energy tensor determined from both forms of (9) are properly generated via a variational principle. Also that the action of the constraints on phase space properly generates spatial diffeomorphisms once matter has been incorporated into the canonical formalism, and (2) and (7) are properly extended to reproduce (1) and (6), respectively, with the appropriate stress-energy tensor.

One of the unfortunate aspects and at the same time strengths of diffeomorphism invariant theories is the great arbitrariness in parameterizing configuration and momentum space fields; any ''all-encompassing'' formalism would have to include canonical transformations that relate different parameterizations. The content of the ''geometric'' derivations of the constraint algebras[8,9,2] is that the algebra is fixed by the reduction of the space–time equations to the evolution of spatial quantities, that is, the algebra is a reflection of the spacetime diffeomorphism invariance. The details of the algebra will change depending on the parameterization of phase space and the choice of constraints, but is always of a fixed form once these choices are made. The point of this work is *not* to advocate any particular parameterization, merely to make it clear that there is actually a wider range of possibilities than heretofore considered. Within this wider class one finds that the Einstein–Dirac system may be treated in a particular limit, and not as an awkward construction that seemingly must be invented solely to deal with Fermion fields. Thus we find that the Einstein–Dirac (E–D) system is operationally no different than the Einstein–Klein Gordon (E–KG) or Einstein–Maxwell (E–M) systems; there is no more derivative-coupling in the E–D system than in the E–M system. Indeed the mixing of canonical variables is necessary in order to retrieve the correct surface frame transformation generators, while retaining the metric compatibility conditions at the level of the field equations.

II. A SCALAR FIELD

This example is a trivial extension of the vacuum GR results, included as a straightforward example of how the extension of the results of Ref. 1 outlined in Sec. I proceeds. Its simplicity is due to the fact that we have chosen canonical coordinates for the scalar field that do not transform under changes of spatial frame. Note that this is *not* necessary; we could just as well chosen to parameterize the scalar field action in terms of a densitized scalar field, in which case some of the structure that we will encounter in Sec. III would have appeared.

The scalar field Lagrangian density with self-coupling potential $V[\phi]$ is

$$\begin{aligned}\underline{\mathcal{L}}^{\phi}&=\mathbf{E}\left(\tfrac{1}{2}\mathrm{g}^{AB}\,{}^{4}\nabla_A[\phi]\,{}^{4}\nabla_B[\phi]-V[\phi]\right)\\&=\frac{\mathbf{E}}{2N}(\partial_t[\phi]-N^a e_a[\phi])^2-\frac{1}{2}\underline{N}\boldsymbol{\gamma}^{ab}\nabla_a[\phi]\nabla_b[\phi]-\underline{N}V[\phi],\end{aligned} \tag{10}$$

from which we derive [from either form of Eq. (9)],

$$T^{\phi}_{AB}={}^4\nabla_A[\phi]{}^4\nabla_B[\phi]-\tfrac{1}{2}g_{AB}g^{CD}\,{}^4\nabla_C[\phi]{}^4\nabla_D[\phi]+g_{AB}V[\phi], \tag{11a}$$

$$\overline{T}^{\phi}_{AB}={}^4\nabla_A[\phi]{}^4\nabla_B[\phi]-g_{AB}V[\phi]. \tag{11b}$$

Using the field equations for ϕ,

$$\frac{\delta S^{\phi}}{\delta\phi}=E\mathbf{g}^{AB}\,{}^4\nabla_A[{}^4\nabla_B[\phi]]-\mathbf{E}\delta_{\phi}[V], \tag{12}$$

it is straightforward to demonstrate that the conservation law ${}^4\nabla_B[T^{\phi}]^B{}_A=0$ is satisfied.

From (10) we see that the phase space of the scalar field may be parameterized by the conjugate pair, $(\phi,\underline{\mathbf{P}})$, where

$$\underline{\mathbf{P}}:=\frac{\mathbf{E}}{N}(\partial_t[\phi]-N^a e_a[\phi])=\mathbf{E}e_{\perp}[\phi]. \tag{13}$$

As a result, the Poisson bracket for the extended system is $\{F,G\}=\{F,G\}_{GR}+\{F,G\}_{\phi}$, where $\{.,.\}_{GR}$ is the *GR* sector Poisson bracket given in Ref. 1 and

$$\{F,G\}_{\phi}:=\int_{\Sigma}d^3x\left(\frac{\delta F}{\delta\phi(x)}\,\frac{\delta G}{\delta\underline{\mathbf{P}}(x)}-\frac{\delta G}{\delta\phi(x)}\,\frac{\delta F}{\delta\underline{\mathbf{P}}(x)}\right). \tag{14}$$

Since neither the extrinsic curvature nor time derivatives of the spatial metric or frame appear in the Lagrangian density (10), we see that the definitions of k_{ab} and k'_{ab} in (3) are undisturbed, and the frame rotation generators are identical to those in vacuum *GR*, $\underline{\mathcal{J}}^a{}_b=\underline{\mathcal{J}}^{GR\,a}{}_b$. Thus we find that

$$\underline{\boldsymbol{\mathcal{H}}}^{\phi}=\underline{\boldsymbol{\mathcal{H}}}'^{\phi}=\tfrac{1}{2}\mathbf{E}(P_{\phi})^2+\tfrac{1}{2}E\boldsymbol{\gamma}^{ab}\nabla_a[\phi]\nabla_b[\phi]+\mathbf{E}V[\phi]=\underline{\mathbf{T}}^{\phi\perp}{}_{\perp}, \tag{15a}$$

$$\underline{\boldsymbol{\mathcal{H}}}^{\phi}_a=\underline{\boldsymbol{\mathcal{H}}}'^{\phi}_a=\underline{\mathbf{P}}_{\phi}\nabla_a[\phi]=\underline{\mathbf{T}}^{\phi\perp}{}_a, \tag{15b}$$

are added to the *GR* constraints.

Hamilton's equations for the scalar field,

$$\{\phi,H\}=NP_{\phi}+N^a\nabla_a[\phi],\quad \{\underline{\mathbf{P}}_{\phi},H\}=E\nabla_a[N\boldsymbol{\gamma}^{ab}\nabla_b[\phi]]-\underline{\mathbf{N}}\delta_{\phi}[V]+E\nabla_a[N^a\mathbf{P}_{\phi}], \tag{16}$$

are equivalent to (12), and noting that $\{\boldsymbol{\gamma}^{ab},\underline{\boldsymbol{\mathcal{H}}}^M[f]\}=\{\boldsymbol{\gamma}^{ab},\underline{\boldsymbol{\mathcal{H}}}^M_a[f^a]\}=0$ and $\{\underline{E}^i{}_a,\underline{\boldsymbol{\mathcal{H}}}^M[f]\}=\{\underline{E}^i{}_a,\underline{\boldsymbol{\mathcal{H}}}^M_a[f^a]\}=0$, we see that neither of (2a) nor (7a) are altered. Using (with similar definitions with primes for Poisson brackets of the primed constraints)

$$\{\mathbf{p}^a{}_i,\underline{\boldsymbol{\mathcal{H}}}^{\phi}[f]\}=\mathbf{Y}^a{}_b[f]E^b{}_i-\tfrac{1}{2}\mathbf{Y}[f]E^a{}_i,\quad \{\underline{\pi}_{ab},\underline{\boldsymbol{\mathcal{H}}}^{\phi}[f]\}=\underline{X}_{ab}[f]-\gamma_{ab}\underline{X}[f], \tag{17a}$$

$$\{\mathbf{p}^a{}_i,\underline{\boldsymbol{\mathcal{H}}}^{\phi}_b[f^b]\}=\mathbf{W}^a{}_b[\vec{f}]E^b{}_i-\tfrac{1}{2}\mathbf{W}[\vec{f}]E^a{}_i,\quad \{\underline{\pi}_{ab},\underline{\boldsymbol{\mathcal{H}}}^{\phi}_c[f^c]\}=\mathbf{Z}_{ab}[\vec{f}]-\gamma_{ab}\mathbf{Z}[\vec{f}], \tag{17b}$$

where

$$W^a{}_b[\vec{f}]=W'^a{}_b[\vec{f}]=-f^a\mathcal{H}^{\phi}_b,\quad Z_{ab}[\vec{f}]=Z'_{ab}[\vec{f}]=0, \tag{18a}$$

$$X_{ab}[f]=X'_{ab}[f]=-\tfrac{1}{2}fT^{\phi}_{ab}=-\tfrac{1}{2}f(\overline{T}_{ab}-\gamma_{ab}\overline{T})-\tfrac{1}{2}f\gamma_{ab}\mathcal{H}^{\phi}, \tag{18b}$$

$$Y_{ab}[f]=Y'_{ab}[f]=-fT^{\phi}_{ab}=-f(\overline{T}_{ab}-\gamma_{ab}\overline{T})-f\gamma_{ab}\mathcal{H}^{\phi}, \tag{18c}$$

it is straightforward to show that the evolution of the extrinsic curvature is correctly generated. That is, we find that $\mathcal{H}^{GR} \rightarrow \mathcal{H}^{GR} + \mathcal{H}^{\phi}$ and $\frac{1}{2} f \overline{T}^{\phi}_{ab}$ is appended to the right-hand sides of (2b), (2c), (7b), and (7c) in accordance with (1) and (6). The unprimed and primed constraint algebras are identical to (44) and (49), respectively, of Ref. 1 with $\mathcal{H} = \mathcal{H}^{GR} + \mathcal{H}^{\phi}$ and $\mathcal{H}_a = \mathcal{H}^{GR}_a + \mathcal{H}^{\phi}_a$, or, equivalently, (40) and (45) combined with $\mathcal{H}^{A}_{U(1)} = 0$.

III. MAXWELL FIELDS

Since the vector potential of the Einstein–Maxwell system is a one-form field, it is associated to $GL\mathbf{M}$ through a (co-)vector representation of $GL(4,\mathbb{R})$, and therefore results in a nontrivial contribution to $\mathcal{J}^a{}_b$. This means that the primed and unprimed cases may no longer be considered together; the unprimed system is considered in Sec. III A and the primed in Sec. III B. We shall see that although the details of the calculation are nontrivial, the results are a fairly straightforward extension of those of vacuum *GR*.

The components of the field strength tensor: $F := dA = \frac{1}{2} F_{AB} \theta^A \wedge \theta^B$, are derived from the one-form $A := A_A \theta^A$ by

$$F_{AB} = {}^4\nabla_A[A]_B - {}^4\nabla_B[A]_A = e_A[A_B] - e_B[A_A] - C_{AB}{}^C A_C, \tag{19}$$

and the standard Lagrangian density is

$$\underline{\mathcal{L}}^A = -\tfrac{1}{4} \mathbf{E} g^{AC} g^{BD} F_{AB} F_{CD} = \underline{\mathbf{N}} \left(\tfrac{1}{2} \gamma^{ab} F_{\perp a} F_{\perp b} - \tfrac{1}{4} F_{ab} F^{ab} \right), \tag{20}$$

where $F_{ab} = \nabla_a A_b - \nabla_b A_a$, and

$$F_{\perp a} = (\partial_t [A_a] - A_b E^b{}_i \partial_t [E^i{}_a] - \mathcal{L}_{\vec{N}}[A]_a)/N - \nabla_a [A_\perp] - A_\perp \nabla_a [\ln(N)]. \tag{21}$$

The Maxwell field equations derived from (20) are

$$\frac{\delta S^M}{\delta A_A} = \mathbf{E}\, {}^4\nabla_B [F]^{BA} = 0, \tag{22}$$

and making use of the variation of the field strength tensor with respect to the spatial frame

$$\delta_E F_{AB} = 2\, {}^4\nabla_C [A]_{[B} E^C{}_\mu \delta E^\mu{}_{A]} + 2 A_C\, {}^4\nabla_{[B} [E^C{}_\mu \delta E^\mu{}_{A]}], \tag{23}$$

we find that the stress-energy tensor as determined by either form of (9) is

$$T^A_{AB} = \overline{T}^A_{AB} = -g^{CD} F_{AC} F_{BD} + \tfrac{1}{4} g_{AB} F^{BC} F_{BC}. \tag{24}$$

From the field equations (22) and using the Bianchi identity, ${}^4\nabla_A [F]_{BC} + {}^4\nabla_C [F]_{AB} + {}^4\nabla_B [F]_{CA} = 0$, we find that the conservation laws, ${}^4\nabla_B [T^A]^B{}_A = 0$, are satisfied. Note that despite the presence of the structure constants $C_{AB}{}^C$ in the field strength tensor (19); the variation of the action (20) using (23) does *not* lead to any terms containing the second-derivative of the frame.

We see from (20) and (21) that the momenta conjugate to A_a are

$$\frac{\delta L^{\mathrm{A}}}{\delta \partial_t [A_a]} = \underline{\gamma}^{ab} F_{\perp b} =: \underline{\mathbf{P}}^a. \tag{25}$$

The phase space of the Maxwell field may be parameterized by the conjugate pair $(A_a, \underline{\mathbf{P}}^a)$, and the Poisson bracket is extended by (summation over a is assumed)

$$\{F,G\}_{\mathrm{A}} := \int_\Sigma d^3x \left(\frac{\delta F}{\delta A_a(x)} \frac{\delta G}{\delta \underline{\mathbf{P}}^a(x)} - \frac{\delta G}{\delta A_a(x)} \frac{\delta F}{\delta \underline{\mathbf{P}}^a(x)} \right). \tag{26}$$

However, we also find that the ''geometrodynamical momentum'' conjugate to $\underline{E}^i{}_a$ as given in (30) of Ref. 1 becomes

$$\mathbf{p}^a{}_i := -2\mathbf{K}^a{}_b E^b{}_i - \mathbf{P}^a A_b E^b{}_i + \tfrac{1}{2} A^b \mathbf{P}_b E^a{}_i , \tag{27}$$

the first term of which is the ''gravitational momentum.''[5] Thus we find that the gravitational and nongravitational fields are mixed, and DeWitt's supermetric is no longer block-diagonal. By defining

$$\kappa_{ab} := -\tfrac{1}{2}(P_{(a}A_{b)} - \tfrac{1}{2}\gamma_{ab}P^c A_c), \tag{28}$$

we now have [cf. the discussion following (33) of Ref. 1],

$$K_{ab} \approx \bar{a} k_{ab} + (1-\bar{a})(k'_{ab} + \kappa_{ab}), \tag{29}$$

where the unprimed case corresponds to $\bar{a}=1$ and the primed case to $\bar{a}=0$. In addition, the generators of frame transformations become

$$\underline{\mathcal{J}}^a{}_b = 2\gamma^{ac}\underline{\pi}_{bc} - \underline{\pi}\delta^a_b - \mathbf{p}^a{}_i \underline{E}^i{}_b + \underline{\mathbf{p}}\delta^a_b - \underline{\mathbf{P}}^a A_b , \tag{30}$$

which correctly generate frame transformations on the vector field sector,

$$\{A_a , \underline{\mathcal{J}}^b{}_c[\omega^c{}_b]\} = \Delta_{\tilde{\omega}}[A]_a , \quad \{\underline{\mathbf{P}}^a , \underline{\mathcal{J}}^b{}_c[\omega^c{}_b]\} = \Delta_{\tilde{\omega}}[\underline{\mathbf{P}}]^a , \tag{31}$$

and satisfy (4).

The Lagrange multiplier $A_\perp$ enforces the U(1) constraint,

$$\underline{\mathcal{H}}^A_{U(1)} = -\mathbf{E}\nabla_a[P]^a, \tag{32}$$

and we find

$$\{\underline{\mathcal{H}}^A_{U(1)}[\alpha], \underline{\mathcal{H}}^A_{U(1)}[\beta]\} = 0, \quad \{\underline{\mathcal{H}}^A_{U(1)}[\alpha], \underline{\mathcal{J}}^a{}_b[\omega^b{}_a]\} = 0, \tag{33}$$

and the nonvanishing brackets,

$$\{A_a , \underline{\mathcal{H}}^A_{U(1)}[\alpha]\} = \nabla_a[\alpha], \quad \{\underline{\mathbf{P}}^a , \underline{\mathcal{H}}^A_{U(1)}[\alpha]\} = 0, \tag{34a}$$

$$\{\mathbf{p}^a{}_i , \underline{\mathcal{H}}^A_{U(1)}[\alpha]\} = -(\mathbf{P}^a E^b{}_i - \tfrac{1}{2}\mathbf{P}^b E^a{}_i)\nabla_b[\alpha]. \tag{34b}$$

This additional constraint is appended to the standard forms of the Hamiltonian via an additional atlas field, $\alpha := NA_\perp$, as (and similarly for the primed system),

$$H^{\mathrm{A}} = \int_\Sigma d^3x (N\underline{\mathcal{H}} + N^a \underline{\mathcal{H}}_a + N^a{}_b \underline{\mathcal{J}}^b{}_a + \alpha \underline{\mathcal{H}}^A_{U(1)}). \tag{35}$$

To determine the form of the other constraints appearing in (35) we follow the development in Ref. 1; choosing $K_{ab} = k_{ab}$ $(\bar{a}=1)$ leads to the unprimed system that closely resembles the coordinate frame approach (Sec. III A), whereas choosing $K_{ab} = k'_{ab} + \kappa_{ab}$ $(\bar{a}=0)$ leads to the primed case (Sec. III B).

A. The unprimed system

Since for the unprimed system $K_{ab} = k_{ab}$, the results follow from identifying the Hamiltonian and momentum constraints from the Maxwell Hamiltonian, leading to the following additions to the unprimed constraints:

$$\underline{\mathcal{H}}^{\mathrm{A}} := \tfrac{1}{2}\mathbf{E}P^a P_a + \tfrac{1}{4}\mathbf{E}F^{ab}F_{ab} = \underline{\mathbf{T}}^{A^\perp}{}_\perp , \tag{36a}$$

$$\underline{\boldsymbol{\mathcal{H}}}_a^A := \underline{\mathbf{P}}^b F_{ab} - \underline{\mathbf{A}}_a \nabla_b [P]^b = \underline{\mathbf{T}}^{A\perp}{}_a + A_a \underline{\boldsymbol{\mathcal{H}}}_{U(1)}^A . \tag{36b}$$

From these it is straightforward to determine

$$\{A_a, \underline{\boldsymbol{\mathcal{H}}}^A[f]\} = f P_a , \quad \{\underline{\mathbf{P}}^a, \underline{\boldsymbol{\mathcal{H}}}^A[f]\} = -\mathbf{E}\nabla_b [f F^{ab}], \tag{37a}$$

$$\{A_a, \underline{\boldsymbol{\mathcal{H}}}_b[f^b]\} = \mathcal{L}_{\vec{f}}[A]_a , \quad \{\underline{\mathbf{P}}^a, \underline{\boldsymbol{\mathcal{H}}}_b[f^b]\} = E\mathcal{L}_{\vec{f}}[\mathbf{P}]^a, \tag{37b}$$

and, using (34), Hamilton's equations for the Maxwell field are found

$$\{A_a, H\} = N P_a + \mathcal{L}_{\vec{N}}[A]_a + \Delta_{\tilde{N}}[A]_a + \nabla_a [\alpha], \tag{38a}$$

$$\{\underline{\mathbf{P}}^a, H\} = -\mathbf{E}\nabla_b [N F^{ab}] + E\mathcal{L}_{\vec{N}}[\mathbf{P}]^a + \Delta_{\tilde{N}}[\underline{\mathbf{P}}]^a. \tag{38b}$$

The combination of these with the $U(1)$ constraint (32) is equivalent to (22).

Using the definitions in (17) we find

$$\mathbf{W}^a{}_b[\vec{f}] = -\mathcal{L}_{\vec{f}}[A_b \mathbf{P}^a] - f^a \underline{\boldsymbol{\mathcal{H}}}_b^A , \quad Z_{ab} = 0, \tag{39a}$$

$$Y^a{}_b[f] = -f\overline{T}^A_{ab} - f P_a P_b - A_b \nabla_c [f F^{ca}], \quad X_{ab}[f] = -\tfrac{1}{2} f \overline{T}^A_{ab} . \tag{39b}$$

Noting that $\overline{T} := \gamma^{ab}\overline{T}_{ab} = \mathcal{H}^A$, it is straightforward to show that $\{k_{ab}, \underline{\boldsymbol{\mathcal{H}}}\} = \tfrac{1}{2} f \overline{T}_{ab} - \tfrac{1}{4}\gamma_{ab}\mathcal{H}^A$, and therefore the correct extension of (2b) is generated. In order to check the extension of (2c), note that we now must compute $\{k'_{ab} + \kappa_{ab}, \underline{\boldsymbol{\mathcal{H}}}[f]\}$, which correctly results in the additional contributions, $\tfrac{1}{2} f \overline{T}_{ab} - \tfrac{1}{4} f \gamma_{ab}\mathcal{H}^A - f k \kappa_{ab} + 2\, f k_{(a}{}^c \kappa_{cb)} + f k_{(a}{}^c \mathcal{J}^A_{[cb)]}$.

It is now possible to compute the algebra of the unprimed constraints, finding (4), (33), (5a), (5b), and

$$\{\underline{\boldsymbol{\mathcal{H}}}_{U(1)}^A[\alpha], \underline{\boldsymbol{\mathcal{H}}}[f]\} = 0, \quad \{\underline{\boldsymbol{\mathcal{H}}}_{U(1)}^A[\alpha], \underline{\boldsymbol{\mathcal{H}}}_a[f^a]\} = -\int_\Sigma d^3x\, \alpha \mathcal{L}_{\vec{f}}[\underline{\boldsymbol{\mathcal{H}}}_{U(1)}^A], \tag{40a}$$

$$\{\underline{\boldsymbol{\mathcal{H}}}[f], \underline{\boldsymbol{\mathcal{H}}}[g]\} = (f\nabla_a[g] - g\nabla_a[f])\,\gamma^{ab}(\underline{\boldsymbol{\mathcal{H}}}_b - A_b \underline{\boldsymbol{\mathcal{H}}}_{U(1)}^A). \tag{40b}$$

B. The primed system

After some algebra, the primed constraints found to be

$$\underline{\boldsymbol{\mathcal{H}}}'^A = \tfrac{1}{2}\mathbf{E} P^a P_a + \tfrac{1}{4}\mathbf{E} F^{ab} F_{ab} + \mathbf{E}(\kappa^a{}_b \kappa^b{}_a - \kappa^2) + 2\mathbf{E}(\kappa^a{}_b k'^b{}_a - k'\kappa), \tag{41a}$$

$$\underline{\boldsymbol{\mathcal{H}}}'^A_a = \underline{\mathbf{P}}^b \nabla_a [A]_b , \tag{41b}$$

which are related to (36) via combinations of the frame rotation generators given in Eq. (45) of Ref. 1, and therefore the combined constraints are equivalent to the Einstein equations, $G^\perp{}_A = \tfrac{1}{2} T^\perp{}_A$. Their action on the Maxwell phase space is found to be

$$\{A_a, \underline{\boldsymbol{\mathcal{H}}}'^A[f]\} = f P_a + f\Delta_{\tilde{k}' + \tilde{\kappa}}[A]_a , \quad \{\underline{\mathbf{P}}^a, \underline{\boldsymbol{\mathcal{H}}}'^A[f]\} = -\mathbf{E}\nabla_b [f F^{ab}] + f\Delta_{\tilde{k}' + \tilde{\kappa}}[\underline{\mathbf{P}}]^a, \tag{42a}$$

$$\{A_a, \underline{\boldsymbol{\mathcal{H}}}'^A_a[f^a]\} = f^b \nabla_b [A]_a = \mathcal{L}'_{\vec{f}}[A]_a , \quad \{\underline{\mathbf{P}}^a, \underline{\boldsymbol{\mathcal{H}}}'^A_a[f^a]\} = \mathbf{E}\nabla_b [f^b P^a] = \mathcal{L}'_{\vec{f}}[\underline{\mathbf{P}}]^a. \tag{42b}$$

Note that the actions (37) and (42) are those of $d_\perp$ and $\mathcal{L}$ and $d'_\perp$ and $\mathcal{L}'$ respectively. Hamilton's equations for the Maxwell sector are straightforward to find, and, as in the case of vacuum *GR* (see (52) of Ref. 1), are weakly equivalent to (38) when one makes the replacements, $N = N'$, $N_a = N'_a$, and $N^a{}_b = N'^a{}_b + \nabla_b [N']^a + N' k^a{}_b$.

Since $\{\underline{E}^i{}_a, \underline{\boldsymbol{\mathcal{H}}}'^A[f]\} = \Delta_{\tilde{\kappa}}[\underline{E}^i{}_a]$, the action of the Hamiltonian constraint on the frame is

$$\{\underline{E}^i{}_a,\underline{\boldsymbol{\mathcal{H}}}'[f]\}=f\Delta_{\tilde{k}'+\tilde{\kappa}}[\underline{E}^i{}_a],\tag{43}$$

properly reproducing the evolution of the vierbein in accordance with (6a). Again making use of the definitions in (17), we find

$$W'_{ab}[\vec{f}]=-f_a\mathcal{H}'^A_b+\nabla_c[f^cP_{[b}A_{a]}]+\nabla_c[P^cA_{(b}f_{a)}]-\nabla_c[A^cf_{(b}P_{a)}],\tag{44a}$$

$$X'_{ab}[f]=-\tfrac{1}{2}f\overline{T}^A_{ab}+\tfrac{1}{2}f\gamma_{ab}(\kappa^a{}_b\kappa^b{}_a-\kappa^2)$$
$$+f\gamma_{ab}(k'^a{}_b\kappa^b{}_a-k'\kappa)+f\kappa_{(a}{}^c\mathcal{J}^{GR}_{[b)c]}+f(k'_{(a}{}^c+\kappa_{(a}{}^c)\mathcal{J}^A_{[b)c]},\tag{44b}$$

$$Y'_{ab}[f]=-f\overline{T}_{ab}-fP_aP_b+A_b\nabla_c[fF_a{}^c]-2f(\kappa_a{}^ck'_{bc}-\kappa k'_{ab})$$
$$-f\gamma_{ab}(\kappa^a{}_b\kappa^b{}_a-\kappa^2)+f(\kappa_a{}^c-\delta^c_a\kappa)\mathcal{J}^{GR}_{[bc]},\tag{44c}$$

$$Z'_{ab}[\vec{f}]=\tfrac{1}{2}(\nabla_c[P^cf_{(a}A_{b)}]+\nabla_c[f^cP_{(a}A_{b)}]-\nabla_c[A^cf_{(a}P_{b)}]),\tag{44d}$$

and, noting that $\kappa_{(a}{}^c\mathcal{J}^A_{[cb)]}=\frac{1}{8}P^cP_cA_aA_b-\frac{1}{8}A^cA_cP_aP_b$, it is straightforward to show that the appropriate additions to (7b) and (7c) are generated from these. The primed constraint algebra consists of (4), (33), (8a), (8d), and

$$\{\underline{\boldsymbol{\mathcal{H}}}^A_{U(1)}[\alpha],\underline{\boldsymbol{\mathcal{H}}}'[f]\}=0,\quad \{\underline{\boldsymbol{\mathcal{H}}}^A_{U(1)}[\alpha],\underline{\boldsymbol{\mathcal{H}}}'_a[f^a]\}=-\int_\Sigma d^3x\,\alpha\mathcal{L}'_{\vec{f}}[\underline{\boldsymbol{\mathcal{H}}}^A_{U(1)}],\tag{45a}$$

$$\{\underline{\boldsymbol{\mathcal{H}}}'[f],\underline{\boldsymbol{\mathcal{H}}}'[g]\}=\int_\Sigma d^3x(f\nabla_a[g]-g\nabla_a[f])(\gamma^{ab}\underline{\boldsymbol{\mathcal{H}}}'_b-E\nabla_b[\boldsymbol{\mathcal{J}}]^{[ab]}-A^a\underline{\boldsymbol{\mathcal{H}}}^A_{U(1)}),\tag{45b}$$

$$\{\underline{\boldsymbol{\mathcal{H}}}'[f],\underline{\boldsymbol{\mathcal{H}}}'_a[g^a]\}=\int_\Sigma d^3x(\vec{f}\mathcal{L}_{\vec{g}}[\underline{\boldsymbol{\mathcal{H}}}']-fg^a\Delta_{\tilde{k}'+\tilde{\kappa}}[\underline{\boldsymbol{\mathcal{H}}}']_a+2g^a\nabla_c[f(k'_{ab}+\kappa_{ab})]\underline{\boldsymbol{\mathcal{J}}}^{[bc]}).\tag{45c}$$

IV. DIRAC SPINORS

In order to consider the introduction of a self-gravitating Dirac spinor, we need to do some preparatory work. The conditions for the existence of a spin structure on a manifold are known,[11,12] however, since we are primarily interested in the initial-value formalism, the relevant result is that any globally hyperbolic spacetime or space–time that admits a Cauchy surface admits a spin structure.[13] The general theory of spinors in curved spacetime (see, for example, Ref. 6) requires the introduction of an SL(2,ℂ) principle bundle—a Dirac spinor is associated to it through the product of a vector and conjugate vector representations of SL(2,ℂ). One may introduce spin frames, and would then develop formalism to deal with the initial-value problem similar to Ref. 14.

Here we will pursue a more straightforward, operationally-oriented path. We note that a Dirac spinor is associated directly to the Lorentz group through a spinor representation,[15] and so we will reduce the general linear frame bundle $GL\mathbf{M}$ to the Lorentz frame bundle $L\mathbf{M}$, with structure group O(1,3). (In fact, for the initial value problem we will be considering time and space oriented frames, and so have in effect reduced this further to $L^+_\uparrow\mathbf{M}$ with structure group the proper Lorentz group, $\mathcal{L}^+_\uparrow(1,3)=\{\Lambda\in O(1,3)|\det(\Lambda)=+1,\Lambda^0{}_0>0\}$.) Elements of $GL(4,\mathbb{R})$ are specialized to $O(1,3)$ as $M^A_B\rightarrow\Lambda^A{}_B$ and $|M^{-1}|^A_B\rightarrow\Lambda_B{}^A=\eta^{AC}\eta_{BD}\Lambda^D{}_C$, where $\Lambda^A{}_C\Lambda_B{}^C=\delta^A_B=\Lambda^C{}_B\Lambda_C{}^A$.

Given a Lorentz transformation Λ, the spinor ψ transforms as $\psi\rightarrow S(\Lambda)\psi$, and the gamma matrices satisfy the Dirac relations (we make use of the definitions and identities in Appendix 2A of Ref. 16), $\{\gamma^A,\gamma^B\}=2\eta^{AB}$, and $\Lambda^A{}_B\gamma^B=S^{-1}(\Lambda)\gamma^AS(\Lambda)$. Writing an infinitesimal Lorentz transformation as $\Lambda^A{}_B \approx \delta^A_B + \omega^A{}_B$, where $\omega^A{}_B$ are the antisymmetric generators of $\mathfrak{so}(1,3)$, for the spinor representation we have $S(\Lambda) \approx 1 - \frac{1}{4} i \hat{\omega}$, where $\hat{\omega} := \omega^A{}_B \sigma^B{}_A$, and $\sigma^{AB} := i \frac{1}{2} [\gamma^A, \gamma^B]$ satisfy the commutation relations

$$[\sigma^{AB}, \sigma^{CD}] = -2i(\sigma^{AC}\eta^{BD} - \sigma^{AD}\eta^{BC} + \sigma^{BD}\eta^{AC} - \sigma^{BC}\eta^{AD}). \tag{46}$$

Therefore the action of the generator of frame rotations Δ on the spinor is

$$\Delta_{\tilde{\omega}}[\psi] = -\tfrac{1}{4} i \hat{\omega} \psi, \tag{47}$$

and, using (46), it is straightforward to show that these generators satisfy the algebra

$$[\Delta_{\tilde{\omega}_1}, \Delta_{\tilde{\omega}_2}]\psi = -\Delta_{\widetilde{[\omega_1, \omega_2]}}[\psi]. \tag{48}$$

The covariant derivative operates on these spinors as

$${}^4\nabla_A[\psi] = e_A[\psi] + \tfrac{1}{4} i \Gamma^B_{AC} \sigma^C{}_B \psi = e_A[\psi] + \tfrac{1}{4} i \, {}^4\tilde{\Gamma}_A \psi, \tag{49}$$

and, as usual, $[{}^4\nabla_A, \gamma^B] = 0$.

The curved space–time Dirac action is (see Sec. 7.10.2 of Ref. 15),

$$S^{\mathrm{D}} = \int d^4x \, \det(E^A{}_\mu) \, \overline{\psi} (i \tfrac{1}{2} \gamma^A \, {}^4\overleftrightarrow{\nabla}_A - m) \psi, \tag{50}$$

where we employ the standard notation $\overline{\psi} \, {}^4\overleftrightarrow{\nabla}_A \psi := \overline{\psi} \, {}^4\nabla_A[\psi] - {}^4\nabla_A[\overline{\psi}] \psi$. A variation of (50) with respect to $\overline{\psi}$ and ψ leads to the Dirac equation

$$i \gamma^A \, {}^4\nabla_A[\psi] - m\psi = 0, \tag{51}$$

and its adjoint, respectively. As a consequence, the vector $J^A := \overline{\psi} \gamma^A \psi$, constructed from a solution to the Dirac equation, is covariantly conserved, ${}^4\nabla_A[J]^A = 0$. In a surface-adapted frame this becomes $\partial_t[E \psi^\dagger \psi] - E \nabla_a[N^a \psi^\dagger \psi] + E \nabla_a[N \psi^\dagger \alpha^a \psi] = 0$, which has the usual quantum-mechanical probabilistic interpretation. Unlike the scalar and Maxwell field examples, the spinor is *not* dimensionless after one has chosen $16\pi G = c = 1$. Indeed, the remaining length scale may be interpreted as the Planck length L_P (or $\hbar$). Here either ψ has dimension $length^{-1/2}$, or we have taken $\hbar = 1$. Note in particular that we will introduce neither $\gamma^\mu := E^\mu{}_A \gamma^A$ nor the coordinate components of the spacetime metric.

Using the variation of the ''spin-connection'' with respect to the vierbein,

$$\delta \hat{\Gamma}_A = \delta E^\mu{}_A E^D{}_\mu \hat{\Gamma}_D + (\eta_{BD} \, {}^4\nabla_A[E^D{}_\mu \delta E^\mu{}_C] - \eta_{BD} \, {}^4\nabla_C[E^D{}_\mu \delta E^\mu{}_A] + \eta_{AD} \, {}^4\nabla_B[E^D{}_\mu \delta E^\mu{}_C]) \sigma^{CB}, \tag{52}$$

and the second form of (9), we find the stress-energy tensor,

$$T^D_{AB} = i \tfrac{1}{2} \overline{\psi} \gamma_{(A} \, {}^4\overleftrightarrow{\nabla}_{B)} \psi, \quad T^D = m \overline{\psi} \psi. \tag{53}$$

Once again we find that there are no terms containing second derivatives of either the spatial metric or frame, albeit trivially because the Dirac action (50) only contains terms that are at mostly linear in first-order derivatives of these fields. The conservation laws, ${}^4\nabla_B[T^D]^B{}_A = 0$, follow from the use of (51), the Bianchi identities of the Riemann tensor, and the wave equation satisfied by the Fermion field, $(\eta^{AB} \, {}^4\nabla_A \, {}^4\nabla_B + m^2 - \frac{1}{4} R) \psi = 0$, where we have used $[{}^4\nabla_A, {}^4\nabla_B]\psi = \frac{1}{4} i R^C{}_{DAB} \sigma^D{}_C \psi = \frac{1}{4} i \hat{R}_{AB} \psi$ and $\hat{R}_{AB} \sigma^{AB} = -2R$.

A. The initial value formalism

To specialize the primed system to orthonormal frames is a simple matter of replacing γ_{ab} by δ_{ab} in all results and ignoring anything related to the $(\boldsymbol{\gamma}^{ab},\underline{\pi}_{ab})$ part of phase space (ending up with a description of canonical tetrad *GR* similar to that discussed in Ref. 17). In addition, it will be useful to introduce (using the conventions of Sec. 1.1 of Ref. 16), $\beta=\gamma^0$ and $\alpha^a=\gamma^0\gamma^a$, in terms of which we have $\sigma^{0a}=i\alpha^a$, and the spin matrices: $\sigma^{ab}:=-i\frac{1}{2}[\alpha^a,\alpha^b]$, satisfying the reduced version of (46),

$$[\sigma^{ab},\sigma^{cd}]=2i(\delta^{bc}\sigma^{da}+\delta^{bd}\sigma^{ac}+\delta^{ca}\sigma^{bd}+\delta^{da}\sigma^{cb}). \tag{54}$$

We define the reduction of the operator (47) that generates infinitesimal O(3) frame rotations by

$$\Delta_{\tilde{\omega}}\psi=i\tfrac{1}{4}\hat{\omega}\psi, \quad \text{where} \quad \hat{\omega}=\omega_{ab}\sigma^{ab}, \tag{55}$$

and the surface-covariant derivative as

$$\nabla_a[\psi]=e_a[\psi]+i\tfrac{1}{4}\hat{\Gamma}_a\psi, \tag{56}$$

where $\hat{\Gamma}_a:=\delta_{bd}\Gamma^b_{ac}\sigma^{bc}$. The commutator of two surface-covariant derivatives acting on a spinor results in $[\nabla_a,\nabla_b]\psi=\frac{1}{4}i\hat{R}_{ab}\psi=\frac{1}{4}iR_{cdab}\sigma^{cd}\psi$.

Decomposing the four-dimensional spin connection results in $i\frac{1}{4}{}^4\hat{\Gamma}_a=i\frac{1}{4}\hat{\Gamma}_a+\frac{1}{2}K_{ab}\alpha^b$ and $i\frac{1}{4}{}^4\hat{\Gamma}_\perp=\frac{1}{2}a_a\alpha^a-i\frac{1}{4}\hat{C}_\perp$, and so we have

$${}^4\nabla_a[\psi]=\nabla_a[\psi]+\tfrac{1}{2}K_{ab}\alpha^b\psi, \tag{57a}$$

$${}^4\nabla_\perp[\psi]=\frac{1}{N}\partial_t[\psi]-\frac{1}{N}N^a\nabla_a[\psi]-\frac{i}{4N}\nabla_a[N]_b\sigma^{ab}\psi+\tfrac{1}{2}a_a\alpha^a\psi+\frac{i}{4N}E_{ai}\partial_t[E^i{}_b]\sigma^{ab}\psi. \tag{57b}$$

(The quantity $\hat{C}_\perp:=\delta_{bc}C_{\perp a}{}^b\sigma^{ac}$, where $C_{\perp a}{}^b$, is given in (18b) of Ref. 1, and its role in the surface covariant normal derivative is discussed in Sec. III B of the same reference.)

Using these and introducing the ''half-densitized'' spinor fields (since $\det(\gamma_{ab})=\det(\delta_{ab})\equiv1$, densities of this type will not appear, and there should be no confusion with the boldfaced quantities of earlier sections),

$$\boldsymbol{\psi}:=E^{1/2}\psi, \quad \boldsymbol{\psi}^\dagger:=E^{1/2}\psi^\dagger, \tag{58}$$

from (50) it is straightforward to deduce the curved spacetime Dirac Lagrangian,

$$L^{\mathrm{D}}=\int_\Sigma d^3x\big(i\tfrac{1}{2}\boldsymbol{\psi}^\dagger\overleftrightarrow{\partial}_t\boldsymbol{\psi}+i\tfrac{1}{2}NE\psi^\dagger\alpha^a\overleftrightarrow{\nabla}_a\psi-mN\boldsymbol{\psi}^\dagger\beta\boldsymbol{\psi}-i\tfrac{1}{2}EN^a\psi^\dagger\overleftrightarrow{\nabla}_a\psi+\tfrac{1}{4}\nabla_a[N]_b\boldsymbol{\psi}^\dagger\sigma^{ab}\boldsymbol{\psi}$$
$$-\tfrac{1}{4}E_{ai}\partial_t[E^i{}_b]\boldsymbol{\psi}^\dagger\sigma^{ab}\boldsymbol{\psi}\big). \tag{59}$$

In order to perform the Legendre transform to obtain the E–D Hamiltonian, we need to choose an appropriate set of canonical coordinates in the Fermionic sector. In a flat space–time this is straightforward; one finds that (discarding a total time derivative from the action) the real and imaginary parts of ψ are canonically conjugate, and one can introduce complex coordinates on phase space and use the quantum-mechanical symplectic form[18] for a four component vector. In a curved space–time one finds that it is the densitized imaginary part of ψ that is conjugate to the real part, however, introducing the half-densitized spinors (58) allows the flat spacetime construction to proceed. Using the inner product on Σ (written first in covariant form, and then specialized to a surface-normal frame),

$$(\psi,\phi)_\Sigma = \int_\Sigma d\Sigma(x) n_A \overline{\psi}\gamma^A \phi = \int_\Sigma dx \, \boldsymbol{\psi}^\dagger \boldsymbol{\phi}, \tag{60}$$

to define the $L^2(\Sigma)$ pairing and therefore the quantum-mechanical symplectic form, the introduction of $(\boldsymbol{\psi},\boldsymbol{\psi}^\dagger)$ as complex coordinates on the Dirac sector of phase space results in the Poisson bracket,

$$\{F,G\}_D = -i\int_\Sigma dx \left(\frac{\delta F}{\delta \boldsymbol{\psi}(x)} \frac{\delta G}{\delta \boldsymbol{\psi}^\dagger(x)} - \frac{\delta G}{\delta \boldsymbol{\psi}(x)} \frac{\delta F}{\delta \boldsymbol{\psi}^\dagger(x)} \right). \tag{61}$$

On a fixed gravitational background, the result of the Legendre transformation would be to drop the first term in (59) and the negative of what remains is H^D. For the E–D system, we find that the final term in (59) contributes to the conjugate momentum of the vierbein, leading to

$$p^a{}_i := -2K^a{}_b E^b{}_i + \tfrac{1}{4}\psi^\dagger \sigma^{ab}\psi E_{bi}, \tag{62}$$

and once again the ''geometrodynamical momentum'' is not equal to the ''gravitational momentum.'' However, unlike the Maxwell case, the definition of the extrinsic curvature on phase space is undisturbed [$K_{ab} = k'_{ab}$ and (7a) is unaltered]; we find only a contribution to the frame rotation generators,

$$\mathcal{J}_{[ab]} := -p_{[ai}E^i{}_{b]} + \tfrac{1}{4}\psi^\dagger \sigma_{ab}\psi. \tag{63}$$

[Note that it is possible, and indeed somewhat simpler, to represent these rotation operators using $\mathcal{S}^a = \frac{1}{2}\epsilon^{abc}\mathcal{J}^{[bc]}$, making use of the spin matrices $s^a := \frac{1}{2}\epsilon^{abc}\sigma^{bc} = \sigma^a \mathbf{1}$, where σ^a are the Pauli spin matrices. We do not do so here simply because we wish to present results that are as similar as possible to those of the GL(3,$\mathbb{R}$) case.] In addition to generating infinitesimal frame rotations on the *GR* sector of phase space, we find that

$$\underline{\mathcal{J}}[\tilde{\omega}] = \int_\Sigma dx\, \omega_{ba} \underline{\mathcal{J}}^{[ab]}, \tag{64}$$

generates frame rotations on the Dirac sector

$$\{\boldsymbol{\psi}, \underline{\mathcal{J}}[\tilde{\omega}]\} = i\tfrac{1}{4}\hat{\omega}\boldsymbol{\psi} = \Delta_{\tilde{\omega}}[\boldsymbol{\psi}], \tag{65}$$

and the $\mathfrak{gl}(3,\mathbb{R})$ algebra (4) is reduced to that of $\mathfrak{so}(3)$,

$$\{\underline{\mathcal{J}}[\tilde{N}], \underline{\mathcal{J}}[\tilde{M}]\} = \int_\Sigma d^3x N_{ba} \Delta_{\tilde{M}}[\underline{\mathcal{J}}]^{[ab]}. \tag{66}$$

The resulting E–D Hamiltonian is once again in standard form,

$$H = \int_\Sigma d^3x (N' \underline{\mathcal{H}}' + N'^a \underline{\mathcal{H}}'_a + N'_{ba}\underline{\mathcal{J}}^{[ab]}), \tag{67}$$

where $\underline{\mathcal{H}}' = \underline{\mathcal{H}}'^{GR} + \underline{\mathcal{H}}^D$ and $\underline{\mathcal{H}}'_a = \underline{\mathcal{H}}'^{GR}_a + \underline{\mathcal{H}}^D_a$ (using the primed *GR* constraints with γ_{ab} replaced by δ_{ab}), and

$$\underline{\mathcal{H}}^D = -\tfrac{1}{2}iE\psi^\dagger \alpha^a \overleftrightarrow{\nabla}_a \psi + m\boldsymbol{\psi}^\dagger \beta \boldsymbol{\psi} = \underline{T}^{D\perp}{}_\perp, \quad \underline{\mathcal{H}}^D_a = i\tfrac{1}{2}E\psi^\dagger \overleftrightarrow{\nabla}_a \psi = \underline{T}^{D\perp}{}_a - E\nabla_b[\mathcal{J}^{D[ab]}], \tag{68}$$

arise from the Legendre transformation of (59). Note that the equivalence of these with the E–D stress-energy tensor components (53) is achieved by making use of the Dirac equation (51) to remove time derivatives of the spinor. Also note that the presence of the derivative of the Dirac frame rotation generator in the momentum constraint is consistent with the relationship between the unprimed and primed constraints in (45b) of Ref. 1. From (68) we find

$$\{\boldsymbol{\psi},\underline{\mathcal{H}}^{D}[f]\}=-fE^{1/2}\alpha^{a}\nabla_{a}[\psi]-\tfrac{1}{2}\nabla_{a}[f]\alpha^{a}\boldsymbol{\psi}-imf\beta\boldsymbol{\psi},\quad \{\boldsymbol{\psi},\underline{\mathcal{H}}_{a}'^{D}[f^{a}]\}=\mathcal{L}_{\mathbf{f}}'[\boldsymbol{\psi}], \tag{69}$$

where $\mathcal{L}'$ acts as described in Ref. 1,

$$\mathcal{L}_{\mathbf{f}}'[\boldsymbol{\psi}]=E^{1/2}f^{a}\nabla_{a}[\psi]+\Delta_{\tilde{\nabla}f}[E^{1/2}]\psi=E^{1/2}f^{a}\nabla_{a}[\psi]+\tfrac{1}{2}\nabla_{a}[f]^{a}\boldsymbol{\psi}. \tag{70}$$

Note that infinitesimal spatial diffeomorphisms act on the frame as $\{E^{i}{}_{a},\underline{\mathcal{H}}_{c}'[f^{c}]\}=-\nabla_{a}[f]^{a}E^{i}{}_{b}$; if one mixes in a spatial frame rotation so that only the symmetric part $\nabla_{(a}[f]_{b)}$ appears, then the resulting generator acting on ψ reproduces the action that is used, for example, in Ref. 19.

From (69) one finds Hamilton's equations for ψ,

$$\{\boldsymbol{\psi},H\}=E^{1/2}(N^{a}-N\alpha^{a})\nabla_{a}[\psi]+\tfrac{1}{2}(\nabla_{a}[N]^{a}-\alpha^{a}\nabla_{a}[N])\boldsymbol{\psi}-imN\beta\boldsymbol{\psi}+\tfrac{1}{4}i\hat{\omega}\boldsymbol{\psi}, \tag{71}$$

which are equivalent to (51). Again making use of the definitions (17) we find (we have used, $\{\sigma^{ab},\alpha^{c}\}=-2\epsilon^{abc}\gamma_{5}$),

$$W^{a}{}_{b}[\vec{f}]=-f^{a}\mathcal{H}_{b}^{D}-\tfrac{1}{4}\nabla_{c}[f^{c}\psi^{\dagger}\sigma_{b}{}^{a}\psi]-\tfrac{1}{4}\nabla_{c}[f^{a}\psi^{\dagger}\sigma^{c}{}_{b}\psi]+\tfrac{1}{4}\nabla_{c}[f_{b}\psi^{\dagger}\sigma^{ac}\psi], \tag{72a}$$

$$Y^{a}{}_{b}[f]=\tfrac{1}{2}if\psi^{\dagger}\alpha^{a}\overleftrightarrow{\nabla}_{b}\psi+\tfrac{1}{8}\nabla_{c}[f\psi^{\dagger}\{\alpha^{c},\sigma_{b}^{a}\}\psi], \tag{72b}$$

and using (53) to find

$$\overline{T}_{ab}^{D}=-\tfrac{1}{2}i\,\delta_{(ac}\psi^{\dagger}\alpha^{c}\overleftrightarrow{\nabla}_{b)}\psi+\tfrac{1}{2}\delta_{(ac}K_{b)d}\psi^{\dagger}\sigma^{cd}\psi+\tfrac{1}{2}\delta_{ab}m\psi^{\dagger}\beta\psi, \tag{73}$$

we find that contributions to (7b) and (7c) are generated to properly reproduce the E–D field equations. The constraint algebra combines (66) with (8), replacing $\gamma_{ab}\rightarrow\delta_{ab}$ everywhere in the latter. [Note that the derivation of the constraint algebra in Ref. 2 could be generalized to the Einstein–Dirac system along the lines of Ref. 20 note however the different parameterization of the spinor part of phase space and the resulting difference in the form of (68).]

V. DISCUSSION

The intent of the preceding[1] and present paper is to describe the Hamiltonian dynamics of general relativistic (vacuum and matter coupled) systems with respect to moving frames. This extended system of dynamical spatial metric *and* frame fields encompasses the following two natural limits: If we choose the diffeomorphism constraints to act on the components of tensors with the usual Lie action and fix the spatial frame to be a coordinate frame, we recover the standard coordinate frame approach to canonical general relativity. If we choose the spatial metric to be equal to the unit matrix and mix the diffeomorphism generators with the frame transformation generators to be compatible with this choice, we recover the orthonormal frame approach to canonical general relativity. Clearly there are also a variety of intermediate choices available, however this enlarged arena provides a bridge over which one may pass from the standard coordinate frame approach to the Lorentz (or orthonormal) frame approach that is more natural for the description of the Einstein–Dirac system.

In this generalized structure, we find that matter fields naïvely appear to be derivative-coupled since the momenta conjugate to the spatial frame fields are not (in general) independent of the matter fields. This sort of derivative-coupling turns out to be benign since the field equations are equivalent to the coordinate frame approach to the model in question. Indeed this feature is a necessary part of the frame approach, and will appear whenever the matter fields behave nontrivially under frame transformations. In this way we understand that the derivative-coupling in the Einstein–Dirac system is a necessary feature, merely reflecting the need to work with local orthonormal frames in order to define the system. We also end up with a simpler description of the Einstein–Dirac system than that given in Ref. 21 (note, however, their use of anticommuting fermions), and a much more natural system than that found in Ref. 22, where primary importance is still given to the coordinate components of the metric tensor. A more detailed examination of the Einstein–Dirac system in adjoint form (in the spirit of Ref. 23) is underway, and by adopting harmonic coordinate conditions and choosing the frame to fix the Local Lorentz gauge, we are able to prove local existence and uniqueness results.

ACKNOWLEDGMENTS

The author was supported by a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada, and thanks J. Légaré for commenting on a preliminary form of this manuscript.

[1] M. A. Clayton, ''Canonical general relativity: Diffeomorphism constraints and spatial frame transformations,'' J. Math. Phys. **39**, 3805–3816 (1998).
[2] M. A. Clayton, ''The tetrad frame constraint algebra,'' Class. Quantum Grav. **14**, 1851–1864 (1997).
[3] J. M. Charap, M. Henneaux, and J. E. Nelson, ''Explicit form of the constraint algebra in tetrad gravity,'' Class. Quantum Grav. **5**, 1405–1414 (1988).
[4] J. E. Nelson and T. Regge, ''Covariant canonical formalism for gravity,'' Ann. Phys. (N.Y.) **166**, 234–249 (1986).
[5] K. Kuchař, ''Geometrodynamics with tensor sources. IV,'' J. Math. Phys. **18**, 1589–1597 (1977).
[6] R. M. Wald, *General Relativity* (The University of Chicago Press, Chicago, 1984).
[7] J. Isenberg and J. Nester, ''Canonical Gravity,'' in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. I.
[8] C. Teitelboim, ''How commutators of constraints reflect the space–time structure,'' Ann. Phys. (N.Y.) **79**, 542–557 (1973).
[9] S. A. Hojman, K. V. Kuchař, and C. Teitelboim, ''Geometrodynamics Regained,'' Ann. Phys. (N.Y.) **96**, 88–135 (1976).
[10] R. Floreanini and R. Percacci, ''Canonical algebra of GL(4): Invariant gravity,'' Class. Quantum Grav. **7**, 975–984 (1990).
[11] R. Geroch, ''Spinor structure of space–times in general relativity. I,'' J. Math. Phys. **9**, 1739–1744 (1968).
[12] C. J. S. Clarke, ''Magnetic charge, holonomy and characteristic classes: Illustrations of the methods of topology in general relativity,'' Gen. Relativ. Gravit. **2**, 43–51 (1971).
[13] R. Geroch, ''Spinor structure of space-times in general relativity. II,'' J. Math. Phys. **11**, 343–348 (1970).
[14] A. Sen, ''Quantum theory of spin-3/2 field in Einstein spaces,'' Int. J. Theor. Phys. **21**, 1–35 (1982).
[15] M. Nakahara, *Geometry, Topology, and Physics* (Adam Hilger, New York, 1990).
[16] B. Thaller, *The Dirac Equation, Texts and Monographs in Physics* (Springer, New York, 1992).
[17] M. Bañados and M. Contreras, ''Darboux coordinates for (first-order) tetrad gravity,'' Class. Quantum Grav. **15**, 1527–1534 (1998).
[18] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Vol. 17 in *Texts in Applied Mathematics* (Springer, New York, 1994).
[19] M. Henneaux, ''Gravitational fields, spinor fields, and groups of motions,'' Gen. Relativ. Gravit. **12**, 137–147 (1980).
[20] K.-H. Lotze, ''The geometrodynamical super-Hamiltonian and supermomentum of the Dirac field in curved space–time,'' Acta Phys. Pol. B **9**, 185–190 (1978).
[21] D. Bao, J. Isenberg, and P. B. Yasskin, ''The dynamics of the Einstein–Dirac system. I. A principle bundle formulation of the theory and its canonical analysis,'' Ann. Phys. (N.Y.) **164**, 103–171 (1985).
[22] J. E. Nelson and C. Teitelboim, ''Hamiltonian formulation of the theory of interacting gravitational and electron fields,'' Ann. Phys. (N.Y.) **116**, 86–104 (1978).
[23] A. E. Fischer and J. E. Marsden, ''The initial value problem and the dynamical formulation of general relativity,'' in *General Relativity: An Einstein Centenary Survey*, edited by S. W. Hawking and W. Israel (Cambridge University Press, New York, 1979).

Reductive decompositions and Einstein–Yang–Mills equations associated to the oscillator group

Raúl Durán Díaz
Instituto de Física Aplicada, CSIC, Serrano 144, 28006-Madrid, Spain

Pedro M. Gadea
Instituto de Matemáticas y Física Fundamental, CSIC, Serrano 123, 28006-Madrid, Spain

José A. Oubiña
Departamento de Xeometría e Topoloxía, Facultade de Matemáticas, Universidade de Santiago de Compostela, 15706-Santiago de Compostela, Spain

(Received 22 February 1999; accepted for publication 10 March 1999)

All of the homogeneous Lorentzian structures on the oscillator group equipped with a bi-invariant Lorentzian metric, and then the associated reductive pairs, are obtained. Some of them are solutions of the Einstein–Yang–Mills equations.
 [S0022-2488(99)00107-3]

I. INTRODUCTION

The oscillator group has interesting features from the viewpoints of both differential geometry and physics (see, for instance, Refs. 1–12). It was introduced by Streater,[1] who named it so because its Lie algebra can be identified to that generated by the differential operators associated with the harmonic oscillator problem, acting on functions of a variable. The oscillator group is the only simply connected four-dimensional non-Abelian solvable Lie group which admits a bi-invariant Lorentzian metric.[3,4] Moreover, it is[8] an example of homogeneous space–time, which as causal space satisfies the so-called causal continuity.

Levichev studied in Ref. 5 the oscillator group with the biinvariant Lorentzian metric also given in Refs. 6 and 8, proving that this group (which geometrically is a Lorentzian symmetric space and physically is related to an isotropic electromagnetic field) provides a solution of the sourceless Einstein–Yang–Mills equations.

On the other hand, generalizing the classical characterization by Cartan[13] of Riemannian symmetric spaces as the spaces of parallel curvature, Ambrose and Singer[14] gave a characterization of Riemannian homogeneous spaces in terms of a (1,2) tensor field S, called homogeneous Riemannian structure in Ref. 15, satisfying certain differential equations (see (1) below). In Refs. 16 and 17, we have extended this concept to the pseudo-Riemannian case.

In Ref. 12, all of the homogeneous Lorentzian structures corresponding to a family of left invariant metrics on the general oscillator groups have been obtained, and the reductive decompositions for the non-bi-invariant metrics in that family have been determined.

In this paper we consider the four-dimensional oscillator group equipped with its usual Lorentzian bi-invariant metric and we determine all the homogeneous Lorentzian structures on it and all the associated reductive pairs. There appear six types of such pairs, and we prove that four of them are solutions of the Einstein–Yang–Mills equations. They have sources except for a particular case, where the symmetric pair appears as a solution of the sourceless Einstein–Yang–Mills equations.

II. PRELIMINARIES

Ambrose and Singer[14] gave a characterization for a connected, simply connected and complete Riemannian manifold to be homogeneous, in terms of a (1,2) tensor field S, called in Ref. 15 a *homogeneous Riemannian structure*. In Ref. 16 it is defined a *homogeneous pseudo-Riemannian* *structure* on a pseudo-Riemannian manifold (M,g) as a tensor field S of type (1,2) such that being ∇ the Levi-Civita connection and R its curvature tensor, the connection $\tilde{\nabla}=\nabla-S$ satisfies the Ambrose–Singer equations

$$\tilde{\nabla}g=0, \quad \tilde{\nabla}R=0, \quad \tilde{\nabla}S=0. \tag{1}$$

In Ref. 16 it is proved that *if* (M,g) *is connected, simply connected and geodesically complete then it admits a homogeneous pseudo-Riemannian structure if and only if it is a reductive homogeneous pseudo-Riemannian manifold*, which means that $M=G/H$, where G is a connected Lie group acting transitively and effectively on M as a group of isometries, H is the isotropy group at a point $o\in M$, and the Lie algebra $\mathfrak{g}$ of G may be decomposed into a vector space direct sum of the Lie algebra $\mathfrak{h}$ of H and an Ad(H)-invariant subspace $\mathfrak{m}$, that is $\mathfrak{g}=\mathfrak{h}\oplus\mathfrak{m}$, $\mathrm{Ad}(H)\mathfrak{m}\subset\mathfrak{m}$. (If G is connected and M is simply connected then H is connected, and the latter condition is equivalent to $[\mathfrak{h},\mathfrak{m}]\subset\mathfrak{m}$).

Let (M,g) be a connected, simply connected and geodesically complete pseudo-Riemannian manifold, and suppose that S is a homogeneous pseudo-Riemannian structure on (M,g). We fix a point $o\in M$ and put $\mathfrak{m}=T_o(M)$. If $\tilde{R}$ is the curvature tensor of the connection $\tilde{\nabla}=\nabla-S$, we can consider the holonomy algebra $\tilde{\mathfrak{h}}$ of $\tilde{\nabla}$ as the Lie subalgebra of antisymmetric endomorphisms of $(\mathfrak{m},g_o)$ generated by the operators $\tilde{R}_{ZW}$, where $Z,W\in\mathfrak{m}$. Then, according to the Ambrose–Singer construction,[14,15] a Lie bracket is defined in the vector space direct sum $\tilde{\mathfrak{g}}=\tilde{\mathfrak{h}}\oplus\mathfrak{m}$ by

$$\begin{aligned}
[U,V]&=UV-VU, \quad U,V\in\tilde{\mathfrak{h}},\\
[U,Z]&=U(Z), \quad U\in\tilde{\mathfrak{h}},\ Z\in\mathfrak{m}, \\
[Z,W]&=\tilde{R}_{ZW}+S_ZW-S_WZ, \quad Z,W\in\mathfrak{m},
\end{aligned}\tag{2}$$

and we say that $(\tilde{\mathfrak{g}},\tilde{\mathfrak{h}})$ is the *reductive pair* associated with the homogeneous pseudo-Riemannian structure S. The connected and simply connected Lie group $\tilde{G}$ whose Lie algebra is $\tilde{\mathfrak{g}}$ acts transitively on M as a group of isometries and $M\equiv\tilde{G}/\tilde{H}$, where $\tilde{H}$ is the connected Lie subgroup of $\tilde{G}$ whose Lie algebra is $\tilde{\mathfrak{h}}$. The set K of the elements of $\tilde{G}$ which act trivially on M is a discrete normal subgroup of $\tilde{G}$, and the Lie group $G=\tilde{G}/K$ acts transitively and effectively on M as a group of isometries, with isotropy group $H=\tilde{H}/K$. Then M is (diffeomorphic to) the reductive homogeneous pseudo-Riemannian manifold G/H.

On the other hand, we consider (see e.g., Ref. 18, p. 134) the Einstein–Yang–Mills equations on a Lorentzian manifold (M,g),

$$r-\tfrac{1}{2}sg+\Lambda g=\kappa T, \tag{3}$$

where g denotes the metric tensor, r the Ricci tensor, and s the scalar curvature of g, Λ and κ the cosmological and gravitational constants, respectively, and T the stress-energy tensor of the gauge field with field strength F, given by

$$T(X,Y)=g^{-1}(i_XF,i_YF)-\tfrac{1}{4}g(X,Y)\|F\|^2, \tag{4}$$

i denoting the interior product.

Let $\mathfrak{g}$ be the Lie algebra with generators P,X,Y,Q, and brackets

$$[X,Y]=P, \quad [Q,X]=Y, \quad [Q,Y]=-X.$$

The corresponding simply connected Lie group G is called the *oscillator group*.

III. LORENTZIAN HOMOGENEOUS STRUCTURES AND REDUCTIVE DECOMPOSITIONS

Consider the bi-invariant Lorentzian metric g on the oscillator group G given in the basis $\{P,X,Y,Q\}$, by

$$g=\begin{pmatrix} & & & 1\\ & 1 & & \\ & & 1 & \\ 1 & & & \end{pmatrix}. \tag{5}$$

Let $\{\eta,\alpha,\beta,\xi\}$ be the dual basis to $\{P,X,Y,Q\}$. The Levi-Civita connection ∇ of g is determined by $2g(\nabla_W U,V)=g([W,U],V)-g([U,V],W)+g([V,W],U)$ for all $U,V,W\in\mathfrak{g}$. So, we obtain that ∇ is given by

$$\nabla_W P=0,$$

$$\nabla_W X=-\tfrac{1}{2}\beta(W)P+\tfrac{1}{2}\xi(W)Y,$$

$$\nabla_W Y=\tfrac{1}{2}\alpha(W)P-\tfrac{1}{2}\xi(W)X,$$

$$\nabla_W Q=\tfrac{1}{2}\beta(W)X-\tfrac{1}{2}\alpha(W)Y,$$

for every $W\in\mathfrak{g}$. The nonvanishing components of the curvature tensor field, for which we adopt the convention $R_{WU}V=\nabla_{[W,U]}V-\nabla_W\nabla_U V+\nabla_U\nabla_W V$, are given by

$$R_{QX}X=-\tfrac{1}{4}P,\quad R_{QX}Q=\tfrac{1}{4}X,\quad R_{QY}Y=-\tfrac{1}{4}P,\quad R_{QY}Q=\tfrac{1}{4}Y.$$

Thus, the only nonvanishing component of the Ricci tensor is $r(Q,Q)=\tfrac{1}{2}$ and the scalar curvature is $s=0$.

In Ref. 12 all the homogeneous Lorentzian structures on the general oscillator group are determined by solving the Ambrose–Singer equations (1). As a consequence, we have all the homogeneous Lorentzian structures on the four-dimensional oscillator group.

Proposition: All the homogeneous Lorentzian structures on (G,g) *are given by*

$$S=\theta\otimes(\alpha\wedge\beta)+\rho\otimes(\alpha\wedge\xi)+\sigma\otimes(\beta\wedge\xi), \tag{6}$$

where ρ, σ, *and* θ *are left invariant 1-forms on G satisfying the conditions*

$$\begin{aligned}\tilde{\nabla}\theta&=0,\\ \tilde{\nabla}\rho&=\sigma\wedge\theta-\tfrac{1}{2}\alpha\otimes\theta+\tfrac{1}{2}\xi\otimes\sigma,\\ \tilde{\nabla}\sigma&=-\rho\wedge\theta-\tfrac{1}{2}\beta\otimes\theta-\tfrac{1}{2}\xi\otimes\rho,\end{aligned} \tag{7}$$

with $\tilde{\nabla}=\nabla-S$.

Proof: Expression (6) follows from the equations $\tilde{\nabla}g=0$ and $\tilde{\nabla}R=0$ in (1). Conditions (7) are equivalent to the equation $\tilde{\nabla}S=0$. (See Theorem 4.1 in Ref. 12.) □

By (6), we can write the following general expression for the homogeneous Lorentzian structures on (G,g):

$$\begin{aligned}S=&(a_1\eta+a_2\alpha+a_3\beta+a_4\xi)\otimes(a\wedge\beta)+(b_1\eta+b_2\alpha+b_3\beta+b_4\xi)\otimes(\alpha\wedge\xi)\\ &+(c_1\eta+c_2\alpha+c_3\beta+c_4\xi)\otimes(\beta\wedge\xi),\quad a_1,\dots,c_4\in\mathbb{R}.\end{aligned}$$

Equations (7) give us, after some long but straightforward calculations, the relations to be satisfied by the coefficients. If $a_1\neq 0$ (we denote this case as (i)), we get the expression of S in terms of a_1,a_2,a_3,a_4,

$$\begin{aligned}S=&(a_1\eta+a_2\alpha+a_3\beta+a_4\xi)\otimes(\alpha\wedge\beta)\\&+\left(-a_3\eta-\frac{a_2a_3}{a_1}\alpha-\left(\frac{a_3^2}{a_1}+\frac{1}{2}\right)\beta+\left(\frac{a_3}{2a_1}-\frac{a_3a_4}{a_1}\right)\xi\right)\otimes(\alpha\wedge\xi)\\&+\left(a_2\eta+\left(\frac{a_2^2}{a_1}+\frac{1}{2}\right)\alpha+\frac{a_2a_3}{a_1}\beta+\left(\frac{a_2a_4}{a_1}-\frac{a_2}{2a_1}\right)\xi\right)\otimes(\beta\wedge\xi).\end{aligned}$$

If $a_1=0$, some tedious computations lead to another five cases, and the coefficients in the six cases (i)–(vi) are as follows:

	(i)	(ii)	(iii)	(iv)	(v)	(vi)
a_1	$\neq 0$	0	0	0	0	0
a_2	$\in\mathbb{R}$	0	0	0	0	0
a_3	$\in\mathbb{R}$	0	0	0	0	0
a_4	$\in\mathbb{R}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\neq\frac{1}{2}$	$\neq\frac{1}{2}$
b_1	$-a_3$	0	0	0	0	0
b_2	$-\frac{a_2a_3}{a_1}$	0	0	$\in\mathbb{R}$	0	0
b_3	$-\frac{a_3^2}{a_1}-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\neq-\frac{1}{2}$	$-\frac{1}{2}$	$-a_4$
b_4	$\frac{a_3}{2a_1}-\frac{a_3a_4}{a_1}$	0	$\in\mathbb{R}$	$\in\mathbb{R}$	$\in\mathbb{R}$	0
c_1	a_2	0	0	0	0	0
c_2	$\frac{a_2^2}{a_1}+\frac{1}{2}$	$\neq\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}-\frac{2b_2^2}{2b_3+1}$	$\frac{1}{2}$	a_4
c_3	$\frac{a_2a_3}{a_1}$	0	0	$-b_2$	0	0
c_4	$-\frac{a_2}{2a_1}+\frac{a_2a_4}{a_1}$	$\in\mathbb{R}$	$\in\mathbb{R}$	$-\frac{2b_2b_4}{2b_3+1}$	$\in\mathbb{R}$	0

In case (i), making the change $x=a_1\neq 0$, $y=a_2/a_1$, $z=a_3/a_1$, $w=a_4-\frac{1}{2}$, we can write the family of homogeneous Lorentzian structures as

$$\begin{aligned}S_{(x,y,z,w)}=&(x\eta+xy\alpha+xz\beta+(w+\tfrac{1}{2})\xi)\otimes(\alpha\wedge\beta)\\&-(xz\eta+xyz\alpha+(xz^2+\tfrac{1}{2})\beta+zw\xi)\otimes(\alpha\wedge\xi)\\&+(xy\eta+(xy^2+\tfrac{1}{2})\alpha+xyz\beta+yw\xi)\otimes(\beta\wedge\xi),\\&x,y,z,w\in\mathbb{R},\quad x\neq 0.\end{aligned}$$

In cases (ii)–(vi) one arrives (after some changes of notations for the coefficients) to the following families:

(ii)

$$S_{(q,c)} = \tfrac{1}{2}\xi \otimes (\alpha \wedge \beta) - \tfrac{1}{2}\beta \otimes (\alpha \wedge \xi) + (q\alpha + c\xi) \otimes (\beta \wedge \xi), \quad q,c \in \mathbb{R}, \quad q \neq \tfrac{1}{2}.$$

(iii)

$$S_{(b,c)} = \tfrac{1}{2}\xi \otimes (\alpha \wedge \beta) + (-\tfrac{1}{2}\beta + b\xi) \otimes (\alpha \wedge \xi) + (\tfrac{1}{2}\alpha + c\xi) \otimes (\beta \wedge \xi), \quad b,c \in \mathbb{R}.$$

(iv)

$$S_{(k,t,b)} = \tfrac{1}{2}\xi \otimes (\alpha \wedge \beta) + (k\alpha + (t - \tfrac{1}{2})\beta + b\xi) \otimes (\alpha \wedge \xi)$$
$$+ ((\tfrac{1}{2} - (k^2/t))\alpha - k\beta - (kb/t)\xi) \otimes (\beta \wedge \xi), \quad k,t,b \in \mathbb{R}, \quad t \neq 0.$$

(v)

$$S_{(a,b,c)} = a\xi \otimes (\alpha \wedge \beta) + (-\tfrac{1}{2}\beta + b\xi) \otimes (\alpha \wedge \xi) + (\tfrac{1}{2}\alpha + c\xi) \otimes (\beta \wedge \xi),$$
$$a,b,c \in \mathbb{R}, \quad a \neq \tfrac{1}{2}.$$

(vi)

$$S_a = a(\xi \otimes (\alpha \wedge \beta) - \beta \otimes (\alpha \wedge \xi) + \alpha \otimes (\beta \wedge \xi)), \quad a \in \mathbb{R}, \quad a \neq \tfrac{1}{2}.$$

Next, we determine $\tilde{\nabla} = \nabla - S$. We however omit the expression of $\tilde{\nabla}$ for the sake of brevity. Then, in each case we calculate the curvature $\tilde{R}$, and so we obtain generators of the holonomy algebra $\tilde{\mathfrak{h}}$ of $\tilde{\nabla}$, which will be expressed in terms of the basis $\{P,X,Y,Q\}$ of $\mathfrak{g}$.

In case (i), we have that the only not always null curvature operators are $\tilde{R}_{XY} = wU$, $\tilde{R}_{QX} = xzU$ and $\tilde{R}_{QY} = -xyU$, where

$$U = \begin{pmatrix} 0 & z & -y & 0 \\ 0 & 0 & 1 & -z \\ 0 & -1 & 0 & y \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

If $y = z = w = 0$ then $\tilde{R} = 0$, but if either y, z, or w is not zero then the holonomy algebra $\tilde{\mathfrak{h}}$ is one-dimensional and generated by U.

In case (ii), the holonomy algebra $\tilde{\mathfrak{h}}$ is generated by the vector

$$V = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and the only nonvanishing curvature operator is $\tilde{R}_{QY} = (q - \tfrac{1}{2})V$.

In case (iii) (and (v)) we deduce $\tilde{R} = 0$.

In case (iv), the holonomy algebra $\tilde{\mathfrak{h}}$ is generated by

$$U=\begin{pmatrix} 0 & -t & k & 0\\ 0 & 0 & 0 & t\\ 0 & 0 & 0 & -k\\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and the curvature of $\tilde{\nabla}$ is determined by $\tilde{R}_{QX}=U$ and $\tilde{R}_{QY}=-(k/t)U$.

In case (vi), the holonomy algebra $\tilde{\mathfrak{h}}$ has two generators

$$U=\begin{pmatrix} 0 & 1 & 0 & 0\\ 0 & 0 & 0 & -1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad V=\begin{pmatrix} 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -1\\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and the curvature is determined by the operators $\tilde{R}_{QX}=(a^2-\frac{1}{4})U$, $\tilde{R}_{QY}=(a^2-\frac{1}{4})V$. If $a=-\frac{1}{2}$ then $\tilde{R}=0$ and if $a\neq-\frac{1}{2}$, then $\tilde{\mathfrak{h}}$ is two-dimensional and Abelian.

Through the usual identification of $\mathfrak{g}$ with the tangent space at the identity of G, we consider $\mathfrak{m}=\mathfrak{g}$ and use (2) in order to obtain the Lie bracket of $\tilde{\mathfrak{g}}=\tilde{\mathfrak{h}}\oplus\mathfrak{g}$ in each case. In particular, we will have the reductive pair $(\tilde{\mathfrak{g}},\tilde{\mathfrak{h}})$ associated with each homogeneous structure.

We next explicitly give those brackets. Some cases are broken in two subcases according to whether the dimension of the holonomy algebra depends on the particular values of the coefficients in the corresponding homogeneous structure.

Case (i$_1$). $(x,y,z,w\in\mathbb{R};\ x\neq 0;\ y,z,w$ not all null). The reductive decomposition associated with $S_{(x,y,z,w)}$ is $\tilde{\mathfrak{g}}_{(x,y,z,w)}=\tilde{\mathfrak{h}}\oplus\mathfrak{g}=\langle\{U,P,X,Y,Q\}\rangle$, with not always null brackets

$$[U,X]=zP-Y,\quad [U,Y]=-yP+X,\quad [U,Q]=-zX+yY,$$

$$[P,X]=-xzP+xY,\quad [P,Y]=xyP-xX,\quad [P,Q]=xzX-xyY,$$

$$[X,Y]=wU+(x(y^2+z^2)+1)P-xyX-xzY,$$

$$[Q,X]=xzU-zwP-xyzX+(w+xy^2+1)Y,$$

$$[Q,Y]=-xyU+ywP-(w+xz^2+1)X+xyzY.$$

Case (i$_2$) $(x\neq 0;\ y=z=w=0)$. The reductive decomposition associated with $S_{(x,0,0,0)}$ is $\tilde{\mathfrak{g}}_x=0\oplus\mathfrak{g}=\langle\{P,X,Y,Q\}\rangle$, with nonvanishing brackets

$$[P,X]=xY,\quad [P,Y]=-xX,\quad [X,Y]=P,\quad [Q,X]=Y,\quad [Q,Y]=-X.$$

Case (ii) $(q,c\in\mathbb{R};\ q\neq\frac{1}{2})$. The reductive decomposition associated to $S_{(q,c)}$ is $\tilde{\mathfrak{g}}_{(q,c)}=\tilde{\mathfrak{h}}\oplus\mathfrak{g}=\langle\{V,P,X,Y,Q\}\rangle$, with nonvanishing brackets

$$[V,Y]=P,\quad [V,Q]=-Y,\quad [X,Y]=(q+\tfrac{1}{2})P,$$

$$[Q,X]=(q+\tfrac{1}{2})Y,\quad [Q,Y]=(q-\tfrac{1}{2})V+cP-X.$$

Case (iii) $(b,c,\in\mathbb{R})$. The reductive decomposition associated with $S_{(b,c)}$ is $\tilde{\mathfrak{g}}_{(b,c)}=0\oplus\mathfrak{g}=\langle\{P,X,Y,Q\}\rangle$, with nonvanishing brackets

$$[X,Y]=P,\quad [Q,X]=bP+Y,\quad [Q,Y]=cP-X.$$

In this case, one obtains the Lie algebra of the oscillator group. In fact, putting $\tilde{X}=X-cP$, $\tilde{Y}=Y+bP$, we have $[\tilde{X},\tilde{Y}]=P$, $[Q,\tilde{X}]=\tilde{Y}$, $[Q,\tilde{Y}]=-\tilde{X}$.

Case (iv) $(k,t,b\in\mathbb{R};\ t\neq 0)$. The reductive decomposition associated with $S_{(k,t,b)}$ is $\tilde{\mathfrak{g}}_{(k,t,b)}=\tilde{\mathfrak{h}}\oplus\mathfrak{g}=\langle\{U,P,X,Y,Q\}\rangle$, with Lie brackets given by

$$[U,X]=-tP,\quad [U,Y]=kP,\quad [U,Q]=tX-kY,\quad [X,Y]=(1-(k^2+t^2)/t)P,$$

$$[Q,X]=U+bP+kX+\left(1-\frac{k^2}{t}\right)Y,\quad [Q,Y]=-\frac{k}{t}U-\frac{kb}{t}P+(t-1)X-kY.$$

Case (v) $(a,b,c\in\mathbb{R};\ a\neq\frac{1}{2})$. The reductive decomposition associated to $S_{(a,b,c)}$ *is* $\tilde{\mathfrak{g}}_{(a,b,c)}=0\oplus\mathfrak{g}=\langle\{P,X,Y,Q\}\rangle$, with nonvanishing brackets

$$[X,Y]=P,\quad [Q,X]=bP+(a+\tfrac{1}{2})Y,\quad [Q,Y]=cP-(a+\tfrac{1}{2})X.$$

Case (vi_1) $(a\in\mathbb{R};\ a\neq\frac{1}{2},-\frac{1}{2})$. The reductive decomposition associated with S_a is $\tilde{\mathfrak{g}}_a=\tilde{\mathfrak{h}}\oplus\mathfrak{g}=\langle\{U,V,P,X,Y,Q\}\rangle$, with nonvanishing brackets

$$[U,X]=P,\quad [U,Q]=-X,\quad [V,Y]=P,\quad [V,Q]=-Y,\quad [X,Y]=2aP,$$

$$[Q,X]=(a^2-\tfrac{1}{4})U+2aY,\quad [Q,Y]=(a^2-\tfrac{1}{4})V-2aX.$$

In particular, if $a=0$, one has $S=0$, and the associated reductive pair $(\tilde{\mathfrak{g}},\tilde{\mathfrak{h}})$ is a symmetric pair, which defines the oscillator group as a symmetric Lorentzian space.

Case (vi_2) $(a=-\frac{1}{2})$. The reductive decomposition associated to $S_{-(1/2)}$ is $\tilde{\mathfrak{g}}_{-(1/2)}=0\oplus\mathfrak{g}=\langle\{P,X,Y,Q\}\rangle$, with nonvanishing brackets

$$[X,Y]=-P,\quad [Q,X]=-Y,\quad [Q,Y]=X,$$

which also gives the Lie algebra of the oscillator group.

IV. CURRENTS, AND EINSTEIN–YANG–MILLS EQUATIONS

We consider the connection form ω on $\tilde{\mathfrak{g}}$, which is the $\tilde{\mathfrak{h}}$-component of the canonical 1-form on $\tilde{\mathfrak{g}}$, and we obtain, in each case, the curvature form $F=d\omega$ and then $\star F$, where $\star$ denotes the Hodge operator with respect to the Lorentzian metric g. The following table provides the values of F and $\star F$ for the cases with nontrivial holonomy,

	$F=d\omega$	$\star F$
(i_1)	$(-w\alpha\wedge\beta+xz\alpha\wedge\xi-xy\beta\wedge\xi)\otimes U$	$-(w\eta\wedge\xi+xz\eta\wedge\beta+xy\eta\wedge\alpha)\otimes U$
(ii)	$(q-\frac{1}{2})(\beta\wedge\xi)\otimes V$	$(q-\frac{1}{2})(\eta\wedge\alpha)\otimes V$
(iv)	$(\alpha\wedge\xi-(k/t)\beta\wedge\xi)\otimes U$	$-(\eta\wedge\beta+(k/t)\eta\wedge\alpha)\otimes U$
(vi_1)	$(a^2-\frac{1}{4})((\alpha\wedge\xi)\otimes U+(\beta\wedge\xi)\otimes V)$	$(a^2-\frac{1}{4})(-(\eta\wedge\beta)\otimes U+(\eta\wedge\alpha)\otimes V)$

Moreover, as the second Yang-Mills equation is $D\star F=J$, where J denotes the current, applying the usual formula[19] one obtains for all the cases

$$J=D\star F=d\star F+\omega\wedge\star F-\star F\wedge\omega=d\star F. \tag{8}$$

Thus, the currents are given, respectively, by

	J
(i_1)	$(xz\eta\wedge\alpha\wedge\xi-xy\eta\wedge\beta\wedge\xi+w\alpha\wedge\beta\wedge\xi)\otimes U$
(ii)	$(q-\frac{1}{2})(\eta\wedge\beta\wedge\xi+c\alpha\wedge\beta\wedge\xi)\otimes V$
(iv)	$(\eta\wedge\alpha\wedge\xi-(k/t)\eta\wedge\beta\wedge\xi+b(1+(k/t)^2)\alpha\wedge\beta\wedge\xi)\otimes U$
(vi_1)	$2a(a^2-\frac{1}{4})((\eta\wedge\alpha\wedge\xi)\otimes U+(\eta\wedge\beta\wedge\xi)\otimes V)$

On the other hand, we consider the metric on $\tilde{\mathfrak{g}}$ which is the direct product of the metric g_0 on $\mathfrak{g}$ and the metric on $\tilde{\mathfrak{h}}$ for which the basic vectors (U and/or V) are unitary.

Now we have $\|F\|^2=2w^2$ in the case (i_1) and $\|F\|^2=0$ in the remaining cases. We next compute the stress-energy tensor T, given by (4), and we express it in terms of the basis $\{P,X,Y,Q\}$ of $\mathfrak{g}$. In case (i_1) we have

$$T=\begin{pmatrix} 0 & 0 & 0 & -\frac{1}{2}w^2 \\ 0 & \frac{1}{2}w^2 & 0 & -xyw \\ 0 & 0 & \frac{1}{2}w^2 & -xzw \\ -\frac{1}{2}w^2 & -xyw & -xzw & x^2(y^2+z^2) \end{pmatrix}.$$

In cases (ii), (iv), and (vi_1), the only nonvanishing component is $T(Q,Q)$, which is, respectively, $(q-\frac{1}{2})^2$, $1+(k/t)^2$, and $2(a^2-\frac{1}{4})^2$.

It is then a matter of calculation to see that the field equations (3) are satisfied in cases (ii), (iv), and (vi_1), and in the case (i_1) if $w=0$ and $y^2+z^2\neq 0$, and to obtain the corresponding cosmological and gravitational constants.

Theorem: *The Lorentzian reductive pairs* $(\tilde{\mathfrak{g}},\tilde{\mathfrak{h}})$ *determined by the homogeneous Lorentzian structures on the four-dimensional oscillator group, in cases* (i_1) *(with* $w=0$ *and* $y^2+z^2\neq 0$*), (ii), (iv) and (vi₁) above, are solutions of the Einstein–Yang–Mills equations for an electromagnetic gauge field with cosmological constant* $\Lambda=0$*, and gravitational constant* κ *given, respectively, by*

$$\frac{1}{2x^2(y^2+z^2)},\quad \frac{1}{2(q-\frac{1}{2})^2},\quad \frac{t^2}{2(k^2+t^2)},\quad \frac{1}{4(a^2-\frac{1}{4})^2}.$$

The gauge algebra $\tilde{\mathfrak{h}}$ *is Abelian and two-dimensional in the last case and one-dimensional in the other three cases. The currents are given in the table following (8). They are nonvanishing except in the case* (vi_1) *for* $a=0$.

If $a=0$ in case (vi_1), the associated symmetric pair $(\tilde{\mathfrak{g}},\tilde{\mathfrak{h}})$ appears as a solution of the Einstein–Yang–Mills equations for a sourceless gauge field with cosmological constant $\Lambda=0$ and gravitational constant $\kappa=4$ (see Levichev[5]).

ACKNOWLEDGMENTS

This work has been partially supported by DGES (Spain) under Project No. PB95-0124, and by Xunta de Galicia, under Project XUGA No. 20703B98.

[1] R. F. Streater, ''The representations of the oscillator group,'' Commun. Math. Phys. **4**, 217–236 (1967).
[2] T. Nomura, ''The Paley–Wiener theorem for the oscillator group,'' J. Math. Kyoto Univ. **22**, 71–96 (1982/83).
[3] A. Medina, ''Groupes de Lie munis de métriques biinvariantes,'' Tôhoku Math. J. **37**, 405–421 (1985).
[4] A. Medina and P. Revoy, ''Les groupes oscillateurs et leurs réseaux,'' Manuscr. Math. **52**, 81–95 (1985).
[5] A. V. Levichev, ''Several symmetric spaces of general relativity theory as solutions of Einstein–Yang–Mills equations,'' in *Group Theoretical Methods in Physics*, Proceedings of the 3rd International Seminar Yurmala, May, 1985 (Nauka, Moscow, 1986), Vol. 1, pp. 145–150 (in Russian).
[6] A. V. Levichev, ''Chronogeometry of an electromagnetic wave given by a biinvariant metric on the oscillator group,'' Siberian Math. J. **27**, 237–245 (1986).
[7] D. Müller and F. Ricci, ''On the Laplace–Beltrami operator on the oscillator group,'' J. Reine Angew. Math. **390**, 193–207 (1988).
[8] A. V. Levichev, ''Methods of investigation of the causal structure of homogeneous Lorentz manifolds,'' Siberian Math. J. **31**, 395–408 (1990).
[9] N. Dörr, ''A note on the oscillator group,'' Seminar Sophus Lie **2**, 31–38 (1991).
[10] O. S. Svidersky, ''Oscillator parallelization of the scalar induced bundle,'' Siberian Math. J. **36**, 965–972 (1995).
[11] O. S. Svidersky, ''Decomposition of a vector field in a parallelized induced bundle,'' Proc. Am. Math. Soc. **125**, 2485–2491 (1997).

[12] P. M. Gadea and J. A. Oubiña, ''Homogeneous Lorentzian structures on the oscillator groups,'' Arch. Math. (Basel) (to be published).
[13] É. Cartan, ''Sur une classe remarquable d'espaces de Riemann,'' Bull. Soc. Math. France **54**, 214–264 (1926).
[14] W. Ambrose and I. M. Singer, ''On homogeneous Riemannian manifolds,'' Duke Math. J. **25**, 647–669 (1958).
[15] F. Tricerri and L. Vanhecke, *Homogeneous Structures on Riemannian Manifolds*, London Math. Soc. Lect. Notes Ser. **83** (Cambridge University Press, Cambridge, 1983).
[16] P. M. Gadea and J. A. Oubiña, ''Homogeneous pseudo-Riemannian structures and homogeneous almost para-Hermitian structures,'' Houston J. Math. **18**, 449–465 (1992).
[17] P. M. Gadea and J. A. Oubiña, ''Reductive homogeneous pseudo-Riemannian manifolds,'' Monatsh. Math. **124**, 17–34 (1997).
[18] D. Bleecker, *Gauge Theories and Variational Principles* (Addison–Wesley, Reading, 1981).
[19] T. Eguchi, P. B. Gilkey, and A. J. Hanson, ''Gravitation, gauge theories and differential geometry,'' Phys. Rep. (Review sect. of Phys. Lett.) **66**, 213–393 (1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Radial conformal motions in Minkowski space–time

Alicia Herrero[a] and Juan Antonio Morales[b]
Departament d'Astronomia i Astrofísica, Universitat de València, E-46100 Burjassot, València, Spain

(Received 11 January 1999; accepted for publication 5 March 1999)

A study of radial conformal Killing fields (RCKF) in Minkowski space–time is carried out, which leads to their classification into three disjointed classes. Their integral curves are straight or hyperbolic lines admitting orthogonal surfaces of constant curvature, whose sign is related to the causal character of the field. Otherwise, the kinematic properties of the timelike RCKF are given and their applications in kinematic cosmology is discussed. © *1999 American Institute of Physics.* [S0022-2488(99)00507-1]

I. INTRODUCTION

The study of vector (and tensor) fields in a Lorentzian metric is a key issue, both from the theoretical and practical points of view. Infinitesimal transformations, fluid flows, eigendirections of a given 2-tensor field, critical points, continuous symmetries, directions attached to coordinate systems, light propagation and polarization in a medium, geodesic and accelerated observers are some examples of basic concepts which are described using vector fields. This work is devoted to analyzing in the Minkowski space–time the main properties of a particular type of fields that we have called *radial* conformal motions.

There are several reasons for carrying out such a study: (i) homothetic and hyperbolic radial motions belong to this kind of fields, (ii) conserved quantities along null geodesics are obtained from conformal Killing vectors, with particular expressions for the radial case, (iii) isotropic distribution functions of photons verifying the Liouville equation can be built from these conserved quantities and, (iv) this study can be easily extended to any conformally flat space–time and used to obtain its conformal factor imposing a given kinematic property of the field (geodesic, homogeneous expansion, etc.); in fact, Infeld–Schild work on kinematic cosmology[1] tacitly involves the concept of timelike radial conformal Killing field (RCKF) in the geodesic case.

Firstly, in Sec. II, we introduce the concept of RCKF and obtain its general expression and the type of subalgebra generated by them; we also present a study of their causal character according to the different domains of the space–time. We continue with a classification of these fields related to the sign of a quantity invariant by internal conformal transformations of the Minkowski metric (Sec. III). The associated integral curves are plotted in Sec. IV and we show that the orthogonal hypersurfaces of the field have constant curvature whose sign is related to its causal character (Sec. V). In Sec. VI, we focus on timelike RCKF and discuss their kinematic properties pointing out their connection with the Milne's interpretation of the cosmological recession velocity in the Minkowski space–time.[2] Finally, in Sec. VII, we comment on several applications of the present study. Some of these results have been presented, without proof, in the E.R.E., annual Spanish relativity meeting.[3]

II. RADIAL CONFORMAL KILLING FIELDS

Let us consider a radial vector field

$$\xi=\alpha(t,r,\theta,\varphi)\frac{\partial}{\partial t}+\beta(t,r,\theta,\varphi)\frac{\partial}{\partial r}$$

[a]Electronic mail: alicia.herrero@uv.es
[b]Electronic mail: antonio.morales@uv.es

in the Minkowski space–time,

$$\eta = -dt\otimes dt + dr\otimes dr + r^2 h, \tag{1}$$

with $h = d\theta\otimes d\theta + \sin^2\theta\, d\varphi\otimes d\varphi$, the metric on the 2-sphere.

The equation $\mathcal{L}_\xi \eta \propto \eta$ expresses that ξ is a conformal Killing field (or conformal motion) of η, where $\mathcal{L}_\xi$ represents the Lie derivative with respect to ξ. This condition leads to that the functions α and β are independent of the angular coordinates θ and φ, resulting in

$$\alpha(t,r) = a(t^2+r^2) + bt + c, \quad \beta(t,r) = r(2at+b), \tag{2}$$

where a, b, and c are arbitrary constants.

Proposition 1: In the Minkowski space–time, the general form of a RCKF is

$$\xi = (a(t^2+r^2)+bt+c)\frac{\partial}{\partial t} + r(2at+b)\frac{\partial}{\partial r}, \tag{3}$$

with a, b, and c as arbitrary constants.

Consequently, ξ can be obtained as linear combination of (the generators of) the timelike translation $\xi_1 = (\partial/\partial t)$, the dilation $\xi_2 = t(\partial/\partial t) + r(\partial/\partial r)$, and the special nonlinear conformal transformation along the t-axis $\xi_3 = (t^2+r^2)(\partial/\partial t) + 2tr(\partial/\partial r)$, that is

$$\xi = a\xi_3 + b\xi_2 + c\xi_1 .$$

The Lie brackets of these generators,

$$[\xi_1,\xi_2] = \xi_1, \quad [\xi_1,\xi_3] = 2\xi_2, \quad [\xi_2,\xi_3] = \xi_3$$

give us the type of the Lie algebra generated by RCKF. In fact, if we consider the vector fields

$$e_1 = \xi_1 - \tfrac{1}{4}\xi_3, \quad e_2 = \xi_2, \quad e_3 = -\xi_1 - \tfrac{1}{4}\xi_3,$$

then, their commutation relations are

$$[e_1,e_2] = -e_3, \quad [e_2,e_3] = e_1, \quad [e_3,e_1] = e_2$$

showing that this Lie algebra is isomorphic to the pseudo-orthogonal algebra AO(1,2); then we have the following result:

Proposition 2: RCKF generate a three-dimensional Lie algebra of Bianchi type VIII.

Note that using null coordinates, $u = t+r$ and $v = t-r$, expression (3) is written in a completely symmetric form

$$\xi = (au^2+bu+c)\frac{\partial}{\partial u} + (av^2+bv+c)\frac{\partial}{\partial v}, \tag{4}$$

where each null coordinate appears separately and in the same way in the corresponding component of the field ξ.

From (3) and (4) we have

$$P \equiv -\eta(\xi,\xi) = [a(t^2+r^2)+bt+c]^2 - r^2(2at+b)^2 = (au^2+bu+c)(av^2+bv+c). \tag{5}$$

The discussion of the sign of P will give the causal character of ξ in the different regions of the space–time (domains of causality of ξ). It is convenient to introduce the quantity

$$\Delta \equiv b^2 - 4ac$$

that make this discussion easier. The results are shown in Table I and plotted in Fig. 1.

TABLE I. Causal character of the field ξ given by (3) or (4) for the different values of the constants a and $\Delta=b^2-4ac$.

		Causal character of a radial conformal Killing vector
$a=0$	$b=0$	timelike everywhere
	$b\neq 0$	null on the light cone at the point $(t=-(c/b),\ r=0)$, timelike inside of the light cone and spacelike outside of the light cone. See Fig. 1(i)
$a\neq 0$	$\Delta<0$	timelike everywhere
	$\Delta=0$	timelike everywhere except for the light cone on $(t=-(b/2a),\ r=0)$ where it is null. See Fig. 1(ii)
	$\Delta>0$	null on the light cones at the points $(t_\pm, r=0)$, $t_\pm=(-b\pm\sqrt{\Delta})/2a$, timelike inside or outside of the two light cones and spacelike in other domains. See Fig. 1(iii)

III. A CLASSIFICATION OF THE RADIAL CONFORMAL MOTIONS

In order to classify the RCKF in equivalence classes, it is convenient to take in account the degree of freedom of the null coordinates $\{u,v\}$. Then we consider the coordinates

$$\overline{u}=\overline{u}(u),\quad \overline{v}=\overline{v}(v),\quad \overline{\theta}=\theta,\quad \overline{\varphi}=\varphi$$

verifying the condition

$$\overline{u}_u\overline{v}_v=\left(\frac{\overline{u}-\overline{v}}{u-v}\right)^2 \tag{6}$$

to obtain a conformally flat form of the Minkowski metric (1),

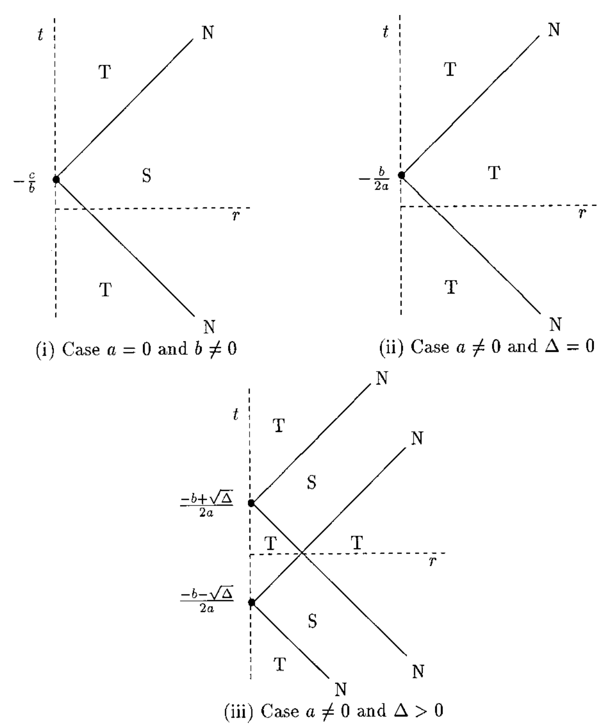

FIG. 1. Domains of causality of a RCKF ξ according to the values of the coefficient a and $\Delta=b^2-4ac$. The different possibilities for the causal character, Timelike, Spacelike or Null, in each domain, are abbreviated with the capital letters T, S, or N, respectively.

$$\eta = F(\bar{u},\bar{v})\left[-\frac{1}{2}(d\bar{u}\otimes d\bar{v}+d\bar{v}\otimes d\bar{u})+\left(\frac{\bar{u}-\bar{v}}{2}\right)^2 h\right],$$

where the *internal conformal factor, F,* is

$$F=\frac{1}{\bar{u}_u\bar{v}_v}=u_{\bar{u}}v_{\bar{v}} \tag{7}$$

with subindexes denoting derivation with respect to the coordinates.

The integration of Eq. (6) with respect to $\bar{u}(u)$, considering the v-coordinate as a parameter, gives

$$\bar{u}(u)=\frac{\bar{v}_v(u-v)}{1+(u-v)A}+\bar{v}, \tag{8}$$

where A is an arbitrary function of v. Now, taking into account that $\bar{u}$ does not depend on v, the derivative of (8) with respect to v gives the following system of equations for A and $\bar{v}$:

$$\begin{cases} A\bar{v}_{vv}-A_v\bar{v}_v+A^2\bar{v}_v=0\\ 2A\bar{v}_v+\bar{v}_{vv}=0. \end{cases} \tag{9}$$

If $A=0$ the solution of Eqs. (8) and (9) is linear, $\bar{u}(u)=pu+q$ and $\bar{v}(v)=pv+q$, where $p\neq 0$ and q are arbitrary constants. In the generic case, $A\neq 0$, the solution has the form

$$\bar{u}(u)=\frac{p}{u+q}+m,\quad \bar{v}(v)=\frac{p}{v+q}+m, \tag{10}$$

where $p\neq 0$, q and m are arbitrary constants; the internal conformal factor F results from Eq. (7),

$$F(\bar{u},\bar{v})=\frac{p^2}{(\bar{u}-m)^2(\bar{v}-m)^2}.$$

Therefore, taking $\bar{u}=\bar{t}+\bar{r}$ and $\bar{v}=\bar{t}-\bar{r}$, we obtain the following proposition:

Proposition 3: The nonlinear coordinate transformations $\bar{t}=\bar{t}(t,r)$, $\bar{r}=\bar{r}(t,r)$, *that maintain invariant the diagonal form of the Minkowski metric, except for an internal conformal factor F are given by*

$$\bar{t}=\frac{-p(t+q)}{r^2-(t+q)^2}+m,\quad \bar{r}=\frac{pr}{r^2-(t+q)^2}$$

with $p\neq 0$, q *and* m *as arbitrary constants. Then, this factor is*

$$F(\bar{t},\bar{r})=\frac{p^2}{(\bar{r}^2-(\bar{t}-m)^2)^2}.$$

Note that, these *internal conformal transformations* in the Minkowski space–time, given by (10), also maintain invariant the form (4) of the RCKF, that is

$$\xi=(\bar{a}\bar{u}^2+\bar{b}\bar{u}+\bar{c})\frac{\partial}{\partial\bar{u}}+(\bar{a}\bar{v}^2+\bar{b}\bar{v}+\bar{c})\frac{\partial}{\partial\bar{v}},$$

where the new constants $\bar{a}$, $\bar{b}$, and $\bar{c}$ are obtained from the following matricial relation:

$$\begin{pmatrix}\overline{a}\\ \overline{b}\\ \overline{c}\end{pmatrix}=\mathcal{M}\begin{pmatrix}a\\ b\\ c\end{pmatrix}\quad\text{with}\quad \mathcal{M}\equiv\frac{1}{p}\begin{pmatrix}-q^2 & q & -1\\ 2q(p+mq) & -(p+2mq) & 2m\\ -(p+mq)^2 & m(p+mq) & -m^2\end{pmatrix}. \tag{11}$$

Moreover $b^2-4ac=\overline{b}^2-4\overline{ac}$ and we have the following proposition:

Proposition 4: The form of a RCKF is invariant by the internal conformal transformations given in Proposition 3. Moreover, the quantity $\Delta=b^2-4ac$ *is also invariant by these transformations.*

Since $\det(\mathcal{M})=1$, the internal conformal transformations of the Minkowski metric are represented as orthogonal transformations on the algebra of the RCKF, considering its Killing form $\mathcal{K}$ as a metric. Hence, Δ is invariantly defined from the scalar product associated with $\mathcal{K}$, that is $\Delta=\mathcal{K}(\xi,\xi)$. The invariance of this quantity suggests us the possibility of a classification of the RCKF depending on the sign of Δ, which will be used to denote these classes. The classes $\Delta=0$, $\Delta>0$ and $\Delta<0$ can be represented by the fields ξ_1, ξ_2, and $\xi_3+\xi_1$, respectively.

Note that the fields ξ_1 and ξ_3 belong to the class $\Delta=0$ because the internal conformal transformation from Proposition 3 with $p=-1$, $q=0$ and m arbitrary lets us write the field ξ_3 as a timelike translation field in the new coordinates

$$\xi_3=(t^2+r^2)\frac{\partial}{\partial t}+2tr\frac{\partial}{\partial r}=\frac{\partial}{\partial \overline{t}}\equiv\overline{\xi_1}$$

as it follows from expression (11). The metric η in these coordinates has the form

$$\eta_{(\overline{t},\overline{r},\theta,\varphi)}=\frac{1}{[\overline{r}^2-(\overline{t}-m)^2]^2}\operatorname{diag}(-1,1,\overline{r}^2,\overline{r}^2\sin^2\theta). \tag{12}$$

Another interesting example is the equivalence between the fields ξ_2 and $\xi_3-\xi_1$ in the class $\Delta>0$. From Proposition 3 and Eq. (11), an internal conformal transformation with $p=-2m\neq 0$ and $q=1$ lets us write $\xi_3-\xi_1$ as the dilation field in the new coordinates $(\overline{t},\overline{r})$,

$$\xi_3-\xi_1=(t^2+r^2-1)\frac{\partial}{\partial t}+2tr\frac{\partial}{\partial r}=\overline{t}\frac{\partial}{\partial \overline{t}}+\overline{r}\frac{\partial}{\partial \overline{r}}\equiv\overline{\xi_2}$$

and now the metric η is written in the form,

$$\eta_{(\overline{t},\overline{r},\theta,\varphi)}=\frac{4m^2}{[\overline{r}^2-(\overline{t}-m)^2]^2}\operatorname{diag}(-1,1,\overline{r}^2,\overline{r}^2\sin^2\theta). \tag{13}$$

Then, as it is shown in last examples, we have the following result:

Proposition 5: The fields $\xi_3-k\xi_1$, *with* $k=0,+1,-1$, *may be taken as representatives of the equivalence classes* $\Delta=0$, $\Delta>0$, *and* $\Delta<0$, *respectively.*

Note that the representatives taken in the above proposition are obtained only from the fields ξ_1 and ξ_3, and then, they are not a basis of the radial conformal Killing algebra. But they generate by commutation the complete algebra, because ξ_2 is, up to a constant factor, the Lie bracket of any pair of these representatives.

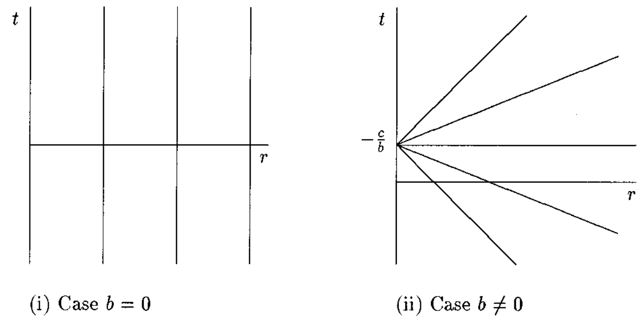

FIG. 2. Integral curves associated with a RCKF ξ given by expression (3) with $a=0$.

IV. INTEGRAL CURVES ASSOCIATED WITH A RADIAL CONFORMAL MOTION

The integral curves of a RCKF ξ given by (3) are the solution of the differential equation

$$\frac{dt}{a(t^2+r^2)+bt+c}=\frac{dr}{r(2at+b)}, \tag{14}$$

which has the following implicit form:

$$a(t^2-r^2)+bt-\omega r+c=0. \tag{15}$$

So, the integral curves are a one-parameter family of straight or hyperbolic lines, depending on the constants a, b, and c of the field and on the parameter ω. When $a=0$, Eq. (15) represents straight lines in the $\{t,r\}$-plane (Fig. 2). When $a\neq 0$, Eq. (15) can be written in the form

$$\left(t+\frac{b}{2a}\right)^2-\left(r+\frac{\omega}{2a}\right)^2=\frac{\Delta-\omega^2}{4a^2},$$

which represents a hyperbolic line for each value of the parameter ω except when $\omega^2=\Delta$ that corresponds to the light cone at the point $(t=-b/2a,\ r=-\omega/2a)$. The vertexes of each hyperbola are the points $[t=-(b/2a),\ r_{\pm}=(-\omega\pm\sqrt{\omega^2-\Delta})/2a]$ if $\Delta\leqslant 0$ or $\omega^2>\Delta>0$ and the points $[t_{\pm}=(-b\pm\sqrt{\Delta-\omega^2})/2a,\ r=-\omega/2a]$ if $\Delta>0$ and $\omega^2<\Delta$. We must consider only the part of the hyperbolic branches with $r>0$. Some of these integral curves are plotted in Fig. 3 for the different values of Δ. Note that in the case $\Delta>0$ there exist a double family of hyperbolic lines, Fig. 3(iii).

The vector field $\xi_2=t(\partial/\partial t)+r(\partial/\partial r)$ is called (the generator of) a dilation transformation since its integral curves are a radial congruence of straight lines [Figs. 1(i) and 2(ii)]. And the vector field $\xi_3=(t^2+r^2)(\partial/\partial t)+2tr(\partial/\partial r)$ is identified as (the generator of) an acceleration transformation along the t-axis, because each integral curve for $\omega\neq 0$ can be seen as a hyperbolic relativistic motion whose acceleration $\mathbf{a}$ has constant length, $|\mathbf{a}|=2|a/\omega|$ [Figs. 1(ii) and 3(ii)].

V. ORTHOGONAL SURFACES TO A RADIAL CONFORMAL KILLING FIELD

Let us consider the covector ξ_* associated by the metric to a RCKF ξ given by (3), that is

$$\xi_*=-(a(t^2+r^2)+bt+c)dt+r(2at+b)dr.$$

This 1-form is integrable, $\xi_*\wedge d\xi_*=0$, and admits as a potential the following function:

$$s(t,r)=\begin{cases} b(t^2-r^2)+2ct & \text{if } a=0\\ \dfrac{a(t^2-r^2)-c}{2at+b} & \text{if } a\neq 0\end{cases} \tag{16}$$

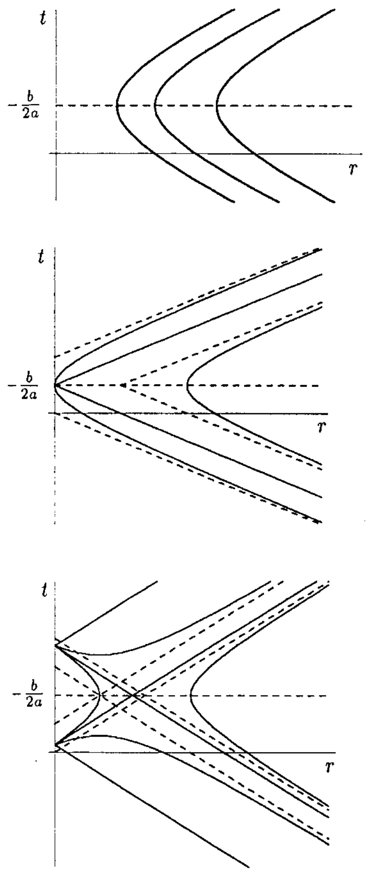

FIG. 3. Integral curves associated with a RCKF ξ given by expression (3) with $a\neq 0$, according to the sign of $\Delta = b^2 - 4ac$.

that is, $\xi_* \propto ds$ and the surfaces $\Sigma_s = \{(t,r,\theta,\varphi)/s = \text{constant}\}$ are orthogonal to the field ξ. Let us consider a domain where the field is not null on any point. The metric may be written using the coordinate s and another coordinate ω orthogonal to ξ, $\eta(d\omega,\xi)=0$. Such a coordinate has the expression

$$\omega(t,r) = \frac{a(t^2 - r^2) + bt + c}{r} \tag{17}$$

for any value of a, according to Eq. (15). In order to express the flat metric in these coordinates, we need the inverse transformation of Eqs. (16) and (17), which is

(i) For $a=0$,

$$t(s,\omega) = \begin{cases} \dfrac{s}{2c} & \text{if } b=0 \\ -\dfrac{c}{b} + \dfrac{\omega}{b}\sqrt{\dfrac{bs+c^2}{\omega^2 - b^2}} & \text{if } b\neq 0, \end{cases}$$

$$r(s,\omega) = \sqrt{\frac{bs+c^2}{\omega^2 - b^2}}. \tag{18}$$

(ii) For $a\neq 0$,

$$t(s,\omega)=\frac{1}{\sigma^2-\omega^2}\left[(\Delta-\omega^2)s-\frac{b}{2a}(\sigma^2-\Delta)\pm\frac{\omega}{2a}\sqrt{(\sigma^2-\Delta)(\omega^2-\Delta)}\right],$$

$$r(s,\omega)=\frac{\sqrt{\sigma^2-\Delta}}{2a(\sigma^2-\omega^2)}[-\sigma\sqrt{\omega^2-\Delta}\pm\omega\sqrt{\sigma^2-\Delta}], \tag{19}$$

where $\sigma\equiv 2as+b$ and the + (−) sign in the t coordinate corresponds to the + (−) sign in the r coordinate.

Then η, written in the coordinates $(s,\omega,\theta,\varphi)$, has the following diagonal form:

$$\eta_{(s,\omega,\theta,\varphi)}=\begin{cases} r^2(s,\omega)\mathrm{diag}\left(\frac{b^2-\omega^2}{4(bs+c^2)^2},\frac{1}{\omega^2-b^2},1,\sin^2\theta\right) & \text{if } a=0 \\ r^2(s,\omega)\mathrm{diag}\left(\frac{\Delta-\omega^2}{4(as^2+bs+c)^2},\frac{1}{\omega^2-\Delta},1,\sin^2\theta\right) & \text{if } a\neq 0. \end{cases} \tag{20}$$

Note that these coordinates $(s,\omega,\theta,\varphi)$ are not, in general, conformally flat coordinates. If we consider the coordinate transformations (18), (19) and Proposition 3, the resulting relation between (s,ω) and $(\bar{t},\bar{r})$ allows us to recover, from expression (20), the metric forms (12) and (13) presented in Sec. III.

From (20), the induced metric on the surfaces Σ_s by the Minkowski metric has the form

$$\gamma_{(\omega,\theta,\varphi)}=r^2(s,\omega)\mathrm{diag}\left(\frac{1}{\omega^2-\Delta},1,\sin^2\theta\right).$$

The Riemann double 2-form of curvature $\mathcal{R}$ of this induced metric can be expressed as $\mathcal{R}=[K(s)/2]\gamma\wedge\gamma$, where $\wedge$ denotes the exterior product of double 1-forms and $K(s)$ is given by

$$K(s)=\begin{cases} \frac{-b^2}{bs+c^2} & \text{for } a=0 \\ \frac{-a}{as^2+bs+c} & \text{for } a\neq 0. \end{cases} \tag{21}$$

Therefore each surface Σ_s (s=constant) has constant sectional curvature, $K(s)$. If we take into account expressions (5) and (16), we obtain

$$P=\begin{cases} bs+c^2 & \text{if } a=0 \\ \frac{as^2+bs+c}{a}(2at+b)^2 & \text{if } a\neq 0 \end{cases} \tag{22}$$

and the sectional curvature of Σ_s can be written in the form

$$K(s)=-\frac{1}{P}(2at+b)^2$$

whose sign depends on the causal character of the field ξ, and the following result follows.

Proposition 6: In the Minkowski space–time, the surfaces orthogonal to a RCKF are three-dimensional spaces with constant curvature, which will be negative or positive if the field is timelike or spacelike, respectively; except for the field $\xi=(\partial/\partial t)$ *(timelike everywhere), whose orthogonal 3-spaces are flat.*

VI. TIMELIKE RADIAL CONFORMAL MOTIONS

The case of timelike RCKF is specially interesting because they are associated with particular, but in general noninertial, observers in Minkowski space–time. We are going to study their kinematical properties. The shear and the vorticity of (the unit vector **u** associated with) a timelike RCKF are zero. The expansion is

$$\theta = \frac{3(2at+b)}{\sqrt{P}}$$

with P given by (5), and the acceleration has the form

$$\mathbf{a} = \frac{2ar}{P}(-r(2at+b)dt + (a(t^2+r^2)+bt+c)dr).$$

Note that $\eta(\mathbf{a},\mathbf{a}) = 4a^2/(\omega^2-\Delta)$ is constant along each integral curve, as we can see from (15). This agrees with the fact that the integral curves associated with a timelike RCKF describe hyperbolic or inertial motions. Then we have the following proposition in the Minkowski space–time:

Proposition 7: In the Minkowski space–time, the acceleration of a timelike RCKF has constant length on each integral curve, that is

$$|\mathbf{a}| = \frac{2|a|}{\sqrt{\omega^2-\Delta}},$$

where ω is given by (15).

Note that, in inertial coordinates, a timelike RCKF (3) is geodesic iff $a=0$. And it will have null expansion (that is, it will be a Killing vector field) iff the constants a and b are equal to zero.

Let us consider the 4-velocity of a timelike RCKF,

$$\mathbf{u} = \frac{\xi}{\sqrt{P}} = \frac{1}{\sqrt{1-\mathrm{v}^2}}\left(\frac{\partial}{\partial t} + \mathrm{v}\frac{\partial}{\partial r}\right),$$

the velocity v relative to the inertial observer $\partial/\partial t$ is given by the quotient between the components ξ^1 and ξ^0 of the field,

$$\mathrm{v} = \frac{r(2at+b)}{a(t^2+r^2)+bt+c}. \tag{23}$$

From Fig. 1, we can clearly see that when we are approaching (inside the timelike regions) to the light cones where the field ξ is null, the relative velocity $\mathrm{v}\to 1$. In particular, when $a=c=0$ we have $\mathrm{v}=r/t$. This corresponds to the geodesic field ξ_2 and adapting coordinates to it, the Minkowski metric is written

$$\eta = -d\tau^2 + \frac{\tau^2}{\left(1-\frac{\rho^2}{4}\right)^2}[d\rho^2 + \rho^2(d\theta^2 + \sin^2\theta d\varphi^2)], \tag{24}$$

which has the form of a Robertson–Walker metric with expansion factor $R(\tau)=\tau$. This is Milne's expression of the flat metric used to give a kinematic interpretation of the Hubble law.[2] The timelike coordinate τ represents the proper time of the geodesic radial congruence associated with the field $\xi_2 = \partial/\partial\tau$; and, according to Proposition 6, the surfaces $\tau=$constant are spaces of negative constant curvature.

Expression (24) can be obtained from (20) redefining the (s,ω) coordinates in the following way:

$$s=\tau^2, \quad \omega=\frac{\rho}{4}+\frac{1}{\rho}.$$

In this sense, Milne's interpretation of the recession velocity of galaxies can be understood adapting coordinates to a RCKF in Minkowski space–time.

VII. DISCUSSION AND COMMENTS

We have analyzed the main properties of the RCKF in Minkowski space–time. In a conformally flat space–time, whose metric can be locally written as $g=e^{2\lambda}\eta$, the form of the RCKF will be given by the same expression (3) or (4) as for the Minkowski space–time. But the acceleration and expansion of these fields will depend on the function λ and its first derivatives. So, additional conditions imposed on these kinematic properties lead to a differential equation for this function λ that can be used to determine it. For instance, imposing that a conformally flat space–time admits a geodesic RCKF, the corresponding differential equation allows one to obtain the conformal factor of the Robertson–Walker metric found by Infeld and Schild[1] in their work on kinematic cosmology.

Therefore, it is natural to wonder whether the existence of a RCKF with certain kinematic properties can characterize the Robertson–Walker metrics and other generalized nonhomogeneous conformally flat cosmological models. In fact, Robertson–Walker universes are those conformally flat space–times which admit a timelike geodesic RCKF.[3] Other kinematic properties over these RCKF (nongeodesic with homogeneous expansion or admitting homogeneous orthogonal 3-spaces) could be used to characterize generalized conformally flat cosmologies. We shall soon develop this idea further.[4]

ACKNOWLEDGMENTS

We wish to thank Dr. Joan J. Ferrando for his useful comments and suggestions. This work has been supported by the Ministerio de Educación y Cultura, Project No. PB96-0797. One of the authors (A.H.) wants to thank the Ministerio de Educación y Cultura for a fellowship.

[1] L. Infeld and A. Schild, ''A new approach to kinematic cosmology,'' Phys. Rev. **68**, 250–272 (1945).
[2] W. Rindler, *Essential Relativity* (Springer, New York, 1977).
[3] A. Herrero and J. A. Morales, ''Radial conformal motions,'' in *Proceedings ERE 98* (World Scientific, Singapore, in press).
[4] A. Herrero and J. A. Morales, ''A kinematic method to obtain conformal factors'' (submitted to J. Math. Phys.).

Dimensionally continued Oppenheimer–Snyder gravitational collapse: Solutions in odd dimensions

Anderson Ilha and Antares Kleber
Departamento de Astrofísica, Observatório Nacional-CNPq, Rua General José Cristino 77, 20921-400 Rio de Janeiro, Brazil

José P. S. Lemos
Departamento de Astrofísica, Observatório Nacional-CNPq, Rua General José Cristino 77, 20921-400 Rio de Janeiro, Brazil, and Departamento de Física, Instituto Superior Técnico, Av. Rovisco Pais 1, 1096 Lisboa, Portugal

(Received 9 July 1998; accepted for publication 4 March 1999)

The Lovelock gravity extends the theory of general relativity to higher dimensions in such a way that the field equations remain of second order. The theory has many constant coefficients with no *a priori* meaning. Nevertheless, it is possible to reduce them to two: the cosmological constant and Newton's constant. In this process one separates theories in even dimensions from theories in odd dimensions. In a previous work, gravitational collapse in even dimensions was analyzed. In this work attention is given to odd dimensions. It is found that black holes also emerge as the final state of gravitational collapse of a regular dust fluid. © *1999 American Institute of Physics.* [S0022-2488(99)03306-X]

I. INTRODUCTION

A generalization of Einstein gravity to other dimensions while keeping the same degrees of freedom (the field equations for the metric remain of second order) is given by the Lovelock action.[1] The theory can also be considered as an extension of Einstein–Hilbert action (see, e.g., Ref. 2), in which new terms make their appearance by taking into the action the Euler densities of the spaces with dimensions lower than the space in consideration.

In a previous work[3] we have studied gravitational collapse in Lovelock gravity for a spacetime with even dimensions, thus extending the Oppenheimer–Snyder collapsing model. Following the work of Refs. 2 and 4, the reason for separating even from odd dimensions in the Lovelock theory comes naturally in a $\mathcal{D}$-dimensional spacetime when one considers embedding the Lorentz group $SO(\mathcal{D}-1,1)$ into the anti-de Sitter group $SO(\mathcal{D}-1,2)$. The Lovelock theory then branches into two distinct classes, with Lagrangians for even dimensions and Lagrangians for odd dimensions. One also finds in this way that the number of constants, that proliferates when one goes to higher and higher dimensions, reduces drastically to two: the cosmological constant Λ and Newton's constant G.

In this work we study gravitational collapse in odd-dimensional spacetimes and show that black holes form from regular initial data consisting of a dust fluid. We follow closely the nomenclature and the division of sections made in Ref. 3. In Sec. II the Lovelock gravity for restricted coefficients in odd-dimensional spacetimes is presented. In Sec. III we display the static solutions in odd dimensions found in Ref. 4. In Sec. IV we find some cosmological or interior matter solutions for perfect fluids. In Sec. V we match the solutions found in Sec. IV to the solutions of Sec. III. In Sec. VI we show that black holes can form through gravitational collapse in Lovelock odd-dimensional gravity. In Sec. VI we comment on the formation of naked singularities and in Sec. VII we present some conclusions. In the paper we usually do $G=c=1$.

II. THE LOVELOCK THEORY

The most general action in $\mathcal{D}\geqslant 3$ spacetime dimensions that yields the same degrees of freedom of Einstein's theory is the so-called Lovelock action, given by[1,2]

$$S=\int \mathcal{L}_{\mathcal{D}}=\kappa \sum_{p=0}^{[(\mathcal{D}-1)/2]} \alpha_p \int_M \epsilon_{a_1\cdots a_{\mathcal{D}}} R^{a_1a_2}\wedge\cdots\wedge R^{a_{2p-1}a_{2p}}\wedge e^{a_{2p+1}}\wedge\cdots\wedge e^{a_{\mathcal{D}}}+S_m, \tag{2.1}$$

where $R^{ab}=d\omega^{ab}+\omega^a_c\wedge\omega^{cb}$ is the curvature two-form, e^a is the local frame one-form, and ω^{ab} is the spin connection, with $a_i=0,1,...,\mathcal{D}-1$. The symbol [ ] over the summation symbol means one should take the integer part of $(\mathcal{D}-1)/2$. S_m is a phenomenological action that describes the macroscopic matter sources.

In general, the constant coefficients α_p are arbitrary. However, it is shown in Ref. 4 that taking certain special choices one is able to get simple meaningful solutions. Following Ref. 4, one first considers embedding the Lorentz group SO$(\mathcal{D}-1,1)$ into the anti-de Sitter group SO$(\mathcal{D}-1,2)$, and then separates into two distinct classes of Lagrangians: Lagrangians for even dimensions and Lagrangians for odd dimensions.

For odd dimensions, $\mathcal{D}=2n-1$, one can find a construction similar to the Chern–Simons action construction in three dimensions. One starts with the Euler density in one dimension above $\mathcal{D}$,

$$E_{2n}=\kappa\epsilon_{A_1\cdots A_{2n}}\hat{R}^{A_1A_2}\wedge\cdots\wedge\hat{R}^{A_{2n-1}A_{2n}}, \tag{2.2}$$

with $A_1,A_2=0,1,...,2n-1$ being the anti-de Sitter indices. $\hat{R}^{AB}$ is the anti-de Sitter curvature two-form constructed with the SO$(\mathcal{D}-1,2)$ connection W^{AB}. Equation (2.2) is a local exact form, and can be written as an exterior derivative of a Lagrangian in $2n-1$ dimensions, i.e., $E_{2n}=d\mathcal{L}_{2n-1}$; see Ref. 4. Decomposing the connection W^{AB} into the connection under $\mathcal{D}$ rotations w^{ab} and inner translations e^a, one finds the anti-de Sitter curvature $\hat{R}$ in terms of the Lorentz curvature R:

$$\hat{R}^{ab}=R^{ab}+\frac{1}{l^2}e^a\wedge e^b, \tag{2.3}$$

where l is a scale factor that is to be related to the cosmological constant $l^2=-1/\Lambda$. Using Eq. (2.3), one finds that the Lagrangian in Eq. (2.2) can be put in the form

$$\mathcal{L}_{2n-1}=\kappa\sum_{p=0}^{n-1}\alpha_p\epsilon_{a_1\cdots a_{\mathcal{D}}}R^{a_1a_2}\wedge\cdots\wedge R^{a_{2p-1}a_{2p}}\wedge e^{a_{2p+1}}\wedge\cdots\wedge e^{a_{\mathcal{D}}}, \tag{2.4}$$

where the coefficients are given by

$$\alpha_p=\frac{1}{\mathcal{D}-2p}\binom{n-1}{p}l^{-\mathcal{D}+2p}, \tag{2.5}$$

and, for convenience, one can choose κ as

$$\kappa=\frac{\mathcal{D}-2}{16\pi Gn}l^{\mathcal{D}-2}. \tag{2.6}$$

Given the action (2.1), the field equations are obtained by the variation with respect to the one-forms e^a. Under the assumption of zero torsion, the variation with respect to the spin connection ω^{ab} vanishes identically. Although the equations have powers in the curvatures, they remain by construction second order in the metric. The field equations are given by

$$-\kappa\sum_{p=0}^{[(\mathcal{D}-1)/2]}\alpha_p(\mathcal{D}-2p)\epsilon_{a_1\cdots a_{\mathcal{D}}}R^{a_1a_2}\wedge\cdots\wedge R^{a_{2p-1}a_{2p}}\wedge e^{a_{2p+1}}\wedge\cdots\wedge e^{a_{\mathcal{D}-1}}=Q_{a_{\mathcal{D}}}, \tag{2.7}$$

where $Q_{a_{\mathcal{D}}}$ is a $(\mathcal{D}-1)$-form associated with the energy momentum tensor T^a_b through the following expression:

$$Q_i=\frac{1}{(\mathcal{D}-1)!}T_i^{a_1}\epsilon_{a_1\cdots a_{\mathcal{D}}}e^{a_2}\wedge\cdots\wedge e^{a_{\mathcal{D}}}. \tag{2.8}$$

III. EXTERIOR VACUUM SOLUTIONS

In the vacuum all components of the energy–momentum tensor vanish, so that the field equations (2.7) are given by

$$-\kappa\sum_{p=0}\alpha_p(\mathcal{D}-2p)\epsilon_{a_1\cdots a_{\mathcal{D}}}R^{a_1a_2}\wedge\cdots\wedge R^{a_{2p-1}a_{2p}}\wedge e^{a_{2p+1}}\wedge\cdots\wedge e^{a_{\mathcal{D}-1}}=0. \tag{3.1}$$

Inserting the coefficients α_p and the constant κ given in (2.5) and (2.6) in Eq. (3.1) one gets for odd dimensions $(\mathcal{D}=2n-1)$,

$$(R^{a_1a_2}+l^{-2}e^{a_1}\wedge e^{a_2})\wedge\cdots\wedge(R^{a_{2n-3}a_{2n-2}}+l^{-2}e^{a_{2n-3}}\wedge e^{a_{2n-2}})\epsilon_{a_1a_2\cdots a_{2n-1}}=0. \tag{3.2}$$

We consider now a static, spherical symmetric spacetime. One can write the metric in the following form:

$$ds_+^2=-g^2(r_+)dt_+^2+g^{-2}(r_+)dr_+^2+r_+^2d\Omega_{\mathcal{D}-2}^2, \tag{3.3}$$

where t and r are the time and radial coordinates and $d\Omega^2_{\mathcal{D}-2}$ is the arc element of a unit $(\mathcal{D}-2)$ sphere. The subscript + reminds that (3.3) is to be viewed as an exterior solution. With metric (3.3) and Eqs. (3.1) and (3.2), Bañados, Teitelboim, and Zanelli found the following exact solution for $\mathcal{D}=2n-1$:[4]

$$ds_+^2=-[1-(M+1)^{2/(\mathcal{D}-1)}+(r_+/l)^2]dt_+^2+\frac{dr_+^2}{1-(M+1)^{2/(\mathcal{D}-1)}+(r_+/l)^2}+r_+^2d\Omega_{\mathcal{D}-2}^2. \tag{3.4}$$

These solutions describe black holes. We will show that they also represent the exterior vacuum solution to a collapsing (or expanding) dust cloud in Lovelock's odd-dimensional theory, as in the even-dimensional case.[3]

IV. INTERIOR MATTER SOLUTIONS

The interior spacetime is modeled by a homogeneous collapsing (or expanding) dust cloud, whose metric is described by Friedmann–Robertson–Walker in $\mathcal{D}$ dimensions,

$$ds^2=-dt^2+a^2(t)\left[\frac{dr^2}{1-kr^2}+r^2d\Omega^2_{\mathcal{D}-2}\right]. \tag{4.1}$$

The coordinates t and r are comoving coordinates (we omit throughout the subscript − to indicate an interior solution). Note that that k has dimension of 1/[length]^2. The energy–momentum tensor for a perfect fluid is given by

$$T_{\alpha\beta}=(\rho+p)u_\alpha u_\beta+pg_{\alpha\beta}, \tag{4.2}$$

where ρ is the energy–density, p the pressure, and u^α is the $\mathcal{D}$ velocity of the fluid. From (4.1)–(4.2) and Lovelock equations (2.7), we obtain

$$-B\frac{d}{dt}\left(\frac{\dot{a}}{a}\right)+\frac{k}{a^2}=\rho+p, \tag{4.3}$$

$$(\mathcal{D}-1)B\left(\frac{\dot{a}}{a}\right)\left[-\frac{k}{a^2}+\frac{d}{dt}\left(\frac{\dot{a}}{a}\right)\right]=\dot{\rho}, \tag{4.4}$$

where

$$B\equiv(\mathcal{D}-2)!\sum_p \alpha_p 2p(\mathcal{D}-2p)\left(\frac{\dot{a}^2+k}{a^2}\right)^{p-1}, \tag{4.5}$$

where the coefficients α_p are given in (2.5), and κ is given in (2.6). Equations (4.3)–(4.4) have a first integral given by

$$\dot{a}^2=-k-\left(\frac{a}{l}\right)^2+\left(\frac{a_0}{l}\right)^2\left[\frac{8\pi l^2\rho_0}{(\mathcal{D}-2)!(\mathcal{D}-2)}\right]^{2/(\mathcal{D}-1)}, \tag{4.6}$$

where ρ_0 and a_0 are constants. Equations (4.3)–(4.4) also have a second integral, i.e., the solution of Eq. (4.6) is given by (see also Ref. 5)

$$a(t/l)=\frac{l}{r_\Sigma}\sqrt{\left\{\left(\frac{1}{l}\right)^{\mathcal{D}-3}\left[\frac{8\pi\rho_0(a_0r_\Sigma)^{\mathcal{D}-1}}{(\mathcal{D}-2)!(\mathcal{D}-2)}\right]\right\}^{2/(\mathcal{D}-1)}-kr_\Sigma^2}\,\sin(b+t/l), \tag{4.7}$$

where b is an arbitrary phase that will be neglected henceforward.

The Ricci quadratic scalar and the Kretschmann scalar are given by

$$R_{ab}R^{ab}=-(\mathcal{D}-1)^2\left(\frac{\ddot{a}}{a}\right)^2+(\mathcal{D}-1)\left[\frac{\ddot{a}}{a}+(\mathcal{D}-2)\frac{\dot{a}^2+k}{a^2}\right]^2, \tag{4.8}$$

$$R_{abcd}R^{abcd}=(\mathcal{D}-1)\left[\left(\frac{\ddot{a}}{a}\right)^2+\left(\frac{\dot{a}^2+k}{a^2}\right)^2\right], \tag{4.9}$$

respectively.

We now assume a dust fluid, $p=0$. For such an equation of state we have

$$\rho=\rho_0\left(\frac{a_0}{a}\right)^{\mathcal{D}-1}, \tag{4.10}$$

where ρ_0 and a_0 are the constants defined above.

Inserting Eq. (4.7) in Eq. (4.10), we obtain the evolution of the density in the dust model:

$$\rho(t/l)=\rho_0\left[\frac{a_0r_\Sigma/l}{\sqrt{\left\{\left(\frac{1}{l}\right)^{\mathcal{D}-3}\left[\frac{8\pi\rho_0(a_0r_\Sigma)^{\mathcal{D}-1}}{(\mathcal{D}-2)!(\mathcal{D}-2)}\right]\right\}^{2/(\mathcal{D}-1)}-kr_\Sigma^2}}\right]^{\mathcal{D}-1}\sin^{-(\mathcal{D}-1)}(t/l). \tag{4.11}$$

We see that the density (4.11) and the curvature scalars (4.8)–(4.9) diverge at $t/l=\pi$, which represents the formation of a singularity.

V. JUNCTION CONDITIONS

Now we match the exterior and interior spacetimes found in Secs. III and IV, respectively, across an interface of separation Σ. The junctions conditions are[6]

$$ds_+^2]_\Sigma = ds_-^2]_\Sigma , \tag{5.1}$$

$$K_{\alpha\beta}^+]_\Sigma = K_{\alpha\beta}^-]_\Sigma , \tag{5.2}$$

where $K_{\alpha\beta}$ is the extrinsic curvature,

$$K_{\alpha\beta}^\pm = -n_\epsilon^\pm \frac{\partial^2 x_\pm^\epsilon}{\partial \xi^\alpha \partial \xi^\beta} - n_\epsilon^\pm \Gamma^\epsilon_{\gamma\delta} \frac{\partial x_\pm^\gamma}{\partial \xi^\alpha} \frac{\partial x_\pm^\delta}{\partial \xi^\beta}, \tag{5.3}$$

and $n_\epsilon^\pm$ are the components of the unit normal vector to Σ in the coordinates $x_\pm$, and ξ represents the intrinsic coordinates in Σ. The subscripts $\pm$ represent the quantities taken in the exterior and interior spacetimes. Both the metrics and the extrinsic curvatures in (5.1)–(5.2) are evaluated at Σ. The metric intrinsic to Σ is written as

$$ds_\Sigma^2 = -d\tau^2 + R^2(\tau) d\Omega_{\mathcal{D}-2}^2, \tag{5.4}$$

where τ is the proper time on Σ and $d\Omega_{\mathcal{D}-2}^2$ denotes the line element on a $\mathcal{D}-2$-dimensional sphere.

Using the junction condition (5.1), metric (5.4), and the exterior metric (3.4), we obtain

$$r_+]_\Sigma = R(\tau), \tag{5.5}$$

and

$$[1-(M+1)^{2/(\mathcal{D}-1)} + (r_+/l)^2]\dot{t}_+^2 - [1-(M+1)^{2/(\mathcal{D}-1)} + (r_+/l)^2]^{-1}\dot{r}_+^2 = 1, \tag{5.6}$$

where $\cdot \equiv d/d\tau$, and both equations are evaluated at Σ. From now on, we will usually omit the subscript Σ to denote evaluation at the interface. Using (5.5) in (5.6), we find

$$\frac{dt_+}{d\tau} = \frac{\sqrt{[1-(M+1)^{2/(\mathcal{D}-1)} + (R/l)^2] + \dot{R}^2}}{[1-(M+1)^{2/(\mathcal{D}-1)} + (R/l)^2]}. \tag{5.7}$$

The unit normal to Σ in the exterior spacetime is

$$n_\epsilon^+ = \left(-\frac{dr_+}{d\tau}, \frac{dt_+}{d\tau}, 0,...,0 \right). \tag{5.8}$$

From (5.3) we then get

$$K_{\theta\theta}^+ = R \sqrt{\left[1-(M+1)^{2/(\mathcal{D}-1)} + \left(\frac{R}{l}\right)^2 \right] + \dot{R}^2}, \tag{5.9}$$

In what follows the other components of K_{ab}^+ are not needed.

The unit normal to Σ in the interior spacetime is

$$n_\epsilon^- = \left(0, \frac{a}{\sqrt{1-kr^2}}, 0,...,0 \right), \tag{5.10}$$

and from (5.3) we have

$$K_{\theta\theta}^- = R(\tau)\sqrt{1-kr_\Sigma^2}. \tag{5.11}$$

Using the junction condition (5.1) for the interior spacetime yields $ar_\Sigma = R(\tau)$. From the condition $K_{\theta\theta}^+ = K_{\theta\theta}^-$, (5.9) and (5.11), we obtain

$$\dot{R}^2+\left(\frac{R}{l}\right)^2+kr_\Sigma^2=(M+1)^{2(\mathcal{D}-1)}. \tag{5.12}$$

Multiplying Eq. (4.6) by r_Σ^2, we get

$$\dot{R}^2+\left(\frac{R}{l}\right)^2+kr_\Sigma^2=\left(\frac{R_0}{l}\right)^2\left[\frac{8\pi l^2\rho_0}{(\mathcal{D}-2)!(\mathcal{D}-2)}\right]^{2/(\mathcal{D}-1)}. \tag{5.13}$$

Comparing Eqs. (5.12) and (5.13), we have

$$M=\left(\frac{1}{l}\right)^{\mathcal{D}-3}\left[\frac{8\pi\rho_0(a_0r_\Sigma)^{\mathcal{D}-1}}{(\mathcal{D}-2)!(\mathcal{D}-2)}\right]-1, \tag{5.14}$$

which is the mass of the cloud expressed in terms of the constants given in the problem.

VI. BLACK HOLE FORMATION

In order to study black hole formation in this theory, we work with the solution found in (4.7). The interior and exterior metrics are given in (4.1) and in (3.4), respectively, and as we have shown in Sec. V, it is possible to make a smooth junction between both spacetimes. To be complete, we treat the cases $\mathcal{D}\geqslant 3$. The case $\mathcal{D}=3$ reduces to the collapse studied in Ref. 7.

For convenience, we rewrite Eqs. (4.6)–(4.9) and (4.11) in terms of the mass M. We thus have

$$a(t/l)=\frac{l}{r_\Sigma}\sqrt{(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2}\,\sin(t/l), \tag{6.1}$$

for the scale factor,

$$\rho(t/l)=\rho_0\left[\frac{a_0r_\Sigma/l}{\sqrt{(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2}}\right]^{\mathcal{D}-1}\sin^{-(\mathcal{D}-1)}(t/l), \tag{6.2}$$

for the density, and

$$R_{ab}R^{ab}=-\frac{(\mathcal{D}-1)^2}{l^4}+\frac{(\mathcal{D}-1)}{l^4}\left\{-1+(\mathcal{D}-2)\frac{[(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2]\cos^2(t/l)+kr_\Sigma^2}{[(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2]\sin^2(t/l)}\right\}^2, \tag{6.3}$$

and

$$R_{abcd}R^{abcd}=\frac{(\mathcal{D}-1)}{l^4}\left\{1+\left[\frac{[(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2]\cos^2(t/l)+kr_\Sigma^2}{[(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2]\sin^2(t/l)}\right]^2\right\}, \tag{6.4}$$

for the quadratic Ricci and Kretschmann scalars, respectively. In this work we restrict the values of the quantity kr_Σ^2, assuming $kr_\Sigma^2=0,\ \pm\frac{1}{2}$. These values have no special meaning, although for kr_Σ^2 positive and large enough there is no solution at all. Note also that the expression (5.14) for the mass is independent of the value chosen for kr_Σ^2.

Gravitational collapse occurs for $\pi/2\leqslant t/l\leqslant\pi$. The time $t/l=\pi/2$ marks the onset of collapse. At this moment there are no singularities in spacetime, as the curvature scalars (6.3)–(6.4) and the density (6.2) indicate. In fact, the singularity appears only at $t/l=\pi$, where all these quantities blow up.

To know whether a black hole has formed or not, one has to search for the appearance of an apparent horizon and an event horizon. The apparent horizon is defined to be the boundary of the region of trapped two-spheres in spacetime. To find this boundary on the interior spacetime, one looks for two spheres $Y \equiv a(t)r = \text{const}$ whose outward normals are null, i.e., $\boldsymbol{\nabla} Y \cdot \boldsymbol{\nabla} Y = 0$. Using metric (4.1), this yields

$$\frac{da(t)}{dt} = -\frac{\sqrt{1-kr^2}}{r}. \tag{6.5}$$

Using (6.1) in (6.5) gives the evolution of the apparent horizon in comoving coordinates,

$$\sqrt{\frac{(M+1)^{2/(\mathcal{D}-1)} - kr_\Sigma^2}{1-kr_\Sigma^2(r/r_\Sigma)^2}}\left(\frac{r}{r_\Sigma}\right) = -\frac{1}{\cos(t/l)}. \tag{6.6}$$

Now, the apparent horizon first forms at the surface r_Σ. Then, for $r = r_\Sigma$, Eq. (6.6) gives the time t/l at which the apparent horizon first forms. On the other hand, one should also be able to find the formation time of the apparent horizon on the surface Σ through an equation on Σ, Eq. (5.12). Indeed, at the junction one has $R = a(t)r_\Sigma$. Then, from junction condition (5.12) and Eq. (6.5) we have that the apparent horizon first forms when

$$\frac{R_{\mathrm{AH}}}{l} = \sqrt{(M+1)^{2/(\mathcal{D}-1)} - 1}. \tag{6.7}$$

Now, using (6.1), the time of formation of the apparent horizon can be found through the equation

$$\frac{R_{\mathrm{AH}}}{l} = a(t_{\mathrm{AH}})\frac{r_\Sigma}{l} = \sqrt{(M+1)^{2/(\mathcal{D}-1)} - kr_\Sigma^2}\,\sin\left(\frac{t_{\mathrm{AH}}}{l}\right). \tag{6.8}$$

Given a dimension $\mathcal{D}$ and an M, one can obtain R_{AH} through Eq. (6.7). Then Eq. (6.8) gives implicitly t_{AH}, the time of the formation of the apparent horizon on the surface Σ for a given k. For instance, for $\mathcal{D}=3$, $M=0.25$, and $kr_\Sigma^2=0$ we find $t_{\mathrm{AH}}/l = 2.68$. Putting this value back in Eq. (6.6) we verify that everything checks.

The event horizon, being a null spherical surface, is determined through the null outgoing lines of the metric (4.1), i.e.,

$$\frac{dt}{dr} = \frac{a(t)}{\sqrt{1-kr^2}}. \tag{6.9}$$

Equation (6.9) can be put in the following integral form:

$$\sqrt{(M+1)^{2/(\mathcal{D}-1)} - kr_\Sigma^2}\int_0^{r_1/r_\Sigma}\frac{d(r/r_\Sigma)}{\sqrt{1-kr_\Sigma^2(r/r_\Sigma)^2}} = \ln\left[\frac{\tan(x_1)}{\tan(x_0)}\right], \tag{6.10}$$

where $x \equiv (1/2)t/l$. Now, the time x_1 is to be precisely equal to the formation time of the apparent horizon, since one expects that in vacuum both horizons coincide.[8] One has then to integrate (6.10) to find the time x_0 at which the event horizon first forms, at $r=0$. For instance, $\mathcal{D}=3$, $M=0.25$, and $kr_\Sigma^2 = -\frac{1}{2}$ we obtain $t_0/l = 1.96$. A plot in comoving coordinates $(t/l, r/r_\Sigma)$ shows the evolution of the apparent and event horizons in Fig. 1. There we repeat the numerical calculations for the same value of $\mathcal{D}$ and M, but with $kr_\Sigma^2=0$ and $kr_\Sigma^2=\frac{1}{2}$, as is shown in lines (b) and (c). In Fig. 2 we show the formation of the apparent and event horizons for $\mathcal{D}=25$, and $M=0.25$ and $kr_\Sigma^2=0$. Intermediate $\mathcal{D}$ dimensions have a similar behavior. Making a matching to the vacuum exterior spacetime one finds the usual Penrose diagram for gravitational collapse and formation of a black hole in an anti-de Sitter background; see Fig. 3.

To study what happens to external observers we note that a light signal emitted from the surface $r_+]_\Sigma$ at the exterior time t_+ obeys the null condition

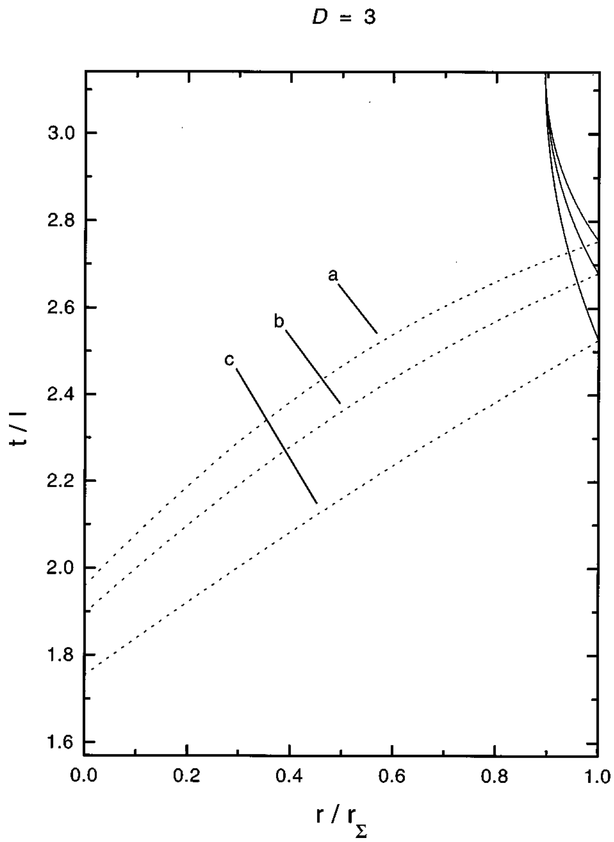

FIG. 1. Gravitational collapse in $\mathcal{D}=3$ dimensions in an asymptotically anti-de Sitter spacetime. The interior dust cloud in comoving coordinates $(t/l, r/r_\Sigma)$ fills the whole diagram. The left side represents the center of the cloud $r/r_\Sigma=0$; the right side the surface of the cloud $r/r_\Sigma=1$. The evolution of the event horizon (dashed line) and apparent horizon (full line) are drawn. The singularity occurs at $t/l=0$. $M=0.25$ was used. The three different cases are (a) $kr_\Sigma^2=-\frac{1}{2}$, (b) $kr_\Sigma^2=0$, and (c) $kr_\Sigma^2=\frac{1}{2}$.

$$\frac{dr_+}{dt_+}=1-(M+1)^{2/(\mathcal{D}-1)}+\left(\frac{r_+}{l}\right)^2 \tag{6.11}$$

[see Eq. (3.4)]. This light ray arrives at a point r_+ at time t_+ given by

$$\frac{t_+}{l}=\frac{t_+]_\Sigma}{l}+\frac{1}{2(M+1)^{2/(\mathcal{D}-1)}-2}\ln\left[\frac{(r_+/l)-[2(M+1)^{2/(\mathcal{D}-1)}-2]}{(r_+/l)+[2(M+1)^{2/(\mathcal{D}-1)}-2]}\right]_{r_+]_\Sigma/l}^{r_+/l}. \tag{6.12}$$

Thus, $t_+/l\to\infty$ when $r_+]_\Sigma/l\to\sqrt{(M+1)^{2/(\mathcal{D}-1)}-1}$, so the collapse to the event horizon appears to take an infinite amount of time to an exterior observer, and the collapse to $r_+=0$ is unobservable from the outside. Also, the redshift from the dust edge is given by

$$z=\frac{dt_+}{dt}-1=\frac{1}{\sqrt{1-kr_\Sigma^2}+\dot{R}}-1. \tag{6.13}$$

When the dust edge crosses the event horizon we have $\dot{R}=-\sqrt{1-kr_\Sigma^2}$, so $z\to\infty$. Thus the collapsing dust will fade from sight, as the redshift of the light from its surface diverges.

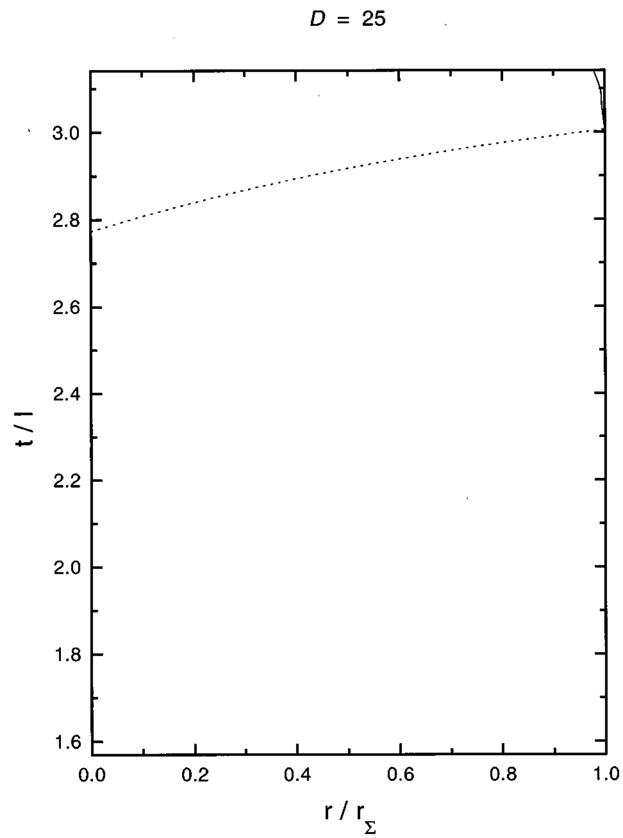

FIG. 2. The dimensionally continued Oppenheimer–Snyder collapse in $\mathcal{D}=25$ dimensions in an asymptotically anti-de Sitter spacetime. $M=0.25$ and $kr_{\Sigma}^{2}=0$ were used. See the subtitle of Fig. 1 for a more detailed explanation.

VII. NAKED SINGULARITIES

To study the presence of naked singularities, i.e., singularities not hidden by an event horizon, we analyze Eqs. (3.4), (6.1)–(6.4), and (5.14). Naked singularities appear only when $M<0$. Although solutions with negative mass are usually considered unphysical, they will be studied here because these generalize the three-dimensional solutions found in Refs. 9, 10, and 7. In the model adopted here, it is useful to separate two distinct classes.

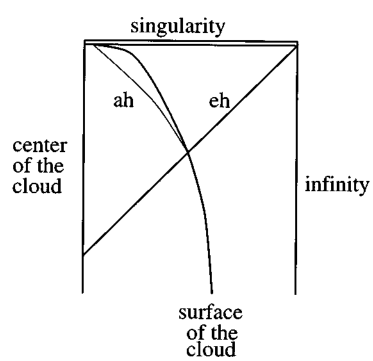

FIG. 3. Penrose diagram for the collapse of a dust cloud in an asymptotically anti-de Sitter spacetime. Each point in the diagram represents a $\mathcal{D}-2$ sphere (eh=event horizon, ah=apparent horizon).

(1) If l remains finite (in which case $\Lambda\neq 0$), for any $\mathcal{D}\geqslant 3$ the curvature scalars (6.3)–(6.4) will blow up when $t/l=\pi$, indicating the formation of a curvature naked singularity.

(2) If we take the limit $l\to\infty$ (in which case $\Lambda=0$), we see from the exterior metric (3.4) that the event horizon is no longer present, and the collapse will form a naked singularity. Taking the limit on Eqs. (6.3)–(6.4), we have

$$R_{ab}R^{ab}=\frac{(\mathcal{D}-1)(\mathcal{D}-2)}{t^4}\left[\frac{(M+1)^{2/(\mathcal{D}-1)}}{(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2}\right]^2, \tag{7.1}$$

$$R_{abcd}R^{abcd}=\frac{\mathcal{D}-1}{t^4}\,\frac{(M+1)^{2/(\mathcal{D}-1)}}{(M+1)^{2/(\mathcal{D}-1)}-kr_\Sigma^2}. \tag{7.2}$$

For any $\mathcal{D}>3$, both (7.1)–(7.2) will vanish because from Eq. (5.14), $M=-1+\mathcal{O}(l^{-\mathcal{D}+3})$, so in the limit we have $M=-1$. Also, from Eq. (6.1) we have in the limit, $a(t)=\sqrt{-k}t$, so that the only possible solution is when $kr_\Sigma^2<0$. Note also that $M=-1$ implies that the exterior metric (3.4) is a Minkowski one, although the interior density (6.2) is nonzero everywhere in the dust cloud. So at $t/l=\pi$ we will have $\rho\to\infty$ in a flat Mikowski spacetime. This is analogous to a Newtonian singularity.

For $\mathcal{D}=3$ we have $M=8\pi\rho_0(a_0r_\Sigma)^2-1$ and $a(t)=\sqrt{8\pi\rho_0a_0^2-k}t$, so that Eqs. (7.1)–(7.2) will be finite but nonzero and the collapse will form a naked conical singularity.[10,7]

VIII. CONCLUSIONS

We have analyzed gravitational collapse in Lovelock gravity for odd-dimensional spacetimes. We have showed that gravitational collapse of a regular initial nonrotating dust cloud proceeds, to form event and apparent horizons, and terminates at a spacelike curvature singularity.

[1] D. Lovelock, J. Math. Phys. **12**, 498 (1971).
[2] C. Teitelboim and J. Zanelli, in *Constraint Theory and Relativistic Dynamics*, edited by G. Longhi and L. Lussana (World Scientific, Singapore 1987).
[3] A. Ilha and J. P. S. Lemos, Phys. Rev. D **55**, 1788 (1997).
[4] M. Bañados, C. Teitelboim, and J. Zanelli, Phys. Rev. D **49**, 975 (1994).
[5] X. Li, Phys. Rev. D **50**, 3787 (1994).
[6] W. Israel, Nuovo Cimento B **44**, 463 (1967).
[7] R. B. Mann and S. F. Ross, Phys. Rev. D **47**, 3319 (1993).
[8] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space and Time* (Cambridge University Press, Cambridge, England, 1973).
[9] S. Deser, R. Jackiw, and G. 't Hooft, Ann. Phys. (N.Y.) **152**, 220 (1984).
[10] S. Giddings, J. Abbot, and K. Kuchar, Gen. Relativ. Gravit. **16**, 751 (1984).

Desingularization of Jacobi metrics and chaos in general relativity

Marek Szydłowski
Astronomical Observatory, Jagiellonian University, ul. Orla 171, 30-244 Cracow, Poland

(Received 29 June 1998; accepted for publication 4 February 1999)

It has been proposed by Wheeler and deWitt to look at the evolution of three-metrics as a geodesic flow on the superspace. Since then a lot of attention has been paid towards better understanding the geometric structure of the superspace. In particular it has been appreciated that the minisuperspace can in a natural way be equipped with the Jacobi metric. However the Jacobi metric is degenerate on certain codimension one hypersurfaces (boundary sets) leading to severe difficulties. The dynamics of minisuperspace models is a special case of dynamics of simple indefinite mechanical systems. It is proved that trajectories of an indefinite mechanical system (M,g,V), with the natural Lagrangian, are pregeodesics with respect to the Jacobi metric $g_E=2|E-V|g$, where M is the configuration space, g the metric defined by the kinetic energy form, V a potential function, and E the total energy of the system. In this paper we also propose to use Eisenhart's principle as an alternative geometrical construction on minisuperspace. Then the dynamics of general relativity is mapped onto a geodesic flow on a smooth manifold without boundary. Hence Eisenhart's proposal seems to be the right way to desingularization of motion in the Jacobi metric which is nontractable in the Jacobi picture. Different methods of desingularizing of the Jacobi metric through the isometric embedding into a more dimensional flat space with the Lorentzian signature is also presented. [S0022-2488(99)01007-5]

I. INTRODUCTION

Mechanical systems with natural Lagrange functions are described by the Lagrangian of the type $L=(1/2)g_{\alpha\beta}\dot{q}^\alpha\dot{q}^\beta-V(p)$, where $V:M\rightarrow\mathbf{R}$ is a potential function. The classical Maupertuis principle states that trajectories of a mechanical system with the total energy $E=(1/2)g_{\alpha\beta}\dot{q}^\alpha\dot{q}^\beta+V(p)$ are geodesics of the Jacobi metric $g_E^J=2(E-V)g$, where g is a Riemannian metric on a smooth manifold M being the configuration space of the system. Let g be given by the kinetic energy form $K(v)=(1/2)g(v,v)$, for every $v\in T_xM$, $x\in M$. The Jacobi metric g_E^J is determined on a submanifold $M_E=\{x\in M:V(x)<E\}$ which is called an *admissible configuration space* (see, for instance, Refs. 1 and 2).

In the previous work,[3] we have generalized the Maupertuis principle by admitting the case when g is a Lorentz metric on M; in such a case the Jacobi metric is of the form $g_E^J=2|E-V|g$, and consequently we have included into the admissible configuration space the regions for which $V(x)\geqslant E$, $x\in M$. According to our previous terminology, the above described mechanical system with the natural Lagrange function is called a *simple mechanical system* (SMS). If the kinetic energy form K is positive definite the system is said to be *classical*; if K is indefinite the system is said to be *relativistic*. In the present work we further study the trajectories of such systems (with respect to the Jacobi metric) and investigate their behavior when they pass through the singular boundary $(V(x)=E)$.

Relativistic systems appears in applications to general relativity and cosmology where the kinetic energy form is indefinite and the metric has the Lorentz signature. For instance, in the so-called ADM formulation of general relativity, the dynamics of space–time reduces to the study of a dynamical system with a suitable potential function which is determined by the geometry of spacelike sections of space–time.[4] In the case of homogeneous anisotropic cosmological models the potential function is determined by the Lie algebras of the corresponding isometry groups acting on spacelike sections of space–time.[5] In all these cases the admissible configuration space contains regions with $V(x)\geqslant E$, and the Jacobi metric g_E^J degenerates on the singular boundary $V(x)=E$.

The status of the Maupertuis principle in classical mechanics is rather ambiguous. (Some authors call this principle the Jacobi principle.[6]) Arnold in his beautiful book,[1] quoting Jacobi who said that ''in almost all textbooks, even in the best ones, this principle is presented in such a way that it is impossible to understand it,'' ironically admitted that also he himself did not want to damage this tradition, and restricted the usefulness of the Maupertuis principle to the cases when $E>V_{\max}(x)$. However, the investigation of classical dynamical systems with degenerate Jacobi metrics (for the cases in which $V(x)\leqslant E$)) began with the works of Wintner[7] and Seifert.[8] Interesting results in this domain were obtained by Kozlov.[9–11] In the present work, we give the proof of a theorem which can be regarded as a rigorous formulation of the generalized (to the case $E\leqslant V(x)$) Maupertuis principle for dynamical systems with the natural Lagrange functions.

It is worthwhile to mention that the Maupertuis principle, for the case with the indefinite kinetic energy form (i.e., for relativistic mechanical systems), was implicitly used by Misner in his minisuperspace construction[12] (and earlier by DeWitt in his concept of superspace).[13] Misner's idea was to define the space of spatially homogeneous and closed solutions of Einstein's equations. Analogously, the Maupertuis principle can be regarded as defining the space of solutions of the Euler–Lagrange equation (geodesics with respect to the Jacobi metric are ''points'' in this space). Other examples of relativistic (low-dimensional) mechanical systems are shown in Table I. In these systems we can find or suspect complex (chaotic) behavior of trajectories in the phase space. One should expect that this behavior has a counterpart in the behavior of geodesics in the space with the Jacobi metric.[14,15]

It was proposed by Wheeler and deWitt to look at the evolution of three-metrics as a geodesic flow on the superspace. Since then a lot of attention has been paid towards better understanding of the geometric structure of the superspace, which turned out to be nontrivial (a metrizable stratified union of manifolds which itself is not a manifold). In particular it has been appreciated that the minisuperspace can in a natural way be equipped with the Jacobi metric. [To see this one should replace, in Misner's formulas (47)–(49), the conformal factor $f^2(g^A)$ by our $2|E-V|$. In this case, Misner's supertime λ coincides with our parameter s.] However the Jacobi metric is degenerate on certain codimension one hypersurfaces (boundary sets) leading to severe difficulties.

It is worthwhile to mention that in the paper by Baierlein *et al.*[16] we can find that general relativity is governed by a Jacobi-type action. This fact was fully discussed in the paper by Barbour[17] in the context of construction of general relativity as a special case of timeless Machian geometrodynamics. Earlier Kuchar[18] showed that it is a possibility of general reduction of geometrodynamical phase space to a mini phase space by the group of motion.

The Hamiltonian function of simple mechanical systems has the form

$$\mathcal{H}(p,q)=\tfrac{1}{2}g^{ab}p_a p_b+V(q). \tag{1}$$

This function determines the first integral of motion, $\mathcal{H}=E=\text{const}$, where E is the total energy of the system. Trajectories of simple mechanical systems live in the phase space

$$\Omega=\{(q,\dot{q})\in R^{2n}:g_{ij}\dot{q}^i\dot{q}^j=2(E-V(q))\}. \tag{2}$$

Positive definiteness of kinetic energy confines the motion of classical systems to a certain domain D of configuration space with nonempty boundary ∂D,

$$D=\partial D\cup\{q\in R^n:E-V(q)>0\}, \tag{3}$$

where $\partial D=\{q\in R^n:E-V(q)=0\}$.

TABLE I. Examples of simple relativistic dynamical systems of two-dimensional configuration space. In these simple, low-dimensional systems we find or suspect complex (chaotic) behavior of trajectories in the phase space. Two identical trajectories of the system starting at slightly different positions (initial conditions) diverge in time. Such sensitive dependence on initial conditions is the main characteristic of chaotic systems and means that they are difficult to predict over long time scales, practically over the Lyapunov characteristic time.

Relativistic system	Hamilton function	Remarks
Friedman–Robertson–Walker cosmology coupled to real free massive scalar field	$\mathcal{H}=\frac{1}{2}(-p_1^2+p_2^2)+\frac{1}{2}(-q_1^2+q_2^2+m^2q_1^2q_2^2)=0$	$m=$ const[a]
Single scalar field evolving in the idealized de Sitter space	$\mathcal{L}=e^{3\nu t}[\frac{1}{2}\dot{\Phi}^2-\frac{1}{2}e^{-2\nu t}(\nabla\Phi)^2+\frac{1}{2}\mu^2\Phi^2-\frac{1}{4}\lambda\Phi^4-(\mu^2/4\lambda)]$	$\Phi(x,y,z,t)$-scalar field[b]
Charged particle in uniform magnetic field and linearly polarized gravitational wave	$\mathcal{H}=\frac{1}{2}(p_1^2-p_2^2)-\frac{1}{2}\frac{(x^1)^2}{1-\alpha\sin[\nu(x^1-x^0)]}\equiv\frac{1}{2}$	α, amplitude of a wave; ω, angular frequency of a wave; Ω, Larmor angular frequency of a charged particle; $\nu=\omega/\Omega$, relative frequency of wave.[c]
Friedmann–Robertson–Walker model with conformally coupled massive, real, self-interacting scalar field	$\mathcal{H}=\frac{1}{2}[-(p_1^2+kq_1^2)+(p_2^2+kq_2^2)+m^2q_1^2q_2^2+\frac{\lambda}{2}q_2^4+\frac{\Lambda}{2}q_1^4]\equiv 0$	$\Lambda,\lambda,m=$ const; $k=0, \pm 1$[d]

[a]Reference 40.
[b]Reference 41.
[c]Reference 42.
[d]Reference 43.

In the case of relativistic systems the whole R^n is accessible as a configuration space and ∂D plays the role of a boundary set, i.e., in any neighborhood of its elements one can find points belonging to D_+ or to D_-, where $D_+=\{q\in R^n:E-V(q)>0\}$, and $D_-=\{q\in R^n:E-V(q)<0\}$.

According to the classical Maupertuis principle the trajectories of SMS are mapped into broken geodesics of the Jacobi metric $g_E^J=2(E-V(q))g$.[3] The matching procedure at the boundary ∂D is provided by Kozlov's theorem.[9–11] These theorems describe the behavior of geodesics (images of physical trajectories) in a close vicinity of the boundary ∂D. Two properties are fundamental here: (1) trajectories reach the boundary transversally and (2) after the reflection from the boundary the motion takes place along the same trajectory (at the vicinity of the boundary).[3] This means, in particular, that not every piecewise geodesic curve in Jacobi geometry represents a physical trajectory of the system, e.g., curves composed with pieces of boundary on which $ds_J^2=0$. Therefore the direct approach from the Jacobi principle leads to unphysical orbits. As an alternative for this approach one can consider motion in the covering space (see Ref. 3) (in QT space, in the Synge terminology, Q is a normal configuration space, T is a space of absolute time).

In other words the mapping of trajectories via the Maupertuis–Jacobi principle does not provide a one-to-one correspondence. The demand of having an affinely parametrized geodesic introduces a new time-parameter s defined as $ds=2(E-V(q))dt$. By virtue of the energy conservation a tangent vector is normalized to unity in $\mathrm{int}\, D$ and is zero at the boundary ∂D. The Maupertuis principle has been formulated only for systems with empty boundary $\partial D=0$ and therefore is useless in realistic applications. In our previous papers[19] the problem of simple classical systems was reduced to the problem of broken geodesics on the manifolds with boundaries. However, the Jacobi metric is degenerate at the boundary. This circumstance is a source of many difficulties. It is convenient to treat the configuration space as a so-called differential space of constant dimension[3] on which geodesics are appropriately continued through the set of degeneracy of the metric. One should bear in mind, however, that behavior of geodesics on spaces with boundaries can be substantially different from that on smooth manifolds.

In the case of simple indefinite mechanical systems (SIMS) one is able to transfer Kozlov's results concerning behavior of geodesics in the neighborhood of ∂D. The Jacobi metric is now analogous to the case of SDMS but the conformal factor is taken with an absolute value $g_E^J = 2|E-V(q)|g$. In the regions D_+, ∂D, and D_- the tangent vector is normalized (i.e., $\|u\|^2 = g^J_{ab}u^a u^b$) to $+1$, 0, and -1, respectively. Now the problem of simple relativistic systems is reduced to the problem of matched geodesics on manifolds with a singular set. First we consider geodesics of the metrics $g_E^+ = 2(E-V(q))$ and $g_E^- = -2(E-V(q))$ separately but in whole R^n. Then we match them along ∂D so that the tangent vector at the boundary ∂D lies on a cone defined by a kinetic energy form.

Generally, if one wishes to see trajectories of simple classical (or relativistic) mechanical systems as geodesics in configuration space one obtains tractable although complicated geometrical structure having curvature singularities at the boundary ∂D.

In this paper we propose to use the so-called Eisenhart's principle as an alternative tool of reducing the dynamics to a geodesic flow. This principle maps the trajectories of Hamiltonian systems into geodesics on a certain (fictitious) extended configuration space. We present the details of this construction and apply it to minisuperspace.

The extended minisuperspace can be treated as a stationary ''space–time,'' where the additional fictitious dimension plays the role of ''time.'' This approach has several advantages:

(i) The dynamics is mapped onto a geodesic flow on a smooth manifold without the boundary, hence one can use standard theory of smooth manifolds;
(ii) The identification of trajectories of the system with geodesics on an extended configuration space is one-to-one (bijective), this property is not shared by the Jacobi metric picture where not every piecewise geodesic motion on the Jacobi space corresponds to the physical motion of the system;
(iii) Eisenhart's proposal seems to be the right way to desingularize motion in the Jacobi metric (e.g., the dynamics of homogeneous cosmological models near the initial singularity) which is nontractable in the Jacobi picture;
(iv) the interrelation between Jacobi's and Eisenhart's pictures is displayed; the latter produces a quotient structure $\mathcal{M}_{\mathrm{Eis}}/G_u$, where G_u is the translation group with respect to the fictitious ''time'' variable u.

As indicated by the line consequently developed in a series of the papers[3,19] our ultimate goal is to provide a geometrical model of the dynamics of mechanical systems which would be useful both for analytical and numerical analysis of complex motions. Special emphasis is put on the invariant indicators of sensitive dependence on initial conditions. Moreover, we find it attractive to see the dynamics of mechanical systems as geodesics in space–times. One can imagine the behavior of a simple mechanical system as a motion of a particle in a certain effective potential. This way of building up physical intuition permits most of the classical physics (including classical field theory). However, one can reproduce such motions in potential wells as the free (geodesic) motions in some fictitious spaces or space–times. This leads us to Eisenhart's geometry—the subject of this paper—in which the motion of a fictitious particle (representing state of the system) is governed by the Fermat principle of the shortest arrival time. The classical results concerning the formulation of dynamics in terms of Riemannian geometry can be found in Ref. 20.

Recently this principle has been extensively used by Pettini *et al.*[21,22] in their efforts to develop appropriate geometric tools for investigating the deterministic chaos and its connections with the foundations of statistical mechanics. Eisenhart's principle is very promising in the sense of being free from many of problems inherent to the standard (Maupertuis–Jacobi) approach. This principle states that the trajectories of SMS are geodesics in a certain $n+1$-dimensional space where the additional dimension is a fictitious one. In other words the motion of SMS can be represented as a free motion in a certain fictitious "space–time" in which the additional dimension is treated as "time."

In Sec. II, we show the existence and uniqueness of geodesics of the Jacobi metric passing through the singular boundary on which the conformal factor $|f|$ vanishes, and the metric $\overline{g}$ is degenerate. In Sec. III, we study trajectories of simple dynamical systems to formulate the generalized Maupertuis–Jacobi theorem. In Secs. IV–VII the advantages and obstacles of the formulation of dynamics in terms of the Eisenhart geometry are presented. It is pointed out that this approach is attractive in the context of the formulation of a local instability criterion. In Sec. VIII we find the one-to-one correspondence between the trajectories of simple mechanical systems and trajectories of particles or photons in the stationary space–times. In Sec. IX the interconnections between integrability of the original dynamics and integrability in its different models is given. In Sec. X we summarize our main results.

II. THE EXISTENCE AND UNIQUENESS OF GEODESICS OF THE JACOBI GEOMETRY THROUGH THE CONFORMAL SINGULARITY

Let us consider a pair (M,g), where M is a smooth manifold M and g a smooth semi-Riemannian metric on M. Let also $\overline{g}=fg$ be a conformal metric on M with $f:M\to\mathbf{R}$ being a smooth function on M. On the set Ξ of points in M at which the function f vanishes the metric $\overline{g}$ is degenerate; this set is called a *conformally singular set*.

Lemma 1: If $\dim M>1$, *and zero is a regular value of f (i.e., for every* $p\in M$ *such that* $f(p)=0$, *one has* $(df)_p\neq 0$*), the conformally singular set* $\Xi=f^{-1}(0)$ *is closed and is a boundary in M.*

Proof: The set Ξ is a $((\dim M)-1)$-dimensional submanifold of M. □

Let us consider a manifold M carrying the metric $\overline{g}=fg$, where $f=2(E-V)$. We assume that $d_p(f)\neq 0$, i.e., that the particle does not enter tangentially into the conformally singular set Ξ. (There are systems in which such a particle tangentially enters the set Ξ, for instance dynamical systems of the mixmaster type in relativistic cosmology.[19] This phenomenon is connected with the appearance of chaos, e.g., in the series of the Kasner epochs.) For simplicity, let us assume that $(df)(v)>0$ (this is not a general restriction since we can always change the direction of motion, multiplying by -1).

Let us notice that Lemma 1 (and in particular the assumption that zero is a regular value of f) is not required in proofs in Sec. III.

Definition 1: A C^0*-curve* $\gamma:(-\epsilon,\epsilon)\to M$ *is a pregeodesic through* $p\in\Xi$ *if* $\gamma(0)=p$, *and* $\gamma|(-\epsilon,0)$ *and* $\gamma|(0,+\epsilon)$ *are pregeodesics in* $M\backslash\Xi$. *If* γ *is a geodesic in* $M\backslash\Xi$ *we call it geodesic through* $p\in\Xi$.

Recall that a pregeodesic in $\mathcal{M}$ is a smooth curve $\gamma:I\to\mathcal{M}$, $I\subset R$ which can be reparametrized to a geodesic.

Now we shall prove the following theorem.

Theorem 1: *For any* $\alpha\in\mathbf{R}\backslash\{0\}$, $v\in T_p\underline{M}$, *such that* $g(v,v)=0$, *and* $v(E-V)>0$, *there exists the unique geodesic of the connection* $\nabla^{\overline{g}}$,

$$\gamma:(-\epsilon,\epsilon)\to M$$

satisfying the following conditions:

(a) $|f|\circ\gamma(s)\gamma'(s)\to v$, for $s\to 0$;
(b) $|f|g(\gamma',\gamma')\equiv\alpha$, for $s\neq 0$ or $g_E(\gamma',\gamma')\equiv\alpha$, for $s\neq 0$;
(c) $|f|\circ\gamma>0$, for $s\neq 0$,

where $\nabla^{\overline{g}}$ denotes a Levi–Civita connection with respect to metric $\overline{g}$; and $v(f)=v\cdot\operatorname{grad} f$ represents a directional derivative along vector v (f is a smooth function).

In the proof of this theorem we shall use Larsen's theorem 4.1 (Ref. 23) which describes the behavior of geodesics in the neighborhood of conformally singular set. However, the present situation is essentially different from that investigated by Larsen. First, we are considering the behavior of geodesics with respect to the Jacobi metric in their passing through the singularity. Geodesics in our sense are more general than those of Larsen, namely, as geodesics we mean continuous curves which, beyond the singular set, are geodesics in the usual sense. Second, we show that a geodesic (with respect to the Jacobi metric) can pass through the singular set with an *a priori* chosen velocity v (not necessarily $v=0$, as in the Larsen case). Let us notice that for g positive definite, if $\alpha=0$, the existence of γ is violated, but this case is excluded by the condition of $v=0$, the uniqueness of γ is violated (the freedom remains to make an affine reparametrization), but this is excluded by the condition $v(E-V)>0$.

Proof: We shall construct a geodesic passing through the singular set by joining together two geodesics in the usual sense. Let $\alpha\in\mathbf{R}$, $v\in T_pM$, $p\in\Xi$, $g(v,v)=0$. Let us also assume that $v(f)=v(2(E-V))>0$. Of course, $(-v)(-f)>0$. From Larsen's theorem it follows that there exists the unique geodesics (with respect to the metric $\overline{g}$) $\tilde{\gamma}_1:[0,\epsilon]\rightarrow M$ such that

$$(1a)\quad (-f)\circ\tilde{\gamma}_1(s)\tilde{\gamma}_1'(s)\rightarrow -v,\quad \text{for}\quad s\rightarrow 0;$$

$$(2a)\quad -\overline{g}(\tilde{\gamma}_1',\tilde{\gamma}_1')\equiv\alpha,$$

$$(3a)\quad -f\circ\tilde{\gamma}_1>0.$$

Let $\gamma_1:(\epsilon,0]\rightarrow M$ be given by $\gamma_1(s)=\tilde{\gamma}_1(-s)$. It can be easily seen that formulas (1a) and (3a) lead to $(-f)\circ\gamma_1(s)\gamma_1'(s)\rightarrow +v$, for $s\rightarrow 0$ and $-f\circ\gamma_1>0$, respectively.

Once more, by using Larsen's theorem, we easily see that there exists the curve $\gamma_2:[0,\epsilon]\rightarrow M$ such that

$$(1b)\quad f\circ\gamma_2(s)\gamma_2'(s)\rightarrow v,\quad \text{for}\quad s\rightarrow 0;$$

$$(2b)\quad \overline{g}(\gamma_2',\gamma_2')\equiv\alpha,$$

$$(3b)\quad f\circ\gamma_2>0.$$

Now, we define the curve $\gamma:(-\epsilon,\epsilon)\rightarrow M$, where $p=\gamma(0)\in\Xi(M)$, by

$$\gamma(s)=\begin{cases}\gamma_1(s) & \text{for}\quad s\in(-\epsilon,0)\\ \gamma_2(s) & \text{for}\quad s\in[0,\epsilon)\end{cases}.$$

Let us notice that γ is a geodesic satisfying conditions (a)–(c) of the theorem. The uniqueness of γ is the consequence of Larsen's theorem, and it should be understood in the following sense: if $\sigma:(-\overline{\epsilon},\overline{\epsilon})\rightarrow M$ is another geodesic satisfying the above conditions then $\sigma=\gamma$ on $(-\overline{\epsilon},\overline{\epsilon})\cap(-\epsilon,\epsilon)$. □

We can also see that from (2a) and (2b) it follows that the velocity vector $\gamma'(s)$ changes the sector of the cone, when passing through the singular set (at $s=0$), i.e., if before reaching the singular set, $\gamma'(s)$ is timelike with respect to the metric $\overline{g}$, then after passing through the singular set it becomes timelike with respect to the metric $-\overline{g}$, i.e., spacelike with respect to the metric $\overline{g}$.

Let the symbol Im( ) denote an image of the set with respect to some mapping, then geodesics γ_1 and γ_2, out of which γ is composed, are such that Im γ_1 is a subset of the subspace $M_-:=\{x\in M: f(x)\leqslant 0\}$, and Im γ_2 is a subset of the subspace $M_+:=\{x\in M: f(x)\geqslant 0\}$.

Let us consider the classical case when g is a Riemann metric and one has $E-V\geqslant 0$. In this case, $g(v,v)=0$ implies that $v=0$, and consequently theorem 1 describes the behavior of the zero velocity geodesics in a neighborhood of the singular set. The condition $E-V\geqslant 0$ means that the geodesic formed from joining together the geodesics γ_1 and γ_2 is such that both Im γ_1 and Im γ_2 are subsets of M_+. From the uniqueness of this geodesic it follows that Im γ_1=Im γ_2, and

$\gamma_1(s)=\gamma_2(-s)$, for $s\in(-\epsilon,0)$. Therefore, the mechanical system approaches the singular set Ξ, attaining at it the zero velocity, and then goes back along the same trajectory. Such a behavior has been demonstrated by Kozlov, for libration motions, with the help of topological methods.[11,24] Examples of classical simple mechanical systems can be found in Refs. 3, 11.

III. TRAJECTORIES IN THE GENERALIZED MAUPERTUIS–JACOBI METRIC

Let $\gamma:I\to M$, where $I=(a,b)$, be a trajectory of a simple mechanical system, i.e., γ satisfies the Euler–Lagrange equation

$$\nabla_{\gamma'(t)}\gamma'(t)=-(\mathrm{grad}\,V)\gamma(t), \quad '\equiv\frac{d}{dt}, \tag{4}$$

with $t\in I$. Equation (4) is defined on the configuration space M (which is a smooth manifold) carrying the metric $g=(1/2)K(v,v)$, $v\in T_xM$, $x\in M$ (which in general is semi-Riemannian). There exists a differentially continuous (this means continuous on M and smooth on $M\backslash\Xi$) reparametrization of $\gamma(t)$, $\gamma_1(s)=\gamma\circ t(s)$, such that $\gamma_1(s)$ is a geodesic with respect to the Jacobi metric $g_E^J=2|E-V|$ (see Ref. 3).

Let us consider the global reparametrization given by

$$\frac{ds}{dt}=2|E-V(\gamma(t))|, \tag{5}$$

i.e.,

$$s=2\int_0^t|E-V(\gamma(t))|d\tau. \tag{6}$$

Let us further consider the set $I_0\subset I$ of singular parameter values of the curve $\gamma:I\to M$,

$$I_0=\{t\in I:\gamma(t)\cap\Xi\neq\emptyset\}=\{t\in I:E-V(\gamma(t))=0\}.$$

A priori we can distinguish the following cases:

(i) I_0 is a set of isolated points;
(ii) I_0 contains intervals and eventually isolated points;
(iii) $I_0=I$.

In each of these cases, we must reparametrize the curve $\gamma(t)$ so as to obtain a geodesic with respect to the Jacobi metric.

(i) In this case, one can see from Eq. (5) that $(ds/dt)|_{I\backslash I_0}\geqslant 0$ and $(ds/dt)|_{I_0}=0$, and we can define the function $t=t(s)$ which is the inverse function of $s=s(t)$. It can be seen that $\gamma_1(s)=(\gamma\circ t)(s)$ is a geodesic with respect to the metric $\overline{g}$.

(ii) Let, for simplicity, I_0 consist of one interval (α,β). In this case, the function $s=s(t)$ is constant on the interval I_0 and strictly increasing elsewhere. We define the function

$$s^*(t)=\begin{cases} s(t) & \text{for } t\leqslant\alpha \\ s(t+\beta-\alpha) & \text{for } t>\alpha \end{cases}. \tag{7}$$

This function is strictly monotonic, therefore $\gamma^*(s^*)=\gamma(t(s^*))$, where $t(s^*)$ is the inverse function of $s^*(t)$, is a geodesic with respect to the metric $\overline{g}$. If I_0 consists of more than one interval we proceed analogously. [Of course, the function $s^*(t)$ can be interpreted as a function of a parameter on the quotient space if the appropriate equivalence relation identifies points of the interval (α,β).] With isolated points of I_0 we proceed as in (i).

(iii) This is a ''singular case'' in which Im γ is contained in the set Ξ and $s\equiv 0$ (and evidently, as the assumption of Lemma 2.1 that zero is the regular value of f is not satisfied). Let us notice that if $V\equiv E$ on an open subset $U\subset M$ then $\overline{g}=0$ on U and all curves on U can be formally regarded as geodesics with respect to the Jacobi metric $\overline{g}$. [Of course, in this case the Maupertuis principle is devoid of any practical meaning since the solutions of the Euler–Lagrange equation are, from the beginning, (nontrivial) geodesics with respect to the original metric g.]

There are reasons to believe that the condition $I_0\subset I$ means the existence of a period in which trajectories of the system behave in a chaotic unpredictable manner. For instance, the evolution of the Bianchi IX world model, according to the first and second Belinsky–Khalatnikov–Lifshitz approximations, consists in periods of evolving along the boundary $V=E$ and oscillating around it.[25] The reparametrization given by (7) is of purely formal character; it reduces the function $s(t)$ to the strictly monotonic function $s^*(t)$. In this way we remove from our analysis a certain, maybe interesting, part of the evolution. However, it is justified because this part is beyond the control of the Jacobi metric. Moreover, this strategy could be regarded as an approximation of a complex (chaotic) behavior by the regular behavior.

We can see that in all cases the trajectory of the original simple mechanical system can be reparametrized to a geodesic with respect to the Jacobi metric. Therefore, we can formulate our main result.

Theorem 2 (Generalized Maupertuis–Jacobi Theorem): *Every trajectory of any simple mechanical system* (M,g,V), *where* $\dim M>1$, *is a pregeodesic with respect to the Jacobi metric* $g_E=2|E-V|g$ *of type (i) or (ii). Every pregeodesic with respect to the Jacobi metric passing through a point of the boundary set* Ξ *is a trajectory of the system* (M,g,V).

Let us notice that the reparametrization of trajectories of a simple mechanical system leads to geodesics with respect to the Jacobi metrics described in the preceding section.

In our model of dynamics the time parameter s is distinguished. It seems to be against the spirit of relativity; however, in agreement with Synge's intuition,[26] the natural parameter s, for classical mechanical systems, has to be regarded as the proper time of the system, analogously to the proper time of the relativity theory (which in this theory is naturally distinguished).

The aim of our construction of geometry of cosmological dynamics is to show the effectiveness of representing trajectories in terms of geodesics for description of complexity in its dynamical behavior. It is very important because the standard dynamical criteria, like the Lyapunov exponents or the Kolmogorov–Sinai entropy, depend on time parametrization. From the other hand one would like to regard the phenomenon of chaos as an intrinsic property of the system (or as an intrinsic property of the orbit which could be recognized by investigating a fragment of this orbit). Such chaos would be by definition invariant with respect to the change of variables (time and phase variables). It was demonstrated by Cornish and Levin[27] that fractal basin boundaries provide an important means of characterizing chaotic systems in an observer independent manner. In our case laws of dynamics are geodesics principles in the configuration space and there is no need for time (for deeper discussion of the role of time in the Jacobi geometry, see Ref. 17).

IV. EISENHART'S GEOMETRY AS A MODEL OF DYNAMICS

The Eisenhart principle states that there is a one-to-one correspondence between trajectories $q^i(t)$ of the system (1) and geodesics of the metric,

$$ds^2=g_{ij}dq^idq^j+\frac{1}{2(V(q)+b))}du^2, \tag{8}$$

where b is constant. The demand that trajectories are mapped bijectively into geodesics of the metric (8) and conservation of energy produce together the following three relations:

(i) The relation between coordinate time t and the metric interval s

$$t=as, \quad a=\text{const}; \tag{9}$$

(ii) The relation between the fictitious ''time'' u and time coordinate t,

$$\frac{du}{dt}=2(V(q)+b); \tag{10}$$

(iii) Interrelation between the constants a and b [obtained by combining (i), (ii), and the Hamiltonian constraint $\mathcal{H}=E-$conservation of energy],

$$\frac{1}{a^2}=g_{ij}\dot{q}^i\dot{q}^j+2(V(q)+b)=2(E+b). \tag{11}$$

Sometimes one may wish to map the trajectories into null geodesics in an extended configuration space. This is, of course possible but then relation (i) loses its sense since the metric interval is null, (ii) and (iii) hold as above but formally $a=\infty$ and hence $b=-E$. In the case of non-null geodesics the relation (ii) rewritten as

$$u=-2\int T dt+2(E+b)t \tag{12}$$

reveals the connection with the Hamilton–Jacobi principle (for details see Ref. 28).

Let us note that $ds^2\geqslant 0$, i.e., Eisenhart's metric (8) is definite for classical as well as relativistic systems in contrast with the Maupertuis–Jacobi principle where the Jacobi metric is indefinite for relativistic systems. It means that the geodesics of Eisenhart's metric are either spacelike or null.

The determinant of Eisenhart's metric is equal to $g^{\mathrm{Eis}}=(2(V(q)+b))^{-1}g$, hence if g_{ij} is Euclidean, g_{ij}^{Eis} will change its signature to Lorentzian whenever $V(q)+b<0$. Eisenhart's metric is singular at $V(q)=-b$. If we reduce the dynamics to null geodesics then $b=-E$ and in this case singularity of Eisenhart's metric coincides with that of the Jacobi metric. If the potential is bounded as in the case of a harmonic oscillator, then we can choose b so that $V(q)+b\neq 0$ and singularity is avoided. If the potential is not bounded then metric singularity inevitably occurs along certain codimension two hypersurfaces. In Eisenhart's geometry with $g_{ij}=\delta_{ij}$ (an appropriate extension to a general case is straightforward) nonzero components of the Riemann tensor are

$$R_{0i0j}=[8(b+V(q))]^{-3}\left[-3\left(\frac{\partial V}{\partial x^i}\right)^2+2(V(q)+b)\frac{\partial^2 V}{\partial x^i\partial x^j}\right] \tag{13}$$

and the Ricci scalar reads

$$R=\frac{(b+V(q))^{-2}}{m}\left[-\frac{3}{2}(\nabla V)^2+(V(q)+b)\Delta V\right], \tag{14}$$

where ∇ and Δ are operators of gradient and Laplacian in $\mathbf{R}^n$, respectively. The last formula remains valid for the general form of the metric g_{ij} with appropriate understanding of gradient and Laplacian. Thus it can be seen that the singular set $D_{\mathrm{sing}}=\{V(q)=-b\}$ introduces a host of curvature singularities. It means that if we wish to have a simple geometric interpretation (geodesics) of physical systems (with often a complex dynamical behavior) the price we have to pay is the appearance of singular structures as was the case in Jacobi geometry. However, Eisenhart's approach has an advantage that the metric g^{Eis} is not singular on crossing D_{sing} in the same manner like the Jacobi metric g^J was along ∂D when the conformal factor was zero. Now, only one component (g_{00}^{Eis}) changes its sign across D_{sing} thus changing the signature of g^{Eis}. Such a behavior is more familiar and can become tractable, as it will be commented later on. Moreover by appropriate choice of the constant b one can move the singular set outside the domain of one's interests (in the Jacobi geometry ∂D was fixed by fixed E) and in the case when $V(q)$ is bounded the problem can be even desingularized.

Let us note that formulas (13) and (14) are explicitly independent of dimension of the space. One can also see that in the vicinity of critical points of the potential the sign of the Ricci scalar is determined by the Laplacian $\Delta V(q)$, i.e., it is negative in the neighborhood of the maximum and positive near the minimum. These properties are the same as for the Ricci scalar in the Jacobi geometry. In the particular case when the potential is a harmonic function the Ricci scalar takes a simple form

$$R=-\tfrac{3}{2}(\nabla \ln V(q))^2, \tag{15}$$

i.e., it is negative everywhere independently of the number of dimensions.

V. MINISUPERSPACE AS A SPACE OF GEODESICS IN EISENHART'S GEOMETRY

The notion of minisuperspace $\mathcal{MS}$ has been introduced about 20 years ago with the aim of providing an adequate state space for the construction of a wave function for gravity under the restriction to homogeneous degrees of freedom.[12] In particular homogeneous anisotropic cosmological models can be viewed as geodesics on this minisuperspace. The minisuperspace is also a natural setting for studying chaos because the evolution equations provided by the Einstein field equations reduce here to the ordinary differential equations. In the case of a full superspace, one could expect complex spatiotemporal behavior leading to the so-called metric turbulence.

The division between space and time in general relativity comes through foliating the space–time manifold M^4 into spacelike hypersurfaces Σ_t. The metric $g_{\mu\nu}$ on M^4 induces a metric γ_{ij} on Σ_t (the first fundamental form of Σ_t) and can be parametrized in the form

$$g_{\mu\nu}=\begin{pmatrix} N_iN^i-N^2 & N_j \\ N_i & \gamma_{ij}\end{pmatrix},$$

where N and N_i are called a lapse function and a shift vector, respectively. In the $3+1$ ADM formalism, dynamical evolution of geometries (three-metrics $g_A=\gamma_{ij}$) on spacelike hypersurfaces of constant time Σ is governed by the Hamilton equations derived from the Hamiltonian[12]

$$\mathcal{H}=\tfrac{1}{2}G_{ijkl}\pi^{ij}\pi^{kl}+V_G, \tag{16}$$

where $G_{AB}=G_{ijkl}=\frac{1}{2}(\gamma_{ik}\gamma_{jl}+\gamma_{il}\gamma_{jk}-\gamma_{ij}\gamma_{kl})$ is the ultralocal metric on $\mathcal{MS}$, the potential is equal $V_G=\frac{1}{2}\gamma^{(3)}R$, where ${}^{(3)}R$ is the Ricci scalar on Σ. The momenta π_{ij} are defined as equal to the second fundamental form on Σ_t: $\pi_{ij}=(1/2N)(N_{i|j}+N_{j|i}-(\partial\gamma_{ij}/\partial t))$ and '$|$' denotes covariant differentiation with respect to γ_{ij}.

Our idea is to transfer the classical construction of Eisenhart's geometry to the minisuperspace. Eisenhart's metric can be viewed as "static" with respect to an additional dimension u treated as a new "time" dimension (G_{00}^{Eis} component refers to this dimension). This extended minisuperspace $\mathcal{MS}^{\mathrm{Eis}}$ admits a Killing vector $\xi_u^\alpha=\delta_0^\alpha$ which is associated with cyclic nature of variable u. In other words $\mathcal{MS}^{\mathrm{Eis}}$ is invariant with respect to translation group $G_u: u\to u+\Delta u$. We can use this fact to build a quotient structure $\mathcal{MS}^{\mathrm{Eis}}/G_u$ with the property that homogeneous cosmological solutions are geodesics on this space. One can show that the metric on this quotient space is

$$d\hat{s}^2=\frac{1+a^{-2}G_{00}^{\mathrm{Eis}}}{-G_{00}^{\mathrm{Eis}}}dl^2, \tag{17}$$

where $dl^2=G_{AB}dg^Adg^B$. If we reduce the problem to null geodesics, i.e., formally $a=\infty$ the metric (17) takes the form of so called optical metric. (A similar construction but for a different purpose has been used by Abramowicz *et al.*[29]) It means that the reduction of $\mathcal{MS}^{\mathrm{Eis}}$ to lower dimensional space occurs according to Fermat's principle; null geodesics in a space $(\mathcal{MS}^{\mathrm{Eis}},G^{\mathrm{Eis}})$

with topology $\mathcal{MS}^{\mathrm{Eis}}=R\times\mathcal{MS}$ will be projected to $\mathcal{MS}$. The metric on $\mathcal{MS}$ is $(G_{AB}/-G_{00})$ and hence the affine parameter λ along the geodesics is exactly the same as coordinate u.

The quotient structure can be identified with the Jacobi geometry with the metric $G^J_{AB}=2V_G G_{AB}$. Closer look at Misner's discussion of minisuperspace[12] reveals that it is exactly the Jacobi geometry with $E=0$. However the occurrence of metric singularities on the set $\partial D=\{V_G=0\}$ is the major obstruction against this idea.

Let us consider a class of homogeneous cosmological models which admit the Hamiltonian formulation (i.e., Bianchi class A types, see also Ref. 30 for Hamiltonian formulation of class B models). In Bogoyavlenskii coordinates[31] we have

$$\mathcal{H}=T(p_i,q_j)+\tfrac{1}{4}V_G(q^i),$$

$$T(p_i,q_j)=2\sum_{i<j}^{3} p_i p_j q_i q_j-\sum_{i=1}^{3} p_i^2 q_i^2,$$

$$V_G(q_i)=2\sum_{i<j}^{3} n_i n_j q_i q_j-\sum_{i=1}^{3} n_i^2 q_i^2, \tag{18}$$

where $n_i\in\{-1,0,+1\}$ are constants which distinguish between various Bianchi types. An important group of models in class A systems is represented by so called Mixmaster models $[B(\mathrm{IX})\text{—}n_1=n_2=n_3=1$ and $B(\mathrm{VIII})\text{—}n_1=n_2=-n_3=1]$. As it has been noticed by Misner, trajectories of these models near the initial singularity concentrate around $V_G\approx 0$. In the Jacobi picture (which is a standard view on the minisuperspace) the trajectories are thus concentrated around singularities of the Jacobi metric. If we represent the dynamics in Eisenhart's picture the problem disappears. As discussed in Sec. IV Eisenhart's metric changes its signature upon crossing the singular set $D_{\mathrm{sing}}=\{V_G(q)=-b\}$. Let us rewrite Eisenhart's metric as

$$ds^2_{\mathrm{Eis}}=2(V_g(q)+b)du^2+G_{AB}dg^A dg^B. \tag{19}$$

In this notation $V_G(q)+b$ plays the role of a lapse function which is positive for some $u>u_0$, zero for $u=u_0$, and negative for $u<u_0$. Hence Eisenhart's space is a space with a variable signature. Such spaces have been contemplated by Ellis *et al.*[32]

VI. LOCAL INSTABILITY OF BIANCHI IX MODELS

Appropriate tools for describing the local instability constitute a necessary prerequisite for discussing the sensitive dependence on initial conditions (SIDC). It is well known that Poincaré and Lyapunov have introduced two different notions of stability. (The discussion of various notions of stability can be found in Szebehely.[33]) Lyapunov was interested in the behavior of the separation vector connecting the points on neighboring trajectories labeled with the same value of time (or arc-length parameter). Poincaré, on the other hand, introduced what is now called orbital stability as measured by the normal component of the separation vector. In our approach we are using the notion of stability in the sense of Poincaré rather than the Lyapunov conception.

The great advantage of the geometrical model lies in a possibility of formulating the criteria of local instability in terms of geometrical invariants. The simplest (but qualitative) invariant measure of separation of initially close geodesics averaged over all two-directions containing tangent and normal vectors is the Ricci scalar R. Since Eisenhart's metric is non-negatively defined, the (sufficient) criterion of local instability is very simple, $R<0$. Local instability implies the property of sensitive dependence on initial conditions.

Let us apply this criterion (in Eisenhart's picture) for dynamics of the Bianchi IX model near the initial singularity. Eisenhart's metric (8) has the following form:

$$ds^2 = G_{AB} dg^A dg^B + \frac{du^2}{2V(x,y,z) + a^{-2}}, \tag{20}$$

where $V(x,y,z) = \frac{1}{4}(2xy + 2xz + 2yz - x^2 - y^2 - z^2)$ is the potential in Bogoyavlenskii coordinates,[31] $g_A = \mathrm{diag}(x,y,z)$ and the minisuperspace metric G_{AB} is read off from the kinetic energy form

$$T = \frac{1}{4}\left(\frac{\dot{x}}{x}\frac{\dot{y}}{y} + \frac{\dot{x}}{x}\frac{\dot{z}}{z} + \frac{\dot{y}}{y}\frac{\dot{z}}{z}\right),$$

i.e.,

$$G_{AB} = \frac{1}{2xyz}\begin{bmatrix} 0 & z & y \\ z & 0 & x \\ y & x & 0 \end{bmatrix} \tag{21}$$

and if $V=0$, metric (20) is a Minkowski metric which is geodesically complete. Note that the b constant is determined by the Hamiltonian constraint to be $b = 1/2a^2$.

The Ricci scalar in Eisenhart's metric takes the form

$$R = -\frac{4a^2}{(1+2a^2V(x,y,z))^2}\{2(1+3a^2)V(x,y,z) + 3a^2[x^2V_{,x}^2 + y^2V_{,y}^2 + z^2V_{,z}^2]$$
$$+ 2V(x,y,z)(1+a^2V(x,y,z)) + 6a^2(xyV_{,x}V_{,y} + xzV_{,x}V_{,z} + yzV_{,y}V_{,z})\}.$$

It can be seen from the above formula that in the vicinity of $V(x,y,z) \approx 0$ states which correspond to a dynamical regime approximated by a series of the Kasner epochs, the Ricci scalar is negative and singularity-free. This allows us to hope that invariant description of the sensitive dependence on initial conditions in terms of curvature invariants of Eisenhart's metric is a right way to desingularize the standard Jacobi approach to minisuperspace.

VII. DESINGULARIZATION OF THE JACOBI METRIC THROUGH AN ADDITIONAL DIMENSION

As it was mentioned before the Jacobi metric is degenerate on the boundary $E - V(q) = 0$ of space admissible for motion which in turn leads to a metric singularity. The occurrence of the metric singularity is the major obstruction against the idea of using the sectional curvature in detecting local instability. Having the Jacobi metric $g_E^J = 2|E - V(q)|g$, one may perform an isometric embedding of the space with the Jacobi metric into the flat Euclidean space with the metric

$$ds^2 = g_{ij}dz^i dz^j + (z^{n+1})^2 - (z^{n+2})^2, \tag{22}$$

where

$$z^i = q^i\sqrt{2|E - V(q)|},$$

$$z^{n+1} = \tfrac{1}{2}(r^2 - 1)\sqrt{2|E - V(q)|},$$

$$z^{n+2} = \tfrac{1}{2}(r^2 + 1)\sqrt{2|E - V(q)|},$$

$$r^2 = g_{ij}q^i q^j.$$

The Jacobi geometry is now realized on the null cone

$$g_{ij}z^iz^j+(z^{n+1})^2-(z^{n+2})^2=0.$$

The vertex of the cone is the relic from the singular boundary set ∂D in the Jacobi metric.

On the other hand there is a consistent description of calculations of the geometrical quantities and invariant functionals of the metric on the conical defects. To elucidate this problem, let us recall that a cone is everywhere flat space except the tip where its curvature R is singular. Obviously, calculations by means of standard formulas of the Riemannian geometry cannot reveal this δ-like singularity, and the other methods must be used to obtain a correct result (for details see Ref. 34). Therefore initially dangerous singularities of the Jacobi metric takes the weak form for which there is some procedure of regularization in calculation of invariants of curvature.

Let us consider the space $\mathcal{M}$ with Eisenhart's metric g^{Eis} for a simple indefinite mechanical system and because metric g^{Eis} is stationary with respect to a given timelike Killing vector Y, and $\mathcal{M}$ admits a global space–time splitting $U\times\mathbf{R}$ adapted to Y. Roughly speaking, this means that $\mathcal{M}$ admits a global coordinate system $(q_1,\ldots,q_n,u)$ with $(q_1,\ldots,q_n)\in U$ open subset of $\mathbf{R}^n$, $t\in\mathbf{R}$ and $Y=\partial/\partial u$. The Killing property of Y is given by the fact that the coefficients of the metric $\mathcal{M}$ do not depend on the ''time'' variable u. The existence of the Killing vector in turn can be used to build a quotient space $\mathcal{M}^{\mathrm{Eis}}/G_u$, where G_u is a group of symmetry generated by the Killing vector. The metric on $\mathcal{M}^{\mathrm{Eis}}/G_u$ can be established on the base of the following theorem:

Theorem 3 (A timelike extension of Fermat's principle in GR): *Let* $(\mathcal{M},g)$ *be a stationary space–time with a Lorentzian signature* $(-,+,\ldots,+)$ *then the problem of geodesics on* $(n+1)$*-dimensional space–time*

$$\delta\left(\int ds^2_{n+1}\right)=0$$

can be reduced to the geodesics problem on n-dimensional fictitious conformally rescaled Riemannian space with the metric

$$d\overline{l^2}=\frac{1+hg_{00}}{-g_{00}}dl^2,\quad \delta\left(\int d\overline{l^2_n}\right)=0,$$

where $dl^2=g_{ij}dq^idq^j$, $g=g_{00}du^2+dl^2$*—metric of space–time,* $h=\mathrm{const}$; *h is positive for timelike geodesics, negative for spacelike geodesics, and* $h=0$ *for null geodesics.*

If we put $h=a^{-2}$ we obtain n-dimensional metric on $\mathcal{M}^{\mathrm{Eis}}/G_u$. It is natural to identify this metric $d\overline{l^2}$ as the Jacobi metric for a simple mechanical system $2(E-V)g_{ij}$. After this comparison we can obtain a theorem which establishes one-to-one correspondence between trajectories of simple definite systems and trajectories of fictitious particles in space–time with the Eisenhart geometry.

VIII. TRAJECTORIES OF SIMPLE MECHANICAL SYSTEMS FROM THE MOTION OF A FICTITIOUS PARTICLE IN A STATIONARY BACKGROUND

The results of the previous section show that the study of a simple mechanical system can be translated into the study of the test particle and photon motion in space–times of general relativity. There are several advantages of doing so. The majority of problems can be reduced to problems of geodesic motions, for which the notion of differential geometry on manifolds often gives more transparent and deeper insight into the underlying symmetry. Moreover, a geodesic motion can be also formulated as a Hamiltonian system, and all the techniques of searching for integrals developed for Hamiltonian dynamics can be used to obtain integrals which do not admit an obvious geometrical interpretation.

On the metric manifold, metric (8) (not necessarily Riemannian) is written in the condensed form, and then the geodesic motion is determined by the Hamiltonian

$$\mathcal{H}=\frac{1}{2m}g(P,P)=\frac{1}{2m}g^{\mu\nu}P_{\mu}P_{\nu} \tag{23}$$

which is equivalent to the geometrical equations of motion in terms of covariant derivatives.

Let us consider a particle trajectory with the momentum components P^{μ} in the space–time manifold with coordinates x^{ν} ($\mu,\nu=0,...,3$ and $x^0=u$, $x^i=q^i$). The corresponding mass parameter m is given by the condition

$$m^2=-g_{\mu\nu}P^{\mu}P^{\nu}, \tag{24}$$

where m represents the rest mass.

From hereafter we assume that space–time $\mathcal{M}$ with Eisenhart metric has the Lorentzian signature $(-+++)$ (the minus sign refers to the ''time'' u direction). The results for the Euclidean signature can be reinterpreted in a simple way. Any trajectory $x^{\mu}(\lambda)$ may be conveniently regarded as an integral curve of the following equations:

$$m\frac{dx^{\mu}}{d\lambda}=P^{\mu}=g^{\mu\nu}P_{\nu}, \tag{25}$$

where λ is an affine parameter along geodesics, and P^{μ} is determined from Eq. (24) as a function of x^{ν}.

Without any loss of generality we can consider only the case of a stationary (static) metric (8). This metric is characterized by the existence of a timelike Killing vector K^{μ}. Thus, it is possible to choose a frame of reference with a fictitious time coordinate

$$x^0=u, \tag{26}$$

with respect to which we have $K^{\mu}=\delta^{\mu}_0$. This means that the corresponding partial derivative of the metric is zero, i.e., $\partial g_{\mu\nu}/\partial x^0=0$. Thus, the vector field P^{μ} is stationary, i.e., $\partial P^{\mu}/\partial x^0=0$. This allows us to include directly a projected trajectory given by

$$m\frac{dx^i}{d\lambda}=P^i, \tag{27}$$

into the n-dimensional quotient manifold $\mathcal{M}/\mathcal{G}$, where $\mathcal{G}$ is a group of ''time'' transformations $u\to u+\Delta u$. The coordinate Killing vector $(\partial/\partial u)$—the generator of the infinitesimal group of isometry—is associated with the action of this group.

This quotient manifold has an induced positively-defined metric with components γ_{ij} which can be read out from the full $(n+1)$-dimensional metric by decomposing it to the form

$$ds^2=g_{00}du^2+2g_{0i}du\,dx^i+\gamma_{ij}dx^idx^j=g_{00}du^2+\gamma_{ij}dx^idx^j=g_{00}du^2+dl^2.$$

This is equivalent to setting

$$g_{00}=A,\quad g_{0i}=0,\quad g_{ij}=\gamma_{ij}.$$

Now, let us introduce a new conformally modified positive definite metric

$$d\hat{l}^2=\hat{g}_{ij}dx^idx^j, \tag{28}$$

on the quotient $\mathcal{M}/\mathcal{G}$ space by setting

$$d\hat{l}^2=\frac{1+hg_{00}}{-g_{00}}dl^2, \tag{29}$$

where $h=$const. For a null geodesic $h=0$, whereas for timelike geodesics $h>0$, and $h<0$ for a spacelike one.

Our aim is to find one-to-one correspondence between the motion of particles or photons in space–time background and trajectories of a simple mechanical system. Thus, it is natural to compare the above metric with the Jacobi metric; then we obtain

$$\frac{1+hg_{00}}{-g_{00}}=2(E-V). \tag{30}$$

Metric (29) is a positive definite metric on the quotient space $\mathcal{M}/\mathcal{G}$ of the dimension n. The constant h in (29) is related to the proper energy defined as the total energy of the particle per its mass $\bar{E}=\mathcal{E}/m$. Notice that u is a cyclic coordinate for the system with Lagrangian $\mathcal{L}=(m/2)\times(ds/d\lambda)^2$.

Thus, the corresponding momentum has to be conserved

$$P_0=\frac{\partial\mathcal{L}}{\partial\left(\frac{du}{d\lambda}\right)}=mg_{00}\frac{du}{d\lambda}=-m\bar{E}=-m\frac{1}{\sqrt{h}}. \tag{31}$$

This implies the relation between h and $\bar{E}$,

$$\mathcal{E}=m\bar{E}=m\frac{1}{\sqrt{h}}\rightarrow h=\bar{E}^{-2}. \tag{32}$$

From (29), (30), and (32) we obtain

$$h=-a^2=\bar{E}^{-2}. \tag{33}$$

Let us notice that in the Eisenhart geometry we study the spacelike geodesics and thus $h<0$ or $\bar{E}$ is pure imaginary.

If $a\rightarrow\infty$, i.e., for the case of null geodesics we have $b=-E$ and $h=0$. Relations (32) and (33) establish the one-to-one correspondence between the Jacobi geometry of simple classical dynamical systems and the geometry of fictitious particles moving in the space–time with the Eisenhart metric.

In the special case $h=0$, the metric (29) coincides with the so-called Fermat or optical metric. Abramowicz *et al.*[29] studied the role of this optical reference geometry for describing a test particle trajectory in the conformally projected three-space with metric (29) for $h=0$. With such a projection in the static space–time null lines of the four-dimensional manifold correspond to the three-dimensional space geodesics. One can easily see this fact considering Fermat's principle in its relativistic formulation.[35] This principle states that if $\mathcal{M}=R\times\Sigma$ is a static space–time with the metric $g=g_{00}dt^2+g_{ij}dx^i dx^j$, where Σ is a 3-manifold of constant time with Riemannian metric ${}^{(3)}g$, and $g_{00}<0$ is a smooth function. Neither function g_{00} nor metric ${}^{(3)}g$ depend on t. Thus, the null geodesics of $(\mathcal{M},g)$, when projected onto Σ, are precisely the Riemannian geodesics of the 3-geometry

$$\left(\Sigma\ ,\frac{{}^{(3)}g}{-g_{00}}\right), \tag{34}$$

and, furthermore, the affine parameter λ (i.e., the arc length) along the projected geodesics in g metric is precisely the static time coordinate t measured along the null geodesics in $(\mathcal{M},g)$. The above principle has a simple generalization to the case of non-null geodesics.[36]

On the other hand, we can regard the variational principle in the reduced space

$$\delta\int d\hat{l}^2=0 \Leftrightarrow \delta\int n^2(x^1,\ldots,x^n)dl^2=0, \tag{35}$$

as the variational principle in geometrical optics considering the problem of a light beam in an inhomogeneous medium characterized by the refraction factor $n(x^j)$ in space with metric dl^2. Therefore, instead of studying the problem of geodesics (null, spacelike or timelike) in the Eisenhart metric, one can equivalently investigate the problem of geodesics in a Riemannian or pseudo-Riemannian manifold with metric (29). The possibility of such a reduction appears as a consequence of ''the static form'' of the space–time metric. From the mathematical point of view, the reduced space corresponds to the conformally adjusted quotient space metric.

IX. INTEGRABILITY OF DYNAMICS IN DIFFERENT GEOMETRIES

There are several definitions of integrability. Generally, integrability means that the system under consideration possesses a sufficiently large number of first integrals. To be more precise, a Hamiltonian system with n degrees of freedom is integrable if it possesses n functionally independent first integrals which are in involution or which form a solvable Lie algebra. It is necessary to specify a class of functions which contains these first integrals as well as to define a domain of their definition.

The Maupertuis or Eisenhart geometry are only some geometrical models of given dynamical behavior. Of course, a given dynamics can admit many models, and different time parameters can be distinguished by different models. The only fact that matters is the existence of an isomorphism between the original dynamics and its model. Unfortunately, discovering a certain property in one of the models (e.g., existence of a first integral) we also find that this property cannot necessarily belong to any model. However if we want to prove nonintegrability in the original dynamics it is enough to find a single model of this dynamics in which the system is nonintegrable.

In this section interconnections between the property of existence of the first integral in different Eisenhart and Jacobi geometries are considered. To illustrate this let us consider a special case of a simple mechanical system given by the Hamiltonian

$$H=\tfrac{1}{2}g^{\alpha\beta}p_\alpha p_\beta+V, \tag{36}$$

where $g^{\alpha\beta}$ is a constant metric. Let,

$$H^J=\frac{1}{4(E-V)}g^{\alpha\beta}p_\alpha p_\beta=E^J=\frac{1}{2} \tag{37}$$

be the Hamiltonian for the Jacobi flow on $\mathcal{M}_E$. Then one can formulate the following theorem:

Theorem 4: *If $F(q^\alpha,p_\alpha)$ is a first integral of (36) then F is the first integral of (37) for H^J* $=1/2$.

Proof: Let $F(q^\alpha,p_\alpha)$ be a first integral of (36). Then,

$$g^{\alpha\beta}p_\beta\frac{\partial F}{\partial q^\alpha}-\frac{\partial V}{\partial q^\alpha}\frac{\partial F}{\partial p_\alpha}=0.$$

Thus,

$$\frac{1}{2(E-V)}g^{\alpha\beta}p_\beta\frac{\partial F}{\partial q^\alpha}-\frac{1}{2(E-V)}\frac{\partial V}{\partial q^\alpha}\frac{\partial F}{\partial p_\alpha}=0,$$

and we can write for points on $\mathcal{M}_E$,

$$\frac{\partial H^J}{\partial p_\alpha}\frac{\partial F}{\partial q^\alpha}-\frac{\partial H^J}{\partial q^\alpha}\frac{\partial F}{\partial p_\alpha}=0$$

for $H^J=1/2$.

From the other hand one can also prove the following:

Theorem 5: *If $F(q^\alpha,p_\alpha)$ is a first integral of (37) then $F(q^\alpha,p_\alpha)$ is a first integral of (36) on set $\mathcal{M}_E=\{(q,p):E-V>0\}$.*

Proof: Because F is a first integral of (37), then

$$\frac{1}{4(E-V)}g^{\alpha\beta}p_\beta\frac{\partial F}{\partial q^\alpha}-\frac{1}{E-V}H^J\frac{\partial V}{\partial q^\alpha}\frac{\partial F}{\partial p_\alpha}=0.$$

Taking points $(q,p)\in\{H^J=1/2\}$ we obtain that

$$g^{\alpha\beta}p_\beta\frac{\partial F}{\partial q^\alpha}-\frac{\partial V}{\partial q^\alpha}\frac{\partial F}{\partial p_\alpha}=0,$$

i.e., that F is a first integral of (36) on $\mathcal{M}_E$.

Finally one can formulate the following conclusions concerning the property of existence of a first integral in the original SMS and its geometrical model based on the Jacobi geometry.

(1) The existence of a first integral of SMS gives us the first integral in the Jacobi geometry but they are only partial first integrals on the energy level $H^J=1/2$.
(2) The first integrals of H^J determines the first integrals of H (i.e., SMS) but they are only partial first integrals on set $H^J=1/2$.
(3) If we find a first integral F for Hamiltonian H^J (i.e., in the Jacobi geometry) and for any value of total energy E, i.e., $\forall E\in R$ the commutator $\{H^J,F\}=0$ then F is also the first integral of H.

The analogous relations can be established in Eisenhart's geometry model of dynamics. Let us consider the Hamiltonian for the Eisenhart metric in the unified form

$$H_E=\frac{1}{2}g^{\alpha\beta}p_\alpha p_\beta,\quad \alpha,\beta=1,...,n+1, \tag{38}$$

where metric $g^{\alpha\beta}$ has the block diagonal form

$$g^{\alpha\beta}=\left[\begin{array}{c|c} g_{ij} & \\ \hline & \frac{1}{2(V+b)}\end{array}\right].$$

One can consider, without loss of the generality, a special form of this metric, namely $g_{ij}=\delta_{ij}$. Then equations of motion take the form

$$\frac{dq^i}{ds}=p_i,\quad \frac{dp_i}{ds}=-\frac{\partial V}{\partial q^i}p_u^2,$$

$$\frac{du}{ds}=2(V+b)p_u,\quad \frac{dp_u}{ds}=0, \tag{39}$$

and

$$ds^2=g_{ij}dq^idq^j+\frac{1}{2(V+b)}du^2.$$

It is easy to notice that our system has the first integral $p_u = g_{uu} dq^u/ds = du/ds/[2(V+b)] = \text{const}$, which gives us the relation between the natural Eisenhart parameter and additional variable $u:(du/ds) \approx 2(V+b)$.

If the original SMS has a first integral, say $F(q^i, \dot{q}^i)$, then

$$\dot{q}^i \frac{\partial F}{\partial q^i} - \frac{\partial V}{\partial q^i} \frac{\partial F}{\partial \dot{q}^i} = 0, \tag{40}$$

where the dot denotes differentiation with respect to mechanical time t.

From (39) we obtain that $\dot{q}^i = p_i/a$ and $\dot{q}^i = \sqrt{2(E+b)}(dq^i/ds)$ which from (40) implies

$$\frac{1}{a^2} p_i \frac{\partial F}{\partial q^i} - \frac{\partial V}{\partial q^i} \frac{\partial F}{\partial p_i} = 0.$$

Therefore if we choose $a=1$ then the first integral of SMS $F(q^i, \dot{q}^i)$ gives us the first integral for geodesic motion in the Eisenhart geometry $\tilde{F}(q^i, p_i)$.

The advantage of Eisenhart's picture of dynamics is that the first integral of geodesic motion determines (in contrast to the Jacobi geometry) all first integrals of the original system. Therefore from searching the first integral for the geodesic flow we simply obtain first integrals of the original system. This fact is a consequence of the following theorem:

Theorem 6: *If $F(q^i, \dot{q}^i)$ is a first integral of equations of motion for SMS, then $F(q^i, \dot{p}_i)$ is a first integral of the Hamiltonian equation with the Eisenhart Hamiltonian $H_E = \frac{1}{2}[\Sigma p_i^2 + 2(V+b)p_u^2]$ which corresponds to $p_u = a^2$.*

Moreover, if $G(q^i, p_i)$ is a first integral of equations of motion for Hamiltonian H_E which is independent on u and p_u, then $G(q^i, \dot{q}^i)$ is the first integral of the original equations of motion for SMS.

Proof: The fact that $G(q^i, p_i)$ is the first integral for Hamiltonian H_E implies the relation

$$p_i \frac{\partial G}{\partial q^i} - \frac{\partial V}{\partial q^i} p_u^2 \frac{\partial G}{\partial p_i} = 0.$$

After substitution of (39) into above formula we obtain

$$a\dot{q}^i \frac{\partial G}{\partial q^i} - \frac{\partial V}{\partial q^i} \frac{p_u^2}{a} \frac{\partial G}{\partial \dot{q}^i} = 0.$$

Now if we put $p_u^2/a^2 = 1$ we obtain the condition for existence of the first integral of SMS.

It seems that our results concerning integrability in the original dynamics and in its models can be generalized to the case of any g_{ij}.

X. CONCLUSIONS

There are reasons to believe that Szydłowski *et al.*'s paper[3] together with the present work provide a strong mathematical basis for the Maupertuis principle as applied to simple mechanical systems. Until now this principle could be only used for $(E-V)>0$ which strongly limited its physical relevance. Now, this limitation has been removed, and in this way the Maupertuis principle has obtained the universal character. Theorem 2 reduces the investigation of the trajectories of a simple mechanical system to the problem of geodesics with respect to the Jacobi metric. These geodesics are obtained by ''gluing together'' two corresponding geodesics (in the usual sense). Owing to this procedure singular points are treated on equal footing with other (regular) points. The Maupertuis principle understood in this way has a global character. Moreover, as it was pointed out by Barbour[17] the Maupertuis principle is more general than Newton's laws since it is, in the natural way, valid also for relativistic systems, which is not true as far as Newton's laws are concerned.

It would be interesting to elucidate interconnection between the topology of the configuration space and the behavior of trajectories of relativistic mechanical systems. Such interconnections, for the case $E>V$ were investigated by Kozlov.[11]

We have demonstrated the attractiveness of Eisenhart's principle from the point of view of constructing invariant criteria of local instability in general relativity (at least as limited to homogeneous degrees of freedom). The key idea is to map the evolution of homogeneous three-metrics under the action of Einstein's equations to a geodesic motion in extended minisuperspace equipped with Eisenhart's metric. The correspondence between the Eisenhart and more familiar Jacobi pictures was discussed. The latter approach is known (together with associated problems) since seminal works of Wheeler and DeWitt[37] and Misner.[12] The transition between $\mathcal{MS}^{\mathrm{Eis}}$ and $\mathcal{MS}^{J}$ is accomplished by constructing a quotient space $\mathcal{MS}^{\mathrm{Eis}}/G_u$, where G_u is the group of translations in u.

We have shown that Eisenhart's approach effectively desingularizes the Jacobi geometry of the minisuperspace in the vicinity of $V\approx 0$ states. This is of a fundamental importance for analysis of Mixmaster cosmological models near the initial singularity. The dynamics of these models can be approximated by a series of Kasner epochs, for which $V\approx 0$. This condition has been creating serious difficulties for the Jacobi approach—the appearance of infinities in the geodesic deviation equation.

Eisenhart's metric can prove useful also on the ground of classical mechanical systems. In particular we are able to restate classical mechanical problems in terms of geodesic motion in a ''static space–time.'' It can produce a feedback between the results obtained in classical mechanics and theory of relativity to build up a better physical intuition of various problems. Appropriate results will be presented in separate papers.

Geometrization according to the Eisenhart's principle should be perceived in a wider context. Namely, one is able to build geometrical models of dynamics in different spaces, e.g., Euclidean, Riemannian, Weyl spaces with affine connection, Einstein–Cartan spaces with torsion, Finsler spaces, etc.[38] The effectiveness of using the Finsler spaces for geometrizing the dynamical problems in rotating frames, where centrifugal forces do not allow us to use the Maupertuis principle directly, was demonstrated by Pettini.[21]

Our considerations were strongly focused on the invariant measures of local instability. The sensitive dependence on initial conditions is the one of the properties appearing in the definition of chaotic behavior.[14] It has been stated by Wiggins as a demand that any two arbitrarily close points evolved under the action of a dynamical system should eventually diverge. This property depends on the metric structure of the space (and also on its topology) in which we embed the trajectories of the system. The advantage of such a formulation of the SDIC is that one is not demanding that the separation be exponentially fast. In fact there are known examples of chaotic systems where the divergence of trajectories is slower than exponential. Moreover the rate of separation depends on the choice of time variable which can be arbitrary in general relativity.

If we construct a geometric model of the dynamics then we can capture this beautiful mathematical idea of the SDIC (see Ref. 14) in terms of geometric curvature invariants. The main problem is to make these quantities regular, i.e., to circumvent the divergencies appearing as the artifacts of the construction. In this respect Eisenhart's picture (although not the only one or by no means the best one) is promising since it allows us to make the singular sets movable and sometimes even regularize the problem completely.

Although it was not our key point here the results of our paper could be perceived in the context of invariant characterization of chaos. Therefore some remarks are in order. It is known that in the case of a manifold without boundary local instability criterion for a geodesic flow can be formulated in terms of sectional curvatures[39] $K_{u;n}g(u,u)<0$, where $K_{u;n}$ is a sectional curvature in two directions spanned on a tangent vector u and a normal n to geodesic. The above condition can be formulated in a more practical way by means of invariant polynomials (one of which is the Ricci scalar). This should hold for every two-direction in every point of the manifold. Whenever $g(u,u)>0$, i.e., positive definite metric (e.g., Eisenhart's metric) or spacelike tangent vector, a standard instability criterion for geodesics is recovered meaning that the system has

SDIC property. Of course, it is not sufficient for chaos to occur. Namely, one demands that the invariant submanifolds be compact (or closed). In our example of the harmonic oscillator there is no chaos despite the negative curvature Eisenhart's space. The reason is that the Eisenhart's space is not compact here. One should stress that although the approach sketched above has already been discussed in the context of the Jacobi geometry by many people (including ourselves) it has not been properly appreciated that the whole idea does not work for spaces with boundaries. In such cases the boundary effects can (and often are) dominant. Moreover very little is known about mathematical properties of the manifolds with boundaries, so it is dangerous to transfer here the intuition acquired by studying smooth (and often compact) manifolds. The Eisenhart's geometry depicted in this paper has an advantage of being boundary free and one may hope that it points toward a right way to proceed in quest for geometric indicators of SDIC.

ACKNOWLEDGMENTS

The author would like to thank Professor A. Staruszkiewicz for constructing remarks clarifying and improving the presentation of our results (especially the nomenclature of definite and indefinite systems). The author is grateful to Professor S. Bazanski and Dr. A. Krawiec for many inspiring discussions. Financial support from the Polish State Committee for Scientific Research (KBN) under Grant No. 2P03D 022 10 is gratefully acknowledged. The author is also very grateful to the referee for drawing the attention to the Baierlein *et al.*, Lanchos, Kuchar papers and for very useful comments.

[1] V. I. Arnold, *Mathematical Methods in Classical Mechanics* (Springer, New York, 1989).
[2] C. P. Ong, Adv. Math. **15**, 269 (1975).
[3] M. Szydłowski, M. Heller, and W. Sasin, J. Math. Phys. **37**, 346 (1996).
[4] R. Arnowitt, S. Deser, and C. W. Misner, ''The dynamics of general relativity,'' in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[5] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, New York, 1971).
[6] S. L. Bażański, J. Phys. A **27**, 3221 (1994).
[7] A. Wintner, *The Analytical Foundations of Celestial Mechanics* (Princeton University Press, Princeton, 1941).
[8] H. Seifert, Math. Z. **51**, 179 (1948).
[9] V. V. Kozlov, Appl. Math. Mech. (in Chinese) **40**, 399 (1976).
[10] V. V. Kozlov, Vestn. Mosk. Inst. Mat. Mech. **5**, 118 (1977).
[11] V. V. Kozlov, Usp. Mat. Nauk **40**, 33 (1985).
[12] C. W. Misner, ''Minisuperspace,'' in *Magic Without Magic*, edited by J. Klauder (Freeman, San Francisco, 1972).
[13] B. S. DeWitt, ''Space–time as a sheaf of geodesics in superspace,'' in *Relativity*, edited by M. Carmeli, S. I. Fickler, and L. Witten (Plenum, New York, 1970).
[14] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos* (Springer, New York, 1990).
[15] H. G. Schuster, *Deterministic Chaos, An Introduction* (VCH, Weinheim, 1987).
[16] R. F. Baierlein, D. H. Sharp, and J. A. Wheeler, Phys. Rev. **126**, 1864 (1962).
[17] J. B. Barbour and B. Bertotti, Proc. R. Soc. London, Ser. A **382**, 295 (1982); J. B. Barbour, Class. Quantum Grav. **11**, 2853 (1994).
[18] K. Kuchar, Phys. Rev. D **4**, 955 (1971).
[19] M. Szydłowski, J. Math. Phys. **35**, 1850 (1994); M. Szydłowski and A. Krawiec, Phys. Rev. D **53**, 6893 (1996); **47**, 5323 (1993); M. Szydłowski and M. Biesiada, *ibid.* **44**, 2369 (1991).
[20] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Cummings, Massachusetts, 1978); C. Lanczos, *The Variational Principles of Mechanics* (University of Toronto Press, Toronto, 1957); A. J. McConnell, *Application of Tensor Analysis* (Dover, New York, 1957); E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge University Press, Cambridge, 1959); J. L. Synge, Philos. Trans. R. Soc. London, Ser. A **226**, 91 (1926); J. Moser, Commun. Pure Appl. Math. **23**, 609 (1970).
[21] M. Pettini, Phys. Rev. E **47**, 826 (1993).
[22] L. Casetti, C. Clementi, and M. Pettini, Phys. Rev. E **54**, 5969 (1996); M. Cerruti-Sola and M. Pettini, *ibid.* **53**, 179 (1996); **51**, 53 (1995); L. Casetti, R. Livi, and M. Pettini, Phys. Rev. Lett. **74**, 375 (1995).
[23] J. C. Larsen, J. Geom. Phys. **9**, 3 (1992).
[24] V. V. Kozlov, Dokl. Akad. Nauk SSSR **249**, 1299 (1979); S. Bolotin, J. Math. Phys. **1**, 275 (1994).
[25] J. D. Barrow, Phys. Rep. **85**, 1 (1982); V. A. Belinski, E. M. Lifszyc, and I. M. Khalatnikov, Adv. Phys. **19**, 525 (1970).
[26] C. Lanczos, *The Variational Principles of Mechanics* (Toronto University Press, Toronto, 1949); J. L. Synge, *Classical Dynamics, Handbuch der Physik III/1*, edited by S. Flüge/Marburg (Springer, Berlin, 1960).
[27] N. J. Cornish and J. J. Levin, Phys. Rev. D **53**, 3022 (1996); Phys. Rev. Lett. **78**, 998 (1997).
[28] L. P. Eisenhart, Ann. Math. **30**, 591 (1929).
[29] M. A. Abramowicz, B. Carter, and J. P. Lasota, Gen. Relativ. Gravit. **20**, 1173 (1988); **20**, 275 (1988).
[30] M. P. Ryan Jr. and S. M. Waller, 1997, gr-qc/9709012.

[31] O. Bogoyavlenskii, *Methods in the Qualitative Theory of Dynamical Systems in Astrophysics and Gas Dynamics* (Springer, New York, 1985).
[32] G. F. R. Ellis, A. Sumeruk, and C. Hellaby, Class. Quantum Grav. **9**, 1535 (1992).
[33] V. Szebehely, Celest. Mech. **34**, 49 (1984).
[34] D. V. Fursaev and S. N. Solodukhin, Phys. Rev. D **52**, 2133 (1995); D. D. Sokolov and A. A. Starobinsky, D A SSSR **22**, 312 (1997).
[35] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[36] M. Szydłowski, J. Math. Phys. **35**, 1850 (1994); M. Szydłowski and A. Krawiec, Phys. Rev. D **53**, 6893 (1996).
[37] B. DeWitt, Phys. Rev. **160**, 1113 (1967).
[38] H. Goenner, E. W. Grafarend, and R. J. You, Manuscr. Geodetica **19**, 339 (1994).
[39] M. Szydłowski, Contemp. Math. **203**, 289 (1997).
[40] E. Calzetta and C. E. Hasi, Class. Quantum Grav. **10**, 1825 (1993).
[41] A. H. Guth and S.-Y. Pi, Phys. Rev. D **32**, 257 (1985).
[42] H. Varvoglis and D. Papadopoulos, Astron. Astrophys. **261**, 664 (1992).
[43] S. Blanco, A. Costa, and O. A. Rosso, Gen. Relat. Grav. **27**, 1295 (1995).

Positive mass conjecture for five-dimensional Lorentzian manifolds

Xiao Zhang[a)]
Morningside Center of Mathematics and Institute of Mathematics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China

(Received 2 September 1997; accepted for publication 26 March 1999)

We extend Witten's proof on positive mass conjecture to five-dimensional Lorentzian manifolds with a rigorous treatment in mathematics. © *1999 American Institute of Physics.* [S0022-2488(99)00707-0]

I. INTRODUCTION

Let N be an $(n+1)$-dimensional Lorentzian manifold with Lorentzian metric $\tilde{g}$ of signature $(-1,1,...,1)$, which satisfies the Einstein equations

$$\tilde{R}_{\alpha\beta}-\frac{\tilde{R}}{2}\tilde{g}_{\alpha\beta}=T_{\alpha\beta}, \tag{1.1}$$

where $\tilde{R}_{\alpha\beta}$, $\tilde{R}$ are the Ricci and scalar curvatures of $\tilde{g}$, respectively, $T_{\alpha\beta}$ is a symmetric tensor field which is interpreted physically as the energy-momentum tensor of matter. Choosing an orthonormal frame $\{e_\alpha\}$ with e_0 timelike. Then, physically, T_{00} is interpreted as the local mass density, and T_{0i} is interpreted as the local angular momentum.

Definition 1.1: A spacelike hypersurface M of N is called asymptotically flat of order τ if there is a compact set $K\subset M$ such that $M-K$ is the disjoint union of a finite number of subsets $M_1,\ldots,M_k$—called the ''ends'' of M—each diffeomorphic to the complement of a contractible compact set in R^n. Under the diffeomorphism the metric of $M_l\subset N$ is of the form

$$g_{ij}=\delta_{ij}+a_{ij} \tag{1.2}$$

in the standard coordinates $\{x^i\}$ on R^n, where

$$a_{ij}=O(r^{-\tau}),\quad \partial_k a_{ij}=O(r^{-\tau-1}),\quad \partial_l\partial_k a_{ij}=O(r^{-\tau-2}). \tag{1.3}$$

Furthermore, the second fundamental forms of M satisfy

$$h_{ij}=O(r^{-\tau-1}),\quad \partial_k h_{ij}=O(r^{-\tau-2}). \tag{1.4}$$

We will often identify the end $M_l\subset M$ with the corresponding set $M_l\subset R^n$.

For spacelike asymptotically flat hypersurface M, we can define the total energy and the total momentum. These quantities include contributions from both the matter and the gravitational field itself. They are defined in each asymptotic end M_l as limits over the sphere $S_{R,l}$ of radius R in $M_l\subset R^n$.

Definition 1.2: Total energy of end M_l is defined as

$$E_l=\lim_{R\to\infty} C_n^{-1}\int_{S_{R,l}}(\partial_j g_{ij}-\partial_i g_{jj})d\Omega^i, \tag{1.5}$$

total linear momentum of end M_l is defined as

[a)]Electronic mail: xzhang@math08.math.ac.cn

$$p_{lk}=\lim_{R\to\infty} C_n^{-1}\int_{S_{R,l}} 2(h_{ik}-\delta_{ik}h_{jj})d\Omega^i, \tag{1.6}$$

where $C_n=4(n-1)\omega_{n-1}$. When the asymptotic order $\tau>(n-2)/2$, these quantities are finite, also Bartnik proved that E_l is independent on the choice of asymptotic coordinates.[1]

Definition 1.3: M is satisfied the dominant energy condition if for each point $p\in M$ and for each timelike vector e_0 at p, $T(e_0,e_0)\geqslant 0$ and $T(e_0,\ )$ is a nonspacelike covector, see Ref. 2. This has the following consequences: if $\{e_\alpha|\alpha=0,1,...,n\}$ is an adapted orthonormal frame field at p $\in M$ with e_0 normal to M and $e_1,\ldots,e_n$ tangent to M, then

$$T^{00}\geqslant|T^{\alpha\beta}|, \quad T^{00}\geqslant(T_i^0T^{0i})^{1/2}. \tag{1.7}$$

(Here, and henceforth, repeated indices are summed with Latin indices running from 1 to n and Greek indices running from 0 to n.)

Now we give the statement of n-dimensional (generalized) positive mass conjecture.

Positive Mass Conjecture I: Let N be an $(n+1)$-dimensional Lorentzian manifold with Lorentzian metric $\tilde{g}$ of signature $(-1,1,...,1)$, $M\subset N$ be an n-dimensional spacelike asymptotically flat hypersurface of order $\tau>(n-2)/2$. If the dominant energy condition holds on M, then, on each end M_l,

$$E_l\geqslant|P_l|\equiv\left(\sum_{k=1}^{n} p_{lk}^2\right)^{1/2}.$$

If $E_{l_0}=0$ for some l_0, then M has only one end and N is flat along M.

When the spacelike hypersurface is maximal, i.e., $H=0$, the (generalized) positive mass conjecture states that

Positive Mass Conjecture II: Let M be an n-dimensional asymptotically flat manifold of order $\tau>(n-2)/2$. If the scalar curvature $R\geqslant 0$, then, on each end M_l, $E_l\geqslant 0$. If $E_{l_0}=0$ for some l_0, then M is isomorphic to R^n.

In general relativity, space–time is a four-dimensional Lorentzian manifold, i.e., $n=3$, the positive mass conjecture was originally conjectured more than 30 years ago by physicists.[3] Subsequently, a great many people worked on this problem and proved various special cases. In 1978, Schoen and Yau used a geometrical method to prove this conjecture for the case of maximal spacelike hypersurface (Conjecture II).[4] Using an auxiliary equation introduced by Jang,[5] they generalized their proof to the nonmaximal spacelike hypersurface case (Conjecture I),[6] and finally solving this long-standing problem. They have also applied their method to prove the positive action conjecture.[7] Two years later, Witten presented a simple new proof of the Conjecture I by using spinors although several points of his argument come from physical intuition and require justification.[8] Soon later, Parker and Taubes gave a complete, rigorous and self-contained proof of the Conjecture I, based on Witten's formulation.[9] The higher dimensional positive mass conjectures have been studied only in the maximal hypersurface case (Conjecture II); Schoen gave a detail n-dimensional proof of his work with Yau which proved the Conjecture II through the use of volume minimizing hypersurface.[10] The proof they gave works for $n\leqslant 7$ in which dimensions they have complete regularity of volume minimizing hypersurfaces. Bartnik proved Conjecture II for n-dimensional spin manifolds following Witten's approach.[1]

However, it has not appeared in the literature for the proof of higher dimensional Positive Mass Conjecture I (nonmaximal hypersurface case). Although, in principle, Witten's proof carries over to all dimensions, by assuming a spin structure, there still exists one technical difficulty: How to prove positivity of induced metric $\langle,\rangle=(e^0.,)$ (see Sec. II for details) on spinor bundle along a spacelike hypersurface? When $n=3$, Parker and Taubes prove it in terms of representation of spin group $SL(2,C)\cong \mathrm{Spin}^0(3,1)$. We shall prove it for $n=4$ in terms of representation of spin group $HU(1,1)\cong\mathrm{Spin}^0(4,1)$ where H denotes the field of quaternious. We expect this fact is true for all spin group $\mathrm{Spin}^0(n,1)$ by direct calculations in the Clifford algebra which do not need to invoke special representations, and we shall address it elsewhere.

The positive mass conjecture for the five-dimensional Lorentzian manifold is interested in the context of the Kaluza–Klein theory. In this paper, we shall prove

Theorem 1.1: *Let N be a five-dimensional Lorentzian manifold with Lorentzian metric $\tilde{g}$ of signature $(-1,1,1,1,1)$, $M\subset N$ be a spin spacelike asymptotically flat hypersurface of the order $\tau>1$. If the dominant energy condition holds on M, then, on each end M_l,*

$$E_l \geqslant |P_l| \equiv \left(\sum_{k=1}^{4} p_{lk}^2 \right)^{1/2}.$$

If $E_{l_0}=0$ for some l_0, then M has only one end and N is flat along M.

We refer to Ref. 11 for the positive mass theorem on a class of ''modified asymptotically flat manifolds'' which satisfy ''modified energy condition,'' and to Ref. 12 for nonspin spacelike hypersurface in five-dimensional Lorentzian manifolds.

II. HU(1,1) REPRESENTATION AND SPINORS

In this section, we assume *N* is a five-dimensional Lorentzian manifold with Lorentzian metric of signature $(-1,1,1,1,1)$, and *M* is a spin spacelike hypersurface in *N*. We shall show there is a globally defined HU(1,1) bundle *S* along *M*. With the help of HU(1,1) Hermitian structure on *S*, we construct a subbundle of End(S) which is exactly the cotangent bundle of *N*. This gives a natural definition of Clifford multiplication, and induce a positively define Spin(4) invariant Hermitian metric on *S*. We describe them first at the level of linear algebra and then globally on the manifold *M*.

Denote *H* as the field of quaternious. The hyperunitary group HU(1,1) is defined to be the subgroup of GL$(2,H)$ that fixes the standard *H*-Hermitian symmetric form

$$(p,q)=\overline{p_1}\cdot q_1-\overline{p_2}\cdot q_2,$$

where $p=(p_1,p_2)^t$, $q=(q_1,q_2)^t\in H^2$. The group $\mathrm{HU}(1,1)=\mathrm{Spin}^0(4,1)$ is the double covering group of connected Lorentz group $SO(4,1)$, see Ref. 13, p. 272. Let *V* be the fundamental representation of HU(1,1) on H^2. For any $X\in\mathrm{End}(V)$, denote X^* the adjoint of *X* under HU(1,1) Hermitian structure. We note that any $A\in\mathrm{HU}(1,1)$ if and only if $AA^*=I$, $A^*A=I$. On End(V), we define the operator

$$RT(X)=\mathrm{Re}\{\mathrm{Trace}(X)\}. \tag{2.1}$$

Proposition 2.1: RT is well-defined, i.e., RT is independent on the choice of basis. Moreover, for any $X,Y\in\mathrm{End}(V)$,

$$RT(X^*Y)=RT(YX^*)=RT(XY^*).$$

Proof: Choosing a basis, we can write

$$X=\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \quad Y=\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix}.$$

Then,

$$X^*=\begin{pmatrix} \overline{x_{11}} & -\overline{x_{21}} \\ -\overline{x_{12}} & \overline{x_{22}} \end{pmatrix},$$

and

$$Y^*=\begin{pmatrix} \overline{y_{11}} & -\overline{y_{21}} \\ -\overline{y_{12}} & \overline{y_{22}} \end{pmatrix},$$

where x_{11}, x_{12}, x_{21}, x_{22}, y_{11}, y_{12}, y_{21}, y_{22} are quaternious numbers. We have $RT(X)=\mathrm{Re}(x_{11}+x_{22})$. Changing basis, *X* changes to $A^{-1}XA$ for some $A\in\mathrm{HU}(1,1)$. So, for proving the first part of the proposition, we need only show that $RT(A^{-1}XA)=RT(X)$.

Let $A=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $A\in\mathrm{HU}(1,1)$ gives that

$$|a|^2-|b|^2=1, \quad |d|^2-|c|^2=1, \quad a\bar{c}-b\bar{d}=0.$$

We note that $\overline{xy}=\bar{y}\bar{x}$, $\mathrm{Re}(x)=\mathrm{Re}(\bar{x})$, $\mathrm{Re}(xy)=\mathrm{Re}(\overline{xy})$ for any quaternious numbers x, y. Therefore,

$$\begin{aligned}RT(A^{-1}XA)&=RT(A^*XA)\\&=\mathrm{Re}(\bar{a}x_{11}a-\bar{c}x_{21}a+\bar{a}x_{12}c-\bar{c}x_{22}c-\bar{b}x_{11}b+\bar{d}x_{21}b-\bar{b}x_{12}d+\bar{d}x_{22}d)\\&=\mathrm{Re}(|a|^2\overline{x_{11}}-|c|^2\overline{x_{22}}-|b|^2\overline{x_{11}}+|d|^2\overline{x_{22}}-c\overline{a x_{21}}+d\bar{b}\overline{x_{21}}+a\overline{c x_{12}}-b\bar{d}\overline{x_{12}})\\&=\mathrm{Re}(\overline{x_{11}}+\overline{x_{22}})=RT(X).\end{aligned}$$

For the proof of the second part, since

$$RT(X^*Y)=\mathrm{Re}(\overline{x_{11}}y_{11}-\overline{x_{12}}y_{12}-\overline{x_{21}}y_{21}+\overline{x_{22}}y_{22}),$$

$$RT(YX^*)=\mathrm{Re}(y_{11}\overline{x_{11}}-y_{12}\overline{x_{12}}-y_{21}\overline{x_{21}}+y_{22}\overline{y_{22}}),$$

$$RT(XY^*)=\mathrm{Re}(x_{11}\overline{y_{11}}-x_{12}\overline{y_{12}}-x_{21}\overline{y_{21}}+x_{22}\overline{y_{22}}).$$

Hence it follows. □

Corollary 2.1: On End(V), inner product

$$\langle X,Y\rangle=-\tfrac{1}{2}RT(X^*Y) \tag{2.2}$$

is independent on the choice of basis.

Set

$$\aleph=\{X\in \mathrm{End}(V):X=X^*\}. \tag{2.3}$$

It is independent on the choice of basis since $(A^*XA)^*=A^*X^*A=A^*XA$ for any $X\in\aleph$, $A\in \mathrm{HU}(1,1)$.

Proposition 2.2: On $\aleph$, $\mathrm{Trace}(X)$ *is independent on the choice of basis.*

Proof: Choosing a basis, let X be given by a matrix as above. Then, $X=X^*$ gives that $x_{11}=\overline{x_{11}}$, $x_{22}=\overline{x_{22}}$, $x_{12}=-\overline{x_{21}}$. Hence x_{11}, x_{22} are real numbers, and $RT(X)=x_{11}+x_{22}=\mathrm{Trace}(X)$. □

Proposition 2.3: Set

$$\aleph_0=\{X\in\aleph, \quad \mathrm{Trace}(X)=0\}, \tag{2.4}$$

then $(\aleph_0,\|\ \|)=(R^{4,1},\tilde{g})$, *where* $\tilde{g}$ *is the standard Lorentzian metric on* $R^{4,1}$.

Proof: Choosing a basis, and take $x_{11}=x_0$, $x_{22}=-x_0$, $x_{12}=x_1+x_2i+x_3j+x_4k\equiv x$, x_0, x_1, x_2, x_3, x_4 are all real numbers. It gives $X=X^*=\begin{pmatrix} x_0 & x\\ -\bar{x} & -x_0\end{pmatrix}$ for any $X\in\aleph_0$. Obviously, $\|X\|^2=\langle X,X\rangle=-x_0^2+x_1^2+x_2^2+x_3^2+x_4^2=\tilde{g}(X,X)$. Hence we can identify any $X=(x_0,x_1,x_2,x_3,x_4)\in R^{4,1}$ as an element in $\aleph_0$ with norm $\|X\|$, under a basis, which is given by the matrix

$$X=\begin{pmatrix} x_0 & x\\ -\bar{x} & -x_0\end{pmatrix}, \tag{2.5}$$

where $x=x_1+x_2i+x_3j+x_4k$. Moreover, this identification does not depend on the choice of basis. □

The spinors for $\mathrm{Spin}^0(4,1)=\mathrm{HU}(1,1)$ structure is defined by V. This space has a $\mathrm{HU}(1,1)$ invariant Hermitian inner product defined by

$$(\phi,\psi)=\overline{\xi_1}\cdot\eta_1-\overline{\xi_2}\cdot\eta_2 \tag{2.6}$$

for $\phi=(\xi_1,\xi_2)^t\in V,\ \psi=(\eta_1,\eta_2)^t\in V$. This inner product is not positive definite.

We define the Clifford multiplication map ‘‘.’’

$$.:R^{4,1}\otimes V\to V$$

$$X.\phi=X\phi,$$

where X is the correspondent element in $\aleph_0$ for point in $R^{4,1}$, choosing a basis, given by the matrix (2.5). Obviously, $X.Y.+Y.X.=-2\tilde{g}(X,Y)\cdot Id$. So by the universal property of Clifford algebra, the map ‘‘·’’ can be extended to a quaternious representation of Clifford algebra Cl(4,1), hence to the group HU(1,1).

Choosing an orthonormal basis $\{e_\alpha\}$ on $R^{4,1}$ with e_0 timelike, let $\{e^\alpha\}$ be its dual basis. The choice of a timelike covector e^0 yields a diagram

$$\begin{array}{ccc}\mathrm{Sp}(1)\times\mathrm{Sp}(1) & \overset{\hat\alpha}{\to} & \mathrm{HU}(1,1)\\ \rho_1\downarrow & & \downarrow\rho_2\\ \mathrm{SO}(4) & \overset{\alpha}{\to} & \mathrm{SO}(4,1),\end{array}$$

where the maps are defined as follows: write $x=x_1+x_2i+x_3j+x_4k\in H\cong R^4$, $X=\begin{pmatrix} x_0 & x\\ \bar{x} & -x_0\end{pmatrix}\in\aleph_0$ $\cong R^{4,1}$, for $p,q\in\mathrm{Sp}(1)$, $A\in\mathrm{HU}(1,1)$, $a\in\mathrm{SO}(4)$,

$$\rho_1((p,q))x=px\overline{q},\qquad \rho_2(A)X=AXA^*,$$

$$\hat\alpha((p,q))=\begin{pmatrix} p & 0\\ 0 & q\end{pmatrix},\qquad \alpha(a)X=\begin{pmatrix} x_0 & ax\\ -\overline{ax} & -x_0\end{pmatrix}.$$

The double-covering map ρ_1 is well-known. Now it is easy to check ρ_1 is also a double-covering map. Moreover, $d\rho_2:\mathrm{hu}(1,1)\cong\mathrm{so}(4,1)$ given by $d\rho_2(e^\alpha e^\beta)=2e^\alpha\wedge e^\beta$. Finally,

$$\rho_2\circ\hat\alpha((p,q))X=\begin{pmatrix} x_0 & px\overline{q}\\ -\overline{qxp} & -x_0\end{pmatrix}=\alpha\circ\rho_1((p,q))X.$$

The above diagram allows us to regard V as Spin(4)=Sp(1)×Sp(1) representation and gives V an another Hermitian structure. The Clifford multiplication $e^0=\begin{pmatrix}1 & \\ & -1\end{pmatrix}:V\to V$ gives an isomorphism $V\cong V$, the new Hermitian structure on V is given by this isomorphism together with the isomorphism $V\cong\overline{V}^*$ given by the HU(1,1) structure. In another word, there is another Hermitian inner product on V given by

$$\langle\phi,\psi\rangle=(e^0.\phi,\psi)=\overline{\xi_1}\cdot\eta_1+\overline{\xi_2}\cdot\eta_2 \tag{2.7}$$

for $\phi=(\xi_1,\xi_2)^t\in V,\ \psi=(\eta_1,\eta_2)^t\in V$. Hence this new inner product is positive definite and Sp(1)×Sp(1) invariant. Now it is easy to derive the following two propositions

Proposition 2.4: For any $X\in R^{4,1}$, spinors $\phi,\psi\in V$, we have $(X.\phi,\psi)=(\phi,X.\psi)$.

Proposition 2.5: For any $x=(x_1,x_2,x_3,x_4)\in R^4$ regarded as an embedding in $X=(0,x)\in R^{4,1}$, we have $\langle x.\phi,\psi\rangle=-\langle\phi,x.\psi\rangle$, and $\langle e^0.\phi,\psi\rangle=\langle\phi,e^0.\psi\rangle$.

If M is a spin spacelike hypersurface in N, the above algebra facts carry over to vector bundles once a spin structure is chosen. Let $F(N)$ denote the $SO(4,1)$ frame bundle of the cotangent bundle of N and let $i:M\to N$ be the inclusion. The required spin structure is a lift of the bundle $i^*F(N)$ to a HU(1,1) bundle over M. But

$$i^*F(N)=F(M)\times_\alpha SO(4,1),$$

so we need only lift the $SO(4)$ frame bundle of M to an Sp(1)×Sp(1) bundle $\widetilde{F(M)}$. The obstruction to such an $\widetilde{F(M)}$ is the Stiefel–Whitney class $\omega_2(M)$.

Since M is spin, $\omega_2(M)=0$, $\widetilde{F(M)}$ exists. The number of such lifts $\widetilde{F(M)}$ is then classified by $H^1(M,Z_2)$. Choosing one, we obtain the desired HU(1,1) bundle

$$i^*\widetilde{F(N)}=\widetilde{F(M)}\times_{\hat{\alpha}}\mathrm{HU}(1,1)$$

over M and the associated spin vector bundle

$$i^*\widetilde{F(N)}\times_\rho V=\widetilde{F(M)}\times_{\bar{\rho}}V,$$

where ρ is the representation V of HU(1,1), and $\bar{\rho}$ is its restriction to Sp(1)×Sp(1). This vector bundle—denoted S—carries the inner products (,) and ⟨,⟩. Sections of S are called hypersurface spinors along M. Proposition 2.3 implies $T^*N\cong\aleph_0(S)$, so the Clifford multiplication is globally-defined on M.

The metric connection $\tilde{\nabla}$ on $F(N)$ determines connections on $i^*\widetilde{F(N)}$ and its associated bundle. The resulting connection (also denoted) $\tilde{\nabla}$ on S is compatible with the metric (,) but not compatible with the metric ⟨,⟩. Let ∇ be the Riemannian connection on $F(M)$. It also induces a connection ∇ on $S=\widetilde{F(M)}\times_{\bar{\rho}}V$. We shall show that ∇ is compatible with ⟨,⟩.

Fix a point $p\in M$ and an orthonormal basis $\{e_\alpha\}$ of T_pN with e_0 normal and e_1, e_2, e_3, e_4 tangent to M. Extend e_1, e_2, e_3, e_4 to an orthonormal frame in a neighborhood of p in M such that $(\nabla_i e_j)_p=0$. Extend this to a local orthonormal $\{e_\alpha\}$ for N with $(\tilde{\nabla}_0 e_j)_p=0$. Let $\{e^\alpha\}$ be the dual coframe. Then

$$(\tilde{\nabla}_i e^j)_p=-h_{ij}e^0,\quad (\tilde{\nabla}_i e^0)_p=-h_{ij}e^j,$$

where $h_{ij}=\langle\tilde{\nabla}_i e_0,e_j\rangle$, $1\leqslant i,\ j\leqslant 4$, are the components of the second fundamental form at p. We have the following relations about two connections on S:

$$\tilde{\nabla}_i=\nabla_i+\tfrac{1}{2}h_{ji}e^0.e^j. \tag{2.8}$$

Proposition 2.6: The induced connection ∇ on S is compatible with the metric ⟨,⟩ and (,).

Proof: In the above frame we have,

$$\begin{aligned}d((\phi,\psi)*e_i)&=((\tilde{\nabla}_i\phi,\psi)+(\phi,\tilde{\nabla}_i\psi))*1\\&=((\nabla_i\phi,\psi)+(\phi,\nabla_i\psi)+(\tfrac{1}{2}h_{ij}e^0.e^j.\phi,\psi)+(\phi,\tfrac{1}{2}h_{ij}e^0.e^j.\psi))*1\\&=((\nabla_i\phi,\psi)+(\phi,\nabla_i\psi))*1.\end{aligned}$$

$$\begin{aligned}d(\langle\phi,\psi\rangle*e_i)&=d((e^0.\phi,\psi)*e_i)\\&=((-h_{ij}e^j.\phi,\psi)+(e^0.\tilde{\nabla}_i\phi,\psi)+(e^0.\phi,\tilde{\nabla}_i\psi))*1\\&=((-h_{ij}e^j.\phi,\psi)+(e^0.\nabla_i\phi,\psi)+(e^0.\phi,\nabla_i\psi)\\&\quad+(e^0.\tfrac{1}{2}h_{ij}e^0.e^j.\phi,\psi)+(e^0.\phi,\tfrac{1}{2}h_{ij}e^0.e^j.\psi))*1\\&=(\langle\nabla_i\phi,\psi\rangle+\langle\phi,\nabla_i\psi\rangle)*1.\end{aligned}$$

□

Finally, let $\Theta : R^4 - K_l \to M_l$ be the diffeomorphism which defines M_l. The pullback bundle $\Theta_l^* \widetilde{F(M)}$ is trivial over the end M_l and the bundle $\Theta_l^* S$ extends trivially over all of R^4. The Θ_l^{-1}—pullbacks of the constant sections of the bundle $R^4 \times S$ over R^4 then provide a set of ''constant spinors'' over the M_l.

III. THE WITTEN–DIRAC OPERATOR

In a local orthonormal coframe $\{e^i\}$ of M, Dirac operator D and the Witten–Dirac operator $\tilde{D}$ are defined by

$$D = e^i . \nabla_i , \quad \tilde{D} = e^i . \tilde{\nabla}_i ,$$

respectively, where ''.'' denotes Clifford multiplication. Obviously, D is self-adjoint under the metric $\langle , \rangle$. And we have the following Lichnerowicz formula:

$$D^* D = D^2 = \nabla^* \nabla + \frac{R}{4}, \tag{3.1}$$

where R is the scalar curvature of M.

Lemma 3.1: For any $\phi \in \Gamma(S)$, *we have*

$$\tilde{D} \phi = D \phi + \frac{H}{2} e^0 . \phi, \tag{3.2}$$

where $H = \Sigma h_{ii}$ *is the mean curvature of M.*

Proof: Since $h_{ij} = h_{ji}$, and $e^i . e^j . = - e^j . e^i .$, for $i \neq j$, then (2.8) gives

$$\tilde{D} \phi = e^i . \tilde{\nabla}_i \phi = e^i . \nabla_i \phi + \frac{1}{2} h_{ij} e^i . e^0 . e^j . \phi = D \phi + \frac{H}{2} e^0 . \phi.$$

In terms of (2.8), (3.2), we can prove

Lemma 3.2:

$$d(\langle e^i . \phi, \psi \rangle * e^i) = (\langle D \phi, \psi \rangle - \langle \phi, D \psi \rangle) * 1 = (\langle \tilde{D} \phi, \psi \rangle - \langle \phi, \tilde{D} \psi \rangle) * 1,$$

$$d(\langle \phi, \tilde{\nabla}_i \psi \rangle * e^i) = (\langle \tilde{\nabla}_i \phi, \tilde{\nabla}_i \psi \rangle - \langle \phi, (-\tilde{\nabla}_i + h_{ij} e^0 . e^j .) \tilde{\nabla}_i \psi \rangle) * 1.$$

Hence,

$$D^* = D, \quad \tilde{D}^* = \tilde{D}, \quad \tilde{\nabla}_i^* = -\tilde{\nabla}_i + h_{ij} e^0 . e^j .$$

Now we derive two Weitzenböck formulas, the second was given by Witten.[8,9]

Theorem 3.1: *For any* $\phi \in \Gamma(S)$,

$$\tilde{D}^2 \phi = \nabla^* \nabla \phi + \tfrac{1}{4}(R + H^2) \phi - \tfrac{1}{2} \nabla_i H e^0 . e^i . \phi \tag{3.3}$$

$$= \tilde{\nabla}^* \tilde{\nabla} \phi + \tfrac{1}{2}(T_{00} + T_{0i} e^0 . e^i .) \phi. \tag{3.4}$$

Proof: Since

$$\nabla_i(e^0.\phi)=(\tilde{\nabla}_i-\tfrac{1}{2}h_{ij}e^0.e^j.)(e^0.\phi)$$
$$=-h_{ij}e^j.\phi+e^0.\tilde{\nabla}_i\phi+\tfrac{1}{2}h_{ij}e^j.\phi$$
$$=e^0.(\tilde{\nabla}_i-\tfrac{1}{2}h_{ij}e^0.e^j.)\phi=e^0.\nabla_i\phi,$$

then Lemma 3.1 and the Lichnerowicz formula (3.1) show

$$\tilde{D}^2\phi=\left(D+\frac{H}{2}e^0.\right)\left(D\phi+\frac{H}{2}e^0.\phi\right)$$
$$=D^2\phi+\frac{H^2}{4}e^0.e^0.\phi+\frac{1}{2}e^i.\nabla_iHe^0.\phi$$
$$=\nabla^*\nabla\phi+\frac{R}{4}\phi+\frac{H^2}{4}\phi-\frac{1}{2}\nabla_iHe^0.e^i.\phi.$$

But

$$\tilde{\nabla}^*\tilde{\nabla}\phi=(-\tilde{\nabla}_i+h_{ij}e^0.e^j.)\tilde{\nabla}_i\phi$$
$$=(-\nabla_i+\tfrac{1}{2}h_{ij}e^0.e^j.)(\nabla_i\phi+\tfrac{1}{2}h_{ik}e^0.e^k.\phi)$$
$$=\nabla^*\nabla\phi-\tfrac{1}{4}h_{ij}h_{ik}e^j.e^k.\phi-\tfrac{1}{2}\nabla_i(h_{ij}e^0.e^j.\phi)+\tfrac{1}{2}h_{ij}e^0.e^j.\nabla_i\phi$$
$$=\nabla^*\nabla\phi+\frac{1}{4}\sum_{i,j}h_{ij}^2\phi-\frac{1}{2}\nabla_ih_{ij}e^0.e^j.\phi.$$

Therefore the second formula can be derived in terms of the following (Gauss and Codazzi) equations

$$T_{00}=\frac{1}{2}\left(R-\sum_{ij}h_{ij}^2+H^2\right),\quad T_{0i}=\sum_j\nabla_jh_{ij}-\nabla_iH.$$

□

The integral forms of Weitzenböck formula are

$$\int_M|\tilde{\nabla}\phi|^2+\langle\phi,\tilde{R}\phi\rangle-|\tilde{D}\phi|^2=\frac{1}{2}\int_{\partial M}\langle\phi,[e^i,e^j].\tilde{\nabla}_j\phi\rangle*e^i,\tag{3.5}$$

where $\tilde{R}=\tfrac{1}{2}(T_{00}+T_{0i}e^0.e_i.)$, and $[e^i,e^j]=e^i.e^j.-e^j.e^i.$

IV. BOUNDARY VALUE PROBLEM

In this section, we assume M is a spin spacelike asymptotically flat hypersurface of order $\tau>1$ in the five-dimensional Lorentzian manifold N. We shall study the infinity boundary value problem for the Witten–Dirac equation. We shall simplify the original arguments in Ref. 9.

First, we recall a lemma in Ref. 9, which can be easily extended to our case.

Lemma 4.1: Suppose M is asymptotically flat and ϕ, $\{\phi_i\}$ are C^1 hypersurface spinors along M which satisfy

$$\tilde{\nabla}\phi=0,\quad\tilde{\nabla}\phi_i=0\quad\textit{for each } i.$$

(i) *If $\lim_{x\to\infty}\phi(x)=0$, where the limit is taken along M in one asymptotic end, then $\phi=0$.*

(ii) *If $\{\phi_i\}$ are linearly independent in some end, then they are linearly independent everywhere on M.*

Proof: We sketch the proof in spirit of Ref. 9. (i) Since $\tilde{\nabla}\phi=0$, i.e., $\nabla_i\phi=-\frac{1}{2}h_{ij}e^0.e^j.\phi$, and $h=O(r^{-\tau-1})$, this gives $|d\ln|\phi||\leqslant Cr^{-\tau-1}$ on the complement of the zero set of ϕ. Integrating this along a path from $x_0\in M$ gives

$$|\phi(x)|\geqslant|\phi(x_0)|e^{C(|x_0|^{-\tau}-|x|^{-\tau})}.$$

Taking x to be the first zero of ϕ along the path of integration, or taking the limit as $|x|\to\infty$ if no such zero exists, shows that $\phi(x_0)=0$. Hence $\phi=0$ on the ends. On the compact set K, since h is bounded, we have

$$|\phi(x)|\geqslant|\phi(x_0)|e^{C(|x_0|-|x|)}.$$

Hence $\phi=0$ on K by taking the path to the ends. (ii) It follows from the first part. □

Now we define the weighted C^k space $C^k_\tau(S)$ as the set of C^k hypersurface spinors ϕ for which the norm

$$\|\phi\|_{C^k_\tau}=\sum_{i=0}^{k}\sup(r^{-\tau+i}|\nabla^i\phi|)$$

is finite. The weighted Hölder space $C^{k,\alpha}_\tau(S)$ is defined for $0<\alpha<1$ as the set of $\phi\in C^k_\tau(S)$ for which the norm

$$\|\phi\|_{C^{k,\alpha}_\tau}=\|\phi\|_{C^k_\tau}+\sup_{x,y}\left\{(\min\{r(x),r(y)\})^{-\tau+k+\alpha}\frac{|\nabla^k\phi(x)-\nabla^k\phi(y)|}{r(x,y)^\alpha}\right\}$$

is finite. Here, the supremum is over all $x\neq y$ such that y is contained in a normal coordinate neighborhood of x, and $\nabla^k\phi(y)$ is the spinor at x obtained by parallel transport along the radial geodesic from x to y. Note that the definitions of the weighted spaces depend on the "distance function" r, and thereby on a choice asymptotic coordinates. However, it is easy to see that r is uniformly equivalent to the geodesic distance from an arbitrary fixed point in M as $r\to\infty$, so all choices of r define equivalent norms.

If M is asymptotically flat of order $\tau>1$ with asymptotic coordinates $\{dx^i\}$ on the end. Orthonormalizing $\{dx^i\}$ yields an orthonormal coframe

$$e^i=dx^i+\tfrac{1}{2}a_{ik}dx^k+O(r^{-\tau-1}). \tag{4.1}$$

The connection coefficients of $\{dx^i\}$ are

$$\Gamma_{kjl}=\tfrac{1}{2}(\partial_j g_{kl}+\partial_l g_{kj}-\partial_k g_{jl})=O(r^{-\tau-1}),$$

thus we obtain

$$\nabla_j=\partial_j-\tfrac{1}{4}\Gamma_{kjl}dx^k.dx^l.+O(r^{-2\tau-1}). \tag{4.2}$$

Denote e^0 as dx^0, then

$$\tilde{D}=dx^j.\partial_j-\frac{1}{4}\Gamma_{kjl}dx^j.dx^k.dx^l.+\frac{H}{2}dx^0.+O(r^{-2\tau-1}). \tag{4.3}$$

Therefore $\tilde{D}$ gives the maps for the following weighted Hölder spaces:

$$C^{2,\alpha}_{-\tau}(S)\xrightarrow{\tilde{D}}C^{1,\alpha}_{-\tau-1}(S)\xrightarrow{\tilde{D}}C^{0,\alpha}_{-\tau-2}(S).$$

For constant spinor ϕ_0, $\partial_j\phi_0=0$, we have

$$\tilde{D}\phi_0=e^j.\nabla_j\phi_0+\frac{H}{2}e^0.\phi_0=-\frac{1}{4}\Gamma_{kjl}dx^j.dx^k.dx^l.\phi_0+\frac{H}{2}dx^0.\phi_0+O(r^{-2\tau-1}).$$

Hence $\tilde{D}\phi_0\in C^{1,\alpha}_{-\tau-1}(S)$, and $\tilde{D}^2\phi_0\in C^{0,\alpha}_{-\tau-2}(S)$.

Lemma 4.2: If M is asymptotically flat of order $\tau>1$ and the dominant energy condition holds on M, then the map

$$\tilde{D}^2:C^{2,\alpha}_{-\tau}(S)\rightarrow C^{0,\alpha}_{-\tau-2}(S) \tag{4.4}$$

is an isomorphism.

Proof: Note that $(\frac{1}{4}(R+H^2)-\frac{1}{2}\nabla_iHe^0.e^i.)\in C^{0,\alpha}_{-\tau-2}(S)$, thus we only need to show that the kernel of $\tilde{D}^2$ on $C^{2,\alpha}_{-\tau}(S)$ is trivial since, by (3.3) and Theorem 9.2(d),[14] the injectivity implies $\tilde{D}^2$ is an isomorphism. Let $\phi\in C^{2,\alpha}_{-\tau}(S)$ satisfy $\tilde{D}^2\phi=\tilde{\nabla}^*\tilde{\nabla}\phi+\tilde{R}\phi=0$. Then

$$\int_M|\tilde{\nabla}\phi|^2+\langle\tilde{R}\phi,\phi\rangle=\int_{\partial M}\langle\phi,\tilde{\nabla}_i\phi\rangle*e^i.$$

But $\langle\phi,\tilde{\nabla}_i\phi\rangle=\langle\phi,(\nabla_i\phi+\frac{1}{2}h_{ij}e^0.e^j.\phi)\rangle=O(r^{-2\tau-1})$, and $\mathrm{Vol}(\partial M)=O(r^{-3})$ by (1.2) and (1.3). Hence the right-hand side of the above integral vanishes. Therefore $\tilde{\nabla}\phi=0$ on *M*. Hence $\phi=0$ by Lemma 4.1 (i), and the proof of the lemma is complete. □

Theorem 4.1: *If M is asymptotically flat of order $\tau>1$ and the dominant energy condition holds on M, then for any constant spinor ϕ_0 on ends, the following boundary value problem has a unique solution $\phi\in C^{2,\alpha}(S)$:*

$$\begin{cases}\tilde{D}\phi=0\\ \lim_{r\rightarrow\infty}\phi=\phi_0.\end{cases} \tag{4.5}$$

Proof: Since $\tilde{D}^2\phi_0\in C^{0,\alpha}_{-\tau-2}(S)$. By Lemma 4.2, there is unique $\phi_1\in C^{2,\alpha}_{-\tau}(S)$ such that $\tilde{D}^2\phi_1=-\tilde{D}^2\phi_0$. Then $\phi=\phi_1+\phi_0$ satisfies $\tilde{D}^2\phi=0$. Let $\psi=\tilde{D}\phi\in C^{1,\alpha}_{-\tau-1}(S)$, then

$$\int_M|\tilde{\nabla}\psi|^2+\langle\tilde{R}\psi,\psi\rangle=\int_{\partial M}\langle\psi,\tilde{\nabla}_i\psi\rangle*e^i=\int_{\partial M}O(r^{-2\tau-3})=0.$$

Therefore $\tilde{\nabla}\psi=0$ on *M*. Hence $\psi=0$ by Lemma 4.1 (i). Thus ϕ is the unique solution of (4.5).□

V. POSITIVE MASS CONJECTURE

In this section, we shall prove the Positive Mass Conjecture.

Theorem 5.1: *Let N be a five-dimensional Lorentzian manifold with Lorentzian metric $\tilde{g}$ of signature (−1,1,1,1,1), $M\subset N$ be a spin spacelike asymptotically flat hypersurface of order $\tau>1$. If the dominant energy condition holds on M, then, on each end M_l,*

$$E_l\geqslant|P_l|\equiv\left(\sum_{k=1}^4 p_{lk}^2\right)^{1/2}.$$

If $E_{l_0}=0$ for some l_0, then M has only one end and N is flat along M.

Proof: For end M_l, let constant spinor $\phi_0 \neq 0$ on M_l, and $\phi_0 = 0$ on other ends. Denote $\phi = \phi_1 + \phi_0$, where $\phi_1 \in C^{2,\alpha}_{-\tau}(S)$, as the corresponding solution of (4.5) for this ϕ_0. Then (3.5) gives, under the coframe $\{e^i\}$ choosen in (4.1),

$$2\int_M |\tilde{\nabla}\phi|^2 + \langle \phi, \tilde{R}\phi\rangle = \int_{\partial M_\infty} \langle \phi, [e^i, e^j].\tilde{\nabla}_j \phi\rangle * e^i = \int_{\partial M_\infty} \langle \phi_0, [e^i, e^j].\tilde{\nabla}_j \phi_0\rangle * e^i + \sum ,$$

where

$$\begin{aligned}
\sum &= \int_{\partial M_\infty} (\langle \phi_1, [e^i, e^j].\tilde{\nabla}_j \phi_0\rangle + \langle \phi_0, [e^i, e^j].\tilde{\nabla}_j \phi_1\rangle + \langle \phi_1, [e^i, e^j].\tilde{\nabla}_j \phi_1\rangle) * e^i \\
&= \int_{\partial M_\infty} \langle \phi_1, [e^i, e^j].\nabla_j \phi_0\rangle * e^i + \int_{\partial M_\infty} \left\langle \phi_1, \frac{1}{2} h_{jk} [e^i, e^j].e^0.e^k.\phi_0 \right\rangle * e^i \\
&\quad + \int_{\partial M_\infty} \langle -\nabla_j \phi_0, [e^i, e^j].\phi_1\rangle * e^i + \int_{\partial M_\infty} d(\langle \phi_0, [e^i, e^j].\phi_1\rangle * (e^i \wedge e^j)) \\
&\quad + \int_{\partial M_\infty} \left\langle \phi_0, \frac{1}{2} h_{jk} [e^i, e^j].e^0.e^k.\phi_1 \right\rangle * e^i + \int_{\partial M_\infty} \langle \phi_1, [e^i, e^j].\tilde{\nabla}_j \phi_1\rangle * e^i \\
&= 2\,\mathrm{Re} \int_{\partial M_\infty} \langle \phi_1, [e^i, e^j].\nabla_j \phi_0\rangle * e^i + \int_{\partial M_\infty} \langle \phi_1, [e^i, e^j].\nabla_j \phi_1\rangle * e^i \\
&\quad + \mathrm{Re} \int_{\partial M_\infty} h_{jk} \langle \phi_1, [e^i, e^j].e^0.e^k.\phi_0\rangle * e^i + \frac{1}{2} \int_{\partial M_\infty} h_{jk} \langle \phi_1, [e^i, e^j].e^0.e^k.\phi_1\rangle * e^i.
\end{aligned}$$

Since $\phi_1 = O(r^{-\tau})$, $\nabla_j \phi_1 = O(r^{-\tau-1})$, $\nabla_j \phi_0 = O(r^{-\tau-1})$, and $h_{ij} = O(r^{-\tau-1})$, then $\Sigma = 0$. Now it is easy to see that $\langle \phi, e^i.e^j.\phi\rangle$ is pure imaginary for $i \neq j$. On the other hand, $|e^i - dx^i| = O(r^{-\tau})$, so we can replace the $\{e^i\}$ by $\{dx^i\}$ in the above integral without changing the value of the limit. And note $\Gamma_{kjl} = \Gamma_{klj}$, we obtain

$$\begin{aligned}
\int_M |\tilde{\nabla}\phi|^2 + \langle \phi, \tilde{R}\phi\rangle &= \frac{1}{2}\,\mathrm{Re} \int_{\partial M_\infty} (\langle \phi_0, [dx^i, dx^j].\tilde{\nabla}_j \phi_0\rangle + O(r^{-2\tau-1})) * dx^i \\
&= \mathrm{Re} \int_{\partial M_\infty} \left\langle \phi_0, -\frac{1}{4} \Gamma_{kjl} (\delta_{ij} + dx^i.dx^j.) dx^k.dx^l.\phi_0 \right\rangle * dx^i \\
&\quad + \mathrm{Re} \int_{\partial M_\infty} \left\langle \phi_0, \frac{1}{2} h_{jk} (\delta_{ij} + dx^i.dx^j.) dx^0.dx^k.\phi_0 \right\rangle * dx^i \\
&= \frac{1}{4}\,\mathrm{Re} \int_{\partial M_\infty} \langle \phi_0, (\Gamma_{jij} + \Gamma_{kjl} dx^i (2\delta_{jk} + dx^k.dx^j.) dx^l.)\phi_0\rangle * dx^i \\
&\quad + \frac{1}{2}\,\mathrm{Re} \int_{\partial M_\infty} \langle \phi_0, h_{jk} (\delta^{ij} + dx^i.dx^j.) dx^0.dx^k.\phi_0\rangle * dx^i \\
&= \frac{1}{4}\,\mathrm{Re} \int_{\partial M_\infty} \langle \phi_0, (\Gamma_{jij} + (2\Gamma_{jjl} dx^i.dx^l. - \Gamma_{kjj} dx^i.dx^k.))\phi_0\rangle * dx^i \\
&\quad + \frac{1}{2} \int_{\partial M_\infty} \langle \phi_0, (h_{ik} - \delta_{ik} h_{jj}) dx^0.dx^k.\phi_0\rangle * dx^i
\end{aligned}$$

$$=\frac{1}{4}\int_{\partial M_\infty}\langle\phi_0,(\gamma_{jij}-2\Gamma_{jji}+\Gamma_{ijj})\phi_0\rangle * dx^i$$

$$+\frac{1}{2}\int_{\partial M_\infty}\langle\phi_0,(h_{ik}-\delta_{ik}h_{jj})dx^0.dx^k.\phi_0\rangle * dx^i$$

$$=\frac{C_4}{4}(\langle\phi_0.E_l\phi_2\rangle+\langle\phi_0,p_{lk}dx^0.dx^k.\phi_0\rangle).$$

But $p_{lk}dx^0.dx^k.=(\begin{smallmatrix}0 & p_l\\ \bar{p}_l & 0\end{smallmatrix})$, where $p_l=p_{l1}+p_{l2}i+p_{l3}j+p_{l4}k$, has real eigenvalues $\lambda=\pm|P_l|$. Now we take ϕ_0 to be the eigenspinor of eigenvalue $-|P_l|$ with $|\phi_0|=1$. In terms of this constant spinor, we finally obtain

$$E_l-|P_l|=4C_4^{-1}\int_M|\tilde{\nabla}\phi|^2+\langle\phi,\tilde{R}\phi\rangle\geqslant 0.$$

Thus the proof of the first part is complete.

Now suppose $E_1=0$. Take constant spinors $\{\psi_{1c}|c=1,2\}$ which form a basis on M_1 and $\psi_{1c}=0$ on all other ends M_l. Let ψ_c be the solutions of $\tilde{D}\psi_c=0$ constructed from this data by Theorem 4.1. The vanishing of E_1 then implies $\tilde{\nabla}\psi_c=0$ and $\psi_c\to 0$ uniformly on each end except M_1. But this contradicts Lemma 4.1 (i) unless M_1 is the only end of M. By Lemma 4.1 (ii), $\{\psi_c\}$ are linearly independent everywhere on M, so in a local frame $\{e_i\}$ of M,

$$0=(\tilde{\nabla}_i\tilde{\nabla}_j-\tilde{\nabla}_j\tilde{\nabla}_i-\tilde{\nabla}_{[e_i,e_j]})\psi_c=-\tfrac{1}{4}\tilde{R}_{\alpha\beta ij}e^\alpha.e^\beta.\psi_c.$$

This implies

$$\tilde{R}_{\alpha\beta ij}=0,\quad 1\leqslant i,j\leqslant 4,\quad 0\leqslant\alpha,\beta\leqslant 4. \tag{5.1}$$

Then Einstein's equations give $T_{00}=\frac{1}{2}\Sigma\tilde{R}_{ijij}=0$. And the dominant energy condition shows $|T_{\alpha\beta}|\leqslant|T_{00}|=0$. Hence $\tilde{R}_{\alpha\beta}=0$. This together with (5.1) implies

$$\tilde{R}_{\alpha\beta\gamma\delta}=0,\quad 0\leqslant\alpha,\beta,\gamma,\delta\leqslant 4.$$

Therefore N is flat along M. And we complete the proof of Theorem. □

ACKNOWLEDGMENTS

This work is a part of author's thesis at The Chinese University of Hong Kong. He would like to express his deep gratitude to Professor S. T. Yau for introducing him to the subjects and the problems, and for his guidance, continuous encouragement and various kinds of help, and to thank Professors S. Y. Cheng, L. F. Tam, and G. Tian for their helpful discussions. He would also like to thank the referee for many valuable suggestions.

[1] R. Bartnik, ''The mass of an asymptotically flat manifold,'' Commun. Pure Appl. Math. **36**, 661–693 (1986).
[2] S. Hawking and S. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973).
[3] S. Arnowitt, S. Deser, and C. Misner, ''Coordinate invariance and energy expressions in general relativity,'' Phys. Rev. **122**, 997-1006 (1961).
[4] R. Schoen and S. T. Yau, ''On the proof of the positive mass conjecture in general relativity,'' Commun. Math. Phys. **65**, 45–76 (1979).
[5] P. S. Jang, ''On the positivity of energy in general relativity,'' J. Math. Phys. **19**, 1152–1155 (1978).
[6] R. Schoen and S. T. Yau, ''Proof of the positive mass theorem. II,'' Commun. Math. Phys. **79**, 231–260 (1981).
[7] R. Schoen and S. T. Yau, ''Proof of the positive action conjecture in quantum relativity,'' Phys. Rev. Lett. **42**, 547–548 (1979).
[8] E. Witten, ''A new proof of the positive energy theorem,'' Commun. Math. Phys. **80**, 381–402 (1981).

[9] T. Parker and C. Taubes, ''On Witten's proof of the positive energy theorem,'' Commun. Math. Phys. **84**, 223–238 (1982).
[10] R. Schoen, *Variational Theory for the Total Scalar Curvature Functional for Riemannian Metric and Related Topics*, Lecture Notes in Math 1365 (Springer, Berlin, 1987), pp. 120–154.
[11] X. Zhang, ''Positive mass theorem for modified energy condition'' (preprint).
[12] X. Zhang, ''Positive mass theorem for hypersurface in five-dimensional Lorentzian manifolds,'' to appear in Commun. Anal. Geom.
[13] R. Harvey, *Spinors and Calibrations* (Academic, New York, 1989).
[14] J. Lee and T. Parker, ''The Yamabe problem,'' Bull. Am. Math. Soc. **17**, 31–81 (1987).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Two-parameter nonstandard deformation of 2×2 matrices

Salih Çelik
Mimar Sinan University, Department of Mathematics, 80690 Besiktas, Istanbul, Turkey

Sultan A. Çelik
Yildiz Technical University, Department of Mathematics, Sisli, Istanbul, Turkey

(Received 11 July 1997; accepted for publication 10 March 1999)

We introduce a two-parameter deformation of 2×2 matrices without imposing any condition on the matrices and give the universal R-matrix of the nonstandard quantum group which satisfies the quantum Yang–Baxter relation. Although in the standard two-parameter deformation the quantum determinant is not central, in the nonstandard case it is central. We note that the quantum group thus obtained is related to the quantum supergroup $GL_{p,q}(1|1)$ by a transformation. © *1999 American Institute of Physics.* [S0022-2488(99)03806-2]

I. INTRODUCTION

Recently the matrix groups of all 2×2 nonsingular matrices like $GL(2)$, $GL(1|1)$, etc., were generalized in two ways as the standard deformation[1–3] and h-deformation.[4–7] Both are based on the deformation of the algebra of functions on the groups generated by coordinate functions that commute.

In standard deformation of matrix groups, these commutation relations are determined by a matrix R so that the functions do not commute but satisfy the equation

$$\hat{R}(T\otimes T)=(T\otimes T)\hat{R},$$

such that, they coincide with the matrix groups for particular values of the deformation parameter. In the h-deformation, this property is the same as the standard deformation. The structure of the matrix groups is important in both deformations since the classical (or super) matrix groups are obtained in some limit of the deformation parameters. In this work we shall construct a two-parameter deformation of 2×2 matrices without imposing any such condition on the matrices just as in Ref. 8 and obtain a two-parameter generalization of their results.

We briefly describe the content of this work. In Sec. II we introduce the group $G_{p,q}$ of the 2×2 matrices by using an R matrix. Section III is devoted to the corresponding Hopf algebra. In Sec. IV we give the universal enveloping algebra of this nonstandard quantum group.

II. $G_{p,q}$ MATRICES

Let

$$T=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

be a 2×2 matrix with entries belonging to an algebra $\mathcal{A}$. We assume that the quantum group equation (no-grading)

$$\hat{R}(T\otimes T)=(T\otimes T)\hat{R} \tag{1}$$

holds, where

$$\hat{R}=\begin{pmatrix} -q & 0 & 0 & 0 \\ 0 & p^{-1}-q & 1 & 0 \\ 0 & qp^{-1} & 0 & 0 \\ 0 & 0 & 0 & p^{-1} \end{pmatrix}. \tag{2}$$

Equation (1) explicitly gives the following relations:

$$\begin{aligned} ab&=-qba, \quad db=qbd, \\ ac&=-pca, \quad dc=pcd, \\ bc&=pq^{-1}cb, \quad b^2=0=c^2, \\ ad&=da+(p^{-1}-q)bc, \end{aligned} \tag{3}$$

where p and q are nonzero complex numbers with $pq\pm 1\neq 0$.

It can be checked that the matrix $R=P\hat{R}$, where P is the usual permutation matrix, satisfies the quantum Yang–Baxter equation

$$R_{12}R_{13}R_{23}=R_{23}R_{13}R_{12} \tag{4}$$

and the matrix $\hat{R}$ satisfies the braid group equation

$$\hat{R}_{12}\hat{R}_{23}\hat{R}_{12}=\hat{R}_{23}\hat{R}_{12}\hat{R}_{23}. \tag{5}$$

We now assume that the matrix elements a and d of T are invertible. Then it is possible to define the inverse of T. To this end, we introduce

$$\Delta_1=ad-p^{-1}bc, \quad \Delta_2=da+q^{-1}cb. \tag{6}$$

Then one obtains

$$T_R^{-1}=\begin{pmatrix} \Delta_1^{-1}d & -q\Delta_1^{-1}b \\ p\Delta_2^{-1}c & \Delta_2^{-1}a \end{pmatrix}, \tag{7a}$$

as the right inverse of T. After some calculations we get

$$\Delta_1 d=d\Delta_1, \quad \Delta_2 a=a\Delta_2,$$

$$\Delta_k b=-q^2 b\Delta_k, \quad \Delta_k c=-p^2 c\Delta_k, \quad k=1,2.$$

Using these relations we obtain

$$T_L^{-1}=\begin{pmatrix} d\Delta_1^{-1} & q^{-1}b\Delta_2^{-1} \\ -p^{-1}c\Delta_1^{-1} & a\Delta_2^{-1} \end{pmatrix}=T_R^{-1}. \tag{7b}$$

Thus the proper left and right inverses of T are equal.

It is easily verified that $a^2\Delta_2^{-1}$ for all values of p and q, commutes with a,d, and anticommutes with b,c. Furthermore $a^2\Delta_2^{-1}$ is invertible. Therefore we obtain

$$S(T)=T^{-1}=\begin{pmatrix} d^{-1} & -a^{-1}ba^{-1} \\ -d^{-1}cd^{-1} & a^{-1} \end{pmatrix}\begin{pmatrix} d^2\Delta_1^{-1} & 0 \\ 0 & a^2\Delta_2^{-1} \end{pmatrix}. \tag{8}$$

We now consider the element

$$D(T)=ad^{-1}-bd^{-1}cd^{-1}=a^2\Delta_2^{-1}. \tag{9}$$

$D(T)$ cannot be regarded as a quantum determinant since it anticommutes with b and c. However, we may regard the element

$$\mathbf{D}(T)=a[D(T)-d^{-1}bd^{-1}c]d^{-1}=[D(T)]^2 \tag{10}$$

as the quantum determinant of T where $D(T)$ is given by (9).

In fact, it is easy to check that the matrix elements of the product matrix TT' satisfy relations (3) for any two commuting quantum matrices T and T' whose elements obey (3). As a consequence of this argument, we have the following relation:

$$D(TT')=D(T)D(T').$$

This result means that $\mathbf{D}(T)$ is central.

This case appears strange from the point of view of quantum group theory.[2] However, it becomes clear from the point of view of the corresponding two-parameter quantum supergroup.[9] We know, from the work of Ref. 9, that the quantum superdeterminant of any supermatrix in $GL_{p,q}(1|1)$ belongs to the center of the algebra generated by the matrix elements of the supermatrix. In the Appendix, we shall show that this nonstandard quantum group is related to the quantum supergroup $GL_{p,q}(1|1)$ by a transformation. So we may expect that the quantum superdeterminant in two-parameter nonstandard deformation must again be a central element.

Now let the nth power of T be

$$T^n=\begin{pmatrix} A_n & B_n \\ C_n & D_n \end{pmatrix}. \tag{11}$$

Then it is easy to check the following relations:

$$\begin{aligned} &A_nB_n=-q^nB_nA_n\,, \quad D_nB_n=q^nB_nD_n\,,\\ &A_nC_n=-p^nC_nA_n\,, \quad D_nC_n=p^nC_nD_n\,,\\ &B_n^2=0=C_n^2, \quad q^nB_nC_n=p^nC_nB_n\,, \end{aligned} \tag{12}$$

and

$$A_nD_n=D_nA_n+(p^{-n}-q^n)C_nB_n\,. \tag{13}$$

The proof of relation (13) is rather lengthy but straightforward.

Let us finally note the following. If the sum $T+T'$ of two $G_{p,q}$ matrices T and T' is required to be a $G_{p,q}$ matrix then the equation

$$\hat{R}'(T\otimes T')=(T'\otimes T)\hat{R}^{-1} \tag{14}$$

holds, where

$$\hat{R}'=\hat{R}^{-1}-(p-q^{-1})I. \tag{15}$$

Equation (14) explicitly reads

$$\begin{aligned} &a'a=pqaa', \quad dd'=pqd'd, \quad d'a=ad',\\ &b'a=-pab', \quad c'a=-qac', \quad bd'=pd'b,\\ &bb'=-b'b, \quad cc'=-c'c, \quad cd'=qd'c, \end{aligned}$$

$$a'b=-qba'+(p^{-1}-q)b'a, \quad a'c=-pca'+(pq-1)ac',$$
$$b'c=pq^{-1}cb'+(q^{-1}-p)ad', \quad bc'=pq^{-1}c'b+(p-q^{-1})d'a,$$
$$db'=qb'd+(q-p^{-1})bd', \quad dc'=pc'd+(pq-1)d'c,$$
$$a'd=da'+(p^{-1}-q)(bc'+b'c). \tag{16}$$

Note that the matrix $R'=P\hat{R}'$ again satisfies the quantum Yang–Baxter relation (4), where P is the usual permutation matrix.

III. THE HOPF ALGEBRA STRUCTURE OF $G_{p,q}$

Let $\mathcal{A}$ be an algebra generated by the elements a, b, c, and d satisfying the relations (3). Then $\mathcal{A}$ is the quoitent algebra

$$\mathcal{A}=\mathcal{C}[a,b,c,d]/J,$$

where $\mathcal{C}[a,b,c,d]$ is the free noncommutative algebra generated by a, b, c, and d and J is the ideal in $\mathcal{C}[a,b,c,d]$ generated by the relations (3).

The usual coproduct on the algebra $\mathcal{A}$ is defined by

$$\Delta:\mathcal{A}\rightarrow\mathcal{A}\otimes\mathcal{A}$$

such that

$$\Delta(t^i{}_j)=t^i{}_k\otimes t^k{}_j, \quad T=(T^i_j) \tag{17}$$

(sum over repeated indices) and the counit

$$\epsilon:\mathcal{A}\rightarrow\mathcal{C}$$

such that

$$\epsilon(t^i{}_j)=\delta^i{}_j. \tag{18}$$

The algebra $\mathcal{A}$ is now the matrix bialgebra generated by 1 and $T=(t^i_j)$, and it is a Hopf algebra with the antipode $S(T)$ which is given by (8). To give a proof of this, one has to verify the following:

$$(id\otimes\Delta)\circ\Delta=(\Delta\otimes id)\circ\Delta,$$
$$(id\otimes\epsilon)\circ\Delta=(\epsilon\otimes id)\circ\Delta, \tag{19}$$
$$m\circ[(id\otimes S)\circ\Delta]=m\circ[(S\otimes id)\circ\Delta],$$

where m denotes the multiplication mapping

$$m(a\otimes b)=ab$$

for any $a,b\in\mathcal{A}$. The proof follows directly.

IV. UNIVERSAL ENVELOPING ALGEBRA OF $G_{p,q}$

In this section we shall construct the quantum enveloping algebra in analogy with the FRT approach.[2]

We consider the matrices $L^{\pm}$ with the generators $U_{\pm}$, $V_{\pm}$, and $X_{\pm}$,

$$L^{+}=\begin{pmatrix} U_{+} & \lambda X_{+} \\ 0 & V_{+} \end{pmatrix}, \quad L^{-}=\begin{pmatrix} U_{-} & 0 \\ -\lambda X_{-} & V_{-} \end{pmatrix}, \tag{20}$$

where $\lambda = q - p^{-1}$. The matrices $L^{\pm}$ satisfy the following relations:

$$\hat{R} L_1^{\pm} L_2^{\pm} = L_2^{\pm} L_1^{\pm} \hat{R}, \tag{21}$$

$$\hat{R} L_1^{+} L_2^{-} = L_2^{-} L_1^{+} \hat{R}, \tag{22}$$

where $L_1 = L \otimes I$ and $L_2 = I \otimes L$. These relations give

$$\begin{aligned} &[U_+, U_-] = [V_+, V_-] = [U_\pm, V_\pm] = 0, \\ &U_+ X_\pm = -q^{\mp 1} X_\pm U_+, \quad V_+ X_\pm = q^{\mp 1} X_\pm V_+, \\ &U_- X_\pm = -p^{\pm 1} X_\pm U_-, \quad V_- X_\pm = p^{\pm 1} X_\pm V_-, \\ &X_+ X_- - q p^{-1} X_- X_+ = \frac{U_+ V_- - V_+ U_-}{q - p^{-1}}, \quad X_\pm^2 = 0. \end{aligned} \tag{23}$$

The coproduct of the generators is given by

$$\Delta(L^{\pm}) = L^{\pm} \dot{\otimes} L^{\pm}, \tag{24}$$

where $\dot{\otimes}$ denotes tensor product and matrix multiplication. Explicity, the action of the coproduct Δ on the generators is

$$\begin{aligned} \Delta(U_\pm) &= U_\pm \otimes U_\pm, \\ \Delta(V_\pm) &= V_\pm \otimes V_\pm, \\ \Delta(X_+) &= X_+ \otimes U_+ + V_+ \otimes X_+, \\ \Delta(X_-) &= X_- \otimes V_- + U_- \otimes X_-. \end{aligned} \tag{25}$$

The co-unit is given by

$$\epsilon(L^{\pm}) = I. \tag{26}$$

Explicitly,

$$\begin{aligned} \epsilon(U_\pm) &= \epsilon(V_\pm) = 1, \\ \epsilon(X_\pm) &= 0. \end{aligned} \tag{27}$$

The co-inverse is given by

$$\begin{aligned} &S(U_\pm) = U_\pm^{-1}, \quad S(V_\pm) = V_\pm^{-1}, \\ &S(X_+) = -U_+^{-1} X_+ V_+^{-1}, \\ &S(X_-) = V_-^{-1} X_- U_-^{-1}. \end{aligned} \tag{28}$$

Therefore one can easily verify that the algebra $\mathcal{U}_{p,q}(U_\pm, V_\pm, X_\pm)$ is a Hopf algebra generated by 1, $U_\pm$, $V_\pm$, $X_\pm$ satisfying the relations (23).

The coproduct of $U_\pm$ and $V_\pm$ together with the fact that they commute implies that they can be written as exponentials of commuting operators,

$$U_+=q^{-(H/2)}p^{(N/2)}, \quad U_-=p^{(H/2)}q^{-(N/2)},$$
$$V_+=q^{-(H/2)}p^{-(N/2)}, \quad V_-=p^{(H/2)}q^{(N/2)}, \tag{29}$$
$$[H,N]=0.$$

The commutation relations of $U_\pm$ and $V_\pm$ with $X_\pm$ in terms of new generators give the following:

$$[H,X_\pm]=\pm 2X_\pm, \quad [N,X_\pm]=0,$$
$$X_+X_- - qp^{-1}X_-X_+ = \left(\frac{p}{q}\right)^{(H+1)/2}[N]_{pq}, \tag{30}$$

where

$$[N]_{pq}=\frac{(pq)^{(N/2)}-(pq)^{(-N/2)}}{(pq)^{(1/2)}-(pq)^{(-1/2)}}. \tag{31}$$

Moreover, the coproduct is now

$$\Delta(H)=H\otimes \mathbf{1}+\mathbf{1}\otimes H,$$
$$\Delta(N)=N\otimes \mathbf{1}+\mathbf{1}\otimes N,$$
$$\Delta(X_+)=X_+\otimes q^{-(H/2)}p^{(N/2)}+q^{-(H/2)}p^{-(N/2)}\otimes X_+, \tag{32}$$
$$\Delta(X_-)=X_-\otimes p^{(H/2)}q^{(N/2)}+p^{(H/2)}q^{-(N/2)}\otimes X_-.$$

ACKNOWLEDGMENTS

This work was supported in part by T. B. T. A. K. the Turkish Scientific and Technical Council. We would like to express our deep gratitude to the referee for critical comments on the manuscript.

APPENDIX

A. Nonstandard quantum planes

In this section, we shall consider quantum planes which are similar to quantum superplanes introduced by Manin.[3]

(1) Quantum plane A_p: This plane, or, rather the polynomial function ring on it is generated by coordinates x and θ with the commutation rules

$$x\theta=-p\theta x, \quad \theta^2=0, \tag{A1}$$

where p is a complex number. The coordinates anticommute for $p=1$ and commute for $p=-1$.

(2) Quantum plane A_q^*: This plane is generated by coordinates φ and y with commutation rules

$$\varphi^2=0, \quad \varphi y=q^{-1}y\varphi, \tag{A2}$$

where q is a complex number. The quantum plane A_q^* is dual to the quantum plane A_p.

Note that the relations (A1) and (A2) are equivalent to the relations

$$\hat{R}(X\otimes X)=-q(X\otimes X),\quad \hat{R}(Y\otimes Y)=p^{-1}(Y\otimes Y).$$

B. Nonstandard quantum deformation of 2×2 matrices with nonstandard quantum planes

Let G be a matrix Lie group of rank 2 and T be any element of G, i.e.,

$$T=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with entries belonging to an algebra $\mathcal{A}$.

We consider linear transformations T with the following properties:

$$T:A_p\rightarrow A_p\,,\quad T:A_q^*\rightarrow A_q^*\,. \tag{A3}$$

The action of T on points of A_p and A_q^* is

$$\begin{pmatrix} \overline{x} \\ \overline{\theta} \end{pmatrix}=\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ \theta \end{pmatrix},\quad \begin{pmatrix} \overline{\varphi} \\ \overline{y} \end{pmatrix}=\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} \varphi \\ y \end{pmatrix}. \tag{A4}$$

We assume that the matrix elements of T commute with the coordinates of A_p and A_q^*. As a consequence of the linear transformations in (A3) the vectors $(\begin{smallmatrix} \overline{x} \\ \overline{\theta} \end{smallmatrix})$ and $(\begin{smallmatrix} \overline{\varphi} \\ \overline{y} \end{smallmatrix})$ should belong to A_p and A_q^*, respectively. This imposes (p,q) commutation relations among the entries of T in (3).

Note that it can be checked that the maps

$$\delta:A_p\rightarrow G\otimes A_p\,,\quad \delta^*:A_q^*\rightarrow G\otimes A_q^* \tag{A5}$$

such that

$$\delta(X)=T\otimes X,\quad \text{i.e.,}\quad \delta(x_i)=t_i^j\otimes x_j\,,\quad X=\begin{pmatrix} x \\ \theta \end{pmatrix},$$

$$\delta(Y)=T\otimes Y,\quad \text{i.e.,}\quad \delta(y_i)=t_i^j\otimes y_j\,,\quad Y=\begin{pmatrix} \varphi \\ y \end{pmatrix} \tag{A6}$$

define the co-action of the quantum group $G_{p,q}$ on the nonstandard quantum planes A_p and A_q^*, respectively.

Finally, one can show that the matrix quantum group (3) is isomorphic to the quantum supergroup $GL_{p,q}(1|1)$. Indeed, if we define the transformation

$$T'=TD, \tag{A7}$$

where T is a matrix whose the matrix elements satisfy (3) and

$$D=\begin{pmatrix} 1 & 0 \\ 0 & g \end{pmatrix},\quad D^2=I, \tag{A8}$$

and we assume that g commutes with a and d, and anticommutes with b and c, then $T'\in GL_{p,q}(1|1)$ as discussed in Ref. 9. In this case $\Delta(T')=T'\dot{\otimes}T'$, etc., are unchanged. One easily sees that when $p=q$, these relations go back to those of Ref. 8.

[1] V. G. Drinfeld, in Proceedings of the IMS, Berkeley, 1986 (unpublished).
[2] N. Y. Reshetikhin, L. A. Takhtajan, and L. D. Faddeev, Leningrad Math. J. **1**, 193 (1990).
[3] Yu I. Manin, Commun. Math. Phys. **123**, 163 (1989).
[4] S. L. Woronowicz, Rep. Math. Phys. **30**, 259 (1991).

[5] S. Zakrzewski, Lett. Math. Phys. **22**, 287 (1991); C. H. Ohn, *ibid.* **25**, 85 (1992); B. A. Kupershmidt, J. Phys. A **25**, L1239 (1992); V. Karimipour, Lett. Math. Phys. **30**, 87 (1994).
[6] A. Aghamohammadi, M. Khorrami, and A. Shariati, J. Phys. A **28**, L225 (1995).
[7] S. Çelik, Lett. Math. Phys. **42**, 299 (1997), and references therein.
[8] S. Majid and M. J. Rodriguez-Plaza, J. Math. Phys. **36**, 7081 (1995).
[9] L. Dabrowski and L. Wang, Phys. Lett. B **266**, 51 (1991).

Closed forms for the exponential mapping on matrix Lie groups based on Putzer's method

F. Silva Leite[a)]
Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra–Portugal and Instituto de Sistemas e Robótica, Pólo de Coimbra, 3030 Coimbra–Portugal

P. Crouch[b)]
Center for Systems Science and Engineering, Arizona State University, Tempe, Arizona 85287

(Received 29 June 1998; accepted for publication 12 March 1999)

We present closed forms for the exponential of some infinitesimal generators of Lie groups which play an important role in physics and engineering applications. These explicit forms are based on the Putzer's method. We also compare this methodology and results with related work by other authors. © *1999 American Institute of Physics.* [S0022-2488(99)00607-6]

I. INTRODUCTION

Lie groups arise most often in engineering applications as the configuration space of mechanical systems and in physics as symmetry groups associated with conservation laws. We refer to Sattinger and Weaver,[1] for an introduction to the role of Lie groups in various fields of mathematics and physics. After Heisenberg in 1932, the neutron and the proton in nuclear physics started being treated as two states of an $su(2)$ doublet, the nucleon. In this sense, the proton and neutron are seen as families of particles that can be transformed into one another by the operations of the symmetry group SU(2). Later, the discovery of quarks from which other particles could be built, replaced the role of SU(2) by that of SU(3). The evolution of particle physics has led to the introduction of larger and more complex Lie groups. For instance, the 21-dimensional symplectic group $SP(3,\mathbb{R})$ is the dynamical group in the microscopic theory of nuclear collective motion (see, for instance, Rowe[2] for additional sources regarding the role of the symplectic group in physics). Many of these applications rely crucially on the use of the exponential mapping. Although the existence of an exponential mapping is guaranteed on any Lie group, finding a closed form for the exponential is a difficult issue. One way to circumvent this problem is by using approximation methods. We refer to the survey paper by Moler and Van Loan[3] for an account of several methods to compute the matrix exponential together with an analysis on the efficiency of the existing algorithms.

For some low-dimensional Lie groups however, there are explicit formulas for the exponential mapping, the most notable one being the Rodrigues' formula for the exponential on the rotation group $SO(3,\mathbb{R})$. The motion of a charged particle under the action of an electromagnetic field is given by a differential equation evolving on the Lorentz group $SO(1,3)$. For this Lie group, Zeni and Rodrigues presented in Ref. 4 an explicit formula for the exponential mapping. More recently we became aware of the work of Barut, Zeni, and Laufer,[5] generalizing the ideas in Ref. 4 to orthogonal groups and presenting a closed form for the exponential mapping for the conformal group O(2,4). As pointed out in their work, the Cayley–Hamilton theorem applied to the generators of orthogonal groups produces either even or odd powers, which amounts to a great simplification for the series defining the exponential. This remarkable characteristic of orthogonal groups is shared by other classical Lie groups as will be explained in Sec. II. Although in the present

[a)]Electronic mail: fleite@mat.uc.pt; Fax: (351)39-832568.
[b)]Electronic mail: peter.crouch@asu.edu; Fax: (1) 602-965 2267.

paper the Cayley–Hamilton theorem also plays an important role, we take a different route from that of Barut, Zeni, and Laufer in Ref. 5 and use instead the Putzer's method.[6] This is key for finding closed forms for the exponential mapping on any matrix Lie group, without having to concern ourselves about the minimal polynomial of a matrix or its Jordan canonical form. One drawback in this method is that one assumes *a priori* that the eigenvalues of the matrix to be exponentiated are known. But this was also an assumption in the work of Barut, Zeni, and Laufer. In Sec. II we define a large class of Lie algebras having a symmetric spectrum. This class contains the Lie algebras that appeared in Refs. 4 and 5. The symmetric property of the spectrum considerably reduces the amount of work to be implemented. In Sec. III we briefly describe Putzer's algorithm and illustrate how to obtain closed forms for the exponential of infinitesimal generators for Lie groups playing a role in physics, including the 21-dimensional symplectic group.

II. PROPERTIES OF *P*-SKEW-SYMMETRIC MATRICES

Let $\mathcal{G}l(n,\mathbb{R})$ denote the set of all $n\times n$ matrices with real entries and P be any $n\times n$ orthogonal matrix, i.e., P satisfies $P^T=P^{-1}$. Define the following set:

$$G=\{X\in\mathcal{G}l(n,\mathbb{R}):X^TPX=P\}, \tag{1}$$

where X^T stands for the transpose X. G is an algebraic closed subgroup of the general linear Lie group $GL(n,\mathbb{R})$, consisting of the invertible matrices in $\mathcal{G}l(n,\mathbb{R})$, and so G is itself a Lie group (Helgason[7]). The infinitesimal generators of G are the matrices belonging to its Lie algebra. This Lie algebra can be identified with the tangent space of G at the identity I_n, that consists of all vectors that are tangent, at the identity, to curves $t\to X(t)$ in G, passing through the identity at time $t=0$. This geometric interpretation of the Lie algebra of a Lie group is enough to find the Lie algebra $\mathcal{L}$ of G. Indeed, differentiating both sides of $X^T(t)PX(t)=P$ with respect to t one gets

$$\frac{dX^T(t)}{dt}PX(t)+X^T(t)P\frac{dX(t)}{dt}=0.$$

Now evaluating at $t=0$ having into consideration that $X(0)=I$ and making $dX(t)/dt|_{t=0}=A$, it follows that the Lie algebra of G is defined by

$$\mathcal{L}=\{A\in\mathcal{G}l(n,\mathbb{R}):A^TP=-PA\}. \tag{2}$$

The Lie bracket on $\mathcal{L}$ is the commutator. It follows from the definition of $\mathcal{L}$ that, if $A\in\mathcal{L}$ then $A^T\in\mathcal{L}$. For the particular situation when $P=I$, $\mathcal{L}$ is the set of skew-symmetric matrices. From now on we will refer to the matrices in $\mathcal{L}$ as P-skew-symmetric.

Now consider the following set of matrices associated with $\mathcal{L}$:

$$\mathcal{J}=\{A\in\mathcal{G}l(n,\mathbb{R}):A^TP=PA\}. \tag{3}$$

It also happens that if $A\in\mathcal{J}$ then $A^T\in\mathcal{J}$. However, $\mathcal{J}$ is not closed under the commutator and so does not have the structure of a Lie algebra. Instead, $\mathcal{J}$ is closed under the product $\{A,B\}=AB+BA$ and possesses another interesting algebraic structure, namely the structure of a Jordan algebra. (For details concerning Jordan algebras see, for instance, Jacobson.[8]) When $P=I,\mathcal{J}$ is just the set of symmetric matrices. We will refer to the matrices in $\mathcal{J}$ as P-symmetric.

The matrices belonging to $\mathcal{L}$ and those belonging to $\mathcal{J}$ have remarkable properties. We start with a decomposition theorem that generalizes a very well known result in matrix theory, namely that every square real or complex matrix can be uniquely decomposed as the sum of a skew-symmetric matrix and a symmetric matrix.

Theorem 1: *If $P^2=\pm I$ and $\mathcal{L}$ and $\mathcal{J}$ are defined, respectively, by (2) and (3), then*

$$\mathcal{G}l(n,\mathbb{R})=\mathcal{L}\oplus\mathcal{J}. \tag{4}$$

Proof: Clearly $\mathcal{L}\cap\mathcal{J}=\{0\}$. Also any matrix $A\in\mathcal{G}l(n,\mathbb{R})$ may be written uniquely as

$$A=\tfrac{1}{2}(A-PA^TP)+\tfrac{1}{2}(A+PA^TP).$$

If we make $\frac{1}{2}(A-PA^TP)=A_1$ and $\frac{1}{2}(A+PA^TP)=A_2$, a trivial calculation shows that if $P^2=I$ then $A_1\in\mathcal{L}$ and $A_2\in\mathcal{J}$, while if $P^2=-I$ then $A_1\in\mathcal{J}$ and $A_2\in\mathcal{L}$. □

Lemma 2: If $A\in\mathcal{L}$ and $B\in\mathcal{J}$, then

(a) $A^{2j}\in\mathcal{J}$ and $A^{2j+1}\in\mathcal{L}$, $\forall j\in\mathbb{N}$;

(b) $B^j\in\mathcal{J}$, $\forall j\in\mathbb{N}$;

(c) $[\mathcal{L},\mathcal{L}]\subset\mathcal{L}$, $[\mathcal{L},\mathcal{J}]\subset\mathcal{J}$, $[\mathcal{J},\mathcal{J}]\subset\mathcal{L}$.

Proof: (a) May be easily proven by induction on *j*, (b) is a consequence of the fact that $\mathcal{J}$ is closed under $\{\cdot,\cdot\}$ and (c) follows easily from the definitions of $\mathcal{L}$ and $\mathcal{J}$. □

Theorem 3: *P*-skew symmetric matrices have a symmetric spectrum.

Proof: This is a consequence of the fact that if $A^TP=-PA$, then $\det(A-\lambda I)=(-1)^n\det(A+\lambda I)$. □

Corollary 4: If $A_{n\times n}$ is a P-skew-symmetric matrix, then

(a) trace $(A)=0$;

(b) If n is odd, $\det(A)=0$;

(c) The characteristic polynomial of A decomposes as

$$\det(A-\lambda I)=\lambda^r(\lambda^{2k}+C_2\lambda^{2k-2}+\cdots+C_{2k-2}\lambda^2+C_{2k}), \tag{5}$$

where r equals the number of zero eigenvalues of A and the coefficients $C_2,\ldots,C_{2k}$ may be written in terms of the nonzero eigenvalues of A, $\pm\lambda_1,\pm\lambda_2,\ldots,\pm\lambda_k$, in the following way:

$$C_{2l}=(-1)^l\sum_{1\leq i_1<i_2<\cdots<i_l\leq k}\lambda_{i_1}^2\lambda_{i_2}^2\cdots\lambda_{i_l}^2,\quad l=1,\ldots,\ k. \tag{6}$$

Proof: (a), (b) and first part of (c) are immediate consequences of theorem 3. Formula (6) can be proven by induction in the order of matrix *A* or, alternatively, we may use the fact that the spectrum of *A* is symmectic associated with the well known result, that the coefficients of the characteristic polynomial of a matrix *A* are the elementary symmetric functions of its eigenvalues. □

The coefficients of the characteristic polynomial of a general matrix may be written in terms of the traces of its powers, as shown in Barut, Zeni, and Laufer.[5] Unfortunately, that formula is very difficult to unravel, even though the traces of its powers may be easily computed if one knows the eigenvalues of the matrix. Equation (6) gives a more usable formula for the coefficients of the characteristic polynomial of a *P*-skew-symmetric matrix. Nevertheless, the following particular cases can be easily proven:

$$C_2=-\sum_{i=1}^{k}\lambda_i^2=-\frac{1}{2}\operatorname{trace}(A^2),$$

$$C_4=\sum_{1\leq i<j\leq k}\lambda_i^2\lambda_j^2=\frac{1}{8}(\operatorname{trace}(A^2))^2-\frac{1}{4}\operatorname{trace}(A^4),$$

$$C_6=\sum_{1\leq i<j<s\leq k}\lambda_i^2\lambda_j^2\lambda_s^2=\frac{1}{48}(\operatorname{trace}(A^2))^3-\frac{1}{8}\operatorname{trace}(A^4)\operatorname{trace}(A^2)+\frac{1}{6}\operatorname{trace}(A^6),$$

$$C_{2k}=(-1)^k\lambda_1^2\lambda_2^2\cdots\lambda_k^2. \tag{7}$$

Note that the formula for C_2 is a particular case of the well known formula $C_2 = \frac{1}{2}[(\mathrm{trace}(A))^2 - \mathrm{trace}(A^2)]$, when $\mathrm{trace}(A)=0$ (see, for instance, Ref. 7, p. 192 and Ref. 5) and $C_{2k}=(-1)^k \lambda_1^2 \lambda_2^2 \cdots \lambda_k^2$, coincides with $(-1)^k \det(A)$, whenever $\det(A)\neq 0$.

Particular cases of P-skew-symmetric matrices are the following:

(1) $P = I_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$, with $p+q=n$.

In this case, G is the orthogonal group $\mathrm{O}(p,q)$. This was the case treated in Barut, Zeni, and Laufer.[5]

The particular situation when $p=1$ and $q=3$, corresponds to the case when the Lie group G is the Lorentz group and has been analyzed by Zeni and Rodrigues in Ref. 4.

When $p=n$ and $q=0$, i.e., $P=I$, the Lie algebra of G consists of the skew-symmetric matrices in $so(n,\mathbb{R})$.

(2) $P = J_m = \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix}$, with $2m=n$.

In this case G is the symplectic group $SP(2n,\mathbb{R})$. The Lie algebra of G consists of the set of Hamiltonian matrices.

(3) Although the Lie algebra of skew-Hermitian matrices $u(n)$ was not considered here, it may be embedded into $so(2n,\mathbb{R})$ through the mapping

$$A+iB \rightarrow \begin{bmatrix} A & B \\ -B & A \end{bmatrix},$$

and, consequently, the results in this section extend directly to the complex Lie groups $\mathrm{U}(n)$ and $\mathrm{SU}(n)$.

The last corollary has great impact in the structure of the exponential of P-skew-symmetric matrices. This was already pointed out by Barut, Zeni, and Laufer in Ref. 5, for the situation when the Lie group G is an orthogonal group. Indeed, if $A \in \mathcal{L}$, the power series defining its exponential may be split into even and odd powers as

$$e^{tA} = \sum_{j=0}^{+\infty} \frac{t^{2j} A^{2j}}{(2j)!} + \sum_{j=0}^{+\infty} \frac{t^{2j+1} A^{2j+1}}{(2j+1)!},$$

where the first sum belongs to $\mathcal{J}$ and the second belongs to $\mathcal{L}$.

Also, if $A \in \mathcal{L}$, we may apply corollary 4 and the Cayley–Hamilton theorem to write

$$A^r (A^{2k} + C_2 A^{2k-2} + \cdots + C_{2k-2} A^2 + C_{2k} I) = 0$$

or

$$A^n = -C_2 A^{n-2} - \cdots - C_{2k-2} A^{n-(2k-2)} - C_{2k} A^r.$$

As a consequence, if n is even (odd), r is also even (odd) and A^{n+2j} is a linear combination of the successive even (odd) powers of A, from $A^r (A^{r+1})$ up to $A^{n-2} (A^{n-1})$.

The coefficients in this linear combination may be obtained recursively from the coefficients appearing in the characteristic polynomial of A. To see these recurrence relations we index the coefficients so that if

$$A^n = -(C_2)_0 A^{n-2} - \cdots - (C_{2k-2})_0 A^{n-(2k-2)} - (C_{2k})_0 A^r,$$

then

$$A^{n+2j} = -(C_2)_j A^{n-2} - \cdots - (C_{2k-2})_j A^{n-(2k-2)} - (C_{2k})_j A^r,$$

$$A^{n+2j+1} = -(C_2)_j A^{n-1} - \cdots - (C_{2k-2})_j A^{n-(2k-3)} - (C_{2k})_j A^{r+1}.$$

A simple calculation shows that

$$(C_2)_j=(C_2)_0(C_2)_{j-1}+(C_4)_{j-1},$$

$$(C_4)_j=(C_4)_0(C_2)_{j-1}+(C_6)_{j-1},$$

$$\vdots$$

$$(C_{2k-2})_j=(C_{2k-2})_0(C_2)_{j-1}+(C_{2k})_{j-1},$$

$$(C_{2k})_j=(C_{2k})_0(C_2)_{j-1}.$$

This was the approach used by Zeni and Rodrigues in Ref. 4 and by Barut, Zeni, and Laufer in Ref. 5 to obtain closed forms for the exponential of $SO(1,3)$, the Lorentz group, and for $SO(p,q)$ when $p+q=6$. We point out that the closed forms presented in Barut, Zeni, and Laufer,[5] for the orthogonal group are only valid when a certain discriminant is nonzero, which corresponds to the case when one does not allow repeated eigenvalues for the matrix A. Using the Putzer's method, finding a closed form for the exponential of a matrix A is a simple and clear process, as long as there are methods to evaluate its eigenvalues. In the next section we illustrate Putzer's method for particular Lie algebras of P-skew-symmetric matrices.

III. APPLICATIONS OF PUTZER'S METHOD

In this section we revisit an old method to calculate the exponential of a square matrix A. The following theorem can be directly derived from the original work of Putzer,[6] later referred as the Putzer's method. Although this method has been around for more than three decades, it has not been fully incorporated into the literature. Among the few references we could find are Apostol[9] and, more recently, Horn and Johnson.[10]

Theorem 5 (Putzer's theorem): *If* $\lambda^n+C_1\lambda^{n-1}+\cdots+C_{n-1}\lambda+C_n=0$ *is the characteristic polynomial of a square matrix A, then*

$$e^{At}=f_1(t)I+f_2(t)A^2+\cdots+f_n(t)A^{n-1},$$

where the vector function $\xi=[f_1\ f_2\cdots f_n]^T$ *is the solution of the differential equation*

$$x^{(n)}+C_1x^{(n-1)}+\cdots+C_{n-1}\dot{x}+C_nx=0 \tag{8}$$

that satisfies

$$x(0)=e_1,\ \ \dot{x}(0)=e_2,\ldots,x^{(n-1)}(0)=e_n.$$

Remark 6: It is clear from Putzer's theorem that the coefficient functions f_i, $i=1,2,\ldots,n$, of the polynomial representation of the matrix exponential e^{tA}, are linearly independent solutions of the same nth order linear differential equation with constant coefficients. The roots of the associated characteristic polynomial are the eigenvalues of A. So, if one knows the eigenvalues of A, finding a closed form for e^{tA} reduces to differentiation of elementary functions and solving a system of n algebraic linear equations in n unknowns, having a coefficient matrix which is invertible. Although the eigenvalues of a general $n\times n$ matrix, with $n>4$, cannot be determined analytically, for matrices with additional structure it may be possible to compute directly the eigenvalues even when $n>4$. In particular, for matrices with a symmetric spectrum, like the ones presented in Sec. II, it is possible to compute directly the eigenvalues of matrices up to order 9, covering a range of very interesting cases.

To compute the polynomial representation of e^{tA}, when

$$\lambda^n+C_1\lambda^{n-1}+\cdots+C_{n-1}\lambda+C_n=0,$$

is the characteristic polynomial of A, it is an immediate consequence of Putzer's theorem that one simply has to follow the steps outlined in the following algorithm:

(1) Construct a fundamental set of solutions, $\{\varphi_1(t),\varphi_2(t),\ldots,\varphi_n(t)\}$ of the scalar differential equation

$$x^{(n)}+C_1x^{(n-1)}+\cdots+C_{n-1}\dot{x}+C_nx=0.$$

(This is a relatively straightforward matter if one knows the eigenvalues of A.)

(2) Construct the invertible matrix

$$M=\begin{bmatrix} \varphi_1(0) & \varphi_2(0) & \ldots & \varphi_n(0) \\ \dot{\varphi}_1(0) & \dot{\varphi}_2(0) & \ldots & \dot{\varphi}_n(0) \\ \vdots & \vdots & \ldots & \vdots \\ \varphi_1^{(n-1)}(0) & \varphi_2^{(n-1)}(0) & \ldots & \varphi_n^{(n-1)}(0) \end{bmatrix}.$$

(3) Compute the $1\times n$ matrix function

$$[\varphi_1(t)\ \varphi_2(t)\cdots\varphi_n(t)]M^{-1}.$$

(4) The coefficients $f_1(t),f_2(t),\ldots,f_n(t)$ for the polynomial representation of e^{tA} form the row of the matrix in step (3).

Next we illustrate the use of this algorithm for finding closed forms for the exponential of the infinitesimal generators of some matrix Lie groups, with emphasis in those important in physics. Our first example was chosen to complement the result presented in Barut, Zeni, and Laufer.[5] As already mentioned earlier, their results are only valid when the matrix to be exponentiated has distinct eigenvalues. Here we consider the repeated eigenvalues case. We also present one of the three nondegenerated cases for the exponential of infinitesimal generators of the 12-dimensional symplectic group $SP(3,\mathbb{R})$.

Any mathematical software package like Mathematica, Maple or Matlab may be used successfully to work out other examples. In some cases the formulas may be further simplified, simply by multiplying by det(M) or some factor in det(M). This seems to play the role of the discriminant in Zeni and Rodrigues[4] and Barut, Zeni, and Laufer.[5]

In the following we assume that a, b, and c are nonzero real numbers.

Example 1: Infinitesimal generators of the orthogonal group $SO(2,4)$. If $\sigma(A)=\{\pm ib,\pm ib,\pm c\}$, then

$$e^{tA}=f_1(t)I+f_2(t)A+f_3(t)A^2+f_4(t)A^3+f_5(t)A^4+f_6(t)A^5,$$

where

$$f_1(t)=\frac{c^2(2b^2+c^2)}{(b^2+c^2)^2}\cos bt+\frac{bc^2}{2(b^2+c^2)}t\sin bt+\frac{b^4}{(b^2+c^2)^2}\cosh ct,$$

$$f_2(t)=-\frac{c^2}{2(b^2+c^2)}t\cos bt+\frac{c^2(5b^2+3c^2)}{2b(b^2+c^2)^2}\sin bt+\frac{b^4}{c(b^2+c^2)^2}\sinh ct,$$

$$f_3(t)=-\frac{2b^2}{(b^2+c^2)^2}\cos bt-\frac{b^2-c^2}{2b(b^2+c^2)}t\sin bt+\frac{b^2}{(b^2+c^2)^2}\cosh ct,$$

$$f_4(t)=\frac{b^2-c^2}{2b^2(b^2+c^2)}t\cos bt-\frac{5b^4-c^4}{2b^3(b^2+c^2)^2}\sin bt+\frac{2b^2}{c(b^2+c^2)^2}\sinh ct,$$

$$f_5(t)=-\frac{1}{(b^2+c^2)^2}\cos bt-\frac{1}{2b(b^2+c^2)}t\sin bt+\frac{1}{(b^2+c^2)^2}\cosh ct,$$

$$f_6(t)=\frac{1}{2b^2(b^2+c^2)}\,t\cos bt-\frac{3b^2+c^2}{2b^3(b^2+c^2)^2}\sin bt+\frac{1}{c(b^2+c^2)^2}\sinh ct.$$

Example 2: Infinitesimal generators of the symplectic group $SP(3,\mathbb{R})$. If $\sigma(A)=\{a\pm ib, -a\pm ib, \pm ci\}$, then

$$e^{tA}=f_1(t)I+f_2(t)A+f_3(t)A^2+f_4(t)A^3+f_5(t)A^4+f_6(t)A^5,$$

where

$$\begin{aligned}f_1(t)=&\frac{c^2(-2a^2+2b^2+c^2)}{(b^2+(a-c)^2)(b^2+(a+c)^2)}\cosh at\cos bt\\&+\frac{c^2(a^4+b^2(b^2+c^2)-a^2(6b^2+c^2))}{2ab(b^2+(a-c)^2)(b^2+(a+c)^2)}\sinh at\sin bt\\&+\frac{(a^2+b^2)^2}{2(b^2+(a-c)^2)(b^2+(a+c)^2)}(\cos ct+\sin ct),\end{aligned}$$

$$\begin{aligned}f_2(t)=&\frac{-c^2(5a^4+b^2(b^2+c^2)-a^2(10b^2+3c^2))}{2a(a^2+b^2)(b^2+(a-c)^2)(b^2+(a+c)^2)}\sinh at\cos bt\\&+\frac{c^2(a^4+5b^4+3b^2c^2-a^2(10b^2+c^2))}{2b(a^2+b^2)(b^2+(a-c)^2)(b^2+(a+c)^2)}\cosh at\sin bt\\&+\frac{(a^2+b^2)^2}{2c(b^2+(a-c)^2)(b^2+(a+c)^2)}(\cos ct-\sin ct),\end{aligned}$$

$$\begin{aligned}f_3(t)=&\frac{2(a^2-b^2)}{(b^2+(a-c)^2)(b^2+(a+c)^2)}\cosh at\cos bt\\&+\frac{-a^4+6a^2b^2-b^4+c^4}{2ab(b^2+(a-c)^2)(b^2+(a+c)^2)}\sinh at\sin bt\\&+\frac{-2(a^2-b^2)}{(b^2+(a-c)^2)(b^2+(a+c)^2)}(\cos ct+\sin ct),\end{aligned}$$

$$\begin{aligned}f_4(t)=&\frac{5a^4-10a^2b^2+b^4-c^4}{2a(a^2+b^2)(b^2+(a-c)^2)(b^2+(a+c)^2)}\sinh at\cos bt\\&+\frac{-a^4+10a^2b^2-5b^4+c^4}{2b(a^2+b^2)(b^2+(a-c)^2)(b^2+(a+c)^2)}\cosh at\sin bt\\&+\frac{-2(a^2-b^2)}{c(b^2+(a-c)^2)(b^2+(a+c)^2)}(\cos ct-\sin ct),\end{aligned}$$

$$\begin{aligned}f_5(t)=&\frac{-1}{(b^2+(a-c)^2)(b^2+(a+c)^2)}\cosh at\cos bt\\&+\frac{a^2-b^2-c^2}{2ab(b^2+(a-c)^2)(b^2+(a+c)^2)}\sinh at\sin bt\\&+\frac{1}{2(b^2+(a-c)^2)(b^2+(a+c)^2)}(\cos ct+\sin ct),\end{aligned}$$

$$f_6(t)=\frac{-3a^2+b^2+c^2}{2a(a^2+b^2)(b^2+(a-c)^2)(b^2+(a+c)^2)}\sinh at\cos bt$$
$$+\frac{a^2-3b^2-c^2}{2b(a^2+b^2)(b^2+(a-c)^2)(b^2+(a+c)^2)}\cosh at\sin bt$$
$$+\frac{1}{2c(b^2+(a-c)^2)(b^2+(a+c)^2)}(\cos ct-\sin ct).$$

ACKNOWLEDGMENTS

We would like to thank an anonymous referee for having brought up to our attention references to Putzer's method. Work supported in part by research network Contract No. ERB FMRXCT-970137 and by a Gulbenkian Foundation grant while the first author was visiting the Systems Science and Engineering Research Center at the Arizona State University.

[1] D. Sattinger and O. Weaver, ''Lie groups and algebras with applications to physics, geometry, and mechanics,'' Appl. Math. Sci. **61** (1986).
[2] D. Rowe, *Noncompact Lie Groups and Some of Their Applications*, edited by E. Tanner and R. Wilson (Kluwer Academic, New York, 1994), pp. 285–300.
[3] C. Moler and C. Van Loan, SIAM (Soc. Ind. Appl. Math.) Rev. **20**, 801–836 (1978).
[4] J. Zeni and W. Rodrigues Jr., Hadronic J. **13**, 317–323 (1990).
[5] A. Barut, J. Zeni, and A. Laufer, J. Phys. A **27**, 5239–5250 (1994).
[6] E. J. Putzer, Am. Math. Monthly **73**, 2–7 (1966).
[7] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces* (Academic, New York, 1978).
[8] N. Jacobson, *Structure and Representation of Jordan Algebras*, American Mathematical Society Coll. Pub. XXXIX (Providence, Rhode Island, 1968).
[9] T. M. Apostol, *Calculus*, 2nd ed. (Wiley, New York, 1969), Vol. II.
[10] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, Cambridge, 1991).

Second-order second-degree Painlevé equations related with Painlevé I–VI equations and Fuchsian-type transformations

U. Muğan[a)] and A. Sakka
Department of Mathematics, Bilkent University, 06533 Bilkent, Ankara, Turkey

(Received 3 February 1998; accepted for publication 2 April 1999)

One-to-one correspondence between the Painlevé I–VI equations and certain second-order second-degree equations of Painlevé type is investigated. The transformation between the Painlevé equations and second-order second-degree equations is the one involving the Fuchsian-type equation. © *1999 American Institute of Physics.* [S0022-2488(99)01507-8]

I. INTRODUCTION

Painlevé,[1] Gambier,[2] and Fuchs[3] addressed a question raised by E. Picard concerning the second-order first-degree ordinary differential equations of the form

$$v''=F(z,v,v'), \tag{1.1}$$

where F is rational in v', algebraic in v, and locally analytic in z, and have the property that all movable singularities of all solutions are poles. Movable means that the position of the singularities varies as a function of initial values. A differential equation is said to have the Painlevé property if all solutions are single valued around all movable singularities. Within the Möbius transformation, Painlevé and his school found 50 such equations. Among all these equations, 6 of them are irreducible and define classical Painlevé transcendents, PI, PII,...,PVI,[4] and the remaining 44 equations are either solvable in terms of known functions or can be transformed into one of the 6 equations. These equations maybe regarded as the nonlinear counterparts of some classical special equations. For example, PII has solution which has similar properties as Airy's functions.[5] Although the Painlevé equations were discovered from strictly mathematical considerations, they have appeared in many physical problems, and possess a rich internal structure. The properties and the solvability of the Painlevé equations have been extensively studied in the literature.[6–11]

The Riccati equation is the only example for the first-order first-degree equation which has the Painlevé property. Before the work of Painlevé and his school, Fuchs[3,4] considered the equation of the form

$$F(z,v,v')=0, \tag{1.2}$$

where F is polynomial in v and v' and locally analytic in z, such that the movable branch points are absent, that is, the generalization of the Riccati equation. Briot and Bouquet[4] considered the subcase of (1.2), that is, the first-order binomial equations of degree $m\in\mathbb{Z}_+$:

$$(v')^m+F(z,v)=0, \tag{1.3}$$

where $F(z,v)$ is a polynomial of degree at most $2m$ in v. It was found out that there are six types of equations of the form (1.3). But, all these equations are either reducible to a linear equation or solvable by means of elliptic functions.[4] Second-order binomial-type equations of degree $m\geqslant 3$,

[a)]Electronic mail: mugan@fen.bilkent.edu.tr

$$(v'')^m + F(z,v,v') = 0, \tag{1.4}$$

where F is polynomial in v and v' and locally analytic in z, were considered by Cosgrove,[12] who found out that there are nine classes. Only two of these classes can have an arbitrary degree m, and the others can have the degrees of three, four, and six. As in the case of first-order binomial-type equations, all nine classes are solvable in terms of the first, second, and fourth Painlevé transcendents, elliptic functions, or by quadratures. Chazy,[13] Garnier,[14] and Bureau[15] considered the third-order differential equations possessing the Painlevé property of the following form:

$$v''' = F(z,v,v',v''), \tag{1.5}$$

where F is assumed to be rational in v,v',v'' and locally analytic in z. But, in Ref. 15 the special form of $F(z,v,v',v'')$,

$$F(z,v,v',v'') = f_1(z,v)v'' + f_2(z,v)(v')^2 + f_3(z,v)v' + f_4(z,v), \tag{1.6}$$

where $f_k(z,v)$, $k=1,...,4$, are polynomials in v of degree k with analytic coefficients in z, was considered. In this class, no new Painlevé transcendent was discovered since, and all of them can be solved either in terms of known functions or one of the six Painlevé transcendents.

Second-order second-degree Painlevé type equations of the following form,

$$(v'')^2 = E(z,v,v')v'' + F(z,v,v'), \tag{1.7}$$

where E and F are assumed to be rational in v,v' and locally analytic in z, were subject the articles.[16,17] A special case of (1.7), given as

$$v'' = M(z,v,v') + \sqrt{N(z,v,v')}, \tag{1.8}$$

was considered in Ref. 16, where M and N are polynomials in v' of degree 2 and 4, respectively, rational in v, and locally analytic in z, and no new Painlevé transcendent was found. In Ref. 17, the special form of (1.7), $E=0$ and thus F polynomial in v and v', was considered and six distinct classes of equations denoted by SD-1,...,SD-VI, were obtained by using the α-method. Also, these classes can be solved in terms of classical Painlevé transcendents (PI,...,PVI), elliptic functions, or solutions of linear equations.

Let $v(z)$ be a solution of any of the 50 Painlevé equations, as listed by Ince,[4] each of which takes the form

$$v'' = P_2(v')^2 + P_1 v' + P_0, \tag{1.9}$$

where P_0, P_1, P_2 are functions of v, z, and a set of parameters α. The transformation, that is, Lie-point symmetry, which preserves the Painlevé property of (1.9), of the form $u(z;\hat{\alpha}) = f(v(z:\alpha),z)$ is the Möbius transformation:

$$u(z;\hat{\alpha}) = \frac{a_1(z)v + a_2(z)}{a_3(z)v + a_4(z)}, \tag{1.10}$$

where $v(z;\alpha)$ solves (1.9) with a set of parameters α and $u(z;\hat{\alpha})$ solves (1.9) with a set of parameters $\hat{\alpha}$. Lie-point symmetry can be generalized by involving $v'(z;\alpha)$, that is, the transformation of the form $u(z;\hat{\alpha}) = F(v'(z;\alpha), v(z;\alpha), z)$. The only transformation which contains v' linearly is the one involving the Riccati equation, that is,

$$u(z;\hat{\alpha}) = \frac{v' + av^2 + bv + c}{dv^2 + ev + f}, \tag{1.11}$$

where a,b,c,d,e,f are functions of z only.

In Ref. 6, the transformation of type (1.11) was used and the aim was to find a,b,c,d,e,f such that (1.11) defines a one-to-one invertible map between solutions v of (1.9) and solutions u of some second-order equations of the Painlevé type. An algorithmic method was developed to investigate the transformation properties of the Painlevé equations, and some new second-degree equations of Painlevé type related with PIII and PVI were also found. Therefore, second-degree equations are important in determining the transformation properties of the Painlevé equations.[18,6] Moreover, second-degree equations of Painlevé type appear in physics.[19–21] Furthermore, second-degree equations also appear as the first-integral of some of the third-order Painlevé-type equations.

Instead of considering the transformation of the form (1.11) one may consider the following transformation:

$$u(z;\hat{\alpha})=\frac{(v')^m+\Sigma_{j=1}^{m}P_j(z,v)(v')^{m-j}}{\Sigma_{j=1}^{m}Q_j(z,v)(v')^{m-j}}, \tag{1.12}$$

where P_j, Q_j are polynomials in v, whose coefficients are meromorphic functions of z and satisfy the Fuchs theorem[4,22] concerning the absence of the movable critical points. A second-order second-degree algebraic differential equation of the form

$$a_1(v'')^2+a_2v''v'+a_3v''v+a_4(v')^2+a_5v'v+a_6v^2=0, \tag{1.13}$$

where a_j, $j=1,2,...,6$, are meromorphic functions of z, was considered by P. Appell.[23] In Ref. 22, it was shown that Appell's condition for solvability of (1.13) is a necessary and sufficient condition for (1.13) to have its solutions free of movable branch points. Also, in Ref. 22, some analogous conditions were applied to irreducible first-order algebraic equations of the second degree, and necessary and sufficient conditions for the solutions of such equations to be free of movable branch points were obtained. A first-order algebraic differential equation of degree $n\geqslant 1$ is given as

$$a_1(z,v)(v')^n+a_2(z,v)(v')^{(n-1)}+\cdots+a_{n-1}(z,v)v'+a_n(z,v)=0, \tag{1.14}$$

where the functions $a_i(z,v)$, $i=1,...,n$, are assumed to be polynomials in v, whose coefficients are analytic functions of z. The necessary and sufficient conditions for the solutions of (1.14) to be free from movable branch points are given by the Fuchs theorem [Ref. 4 (Chap. XIII) and Ref. 22 (theorem 1.1)]. The Fuchs theorem shows that, apart from the other conditions, the irreducible form of the first-order algebraic differential equation of the second degree is

$$a_1(z)(v')^2+[a_2(z)v^2+a_3(z)v+a_4(z)]v'+[a_5(z)v^4+a_6(z)v^3+a_7(z)v^2+a_8(z)v+a_9(z)]=0, \tag{1.15}$$

where $a_i(z)$, $i=1,2,...,9$, are analytic functions of z and $a_1(z)\neq 0$. Let

$$F(v):=A_0v^4+A_1v^3+A_2v^2+A_3v+A_4, \tag{1.16}$$

where

$$A_0=4a_1a_5-a_2^2,\quad A_1=4a_1a_6-2a_2a_3,$$
$$A_2=4a_1a_7-2a_2a_4-a_3^2,\quad A_3=4a_1a_8-2a_3a_4, \tag{1.17}$$
$$A_4=4a_1a_9-a_4^2,$$

It is known that when $F(v)\neq 0$, there are unique monic polynomials $F_1(v),F_2(v)$ such that

$$F(v)\equiv A(z)F_1(v)[F_2(v)]^2, \tag{1.18}$$

where $A(z)$ is an analytic function and $F_1(z)$ has no multiple roots. In Ref. 22 it was shown (theorem 6.2) that the solutions of the equation (1.15) are free of movable branch points if and only if the following conditions hold:

$$\begin{aligned}
&(i)\ \ F_1(v) \quad \text{divides} \quad G_1(v) := (a_2v^2+a_3v+a_4)\frac{\partial F_1}{\partial v} - 2a_1\frac{\partial F_1}{\partial z},\\
&(ii)\ \ A_0=0 \quad \text{and } A_1\neq 0 \quad \text{imply } \ a_2=0,\\
&(iii)\ \ A_0=A_1=A_2=0 \quad \text{and } A_3\neq 0 \quad \text{imply } \ a_2=0.
\end{aligned} \tag{1.19}$$

The conditions of the Fuchs theorem are satisfied if and only if the conditions (1.19) are satisfied.

In this article, we investigate one-to-one correspondence between PI–PVI and some second-order second-degree Painlevé-type equations such that the transformation involving Eq. (1.15) is used and given by

$$u = \frac{(v')^2 + (a_2v^2+a_1v+a_0)v' + b_4v^4+b_3v^3+b_2v^2+b_1v+b_0}{(c_2v^2+c_1v+c_0)v' + d_4v^4+d_3v^3+d_2v^2+d_1v+d_0}, \tag{1.20}$$

where a_j, b_k, c_j, d_k, $j=0,1,2$, $k=0,1,2,3,4$, are functions of z and a set of parameters α. By using the transformations of the form (1.20), second-order second-degree Painlevé-type equations which are labeled as SD-I.a, SD-I.b, SD-I.c, SD-I.d, and SD-I.e in Ref. 17, can be obtained from PVI, PIII and PV, PIV, PII, PI, respectively. In the following sections, we first present the procedure to obtain these known equations, and for each Painlevé equation we provide an example of a second-order second-degree Painlevé-type equation that has not been considered in the literature.

The procedure used to obtain second-degree Painlevé-type equations and one-to-one correspondence with PI–PVI is as follows: Given Eq. (1.9), determine a_j, b_k, c_j, d_k, $j=0,1,2,3$, $k=0,1,2,3,4$, by requiring that (1.20) defines a one-to-one map between the solution v of (1.9) and solution u of some second-degree equation of the Painlevé type. Let $A_j := c_ju - a_j$, $B_k := d_ku - b_k$. Then the transformation (1.20) can be written as

$$(v')^2 = (A_2v^2+A_1v+A_0)v' + B_4v^4+B_3v^3+B_2v^2+B_1v+B_0. \tag{1.21}$$

It should be noted that if Eq. (1.21) is reducible, that is, if there exits a nontrivial factorization, then it can be reduced to a Riccati equation. If it is not reducible, then its solutions are free of movable branch points provided that the conditions given in (1.19) are satisfied. Differentiating Eq. (1.21) and using (1.9) to replace v'' and (1.21) to replace $(v')^2$, one gets

$$\Phi v' + \Psi = 0, \tag{1.22}$$

where

$$\begin{aligned}
\Phi &= (P_1 - 2A_2v - A_1)(A_2v^2+A_1v+A_0) + P_2(A_2v^2+A_1v+A_0)^2 + 2P_0 - 4B_4v^3 - (3B_3+A_2')v^2\\
&\quad - (2B_2+A_1')v - (B_1+A_0') + 2P_2(B_4v^4+B_3v^3+B_2v^2+B_1v+B_0),\\
\Psi &= (B_4v^4+B_3v^3+B_2v^2+B_1v+B_0)[P_2(A_2v^2+A_1v+A_0) + 2P_1 - 2A_2v - A_1]\\
&\quad - P_0(A_2v^2+A_1v+A_0) - (B_4'v^4+B_3'v^3+B_2'v^2+B_1'v+B_0').
\end{aligned} \tag{1.23}$$

There are two cases to be distinguished:

(I) $\Phi=0$: Equation (1.22) becomes

$$\Psi = 0. \tag{1.24}$$

If the solutions of the equation (1.21) are free of movable branch points, that is, the conditions given in (1.19) are satisfied, then one obtains the Painlevé-type equation of degree $n>1$ related with PI–PVI equations. To obtain the second-degree Painlevé-type equations, one should reduce the equation (1.24) to a linear equation in v. If (1.24) is reduced to an equation which is quadratic in v, then one obtains the second-order fourth-degree Painlevé-type equations related with PI–PVI, which are not considered in this article. Hence, one can find a_j, b_k, c_j, d_k such that (1.24) reduces to a linear equation in v,

$$A(u',u,z)v+B(u',u,z)=0, \tag{1.25}$$

then, substitute $v=-B/A$ into Eq. (1.21) to determine the second-order second-degree equation of the Painlevé type for u.

(II) $\Phi\neq 0$: If Φ divides Ψ, then (1.21) can be reduced to a Riccati equation and hence its solutions are free of movable branch points. Then, one can substitute $v'=-\Psi/\Phi$ in Eq. (1.21) and obtain the following equation for v:

$$\Psi^2+(A_2v^2+A_1v+A_0)\Phi\Psi-\Phi^2(B_4v^4+B_3v^3+B_2v^2+B_1v+B_0)=0. \tag{1.26}$$

Finding a_j, b_k, c_j, and d_k such that (1.26) reduces to a quadratic equation in v,

$$A(u',u,z)v^2+B(u',u,z)v+C(u',u,z)=0. \tag{1.27}$$

Solving the equation (1.27) for v and substituting into equation (1.22) yields a second-order second-degree Painlevé-type equation for u.

It turns out that PI admits transformations discussed in cases I and II, and PII–PVI admit only transformations of case II.

Second-order second-degree Painlevé-type equations were studied mainly by Bureau and Cosgrove.[16,17] But, as mentioned before, in both articles the special form of the second-degree Painlevé-type equations was considered, and no new Painlevé transcendent was found. In Refs. 24 and 25 the transformation (1.11) was used to obtain one-to-one correspondence between PI–PVI and certain second-degree Painlevé-type equations. Some of these second-degree equations had been obtained previously, but most of them had not been considered in the literature before. In this article, we investigate the transformation of type (1.20) to obtain the one-to-one correspondence between PI–VI and the second-order second-degree Painlevé-type equations. By using the transformation of type (1.11) and the procedure described above, it is possible to obtain all of the second-degree equations given in Ref. 17 except the ones which can be solvable in terms of elliptic functions or solutions of linear equations. In addition to known equations which are related with Painlevé equations through the transformation (1.20), it is possible to obtain some new second-degree equations of the Painlevé type. Since the calculations are extremely tedious, one new second-degree Painlevé-type equation for each Painlevé equation, PI–PVI, is given. Throughout this article $'$ denotes the derivative with respect to z and $\cdot$ denotes the derivative with respect to x.

II. PAINLEVÉ I

Let $v(z)$ be a solution of PI equation,

$$v''=6v^2+z. \tag{2.1}$$

Then, for PI the equation (1.22) takes the form of

$$(\phi_3v^3+\phi_2v^2+\phi_1v+\phi_0)v'+\psi_5v^5+\psi_4v^4+\psi_3v^3+\psi_2v^2+\psi_1v+\psi_0=0, \tag{2.2}$$

where

$$\phi_3=2(A_2^2+2B_4),\quad \phi_2=A_2'+3B_3+3A_1A_2-12,$$

$$\phi_1=A_1'+2B_2+A_1^2+2A_0A_2,\quad \phi_0=A_0'+B_1+A_0A_1-2z,$$

$$\psi_5=2A_2B_4,\quad \psi_4=B_4'+A_1B_4+2A_2B_3+6A_2, \tag{2.3}$$

$$\psi_3=B_3'+A_1B_3+2A_2B_2+6A_1,\quad \psi_2=B_2'+A_1B_2+2A_2B_1+6A_0+zA_2,$$

$$\psi_1=B_1'+A_1B_1+2A_2B_0+zA_1,\quad \psi_0=B_0'+A_1B_0+zA_0.$$

Case I: $\Phi=0$: One should choose $c_j=0$, $j=0,1,2$, $d_k=0$, $k=1,2,3,4$, $b_4=\frac{1}{2}a_2^2$, $b_3=-\frac{1}{3}a_2'+a_1a_2-4$, $b_2=-\frac{1}{2}a_1'+\frac{1}{2}a_1^2+a_0a_2$, $b_1=-a_0'+a_0a_1-2z$. One can always absorb b_0 and d_0 in u by a proper Möbius transformation. Hence, without loss of generality, one can set $b_0=0$ and $d_0=2$. The only possibility to reduce the equation $\Psi=0$ to a linear equation in v is to set $\psi_5=\psi_4=\psi_3=\psi_2=0$. Therefore, one obtains $a_2=a_1=a_0=b_4=b_2=0$, $b_3=-4$, and $b_1=-2z$. Then the equation (1.20) becomes

$$2u=(v')^2-4v^3-2zv, \tag{2.4}$$

and the linear equation for v reads

$$v+u'=0. \tag{2.5}$$

Equation (2.4) with the condition (2.5) satisfies corollary 6.3 in Ref. 22, and hence its solutions are free of movable branch points. By following the procedure discussed in the Introduction, one can get the following second-order second-degree equation for $u(z)$:

$$(u'')^2=-4(u')^3-2(zu'-u). \tag{2.6}$$

Equation (2.6) was first obtained by Cosgrove[17] and labeled as SD-I.e.

Case II: $\Phi\neq 0$: As an example, let $\phi_i=0$, $i=1,2,3$, $\phi_0\neq 0$, and $\psi_l=0$, $l=2,3,4,5$. These choices imply that $A_j=0$, $j=0,1,2$, $B_4=B_2=0$, and $B_3=4$. Then, Eq. (1.26) becomes

$$(B_1'v+B_0')^2-(B_1-2z)^2(4v^3+B_1v+B_0)=0. \tag{2.7}$$

To reduce the equation (2.7) to a quadratic equation for v, one has to take $d_1\neq 0$ and, hence, without loss of generality, $b_1=0$ and $d_1=1$. Moreover, d_0 and b_0 are the solutions of the following equations:

$$d_0'(b_0'-2zd_0')=0,\quad (d_0')^2+4d_0^3+b_0=0,\quad (b_0')^2-4z^2(d_0')^2=0. \tag{2.8}$$

Here, we only consider the case $d_0'=0$; then $d_0=\mu$ and $b_0=-4\mu^3$, where μ is a constant. Therefore, the equations (1.21) and (1.22) become

$$(v')^2=4v^3+uv+\mu(u+4\mu^2) \tag{2.9}$$

and

$$v'=\frac{-u'}{(u-2z)}(v+\mu), \tag{2.10}$$

respectively, and the quadratic equation for v takes the form of

$$4(u-2z)^2v^2-[(u')^2+4\mu(u-2z)^2]v-\mu(u')^2+(u+4\mu^2)(u-2z)^2=0. \tag{2.11}$$

Let $u(z)=-2(e^x/(y-1)+6\mu^2)$ and $z=e^x-6\mu^2$. Then the equations (2.9) and (2.11) give one-to-one correspondence between solutions $v(z)$ of PI and solutions $y(x)$ of the following second-order second-degree Painlevé-type equation

$$\{4y(y-1)(\ddot{y}-\dot{y})-(\dot{y}-y+1)[(7y-4)\dot{y}+5y(y-1)]+12\mu e^{2x}y^3(y-1)^2\}^2$$
$$=(y+2)^2\{[(\dot{y}-y+1)^2+12\mu e^{2x}y^2(y-1)^2]^2+32e^{5x}y^4(y-1)^3\}. \tag{2.12}$$

III. PAINLEVÉ II

Let $v(z)$ be a solution of PII equation

$$v''=2v^3+zv+\alpha. \tag{3.1}$$

Then, for PII, the equation (1.22) takes the following form:

$$(\phi_3v^3+\phi_2v^2+\phi_1v+\phi_0)v'+\psi_5v^5+\psi_4v^4+\psi_3v^3+\psi_2v^2+\psi_1v+\psi_0=0, \tag{3.2}$$

where

$$\phi_3=4(B_4+\tfrac{1}{2}A_2^2-1), \quad \phi_2=A_2'+3B_3+3A_1A_2,$$
$$\phi_1=A_1'+2B_2+2A_0A_2+A_1^2-2z, \quad \phi_0=A_0'+B_1+A_0A_1-2\alpha,$$
$$\psi_5=2A_2(B_4+1), \quad \psi_4=B_4'+A_1B_4+2A_2B_3+2A_1, \tag{3.3}$$
$$\psi_3=B_3'+A_1B_3+2A_2B_2+2A_0+zA_2, \quad \psi_2=B_2'+A_1B_2+2A_2B_1+zA_1+\alpha A_2,$$
$$\psi_1=B_1'+A_1B_1+2A_2B_0+zA_0+\alpha A_1, \quad \psi_0=B_0'+A_1B_0+\alpha A_0.$$

Here, we only consider the case $\phi_i=0$, $i=1,2,3$, $\phi_0\neq 0$, and $\psi_l=0$, $l=3,4,5$. $\psi_5=0$ implies that either $A_2=0$ or $B_4=-1$.

Case i: If $A_2=0$, then one obtains $A_1=A_0=0$, $B_4=1$, $B_3=0$, $B_2=z$ and $\phi_0=B_1-2\alpha$, $\psi_2=1$, $\psi_1=B_1'$, $\psi_0=B_0'$. With these choices, the equation (1.26) yields

$$(v^2+B_1'v+B_0')^2-(B_1-2\alpha)^2(v^4+zv^2+B_1v+B_0)=0. \tag{3.4}$$

To reduce the equation (3.4) to a quadratic equation in v, one possibility is to set the coefficients of v^4 and v^3 to zero. Then, one obtains $B_1=2\alpha+\epsilon$, where $\epsilon=\pm 1$, and by using the proper Möbius transformation, one may take $B_0=2u+\frac{1}{4}z^2$. Hence, the equations (1.21) and (1.22) become

$$(v')^2=v^4+zv^2+(2\alpha+\epsilon)v+2u+\frac{z^2}{4} \tag{3.5}$$

and

$$v'=\epsilon\left(v^2+2u'+\frac{z}{2}\right), \tag{3.6}$$

respectively. The quadratic equation in v is

$$4u'v^2-(2\alpha+\epsilon)v+4(u')^2+2(zu'-u)=0. \tag{3.7}$$

The equations (3.5) and (3.7) give one-to-one correspondence between solutions $v(z)$ of PII and solutions $u(z)$ of the equation

$$(u'')^2=-4(u')^3-2u'(zu'-u)+\tfrac{1}{16}(2\alpha+\epsilon)^2. \tag{3.8}$$

The equation (3.8) was first obtained by Cosgrove[17] and labeled as SD-I.d.

Case ii: If $B_4=-1$, then one obtains $A_2=2\epsilon$, $A_1=0$, $A_0=\epsilon z$, $B_3=0$, $B_2=-z$ and $\phi_0=B_1-2\alpha+\epsilon\neq 0$, $\psi_2=4\epsilon B_1+2\epsilon\alpha-1$, $\psi_1=B_1'+4\epsilon B_0+\epsilon z^2$, $\psi_0=B_0'+\epsilon\alpha z$, where $\epsilon=\pm 1$. Then the equation (1.26) becomes

$$(\psi_2 v^2+\psi_1 v+\psi_0)^2+2\epsilon\phi_0(v^2+\tfrac{1}{2}z)(\psi_2 v^2+\psi_1 v+\psi_0)+\phi_0^2(v^4+zv^2-B_1 v-B_0)=0. \tag{3.9}$$

To reduce the equation (3.9) to a quadratic equation in v one may set the coefficients of v^4 and v^3 to zero. Thus one obtains $B_1=0$, and without loss of generality one may take $B_0=\frac{1}{4}(u-z^2)$. Therefore, the equations (1.21) and (1.22) give

$$(v')^2=\epsilon(2v^2+z)v'-v^4-zv^2+\tfrac{1}{4}(u-z^2), \tag{3.10}$$

and

$$v'=\frac{\epsilon}{(2\alpha-\epsilon)}\left[(2\alpha-\epsilon)v^2+uv+\frac{\epsilon}{4}u'+\frac{1}{2}(2\alpha-\epsilon)z\right], \tag{3.11}$$

respectively, and the quadratic equation in v is

$$(4uv+\epsilon u')^2=4(2\alpha-\epsilon)^2 u. \tag{3.12}$$

The equations (3.10) and (3.12) give one-to-one correspondence between solutions $v(z)$ of PII and solutions $u(z)$ of the following second-order second-degree Painlevé-type equation:

$$[4uu''-3(u')^2+8zu^2+4(2\alpha-\epsilon)^2 u]^2=64u^5. \tag{3.13}$$

IV. PAINLEVÉ III

Let $v(z)$ be a solution of PIII equation

$$v''=\frac{1}{v}(v')^2-\frac{1}{z}v'+\gamma v^3+\frac{1}{z}(\alpha v^2+\beta)+\frac{\delta}{v}. \tag{4.1}$$

Then, for PIII, the equation (1.22) takes the following form:

$$(\phi_4 v^4+\phi_3 v^3+\phi_2 v^2+\phi_1 v+\phi_0)v'+\psi_6 v^6+\psi_5 v^5+\psi_4 v^4+\psi_3 v^3+\psi_2 v^2+\psi_1 v+\psi_0=0, \tag{4.2}$$

where

$$\phi_4=2\gamma-2B_4-A_2^2,\quad \phi_3=\frac{2\alpha}{z}-B_3-A_1A_2-A_2'-\frac{1}{z}A_2,\quad \phi_2=-\left(A_1'+\frac{1}{z}A_1\right),$$

$$\phi_1=\frac{2\beta}{z}+B_1+A_0A_1-A_0'-\frac{1}{z}A_0,\quad \phi_0=A_0^2+2B_0+2\delta,$$

$$\psi_6=-A_2(B_4+\gamma),\quad \psi_5=-\left(B_4'+\frac{2}{z}B_4+A_2B_3+\gamma A_1+\frac{\alpha}{z}A_2\right),$$

$$\psi_4=A_0B_4-B_3'-\frac{2}{z}B_3-A_2B_2-\gamma A_0-\frac{\alpha}{z}A_1, \tag{4.3}$$

$$\psi_3 = A_0 B_3 - B_2' - \frac{2}{z} B_2 - A_2 B_1 - \frac{\beta}{z} A_2 - \frac{\alpha}{z} A_0,$$

$$\psi_2 = A_0 B_2 - B_1' - \frac{2}{z} B_1 - A_2 B_0 - \frac{\beta}{z} A_1 - \delta A_2,$$

$$\psi_1 = A_0 B_1 - B_0' - \frac{2}{z} B_0 - \frac{\beta}{z} A_0 - \delta A_1, \quad \psi_0 = A_0 (B_0 - \delta).$$

As an example, let $A_0 = 0$, $A_1 = 2/z$, $B_1 = -2\beta/z$, and $B_0 = -\delta$. Then one gets $\phi_0 = \phi_1 = \phi_2 = 0$ and $\psi_0 = \psi_1 = \psi_2 = 0$. Moreover, if $\phi_4 = \psi_6 = 0$, then either $A_2 = 0$, $B_4 = \gamma$ or $A_2 = 2\sqrt{\gamma}$, $B_4 = -\gamma$, where γ can be taken with either sign.

Case i: If $A_2 = 0$, $B_4 = \gamma$, then equation (1.26) takes the following form:

$$(\psi_5 v^2 + \psi_4 v + \psi_3)^2 + \frac{2}{z} \phi_3 v (\psi_5 v^2 + \psi_4 v + \psi_3) - \phi_3^2 \left(\gamma v^4 + B_3 v^3 + B_2 v^2 - \frac{2\beta}{z} v - \delta \right) = 0. \tag{4.4}$$

To reduce the equation (4.4) to a quadratic equation in v one may set the coefficients of v^4 and v^3 to zero. Then, one obtains $B_3 = (2/z)(\alpha + 2\sqrt{\gamma})$, and without loss of generality, one may take $B_2 = u$. With these choices the quadratic equation in v takes the following form,

$$\begin{aligned} &8[\gamma z^3 u' + 2(\alpha + \sqrt{\gamma})(\alpha + 3\sqrt{\gamma})] v^2 + 8[(\alpha + 2\sqrt{\gamma})(zu' + u) + 4\gamma\beta] v \\ &\quad + z^2 (zu' + 2u)^2 + 16\gamma\delta z^2 = 0, \end{aligned} \tag{4.5}$$

and the transformations (1.21) and (1.22) become

$$(v')^2 = \frac{2}{z} v v' + \gamma v^4 + \frac{2}{z} (\alpha + 2\sqrt{\gamma}) v^3 + u v^2 - \frac{2\beta}{z} v - \delta \tag{4.6}$$

and

$$v' = \frac{-1}{4z\sqrt{\gamma}} [4\gamma z v^2 + 4(\alpha + \sqrt{\gamma}) v + z^2 u' + 2zu], \tag{4.7}$$

respectively. Then, the transformations (4.5) and (4.6) give one-to-one correspondence between solutions $v(z)$ of PIII and solutions $y(x)$ of the following second-order second-degree Painlevé-type equation

$$x^2 (\ddot{y})^2 = -4(\dot{y})^2 (x\dot{y} - y) - \frac{\gamma\delta}{16} (x\dot{y} - y) + \frac{\beta}{16} (\alpha + 2\sqrt{\gamma}) \dot{y} + \frac{1}{256} [\gamma\beta^2 - \delta(\alpha + 2\sqrt{\gamma})^2], \tag{4.8}$$

where $y(x) = \frac{1}{16}[z^2 u(z) + 1]$ and $x = z^2$. The equation (4.8) was first obtained by Cosgrove[17] and labeled as SD-I.b (with $A_1 = 0$).

Case ii: $A_2 = 2\sqrt{\gamma}$, $B_4 = -\gamma$: The equation (1.26) takes the form of

$$\begin{aligned} &(\psi_5 v^2 + \psi_4 v + \psi_3)^2 + \frac{2}{z} \phi_3 v (\sqrt{\gamma} z v + 1)(\psi_5 v^2 + \psi_4 v + \psi_3) \\ &\quad + \phi_3^2 \left(\gamma v^4 - B_3 v^3 - B_2 v^2 + \frac{2\beta}{z} v + \delta \right) = 0. \end{aligned} \tag{4.9}$$

One may set the coefficients of v^4 and v^3 to zero in order to reduce the equation (4.9) to a quadratic equation in v. Then, one obtains $B_3=-2\sqrt{\gamma}/z$, and without loss of generality one can take $B_2=u$. Then, the equation (1.21) becomes

$$(v')^2=\frac{2}{z}v(\sqrt{\gamma}zv+1)v'-\gamma v^4-\frac{2\sqrt{\gamma}}{z}v^3+uv^2-\frac{2\beta}{z}v-\delta. \tag{4.10}$$

By using the linear transformation $y(x)=z^2u(z)+1$, $2x=z^2$, and $\mu=\alpha-2\sqrt{\gamma}$, the equation (1.22) can be written as

$$v'=\sqrt{\gamma}v^2+\frac{1}{z}\left(\frac{\sqrt{\gamma}}{\mu}y+1\right)v+\frac{1}{2\mu}(\dot{y}-2\beta\sqrt{\gamma}), \tag{4.11}$$

and the quadratic equation for v is

$$4y(\gamma y-\mu^2)v^2+4z[\sqrt{\gamma}y(\dot{y}-2\beta\sqrt{\gamma})+2\beta\mu^2]v+z^2[(\dot{y}-2\beta\sqrt{\gamma})^2+4\delta\mu^2]=0. \tag{4.12}$$

The equations (4.10) and (4.12) give one-to-one correspondence between solutions $v(z)$ of PIII and solutions $y(x)$ of the following equation:

$$x^2[2y^2\ddot{y}-y\dot{y}^2-4(\delta\mu^2-\gamma\beta^2)y-8\beta^2\mu^2]^2=(y^2+4\beta\mu x)^2[y(\dot{y})^2-4(\gamma y-\mu^2)(\delta y+\beta^2)]. \tag{4.13}$$

V. PAINLEVÉ IV

Let $v(z)$ be a solution of PIV

$$v''=\frac{1}{2v}(v')^2+\frac{3}{2}v^3+4zv^2+2(z^2-\alpha)v+\frac{\beta}{v}. \tag{5.1}$$

Then, for PIV the equation (1.22) takes the following form,

$$(\phi_4v^4+\phi_3v^3+\phi_2v^2+\phi_1v+\phi_0)v'+\psi_6v^6+\psi_5v^5+\psi_4v^4+\psi_3v^3+\psi_2v^2+\psi_1v+\psi_0=0, \tag{5.2}$$

where

$$\begin{gathered}
\phi_4=3(1-B_4-\tfrac{1}{2}A_2^2),\quad \phi_3=8z-2B_3-2A_1A_2-A_2',\\
\phi_2=4(z^2-\alpha)-B_2-\tfrac{1}{2}A_1^2-A_0A_2-A_1',\quad \phi_1=-A_0',\quad \phi_0=\tfrac{1}{2}A_0^2+B_0+2\beta,\\
\psi_6=-\tfrac{3}{2}A_2(B_4+1),\quad \psi_5=-(B_4'+\tfrac{1}{2}A_1B_4+\tfrac{3}{2}A_2B_3+4zA_2+\tfrac{3}{2}A_1),\\
\psi_4=\tfrac{1}{2}A_0B_4-B_3'-\tfrac{1}{2}A_1B_3-\tfrac{3}{2}A_2B_2-2(z^2-\alpha)A_2-\tfrac{3}{2}A_0-4zA_1,\\
\psi_3=\tfrac{1}{2}A_0B_3-B_2'-\tfrac{1}{2}A_1B_2-\tfrac{3}{2}A_2B_1-2(z^2-\alpha)A_1-4zA_0,\\
\psi_2=\tfrac{1}{2}A_0B_2-B_1'-\tfrac{1}{2}A_1B_1-\tfrac{3}{2}A_2B_0-\beta A_2-2(z^2-\alpha)A_0,\\
\psi_1=\tfrac{1}{2}A_0B_1-B_0'-\tfrac{1}{2}A_1B_0-\beta A_1,\quad \psi_0=\tfrac{1}{2}A_0(B_0-2\beta).
\end{gathered} \tag{5.3}$$

As an example, let $A_0=0$ and $B_0=-2\beta$. Then one gets $\phi_0=\phi_1=\psi_0=\psi_1=0$. Moreover, setting $\phi_4=\phi_3=\psi_6=\psi_5=0$, one has the following two distinct cases: (i) $A_2=0$, $A_1=0$, $B_4=1$, $B_3=4z$ or (ii) $A_2=2\epsilon$, $A_1=4\epsilon z$, $B_4=-1$, $B_3=-4z$, where $\epsilon=\pm1$.

Case i: In this case Eq. (1.26) takes the form of

$$(\psi_4 v^2+\psi_3 v+\psi_2)^2-\phi_2^2(v^4+4zv^3+B_2v^2+B_1v-2\beta)=0. \tag{5.4}$$

To reduce the equation (5.4) to a quadratic equation in v, one may set the coefficients of v^4 and v^3 to zero. Then, one obtains $B_2=4(z^2-\alpha+\epsilon)$ and, hence, without loss of generality, one can choose $B_1=u$. Then the equations (1.21) and (1.22) become

$$(v')^2=v^4+4zv^3+4(z^2-\alpha+\epsilon)v^2+uv-2\beta \tag{5.5}$$

and

$$v'=\frac{-\epsilon}{4}(4v^2+8zv+u'), \tag{5.6}$$

respectively. The equations (5.5) and

$$8[u'+8(\alpha-\epsilon)]v^2+16(zu'-u)v+(u')^2+32\beta=0 \tag{5.7}$$

give one-to-one correspondence between solutions $v(z)$ of PIV and solutions $u(z)$ of the following equation:

$$(u'')^2=4(zu'-u)^2-\tfrac{1}{2}[(u')^2+32\beta](u'+8\alpha-8\epsilon). \tag{5.8}$$

The transformation $u=8(y-\mu z)$, where $\mu=\frac{1}{3}(\alpha-\epsilon)$, transforms the equation (5.8) to the following equation,

$$(y'')^2=-4(y')^3+4(zy'-y)^2+2(6\mu^2-\beta)y'-4\mu(2\mu^2+\beta), \tag{5.9}$$

which was first obtained by Cosgrove[17] and labeled as SD-I.c.

Case ii: In this case Eq. (1.26) can be written as follows:

$$[(\psi_4+\epsilon\phi_2)v^2+(\psi_3+2\epsilon z\phi_2)v+\psi_2]^2=\phi_2^2[(B_2+4z^2)v^2+B_1v-2\beta]. \tag{5.10}$$

It is clear that if one sets $\psi_4+\epsilon\phi_2=0$, then the equation (5.10) reduces to a quadratic equation in v. Thus, one should take $B_2=-4z^2$ and, without loss of generality, $B_1=u$. Then, the equations (1.21) and (1.22) become, respectively,

$$(v')^2=2\epsilon v(v+2z)v'-v^4-4zv^3-4z^2v^2+uv-2\beta \tag{5.11}$$

and

$$v'=\frac{\epsilon}{12\mu}[12\mu v^2-3(u-8\mu z)v-(\epsilon u'+2zu-4\beta)], \tag{5.12}$$

where $\mu=\frac{1}{3}(\alpha+\epsilon)$. The equations (5.11) and

$$9u^2v^2+2u(3\epsilon u'+6zu-12\beta-72\mu^2)v+(\epsilon u'+2zu-4\beta)^2+288\beta\mu^2=0 \tag{5.13}$$

give one-to-one correspondence between solutions $v(z)$ of PIV and solutions $u(z)$ of the following second-order second-degree Painlevé-type equation:

$$\begin{aligned}&[3uu''-2(u')^2-2\epsilon(zu-2\beta+12\mu^2)u'-2(4z^2-3\epsilon)u^2-8(6\mu^2-\beta)zu+16(6\mu^2-\beta)^2]^2\\&\quad=-27[u^2-16\mu(2\mu^2+\beta)]^2[\epsilon u'+2zu+2\beta-12\mu^2].\end{aligned} \tag{5.14}$$

VI. PAINLEVÉ V

Let $v(z)$ be a solution of PV:

$$v''=\frac{3v-1}{2v(v-1)}(v')^2-\frac{1}{z}v'+\frac{\alpha}{z^2}v(v-1)^2+\frac{\beta(v-1)^2}{z^2v}+\frac{\gamma}{z}v+\frac{\delta v(v+1)}{v-1}. \tag{6.1}$$

Then, for PV, Eq. (1.22) takes the form of

$$(\phi_5v^5+\phi_4v^4+\phi_3v^3+\phi_2v^2+\phi_1v+\phi_0)v'+\psi_7v^7+\psi_6v^6$$
$$+\psi_5v^5+\psi_4v^4+\psi_3v^3+\psi_2v^2+\psi_1v+\psi_0=0, \tag{6.2}$$

where

$$\phi_5=\frac{2\alpha}{z^2}-B_4-\frac{1}{2}A_2^2,\quad \phi_4=3B_4+\frac{3}{2}A_2^2-\frac{6\alpha}{z^2}-A_2'-\frac{1}{z}A_2,$$

$$\phi_3=2B_3+B_2+\frac{1}{2}A_1^2+2A_1A_2+A_0A_2+A_2'+\frac{1}{z}A_2-A_1'-\frac{1}{z}A_1+\frac{2}{z^2}[3\alpha+\beta+\gamma z+\delta z^2],$$

$$\phi_2=2B_1+B_2+\frac{1}{2}A_1^2+2A_0A_1+A_0A_2+A_1'+\frac{1}{z}A_1-A_0'-\frac{1}{z}A_0-\frac{2}{z^2}[\alpha+3\beta+\gamma z-\delta z^2],$$

$$\phi_1=3B_0+\frac{3}{2}A_0^2+\frac{6\beta}{z^2}+A_0'+\frac{1}{z}A_0,\quad \phi_0=-\left(\frac{2\beta}{z^2}+B_0+\frac{1}{2}A_0^2\right),$$

$$\psi_7=-\frac{1}{2}A_2\left(B_4+\frac{2\alpha}{z^2}\right),\quad \psi_6=B_4\left(\frac{3}{2}A_2+\frac{1}{2}A_1-\frac{2}{z}\right)-\frac{1}{2}A_2B_3+\frac{\alpha}{z^2}(3A_2-A_1)-B_4', \tag{6.3}$$

$$\psi_5=B_4\left(\frac{1}{2}A_1+\frac{3}{2}A_0+\frac{2}{z}\right)+B_3\left(\frac{3}{2}A_2+\frac{1}{2}A_1-\frac{2}{z}\right)-\frac{1}{2}A_2B_2$$
$$-\frac{A_2}{z^2}(3\alpha+\beta+\gamma z+\delta z^2)+\frac{\alpha}{z^2}(3A_1-A_0)+B_4'-B_3',$$

$$\psi_4=B_3\left(\frac{1}{2}A_1+\frac{3}{2}A_0+\frac{2}{z}\right)+B_2\left(\frac{3}{2}A_2+\frac{1}{2}A_1-\frac{2}{z}\right)-\frac{1}{2}A_2B_1-\frac{1}{2}A_0B_4$$
$$-\frac{A_1}{z^2}(3\alpha+\beta+\gamma z+\delta z^2)+\frac{A_2}{z^2}(\alpha+3\beta+\gamma z-\delta z^2)+\frac{3\alpha}{z^2}A_0+B_3'-B_2',$$

$$\psi_3=B_2\left(\frac{1}{2}A_1+\frac{3}{2}A_0+\frac{2}{z}\right)+B_1\left(\frac{3}{2}A_2+\frac{1}{2}A_1-\frac{2}{z}\right)-\frac{1}{2}A_2B_0-\frac{1}{2}A_0B_3$$
$$-\frac{A_0}{z^2}(3\alpha+\beta+\gamma z+\delta z^2)+\frac{A_1}{z^2}(\alpha+3\beta+\gamma z-\delta z^2)-\frac{3\beta}{z^2}A_2+B_2'-B_1',$$

$$\psi_2=B_1\left(\frac{1}{2}A_1+\frac{3}{2}A_0+\frac{2}{z}\right)+B_0\left(\frac{3}{2}A_2+\frac{1}{2}A_1-\frac{2}{z}\right)$$
$$+\frac{A_0}{z^2}(\alpha+3\beta+\gamma z-\delta z^2)+\frac{\beta}{z^2}(A_2-3A_1)+B_1'-B_0',$$

$$\psi_1 = B_0\left(\frac{3}{2}A_0 + \frac{1}{2}A_1 + \frac{2}{z}\right) - \frac{1}{2}A_0B_1 + \frac{\beta}{z^2}(A_1 - 3A_0) + B_0', \quad \psi_0 = \frac{-1}{2}A_0\left(B_0 - \frac{2\beta}{z^2}\right).$$

As an example, let

$$A_1 = \frac{-2}{z}(zA_0 + 1), \quad A_2 = \frac{1}{z}(zA_0 + 2),$$
$$B_3 = -\left(2B_2 + 3B_1 + 4B_0 - \frac{2\gamma}{z} + 4\delta\right), \quad B_4 = B_2 + 2B_1 + 3B_0 - \frac{2\gamma}{z} + 2\delta, \tag{6.4}$$

and let $\phi_0 = \phi_1 = \psi_0 = \psi_1 = 0$. Then, Eq. (7.2) can be written as

$$\phi_5 v' + \psi_7 v^2 + (\psi_6 + 3\psi_7)v - \psi_2 = 0, \tag{6.5}$$

and the equation (1.26) can be written as

$$\begin{aligned}[&\psi_7 v^2 + (\psi_6 + 3\psi_7 v - \psi_2 + \tfrac{1}{2}\phi_5(A_2 v^2 + A_1 v + A_0)]^2 \\ &= \phi_5^2[(B_4 + \tfrac{1}{4}A_2^2)v^4 + (B_3 + \tfrac{1}{2}A_1A_2)v^3 + (\tfrac{1}{4}B_2A_1^2 + \tfrac{1}{2}A_0A_2)v^2 \\ &\quad + (B_1 + \tfrac{1}{2}A_0A_1)v + (B_0 + \tfrac{1}{4}A_0^2)].\end{aligned} \tag{6.6}$$

Here $\psi_0 = 0$ implies that either $A_0 = 0$ or $B_0 = 2\beta/z^2$.

Case i: $A_0 = 0$: Equations (6.4) and $\phi_0 = \phi_1 = \psi_1 = 0$ imply that $A_1 = -2/z$, $A_2 = 2/z$, and $B_0 = -2\beta/z^2$. If $B_4 = (\mu^2 - 1)/z^2$, where $\mu = 1 - \sqrt{2\alpha}$ and $\sqrt{2\alpha}$ can take either sign, and without loss of generality $B_1 = (1/z^2)(4u + \gamma z - \mu^2 + 6\beta)$, then Eq. (6.6) reduces to the following quadratic equation for v:

$$Av^2 + Bv + C = 0, \tag{6.7}$$

where

$$\begin{aligned} A &= 8\mu^2[2(zu' + u) + \delta z^2 - \mu^2 + 2\beta] + (4u - \gamma z - 3\mu^2 + 2\beta)^2, \\ B &= 2(4u - \gamma z - 3\mu^2 + 2\beta)[4(zu' - u) + \mu^2 - 2\beta] - 4\mu^2(4u + \gamma z - \mu^2 + 6\beta), \\ C &= [4(zu' - u) + \mu^2 - 2\beta]^2 + 8\beta\mu^2. \end{aligned} \tag{6.8}$$

The equations (1.21) and (1.22) respectively become

$$\begin{aligned}(v')^2 = \frac{1}{z^2}[&2zv(v-1)v' + (\mu^2 - 1)v^4 + (4u - \gamma z - 3\mu^2 + 2\beta + 2)v^3 \\ &- (8u + 2\delta z^2 - 3\mu^2 + 6\beta + 1)v^2 + (4u + \gamma z - \mu^2 + 6\beta)v - 2\beta],\end{aligned} \tag{6.9}$$

and

$$v' = \frac{1}{2\mu z}[2\mu\sqrt{2\alpha}v^2 - (4u - \gamma z - 3\mu^2 + 2\beta + 2\mu)v - (4zu' - 4u + \mu^2 - 2\beta)]. \tag{6.10}$$

The equations (6.9) and (6.7) define one-to-one correspondence between solutions $v(z)$ of PV and solutions $u(z)$ of the following second-order second-degree Painlevé-type equation:

$$z^2(u'')^2=-4(u')^2(zu'-u)-2\delta(zu'-u)^2-[\delta(\mu^2-2\beta)-\tfrac{1}{4}\gamma^2](zu'-u)$$
$$+\tfrac{1}{2}\gamma(\mu^2+2\beta)u'+\tfrac{1}{8}[\gamma^2(\mu^2-2\beta)-\delta(\mu^2+2\beta)^2]. \tag{6.11}$$

The equation (6.11) was first obtained by Cosgrov[17] and labeled as SD-I.b.

Case ii: $B_0=2\beta/z^2$: Then $A_0=2(\mu-1)/z$, $B_1=-\frac{1}{2}A_0A_1$, where $(\mu-1)^2=-2\beta$. With out loss of generality, let $B_2=(1/z^2)[u-6(\mu-1)^2-6(\mu-1)-1+2\gamma z-2\delta z^2]$. Then Eq. (6.4) implies that $B_4=(1/z^2)(u-\mu^2)$, $B_3=(-2/z^2)[u+\gamma z-\mu(2\mu-1)]$. With these choices, the equation (6.6) becomes

$$Av^2+Bv+C=0, \tag{6.12}$$

where

$$A=u[u^2-2(\mu^2+2\alpha)u+(\mu^2-2\alpha)^2],$$
$$B=-4\mu zuu'-2(u+\gamma z)[u^2-2(\mu^2+2\alpha)u+(\mu^2-2\alpha)^2], \tag{6.13}$$
$$C=-[zu'-2\mu(u+\gamma z)]^2+(u-2\delta z^2+2\gamma z)(u+\mu^2-2\alpha)^2.$$

The equation (1.21) can be written as follows,

$$[zv'-(v-1)(\mu v-\mu-1)]^2=uv^4-2(u+\gamma z)v^3+(u-2\delta z^2+2\gamma z)v^2, \tag{6.14}$$

and the equation (1.22) becomes

$$v'=-\frac{1}{z(u+\mu^2-2\alpha)}\{\mu(u-\mu^2+2\alpha)v^2$$
$$+[zu'-u-2\gamma\mu z+(2\mu-1)(\mu^2-2\alpha)]v-(\mu-1)(u+\mu^2-2\alpha)\}. \tag{6.15}$$

Let $u(z)$ be a solution of the following second-order second-degree equation of Painlevé type:

$$[2uu''-(u')^2+2\delta u^2+2\gamma u-2\delta(\mu^2-2\alpha)^2-2\gamma^2(\mu^2+2\alpha)]^2$$
$$=8[u^2-\gamma z(\mu^2-2\alpha)^2]^2\{u(u')^2+(2\delta u+\gamma^2)$$
$$\times[u^2-2(\mu^2+2\alpha)u+(\mu^2-2\alpha)^2]\}. \tag{6.16}$$

Then Eqs. (6.12) and (6.14) give one-to-one correspondence between solutions $v(z)$ of PV and $u(z)$ of the equation (6.16).

VII. PAINLEVÉ VI

Let $v(z)$ be a solution of PVI

$$v''=\frac{1}{2}\left(\frac{1}{v}+\frac{1}{v-1}+\frac{1}{v-z}\right)(v')^2-\left(\frac{1}{z}+\frac{1}{z-1}+\frac{1}{v-z}\right)v'$$
$$+\frac{v(v-1)(v-z)}{z^2(z-1)^2}\left(\alpha+\frac{\beta z}{v^2}+\frac{\gamma(z-1)}{(v-1)^2}+\frac{\delta z(z-1)}{(v-z)^2}\right). \tag{7.1}$$

Then, for PVI, Eq. (1.22) takes the form of

$$(\phi_6v^6+\phi_5v^5+\phi_4v^4+\phi_3v^3+\phi_2v^2+\phi_1v+\phi_0)v'+\psi_8v^8$$
$$+\psi_7v^7+\psi_6v^6+\psi_5v^5+\psi_4v^4+\psi_3v^3+\psi_2v^2+\psi_1v+\psi_0=0, \tag{7.2}$$

where

$$\phi_6=\frac{2\alpha}{z^2(z-1)^2}-B_4-\frac{1}{2}A_2^2,\quad \phi_5=2(z+1)B_4+(z+1)A_2^2-\frac{4\alpha(z+1)}{z^2(z-1)^2}-A_2'-\frac{(2z-1)}{z(z-1)}A_2,$$

$$\phi_4=\frac{z}{(z-1)}A_2-\frac{(2z-1)}{z(z-1)}(A_1-A_2)+\frac{1}{2}A_1^2+A_0A_2+(z+1)A_1A_2-\frac{3}{2}zA_2^2+B_2+(z+1)B_3$$
$$-3zB_4+(z+1)A_2'-A_1'+\frac{2}{z^2(z-1)^2}[\alpha(z^2+4z+1)+\beta z+(\delta z+\gamma)(z-1)],$$

$$\phi_3=\frac{z}{(z-1)}(A_1-A_2)-\frac{(2z-1)}{z(z-1)}(A_0-A_1)+2A_0A_1-2zA_1A_2+2B_1-2zB_3$$
$$+(z+1)A_1'-A_0'-zA_2'-\frac{4}{z(z-1)^2}[(\alpha+\beta)(z+1)+(\gamma+\delta)(z-1)],$$

$$\phi_2=\frac{z}{(z-1)}(A_0-A_1)+\frac{(2z-1)}{z(z-1)}A_0+\frac{3}{2}A_0^2-(z+1)A_0A_1-zA_0A_2-\frac{1}{2}zA_1^2+3B_0-(z+1)B_1$$
$$-zB_2-zA_1'+(z+1)A_0'-\frac{2}{z(z-1)^2}[\alpha z+\beta(z^2+4z+1)+(\gamma z+\delta)(z-1)],$$

$$\phi_1=-\left[2(z+1)B_0+(z+1)A_0^2+\frac{4\beta(z+1)}{(z-1)^2}+zA_0'+\frac{z}{(z-1)}A_0\right],$$

$$\phi_0=z\left[B_0+\frac{1}{2}A_0^2+\frac{2\beta}{(z-1)^2}\right],$$

$$\psi_8=-\frac{1}{2}A_2\left[B_4+\frac{2\alpha}{z^2(z-1)^2}\right],$$

$$\psi_7=B_4\left[(z+1)A_2+\frac{1}{2}A_1-\frac{(2z-1)}{z(z-1)}\right]-\frac{1}{2}A_2B_3+\frac{\alpha}{z^2(z-1)^2}[2(z+1)A_2-A_1]-B_4',\quad (7.3)$$

$$\psi_6=B_4\left[\frac{3}{2}(A_0-zA_2)+\frac{2z}{(z-1)}+\frac{2(2z-1)}{z(z-1)}\right]+B_3\left[(z+1)A_2+\frac{1}{2}A_1-\frac{2(2z-1)}{z(z-1)}\right]$$
$$-\frac{1}{2}A_2B_2+(z+1)B_4'-B_3'+\frac{\alpha}{z^2(z-1)^2}[2(z+1)A_1-A_0]$$
$$-\frac{A_2}{z^2(z-1)^2}[\alpha(z^2+4z+1)+\beta z+(\delta z+\gamma)(z-1)],$$

$$\psi_5=B_3\left[\frac{3}{2}(A_0-zA_2)+\frac{2z}{(z-1)}+\frac{2(2z-1)}{z(z-1)}\right]+B_2\left[(z+1)A_2+\frac{1}{2}A_1-\frac{2(2z-1)}{z(z-1)}\right]-\frac{1}{2}A_2B_1$$
$$-B_4\left[\frac{1}{2}zA_1+(z+1)A_0+\frac{2z}{(z-1)}\right]+\frac{2\alpha(z+1)}{z^2(z-1)^2}A_0-\frac{A_1}{z^2(z-1)^2}[\alpha(z^2+4z+1)+\beta z$$
$$+(\delta z+\gamma)(z-1)]+\frac{2A_2}{z(z-1)^2}[(\alpha+\beta)(z+1)+(\gamma+\delta)(z-1)]+(z+1)B_3'-zB_4'-B_2',$$

$$\psi_4 = B_2\left[\frac{3}{2}(A_0 - zA_2) + \frac{2z}{(z-1)} + \frac{2(2z-1)}{z(z-1)}\right] + B_1\left[(z+1)A_2 + \frac{1}{2}A_1 - \frac{2(2z-1)}{z(z-1)}\right] - \frac{1}{2}A_2B_0$$
$$- B_3\left[\frac{1}{2}zA_1 + (z+1)A_0 + \frac{2z}{(z-1)}\right] + \frac{1}{2}zA_0B_4 - \frac{A_0}{z^2(z-1)^2}[\alpha(z^2+4z+1) + \beta z$$
$$+ (\delta z + \gamma)(z-1)] + \frac{2A_1}{z(z-1)^2}[(\alpha+\beta)(z+1) + (\gamma+\delta)(z-1)]$$
$$- \frac{A_2}{z(z-1)}[\alpha z + \beta(z^2+4z+1) + \gamma z(z-1) + \delta(z-1)] + (z+1)B_2' - zB_3' - B_1',$$

$$\psi_3 = B_1\left[\frac{3}{2}(A_0 - zA_2) + \frac{2z}{(z-1)} + \frac{2(2z-1)}{z(z-1)}\right] + B_0\left[(z+1)A_2 + \frac{1}{2}A_1 - \frac{2(2z-1)}{z(z-1)}\right] + \frac{1}{2}zA_0B_3$$
$$- B_2\left[\frac{1}{2}zA_1 + (z+1)A_0 + \frac{2z}{(z-1)}\right] + \frac{2\beta(z+1)}{(z-1)^2}A_2 + \frac{2A_0}{z(z-1)^2}[(\alpha+\beta)(z+1)$$
$$+ (\gamma+\delta)(z-1)] - \frac{A_1}{z(z-1)}[\alpha z + \beta(z^2+4z+1) + (\gamma z+\delta)(z-1)] + (z+1)B_1' - zB_2' - B_0',$$

$$\psi_2 = B_0\left[\frac{3}{2}(A_0 - zA_2) + \frac{2z}{(z-1)} + \frac{2(2z-1)}{z(z-1)}\right] - B_1\left[\frac{1}{2}zA_1 + (z+1)A_0 + \frac{2z}{(z-1)}\right]$$
$$+ \frac{1}{2}zA_0B_2 + (z+1)B_0' - zB_1' + \frac{\beta}{(z-1)^2}[2(z+1)A_1 - zA_2]$$
$$- \frac{A_0}{z(z-1)}[\alpha z + \beta(z^2+4z+1) + (\gamma z+\delta)(z-1)],$$

$$\psi_1 = \frac{\beta}{(z-1)^2}[2(z+1)A_0 - zA_1] + \frac{1}{2}zA_0B_1 - B_0\left[(z+1)A_0 + \frac{1}{2}zA_1 + \frac{2z}{(z-1)}\right] - zB_0',$$

$$\psi_0 = \frac{z}{2}A_0\left[B_0 - \frac{2\beta}{(z-1)^2}\right].$$

As an example, let

$$A_1 = \frac{-1}{z(z-1)}[(z^2-1)A_0 + 2], \quad A_2 = \frac{1}{z(z-1)}[(z-1)A_0 + 2],$$

$$B_3 = \frac{-1}{z^3(z-1)}[z^2(z^2-1)B_2 + z(z-1)(z^2+z+1)B_1 + (z-1)(z^3+z^2+z+1)B_0 - 2\gamma z^2 - 2\delta], \tag{7.4}$$

$$B_4 = \frac{1}{z^3(z-1)}[z^2(z-1)B_2 + z(z^2-1)B_1 + (z-1)(z^2+z+1)B_0 - 2\gamma z - 2\delta],$$

and $\phi_0 = \phi_1 = \psi_0 = \psi_1 = 0$. Then, the equation (7.2) takes the following form,

$$\phi_6 v' + \psi_8 v^2 + [\psi_7 + (z+1)\psi_8]v + \frac{1}{z^2}\psi_2 = 0, \tag{7.5}$$

and the equation (1.26) can be written as

$$\left(\psi_8 v^2+[\psi_7+(z+1)\psi_8]v+\frac{1}{z^2}\psi_2+\frac{1}{2}\phi_6(A_2v^2+A_1v+A_0)\right)^2$$
$$=\phi_6^2[(B_4+\tfrac{1}{4}A_2^2)v^4+(B_3+\tfrac{1}{2}A_1A_2)v^3+(B_2+\tfrac{1}{4}A_1^2+\tfrac{1}{2}A_0A_2)v^2$$
$$+(B_1+\tfrac{1}{2}A_0A_1)v+(B_0+\tfrac{1}{4}A_0^2)]. \tag{7.6}$$

The equation $\psi_0=0$ implies that either $A_0=0$ or $B_0=2\beta/(z-1)^2$.

Case i: $A_0=0$: Then, the equation $\phi_0=0$ implies that $B_0=-2\beta/(z-1)^2$ and then the equations $\phi_1=\psi_1=0$ are satisfied identically. Let $B_4=(\mu^2-1)/z^2(z-1)^2$, where $\mu=1-\sqrt{2\alpha}$ and $\sqrt{2\alpha}$ can take either sign, and without loss of generality, let $B_2=-[1/z^2(z-1)^2][4(z+1)u+(\beta-\alpha+\sqrt{\alpha})(3z+1)+(\gamma-\delta)(3z-1)]$. Then the equation (7.6) reduces to the following quadratic equation for v:

$$Av^2+Bv+C=0,$$
$$A=4\mu^2[4z(z-1)u'+4u+2\nu z-\kappa]+[4u-2\lambda(z-1)+\nu-\mu^2]^2, \tag{7.7}$$
$$B=2z[4u-2\lambda(z-1)+\nu-\mu^2][4(z-1)u'-4u-\nu]-4\mu^2z[4u+2(\gamma+\beta)(z-1)+\nu+4\beta],$$
$$C=z^2[4(z-1)u'-4u-\nu]^2+8\beta\mu^2z^2,$$

where $\kappa=\alpha-\beta+\gamma-\delta-\sqrt{2\alpha}+1$, $\lambda=\alpha+\delta-\sqrt{2\alpha}$, and $\nu=\beta+\gamma-\alpha-\delta+\sqrt{2\alpha}$. The equation (1.21) can be written as

$$[z(z-1)v'-v(v-1)]^2=\mu^2v^4+[4u-2\lambda(z-1)+\nu-\mu^2]v^3-[4(z+1)u+3\nu z-\kappa]v^2$$
$$+z[4u+2(\gamma+\beta)(z-1)+\nu+4\beta]v-2\beta z^2, \tag{7.8}$$

and the equation (1.22) becomes

$$v'=\frac{1}{2\mu z(z-1)}\{2\mu\sqrt{2\alpha}v^2-[4u-2\lambda(z-1)+\nu-\mu^2+2\mu]v-z[4(z-1)u'-4u-\nu]\}. \tag{7.9}$$

Equations (7.7) and (7.8) give one-to-one correspondence between solutions $v(z)$ of PVI and solutions $u(z)$ of the following second-order second degree equation of Painlevé type:

$$z^2(z-1)^2(u'')^2=-4u'(zu'-u)^2+4(u')^2(zu'-u)+\kappa(u')^2+\lambda(\gamma+\beta)(zu'-u)$$
$$+\tfrac{1}{4}[4(\gamma-\beta)(\mu^2-\lambda)+\nu^2]u'+\tfrac{1}{4}[\lambda^2(\gamma-\beta)+(\gamma+\beta)^2(\mu^2-\lambda)]. \tag{7.10}$$

The equation (7.10) was first obtained by Cosgrove[17] and labeled as SD-I.a.

Case ii: $B_0=2\beta/(z-1)^2$: Then $A_0=2(\mu-1)/(z-1)$, $B_0=\tfrac{1}{4}A_0^2$, and $B_1=-\tfrac{1}{2}A_0A_1$, where $(\mu-1)^2=-2\beta$. Without loss of generality, let

$$B_2=\frac{1}{z^2(z-1)^2}[zu-\mu^2(z^2+4z+1)+2\mu z(z+2)-(z^2+z-1)+2\gamma z(z-1)+2\delta(z-1)]. \tag{7.11}$$

Then one obtains $B_4=[1/z^2(z-1)^2](u-\mu^2)$ and $B_3=[-1/z^2(z-1)^2][(z+1)(u-2\mu^2)+2\mu z+\lambda(z-1)]$, where $\lambda=2\gamma+2\delta-1$. With these choices the equation (7.6) yields the following quadratic equation for v:

$$Av^2+Bv+C=0, \tag{7.12}$$

where

$$\begin{aligned} A&=u[u^2-2(\mu^2+2\alpha)u+(\mu^2-2\alpha)^2],\\ B&=-(4\mu z(z-1)uu'+[(z+1)u+\lambda(z-1)][u^2-2(\mu^2+2\alpha)u+(\mu^2-2\alpha)^2]),\\ C&=-(z(z-1)u'-\mu[(z+1)u+\lambda(z-1)])^2+[zu+2\gamma(z-1)^2+\lambda(z-1)](u+\mu^2-2\alpha)^2. \end{aligned} \tag{7.13}$$

The equations (1.21) and (1.22) become

$$\begin{aligned} &[z(z-1)v'-\mu v^2+(\mu z-z+\mu)v-(\mu-1)z]^2\\ &\quad =uv^4-[(z+1)u+\lambda(z-1)]v^3+[zu+2\gamma(z-1)^2+\lambda(z-1)]v^2, \end{aligned} \tag{7.14}$$

and

$$\begin{aligned} v'=\frac{-1}{z(z-1)(u+\mu^2-2\alpha)}&\{\mu(u-\mu^2+2\alpha)v^2+[z(z-1)u'-z(u+\mu^2-2\alpha)\\ &-\mu\lambda(z-1)+\mu(\mu^2-2\alpha)(z+1)]v-(\mu-1)z(u+\mu^2-2\alpha)\}, \end{aligned} \tag{7.15}$$

respectively. Let $u(z)$ be a solution of the following second-order second-degree equation of Painlevé type:

$$\begin{aligned} &[4z^2u^2u''-2z^2u(u')^2+4zu^2u'+P_4(u)]^2\\ &\quad=\left[\frac{(z+1)}{(z-1)}u^2-\lambda(\mu^2-2\alpha)\right]^2[4z^2u(u')^2+Q_4(u)],\\ &P_4(u):=u^4+(\lambda-4\gamma-\mu^2-2\alpha)u^3+[\lambda^2(\mu^2+2\alpha)+(\lambda-4\gamma)(\mu^2-2\alpha)^2]u-\lambda^2(\mu^2-2\alpha)^2,\\ &Q_4(u):=[u^2+2(\lambda-4\gamma)u+\lambda^2][u^2-2(\mu^2+2\alpha)u+(\mu^2-2\alpha)^2]. \end{aligned} \tag{7.16}$$

Then, the equations (7.12) and (7.14) gives one-to-one correspondence between solutions $v(z)$ of PVI and $u(z)$ of the equation (7.16).

[1] P. Painlevé, Bull. Soc. Math. Fr. **28**, 214 (1900); Acta. Math. **25**, 1 (1912).
[2] B. Gambier, Acta Math. **33**, 1 (1909).
[3] R. Fuchs, Math. Ann. **63**, 301 (1907).
[4] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956).
[5] N. P. Erugin, Dokl. Akad. Nauk BSSR **2** (1958).
[6] A. S. Fokas and M. J. Ablowitz, J. Math. Phys. **23**, 2033 (1982).
[7] U. Muğan and A. S. Fokas, J. Math. Phys. **33**, 2031 (1992).
[8] U. Muğan, A. Sakka, J. Math. Phys. **36**, 1284 (1995).
[9] A. S. Fokas and X. Zhou, Commun. Math. Phys. **144**, 601 (1992).
[10] A. S. Fokas, U. Muğan, and X. Zhou, Inverse Probl. **8**, 757 (1992).
[11] U. Muğan and A. Sakka, J. Phys. A **28**, 4109 (1995).
[12] C. M. Cosgrove, Stud. Appl. Math. **90**, 119 (1993).
[13] J. Chazy, Acta Math. **34**, 317 (1911).
[14] R. Garnier, Ann. Sci. Ec. Norm. Super., **29**, 1 (1912).
[15] F. Bureau, Ann. Math. **66**, 1 (1964).
[16] F. Bureau, Ann. Math. **91**, 163 (1972).
[17] C. M. Cosgrove and G. Scoufis, Stud. Appl. Math. **88**, 25 (1993).
[18] A. S. Fokas and Y. C. Yortsos, Lett. Nuovo Cimento **30**, 539 (1981).
[19] C. M. Cosgrove, J. Phys. A **10**, 1481 (1977).
[20] C. M. Cosgrove, J. Phys. A **10**, 2093 (1977).
[21] C. M. Cosgrove, J. Phys. A **11**, 2405 (1978).

[22] R. Chalkley, J. Diff. Eqns. **68**, 72 (1987).
[23] P. Appell, J. Math. Pures Appl. **5**, 361 (1889).
[24] A. Sakka and U. Muğan, J. Phys. A **30**, 5159 (1997).
[25] A. Sakka and U. Muğan, J. Phys. A **31**, 2471 (1998).

Classification of differential calculi on $U_q(\mathfrak{b}_+)$, classical limits, and duality

Robert Oeckl
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 9EW, United Kingdom

(Received 27 July 1998; accepted for publication 5 March 1999)

We give a complete classification of bicovariant first order differential calculi on the quantum enveloping algebra $U_q(\mathfrak{b}_+)$ which we view as the quantum function algebra $C_q(B_+)$. Here, $\mathfrak{b}_+$ is the Borel subalgebra of $\mathfrak{sl}_2$. We do the same in the classical limit $q \to 1$ and obtain a one-to-one correspondence in the finite dimensional case. It turns out that the classification is essentially given by finite subsets of the positive integers. We proceed to investigate the classical limit from the dual point of view, i.e., with ''function algebra'' $U(\mathfrak{b}_+)$ and ''enveloping algebra'' $C(B_+)$. In this case there are many more differential calculi than coming from the q-deformed setting. As an application, we give the natural intrinsic four-dimensional calculus of κ-Minkowski space and the associated formal integral.
 [S0022-2488(99)00807-5]

I. INTRODUCTION

One of the fundamental ingredients in the theory of noncommutative or quantum geometry is the notion of a differential calculus. In the framework of quantum groups the natural notion is that of a bicovariant differential calculus as introduced by Woronowicz.[1] Due to the allowance of noncommutativity the uniqueness of a canonical calculus is lost. It is therefore desirable to classify the possible choices. The most important piece is the space of one-forms or ''first order differential calculus'' to which we will restrict our attention in the following. (From this point on we will use the term ''differential calculus'' to denote a bicovariant first order differential calculus.)

Much attention has been devoted to the investigation of differential calculi on quantum groups $C_q(G)$ of function algebra type for G a simple Lie group. Natural differential calculi on matrix quantum groups were obtained by Jurco[2] and Carow-Watamura *et al.*[3] A partial classification of calculi of the same dimension as the natural ones was obtained by Schmüdgen and Schüler.[4] More recently, a classification theorem for factorisable cosemisimple quantum groups was obtained by Majid,[5] covering the general $C_q(G)$ case. A similar result was obtained later by Baumann and Schmitt.[6] Also, Heckenberger and Schmüdgen[7] gave a complete classification on $C_q(SL(N))$ and $C_q(Sp(N))$.

In contrast, for G not simple or semisimple the differential calculi on $C_q(G)$ are largely unknown. A particularly basic case is the Lie group B_+ associated with the Lie algebra $\mathfrak{b}_+$ generated by two elements X,H with the relation $[H,X]=X$. The quantum enveloping algebra $U_q(\mathfrak{b}_+)$ is self-dual, i.e., is nondegenerately paired with itself.[8] This has an interesting consequence: $U_q(\mathfrak{b}_+)$ may be identified with (a certain algebraic model of) $C_q(B_+)$. The differential calculi on this quantum group and on its ''classical limits'' $C(B_+)$ and $U(\mathfrak{b}_+)$ will be the main concern of this paper. We pay hereby equal attention to the dual notion of ''quantum tangent space.''

In Sec. II we obtain the complete classification of differential calculi on $C_q(B_+)$. It turns out that (finite dimensional) differential calculi are characterized by finite subsets $I \subset \mathbb{N}$. These sets determine the decomposition into coirreducible (i.e., not admitting quotients) differential calculi characterized by single integers. For the coirreducible calculi the explicit formulas for the commutation relations and braided derivations are given.

In Sec. III we give the complete classification for the classical function algebra $C(B_+)$. It is essentially the same as in the q-deformed setting and we stress this by giving an almost one-to-one correspondence of differential calculi to those obtained in the previous section. In contrast, however, the decomposition and coirreducibility properties do not hold at all. (One may even say that they are maximally violated.) We give the explicit formulas for those calculi corresponding to coirreducible ones.

More interesting perhaps is the ''dual'' classical limit, i.e., we view $U(\mathfrak{b}_+)$ as a quantum function algebra with quantum enveloping algebra $C(B_+)$. This is investigated in Sec. IV. It turns out that in this setting we have considerably more freedom in choosing a differential calculus since the bicovariance condition becomes much weaker. This shows that this dual classical limit is in a sense ''unnatural'' as compared to the ordinary classical limit of Sec. III. However, we can still establish a correspondence of certain differential calculi to those of Sec. II. The decomposition properties are conserved while the coirreducibility properties are not. We give the formulas for the calculi corresponding to coirreducible ones.

Another interesting aspect of viewing $U(\mathfrak{b}_+)$ as a quantum function algebra is the connection to quantum deformed models of space–time and its symmetries. In particular, the κ-deformed Minkowski space coming from the κ-deformed Poincaré algebra[9,10] is just a simple generalization of $U(\mathfrak{b}_+)$. We use this in Sec. V to give a natural four-dimensional differential calculus. Then we show (in a formal context) that integration is given by the usual Lesbegue integral on $\mathbb{R}^n$ after normal ordering. This is obtained in an intrinsic context different from the standard κ-Poincaré approach.

A further important motivation for the investigation of differential calculi on $U(\mathfrak{b}_+)$ and $C(B_+)$ is the relation of those objects to the Planck-scale Hopf algebra.[11,12] This shall be developed elsewhere.

In the remaining parts of this introduction we will specify our conventions and provide preliminaries on the quantum group $U_q(\mathfrak{b}_+)$, its deformations, and differential calculi.

A. Conventions

Throughout, $\mathbb{k}$ denotes a field of characteristic 0 and $\mathbb{k}(q)$ denotes the field of rational functions in one parameter q over $\mathbb{k}$. $\mathbb{k}(q)$ is our ground field in the q-deformed setting, while $\mathbb{k}$ is the ground field in the ''classical'' settings. Within Sec. II one could equally well view $\mathbb{k}$ as the ground field with $q \in \mathbb{k}^*$ not a root of unity. This point of view is problematic, however, when obtaining ''classical limits'' as in Secs. III and IV.

The positive integers are denoted by $\mathbb{N}$ while the non-negative integers are denoted by $\mathbb{N}_0$. We define q-integers, q-factorials, and q-binomials as follows:

$$[n]_q = \sum_{i=0}^{n-1} q^i, \quad [n]_q! = [1]_q [2]_q \cdots [n]_q, \quad \begin{bmatrix} n \\ m \end{bmatrix}_q = \frac{[n]_q!}{[m]_q! [n-m]_q!}.$$

For a function of several variables (among them x) over $\mathbb{k}$ we define

$$(T_{a,x} f)(x) = f(x+a), \quad (\nabla_{a,x} f)(x) = \frac{f(x+a) - f(x)}{a},$$

with $a \in \mathbb{k}$ and similarly over $\mathbb{k}(q)$

$$(Q_{m,x} f)(x) = f(q^m x), \quad (\partial_{q,x} f)(x) = \frac{f(x) - f(qx)}{x(1-q)},$$

with $m \in \mathbb{Z}$.

We frequently use the notion of a polynomial in an extended sense. Namely, if we have an algebra with an element g and its inverse g^{-1} [as in $U_q(\mathfrak{b}_+)$] we will mean by a polynomial in g, g^{-1} a finite power series in g with exponents in $\mathbb{Z}$. The length of such a polynomial is the difference between highest and lowest degree.

If H is a Hopf algebra, then H^{op} will denote the Hopf algebra with the opposite product.

B. $U_q(\mathfrak{b}_+)$ and its classical limits

We recall that, in the framework of quantum groups, the duality between enveloping algebra $U(\mathfrak{g})$ of the Lie algebra and algebra of functions $C(G)$ on the Lie group carries over to q-deformations. In the case of $\mathfrak{b}_+$, the q-deformed enveloping algebra $U_q(\mathfrak{b}_+)$ defined over $\mathbb{k}(q)$ as

$$U_q(\mathfrak{b}_+)=\mathbb{k}(q)\langle X,g,g^{-1}\rangle \quad \text{with relations,}$$

$$gg^{-1}=1, \quad Xg=qgX,$$

$$\Delta X=X\otimes 1+g\otimes X, \quad \Delta g=g\otimes g,$$

$$\epsilon(X)=0, \quad \epsilon(g)=1, \quad \mathrm{S}X=-g^{-1}X, \quad \mathrm{S}g=g^{-1}$$

is self-dual. Consequently, it may alternatively be viewed as the quantum algebra $C_q(B_+)$ of functions on the Lie group B_+ associated with $\mathfrak{b}_+$. It has two classical limits, the enveloping algebra $U(\mathfrak{b}_+)$ and the function algebra $C(B_+)$. The transition to the classical enveloping algebra is achieved by replacing q by e^{-t} and g by e^{tH} in a formal power series setting in t, introducing a new generator H. Now, all expressions are written in the form $\Sigma_j a_j t^j$ and only the lowest order in t is kept. The transition to the classical function algebra on the other hand is achieved by setting $q=1$. This may be depicted as follows:

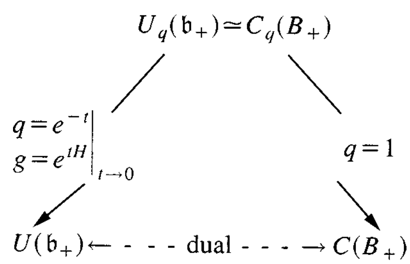

The self-duality of $U_q(\mathfrak{b}_+)$ is expressed as a pairing $U_q(\mathfrak{b}_+)\times U_q(\mathfrak{b}_+)\rightarrow\mathbb{k}$ with itself,

$$\langle X^n g^m, X^r g^s\rangle=\delta_{n,r}[n]_q!q^{-n(n-1)/2}q^{-ms} \quad \forall n,r\in\mathbb{N}_0 \; m,s\in\mathbb{Z}.$$

In the classical limit this becomes the pairing $U(\mathfrak{b}_+)\times C(B_+)\rightarrow\mathbb{k}$,

$$\langle X^n H^m, X^r g^s\rangle=\delta_{n,r}n!s^m \quad \forall n,m,r\in\mathbb{N}_0 \; s\in\mathbb{Z}. \tag{1}$$

C. Differential calculi and quantum tangent spaces

In this section we recall some facts about differential calculi along the lines of Majid's treatment in Ref. 5.

Following Woronowicz,[1] first order bicovariant differential calculi on a quantum group A (of function algebra type) are in one-to-one correspondence to submodules M of $\ker\epsilon\subset A$ in the category ${}^A_A\mathcal{M}$ of (say) left crossed modules of A via left multiplication and left adjoint coaction,

$$a\triangleright v=av, \quad \mathrm{Ad_L}(v)=v_{(1)}\mathrm{S}v_{(3)}\otimes v_{(2)} \quad \forall a\in A, v\in M.$$

More precisely, given a crossed submodule M, the corresponding calculus is given by $\Gamma=\ker\epsilon/M\otimes A$ with $\mathrm{d}a=\pi(\Delta a-1\otimes a)$ (π the canonical projection). The right action and coaction on Γ are given by the right multiplication and coproduct on A, the left action and coaction by the tensor product ones with $\ker\epsilon/M$ as a left crossed module. In all of what follows, ''differential calculus'' will mean ''bicovariant first order differential calculus.''

Alternatively,[5] given in addition a quantum group H dually paired with A (which we might think of as being of enveloping algebra type), we can express the coaction of A on itself as an action of H^{op} using the pairing

$$h \triangleright v = \langle h, v_{(1)} \mathrm{S} v_{(3)} \rangle v_{(2)} \quad \forall h \in H^{\mathrm{op}}, v \in A.$$

Thereby we change from the category of (left) crossed A-modules to the category of left modules of the quantum double $A \bowtie H^{\mathrm{op}}$.

In this picture the pairing between A and H descends to a pairing between $A/\mathbb{k}1$ (which we may identify with $\ker\epsilon \subset A$) and $\ker\epsilon \subset H$. Further quotienting $A/\mathbb{k}1$ by M (viewed in $A/\mathbb{k}1$) leads to a pairing with the subspace $L \subset \ker\epsilon c H$ that annihilates M. L is called a ''quantum tangent space'' and is dual to the differential calculus Γ generated by M in the sense that $\Gamma \cong \mathrm{Lin}(L,A)$ via

$$A/(\mathbb{k}1 + M) \otimes A \to \mathrm{Lin}(L,A) \quad v \otimes a \mapsto \langle \cdot , v \rangle a \tag{2}$$

if the pairing between $A/(\mathbb{k}1+M)$ and L is nondegenerate.

The quantum tangent spaces are obtained directly by dualising the (left) action of the quantum double on A to a (right) action on H. Explicitly, this is the adjoint action and the coregular action

$$h \triangleright x = h_{(1)} x \mathrm{S} h_{(2)}, \quad a \triangleright x = \langle x_{(1)}, a \rangle x_{(2)} \quad \forall a \in A^{\mathrm{op}}, h, x \in H,$$

where we have converted the right action to a left action by going from $A \bowtie H^{\mathrm{op}}$-modules to $H \bowtie A^{\mathrm{op}}$-modules. Quantum tangent spaces are subspaces of $\ker\epsilon \subset H$ invariant under the projection of this action to $\ker\epsilon$ via $x \mapsto x - \epsilon(x)1$. Alternatively, the left action of A^{op} can be converted to a left coaction of H being the comultiplication (with subsequent projection onto $H \otimes \ker\epsilon$).

We can use the evaluation map (2) to define a ''braided derivation'' on elements of the quantum tangent space via

$$\partial_x : A \to A \quad \partial_x(a) = \mathrm{d}a(x) = \langle x, a_{(1)} \rangle a_{(2)} \quad \forall x \in L, a \in A.$$

This obeys the braided derivation rule

$$\partial_x(ab) = (\partial_x a) b + a_{(2)} \partial_{a_{(1)} \triangleright x} b \quad \forall x \in L, a \in A.$$

Given a right invariant basis $\{\eta_i\}_{i \in I}$ of Γ with a dual basis $\{\phi_i\}_{i \in I}$ of L we have

$$\mathrm{d}a = \sum_{i \in I} \eta_i \cdot \partial_i(a) \quad \forall a \in A,$$

where we denote $\partial_i = \partial_{\phi_i}$. (This can be easily seen to hold by evaluation against $\phi_i \forall i$.)

II. CLASSIFICATION ON $C_q(B_+)$ AND $U_q(\mathfrak{b}_+)$

In this section we completely classify differential calculi on $C_q(B_+)$ and, dually, quantum tangent spaces on $U_q(\mathfrak{b}_+)$. We start by classifying the relevant crossed modules and then proceed to a detailed description of the calculi.

Lemma II.1: (a) Left crossed $C_q(B_+)$-submodules $M \subseteq C_q(B_+)$ by left multiplication and left adjoint coaction are in one-to-one correspondence to pairs (P,I), where $P \in \mathbb{k}(q)[g]$ is a polynomial with $P(0)=1$ and $I \subset \mathbb{N}$ is finite. $\mathrm{codim} M < \infty$ iff $P=1$. *In particular* $\mathrm{codim} M = \Sigma_{n \in I} n$ *if* $P=1$.

(b) The finite codimensional maximal M correspond to the pairs $(1,\{n\})$ with n the codimension. The infinite codimensional maximal M are characterized by $(P,\emptyset)$ with P irreducible and $P(g) \neq 1 - q^{-k} g$ for any $k \in \mathbb{N}_0$.

(c) Crossed submodules M of finite codimension are intersections of maximal ones. In particular $M = \cap_{n \in I} M^n$, with M^n corresponding to $(1,\{n\})$.

Proof: (a) Let $M \subseteq C_q(B_+)$ be a crossed $C_q(B_+)$-submodule by left multiplication and left adjoint coaction and let $\Sigma_n X^n P_n(g) \in M$, where P_n are polynomials in g, g^{-1} (every element of $C_q(B_+)$ can be expressed in this form). From the formula for the coaction ((A1), see Appendix) we observe that for all n and for all $t \leqslant n$ the element

$$X^t P_n(g) \prod_{s=1}^{n-t} (1-q^{s-n}g)$$

lies in *M*. In particular this is true for $t=n$, meaning that elements of constant degree in *X* lie separately in *M*. It is therefore enough to consider such elements.

Let now $X^n P(g) \in M$. By left multiplication $X^n P(g)$ generates any element of the form $X^k P(g) Q(g)$, where $k \geqslant n$ and Q is any polynomial in g, g^{-1}. (Note that $Q(q^k g) X^k = X^k Q(g)$.) We see that *M* contains the following elements:

$$\begin{array}{l}
\vdots \\
X^{n+2}P(g) \\
X^{n+1}P(g) \\
X^{n}P(g) \\
X^{n-1}P(g)(1-q^{1-n}g) \\
X^{n-2}P(g)(1-q^{1-n}g)(1-q^{2-n}g) \\
\vdots \\
XP(g)(1-q^{1-n}g)(1-q^{2-n}g)\ldots(1-q^{-1}g) \\
\quad P(g)(1-q^{1-n}g)(1-q^{2-n}g)\ldots(1-q^{-1}g)(1-g).
\end{array}$$

Moreover, if *M* is generated by $X^n P(g)$ as a module then these elements generate *M* as a vector space by left multiplication with polynomials in g, g^{-1}. (Observe that the application of the coaction to any of the elements shown does not generate elements of new type.)

Now, let *M* be a given crossed submodule. We pick, among the elements in *M* of the form $X^n P(g)$ with *P* of minimal length, one with lowest degree in *X*. Then certainly the elements listed above are in *M*. Furthermore for any element of the form $X^k Q(g)$, Q must contain *P* as a factor and for $k<n$, Q must contain $P(g)(1-q^{1-n}g)$ as a factor. We continue by picking the smallest n_2, so that $X^{n_2}P(g)(1-q^{1-n}g) \in M$. Certainly $n_2<n$. Again, for any element of $X^l Q(g)$ in *M* with $l<n_2$, we have that $P(g)(1-q^{1-n}g)(1-q^{1-n_2}g)$ divides $Q(g)$. We proceed by induction, until we arrive at degree zero in *X*.

We obtain the following elements generating *M* as a vector space by left multiplication with polynomials in g, g^{-1} (rename $n_1=n$):

$$\begin{array}{l}
\vdots \\
X^{n_1+1}P(g) \\
X^{n_1}P(g) \\
X^{n_1-1}P(g)(1-q^{1-n_1}g) \\
\vdots
\end{array}$$

$$X^{n_2}P(g)(1-q^{1-n_1}g)$$

$$X^{n_2-1}P(g)(1-q^{1-n_1}g)(1-q^{1-n_2}g)$$

$$\vdots$$

$$X^{n_3}P(g)(1-q^{1-n_1}g)(1-q^{1-n_2}g)$$

$$X^{n_3-1}P(g)(1-q^{1-n_1}g)(1-q^{1-n_2}g)(1-q^{1-n_3}g)$$

$$\vdots$$

$$P(g)(1-q^{1-n_1}g)(1-q^{1-n_2}g)(1-q^{1-n_3}g)\cdots(1-q^{1-n_m}g).$$

We see that the integers $n_1,\ldots,n_m$ uniquely determine the shape of this picture. The polynomial $P(g)$ on the other hand can be shifted (by g and g^{-1}) or renormalized. To determine M uniquely we shift and normalize P in such a way that it contains no negative powers and has unit constant coefficient. P can then be viewed as a polynomial $\in \Bbbk(q)[g]$.

We see that the codimension of M is the sum of the lengths of the polynomials in g over all degrees in X in the above picture. Finite codimension corresponds to $P=1$. In this case the codimension is the sum $n_1+\cdots+n_m$.

(b) We observe that polynomials of the form $1-q^j g$ have no common divisors for distinct j. Therefore, finite codimensional crossed submodules are maximal if and only if there is just one integer $(m=1)$. Thus, the maximal left crossed submodule of codimension k is generated by X^k and $1-q^{1-k}g$. For an infinite codimensional crossed submodule we certainly need $m=0$. Then, the maximality corresponds to irreducibility of P.

(c) This is again due to the distinctness of factors $1-q^j g$. □

Corollary II.2: (a) Left crossed $C_q(B_+)$-submodules $M\subseteq \ker\epsilon\subset C_q(B_+)$ are in one-to-one correspondence to pairs (P,I) as in Lemma II.1 with the additional constraint $(1-g)$ divides $P(g)$ or $1\in I$. $\operatorname{codim} M<\infty$ iff $P=1$. *In particular* $\operatorname{codim}M=(\Sigma_{n\in I}n)-1$ *if* $P=1$.

(b) The finite codimensional maximal M correspond to the pairs $(1,\{1,n\})$ with $n\geqslant 2$ the codimension. The infinite codimensional maximal M correspond to pairs $(P,\{1\})$ with P irreducible and $P(g)\neq 1-q^{-k}g$ for any $k\in\mathbb{N}_0$.

(c) Crossed submodules M of finite codimension are intersections of maximal ones. In particular $M=\cap_{n\in I}M^n$, with M^n corresponding to $(1,\{1,n\})$.

Proof: First observe that $\Sigma_n X^n P_n(g)\in\ker\epsilon$ if and only if $(1-g)$ divides $P_0(g)$. This is to say that that $\ker\epsilon$ is the crossed submodule corresponding to the pair $(1,\{1\})$ in Lemma II.1. We obtain the classification from the one of Lemma II.1 by intersecting everything with this crossed submodule. In particular, this reduces the codimension by one in the finite codimensional case. □

Lemma II.3: (a) Left crossed $U_q(\mathfrak{b}_+)$-submodules $L\subseteq U_q(\mathfrak{b}_+)$ via the left adjoint action and left regular coaction are in one-to-one correspondence to the set $\{\mathbb{N}_0\rightarrow\{1,2,3\}\}\times\{\mathbb{N}\rightarrow\{1,2\}\}$. Finite dimensional L are in one-to-one correspondence to finite sets $I\subset\mathbb{N}$ and $\dim L=\Sigma_{n\in I}n$.

(b) Finite dimensional irreducible L correspond to $\{n\}$ with n the dimension.

(c) Finite dimensional L are direct sums of irreducible ones. In particular $L=\oplus_{n\in I}L^n$ with L^n corresponding to $\{n\}$.

Proof: (a) The action takes the explicit form

$$g\triangleright X^n g^k=q^{-n}X^n g^k,\quad X\triangleright X^n g^k=X^{n+1}g^k(1-q^{-(n+k)}),$$

while the coproduct is

$$\Delta(X^n g^k)=\sum_{r=0}^{n}\begin{bmatrix} n\\ r\end{bmatrix}_q q^{-r(n-r)}X^{n-r}g^{k+r}\otimes X^r g^k,$$

which we view as a left coaction here. Let now $L \subseteq U_q(\mathfrak{b}_+)$ be a crossed $U_q(\mathfrak{b}_+)$-submodule via this action and coaction. For $\Sigma_n X^n P_n(g) \in L$ invariance under the action by g clearly means that $X^n P_n(g) \in L\ \forall n$. Then from invariance under the coaction we can conclude that if $X^n \Sigma_j a_j g^j \in L$ we must have $X^n g^j \in L\ \forall j$, i.e., elements of the form $X^n g^j$ lie separately in L and it is sufficient to consider such elements. From the coaction we learn that if $X^n g^j \in L$ we have $X^m g^j \in L\ \forall m \leqslant n$. The action by X leads to $X^n g^j \in L \Rightarrow X^{n+1} g^j \in L$ except if $n+j=0$. The classification is given by the possible choices we have for each power in g. For every positive integer j we can choose whether or not to include the span of $\{X^n g^j | \forall n\}$ in L and for every nonpositive integer we can choose to include either the span of $\{X^n g^j | \forall n\}$ or just $\{X^n g^j | \forall n \leqslant -j\}$ or neither, i.e., for positive integers ($\mathbb{N}$) we have two choices while for nonpositive (identified with $\mathbb{N}_0$) ones we have three choices.

Clearly, the finite dimensional L are those where we choose only to include finitely many powers of g and also only finitely many powers of X. The latter is only possible for the nonpositive powers of g. By identifying positive integers n with powers $1-n$ of g, we obtain a classification by finite subsets of $\mathbb{N}$.

(b) Irreducibility clearly corresponds to just including one power of g in the finite dimensional case.

(c) The decomposition property is obvious from the discussion. □

Corollary II.4: (a) Left crossed $U_q(\mathfrak{b}_+)$*-submodules* $L \subseteq \ker\epsilon \subset U_q(\mathfrak{b}_+)$ *via the left adjoint action and left regular coaction (with subsequent projection to* $\ker\epsilon$ *via* $x \mapsto x - \epsilon(x)1$*) are in one-to-one correspondence to the set* $\{\mathbb{N} \to \{1,2,3\}\} \times \{\mathbb{N}_0 \to \{1,2\}\}$. *Finite dimensional L are in one-to-one correspondence to finite sets* $I \subset \mathbb{N} \backslash \{1\}$ *and* $\dim L = \Sigma_{n \in I} n$.

(b) Finite dimensional irreducible L correspond to $\{n\}$ *with* $n \geqslant 2$ *the dimension.*

(c) Finite dimensional L are direct sums of irreducible ones. In particular $L = \oplus_{n \in I} L^n$ *with* L^n *corresponding to* $\{n\}$.

Proof: Only a small modification of Lemma II.3 is necessary. Elements of the form $P(g)$ are replaced by elements of the form $P(g)-P(1)$. Monomials with nonvanishing degree in X are unchanged. The choices for elements of degree 0 in g are reduced to either including the span of $\{X^k | \forall k>0\}$ in the crossed submodule or not. In particular, the crossed submodule characterized by $\{1\}$ in Lemma II.3 is projected out. □

Differential calculi in the original sense of Woronowicz are classified by Corollary II.2 while from the quantum tangent space point of view the classification is given by Corollary II.4. In the finite dimensional case the duality is strict in the sense of a one-to-one correspondence. The infinite dimensional case on the other hand depends strongly on the algebraic models we use for the function or enveloping algebras. It is therefore not surprising that in the present purely algebraic context the classifications are quite different in this case. We will restrict ourselves to the finite dimensional case in the following description of the differential calculi.

Theorem II.5: *(a) Finite dimensional differential calculi* Γ *on* $C_q(B_+)$ *and corresponding quantum tangent spaces L on* $U_q(\mathfrak{b}_+)$ *are in one-to-one correspondence to finite sets* $I \subset \mathbb{N} \backslash \{1\}$. *In particular* $\dim \Gamma = \dim L = \Sigma_{n \in I} n$.

(b) Coirreducible Γ *and irreducible L correspond to* $\{n\}$ *with* $n \geqslant 2$ *the dimension. Such a* Γ *has a right invariant basis* $\eta_0, \ldots, \eta_{n-1}$ *so that the relations*

$$\mathrm{d}X = \eta_1 + (q^{n-1}-1)\eta_0 X, \quad \mathrm{d}g = (q^{n-1}-1)\eta_0 g,$$

$$[a, \eta_0] = \mathrm{d}a \quad \forall a \in C_q(B_+),$$

$$[g, \eta_i]_{q^{n-1-i}} = 0 \quad \forall i, \quad [X, \eta_i]_{q^{n-1-i}} = \begin{cases} \eta_{i+1} & if \quad i < n-1 \\ 0 & if \quad i = n-1 \end{cases}$$

hold, where $[a,b]_p := ab - pba$. *By choosing the dual basis on the corresponding irreducible L we obtain the braided derivations*

$$\partial_i:f:=:Q_{n-1-i,g}Q_{n-1-i,X}\frac{1}{[i]_q!}(\partial_{q,X})^i f: \quad \forall i\geq 1,$$

$$\partial_0:f:=:Q_{n-1,g}Q_{n-1,X}f-f:$$

for $f\in \Bbbk(q)[X,g,g^{-1}]$ *with normal ordering* $\Bbbk(q)[X,g,g^{-1}]\to C_q(B_+)$ *given by* $g^nX^m\mapsto g^nX^m$.

(c) Finite dimensional Γ *and* L *decompose into direct sums of coirreducible, respectively, irreducible ones. In particular* $\Gamma=\oplus_{n\in I}\Gamma^n$ *and* $L=\oplus_{n\in I}L^n$ *with* Γ^n *and* L^n *corresponding to* $\{n\}$.

Proof: (a) We observe that the classifications of Lemma II.1 and Lemma II.3 or Corollary II.2 and Corollary II.4 are dual to each other in the finite (co)dimensional case. More precisely, for $I\subset\mathbb{N}$ finite the crossed submodule M corresponding to $(1,I)$ in Lemma II.1 is the annihilator of the crossed submodule L corresponding to I in Lemma II.3 and vice versa. $C_q(B_+)/M$ and L are dual spaces with the induced pairing. For $I\subset\mathbb{N}\setminus\{1\}$ finite this descends to M corresponding to $(1,I\cup\{1\})$ in Corollary II.2 and L corresponding to I in Corollary II.4. For the dimension of Γ observe $\dim\Gamma=\dim\ker\epsilon/M=\operatorname{codim} M$.

(b) Coirreducibility (having no proper quotient) of Γ clearly corresponds to maximality of M. The statement then follows from parts (b) of Corollaries II.2 and II.4. The formulas are obtained by choosing the basis $\eta_0,\ldots,\eta_{n-1}$ of $\ker\epsilon/M$ as the equivalence classes of

$$(g-1)/(q^{n-1}-1),X,\ldots,X^{n-1}.$$

The dual basis of L is then given by

$$g^{1-n}-1,Xg^{1-n},\ldots,q^{k(k-1)}\frac{1}{[k]_q!}X^kg^{1-n},\ldots,q^{(n-1)(n-2)}\frac{1}{[n-1]_q!}X^{n-1}g^{1-n}.$$

(c) The statement follows from Corollaries II.2 and II.4 parts (c) with the observation

$$\ker\epsilon/M=\ker\epsilon/\bigcap_{n\in I}M^n=\oplus_{n\in I}\ker\epsilon/M^n.$$

□

Corollary II.6: There is precisely one differential calculus on $C_q(B_+)$ *which is natural in the sense that it has dimension 2. It is coirreducible and obeys the relations*

$$[g,\mathrm{d}X]=0,\quad [g,\mathrm{d}g]_q=0,\quad [X,\mathrm{d}X]_q=0,\quad [X,\mathrm{d}g]_q=(q-1)(\mathrm{d}X)g,$$

with $[a,b]_q:=ab-qba$. In particular we have

$$\mathrm{d}:f:=\mathrm{d}g:\partial_{q,g}f:+\mathrm{d}X:\partial_{q,X}f:\quad \forall f\in\Bbbk(q)[X,g,g^{-1}].$$

Proof: This is a special case of Theorem II.5. The formulas follow from (b) with $n=2$. □

III. CLASSIFICATION IN THE CLASSICAL LIMIT

In this section we give the complete classification of differential calculi and quantum tangent spaces in the classical case of $C(B_+)$ along the lines of the previous section. We pay particular attention to the relation to the q-deformed setting.

The classical limit $C(B_+)$ of the quantum group $C_q(B_+)$ is simply obtained by substituting the parameter q with 1. The classification of left crossed submodules in part (a) of Lemma II.1 remains unchanged, as one may check by going through the proof. In particular, we get a correspondence of crossed modules in the q-deformed setting with crossed modules in the classical setting as a map of pairs $(P,I)\mapsto(P,I)$ that converts polynomials $\Bbbk(q)[g]$ to polynomials $\Bbbk[g]$ (if defined) and leaves sets I unchanged. This is one-to-one in the finite dimensional case. However, we did use the distinctness of powers of q in part (b) and (c) of Lemma II.1 and have to account for changing this. The only place where we used it, was in observing that factors $1-q^j g$ have no common divisors for distinct j. This was crucial to conclude the maximality (b) of certain finite codimensional crossed submodules and the intersection property (c). Now, all those factors become $1-g$.

Corollary III.1: (a) Left crossed $C(B_+)$-submodules $M\subseteq C(B_+)$ by left multiplication and left adjoint coaction are in one-to-one correspondence to pairs (P,I), where $P\in \Bbbk[g]$ is a polynomial with $P(0)=1$ and $I\subset\mathbb{N}$ is finite. $\operatorname{codim} M<\infty$ iff $P=1$. *In particular* $\operatorname{codim} M=\Sigma_{n\in I}n$ *if* $P=1$.

(b) The infinite codimensional maximal M are characterized by $(P,\emptyset)$ with P irreducible and $P(g)\neq 1-g$ for any $k\in\mathbb{N}_0$.

In the restriction to $\ker\epsilon\subset C(B_+)$ corresponding to Corollary II.2 we observe another difference to the q-deformed setting. Since the condition for a crossed submodule to lie in $\ker\epsilon$ is exactly to have factors $1-g$ in the X-free monomials this condition may now be satisfied more easily. If the characterizing polynomial does not contain this factor it is now sufficient to have just any nonempty characterizing integer set I and it need not contain 1. Consequently, the map $(P,I)\mapsto(P,I)$ does not reach all crossed submodules now.

Corollary III.2: (a) Left crossed $C(B_+)$-submodules $M\subseteq\ker\epsilon\subset C(B_+)$ are in one-to-one correspondence to pairs (P,I) as in Corollary III.1 with the additional constraint $(1-g)$ divides $P(g)$ or I nonempty. $\operatorname{codim} M<\infty$ iff $P=1$. *In particular* $\operatorname{codim} M=(\Sigma_{n\in I}n)-1$ *if* $P=1$.

(b) The infinite codimensional maximal M correspond to pairs $(P,\{1\})$ with P irreducible and $P(g)\neq 1-g$.

Let us now turn to quantum tangent spaces on $U(\mathfrak{b}_+)$. Here, the process to go from the q-deformed setting to the classical one is not quite so straightforward.

Lemma III.3: Proper left crossed $U(\mathfrak{b}_+)$-submodules $L\subset U(\mathfrak{b}_+)$ via the left adjoint action and left regular coaction are in one-to-one correspondence to pairs (l,I) with $l\in\mathbb{N}_0$ and $I\subset\mathbb{N}$ finite. $\dim L<\infty$ iff $l=0$. *In particular* $\dim L=\Sigma_{n\in I}n$ *if* $l=0$.

Proof: The left adjoint action takes the form

$$X\triangleright X^nH^m=X^{n+1}(H^m-(H+1)^m),\quad H\triangleright X^nH^m=nX^nH^m,$$

while the coaction is

$$\Delta(X^nH^m)=\sum_{i=1}^{n}\sum_{j=1}^{m}\binom{n}{i}\binom{m}{j}X^iH^j\otimes X^{n-1}H^{m-j}.$$

Let L be a crossed submodule invariant under the action and coaction. The (repeated) action of H separates elements by degree in X. It is therefore sufficient to consider elements of the form $X^nP(H)$, where P is a polynomial. By acting with X on an element $X^nP(H)$ we obtain $X^{n+1}(P(H)-P(H+1))$. Subsequently applying the coaction and projecting on the left-hand side of the tensor product onto X (in the basis X^iH^j of $U(\mathfrak{b}_+)$) leads to the element $X^n(P(H)-P(H+1))$. Now the degree of $P(H)-P(H+1)$ is exactly the degree of $P(H)$ minus 1. Thus we have polynomials $X^nP_i(H)$ of any degree $i=\deg(P_i)\leqslant\deg(P)$ in L by induction. In particular, $X^nH^m\in L$ for all $m\leqslant\deg(P)$. It is thus sufficient to consider elements of the form X^nH^m. Given such an element, the coaction generates all elements of the form X^iH^j with $i\leqslant n$, $j\leqslant m$.

For given n, the characterizing datum is the maximal m so that $X^nH^m\in L$. Due to the coaction this cannot decrease with decreasing n and due to the action of X this can decrease at most by 1 when increasing n by 1. This leads to the classification given. For $l\in\mathbb{N}_0$ and $I\subset\mathbb{N}$ finite, the corresponding crossed submodule is generated by

$$X^{n_m-1}H^{l+m-1},X^{n_m+n_{m-1}-1}H^{l+m-2},\ldots,X^{(\Sigma_i n_i)-1}H^l\quad\text{and}\quad X^{(\Sigma_i n_i)+k}H^{l-1}\quad\forall k\geqslant 0\quad\text{if}\quad l>0$$

as a crossed module. □

For the transition from the q-deformed (Lemma II.3) to the classical case we observe that the space spanned by $g^{s_1},\ldots,g^{s_m}$ with m different integers $s_i\in\mathbb{Z}$ maps to the space spanned by

$1,H,\ldots,H^{m-1}$ in the prescription of the classical limit (as described in Sec. I B, i.e., the classical crossed submodule characterized by an integer l and a finite set $I\subset\mathbb{N}$ comes from a crossed submodule characterized by this same I and additionally l other integers $j\in\mathbb{Z}$ for which $X^k g^{1-j}$ is included. In particular, we have a one-to-one correspondence in the finite dimensional case.

To formulate the analog of Corollary II.4 for the classical case is essentially straightforward now. However, as for $C(B_+)$, we obtain more crossed submodules than those from the q-deformed setting. This is due to the degeneracy introduced by forgetting the powers of g and just retaining the number of different powers.

Corollary III.4: (a) Proper left crossed $U(\mathfrak{b}_+)$*-submodules* $L\subset\ker\epsilon\subset U(\mathfrak{b}_+)$ *via the left adjoint action and left regular coaction (with subsequent projection to* $\ker\epsilon$ *via* $x\mapsto x-\epsilon(x)1$*) are in one-to-one correspondence to pairs (l,I) with* $l\in\mathbb{N}_0$ *and* $I\subset\mathbb{N}$ *finite where* $l\neq 0$ *or* $I\neq\emptyset$. $\dim L<\infty$ iff $l=0$. *In particular* $\dim L=(\Sigma_{n\in I}n)-1$ *if* $l=0$.

As in the q-deformed setting, we give a description of the finite dimensional differential calculi where we have a strict duality to quantum tangent spaces.

Proposition III.5: (a) Finite dimensional differential calculi Γ *on* $C(B_+)$ *and finite dimensional quantum tangent spaces L on* $U(\mathfrak{b}_+)$ *are in one-to-one correspondence to nonempty finite sets* $I\subset\mathbb{N}$. *In particular* $\dim\Gamma=\dim L=(\Sigma_{n\in I}n)-1$.

The Γ *with* $1\in\mathbb{N}$ *are in one-to-one correspondence to the finite dimensional calculi and quantum tangent spaces of the q-deformed setting (Theorem II.5(a)).*

(b) The differential calculus Γ *of dimension* $n\geq 2$ *corresponding to the coirreducible one of* $C_q(B_+)$ *(Theorem II.5(b)) has a right invariant basis* $\eta_0,\ldots,\eta_{n-1}$ *so that*

$$\mathrm{d}X=\eta_1+\eta_0 X,\quad \mathrm{d}g=\eta_0 g,$$

$$[g,\eta_i]=0\,\forall i,\quad [X,\eta_i]=\begin{cases}0 & \text{if } i=0 \text{ or } i=n-1\\ \eta_{i+1} & \text{if } 0<i<n-1\end{cases}$$

hold. The braided derivations obtained from the dual basis of the corresponding L are given by

$$\partial_i f=\frac{1}{i!}\left(\frac{\partial}{\partial X}\right)^i f\quad \forall i\geq 1,$$

$$\partial_0 f=\left(X\frac{\partial}{\partial X}+g\frac{\partial}{\partial g}\right)f$$

for $f\in C(B_+)$.

(c) The differential calculus of dimension $n-1$ *corresponding to the one in (b) with 1 removed from the characterizing set is the same as the one above, except that we set* $\eta_0=0$ *and* $\partial_0=0$.

Proof: (a) We observe that the classifications of Corollary III.1 and Lemma III.3 or Corollary III.2 and Corollary III.4 are dual to each other in the finite (co)dimensional case. More precisely, for $I\subset\mathbb{N}$ finite the crossed submodule M corresponding to $(1,I)$ in Corollary III.1 is the annihilator of the crossed submodule L corresponding to $(0,I)$ in Lemma III.3 and vice versa. $C(B_+)/M$ and L are dual spaces with the induced pairing. For nonempty I this descends to M corresponding to $(1,I)$ in Corollary III.2 and L corresponding to $(0,I)$ in Corollary III.4. For the dimension of Γ note $\dim\Gamma=\dim\ker\epsilon/M=\operatorname{codim}M$.

(b) For $I=\{1,n\}$ we choose in $\ker\epsilon\subset C(B_+)$ the basis $\eta_0,\ldots,\eta_{n-1}$ as the equivalence classes of $g-1,X,\ldots,X^{n-1}$. The dual basis in L is then $H,X,\ldots,(1/k!)X^k,\ldots,1/(n-1)!X^{n-1}$. This leads to the formulas given.

(c) For $I=\{n\}$ we get the same as in (b) except that η_0 and ∂_0 disappear. □

The classical commutative calculus is the special case of (b) with $n=2$. It is the only calculus of dimension 2 with $\mathrm{d}g\neq 0$. Note that it is not coirreducible.

IV. THE DUAL CLASSICAL LIMIT

We proceed in this section to the more interesting point of view where we consider the classical algebras, but with their roles interchanged, i.e., we view $U(\mathfrak{b}_+)$ as the "function algebra" and $C(B_+)$ as the "enveloping algebra." Due to the self-duality of $U_q(\mathfrak{b}_+)$, we can again view the differential calculi and quantum tangent spaces as classical limits of the q-deformed setting investigated in Sec. II.

In this dual setting the bicovariance constraint for differential calculi becomes much weaker. In particular, the adjoint action on a classical function algebra is trivial due to commutativity and the adjoint coaction on a classical enveloping algebra is trivial due to cocommutativity. In effect, the correspondence with the q-deformed setting is much weaker than in the ordinary case of Sec. III. There are much more differential calculi and quantum tangent spaces than in the q-deformed setting.

We will not attempt to classify all of them in the following but essentially contend ourselves with those objects coming from the q-deformed setting.

Lemma IV.1: Left $C(B_+)$-subcomodules $\subseteq C(B_+)$ via the left regular coaction are $\mathbb{Z}$-graded subspaces of $C(B_+)$ with $|X^n g^m| = n+m$, stable under formal derivation in X.

By choosing any ordering in $C_q(B_+)$, left crossed submodules via left regular action and adjoint coaction are in one-to-one correspondence to certain subcomodules of $C(B_+)$ by setting $q=1$. Direct sums correspond to direct sums.

This descends to $\ker\epsilon \subset C(B_+)$ *by the projection* $x \mapsto x - \epsilon(x)1$.

Proof: The coproduct on $C(B_+)$ is

$$\Delta(X^n g^k) = \sum_{r=0}^{n} \binom{n}{r} X^{n-r} g^{k+r} \otimes X^r g^k,$$

which we view as a left coaction. Projecting on the left-hand side of the tensor product onto g^l in a basis $X^n g^k$, we observe that coacting on an element $\Sigma_{n,k} a_{n,k} X^n g^k$ we obtain elements $\Sigma_n a_{n,l-n} X^n g^{l-n}$ for all l, i.e., elements of the form $\Sigma_n b_n X^n g^{l-n}$ lie separately in a subcomodule and it is sufficient to consider such elements. Writing the coaction on such an element as

$$\sum_t \frac{1}{t!} X^t g^{l-t} \otimes \sum_n b_n \frac{n!}{(n-t)!} X^{n-t} g^{l-n},$$

we see that the coaction generates all formal derivatives in X of this element. This gives us the classification, $C(B_+)$-subcomodules $\subseteq C(B_+)$ under the left regular coaction are $\mathbb{Z}$-graded subspaces with $|X^n g^m| = n+m$, stable under formal derivation in X given by $X^n g^m \mapsto n X^{n-1} g^m$.

The correspondence with the $C_q(B_+)$ case follows from the trivial observation that the coproduct of $C(B_+)$ is the same as that of $C_q(B_+)$ with $q=1$.

The restriction to $\ker\epsilon$ is straightforward. □

Lemma IV.2: The process of obtaining the classical limit $U(\mathfrak{b}_+)$ from $U_q(\mathfrak{b}_+)$ is well defined for subspaces and sends crossed $U_q(\mathfrak{b}_+)$-submodules $\subset U_q(\mathfrak{b}_+)$ by regular action and adjoint coaction to $U(\mathfrak{b}_+)$-submodules $\subset U(\mathfrak{b}_+)$ by regular action. This map is injective in the finite codimensional case. Intersections and codimensions are preserved in this case.

This descends to $\ker\epsilon$.

Proof: To obtain the classical limit of a left ideal it is enough to apply the limiting process (as described in Sec. I B) to the module generators. (We can forget the additional comodule structure.) On the one hand, any element generated by left multiplication with polynomials in g corresponds to some element generated by left multiplication with a polynomial in H, that is, there will be no more generators in the classical setting. On the other hand, left multiplication by a polynomial in H comes from left multiplication by the same polynomial in $g-1$, that is, there will be no fewer generators.

The maximal left crossed $U_q(\mathfrak{b}_+)$-submodule $\subseteq U_q(\mathfrak{b}_+)$ by left multiplication and adjoint coaction of codimension n $(n\geqslant 1)$ is generated as a left ideal by $\{1-q^{1-n}g,X^n\}$ (see Lemma II.1). Applying the limiting process to this leads to the left ideal of $U(\mathfrak{b}_+)$ (which is not maximal for $n\neq 1$) generated by $\{H+n-1,X^n\}$ having also codimension n.

More generally, the picture given for arbitrary finite codimensional left crossed modules of $U_q(\mathfrak{b}_+)$ in terms of generators with respect to polynomials in g,g^{-1} in Lemma II.1 carries over by replacing factors $1-q^{1-n}g$ with factors $H+n-1$ leading to generators with respect to polynomials in H. In particular, intersections go to intersections since the distinctness of the factors for different n is conserved.

The restriction to $\ker\epsilon$ is straightforward. □

We are now in a position to give a detailed description of the differential calculi induced from the q-deformed setting by the limiting process.

Proposition IV.3: (a) Certain finite dimensional differential calculi Γ *on* $U(\mathfrak{b}_+)$ *and quantum tangent spaces L on* $C(B_+)$ *are in one-to-one correspondence to finite dimensional differential calculi on* $U_q(\mathfrak{b}_+)$ *and quantum tangent spaces on* $C_q(B_+)$. *Intersections correspond to intersections.*

(b) In particular, Γ *and L corresponding to coirreducible differential calculi on* $U_q(\mathfrak{b}_+)$ *and irreducible quantum tangent spaces on* $C_q(B_+)$ *via the limiting process are given as follows:* Γ *has a right invariant basis* $\eta_0,\ldots,\eta_{n-1}$ *so that*

$$\mathrm{d}X=\eta_1,\quad \mathrm{d}H=(1-n)\eta_0,$$

$$[H,\eta_i]=(1-n+i)\eta_i\ \ \forall i,\quad [X,\eta_i]=\begin{cases}\eta_{i+1} & \text{if } i<n-1\\ 0 & \text{if } i=n-1\end{cases}$$

holds. The braided derivations corresponding to the dual basis of L are given by

$$\partial_i:f:=:T_{1-n+i,H}\frac{1}{i!}\left(\frac{\partial}{\partial X}\right)^i f:\quad \forall i\geqslant 1,$$

$$\partial_0:f:=:T_{1-n,H}f-f:$$

for $f\in \Bbbk[X,H]$ *with the normal ordering* $\Bbbk[X,H]\to U(\mathfrak{b}_+)$ *via* $H^nX^m\mapsto H^nX^m$.

Proof: (a) The strict duality between $C(B_+)$-subcomodules $L\subseteq\ker\epsilon$ given by Lemma IV.1 and Corollary II.4 and $U(\mathfrak{b}_+)$-modules $U(\mathfrak{b}_+)/(\Bbbk 1+M)$ with M given by Lemma IV.2 and Corollary II.2 can be checked explicitly. It is essentially due to mutual annihilation of factors $H+k$ in $U(\mathfrak{b}_+)$ with elements g^k in $C(B_+)$.

(b) L is generated by $\{g^{1-n}-1,Xg^{1-n},\ldots,X^{n-1}g^{1-n}\}$ and M is generated by $\{H(H+n-1),X(H+n-1),X^n\}$. The formulas are obtained by denoting with $\eta_0,\ldots,\eta_{n-1}$ the equivalence classes of $H/(1-n),X,\ldots,X^{n-1}$ in $U(\mathfrak{b}_+)/(\Bbbk 1+M)$. The dual basis of L is then

$$g^{1-n}-1,Xg^{1-n},\ldots,\frac{1}{(n-1)!}X^{n-1}g^{1-n}.$$

□

In contrast to the q-deformed setting and to the usual classical setting the many freedoms in choosing a calculus leave us with many two-dimensional calculi. It is not obvious which one we should consider to be the ''natural'' one. Let us first look at the two-dimensional calculus coming from the q-deformed setting as described in (b). The relations become

$$[\mathrm{d}H,a]=da,\quad [\mathrm{d}X,a]=0,\quad \forall a\in U(\mathfrak{b}_+),$$

$$\mathrm{d}:f:=\mathrm{d}H:\nabla_{1,H}f:+\mathrm{d}X:\frac{\partial}{\partial X}f:$$

for $f\in\mathbb{k}[X,H]$.

We might want to consider calculi which are closer to the classical theory in the sense that derivatives are not finite differences but usual derivatives. Let us therefore demand

$$\mathrm{d}P(H)=\mathrm{d}H\,\frac{\partial}{\partial H}P(H)\ \text{ and }\ \mathrm{d}P(X)=\mathrm{d}X\,\frac{\partial}{\partial X}P(X)$$

for polynomials P and $\mathrm{d}X\neq 0$ and $\mathrm{d}H\neq 0$.

Proposition IV.4: There is precisely one differential calculus of dimension 2 meeting these conditions. It obeys the relations

$$[a,\mathrm{d}H]=0,\quad [X,\mathrm{d}X]=0,\quad [H,\mathrm{d}X]=\mathrm{d}X,$$

$$\mathrm{d}{:}f{:}=\mathrm{d}H{:}\,\frac{\partial}{\partial H}f{:}+\mathrm{d}X{:}\,\frac{\partial}{\partial X}f{:}$$

where the normal ordering $\mathbb{k}[X,H]\to U(\mathfrak{b}_+)$ *is given by* $X^nH^m\mapsto X^nH^m$.

Proof: Let M be the left ideal corresponding to the calculus. It is easy to see that for a primitive element a the classical derivation condition corresponds to $a^2\in M$ and $a\notin M$. In our case $X^2,H^2\in M$. If we take the ideal generated from these two elements we obtain an ideal of $\ker\epsilon$ of codimension 3. Now, it is sufficient without loss of generality to add a generator of the form $\alpha H+\beta X+\gamma XH$. α and β must then be zero in order not to generate X or H in M, i.e., M is generated by H^2,XH,X^2. The relations stated follow. □

V. REMARKS ON κ-MINKOWSKI SPACE AND INTEGRATION

There is a straightforward generalization of $U(\mathfrak{b}_+)$. Let us define the Lie algebra $\mathfrak{b}_{n+}$ as generated by $x_0,\ldots,x_{n-1}$ with relations

$$[x_0,x_i]=x_i,\quad [x_i,x_j]=0\quad \forall i,j\geqslant 1.$$

Its enveloping algebra $U(\mathfrak{b}_{n+})$ is nothing but (rescaled) κ-Minkowski space as introduced in Ref. 10. In this section we make some remarks about its intrinsic geometry.

We have an injective Lie algebra homomorphism $b_{n+}\to b_+$ given by $x_0\mapsto H$ and $x_i\mapsto X$. This is an isomorphism for $n=2$. The injective Lie algebra homomorphism extends to an injective homomorphism of enveloping algebras $U(\mathfrak{b}_+)\to U(\mathfrak{b}_{n+})$ in the obvious way. This gives rise to an injective map from the set of submodules of $U(\mathfrak{b}_+)$ to the set of submodules of $U(\mathfrak{b}_{n+})$ by taking the preimage. In particular this induces an injective map from the set of differential calculi on $U(\mathfrak{b}_+)$ to the set of differential calculi on $U(\mathfrak{b}_{n+})$ which are invariant under permutations of the x_i $i\geqslant 1$.

Corollary V.1: There is a natural n-dimensional differential calculus on $U(\mathfrak{b}_{n+})$ *induced from the one considered in Proposition IV.4. It obeys the relations*

$$[a,\mathrm{d}x_0]=0\quad \forall a\in U(\mathfrak{b}_{n+}),\quad [x_i,\mathrm{d}x_j]=0,\quad [x_0,\mathrm{d}x_i]=\mathrm{d}x_i\quad \forall i,j\geqslant 1$$

$$\mathrm{d}{:}f{:}=\sum_{\mu=0}^{n-1}dx_\mu{:}\,\frac{\partial}{\partial x_\mu}f{:}$$

where the normal ordering is given by

$$\mathbb{k}[x_0,\ldots,x_{n-1}]\to U(\mathfrak{b}_{n+1})\ \text{ via }\ x_{n-1}^{m_{n-1}}\cdots x_0^{m_0}\mapsto x_{n-1}^{m_{n-1}}\cdots x_0^{m_0}.$$

Proof: The calculus is obtained from the ideal generated by

$$x_0^2,x_ix_j,x_ix_0\quad \forall i,j\geqslant 1$$

being the preimage of X^2, XH, X^2 in $U(\mathfrak{b}_+)$. □

Let us try to push the analogy with the commutative case further and take a look at the notion of integration. The natural way to encode the condition of translation invariance from the classical context in the quantum group context is given by the condition

$$\left(\int \otimes \mathrm{id}\right)\circ\Delta a = 1\int a \quad \forall a\in A$$

which defines a right integral on a Hopf algebra A.[13] (Correspondingly, we have the notion of a left integral.) Let us formulate a slightly weaker version of this equation in the context of a Hopf algebra H dually paired with A. We write

$$\int (h-\epsilon(h))\triangleright a = 0 \quad \forall h\in H, a\in A,$$

where the action of H on A is the coregular action $h\triangleright a = a_{(1)}\langle a_{(2)}, h\rangle$ given by the pairing.

In the present context we set $A=U(\mathfrak{b}_{n+})$ and $H=C(B_{n+})$. We define the latter as a generalization of $C(B_+)$ with commuting generators $g, p_1, \ldots, p_{n-1}$ and coproducts

$$\Delta p_i = p_i\otimes 1 + g\otimes p_i, \quad \Delta g = g\otimes g.$$

This can be identified (upon rescaling) as the momentum sector of the full κ-Poincaré algebra (with $g=e^{p_0}$). The pairing is the natural extension of (1),

$$\langle x_{n-1}^{m_{n-1}}\cdots x_1^{m_1}x_0^k, p_{n-1}^{r_{n-1}}\cdots p_1^{r_1}g^s\rangle = \delta_{m_{n-1},r_{n-1}}\cdots\delta_{m_1,r_1} m_{n-1}!\cdots m_1! s^k.$$

The resulting coregular action is conveniently expressed as (see also Ref. 10)

$$p_i\triangleright :f: = :\frac{\partial}{\partial x_i}f:, \quad g\triangleright :f: = :T_{1,x_0}f:$$

with $f\in\mathbb{k}[x_0,\ldots,x_{n-1}]$. Due to cocommutativity, the notions of left and right integral coincide. The invariance conditions for integration become

$$\int :\frac{\partial}{\partial x_i}f: = 0 \quad \forall i\in\{1,\ldots,n-1\} \text{ and } \int :\nabla_{1,x_0}f: = 0.$$

The condition on the left is familiar and states the invariance under infinitesimal translations in the x_i. The condition on the right states the invariance under integer translations in x_0. However, we should remember that we use a certain algebraic model of $C(B_{n+})$. We might add, for example, a generator p_0 to $C(B_{n+})$ that is dual to x_0 and behaves as the ''logarithm'' of g, i.e., acts as an infinitesimal translation in x_0. We then have the condition of infinitesimal translation invariance

$$\int :\frac{\partial}{\partial x_\mu}f: = 0$$

for all $\mu\in\{0,1,\ldots,n-1\}$.

In the present purely algebraic context these conditions do not make much sense. In fact they would force the integral to be zero on the whole algebra. This is not surprising, since we are dealing only with polynomial functions which would not be integrable in the classical case either. In contrast, if we had for example the algebra of smooth functions in two real variables, the conditions just characterize the usual Lesbegue integral (up to normalization). Let us assume $\mathbb{k}=\mathbb{R}$ and suppose that we have extended the normal ordering vector space isomorphism $\mathbb{R}[x_0,\ldots,x_{n-1}]\cong U(\mathfrak{b}_{n+})$ to a vector space isomorphism of some sufficiently large class of functions on $\mathbb{R}^n$ with a suitable completion $\hat{U}(\mathfrak{b}_{n+})$ in a functional analytic framework (embedding $U(\mathfrak{b}_{n+})$ in some operator algebra on a Hilbert space). It is then natural to define the integration on $\hat{U}(\mathfrak{b}_{n+})$ by

$$\int :f:=\int_{\mathbb{R}^n} f\, dx_0 \cdots dx_{n-1},$$

where the right-hand side is just the usual Lesbegue integral in n real variables $x_0,\ldots,x_{n-1}$. This integral is unique (up to normalization) in satisfying the covariance condition since, as we have seen, these correspond just to the usual translation invariance in the classical case via normal ordering, for which the Lesbegue integral is the unique solution. It is also the $q\to 1$ limit of the translation invariant integral on $U_q(\mathfrak{b}_+)$ obtained in Ref. 14.

We see that the natural differential calculus in Corollary V.1 is compatible with this integration in that the appearing braided derivations are exactly the actions of the translation generators p_μ. However, we should stress that this calculus is not covariant under the full κ-Poincaré algebra, since it was shown in Ref. 15 that in $n=4$ there is no such calculus of dimension 4. Our results therefore indicate a new intrinsic approach to κ-Minkowski space that allows a bicovariant differential calculus of dimension 4 and a unique translation invariant integral by normal ordering and Lesbegue integration.

ACKNOWLEDGMENTS

I would like to thank S. Majid for proposing this project, and for fruitful discussions during the preparation of this paper.

APPENDIX: THE ADJOINT COACTION ON $U_q(\mathfrak{b}_+)$

The coproduct on X^n is

$$\Delta(X^n)=\sum_{r=0}^{n}\begin{bmatrix} n \\ r \end{bmatrix}_q g^r X^{n-r}\otimes X^r,$$

$$(\mathrm{id}\otimes\Delta)\Delta(X^n)=\sum_{r=0}^{n}\sum_{i=0}^{r}\begin{bmatrix} n \\ r \end{bmatrix}_q \begin{bmatrix} r \\ i \end{bmatrix}_q g^r X^{n-r}\otimes g^i X^{r-i}\otimes X^i.$$

From this we get

$$\begin{aligned}
\mathrm{Ad_L}(X^n) &= \sum_{r=0}^{n}\sum_{s=0}^{r}\begin{bmatrix} n \\ r \end{bmatrix}_q \begin{bmatrix} r \\ s \end{bmatrix}_q g^r X^{n-r} \mathrm{S}X^s \otimes g^s X^{r-s} \\
&= \sum_{r=0}^{n}\sum_{s=0}^{r}\begin{bmatrix} n \\ r \end{bmatrix}_q \begin{bmatrix} r \\ s \end{bmatrix}_q g^r X^{n-r}(-g^{-1}X)^s \otimes g^s X^{r-s} \\
&= \sum_{t=0}^{n}\sum_{s=0}^{n-t}\begin{bmatrix} n \\ t+s \end{bmatrix}_q \begin{bmatrix} t+s \\ s \end{bmatrix}_q g^{t+s} X^{n-t-s}(-g^{-1}X)^s \otimes g^s X^{t} \\
&= \sum_{t=0}^{n}\sum_{s=0}^{n-t}\begin{bmatrix} n \\ t \end{bmatrix}_q \begin{bmatrix} n-t \\ s \end{bmatrix}_q g^{t+s} X^{n-t-s}(-g^{-1}X)^s \otimes g^s X^{t} \\
&= \sum_{t=0}^{n}\begin{bmatrix} n \\ t \end{bmatrix}_q g^t X^{n-t}\otimes X^t \sum_{s=0}^{n-t}\begin{bmatrix} n-t \\ s \end{bmatrix}_q q^{s(s+1)/2}(-q^{-n}g)^s \\
&= \sum_{t=0}^{n}\begin{bmatrix} n \\ t \end{bmatrix}_q g^t X^{n-t}\otimes X^t \prod_{u=1}^{n-t}(1-q^{u-n}g),
\end{aligned}$$

where we have used

$$\sum_{i=0}^{n} \begin{bmatrix} n \\ i \end{bmatrix}_q q^{i(i+1)/2} x^i = \prod_{j=1}^{n} (1+q^j x),$$

which can be easily checked by induction. Using the property

$$\mathrm{Ad_L}(ag^n) = \mathrm{Ad_L}(a)(1 \otimes g^n) \quad \forall n \in \mathbb{Z},$$

we obtain for any polynomial P in g, g^{-1},

$$\mathrm{Ad_L}(X^n P(g)) = \sum_{t=0}^{n} \begin{bmatrix} n \\ t \end{bmatrix}_q g^t X^{n-t} \otimes X^t P(g) \prod_{u=1}^{n-t} (1 - q^{u-n} g). \tag{A1}$$

[1] S. L. Woronowicz, ''Differential calculus on compact matrix pseudogroups (quantum groups),'' Commun. Math. Phys. **122**, 125–170 (1989).
[2] B. Jurco, ''Differential calculus on quantized simple Lie groups,'' Lett. Math. Phys. **22**, 177–186 (1991).
[3] U. Carow-Watamura, M. Schlieker, S. Watamura, and W. Weich, ''Bicovariant differential calculus on quantum groups $SU_q(N)$ and $SO_q(N)$,'' Commun. Math. Phys. **142**, 605–641 (1991).
[4] K. Schmüdgen and A. Schüler, ''Classification of bicovariant differential calculi on quantum groups,'' Commun. Math. Phys. **170**, 315–335 (1995).
[5] S. Majid, ''Classification of bicovariant differential calculi,'' J. Geom. Phys. **25**, 119–140 (1998).
[6] P. Baumann and F. Schmitt, ''Classification of bicovariant differential calculi on quantum groups (a representation-theoretic approach),'' Commun. Math. Phys. **194**, 71–86 (1998).
[7] I. Heckenberger and K. Schmüdgen, ''Bicovariant differential calculi on $SL_q(N)$ and $Sp_q(N)$,'' Czech. J. Phys. **47**, 1145–1151 (1997).
[8] V. G. Drinfeld, ''Quantum groups,'' in *Proceedings of the International Congress of Mathematicians* (American Mathematical Society, Providence, Rhode Island, 1987).
[9] J. Lukierski, A. Nowicki, and H. Ruegg, ''New quantum Poincaré algebra and κ-deformed field theory,'' Phys. Lett. B **293**, 344–352 (1992).
[10] S. Majid and H. Ruegg, ''Bicrossproduct structure of κ-Poincaré group and noncommutative geometry,'' Phys. Lett. B **334**, 348–352 (1994).
[11] S. Majid, ''Hopf algebras for physics at the planck scale,'' Class. Quantum Grav. **5**, 1587–1606 (1988).
[12] S. Majid, *Foundations of Quantum Group Theory* (Cambridge University Press, Cambridge, 1995).
[13] M. E. Sweedler, *Hopf Algebras* (Benjamin, New York, 1969).
[14] S. Majid, ''On q-regularization,'' Int. J. Mod. Phys. A **5**, 4689–4696 (1990).
[15] C. Gonera, P. Kosinski, and P. Maslanka, ''Differential calculi on quantum Minkowski space,'' J. Math. Phys. **37**, 5820–5827 (1996).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Representations of the Weyl group and Wigner functions for SU(3)

D. J. Rowe and B. C. Sanders[a)]
Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada

H. de Guise
Centre de Recherches Mathématiques, Unversité de Montréal,
C.P. 6128 Succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada

(Received 30 September 1998; accepted for publication 11 November 1998)

Bases for SU(3) irreps are constructed on a space of three-particle tensor products of two-dimensional harmonic oscillator wave functions. The Weyl group is represented as the symmetric group of permutations of the particle coordinates of these spaces. Wigner functions for SU(3) are expressed as products of SU(2) Wigner functions and matrix elements of Weyl transformations. The constructions make explicit use of dual reductive pairs which are shown to be particularly relevant to problems in optics and quantum interferometry. © *1999 American Institute of Physics.* [S0022-2488(99)01506-6]

I. INTRODUCTION

Considerable progress has been made in the development of systematic algorithms for computing matrix elements of the infinitesimal generators of Lie groups in an arbitrary representation. Much less is known about the matrices of finite group elements other than those of SU(2), and the related groups E(2), HW(1), and SU(1,1).[1]

The matrix elements of finite SU(2) transformations are the well-known Wigner $\mathcal{D}$ functions. These functions are used in many areas of physics, notably in nuclear, atomic, and molecular spectroscopy. Recently, it has been shown that the Wigner functions of SU(2) (Ref. 2) and higher unitary groups[3] are needed in the analysis of quantum interferometers. Because of the Peter–Weyl theorem, Wigner functions also play a central role in the theory of harmonic analysis.

We consider here the Wigner functions for SU(3); such functions are needed, for example, in computing SU(3) Clebsch–Gordan coefficients in an SO(3) basis.[4] Expressions for SU(3) Wigner functions were first derived, to our knowledge, by Chacón and Moshinsky,[5] in terms of SU(2) Wigner functions and matrix elements of Weyl reflection operators. Matrix elements of some Weyl reflections were derived by Macfarlane *et al.*[6] and Mukunda and Pandit.[7] The latter gave the matrix elements as products of three SU(2) Clebsch–Gordan coefficients. Chacón and Moshinsky gave expressions for matrix elements of other Weyl reflections as SU(2) Racah coefficients. These results raise the question: what does the Weyl group have to do with SU(2)? The answer appears to be that basis states for SU(3) irreps (irreducible representations) are naturally expressed in an SU(2)-coupled basis, and elements of the Weyl group for SU(3), which is isomorphic to the permutation group S_3, act on such states as SU(2) recoupling operators. More explicitly, if one constructs basis states for SU(3) by SU(2) coupling the wave functions for three particles in two-dimensional harmonic oscillator states, then the Weyl reflection operators permute the coordinates of the particles. A similar interpretation of the Weyl reflections was given by Gal and Lipkin[8] as the permutations of a coupled system of three spin-$\frac{1}{2}$ quarks.

In deriving our results, we make use of two mutually commuting subgroups, U(3) and U(2), of U(6). When acting within the space of a fully symmetric representation of U(6), these subgroups are said to form a *dual reductive pair*.[9] Such dual pairs are particularly relevant for

[a)]Permanent address: Department of Physics, Macquarie University, Sydney, New South Wales 2109, Australia.

describing the properties of three particles in a two-dimensional harmonic oscillator or three spin-half quarks. An overview of these and other dual pairs and their uses in optics and quantum interferometry is given in the Discussion section at the end of this paper.

II. PARAMETRIZATION OF SU(3)

Many parametrizations of SU(3) elements are possible. The most useful ones would appear to arise from factorization of SU(3) group elements into products of subgroup elements whose Wigner functions are known. Three obvious candidates for suitable subgroups are the groups $SU(2)_{12}$, $SU(2)_{13}$, and $SU(2)_{23}$, the three SU(2) subgroups whose root systems are subsystems of the SU(3) root system shown in Fig. 1. We denote an element of $SU(2)_{ij}$ by $R_{ij}(\alpha,\beta,\gamma)$, where (α,β,γ) are the standard Euler angles.

Murnaghan[10] has shown that a possible parametrization of an element $g\in SU(3)$ is given by

$$g(\alpha_1,\beta_1,\alpha_2,\beta_2,\alpha_3,\beta_3,\delta_1,\delta_2)=e^{-i(h_1\delta_1+h_2\delta_2)}R_{23}(\alpha_1/2,\beta_1,-\alpha_1/2) \times R_{13}(\alpha_2/2,\beta_2,-\alpha_2/2)R_{12}(\alpha_3/2,\beta_3,-\alpha_3/2), \tag{1}$$

where h_1 and h_2 are elements of the Cartan subalgebra.

A similar parametrization, with a different ordering, was proposed by Reck *et al.*[3] These authors showed that one can factor a general $N\times N$ unitary matrix as a product of U(2) matrices and an overall phase, with the added insight that each U(2) transformation can be realized experimentally as an optical element.

In this paper, we choose a parametrization that takes advantage of the fact that, in a canonical basis, one constructs $U(N)$ irreps in a basis that reduces a particular $U(N-1)$ subgroup. Thus, an arbitrary $SU(N)$ matrix is factored

$$\left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & X_{N-1} & \\ 0 & & & \end{array}\right)\left(\begin{array}{cc|c} e^{i\alpha}\cos(\beta/2) & -\sin(\beta/2) & 0 \\ \sin(\beta/2) & e^{-i\alpha}\cos(\beta/2) & \\ \hline & 0 & I_{N-2} \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & Y_{N-1} & \\ 0 & & & \end{array}\right) \tag{2}$$

where X_{N-1} and Y_{N-1} are $SU(N-1)$ matrices; I_{N-2} is the $(N-2)\times(N-2)$ identity matrix. For SU(2) [with the indices ordered (z,x,y)] this gives the usual factorization $R(\alpha,\beta,\gamma) =R_z(\alpha)R_y(\beta)R_z(\gamma)$. For $g\in SU(3)$, we obtain

$$g(\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\alpha_3,\beta_3,\gamma_3)=R_{23}(\alpha_1,\beta_1,\gamma_1)R_{12}(\alpha_2,\beta_2,\alpha_2)R_{23}(\alpha_3,\beta_3,\gamma_3). \tag{3}$$

The parameters in this expression are derived for an arbitrary $g\in SU(3)$ in the Appendix, by a method communicated to us by J. Repka.

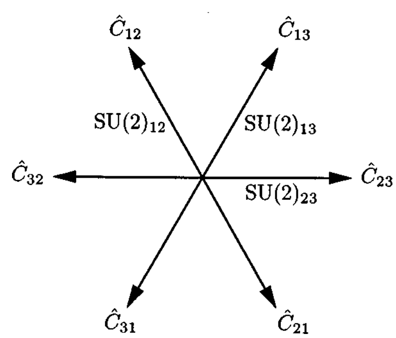

FIG. 1. Three SU(2) subsystems of the SU(3) root system.

All of the above factorizations enable one to express the SU(3) Wigner functions in terms of matrix elements of finite SU(2) transformations.

III. BASIS STATES

A. Highest weight states

An SU(3) irrep is characterized by a highest weight (λ,μ) and a corresponding highest weight state $|\phi(\lambda,\mu)\rangle$, defined as follows. The su(3) Lie algebra is spanned in the usual way by the subset of u(3) operators

$$\begin{aligned}&\hat{C}_{ij}, \quad i<j \ \text{ raising operators},\\ &\hat{C}_{ij}, \quad i>j \ \text{ lowering operators},\\ &\hat{h}_1=\hat{C}_{11}-\hat{C}_{22}, \quad \hat{h}_2=\hat{C}_{22}-\hat{C}_{33}, \quad \text{Cartan operators},\end{aligned} \tag{4}$$

where the $\{\hat{C}_{ij}\}$ operators satisfy the commutation relations

$$[\hat{C}_{ij},\hat{C}_{kl}]=\delta_{jk}\hat{C}_{il}-\delta_{il}\hat{C}_{kj}. \tag{5}$$

The highest weight state $|\phi(\lambda,\mu)\rangle$ then satisfies the equations

$$\begin{aligned}&\hat{C}_{ij}|\phi(\lambda,\mu)\rangle=0, \quad i<j,\\ &\hat{h}_1|\phi(\lambda,\mu)\rangle=\lambda|\phi(\lambda,\mu)\rangle, \quad \hat{h}_2|\phi(\lambda,\mu)\rangle=\mu|\phi(\lambda,\mu)\rangle.\end{aligned} \tag{6}$$

Without loss of generality, we suppose that $|\phi(\lambda,\mu)\rangle$ is also an eigenstate of the operator $\hat{C}_{33}$ with zero eigenvalue. It then satisfies the equations

$$\hat{C}_{11}|\phi(\lambda,\mu)\rangle=(\lambda+\mu)|\phi(\lambda,\mu)\rangle, \quad \hat{C}_{22}|\phi(\lambda,\mu)\rangle=\mu|\phi(\lambda,\mu)\rangle, \quad \hat{C}_{33}|\phi(\lambda,\mu)\rangle=0. \tag{7}$$

The Hilbert space, $\mathbb{H}^{(\lambda,\mu)}$, for the SU(3) irrep with highest weight (λ,μ) thereby becomes a Hilbert space for a U(3) irrep of highest weight $(\lambda+\mu,\mu,0)$.

B. The Gel'fand–Tsetlin basis

To use the factorization of Eq. (3) in computing Wigner functions, we need a basis for the Hilbert space $\mathbb{H}^{(\lambda,\mu)}$ that reduces the $\mathrm{SU}(3)\supset\mathrm{SU}(2)_{23}$ subgroup chain. Such a basis is the so-called canonical or Gel'fand–Tsetlin basis;[11]

$$\left\{\left|\begin{matrix}p & q\\ & r\end{matrix}\right\rangle \equiv \left|\begin{matrix}\lambda+\mu & \mu & 0\\ & p \quad q & \\ & r & \end{matrix}\right\rangle; \begin{matrix}\lambda+\mu\geqslant p\geqslant\mu\geqslant q\geqslant 0\\ p\geqslant r\geqslant q\end{matrix}\right\}, \tag{8}$$

which reduces the chain

$$\begin{matrix}\mathrm{U}(3)\\ (\lambda+\mu,\mu,0)\end{matrix} \supset \begin{matrix}\mathrm{U}(2)_{23}\\ (p,q)\end{matrix} \supset \begin{matrix}\mathrm{U}(1)_3\\ r\end{matrix}, \tag{9}$$

where $\mathrm{U}(1)_3\subset\mathrm{U}(2)_{23}$ is the subgroup whose Lie algebra is spanned by $\hat{C}_{33}$.

The Gel'fand states are eigenstates of the weight operators; i.e.,

$$\hat{C}_{ii}\left|\begin{matrix}p & q\\ & r\end{matrix}\right\rangle=\nu_i\left|\begin{matrix}p & q\\ & r\end{matrix}\right\rangle, \quad i=1,2,3, \tag{10}$$

with

$$\nu_1=\lambda+2\mu-p-q, \quad \nu_2=p+q-r, \quad \nu_3=r. \tag{11}$$

One sees that the components of a weight $\nu=(\nu_1,\nu_2,\nu_3)$ add up to $\lambda+2\mu$. They are linearly dependent and insufficient to define a state uniquely. However, the Gel'fand-Tsetlin states also reduce the subgroup chain

$$\begin{matrix} \mathrm{U}(3) \\ (\lambda+\mu,\mu,0) \end{matrix} \supset \begin{matrix} \mathrm{SU}(2)_{23} \\ I \end{matrix} \supset \begin{matrix} \mathrm{U}(1)_{23} \\ M \end{matrix}, \tag{12}$$

and have $\mathrm{SU}(2)_{23}$ quantum numbers, I and M, related to p, q, and r by

$$I=\tfrac{1}{2}(p-q), \quad M=\tfrac{1}{2}(\nu_2-\nu_3)=\tfrac{1}{2}(p+q)-r. \tag{13}$$

Thus, the weight ν and the $\mathrm{SU}(2)_{23}$ angular momentum I together uniquely define a basis state and, with the above relationships between ν, I and p, q, r, we can relabel a Gel'fand–Tsetlin state

$$|\nu I\rangle \equiv \left| \begin{matrix} p & & q \\ & r & \end{matrix} \right\rangle. \tag{14}$$

We shall refer to the basis $\{|\nu I\rangle\}$ either as a Gel'fand–Tsetlin basis or as a weight basis.

C. An SU(2)-coupled realization

The Gel'fand–Tsetlin states can be constructed explicitly as three-particle SU(2)-coupled products of two-dimensional harmonic-oscillator states.

The construction makes use of a well-known duality relationship (discussed by Moshinsky and Chacón[5]) between U(3) and $\mathcal{U}(2)$ as commuting subgroups of U(6). Let $\{a^\dagger_{im}, a_{im};\ i=1,...,3,\ m=1,2\}$ denote (two-dimensional) harmonic oscillator raising and lowering operators for three particles. The operators $\{a^\dagger_{im}a_{jn}\}$ then span a u(6) Lie algebra. This algebra has two mutually commuting subalgebras: u(3) spanned by the operators

$$\hat{C}_{ij}=\sum_{m=1}^{2} a^\dagger_{im}a_{jm}, \tag{15}$$

and $u(2)$ spanned by

$$\hat{B}_{mn}=\sum_{i=1}^{3} a^\dagger_{im}a_{in}. \tag{16}$$

The algebras u(3) and $u(2)$ are examples of a so-called *dual pair*.[9] The use of a dual pair $(\mathrm{u}(N),u(n))$ and the corresponding direct sum subalgebra $\mathrm{u}(N)+u(n)\subset \mathrm{u}(Nn)$ are well known, for example, in the classification of states of N particles in an n-dimensional harmonic oscillator; cf., for example, the paper by Hagen and MacFarlane[12] which presents a method for deriving the $\mathrm{SU}(m)\times\mathcal{SU}(n)$ content of $\mathrm{SU}(mn)$ and provides tables for the $\mathrm{SU}(6)\to\mathrm{SU}(3)\times\mathcal{SU}(2)$ branching rules.

Now observe that, if $|0\rangle$ is the state in which all particles are in their respective harmonic oscillator ground states, the state

$$|\phi(\lambda,\mu)\rangle=(a^\dagger_{11})^\lambda(a^\dagger_{11}a^\dagger_{22}-a^\dagger_{12}a^\dagger_{21})^\mu|0\rangle \tag{17}$$

satisfies all the conditions of Eq. (6). Thus, $|\phi(\lambda,\mu)\rangle$ is an (unnormalized) SU(3) highest weight state. But it also satisfies

$$\hat{B}_{12}|\phi(\lambda,\mu)\rangle=0,$$
$$\hat{B}_{11}|\phi(\lambda,\mu)\rangle=(\lambda+\mu)|\phi(\lambda,\mu)\rangle,\quad \hat{B}_{22}|\phi(\lambda,\mu)\rangle=\mu|\phi(\lambda,\mu)\rangle, \tag{18}$$

which means that $|\phi(\lambda,\mu)\rangle$ is simultaneously a highest weight state for $u(2)$ with highest weight $(\lambda+\mu,\mu)$ and a highest weight state for u(3) with highest weight $(\lambda+\mu,\mu,0)$, cf. Eq. (7). Moreover, since the u(3) and $u(2)$ operators commute with one another, we can identify all the desired SU(3) basis states with those of the subset of $\mathrm{U}(3)\times\mathcal{U}(2)$ states that are of $\mathcal{U}(2)$ highest weight. This result is a special case of a general result for dual pairs;[9] for any N and n, the commuting algebras u(N) and $u(n)$ have a complete set of highest weight states in common within the carrier space of a fully symmetric irrep of the Lie algebra u(Nn) [i.e., an irrep of highest weight $(\sigma,0,...)$, where σ, equal to $\lambda+2\mu$ in the present case, is the total number of harmonic oscillator quanta].

It is well known that basis states for an $su(2)$ irrep of spin s_i are given by

$$|s_i,m_i\rangle=\frac{(a_{i1}^\dagger)^{s_i+m_i}(a_{i2}^\dagger)^{s_i-m_i}}{\sqrt{(s_i+m_i)!(s_i-m_i)!}}|0\rangle. \tag{19}$$

These states are also a basis for a $u(2)$ irrep of highest weight $(2s_i,0)$. They are tensor products of pairs of $u(1)$ irreps of $u(1)$ spin (s_i+m_i) and $-(s_i-m_i)$, respectively. A Gel'fand basis for SU(3) can likewise be constructed from triple tensor products of $su(2)$ irreps.

Theorem: The weight basis, defined by Eqs. (8)–(14), can be expressed, to within arbitrary phase factors:

$$|\nu I\rangle=[|\tfrac{1}{2}\nu_1\rangle\otimes[|\tfrac{1}{2}\nu_2\rangle\otimes|\tfrac{1}{2}\nu_3\rangle]^I]^{\lambda/2}_{\lambda/2},$$

$$=\sum_{m_1m_2m_3(N)}(\tfrac{1}{2}\nu_3,m_3;\tfrac{1}{2}\nu_2,m_2|I,N)(I,N;\tfrac{1}{2}\nu_1,m_1|\tfrac{1}{2}\lambda,\tfrac{1}{2}\lambda)|\tfrac{1}{2}\nu_1,m_1\rangle|\tfrac{1}{2}\nu_2,m_2\rangle|\tfrac{1}{2}\nu_3,m_3\rangle, \tag{20}$$

with $\nu=(\nu_1,\nu_2,\nu_3)$.

Proof: It follows, from Eq. (15), that

$$\hat{C}_{ii}|\nu I\rangle=\nu_i|\nu I\rangle. \tag{21}$$

Thus, the states $|\nu I\rangle$ have the same weights as their Gel'fand–Tsetlin counterparts. It remains to show that a state $|\nu I\rangle$, defined by Eq. (20), has $\mathrm{SU}(2)_{23}$ angular momentum I.

Consider a set of states for particles 2 and 3 which span an irrep of $\mathrm{u}(2)\times u(2)\subset \mathrm{u}(3)\times u(2)$, where the $\mathrm{u}(2)\subset\mathrm{u}(3)$ subalgebra is spanned by the operators $\{\hat{C}_{23},\hat{C}_{32},\hat{C}_{22},\hat{C}_{33}\}$. If the two-particle states transform according to a u(2) irrep (p,q) then, by duality, they also belong to $u(2)$ irreps of the same highest weight, (p,q). Thus, if a state has su(2) angular momentum $I=(p-q)/2$, it also has $su(2)$ angular momentum I. It follows that the $su(2)$-coupled two-particle state,

$$[|\tfrac{1}{2}\nu_2\rangle\otimes\tfrac{1}{2}\nu_3]^I_N, \tag{22}$$

belongs to a $u(2)$ irrep (p,q) with

$$p+q=\nu_2+\nu_3,\quad p-q=2I, \tag{23}$$

and therefore to the u(2) irrep with the same labels (p,q) and to the irrep with angular momentum $I=\frac{1}{2}(p-q)$ of the subalgebra $\mathrm{su}(2)\subset\mathrm{u}(2)$. This completes the proof.

IV. MATRIX ELEMENTS OF WEYL OPERATORS

The Weyl group is generated by reflections of the roots in the hyperplanes perpendicular to each of the roots. Let α_{ij} denote the SU(3) root whose root vector is $\hat{C}_{ij}$ and let P_{ij} denote the reflection in the line perpendicular to α_{ij}. Then, for example,

$$P_{12}:\alpha_{12}\rightarrow\alpha_{21},\quad \alpha_{13}\rightarrow\alpha_{23},\quad \alpha_{32}\rightarrow\alpha_{31}, \tag{24}$$

and $P_{12}^2=1$. Thus, one obtains the known result that the Weyl group for SU(3) is isomorphic to the symmetric group S_3 of permutations of three objects and that the subset of reflections correspond to transpositions.

By writing Eq. (20) in the form

$$\Psi_{\nu I}(123)\equiv\langle 123|\nu I\rangle=[\psi_{\nu_1}(1)\otimes[\psi_{\nu_2}(2)\otimes\psi_{\nu_3}(3)]^I]^{\lambda/2}_{\lambda/2}, \tag{25}$$

we obtain representations of the Weyl group for SU(3) in which, for example,

$$[P_{12}\Psi_{\nu I}](123)=\langle 123|P_{12}|\nu I\rangle=\Psi_{\nu I}(213),$$

$$[P_{13}\Psi_{\nu I}](123)=\Psi_{\nu I}(321), \tag{26}$$

$$[P_{132}\Psi_{\nu I}](123)=[P_{12}P_{13}\Psi_{\nu I}](123)=\Psi_{\nu I}(312).$$

It follows that

$$\begin{aligned}[P_{12}\Psi_{\nu I}](123)&=[\psi_{\nu_1}(2)\otimes[\psi_{\nu_2}(1)\otimes\psi_{\nu_3}(3)]^I]^{\lambda/2}_{\lambda/2}\\ &=\sum_{I'}(-1)^{(\nu_3-2I-2I'+2\mu-\lambda)/2}\sqrt{(2I+1)(2I'+1)}\\ &\quad\times\begin{Bmatrix}\nu_1/2 & \nu_3/2 & I'\\ \nu_2/2 & \lambda/2 & I\end{Bmatrix}[\psi_{\nu_2}(1)\otimes[\psi_{\nu_1}(2)\otimes\psi_{\nu_3}(3)]^{I'}]^{\lambda/2}_{\lambda/2},\end{aligned} \tag{27}$$

where $\{{abc \atop def}\}$ is a Wigner $6\text{-}j$ symbol. Thus, we obtain the matrix elements

$$\langle\nu' I'|P_{12}|\nu I\rangle=\delta_{\nu_1',\nu_2}\delta_{\nu_2',\nu_1}\delta_{\nu_3',\nu_3}(-1)^{(\nu_3-2I-2I'+2\mu-\lambda)/2}\sqrt{(2I+1)(2I'+1)}\begin{Bmatrix}\nu_1/2 & \nu_3/2 & I'\\ \nu_2/2 & \lambda/2 & I\end{Bmatrix}. \tag{28}$$

In a similar way one determines that

$$\langle\nu' I'|P_{123}|\nu I\rangle=\delta_{\nu_1',\nu_3}\delta_{\nu_2',\nu_1}\delta_{\nu_3',\nu_2}(-1)^{(\nu_1+\nu_2-2I'+2\lambda)/2}\sqrt{(2I+1)(2I'+1)}\begin{Bmatrix}\nu_1/2 & \nu_2/2 & I'\\ \nu_3/2 & \lambda/2 & I\end{Bmatrix} \tag{29}$$

and

$$\langle\nu' I'|P_{132}|\nu I\rangle=\delta_{\nu_1',\nu_2}\delta_{\nu_2',\nu_3}\delta_{\nu_3',\nu_1}(-1)^{(\nu_1+2I+2\mu+\lambda)/2}\sqrt{(2I+1)(2I'+1)}\begin{Bmatrix}\nu_1/2 & \nu_3/2 & I'\\ \nu_2/2 & \lambda/2 & I\end{Bmatrix}. \tag{30}$$

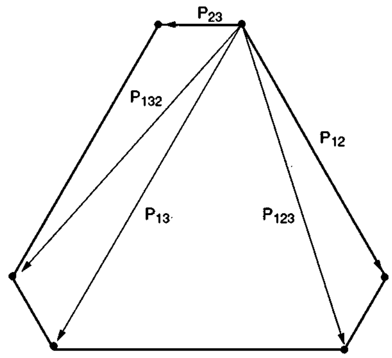

FIG. 2. The action of Weyl group elements on the highest weight of an SU(3) irrep.

To check these results, it is useful to apply them to the highest weight state. We find that

$$P_{12}\left|(\lambda+\mu,\mu,0)\,\frac{\mu}{2}\right\rangle=(-1)^{\mu}\left|(\mu,\lambda+\mu,0)\,\frac{\lambda+\mu}{2}\right\rangle,$$

$$P_{123}\left|(\lambda+\mu,\mu,0)\,\frac{\mu}{2}\right\rangle=\left|(0,\lambda+\mu,\mu)\,\frac{\lambda}{2}\right\rangle, \tag{31}$$

$$P_{132}\left|(\lambda+\mu,\mu,0)\,\frac{\mu}{2}\right\rangle=(-1)^{\mu}\left|(\mu,0,\lambda+\mu)\,\frac{\lambda+\mu}{2}\right\rangle,$$

consistent with the known action on the highest weight shown in Fig. 2. As expected, Weyl group elements map extremal states into other extremal states.

V. WIGNER FUNCTIONS

Matrix elements of $\mathrm{SU}(2)_{23}$ group elements are given immediately in the $\{|\nu I\rangle\}$ basis as SU(2) Wigner functions; viz.,

$$\langle \nu' I'|R_{23}(\alpha,\beta,\gamma)|\nu I\rangle=\delta_{\nu_1',\nu_1}\delta_{I'I}\mathcal{D}^{I}_{(\nu_2'-\nu_3')/2,(\nu_2-\nu_3)/2}(\alpha,\beta,\gamma), \tag{32}$$

where $\mathcal{D}^{I}_{M,N}$ is a standard SU(2) Wigner function.

To evaluate matrix elements of the other $\mathrm{SU}(2)_{ij}$ subgroups, we make use of the fact (noted by Chacón and Moshinsky[5]) that the different $\mathrm{SU}(2)_{ij}$ subgroups are Weyl transforms of one another. Thus, for example, the infinitesimal generators of $\mathrm{SU}(2)_{12}$

$$\hat{C}_{12},\quad \hat{C}_{21},\quad \tfrac{1}{2}(\hat{C}_{11}-\hat{C}_{22}), \tag{33}$$

are related to those of $\mathrm{SU}(2)_{23}$ by

$$\hat{C}_{12}=P_{132}\hat{C}_{23}P_{132}^{-1}=P_{132}\hat{C}_{23}P_{123}. \tag{34}$$

It follows that

$$R_{12}(\alpha,\beta,\gamma)=P_{132}R_{23}(\alpha,\beta,\gamma)P_{123}. \tag{35}$$

Similarly, one finds that

$$R_{13}(\alpha,\beta,\gamma)=P_{12}R_{23}(\alpha,\beta,\gamma)P_{12}. \tag{36}$$

Thus, with the parametrization given by Eq. (3), we obtain the SU(3) Wigner functions

$$D^{(\lambda\mu)}_{\nu' I',\nu I}(\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\alpha_3,\beta_3,\gamma_3)$$

$$=\sum \mathcal{D}^{I'}_{(\nu_2'-\nu_3')/2,(\tau_2-\sigma_3)/2}(\alpha_1,\beta_1,\gamma_1)\langle(\nu_1',\tau_2,\sigma_3)I'|P_{132}|(\sigma_3,\nu_1',\tau_2)J\rangle$$

$$\times\mathcal{D}^{J}_{(\nu_1'-\tau_2)/2,(\nu_1-\sigma_2)/2}(\alpha_2,\beta_2,\alpha_2)\langle(\sigma_3,\nu_1,\sigma_2)J|P_{123}|(\nu_1,\sigma_2,\sigma_3)I\rangle$$

$$\times\mathcal{D}^{I}_{(\sigma_2-\sigma_3)/2,(\nu_2-\nu_3)/2}(\alpha_3,\beta_3,\gamma_3), \tag{37}$$

where the sum is over all σ, τ, and J values allowed by Eqs. (11), (13), and the betweenness conditions (8).

VI. MATRIX ELEMENTS OF SO(3)

If SO(3)⊂SU(3) is the subgroup whose infinitesimal generators are the angular momentum operators

$$\hat{L}_z=-i(\hat{C}_{23}-\hat{C}_{32}),\quad \hat{L}_x=-i(\hat{C}_{31}-\hat{C}_{13}),\quad \hat{L}_y=-i(\hat{C}_{12}-\hat{C}_{21}), \tag{38}$$

then we have the identities

$$\hat{L}_z=2\hat{I}_y,\quad \hat{L}_x=-2\hat{F}_y,\quad \hat{L}_y=2\hat{T}_y, \tag{39}$$

where $\hat{I}_y$, $\hat{T}_y$ and $\hat{F}_y$ belong to the Lie algebras of $\mathrm{SU}(2)_{23}$, $\mathrm{SU}(2)_{13}$, and $\mathrm{SU}(2)_{12}$, respectively. Thus, with the standard parametrization of an SO(3) element

$$\Omega(\alpha,\beta,\gamma)=e^{-i\alpha\hat{L}_z}e^{-i\beta\hat{L}_y}e^{-i\gamma\hat{L}_x}, \tag{40}$$

we have the identity

$$\Omega(\alpha,\beta,\gamma)=R_{23}(0,2\alpha,0)R_{12}(0,2\beta,0)R_{23}(0,2\gamma,0)=R_{23}(0,2\alpha,0)P_{132}R_{23}(0,2\beta,0)P_{123}R_{23}(0,2\gamma,0), \tag{41}$$

and the matrix elements

$$\langle\nu' I'|\Omega(\alpha,\beta,\gamma)|\nu I\rangle=\sum_{\sigma\tau J} d^{I'}_{(\nu_2'-\nu_3')/2,(\tau_3-\sigma_3)/2}(2\alpha)$$

$$\times\langle(\nu_1',\tau_3,\sigma_3)I'|P_{132}|(\sigma_3,\nu_1',\tau_3)J\rangle d^{J}_{(\nu_1'-\tau_3)/2,(\nu_1-\sigma_2)/2}(2\beta)$$

$$\times\langle(\sigma_3,\nu_1,\sigma_2)J|P_{123}|(\nu_1,\sigma_2,\sigma_3)I\rangle d^{L}_{(\sigma_2-\sigma_3)/2,(\nu_2-\nu_3)/2}(2\gamma), \tag{42}$$

where d^I_{MN} is a reduced SU(2) Wigner function.

VII. DISCUSSION

We have derived matrix elements of Weyl group elements and expressions for SU(3) Wigner functions by making use of the dual actions of U(3) and U(2) on the carrier spaces of symmetric representations of U(6).

The groups U(3) and U(2) are special cases of U(N) and U(n) groups that form dual pairs on the carrier spaces of fully symmetric irreps [i.e., irreps of highest weight (σ,0,...)] of U($N\times n$); they are also dual on a direct sum of such spaces.

The essential property of a dual pair[9,13] is that the constituent groups are the centralizers of each other's actions on a specified vector space. The classic example is the Schur–Weyl pair[14] of unitary, $U(n)$, and symmetric, $\mathbf{S}_N$, groups which have commuting actions on the N-fold tensor product, $\mathbb{C}^{N\times n}$, of a complex n-dimensional vector space, $\mathbb{C}^n$. The Schur–Weyl duality has been used effectively to relate the characters of unitary groups, which are infinite Lie groups, to those of the finite symmetric groups. It also underlies the famous Littlewood–Richardson rules[15] for tensor products and the methods of King, Wybourne, and others[16] for inferring branching rules.

Another famous dual pair comprises the orthogonal, $O(N)$, and symplectic, $Sp(n,\mathbb{R})$, groups acting on the N-fold tensor product $\mathbb{H}^{N\times n}$ of the n-dimensional harmonic oscillator Hilbert space $\mathbb{H}^n$.[17] Whereas the Schur–Weyl duality relates the properties of a finite-dimensional irrep of a Lie group to those of a discrete group, the symplectic-orthogonal duality relates the properties of an infinite-dimensional irrep of a noncompact Lie group to those of a compact Lie group. This duality was used, for example, to infer the $Sp(n,\mathbb{R})\to U(n)$ branching rules from known properties of $O(N)$.[18]

It is interesting to note that $U(n)\times U(N)$ and $Sp(n)\times O(N)$ are both direct products of dual pairs on a common harmonic oscillator Hilbert space $\mathbb{H}^{N\times n}$. Thus, one has the useful concept of dual subgroup chains

$$Sp(n,\mathbb{R})\supset U(n)\leftrightarrow O(N)\supset U(N), \tag{43}$$

involving the two dual pairs $Sp(n,\mathbb{R})\times O(N)$ and $U(n)\times U(N)$. These duality relations have been used[19] to relate the representations and tensor products of $U(N)$ in an $O(N)$ basis to those of $Sp(n,\mathbb{R})$ in a $U(n)$ basis. They also play an essential role in the microscopic theory of nuclear collective motion[20] with $\mathbb{H}^{N\times n}$ regarded as the Hilbert space for N-particles in an n-dimensional space.

It should be mentioned that dual subgroup chains were discovered long ago by Brauer[21] who extended the Schur–Weyl duality by observing that the centralizer of the orthogonal subgroup $O(n)\subset U(n)$ on $\mathbb{C}^{N\times n}$ is a group (also an algebra) that contains the symmetric group S_N as a subgroup [cf. Ref. 13 for a discussion of the $O(n)$–Brauer duality].

The physical significance of several of the above dual pairs is illustrated effectively by applications to optics and quantum interferometry, applications which motivated the present investigation.

It has long been known that geometrical optics is an application of Hamiltonian mechanics. Moreover, in the linear approximation, the transformation of a light beam by an optical element, such as a lens, is an $Sp(2,\mathbb{R})$ transformation. This observation is important because it means that the combined effects of many optical elements can be inferred by matrix multiplication. More importantly, one can go beyond the linear approximation to compute the aberrations of an optical system and how to correct them. The techniques for doing this have been developed into a fine art by Dragt and his students[22] and have revolutionized the design of charged-particle and optical beam systems; an introduction to the subject has been given by Guillemin and Sternberg.[23]

We note that there also exists a dual group action on optical systems. If a beam of light or charged particles is polarizable or has intrinsic spin degrees of freedom, then, in addition to the symplectic group action on its spatial phase-space coordinates, there is a dual orthogonal group action on its polarization state. Thus, for example, for light, with two linearly independent polarizations, or for spin-half particle beams, one has a dual $Sp(2,\mathbb{R})\times O(2)$ action on the combined space-spin degrees of freedom. [Note that we mean by $Sp(2,\mathbb{R})$ the rank-2 group of real canonical transformation of a four-dimensional phase space; some authors denote the same group by $Sp(4,\mathbb{R})$.] Thus, one can extend the dynamical group for an optical system from $Sp(2,\mathbb{R})$ to the direct product group $Sp(2,\mathbb{R})\times O(2)$ and thereby admit polarizing (spin rotation) as well as focusing elements. One can further extend the dynamical group to $Sp(4,\mathbb{R})\supset Sp(2,\mathbb{R})\times O(2)$ to include combinations of the two. [It is of interest to note that a general polarizing element is not restricted to $O(2)$ and may induce a $U(2)$ transformation that lies inside $Sp(4,\mathbb{R})$ but which does not commute with the group $Sp(2,\mathbb{R})$ of spatial transformations.]

Such extensions are relevant for describing the quantum interference of light or particle beams. In this case, one is interested in the detailed quantum states of many-photon (many-particle) system. Thus, one is interested in the unitary representations of the dynamical group and, as we have shown explicitly for $U(3)\times U(2)$ in Sec. III C, the irreps of a dynamical group are determined by those of its dual and vice versa.

It has recently been proposed that quantum interferometers should be analyzed in terms of unitary groups.[2,3] A typical quantum interferometer comprises a sequence of elements in which two input modes of the electromagnetic field (beams) are transformed linearly into two output modes. It has been shown that the transformation of the two modes by such an optical element is a $U(2)$ transformation [an $SU(2)$ transformation together with a phase shift].[2] It has also been shown[2] that a so-called *active* interferometer can similarly be represented by an $SU(1,1)$ transformation [note that $SU(1,1)$ is isomorphic to $Sp(1,\mathbb{R})$] and that a linear optical system, comprising n input modes, is represented by an $SU(n)$ transformation.[3]

The use of dual pairs provides a natural framework for the extension of these methods to include polarization and optical elements whose parameters depend on the polarization state of the input fields. To include polarization, one simply extends the $U(n)$ group to $U(n)\times U(2)$ and to include combinations of polarizers and beam splitters, for example, one extends to $U(2n)\supset U(n)\times U(2)$. This is particularly relevant in the quantal context because the input states available to σ photons, when there are n input modes and two linearly independent polarizations for each photon, span an irrep of highest weight $(\sigma,0,...)$ of the group $U(2n)$. The duality properties imply that the subrepresentations available to the subgroup $U(n)\times U(2)$, on restriction of the $U(2n)$ representation $(\sigma,0,...)$, are the so-called two-rowed irreps of type $(\lambda_1,\lambda_2,0,...)\times(\lambda_1,\lambda_2)$ (i.e., irreps whose highest weights have no more than two nonzero components). This follows simply because a $U(2)$ weight has only two components and the two subgroups, $U(n)$ and $U(2)$, being each other's duals, have highest weight states in common. This results in an enormous simplification in the analysis of a multi-mode interferometer. [Note that, as usual, the $SU(n)$ labels are obtained by taking differences of $U(n)$ labels, so that the $U(n)$ irrep $(\lambda_1,\lambda_2,0,...)$ restricts to the $SU(n)$ irrep $(\lambda_1-\lambda_2,\lambda_2,0,...)$.]

An important application of $SU(3)$ interferometry is the experimental test of Bell's theorem without inequalities, known as the GHZ test.[24] Standard tests of Bell's theorem, designed to test the hypotheses of local realism against quantum theory, involve spacelike-separated measurements of two polarization-correlated fields, and local realism establishes an upper bound on the possible degree of correlations between the two fields. The GHZ test, in its ideal form, yields one experimental result for local realism and an entirely different result for quantum theory. Thus, a particular observation determines which theory is correct, and an inequality is not necessary. In the context of $SU(3)$ Wigner functions, the important aspect of the GHZ test is that three polarization-correlated fields are used, and therefore $U(3)\times U(2)$, accounting for three fields and two polarizations, is appropriate here.

Consider, for example, the $SU(n)$ transformations of a one-rowed irrep, $(\lambda,0,...)$, by a system which ignores the polarization. For such an irrep, the highest weight state can be identified with the state

$$|\phi(\lambda,0,...)\rangle=(a_{11}^{\dagger})^{\lambda}|0\rangle. \tag{44}$$

of maximum polarization. Hence, all states of the $SU(n)$ irrep with this highest weight state have maximum polarization. Thus, the $SU(2)$ coupling becomes trivial and basis states for the irrep are labelled simply and uniquely by their weights. It follows that the basis states of the generalized version of the theorem of Sec. III C are simply the states

$$|\nu\rangle=\frac{(a_{11}^{\dagger})^{\nu_1}}{\sqrt{\nu_1!}}\frac{(a_{21}^{\dagger})^{\nu_2}}{\sqrt{\nu_2!}}\cdots\frac{(a_{n1}^{\dagger})^{\nu_n}}{\sqrt{\nu_n!}}|0\rangle. \tag{45}$$

The elements of the Weyl group are seen to act on such states by simply permuting the components $\{\nu_i\}$ of the weights.

For the general two-rowed irreps one must include explicit SU(2) coupling, as shown for SU(3) in Sec. III C. For example, basis states for a two-rowed irrep of U(4) are highest weight states of the dual algebra U(2) and have the general form

$$[|\tfrac{1}{2}\nu_1\rangle\otimes[|\tfrac{1}{2}\nu_2\rangle\otimes[|\tfrac{1}{2}\nu_3\rangle\otimes|\tfrac{1}{2}\nu_4\rangle]^I]^J]^{\lambda/2}_{\lambda/2}. \tag{46}$$

Thus, computing matrix elements of Weyl group elements for any two-rowed SU(n) irrep never involves more than SU(2) recoupling.

ACKNOWLEDGMENTS

The authors are pleased to acknowledge important contributions from J. Repka and C. Bahri.

APPENDIX: FACTORIZATION OF AN SU(3) ELEMENT

Claim: Any element $g\in \mathrm{SU}(3)$ can be parametrized and expressed as a product

$$g(\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\alpha_3,\beta_3,\gamma_3)=R_{23}(\alpha_1,\beta_1,\gamma_1)R_{12}(\alpha_2,\beta_2,\alpha_2)R_{23}(\alpha_3,\beta_3,\gamma_3), \tag{A1}$$

where $R_{23}(\alpha,\beta,\gamma)\in \mathrm{SU}(2)_{23}$, $R_{12}(\alpha,\beta,\alpha)\in \mathrm{SU}(2)_{12}$, and the $\{\mathrm{SU}(2)_{ij}\}$ are the subgroups of SU(3) defined by the subsystems of roots shown in Fig. 1.

Proof: First observe that any SU(3) matrix can be brought to the form

$$\begin{pmatrix} * & * & * \\ * & * & * \\ 0 & * & * \end{pmatrix} \tag{A2}$$

by an $\mathrm{SU}(2)_{23}$ transformation; viz.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & Y^* & Z^* \\ 0 & -Z & Y \end{pmatrix}\begin{pmatrix} x & * & * \\ y & * & * \\ z & * & * \end{pmatrix}=\begin{pmatrix} x & * & * \\ \sqrt{1-|x|^2|} & * & * \\ 0 & * & * \end{pmatrix}, \tag{A3}$$

where $Y=y(1-|x|^2)^{-\frac{1}{2}}$, and $Z=z(1-|x|^2)^{-\frac{1}{2}}$ and we have used the fact that $|x|^2+|y|^2+|z|^2=1$. A subsequent $\mathrm{SU}(2)_{12}$ transformation then brings the matrix to $\mathrm{SU}(2)_{23}$ form; i.e.,

$$\begin{pmatrix} x^* & \sqrt{1-|x|^2|} & 0 \\ -\sqrt{1-|x|^2|} & x & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x & * & * \\ \sqrt{1-|x|^2|} & * & * \\ 0 & * & * \end{pmatrix}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{pmatrix}. \tag{A4}$$

Thus, we determine that

$$\begin{pmatrix} x^* & \sqrt{1-|x|^2|} & 0 \\ -\sqrt{1-|x|^2|} & x & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & Y^* & Z^* \\ 0 & -Z & Y \end{pmatrix}\begin{pmatrix} x & * & * \\ y & * & * \\ z & * & * \end{pmatrix}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{pmatrix}. \tag{A5}$$

Inversion of this equation gives

$$\begin{pmatrix} x & * & * \\ y & * & * \\ z & * & * \end{pmatrix}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & Y & -Z^* \\ 0 & Z & Y^* \end{pmatrix}\begin{pmatrix} x & -\sqrt{1-|x|^2} & 0 \\ \sqrt{1-|x|^2} & x^* & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{pmatrix}, \tag{A6}$$

which proves the claim with suitably chosen parameter values; e.g.,

$$x = e^{-i\alpha_2} \cos(\beta_2/2), \quad \sqrt{1-|x|^2} = \sin(\beta_2/2). \tag{A7}$$

[1] N. Ya Vilenkin, *Special Functions and Theory of Group Representations*, Translations of Math. Monogr. **22** (American Mathematical Society, Providence, RI, 1968); W. Miller, *Lie Theory and Special Functions* (Academic, New York, 1968); J. Talman, *Special Functions, A Group Theoretic Approach* (Benjamin, New York, 1968); H. Ui, Prog. Theor. Phys. **44**, 703 (1970).
[2] B. Yurke, S. L. McCall, and J. R. Klauder, Phys. Rev. A **33**, 4033 (1986); B. C. Sanders and G. J. Milburn, Phys. Rev. Lett. **75**, 2944 (1995); B. C. Sanders, G. J. Milburn, and Z. Zhang, J. Mod. Opt. **44**, 1309 (1997); R. Walser, Phys. Rev. Lett. **79**, 4724 (1997).
[3] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, Phys. Rev. Lett. **73**, 58 (1994); B. C. Sanders and A. Mann, APS Eprint aps1998apr27_002.
[4] J. P. Draayer and S. A. Williams, Nucl. Phys. A **129**, 647 (1969).
[5] E. Chacón and M. Moshinsky, Phys. Lett. **23**, 567 (1966); M. Moshinsky and E. Chacón, in *Spectroscopic and Group Theoretic Methods in Physics* (North-Holland, Amsterdam, 1968), p. 99.
[6] A. J. Macfarlane, E. C. G. Sudarshan, and C. Dullemond, Nuovo Cimento **30**, 845 (1963).
[7] N. Mukunda and L. K. Pandit, J. Math. Phys. **6**, 746 (1965).
[8] A. Gal and H. J. Lipkin, in *Spectroscopic and Group Theoretic Methods in Physics* (North-Holland, Amsterdam, 1968), p. 119.
[9] R. Howe, Proc. Symposia in Pure Math. **33**, 275 (1979).
[10] F. D. Murnaghan, *The Unitary and Rotation Groups* (Spartan, Washington, DC, 1962).
[11] I. M. Gel'fand and M. A. Tsetlin, Dokl. Akad. Nauk. USSR **71**, 825 (1950).
[12] C. R. Hagen and A. J. MacFarlane, J. Math. Phys. **6**, 1355 (1965).
[13] G. Benkart, Contemp. Math. **194**, 1 (1996).
[14] H. Weyl, *Classical Groups* (Princeton U.P., Princeton, NJ, 1939).
[15] D. E. Littlewood, *The Theory of Group Characters* (Clarendon, Oxford, 1950), 2nd ed.
[16] R. C. King, J. Phys. A **8**, 429 (1975); G. R. E. Black, R. C. King, and B. G. Wybourne, *ibid.* **16**, 1555 (1983); G. R. E. Black and B. G. Wybourne, *ibid.* **16**, 2405 (1983).
[17] M. A. Lohe and C. A. Hurst, J. Math. Phys. **12**, 1882 (1971); M. Moshinsky and C. Quesne, *ibid.* **12**, 1772 (1971); M. Kashiwara and M. Vergne, Invent. Math. **44**, 1 (1978).
[18] D. J. Rowe, B. G. Wybourne, and P. H. Butler, J. Phys. A **18**, 939 (1985).
[19] J. Deenen and C. Quesne, J. Phys. A **16**, 2095 (1983); C. Quesne, *ibid.* **17**, 777, 791 (1984); R. Le Blanc and D. J. Rowe, *ibid.* **18**, 1891, 1905 (1985).
[20] D. J. Rowe, Rep. Prog. Phys. **48**, 1419 (1985); Prog. Part. Nucl. Phys. **37**, 265 (1996).
[21] R. Brauer, Ann. Math. **38**, 857 (1937).
[22] A. J. Dragt, ''Lectures on nonlinear orbit dynamics,'' in *Physics of High Energy Particle Accelerators*, AIP Conf. Proc. **87**, 147 (1982).
[23] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics* (Cambridge U.P., Cambridge, 1984).
[24] D. M. Greenberger, M. A. Horne, and A. Zeilinger, ''Going Beyond Bell's Theorem,'' in *Bell's Theorem, Quantum Theory, and Conceptions of the Universe*, edited by M. Kafatos (Kluwer Academic, Dordrecht, 1989), p. 73; D. M. Greenberger, M. A. Horne, A. Shimony, and A. Zeilinger, Am. J. Phys. **58**, 1131 (1990).

A special irreducible matrix representation of the real Clifford algebra $C(3,1)$

K. Scharnhorst[a)]
Humboldt-Universität zu Berlin, Institut für Physik, Invalidenstrasse 110, D-10115 Berlin, Germany

(Received 1 May 1998; accepted for publication 29 October 1998)

4×4 Dirac (gamma) matrices [irreducible matrix representations of the Clifford algebras $C(3,1)$, $C(1,3)$, $C(4,0)$] are an essential part of many calculations in quantum physics. Although the final physical results do not depend on the applied representation of the Dirac matrices (e.g., due to the invariance of traces of products of Dirac matrices), the appropriate choice of the representation used may facilitate the analysis. The present paper introduces a particularly symmetric real representation of 4×4 Dirac matrices (Majorana representation) which may prove useful in the future. As a by-product, a compact formula for (transformed) Pauli matrices is found. The consideration is based on the role played by isoclinic 2-planes in the geometry of the real Clifford algebra $C(3,0)$ which provide an invariant geometric frame for it. It can be generalized to larger Clifford algebras.
 [S0022-2488(99)04606-X]

I. INTRODUCTION

Dirac (gamma) matrices used within many calculations in quantum physics can be understood as representations of Clifford algebras. In four-dimensional Minkowski or Euclidean space they are representations of the Clifford algebras $C(3,1)$, $C(1,3)$ or $C(4,0)$, respectively. While there is no problem to write down sets of complex 4×4 Dirac matrices which form irreducible representations of these Clifford algebras, a set of real 4×4 Dirac matrices (Majorana representation), which we will be interested in, can only be obtained for the Clifford algebra $C(3,1)$[1–4] (further material on real Clifford algebras can be found in Ref. 5, Chap. 13, Refs. 6–11). These matrices obey the standard relation

$$\gamma_\mu\gamma_\nu+\gamma_\nu\gamma_\mu=2\,\eta_{\mu\nu}\mathbf{1}, \tag{1}$$

where $\eta_{\mu\nu}$, $\mu,\nu=1,...,4$ are the elements of the diagonal matrix η with $\mathrm{diag}(\eta)=(1,1,1,-1)$ and $\mathbf{1}$ is the 4×4 unit matrix. An explicit representation of real gamma matrices is provided by the following expressions (adapted from Ref. 4):

$$\gamma_1=\begin{pmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{pmatrix}, \tag{2}$$

$$\gamma_2=\begin{pmatrix}0&0&1&0\\0&0&0&-1\\1&0&0&0\\0&-1&0&0\end{pmatrix}, \tag{3}$$

[a)]Electronic mail: scharnh@physik.hu-berlin.de

$$\gamma_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{4}$$

$$\gamma_4 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}. \tag{5}$$

But, Eq. (1) is invariant under orthogonal transformations O of the gamma matrices

$$\gamma'_\mu = O\gamma_\mu O^T \tag{6}$$

and any other set of congruent [by virtue of (6)] gamma matrices γ'_μ will also be equally appropriate as representation of $C(3,1)$. (The general situation is described by Pauli's fundamental theorem.[12,13]) Now, let us denote the real linear vector space $\mathbf{R}_4$ in which the elements of the Clifford algebra $C(3,1)$ act as operators by V (spinor space). Then, the matrices γ_μ can be understood as representations of the generators of $C(3,1)$ with respect to a certain orthonormal basis in V which defines in it a rectangular coordinate system. Any transformation (6) of the gamma matrices corresponds to an orthogonal transformation in V and consequently to a change of the coordinate system in V. The concrete shape of the gamma matrices changes in performing these transformations. In explicit calculations in which gamma matrices occur the required effort may depend on the explicit shape of the gamma matrices. Therefore, in dependence on the physical problem under consideration one may ask whether it is possible to find a coordinate system in which the gamma matrices assume a particularly convenient shape for some calculational purpose. The detailed requirements certainly may depend on the purpose. From such a problem, recently we have been led to ask ourselves whether it is possible to find an irreducible representation of the real Clifford algebra $C(3,1)$ which is particularly symmetric with respect to the index μ of the gamma matrices γ'_μ. Indeed, it is possible to find an orthogonal transformation which transforms the gamma matrices (2)–(5) into the following expressions which are obviously particularly symmetric with respect to the index of the gamma matrices $k=1,2,3$ ($\mathbf{1}$ and $\mathbf{0}$ are the 2×2 unit and null matrices, respectively; φ_0 is some arbitrary real constant; cf. Sec. V),

$$\gamma'_k = \frac{1}{\sqrt{3}}\begin{pmatrix} \mathbf{1} & \mathbf{F}_k \\ \mathbf{F}_k & -\mathbf{1} \end{pmatrix}, \quad \mathbf{F}_k = \begin{pmatrix} f(-\varphi_k) & f(\varphi_k) \\ f(\varphi_k) & -f(-\varphi_k) \end{pmatrix}, \tag{7}$$

$$f(\varphi) = \cos\varphi + \sin\varphi = \sqrt{2}\cos\left(\varphi - \frac{\pi}{4}\right), \tag{8}$$

$$\varphi_k = \varphi(k) = \varphi_0 + \frac{2\pi}{3}k, \tag{9}$$

$$\gamma'_4 = \begin{pmatrix} \mathbf{0} & -\mathbf{1} \\ \mathbf{1} & \mathbf{0} \end{pmatrix}. \tag{10}$$

As a by-product, from the above expressions one obtains the following compact formula for transformed Pauli matrices [irreducible matrix representations of the complex Clifford algebra $C(3,0)$; cf. Appendix B].

$$\sigma'_k = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & \sqrt{2}e^{-i\varphi_k} \\ \sqrt{2}e^{i\varphi_k} & -1 \end{pmatrix}. \tag{11}$$

It is the purpose of the present article to systematically derive the above expressions relying on certain information not applied previously within the present context. The discussion is accompanied by references to the relevant but scattered literature.

Our considerations will be guided by the following idea. Related to the Clifford algebra $C(3,0)$, it should be possible to find an expression for the set of the gamma matrices γ'_k, $k=1,2,3$ which is particularly symmetric with respect to the index k. We approach the problem by noting that each gamma matrix γ_k has 2 two-dimensional eigenspaces related to the eigenvalues $\rho=1$ and $\rho=-1$ (which are orthogonal to each other). Any coordinate system in V stands in a certain geometric relation to all the eigenspaces of the gamma matrices whose mutual relation is an invariant under any transformation (6). Now, the idea consists in finding such a coordinate system in V with respect to which all the eigenspaces of the gamma matrices lie in a particularly symmetric way. Then, one may expect that the explicit expressions for the gamma matrices γ'_k reflect this symmetry. Therefore, in Sec. II we start with some observations concerning the eigenspaces of the generators of the Clifford algebra $C(3,0)$ (more precisely, in using this term we always mean the generators of its irreducible representations).

II. ISOCLINIC 2-PLANES IN $\mathbf{R}_4$

To begin with, let us discuss some aspects of the geometry of 2-planes in the affine space $\mathbf{R}_4$ which we also denote by V for simplicity. We restrict our consideration to 2-planes containing the point $\mathbf{x}=(0,0,0,0)$ [i.e., to the Grassmann manifold $G(2,4)$, for a related review see Ref. 14]. We will rely here on the general multidimensional matrix formalism presented in Ref. 15, Chap. 3, Sec. 3 (also see Ref. 16, Chap. III, Sec. 3.3) which we specialize to $\mathbf{R}_4$. In the following we will start with some material which provides the necessary information on those aspect of the formalism of Refs. 15 and 16 which is relevant for the present paper.

For our purposes, a point $\mathbf{x}$ of a given 2-plane A can be described in terms of the equation

$$\mathbf{x}=\mathbf{A}\mathbf{t}, \tag{12}$$

where $\mathbf{A}$ is a 4×2 matrix whose two columns are given by the coordinates of two linearly independent vectors spanning the 2-plane A while $\mathbf{t}$ is the two-component vector of the coordinates of the point $\mathbf{x}\in A$. Two 2-planes A and B can intersect in V in various ways. In order to study their relation, to each pair of lines $X\subset A$, $Y\subset B$ the angle they enclose can be calculated. Once a line $X\subset A$ is fixed, for any arbitrary line $Y\subset B$ the angle enclosed assumes values between some $\alpha_0\geqslant 0$ and $\pi/2$. In general, α_0 may lie between some minimal and some maximal value—the so-called *stationary angles (principal angles)* $\alpha_{\min}$, $\alpha_{\max}$—which are characteristic for the geometry of the pair of 2-planes A, B. Now, from an extremum principle a 2×2 matrix

$$\mathbf{W}=(\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{B})(\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}) \tag{13}$$

can be constructed[17] for whose eigenvalues w_1 and w_2 the equations

$$w_1=\cos^2\alpha_{\max}, \tag{14}$$

$$w_2=\cos^2\alpha_{\min} \tag{15}$$

apply. If the 2-planes A, B are given by means of Eq. (12) in terms of two orthonormal vectors each, Eq. (13) simplifies to the form[35]

$$\mathbf{W}=(\mathbf{A}^T\mathbf{B})(\mathbf{B}^T\mathbf{A}). \tag{16}$$

If the matrix $\mathbf{W}$ is proportional to the unit matrix (i.e., $w_1=w_2=w$)

$$\mathbf{W}=w\mathbf{1}, \tag{17}$$

the 2-planes A and B are said to be (mutually) *isoclinic*.[41] Then, to each vector $\mathbf{x}\in A$ a unique line in B exists (determined by the orthogonal projection of $\mathbf{x}$ onto B) which encloses with $\mathbf{x}$ the (stationary) angle $\alpha=\arccos\sqrt{w}$.[48] Finally, we would like to mention that under some natural bijection between $\mathbf{R}_4$ and $\mathbf{C}_2$ $((z_1,z_2)=(x_1+ix_2,x_3+ix_4)\in\mathbf{C}_2,\ (x_1,x_2,x_3,x_4)\in\mathbf{R}_4)$ two isoclinic 2-planes in $\mathbf{R}_4$ correspond to two lines through the origin in $\mathbf{C}_2$ (Ref. 23, Sec. 1-7, p. 51, theorem 1-7.4).

Now, the above formalism can be used to analyze the geometry of the set of 6 two-dimensional eigenspaces of the generators of the Clifford algebra $C(3,0)$ (i.e., more precisely, the generators of its irreducible representation). After some calculation using, e.g., the explicit representations of the gamma matrices (2)–(4) one finds that all their six eigenspaces are pairwise isoclinic 2-planes (some choice for the matrices $\mathbf{A}$ describing the eigenspaces is given in Appendix A). Of course, the two eigenspaces of a given gamma matrix γ_k are orthogonal to each other. But, any other two eigenspaces are pairwise isoclinic with an (stationary) angle $\alpha=\pi/4$. Consequently, we can find, at maximum, a set of three eigenspaces of the gamma matrices γ_k, $k=1,2,3$, whose elements are pairwise isoclinic with the angle $\pi/4$.[54] Such a set of 2-planes is called an *equiangular frame* (Ref. 22, Pt. I, Sec. 5, p. 40). With respect to the aim of the present paper, in the following we will just be interested in such sets.

III. THE CLIFFORD ALGEBRA $C(3,0)$ AND EQUIANGULAR FRAMES

We begin this section with some necessary information taken from Ref. 22[58] and specialized to the present needs (in the following the term ''adapted quote'' always means that the original text is quoted exactly except that any reference to the general multidimensional space $\mathbf{R}_{2n}$ has been specialized to $\mathbf{R}_4$). The following definition will be used: ''A *set of mutually isoclinic 2-planes* in $\mathbf{R}_4$ is characterized by the property that every two 2-planes of the set are isoclinic with each other. A set of mutually isoclinic 2-planes in $\mathbf{R}_4$ is called a *maximal* set if it is not subset of a larger set of mutually isoclinic 2-planes'' (this is an adapted quote from Ref. 22, Pt. I, Sec. 3, p. 19).[60]

In order to make contact with the formalism used in Ref. 22 which we will rely on in the further discussion we need to rewrite the defining equation (12) for a 2-plane A in one of the following two (alternative) ways:

$$\mathbf{x}_{(3,4)}=\tilde{\mathbf{A}}\mathbf{x}_{(1,2)}, \quad \tilde{\mathbf{A}}=\bar{\bar{\mathbf{A}}}(\overline{\mathbf{A}})^{-1}, \tag{18}$$

$$\mathbf{x}_{(1,2)}=\tilde{\tilde{\mathbf{A}}}\mathbf{x}_{(3,4)}, \quad \tilde{\tilde{\mathbf{A}}}=\overline{\mathbf{A}}(\bar{\bar{\mathbf{A}}})^{-1}. \tag{19}$$

Here, the notation $\mathbf{x}_{(1,2)}=(x_1,x_2)^T$, $\mathbf{x}_{(3,4)}=(x_3,x_4)^T$ is used and the 2×2 matrices $\overline{\mathbf{A}}$, $\bar{\bar{\mathbf{A}}}$ are related to the matrix $\mathbf{A}$ in the following way:

$$\mathbf{A}=\begin{pmatrix}\overline{\mathbf{A}}\\ \bar{\bar{\mathbf{A}}}\end{pmatrix}. \tag{20}$$

Equation (18) [(19)] is valid for any 2-plane which is isoclinic but not identical to the 2-plane $O_{(3,4)}$: $\mathbf{x}_{(1,2)}=0$ [$O_{(1,2)}$: $\mathbf{x}_{(3,4)}=0$] (this entails that the 2-plane A intersects the 2-plane $O_{(3,4)}$ [$O_{(1,2)}$] in the point $\mathbf{x}=(0,0,0,0)$ only and, therefore, ensures the invertibility of $\overline{\mathbf{A}}$ [$\bar{\bar{\mathbf{A}}}$]).

According to Wong (Ref. 22, Pt. I, Sec. 7, p. 54, theorem 7.2; also see Ref. 23, Sec. 1-7, p. 43), every maximal set of mutually isoclinic 2-planes in $\mathbf{R}_4$ is of dimension 2 and is congruent (i.e., related by an orthogonal transformation in $\mathbf{R}_4$) to the maximal set given by

$$\mathbf{x}_{(3,4)}=\tilde{\mathbf{B}}(\lambda_0,\lambda_1)\mathbf{x}_{(1,2)}=[\lambda_0\tilde{\mathbf{B}}_0+\lambda_1\tilde{\mathbf{B}}_1]\mathbf{x}_{(1,2)}, \tag{21}$$

$$\tilde{\mathbf{B}}_0=\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix},\quad \tilde{\mathbf{B}}_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \tag{22}$$

or

$$\mathbf{x}_{(1,2)}=\tilde{\tilde{\mathbf{B}}}(\lambda_0,\lambda_1)\mathbf{x}_{(3,4)}, \tag{23}$$

$$\tilde{\tilde{\mathbf{B}}}(\lambda_0,\lambda_1)=\tilde{\mathbf{B}}(\lambda_0,\lambda_1)^{-1}=\frac{1}{\lambda_0^2+\lambda_1^2}\tilde{\mathbf{B}}(\lambda_0,\lambda_1)=\tilde{\mathbf{B}}(\lambda_0',\lambda_1'),$$
$$\lambda_n'=\frac{\lambda_n}{\lambda_0^2+\lambda_1^2},\quad n=1,2, \tag{24}$$

where λ_0, λ_1 are two real parameters.[64]

Both of the 2-planes $O_{(1,2)}:\mathbf{x}_{(3,4)}=0$ and $O_{(3,4)}:\mathbf{x}_{(1,2)}=0$ belong to this maximal set (Ref. 22, Pt. I, Sec. 2, p. 16, lemma 2.2). Equations (21) and (23) entail that the matrix $\mathbf{B}$ to be inserted in the corresponding Eq. (12) reads, e.g. (we have chosen particularly simple expressions),

$$\mathbf{B}(\lambda_0,\lambda_1)=\frac{1}{\sqrt{1+\lambda_0^2+\lambda_1^2}}\begin{pmatrix}\mathbf{1}\\ \tilde{\mathbf{B}}(\lambda_0,\lambda_1)\end{pmatrix}, \tag{25}$$

or

$$\mathbf{B}(\lambda_0',\lambda_1')=\frac{1}{\sqrt{1+\lambda_0'^2+\lambda_1'^2}}\begin{pmatrix}\tilde{\mathbf{B}}(\lambda_0',\lambda_1')\\ \mathbf{1}\end{pmatrix}. \tag{26}$$

Furthermore, Wong finds that (adapted quote) ''in $\mathbf{R}_4$, any maximal set of mutually isoclinic 2-planes which contains the 2-plane $O_{(1,2)}$ corresponds to a linear subspace of the linear space of all 2×2 matrices'' (Ref. 22, Pt. I, Sec. 3, p. 20, lemma 3.2). Now, in this two-dimensional subspace a matrix basis can be chosen in such a way that the 2-planes described by the elements of the basis and the 2-plane $O_{(1,2)}$ (or $O_{(3,4)}$) form an equiangular frame (Ref. 22, Pt. I, Sec. 3, p. 24, lemma 3.3 and p. 40). As one may convince oneself easily by means of the explicit expressions given in Appendix A, each equiangular frame built from the eigenspaces of the gamma matrices contains a basis of one and the same maximal set of mutually isoclinic 2-planes.

For the purpose of the present paper it appears to be useful to consider two disjoint equiangular frames Ω connected with the gamma matrices (2)–(4)—one (Ω_1) related to the three eigenspaces to the eigenvalue $\rho=1$, and the other one (Ω_{-1}) related to the three eigenspaces to the eigenvalue $\rho=-1$. The following theorem by Wong will be helpful then (Φ is any maximal set of mutually isoclinic 2-planes in $\mathbf{R}_4$; the following is an adapted quote; the indices have also been changed to conform to the notation used in the present article): ''If the angles between any 2-plane of Φ and the three 2-planes of an equiangular frame are θ_k $(1\leq k\leq 3)$, then

$$\cos^2 2\theta_1+\cos^2 2\theta_2+\cos^2 2\theta_3=1. \tag{27}$$

Conversely, given any set of three angles θ_k $(1\leq k\leq 3)$ such that $0\leq\theta_k\leq\pi$ and $\Sigma\cos^2 2\theta_k=1$, then there exists a unique 2-plane isoclinic to each of the three 2-planes of a given equiangular frame, making angles θ_k with them, and this 2-plane belongs to Φ'' [Ref. 22, Pt. I, Sec. 5, p. 41, theorem 5.3 (b)]. From this insight we conclude that, obviously, to each equiangular frame Ω_1 [Ω_{-1}] two uniquely determined 2-planes $A_{1\pm}$ [$A_{-1\pm}$] exist which lie in a particularly symmetric way (isoclinic) relative to the elements of Ω_1 [Ω_{-1}]. For $A_{1\pm}$, $A_{-1\pm}$ it holds

$$\theta_1 = \theta_2 = \theta_3 = \theta_{\text{sym}}, \quad \cos 2\theta_{\text{sym}} = \pm \frac{1}{\sqrt{3}}. \tag{28}$$

For the corresponding eigenvalue of the matrix **W**, Eq. (17), one obtains

$$w = \cos^2 \theta_{\text{sym}} = \frac{1}{2}(1 + \cos 2\theta_{\text{sym}}) = \frac{1}{2}\left(1 \pm \frac{1}{\sqrt{3}}\right) = w_\pm . \tag{29}$$

The two different values of θ_{sym} (and w)) will not cause any major difference in the following considerations as both cases are related by a simple permutation of the indices of the gamma matrices.

IV. CHANGE OF THE COORDINATE SYSTEM

We may now set out to determine the position of the 2-planes $A_{1\pm}$, $A_{-1\pm}$ using the formulas given in Secs. II and III. For the 2-planes $A_{1\pm}$, $A_{-1\pm}$ we can apply a general ansatz according to Eqs. (21), (23), (25), (26) and calculate the eigenvalue of the matrix **W** for each of the three pairs given by one of the elements of the equiangular frame Ω_1 $[\Omega_{-1}]$ and $A_{1\pm}$ $[A_{-1\pm}]$. For each eigenvalue ρ of the gamma matrices (2)–(4), this leads to a set of three equations for the parameters λ_0, λ_1 which have to be solved simultaneously taking into account Eq. (29). These equations read for $\rho = 1$ (in sequence for the indices $k = 1$, $k = 2$ and $k = 3$ of the gamma matrices, respectively)

$$w_\pm = \frac{\lambda_0'^2 + (1 + \lambda_1')^2}{2(1 + \lambda_0'^2 + \lambda_1'^2)}, \tag{30}$$

$$w_\pm = \frac{(1 - \lambda_0')^2 + \lambda_1'^2}{2(1 + \lambda_0'^2 + \lambda_1'^2)}, \tag{31}$$

$$w_\pm = \frac{\lambda_0'^2 + \lambda_1'^2}{1 + \lambda_0'^2 + \lambda_1'^2}, \tag{32}$$

and for $\rho = -1$,

$$w_\pm = \frac{\lambda_0^2 + (1 - \lambda_1)^2}{2(1 + \lambda_0^2 + \lambda_1^2)}, \tag{33}$$

$$w_\pm = \frac{(1 + \lambda_0)^2 + \lambda_1^2}{2(1 + \lambda_0^2 + \lambda_1^2)}, \tag{34}$$

$$w_\pm = \frac{\lambda_0^2 + \lambda_1^2}{1 + \lambda_0^2 + \lambda_1^2}. \tag{35}$$

(Equations (30)–(32) [(33)–(35)] have been derived using the expressions given in Appendix A and Eq. (26) [(25)].) The solution of the above equations reads for $\rho = 1$,

$$\lambda_0' = -\lambda_1' = -\lambda_\pm , \tag{36}$$

and for $\rho = -1$,

$$\lambda_0 = -\lambda_1 = \lambda_\pm . \tag{37}$$

Here,

$$\lambda_{\pm}=\pm\sqrt{3}w_{\pm}, \tag{38}$$

which entails

$$2\lambda_{\pm}\lambda_{\mp}=-1. \tag{39}$$

Now, we may assume that the explicit representations for the gamma matrices (2)–(5) are related to the natural basis in V from which two pairs of basis vectors can be selected which define the orthogonal 2-planes $O_{(1,2)}$, $O_{(3,4)}$. In order to obtain a particularly symmetric representation for the gamma matrices it appears to be advantageous now to go over to an orthonormal basis from which two pairs of basis vectors can be chosen which define the orthogonal 2-planes $A_{1\pm}$, $A_{-1\pm}$. This change of the basis in V is associated with an orthogonal transformation O in V which transforms the gamma matrices in accordance with Eq. (6). We start by choosing an appropriate orthonormal basis in V from which the matrices $\mathbf{A}_{1\pm}$, $\mathbf{A}_{-1\pm}$ describing the 2-planes $A_{1\pm}$, $A_{-1\pm}$ can be built [we simply insert the solutions (36) and (37) into Eqs. (26) and (25), respectively],

$$\mathbf{A}_{1\pm}=\frac{1}{\sqrt{1+2\lambda_{\pm}^2}}\begin{pmatrix}\lambda_{\pm} & \lambda_{\pm}\\ \lambda_{\pm} & -\lambda_{\pm}\\ 1 & 0\\ 0 & 1\end{pmatrix}, \tag{40}$$

$$\mathbf{A}_{-1\pm}=\frac{1}{\sqrt{1+2\lambda_{\pm}^2}}\begin{pmatrix}1 & 0\\ 0 & 1\\ -\lambda_{\pm} & -\lambda_{\pm}\\ -\lambda_{\pm} & \lambda_{\pm}\end{pmatrix}. \tag{41}$$

One immediately recognizes that the 2-planes $A_{1\pm}$, $A_{-1\pm}$ are orthogonal to each other. Furthermore, by virtue of Eq. (39) it holds $A_{1\pm}=A_{-1\mp}$. Of course, the above choice for the matrices $\mathbf{A}_{1\pm}$, $\mathbf{A}_{-1\pm}$ is not unique and any orthonormal basis which is related to the basis used in the above equations by a rotation within the 2-planes $A_{1\pm}$, $A_{-1\pm}$ is equally well suited. In fact, further below we will use exactly this freedom to obtain our final result (7)–(10).

The transition from the natural basis in V which is related to the 2-planes $O_{(1,2)}$, $O_{(3,4)}$ to the basis which is given in terms of Eqs. (40) and (41) and which is related to the 2-planes $A_{1\pm}$, $A_{-1\pm}$ is described by the orthogonal transformation $O_{\pm}$,

$$O_{\pm}=\frac{1}{\sqrt{1+2\lambda_{\pm}^2}}\begin{pmatrix}\lambda_{\pm} & \lambda_{\pm} & 1 & 0\\ \lambda_{\pm} & -\lambda_{\pm} & 0 & 1\\ 1 & 0 & -\lambda_{\pm} & -\lambda_{\pm}\\ 0 & 1 & -\lambda_{\pm} & \lambda_{\pm}\end{pmatrix}, \tag{42}$$

which leads via $\gamma_{\mu}''=O_{\pm}\gamma_{\mu}O_{\pm}^T$ to the correspondingly transformed expressions for the gamma matrices γ_{μ} [of course, for our choice (42) it holds $O_{\pm}=O_{\pm}^T$]. After some algebra [taking into account Eq. (39)] one finds

$$\gamma_{1\pm}''=-\gamma_{2\mp}''=\pm\frac{1}{\sqrt{3}}\begin{pmatrix}1 & 0 & -\lambda_{\pm} & -\lambda_{\mp}\\ 0 & 1 & -\lambda_{\mp} & \lambda_{\pm}\\ -\lambda_{\pm} & -\lambda_{\mp} & -1 & 0\\ -\lambda_{\mp} & \lambda_{\pm} & 0 & -1\end{pmatrix}, \tag{43}$$

$$\gamma''_{3\pm}=\pm\frac{1}{\sqrt{3}}\begin{pmatrix}1&0&1&1\\0&1&1&-1\\1&1&-1&0\\1&-1&0&-1\end{pmatrix},\tag{44}$$

$$\gamma''_4=\begin{pmatrix}0&0&-1&0\\0&0&0&-1\\1&0&0&0\\0&1&0&0\end{pmatrix}.\tag{45}$$

From Eq. (43) one immediately recognizes that the two cases differing by the sign in Eq. (28) are related to each other by a permutation of the gamma matrices with the indices $k=1$ and $k=2$.

V. RESIDUAL ROTATIONS

Although in Sec. IV we have performed the transformation to a coordinate system which lies in a particularly symmetric way with respect to the equiangular frames Ω_1, Ω_{-1} built from the eigenspaces of the gamma matrices, at first glance the transformed expressions (43), (44) do not seem to exhibit any particular symmetry with respect to the index $k=1,2,3$ of the gamma matrices. However, the expected symmetry is there and we are going to reveal it now. Let us remind ourselves that the choice of the new basis (coordinate system) was not unique and we have disregarded for the moment the remaining freedom to perform rotations within the 2-planes $A_{1\pm}$, $A_{-1\pm}$. Any such rotation can be described by the orthogonal transformation

$$O(\beta_1,\beta_{-1})=\begin{pmatrix}\cos\beta_1&-\sin\beta_1&0&0\\ \sin\beta_1&\cos\beta_1&0&0\\0&0&\cos\beta_{-1}&-\sin\beta_{-1}\\0&0&\sin\beta_{-1}&\cos\beta_{-1}\end{pmatrix},\tag{46}$$

where β_1 and β_{-1} are the independent rotation angles within the orthogonal 2-planes $A_{1\pm}$ and $A_{-1\pm}$, respectively (for the sake of completeness we mention that in addition to the above rotations an inversion within one of the 2-planes $A_{1\pm}$, $A_{-1\pm}$ may be considered). Again, we can write down the further transformed gamma matrices $\gamma'_\mu=O(\beta_1,\beta_{-1})\gamma''_\mu O(\beta_1,\beta_{-1})^T$. For brevity, we give the relatively simple expressions for $\gamma'_{3\pm}$ and γ'_4 only,

$$\gamma'_{3\pm}(\varphi)=\pm\frac{1}{\sqrt{3}}\begin{pmatrix}1&0&f(-\varphi)&f(\varphi)\\0&1&f(\varphi)&-f(-\varphi)\\f(-\varphi)&f(\varphi)&-1&0\\f(\varphi)&-f(-\varphi)&0&-1\end{pmatrix},\tag{47}$$

$$\gamma'_4(\overline{\varphi})=\begin{pmatrix}0&0&-\cos\overline{\varphi}&\sin\overline{\varphi}\\0&0&-\sin\overline{\varphi}&-\cos\overline{\varphi}\\ \cos\overline{\varphi}&\sin\overline{\varphi}&0&0\\-\sin\overline{\varphi}&\cos\overline{\varphi}&0&0\end{pmatrix}.\tag{48}$$

Here, $\varphi=\beta_1+\beta_{-1}$ and $\overline{\varphi}=\beta_1-\beta_{-1}$. The gamma matrices $\gamma'_{k\pm}$, $k=1,2,3$, do not depend on $\overline{\varphi}$ while γ'_4 does not depend on φ. The function f is given by

$$f(\varphi)=\cos\varphi+\sin\varphi=\sqrt{2}\cos\left(\varphi-\frac{\pi}{4}\right). \tag{49}$$

Symmetry considerations now suggest that any set of (three) rotations $O(\beta_1,\beta_{-1})$ among whose elements $\varphi=\beta_1+\beta_{-1}$ changes by a multiple of $2\pi/3$ (mod 2π) will lead to a set of three gamma matrices with the indices $k=1,2,3$. Consequently, in order to describe this set we can write

$$\varphi(k)=\varphi_0+\frac{2\pi}{3}k=\varphi_k, \tag{50}$$

where φ_0 is some real constant. Any three gamma matrices given by Eqs. (47), (49), and (50) can be chosen to serve as an irreducible representation of the real Clifford algebra $C(3,0)$. If we choose $\varphi_0=0$, Eqs. (47), (49), and (50) exactly reproduce the set of gamma matrices (43), (44), i.e.,

$$\gamma'_{3\pm}\left(\frac{2\pi}{3}\right)=\gamma''_{1\pm}, \quad \gamma'_{3\pm}\left(\frac{4\pi}{3}\right)=\gamma''_{2\pm}. \tag{51}$$

Furthermore, for the sake of simplicity it seems to be convenient to set $\overline{\varphi}=0$ and to vary φ exclusively. [Such an orthogonal transformation is called a *Clifford translation* (Ref. 23, Sec. 2-6, p. 102) and has special properties. In this context, also note Ref. 66.] This way the final result [Eqs. (7)–(10), also see Appendix B for some related consideration] quoted in Sec. I is obtained [where we have omitted, for simplicity, the $\pm$ sign on the right-hand side of Eq. (47) which relates to the two inequivalent irreducible representations of $C(3,0)$ (Ref. 2, p. 1657)]. The generators of the real Clifford algebra $C(3,0)$ are found from one of them by means of a discrete $\mathbf{Z}_6\sim\mathbf{Z}_2\times\mathbf{Z}_3$ subgroup of the orthogonal group $O(4)$ (in other words, the $\mathbf{Z}_6$ subgroup realizes a permutation among the gamma matrices). The Clifford translation in the spinor space V with $\beta_1=\beta_{-1}=\pi/3$ corresponds to a rotation by $2\pi/3$ around the axis (1,1,1) in the vector space $\mathbf{R}_{3,0}$ associated with the Clifford algebra $C(3,0)$ (it is an element of the group Spin(3)).

We want to extend our discussion now to the real Clifford algebra $C(3,2)$, which is the largest Clifford algebra admitting an irreducible representation by means of 4×4 matrices. From Eqs. (47), (51) we can calculate the product

$$\gamma'_{3\pm}(\varphi_1)\,\gamma'_{3\pm}(\varphi_2)\,\gamma'_{3\pm}(\varphi_3)=\begin{pmatrix}0&1&0&0\\-1&0&0&0\\0&0&0&-1\\0&0&1&0\end{pmatrix}, \tag{52}$$

which is found to be independent of the choice of φ_0. Allowing an arbitrary value for $\overline{\varphi}$, γ'_5 can then be calculated and reads

$$\begin{aligned}\gamma'_5=\gamma'_5(\overline{\varphi})&=\gamma'_{3\pm}(\varphi_1)\,\gamma'_{3\pm}(\varphi_2)\,\gamma'_{3\pm}(\varphi_3)\,\gamma'_4(\overline{\varphi})\\&=\gamma'_4\left(\overline{\varphi}-\frac{\pi}{2}\right)\\&=\begin{pmatrix}0&0&-\sin\overline{\varphi}&-\cos\overline{\varphi}\\0&0&\cos\overline{\varphi}&-\sin\overline{\varphi}\\\sin\overline{\varphi}&-\cos\overline{\varphi}&0&0\\\cos\overline{\varphi}&\sin\overline{\varphi}&0&0\end{pmatrix},\\{\gamma'_5}^2&=-\mathbf{1}.\end{aligned} \tag{53}$$

Finally, the charge conjugation operator C ($C^T=-C$, $C\gamma'_\mu C^{-1}=-\gamma'^T_\mu$) can be given by $C=\gamma'_4(\overline{\varphi})$. In difference to the $C(3,0)$ subalgebra of the Clifford algebra $C(3,2)$, which is generated by relying on Eq. (50) [a variation of φ by 2π leads to just one copy of the generators of $C(3,0)$], the $C(0,2)$ subalgebra can be represented by $\gamma'_4=\gamma'_4(\overline{\varphi})$, $\gamma'_5=\gamma'_4(\overline{\varphi}\pm\pi/2)$ [a variation of $\overline{\varphi}$ by 2π leads to two copies of the generators of $C(0,2)$].[67] In this context, note

$$\gamma'_4(\overline{\varphi})=-\gamma'_4(\overline{\varphi}+\pi). \tag{54}$$

For $\varphi=0$, the second generator of the real Clifford algebra $C(0,2)$ is obtained from the first by means of a discrete $\mathbf{Z}_8\sim(\mathbf{Z}_2)^3$ subgroup of the orthogonal group $O(4)$. A rotation (46) in the spinor space V with $\beta_1=-\beta_{-1}=\pi/4$ corresponds to a rotation by $\pi/2$ in the vector space $\mathbf{R}_{0,2}$ associated with the Clifford algebra $C(0,2)$ [it is an element of the group Spin(2)].

VI. DISCUSSION

According to Pauli's fundamental theorem[12,13] any set of (in general, complex) 4×4 gamma matrices γ_μ, which represent the Clifford algebra $C(3,1)$, is related to our expressions for γ'_μ [Eqs. (7)–(10)] by means of a nonsingular transformation S ($\gamma_\mu=S\gamma'_\mu S^{-1}$). Therefore, any such set can, in principle, be written in a form analogous to Eqs. (7)–(10) (of course, in general such a representation may look fairly cumbersome). It is clear, that this consideration of the (complex) Clifford algebra $C(3,1)$ immediately carries over with little change to the Clifford algebra $C(1,3)$ and does not require any further special investigation. Furthermore, it seems natural to expect that the discussion of the real Clifford algebra $C(3,1)$ performed in the present paper can appropriately be generalized also to other Clifford algebras. Of course, the simpler and rather trivial case of the real Clifford algebra $C(2,1)$ which is presented in Appendix C carries the traces of the structures found for $C(3,0)$. On the other hand, one should expect that these structures themselves are also traces of more general structures of Clifford algebras which contain $C(3,0)$ as a subalgebra. Let us emphasize at this point that the mathematical tools we have relied on in Secs. II and III are not specific to the present case (although, we have specialized them to the present case, for simplicity) and they can also be used in more general situations. As interesting as this may be, it goes far beyond the purpose of the present study and, therefore, will not be investigated here.

ACKNOWLEDGMENTS

I am grateful to B. A. Rosenfeld for kindly informing me of Ref. 16. The present work has been performed under the EC Training and Mobility of Researchers Program, return fellowship Contract No. ERBFMBICT961197.

APPENDIX A: DESCRIPTION OF THE EIGENSPACES OF γ_k

In this Appendix we give some explicit expressions for the matrices $\mathbf{A}_{k,\rho}$ which define via Eq. (12) the eigenspace (i.e., the 2-plane $A_{k,\rho}$) of the gamma matrix γ_k, $k=1,2,3$, to the eigenvalue $\rho=1,-1$. We rely on orthonormal basis vectors for each eigenspace.

$$\mathbf{A}_{1,1}=\frac{1}{\sqrt{2}}\begin{pmatrix}1&0\\0&1\\0&1\\1&0\end{pmatrix}, \tag{A1}$$

$$\mathbf{A}_{1,-1}=\frac{1}{\sqrt{2}}\begin{pmatrix}1&0\\0&1\\0&-1\\-1&0\end{pmatrix}, \tag{A2}$$

$$\mathbf{A}_{2,1}=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 0\\ 0 & 1\\ 1 & 0\\ 0 & -1\end{pmatrix}, \tag{A3}$$

$$\mathbf{A}_{2,-1}=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 0\\ 0 & 1\\ -1 & 0\\ 0 & 1\end{pmatrix}, \tag{A4}$$

$$\mathbf{A}_{3,1}=\begin{pmatrix}1 & 0\\ 0 & 1\\ 0 & 0\\ 0 & 0\end{pmatrix}, \tag{A5}$$

$$\mathbf{A}_{3,-1}=\begin{pmatrix}0 & 0\\ 0 & 0\\ 1 & 0\\ 0 & 1\end{pmatrix}. \tag{A6}$$

From Eqs. (12), (18)–(20) one easily recognizes that for the 2-planes $A_{3,1}$, $A_{3,-1}$ holds $A_{3,1}=O_{(1,2)}$, $A_{3,-1}=O_{(3,4)}$ ($O_{(1,2)}:\mathbf{x}_{(3,4)}=0$, $O_{(3,4)}:\mathbf{x}_{(1,2)}=0$).

APPENDIX B: TRANSFORMED PAULI MATRICES

As Pauli matrices [irreducible matrix representations of the complex Clifford algebra $C(3,0)$] play a significant role in theoretical physics, in this Appendix we wish to comment on the derivation of a particularly symmetric expression for these 2×2 matrices by means of the approach discussed in the main part of the paper. The standard expressions for the Pauli matrices read

$$\sigma_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix},\quad \sigma_2=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix},\quad \sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{B1}$$

In order to make contact with the main part of the paper it turns out to be useful to represent the complex numbers which are entries of the matrices (B1) by means of 2×2 matrices using the rule

$$z=a+ib\rightarrow\begin{pmatrix}a & -b\\ b & a\end{pmatrix}. \tag{B2}$$

This leads to a set of three real 4×4 matrices which are congruent to the gamma matrices given by Eq. (7). In order to obtain the desired final result we have to subject the latter gamma matrices to a further orthogonal transformation—an inversion [mentioned below Eq. (46)]. Then the rule (B2) can be reversed yielding the following transformed Pauli matrices ($k=1,2,3$):

$$\sigma'_k=\frac{1}{\sqrt{3}}\begin{pmatrix}1 & \sqrt{2}e^{-i\varphi_k}\\ \sqrt{2}e^{i\varphi_k} & -1\end{pmatrix}, \tag{B3}$$

$$\varphi_k=\varphi(k)=\varphi_0+\frac{2\pi}{3}k. \tag{B4}$$

Here, φ_0 is some arbitrary real constant which, however, has been shifted with respect to Eq. (9).

APPENDIX C: THE CASE OF THE REAL CLIFFORD ALGEBRA $C(2,1)$

In the present Appendix we want to illustrate the formalism used in the main part of the paper in the rather trivial case of the real Clifford algebra $C(2,1)$. We display the equations (including the notation) in close analogy to the discussion performed in the main part of the paper. We start with some explicit expressions for the gamma matrices [σ_k are the standard Pauli matrices (B1)],

$$\gamma_1=\sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \tag{C1}$$

$$\gamma_2=\sigma_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \tag{C2}$$

$$\gamma_3=i\sigma_2=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}. \tag{C3}$$

The eigenspaces of the gamma matrices γ_1, γ_2 are described by the following matrices:

$$\mathbf{A}_{1,1}=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\ 1\end{pmatrix}, \tag{C4}$$

$$\mathbf{A}_{1,-1}=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\ -1\end{pmatrix}, \tag{C5}$$

$$\mathbf{A}_{2,1}=\begin{pmatrix}1\\ 0\end{pmatrix}, \tag{C6}$$

$$\mathbf{A}_{2,-1}=\begin{pmatrix}0\\ 1\end{pmatrix}. \tag{C7}$$

It is clear that the angle between the eigenspaces (lines, 1-planes) which relate to different gamma matrices γ_1, γ_2 is $\pi/4$ (cf. Fig. 1).[68] Each line through the origin $\mathbf{x}=(0,0)$ is (trivially) isoclinic to each other such line. Therefore, the analogs of Eqs. (21), (23) are

$$x_2=\lambda x_1, \tag{C8}$$

$$x_1=\lambda' x_2, \quad \lambda'=\lambda^{-1}. \tag{C9}$$

Equations (25), (26) are mirrored by

$$\mathbf{B}(\lambda)=\frac{1}{\sqrt{1+\lambda^2}}\begin{pmatrix}1\\ \lambda\end{pmatrix}, \tag{C10}$$

and

$$\mathbf{B}(\lambda')=\frac{1}{\sqrt{1+\lambda'^2}}\begin{pmatrix}\lambda'\\ 1\end{pmatrix}. \tag{C11}$$

Of course, to each set of the eigenspaces $\{A_{1,1}, A_{2,1}\}$, $\{A_{1,-1}, A_{2,-1}\}$ two lines $A_{1\pm}$, $A_{-1\pm}$ exist, respectively, which lie symmetric with respect to the elements of the set (cf. Fig. 1). The analog of Eq. (27) reads (θ_k are the angles between the two eigenspaces to the eigenvalue $\rho=1$ [$\rho=-1$] and $A_{1\pm}$ [$A_{-1\pm}$])

$$\cos^2 2\theta_1 + \cos^2 2\theta_2 = 1. \tag{C12}$$

For $A_{1\pm}$, $A_{-1\pm}$ the relations [in analogy to Eqs. (28), (29)]

$$\theta_1 = \theta_2 = \theta_{\text{sym}}, \quad \cos 2\theta_{\text{sym}} = \pm\frac{1}{\sqrt{2}}, \tag{C13}$$

and

$$w = \cos^2\theta_{\text{sym}} = \frac{1}{2}(1+\cos 2\theta_{\text{sym}}) = \frac{1}{2}\left(1 \pm \frac{1}{\sqrt{2}}\right) = w_\pm \tag{C14}$$

are valid.

In analogy to Eqs. (30)–(35), in order to determine the lines $A_{1\pm}$, $A_{-1\pm}$ we have to solve the following equations for $\rho=1$ (in sequence for the indices $k=1$, $k=2$ of the gamma matrices, respectively)

$$w_\pm = \frac{(1+\lambda')^2}{2(1+\lambda'^2)}, \tag{C15}$$

$$w_\pm = \frac{\lambda'^2}{1+\lambda'^2}, \tag{C16}$$

and for $\rho=-1$,

$$w_\pm = \frac{(1-\lambda)^2}{2(1+\lambda^2)}, \tag{C17}$$

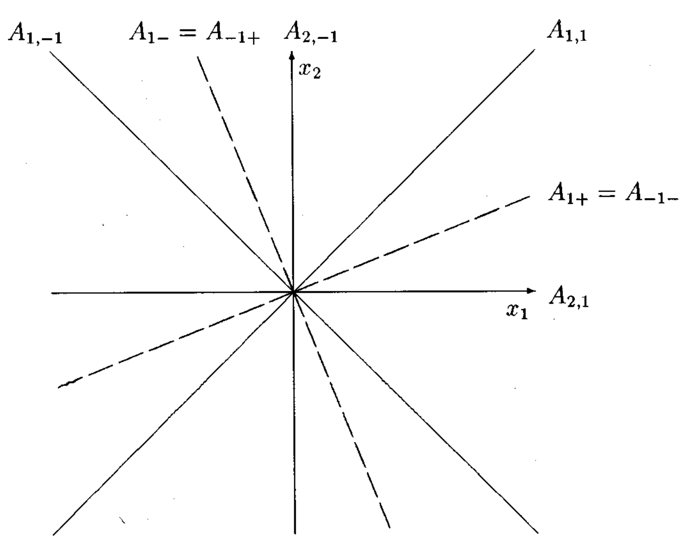

FIG. 1. Geometry of the eigenspaces of the gamma matrices γ_1, γ_2 [(C1), (C2)].

$$w_{\pm}=\frac{\lambda^2}{1+\lambda^2}. \tag{C18}$$

[Equations (C15), (C16) [(C17), (C18)] have been derived using Eq. (C11) [(C10)].] The solution of the above equations reads for $\rho=1$,

$$\lambda'=\lambda_{\pm}, \tag{C19}$$

and for $\rho=-1$,

$$\lambda=-\lambda_{\pm}. \tag{C20}$$

Here,

$$\lambda_{\pm}=\pm 2\sqrt{2}w_{\pm}, \tag{C21}$$

which entails

$$\lambda_{\pm}\lambda_{\mp}=-1. \tag{C22}$$

Inserting Eqs. (C20) and (C22) into Eqs. (C11) and (C10), respectively, one finds

$$\mathbf{A}_{1\pm}=\frac{1}{\sqrt{1+\lambda_{\pm}^2}}\begin{pmatrix}\lambda_{\pm}\\ 1\end{pmatrix}, \tag{C23}$$

$$\mathbf{A}_{-1\pm}=\frac{1}{\sqrt{1+\lambda_{\pm}^2}}\begin{pmatrix}1\\ -\lambda_{\pm}\end{pmatrix} \tag{C24}$$

(cf. Fig. 1; it holds $A_{1+}=A_{-1-}$, $A_{1-}=A_{-1+}$). The orthogonal transformation leading to the new coordinate system consequently reads

$$O_{\pm}=\frac{1}{\sqrt{1+\lambda_{\pm}^2}}\begin{pmatrix}\lambda_{\pm} & 1\\ 1 & -\lambda_{\pm}\end{pmatrix}. \tag{C25}$$

This way the following final result is obtained:

$$\gamma'_{1\pm}=\pm\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}, \tag{C26}$$

$$\gamma'_{2\pm}=\pm\frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ -1 & -1\end{pmatrix}, \tag{C27}$$

$$\gamma'_{3}=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}. \tag{C28}$$

It is clear that in the present case there is no residual continuous symmetry which has been exploited in Sec. V of the main part of the paper which is dealing with the real Clifford algebra $C(3,1)$.

[1] S. Okubo, in *Quarks, Symmetries and Strings*, Symposium in Honour of Bunji Sakita's 60th Birthday, New York, 1–2 October 1990, edited by M. Kaku, A. Jevicki, and K. Kikkawa (World Scientific, Singapore, 1991), p. 256.
[2] S. Okubo, J. Math. Phys. **32**, 1657 (1991); **32**, 1669 (1991).
[3] S. Okubo, Math. Japonica **41**, 59 (1995).

[4] Y. Brihaye, P. Maslanka, S. Giler, and P. Kosinski, J. Math. Phys. **33**, 1579 (1992).
[5] I. R. Porteous, *Topological Geometry*, 2nd ed. (Cambridge University Press, Cambridge, 1981).
[6] I. R. Porteous, *Clifford Algebras and the Classical Groups*, Cambridge Studies in Advanced Mathematics Vol. 50 (Cambridge University Press, Cambridge, 1995).
[7] P. Lounesto, *Clifford Algebras and Spinors*, London Mathematical Society Lecture Notes Series Vol. 239 (Cambridge University Press, Cambridge, 1997).
[8] R. Coquereaux, Phys. Lett. B **115**, 389 (1982).
[9] A. Dimakis, J. Phys. A **22**, 3171 (1989).
[10] J. Rembieliński, in *Clifford Algebras and their Applications in Mathematical Physics*, Proceedings of the Second Workshop held at Montpellier, France, 1989, edited by A. Micali, R. Boudet, and J. Helmstetter, Fundamental Theories of Physics Vol. 47 (Kluwer, Dordrecht, 1992), p. 97.
[11] R. Shaw, in *Clifford Algebras and Spinor Structures—A Special Volume Dedicated to the Memory of Albert Crumeyrolle (1919–1992)*, edited by R. Abłamowicz and P. Lounesto, Mathematics and Its Applications Vol. 321 (Kluwer, Dordrecht, 1995), p. 59.
[12] W. Pauli, Ann. Phys. (Leipzig) **18**, 337 (1933) [in German]; in *Pieter Zeeman, 1865—25 Mei—1935: Verhandelingen op 25 Mei 1935 Aangeboden aan Prof. Dr. P. Zeeman [Pieter Zeeman, 1865—May 25—1935: Proceedings of May 25, 1935 in Honour of Prof. Dr. P. Zeeman]* (Nijhoff, The Hague, 1935), p. 31 [in German]; Ann. Inst. Henri Poincaré **6**, 109 (1936) [in French]. Reprinted in *Collected Scientific Papers by Wolfgang Pauli*, edited by R. Kronig and V. F. Weisskopf (Interscience, New York, 1964), Vol. 2, pp. 662, 724, 753.
[13] R. H. Good, Jr., Rev. Mod. Phys. **27**, 187 (1955).
[14] A. A. Borisenko and Yu. A. Nikolaevskii, Usp. Mat. Nauk **46** (2), 41 (1991) [Russ. Math. Surveys **46** (2), 45 (1991)].
[15] B. A. Rozenfel'd, *Mnogomernye Prostranstva [Multidimensional Spaces]* (Nauka, Moscow, 1966) [in Russian].
[16] B. Rosenfeld, *Geometry of Lie Groups*, Mathematics and Its Applications Vol. 393 (Kluwer, Dordrecht, 1997).
[17] Equation (3.109) in Ref. 15, Sec. 3.3.15, p. 107 and Eq. (3.13) in Ref. 16, Sec. 3.3.3, p. 179; this matrix has also been considered in Ref. 18, Sec. 5, p. 807, Ref. 19, Ref. 20, Sec. 4, p. 1195, and Ref. 21, p. 242, Eq. (3). An equivalent definition based on a parametrization like (18), (19) is given in Ref. 22, Pt. I, Sec. 2, p. 13 and in Ref. 23, Sec. 1-3, p. 16, theorem 1-3.5. Further useful information with respect to the angles between planes can be found in Ref. 24, Sec. 31, 5., p. 392, Ref. 25, Ref. 26, Chap. 1, p. 23, Ref. 27, Sec. 7.5, p. 421, and Ref. 28 (also in Ref. 29 which, however, overlaps with Refs. 23 and 24). Also note Ref. 30 and references therein (Sec. 3, p. 18, bibliographical note). Finally, for the convenience of the reader we would like to mention some further recent references dealing with the angles between subspaces—Refs. 31–34.
[18] S. N. Afriat, Proc. Cambridge Philos. Soc. **53**, 800 (1957).
[19] H. Zassenhaus, Am. Math. Monthly **71**, 218 (1964).
[20] A. Ben-Israel, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **15**, 1184 (1967).
[21] A. Lanteri, Ist. Lombardo (Rend. Sci.) A **110**, 242 (1976) [in Italian].
[22] Y. C. Wong, *Isoclinic n-Planes in Euclidean 2n-Space, Clifford Parallels in Elliptic (2n-1)-Space, and the Hurwitz Matrix Equations*, Mem. American Mathematical Society Vol. 41 (1961) (AMS, Providence, RI, 1971), 2nd printing.
[23] Y. C. Wong, *Linear Geometry in Euclidean 4-Space*, Southeast Asian Mathematical Society Monograph No. 1 (SEAMS, Hong Kong, 1977).
[24] P. A. Shirokov, *Tenzornoe Ischislenie. Algebra Tenzorov [Tensor Calculus. Tensor Algebra]*, 2nd ed. (Kazan University Press, Kazan, 1961) [in Russian].
[25] Y. C. Wong, Proc. Natl. Acad. Sci. USA **57**, 589 (1967).
[26] L. D. Ivanov, *Variatsii Mnozhestv i Funktsii [Variations of Sets and Functions]* (Nauka, Moscow, 1975) [in Russian].
[27] D. S. Watkins, *Fundamentals of Matrix Computations* (Wiley, New York, 1991).
[28] I. C. F. Ipsen and C. D. Meyer, Am. Math. Monthly **102**, 904 (1995).
[29] K. Y. Woo, Southeast Asian Bull. Math. **3**, 287 (1979).
[30] C. Davis and W. M. Kahan, SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **7**, 1 (1970).
[31] P. A. Wedin, in *Matrix Pencils*, Proceedings of a Conference Held at Pite Havsbad, Sweden, 22–24 March 1982, edited by B. Kågström and A. Ruhe, Lecture Notes in Mathematics Vol. 973 (Springer, Berlin, 1983), p. 263.
[32] J. Miao and A. Ben-Israel, Linear Algebr. Appl. **171**, 81 (1992).
[33] J. Miao and A. Ben-Israel, Linear Algebr. Appl. **237/238**, 71 (1996).
[34] S. Jiang, Geom. Dedicata **63**, 113 (1996).
[35] Equation (3.110) in Ref. 15, Sec. 3.3.15, p. 108 and Eq. (3.15) in Ref. 16, Sec. 3.3.3, p. 179; this matrix has also been considered in Ref. 36, p. 136, Ref. 37, Sec. 2, Ref. 38, Chap. V, Sec. 3, p. 283, Ref. 39, p. 416, Eq. (5′), and Ref. 40, Sec. 1; also see Ref. 27, Sec. 7.5, corollary 7.5.9, p. 425.
[36] D. A. Flanders, in *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948* (Interscience, New York, 1948), p. 129.
[37] G. Venticos, Bull. Greek Math. Soc. **30**, 85 (1956) [in Greek, English abstract].
[38] H. Reichardt, *Vorlesungen über Vektor- und Tensorrechnung [Lectures on Vector and Tensor Calculus]*, Hochschulbücher für Mathematik Vol. 34 (Deutscher Verlag der Wissenschaften, Berlin, 1957) [in German].
[39] Č. Vitner, Časopis Pěst. Mat. **87**, 415 (1962) [in Czech, Russian and German abstracts].
[40] W. Degen, Rev. Roum. Math. Pures Appl. **21**, 503 (1976) [in German].
[41] Reference 15, Sec. 3.3.16, p. 109. Equation (17) has also been considered in Ref. 42, p. 99, Sec. 2.3; a related equation can be found in Ref. 43, Sec. 4, p. 144 and in Ref. 44, Sec. 3, p. 533, Eq. (3.2); also see Ref. 45, Sec. 2, p. 299, Eq. (2.3), and Ref. 46, Sec. 1, p. 481, (1.2). Instead of the term isoclinic also the terms *isocline* (Ref. 47) and *isogonal* (Ref. 18) have been used.
[42] P. W. H. Lemmens and J. J. Seidel, Proc. K. Ned. Akad. Wet., Ser. A: Math Sci. **76**, 98 (1973).

[43] M. Bruni, Ann. Mat. Pura Appl. **133**, 141 (1983) [in Italian].
[44] M. Bruni, Rend. Mat. Appl. **4**, 531 (1984) [in Italian, English abstract].
[45] J. Gunson, Ann. Inst. Henri Poincaré, Sect. A **17**, 295 (1972).
[46] D. B. Shapiro, Geom. Dedicata **7**, 479 (1978).
[47] H. P. Manning, *Geometry of Four Dimensions* (Macmillan, New York, 1914). Reprint: (Dover, New York, 1956).
[48] For some further information with respect to isoclinic 2-planes the reader is referred to Ref. 23 (which, however, does not contain any references) and Ref. 47, in particular Chap. III, Sec. III and Chap. IV, Sec. VI. Manning (Ref. 47) cites as his main source a comprehensive article by Stringham (Ref. 49) which contains a list of related references from the nineteenth century literature. A comprehensive overview of the literature on multidimensional Euclidean geometry before 1911 can be found in Ref. 50. Finally, the subject of isoclinic 2-planes has also been dealt with in monographs 51, 52, and 53 (the sequence indicates the depth of the discussion with Ref. 51 being the most comprehensive one among these three titles).
[49] I. Stringham, Trans. Am. Math. Soc. **2**, 183 (1901).
[50] D. M. Y. Sommerville, *Bibliography of Non-Euclidean Geometry, Including the Theory of Parallels, the Foundation of Geometry, and Space of N Dimensions* (Harrison, London, 1911); supplemented reprint: *Bibliography of Non-Euclidean Geometry* (Chelsea, New York, 1970).
[51] A. R. Forsyth, *Geometry of Four Dimensions* (Cambridge University Press, Cambridge, 1930), Vol. 1.
[52] P. H. Schoute, *Mehrdimensionale Geometrie [Multidimensional Geometry]*, 2 Vols., Sammlung Schubert, Vols. 35, 36 (Göschen, Leipzig, 1902, 1905), Vol. 1: *Erster Teil: Die linearen Räume [First Part: The Linear Spaces]* [in German].
[53] D. M. Y. Sommerville, *An Introduction to the Geometry of N Dimensions* (Methuen, London, 1929); reprint (Dover, New York, 1958).
[54] This is the maximal possible number of equi-isoclinic 2-planes with angle $\pi/4$ in $\mathbf{R}_4$ in general, cf. Ref. 42, Sec. 5, (5.1); incidentally, note some generalizations of this result obtained in Refs. 55–57 for complex and quaternionic spaces V.
[55] S. G. Hoggar, Ars. Comb. **2**, 11 (1976).
[56] S. G. Hoggar, in *Congressus Numerantium XV*, Proceedings of the Fifth British Combinatorial Conference 1975, University of Aberdeen, 14–18 July 1975, edited by C. St. J. A. Nash-Williams and J. Sheehan (Utilitas Mathematicas, Winnipeg, 1975), p. 319.
[57] S. G. Hoggar, Proc. Edinburgh Math. Soc. **20**, 287 (1976-77).
[58] The results of this paper have been reformulated from a different point of view in Ref. 59.
[59] J. A. Tyrrell and S. G. Semple, *Generalized Clifford Parallelism*, Cambridge Tracts in Mathematics and Mathematical Physics Vol. 61 (Cambridge University Press, Cambridge, 1971).
[60] In Ref. 61 the shorter term *isoclinic set of 2-planes* is used instead of the term ''maximal set of mutually isoclinic 2-planes.'' As we mainly rely on Ref. 22, in the present article we use the original term. Furthermore, every maximal set of mutually isoclinic 2-planes is also a *maximal normal set of 2-planes*—Refs. 61 and 62. The term *normal set* denotes a set of 2-planes which are pairwise either orthogonal or normally related to each other. Two 2-planes A, B are *normally related* if $A\cap B=A\cap B^{\perp}=\{0\}$. Finally, the relation of maximal sets of mutually isoclinic 2-planes to the Hopf map has been considered in Ref. 63.
[61] Y. C. Wong and K. P. Mok, Linear Algebr. Appl. **139**, 31 (1990).
[62] P. Yiu, Linear Algebr. Appl. **181**, 241 (1993).
[63] Y. C. Wong and K. P. Mok, Enseign. Math. **34**, 167 (1988).
[64] The corresponding expression of Wong—Ref. 22 is related to Eq. (21) by an orthogonal transformation (inversion) $(x_1,x_2,x_3,x_4)\rightarrow(-x_1,x_2,x_3,x_4)$. Equation (21) has been rederived in Ref. 65 by means of fairly elementary considerations. $\tilde{\mathbf{B}}(\lambda_0,\lambda_1)$ can be written as $\tan\theta\,\tilde{\mathbf{D}}$ ($\tan\theta=\sqrt{\lambda_0^2+\lambda_1^2}$), where θ is the angle between the 2-planes $O_{(1,2)}$, B, and $\tilde{\mathbf{D}}$ is an orthogonal matrix ($\tilde{\mathbf{D}}\in O(2)$)—Ref. 44.
[65] W. Reye, Rev. Colomb. Mat. **23**, 1 (1989) [in Spanish].
[66] J. E. Mebius, *Isocliny, Souillart matrices, quaternions and four-dimensional rotations*, Report No. 92-57 (Delft University of Technology, Faculty of Technical Mathematics and Informatics, Delft, 1992), available at the following URL: ftp://ftp.twi.tudelft.nl/pub/publications/tech-reports/1992/DUT-TWI-92-57.ps.gz.
[67] This difference between the Clifford subalgebras $C(3,0)$ and $C(0,2)$ is caused by the fact that during a full 2π turn around the axis (1,1,1) in the vector space $\mathbf{R}_{3,0}$ associated with $C(3,0)$ the rectangular coordinate system is mapped 3 times onto itself while during a full 2π turn in the vector space $\mathbf{R}_{0,2}$ associated with $C(0,2)$ the rectangular coordinate system is mapped 4 times onto itself (allowing reflections). Taking into account the different number of generators of $C(3,0)$ and $C(0,2)$, 3 and 2, respectively, this leads to a natural explanation for the difference.
[68] A related discussion can be found in Ref. 69, Vol. 2, *Mathematische Zusätze und Ergänzungen [Mathematical Supplements]*, no. 13: Zwei- und vierreihige Matrizen. Darstellung der hyperkomplexen γ-Einheiten durch Matrizen. Zu Kap. IV, § 5 [Two- and four-row matrices. Representation of the hypercomplex γ units by matrices. To Chap. IV, § 5], p. 780 of the cited German edition, also earlier editions of Ref. 69, Vol. 2, contain this material in some appendix.
[69] A. Sommerfeld, *Atombau und Spektrallinien*, 3rd ed. (Vieweg, Braunschweig, 1951); reprint (Verlag Harri Deutsch, Thun, 1978); English translation of an earlier edition: *Atomic Structure and Spectral Lines* (Methuen, London, 1934) (Dutton, New York, 1934), 3rd rev. ed.

Congruences and canonical forms for a positive matrix: Application to the Schweinler–Wigner extremum principle

R. Simon[a)]
The Institute of Mathematical Sciences, C. I. T. Campus, Chennai 600 113, India

S. Chaturvedi[b)] and V. Srinivasan[c)]
School of Physics, University of Hyderabad, Hyderabad 500 046, India

(Received 30 November 1998; accepted for publication 5 March 1999)

It is shown that a $N\times N$ real symmetric [complex Hermitian] positive definite matrix V is congruent to a diagonal matrix modulo a pseudo-orthogonal [pseudo-unitary] matrix in $\mathrm{SO}(m,n)[\mathrm{SU}(m,n)]$, for any choice of partition $N=m+n$. It is further shown that the method of proof in this context can easily be adapted to obtain a rather simple proof of Williamson's theorem which states that if N is even then V is congruent also to a diagonal matrix modulo a symplectic matrix in $\mathrm{Sp}(N,\mathcal{R})[\mathrm{Sp}(N,\mathcal{C})]$. Applications of these results considered include a generalization of the Schweinler–Wigner method of ‘‘orthogonalization based on an extremum principle’’ to construct pseudo-orthogonal and symplectic bases from a given set of linearly independent vectors.
[S0022-2488(99)00307-2]

I. INTRODUCTION

It is well known that an N-dimensional real symmetric [complex Hermitian] matrix V is congruent to a diagonal matrix modulo an orthogonal [unitary] matrix.[1] That is, $V=S^\dagger DS$, where D is diagonal and $S\in\mathrm{SO}(N)[S\in\mathrm{SU}(N)]$. If, in addition, V is also positive definite, new possibilities arise for establishing its congruence to a diagonal matrix. For N even, it was shown by Williamson[2] some 60 years ago, and subsequently by several authors,[3,4] that such a V is also congruent to a diagonal matrix modulo a symplectic matrix in $\mathrm{Sp}(N,\mathcal{R})[\mathrm{Sp}(N,\mathcal{C})]$. That is, $V>0$ implies $V=S^\dagger D' S$, where D' is diagonal and $S\in\mathrm{Sp}(N,\mathcal{R})[S\in\mathrm{Sp}(N,\mathcal{C})]$. Williamson's theorem has recently been exploited in defining quadrature squeezing and symplectically covariant formulation of the uncertainty principle for multimode states.[5] In this work we establish yet another kind of congruence of a real symmetric [complex Hermitian] positive definite matrix to a diagonal matrix valid, for both odd and even dimensions. We show that an N-dimensional real symmetric [complex Hermitian] positive definite matrix V is congruent to a diagonal matrix modulo a pseudo-orthogonal [pseudo-unitary] matrix. That is, $V>0$ implies $V=S^\dagger D'' S$, where D'' is diagonal and $S\in\mathrm{SO}(m,n)[S\in\mathrm{SU}(m,n)]$, for any choice of partition $N=m+n$. A simple proof of this result is given. The strategy adopted in proving this result, with appropriate modification, works for the Williamson case as well, and affords a particularly simple proof of Williamson's theorem. Needless to add that the diagonal entries of neither D' nor D'' correspond to the eigenvalues of V.

The theorems established here play a crucial role in enabling one to construct pseudo-orthogonal and symplectic bases from a given set of linearly independent vectors via an extremum principle in the spirit of the work of Schweinler and Wigner.[6] In an important contribution to the age old ‘‘orthogonalization problem’’—the problem of constructing an orthonormal set of vectors

[a)]Electronic mail: simon@imsc.ernet.in
[b)]Electronic mail: scsp@uohyd.ernet.in
[c)]Electronic mail: vssp@uohyd.ernet.in

from a given set of linearly independent vectors—Schweinler and Wigner proposed an orthonormal basis which, unlike the familiar Gram–Schmidt basis (which depends on the particular initial order in which the given linearly independent vectors are listed), treats all the linearly independent vectors on an equal footing and has since found important application in wavelet analysis.[7] More significantly, they showed that this special basis follows from *an extremum principle*. In this work, we exploit our results on congruence to obtain generalizations of the Schweinler–Wigner extremum principle leading to pseudo-orthogonal and symplectic bases from a given set of linearly independent vectors. Conversely, the extremum principle, once formulated, can be interpreted as a procedure for finding the appropriate congruence transformation to effect the desired diagonalization.

II. CONGRUENCE OF A POSITIVE MATRIX UNDER PSEUDO-ORTHOGONAL (PSEUDO-UNITARY) TRANSFORMATIONS

The fact that a real symmetric [complex Hermitian] matrix is congruent to a diagonal matrix modulo an orthogonal [unitary] matrix is well known. While congruence coincides with conjugation in the real orthogonal and complex unitary cases, they become distinct when more general sets of transformations are involved. A question which naturally arises is whether congruence to a diagonal form can also be achieved through a pseudo-orthogonal [pseudo-unitary] transformation. The answer to this question turns out to be in the affirmative with the caveat that the matrix in question be positive definite, and can be formulated as the following theorem:

Theorem 1: Let V be a real symmetric positive definite matrix of dimension N. Then, for any choice of partition $N=m+n$, there exists an $S\in SO(m,n)$ such that

$$S^TVS=D^2=\text{diagonal} \ \ (\text{and} \ >0). \tag{1}$$

Proof: We begin by recalling that the group $\mathrm{SO}(m,n)$ consists of all real matrices which satisfy $S^TgS=g$, $\det S=1$, where

$$g=\mathrm{diag}(\underbrace{1,1,\ldots,1}_{m},\underbrace{-1,\ldots,-1}_{n}).$$

Consider the matrix $V^{-1/2}gV^{-1/2}$ constructed from the given matrix V. Since $V^{-1/2}gV^{-1/2}$ is real symmetric, there exists a rotation matrix $R\in\mathrm{SO}(N)$ which diagonalizes $V^{-1/2}gV^{-1/2}$,

$$R^TV^{-1/2}gV^{-1/2}R=\text{diagonal}\equiv\Lambda. \tag{2}$$

This may be viewed also as a congruence of g using $V^{-1/2}R$, and signatures are preserved under congruence. (Indeed, signatures are the only invariants if we allow congruence over the full linear group $GL(N,\mathcal{R})$.) As a consequence, the diagonal matrix Λ can be expressed as the product of a positive diagonal matrix and g,

$$R^TV^{-1/2}gV^{-1/2}R=D^{-2}g=D^{-1}gD^{-1}. \tag{3}$$

Here D is diagonal and positive definite.

Taking the inverse of the matrices on both sides of (3) we find that the diagonal entries of $gD^2=D^2g$ are the eigenvalues of $V^{1/2}gV^{1/2}$ and that the columns of R are the eigenvectors of $V^{1/2}gV^{1/2}$. Since $V^{1/2}gV^{1/2}$, gV, and Vg are conjugate to one another, we conclude that D^2 is determined by the eigenvalues of $gV\sim Vg$.

Define $S=V^{-1/2}RD$. It may be verified that S satisfies the following two equations:

$$S^TgS=g,$$

$$S^TVS=D^2=\text{diagonal}. \tag{4}$$

The first equation says that $S\in SO(m,n)$ and the second says that V is diagonalized through congruence by S. Hence the proof.

By replacing the superscript T by †, the group SO(m,n) by SU(m,n), and $R\in SO(N)$ by $U\in \mathrm{SU}(N)$ in the statement and proof of the above theorem, we have the following theorem which applies to the complex case.

Theorem 2: Let V be a Hermitian positive definite matrix of dimension N. Then, for any partition $N=m+n$, there exists an $S\in \mathrm{SU}(m,n)$ such that

$$S^{\dagger}VS=D^2=\text{diagonal (and } >0). \tag{5}$$

III. A SIMPLE PROOF OF WILLIAMSON'S THEOREM

It turns out that the above procedure when applied to the real symplectic group of linear canonical transformations leads a particularly simple proof of Williamson's theorem.

Theorem 3: Let V be a $2n$-dimensional real symmetric positive definite matrix. Then there exists an $S\in \mathrm{Sp}(2n,\mathcal{R})$ such that

$$S^TVS=D^2>0,$$

$$D^2=\mathrm{diag}(\kappa_1,\kappa_2,\ldots,\kappa_n,\kappa_1,\kappa_2,\ldots,\kappa_n). \tag{6}$$

Proof: Note that the $2n$-dimensional diagonal matrix D has only n independent entries. The group Sp$(2n,\mathcal{R})$ consists of all real matrices S which obey the condition

$$S^T\beta S=\beta,\quad \beta=\begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}, \tag{7}$$

with 1 and 0 denoting the $n\times n$ unit and zero matrices, respectively. Even though $S^T\beta S=\beta$ may appear to suggest that $\det S=\pm 1$, it turns out that $\det S=1$. In other words, Sp$(2n,\mathcal{R})$ consists of just one connected (though not simply connected) piece. Indeed, for every $n\geqslant 1$ the connectivity property of Sp$(2n,\mathcal{R})$ is the same as that of the circle.

The most general $S\in GL(2n,\mathcal{R})$ which solves $S^TVS=D^2$ is $S=V^{-1/2}RD$, where $R\in O(2n)$. Note that none of the factors D, R or $V^{-1/2}$ is an element of Sp$(2n,\mathcal{R})$. However, a V-dependent choice of D,R can be so made that the product $V^{-1/2}RD$ is an element of Sp$(2n,\mathcal{R})$ as we shall now show.

Since $\beta^T=-\beta$, it follows that $\mathcal{M}=V^{-1/2}\beta V^{-1/2}$ is antisymmetric. Hence there exists an $R\in SO(2n)$ such that[8]

$$R^TV^{-1/2}\beta V^{-1/2}R=\begin{pmatrix} 0 & \Omega\\ -\Omega & 0\end{pmatrix},\quad \Omega=\text{diagonal}>0. \tag{8}$$

Define a diagonal positive definite matrix

$$D=\begin{pmatrix} \Omega^{-1/2} & 0\\ 0 & \Omega^{-1/2}\end{pmatrix}. \tag{9}$$

Then we have

$$DR^TV^{-1/2}\beta V^{-1/2}RD=\beta. \tag{10}$$

Now define $S=V^{-1/2}RD$. It may be verified that S enjoys the following properties:

$$S^T\beta S=\beta,$$

$$S^TVS=D^2=\text{diagonal}. \tag{11}$$

The first equation says that $S \in \mathrm{Sp}(2n,\mathcal{R})$ and the second one says that V is diagonalized by congruence through S. This completes the proof of the Williamson theorem. To appreciate the simplicity of the present the reader may like to compare it with two recently published proofs of the Williamson theorem.[4]

We wish to explore the structure underlying the above proof a little further so that the relationship between D and S in (11) on the one hand and the eigenvalues and eigenvectors of βV^{-1} (or $V^{-1/2}\beta V^{-1/2}$) on the other becomes transparent. Again consider the matrix $\mathcal{M} = V^{-1/2}\beta V^{-1/2}$. It is a real, nonsingular, antisymmetric matrix and hence its eigenvalues $i\omega_\alpha$ and eigenvectors η_α have the following properties:

$$\mathcal{M}\eta_\alpha = i\omega_\alpha \eta_\alpha, \quad \alpha = 1,...,2n;$$
$$\omega_k > 0, \quad k=1,...,n; \quad \omega_{n+k} = -\omega_k;$$
$$\eta_{n+k} = \eta_k^*; \quad k=1,...,n. \tag{12}$$

The eigenvectors η_α can be chosen to be orthonormal even when the eigenvalues $i\omega_\alpha$ are degenerate. Arrange the eigenvectors η_α as columns of a matrix U. The matrix U thus obtained clearly belongs to the unitary group $U(2n)$, and satisfies

$$U^\dagger \mathcal{M} U = \Lambda, \quad \Lambda = \begin{pmatrix} i\Omega & 0 \\ 0 & -i\Omega \end{pmatrix}, \tag{13}$$

where $\Omega = \mathrm{diag}(\omega_1,...,\omega_n) > 0$. Now define the following $2n \times 2n$ unitary matrices:

$$\Sigma = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \Delta = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}. \tag{14}$$

These two matrices have the properties $\Sigma^2 = 1$, $U\Sigma = U^*$, and $\Sigma\Delta = \Delta^*$ ($*$ denotes the complex conjugate of a matrix). As a useful consequence of these properties we have

$$U^*\Delta^* = U^*\Sigma\Sigma\Delta^* = U\Delta. \tag{15}$$

We find that the unitary matrix $U\Delta$ is real; $U\Delta \in O(2n)$.

Now consider $S = V^{-1/2}U\Delta D$, where D is a diagonal matrix to be determined. It follows from the definition of S and the reality of $U\Delta \in O(2n)$ that

$$S^T V S = S^\dagger V S = D^2. \tag{16}$$

Further, recalling that $U^\dagger \mathcal{M} U = \Lambda$ we obtain

$$S^T \beta S = S^\dagger \beta S = D\Delta^\dagger U^\dagger \mathcal{M} U \Delta D = D\Delta^\dagger \Lambda \Delta D = D\begin{pmatrix} O & \Omega \\ -\Omega & O \end{pmatrix} D. \tag{17}$$

It is now evident that the following choice for D ensures that S is an element of $S \in \mathrm{Sp}(2n,\mathcal{R})$:

$$D = \begin{pmatrix} \Omega^{-1/2} & O \\ O & \Omega^{-1/2} \end{pmatrix}. \tag{18}$$

This completes our analysis of the manner in which S and D are related to the eigenvalues and eigenvectors of the matrix βV^{-1}.

As in the pseudo-orthogonal case, by replacing the superscript T by † in the statement and proof of Theorem 3, one obtains the following result:

Theorem 4: Let V be a $2n$-dimensional Hermitian positive definite matrix. Then there exists an $S\in \mathrm{Sp}(2n,\mathcal{C})$ such that

$$S^{\dagger}VS=D^2>0,$$
$$D^2=\mathrm{diag}(\kappa_1,\kappa_2,\ldots,\kappa_n,\kappa_1,\kappa_2,\ldots,\kappa_n). \tag{19}$$

An immediate consequence of the theorems stated above is that for a real symmetric [complex Hermitian] positive definite matrix we can not talk about *the* canonical form under congruence, for there are $m+n$ possible choices of $\mathrm{SO}(m,n)[\mathrm{SU}(m,n)]$, and in the case of even dimension one more choice coming from Williamson's theorem. Needless to add that for the same matrix V, the diagonal matrix D will be different for different choices.

IV. ORTHOGONALIZATION PROCEDURES

Assume that we are given a set of linearly independent N-dimensional vectors $v_1,\ldots,v_N$. Let G denote the associated Gram matrix of pairwise inner products, $G_{ij}=(v_i,v_j)$. The Gram matrix is Hermitian by construction, and positive definite by virtue of the linear independence of the given vectors. The orthogonalization problem, i.e., constructing a set of orthonormal vectors out of the given set of linearly independent vectors, amounts to finding a matrix S that solves

$$S^{\dagger}GS=1, \quad \text{i.e.,} \quad G^{-1}=SS^{\dagger}. \tag{20}$$

Each such S defines an orthogonalization procedure.

Let us arrange the set of N vectors as the entries of a row $\mathbf{v}=(v_1,v_2,\ldots,v_N)$, and let $\mathbf{z}=(z_1,z_2,\ldots,z_N)$ represent a generic orthonormal basis. The orthonormal set of vectors $\mathbf{z}$ corresponding to a chosen S are related to the given set of linearly independent vectors through $\mathbf{z}=\mathbf{v}S$. Clearly, there are infinitely many choices for S satisfying (20); given an S satisfying (20), any $S'=SU$, where U is an arbitrary unitary matrix also satisfies (20). Thus the freedom available for the solution of the orthonormalization problem is exactly as large as the unitary group $\mathrm{U}(N)$, and this was to be expected.

Schweinler and Wigner[6] posed and answered the following question: is there a way of discriminating between various choices of S that solves (20) and hence between various orthogonalization procedures? They argued that a particular choice of orthogonalization procedure should correspond ultimately to the extremization of a suitable scalar function over the manifold of all orthonormal bases, with the given linearly independent vectors appearing as parameters in the function. Different choices of onthonormal bases will then correspond to different functions to be extremized. They preferred the function to be symmetric under permutation of the given vectors. As an example they considered the following function which is quartic in the given vectors:

$$m(\mathbf{z})=\sum_k \left(\sum_l |(z_k,v_l)|^2\right)^2. \tag{21}$$

They showed that the extremum (maximum in this case) value of $m(\mathbf{z})$ is given by $\mathrm{tr}(G^2)$, and this value corresponds to the orthonormal basis $\mathbf{z}=\mathbf{v}U_0P^{-1/2}$, where U_0 is the unitary matrix which diagonalizes G: $U_0{}^{\dagger}GU_0=P$. We may refer to this as the Schweinler–Wigner basis, and the function $m(\mathbf{z})$ as the Schweinler–Wigner quartic form. It is clear that U_0 and hence the Schweinler–Wigner basis is essentially unique if the eigenvalues of the Gram matrix G are all distinct. We may note in passing that, unlike the Gram–Schmidt orthogonalization procedure, the Schweinler–Wigner procedure is democratic in that it treats all the linearly independent vectors $\mathbf{v}$ on an equal footing.

The content of the work of Schweinler and Wigner has recently been reformulated[9] in a manner that offers a clearer and more general picture of the Schweinler–Wigner quartic form

$m(\mathbf{z})$ and of the orthonormal basis which maximizes it. This perspective on the orthogonalization problem plays an important role in our generalizations of the Schweinler–Wigner extremum principle, and hence we summarize it briefly.

Since every orthonormal basis is the eigenbasis of a suitable Hermitian operator, it is of interest to characterize the Schweinler–Wigner basis in terms of such an operator. Given linearly independent N-dimensional vectors $\mathbf{v}=(v_1,v_2,\ldots,v_N)$, the operator $\hat{M}=\Sigma_j v_j v_j^\dagger$ is Hermitian positive definite. In a *generic orthonormal* basis $\mathbf{z}$, it is represented by a Hermitian positive definite matrix $M(\mathbf{z})$:$M(\mathbf{z})_{ij}=(z_i,\hat{M}z_j)$. Under a change of orthonormal basis $\mathbf{z}\rightarrow\mathbf{z}'=\mathbf{z}S$, $M(\mathbf{z})$ transforms as follows:

$$M(\mathbf{z})\rightarrow M(\mathbf{z}')=S^\dagger M(\mathbf{z})S,\quad S\in U(N). \tag{22}$$

Recall that $U(N)$ acts transitively on the set of all orthonormal bases and that $\mathrm{tr}(M(\mathbf{z})^2) = \sum_{j,k}|M(\mathbf{z})_{jk}|^2$ is invariant under such a change of basis, and hence is independent of $\mathbf{z}$. The Schweinler–Wigner quartic form $m(\mathbf{z})$ can easily be identified as $\sum_k(M(\mathbf{z})_{kk})^2$. In view of the above invariance, maximization of $\sum_k(M(\mathbf{z})_{kk})^2$ is the same as minimization of $\sum_{j\neq k}|M(\mathbf{z})_{jk}|^2$. The absolute minimum of $\sum_{j\neq k}|M(\mathbf{z})_{jk}|^2$ equals zero, and obtains when $M(\mathbf{z})$ is diagonal. Thus, the orthonormal basis which maximizes $\sum_k(M(\mathbf{z})_{kk})^2$ is the same as the one in which $\hat{M}$ is diagonal, and we arrive at the following important conclusion of Ref. 9:

Theorem 5: The distinquished orthonormal basis which extremizes the Schweinler–Wigner quartic form $m(\mathbf{z})$ over the manifold of all orthonormal bases is the same as the orthonormal basis in which the positive definite matrix $M(\mathbf{z})$ becomes diagonal.

Important for the above structure is the fact that the invariant $\mathrm{tr}(M(\mathbf{z})^2)$ is the sum of non-negative quantities, and therefore a part of it is necessarily bounded. It is precisely this property, which can be traced to the underlying unitary symmetry, that is not available when we try to generalize the Schweinler–Wigner procedure to construct pseudo-orthonormal and symplectic bases wherein the underlying symmetries are the noncompact groups $SO(m,n)$ and $\mathrm{Sp}(2n,\mathcal{R})$ respectively.

V. LORENTZ BASIS WITH AN EXTREMUM PROPERTY

In this section we show how the Schweinler–Wigner procedure can be generalized to construct pseudo-orthonormal basis based on an extremum principle. We begin with the case of real vectors.

We are given a set of linearly independent real N-dimensional vectors $\mathbf{v}=(v_1,\ldots,v_N)$ and we want to construct out of it a pseudo-orthonormal basis $[SO(m,n)$ Lorentz basis with $N=m+n]$, i.e., a set of vectors $\mathbf{z}=(z_1,\ldots,z_N)$ satisfying

$$(z_k,gz_l)=g_{kl},\quad g=\mathrm{diag}(\underbrace{1,1,\ldots,1}_{m},\underbrace{-1,\ldots,-1}_{n}).$$

Let $\hat{M}=\sum_j v_j v_j^T$ as before, and let the symmetric positive definite matrix $M(\mathbf{z})$:$M(\mathbf{z})_{ij} =(z_i,\hat{M}z_j)$ represent $\hat{M}$ in a *generic pseudo-orthonormal* basis $\mathbf{z}$. Under a pseudo-orthogonal change of basis $\mathbf{z}\rightarrow\mathbf{z}'=\mathbf{z}S$, the matrix $M(\mathbf{z})$ transforms as follows:

$$M(\mathbf{z})\rightarrow M(\mathbf{z}')=S^TM(\mathbf{z})S,\quad S\in SO(m,n). \tag{24}$$

Since $S^TgS=g$ (or $gS^T=S^{-1}g$) by definition, we have

$$S:\quad gM(\mathbf{z})\rightarrow gM(\mathbf{z}')=S^{-1}gM(\mathbf{z})S. \tag{25}$$

That is, as $M(\mathbf{z})$ undergoes congruence, $gM(\mathbf{z})$ undergoes conjugation. Thus, $\mathrm{tr}(gM(\mathbf{z}))^l$, $l=1,2,...$, are invariant. In what follows we shall often leave implicit the dependence of M on the generic pseudo-orthonormal basis $\mathbf{z}$.

Consider the invariant $\mathrm{tr}(gM(\mathbf{z})gM(\mathbf{z}))$ corresponding to $l=2$. Write $M=M^{\mathrm{even}}+M^{\mathrm{odd}}$, where

$$M^{\mathrm{even}}=\tfrac{1}{2}(M+gMg), \quad M^{\mathrm{odd}}=\tfrac{1}{2}(M-gMg). \tag{26}$$

In the above decomposition we have exploited the fact that g is, like parity, an *involution*.

With M expressed in the (m,n) block form

$$M=\begin{pmatrix} A & C \\ C^T & B \end{pmatrix}, \quad A^T=A, \quad B^T=B, \tag{27}$$

we have

$$M^{\mathrm{even}}=\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \quad M^{\mathrm{odd}}=\begin{pmatrix} 0 & C \\ C^T & 0 \end{pmatrix}. \tag{28}$$

Symmetry of M implies that M^{odd} and M^{even} are symmetric. Further, M^{odd} and M^{even} are trace orthogonal; $\mathrm{tr}(M^{\mathrm{odd}}M^{\mathrm{even}})=0$. Thus,

$$\mathrm{tr}(gMgM)=\mathrm{tr}(M^{\mathrm{even}})^2-\mathrm{tr}(M^{\mathrm{odd}})^2, \tag{29}$$

which can also be written as

$$\mathrm{tr}(MgMg)=\mathrm{tr}(M^2)-2\,\mathrm{tr}(M^{\mathrm{odd}})^2. \tag{30}$$

A few observations are in order:

(1) In contradistinction to the original unitary case, the invariant in the present case is no more a sum of squares. This can be traced to the noncompactness of the underlying $SO(m,n)$ symmetry. As one consequence, $\sum_k (M_{kk})^2$ is not bounded. As an example, consider the simplest case $m=1$, $n=1$ and let

$$M=\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad a,b>0. \tag{31}$$

Under congruence by the $SO(1,1)$ element

$$S=\begin{pmatrix} \cosh\mu & \sinh\mu \\ \sinh\mu & \cosh\mu \end{pmatrix}, \tag{32}$$

the value of $\sum_k (M_{kk})^2$ changes from a^2+b^2 to $a^2+b^2+2ab\sinh^2\mu\cosh^2\mu$, which grows with μ without bounds, showing that $\sum_k (M_{kk})^2$ and hence $\mathrm{tr}(M^2)$ is not bounded. Thus, in contrast to the unitary case, extremization of the Schweinler–Wigner quartic form $\sum_k (M_{kk})^2$ will make no sense in the absence of further restrictions.

(2) The structure of the invariant $\mathrm{tr}(gMgM)$ in (30) suggests the further restriction needed to be imposed; within the submanifold of pseudo-orthogonal bases $\mathbf{z}$ which keep $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2$ (and hence $\mathrm{tr}(M(\mathbf{z})^2)$) at a fixed value we can maximize $\sum_k M(\mathbf{z})_{kk}^2$. In particular we can do this within the submanifold which minimizes $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2$, and hence $\mathrm{tr}(M(\mathbf{z})^2)$. Clearly, zero is

the absolute minimum of the nonnegative object $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2$. But by Theorem 1 there exists a Lorentz basis $\mathbf{z}$ in which $M(\mathbf{z})$ is diagonal and hence $M(\mathbf{z})^{\mathrm{odd}}=0$. Thus the minimum $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2=0$, and hence the minimum of $\mathrm{tr}(M(\mathbf{z})^2)$, namely, $\mathrm{tr}(gM(\mathbf{z})gM(\mathbf{z}))$, is attainable.

The above observations suggest the following *two step analog of the Schweinler–Wigner extremum principle for Lorentz bases*. Choose the submanifold of Lorentz bases which minimize the quartic form $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2$, and maximize the Schweinler–Wigner quartic form $m(\mathbf{z})=\sum_k (M(\mathbf{z})_{kk})^2$ within this submanifold. Clearly, the first step takes M to a block-diagonal form, and the second one diagonalizes it. Thus we have established the following generalization of Theorem 5 to the pseudo-orthonormal case:

Theorem 6: The distinquished pseudo-orthonormal basis which extremizes the ''Schweinler–Wigner'' quartic form $m(\mathbf{z})$ over the submanifold of pseudo-orthonormal bases which minimize the quartic form $\mathrm{tr}(M(\mathbf{z})^2)$ is the same as the pseudo-orthonormal basis in which the positive definite matrix $M(\mathbf{z})$ becomes diagonal.

The submanifold under reference consists of Lorentz bases which are related to one another through the maximal compact (connected) subgroup of $\mathrm{SO}(m,n)$, namely $\mathrm{SO}(m)\times SO(n)$. This subgroup consists of matrices of the block-diagonal form

$$\begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix}, \quad R_1\in \mathrm{SO}(m), \quad R_2\in \mathrm{SO}(n), \tag{33}$$

and this is precisely the subgroup of $\mathrm{SO}(m,n)$ transformations that do not mix the even and odd parts of $M(\mathbf{z})$.

To conclude this section we may note that the above construction carries over to the complex case, with obvious changes like replacing T by † and $\mathrm{SO}(m,n)$ by $\mathrm{SU}(m,n)$.

VI. SYMPLETIC BASIS WITH AN EXTREMUM PROPERTY

Our construction in the pseudo-orthogonal case suggests a scheme by which the Schweinler–Wigner extremum principle can be generalized to construct a symplectic basis. Suppose that we are given a set of linearly independent vectors $\mathbf{v}=(v_1,v_2,\ldots,v_{2n})$ in $\mathcal{R}^{2n}$. The natural symplectic structure in R^{2n} is specified by the standard symplectic ''metric'' β defined in (7). Let $\mathbf{z}=(z_1,z_2,\ldots,z_{2n})$ denote a generic symplectic basis. That is, $(z_j,\beta z_k)=\beta_{jk}$, $j,k=1,2,...,2n$. The real symplectic group $\mathrm{Sp}(2n,R)$ acts transitively on the set of all symplectic bases.

To generalize the Schweinler–Wigner principle to the symplectic case, we begin be defining $\hat{M}=\sum_{j=1}^{2n} v_j v_j^T$. Let $M(\mathbf{z})$:$M(\mathbf{z})_{ij}=(z_i,\hat{M}z_j)$ be the symmetric positive definite matrix representing the operator $\hat{M}$ in a *generic symplectic* basis $\mathbf{z}$. Under a symplectic change of basis $\mathbf{z}\to\mathbf{z}'=\mathbf{z}S$, $S\in\mathrm{Sp}(2n,\mathcal{R})$, the matrix $M(\mathbf{z})$ undergoes the following transformation:

$$M(\mathbf{z})\to M(\mathbf{z}')=S^T M(\mathbf{z})S, \quad S\in \mathrm{Sp}(2n,R). \tag{34}$$

Since $S^T\beta S=\beta$ implies $\beta S^T=S^{-1}\beta$, we have

$$S: \quad \beta M(\mathbf{z})\to\beta M(\mathbf{z}')=S^{-1}\beta M(\mathbf{z})S. \tag{35}$$

That is, under a symplectic change of basis $M(\mathbf{z})$ undergoes congruence, but $\beta M(\mathbf{z})$ undergoes conjugation. Hence $\mathrm{tr}(\beta M(\mathbf{z}))^{2l}$, $l=1,2,...,n$ are invariant. (Note that $\mathrm{tr}(\beta M(\mathbf{z}))^{2l+1}=0$ in view of $\beta^T=-\beta$, $M(\mathbf{z})^T=M(\mathbf{z})$.)

Since $i\beta$ is an *involution* we can use it to separate $M(\mathbf{z})$ into even and odd parts,

$$M(\mathbf{z})=M(\mathbf{z})^{\mathrm{even}}+M(\mathbf{z})^{\mathrm{odd}},$$

$$M(\mathbf{z})^{\mathrm{even}}=\tfrac{1}{2}(M(\mathbf{z})+\beta M(\mathbf{z})\beta^T),$$
$$M(\mathbf{z})^{\mathrm{odd}}=\tfrac{1}{2}(M(\mathbf{z})-\beta M(\mathbf{z})\beta^T). \tag{36}$$

The even and odd parts of $M(\mathbf{z})$ satisfy the symmetry properties

$$\beta M(\mathbf{z})^{\mathrm{even}}\beta^T=M(\mathbf{z})^{\mathrm{even}},\quad \beta M(\mathbf{z})^{\mathrm{odd}}\beta^T=-M(\mathbf{z})^{\mathrm{odd}}. \tag{37}$$

Further. $M(\mathbf{z})^{\mathrm{odd}}$ and $M(\mathbf{z})^{\mathrm{even}}$ are trace orthogonal; $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}}M(\mathbf{z})^{\mathrm{even}})=0$.

The structure of the even and odd parts of $M(\mathbf{z})$ may be appreciated by writing $M(\mathbf{z})$ in the block form

$$M(\mathbf{z})=\begin{pmatrix} A & C \\ C^T & B \end{pmatrix},\quad A^T=A,\quad B^T=B. \tag{38}$$

We have

$$M(\mathbf{z})^{\mathrm{even}}=\begin{pmatrix} \tfrac{1}{2}(A+B) & \tfrac{1}{2}(C-C^T) \\ -\tfrac{1}{2}(C-C^T) & \tfrac{1}{2}(A+B) \end{pmatrix},$$
$$M(\mathbf{z})^{\mathrm{odd}}=\begin{pmatrix} \tfrac{1}{2}(A-B) & \tfrac{1}{2}(C+C^T) \\ \tfrac{1}{2}(C+C^T) & \tfrac{1}{2}(B-A) \end{pmatrix}. \tag{39}$$

Now consider the invariant $-\mathrm{tr}(\beta M(\mathbf{z})\beta M(\mathbf{z}))=\mathrm{tr}(\beta^T M(\mathbf{z})\beta M(\mathbf{z}))$. We have

$$\mathrm{tr}(\beta^T M(\mathbf{z})\beta M(\mathbf{z}))=\mathrm{tr}(M(\mathbf{z})^{\mathrm{even}})^2-\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2, \tag{40}$$

which can also be written as

$$\mathrm{tr}(\beta^T M(\mathbf{z})\beta M(\mathbf{z}))=\mathrm{tr}(M(\mathbf{z})^2)-2\,\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2. \tag{41}$$

The structural similarity of this invariant to that in the pseudo-orthogonal case should be appreciated.

Now, by an argument similar to the pseudo-orthogonal case one finds that, owing to the noncompactness of $\mathrm{Sp}(2n,\mathcal{R})$, the function $\mathrm{tr}(M(\mathbf{z})^2)$ and hence the Schweinler–Wigner quartic form $\sum_{k=1}^{2n}(M(\mathbf{z})_{kk})^2$ is unbounded if $\mathbf{z}$ is allowed to run over the entire manifold of all symplectic bases. For instance, in the lowest dimensional case $n=1$ with M chosen to be

$$M=\begin{pmatrix} a & u \\ d & b \end{pmatrix},\quad a,b>0,\quad ab-ud>0, \tag{42}$$

under congruence by the $\mathrm{Sp}(2,,\mathcal{R})$ matrix

$$S=\begin{pmatrix} \mu & 0 \\ 0 & 1/\mu \end{pmatrix}, \tag{43}$$

the value of $\sum_k(M_{kk})^2$ changes from a^2+b^2 to $\mu^2a^2+(1/\mu^2)b^2$ which, by an appropriate choice of μ, can be made as large as one wishes.

However, it follows from (41) that over the submanifold of symplectic bases which leave $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2$ fixed, the function $\mathrm{tr}(M(\mathbf{z})^2)$ remains invariant and so the quartic form $\Sigma(M(\mathbf{z})_{kk})^2$ is bounded within this restricted class of symplectic bases and hence can be maximized. In particular the nonnegative $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2$ can be chosen to take its minimum value. Williamson theorem implies that there are symplectic bases which realize the absolute minimum $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2=0$.

We can now formulate the *analog of the Schweinler–Wigner extremum principle for symplectic bases* in the following way: Take the subfamily of symplectic bases in which $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2$ and hence $\mathrm{tr}(M(\mathbf{z})^2)$ is minimum. [This minimum of $\mathrm{tr}(M(\mathbf{z})^2)$ equals the invariant $\mathrm{tr}(\beta^T M(\mathbf{z})\beta M(\mathbf{z}))$]. Then maximize the Schweinler–Wigner quartic form $m(\mathbf{z})=\sum_k (M(\mathbf{z})_{kk})^2$ within this submanifold of symplectic bases. This will lead, not just to a basis in which $M(\mathbf{z})$ is diagonal, but to one where $M(\mathbf{z})$ has the Williamson canonical form $M(\mathbf{z})=\mathrm{diag}(\kappa_1,...,\kappa_n;\kappa_1,...,\kappa_n)$. We have thus established the following generalization of the Schweinler–Wigner extremum principle to the symplectic case.

Theorem 7: The distinguished symplectic basis which extremizes the ''Schweinler–Wigner'' quartic form $m(\mathbf{z})$ over the submanifold of symplectic bases which minimize the quartic form $\mathrm{tr}(M(\mathbf{z})^2)$ is the same as the symplectic basis in which the positive definite matrix $M(\mathbf{z})$ assumes the Williamson canonical diagonal.

Note that once $M(\mathbf{z})^{\mathrm{odd}}=0$ is reached, as implied by $\mathrm{tr}(M(\mathbf{z})^{\mathrm{odd}})^2=0$, $M(\mathbf{z})$ has the special even form

$$\begin{pmatrix} A & C \\ -C & A \end{pmatrix}, \quad A^T=A, \quad C^T=-C, \tag{44}$$

so that $A+iC$ is Hermitian. The subgroup of symplectic transformations which do not mix $M(\mathbf{z})^{\mathrm{even}}$ with $M(\mathbf{z})^{\mathrm{odd}}$, and hence maintain the property $M(\mathbf{z})^{\mathrm{odd}}=0$ have the special form

$$S=\begin{pmatrix} X & Y \\ -Y & X \end{pmatrix}, \quad X+iY\in \mathrm{U}(n). \tag{45}$$

This subgroup, isomorphic to the unitary group $\mathrm{U}(n)$, is the maximal compact subgroup[10] of $\mathrm{Sp}(2n,\mathcal{R})$. Thus, diagonalizing $M(\mathbf{z})$ using symplectic change of basis, after it has reached the even form, is the same as diagonalizing an n-dimensional Hermitian matrix using unitary transformations.

VII. CONCLUDING REMARKS

To conclude, we have shown that an $N\times N$ real symmetric [complex Hermitian] positive definite matrix is congruent to a diagonal form modulo a pseudo-orthogonal [pseudo-unitary] matrix belonging to $\mathrm{SO}(m,n)[\mathrm{SU}(m,n)]$, for any choice of partition $N=m+n$. The method of proof of this result is adapted to provide a simple proof of Williamson's theorem. An important consequence of these theorems is that while a real-symmetric [complex-Hermitian] positive definite matrix has a unique diagonal form under conjugation, it has several different canonical diagonal forms under congruence. The theorems developed here are used to formulate an extremum principle a lá Schweinler and Wigner for constructing pseudo-orthonormal[pseudo-unitary] and symplectic bases from a given set of linearly independent vectors. Conversely, the extremum principle thus formulated can be used for finding the congruence transformation which brings about the desired diagonalization.

It is interesting that pseudo-orthonormal basis and symplectic basis could be constructed by extremizing *precisely the same Schweinler–Wigner quartic form* $m(\mathbf{z})=\sum_k (M(\mathbf{z})_{kk})^2$ that was originally used to construct orthonormal basis in the unitary case. However, it must be borne in mind that the similarity in the structure of the quartic form to be extremized in the three cases considered is only at a formal level. In reality, the three quartic forms are very different objects, for they are functions over topologically very different manifolds; $\mathbf{z}$ runs over the group manifold $\mathrm{U}(N)$ of orthogonal frames in the original Schweinler–Wigner case, the group manifold $\mathrm{SO}(m,n)$

of pseudo-orthogonal frames in the Lorentz case, and over the group manifold $Sp(2n,\mathcal{R})$ in the symplectic case. This has the consequence that, unlike the orthogonal case, this quartic form is unbounded in the noncompact $SO(m,n)[SU(m,n)]$ and $Sp(2n,\mathcal{R})$ cases. Insight into the structure of these groups was used to achieve constrained extremization within a natural maximal compact submanifold.

[1] See, for instance, F. C. Gantmacher, *The Theory of Matrices* (Chelsea, New York, 1960), Vol. 1.
[2] J. Williamson, Am. J. Math. **58**, 141 (1936); **59**, 599 (1936); **61**, 897 (1936). Williamson's results are more general than the theorem quoted, and obtain all the different canonical forms a real symmetric (not necessarily positive definite) matrix can take under congruence by the real symplectic group. The results of Williamson are summarized in a manner that should appeal to physicists, in V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978), Appendix 6.
[3] J. Moser, Commun. Pure Appl. Math. **11**, 81 (1958); A. Weinstein, Bull. Am. Math. Soc. **75**, 814 (1971); N. Burgoyne and R. Cushman, Celest. Mech. **8**, 435 (1974); J. Laub and K. Meyer, *ibid.* **9**, 213 (1974).
[4] A. J. Dragt, F. Neri, and G. Rangarajan, Phys. Rev. A **45**, 2572 (1992); E. C. G. Sudarshan, C. B. Chiu, and G. Bhamathi, *ibid.* **52**, 43 (1995).
[5] R. Simon, E. C. G. Sudarshan, and N. Mukunda, Phys. Rev. A **36**, 3868 (1987); R. Simon, N. Mukunda, and B. Dutta, *ibid.* **49**, 1567 (1994); Arvind, B. Dutta, N. Mukunda, and R. Simon, Pramana, J. Phys. **45**, 471 (1995); Arvind, B. Dutta, N. Mukunda, and R. Simon, Phys. Rev. A **52**, 1609 (1995).
[6] H. C. Schweinler and E. P. Wigner, J. Math. Phys. **11**, 1693 (1970).
[7] See, for instance, C. K. Chui, *Wavelet Analysis and its Applications* (Academic, San Diego, 1992).
[8] Just as the diagonal form is the canonical form for real symmetric matrices under rotation, $i\sigma_2 \otimes K$, with K diagonal, is the canonical form for a real antisymmetric matrix under rotation. Further K can be chosen to be non-negative, in general, and positive definite when the antisymmetric matrix is nonsingular.
[9] S. Chaturvedi, A. K. Kapoor, and V. Srinivasan, J. Phys. A **31**, L367 (1998).
[10] R. Simon, N. Mukunda, and B. Dutta, Phys. Rev. A **49**, 1567 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 7 JULY 1999

Invariant Painlevé analysis and coherent structures of two families of reaction-diffusion equations

Ugur Tanriver and S. Roy Choudhury[a)]
Department of Mathematics, University of Central Florida, Orlando, Florida 32816-1364

(Received 25 June 1998; accepted for publication 12 March 1999)

Exact closed-form coherent structures (pulses/fronts/domain walls) having the form of complicated traveling waves are constructed for two families of reaction–diffusion equations by the use of invariant Painlevé analysis. These analytical solutions, which are derived directly from the underlying PDE's, are investigated in the light of restrictions imposed by the ODE that any traveling wave reduction of the corresponding PDE must satisfy. In particular, it is shown that the coherent structures (a) asymptotically satisfy the ODE governing traveling wave reductions, and (b) are accessible to the PDE from compact support initial conditions. The solutions are compared with each other, and with previously known solutions of the equations. © *1999 American Institute of Physics.* [S0022-2488(99)01907-6]

I. INTRODUCTION

There has been considerable interest in coherent structure solutions of nonintegrable nonlinear partial differential equations (NLPDEs)[1–10] since these provide an organizing structure to the space of solutions. In a very rough sense, this is somewhat analogous to the way in which families of soliton solution act as basic building blocks for the solution space of integrable equations. Recent work, primarily in the context of generalized Ginzburg–Landau amplitude equations in pattern-forming systems, has included the existence of pulse (solitary wave), front (shock) and domain wall coherent structures using center manifold techniques,[11,12] as well as investigations of periodic and quasi-periodic solutions.[13–17] Another, more physics-oriented, approach was developed by van Saarloos[18,19] to investigate linear and nonlinear marginal stability of fronts. This approach has been comprehensively reviewed by van Saarloos and Hohenberg[20] in the context of generalized Ginzburg–Landau equations. Using the idea that spatio-temporal coherent structure solutions of NLPDEs, whether periodic, quasi-periodic, or chaotic, must obey the underlying singularity structure, Conte and co-workers[21,22] have used ideas related to the Painlevé test for integrability[23,24] and its modifications[25] to derive families of solutions of the complex cubic and quintic Ginzburg–Landau equation. Also, using phase-plane techniques on the ordinary differential equation (ODE) which must be satisfied by any traveling wave solution to the real Ginzburg–Landau equation, Powell *et al.*[26] have rederived and significantly elucidated several of van Saarloos' results[18,19] in a completely different manner. In addition, they use simple analytic solutions of the PDE obtained using truncated Painlevé expansions,[27] together with ideas from phase-plane analysis, as well as absolute versus convective instability of waves.[28] As a result, they show that front/pulse solution of the PDE must satisfy the traveling wave reduced ODE asymptotically. They also derive conditions for the accessibility of the solutions from compact support initial conditions.

In this paper, we consider coherent structures of the reaction-diffusion equation

$$u_t = u_{xx} + \frac{u}{b}(b+u)(1-u), \tag{1}$$

which has been considered in Refs. 4, 18 and 19. Note that other work on coherent structures of

[a)]Author for correspondence. Electronic mail: choudhur@longwood.cs.ucf.edu various reaction-diffusion equations is summarized in those papers, as well as in Section 1 of Ref. 26. For the purposes of comparison, we shall also consider coherent structures of the two families of reaction-diffusion equations

$$u_t=\beta u^2(1-u)+Du_{xx} \tag{2}$$

and

$$u_t=\beta u(1-u)+Du_{xx}\,. \tag{3}$$

Of these, (3) is the famous Fisher–Kolmogorov equation,[29,30] while (2) has second- and third-order nonlinearities, which is also true of (1). The primary difference between (1) and (2)/(3) is that the parameter b in the former adjusts the relative strength of the second- and third-order nonlinearities, while these strengths are fixed in (2) and (3).

To date, the approaches to the treatment of coherent structures may broadly be classified into three groups. First, there is the qualitative phase-plane/center manifold analysis of the traveling wave reduced ODEs to prove the existence of coherent structures. The second approach consists of actual construction of coherent structures via numerical simulation of the traveling wave reduced ODEs. The third approach comprises containment arguments wherein, starting from the correct boundary condition at one end of the interval, one shows that at the other end the solution asymptotes to a constant value. It thus corresponds to a coherent structure, rather than shooting off to infinity. Such containment arguments may often involve delicate analysis. The coherent structures derived in this paper are, in a sense, an attempt to connect the first two approaches by providing quantitative analytical expressions for nontrivial coherent structures. Clearly, these coherent structures are also of relevance in modeling the physics of the problems under consideration, although that is not the purpose of this paper. In fact, the next natural question to consider is their actual use in modeling applications. Some discussion regarding this follows the derivation of the coherent structures in Sec. IV.

In Sec. II we use invariant Painlevé expansions truncated at different orders to obtain nontrivial families of analytic solutions of the reaction-diffusion equations. Two points are worth noting in this context. First, the invariant Painlevé analysis[24] builds in invariance to the Möbius or homographic group ''*a priori*.'' In turn, this leads to simpler compatibility equations for the coefficients (the so-called Painlevé–Bäcklund equations) yielding more general solutions than obtained for (1) from the use of truncated noninvariant Painlevé expansions.[31] Second, although truncated invariant Painlevé expansions have been used fairly widely in recent years to derive analytic solutions (see Refs. 32–34, for instance), the Painlevé–Bäcklund equations which result from (1) and which are solved to obtain analytic solutions are quite complicated. Having obtained analytic solutions of the PDE in Sec. II, we next consider the properties of the ordinary differential equation governing traveling wave solutions in Sec. III. In Sec. IV, we discuss the compatibility of the PDE solutions and solutions derived earlier with those of the ODE, as well as accessibility from initial conditions. We also give numerical examples of various coherent structure solutions.

II. INVARIANT TRUNCATION PROCEDURE AND SPECIAL SOLUTIONS

A. Truncation procedure

For a NLPDE that is algebraic in u and its derivatives

$$E(u,x,t)=0 \tag{4a}$$

around a movable singular manifold

$$\Phi-\Phi_0=0, \tag{4b}$$

one looks, in the invariant Painlevé formulation,[24] for a solution as an expansion of the form

$$u=\chi^{-\alpha}\sum_{j=0}^{\infty} u_j\chi^j, \tag{5}$$

where the coefficients u_j are invariant under a group of homographic (Möbius) transformations on Φ. The expansion variable χ, which must vanish as $(\Phi-\Phi_0)$, is chosen to be

$$\chi\equiv\frac{\psi}{\psi_x}=\left(\frac{\Phi_x}{\Phi-\Phi_0}-\frac{\Phi_{xx}}{2\Phi_x}\right)^{-1}, \quad \psi=(\Phi-\Phi_0)\Phi_x^{-1/2}.$$

The variable χ satisfies the Riccati equations

$$\chi_x=1+\tfrac{1}{2}S\chi^2, \tag{6a}$$

$$\chi_t=-C+C_x\chi-\tfrac{1}{2}(CS+C_{xx})\chi^2, \tag{6b}$$

and the variable ψ satisfies the linear equations

$$\psi_{xx}=-\tfrac{1}{2}S\psi, \tag{7a}$$

$$\psi_t=\tfrac{1}{2}C_x\psi-C\psi_x. \tag{7b}$$

Note that the systems of equations (6) and (7) are equivalent to each other. In (6) and (7), the quantities S (Schwarzian derivative) and C (the ''dimension of velocity'' or celerity) are defined by

$$S=\frac{\Phi_{xxx}}{\Phi_x}-\frac{3}{2}\left(\frac{\Phi_{xx}}{\Phi_x}\right)^2, \tag{8a}$$

$$C=-\frac{\Phi_t}{\Phi_x}, \tag{8b}$$

and are invariant under the group of homographic (Möbius) or fractional linear transformations[18]

$$\Phi\to\frac{a\Phi+b}{c\Phi+d}, \quad ad-bc=1. \tag{9}$$

These homographic invariants are linked by the cross-derivative condition ($\Phi_{xxxt}=\Phi_{txxx}$)

$$S_t+C_{xxx}+2C_xS+CS_x=0. \tag{10}$$

B. Solutions via invariant Painlevé analysis

We apply the above formalism to (1). The leading-order dominant balance yields $\alpha=2$. Using (5), truncated at the constant term

$$u=u_0\chi^{-2}+u_1\chi^{-1}+u_2, \tag{11}$$

in (1), and eliminating the derivatives of χ using (6) yields a set of coupled nonlinear partial differential equations (the Painlevé–Bäcklund equations) order by order in powers of χ. These are given in Appendix A. The first four equations yield

$$u_0=0,$$

$$u_1=0 \text{ or } \pm\sqrt{2b}. \tag{12}$$

Inspection of (A5)–(A7) in Appendix A shows the need for further assumptions to allow their solutions. Making the further assumption[31–34] that C is a constant yields

$$\text{(a)}\quad 2+3b-3b^2-2b^3+(3\sqrt{2b}+3\sqrt{2}b^{5/2}+3\sqrt{2}b^{3/2})C-2\sqrt{2}b^{3/2}C^3=0,$$

$$\text{or}\ \ C=C_1\equiv\frac{-\sqrt{2}-2\sqrt{2}b}{2\sqrt{b}},\quad C=C_2\equiv\frac{-\sqrt{2}+\sqrt{2}b}{2\sqrt{b}}\tag{13a}$$

$$\text{or}\ \ C=C_3\equiv\frac{2\sqrt{2}+\sqrt{2}b}{2\sqrt{b}}\quad\text{for}\ \ u_1=\sqrt{2b},$$

with

$$S=\frac{C^2}{6}-\frac{b}{3}-\frac{1}{3b}-\frac{1}{3}=2Q^2,\tag{13b}$$

or

$$\text{(b)}\quad u_2=0,\ 1,\quad \text{or}\quad -b\quad \text{for}\ \ u_1=0\tag{14}$$

[this yields only trivial constant solutions using (11)]. The cross-derivative condition (10) is now satisfied identically. The Schrödinger equation (6a) yields

$$\psi(x,t)=A(t)\cos Qx+B(t)\sin Qx$$

and hence, using (6b),

$$\chi\equiv\frac{\psi}{\psi_x}=\frac{c_1\cos(Q\xi)-c_2\sin(Q\xi)}{-Q[c_2\cos(Q\xi)+c_1\sin(Q\xi)]}\tag{15a}$$

with

$$\xi=x-Ct.\tag{15b}$$

Hence, using (12) and (13) in (11), traveling wave special solutions of (1) (for $u_1=\sqrt{2b}$) are

$$u^{(1)}=\pm\frac{\sqrt{2b}}{\chi}+\frac{1}{6}(2-2b-C\sqrt{2b}),\tag{16}$$

where χ is given by (15a), and C has one of the values in (13a). A solution may be derived analogously for $u_1=-\sqrt{2b}$, and a similar, less interesting, solution may be obtained using (11), (12), and (14) (for $u_2=0$, 1 or $-b$). Note that C and Q are connected via (13b), and hence Q is (implicitly) a function of b.

A similar process applied to (2) yields a solution

$$u_1^{(2)}=\pm\frac{\sqrt{2D/\beta}}{\chi}+\left(\frac{2D\mp C\sqrt{2D/\beta}}{6D}\right),\tag{17a}$$

where χ is given by (15), with

$$S=\frac{C^2-2D\beta}{6D^2}\equiv 2Q^2,\tag{17b}$$

and C is a solution of the cubic

$$2\beta\left(\frac{1}{3}\mp\frac{C}{3\sqrt{2\beta D}}\right)^3-2\beta\left(\frac{1}{3}\mp\frac{C}{3\sqrt{2\beta D}}\right)^2\pm\sqrt{\frac{2D}{\beta}}C\left(\frac{C^2-2D\beta}{6D^2}\right)=0. \tag{17c}$$

For instance, with $D=1$, $\beta=2$, we obtain $C=-1$, or $C=2$.

By contrast, the same procedure applied to (3) yields simpler, relatively trivial, traveling wave solutions. For completeness, we include a solution of (2) obtained earlier[31] using noninvariant Painlevé analysis:

$$u_{\text{II}}^{(2)}=\gamma\sqrt{\frac{2D}{\beta}}\left[\frac{\alpha+\sqrt{2D/\beta}e^{\sqrt{\beta/2D}\,\xi_1}}{\alpha\gamma\xi_1+(2D\gamma/\beta)e^{\sqrt{\beta/2D}\,\xi_1}+\delta}\right], \tag{18a}$$

where

$$\xi_1\equiv x+\sqrt{\frac{\beta D}{2}}t, \tag{18b}$$

and α, γ, and δ are constants. Note that the discussion of this solution later in this paper is new, as is the framework of that discussion.

C. Preliminary discussion

We shall consider the behavior of the solutions (16)–(18) further in the next two sections. However, we first need to consider some results for the ODEs, derived from traveling wave reductions, of the underlying PDEs (1) and (2). This will be done in Sec. III. At this point, we make some preliminary observations regarding the solutions (16)–(18) which will be needed in Sec. III.

We note that we have aperiodic hyperbolic functions in (16) (corresponding to a coherent structure) for Q imaginary or, from (13a) and (13b),

$$\begin{gathered}0<b<\sqrt{\tfrac{3}{2}}\quad\text{or}\quad b<-\sqrt{\tfrac{3}{2}}\quad\text{for}\;\; C=C_1,\\ b>0\quad\text{or}\quad -3-\sqrt{6}<b<-3+\sqrt{6}\quad\text{for}\;\; C=C_2,\\ b>0\quad\text{for}\;\; C=C_3.\end{gathered} \tag{19}$$

For these cases, $\lim_{\xi\to\pm\infty}\chi=\pm 1/|Q|$ from (15), so that (16) yields

$$\lim_{\xi\to\pm\infty}u^{(1)}=\pm\sqrt{2b}|Q|+\tfrac{1}{6}(2-2b-C\sqrt{2b}). \tag{20}$$

For instance, for the first value of $C\equiv C_1$ given by (13a), this becomes

$$\lim_{\xi\to+\infty}u^{(1)}=1,\qquad \lim_{\xi\to-\infty}u^{(1)}=0. \tag{21}$$

Note that the solution $u^{(1)}$ tends to different values as $\xi\to\pm\infty$, and these values are independent of the constants c_1 and c_2 in (15a). Thus, the solutions (16) represent front solutions of (1).

Similarly, considering the solution $u_1^{(2)}$ of (2), if the roots of (17c) for C are C_4, C_5, and C_6 (say), then Q in (17b) may be imaginary for some ranges of β and D. For these cases, $Q\equiv i|Q|$, and $\lim_{\xi\to\pm\infty}\chi=\pm 1/|Q|$, so that (17) yields

$$\lim_{\xi\to\pm\infty}u_1^{(2)}=\pm\sqrt{\frac{2D}{\beta}}|Q|+\frac{1}{3}\pm\frac{C}{6D}\sqrt{\frac{2D}{\beta}}. \tag{22}$$

Note that $u_1^{(2)}$ tends to different values as $\xi \to \pm\infty$ and thus corresponds to a front solution of (2). For instance, with $D=1$, $\beta=2$, $C=\pm 1$ and the upper sign,

$$\lim_{\xi \to = -\infty} u_1^{(2)} = 0, \quad \lim_{\xi \to \infty} u_1^{(2)} = 1.$$

It may be shown in an analogous manner that the solutions $u_{\mathrm{II}}^{(2)}$ in (18) are also front solutions of (2). In the next section, we shall consider some properties of the ODE obtained by performing traveling wave reductions on the reaction-diffusion equations (1) and (2) before returning to discuss the solutions in this section further. We shall refer to (20)–(22) further during that discussion.

III. ANALYSIS OF TRAVELING WAVE REDUCED ODE

We shall look for traveling wave reductions of the PDE's (1) and (2). We present the results for (2) since the algebra is somewhat easier. The treatment for (1) is analogous.

Looking for traveling wave solutions of (2) of the form

$$u(x,t) = u(z) \equiv u(x - Ct) \tag{23a}$$

yields the ordinary differential equation

$$-Cu' = \beta u^2(1-u) + Du'', \tag{23b}$$

where the prime denotes d/dz. Note that we use z as an explicit traveling wave variable ''*a priori*,'' as distinct from the analogous variable ξ which emerged ''*a posteriori*'' in Sec. II from the Painlevé analysis.

Treating (23) as a dynamical system in the $(u,u') \equiv (u,v)$ plane in the standard way, we find the fixed (critical) points in the (u,v) plane:

$$(u_0,0) \equiv (0,0), \tag{24a}$$

$$(u_+,0) \equiv (1,0), \tag{24b}$$

whose linear stability is governed by the eigenvalues (which are also the spatial wave numbers in z space)

$$\lambda_0^{1,2} = 0,\ \frac{-C}{D}, \tag{25a}$$

$$\lambda_+^{1,2} = \frac{-C/D \pm \sqrt{C^2/D^2 + 4\beta/D}}{2}. \tag{25b}$$

Since β and D are positive, the fixed point $(u_+,0)$ is thus a saddle-point, while $(u_0,0)$ is a nonhyperbolic fixed point. The system (23) may thus have a heteroclinic orbit connecting $(u_0,0)$ and $(u_+,0)$. In the context of the underlying PDE (2), this corresponds to a front solution connecting $u_0(u_+)$ and $u_+(u_0)$ as z goes from $-\infty$ to $+\infty$. From (22), we see that the solutions (17) of the PDE (directly obtained from the PDE) are indeed heteroclinic orbits of (23). For instance, with $D=1$, $\beta=1$, $C=\pm 1$, and the lower sign in (22), $\lim_{\xi \to +\infty} u_1^{(2)} = 0$ and $\lim_{\xi \to -\infty} u_1^{(2)} = 1$, so that this front solution $u_{\mathrm{II}}^{(2)}$ is a heteroclinic orbit of (23) joining $(u_0,0)$ to $(u_+,0)$ as $z \to \pm\infty$. Note that for some PDEs (such as the long-wave equations[35]) this may not happen automatically—the integrated version of the traveling wave reduced ODE may contain unknown constants of integration which must be chosen to ensure this. An analogous treatment of (1) yields the traveling wave reduced ODE

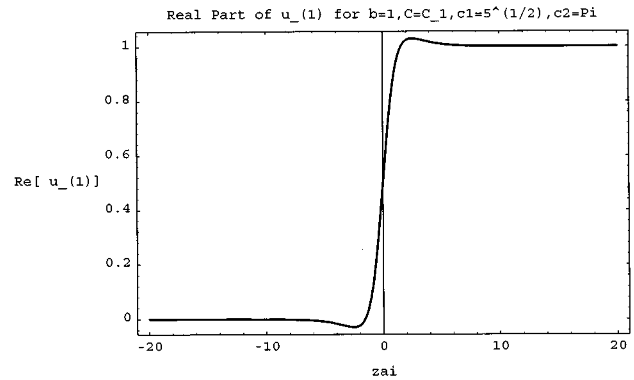

FIG. 1. Real part of $u\_(1)$ for $b=1$, $C=C\_1$, $c1=5\hat{}(1/2)$, $c2=Pi$.

$$-Cu'=u''+\frac{u}{b}(b+u)(1-u), \tag{26}$$

which, treated as a dynamical system in $(u,u')\equiv(u,v)$, has fixed points $(u_0,0)=(0,0)$, $(u_+,0)=(1,0)$ and $(u_-,0)=(-b,0)$. As for (2), it may be shown using (20) and (21) that the front solutions $u^{(1)}$ of (1) correspond to heteroclinic orbits of (26) joining two of the above fixed points. For instance, for $C\equiv C_1$, $\lim_{\xi\to-\infty}u^{(1)}=0$, $\lim_{\xi\to\infty}u^{(1)}=1$, so that the front solution $u^{(1)}$ of (1) is a heteroclinic orbit of (26) joining $(u_0,0)$ to $(u_+,0)$ as $z\to\pm\infty$. Once again, this need not happen automatically, as it does not for the long-wave equations,[35] for instance.

As extensively investigated and stressed by Powell *et al.*[26] the front solutions represented by the heteroclinic orbits of the traveling wave reduced ODEs (23) and (26) need not correspond to fronts obtained directly from the PDE (1). We shall now consider this further.

IV. DISCUSSION

In this section, we consider further features of the solutions $u^{(1)}$ and $u_1^{(2)}$ [in (16) and (17)] of (1) and (2) obtained by use of invariant Painlevé expansions. Powell *et al.*[26] have, among numerous other things, made the points that coherent structure solutions such as (16) and (17), which are directly obtained from a PDE, (a) must asymptotically satisfy the ODE governing traveling wave reductions, and (b) be accessible to the PDE from compact support initial conditions. Considering the traveling wave reduced ODE (23) obtained from (2), we have $z\equiv x-Ct\to-\infty$ as $t\to\infty$ (for $C>0$), and so u tends to the saddle-point $(u_+,0)$ in (24b) along its unstable manifold. From (25b), the eigenvalue along this direction is

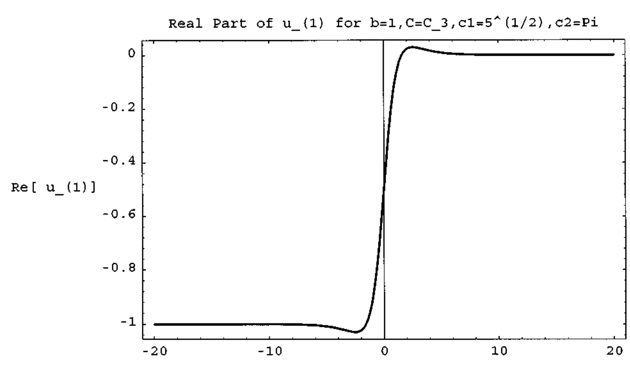

FIG. 2. Real part of $u\_(1)$ for $b=1$, $C=C\_3$, $c1=5\hat{}(1/2)$, $c2=Pi$.

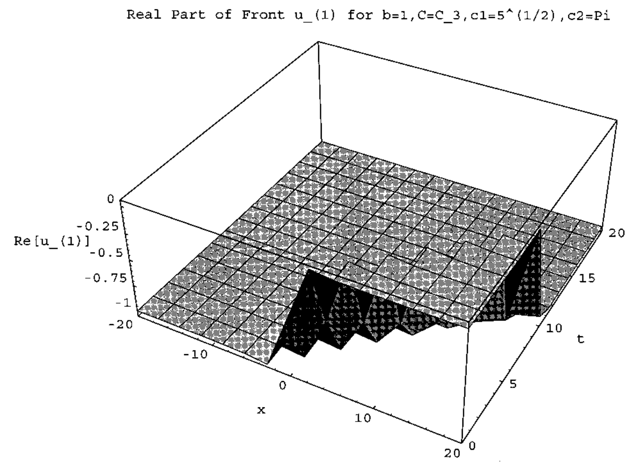

FIG. 3. Real part of Front $u\_(1)$ for $b=1$, $C=C\_3$, $c1=5\hat{}(1/2)$, $c2=Pi$.

$$\lambda_{\mathrm{ODE}}=\lambda_{+}^{1}=\frac{-C/D+\sqrt{C^2/D^2+4\beta/D}}{2}. \tag{27}$$

By the Unstable Manifold Theorem,[36] (27) gives the time asymptotic spatial wave number of the front solutions (2) [along the global unstable manifold of $(u_+,0)$] satisfying the ODE (23) and with solution values u_0 and u_+ for $t\rightarrow\mp\infty$. Inspection of the solutions $u_1^{(2)}$ in (17) reveals that the wave number ($\lambda_{\mathrm{PDE}}\equiv|Q|$) of these solutions obtained directly from the PDE are exactly the same as λ_{ODE} [this may be seen from (17b) and (27), using the fact that C satisfies (17c)]. In Sec. III, we verified that the values of (a) u_{PDE} for $\xi\rightarrow\pm\infty$, and (b) u_{ODE} for $z\rightarrow\pm\infty$ are matched; here we see that the resulting time asymptotic wavenumbers in the ODE and PDE solutions are also the same. Thus, as conjectured in Powell *et al.*, the solutions obtained via Painlevé analysis are indeed the so-called nonlinear solutions;[19] note that Powell *et al.* equivalently think of C as a function of λ, instead of λ as a function of C as done here. As pointed out by both Powell *et al.*[26] and Marcq *et al.*[22] for the GL equation, this is because the Painlevé analysis builds in ''*a priori*'' the singularity structure which must be satisfied by any coherent structure solution of the PDE.

Although the front solutions (16) and (17) satisfy the traveling wave reduced ODEs (26) and (23), we must also check the accessibility of these solutions to the PDEs (1) and (2) from compact support initial conditions as stressed in Ref. 26. Following the treatment of absolute versus convective instability in Ref. 28, we find the temporal growth rate at any fixed x spatial position

$$\sigma=|-C\lambda_{\mathrm{PDE}}|\equiv\left|-C|Q|=-C\left[\frac{2D\beta-C^2}{12D^2}\right]^{1/2}\right|. \tag{28}$$

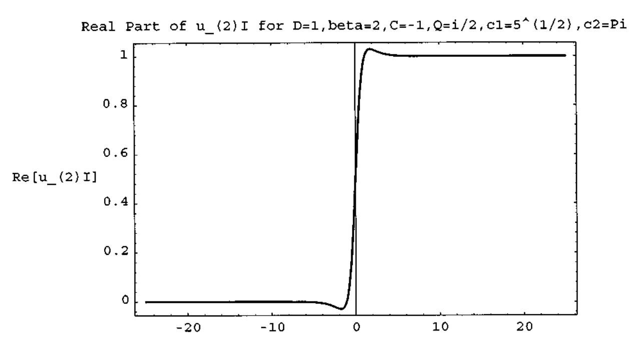

FIG. 4. Real part of $u\_(2)I$ for $D=1$, beta$=2$, $C=-1$, $Q=i/2$, $c1=5\hat{}(1/2)$, $c2=Pi$.

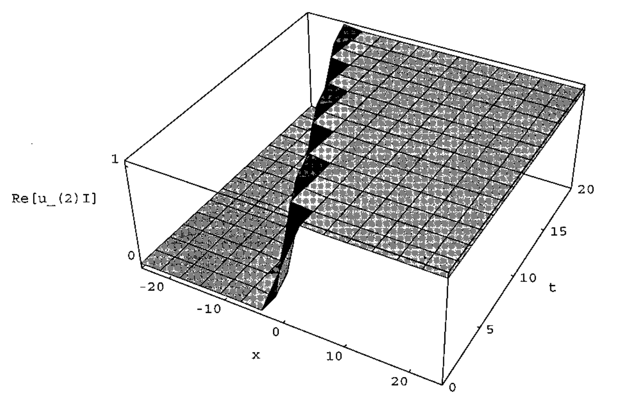

FIG. 5. Real part of $u\_(2)I$ for $D=1$, beta=2, $C=-1$, $Q=i/2$, $c1=5\hat{}(1/2)$, $c2=Pi$.

Thus, we expect that the front which emerges asymptotically in time from compact support initial conditions corresponds to the root of (17c) for which we have the maximum temporal growth rate. For instance, with $D=1$, $\beta=2$, we obtain $C=-1$ or 2 from (17c). From (17b), $|Q|=\frac{1}{2}$ for $C=-1$, and $|Q|=0$ for $C=2$. Thus the maximum σ occurs for $C=-1$.

Analogous results apply to the wave numbers and temporal growth rates of the solution (16) obtained directly from the PDE (1) and its traveling wave reduced ODE (26). The algebra is harder, but is tractable using a computer algebra system.

One should note that some of the results obtained by phase-plane analysis of the traveling wave reduced ODE are equivalent to those obtained by van Saarloos' linear and nonlinear marginal stability analysis and steepest envelope technique.[29] For completeness, this treatment is summarized in Appendix B for Eq. (2). An analogous treatment holds for (1).

Finally, let us consider plots of the solutions of (1) and (2) given by (16) and (17), and compare them further to predictions from the traveling wave reduced ODEs. For both (16) and (17), we choose representative parameter values corresponding to the front solutions discussed earlier. Note that we may choose the constants c_1 and c_2 to make (16) and (17) correspond to physically relevant real solutions. We pick arbitrary c_1 and c_2 values instead, and plot the real parts of the solutions.

Figures 1 and 2 show the real part of $u^{(1)}$ [given by (16)] for $b=1$, $c_1=\sqrt{5}$, $c_2=\pi$, and (a) $C=C_1$ for Fig. 1 and (b) $C=C_3$ for Fig. 2. Note that the primary difference between these plots is that the former, with $C=C_1$, corresponds to a front connecting the states 0 and 1 or u_0 and u_+ as $\xi\rightarrow\pm\infty$, while the latter front with $C=C_3$ connects the state -1 and 0 [note that the third fixed point $(u_-,0)=(-b,0)$ of (26) is $(-1,0)$ for $b=1$]. Figure 3 shows the same front as Fig. 2, but in (x,t) coordinates. Note the rightward propagation of the front (towards larger x) as t increases due to the phase speed $C=C_3=3/\sqrt{2}$ being positive.

Figures 4 and 5 show the solution (17) of the PDE (2) for $c_1=\sqrt{5}$, $c_2=\pi$, $D=1$, $\beta=2$, $Q=i/2$, and $C=-1$ [note that $Q=\pm i/2$, and $C=-1$ or 2 by (17b) and (17c)—we pick $Q=i/2$ and $C=-1$]. The solution corresponds to a front joining the states 0 and 1 [or the fixed points $(u_0,0)$ and $(u_+,0)$ of Eq. (23) given in (24)]. In Fig. 5, note the leftward propagation of the front due to the negative phase speed $C=-1$.

In conclusion, we have derived two nontrivial families of analytical solutions of (1) and (2), which may sometimes be coherent structures, and analyzed several of their properties. As mentioned in Sec. I, these analytical solutions act as a sort of bridge between two of the common approaches to the analysis of coherent structures. These two approaches are, first, proofs of the existence of coherent structure solutions of the traveling wave reduced ODEs, and, second, construction of coherent structures by numerical simulations of these ODEs. The analytical solutions may also be of relevance in modeling the physics of the problem under consideration. Although it is not the purpose of this article to consider detailed modeling issues, some of the approaches which may be relevant to modeling of reaction-diffusion equations include those in Ref. 29, Chap. 3 of Ref. 37, Chap. 6 of Ref. 38, as well as numerous research papers. Related modeling issues for other nonlinear PDEs are discussed, for instance, in Chap. 10 of Ref. 39 (this also discusses nonintegrable equations, not just integrable equations as the chapter title might seem to indicate), and the recent review article by Balmforth.[40]

ACKNOWLEDGMENT

The authors acknowledge the helpful comments of the referee regarding stylistic matters, inclusion of Appendix B, as well as a discussion of the relevance of the results in this paper. These improved the clarity and completeness of the paper greatly.

APPENDIX A

The equations obtained at different powers of χ are

$$O(\chi^{-6}): \quad u_0=0, \tag{A1}$$

$$O(\chi^{-5}): \quad u_0^2 u_1=0, \tag{A2}$$

$$O(\chi^{-4}): \quad 6u_0-u_0^2+\frac{u_0^2}{b}-\frac{3u_0u_1^2}{b}-\frac{3u_0^2u_2}{b}=0, \tag{A3}$$

$$O(\chi^{-3}): \quad -2Cu_0+2u_1-2u_0u_1+\frac{2u_0u_1}{b}-\frac{u_1^3}{b}-\frac{6u_0u_1u_2}{b}-4u_{0x}=0, \tag{A4}$$

$$O(\chi^{-2}): \quad u_0+4Su_0-Cu_1-u_1^2+\frac{u_1^2}{b}-2u_0u_2+\frac{2u_0u_2}{b}-\frac{3u_1^2u_2}{b}-\frac{3u_0u_2^2}{b}-u_{0t}+2u_0C_x-2u_{1x} + u_{0xx}=0, \tag{A5}$$

$$O(\chi^{-1}): \quad -CSu_0+u_1+Su_1-2u_1u_2+\frac{2u_1u_2}{b}-\frac{3u_1u_2^2}{b}-u_{1t}+u_1C_x-u_0S_x-2Su_{0x}-u_0C_{xx} + u_{1xx}=0, \tag{A6}$$

$$O(\chi^0): \quad \frac{1}{2}S^2u_0-\frac{1}{2}CSu_1+u_2-u_2^2+\frac{u_2^2}{b}-\frac{u_2^3}{b}-u_{2t}-\frac{1}{2}u_1S_x-Su_{1x}-\frac{1}{2}u_1C_{xx}+u_{2xx}=0. \tag{A7}$$

APPENDIX B: VAN SAARLOOS' TECHNIQUE

The tail of the coherent structure must obey the linear RD equation (2) as $x\rightarrow\infty$ (since $u\rightarrow 0$ or 1) so that

$$u_t=Du_{xx}\,. \tag{B1}$$

Consider the behavior of a linear mode

$$u=\exp[i(\omega-i\sigma)t+i(k-i\lambda)z],$$

where $z=x-Ct$ and C is the front speed. For this mode to be part of a persistent front, it must have zero temporal growth $\sigma=0$ in a frame moving at speed C. This gives the dispersion relation

$$i\omega-iC(k-i\lambda)=-D(k-i\lambda)^2. \tag{B2}$$

For a nonoscillatory mode with $\omega = 0$, separating the real and imaginary parts of (B2) yields either (a) $k = 0$, $\lambda = -C/D$, or (b) $\lambda = -C/(2D)$, $k^2 = -(C/2D)^2$. Case (b) with imaginary k implies no coherent structures, so we have

$$\lambda = -C/D,$$

which is equal to λ_{PDE}. Various other features discussed in the text may also be derived by this approach (see Refs. 18, 19, and 26).

[1] A. Kolmogorof, I. Petrovsky, and N. Piskunov, Bull. Univ. Moskou. Ser. Internat., Sec. A **1**, 1 (1937).
[2] W. E. Fitzgibbon (editor), *Partial Differential Equations and Dynamical Systems* (Pitman, Boston, 1984).
[3] G. Dee and J. S. Langer, Phys. Rev. Lett. **50**, 383 (1983).
[4] E. Ben-Jacob, H. R. Brand, G. Dee, L. Kramer, and J. S. Langer, Physica D **14**, 348 (1985).
[5] G. Dee, J. Stat. Phys. **39**, 705 (1985).
[6] D. G. Aronson and H. F. Weinberger, in *Partial Differential Equations and Related Topics*, edited by J. A. Goldstein (Springer, Heidelberg, 1975).
[7] D. G. Aronson and H. F. Weinberger, Adv. Math. **30**, 33 (1978).
[8] P. Fife, in *Mathematical Aspects of Reacting and Diffusing Systems*, edited by S. Levin (Springer, New York, 1979).
[9] R. A. Fisher, Annals Eugenics **7**, 355 (1937).
[10] R. J. Field and M. Burger (editors) *Oscillations and Traveling Waves in Chemical Systems* (Wiley, New York, 1985).
[11] C. K. R. T. Jones, T. M. Kapitula, and J. A. Powell, Proc. R. Soc. Edinburgh A **116**, 193 (1990).
[12] J. Duan and P. Holmes, Proc. R. Soc. Edinburg **38**, 77 (1995).
[13] A. Doelman, Physica D **401**, 156 (1989).
[14] A. Doelman, J. Nonlinear Sci. **3**, 225 (1993).
[15] A. Doelman and W. Eckhaus, Physica D **53**, 249 (1991).
[16] W. Eckhaus and G. Iooss, Physica D **39**, 124 (1989).
[17] N. Bekki and B. Nozaki, Phys. Lett. A **110**, 133 (1985).
[18] W. van Saarloos, Phys. Rev. A **37**, 211 (1988).
[19] W. van Saarloos, Phys. Rev. A **39**, 6367 (1989).
[20] W. van Saarloos and P. C. Hohenberg, Physica D **56**, 303 (1992).
[21] R. Conte and M. Musette, Physica D **69**, 1 (1993).
[22] P. Marcq, H. Chate, and R. Conte, Physica D **73**, 305 (1994).
[23] M. Tabor, *Chaos and Integrability in Nonlinear Dynamics: An Introduction* (Wiley, New York, 1989).
[24] R. Conte, Phys. Lett. A **140**, 383 (1989).
[25] R. Conte, in *Introduction to Methods of Complex Analysis and Geometry for Classical Mechanics and Nonlinear Waves*, edited by D. Benest and C. Froeschle (Editions Fronteres, Gif-sur Yvette, 1994).
[26] J. A. Powell, A. C. Newell, and C. K. R. T. Jones, Phys. Rev. A **44**, 3636 (1991).
[27] F. Cariello and M. Tabor, Physica D **39**, 77 (1989).
[28] G. Schmidt, *Physics of High Temperature Plasmas*, 2nd ed. (Academic, New York, 1966).
[29] J. D. Murray, *Mathematical Biology* (Springer, Berlin, 1989).
[30] J. P. Pauwelussen and L. A. Peletier, J. Math. Biol. **11**, 207 (1981).
[31] B-Yu. Guo and Z. C. Xiong, J. Phys. A **24**, 645 (1991).
[32] N. Isidore, J. Phys. A **29**, 3679 (1996).
[33] A. Pickering, J. Phys. A **26**, 4395 (1993).
[34] N. Isidore and W. Malfliet, J. Phys. A **30**, 5151 (1997).
[35] S. R. Choudhury, Can. J. Phys. **70**, 595 (1992).
[36] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields* (Springer, New York, 1983).
[37] P. Grindrod, *Patterns and Waves* (Clarendon, Oxford, 1991).
[38] J. D. Logan, *An Introduction to Nonlinear Partial Differential Equations* (Wiley, New York, 1994).
[39] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. C. Morris, *Solitons and Nonlinear Wave Equations* (Academic, London, 1990).
[40] N. J. Balmforth, Annu. Rev. Fluid Mech. **27**, 335 (1995).

Frequency analysis based on general periodic functions

Yuchuan Wei[a)]
Department of Electrical Engineering, No. 202, Beijing University of Aeronautics and Astronautics, Beijing 100083, People's Republic of China and Department of Physics, Beijing University of Science and Technology, Beijing 100083, People's Republic of China

(Received 21 August 1998; accepted for publication 14 February 1999)

Following the sine–cosine function, the sawtooth wave, square wave, triangular wave, trapezoidal wave, and so on become new easily generated periodic functions in modern electronics. Similar to Fourier's idea, a natural question is whether a signal can be considered as a superposition of easily generated functions with different frequencies. Therefore it is necessary to generalize Fourier analysis based on sine–cosine functions into frequency analysis based on general periodic functions. In this paper, we introduce the frequency series and frequency transformation based on general periodic functions. We discuss when a frequency system is a complete system or an unconditional basis in $L^2[-\pi,\pi]$, and when a frequency transformation can be carried out in $L^2(-\infty,+\infty)$. For practical convenience almost all easily generated functions in electronics are considered carefully as examples. As a new and practical generalization of classical Fourier analysis, these results will become a theoretical foundation for the technique of easily generated function analysis in signal processing. [S0022-2488(99)02706-1]

I. INTRODUCTION

In 1807 the French mathematician J. B. J. Fourier asserted that any function with period 2π can be expressed as a trigonometric series, and that any nonperiodic function can be expressed as a trigonometric integral. This great idea has had an important influence upon science and technology. In electronics, Fourier analysis has been playing an important role, and a signal is often considered as a superposition of many sine and cosine functions with different frequencies. However, with the development of electronic technique, the sawtooth wave, square wave, triangular wave, trapezoidal wave, and so on become new easily generated periodic functions. For example, it is quite convenient to get a system of square waves with different frequencies from a common high-frequency pulse by counters. A practical question arises naturally: *Can a signal be considered as a superposition of other periodic functions (such as square waves) with different frequencies*?

In order to answer this practical question, we need to introduce the frequency analysis based on general periodic functions in mathematics, which is a natural generalization of Fourier analysis based on the sine and cosine functions. Considering the fact that the theory for Fourier series in $L^2[-\pi,\pi]$ and the theory for Fourier transformation in $L^2(-\infty,+\infty)$ are both simple and important, we shall develop a theory for general frequency series in $L^2[-\pi,\pi]$ and a theory for general transformation in $L^2(-\infty,+\infty)$. In the following, we shall discuss when a frequency system is a complete system or an unconditional basis in $L^2[-\pi,\pi]$, and when a frequency transformation can be carried out in $L^2(-\infty,+\infty)$. For practical convenience, almost all easily generated functions in electronics are considered carefully as examples. The conclusions in this paper will provide a theoretical foundation for the technique of easily generated function analysis in electronics.

[a)]Electronic mail: weiyuch@bltda.com.bta.net.cn

The results show that frequency analysis based on general periodic functions has a close relation with the Dirichlet multiplication in number theory.

In this paper $L^2[-\pi,\pi]$ denotes the real Hilbert space of functions with the period 2π which are quadratically integrable on $[-\pi,\pi]$. It is well known that $L^2[-\pi,+\pi]$ can be decomposed into three mutually orthogonal subspaces:

$$L^2[-\pi,+\pi] = \text{constant function subspace} \oplus L^2_{\text{even}}[-\pi,+\pi] \oplus L^2_{\text{odd}}[-\pi,+\pi]$$

$$= \text{span}\{1\} \oplus \overline{\text{span}\{\cos nx\}_{n=1}^{\infty}} \oplus \overline{\text{span}\{\sin nx\}_{n=1}^{\infty}},$$

where $L^2_{\text{even}}[-\pi,+\pi]$ is the subspace of even functions with the mean-value 0, and $L^2_{\text{odd}}[-\pi,+\pi]$ is the subspace of odd functions.

II. A FREQUENCY SYSTEM IN $L^2[-\pi,\pi]$

In this section we suppose that the function

$$F(x) = \sum_{n=1}^{\infty} A(n)\cos(nx) + B(n)\sin(nx) \tag{1}$$

is a given nonzero element in $L^2[-\pi,\pi]$, where $A(n)$ and $B(n)$ denote its Fourier coefficients. We shall show that the frequency system based on the function $F(x)$,

$$F(x), F(2x), F(3x), \ldots, F(nx), \ldots, \tag{2}$$

is linearly independent, weakly convergent to zero, and incomplete in $L^2[-\pi,\pi]$.

Lemma: Suppose that f and g are two arithmetical functions and h is their Dirichlet product,[1] i.e.,

$$h = f*g. \tag{3}$$

Then we have

$$h=0 \Leftrightarrow f=0 \quad \text{or} \quad g=0. \tag{4}$$

In this paper, as an example, the notation $h=0$ means that

$$h(n)=0 \quad \text{for} \quad n=1,2,3,\ldots. \tag{5}$$

Proof: From the definition of Dirichlet multiplication, it follows obviously that

$$f=0 \quad \text{or} \quad g=0 \Rightarrow h=0. \tag{6}$$

Now let us consider

$$h=0 \Rightarrow f=0 \quad \text{or} \quad g=0. \tag{7}$$

There are three cases.

(1) If $f(1)\neq 0$, then $g=h*f^{-1}=0$. Here -1 denotes the inverse of an arithmetical function.[1] Similarly, if $g(1)\neq 0$, then $f=h*g^{-1}=0$.
Therefore relation (7) holds in this case.

(2) Suppose that

$$f(n)=g(n)=0 \quad \text{for} \quad n<k \tag{8}$$

and that $f(k)\neq 0$, where k is a positive integer greater than 1.
Then we have

$$h(k^2)=f(k)g(k)=0\Rightarrow g(k)=0,$$

$$h(k(k+1))=f(k)g(k+1)=0\Rightarrow g(k+1)=0,$$

$$\cdots\ \cdots\ \cdots,$$

$$h(k(k+i))=f(k)g(k+i)=0\Rightarrow g(k+i)=0,$$

$$\cdots\ \cdots\ \cdots,$$

where i is any non-negative integer. Therefore we have $g=0$.

Similarly, suppose that

$$f(n)=g(n)=0 \quad \text{for } n<k \tag{9}$$

and that $g(k)\neq 0$, where k is a positive integer greater than 1.
For the same reason, we have $f=0$. Therefore relation (7) holds in this case as well.

(3) Suppose that

$$f(n)=g(n)=0 \quad \text{for any positive integer } n. \tag{10}$$

In fact we already have $f=g=0$ and therefore relation (7) holds obviously in this case. In one word, relation (7) holds always.
This completes the proof of the lemma. □

Proposition 1: We have

$$\sum_{n=1}^{\infty} c(n)F(nx)=0\Leftrightarrow c=0, \tag{11}$$

where $c(n)$ is a real-valued arithmetical function.

Proof: If

$$c=0, \tag{12}$$

the conclusion is trivially true.

Let us consider the case

$$\sum_{n=1}^{\infty} c(n)F(nx)=0. \tag{13}$$

Putting (1) in (13), we have

$$\sum_{n=1}^{\infty} (a(n)\cos nx+b(n)\sin nx)=0, \tag{14}$$

where

$$a(n)=\sum_{d|n} A\left(\frac{n}{d}\right)c(d)=A*c(n),$$

$$b(n)=\sum_{d|n} B\left(\frac{n}{d}\right)c(d)=B*c(n),$$

Here $d|n$ means that d is a factor of n.

From (14), we obtain that

$$a=b=0.$$

Since $F(x)$ is a nonzero element, we have

$$A\neq 0 \quad \text{or} \quad B\neq 0. \tag{15}$$

By the lemma, we obtain that

$$c=0, \tag{16}$$

This completes the proof. □

Consequence: If a function $\psi(x)\in L^2[-\pi,\pi]$ can be expressed as a frequency series based on $F(x)$:

$$\psi(x)=c(1)F(x)+c(2)F(2x)+c(3)F(3x)+\cdots+c(n)F(nx)+\cdots, \tag{17}$$

then the coefficients $c(n)$ are unique.

Proposition 2: The frequency system $\{F(nx)\}_{n=1}^{\infty}$ converges weakly to 0.

Proof: Suppose that $\psi(x)\in L^2[-\pi,+\pi]$ is an arbitrary function, and its Fourier series is

$$\psi(x)=a_0+\sum_{l=1}^{\infty} a(l)\cos(lx)+b(l)\sin(lx). \tag{18}$$

We have

$$\begin{aligned}|\langle F(nx),\psi(x)\rangle| &= \pi\left|\sum_{l=1}^{\infty} A(l)a(nl)+\sum_{l=1}^{\infty} B(l)b(nl)\right| \\ &\leqslant \pi\sqrt{\sum_{l=1}^{\infty} A^2(l)}\sqrt{\sum_{l=1}^{\infty} a^2(nl)}+\pi\sqrt{\sum_{l=1}^{\infty} B^2(l)}\sqrt{\sum_{l=1}^{\infty} b^2(nl)} \\ &\leqslant \pi\sqrt{\sum_{l=1}^{\infty} A^2(l)}\sqrt{\sum_{l=n}^{\infty} a^2(l)}+\pi\sqrt{\sum_{l=1}^{\infty} B^2(l)}\sqrt{\sum_{l=n}^{\infty} b^2(l)}\rightarrow 0 (\text{as } n\rightarrow\infty).\end{aligned}$$

Therefore we have

$$\langle F(nx),\psi(x)\rangle\rightarrow 0 \quad (\text{as } n\rightarrow\infty). \tag{19}$$

This completes the proof. □

Proposition 3: The frequency system $\{F(nx)\}_{n=1}^{\infty}$ is incomplete in $L^2[-\pi,+\pi]$.

Proof: There are two cases.

(1) If $A(1)\neq 0$ or $B(1)\neq 0$, then there exists a nonzero element $\psi(x)=-B(1)\cos x+A(1)\sin x$ such that

$$\int_{-\pi}^{\pi} F(nx)\psi(x)dx=0 \quad \text{for } n=1,2,3,\ldots. \tag{20}$$

(2) If $A(1)=B(1)=0$, we have

$$\int_{-\pi}^{\pi} F(nx)\cos x\, dx=0 \quad \text{for } n=1,2,3,\ldots. \tag{21}$$

In one word, the frequency system $\{F(nx)\}_{n=1}^{\infty}$ is incomplete in $L^2[-\pi,+\pi]$. □

Now that the frequency system is incomplete in $L^2[-\pi,+\pi]$, we shall develop the problem in its odd function subspace $L^2_{\mathrm{odd}}[-\pi,+\pi]$ in the following section.

III. A FREQUENCY SYSTEM IN $L^2_{\mathrm{odd}}[-\pi,+\pi]$

In this section we suppose that the function

$$Y(x)=\sum_{n=1}^{\infty} B(n)\sin nx \tag{22}$$

is a given nonzero element in $L^2_{\mathrm{odd}}[-\pi,+\pi]$, where $B(n)$ denotes its Fourier coefficients.

We shall show that the frequency system $\{Y(nx)\}_{n=1}^{\infty}$ may be an orthogonal basis, an unconditional basis, a complete system, or an incomplete system in $L^2_{\mathrm{odd}}[-\pi,+\pi]$. Let us see a simple example.

Example: If $Y(x)=\sin x+k\sin 2x$, where k is a real number, then in $L^2_{\mathrm{odd}}[-\pi,+\pi]$ the frequency system $\{Y(nx)\}_{n=1}^{\infty}$ is:

(1) an orthogonal basis when $k=0$;
(2) an unconditional basis when $0<|k|<1$;
(3) a complete system, but not a basis, when $|k|=1$;
(4) an incomplete system when $|k|>1$.

Proof of the example:
(1) This is clear.
(2) This follows from Proposition 18, which we shall introduce.
(3) Let us consider the completeness first.
Suppose that $\psi(x)\in L^2_{\mathrm{odd}}[-\pi,+\pi]$ is a function such that

$$\int_{-\pi}^{\pi}\psi(x)Y(nx)dx=0 \quad \text{for } n=1,2,3,\ldots. \tag{23}$$

The Fourier series of $\psi(x)$ can be expressed as

$$\psi(x)=\sum_{n=1}^{\infty} b(n)\sin nx. \tag{24}$$

From relation (23), we obtain that

$$b(n)+kb(2n)=0, \tag{25}$$

or

$$b(n)=(-k)b(2n), \tag{26}$$

for all positive integer n.

Thus for an arbitrary positive integer l, we have

$$b(l)=(-k)b(2l)=(-k)^2b(2^2l)=\cdots=(-k)^m b(2^m l)=\cdots. \tag{27}$$

Since $b(n)\to 0$ as $n\to\infty$, we have

$$b(l)=0. \tag{28}$$

Since l is an arbitrary positive integer, we obtain $\psi(x)=0$. This completes the proof of the completeness.
However, the frequency system is not a basis of $L^2_{\mathrm{odd}}[-\pi,+\pi]$.

If it is a basis, the function $\sin x$ can be expressed as a frequency series:

$$\sin x=\sum_{n=1}^{\infty} D(n)Y(nx), \tag{29}$$

where $D(n)$ are coefficients.

Putting $Y(x)=\sin x+k\sin 2x$ in (29), we obtain that

$$\sin x=D(1)\sin x+(kD(1)+D(2))\sin 2x+D(3)\sin 3x+(kD(2)+D(4))\sin 4x+\cdots. \tag{30}$$

Considering the coefficients of the terms $\sin 2^m x$ $(m=0,1,2,...)$, we obtain that

$$D(1)=1, \tag{31}$$

and

$$D(2^m)+kD(2^{m-1})=0 \quad \text{for } m=1,2,3,... . \tag{32}$$

Thus we obtain that

$$D(2^m)=(-k)^m \quad \text{for } m=1,2,3,... . \tag{33}$$

Therefore $D(2^m)$ does not converge to 0 as m increases, and hence $D(n)$ does not converge to 0 either as n increases. This is a contradiction.

So the conclusion should be that the frequency system is not a basis of $L^2_{\text{odd}}[-\pi,+\pi]$.

(4) The frequency system is not complete in $L^2_{\text{odd}}[-\pi,+\pi]$, because there exists a function

$$\psi(x)=\sin x+\left(-\frac{1}{k}\right)\sin 2x+\left(-\frac{1}{k}\right)^2\sin 4x+\left(-\frac{1}{k}\right)^3\sin 8x+\cdots+\left(-\frac{1}{k}\right)^m\sin 2^m x+\cdots \tag{34}$$

such that

$$\int_{-\pi}^{\pi}\psi(x)Y(nx)dx=0 \quad \text{for } n=1,2,3,... . \tag{35}$$

This completes the proof of the example. □

Proposition 4: The frequency system $\{Y(nx)\}_{n=1}^{\infty}$ is an orthonormal basis of $L^2_{\text{odd}}[-\pi,+\pi]$ if and only if

$$Y(x)=\pm\frac{1}{\sqrt{\pi}}\sin x. \tag{36}$$

Proof: If part is clear.

Next let us consider the only-if part. If $\{Y(nx)\}_{n=1}^{\infty}$ is an orthonormal basis of $L^2_{\text{odd}}[-\pi,+\pi]$, then the function $\sin x$ has a frequency series based on $Y(x)$:

$$\sin x=\sum_{n=1}^{\infty} D(n)Y(nx), \tag{37}$$

where the coefficients $D(n)$ are determined by

$$D(n)=\int_{-\pi}^{\pi}\sin x\,Y(nx)dx. \tag{38}$$

It follows that

$$D(1)=\pi B(1), \tag{39}$$

$$D(n)=0 \quad \text{for } n=2,3,4,\dots . \tag{40}$$

Thus we have

$$\sin x=\pi B(1)Y(x) \tag{41}$$

$$=\pi B(1)\sum_{n=1}^{\infty} B(n)\sin nx. \tag{42}$$

From (42), we obtain that

$$B(1)=\pm\frac{1}{\sqrt{\pi}}, \tag{43}$$

$$B(n)=0 \quad \text{for } n=2,3,4,\dots, \tag{44}$$

i.e.,

$$Y(x)=\pm\frac{1}{\sqrt{\pi}}\sin x \qquad \square \tag{45}$$

Proposition 5: The frequency system $\{Y(nx)\}_{n=1}^{\infty}$ is incomplete in $L^2_{\text{odd}}[-\pi,+\pi]$ if $Y(x)$'s first Fourier coefficient $B(1)=0$.

Proof: It is because

$$\int_{-\pi}^{\pi} Y(nx)\sin x\, dx=0, \tag{46}$$

for $n=1,2,3,\dots$. □

Proposition 6: If $B(1)\neq 0$, then the frequency system

$$Y(x),Y(2x),Y(3x),\cdots,Y(nx),\dots \tag{47}$$

and the function system

$$g_1(x),g_2(x),g_3(x),\dots,g_n(x),\dots \tag{48}$$

are biorthogonal, where the functions $g_n(x)$ are defined by

$$g_n(x)=\frac{1}{\pi}\sum_{d|n} B^{-1}\left(\frac{n}{d}\right)\sin dx \tag{49}$$

for $n=1,2,3,\dots$. Here B^{-1} denotes the Dirichlet inverse of $B(n)$.

Proof: For any two positive integers m and n, we have

$$\langle Y(mx),g_n(x)\rangle=\left\langle \sum_{k=1}^{\infty} B(k)\sin(mkx), \frac{1}{\pi}\sum_{d|n} B^{-1}\left(\frac{n}{d}\right)\sin dx\right\rangle$$

$$=\sum_{m|d,d|n} B^{-1}\left(\frac{n}{d}\right)B\left(\frac{d}{m}\right)$$

$$=\begin{cases} 0 & m\dagger n \\ 0 & \dfrac{n}{m}=2,3,... \\ 1 & \dfrac{n}{m}=1 \end{cases}$$

$$=\begin{cases} 0 & m\neq n \\ 1 & m=n \end{cases}=\delta_{mn},$$

where $m\dagger n$ denotes that m is not a factor of n, and δ_{mn} is the Kronecher's delta symbol. This completes the proof. □

Proposition 7: If a function $\psi(x)\in L^2_{\text{odd}}[-\pi,+\pi]$ has a frequency series based on the function $Y(x)$ (with $B(1)\neq 0$):

$$\psi(x)=\sum_{n=1}^{\infty} D(n)Y(nx), \tag{50}$$

then the coefficients $D(n)$ can be determined uniquely by

$$D(n)=\int_{-\pi}^{\pi}\psi(x)g_n(x)dx, \tag{51}$$

where $g_n(x)$ are the biorthogonal functions defined by (49).

Proof: This proposition follows from Proposition 6 obviously. □

In the following sections, we shall discuss what properties the function $Y(x)$ need possess so that its frequency system is a complete system or an unconditional basis.

IV. A COMPLETE SYSTEM IN $L^2_{\text{odd}}[-\pi,+\pi]$

We shall discuss what properties the function $Y(x)$ need possess so that its frequency system is a complete system in $L^2_{\text{odd}}[-\pi,+\pi]$.

Definition: The subset W of $L^2_{\text{odd}}[-\pi,+\pi]$ is defined by

$$W=\left\{ Y(x)\in L^2_{\text{odd}}[-\pi,+\pi]:Y(x)=\sum_{n=1}^{\infty} B(n)\sin nx, \quad \text{with } \sum_{l=1}^{\infty}\left(\sum_{mn=1}|B^{-1}(n)B(m)|\right)^2<\infty.\right\}$$

Proposition 8: If $Y(x)\in W$, then

(1) the series

$$\sin x=\sum_{n=1}^{\infty} B^{-1}(n)Y(nx) \tag{52}$$

converges unconditionally in the sense of the norm of $L^2[-\pi,+\pi]$;
(2) the frequency system $\{Y(nx)\}_{n=1}^{\infty}$ is complete in $L^2_{\text{odd}}[-\pi,+\pi]$.

Proof:

(1) In fact we have

$$\sum_{n=1}^{\infty} B^{-1}(n)Y(nx) \tag{53}$$

$$=\sum_{n=1}^{\infty} B^{-1}(n)\sum_{m=1}^{\infty} B(m)\sin(mnx) \tag{54}$$

$$=\sum_{l=1}^{\infty} \sum_{mn=l} B^{-1}(n)B(m)\sin lx \tag{55}$$

$$=\sin x. \tag{56}$$

From (54) to (55), the summation order was changed. But this does not matter for the series encountered here are unconditionally convergent in the sense of the norm of $L^2[-\pi,\pi]$. The reason for unconditional convergence is that the series

$$\sum_{l=1}^{\infty} \sum_{mn=l} |B^{-1}(n)B(m)|\sin lx$$

is convergent, and its norm square

$$\pi\sum_{l=1}^{\infty} \left(\sum_{mn=l} |B^{-1}(n)B(m)|\right)^2<\infty. \tag{57}$$

(2) From (52), it follows that

$$\sin mx=\sum_{n=1}^{\infty} B^{-1}(n)Y(mnx) \tag{58}$$

for $m=1,2,3,\ldots$.

Since the sine function system $\{\sin nx\}_{n=1}^{\infty}$ is complete in $L^2_{\rm odd}[-\pi,+\pi]$, so is the frequency system $\{Y(nx)\}_{n=1}^{\infty}$. □

Definition: The subset Y of $L^2_{\rm odd}[-\pi,+\pi]$ is defined by

$$Y=\left\{Y(x):\ Y(x)=\sum_{n=1}^{\infty} B(n)\sin x,\right.$$

$$\left.\text{where } B(n) \text{ is quadratically summable and completely multiplicative.}\right\}$$

If $Y(x)\in Y$, one says that $Y(x)$ is an odd function with quadratically summable and completely multiplicative Fourier coefficients, or an odd WQSCMFC function. Here the term that $B(n)$ is quadratically summable means that

$$\sum_{n=1}^{\infty} B^2(n)<\infty.$$

The term that $B(n)$ is completely multiplicative means that $B(1)=1$ and $B(m)B(n)=B(mn)$ for any two positive integers.[1]

Proposition 9: Y is a subset of W, i.e., $Y\subset W$.

Proof: Since $B(n)$ is completely multiplicative, we have and $B^{-1}(n)=\mu(n)B(n)$.[1] It follows that

$$\sum_{l=1}^{\infty}\left(\sum_{mn=l}|B^{-1}(n)B(m)|\right)^2=\sum_{l=1}^{\infty}B^2(l)\left(\sum_{mn=l}|\mu(n)|\right)^2$$
$$\leqslant\sum_{l=1}^{\infty}B^2(l)d^2(l)$$
$$\leqslant\sum_{l=1}^{\infty}B^2(l)d*d(l)=\left(\sum_{l=1}^{\infty}B^2(l)\right)^4<\infty.$$

Here $d(n)$ denotes the number of positive divisors of a positive integer n, called the divisor function.[1] We have used the relation

$$d*d(n)=\sum_{m|n}d(m)d\left(\frac{n}{m}\right)\geqslant\sum_{m|n}d(n)=d^2(n). \tag{59}$$

Therefore we obtain that $Y\subset W$. □

Proposition 10: If $Y(x)\in Y$, and

$$\psi(x)=\sum_{n=1}^{\infty}b(n)\sin nx$$

is a function in $L^2_{\mathrm{odd}}[-\pi,+\pi]$ with

$$\sum_{n=1}^{\infty}b^2(n)d(n)<\infty, \tag{60}$$

then $\psi(x)$ can be expressed as a frequency system based on $Y(x)$:

$$\psi(x)=\sum_{n=1}^{\infty}D(n)Y(nx),$$

where

$$D(n)=\sum_{d|n}\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)b(d)=B^{-1}*b(n). \tag{61}$$

Proof: This proof is similar to that in Ref. 2. With the help of Ref. 2, readers should have no difficulties in constructing a complete proof except for the use of a new inequality

$$\sum_{j=1}^{\infty}B^2(j)d^3(j)\leqslant\left(\sum_{j=1}^{\infty}B^2(j)\right)^8<\infty. \tag{62}$$

It is because

$$d*d*d*d(n)\geqslant d^2*d^2(n)=\sum_{m|n}d^2(m)d^2\left(\frac{n}{m}\right)\geqslant\sum_{m|n}d^2(n)=d^3(n). \tag{63}$$

Proposition 11: If $Y(x)\in Y$, the orthonormalized functions of the frequency system $\{Y(nx)\}$ can be given explicitly by

$$e_n(x)=\frac{1}{\|Y(x)\|}\frac{1}{\sqrt{\beta(n)}}\sum_{d|n}\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)Y(dx), \tag{64}$$

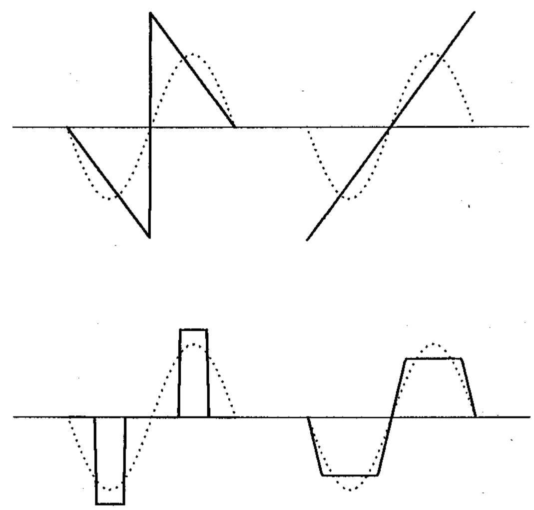

FIG. 1. The I-sawtooth wave, the II-sawtooth wave, the three-valued function, and the trapezoidal wave.

where the arithmatical function $\beta(n)$ is defined by

$$\beta(n)=\sum_{d|n}\mu(d)B^2(d). \tag{65}$$

Proof: Since this proof has little difference from that in Ref. 2, we omit it. Obviously the orthonormal system $\{e_n\}_{n=1}^{\infty}$ is an orthonormal basis of $L^2_{\text{odd}}[-\pi,+\pi]$. □

Proposition 12: If $Y_1(x)\in Y$, $Y_2(x)\in Y$, and that their Fourier coefficients are $B_1(n)$ and $B_2(n)$, respectively, then the two functions are related by

$$Y_1(x)=\sum_{n=1}^{\infty}D_{12}(n)Y_2(nx), \tag{66}$$

$$Y_2(x)=\sum_{n=1}^{\infty}D_{21}(n)Y_1(nx), \tag{67}$$

where

$$D12(n)=\sum_{d|n}\mu\left(\frac{n}{d}\right)B_2\left(\frac{n}{d}\right)B_1(d)=B_1*B_2^{-1}(n), \tag{68}$$

$$D21(n)=\sum_{d|n}\mu\left(\frac{n}{d}\right)B_1\left(\frac{n}{d}\right)B_2(d)=B_2*B_1^{-1}(n). \tag{69}$$

Proof: From Proposition 10, we can obtain this proposition easily. □

As examples, let us consider several basic wave forms, which are frequently used and easily generated in electronics.

Example 1: The I-sawtooth wave. The I-sawtooth wave $Y_{\text{Isa}}(x)$ is a function defined on $(-\infty,+\infty)$ with the period 2π, whose value in one period is given by

$$Y_{\text{Isa}}(x)=\begin{cases}(\pi-x)/2 & 0<x<2\pi\\ 0 & x=0\end{cases}. \tag{70}$$

See Fig. 1. By the way, in Fig. 1 the graphs of four functions are arranged in the following order:

$$\begin{pmatrix} Y_{\mathrm{Isa}}(x) & Y_{\mathrm{IIsa}}(x) \\ Y(\alpha,a;x) & Y_{\mathrm{tra}}(\alpha,a;x) \end{pmatrix}.$$

Its Fourier series is[3]

$$Y_{\mathrm{Isa}}(x)=\sin x+\tfrac{1}{2}\sin(2x)+\tfrac{1}{3}\sin(3x)+\cdots+\frac{1}{n}\sin(nx)+\cdots \tag{71}$$

$$=\sum_{n=1}^{\infty} B_{\mathrm{Isa}}(n)\sin(nx). \tag{72}$$

Its Fourier coefficient

$$B_{\mathrm{Isa}}(n)=1/n \tag{73}$$

is completely multiplicative, so the I-sawtooth wave $Y_{\mathrm{Isa}}(x)$ is an odd WQSCMFC function, i.e., $Y_{\mathrm{Isa}}(x)\in Y$.

Therefore the sine function can be considered as a superposition of I-sawtooth waves with different frequencies:

$$\sin x=Y_{\mathrm{Isa}}(x)-\tfrac{1}{2}Y_{\mathrm{Isa}}(2x)-\tfrac{1}{3}Y_{\mathrm{Isa}}(3x)+\cdots+\frac{\mu(n)}{n}Y_{\mathrm{Isa}}(nx)+\cdots. \tag{74}$$

The frequency system based on the I-sawtooth

$$Y_{\mathrm{Isa}}(x),\quad Y_{\mathrm{Isa}}(2x),\quad Y_{\mathrm{Isa}}(3x),\ldots,Y_{\mathrm{Isa}}(nx),\ldots \tag{75}$$

is complete in $L^2_{\mathrm{odd}}[-\pi,+\pi]$. (However whether it is a basis of $L^2_{\mathrm{odd}}[-\pi,+\pi]$ is still unknown.) Its biorthogonal system is

$$\frac{1}{\pi}\sin x,\ \frac{1}{\pi}(\sin 2x-\tfrac{1}{2}\sin x),\ \frac{1}{\pi}\left(\sin 3x-\frac{1}{3}\sin x\right),\ldots,\frac{1}{\pi}\sum_{d|n}\frac{d}{n}\mu\left(\frac{n}{d}\right)\sin dx,\ldots, \tag{76}$$

and its orthonormalized system is

$$\sqrt{\frac{6}{\pi^3}}Y_{\mathrm{Isa}}(x),\quad \sqrt{\frac{8}{\pi^3}}\left(Y_{\mathrm{Isa}}(2x)-\frac{1}{2}Y_{\mathrm{Isa}}(x)\right),\quad \sqrt{\frac{27}{4\pi^3}}\left(Y_{\mathrm{Isa}}(3x)-\frac{1}{3}Y_{\mathrm{Isa}}(x)\right),\ldots,$$

$$\sqrt{\frac{6}{\pi^3}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|n}\frac{d}{n}\mu\left(\frac{n}{d}\right)Y_{\mathrm{Isa}}(dx),\ldots,$$

where

$$\beta(n)=\sum_{d|n}\frac{\mu(d)}{d^2}. \tag{77}$$

Example 2: The II-sawtooth wave. The II-sawtooth wave $Y_{\mathrm{IIsa}}(x)$ is a function defined on $(-\infty,+\infty)$ with the period 2π (Fig. 1), whose value in one period is given by

$$Y_{\mathrm{IIsa}}(x)=\begin{cases} x/2 & -\pi<x<\pi \\ 0 & x=-\pi,\pi. \end{cases} \tag{78}$$

Its Fourier series is[3]

$$Y_{\mathrm{IIsa}}(x)=\sin x-\frac{1}{2}\sin(2x)+\frac{1}{3}\sin(3x)+\cdots+\frac{(-1)^{n+1}}{n}\sin(nx)+\cdots \tag{79}$$

$$=\sum_{n=1}^{\infty} B_{\mathrm{IIsa}}(n)\sin(nx). \tag{80}$$

Its Fourier coefficient

$$B_{\mathrm{IIsa}}(n)=\frac{(-1)^{n+1}}{n}=\frac{\theta(n)}{n} \tag{81}$$

is not completely multiplicative, so the II-sawtooth wave $Y_{\mathrm{IIsa}}(x)$ is not a WQSCMFC function, i.e., $Y_{\mathrm{IIsa}}(x)$ is not in the set Y. Here θ is the arithmetical function defined by

$$\theta(n)=(-1)^{n+1}. \tag{82}$$

Furthermore, it is not difficult to prove that $Y_{\mathrm{IIsa}}(x)$ is not in the set W either [hint: see (94)]. Still, the frequency system based on the II-sawtooth

$$Y_{\mathrm{IIsa}}(x),\ \ Y_{\mathrm{IIsa}}(2x),\ \ Y_{\mathrm{IIsa}}(3x),\ldots,Y_{\mathrm{IIsa}}(nx),\ldots \tag{83}$$

is a complete system in $L^2_{\mathrm{odd}}[-\pi,+\pi]$ (see Proposition 13). Its biorthogonal system is

$$\frac{1}{\pi}\sin x,\ \ \frac{1}{\pi}\left(\sin 2x+\frac{1}{2}\sin x\right),\ \ \frac{1}{\pi}\left(\sin 3x-\frac{1}{3}\sin x\right),\ldots,\frac{1}{\pi}\sum_{d|n}\frac{d}{n}\,\theta^{-1}\left(\frac{n}{d}\right)\sin dx,\ldots, \tag{84}$$

and its orthonormalized system is

$$\sqrt{\frac{6}{\pi^3}}\,Y_{\mathrm{IIsa}}(x),\ \ 4\sqrt{\frac{2}{5\pi^3}}\left(Y_{\mathrm{IIsa}}(2x)+\frac{1}{4}Y_{\mathrm{IIsa}}(x)\right),\ \ \sqrt{\frac{27}{4\pi^3}}\left(Y_{\mathrm{IIsa}}(3x)-\frac{1}{3}Y_{\mathrm{IIsa}}(x)\right),$$

$$\sqrt{\frac{20}{3\pi^3}}\left(Y_{\mathrm{IIsa}}(4x)+\frac{3}{10}Y_{\mathrm{IIsa}}(2x)+\frac{1}{5}Y(x)\right),\ldots .$$

Here θ^{-1} is the Dirichlet inversion of θ, which is given explicitly by

$$\theta^{-1}(n)=\begin{cases}\mu(n) & l=0\\ 2^{l-1}\mu(2k+1) & l=1,2,3\ldots .\end{cases} \tag{85}$$

Here l and k are determined by

$$n=2^l(2k+1). \tag{86}$$

Proposition 13: The frequency system based on the II-sawtooth wave, i.e., $\{Y_{\mathrm{IIsa}}(nx)\}_{n=1}^{\infty}$, is a complete system in $L^2_{\mathrm{odd}}[-\pi,+\pi]$, but it is not a basis of $L^2_{\mathrm{odd}}[-\pi,+\pi]$.

Proof:

(1) It is complete in $L^2_{\mathrm{odd}}[-\pi,+\pi]$.

Suppose that there exists a function $\psi(x)\in L^2_{\mathrm{odd}}[-\pi,+\pi]$ such that

$$\langle Y_{\mathrm{IIsa}}(nx),\psi(x)\rangle=0 \quad \text{for } n=1,2,3,\ldots . \tag{87}$$

From the definitions of the two sawtooth waves, we have the relation

$$Y_{\rm IIsa}(x)=Y_{\rm Isa}(x)-Y_{\rm Isa}(2x). \tag{88}$$

Thus we obtain that

$$\langle Y_{\rm Isa}(nx),\psi(x)\rangle=\langle Y_{\rm Isa}(2nx),\psi(x)\rangle \quad \text{for} \;\; n=1,2,3,... . \tag{89}$$

Furthermore, for an arbitrary positive integer l we have

$$\langle Y_{\rm Isa}(lx),\psi(x)\rangle=\langle Y_{\rm Isa}(2lx),\psi(x)\rangle=\cdots=\langle Y_{\rm Isa}(2^m lx),\psi(x)\rangle=\cdots, \tag{90}$$

where $m=0,1,2,...$. Due to Proposition 2, $Y_{\rm Isa}(nx)$ converges weakly to zero, and we have

$$\langle Y_{\rm Isa}(lx),\psi(x)\rangle=0. \tag{91}$$

Since $\{Y_{\rm Isa}(nx)\}_{n=1}^{\infty}$ is complete, we have

$$\psi(x)=0. \tag{92}$$

This completes the proof about the completeness.

(2) It is not a basis of $L^2_{\rm odd}[-\pi,+\pi]$.

If it is a basis of $L^2_{\rm odd}[-\pi,+\pi]$, by Proposition 7, $\sin x$ can be expressed as a frequency series:

$$\sin x=Y_{\rm IIsa}(x)+\frac{1}{2}Y_{\rm IIsa}(2x)-\frac{1}{3}Y_{\rm IIsa}(3x)+\cdots+\frac{\theta^{-1}(n)}{n}Y_{\rm IIsa}(nx)+\cdots. \tag{93}$$

But this is impossible, because $\theta^{-1}(n)/n$ does not converge to zero. In fact, for $n=2^m$, $m=1,2,3,...$, we always have

$$\frac{\theta^{-1}(n)}{n}=\frac{1}{2}. \tag{94}$$

Therefore it is not a basis of $L^2_{\rm odd}[-\pi,+\pi]$.

Example 3: A periodic three-valued function. Suppose that $Y(\alpha,a;x)$ is a three-valued function of the period 2π (Fig. 1), the value in one period of which is given by

$$Y(\alpha,a;x)=\begin{cases} 0 & -\pi<x<-\pi+\alpha,-\alpha<x<\alpha,-\pi-\alpha<x<\pi \\ -a & -\pi+\alpha<x<-\alpha \\ a & \alpha<x<\pi-\alpha, \end{cases} \tag{95}$$

where $a>0$ and $0\leqslant\alpha<\pi/2$.

The Fourier series of $Y(\alpha,a;x)$ is[3]

$$Y(\alpha,a;x)=\frac{4a}{\pi}\left(\cos\alpha\sin x+\frac{1}{3}\cos 3\alpha\sin 3x+\frac{1}{5}\cos 5\alpha\sin 5x+\cdots\right) \tag{96}$$

$$=\sum_{n=1}^{\infty}B(n)\sin nx, \tag{97}$$

where the Fourier coefficients

$$B(n)=\begin{cases} \dfrac{4a}{\pi}\dfrac{\cos n\alpha}{n}, & n=1,3,5,... \\ 0, & n=2,4,6,... . \end{cases} \tag{98}$$

Proposition 14: $Y(\alpha,a;x)$ is an odd WQSCMFC function, i.e., $Y(\alpha,a;x)\in Y$ if and only if

$$(0)\quad \alpha=0,\quad a=\frac{\pi}{4}, \tag{99}$$

$$(1)\quad \alpha=\frac{\pi}{4},\quad a=\frac{\sqrt{2}\pi}{4}, \tag{100}$$

or

$$(2)\quad \alpha=\frac{\pi}{6},\quad a=\frac{\sqrt{3}\pi}{6}. \tag{101}$$

Proof: It is easy to verify that in the case (0), (1) or (2), the function $Y(\alpha,a;x)$ is an odd WQSCMFC function indeed.

Now let us prove that except for the three cases $Y(\alpha,a;x)$ is not an odd WQSCMFC function. If $Y(\alpha,a;x)$ is an odd WQSCMFC function, then its Fourier coefficients

$$B(n)=\begin{cases}\dfrac{4a}{\pi}\dfrac{\cos n\alpha}{n}, & n=1,3,5,\ldots\\ 0, \quad n=2,4,6,\ldots\end{cases} \tag{102}$$

are a completely multiplicative arithmetical function. From $B(1)=1$ it follows that

$$a=\frac{\pi}{4}\frac{1}{\cos\alpha}.$$

From $B^2(3)=B(9)$ it follows that

$$\left(\frac{\cos 3\alpha}{\cos\alpha}\right)^2=\frac{\cos 9\alpha}{\cos\alpha}. \tag{103}$$

Equation (103) has three solutions in $[0,\pi/2)$:

$$\alpha=0,\quad \alpha=\frac{\pi}{4},\quad \text{or}\quad \alpha=\frac{\pi}{6}.$$

The associated value of a is

$$a=\frac{\pi}{4},\quad a=\frac{\sqrt{2}\pi}{4},\quad \text{or}\quad a=\frac{\sqrt{3}\pi}{6}.$$

Thus there exist at most three WQSCMFC function among all $Y(\alpha,a;x)$ with $a>0$ and $0\leqslant\alpha<\pi/2$.
This completes the proof. □

For convenience, we call the three functions

$$Y\left(0,\frac{\pi}{4};x\right)=Y_{\mathrm{sq}}(x), \tag{104}$$

$$Y\left(\frac{\pi}{4},\frac{\sqrt{2}\pi}{4};x\right)=Y_{\pi/4}(x), \tag{105}$$

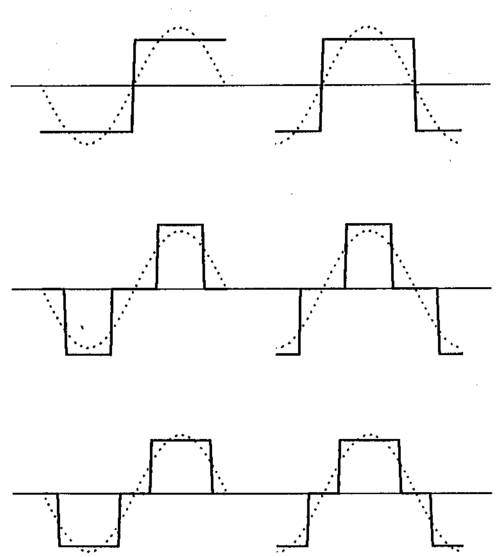

FIG. 2. The odd and even square waves, $\pi/4$ three-valued functions, and $\pi/6$ three-valued functions.

$$Y\left(\frac{\pi}{6},\frac{\sqrt{3}\pi}{6};x\right)=Y_{\pi/6}(x), \tag{106}$$

odd square wave, $\pi/4$ three-valued function, and $\pi/6$ three-valued function, respectively, which are common in electronics. See Fig. 2. By the way, in Fig. 2 the graphs of six functions are arranged in the following order:

$$\begin{pmatrix} Y_{\text{sq}}(x) & X_{\text{sq}}(x) \\ Y_{\pi/4}(x) & X_{\pi/4}(x) \\ Y_{\pi/6}(x) & X_{\pi/6}(x) \end{pmatrix}.$$

Regarding the three-valued functions the following questions are still open.

Questions:

(1) In the case (0), (1), or (2), is the frequency system $\{Y(\alpha,a;nx)\}$ a basis of $L^2_{\text{odd}}[-\pi,+\pi]$?

(2) Except for the case (0), (1), and (2), is the frequency system $\{Y(\alpha,a;nx)\}$ a complete system in $L^2_{\text{odd}}[-\pi,+\pi]$?

By Proposition 12, the sawtooth wave $Y_{\text{Isa}}(x)$ and the odd square waves $Y_{\text{sq}}(x)$ are related by

$$Y_{\text{Isa}}(x)=Y_{\text{sq}}(x)+\frac{1}{2}Y_{\text{sq}}(2x)+\frac{1}{4}Y_{\text{sq}}(4x)+\cdots+\frac{1}{2^m}Y_{\text{sq}}(2^m)+\ldots, \tag{107}$$

and

$$Y_{\text{sq}}(x)=Y_{\text{Isa}}(x)-\tfrac{1}{2}Y_{\text{Isa}}(2x). \tag{108}$$

V. AN UNCONDITIONAL BASIS IN $L^2_{\text{odd}}[-\pi,+\pi]$

We shall discuss what properties a function need possess so that its frequency system is an unconditional basis of $L^2_{\text{odd}}[-\pi,+\pi]$.

Definition: A function set B is defined by

$$B=\left\{Y(x):\ Y(x)=\sum_{n=1}^{\infty} B(n)\sin nx \text{ where } B(n) \text{ satisfies } \sum_{n=1}^{\infty}|B(n)|<\infty,\quad \sum_{n=1}^{\infty}|B^{-1}(n)|<\infty\right\}.$$

Obviously, B is a subset of W.

Proposition 15: If $Y(x)\in B$, then

(1) the series

$$\sin x=\sum_{n=1}^{\infty} B^{-1}(n)Y(nx) \tag{109}$$

converges absolutely and uniformly in $[-\pi,\pi]$;

(2) the frequency system $\{Y(nx)\}$ is an unconditional basis of $L^2_{\text{odd}}[-\pi,+\pi]$.

Proof:

(1) In fact we have

$$\sum_{n=1}^{\infty} B^{-1}(n)Y(nx) \tag{110}$$

$$=\sum_{n=1}^{\infty} B^{-1}(n)\sum_{m=1}^{\infty} B(m)\sin(mnx) \tag{111}$$

$$=\sum_{l=1}^{\infty}\sum_{mn=l} B^{-1}(n)B(m)\sin lx \tag{112}$$

$$=\sin x. \tag{113}$$

Since

$$\sum_{n=1}^{\infty}|B^{-1}(n)|\sum_{m=1}^{\infty}|B(m)\sin(mnx)|\le\sum_{n=1}^{\infty}|B^{-1}(n)|\sum_{m=1}^{\infty}|B(m)|<\infty,$$

by Weierstrass M-test, we obtain that the series (109) converges absolutely and uniformly in $[-\pi,\pi]$.

(2) Suppose that $\psi(x)\in L^2_{\text{odd}}[-\pi,+\pi]$ is an arbitrary function and that $b(n)$ denotes its Fourier coefficients. In fact, we have

$$\psi(x)=\sum_{n=1}^{\infty} b(n)\sin nx \tag{114}$$

$$=\sum_{n=1}^{\infty} b(n)\sum_{l=1}^{\infty} B^{-1}(l)Y(nlx) \tag{115}$$

$$=\sum_{n=1}^{\infty}\left(\sum_{d|n} b(d)B^{-1}\left(\frac{n}{d}\right)\right)Y(nx) \tag{116}$$

$$=\sum_{n=1}^{\infty} D(n)Y(nx), \tag{117}$$

where

$$D(n)=\sum_{d|n} b(d)B^{-1}\left(\frac{n}{d}\right)=B^{-1}*b(n). \tag{118}$$

From (115) to (116), the summation orders were changed. But this does not matter because all the following series

$$\sum_{n=1}^{\infty} b(n)\sum_{l=1}^{\infty} B^{-1}(l)Y(nlx)=\sum_{n=1}^{\infty} b(n)\sum_{l=1}^{\infty} B^{-1}(l)\sum_{m=1}^{\infty} B(m)\sin(nlmx)$$
$$=\sum_{k=1}^{\infty}\left(\sum_{nlm=k} b(n)B^{-1}(l)B(m)\right)\sin(kx)$$

are unconditionally convergent in $L^2[-\pi,\pi]$, or one can say, they have sums independent of order. The reason for the unconditional convergence is that the series

$$\sum_{k=1}^{\infty}\left(\sum_{nlm=k}|b(n)B^{-1}(l)B(m)|\right)\sin(kx) \tag{119}$$

is convergent, the norm of which is

$$\left\|\sum_{k=1}^{\infty}\left(\sum_{nlm=k}|b(n)B^{-1}(l)B(m)|\right)\sin(kx)\right\| \tag{120}$$

$$=\left\|\sum_{m=1}^{\infty}|B(m)|\sum_{l=1}^{\infty}|B^{-1}(l)|\sum_{n=1}^{\infty}|b(n)|\sin(nlmx)\right\| \tag{121}$$

$$\leqslant\sum_{m=1}^{\infty}|B(m)|\sum_{l=1}^{\infty}|B^{-1}(l)|\left\|\sum_{n=1}^{\infty}|b(n)|\sin(nlmx)\right\| \tag{122}$$

$$=\sum_{m=1}^{\infty}|B(m)|\sum_{l=1}^{\infty}|B^{-1}(l)|\sqrt{\pi\sum_{n=1}^{\infty}b^2(n)} \tag{123}$$

$$<\infty. \tag{124}$$

We can obtain the uniqueness of the series (117) by the consequence of Proposition 1.
Therefore $\{Y(nx)\}$ is an unconditional basis of $L^2[-\pi,\pi]$.
This completes the proof. □

Example: The function

$$Y(x)=\sin x-\frac{1}{2^2}\sin 2x+\frac{1}{3^2}\sin 3x+\cdots+\frac{(-1)^{n+1}}{n^2}\sin nx+\ldots, \tag{125}$$

is an element of *B*, since $B(n)$ satisfy that

$$B(1)=1,\quad \sum_{n=1}^{\infty}|B(n)|=\frac{\pi^2}{6},\quad \sum_{n=1}^{\infty}|B^{-1}(n)|=\frac{18}{\pi^2}. \tag{126}$$

Definition: The function set Y_a is defined by

$$Y_a=\left\{Y(x):Y(x)=\sum_{n=1}^{\infty}B(n)\sin nx,\right.$$

$$\left.\text{where } B(n) \text{ is absolutely summable and completely multiplicative}\right\}.$$

If $Y(x)\in Y_a$, one says that $Y(x)$ is an odd function with absolutely summable and completely multiplicative Fourier coefficients, or an odd WASCMFC function.

Here the term that $B(n)$ is absolutely summable means that

$$\sum_{n=1}^{\infty}|B(n)|<\infty.$$

Proposition 16: Y_a is a subset of *B*, i.e., $Y_a\subset B$.

Proof: If $B(n)$ is completely multiplicative, we have

$$\sum_{n=1}^{\infty}|B^{-1}(n)|=\sum_{n=1}^{\infty}|\mu(n)B(n)|\leqslant\sum_{n=1}^{\infty}|B(n)|<\infty. \tag{127}$$

Additionally, we have $Y_a\subset Y$.

Example: Trapezoidal wave. The trapezoidal wave $Y_{\text{tra}}(a,\alpha;x)$ is a function of the period 2π (Fig. 1), the value in one period of which is given by

$$Y_{\text{tra}}(\alpha,a;x)=\begin{cases}ax/\alpha & \text{for}-\alpha\leqslant x\leqslant\alpha\\ a & \text{for }\ \alpha\leqslant x\leqslant\pi-\alpha\\ -a & \text{for}-\pi+\alpha\leqslant x\leqslant-\alpha\\ a(\pi-x)/\alpha & \text{for }\ \pi-\alpha\leqslant x\leqslant\pi\\ -a(\pi+x)/\alpha & \text{for}-\pi\leqslant x\leqslant-\pi+\alpha,\end{cases} \tag{128}$$

where

$$0<\alpha\leqslant\pi/2, a>0. \tag{129}$$

Its Fourier series is[3]

$$Y_{\text{tra}}(\alpha,a;x)=\frac{4a}{\pi\alpha}\left(\sin\alpha\sin x+\frac{1}{3^2}\sin 3\alpha\sin 3x+\frac{1}{5^2}\sin 5\alpha\sin 5x+\cdots\right)=\sum_{n=1}^{\infty}B(n)\sin nx,$$

where

$$B(n)=\begin{cases}\dfrac{4a}{\pi\alpha}\dfrac{1}{n^2}\sin n\alpha, & n=1,3,5,\ldots\\ 0, & n=2,4,6,\ldots.\end{cases} \tag{130}$$

Proposition 17: The function $Y_{\text{tra}}(\alpha,a;x)$ is an odd WASCMFC function if and only if

$$(0)\quad \alpha=\frac{\pi}{2},\quad a=\frac{\pi^2}{8}, \tag{131}$$

$$(1)\quad \alpha=\frac{\pi}{4},\quad a=\frac{\sqrt{2}\pi^2}{16}, \tag{132}$$

or

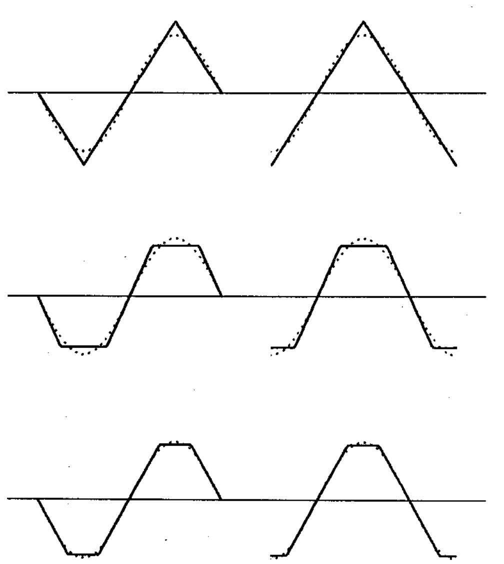

FIG. 3. The odd and even triangular waves, $\pi/4$ trapezoidal waves, and $\pi/3$ trapezoidal waves.

$$(2) \quad \alpha=\frac{\pi}{3}, \quad a=\frac{\sqrt{3}\pi^2}{18}. \tag{133}$$

Proof: This proof is similar to that of Proposition (14). □

For convenience, we call the three functions

$$Y_{\rm tra}\left(\frac{\pi}{2},\frac{\pi^2}{8};x\right)=Y_{\rm tri}(x), \tag{134}$$

$$Y_{\rm tra}\left(\frac{\pi}{4},\frac{\sqrt{2}\pi^2}{16};x\right)=Y_{\pi/4{\rm tra}}(x), \tag{135}$$

$$Y_{\rm tra}\left(\frac{\pi}{3},\frac{\sqrt{3}\pi^2}{18};x\right)=Y_{\pi/3{\rm tra}}(x), \tag{136}$$

odd triangular wave, $\pi/4$ trapezoidal wave, and $\pi/3$ trapezoidal wave, respectively, which are common in electronics. See Fig. 3. By the way, in Fig. 3 the graphs of six functions are arranged in the following order:

$$\begin{pmatrix} Y_{\rm tri}(x) & X_{\rm tri}(x) \\ Y_{\pi/4{\rm tra}}(x) & X_{\pi/4{\rm tra}}(x) \\ Y_{\pi/3{\rm tra}}(x) & X_{\pi/3{\rm tra}}(x) \end{pmatrix}.$$

Proposition 18: If the Fourier coefficients $B(n)$ of $Y(x)\in L^2_{\rm odd}[-\pi,+\pi]$ satisfy that

$$B(1)=1, \quad \sum_{n=2}^{\infty}|B(n)|<1, \tag{137}$$

then $Y(x)\in B$.

Proof: In fact we have

$$\left(\sum_{n=1}^{\infty} B(n)\right)^{-1} = \left(1+\sum_{n=2}^{\infty} B(n)\right)^{-1}$$
$$= 1+(-1)\sum_{n=2}^{\infty} B(n)+\cdots+(-1)^m\left(\sum_{n=2}^{\infty} B(n)\right)^m+\cdots = \sum_{n=1}^{\infty} B^{-1}(n).$$

Therefore we obtain that

$$\sum_{n=1}^{\infty} |B^{-1}(n)| \leqslant 1+\sum_{n=2}^{\infty} |B(n)|+\cdots+\left(\sum_{n=2}^{\infty} |B(n)|\right)^m+\cdots = \frac{1}{1-\Sigma_{n=2}^{\infty}|B(n)|} < \infty.$$

This completes the proof. □

Proposition 19: For $\pi/2 \geqslant \alpha > \arcsin(\pi^2/8-1)$, the frequency system $\{Y_{\mathrm{tra}}(\alpha,a;nx)\}$ is an unconditional basis of $L^2_{\mathrm{odd}}[-\pi,+\pi]$.

Proof: It does not matter taking

$$a = \frac{\pi}{4}\,\frac{\alpha}{\sin\alpha}. \tag{138}$$

The Fourier coefficients of $Y_{\mathrm{tra}}(\alpha,(\pi/4)\alpha/\sin\alpha;x)$ is

$$Y_{\mathrm{tra}}\left(\alpha,\frac{\pi}{4}\,\frac{\alpha}{\sin\alpha};x\right) = \sin x+\frac{1}{3^2}\,\frac{\sin 3\alpha}{\sin\alpha}\sin 3x+\frac{1}{5^2}\,\frac{\sin 5\alpha}{\sin\alpha}\sin 5x+\cdots. \tag{139}$$

$$= \sum_{n=1}^{\infty} B(n)\sin nx. \tag{140}$$

Obviously we have $B(1)=1$. Also we have

$$\sum_{n=2}^{\infty} |B(n)| = \frac{1}{\sin\alpha}\left(\frac{|\sin 3\alpha|}{3^2}+\left(\frac{|\sin 5\alpha|}{5^2}\right)+\cdots\right) \leqslant \frac{1}{\sin\alpha}\left(\frac{1}{3^2}+\frac{1}{5^2}+\cdots\right) = \frac{1}{\sin\alpha}\left(\frac{\pi^2}{8}-1\right) < 1.$$

By Proposition (18), we obtain this proposition. Here we have

$$\arcsin\left(\frac{\pi^2}{8}-1\right) \doteq \arcsin(0.233\,70) \doteq 13°31'. \tag{141}$$

VI. A COMBINATIVE FREQUENCY SYSTEM IN $L^2[-\pi,\pi]$

Since the case of the even function subspace $L^2_{\mathrm{even}}[-\pi,\pi]$ is similar to the case of the odd function subspace $L^2_{\mathrm{odd}}[-\pi,+\pi]$, let us return directly to the whole space $L^2[-\pi,\pi]$, and draw our general conclusions.

Suppose that

$$X(x) = \sum_{n=1}^{\infty} A(n)\cos nx \tag{142}$$

is a given function in $L^2_{\mathrm{even}}[-\pi,+\pi]$, and

$$Y(x) = \sum_{n=1}^{\infty} B(n)\sin nx \tag{143}$$

is a given function in $L^2_{\mathrm{odd}}[-\pi,+\pi]$.

Generally, if the frequency system

$$Y(x),Y(2x),Y(3x),\ldots,Y(nx),\ldots \tag{144}$$

is a complete system (an unconditional basis) in $L^2_{\mathrm{odd}}[-\pi,+\pi]$, and the frequency system

$$X(x),X(2x),X(3x),\ldots,X(nx),\ldots \tag{145}$$

is a complete system (an unconditional basis) in $L^2_{\mathrm{even}}[-\pi,+\pi]$ as well, then the combinative frequency system

$$1,X(x),Y(x),X(2x),Y(2x),\ldots,X(nx),Y(nx),\ldots \tag{146}$$

is a complete system (an unconditional basis) in the whole space $L^2[-\pi,\pi]$.

If the combinative frequency system (146) is a complete system in $L^2[-\pi,\pi]$, then its orthonormalized system is an orthonormal basis of $L^2[-\pi,\pi]$, and any function $f(x)\in L^2[-\pi,\pi]$ can be approximated by a linear combination of finite functions in (146) with an arbitrary mean-square error.

If the combinative frequency system (146) is an unconditional basis, then any function $f(x)\in L^2[-\pi,\pi]$ can be expressed as a combinative frequency series based on $X(x)$ and $Y(x)$:

$$f(x)=C_0+\sum_{n=1}^{\infty} C(n)X(nx)+D(n)Y(nx), \tag{147}$$

where the coefficients C_0, $C(n)$, and $D(n)$ can be determined by

$$C_0=\int_{-\pi}^{\pi} f(x)dx, \tag{148}$$

$$C(n)=\int_{-\pi}^{\pi} f(x)h_n(x)dx, \tag{149}$$

$$D(n)=\int_{-\pi}^{\pi} f(x)g_n(x)dx. \tag{150}$$

Here

$$h_n(x)=\sum_{d|n} A^{-1}\left(\frac{n}{d}\right)\cos dx, \tag{151}$$

$$g_n(x)=\sum_{d|n} B^{-1}\left(\frac{n}{d}\right)\sin dx \tag{152}$$

are the biorthogonal functions of $X(nx)$ and $Y(nx)$, respectively.

Speaking concretely, the complete combinative frequency systems of practical importance are (1) the square wave system[2]

$$1,X_{\mathrm{sq}}(x),Y_{\mathrm{sq}}(x),X_{\mathrm{sq}}(2x),Y_{\mathrm{sq}}(2x),\ldots,X_{\mathrm{sq}}(nx),Y_{\mathrm{sq}}(nx),\ldots, \tag{153}$$

(2) the $\pi/4$ three-valued function system

$$1,X_{\pi/4}(x),Y_{\pi/4}(x),X_{\pi/4}(2x),Y_{\pi/4}(2x),\ldots,X_{\pi/4}(nx),Y_{\pi/4}(nx),\ldots, \tag{154}$$

(3) the $\pi/6$ three-valued function system

$$1, X_{\pi/6}(x), Y_{\pi/6}(x), X_{\pi/6}(2x), Y_{\pi/6}(2x), \ldots, X_{\pi/6}(nx), Y_{\pi/6}(nx), \ldots, \tag{155}$$

and so on. Here the functions

$$X_{\mathrm{sq}}(x) = Y_{\mathrm{sq}}\left(x + \frac{\pi}{2}\right), \tag{156}$$

$$X_{\pi/4}(x) = Y_{\pi/4}\left(x + \frac{\pi}{2}\right), \tag{157}$$

$$X_{\pi/6}(x) = Y_{\pi/6}\left(x + \frac{\pi}{2}\right), \tag{158}$$

are called even square wave, $\pi/4$ three-valued function, $\pi/6$ three-valued function, respectively. See Fig. 2.

The unconditional bases of the practical importance are
(1) the triangular wave basis

$$1, X_{\mathrm{tri}}(x), Y_{\mathrm{tri}}(x), X_{\mathrm{tri}}(2x), Y_{\mathrm{tri}}(2x), \ldots, X_{\mathrm{tri}}(nx), Y_{\mathrm{tri}}(nx), \ldots, \tag{159}$$

(2) the $\pi/4$ trapezoidal wave basis

$$1, X_{\pi/4\mathrm{tra}}(x), Y_{\pi/4\mathrm{tra}}(x), X_{\pi/4\mathrm{tra}}(2x), Y_{\pi/4\mathrm{tra}}(2x), \ldots, X_{\pi/4\mathrm{tra}}(nx), Y_{\pi/4\mathrm{tra}}(nx), \ldots, \tag{160}$$

(3) the $\pi/3$ trapezoidal wave basis

$$1, X_{\pi/3\mathrm{tra}}(x), Y_{\pi/3\mathrm{tra}}(x), X_{\pi/3\mathrm{tra}}(2x), Y_{\pi/3\mathrm{tra}}(2x), \ldots, X_{\pi/3\mathrm{tra}}(nx), Y_{\pi/3\mathrm{tra}}(nx), \ldots, \tag{161}$$

and so on.

Here the functions

$$X_{\mathrm{tri}}(x) = Y_{\mathrm{tri}}\left(x + \frac{\pi}{2}\right), \tag{162}$$

$$X_{\pi/4\mathrm{tra}}(x) = Y_{\pi/4\mathrm{tra}}\left(x + \frac{\pi}{2}\right), \tag{163}$$

$$X_{\pi/3\mathrm{tra}}(x) = Y_{\pi/3\mathrm{tra}}\left(x + \frac{\pi}{2}\right), \tag{164}$$

are called even triangular wave, $\pi/4$ trapezoidal wave, and $\pi/3$ trapezoidal wave. See Fig. 3. In fact, for $\alpha > 13°31'$, a general trapezoidal wave system

$$1, X_{\mathrm{tra}}(\alpha, a; x), Y_{\mathrm{tra}}(\alpha, a; x), X_{\mathrm{tra}}(\alpha, a; 2x), Y_{\mathrm{tra}}(\alpha, a; 2x), \ldots X_{\mathrm{tra}}(\alpha, a; nx), Y_{\mathrm{tra}}(\alpha, a; nx), \ldots \tag{165}$$

is an unconditional basis of $L^2[-\pi, \pi]$, where the function

$$X_{\mathrm{tra}}(\alpha, a; x) = Y_{\mathrm{tra}}\left(\alpha, a; x + \frac{\pi}{2}\right) \tag{166}$$

is called even trapezoidal wave.

VII. A FREQUENCY TRANSFORMATION IN $L^2(R)$

In this section, we shall introduce in $L^2(R)$ the frequency transformation based on general periodic functions, which is a generalization of Fourier transformation based on sine–cosine functions.

$L^2(-\infty,+\infty)$ or $L^2(R)$ denotes the space of quadratically integrable functions on $R=(-\infty,+\infty)$. It is well known that $L^2(R)$ can be decomposed into an even function subspace $L^2_{\text{even}}(R)$ and an odd function subspace $L^2_{\text{odd}}(R)$:

$$L^2(R)=L^2_{\text{even}}(R)\oplus L^2_{\text{odd}}(R). \tag{167}$$

By Plancherel's theory,[4] Fourier transformation $\mathbf{F}$ and its inverse transformation $\mathbf{F}^{-1}$ are two bounded linear operators from $L^2(R)$ to itself, i.e., $\mathbf{F}\in B(L^2(R))$, $\mathbf{F}^{-1}\in B(L^2(R))$. $B(L^2(R))$ denotes the Banach space of all the bounded linear operators from $L^2(R)$ to itself.

That is to say, for any two functions $f(x)$, $\hat{f}(\omega)\in L^2(R)$, we have

$$\hat{f}(\omega)=(\mathbf{F}f)(\omega)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}f(x)e^{-i\omega x}\,dx \tag{168}$$

$$\Leftrightarrow$$

$$f(x)=(\mathbf{F}^{-1}\hat{f})(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\hat{f}(\omega)e^{i\omega x}\,d\omega. \tag{169}$$

Strictly speaking, when the integrations in (168) and (169) do not exist, $\int_{-\infty}^{+\infty}$ should be considered to be $\lim_{N\to\infty}\int_{-N}^{N}$, where the limit is in the sence of the norm in $L^2(R)$, see Ref. 4.

In the even function subspace $L^2_{\text{even}}(R)$, Fourier transformation $\mathbf{F}$ and its inversion $\mathbf{F}^{-1}$ become Fourier cosine transformation $\mathbf{F}_{\cos}$ and its inversion $\mathbf{F}_{\cos}^{-1}$. That is to say, for any two functions $f_{\text{even}}(x)$, $a(\omega)\in L^2_{\text{even}}(R)$, we have

$$a(\omega)=(\mathbf{F}_{\cos}f_{\text{even}})(\omega)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}f_{\text{even}}(x)\cos(\omega x)dx \tag{170}$$

$$\Leftrightarrow$$

$$f_{\text{even}}(x)=(\mathbf{F}_{\cos}^{-1}a)(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}a(\omega)\cos(\omega x)d\omega. \tag{171}$$

In the odd function subspace $L^2_{\text{odd}}(R)$, there exist similar Fourier sine transformation and its inversion.

Next let us introduce the transformation based on general periodic functions step by step.

In this section we suppose that $X(x)\in L^2[-\pi,\pi]$ and $Y(x)\in L^2[-\pi,\pi]$ are a given even function and a given odd function with period 2π. Their Fourier series are

$$X(x)=\sum_{n=1}^{\infty}A(n)\cos nx,\qquad Y(x)=\sum_{n=1}^{\infty}B(n)\sin nx.$$

Their Fourier coefficients satisfy

$$\sum_{n=1}^{\infty} |A(n)| \frac{1}{\sqrt{n}} < \infty, \quad \sum_{n=1}^{\infty} |A^{-1}(n)| \frac{1}{\sqrt{n}} < \infty,$$

$$\sum_{n=1}^{\infty} |B(n)| \frac{1}{\sqrt{n}} < \infty, \quad \sum_{n=1}^{\infty} |B^{-1}(n)| \frac{1}{\sqrt{n}} < \infty.$$

Definition: The functions $\tilde{X}(x)$ and $\tilde{Y}(x)$ are called the dual functions of $X(x)$ and $Y(x)$, respectively, which are defined on $(-\infty,\infty)$ by

$$\tilde{X}(x) = \frac{1}{2\pi} \sum_{n=1}^{\infty} A^{-1}(n) \frac{1}{n} \cos\left(\frac{x}{n}\right), \tag{172}$$

$$\tilde{Y}(x) = \frac{1}{2\pi} \sum_{n=1}^{\infty} B^{-1}(n) \frac{1}{n} \sin\left(\frac{x}{n}\right). \tag{173}$$

Proposition 20: The functions $\tilde{X}(x)$ and $\tilde{Y}(x)$ are bounded and infinitely differentiable in $(-\infty,+\infty)$.

Proof: We have

$$|\tilde{X}(x)| \leqslant \frac{1}{2\pi} \sum_{n=1}^{\infty} \left| A^{-1}(n) \frac{1}{n} \cos\left(\frac{x}{n}\right) \right| \tag{174}$$

$$\leqslant \frac{1}{2\pi} \sum_{n=1}^{\infty} |A^{-1}(n)| \frac{1}{n}. \tag{175}$$

Therefore the function $\tilde{X}(x)$ is bounded. By Weierstrass M-test, the series (172) converges absolutely and uniformly in $(-\infty,\infty)$. By properties of a uniformly convergent series, the function $\tilde{X}(x)$ is infinitely differentiable.

The case of the function $\tilde{Y}(x)$ is similar.

Definition: The operator $\mathbf{T}(n): L^2(R) \to L^2(R)$ is defined by

$$\mathbf{T}(n) f(x) = f\left(\frac{x}{n}\right), \tag{176}$$

where $f(x) \in L^2(R)$ and n is any positive integer.

Proposition 21:

(1) The operator $\mathbf{T}(n)$ is a bounded linear operator, i.e., $\mathbf{T}(n) \in B(L^2(R))$, and its norm $\|\mathbf{T}(n)\| = \sqrt{n}$.

(2) For any two positive integers m and n, we have

$$\mathbf{T}(m)\mathbf{T}(n) = \mathbf{T}(mn), \tag{177}$$

that is to say $\mathbf{T}(n)$ is completely multiplicative.

Proof: This proposition is clear from the definition. □

Proposition 22: The operator series

$$\mathbf{S} = \sum_{n=1}^{\infty} A(n) \frac{1}{n} \mathbf{T}(n), \tag{178}$$

$$\mathbf{S}^{-1} = \sum_{n=1}^{\infty} A^{-1}(n) \frac{1}{n} \mathbf{T}(n) \tag{179}$$

converge absolutely in the sense of the norm in $B(L^2(R))$, and are mutually inverse.

Proof: Because

$$\sum_{n=1}^{\infty}\left\|A(n)\frac{1}{n}\mathbf{T}(n)\right\|\leqslant\sum_{n=1}^{\infty}|A(n)|\frac{1}{n}\|\mathbf{T}(n)\|=\sum_{n=1}^{\infty}|A(n)|\frac{1}{\sqrt{n}}<\infty, \tag{180}$$

$$\sum_{n=1}^{\infty}\left\|A^{-1}(n)\frac{1}{n}\mathbf{T}(n)\right\|\leqslant\sum_{n=1}^{\infty}|A^{-1}(n)|\frac{1}{n}\|\mathbf{T}(n)\|=\sum_{n=1}^{\infty}|A^{-1}(n)|\frac{1}{\sqrt{n}}<\infty, \tag{181}$$

the two operator series converge absolutely in the sense of the norm in $B(L^2(R)$. Thus $\mathbf{S}$ and $\mathbf{S}^{-1}$ are two bounded linear operators defined on $L^2(R)$, i.e., $\mathbf{S}\in B(L^2(R))$, $\mathbf{S}^{-1}\in B(L^2(R))$.

Furthermore, we have

$$\mathbf{S}\mathbf{S}^{-1}=\mathbf{S}^{-1}\mathbf{S} \tag{182}$$

$$=\sum_{n=1}^{\infty}A(n)\frac{1}{n}\mathbf{T}(n)\sum_{m=1}^{\infty}A^{-1}(m)\frac{1}{m}\mathbf{T}(m) \tag{183}$$

$$=\sum_{k=1}^{\infty}\frac{1}{k}\mathbf{T}(k)\sum_{n|k}A(n)A^{-1}\left(\frac{k}{n}\right) \tag{184}$$

$$=\mathbf{T}(1)=\mathbf{I}, \tag{185}$$

where $\mathbf{I}$ is the identical operator. In other words, the operators $\mathbf{S}$ and $\mathbf{S}^{-1}$ are mutually inverse. □

Proposition 23: For any two functions $f_{\mathrm{even}}(x)$, $C(\omega)\in L^2_{\mathrm{even}}(R)$, we have

$$C(\omega)=(\mathbf{W}_{\mathrm{even}}f_{\mathrm{even}})(\omega)=\int_{-\infty}^{+\infty}f_{\mathrm{even}}(x)\tilde{X}(\omega x)\,dx \tag{186}$$

$$\Leftrightarrow$$

$$f_{\mathrm{even}}(x)=(\mathbf{W}_{\mathrm{even}}^{-1}C)(x)=\int_{-\infty}^{+\infty}C(\omega)X(\omega x)\,d\omega. \tag{187}$$

Here $\mathbf{W}_{\mathrm{even}}:L^2_{\mathrm{even}}(R)\to L^2_{\mathrm{even}}(R)$ is called the transformation based on the even period function $X(x)$, and $\mathbf{W}_{\mathrm{even}}^{-1}:L^2_{\mathrm{even}}(R)\to L^2_{\mathrm{even}}(R)$ is its inverse.

Proof: Let us consider the case ⇒.

First we have

$$\begin{aligned}C(\omega)&=(\mathbf{W}_{\mathrm{even}}f_{\mathrm{even}})(\omega)\\&=\int_{-\infty}^{+\infty}f_{\mathrm{even}}(x)\tilde{X}(\omega x)\,dx\\&=\frac{1}{2\pi}\int_{-\infty}^{+\infty}f_{\mathrm{even}}(x)\sum_{n=1}^{\infty}A^{-1}(n)\frac{1}{n}\cos\left(\frac{x\omega}{n}\right)dx,\\&=\frac{1}{2\pi}\sum_{n=1}^{\infty}A^{-1}(n)\frac{1}{n}\int_{-\infty}^{+\infty}f_{\mathrm{even}}(x)\cos\left(\frac{x\omega}{n}\right)dx\\&=\frac{1}{2\pi}\sum_{n=1}^{\infty}A^{-1}(n)\frac{1}{n}\mathbf{T}_n\int_{-\infty}^{+\infty}f_{\mathrm{even}}(x)\cos(x\omega)\,dx\\&=\frac{1}{\sqrt{2\pi}}\sum_{n=1}^{\infty}A^{-1}(n)\frac{1}{n}\mathbf{T}_n\mathbf{F}_{\cos}f_{\mathrm{even}}(x)=\frac{1}{\sqrt{2\pi}}\mathbf{S}^{-1}\mathbf{F}_{\cos}f_{\mathrm{even}}(x).\end{aligned}$$

In one word, in $L^2_{\text{even}}(R)$ the operator $\mathbf{W}_{\text{even}}$ can be decomposed into two operators:

$$\mathbf{W}_{\text{even}}=\frac{1}{\sqrt{2\pi}}\mathbf{S}^{-1}\mathbf{F}_{\cos}.$$

Since both operators $\mathbf{S}^{-1}$ and $\mathbf{F}_{\cos}$ have inverse operators, the operator $\mathbf{W}_{\text{even}}$ has inverse operator as well:

$$\mathbf{W}^{-1}_{\text{even}}=\sqrt{2\pi}\mathbf{F}^{-1}_{\cos}\mathbf{S}. \tag{188}$$

Since all the operators $\mathbf{S}$, $\mathbf{S}^{-1}$, $\mathbf{F}_{\cos}$, and $\mathbf{F}^{-1}_{\cos}$ are bounded linear operators, so are $\mathbf{W}_{\text{even}}$ and $\mathbf{W}^{-1}_{\text{even}}$, i.e.,

$$\mathbf{W}_{\text{even}},\mathbf{W}^{-1}_{\text{even}}\in B(L^2_{\text{even}}(R)). \tag{189}$$

Therefore we have

$$\begin{aligned}
f_{\text{even}}(x)&=\mathbf{W}^{-1}_{\text{even}}C(\omega)\\
&=\sqrt{2\pi}\mathbf{F}^{-1}_{\cos}\mathbf{S}C(\omega)\\
&=\sqrt{2\pi}\mathbf{F}^{-1}_{\cos}\sum_{n=1}^{\infty}A(n)\frac{1}{n}\mathbf{T}(n)C(\omega)\\
&=\sqrt{2\pi}\mathbf{F}^{-1}_{\cos}\sum_{n=1}^{\infty}A(n)\frac{1}{n}C\left(\frac{\omega}{n}\right)\\
&=\int_{-\infty}^{\infty}\cos(\omega x)\sum_{n=1}^{\infty}A(n)\frac{1}{n}C\left(\frac{\omega}{n}\right)d\omega\\
&=\sum_{n=1}^{\infty}A(n)\frac{1}{n}\int_{-\infty}^{\infty}\cos(\omega x)C\left(\frac{\omega}{n}\right)d\omega\\
&=\sum_{n=1}^{\infty}A(n)\int_{-\infty}^{\infty}\cos(n\omega x)C(\omega)d\omega\\
&=\int_{-\infty}^{\infty}C(\omega)\sum_{n=1}^{\infty}A(n)\cos(n\omega x)d\omega\\
&=\int_{-\infty}^{\infty}C(\omega)X(\omega x)d\omega.
\end{aligned}$$

The case $\Leftarrow$ is similar. This completes the proof.

Similarly, we have the following two propositions.

Proposition 24: The operator series

$$\mathbf{R}=\sum_{n=1}^{\infty}B(n)\frac{1}{n}\mathbf{T}(n), \tag{190}$$

$$\mathbf{R}^{-1}=\sum_{n=1}^{\infty}B^{-1}(n)\frac{1}{n}\mathbf{T}(n) \tag{191}$$

converge absolutely in the sense of the norm in $B(L^2(R))$, and are mutually inverse.

Proposition 25: For any two functions $f_{\text{odd}}(x)$, $D(\omega)\in L^2_{\text{odd}}(R)$, we have

$$D(\omega)=(\mathbf{W}_{\text{odd}}f_{\text{odd}})(\omega)=\int_{-\infty}^{+\infty}f_{\text{odd}}(x)\tilde{Y}(\omega x)dx \tag{192}$$

$$\Leftrightarrow$$

$$f_{\text{odd}}(x)=(\mathbf{W}_{\text{odd}}^{-1}D)(x)=\int_{-\infty}^{+\infty}D(\omega)Y(\omega x)d\omega. \tag{193}$$

where $\mathbf{W}_{\text{odd}}:L^2_{\text{odd}}(R)\to L^2_{\text{odd}}(R)$ denotes the transformation based on odd periodic function $Y(x)$, and $\mathbf{W}_{\text{odd}}^{-1}:L^2_{\text{odd}}(R)\to L^2_{\text{odd}}(R)$ denotes its inverse.

Proposition 26: For any function $f(x)\in L^2(-\infty,+\infty)$, we have

$$f(x)=\int_{-\infty}^{\infty}C(\omega)X(\omega x)+D(\omega)Y(\omega x)d\omega, \tag{194}$$

where

$$C(\omega)=\int_{-\infty}^{\infty}f(x)\tilde{X}(\omega x)dx, \tag{195}$$

$$D(\omega)=\int_{-\infty}^{\infty}f(x)\tilde{Y}(\omega x)dx. \tag{196}$$

Proof: Combining Propositions 23 and 25, we obtain this one easily.

Definition: The two functions $W(x)$ and $\tilde{W}(x)$ are defined on $(-\infty,+\infty)$ by

$$W(x)=X(x)+Y(x), \tag{197}$$

$$\tilde{W}(x)=\tilde{X}(x)+\tilde{Y}(x), \tag{198}$$

Proposition 27: For any two functions $f(x)$, $f'(\omega)\in L^2(-\infty,+\infty)$, we have

$$f'(\omega)=\int_{-\infty}^{+\infty}f(x)\tilde{W}(\omega x)dx \tag{199}$$

$$\Leftrightarrow$$

$$f(x)=\int_{-\infty}^{+\infty}f'(\omega)W(\omega x)d\omega. \tag{200}$$

Proof: Combining Propositions 23 and 25, we obtain this one easily. By the way, this transformation is a generalization of Hartley transformation.[5]

Definition: The two complex-valued functions $Z(x)$ and $\tilde{Z}(x)$ are defined on $(-\infty,+\infty)$ by

$$Z(x)=X(x)+iY(x), \tag{201}$$

$$\tilde{Z}(x)=\tilde{X}-i\tilde{Y}. \tag{202}$$

Proposition 28: For any two functions $f(x)$, $\overline{f(\omega)}\in L^2(-\infty,+\infty)$, we have

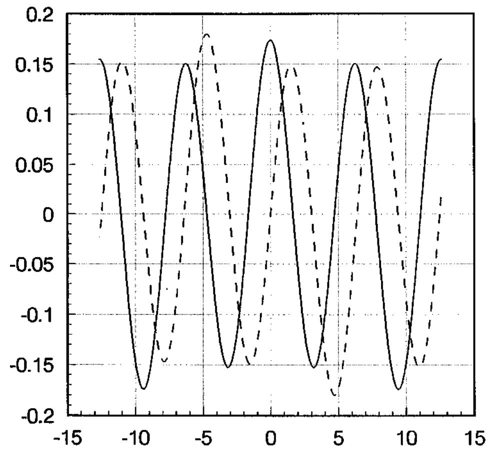

FIG. 4. The dual functions of square waves, $\tilde{X}_{sq}(x)$ and $\tilde{Y}_{sq}(x)$(---).

$$\overline{f(\omega)}=(\mathbf{W}f)(\omega)=\int_{-\infty}^{+\infty}f(x)\tilde{Z}(\omega x)dx \tag{203}$$

$\Leftrightarrow$

$$f(x)=(\mathbf{W}^{-1}\overline{f})(x)=\int_{-\infty}^{+\infty}\overline{f}(\omega)Z(\omega x)d\omega. \tag{204}$$

$\mathbf{W}:L^2(R)\rightarrow L^2(R)$ is called the transformation based on the periodic functions $X(x)$ and $Y(x)$, and $\mathbf{W}^{-1}:L^2(R)\rightarrow L^2(R)$ is its inversion.

Proof: Combining Propositions 23 and 25, we obtain this one easily.

The relation between the transform based on $X(x)$ and $Y(x)$ and Fourier transform of the same function $f(x)\in L^2(R)$ is

$$\overline{f}=C(\omega)-iD(\omega) \tag{205}$$

$$=\frac{1}{\sqrt{2\pi}}\mathbf{S}^{-1}a(\omega)-i\mathbf{R}^{-1}b(\omega) \tag{206}$$

$$=\frac{1}{\sqrt{2\pi}}\sum_{n=1}^{\infty}\left[A(n)^{-1}\frac{1}{n}a\left(\frac{\omega}{n}\right)-iB(n)^{-1}\frac{1}{n}b\left(\frac{\omega}{n}\right)\right], \tag{207}$$

where the even function $a(\omega)$ and the odd function $b(\omega)$ are defined by $f(x)$'s Fourier transform

$$\hat{f}(\omega)=a(\omega)-ib(\omega). \tag{208}$$

The dual functions of square waves, triangular waves, and $\pi/4$ trapezoidal waves are shown in Figs. 4, 5, 6, respectively.

VIII. CONCLUSIONS AND QUESTIONS

From the above discussion, under certain conditions we conclude that a function with period 2π may be expressed as a frequency series of general periodic functions, and that a nonperiodic

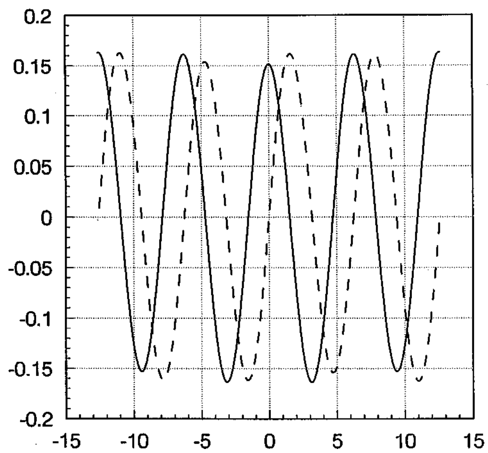

FIG. 5. The dual functions of triangular waves, $\tilde{X}_{\mathrm{tri}}(x)$ and $\tilde{Y}_{\mathrm{tri}}(x)$(---).

function can be expressed as a frequency integral of general periodic functions. This is a new and practical generalization of classical Fourier analysis based on sine–cosine functions.[6] In electronics, we can say in many cases that a signal can be considered as a superposition of easily generated functions with different frequencies. The results in this paper make it possible to represent a signal by use of square waves, triangular waves, and trapezoidal waves. This forms a theoretical foundation for the technique of easily generated function analysis in modern electronics.

However, regarding the frequency analysis based on general periodic functions, there is a lot of work to do.[7] The central questions are:

(1) What is the sufficient and necessary condition for a combinative frequency system $1,\{X(nx),Y(nx)\}_{n=1}^{\infty}$ to be

(1) a complete system;

(2) a basis;

(3) an unconditional basis;

in $L^2[-\pi,\pi]$?

(2) What is the sufficient and necessary condition for a function in $L^2(R)$ to be expressed as a frequency integral (see (194))?

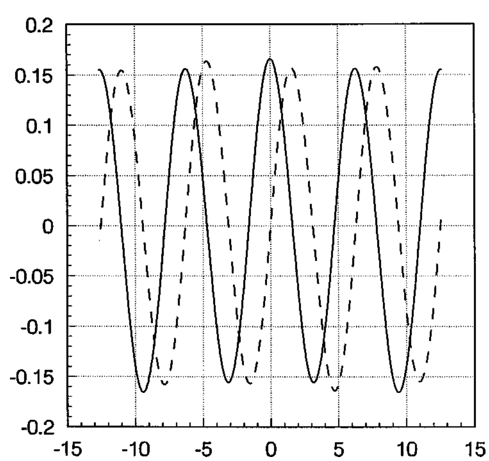

FIG. 6. The dual functions of $\pi/4$ trapezoidal waves, $\tilde{X}_{\pi/4\mathrm{tra}}(x)$ and $\tilde{Y}_{\pi/4\mathrm{tra}}(x)$(---).

Furthermore what about $L^p[-\pi,+\pi]$ and $L^p(R)$ (such as $p=1$) instead of $L^2[-\pi,+\pi]$ and $L^2(R)$?

Obviously these questions are quite challenging ones for mathematicians.

ACKNOWLEDGMENTS

The author thanks Professor Q. Zhang and Professor Z. Liu. Also the author thanks Hui Li at Yale University, and Wenxiang Cong at the Chinese Academy for their help. This work is supported partly by the Chinese National Nature Science Foundation and the Chinese Postdoctoral Science Foundation.

[1] T. M. Apostol, *Introduction to Analytic Number Theory* (Springer, New York, 1976), Chap. 2.
[2] Y. Wei and N. Chen, ''Square Wave Analysis,'' J. Math. Phys. **39**, 4226–4245 (1998).
[3] I. N. Bronshtein and K. A. Semendyayev, *Handbook of Mathematics*, English translation, 3rd ed. (Van Nostrand, New York, 1985), pp. 574–578.
[4] C. Chui, *An Introduction to Wavelets* (Academic, New York, 1992), Chap. 2.
[5] R. N. Bracewell, ''Discrete Hartley transform,'' J. Opt. Soc. Am. **73**, 1832–1835 (1983).
[6] B. Torresani, ''Introduction to the special issue on wavelet and time frequency analysis,'' J. Math. Phys. **39**, 3952–3953 (1998).
[7] Y. Wei, Doctor thesis, ''Easily-generated Function Analysis,'' Beijing University of Science and Technology, 1998.

Comment on ''Generalized W_∞ symmetry algebra of the conditionally integrable nonlinear evolution equation'' [J. Math. Phys. 36, 3492 (1995)]

Wen-Xiu Ma[a)]
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, People's Republic of China

(Received 5 August 1997; accepted for publication 15 December 1998)

Two remarks on an inverse operator of a differential operator ∂_x and on symmetries of two kinds of differential equations $\Delta(u)=0$ and $\partial_x\Delta(u)=0$ are pointed out and then a few of remarks about the paper ''Generalized W_∞ symmetry algebra of the conditionally integrable nonlinear evolution equation'' [Y. Lou and J. P. Weng, J. Math. Phys. **36**, 3492 (1995)] are presented. It is in particular shown that if we consider ∂_x^{-1} to be a linear and right inverse operator of ∂_x, the vector fields $\sigma_n^u(f)$ proposed in the above paper are not certain to be symmetries of the Jimbo–Miwa–Kadomtsev–Petviashvili system under consideration. Moreover the commutators of a generalized W_∞ algebra established among these vector fields do not always hold. Therefore the vector fields $\sigma_n^u(f)$ do not provide an example of application of the formal series ansatz in the above paper. © *1999 American Institute of Physics.* [S0022-2488(99)03606-3]

I. INTRODUCTION

It is interesting to construct a kind of symmetries involving an arbitrary function of certain independent variable for nonlinear integrable equations. They often constitute a class of generalized W_∞ algebra

$$[\sigma_m^u(f_1),\sigma_n^u(f_2)]=\alpha\sigma_{m+n-l}^u((m+\beta)f_1f_{2t}-(n+\beta)f_{1t}f_2), \tag{1}$$

where l is a natural number, α, β are two arbitrary constants, and we assume that an involved specific independent variable is the time variable t. Some typical examples have been considered, including the Kadomtsev–Petviashvili equation,[1] the Davey–Stewartson equation,[2] the Hirota's bilinear equations,[3] and three-wave resonant interaction system,[4] etc.

Recently Lou and Weng considered the above-mentioned generalized W_∞ symmetry algebras for conditionally integrable nonlinear equations.[5] They took the Jimbo–Miwa–Kadomtsev–Petviashvili (JMKP) system of equations (see Ref. 6 for more information about the JMKP system) as an illustrative example. This system is a combination of the so-called Jimbo–Miwa equation and the Kadomtsev–Petviashvili equation:

$$\begin{aligned}K_0(u)&=u_{xxxy}+3(uu_y)_x+3u_{xx}\partial_x^{-1}u_y+3u_xu_y+2u_{yt}-3u_{xz}=0,\\ K_1(u)&=(u_{xxx}+6uu_x)_x+3u_{yy}-4u_{xt}=0.\end{aligned} \tag{2}$$

For the JMKP system (2), Lou and Weng[5] proposed a hierarchy of symmetries $\sigma_n^u(f)$ involving an arbitrary function $f=f(z)$ defined by:

$$\sigma_n^u(f)=\partial_x\sigma_n^w(f)|_{u=w_x},\quad n\geqslant 0, \tag{3}$$

[a)]Electronic mail: mawx@math.cityu.edu.hk

$$\sigma_n^w(f)=\frac{A_n}{n!}\sum_{k=0}^{n+1} f^{(n+1-k)}(z)\big(\tfrac{1}{3}\partial_x^2\partial_y+\partial_x^{-1}w_x\partial_x\partial_y+\partial_x^{-1}w_{xy}\partial_x+\partial_x^{-1}w_{xx}\partial_y$$
$$+\partial_x^{-1}w_y\partial_x^2+\tfrac{2}{3}\partial_x^{-1}\partial_y\partial_t-\partial_z\big)^k(yt^n)\big|_{H_0=0,H_1=0},\quad n\geqslant 0, \tag{4}$$

where A_n are constants and $f^{(i)}(z)=(\partial_z)^i f(z),\ i\geqslant 0$, and they pointed out that these symmetries constituted a generalized W_∞ algebra:

$$[\sigma_m^u(f_1),\sigma_n^u(f_2)]=\tfrac{1}{4}\sigma_{m+n-3}^u((m+1)f_1\dot{f}_2-(n+1)f_2\dot{f}_1), \tag{5}$$

where $\dot{f}_i=\partial_z f_i,\ i=1,2$. The vector fields $\sigma_n^w(f)$ defined by (4) are also regarded as symmetries of the following potential JMKP system

$$H_0=H_0(w)=w_{xz}-\tfrac{1}{3}w_{xxxy}-w_{xy}w_x-w_yw_{xx}-\tfrac{2}{3}w_{yt}=0,$$
$$H_1=H_1(w)=w_{xxxx}+6w_xw_{xx}+3w_{yy}-4w_{xt}=0, \tag{6}$$

which is a starting point to construct the symmetries $\sigma_n^u(f)$ of the JMKP system (2).

In this comment, we intend to provide two remarks on an inverse operator of a differential operator ∂_x and on relations between symmetries of two kinds of differential equations $\Delta(u)=0$ and $\partial_x\Delta(u)=0$, and then a couple of remarks about Lou and Weng's work in Ref. 5 It is shown that if we consider the inverse operator ∂_x^{-1} to be a linear operator and a right inverse operator of ∂_x, the vector fields $\sigma_n^u(f)$ and $\sigma_n^w(f)$ defined by (3) and (4) are not certain to be symmetries of the JMKP system (2) and the potential JMKP system (6), respectively, whatever boundary condition is imposed on the potential u. Moreover it is verified that the symmetry algebra defined by (5) does not always hold, and that a symmetry relation

$$\sigma^u(f)=\partial_x\sigma^w(f)\big|_{u=w_x} \tag{7}$$

is not always correct between the systems (2) and (6), although the system (6) may be generated from the system (2) after setting $u=w_x$. Only additional natural conditions $\partial_y\partial_x^{-1}=\partial_x^{-1}\partial_y$, $\partial_t\partial_x^{-1}=\partial_x^{-1}\partial_t$ and $\partial_z\partial_x^{-1}=\partial_x^{-1}\partial_z$ are needed in proving our above statements. Therefore the vector fields $\sigma_n^u(f)$ do not provide an example of application of the formal series ansatz in Ref. 5.

II. TWO OBSERVATIONS

Observation 1: Let us first make a remark on definitions of an inverse operator of ∂_x, where x is a coordinate of $\mathbb{R}^n$. Since we have a nonempty kernel of ∂_x, we cannot make any definition of an inverse operator ∂_x^{-1} such that the inverse operator ∂_x^{-1} maps $C^\infty(\mathbb{R}^n)$ into $C^\infty(\mathbb{R}^n)$ and satisfies the conditions of $\partial_x^{-1}\partial_x=\partial_x\partial_x^{-1}=1$, the most natural left and right inverse conditions on an inverse operator ∂_x^{-1}. However we can take a quotient space of the space $C^\infty(\mathbb{R}^n)$ under $\ker(\partial_x)$ as a new domain of definition of ∂_x. If we view the differential operator ∂_x as a mapping from the quotient space to the same quotient space, $\partial_x:[f]\mapsto[\partial_x f]$, where

$$[f]=\{g\in C^\infty(\mathbb{R}^n)\,|\,\partial_x g=\partial_x f\},$$

the operator ∂_x is still not injective and thus an inverse operator of ∂_x will not be well defined over the whole quotient space. In order to obtain one-to-one correspondence of ∂_x, we can make the following:

$$\partial_x:C^\infty(\mathbb{R}^n)/\ker(\partial_x)\to C^\infty(\mathbb{R}^n),\quad [f]\mapsto\partial_x f; \tag{8}$$

$$\partial_x^{-1}: C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n)/\ker(\partial_x), \quad f \mapsto \left[\int_0^x f(\cdots, x', \cdots) dx' \right]. \tag{9}$$

This definition satisfies the left and right inverse conditions: $\partial_x^{-1}\partial_x = \partial_x\partial_x^{-1} = 1$. But over the quotient space, we cannot keep the same differential operation ∂_x: $\partial_x(fg) = \partial_x(f)g + f\partial_x(g)$. This is to say, the equality

$$\partial_x([f][g]) = (\partial_x f)[g] + [f](\partial_x g) = \partial_x[h],$$

where $[f][g] = [h]$ is a multiplication on $C^\infty(\mathbb{R}^n)/\ker(\partial_x)$, will make no sense. There is an unavoidable problem: how to define $[f][g] = [h]$, $f[g]$ and $[f]g$ such that $(\partial_x f)[g] + [f](\partial_x g)$ is the same function as $\partial_x h (= \partial_x[h])$.

Nevertheless we can define a left and right inverse operator of ∂_x from some subspace of $C^\infty(\mathbb{R}^n)$ to the same subspace, or a right inverse operator of ∂_x from the whole space of smooth functions to the same whole space of smooth functions, and the differential operation ∂_x does not need to be changed over function spaces. In Sec. III, we use the latter definition to explain some problems because the former definition is not suitable for our discussion.

Observation 2: Let us secondly make a remark on symmetries of two kinds of systems of differential equations. Differentiating a given system of differential equations $\Delta(u, u_x, \ldots) = 0$ with respect to x leads to the system of differential equations $\partial_x\Delta(u, u_x, \ldots) = 0$. We call it a derivative system of the original system. For their symmetry algebras, we can have the following result.

Two symmetry algebras of the systems of differential equations $\Delta(u, u_x, \ldots) = 0$ *and* $\partial_x\Delta(u, u_x, \cdots) = 0$ *are not certain to be the same.*

Let us take a scalar equation $u_t = u_x$ as an illustrative example to show the above statement. This moment, the derivative equation reads as $u_{tx} = u_{xx}$. We choose a symmetry $\sigma(u) = u^2$ of the equation $u_t = u_x$. A direct computation gives rise to

$$\sigma_{tx} - \sigma_{xx} = (u^2)_{tx} - (u^2)_{xx} = 2u(u_{tx} - u_{xx}) + 2u_x(u_t - u_x).$$

This does not equal to zero if we choose a solution $u(x,t) = e^{x+t} + e^t$ to the derivative equation $u_{tx} = u_{xx}$. Therefore u^2 is not a symmetry of the derivative equation $u_{tx} = u_{xx}$.

On the other hand, we choose a symmetry $\sigma(u) = u_x + e^t$ of the derivative equation $u_{tx} = u_{xx}$. But the $\sigma(u) = u_x + e^t$ does not satisfy the linearized equation of the equation $u_t = u_x$ and thus it is not a symmetry of the equation $u_t = u_x$. Therefore it follows that the above statement is true.

III. COUNTEREXAMPLES

Let us now turn to discussion on the results in Lou and Weng's work of Ref. 5. We claim the conditions on ∂_x^{-1}:

$$\begin{gathered} \partial_x^{-1}: C^\infty(\mathbb{R}^4) \to C^\infty(\mathbb{R}^4) \ \text{being linear}, \ \partial_x\partial_x^{-1} = 1, \\ \partial_y\partial_x^{-1} = \partial_x^{-1}\partial_y, \partial_t\partial_x^{-1} = \partial_x^{-1}\partial_t, \ \partial_z\partial_x^{-1} = \partial_x^{-1}\partial_z, \end{gathered} \tag{10}$$

which may be achieved, for example, by defining

$$(\partial_x^{-1} f)(x,y,t,z) = \int_0^x f(x', y, t, z) dx'. \tag{11}$$

Note that we have $\partial_x^{-1} 0 = 0$, since ∂_x^{-1} is a linear mapping.

What we would like to show now is that under the definition of (10) for ∂_x^{-1}, both the vector fields given in Ref. 5,

$$\sigma_0^u(f)=fu_x \quad \text{with} \quad \dot{f}=\partial_z f\neq 0, \tag{12}$$

$$\sigma_u^4(1)=2u_{xxt}+4\partial_x^{-1}u_{tt}+3u_{xyy}+3\partial_x^{-1}u_{yz}+6u_x\partial_x^{-1}u_t+6uu_t \\ +12u_y\partial_x^{-1}u_y+3u_x\partial_x^{-2}u_{yy}+3u\partial_x^{-1}u_{yy}, \tag{13}$$

are not symmetries of the JMKP system (2). Actually from the JMKP system (2) itself, we easily find its linearized system:

$$K_0'(u)[\sigma]:=\sigma_{xxxy}+3(u\sigma_y+\sigma u_y)_x+3\sigma_{xx}\partial_x^{-1}u_y+3u_{xx}\partial_x^{-1}\sigma_y+3\sigma_x u_y+3u_x\sigma_y+2\sigma_{yt}-3\sigma_{xz}=0, \\ K_1'(u)[\sigma]:=(\sigma_{xxx}+6u\sigma_x+6\sigma u_x)_x+3\sigma_{yy}-4\sigma_{xt}=0. \tag{14}$$

The substitution of the vector field (12) into the expression of $K_0'(u)[\sigma]$ in the first equation of (14) leads to

$$K_0'(u)[\sigma_0^u(f)]=f\partial_x K_0(u)+3fu_{xx}\partial_x^{-1}u_{xy}-3fu_{xx}u_y-3\dot{f}u_{xx}. \tag{15}$$

Choose a solution $v(x,t)=2\ \mathrm{sech}^2(x+t)$ to the Korteweg–de Vries equation

$$v_t=\tfrac{1}{4}v_{xxx}+\tfrac{6}{4}vv_x. \tag{16}$$

This moment, $v_{xx}\neq 0$ and v_{xx} has and only has two zero points. The function $u(x,y,t,z)=v(x,t)$ is a solution to the JMKP system (2), and hence from (15) we have

$$(K_0'(u)[\sigma_0^u(f)])|_{u=v}=-3\dot{f}v_{xx}\neq 0, \tag{17}$$

noticing that $\partial_x^{-1}0=0$. This implies that $\sigma_0^u(f)$ is not a symmetry while $\dot{f}\neq 0$.

Let us now choose a solution $u(x,y,t,z)=g(t,z)$ where g is an arbitrary function satisfying $\dot{g}_{tt}=\partial_z g_{tt}\neq 0$. This moment, we have

$$\begin{aligned}(K_0'(u)[\sigma_4^u(1)])|_{u=g}&=(K_0'(u)[4\partial_x^{-1}g_{tt}+6gg_t])|_{u=g}\\&=(2\partial_y\partial_t-3\partial_x\partial_z)(4\partial_x^{-1}g_{tt}+6gg_t)\\&=8\partial_y\partial_t\partial_x^{-1}g_{tt}-12\dot{g}_{tt}\\&=8\partial_t\partial_x^{-1}\partial_y g_{tt}-12\dot{g}_{tt}=-12\dot{g}_{tt}\neq 0,\end{aligned} \tag{18}$$

which shows that $\sigma_4^u(1)$ is not a symmetry of the JMKP system (2), either.

Moreover, we can show that the vector field given in Ref. 5,

$$\sigma_4^w(1)=2w_{xxt}+4\partial_x^{-1}w_{tt}+3w_{xyy}+3\partial_x^{-1}w_{yz}+6w_xw_t+6w_y^2+3w_x\partial_x^{-1}w_{yy},$$

is itself not a symmetry of the potential JMKP system (6) at all. This is obvious by observing that

$$\sigma_4^w(1)|_{w=g(t,z)}=4\partial_x^{-1}g_{tt}, \quad (H_0'(w)[\sigma_4^w(1)])|_{w=g(t,z)}=4\dot{g}_{tt}, \tag{19}$$

where an arbitrary function $g=g(t,z)$ is always a solution to the potential JMKP system (6). It also implies that the formal series does not truncate in this case.

We point out that the explicit expressions for $\sigma_n^u(f)$ and $\sigma_n^w(f)$ for $0\leqslant n\leqslant 4$ in Ref. 5 may not be generated from the general formulas (3) and (4), if we view the involved inverse operator ∂_x^{-1} as a normal inverse operator, for example, as in (11). Nevertheless they satisfy the recursion relation [see (11) in Ref. 5] in the formal series ansatz, which is one of two conditions for a formal series to be a symmetry (the other is that the series must truncate). This implies that they can be generated from the formulas (3) and (4) but we have to choose a special integration constant for ∂_x^{-1} each time.

Next, we want to verify that the commutator relation (5) is not always correct for $m+n\geqslant 3$, besides its being incomplete due to making no sense for $m+n<3$. For example, we can work out

$$[\sigma_0^u(f),\sigma_4^u(1)]|_{u=g(t,z)}=0 \tag{20}$$

and can further find that

$$[\sigma_0^u(f),\sigma_4^u(1)]\neq \tfrac{1}{4}\sigma_1^u(-5\dot{f}), \tag{21}$$

in virtue of the quantities

$$\tfrac{1}{4}\sigma_1^u(-5\dot{f})=-\tfrac{5}{4}\dot{f}u_y-\tfrac{15}{16}\ddot{f}(tu_x+\tfrac{2}{3}),\quad \tfrac{1}{4}\sigma_1^u(-5\dot{f})|_{u=g(t,z)}=-\tfrac{5}{8}\ddot{f},$$

where $\ddot{f}=(\partial_z)^2 f$. Hence two sides of (21) cannot be balanced. This implies that our statement is true.

In addition, the vector field

$$\sigma_0^w(f)=fw_x+y\dot{f}$$

is always a symmetry of (6) for any f. Therefore we see that the relation (7), i.e., $\sigma^u(f)=\partial_x\sigma^w(f)|_{u=w_x}$, between symmetries of the JMKP system (2) and the potential JMKP system (6) is not appropriate. The reason is that the system (6) is generated from the system (2) through not only a Miura transformation $u=w_x$, but also an integration of x and an application of the left inverse condition $\partial_x^{-1}\partial_x=1$ (which actually could not be imposed). Let us now recall the second observation in Sec. II and the nonexistence of a left and right inverse linear operator ∂_x^{-1} on the whole space of smooth functions $C^\infty(\mathbb{R}^n)$. Therefore we cannot obtain the relation $\sigma^u(f)=\partial_x\sigma^w(f)|_{u=w_x}$ between symmetries of the JMKP system (2) and the potential JMKP system (6).

What we can get under $u=w_x$ is that if σ^w is a symmetry of the derivative potential JMKP system $\partial_x H_0(w)=0$ and $\partial_x H_1(w)=0$, where H_0 and H_1 are defined by (6), then $\sigma^u=\sigma^w|_{u=w_x}$ is a symmetry of the JMKP system (2). In the work of Ref. 5, the authors mix up two systems of the potential JMKP system and the derivative potential JMKP system above, which actually have different symmetry algebras in view of the second observation in Sec. II. The similar confusion appeared in the discussion of symmetries of the Kadomtsev–Petviashvili equation in Ref. 7.

IV. DISCUSSIONS

We remark that the JMKP system (2) is not a system of evolution equations with an evolution variable z, and so is not the potential JMKP system (6), either. For an evolution equation or a system of evolution equations: $u_t=K(x,u)$, which does not depend explicitly on the evolution variable t, a kind of more general symmetry algebras than Virasoro symmetry algebras has been considered in Ref. 8. However it has also been shown[9] that any Laurent polynomial time-dependent symmetry of $u_t=K(x,u)$ has to be of polynomial time-dependent form. Therefore for $u_t=K(x,u)$, there does not exist any symmetry of the following form:

$$\sigma(f)=\sum_{i=0}^{N}\frac{\partial^{N-i}f(t)}{\partial t^{N-i}}S_i(x,u),$$

where f is an arbitrary function of the time variable but $S_i(x,u)$, $0\leqslant i\leqslant N$, do not depend explicitly upon the time variable except the space variables. If we consider a t variable-coefficient system of evolution equations: $u_t=K(t,x,u)$, there appears a different situation.[10,11] For example, there exist Laurent polynomial time-dependent symmetries and much higher-degree polynomial time-dependent symmetries, even for a system of evolution equations in $1+1$ dimensions.[10,11]

One might ask whether or not the vector fields defined by (3) and (4) could become symmetries of the corresponding systems, under other conditions on ∂_x^{-1}. The answer is still no. Actually our conditions (10) on ∂_x^{-1} are the best possible set of conditions which gives a well-defined inverse operator of ∂_x. If we consider ∂_x^{-1} to be an indefinite integration operator as in Ref. 5, i.e., $\partial_x^{-1}: f \mapsto [\int_0^x f(\ldots,x',\ldots)dx']$ by using the notation in Sec. II, then we will meet some unsolvable difficulties in keeping reasonable differential operations as specified in Sec. II. But it is essential to keep differential operations when we make computation about, for example, Lie–Bäcklund symmetries. Therefore strictly speaking, there is no possibility which gives a definition of ∂_x^{-1} being an indefinite integration operation while discussing symmetries. Nevertheless one might disregard differential operations such as $\partial_x(fg)=\partial_x(f)g+f\partial_x(g)$ in order to view ∂_x^{-1} as an indefinite integration operation. But the vector fields defined by (3) and (4) are still not able to become symmetries of the corresponding systems, because our counterexamples can hold if the integration constants are chosen to be zero, one of all choices of integration constants. However one can even require some special integration constants to be introduced such that the symmetry conditions are satisfied. If so, there are at least two problems which cannot be solved. The first problem is that the vector fields involving ∂_x^{-1} are not already well defined, since the integration constants have not been chosen before checking symmetry conditions. The second problem is that the integration constants selected in this way will most likely depend on solutions of the original systems, because the symmetry conditions always contain solutions. Thus the symmetry vector fields themselves will sometimes be determined only after solutions of the original systems are given. This causes the symmetry vector fields not to be well defined. For the case of the JMKP and the potential JMKP systems, the vector fields defined by (3) and (4) are exactly such examples, if ∂_x^{-1} is viewed as an indefinite integration operator. There are some other similar situations, e.g., the situation of the modified KP equation, which is pointed out in Ref. 12.

In conclusion, what we have pointed out so far is that the vector fields defined by (3) and (4) cannot always become symmetries of the corresponding systems in any possible sense of ∂_x^{-1}. We are in doubt that such formal series as defined by (3) and (4) could become Lie–Bäcklund symmetries (except Lie-point symmetries) of differential equations. At least for evolution equations not explicitly depending on the evolution variable, there must be no such kind of symmetries, in Lie-point form or Lie–Bäcklund form, involving an arbitrary function of the evolution variable (see Ref. 9).

ACKNOWLEDGMENTS

The author would like to thank the City University of Hong Kong and the Research Grants Council of Hong Kong for financial support. He is also grateful to the referee for many valuable comments.

[1] F. Schwarz, J. Phys. Soc. Jpn. **51**, 2387 (1982); D. David, N. Kamran, D. Levi, and P. Winternitz, Phys. Rev. Lett. **55**, 2111 (1985); J. Math. Phys. **27**, 1225 (1986); D. David, D. Levi, and P. Winternitz, Phys. Lett. A **118**, 390 (1986); S. Y. Lou, J. Phys. A **26**, 4387 (1993).
[2] B. Champagne and P. Winternitz, J. Math. Phys. **29**, 1 (1988); S. Y. Lou and X. B. Hu, J. Phys. A **27**, L207 (1994).
[3] K. M. Tamizhmani, A. Ramani, and B. Grammaticos, J. Math. Phys. **32**, 2635 (1991).
[4] R. A. Leo, L. Martina, and G. Soliani, J. Math. Phys. **27**, 2623 (1986); L. Martina and P. Winternitz, Ann. Phys. (N.Y.) **196**, 231 (1989).
[5] S. Y. Lou and J. P. Weng, J. Math. Phys. **36**, 3492 (1995).
[6] B. Dorrizzi, B. Grammaticos, A. Ramani, and P. Winternitz, J. Math. Phys. **27**, 2848 (1986); J. Rubin and P. Winternitz, *ibid.* **31**, 2085 (1990).
[7] S. Y. Lou, J. Phys. A **26**, 4387 (1993).
[8] W. X. Ma, Acta Math. Appl. Sinica **17**, 388 (1994).
[9] W. X. Ma, Sci. China, Ser. A **34**, 769 (1991).
[10] W. X. Ma, R. K. Bullough, P. J. Caudrey, and W. I. Fushchych, J. Phys. A **30**, 5141 (1997).
[11] W. X. Ma, R. K. Bullough, and P. J. Caudrey, J. Nonlin. Math. Phys. **4**, 293 (1997) solv-int/9705015.
[12] W. X. Ma, Lett. Math. Phys. **41**, 237 (1997).

Response to "Comment on 'Generalized W_∞ symmetry algebra of the conditionally integrable nonlinear evolution equation' " [J. Math. Phys. 40, 3685 (1999)]

Sen-yue Lou[a)]
Applied Physics Department, Shanghai Jiao Tong University, Shanghai 200030, People's Republic of China and Institute of Mathematical Physics, Ningbo University, Ningbo 315211, People's Republic of China

Jian-ping Weng
Department of Physics, Lishui Normal College, Lishui 323000, People's Republic of China

(Received 24 October 1998; accepted for publication 15 December 1998)

To verify the correctness of the high order symmetries for the Jimbo–Miwa–Kadomtsev–Petviashvilli system (JMKP) and the potential JMKP (PJMKP) system given in our paper [J. Math. Phys. **36**, 3492–3497 (1995)], Ma's definition of ∂_x^{-1} may be used only for the general *nonkernel* solutions of the models. For some types of special solutions which are the kernels of some differential operators, one has to use ∂_x^{-1} as the indefinite operator and selected the integral functions appropriately. [S0022-2488(99)01207-4]

In Ref. 1, the author tries to solve the ∂_x^{-1} problem and give some remarks on Lou and Weng's paper.[2] However, the main results of the paper are not correct or not self-consistent.

(i) On one hand, the author claims that it is sufficient to define a right inverse of ∂_x and

$$\begin{cases} \partial_x^{-1}: C^\infty(\|\mathbb{R}^4) \to C^\infty(\|\mathbb{R}^4) \ \text{ being linear, } \ \partial_x\partial_x^{-1} = -1, \\ \partial_y\partial_x^{-1} = \partial_x^{-1}\partial_y, \ \ \partial_y\partial_x^{-1} = \partial_x^{-1}\partial_y, \ \ \partial_y\partial_x^{-1} = \partial_x^{-1}\partial_y, \end{cases} \tag{1}$$

is the only condition on the operator ∂_x^{-1} and

$$\partial_x^{-1} 0 = 0. \tag{2}$$

On the other hand, all the results of the paper are illustrated by taking the solutions as some special forms which are the kernels of the differential operators (∂_x, ∂_y, ∂_z and ∂_ξ, $\xi = x + t$). It is impossible to obtain these solutions if one uses (2) and (1) only though they are really solutions.

(ii) $\sigma_4^w(1)$ and $\sigma_4^u(1)$ are really symmetries of the PJMKP and JMKP, respectively. If one uses Ma's definition of the operator to check the correctness of $\sigma_4^w(1)$ and $\sigma_4^u(1)$, one must take w and u being general *nonkernel* solutions of the PJMKP and JMKP systems. We do check the conclusion both by hand and by computer algebras. Because it is too long to write the detail verification procedures down here, I suggest the readers who are interested in this problem to check the conclusions by using some nontrivial (nonkernel) solutions, say, general multisoliton solutions. The single plane soliton solutions read

$$w = 2k_1 \tanh\left(k_1 x + k_2 y + \frac{1}{2}\,\frac{k_2(4k_1^4 + k_2^2)}{k_1^2} z + \frac{1}{4}\,\frac{4k_1^4 + 3a_2^2}{k_1} t \right) \tag{3}$$

[a)]Electronic mail: sylou@public.nbptt.zj.cn

for potential JMKP and

$$u=2k_1^2\ \mathrm{sech}^2\left(k_1x+k_2y+\frac{1}{2}\frac{k_2(4k_1^4+k_2^2)}{k_1^2}z+\frac{1}{4}\frac{4k_1^4+3a_2^2}{k_1}t\right) \tag{4}$$

for JMKP. The corresponding $\sigma_4^w(1)$ and $\sigma_4^u(1)$ read

$$\sigma_4^w(1)=\frac{3}{2}\frac{16k_1^8+40k_1^4k_2^2+5k_2^4}{k_1^2}\ \mathrm{sech}^2\left(k_1x+k_2y+\frac{1}{2}\frac{k_2(4k_1^4+k_2^2)}{k_1^2}z+\frac{1}{4}\frac{4k_1^4+3a_2^2}{k_1}t\right), \tag{5}$$

$$\sigma_4^u(1)=-3\frac{16k_1^8+40k_1^4k_2^2+5k_2^4}{k_1}\ \mathrm{sech}^2\left(k_1x+k_2y+\frac{1}{2}\frac{k_2(4k_1^4+k_2^2)}{k_1^2}z+\frac{1}{4}\frac{4k_1^4+3a_2^2}{k_1}t\right)$$
$$\times\tanh\left(k_1x+k_2y+\frac{1}{2}\frac{k_2(4k_1^4+k_2^2)}{k_1^2}z+\frac{1}{4}\frac{4k_1^4+3a_2^2}{k_1}t\right). \tag{6}$$

The general multisoliton solutions of the models can be found in literature, say, in Ref. 3.

(iii) The explanation of the observation 2 in Ref. 1 is not correct. It is known that if two equations $F(u)=0$ and $G(v)=0$ are related by a Bäcklund transformation $B(u,v)=0$, then the symmetries of two equations are related by $B_u'\sigma^u+B_v'\sigma^v=0$. The author uses $u_t=u_x$ and $v_{tx}=v_{xx}$ as a simple illustrative example (the author used the same notation for two equation). The solutions of these two equations may be related in two ways. (1) $v=u+f(t)$ for arbitrary $f(t)$ that means if σ^u is a symmetry of $u_t=u_x$ then it must also be a symmetry of $v_{tx}=v_{xx}$. The author's special example is $\sigma^u=u^2$. So u^2 (but not v^2!!) must also be a symmetry of $v_{tx}=v_{xx}$. (2) $u=v_x$ that means if σ^v is a symmetry of $v_{tx}=v_{xx}$, then $\sigma^u=\sigma_x^v$ must be a symmetry of the u equation. The author's confusion comes from his using the same notations of u and v.

(iv) If one uses the trivial (kernel) solutions as a special examples and the authors' definition of ∂_x, many important traditional conclusions related to symmetries and Bäcklund transformations will be destroyed. Say, in the proof procedures of the hereditary property of the recursion operators of (1+1)-dimensional integrable models, one has to use the condition, $u(x\to\infty,t)=0$, which is not valid for u is selected as a special kernel of differential operators.[4]

In summary, to verify the correctness of the high order symmetries for the JMKP and the PJMKP system Ma's definition of ∂_x^{-1} may be used only for the general *nonkernel* solutions of the models. For some types of special solutions which are the kernels of some differential operators, one has to use ∂_x^{-1} as an indefinite operator and selected the integral functions appropriately.

[1] W-x. Ma, ''Comment on 'Generalized W_∞ symmetry algebra of the conditionally integrable nonlinear evolution equation,' '' J. Math. Phys. **40**, 3685 (1999).
[2] S-y. Lou and J-p. Weng, ''Generalized W_∞ symmetry algebra of the conditionally integrable nonlinear evolution equation,'' J. Math. Phys. **36**, 3492 (1995).
[3] H-y. Ruan and S-y. Lou, ''Higher dimensional dromion structures: Jimbo–Miwa–Kadomtsev–Petviashvili system,'' J. Math. Phys. **38**, 3123 (1997).
[4] C-h. Gu, B-l. Guo, Y-s. Li, C-w. Cao, C. Tian, G-z. Tu, and M-l. Ge, *Soliton Theory and its Application* (Zhejiang, China, 1990), pp. 216–267.

Comment on ''Geometric phase, bundle classification, and group representation'' [J. Math. Phys. 37, 1218 (1996)]

H. Azad
Department of Mathematical Sciences, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia

M. N. Qureshi
Azad Jammu & Kashmir University, Muzaffarabad, Pakistan

M. Ziad
Department of Mathematics, Quaid-i-Azam University, Islamabad, Pakistan

(Received 30 April 1998; accepted for publication 11 January 1999)

This note is a comment on the mathematical aspects of the paper by Mostafazadeh [J. Math. Phys. **37**, 1218 (1996)]. Its aim is to remove some of the ambiguities and mistakes of the paper by making more transparent the mathematics involved in the calculation of topological charges.
[S0022-2488(99)03406-4]

I. INTRODUCTION

In his paper,[1] Mostafazadeh discusses at length the relevance of the Borel–Weil–Bott theorem to Berry–Simon theory[2] and computations involving Chern classes as well as generators for the second de Rham cohomology of flag manifolds. However, there are some conceptual errors in the paper which may be a hindrance to a proper understanding. For example, on p. 1226 (line 2) of Ref. 1, one finds the statement that a flag manifold K/L of a compact Lie group K is a submanifold of K/T, T being a maximal torus of K, which is not true as both are homogeneous spaces of the same group K. At the infinitesimal level, i.e., the level of tangent spaces, the above statement holds. However, as a vector space has no topology, one must work directly with the manifold if one wants to compute Chern numbers. Also, the Chern number for the determinant bundle of $\mathbb{CP}^N$ is calculated to be 2 whereas it is $(N+1)$. In this note we rectify these mistakes. As the concepts relating to Chern numbers and their actual computations are important for physics, it is desirable to explain them in a clearer and more rigorous manner. We do this in Secs. II and III. We review two key examples in detail; later we give the precise statements in the most general form, omitting all proofs, and refer instead to the literature. All of this goes back essentially to Borel–Hirzebruch.[3]

II. TWO BASIC METHODS FOR OBTAINING GENERATORS OF COHOMOLOGY

The following examples illustrate two basic methods for obtaining cohomology generators:

(I) Let M be a complex manifold, $\{U_i\}_{i\in I}$ an open covering of M and $f_i: U_i \Rightarrow \mathbb{C}$ holomorphic functions such that on the intersection $U_{ij}=U_i\cap U_j$, $g_{ij}=f_i/f_j$ is nonzero. The system of functions defines a line bundle $\mathcal{L}$ on M,[4] namely, we identify $(p,z)\in U_i\times\mathbb{C}$ with $(q,w)\in U_j\times\mathbb{C}$ if and only if $p=q$ and $w=g_{ji}(p)z$. The functions $\{f_i\}$ represent a section σ, called the canonical section, of this bundle.[4] A norm on this bundle is given by a family of functions

$$\varphi_i : U_i \rightarrow \mathbb{R}^{>0}$$

such that on $U_i\cap U_j$, we have

$$\varphi_i/\varphi_j = |g_{ji}|.$$

The form

$$C(\mathcal{L})=\frac{i}{2\pi}\partial\overline{\partial}\log[N(\sigma)^{-2}]=\frac{i}{2\pi}\sum_{\alpha,\beta}\frac{\partial^2}{\partial z_\alpha\partial\overline{z}_\beta}[\log N(\sigma)]^{-2}dz_\alpha\wedge d\overline{z_\beta}$$

represents the Chern class of $\mathcal{L}$.

A standard way of obtaining line bundles is by using the functions which define a submanifold of codimension 1. For example, consider complex projective space $\mathbb{P}^1(\mathbb{C})$ with homogeneous coordinates $[z_0:z_1]$. Let $U_0=\{[z_0:z_1]:z_0\neq 0\}$, $U_1=\{[z_0:z_1]:z_1\neq 0\}$. Let $p=[0:1]$. To the point p there correspond the functions $f_0=1$ on U_0 and $f_1=z_0/z_1$ on U_1, so $g_{ij}=(z_j/z_i)$ $(i,j=0,1)$ and the norm of the canonical section σ represented by the functions $\{f_0,f_1\}$ is $\|\sigma\|=|z_0|/|z|$. So

$$C(\mathcal{L})=\frac{i}{2\pi}\partial\overline{\partial}\log\frac{|z|^2}{|z_0|^2}=\frac{i}{2\pi}\partial\overline{\partial}\log\left(1+\left|\frac{z_1}{z_0}\right|^2\right).$$

It represents a generator of $H^2(\mathbb{P}^1,\mathbb{Z})$. All generators of $H^2(M,\mathbb{Z})$, M a flag manifold—we define this term in Sec. III—are obtained in an analogous manner.

(II) Another way of obtaining generators of cohomology uses representation theory of groups. It is illustrated by the following example.

Consider the flag manifold

$$F_{1,2}=\{(\ell,\pi):\ell\subset\pi,\quad \ell \text{ and } \pi \text{ being lines and planes in } \mathbb{C}^3 \text{ passing through the origin}\}.$$

Let e_0,e_1,e_2 be the standard basis of $\mathbb{C}^3$, let $\ell_0=\mathbb{C}e_0$, $\pi_0=\mathbb{C}\langle e_0,e_1\rangle$, and $\xi_0=(\ell_0,\pi_0)$. So $F_{1,2}$ is the orbit of ξ_0 under the natural action of SL(3,$\mathbb{C}$). The stabilizer of ξ_0 in $G=\mathrm{SL}(3,\mathbb{C})$ is the group B of upper triangular matrices, so $G/B\cong G\cdot\xi_0=F_{1,2}$.

Now the basic representations of SL(3,$\mathbb{C}$) are $\mathbb{C}^3=V$ and $\wedge^2(V)$ and their highest weight vectors (i.e., those lines which are B invariant) are v_1 and $v_1\wedge v_2$, where $v_1=(1,0,0)^T$, $v_2=(0,1,0)^T$, T being the transpose. Consider the forms $\overline{\omega_\alpha}$ and $\overline{\omega_\beta}$ on G defined by

$$\overline{\omega_\alpha}=i\partial\overline{\partial}\log\|g\cdot v_1\|,$$

$$\overline{\omega_\beta}=i\partial\overline{\partial}\log\|g\cdot(v_1\wedge v_2)\|.$$

Let $\pi:G\to G/B=F_{1,2}$ be the natural map. The forms $\overline{\omega_\alpha},\overline{\omega_\beta}$ descend to forms ω_α and ω_β on $F_{1,2}$ in the sense that $\pi^*\omega_\alpha=\overline{\omega_\alpha}$, $\pi^*\omega_\beta=\overline{\omega_\beta}$. These forms generate $H^2(G/B,\mathbb{R})$ and the projective lines dual to these forms have the following explicit description.

Let

$$L_\alpha=\left\langle\begin{pmatrix}x & y & 0\\ z & t & 0\\ 0 & 0 & 1\end{pmatrix}:\begin{pmatrix}x & y\\ z & t\end{pmatrix}\in\mathrm{SL}(2,\mathbb{C})\right\rangle,$$

$$L_\beta=\left\langle\begin{pmatrix}1 & 0 & 0\\ 0 & x & y\\ 0 & z & t\end{pmatrix}:\begin{pmatrix}x & y\\ z & t\end{pmatrix}\in\mathrm{SL}(2,\mathbb{C})\right\rangle.$$

We have $\mathbb{P}_\alpha=L_\alpha$, $\xi_0\cong\mathbb{P}^1(\mathbb{C})$ and $\mathbb{P}_\beta=L_\beta\cdot\xi_0\cong\mathbb{P}^1(\mathbb{C})$ and moreover $\int_{\mathbb{P}_\beta}\omega_\alpha=\delta_{\alpha_\beta}$: see Ref. 5.

III. FLAG MANIFOLDS AND THEIR COHOMOLOGY: TOPOLOGICAL CHARGES

Let G be a complex semisimple Lie group and B a Borel subgroup of G; so B is a maximal connected solvable subgroup of G. A homogeneous space of the form G/P, where the group $P\supset B$, is a flag manifold. These are precisely those complex compact homogeneous spaces which are simply connected and are embeddable in some projective space.

In this section, we give a precise description of the second de Rham cohomology of flag manifolds and give a formula for the Chern class of line bundles on flag manifolds; this formula involves the so-called *topological charges* associated to line bundles. The results are due originally to Borel–Hirzebruch.[3] We will just state the results, as all the proofs are in Ref. 3; they are also given, from a different point of view, in Ref. 5. We assume familiarity with the theory of weights and roots as set forth in Refs. 6 and 7.

Let G be a complex reductive group, B a Borel subgroup of G,T a maximal torus of G contained in B,R the roots of T in G, R^+ the positive system of roots defined by the pair (B,T), and S the corresponding system of simple roots.

One knows that for each $\alpha\in\mathbb{R}^+$ there exist $X_\alpha, X_{-\alpha}\in \mathrm{Lie}(G)$ such that the map

$$\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}\mapsto X_\alpha, \quad \begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}\mapsto X_{-\alpha}$$

extends to an isomorphism of $\mathrm{SL}(2,\mathbb{C})$ onto the Lie algebra generated by $X_\alpha, X_{-\alpha}$. Hence there exists a homomorphism φ_α from $\mathrm{SL}(2,\mathbb{C})$ onto a subgroup L_α of G whose Lie algebra is generated by $X_\alpha, X_{-\alpha}$.

The group K generated by $\varphi_\alpha(\mathrm{SU}(2))$ (α simple) is a maximal compact subgroup of G. Let $\pi\subset S$ and $P_\pi = P$ be the corresponding parabolic subgroup. So P_π is generated by B and $L_\alpha(\alpha\in\pi)$. Let $\xi_0 = eP\in G/P$. For $\alpha\in S\setminus\pi$ we have $L_\alpha\cdot\xi_0\cong\mathbb{P}^1(\mathbb{C})\cong S^2$: denote this projective line by $\mathbb{P}_\alpha$. Let $\{\omega_\alpha:\alpha\in S\}$ be the fundamental dominant weights, ρ_α the irreducible representation with highest weight ω_α, and v a highest weight vector therein. Let $\overline{\omega_\alpha}$ be the form on G defined by

$$\overline{\omega_\alpha} = i\partial\bar\partial \log\|\rho_\alpha(g)\cdot v\|^2.$$

For $\alpha\in S\setminus\pi$, the form $\overline{\omega_\alpha}$ is the pull-back of a K-invariant form, which we also denote by ω_α, on G/P; namely, if s is a local cross section of the natural map $G\to G/P$, then $\omega_\alpha(\xi) = i\partial\bar\partial\log\|\rho_\alpha(s(\xi))\cdot v\|^2$. Using these forms, we have the following description of $H_2(G/P)$ and $H^2(G/P)$:

Theorem: *The projective lines* $\mathbb{P}_\alpha(\cong S^2)(\alpha\in S\setminus\pi)$ *form a basis of* $H_2(G/P)$ *and the forms* $\omega_\alpha(\alpha\in S\setminus\pi)$ *form a basis of* $H^2(G/P)$. Moreover,

$$\frac{1}{2\pi}\int_{\mathbb{P}_\beta}\omega_\alpha = \delta_{\alpha,\beta}.$$

For a proof, see Refs. 3 and 5.

Now consider a character χ of the parabolic group P of G. This character gives rise to a line bundle L_χ (over G/P) which is by definition $G\times\mathbb{C}/\sim$, where the equivalence relation $\sim$ is defined by

$$(g,z)\sim(g_1,z_1)\Leftrightarrow g_1 = gp, \quad z_1 = \chi(p)^{-1}z$$

for some $p\in P$.

Take a local nonvanishing section σ of the bundle L_χ. The Chern class $C(L_\chi)$ of L_χ is represented by

$$\frac{i}{2\pi}\,\partial\bar\partial\log\,[N(\sigma)]^{-2},$$

N being a norm on L_χ. So $C(L_\chi)=\Sigma_{\alpha\in S\setminus\pi} n_\alpha[\omega_\alpha]$.

Proposition: We have $n_\alpha=-\langle\chi,\check{\alpha}\rangle\in\mathbb{Z}$.

For a proof, see Ref. 5.

These integers are what are called topological charges in the physics literature and to which Mostafazadeh refers toward the end of Ref. 1.

Finally, let us determine the charge of the anticanonical bundle of $\mathbb{P}^n(\mathbb{C})$, i.e., the Chern class of the line bundle $\det T\mathbb{CP}^n$. Certainly this can be done using the above formula. However, it is more instructive to proceed directly. This is quite standard (see, e.g., Ref. 4). For the sake of completeness, we give a derivation in the present setup. For this, notice that if we are given local nonvanishing sections s_α of a line bundle L, defined on open sets U_α which cover the manifold M, then the map

$$(\xi,z)\mapsto(zs_\alpha(\xi))$$

is an isomorphism of $U_\alpha\times\mathbb{C}$ onto $\pi^{-1}(U_\alpha)$, $\pi:L\to M$ being the projection all of whose fibers are lines, so the transition functions of L are given by

$$g_{\beta\alpha}(\xi)=s_\alpha(\xi)/s_\beta(\xi).$$

For the computation of the Chern class $C(K(\mathbb{P}^n))$ of the canonical bundle $K(\mathbb{P}^n)$ of $\mathbb{P}^n$, we take the open sets

$$U_i=\{[z_0:\cdots:z_n]:z_i\neq 0\},\quad i=0,1,...,n$$

and on U_i the differential form ω_i defined by

$$\omega_i=d(z_0/z_i)\wedge\cdots\wedge d(z_{i-1}/z_i)\wedge d(z_{i+1}/z_i)\cdots\wedge d(z_n/z_i).$$

So $\omega_i=(z_i/z_0)^{-(n+1)}(-1)^i\omega_0$. Rescaling ω_i may assume that $\omega_i=(z_i/z_0)^{-(n+1)}\omega_0$. Therefore $\omega_i/\omega_0=(z_i/z_0)^{-(n+1)}$ and so $g_{0i}=(z_0/z_i)^{n+1}$. Hence

$$g_{ij}=g_{i0}g_{0j}=(z_i/z_j)^{n+1}.$$

On the other hand, the line bundle $[H]$ defined by the hyperplane $z_0=0$ is given on U_i by the functions z_0/z_i, so its transition functions $g_{\widetilde{ij}}$ are $g_{\widetilde{ij}}=(z_0/z_i)/(z_0/z_j)=(z_j/z_i)$. Therefore $g_{ij}=(g_{\widetilde{ij}})^{-(n+1)}$, so

$$K(\mathbb{P}^n)=-(n+1)[H]$$

and hence $\wedge^n T\mathbb{P}^n=(n+1)[H]$.

So the Chern number of the determinant bundle of $\mathbb{P}^n$ is $(n+1)$.

[1] A. Mostafazadeh, J. Math. Phys. **3**, 1218 (1996).
[2] M. V. Berry, Proc. R. Soc. London, Ser. A **392**, 392 (1984).
[3] A. Borel and F. Hirzebruch, Am. J. Math. **80**, 459 (1958).
[4] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1994).
[5] H. Azad, R. Kobayashi, and M. N. Qureshi, J. Algebra **196**, 620 (1997).
[6] R. Steinberg, ''Lectures on Chavalley groups,'' Yale University 1967.
[7] R. Slansky, Phys. Rep. **79**, 1 (1981).

Erratum: ''Geometric phase, bundle classification, and group representation'' [J. Math. Phys. 37, 1218 (1996)]

Ali Mostafazadeh
Koç University, Istinye 80860 Istanbul, Turkey

(Received 19 October 1998; accepted for publication 11 January 1999)

[S0022-2488(99)03506-9]

Azad *et al.*,[1] have informed me of a mistake in Eq. (64) of my paper, Ref. 2, which has led to another mistake in Eq. (61). These equations only hold for $N=1$. For arbitrary N, they must be changed to

$$c_1(T\mathbb{C}P^N)=\chi(\mathbb{C}P^N)=N+1, \tag{1}$$

$$\mathrm{Det}(T\mathbb{C}P^N)=\underbrace{E^*\otimes E^*\otimes\cdots\otimes E^*}_{N+1\ \text{times}}, \tag{2}$$

respectively. Equation (1) can be easily inferred from the computation of the total Chern class of $T\mathbb{C}P^N$ which can be found in Example 6.3 of Ref. 3. Equation (2) follows from Eq. (1) using the arguments given in Sec. IV of Ref. 2, below Eq. (61).

Note that the mistake that occurred in Eqs. (61) and (64) does not change any of the conclusions of Ref. 2.

[1] H. Azad, M. N. Qureshi, and M. Ziad (unpublished).
[2] A. Mostafazadeh, J. Math. Phys. **37**, 1218–1233 (1996).
[3] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Asymptotics of the scattering coefficients for a generalized Schrödinger equation

Tuncay Aktosun
Department of Mathematics, North Dakota State University, Fargo, North Dakota 58105

Martin Klaus
Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

(Received 5 October 1998; accepted for publication 11 February 1999)

The generalized Schrödinger equation $d^2\psi/dx^2+F(k)\psi=[ikP(x)+Q(x)]\psi$ is considered, where P and Q are integrable potentials with finite first moments and F satisfies certain conditions. The behavior of the scattering coefficients near zeros of F is analyzed. It is shown that in the so-called exceptional case, the values of the scattering coefficients at a zero of F may be affected by $P(x)$. The location of the k-values in the complex plane where the exceptional case can occur is studied. Some examples are provided to illustrate the theory. © *1999 American Institute of Physics.* [S0022-2488(99)03007-8]

I. INTRODUCTION

In this paper we consider the generalized Schrödinger equation

$$\frac{d^2\psi(k,x)}{dx^2}+F(k)\psi(k,x)=[ikP(x)+Q(x)]\psi(k,x),\quad x\in\mathbf{R}, \tag{1.1}$$

where the properties of F will be detailed below. The functions P and Q satisfy

$$P\in L_1^1(\mathbf{R}),\quad Q\in L_1^1(\mathbf{R}), \tag{1.2}$$

where $L_1^1(\mathbf{R})$ is the class of measurable functions f such that $\int_{-\infty}^{\infty}dx\,|f(x)|(1+|x|)<+\infty$. For the majority of the paper, P and Q need not be real valued; if they are, this will be stated explicitly. In applications, k may correspond to a wave number while $F(k)$ may represent energy. The coefficient $P(x)$ may represent the absorptive properties of a medium, and $Q(x)$ may be a restoring force density or a potential for an external force. Some special cases of (1.1) are

(A) $F(k)=k^2$ with $P(x)\equiv 0$,
(B) $F(k)=k^2$ with $P(x)\not\equiv 0$,
(C) $F(k)=k^2+1/(4\beta^2)$ with $\beta>0$.

Case (A) corresponds to the well-known quantum-mechanical case of the Schrödinger equation on the line with potential $Q(x)$. Case (B) was studied by Jean and Jaulent,[1–4] and more recently by Sattinger and Szmigielski,[5] and by us[6] when P is real valued. Case (C) has been investigated by Kaup[7] in connection with the inverse scattering transform for an evolution equation (a long-wave water equation resembling the Boussinesq equation) by Tsutsumi[8] and, more recently, under the assumption that $\int_{-\infty}^{\infty}dx\,P(x)=0$, by Sattinger and Szmigielski.[9]

Our interest in (1.1) is motivated by various inverse problems associated with (1.1). In studying such problems, one needs to know the asymptotics of various quantities as the parameter k approaches certain special values, in particular as $F(k)\to\infty$ or as $k\to k_0$, where k_0 is a zero of F. In this paper we will only be concerned with the second situation. We will call $k_0\in\mathbf{C}$ a *critical* value of (1.1) if $F(k_0)=0$. Here, $\mathbf{C}$ denotes the complex plane. The quantities whose asymptotics we will study are the transmission and reflection coefficients associated with (1.1). Before we define these quantities we list the assumptions on F:

(H1) Supposing k_0 is a critical value of (1.1), there exists a set $\mathcal{S}\subset \mathbf{C}$ such that $F(k)$ is continuous on $\mathcal{S}$, $F(k)\neq 0$ on $\mathcal{S}\setminus\{k_0\}$, and the map $k\mapsto\mu(k)=\sqrt{F(k)}$ is one-to-one for $k\in S$. Here the branch of the square root is such that $0\leqslant\arg\mu<\pi$, where $\mu=\mu(k)$.

(H2) There is a path $\mathcal{P}(k_0)$ in $\mathcal{S}$ containing k_0 on which μ takes on real non-negative values.

Note that, by (H1), $\mathcal{D}=\mu(\mathcal{S})$ is a subset of the closed upper-half complex plane $\overline{\mathbf{C}^+}$. (H2) indicates that there is an $\epsilon>0$ so that $[0,\epsilon]\in\mathcal{D}$. In cases (A) and (B), $k_0=0$ is the only critical value. We may then choose $\mathcal{S}=\{k:0\leqslant\arg k<\pi\}\cup\{0\}$, so that $\mu(k)=k$ and $\mathcal{D}=\mathcal{S}$. For the path $\mathcal{P}(k_0)$ we may take the interval $[0,+\infty)$. In case (C) the critical values are $k_0=\pm i/(2\beta)$. The disk $\{k: |k-i/(2\beta)|\leqslant 1/(2\beta)\}$ can then be used as $\mathcal{S}$ near the critical point $+i/(2\beta)$ and we have $\mathcal{D}=\{\mu: |\mu|\leqslant 1/(2\beta),\ 0\leqslant\arg\mu<\pi\}\cup\{0\}$. As the path $\mathcal{P}(k_0)$ we can take the imaginary interval $i[0,1/(2\beta)]$. The modifications for the other critical point are obvious.

For $k\in\mathcal{S}$, (1.1) possesses the solutions $f_l(k,x)$ and $f_r(k,x)$, the so-called Jost solutions from the left and from the right, respectively, that are uniquely defined by their spatial asymptotics, namely,

$$f_l(k,x)=e^{i\mu x}[1+o(1)],\quad f_l'(k,x)=i\mu e^{i\mu x}[1+o(1)],\quad x\to+\infty, \tag{1.3}$$

$$f_r(k,x)=e^{-i\mu x}[1+o(1)],\quad f_r'(k,x)=-i\mu e^{-i\mu x}[1+o(1)],\quad x\to-\infty, \tag{1.4}$$

where the prime indicates the derivative with respect to the spatial variable x. For $k\in\mathcal{P}(k_0)\setminus\{k_0\}$ the Jost solutions obey

$$f_l(k,x)=\frac{1}{T(k)}e^{i\mu x}+\frac{L(k)}{T(k)}e^{-i\mu x}+o(1),\quad x\to-\infty, \tag{1.5}$$

$$f_r(k,x)=\frac{1}{T(k)}e^{-i\mu x}+\frac{R(k)}{T(k)}e^{i\mu x}+o(1),\quad x\to+\infty, \tag{1.6}$$

which define the transmission coefficient T and the reflection coefficients R from the right and L from the left, respectively. These quantities will collectively be referred to as scattering coefficients. It is also possible to define the scattering coefficients in terms of certain Wronskians of the Jost solutions. For example, letting $[f;g]=fg'-f'g$ denote the Wronskian, from (1.1) and (1.3)–(1.6) we get

$$\frac{2i\mu}{T(k)}=[f_r(k,\cdot);f_l(k,\cdot)]. \tag{1.7}$$

In analogy with the usual Schrödinger equation, given a critical value k_0 we will distinguish between two cases: We say that the generic (exceptional) case occurs at $k=k_0$ if and only if $f_l(k_0,x)$ and $f_r(k_0,x)$ are linearly independent (dependent). In the exceptional case, we let γ denote the nonzero constant defined as,

$$\gamma=\frac{f_l(k_0,x)}{f_r(k_0,x)}. \tag{1.8}$$

From (1.7) we see that k_0 corresponds to the exceptional case if and only if

$$F(k_0)=0,\quad \lim_{k\to k_0}\frac{\mu(k)}{T(k)}=0.$$

In short, we will say that k_0 is an exceptional value if it corresponds to the exceptional case for (1.1).

The behavior of the scattering coefficients of (1.1) at the critical values k_0 does not seem to have been studied in detail before, except in cases (A) and (B). In these two cases it is known[6,10–13] that there are two ways in which $T(k)$ can behave as $k\to 0$: either $T(k)=ick+o(k)$ for some nonzero c, or $T(k)=T(0)+o(1)$ with $T(0)\neq 0$. The former corresponds to the generic case and the latter corresponds to the exceptional case. In case (C) a detailed investigation of the behavior of the scattering coefficients near $k_0=\pm 1/(2\beta)$ does not seem to have been done before. This is one of our goals in this paper, and particular attention will be paid to the exceptional case. In connection with a statement made in Theorem 2.6 of Ref. 9 regarding case (C) with $\beta=\frac{1}{2}$, we would like to comment that while it is true that for reflectionless potentials only the exceptional case can occur at $k=\pm i$, there are also potentials $P(x)$ and $Q(x)$, in particular real ones, which are not reflectionless and for which the exceptional case occurs. This will be discussed in more detail in Sec. III.

This paper is organized as follows. In Sec. II we prove our main result concerning the behavior of the scattering coefficients at a critical value (Theorem 2.2) and apply it to cases (A)–(C) (Corollary 2.3). We also present some information about the location of the exceptional k-values in the complex plane. In Sec. III we consider case (C) in more detail, show that one must not identify the exceptional case with the reflectionless case, and provide four examples illustrating the location of the exceptional k-values and other aspects of the theory.

II. ASYMPTOTICS OF THE SCATTERING COEFFICIENTS

In this section we study the asymptotic behavior of the scattering coefficients as $k\to k_0$, where k_0 is a critical value of (1.1). In doing so we will only be concerned with the leading terms of the asymptotic expansions. Our main result is presented in Theorem 2.2. For its proof, we first need some results about the usual Schrödinger equation.

Consider the pair of Schrödinger equations

$$\frac{d^2\phi_j(\mu,x)}{dx^2}+\mu^2\phi_j(\mu,x)=V_j(x)\phi_j(\mu,x),\quad j=1,2, \tag{2.1}$$

where $V_j\in L^1(\mathbf{R})$. Here μ is allowed to range over all of $\overline{\mathbf{C}^+}$; it is not restricted to $\mathcal{D}$ defined earlier. Let t_j denote the transmission coefficient and r_j and l_j denote the reflection coefficients from the right and left, respectively, for the potential V_j. Let $g_{j;l}(\mu,x)$ and $g_{j;r}(\mu,x)$ denote the corresponding Jost solutions of (2.1) from the left and right, respectively. It is known[10,11,13] that

$$g_{j;l}(-\mu,x)=t_j(\mu)g_{j;r}(\mu,x)-r_j(\mu)g_{j;l}(\mu,x),\quad \mu\in\mathbf{R},$$

$$g_{j;r}(-\mu,x)=t_j(\mu)g_{j;l}(\mu,x)-l_j(\mu)g_{j;r}(\mu,x),\quad \mu\in\mathbf{R}. \tag{2.2}$$

Since μ appears as μ^2 in (2.1), $g_{j;l}(-\mu,x)$ and $g_{j;r}(-\mu,x)$ are also solutions of (2.1), and $g_{j;l}(-\mu,x)=e^{-i\mu x}[1+o(1)]$ as $x\to+\infty$ and $g_{j;r}(-\mu,x)=e^{i\mu x}[1+o(1)]$ as $x\to-\infty$.

Proposition 2.1: Suppose that $V_j\in L^1(\mathbf{R})$ for $j=1,2$. Then the scattering coefficients of (2.1) satisfy

$$\frac{1}{t_2(\mu)}=\frac{1}{t_1(\mu)}+\frac{i}{2\mu}\int_{-\infty}^{\infty}dx\,[V_2(x)-V_1(x)]g_{2;l}(\mu,x)g_{1;r}(\mu,x),\quad \mu\in\overline{\mathbf{C}^+}\setminus\{0\}, \tag{2.3}$$

$$\frac{l_2(\mu)}{t_2(\mu)}=\frac{l_1(\mu)}{t_1(\mu)}-\frac{i}{2\mu}\int_{-\infty}^{\infty}dx\,[V_2(x)-V_1(x)]g_{2;l}(\mu,x)g_{1;r}(-\mu,x),\quad \mu\in\mathbf{R}\setminus\{0\}, \tag{2.4}$$

$$\frac{1}{t_2(\mu)}=\frac{1}{t_1(\mu)}+\frac{i}{2\mu}\int_{-\infty}^{\infty}dx\,[V_2(x)-V_1(x)]g_{2;r}(\mu,x)g_{1;l}(\mu,x),\quad \mu\in\overline{\mathbf{C}^+}\setminus\{0\}, \tag{2.5}$$

$$\frac{r_2(\mu)}{t_2(\mu)}=\frac{r_1(\mu)}{t_1(\mu)}-\frac{i}{2\mu}\int_{-\infty}^{\infty}dx\,[V_2(x)-V_1(x)]g_{2;r}(\mu,x)g_{1;l}(-\mu,x),\quad \mu\in\mathbf{R}\setminus\{0\}. \tag{2.6}$$

Proof: First, let us note that (2.3) and (2.5) are given on p. 329 of Ref. 13, and some formulas related to (2.4) and (2.6) can be also found there. We will give a different proof which yields both (2.3) and (2.4) simultaneously. The proof of (2.5) and (2.6) is similar and hence will be omitted. By the variation of parameters formula, $g_{2;l}(\mu,x)$ obeys the integral equation

$$g_{2;l}(\mu,x)=g_{1;l}(\mu,x)+\int_x^\infty dy\,\mathcal{G}(\mu;x,y)[V_2(y)-V_1(y)]g_{2;l}(\mu,y), \tag{2.7}$$

where

$$\mathcal{G}(\mu;x,y)=\frac{g_{1;l}(\mu,x)g_{1;r}(\mu,y)-g_{1;r}(\mu,x)g_{1;l}(\mu,y)}{[g_{1;l}(\mu,\cdot);g_{1;r}(\mu,\cdot)]}. \tag{2.8}$$

Note that the Wronskian in (2.8) is related to the transmission coefficient as

$$t_j(\mu)=-\frac{2i\mu}{[g_{j;l}(\mu,\cdot);g_{j;r}(\mu,\cdot)]}. \tag{2.9}$$

Now (2.3) and (2.4) follow by letting $x\to-\infty$ in (2.7) and using (1.4), (1.5), (2.2), and (2.9). ■

In the next theorem, the behavior of the scattering coefficients of (1.1) is analyzed at critical k-values.

Theorem 2.2: Suppose $P,Q\in L_1^1(\mathbf{R})$ and $F(k)$ satisfies (H1) and (H2). If $k_0\in\mathbf{C}$ is a critical value of (1.1), then we have the following.

(i) In the generic case we have

$$T(k)=-\frac{2i\mu}{[f_l(k_0,\cdot);f_r(k_0,\cdot)]}+o(\mu),\quad k\to k_0\quad\text{in}\quad\mathcal{S}, \tag{2.10}$$

$$L(k)=-1+o(1),\quad R(k)=-1+o(1),\quad k\to k_0\quad\text{in}\quad\mathcal{P}(k_0).$$

(ii) In the exceptional case, using the constants α and ω defined by

$$\alpha=\lim_{k\to k_0}\frac{k-k_0}{\mu(k)},\qquad \omega=\gamma^2+1-\alpha\int_{-\infty}^\infty dx\,P(x)f_l(k_0,x)^2,$$

we distinguish two subcases: (a) If α exists and is finite and $\omega\neq0$, then

$$T(k)=\frac{2\gamma}{\omega}+o(1),\quad k\to k_0\quad\text{in}\quad\mathcal{S}, \tag{2.11}$$

$$L(k)=\frac{2\gamma^2-\omega}{\omega}+o(1),\quad k\to k_0\quad\text{in}\quad\mathcal{P}(k_0), \tag{2.12}$$

$$R(k)=\frac{2-\omega}{\omega}+o(1),\quad k\to k_0\quad\text{in}\quad\mathcal{P}(k_0). \tag{2.13}$$

(b) If $\lim_{k\to k_0}|(k-k_0)/\mu(k)|=+\infty$ and $\int_{-\infty}^\infty dx\,P(x)f_l(k_0,x)^2\neq0$, then

$$T(k_0)=0,\quad L(k_0)=-1,\quad R(k_0)=-1.$$

In the exceptional case, if α exists and $\omega=0$, then the scattering coefficients are not continuous at k_0; if α does not exist, then, in general, the scattering coefficients are not continuous at k_0.

Proof: In (2.3)–(2.6) we replace $V_1(x)$ by $ik_0P(x)+Q(x)$ and $V_2(x)$ by $ikP(x)+Q(x)$ and note that because of (1.2) we have $V_j\in L_1^1(\mathbf{R})$ for $j=1,2$ instead of just $V_j\in L^1(\mathbf{R})$. The stronger assumption allows us to take the limit $\mu\to 0$ in (2.3)–(2.6). Thanks to Proposition 2.1 we can make full use of the results[12] known in the case $P(x)\equiv 0$. For $g_{1;l}(\mu,x)$ and $g_{2;r}(\mu,x)$ in (2.3)–(2.6), we substitute $f_l^{[0]}(\mu,x)$ and $f_r^{[0]}(\mu,x)$, respectively, where the latter two are the Jost solutions of

$$\frac{d^2\varphi(\mu,x)}{dx^2}+\mu^2\varphi(\mu,x)=[ik_0P(x)+Q(x)]\varphi(\mu,x). \tag{2.14}$$

Let $T^{[0]}(\mu)$, $R^{[0]}(\mu)$, and $L^{[0]}(\mu)$ denote the scattering coefficients associated with (2.14). Then from (2.3) we get

$$\frac{1}{T(k)}=\frac{1}{T^{[0]}(\mu)}\left[1-\frac{k-k_0}{2\mu}T^{[0]}(\mu)\int_{-\infty}^{\infty}dx\,P(x)f_r^{[0]}(\mu,x)f_l(k,x)\right],\quad \mu\in\mathcal{D}\setminus\{0\}. \tag{2.15}$$

When $P,Q\in L_1^1(\mathbf{R})$, we have

$$|f_r^{[0]}(\mu,x)|\leqslant C(1+\max\{0,x\})e^{(\mathrm{Im}\,\mu)x},\quad \mu\in\overline{C^+}, \tag{2.16}$$

$$|f_l(k,x)|\leqslant C(1+\max\{0,-x\})e^{-(\mathrm{Im}\,\mu)x},\quad k\in\mathcal{S}, \tag{2.17}$$

where C is a constant independent of x and k. Hence, by the Lebesgue dominated convergence theorem, the integral on the right-hand side in (2.15) converges as $k\to k_0$. Now (2.10) follows from (2.9), (2.15), and the fact that in the generic case we have

$$[f_l(k_0,\cdot);f_r(k_0,\cdot)]=[f_l^{[0]}(k_0,\cdot);f_r^{[0]}(k_0,\cdot)]\neq 0.$$

In the exceptional case we obtain (2.11) by using (2.15)–(2.17) along with the fact that

$$f_r^{[0]}(0,x)=f_r(k_0,x)=\frac{1}{\gamma}f_l(k_0,x), \tag{2.18}$$

and (cf. Ref. 12)

$$T^{[0]}(0)=\frac{2\gamma}{\gamma^2+1},$$

where γ is the constant in (1.8). The statement $T(k_0)=0$ in part (b) follows directly from (2.15). Turning to $L(k)$, from (2.4) we get

$$\frac{L(k)}{T(k)}=\frac{L^{[0]}(\mu)}{T^{[0]}(\mu)}+\frac{k-k_0}{2\mu}\int_{-\infty}^{\infty}dx\,P(x)f_r^{[0]}(-\mu,x)f_l(k,x). \tag{2.19}$$

Using (2.16) and (2.17) one can show that the integral in (2.19) has a finite limit as $k\to k_0$. In the generic case, we have $L^{[0]}(0)=-1$ and

$$\lim_{k\to k_0}\frac{T(k)}{T^{[0]}(\mu)}=\frac{[f_l^{[0]}(0,\cdot);f_r^{[0]}(0,\cdot)]}{[f_l(k_0,\cdot);f_r(k_0,\cdot)]}=1,$$

$$\lim_{k\to k_0}\frac{T(k)}{\mu}=\frac{-2i}{[f_l(k_0,\cdot);f_r(k_0,\cdot)]}.$$

Thus (2.19) implies that $L(k_0)=-1$. To prove (2.12) we use (2.18), (2.19), and the fact that in the exceptional case we have

$$L^{[0]}(0)=\frac{\gamma^2-1}{\gamma^2+1}.$$

The arguments leading to (2.13) and in case (b) are similar. ■

The implications of Theorem 2.2 for the special cases (A)–(C) are as follows.

Corollary 2.3: Suppose $P,Q\in L_1^1(\mathbf{R})$ and $F(k)$ satisfies (H1) and (H2). If $k_0\in\mathbf{C}$ is an exceptional value, then we have the following.

(i) In case (A), we have $k_0=0$ and

$$T(k_0)=\frac{2\gamma}{\gamma^2+1},\quad L(k_0)=\frac{\gamma^2-1}{\gamma^2+1},\quad R(k_0)=\frac{1-\gamma^2}{\gamma^2+1}. \tag{2.20}$$

(ii) In case (B), we have $k_0=0$, $\mu(k)=k$, $\alpha=1$, and (2.11)–(2.13) hold.

(iii) In case (C), $k_0=\pm i/(2\beta)$ and $F(k)$ vanishes linearly at k_0; hence $\alpha=0$. In this case (2.20) holds.

Next we address the question of where in the complex plane the possible exceptional k-values can occur. Of course, in order for a k-value to correspond to the exceptional case, it must first be a critical value, and this depends on $F(k)$. In the next proposition, without referring to any specific form of $F(k)$, we present some sufficient conditions which ensure that the exceptional k-values cannot occur off the imaginary axis.

Proposition 2.4: Assume $P(x)\not\equiv 0$, $Q(x)$ and $P(x)$ are real valued, and $P,Q\in L_1^1(\mathbf{R})$. If k_0 is an exceptional value but not purely imaginary, then $\int_{-\infty}^{\infty}dx\,P(x)|f_1(k_0,x)|^2=0$. If $Q(x)\geqslant 0$, or $P(x)\leqslant 0$, or $P(x)\geqslant 0$, then the exceptional k-values for (1.1) can occur only on the imaginary axis.

Proof: Recall that in the exceptional case the Jost solutions of (1.1), $f_l(k_0,x)$ and $f_r(k_0,x)$, are linearly dependent and hence $f_l(k_0,x)$ remains bounded as $x\to\pm\infty$. Moreover, since $F(k_0)=0$, one can show [cf. (2.11) of Ref. 14] that $f_l'(k_0,x)=o(1/x)$ as $x\to\pm\infty$. Thus, from (1.1), after integrating by parts and using

$$\lim_{x\to\pm\infty} f_l'(k_0,x)f_l(k_0,x)^*=0,$$

where $*$ denotes complex conjugation, we obtain

$$\int_{-\infty}^{\infty}dx\,|f_l'(k_0,x)|^2+\int_{-\infty}^{\infty}dx\,Q(x)|f_l(k_0,x)|^2=-ik_0\int_{-\infty}^{\infty}dx\,P(x)|f_l(k_0,x)|^2.$$

Since the right-hand side has to be real, both assertions follow. ■

III. SPECIAL CASE (C) AND EXAMPLES

We first consider case (C) in some more detail and discuss the implications of our results for the work of Sattinger and Szmigielski.[9] To establish the connection between the notation used here and that used in Ref. 9, we note that in Ref. 9 the special case $F(k)=k^2+1$ was considered with the notation $E^2=k^2+1$ (i.e., E in Ref. 9 corresponds to μ here), and a complex uniformization parameter z was used to express E and k as

$$E=\frac{1}{2}\left(z+\frac{1}{z}\right),\quad k=\frac{1}{2}\left(z-\frac{1}{z}\right).$$

Then two sets of solutions of (1.1), $\psi_\pm(x,z)$ and $\phi_\pm(x,z)$, having specific asymptotic behaviors were defined. We state here only their connection with the Jost solutions of (1.1). We have

$$\psi_+(x,z)=f_l(k,x),\quad \psi_-(x,z)=T(k)f_r(k,x)-R(k)f_l(k,x),$$

$$\phi_+(x,z)=f_r(k,x),\quad \phi_-(x,z)=T(k)f_l(k,x)-L(k)f_r(k,x).$$

These definitions imply that

$$\phi_+(x,z)=a(z)\psi_-(x,z)+b(z)\psi_+(x,z),$$

$$\phi_-(x,z)=c(z)\psi_-(x,z)+d(z)\psi_+(x,z),$$

with

$$a(z)=\frac{1}{T(k)},\quad b(z)=\frac{R(k)}{T(k)},\quad c(z)=-\frac{L(k)}{T(k)},\quad d(z)=\frac{T(k)^2-L(k)R(k)}{T(k)}.$$

The quantities

$$r_+(z)=\frac{b(z)}{a(z)},\quad r_-(z)=\frac{c(z)}{d(z)},$$

were called generalized reflection coefficients in Ref. 9. In terms of our scattering coefficients we have

$$r_+(z)=R(k),\quad r_-(z)=\frac{L(k)}{L(k)R(k)-T(k)^2}.$$

Now let us apply Theorem 2.2 and Corollary 2.3 to the problem studied in Ref. 9. The critical points are $k=\pm i$, corresponding to $z=\pm i$. In the notation of Ref. 9, generically one has $r_\pm(i)=r_\pm(-i)=-1$; on the other hand, in the exceptional case, one has

$$r_+(\pm i)=\frac{1-\gamma_\pm^2}{\gamma_\pm^2+1},\quad r_-(\pm i)=\frac{1-\gamma_\pm^2}{\gamma_\pm^2+1}.$$

Here γ_+ and γ_- are the constants in (1.8) at the critical points i and $-i$, respectively. Thus, we see that potentials need not necessarily be reflectionless in order to violate $r_+(\pm i)=-1$ or $r_-(\pm i)=-1$. In fact, in the next example we show that even rather simple potentials may cause nontrivial reflection in the exceptional case. The following examples involve potentials of the form

$$P(x)=\begin{cases} b_+, & 0<x<1,\\ b_-, & -1<x<0,\\ 0, & \text{elsewhere,}\end{cases}\qquad Q(x)=\begin{cases} a_+, & 0<x<1,\\ a_-, & -1<x<0,\\ 0, & \text{elsewhere,}\end{cases}\tag{3.1}$$

where $a_\pm$ and $b_\pm$ are parameters.

Example 3.1: In (3.1) let us use $b_+=2$, $b_-=b$ with $b\geqslant 0$, $a_+=1$, $a_-=0$, and choose $F(k)=k^2+1$. We can solve (1.1) and evaluate the scattering coefficients explicitly. The critical points are $k=\pm i$. Letting $k=i(1-\epsilon)$, as $k\to i$ so that $\epsilon\to 0$ through positive values, we obtain

$$\frac{2i\mu}{T(k)}=\sqrt{b}\cos 1\sin\sqrt{b}+\sin 1\cos b+O(\sqrt{\epsilon}).\tag{3.2}$$

There are an infinite number of positive b-values that cause the leading term in (3.2) to vanish, and each such b-value causes $k=i$ to yield the exceptional case. The smallest is $b=6.77194$, the next two values are $b=36.3663$ and $b=85.7127$ (the overline means that the last digit may have been affected by round-off). For $b=6.77194$ we get $T(i)=-0.901744$ and $L(i)=-0.432166$; for $b=36.3663$ we get $T(i)=0.851046$ and $L(i)=-0.525091$; and for $b=85.7127$ we get $T(i)=-0.842793$ and $L(i)=-0.538224$.

Example 3.2: In (3.1) let $b_+=b_-=b$, $a_+=a_-=0$, and choose $F(k)=k^2$. Then we are in the exceptional case for every $b\geqslant 0$. The only critical value is $k_0=0$ and we have $f_l(k_0,x)=1$ and $\gamma=1$ [cf. (1.8)], and

$$\frac{1}{T(k)} = e^{2ik}\left[\cos(2\sigma) - \frac{2ik+b}{2\sigma}\sin(2\sigma)\right], \tag{3.3}$$

where $\sigma = \sqrt{k^2 - ikb}$. Hence

$$\frac{1}{T(0)} = 1 - b,$$

which is in agreement with (2.11). If $b=1$, then $\omega=0$ and $\alpha=1$ in Theorem 2.2, and (3.3) gives

$$T(k) = \frac{3i}{2k} + O(1), \quad k \to 0.$$

This shows that $T(k)$ can be discontinuous at a critical value.

We conclude with two examples illustrating the location of possible exceptional k-values; in these examples, unless otherwise indicated, $F(k)$ is not assumed to have any special form.

Example 3.3: In (3.1) let $b_+ = 1$ and $b_- = a_+ = a_- = -1$. Setting $k = k_0$ and $F(k_0) = 0$ we solve (1.1) to find the Jost solution $f_l(k_0, x)$ and then impose the condition that $f_l(k_0, x)$ be bounded as $x \to -\infty$; that is, we demand that $f_l'(k_0, -1) = 0$. This is a necessary condition for k_0 to be an exceptional value for any given function $F(k)$. A straightforward calculation shows that the (possibly) exceptional values are given by the solutions of the equation

$$\sqrt{-1+ik_0}\tanh\sqrt{-1+ik_0} + \sqrt{-1-ik_0}\tanh\sqrt{-1-ik_0} = 0.$$

This equation has infinitely many roots on the imaginary axis located symmetrically about the origin and, as can be seen numerically, one symmetric pair of roots on the real axis. The two real roots are $k_0 = \pm 1.355$, and the imaginary roots closest to zero are $k_0 = \pm 14.139i$. The corresponding Jost solution $f_l(k_0, x)$ is given by

$$f_l(k_0, x) = \begin{cases} \cosh(\sqrt{-1+ik_0}(1-x)), & 0 \leqslant x \leqslant 1, \\ \dfrac{\cosh\sqrt{-1+ik_0}}{\cosh\sqrt{-1-ik_0}}\cosh(\sqrt{-1-ik_0}(x+1)), & -1 \leqslant x \leqslant 0, \end{cases}$$

and on each of the intervals $(-\infty, -1)$ and $(1, +\infty)$, $f_l(k_0, x)$ is constant and obtained by continuity. This example shows the possibility of real as well as purely imaginary exceptional values. The two imaginary roots above would be critical values for case (C) if $\beta = 0.035$. The two real roots are not critical values for any of cases (A)–(C). In accordance with Proposition 2.4, one can verify that $\int_{-1}^{1} dx\, P(x) |f_l(k_0, x)|^2 = 0$.

Example 3.4: In (3.1) let $b_+ = b_- = 1$, $a_+ = 0$, and $a_- = -1$. Then the (possibly) exceptional values satisfy

$$\sqrt{ik_0}\tanh\sqrt{ik_0} + \sqrt{-1-ik_0}\tanh\sqrt{-1-ik_0} = 0.$$

There are again infinitely many purely imaginary roots; there are also complex roots, one pair of which is $k_0 = \pm 1.1008 + 0.5i$. The corresponding Jost solution $f_l(k_0, x)$ is given by

$$f_l(k_0, x) = \begin{cases} \cosh(\sqrt{ik_0}(1-x)), & 0 \leqslant x \leqslant 1, \\ \dfrac{\cosh\sqrt{ik_0}}{\cosh\sqrt{-1-ik_0}}\cosh(\sqrt{-1-ik_0}(x+1)), & -1 \leqslant x \leqslant 0, \end{cases}$$

and for $|x|>1$, $f_l(k_0,x)$ is constant and obtained by continuity. This example shows the possibility of exceptional values that are neither real nor purely imaginary. In case (C) only the purely imaginary roots could be critical values for suitable β.

ACKNOWLEDGMENT

The research leading to this article was supported in part by the National Science Foundation under Grant Nos. DMS-9501053 and DMS-980321.

[1] M. Jaulent and C. Jean, ''The inverse *s*-wave scattering problem for a class of potentials depending on energy,'' Commun. Math. Phys. **28**, 177–220 (1972).
[2] M. Jaulent and C. Jean, ''The inverse problem for the one-dimensional Schrödinger equation with an energy-dependent potential. I,'' Ann. Inst. Henri Poincaré, Sect. A **25**, 105–118 (1976).
[3] M. Jaulent and C. Jean, ''The inverse problem for the one-dimensional Schrödinger equation with an energy-dependent potential. II,'' Ann. Inst. Henri Poincaré, Sect. A **25**, 119–137 (1976).
[4] M. Jaulent, ''Inverse scattering problems in absorbing media,'' J. Math. Phys. **17**, 1351–1360 (1976).
[5] D. H. Sattinger and J. Szmigielski, ''Energy dependent scattering theory,'' Differ. Integral Eqs. **8**, 945–959 (1995).
[6] T. Aktosun, M. Klaus, and C. van der Mee, ''Wave scattering with absorption,'' J. Math. Phys. **39**, 1957–1992 (1998).
[7] J. Kaup, ''A higher-order water-wave equation and the method for solving it,'' Prog. Theor. Phys. **54**, 396–408 (1975).
[8] M. Tsutsumi, ''On the inverse scattering problem for the one-dimensional Schrödinger equation with an energy dependent potential,'' J. Math. Anal. Appl. **83**, 316–350 (1981).
[9] D. H. Sattinger and J. Szmigielski, ''A Riemann-Hilbert problem for an energy dependent Schrödinger operator,'' Inverse Probl. **12**, 1003–1025 (1996).
[10] L. D. Faddeev, ''Properties of the *S*-matrix of the one-dimensional Schrödinger equation,'' Am. Math. Soc. Transl. **2**, 139–166 (1964) [Trudy Mat. Inst. Steklova **73**, 314–336 (1964) (in Russian)].
[11] P. Deift and E. Trubowitz, ''Inverse scattering on the line,'' Commun. Pure Appl. Math. **32**, 121–251 (1979).
[12] M. Klaus, ''Low-energy behaviour of the scattering matrix for the Schrödinger equation on the line,'' Inverse Probl. **4**, 505–512 (1988).
[13] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. (Springer-Verlag, New York, 1989).
[14] T. Aktosun, M. Klaus, and C. van der Mee, ''On the Riemann-Hilbert problem for the one-dimensional Schrödinger equation,'' J. Math. Phys. **34**, 2651–2690 (1993).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

On the dynamics of the Holstein model from the anticontinuous limit

Dario Bambusi[a)]
Dipartimento di Matematica dell'Università, Via Saldini 50, 20133 Milano, Italy

(Received 17 December 1998; accepted for publication 18 March 1999)

We consider the Holstein model describing an electron interacting with a lattice of identical oscillators. We remark that the on site system (i.e., the system in which the interaction between the different sites of the lattice vanishes) is integrable and anisocronous. This allows us to apply some recent Nekhoroshev-type results to show that corresponding to the majority of initial data in which the electron probability is concentrated on a finite number of sites, the electron probability distribution is approximatively constant for times growing exponentially with the inverse of the coupling parameter. Moreover, for the same times, the total energy of the oscillator system is approximatively constant. © *1999 American Institute of Physics.* [S0022-2488(99)00308-4]

I. INTRODUCTION

In this paper we consider the adiabatic Holstein model describing one electron in interaction with an infinite lattice of identical oscillators. It will be studied as a classical (infinite dimensional) nonlinear dynamical system. Scaling suitably time and the oscillator's variables, the equations of the system can be written in the form

$$\left.\begin{array}{c} -i\dot{\psi}_i=-q_i\psi_i-\epsilon[(\psi_i-\psi_{i-1})+(\psi_i-\psi_{i+1})] \\ \ddot{q}_i=-\omega^2 q_i+|\psi_i|^2 \end{array}\right., \quad i\in\mathbf{Z} \tag{1.1}$$

where q_i are the oscillator's variables, and $|\psi_i|^2$ is the probability of finding the electron at site i; ϵ is the coupling between the different sites, and ω the frequency of the oscillators.

The present work was stimulated by a paper by Hennig,[1] where the dynamics of the Holstein model in the case $\omega \gg 1$ was studied. Hennig applied the Nekhoroshev-type techniques of Ref. 2 to show that if ω is large enough, then there is no exchange of energy between the oscillator's system and the electron system for times growing exponentially with ω. He also presented some numerical simulations showing that if ϵ is small enough also each of the probability amplitudes $|\psi_i|^2$ remain constant. Finally he gave an heuristic explanation of such a phenomenon.

In the present paper we prove rigorously, at least in some quite general situation, and for long times, the constancy of the electron probability distribution, i.e., of the quantities $|\psi_i|^2$. More precisely, we will consider the class of initial data in which the electron probability distribution $|\psi_i|^2$ is essentially localized at a finite number of sites. We will prove that, provided ϵ is below some threshold (which could be explicitly computed), then corresponding to the majority of such initial data the quantities $|\psi_i|^2$ are approximately constant, and moreover the total harmonic energy of the oscillators is approximately constant for very long times, namely, for times of order $\exp(C\epsilon^{-a})$ with positive C and a.

The result is obtained in two steps. First, we consider the anticontinuous limit $\epsilon=0$ in which the system is decoupled into infinitely many identical systems with two degrees of freedom (on site system). Our starting point is the remark that the on site system is an integrable Hamiltonian

[a)]Electronic mail: BAMBUSI@MAT.UNIMI.IT

system. So, first of all we introduce its action angle variables. In this way it becomes clear that the on site system is anysocronous, i.e., the frequencies of its motion depend on the initial data. We come to the second step; to explain in a simple way the key idea consider initial data in which the electron probability is concentrated at one single site of the lattice, say the site $\bar{i}$. It turns out that, in the on site system, the phase of the electron's probability $\psi_{\bar{i}}$ rotates with a frequency which depends on the initial datum and therefore, in general, is nonresonant with ω. Moreover, the phases related to ψ_i $i\neq\bar{i}$ do not move at all. Then one can hope to use perturbation theory to show that also when $\epsilon\neq 0$ the total energy of the oscillator's system and the electron probabilities $|\psi_i|^2$ are essentially constant. The situation is similar when the electron probability is concentrated at a finite number of sites.

To show that this is actually what happens we use a Nekhoroshev-type result which was recently proved by the author[3] (for an improved version applicable to PDE's, see Ref. 4). In Ref. 3 a perturbation of an infinite dimensional system with $n<\infty$ frequencies was considered, and it was proven that there exists n functions which remain approximatively constant when the system is subjected to a smooth perturbation. In the present case such functions are $n-1$ electron probabilities and the total harmonic energy of the oscillator system. Exploiting also the conservation of the total probability of finding an electron in the lattice we are able to conclude that the whole probability distribution of the electron is approximatively constant for the considered times.

Finally we recall that Holstein model was already studied from the anticontinuous limit by Aubry[5] who proved that, provided ϵ is small enough, the Holstein model has quasiperiodic breathers, i.e., quasiperiodic localized solutions. Such solutions physically correspond to the situation in which one of the oscillators perform large amplitude oscillations while the other is approximatively at rest; the electron probability distribution is constant, but the phases rotate uniformly.

The paper is organized as follows: in Sec. II we discuss the on site dynamics; in Sec. III we recall the main abstract theorem of Ref. 3, state and prove the main result of the present paper (freezing of the electron's probabilities, and of the oscillator's energy). Finally in Sec. IV we discuss some possible extensions of these results.

II. THE ON SITE DYNAMICS

Consider the adiabatic Holstein model (1.1), define x_i, and y_i as the real and imaginary parts of ψ_i; then it is easy to verify that (1.1) are the Hamilton's equations of

$$H:=\sum_i H_{os}(p_i,q_i,x_i,y_i)+\frac{1}{2}\epsilon\sum_i [(x_i-x_{i-1})^2+(y_i-y_{i-1})^2], \tag{2.1}$$

where

$$H_{os}(p_i,q_i,x_i,y_i):=\frac{p_i^2+\omega^2 q_i^2}{2}+q_i\frac{x_i^2+y_i^2}{2}, \tag{2.2}$$

and $(p_k,q_k),(x_i,y_i)$ are canonically conjugated variables.

Remark 2.1: The quantity $\Sigma_i(x_i^2+y_i^2)/2$ *is a constant of motions for the dynamics of the Holstein model. It represents the probability of finding an electron in the lattice, so it should be subjected to the normalization condition*

$$\sum_i \frac{x_i^2+y_i^2}{2}=1.$$

For simplicity in the following we will drop this condition.

We begin by studying the on site dynamics described by the Hamiltonian (2.2). It is immediate to realize that the system (2.2) is integrable. Indeed the action $(x_i^2+y_i^2)/2$, is a constant of motion, and the Hamiltonian itself is the second constant.

Fix *i*, then one can explicitly construct action angle variables for H_{os} in two steps. First define the variables (J_i, ϕ_i) by

$$x_i = \sqrt{J_i}\cos\phi_i\,, \quad y_i = \sqrt{J_i}\sin\phi_i\,;$$

the Hamiltonian of the on site system takes the form

$$H_{os} = \frac{p_i^2 + \omega^2 q_i^2}{2} + q_i J_i\,. \tag{2.3}$$

Then the canonical transformation

$$Q_i = q_i + \frac{J_i}{\omega^2}, \quad P_i = p_i\,, \quad J_i' = J_i\,, \quad \phi_i' = \phi_i + \frac{P_i}{\omega^2}$$

(with generating function $P_i(q_i + (J_i'/\omega^2)) + J_i'\phi_i$) reduces the Hamiltonian (2.3) to the form

$$\frac{P_i^2 + \omega^2 Q_i^2}{2} - \frac{1}{2\omega^2} J_i^2, \tag{2.4}$$

where we omitted the prime from *J*. The main feature of this Hamiltonian is that it clearly shows that it is integrable and that the frequency of the electron's system is J_i/ω^2, which depends on the electron density at the considered site.

Remark 2.2: The quantity

$$J_i = \frac{x_i^2 + y_i^2}{2}$$

is the probability of finding an electron at the site i of the lattice.

The integration of this system is immediate and gives

$$P_i(t) = P_{i,0}\cos(\omega t) - \omega Q_{i,0}\sin(\omega t),$$

$$Q_i(t) = Q_{i,0}\cos(\omega t) + \frac{P_{i,0}}{\omega}\sin(\omega t),$$

$$J_i(t) = J_{i,0}, \qquad \phi_i'(t) = \frac{J_i}{\omega^2}t + \phi_{i,0}'.$$

So, along with the solutions the electron density J_i is a constant of motion, the corresponding phase ϕ' rotates uniformly, and the variables P_i, Q_i oscillate with frequency ω around zero. In terms of the original variables the solution appears as follows: the phase of the electron rotates and moreover performs an oscillation with frequency ω, while the oscillator still oscillates with frequency ω, but around the shifted equilibrium position $q_i = -J_i/\omega^2$.

III. THE COMPLETE SYSTEM

We use now perturbation theory to study the dynamics of the complete system. We will concentrate on the situation in which the electron's probability is concentrated at a finite number of sites of the lattice. Moreover, in order to avoid the singularity introduced by the action angle coordinates we prefer to give the main statement in terms of the original variables *p,q,x,y*. So, in the statement of the forthcoming theorem J_i must be interpreted as a short notation for $(x_i^2 + y_i^2)/2$; moreover we will denote by

$$h_0 := \sum_i \frac{1}{2}\left(p_i^2 + \omega^2\left(q_i + \frac{J_i}{\omega^2}\right)^2\right)$$

the total harmonic energy of the oscillators.

We fix a subset $\mathcal{L} \subset \mathbf{Z}$ with cardinality $n < \infty$ and an n-dimensional vector $\nu = \{\nu_i\}_{i \in \mathcal{L}}$ which fulfills the following diophantine condition

$$|\nu \cdot k + \omega l| \geq \frac{\gamma}{|(k,l)|^\tau}, \quad \forall (k,l) \in \mathbf{Z}^{n+1} \setminus \{0\}, \tag{3.1}$$

where $|(k,l)| := |l| + \Sigma |k_i|$, and γ and τ are strictly positive constants.

We will prove the following:

Theorem 3.1: *There exists a positive* ϵ_**, and positive constants* $C_1, \ldots, C_4$ *with the following properties: assume* $|\epsilon| < \epsilon_*$*, and consider an initial datum satisfying*

$$|J_i(0) - \nu_i \omega^2| \leq C_1 \epsilon, \quad \forall i \in \mathcal{L}; \quad \sum_{i \notin \mathcal{L}} J_i(0) \leq C_1 \epsilon, \quad h_0(0) < \infty; \tag{3.2}$$

then, along the corresponding solution one has

$$|J_i(t) - J_i(0)| \leq C_2 \epsilon^{1/2(\tau+1)} \epsilon, \quad \forall i \in \mathcal{L},$$

$$\sum_{i \notin \mathcal{L}} J_i(t) \leq C_2 \epsilon, \quad |h_0(t) - h_0(0)| \leq C_2 \epsilon^{1/2(\tau+1)} h_0(0), \tag{3.3}$$

for exponentially long times, namely, for the times t fulfilling

$$|t| \leq C_3 \exp\left(\frac{C_4}{\epsilon}\right)^{1/2(\tau+1)}. \tag{3.4}$$

The proof will be obtained by applying the normal form theorem of Ref. 3. Before going to the details we recall the result of Ref. 3 that will be applied here. We simplify slightly its statement since for the present application we need only to deal with a strongly symplectic space.

Consider a strongly symplectic Banach space $\mathcal{P}$, i.e., a Banach space endowed by a skew symmetric nondegenerate form Ω. Then one can define the application $\mathcal{P} \ni X \mapsto \Omega(X,.) \in \mathcal{P}^*$, which is injective. We recall that Ω is said to be strongly nondegenerate if such a map is also surjective. In this case the Poisson tensor J which is the inverse of this map is a linear *bounded* operator. It follows that the Hamiltonian vector field X_f of a C^∞ function f (which is defined by $X_f := J df$) is also of class C^∞. Consider the complexification $\mathcal{P}^{\mathbf{C}}$ of $\mathcal{P}$; we will denote by $B_R(z) \subset \mathcal{P}^{\mathbf{C}}$ the ball of radius R with center at z. Moreover, given a set $\mathcal{G} \subset \mathcal{P}$ we will denote

$$\mathcal{G}_R := \bigcup_{z \in \mathcal{G}} B_R(z).$$

On a set $\mathcal{G}$ consider a Hamiltonian function

$$H = h_\omega + f. \tag{3.5}$$

Fix a positive R; we will assume that both h_ω and f extend to complex analytic functions on $\mathcal{G}_R$. The main abstract result of Ref. 3 ensures that, if h_ω generates a quasiperiodic flow which leaves invariant $\mathcal{G}_R$, and f is small, the system has a certain number of approximate constants of motion, for times growing exponentially with a power of the inverse of the size of f.

To give a precise statement we write more explicitly the assumptions. We will denote by Φ_t the flow generated by the Hamiltonian vector field of h_ω. We assume

(1) $\Phi_t(\mathcal{G}_r)\subset\mathcal{G}_r$, for any $r\leq R$;

(2) Φ_t is quasiperiodic, namely, there exists a group action $\Psi:\mathbf{T}^n\times\mathcal{G}_R\to\mathcal{G}_R$ and an n dimensional real vector ω such that

$$\Phi_t(z)=\Psi_{\omega_1 t,\omega_2 t,\ldots,\omega_n t}(z),\ \ \forall z\in\mathcal{G}_R .$$

Moreover, there exists functions $I_1,\ldots,I_n$ analytic in $\mathcal{G}_R$ whose Hamiltonian vector fields generate the flows $\Psi_{\phi_1,0,0,\ldots,0}\cdots\Psi_{0,0,\ldots,0,\phi_n}$.

(3) The frequency vector is diophantine, namely there exist positive γ and τ such that

$$|\omega\cdot k|\geq\frac{\gamma}{|k|^\tau},\quad \forall k\in\mathbf{Z}^n\setminus\{0\}.$$

(4) Denote by $\mathbf{T}^n+i\sigma$ the set of the ϕ's belonging to the complexified torus such that $|\mathrm{Im}\,\phi_i|\leq\sigma$. There exist positive σ and $r_*>R/2$ such that Ψ extends to an analytic map $\Psi:\mathbf{T}^n+i\sigma\times\mathcal{G}_{r_*}\to\mathcal{G}_R$.

(5) Finally we assume that there exists a finite constant ϵ_f such that

$$\frac{1}{R}\sup_{z\in\mathcal{G}_R}\|X_f(z)\|\leq\epsilon_f . \tag{3.6}$$

Then Corollary 4.2 of Ref. 3 states that

Theorem 3.2: *Under the above assumptions there exists positive constants ϵ_*, C_5,C_6,C_7, independent of R, such that, provided $\epsilon_f\leq\epsilon_*$, then along the solutions of the Cauchy problem for (3.5), with initial data in $\mathcal{G}_{R/4}$, one has*

$$|I_i(z(t))-I_i(z(0))|\leq C_5\epsilon_f^{1/(\tau+1)}\sup_{z\in\mathcal{G}_R}|I_i(z)|,\quad i=1,\ldots,n$$

for all times $t\leq\min\{T_0,T_*\}$, where T_0 is the escape time of the solution from the domain $\mathcal{G}_{3R/8}$, and

$$T_*:=\frac{C_7}{\epsilon_f}\exp\left[\frac{C_6}{\epsilon_f}\right]^{1/(\tau+1)}.$$

Proof of theorem 3.1: We will apply the above theorem 3.2, and then we will show that the escape time T_0 is larger than T_*. We prepare the system by introducing ''action angle'' variables $(J_i,\phi_i),(P_i,Q_i)$ for the on site system at the sites $i\in\mathcal{L}$; making a translation of the action J_i, namely, introducing the variables

$$I_i:=J_i-\nu_i\omega^2,$$

the Holstein model assumes the form

$$H=h_\omega+f_1+f_2+f_3,$$

where

$$h_\omega:==h_0-\sum_i\ \nu_i I_i,$$

$$f_1:=\epsilon H_{\mathrm{int}},\quad f_2:=-\sum_{i\in\mathcal{L}}\frac{I_i^2}{2\omega^2},\ \ f_3:=-\sum_{i\notin\mathcal{L}}\frac{J_i^2}{2\omega^2},$$

where

$$H_{\rm int}=\frac{1}{2}\sum_i \,[(x_i-x_{i-1})^2+(y_i-y_{i-1})^2],$$

and, for $i\in\mathcal{L}$, x_i and y_i must be intended as functions of the action angle variables, while, for $i\notin\mathcal{L}$, the symbols P_i, Q_i, and J_i must be intended as short notations for their complete expression. We fix a parameter R and define the norm of a sequence

$$z=(\{P_i,Q_i,I_i,\phi_i\}_{i\in\mathcal{L}},\{p_i,q_i,x_i,y_i\}_{i\notin\mathcal{L}})$$

by

$$\|z\|^2:=\sum_{i\in\mathcal{L}}\frac{|P_i|^2+\omega^2|Q_i|^2}{2}+\sum_{i\notin\mathcal{L}}\frac{|p_i|^2+\omega^2|q_i|^2}{2}+\max_{i\in\mathcal{L}}|I_i|^2+\max_{i\in\mathcal{L}}R^2|\phi_i|^2+R\sum_{i\notin\mathcal{L}}\frac{|x_i|^2+|y_i|^2}{2}$$

and define $\mathcal{P}$ as the completion of set of sequences with compact support in this norm. We choose the domain $\mathcal{G}\subset\mathcal{P}$ to be the set

$$\left\{z\in\mathcal{P}:h_0(z)\leqslant 2E,\quad I_i=0,\quad \forall i\in\mathcal{L},\quad J_i=\frac{x_i^2+y_i^2}{2}=0,\quad \forall i\notin\mathcal{L}\right\}.$$

And we extend it to the complex using again the parameter R.

It is now easy to compute ϵ_f. To this end remark that $H_{\rm int}$ is a bounded polynomial of the variables x,y, so it is analytic as a function of these variables. Since the change of variables involving action angle variables is also analytic, $H_{\rm int}$ is analytic also as a function of these variables. Moreover $H_{\rm int}$ is independent of the p,q variables. It follows that there exists a constant independent of ϵ and R which bounds the norm of the Hamiltonian vector field of f_1. Concerning the vector field of f_2, it has only ϕ components. It is linear in J, and therefore it is bounded by a constant times R^2. Finally consider f_3; the corresponding vector field is a cubic polynomial in x,y and therefore the square of its norm is bounded by a constant times R times R^3. It follows that we have

$$\frac{1}{R}\|X_f\|\leqslant C\left(\frac{\epsilon}{R}+\epsilon+R\right)$$

with a suitable constant C. Choosing $R=\sqrt{\epsilon}$ we have

$$\epsilon_f=C'\sqrt{\epsilon}. \tag{3.7}$$

We define now the group action Ψ. We denote by ψ_1,ψ_2,ψ_{n+1} the angles on the torus. Then $(P,Q,J,\phi,p,q,x,y):=\Psi_{\psi_1,\psi_2,\psi_{n+1}}(P_0,Q_0,J_0,\phi_0,p_0,q_0,x_0,y_0)$ is defined by

$$I_i=I_{i,0},\quad i\in\mathcal{L},\quad \phi_i(\psi_{j_i})=\phi_{0,i}+\psi_{j_i},\quad i\in\mathcal{L},$$

$$P_i(\psi_1)=P_{0,i}\cos\psi_1-\omega Q_{0,i}\sin\psi_1,\quad i\in\mathcal{L},$$

$$Q_i(\psi_1)=Q_{0,i}\cos\psi_1+\frac{P_{0,i}}{\omega}\sin\psi_1,\quad i\in\mathcal{L},$$

$$q_i(\psi_1)=\left(q_{0,i}+\frac{J_i^2}{\omega^2}\right)\cos\psi_1+\frac{p_{0,i}}{\omega}\sin\psi_1-\frac{J_i}{\omega^2},\quad i\notin\mathcal{L},$$

$$p_i(\psi_1)=-\omega\left(q_{0,i}+\frac{J_i}{\omega^2}\right)\sin\psi_1+\frac{p_{0,i}}{\omega}\cos\psi_1,\quad i\notin\mathcal{L},$$

$$x_i(\psi_1)=x_{0,i}\cos\left(\frac{p_i(\psi_1)-p_{i,0}}{\omega^2}\right)+y_{0,i}\sin\left(\frac{p_i(\psi_1)-p_{i,0}}{\omega^2}\right),\quad i\notin\mathcal{L},$$

$$y_i(\psi_1)=x_{0,i}\sin\left(\frac{p_i(\psi_1)-p_{i,0}}{\omega^2}\right)+y_{0,i}\cos\left(\frac{p_i(\psi_1)-p_{i,0}}{\omega^2}\right),\quad i\notin\mathcal{L},$$

where j_i is defined by $j_i\in[2,...,n+1]$, $i<i'$ implies $j_i<j_{i'}$.

So it is possible to apply theorem 3.2 obtaining the estimates (3.3) for times which are the minimum between the escape time from $\mathcal{G}_{3R/8}$ and the times (3.4). To bound the escape time remark that the function h_0 can be used as a Lyapunov functions to show that the variables P,Q cannot leave their domains for the times we are interested in. The same is obviously true for the variables I_i. To bound also the escape time for the variables $\{x_i,y_i\}_{i\notin\mathcal{L}}$ just use the information that $\Sigma_i(x_i^2+y_i^2)/2$ is a constant of motion. This allows us to conclude that

$$\sum_{i\notin\mathcal{L}}\frac{x_i(t)^2+y_i(t)^2}{2}\leqslant\sum_{i\notin\mathcal{L}}\frac{x_i(0)^2+y_i(0)^2}{2}+\sum_{i\in\mathcal{L}}|I_i(t)-I_i(0)|$$

but the latter quantity is estimated by (3.3), and therefore also the variables x_i,y_i cannot escape from their domain. So the theorem is proven. ■

IV. DISCUSSION

So, we have proved that, provided ϵ is small enough, and corresponding to the majority of initially localized electron probability distributions, the electron probabilities $|\psi_i|^2$ and the total energy of the oscillators are essentially constant for very long times. One can ask what happens corresponding to the other initial data.

We expect that the electron probability distribution should be essentially constant corresponding to *any* initial datum with $|\psi_i|^2$ concentrated at a finite number of sites. Namely we expect that there is no need of the diophantine condition (3.1). The reason is that the Holstein model appears as a perturbation of the Hamiltonian

$$h_0-\sum_i\frac{J_i^2}{2\omega^2},$$

which is quasiconvex as a function of J_i and h_0. For quasiconvex systems with finitely many degrees of freedom Nekhoroshev theory ensures that the actions are approximatively constant under perturbation, and for exponentially long times. This of course requires the development of the geometric part of Nekhoroshev theorem. The Holstein model however is infinite dimensional; we recall that the extension of the geometric part of Nekhoroshev's theorem to some infinite dimensional systems has been carried out (following the finite dimensional proof by Lochak[6]) Refs. 2 and 7. We expect the techniques of Ref. 7 to be applicable to the present case. However, since there is not a general theorem which is directly applicable to Holstein model, one should repeat the whole argument of Ref. 7. For this reason we limited ourselves to deal with the much simpler nonresonant case.

For what concerns probability distributions which are not concentrated, we have not a precise idea, but we expect that (in general) there should not be an analog of theorem 3.1.

In the present paper we limited ourselves to the one electron Holstein model. However it is easy to see that also the on site dynamics of the M electron Holstein model is integrable, and that its Hamiltonian, in terms of action angle variables has exactly the form (2.4). So, one can expect theorem 3.1 (and its possible extensions) to hold also in the M electron case. However, its proof is much more difficult in the case of M electrons. This is due to the singularities that appear introducing action angle variables, which are much harder in the M electron case. We expect that it should be possible to overcome such difficulties using the techniques developed by Niedermann[8] which avoid completely the introduction of action angle variables (see also Ref. 9).

Finally we remark that the case where the linear restoring force for the oscillators is substituted by a nonlinear force (as Ref. 5), can also be dealt with by the same techniques developed here. Indeed also in this case the on site system is integrable. However, since the explicit introduction of action angle variables cannot (in general) be done, the proof of theorem 3.1 would turn out to be slightly more difficult. On the contrary the case of a nonlinear interaction between different sites can be treated without additional difficulties.

[1] D. Hennig, ''Exponential stability for the energy exchange dynamics of the Holstein model, J. Math. Phys. **39**, 3568–3576 (1998).
[2] D. Bambusi and A. Giorgilli, ''Exponential stability of states close to resonance in infinite dimensional Hamiltonian systems,'' J. Stat. Phys. **71**, 569–606 (1993).
[3] D. Bambusi, ''Exponential stability of breathers in Hamiltonian chains of weakly coupled oscillators,'' Nonlinearity **9**, 433–457 (1996).
[4] D. Bambusi, ''Long time stability of some small amplitude solutions in nonlinear Schrödinger equations,'' Commun. Math. Phys. **189**, 205–226 (1997).
[5] S. Aubry, ''Breathers in nonlinear lattices: Existence, linear stability, and quantization,'' Physica D **103**, 201–250 (1997).
[6] P. Lochak, ''Canonical perturbation theory via simultaneous approximation,'' Usp. Mat. Nauk **47**, 59–140 (1992); English translation, Russ. Math. Surv. **47**, 57–133 (1992).
[7] D. Bambusi, ''Nekhoroshev theorem for small amplitude solutions in nonlinear Schrödinger equations,'' Math. Z. **230**, 345–387 (1999).
[8] L. Niederman, ''Nonlinear stability around an elliptic equilibrium point in a Hamiltonian system,'' Nonlinearity **11**, 1465–1479 (1998).
[9] J. Pöschel, ''On Nekhoroshev estimates for a nonlinear Schrödinger equation and a theorem by Bambusi'' (preprint, 1998).

Phase-space representation of quantum state vectors: The relative-state approach and the displacement-operator approach

Masashi Ban
Advanced Research Laboratory, Hitachi, Ltd., Akanuma 2520, Hatoyama, Saitama 350-0395, Japan

(Received 17 August 1998; accepted for publication 7 April 1999)

Phase-space representation of quantum state vectors has been recently formulated by means of the relative-state method developed by the present author [J. Math. Phys. **39**, 1744 (1998)]. It is, however, pointed out by Møller that the displacement-operator method provides another basis of phase-space representation of quantum state vectors [J. Math. Phys. (to appear)]. Hence the relation between the relative-state approach and the displacement-operator approach is discussed, both of which yield equivalent phase-space representations. © *1999 American Institute of Physics.* [S0022-2488(99)02407-X]

I. INTRODUCTION

Phase-space functions that represent quantum state vectors are very useful for investigating physical and chemical properties of quantum mechanical systems.[1] Furthermore, measurement processes of quantum states of light can be described in terms of phase-space functions. Hence it is important to formulate a phase-space representation of quantum state vectors from the first principles of quantum mechanics without introducing additional assumptions. The present author has recently formulated the phase-space representation by means of the relative-state method.[2] It has been shown that under certain conditions, the results become equivalent to those obtained by Torres-Vega and Frederick[3] and Harriman.[4] The phase-space representation obtained by means of the relative-state method provides the mathematical and physical basis of their results. Møller has recently pointed out that besides the relative-state method, there is another method for obtaining a phase-space representation of quantum state vectors.[5] In fact, he has formulated the phase-space representation by means of the displacement-operator method and shown that under certain conditions, the relative-state method and the displacement-operator method yield the equivalent phase-space representations. This short note that discusses the relation between these two methods is a response to the comment by Møller.[5]

II. PHASE-SPACE REPRESENTATIONS

A. The relative-state approach

We first summarize the phase-space representation of quantum state vectors formulated by means of the relative-state method.[2] Let $\mathcal{S}$ be a relevant quantum system that we describe the physical properties, and let $\mathcal{S}_r$ be a reference system. In some cases, a reference system represents a measurement apparatus for the relevant quantum system. The sets of the position eigenstates of the relevant and reference quantum systems are denoted as $\{|x\rangle | x\in \mathbf{R}\}$ and $\{|x\rangle_r | x\in \mathbf{R}\}$, where $\mathbf{R}$ stands for the set of all real numbers.[6] Then we introduce a state vector $|\omega(r,k;s)\rangle\rangle$ of the compound quantum system $\mathcal{S}+\mathcal{S}_r$,[2]

$$|\omega(r,k;s)\rangle\rangle=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dx\,e^{ikx}\left|x+\frac{1}{2}(1+s)r\right\rangle\otimes\left|x-\frac{1}{2}(1-s)r\right\rangle_r$$

$$=\frac{1}{\sqrt{2\pi}}e^{-i(1+s)kr/2}\int_{-\infty}^{\infty}dx\,e^{ikx}|x\rangle\otimes|x-r\rangle_r, \tag{1}$$

which is the simultaneous eigenstate of operators $\hat{x}\otimes\hat{1}_r-\hat{1}\otimes\hat{x}_r$ and $\hat{p}\otimes\hat{1}_r+\hat{1}\otimes\hat{p}_r$ with eigenvalues r and k. The state vector $|\omega(r,k;s)\rangle\rangle$ is called the relative-state vector.[2]

Since we investigate the properties of the relevant quantum system $\mathcal{S}$ and the reference quantum system $\mathcal{S}_r$ is irrelevant, we fix some state vector $|\phi\rangle_r$ of the reference quantum system that may be arbitrarily chosen. In this case, the relevant quantum system is described in terms of the reduced relative-state vectors,

$$|\omega(r,k;s)\rangle={}_r\langle\phi|\omega(r,k;s)\rangle\rangle=\frac{1}{\sqrt{2\pi}}e^{-i(1+s)kr/2}\int_{-\infty}^{\infty}dx\,e^{ikx}\phi^*(x-r)|x\rangle, \tag{2}$$

where $\phi(x)={}_r\langle x|\phi\rangle_r$ is an arbitrary wave function normalized as $\int_{-\infty}^{\infty}dx\,|\phi(x)|^2=1$. It is easy to see that the set of the reduced relative-state vectors, $\{|\omega(r,k;s)\rangle|r,k\in\mathbf{R}\}$, becomes an overcomplete system.[2] Hence any quantum state vector $|\psi\rangle$ of the relevant system is represented by the phase-space function $\psi_s(r,k;s)=\langle\omega(r,k;s)|\psi\rangle$ which is normalized as $\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk|\psi_s(r,k)|^2=1$. The position and momentum operators, $\hat{x}$ and $\hat{p}$, are represented by

$$\langle\omega(r,k;s)|\hat{x}|\psi\rangle=\left[\frac{1}{2}(1+s)r+i\frac{\partial}{\partial k}\right]\psi_s(r;k), \tag{3}$$

$$\langle\omega(r,k;s)|\hat{p}|\psi\rangle=\left[\frac{1}{2}(1-s)r-i\frac{\partial}{\partial r}\right]\psi_s(r;k). \tag{4}$$

The average values and fluctuations of the phase-space variables, r and k, with respect to the probability density $|\psi_s(r,k)|^2$, are given by $\overline{r}=x_\psi-x_\phi$, $\overline{k}=p_\psi+p_\phi$, $(\Delta r)^2=(\Delta x_\psi)^2+(\Delta x_\phi)^2$ and $(\Delta k)^2=(\Delta p_\psi)^2+(\Delta p_\phi)^2$, where we set $x_\psi=\langle\psi|\hat{x}|\psi\rangle$, $p_\psi=\langle\psi|\hat{p}|\psi\rangle$, $(\Delta x_\psi)^2=\langle\psi|\hat{x}^2|\psi\rangle-(\langle\psi|\hat{x}|\psi\rangle)^2$, and so on.

The phase-space function $\psi_s(r;k)$ depends on the quantum state $|\phi\rangle_r$ of the reference system. Thus, besides the intrinsic fluctuation of the relevant system, the additional fluctuation is introduced in the phase-space representation. For example, the fluctuation of the phase-space variable r is given by $(\Delta x_\psi)^2+(\Delta x_\phi)^2$, but not $(\Delta x_\psi)^2$, where $(\Delta x_\psi)^2$ is the intrinsic fluctuation and $(\Delta x_\phi)^2$ is the additional one. Such an additional fluctuation can be attributed to the effect of the measurement apparatus which is used to obtain the information about the position and momentum of the relevant system. In the fuzzy-space formulation,[7,8] the quantum state $|\phi\rangle_r$ characterizes the finite accuracy of the position and momentum variables. Therefore, the phase-space representation is the operational description of the relevant quantum system rather than the intrinsic description.[9]

We can investigate the physical and chemical properties of the relevant quantum system in terms of the phase-space function $\psi_s(r,k)$ and the differential operators $\frac{1}{2}(1+s)+i\partial/\partial k$ and $\frac{1}{2}(1-s)-i\partial/\partial r$. It is shown that under certain conditions,[2] these results become equivalent to those obtained by Torres-Vega and Frederick[3] and Harriman.[4] The properties of the phase-space representation obtained by the relative-state method have been investigated in detail.[2]

B. The displacement-operator approach

We next introduce the phase-space representation of quantum state vectors obtained by the displacement-operator approach.[1,5] The displacement operator $\hat{D}_s(r,k)$ used in this approach is given by

$$\hat{D}_s(r,k)=\exp[i(k\hat{x}-r\hat{p}-\tfrac{1}{2}skr)], \tag{5}$$

where the parameter s determines the operator ordering; the standard ordering $\hat{D}_1(r,k) = e^{-ir\hat{p}}e^{ik\hat{x}}$ for $s=\mathbf{1}$, the antistandard ordering $\hat{D}_{-1}(r,k)=e^{ik\hat{x}}e^{-ir\hat{p}}$ for $s=-1$, and the symmetric ordering $\hat{D}_0(r,k)=e^{i(k\hat{x}-r\hat{p})}$ for $s=0$.[10] Using the displacement operator $\hat{D}_s(r,k)$, Mølller introduces a state vector $|\Omega(r,k;s)\rangle$ to formulate the phase-space representation of the relevant quantum system,

$$|\Omega(r,k;s)\rangle = \frac{1}{\sqrt{2\pi}}\hat{D}_s(r,k)|\chi\rangle = \hat{D}_s(r,k)|\Omega(0,0;s)\rangle, \tag{6}$$

where $|\chi\rangle$ is an arbitrary normalized state vector of the relevant quantum system $\mathcal{S}$. In particular, if $|\chi\rangle$ is a vacuum state, $|\Omega(r,k;s)\rangle$ becomes the Glauber coherent state, except for the phase factor and the normalization constant.[11] It is easy to see that the set $\{|\Omega(r,k;s)\rangle | r,k\in\mathbf{R}\}$ becomes an overcomplete system. Thus a quantum state vector $|\psi\rangle$ of the relevant system is represented by the phase-space function $\tilde{\psi}_s(r,k;s)=\langle\Omega(r,k;s)|\psi\rangle$ which is normalized as $\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk|\tilde{\psi}_s(r,k)|^2=1$. We obtain the phase-space representation of the canonical position and momentum operators,

$$\langle\Omega(r,k;s)|\hat{x}|\psi\rangle = \left[\frac{1}{2}(1+s)r + i\frac{\partial}{\partial k}\right]\tilde{\psi}_s(r;k), \tag{7}$$

$$\langle\Omega(r,k;s)|\hat{p}|\psi\rangle = \left[\frac{1}{2}(1-s)r - i\frac{\partial}{\partial r}\right]\tilde{\psi}_s(r;k), \tag{8}$$

which are identical with those obtained by the relative-state approach [see Eqs. (3) and (4)]. This result indicates that the relative-state approach and the displacement-operator approach yield the equivalent phase-space representations of quantum state vectors of the relevant system. We finally remark that the phase-space representation by the displacement-operator approach depends on the quantum state $|\chi\rangle$. This arbitrariness is equivalent to that in the relative-state approach. Thus, the displacement-operator approach also yields the operational description of the relevant system.

III. RELATION BETWEEN THE RELATIVE-STATE APPROACH AND THE DISPLACEMENT-OPERATOR APPROACH

We have found that the phase-space representation obtained by the relative-state method is equivalent to that obtained by the displacement-operator method. The equivalence results from fact that the state vector $|\Omega(r,k;s)\rangle$ introduced by Mølller is identical with the reduced relative-state vector $|\omega(r,k;s)\rangle$. In fact, using the normalizability of the wave function $\phi(x)$, we obtain

$$\begin{aligned}
|\omega(r,k;s)\rangle &= \frac{1}{\sqrt{2\pi}}e^{-i(1+s)kr/2}\int_{-\infty}^{\infty}dx\, e^{ikx}\phi^*(x-r)|x\rangle \\
&= e^{-ir\hat{p}}\frac{1}{\sqrt{2\pi}}e^{i(1-s)kr/2}\int_{-\infty}^{\infty}dx\, e^{ikx}\phi^*(x)|x\rangle \\
&= e^{-ir\hat{p}}e^{-k\hat{x}}\frac{1}{\sqrt{2\pi}}e^{i(1-s)kr/2}\int_{-\infty}^{\infty}dx\, \phi^*(x)|x\rangle \\
&= e^{i(k\hat{x}-r\hat{p}-skr/2)}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dx\, \phi^*(x)|x\rangle \\
&= \hat{D}_s(r,k)|\omega(0,0;s)\rangle = \frac{1}{\sqrt{2\pi}}\hat{D}_s(r,k)|\gamma\rangle,
\end{aligned} \tag{9}$$

where the normalized state vector $|\gamma\rangle$ of the relevant system is given by

$$|\gamma\rangle=\int_{-\infty}^{\infty}dx\,\phi^*(x)|x\rangle. \tag{10}$$

Hence, if we identify the state vector $|\gamma\rangle$ in the relative-state method with the state vector $|\chi\rangle$ in the displacement-operator method [or equivalently $\phi^*(x)=\chi(x)$ (Ref. 5)], we obtain the equality $|\omega(r,k;s)\rangle=|\Omega(r,k;s)\rangle$ and thus the equivalence is obvious.

We now consider the relation between the relative-state approach and the displacement-operator approach from the probabilistic point of view. Suppose that the relevant system is in a pure quantum state described by a statistical operator $\hat{\rho}=|\psi\rangle\langle\psi|$. Although we can consider a mixed quantum state, we confine ourselves to investigating a pure quantum state since we are interested in the phase-space representation of quantum state vectors. From the general theory of the operational quantum mechanics,[12,13] the probability $P(X,Y)$ that the phase-space variables r and k take values respectively in the subsets X and Y of the set $\mathbf{R}$ is given by

$$P(X,Y)=\mathrm{Tr}[\hat{\mathcal{M}}(X,Y)\hat{\rho}]=\langle\psi|\hat{\mathcal{M}}(X,Y)|\psi\rangle. \tag{11}$$

The operator $\hat{\mathcal{M}}(X,Y)$ defined on the Hilbert space $\mathcal{H}$ is a positive operator-valued measure (POVM), but not a projection-valued measure in general. In the phase-space representation obtained by the relative-state method or the displacement-operator method, the POVM is given by

$$\hat{\mathcal{M}}(X,Y)=\int_{r\in X}dr\int_{k\in Y}dk\,|\omega(r,k;s)\rangle\langle\omega(r,k;s)|=\int_{r\in X}dr\int_{k\in Y}dk\,|\Omega(r,k;s)\rangle\langle\Omega(r,k;s)|. \tag{12}$$

Then we obtain the phase-space probability

$$P(X,Y)=\int_{r\in X}dr\int_{k\in Y}dk\,|\psi_s(r,k)|^2=\int_{r\in X}dr\int_{k\in Y}dk\,|\tilde{\psi}_s(r,k)|^2, \tag{13}$$

which indicates that the probability amplitude in phase space is given by $\psi_s(r,k)$ or $\tilde{\psi}_s(r,k)$. It is important to note that the POVM $\hat{\mathcal{M}}(X,Y)$ given by Eq. (12) is not a projection-valued measure because of the nonorthogonality of the state vectors $|\omega(r,k;s)\rangle$ and $|\Omega(r,k;s)\rangle$.

It is well known that any POVM is made some projection-valued measure by extending a Hilbert space (the Naimark theorem).[13] The operators $\hat{x}\otimes\hat{1}_r-\hat{1}\otimes\hat{x}_r$ and $\hat{p}\otimes\hat{1}_r+\hat{1}\otimes\hat{p}_r$ of the compound system $\mathcal{S}+\mathcal{S}_r$ are commutable and thus represent simultaneously measurable quantities, the eigenstate of which is given by the relative-state vector $|\omega(r,k;s)\rangle\rangle$ [see Eq. (1)] in the sense of Ref. 6. The projection operator[14]

$$\hat{\mathcal{N}}(X,Y)=\int_{r\in X}dr\int_{k\in Y}dk\,|\omega(r,k;s)\rangle\rangle\langle\langle\omega(r,k;s)| \tag{14}$$

describes the simultaneous measurement, where the measurement outcomes of $\hat{x}\otimes\hat{1}_r-\hat{1}\otimes\hat{x}_r$ and $\hat{p}\otimes\hat{1}_r+\hat{1}\otimes\hat{p}_r$ belong to X and Y. Then it is easy to see that this operator satisfies the relation

$$P(X,Y)=\mathrm{Tr}[\hat{\mathcal{M}}(X,Y)\hat{\rho}]=\mathrm{Tr}\,\mathrm{Tr}_r[\hat{\mathcal{N}}(X,Y)\hat{\rho}\otimes\hat{\sigma}_r], \tag{15}$$

where we have used $|\omega(r,k;s)\rangle={}_r\langle\phi|\omega(r,k;s)\rangle\rangle$ and Tr_r is the trace operation over the reference system $\mathcal{S}_r$. This result indicates that the projection operator $\hat{\mathcal{N}}(X,Y)$ is the Naimark extension of the POVM $\hat{\mathcal{M}}(X,Y)$ in the sense of Ref. 6. In Eq. (15), $\hat{\sigma}_r=|\phi\rangle_{rr}\langle\phi|$ is called the Naimark state. In particular, when $|\phi\rangle_r$ is the vacuum state, the projection operator $\hat{\mathcal{N}}(X,Y)$ describes the balanced homodyne detection and the heterodyne detection.[15–17] Therefore, we have found that the relative-state approach corresponds to the Naimark extension of the displacement-operator approach.

In conclusion, it is shown that the quantum state vector $|\Omega(r,k;s)\rangle$ used in the displacement-operator approach is equivalent to the reduced relative-state vector $|\omega(r,k;s)\rangle$. Thus it is obvious that the relative-state approach and the displacement-operator approach yield equivalent phase-space representations of quantum state vectors. From the probabilistic point view, the relative-state approach corresponds to the Naimark extension of the displacement-operator approach.

[1] K. B. Møller, T. G. Jørgensen, and G. Torres-Vega, J. Chem. Phys. **106**, 7228 (1997) and references therein.
[2] M. Ban, J. Math. Phys. **39**, 1744 (1998).
[3] G. Torres-Vega and J. H. Frederick, J. Chem. Phys. **98**, 187 (1993).
[4] J. E. Harriman, J. Chem. Phys. **100**, 3651 (1994).
[5] K. B. Møller, ''Comment on phase-space representation of quantum state vectors,'' J. Math. Phys. **40**, 2531 (1999).
[6] Any eiganstate of an unbounded operator with continuous spectrum, such as position and momentum operators, cannot be an element of a Hilbert space. Such a state vector is rigorously treated by introducing a rigged Hilbert space (or Gel'fand triplets) [see A. Böhm, *The Rigged Hilbert Space and Quantum Mechanics* (Springer-Verlag, Berlin, 1978); A. Böhm and M. Gadella, *Dirac Kets, Gamov Vectors, and Gel'fand Triplets* (Springer-Verlag, Berlin, 1989)].
[7] E. Prugovečki, J. Math. Phys. **17**, 517 (1976); **17**, 1673 (1976).
[8] S. Twareque Ali and E. Prugovečki, J. Math. Phys. **17**, 1105 (1976); **18**, 219 (1977).
[9] M. Ban, Int. J. Theor. Phys. **36**, 2583 (1997).
[10] S. G. Agarwal and E. Wolf, Phys. Rev. D **2**, 2161 (1970); **2**, 2187 (1970); **2**, 2206 (1970).
[11] J. R. Klauder and B.-S. Skagerstam, *Coherent States* (World Scientific, Singapore, 1985).
[12] E. B. Davies, *Quantum Theory of Open Systems* (Academic, New York, 1976).
[13] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).
[14] C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic, New York, 1976).
[15] S. Stenholm, Ann. Phys. (N.Y.) **218**, 233 (1992).
[16] J. H. Shapiro and S. S. Wagner, IEEE J. Quantum Electron. **QE-20**, 803 (1984).
[17] Y. Lai and H. A. Haus, Quantum Opt. **1**, 99 (1989).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

A one-dimensional model for *n*-level atoms coupled to an electromagnetic field

Zorawar S. Bassi[a] and André LeClair
Newman Laboratory, Cornell University, Ithaca, New York 14853

(Received 1 February 1999; accepted for publication 4 May 1999)

A model for *n*-level atoms coupled to quantized electromagnetic fields in a fibrillar geometry is constructed. In the slowly varying envelope and rotating wave approximations, the equations of motion are shown to satisfy a zero curvature representation, implying integrability of the quantum system. © *1999 American Institute of Physics.* [S0022-2488(99)02808-X]

I. INTRODUCTION

The interaction of radiation with two-level atoms has been extensively studied under various approximations. In one spatial dimension, the reduced Maxwell–Bloch equations resulting from the slowly varying envelope and rotating wave approximations are known to be quantum integrable.[1] In this paper we generalize the one-dimensional case of two-level atoms to that of *n*-level atoms.

In the first of two main parts, we construct the fully quantum *n*-level model. The system consists of *n*-level atoms distributed in a fibrillar geometry, interacting with radiation through a minimally coupled Hamiltonian. In the remaining section we apply the approximations, and show that the Heisenberg equations of motion for the reduced system satisfy the so-called zero curvature representation. This implies that the system is integrable and can be solved by the quantum inverse scattering method.

II. MATHEMATICAL BACKGROUND

Let us first recall the sl_n Lie algebra. The n^2-1 generators, written as

$$\{E_{ij},H_m|1\leq i\neq j\leq n,\ 1\leq m\leq r\},\tag{2.1}$$

where $r=n-1$ is the rank, satisfy the following brackets (in the Chevalley basis):

$$[E_{ij},E_{kl}]=\begin{Bmatrix}\delta_{kj}E_{il}-\delta_{il}E_{kj} & \text{if}\quad \delta_{il}\delta_{kj}=0\\ \sum_{m=i}^{j-1}H_m(i<j) & \text{if}\quad \delta_{il}\delta_{kj}=1\end{Bmatrix},\tag{2.2a}$$

$$[E_{ij},H_m]=(\delta_{jm}-\delta_{im}-\delta_{jm+1}+\delta_{im+1})E_{ij},\tag{2.2b}$$

$$[H_a,H_b]=0.\tag{2.2c}$$

The set spanned by $\{H_m\}$ is the Cartan subalgebra. [Note that in the bracket (2.2a) a term of the form $\delta_{i\neq j}E_{kk}$ is formally equal to zero, even though E_{kk} has not been defined. The set $\{E_{ij}|1\leq i,j\leq n\}$ satisfying the first relation in (2.2a) is a basis for the algebra gl_n.] A representation ρ of sl_n will be denoted as

$$\{E^{\rho}_{ij}=\rho(E_{ij}),H^{\rho}_m=\rho(H_m)\}.\tag{2.3}$$

---
[a]Electronic mail: zorawar@mail.lns.cornell.edu

The $r\times r$ Cartan matrix A has the explicit form

$$A_{uv}=2\,\delta_{uv}-\delta_{uv-1}-\delta_{uv+1}\,. \tag{2.4}$$

It is a symmetric matrix with diagonal elements 2 and nearest off-diagonal elements -1. We now proceed to build our quantum system.

III. THE INTERACTING *n*-LEVEL HAMILTONIAN

We model a free n-level atom as having a single electron with eigenstates $|i\rangle$, $i=1,2,...,n$, and energies $\epsilon_1>\epsilon_2>\cdots>\epsilon_n$. The energy splitting between states $|i\rangle$ and $|i+1\rangle$ will be denoted by ω_i or ω_{ii+1},

$$\omega_i=\omega_{ii+1}=\epsilon_i-\epsilon_{i+1}\,,\quad 1\leqslant i<n. \tag{3.1}$$

The second notation can be generalized as follows:

$$\omega_{ij}=\epsilon_i-\epsilon_j\,,\quad 1\leqslant i<j\leqslant n. \tag{3.2}$$

The notation ω_i is only defined for the energy splitting between successive states.

To define various atomic operators, we first introduce fermion creation and destruction operators $\{b_i\,,b_i^\dagger\}$ for $1\leqslant i\leqslant n$. The operator $b_i^\dagger(b_i)$ creates (destroys) an electron in the ith level. These operators satisfy the algebra

$$\{b_i\,,b_j\}=\{b_i^\dagger\,,b_j^\dagger\}=0,\quad \{b_i\,,b_j^\dagger\}=\delta_{ij}\,. \tag{3.3}$$

The atomic operators can now be written as

$$\mathcal{O}_{ij}=b_i^\dagger b_j\,,\quad 1\leqslant i,\ j\leqslant n \tag{3.4}$$

or linear combinations of the $\mathcal{O}_{ij}$'s. The action of $\mathcal{O}_{ij}$ on an atomic state $|k\rangle$ is given by

$$\mathcal{O}_{ij}|k\rangle=b_i^\dagger b_j|k\rangle=\delta_{kj}|i\rangle. \tag{3.5}$$

From (3.3) the general commutator for the $\mathcal{O}$ operators is

$$[\mathcal{O}_{ij}\,,\mathcal{O}_{kl}]=\mathcal{O}_{il}\delta_{jk}-\mathcal{O}_{kj}\delta_{il}\,. \tag{3.6}$$

Operators of the form $\mathcal{O}_{i<j}$ are referred to as raising operators. These cause a transition from the lower energy state $|j\rangle$ to the higher energy state $|i\rangle$. Similarly the operators $\mathcal{O}_{j>i}$ are lowering operators. We also define a set of commuting operators, denoted $\mathcal{H}_m$, as follows:

$$\mathcal{H}_m=\mathcal{O}_{mm}-\mathcal{O}_{m+1m+1}\,,\quad 1\leqslant m\leqslant r. \tag{3.7}$$

The set $\{\mathcal{O}_{i\neq j}\,,\mathcal{H}_m\}$ satisfies (2.2), thus forming a representation of sl_n. (We shall often use the notation $X_{a<b}(X_{a>b})$ to mean X_{ab} with $a<b(a>b)$ for X any quantity, operator, c number, etc.)

The free atomic Hamiltonian can be written as

$$H_0^{\text{atom}}=\sum_{u=1}^{r}\left(\sum_{v=1}^{u}\epsilon_v\right)\mathcal{H}_u+\left(\sum_{v=1}^{n}\epsilon_v\right)b_n^\dagger b_n\,. \tag{3.8}$$

We will choose the arbitrary lowest state energy ϵ_n to be such that $\Sigma_{v=1}^{n}\epsilon_v=0$. In terms of the ω_i's and the inverse Cartan matrix A^{-1}, this means setting

$$\epsilon_n=-\sum_{v=1}^{r}A_{rv}^{-1}\omega_v\,, \tag{3.9}$$

which implies

$$\sum_{v=1}^{u} \epsilon_v = \sum_{v=1}^{r} A_{uv}^{-1} \omega_v . \tag{3.10}$$

The atomic Hamiltonian (3.8) thus takes the form

$$H_0^{\text{atom}} = \sum_{1 \leqslant u,v \leqslant r} A_{uv}^{-1} \omega_v \mathcal{H}_u . \tag{3.11}$$

To couple the atom to an electromagnetic field we make use of the minimal coupling prescription (see any standard text on quantum mechanics, e.g., Ref. 2). The standard Hamiltonian is

$$H = H_0^{\phi} + H_0^{\text{atom}} + H_{\text{int}} , \tag{3.12}$$

where

$$H_0^{\text{atom}} = \frac{1}{2m_e} \vec{p} \cdot \vec{p} + V(\vec{x}), \tag{3.13}$$

$$H_{\text{int}} = -\frac{e}{2m_e} (\vec{p} \cdot \vec{A}(\vec{x}) + \vec{A}(\vec{x}) \cdot \vec{p}) + \frac{e^2}{2m} \vec{A}(\vec{x}) \cdot \vec{A}(\vec{x}), \tag{3.14}$$

and H_0^{ϕ} is the free field Hamiltonian. The Hamiltonian (3.13) is identified with (3.11), i.e., the exact eigenstates and energies of (3.13) are taken to be $\{|i\rangle, \epsilon_i\}$, $i=1,...,n$. If the spatial variation of the vector potential $\vec{A}$ is small across the atom, we can take its value at a fixed point $\vec{x}_0$ inside the atom. Using

$$\vec{p} = -i m_e [\vec{x}, H_0^{\text{atom}}], \tag{3.15}$$

we get

$$-\frac{e}{2m_e} \langle a | \vec{p} \cdot \vec{A}(\vec{x}) + \vec{A}(\vec{x}) \cdot \vec{p} | b \rangle = i(\epsilon_b - \epsilon_a) \vec{A}(\vec{x}_0) \cdot \langle a | \vec{d} | b \rangle, \tag{3.16}$$

where $\vec{d} = e\vec{x}$ is the electric dipole operator. Since $\vec{d}$ is a vector operator and the atomic states are assumed to be parity eigenstates, we have $\langle i | \vec{d} | i \rangle = 0$. The nonzero matrix elements can be parametrized as follows $(i<j)$:

$$\langle i | \vec{d} | j \rangle = d_{ij} e^{i\alpha_{ij}} \hat{n}_{ij} , \quad \langle j | \vec{d} | i \rangle = d_{ij} e^{-i\alpha_{ij}} \hat{n}_{ij} , \tag{3.17}$$

where $d_{ij} \geqslant 0$ and the $\hat{n}_{ij}$'s are unit vectors. We will consider the situation where $\hat{n}_{ij}$ is independent of i,j due to some symmetry of the system and write $\hat{n} \equiv \hat{n}_{ij}$. The dipole operator can be expanded in terms of the raising and lowering operators as

$$\vec{d} = \hat{n} \sum_{i<j} (d_{ij} e^{i\alpha_{ij}} \mathcal{O}_{ij} + d_{ij} e^{-i\alpha_{ij}} \mathcal{O}_{ji}), \tag{3.18}$$

which gives for the interaction Hamiltonian

$$H_{\text{int}} = -i \vec{A}(\vec{x}_0) \cdot \hat{n} \sum_{i<j} \omega_{ij} d_{ij} (\mathcal{O}_{ij} e^{i\alpha_{ij}} - \mathcal{O}_{ji} e^{-i\alpha_{ij}}) + \frac{e^2}{2m_e} \vec{A}(\vec{x}_0) \cdot \vec{A}(\vec{x}_0). \tag{3.19}$$

To reduce this system to a one-dimensional model, we make use of the fibrillar geometry. The atom can be thought of as an impurity in an optical fiber of cross-sectional area $\mathcal{A}$ and length L, with $L \gg \sqrt{\mathcal{A}}$. Taking the fiber along the $\hat{x}$ direction, the reduced field action is found to be (see Ref. 3 for details)

$$S_{\text{Maxwell}} = \int dx\, dt \tfrac{1}{2}(\partial_t \phi \partial_t \phi - \partial_x \phi \partial_x \phi), \tag{3.20}$$

where ϕ is a dimensionless scalar field defined through

$$\vec{A} \cdot \hat{n} = \sqrt{\frac{4\pi}{\mathcal{A}_{\text{eff}}}}\, \phi. \tag{3.21}$$

Here $\vec{A}$ is the vector potential depending only on the x coordinate and $\mathcal{A}_{\text{eff}}$ is the effective fiber cross-sectional area. The field ϕ satisfies the commutation relation

$$[\phi(x,t), \partial_t \phi(x',t)] = i\,\delta(x-x'). \tag{3.22}$$

From the action the free field Hamiltonian is found to be

$$H_0^{\phi} = \int dx \frac{1}{2}[(\partial_t \phi)^2 + (\partial_x \phi)^2] + \frac{2\pi e^2}{m_e \mathcal{A}_{\text{eff}}} \phi^2(x_0), \tag{3.23}$$

where the last term is the quadratic potential term taken from H_{int}. Now the interaction Hamiltonian is

$$H_{\text{int}} = -\frac{i}{2} \phi(x_0) \sum_{i<j} \omega_{ij} \beta_{ij} (\mathcal{O}_{ij} e^{i\alpha_{ij}} - \mathcal{O}_{ji} e^{-i\alpha_{ij}}), \tag{3.24}$$

where (explicitly showing $\hbar$ and c)

$$\beta_{i<j} = \sqrt{\frac{16\pi}{\hbar c \mathcal{A}_{\text{eff}}}}\, d_{ij}. \tag{3.25a}$$

For $1 \leqslant m \leqslant r$ we define

$$\beta_m = \sqrt{\frac{16\pi}{\hbar c \mathcal{A}_{\text{eff}}}}\, d_{mm+1} = \beta_{mm+1}. \tag{3.25b}$$

The β parameters are the important dimensionless coupling constants of the model. The spontaneous decay rate $\Gamma_{ij}^s = 1/\tau_{ij}^s$ of a single excited atom from the state $|i\rangle$ to the state $|j\rangle$ is given by

$$\Gamma_{ij}^s = \frac{\beta_{ij}^2}{4} \omega_{ij}. \tag{3.26}$$

Next to make the transition to a continuous system. For N atoms positioned at $x = x_m$, $m = 1,...,N$, let $\vec{d}_m = e(\vec{x} - \vec{x}_m)$ and $\mathcal{O}_{ij}(x_m)$ be the dipole and transition operators for the atom at x_m. The matrix elements of $\vec{d}_m$ are independent of the position, however, the orientation $\hat{n}$ can vary from atom to atom. The operator $\vec{d}_m$ can be written in terms of the single atom matrix elements as

$$\vec{d}_m = \hat{n}_m \sum_{i<j} d_{ij} (\mathcal{O}_{ij}(x_m) e^{i\alpha_{ij}} + \mathcal{O}_{ji}(x_m) e^{-i\alpha_{ij}}). \tag{3.27}$$

For simplification, we suppose the situation where all atoms are aligned $\hat{n}_m = \hat{n}$ (e.g., by an external electric field), giving

$$H_{\rm int} = -\frac{i}{2}\int dx\, \phi(x)\sum_{i<j} \omega_{ij}\beta_{ij}(\mathcal{O}_{ij}(x,t)e^{i\alpha_{ij}} - \mathcal{O}_{ji}(x,t)e^{-i\alpha_{ij}}), \tag{3.28}$$

where we have introduced the space-dependent (continuous) transition operators

$$\mathcal{O}_{ij}(x,t) = \sum_{m=1}^{N} \mathcal{O}_{ij}(x_m,t)\,\delta(x-x_m). \tag{3.29}$$

The discrete operator $\mathcal{O}_{ij}(x_m,t)$ acts only on the atom at x_m to cause a transition from $|j\rangle$ to $|i\rangle$. [Note the abuse of notation, we are using the same symbol $\mathcal{O}_{ij}$ for both the continuous and discrete operators. Henceforth only the continuous version $\mathcal{O}_{ij}(x,t)$ will appear.] A time dependence in (3.29) signifies that the $\mathcal{O}_{ij}$ operators are to be treated as Heisenberg operators. The continuous $\mathcal{H}_m(x,t)$ operators are defined similarly. The general commutator for the space–time transition operators is

$$[\mathcal{O}_{ij}(x,t),\mathcal{O}_{kl}(x',t)] = (\mathcal{O}_{il}(x,t)\,\delta_{jk} - \mathcal{O}_{kj}(x,t)\,\delta_{il})\,\delta(x-x'), \tag{3.30}$$

from which it is easily seen that the algebra (or now more appropriately current algebra) satisfied by the set $\{\mathcal{O}_{i\neq j}(x,t),\mathcal{H}_m(x,t)\}$ is identical to the sl_n algebra (2.2). The free atomic Hamiltonian takes the form

$$H_0^{\rm atom} = \int dx \sum_{1\leq u,v\leq r} A_{uv}^{-1}\omega_v \mathcal{H}_u(x,t). \tag{3.31}$$

The complete Hamiltonian for the system is therefore $H_0^{\rm atom} + H_0^{\phi} + H_{\rm int}$, with $H_0^{\rm atom}$, H_0^{ϕ}, and $H_{\rm int}$ given by (3.31), (3.23), and (3.28), respectively.

IV. TWO APPROXIMATIONS AND INTEGRABILITY

We now make use of two approximations common in quantum optics to further simplify $H_{\rm int}$—these being the slowly varying envelope and rotating wave approximations.[4]

In the slowly varying envelope approximation, one assumes that near resonant photons with energies $\approx\omega_{ij}$ are most relevant. Then the scalar field ϕ can be expanded about the various resonances as

$$\phi(x,t) \approx \sum_{i<j} (e^{-i\omega_{ij}(t-x)}\psi_{ij}(x,t) + e^{i\omega_{ij}(t-x)}\psi_{ij}^{\dagger}(x,t)), \tag{4.1}$$

where $\psi_{ij}(x,t)$ and $\psi_{ij}^{\dagger}(x,t)$ are destruction and creation fields with mode expansions

$$\psi_{i<j}(x,0) = \frac{1}{\sqrt{2\omega_{ij}}}\int \frac{dk_e}{\sqrt{2\pi}}\hat{a}_{ij}(k_e)e^{ik_e x}, \tag{4.2a}$$

$$\psi_{i<j}^{\dagger}(x,0) = \frac{1}{\sqrt{2\omega_{ij}}}\int \frac{dk_e}{\sqrt{2\pi}}\hat{a}_{ij}^{\dagger}(k_e)e^{-ik_e x}, \tag{4.2b}$$

and

$$\hat{a}_{ij}(k_e) = a_{ij}(k_e+\omega_{ij}), \quad \hat{a}_{ij}^{\dagger}(k_e) = a_{ij}^{\dagger}(k_e+\omega_{ij}). \tag{4.3}$$

Here a_{ij} ($a^\dagger_{ij}$) is the usual photon destruction (creation) operator. (Note that we are only considering right-moving plane waves.) The operator $\hat{a}_{ij}(k_e)$ ($\hat{a}^\dagger_{ij}(k_e)$) destroys (creates) a photon with energy

$$|k|=|k_e+\omega_{ij}|\approx\omega_{ij}\,. \tag{4.4}$$

Thus k_e acts as an ''envelope'' vector about the ω_{ij} resonance. The photon operators satisfy standard commutation relations

$$[\hat{a}_{ij}(k),\hat{a}^\dagger_{kl}(k')]=[a_{ij}(k),a^\dagger_{kl}(k')]=\delta_{ik}\delta_{jl}\delta(k-k'). \tag{4.5}$$

In writing (4.1) and the above commutation relations, we have assumed that all resonances ω_{ij} $\{\frac{1}{2}(n^2-n)$ in total) are distinct (i.e., $\omega_{ij}=\omega_{kl}\Leftrightarrow i=k$ and $j=l$) and sharp. From (4.5) the component fields satisfy

$$[\psi_{ij}(x,t),\psi^\dagger_{kl}(x',t)]=\frac{1}{2\omega_{ij}}\,\delta_{ik}\delta_{jl}\delta(x-x'), \tag{4.6}$$

with all other commutators zero.

The rotating wave approximation reduces the number of interactions in H_{int}. Using (4.1) in H_{int} we obtain terms with both photon creation ($a^\dagger$) and atomic raising ($\mathcal{O}_{i<j}$) operators (or photon destruction and atomic lowering $\{a,\mathcal{O}_{j>i}\}$), or two photon creation (destruction) operator terms in H_0^ϕ. Such high frequency terms lead to vacuum fluctuations and higher order processes. The rotating wave approximation sets these processes to zero. We also set to zero terms of the form $\psi_{i<j}\mathcal{O}_{k<l}$ (or $\psi^\dagger_{i<j}\mathcal{O}_{l>k}$) for $(i,j)\neq(k,l)$ since they give no contribution to lowest order in perturbation theory. So we only retain those terms which pair creation/lowering or destruction/raising operators (and creation/destruction operators in H_0^ϕ) and connect states with approximately equal energy.

Combining these two approximations we get

$$H_{\text{int}}=-\frac{i}{2}\int dx\sum_{i<j}\beta_{ij}\omega_{ij}(\psi_{ij}e^{i\alpha_{ij}}\mathcal{O}_{ij}e^{-i\omega_{ij}(t-x)}-\psi^\dagger_{ij}e^{-i\alpha_{ij}}\mathcal{O}_{ji}e^{i\omega_{ij}(t-x)}), \tag{4.7}$$

$$H_0^\phi=-2i\int dx\sum_{i<j}\omega_{ij}\psi^\dagger_{ij}\partial_x\psi_{ij}\,. \tag{4.8}$$

The free field Hamiltonian follows from the field action, which using

$$|k_e|^2\ll\omega_{ij}^2\Rightarrow|\partial_x\psi_{ij}|\ll\omega_{ij}|\psi_{ij}|,\quad |\partial_t\psi_{ij}|\ll\omega_{ij}|\psi_{ij}|, \tag{4.9}$$

approximates to

$$\int dx\,dt\tfrac{1}{2}((\partial_t\phi)^2-(\partial_x\phi)^2)\approx 2i\int dx\,dt\sum_{i<j}\omega_{ij}\psi^\dagger_{ij}(\partial_x+\partial_t)\psi_{ij}\,. \tag{4.10}$$

[In (4.8) we have dropped the quadratic term that appears in (3.23). For electric fields small compared with e/a_0^2, this term is negligible in relation to H_{int}.] The phases $e^{\pm i\alpha_{ij}}$ and $e^{\pm i\omega_{ij}x}$ can be absorbed into $\{\psi_{ij},\psi^\dagger_{ij}\}$ and $\{\mathcal{O}_{ij},\mathcal{O}_{ji}\}_{i<j}$, respectively, without changing the commutation relations. The time-dependent phase $e^{-i\omega_{ij}t}$ ($e^{+i\omega_{ij}t}$) cancels the time dependence of $\mathcal{O}_{i<j}$ ($\mathcal{O}_{j>i}$) coming from the free atomic Hamiltonian. Thus we can set H_0^{atom} to zero and consider the model defined by the complete Hamiltonian

$$H=-2i\int dx\sum_{i<j}\omega_{ij}\psi^\dagger_{ij}\partial_x\psi_{ij}-\frac{i}{2}\int dx\sum_{i<j}\omega_{ij}\beta_{ij}(\psi_{ij}\mathcal{O}_{ij}-\psi^\dagger_{ij}\mathcal{O}_{ji}). \tag{4.11}$$

Finally we rescale ψ_{ij} and $\psi_{ij}^\dagger$ as

$$\psi_{ij}\rightarrow\frac{\psi_{ij}}{\sqrt{2\omega_{ij}}},\quad \psi_{ij}^\dagger\rightarrow\frac{\psi_{ij}^\dagger}{\sqrt{2\omega_{ij}}}, \tag{4.12}$$

which gives for the commutator (4.6),

$$[\psi_{ij}(x,t),\psi_{kl}^\dagger(x',t)]=\delta_{ik}\delta_{jl}\delta(x-x'), \tag{4.13}$$

and defining

$$\tilde{\omega}_{i<j}=\frac{1}{2\sqrt{2}}\beta_{ij}\sqrt{\omega_{ij}}=\sqrt{\frac{\Gamma_{ij}^s}{2}},\quad \tilde{\omega}_m=\tilde{\omega}_{mm+1}\quad (1\leqslant m\leqslant r), \tag{4.14}$$

the Hamiltonian (4.11) becomes

$$H=-i\int dx\sum_{i<j}\psi_{ij}^\dagger\partial_x\psi_{ij}-i\int dx\sum_{i<j}\tilde{\omega}_{ij}(\psi_{ij}\mathcal{O}_{ij}-\psi_{ij}^\dagger\mathcal{O}_{ji}). \tag{4.15}$$

The first interaction term, $\psi_{ij}\mathcal{O}_{ij}$, causes an atomic transition from a lower energy state $|j\rangle$ to a higher energy state $|i\rangle$, along with the absorption of a photon of energy $\approx\omega_{ij}$. The second term, $\psi_{ij}^\dagger\mathcal{O}_{ji}$, causes a transition from a higher energy state $|i\rangle$ to a lower energy state $|j\rangle$, along with the creation of a photon of energy $\approx\omega_{ij}$.

From H we can obtain the Heisenberg operator equations of motion. Explicitly we find

$$\partial_t\mathcal{O}_{k<l}=-\sum_{j>l}\tilde{\omega}_{lj}\psi_{lj}\mathcal{O}_{kj}+\sum_{i<k}\tilde{\omega}_{ik}\psi_{ik}\mathcal{O}_{il}+\sum_{\substack{i<l\\ i\neq k}}\tilde{\omega}_{il}\psi_{il}^\dagger\mathcal{O}_{ki}-\sum_{\substack{j>k\\ j\neq l}}\tilde{\omega}_{kj}\psi_{kj}^\dagger\mathcal{O}_{jl}+\tilde{\omega}_{kl}\psi_{kl}^\dagger\sum_{m=k}^{l-1}\mathcal{H}_m, \tag{4.16a}$$

$$\partial_t\mathcal{O}_{l>k}=\partial_t\mathcal{O}_{k<l}^\dagger=-\sum_{j>l}\tilde{\omega}_{lj}\psi_{lj}^\dagger\mathcal{O}_{jk}+\sum_{i<k}\tilde{\omega}_{ik}\psi_{ik}^\dagger\mathcal{O}_{li}+\sum_{\substack{i<l\\ i\neq k}}\tilde{\omega}_{il}\psi_{il}\mathcal{O}_{ik}$$
$$-\sum_{\substack{j>k\\ j\neq l}}\tilde{\omega}_{kj}\psi_{kj}\mathcal{O}_{lj}+\tilde{\omega}_{kl}\psi_{kl}\sum_{m=k}^{l-1}\mathcal{H}_m, \tag{4.16b}$$

$$\partial_t\mathcal{H}_m=-2\tilde{\omega}_{mm+1}(\psi_{mm+1}\mathcal{O}_{mm+1}+\psi_{mm+1}^\dagger\mathcal{O}_{m+1m})-\sum_{j>m+1}\tilde{\omega}_{mj}(\psi_{mj}\mathcal{O}_{mj}+\psi_{mj}^\dagger\mathcal{O}_{jm})$$
$$-\sum_{i<m}\tilde{\omega}_{im+1}(\psi_{im+1}\mathcal{O}_{im+1}+\psi_{im+1}^\dagger\mathcal{O}_{m+1i})+\sum_{i<m}\tilde{\omega}_{im}(\psi_{im}\mathcal{O}_{im}+\psi_{im}^\dagger\mathcal{O}_{mi})$$
$$+\sum_{j>m+1}\tilde{\omega}_{m+1j}(\psi_{m+1j}\mathcal{O}_{m+1j}+\psi_{m+1j}^\dagger\mathcal{O}_{jm+1}), \tag{4.16c}$$

$$(\partial_t+\partial_x)\psi_{k<l}=\tilde{\omega}_{kl}\mathcal{O}_{lk}, \tag{4.16d}$$

$$(\partial_t+\partial_x)\psi_{k<l}^\dagger=\tilde{\omega}_{kl}\mathcal{O}_{kl}. \tag{4.16e}$$

Each term in the equations of motion for the atomic operators has a simple physical interpretation. For example, consider the raising operator $\mathcal{O}_{k<l}$. The first summation in (4.16a) is a sum over all atomic and photon operator pairs, where the photon field destroys a photon of energy $\omega_{l<j}$, and

the atomic (raising) operator causes a transition from the lower state $|j\rangle$ to the higher state $|k\rangle$. The change in energy of the system corresponding to the atomic/field pair being ω_{kl} [as it is for every other term in (4.16a)]. If we think of a field destroying (creating) a photon of energy $\omega_{i<j}$ as ''connecting'' atomic states $|i\rangle$ to $|j\rangle$ ($|j\rangle$ to $|i\rangle$), along with $\mathcal{O}_{i<j}$ ($\mathcal{O}_{j>i}$) connecting $|j\rangle$ to $|i\rangle$ ($|i\rangle$ to $|j\rangle$) and $\mathcal{H}_m$ connecting $|m\rangle$ to $|m\rangle$ and $|m+1\rangle$ to $|m+1\rangle$, then (4.16a) is, aside from c-number factors, a sum over all atomic/field pairs connecting $|l\rangle$ to $|k\rangle$ through some intermediate state, i.e., $|l\rangle\rightarrow|j\rangle\rightarrow|k\rangle$. A similar interpretation follows for $\mathcal{O}_{l>k}$ and $\mathcal{H}_m$.

These equations of motion have a zero curvature representation (Refs. 5 and 6)

$$\partial_t A_x - \partial_x A_t = [A_x, A_t], \tag{4.17}$$

where A_x and A_t are matrices of quantum operators given by

$$A_x = \mu \sum_{1\leq m,n\leq r} A^{-1}_{mn}\mathcal{H}_n H^\rho_m + \mu \sum_{i<j} (\mathcal{O}_{ij}E^\rho_{ij} + \mathcal{O}_{ji}E^\rho_{ji}) + \sum_{i<j} \tilde{\omega}_{ij}(-\psi^\dagger_{ij}E^\rho_{ij} + \psi_{ij}E^\rho_{ji}) - \frac{1}{\mu}\sum_{1\leq m,n\leq r} A^{-1}_{mn}\tilde{\omega}^2_n H^\rho_m, \tag{4.18a}$$

$$A_t = \frac{1}{\mu}\sum_{1\leq m,n\leq r} A^{-1}_{mn}\tilde{\omega}^2_n H^\rho_m - \sum_{i<j} \tilde{\omega}_{ij}(-\psi^\dagger_{ij}E^\rho_{ij} + \psi_{ij}E^\rho_{ji}), \tag{4.18b}$$

provided that the $\tilde{\omega}_{i<j}$'s satisfy

$$\tilde{\omega}^2_{i<j} = \tilde{\omega}^2_{i<k} + \tilde{\omega}^2_{k<j}, \tag{4.19}$$

for any intermediate value of $k, i<k<j$. This is trivially satisfied by $\tilde{\omega}_{ii+1}$, in which case there is no such k. Here A^{-1}_{mn} is the inverse Cartan matrix, and $\{E^\rho_{i\neq j}, H^\rho_m\}$ are matrices in any representation ρ of sl_n satisfying (2.2). Requiring (4.17) to be valid for all values of the arbitrary spectral parameter μ, is equivalent to the equations of motion (4.16) [of course provided (4.19) is satisfied]. This equivalence can be shown by making use of the commutation relations (2.2). The equations of motion (4.16c) for the $\mathcal{H}_m$ operators can be rewritten in a more compact form

$$\partial_t \mathcal{H}_m = -\sum_{u=1}^{r} A_{mu} \sum_{u<l\leq n} \sum_{k=1}^{u} \tilde{\omega}_{kl}(\mathcal{O}_{kl}\psi_{kl} + \mathcal{O}_{lk}\psi^\dagger_{kl}). \tag{4.20}$$

The constraint (4.19) arises in forming a zero curvature representation for the field ($\{\psi_{ij}, \psi^\dagger_{ij}\}$) equations of motion, and reduces the number of free parameters to $r=n-1$, these being $\{\tilde{\omega}_m\}_{1\leq m\leq r}$. The definition (4.14) shows that the constraint is equivalent to the requirement that the spontaneous decay rate (for a single atom) from $|i\rangle$ to $|j\rangle$, Γ^s_{ij}, be equal to the sum of the decay rates Γ^s_{ik} and Γ^s_{kj} for any intermediate state $|k\rangle$,

$$\Gamma^s_{i<j} = \Gamma^s_{i<k} + \Gamma^s_{k<j}. \tag{4.21}$$

The equations of motion for the atomic operators have the zero curvature representation (4.17) independent of the constraint.

Having obtained a zero curvature representation, the system can now be solved using the quantum inverse scattering method (see Ref. 7 and references therein). In particular, from (4.17) and (4.18) one can construct a quantum monodromy matrix $T(\mu)$ depending on the spectral parameter μ, with the property that its trace, i.e., the transfer matrix $\tau(\mu) = \mathrm{Tr}\, T(\mu)$, commutes for arbitrary values of μ, $[\tau(\mu), \tau(\lambda)] = 0$. Furthermore, this commutation relation implies that $[H, \tau(\mu)] = 0$, which when expanded in μ leads to an infinite set of quantum conservation laws, hence integrability of the model. Eigenstates and eigenvalues of H can then be constructed with the algebraic Bethe ansatz.

[1] V. I. Rupasov, Pis'ma Zh. Eksp. Teor. Fiz. **36**, 115 (1982) [JETP Lett. **36**, 142 (1982)].
[2] G. Baym, *Lectures on Quantum Mechanics* (Addison–Wesley, New York, 1990).
[3] A. LeClair, Phys. Rev. A **56**, 782 (1997).
[4] L. Allen and J. H. Eberly, *Optical Resonance and Two-Level Atoms* (Dover, New York, 1987).
[5] L. A. Dickey, Commun. Math. Phys. **88**, 27 (1983).
[6] L. A. Dickey, Commun. Math. Phys. **87**, 505 (1983).
[7] V. E. Korepin, N. M. Bogoliubov, and A. G. Izergin, *Quantum Inverse Scattering Method and Correlation Functions* (Cambridge University Press, Cambridge, 1993).

Magnetic monopole in the Feynman's derivation of Maxwell equations

A. Bérard and Y. Grandati
L.P.L.I. Institut de Physique, 1 Boulevard D. Arago, F-57070, Metz, France

H. Mohrbach
L.P.L.I. Institut de Physique, 1 Boulevard D. Arago, F-57070, Metz, France and M.I.T, Center for Theoretical Physics, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307

(Received 25 February 1999; accepted for publication 21 April 1999)

In 1992, Dyson published Feynman's proof of the homogeneous Maxwell equations assuming only the Newton's law of motion and the commutation relations between position and velocity for a nonrelativistic particle. Recently Tanimura gave a generalization of this proof in a relativistic context. Using the Hodge duality we extend his approach in order to derive the two groups of Maxwell equations with a magnetic monopole in flat and curved spaces. [S0022-2488(99)00408-9]

I. INTRODUCTION

Various ways exist to present the Maxwell equations. The usual one is the historical approach in which the empirical basis for each equation is first given. Another remarkable way is exposed in an old unpublished work of Feynman, reported more recently in an elegant paper by Dyson.[1] His proof of Maxwell equations is shown assuming only the Newton's law of motion:

$$F^i = m\frac{d^2x^i}{dt^2}, \tag{1}$$

the commutation relations between position and velocity for a simple nonrelativistic particle, without reference to an action or variational principle:

$$[x^i, x^j] = 0, \tag{2}$$

and

$$m[x^i, \dot{x}^j] = i\hbar\,\delta^{ij}. \tag{3}$$

From this assumption he deduced that $\vec{F}$ has the form of the Lorentz force:

$$F^i(x^k, \dot{x}^k, t) = q\{E^i(x^k, t) + [\vec{\dot{x}} \wedge \vec{B}(x^k, t)]^i\}, \tag{4}$$

where the fields satisfy the first group of Maxwell equations:

$$\begin{gathered} \operatorname{div}\vec{B} = 0, \\ \frac{\partial\vec{B}}{\partial t} + \overrightarrow{\mathrm{rot}}\,\vec{E} = 0. \end{gathered} \tag{5}$$

The remaining two other equations are claimed to be a definition of charge density ϱ and current density $\vec{j}$. This is a nontrivial and very interesting result because first, the basic assumption appears to restrict the interaction of a nonrelativistic particle to the electromagnetic one and second, there is no need for the existence of Hamiltonian or Lagrangian formalism. Note however that Hojman and Shepley[2] showed how under certain conditions, an action could be associated to the Feynman's commutation relations using a Helmholtz variational inverse problem.[3]

Since then, several other authors have contributed to the success of this Feynman's presentation. In particular, Tanimura[4] has proposed a generalization in a Lorentz covariant form with a scalar evolution parameter τ and an extension to the case of non-Abelian gauge theories. In a recent letter[5] we have explained how, if a sO_3 Lie algebra structure is required, a Poincaré magnetic momentum and a Dirac magnetic monopole can be introduced in three dimensions.

In this letter, we extend Tanimura's work in the four-dimensional formalism to find the four Maxwell equations which appear as a consequence of the Jacobi identities. The main part is devoted to the demonstration of the generalized Maxwell equations with a magnetic monopole which requires the Hodge duality of the electromagnetic tensor $F^{\mu\nu}$. In this context the duality between the electric and magnetic fields is also due to the Jacobi identities. The case of a curved space is rapidly treated as a direct application of this formalism.

In this approach we can naturally introduce the fiber bundle formalism and particularly the tangent bundle structure (not the cotangent fiber bundle). This is, however, not essential for our purpose. The space where we work is the configuration space in the physicist's language.

II. MAXWELL EQUATIONS IN $D=4$

Assume a particle of mass m moving in the Minkowski space with position $x^\mu(\tau)$ ($\mu=1,2,3,4$) depending on the parameter τ. Contrary to Dyson, we do not postulate the Newton equation of motion. Instead we consider at the starting point a natural commutation relation:

$$[x^\mu,x^\nu]=0, \tag{6}$$

with the following property:

$$\frac{d}{d\tau}[x^\mu,x^\nu]=[\dot{x}^\mu,x^\nu]+[x^\mu,\dot{x}^\nu]. \tag{7}$$

Differentiating (6) with respect to time gives:

$$[\dot{x}^\mu,x^\nu]+[x^\mu,\dot{x}^\nu]=0, \tag{8}$$

which implies

$$[x^\mu,\dot{x}^\nu]=\frac{g^{\mu\nu}(x)}{m}, \tag{9}$$

where $g^{\mu\nu}(x)$ is a symmetric tensor. A symplectic structure is then defined on the tangent bundle like Souriau[6] in his symplectic classical mechanics.

Following Tanimura[4] we consider in Secs. I and II the case where $g^{\mu\nu}(x)$ is the metric tensor of the Minkowski space:

$$g^{\mu\nu}(x)=g^{\mu\nu}. \tag{10}$$

We also require that the brackets satisfy the relations:

$$\begin{gathered}[A,B]=-[B,A],\\ [A,BC]=[A,B]C+B[A,C]\quad\text{(Leibnitz rule)},\\ [A,[B,C]]+[B,[C,A]]+[C,[A,B]]=0\quad\text{(Jacobi identity)},\end{gathered} \tag{11}$$

where the quantities A, B, and C can be equal to x^μ or $\dot{x}^\mu$.

Combining Eqs. (9)–(11) we get:

$$[x^\mu, f(x^\rho)] = 0,$$
$$[x^\mu, f(x^\rho, \dot{x}^\rho)] = \frac{1}{m}\frac{\partial f(x^\rho, \dot{x}^\rho)}{\partial \dot{x}^\mu}, \tag{12}$$
$$[\dot{x}^\mu, f(x^\rho)] = -\frac{1}{m}\frac{\partial f(x^\rho)}{\partial x^\mu}.$$

On another hand, differentiating (9) gives:

$$[x^\mu, \ddot{x}^\nu] = -[\dot{x}^\mu, \dot{x}^\nu] = \frac{q}{m^2}F^{\mu\nu}, \tag{13}$$

where F is an antisymmetric tensor and q will be associated to the electric charge of the particle in the following.

Using the Jacobi identity (11) and the relations (12) we easily find that $F^{\mu\nu}$ is only position coordinates dependent and then deduce the first group of Maxwell equations:

$$\partial^\mu F^{\nu\rho} + \partial^\nu F^{\rho\mu} + \partial^\rho F^{\mu\nu} = 0. \tag{14}$$

Contracting two indices we find:

$$\partial^\nu F^\mu_\nu = -\partial^\nu F^\mu{}_\nu = -j^\mu, \tag{15}$$

which is the second group of Maxwell equations. Contrary to the three-dimensional case, the second group of Maxwell equations is not imposed as a definition of the matter as in Ref. 1. Here it is a direct consequence of the formalism.

Note that by integrating (13) we get also the equation of motion:

$$m\ddot{x}^\mu = qF^{\mu\nu}(x)\dot{x}_\nu + G^\mu(x), \tag{16}$$

where the field $G^\mu(x)$ satisfies

$$\partial^\mu G^\nu - \partial^\nu G^\mu = 0. \tag{17}$$

Taking the divergence of the "current density" we obtain the current density conservation law:

$$\partial_\mu j^\mu = [\dot{x}_\mu, j^\mu] = [\dot{x}_\mu, [\dot{x}^\nu [\dot{x}^\mu, \dot{x}_\nu]]] = 0. \tag{18}$$

We finally see that Tanimura's approach contains all the Maxwell equations in the presence of an electric source. In Sec. III we extend this formalism by introducing the Hodge duality of the electromagnetic tensor.

III. MAXWELL EQUATIONS WITH MAGNETIC MONOPOLE

Instead of (13) we choose for the definition of the gauge curvature:

$$[x^\mu, \ddot{x}^\nu] = -[\dot{x}^\mu, \dot{x}^\nu] = \frac{1}{m^2}(qF^{\mu\nu} + g{}^*F^{\mu\nu}), \tag{19}$$

where g will be interpreted as the magnetic charge of the Dirac monopole. The * operation is the Hodge duality.

As usually we deduce the equation of motion by integrating (19),

$$m\ddot{x}^{\mu}=(qF^{\mu\nu}(x)+g{}^{*}F^{\mu\nu}(x))\dot{x}_{\nu}+G^{\mu}(x)$$

and in the same manner as Sec. II we find:

$$\partial^{\mu}G^{\nu}-\partial^{\nu}G^{\mu}=0. \tag{20}$$

The Jacobi identity between the velocities (11) gives the following relation between the electromagnetic tensor and his dual:

$$q(\partial^{\mu}F^{\nu\rho}+\partial^{\nu}F^{\rho\mu}+\partial^{\rho}F^{\mu\nu})+g(\partial^{\mu}{}^{*}F^{\nu\rho}+\partial^{\nu}{}^{*}F^{\rho\mu}+\partial^{\rho}{}^{*}F^{\mu\nu})=0, \tag{21}$$

that is:

$$\begin{aligned}\partial^{\mu}F^{\nu\rho}+\partial^{\nu}F^{\rho\mu}+\partial^{\rho}F^{\mu\nu}&=gN^{\mu\nu\rho},\\ \partial^{\mu}{}^{*}F^{\nu\rho}+\partial^{\nu}{}^{*}F^{\rho\mu}+\partial^{\rho}{}^{*}F^{\mu\nu}&=-qN^{\mu\nu\rho},\end{aligned} \tag{22}$$

where $N^{\mu\nu\rho}$ is a tensor to be interpreted.

Using the differential forms language defined on the Minkowski space (M_4) we write the preceding equations in a compact form:

$$\begin{aligned}dF&=gN,\\ d{}^{*}F&=-qN,\end{aligned} \tag{23}$$

where F and ${}^{*}F\in\wedge^{2}(M_4)$ and $N\in\wedge^{3}(M_4)$.

If we put:

$$\begin{aligned}gN&=-{}^{*}k,\\ qN&={}^{*}j,\end{aligned} \tag{24}$$

where j and $k\in\wedge^{1}(M_4)$, we deduce:

$$\begin{aligned}\delta F&=j,\\ dF&=-{}^{*}k,\end{aligned} \tag{25}$$

δ being the usual codifferential,

$$\delta:\wedge^{k}(M_4)\rightarrow\wedge^{k-1}(M_4),$$

defined here as

$$\delta=(-)^{k(4-k+1)+1}({}^{*}d{}^{*}).$$

Interpreting the 1-forms j and k as the electric and magnetic four-dimensional current densities we obtained the two groups of Maxwell equations in the presence of a magnetic monopole. Note that the Jacobi identity imposes a proportionality relation between the two currents.

As in the absence of monopole we find the conservation of the two current densities:

$$\begin{aligned}\partial_{\mu}j^{\mu}&=[\dot{x}_{\mu},j^{\mu}]=\delta j=0,\\ \partial_{\mu}k^{\mu}&=[\dot{x}_{\mu},k^{\mu}]=\delta k=0,\end{aligned} \tag{26}$$

giving the following possible choice for the currents:

$$k^\mu=\rho_m\dot{x}^\mu,$$
$$j^\mu=\rho_e\dot{x}^\mu, \tag{27}$$

where ρ_m and ρ_e are evidently the magnetic and the electric charge densities.

Remark: The differential forms used in this part are of course singular differential forms which are naturally associated to the electric and magnetic current densities of point particles. These densities are then represented also by singular 1-forms defined throughout the whole Minkowski space.

IV. APPLICATION TO A CURVED SPACE

Like Tanimura[4] we now consider in (9) a general metric tensor,

$$[x^\mu,\dot{x}^\nu]=-\frac{g^{\mu\nu}(x)}{m}.$$

Differentiating this last equation we obtain

$$m[x^\mu,\ddot{x}^\nu]=-m[\dot{x}^\mu,\dot{x}^\nu]=-\partial_\rho g^{\mu\nu}\dot{x}^\rho.$$

Following the previous parts we define a new tensor $W^{\mu\nu}(x,\dot{x})$ and its dual, both position and velocity dependent, such that

$$[x^\mu,\ddot{x}^\nu]=-[\dot{x}^\mu,\dot{x}^\nu]=\frac{1}{m^2}(qW^{\mu\nu}(x,\dot{x})+g{}^*W^{\mu\nu}(x,\dot{x})), \tag{28}$$

which implies

$$q[x^\rho,W^{\mu\nu}]+g[x^\rho,{}^*W^{\mu\nu}]=m^2(\partial^\mu g^{\rho\nu}-\partial^\nu g^{\rho\mu}) \tag{29}$$

or

$$qW^{\mu\nu}+g{}^*W^{\mu\nu}=m(\partial^\mu g^{\rho\nu}-\partial^\nu g^{\rho\mu})\dot{x}_\rho+qF^{\mu\nu}(x)+g{}^*F^{\mu\nu}(x). \tag{30}$$

Integrating (28) and using (30) we recover the Lorentz force:

$$F^\mu(x,\dot{x},\tau)=m\frac{d\dot{x}^\mu}{d\tau}=-m\Gamma^{\nu\rho\mu}(x)\dot{x}_\nu\dot{x}_\rho-qF^{\nu\mu}\dot{x}_\nu-g{}^*F^{\nu\mu}\dot{x}_\nu+G^\mu(x,\tau). \tag{31}$$

The Christoffel symbols are defined by

$$\Gamma^{\nu\rho\mu}=\tfrac{1}{2}(\partial^\rho g^{\nu\mu}-\partial^\nu g^{\rho\mu}-\partial^\mu g^{\rho\nu}) \tag{32}$$

and the Hodge duality is

$${}^*F^{\mu\nu}=\frac{1}{2\sqrt{-g}}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}.$$

On another side, like Tanimura in the case without Hodge duality, we easily obtain

$$[\dot{x}_\mu,\dot{x}_\nu]=-\frac{1}{m^2}(qF_{\mu\nu}+g{}^*F_{\mu\nu}).$$

Requiring that the Jacobi identities (11) are also satisfied in this context, and introducing $x_\mu=g_{\mu\nu}x^\nu$, we can then demonstrate that

$$[\dot{x}_\mu,[\dot{x}_\nu,\dot{x}_\rho]]+[\dot{x}_\nu,[\dot{x}_\rho,\dot{x}_\mu]]+[\dot{x}_\rho,[\dot{x}_\mu,\dot{x}_\nu]]=0,$$
$$[x_\mu,[\dot{x}_\nu,\dot{x}_\rho]]+[\dot{x}_\nu,[\dot{x}_\rho,x_\mu]]+[\dot{x}_\rho,[x_\mu,\dot{x}_\nu]]=0, \tag{33}$$

and that the antisymmetric tensors $F^{\nu\mu}$ obeys the same equation as in the Minkowski case:

$$q(\partial_\mu F_{\nu\rho}+\partial_\nu F_{\rho\mu}+\partial_\rho F_{\mu\nu})+g(\partial_\mu^* F_{\nu\rho}+\partial_\nu^* F_{\rho\mu}+\partial_\rho^* F_{\mu\nu})=0.$$

Such an equation is not possible for the tensor $W^{\mu\nu}$, because it is also velocity dependent.

Repeating the same computation as in Sec. III we again arrive at

$$\delta F=j,$$
$$dF=-{}^*k. \tag{34}$$

V. CONCLUSION

The Tanimura covariant extension of Feynman's formalism is a good approach for the study of gauge theories with magnetic monopoles, the two groups of Maxwell equations being a direct consequence of the different Jacobi identities. We have presented two ways to find this result. The first one is an improvement of Tanimura's deduction. The second one is narrowly connected to the Hodge duality and to the presence of a magnetic monopole. The case without Hodge duality can be retrieved by setting the magnetic current density equal to zero. The generalization to a curved space is direct. The problem of the physical meaning of this τ parameter is still open.

[1] F. Dyson, Am. J. Phys. **58**, 209 (1990).
[2] S. A. Hojman and L. C. Shepley, J. Math. Phys. **32**, 142 (1991).
[3] H. Helmholtz and J. Reine, Angew. Math. **100**, 137 (1887).
[4] S. Tanimura, Ann. Phys. (N.Y.) **220**, 229 (1992).
[5] A. Bérard, Y. Grandati, and H. Mohrbach, Phys. Lett A **254**, 133 (1999).
[6] J. M. Souriau, *Structure des Systèmes Dynamiques* (Dunod, Paris, 1970).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

The generalized Moyal–Nahm and continuous Moyal–Toda equations

Carlos Castro[a)]
Departamento de Fisica, Centro de Investigacion y Estudios Avanzados del IPN, Apdo. 14-740 CP 07000, Mexico DF, Mexico and Center for Theoretical Studies of Physical Systems, Clark Atlanta University, Atlanta, Georgia 30314

Jerzy Plebański
Departamento de Fisica, Centro de Investigacion y Estudios Avanzados del IPN, Apdo. 14-740 CP 07000, Mexico DF, Mexico

(Received 27 October 1997; accepted for publication 1 April 1999)

We present in detail a class of solutions to the 4D SU(∞) Moyal–anti-self-dual Yang–Mills (ASDYM) equations (an effective 6D theory) that are related to *reductions* of the generalized Moyal–Nahm equations using the Ivanova–Popov ansatz. The former yields solutions to the ASDYM/SDYM equations for arbitrary gauge groups in four dimensions. A further dimensional reduction of the above effective 6D equations yields solutions to the Moyal–anti-self-dual gravitational equations in four dimensions. The self-dual Yang–Mills/self-dual gravity case requires a separate study. The SU(2) Toda lattice and SU(∞) (continuous) Moyal–Toda lattice equations are derived from the Moyal–Nahm equations. An explicit map taking the Moyal heavenly form (after a rotational Killing symmetry reduction of the Moyal heavenly equations) into the SU(2) Toda lattice field is found. Finally, the generalized Moyal–Nahm equations are conjectured that contain the (continuous) SU(∞) Moyal–Toda lattice equations, after a suitable reduction process. Embeddings of the different types of Moyal–Toda lattice equations into the Moyal–Nahm equations are described.
[S0022-2488(99)00508-3]

I. INTRODUCTION

The quantization program of the 3D continuous Toda theory (2D Toda molecule) is a challenging enterprise that we believe would enable us to understand many of the features of the quantum dynamics and spectra of the quantum self-dual membrane.[1] This is based on the observation that the light-cone-gauge (spherical) supermembrane moving in a D-dimensional flat space–time background has a correspondence with a $D-1$ SU(∞) super Yang–Mills (SYM) theory, dimensionally reduced to one temporal dimension; i.e., with an SU(∞) supersymmetric gauge quantum mechanical model (matrix model).[2,3]

It was shown in Ref. 1 that exact (particular) solutions of the $D=11$ light-cone (spherical) supermembrane, related to the $D=10$ SU(∞) SYM theory, could be constructed based on a particular class of reductions of the SYM equations from higher dimensions to four dimensions.[4] In particular, solutions of the $D=10$ YM equations given by the $D=10$ YM potentials, $\mathcal{A}_\mu$, can be obtained in terms of the 4D YM potentials, A_i, that obey the $D=4$ self-dual YM (SDYM) equations. Dimensional reductions of the latter SU(∞) SDYM equations to one temporal dimension are equivalent to the SU(∞) Nahm equations in the temporal gauge $A_0=0$.

Finally, the embedding of the continuous SU(∞) Toda equation into the SU(∞) Nahm equations was performed in Ref. 1 based on the connection between the $D=5$ self-dual membrane and the SU(2) Toda molecule/chain equations.[5] A continuous Toda theory in connection to self-dual

[a)]Electronic mail: castro@ctsps.cau.edu

gravity was also found by Chapline and Yamagishi[6] in the description of a three-dimensional version of anyon superconductivity. Based on the theory of gravitational instantons, a 3D model describing the condensation of quasi-particles (*chirons*) with properties related to fractional statistics was found.

The classical Toda theory can be obtained also from a rotational Killing symmetry reduction[7] of the 4D self-dual gravitational (SDG) equations expressed in terms of first heavenly form that furnish (complexified) self-dual metrics of the form: $ds^2=(\partial_{x^i}\partial_{\tilde{x}^j}\Omega)dx^i d\tilde{x}^j$ for $x^i=y,z$; $\tilde{x}^j=\tilde{y},\tilde{z}$ and Ω is the first heavenly form. The latter equations can, in turn, be obtained from a dimensional reduction of the 4D SU(∞) SDYM equations, an effective 6D theory (Refs. 8 and 9 and references therein). The Lie algebra su(∞) was shown to be isomorphic (in a basis-dependent limit) to the Lie algebra of area-preserving diffeomorphisms (diffs) of a 2D surface, $sdiff(\Sigma)$ by Hoppe.[2] It is for this reason that a Weyl–Wigner–Moyal (WWM) quantization of the reductions of the first heavenly equation will be used in this letter.

Using our results of Ref. 10 based on Ref. 11 we have shown that a Weyl–Wigner–Groenwold–Moyal (WWGM)[12] quantization approach yields a straightforward quantization scheme for the 3D continuous Toda theory (2D Toda molecule). Supersymmetric extensions can be carried out following Ref. 8 where we wrote down the supersymmetric analog of the heavenly equations for SD supergravity.

There are some differences between our results and those which in general have appeared in the literature. Among these are (i) one is *not* taking the limit of $\hbar\rightarrow 0$ while having $N=\infty$ in the classical SU(N). Recently, Fairlie[13] has written solutions to Moyal–Nahm equations, with $\hbar\neq 0$ for the eight transverse membrane coordinates in $D=11$ in terms of spinors using the WWM formulation. (ii) We are working with the generalized Moyal–Nahm equations involving a Moyal bracket wrt an *enlarged* phase space, q,p,q',p' and not with the standard SU(2) Moyal–Nahm equations involving a Moyal bracket wrt q,p only. We have $\hbar\neq 0$, $N=\infty$ simultaneously. (iii) The connection with the self-dual membrane and W_∞ algebras was proposed in Ref. 1. The results of Ref. 11 become very useful in the implementation of the WWM quantization program and in the embedding of the SU(2) Moyal–Nahm solutions into the generalized Moyal–Nahm equations studied in the present work.

In the next section we present in detail a class of solutions to the 4D SU(∞) Moyal anti-self-dual Yang–Mills equations that are related to *reductions* of the generalized Moyal–Nahm equations using the Ivanova–Popov ansatz. The former yields solutions to the anti-self-dual Yang–Mills (ASDYM)/SDYM equations for arbitrary gauge groups. A further dimensional reduction yields solutions to the Moyal anti-self-dual gravitational equations. The self-dual Yang–Mills/self-dual gravity case requires a separate study.

We write down the explicit Moyal quantization of all the equations involved in the Ivanova–Popov construction.[4] In particular, deformations of the ordinary Laplace equation in four dimensions are required. This is necessary in order to write down the equation governing the deformations of the scalar field involved in the construction of Ref. 4, and which is mapped into the Moyal deformed continuous Toda field via deformations of the twistor transform. Deformations of twistor surfaces have yet to be constructed. For comments in that direction we refer to the work of Strachan and Takasaki[14,15] in their study of higher-dimensional integrable models.

In Sec. III the SU(2) Toda lattice and (continuous) SU(∞) Moyal–Toda lattice-type equations are derived from Moyal–Nahm equations. By ''SU(∞) Moyal'' one means the Moyal deformations of the symplectic diffeomorphisms in four dimension instead of two. This will be explained in detail in the text. The Legendre-like transform between solutions to the rotational Killing symmetry reductions of the 4D Moyal heavenly equations and the SU(2) Toda lattice equations is studied, in particular, the explicit map taking the Moyal heavenly form into the SU(2). The Toda lattice field is found.

Finally, in Sec. IV, the generalized Moyal–Nahm equations are provided that *contain* the (continuous) SU(∞) Moyal–Toda lattice-type equations after a suitable reduction. Embeddings of the different types of Moyal–Toda lattice equations into the Moyal–Nahm equations are provided.

The most salient feature of the generalized Moyal–Nahm equation is that it involves an effective 8D theory associated with symplectic diffeomorphisms of a 4D manifold. This 8D effective theory may have an important role in understanding the quantum dynamics of the 11D supermembrane based on the the m(atrix) membrane models and their integrability properties.[3,13]

In the conclusion we discuss (among other things) briefly how deformation quantization techniques for higher-dimensional volume forms, the Zariski quantization,[16] may be used to quantize p branes. Other types of deformations are possible, like those which give up the *associative* character of the Moyal product, the so-called q deformations, and which can be used to construct deformations of the self-dual membrane (we refer to Ref. 17). We expect that a q-Moyal deformation program of the self-dual membrane might yield important information about how to quantize the full membrane theory beyond the self-dual exactly integrable sector and that the particle/soliton spectrum of the underlying conformal affine Toda models will shed some light onto the particle content of the more general theory.[1,10]

Note: Sometime after this work was completed we were informed[18] based on Strachan's work[14] that a map from the master equation involving the Moyal deformations of the SDYM equations to Strachan's SU(2) Toda lattice equation (a Toda lattice model whose discrete spacing is given by multiples of $\hbar$) could be obtained by a suitable dimensional reduction. However, the latter reduction is *not* a rotational Killing symmetry reduction studied in this work; i.e, the particular dimensional reduction of the effective 6D master equation is *not* a Moyal deformation of the heavenly equation (although the heavenly equation can be obtained from the master equation). The results of Ref. 18 are based on a *two*-step process where *after* performing the WWGM map, an explicit introduction of $\hbar$ is put in by hand. More on this shall be explained in the text.

Remark: In this paper we use the following abbreviations: ''Moyal–Nahm equation'' stands for ''the Moyal deformation of the Nahm equation;'' ''Moyal–Toda equation'' means ''the Moyal deformation of the Toda equations,'' etc.

II. THE SU(∞) MOYAL ASDYM/SDYM EQUATIONS

We will study in this section the Moyal 4D SU(∞) ASDYM/SDYM in connection to the ASDG/SDG equations and the continuous 2D Toda molecule. To begin with, the Moyal bracket of two YM potentials $A_y, A_{\tilde{y}}$, for example, can be expanded in powers of $\hbar$ as in Ref. 14:

$$\{A_y, A_{\tilde{y}}\}_{q,p} \equiv \sum_{s=0}^{\infty} \frac{(-1)^2 \hbar^{2s}}{(2s+1)!} \sum_{l=0}^{2s+1} (-1)^l (C_l^{2s+1}) [\partial_q^{2s+1-l} \partial_p^l A_y][\partial_p^{2s+1-l} \partial_q^l A_{\tilde{y}}], \tag{1}$$

where C_l^{2s+1} are the binomial coefficients.

The crucial difference between the solutions of the SU(2) Moyal–Nahm equations[11] and the generalized Moyal–Nahm case is that one *must* have an *extra* explicit dependence on another *set* of phase space variables, q', p'. In particular, those *reductions* of the generalized Moyal–Nahm equations that are linked to the Moyal–Toda equations must have an extra t dependence for the YM potentials. The continuous Toda molecule equation as well as the usual Toda system may be written in the double commutator form of the Brockett equation:[19]

$$\frac{\partial L(\tau,t)}{\partial \tau} = [L, [L, H]]. \tag{2}$$

Here L has the form

$$L \equiv A_+ + A_- = X_o(-iu) + X_{+1}(e^{(\rho/2)}) + X_{-1}(e^{(\rho/2)}), \tag{3}$$

with the connections $A_\pm$ taking values in the subspaces $\mathcal{G}_o \oplus \mathcal{G}_{\pm 1}$ of some $\mathbf{Z}$-graded continuum Lie algebra $\mathcal{G} = \oplus_m \mathcal{G}_m$ of a novel class. Here $H = X_o(\kappa)$ is a continuous limit of the Cartan element of the principal sl(2) subalgebra of $\mathcal{G}$. The functions $\kappa(\tau,t), u(\tau,t), \rho(\tau,t)$ satisfied certain equations given in Ref. 19.

Upon the elimination of u one obtains the Toda equation. A naive look at (2) might beg the question: Where is the t dependence? The t dependence is *encoded* in the continuum algebra commutation relations of the generators that are parametrized by a family of functions depending on both τ and t. It is in this fashion that the t dependence makes its presence in (2).

To implement the WWM prescription, one may consider the case when **G** is a group of unitary operators acting in the Hilbert space of square integrable functions on the line, $L^2(R^1)$. Then, $\mathcal{G}$ is now the associated (continuum) Lie algebra of self-adjoint operators acting in the Hilbert space, $L^2(R^1)$. The following operator-valued quantities depend on the two coordinates, τ and t, obey the operator version of the Brockett equation, and, after the WWM map, read

$$\frac{\partial \hat{L}(\tau,t)}{\partial \tau}=\frac{1}{i\hbar}[\hat{L},\frac{1}{i\hbar}[\hat{L},\hat{H}]]\leftrightarrow\frac{\partial\mathcal{L}}{\partial\tau}=\{\mathcal{L},\{\mathcal{L},\mathcal{H}\}\}, \tag{4}$$

where $\mathcal{L}(\tau,t,q,p;\hbar)$ and $\mathcal{H}(\tau,t,q,p;\hbar)$ are the corresponding elements in the phase space after performing the WWM map. The main problem with this approach is that we do *not* have representations of the continuum **Z**-graded Lie algebras in the Hilbert space, $L^2(R^1)$, and, consequently, we cannot evaluate the matrix elements $\langle q-(\xi/2)|\hat{L}(r,\tau)|q+(\xi/2)\rangle;\langle\hat{H}\rangle$. For this reason we have to return to another method to solve this problem.

The quantities $\mathcal{L}$ and $\mathcal{H}$ are defined as

$$\mathcal{L}(\tau,t,q,p;\hbar)\equiv\int_{-\infty}^{\infty}d\xi\left\langle q-\frac{\xi}{2}\right|\hat{L}(\tau,t)\left|q+\frac{\xi}{2}\right\rangle\exp\left[\frac{i\xi p}{\hbar}\right], \tag{5}$$

$$\mathcal{H}(\tau,t,q,p;\hbar)\equiv\int_{-\infty}^{\infty}d\xi\left\langle q-\frac{\xi}{2}\right|\hat{H}(\tau,t)\left|q+\frac{\xi}{2}\right\rangle\exp\left[\frac{i\xi p}{\hbar}\right]. \tag{6}$$

The latter matrix elements, if known, suffice to determine the quantities $\mathcal{L}$ and $\mathcal{H}$ associated with the 2D continuous Toda molecule equation.

Despite not knowing the explicit operator form of $\hat{L}(\tau,t)$ and $\hat{H}(\tau,t)$ acting in the Hilbert space, $L^2(R^1)$, one may still write down solutions for the continuous Moyal–Toda equation. This can be achieved starting from the original Moyal SDYM/ASDYM equations associated with the SU(∞) group in $D=4$ and looking for solutions. Ivanova and Popov,[4] in a summary of YM equations in $D\geqslant 4$, have discussed that solutions to the ASDYM/SDYM equations in $D=4$ for an arbitrary Lie group, G, which are *linked* to the Nahm equations, may be obtained from the ansatz:

$$A_\mu=-\eta^\alpha_{\mu\nu}T_\alpha(\phi(x^\mu))\partial_\nu\phi,\quad \eta^\alpha_{\mu\nu}=\epsilon^\alpha_{\beta\gamma},\quad \eta^\alpha_{\mu 0}=-\eta^\alpha_{0\mu}=\delta^\alpha_\mu. \tag{7}$$

The t'Hooft matrices obey the quaternionic algebra: $\eta^\alpha_{\mu\lambda}\eta^\beta_{\lambda\nu}=-\delta^{\alpha\beta}\delta_{\mu\nu}-\epsilon^{\alpha\beta\gamma}\eta^\gamma_{\mu\nu}$. The function ϕ obeys $\partial_\mu\partial^\mu\phi=0$ (ASDYM) and $\phi=x_\mu x^\mu$ (SDYM), and the three Lie algebra-valued *scalar* functions $T_\alpha(\phi)=T^A_\alpha(\phi)L_A$, for $\alpha=1,2,3$, satisfy the Nahm equations wrt the ϕ function:

$$\epsilon_{\alpha\beta\gamma}\frac{dT_\gamma}{d\phi}=\pm[T_\alpha,T_\beta], \tag{8}$$

where the $\pm$ corresponds to the SDYM/ASDYM case. Notice that the simple reflection $T_\alpha\to -T_\alpha$ converts the SDYM to the ASDYM solutions with the proviso that now ϕ obeys the 4D Laplace equation. It is very important to emphasize that Ivanova and Popov used a Euclidean space–time signature. This will become important later on when we discuss other results obtained in a $(++,--)$ signature.

The Ivanova–Popov ansatz, for Euclidean signatures, will yield solutions to the Moyal deformations of the anti-self-dual gravitational equations in four dimensions from dimensional reductions of the SU(∞) ASDYM equations. However, this will not be the case for the self-dual gravitational equations that can be obtained from reductions of the SU(∞) SDYM equations. The

Ivanova–Popov ansatz will not yield solutions to the Moyal deformations of the SDG equations. Another type of solutions will be required. For signatures $2+2$, the situation is reversed. More on this issue will be clarified in the next sections.

A WWM quantization requires writing down the symbol map of the *operators* acting in the Hilbert space, $L^2(R^1)$, associated with the three Lie-algebra-valued functions, $\hat{T}_\alpha$, so the Moyal–Nahm equations are

$$\epsilon_{\alpha\beta\gamma}\frac{\partial \mathcal{T}_\gamma}{\partial \phi}=\pm\{\mathcal{T}_\alpha,\mathcal{T}_\beta\}_{\text{Moyal}}, \tag{9a}$$

with $\mathcal{T}_\alpha(\phi(x^\mu;q,p,\hbar);q,p,\hbar)=\text{symbol}[\hat{T}_\alpha]$, where $\hat{T}_\alpha$ is a representation of the Lie-algebra-valued operators in $L^2(R^1)$. Shortly we will explain why ϕ acquires also a $q,p,\hbar$ dependence.

Rigorously speaking one should write $\mathcal{T}_\alpha[\mathcal{G}]$ to include the explicit dependence of $\mathcal{T}_\alpha$ on the Lie algebra $\mathcal{G}$ involved initially in the construction. In the case that $G=\text{SU}(\infty)$, one is required then to *extend* the symplectic diffeomorphisms in two to four dimensions. The Moyal bracket involves now a generalized phase space q_i,p_i for $i=1,2$. For $G=\text{SU}(2)$ the fact that the dual of the SU(2) Lie algebra is R^3 allows us to establish the *correspondence* among the three scalars $\mathcal{T}_\alpha$ with the three components X,Y,Z of a four-vector, after fixing the gauge $A_0=0$, replacing ϕ by τ and making the correspondence $A_x\leftrightarrow X$, $A_y\leftrightarrow Y$, $A_z\leftrightarrow Z$. In this way one retrieves the SU(2) Moyal–Nahm equations

$$\frac{\partial X}{\partial \tau}=\{Y,Z\},\ \ldots$$

for the three vector-valued functions $X_i=X_i(\tau;q,p,\hbar)$. Notice that, before, the Moyal–Nahm functions, $\mathcal{T}_\alpha$, are *scalar*-valued as well as ϕ.

Furthermore, if Moyal deformations of the Toda equations are related to the Moyal–Nahm equations via the Lax pair formalism, the scalar $\phi(x^\mu)$ will require, in general, a deformation of the type $\phi(x^\mu;q_i,p_i,\hbar)$ obeying a deformation of the Laplace equation. The latter equation is, in general, modified to include $\hbar$ corrections and derivatives acting on the phase space coordinates as well. A natural deformation of the Laplace equation is simply given by $\partial_\mu * \partial^\mu \phi(x^\mu;q,p,\hbar)=0$. Writing explicitly, the above equation yields

$$\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{i\hbar}{2}\right)^n \omega^{i_1j_1}\omega^{i_2j_2}\cdots\omega^{i_nj_n}(\partial_{i_1}\partial_{i_2}\cdots\partial_{i_n}\partial_\mu)(\partial_{j_1}\partial_{j_2}\cdots\partial_{j_n}\partial^\mu)\phi(x^\mu;q,p,\hbar)=0. \tag{9b}$$

Equations (9a) and (9b) are the *defining* Moyal deformations of the Ivanova–Popov equations. The deformed YM potential A_μ obtained from the ansatz (7) will then be equal to the Moyal star product:

$$A_\mu=-\eta^\alpha_{\mu\nu}T_\alpha[\phi(x^\mu,q,p,\hbar);q,p,\hbar]*\partial_\nu\phi(x^\mu,q,p,\hbar), \quad \eta^\alpha_{\mu\nu}=\epsilon^\alpha_{\beta\gamma}, \quad \eta^\alpha_{\mu 0}=-\eta^\alpha_{0\mu}=\delta^\alpha_\mu. \tag{9c}$$

Equations (9a)–(9c) are the Moyal deformations of the Ivanov–Popov construction. A particular solution to the deformed Laplace equation could be of the type $\phi=\Sigma\hbar^n\phi_n$. The $\hbar=0$ limit yields the ordinary Laplace equation for the function $\phi_0(x^\mu)$ and the others $\phi_n(x^\mu,q,p)$ contain an extra q,p dependence also.

This can be explicitly seen when one performs a direct Moyal quantization program of all the equations involved in the Ivanova–Popov construction. Operators are mapped into functions of phase space via the symbol map. Products of operators are mapped into the Moyal star product of their corresponding symbols and this involves derivatives of arbitrary order wrt q,p. Furthermore, a suitable ordering prescription must be specified *a priori*. Therefore, a Moyal quantization of the Ivanov–Popov construction induces a $q,p,\hbar$ dependence on the scalar ϕ and it deforms the original Laplace equation. We shall go back to this issue when we discuss the twistor transform mapping the nonlinear 3D Toda equation into the 3D Laplace equation for the scalar ϕ.

A particular class of solutions of the SU(2) Moyal–Nahm equations, Eq. (9a), in terms of the Jacobi elliptic functions wrt the *undeformed* ϕ has been given by[11]

$$\mathcal{T}_1 = sn[\phi]\left[\frac{i}{2}p(q^2-1) - \hbar\left(\beta+\frac{1}{2}\right)q\right], \quad \mathcal{T}_2 = dn[\phi]\left[-\frac{1}{2}p(q^2+1) - i\hbar\left(\beta+\frac{1}{2}\right)q\right],$$
$$\mathcal{T}_3 = cn[\phi]\left[-ipq - \hbar\left(\beta+\frac{1}{2}\right)\right], \quad \beta=\text{const}, \tag{10}$$

where the (undeformed) scalar function ϕ has a correspondence with *one*, and only one, temporal parameter, $\phi \to \tau$. When one replaces $\phi \to \tau$, the ansatz of (7) gives that $A_0=0$, $A_i \sim \mathcal{T}_i$, and, as expected, the SU(2) Moyal–Nahm equations involve the three components of a *four*-vector and the derivatives are taken wrt the temporal variable τ that does *not* transform as a scalar like ϕ.

It is important also to remark that if ϕ acquires a deformation $\phi(x^\mu;q,p,\hbar)$, the solutions to (9a) given by Eqs. (10) *no* longer hold. This is because the Moyal brackets affect also the q,p terms contained in $\phi=\phi(x^\mu,q,p,\hbar)$. To find solutions to the set of Eqs. (9a) and (9b) is a very difficult enterprise, mainly because the *deformed* scalar field ϕ obeys a very complicated differential equation. In the undeformed case, ϕ obeys the ordinary Laplace equation and the solutions of (9a) given by Eqs. (10) will be valid then. For the deformed $\phi=\phi(x^\mu,q,p,\hbar)$ case, the Moyal–Ivanov–Popov equations read after using the chain rule

$$\epsilon_{\alpha\beta\gamma}\frac{\partial \mathcal{T}_\gamma}{\partial \phi} = \pm\{\mathcal{T}_\alpha,\mathcal{T}_\beta\}_{\text{Moyal}} = \mathcal{T}_{\alpha\beta}[(\partial_q\phi)(\partial_\phi \mathcal{T}_\alpha);(\partial_p\phi)(\partial_\phi \mathcal{T}_\beta);\ldots;\partial_q \mathcal{T}_\alpha,\partial_p \mathcal{T}_\beta,\ldots], \tag{11a}$$

where $\mathcal{T}_{\alpha\beta}$ is a very complicated function of the derivatives of $\mathcal{T}_\alpha,\mathcal{T}_\beta$ wrt the ϕ function and the q,p variables. Also required is the knowledge of the derivatives of ϕ wrt the q,p variables. This involves first finding a solution to Eq. (9b) prior to writing (11). There is, however, a crucial condition needed to be imposed on the solutions to the deformed Laplace equation if Eq. (11a) is a meaningful set of differential equations involving *solely* functions of $\phi,q,p,\hbar$. The lhs of (11a) involves solely $\phi,q,p,\hbar$; therefore, the rhs of (11a) must also. This requires to set the derivatives $\partial_q\phi\partial_p\phi$ to be of the form (11c) such that the Moyal–Nahm equations have indeed the correct homogeneous form:

$$\epsilon_{\alpha\beta\gamma}\frac{\partial \mathcal{T}_\gamma}{\partial \phi} = \pm\{\mathcal{T}_\alpha,\mathcal{T}_\beta\}_{\text{Moyal}} = \mathcal{F}_{\alpha\beta}[\phi,q,p,\hbar]. \tag{11b}$$

For this last equation to hold one requires to fix the *first* derivatives of ϕ wrt the q,p variables to have the form

$$\partial_q\phi = F_q(\phi), \quad \partial_p\phi = F_p(\phi) \Rightarrow \partial_q^2\phi = F_q(\phi)\partial_\phi F_q(\phi), \quad \partial_p^2\phi = F_p(\phi)\partial_\phi F_p(\phi)\cdots,$$
$$\partial_p\partial_q\phi = \partial_q\partial_p\phi \Rightarrow F_p(\phi)\partial_\phi F_q(\phi) = F_q(\phi)\partial_\phi F_p(\phi). \tag{11c}$$

The mixed space–time derivatives are

$$\partial_\mu\partial_q\phi = (\partial_\mu\phi)(\partial_\phi F_q(\phi)), \quad \partial_\mu\partial_p\phi = (\partial_\mu\phi)(\partial_\phi F_p(\phi))\cdots. \tag{11d}$$

Hence, the equation $(\partial_\mu * \partial^\mu)\phi = 0$ will involve an infinite number of derivative terms of the form

$$\partial_\mu^2\phi, \quad (\partial_\mu\phi)^2, \quad \partial_\phi^n F_q(\phi), \quad \partial_\phi^n F_p(\phi), \quad \ldots. \tag{11e}$$

In other words, the meaningful set of solutions to the Moyal–Nahm equations (9a), (11a), and (11b) are parametrized by a family of solutions to the deformed Laplace equation and these, in turn, are given by a family (moduli) of functions $F_q(\phi), F_q(\phi)$ that restrict the *first* derivatives of ϕ wrt the q,p variables to be of the form (11c). Without these homogeneity conditions, Eq. (9a) could not have the desired form (11b).

Things simplify considerably in the undeformed scalar ϕ case because the rhs's of the Moyal–Nahm equations yield *ordinary* derivatives wrt the q,p variables so that Eq. (10), for example, will be a valid solution for any scalar ϕ obeying the 4D Laplace equation: one has *factorized* completely the ϕ dependence from the q,p one in the Moyal–Nahm functions $\mathcal{T}_\alpha$. No restrictions like (11c) will be needed.

By SU(2) Moyal–Nahm it is meant that the solutions (10) in Ref. 11 were obtained by performing the WWM map which takes su(2) Lie-algebra-valued operators belonging to the Hilbert space, $L^2(R^1)$, into functions of q,p. Before hand, the Wolf representation of the su(2) Lie-algebra-valued generators, in matrix form using the Pauli spin matrices, $A_i^a \tau_a$, needs to be used prior to evaluating the WWM map, i.e, it was essential to use a representation which takes the three Pauli spin SU(2) matrices into three known operators in $\hat{q},\hat{p}$. It is in this sense that one may speak of the solutions (10) to the SU(2) Moyal–Nahm equations.

Since representations of SU(∞) in terms of operators in the Hilbert space $L^2(R)$ are *not* known (as far as we know), one cannot evaluate explicitly the WWM map. In addition, SU(∞) requires use of the *extended* phase space which implies that the Moyal bracket to be used in (9) will be the one wrt the q_i, p_i phase space coordinates rather than to q,p. We have symplectic diffeomorphisms in four dimensions instead of two dimensions. Therefore, the generalized Moyal–Nahm equations require an *extra* set of variables q', p' that must be introduced to account for the area-preserving diffeomorphisms algebra associated with a 4D manifold instead of a two-dimensional surface (sphere, torus). So, now we have $\mathcal{T}_\alpha(\phi; q,p,\hbar; q',p')$. The Moyal brackets are then computed wrt the *enlarged* phase space involving the q,p and q',p'. We shall discuss this in detail in the last section.

Nevertheless, the generalized ''SU(∞)'' Moyal–Nahm equations admit *reductions* to the continuous Moyal–Toda equations and equations directly linked to the 4D SU(∞) Moyal ASDYM/SDYM equations. We refrain from using the term SU(∞) Moyal–Nahm because it is *not* really a WWM quantization of the classical SU(∞) Nahm equations but, instead, one has Moyal deformations of the algebra of symplectic diffeomorphisms in 4D. Such algebra is an infinite dimensional extension of the area-preserving diffeomorphisms of a two-dimensional surface.[20]

The ASDYM equations in 4D Euclidean space–time read $F_{y\tilde{y}} + F_{z\tilde{z}} = 0$ and $F_{yz} = 0$, $F_{\tilde{y}\tilde{z}} = 0$. For signature $(++,--)$ the situation is *reversed*; the SDYM equations are the ones of the form $F_{y\tilde{y}} + F_{z\tilde{z}} = 0$ instead. The Moyal deformed YM potentials are related to the Moyal heavenly form as follows:[9,10]

$$\partial_{\tilde{z}} A_{\tilde{y}} = \partial_{\tilde{w}} A_{\tilde{y}} = \partial_{\tilde{w}} \partial_w \Omega + \tfrac{1}{2}, \tag{12a}$$

$$\partial_{\tilde{y}} A_{\tilde{z}} = \partial_w A_{\tilde{z}} = \partial_{\tilde{w}} \partial_w \Omega - \tfrac{1}{2}, \tag{12b}$$

$$w = z + \tilde{y}, \quad \partial_w = \partial_z = \partial_{\tilde{y}}, \quad \tilde{w} = \tilde{z} - y, \quad \partial_{\tilde{w}} = \partial_{\tilde{z}} = -\partial_y, \tag{13}$$

where Ω is a solution of the Moyal heavenly equation:

$$\{\Omega_{,w}, \Omega_{,\tilde{w}}\}_{q,p} = \pm 1 \leftrightarrow \{\partial_{\tilde{z}} + A_{\tilde{z}}, \partial_{\tilde{y}} + A_{\tilde{y}}\} = 0. \tag{14}$$

The -1 is assigned to the ASDYM in four dimensions with signature $2+2$ related to anti-self-dual gravity. The $+1$ is assigned to the SDYM in four dimensions with signature $2+2$ related to self-dual gravity. For Euclidean signatures, the situation is *reversed*: The -1 is assigned to the SDYM in four dimensions with signature $4+0$ related to self-dual gravity. The $+1$ is assigned to the ASDYM in four dimensions with signature $4+0$ related to anti-self-dual gravity. The important thing to remember is that the scalar field obeying deformations of the Laplace equation is the solution connected to self-dual gravity in $2+2$ that corresponds precisely to the ASDYM equations in $4+0$ given by Ivanova and Popov.

The Moyal deformed YM potentials in terms of the Moyal–Nahm functions $\mathcal{T}_\alpha$ are obtained directly from the ansatz (9c) and, finally, the Moyal heavenly form, up to an integration ''constant'' $f(q,p)$ is

$$\Omega(w,\tilde{w},q,p,\hbar)=\int \left(A_{\tilde{y}}-\frac{1}{2}\tilde{w}\right) dw+\int \left(A_{\tilde{z}}+\frac{1}{2}w\right) d\tilde{w} \tag{15}$$

with

$$\tilde{y}=x^2-ix^3, \quad \tilde{z}=x^0+ix^1, \quad y=x^2+ix^3, \quad z=x^0-ix^1, \tag{16a}$$

and

$$A_{\tilde{y}}=A_2-iA_3, \quad A_{\tilde{z}}=A_0+iA_1. \tag{16b}$$

Equation (15) is the main equation of this section. It establishes the direct relation between solutions of the 4D Moyal heavenly equation in terms of the Moyal YM potentials via solutions of the Moyal–Nahm equations (9a). The relation between the YM potentials and the Moyal–Nahm functions is provided by Eqs. (9c) and (16b). We shall go back to Eq. (15) in Sec. III C when we discuss the Legendre-like transform relating rotationally Killing symmetry reductions of the 4D Moyal heavenly equations (from four to three dimensions), given by Ω, to solutions of the $2+1$ SU(2) Lattice equations. The latter are obtained from Moyal deformations of the SU(2) Nahm equations derived by Strachan.[14]

These reductions of the Moyal SU(2) SDYM/ASDYM equations in four dimensions, an *effective* 6D theory, to a final 4D equation, the Moyal heavenly equation given in Eq. (13), are also compatible with the fact that the Toda equations for SU(N) are obtained from particular reductions of Nahm equations which, in turn, can be represented in a Lax pair form:

$$L=T_1(\phi)+iT_2(\phi), \quad iT_3(\phi)=M, \quad \frac{dL}{d\phi}=[L,M]. \tag{17}$$

Consequently, in the $N=\infty$ limit, one can recast the continuous Moyal–Toda equations in the double commutator form after establishing the following *correspondence* (which are *not* identifications) [see (4)]:

$$\mathcal{L}(u,\rho)\leftrightarrow \mathcal{T}_1+i\mathcal{T}_2, \quad i\mathcal{T}_3=\mathcal{M}\leftrightarrow\{\mathcal{L},\mathcal{H}\}, \quad \frac{\partial\mathcal{L}}{\partial\tau}=\{\mathcal{L},\{\mathcal{L},\mathcal{H}\}\}, \tag{18}$$

where the $\mathcal{T}_\alpha$ obey (9).

The meaning of Eq. (18) is the following: The data $\{\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3\}$, the Moyal–Nahm functions, can be mapped into the three functions $\{\mathcal{L},\mathcal{H},\mathcal{M}\}$ belonging to the Lax–Brockett formalism of continuum **Z**-graded Lie algebras.[19] Equation (18) establishes the correspondence among these data: a functional embedding of the data $\mathcal{T}_\alpha$, parametrized by one function ϕ, into the data $\{\mathcal{L},\mathcal{H},\mathcal{M}\}$ parametrized by the three functions: u,ρ,κ. With ρ being the (Moyal deformations of the) continuous Toda field. Finding the differential equations for the remaining two functions u,κ in the Moyal deformed case and eliminating them to arrive at the final equation involving solely the continuous Moyal–Toda field ρ is a highly nontrivial matter. This will involve Moyal deformations of continuum **Z**-graded Lie algebras. This subject has not been studied to our knowledge. In Sec. IV we shall go back to the Lax–Brockett formalism.

The reason that one can make the correspondences (which are *not* identifications) given by (18) is because there are continuum Lie algebras that are isomorphic to Poisson bracket algebras

$\sim$su(∞), which correspond to the Lie algebras of area-preserving diffeomorphisms of the sphere or torus.[19]. It is in this sense that the correspondence of Eq. (18) is implemented. There are two ways to retrieve Moyal–Toda equations: one way is to use the Lax–Brockett double commutator form, and another is to use the Lax representations for the SU(∞) Nahm equations. The correspondence between these two constructions of the Moyal equations originates from the fact that there is a Legendre-like transform that maps solutions of the $2+1$ continuous Toda equation to those of the 3D Laplace equation, i.e., $\rho \rightarrow \phi$.

The $2+1$ continuous Toda equation occurs in the theory of self-dual Einstein spaces and has a well-known Eguchi–Hanson solution.[21] Prasad[22] (and discussed by Chapline in Ref. 6) has shown that by a change of variables one can transform the Toda equation into a three-dimensional Laplace equation for a certain function $\phi=\phi(\rho)$ related to the two-center gravitational instanton of Gibbons and Hawking.[21] So the correspondence dictated by Eq. (18) is a reflection of the Legendre-like map which takes the Toda field ρ to ϕ obeying the Laplace equation after a dimensional reduction from $2+1$ to 2 dimensions. The Legendre-like transform determines the correspondence given in (18), once the suitable maps from the remaining u,κ functions to ϕ are found.

As stated earlier, a subtlety will now arise. Due to the Moyal deformations of the Toda equations, it is expected then that an accompanying *deformation* of the Laplace equation for the scalar field ϕ follows. The Prasad map taking $\phi \rightarrow \rho$ must be deformed as well. Therefore one should be forced to include $\hbar$ corrections to the scalar $\phi(x;\hbar)$. Similar considerations have been found in Ref. 18 *after* employing the WWGM map: an explicit introduction of $\hbar$ was made afterwards. Nevertheless, the correspondence given by Eq. (18) still holds once the deformed map from ϕ (obeying deformations of Laplace equation) to ρ (obeying deformations of the Toda equation) is found. Similar arguments apply to remaining functions u,κ; these also acquire $\hbar$ corrections. The Prasad[6,22] map will be studied in Sec. III A.

The sort of dimensional reductions of the 4D SU(∞) ASDYM/SDYM equations that we are studying in this section requires that one perform a series of coordinate redefinitions and dimensional reductions from the effective 6D theory to a 4D one:

$$\{x^0,x^1,x^2,x^3;q,p\} \rightarrow \{y,z,\tilde{y},\tilde{z};q,p\} \rightarrow \{w,\tilde{w};q,p\} \rightarrow \{\tau,t;q,p\}. \tag{19}$$

The Laplace equation fixes the family of (undeformed) functions ϕ. Due to the dimensional reduction, the 4D Laplace operator *degenerates* to zero. This can be verified by simple inspection:

$$w=x^0-ix^1+x^2-ix^3, \quad \tilde{w}=x^0+ix^1-x^2-ix^3, \quad w^* \neq \tilde{w},$$
$$\partial_0=\partial_w+\partial_{\tilde{w}}, \quad \partial_1=-i\partial_w+i\partial_{\tilde{w}}, \quad \partial_2=\partial_w-\partial_{\tilde{w}}, \quad \partial_3=-i\partial_w-i\partial_{\tilde{w}}. \tag{20}$$

One can verify the Laplace operator acting on (the undeformed) ϕ:

$$\partial_0^2+\partial_1^2+\partial_2^2+\partial_3^2=\partial_y\partial_{\tilde{y}}+\partial_z\partial_{\tilde{z}}=\partial_w\partial_{\tilde{w}}-\partial_w\partial_{\tilde{w}}\equiv 0, \tag{21}$$

hence, as a result of the dimensional reduction, $\phi(w,\tilde{w})$, the 4D Laplace operator acting on (the undeformed) ϕ degenerates to zero, i.e., *any* function of the form $\phi=\phi(w,\tilde{w})$ *obeys* automatically the 4D Laplace equation.

If one wishes, one may restrict the solutions of the 4D Laplace equation, $\partial_\mu\partial^\mu\phi(w,\tilde{w})=0$, for arbitrary functions ϕ to those obeying the 2D Laplace equation instead:

$$\partial_\tau^2\phi+\partial_t^2\phi=\partial_w\partial_{\tilde{w}}\phi=0 \Rightarrow \phi=f(w)+g(\tilde{w}), \quad w\equiv\tau+it, \quad \tilde{w}=\tau-it. \tag{22}$$

In this fashion one can remove the arbitrariness of the (undeformed) ϕ.

It is important to emphasize that $\tilde{w}\neq w^*$ and that τ,t are *complex* valued since the Moyal heavenly equations are defined in *complexified* 4D space–time. A real slice may be taken by choosing $\tilde{w}=w^*$, which implies that τ,t must be real and from (20) we learn in this case that

$(x^0)^* = x^0$; $(x^1)^* = x^1$ and $(x^2)^* = -x^2$; $(x^3)^* = -x^3$. The general solution to the 2D (complexified in general) (undeformed) Laplace equation for the (undeformed) ϕ is $\phi = f(\tau + it) + g(\tau - it)$ for f,g arbitrary.

The SDYM equations require a separate study since in this case (the undeformed) $\phi = y\tilde{y} + z\tilde{z}$ is fixed and does not obey the Laplace equation. A different kind of reduction other than the Ivanova–Popov ansatz, via the Moyal–Nahm equations, is necessary to obtain the Moyal SDG equations from the 4D SU(∞) SDYM equations.

To finalize this section we point out that the SU(2) Toda lattice equations can be derived from the SU(2) Moyal–Nahm equations involving a field $\Psi = \Psi(\tau, t, \hbar)$ with $\tau = z_+ + z_-$. This has been shown in Ref. 14 and discussed in Ref. 18 in the constructions of a master integrable equation that contains the Kadomtsev–Petviashvili (KP) and Korteweg–de Vries (KdV) hierarchies. The explict relation between rotational Killing symmetry reductions of Ω and the SU(2) lattice Toda field, Ψ, will be studied further in Sec. III C. We proceed now to study the Moyal–Toda equations in the next section.

III. THE MOYAL–TODA EQUATIONS

A. A continuous Moyal–Toda equation and the Legendre transform

In this section we shall display the different forms of the Moyal–Toda equations that are related to the quantities $\mathcal{L}(u,\rho)$ and $\mathcal{H}(\kappa)$ in Eqs. (5), (6), and (18) and the $\mathcal{T}_\alpha[\phi, q, p, \hbar]$. We return now to this discussion. After the WWM map is performed, u, ρ, κ *acquire* an additional dependence on $q, p, \hbar$. To illustrate this, let us look at the operator form of the original continuous Toda equation:

$$\frac{\partial^2 \hat{\rho}}{\partial \tau^2} = \frac{\partial^2 e^{\hat{\rho}}}{\partial t^2}. \tag{23}$$

Given an operator, $\hat{\rho}(\tau, t)$, acting in the Hilbert space of square integrable functions on the line, of the form

$$\hat{\rho} = \sum_{mn} \rho_{mn}(\tau, t)(\hat{q}^m \hat{p}^n + \hat{p}^m \hat{q}^n + \cdots), \tag{24a}$$

with a Weyl ordering prescription imposed on the monomials in $\hat{q}^m \hat{p}^n$,

$$\hat{q}\hat{p} \rightarrow \hat{q}\hat{p} + \hat{p}\hat{q}, \quad \hat{q}\hat{p}^2 \rightarrow \hat{q}\hat{p}^2 + \hat{p}^2\hat{q} + 2\hat{p}\hat{q}\hat{p} \cdots. \tag{24b}$$

More complicated operators are also possible that are not necessary sums of monomials. The WWM map converting operators, $\hat{\rho}(\tau, t)$, into functions in phase space (making use of the symbol map) yields

$$\text{symbol } [\hat{\rho}] = \rho(t, \tau, q, p, \hbar), \quad \text{symbol } [e^{\hat{\rho}}] = e^{*\rho} = 1 + \rho + \frac{\rho * \rho}{2!} + \frac{\rho * \rho * \rho}{3!} + \cdots. \tag{25}$$

Hence, the putative Moyal continuous Toda molecule equations reads

$$\frac{\partial^2 \rho}{\partial \tau^2} = \frac{\partial^2 e^{*\rho}}{\partial t^2}. \tag{26}$$

The Moyal star product of two functions of phase space of dimension $2n$ whose symplectic form has the inverse ω^{IJ} is defined

$$f * g = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{i\hbar}{2} \right)^n \omega^{i_1 j_1} \omega^{i_2 j_2} \cdots \omega^{i_n j_n} (\partial_{i_1} \partial_{i_2} \cdots \partial_{i_n} f)(\partial_{j_1} \partial_{j_2} \cdots \partial_{j_n} g). \tag{27}$$

When ω^{ij} is the inverse of the symplectic form in two dimensions, the derivatives are taken wrt the q,p variables only. To recover the $2+1$ continuous Toda requires replacing the lhs of (26) by $(\partial_\tau)^2$ by $\partial_{z_+}\partial_{z_-}$ and setting $\rho(z_+,z_-,t,q,p,\hbar)$:

$$\frac{\partial^2\rho}{\partial z_+\partial z_-}=\frac{\partial^2 e^{*\rho}}{\partial t^2}. \tag{28}$$

The Prasad transformation[6,22] maps the 3D Laplace equation (obtained from the solutions associated with the translational Killing symmetry reductions of a self-dual Euclidean Einstein space[7]) to the continuous Toda equation:

$$(\partial_z\partial_{\overline{z}}+\partial_U^2)S=0\leftrightarrow\frac{\partial^2\rho}{\partial z_+\partial z_-}=-\frac{\partial^2 e^{\rho}}{\partial t^2}. \tag{29}$$

[Notice the *sign* difference in Eq. (29) as compared with Eq. (28).] Here S is now a function of $z=x+iy$, $\overline{z}=x-iy$, and U. Chapline[6] has performed the Legendre transform between S and ρ for those solutions of the continuous Toda equation related with the Eguchi–Hanson instanton:

$$\rho(z,\overline{z},t)=\ln\frac{t^2-a^2}{(1+z\overline{z})^2},\quad e^{\rho}=(t^2-a^2)e^{-2\ln(1+z\overline{z})},\quad t>a. \tag{30a}$$

The reason we exponentiated the solution (30a) will become clear next when we discuss the embedding of the Liouville equation into the continuous Toda one. The Liouville field will be, in this case, $\phi_L=-\ln(1+z\overline{z})$.

The Prasad change of variables[22] allows us to transform the Toda equation into the 3D Laplace equation as Eq. (29) shows. The function $S=S(\rho)$ corresponding to the Eguchi–Hanson instanton solution (30a) written in terms of the new coordinates (after the Prasad transform) is

$$S=\frac{1}{2}\sum_{i=1}^{2}\ln\frac{R_i+U-U_i}{R_i-U+U_i},\quad -\infty<U<\infty;\quad R_i^2=(U-U_i)^2+4(z-z_i)(\overline{z}-\overline{z_i}). \tag{30b}$$

Here x_i,y_i,U_i are the locations of the two-quasiparticles (two-center instanton of Gibbons-Hawking[21]) for $i=1,2$.

The most general solution to the ordinary classical continuous Toda equations has been found by Saveliev.[19] The Eguchi–Hanson instanton solution is precisely the one related to the well-known *embedding* of the Liouville equation into the continuous Toda equations.[19] It is given that

$$\frac{\partial^2\psi_0}{\partial\tau^2}=\frac{\partial^2 e^{\psi_0}}{\partial t^2}. \tag{31}$$

One may plug in the ansatz which will automatically reproduce the Liouville equation:

$$e^{\psi_0}=(\tfrac{1}{2}t^2+bt+c)e^{\phi_L}\Rightarrow\partial_{z_+}\partial_{z_-}\phi_L(z_+,z_-)=e^{\phi_L}, \tag{32a}$$

after performing the dimensional reduction $\tau=z_++z_-$. So, in the $\hbar=0$ limit, Eqs. (26) and (28) reduce to the ordinary classical continuous Toda molecule (chain) equations. Equation (31) is the the relevant one to map Moyal deformations of ρ into the 3D deformed Laplace equation (29). The embedding relation between ρ and ϕ_L in the Moyal deformed case is

$$\frac{\partial^2\rho}{\partial z_+\partial z_-}=\frac{\partial^2 e^{*\rho}}{\partial t^2},\quad e^{*\rho}=\left(\frac{1}{2}t^2+bt+c\right)e^{*\phi_L}\Rightarrow\partial_{z_+}\partial_{z_-}\phi_L(z_+,z_-,q,p,\hbar)=e^{*\phi_L} \tag{32b}$$

with $\phi_L(z_+,z_-,q,p,\hbar)$ and $\rho(z_+,z_-,t,q,p,\hbar)$ obeying Eq. (28). The last expression of Eq. (32b) is the Moyal deformation of the Liouville equation.

The *star* logarithm operation allows us to express

$$\rho=\ln_*\left[\left(\frac{1}{2}t^2+bt+c\right)e^{*\phi_L}\right]=\ln_*\left[\left(\frac{1}{2}t^2+bt+c\right)\right]+\ln_*[e^{*\phi_L}]$$
$$=\ln_*\left[\left(\frac{1}{2}t^2+bt+c\right)\right]+\phi_L\Rightarrow\frac{\partial^2\rho}{\partial z_+\partial z_-}=\partial_{z_+}\partial_{z_-}\phi_L. \qquad (32c)$$

Equations (32) are the relevant ones to extend the results of Eqs. (29) and (30) to the Moyal case when we deform the 3D Laplace equation $(\partial_i*\partial^i)S=0$ for a Moyal deformed function $S=(z,\overline{z},U,q,p,\hbar)$. The Moyal–Prasad map yields the continuous Moyal–Toda equation (28) whose particular Eguchi–Hanson instanton-inspired solutions will be deformations of Eq. (30a) and (30b). The natural expansions $\phi_L=\Sigma\hbar^n\phi_n(z_+,z_-,q,p)$ and $\rho=\Sigma\hbar^n\rho_n(z_+,z_-,t,q,p)$ will generate solutions to (32b) and (28) to all orders in $\hbar$ by an infinite number of iterations. To each order in $\hbar^n$ one performs the Moyal–Prasad map (change of coordinates[22]) yielding the values of $S=\Sigma\hbar^n S_n(z,\overline{z},U,q,p)$ which solve the deformed 3D Laplace equation (to that given order in $\hbar^n$). It is a very difficult and challenging enterprise to achieve this in practice; i.e, to find the Moyal–Prasad transform to *all* orders in $\hbar$ in *closed* form!

The above equations (26), (28), and (32) were obtained from the operator form of the original continuous Toda equation. In general, at the quantum level, the forms of the operator equations of motion are *not* the same as those of the original classical field. A modification of (26) will be presented shortly where the rhs is modified; i.e., there will be derivatives of infinite order wrt the t variable that originate from deformations of the continuum graded Lie algebras as a result of replacing ordinary Poisson bracket by Moyal brackets.

The $\hbar=0$ limit of (26) yields on the rhs the ordinary exponential, e^ρ, because in the classical limit the Moyal star product becomes the ordinary pointwise product of functions. Since in the classical limit Eq. (26) involves a differential equation wrt the variables τ,t only, the classical limit of (26) does not determine the q,p dependence of $\rho(\tau,t,q,p,\hbar=0)$, which may be completely arbitrary. Assuming that ρ admits an expansion in powers of $\hbar$: $\rho=\Sigma\hbar^n\rho_n(\tau,t,q,p)$, one can impose the condition that the zeroth-order term does not depend on q,p: $\rho_0(\tau,t,q,p)\equiv\rho_{\text{class}}(\tau,t)$. In this fashion the $\hbar=0$ limit reproduces the classical continuous Toda equation for $\rho_{\text{class}}(\tau,t)$.

Another way to obtain the continuous Moyal–Toda equations directly should be to perform the master Legendre transform mapping, if indeed it exists, between $\Omega(y',\tilde{y}',z',\tilde{z}',\hbar)$ obeying the rotationally Killing symmetry reductions of the Moyal SDG equations[9,21] to $\rho(\tau\equiv z_++z_-,t,q,p,\hbar)$[1,9] obeying the continuous Moyal–Toda equations. The immediate problem with this is that the *number* of variables does not match. ρ has five whereas Ω has four. Nevertheless, an embedding of one into the other is possible as we shall see. Strachan[14] has shown that the SU(2) Moyal–Nahm equations admit a reduction to the SU(2) Toda lattice (which yields the *classical* continuous Toda equation in the $\hbar=0$ limit). Therefore, reductions of the SU(2) generalized Moyal–Nahm equations should yield the continuous Moyal–Toda lattice-type equations. This shall be studied in Sec. IV.

We shall continue shortly with the Strachan ansatz and write down a more general equation than (26) and (28) which contains derivatives of infinite order wrt the t variable; i.e., the operator equations of motion for the quantized Toda field *differ* from the classical counterpart. The Legendre-like transform will be discussed also. The study of the geometry associated with these Moyal deformations has been given in Ref. 23.

B. Strachan's reduction of the SU(2) Moyal–Nahm equations to the SU(2) lattice Toda equations

It is known that the ordinary continuous Toda equation may be obtained from axial-symmetry reductions of the SU(∞) classical Nahm equations. This fact permitted Strachan to construct SU(2) Toda lattice equations from Moyal deformations of the SU(2) Nahm equations by replacing the

Poisson bracket by the Moyal bracket.[14] If one writes the Moyal–Nahm equations of the type given (with the plus sign) by Eq. (8), replacing $\phi \rightarrow \tau$ and imposing the axial-symmetry reductions of the form

$$T_1 = X_1 = h(\tau, t = q, \hbar)\cos p, \quad T_2 = X_2 = h(\tau, t = q, \hbar)\sin p, \quad T_3 = X_3 = z(\tau, t, \hbar), \tag{33}$$

allows the *decoupling* of the $\cos p, \sin p$ terms after computing the Moyal bracket and after eliminating the function $z(\tau, t)$. Strachan arrived at

$$\frac{\partial^2 \psi}{\partial \tau^2} = \frac{1}{4}\left[\frac{\Delta - \Delta^{-1}}{\hbar}\right]^2 e^{\psi} = \frac{\partial^2}{\partial t^2} e^{\psi} + \frac{\hbar^2}{3}\frac{\partial^4}{\partial t^4} e^{\psi} + O(\hbar^{2n}) \cdots . \tag{34}$$

The shift operators Δ, Δ^{-1} and $\psi(\tau, t, \hbar)$ are defined:

$$\Delta \psi = \psi(t + \hbar), \quad \Delta^{-1} \psi = \psi(t - \hbar), \quad e^{\psi/2} \equiv h(\tau, t). \tag{35}$$

In general, the $2+1$ Toda lattice equation for the field $\psi(z_+, z_-, t, \hbar)$ reads

$$\frac{\partial^2 \psi}{\partial z_+ \partial z_-} = \frac{1}{4}\left[\frac{\Delta - \Delta^{-1}}{\hbar}\right]^2 e^{\psi}, \tag{36}$$

upon the reduction $2+1$ to 2 dimensions, $\tau = z_+ + z_-$, one recovers (34).

Equations (34) and (36) will be referred to from now on as the $D=2$ and $D=2+1$ SU(2) lattice Toda equations, respectively. They involve one field only, ψ, in the same way that the Liouville equation is tantamount to an sl(2) Toda field equation. In the classical limit, $\hbar = 0$ one recovers the *classical continuous* SU(∞) Toda molecule equation as expected. This can be seen by expanding $\psi(\tau, t, \hbar) = \psi_0 + \hbar^2 \psi_2 + \cdots$. The $\hbar = 0$ limit reproduces again the classical continuous Toda equation for the field $\psi_0(\tau, t)$, as Eq. (26) does for $\rho(\tau, t, q, p, \hbar = 0) = \rho_{\text{class}}(\tau, t)$.

It is an interesting question (although not the right question to ask) if one could find a representation of the su(2) Lie algebra in terms of $\hat{q}, \hat{p}$ operators that would yield Strachan's solutions after performing the WWM map of the operator equations associated with the su(2) Nahm equations. The solutions (33) clearly differ from those presented in Eq. (10), not only in the τ functional dependence implicit in the elliptic functions but also in the dependence of the phase space variables.

The reason that one should *not* view Strachan's construction as a direct WWGM quantization of the SU(2) Toda field, a Liouville theory, is because the discrete-differential equations (34) and (36) represent really a SU(2)/SL(2) Toda lattice theory which discrete-spatial spacings in multiples of $\hbar$: the field ψ is evaluated at discrete jumps $t, t \pm \hbar,$. However, there is continuous temporal dynamics represented by the ∂_τ derivatives. See the important work of Dimakis *et al.*[24] on this respect. The rhs of (34) and (36) is a forward/backward discrete difference operator in jumps of $\hbar$ (although the variable t is continuous) which can be expanded into an infinite number of derivatives, a nonlocal expression. The Moyal product is also nonlocal due to the infinite number of derivatives.

Therefore, rigorously speaking one must refer to the discrete-differential Strachan's equation as the SU(2) lattice Toda equation obtained from axial-symmetry reductions of the SU(2) Moyal–Nahm equations and *not* as a direct WWM quantization of the SU(2) Liouville equation. It has been speculated by Reuter[23] that the discrete operator arising from the Moyal brackets might be one source of nonlocalities in quantum mechanics.

From the form of (10) one should notice that a WWGM quantization program of Ref. 11 is *not* the same as the Strachan construction[14] and that of Ref. 18. In general, one may alter the operator form of the quantum field equations of motion from the original classical field equations. For example, the operator equations may be different from Eqs. (23) and read instead

$$\frac{\partial^2 \hat{\rho}}{\partial \tau^2} = \frac{1}{4}\left[\frac{\Delta - \Delta^{-1}}{\hbar}\right]^2 e^{\hat{\rho}} = \frac{\partial^2}{\partial t^2} e^{\hat{\rho}} + O(\hbar^{2n}) \cdots . \tag{37}$$

We may have again a discrete-differential operator equations of motion: the rhs involves derivatives of *infinite* order of the operator $\hat{\rho}(\tau,t)$ wrt the t variable. In general, upon quantization the structure of the Lie algebra could itself be modified as well. This occurs in the study of quantum groups and quantum Lie algebras where quantum integrability requires deformations of the classical Lie algebraic structures.

In view of this, a modification of the Moyal continuous Toda equation (a further deformation of the Strachan's Toda lattice equations by introducing the star exponential) is

$$\frac{\partial^2 \rho}{\partial \tau^2} = \frac{1}{4}\left[\frac{\Delta - \Delta^{-1}}{\hbar}\right]^2 e^{*\rho} = \frac{\partial^2}{\partial t^2} e^{*\rho} + O(\hbar^{2n}) \ldots , \tag{38}$$

and is obtained after performing the WWM map of Eq. (37). The Moyal star product is taken wrt the q,p variables only. The extra dependence on the two phase space variables, q,p, is due to the WWM symbol map taking operators into functions in phase space. The t parameter is the one that encodes the continuum Lie algebra generators and commutation relations.

An even further deformation of (38) is as well

$$(\partial_\tau * \partial_\tau)\rho(\tau,t,q,p,\hbar) = \frac{1}{4}\left[\frac{\Delta - \Delta^{-1}}{\hbar}\right]^2 e^{*\rho} = \frac{\partial^2}{\partial t^2} e^{*\rho} + O(\hbar^{2n}) \ldots , \tag{39}$$

where the operators on the lhs of (39) are deformed as well, similarly to the deformations of Laplace equation (9a) using Moyal star products of differentials. Equation (38) is an extension of Strachan's SU(2) lattice Toda equation for the field $\psi(\tau,t,\hbar)$; in particular, the latter SU(2) lattice Toda equation can be *embedded* into Eq. (38). This shall be studied in the next subsection. Equation (39) is an extension, on both sides of the equation, of the original Moyal–Toda equation given by Eq. (26). Both the rhs and the lhs of (26) have been extended in Eq. (39).

Once again, reinserting the z_+, z_- dependence on the ρ we have finally the following discrete-differential equation:

$$(\partial_{z_+} * \partial_{z_-})\rho(z_+, z_-, t, q, p, \hbar) = \frac{1}{4}\left[\frac{\Delta - \Delta^{-1}}{\hbar}\right]^2 e^{*\rho}. \tag{40}$$

This is the more general equation of all the equations presented in this subsection containing all others as special limiting cases.

Concluding this subsection, we shall refer to the three discrete-differential Equations (38)-(40) as equations of the continuous Moyal–Toda lattice type. Equation (26) is the continuous Moyal–Toda equation obtained directly from a WWM quantization of the classical continuous Toda equation. Equation (32b) is the Moyal–Liouville equation and, finally, Eqs. (34) and (36) are the SU(2) Toda lattice equations in 2 and $2+1$ dimensions obtained from axial symmetry reductions of the SU(2) Moyal–Nahm equations *a la* Strachan.[16] It is very important to *distinguish* among these families of equations. The continuous Moyal–Toda lattice equation given by (40) contains all others as special cases. In this sense Eq. (40) is a sort of master discrete-differential equation.

We have learned from Strachan that the SU(2) Toda lattice equation is *contained* in the SU(2) Moyal–Nahm equations and, similarly, one would expect that the SU(∞) (continuous) Moyal–Toda lattice equations should be *contained* in the *generalized* Moyal–Nahm equations. It is precisely for this reason that the generalized Moyal–Nahm equations must depend on an *extra* set of phase space variables as argued earlier. This will be studied in Sec. IV where, in particular, we shall provide a plausible embedding of the Moyal–Toda lattice equations into the generalized Moyal–Nahm equations.

In conclusion, after these steps are taken, the Strachan ansatz is a very special case of the most general Ivanova–Popov ansatz for the three functions, $\mathcal{T}_\alpha$. Solutions to (34) and (36) may be obtained through iterations after expanding in even powers of $\hbar$: $\psi_o+\hbar^2\psi_2+\cdots$. In this way an infinite, but known, number of differential equations yields iteratively the solutions for ψ_{2n}. To solve this system is another matter.

C. The Moyal heavenly equation and the SU(2) Toda lattice

Before we begin this section we must emphasize that the results of Refs. 14 and 18 are very different from the results of this section, mainly because the results of Ref. 18 do *not* involve a rotational Killing symmetry reduction of the Moyal heavenly equations; i.e., this involves a particular class of dimensional reductions of the Moyal SDYM master equations that does *not* lead to the Moyal heavenly equations (although the Moyal heavenly equation can be obtained by another reduction). The authors of Ref. 18 based their work on the results by Strachan[14] on the Toda/KP hierarchies and obtained the SU(2) Toda lattice equation from a *different* route than the one described here. The equations in Refs. 14 and 18 are the SU(2) Toda lattice equations whose discrete spacing is a multiple of $\hbar$. When $\hbar=0$ one recovers the continuous Toda equations, or the so-called Boyer–Finley equations obtained originally from Killing symmetry reductions of the heavenly equations.

As discussed earlier in Sec. II, we can conclude that Eqs. (15) and (33) already *contain* the required map from the Ω heavenly form obeying the Moyal heavenly equation, after a suitable rotational Killing symmetry reduction from four dimensions to three, to the $2+1$ SU(2) Toda lattice field equations for a field ψ obeying Strachan equation (36). This can also be achieved via the Lax pair formalism. This automatically solves the problem of finding the Legendre-like transform from (rotational Killing symmetry reductions of) Ω to the $2+1$ SU(2) Toda lattice field $\psi(z_+,z_-,t;\hbar)$.

Hence, Eq. (15) yields $\Omega=\Omega[A_i]$. The Strachan solution corresponds to the particular case when ϕ is replaced by τ, which implies that A_i is replaced by X_i and, finally, the remaining Eqs. (33) give the explicit relations $A_i=X_i$ in terms of $\psi(\tau\equiv z_++z_-,t,\hbar),z(\tau\equiv z_++z_-,t,\hbar)$ with $\exp\psi/2\equiv h(\tau\equiv z_++z_-,t,\hbar)$.

To relate Ω with ρ is not that straightforward. There is an obstruction if one wishes to study the map from Ω to the continuous Toda field $\rho(z_+,z_-,t,q,p,\hbar)$. The *number* of variables does *not* match. However, one can ''embed'' both Ω and ψ into $\rho(z_+,z_-,t,q,p,\hbar)$; i.e., embedding the solutions to the SU(2) Toda lattice equations into the continuous Moyal–Toda lattice ones. The embedding process is based on the following facts:

It has been known for some time that rotational Killing symmetry reductions of the self-dual gravitational or heavenly equations lead to the continuous Toda equations.[7] The relation between rotational Killing symmetry reductions of $\Omega(y',\tilde{y}',z',\tilde{z}',\hbar)$ obeying the 4D Moyal SDG equations and $\rho(z_+,z_,,t,q,p,\hbar)$ obeying the $2+1$ continuous Moyal–Toda lattice equation should be defined as the embedding map:[10,25]

$$\begin{aligned}\Omega(y',\tilde{y}',z',\tilde{z}',\hbar)&\equiv\sum_{n=0}^{\infty}\left[\frac{\hbar}{\tilde{y}'}\right]^n\Omega_n(r\equiv y'\tilde{y}';z',\tilde{z}')\\&\to\rho(z_+,z_-,t,q,p,\hbar)\\&=\sum_{n=0}^{n=\infty}\sum_{l=0}^{l=n}\sum_{m=-l}^{m=+l}\hbar^n\rho_n^{lm}(z_+,z_-,t)Y_{lm}(\theta,\varphi),\\ q=\cot(\theta/2)\cos\varphi,&\quad p=\cot(\theta/2)\sin\varphi.\end{aligned}\tag{41}$$

The clear problem with Eq. (41) is that the *numbers* of variables do *not* match: ρ requires *five* whereas Ω requires *four*. For this reason one should refer to Eq. (41) as an embedding relation between Ω and ρ. Let us proceed with the terms appearing in the expansion of the ρ field in spherical harmonics. The limits in l are defined with the proviso that in the limit $\hbar=0$ the zeroth-order terms will survive only giving

$$\rho_0^{00}(\tau,t)Y_{00}(\theta,\varphi)\equiv\rho_{\text{class}}(\tau,t),\quad \lim_{\hbar=0} e^{*\rho}=e^{\rho_{\text{class}}(\tau,t)},\quad \lim_{\hbar=0}\rho=\rho_{\text{class}}(\tau,t), \tag{42}$$

and, as expected, the zeroth-order terms do *not* depend on q,p. We have also performed the stereographic projection mapping the sphere into the complex plane (phase space of q,p). This implies that the embedding relation, up to the nth order, map must establish the correspondence:

$$\left[\frac{1}{\tilde{y}'}\right]^n \Omega_n(r\equiv y'\tilde{y}';z',\tilde{z}')\rightarrow\sum_{l=0}^{l=n}\sum_{m=-l}^{m=+l}\rho_n^{lm}(z_+,z_-,t)Y_{lm}(\theta,\varphi). \tag{43}$$

It is essential not to confuse the prime variables $y',z'\ldots$ with the variables $y,z,\ldots$. A detailed discussion of the maps between the primed and unprimed variables was given in Ref. 10 based on Ref. 25:

$$\{\Omega_{,z'},\Omega_{,y'}\}_{\tilde{z}'\tilde{y}'}=1\leftrightarrow\{\Omega_{,w},\Omega_{,\tilde{w}}\}_{q,p}=1. \tag{44}$$

The inverse symbol map takes functions of phase space into operators in $L^2(R)$ and Moyal brackets into commutators leading to the operator equations

$$\frac{1}{i\hbar}[\hat{\Omega}_{,z'},\hat{\Omega}_{,y'}]=\hat{1}\leftrightarrow\frac{1}{i\hbar}[\hat{\Omega}_{,w},\hat{\Omega}_{,\tilde{w}}]=\hat{1}, \tag{45a}$$

where

$$\Omega(w,\tilde{w},q,p,\hbar)=\sum(\hbar)^n\Omega_n(w,\tilde{w},q,p). \tag{45b}$$

Extreme care must be taken *not* to set Ω_n as a function of $u\equiv w\tilde{w}$ and q,p. If this is wrongly assumed, then the rhs of (45a) will be *zero* instead of 1. This corrects an erratum in Ref. 10. An example of the coordinate transformation (a field-dependent coordinate transformation up to zeroth order) between the primed and unprimed variables associated with the particular solution $\Omega=\Omega_0=z'\tilde{z}'+y'\tilde{y}'$ is[9]

$$\tilde{z}'=q,\quad \tilde{y}'=p,\quad z'=w+\frac{\lambda}{q},\quad y'=\tilde{w}-\frac{\lambda}{p}, \tag{46}$$

with λ a complex constant. As n runs the coordinate transformation varies and one speaks of a field-dependent transformation to the nth order. Notice in (46) that $r=\tilde{y}'y'=p\tilde{w}-\lambda$ which is not equal to $w\tilde{w}$.

The discrepancy, which at first sight might be troublesome, is that the number of variables of Ω and ρ does not match. It is only after the q,p dependence is factorized in (41) that one can match the variables in $\Omega_n(r,z',\tilde{z}')$ with those in $\rho_n^{lm}(t,z_+,z_-)$. However, the $\psi_n(z_+,z_-,t)$ belonging to the $2+1$ SU(2) Toda lattice equation admits a perfect match with the number of variables $\Omega_n(r,z',\tilde{z}')$. For this reason, when one speaks of the Legendre-like transform from the Ω, obeying the rotational Killing symmetry reductions of the Moyal heavenly equation, to the Moyal–Toda lattice-type equations, one must refer to Strachan's SU(2) Toda lattice equation (36) as discussed at the beginning of this subsection.

The embedding of the SU(2) Toda lattice equations (34) and (36) into the continuous Moyal–Toda lattice-type equations (38)–(40) requires us to find a particular subset of solutions to the ρ_n^{lm} such that the following factorization condition occurs:

$$\sum_{l=0}^{l=n}\sum_{m=-l}^{m=+l}\rho_n^{lm}Y_{lm}=\psi_n f_n(q,p)=\rho_n(z_+,z_-,t,q,p) \tag{47}$$

(no sum over n), where the functions ψ_n are precisely those solving Strachan equations (34) and (36) after expanding ψ in powers of $\hbar$; i.e., therefore, one has embedded the SU(2) Toda lattice equation into the continuous Moyal–Toda lattice ones. Solving for ρ_n^{lm} in (47) yields

$$\rho_n^{lm}(z_+,z_-,t)\sim\psi_n(z_+,z_-,t)\int d\varphi\, d(\cos\theta)Y_{lm}(\varphi,\theta)f_n(q,p). \tag{48}$$

Here q,p were given in terms of φ,θ in (41). The $f_n(q,p)$ functions in principle can be determined from the differential equations obtained after inserting Eq. (47) into (38) and after replacing ∂_τ^2 by $\partial_{z_+}\partial_{z_-}$ on the lhs of (38).

Concluding this subsection, by sorting out the signature subtleties used by different authors,[3,8] the transformation of rotational Killing symmetry reduction of the 4D Moyal heavenly equations into the $2+1$ SU(2) Toda lattice field equations is attained in a two-step process by using Eqs. (15) and (33). In this fashion one establishes the sought-after relation between (reductions of) Ω and $(\psi(\tau,t,\hbar),z(\tau,t,\hbar))$ after the three YM potentials A_i involved in (15) are replaced by the Moyal–Nahm functions X_i defined in (33). This is probably one of the most relevant results of this work. The embedding of ψ obeying the SU(2) lattice equations (34) and (36) into the ρ obeying the continuous Moyal–Toda lattice equation (38) was also provided through the factorization conditions (47) and (48).

IV. THE GENERALIZED MOYAL–NAHM EQUATIONS, MOYAL DEFORMATIONS OF LOOP ALGEBRAS, SYMPLECTIC DIFFEOMORPHISMS IN 4D AND CONTINUUM LIE ALGEBRAS

A. Reductions of Moyal–Nahm equations

We have been studying *reductions* of the generalized Moyal–Nahm equations related to the Moyal deformations of the SU(∞) ASDYM/SDYM equations in four dimensions, an effective 6D theory. The most general Moyal–Nahm equations require, at least, an extra set of q',p' coordinates and hence one has an effective 8D theory where the Moyal bracket is taken wrt the enlarged phase space, i.e., the Weyl–Wigner–Moyal formalism involves mapping operator-valued quantities living in a 4D space–time [belonging to a Hilbert space of $L^2(R^2)$, instead of $L^2(R^1)$] into functions of the enlarged phase space, q,p,q',p'. Instead of dealing with Moyal deformed symplectic diffeomorphisms of a two-dimensional surface, one is now dealing with Moyal deformed symplectic diffeomorphisms in four dimensions. Therefore the effective theory is now 8D instead of 6D!

We *define* the generalized Moyal–Nahm equations as

$$\epsilon_{\alpha\beta\gamma}\frac{\partial\mathcal{T}_\gamma}{\partial\phi}=\{\{\mathcal{T}_\alpha,\mathcal{T}_\beta\}\},\quad \mathcal{T}_\alpha[\phi(x^\mu;q,p,q',p',\hbar);q,p,q',p';\hbar], \tag{49}$$

where now the Moyal bracket must be taken wrt an enlarged phase space q,p,q',p'. There is an explict and implicit $q_i,p_i,\hbar$ dependence in the Moyal–Nahm functions. The Moyal bracket of f,g wrt the enlarged phase space is compactly written as

$$\{\{f,g\}\}\equiv\frac{1}{\hbar}f(\sin[\hbar(\overleftarrow{\partial}_q\vec{\partial}_p-q\leftrightarrow p+\overleftarrow{\partial}_{q'}\vec{\partial}_{p'}-q'\leftrightarrow p')])g. \tag{50}$$

Expanding the sine function in powers of $\hbar$ one retrieves the infinite derivative terms. The WWM map takes self-adjoint operator-valued quantities, living in the Hilbert space $L^2(R^2),\hat{A}(x^\mu)$, into real-valued functions in phase space $\mathcal{A}(x^\mu;q,p,q',p';\hbar)$:

$$\mathcal{A}(x^\mu;q,p,q',p';\hbar)\equiv\int_{-\infty}^{\infty}d^2\xi\left\langle \vec{q}-\frac{\vec{\xi}}{2}\right|\hat{A}(x^\mu)\left|\vec{q}+\frac{\vec{\xi}}{2}\right\rangle \exp\left[\frac{i\vec{\xi}.\vec{p}}{\hbar}\right],\quad \vec{q}=(q,q'),\quad \vec{p}=(p,p'), \tag{51}$$

$$\left|\vec{q}+\frac{\vec{\xi}}{2}\right\rangle=\left|q_1+\frac{\xi_1}{2}\right\rangle\otimes\left|q_2+\frac{\xi_2}{2}\right\rangle\cdots. \tag{52}$$

Imagine representing now the SU(2) Lie-algebra-valued YM potentials (matrices) in terms of operators in $\hat{q},\hat{p},\hat{q}',\hat{p}'$ and performing afterwards the WWM map.

The relevant algebra is now the Moyal deformations of symplectic diffeomorphisms in four dimensions instead of two. For this reason it is incorrect to say that one has ''SU(∞)'' Moyal–Nahm equations. For example, the generators are labeled as $V_m^{l,\vec{k}}$, where $\vec{k}=(k_1,k_2)$,[20] and obey the infinite-dimensional generalization of the w_∞ algebra (area-preserving diffeomorphisms of the plane):

$$[V_m^{l,\vec{k}},V_n^{j,\vec{l}}]=[(j+1)m-(l+1)n]V_{m+n}^{l+j,\vec{k}+\vec{l}}+\vec{k}\times\vec{l}\,V_{m+n}^{l+j+1,\vec{k}+\vec{l}}. \tag{53}$$

This algebra of symplectic diffeomorphisms in four dimensions has a realization in terms of ordinary Poisson brackets wrt the q,p,q',p' enlarged phase space variables. The Moyal deformations are obtained by replacing ordinary Poisson brackets by Moyal ones. If one had a presentation of the SU(2) algebra as linear operators in $L^2(R^2)$, instead of the known representations in $L^2(R^1)$,[11] one could then evaluate the WWM map (51) and obtain solutions to the generalized Moyal–Nahm equations as it was done in Ref. 11.

Another type of generalized Moyal–Nahm equation one could write is such where instead of having a Moyal bracket wrt the enlarged phase space, one has a *partial* Moyal bracket wrt one set of q,p variables:

$$\epsilon_{\alpha\beta\gamma}\frac{\partial\mathcal{T}_\gamma}{\partial\phi}=\{\mathcal{T}_\alpha,\mathcal{T}_\beta\}_{q,p},\quad \mathcal{T}_\alpha[\phi(x^\mu;q,p,q',p',\hbar);q,p,q',p';\hbar]. \tag{54}$$

The *problem* with Eq. (54) as such is that it does *not* determine the functional dependence on the q',p' variables since there is no differential operators which involve now the q',p' variables. For this reason we should *disregard* (54) as a valid equation.

We shall study the reductions of (49) with the goal of obtaining the continuous Moyal–Toda lattice-type equations (38)–(40). First, one replaces ϕ by τ and, as usual, $A_0=0$, $A_i\sim\mathcal{T}_i$. Second, one imposes the reduction condition $t=q'$ while recurring to an ansatz that allows one to decouple the $\cos p'$, $\sin p'$ terms, after computing the Moyal bracket in (49) and (50), giving, finally, an equation involving *only* the τ,t,q,p variables. If we set

$$\mathcal{T}_1=R(\tau,t=q',q,p,\hbar)\cos p',\quad \mathcal{T}_2=R(\tau,t=q',q,p,\hbar)\sin p',\quad \mathcal{T}_3=z(\tau,t=q',\hbar). \tag{55}$$

After plugging (55) into (49), the terms in $\cos p'$, $\sin p'$ decouple and, eliminating z, one obtains the following highly nontrivial equation for the function R, after computing the Moyal bracket wrt the enlarged phase space variables:

$$\frac{\partial^2\ln R}{\partial\tau^2}=\alpha\left(\frac{\Delta-\Delta^{-1}}{\hbar}\right)^2R^2+\beta\left(\frac{\Delta-\Delta^{-1}}{\hbar}\right)\sum_{n=1}^{\infty}(\hbar)^nC_{i_1\ldots i_n}(\partial_{i_1}\cdots\partial_{i_{n-1}}R)(\partial_{j_1}\cdots\partial_{j_n}R). \tag{56}$$

The second terms on the rhs of (56) contain mixed derivatives of infinite order wrt the t,q,p variables. Here α, β and $C_{i_1\ldots i_n}$ are constants. The Strachan SU(2) Toda lattice equation is recovered automatically by dropping the extra q,p dependence on R [so that the second term on the rhs of (56) becomes zero] and by equating $R^2=e^\rho$.

Due to the extra q,p dependence the situation changes drastically. In order to obtain an equation like (38) it is required to establish the *new* functional relation between the R and ρ functions. The previous relation $R^2=e^\rho$ does not work, for example, if one sets

$$R*R=e^{*\rho}\Rightarrow\frac{\rho}{2}\neq\ln R. \tag{57}$$

The lhs of Eq. (56) will no longer be $(1/2)\partial_\tau^2\rho$. Furthermore, if (57) was the correct relation between R and ρ, the rhs of (56) would *not* equal the rhs of (38) despite the fact that the rhs of (56) contains the same type of infinite derivative terms as Eq. (38) does. This is because [after using the relationship (57)] the *coefficients* on the rhs of Eqs. (56) and (38) *differ*. This could be fixed easily by adding terms on *both* sides of Eq. (56) with the appropriate coefficients so that the rhs of (56) equals precisely the corresponding term of Eq. (38):

$$\left(\frac{\Delta-\Delta^{-1}}{\hbar}\right)^2 R*R=\left(\frac{\Delta-\Delta^{-1}}{\hbar}\right)^2 e^{*\rho}.$$

However, there is still no assurance that the lhs of (56), after the addition of those terms which rendered the rhs of (56) in the required form given by Eq. (38), will have *also* the sought-after form,

$$\frac{\partial^2\ln_* R}{\partial\tau^2}=\frac{1}{2}\,\frac{\partial^2\rho}{\partial\tau^2}, \tag{58a}$$

where the *star* logarithm and *star* square root are defined as

$$\ln_*(e^{*\rho})\equiv\rho,\quad R*R=e^{*\rho}\Rightarrow R=[e^{*\rho}]_*^{1/2},\quad \ln_* R=\ln_*[e^{*\rho}]_*^{1/2}=\tfrac{1}{2}\rho. \tag{58b}$$

In other words, it is not clear that one can adjust *both* sides of Eq. (56), by adding terms on both sides simultaneously, to match the required form of the continuous Moyal–Toda lattice equation (38). It would have been very fortuitous that the simultaneous addition of terms on both sides of the equation (56) would lead exactly to the continuous Moyal–Toda lattice equation (38).

Therefore, by inspection, the ansatz proposed in (55) does not seem to work if one wishes to find a direct relation between R and ρ. This could be remedied by introducing two auxiliary functions as we shall see shortly. Although it is *undesirable* to introduce extraneous functions into this construction, the use of two auxiliary functions seems to work in principle. The source of the problem is to find to the new functional relationship between R and ρ that will render both sides of Eq. (56) correctly. Nevertheless, matters are not final. Equation (56) is *per se* satisfactory in the sense that it yields a well-defined differential equation for the R function with an infinite number of derivative terms. Imposing the condition

$$\frac{1}{2}\,\frac{\partial^2\rho}{\partial\tau^2}=\frac{\partial^2\ln R}{\partial\tau^2}\Rightarrow\frac{1}{2}\rho=\ln R+F(t,q,p,\hbar)\tau+G(t,q,p,\hbar) \tag{59}$$

fixes ρ in terms of R and two auxiliary functions, F,G. Solving for R yields

$$R=e^{\rho/2-F\tau-G}=e^{\rho/2}e^{-F(t,q,p,\hbar)\tau-G(t,q,p,\hbar)}. \tag{60}$$

Inserting this new value for R into the rhs of (56) and equating it to the rhs of (38) determines another differential equation for the F,G functions in conjunction with the original Moyal–Toda lattice equation. A coupled set of differential equations for ρ and F,G is obtained in this fashion: Eqs. (38), (56), and (59), three equations for three functions.

The main question now will be to see whether or not this system of coupled differential equations is self-consistent and has well-defined nontrivial and nonsingular solutions, in particular, whether, or not Eq. (59) is in fact *compatible* with the remaining two equations (38) and (56). This very difficult question remains to be answered.

B. Embeddings of the Moyal–Toda equations

It is also reasonable to look for other ways of embedding the continuous Moyal–Toda lattice equations into the Moyal–Nahm equations that do *not* require a direct decoupling of the $\cos p'$, $\sin p'$ functions, for example, integrating out the p' variable without the need to reduce the generalized Moyal–Nahm equations.

The use of the $\mathcal{T}_\alpha$ was successful earlier in obtaining the SU(2) Toda lattice equation from the Moyal–Nahm equations after one imposing the reduction $q=q'$, $p=p'$ conditions and replacing ϕ by τ and setting $t=q$. It is reasonable to ask if other embeddings/reductions are possible. The continuous Moyal–Toda equation (26) admits the following embedding into the Moyal–Nahm equations for those *special* solutions of ρ satisfying the *reduction* condition: $\rho(\tau,t,q,p,\hbar)=\rho(\tau \pm it,q,p,\hbar)$, where in addition one restricts the solutions of the deformed Laplace equation for the scalar ϕ to have the trivial values: $\tau\pm it$. For this very restricted case one has the following relations:

$$\frac{\partial \mathcal{T}_3}{\partial \phi}=\frac{\partial^2\rho}{\partial\phi^2}\Rightarrow\rho(\phi,q,p,\hbar)=\int \mathcal{T}_3\, d\phi+F(q,p,\hbar)\phi+G(q,p,\hbar), \tag{61a}$$

$$\{\mathcal{T}_+,\mathcal{T}_-\}=-\frac{\partial^2}{\partial\phi^2}e^{*\rho(\phi,q,p,\hbar)},\quad \rho=\rho(\phi,q,p,\hbar),\quad q=q',p=p', \tag{61b}$$

$$\frac{\partial^2\rho}{\partial\phi^2}=-\frac{\partial^2}{\partial\phi^2}e^{*\rho(\phi,q,p,\hbar)}. \tag{61c}$$

Following similar arguments as above yields a system of differential equations to solve for the three functions $\rho(\phi,q,p,\hbar),F,G$. One can infer directly from Eq. (61c)

$$\rho+e^{*\rho}=A(q,p,\hbar)\phi+B(q,p,\hbar). \tag{62}$$

In this simpler case we have an implicit equation for ρ. If one were able to invert this last equation, one will have then an expression of the form $\rho=\rho(\phi,q,p,\hbar)$ that would have been a particular solution to Eq. (61c) related to the embedding displayed by Eqs. (61a) and (61b). It is not an easy task to invert the latter equation. Furthermore, the full-fledged theory requires us to use more general solutions for the deformed scalar ϕ obeying the deformations of Laplace equation (9b). The choice $\tau\pm it$ was the trivial one.

Finally, a natural embedding of the continuous Moyal–Toda lattice-type equations (38)–(40) into the generalized Moyal–Nahm equations (49) is to set $q'=t$ and to replace as usual the function ϕ by τ so that the ansatz in (7) yields $A_0=0$ and $A_i(\tau,q'=t,p',q,p,\hbar)$ is replaced by $\mathcal{T}_i$ for $i=1,2,3$. Given a solution to the generalized SU(2) Moyal–Nahm equations (using Moyal brackets wrt the extended phase space variables) for the three potentials $\mathcal{A}_i$, one selects the particular equation

$$\frac{\partial\mathcal{A}_3}{\partial\tau}=\{\{\mathcal{A}_+,\mathcal{A}_-\}\},\quad \mathcal{A}_\pm=\frac{1}{\sqrt{2}}(\mathcal{A}_1\pm i\mathcal{A}_2). \tag{63}$$

A partial integration taken wrt the p' variable only yields the embedding relations

$$\int_{-\infty}^{+\infty}dp'\frac{\partial\mathcal{A}_3}{\partial\tau}=\frac{\partial^2\rho(\tau,t,q,p,\hbar)}{\partial\tau^2}, \tag{64}$$

$$\int_{-\infty}^{+\infty} dp' \{\{\mathcal{A}_+,\mathcal{A}_-\}\} = \frac{1}{4}\left(\frac{\Delta-\Delta^{-1}}{\hbar}\right)^2 e*^{\rho} = \partial_t^2 e*^{\rho(\tau,t,q,p,\hbar)} + \frac{1}{3}(\hbar)^2\partial_t^4 e*^{\rho} + \cdots. \tag{65}$$

There are some total derivative terms wrt p' which vanish after integration, but not all of the terms appearing on the lhs of (65) are total derivatives. From Eq. (64) one learns that

$$\rho = \int_{-\infty}^{+\infty} dp' \int_{\tau'=0}^{\tau'=\tau} \mathcal{A}_3\, d\tau' + \mathcal{F}(t,q,p,\hbar)\tau + \mathcal{G}(t,q,p,\hbar). \tag{66}$$

The functions $\mathcal{F}$, $\mathcal{G}$ are not arbitrary, but are part of the system of three coupled differential equations given by Eqs. (63)–(65). The embedding is characterized by integrating out the variable p' without imposing a reduction on the Moyal–Nahm equations (49) to decouple the p' variable directly. In all these embeddings, two auxiliary functions are required.

It is warranted to see whether or not is possible to find a Killing symmetry reduction of the generalized Moyal–Nahm equations (49) directly to the continuous Moyal–Toda lattice-type equations (38)–(40) *without* the need to recur to auxiliary functions and avoid the complicated set of coupled differential equations. The essence of the problem lies in the fact that the t variable plays two different roles. In one case, like in the rotational Killing symmetry reduction of the heavenly equation, it behaves like an ordinary space–time variable and, in another, it behaves like an *internal* phase space variable associated with the Lie algebra of the area-preserving diffeomorphisms of the sphere. It seems very difficult to reconcile both roles within the WWGM formalism without returning to the coupled system of differential equations. We believe that there maybe a reduction (other than the axial symmetry reduction proposed above) of the generalized Moyal–Nahm equations that successfully decouples the p' variable and that reproduces the continuous Moyal–Toda lattice equations *without* the introduction of auxiliary functions.

C. Continuum Lie algebras and the Lax–Brockett formalism

To finalize matters we recall again the Lax–Brockett formalism of **Z**-graded continuum Lie algebras[19] which is perhaps the most geometrical of all approaches.

The data $\{\mathcal{T}_\alpha\}$ for three Moyal–Nahm functions need to be related to the three functions $\{\mathcal{L},\mathcal{H},\mathcal{M}\}$ required in the Lax–Brockett formalism of continuum **Z**-graded Lie algebras.[19] Equation (18) establishes the correspondence among these data: a functional embedding of the data $\{\mathcal{T}_\alpha\}$ (parametrized by one function, ϕ) into the data $\{\mathcal{L},\mathcal{H},\mathcal{M}\}$ (parametrized by three functions, u,ρ,κ). However, this is only possible if one can construct the three twistorlike transformations which map the now-deformed scalar field $\phi(x^\mu;q_i,p_i,\hbar)$ into the now-deformed three functions appearing in Eqs. (2)–(4): ρ,u,κ depending on $\tau,t;q_i,p_i,\hbar$. One of them is the Moyal deformed continuous Toda field: ρ.

Presumably this should be related to the problem of deformations of twistor surfaces and Kodaira–Spencer deformation theory.[14] Unfortunately, we cannot say more on this matter. A start will be in constructing Moyal deformations of continuum Lie algebras. We are unaware if this has ever been done. A construction will provide the functional relations of $\{\mathcal{L},\mathcal{H},\mathcal{M}\}$ in terms of u,ρ,κ. The Lax–Brockett formalism of the Moyal–Nahm equations should yield the continuous Moyal–Toda equation for ρ (26) after the elimination of u,κ as it occurs in the undeformed case.[19]

For this reason, it is of tantamount importance to construct representations of continuum Lie algebras and for that matter representations of SU(∞) as well. A WWM quantization will then provide the functional relations of $\{\mathcal{L},\mathcal{H},\mathcal{M}\}$ in terms of u,ρ,κ.

The main solution to the problem of writing down Moyal–Toda equations from reductions of Moyal–Nahm equations requires the construction of Moyal deformations of the graded continuum Lie algebras. There are other Moyal deformations related to the infinite-dimensional loop algebras associated with w_∞ algebras. For example, the loop algebra of $sdiff(R^2)$, the algebra of maps of the circle into w_∞, in the basis of functions $x^{s+m}y^{s-m}$ is

$$[v_m^s(\sigma),v_n^t(\sigma')]=[(t-n)(s+m)-(s-m)(t+n)]v_{m+n}^{s+t-1}(\sigma)\frac{\partial}{\partial\sigma}\delta(\sigma-\sigma'). \tag{67}$$

These loop algebras may admit Moyal deformations as well since the Moyal deformation of the centerless w_∞ algebra *is* the centerless W_∞ algebra.[26–28] Central extensions can be added as well. Hence, the Moyal deformations of the algebra (67) will be just the infinite-dimensional loop algebra associated with W_∞.

Moyal deformations of the **Z**-graded continuum Lie algebras[19] ought to be very relevant in the Moyal quantization program of the continuous Toda theory. Especially in regards to determining the differential equations for the $u(\tau,t,q,p,\hbar)$ and $\kappa(\tau,t,q,p,\hbar)$ appearing directly in the Lax–Brockett double commutator formalism. More on this shall be said in a forthcoming publication.

V. CONCLUSION

We have explicitly presented in Sec. II a class of solutions to the Moyal SU(∞) ASDYM equations in four dimensions that are related to the *reductions* of the generalized Moyal–Nahm equations via the Ivanova–Popov ansatz. A dimensional reduction yields solutions to the Moyal deformations of the ASDG equations. The SDYM and SDG case requires a separate study.

Since the ASDYM equations studied by Ivanova and Popov[4] in Euclidean 4D correspond to the SDYM equations in $2+2$ dimensions studied in Ref. 9, one can write down the transformation that maps the rotational Killing symmetry reductions of the 4D Moyal heavenly equations given by Eq. (14) into the $2+1$ SU(2) Toda lattice equations given by Eq. (36). A two-step process is required to attain such a map and it is explicitly given by Eqs. (15) and (33). This is one of the most relevant results of this work.

Three *different* types of Toda equations have been studied. (i) The continuous Moyal–Toda equation (26) and the Moyal–Liouville equation (32b). Two types of discrete-differential equations, (ii) the SU(2) Toda lattice (34) and (36) and (iii) the continuous Moyal–Toda lattice equations, Eqs. (38)–(40). The generalization of the Prasad transformation[6,22] to the Moyal case mapping deformations of the 3D Laplace equation to the continuous Moyal–Toda equation (26) was discussed in Sec. III A.

Finally, the generalized Moyal–Nahm equations (49) have been provided that should *contain* the continuous Moyal–Toda lattice-type equations after a suitable reduction, similar to the one performed by Strachan[14] which yields the SU(2) Toda lattice from the SU(2) Moyal–Nahm equations. Unfortunately, this reduction requires the introduction of two auxiliary fields. Further details of the reduction *without* auxiliary fields is currently under investigation. Embeddings of the various forms of the Moyal–Toda equations into the Moyal–Nahm equation were also provided and, again, the introduction of two auxiliary fields was required.

The project for the future is to study the Lax–Brockett formalism of the generalized Moyal–Nahm equations. Unfortunately we lack an explicit knowledge of the form of the $\mathcal{L}$, $\mathcal{M}$ functions in terms of the deformed ρ,u,κ fields. If we did, then one could apply the Lax–Brockett formalism (18) to the Moyal–Nahm equations to obtain directly the Moyal–Toda equations for ρ upon eliminating the u,κ fields. Presumably, this could be a realization of deformations of twistor surfaces.[15] The connection to Kodaira–Spencer deformation theory[14] is unknown at the moment.

Other Moyal deformations applied to higher extended objects, p-branes, remain to be studied: the so-called Moyal–Nambu–Poisson algebras related to deformations of the volume forms. The natural deformation quantization technique is the Zariski product[16] which generalizes the Moyal product to p-branes. Octonionic[29] and quaternionic Moyal–Nahm equations can be constructed as well using the octonionic/quaternionic structure constants instead of the $\epsilon_{\alpha\beta\gamma}$ tensor density. The fact that the generalized Moyal–Nahm equations require eight dimensions may have an important role in understanding the quantum dynamics of the 11D membrane[13] and the role of W_∞ algebras.[25]

ACKNOWLEDGMENTS

We are indebted to M. Przanowski for valuable discussions, to T. Ivanova for useful remarks, and to J. Boedo and L. Russo for their help and hospitality at the University of California, San Diego. This work was supported in part by the CONACyT, México.

[1] C. Castro, Phys. Lett. B **288**, 291 (1992); Chaos Solitons Fractals **7**, 711 (1996); ''The Noncritical W_∞ String Sector of the Membrane,'' hep-th/9612160.
[2] J. Hoppe, ''Quantum Theory of a Relativistic Surface,'' Ph.D thesis, M.I.T. (1982).
[3] E. Bergshoeff, E. Sezgin, Y. Tanni, and P. K. Townsend, Ann. Phys. **199**, 340 (1990); B. de Wit, J. Hoppe, and H. Nicolai, Nucl. Phys. B **305**, 545 (1988); B. Biran, E. G. Floratos, and G. K. Saviddy, Phys. Lett. B **198**, 32 (1987); T. Banks, W. Fischler, S. Shenker, and L. Susskind, Phys. Rev. D **55**, 5112 (1997).
[4] T. A. Ivanova and A. D. Popov, J. Math. Phys. **34**, 674 (1993); Theor. Math. Phys. **94**, 225 (1993); Lett. Math. Phys. **23**, 29 (1991); JETP Lett. **54**, 69 (1991).
[5] E. G. Floratos and G. K. Leontaris, Phys. Lett. B **223**, 153 (1989).
[6] G. Chapline and K. Yamagishi, Class. Quantum Grav. **8**, 427 (1991); Phys. Rev. Lett. **66**, 3064 (1991); G. Chapline, Mod. Phys. Lett. A **7**, 1959 (1992).
[7] C. D. Boyer and J. Finley III, J. Math. Phys. **23**, 1126 (1982); Q. H. Park, Int. J. Mod. Phys. A **7**, 1415 (1991).
[8] C. Castro, J. Math. Phys. **34**, 681 (1993); **35**, 3013 (1994).
[9] J. Plebański and M. Przanowski, Phys. Lett. A **219**, 249 (1996).
[10] C. Castro, Phys. Lett. B **413**, 53 (1997).
[11] H. Garcia-Compeán and J. Plebański, Phys. Lett. A **234**, 5 (1997).
[12] H. Weyl, Z. Phys. **46**, 1 (1927); E. Wigner, Phys. Rev. **40**, 749 (1932); J. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1945); H. J Groenwold, Physica (Amsterdam) **12**, 405 (1946).
[13] D. Fairlie, Mod. Phys. Lett. A **13**, 263 (1998); D. Fairlie, C. Zachos, and T. Curtright, ''Matrix Membranes and Integrability,'' hep-th/9709042.
[14] I. A. B. Strachan, Phys. Lett. B **282**, 63 (1992); J. Geom. Phys. **21**, 255 (1997); J. Phys. A **29**, 6117 (1996); private communication.
[15] K. Takasaki and P. Guha, ''Dispersionless Hierarchies, Hamilton Jacobi Theory and Twistor Correspondences,'' solv-int/9705013.
[16] G. Dito, M. Flato, D. Sternheimer, and L. Takhtajan, ''Deformation Quantization and Nambu Mechanics,'' hep-th/9602016.
[17] C. Castro, in preparation.
[18] M. Przanowski and S. Formański, ''Searching for a universal integrable system,'' University of Lodz preprint, May 1997, to be published in Acta Phys. Pol. B.
[19] M. Saveliev, Theor. Math. Phys. **92**, 457 (1992); M. Saveliev and A. Vershik, Phys. Lett. A **143**, 121 (1990).
[20] I. Bakas and E. Kiritsis, in ''*Common Trends in Mathematics and Quantum Field Theories*,'' edited by T. Eguchi, T. Inami, and T. Miwa (Prog. Theor. Phys. Supplement, 1990).
[21] G. W Gibbons and S. W Hawking, Phys. Lett. B **78**, 430 (1978); T. Eguchi and A. J. Hanson, Phys. Lett. B **74**, 249 (1978).
[22] M. K Prasad, Phys. Lett. B **83**, 310 (1979).
[23] B. Fedosov, J. Diff. Geom. **40**, 213 (1994); M. Reuter, ''Non-Commutative Geometry of Quantum Phase Space,'' hep-th/9510011; H. Garcia-Compean, J. Plebański, and M. Przanowski, Acta Phys. Pol. B **29**, 549 (1998).
[24] A. Dimakis and F. Muller-Hoissen, ''Integrable discretizations of chiral models via deformation of the differential calculus,'' hep-th/9512007; ''Deformations of classical geometries and integrable systems,'' physics/9712002.
[25] C. Castro, Phys. Lett. B **353**, 201 (1995).
[26] D. B. Fairlie and J. Nuyts, Commun. Math. Phys. **134**, 413 (1990).
[27] I. Bakas, B. Khesin, and E. Kiritsis, Commun. Math. Phys. **151**, 233 (1993).
[28] C. Pope, L. Romans, and X. Shen, Phys. Lett. B **236**, 173 (1990).
[29] D. Fairlie and T. Ueno, ''Higher dimensional generalizations of the Euler Top Equations,'' hep-th/9710079; T. Ueno, ''General solution of the 7D octonionic top equations,'' hep-th/9801079.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Central charge and the Andrews–Bailey construction

Leung Chim
Department of Physics and Mathematical Physics, The University of Adelaide, Adelaide, SA, 5005, Australia

(Received 11 November 1998; accepted for publication 9 April 1999)

From the equivalence of the bosonic and fermionic representations of finitized characters in conformal field theory, one can extract mathematical objects known as Bailey pairs. Recently Berkovich, McCoy, and Schilling have constructed a ''generalized'' character formula depending on two parameters ρ_1 and ρ_2, using the Bailey pairs of the unitary model $M(p-1,p)$. By taking appropriate limits of these parameters, they were able to obtain the characters of model $M(p,p+1)$, $N=1$ model $SM(p,p+2)$, and the unitary $N=2$ model with central charge $c=3(1-2/p)$. In this letter we computed the effective central charge associated with this ''generalized'' character formula using a saddle point method. The result is a simple expression in dilogarithms which interpolates between the central charges of these unitary models. © *1999 American Institute of Physics.*
[S0022-2488(99)02307-5]

I. INTRODUCTION

More than a decade since its creation, two-dimensional conformal field theory (CFT)[1] and its integrable perturbations[2] still remain as one of the most active research topics in modern physics. A current focus is in the study of various bases of the Hilbert space in CFT. Different choices of the basis would lead to a different representation for the partition function of the CFT defined on a compact manifold such as a torus or a cylinder. This partition function is usually written in terms of characters of the Virasoro or some extended algebras. The ''bosonic'' form of these character formulas are well known for the minimal models of CFT (e.g., see Refs. 3–8, and references therein). Recently the Stony Brook group has constructed numerous new character formulas for minimal CFT based on fermionic quasiparticles.[9–12] For several CFTs, more than one ''fermionic'' expression exists for the same conformal character. In these cases, the different expressions are related to the different integrable perturbations of the same CFT. These developments all lend support to the idea of a massless scattering S-matrix description of CFT.[13–16] The construction of the quasiparticle basis of the Hilbert space is also apparently related to the problem of diagonalizing the infinite set of local integrals of motion in CFT.[17,18] For a description in terms of other bases see Refs. 19–21.

The equivalence of the bosonic and fermionic character formulas gives rise to beautiful q-series identities of the Rogers–Ramanujan type.[22–24] Such identities were proven by Schur[25] and Andrews[26] by establishing identities for families of polynomials whose limiting case yields the Rogers–Ramanujan identities. This method was applied to the character identities of the unitary minimal models by Melzer in Ref. 27, where he proposed a finitization of fermionic character formulas to match the bosonic polynomials of Ref. 28. In Ref. 29, several classes of q-series identities were proven using Andrews' generalization[30] of Bailey's lemma.[31] The key observations in Ref. 29 were that Bailey pairs can be extracted from finitized characters (which must be in the general bosonic form of Ref. 32)[33] and several series of CFT [including subsets of $M(p,p+1)$, $M(p,p+2)$, and $M(p,kp+1)$] are ''linked'' on a so-called Bailey's chain.[30] The equivalence proof for all members of a series is a straightforward application of Bailey's lemma, once a proof is established for a single member.[30]

In a remarkable paper,[34] the procedure of Ref. 29 was repeated using a more general form of the Andrews–Bailey construction which contains two parameters ρ_1 and ρ_2.[30,31] From the (dual)

Bailey pairs for the unitary minimal model $M(p-1,p)$, the more general construction gives a ''generalized'' character formula [Eq. (4.19) of Ref. 34] depending on the two parameters. Three specializations of these parameters lead to known results:

$$\begin{aligned}&(\mathrm{i}): \rho_1 \to \infty, \quad \rho_2 \to \infty;\\&(\mathrm{ii}): \rho_1 \to \infty, \quad \rho_2 = \text{finite};\\&(\mathrm{iii}): \rho_1 = \text{finite}, \quad \rho_2 = \text{finite}.\end{aligned} \tag{1.1}$$

In the first case, the ''generalized'' character formula becomes the character for the next model in the unitary series, i.e., $M(p,p+1)$ with central charge $1-6/p(p+1)$. Case (ii) leads to fermionic character formula for $N=1$ supersymmetric model $SM(p,p+2)$ with central charge $\frac{3}{2}-12/p(p+1)$, while case (iii) gives the fermionic character of unitary $N=2$ model with central charge $c=3(1-2/p)$. It is amazing that this ''generalized'' character formula connects the unitary models with their supersymmetric counterparts. In fact, this construction (and its generalizations) can also be applied to other minimal models $M(p,p')$.[29,11,35–37] A natural question (raised in Ref. 34) which needs addressing is whether this construction has any connection to massless renormalization group flows between these CFT.[13,38]

In this paper, we shall attempt to understand this ''generalized'' character formula for the unitary series by computing the associated ''generalized'' effective central charge. In Sec. II we give a brief review of the construction detailed in Ref. 34 and establish our notations. The generalized effective central charge is calculated in Sec. III via a saddle point approximation following Refs. 39, 40, 9, and special cases are treated. A discussion of our result is given in Sec. IV.

II. THE ANDREWS–BAILEY CONSTRUCTION OF THE UNITARY MODELS

The unitary CFT $M(p-1,p)$ is the continuum limit of the $(p-1)$-states RSOS lattice model[28] at its critical point between regimes III and IV.[41] The equivalence of the associated bosonic and fermionic finitized characters can be written as[27]

$$B_{r,s}^{(L,p)} = F_{r,s}^{(L,p)}. \tag{2.1}$$

The bosonic side has the form

$$B_{r,s}^{(L,p)}(q) = \sum_{j=-\infty}^{\infty} \left(q^{j(jp(p-1)+pr-(p-1)s)} \begin{bmatrix} L \\ [\frac{1}{2}(L+s-r)]-pj \end{bmatrix}_q - q^{(jp-s)(j(p-1)-r)} \times \begin{bmatrix} L \\ [\frac{1}{2}(L-s-r)]+pj \end{bmatrix}_q \right), \tag{2.2}$$

where $[n]$ denotes the integer part of n, and

$$\begin{bmatrix} n \\ m \end{bmatrix}_q = \begin{cases} \dfrac{(q)_n}{(q)_m (q)_{n-m}} & \text{for } 0 \leqslant m \leqslant n \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

is the usual q-binomial coefficient with

$$(a)_n = \frac{(a)_\infty}{(aq^n)_\infty}, \quad (a)_\infty = \prod_{l=0}^{\infty} (1-aq^l). \tag{2.4}$$

One should note that this finitized character is equal to the one-dimensional configuration sum of the underlying RSOS model in regime III, defined on a square lattice of size L,[28] and can be used to compute local height probabilities. In the limit $L\to\infty$, (2.2) becomes the character formula[3] for the irreducible representation generated by the primary field $\Phi_{(r,s)}$, with normalization $\chi_{r,s}(q)=1+\Sigma_{N\geqslant 1}a_N q^N$.

There are two forms for the fermionic character $F_{r,s}^{(L,p)}$, depending on the finite parameter L. Let C_{p-3} and I_{p-3} stand for, respectively, the Cartan and incidence matrices of the Lie algebra A_{p-3}. Furthermore denote the i unit vector in $\mathbb{R}^{p-3}$ as $\vec{e}_i$, and set $\vec{e}_i=\vec{0}$ for $i<1$ or $i>(p-3)$. Then

$$F_{r,s}^{(L,p)}(q) = q^{-(1/4)(s-r)(s-r-1)} \sum_{\vec{m}\in 2\mathbb{Z}^{p-3}+\vec{Q}_{r,s}} q^{(1/4)\vec{m}^T C_{p-3}\vec{m}-(1/2)\vec{A}_{r,s}\vec{m}} \prod_{i=1}^{p-3} \begin{bmatrix} \frac{1}{2}(I_{p-3}\vec{m}+\vec{u}_{r,s}+L\vec{e}_1)_i \\ m_i \end{bmatrix}_q , \tag{2.5}$$

where $\vec{m}^T=(m_1,\ldots,m_{p-3})$. When $L+r-s$ is even,

$$\vec{A}_{r,s}=\vec{e}_{s-1}, \quad \vec{u}_{r,s}=\vec{e}_{s-1}+\vec{e}_{p-r-1}, \tag{2.6}$$

$$\vec{Q}_{r,s}=(r-1)\sum_{i=1}^{p-3}\vec{e}_i+(\vec{e}_{s-2}+\vec{e}_{s-4}+\cdots)+(\vec{e}_{p-r}+\vec{e}_{p+2-r}+\cdots);$$

and when $L+r-s$ is odd,

$$\vec{A}_{r,s}=\vec{e}_{p-s-1}, \quad \vec{u}_{r,s}=\vec{e}_{p-s-1}+\vec{e}_r, \tag{2.7}$$

$$\vec{Q}_{r,s}=(s-1)\sum_{i=1}^{p-3}\vec{e}_i+(\vec{e}_{r-1}+\vec{e}_{r-3}+\cdots)+(\vec{e}_{p-s}+\vec{e}_{p+2-s}+\cdots).$$

These two forms yield the same q-series in the limit $L\to\infty$. Proofs of the fermionic sums are given in Refs. 42–46.

Two sequences $\{\alpha_n\}$ and $\{\beta_n\}$ form a (bilateral) Bailey pair relative to a if they satisfy the relation

$$\beta_n=\sum_{j=-\infty}^{n}\frac{\alpha_j}{(q)_{n-j}(aq)_{n+j}}. \tag{2.8}$$

If we set $L=2l+r-s+2x$, then from (2.1) we can read off a (bilateral) Bailey pair relative to $a=q^{r-s+2x}$ as

$$\alpha_n=\begin{cases} q^{j(jp(p-1)+pr-(p-1)s)} & \text{for } n=pj-x \\ -q^{(jp-s)(j(p-1)-r)} & \text{for } n=pj-r-x, \\ 0 \quad \text{for otherwise} & \end{cases} \tag{2.9a}$$

$$\beta_n=\begin{cases} \dfrac{1}{(aq)_{2n}}F_{r,s}^{(2n+r-s+2x,p)}(q) & \text{for } \geqslant 0 \\ 0 \quad \text{otherwise.} & \end{cases} \tag{2.9b}$$

An important step in the Andrews–Bailey construction of the unitary models is to define another Bailey pair relative to a,

$$A_n=\begin{cases} q^{j^2p-sj+x(s-r-x)} & \text{for } n=pj-x \\ -q^{j^2p-sj+x(s-r-x)} & \text{for } n=pj-r-x \\ 0 & \text{otherwise} \end{cases} \tag{2.10a}$$

$$B_n=\begin{cases} \frac{1}{(aq)_{2n}} q^{n^2}a^nF_{r,s}^{(2n+r-s+2x,p)}(q^{-1}) & \text{for } n\geqslant 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.10b}$$

which are dual to (2.9).[30,34] Here

$$F_{r,s}^{(L,p)}(q^{-1}) = q^{(1/4)(s-r)(s-r-1)} \sum_{\vec{m}\in 2\mathbb{Z}^{p-3}+\vec{Q}_{r,s}} q^{(1/4)\vec{m}^TC_{p-3}\vec{m}+(1/2)(\vec{A}_{r,s}-\vec{u}_{r,s}-L\vec{e}_1)\vec{m}} \times \prod_{i=1}^{p-3} \begin{bmatrix} \frac{1}{2}(I_{p-3}\vec{m}+\vec{u}_{r,s}+L\vec{e}_1)_i \\ m_i \end{bmatrix}_q . \tag{2.11}$$

This dual transformation ($q\to q^{-1}$) takes us from the $M(p-1,p)$ finitized characters to the $M(1,p)$ finitized characters.[10,44] The nonunitary minimal model $M(1,p)$ has actually zero operator content in the usual range of r and s, but admits nontrivial finitizations. In fact, up to some prefactors, (2.11) is the finitization[47] of the $\mathbb{Z}_{p-2}$ parafermions[48] which describe the critical point between regimes I and II of the RSOS model.[28]

The Andrews–Bailey construction tells us that if (2.10) is a (bilateral) Bailey pair, then

$$A_n'=\left(\frac{(\rho_1)_n(\rho_2)_n(aq/\rho_1\rho_2)^n}{(aq/\rho_1)_n(aq/\rho_2)_n}\right)A_n , \tag{2.12a}$$

$$B_n'=\sum_{m=-\infty}^{n}\left(\frac{(\rho_1)_m(\rho_2)_m(aq/\rho_1\rho_2)_{n-m}(aq/\rho_1\rho_2)^m}{(q)_{n-m}(aq/\rho_1)_n(aq/\rho_2)_n}\right)B_m \tag{2.12b}$$

also forms a (bilateral) Bailey pair with respect to a. Now using the defining relation (2.8) with this new Bailey pair and taking the limit $n\to\infty$, one easily obtains the formula

$$\frac{(aq/\rho_1)_\infty(aq/\rho_2)_\infty}{(aq)_\infty(aq/\rho_1\rho_2)_\infty}\sum_{j=-\infty}^{\infty} q^{j(jp-s)+x(s-r-x)}\left(\frac{(\rho_1)_{pj-x}(\rho_2)_{pj-x}(aq/\rho_1\rho_2)^{pj-x}}{(aq/\rho_1)_{pj-x}(aq/\rho_2)_{pj-x}} - \frac{(\rho_1)_{pj-r-x}(\rho_2)_{pj-r-x}(aq/\rho_1\rho_2)^{pj-r-x}}{(aq/\rho_1)_{pj-r-x}(aq/\rho_2)_{pj-r-x}}\right) = \sum_{n=0}^{\infty}(\rho_1)_n(\rho_2)_n(aq/\rho_1\rho_2)^n\frac{q^{n^2}a^n}{(aq)_{2n}}F_{r,s}^{(2n+r-s+2x,p)}(q^{-1}). \tag{2.13}$$

We shall refer to the expression in (2.13) as the ‘‘generalized’’ character formula, and we will compute the effective central charge associated with it in Sec. III. In the limiting case (i) (and setting $x=0$), (2.13) becomes the character formula for $\chi_{s,r}^{(p,p+1)}$, where $1\leqslant r\leqslant(p-2)$ and $1\leqslant s\leqslant(p-1)$. Similarly, the ‘‘generalized’’ character (2.13) yields characters for the $N=1$ supersymmetric model $SM(p,p+2)$ in the case (ii) (with $\rho_2=-q^{r-s+1/2}$), while case (iii) (for example with $\rho_1=-yq^{1/2}$ and $\rho_2=-y^{-1}q^{1/2}$ for $r=s=1$ in the Neveu–Schwarz sector) leads to characters of the $N=2$ model with central charge $c=3(1-2/p)$.[34] Repeating the Andrews–Bailey construction starting from (2.1) and taking $L=2l+r-s+2x+1$, another ‘‘generalized’’ character can be obtained. The latter becomes the character formula for $\chi_{s,r+2}^{(p,p+1)}$ in case (i), and gives more characters for the supersymmetric models in the cases of (ii) and (iii) (please see Ref. 34 for more details). Since this second ''generalized'' character leads to the same central charge as (2.13), it will not be considered further in this work.

III. EFFECTIVE CENTRAL CHARGE

In this section, we shall calculate the asymptotic behavior of (2.13) as $q \to 1^-$. This method of computing the effective central charge for CFT fermionic characters are by now standard.[39,40,9,43] Therefore we will be brief with the procedure, but detailing in places where our calculation differs from the norm. First we shall predict the asymptotic growth of the ''generalized'' character using a physical argument. Subsequently we will derive this asymptotic behavior directly from the character formula.

A. Asymptotic behavior of the ''generalized'' character

Characters in CFT admit an interpretation as the partition function of the model defined on a cylinder with conformal boundary conditions on its rims.[49] The modular invariance property of these character formulas give us precise information about their asymptotic behavior. The working assumption in this section will be that (2.13) also gives the cylindrical partition function for some quantum field theory with appropriate boundary conditions labeled a and b. Note that a and b depend on the values of r and s, as well as ρ_1 and ρ_2. Define the modular parameter

$$q = e^{2\pi i \tau}, \quad \tilde{q} = e^{-2\pi i/\tau},$$

with

$$\tau = \frac{iR}{2\pi L} \tag{3.1}$$

for a cylinder of length L and circumference R. If we take the (imaginary) time coordinate to be in the R direction and space in the L direction, the generalized character (2.13) can be written as

$$\chi_{s,r}^{(p)}(\rho_1,\rho_2;\tau) = \mathrm{Tr}_P e^{-R\mathcal{H}_{ab}(\rho_1,\rho_2|p)/L} = \mathrm{Tr}_P q^{\mathcal{H}_{ab}(\rho_1,\rho_2|p)}, \tag{3.2}$$

where $\mathcal{H}_{ab}$ is the (normalized) dimensionless Hamiltonian of the field theory with open boundary conditions a and b. The trace is taken over the sector of the Hilbert space with boundary condition P along the circumference of the cylinder. For the cases (1.1), $\mathcal{H}_{ab}$ becomes $L_0 - \Delta(s,r)$ where $\Delta(s,r)$ is the appropriate conformal dimension for each unitary model. If instead we take space to be compactified in the R direction and time to evolve in the L direction, then the partition function will be

$$\chi_{s,r}^{(p)}(\rho_1,\rho_2;\tau) = \langle a | e^{-L\mathcal{H}_P(\rho_1,\rho_2|p)/R} | b \rangle, \tag{3.3}$$

where $\mathcal{H}_P$ is the dimensionless Hamiltonian with closed boundary condition P. Here $\langle a|$ and $|b\rangle$ represent the boundary states at the ends of the cylinder. Note that extra prefactors (in powers of q) due to normalization have been dropped from the right-hand side of (3.3), as they are irrelevant in the following limit (where $q \to 1^{-1}$). In the limit $L \to \infty$, the inner product (3.3) is dominated by the ground state of $\mathcal{H}_P$ with energy E_0,

$$\lim_{L\to\infty} \chi_{s,r}^{(p)}(\rho_1,\rho_2;\tau) \sim g_a g_b \tilde{q}^{E_0/4\pi^2}, \tag{3.4}$$

where we denote the contributions from each boundary as g_a and g_b. (3.4) is our prediction of the asymptotic behavior of (2.13).

The fermionic form of the generalized character (2.13) is most suitable for taking the asymptotic limit $q \to 1^-$. The important thing to notice here is that for fixed p, E_0 depends only on the parameters ρ_1 and ρ_2, and is independent of r and s. Standard arguments[9] (related to the r and s independence of E_0) give us the freedom to remove the restriction $\vec{Q}_{r,s}$ and linear terms in the exponent of q from the fermionic sum in this limit. Thus to compute the asymptotic behavior of the generalized character, i.e., to obtain the leading exponent of $\tilde{q}$, we could just concentrate on the simplest case of the identity representation $\chi_{1,1}^{(p)}(\rho_1,\rho_2;\tau)$. To implement the special limits (1.1) in this case, let us parametrize ρ_1 and ρ_2 as

$$\rho_1 = -\frac{q^{1/2}}{A}, \quad \rho_2 = -\frac{q^{1/2}}{B}, \tag{3.5}$$

thus we have

$$\begin{aligned} &\text{(i)} \quad A=0, \quad B=0, \\ &\text{(ii)} \quad A=0, \quad B=1, \\ &\text{(iii)} \quad A=1, \quad B=1, \end{aligned} \tag{3.6a}$$

with $x=0$ and $a=1$ for all three cases. With our choice of parametrization, the limits (1.1) actually become (i) $\rho_1 \to -\infty$, $\rho_2 \to -\infty$ and (ii) $\rho_1 \to -\infty$, $\rho_2 =$ finite. However, we can still obtain the same conformal characters from (2.13). Here we specialized to the case of $y=1$ for the $N=2$ CFT. The limits (ii) and (iii) used here lead to the Neveu—Schwarz characters for the supersymmetric models. We will not consider the Ramond sector, although it can be treated by a straightforward generalization of the computation presented here. A convenient parametrization of the ground state energy E_0 is

$$E_0(A,B|p) = -\frac{\pi^2}{6}\tilde{c}(A,B|p). \tag{3.7}$$

From (3.4), we shall interpret $\tilde{c}$ as the ''generalized'' effective central charge, and expect that in the limits (3.6), it will take on the values of $1-6/p(p+1)$, $3/2-12/p(p+1)$, and $c=3(1-2/p)$, respectively.

B. Effective central charge

After all the simplifications mentioned above, the q-series we shall consider is

$$\begin{aligned} \tilde{\chi}_{1,1}^{(p)}(A,B;q) = \sum_{n=0}^{\infty} \sum_{m_1,\ldots,m_{p-3}=0}^{\infty} \left(-\frac{q^{1/2}}{A}\right)_n \\ \times \left(-\frac{q^{1/2}}{B}\right)_n (AB)^n \frac{q^{\vec{m}^T C_{p-3}\vec{m} - 2nm_1 + n^2}}{(q)_{2n}} \prod_{i=1}^{p-3} \begin{bmatrix} (I_{p-3}\vec{m} + n\vec{e}_1)_i \\ 2m_i \end{bmatrix}_q . \end{aligned} \tag{3.8}$$

If the coefficients in this series $\tilde{\chi}_{1,1}^{(p)} = \sum a_M q^M$ behave like $a_M \sim e^{2\pi\sqrt{M\tilde{c}/6}}$ for large M, then as $q \to 1^-$, $\tilde{\chi}_{1,1}^{(p)}$ diverges like $\tilde{q}^{-\tilde{c}/24}$. In other words one can obtain the ''generalized'' central charge $\tilde{c}$ from the asymptotic growth of the coefficient a_M. The latter is computed by applying the saddle point method to

$$a_{M-1} = \oint \frac{dq}{2\pi i} \tilde{\chi}_{1,1}^{(p)}(A,B;q) q^{-M} = \oint \frac{dq}{2\pi i} \sum_n \sum_{\vec{m}} f(n,\vec{m};q). \tag{3.9}$$

The saddle point occurs at the point where the derivatives of

$$\log f(n,\vec{m};q) \approx \int_0^n \log\left(1+\frac{q^k}{A}\right)dk + \int_0^n \log\left(1+\frac{q^k}{B}\right)dk + n\log(AB)$$
$$-\int_0^{2n}\log(1-q^k)\,dk + (n^2-2nm_1+\vec{m}^T C_{p-3}\vec{m}-M)\log q$$
$$+\sum_{i=1}^{p-3}\left(\int_0^{(I_{p-3}\vec{m}+n\vec{e}_1)_i} - \int_0^{(I_{p-3}\vec{m}+n\vec{e}_1-2\vec{m})_i} - \int_0^{2m_i}\right)\log(1-q^k)\,dk \tag{3.10}$$

with respect to n, $m_1,\ldots,m_{p-3}$ and q are all zero. In deriving the expression in (3.10), sums such as $\log\{(q)_n\}$ and $\log\{(-q^{1/2}/A)_n\}$ were approximated by integrals. There are several ways to make this approximation. Ultimately, the difference between the various approximation schemes is equivalent to a difference in the linear terms in the exponent of q, and do not influence the quadratic terms. Since the asymptotic growth is not expected to depend on the linear terms as explained above, we have the freedom to use the following two (different) approximations:

$$\log\{(q)_n\} \sim \int_0^n \log(1-q^k)\,dk, \tag{3.11a}$$

$$\log\left\{\left(-\frac{q^{1/2}}{A}\right)_n\right\} \sim \int_0^n \log\left(1+\frac{q^k}{A}\right)dk. \tag{3.11b}$$

This combination of approximations was chosen to simplify the algebra after differentiation.

Let us define

$$v_i = q^{-2m_i}, \quad w_i = q^{(I_{p-3}\vec{m}+n\vec{e}_1)_i}. \tag{3.12}$$

The differentiation with respect to m_i and n produced the following set of relations for their saddle point values $\overline{m_i}$ and $\overline{n}$:

$$(1-y_i)^2 = \prod_{j=1}^{p-3} y_j^{I_{ij}}, \tag{3.13a}$$

$$q^{2\overline{n}\delta_{1,i}}(1-x_i)^2 = \prod_{j=1}^{p-3} x_j^{I_{ij}}, \tag{3.13b}$$

$$(1-q^{2\overline{n}})^2 = (A+q^{\overline{n}})(B+q^{\overline{n}})q^{2\overline{n}}x_1, \tag{3.13c}$$

where

$$x_i = \frac{(1-\overline{w_i})\overline{v_i}}{1-\overline{v_i}\,\overline{w_i}}, \quad y_i = \frac{(1-\overline{w_i})}{1-\overline{v_i}\,\overline{w_i}}. \tag{3.14}$$

It is easy to show that in the special cases of (3.6), (3.13) reduces to a system of algebraic equations governed by the algebras A_{p-2}, A_{p-1}, and D_{p-1}, respectively. For these algebras, the corresponding systems of equations are solved in the literature, and are known to be related to the Thermodynamic Bethe Ansatz (TBA) approach (see, e.g., Refs. 50–54). Here we can easily write down the solution for y_i as

$$y_i = \frac{\sin^2 (1+i) \dfrac{\pi}{p}}{\sin^2 \dfrac{\pi}{p}}. \tag{3.15}$$

One can also show that

$$x_i = \frac{\sin^2 (p-1-i)\theta}{\sin^2 \theta} \tag{3.16}$$

satisfies (3.13) with the closure conditions

$$x_{p-2} = 1, \tag{3.17a}$$

$$x_0 = \frac{\sin^2 (p-1)\theta}{\sin^2 \theta} = q^{-2\bar{n}}, \tag{3.17b}$$

$$x_{-1} = \frac{\sin^2 p\theta}{\sin^2 \theta} = (1 + A q^{-\bar{n}})(1 + B q^{-\bar{n}}), \tag{3.17c}$$

(3.17d)

The parameter θ is related to A and B by the relation

$$(A+B)\sin\theta + AB \sin(p-1)\theta = \sin(p+1)\theta. \tag{3.18}$$

To compute $\log f(n,\vec{m};q)$ at the stationary point with respect to m_i and n, we first rewrite it using the relations

$$\int_0^{\bar{z}} \log(1-q^k)\,dk = \frac{1}{\log q}\left[L(1-q^{\bar{z}}) + \frac{1}{2}\log(1-q^{\bar{z}})\log q^{\bar{z}} \right], \tag{3.19a}$$

$$\int_0^{\bar{z}} \log\left(1+\frac{q^k}{A}\right) dk = \frac{1}{\log q}\left[L\left(\frac{q^{\bar{z}}}{q^{\bar{z}}+A}\right) - L\left(\frac{1}{1+A}\right) + \frac{1}{2}\log A \log\left(\frac{1+A}{q^{\bar{z}}+A}\right)\right] + \frac{\bar{z}}{2}\log\left(1+\frac{q^{\bar{z}}}{A}\right). \tag{3.19b}$$

The Rogers dilogarithm in (3.19) is defined by[55]

$$L(z) = Li_2(z) + \frac{1}{2}\log z \log(1-z), \quad Li_2(z) = -\int_0^z \frac{\log(1-w)}{w}\,dw \tag{3.20}$$

and $L(1) = \pi^2/6$. The five terms relation for the dilogarithm in our case can be written as

$$L(1-w_i) - L(1-v_i w_i) - L(1-v_i^{-1}) = L(1-y_i^{-1}) - L(1-x_i^{-1}). \tag{3.21}$$

Hence we have

$$\log f(n,\vec{m};q)\Big|_{\substack{\vec{m}=\bar{\vec{m}} \\ n=\bar{n}}} \approx -M \log q - \frac{\pi^2 \tilde{c}(A,B|p)}{6 \log q}, \tag{3.22}$$

where

$$\tilde{c}(A,B|p)=\frac{1}{L(1)}\left(L\left(\frac{1}{1+A}\right)+L\left(\frac{1}{1+B}\right)+L\left(1-\frac{1}{x_0}\right)-L\left(\frac{1}{1+\sqrt{x_0}A}\right)-L\left(\frac{1}{1+\sqrt{x_0}B}\right)\right.$$
$$\left.+\sum_{i=1}^{p-3}\left[L\left(1-\frac{1}{x_i}\right)-L\left(1-\frac{1}{y_i}\right)\right]+\frac{1}{2}\log A\log\left(\frac{1+\sqrt{x_0}A}{1+A}\right)+\frac{1}{2}\log B\log\left(\frac{1+\sqrt{x_0}B}{1+B}\right)\right). \tag{3.23}$$

By differentiating (3.22) with respect to q, we found the saddle point value of q to be

$$\overline{q}=e^{-\sqrt{\pi^2\tilde{c}/6M}}. \tag{3.24}$$

This leads to the expected asymptotic behavior of a_M for large M, and hence we can interpret $\tilde{c}(A,B|p)$ as a ''generalized'' effective central charge for (2.13). The sums in (3.23) can be further simplified using dilogarithm sum rules[55,56] to yield

$$\sum_{i=1}^{p-3}L\left(1-\frac{1}{y_i}\right)=\left(p-5+\frac{6}{p}\right)L(1), \tag{3.25}$$

$$L\left(1-\frac{1}{x_0}\right)+\sum_{i=1}^{p-3}\left[L\left(1-\frac{1}{x_i}\right)\right]=(p-1)L(1)-p(p-1)\theta^2+2Li_2\left(-\frac{\sin(p-1)\theta}{\sin\theta},p\theta\right)$$
$$+\log\left(\frac{\sin(p-1)\theta}{\sin\theta}\right)\log\left(\frac{\sin p\theta}{\sin\theta}\right), \tag{3.26}$$

where

$$Li_2(r,\theta)=\mathrm{Re}\{Li_2(re^{i\theta})\}=-\frac{1}{2}\int_0^r\frac{\log(1-2x\cos\theta+x^2)}{x}\,dx. \tag{3.27}$$

The resultant expression for the ''generalized'' central charge is

$$\tilde{c}(A,B|p)=\frac{1}{L(1)}\left[Li_2(-A)+Li_2(-B)-Li_2\left(-A\frac{\sin(p-1)\theta}{\sin\theta}\right)-Li_2\left(-B\frac{\sin(p-1)\theta}{\sin\theta}\right)\right.$$
$$\left.+\left(4-\frac{6}{p}\right)L(1)-p(p-1)\theta^2+2Li_2\left(-\frac{\sin(p-1)\theta}{\sin\theta},p\theta\right)\right]. \tag{3.28}$$

The simple expression in (3.28) is the main result of this letter. It gives the effective central charge associated with the ''generalized'' character formula (2.13) in terms of dilogarithm functions. The expressions in (3.23) and (3.28) are valid for $A\geqslant 0$ and $B\geqslant 0$.

C. Special cases

Consider the domain $A=0$, and $B=\sin(p+1)\theta/\sin\theta$ follows from (3.18). To implement the special case (i), we take the limit $B\to 0$, thus yielding $\theta=\pi/(p+1)$. Consequently by using the identity[55]

$$Li_2(2\cos\theta,\theta)=\left(\frac{\pi}{2}-\theta\right)^2,$$

we obtained $\tilde{c}(0,0|p)=1-6/p(p+1)$, which is the central charge of the unitary model $M(p,p+1)$. In the limit (ii), taking $B=1$ we found $\theta=\pi/(p+2)$. Using the limit

$$Li_2\left(-\frac{\sin(p-1)\theta}{\sin\theta}\right)+2Li_2\left(-\frac{\sin(p-1)\theta}{\sin\theta},p\theta\right)\bigg|_{\theta=\pi/(p+2)}=\frac{2}{3}\pi^2-2\frac{(2p+1)}{(p+2)^2}\pi^2,$$

we recover the central charge of the $N=1$ unitary model $\tilde{c}(0,1|p)=\frac{3}{2}-12/p(p+2)$.

It is interesting that $\tilde{c}(0,B|p)$ is a smooth monotonic function of B between the above two limits. In particular for the case of $p=3$, we have a function which connects the central charges of the Ising and tricritical Ising model, while the ''generalized'' character (2.13) takes us from the tricritical Ising character $\chi_{1,1}^{(4,5)}+\chi_{1,4}^{(4,5)}$ to the Ising character $\chi_{1,1}^{(3,4)}$. Hence it is desirable to compare $\tilde{c}(0,B|3)$ with the known ground state scaling function $\mathcal{C}(r)$ obtained from TBA.[13] Recall that the latter is a function of a scaling parameter r, with UV limit $(r\rightarrow 0)\frac{7}{10}$ and IR limit $(r\rightarrow\infty)\frac{1}{2}$, respectively. Therefore to compare the two expressions, we need to find a parametrization of the variable B in terms of r. This can always be done since one can in principle invert the function $\tilde{c}(0,B|3)$ to obtain the parametrization $B(r)=\tilde{c}^{-1}(\mathcal{C}(r))$. However we were unable to express $B(r)$ in a simple and closed form. This is perhaps not surprising since $\mathcal{C}(r)$ is written as an integral involving two pseudoenergies ϵ_1 and ϵ_2, which in turn are given by two coupled integral equations involving r. Only in the UV or IR limits do we get a simplification of the integral equations, which then allow us to write $\mathcal{C}$ in terms of dilogarithms.[13] Hence the parametrization $B(r)$, which yields an expression for $\mathcal{C}(r)$ in terms of dilogarithms for general r, is likely to be complicated. It is also unclear at this stage whether this parametrization admits any physical interpretations.

Now let us focus on the other domain $A=B$. The relation (3.18) tells us $A=[\cos(p+1)\theta/2]/[\cos(p-1)\theta/2]$. Of course in the limiting case (i), $A\rightarrow 0$, we found $\theta=\pi/(p+1)$ as before. Once again $\tilde{c}(A,A|p)$ is a smooth monotonic function of A. For the $N=2$ supersymmetric limit (iii), taking $A\rightarrow 1$, we get $\theta=0$ and $\tilde{c}(1,1|p)=3(1-2/p)$ as expected. Hence $\tilde{c}(A,B|p)$ indeed give us a function which interpolates between the central charges of an unitary model and its supersymmetric counterparts.

IV. DISCUSSION

In this work, we have studied the asymptotic behavior of the ''generalized'' character formula $\chi_{s,r}^{(p)}(\rho_1,\rho_2;q)$ (2.13) in the limit $q\rightarrow 1^-$. In this limit, we show that the q-series diverges like $\tilde{q}^{-\tilde{c}/24}$ and we found a simple expression (3.28) for the ''generalized'' effective central charge $\tilde{c}$ in terms of dilogarithms. In the limiting cases (i), (ii), and (iii), $\tilde{c}$ yields the central charges of the unitary models and their supersymmetric counterparts.

Having stated our conclusion, we shall take the liberty to indulge in some (pure) speculations. Of course it is not surprising that we can find a function which reproduces the correct central charges in the various limits. Indeed $\chi_{s,r}^{(p)}(\rho_1,\rho_2;q)$ also becomes the corresponding CFT characters in these limits. But what is *a priori* not expected from the Andrews–Bailey construction is that the ''generalized'' character (2.13) would exhibit the asymptotic behavior found in Sec. III. The prediction for this behavior was based on the assumption that (2.13) gives the partition function for some quantum field theory. This field theory must be invariant under interchanging roles of space and time. This suggests that $\chi_{s,r}^{(p)}(\rho_1,\rho_2;q)$, when multiplied by a suitable factor $q^{\mathcal{D}_{s,r}(\rho_1,\rho_2|p)}$, may be modular covariant. It would be very interesting to show directly from (2.13) that

$$q^{\mathcal{D}_{s,r}(\rho_1,\rho_2|p)}\chi_{s,r}^{(p)}(\rho_1,\rho_2;q)=\sum_{s',r'}S_{s,r}^{s',r'}(\rho_1,\rho_2|p)\tilde{q}^{\mathcal{D}_{s',r'}(\rho_1,\rho_2|p)}\chi_{s',r'}^{(p)}(\rho_1,\rho_2;\tilde{q})$$

for some ''generalized'' S-matrix. Presumably the elements of this S-matrix (if it exits) can be calculated from the nonperturbative corrections to the saddle point.[57,58] This computation would be much more involved than that in Sec. III since the elements of S depend on r and s

Another interesting puzzle is the nature of the quantum field theory with the Hamiltonian $\mathcal{H}_P$ discussed in Sec. III. From (3.7), the ground state energy of $\mathcal{H}_P(\rho_1,\rho_2|p)$ is proportional to

$\tilde{c}(A,B|p)$ which is written in terms of dilogarithms. This seems to indicate that $\mathcal{H}_P(\rho_1,\rho_2|p)$ is the Hamiltonian for a (maybe irrational) CFT which interpolates between the $N=0$, $N=1$ and $N=2$ unitary models. Last, it is worth investigating whether the approach taken here can be apply to the trinomial analog of Bailey's lemma.[37,12]

ACKNOWLEDGMENTS

The author is grateful to O. Foda and O. Warnaar for reading the manuscript and many valuable discussions. I would also like to thank R. Kedem, A. Kirillov, C. Pisani, and D. Pihan for useful conversations. This work is supported by the Australian Research Council.

[1] A. A. Belavin, A. M. Polyakov, and A. B. Zamolodchikov, Nucl. Phys. B **241**, 333 (1984).
[2] A. B. Zamolodchikov, Int. J. Mod. Phys. A **3**, 743 (1988).
[3] A. Rocha-Caridi, in *Vertex Operators in Mathematics and Physics*, edited by J. Lepowsky, S. Mandelstam, and I. M. Singer, (Springer, Berlin, 1985).
[4] V. G. Kac, *Infinite Dimensional Lie Algebras* (Cambridge U.P., Cambridge, 1990), 3rd ed.
[5] B. L. Feigin and D. B. Fuchs, Funct. Anal. Appl. **16**, 114 (1982); **17**, 241 (1983).
[6] V. G. Kac and D. H. Peterson, Adv. Math. **53**, 125 (1984).
[7] P. Goddard, A. Kent, and D. Olive, Commun. Math. Phys. **103**, 105 (1986).
[8] P. Bouwknegt, J. McCarthy, and K. Pilch, Nucl. Phys. **352**, 139 (1991); P. Bouwknegt and K. Schoutens, Phys. Rep. **223**, 183 (1993).
[9] S. Dasmahapatra, R. Kedem, T. R. Klassen, B. M. McCoy, and E. Melzer, Int. J. Mod. Phys. B **7**, 3617 (1993).
[10] A. Berkovich and B. M. McCoy, Lett. Math. Phys. **37**, 49 (1996).
[11] A. Berkovich, B. M. McCoy, and A. Schilling, Commun. Math. Phys. **191**, 325 (1998).
[12] A. Berkovich, B. M. McCoy, and P. A. Pearce, Nucl. Phys. B **19**, 597 (1998).
[13] Al. B. Zamolodchikov, Nucl. Phys. B **358**, 524 (1991).
[14] A. B. Zamolodchikov and Al. B. Zamolodchikov, Nucl. Phys. B **379**, 602 (1992).
[15] P. Fendley and H. Saleur, preprint USC-93-022, hepth/9310058.
[16] E. Melzer, Int. J. Mod. Phys. A **9**, 5753 (1994).
[17] V. V. Bazhanov, S. L. Lukyanov, and A. B. Zamolodchikov, Commun. Math. Phys. **177**, 381 (1996); **190**, 247 (1997).
[18] D. Fioravanti, F. Ravanini, and M. Stanishkov, Phys. Lett. B **367**, 113 (1996).
[19] P. Bouwknegt, A. W. W. Ludwig, and K. Schoutens, Phys. Lett. B **359**, 304 (1995).
[20] S. Iso, Nucl. Phys. **A443**, 581 (1995).
[21] D. Bernard, V. Pasquier, and D. Serban, Nucl. Phys. B **428**, 612 (1994).
[22] L. J. Rogers, Proc. London Math. Soc. **25**, 318 (1894).
[23] L. J. Rogers, S. Ramanujan, and G. H. Hardy, Proc. Cambridge Philos. Soc. **19**, 211 (1919).
[24] G. E. Andrews, *The Theory of Partitions* (Addison–Wesley, Reading, MA, 1976).
[25] I. Schur, S.-B. Preuss. Akad. Wiss. Phys.-Math. KI., 302 (1917).
[26] G. E. Andrews, Scr. Math. **28**, 297 (1970).
[27] E. Melzer, Int. J. Mod. Phys. A **9**, 1115 (1994).
[28] G. E. Andrews, R. J. Baxter, and P. J. Forrester, J. Stat. Phys. **35**, 193 (1984).
[29] O. Foda and Y-H. Quano, Int. J. Mod. Phys. A **12**, 1651 (1997).
[30] G. E. Andrews, Pac. J. Math. **114**, 267 (1984).
[31] W. N. Bailey, Proc. London Math. Soc. **49**, 421 (1947); **50**, 1 (1949).
[32] P. J. Forrester and R. J. Baxter, J. Stat. Phys. **38**, 435 (1985).
[33] B. M. McCoy (private communication).
[34] A. Berkovich, B. M. McCoy, and A. Schilling, Physica A **228**, 33 (1996).
[35] A. Berkovich, B. M. McCoy, A. Schilling, and S. O. Warnaar, Nucl. Phys. B **499**, 621 (1997).
[36] A. Schilling and S. O. Warnaar, preprint, q-alg/9607014.
[37] G. E. Andrews and A. Berkovich, Commun. Math. Phys. **192**, 245 (1998).
[38] P. Fendley, H. Saleur, and Al. B. Zamolodchikov, Int. J. Mod. Phys. A **8**, 5751 (1993).
[39] B. Richmond and G. Szekeres, J. Aust. Math. Soc., Ser. A **31**, 362 (1981).
[40] W. Nahm, A. Recknagel, and M. Terhoeven, Mod. Phys. Lett. A **8**, 1835 (1993).
[41] D. A. Huse, Phys. Rev. B **30**, 3908 (1984).
[42] A. Berkovich, Nucl. Phys. B **431**, 315 (1994).
[43] S. O. Warnaar, J. Stat. Phys. **82**, 657 (1996).
[44] S. O. Warnaar, J. Stat. Phys. **84**, 49 (1996).
[45] A. Schilling, Nucl. Phys. B **459**, 393 (1996).
[46] O. Foda and T. A. Welsh, Melzer's Identities Revisited, math.QA/9811156.
[47] O. Foda and S. O. Warnaar (private communication).
[48] A. B. Zamolodchikov and V. A. Fateev, Sov. Phys. JETP **62**, 215 (1985).
[49] J. L. Cardy, Nucl. Phys. B **324**, 581 (1989).
[50] Al. B. Zamolodchikov, Nucl. Phys. B **358**, 497 (1991).
[51] T. R. Klassen and E. Melzer, Nucl. Phys. B **370**, 511 (1992).

[52] A. Klumper and P. A. Pearce, Physica A **183**, 304 (1992).
[53] V. Bazhanov and N. Reshetikhin, J. Phys. A **23**, 1477 (1990).
[54] M. J. Martins, Phys. Rev. Lett. **22**, 419 (1991).
[55] L. Lewin, *Polylogarithms and Associated Functions* (North-Holland, New York, 1981).
[56] A. N. Kirillov, Prog. Theor. Phys. Suppl. **118**, 61 (1995).
[57] M. Terhoeven, Mod. Phys. Lett. A **9**, 133 (1994).
[58] W. Nahm, in *Interface between Physics and Mathematics*, edited by W. Nahm and J. Shen (World Scientific, Singapore, 1994).

Modular invariance on the torus and Abelian Chern–Simons theory

J. Guerrero
Instituto de Astrofísica de Andalucía, Apartado Postal 3004, 18080 Granada, Spain, and Instituto de Física Teórica y Computacional Carlos I, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva, Granada 18002, Spain, and Departimento di Scienze Fisiche, Mostra d'Oltremare Pad. 19, 80125 Napoli, Italy

M. Calixto
Instituto de Física Teórica y Computacional Carlos I, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva, Granada 18002, Spain and Department of Physics, University of Wales Swansea, Singleton Park, SA2 8PP, United Kingdom

V. Aldaya
Instituto de Astrofísica de Andalucía, Apartado Postal 3004, 18080 Granada, Spain and Instituto de Física Teórica y Computacional Carlos I, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva, Granada 18002, Spain

(Received 23 September 1998; accepted for publication 16 April 1999)

The implementation of modular invariance on the torus as a phase space at the quantum level is discussed in a group-theoretical framework. Unlike the classical case, at the quantum level some restrictions on the parameters of the theory should be imposed to ensure modular invariance. Two cases must be considered, depending on the cohomology class of the symplectic form on the torus. If it is of integer cohomology class n, then full modular invariance is achieved at the quantum level only for those wave functions on the torus which are periodic if n is even, or antiperiodic if n is odd. If the symplectic form is of rational cohomology class n/r, a similar result holds—the wave functions must be either periodic or antiperiodic on a torus r times larger in both directions, depending on the parity of nr. Application of these results to the Abelian Chern–Simons theory is discussed. [S0022-2488(99)04007-4]

I. INTRODUCTION

Since the pioneer work by Dirac[1] on the quantization of constrained systems, much work has been done on this subject, and plenty of methods have been developed to face this interesting and, many times, difficult problem. Roughly speaking, the different methods can be classified into two types, depending on whether the quantization of the corresponding unconstrained system is first performed and then the constraints imposed at the quantum level (the ''quantize-first'' method) or the constraints are first imposed and then the quantization of the resulting ''reduced'' system is performed (the ''constrain-first'' method). An example of the former is given by the above-mentioned paper by Dirac,[1] while the latter was originated by the work of Faddeev.[2] Many other procedures derive from these two, adapted to the properties of the particular system under consideration. Thus, for instance, the Becchi–Rouet–Stora–Tyutin (BRST) quantization is a ''quantize-first'' technique adapted to the covariant quantization of gauge invariant systems.[3] Also, the method proposed by Ashtekar[4] was designed to simplify the form of the quantum constraints in quantizing gravity. Alternatively, symplectic or Marsden–Weinstein reduction[5] is a specific technique developed to obtain a reduced classical phase space, which is the starting point for (some sort of) geometric quantization.[6]

The main drawback of the ''constrain-first'' method lies in the fact that the classical phase space could not be properly defined as a differential manifold or, even more, the classical equation of motion might have no general solution. In addition, all the problems that geometric quantization encounters in dealing with nontrivial phase spaces must be considered (anomalies, i.e., the lack of invariant polarizations, the search for operators compatible with the polarization, etc.).

The troubles with the ''quantize-first'' methods appear in the implementation of the quantum constraints; only quadratic constraints can be directly imposed due to normal-order ambiguities. Besides, finding the operators that preserve the quantum constraints is a nontrivial problem.

In Ref. 7 a method for studying quantum systems with constraints on a group-theoretical framework, Algebraic Quantization on a Group (AQG), was introduced. AQG is a ''quantize-first'' method in which both the unconstrained systems and the constraints are supposed to be dealt with in a group setting. This could seem, at first instance, a severe restriction but, in practice, most of the interesting cases can be treated with this formalism, and the advantages it provides are numerous. In particular, there are no ambiguities in the imposition of quantum constraints (even for nonpolynomical ones), and there is an operative characterization for the operators that preserves the quantum constraints.

Another advantage of AQG is the possibility of implementing the nontrivial topology of a phase space as a ''contraint subgroup'' that contains the first homotopy group of the phase space, made of discrete transformations, which can be easily addressed in this formalism.

In Ref. 8 the quantization of the Heisenberg–Weyl (HW) group with constraints was considered and the particular case of the HW group on the torus was studied. Now, we wish to implement modular invariance on the torus at the quantum level. In general, the modular invariance of a conformal theory formulated on a Riemannian surface of genus g,Σ_g, refers to the quotient group $\mathrm{Diff}(\Sigma_g)/\mathrm{Diff}_0(\Sigma_g)$, where $\mathrm{Diff}(\Sigma_g)$ is the group of diffeomorphisms of Σ_g and the subscript 0 designates the normal subgroup of diffeomorphisms connected to the identity (see, e.g., Refs. 9 and 10).

Clearly, modular transformations on the torus are the $SL(2,Z)$ subgroup of the group $SL(2,R)\approx Sp(2,R)$ of linear symplectic transformations of the plane that preserves the torus. Therefore we can implement them in the formalism of Algebraic Quantization on a Group by considering the Schrödinger group (or Weyl-Symplectic group, see Ref. 11) $WSp(2,R)$ as the symmetry group of the unconstrained system and imposing the appropriate constraints to obtain a torus as the (reduced) symplectic manifold, pretty much in the same manner as in Ref. 8. Then we expect to obtain modular transformations as good operators, i.e., those preserving the Hilbert space of wave functions satisfying the constraints. However, to obtain full modular invariance, we must impose some restrictions on the parameters of the theory. As in Ref. 8, three different cases should be considered, depending on the cohomology class of the symplectic form on the torus, which can be integer, fractional, or irrational. Only the integer and fractional cases will be considered here, since the irrational one requires techniques from noncommutative geometry[12] and lies beyond the scope of this paper.

These results are applied to $2+1$D Abelian Chern–Simons theory and compared with the ones obtained in the literature.

The present paper is organized as follows: In Sec. III we study the Schrödinger group without constraints and compute the metaplectic (or spinor) representation with the help of a higher-order polarization. Section IV is devoted to the determination of the constrained Hilbert space and good operators when the phase-space is constrained to be a torus. Two cases are considered, the one for which the symplectic form on the torus is of integer cohomology class n (Sec. IV A), where full modular invariance is obtained only when the wave functions are periodic for n even or antiperiodic for n odd, and the case of symplectic form of rational cohomology class n/r (Sec. IV B), where full modular invariance is obtained only when the wave functions are periodic for nr even or antiperiodic for nr odd. Here periodicity and antiperiodicity are understood in a torus which is r times larger in both directions. Finally, Sec. V is devoted to the application of our study to $2+1$D Abelian Chern–Simons theory.

In the Appendix, we study the representations of the subgroup T of constraints both for the integral and fractional case.

II. ALGEBRAIC QUANTIZATION ON A GROUP

Algebraic Quantization on a Group (AQG) (see Refs. 7 and 8) is a group-theoretical procedure developed for quantizing systems with constraints (both first and second class) in a first-quantize-then-constrain basis. The starting point is the group $\tilde{G}$ of quantum symmetries of the unconstrained system, which is a central extension by U(1) of the group G of classical symmetries of the unconstrained system. From $\tilde{G}$, a subgroup T, called the structure group, is selected for defining the constraints. For convenience, T is chosen to include the U(1) subgroup of the central extension, which accounts for the phase invariance of quantum mechanics [U(1) equivariance], in such a way that $\tilde{G}/T$ is the classical reduced phase space of the constrained system. (To be precise, $\tilde{G}$ contains in general symmetries without symplectic content, like time translations or rotations, so that $\tilde{G}/T$ is the reduced presymplectic manifold of the constrained system.)

The quantum Hilbert space $\mathcal{H}_T$ for the constrained system is defined by selecting, from the Hilbert space $\mathcal{H}$ associated with a unitary irreducible representation $U(\tilde{G})$ of $\tilde{G}$, those wave functions that transform irreducibly under a given unitary irreducible representation $D(T)$ of T. We shall say that these wave functions satisfy the T-function condition (or T-equivariance condition), which has the general form:

$$\Psi^{\alpha}(g_T * g) = D^{\alpha}(g_T)\Psi^{\alpha}(g), \quad \forall g_T \in T, \tag{1}$$

where the index α in D ranges over the set $\hat{T}$, the Pontryagin dual of T—that is, the set of all unitary irreducible representations of T. Precisely stated, α will be allowed to vary along the subset $\hat{T}_U \subset \hat{T}$ of those representations which are contained in the restriction of $U(\tilde{G})$ to T; otherwise the constraints would be inconsistent and the constrained Hilbert space $\mathcal{H}_T$ would be trivial. In particular, the representation D^{α}, when restricted to the subgroup $U(1) \subset T$, should be the natural (faithful) representation of U(1), $D^{\alpha}(\zeta) = \zeta, \ \forall \zeta \in U(1)$. That is, the T-equivariance condition must contain the U(1)-equivariance condition. Complex functions on the group satisfying the T-equivariance condition can be identified with sections of the vector bundle associated with the principal bundle $T \to \tilde{G} \to \tilde{G}/T$ through the representation D^{α} of T.[13]

Both the unitary irreducible representations $U(\tilde{G})$ and $D(T)$ can be obtained, for instance, by using the Group Approach to Quantization (GAQ) technique (see Ref. 7 and references therein), which uses the method of polarizations (see below) to reduce the left-regular representation of the group acting on U(1)-equivariant complex functions on the group $\tilde{G}$.

An important concept that we are forced to introduce is the notion of *good operators*, defined as those preserving the constrained Hilbert space $\mathcal{H}_T$. It is clear that, since $\mathcal{H}_T$ is in general smaller than $\mathcal{H}$, not all operators in $\tilde{G}$ will preserve it; otherwise the representation $U(\tilde{G})$ would be reducible. It is difficult to give a general characterization of these operators (for instance, there can be operators preserving $\mathcal{H}_T$ which belong neither to $\tilde{G}$ nor to its enveloping algebra, escaping to any algebraic or differential characterization), but we can find all good operators in $\tilde{G}$ simply by considering the *little group* of the representation $D^{\alpha}(T)$ of T—that is, the subgroup G_{good} of elements g_g that send the representation $D^{\alpha}(T)$ to an equivalent one under the adjoint action:

$$D^{\alpha}_{g_g}(g_T) \equiv D^{\alpha}(g_g * g_T * g_g^{-1}) \approx D^{\alpha}(g_T), \quad \forall g_T \in T, \quad \forall g_g \in G_{\text{good}}. \tag{2}$$

Note that this definition generalizes the (sufficient) ones given in Refs. 7 and 8. For instance, in the case in which the representation $D^{\alpha}(T)$ is one dimensional (in particular, if T is Abelian), the definition above gives $D^{\alpha}_{g_g}(T) = D^{\alpha}(T)$, and the sufficient condition given in Ref. 8,

$$[G_{\text{good}}, T] \subset \ker D^{\alpha}(T), \tag{3}$$

also proves to be necessary. This characterization reproduces the standard one for the case of first-class constraints, for which $T=C\times U(1)$, where C is the subgroup of constraints [U(1) only accounts for the phase-invariance of quantum mechanics]. If we choose for C the trivial representation [for U(1) the natural representation must always be chosen], then

$$[G_{\rm good},C]\subset C. \tag{4}$$

This condition gives $G_{\rm good}$ as the normalizer of the constraints, as is usually the case (see, e.g., Ref. 4). However, if a nontrivial representation of C is chosen, the subgroup of good operators can be smaller that the normalizer of the constraints, revealing a strong dependence of $G_{\rm good}$ on the representation $D^{\alpha}(T)$ of T and, therefore, we should use the more precise notation $G^{\alpha}_{\rm good}$ for the subgroup of operators preserving the reduced Hilbert space $\mathcal{H}^{\alpha}_{T}$. Note that, from the very definition of little group, $G^{\alpha}_{\rm good}\subset N_T$, $\forall\alpha\in\hat{T}_U$, where N_T is the normalizer of T in $\tilde{G}$, so that the appropriate place to look for good operators will be in N_T.

It is useful to examine the case in which C is an invariant subgroup of $\tilde{G}$ and we choose $D(T)$ to be the restriction of $U(\tilde{G})$ to T [or $U(\tilde{G})$ to be the induced representation by $D(T)$]. Then the constraints are trivial, i.e., they do not imply additional restrictions on the wave functions, and the constrained and unconstrained Hilbert spaces coincide. Moreover, the subgroup of good operators turns out to be the whole $\tilde{G}$. In this case, C is called a gauge group (see Ref. 14).

A separate study is warranted by the case when T cannot be written as $C\times U(1)$, for instance when T is a nontrivial central extension of C by U(1). In this case, C contains canonically conjugated variables, and the constraints are of *second class*. This case, also contemplated in Refs. 7 and 8, will be studied in Sec. IV B.

It should be noted that the same program can be carried out considering the Lie algebras $\tilde{\mathcal{G}}$ of $\tilde{G}$ and $\mathcal{T}$ of T, when these are simply connected groups. In this case, the treatment becomes simpler, since the representations $dU(\tilde{\mathcal{G}})$ and $dD(\mathcal{T})$ are easier to obtain. In general, however, the treatment is more involved, not only because the good operators can lie in the enveloping algebra, but also because the constraints themselves can be defined through higher-order differential equations.[15] But all these cases can be handled with a direct generalization of AQG.

Thus, AQG can be applied to constrained systems, irrespective of the type (first or second class) of constraints. Some examples of application of AQG can be found in Ref. 7, where parity in a two-particle system was introduced to obtain both bosonic and fermionic quantizations, and diffeomorphism constraints to obtain the bosonic string.

Other interesting examples for applying AQG are those systems in which the configuration or phase spaces are multiply connected and the group $\tilde{G}$ of quantum symmetries of the simply connected counterpart (universal covering) is known. If P is a multiply connected phase space which is homogeneous under a group G of symmetries, then P is locally diffeomorphic to a coadjoint orbit of G, or to a coadjoint orbit of a central extension of G by U(1) or R, $\tilde{G}$.[16] For the first case, if H is the isotropy group of P, G/H is locally diffeomorphic to P. If we choose G appropriately (taking coverings, if necessary) in such a way that G/H is simply connected, then P is the quotient of G/H by $\pi_1(P)$, the first homotopy group of P. For the cases in which P is locally diffeomorphic to a coadjoint orbit of a central extension $\tilde{G}$ of G, and if $\tilde{G}$ is chosen (taking coverings) in such a way that this orbit is simply connected, then P is the quotient of $\tilde{G}/H$ by $\pi_1(P)\times U(1)$ (or R). Then $C=\pi_1(P)$ and $T=C\times U(1)$.

However, if P is not the cotangent bundle of any configuration space (as, for instance, the sphere or the torus as symplectic manifolds), then it could well happen that $\pi_1(P)$, as a subset of $\tilde{G}$ (we should not forget that all operations of taking quotients are done in $\tilde{G}$, and therefore we must consider the embeding of $\pi_1(P)$ in $\tilde{G}$, and this could not be a group), contains canonically conjugated pairs. In this case, T is a central extension of C by U(1) and the constraints are of second class. However, if the representation D of T is finite dimensional (see the Appendix), even though T defines second-class constraints, the treatment follows as though they were first-class, yet non-Abelian.

III. THE SCHRÖDINGER GROUP

As mentioned in Sec. I, we shall replace the HW group used in Ref. 8 with the Schrödinger group, which coincides with the Weyl-Symplectic group W $Sp(2,R)$ in one dimension. It was first studied by Niederer[17] as the maximal kinematical invariance group of the Schrödinger equation with general quadratic potential. The complete classification of its unitary irreducible representations was given in Ref. 18. Mathematically it can be obtained from the Galilei (or from the Newton) group by replacing the time parameter with the three-parameter group $SL(2,R)$. The interest of the $SL(2,R)$ group in the present work lies in the fact that it constitutes the maximal finite subgroup of the diffeomorphisms (in fact symplectomorphisms) group of the phase space R^2 (see Ref. 19, where some physical meaning is given to the representations considered "unphysical" in Ref. 18).

To perform a global-coordinate treatment of the problem, we shall start by considering matrices $S \in GL(2,R)$ instead of $SL(2,R)$, and the condition for these matrices to belong to $SL(2,R)$ will appear naturally. A group law for the Schrödinger group can be written as:

$$
\begin{aligned}
\vec{x}'' &= \vec{x}' + \frac{S'}{|S'|^{1/2}}\vec{x}, \\
S'' &= S'S, \\
\zeta'' &= \zeta'\zeta \exp\frac{im\omega}{2\hbar}\left[\frac{-A'x_2'x_1 - B'x_2'x_2 + C'x_1'x_1 + D'x_1'x_2}{|S'|^{1/2}}\right],
\end{aligned} \tag{5}
$$

where

$$
\vec{x} = (x_1, x_2) \in R^2, \quad S = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in GL(2,R), \quad |S| \equiv AD - BC
$$

and $m\omega/\hbar$. is an adimensional constant parametrizing the central extensions of the HW group (we write it in this form for later convenience). The factor $|S'|^{-1/2}$ in the semidirect action of $GL(2,R)$ is needed in order to have a proper central extension.

Let us quantize this system using GAQ, whose principal ingredients will be introduced as needed (see Ref. 7 for details). From the group law, the left-invariant vector fields associated with the coordinates $x_1, x_2, A, B, C, D, \zeta$,

$$
\begin{aligned}
\tilde{X}^L_{x_1} &= |S|^{-1/2}\left[A\frac{\partial}{\partial x_1} + C\frac{\partial}{\partial x_2} + \frac{m\omega}{2\hbar}(-Ax_2 + Cx_1)\Xi\right], \\
\tilde{X}^L_{x_2} &= |S|^{-1/2}\left[D\frac{\partial}{\partial x_2} + B\frac{\partial}{\partial x_1} + \frac{m\omega}{2\hbar}(-Bx_2 + Dx_1)\Xi\right], \\
\tilde{X}^L_A &= A\frac{\partial}{\partial A} + C\frac{\partial}{\partial C}, \quad \tilde{X}^L_B = A\frac{\partial}{\partial B} + C\frac{\partial}{\partial D}, \quad \tilde{X}^L_C = B\frac{\partial}{\partial A} + D\frac{\partial}{\partial C}, \\
\tilde{X}^L_D &= B\frac{\partial}{\partial B} + D\frac{\partial}{\partial D}, \quad \tilde{X}^L_\zeta = i\zeta\frac{\partial}{\partial \zeta} \equiv \Xi,
\end{aligned} \tag{6}
$$

as well as the right-invariants ones,

$$
\begin{aligned}
&\tilde{X}^R_{x_1}=\frac{\partial}{\partial x_1}+\frac{m\omega}{2\hbar}x_2\Xi, \quad \tilde{X}^R_{x_2}=\frac{\partial}{\partial x_2}-\frac{m\omega}{2\hbar}x_1\Xi,\\
&\tilde{X}^R_A=A\frac{\partial}{\partial A}+B\frac{\partial}{\partial B}+\frac{1}{2}x_1\frac{\partial}{\partial x_1}-\frac{1}{2}x_2\frac{\partial}{\partial x_2},\\
&\tilde{X}^R_B=C\frac{\partial}{\partial A}+D\frac{\partial}{\partial B}+x_2\frac{\partial}{\partial x_1},\\
&\tilde{X}^R_C=A\frac{\partial}{\partial C}+B\frac{\partial}{\partial D}+x_1\frac{\partial}{\partial x_2},\\
&\tilde{X}^R_D=C\frac{\partial}{\partial C}+D\frac{\partial}{\partial D}-\frac{1}{2}x_1\frac{\partial}{\partial x_1}+\frac{1}{2}x_2\frac{\partial}{\partial x_2},\\
&\tilde{X}^R_\zeta=\Xi,
\end{aligned}
\tag{7}
$$

can be obtained. The commutation relations for the (left) Lie algebra are:

$$
\begin{aligned}
&[\tilde{X}^L_A,\tilde{X}^L_B]=\tilde{X}^L_B, && [\tilde{X}^L_A,\tilde{X}^L_{x_2}]=-\tfrac{1}{2}\tilde{X}^L_{x_2},\\
&[\tilde{X}^L_A,\tilde{X}^L_C]=-\tilde{X}^L_C, && [\tilde{X}^L_B,\tilde{X}^L_{x_1}]=0,\\
&[\tilde{X}^L_A,\tilde{X}^L_D]=0, && [\tilde{X}^L_B,\tilde{X}^L_{x_2}]=\tilde{X}^L_{x_1},\\
&[\tilde{X}^L_B,\tilde{X}^L_C]=\tilde{X}^L_A-\tilde{X}^L_D, && [\tilde{X}^L_C,\tilde{X}^L_{x_1}]=\tilde{X}^L_{x_1},\\
&[\tilde{X}^L_B,\tilde{X}^L_D]=\tilde{X}^L_B, && [\tilde{X}^L_C,\tilde{X}^L_{x_2}]=0,\\
&[\tilde{X}^L_C,\tilde{X}^L_D]=-\tilde{X}^L_C, && [\tilde{X}^L_D,\tilde{X}^L_{x_1}]=-\frac{1}{2}\tilde{X}^L_{x_1},\\
&[\tilde{X}^L_{x_1},\tilde{X}^L_{x_2}]=\frac{m\omega}{\hbar}\Xi, && [\tilde{X}^L_D,\tilde{X}^L_{x_2}]=\tfrac{1}{2}\tilde{X}^L_{x_2}.\\
&[\tilde{X}^L_A,\tilde{X}^L_{x_1}]=\tfrac{1}{2}\tilde{X}^L_{x_1}, &&
\end{aligned}
\tag{8}
$$

From these commutation relations we see that two linear combinations of vector fields can be introduced, $\tilde{X}^L_A-\tilde{X}^L_D$ and $\tilde{X}^L_A+\tilde{X}^L_D$ (the same for the right-invariant vector fields), in such a way that $\tilde{X}^L_A+\tilde{X}^L_D$ is a central generator, which is also horizontal (see below), and therefore is a gauge generator (see Ref. 14). In fact, it coincides with its right version, as is always the case for a central generator.

We define the Quantization 1-form Θ as the vertical component (dual to the vertical generator, Ξ, in this basis) of the canonical 1-form of the Lie algebra:

$$
\Theta=\frac{m\omega}{2\hbar}(x_2dx_1-x_1dx_2)+\frac{d\zeta}{i\zeta}. \tag{9}
$$

The 2-form $d\Theta$ defines a presymplectic form on $\tilde{G}$, and its value at the identity, $\Sigma = d\Theta|_e$, is a 2-co-cycle of the Lie-algebra, and it can be used to characterize the central extension (when the group $\tilde{G}$ is simply connected). A subalgebra is said to be *horizontal* if it lies in the kernel of Θ. The *characteristic subalgebra* is defined as $\mathcal{G}_\Theta = \mathrm{Ker}\,\Theta \cap \mathrm{Ker}\,d\Theta$, and in this case it has the form:

$$\mathcal{G}_\Theta = \langle \tilde{X}_A^L + \tilde{X}_D^L, \tilde{X}_A^L - \tilde{X}_D^L, \tilde{X}_B^L, \tilde{X}_C^L \rangle. \tag{10}$$

Note that $d\Theta/(\mathrm{Ker}\,d\Theta)$ defines a true symplectic form in R^2.

We define the representation $\mathrm{U}(\tilde{G})$ of $\tilde{G}$ to be given by the left regular representation on complex wave functions over $\tilde{G}$, satisfying the U(1)-function condition $\Xi\Psi = i\Psi$ (phase invariance of quantum mechanics). This representation is obviously reducible, and additional restrictions should be imposed on the wave functions in order to obtain an irreducible representation. These are accomplished by the polarization $\mathcal{P}$, defined as a maximal horizontal left subalgebra of $\tilde{\mathcal{G}}$. The condition $X^L\Psi = 0, \forall X^L \in \mathcal{P}$ leads, in most of the cases, to an irreducible representation $\mathrm{U}(\tilde{G})$ acting on the Hilbert space $\mathcal{H}$ of complex polarized functions on the group satisfying the U(1)-function condition.

However, there are groups, called anomalous (Ref. 11), for which this representation $\mathrm{U}(\tilde{G})$ so obtained is not irreducible, and a generalization of the concept of polarization is required for them. This task is acomplished by means of higher-order polarizations (see Refs. 11, 20, and 21), which admit elements of the left enveloping algebra to enter into them.

The system we are studying is an example of an anomalous system (see Refs. 11 and 21), and a higher-order polarization is required to obtain an irreducible representation. There are essentially two of them,[2] given by

$$\mathcal{P}^{\mathrm{HO}} = \Big\langle \tilde{X}_A^L + \tilde{X}_D^L, \tilde{X}_A^L - \tilde{X}_D^L - \frac{i\hbar}{2m\omega}(\tilde{X}_{x_1}^L \tilde{X}_{x_2}^L + \tilde{X}_{x_2}^L \tilde{X}_{x_1}^L), \tilde{X}_B^L + \frac{i\hbar}{2m\omega}(\tilde{X}_{x_1}^L)^2,$$
$$\tilde{X}_C^L - \frac{i\hbar}{2m\omega}(\tilde{X}_{x_2}^L)^2, \tilde{X}_{x_1}^L \quad \text{or} \quad \tilde{X}_{x_2}^L \Big\rangle. \tag{11}$$

(There are another two, if we allow for complex coordinates, but all of them lead to equivalent representations.) If we choose, for instance, $\tilde{X}_{x_1}^L$ to be in the polarization, the polarization equations are

$$(\tilde{X}_A^L + \tilde{X}_D^L)\Psi = 0, \quad \tilde{X}_B^L\Psi = 0,$$
$$(\tilde{X}_A^L - \tilde{X}_D^L)\Psi = -\tfrac{1}{2}\Psi, \tag{12}$$
$$\tilde{X}_{x_1}^L\Psi = 0, \quad \tilde{X}_C^L\Psi = \frac{i\hbar}{2m\omega}(\tilde{X}_{x_2}^L)^2\Psi.$$

The first of these equations has as solutions those complex wave functions on the group $\mathrm{GL}(2,R)$ which are defined on $\mathrm{SL}(2,R)$, as expected. Therefore, the solutions of this equation have the form:

$$\Psi = \Psi(a,b,c,d,x_1,x_2), \tag{13}$$

where

$$a \equiv \frac{A}{\sqrt{AD-BC}}, \quad b \equiv \frac{B}{\sqrt{AD-BC}}, \quad c \equiv \frac{C}{\sqrt{AD-BC}}, \quad \text{and } d \equiv \frac{D}{\sqrt{AD-BC}}, \text{ with } ad-bc=1.$$

To proceed further in solving the polarization equations, it is convenient to introduce local charts on SL(2,R). We choose them to be the ones defined by $a\neq 0$ and $c\neq 0$, respectively. (Certainly they really correspond to four contractible charts: $a>0$, $a<0$ and $c<0$, $c>0$, but the transition functions between each pair of these charts are trivial, so we shall consider them as only one chart.) The first chart contains the identity element I_2 of SL(2,R), and the second contains

$$J\equiv\begin{pmatrix} \mathbf{0} & 1 \\ -1 & 0 \end{pmatrix}.$$

The solutions to the polarization equations are given by:
For $a\neq 0$:

$$\Psi=\zeta a^{-1/2}\exp\left(\frac{im\omega}{2\hbar}xy\right)\chi(\tau,y), \tag{14}$$

where $x\equiv x_1$, $y\equiv x_2-\tau x_1$, and $\tau\equiv c/a$, with χ satisfying the Schrödinger-like equation

$$\frac{\partial\chi}{\partial\tau}=\frac{i\hbar}{2m\omega}\frac{\partial^2\chi}{\partial y^2}. \tag{15}$$

For $c\neq 0$:

$$\tilde{\Psi}=\zeta c^{-1/2}\exp\left(-\frac{im\omega}{2\hbar}\tilde{x}\tilde{y}\right)\tilde{\chi}(\tilde{\tau},\tilde{y}), \tag{16}$$

where $\tilde{x}\equiv x_2$, $\tilde{y}\equiv x_1-\tilde{\tau}x_2$, and $\tilde{\tau}\equiv a/c$, with $\tilde{\chi}$ satisfying the Schrödinger-like equation

$$\frac{\partial\tilde{\chi}}{\partial\tilde{\tau}}=-\frac{i\hbar}{2m\omega}\frac{\partial^2\tilde{\chi}}{\partial\tilde{y}^2}. \tag{17}$$

The element J represents a rotation of $\pi/2$ in the plane (x_1,x_2), and takes the wave function from one local chart to the other. (In fact, up to a factor, J represents the Fourier transform passing from the x_1 representation to the x_2 representation.) Obviously, $J^4=I_2$, but acting with J on the wave functions we obtain:

$$\Psi(J*g)=(-1)^{1/4}\tilde{\Psi}(g), \tag{18}$$

from which the result $\Psi(J^4*g)=-\Psi(g)$ follows, that is, the representation obtained for the subgroup SL(2,R) is two-valued. This representation is the well-known *metaplectic* or *spinor representation*. The metaplectic representation is for SL(2,R) as the $\frac{1}{2}$-spin representation is for SO(3) (see Ref. 22 and references therein, and also Ref. 16). We refer the reader to Ref. 19 for a detailed study of the Schrödinger group, including the nonanomalous representations and a physical interpretation for them.

IV. THE SCHRÖDINGER GROUP ON THE TORUS

Once we have obtained the polarized wave functions and therefore fixed the unitary and irreducible representation U($\tilde{G}$) of $\tilde{G}$ and the unconstrained Hilbert space $\mathcal{H}$, we have to impose the appropriate constraints to reduce the phase space to a torus. This task is achieved by the structure group T, which is a fiber bundle with base $\Gamma_{\vec{L}}\equiv\{e_{\vec{k}},\vec{k}\in Z\times Z\}$ and fiber U(1), where $e_{\vec{k}}$ are translations of $\vec{x}$ by an amount of $\vec{L}_{\vec{k}}\equiv(k_1L_1,k_2L_2)$, in such a way that $\tilde{G}/T$ is essentially the torus. [As was commented before, $\tilde{G}/T$ in this case is a presymplectic manifold, which, once the kernel of the (pre)symplectic form $d\Theta$ (containing the $SL(2,R)$ subalgebra) is removed, turns out to be a torus.] The fibration of T by U(1) depends on the values of m, ω, L_1, and L_2, and is, in general, nontrivial.

The following task is to obtain the irreducible representations of T. These are studied in detail in the Appendix, and here we shall report only the main results. The form of the representations of T depends strongly on its structure as U(1) bundle with base $\Gamma_{\vec{L}}$ (which plays the role of constraints C), and this is determined by the character of the adimensional parameter $m\omega L_1L_2/2\pi\hbar$, in such a way that:

(i) Integer Case: $m\omega L_1L_2/2\pi\hbar = n\in Z$. In this case T is Abelian, $T=\Gamma_{\vec{L}}\times U(1)$, and therefore all its representations are one dimensional.

(ii) Fractional Case: $m\omega L_1L_2/2\pi\hbar = n/r$, where n and r are relative prime integers (with $r>1$). In this case T is not Abelian, but its representations are of finite dimension.

(iii) Irrational Case: $m\omega L_1L_2/2\pi\hbar=\rho$, where ρ is an irrational number. In this case, T is not Abelian and possesses representations (the ones which are compatible with the U(1)-function condition) of infinite dimension.

The irrational case will not be considered here, since its study requires techniques from noncommutative geometry,[12] and therefore lies beyond the scope of the present work. The most interesting properties of this case, in particular, of the group C^*-algebra generated by the elements of T, denoted *irrational rotation algebra*, is that it is not a type I algebra[23] (in fact it is a type II_∞ algebra).

Although normally the integer and fractional cases are used in physical applications, like Abelian Chern–Simons theory (see Sec. V) or the quantum Hall effect (integer and fractional), there exists works[24] where the irrational case has been use to study the quantum Hall effect, using techniques of C^*-algebras and cyclic cohomology to explain the integrality of the conductance on the quantum Hall effect (see also Ref. 12).

A. The integer case

We shall consider first the integer case, for which $m\omega L_1L_2/2\pi\hbar = n\in Z$ and the structure group is $T=\Gamma_{\vec{L}}\times U(1)$, $\Gamma_{\vec{L}}$ being a subgroup isomorphic to $Z\times Z$. This case leads to a symplectic form on the torus of integer cohomology class n (and therefore the torus is quantizable according to Geometric Quantization), and n can be interpreted as the Chern number of a U(1)-bundle over the torus (see Ref. 8).

The representations of T [compatible with the U(1)-function condition] for the integer case are easily computed (see the Appendix), and have the form:

$$D^{\vec{\varphi}}(I_2,k_1L_1,k_2L_2,\zeta)=\zeta e^{i(\varphi_1k_1+\varphi_2k_2)}e^{-i\pi nk_1k_2}, \tag{19}$$

where $\varphi_1,\varphi_2\in[0,2\pi)$ parametrize the inequivalent representations of the subgroup $\Gamma_{\vec{L}}\approx Z\times Z$. They are the analog of vacuum angles in quantum chromodynamics (see, e.g., Ref. 25).

The T-function conditions are written as $\Psi^{\vec{\varphi}}(g_T*g)=D^{\vec{\varphi}}(g_T)\Psi^{\vec{\varphi}}(g)$, $\forall g_T\in T$. They can be interpreted as periodic boundary conditions, selecting those wave functions in $\mathcal{H}$ which are quasiperiodic, i.e., picking up a phase $e^{i\varphi_1}$ when translated by L_1 and $e^{i\varphi_2}$ when translated by L_2. This condition reduces to $\Psi^0(g_T*g)=\zeta\Psi^0(g)$ if the trivial representation for $\Gamma_{\vec{L}}$ is chosen (strictly periodic boundary conditions). As in Ref. 8, the rest of nonequivalent representations can be obtained by acting with those finite translations which are not good operators. We are not interested in their explicit form, so we refer the interested reader to Ref. 8 for the details of the computations.

The solutions to the T-function condition for the trivial representation are those functions Ψ of the form (14) for which $\chi(\tau,y)$ is of the form:

$$\chi^0(\tau,y)=\sum_{k=0}^{n-1} a_k\Delta_k^0(\tau,y), \tag{20}$$

with

$$\Delta_k^0(\tau,y)=\exp\left[i2\pi k\left(\frac{y}{L_2}-\frac{\delta}{2n}\tau k\right)\right]\sum_{q\in Z}\exp\left(i2\pi nq\left[\frac{y}{L_2}-\frac{\delta}{2n}\tau(k^2+nq)\right]\right), \tag{21}$$

$\delta\equiv L_1/L_2$ and a_k, $k=0,1,...,n-1$ being arbitrary coefficients. We can write these in a form which resembles the one obtained in Ref. 8, where we considered only the Heisenberg–Weyl subgroup:

$$\Delta_k^0(\tau,y)=\exp\left(i\frac{\delta L_2^2\tau}{4\pi n}\frac{d^2}{dy^2}\right)\Delta_k^0(y), \tag{22}$$

with $\Delta_k^0(y)=e^{i2\pi k(y/L_2)}\sum_{q\in Z}\delta(y-(q/n)L_2)$.

For the local chart at $J(c\neq 0)$, we could follow the same procedure or simply transform the wave function acting with J. The result obtained is completely analogous to the one obtained in the local chart at the identity. Therefore, the constrained Hilbert space $\mathcal{H}_T$ is finite dimensional, with a basis of n independent functions, $\{\Delta_k^0\}_{k=0}^{n-1}$.

Now we have to compute the *good operators*, those preserving the Hilbert space $\mathcal{H}_T$ of polarized wave functions verifying the T-function condition. We should look for good operators in the normalizer of T in $\tilde{G}$. In this case (this result is also valid for the fractional case), we have:

$$N_T=\left\{\left(\begin{pmatrix} a & b\\ c & d\end{pmatrix},x_1,x_2,\zeta\right)\in\tilde{G}\quad\text{such that } a,b\delta^{-1},c\delta,d\in Z,x_1,x_2\in R,\zeta\in \mathrm{U}(1)\right\}, \tag{23}$$

which implies that N_T is the semidirect product of SL(2,Z) by the HW group.

Since T is Abelian, the characterization (2) reduces to (3), and this leads to the condition:

$$(a-1)\frac{\varphi_1}{2\pi}+c\delta\frac{\varphi_2}{2\pi}+n\left(-a\frac{x_2}{L_2}+c\delta\frac{x_1}{L_1}-\frac{1}{2}ac\delta\right)=k\in Z,$$

$$(d-1)\frac{\varphi_2}{2\pi}+b\delta^{-1}\frac{\varphi_1}{2\pi}+n\left(-b\delta^{-1}\frac{x_2}{L_2}+d\frac{x_1}{L_1}-\frac{1}{2}db\delta^{-1}\right)=k'\in Z. \tag{24}$$

With regard to the HW subgroup (i.e., with $a=d=1$ and $b=c=0$), we get the same result as in Ref. 8: $x_1=k_1(L_1/n)$ and $x_2=k_2(L_2/n)$, with $k_1,k_2\in Z$. This implies that

$$W\equiv\{\zeta(\hat{\eta}_1)^{k_1/n}(\hat{\eta}_2)^{k_2/n},k_1,k_2\in Z,\zeta\in \mathrm{U}(1)\}\subset G_{\text{good}}^{\vec{\varphi}}, \tag{25}$$

with $\hat{\eta}_1\equiv e_{(1,0)}$ and $\hat{\eta}_2\equiv e_{(0,1)}$, for any values of the vacuum angles φ_1 and φ_2. These operators can be interpreted as the Wilson loops in a Chern–Simons theory on the torus (see Sec. V and Refs. 25 and 26).

When studying the SL(2,R) subgroup (i.e., with $x_1=x_2=0$), we can proceed in two ways. Either we can determine for which values of φ_1 and φ_2 we obtain the full modular group SL(2,Z) as good operators, or we can compute $G_{\text{good}}^{\vec{\varphi}}$ for given values of φ_1 and φ_2.

In the first case, from (24) we easily deduce that modular invariance is achieved for $\varphi_1=2\pi m_1$, $\varphi_2=2\pi m_2$ if n is even and for $\varphi_1=\pi(2m_1+1)$, $\varphi_2=\pi(2m_2+1)$ if n is odd, with $m_1,m_2\in Z$. Clearly, since the vacuum angles are defined modulo 2π these correspond to $\varphi_1=\varphi_2=0$, or periodic boundary conditions for n even and to $\varphi_1=\varphi_2=\pi$, or antiperiodic boundary conditions for n odd. This is an interesting result, since it reflects the fact that good operators really depend on the particular representation $D^{\vec{\varphi}}$ of T we are considering.

The group G_{good} of good operators for these cases would be obtained by taking the product of elements of SL(2,Z) with those of W given by (25). But from (24) we see that there are a few more good operators which cannot be obtained in this way. Altogether, we obtain the following group of good operators for $\varphi_1=\varphi_2=0$ with n even and for $\varphi_1=\varphi_2=\pi$ with n odd:

$$G_{\rm good}=\left\{\left(S,\frac{1}{n}SJ^m\vec{L}_{\vec{k}},\zeta\right)\ \text{such that}\ S\in \mathrm{SL}(2,Z),m=0,1,2,3,\vec{k}\in Z\times Z,\zeta\in \mathrm{U}(1)\right\}. \qquad (26)$$

The computation of $G_{\rm good}^{\vec{\varphi}}$ for arbitrary values of φ_1,φ_2 is a bit more involved. We have seen that the subgroup W given in (25) is always included in $G_{\rm good}^{\vec{\varphi}}$, so we have only to consider the SL(2,Z) subgroup. It is easy to see that if both $\varphi_1/2\pi$ and $\varphi_2/2\pi$ are irrational, then only the identity matrix in SL(2,Z) is a good operator, so there is no hint of modular invariance for this case. If $\varphi_1/2\pi$ is irrational and $\varphi_2/2\pi=p/q$ is rational (the case obtained by interchanging 1 and 2 is analogous), then only the subgroup of modular transformations of the form

$$\begin{pmatrix} 1 & \epsilon q\delta^{-1}k \\ 0 & 1 \end{pmatrix}$$

are good operators, with $k\in Z$ and with $\epsilon=1$ for n even and $\epsilon=2$ for n odd. If $\varphi_2/2\pi=p_1/q_1$ and $\varphi_2/2\pi=p_2/q_2$ are rational, then the good operators are given by the subgroup of modular transformations satisfying the following diophantine equations:

$$\begin{aligned}(a-1)\frac{p_1}{q_1}+c\delta\frac{p_2}{q_2}-n\frac{ac\delta}{2}&=k\in Z,\\ b\delta^{-1}\frac{p_1}{q_1}+(d-1)\frac{p_2}{q_2}-n\frac{db\delta^{-1}}{2}&=k'\in Z.\end{aligned} \qquad (27)$$

B. The fractional case

For the fractional case, we shall restrict ourselves to the determination of the subgroup of good operators. The computation of the explicit form of the constrained wave functions can be performed along the guidelines of Sec. IV A (they are essentially the ones given in Ref. 8 for the HW group), using the representations of T given in the Appendix. The dimension of the Hilbert space turns out to be nr, and it can be considered to be an n dimensional Hilbert space made of vector-valued wave functions, r being the dimension of the vector space.

To determine the subgroup $G_{\rm good}$ of good operators, we make use of the characterization (2) for the little group, where now, since the representations are of dimension r, the equivalence can be established through a nontrivial unitary matrix $V(g_g)$.

First, we compute, for

$$g_g=\left(\begin{pmatrix} a & b \\ c & d\end{pmatrix},x_1,x_2,\zeta'\right)\in N_T,$$

$$\begin{aligned}g_g*(I_2,k_1L_1,k_2L_2,\zeta)*g_g^{-1}=&(I_2,(ak_1+b\delta^{-1}k_2)L_1,(c\delta k_1+dk_2)\\ &\times L_2,\zeta e^{i2\pi(n/r)[(-a(x_2/L_2)+c\delta(x_1/L_1))k_1+(-b\delta^{-1}(x_2/L_2)+d(x_1/L_1))k_2]}),\end{aligned} \qquad (28)$$

and then we must find for which $g_g\in N_T$ we have

$$D^{\vec{\varphi}}(g_g*g_T*g_g^{-1})=V(g_g)D^{\vec{\varphi}}(g_T)V(g_g)^\dagger,\quad \forall g_T\in T, \qquad (29)$$

where the representations $D^{\vec{\varphi}}$ for the fractional case (obtained in the Appendix), are given by

$$D^{\vec{\varphi}}(I_2,k_1L_1,k_2L_2,\zeta)=\zeta e^{i(\varphi_1k_1+\varphi_2k_2)}e^{-i\pi(n/r)k_1k_2}A_r^{k_1}B_r^{k_2}. \qquad (30)$$

We proceed as in the integer case, computing first the good operators in the HW subgroup. Then the previous equation is written:

$$\exp\left[i2\pi\frac{n}{r}\left(-\frac{x_2}{L_2}k_1+\frac{x_1}{L_1}k_2\right)\right]A_r^{k_1}B_r^{k_2}=V(g_g)A_r^{k_1}B_r^{k_2}V(g_g)^{\dagger}. \tag{31}$$

This equation is the same one which states the equivalence of the representations $D^{(\mu_1,\mu_2)}$ and $D^{(0,0)}$ and, therefore, making use of the results given in the Appendix, we find that $x_1 = k(L_1/n)$ and $x_2=k'(L_2/n)$, with $k,k'\in Z$. This implies that the subgroup W given in (25) is included in $G_{\rm good}^{\vec{\varphi}}$ for all values of $\varphi_1,\varphi_2\in[0,2\pi/r)$.

As far as the SL$(2,Z)$ subgroup is concerned, we shall determine only the conditions under which full modular invariance is obtained as good operators, and for this purpose we shall make use of the fact that SL$(2,Z)$ is generated by two modular transformations:

$$g_1\equiv\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix},\quad g_2\equiv\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}. \tag{32}$$

Determining under which conditions these two transformations are good operators will tell us when the theory is fully modular invariant. For g_1 we obtain the condition:

$$\exp\left[i2\pi k_2\left(\frac{\varphi_1}{2\pi}-\frac{nk_2}{2r}\right)\right]A_r^{k_1+k_2}B_r^{k_2}=V(g_1)A_r^{k_1}B_r^{k_2}V(g_1)^{\dagger},\quad \forall k_1,k_2\in Z. \tag{33}$$

For this condition to hold, it is necessary that $\varphi_1=0$ if nr is even, or $\varphi_1=\pi/r$ if nr is odd. For the first case, the unitary matrix $V(g_1)$ has the form $V(g_1)_{ij}=\omega_r^{(i-1)^2/2}\delta_{ij}$, and, for the second, we have $V(g_1)_{ij}=\omega_r^{(i-1)/2n}\omega_r^{(i-1)^2/2}\delta_{ij}$.

For g_2 to be a good operator, we obtain the condition:

$$\exp\left[i2\pi k_1\left(\frac{\varphi_2}{2\pi}-\frac{nk_1}{2r}\right)\right]A_r^{k_1}B_r^{k_1+k_2}=V(g_2)A_r^{k_1}B_r^{k_2}V(g_2)^{\dagger},\quad \forall k_1,k_2\in Z. \tag{34}$$

Again, for this condition to hold it has to be $\varphi_2=0$ if nr is even or $\varphi_2=\pi/r$ if nr is odd. The unitary matrix $V(g_2)$ has the form:

$$V(g_2)=V=\frac{1}{\sqrt{r}}\begin{pmatrix}1 & \omega_r^{(r-1)(r-2)/2} & \cdots & \cdots & 1\\ 1 & 1 & \cdots & \cdots & \\ \omega_r & 1 & \cdots & \cdots & \\ \omega_r^3 & \omega_r & \cdots & \cdots & \vdots\\ \vdots & \vdots & \vdots & \vdots & \omega_r^{(r-1)(r-2)/2}\\ \omega_r^{(r-1)(r-2)/2} & \cdots & \cdots & 1 & 1\end{pmatrix}\quad \text{if } nr \text{ is even}$$

$$V(g_2)=A_r^{1/2n}V\quad \text{if } nr \text{ is odd},$$

where $(A_r^{1/2n})_{ij}=e^{i\pi[(i-1)/r]}\delta_{ij}$.

It should be stressed that the values of $\vec{\varphi}$ for which full modular invariance is obtained correspond to wave functions which are periodic if nr is even, or antiperiodic if nr is odd, where these boundary conditions should be understood with respect to translations by rL_1 and rL_2.

Note also that the matrix representation $V(g_1)$ and $V(g_2)$ obtained for g_1 and g_2 [and therefore for the whole SL$(2,Z)$ group] corresponds to their action on the r-dimensional vector space.

The complete action of any modular transformation on the wave functions (through $nr\times nr$ matrices) decomposes, thus, in a tensor product of an $n\times n$ matrix and an $r\times r$ matrix, each one acting on different indices of the wave functions.[26]

This structure of tensor product of the Hilbert space suggests a duality under the interchange of n and r. Indeed, the set of Wilson loops (25) for the theories characterized by n/r and r/n are isomorphic. Since all the information of the theory is contained in the Wilson loops, we could say that the two theories are equivalent. The case $n/r=1$ would, of course, be self-dual. Moreover, as pointed out in Ref. 27, if we denote by $\mathcal{A}_{n/r}$ the (group) algebra generated by A and B satisfying

$$AB=e^{i2\pi(n/r)}BA, \tag{35}$$

then we have $\mathcal{A}_{1/(nr)}=\mathcal{A}_{n/r}\times\mathcal{A}_{r/n}$. Therefore, the algebra of Wilson loops, besides being the same for a theory with $T=\mathcal{A}_{n/r}$ and $T=\mathcal{A}_{r/n}$, is given by the direct product of both (commuting) algebras. From the point of view of noncommutative C^*-algebras, the algebras $\mathcal{A}_{n/r}$ and $\mathcal{A}_{r/n}$ are strongly Morita equivalent, which means, in particular, that they possess the same representation theory[23] (see also Ref. 12).

V. 2+1D ABELIAN CHERN–SIMONS THEORY

As a first application of our results, let us consider a pure topological field theory on the torus.

Let M be a globally hyperbolic three-dimensional manifold, $M=\Sigma\times R$, where Σ is an orientable two-dimentional manifold.

The action for an Abelian Chern–Simons (ACS) theory is given by[28,25,26]

$$S_{\rm ACS}=\frac{k}{4\pi}\int_M (A\wedge dA), \tag{36}$$

where A is a one-form in M which takes values on the Lie algebra $\mathcal{K}$ of an Abelian Lie group K. It is straightforward to check that the action $S_{\rm ACS}$ is invariant under gauge transformations $A\rightarrow A+iU^{-1}dU$ for any (single-valued) $U:M\rightarrow K$.

The equations of motion are:

$$dA\equiv F=0, \tag{37}$$

the solution of which is the vector space $\mathcal{V}_{\rm ACS}$ of all flat connections on M. A generic element $A\in\mathcal{V}_{\rm ACS}$ can be written in the form $(A_0,iU^{-1}\nabla U+a(t))$, where a is a map from R to the fiber of $T^*(\Sigma)\otimes\mathcal{K}$.

This vector space of solutions can be endowed with a (pre-)symplectic structure by means of a (pre-)symplectic form

$$\Omega_{\rm ACS}(A',A)=\int_\Sigma J=\frac{k}{4\pi}\int_\Sigma A'\wedge A, \tag{38}$$

where $J^\mu\equiv(k/4\pi)\epsilon^{\mu\nu\sigma}A'_\nu A_\sigma$ is a divergenceless current which ensures the independence of $\Omega_{\rm ACS}(A',A)$ on the chosen Cauchy factorization of M, $M=\Sigma\times R$.

Since the exterior derivative d commutes with the pullback operator $*$, if f is a diffeomorphism of M, and A', $A\in\mathcal{V}_{\rm ACS}$, then $A'+f^*A$ is also a solution of (37).

With this information, we can propose a quantizing group $\tilde{G}_{\rm ACS}$ for this theory, the composition law of which is

$$f''=f'\circ f,\quad f,f',f''\in{\rm Diff}(M),$$
$$A''=f^{-1*}A'+A, \tag{39}$$

$$\zeta''=\zeta\zeta' \exp \Omega_{\mathrm{ACS}}(f^{-1*}A',A),$$

i.e., the extension by U(1) of the semidirect product $\mathcal{V}_{\mathrm{ACS}}\otimes_s \mathrm{Diff}(M)$. The characteristic subgroup (generated by the kernel of Ω_{ACS}, see Sec. III) of this group proves to be $G_\Omega=\{(f,A,1)/A=(A_0,iU^{-1}\nabla U)$ for some $U:M\rightarrow K\}\subset\tilde{G}_{\mathrm{ACS}}$, which contains the *gauge group* G_{gauge} of the theory, constituted by all (single-valued) $U:M\rightarrow K$. [To be precise, here, G_{gauge} is the orbit at the identity of Map(M,K) on $\mathcal{V}_{\mathrm{ACS}}$, under the action $A\rightarrow A+iU^{-1}dU$. Including the group Map(M,K) explicitly in $\tilde{G}_{\mathrm{ACS}}$ requires a slight modification of the notion of gauge transformation[29] (see also Ref. 14 for a discussion on the definition of *gauge group*).] Thus, the polarization conditions (which contain the characteristic subgroup) imply that wave functions depend only on topological and gauge invariant quantities. For this kind of theory, standard approaches claim that all gauge-invariant information of a connection can be extracted from the *Wilson loops* defined by

$$W(A,\gamma)=\exp\int_\gamma A, \tag{40}$$

for any loop γ in Σ. Since connections A are flat, the Wilson loops will depend only on the homotopy class $[\gamma]\in\pi_1(\Sigma)$ of the corresponding loop γ. For this reason the normal subgroup $\mathrm{Diff}_0(M)\subset\mathrm{Diff}(M)$ of diffeomorphism of M connected to the identity acts trivially on the Wilson loops. Therefore the diffeomorphisms that really matter in (39) are the quotient $\mathrm{Diff}(M)/\mathrm{Diff}_0(M)$ called the *modular group* (see Refs. 9 and 10) of the Riemann surface Σ (note that all diffeomorphisms of the R part of M are connected to the identity).

It should be stressed that if $\pi_1(\Sigma)=0$ [which implies that $H^1(\Sigma)=0$] then the ACS theory is trivial since all connections are of the form $A=iU^{-1}dU$ for some (always single-valued) $U:M\rightarrow K$. This implies that $G_\Omega=G_{\mathrm{ACS}}\equiv\mathcal{V}_{\mathrm{ACS}}\otimes_s\mathrm{Diff}(M)$ and $\tilde{G}_{\mathrm{ACS}}=G_{\mathrm{ACS}}\times \mathrm{U}(1)$, that is, the central extension is trivial (another way of seeing this is that since the homotopy group of M is trivial, all Wilson loops are trivial). Therefore we assume that Σ is a multiply connected oriented two-dimensional manifold. What makes the theory nontrivial in this case is the fact that the gauge group, G_{gauge}, is smaller than its simply connected counterpart $\overline{G}_{\mathrm{gauge}}$ in the universal covering space $\overline{M}$ of M, constituted by all $\overline{U}:\overline{M}\rightarrow K$, with $\overline{M}=\overline{\Sigma}\times R$, and $\overline{\Sigma}$ the universal covering space of Σ. In fact, the group $G_{\mathrm{gauge}}\subset\overline{G}_{\mathrm{gauge}}$ is made of those elements $\overline{U}\in\overline{G}_{\mathrm{gauge}}$ verifying $\overline{U}\circ[\gamma]=\overline{U}$, where here $[\gamma]$ represents the natural action (as diffeomorphism) of the homotopy class $[\gamma]\in\pi_1(M)$ on $\overline{M}$.

For the present case, $\Sigma=S^1\times S^1$ and $K=\mathrm{U}(1)$. The space $\mathcal{V}_{\mathrm{ACS}}$ is made of connections of the form $(A_0,ig^{-1}\nabla g+a(t))$, where U is single-valued on the torus. The solution manifold, that which remains once the quotient by the characteristic subgroup G_Ω is taken, is parametrized by the variables $a_1(t)$, $a_2(t)$ modulo an integer, defining a torus. The reason is that G_Ω also contains the *global (large) gauge transformations* (see for instance Ref. 26 and references therein),

$$a_j\rightarrow a_j+k_j,\quad k_j\in Z. \tag{41}$$

These large gauge transformations are clearly seen to come from transformations of the form $U=\exp(ik_jx^j)$, with $k_j\in Z$ and $\{x^j\}$ a set of local coordinates on the torus, in such a way that 0 and 2π are identified. The reason of the restriction of k_j to integers is the condition $\overline{U}\circ[\gamma]=\overline{U}$. This indicates that the gauge group G_{gauge} is a disconnected group, with $G_{\mathrm{gauge}}/G^0_{\mathrm{gauge}}=Z\times Z$, G^0_{gauge} being the connected component of the identity.

For this reason, in the quantum theory the operator associated with the variables A (more precisely, a) are not properly defined (they are *bad operators*, see Sec. II). We must resort to single-valued (*good*) operators of the form:

$$W(A)=\exp\left(2\pi i\sum_{j=1}^{2} n_j a_j\right), \tag{42}$$

where $n_j \in Z$ should be interpreted as the winding number of a path γ around the cycle j. Remember that for the torus, the homotopy classes $[\gamma]$ are generated by two elements, $[\gamma_j]$, $j=1,2$, representing loops (with winding number one) around each one of the two cycles of the torus. The modular group proves to be $\mathrm{Diff}(T^2)/\mathrm{Diff}_0(T^2)=\mathrm{SL}(2,Z)$.

At this point it should be stressed that the resulting theory corresponds to a quantum mechanical system with phase space a torus parametrized by $(q,p)\equiv(a_1 \bmod 1, a_2 \bmod 1)$.

According to this equivalence, we could have studied this system in the framework of AQG by starting with the group $\mathrm{HW}\otimes_s\mathrm{Diff}(R^2\times R)$ with structure group T a fiber bundle with base $Z\times Z$ and fiber U(1), where $Z\times Z$ is the subgroup of $\mathrm{Diff}(R^2\times R)$ of translations by (k_1L_1,k_2L_2), with $k_1,k_2\in Z$. Since the only relevant diffeomorphisms at the final theory on the torus will be the modular transformations $\mathrm{SL}(2,Z)\subset\mathrm{SL}(2,R)$, it is enough to start with $\mathrm{HW}\otimes_s\mathrm{SL}(2,R)$, which is the Schrödinger group. Thus, all the results of Sec. IV apply here. Of course, we could have started directly with $\mathrm{HW}\otimes_s\mathrm{SL}(2,Z)$, but this group, being disconnected, is more difficult to quantize than the Schrödinger group (in particular, finding a polarization for this group is a difficult task). In addition, we think that showing how SL(2,Z) emerges as good operators is a very illustrative way of studying the problem.

In summary

(1) The coupling constant k plays the same role as the quantity $\phi\equiv m\omega L_1L_2/2\pi\hbar$ in the Schrödinger group on the torus, determining the character of the resulting (finite-dimensional) Hilbert space.
(2) The set of Wilson loops (42) takes part of the set of good operators in our language. More precisely, they are the analog of the set W given in (25).
(3) The group of large gauge transformations is the analog of the structure group T. When the coupling constant k is fractional, this gauge group is called *anomalous*[26] because of its non-Abelian character due to the nontrivial fibration for this case, as oposed to the original Abelian gauge group K.
(4) The nonequivalent representations of T, parametrized by the indices φ_1,φ_2 (vacuum angles), characterize the nonequivalent quantizations of the theory.

The Chern–Simons theory constitutes a particular example of a drastic reduction of the number of original infinite (field) degrees of freedom to a finite number (which, in addition, contain a finite number of states, due to the compactness of the phase space when restricted to the torus), as a consequence of a huge gauge invariance which kills all of them except for the topologic ones.

A. Further comments

Comparing our results with those in the literature, we find full agreement with Ref. 25, in the context of U(1) Chern–Simon theory on the torus, as far as the integer case is concerned. For the fractional case, an apparent discrepance with the results in Ref. 25 appears: In our notation, modular invariance is obtained only for n even (and any value of r) and vacuum angles $\varphi_1=\varphi_2=0$. However, the agreement is achieved if we realize that the proper range of inequivalent vacuum angles in Ref. 25 should be $[0,2\pi/r)$.

This problem was also studied in Ref. 26 [also in the context of U(1) Chern–Simons theory and anyons on the torus], where full modular invariance was obtained for both the integer and fractional case, but they claimed that the vacuum angles always have to be $\frac{1}{2}$ disregarding the parity of the coupling constant (the equivalent of our n/r). A more detailed analysis of their results reveals that the vacuum angles they introduce are defined modulo $1/nr$, and $(\frac{1}{2} \bmod 1/nr)$ is 0 for even nr and $1/2nr$ for odd nr, corresponding to periodic and antiperiodic boundary conditions, respectively. Therefore, their results completely agree with ours.

In Ref. 27, a non-Abelian Chern–Simons theory is considered, with gauge group $SL(2,R)$. When restricted to the torus, they obtained essentially the same results as ours and those of Refs. 25 and 26 (with the above-mentioned remarks) with respect to the Hilbert space and the set of observables (good operators), because the reduced phase space of the theory is almost the space of flat connections of an Abelian gauge group.

ACKNOWLEDGMENTS

J.G. thanks the Spanish MEC for a Postdoctoral grant and the Department of Physics of Naples-INFN for its hospitality and financial support, and M.C. thanks the Spanish MEC for a FPI grant. Work partially supported by the DGICYT.

APPENDIX: UNITARY AND IRREDUCIBLE REPRESENTATIONS OF *T*

The structure subgroup *T*, as defined in Sec. IV, is a U(1) bundle with base $\Gamma_L^{\vec{}}$, and can be written as

$$T=\{(I_2,k_1L_1,k_2L_2,\zeta)\in\tilde{G}/k_1,k_2\in Z,\zeta\in U(1)\}, \tag{A1}$$

with group law derived from the group law of the Schrödinger group:

$$(I_2,k_1'L_1,k_2'L_2,\zeta')*(I_2,k_1L_1,k_2L_2,\zeta)=\left(I_2,(k_1'+k_1)L_1,(k_2'+k_2)L_2,\zeta'\zeta \times\exp\left[i\frac{m\omega L_1L_2}{2\hbar}(k_1'k_2-k_2'k_1)\right]\right). \tag{A2}$$

To determine the structure of *T*, we compute the group commutator of two elements:

$$[(I_2,k_1'L_1,k_2'L_2,\zeta'),(I_2,k_1L_1,k_2L_2,\zeta)]=\left(I_2,0,0,\exp\left[i\frac{m\omega L_1L_2}{\hbar}(k_1'k_2-k_2'k_1)\right]\right), \tag{A3}$$

from which we see that its structure depends on the value of $m\omega L_1L_2/2\pi\hbar$, in such a way that there are three possibilities:

(i) Integer Case: $m\omega L_1L_2/2\pi\hbar=n\in Z$.
(ii) Fractional Case: $m\omega L_1L_2/2\pi\hbar=n/r$, $n,r\in Z$ and relative prime (with $r>1$).
(iii) Irrational Case: $m\omega L_1L_2/2\pi\hbar=\rho$, with ρ an irrational number.

Let us study the integer and fractional case separately. The irrational case will not be considered here (see Ref. 12 for a detailed study of this case).

1. Integer case

In this case, *T* is an Abelian group, and therefore $T=\Gamma_L^{\vec{}}\times U(1)$ and all its representations are of dimension 1. As stated above, we shall consider only those representations, which restricted to U(1), are the natural representations, and these have the form:

$$D^{\vec{\varphi}}(I_2,k_1L_1,k_2L_2,\zeta)=\zeta e^{i(\varphi_1k_1+\varphi_2k_2)}e^{-i\pi nk_1k_2},\quad \forall k_1,k_2\in Z,\quad \forall\zeta\in U(1), \tag{A4}$$

where the range of inequivalent representations, since they are one-dimensional, is given simply by $\varphi_1,\varphi_2\in[0,2\pi)$. Note that, except for the term $e^{-i\pi nk_1k_2}$ this is the product of the natural representation of U(1) times a representation of $\Gamma_L^{\vec{}}\approx Z\times Z$. This extra term is only a coboundary coming from the fact that we have used Bargmann's cocycle in the group law of the Schrödinger group, and Bargmann's cocycle does not satisfy the conditions given in Ref. 8 for the possible cocycles for the HW group on the torus. Note, thus, that this restriction can be relaxed by introducing this coboundary term in the representations of *T*.

2. Fractional case

In this case, T is not Abelian, and the commutator of two elements has the form:

$$[(I_2,k_1'L_1,k_2'L_2,\zeta'),(I_2,k_1L_1,k_2L_2,\zeta)]=(I_2,0,0,\omega_r^{k_1'k_2-k_2'k_1}), \tag{A5}$$

where $\omega_r\equiv e^{i2\pi(n/r)}$ is an rth root of unity. Note that if $|n|>r$, then $\omega_r=e^{i2\pi(n/r)}=e^{i2\pi(q/r)}$, where $q=n \bmod r$. Since n and r are relative prime, q and r turn out also to be relative prime and, therefore, we can use either of the two pairs to characterize T.

The group T admits a nontrivial characteristic subgroup (see Ref. 8), of the form:

$$G_C=\{(I_2,rk_1L_1,rk_2L_2,e^{i\pi nrk_1k_2})/k_1,k_2\in Z\}. \tag{A6}$$

The characteristic subgroup can be identified in this case with the Casimir elements of T which are not in U(1), i.e., those elements of T [not belonging to U(1)] which commute with all other elements in T. In fact, the center of T is given by $G_C\times \mathrm{U}(1)$.

If we quotient T by G_C, we obtain a group which is a *generalized Clifford group* G_2^r (see Ref. 30 for the definition and the study of representations of generalized Clifford groups) times U(1). Therefore, the representations of T can be obtained from those of G_C and G_2^r [and the natural representation of U(1)].

The representations of G_C, being isomorphic to $Z\times Z$, are characterized by two ''vacuum angles'' φ_1,φ_2, whose range of nonequivalence should be determined. The representations of G_2^r are studied in detail in Ref. 30, so that here we shall give only the results. It should be remarked, however, that in Ref. 30 ω_r is an arbitrary rth root of unity, and different choices for it give inequivalent representations of G_2^r, whereas here the value of ω_r is given *a priori* (it is determined by the fact that T is a subgroup of $\tilde{G}$), so that the representation of G_2^r is uniquely determined. In addition, since n and r are relative prime, ω_r is a primitive rth root of unity, implying that the representation of G_2^r associated with it is of dimension r, either for prime or nonprime r.

The r-dimensional unitary irreducible representation of G_2^r can be constructed with the aid of two $r\times r$ matrices, A_r and B_r:

$$(A_r)_{ij}=\omega_r^{i-1}\delta_{ij},\quad (B_r)_{ij}=\delta_{i,(j \bmod r)+1},\quad i,j=1,2,...,r, \tag{A7}$$

verifying $A_rB_r=\omega_rB_rA_r$, and $A_r^r=B_r^r=I_r$. Putting together this representation of G_2^r and that of $G_C\approx Z\times Z$, we can build a representation for the entire T, of the form:

$$D^{\vec{\varphi}}(I_2,k_1L_1,k_2L_2,\zeta)=\zeta e^{i(\varphi_1k_1+\varphi_2k_2)}e^{-i\pi(n/r)k_1k_2}A_r^{k_1}B_r^{k_2},\quad \forall k_1,k_2\in Z,\forall\zeta\in \mathrm{U}(1). \tag{A8}$$

One would naïvely expect that the range of nonequivalent representations would be $\varphi_1,\varphi_2 \in[0,2\pi)$, as in the integer case. However, since the representations are not one dimensional, there could be nontrivial unitary transformations relating representations in this interval and, therefore, reducing the range on nonequivalence.

Thus, we have to determine the minimum values of μ_1,μ_2 for which the representation $D^{(\mu_1,\mu_2)}$ is equivalent to the trivial representation $D^{(0,0)}$, i.e., there exists a unitary matrix V such that $D^{(\mu_1,\mu_2)}=VD^{(0,0)}V^\dagger$. Studying separately the cases $(\mu_1,0)$ and $(0,\mu_2)$, and after a few computations, we obtain:

(i) $D^{(\mu_1,0)}$ is equivalent to $D^{(0,0)}$ for $\mu_1=2\pi/r$, with $V=B_r^{m_0}$.
(ii) $D^{(0,\mu_2)}$ is equivalent to $D^{(0,0)}$ for $\mu_2=2\pi/r$, with $V=A_r^{1/n}$.

Here $(A_r^{1/n})_{ij}\equiv\omega_r^{(i-1)/n}\delta_{ij}=e^{i2\pi[(i-1)/r]}\delta_{ij}$, and $0<m_0<r$ is an integer solution of the diophantine equation:

$$1+nm_0=rk,\quad k\in Z, \tag{A9}$$

which has always two solutions in the range $\{-(r-1),\ldots,0,...,r-1\}$, provided n and r are relative prime (this is a particular case of the Bezout lemma, for $gcd(n,r)=1$, which in turn can be proven using Euclidean division of integers, see, for instance, Ref. 31). Note that $(A_r^{1/n})^n = A_r$ and $(B^{m_0})^n = B_r^{-1}$, so that these matrices can be considered as the nth roots of the matrices A_r and B_r^{-1}, respectively.

Therefore, the range of nonequivalent representations of T is reduced to $\varphi_1,\varphi_2\in[0,2\pi/r)$. This fact will be of extreme importance for the determination of the good operators in the fractional case.

[1] P. A. M. Dirac, *Lectures on Quantum Mechanics* (Yeshiva University, New York, 1964).
[2] L. D. Faddeev, Theor. Math. Phys. **1**, 1 (1970).
[3] C. Becchi, A. Rouet, and R. Stora, Phys. Lett. **25B**, 29 (1967).
[4] A. Ashtekar and R. Tate, J. Math. Phys. **35**, 6434 (1994).
[5] J. Marsden and A. Weinstein, Rep. Math. Phys. **5**, 121 (1974).
[6] N. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1980).
[7] V. Aldaya, J. Navarro-Salas, and A. Ramírez, Commun. Math. Phys. **121**, 541 (1989).
[8] V. Aldaya, M. Calixto, and J. Guerrero, Commun. Math. Phys. **178**, 399 (1996).
[9] M. Schlichenmaier, *An Introduction to Riemann Surfaces, Algebraic Curves and Moduli Spaces*, Lecture Notes Physics, Vol. 322 (Springer, Berlin, 1984).
[10] M. Nakahara, *Geometry, Topology and Physics* (IOP, Bristol, 1990).
[11] V. Aldaya, J. Navarro-Salas, J. Bisquert, and R. Loll, J. Math. Phys. **33**, 3087 (1992).
[12] A. Connes, *NonCommutative Geometry* (Academic, New York, 1994).
[13] N. P. Landsman and N. Linden, Nucl. Phys. B **365**, 121 (1991).
[14] M. Navarro, V. Aldaya, and M. Calixto, J. Math. Phys. **38**, 1454 (1997).
[15] V. Aldaya, M. Calixto, and J. M. Cerveró, Commun. Math. Phys. **200**, 325 (1999).
[16] A. A. Kirillov, *Elements of the Theory of Representations* (Springer, Berlin, 1976).
[17] U. Niederer, Helv. Phys. Acta **45**, 802 (1972); **46**, 191 (1973); **47**, 167 (1974).
[18] M. Perroud, Helv. Phys. Acta **50**, 233 (1977).
[19] M. Calixto, V. Aldaya, and J. Guerrero, Int. J. Mod. Phys. A **13**, 4889 (1998).
[20] V. Aldaya and J. Navarro-Salas, Commun. Math. Phys. **139**, 433 (1991).
[21] V. Aldaya, J. Guerrero, and G. Marmo, Int. J. Mod. Phys. A **12**, 3 (1997); math-ph/9811015 (1998).
[22] G. B. Folland, *Harmonic Analysis in Phase Space*, Annals of Mathematical Studies, Vol. 122 (Princeton University Press, Princeton, 1989).
[23] M. A. Rieffel, Pac. J. Math. **93**, 415 (1981).
[24] J. Bellissard, *Ordinary Quantum Hall Effect and Noncommutative Cohomology*, in Proceedings on Localization in Disordered Systems, Bad Schandau (Teubner, Leipzig, 1988).
[25] A. P. Polychronakos, Ann. Phys. (N.Y.) **203**, 231 (1990).
[26] R. Iengo and K. Lechner, Phys. Rep. **213**, 179 (1992).
[27] C. Imbimbo, Nucl. Phys. B **384**, 484 (1992).
[28] E. Witten, Commun. Math. Phys. **121**, 377 (1989).
[29] V. Aldaya, M. Calixto, and M. Navarro, Int. J. Mod. Phys. A **12**, 3609 (1997).
[30] R. Jagannathan and N. R. Ranganathan, Rep. Math. Phys. **5**, 131 (1974); **7**, 229 (1975).
[31] S. Maclane and G. Birkhoff, *Algebra* (Macmillan, New York, 1967).

Free field approach to the dilute A_L models

Yuji Hara[a)]
Institute of Physics, Graduate School of Arts and Sciences, University of Tokyo, Tokyo 153-8902, Japan

Michio Jimbo[b)]
Division of Mathematics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan

Hitoshi Konno[c)]
Department of Mathematics, Faculty of Integrated Arts and Sciences, Hiroshima University, Higashi-Hiroshima 739-8521, Japan

Satoru Odake[d)]
Department of Physics, Faculty of Science, Shinshu University, Matsumoto 390-8621, Japan

Jun'ichi Shiraishi[e)]
Institute for Solid State Physics, University of Tokyo, Tokyo 106-0032, Japan

(Received 8 March 1999; accepted for publication 1 April 1999)

We construct a free field realization of vertex operators of the dilute A_L models along with the Felder complex. For $L=3$, we also study an E_8 structure in terms of the deformed Virasoro currents.
[S0022-2488(99)01308-0]

I. INTRODUCTION

The dilute A_L model[1,2] is an integrable lattice model obtained by an RSOS restriction of the face model of type $A_2^{(2)}$.[3] It possesses several intriguing features. Among others, we are interested in the following points.

(i) At criticality, the dilute A_L model in regime 2^+ is described by conformal field theory (CFT), which belongs to the Virasoro minimal unitary series with the central charge $c=1-6/(L(L+1))$. The Andrews–Baxter–Forrester (ABF) model in regime III is known to be a different off-critical lattice model having the same critical behavior. While the ABF model corresponds to the (1,3)-perturbation of the minimal unitary CFT, the dilute A_L model corresponds to the (1,2)-perturbation of the same CFT.

(ii) In the particular case of $L=3$, the model falls within the same universality class as the two-dimensional Ising model in a magnetic field. The elliptic nome in the Boltzmann weights plays the role of a magnetic field, as opposed to the usual role as a temperaturelike variable. In the field theory limit, the scattering process of particles exhibits an E_8 structure.

In Ref. 4, bosonization of the ABF model in regime III was achieved. By ''bosonization'' we mean a free field realization of vertex operators (VOs) of the model as formulated in Ref. 5. It was also found that the deformed Virasoro algebra (DVA) proposed earlier in Ref. 6 arises naturally, in such a way that the VOs play the role of deformed chiral primary fields for DVA. (More specifically, the VOs of type I and type II correspond to the simplest primary fields ϕ_{21} and ϕ_{12}, respectively. Analogs of general ϕ_{mn} are obtained by a fusion procedure.[7] For an interpretation of the VOs as intertwiners for elliptic algebras, see Ref. 8.) The Becchi–Rouet–Stora–Tyutin

a)Electronic mail: snowy@gokutan.c.u-tokyo.ac.jp
b)Electronic mail: jimbo@kusm.kyoto-u.ac.jp
c)Electronic mail: konno@mis.hiroshima-u.ac.jp
d)Electronic mail: odake@azusa.shinshu-u.ac.jp
e)Electronic mail: shiraish@momo.issp.u-tokyo.ac.jp

(BRST) resolution of Fock spaces, which singles out irreducible representations of the Virasoro algebra, carries over to the deformed version as well. Thus Ref. 4 presents an off-critical lattice version of the (1,3)-perturbation of the minimal unitary CFT in the free field picture. In this paper we study a similar problem for the (1,2)-perturbed CFT by bosonizing the dilute A_L model. For this purpose we adapt the construction of Refs. 5, 4, and 9 to the case of $A_2^{(2)}$.

In the trigonometric limit, bosonization of VOs has been given in Refs. 10 and 11 using representation theory of the quantum affine algebra $U_q(A_2^{(2)})$. In principle, we are to follow its elliptic analog on the basis of the face-type elliptic algebra.[12,8] Because of some technical difficulties in dealing with the latter, we take here a more pedestrian way and solve the exchange relations directly to obtain the bosonic realizations of VOs. In the course, we use the elliptic Drinfeld currents obtained by a ''dressing'' procedure.[8]

In view of points (i) and (ii) above, it is natural to expect that a different deformation of the Virasoro algebra arises from the dilute A_L models. Such a deformation was found in Ref. 13 through bosonization of the $A_2^{(2)}$ affine Toda field theory. (See Refs. 14 and 15 for more general deformed W algebras including the case $A_{2l}^{(2)}$.) To make a distinction from the original DVA of Ref. 6, we use the symbol $\mathcal{V}_{x,r}(A_2^{(2)})$ to denote the DVA of Ref. 13. In the original case, the generating function of DVA (hereafter referred to as the ''DVA current'') can also be obtained from ''fusion'' of the VOs.[16–18] In the same way, we reproduce the current of $\mathcal{V}_{x,r}(A_2^{(2)})$ by taking residues of products of the bosonized VOs in the present case.

For the dilute A_L models in regime 2^+, the space of states of the corner transfer matrix is an analog of the minimal unitary representation[2] for $\mathcal{V}_{x,r}(A_2^{(2)})$. In order to obtain them from the Fock spaces, we consider a Felder-type resolution using the elliptic currents of type $A_2^{(2)}$ as screening currents. Unlike the case of the ABF models, the BRST charges are not simply a power of screening operators. Such a complication seems to be common in the higher rank situations.[19] We prove the nilpotency of BRST charges with the help of the Feigin–Odesskii algebra.[19] Assuming a cohomological property of the resulting complex, we write down integral expressions for the two-point local height probabilities (LHPs) and traces of general product of the VOs.

In the case of the dilute A_3 model, it is of some interest to see how Zamolodchikov's E_8 structure of scattering process[20] looks like that in the free field picture. What plays the role of the Zamolodchikov–Faddeev operator for creation/annihilation of bound states is the DVA current.[16] Specializing to $L=3$, we introduce eight kinds of DVA currents by suitably fusing the elementary current of $\mathcal{V}_{x,r}(A_2^{(2)})$. We then find a curious similarity between the operator product expansions of these currents and the so-called T-system of E_8 type for the transfer matrix.[21,22] At present we do not understand its proper meaning.

This paper is organized as follows. In Sec. II, we review the definition of the dilute A_L model and give a brief description of the vertex operator approach. In Sec. III, we give bosonization of the VOs. We derive the current for $\mathcal{V}_{x,r}(A_2^{(2)})$ from them and state a conjecture for the Kac determinant formula for $\mathcal{V}_{x,r}(A_2^{(2)})$. We also present a Felder-type BRST complex of the Fock spaces. Section IV is devoted to an application of the bosonization to the calculation of the LHP. In Sec. V, regarding the $\mathcal{V}_{x,r}(A_2^{(2)})$ current as the ZF operator for the particles in the A_3 model, we discuss a similarity between the T-system and the ''bootstrap'' of the ZF operators. Appendix A is a summary of the operator product expansion formulas for the elliptic currents and VOs. In Appendix B, we prove the nilpotency of the BRST charges. Appendix C is devoted to an exposition of the fusion properties of the deformed W currents for $A_{N-1}^{(1)}$, which should be compared with those for the DVA currents associated with the dilute A_3 model in Sec. V.

II. THE DILUTE A_L MODELS

A. Boltzmann weights

Throughout this paper we fix a positive integer $L\geqslant 3$. In the dilute A_L model, the local fluctuation variables $a,b,\ldots,$ take one of the L states $1,2,\ldots,L$, and those on neighboring lattice sites are subject to the condition $a-b=0,\pm 1$. The Boltzmann weights can be found in Ref. 2, Eq.

(3.1). For our purpose it is convenient to use the parametrization given in Appendix A of Ref. 2, which is suitable in the ''low-temperature'' regime. With some change of notation we recall the formula below.

Let $x=e^{-2\pi\lambda/\epsilon}$, $r=\pi/(2\lambda)$, and $u=-u_{\mathrm{orig}}/(2\lambda)$, where λ,ϵ are the variables used in Ref. 2 and u_{orig} stands for ''u'' there. We shall restrict ourselves to the ''regime 2^+'' defined by

$$0<x<1, \quad r=2\frac{L+1}{L+2}, \quad -\frac{3}{2}<u<0. \tag{2.1}$$

Along with the variable u, we often use the multiplicative variable

$$z=x^{2u}.$$

Changing an overall scalar factor we put the Boltzmann weights in the form

$$W\left(\begin{matrix} a & b \\ c & d \end{matrix}\middle| u\right)=\rho(u)\overline{W}\left(\begin{matrix} a & b \\ c & d \end{matrix}\middle| u\right),$$

where $\rho(u)$ will be specified in (2.4) below. To give the formula for the $\overline{W}$ factors, let us set

$$[u]=x^{u^2/r-u}\Theta_{x^{2r}}(x^{2u}), \quad [u]_+=x^{u^2/r-u}\Theta_{x^{2r}}(-x^{2u}), \tag{2.2}$$

where

$$\Theta_p(z)=(z;p)_\infty(pz^{-1};p)_\infty(p;p)_\infty,$$

$$(z;p_1,\ldots,p_k)_\infty=\prod_{n_1,\ldots,n_k=0}^{\infty}(1-p_1^{n_1}\cdots p_k^{n_k}z).$$

Then we have

$$\overline{W}\left(\begin{matrix} a\pm1 & a \\ a & a\mp1 \end{matrix}\middle| u\right)=1,$$

$$\overline{W}\left(\begin{matrix} a & a\pm1 \\ a & a\pm1 \end{matrix}\middle| u\right)=\overline{W}\left(\begin{matrix} a\pm1 & a\pm1 \\ a & a \end{matrix}\middle| u\right)=-\left(\frac{[\pm a+3/2]_+[\pm a-1/2]_+}{[\pm a+1/2]_+^2}\right)^{1/2}\frac{[u]}{[1+u]},$$

$$\overline{W}\left(\begin{matrix} a\pm1 & a \\ a & a \end{matrix}\middle| u\right)=\overline{W}\left(\begin{matrix} a & a \\ a & a\pm1 \end{matrix}\middle| u\right)=\frac{[\pm a+1/2+u]_+}{[\pm a+1/2]_+}\frac{[1]}{[1+u]},$$

$$\overline{W}\left(\begin{matrix} a & a\mp1 \\ a\pm1 & a \end{matrix}\middle| u\right)=(G_a^+G_a^-)^{1/2}\frac{[1/2+u]}{[3/2+u]}\frac{[u]}{[1+u]},$$

$$\overline{W}\left(\begin{matrix} a & a \\ a\pm1 & a \end{matrix}\middle| u\right)=\overline{W}\left(\begin{matrix} a & a\pm1 \\ a & a \end{matrix}\middle| u\right)=-(G_a^\pm)^{1/2}\frac{[\pm a-1-u]_+}{[\pm a+1/2]_+}\frac{[1]}{[1+u]}\frac{[u]}{[3/2+u]},$$

$$\overline{W}\left(\begin{matrix} a & a\pm1 \\ a\pm1 & a \end{matrix}\middle| u\right)=\frac{[\pm 2a+1-u]}{[\pm 2a+1]}\frac{[1]}{[1+u]}-G_a^\pm\frac{[\pm 2a-1/2-u]}{[\pm 2a+1]}\frac{[u]}{[3/2+u]}\frac{[1]}{[1+u]},$$

$$\overline{W}\left(\begin{matrix} a & a \\ a & a \end{matrix}\middle| u\right)=\frac{[3+u]}{[3]}\frac{[1]}{[1+u]}\frac{[3/2-u]}{[3/2+u]}+H_a\frac{[1]}{[3]}\frac{[u]}{[1+u]}.$$

Here

$$G_a^{\pm}=\frac{S(a\pm 1)}{S(a)},\quad S(a)=(-1)^a\frac{[2a]}{[a]_+},\quad H_a=G_a^{+}\frac{[a-5/2]_+}{[a+1/2]_+}+G_a^{-}\frac{[a+5/2]_+}{[a-1/2]_+}. \tag{2.3}$$

We choose $\rho(u)$ so that the partition function per site of the model equals 1. Explicitly it is given by[2]

$$z^{(r-1)/r}\rho(u)=\frac{\rho_+(u)}{\rho_+(-u)},\quad \rho_+(u)=\frac{(x^2z,x^3z,x^{2r+3}z,x^{2r+4}z;x^6,x^{2r})_\infty}{(x^5z,x^6z,x^{2r}z,x^{2r+1}z;x^6,x^{2r})_\infty}, \tag{2.4}$$

where $z=x^{2u}$, and we have introduced the notation

$$(a_1,\ldots,a_n;p_1,\ldots,p_k)_\infty=\prod_{j=1}^{n}(a_j;p_1,\ldots,p_k)_\infty. \tag{2.5}$$

Graphically we represent the Boltzmann weights as follows:

$$W\begin{pmatrix}a & b\\ c & d\end{pmatrix}\!\!\left|u_1-u_2\right) =$$

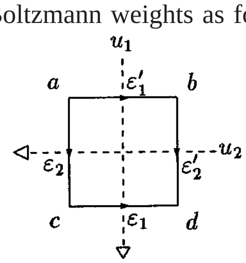

$$b=a+\varepsilon_1',\quad c=a+\varepsilon_2,\quad d=b+\varepsilon_2'=c+\varepsilon_1.$$

For definiteness we list below the basic properties of the Boltzmann weights:
The Yang–Baxter equation

$$\sum_g W\left(\begin{matrix}a & b\\ g & c\end{matrix}\middle| u\right)W\left(\begin{matrix}a & g\\ f & e\end{matrix}\middle| v\right)W\left(\begin{matrix}g & c\\ e & d\end{matrix}\middle| u+v\right)=\sum_g W\left(\begin{matrix}a & b\\ f & g\end{matrix}\middle| u+v\right)W\left(\begin{matrix}b & c\\ g & d\end{matrix}\middle| v\right)W\left(\begin{matrix}f & g\\ e & d\end{matrix}\middle| u\right);$$

unitarity

$$\sum_g W\left(\begin{matrix}a & b\\ g & c\end{matrix}\middle| u\right)W\left(\begin{matrix}a & g\\ d & c\end{matrix}\middle| -u\right)=\delta_{bd};$$

crossing symmetry

$$W\left(\begin{matrix}b & d\\ a & c\end{matrix}\middle| -\frac{3}{2}-u\right)=\sqrt{\frac{S(a)S(d)}{S(b)S(c)}}\,W\left(\begin{matrix}a & b\\ c & d\end{matrix}\middle| u\right);$$

and initial condition

$$W\left(\begin{matrix}a & b\\ c & d\end{matrix}\middle| 0\right)=\delta_{bc}.$$

B. Vertex operators

Hereafter we assume that L is odd. The model has ground states labeled by odd integers $l=1,3,\ldots,L-2$.[2] They are characterized as configurations in which all heights take the same value b. If $L=4m\pm 1$, then the possible values are $b=l\,(1\leqslant l\leqslant 2m-1, l\text{: odd})$ or $b=l+1\,(2m+1\leqslant l\leqslant L-2, l\text{: odd})$.

Consider the corner transfer matrices $A(z),B(z),C(z),D(z)$ corresponding to the NW, SW, SE, and NE quadrants, respectively. In the infinite volume limit, we have

$$C(z)=A(z)=z^{-\mathcal{H}},\quad B(z)=D(z)=\sqrt{S(k)}x^{3\mathcal{H}}z^{\mathcal{H}},$$

with k denoting the value of the central height. The operator $\mathcal{H}$ (the corner Hamiltonian) is independent of z. We denote by $\mathcal{L}_{l,k}$ the space of eigenstates of $\mathcal{H}$ in the sector where the central height is fixed to k and the boundary heights are in the ground state l. It was found in Ref. 2 that the generating function of the spectrum of $\mathcal{H}$ coincides with the character of the Virasoro minimal unitary series. Namely,

$$\mathrm{tr}_{\mathcal{L}_{l,k}}(q^{\mathcal{H}})=\chi_{l,k}(q), \tag{2.6}$$

where

$$\chi_{l,k}(q)=\frac{q^{\Delta_{l,k}-c/24}}{(q;q)_\infty}\sum_{j\in\mathbb{Z}}(q^{L(L+1)j^2+((L+1)l-Lk)j}-q^{L(L+1)j^2+((L+1)l+Lk)j+lk}),$$

$$c=1-\frac{6}{L(L+1)},\quad \Delta_{l,k}=\frac{((L+1)l-Lk)^2-1}{4L(L+1)}.$$

Consider next the half-infinite transfer matrix extending to infinity in the north. We denote it by

$$\Phi^{(k,k+\epsilon)}(z):\mathcal{L}_{l,k+\epsilon}\to\mathcal{L}_{l,k}\quad(\epsilon=0,\pm1).$$

Likewise we denote by

$$\Phi^{*(k+\epsilon,k)}(z^{-1}):\mathcal{L}_{l,k}\to\mathcal{L}_{l,k+\epsilon}\quad(\epsilon=0,\pm1)$$

the half-infinite transfer matrix extending to infinity in the west. We shall also write

$$\Phi^{(k,k+\epsilon)}(z)=\Phi_\epsilon(z),\quad \Phi^{*(k+\epsilon,k)}(z)=\Phi^*_\epsilon(z),$$

and call them vertex operators (VOs) of type I.

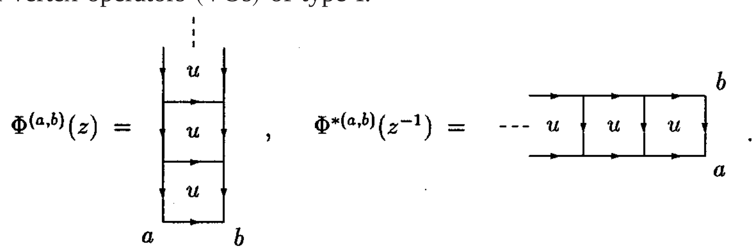

Intuitive graphical arguments based on the properties of the Boltzmann weights lead to the following formulas. For details we refer the reader to Refs. 23 and 4,

$$\Phi^{(a,c)}(z_2)\Phi^{(c,d)}(z_1)=\sum_g W\left(\begin{matrix}a & g\\ c & d\end{matrix}\middle| u_1-u_2\right)\Phi^{(a,g)}(z_1)\Phi^{(g,d)}(z_2)\quad(z_j=x^{2u_j}),$$

$$w^{\mathcal{H}}\Phi^{(a,b)}(z)w^{-\mathcal{H}}=\Phi^{(a,b)}(wz),\quad \Phi^{*(b,a)}(z)=\sqrt{\frac{S(a)}{S(b)}}\Phi^{(b,a)}(x^{-3}z),$$

$$\sum_g \Phi^{*(a,g)}(z)\Phi^{(g,a)}(z)=1,\quad \Phi^{(a,b)}(z)\Phi^{*(b,c)}(z)=\delta_{ac}.$$

As explained in Ref. 5, multipoint local height probabilities are expressed as traces of VOs. Consider neighboring $n+1$ lattice sites in a row. Let $P_l(a_0,\ldots,a_n)$ denote the probability of finding these local variables to be $(a_0,\ldots,a_n)$. Then we have

$$P_l(a_0,\ldots,a_n)=Z_l^{-1}S(a_n)\mathrm{tr}_{\mathcal{L}_{l,a_n}}(x^{6\mathcal{H}}\Phi^{*(a_n,a_{n-1})}(z)\cdots\Phi^{*(a_1,a_0)}(z)\Phi^{(a_0,a_1)}(z)\cdots\Phi^{(a_{n-1},a_n)}(z)). \tag{2.7}$$

Here the nomalization factor Z_l is

$$Z_l=\sum_{k=1}^{L} S(k)\chi_{l,k}(x^6),$$

which can be expressed in product of theta functions with conjugate modulus.[2] In the simplest case $n=0$, the one-point function $P_l(k)$ is given by $P_l(k)=Z_l^{-1}S(k)\chi_{l,k}(x^6)$.

Remark: We follow mostly the notation of Ref. 9 but there are minor changes. The $\Phi(\zeta^{-1})$ in Ref. 9 corresponds to $\Phi(z)$ in the present notation. We have also reversed the orientation of edges of the Boltzmann weights.

III. BOSONIZATION OF VERTEX OPERATORS

A. Bosons

In this section we present a bosonic realization of vertex operators. The working closely follows Refs. 4 and 9. We set

$$[n]_x=\frac{x^n-x^{-n}}{x-x^{-1}},$$

and introduce the oscillators α_n $(n\neq 0)$ and P,Q satisfying the commutation relations

$$\begin{gathered}[\alpha_n,\alpha_m]=\frac{[n]_x([2n]_x-[n]_x)}{n}\frac{[rn]_x}{[(r-1)n]_x}\delta_{n+m,0},\\ [P,iQ]=1.\end{gathered} \tag{3.1}$$

Notice that $[2n]_x-[n]_x=[3n]_x[n/2]_x/[3n/2]_x$. We shall also use

$$\alpha_n'=(-1)^n\frac{[(r-1)n]_x}{[rn]_x}\alpha_n.$$

We denote by

$$\mathcal{F}_{l,k}=\mathbb{C}[\alpha_{-1},\alpha_{-2},\ldots]|l,k\rangle$$

the Fock space generated by

$$|l,k\rangle=e^{p_{l,k}iQ}|0,0\rangle,\quad P|l,k\rangle=p_{l,k}|l,k\rangle,$$

where $p_{l,k}$ is

$$p_{l,k}=-\frac{l}{2}\sqrt{\frac{r}{r-1}}+k\sqrt{\frac{r-1}{r}}=-l\sqrt{\frac{L+1}{2L}}+k\sqrt{\frac{L}{2(L+1)}}. \tag{3.2}$$

[Recall that $r=2(L+1)/(L+2)$.] These Fock spaces are graded by

$$d=\sum_{n=1}^{\infty}\frac{n^2}{[n]_x([2n]_x-[n]_x)}\frac{[(r-1)n]_x}{[rn]_x}\alpha_{-n}\alpha_n+\frac{1}{2}P^2-\frac{1}{24},$$

which satisfies $[d,\alpha_n]=-n\alpha_n$, $[d,iQ]=P$ and $d|l,k\rangle=(\Delta_{l,k}-c/24)|l,k\rangle$. For later use, we define the operators $\hat{l},\hat{k}:\mathcal{F}_{l,k}\to\mathcal{F}_{l,k}$ by

$$\hat{l}|_{\mathcal{F}_{l,k}}=l\times\mathrm{id}_{\mathcal{F}_{l,k}},\quad \hat{k}|_{\mathcal{F}_{l,k}}=k\times\mathrm{id}_{\mathcal{F}_{l,k}}.$$

B. Vertex operators

In Refs. 10 and 11, a bosonic realization of the level-one representation of the quantum affine algebra $U_q(A_2^{(2)})$ and associated vertex operators have been obtained. We shall consider their elliptic counterparts.

The elliptic version of the Drinfeld currents are constructed from the trigonometric ones by a ‘‘dressing’’ procedure described in Ref. 8. Applying it to the present case of $U_q(A_2^{(2)})$, we obtain

$$x_+(z):\mathcal{F}_{l,k}\to\mathcal{F}_{l-2,k},\quad x_-(z):\mathcal{F}_{l,k}\to\mathcal{F}_{l,k-1},$$

$$x_+(z)=:\exp\left(-\sum_{n\neq 0}\frac{\alpha_n}{[n]_x}z^{-n}\right):\times e^{\left(\sqrt{\frac{r}{r-1}}iQz^{\sqrt{\frac{r}{r-1}}P+\frac{r}{2(r-1)}}\right)},\tag{3.3}$$

$$x_-(z)=:\exp\left(\sum_{n\neq 0}\frac{\alpha'_n}{[n]_x}z^{-n}\right):\times e^{\left(-\sqrt{\frac{r-1}{r}}iQz^{-\sqrt{\frac{r-1}{r}}P+\frac{r-1}{2r}}\right)}.\tag{3.4}$$

The elliptic version of VOs (of type I and type II) are defined in terms of their trigonometric ones and a ‘‘twistor’’ given by an infinite product of the universal R matrix.[24] They satisfy the commutation relations of the type (3.11)–(3.13) in Sec. III C. As we do not know how to evaluate the twistor in the bosonic realization, we have solved relations (3.11)–(3.13) directly for $\Phi_\epsilon(z),\Psi^*_\epsilon(z)$. We obtain the following.

Type I:

$$\Phi_\epsilon(z):\mathcal{F}_{l,k}\to\mathcal{F}_{l,k-\epsilon},$$

$$\Phi_-(z)=:\exp\left(-\sum_{n\neq 0}\frac{\alpha'_n}{[2n]_x-[n]_x}z^{-n}\right):\times e^{\left(\sqrt{\frac{r-1}{r}}iQz^{\sqrt{\frac{r-1}{r}}P+\frac{r-1}{2r}}\right)},\tag{3.5}$$

$$\Phi_0(z)=x^{(1-r)/2r}\oint_{C_0}\underline{dz_1}\Phi_-(z)x_-(z_1)\frac{1}{\sqrt{[\hat{k}+1/2]_+[\hat{k}-1/2]_+}}\frac{[u-u_1+\hat{k}]_+}{[u-u_1+1/2]},\tag{3.6}$$

$$\Phi_+(z)=x^{(1-r)/r}\oint\oint_{C_+}\underline{dz_1}\underline{dz_2}\Phi_-(z)x_-(z_1)x_-(z_2)$$
$$\times\sqrt{\frac{S(\hat{k}-1)}{S(\hat{k})}}\frac{1}{[\hat{k}-1/2]_+[2\hat{k}-2]}\frac{[u-u_1+2\hat{k}-3/2]}{[u-u_1+1/2]}\frac{[u_1-u_2+\hat{k}]_+}{[u_1-u_2+1/2]}.\tag{3.7}$$

Type II:

$$\Psi^*_\epsilon(z):\mathcal{F}_{l,k}\to\mathcal{F}_{l-2\epsilon,k},$$

$$\Psi_-^*(z) = :\exp\left(\sum_{n\neq 0} \frac{\alpha_n}{[2n]_x - [n]_x} z^{-n}\right): \times e^{\left(-\sqrt{\frac{r}{r-1}} iQz - \sqrt{\frac{r}{r-1}} P + \frac{r}{2(r-1)}\right)}, \tag{3.8}$$

$$\Psi_0^*(z) = ix^{r/2(r-1)} \oint_{C_0^*} \underline{dz_1} \Psi_-^*(z) x_+(z_1) \frac{1}{\sqrt{[(\hat{l}+1)/2]_+^* [(\hat{l}-1)/2]_+^*}} \frac{[u-u_1-\hat{l}/2]_+^*}{[u-u_1-1/2]^*}, \tag{3.9}$$

$$\Psi_+^*(z) = x^{r/(r-1)} \oint \oint_{C_+^*} \underline{dz_1} \underline{dz_2} \Psi_-^*(z) x_+(z_1) x_+(z_2)$$
$$\times \sqrt{\frac{S^*(\hat{l}/2-1)}{S^*(\hat{l}/2)}} \frac{1}{[(\hat{l}-1)/2]_+^* [\hat{l}-2]^*} \frac{[u-u_1-\hat{l}+3/2]^*}{[u-u_1-1/2]^*} \frac{[u_1-u_2-\hat{l}/2]_+^*}{[u_1-u_2-1/2]^*}. \tag{3.10}$$

Here $z = x^{2u}$, $z_j = x^{2u_j}$, $\underline{dz_j} = dz_j/(2\pi i z_j)$ and

$$[u]^* = x^{u^2/(r-1)-u} \Theta_{x^{2r-2}}(x^{2u}), \quad [u]_+^* = x^{u^2/(r-1)-u} \Theta_{x^{2r-2}}(-x^{2u}), \quad S^*(a) = (-1)^a \frac{[2a]^*}{[a]_+^*}.$$

The poles of the integrand of (3.6)–(3.10) and the integration contours are listed in the following table ($n = 0,1,2,...$). For example, C_0 is a simple closed contour that encircles $x^{1+2rn}z$ $(n \geqslant 0)$ but not $x^{-1-2rn}z$ $(n \geqslant 0)$.

	Inside	Outside
C_0	$z_1 = x^{1+2rn}z$	$z_1 = x^{-1-2rn}z$
C_+	$z_1 = x^{1+2rn}z$	$z_1 = x^{-1-2rn}z$
	$z_2 = x^{1+2rn}z_1$	$z_2 = x^{-1-2rn}z, x^{-1-2rn}z_1, x^{2-2r(n+1)}z_1$
C_0^*	$z_1 = x^{-1+2(r-1)n}z$	$z_1 = x^{1-2(r-1)n}z$
C_+^*	$z_1 = x^{-1+2(r-1)n}z$	$z_1 = x^{1-2(r-1)n}z$
	$z_2 = x^{-1+2(r-1)n}z_1$	$z_2 = x^{1-2(r-1)n}z, x^{1-2(r-1)n}z_1, x^{-2-2(r-1)(n+1)}z_1$

C. Commutation relations and inversion identities

The VOs given above satisfy the following commutation relations:

$$\Phi_{\epsilon_2}(z_2)\Phi_{\epsilon_1}(z_1) = \sum_{\substack{\epsilon_1', \epsilon_2' \\ \epsilon_1'+\epsilon_2' = \epsilon_1+\epsilon_2}} W\left(\begin{matrix} \hat{k} & \hat{k}+\epsilon_1' \\ \hat{k}+\epsilon_2 & \hat{k}+\epsilon_1+\epsilon_2 \end{matrix} \middle| u_1 - u_2\right) \Phi_{\epsilon_1'}(z_1)\Phi_{\epsilon_2'}(z_2), \tag{3.11}$$

$$\Psi_{\epsilon_1}^*(z_1)\Psi_{\epsilon_2}^*(z_2) = \sum_{\substack{\epsilon_1', \epsilon_2' \\ \epsilon_1'+\epsilon_2' = \epsilon_1+\epsilon_2}} W^*\left(\begin{matrix} \hat{l}/2 & \hat{l}/2+\epsilon_1 \\ \hat{l}/2+\epsilon_2' & \hat{l}/2+\epsilon_1+\epsilon_2 \end{matrix} \middle| u_1 - u_2\right) \Psi_{\epsilon_2'}^*(z_2)\Psi_{\epsilon_1'}^*(z_1), \tag{3.12}$$

$$\Phi_{\epsilon_2}(z_2)\Psi_{\epsilon_1}^*(z_1) = \tau(u_1-u_2)\Psi_{\epsilon_1}^*(z_1)\Phi_{\epsilon_2}(z_2). \tag{3.13}$$

Here we have set (for $z = x^{2u}$)

$$W^*\left(\begin{matrix} a & b \\ c & d \end{matrix}\middle| u\right) = \overline{W}\left(\begin{matrix} a & b \\ c & d \end{matrix}\middle| u\right)\Bigg|_{r\to r-1} \times \rho^*(u), \tag{3.14}$$

$$z^{-r/(r-1)}\rho^*(u) = \frac{\rho_+^*(u)}{\rho_+^*(-u)}, \quad \rho_+^*(u) = \frac{(x^3z,x^4z,x^{2r}z,x^{2r+1}z;x^6,x^{2r-2})_\infty}{(z,xz,x^{2r+3}z,x^{2r+4}z;x^6,x^{2r-2})_\infty}, \tag{3.15}$$

$$\tau(u) = z\,\frac{\Theta_{x^6}(-xz^{-1})\Theta_{x^6}(-x^2z^{-1})}{\Theta_{x^6}(-xz)\Theta_{x^6}(-x^2z)}. \tag{3.16}$$

Note that

$$\rho^*(u) = -\rho(u)|_{r\to r-1} \times z\,\frac{\Theta_{x^6}(xz^{-1})\Theta_{x^6}(x^2z^{-1})}{\Theta_{x^6}(xz)\Theta_{x^6}(x^2z)}.$$

We do not present the tedious but straightforward verification of (3.11)–(3.13).

For the description of correlation functions we need also the ''dual'' VOs. Define

$$\Phi_\epsilon^*(z) = g\sqrt{S(\hat{k})}^{-1}\Phi_{-\epsilon}(x^{-3}z)\sqrt{S(\hat{k})}, \tag{3.17}$$

$$\Psi_\epsilon(z) = g^{*-1}\sqrt{S^*(\hat{l}/2)}\Psi_{-\epsilon}^*(x^{-3}z)\sqrt{S^*(\hat{l}/2)}^{-1}, \tag{3.18}$$

where

$$g^{-1} = \frac{(x;x^{2r})_\infty}{(x^2;x^{2r})_\infty^2(x^{2r-1};x^{2r})_\infty(x^{2r};x^{2r})_\infty^4}\,\frac{(x^5,x^6,x^{2r},x^{2r+1};x^6,x^{2r})_\infty}{(x^2,x^3,x^{2r+3},x^{2r+4};x^6,x^{2r})_\infty},$$

$$g^* = \frac{(x^{-1};x^{2r-2})_\infty}{(x^{-2};x^{2r-2})_\infty^2(x^{2r-1};x^{2r-2})_\infty(x^{2r-2};x^{2r-2})_\infty^5}\,\frac{(x^3,x^4,x^{2r},x^{2r+1};x^6,x^{2r-2})_\infty}{(x,x^6,x^{2r+3},x^{2r+4};x^6,x^{2r-2})_\infty}.$$

Then we have

$$\Phi_{\epsilon_2}(z)\Phi_{\epsilon_1}^*(z) = \delta_{\epsilon_1,\epsilon_2}\times \mathrm{id}, \quad \Psi_{\epsilon_1}(z_1)\Psi_{\epsilon_2}^*(z_2) = \frac{\delta_{\epsilon_1,\epsilon_2}}{1-z_1/z_2} + \cdots \quad (z_1\to z_2), \tag{3.19}$$

$$\sum_\epsilon \Phi_\epsilon^*(z)\Phi_\epsilon(z) = \mathrm{id}, \quad \sum_\epsilon \Psi_\epsilon^*(z_2)\Psi_\epsilon(z_1) = \frac{1}{1-z_1/z_2} + \cdots \quad (z_1\to z_2). \tag{3.20}$$

D. Deformed Virasoro algebra

Brazhnikov and Lukyanov[13] pointed out that one can associate to the algebra $A_2^{(2)}$ a deformed Virasoro algebra (DVA) which is different from the one found in Ref. 6. The original DVA of Ref. 6, associated with $A_1^{(1)}$, arises also as a ''fusion'' of VOs.[16] Let us discuss this point in the present case of $A_2^{(2)}$.

Let

$$\Lambda_\pm(z) = :\exp\left(\pm\sum_{n\neq 0}\lambda_n(x^{\pm 3/2}z)^{-n}\right):\times x^{\pm 2\sqrt{r(r-1)}P},$$

$$\Lambda_0(z) = -\frac{[r-1/2]_x}{[1/2]_x}:\exp\left(\sum_{n\neq 0}\lambda_n(x^{-n/2}-x^{n/2})z^{-n}\right):, \tag{3.21}$$

$$T(z) = \Lambda_+(z) + \Lambda_0(z) + \Lambda_-(z),$$

where

$$\lambda_n=(-1)^n(x-x^{-1})\frac{[(r-1)n]_x}{[2n]_x-[n]_x}\alpha_n=(x-x^{-1})\frac{[rn]_x}{[2n]_x-[n]_x}\alpha'_n,$$

$$[\lambda_n,\lambda_m]=(x-x^{-1})^2\frac{1}{n}\frac{[n]_x[rn]_x[(r-1)n]_x}{[2n]_x-[n]_x}\delta_{m+n,0}.$$

Then $T(z)$ is obtained from VOs by fusing them,

$$\Phi_{\epsilon_2}(x^{r+3/2}z')\Phi^*_{\epsilon_1}(x^{-r+3/2}z)=\left(1-\frac{z}{z'}\right)(-1)^{\epsilon_1+1}\delta_{\epsilon_1,\epsilon_2}T(z)\cdot x^{1-r}\frac{(x,x^6,x^{5-2r},x^{6-2r};x^6)_\infty}{(x^3,x^4,x^{2-2r},x^{3-2r};x^6)_\infty}+\cdots\quad(z'\to z). \tag{3.22}$$

The $T(z)$ satisfies the DVA of Ref. 13

$$\begin{aligned}&f\left(\frac{z_2}{z_1}\right)T(z_1)T(z_2)-f\left(\frac{z_1}{z_2}\right)T(z_2)T(z_1)\\&=(x-x^{-1})\frac{[r+1/2]_x[r]_x[r-1]_x[r-3/2]_x}{[1/2]_x[3/2]_x}\left(\delta\left(x^3\frac{z_2}{z_1}\right)-\delta\left(x^{-3}\frac{z_2}{z_1}\right)\right)\\&\quad+(x-x^{-1})\frac{[r]_x[r-1/2]_x[r-1]_x}{[1/2]_x}\left(\delta\left(x^2\frac{z_2}{z_1}\right)T(xz_2)-\delta\left(x^{-2}\frac{z_2}{z_1}\right)T(x^{-1}z_2)\right),\end{aligned} \tag{3.23}$$

where $f(z)$ is

$$\begin{aligned}f(z)&=\exp\left(-\sum_{n>0}(x-x^{-1})^2\frac{1}{n}\frac{[n]_x[rn]_x[(r-1)n]_x}{[2n]_x-[n]_x}z^n\right)\\&=\frac{1}{1-z}\frac{(x^{2-2r}z,x^{3-2r}z,x^4z,x^5z,x^{2r}z,x^{2r+1}z;x^6)_\infty}{(x^{5-2r}z,x^{6-2r}z,xz,x^2z,x^{2r+3}z,x^{2r+4}z;x^6)_\infty}.\end{aligned} \tag{3.24}$$

The notation of Ref. 13 is related to ours by $x_{\mathrm{BL}}=x^{3/2}$, $b/Q=r$, $1/(Qb)=1-r$, $g(z)=f(z)$, $\mathbf{V}(z)=T(z)$. In what follows we call this algebra $\mathcal{V}_{x,r}(A_2^{(2)})$. The relation (3.23) is invariant under

$$r\mapsto 1-r,\quad x\mapsto x,\quad T(z)\mapsto -T(z). \tag{3.25}$$

We remark that $\tilde{T}(z)=-\Lambda_+(z)+\Lambda_0(z)-\Lambda_-(z)$ also satisfies (3.23), which is obtained from type II VOs,

$$\Psi_{\epsilon_1}(x^{r-1+3/2}z')\Psi^*_{\epsilon_2}(x^{-(r-1)+3/2}z)=\frac{1}{1-z'/z}(-1)^{\epsilon_1+1}\delta_{\epsilon_1,\epsilon_2}\tilde{T}(-z)\times(-x^{-r})\frac{(x^2,x^3,x^{5-2r},x^{6-2r};x^6)_\infty}{(x^5,x^6,x^{2-2r},x^{3-2r};x^6)_\infty}+\cdots\quad(z'\to z).$$

Let us discuss some features of $\mathcal{V}_{x,r}(A_2^{(2)})$.

Conformal limit: In the conformal limit ($x=e^{\hbar}\to 1,r$: fixed), (3.23) admits two limits[13] related by (3.25),

$$T(z)=3-2r+\hbar^2(8r(r-1)z^2L(z)+\tfrac{1}{6}r(r-1)(1-2r)+(2-r)^2)+O(\hbar^4), \tag{3.26}$$

$$T(z)=-1-2r+\hbar^2(-8r(r-1)z^2\tilde{L}(z)+\tfrac{1}{6}r(r-1)(1-2r)-(1+r)^2)+O(\hbar^4), \tag{3.27}$$

where $L(z),\tilde{L}(z)$ are the Virasoro currents with the central charges $c,\tilde{c}$, respectively,

$$c=1-\frac{3(2-r)^2}{r(r-1)}=1-\frac{6}{L(L+1)},\quad \tilde{c}=1-\frac{3(1+r)^2}{r(r-1)}.$$

In the free boson realization (3.21), $T(z)$ and $\tilde{T}(z)$ have ''natural'' expansions (3.26) and (3.27), respectively, by the following identification:

$$\lambda_n=2\hbar\sqrt{r(r-1)}\sqrt{\frac{x^{rn}-x^{-rn}}{2\hbar rn}\frac{x^{(r-1)n}-x^{-(r-1)n}}{2\hbar(r-1)n}\frac{1}{x^n-1+x^{-n}}}a_n,$$

$$P=a_0-\frac{2-r}{2\sqrt{r(r-1)}}=\tilde{a}_0-\frac{r+1}{2\sqrt{r(r-1)}},$$

$$L(z)=:\frac{1}{2}(\partial\phi(z))^2:+\frac{2-r}{2\sqrt{r(r-1)}}\partial^2\phi(z),$$

$$\tilde{L}(z)=:\frac{1}{2}(\partial\tilde{\phi}(z))^2:+\frac{r+1}{2\sqrt{r(r-1)}}\partial^2\tilde{\phi}(z),$$

where $[a_n,a_m]=n\delta_{n+m,0}$, $\partial\phi(z)=\Sigma_{n\in\mathbb{Z}}a_nz^{-n-1}$ and $\partial\tilde{\phi}(z)=\partial\phi(z)|_{a_0\to\tilde{a}_0}$. On the other hand, $T(z)$ has an expansion of the form (3.27) with

$$P=\tilde{a}_0-\frac{r+1}{2\sqrt{r(r-1)}}+\frac{i\pi(2n+1)}{2\hbar\sqrt{r(r-1)}}\quad(n\in\mathbb{Z}).$$

Kac determinant: Let $T(z)=\Sigma_{n\in\mathbb{Z}}T_nz^{-n}$, and let $U_\pm$ be the algebra generated by $\{T_n\}_{\pm n>0}$. As usual, the Verma module of highest weight $\lambda\in\mathbb{C}$ is defined as the free left U_--module generated by a vector $|\lambda\rangle$ such that $T_n|\lambda\rangle=0$ $(n>0)$ and $T_0|\lambda\rangle=\lambda|\lambda\rangle$. Likewise the right Verma module is defined by $\langle\lambda|T_n=0$ $(n<0)$, $\langle\lambda|T_0=\lambda\langle\lambda|$, and $\langle\lambda|\lambda\rangle=1$. At level N there are $p(N)$ (the number of partition) independent states, $T_{-n_1}T_{-n_2}\cdots T_{-n_l}|\lambda\rangle$ $(n_1\geqslant n_2\geqslant\cdots\geqslant n_l>0,\ \Sigma_{i=1}^l n_i=N)$. Let us number these states by the reverse lexicographic ordering for $(n_1,n_2,\ldots,n_l)$, i.e., $|\lambda;N,1\rangle=T_{-N}|\lambda\rangle$, $|\lambda;N,2\rangle=T_{-N+1}T_{-1}|\lambda\rangle,\ldots,|\lambda;N,p(N)\rangle=T_{-1}^N|\lambda\rangle$. Similarly we define $\langle\lambda;N,1|=\langle\lambda|T_N$, $\langle\lambda;N,2|=\langle\lambda|T_1T_{N-1},\ldots,\langle\lambda;N,p(N)|=\langle\lambda|T_1^N$.

We conjecture that the Kac determinant at level N is given by

$$\det(\langle\lambda;N,i|\lambda;N,j\rangle)_{1\leqslant i,j\leqslant p(N)}=\prod_{\substack{l,k\geqslant1\\lk\leqslant N}}\left((\lambda-\lambda_{l,k})(\lambda-\tilde{\lambda}_{l,k})\times\frac{(x^{rl}-x^{-rl})(x^{(r-1)l}-x^{-(r-1)l})}{x^l-1+x^{-l}}\right)^{p(N-lk)},\tag{3.28}$$

where

$$\lambda_{l,k}=x^{-lr+2k(r-1)}+x^{lr-2k(r-1)}-\frac{[r-1/2]_x}{[1/2]_x},$$

$$\tilde{\lambda}_{l,k}=-x^{l(r-1)-2kr}-x^{-l(r-1)+2kr}-\frac{[r-1/2]_x}{[1/2]_x}.$$

We remark that in the free boson realization $T_0|l,k\rangle=\lambda_{l,k}|l,k\rangle$ and $T_0 e^{\alpha iQ}|2k,l/2\rangle =\tilde{\lambda}_{l,k}e^{\alpha iQ}|2k,l/2\rangle$, where

$$\alpha=\frac{i\pi(2n+1)}{2\hbar\sqrt{r(r-1)}}\quad (n\in\mathbb{Z}).$$

E. Felder complex

The Fock spaces $\mathcal{F}_{l,k}$ themselves do not give a bosonic realization of the space of states $\mathcal{L}_{l,k}$ of the corner Hamiltonian. For this we need a cohomological construction using an analog of the Felder complex:[25]

$$\cdots \xrightarrow{X_{-2}} \mathcal{F}_{2L-l,k} \xrightarrow{X_{-1}} \mathcal{F}_{l,k} \xrightarrow{X_0} \mathcal{F}_{-l,k} \xrightarrow{X_1} \mathcal{F}_{l-2L,k} \xrightarrow{X_2} \cdots, \tag{3.29}$$

$$X_j X_{j-1}=0.$$

In the case of the algebra $A_1^{(1)}$, Lukyanov and Pugai constructed the coboundary map X_j as a power of a single operator (Ref. 4, see also Ref. 26). In our case, the formula for X_j is a little more involved.

Set

$$Q_1=\oint_{|z|=1}\underline{dz}\, x_+(z)\frac{[u+\hat{l}/2]^*}{[u+1/2]^*},$$

$$Q_2^{(a)}=\oint\oint_{|z_1|=|z_2|=1}\underline{dz_1}\,\underline{dz_2}\, x_+(z_1)x_+(z_2)\frac{1}{[u_1+1/2]^*[u_2+1/2]^*}$$
$$\times\frac{[u_1-u_2]^*}{[u_1-u_2+1]^*[u_1-u_2-1/2]^*} f_2^{(a)}(u_1+\hat{l}/2,u_2+\hat{l}/2),$$

where

$$f_2^{(a)}(u_1,u_2)=[2a+1]^*[a-1/2]^*[u_1-a]^*[u_2+a-1]^*[u_1-u_2+a-1/2]^*$$
$$-[2a-1]^*[a+1/2]^*[u_1+a]^*[u_2-a-1]^*[u_1-u_2-a-1/2]^*.$$

These operators are mutually commutative (see Lemma B1). We define ''BRST charges'' Q_l $(1\le l\le L-1)$ as follows:

$$Q_l=\begin{cases} Q_1 Q_2^{(1)}\cdots Q_2^{(m)} & (l=2m+1)\\ Q_2^{((L+1)/2-m)}\cdots Q_2^{((L-3)/2)}Q_2^{((L-1)/2)} & (l=2m)\end{cases}.$$

Note that $Q_L=Q_l Q_{L-l}$.

We prove the following propositions in Appendix B (Propositions B2 and B3).

Proposition 3.1: Suppose l' is odd and $l'\equiv l \bmod L$ $(1\le l\le L-1)$. On the space $\mathcal{F}_{l',k}$, Q_l is expressed as

$$Q_l=\oint\cdots\oint_{|z_1|=\cdots=|z_l|=R}\underline{dz_1}\cdots\underline{dz_l}\, x_+(z_1)\cdots x_+(z_l)H_l(u_1,\ldots,u_l), \tag{3.30}$$

$$H_l(u_1,\ldots,u_l)=\pm\overline{h_l}(u_1,\ldots,u_l)\prod_{1\le i<j\le l}\frac{[u_i-u_j]^*}{[u_i-u_j+1]^*[u_i-u_j-1/2]^*}, \tag{3.31}$$

where $\overline{h_l}(u_1,\ldots,u_l)$ *is holomorphic, symmetric and satisfies* $\overline{h_l}(u_1+v,\ldots,u_l+v)=\overline{h_l}(u_1,\ldots,u_l)$. *We have*

$$H_l(u_1+r-1,...,u_l)=H_l(u_1,\ldots,u_l), \tag{3.32}$$

$$H_l(u_1+\tau,\ldots,u_l)=H_l(u_1,\ldots,u_l)e^{-\pi i(l-1)/(r-1)}, \tag{3.33}$$

where $\tau=\pi i/\log x$.

Hence (3.30) does not depend on $R>0$.

Proposition 3.2: Under the same condition as above, we have

$$Q_l Q_{L-l}=0 \quad (1\leqslant l\leqslant L-1).$$

Let us call $C_{l,k}$ the cochain complex (3.29) defined by

$$X_{2j}=Q_l: \quad \mathcal{F}_{l-2jL,k}\rightarrow\mathcal{F}_{-l-2jL,k},$$

$$X_{2j+1}=Q_{L-l}: \quad \mathcal{F}_{-l-2jL,k}\rightarrow\mathcal{F}_{l-2(j+1)L,k}.$$

In the conformal limit where $x\rightarrow 1$ and $z=x^{2u}$ are kept fixed, this complex formally tends to Felder's complex[25] for the minimal unitary series. In view of this, it is natural to expect that

$$H^j(C_{l,k})=\mathrm{Ker}\,X_j/\mathrm{Im}\,X_{j-1}=0 \quad (j\neq 0). \tag{3.34}$$

By the Euler–Poincaré principle, the 0th cohomology $H^0(C_{l,k})$ then has the same character as the space of states $\mathcal{L}_{l,k}$ [see (2.6)],

$$\mathrm{tr}_{H^0(C_{l,k})}(q^d)=\mathrm{tr}_{\mathcal{L}_{l,k}}(q^{\mathcal{H}}).$$

Henceforth we assume (3.34) and make an identification

$$H^0(C_{l,k})=\mathcal{L}_{l,k}, \quad d=\mathcal{H}.$$

Proposition 3.3: Under the same assumption as in Proposition 3.1, we have on $\mathcal{F}_{l',k}$

$$[\Phi_\epsilon(z),Q_l]=0 \quad (\epsilon=0,\pm), \tag{3.35}$$

$$[T(z),Q_l]=0. \tag{3.36}$$

Proof: (3.36) is a consequence of (3.35) and (3.22). Let us show

$$\Phi_-(z)Q_l=Q_l\Phi_-(z).$$

The left (respectively, right) hand side is well defined if we choose $R\ll 1$ (respectively, $R\gg 1$) in (3.30). As meromorphic functions we have $\Phi_-(z)x_+(z_j)=x_+(z_j)\Phi_-(z)$, and the product has no poles. Since Q_l does not depend on R, the conclusion follows.

Next let us prove (3.35) with $\epsilon=0$. The case $\epsilon=+$ can be shown similarly. Dropping irrelevant constants we consider

$$\Phi_0'(z)=\oint_{C_0}\underline{dz'}\,\Phi_-(z)x_-(z')\frac{[u-u'+\hat{k}]_+}{[u-u'+1/2]}.$$

As meromorphic functions we have

$$x_-(z)x_+(z')=x_+(z')x_-(z)=\frac{z+z'}{(z+xz')(z+x^{-1}z')}:x_-(z)x_+(z'):.$$

We use the expression (3.30) with $x^2|z|<R<x^{-2}|z|$. Taking into account the symmetry in the integration variables $z_1,\ldots,z_l$, we obtain

$$[\Phi_0'(z),Q_l]=l\oint\cdots\oint_{|z_1|=\cdots=|z_l|=R}\underline{dz_1}\cdots\underline{dz_l}H_l(u_1,\ldots,u_l)(\mathrm{res}_{z'=-xz_1}+\mathrm{res}_{z'=-x^{-1}z_1})\Phi_-(z)x_-(z')x_+(z_1)\cdots x_+(z_l)\frac{[u-u'+\hat{k}]_+}{[u-u'+1/2]}\underline{dz'}.$$

By noting the identity

$$:x_+(z)x_-(-x^{-1}z):=x^{-2r+1}:x_+(z^{2r-2}z)x_-(-x^{2r-1}z):,$$

we can rewrite the right-hand side as follows:

$$l\oint\cdots\oint_{|z_2|=\cdots=|z_l|=R}\underline{dz_2}\cdots\underline{dz_l}\left(\oint_{C_1}\underline{dz_1}A(z_1,z)x_+(z_2)\cdots x_+(z_l)H_l(u_1,\ldots,u_l)-\oint_{C_2}\underline{dz_1}A(x^{2r-2}z_1,z)x_+(z_2)\cdots x_+(z_l)H_l(u_1,\ldots,u_l)\right), \tag{3.37}$$

where

$$A(z_1,z)=\mathrm{res}_{z'=-xz_1}\Phi_-(z)x_-(z')x_+(z_1)\frac{[u-u'+\hat{k}]_+}{[u-u'+1/2]}\underline{dz'}$$
$$=\frac{\text{holomorphic function}}{(-x^2z_1/z,-x^{2r}z/z_1;x^{2r})_\infty}:\Phi_-(z)x_-(-xz_1)x_+(z_1):.$$

The contours for z_1 are $(n\geqslant 0)$

	Inside	Outside
C_1	$z_1=-x^{2r(n+1)}z$	$z_1=-x^{-2-2rn}z$
C_2	$z_1=-x^{2+2rn}z$	$z_1=-x^{-2r(n+1)}z$

Moreover the product $:x_-(-xz_1)x_+(z_1):x_+(z_j)$ is holomorphic in z_1 for $|x^{2r}z_j|<|z_1|$. In view of the periodicity (3.32), the two terms of (3.37) cancel out by shifting the contour $z_1\rightarrow x^{2r-2}z_1$. □

IV. LOCAL HEIGHT PROBABILITIES

We present here a calculation of the local height probabilities (LHPs) for the dilute A_L models in the regime 2^+.

A. Two-point LHP

We have already mentioned the result (2.6) about the one-point function. As the next simplest case, let us consider the probability $P_l(a-\epsilon,a)$ of finding two neighboring local height variables to be $a-\epsilon$, $a(\epsilon=0,\pm)$:

$$P_l(a-\epsilon,a)=\frac{1}{Z_l}S(a)\mathrm{tr}_{\mathcal{L}_{l,a}}(x^{6\mathcal{H}}\Phi_\epsilon^*(z)\Phi_\epsilon(z))$$
$$=\frac{1}{Z_l}g\sqrt{S(a)S(a-\epsilon)}\mathrm{tr}_{\mathcal{L}_{l,a}}(x^{6\mathcal{H}}\Phi_{-\epsilon}(x^{-3}z)\Phi_\epsilon(z)). \tag{4.1}$$

Note that $P_l(a-\epsilon,a)$ is independent of z. From (4.1) and the property of the type I VO (3.20), we have the following relations:

$$\sum_{\epsilon=\pm1,0} P_l(a-\epsilon,a)=\frac{S(a)\chi_{l,a}(x^6)}{Z_l},\quad P_l(a-\epsilon,a)=P_l(a,a-\epsilon),\quad P_l(0,1)=0.$$

The evaluation of the trace yields the following expressions:

$$P_l(a-1,a)=-\frac{S(a-1)x^{(1-r)/r}}{[a-\frac{1}{2}]_+[2a-2]}\oint\oint_{C_+(1)}\underline{dw_1}\,\underline{dw_2}\,\mathcal{I}(w_1,w_2)\times\frac{[v_1-\frac{1}{2}][v_1-2a+\frac{3}{2}][v_1-v_2+a]_+}{[v_1+\frac{1}{2}][v_1-\frac{1}{2}][v_1-v_2+\frac{1}{2}]},$$

$$P_l(a,a)=\frac{S(a)x^{(1-r)/r}}{[a+\frac{1}{2}]_+[a-\frac{1}{2}]_+}\oint_{C_0(x^{-3})}\underline{dw_1}\oint_{C_0(1)}\underline{dw_2}\,\mathcal{I}(w_1,w_2)\frac{[v_1-a+\frac{3}{2}]+[v_2-a]_+}{[v_1+1][v_2-\frac{1}{2}]},$$

$$P_l(a+1,a)=-\frac{S(a)x^{(1-r)/r}}{[a+\frac{1}{2}]_+[2a]}\oint\oint_{C_+(x^{-3})}\underline{dw_1}\,\underline{dw_2}\,\mathcal{I}(w_1,w_2)\times\frac{[v_1-2a+1][v_1-v_2+a+1]_+[v_2+\frac{1}{2}]}{[v_1+1][v_1-v_2+\frac{1}{2}][v_2-\frac{1}{2}]}.$$

Here $w_i=x^{2v_i}$ $(i=1,2)$ and

$$\begin{aligned}\mathcal{I}(w_1,w_2)&=\mathrm{tr}_{\mathcal{L}_{l,a}}(x^{6\mathcal{H}}\Phi_-(x^{-3})x_-(w_1)\Phi_-(1)x_-(w_2))\\&=\mathcal{O}_{l,a}(w_1,w_2)\frac{(x^5,x^6;x^6,x^{2r})^2_\infty}{(x^{2r+3},x^{2r+4};x^6,x^{2r})^2_\infty}\\&\quad\times\frac{(x^{2r+2}w_1,x^{2r+2}/w_1,x^{2r-1}w_2,x^{2r+2}/w_2;x^3,x^{2r})_\infty}{(x^4w_1,x^4/w_1,xw_2,x^4/w_2;x^3,x^{2r})_\infty}(x^6,w_2/w_1,x^6w_1/w_2;x^6)_\infty\\&\quad\times\frac{G_{x^6}(x^{2r-1},w_2/w_1)G_{x^6}(x^2,w_2/w_1)}{G_{x^6}(x^{2r-2},w_2/w_1)G_{x^6}(x,w_2/w_1)},\end{aligned}$$

where $\mathcal{O}_{l,a}(w_1,w_2)$ is the zero-mode contribution

$$\begin{aligned}\mathcal{O}_{l,a}(w_1,w_2)&=(x^6)^{\Delta_{l,a}-c/24}(x^3w_1w_2)^{(l(L+1)-aL)/(2(L+1))+L/(4(L+1))}\\&\quad\times\sum_{j\in\mathbb{Z}}(x^3w_1w_2)^{-Lj}((x^6)^{L(L+1)j^2-(l(L+1)-aL)j}\\&\quad-(x^6)^{L(L+1)j^2+(l(L+1)+aL)j+la}(x^3w_1w_2)^{-l}),\end{aligned}$$

and

$$G_{x^6}(A,z)=(x^6A;x^6,x^{2r})^2_\infty(Az;x^6,x^{2r})_\infty(x^6A/z;x^6,x^{2r})_\infty.$$

The contours $C_+(1),C_0(x^{-3})\cup C_0(1),C_+(x^{-3})$ are chosen as follows $(n,m\geqslant0)$: For all the contours, the poles $w_1=x^{4+3m+2rn}$, $w_2=x^{4+3m+2rn}$, $x^{4+6m+2r(n+1)}w_1$, $x^{1+6(m+1)+2rn}w_1$ are inside and the poles $w_1=x^{-4-3m-2rn}$, $w_2=x^{-1-3m-2rn}$, $x^{2-6m-2r(n+1)}w_1$, $x^{-1-6m-2rn}w_1$ are outside. In addition,

	Inside	Outside
$C_+(1)$	$w_1=x^{-1+2r(n+1)}$	$w_1=x^{-1-2rn}$
	$w_2=x^{1+2rn}w_1$	$w_2=x^{-1-2rn},x^{-1-2rn}w_1$, $x^{2-2r(n+1)}w_1$
$C_0(x^{-3})\cup C_0(1)$	$w_1=x^{-2+2rn}$	$w_1=x^{-4-2rn}$
	$w_2=x^{1+2rn}$	$w_2=x^{-1-2rn}$
$C_+(x^{-3})$	$w_1=x^{-2+2rn}$	$w_1=x^{-4-2rn}$
	$w_2=x^{1+2rn}$, $x^{1+2rn}w_1$	$w_2=x^{1-2r(n+1)}$, $x^{-1-2rn}w_1$, $x^{2-2r(n+1)}w_1$

B. General case

Integral representation of the N-point correlation functions can be derived in a similar manner. It is written in terms of the traces of the type I vertex operators as in (2.7):

$$Z_l^{-1}S(k)\mathrm{tr}_{\mathcal{L}_{l,k}}(\Phi^*_{\epsilon_1}(x^6z_1)\cdots\Phi^*_{\epsilon_N}(x^6z_N)\Phi_{\epsilon_N}(x^6z_N)\cdots\Phi_{\epsilon_1}(x^6z_1)x^{6\mathcal{H}}). \tag{4.2}$$

Here we give only the integral formula for the traces over the Fock module in a general situation

$$\mathrm{tr}_{\mathcal{F}_{l,k}}(\Phi_{\epsilon_1}(z_1)\cdots\Phi_{\epsilon_N}(z_N)x^{6\mathcal{H}}). \tag{4.3}$$

We assume $\Sigma_{t=1}^N \epsilon_t=0$. Otherwise (4.3) vanishes.

First we prepare several functions,

$$F(z)=\frac{(x^{5+2r}z;x^6,x^{2r})}{(x^7z;x^6,x^{2r})},\quad G(z)=\frac{F(z)}{F(xz)F(x^{-1}z)},$$

$$H(z)=\frac{(x^8z,x^9z,x^{9+2r}z,x^{10+2r}z;x^6,x^6,x^{2r})}{(x^{11}z,x^{12}z,x^{6+2r}z,x^{7+2r}z;x^6,x^6,x^{2r})}.$$

Define $h_{\epsilon_n}(z_n,\{w_{n,i}\},\hat{k})$ $(\epsilon_n=0,+)$ by normal-ordering the integrand of (3.6) and (3.7),

$$\Phi_{\epsilon_n}(z_n)=\oint\left(\prod_{i\in I(\epsilon)}\underline{dw_{n,i}}\right):\Phi_-(z_n)\prod_{i\in I(\epsilon)}x_-(w_{n,i}):h_{\epsilon_n}(z_n,\{w_{n,i}\},\hat{k}),$$

$$I(0)=\{1\},I(+)=\{1,2\}.$$

Explicitly we have

$$h_0(z,w,k)=\frac{x^{(2u-2v+k+1/2)(k-1/2)/r^2-k+1/2}(xz^2)^{(1-r)/2r}}{\sqrt{[k+1/2]_+[k-1/2]_+}}\,\frac{(-x^{2k}z/w,-x^{2r-2k}w/z;x^{2r})_\infty}{(xw/z,xz/w;x^{2r})_\infty},$$

$$h_+(z,w_1,w_2,k)$$

$$=\sqrt{\frac{S(k-1)}{S(k)}}\frac{x^{\{4(k-1)(u-v_1+k-1/2)+(k-1/2)(2v_1-2v_2+k+1/2)\}/r^2-3k+5/2}(xz^2/w_1)^{(1-r)/r}}{[k-1/2]_+[2k-2]}$$

$$\times\frac{(x^{2r-1}w_2/z,x^{4k-3}z/w_1,x^{2r-4k+3}w_1/z;x^{2r})_\infty}{(xw_2/z,xw_1/z,xz/w_1;x^{2r})_\infty}$$

$$\times\frac{(x^2w_2/w_1,-x^{2k}w_1/w_2,-x^{2r-2k}w_2/w_1;x^{2r})_\infty}{(xw_2/w_1,x^{2r-2}w_2/w_1,xw_1/w_2;x^{2r})_\infty}\left(1-\frac{w_1}{w_2}\right).$$

We use the symbol $\langle\langle A(z)B(w)\rangle\rangle$ to denote the normal ordering factors

$$A(z)B(w)=\langle\langle A(z)B(w)\rangle\rangle:A(z)B(w):.$$

(See the list in Appendix A.)

With this notation we have

$$\mathrm{tr}_{\mathcal{F}_{l,k}}(\Phi_{\epsilon_1}(z_1)\cdots\Phi_{\epsilon_N}(z_N)x^{6\mathcal{H}})$$

$$=\oint\cdots\oint\prod_{\substack{1\leq m\leq N\\ \epsilon_m\neq -}}\left(\prod_{j\in I(\epsilon_m)}\underline{dw_{m,j}}\right)h_{\epsilon_m}\left(z_m,\{w_{m,j}\},k+\sum_{t=1}^{m}\epsilon_t\right)$$

$$\times\prod_{1\leq m<n\leq N}\langle\langle\Phi_-(z_m)\Phi_-(z_n)\rangle\rangle\prod_{\substack{1\leq m<n\leq N\\ i\in I(\epsilon_n);j\in I(\epsilon_m)}}\langle\langle x_-(w_{m,j})x_-(w_{n,i})\rangle\rangle$$

$$\times\prod_{\substack{1\leq n<m\leq N\\ j\in I(\epsilon_m)}}\langle\langle\Phi_-(z_n)x_-(w_{m,j})\rangle\rangle\prod_{\substack{1\leq m<n\leq N\\ j\in I(\epsilon_m)}}\langle\langle x_-(w_{m,j})\Phi_-(z_n)\rangle\rangle$$

$$\times\mathrm{tr}_{\mathcal{F}_{l,k}}\left(:\Phi_-(z_1)\cdots\Phi_-(z_N)\prod_{\substack{1\leq m\leq N\\ j\in I(\epsilon_m)}}x_-(w_{m,j}):x^{6\mathcal{H}}\right),\tag{4.4}$$

where

$$\mathrm{tr}_{\mathcal{F}_{l,k}}\left(:\Phi_-(z_1)\cdots\Phi_-(z_N)\prod_{\substack{1\leq m\leq N\\ j\in I(\epsilon_m)}}x_-(w_{m,j}):x^{6\mathcal{H}}\right)$$

$$=\prod_{1\leq m,n\leq N}H(z_n/z_m)\prod_{\substack{1\leq m,n\leq N\\ i\in I(\epsilon_n),j\in I(\epsilon_m)}}G(w_{n,i}/w_{m,j})$$

$$\times\prod_{\substack{1\leq n,m\leq N\\ j\in I(\epsilon_m)}}F(z_n/w_{m,j})F(w_{m,j}/z_n)\times\frac{x^{6(p_{l,k}^2/4-1/24)}}{(x^6;x^6)_\infty}$$

$$\times\left(\left(\prod_{\substack{1\leq m\leq N\\ j\in I(\epsilon_m)}}w_{m,j}\right)\Bigg/\left(\prod_{1\leq n\leq N}z_n\right)\right)^{\sqrt{(r-2)/r}\,p_{l,k}}$$

$$\times\left(\prod_{\substack{1\leq m\leq N\\ j\in I(\epsilon_m)}}w_{m,j}\prod_{1\leq n\leq N}z_n\right)^{(r-1)/(2r)},\tag{4.5}$$

where $p_{l,k}$ is given in (3.2). The following are the list of poles of the integrand as functions of $w_{m,j}$. The contour for $\underline{dw}_{m,j}$ encircles only those denoted ''inside'' $(a,b\in\mathbb{Z}_{\geqslant 0})$:

	Inside	Outside
$h_0(z_m,w_{m,1},k)$	$w_{m,1}=x^{1+2rb}z_m$	$w_{m,1}=x^{-1-2rb}z_m$
$h_+(z_m,w_{m,1},w_{m,2},k)$	$w_{m,1}=x^{1+2rb}z_m$	$w_{m,1}=x^{-1-2rb}z_m$
	$w_{m,2}=x^{1+2rb}w_{m,1}$	$w_{m,2}=x^{-1-2rb}w_{m,1},x^{-1-2rb}z_m$
$\langle\langle x_-(w_{m,j})\Phi_-(z_n)\rangle\rangle$	$w_{m,j}=x^{1+2rb}z_n$	
$\langle\langle \Phi_-(z_n)x_-(w_{m,j})\rangle\rangle$		$w_{m,j}=x^{-1-2rb}z_n$
$\langle\langle x_-(w_{m,j})x_-(w_{n,i})\rangle\rangle$	$w_{m,j}=x^{1+2rb}w_{n,i}$	
	$w_{m,j}=x^{-2+2r(1+b)}w_{n,i}$	
$G(w_{n,i}/w_{m,j})$	$w_{m,j}=x^{7+2rb+6a}w_{n,i}$	
	$w_{m,j}=x^{2r(1+b)+6(1+a)}w_{n,i}$	
	$w_{m,j}=x^{4+2r(1+b)+6a}w_{n,i}$	
$F(z_n/w_{m,j})$	$w_{m,j}=x^{7+2rb+6a}z_n$	$w_{m,j}=x^{-7-2rb-6a}z_n$

The formula for the N-point correlation function (4.2) can be obtained through specializing (4.5) and noting

$$H(x^3z)H(z)=\frac{1}{F(x^2z)F(xz)}.$$

Since the result is lengthy we do not present it here.

V. DISCUSSION

As was discussed in the main text, the DVA for the dilute A_L model [which we have denoted by $\mathcal{V}_{x,r}(A_2^{(2)})$] exactly coincides with the one found by Brazhnikov and Lukyanov.[13] In Ref. 13, $\mathcal{V}_{x,r}(A_2^{(2)})$ with $|x|=1$ was treated as the Zamolodchikov–Faddeev (ZF) algebra for the Bullough–Dodd model ($A_2^{(2)}$ Toda field theory). We regard $\mathcal{V}_{x,r}(A_2^{(2)})$ with $0<x<1$, $r=2(L+1)/(L+2)$ as the ZF algebra for the dilute A_L model (restricted face model), and apply the idea of bootstrap method to study the fusion of the $\mathcal{V}_{x,r}(A_2^{(2)})$ current $T(z)$.

The two-dimensional Ising model at the critical temperature $T=T_c$ is described by the $c=1/2$ minimal CFT. Perturbing it by a magnetic field while keeping the same temperature ($T=T_c$), an off-critical integrable model is obtained.[20] A fascinating feature of this theory is that the Lie algebra E_8 appears as a hidden symmetry; one can check that the integrals of motion P_s appear at the exponents of E_8, $s=1,7,11,13,17,19,\ldots$, the bootstrap program closes within eight particles, the mass ratios are given by the Perron–Frobenius vector for the incidence matrix of E_8, and so on. Further discussions of the model as the $\phi_{1,2}$-perturbation of the $c=1/2$ CFT can be found in Ref. 27. It is argued that the dilute A_3 model is in the universality class of the magnetic-perturbed Ising model.[1] As in the case of the ABF model,[4] our free field realization for the dilute A_3 model properly reduces to that of the $c=1/2$ CFT, including the VOs, $\mathcal{V}_{x,r}(A_2^{(2)})$, and the Felder complex. Our description of the dilute A_3 model, therefore, provides a lattice analog of the $\phi_{1,2}$-perturbation of the $c=1/2$ CFT.

In this section, we study an E_8 structure arising from $\mathcal{V}_{x,r}(A_2^{(2)})$ for the dilute A_3 model ($r=8/5$). We construct eight fused DVA currents $T^{(a)}(u)$ $(a=1,2,\ldots,8)$ from the fundamental $\mathcal{V}_{x,r}(A_2^{(2)})$ current $T(z)$ using a bootstrap procedure. We show that these fused currents obey a set of relations which resembles the so-called level-two restricted T-system of type $E_8^{(1)}$.[21]

The T-system of type $E_8^{(1)}$[21] is written as

$$T_m^{(a)}(u-\tfrac{1}{20})T_m^{(a)}(u+\tfrac{1}{20})=T_{m-1}^{(a)}(u)T_{m+1}^{(a)}(u)+g_m^{(a)}(u)\prod_{b\sim a}T_m^{(b)}(u), \tag{5.1}$$

where $T_m^{(a)}(u)$ $(a=1,2,...,8)$ denotes the eigenvalues of the transfer matrix, the symbol $b\sim a$ means that b and a are adjacent nodes in the Dynkin diagram of E_8, and $g^{(a)}(u)$'s are some functions. (We have rescaled the u variable of Ref. 21 to fit the present notation.) If the face model is restricted, then we have the truncation $1\leqslant m\leqslant \ell$. The integer ℓ is called level. If $\ell=2$, $T_2^{(a)}(u)$ becomes proportional to the identity, and (5.1) reduces to

$$T_1^{(a)}(u-\tfrac{1}{20})T_1^{(a)}(u+\tfrac{1}{20})=\phi^{(a)}(u)+g_1^{(a)}(u)\prod_{b\sim a}T_1^{(b)}(u), \tag{5.2}$$

with some functions $\phi^{(a)}(u)$. This is called the level-two restricted T-system of type $E_8^{(1)}$. In Ref. 22, the T-system in (5.2) is realized in terms of the ''quantum'' transfer matrix for the dilute A_3 model.

Now we come back to the deformed Virasoro algebra $\mathcal{V}_{x,r}(A_2^{(2)})$ for the dilute A_3 model. Before going into the technical details, let us roughly state the type of formulas we find for $\mathcal{V}_{x,r}(A_2^{(2)})$ with $r=8/5$.

Definition 5.1: Define the eight DVA currents $T^{(a)}(u)$ $(a=1,2,...,8)$ corresponding to the simple roots of E_8 by

$$T^{(a)}(u)=T_{\bar{a}}(u),\quad a=1,2,3,4,5,$$

$$T^{(6)}(u)=f_{1\bar{3}}(u_2-u_1)T_1(u_1)T_{\bar{3}}(u_2)\Big|_{\substack{u_1=u+11/20\\ u_2=u-3/20}},$$

$$T^{(7)}(u)=T_2(u),$$

$$T^{(8)}(u)=f_{1\bar{2}}(u_2-u_1)T_1(u_1)T_{\bar{2}}(u_2)\Big|_{\substack{u_1=u+9/20\\ u_2=u-4/20}}.$$

Here the fused DVA currents $T_n(u)$, $T_{\bar{n}}(u)$ and the structure function $f_{m\bar{n}}(u)$ are defined in Definition 5.5 below.

Proposition 5.2: The following relations hold:

$$f^{(1)}(u_1,u_2)T^{(1)}(u_1)T^{(1)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=T^{(2)}(u), \tag{5.3}$$

$$f^{(2)}(u_1,u_2)T^{(2)}(u_1)T^{(2)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=g^{(2)}(u_1,u_2)T^{(1)}(u_1)T^{(3)}(u_2)\Big|_{\substack{u_1=u\\ u_2=u}}, \tag{5.4}$$

$$f^{(3)}(u_1,u_2)T^{(3)}(u_1)T^{(3)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=g^{(3)}(u_1,u_2)T^{(2)}(u_1)T^{(4)}(u_2)\Big|_{\substack{u_1=u\\ u_2=u}}, \tag{5.5}$$

$$f^{(4)}(u_1,u_2)T^{(4)}(u_1)T^{(4)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=g^{(4)}(u_1,u_2)T^{(3)}(u_1)T^{(5)}(u_2)\Big|_{\substack{u_1=u\\ u_2=u}}, \tag{5.6}$$

$$f^{(5)}(u_1,u_2)T^{(5)}(u_1)T^{(5)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=g^{(5)}(u_1,u_2,u_3)T^{(4)}(u_1)T^{(6)}(u_2)T^{(8)}(u_3)\Big|_{\substack{u_1=u\\ u_2=u\\ u_3=u}}, \tag{5.7}$$

$$f^{(6)}(u_1,u_2)T^{(6)}(u_1)T^{(6)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=g^{(6)}(u_1,u_2)T^{(5)}(u_1)T^{(7)}(u_2)\Big|_{\substack{u_1=u\\ u_2=u}}, \tag{5.8}$$

$$f^{(7)}(u_1,u_2)T^{(7)}(u_1)T^{(7)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=T^{(6)}(u), \tag{5.9}$$

$$f^{(8)}(u_1,u_2)T^{(8)}(u_1)T^{(8)}(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}=T^{(5)}(u), \tag{5.10}$$

with appropriate functions $f^{(a)}$, $g^{(a)}$ (*see Definition 5.15*). *Both sides are regarded as operators on the cohomology* $H^0(C_{l,k})$. *Likewise we have the relations*

$$[\![u_2-u_1-\tfrac{3}{2}]\!] f^{(a)}(u_1,u_2)T^{(a)}(u_1)T^{(a)}(u_2)\Big|_{\substack{u_1=u+1/20-r/2\\ u_2=u-1/20+r/2}} = c^{(a)}\mathrm{id}, \tag{5.11}$$

for $a=1,2,...,8$, *where* $c^{(a)}$*'s are some constants and the symbol* $[\![\;]\!]$ *is defined below in* (5.14).

We notice that (5.3)–(5.11) for the fused $\mathcal{V}_{x,r}(A_2^{(2)})$ currents look very similar to the *T*-system (5.2) arising from the analytic Bethe ansatz. There is, however, an obvious discrepancy between them; while the *T*-system (5.2) comprises two terms, (5.3)–(5.11) consists of only one term. More precisely, the right-hand side of (5.3)–(5.10) corresponds to the second term in (5.2), whereas that of (5.11) corresponds to the first term. In the left-hand side, the spectral parameters u_1-u_2 of (5.3)–(5.10) and (5.11) differ by *r*. Such a shift by *r* is irrelevant in the *T*-system (5.2), because the transfer matrix eigenvalues $T_m^{(a)}(u)$ (with appropriate normalization) are periodic, $T_m^{(a)}(u+r)=T_m^{(a)}(u)$. This is a reflection of the quasiperiodicity of the Boltzmann weights. On the other hand, the $\mathcal{V}_{x,r}(A_2^{(2)})$ currents $T^{(a)}(u)$ *are by no means* doubly quasiperiodic; we have $T^{(a)}(u+\pi i/\log x)=T^{(a)}(u)$ but $T^{(a)}(u+r)\neq T^{(a)}(u)$.

We have not understood yet the reason why we have such similarities between the *T*-system for the Bethe ansatz and the exchange relations for the DVA. For the purpose of comparison, we summarize in Appendix C the fusions of the DVA current and the ''*T*-system'' for the algebra $A_{N-1}^{(1)}$.

In the rest of this section, we briefly sketch the derivation of Proposition 5.2.

A. Operator product expansions

Let *r* be generic for a while. We set

$$r^*=r-1, \tag{5.12}$$

$$[u]_x=\frac{x^u-x^{-u}}{x-x^{-1}}, \tag{5.13}$$

$$[\![u]\!]=\frac{[u]_x}{[u+r^*+1]_x}=\frac{1}{[\![-u-r^*-1]\!]}. \tag{5.14}$$

In this section we prefer to use the additive notation and write $f(u)$ for the structure function $f(z)$ $(z=x^{2u})$ in (3.24). It satisfies the relations:

Lemma 5.3:

$$(i)\qquad \frac{f(u-\frac{1}{2})f(u+\frac{1}{2})}{f(u)}=\frac{[\![u-r^*-1/2]\!]}{[\![u-1/2]\!]},$$

$$(ii)\qquad f(u)f\left(u\pm\frac{3}{2}\right)=\frac{[\![\pm u-r^*]\!]}{[\![\pm u]\!]}\frac{[\![\pm u-r^*+1/2]\!]}{[\![\pm u+1/2]\!]},$$

$$(iii)\qquad f(u-1)f(u)f(u+1)=\frac{[\![u-r^*-1]\!]}{[\![u-1]\!]}\frac{[\![u-r^*-1/2]\!]}{[\![u-1/2]\!]}\frac{[\![u-r^*]\!]}{[\![u]\!]}.$$

The operators $\Lambda_\pm(z)$, $\Lambda_0(z)$ are defined by (3.21).

Lemma 5.4: The operator product expansions (OPEs) among $\Lambda_i(u)$ *are*

$$f(u_2-u_1)\Lambda_i(x^{2u_1})\Lambda_j(x^{2u_2})=:\Lambda_i(x^{2u_1})\Lambda_j(x^{2u_2}):\times\begin{cases}1 & (i,j)=(+,+)\\ \frac{[\![u_1-u_2-r^*]\!]}{[\![u_1-u_2]\!]} & (i,j)=(+,0)\\ \frac{[\![u_1-u_2-r^*]\!]}{[\![u_1-u_2]\!]}\frac{[\![u_1-u_2-r^*+1/2]\!]}{[\![u_1-u_2+1/2]\!]} & (i,j)=(+,-)\\ \frac{[\![u_1-u_2-r^*-1]\!]}{[\![u_1-u_2-1]\!]} & (i,j)=(0,+)\\ \frac{[\![u_1-u_2-r^*-1/2]\!]}{[\![u_1-u_2-1/2]\!]} & (i,j)=(0,0)\\ \frac{[\![u_1-u_2-r^*]\!]}{[\![u_1-u_2]\!]} & (i,j)=(0,-)\\ \frac{[\![u_1-u_2-r^*-1]\!]}{[\![u_1-u_2-1]\!]}\frac{[\![u_1-u_2-r^*-3/2]\!]}{[\![u_1-u_2-3/2]\!]} & (i,j)=(-,+)\\ \frac{[\![u_1-u_2-r^*-1]\!]}{[\![u_1-u_2-1]\!]} & (i,j)=(-,0)\\ 1 & (i,j)=(-,-)\end{cases}.$$

B. Fused $\mathcal{V}_{x,r}(A_2^{(2)})$ currents $T_n(u)$ and $T_{\bar{n}}(u)$

Suggested by the bootstrap program for general r, we introduce the following fused currents of $\mathcal{V}_{x,r}(A_2^{(2)})$.

Definition 5.5: Define the fused DVA currents $T_n(u)$ and $T_{\bar{n}}(u)$ by

$$T_0(u)=T_{\bar{0}}(u)=\mathrm{id},\quad T_1(u)=T_{\bar{1}}(u)=T(x^{2u}),$$

$$T_n(u)=\prod_{1\le i<j\le n} f(u_j-u_i)\cdot T_1(u_1)T_1(u_2)\cdots T_1(u_n)\Big|_{\substack{u_i=u+((n-1)/2)r^*-(i-1)r^*\\ 1\le i\le n}},$$

$$T_{\bar{n}}(u)=\prod_{1\le i<j\le n} f(u_j-u_i)\cdot T_1(u_1)T_1(u_2)\cdots T_1(u_n)\Big|_{\substack{u_i=u+((n-1)/2)(r^*-1/2)-(i-1)(r^*-1/2)\\ 1\le i\le n}}.$$

Define the structure functions for $T_n(u)$, $T_{\bar{n}}(u)$ by

$$f_{mn}(u)=\prod_{i=1}^{m}\prod_{j=1}^{n} f\left(u+\frac{n-m}{2}r^*-(j-i)r^*\right),$$

$$f_{\overline{mn}}(u)=\prod_{i=1}^{m}\prod_{j=1}^{n} f\left(u+\frac{n-m}{2}\left(r^*-\frac{1}{2}\right)-(j-i)\left(r^*-\frac{1}{2}\right)\right),$$

$$f_{\bar{m}n}(u)=f_{n\bar{m}}(u)=\prod_{i=1}^{m}\prod_{j=1}^{n} f\left(u+\frac{n+1}{2}r^*-\frac{m+1}{2}\left(r^*-\frac{1}{2}\right)-jr^*+i\left(r^*-\frac{1}{2}\right)\right).$$

These fused DVA currents enjoy the ZF exchange relations.

Lemma 5.6: As meromorphic functions, the following exchange relations hold:

$$f_{mn}(v-u)T_m(u)T_n(v)=f_{nm}(u-v)T_n(v)T_m(u),$$

$$f_{\overline{mn}}(v-u)T_{\overline{m}}(u)T_{\overline{n}}(v)=f_{\overline{nm}}(u-v)T_{\overline{n}}(v)T_{\overline{m}}(u),$$

$$f_{m\overline{n}}(v-u)T_{m}(u)T_{\overline{n}}(v)=f_{\overline{n}m}(u-v)T_{\overline{n}}(v)T_{m}(u).$$

Lemma 5.7:

$$(i)\quad [\![u-v\pm 1]\!]f(v-u)T_1(u)T_1(v)|_{u=v\mp 1}=\frac{\mp T_1(v\mp 1/2)}{[\![-1/2]\!][\![-1]\!]},$$

$$(ii)\quad [\![u-v\pm 3/2]\!]f(v-u)T_1(u)T_1(v)|_{u=v\mp 3/2}=\frac{\mp \mathrm{id}}{[\![1/2]\!][\![-1]\!][\![-3/2]\!]},$$

$$(iii)\quad [\![u_1-u_2-1]\!][\![u_2-u_3-1]\!]f(u_2-u_1)f(u_3-u_1)f(u_3-u_2)T_1(u_1)T_1(u_2)T_1(u_3)|_{\substack{u_2=u_1-1\\ u_3=u_2-1}}$$

$$=\frac{\mathrm{id}}{[\![1]\!]\left[\!\!\left[\frac{1}{2}\right]\!\!\right]\left[\!\!\left[-\frac{1}{2}\right]\!\!\right][\![-1]\!][\![-1]\!]\left[\!\!\left[-\frac{3}{2}\right]\!\!\right][\![-2]\!]}.$$

We need the analyticity properties of the operator products. (Some details of the derivation are given in Appendix C for the case of $A_{N-1}^{(1)}$.)

Lemma 5.8: The product $f_{mn}(v-u)T_m(u)T_n(v)$ has poles only at

$$u-v=\begin{cases}\left(\frac{m+n}{2}-k\right)r^*-1\\ \left(\frac{m+n}{2}-k\right)r^*-\frac{3}{2}\\ -\left(\frac{m+n}{2}-k\right)r^*+1\\ -\left(\frac{m+n}{2}-k\right)r^*+\frac{3}{2}\end{cases}\quad k=1,2,...,\min(m,n).$$

All the poles are simple.

Lemma 5.9: The product $f_{\overline{mn}}(v-u)T_{\overline{m}}(u)T_{\overline{n}}(v)$ has poles only at

simple pole:

$$u-v=\begin{cases}\left(\frac{m+n}{2}-k\right)\left(r^*-\frac{1}{2}\right)-\frac{3}{2}\\ -\left(\frac{m+n}{2}-k\right)\left(r^*-\frac{1}{2}\right)+\frac{3}{2}\end{cases}\quad k=1,2,...,\min(m,n),$$

and

poles with multiplicity min(min(m,n),min(l,$m+n-l$)):

$$u-v=\begin{cases}\left(\frac{m+n}{2}-l\right)\left(r^*-\frac{1}{2}\right)-1\\ -\left(\frac{m+n}{2}-l\right)\left(r^*-\frac{1}{2}\right)+1\end{cases}\quad l=1,2,...,m+n-1.$$

Lemma 5.10: The product $f_{\overline{m}n}(v-u)T_{\overline{m}}(u)T_n(v)$ has poles only at

$$u-v=\begin{cases} \frac{n-1}{2}r^*+\left(\frac{m+1}{2}-l\right)\left(r^*-\frac{1}{2}\right)-1 & \\ \frac{n-1}{2}r^*+\left(\frac{m+1}{2}-k\right)\left(r^*-\frac{1}{2}\right)-\frac{3}{2} & l=1,2,\dots m, \\ -\frac{n-1}{2}r^*-\left(\frac{m+1}{2}-l\right)\left(r^*-\frac{1}{2}\right)+1 & k=1,2\dots,\min(m,n). \\ -\frac{n-1}{2}r^*-\left(\frac{m+1}{2}-k\right)\left(r^*-\frac{1}{2}\right)+\frac{3}{2} & \end{cases}$$

All the poles are simple.

C. The case of the dilute A_3 model

The parameters for the dilute A_3 model are given by

$$L=3, \quad r=2\frac{L+1}{L+2}=\frac{8}{5}, \quad r^*=\frac{3}{5}.$$

For $L=3$, we expect to have the following extra symmetry for $T_n(u)$.

Conjecture 5.11: As operators acting on the BRST cohomology $H^0(C_{l,k})$ $(k=1,2,3,\ l=1,2)$, we have

$$(i) \quad T_3(u)=\frac{T_2(u)}{[\![\frac{2}{10}]\!][\![\frac{3}{10}]\!]},$$

$$(ii) \quad T_4(u)=\frac{-T_1(u)}{[\![-\frac{9}{10}]\!][\![-\frac{4}{10}]\!][\![-\frac{3}{10}]\!][\![\frac{3}{10}]\!][\![\frac{2}{10}]\!][\![\frac{8}{10}]\!]},$$

$$(iii) \quad T_5(u)=\frac{-\mathrm{id}}{[\![-\frac{15}{10}]\!][\![-\frac{10}{10}]\!][\![-\frac{9}{10}]\!][\![-\frac{4}{10}]\!][\![-\frac{3}{10}]\!][\![\frac{2}{10}]\!][\![\frac{3}{10}]\!][\![\frac{8}{10}]\!][\![\frac{9}{10}]\!][\![\frac{14}{10}]\!]}.$$

One of the grounds for this Conjecture 5.11 is the degeneration of the structure functions.

Lemma 5.12: For $r^*=3/5$, we have

$$(i) \quad f_{13}(u)=f_{12}(u)\frac{[\![u-\frac{7}{10}]\!][\![u-\frac{12}{10}]\!]}{[\![u-\frac{4}{10}]\!][\![u-\frac{9}{10}]\!]},$$

$$(ii) \quad f_{14}(u)=f_{11}(u)\frac{[\![u-\frac{10}{10}]\!][\![u-\frac{15}{10}]\!]}{[\![u-\frac{1}{10}]\!][\![u-\frac{6}{10}]\!]},$$

$$(iii) \quad f_{15}(u)=\frac{[\![u-\frac{13}{10}]\!][\![u-\frac{18}{10}]\!]}{[\![u+\frac{2}{10}]\!][\![u-\frac{3}{10}]\!]}.$$

Using Lemmas 5.12, 5.8, 5.9, and 5.10, we can check that the replacement $T_m(u)\leftrightarrow T_{5-m}(u)$ will not affect the analyticity in any of the OPEs acting on the BRST cohomology space.

To obtain the correct proportionality constants in Conjecture 5.11, we calculated $\langle l,k|T_m(u)|l,k\rangle$ for $k=1,2,3$, $l=1,2$.

D. Fusions of $T_2(u)$ at $r^*=3/5$

If we study the bootstrap for E_8-symmetric particles carefully,[20] we realize that it is helpful to consider the fusions of $T_2(u)$.

Lemma 5.13: For $r^*=3/5$, *the following equalities hold.*

$$(i)\quad f_{22}(u_2-u_1)T_2(u_1)T_2(u_2)\Big|_{\substack{u_1=u+1/20\\ u_2=u-1/20}}$$

$$=\frac{[\![-\frac{8}{10}]\!][\![-\frac{7}{10}]\!][\![-\frac{2}{10}]\!][\![\frac{3}{10}]\!]}{[\![-\frac{5}{10}]\!][\![-\frac{4}{10}]\!][\![-\frac{3}{10}]\!]} f_{1\bar{3}}(u_2-u_1)T_1(u_1)T_{\bar{3}}(u_2)\Big|_{\substack{u_1=u+11/20\\ u_2=u-3/20}}$$

$$=\frac{[\![-\frac{8}{10}]\!][\![-\frac{7}{10}]\!][\![-\frac{2}{10}]\!][\![\frac{3}{10}]\!]}{[\![-\frac{5}{10}]\!][\![-\frac{4}{10}]\!][\![-\frac{3}{10}]\!]} f_{\bar{3}1}(u_2-u_1)T_{\bar{3}}(u_1)T_1(u_2)\Big|_{\substack{u_1=u+3/20\\ u_2=u-11/20}},$$

$$(ii)\quad f_{22}(u_2-u_1)T_2(u_1)T_2(u_2)\Big|_{\substack{u_1=u+7/20\\ u_2=u-7/20}}$$

$$=\frac{[\![-\frac{8}{10}]\!]}{[\![-\frac{5}{10}]\!][\![-\frac{3}{10}]\!][\![-\frac{2}{10}]\!]} f_{1\bar{2}}(u_2-u_1)T_1(u_1)T_{\bar{2}}(u_2)\Big|_{\substack{u_1=u+9/20\\ u_2=u-4/20}}$$

$$=\frac{[\![-\frac{8}{10}]\!]}{[\![-\frac{5}{10}]\!][\![-\frac{3}{10}]\!][\![-\frac{2}{10}]\!]} f_{\bar{2}1}(u_2-u_1)T_{\bar{2}}(u_1)T_1(u_2)\Big|_{\substack{u_1=u+4/20\\ u_2=u-9/20}}.$$

To prove these, we use

$$\tfrac{5}{2}r^*-\tfrac{3}{2}=0,\quad T_2(u)=[\![\tfrac{2}{10}]\!][\![\tfrac{3}{10}]\!]T_3(u),$$

and Lemma 5.7.

E. $T_{\bar{5}}(u)$, $T_{\bar{6}}(u)$, $T_{\bar{7}}(u)$ at $r^*=3/5$

The fused DVA currents $T_{\bar{5}}(u)$, $T_{\bar{6}}(u)$, $T_{\bar{7}}(u)$ for $r^*=3/5$ can be rewritten as follows.

Lemma 5.14: For $r^*=3/5$, *we have*

$$T_{\bar{5}}(u)=\left[\!\!\left[-\frac{1}{2}\right]\!\!\right][\![-1]\!]\left(\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!]}\right)^2[\![u_2'-u_2-1]\!]f_{\bar{2}1}(u_2-u_1)f_{\bar{2}1}(u_2'-u_1)f_{\bar{2}\bar{2}}(u_3-u_1)$$

$$\times f_{11}(u_2'-u_2)f_{1\bar{2}}(u_3-u_2')f_{1\bar{2}}(u_3-u_2)T_{\bar{2}}(u_1)T_1(u_2)T_1(u_2')T_{\bar{2}}(u_3)\Big|_{\substack{u_1=u+3/20\\ u_2=u-1/2\\ u_2'=u+1/2\\ u_3=u-3/20}},$$

$$T_{\bar{6}}(u)=\left[\!\!\left[-\frac{1}{2}\right]\!\!\right][\![-1]\!]\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!]}\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!][\![-\frac{8}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!][\![-\frac{14}{10}]\!]}[\![u_2'-u_2-1]\!]f_{\bar{3}1}(u_2-u_1)$$

$$\times f_{\bar{3}1}(u_2'-u_1)f_{\bar{3}\bar{2}}(u_3-u_1)f_{11}(u_2'-u_2)f_{1\bar{2}}(u_3-u_2')$$

$$\times f_{1\bar{2}}(u_3-u_2)T_{\bar{3}}(u_1)T_1(u_2)T_1(u_2')T_{\bar{2}}(u_3)\Big|_{\substack{u_1=u+3/20\\ u_2=u-11/20\\ u_2'=u+9/20\\ u_3=u-4/20}},$$

$$T_{\bar{7}}(u)=f_{5\bar{2}}(u_2-u_1)T_{\bar{5}}(u_1)T_2(u_2)|_{u_1=u, u_2=u}=\left[\!\left[-\frac{1}{2}\right]\!\right][\![-1]\!]\left(\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!][\![-\frac{8}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!][\![-\frac{14}{10}]\!]}\right)^2$$

$$\times[\![u_2'-u_2-1]\!]f_{\bar{3}1}(u_2-u_1)f_{\bar{3}1}(u_2'-u_1)f_{\bar{3}\bar{3}}(u_3-u_1)f_{11}(u_2'-u_2)$$

$$\times f_{1\bar{3}}(u_3-u_2')f_{1\bar{3}}(u_3-u_2)T_{\bar{3}}(u_1)T_1(u_2)T_1(u_2')T_{\bar{3}}(u_3)|_{u_1=u+4/20, u_2=u-1/2, u_2'=u+1/2, u_3=u-4/20}.$$

F. Structure functions

In accordance with the currents $T^{(a)}(u)$, we introduce the following structure functions.
Definition 5.15: Define $f^{(a)}$ $(a=1,2,...,8)$, $g^{(a)}$ $(a=2,3,...,6)$ *by*

$$f^{(a)}(u_1,u_2)=f_{\bar{a}\bar{a}}(u_2-u_1) \quad (1\leqslant a\leqslant 5),$$

$$g^{(a)}(u_1,u_2)=f_{\overline{a-1}\,\overline{a+1}}(u_2-u_1) \quad (2\leqslant a\leqslant 4),$$

$$g^{(5)}(u_1,u_2,u_3)=\left[\!\left[-\frac{1}{2}\right]\!\right][\![-1]\!]\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!]}\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!][\![-\frac{8}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!][\![-\frac{14}{10}]\!]}\frac{[\![-\frac{4}{10}]\!][\![-\frac{5}{10}]\!][\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!]}{[\![-\frac{10}{10}]\!][\![-\frac{11}{10}]\!][\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!]}$$

$$\times f_{\bar{4}\bar{3}}\left(u_2-u_1+\frac{3}{20}\right)f_{\bar{4}1}\left(u_2-u_1-\frac{11}{20}\right)f_{\bar{4}1}\left(u_3-u_1+\frac{9}{20}\right)f_{\bar{4}\bar{2}}\left(u_3-u_1-\frac{4}{20}\right)$$

$$\times[\![u_3-u_2]\!]f_{\bar{3}1}\left(u_3-u_2+\frac{6}{20}\right)f_{\bar{3}\bar{2}}\left(u_3-u_2-\frac{7}{20}\right)f_{11}(u_3-u_2+1)$$

$$\times f_{1\bar{2}}\left(u_3-u_2+\frac{7}{20}\right),$$

$$f^{(6)}(u_1,u_2)=\left[\!\left[-\frac{1}{2}\right]\!\right][\![-1]\!]\left(\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!][\![-\frac{8}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!][\![-\frac{14}{10}]\!]}\right)^2\frac{[u_1-u_2-\frac{1}{10}]_x}{[\frac{16}{10}]_x}\left[\!\left[u_1-u_2-\frac{1}{10}\right]\!\right]$$

$$\times f_{1\bar{3}}\left(u_2-u_1+\frac{4}{10}\right)f_{\bar{3}\bar{3}}\left(u_2-u_1-\frac{3}{10}\right)f_{11}\left(u_2-u_1+\frac{11}{10}\right)f_{\bar{3}1}\left(u_2-u_1+\frac{4}{10}\right),$$

$$g^{(6)}(u_1,u_2)=f_{5\bar{2}}(u_2-u_1),$$

$$f^{(7)}(u_1,u_2)=\frac{[\![-\frac{5}{10}]\!][\![-\frac{4}{10}]\!][\![-\frac{3}{10}]\!]}{[\![-\frac{8}{10}]\!][\![-\frac{7}{10}]\!][\![-\frac{2}{10}]\!][\![\frac{3}{10}]\!]}\times f_{22}(u_2-u_1),$$

$$f^{(8)}(u_1,u_2)=\left[\!\left[-\frac{1}{2}\right]\!\right][\![-1]\!]\left(\frac{[\![-\frac{6}{10}]\!][\![-\frac{7}{10}]\!]}{[\![-\frac{12}{10}]\!][\![-\frac{13}{10}]\!]}\right)^2\left[\!\left[u_1-u_2-\frac{1}{10}\right]\!\right]f_{1\bar{2}}\left(u_2-u_1-\frac{5}{20}\right)$$

$$\times f_{\bar{2}\bar{2}}\left(u_2-u_1+\frac{8}{20}\right)f_{11}\left(u_2-u_1-\frac{18}{20}\right)f_{\bar{2}1}\left(u_2-u_1-\frac{5}{20}\right).$$

Collecting all the information together, we easily obtain the ‘‘T-system’’ with the E_8 symmetry stated in Proposition 5.2.

ACKNOWLEDGMENTS

We thank Atsuo Kuniba, Yaroslav Pugai, and Junji Suzuki for discussions and their interest in this work. M.J. is grateful to Olivier Babelon and Jean Avan for their kind invitation and hospitality during his stay in Paris VI, where part of this work was done.

APPENDIX A: OPERATOR PRODUCT EXPANSIONS

We list the normal ordering relations. For operators $A(z),B(w)$ that have the form :exp(linear in boson):, we use the notation

$$A(z)B(w)=\langle\langle A(z)B(w)\rangle\rangle:A(z)B(w):$$

and write down only the part $\langle\langle A(z)B(w)\rangle\rangle$,

$$\langle\langle x_+(z_1)x_+(z_2)\rangle\rangle=z_1^{r/(r-1)}(1-z_2/z_1)\,\frac{(x^{-2}z_2/z_1,x^{2r-1}z_2/z_1;x^{2r-2})_\infty}{(x^{-1}z_2/z_1,x^{2r}z_2/z_1;x^{2r-2})_\infty},$$

$$\langle\langle x_-(z_1)x_-(z_2)\rangle\rangle=z^{(r-1)/r}(1-z_2/z_1)\,\frac{(x^{2}z_2/z_1,x^{2r-1}z_2/z_1;x^{2r})_\infty}{(xz_2/z_1,x^{2r-2}z_2/z_1;x^{2r})_\infty},$$

$$\langle\langle x_\pm(z_1)x_\mp(z_2)\rangle\rangle=z_1^{-1}\frac{1+z_2/z_1}{(1+xz_2/z_1)(1+x^{-1}z_2/z_1)},$$

$$\langle\langle \Phi_-(z_1)x_-(z_2)\rangle\rangle=z_1^{-(r-1)/r}\frac{(x^{2r-1}z_2/z_1;x^{2r})_\infty}{(xz_2/z_1;x^{2r})_\infty},$$

$$\langle\langle x_-(z_2)\Phi_-(z_1)\rangle\rangle=z_2^{-(r-1)/r}\frac{(x^{2r-1}z_1/z_2;x^{2r})_\infty}{(xz_1/z_2;x^{2r})_\infty},$$

$$\langle\langle \Phi_-(z_1)x_+(z_2)\rangle\rangle=\langle\langle x_+(z_2)\Phi_-(z_1)\rangle\rangle=(z_1+z_2),$$

$$\langle\langle \Psi^*_-(z_1)x_+(z_2)\rangle\rangle=z_1^{-r/(r-1)}\frac{(x^{2r-1}z_2/z_1;x^{2r-2})_\infty}{(x^{-1}z_2/z_1;x^{2r-2})_\infty},$$

$$\langle\langle x_+(z_2)\Psi^*_-(z_1)\rangle\rangle=z_2^{-r/(r-1)}\frac{(x^{2r-1}z_1/z_2;x^{2r-2})_\infty}{(x^{-1}z_1/z_2;x^{2r-2})_\infty},$$

$$\langle\langle \Psi^*_-(z_1)x_-(z_2)\rangle\rangle=\langle\langle x_-(z_2)\Psi^*_-(z_1)\rangle\rangle=(z_1+z_2),$$

$$\langle\langle \Phi_-(z_1)\Phi_-(z_2)\rangle\rangle=z_1^{(r-1)/r}\frac{(x^2z_2/z_1,x^3z_2/z_1,x^{2r+3}z_2/z_1,x^{2r+4}z_2/z_1;x^6,x^{2r})_\infty}{(x^5z_2/z_1,x^6z_2/z_1,x^{2r}z_2/z_1,x^{2r+1}z_2/z_1;x^6,x^{2r})_\infty},$$

$$\langle\langle \Psi^*_-(z_1)\Psi^*_-(z_2)\rangle\rangle=z_1^{r/(r-1)}\frac{(z_2/z_1,xz_2/z_1,x^{2r+3}z_2/z_1,x^{2r+4}z_2/z_1;x^6,x^{2r-2})_\infty}{(x^3z_2/z_1,x^4z_2/z_1,x^{2r}z_2/z_1,x^{2r+1}z_2/z_1;x^6,x^{2r-2})_\infty},$$

$$\langle\langle \Phi_-(z_1)\Psi^*_-(z_2)\rangle\rangle=z_1^{-1}\frac{(-x^4z_2/z_1,-x^5z_2/z_1;x^6)_\infty}{(-xz_2/z_1,-x^2z_2/z_1;x^6)_\infty},$$

$$\langle\langle \Psi^*_-(z_2)\Phi_-(z_1)\rangle\rangle=z_2^{-1}\frac{(-x^4z_1/z_2,-x^5z_1/z_2;x^6)_\infty}{(-xz_1/z_2,-x^2z_1/z_2;x^6)_\infty}.$$

As meromorphic functions we have

$$x_+(z_1)x_+(z_2)=\frac{[u_1-u_2+1]^*}{[u_1-u_2-1]^*}\frac{[u_1-u_2-1/2]^*}{[-u_1+u_2-1/2]^*}x_+(z_2)x_+(z_1),$$

$$x_-(z_1)x_-(z_2)=\frac{[u_1-u_2-1]}{[u_1-u_2+1]}\frac{[u_1-u_2+1/2]}{[-u_1+u_2+1/2]}x_-(z_2)x_-(z_1),$$

$$x_\pm(z_1)x_\mp(z_2)=x_\mp(z_2)x_\pm(z_1),$$

$$\Phi_-(z_1)x_-(z_2)=\frac{[u_1-u_2+1/2]}{[-u_1+u_2+1/2]}x_-(z_2)\Phi_-(z_1),$$

$$\Phi_-(z_1)x_+(z_2)=x_+(z_2)\Phi_-(z_1),$$

$$\Psi_-^*(z_1)x_+(z_2)=\frac{[u_1-u_2-1/2]^*}{[-u_1+u_2-1/2]^*}x_+(z_2)\Psi_-^*(z_1),$$

$$\Psi_-^*(z_1)x_-(z_2)=x_-(z_2)\Psi_-^*(z_1),$$

$$\Phi_-(z_1)\Phi_-(z_2)=\rho(u_2-u_1)\Phi_-(z_2)\Phi_-(z_1),$$

$$\Psi_-^*(z_1)\Psi_-^*(z_2)=\rho^*(u_1-u_2)\Psi_-^*(z_2)\Psi_-^*(z_1),$$

$$\Phi_-(z_1)\Psi_-^*(z_2)=\tau(u_2-u_1)\Psi_-^*(z_2)\Phi_-(z_1).$$

Here $\rho(u)$, $\rho^*(u)$, and $\tau(u)$ are given respectively by (2.4), (3.15), and (3.16).

APPENDIX B: BRST CHARGES

We give here a proof of the properties of BRST charges stated in Sec. III E. The method is a proper adaptation of Ref. 19 to the present situation.

1. Feigin–Odesskii algebra

First let us prepare the notation. Let $\widetilde{A}_n$ be the set of all functions $F(u_1,\ldots,u_n)$ which is holomorphic on $\mathbb{C}^n$, symmetric in $u_1,\ldots,u_n$, and enjoys the quasiperiodicity properties ($r^*=r-1$),

$$F(u_1+r^*,u_2,\ldots,u_n)=(-1)^nF(u_1,u_2,\ldots,u_n), \tag{B1}$$

$$F(u_1+\tau,u_2\ldots,u_n)=(-1)^nF(u_1,u_2,\ldots,u_n)\exp\left(\frac{2\pi i}{r^*}\left(nu_1-\sum_{j=2}^n u_j-\frac{n-1}{2}+n\frac{\tau}{2}\right)\right), \tag{B2}$$

where $\tau=\pi i/\log x$. Clearly,

$$\widetilde{A}_1=\mathbb{C}f_1,\quad f_1(u)=[u]^*.$$

We have also $\dim\widetilde{A}_2=2$. If $F\in\widetilde{A}_n$ is not identically 0, then it has n zeroes $\{u_1^{(j)}\}_{j=1}^n \bmod \mathbb{Z}r^*\oplus\mathbb{Z}\tau$ satisfying

$$\sum_{j=1}^n u_1^{(j)}=\sum_{j=2}^n u_2+\frac{n-1}{2}.$$

Let $F\in\widetilde{A}_m$, $G\in\widetilde{A}_n$. Following the line of Ref. 19, we define the $*$-product $F*G\in\widetilde{A}_{m+n}$ by

$$(F*G)(u_1,\ldots,u_{m+n})$$
$$=\mathrm{Sym}\left(F(u_1-n,\ldots,u_m-n)G(u_{m+1},\ldots,u_{m+n})\prod_{\substack{1\le i\le m\\ m+1\le j\le m+n}}\frac{[u_i-u_j+1]^*[u_i-u_j-1/2]^*}{[u_i-u_j]^*}\right).$$

Here the symbol Sym stands for the symmetrization. $\widetilde{A}=\oplus_{n=0}^{\infty}\widetilde{A}_n$ equipped with the $*$-product is an associative graded algebra with unit. We denote by $A=\oplus_{n=0}^{\infty}A_n$ $(A_n=\widetilde{A}_n\cap A)$ the subalgebra of $\widetilde{A}$ generated by $\widetilde{A}_1$ and $\widetilde{A}_2$.

Let us say that a function $F(u_1,\ldots,u_n)$ has the property (P) if either $n=1,2$, or $n\geqslant 3$ and

$$F(u_1,\ldots,u_n)|_{u_j-u_i=u_k-u_j=1/2}\equiv 0 \quad \text{for any distinct } i,j,k.$$

Lemma B.1: (i) Elements of A have the property (P).

(ii) A is commutative.

Proof: From the definition of $*$ we can verify that, if $F\in\widetilde{A}_m$ and $G\in\widetilde{A}_n$ have the property (P), then so does $F*G$. Hence (i) follows.

Let $g\in A_2$ be an element linearly independent from $f_1*f_1\in A_2$. To see (ii), it suffices to show that $f_1*g=g*f_1$. Set $h=f_1*g-g*f_1$. A simple check shows that $h(u+1,u,u-1)=0$. By symmetry and the property (P), $h(u_1,u_2,u_2-1)$ viewed as a function of u_1 has zeroes at $u_2+1,u_2-1/2,u_2-2$. Since their sum is different from $2u_2 \bmod \mathbb{Z}r^*\oplus\mathbb{Z}\tau$, we have $h(u_1,u_2,u_2-1)\equiv 0$. By symmetry, this implies that $h(u_1,u_2,u_3)$ has four zeroes $u_1=u_2\pm 1,u_3\pm 1$. Hence $h\equiv 0$. □

2. BRST charges

For $F\in A_n$, we define

$$Q(F)=\oint\cdots\oint\prod_{j=1}^{n}\frac{dz_j}{[u_j+\frac{1}{2}]^*}x_+(z_1)\cdots x_+(z_n)$$
$$\times\left(\prod_{1\le i<j\le n}\frac{[u_i-u_j]^*}{[u_i-u_j+1]^*[u_i-u_j-1/2]^*}\right)F(u_1+\hat{l}/2,\ldots,u_n+\hat{l}/2),$$

where the contour is $|z_1|=\cdots=|z_n|=1$. Because of the quasiperiodicity (B2), the integrand is single valued. Using the exchange relation

$$x_+(z_1)x_+(z_2)=-\frac{[u_1-u_2+1]^*}{[u_2-u_1+1]^*}\frac{[u_1-u_2-1/2]^*}{[u_2-u_1-1/2]^*}x_+(z_2)x_+(z_1)$$

along with $\hat{l}x_+(z)=x_+(z)(\hat{l}-2)$, we find

$$Q(F)Q(G)=Q(F*G). \tag{B3}$$

Now set

$$f_1(u)=[u]^*,$$

$$f_2^{(a)}(u_1,u_2)=[2a+1]^*[a-1/2]^*[u_1-a]^*[u_2+a-1]^*[u_1-u_2+a-1/2]^*$$
$$-[2a-1]^*[a+1/2]^*[u_1+a]^*[u_2-a-1]^*[u_1-u_2-a-1/2]^*,$$

and introduce $h_l\in A_l$ $(1\leqslant l\leqslant L)$ by

$$h_{2m+1}=f_1*f_2^{(1)}*\cdots*f_2^{(m)}\quad\left(0\leqslant m\leqslant\frac{L-1}{2}\right),$$

$$h_{2m}=f_2^{((L+1)/2-m)}*\cdots*f_2^{((L-3)/2)}*f_2^{((L-1)/2)}\quad\left(1\leqslant m\leqslant\frac{L-3}{2}\right).$$

We have $h_L=h_l*h_{L-1}$ $(1\leqslant l\leqslant L-1)$. We define the BRST charges by

$$Q_l=Q(h_l).$$

Propositions 3.1 and 3.2 reduce to the following assertions.

Proposition B.2: $h_l(u_1,\ldots,u_l)$ can be written as

$$h_l(u_1,\ldots,u_l)=\overline{h_l}(u_1,\ldots,u_l)\times\begin{cases}\prod_{i=1}^{l}\left[u_i+\dfrac{-l+1}{2}\right]^* & (l:\ \text{odd})\\ \prod_{i=1}^{l}\left[u_i+\dfrac{L-l+1}{2}\right]^* & (l:\ \text{even})\end{cases},\tag{B4}$$

where $\overline{h_l}$ are holomorphic and satisfies

$$\overline{h_l}(u_1+r^*,\ldots,u_l)=(-1)^{l-1}\overline{h_l}(u_1,\ldots,u_l),$$

$$\overline{h_l}(u_1+\tau,\ldots,u_l)=\overline{h_l}(u_1,\ldots,u_l)\times\exp\left(\frac{2\pi i}{r^*}\sum_{j=2}^{l}\left(u_1-u_j+\frac{\tau}{2}\right)\right).$$

Moreover it is translationally invariant, i.e.,

$$\overline{h_l}(u_1+v,\ldots,u_l+v)=\overline{h_l}(u_1,\ldots,u_l).$$

Proposition B.3: We have $h_L\equiv 0$.

Proof of Proposition B.3: Let $0\leqslant m\leqslant(L-3)/2$. In the equality $h_L=h_{2m+1}*h_{L-2m-1}$ we set $u_1=-m-1$. Using Proposition B.2, we find that each summand in the symmetrization (B.3) vanishes. Similarly, if we set $u_1=m$, then each summand of $h_L=h_{L-2m-1}*h_{2m+1}$ vanishes. Therefore h_L has $L-1$ zeroes $u_1=0,1,\cdots,(L-3)/2,-1,-2,\ldots,-(L-1)/2$.

Suppose h_L did not vanish identically. From the quasiperiodicity, h_L has a zero at $u_1=\Sigma_{j=2}^{L}u_j+L-1$. By symmetry, $u_1=u_2-\Sigma_{j=3}^{L}u_j-(L-1)$ must also be a zero. This is a contradiction. □

In the next subsection we prove Proposition B.2.

3. Proof of proposition B.2

We prove Proposition B.2 for odd $l=2m+1$. The statement is obvious for $m=0$. Assuming $m\geqslant 1$ we proceed by induction on m.

Lemma B.4: For $m=1,\ldots,(L-1)/2$ we have

$$h_{2m+1}(m,m\pm 1,u_3,\ldots,u_{2m+1})\equiv 0.\tag{B5}$$

Proof: We use the property

$$f_2^{(a)}(\pm a,\pm a+1)=0.\tag{B6}$$

In the definition of $h_{2m+1}=h_{2m-1}*f_2^{(m)}$, set $u_1=m,u_2=m+1$. Using the induction hypothesis $h_{2m-1}(m-1,\ldots)\equiv 0$ and $f_2^{(m)}(m,m+1)=0$, we see that each summand vanishes. Similarly if we set $u_1=m,u_2=m-1$ in $h_{2m+1}=f_2^{(m)}*h_{2m-1}$ and use $f_2^{(m)}(-m,-m+1)=0$, the result is zero.□

Lemma B.5: For $t=2,3,\ldots$, we have

$$h_{2m+1}(m,u_2,\ldots,u_{2m+1})\Big|_{\substack{u_{2s+1}=u_{2s}+1/2\\ t\leqslant s\leqslant m}}\equiv 0. \tag{B7}$$

Taking $t=m+1$ we obtain the first assertion of Proposition B.2.

Proof: Denote the left-hand side of (B7) by g_t. We show $g_t\equiv 0$ by induction on t.

Let $t=2$, and consider first

$$h_{2m+1}(m,m\pm 1/2,u_3,\ldots,u_{2m+1})\Big|_{\substack{u_{2s+1}=u_{2s}+1/2\\ 2\leqslant s\leqslant m}}. \tag{B8}$$

As a function of u_3, (B8) has $2m+1$ zeroes at $m+1,m-1,m\mp 1/2$ and $u_{2s}+1,u_{2s}-1/2$ $(2\leqslant s \leqslant m)$. Comparing with the quasiperiodicity, we conclude that (B8) vanishes identically. This means that g_2 as a function of u_2 has $2m+2$ zeroes at $m\pm 1,m\pm 1/2$ and $u_{2s}+1,u_{2s}-1/2$ $(2 \leqslant s\leqslant m)$. Therefore $g_2\equiv 0$.

Suppose we have shown $g_t\equiv 0$, and consider

$$g_{t+1}=h_{2m+1}(m,u_2,\ldots,u_{2t},u_{2t+1},u_{2t+2},u_{2t+2}+1/2,\ldots,u_{2m},u_{2m}+1/2). \tag{B9}$$

From the induction hypothesis, it vanishes for $u_{2t+1}=u_{2t}+1/2$. By symmetry it vanishes for

$$u_2=u_3\pm 1/2,\ldots,u_{2t+1}\pm 1/2.$$

It also vanishes for $u_2=m\pm 1$ and $u_2=u_{2s}-1/2,u_{2s}+1$ $(t+1\leqslant s\leqslant m)$. Since the number of zeroes exceed $2m+1$, we conclude $g_{t+1}\equiv 0$. □

In the case of even l, we note that

$$f_2^{((L-1)/2)}(u_1,u_2)=[2]^*\left[\frac{L}{2}\right]^*\left[u_1+\frac{L-1}{2}\right]^*\left[u_2+\frac{L-1}{2}\right]^*\left[u_1-u_2-\frac{L}{2}\right]^*$$

has a zero at $u_1=(1-L)/2$. Using this the proof goes similarly.

Lemma B.6: The function $\overline{h_l}$ in (B4) is translationally invariant.

Proof: Consider

$$\overline{h_l}(u_1+v,\ldots,u_l+v).$$

It is holomorphic and doubly periodic in v, hence it is a constant. The conclusion follows by setting $v=0$. □

APPENDIX C: DEFORMED *W* ALGEBRA FOR $\mathfrak{sl}_N$

We discuss here the fusion of the deformed W algebra (DWA) associated with $\mathfrak{sl}_N$.[28,29]

1. Basic current

Fix complex numbers $x,r^*\in\mathbb{C}$, $0<|x|<1$. We keep the notation $[u]_x$ (5.13) and $[\![u]\!]$ (5.14). In this appendix we consider the case of "generic" r^* i.e., we assume that $m,n\in\mathbb{Z},[m+nr^*]_x=0$ implies $m=n=0$.

In the free field realization, the simplest current of DWA for the algebra $\mathfrak{sl}_N$ is presented in the form

$$W_{(1)}(u)=\sum_{i=1}^{N}\Lambda_i(u).$$

Each $\Lambda_i(u)$ is a normally ordered exponential of bosonic oscillators. Their explicit formula is irrelevant here [see, e.g., Ref. 28, Eq. (2), wherein $z=x^{2u}$, $q=x^{2r^*+2}$, $t=x^{2r^*}$ in the present notation]. We need only the following normal ordering rule for their products:

$$f(u,v)\Lambda_i(u)\Lambda_j(v)=:\Lambda_i(u)\Lambda_j(v):\times\begin{cases}\dfrac{[\![u-v-1-r^*]\!]}{[\![u-v-1]\!]} & (i<j)\\ 1 & (i=j)\\ \dfrac{[\![u-v-r^*]\!]}{[\![u-v]\!]} & (i>j),\end{cases} \tag{C1}$$

where the structure function $f(u,v)=f(v-u)$ is given by

$$f(u)=\frac{1}{(1-x^{2u})}\frac{(x^{2(u+N-1)},x^{2(u+1+r^*)},x^{2(u-r^*)};x^{2N})_\infty}{(x^{2(u+1)},x^{2(u+N+r^*)},x^{2(u+N-r^*-1)};x^{2N})_\infty}.$$

Lemma C.1: We have the exchange relation as meromorphic functions

$$f(u,v)W_{(1)}(u)W_{(1)}(v)=f(v,u)W_{(1)}(v)W_{(1)}(u). \tag{C2}$$

Both sides are regular except for simple poles at $u-v=\pm 1 \bmod \Gamma$, *where* $\Gamma=(\pi i/\log x)\mathbb{Z}$. Notice that the pole $u=v$ which appears in (C1) is canceled in (C2). In general, each matrix element of the product

$$\prod_{1\le s<t\le m} f(u_s,u_t)\times W_{(1)}(u_1)\cdots W_{(1)}(u_m)$$

is a rational function of x^{2u_i} with at most simple poles at $u_i-u_j=\pm 1$ $(i<j)$.

2. Fused currents

Let $\lambda=(\lambda_1,\ldots,\lambda_l)(\lambda_1\geqslant\cdots\geqslant\lambda_l>0)$ be a partition. We identify λ with a Young diagram. For $j,s=1,2,\ldots$, we attach a variable $u(j,s)$ to the box on the jth row and sth column of λ:

$\lambda =$ 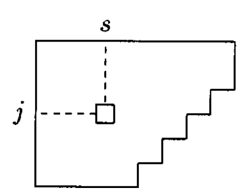 , $\quad u(j,s)=u-(j-1)-(s-1)r^*.$

For partitions $\lambda=(\lambda_1,\ldots,\lambda_l)$, $\mu=(\mu_1,\ldots,\mu_m)$, we set

$$f_{\lambda,\mu}(u,v)=\prod_{\substack{1\le j\le l\\ 1\le s\le\lambda_j}}\prod_{\substack{1\le k\le m\\ 1\le t\le\mu_k}} f(u(j,s),v(k,t)). \tag{C3}$$

We shall associate "fused" currents $W_\lambda(u)$ with each λ. First consider the case of a single row diagram $\lambda=(m)$.

Definition C.2:

$$W_{(m)}(u)=\left(\prod_{1\le s<t\le m} f(u_s,u_t)\times W_{(1)}(u_1)\cdots W_{(1)}(u_m)\right)\Bigg|_{\substack{u_s=u-(s-1)r^*\\ 1\le s\le m}}.$$

In view of the remark after Lemma C.1, the right-hand side is well-defined. Alternatively $W_{(m)}(u)$ can be defined inductively as

$$W_{(m)}(u)=f_{(m-1),(1)}(u,u')W_{(m-1)}(u)W_{(1)}(u')\Big|_{u'=u-(m-1)r^*} \tag{C4}$$

$$= f_{(1),(m-1)}(u,u')W_{(1)}(u)W_{(m-1)}(u')\Big|_{u'=u-r^*}. \tag{C5}$$

Lemma C.3: We have

$$f_{(1),(m)}(u,v)W_{(1)}(u)W_{(m)}(v) = f_{(m),(1)}(v,u)W_{(m)}(v)W_{(1)}(u). \tag{C6}$$

Both sides of (C6) are regular except for simple poles $(\mathrm{mod}\,\Gamma)$ at $u-v=-1,\ 1-(m-1)r^*$.

Proof: The exchange relation (C6) is obvious. Let us verify the statement about the position of poles by induction on *m*. The case $m=1,2$ can be verified by direct calculation. Suppose it is true for $m-1$. Using the expressions (C4) and (C5) and the induction hypothesis, we see that the possible poles in u are confined to

$$\{v-1, v+1-(m-2)r^*, v\pm 1-(m-1)r^*\}\cap\{v\pm 1, v-r^*-1, v+1-(m-1)r^*\}$$
$$=\{v-1, v+1-(m-1)r^*\}.$$

□

Arguing similarly, we have
Lemma C.4:

$$f_{(m),(n)}(u,v)W_{(m)}(u)W_{(n)}(v) = f_{(n),(m)}(v,u)W_{(n)}(v)W_{(m)}(u). \tag{C7}$$

Both sides of (C7) are regular except for simple poles $(\mathrm{mod}\,\Gamma)$ *at*

$$u-v = 1-jr^* \quad (\max(0,n-m)\leqslant j\leqslant n-1)$$
$$= -1+jr^* \quad (\max(0,m-n)\leqslant j\leqslant m-1).$$

For a general partition $\lambda=(\lambda_1,\ldots,\lambda_l)$, we define
Definition C.5:

$$W_\lambda(u) = \left(\prod_{i=1}^{l-1}[\![u_i-u_{i+1}-1]\!]\cdot\prod_{1\leqslant i<j\leqslant l} f_{(\lambda_i),(\lambda_j)}(u_i,u_j)\cdot W_{(\lambda_1)}(u_1)\cdots W_{(\lambda_l)}(u_l)\right)\Bigg|_{\substack{u_i=u-(i-1)\\ 1\leqslant i\leqslant l}}.$$

This definition makes sense by Lemma C.3. We have

$$f_{\lambda,\mu}(u,v)W_\lambda(u)W_\mu(v) = W_\mu(v)W_\lambda(u)f_{\mu,\lambda}(v,u). \tag{C8}$$

In the case of a single column diagram $\lambda=(1^a)$, $W_{(1^a)}(u)$ coincides with the fundamental DWA currents $W_a(z)$ in Refs. 28 and 29 up to a numerical factor and a shift of u [see (C10) below].

We remark that another fused current can be constructed similarly by replacing $u(j,s)=u-(j-1)-(s-1)r^*$ with $u-(j-1)+(s-1)r$ $(r=r^*+1)$.

3. Tableaux sum

Let λ be a partition. Denote by $\mathrm{SST}(\lambda)$ the set of semistandard tableaux of shape λ on the letters $\{1,2,\ldots,N\}$. For $T\in\mathrm{SST}(\lambda)$, we set

$$\Lambda_T(u) = :\prod_{\substack{1\leqslant j\leqslant l\\ 1\leqslant s\leqslant\lambda_j}}\Lambda_{T(j,s)}(u(j,s)):,$$

where $T(j,s)\in\{1,\ldots,N\}$ signifies the letter in the (j,s)th position of T.

The current $W_\lambda(u)$ is given explicitly as follows.

Proposition C.6: We have

$$W_\lambda(u)=d_\lambda \sum_{T\in \mathrm{SST}(\lambda)} c_T\cdot \Lambda_T(u).$$

The coefficients d_λ, c_T *are given by*

$$d_\lambda=\prod_{1+j<k}\frac{[\![k-j-1-\lambda_j r^*]\!]_{\lambda_k}}{[\![k-j-1]\!]_{\lambda_k}}\cdot\prod_{k=2}^{l}\frac{[\![-\lambda_{k-1}r^*]\!]_{\lambda_k}}{[\![r^*]\!]_{\lambda_k-1}},$$

$$c_T=\prod_{j=1}^{l}\frac{\Pi_{i=1}^{N}[\![-1]\!]_{w_{ji}}}{[\![-1]\!]_{\lambda_j}}\cdot\prod_{j<k}\frac{[\![k-j-\lambda_j r^*]\!]_{\lambda_k}}{[\![k-j-1-\lambda_j r^*]\!]_{\lambda_k}}\prod_{j<k}\prod_{i=1}^{N}\frac{[\![k-j-1+(s_{k,i-1}-s_{j,i-1})r^*]\!]_{w_{k,i}}}{[\![k-j+(s_{k,i-1}-s_{j,i})r^*]\!]_{w_{k,i}}},$$

where w_{ji} *is the number of the letter* i *in the* jth *row of* T $(1\leq j\leq l,\ 1\leq i\leq N)$, $s_{ji}=w_{j1}+\cdots+w_{ji}$, and

$$[\![u]\!]_n=[\![u]\!][\![u+r^*]\!]\cdots[\![u+(n-1)r^*]\!].$$

We omit the proof. Notice that $c_{T_0}=1$ for the tableau T_0 with $T_0(j,s)=j$ for all j,s.

Example.

$$W_{(m)}(u)=\sum_{\substack{w_1,\dots,w_N\geq 0\\ w_1+\cdots+w_N=m}}\frac{\Pi_{i=1}^{N}[\![-1]\!]_{w_i}}{[\![-1]\!]_m}\cdot\Lambda_T(u),\tag{C9}$$

where $T=(1^{w_1},2^{w_2},\dots,N^{w_N})$,

$$W_{(1^a)}(u)=d_{(1^a)}\sum_{1\leq i_1<\cdots<i_a\leq N}:\Lambda_{i_1}(u)\cdots\Lambda_{i_a}(u-a+1):.\tag{C10}$$

4. $W_\lambda(u)$ in terms of $W_{(1^a)}(u)$

$W_\lambda(u)$ can also be obtained from $W_{(1^a)}(u)$. First note the following fact which can be shown similarly as Lemma C.4.

Lemma C.7: *We have*

$$\tilde f_{(1^a),(1^b)}(u,v)W_{(1^a)}(u)W_{(1^b)}(v)=\tilde f_{(1^b),(1^a)}(v,u)W_{(1^b)}(v)W_{(1^a)}(u),\tag{C11}$$

where $\tilde f_{\lambda,\mu}(u,v)$ *is defined similarly as in (C3) with* $f(u,v)$ *replaced by*

$$\tilde f(u,v)=\frac{[\![u-v-1]\!]}{[\![u-v-r^*]\!]}f(u,v)=\frac{[\![v-u-1]\!]}{[\![v-u-r^*]\!]}f(u,v).$$

Both sides of (C11) are regular except for simple poles $(\mathrm{mod}\,\Gamma)$ *at*

$$\begin{aligned}u-v&=r^*-j\quad(\max(0,b-a)\leq j\leq b-1)\\&=-r^*+j\quad(\max(0,a-b)\leq j\leq a-1).\end{aligned}$$

We remark that the exchange relation (C8) holds true with $f_{\lambda,\mu}(u,v)$ replaced by $\tilde f_{\lambda,\mu}(u,v)$.

Returning to the general λ, denote by $\mu_1\geq\cdots\geq\mu_m$ its column lengths (hence the transposed diagram is $\lambda'=(\mu_1,\dots,\mu_m)$).

Lemma C.8:

$$W_\lambda(u)=\left(\prod_{i=2}^{m}[\![u_{i-1}-u_i-r^*]\!]^{-\mu_i+1}\right.$$
$$\left.\left.\times\prod_{1\leqslant i<j\leqslant l}f_{(1^{\mu_i}),(1^{\mu_j})}(u_i,u_j)\cdot W_{(1^{\mu_1})}(u_1)\cdots W_{(1^{\mu_m})}(u_m)\right)\right|_{\substack{u_i=u-(i-1)r^*\\1\leqslant i\leqslant m}}.$$

Proof: Let $\overline{\lambda}$ be the diagram obtained by removing the last column of λ so that $\lambda'=(\overline{\lambda}',\mu_m)$. We show

$$W_\lambda(u)=[\![u-v-(m-1)r^*]\!]^{-\mu_m+1}f_{\overline{\lambda},(1^{\mu_m})}(u,v)W_{\overline{\lambda}}(u)W_{(1^{\mu_m})}(v)|_{v=u-(m-1)r^*}\tag{C12}$$

by induction on μ_m. The lemma follows by repeated use of this equation.

If $\mu_m=1$, then (C12) is immediate from the definition. Assuming the statement is true for μ_m $(\mu_{m-1}\geqslant\mu_m+1\geqslant 2)$ we consider (C12) with $\lambda'=(\overline{\lambda}',\mu_m+1)$. We have

$$\begin{aligned}&[\![u-v-(m-1)r^*]\!]^{-\mu_m}f_{\overline{\lambda},(1^{\mu_m+1})}(u,v)W_{\overline{\lambda}}(u)W_{(1^{\mu_m+1})}(v)\\&\quad=[\![u-v-(m-1)r^*]\!]^{-\mu_m+1}[\![u-v'-\mu_m-(m-1)r^*]\!]^{-1}f_{\overline{\lambda},(1^{\mu_m})}(u,v)f_{\overline{\lambda},(1)}(u,v')\\&\quad\times[\![v-v'-\mu_m]\!]f_{(1^{\mu_m}),(1)}(v,v')W_{\overline{\lambda}}(u)W_{(1^{\mu_m})}(v)W_{(1)}(v')|_{v'=v-\mu_m}.\end{aligned}\tag{C13}$$

Let us verify that the right-hand side of (C13) (before specialization $v'=v-\mu_m$) is regular at $v=u-(m-1)r^*$, $v'=u-\mu_m-(m-1)r^*$. From the induction hypothesis,

$$[\![u-v-(m-1)r^*]\!]^{-\mu_m+1}f_{\overline{\lambda},(1^{\mu_m})}(u,v)W_{\overline{\lambda}}(u)W_{(1^{\mu_m})}(v)$$

is regular at $v=u-(m-1)r^*$, and

$$[\![v-v'-\mu_m]\!]f_{(1^{\mu_m}),(1)}(v,v')W_{(1^{\mu_m})}(v)W_{(1)}(v')$$

is regular at $v'=v-\mu_m$. Finally Lemma C.7 implies that

$$\tilde{f}_{\overline{\lambda},(1)}(u,v')W_{\overline{\lambda}}(u)W_{(1)}(v')$$

is regular at $v'=u-\mu_m-(m-1)r^*$, and

$$\begin{aligned}&[\![u-v'-\mu_m-(m-1)r^*]\!]^{-1}\frac{f_{\overline{\lambda},(1)}(u,v')}{\tilde{f}_{\overline{\lambda},(1)}(u,v')}\\&\quad=[\![u-v'-\mu_m-(m-1)r^*]\!]^{-1}\prod_{\substack{1\leqslant k\leqslant m-1\\1\leqslant j\leqslant\mu_k}}\frac{[\![u-v'-(j-1)-kr^*]\!]}{[\![u-v'-j-(k-1)r^*]\!]}\end{aligned}$$

is also regular (since $\mu_{m-1}\geqslant\mu_m+1$).

We let $v=u-(m-1)r^*$ in (C13) and change the order of specialization. Using the induction hypothesis for the diagram $\tilde{\lambda}'=(\overline{\lambda}',\mu_m)$, we obtain

$$\begin{aligned}&[\![u-v'-\mu_m-(m-1)r^*]\!]^{-1}[\![u-(m-1)r^*-v'-\mu_m]\!]\\&\quad\times f_{\overline{\lambda},(1)}(u,v')f_{(1^{\mu_m}),(1)}(u-(m-1)r^*,v')W_{\tilde{\lambda}}(u)W_{(1)}(v')|_{v'=u-\mu_m-(m-1)r^*}\\&\quad=f_{\tilde{\lambda},(1)}(u,v')W_{\tilde{\lambda}}(u)W_{(1)}(v')|_{v'=u-\mu_m-(m-1)r^*}=W_\lambda(u).\end{aligned}$$

□

5. Rectangular diagrams

For a rectangular Young diagram $\lambda=(m^a)$, we write $W_m^{(a)}(u)=W_{(m^a)}(u)$. The following relations may be viewed as an analog of the T-system for the transfer matrices discussed in Ref. 21.

Proposition C.9:

$$f_{(m^a),(m^a)}(u,v)W_m^{(a)}(u)W_m^{(a)}(v)|_{v=u-r*}$$
$$=(-1)^{a-1}f_{((m+1)^a),((m-1)^a)}(u,v)W_{m+1}^{(a)}(u)W_{m-1}^{(a)}(v)|_{v=u-r*}, \tag{C14}$$

$$\tilde{f}_{(m^a),(m^a)}(u,v)W_m^{(a)}(u)W_m^{(a)}(v)|_{v=u-1}$$
$$=(-1)^{m-1}C_m^{(a)}\tilde{f}_{(m^{a+1}),(m^{a-1})}(u,v)W_m^{(a+1)}(u)W_m^{(a-1)}(v)|_{v=u-1}, \tag{C15}$$

where

$$C_m^{(a)}=\prod_{1\leqslant s,t\leqslant m}\frac{[\![a-1-(s-t)r*]\!]}{[\![a-(1+s-t)r*]\!]}.$$

Both sides of (C14), (C15) are well defined.

We sketch below the proof of (C14). First we check the regularity of both sides at $v=u-r*$. For the right-hand side, this can be shown from Lemma C.4. For the left-hand side, we use Lemma C.7 to find that

$$\tilde{f}_{(m^a),(m^a)}(u,v)W_m^{(a)}(u)W_m^{(a)}(v)$$

has poles of order at most $2(m-1)$ at $u=v-r*$. Since

$$\frac{f_{(m^a),(m^a)}(u,v)}{\tilde{f}_{(m^a),(m^a)}(u,v)}=O([u-v+r*]_x^{2(m-1)}) \quad (u\to v-r*),$$

the desired regularity follows. In the same way (using Lemma C.4) we see that

$$f_{(m^a),((m-1)^a)}(u,v)W_m^{(a)}(u)W_{m-1}^{(a)}(v)$$

has poles of order at most $(a-1)$ at $u=v-r*$.

Consider the expression

$$A\equiv f_{(m^a),((m-1)^a)}(u,u')f_{(m^a),(1^a)}(u,v)f_{((m-1)^a),(1^a)}(u',v)W_m^{(a)}(u)W_{m-1}^{(a)}(u')W_1^{(a)}(v).$$

From the definition of $W_\lambda(u)$, we have

$$A=[\![u'-v-(m-1)r*]\!]^{a-1}f_{(m^a),(m^a)}(u,u')W_m^{(a)}(u)W_m^{(a)}(u')$$
$$+O([\![u'-v-(m-1)r*]\!]^a) \quad (v\to u'-(m-1)r*)$$

$$=[\![u-v-mr*]\!]^{a-1}f_{((m+1)^a),((m-1)^a)}(u,u')W_{m+1}^{(a)}(u)W_{m-1}^{(a)}(u')$$
$$+O([\![u-v-mr*]\!]^a) \quad (v\to u-mr*).$$

Writing $y=u-v-mr*$, $y'=u'-v-(m-1)r*$ and multiplying both sides by $[\![u-u'-r*]\!]^{a-1}$ we have the equality of the form

$$\varphi(y,y')=[\![y]\!]^{a-1}[\![y-y']\!]^{a-1}\psi(y-y')+O(y^a)\quad (y\to 0)$$
$$=[\![y']\!]^{a-1}[\![y-y']\!]^{a-1}\psi'(y-y')+O(y'^a)\quad (y'\to 0),$$

where $\varphi(y,y')$, $\psi(y-y')$, and $\psi'(y-y')$ are regular near $y=y'=0$. This implies that $(-1)^{a-1}\psi(0)=\psi'(0)$, and (C14) follows.

[1] S. O. Warnaar, B. Nienhuis, and K. A. Seaton, ''New construction of solvable lattice models including an Ising model in a field,'' Phys. Rev. Lett. **69**, 710–712 (1992).

[2] S. O. Warnaar, P. A. Pearce, K. A. Seaton, and B. Nienhuis, ''Order parameters of the dilute *A* models,'' J. Stat. Phys. **74**, 469–531 (1994).

[3] A. Kuniba, ''Exact solution of solid-on-solid models for twisted affine Lie algebras $A^{(2)}_{2n}$ and $A^{(2)}_{2n-1}$,'' Nucl. Phys. B **355**, 801–821 (1991).

[4] S. Lukyanov and S. Pugai, ''Multi-point local height probabilities in the integrable RSOS model,'' Nucl. Phys. B: Field Theory Stat. Syst. **473[FS]**, 631–658 (1996).

[5] O. Foda, M. Jimbo, T. Miwa, K. Miki, and A. Nakayashiki, ''Vertex operators in solvable lattice models,'' J. Math. Phys. **35**, 13–46 (1994).

[6] J. Shiraishi, H. Kubo, H. Awata, and S. Odake, ''A quantum deformation of the Virasoro algebra and the Macdonald symmetric functions,'' Lett. Math. Phys. **38**, 33–57 (1996).

[7] A. A. Kadeishvili, ''Vertex operators for deformed Virasoro algebra,'' hep-th/9604153, 1996.

[8] M. Jimbo, H. Konno, S. Odake, and J. Shiraishi, ''Elliptic algebra $U_{q,p}(\widehat{\mathfrak{sl}}_2)$: Drinfeld currents and vertex operators,'' Commun. Math. Phys. **199**, 605–647 (1999).

[9] T. Miwa and R. Weston, ''Boundary ABF models,'' Nucl. Phys. B **486**, 517–545 (1997).

[10] Y. Matsuno, ''Calculation of vertex operators for a level one integrable representation of $U_q(A^{(2)}_2)$,'' Master's thesis, Kyoto University, 1997.

[11] N. H. Jing and K. C. Misra, ''Vertex operators for twisted quantum affine algebras,'' q-alg/9701034, Trans. Am. Math. Soc. (to be published).

[12] H. Konno, ''An elliptic algebra $U_{q,p}(\widehat{\mathfrak{sl}}_2)$ and the fusion RSOS model,'' Commun. Math. Phys. **195**, 373–403 (1998).

[13] V. Brazhnikov and S. Lukyanov, ''Angular quantization and form factors in massive integrable models,'' Nucl. Phys. B **512**, 616–636 (1998).

[14] E. Frenkel and N. Reshetikhin, ''Deformations of *W*-algebras associated to simple Lie algebras,'' q-alg/9708006, 1997.

[15] P. Bouwknegt and K. Pilch, ''On deformed $\mathcal{W}$-algebras and quantum affine algebras,'' math.QA/9801112, 1998.

[16] S. Lukyanov, ''A note on the deformed Virasoro algebra,'' Phys. Lett. B **367**, 121–125 (1996).

[17] M. Jimbo, H. Konno, and T. Miwa, ''Massless *XXZ* model and degeneration of the elliptic algebra $\mathcal{A}_{q,p}(\widehat{\mathfrak{sl}}_2)$,'' in *Deformation Theory and Symplectic Geometry*, edited by D. Sternheimer, J. Rawnsley, and S. Gutt, Math. Phys. Studies Vol. 20 (Kluwer, Dordrecht, 1997), pp. 117–138.

[18] M. Jimbo and J. Shiraishi, ''A coset-type construction for the deformed Virasoro algebra,'' Lett. Math. Phys. **43**, 173–185 (1998); errata: **44**, 349–352 (1998).

[19] B. Feigin, M. Jimbo, T. Miwa, A. Odesskii, and Y. Pugai, ''Algebra of screening operators for the deformed W_n algebra,'' Commun. Math. Phys. **191**, 501–541 (1998).

[20] A. B. Zamolodchikov, ''Integrals of motion and *S*-matrix of the (scaled) $T=T_c$ Ising model with magnetic field,'' Int. J. Mod. Phys. A **4**, 4235–4248 (1989).

[21] A. Kuniba, T. Nakanishi, and J. Suzuki, ''Functional relations in solvable lattice models. I. Functional relations and representation theory,'' Int. J. Mod. Phys. A **9**, 5215–5266 (1994).

[22] J. Suzuki, ''Quantum Jacobi–Trudi formula and E_8 structure in the Ising model in a field,'' Nucl. Phys. B **528**, 683–700 (1998).

[23] M. Jimbo and T. Miwa, *Algebraic Analysis of Solvable Lattice Models*, CBMS Regional Conference Series in Mathematics Vol. 85 (American Mathematical Society, Providence, 1994).

[24] M. Jimbo, H. Konno, S. Odake, and J. Shiraishi, ''Quasi-Hopf twistors for elliptic quantum groups,'' q-alg/9712029, Transform. Groups (to be published).

[25] G. Felder, ''BRST approach to minimal models,'' Nucl. Phys. B **317**, 215–236 (1989).

[26] M. Jimbo, M. Lashkevich, T. Miwa, and S. Pugai, ''Lukyanov's screening operators for the deformed Virasoro algebra,'' Phys. Lett. A **229**, 285–292 (1997).

[27] F. Smirnov, ''Exact *S*-matrices for $\phi_{1,2}$-perturbed minimal models of conformal field theory,'' Int. J. Mod. Phys. A **6**, 1407–1428 (1991).

[28] H. Awata, H. Kubo, S. Odake, and J. Shiraishi, ''Quantum $\mathcal{W}_N$ algebras and Macdonald polynomials,'' Commun. Math. Phys. **179**, 401–416 (1996).

[29] B. L. Feigin and E. V. Frenkel, ''Quantum $\mathcal{W}$-algebras and elliptic algebras,'' Commun. Math. Phys. **178**, 653–678 (1996).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Exact solutions of the Dirac equation in a nonfactorizable metric

M. N. Hounkonnou[a)] and J. E. B. Mendy
Unité de Recherche en Physique Théorique (URPT), Institut de Mathématiques et de Sciences Physiques (IMSP), B.P. 613 Porto-Novo, Bénin

(Received 10 September 1998; accepted for publication 27 January 1999)

We present the covariant generalization of the Dirac equation in a nonfactorizable metric and give the corresponding exact solutions in terms of special functions as well as the explicit form of the spinor solution. Then we treat the particular case of the Weyl equation for the neutrinos.
[S0022-2488(99)02807-8]

I. INTRODUCTION

One of the most exciting challenges of the present day relativistic quantum theory remains, undoubtedly, the task of finding exact solutions of the Dirac equation in the presence of external fields. This can be explained by the intense research made in the subject.[1–13] The motivations for these activities are diverse and evident in view of the wide role of the Dirac equation in modern physics, for example, for investigating the spin $-\frac{1}{2}$ particle and for the necessity of analysis of synchrotronic radiation.[6]

To this purpose, many systems have been subjects of considerable interest and studies. The pioneering investigation could be the work by Brill and Wheeler in 1957,[1] who considered the Dirac equation in a central gravitational field associated with a diagonal metric. Using a normal diagonal tetrad, these authors constructed the generalized angular momentum operator separating the variables in the Dirac equation.

In a remarkable paper entitled ‘‘*Criteria of separability of the variables in the Dirac equation in gravitational fields*,’’ which appeared in 1987,[10] Shishkin and Andrushkevich provided the necessary and sufficient conditions, based on rigorous theorems, for separability of the variables for a diagonal tetrad gauge, and deduced the operators that determine the dependence of the wave function on the separated variables. In the same year, Barut and Duru[2] gave exact solutions of the Dirac equation in spatially flat Robertson–Walker space–times for models of expanding universes and discussed the current decomposition. In 1989, Shishkin and Villalba analyzed the possibilities of using the method of separation of variables in the Dirac equation in the presence of external vector fields.[6] Later, Villalba and Percoco[3] presented exact solutions and separation of variables in the Dirac and Weyl equations in a universe filled with radiation, an arbitrary expansion of the Robertson–Walker metric and in open flat and closed expanding cosmological Robertson–Walker universes. In Ref. 4, the author gave exact solutions of the Dirac equation in a static reducible Einstein space and analyzed the asymptotic behavior of the spinor solution. In Ref. 11, Shishkin and Cabos considered the separation of variables in the Dirac equation for the case of a general set of connections of the Dirac particle with the external fields, using Cartesian coordinates, whereas in Ref. 7 there are presented the possibilities of using the method of algebraic separation of variables in the Dirac equation in the presence of gravitational fields. The same technique was used[5] in the local rotating diagonal gauge in spherical coordinates and exact solutions are obtained, the energy spectrum computed and its dependence in the intensity of the Aharonov–Bohm and the magnetic monopole strengths analyzed. Let us also mention that exact solutions of the Dirac equation and the structure of metric functions were given[8] in the presence of gravitational

[a)]Electronic mail: hounkon@syfed.bj.refer.org

fields for massless neutrinos and in terms of special functions for electric neutral particles with anomalous electric and magnetic moments.[9]

Here, we provide exact solutions of the Dirac equation in a nonfactorizable metric, using the method of separation of variables adopted in Ref. 7. The general study of solutions to the Dirac equation in a curved space–time is clearly of physical interest. Indeed, such analysis is the prerequisite for the quantization of the corresponding quantum field theory. In turn, the quantized field theory enables the representation of associated quantum processes, such as particle creation through gravitational acceleration, or pair creation in the presence of strong gravitational fields, existent either close to black holes or during stellar gravitational collapse, or at the initial conditions of the Universe. Even though these issues have been well studied in specific curved space–times, such as the Schwarschild solution,[14] an analysis within a general class of space–time metrics could add new insight into the physics of such phenomena.

In Sec. III, the general class of metrics used in the present paper is introduced, for which the Dirac equation may explicitly be solved through separation of variables. These metrics are represented by

$$ds^2=-dt^2+a^2(t)(dx^2+b^2(x)[dy^2+c^2(y)dz^2]), \tag{1}$$

where $a(t)$, $b(x)$ and $c(y)$ are *a priori* arbitrary functions of the local coordinates t, x and y. This general class of metrics includes as particular cases well-established examples, which themselves could belong to other general classes of models. Thus, for example, the usual Friedman–Lemaître–Roberston–Walker homogeneous and isotropic metric of standard cosmology is of the above form (whether in Cartesian or spherical coordinates), and, more generally, the choice (1) also includes general classes of Kantowski–Sachs metrics[15] for anisotropic cosmologies. Moreover, some examples of metrics used in models for stellar gravitational collapse are[16] of the form (1). As a matter of fact, even when particularizing to specific functions $a(t)$, $b(x)$ and $c(y)$ later on, some of the considered examples correspond to these well-established curved space–times. Finally, it may also be worth pointing out that *a priori*, the above class of metrics solves Einstein's equations for specific distributions of energy-momentum of matter in space–time, in the presence of which the study of the quantized Dirac field may be of interest. Though no specific example will be provided, this avenue could be pursued in future work. The paper is organized as follows. In Sec. II, we present the covariant generalization of the Dirac equation in the nonfactorizable metric. In Sec. III, exact solutions of the Dirac equation and the explicit form of the spinor solution are obtained. Finally, in Sec. IV, we give an exact solution of the Weyl equation considering the same nonfactorizable metric.

II. DIRAC EQUATION IN A NONFACTORIZABLE METRIC

We consider the nonfactorizable metric[12]

$$ds^2=-dt^2+a^2(t)(dx^2+b^2(x)[dy^2+c^2(y)dz^2]), \tag{2}$$

and the covariant generalization of the Dirac equation

$$\{\overline{\gamma}^{\mu}(\partial_{\mu}-\Gamma_{\mu})+m\}\tilde{\Psi}=0. \tag{3}$$

The Γ_{μ} are the spinor connections satisfying the relation

$$\Gamma_{\mu}=-\tfrac{1}{4}(\partial_{\mu}h_a^{\rho}+\Gamma^{\rho}_{\sigma\mu}h_a^{\sigma})g_{\nu\rho}h_b^{\nu}\gamma^b\gamma^a. \tag{4}$$

The $\overline{\gamma}^{\mu}$ are the Dirac matrices associated with the line element (2) and are related to the γ^a-standard flat Dirac matrices as follows:

$$\overline{\gamma}^{\mu}=h_a^{\mu}\gamma^a, \quad \overline{\gamma}_{\mu}=h_{\mu}^a\gamma_a. \tag{5}$$

The tetrad components h_a^μ are defined as[7]

$$g^{\mu\nu}=h_a^\mu h_b^\nu \eta^{ab}, \quad g_{\mu\nu}=h_\mu^a h_\nu^b \eta_{ab}, \tag{6}$$

where $g_{\mu\nu}$ and η_{ab} are the metrics given by (2) and the Minkowski space–time, respectively. The following anticommutation relations hold:

$$[\overline{\gamma}^\mu,\overline{\gamma}^\nu]_+=2g^{\mu\nu}\mathbb{1}, \quad [\gamma^a,\gamma^b]_+=2g^{ab}\mathbb{1}. \tag{7}$$

We can write

$$h_\mu^a=\mathrm{diag}(1,a,ab,abc), \quad g_{\mu\nu}=\mathrm{diag}(1,a^2,a^2b^2,a^2b^2c^2). \tag{8}$$

So, Eq. (4) can be decompose as

$$\Gamma_0=0, \quad \Gamma_1=\tfrac{1}{2}a_{.t}\gamma^0\gamma^1, \quad \Gamma_2=\tfrac{1}{2}ba_{.t}\gamma^0\gamma^2+\tfrac{1}{2}b_{.x}\gamma^1\gamma^2, \tag{9}$$

$$\Gamma_3=\tfrac{1}{2}bca_{.t}\gamma^0\gamma^3+\tfrac{1}{2}cb_{.x}\gamma^1\gamma^3+\tfrac{1}{2}c_{.y}\gamma^2\gamma^3, \tag{10}$$

and the Dirac equation (3) rewrites as

$$\left\{\gamma^0\partial_t+\frac{1}{a}\gamma^1\partial_x+\frac{1}{ab}\gamma^2\partial_y+\frac{1}{abc}\gamma^3\partial_z+m\right\}\Psi=0, \tag{11}$$

where $\tilde{\Psi}$ is related to Ψ by

$$\tilde{\Psi}=a^{-3/2}b^{-1}c^{-1/2}\Psi. \tag{12}$$

Since the metric (2) is a function only of variables t, x and y, we can set

$$\Psi=\Psi(t,x,y)\exp(ik_z z). \tag{13}$$

Using the method of separation of variables,[7] we transform (11) as a sum of two first-order differential operators commuting as follows:

$$[\hat{K}_1,\hat{K}_2]_+=0, \quad \{\hat{K}_1+\hat{K}_2\}\Phi=0, \quad -\hat{K}_1\Phi=\hat{K}_2\Phi=k\Phi, \tag{14}$$

where k is a constant of separation, and

$$\hat{K}_1=-i(ab\gamma^0\partial_t+b\gamma^1\partial_x+abm)\gamma^1\gamma^0, \quad \hat{K}_2=-i\left(\gamma^2\partial_y+\frac{1}{c}\gamma^3\partial_z\right)\gamma^1\gamma^0, \tag{15}$$

$$\Psi=\gamma^1\gamma^0\Phi. \tag{16}$$

Let consider the suitable representation of Dirac matrices:[13]

$$\gamma^0=\begin{pmatrix}-i & 0\\ 0 & i\end{pmatrix}, \quad \gamma^k=\begin{pmatrix}0 & \sigma^k\\ \sigma^k & 0\end{pmatrix}, \quad k=1,2,3. \tag{17}$$

There exists a unitary transformation S that connects Dirac matrices and the spinor Φ in the diagonal gauge to Dirac matrices and the spinor Φ':

$$\gamma^\mu\to\gamma'^\mu=S^{-1}\gamma^\mu S, \quad \Phi\to\Phi'=S^{-1}\Phi, \tag{18}$$

such that the matrix S is given by

$$S=\tfrac{1}{2}(I+\gamma^1\gamma^2+\gamma^2\gamma^3+\gamma^3\gamma^1). \tag{19}$$

Substituting (18) into (14) and expressing γ'^{μ} in terms of γ^{μ}, we obtain

$$(ab\gamma^3\partial_t+b\gamma^0\partial_x+abm\gamma^3\gamma^0+ik)\Phi'=0, \tag{20}$$

$$\left(\gamma^1\gamma^3\gamma^0\partial_y+\frac{1}{c}\gamma^2\gamma^3\gamma^0\partial_z-ik\right)\Phi'=0. \tag{21}$$

Equation (21) splits into two first-order differential equations as

$$\left(\sigma^2\partial_y+\frac{ik_z}{c}\sigma^1-ik\right)\Phi_1'=0, \quad \left(-\sigma^2\partial_y-\frac{ik_z}{c}\sigma^1-ik\right)\Phi_2'=0, \tag{22}$$

where

$$\Phi'=\begin{pmatrix}\Phi_1'\\ \Phi_2'\end{pmatrix}. \tag{23}$$

Using the algebra of Pauli matrices and taking into account the form of (22), we obtain

$$\left(\partial_y+\frac{k_z}{c}\right)\chi_1-k\chi_2=0, \quad \left(\partial_y-\frac{k_z}{c}\right)\chi_2+k\chi_1=0. \tag{24}$$

It is easy from (22) to see that the spinor Φ' has the following structure:

$$\Phi'=\begin{pmatrix}U(t,x)\chi_1(y)\\ U(t,x)\chi_2(y)\\ V(t,x)\chi_1(y)\\ -V(t,x)\chi_2(y)\end{pmatrix}\exp(ik_z z). \tag{25}$$

Eliminating χ_1 and χ_2 from the system (24), respectively, we obtain

$$d_{yy}^2\chi_1+\left(-\frac{k_z}{c^2}d_y c-\frac{k_z^2}{c^2}+k^2\right)\chi_1=0, \quad d_{yy}^2\chi_2+\left(\frac{k_z}{c^2}d_y c-\frac{k_z^2}{c^2}+k^2\right)\chi_2=0. \tag{26}$$

To apply the method of the separation of variables, it is convenient to rewrite Eq. (20) as

$$\left[(a\gamma^0\partial_t+am)\gamma^3\gamma^0+\left(\gamma^0\partial_x+\frac{ik}{b}\right)\right]\Phi'=0. \tag{27}$$

The auxiliary function ξ defined by

$$\Phi'=\left[(a\gamma^0\partial_t+am)\gamma^3\gamma^0+\left(\gamma^0\partial_x-\frac{ik}{b}\right)\right]\xi, \tag{28}$$

and the substitution of Eq. (28) into Eq. (27) give

$$\left[(a\gamma^0\partial_t+am)(-a\gamma^0\partial_t+am)+\left(\gamma^0\partial_x+\frac{ik}{b}\right)\left(\gamma^0\partial_x-\frac{ik}{b}\right)\right]\xi=0. \tag{29}$$

Equation (29) reduces to

$$[\hat{K}_3,\hat{K}_4]_+=0, \quad \{\hat{K}_3+\hat{K}_4\}\xi=0, \tag{30}$$

with

$$\hat{K}_3=(a\gamma^0\partial_t+am)(-a\gamma^0\partial_t+am),\quad \hat{K}_4=\left(\gamma^0\partial_x+\frac{ik}{b}\right)\left(\gamma^0\partial_x-\frac{ik}{b}\right), \tag{31}$$

$$-\hat{K}_3\xi=\hat{K}_4\xi=\lambda^2\xi. \tag{32}$$

From Eq. (32), we can express ξ as follows

$$\xi=\begin{pmatrix}\Sigma_1(t)R_1(x)C_1(y,z)\\ \Sigma_1(t)R_1(x)C_2(y,z)\\ \Sigma_2(t)R_2(x)C_3(y,z)\\ \Sigma_2(t)R_2(x)C_4(y,z)\end{pmatrix}, \tag{33}$$

such that the spinor Φ' can be written as

$$\Phi'=A_0\begin{pmatrix}\Sigma_1(t)R_1(x)\chi_1(y)\\ \Sigma_1(t)R_1(x)\chi_2(y)\\ \Sigma_2(t)R_2(x)\chi_1(y)\\ -\Sigma_2(t)R_2(x)\chi_2(y)\end{pmatrix}\exp(ik_z z), \tag{34}$$

where $\Sigma_1(t)$, $\Sigma_2(t)$, $R_1(x)$ and $R_2(x)$ satisfy

$$\left(\partial_x+\frac{k}{b}\right)R_1=\lambda R_2,\quad \left(\partial_x+\frac{k}{b}\right)R_2=-\lambda R_1, \tag{35}$$

$$(a\partial_t+iam)\Sigma_2=-i\lambda\Sigma_1,\quad (a\partial_t-iam)\Sigma_1=-i\lambda\Sigma_2. \tag{36}$$

Eliminating R_1,R_2 from (35), and Σ_1,Σ_2 from (36), we have

$$\left(\partial_{xx}^2-\frac{k}{b^2}\partial_x b-\frac{k^2}{b^2}+\lambda^2\right)R_1=0,\quad \left(\partial_{xx}^2+\frac{k}{b^2}\partial_x b-\frac{k^2}{b^2}+\lambda^2\right)R_2=0, \tag{37}$$

$$\left(\partial_{tt}^2+m^2+\frac{\lambda^2}{a^2}\right)\Sigma_1=0,\quad \left(\partial_{tt}^2+m^2+\frac{\lambda^2}{a^2}\right)\Sigma_2=0. \tag{38}$$

III. EXACT SOLUTIONS OF DIRAC EQUATION

As already discussed in Sec. I, many works have been devoted to the problem of finding exact solutions to the Dirac equation in external fields and recently some exact solutions have been obtained using different techniques and methods.[2,4] One of the most effective and powerful tools in solving systems of partial differential equations is the method of separation of variables.[6] That is, of course, not just restricted to scalar equations such as the Hamilton–Jacobi, Helmholtz and Laplace–Beltrami equations of mathematical physics, but also applies to systems of partial differential equations such as the Dirac equation in flat or curved space–time. In the last case, this method allows us to reduce the problem to solving a system of ordinary differential equations.

The study of the behavior of relativistic particles obeying the Dirac equation in curved space, in particular in expanding universes, is of considerable importance in astrophysics and cosmology. Such investigations go back to Fock,[17] Tetrode,[18] Schrodinger,[19,20] and McVittie[21] and enable us to quantize the relativistic spin $-\frac{1}{2}$ electron field in curved background and study the effect of gravity in atomic spectra.[22] Thus, the Dirac equation in the zero-momentum limit has been solved by Isham and Nelson[23] who proposed a quantization and obtained a mass of the order of the Universe, so the equation has a different interpretation than the electron. Barut and Duru[2] have been interested in the behavior of the electrons and neutrino in curved spaces and obtained an exact solution for arbitrary momentum and mass, in spatially flat Robertson–Walker space–times. Parker[24] studied the pair creation of $-\frac{1}{2}$ particles (massive and massless) in Robertson–Walker universes. He showed that for massless neutrinos, as a result of conformal invariance, there was no pair creation. Here, we consider the four-dimensional nonfactorizable metric (2) and the final equations (37) and (38). To find exact solutions to these equations in terms of special functions is not always easy. It depends on the analytical expressions of the metric functions $a(t)$, $b(x)$ and $c(y)$. In this section, we restrict our analysis to some interesting particular cases corresponding to the most relevant situations in astrophysics and cosmology[2,17–21,24,25] such as the Robertson–Walker space–times or the Friedman–Robertson–Walker metric, which lead to exact solutions in terms of special functions.

(*i*) $c(y)=\beta y$, where β is arbitrary constant. Substituting this expression into (26), we obtain

$$d_{yy}^2\chi_1+\left(-\frac{k_z}{\beta y^2}-\frac{k_z^2}{\beta^2 y^2}+k^2\right)\chi_1=0,\quad d_{yy}^2\chi_2+\left(\frac{k_z}{\beta y^2}-\frac{k_z^2}{\beta^2 y^2}+k^2\right)\chi_2=0, \tag{39}$$

which have solutions

$$\chi_1=d_1 y^{1/2}\mathcal{C}_{k_z/\beta+1/2}(ky)\quad\text{and}\quad\chi_2=d_2 y^{1/2}\mathcal{C}_{k_z/\beta-1/2}(ky), \tag{40}$$

where $\mathcal{C}$ is the general solution of the cylindrical Bessel equation.[26] The relation between the constants d_1 and d_2 can be obtained in a straightforward way, by substituting (40) into (24).

(*ii*) $c(y)=\beta\exp(\alpha y)$, where α and β are arbitrary constants. The system (26) takes the form

$$d_{yy}^2\chi_1+\left(-\frac{k_z}{\beta\exp(\alpha y)}-\frac{k_z^2}{\beta^2\exp(2\alpha y)}+k^2\right)\chi_1=0, \tag{41}$$

$$d_{yy}^2\chi_2+\left(\frac{k_z}{\beta\exp(\alpha y)}-\frac{k_z^2}{\beta^2\exp(2\alpha y)}+k^2\right)\chi_2=0. \tag{42}$$

Making the change of variables

$$z=2\beta^{-1}k_z\alpha^{-1}\exp(-\alpha y),\quad \chi_{1,2}=z^{1/2}\tilde{\chi}_{1,2}, \tag{43}$$

Eqs. (41) and (42) take the form

$$d_{yy}^2\tilde{\chi}_1+\left(-\frac{1}{4}-\frac{1}{2z}+\frac{\frac{1}{4}+k^2/\alpha^2}{z^2}\right)\tilde{\chi}_1=0,\quad d_{yy}^2\tilde{\chi}_2+\left(-\frac{1}{4}+\frac{1}{2z}+\frac{\frac{1}{4}+k^2/\alpha^2}{z^2}\right)\tilde{\chi}_2=0. \tag{44}$$

Taking into account (43), we get the solutions[26]

$$\chi_1=d_1(\exp(-\alpha y))^{ik/\alpha}\exp(-pe^{-\alpha y})M\left(1+\frac{ik}{\alpha},1+\frac{2ik}{\alpha},2pe^{-\alpha y}\right), \tag{45}$$

$$\chi_2=d_2(\exp(-\alpha y))^{ik/\alpha}\exp(-pe^{-\alpha y})M\left(\frac{ik}{\alpha},1+\frac{2ik}{\alpha},2pe^{-\alpha y}\right), \tag{46}$$

$$p=k_z\beta^{-1}\alpha^{-1}, \tag{47}$$

where $M(a,b,z)$ is the confluent hypergeometric function.

(*iii*) $c(y)=\tan(\alpha y)$. In this case, we get

$$d_{yy}^2\chi_1+\left(-\frac{k_z\alpha+k_z^2}{\sin^2(\alpha y)}+k_z^2+k^2\right)\chi_1=0,\quad d_{yy}^2\chi_2+\left(\frac{k_z\alpha-k_z^2}{\sin^2(\alpha y)}+k_z^2+k^2\right)\chi_2=0. \tag{48}$$

Making the change of variables

$$u = \sin^2(\alpha y), \tag{49}$$

and putting

$$\chi_1 = \sin^{q_1}(\alpha y) f(y), \quad \chi_2 = \sin^{q_2}(\alpha y) g(y), \tag{50}$$

where

$$q_1 = \frac{1}{2} + \left(\frac{1}{4} + \frac{k_z}{\alpha} + \frac{k_z^2}{\alpha^2}\right)^{1/2}, \quad q_2 = \frac{1}{2} + \left(\frac{1}{4} - \frac{k_z}{\alpha} + \frac{k_z^2}{\alpha^2}\right)^{1/2}, \tag{51}$$

the system (48) becomes

$$u(1-u) d_{uu}^2 f + \left(\frac{1}{2} + q_1 - (1+q_1)u\right) d_u f - \left(\frac{q_1^2}{4} - \frac{k_z^2 + k^2}{4\alpha^2}\right) f = 0, \tag{52}$$

$$u(1-u) d_{uu}^2 g + \left(\frac{1}{2} + q_2 - (1+q_2)u\right) d_u g - \left(\frac{q_2^2}{4} - \frac{k_z^2 + k^2}{4\alpha^2}\right) g = 0. \tag{53}$$

The solutions of these equations can be expressed in terms of hypergeometric functions of Gauss $F(a,b,z)$.[26] So, taking into account Eq. (49), they are given by

$$\chi_1 = d_1 \sin^{q_1}(\alpha y) F\left(\frac{1}{2}q_1 + \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, \frac{1}{2}q_1 - \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, q_1 + \frac{1}{2}, \sin^2(\alpha y)\right), \tag{54}$$

$$\chi_2 = d_2 \sin^{q_2}(\alpha y) F\left(\frac{1}{2}q_2 + \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, \frac{1}{2}q_2 - \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, q_2 + \frac{1}{2}, \sin^2(\alpha y)\right). \tag{55}$$

(*iv*) $c(y) = \cot(\alpha y)$. The system (26) transforms to

$$d_{yy}^2 \chi_1 + \left(\frac{k_z \alpha - k_z^2}{\cos^2(\alpha y)} + k_z^2 + k^2\right)\chi_1 = 0, \quad d_{yy}^2 \chi_2 + \left(-\frac{k_z \alpha + k_z^2}{\cos^2(\alpha y)} + k_z^2 + k^2\right)\chi_2 = 0. \tag{56}$$

The solutions of (56) can be obtained from Eqs. (54) and (55) by making the identification

$$\alpha y \to \alpha y + \frac{\pi}{2}. \tag{57}$$

Thus, these solutions take the forms

$$\chi_1 = d_1 \cos^{q_1}(\alpha y) F\left(\frac{1}{2}q_1 + \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, \frac{1}{2}q_1 - \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, q_1 + \frac{1}{2}, \cos^2(\alpha y)\right), \tag{58}$$

$$\chi_2 = d_2 \cos^{q_2}(\alpha y) F\left(\frac{1}{2}q_2 + \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, \frac{1}{2}q_2 - \frac{(k_z^2+k^2)^{1/2}}{2\alpha}, q_2 + \frac{1}{2}, \cos^2(\alpha y)\right), \tag{59}$$

where

$$q_1 = \frac{1}{2} + \left(\frac{1}{4} - \frac{k_z}{\alpha} + \frac{k_z^2}{\alpha^2}\right)^{1/2}, \quad q_2 = \frac{1}{2} + \left(\frac{1}{4} + \frac{k_z}{\alpha} + \frac{k_z^2}{\alpha^2}\right)^{1/2}. \tag{60}$$

(*v*) $c(y) = \tanh(\alpha y)$. In this case, the system (26) rewrites as

$$d_{yy}^2\chi_1+\left(-\frac{k_z\alpha+k_z^2}{\sinh^2(\alpha y)}-k_z^2+k^2\right)\chi_1=0,\quad d_{yy}^2\chi_2+\left(\frac{k_z\alpha-k_z^2}{\sin^2(\alpha y)}-k_z^2+k^2\right)\chi_2=0. \tag{61}$$

Making the change of variables

$$v=-\sinh^2(\alpha y), \tag{62}$$

and following the analog way as in (*iii*), we transform (61) to get

$$v(1-v)d_{vv}^2f+\left(\frac{1}{2}+q_1-(1+q_1)v\right)d_vf-\left(\frac{q_1^2}{4}-\frac{k_z^2-k^2}{4\alpha^2}\right)f=0, \tag{63}$$

$$v(1-v)d_{vv}^2g+\left(\frac{1}{2}+q_2-(1+q_2)v\right)d_vg-\left(\frac{q_2^2}{4}-\frac{k_z^2-k^2}{4\alpha^2}\right)g=0. \tag{64}$$

The solutions of these equations can be also expressed in terms of hypergeometric functions of Gauss $F(a,b,z)$. Thus, taking into account Eq. (62), they are given as

$$\chi_1=d_1\sinh^{q_1}(\alpha y)F\left(\frac{1}{2}q_1+\frac{(k_z^2-k^2)^{1/2}}{2\alpha},\frac{1}{2}q_1-\frac{(k_z^2-k^2)^{1/2}}{2\alpha},q_1+\frac{1}{2},-\sinh^2(\alpha y)\right), \tag{65}$$

$$\chi_2=d_2\sinh^{q_2}(\alpha y)F\left(\frac{1}{2}q_2+\frac{(k_z^2-k^2)^{1/2}}{2\alpha},\frac{1}{2}q_2-\frac{(k_z^2-k^2)^{1/2}}{2\alpha},q_2+\frac{1}{2},-\sinh^2(\alpha y)\right). \tag{66}$$

(*vi*) $c(y)=\coth(\alpha y)$. The system (26) rewrites as

$$d_{yy}^2\chi_1+\left(\frac{k_z\alpha+k_z^2}{\cosh^2(\alpha y)}-k_z^2+k^2\right)\chi_1=0,\quad d_{yy}^2\chi_2+\left(\frac{-k_z\alpha+k_z^2}{\cosh^2(\alpha y)}-k_z^2+k^2\right)\chi_2=0. \tag{67}$$

We can obtain from Eqs. (65) and (66) the solutions of (67) by making the identification

$$\alpha y\rightarrow\alpha y+\frac{i\pi}{2}. \tag{68}$$

Thus, we get

$$\chi_1=d_1\cosh^{q_1}(\alpha y)F\left(\frac{1}{2}q_1+\frac{(k_z^2-k^2)^{1/2}}{2\alpha},\frac{1}{2}q_1-\frac{(k_z^2-k^2)^{1/2}}{2\alpha},q_1+\frac{1}{2},\cosh^2(\alpha y)\right), \tag{69}$$

$$\chi_2=d_2\cosh^{q_2}(\alpha y)F\left(\frac{1}{2}q_2+\frac{(k_z^2-k^2)^{1/2}}{2\alpha},\frac{1}{2}q_2-\frac{(k_z^2-k^2)^{1/2}}{2\alpha},q_2+\frac{1}{2},\cosh^2(\alpha y)\right), \tag{70}$$

where q_1 and q_2 are given by (51).

(*vii*) $c(y)=\sin(\alpha y)$. The system (24) becomes

$$\left(d_y+\frac{k_z}{\sin(\alpha y)}\right)\chi_1-k\chi_2=0\quad\text{and}\quad\left(d_y-\frac{k_z}{\sin(\alpha y)}\right)\chi_2+k\chi_1=0. \tag{71}$$

It is enough to consider the solutions of the system (71) when k and k_z are positive; the other three cases can be obtained by interchanging the roles of χ_1 and χ_2. Considering the ansatz

$$\chi_1=(\sin(\alpha y))^{k_z/\alpha}\sin\left(\frac{\alpha y}{2}\right)f(y)\quad\text{and}\quad\chi_2=(\sin(\alpha y))^{k_z/\alpha}\cos\left(\frac{\alpha y}{2}\right)g(y), \tag{72}$$

and making the change of variables

$$u=\cos(\alpha y), \tag{73}$$

we obtain the differential equations

$$(1-u^2)d_{uu}^2 f+\left(-1-2\left(1+\frac{k_z}{\alpha}\right)u\right)d_u f-\left(\left(\frac{k_z}{\alpha}+\frac{1}{2}\right)^2-\frac{k_z^2}{\alpha^2}\right)f=0, \tag{74}$$

$$(1-u^2)d_{uu}^2 g+\left(1-2\left(1+\frac{k_z}{\alpha}\right)u\right)d_u g-\left(\left(\frac{k_z}{\alpha}+\frac{1}{2}\right)^2-\frac{k_z^2}{\alpha^2}\right)g=0, \tag{75}$$

whose solutions are given in terms of Jacobi polynomials $P_n^{(\alpha,\beta)}$.[26] Thus, taking into account Eqs. (72) and (73), we get

$$\chi_1=d_1(\sin(\alpha y))^{k_z/\alpha}\sin\left(\frac{\alpha y}{2}\right)P_n^{(k_z/\alpha+1/2,k_z/\alpha-1/2)}(\cos(\alpha y)), \tag{76}$$

$$\chi_2=d_2(\sin(\alpha y))^{k_z/\alpha}\cos\left(\frac{\alpha y}{2}\right)P_n^{(k_z/\alpha-1/2,k_z/\alpha+1/2)}(\cos(\alpha y)), \tag{77}$$

where n reads as

$$n=\frac{k}{\alpha}-\frac{k_z}{\alpha}-\frac{1}{2}. \tag{78}$$

(*viii*) $c(y)=\cos(\alpha y)$. The system (37) becomes

$$\left(d_y+\frac{k_z}{\cos(\alpha y)}\right)\chi_1-k\chi_2=0 \quad \text{and} \quad \left(d_y-\frac{k_z}{\cos(\alpha y)}\right)\chi_2+k\chi_1=0. \tag{79}$$

Their solutions can be obtained by analogy with (76) and (77), making the identification

$$\alpha y\to\alpha y-\frac{\pi}{2}, \tag{80}$$

such that

$$\chi_1=d_1(-\cos(\alpha y))^{k_z/\alpha}\left[\sin\left(\frac{\alpha y}{2}\right)-\cos\left(\frac{\alpha y}{2}\right)\right]P_n^{(k_z/\alpha+1/2,k_z/\alpha-1/2)}(\sin(\alpha y)), \tag{81}$$

$$\chi_2=d_2(-\cos(\alpha y))^{k_z/\alpha}\left[\cos\left(\frac{\alpha y}{2}\right)+\sin\left(\frac{\alpha y}{2}\right)\right]P_n^{(k_z/\alpha-1/2,k_z/\alpha+1/2)}(\sin(\alpha y)), \tag{82}$$

where n is given by (78).

(*ix*) $c(y)=\sinh(\alpha y)$. Substituting it into (24), we obtain

$$\left(d_y+\frac{k_z}{\sinh(\alpha y)}\right)\chi_1-k\chi_2=0 \quad \text{and} \quad \left(d_y-\frac{k_z}{\sinh(\alpha y)}\right)\chi_2+k\chi_1=0. \tag{83}$$

Considering the ansatz

$$\chi_1=(\sinh(\alpha y))^{k_z/\alpha}\sinh\left(\frac{\alpha y}{2}\right)f(y), \quad \chi_2=(\sinh(\alpha y))^{k_z/\alpha}\cosh\left(\frac{\alpha y}{2}\right)g(y), \tag{84}$$

and making the change of variables

$$u=-\cosh(\alpha y), \tag{85}$$

we obtain the differential equations

$$(1-u^2)d_{uu}^2 f+\left(1-2\left(1+\frac{k_z}{\alpha}\right)u\right)d_u f-\left(\left(\frac{k_z}{\alpha}+\frac{1}{2}\right)^2+\frac{k_z^2}{\alpha^2}\right)f=0, \tag{86}$$

$$(1-u^2)d_{uu}^2 g+\left(-1-2\left(1+\frac{k_z}{\alpha}\right)u\right)d_u g-\left(\left(\frac{k_z}{\alpha}+\frac{1}{2}\right)^2+\frac{k_z^2}{\alpha^2}\right)g=0, \tag{87}$$

whose solutions are also given in terms of Jacobi polynomials $P_n^{(\alpha,\beta)}$, taking into account Eqs. (84) and (85), as

$$\chi_1=d_1(\sinh(\alpha y))^{k_z/\alpha}\sinh\left(\frac{\alpha y}{2}\right)P_n^{(k_z/\alpha-1/2,k_z/\alpha+1/2)}(-\cosh(\alpha y)), \tag{88}$$

$$\chi_2=d_2(\sinh(\alpha y))^{k_z/\alpha}\cosh\left(\frac{\alpha y}{2}\right)P_n^{(k_z/\alpha+1/2,k_z/\alpha-1/2)}(-\cosh(\alpha y)), \tag{89}$$

where n reads as

$$n=\frac{ik}{\alpha}-\frac{k_z}{\alpha}-\frac{1}{2}. \tag{90}$$

(*x*) $c(y)=\cosh(\alpha y)$. Here the system (24) becomes

$$\left(d_y+\frac{k_z}{\cosh(\alpha y)}\right)\chi_1-k\chi_2=0 \quad \text{and} \quad \left(d_y-\frac{k_z}{\cosh(\alpha y)}\right)\chi_2+k\chi_1=0. \tag{91}$$

As in the previous case, using Eqs. (88) and (89) and making the identifications

$$\alpha y\to\alpha y-\frac{i\pi}{2}, \quad k_z\to ik_z, \tag{92}$$

we get

$$\chi_1=d_1(-i\cosh(\alpha y))^{ik_z/\alpha}\left[\sinh\left(\frac{\alpha y}{2}\right)-i\cosh\left(\frac{\alpha y}{2}\right)\right]P_n^{(ik_z/\alpha-1/2,ik_z/\alpha+1/2)}(i\sinh(\alpha y)), \tag{93}$$

$$\chi_2=d_2(-i\cosh(\alpha y))^{ik_z/\alpha}\left[\cosh\left(\frac{\alpha y}{2}\right)-i\sinh\left(\frac{\alpha y}{2}\right)\right]P_n^{(ik_z/\alpha+1/2,ik_z/\alpha-1/2)}(i\sinh(\alpha y)), \tag{94}$$

where n is given by

$$n=\frac{ik}{\alpha}-\frac{ik_z}{\alpha}-\frac{1}{2}. \tag{95}$$

The metrics associated to cases (*vii*)–(*x*) can be identified as particular cases of Robertson–Walker spaces–times.[3]

When n is not an integer value, we have to express the Jacobi polynomials in terms of the Gauss hypergeometric functions, by means of the relation[26]

$$P_n^{(\alpha,\beta)}(x)=\frac{\Gamma(n+\alpha+1)}{\Gamma(n+1)\Gamma(\alpha+1)}F\left(-n,n+\alpha+\beta+1,\alpha+1,\frac{1-x}{2}\right). \tag{96}$$

Now we proceed to analyze the system of Eq. (35). These equations can be solved in terms of special functions for the following cases.

(*i*) $b(x)=\beta x$, where β is an arbitrary constant. Substituting this expression into (37), we arrive at

$$d_{xx}^2R_1+\left(-\frac{k}{\beta x^2}-\frac{k^2}{\beta^2x^2}+\lambda^2\right)R_1=0 \quad \text{and} \quad d_{xx}^2R_2+\left(\frac{k}{\beta x^2}-\frac{k^2}{\beta^2x^2}+\lambda^2\right)R_2=0, \tag{97}$$

which have solutions

$$R_1=c_0x^{1/2}\mathcal{C}_{k/\beta+1/2}(\lambda x) \quad \text{and} \quad R_2=c_0x^{1/2}\mathcal{C}_{k/\beta-1/2}(\lambda x), \tag{98}$$

where $\mathcal{C}$ is the general solution of the cylindrical Bessel equation.

(*ii*) $b(x)=\beta\exp(\alpha x)$, where α and β are arbitrary constants. The system (37) leads to

$$d_{xx}^2R_1+\left(-\frac{k}{\beta\exp(\alpha x)}-\frac{k^2}{\beta^2\exp(2\alpha x)}+\lambda^2\right)R_1=0, \tag{99}$$

$$d_{xx}^2R_2+\left(\frac{k}{\beta\exp(\alpha x)}-\frac{k^2}{\beta^2\exp(2\alpha x)}+\lambda^2\right)R_2=0. \tag{100}$$

Since Eqs. (99) and (100) have the same structure as the system (41) and (42), we obtain

$$R_1=c_1(\exp(-\alpha x))^{i\lambda/\alpha}\exp(-pe^{-\alpha x})M\left(1+\frac{i\lambda}{\alpha},1+\frac{2i\lambda}{\alpha},2pe^{-\alpha x}\right), \tag{101}$$

$$R_2=c_1(\exp(-\alpha x))^{i\lambda/\alpha}\exp(-pe^{-\alpha x})M\left(\frac{i\lambda}{\alpha},1+\frac{2i\lambda}{\alpha},2pe^{-\alpha x}\right), \tag{102}$$

where $M(a,b,z)$ is the confluent hypergeometric function, and p reads as

$$p=k\beta^{-1}\alpha^{-1}. \tag{103}$$

(*iii*) $b(x)=\sin(\alpha x)$. We have the new system

$$\left(d_x+\frac{k}{\sin(\alpha x)}\right)R_1-kR_2=0 \quad \text{and} \quad \left(d_x-\frac{k}{\sin(\alpha x)}\right)R_2+kR_1=0, \tag{104}$$

from which we deduce

$$R_1=c_2(\sin(\alpha x))^{k/\alpha}\sin\left(\frac{\alpha x}{2}\right)P_n^{(k/\alpha+1/2,k/\alpha-1/2)}(\cos(\alpha x)), \tag{105}$$

$$R_2=c_2(\sin(\alpha x))^{k/\alpha}\cos\left(\frac{\alpha x}{2}\right)P_n^{(k/\alpha-1/2,k/\alpha+1/2)}(\cos(\alpha x)), \tag{106}$$

where n reads as

$$n=\frac{\lambda}{\alpha}-\frac{k}{\alpha}-\frac{1}{2}. \tag{107}$$

(*iv*) $b(x)=\cos(\alpha x)$. Substituting this expression into (35), and making the identification with (105) and (106),

$$\alpha x \rightarrow \alpha x - \frac{\pi}{2}, \tag{108}$$

the solutions take the forms

$$R_1 = c_3(-\cos(\alpha x))^{k/\alpha}\left[\sin\left(\frac{\alpha x}{2}\right) - \cos\left(\frac{\alpha x}{2}\right)\right] P_n^{(k/\alpha+1/2, k/\alpha-1/2)}(\sin(\alpha x)), \tag{109}$$

$$R_2 = c_3(-\cos(\alpha x))^{k/\alpha}\left[\cos\left(\frac{\alpha x}{2}\right) + \sin\left(\frac{\alpha x}{2}\right)\right] P_n^{(k/\alpha-1/2, k/\alpha+1/2)}(\sin(\alpha x)), \tag{110}$$

where n is given by (107).

(*v*) $b(x)=\sinh(\alpha x)$. The system (35) gives

$$\left(d_x + \frac{k}{\sinh(\alpha x)}\right) R_1 - kR_2 = 0 \quad \text{and} \quad \left(d_x - \frac{k}{\sinh(\alpha x)}\right) R_2 + kR_1 = 0. \tag{111}$$

As from the system (83), we obtain

$$R_1 = c_4(\sinh(\alpha x))^{k/\alpha} \sinh\left(\frac{\alpha x}{2}\right) P_n^{(k/\alpha-1/2, k/\alpha+1/2)}(-\cosh(\alpha x)), \tag{112}$$

$$R_2 = c_4(\sinh(\alpha x))^{k/\alpha} \cosh\left(\frac{\alpha x}{2}\right) P_n^{(k/\alpha+1/2, k/\alpha-1/2)}(-\cosh(\alpha x)), \tag{113}$$

where n is given by

$$n = \frac{i\lambda}{\alpha} - \frac{k}{\alpha} - \frac{1}{2}. \tag{114}$$

(*vi*) $b(x)=\cosh(\alpha x)$. Substituting this expression into (35), and making the identification with (112) and (113),

$$\alpha x \rightarrow \alpha x - \frac{i\pi}{2}, \quad k \rightarrow ik, \tag{115}$$

the solutions take the form

$$R_1 = c_5(-i\cosh(\alpha x))^{ik/\alpha}\left[\sinh\left(\frac{\alpha x}{2}\right) - i\cosh\left(\frac{\alpha x}{2}\right)\right] P_n^{(ik/\alpha-1/2, ik/\alpha+1/2)}(i\sinh(\alpha x)), \tag{116}$$

$$R_2 = c_5(-i\cosh(\alpha x))^{ik/\alpha}\left[\cosh\left(\frac{\alpha x}{2}\right) - i\sinh\left(\frac{\alpha x}{2}\right)\right] P_n^{(ik/\alpha+1/2, ik/\alpha-1/2)}(i\sinh(\alpha x)), \tag{117}$$

where n is given by

$$n = \frac{i\lambda}{\alpha} - \frac{ik}{\alpha} - \frac{1}{2}. \tag{118}$$

For the cases (vii) $b(x)=\tan(\alpha x)$, $(viii)$ $b(x)=\cot(\alpha x)$, (ix) $b(x)=\tanh(\alpha x)$ and (x) $b(x)=\tan(\alpha x)$, we obtain the same solutions as in (iii) $c(y)=\tan(\alpha y)$, (iv) $c(y)=\cot(\alpha y)$, (v) $c(y)=\tanh(\alpha y)$ and (vi) $c(y)=\coth(\alpha y)$ by changing

$$k\rightarrow\lambda \quad \text{and} \quad k_z\rightarrow k. \tag{119}$$

Note that the different cases (iii) $b(x)=\sin(\alpha x)$, (v) $b(x)=\sinh(\alpha x)$, (vii) $c(y)=\sin(\alpha y)$ and (ix) $c(y)=\sinh(\alpha y)$ correspond to the Friedman–Robertson–Walker metric,[12] where we have introduced an arbitrary constant α.

Now we proceed to analyze the systems (38). We consider three special models for expansion.[2]

(i) $a(t)=\beta t$. This model also considered by Schrödinger[2] here leads to

$$\left(d_{tt}^2+m^2+\frac{\lambda^2}{\beta^2t^2}\right)\Sigma_1=0. \tag{120}$$

Changing the variable t to

$$z=2imt, \tag{121}$$

Eq. (120) becomes

$$\left(d_{zz}^2-\frac{1}{4}+\frac{\lambda^2}{\beta^2z^2}\right)\Sigma_1=0. \tag{122}$$

We recognize the Whittaker differential equation which gives the solution, by taking into account (121), as follows

$$\Sigma_1=b_1W_{k,\mu}(2imt), \quad \text{where} \quad k=0, \quad \mu=\sqrt{\frac{1}{4}-\frac{\lambda^2}{\beta^2}}. \tag{123}$$

Using the Whittaker function identities[26]

$$zW'_{k,\mu}(z)=\left(\frac{z}{2}-k\right)W_{k,\mu}(z)-W_{k+1,\mu}(z), \tag{124}$$

from (36), we get

$$\Sigma_2=-\frac{ib_1\beta}{\lambda}W_{1,\mu}(2imt). \tag{125}$$

(ii) $a(t)=e^{\alpha t}$, an inflationary universe. Substituting the above expression into (38), we arrive at

$$(d_{tt}^2+m^2+\lambda^2e^{-2\alpha t})\Sigma_1=0. \tag{126}$$

Introducing the new coordinates z by

$$z=\frac{\lambda}{\alpha}e^{-\alpha t}, \tag{127}$$

Eq. (126) becomes the Bessel equation

$$(z^2d_{zz}^2+zd_z+z^2+m^2)\Sigma_1=0. \tag{128}$$

Taking into account Eq. (127), we obtain

$$\Sigma_1 = b_2 J_{im}\left(\frac{\lambda}{\alpha} e^{-\alpha t}\right). \tag{129}$$

Using the Bessel function identities[26]

$$\frac{d}{dz} J_\nu(z) = -J_{\nu+1}(z) + \frac{\nu}{z} J_\nu(z), \tag{130}$$

and from (36) we get

$$\Sigma_2 = \frac{ib_2}{\lambda}\left(\lambda J_{im+1}\left(\frac{\lambda}{\alpha} e^{-\alpha t}\right) - im(\alpha+1) J_{im}\left(\frac{\lambda}{\alpha} e^{-\alpha t}\right)\right). \tag{131}$$

(*iii*) $a(t) = \beta\sqrt{t}$, a model of a radiation-dominated universe. Substituting the above expression into (38), we obtain

$$\left(d_{tt}^2 + m^2 + \frac{\lambda^2}{\beta^2 t}\right)\Sigma_1 = 0. \tag{132}$$

Making a change of variables

$$t = iz^2, \tag{133}$$

Eq. (132) takes the form

$$\left(d_{zz}^2 - \frac{1}{z} d_z - 4m^2 z^2 + \frac{4i\lambda^2}{\beta^2}\right)\Sigma_1 = 0. \tag{134}$$

Putting

$$\Sigma_1(z) = e^{-v/2} F(v) \quad \text{and} \quad v = 2mz^2, \tag{135}$$

we can write Eq. (134) as follows:

$$\left(v d_{vv}^2 - v d_v + \frac{i\lambda^2}{m\beta^2}\right) F = 0. \tag{136}$$

Taking into account (133) and (135), the solution of (136) is given by the confluent hypergeometric function ${}_1F_1(a,b,z)$:

$$\Sigma_1(t) = b_3 e^{imt}\, {}_1F_1\left(0, \frac{-i\lambda^2}{m\beta^2}, -2imt\right). \tag{137}$$

Using the confluent hypergeometric function identities[26]

$$b\ {}_1F_1'(a,b,z) = b\ {}_1F_1(a,b,z) - (b-a)\, {}_1F_1(a,b+1,z), \tag{138}$$

and from (36) we get

$$\Sigma_2 = \frac{\beta b_3}{\sqrt{4i\lambda^2}} e^{imt}(\sqrt{4it} - im - 1)\, {}_1F_1\left(0, \frac{-i\lambda^2}{m\beta^2}, -2imt\right) - \frac{\lambda b_3}{\beta m^2}\sqrt{\frac{t}{4i}}\, {}_1F_1\left(0, \frac{-i\lambda^2}{m\beta^2} + 1, -2imt\right). \tag{139}$$

The different cases (*i*), (*ii*) and (*iv*) are analyzed by Barut *et al.*,[2] considering $b = 1$ and $c = 1$.

IV. SOLUTION OF THE WEYL EQUATION

Neutrinos in the nonfactorizable metric (2) are described by the Weyl equation which corresponds to the massless limit of the Dirac equation. The chirality condition, in the representation (17), reads

$$(1-i\gamma_5)\tilde{\Psi}'=0, \quad \gamma_5=\gamma_0\gamma_1\gamma_2\gamma_3. \tag{140}$$

The Weyl spinor can be expressed in terms of the solution of the Dirac equation as follows:

$$(1+i\gamma_5)\tilde{\Psi}'=\Psi_w. \tag{141}$$

The matrix γ_5 commutes with the matrix transformation S. Therefore, we can write

$$(1+i\gamma_5)\tilde{\Psi}=\Psi_0. \tag{142}$$

Taking into account (12), (16) and (34), the relation (142) takes the form

$$\Psi_0=A_0 a^{-3/2}b^{-1}c^{-1/2}e^{ik_z z}\begin{pmatrix} h \\ ih \end{pmatrix}, \tag{143}$$

where

$$h=\begin{pmatrix} (i-1)(\Sigma_2 R_1+i\Sigma_1 R_2)\chi_1-(1+i)(\Sigma_2 R_1-i\Sigma_1 R_2)\chi_2 \\ (i-1)(\Sigma_2 R_1+i\Sigma_1 R_2)\chi_1+(1+i)(\Sigma_2 R_1-i\Sigma_1 R_2)\chi_2 \end{pmatrix}. \tag{144}$$

The solution in this gauge can be obtained applying the unitary transformation (18) to the spinor (143).

In this paper, we have used the algebraic method of separation of variables to show the possibility of separating Dirac equation in a nonfactorizable metric in terms of two commuting first-order differential operators. We have given analytical solutions in terms of special functions, setting conditions on the metric functions $a(t)$, $b(x)$ and $c(y)$. Some particular cases correspond to the Friedman–Robertson–Walker metric[12] and to the Robertson–Walker spaces–times.[3]

ACKNOWLEDGMENTS

The authors thank the referee for his positive comments on the manuscript, and are grateful for his suggestion to emphasize better the physics motivations for their study. They also thank Professor Jan Govaerts of Unité FYNU, Université Catholique de Louvain, Belgium, for useful discussions, and gratefully acknowledge the UNB/CUD-CIUF-IMSP*I* Belgian Cooperation for financial support.

[1] D. Brill and J. A. Wheeler, Rev. Mod. Phys. **29**, 465 (1987).
[2] A. O. Barut and I. H. Duru, Phys. Rev. D **36**, 3705 (1987).
[3] V. M. Villalba and U. Percoco, J. Math. Phys. **31**, 715 (1990).
[4] V. M. Villalba, J. Math. Phys. **31**, 1483 (1990).
[5] V. M. Villalba, J. Math. Phys. **36**, 3332 (1995).
[6] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **30**, 2132 (1989).
[7] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **33**, 2093 (1992).
[8] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **33**, 4037 (1992).
[9] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **34**, 5037 (1993).
[10] G. V. Shishkin and I. E. Andrushkevich, Theor. Math. Phys. **70**, 204 (1987).
[11] G. V. Shishkin and W. D. Cabos, J. Math. Phys. **32**, 3184 (1991).
[12] M. Fels and N. Kamran, Proc. R. Soc. London, Ser. A **428**, 229 (1990).
[13] J. M. Jauch and F. Rohrlich, *Theory of Photons and Electrons* (Springer, New York, 1976).
[14] N. D. Birrel and P. C. W. Davies, *Quantum fields in curved space times* (Cambridge U.P., Cambridge, 1982).
[15] M. A. H. MacCallum, in *General Relativity, An Einstein Centenary Survey*, edited by S. W. Hawking and W. Israel (Cambridge U.P., Cambridge, 1979), pp. 533–580.
[16] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), pp. 842–871.

[17] V. Fock, Z. Phys. **53**, 592 (1928).
[18] H. Tetrode, Z. Phys. **50**, 336 (1928).
[19] E. Schrodinger, Preuss. akad. Wiss Berlin, Ber. **10**, 11 (1932).
[20] E. Schrodinger, Physic. **6**, 899 (1939); Proc. R. Ir. Acad. Sect. A, Math. Astron. Phys. Sci. **46A**, 9 (1940).
[21] G. C. McVittie, Mon. Not. R. Astron. Soc. **92**, 868 (1932).
[22] L. Parker, Phys. Rev. Lett. **44**, 1559 (1980).
[23] C. J. Isham and J. E. Nelson, Phys. Rev. D **10**, 3226 (1974).
[24] L. Parker, Phys. Rev. **183**, 1057 (1969); Phys. Rev. D **3**, 346 (1971).
[25] S. Weinberg, *Gravitation and Cosmology, Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972).
[26] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1964).

One-loop stress-tensor renormalization in curved background: The relation between ζ-function and point-splitting approaches, and an improved point-splitting procedure

Valter Moretti[a)]
Department of Mathematics, Trento University, and Istituto Nazionale di Fisica Nucleare, Gruppo Collegato di Trento, I-38050 Povo (TN), Italy

(Received 12 February 1999; accepted for publication 22 March 1999)

We conclude the rigorous analysis of a previous paper [V. Moretti, Commun. Math. Phys. **201**, 327 (1999)] concerning the relation between the (Euclidean) point-splitting approach and the local ζ-function procedure to renormalize physical quantities at one-loop in (Euclidean) Quantum Field Theory in curved space–time. The case of the stress tensor is now considered in general D-dimensional closed manifolds for positive scalar operators $-\Delta+V(x)$. Results obtained formally in previous works [in the case $D=4$ and $V(x)=\xi R(x)+m^2$] are rigorously proven and generalized. It is also proven that, in static Euclidean manifolds, the method is compatible with Lorentzian-time analytic continuations. It is proven that the result of the ζ-function procedure is the same obtained from an improved version of the point-splitting method which uses a particular choice of the term $w_0(x,y)$ in the Hadamard expansion of the Green's function, given in terms of heat-kernel coefficients. This version of the point-splitting procedure works for any value of the field mass m. If D is even, the result is affected by an arbitrary one-parameter class of (conserved in absence of external source) symmetric tensors, dependent on the geometry locally, and it gives rise to the general correct trace expression containing the renormalized field fluctuations as well as the conformal anomaly term. Furthermore, it is proven that, in the case $D=4$ and $V(x)=\xi R(x)+m^2$, the given procedure reduces to the Euclidean version of Wald's improved point-splitting procedure provided the arbitrary mass scale present in the ζ-function is chosen opportunely. It is finally argued that the found point-splitting method should work generally, also dropping the hypothesis of a closed manifold, and not depending on the ζ-function procedure. This fact is indeed checked in the Euclidean section of Minkowski space–time for $A=-\Delta+m^2$ where the method gives rise to the correct Minkowski stress tensor for $m^2\geqslant 0$ automatically. [S0022-2488(99)01008-7]

I. INTRODUCTION

In a previous paper,[1] we have considered the relationship between the ζ-function and the point-splitting procedures in renormalizing some physical quantities: effective Lagrangian, effective action, and field fluctuations. The more interesting quantity, namely, the stress tensor, is the object of the present paper. The aim of this paper is hence twofold. First we want to give a rigorous mathematical foundation as well as a generalization of several propositions contained in Ref. 2 where they have been stated without rigorous proof. This is a quite untrivial task because it involves an extension of the heat kernel theory considering the derivatives of its usual ''asymptotic'' expansion. As we shall see shortly, this is the core of all the analyticity properties of the generalized tensorial ζ-functions involved in the stress-tensor renormalization procedure.

[a)]Electronic mail: moretti@science.unitn.it

Second, we want to study the relation between our technique and the more usual point-splitting procedure in depth. This is another open issue after the appearance of Ref. 2. We know, through practical examples, that these two approaches agree essentially in several concrete cases, but up to now, no general proof of this fact has been given. Anyhow, it was conjectured by Wald[3] that, in general, these two approaches should lead to the same results. The extension of the ζ-function approach to the stress tensor has been introduced in Ref. 2 formally. This paper contains a proof of mathematical consistence of the method in a generalized case as well as a general proof of the agreement between the two approaches under our hypotheses on the manifold and the field operator.

It is a well-known fact that the point-splitting procedure faces some difficulties in the case of a field which is massless. Indeed, in such a case, one cannot make use of the Schwinger–deWitt algorithm to fix w_0 in the Hadamard expansion[3,4] and, at least in the massless conformally coupled case, the point-splitting procedure has been improved in order to get both the conformal anomaly and the conservation of the renormalized stress tensor.[3] Recently, Wald has argued that such an improved procedure, which picks out $w_0 \equiv 0$, can be generalized in more general cases.[5] Differently from the point-splitting procedure, the local ζ-function approach seems to work without to distinguish between different values of mass and coupling with the curvature. This fact makes more intriguing the issue of the relation between the two procedures.

This paper is organized as follows. In the first part, we shall recall the main features of the classical theory of the stress tensor to the reader and we shall introduce the main ideas concerning the renormalization of the stress tensor via the ζ-function. In a second part, first we shall develop further the theory of the heat-kernel expansion in order to build up the theory of the ζ-function of the stress tensor. All the work is developed in a closed D-dimensional manifold for a quite general Euclidean motion operator of a real scalar field. Successively, we shall state and prove several theorems concerning generalizations of several mathematical conjecture employed in Ref. 2. The final part of this work is devoted to investigating the relation between the two considered techniques and to introduce a generalized point-splitting procedure. Indeed, within that part, we shall give a proof of the agreement of the two approaches, introducing an improved point-splitting procedure which is quite similar and generalizes that pointed out in Refs. 3 and 5. We shall see that our prescription gives all the expected results (it gives the trivial stress tensor in Minkowski space–time, the conformal anomaly, and a conserved stress tensor, in general, producing agreement with the result of the field fluctuations renormalization). A final summary ends the work. In the Appendix, the proofs of some theorems and lemmata are reported.

II. PRELIMINARIES

Within this section, we state the general mathematical hypotheses we shall deal with and, very quickly, we review the main physical ideas concerning the classical stress tensor and its one-loop renormalization via point-splitting[4–6] and via local ζ-function.[2]

We assume all the definitions and theorems given in Ref. 1 and we shall refer to those definitions and theorems throughout all parts of this work.

A. General hypotheses and notations

The hypotheses we shall deal with in this work are the same as in Ref. 1. Therefore, from now on, $\mathcal{M}$ is a Hausdorff, connected, oriented, C^∞ Riemannian D-dimensional manifold. We suppose also that $\mathcal{M}$ is compact without boundary (namely, is ''closed''). Concerning the operators, we shall consider real elliptic differential operators with the Schrödinger form ''Laplace–Beltrami operator plus potential''

$$A' = -\Delta + V : C^\infty(\mathcal{M}) \to L^2(\mathcal{M}, d\mu_g), \tag{1}$$

where, locally, $\Delta = \nabla_a \nabla^a$, and ∇ means the covariant derivative associated to the metric connection, $d\mu_g$ is the Borel measure induced by the metric, and V is a *real* function belonging to

$C^\infty(\mathcal{M})$. We assume that A' is bounded below by some $C \geq 0$ (namely A is positive). (See sufficient conditions in Ref. 1.) These are the **general hypotheses** which we shall refer to throughout the paper.

Moreover, in the most part of the theorems, we shall use also the fact that the *injectivity radius* of the manifold r is strictly positive in closed manifolds (see Ref. 1 for further comments on this point).

As general remarks, we remind the reader that, as in the previous work, ''holomorphic'' and ''analytic'' are synonyms throughout this paper, natural units $\hbar = c = 1$ are used, and the symbol A indicates the only self-adjoint extension of the essentially self-adjoint operator A'. In the practice, as seen in Ref. 1, A coincides with the Friedrichs self-adjoint extension of A'. Also, R indicates the scalar curvature. Moreover, the symbol $\sigma(x,y)$ means one-half the squared geodesical distance of x from y which is continuous everywhere and C^∞ in any convex normal neighborhood.

Concerning derivative operators, we shall employ the notations in a fixed local coordinate system,

$$D_x^\alpha := \left. \frac{\partial^{|\alpha|}}{\partial x^{1\alpha_1} \cdots \partial x^{D\alpha_D}} \right|_x , \tag{2}$$

where the *multindex* α is defined by $\alpha := (\alpha_1 \ldots, \alpha_D)$, $\alpha_i \geq 0$ is any natural number ($i = 1,...,D$), and $|\alpha| := \alpha_1 + \cdots + \alpha_D$. Moreover, n_k will indicate the multindex $(0,\cdots 0,n,0,\cdots 0)$ where the only nonvanishing number is $n \in \mathbb{N}$ which takes the kth position.

Concerning the definitions of the metric-connection symbols and curvature tensors, we shall follow the notations and conventions employed in Ref. 1 which are the same employed in Ref. 6, either for Riemannian or Lorentzian signature.

B. Physical background and classical definitions

All quantities related to A' we have considered in Ref. 1 and the averaged stress tensor we consider here, for $D = 4$, appears in (Euclidean) Quantum Field Theory (QFT) in curved background and concerns the theory of quasifree scalar fields. In several concrete cases of QFT, the form of $V(x)$ is $m^2 + \xi R(x)$, where m^2 is the square mass of the considered field, R is the scalar curvature of the manifold, and ξ is a real parameter. As usual the *conformal coupling* is defined by[4,6,7]

$$\xi_D := (D-2)/[4(D-1)]. \tag{3}$$

Similarly to the physical quantities considered in Ref. 1, also the stress tensor is formally obtained from the Euclidean functional integral

$$Z[A'] := \int \mathcal{D}\phi e^{-(1/2)\int_{\mathcal{M}} \phi A'[\mathbf{g}]\phi d\mu_g} =: e^{-S_{\text{eff}}[A']}. \tag{4}$$

Here S_{eff} is the (Euclidean) effective action of the field. (Here, we use the opposite sign conventions in defining the effective action and thus the stress tensor, with respect the conventions employed in Ref. 2. Our conventions are the same as used in Ref. 8.)

The integral above can be considered as a partition function of a field in a particular quantum state corresponding to a canonical ensemble.[8,9] The direct physical interpretation as a partition function should work provided the manifold has a static Lorentzian section obtained by analytically continuing some global temporal coordinate $x^0 = \tau$ of some global chart into imaginary values $\tau \to it$ and considering (assuming that they exist) the induced continuations of the metric and relevant quantities. It is required also that ∂_τ is a global Killing field of the Riemannian manifold generated by an isometry group S_1. Finally, it is required that ∂_τ can be continued into a (generally local) timelike Killing field ∂_t in the Lorentzian section (see Refs. 9 and 8). Then one assumes that $k_B \beta$ is the inverse of the temperature of the canonical ensemble quantum state, β

being the period of the coordinate τ. The limit case of vanishing temperature is also considered and, in that case, the manifold cannot be compact. Similar interpretations hold for the (analytic continuations of) the stress tensor.

Formally[4,8–11] we have $Z[A',\mathbf{g}]:=[\det(A'/\mu^2)]^{-1/2}$, where our definition of the determinant of the operator A' is given by the ζ-function approach[4,8–11] as pointed out in Ref. 1. The scale μ^2 present in the determinant is necessary for dimensional reasons[9] and plays a central role in the ζ-function interpretation of the determinant and in the consequent theory. Such a scale introduces an ambiguity which remains in the finite renormalization parts of the renormalized quantities and, dealing with the renormalization of the stress tensor within the semiclassical approach to the quantum gravity, it determines the presence of quadratic-curvature terms in effective Einstein equations.[2] Similar results are discussed in Refs. 3–6 employing other renormalization procedures (point-splitting).

Coming to the (Euclidean) *classical* stress tensor $T_{ab}(x)$, it is defined (e.g., see Ref. 7) as the locally quadratic form of the field obtained by the usual functional derivative once the field ϕ is fixed:

$$T_{ab}[\phi,\mathbf{g}](x):=\frac{2}{\sqrt{g}}\,\frac{\delta}{\delta g_{ab}(x)}\left(\frac{1}{2}\int_{\mathcal{I}}\sqrt{g(x)}\,\phi A'[\mathbf{g}]\phi\, d^Dx\right). \tag{5}$$

This functional derivative can be rigorously understood in terms of a Gâteaux derivative for functionals on real $C^\infty(\mathcal{I})$ *symmetric* tensor fields g_{ab} and the integration above is performed in the open set $\mathcal{I}$ containing x where the considered coordinate system is defined. Equation (5) means that, and this is the rigorous definition of the *symmetric* tensor field $T_{ab}[\phi,\mathbf{g}](x)$, for any C^∞ *symmetric* tensor field h_{ab} with compact support contained in $\mathcal{I}$

$$2\,\frac{d}{d\alpha}\bigg|_{\alpha=0} S_{\mathcal{I}}[\mathbf{g}+\alpha\mathbf{h}]=\frac{1}{2}\int_{\mathcal{I}}\sqrt{g(x)}\,T_{ab}[\phi,\mathbf{g}](x)h^{ab}(x)\,d^Dx, \tag{6}$$

where

$$S_{\mathcal{I}}[\phi,\mathbf{g}]:=\frac{1}{2}\int_{\mathcal{I}}\sqrt{g(x)}\,\phi A'[\mathbf{g}]\phi\, d^Dx. \tag{7}$$

In the case

$$A'=-\Delta+m^2+\xi R(x)+V'(x), \tag{8}$$

V' being a C^∞ function which does not depend on the metric, a direct computation of $T_{ab}(x)$ through this procedure gives

$$\begin{aligned}T_{ab}[\phi,\mathbf{g}](x)={}&\nabla_a\phi(x)\nabla_b\phi(x)-\tfrac{1}{2}g_{ab}(x)[\nabla_c\phi(x)\nabla^c\phi(x)+(m^2+V'(x))\,\phi^2(x)]\\&+\xi[(R_{ab}(x)-\tfrac{1}{2}g_{ab}(x)R(x))\phi^2(x)+g_{ab}(x)\nabla_c\nabla^c\phi^2-\nabla_a\nabla_b\phi^2(x)].\end{aligned} \tag{9}$$

As is well known, T_{ab} given in (9) and evaluated on ϕ is *conserved* ($\nabla_a T^{ab}\equiv 0$) provided ϕ is a sufficiently smooth (customary C^∞) solution of the Euclidean motion, namely, $A'\phi\equiv 0$, and $V'\equiv 0$. More generally for the solution of Euclidean motion, in local coordinates and for any point $x\in M$ one finds

$$\nabla_a T^{ab}[\phi,\mathbf{g}](x)=-\tfrac{1}{2}\phi^2(x)\nabla^b V'(x). \tag{10}$$

Another important classical property is the following one. Whenever the field ϕ is massless and conformally coupled [i.e., $V'(x)\equiv m^2=0$ and $\xi=\xi_D$], the Euclidean action $S_{\mathcal{M}}$ is invariant under local conformal transformations and it holds also

$$g_{ab}T^{ab}[\phi,\mathbf{g}](x)=0 \tag{11}$$

everywhere, for smooth fields ϕ which are solutions of the (Euclidean) motion equations. Equations (10) and (11) can be checked for the tensor in (9) directly, holding our general hypotheses.

Actually, the requirement of A' positive is completely unnecessary for all the definitions and results given above which hold true in any C^∞ Riemannian as well as *Lorentzian* manifold. In our approach, the Lorentzian stress tensor is obtained by analytic continuation of the Euclidean time as pointed out above.

Passing to the quantum-averaged quantities, following Schwinger,[12] the *one-loop stress tensor* averaged on the quantum state determined by the Feynman propagator obtained from the Green's function of A' can be *formally* defined by[4,5,8,9]

$$\langle T_{ab}(x|A')\rangle := \frac{2}{\sqrt{g}}\frac{\delta}{\delta g_{ab}(x)}S_{\rm eff} = Z[A',\mathbf{g}]^{-1}\int \mathcal{D}\phi e^{-(1/2)\int_{\mathcal{M}}\phi A'[\mathbf{g}]\phi d\mu_g}T_{ab}[\phi,\mathbf{g}](x). \tag{12}$$

It is well known that the right-hand sides of (12) and the corresponding quantity in the Lorentzian section are affected by divergences whenever one tries to compute them by trivial procedures.[4,6,8] For instance, proceeding as usual (e.g., see Ref. 5), interpreting the functional integral of $\phi(x)\phi(y)$ as a Green's function of A' (the analytic continuation of the Feynman propagator), $G(x,y)$, and then defining an off-diagonal quantum averaged stress tensor,

$$\langle T_{ab}(x,y)\rangle = Z[A',\mathbf{g}]^{-1}\int \mathcal{D}\phi e^{-(1/2)\int_{\mathcal{M}}\phi A'[\mathbf{g}]\phi d\mu_g}O_{ab}(x,y)\phi(x)\phi(y) = O_{ab}(x,y)G(x,y), \tag{13}$$

where $O_{ab}(x,y)$ is an opportune bivectorial differential operator (see Ref. 5), the limit of coincidence of arguments *x* and *y*, necessary to get $\langle T_{ab}(x)\rangle$, trivially diverges. One is therefore forced to remove these divergences *by hand* and this is nothing but the main idea of the *point-splitting procedure*. Within the point-splitting procedure (10) is requested also for the quantum-averaged stress tensor at least in the case $V'=0$. Conversely, the property (11) generally does not hold in the case of a conformally coupled massless field: a *conformal anomaly* appears.[3–6]

Another approach to interpret the left-hand side of (12) in terms of local ζ-function was introduced in Ref. 2 without rigorous mathematical discussion. Anyhow, this approach has produced correct results and agreement with point-splitting procedures in several concrete cases[2,13] and it has pointed out a strong self-consistence and a general agreement[2] with the general axiomatic theory of the stress tensor renormalization built up by Wald.[5] (It is anyway worth stressing that Wald's axiomatic approach concerns the Lorentzian theory and thus any comparison involves an analytical continuation of the Euclidean theory. In such a way all the issues related to the locality of the theory cannot be compared directly with the general ζ-function approach.) Moreover, differently from the known point-splitting techniques, no difficulty arises dealing with the case of a massless conformally coupled field.

Similarly to the cases treated in Ref. 1, the *definition* of the formal quantity on the left-hand side of (12)[2] given in terms of the ζ-function and heat kernel contains an implicit *infinite renormalization* procedure in the sense that the result is finally free from divergences.

C. The key idea of the ζ-function regularization of the stress tensor

The key idea of ζ-function regularization of the stress tensor concerns the extension of the use of the ζ-function from the effective action to the stress tensor employing some manipulations of the series involved in the ζ-function technique. We remind the reader that formally one has[9,8,1]

$$S_{\rm eff}[A]_{\mu^2} = \frac{1}{2}\frac{d}{ds}\bigg|_{s=0}\left\{-\sum_{j\in\mathbb{N}}{}'\left(\frac{\lambda_j}{\mu^2}\right)^{-s}\right\}. \tag{14}$$

Here μ is the usual arbitrary mass scale necessary for dimensional reasons. Actually, the identity above holds true in the sense of the analytic continuation. Then, one can try to give some meaning to the following formal passages:

$$\langle T_{ab}(x)\rangle_{\mu^2}=\frac{2}{\sqrt{g(x)}}\frac{\delta}{\delta g_{ab}(x)}\frac{1}{2}\frac{d}{ds}\bigg|_{s=0}\left\{-\sum_{j\in\mathbb{N}}{}'\left(\frac{\lambda_j}{\mu^2}\right)^{-s}\right\}$$
$$=\frac{1}{2}\frac{d}{ds}\bigg|_{s=0}\left\{-\sum_{j\in\mathbb{N}}{}'\frac{2}{\sqrt{g(x)}}\frac{\delta}{\delta g_{ab}(x)}\left(\frac{\lambda_j}{\mu^2}\right)^{-s}\right\}$$
$$=\frac{1}{2}\frac{d}{ds}\bigg|_{s=0}\left\{\frac{s}{\mu^2}\sum_{j\in\mathbb{N}}{}'\left(\frac{\lambda_j}{\mu^2}\right)^{-(s+1)}\frac{2}{\sqrt{g(x)}}\frac{\delta\lambda_j}{\delta g_{ab}(x)}\right\}. \tag{15}$$

The functional derivative of λ_j has been computed in Ref. 2, at least formally. The passages above are mathematically incorrect most likely; anyhow, in Ref. 2 it was conjectured that the series in the last line of (15) converges and it can be analytically continued into a regular function $Z_{ab}(s,x|A/\mu^2)$ in a neighborhood of $s=0$. Then, one can *define* the renormalized averaged one-loop stress tensor as

$$\langle T_{ab}(x)\rangle_{\mu^2}:=\frac{1}{2}\frac{d}{ds}\bigg|_{s=0}Z_{ab}(s,x|A/\mu^2). \tag{16}$$

The explicit form of Z_{ab} found in Ref. 2 following the route above was

$$Z_{ab}(s,x|A/\mu^2)=2\frac{s}{\mu^2}\zeta_{ab}(s+1,x|A/\mu^2)+sg_{ab}(x)\zeta(s,x|A/\mu^2), \tag{17}$$

where $\zeta_{ab}(s,x|A/\mu^2)$ is the analytic continuation of the series

$$\sum_{j\in\mathbb{N}}{}'\left(\frac{\lambda_j}{\mu^2}\right)^{-s}T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x), \tag{18}$$

and

$$T_{ab}[\phi,\phi^*,\mathbf{g}](x):=-\frac{2}{\sqrt{g(x)}}\frac{\delta}{\delta g_{ab}(x)}\frac{1}{2}\int_{\mathcal{I}}\phi A'[\mathbf{g}]\phi^*\,d\mu_g. \tag{19}$$

However, no proof of the convergence of the series above was given in Ref. 2 for the general case, but the method was checked in concrete cases, where it was found that the series above converges really as supposed. In Ref. 2, it was shown also that, assuming reasonable mathematical properties of the involved functions, this approach in four-dimensional operators $A'=-\Delta+\xi R(x)+m^2$ should produce a stress tensor which is conserved and gives rise to the conformal anomaly. In Ref. 2, it was also (not rigorously) proven that the ambiguity arising from the presence of the arbitrary scale μ^2 gives rise to conserved geometric terms added to the stress tensor, in agreement with Wald's axioms.

We expect that the not completely rigorous procedures employed in Ref. 2 make sense provided the usual heat-kernel ''asymptotic'' expansion at $t\to 0$ can be derived in the variables which range in the manifold producing a similar expansion (this result is not trivial at all) and provided the series (18) can be derived under the symbol of summation (also this fact is not so obvious). Therefore, in the next parts of this work, we shall investigate also similar issues before we prove and generalize results found in Ref. 2.

III. THE LOCAL ζ-FUNCTION AND THE ONE-LOOP STRESS TENSOR

In this part and within our general hypotheses, we develop a rigorous theory of the ζ-function of the stress tensor and give a rigorous proof of some properties of particular tensorial ζ-functions introduced in Ref. 2.

The first subsection is devoted to generalizing some properties of the heat-kernel concerning the smoothness of several heat-kernel expansions necessary in the second subsection.

A. The smoothness of the heat-kernel expansion and the ζ-function

A first very useful result, which we state in the form of a lemma, concerns the smoothness of the heat-kernel expansion for $t\to 0$ (Theorem 1.3 of Ref. 1) and the possibility of deriving term by term such an "asymptotic expansion."

Before stating the result it is worth stressing that, in the trivial case $|\alpha|=|\beta|=0$, the statement of the lemma below and the corresponding proof include the point $(a2)$ in Theorem 1.3 of Ref. 1 given without proof there.

Lemma 3.1: Let us assume our general hypotheses on $\mathcal{M}$ *and* A'. *For any* $u\in\mathcal{M}$ *there is an open neighborhood* I_u *centered on u such that, for any local coordinate system defined therein, for any couple of points* x, $y\in I_u$, *for any couple of multindices* α, β, *and for any integer* $N>D/2+2|\alpha|+2|\beta|$ $(D/2+2$ *if* $|\alpha|=|\beta|=0)$ *the heat-kernel expansion (a) of Theorem 1.3 in Ref. 1 can be derived term by term obtaining* $[\eta\in(0,1)$ *is fixed arbitrarily as usual*$]$

$$D_x^\alpha D_y^\beta K(t,x,y)=D_x^\alpha D_y^\beta\left\{\frac{e^{-\sigma(x,y)/2t}}{(4\pi t)^{D/2}}\sum_{j=0}^{N} a_j(x,y|A)t^j\right\}+\frac{e^{-\eta\sigma(x,y)/2t}}{(4\pi t)^{D/2}}t^{N-|\alpha|-|\beta|}O_{\eta,N}^{(\alpha,\beta)}(t;x,y), \tag{20}$$

where the derivatives are computed in the common coordinate system given above and the function $(t,x,y)\mapsto O_{\eta,N}^{(\alpha\beta)}(t;x,y)$ *belongs to* $C^0([0,+\infty)\times I_u\times I_u)$ *at least, and for any positive constant* $K_{\eta,N}^{(\alpha,\beta)}$ *and* $0\leq t<K_{\eta,N}^{(\alpha,\beta)}$, *one gets*

$$|O_{\eta,N}^{(\alpha,\beta)}(t;x,y)|<M_{K_{\eta,N}^{(\alpha,\beta)}}|t|, \tag{21}$$

$M_{K_{\eta,N}^{(\alpha,\beta)}}$ *being a corresponding positive constant not dependent on* $x,y\in I_u$ *and* t.

Proof: See the Appendix. □

The next lemma concerns the possibility of interchanging the operators D_x^α, D_y^β with the symbol of series in the eigenvector expansion for the heat-kernel given in (b) of Theorem 1.1 in Ref. 1.

Lemma 3.2: Within our hypotheses on M and A', *the eigenvector expansion of the heat kernel given in (b) of Theorem 1.1 in* Ref. 1,

$$K(t,x,y|A)=\sum_{j=0}^{\infty} e^{-\lambda_j t}\phi_j(x)\phi_j^*(y), \tag{22}$$

where $t\in(0,+\infty)$, $x,y\in\mathcal{M}$, *and the real numbers* $\lambda_j(0\leq\lambda_0\leq\lambda_1,\leq\lambda_2,\leq\cdots)$ *are the eigenvalues of A with corresponding orthogonal normalized eigenvector* ϕ_j, *can be derived in x and y passing the derivative operators under the symbol of series. Indeed, in a coordinate system defined in a sufficiently small neighborhood* I_u *of any point* $u\in\mathcal{M}$, *for* $x,y\in I_u$, *for* $t\in(0,+\infty)$ *and for any couple of multindices* α,β,

$$D_x^\alpha D_y^\beta K(t,x,y|A)=\sum_{j=0}^{\infty} e^{-\lambda_j t}D_x^\alpha\phi_j(x)D_y^\beta\phi_j^*(y). \tag{23}$$

Moreover, for any $T>0$ *the following upper bounds hold:*

$$|e^{-\lambda_j t} D_x^\alpha \phi_j(x) D_y^\beta \phi_j^*(y)| \leq P_T^{(\alpha,\beta)} e^{-\lambda_j(t-2T)}, \quad (24)$$

$$|D_x^\alpha D_y^\beta K(t,x,y|A) - D_x^\alpha D_y^\beta P_0(x,y|A)| \leq P_T^{(\alpha,\beta)} \sum_{j\in\mathbb{N}}{}' e^{-\lambda_j(t-2T)} \quad (25)$$

$$\leq Q_T^{(\alpha,\beta)} e^{-\lambda(t-2T)}, \quad (26)$$

where $x,y \in I_u$ *and* $t \in (2T,+\infty)$, $P_T^{(\alpha,\beta)}$ *and* $Q_T^{(\alpha,\beta)}$ *are positive constants which do not depend on t,x,y,* $P_0(x,y|A)$ *is the integral kernel of the projector onto the kernel of A,* λ *is the value of the first strictly positive eigenvalue, the prime on the summation symbol indicates that the summation on the vanishing eigenvalues is not considered, and, finally,*

$$P_T^{(\alpha,\beta)} = [\sup_{x\in \overline{I_u}} \|D_x^\alpha K(T,x,.|A)\|_{L^2(\mathcal{M},d\mu_g)}][\sup_{y\in \overline{I_u}} \|D_y^\beta K(T,.,y|A)\|_{L^2(\mathcal{M},d\mu_g)}]. \quad (27)$$

Therefore, the convergence of the series in (23) is absolute in a uniform sense for (t,x,y) belonging in any set $[\gamma,+\infty)\times I_u \times I_u$, $\gamma > 0$.

Proof: See the Appendix. □

Remark: The right-hand side of (25) can be also written down as

$$P_T^{(\alpha,\beta)} \int_{\mathcal{M}} d\mu_g(z) \{K(t-2T,z,z|A) - P_0(z,z|A)\} = P_T^{(\alpha,\beta)} \,\mathrm{Tr}\{K_{(t-2T)} - P_0\}. \quad (28)$$

The two lemmata above enable us to state and prove a theorem concerning the derivability of the ζ-function. First of all, let us give some definitions (in the following we shall refer to Definitions 2.1 and 2.2 in Ref. 1).

Definition 3.1: Let us assume our general hypotheses on $\mathcal{M}$ *and* A'. *Fixing a sufficiently small neighborhood* I_u *of any point* $u \in \mathcal{M}$, *considering a coordinate system defined in* I_u *and choosing a couple of multindices* α,β, *the* **off-diagonal derived local ζ-function** *of the operator A is defined for* $x,y \in I_u$, $\mathrm{Re}\, s > D/2 + |\alpha| + |\beta|$, *as*

$$\zeta^{(\alpha,\beta)}(s,x,y|A/\mu^2) := D_x^\alpha D_y^\beta \zeta(s,x,y|A/\mu^2), \quad (29)$$

provided the right-hand side exists, where both derivatives are computed in the coordinate system defined above.

Definition 3.2: Let us assume our general hypotheses on $\mathcal{M}$ *and* A'. *Fixing a sufficiently small neighborhood* I_u *of any point* $u \in \mathcal{M}$, *considering a coordinate system defined in* I_u *and choosing a couple of multindices* α,β, *the* **derived local ζ-function** *of the operator A is defined for* $x \in I_u$, $\mathrm{Re}\, s > D/2 + |\alpha| + |\beta|$, *as*

$$\zeta^{(\alpha,\beta)}(s,x|A/\mu^2) := \{D_x^\alpha D_y^\beta \zeta(s,x,y|A/\mu^2)\}_{x=y}, \quad (30)$$

provided the right-hand side exists, where both derivatives are computed in the coordinate system defined above.

Remark: The use of a common coordinate system either for x and y is essential in these definitions.

The following theorem proves that the given definitions make sense.

Theorem 3.1: *Let us assume our general hypotheses on* $\mathcal{M}$ *and* A'. *The local off-diagonal* ζ*-function of the operator A defined for* $x,y \in \mathcal{M}$, $\mathrm{Re}\, s > D/2$, $\mu > 0$ (μ *being a constant with the dimension of a mass*),

$$\zeta(s,x,y|A/\mu^2) = \frac{1}{\Gamma(s)} \int_0^{+\infty} d(\mu^2 t)(\mu^2 t)^{s-1} \{K(t,x,y|A) - P_0(x,y|A)\}, \quad (31)$$

can be derived in x and y under the symbol of integration in a common coordinate system defined in a sufficiently small neighborhood I_u of any point $u \in \mathcal{M}$, provided $\mathrm{Re}\, s$ is sufficiently large. In particular, for any choice of multindices α,β and $x,y \in I_u$, it holds
(a) *for $\mathrm{Re}\, s > D/2 + |\alpha| + |\beta|$ the derived local ζ-functions are well defined holding*

$$D_x^\alpha D_y^\beta \zeta(s,x,y|A/\mu^2) = \frac{1}{\Gamma(s)} \int_0^{+\infty} d(\mu^2 t)(\mu^2 t)^{s-1} D_x^\alpha D_y^\beta \{K(t,x,y|A) - P_0(x,y|A)\}. \quad (32)$$

Moreover, the right-hand side of (32) defines an s-analytic function which belongs to $C^0(\{s \in \mathbb{C} | \mathrm{Re}\, s > D/2 + |\alpha| + |\beta|\} \times I_u \times I_u)$ together with all its s-derivatives.
(b) *Whenever $x \neq y$ are fixed in I_u,*

(1) *the right-hand side of (32) can be analytically continued in the variable s in the whole complex plane.*

(2) *Varying $s \in \mathbb{C}$ and $(x,y) \in (I_u \times I_u) - \mathcal{D}_{I_u}$, the s-continued function in (31) defines an everywhere s-analytic function which belongs to $C^\infty(\mathbb{C} \times \{(I_u \times I_u) - \mathcal{D}_{Iu}\})$ [where $\mathcal{D}_{I_u} := \{(x,y) \in I_u \times I_u | x = y\}$] and it holds in $\mathbb{C} \times \{(I_u \times I_u) - \mathcal{D}_{Iu}\}$*

$$D_x^\alpha D_y^\beta \zeta(s,x,y|A/\mu^2) = \zeta^{(\alpha,\beta)}(s,x,y|A/\mu^2), \quad (33)$$

where the function ζ on the left-hand side and the function $\zeta^{(\alpha,\beta)}$ on the right-hand side are the respective s-analytic continuations of the initially defined ζ-function (31) and the right-hand side of (32).

(3) *Equation (32) holds also when the left-hand side is replaced by the s-continued function $\zeta^{(\alpha,\beta)}$ for $\mathrm{Re}\, s > 0$, or everywhere provided $D_x^\alpha D_y^\beta P_0(x,y|A) = 0$ in the considered point (x,y).*
(c) *Whenever $x = y$ is fixed in I_u,*

(1) *the right-hand side of (30) can be analytically continued in the variable s in the complex plane obtaining a meromorphic function with possible poles, which are simple poles only, situated in the points*

$$s_j^{(\alpha,\beta)} = D/2 + |\alpha| + |\beta| - j, \quad j = 0,1,2,..., \quad \text{if } D \text{ is odd};$$

$$s_j^{(\alpha,\beta)} = D/2 + |\alpha| + |\beta| - j, \quad j = 0,1,2,...,D/2 - 1 + |\alpha| + |\beta|, \quad \text{if } D \text{ even}.$$

These poles and the corresponding residues are the same as for the set of analytic functions, labeled by the integer $N > D/2 + 2|\alpha| + 2|\beta|$ ($N > D/2 + 2$ if $|\alpha| = |\beta| = 0$),

$$R_N(s,x)_{\mu_0^{-2}} := \frac{\mu^{2s}}{(4\pi)^{D/2}\Gamma(s)} \sum_{j=0}^{N} \int_0^{\mu_0^{-2}} dt \times \{D_x^\alpha D_y^\beta e^{-\sigma(x,y)/2t} a_j(x,y|A)\}_{x=y} t^{s-1-D/2+j}, \quad (34)$$

defined for $x \in I_u$ and $\mathrm{Re}\, s > D/2 + |\alpha| + |\beta|$ and then continued in the s-complex plane. μ_0 is an arbitrary strictly positive mass scale which does not appear in the residues.

(2) *Varying $x \in I_u$, the s-continued function belongs to $C^0((\mathbb{C} - \mathcal{P}^{(\alpha,\beta)}) \times \mathcal{M})$ together with all its s derivatives, $\mathcal{P}^{(\alpha,\beta)}$ being the set of the actual poles (each for some x) among the points listed above. Moreover, for any coordinate x^k and $(s,x) \in \{\mathbb{C} - (\mathcal{P}^{(\alpha,\beta)} \cup \mathcal{P}^{(\alpha+1_k,\beta)} \cup \mathcal{P}^{(\alpha,\beta+1_k)})\} \times I_u$, $(\partial/\partial x^k)\zeta^{(\alpha,\beta)}(s,x|A/\mu^2)$ exists, is continuous in (s,x) with all of its s derivatives, analytic in the variable s and*

$$\frac{\partial}{\partial x^k} \zeta^{(\alpha,\beta)}(s,x|A/\mu^2) = \zeta^{(\alpha+1_k,\beta)}(s,x|A/\mu^2) + \zeta^{(\alpha,\beta+1_k)}(s,x|A/\mu^2), \quad (35)$$

where $\zeta^{(\alpha,\beta)}$ is the analytic continuation of the initially defined function (30).

(d) *For* $x,y \in I_u$, *the analytic continuations of the right-hand sides of (29) and (30) are well defined in a neighborhood of* $s=0$ *and it holds, and the result does not depend on the values of* $\mu_0>0$ *and* $\mu>0$,

$$[D_x^\alpha D_y^\beta \zeta(s,x,y|A/\mu^2)]_{s=0} + D_x^\alpha D_y^\beta P(x,y) = \delta_D \delta_{x,y} \lim_{s\to 0} R_N(s,x)_{\mu_0^{-2}}, \tag{36}$$

where $|_{s=0}$ *means the analytic continuation from* $\mathrm{Re}\, s>D/2+|\alpha|+|\beta|$ *to* $s=0$ *of the considered function and N is any integer* $>D/2+2|\alpha|+2|\beta|$ ($D/2$ *whenever* $|\alpha|=|\beta|=0$). *Finally, we have defined* $\delta_D=0$ *if D is odd and 1 otherwise,* $\delta_{x,y}=0$ *if* $x\neq y$ *or 1 otherwise.*

Sketch of Proof: The proof of this theorem is a straightforward generalization of the proof of Theorem 2.2 in Ref. 1, so we just sketch this proof. As in the proof of Theorem 2.2 in Ref. 1, the main idea is to break off the integration in (31) for $\mathrm{Re}\, s>D/2$ as

$$\zeta(s,x,y|A/\mu^2) = \frac{\mu^{2s}}{\Gamma(s)} \int_0^{+\infty} dt\, t^{s-1}[K(t,x,y|A) - P_0(x,y|A)] \tag{37}$$

$$= \frac{\mu^{2s}}{\Gamma(s)} \int_0^{\mu_0^{-2}} \{\cdots\} + \frac{\mu^{2s}}{\Gamma(s)} \int_{\mu_0^{-2}}^{+\infty} \{\cdots\}, \tag{38}$$

where $\mu_0>0$ is an arbitrary mass cutoff. Then one studies the possibility of computing the derivative passing D_x^α and D_y^β under the symbol of integration in both integrals on the right-hand-side above. This is possible provided the absolute values of the derived integrand are *x*,*y* uniformly bounded by integrable functions dependent on α,β in general, for any choice of α and β. This assures also the continuity of the derivatives because the derivatives of the integrands are continuous functions. The analyticity in *s* can be proved by checking the Cauchy–Riemann conditions passing the derivative under the symbol of integration once again. The *s*-derivatives of the integrand at any order can still be proven to be bounded with the same procedure. Then the proof is similar to the proof of Theorem 2.2 of Ref. 1. One uses Lemma 3.2 and (26) (choosing $2T<\mu_0^{-2}$) in place of the corresponding formula (99) of Ref. 1, to prove that the latter integral on the right-hand side of (38) can be derived under the symbol of integration obtaining an *s*-analytic function continuous with all of its *s*-derivatives, for $s\in\mathbb{C}$ and $x,y\in I_u$. The former integral can be studied employing Lemma 3.1 and, in particular, (20). The requirement $N>D/2+2$ in the expansion in Ref. 1 has to be changed $N>D/2+2|\alpha|+2|\beta|$ in the present case. The requirement in the point (a) $\mathrm{Re}\, s>D/2+|\alpha|+|\beta|$ arises by the term with $j=0$ in the heat-kernel expansion when all the derivatives either in *x* and in *y* act on the exponential producing a factor $t^{-|\alpha|-|\beta|}$ and posing $x=y$ in the end. Equation (101) and the successive ones of Ref. 1 have to be changed employing $O_\eta^{(\alpha,\beta)}$ in place of O_η and $t^{s-1+N-D/2-|\alpha|-|\beta|}$ in place of $t^{s-1+N-D/2}$.

The requirement $D_x^\alpha D_y^\beta P_0(x,y|A)=0$ in (b3) is simply due to the divergence of the integral $\int_0^{\mu_0^{-2}} dt t^{s-1}$ for $s\leqslant 0$.

Equation (36) is essentially due to the presence of the factor $1/\Gamma(s)$ in all considered integrals, which vanishes with a simple zero as $s\to 0$. □

Comments:

(1) The right-hand side of (36), for $x=y$ and when *D* is even, has the form

$$\frac{D_x^\alpha D_y^\beta a_{D/2}(x,y|A)+\cdots}{(4\pi)^{D/2}}, \tag{39}$$

where the dots indicate a finite number of further terms consisting of derivatives of product of heat-kernel coefficients and powers of $\sigma(x,y)$, computed in the coincidence limit of the arguments. In the case $|\alpha|=|\beta|=0$, this agrees with the found result for the simple local ζ-function given in Ref. 1.

(2) It is worth noticing that the right-hand side of (36) proves that the procedures of s-continuing $D_x^\alpha D_y^\beta \zeta(s,x,y|A/\mu^2)$ and that of taking the coincidence limit of arguments x,y *generally do not commute*. This means that, understanding both sides in the sense of the analytic continuation, in general

$$\zeta^{(\alpha,\beta)}(s,x,y|A/\mu^2)|_{x=y} \neq \zeta^{(\alpha,\beta)}(s,x|A/\mu^2). \tag{40}$$

Above the coincidence limit is taken *after* the analytic continuation. Obviously, whenever $\mathrm{Re}\, s > D/2+|\alpha|+|\beta|$,

$$\zeta^{(\alpha,\beta)}(s,x,y|A/\mu^2)|_{x=y} = \zeta^{(\alpha,\beta)}(s,x|A/\mu^2). \tag{41}$$

(3) The point (b2) proves that, for $x\neq y$, the Green's function of any operator A^n, $n=0,1,2,...$, defined in Ref. 1 via local ζ-function, is C^∞ as one could have to expect.

A second and last theorem concerns the possibility of computing the derived local ζ-functions through a series instead of an integral.

Theorem 3.2: *Within our hypotheses on $\mathcal{M}$ and A' and $\mu>0$, the (off-diagonal and not) derived local ζ-function can be computed as the sum of a series. Indeed, choosing a couple of multindices α,β, in a common coordinate system defined in a sufficiently small neighborhood of any point $u\in\mathcal{M}$, one has, in the sense of the punctual convergence,*

$$\zeta^{(\alpha,\beta)}(s,x,y|A/\mu^2) = \sum_{j\in\mathbb{N}}{}' \left(\frac{\lambda}{\mu^2}\right)^{-s} D_x^\alpha \phi_j(x) D_y^\beta \phi_j^*(y), \tag{42}$$

$$\zeta^{(\alpha,\beta)}(s,x|A/\mu^2) = \sum_{j\in\mathbb{N}}{}' \left(\frac{\lambda}{\mu^2}\right)^{-s} D_x^\alpha \phi_j(x) D_x^\beta \phi_j^*(x), \tag{43}$$

provided $\mathrm{Re}\, s>3D/2+|\alpha|+|\beta|$ *and* $(x,y)\in I_u\times I_u$.

Proof: First of all it is worth stressing that, in the considered domain for s, the functions are continuous in all variables and

$$\zeta^{(\alpha,\beta)}(s,x,y|A/\mu^2)|_{x=y} = \zeta^{(\alpha,\beta)}(s,x|A/\mu^2). \tag{44}$$

So, we perform our proof in the general case $x\neq y$, and then consider the coincidence limit of arguments. Therefore, from Theorem 3.1., for $\mathrm{Re}\, s>D/2+|\alpha|+|\beta|$, one has

$$\zeta^{(\alpha,\beta)}(s,x,y|A/\mu^2) = \frac{\mu^{2s}}{\Gamma(s)}\int_0^{\mu_0^{-2}} dt\, t^{s-1} D_x^\alpha D_y^\beta \{K(t,x,y|A) - P_0(x,y|A)\}$$
$$+ \frac{\mu^{2s}}{\Gamma(s)}\int_{\mu_0^{-2}}^{+\infty} dt\, t^{s-1} D_x^\alpha D_y^\beta \{K(t,x,y|A) - P_0(x,y|A)\}. \tag{45}$$

Here $\mu_0>0$ arbitrarily. Let us focus attention on the second integral. It can be written also

$$\frac{\mu^{2s}}{\Gamma(s)}\int_{\mu_0^{-2}}^{+\infty} dt \sum_{j\in\mathbb{N}}{}' t^{s-1} D_x^\alpha \phi_j(x) D_y^\beta \phi_j^*(y) e^{-\lambda_j t}, \tag{46}$$

where we have used Lemma 3.2. We want to show that it is possible to interchange the symbol of series with that of integration. We shall prove a similar fact for the other integral in (45), then the well-known formula $(a>0)$

$$a^{-s} = \frac{1}{\Gamma(s)}\int_0^{+\infty} dt\, t^{s-1} e^{-at} \tag{47}$$

will complete the proof of the theorem.

To prove the possibility of interchanging the integration with the summation in the integral (46), it is sufficient to show that the absolute value of the function after the summation symbol is integrable in the measure $\int dt \times \Sigma_j$. Then Fubini's theorem allows one to interchange the integrations. From Lemma 3.2, we know that, for $t > 2T > 0$,

$$\sum_{j\in\mathbb{N}}{}' |t^{s-1} D_x^{\alpha}\phi_j(x) D_y^{\beta}\phi_j^*(y) e^{-\lambda_j t}| \leqslant P_T^{(\alpha,\beta)} \sum_{j\in\mathbb{N}}{}' t^{\mathrm{Re}\, s-1} e^{-\lambda_j(2t-2T)}$$
$$\leqslant Q_T^{(\alpha,\beta)} t^{\mathrm{Re}\, s-1} e^{-\lambda(t-2T)}, \tag{48}$$

where λ is the first strictly positive eigenvalue of A. We choose the constant $T < \mu_0^{-2}/2$. The t-integration in $[\mu_0^{-2}, +\infty)$ of the last line above is finite for any $s \in \mathbb{C}$. Thus, a part of Fubini's theorem proves that the function after the summation symbol in (46) is integrable in the product measure.

Let us perform a similar proof for the first integral on the right-hand side of (45). It can be written down

$$\frac{\mu^{2s}}{\Gamma(s)} \int_0^{\mu_0^{-2}} dt \sum_{j\in\mathbb{N}}{}' t^{s-1} D_x^{\alpha}\phi_j(x) D_y^{\beta}\phi_j^*(y) e^{-\lambda_j t}. \tag{49}$$

We want to show that it is possible to interchange the symbol of series with that of integration. Posing $T = t/4$ we have, for $t \in (0, \mu_0^{-2}]$,

$$\sum_{j\in\mathbb{N}}{}' |t^{s-1} D_x^{\alpha}\phi_j(x) D_y^{\beta}\phi_j^*(y) e^{-\lambda_j t}| \leqslant P_{t/4}^{(\alpha,\beta)} t^{\mathrm{Re}\, s-1}\, \mathrm{Tr}\{K_{t/2} - P_0\}, \tag{50}$$

where for (27)

$$P_T^{(\alpha,\beta)} := [\sup_{x\in \bar{I}_u} \|D_x^{\alpha} K(T,x,.|A)\|][\sup_{y\in \bar{I}_u} \|D_y^{\beta} K(T,.,y|A)\|].$$

Employing (20) of Lemma 3.1 and taking account of the finite volume of the manifold one finds that there is a positive constant A such that, for $t \in (0, \mu_0^{-2}]$,

$$P_T^{(\alpha,\beta)} \leqslant A T^{-D-|\alpha|-|\beta|}. \tag{51}$$

This is due to the leading order for $t \to 0$ of the heat-kernel expansion (20). This upper bound, inserted in (50) with $T = t/4$, together with the x-integral of the heat-kernel expansion (19) in Theorem 1.3 of Ref. 1, entails

$$\sum_{j\in\mathbb{N}}{}' |t^{s-1} D_x^{\alpha}\phi_j(x) D_y^{\beta}\phi_j^*(y) e^{-\lambda_j t}| \leqslant B t^{\mathrm{Re}\, s-1-3D/2-|\alpha|-|\beta|}, \tag{52}$$

where B is a positive constant.

As in the previously considered case, for $\mathrm{Re}\, s > 3D/2 + |\alpha| + |\beta|$, we can interchange the symbol of integral with that of series also in the second integral of (45). Then (47) entails the thesis. □

Notice that, in the case $|\alpha| = |\beta| = 0$, the convergence of the series (43) arises for $\mathrm{Re}\, s > D/2$, and it is uniform as is well known.[1] Actually, our theorem uses a quite rough hypothesis. Nevertheless, this is enough for the use we shall make of the theorem above.

Following the way traced out in Theorem 1.3, we can give a precise definition concerning the ζ-function of the stress tensor. We shall assume, more generally than in Ref. 2, $A' := -\Delta + V$ where $V(x) := m^2 + \xi R + V'(x)$ and V' is real belongs to $C^{\infty}(\mathcal{M})$ and does not depend on the metric. Moreover, in this paper we consider a general D-dimensional manifold rather than the more physical case $D=4$ studied in Ref. 2. Also, as required by our general hypotheses, A' must be positive. It is worth stressing that this does not entail necessarily $m^2+\xi R(x)>0$ everywhere also when $V'\equiv 0$, and neither $V(x)>0$ everywhere in the general case (see Ref. 1).

B. The ζ-regularized stress tensor and its properties

For future convenience, let us define the symmetric tensorial field in a local coordinate system,

$$T_{ab}[\phi,\phi^*,\mathbf{g}](x) := \frac{2}{\sqrt{g(x)}}\frac{\delta}{\delta g_{ab}(x)}\frac{1}{2}\int_{\mathcal{I}}\phi A'[\mathbf{g}]\phi^*\, d\mu_g, \tag{53}$$

where $\phi\in C^\infty(\mathcal{M})$ and the functional derivative has been defined in Theorem 1.2. The precise form of $T_{ab}[\phi,\phi^*,\mathbf{g}](x)$ reads, in our case,

$$\begin{aligned}T_{ab}[\phi,\phi^*,\mathbf{g}](x) = {}& \tfrac{1}{2}(\nabla_a\phi(x)\nabla_b\phi^*(x)+\nabla_a\phi^*(x)\nabla_b\phi(x))\\ &-\tfrac{1}{2}g_{ab}(x)[\nabla_c\phi(x)\nabla^c\phi^*(x)+(m^2+V'(x))|\phi|^2(x)]+\xi[(R_{ab}(x)\\ &-\tfrac{1}{2}g_{ab}(x)R(x))|\phi|^2(x)+g_{ab}(x)\nabla_c\nabla^c|\phi|^2(x)-\nabla_a\nabla_b|\phi|^2(x)].\end{aligned} \tag{54}$$

A few trivial manipulations which make use of $A'\phi_j=\lambda_j\phi_j$ lead us to a simpler form for $T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x)$, namely

$$\begin{aligned}T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x) = {}& \tfrac{1}{2}(\nabla_a\phi_j(x)\nabla_b\phi_j^*(x)+\nabla_a\phi_j^*(x)\nabla_b\phi_j(x))-\xi\nabla_a\nabla_b|\phi_j|^2\\ &+\left(\xi-\frac{1}{4}\right)g_{ab}(x)\Delta|\phi_j|^2(x)+\xi R_{ab}(x)|\phi_j|^2-\frac{g_{ab}(x)}{2}\lambda_j|\phi_j|^2(x).\end{aligned} \tag{55}$$

Following the insights given in Sec. II C of this work as well as Ref. 2, we can give the following definition.

Definition 3.3: Within our hypotheses on $\mathcal{M}$ and $A' := -\Delta+m^2+\xi R+V'(x)$ defined above $(m,\xi\in\mathbb{R})$, the **local ζ-function of the stress tensor** *is the symmetric tensorial field defined in local coordinates as*

$$Z_{ab}(s,x|A/\mu^2) := 2\frac{s}{\mu^2}\zeta_{ab}(s+1,x|A/\mu^2)+s g_{ab}(x)\zeta(s,x|A/\mu^2), \tag{56}$$

where $\zeta_{ab}(s,x|A/\mu^2)$ is defined as the sum of the series below, in a sufficiently small neighborhood I_u of any point $u\in\mathcal{M}$ for $\mathrm{Re}\, s>3D/2+2$,

$$\sum_{j\in\mathbb{N}}{}'\left(\frac{\lambda_j}{\mu^2}\right)^{-s}T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x), \tag{57}$$

and $T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x)$ is defined in (53) and (55) with respect to a base of smooth orthogonal normalized eigenvector of A.

Comments:

(1) The definition given above makes sense since the relevant series converges for $\mathrm{Re}\, s>3D/2+2$ because of Theorem 3.2. Notice that the given definition does not depend on the base of smooth orthogonal normalized eigenvectors of A (take account that each eigenspace has finite dimension as follows from Theorem 1.1 in Ref. 1).

(2) The fact that the coefficients $Z_{ab}(s,x|A/\mu^2)$ do define a tensor is a direct consequence of (55) and (57). This can be trivially proven for $\mathrm{Re}\, s>3D/2+2$, where one can make use of the

series (57) which trivially defines a tensor since all terms of the sum are separately components of a tensor. Then, the proven property remains unchanged after the analytic continuation for values of s where the series does not converge.

(3) It is worthwhile noticing that the final expression of T_{ab}, and thus Z_{ab} self, does not contain either m^2 or V' explicitly.

(4) A definition trivially equivalent to (56) (up to analytic continuations in the variable s) is given by posing directly, for $\mathrm{Re}\, s > 3D/2+2$,

$$Z_{ab}(s,x|A/\mu^2) = s \sum_{j\in\mathbb{N}}{}' \left\{ \frac{2}{\mu^2}\left(\frac{\lambda_j}{\mu^2}\right)^{-(s+1)} T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x) + g_{ab}(x)\left(\frac{\lambda_j}{\mu^2}\right)^{-s}\phi_j(x)\phi_j^*(x)\right\}. \tag{58}$$

(5) Due to the uniqueness theorem for analytic functions and Theorem 3.2, each component of Z_{ab} can be built up, within opportune regions, employing the heat-kernel in the fashion of Theorem 3.1. Following this route, by Theorem 3.1, one proves trivially that the symmetric tensorial field $Z_{ab}(s,x|A/\mu^2)$ is continuous together for all of its s derivatives as a function of (s,x). In particular, each component defines a meromorphic function of the variable s whenever x is fixed. Therefore, let us consider the *simple* poles which may appear in the s-continued components of $Z_{ab}(s,x|A/\mu^2)$. One has to rewrite each component of this tensorial field in terms of simple ζ-functions and derived ζ-functions via Theorems 1.1 and 1.2. Once this has been done, one sees that the only functions which really appear in Z_{ab} are $\zeta^{(1_a,1_b)}$, $\zeta^{(1_b,1_a)}$, and ζ, each function evaluated at $s+1$ [see (59)]. The simple local ζ-function evaluated in $s+1$ admits possible simple poles in the points s_j with $s_j = D/2-j-1$ where $j=0,1,\ldots$, whenever D is odd, otherwise $j=0,1,\ldots,D/2-1$ whenever D is even. Anyhow, the factor s in (56) cancels out the possible pole at $s=0$, which may appear in $\zeta(s+1)$ when D is even. The functions $\zeta^{(1_c,1_d)}(s+1,x|A/\mu^2)$ have been considered in Theorem 1.1 and their possible simple poles may arise in s_j with $s_j = D/2+|1_a|+|1_b|-1-j = D/2-j+1$ and $j=0,1,\ldots$ whenever D is odd, otherwise $j=0,1,\ldots,D/2+1$ whenever D is even. Once again, because of the factor s on the right-hand side of (56), any possible simple pole at $s=0$ is canceled out.

The last comment above can be stated as a theorem.

Theorem 3.3: *In our general hypotheses on* $\mathcal{M}$ *and* A', *(a) each component of* $Z_{ab}(s,x|A/\mu^2)$ *can be analytically continued into a mesomorphic function of s whenever x is fixed. In particular, in the sense of the analytic continuation, it holds, for x belonging to a sufficiently small neighborhood of any point* $u\in\mathcal{M}$,

$$\begin{aligned} Z_{ab}(s,x|A/\mu^2) &= \frac{s}{\mu^2}[\zeta^{(1_a,1_b)}(s+1,x|A/\mu^2) + \zeta^{(1_b,1_a)}(s+1,x|A/\mu^2)] \\ &\quad + \frac{2s}{\mu^2}\left[\left(\xi-\frac{1}{4}\right)g_{ab}(x)\Delta + \xi R_{ab}(x) - \xi\nabla_a\nabla_b\right]\zeta(s+1,x|A/\mu^2). \end{aligned} \tag{59}$$

(b) The possible poles of each component of $Z_{ab}(s,x|A/\mu^2)$, *which are simple poles only, are situated in the points*

$$s_j = D/2-j+1, \quad j=0,1,2,\ldots, \quad \text{if } D \text{ is odd};$$

$$s_j = D/2-j+1, \quad j=0,1,2,\ldots,D/2, \quad \text{if } D \text{ even}.$$

(c) Varying $x\in I_u$ *and* $s\in\mathbb{C}$ *the s-analytically continued symmetric tensorial field* $(s,x)\mapsto Z_{ab}(s,x|A/\mu^2)$ *defines an s-analytic tensorial field of* $C^0((\mathbb{C}-\mathcal{P})\times I_u)$ *together with all its s derivatives, where* $\mathcal{P}$ *is the set of the actual poles (each for some x) among the points listed above.*

Proof: Sketched above. □

Remark: Equation (59) could be used as an independent definition of the ζ-function of the stress tensor. The important point is that it does not refer to any series of eigenvectors. It could be considered as the starting point for the generalization of this theory in the case where the spectrum of the operator A is continuous provided the functions on the right-hand side of (59) are defined in terms of t integrations of derivatives of the heat kernel.

Definition 3.4: In our general hypotheses on $\mathcal{M}$ and A' and for $x \in I_u$ where I_u is a sufficiently small neighborhood of $u \in \mathcal{M}$, the **one-loop renormalized stress tensor** *is defined in a local coordinate system in I_u by the set of functions $(a,b=1,...,D)$,*

$$\langle T_{ab}(x|A)\rangle_{\mu^2} := \frac{1}{2}\frac{d}{ds}\bigg|_{s=0} Z_{ab}(s,x|A/\mu^2), \tag{60}$$

where the tensorial field Z_{ab} which appears on the right-hand side is the s-analytic continuation of that defined above and $\mu^2>0$ is any fixed constant with the dimensions of a squared mass.

We can state and prove the most important properties of $\langle T_{ab}(x|A)\rangle_{\mu^2}$ in the following theorem. These results generalize previously obtained results[2,13] for a more general operator A and for any dimension $D>0$.

Theorem 3.4: *In our general hypotheses on $\mathcal{M}$ and A', the functions $x \mapsto \langle T_{ab}(x|A)\rangle_{\mu^2}$ defined above satisfy the following properties.*

(a) The functions $x \mapsto \langle T_{ab}(x|A)\rangle_{\mu^2}$ $(a,b=1,2,...,D)$ define a C^∞ symmetric tensorial field on $\mathcal{M}$.

(b) This tensor is conserved *for $V' \equiv 0$, and, more generally,*

$$\nabla^a \langle T_{ab}(x|A)\rangle_{\mu^2} = -\tfrac{1}{2}\langle \phi^2(x|A)\rangle_{\mu^2}\nabla_b V'(x) \tag{61}$$

everywhere in $\mathcal{M}$.

(c) For any rescaling $\mu^2 \mapsto \alpha\mu^2$, where $\alpha>0$ is a pure number, one has

$$\langle T_{ab}(x|A)\rangle_{\mu^2} \to \langle T_{ab}(x|A)\rangle_{\alpha\mu^2} = \langle T_{ab}(x|A)\rangle_{\mu^2} + (\ln\alpha)t_{ab}(x|A), \tag{62}$$

where $t_{ab}(x|A)=Z_{ab}(0,x|A)/2$, which coincides also with the residue of the pole of $\zeta_{ab}(s+1,x|A)$ at $s=0$, is a, conserved for $V' \equiv 0$, symmetric tensor not dependent on μ built up by a linear combination of product of the metric, curvature tensors, $V'(x)$ and their covariant derivatives evaluated at the point x. In general, it satisfies

$$\nabla^a t_{ab}(x|A) = -\delta_D \frac{a_{D/2-1}(x,x|A)}{2(4\pi)^{D/2}}\nabla_b V'(x), \tag{63}$$

where $\delta_D=0$ when D is odd and $\delta_D=1$ otherwise. In terms of heat-kernel coefficients one has also

$$t_{ab}(x|A) = \frac{\delta_D}{(4\pi)^{D/2}}\left\{ a_{D/2-1,(ab)}(x,x|A) + \frac{g_{ab}(x)}{2}a_{D/2}(x,x|A) + \left[\left(\xi-\frac{1}{4}\right)g_{ab}(x)\Delta + \xi R_{ab}(x) - \xi\nabla_a\nabla_b\right]a_{D/2-1}(x,x|A)\right\}, \tag{64}$$

where we have employed the notations (using the same coordinate system both for x and y)

$$a_{j,(ab)}(x,x|A) := \tfrac{1}{2}[(\nabla_{(x)a}\nabla_{(y)b} + \nabla_{(y)a}\nabla_{(x)b})a_j(x,y|A)]|_{x=y}. \tag{65}$$

(d) *Concerning the trace of $\langle T_{ab}(x|A)\rangle_{\mu^2}$ one has*

$$g^{ab}(x)\langle T_{ab}(x|A)\rangle_{\mu^2}=\left(\frac{\xi_D-\xi}{4\xi_D-1}\Delta-m^2-V'(x)\right)\langle\phi^2(x|A)\rangle_{\mu^2}+\delta_D\frac{a_{D/2}(x,x|A)}{(4\pi)^{D/2}}-P_0(x,x|A). \tag{66}$$

Above, $\langle\phi^2(x|A)\rangle_{\mu^2}$ *is the value of the averaged quadratic fluctuations of the field computed by the* ζ*-function approach.*[1,13]

[The coefficient $(4\xi_D-1)^{-1}$ above is misprinted in Ref. 13 where $(2\xi_D)^{-1}$ appears in place of it.]

Sketch of Proof: Barring the issue concerning the smoothness, the property (a) is a trivial consequence of the corresponding fact for $Z_{ab}(s,x|A/\mu^2)$ discussed in Comment (2) after Definition 3.3. The tensorial field belongs to C^∞ because of the C^∞ smoothness of the functions $(s,x)\mapsto Z_{ab}(s,x|A/\mu^2)$ for $(s,x)\in J_0\times I_u$, where J_0 and I_u are respectively neighborhoods of $s=0$ in $\mathbb{C}$ and $u\in\mathcal{M}$. Indeed, first of all, no pole at $s=0$ arises in the functions $(s,x)\mapsto Z_{ab}(s,x|A/\mu^2)$ and in their x derivatives. This is because, considering (56) and (c2) of Theorem 3.1, one notices that if any pole appears in the various $\zeta^{(\alpha,\beta)}$ functions used building up Z_{ab}, it has to be a simple pole. Anyhow, the factor s makes the global functions Z_{ab} regular at $s=0$. Using recursively (c2) of Theorem 3.1 one has that each function $x\mapsto Z_{ab}(s,x|A/\mu^2)$ is C^∞ in a neighborhood of u for any fixed $u\in\mathcal{M}$ and $s=0$. More generally, this result holds for s fixed in a neighborhood of 0 because, by (c1) of Theorem 3.1, one has that no pole can arise in an open disk centered in $s=0$ with radius $\rho=\frac{1}{2}$. The functions Z_{ab} and all their x derivatives are also s-analytic for x fixed in a neighborhood of 0. Then, we can conclude that any function $(s,x)\mapsto Z_{ab}(s,x|A/\mu^2)$ is C^∞ in a neighborhood of $(0,u)$ for any fixed $u\in\mathcal{M}$. The C^∞ smoothness of the stress tensor then follows trivially from (60) directly.

The property (b) can be proved as follows. From the point (c2) of Theorem 3.1 and taking account of (a) of Theorem 3.3 and the definition (60), we have that (b) holds true if $\nabla^a Z_{ab}(s,x|A/\mu^2)=-Z(s,x|A/\mu^2)\nabla_b V'(x)$ for the considered point x and $s\in\mathbb{C}$ away from the poles, the function Z on the right-hand side of the field fluctuations (see Definition 2.7 in Ref. 1). By the theorem of the uniqueness of the analytic continuation, if one is able to prove such an identity for $\mathrm{Re}\, s$ sufficiently large, this assures also the validity of $\nabla^a Z_{ab}(s,x|A/\mu^2)=-Z(s,x|A/\mu^2)\nabla_b V'(x)$ everywhere in the variable s. Therefore, let us prove that there is an $M>0$ such that $\nabla^a Z_{ab}(s,x|A/\mu^2)=-Z(s,x|A/\mu^2)\nabla_b V'(x)$ for $\mathrm{Re}\, s>M$ and this will be enough to prove the point (b). To get this goal we represent $\nabla^a Z_{ab}(s,x|A/\mu^2)$, employing (59) for each function Z_{ab}. Then we make recursive use of (35) of Theorem 3.1 and obtain $\nabla^a Z_{ab}(s,x|A/\mu^2)$ written as a linear combination of functions $\zeta^{(\alpha,\beta)}(s+1,x|A/\mu^2)$. Finally we can expand all these functions in series of the form (43) of Theorem 3.2, provided $\mathrm{Re}\, s>M$ for an opportune $M>0$. Taking account of the comment (4) after Definition 3.3, the explicit expression of the final series of $\nabla^a Z_{ab}(s,x|A/\mu^2)$ reads, for $\mathrm{Re}\, s>M$,

$$\nabla^a Z_{ab}(s,x|A/\mu^2)=s\sum_{j\in\mathbb{N}}{}'\frac{2}{\lambda_j}\left(\frac{\lambda_j}{\mu^2}\right)^{-s}\nabla^a\left\{T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x)+\frac{\lambda_j g_{ab}(x)}{2}\phi_j(x)\phi_j^*(x)\right\}. \tag{67}$$

Finally, using the form of the ζ-function of the field fluctuations given in Ref. 1, one has to prove that for any $x\in\mathcal{M}$,

$$\nabla^a\left\{T_{ab}[\phi_j,\phi_j^*,\mathbf{g}](x)+\frac{\lambda_j g_{ab}(x)}{2}\phi_j(x)\phi_j^*(x)\right\}=-\frac{1}{2}\phi_j(x)\phi_j^*(x)\nabla_b V'(x). \tag{68}$$

This is nothing but the generalized ''conservation law'' of the stress tensor for the action

$$S_j[\phi,\phi^*]=\frac{1}{2}\int_{\mathcal{M}}[\phi A'[\mathbf{g}]\phi^*-\lambda_j\phi\phi^*]d\mu_g. \tag{69}$$

Indeed, (68) holds when the field ϕ satisfies the motion equations for the action above $A'\phi=\lambda_j\phi$. This is satisfied by the C^∞ eigenfunctions of $A\phi_j$ with eigenvalue λ_j. Therefore (68) holds true and (b) is proven.

Concerning the point (c), (62) with $t_{ab}(x|A)=Z_{ab}(0,x|A)/2$ arises as a direct consequence of the definition (60), noticing that, from Theorem 2.3, $Z(s,x|A/\mu^2)$ is analytic at $s=0$ and, from (56), (60) can be written down also:

$$\langle T_{ab}(x|A)\rangle_{\mu^2}=\frac{1}{2}\left.\frac{d}{ds}\right|_{s=0}Z_{ab}(s,x|A)+\frac{1}{2}Z_{ab}(0,x|A)\ln\mu^2, \tag{70}$$

Similarly, from Definition 3.3, one sees also that $Z(s,x|A/\mu^2)/2$ evaluated at $s=0$ takes contribution only from the possible pole at $s=1$ of the function $s\mapsto\zeta_{ab}(s,x|A/\mu^2)$ (the remaining simple ζ function is regular for $s=0$) and coincides with the value of residue of the pole of this function at $s=0$. When D is odd, no pole of $s\mapsto\zeta_{ab}(s,x|A/\mu^2)$ arises at $s=1$ because of (c) of Theorem 3.1. This is the reason for the δ_D on the right-hand side of (64). The form (64) of t_{ab} assures that it is built up as a linear combination of product of the metric, curvature tensors, $V'(x)$, and their covariant derivatives, everything evaluated at the same point x. The property (63) is consequence of $\nabla^a Z_{ab}(s,x|A/\mu^2)=-Z(s,x|A/\mu^2)\nabla_b V'(x)$ proven during the proof of (b), employing the pole structure of the function $Z(s,x)$ given in Ref. 1. Notice also that t_{ab} is symmetric by construction. Therefore, we have to prove (64) and this concludes the proof of (c). It is sufficient to show that

$$\lim_{s\to 0} s\zeta_{ab}(s+1,x|A)=\frac{\delta_D}{(4\pi)^{D/2}}\left\{a_{D/2-1,(ab)}(x,x|A)+\frac{g_{ab}(x)}{2}a_{D/2}(x,x|A)\right.$$
$$\left.+\left[\left(\xi-\frac{1}{4}\right)g_{ab}(x)\Delta+\xi R_{ab}(x)-\xi\nabla_a\nabla_b\right]a_{D/2-1}(x,x|A)\right\}.$$

From (a) of Theorem 3.3 this is equivalent to

$$\lim_{s\to 0} s\,\frac{1}{2}[\zeta^{(1_a,1_b)}(s+1,x|A)+\zeta^{(1_b,1_a)}(s+1,x|A)]$$
$$+\lim_{s\to 0} s\left[\left(\xi-\frac{1}{4}\right)g_{ab}(x)\Delta+\xi R_{ab}(x)-\xi\nabla_a\nabla_b\right]\zeta(s+1,x|A/\mu^2)$$
$$=\frac{\delta_D}{(4\pi)^{D/2}}\left\{a_{D/2-1,(ab)}(x,x|A)+\frac{g_{ab}(x)}{2}a_{D/2}(x,x|A)\right.$$
$$\left.+\left[\left(\xi-\frac{1}{4}\right)g_{ab}(x)\Delta+\xi R_{ab}(x)-\xi\nabla_a\nabla_b\right]a_{D/2-1}(x,x|A).\right. \tag{71}$$

The proof of this identity is very straightforward so we sketch its way only. Using Theorem 3.1, it is sufficient to consider the decomposition for large Re s (and a similar decomposition interchanging a with b):

$$s\zeta^{(1_a,1_b)}(s+1,x,y|A/\mu^2)|_{x=y}=\frac{s}{\Gamma(s+1)}\int_0^{+\infty}dt\,t^s D_x^{1_a}D_y^{1_b}[K(t,x,y|A)-P_0(x,y|A)]|_{x=y}$$
$$=\frac{s}{\Gamma(s+1)}\int_0^{\mu_0^{-2}}D_x^{1_a}D_y^{1_b}\{\cdots\}\Big|_{x=y}+\frac{s}{\Gamma(s+1)}\int_{\mu_0^{-2}}^{+\infty}D_x^{1_a}D_y^{1_b}\{\ldots\}\Big|_{x=y}. \tag{72}$$

Then, we can expand the integrand of the first integral in the second line of (72) using (20) of Lemma 3.1 and we can continue both integrals in the second line of (72) as far as $s=0$. A direct computation proves that, because of the presence of the factor s, only the first integral in (72), expanded as said above, gives a contribution. The contribution arises from the terms of the heat-kernel expansion which, once integrated in t (taking account of the factor t^s in the integrand), have a pole for $s=0$. This pole is canceled out by the factor s giving a finite result. The terms which have no pole at $s=0$ vanish due to the factor s, in the limit $s\to 0$. A similar procedure can be employed concerning the second limit in (71). In performing calculations, it is worth to remember that $\nabla_{(x)a}\sigma(x,y)$ and $\nabla_{(y)b}\sigma(x,y)$ *vanish* in the limit $x\to y$, and, furthermore, $\nabla_{(x)a}\nabla_{(y)b}\sigma(x,y)|_{x=y}=-g_{ab}(y)$. Summing all contributions, one obtains (64).

Concerning the point (d), the proof is dealt with as follows. Starting from (55) one finds

$$g^{ab}T_{ab}[\phi_j,\phi_j^*,\mathbf{g}]=\nabla_c\phi_j\nabla^c\phi_j^*+\left\{\xi R+\left[\xi(D-1)-\frac{D}{4}\right]\Delta\right\}|\phi_j|^2-\frac{D}{2}\lambda_j|\phi_j|.$$

Then, employing the identities $2\nabla_c\phi^*\nabla^c\phi=\Delta|\phi|^2-\phi\Delta\phi^*-\phi^*\Delta\phi$ and $(-\Delta+\xi R+m^2+V')\phi_j=\lambda_j\phi_j$, we have also

$$g^{ab}T_{ab}[\phi_j,\phi_j^*,\mathbf{g}]=\left[\xi(D-1)-\frac{D-2}{4}\right]\Delta|\phi_j|^2-(m^2+V')|\phi|^2+\frac{2-D}{2}\lambda_j|\phi_j|.$$

Since $\xi_D=(D-2)/[4(D-1)]$, we have finally

$$g^{ab}T_{ab}[\phi_j,\phi_j^*,\mathbf{g}]=\left[\frac{\xi_D-\xi}{4\xi_D-1}\Delta-m^2-V'\right]|\phi_j|^2+\frac{2-D}{2}\lambda_j|\phi_j|.$$

From Definition 3.3, this entails that, for $\mathrm{Re}\, s$ sufficiently large,

$$g^{ab}Z_{ab}(s,x|\mu^2)=\frac{2}{\mu^2}\left[\frac{\xi_D-\xi}{4\xi_D-1}\Delta-m^2-V'\right]s\zeta(s+1,x|A/\mu^2)+2s\zeta(s,x|A/\mu^2). \tag{73}$$

The function $(s,x)\mapsto s\zeta(s+1,x|A/\mu^2)$ is C^∞ in a neighborhood of $(0,u)$ for any $u\in\mathcal{M}$. The proof is similar to that given in (b) above for $(s,x)\mapsto Z_{ab}(s,x|\mu^2)$. Finally, employing Definition 3.4 also taking account of Definition 2.7 in Ref. 1 and (34) in Theorem 2.2 in Ref. 1 [i.e., (74) below], one finds (66). □

Comments:

(1) Concerning the point (b) which generalizes the classical law (10), we stress that this result is strongly untrivial. We have not put this result somewhere ''by hand'' in the definitions and hypotheses we have employed. Notice also that, in the case $V'\equiv 0$, the tensor $T_{ab}[\phi_j,\phi_j^*,\mathbf{g}]$ we have used in the definitions is not conserved. Nevertheless, the final stress tensor is conserved. This should mean that the local ζ-function approach is quite a deep approach.

(2) Concerning the point (c), we notice that this result is in agreement with Wald's axioms [5] and, on a purely mathematical ground, it reduces the ambiguity allowed by Wald's theorem. Indeed, Wald's theorem involves at least two arbitrary terms dependent on two free parameters. Recently it has been proven that in the case of massive fields which are not conformally coupled such an ambiguity should be much larger.[14] The point (d) proves that the corresponding ambiguity related to the field fluctuations is consistent with that which arises from the stress tensor. Assuming the ζ-function procedure, the only ambiguity remaining is just that related to the initial arbitrary mass scale μ. On the other hand, there is no physical evidence that the ζ-function procedure is the physically correct one and thus one cannot conclude that this method gets rid of the ambiguity pointed out by Wald *et al.*

(3) Concerning the point (d), we notice that, in the case $\xi=\xi_D$ and $V'\equiv m^2=0$, the usual conformal anomaly[8,9] arises provided D is even and the kernel of A is trivial. Anyhow, in the case

Ker A is untrivial, the trace anomaly takes a contribution from the null modes also when D is odd. In any cases, for the anomalous term, it holds [(34) in Theorem 2.2 in Ref. 1]

$$\delta_D \frac{a_{D/2}(x,x|A)}{(4\pi)^{D/2}} - P_0(x,x|A) = \zeta(0,x|A/\mu^2) \tag{74}$$

also for $\xi \neq \xi_D$.

Let us consider some issues related to the physical interpretations of the theory. Suppose S_1 acts as a globally one-parameter isometry group on the Riemannian manifold $\mathcal{M}$ giving rise to closed orbits with period $\beta > 0$. Suppose also that there exist a $D-1$ embedded submanifold Σ which, barring fixed points, intersects each orbit just once and is orthogonal to the Killing vector field of the isometry group K (notice that any submanifold Σ_τ, obtained by the action on Σ of the isometry group on the points of Σ, remains orthogonal to the Killing vector field). In this case the Riemannian metric is said to be *static*, the parameter of the group τ is said to be the *Euclidean time* of the manifold with period β, and the submanifold Σ is said to be the *Euclidean space* of the manifold.

As is well known, β is interpreted as the ''statistical mechanics'' inverse temperature of the quantum state; anyway, it has no direct physical meaning because it can be changed by rescaling the normalization of the Killing vector K everywhere by a constant factor. The physical temperature, which, in principle, may be measured by a thermometer, is the local rescaling-invariant Tolman temperature $T_{\mathrm{T}} := 1/[\sqrt{(K,K)}\beta]$.

Whenever $\mathcal{M}$ is static and Σ is endowed with a global coordinate system $(x^1,\ldots,x^{D-1}) \equiv \vec{x}$, $\mathcal{M}$ is endowed with a natural coordinate system $(\tau,\vec{x})$, $\tau \in (0,\beta)$, $\vec{x} \in \Sigma - \mathcal{F}$, where $\mathcal{F}$ is the set of the fix points of the group (which, anyhow, may be empty). This coordinate system is obtained by the evolution of the coordinates on Σ along the orbits of the isometry group and is almost global in the sense that is defined everywhere on $\mathcal{M}$ except for the set of the (coincident) endpoints of each orbit at $\vec{x}$ constant including the fix points of the group. This set has anyway negligible measure. Coordinates $(\tau,\vec{x})$ given above are said to be *static coordinates*. Notice that, in these coordinates, $\partial_\tau g_{ab} = 0$ and $g_{\tau\alpha} = 0$ for $\alpha = 1,...,D-1$ everywhere. Local static coordinates are defined similarly.

The important result is that, under our general hypotheses, supposing also that $\mathcal{M}$ is static and admits static coordinates $(\tau,\vec{x})$ and V' does not depend on τ, one has that the stress tensor *depends on* $\vec{x}$ *only* and satisfies everywhere

$$\langle T_{\tau\alpha}(\vec{x}|A)\rangle_{\mu^2} = \langle T_{\alpha\tau}(\vec{x}|A)\rangle_{\mu^2} = 0 \tag{75}$$

for $\alpha = 1,...,D-1$. The remarkable point as far as the physical ground is concerned is that this result allows one to look for analytic continuations towards Lorentzian metrics performing the analytical continuation $\tau \to it$ and without encountering imaginary components of the continued stress tensor. Notice that also $\langle \phi^2(x|A)\rangle_{\mu^2}$ and the effective Lagrangian $\mathcal{L}_{\mathrm{eff}}(x|A)_{\mu^2}$ (see Definition 2.5 in Ref. 1) do not depend on the temporal coordinate and, moreover, all results contained in Theorem 3.4 hold true in the Lorentzian section of the manifold, considering the trivial analitic continuations of all the terms which appear in the thesis. One has the following theorem.

Theorem 3.5: *Within our hypotheses on* $\mathcal{M}$ *and A, suppose* $\mathcal{M}$ *is static with Euclidean time* $\tau \in (0,\beta)$ $(\beta > 0)$ *and* V' *is invariant under Euclidean time displacements. In this case, for any* $\mu^2 > 0$ *and any point* $x \in \mathcal{M}$,

$$\langle T_{ab}(x|A)\rangle_{\mu^2} \mathrm{K}^a(x)\sigma^b(x) = 0, \tag{76}$$

where K *is the Killing vector field associated to the time* τ *and* $\sigma(x)$ *is any vector orthogonal to* K *at x. Furthermore, denoting the Lie derivative along* K *by* $\mathcal{L}(\mathrm{K})$, *it holds everywhere on* $\mathcal{M}$ *(away from fixed points concerning (77))*

$$\mathcal{L}(\mathrm{K})_c \langle T^{ab}(x|A)\rangle_{\mu^2} = 0, \tag{77}$$

$$\mathcal{L}(\mathrm{K}) \langle \phi^2(x|A)\rangle_{\mu^2} = 0, \tag{78}$$

$$\mathcal{L}(\mathrm{K})\mathcal{L}_{\mathrm{eff}}(x|A)_{\mu^2}=0. \tag{79}$$

Finally, all the results of Theorem 3.4 hold in the Lorentzian section provided one considers the Lorentzian-time-continued quantities in place of the corresponding Riemannian ones everywhere.

Sketch of Proof: In the given hypotheses and fixed $x\in\mathcal{M}$ [x different from any fixed point in such a case, the thesis being trivial since $\mathrm{K}(x)=0$], let us consider a generally local coordinate system $\vec{x}$ on the Euclidean space Σ around the intersection of the orbit passing from $x\in\mathcal{M}$. This induces a natural local coordinate system on $\mathcal{M}$, $(\tau,\vec{x})$ [where $\tau\in(0,\beta)$], which includes the same point x. In our hypotheses (76) is trivially equivalent to (75) in the considered coordinate system.

Concerning the form of the ζ-function of the stress tensor given in Definition 3.3, taking account of (54), since $g_{\tau\alpha}(x)=0$ and $\partial_\tau g_{ab}(x)=0$, only the first line of (54) and the last term in the last line may produce the considered components of the stress tensor. Actually, the dependence from τ of the eigenfunctions ϕ_j of the operator A can be taken of the form $e^{i\omega\tau}$ with $\omega\in\mathbb{R}$ just because $\partial_\tau=\mathrm{K}$ is a Killing field as we shall prove shortly. Then, the last term in the last line of (54) immediately vanishes concerning the considered components because the argument of the covariant derivatives (which commute on scalar fields) does not depend on τ; furthermore, taking account that A is real and thus ϕ_j and ϕ_j^* correspond to the same eigenvalue, one sees that the contribution coming from the first line of (54) computed for $b\neq a=\tau$ and $a\neq b=\tau$ vanishes when one sums over j to get the stress tensor ζ-function. The validity of (77)–(79) is also obvious working in local static coordinates where the Lie derivative reduces to the ordinary τ derivative and taking account of the imaginary exponential dependence form τ of the modes. In fact, this dependence is canceled out directly in the various ζ-functions due to the product of ϕ_j and ϕ_j^* (or corresponding derivatives) which appear in their definitions.

Let us finally prove that one can define the normalized orthogonal eigenvectors of A' (and thus A) in order to have the dependence from τ said above. Remembering that each eigenfunction of A is a $C^\infty(\mathcal{M})$ function, and working in the local coordinate system around the orbit of x considered above where g_{ab} does not depend on τ, one trivially has that

$$A'\partial_\tau\phi_{jk_j}=\partial_\tau A'\phi_{jk_j}=\lambda_j\partial_\tau\phi_{jk_j}, \tag{80}$$

where ϕ_{jk_j} is an eigenvector of A with eigenvalue λ_j. This holds in the considered coordinates and, therefore, choosing different local coordinate systems in Σ and reasoning similarly, the above identity can be proven to hold *almost* everywhere on $\mathcal{M}$ provided $\partial_\tau\phi_{jk_j}$ is interpreted as the $C^\infty(\mathcal{M})$ scalar field $(\mathrm{K},\nabla\phi_{jk_j})$. Remembering that the dimension of each eigenspace d_j is finite (Theorem 1.1. in Ref. 1), it must be

$$\partial_\tau\phi_{jk_j}(x)=\sum_{l_j=1}^{d_j}c_{k_jl_j}\phi_{jl_j}(x), \tag{81}$$

almost everywhere. Remembering that locally $\partial_\tau g_{ab}=0$ and, since $g_{\tau\alpha}=0$, $g=(\mathrm{K},\mathrm{K})h$ (where h is the determinant of the metric induced on Σ), one finds from (81)

$$\begin{aligned}c^*_{k_jh_j}+c_{h_jk_j}&=\int_{\mathcal{M}}\partial_\tau\{\phi^*_{jk_j}(x)\phi_{jh_j}(x)\}d\mu_g(x)\\&=\int_0^\beta d\tau\frac{d}{dt}\int_\Sigma\phi^*_{jk_j}(\tau,p)\phi_{jh_j}(\tau,p)(K(p),K(p))^{1/2}d\nu(p),\end{aligned}$$

where p is any point on the submanifold Σ and ν is its (finite) Riemannian measure induced there from the metric. We have passed the derivative through the symbol of integration employing Fubini's theorem and Lebesgue's dominate convergence theorem. The right-hand side of the identity above vanishes, taking account that, for any fixed p,

$$\lim_{\tau\to 0^+} \phi(\tau,p)_{jk_j} = \lim_{\tau\to\beta} \phi(\tau,p)_{jk_j} \tag{82}$$

because the orbits of the coordinate τ are closed and the functions are continuous in the whole manifold. Therefore, the matrix of the coefficients c_{pq} is anti-Hermitian. Finally, in the considered eigenspace, we can choose an orthogonal base of smooth normalized eigenfunctions where the matrix above is represented by a diagonal matrix, the eigenvalues being $i\omega_{jl_j}$, $\omega_{jl_j}\in\mathbb{R}$ and $l_j = 1,2,...,d_j$. In the new base one rewrites (81), in local coordinates,

$$\partial_\tau \phi_{jk_j}(x) = i\omega_{jk_j}\phi_{jk_j}(x), \tag{83}$$

and this entails trivially, with $\omega_{jk_j} = 2\pi n_{jk_j}/\beta$, $n_{jk_j}\in\mathbb{Z}$ by (82),

$$\phi_{jk_j}(x) = e^{i\omega_{jk_j}\tau}\varphi_{jk_j}(x^1,\ldots,x^{D-1}). \tag{84}$$

We leave to the reader the simple proof of the last statement of our theorem which can be carried out in local coordinates. □

As a final remark notice that changes in the period β of the manifold which correspond to actual increases of the proper length of the orbits (and not to a simple rescaling of the normalization of the Killing vector), in general produce *conical singularities* in the fixed points of the Lie group provided they exist. In such a case the manifold fails to be smooth and, in general, the theorems proven in this work and in Ref. 1 may not hold.

IV. THE RELATION BETWEEN THE ζ-FUNCTION AND THE POINT-SPLITTING TO RENORMALIZE THE STRESS TENSOR. AN IMPROVED FORMULA FOR THE POINT-SPLITTING PROCEDURE

Similarly to the previous work, we prove here that, in our general hypotheses, a particular (improved) form of the point-splitting procedure can be considered as a consequence of the ζ-function technique.

A. The point-splitting renormalization

Let us summarize the point-splitting approach to renormalize the one-loop stress tensor[4–6] in the Euclidean case. First of all, we want to rewrite (9) into a more convenient form. Employing the motion equations $A'\phi\equiv 0$ one can rewrite the right-hand side of (9) as

$$\begin{aligned}
T_{ab}[\phi,\mathbf{g}](x) = {} & (1-2\xi)[\nabla_a\phi(x)\nabla_b\phi(x) + \phi(x)\nabla_a\nabla_b\phi(x)] \\
& + \left(2\xi - \frac{1}{2}\right) g_{ab}(x)[\nabla_c\phi(x)\nabla^c\phi(x) + \phi(x)\Delta\phi(x)] \\
& + \left[\frac{g_{ab}(x)}{D}\phi(x)\Delta\phi(x) - \phi(x)\nabla_a\nabla_b\phi(x)\right] \\
& - \xi\left[R_{ab}(x) - \frac{g_{ab}(x)}{D}R(x)\right]\phi^2(x) - \frac{V'(x)+m^2}{D} g_{ab}(x)\phi^2(x).
\end{aligned} \tag{85}$$

Notice that the first two lines on the right-hand side of (85) produce a vanishing trace in the case of $\xi = \xi_D (:= (D-2)/[4(D-1)])$; the third and the fourth lines have separately a vanishing trace not depending on ξ. Finally, the trace of the last line is $-[V'(x)+m^2]\phi^2(x)$ trivially. It is obvious that, in the case of conformal coupling ($\xi=\xi_D$, $V'\equiv m^2 = 0$), the trace of the stress tensor vanishes. Conversely, for $\xi\neq\xi_D$ one also obtains

$$g^{ab}(x)T_{ab}(x) = \left(\frac{\xi_D-\xi}{4\xi_D - 1}\Delta - m^2 - V'(x)\right)\phi^2(x). \tag{86}$$

This is nothing but the *classical* version of (66). The most important difference is the lack of the trace anomaly term which is related to the last two terms on the right-hand side of (66).

The point-splitting procedure can be carried out employing the expression above for the stress tensor (actually one expects that the same final result should arise starting from different but equivalent expressions of the stress tensor). The basic idea is very simple.[4–6,15,16] One *defines* the a,b component of the one-loop renormalized stress tensor in the point y as the result of the following limit:

$$\langle T_{ab}(y)\rangle := \lim_{x\to y} \mathcal{D}_{ab}(x,y)\{\langle \phi(x)\phi(y)\rangle - \mathrm{H}(x,y)\}, \tag{87}$$

where the quantum average of the couple of fields is interpreted as the Green's function of the field equation corresponding to the quantum state one is considering, and $\mathrm{H}(x,y)$ is a Hadamard local fundamental solution[5,6,17] which has just the task of removing the argument-coincidence divergences from the Green's function *and from its derivatives* and does not depend on the quantum state. The operator $\mathcal{D}_{ab}(x,y)$ ''splits'' the point y and it is written down following (85), after an opportune symmetrization of the arguments (again, the final result should not depend on this symmetrization procedure),

$$\begin{aligned}\mathcal{D}_{ab}(x,y) = {} & \frac{1-2\xi}{2}[I_a^{a'}\nabla_{(x)a'}\nabla_{(y)b} + I_b^{b'}\nabla_{(x)b'}\nabla_{(y)a} + \nabla_{(y)a}\nabla_{(y)b} + I_a^{a'}I_b^{b'}\nabla_{(x)a'}\nabla_{(x)b'}] \\ & + \left(2\xi - \frac{1}{2}\right)\frac{g_{ab}(y)}{2}[2I_c^{c'}\nabla_{(x)c'}\nabla^{(y)c} + \Delta_x + \Delta_y] \\ & + \frac{1}{2}\left[\frac{g_{ab}(y)}{D}(\Delta_x + \Delta_y) - \nabla_{(y)a}\nabla_{(y)b} - I_a^{a'}I_b^{b'}\nabla_{(x)a'}\nabla_{(x)b'}\right] \\ & + \xi\left[R_{ab}(y) - \frac{g_{ab}(y)}{D}R(y)\right] - \frac{V'(y)+m^2}{D}g_{ab}(y).\end{aligned} \tag{88}$$

Above $I_a^{b'} = I_{(y)a}^{(x)b'}(y,x)$ is a generic component of the bitensor of parallel displacement from y to x, so the (co)tangent space at the point x is identified with the fixed (co)tangent space at the point y.

What one has to fix, in order to use (87) on a particular quantum state, is the Hadamard solution H. It is known that, in the case D is even, this solution is not unique[5,6] and is determined once one has fixed the term $w_0(x,y)$ [see Comment (2) of Theorem 2.6 in Ref. 1]. This term, differently from the case of the renormalization of the field fluctuations, is not completely arbitrary. Indeed, it is possible to show that there are terms w_0 producing a left-hand side of (87) which is *not* conserved.[5] Moreover, the massless conformally coupled case, and, more generally, the case $m=0$ and $V'\equiv 0$, involves some difficulties for the choice of w_0. For $m\neq 0$, it is possible to fix w_0 through the Schwinger–deWitt algorithm[4] obtaining a conserved renormalized stress tensor.[4,5] This is not possible for $m=0$ because Schwinger–deWitt's algorithm becomes singular in that case. Anyhow, there is a further prescription due to Adler, Lieberman, and Ng[18] (see also Refs. 3 and 5) which seems to overcome this drawback: this is the simplest choice $w_0(x,y)\equiv 0$. However, in the case of a massless conformally coupled field at least, as pointed out by Wald,[3] another drawback arises: the above prescription cannot produce a conserved stress tensor. Nevertheless, as proven in Ref. 3, in the case of (analytic in the cited reference) either a Lorentzian or Riemannian manifold, it is still possible to add a finite term on the right-hand side of (87) which takes account of the failure of the conservation law in order to have a conserved final left-hand side. This further term carries also a contribution to the trace of the final tensor which then fails to vanish and coincides with the well-known conformal anomaly. In Ref. 5, it has been argued that such an improved procedure can be generalized to any value of m and ξ getting

$$\langle T_{ab}(y)\rangle := g_{ab}(y)Q(y) + \lim_{x\to y} \mathcal{D}_{ab}(x,y)\{\langle \phi(x)\phi(y)\rangle - \mathrm{H}^{(0)}(x,y)\}, \tag{89}$$

where $\mathrm{H}^{(0)}$ is the Hadamard solution determined by the choice $w_0 \equiv 0$ and Q is a term fixed by imposing both the conservation of the left-hand side of (89) and the request that the renormalized stress tensor vanishes in the Minkowski vacuum. Employing the local ζ-function approach, we shall find out a point-splitting procedure which, in the case of a compact manifold, generalizes Wald's one for a general operator $-\Delta+V$ in $D>1$ dimensions in a Riemannian, not necessarily analytic, manifold and gives an explicit expression for Q automatically.

B. Local ζ-function and point-splitting procedure. An improved point-splitting prescription

In this part of the work we shall state a theorem concerning the relation between the two considered techniques proving their substantial equivalence within our general hypotheses.

Theorem 4.1: *Let us assume our general hypotheses on $\mathcal{M}$ and A' and suppose also $D>1$.*

(a) The renormalized strees tensor $\langle T_{ab}(y|A)\rangle_{\mu^2}$ defined in Definition 3.4 can be also computed as the result of a point-splitting procedure. Indeed one has, for any $\mu^2>0$,

$$\langle T_{ab}(y|A)\rangle_{\mu^2} = \lim_{x\to y} \mathcal{D}_{ab}(x,y)\{G(x,y|A) - \mathrm{H}_{\mu^2}(x,y)\} + \frac{g_{ab}(y)}{D}\left(\delta_D \frac{a_{D/2}(y,y|A)}{(4\pi)^{D/2}} - P_0(y,y|A)\right), \tag{90}$$

where $\mathcal{D}_{ab}$ is defined in (88), $G(x,y|A) := \mu^{-2}\zeta(1,x,y|A/\mu^2)$ is the μ^2-independent "Green's function" of A defined in Ref. 1, $P_0(y,y|A)$ is the C^∞ integral kernel of the projector on the kernel of A and $\mathrm{H}_{\mu^2}(x,y)$ is defined as (the summation appears for $D\geqslant 4$ only)

$$\mathrm{H}_{\mu^2}(x,y) = \sum_{j=0}^{D/2-2} (D/2-j-2)!\left(\frac{2}{\sigma}\right)^{D/2-j-1} \frac{a_j(x,y|A)}{(4\pi)^{D/2}} - \frac{a_{D/2-1}(x,y|A)}{(4\pi)^{D/2}}(2\gamma + \ln\mu^2)$$
$$- \frac{2a_{D/2-1}(x,y|A) - a_{D/2}(x,y|A)\sigma}{2(4\pi)^{D/2}} \ln\left(\frac{\sigma}{2}\right). \tag{91}$$

if D is even, and (the summation appears for $D\geqslant 5$ only)

$$\mathrm{H}_{\mu^2}(x,y) = \sum_{j=0}^{(D-5)/2} \frac{(D-2j-4)!!\sqrt{\pi}}{2^{(D-3)/2-j}}\left(\frac{2}{\sigma}\right)^{D/2-j-1} \frac{a_j(x,y|A)}{(4\pi)^{D/2}}$$
$$+ \frac{a_{(D-3)/2}(x,y|A)}{(4\pi)^{D/2}}\sqrt{\frac{2\pi}{\sigma}} - \frac{a_{(D-1)/2}(x,y|A)}{(4\pi)^{D/2}}\sqrt{2\pi\sigma} \tag{92}$$

if D is odd.

(b) H_{μ^2} is a particular Hadamard local solution of the operator A' truncated at the orders L,M,N. Indeed, one has

$$\mathrm{H}_{\mu^2}(x,y) = \frac{\Theta_D}{(4\pi)^{D/2}(\sigma/2)^{D/2-1}} \sum_{j=0}^{L} u_j(x,y)\sigma^j(x,y)$$
$$+ \delta_D\left(\sum_{j=0}^{M} v_j(x,y)\sigma^j\right)\ln\left(\frac{\sigma}{2}\right) + \delta_D \sum_{j=0}^{N} w_j(x,y)\sigma^j, \tag{93}$$

where $\delta_D=0$ if D is odd and $\delta_D=1$ if D is even; $\Theta_D=0$ for $D=2$ and $\Theta_D=1$ otherwise; and furthermore,

(1) *$L=D/2-2$, $M=1$, and $N=0$ for D even, and $L=(D-1)/2$ when D is odd.*

(2) *The coefficients u_j and v_j of the above Hadamard expansion are completely determined by fixing the value as $x\rightarrow y$ of the coefficient of the leading divergent term in order that this expansion for $L,M,N\rightarrow +\infty$ defines a Green's function formally. Using our conventions, this means*

$$u_0(y,y)=\frac{4\pi^{D/2}}{D(D-2)\omega_D}, \tag{94}$$

for $D\geqslant 3,\omega_D$ being the volume of the unitary D-dimensional disk, and

$$v_0(y,y)=\frac{1}{4\pi} \tag{95}$$

for $D=2$.

(3) *The coefficients w_j, when D is even, are completely determined by posing*

$$w_0(x,y):=-\frac{a_{D/2-1}(x,y|A)}{(4\pi)^{D/2}}[2\gamma+\ln\mu^2]. \tag{96}$$

Proof: See the Appendix. □

Comments:

(1) Whenever D is even, the logarithm in (93) contains a dimensional quantity. At first sight, this may look like a mistake. Actually, this apparent drawback means that the third summation in (93) has to contain terms proportional to $\ln\mu^2$ which can be reabsorbed in the second summation transforming the argument of the logarithm from $\sigma/2$ into the nondimensional one $\sigma\mu^2/2$. Indeed, the term on the right-hand side of (96) makes this job concerning the term v_0 in (93). Since (93) is computed up to $M=1$, one may expect the presence of a corresponding term w_1 in the last summation in (93). Actually, this term gives no contribution to the stress tensor employing (90) and (88) as one can check directly, taking into account that in any coordinate system around any $x\in\mathcal{M}$ (with an obvious meaning of the notations)

$$I_b^{a'}(x,y)|_{x=y}=\delta_b^{a'} \tag{97}$$

and

$$\begin{aligned}\nabla_{(x)a}\nabla_{(x)b}\sigma(x,y)|_{x=y}&=-\nabla_{(x)a}\nabla_{(y)b}\sigma(x,y)|_{x=y}\\&=\nabla_{(y)a}\nabla_{(y)b}\sigma(x,y)|_{x=y}=-\nabla_{(y)a}\nabla_{(x)b}\sigma(x,x')|_{x=y}=g_{ab}(y).\end{aligned} \tag{98}$$

In particular, one can check that each line on the right-hand side of (88) vanishes separately when it is evaluated for $x\rightarrow y$ on the considered terms of the Hadamard expansion. This is the reason we have put $N=0$ in (93) and we have omitted the corresponding term in (91). Notice that, conversely, in the usual version of point-splitting procedure,[4,3,5] the term $w_1(x,y)$ is necessary. Similarly, the terms of order $\sigma^n\ln\sigma$ with $n>1$ give no contribution to the stress tensor and thus we have omitted them in (93).

(2) In the case D is odd, the expansions (92) and (93) do not consider terms corresponding to $\sigma^{k+(1/2)}$ with $k=1,2,\ldots$. In fact, these terms give no contribution to the stress tensor via (90) and (88). Also in this case, each line on the right-hand side of (88) gives a contribution which vanishes separately for $x\rightarrow y$. Since (88) and (90) involve that the result does not depend on the coordinate system, one can check this fact working in Riemannian normal coordinates centered in y.

(3) The point-splitting procedure suggested in Ref. 5 for $D=4$, differently from our procedure, requires $w_0\equiv 0$ rather than (96). Actually, the function σ which appears in Ref. 5 is defined as two times our function σ. Therefore, taking into account that the argument of the logarithm in the second line of (91) is a quarter of Wald's one, Wald's prescription corresponds to taking

$w_0(x,y)-v_0(x,y)\ln 4=0$ in our case. Actually, as clarified in Ref. 3, the logarithm argument which appear in Wald's prescription has to be understood as $\ln(\sigma/u^2)$, where u is the unit of length employed. In our formalism, this corresponds, in particular, to performing the changes $\ln(\sigma/2)\rightarrow\ln[\sigma/(2u^2)]$ and $\ln\mu^2\rightarrow\ln(\mu^2 u^2)$ in (91). Equation (91) with the changes above entails that Wald's prescription, namely $w_0(x,y)-v_0(x,y)\ln 4=0$, is satisfied provided one fixes a μ^2 such that $2\gamma+\ln(\mu^2 u^2/4)=0$, namely $\mu=2e^{-\gamma}/u$.

This proves that, under our hypotheses, our prescription generalizes Wald's when the latter is understood in the Euclidean section of the manifold. Moreover, our prescription, different from Ref. 5, gives explicitly the form of the Hadamard local function to subtract to the Green's function in the general case as well as an explicit expression for the term Q in (89), in terms of heat-kernel coefficients and, trivially, for any choice of the value of μ^2.

(4) Rescaling the parameter μ^2, the expression of the final stress tensor changes by taking a term $(\ln\alpha)t_{ab}(y)$. We know the explicit form of such a term; indeed, it must be that given in the point (c) of Theorem 3.4. Notice also that the obtained point-splitting method, also concerning the rescaling of μ^2, agrees with the corresponding point-splitting procedure for computing the field fluctuations given in Ref. 1. For example, the point (d) of Theorem 3.4 holds, provided both sides are renormalized with the point-splitting procedures above and the same value of μ^2 is fixed.

(5) The point-splitting procedure we have found out uses the heat-kernel expansion in Theorem 1.3 of Ref. 1 and nothing further. This expansion can be built up also either in noncompact manifolds or manifolds containing boundary, essentially because it is based upon local considerations (see discussion in Ref. 3 concerning Schwinger–deWitt's expansion). Therefore, it is natural to expect that the obtained procedure, not depending on the ζ-function approach, may work in the general case (namely, it should produce a symmetric conserved stress tensor with the known properties of the trace also in noncompact or containing boundary manifolds), provided the Green's function of the considered quantum state has the Hadamard behavior.

(6) As a final comment, let us check on the found point-splitting method in the Euclidean section of Minkowski space–time which is out of our general hypotheses, without referring to the ζ-function approach. In this case, for $A=-\Delta+m^2$, one has that the heat-kernel referred to globally flat coordinates reads

$$\mathrm{K}(t,x,y|A)=\frac{e^{-\sigma/2t}}{(4\pi)^2t^2}e^{-m^2t}, \tag{99}$$

and thus, supposing $m^2>0$,

$$a_j(x,y|A)=\frac{(-1)^j m^{2j}}{j!}. \tag{100}$$

As is well known, the (Euclidean) Green's function of Minkowski vacuum can be computed directly:

$$G(x,y|A)=\int_0^{+\infty}\mathrm{K}(t,x,y)\,dt=\frac{2m}{(4\pi)^2\sqrt{\sigma/2}}\mathrm{K}_1\left(2\sqrt{\frac{\sigma m^2}{2}}\right). \tag{101}$$

Expanding K_1 in powers and logarithms of σ one gets

$$G(x,y|A)=\frac{2}{(4\pi)^2\sigma}+\frac{1}{(4\pi)^2}\left\{m^2+\frac{m^4}{4}\sigma+\sigma^2 f(\sigma)\right\}\ln\left(\frac{\sigma}{2}\right)+\frac{m^2}{(4\pi)^2}(2\gamma-1+\ln m^2)+\sigma g(\sigma), \tag{102}$$

where f and g are smooth bounded functions. Then, employing (90), (93), and (100), it is a trivial task to prove that, *provided* the choice $\mu=me^{-3/4}$ is taken in (96), one gets $\langle T_{ab}(y)\rangle\equiv 0$ as it is expected. In particular, one finds also $Q(y)\equiv m^4/(128\pi^2)$ for the coefficient of $g_{ab}(y)$ in (89) and in the last line of (90). For a general value of μ^2, the computation of the stress-tensor trace via the formula in (d) in Theorem 3.4 reproduces the correct Coleman–Weinberg results [19] for the field fluctuations still obtained by the local ζ function approach[13] as well as by using the point-splitting formula given in Theorem 2.6 in Ref. 1.

The case $m=0$ is much more trivial. In this case the heat kernel is given by (99) with $m=0$, and thus only $a_0(x,y)\equiv 1$ survives in the heat-kernel expansion. In this case, A is not positive defined but positive only, the manifold is not compact, and the Minkowski vacuum Green's function can still be computed integrating the heat-kernel despite that the local ζ-function does not exist. Moreover, the Green's function coincides with the Hadamard local solution $2/[(4\pi)^2\sigma]$. Furthermore, $Q(y)\equiv 0$, and thus our procedure gives a vanishing stress tensor as well.

V. SUMMARY AND OUTLOOKS

In this paper we have concluded the rigorous analysis started in Ref. 1 concerning the mathematical foundation of the theory of the local ζ-function renormalization of the one-loop stress tensor introduced in Ref. 2. The other important point developed herein has been the relation between the local ζ-function approach and the (Euclidean) point-splitting procedure.

Concerning the first point, we have proven that the ζ-function theory of the stress tensor can be rigorously defined, at least in closed manifold, giving results which agree with and generalize previous results concerning the ζ-function renormalization of the field fluctuations.[1] On the mathematical ground, we have also proven a few new theorems about the smoothness of the heat-kernel expansions.

Concerning the second proposed goal, we have found out that the two methods (ζ-function and point-splitting) agree essentially, provided a particular form of point-splitting procedure is employed. Within the hypotheses of a Riemannian compact C^∞ manifold, this point-splitting procedure is a natural generalization (in any $D>1$ and for a larger class of Euclidean motion operators) of Wald's improved procedure presented in Ref. 3 and also discussed in Ref. 5 defined in a Lorentzian manifold (but the same arguments employed can be trivially extended to Riemannian manifolds). Our procedure also gives explicitly the form of the various terms which are employed in the point-splitting procedure in terms of the heat-kernel expansion.

In our opinion, the found point-splitting procedure should work also without the employed hypotheses and independently from the ζ-function procedure. We have anyhow checked this conjecture in the Euclidean Minkowski space–time, proving that it holds true as expected either in the case $m=0$ or $m>0$. Moreover, the obtained results concerning the point splitting procedure should be trivially generalized for static Lorentzian manifolds at least.

ACKNOWLEDGMENTS

I am very grateful to D. Luminati for useful discussions. This work has been financially supported by a Postdoctoral Research Fellowship of the Department of Mathematics of the University of Trento.

APPENDIX: PROOF OF SOME LEMMATA AND THEOREMS

Proof of Lemma 3.1: Let us consider the form of the heat-kernel as it was built up in Ref. 20, Sec. 4, Chap. VI. This construction holds also in the case of an operator $A' := -\Delta + V$ and not only $A' := -\Delta$ as pointed out in Ref. 1. In our notations, one has by (45) in Sec. 4 of Chap. VI of Ref. 20

$$F_N(t,x,y)=\frac{e^{-\sigma(x,y)/2t}}{(4\pi t)^{D/2}}\chi(\sigma(x,y))\sum_{j=0}^{N} a_j(x,y|A)t^j, \tag{A1}$$

$N>D/2+2$ is a fixed integer.

[Actually, equation (45) in Sec. 4 of Chap. VI of Ref. 20 is misprinted in Ref. 20 because of the unnecessary presence of the operator L_x on the right-hand side of the first line of (45) in Sec.

4 of Chap. VI of Ref. 20. Since the absence of this operator in the correct formula, we cannot get the second line of (45) in a direct way. In fact, in Ref. 1 we have used a different (but equivalent in practice) form of the remaining of the heat-kernel expansion with respect to that which appears in (45). Some other parts of Sec. 4 of Chap. VI in Ref. 20 contain several other misprints like the requirement $F \in C^0(M \times M \times [0+\infty))$ in Lemma 2 in Ref. 20, which has to be corrected into $F \in C^1(M \times M \times [0+\infty))$.] Also,

$$\mathrm{K}(t,x,y|A) = F_N(t,x,y) + (F_N * F)(t,x,y,), \tag{A2}$$

where F_N is the $C^\infty((0,+\infty)\times\mathcal{M}+\mathcal{M})$ parametrix defined in Ref. 1.

The remaining proportional to $O_{\eta,N}$ in (20) of Lemma 3.1 of Ref. 1 is therefore $(F_N * F) \times (t,x,y)$. We remind the reader that $\sigma(x,y)$ is one-half the squared geodesical distance $(d(x,y))$ from x to y and defines an everywhere continuous function on $\mathcal{M}\times\mathcal{M}$ which is also C^∞ in the set of the points x,y such that $d(x,y)<r$. Here $\chi(u)$ is a non-negative $C^\infty([0,+\infty))$ function which takes the constant value 1 for $|u|<r^2/16$ and vanishes for $|u|\geqslant r^2/4$, r being the injectivity radius of the manifold. The convolution $*$ has been defined in Sec. 4 of Chap. VI of Ref. 20,

$$(G*\mathrm{H})(t,x,y) := \int_0^t d\tau \int_{\mathcal{M}} d\mu_g(z) G(\tau,x,z)\mathrm{H}(t-\tau,z,y), \tag{A3}$$

whenever the right-hand side makes sense.

Finally, the function F which appears in (A2) is defined by a uniformly convergent series in $[0,T]\times\mathcal{M}\times\mathcal{M}$ for any $T>0$ [see (43) in Sec. 4 of Chap. VI of Ref. 20].

$$F(t,x,y) := \sum_{l=1}^{\infty} [(A'_x - \partial/\partial t)F_N]^{*l}(t,x,y) \tag{A4}$$

(B^{*l} means $B*B*\ldots*Bl$ times). This function belongs to $C^L([0,+\infty)\times\mathcal{M}\times\mathcal{M})$ provided $M > D/2+2L$ (see Ref. 20). Equation (A2) satisfies the heat-kernel equation provided F is C^1 in all variables, namely $N>D/2+2$.

Let us consider (A2). The remainder of the ''asymptotic'' expansion of the heat kernel computed up to the coefficient $a_N(x,y|A)(N>D/2+2)$ is just the second term on the right-hand side. It can be explicitly written down (see Ref. 20)

$$(F_N*F)(t,x,y) = \int_0^t d\tau\, \tau^{-D/2}(t-\tau)^{N-D/2} \int_{\mathcal{M}} d\mu_g(z) \mathcal{F}_N(\tau,x,z)\mathcal{F}(t-\tau,z,y) \times e^{-\sigma(x,z)/2\tau} e^{-\sigma(z,y)/2(t-\tau)}. \tag{A5}$$

Here $\mathcal{F}_N(t,x,z)$ defines a function which belongs to $C^\infty([0,+\infty)\times\mathcal{M}\times\mathcal{M})$ and vanishes smoothly whenever the geodesical distance between x and z is sufficiently large, i.e., $d(x,z) \geqslant r/2$, due to the presence of the function χ in the expression of the parametrics (A1). $\mathcal{F}$ defines an everywhere continuous function which belongs also to C^L provided $N>D/2+2L$ and the geodesical distance between y and z is sufficiently short, i.e., $d(y,z)<r$, and $t\in[0,+\infty)$.

Then let us pick out a point $u\in\mathcal{M}$. We can find a geodesically spherical open neighborhood of u,J_u, with a geodesic radius $r_0<r/8$. By the definition of the function χ, it holds $\chi(\sigma(x,y)) = 1$ whenever $x,y\in J_u$ and thus the coefficient χ can be omitted in the heat-kernel expansion working with any coordinate system defined in a neighborhood of J_u (e.g., normal Reimannian coordinates). From now on concerning the points x and y we shall work within such a coordinate system in the neighborhood J_u. Notice also that, by the triangular inequality $d(x,y) [=\sqrt{2\sigma(x,y)}]<r/4$ whenever $x,y\in J_u$.

Now, let us suppose $N>D/2+2|\alpha|+2|\beta|$. This entails $F\in C^{|\alpha|+|\beta|}$ and thus also $\mathcal{F} \in C^{|\alpha|+|\beta|}$ provided the distance of its arguments defined on the manifold is less than r and $t \in (0,+\infty)$. We can apply operators D_x^α and D_y^β to both sides of (A2). The action of the derivatives

(A2) produces the first term on the right-hand side of (20) at least (notice that $\chi\equiv 1$ in our hypotheses). Let us focus attention on the action of the derivatives on the remaining in (A2). Our question concerns the possibility of passing these under the integration symbol in (A5). The action of the derivatives can be carried under integration symbols (obtaining also an x,y-continuous final function if the derivatives of the integrand are continuous) provided, for any fixed choice of a couple of of multindices α,β, the derivatives of the integrand are locally x,y-uniformly bounded by an integrable function (dependent on the multindices in general). We shall see that this is the case after we have manipulated the integral opportunely. Notice that the derivatives (with respect to the manifold variables) of the function $\mathcal{F}$ do exist because the second integral on the right-hand side of (A5) takes contribution only from the points z such as both $d(y,z)<r$ and $d(x,z)<r$ are fulfilled as required above. Indeed, it must be $d(x,y)<r/2$, otherwise, $\mathcal{F}_N(\tau,x,z)$ smoothly vanishes as pointed out above, and, taking account of $d(x,y)<r/4$, the triangular inequality entails also $d(y,z)\leqslant d(x,z)+d(x,y)<r/2+r/4=3r/4$.

Now, let us fix a new open neighborhood of u, I_u, such that its closure is contained in J_u, and fix $T>0$. Barring $\tau\mapsto\tau^{-D/2}$, all functions of τ,x,y,z and all their (x,y,z-)derivatives we shall consider are bounded in the compact $[0,T]\times\overline{I_u}\times\overline{I_u}\times\mathcal{M}$ where we are working because these are continuous therein. We can rearrange the expression (A5) into

$$(F_N*F)(t,x,y)=\int_0^t d\tau\tau^{-D/2}(t-\tau)^{N-D/2}\int_{S^{D-1}}d\vec{v}\int_0^{+\infty}d\rho\rho^{D-1}J(x,\vec{v},\rho)$$
$$\times e^{-\rho^2/2\tau}\mathcal{F}_M(\tau,x,z(x,\rho,\vec{v}))\mathcal{F}(t-\tau,z(x,\rho,\vec{v}),y)\times e^{-\sigma(z(x,\rho,\vec{v}),y)/2(t-\tau)}. \tag{A6}$$

where, to determine the position of z, we have employed a spherical system of normal coordinates $\rho,\vec{v}$ centered in any x, ρ is the distance of z from x, its range is maximized in the integrals above because the integrand vanishes smoothly for $\rho>r/2$, and thus all the functions contained in the integrand are well-defined within $\{\rho\in[0,+\infty)\}$. $\vec{v}$ is a unitary $(D-1)$-dimensional vector and $d\mu_g(z)=d\rho d\vec{v}\rho^{D-1}J(x,\vec{v},\rho)$, $d\vec{v}$ is the natural measure on S^{D-1}. The function J is continuous and bounded in $\overline{I_u}\times S^{D-1}\times\{\rho\in[0,r/2]\}$ together with all derivatives.

Then, we can change variables $\rho\mapsto\rho/\sqrt{\tau}=:\rho'$ obtaining

$$(F_N*F)(t,x,y)=\int_0^t d\tau(t-\tau)^{N-D/2}\int_{S^{D-1}}d\vec{v}\int_0^{+\infty}d\rho'\rho'^{D-1}J(x,\vec{v},\tau^{1/2}\rho')$$
$$\times e^{-\rho'^2/2}\mathcal{F}_M(\tau,x,z(x,\tau^{1/2}\rho',\vec{v}))\mathcal{F}(t-\tau,z(x,\tau^{1/2}\rho',\vec{v}),y)$$
$$\times e^{-\sigma(z(x,\tau^{1/2}\rho',\vec{v}),y)/2(t-\tau)}. \tag{A7}$$

The formal action of the operators D_x^α and D_y^β under the integration produces a sum of continuous and bounded functions (now the function $\tau\mapsto\tau^{-D/2}$ has disappeared and the remaining functions and their x,y,z-derivatives are bounded since they are product of bounded functions). Also, it changes $(t-\tau)^{N-D/2}$ into several terms of the form $(t-\tau)^{N-D/2-L_i}$ (where each $L_i\leqslant|\alpha|+|\beta|$), because of the derivatives of the second exponential. These functions of τ are continuous and bounded being $N>D/2+|\alpha|+|\beta|\geqslant L_i$ in our hypotheses. We can bound the absolute value of these functions by $Ce^{-\rho'^2/2}$, where C is a sufficiently large constant. This function is trivially integrable in the measure we are considering. This x,y,t-uniform bound assures that, concerning the x,y-derivatives of F_N*F, one can interchange the symbols of derivatives with those of integrals and also that the derivative of $(t,x,y)\mapsto(F_N*F)(t,x,y)$ are continuous functions in $(0,+\infty)\times\mathcal{I}_u\times\mathcal{I}_u$.

In order to finish this proof, let us consider a finer estimate of $O_{\eta,N}^{(\alpha,\beta)}(x,y)$. We have the inequality[20] for $\tau\in[0,t]$

$$\frac{d^2(x,y)}{t} \leqslant \frac{d^2(x,z)}{\tau} + \frac{d^2(z,y)}{t-\tau}, \tag{A8}$$

and thus, picking out any $\eta \in (0,1)$ and posing $\delta := 1-\eta \in (0,1)$, we get (notice that $t-\tau \geqslant 0$)

$$e^{-\sigma(x,z)/2\tau} e^{-\sigma(z,y)/2(t-\tau)} \leqslant e^{-\eta\sigma(x,y)/2t} (e^{-\delta\sigma(x,z)/2\tau} e^{-\delta\sigma(z,y)/2(t-\tau)}) \leqslant e^{-\eta\sigma(x,y)/2t} e^{-\delta\sigma(x,z)/2\tau}. \tag{A9}$$

We can use this relation in the x,y-derivatives of (A7), obtaining

$$|D_x^\alpha D_y^\beta (F_N * F)(t,x,y)| \leqslant \sum_i e^{-\eta\sigma(x,y)/2t} \int_0^t d\tau (t-\tau)^{N-D/2-L_i} \int_{S^{D-1}} d\vec{v} \int_0^{+\infty} d\rho' \rho'^{D-1} e^{-\delta\rho'^2/2} C_i, \tag{A10}$$

where the coefficients C_i are upper bounds of the absolute values of the continuous functions missed in the integrand and $L_i \leqslant |\alpha|+|\beta|$. We can execute the integral in τ obtaining, for $0<t \leqslant T$ and $x,y \in I_u$ (remember that $N>D/2+|\alpha|+|\beta|$)

$$|D_x^\alpha D_y^\beta (F_N * F)(t,x,y)| \leqslant e^{-\eta\sigma(x,y)/2t} \sum_i C'_{i,\delta} t^{N+1-L_i-D/2} \leqslant \frac{C_\delta}{(4\pi)^{D/2}} e^{-\eta\sigma(x,y)/2t} t^{N+1-D/2-|\alpha|-|\beta|}, \tag{A11}$$

where C_δ is a positive constant sufficiently large which depends on T,α,β in general. This proves the remaining part of the thesis posing $\mathrm{K}_{\eta,N}^{(\alpha,\beta)} := T$ and $M_{\eta,N}^{(\alpha,\beta)} := C_\delta$. Indeed, the remaining $O_{\eta,N}^{(\alpha,\beta)}$ we wanted to compute coincides with $D_x^\alpha D_y^\beta (F_N * F)$ just up to the factor $(4\pi t)^{-D/2} t^{N-|\alpha|-|\beta|} \exp(-\eta\sigma/2t)$. $O_{\eta,N}^{(\alpha,\beta)}$ can be defined in $t=0$ as $O_{\eta,N}^{(\alpha,\beta)}(0;x,y)=0$, obtaining a continuous function in $[0,+\infty) \times I_u \times I_u$. □

Proof of Lemma 3.2: Let us consider an eigenvector ϕ_j and fix $T \in (0,+\infty)$ and consider a neighborhood of $u \in \mathcal{M}, J_u$ where a coordinate system is defined. In the following, x and y are points in a new neighborhood of u, I_u, such that its closure is contained in J_u. These points are represented by the coordinate system given above and the derivative operators are referred to these coordinates. From Theorem 1.3 of Ref. 1, it holds

$$e^{-T\lambda_j} \phi_j(x) = \int_{\mathcal{M}} d\mu_g(z) \mathrm{K}(T,x,z|A) \phi_j(z). \tag{A12}$$

We can derive both sides of the equation above employing operators D_x^α. Since, for a fixed T the derivatives of K are bounded $[(x,z) \mapsto \mathrm{K}(T,x,z|A)$ is C^∞ and $\overline{I_u} \times \mathcal{M}$ is compact in our hypotheses[1]], we can pass the derivative operator under the integral symbol obtaining

$$|D_x^\alpha \phi_j(x)| = |e^{\lambda_j T} \int_{\mathcal{M}} d\mu_g(z) D_x^\alpha \mathrm{K}(t,x,z|A) \phi_j(z)| \leqslant e^{\lambda l_j T} \|D_x^\alpha \mathrm{K}(T,x,.|A)\|_{L^2(\mathcal{M},d\mu_g)}, \tag{A13}$$

where we have made use of the Cauchy–Schwarz inequality and we have taken account of $\|\phi_j\|=1$ [from now on we omit the index $L^2(\mathcal{M},d\mu_g)$ in the norms because there is no ambiguity]. The function $x \mapsto \|D_x^\alpha \mathrm{K}(T,x,.|A)\|$, for $x \in \overline{I_u}$ is continuous from Lebesgue's dominate convergence theorem since $D_x^\alpha \mathrm{K}(T,x,y|A)$ defines a continuous function in x and y and there is a constant (dependent on T, in general) M_T such that $|D_x^\alpha \mathrm{K}(T,x,z|A)|^2 \leqslant M_T$ for (x,z) which belong in the compact $\overline{I_u} \times \mathcal{M}$ and the measure of the manifold is finite. The same results holds whenever one keeps fixed y in I_u and integrates over x. Therefore, let us define

$$P_T^{(\alpha,\beta)} := [\sup_{x \in \overline{I_u}} \|D_x^\alpha \mathrm{K}(T,x,.|A)\|][\sup_{y \in \overline{I_u}} \|D_y^\beta \mathrm{K}(T,.,y|A)\|] \tag{A14}$$

and we have, for any $x,y \in I_u$, the λ_j-uniform upper bound

$$|e^{-\lambda_j t} D_x^\alpha \phi_j(x) D_y^\beta \phi_j^*(y)| \leqslant P_T^{(\alpha,\beta)} e^{-\lambda_j(t-2T)}. \tag{A15}$$

The found inequality proves that the absolute values of the terms of the series

$$\sum_{j\in\mathbb{N}}{}' e^{-\lambda_j t} D_x^\alpha \phi_j(x) D_y^\beta \phi_j^*(y) \tag{A16}$$

are x,y-uniformly bounded, for $(x,y,t)\in I_u \times I_u \times (2T,+\infty)$, by terms of the convergent series (see (30) in Ref. 1)

$$\sum_{j\in\mathbb{N}}{}' e^{-\lambda_j(t-2T)} P_T^{(\alpha,\beta)} = P_T^{(\alpha,\beta)} \int_{\mathcal{M}} d\mu_g(z) \{ \mathrm{K}(t-2T,z,z|A) - P_0(z,z|A)\}$$
$$= P_T^{(\alpha,\beta)} \,\mathrm{Tr}\{\mathrm{K}_{(t-2T)} - P_0\}.$$

This holds for any choice of the multindices α,β and this entails (23), (25), and (26). The final upper bound (26) is a trivial consequence of (99) in Ref. 1 and the fact that the manifold has a finite measure. □

Sketch of Proof of Theorem 4.1: Let us fix a coordinate system in a neighborhood I_u of a point $u\in\mathcal{M}$. All the following considerations will refer to these coordinates, and in particular to a couple of points x,y within that neighborhood. Then, let us consider the expression (58) for the ζ-function of the stress tensor. Employing the eigenvalue equation $A\phi_j = \lambda_j \phi_j$ one can rearrange (58) into

$$Z_{ab}(s,y|A/\mu^2) = \sum_{j\in\mathbb{N}}{}' \frac{2s}{\mu^2}\left(\frac{\lambda_j}{\mu^2}\right)^{-(s+1)} T'_{ab}[\phi_j,\phi_j^*,\mathbf{g}](y), \tag{A17}$$

where (C.C. means the complex conjugation of the terms already written)

$$\begin{aligned} T'_{ab}[\phi_j,\phi_j^*,\mathbf{g}](y) = {}&(1-2\xi)\frac{1}{2}(\nabla_a\phi_j(y)\nabla_b\phi_j^*(y) + \phi_j(y)\nabla_a\nabla_b\phi_j^*(y) + \text{C.C.})\\ &+\left(2\xi - \frac{1}{2}\right)\frac{g_{ab}(x)}{2}(\nabla_c\phi_j(y)\nabla^c\phi_j^*(y) + \phi_j(y)\Delta\phi_j^*(y) + \text{C.C.})\\ &+\frac{1}{2}\left(\frac{g_{ab}(y)}{D}\phi_j(y)\Delta\phi_j^*(y) - \phi_j(y)\nabla_a\nabla_b\phi_j^*(y) + \text{C.C.}\right)\\ &+\xi\left(R_{ab}(y) - \frac{g_{ab}(y)}{D}R(y)\right)|\phi_j(y)|^2 - \frac{V'(y)+m^2}{D}g_{ab}(y)|\phi_j(y)|^2\\ &+\frac{\lambda_j}{D}g_{ab}(y)|\phi_j(y)|^2. \end{aligned} \tag{A18}$$

The stress tensor is then given by (60) in Definition 3.4 after the analytic continuation in the variable s of $Z_{ab}(s,y|A/\mu^2)$ given in (A17). Employing Theorem 3.2 and (A18) and (A17), we can write down the expression of $Z_{ab}(s,y|A/\mu^2)$, employing also functions $\zeta^{[\alpha,\beta]}(s,y|A/\mu^2)$ defined as in Definition 3.2 with the difference that covariant derivatives are employed instead of ordinary derivatives. We get, omitting the arguments y and A/μ^2 in the various ζ-functions for the sake of brevity,

$$
\begin{aligned}
Z_{ab}(s,y|A/\mu^2)=&(1-2\xi)\,\frac{s}{\mu^2}\,(\zeta^{(1_a,1_b)}(s+1)+\zeta^{(1_b,1_a)}(s+1)+\zeta^{[1_a+1_b,0]}(s+1)\\
&+\zeta^{[0,1_a+1_b]}(s+1))+\left(2\xi-\frac{1}{2}\right)\frac{s g_{ab}(y)g^{cd}(y)}{\mu^2}\,(2\zeta^{(1_c,1_d)}(s+1)\\
&+\zeta^{[0,1_c+1_d]}(s+1)+\zeta^{[1_c+1_d,0]}(s+1))+\frac{s}{\mu^2}\left[\frac{g_{ab}(y)g^{cd}(y)}{D}\,(\zeta^{[0,1_c+1_d]}(s+1)\right.\\
&\left.+\zeta^{[1_c+1_d,0]}(s+1))-\zeta^{[0,1_a+1_b]}(s+1)-\zeta^{[1_a+1_b,0]}(s+1)\right]\\
&+\frac{2s\xi}{\mu^2}\left(R_{ab}(y)-\frac{g_{ab}(y)}{D}R(y)\right)\zeta(s+1)-\frac{V'(y)+m^2}{D}\,\frac{2sg_{ab}(y)}{\mu^2}\,\zeta(s+1)\\
&+\frac{2sg_{ab}(y)}{D}\,\zeta(s). \qquad \text{(A19)}
\end{aligned}
$$

First of all, we notice that the term proportional to $g_{ab}(y)$ in (90) arises from the last term above via item (c) of Theorem 2.2 in Ref. 1.

Let us consider the contribution to the stress tensor due to the terms $\zeta^{(1_a,1_b)}(s+1,y|A/\mu^2)$. Similarly to (101) in Ref. 1, we can define, for any $\mu_0^2>0$ fixed and N integer $>D/2+4$, taking account of Lemma 3.1 above:

$$
\begin{aligned}
\zeta^{(1_a,1_b)}(N,s+1,x,y|A/\mu^2,\mu_0^{-2}):=&\frac{\mu^{2s}}{\Gamma(s+1)}\int_0^{\mu_0^{-2}}dt\,t^s\frac{e^{-\eta\sigma(x,y)/2t}}{(4\pi t)^{D/2}}\,t^{N-2}O^{(1_a,1_b)}_{\eta,N}(t;x,y)\\
&+\frac{\mu^{2s+2}}{\Gamma(s+1)}\int_{\mu_0^{-2}}^{+\infty}dt\,t^s\nabla_{(x)a}\nabla_{(y)b}[\mathrm{K}(t,x,y|A)-P_0(x,y|A)]. \qquad \text{(A20)}
\end{aligned}
$$

Similarly to Lemma 2.1 in Ref. 1, one can prove that the function of s,x,y defined above is continuous in a neighborhood $I\times I_u\times I_u$, where I is a complex neighborhood of $s=0$ with all of its s derivatives, in particular, it is s-analytic therein. Employing Lemma 3.1 and the item (a) of Theorem 3.1 one can write also, for $\mathrm{Re}\,s+1>D/2+4$,

$$
\begin{aligned}
\zeta^{(1_d,1_b)}(s+1,x,y|A/\mu^2)=&\zeta^{(1_a,1_b)}(N,s+1,x,y|A/\mu^2,\mu_0^{-2})-\left(\frac{\mu}{\mu_0}\right)^{2s+2}\frac{\nabla_{(x)a}\nabla_{(x)b}P_0(x,y|A)}{(s+1)\Gamma(s+1)}\\
&+\frac{\mu^{2s+2}}{(4\pi)^{D/2}\Gamma(s+1)}\sum_{j=0}^{N}\nabla_{(x)a}\nabla_{(y)b}\left(\int_0^{\mu_0^{-2}}dt\,t^{s-D/2+j}e^{-\sigma/2t}a_j(x,y|A)\right). \qquad \text{(A21)}
\end{aligned}
$$

In particular, for $\mathrm{Re}\,s+1>D/2+4$, the left-hand side above is continuous in x,y and thus we can take the coincidence limit for $x\to y$. Noticing that one can also pass the derivatives under the sign of integration on the right-hand side and that $\nabla_{(x)a}\nabla_{(y)b}\sigma(x,y)|_{x=y}=-g_{ab}(y)$ and $\nabla_c\sigma(x,y)|_{x=y}=0$, we get for the right-hand side of the expression above multiplied by $2s/\mu^2$ and evaluated for $x=y$

$$\frac{2s}{\mu^2}\zeta^{(1_a,1_b)}(N,s+1,y,y|A/\mu^2,\mu_0^{-2})-\frac{2s}{\mu^2}\left(\frac{\mu}{\mu_0}\right)^{2s+2}\frac{P_{0ab}(x,y|A)}{(s+1)\Gamma(s+1)}$$
$$+\frac{2s}{\mu^2}\frac{\mu^{2s+2}}{(4\pi)^{D/2}\Gamma(s+1)}\sum_{j=0}^{N}\left\{\frac{a_{jab}(y,y|A)\mu_0^{-2(s-D/2+j+1)}}{s-D/2+j+1}\right.$$
$$\left.+\frac{g_{ab}(y)}{2}\frac{a_j(y,y|A)\mu_0^{-2(s-D/2+j)}}{s-D/2+j}\right\},$$

where $a_{jab}(y,y|A):=\nabla_{(x)a}\nabla_{(y)b}a_j(x,y|A)|_{x=y}$, and $P_{jab}(y,y|A):=\nabla_{(x)a}\nabla_{(y)b}P_0(x,y|A)|_{x=y}$. The contribution to the stress tensor, namely, to $(d/ds)|_{s=0}Z_{ab}(s,y|A/\mu^2)/2$ of the considered term, is then obtained by continuing the result above as far as $s=0$, executing the s derivative and multiplying for $(1-2\xi)/2$ the final result. Taking account that $\zeta(N,s,x,y|A/\mu^2,\mu_0^{-2})$ is smooth in a neighborhood of $s=0$, this leads to, apart from the unessential factor $(1-2\xi)$,

$$\left.\frac{d}{ds}\right|_{s=0}\frac{2s}{\mu^2}\zeta^{(1_a,1_b)}(s+1,y|A/\mu^2)$$
$$=\frac{1}{\mu^2}\zeta^{(1_a,1_b)}(N,1,y,y|A/\mu^2,\mu_0^{-2})-\frac{P_{0ab}(y,y|A)}{\mu^2}$$
$$+\frac{1}{(4\pi)^{D/2}}\sum_{j=0,j\neq D/2-1}^{N}\frac{a_{jab}(y,y|A)}{\mu_0^{2j-D+2}(j-D/2+1)}+\delta_D\left(\gamma+2\ln\frac{\mu}{\mu_0}\right)\frac{a_{(D/2-1)ab}(y,y|A)}{(4\pi)^{D/2}}$$
$$+\frac{g_{ab}(y)}{(4\pi)^{D/2}}\sum_{j=0,j\neq D/2}^{N}\frac{a_j(y,y|A)}{\mu_0^{2j-D}(j-D/2)}+\delta_D g_{ab}(y)\left(\gamma+2\ln\frac{\mu}{\mu_0}\right)\frac{a_{D/2}(y,y|A)}{(4\pi)^{D/2}}. \tag{A22}$$

Let us consider the first line on the right-hand side of (A19) for a moment. The other terms different from $\zeta^{(1_a,1_b)}(s+1)$ can be undertaken to a procedure similar to that developed above. The important point is that, once one has performed such a procedure, all terms with a factor $g_{ab}(y)$ similar to the terms in the last line of (A22) cancel out each other, and thus, in the final expression of the first line on the right-hand side of (A18), no term with a factor $g_{ab}(y)$ survives. The same fact happens for the second and third lines of (A19). Since $\zeta^{(1_a,1_b)}$ $\times(N,1,y,y|A/\mu^2,\mu_0^{-2})$ and the derivatives of heat-kernel coefficients are continuous in x,y we can compute the right-hand side of (A22) as a limit of coincidence:

$$\left.\frac{d}{ds}\right|_{s=0}\frac{2s}{\mu^2}\zeta^{(1_a,1_b)}(s+1,y|A/\mu^2)$$
$$=\lim_{x\to y}\left\{\frac{1}{\mu^2}\zeta^{(1_a,1_b)}(N,1,x,y|A/\mu^2,\mu_0^{-2})-\frac{P_{0ab}(x,y|A)}{\mu^2}\right.$$
$$+\frac{1}{(4\pi)^{D/2}}\sum_{j=0,j\neq D/2-1}^{N}\frac{a_{jab}(x,y|A)}{\mu_0^{2j-D+2}(j-D/2+1)}+\delta_D\left(\gamma+2\ln\frac{\mu}{\mu_0}\right)\frac{a_{(D/2-1)ab}(x,y|A)}{(4\pi)^{D/2}}$$
$$+\frac{g_{ab}(y)}{(4\pi)^{D/2}}\sum_{j=0,j\neq D/2}^{N}\frac{a_j(x,y|A)}{\mu_0^{2j-D}(j-D/2)}+\delta_D g_{ab}(y)\left(\gamma+2\ln\frac{\mu}{\mu_0}\right)\frac{a_{D/2}(x,y|A)}{(4\pi)^{D/2}}. \tag{A23}$$

Moreover, since the function in the limit is continuous, the same limit can be computed by identifying the tangent space at x with the tangent space at y and thus introducing the bitensor $I_a^{a'}=I_{(y)a}^{(x)a'}(y,x)$ of parallel displacement from y to x as usual. Employing (A21) we finally get

$$(1-2\xi)\frac{d}{ds}\bigg|_{s=0}\frac{2s}{\mu^2}\zeta^{(1_a,1_b)}(s+1,y|A/\mu^2)$$
$$=\lim_{x\to y}(1-2\xi)I_a^{a'}\nabla_{(x)a'}\nabla_{(y)b}\left\{\frac{1}{\mu^2}\zeta(1,x,y|A/\mu^2)-\mathrm{H}_N(x,y)\right\}+(1-2\xi)g_{ab}(y)\mathrm{H}'(y), \tag{A24}$$

where, as we said above, the final term proportional to $g_{ab}(y)$ gives no contribution to the final stress tensor because it cancels against similar terms in the first line of (A18). The explicit form of H_N reads

$$\mathrm{H}_N(x,y)=\sum_{j=0}^{N}\frac{a_j(x,y|A)}{(4\pi)^{D/2}}\int_0^{\mu_0^{-2}}t^{j-D/2}e^{-\sigma(x,y)/2t}-\frac{1}{(4\pi)^{D/2}}\sum_{j=0,j\neq D/2-1}^{N}\frac{a_j(x,y|A)}{\mu_0^{2j-D+2}(j-D/2+1)}$$
$$-\delta_D\frac{a_{D/2-1}(x,y|A)}{(4\pi)^{D/2}}\left[\gamma+\ln\left(\frac{\mu}{\mu_0}\right)^2\right]. \tag{A25}$$

The same procedure has to be used for each term on the right-hand side of (A18) except for the last term which, as it stands, produces the last term on the right-hand side of (90). Summing all contributions, one gets (90) with H_N in place of H_{μ^2}. Anyhow, executing the integrations above using the results (52)–(58) in Ref. 1 $(D>1)$, expanding H_N in powers and logarithm of σ and taking account of Comments (1) and (2) after Theorem 4.1 above, we have that, in the expansion of H_N one can take account only of the terms pointed out in the item (a) of Theorem 4.1; these are the only terms which do not contain the arbitrary parameter μ_0^2 (which cannot remain in the final result). Therefore, the part of H_N which gives contributions to the final stress tensor coincides with H_{μ^2} given in (91) and (92).

This proves the point (a) of Theorem 4.1. The point (b) is trivially proven by a direct comparison between (23)–(25) in Ref. 1 and the equation for the coefficients of the Hadamard local solution given in Chap. 5 of Ref. 17 which determines completely the coefficients u_j and v_j of the local solution once the values of the coefficients of the leading divergences are fixed for $x\to y$, and the coefficients w_j once w_0 has been fixed. In performing this comparison, concerning the normalization conditions (94) and (95) in particular, notice that the measure used in the integrals employed in Ref. (17) is the Euclidean one $d^n x$ instead of our measure $\sqrt{g(x)}d^n x$. □

[1] V. Moretti, Commun. Math. Phys. **201**, 327 (1999).
[2] V. Moretti, Phys. Rev. D **56**, 7797 (1997).
[3] R. M. Wald, Phys. Rev. D **17**, 1477 (1978).
[4] N. D. Birrel and P. C. W. Davies, *Quantum Fields in Curved Space* (Cambridge U.P., Cambridge, 1982).
[5] R. M. Wald, *Quantum Field Theory and Black Hole Thermodynamics in Curved Spacetime* (Univ. Chicago, Chicago, 1994).
[6] S. A. Fulling, *Aspects of Quantum Field Theory in Curved Space-Time* (Cambridge U.P., Cambridge, 1991).
[7] R. M. Wald, *General Relativity* (Univ. Chicago, Chicago, 1984).
[8] R. M. Wald, Commun. Math. Phys. **70**, 226 (1979).
[9] S. W. Hawking, Commun. Math. Phys. **55**, 133 (1977).
[10] E. Elizalde, S. D. Odintsov, A. Romeo, A. A. Bytsenko, and S. Zerbini, *Zeta Regularization Techniques with Applications* (World Scientific, Singapore, 1994).
[11] E. Elizalde, *Ten Physical Applications of Spectral Zeta Functions* (Springer-Verlag, Berlin, 1995).
[12] J. Schwinger, Proc. Natl. Acad. Sci. USA **37**, 452 (1951).
[13] D. Iellici and V. Moretti, Phys. Lett. B **425**, 33 (1998).
[14] W. Tichy and E. E. Flanagan, Phys. Rev. D **58**, 124007 (1998).
[15] S. M. Christensen, Phys. Rev. D **14**, 2490 (1976).
[16] S. M. Christensen, Phys. Rev. D **17**, 946 (1978).
[17] P. R. Garabedian, *Partial Differential Equations* (Wiley, New York, 1964).
[18] S. L. Adler, J. Lieberman, and Y. J. Ng, Ann. Phys. (N.Y.) **106**, 279 (1977).
[19] S. Coleman and E. Weinberg, Phys. Rev. D **7**, 1888 (1973).
[20] I. Chavel, *Eigenvalues in Riemannian Geometry* (Academic, Orlando, 1984).

Inverse problem for an inhomogeneous Schrödinger equation

A. G. Ramm[a)]
Department of Mathematics, Kansas State University, Manhattan, Kansas 66506-2602

(Received 8 March 1999; accepted for publication 27 April 1999)

Let $(l-k^2)u=-u''+q(x)u-k^2u=\delta(x)$, $x\in\mathbb{R}$, $\partial u/\partial|x|-iku\to 0$, $|x|\to\infty$. Assume that the potential $q(x)$ is real valued and compactly supported: $q(x)=\overline{q(x)}$, $q(x)=0$ for $|x|\geqslant 1$, $\int_{-1}^{1}|q|dx<\infty$, and that $q(x)$ produces no bound states. Let $u(-1,k)$ and $u(1,k)$, $\forall k>0$ be the data. It is shown that under the above assumptions these data determine $q(x)$ uniquely. © *1999 American Institute of Physics.* [S0022-2488(99)02108-8]

I. INTRODUCTION

For several decades, the following inverse problems of practical interest are open. Let

$$\nabla^2 u+k^2u+k^2v(x)u=-\delta(x), \quad \text{in } \mathbb{R}^3, \tag{1.1}$$

u satisfies the radiation condition at infinity, and $v(x)$ is a compactly supported piecewise-smooth function, supp $v\subset\mathbb{R}^3_-:=\{x:x_3<0\}$.

The data are the values $u(x_1,x_2,0,k)$ for all $\hat{x}:=(x_1,x_2)\in\mathbb{R}^2$ and $k>0$.

(IP1) The inverse problem is the following.

Given the data, find $v(x)$.

Uniqueness of the solution to this problem is not proved. IP1 is not overdetermined: the data is a function of three variables, and $v(x)$ is also.

A similar inverse problem can be formulated: Let

$$\nabla^2 u+k^2u-q(x)u=0, \quad \text{in } \mathbb{R}^3, \tag{1.2}$$

$$u=e^{ik\alpha\cdot x}+A(\alpha',\alpha,k)\frac{e^{ikr}}{r}+o\left(\frac{1}{r}\right), \quad r:=|x|\to\infty, \quad \alpha'=\frac{x}{r}, \tag{1.3}$$

where $\alpha\in S^2$ is a given unit vector, $q(x)$ is a real-valued piecewise-smooth function, supp $q(x)\subset B_a:=\{x:|x|\leqslant a\}$, and S^2 is the unit sphere.

(IP2) *Given* $A(\alpha',\alpha_0,k)$ *for all* $\alpha'\in S^2$, *all* $k>0$ *and a fixed* $\alpha=\alpha_0\in S^2$, *find* $q(x)$.

The uniqueness of the solution to (IP2) is not proved.

The third problem is the following.

Let

$$\nabla^2 u+k^2u-q(x)u=-\delta(x), \quad \text{in } \mathbb{R}^3; \tag{1.4}$$

u satisfies the radiation condition, and $q(x)$ is the same as in (IP2).

The data are the values $u(x,k)|_{|x|=a}$.

(IP3) *Given the data* $u(x,k)|_{|x|=a}$ *for all* $k>0$ *and all* x *on the sphere* $S_a:=\{x:|x|=a\}$, *find* $q(x)$.

Uniqueness of the solution to (IP3) is not proved.

An overview of inverse problems and references one can find in Refs. 1–3.

[a)]Electronic mail: ramm@math.ksu.edu

Our purpose in this paper is to study the one-dimensional analog of (IP3) and to prove for this analog a uniqueness theorem. The one-dimensional analog of (IP3) corresponds to a plasma equation in a layer.

Let

$$lu-k^2u:=-u''+q(x)u-k^2u=\delta(x), \quad x\in\mathbb{R}^1, \tag{1.5}$$

$$\frac{\partial u}{\partial|x|}-iku\to 0, \quad |x|\to\infty. \tag{1.6}$$

Assume that $q(x)$ is a real-valued function,

$$q(x)=0, \quad \text{for } |x|>1, \quad q\in L^\infty[-1,1]. \tag{1.7}$$

Suppose that the data,

$$\{u(-1,k),u(1,k)\}, \quad \forall k>0, \tag{1.8}$$

are given.

The inverse problem analogous to (IP3) is the following.

(IP) Given the data (1.8), find $q(x)$.

This problem, as well as (IP1)–(IP3), is of practical interest. One can think about finding the properties of an inhomogeneous slab (the governing equation is a plasma equation) from the boundary measurements of the field, generated by a point source inside the slab.

In the literature there are many results concerning various inverse problems for the homogeneous version of Eq. (1.5), but it seems that no results concerning (IP) are known.

Assume that the self-adjoint operator $l=-d^2/dx^2+q(x)$ in $L^2(\mathbb{R})$ has no negative eigenvalues [this is the case when $q(x)\geqslant 0$, for example]. The operator l is the closure in $L^2(\mathbb{R})$ of the symmetric operator l_0 defined on $C_0^\infty(\mathbb{R}^1)$ by the formula $l_0u=-u''+q(x)u$. Our result is the following.

Theorem 1: *Under the above assumptions IP has, at most, one solution.*

II. PROOF OF THEOREM 1

The solution to (1.5)–(1.6) is

$$u=\begin{cases}\dfrac{g(k)}{[f,g]}f(x,k), & x>0,\\[2mm] \dfrac{f(k)}{[f,g]}g(x,k), & x<0.\end{cases} \tag{2.1}$$

Here $f(x,k)$ and $g(x,k)$ solve homogeneous version of Eq. (1.5) and have the following asymptotics:

$$f(x,k)\sim e^{ikx}, \quad x\to+\infty, \quad g(x,k)\sim e^{-ikx}, \quad x\to-\infty, \tag{2.2}$$

$$f(k):=f(0,k), \quad g(k):=g(0,k), \tag{2.3}$$

$$[f,g]:=fg'-f'g=-2ika(k), \tag{2.4}$$

where the prime denotes differentiation with respect to the x variable, and $a(k)$ is defined by the equation

$$f(x,k)=b(k)g(x,k)+a(k)g(x,-k). \tag{2.5}$$

It is known (see, for example, Ref. 4) that

$$g(x,k)=-b(-k)f(x,k)+a(k)f(x,-k), \tag{2.6}$$

$$a(-k)=\overline{a(k)}, \quad b(-k)=\overline{b(k)}, \quad |a(k)|^2=1+|b(k)|^2, \quad k\in\mathbb{R}, \tag{2.7}$$

$$a(k)=1+O\left(\frac{1}{k}\right), \quad k\to\infty, \quad k\in\mathbb{C}_+; \quad b(k)=O\left(\frac{1}{k}\right), \quad |k|\to\infty, \quad k\in\mathbb{R}, \tag{2.7$'$}$$

$$[f(x,k),g(x,-k)]=2ikb(k), \quad [f(x,k),g(x,k)]=-2ika(k), \tag{2.8}$$

$a(k)$ is analytic in $\mathbb{C}_+$, $b(k)$, in general, does not admit analytic continuation from $\mathbb{R}$, but if $q(x)$ is compactly supported, then $a(k)$ and $b(k)$ are analytic functions of $k\in\mathbb{C}\backslash 0$.

The functions

$$A_1(k):=\frac{g(k)f(1,k)}{-2ika(k)}, \quad A_2(k):=\frac{f(k)g(-1,k)}{-2ika(k)}, \tag{2.9}$$

are the data; they are known for all $k>0$. Therefore one can assume the functions

$$h_1(k):=\frac{g(k)}{a(k)}, \quad h_2(k):=\frac{f(k)}{a(k)}, \tag{2.10}$$

to be known for all $k>0$, because

$$f(1,k)=e^{ik}, \quad g(-1,k)=e^{ik}, \tag{2.11}$$

as follows from the assumption (1.7) and from (2.2).

From (2.10), (2.6), and (2.5), it follows that

$$a(k)h_1(k)=-b(-k)f(k)+a(k)f(-k)=-b(-k)h_2(k)a(k)+h_2(-k)a(-k)a(k), \tag{2.12}$$

$$a(k)h_2(k)=b(k)a(k)h_1(k)+a(k)h_1(-k)a(-k). \tag{2.13}$$

From (2.12) and (2.13), it follows that

$$-b(-k)h_2(k)+h_2(-k)a(-k)=h_1(k), \tag{2.14}$$

$$b(k)h_1(k)+a(-k)h_1(-k)=h_2(k). \tag{2.15}$$

Eliminating $b(-k)$ from (2.14) and (2.15), one gets

$$a(k)h_1(k)h_2(k)+a(-k)h_1(-k)h_2(-k)=h_1(k)h_1(-k)+h_2(-k)h_2(k), \tag{2.16}$$

or

$$a(k)=m(k)a(-k)+n(k), \quad k\in\mathbb{R}, \tag{2.17}$$

where

$$m(k):=-\frac{h_1(-k)h_2(-k)}{h_1(k)h_2(k)}, \quad n(k):=\frac{h_1(-k)}{h_2(k)}+\frac{h_2(-k)}{h_1(k)}. \tag{2.18}$$

Problem (2.17) is a Riemann problem (see Ref. 5 for the theory of this problem) for the pair $\{a(k),a(-k)\}$, the function $a(k)$ is analytic in $\mathbb{C}_+ := \{k : k\in\mathbb{C}, \operatorname{Im} k>0\}$ and $a(-k)$ is analytic in $\mathbb{C}_-$. The functions $a(k)$ and $a(-k)$ tend to one as k tends to infinity in $\mathbb{C}_+$ and, respectively, in $\mathbb{C}_-$; see Eq. (2.7′).

The function $a(k)$ has finitely many simple zeros at the points ik_j, $1\le j\le J$, $k_j>0$, where $-k_j^2$ are the negative eigenvalues of the operator l defined by the differential expression $lu = -u'' + q(x)u$ in $L^2(\mathbb{R})$.

The zeros ik_j are the only zeros of $a(k)$ in the upper half-plane k.

Define

$$\operatorname{ind} a(k) := \frac{1}{2\pi i}\int_{-\infty}^{\infty} d\ln a(k). \tag{2.19}$$

One has

$$\operatorname{ind} a = J, \tag{2.20}$$

where J is the number of negative eigenvalues of the operator l, and, using (2.10), (2.20) and (2.18), one gets

$$\operatorname{ind} m(k) = -2[\operatorname{ind} h_1(k) + \operatorname{ind} h_2(k)] = -2[\operatorname{ind} g(k) + \operatorname{ind} f(k) - 2J]. \tag{2.21}$$

Since l has no negative eigenvalues, it follows that $J=0$.

In this case $\operatorname{ind} f(k) = \operatorname{ind} g(k) = 0$ (see Lemma 1 below), so $\operatorname{ind} m(k)=0$, and $a(k)$ is uniquely recovered from the data as the solution of (2.17), which tends to one at infinity; see Eq. (2.7′). If $a(k)$ is found, then $b(k)$ is uniquely determined by Eq. (2.15), and so the reflection coefficient $r(k) := b(k)/a(k)$ is found. The reflection coefficient determines a compactly supported $q(x)$ uniquely (see Ref. 2).

To make this paper self-contained, let us outline a proof of the last claim using an argument different from the one given in Ref. 2.

If $q(x)$ is compactly supported, then the reflection coefficient $r(k) := b(k)/a(k)$ is meromorphic. Therefore, its values for all $k>0$ determine uniquely $r(k)$ in the whole complex k plane as a meromorphic function. The poles of this function in the upper half-plane are the numbers ik_j, $j=1,2,\ldots,J$. They determine uniquely the numbers k_j, $1\le j\le J$, which are a part of the standard scattering data $\{r(k),k_j,s_j,1\le j\le J\}$, where s_j are the norming constants.

Note that if $a(ik_j)=0$ then $b(ik_j)\neq 0$: otherwise Eq. (2.5) would imply $f(x,ik_j)\equiv 0$, in contradiction to the first relation (2.2).

If $r(k)$ is meromorphic, then the norming constants can be calculated by the formula $s_j = -i[b(ik_j)/\dot{a}(ik_j)] = -i\operatorname{Res}_{k=ik_j} r(k)$, where the dot denotes differentiation with respect to k, and Res denotes the residue. So, for compactly supported potential, the values of $r(k)$ for all $k>0$ determine uniquely the standard scattering data, that is, the reflection coefficient, the bound states $-k_j^2$, and the norming constants s_j, $1\le j\le J$. These data determine the potential uniquely.

Theorem 1 is proved. □

Lemma 1: If $J=0$ then $\operatorname{ind} f = \operatorname{ind} g = 0$.

Proof: We prove $\operatorname{ind} f=0$. The proof of the equation $\operatorname{ind} g=0$ is similar. Since $\operatorname{ind} f(k)$ equals the number of zeros of $f(k)$ in $\mathbb{C}_+$, we have to prove that $f(k)$ does not vanish in $\mathbb{C}_+$. If $f(z)=0$, $z\in\mathbb{C}_+$, then $z=ik$, $k>0$, and $-k^2$ is an eigenvalue of the operator l in $L^2(0,\infty)$, with the boundary condition $u(0)=0$.

From the variational principle one can find the negative eigenvalues of the operator l in $L^2(\mathbb{R}_+)$ with the Dirichlet condition at $x=0$ as consecutive minima of the quadratic functional. The minimal eigenvalue is

$$-k^2 = \inf\int_0^{\infty}[u'^2 + q(x)u^2]dx := \kappa_0, \quad u\in \mathring{H}^1(\mathbb{R}_+), \quad \|u\|_{L^2(\mathbb{R}_+)} = 1, \tag{2.22}$$

where $\mathring{H}^1(\mathbb{R}_+)$ is the Sobolev space of $H^1(\mathbb{R}_+)$ functions, satisfying the condition $u(0)=0$.

On the other hand, if $J=0$, then

$$0\leqslant \inf\int_{-\infty}^{\infty}[u'^2+q(x)u^2]dx := \kappa_1, \quad u\in H^1(\mathbb{R}), \quad \|u\|_{L^2(\mathbb{R})}=1. \tag{2.23}$$

Since any element u of $\mathring{H}^1(\mathbb{R}_+)$ can be considered as an element of $H^1(\mathbb{R})$ if one extends u to the whole axis by setting $u=0$ for $x<0$, it follows from the variational definitions (2.22) and (2.23) that $\kappa_1\leqslant\kappa_0$. Therefore, if $J=0$, then $\kappa_1\geqslant 0$, and therefore $\kappa_0\geqslant 0$. This means that operator l on $L^2(\mathbb{R}_+)$ with the Dirichlet condition at $x=0$ has no negative eigenvalues. This means that $f(k)$ does not have zeros in $\mathbb{C}_+$, if $J=0$. Thus $J=0$ implies $\operatorname{ind} f(k)=0$.

Lemma 1 is proved. □

Remark 2: The above argument shows that, in general,

$$\operatorname{ind} f\leqslant J \quad \text{and} \quad \operatorname{ind} g\leqslant J, \tag{2.24}$$

so that (2.21) implies

$$\operatorname{ind} m(k)\geqslant 0. \tag{2.25}$$

Therefore the Riemann problem (2.17) is always solvable.

[1] A. G. Ramm, *Multidimensional Inverse Scattering Problems* (Longman/Wiley, New York, 1992), pp. 1–385; Russian translation of the expanded monograph (Mir, Moscow, 1994), pp. 1–496.
[2] A. G. Ramm, ‘‘Property C for ODE and applications to inverse scattering,’’ Z. Anal. Angew. **18**, 331–348 (1999).
[3] A. G. Ramm, *Property C for ODE and Applications to Inverse Problems*, to appear in Proceedings of the International Conference on Operator Theory, Fields Institute Communications.
[4] V. A. Marchenko, *Sturm–Liouville Operators and Applications* (Birkhäuser, Boston, 1986).
[5] F. Gakhov, *Boundary Value Problems* (Pergamon, New York, 1966).

Spin-mass content of the Bhabha particle and the group chain $SU(4)\supset Sp(4)\supset SU(2)\times U(1)$

Yu. F. Smirnov[a)]
Instituto de Ciencias Nucleares, UNAM, Apartado Postal 70-543, 04510 México D.F., México

Anju Sharma[b)]
Instituto de Física, UNAM, Apartado Postal 20-364, 01000 México D.F., México

(Received 24 August 1998; accepted for publication 11 March 1999)

The relativistic particles described by the Bhabha equation (Bhabha particles) can possess several values of masses ($M_\tau = n/2\tau$) and spins $s \leq n/2$. In order to find the spin-mass content of such particles corresponding to the general irrep $\langle n^1 n^2 \rangle$ of the group $Sp(4)\sim O(5)$ which is the symmetry group of the Bhabha equation, the noncanonical chain of groups $Sp(4)\supset SU_s(2)\times U_\tau(1)$ is considered. In this paper, we construct the basis of the irrep $\langle n^1 n^2 \rangle$ in a boson realization form using the method of elementary permissible diagrams. This is achieved by the introduction of special ''symplectic'' bosons. The maximum and minimum values of spin s are established for each value of τ. The multiplicity of degenerate pairs (s,t) can be easily calculated with the help of the generating function obtained in this paper.
 [S0022-2488(99)01908-8]

I. INTRODUCTION

The Lorentz-invariant Bhabha equation[1] describes a free relativistic particle of arbitrary spin in terms of a single equation linear in 4-momentum, with the orthogonal group $O(5)$ as its symmetry group. The equation and its properties were extensively studied in the 1970's by Krajcik and Nieto in a series of papers.[2] Interest in the Bhabha equation has been increased recently[3] by the fact that all relativistic equations for particles with spin $s>1/2$ (Duffin–Kemmer, Fierz–Pauli, Bargmann–Wigner equations), are particular cases of the Bhabha equation (sometimes with additional conditions). Recently, a simpler way was proposed[4] by a research group including the present authors, to ''derive'' the Bhabha equation in a Lorentz-covariant way, and to study its properties using the canonical chain of orthogonal groups $O(5)\supset O(4)\supset O(3)\supset O(2)$. The equation is written in terms of the generators $\Lambda_{\mu 5}$ ($\mu=1,2,3,4$) of the group $O(5)$ as[4]

$$(2ic\Lambda_{i5}p_i + 2c\Lambda_{45}p_0 + nmc^2)\psi = 0, \tag{1}$$

with the index i taking values $i=1,2,3$. The quantity n is related to the spin such that $n/2$ gives the maximum possible spin value the particle can possess though the general spin ranges from $n/2,\ n/2-1\cdots$ to 0 or $\frac{1}{2}$ depending on whether n is even or odd.

It is well known[1,2,4] that Eq. (1) is not a single mass equation but describes several states of the particle (Bhabha particle) with different masses associated with each spin s.[1,2,4] The masses were found to be inversely proportional to the eigenvalues τ of the operator Λ_{45} such that[4]

$$M_\tau = \frac{n}{2\tau}. \tag{2}$$

[a)]Electronic mail: smirnov@nuclecu.unam.mx
[b)]Electronic mail: anju@fenix.ifisicacu.unam.mx

The general algorithm to find out the values of spin s compatible with given mass M_τ, was suggested in Refs. 1, 2, 4. But to the best of our knowledge, a closed group theoretical solution of this problem does not exist in the literature.

The difficulty in finding such a general solution arises due to the fact that the generator Λ_{45} responsible for several mass states of the Bhabha particle, is nondiagonal in the canonical group chain $O(5)\supset O(4)\supset O(3)\supset O(2)$ used in Refs. 2, 4. In order to find the spin-mass content of the Bhabha particles, it is necessary to use a basis adapted to a chain of subgroups for which the spin operator S^2 and the generator Λ_{45} would simultaneously be diagonal. Therefore in this paper, we propose to use the noncanonical group chain,

$$Sp(4)\supset SU_s(2)\times U_\tau(1), \tag{3}$$

to find out the solution of the spin-mass content problem mentioned above.

In group reduction (3), $Sp(4)$ is a group isomorphic to $O(5)$, $SU_s(2)$ is the group of the usual spin and the group $U_\tau(1)$ is generated by the operator Λ_{45} which is identical to the operator T_0 corresponding to the sign-spin projection considered in Ref. 5 in the framework of the supermultiplet scheme $SU(4)\supset SU_s(2)\times SU_t(2)$.

To the best of our knowledge the group reduction (3) has so far not been used in relativistic problems, although it was used widely in the nuclear shell model to describe the effects of the isotopic invariant pairing interaction between nucleons of both kinds in the same nuclear shell.[6]

In order to construct the basis adapted to the chain (3), we use in contrast to the previous works,[6] the classical method of elementary permissible diagrams (epd's)[7–9] applied to the boson representation of the states corresponding to the group chain,

$$U(4)\supset Sp(4)\supset SU_s(2)\times U_\tau(1). \tag{4}$$

The paper is organized as follows: In Sec. II, we give a realization of the generators of groups belonging to chain (4) in terms of the boson creation and annihilation operators. In the process, the generators of the group complementary to $Sp(4)$ are also constructed. In Sec. III, we calculate the epd's using a ''trial and error'' method by considering some of the lowest irreps of $U(4)$.[5] In order to obtain the basis vectors with a definite $Sp(4)$ symmetry, the so called ''symplectic bosons'' are used which correspond to a modified form of the bosons introduced earlier in Refs. 10, 11 for systems with the orthogonal and symplectic symmetry. Section IV is concerned with the structure of the basis corresponding to the chain (4). The generating function for this basis is constructed, and the completeness of the latter is verified by the calculation of its dimension. In Sec. V we deal with the spin-mass content of the Bhabha particle, i.e., the maximum and minimum values of spin s for a given mass M_τ are established. The multiplicity of each s,τ combination can be easily found from the generating function. Some final remarks are given in the conclusion.

II. GROUP CHAIN $SU(4)\supset Sp(4)\supset SU_s(2)\times U_\tau(1)$

A. Boson realization

In order to construct the basis for the $Sp(4)$ irrep $\langle\omega^1,\omega^2\rangle$ $(\omega^1\geqslant\omega^2\geqslant 0)$ in an explicit form, it is convenient to embed the group $Sp(4)$ into the group $U(4)$ and to use the standard boson realization for the latter. To obtain the general $Sp(4)$ irrep $\langle\omega^1,\omega^2\rangle$, it is sufficient to consider its embedding into the $U(4)$ irrep $\{h\}=\{h_1h_2h_3h_4\}$ with the Young scheme containing only two rows of length ω^1 and ω^2, i.e.,

$$\{h\}\equiv\{\omega^1\omega^2\}=\{\omega^1\omega^2 00\}. \tag{5}$$

For the realization of such a $U(4)$ irrep it is enough to use two kinds of creation and annihilation boson operators $\eta^1_{\sigma\tau},\eta^2_{\sigma\tau}$ and $\xi^1_{\sigma\tau},\xi^2_{\sigma\tau}$, where the projections σ and τ of the usual spin s and the sign-spin t take the values $\pm 1/2$.

The boson operators satisfy the standard commutation relations

$$[\eta^{\kappa}_{\sigma\tau},\eta^{\kappa'}_{\sigma'\tau'}]=[\xi^{\kappa}_{\sigma\tau},\xi^{\kappa'}_{\sigma'\tau'}]=0,\quad [\xi^{\kappa}_{\sigma\tau},\eta^{\kappa'}_{\sigma'\tau'}]=\delta_{\kappa\kappa'}\delta_{\sigma\sigma'}\delta_{\tau\tau'}\,. \tag{6}$$

The generators of the $U(4)$ group are expressed in terms of boson operators as follows:

$$A_{\sigma\tau,\sigma'\tau'}=\sum_{k=1,2}\eta^k_{\sigma\tau}\xi^k_{\sigma'\tau'}\,. \tag{7}$$

It was shown in Ref. 12 that they can also be chosen in a tensor form,

$$S_\mu=\sum_{k=1}^{2}[\eta^\kappa\times\xi^\kappa]^{10}_{\mu 0},\quad T_\nu=\sum_{k=1}^{2}[\eta^\kappa\times\xi^\kappa]^{01}_{0\nu},$$

$$R_{\mu\nu}=\sum_{k=1}^{2}[\eta^\kappa\times\xi^\kappa]^{11}_{\mu\nu},\quad \mu,\nu=0,\pm 1, \tag{8}$$

$$N=2\sum_{k=1}^{2}[\eta^\kappa\times\xi^\kappa]^{00}_{00}\,.$$

Here S_μ and T_ν are the operators of the ordinary and sign spin, respectively. The creation operators $\eta_{\sigma\tau}$ are the components of a double tensor operator of rank 1/2 with respect to both $SU_s(2)$ and $SU_t(2)$ groups,

$$\eta_{\sigma\tau}=V^{(1/2)(1/2)}_{\sigma\tau}\,. \tag{9}$$

The annihilation operators are connected with another tensor of the same rank 1/2, 1/2,

$$\xi_{\sigma\tau}=(-1)^{1-\sigma-\tau}W^{(1/2)(1/2)}_{-\sigma-\tau}=\eta^\dagger_{\sigma\tau}\,. \tag{10}$$

The tensor product of two operators into a common double tensor of the ranks r_1 and r_2 with respect to the groups $SU_s(2)$ and $SU_t(2)$ is denoted as follows:

$$[\eta\times\xi]^{r_1r_2}_{\rho_1\rho_2}=\sum_{\sigma_1\sigma_2,\tau_1\tau_2}\left\langle\frac12\sigma_1\frac12\sigma_2\middle|r_1\rho_1\right\rangle\left\langle\frac12\tau_1\frac12\tau_2\middle|r_2\rho_2\right\rangle V^{(1/2)(1/2)}_{\sigma_1\tau_1}W^{(1/2)(1/2)}_{\sigma_2\tau_2}\,, \tag{11}$$

where $\langle j_1m_1j_2m_2|jm\rangle$ are the usual Clebsch–Gordan coefficients. The commutation relations of the operators (8) are of the form

$$\begin{aligned}
[S_\mu,R_{\mu'\nu}]&=\sqrt{2}\langle 1\mu' 1\mu|1\mu+\mu'\rangle R_{\mu+\mu'\nu}\,,\\
[T_\nu,R_{\mu\nu'}]&=\sqrt{2}\langle 1\nu' 1\nu|1\nu+\nu'\rangle R_{\mu,\nu+\nu'}\,,\\
[R_{\mu\nu},R_{\mu'\nu'}]&=\delta_{\nu,-\nu'}\frac{1}{4\sqrt{2}}(-1)^{1-\nu}S_{\mu+\mu'}\langle 1\mu 1\mu'|1\mu+\mu'\rangle\\
&\quad+\delta_{\mu,-\mu'}\frac{1}{4\sqrt{2}}(-1)^{1-\mu}T_{\nu+\nu'}(1\nu 1\nu'|1\nu+\nu').
\end{aligned} \tag{12}$$

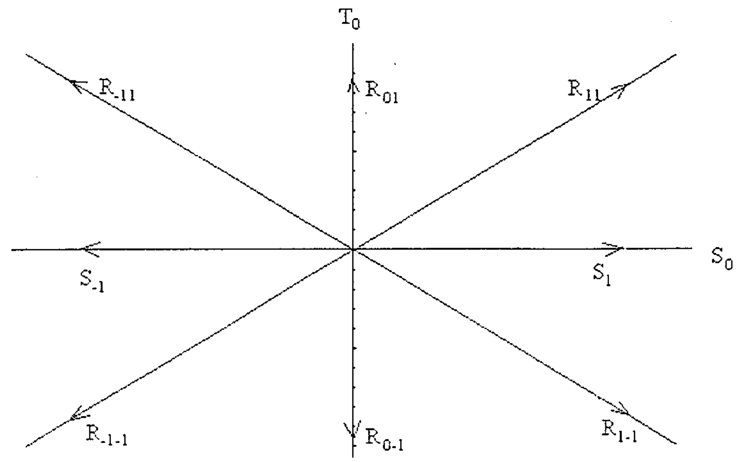

FIG. 1. Root diagram for the *Sp*(4) group.

The operator N in (8) commutes with the rest of the generators. If we remove from the set (8) the operators $N, T_1, T_{-1}, R_{10}, R_{00}$, and R_{-10}, the remaining 10 operators form a closed Lie algebra. It follows from (12) that

$$[S_0,[\eta\times\xi]^{r_1r_2}_{\mu\nu}]=\mu[\eta\times\xi]^{r_1r_2}_{\mu\nu},$$
$$[T_0,[\eta\times\xi]^{r_1r_2}_{\mu\nu}]=\nu[\eta\times\xi]^{r_1r_2}_{\mu\nu}. \tag{13}$$

Thus, if we take S_0,T_0 as the weight (Cartan) generators, the generator $[\eta\times\xi]^{r_1r_2}_{\mu\nu}$ corresponds to the root $(\mu\nu)$. As a result, we obtain the root diagram for these 10 generators given in Fig. 1. The two Cartan generators correspond to the center of this figure. It is known that such a diagram corresponds to the Lie algebra for the $Sp(4)\simeq SO(5)$ group.

Using the expression (12) we can verify that all commutation relations necessary to form the $Sp(4)$ group, are satisfied by the operators S_μ, T_0, $R_{\mu,\pm1}$, $\mu=0,\pm1$. Therefore these operators can be considered as the generators of the subgroup $Sp(4)$ of the group $U(4)$.

It should be noted that the antisymmetric bilinear combination,

$$I=[\eta^1\times\eta^2]^{01}_{00}=\tfrac{1}{2}(\eta^1_{++}\eta^2_{--}-\eta^1_{-+}\eta^2_{+-}+\eta^1_{+-}\eta^2_{-+}-\eta^1_{--}\eta^2_{++}), \tag{14}$$

commutes with all the $Sp(4)$ generators, i.e., it is an $Sp(4)$ invariant.

B. Group complementary to *Sp*(4)

In order to construct the polynomial basis in terms of epd's corresponding to the chain $Sp(4)\supset SU_s(2)\times U_\tau(1)$, it is necessary first to find out the form of the group G complementary to $Sp(4)$.[9,12] It should be noted in this connection that the group $U(4)$ is complementary to the group $U(2)$ defined by the generators

$$F^+=A^{12},\quad F^-=A^{21},\quad F^0=\tfrac{1}{2}(A^{11}-A^{22})=\tfrac{1}{2}(N^1-N^2),\quad N=A^{11}+A^{22}, \tag{15}$$

which satisfy the standard commutation relations for the $SU(2)$ group,

$$[F^0,F^\pm]=\pm F^\pm,\quad [F^+,F^-]=2F^0, \tag{16}$$

and commute with the number operator N,

$$N=N^1+N^2, \quad N^i=A^{ii}. \tag{17}$$

Here and below we use the notation

$$A^{\kappa\kappa'}=\sum_{\sigma\tau} \eta^{\kappa}_{\sigma\tau}\xi^{\kappa'}_{\sigma\tau}. \tag{18}$$

The irrep of the $U(2)$ group is characterized by the same partition $[n^1n^2]$ as the irreps of the $U(4)$ group[9,12] where (n^1n^2) is the highest weight of this irrep. Alternatively, this irrep can also be enumerated by the quantum numbers $N=n^1+n^2$ and

$$f=\tfrac{1}{2}(n^1-n^2), \tag{19}$$

giving the eigenvalue $f(f+1)$ for the $SU_F(2)$ Casimir operator,

$$F^2=F^-F^+ +F^0(F^0+1). \tag{20}$$

The basis vectors of the irrep are labeled by the quantum number ϕ which is the eigenvalue of the operator F^0. Since $Sp(4)$ is a subgroup of $U(4)$, it commutes with the generators of the $U(2)$ group mentioned above.

Thus the group G complementary to $Sp(4)$ must contain this group $U(2)$ [or $SU_F(2)$] as its subgroup. The additional operators commuting with the $Sp(4)$ generators are of the form

$$\begin{aligned} K_+ &= I=[\eta^1\times\eta^2]^{01}_{00}, \\ K_- &= (I)^\dagger=-[\xi^1\times\xi^2]^{01}_{00}, \\ K_0 &= \tfrac{1}{2}(N+4). \end{aligned} \tag{21}$$

They satisfy the commutation relations

$$[K_0,K_\pm]=K_\pm, \quad [K_-,K_+]=2K_0, \tag{22}$$

which are typical for the generators of the noncompact group $SU_K(1,1)$. The generators K commute with the generators F given in (15). Therefore we can consider the direct product of the groups $SU_F(2)$ and $SU_K(1,1)$ as a maximal group commuting with the group $Sp(4)$. It is known that this direct product is identical to the noncompact group $O^*(4)$. Thus we can say that the group G complementary to $Sp(4)$ is of the form

$$G=O^*(4)=SU_K(1,1)\times SU_F(2), \tag{23}$$

with κ and f giving the irreps of subgroups $SU_K(1,1)$ and $SU_F(2)$, respectively. Now it is necessary to prove that the characteristics of the irreps of G are connected unambiguously to the signature $\langle n^1n^2\rangle$ of the corresponding $Sp(4)$ irrep. It should be noted that the reduction of the $U(4)$ irrep $\{n^1n^2\}$ into the $Sp(4)$ irrep is of the form

$$\{n^1n^2\}=\langle n^1n^2\rangle+\langle n^1-1,n^2-1\rangle+\cdots+\langle n^1-n^2,0\rangle. \tag{24}$$

This means that the ket,

$$|\{n^1+r,n^2+r\}\langle n^1n^2\rangle\cdots\rangle\sim K_+^r|\{n^1n^2\}\langle n^1n^2\rangle\cdots\rangle, \tag{25}$$

with fixed values of n^1, n^2 and $r=0,1,2,\ldots,\infty$ (the remaining quantum numbers are substituted by dots), gives the basis of the infinite dimensional unitary irrep of the group $SU_K(1,1)$ belonging to the positive discrete series. By action of the $SU_K(1,1)$ Casimir operator,

$$K^2 = K_+ K_- - K_0(K_0-1), \tag{26}$$

on ket (25), and by taking into account that the lowest weight vector $|\{n^1n^2\}\langle n^1n^2\rangle\cdots\rangle$ is annihilated by the lowering generator K_-,

$$K_-|\{n^1n^2\}\langle n^1n^2\rangle\cdots\rangle = 0, \tag{27}$$

we obtain

$$K^2|\{n^1+r,n^2+r\}\langle n^1n^2\rangle\cdots\rangle = -k(k-1)|\{n^1+r,n^2+r\}\langle n^1n^2\rangle\cdots\rangle, \tag{28}$$

where

$$k = \tfrac{1}{2}(n^1+n^2+4). \tag{29}$$

Thus the ket (25) corresponding to the chain $U(4)\supset Sp(4)\supset SU_s(2)\times U_\tau(1)$, which is labeled by the symbols $\{n^1+r,n^2+r\}$ and $\langle n^1n^2\rangle$ for the irreps of these groups, can be characterized simultaneously by the quantum numbers

$$k = \tfrac{1}{2}(n^1+n^2+4), \quad f = \tfrac{1}{2}(n^1-n^2), \tag{30}$$

giving the irreps of the $SU_K(1,1)$ and $SU_F(2)$ groups (The value of r determines the ''projection'' K_0 of the non-compact moment K.)

According to (30) the symbols k and f of the $O^*(4)$ irreps can be expressed in a unique manner through the $Sp(4)$ irrep $\langle n^1n^2\rangle$, and thus the groups $Sp(4)$ and $O^*(4)$ are complementary. Our result agrees with that of Quesne[13] where she dealt with the general symplectic group $Sp(2n)$ and found $O^*(2n)$ as its complementary group.

Returning to our problem of the construction of the basis for the $Sp(4)$ irrep $\langle n^1n^2\rangle$, we want to construct at first the boson basis for the $U(4)$ irrep $\{n^1n^2\}$, and then to select in this basis the part corresponding to the first term $\langle n^1n^2\rangle$ in the expansion (24), i.e., to construct the vectors (25) with $r=0$ satisfying the relation (27).

C. The basis corresponding to the chain $SU(4)\supset Sp(4)\supset SU_s(2)\times U_\tau(1)$

In view of the previous section, the basis vectors of the $Sp(4)$ irrep $\langle n^1n^2\rangle$ corresponding to the chain $U(4)\supset Sp(4)\supset SU_s(2)\times U_\tau(1)$ can be written in the form of the following ket:

$$|\{n^1n^2\}\langle n^1n^2\rangle \Gamma s\sigma\tau; f\phi\rangle, \tag{31}$$

where the quantum numbers n^1 and n^2 label the irreps of the $U(4)$ and $Sp(4)$ group as well as the $O^*(4)$ group complementary to $Sp(4)$. The multiple irreps (s,τ) of the subgroups $SU_s(2)$ and $U_\tau(1)$ belonging to the given $Sp(4)$ irrep $\langle n^1n^2\rangle$ are distinguished by the symbol Γ. The quantum number f given in (30) represents the irrep of the $SU_F(2)$ group, ϕ is the weight of the vector (31) with respect to this group $(-f\leqslant\phi\leqslant f)$.

We shall construct the vectors (31) of the highest weight with respect to both $SU_S(2)$ and $SU_F(2)$ groups (i.e., $\sigma=s, \phi=f$). The vectors with smaller values of σ and ϕ can be obtained by applying the lowering operators S_{-1} and F_- to the highest weight vector $P^{s\tau}_{n^1n^2}(\eta^1,\eta^2)|0\rangle$.

The polynomial $P^{s\tau}_{n^1n^2}(\eta^1,\eta^2)$ which for brevity we denote by P, satisfies the following conditions:

$$N^1P|0\rangle = n^1P|0\rangle, \quad N^2P|0\rangle = n^2P|0\rangle,$$

$$S_+P|0\rangle = 0, \quad S_0P|0\rangle = sP|0\rangle,$$

$$F_+P|0\rangle = 0, \quad F_0P|0\rangle = fP|0\rangle,$$

$$K_0 P|0\rangle = \tfrac{1}{2}(n^1+n^2+4)P|0\rangle, \tag{32}$$

$$K_- P|0\rangle = 0. \tag{33}$$

By substituting $\xi \to \partial/\partial\eta$ in the expression for each generator, the set of above equations (32), (33) can be considered as differential equations on P-polynomials and then solved. In a way similar to Refs. 7–9, we shall seek the solution of the polynomial P as a product of the powers of the so called elementary permissible diagrams (epd's) W_i such that

$$P = \prod_i W_i^{k_i}(\eta^1, \eta^2), \tag{34}$$

where $W_i^{k_i}(\eta^1 \eta^2)$ are the polynomials of the lowest powers n_i^1, n_i^2 satisfying Eqs. (32) with definite values of spin s_i and the projection of the sign-spin τ_i, which cannot be reduced to a product of the powers of simpler solutions. Since all generators in (32) are first-order differential operators, it is clear that the monomial (34) satisfies Eqs. (32) if they are satisfied by the epd's $W_i(\eta^1, \eta^2)$.

Since $s = \Sigma_i k_i s_i$, $\tau = \Sigma_i k_i \tau_i$, it is obvious that the independent vectors (31) with the same values of s and τ are characterized by different sets of the epd powers $k_1, k_2 \dots$. A full set of epd's, their powers, spins and τ-values will be established below. In the next section, we consider only the solution of the equations (32). The problem of the construction of polynomials (monomials) P satisfying the condition (33) is postponed until Sec. IV.

III. ELEMENTARY PERMISSIBLE DIAGRAMS

In order to find out the epd's, we use the ''trial and error'' approach and start with few of the lowest irreps {1}, {2}, {11} of $U(4)$.

The irrep {1} is realized in a space of four boson operators $\eta^1_{\sigma\tau}(\sigma, \tau = \pm\frac{1}{2})$. As mentioned above, we are interested in the epd's with the maximum spin projection $\sigma_i = s_i$ and the maximum value of $\phi_i = f_i$. Therefore we select as the first two epd's the operators

$$a = \eta^1_{++} \quad \text{and} \quad b = \eta^1_{+-}, \tag{35}$$

corresponding to $s = \sigma = +1/2$, $f = \phi = +1/2$, $\tau = +1/2$ and $-1/2$, respectively. In (35) and below the values of σ and τ are substituted by the signs $+$ or $-$ of these projections. The operators $\eta^2_{\sigma\tau}$ can not be used as epd's because they correspond to $\phi = -1/2$.

For $n=2$, there exist two $U(4)$ irreps with partitions {2} and {11}. The symmetric irrep {2} is characterized by the values of the ordinary and sign spins $s=t=0$ and $s=t=1$.[5] Thus the list of the states with various τ and the maximal spin projections $\sigma = s$ is of the form

τ	$s=\sigma$	Expression in terms of epd's
1	1	a^2
0	1	ab
-1	1	b^2
0	0	c

. (36)

It is clear from this list that the first three vectors can be expressed in terms of the epd's a and b introduced in (35). However, the last vector with $s=t=0$ can not be reduced to the epd's (35). It should be considered as a new independent epd of the form

$$c=\sum_{\sigma\tau}\left\langle \frac{1}{2}\sigma\frac{1}{2}-\sigma\middle|00\right\rangle\left\langle\frac{1}{2}\tau\frac{1}{2}-\tau\middle|00\right\rangle \eta^1_{\sigma\tau}\eta^1_{-\sigma-\tau}=\eta^1_{++}\eta^1_{--}-\eta^1_{+-}\eta^1_{-+}. \tag{37}$$

The $U(4)$ irrep $\{2\}$ corresponds to the value $f=1$, and thus the vector (37) is characterized by $\phi=f=1$. The combinations $\eta^1\eta^2, \eta^2\eta^2$ correspond to $\phi=0$ and -1, respectively, and should thus be omitted.

For the realization of the antisymmetric $U(4)$ irrep $\{11\}$, both kinds of boson operators η^1 and η^2 are necessary. This irrep contains the states with $s=1, t=0$ and $s=0$, $t=1$.[5] They cannot be expressed in terms of epd's (35) and (37). Therefore all of them should be considered as new epd's. Their list with the corresponding $(\tau,s)-$ values is given below:

τ	$s=\sigma$	epd
1	0	d
0	0	I
-1	0	e
0	1	f

. (38)

The expressions for the epd's d, e and f in terms of the boson operators can be found in Table I.

As for the epd I, it coincides with the $Sp(4)$ invariant (14). It is clear that the inclusion of this epd in the monomial (34) does not change its $Sp(4)$ symmetry, but only increases its power. Therefore this epd can not be present in the construction of the vectors (31) we are interested in. It can be contained in the vectors $|\{h_1h_2\}\langle n^1n^2\rangle\cdots\rangle$ with $n^1+n^2<h_1+h_2$. Thus we do not consider the epd I further.

Using a procedure similar to what is done for the cases $n=1,2$, it can be easily shown that all $U(4)$ irreps corresponding to $n=3,4$ do not give rise to any new epd's but can be constructed using a, b, c, d, e and f. However, for the partition $\{31\}$ for $n=4$, we notice that there are two states occurring with the spin $s=1$ and $\tau=0$, which can be constructed using epd's in three possible ways: ea^2, db^2 and cf. This implies that out of these three possibilities only two are independent. In fact, it can be easily proved that

$$cf=Iab-ea^2-db^2. \tag{39}$$

TABLE I. Elementary permissible diagrams for the $U(4)$ group.

epd	n^1	n^2	s	τ	Relation with quantities in Ref. 14[a]
$a=\eta^1_{++}$	1	0	$\frac{1}{2}$	$\frac{1}{2}$	$AS^{1/2}T^{1/2}$
$b=\eta^1_{+-}$	1	0	$\frac{1}{2}$	$-\frac{1}{2}$	$AS^{1/2}T^{-1/2}$
$c=\eta^1_{++}\eta^1_{--}-\eta^1_{+-}\eta^1_{-+}$	2	0	0	0	A^2
$d=\eta^1_{++}\eta^2_{-+}-\eta^1_{-+}\eta^2_{++}$	1	1	0	1	BT
$e=\eta^1_{+-}\eta^2_{--}-\eta^1_{--}\eta^2_{+-}$	1	1	0	-1	BT^{-1}
$f=\eta^1_{++}\eta^2_{+-}-\eta^1_{+-}\eta^2_{++}$	1	1	1	0	BS

[a]For details see Secs. IV B and IV C of the text.

This suggests that out of 6 epd's given in Table I only five are independent. In other words, the relation (39) means that the monomials (34) can not contain epd's c and f simultaneously. Further analysis shows (see Sec. IV) that the set of epd's given in Table I, is complete. In this connection we can conclude that the basis of the $Sp(4)$ irrep $\langle n^1 n^2\rangle$ can be split into two parts. The first part consists of monomials (34) which do not contain the epd f; the second one includes monomials in which the epd c is absent.

Symplectic bosons

We now turn our attention to the problem of how the monomials of the form (34) satisfying the ''symplecticity'' condition (33) can be constructed.

Here we follow the concept of the modified or ''traceless'' bosons for the orthogonal and symplectic groups suggested and developed in Refs. 9–11. We introduce new creation operators which we call the ''symplectic bosons,'' viz.,

$$a^{1+}_{\sigma\tau}=\eta^1_{\sigma\tau}-(-1)^{1/2-\sigma}K_+\frac{1}{N+1}\xi^2_{-\sigma-\tau},$$
$$a^{2+}_{\sigma\tau}=\eta^2_{\sigma\tau}-(-1)^{1/2+\sigma}K_+\frac{1}{N+1}\xi^1_{-\sigma-\tau}, \tag{40}$$

where operators K_+ and N are given in Eqs. (21) and (17), respectively. They satisfy the commutation relations

$$[a^{k+}_{\sigma\tau},a^{k'+}_{\sigma'\tau'}]=0,\quad k,k'=1,2. \tag{41}$$

We then replace the ordinary boson operators $\eta^{1,2}$ by the symplectic ones $a^{1,2+}$ in the epd's given in Table I (and denote epd's with primes), and thus in the process we modify $P^{s\tau}_{n^1n^2}(\eta^1,\eta^2)$ to $P^{s\tau}_{n^1n^2}(a^{1+},a^{2+})$. Similarly, the raising $SU_K(1,1)$ generator K_+ gets modified into

$$K'_+=\sum_{\sigma\tau}(-1)^{1/2-\sigma}a^{1+}_{\sigma\tau}a^{2+}_{-\sigma-\tau}, \tag{42}$$

which can be simplified to

$$K'_+=\frac{1}{(N-1)(N-2)}K^2_+K_-. \tag{43}$$

Since K'_+ commutes with $P^{s\tau}_{n^1n^2}(a^{1+}a^{2+})$ both being constructed out of the commuting operators $a^{1,2+}$, we can write

$$K'_+P^{s\tau}_{n^1n^2}(a^{1+}a^{2+})|0\rangle=P^{s\tau}_{n^1n^2}(a^{1+}a^{2+})K'_+|0\rangle=0, \tag{44}$$

which is equated to zero because of (43). Therefore we obtain

$$\frac{1}{(N-1)(N-2)}K^2_+K_-P|0\rangle=0. \tag{45}$$

Since $[(N-1)(N-2)]^{-1}K^2_+=[(N-1)(N-2)]^{-1}I^2$ does not vanish identically, the states $P^{s\tau}_{n^1n^2}(a^{1+},a^{2+})|0\rangle$ satisfy the symplecticity condition (33).

Thus these states correspond to the irrep $\langle n^1n^2\rangle$ of $Sp(4)$ and $\{n^1n^200\}$ of $U(4)$, as we require a partition involving two numbers to characterize the general irreps of $Sp(4)$ and four numbers for irreps $U(4)$.[5,9] It can be easily verified that the other conditions (32) on polynomials

$P^{s\tau}_{n^1n^2}(a^{1+},a^{2+})$ remain unchanged by the introduction of symplectic bosons in the epd's of Table I. Our task now is to construct the basis vectors of the $Sp(4)$ irrep $\langle n^1n^2\rangle$ as polynomials (monomials) of these epd's.

IV. STRUCTURE OF THE BASIS, GENERATING FUNCTION AND THE DIMENSION OF THE $Sp(4)$ IRREP

A. Structure of the basis

As mentioned earlier in Sec. III, the epd's c' and f' can not occur simultaneously in a basis vector, i.e., if the power γ of c' is nonvanishing then the power φ of the epd f' is vanishing, and vice versa $\gamma=0$ if $\varphi\neq 0$. Thus the polynomial basis states can be distributed into two subsets:

$$I. \quad P_{n^1n^2}(a^{1+},a^{2+})|0\rangle = a'^{\alpha}b'^{\beta}c'^{\gamma}d'^{\delta}e'^{\epsilon}|0\rangle, \tag{46}$$

$$II. \quad P_{n^1n^2}(a^{1+},a^{2+})|0\rangle = a'^{\alpha}b'^{\beta}d'^{\delta}e'^{\epsilon}f'^{\phi}|0\rangle, \tag{47}$$

where $\alpha,\beta,\gamma,\delta,\epsilon,\varphi\geqslant 0$ are non-negative integers. Our aim now is to find out these exponents in terms of the fixed values of quantum numbers n^1, n^2, s and τ. We consider the cases (46) and (47) separately.

(i) Case I. From Table I and expression (46), we obtain the following relations:

$$\begin{gathered} n^1=\alpha+\beta+2\gamma+\delta+\epsilon, \quad n^2=\delta+\epsilon, \\ s=\tfrac{1}{2}(\alpha+\beta), \quad \tau=\tfrac{1}{2}(\alpha-\beta)+\delta-\epsilon. \end{gathered} \tag{48}$$

Solving them with respect to $\alpha,\gamma,\delta,\epsilon$ we obtain

$$\begin{gathered} \alpha=2s-\beta, \quad \gamma=\tfrac{1}{2}(n^1-n^2)-s, \\ \delta=\tfrac{1}{2}(n^2+\tau-s+\beta), \quad \epsilon=\tfrac{1}{2}(n^2-\tau+s-\beta), \end{gathered} \tag{49}$$

where the non-negative integer $\beta\geqslant 0$ is a free parameter here. Since $\gamma\geqslant 0$ is also a non-negative integer, the spin values s are restricted in this part of the basis by the condition

$$s\leqslant \tfrac{1}{2}(n^1-n^2). \tag{50}$$

(ii) Case II. Similar to (48) we have in this case the relations

$$\begin{gathered} n^1=\alpha+\beta+\delta+\epsilon+\varphi, \quad n^2=\delta+\epsilon+\varphi, \\ s=\tfrac{1}{2}(\alpha+\beta)+\varphi, \quad \tau=\tfrac{1}{2}(\alpha-\beta)+\delta-\epsilon. \end{gathered} \tag{51}$$

From here we obtain the following expressions for $\alpha,\delta,\epsilon,\varphi$:

$$\begin{gathered} \alpha=n^1-n^2-\beta, \quad \delta=\tfrac{1}{2}(n^2+\tau-s+\beta), \\ \epsilon=\tfrac{1}{2}(n^1-\tau-s-\beta), \quad \varphi=s-\tfrac{1}{2}(n^1-n^2), \end{gathered} \tag{52}$$

where the non-negative integer $\beta\geqslant 0$ is again a dummy variable here. Since φ is also a non-negative integer, we obtain for this part of the basis the condition

$$s\geqslant \tfrac{1}{2}(n^1-n^2). \tag{53}$$

Thus the two parts (46) and (47) of the basis of the $Sp(4)$ irrep $\langle n^2 n^2\rangle$ can be rewritten in the following form:

$$P_{n^1n^2}(a^{1+},a^{2+})|0\rangle = a'^{2s-\beta}b'^{\beta}c'^{(n^1-n^2)/2-s}d'^{(n^2+\tau-s+\beta)/2}e'^{(n^2-\tau+s-\beta)/2}|0\rangle,$$
$$\text{for}\ \ s\leqslant\frac{(n^1-n^2)}{2}, \tag{54}$$

$$= a'^{n^1-n^2-\beta}b'^{\beta}d'^{(n^2+\tau-s+\beta)/2}e'^{(n^1-\tau-s-\beta)/2}f'^{s-(n^1-n^2)/2}|0\rangle,$$
$$\text{for}\ \ s>\frac{(n^1-n^2)}{2}, \tag{55}$$

with the non-negative integer β being a dummy variable. It plays the role of an additional quantum number to distinguish independent polynomials $P_{n^1n^2}(a^{1+},a^{2+})$ corresponding to the same values of the quantum numbers s and τ. The possible values of β are determined by the requirement that all powers of the epd's in expressions (54), (55) are non-negative integers. We shall discuss this in more detail in Sec. V.

We now turn our attention to the generating function for the basis (54), (55).

B. The generating function

The generating function $F(AB;ST)$ for the group reduction $Sp(4)\supset SU_s(2)\times U_\tau(1)$ is defined in the following way:[14]

$$F(AB;ST)=\sum_{n^1n^2s\tau}\nu(n^1n^2s\tau)A^{n^1-n^2}B^{n^2}S^sT^\tau, \tag{56}$$

such that the coefficient $\nu(n^1n^2s\tau)$ in its power expansion gives the number of states with given values of s and τ in the $Sp(4)$ irrep $\langle n^1n^2\rangle$.

To construct the generating function (56) we start with the basis states (46), (47). The contribution of each epd into the generating function can be characterized by a definite combination of the quantities A,B,S,T which are given in the last column of Table I. As a result, the contributions of vectors (46) and (47) into a generating function can be represented by the expressions

$$A^{\alpha+\beta+2\gamma}B^{\delta+\epsilon}S^{(\alpha+\beta)/2}T^{(\alpha-\beta)/2+\delta-\epsilon} \tag{57}$$

and

$$A^{\alpha+\beta}B^{\delta+\epsilon+\varphi}S^{(\alpha+\beta)/2+\varphi}T^{(\alpha-\beta)/2+\delta-\epsilon}, \tag{58}$$

respectively. The independent summation of both of these expressions on α, β, γ, δ, ϵ and φ from 0 to infinity and unification of the two results, gives the following formula for the generating function:

$$F(AB;ST)=\frac{1}{(1-AS^{1/2}T^{1/2})(1-AS^{1/2}T^{-1/2})}\,\frac{1}{(1-BT)(1-BT^{-1})}\left\{\frac{1}{1-A^2}+\frac{1}{1-BS}-1\right\}. \tag{59}$$

The last term (-1) is included in the curly brackets in order to avoid the double counting of the states with $\gamma=\varphi=0$. The above formula is in agreement with the corresponding generating function (2.6) found in Ref. 14 using a different approach. This coincidence means that our basis (54), (55) is complete. However, it can also be verified by the calculation of the dimension of the $Sp(4)$ irrep $\langle n^1n^2\rangle$.

C. The dimension

It was demonstrated above that the structure of the basis (54), (55) is described by the generating function (59). We now calculate the dimension of the irrep $\langle n^1 n^2\rangle$. It is clear that this dimension can be found by the formula

$$\dim\langle n^1 n^2\rangle = \sum_{s\tau} \nu(n^1 n^2 s\tau)(2s+1). \tag{60}$$

In order to calculate it, we can do the power expansion of the function (59) in terms of the monomials (57), (58), and then select the terms with the fixed powers n^1-n^2 and n^2 of parameters A and B, respectively. As a result the number of summation indices decreases up to 3.

At last we add the factor $(2s+1)$ into the resulting sum in correspondence with (60), and substitute $A=B=S=T=1$. Such a procedure is first applied to (58) given the following restrictions: $\alpha+\beta=n^1-n^2$, $\delta+\epsilon+\varphi=n^2$. Using α, φ and $\xi=\epsilon+\varphi$ as summation indices and taking into account that $2s+1=n^1-n^2+2\varphi+1$, we obtain the following expression:

$$N_1=\sum_{\alpha=0}^{n^1-n^2}\sum_{\xi=0}^{n^2}\sum_{\varphi=0}^{\xi}(n^1-n^2+2\varphi+1)=\frac{1}{6}(n^1-n^2+1)(n^2+1)(n^2+2)(3n^1-n^2+3), \tag{61}$$

for the contribution of the vectors (47) into the dimension of the $Sp(4)$ irrep $\langle n^1 n^2\rangle$. To obtain (61), the well known identities were used,

$$\sum_{\kappa=1}^{n}\kappa=\frac{1}{2}n(n+1), \quad \sum_{\kappa=1}^{n}\kappa^2=\frac{1}{6}n(n+1)(2n+1). \tag{62}$$

In a similar manner, the contribution N_2 of the vectors (46) into $\dim\langle n^1 n^2\rangle$ can be found to give the following result:

$$N_2=\tfrac{1}{6}(n^1-n^2)(n^1-n^2+1)(n^1-n^2-1)(n^2+1). \tag{63}$$

We obtain the total dimension as

$$\dim\langle n^1 n^2\rangle = N_1+N_2=\tfrac{1}{6}(n^1-n^2+1)(n^2+1)(n^1+n^2+3)(n^1+2), \tag{64}$$

which coincides with the standard expression for the $Sp(4)$ irrep dimension.

V. SPIN-MASS CONTENT OF THE BHABHA PARTICLES

As mentioned in the Introduction, the Bhabha equation (1) describing a free particle of arbitrary spin,[1–4] with $Sp(4)$ as its symmetry group, represents several mass states occurring for each spin which can take values $\frac{1}{2}(n^1+n^2), \frac{1}{2}(n^1+n^2)-1,\ldots,0$ or $\frac{1}{2}$ depending on whether n^1+n^2 is even or odd. In this section, we find all possible values of spin s which a particle of mass M_τ can take. Since M_τ is proportional to $1/\tau$ [see Eq. (2)] we need to find the maximum and minimum values of s for a given τ in order to find spin-mass content of the Bhabha particles. This is facilitated by taking into account various limitations on the non-negative integer powers of each epd appearing in the polynomials (54), (55).

First of all, it should be noted that the root diagram shown in Fig. 1 is symmetric with respect to the substitution T_0 by $-T_0$ and T_0 by S_0 [these transformations are the elements of the group of Weyl reflections for $Sp(4)$]. It means that at $-\tau$ the same values of spin s are as admissible as for τ. Therefore we can restrict ourselves to the consideration of only the non-negative values of τ. Besides, the maximum possible value of τ is identical to the largest possible value of spin s in the $Sp(4)$ irrep $\langle n^1 n^2\rangle$. Both τ and s are simultaneously integer or half-integer.

The maximum possible value of spin, S_{max}, for a given τ can be calculated using the restrictions inherent to the part of the basis corresponding to $s \geqslant \frac{1}{2}(n^1 - n^2)$. By noting the fact that each of the powers in (55) are non-negative integers, we have

$$(1)\ \beta \geqslant 0, \quad (2)\ \beta \leqslant n^1 - \tau - s, \quad (3)\ \beta \geqslant -n^2 - \tau + s, \quad (4)\ \beta \leqslant n^1 - n^2. \tag{65}$$

It follows from (2) and (3) above that the largest possible value of s is equal to $(n^1+n^2)/2$. Therefore for τ we obtain

$$-\frac{1}{2}(n^1+n^2) \leqslant \tau \leqslant \frac{1}{2}(n^1+n^2). \tag{66}$$

In the interval $\frac{1}{2}(n^1-n^2) \leqslant \tau \leqslant \frac{1}{2}(n^1+n^2)$, the maximum value of spin S_{max} corresponds to the cross point of the boundaries determined by inequalities (1) and (2), i.e.,

$$S_{max} = n^1 - \tau. \tag{67}$$

For $0 \leqslant \tau \leqslant \frac{1}{2}(n^1-n^2)$ the value S_{max} corresponds to the cross-point of the boundaries determined by inequalities (2) and (3), i.e.,

$$S_{max} = \frac{1}{2}(n^1+n^2). \tag{68}$$

If $\tau < 0$ we obtain

$$S_{max} = n^1 - |\tau|, \quad \text{for } \frac{1}{2}(n^1-n^2) \leqslant |\tau| \leqslant \frac{1}{2}(n^1+n^2) \tag{69}$$

and

$$S_{max} = \frac{1}{2}(n^1+n^2), \quad \text{for } 0 \leqslant |\tau| \leqslant \frac{1}{2}(n^1-n^2), \tag{70}$$

because of symmetry with respect to the reflection $\tau \to -\tau$ mentioned before.

The minimum value of spin S_{min} at given τ is contained in the part of the basis (54) corresponding to $s \leqslant \frac{1}{2}(n^1-n^2)$. In this case, there exists the following limitations:

$$(1)\ \beta \geqslant 0, \quad (2)\ \beta \leqslant n^2 - \tau + s, \quad (3)\ \beta \geqslant -n^2 - \tau + s, \quad (4)\ \beta \leqslant 2s. \tag{71}$$

All possible (non-negative) values of τ can be distributed in two parts:

$$(i)\ \frac{1}{2}(n^1+n^2) \geqslant \tau \geqslant n^2, \quad (ii)\ n^2 \geqslant \tau \geqslant 0. \tag{72}$$

In the first case, the value S_{min} corresponds to the cross-point of the boundaries determined by the inequalities (1) and (2), i.e.,

$$S_{min} = \tau - n^2, \quad \text{for } \frac{1}{2}(n^1+n^2) \geqslant \tau \geqslant n^2. \tag{73}$$

In the second case the value S_{min} corresponds to an admissible point which is closest to the cross-point of boundaries determined by the inequalities (1) and (4), i.e., $S_{min} = 0, 1/2$ or 1 depending on the parity of n^1 and n^2. In order to establish the exact value of S_{min} it is necessary to take into account that in correspondence with (49) β can ''jump'' only by 2 units and has the same parity as $n^2 + \tau - s$. Therefore the value $S_{min} = 0$, i.e., $\beta_{min} = 0$, exists only if $n^2 + \tau$ is even. At $n^2 + \tau = \text{odd}$, we obtain $S_{min} = 1$. If $n^1 + n^2 = \text{odd}$, when all spins are half-integers, we have $S_{min} = 1/2$. We can summarize these results as follows:

$$S_{min} = |\tau| - n^2, \quad \text{for } |\tau| \geqslant \frac{1}{2}(n^1-n^2), \tag{74}$$

$$S_{min} = 0, \quad \text{if } n^1 + n^2 = \text{even}, \quad n^2 + |\tau| = \text{even}, \tag{75}$$

$$S_{\min}=1, \quad \text{if } n^1+n^2=\text{even}, \quad n^2+|\tau|=\text{odd}, \tag{76}$$

$$S_{\min}=\tfrac{1}{2}, \quad \text{if } n^1+n^2=\text{odd}. \tag{77}$$

Thus each mass state can take spin s whose value ranges from $S_{\max}, S_{\max}-1,\ldots,S_{\min}$, given by (69), (70), (74)–(77). The exclusion to these rules occurs for the irrep $\langle n^1 n^1\rangle$. In this case, the limitation (4) in (65) allows only one value $\beta=0$. It follows from (49) that the spin s jumps by 2 units in this case. Thus in correspondence with limitation (2) in (65) we obtain

$$s=n^1-|\tau|, n^1-|\tau|-2,\ldots,1 \quad \text{or} \quad 0, \tag{78}$$

for $n^1-|\tau|=$ odd or even, respectively.

In general, all the values of s except $S_{\min}$ and $S_{\max}$ are degenerate. The multiplicity of each pair (s,τ) is equal to the values of β at the given values of n^1, n^2, s and τ. It can also be found using the generating function described in Sec. IV. Because of the unique possible value $\beta=0$, all combinations of (s,τ) in the irrep $\langle n^1 n^1\rangle$ are not degenerate.

VI. CONCLUSION

In the present paper, the basis of the general $Sp(4)$ irrep $\langle n^1 n^2\rangle$ adapted to the noncanonical group chain $Sp(4)\supset SU_s(2)\times U_\tau(1)$ is constructed. Since the group $Sp(4)\simeq SO(5)$ is a symmetry group of the Bhabha equation,[1–4] the study of such a basis allows us to solve naturally the problem of finding the spin-mass content of the Bhabha particles, i.e., to determine what values of spin s are compatible with a definite mass value $M_\tau=n/2\tau$. The multiplicity of each pair (s,τ) can be easily calculated using the generating function for this basis found in Sec. IV. The noncanonical basis discussed in this paper is also useful to consider the nonrelativistic limit of the Bhabha equation in an interaction with the external fields and to obtain some intrinsic properties of the Bhabha particles (magnetic moment etc.).[15] After knowing the explicit form of the basis, it is possible to calculate the matrix elements of the $Sp(4)$ group generators in order to rewrite the Bhabha equation (1) in a matrix form. Because of the shortage of the space we do not discuss it in this paper.

ACKNOWLEDGMENTS

The authors are thankful to Professor M. Moshinsky and Professor A. G. Nikitin for valuable discussions and remarks. The paper is supported partly by the Russian Foundation of Fundamental Research (Grant No. RFFI N 99-01-01163).

[1] H. J. Bhabha, Rev. Mod. Phys. **17**, 200 (1945).
[2] R. A. Krajcik and M. M. Nieto, Phys. Rev. D **10**, 4049 (1974); **11**, 1442 (1975); **11**, 1459 (1975); **13**, 924 (1976); **14**, 418 (1976); **15**, 433 (1977).
[3] R. K. Loide, I. Ots, and R. Saar, J. Phys. A **30**, 4005 (1977).
[4] M. Moshinsky, A. G. Nikitin, A. Sharma, and Yu. F. Smirnov, J. Phys. A **31**, 6045 (1998).
[5] M. Moshinsky and Yu. F. Smirnov, J. Phys. A **29**, 6027 (1996).
[6] B. H. Flowers, Proc. R. Soc. London, Ser. A **212**, 248 (1952); K. T. Hecht, Nucl. Phys. **63**, 177 (1965); J. C. Parikh, Nucl. Phys. **63**, 214 (1965); K. Ahmed and R. T. Sharp, J. Math. Phys. **11**, 1112 (1970); Yu. F. Smirnov and V. N. Tolstoy, Rep. Math. Phys. **4**, 97 (1973).
[7] M. Moshinsky and V. Bargmann, Nucl. Phys. **23**, 77 (1961).
[8] M. Moshinsky, J. Patera, R. T. Sharp, and P. Winternitz, Ann. Phys. (N.Y.) **95**, 139 (1975).
[9] M. Moshinsky and Yu. F. Smirnov, *The Harmonic Oscillator in Modern Physics* (Harwood Academic, Amsterdam, Netherlands, 1996).
[10] N. Y. Vilenkin, *Special Functions and Theory of Group Representations* (American Mathematical Society Translations, Providence, RI, 1968).
[11] M. A. Lohe and C. A. Hurst, J. Math. Phys. **12**, 1882 (1971).
[12] M. Moshinsky and C. Quesne, Phys. Lett. B **29**, 482 (1969).
[13] C. Quesne, J. Math. Phys. **14**, 366 (1973).
[14] N. Hambli, J. Michelson, and R. T. Sharp, J. Math. Phys. **37**, 3022 (1996).
[15] A. Sharma, M. Moshinsky, and Yu. F. Smirnov, J. Phys. A **31**, 10017 (1998).

Time-asymptotic traveling-wave solutions to the nonlinear Vlasov–Poisson–Ampère equations

Carlo Lancellotti
University of Virginia, Charlottesville, Virginia 22903-2442

J. J. Dorning[a)]
Mathématiques pour l'Industrie et la Physique (UMR CNRS 5640), Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France

(Received 12 January 1999; accepted for publication 28 April 1999)

We consider the Vlasov–Poisson–Ampère system of equations, and we seek solutions for the electric field $E(x,t)$ that are periodic in space and asymptotically almost periodic in time. Introducing the representation $E(x,t)=T(x,t)+A(x,t)$ (where T and A are, respectively, the transient and time-asymptotic parts of E) enables us to decompose the nonlinear Poisson equation into a *transient equation* and a *time-asymptotic equation*. We then study the latter in isolation as a bifurcation problem for A with the initial condition and T as parameters. We show that the Fréchet derivative at a generic bifurcation point has a nontrivial null space determined by the roots of a *Vlasov dispersion relation*. Hence, the bifurcation analysis leads to a general solution for A given (at leading order) by a discrete superposition of traveling-wave modes, whose frequencies and wave numbers satisfy the Vlasov dispersion relation, and whose amplitudes satisfy a system of nonlinear algebraic equations. In applications, there is usually a finite number of roots to the dispersion relation, and the equations for the time-asymptotic wave amplitudes reduce to a finite dimensional bifurcation problem in terms of the amplitude of the initial condition. © *1999 American Institute of Physics.* [S0022-2488(99)01208-6]

I. INTRODUCTION

The subject of this work is the well-known Vlasov–Poisson–Ampère (VPA) system of equations

$$\frac{\partial f_\alpha}{\partial t}+v\frac{\partial f_\alpha}{\partial x}+\frac{q_\alpha}{m_\alpha}E\frac{\partial f_\alpha}{\partial v}=0, \tag{1a}$$

$$\frac{\partial E}{\partial x}=4\pi\sum_\alpha q_\alpha\int dv\, f_\alpha, \tag{1b}$$

$$\frac{\partial E}{\partial t}=-4\pi\sum_\alpha q_\alpha\int dv\, v f_\alpha, \tag{1c}$$

for the propagation of longitudinal electric signals in a collisionless plasma. Here, $E(x,t)$ is the macroscopic longitudinal electric field, and $\alpha=1,\ldots,N_S$ are the indexes of the N_S different species of particles present in the plasma, each with charge q_α, mass m_α and single-particle distribution function $f_\alpha(x,v,t)$. A large body of recent mathematical literature[1–4] has produced extensive results on existence, uniqueness, and regularity of both classical and weak solutions to these equations. However, any concrete calculation of these solutions is made very difficult, especially in the long-time limit, by the strong nonlinearity of the problem.

[a)]Permanent address: University of Virginia, Charlottesville, Virginia 22903-2442.

Physically, the nonlinearity in the VPA system corresponds to the self-consistent feedback from the electric field on the distribution functions. Interestingly, in many wave-propagation problems we expect this feedback to become less and less important in the long-time limit, because the anharmonic mixing of the single-particle trajectories makes the plasma less and less able to exchange energy with the electric field in any coherent fashion. This fact was first pointed out in the physics literature by O'Neil.[5] In a classic paper, he argued that a sinusoidal perturbation to a thermal equilibrium will, in general, either Landau damp to a zero electric field[6] or lead to a nonzero time-asymptotic solution for the electric field, depending on the amplitude of the initial disturbance and on the magnitude of the Landau damping rate. O'Neil assumed the nonzero time-asymptotic solution for E to be given by a single traveling wave of constant amplitude, but in more realistic physical situations one can reasonably conjecture that the final state will be comprised of more than one wave mode. In fact, this scenario has been confirmed by a significant amount of experimental and numerical evidence.[7–9] Recent rigorous investigations[10–12] have shown that the nonlinear VPA system does indeed admit undamped traveling-wave solutions like those suggested in the physics literature. Particularly relevant to what will follow are the nonlinearly superimposed traveling-wave BGK-type[13] solutions obtained by Buchanan and Dorning.[12] However, it is not known whether any of these solutions can be reached as time-asymptotic limits of solutions to the initial value problem for the VPA system.

In this article we develop a new procedure for the analysis of the long-time behavior of the solutions to the VPA system for a certain class of initial conditions. Our method is based on the representation of the electric field as the sum of a transient term and a time-asymptotic term, and correspondingly, on the decomposition of the VPA problem into a transient part and a time-asymptotic part. This decomposition turns out to be fruitful in a number of ways, not all of which will be explored in this paper. For instance, we shall not analyze the transient problem, which seems amenable to a relatively straightforward perturbation analysis[14] since most of the well-known secularities in the VPA problem are associated with time-asymptotic evolution. We shall focus, instead, on the time-asymptotic part of the problem; the basic idea will be to show that this time-asymptotic part can be studied *in isolation* as a bifurcation problem for the time-asymptotic electric field with the initial condition and the transient field playing the role of parameters. Our main result shows that *if* the VPA system possesses a nonzero small-amplitude time-asymptotic solution (in a sense that will be defined below), then the corresponding electric field is given at leading order by a superposition of traveling-wave modes associated with the roots of a ''time-asymptotic'' Vlasov dispersion relation. This dispersion relation is completely determined by the initial condition and by the transient electric field, and the same is true for the amplitudes of these traveling-wave modes, which satisfy a nonlinear system of algebraic equations.

II. PRELIMINARIES

Throughout this study E and f_α will be assumed to be bounded $\mathcal{C}^1$ functions of their arguments, with $f_\alpha \geqslant 0$. In fact, it is enough to assume f_α to be bounded and non-negative at time zero, since then the Vlasov equation implies the *a priori* bounds

$$0 \leqslant f_\alpha(x,v,t) \leqslant \sup_{(x,v)} f_\alpha(x,v,0) \equiv M_\alpha . \tag{2}$$

We also shall assume f_α and $v f_\alpha$ to be integrable functions of v on $\mathbb{R}$ so that the charge and current densities are well defined. Both E and the f_α will be *spatially 2π-periodic*. In this case it is easy to show that replacing the Ampère equation by its spatially averaged form

$$\frac{dE_0}{dt} = -2 \sum_\alpha q_\alpha \int_{-\pi}^{+\pi} dx \int dv\, v f_\alpha \tag{1c$'$}$$

(where $E_0(t)$ is the $k=0$ spatial Fourier component of $E(x,t)$) yields a system of equations which is completely equivalent to Eqs. (1). Hence, we shall write the VPA system in the compact form

$$\frac{\partial f_\alpha}{\partial t}+v\frac{\partial f_\alpha}{\partial x}+\frac{q_\alpha}{m_\alpha}E\frac{\partial f_\alpha}{\partial v}=0, \tag{3a}$$

$$E=\Lambda(f_1,\ldots,f_{N_S}), \tag{3b}$$

where Λ is defined by its spatial Fourier components

$$\Lambda_k(f_1,\ldots,f_{N_S})\equiv\begin{cases}\dfrac{4\pi}{ik}\sum_\alpha q_\alpha\int dv f_{\alpha,k} & k\neq 0\\ E_0(0)-4\pi\sum_\alpha q_\alpha\int_0^t d\tau\int dv\, v f_{\alpha,0} & k=0\end{cases} \tag{4}$$

which have been obtained from Eq. (1b) and (1c′) for $k\neq 0$ and $k=0$, respectively. Here, the $f_{\alpha,k}$ are the spatial Fourier components of f_α, and $E_0(0)$ is an assigned initial condition for $E_0(t)$. The spatially uniform part of the Poisson equation, Eq. (1b), reduces to the zero-charge condition

$$\sum_\alpha q_\alpha\int_{-\pi}^{+\pi}dx\int dv f_\alpha=0. \tag{5}$$

From the Vlasov equation it follows immediately that this condition is satisfied as long as it holds at $t=0$.

The distribution functions satisfy initial conditions of the form

$$f_\alpha(x,v,0)=\mathcal{F}_\alpha(x,v)\equiv F_\alpha(v)+h_\alpha(x,v), \tag{6}$$

where the $F_\alpha(v)$ correspond to a Vlasov equilibrium and satisfy the conditions

$$q=\sum_\alpha q_\alpha\int F_\alpha(v)dv=0, \tag{7}$$

$$j=\sum_\alpha q_\alpha\int vF_\alpha(v)dv=0. \tag{8}$$

The function h_α in Eq. (6) will be taken to have no spatially uniform part, so that

$$\int_{-\pi}^{+\pi}dx h_\alpha(x,v)=0. \tag{9}$$

Equations (7) and (9) ensure that the $f_\alpha(x,v,0)$, and thus the $f_\alpha(x,v,t)$, satisfy Eq. (5). Clearly, once the $f_\alpha(x,v,0)$ are chosen, all the initial Fourier components for the field $E_k(0)$ with $k\neq 0$ are automatically assigned via the Poisson equation.

The characteristic system for the Vlasov equation, Eq. (3a), is given by Newton's equations $(dx/dt)=v$, $(dv/dt)=(q_\alpha/m_\alpha)E(x,t)$. All the electric fields E appearing in this paper will be such that these equations have global classical solutions that can be extended indefinitely in t according to classic theorems on ODEs.[15] Then, the general solution to the Vlasov equation can be written as

$$f_\alpha(x,v,t)=\mathcal{F}_\alpha(x_0^E(x,v,t),v_0^E(x,v,t))\equiv\mathbf{f}_\alpha(E,\mathcal{F}_\alpha), \tag{10}$$

where we have introduced the the ''inverse trajectories'' $x_0^E(x,v,t),v_0^E(x,v,t)$ determined by Newton's equations; these functions associate with each phase-point (x,v) the initial condition

(x_0^E, v_0^E) at time zero that leads to (x,v) at time t. We shall use the notation $f_\alpha(x,v,t) = \mathbf{f}_\alpha(E,\mathcal{F}_\alpha)$ whenever we want to emphasize the functional dependence of the distribution function on E and the initial condition. Similarly, we shall write

$$\mathbf{f}'_\alpha(E,\mathcal{F}_\alpha) = \frac{\partial}{\partial v}[\mathcal{F}_\alpha(x_0^E(x,v,t), v_0^E(x,v,t))]. \tag{11}$$

Substituting Eq. (10) into Eq. (3b) reduces the problem to a single nonlinear equation for E

$$E = \mathcal{N}(E, F_\alpha, h_a), \tag{12}$$

where we have defined

$$\mathcal{N}(E,F_\alpha,h_\alpha) \equiv \Lambda(\mathbf{f}_1(E,\mathcal{F}_1),\ldots,\mathbf{f}_{N_S}(E,\mathcal{F}_{N_S})). \tag{13}$$

Equation (12) will be called the VPA equation; here, of course, the Vlasov equation has become part of the definition of $\mathcal{N}$ (through Eqs. (10) and (4)).

III. *A–T* DECOMPOSITION

As mentioned in our introductory discussion, we shall seek solutions for E that are the sum of a *transient* part and a *time-asymptotic* part, of the form

$$E(x,t) = T(x,t) + A(x,t). \tag{14}$$

Since our interest is in periodic traveling-wave solutions, all the functions involved will be assumed to be continuously differentiable and periodic in x. As far as the time variable is concerned, a very general class of functions that have the representation in Eq. (14) is given by the *asymptotically almost periodic continuous* (a.a.p.c.) functions of t.[16] Let $\mathcal{AP}$ be the set of all the *almost periodic continuous* (a.p.c.) functions of t uniformly with respect to x.[17] Let $\mathcal{T}$ be the set of all the continuous functions of t on $\mathbb{R}^+ g(x,t)$ such that $\lim_{t\to+\infty} g(x,t) = 0$ uniformly in x. The space of the a.a.p.c. functions of t on $\mathbb{R}^+$ (uniformly with respect to x, periodic and continuously differentiable in x) is given by the direct sum $\mathcal{AP}+\mathcal{T}$. Here, we shall focus on the subspace $\mathcal{W}\subset\mathcal{AP}+\mathcal{T}$ of the functions that are also continuously differentiable in t, i.e.,

$$\mathcal{W} \equiv \{W \in \mathcal{C}^1(\mathbb{R}\times\mathbb{R}^+) \;\; s.t. \;\; W = A + T, \quad A \in \mathcal{AP}, \quad T \in \mathcal{T}\}. \tag{15}$$

As shown by Fréchet,[16] a.a.p.c. functions enjoy many of the classic properties of almost periodic functions. In particular, if $W(x,t)$ is a.a.p.c. in t then it is bounded and uniformly continuous on $\mathbb{R}^+$; in fact, $\mathcal{AP}$, $\mathcal{T}$ and $\mathcal{W}$ can be made into Banach spaces with the supnorm

$$\|W\| = \sup_{I,\mathbb{R}^+} |W|, \tag{16}$$

where $I \equiv [-\pi, +\pi]$. Moreover, the *mean value* of W,

$$\mathbf{M}_t[W(x,t)] = \lim_{\sigma\to\infty} \frac{1}{\sigma} \int_0^\sigma W(x,t)\,dt \tag{17}$$

is always well defined and coincides with the mean value of the a.p.c. part of W, as follows from the fact that $\mathbf{M}_t[T(x,t)] = 0 \;\; \forall T \in \mathcal{T}$ (this will be referred to as "Fréchet's Lemma"[16] in what follows). The existence of the mean is very important because it makes it possible to associate with each a.a.p.c. W its *Fourier–Bohr coefficients*,

$$w(\lambda,x)=\mathbf{M}_t[W(x,t)e^{-i\lambda t}]=\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma W(x,t)e^{-i\lambda t}dt \tag{18}$$

which, of course, coincide with the Fourier–Bohr coefficients of the a.p.c. part of W. For this latter part we know that there exist at most a countably infinite set of real numbers λ_i such that $w(\lambda,x)\equiv 0\ \ \forall\lambda\neq\lambda_i$.[17] Hence, with each a.a.p.c. function W there is associated a unique Fourier–Bohr series

$$W(x,t)\sim\sum w_i(x)e^{i\lambda_i t} \tag{19}$$

(here $w_i(x)\equiv w(\lambda_i,x)$). The series in Eq. (19) coincides with the Fourier–Bohr series for the a.p.c. (time-asymptotic) part of W and determines it uniquely. Of course, it *does not* determine the *whole* function W, whose transient part has been averaged away in the computation of the coefficients w_i. In fact, the Fourier–Bohr series of an a.a.p.c. function gives an explicit representation of the fundamental projection operator $P_a:\mathcal{W}\to\mathcal{AP}$ such that

$$P_aE=A. \tag{20}$$

By associating with any $E\in\mathcal{W}$ its a.p.c. part A, P_a will play a major role in what follows, enabling us to ''sort out'' the essential features of long-time wave propagation from less interesting (and more complicated) transient phenomena.

Of course, the validity of the assumption that $E(x,t)\in\mathcal{W}$ is far from obvious, since there seems to be no easy way to prove rigorously *a priori* that the problem admits an asymptotic a.p.c. state. On the contrary, it could be argued that the intricacies of the nonlinear particle dynamics will always generate some ''noise,'' which cannot be reasonably expected to be a.a.p.c. in time. We are thinking here of the particles belonging to the thin stochastic layers generated by multiwave resonances in the phase plane, as described by Rechester and Stix[18] and by Buchanan and Dorning.[12] On the other hand, Buchanan and Dorning[12] obtained solutions that *are* a.p.c. to *leading order* in the field amplitude, and showed that the noise coming from the stochastic layers vanishes exponentially with that amplitude. Motivated by these results, we study the evolution of initial conditions that produce a *small* time-asymptotic state, and show that the *approximate* solution for $E(x,t)$ is a.a.p.c. in time, with an error that is negligible with respect to the time-asymptotic field amplitude.

Let us substitute the a.a.p.c. representation of E, Eq. (14), into the nonlinear VPA equation, Eq. (12); in order to ensure that this equation is well defined, we need to assume that the integral in dt in Eq. (4) is bounded (and thus a.a.p.c. (Ref. 17)) in time. This implies the necessary condition

$$4\pi\sum_\alpha q_\alpha \lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma d\tau\int dv\, v\, f_{\alpha,0}(v,\tau)=0 \tag{21}$$

which can also be obtained directly by taking the time-average of the Ampère equation, Eq. (1c). We now apply the projection operator P_a introduced above to both sides of Eq. (12) in order to decompose the problem into its transient and time-asymptotic components. This yields the following system of two coupled equations for the time-asymptotic field A and the transient field T:

$$A=P_\alpha\,\mathcal{N}(A+T,F_\alpha,h_\alpha), \tag{22a}$$

$$T=(I-P_a)\mathcal{N}(A+T,F_\alpha,h_\alpha). \tag{22b}$$

Our strategy will be to focus on the ''asymptotic equation,'' Eq. (22a). We shall show that it is possible to obtain important information about the structure of solutions to Eq. (22a) *independently* of the details of T. This will be done in the following steps:

(1) First, we shall show that the asymptotic equation can be studied as an infinite-dimensional bifurcation problem, because it possesses a manifold of vanishing solutions corresponding to different choices of the initial perturbation h_α and of the corresponding transient field T.

(2) Then, we shall show that there is a natural way to linearize Eq. (22a), which leads to a *time-asymptotic linear equation*, whose solutions are very different from those found in the traditional linear theory. In particular, the time-asymptotic linear theory will yield results that are perfectly consistent with the undamped nonlinear multiple traveling-wave solutions recently discovered by Buchanan and Dorning.[12]

(3) Finally, we shall exploit the properties of the time-asymptotic linear operator in order to reduce the original nonlinear problem to a lower-dimensional system of bifurcation equations for the amplitudes of the leading-order Fourier–Bohr coefficients of A.

IV. PURELY TRANSIENT ANALYSIS

We want to show that the time-asymptotic equation, Eq. (22a), is satisfied by the zero time-asymptotic field $A\equiv 0$ independently of the choice of the initial condition. Substituting $A\equiv 0$ into Eq. (22a) gives the equation

$$P_a \mathcal{N}(T,F_\alpha,h_\alpha)=0, \tag{23}$$

which in turn entails the analysis of the Vlasov equation for a purely transient field

$$\frac{\partial f_\alpha}{\partial t}+v\frac{\partial f_\alpha}{\partial x}+\frac{q_\alpha}{m_\alpha}T\frac{\partial f_\alpha}{\partial v}=0. \tag{24}$$

Of course, the general solution of this equation has the form $f_\alpha^T(x,v,t) =\mathcal{F}_\alpha(x_0^T(x,v,t),v_0^T(x,v,t))$, where the $[x_0^T(x,v,t),v_0^T(x,v,t)]$ give the ''inverse particle trajectory,'' i.e., the starting point at time zero for a particle that arrives at the point (x,v) at time t under the influence of the electric field T. The fundamental observation is that, as the *transient T* dies away, it acts on the particles more and more weakly, so that the trajectories tend asymptotically to straight lines as $t\to\infty$. This suggests that in the long-time limit the distribution function must approach some kind of Vlasov equilibrium, at least in a coarse-grained sense, so that Eq. (23) is satisfied.

To establish this, let us consider Newton's equations for the purely transient, spatially periodic field T: $\dot{x}=v$, $\dot{v}=T(x,t)$ (where we have set $q_\alpha/m_\alpha=1$ to avoid excessive notational detail). For the initial value problem with the initial conditions $x(0)=x_0$, $v(0)=v_0$, these equations can be written in the integral form

$$x(t)-x_0=v_0t+\int_0^t d\tau\int_0^\tau d\tau'\,T(x(\tau'),\tau'), \tag{25a}$$

$$v(t)-v_0=\int_0^t d\tau\;T(x,(\tau),\tau). \tag{25b}$$

We want the field $T(x,t)$ to decay fast enough, as $t\to\infty$, that both $\sup_x|T(x,t)|$ and $\int_t^\infty d\tau'\,\sup_x|T(x,\tau')|$ be integrable over the positive real axis. This will be ensured when $\sup_x|T(x,t)|$ tends to zero at least as fast as $t^{-\eta}$ as $t\to\infty$, with $\eta>2$. Under this condition, it follows immediately from Eq. (25b) that

$$\lim_{t\to\infty} v(t)=v_0+\int_0^\infty d\tau\;T(x(\tau),\tau), \tag{26}$$

where the integral on the right-hand side is a function only of the initial point (x_0,v_0) of the trajectory $x(t)$ in phase space. Hence, we shall write Eq. (26) in the form

$$\lim_{t\to\infty} v(t) = v_0 + H(x_0, v_0) \equiv v_\infty . \tag{27}$$

Then, Eq. (25a) can be rewritten immediately as

$$x(t) - x_0 = v_\infty t + \int_0^t d\tau \int_\tau^\infty d\tau' \, T(x(\tau'), \tau'). \tag{28}$$

Thus, under the above conditions on T,

$$\lim_{t\to\infty} [x(t) - v_\infty t] = x_0 + \int_0^\infty d\tau \int_\tau^\infty d\tau' \, T(x(\tau')\tau') \tag{29}$$

which will be written in the form

$$\lim_{t\to\infty} [x(t) - v_\infty t] = x_0 + G(x_0, v_0) \equiv x_\infty . \tag{30}$$

Equations (27) and (30) express the fact that, as t goes to infinity and $T(x,t)$ tends to zero, each trajectory tends asymptotically to the straight line trajectory starting at $t=0$ from the ''fictitious'' phase point (x_∞, v_∞). That is,

$$[x(t), v(t)] \xrightarrow[t\to\infty]{} [x_\infty + v_\infty t, v_\infty]. \tag{31}$$

Now, we want to obtain an analogous result for the ''inverse'' trajectories $[x^T(x,v,t), v^T(x,v,t)]$. This requires a slightly more sophisticated analysis, which leads to the following theorem:

Theorem 1: *If there is a number $\eta > 2$ such that both $\sup_x |T(x,t)|$ and $\sup_x |(dT/dx)(x,t)|$ go to zero as $t^{-\eta}$ as $t \to \infty$, then the corresponding inverse trajectories can be written in the form,*

$$x_0^T(x,v,t) = x - vt + \tilde{G}(x - vt, v) + \tau_1(x,v,t), \tag{32a}$$

$$v_0^T(x,v,t) = v + \tilde{H}(x - vt, v) + \tau_2(x,v,t), \tag{32b}$$

where $\tilde{G}, \tilde{H} \in \mathcal{C}(I \times \mathbb{R}^+)$ and $\lim_{t\to\infty} \tau_1(x,v,t) = \lim_{t\to\infty} \tau_2(x,v,t) = 0$, uniformly in x and v.

Proof: The proof of this result is based on the fact that, by definition, $x_0^T(x,v,t)$ and $v_0^T(x,v,t)$ must satisfy the Vlasov equations

$$\frac{\partial x_0^T}{\partial t} + v \frac{\partial x_0^T}{\partial x} + T \frac{\partial x_0^T}{\partial v} = 0, \tag{33}$$

$$\frac{\partial v_0^T}{\partial t} + v \frac{\partial v_0^T}{\partial x} + T \frac{\partial v_0^T}{\partial v} = 0, \tag{34}$$

which can be written in the integral form

$$x_0^T(x,v,t) = x - vt - \int_0^t d\tau \left[T \frac{\partial x_0^T}{\partial v} \right]_{(x - v(t-\tau), v)}, \tag{35}$$

$$v_0^T(x,v,t) = v - \int_0^t d\tau \left[T \frac{\partial v_0^T}{\partial v} \right]_{(x - v(t-\tau), v)}. \tag{36}$$

In this case, it is not immediately obvious that one can take $t \to \infty$ in the limits of integration, as was done for Eqs. (25). The core of the proof will be precisely the determination of appropriate bounds on the growth in time of $(\partial x_0^T/\partial v)$ and $(\partial v_0^T/\partial v)$, in order to show that the integrands on the right-hand sides of Eqs. (35) and (36) are, indeed, integrable at infinity. The superscript ''T'' will be dropped for the remainder of this proof.

Let us consider, first, the ''direct'' trajectories, written in the form $x(x_0,v_0,t),v(x_0,v_0,t)$. We shall exploit the decay properties of T in order to obtain bounds on the growth in time of the partial derivatives $\partial x/\partial x_0$, $\partial v/\partial x_0$, $\partial x/\partial v_0$, and $\partial v/\partial v_0$. To begin, we take the derivative with respect to x_0 of both sides of Eq. (25a),

$$\frac{\partial x}{\partial x_0}(x_0,v_0,t)=1+\int_0^t d\tau \int_0^\tau d\tau' \frac{\partial x}{\partial x_0}(x_0,v_0,\tau')T_x(x(x_0,v_0,\tau'),\tau'), \tag{37}$$

where $T_x(x,t)\equiv(\partial T/\partial x)(x,t)$. Now, let us multiply both sides of Eq. (37) by $T_x(x(x_0,v_0,t),t)$, rename the variable t as t' and integrate in dt' from 0 to t. We obtain

$$g(x_0,v_0,t)=h(x_0,v_0,t)+\int_0^t dt' T_x(x(x_0,v_0,t'),t')\int_0^{t'} d\tau\, g(x_0,v_0,\tau), \tag{38}$$

where we have defined

$$g(x_0,v_0,t)\equiv\int_0^t dt' \frac{\partial x}{\partial x_0}(x_0,v_0,t')T_x(x(x_0,v_0,t'),t') \tag{39}$$

and

$$h(x_0,v_0,t)\equiv\int_0^t dt' T_x(x(x_0,v_0,t'),t'). \tag{40}$$

The second term on the right-hand side of Eq. (38) is then integrated by parts, yielding

$$\begin{aligned}\int_0^t dt' T_x(x(x_0,v_0,t'),t')\int_0^{t'} d\tau\, g(x_0,v_0,\tau)&=h(x_0,v_0,t)\int_0^t dt' g(x_0,v_0,t')\\&\quad-\int_0^t dt' g(x_0,v_0,t')h(x_0,v_0,t')\\&=\int_0^t dt' g(x_0,v_0,t')\int_{t'}^t d\tau\, T_x(x(x_0,v_0,\tau),\tau).\end{aligned} \tag{41}$$

Thus, Eq. (38) can be written as a Volterra equation, i.e.,

$$g(x_0,v_0,t)=h(x_0,v_0,t)+\int_0^t dt' K(x_0,v_0,t,t')g(x_0,v_0,t'), \tag{42}$$

where the ''kernel'' $K(x_0,v_0,t,t')$ is given by

$$K(x_0,v_0,t,t')\equiv\int_{t'}^t d\tau\, T_x(x(x_0,v_0,\tau),\tau). \tag{43}$$

Due to the integrability of $|T_x|$,

$$|K(x_0,v_0,t,t')|\leqslant\int_{t'}^{t}d\tau|T_x(x(x_0,v_0,\tau),\tau)|\leqslant\int_{t'}^{\infty}d\tau\,|T_x(x(x_0,v_0,\tau),\tau)|\equiv H(x_0,v_0,t'). \tag{44}$$

Hence, Eq. (42) implies the inequality

$$|g(x_0,v_0,t)|\leqslant|h(x_0,v_0,t)|+\int_0^t dt'\,H(x_0,v_0,t')|g(x_0,v_0,t')|. \tag{45}$$

Then, from Gronwall's inequality it follows that

$$|g(x_0,v_0,t)|\leqslant C(x_0,v_0)e^{B(x_0,v_0)}\equiv M(x_0,v_0)<\infty, \tag{46}$$

where

$$C(x_0,v_0)\leqslant\int_0^{\infty}dt'\,|T_x(x(x_0,v_0,t'),t')| \tag{47}$$

and

$$B(x_0,v_0)\equiv\int_0^{\infty}dt'\,H(x_0,v_0,t'). \tag{48}$$

Now, let us go back to Eq. (37), and to the corresponding equation for $\partial v/\partial x_0$ which is obtained by taking the partial derivative of Eq. (25b) with respect to x_0. According to Eq. (39), these equations can be written in the form

$$\frac{\partial x}{\partial x_0}(x_0,v_0,t)=1+\int_0^t d\tau\; g(x_0,v_0,\tau), \tag{49a}$$

$$\frac{\partial v}{\partial x_0}(x_0,v_0,t)=g(x_0,v_0,t). \tag{49b}$$

From the bound on $|g(x_0,v_0,t)|$ in Eq. (46) it follows immediately that

$$\left|\frac{\partial x}{\partial x_0}(x_0,v_0,t)\right|\leqslant 1+Mt, \tag{50a}$$

$$\left|\frac{\partial v}{\partial x_0}(x_0,v_0,t)\right|\leqslant M. \tag{50b}$$

Here, M is defined as $\sup_{x_0,v_0}M(x_0,v_0)$ which exploits the fact that $M(x_0,v_0)$ is a uniformly bounded function of (x_0,v_0), as follows from Eqs. (46), (47), (48), and from the integrability properties of T_x. Equations (50) give the result sought: under the conditions on the time integrability of $T_x(x,t)$, we have found that $\partial x/\partial x_0$ and $\partial v/\partial x_0$ can grow at most linearly in time.

Using a completely analogous procedure to find similar bounds on the derivatives $\partial x/\partial v_0$ and $\partial v/\partial v_0$ we obtain

$$\left|\frac{\partial x}{\partial v_0}(x_0,v_0,t)\right|\leqslant Nt, \tag{51a}$$

$$\left|\frac{\partial v}{\partial v_0}(x_0,v_0,t)\right|\leqslant N. \tag{51b}$$

Here, $N \equiv \sup_{x_0,v_0} N(x_0,v_0)$, where

$$N(x_0,v_0) \equiv 1 + D(x_0,v_0) e^{B(x_0,v_0)}. \tag{52}$$

$B(x_0,v_0)$ is the same quantity that was defined in Eq. (48) and $D(x_0,v_0)$ satisfies

$$D(x_0,v_0) \leq \int_0^\infty dt' t' |T_x(x(x_0,v_0,t'),t')|. \tag{53}$$

From the inequalities in Eqs. (50) and (51), we easily obtain similar bounds on the time growth of the derivatives of the ''inverse'' functions x_0 and v_0. These bounds follow from the fact that the Jacobian matrix of $[x_0(x,v,t),v_0(x,v,t)]$ (viewed as a function from $\mathbb{R}^2$ onto $\mathbb{R}^2$) is the inverse of the Jacobian of the inverse function $[x(x_0,v_0,t),v(x_0,v_0,t)]$ for each given t, and that this latter has determinant equal to one. We find

$$\left|\frac{\partial x_0}{\partial x}(x,v,t)\right| \leq N, \tag{54a}$$

$$\left|\frac{\partial v_0}{\partial x}(x,v,t)\right| \leq M, \tag{54b}$$

$$\left|\frac{\partial x_0}{\partial v}(x,v,t)\right| \leq Nt, \tag{55a}$$

$$\left|\frac{\partial v_0}{\partial v}(x,v,t)\right| \leq 1 + Mt. \tag{55b}$$

By combining the decay properties of T with Eqs. (55), it is easy to see that the integrands $T(\partial x_0/\partial v)$ and $T(\partial v_0/\partial v)$ in the integrals in Eqs. (35) and (36) are integrable over $\mathbb{R}^+$. Hence, by adding and subtracting the corresponding integrals on $[t,\infty)$, we can write

$$x_0(x,v,t) \approx x - vt + \tilde{G}(x-vt,v), \tag{56a}$$

$$v_0(x,v,t) \approx v + \tilde{H}(x-vt,v), \tag{56b}$$

where the symbol $\approx$ is used to indicate that the two sides are equal up to certain transient functions of (x,v,t) that disappear in the time-asymptotic limit, uniformly with respect to x and v. The functions $\tilde{G}$ and $\tilde{H}$ are given by

$$\tilde{G}(x,v) \equiv \int_0^\infty d\tau \; T(x+v\tau,\tau) \frac{\partial x_0}{\partial v}(x+v\tau,v,\tau), \tag{57a}$$

$$\tilde{H}(x,v) \equiv \int_0^\infty d\tau \; T(x+v\tau,\tau) \frac{\partial v_0}{\partial v}(x+v\tau,v,\tau), \tag{57b}$$

and are clearly continuous (x_0 and v_0 being $\mathcal{C}^1$ according to standard theorems on ordinary differential equations). □

Equations (32) tell us that the inverse particle trajectories at long times tend asymptotically to the straight-line paths $(x-vt,v)$, but with one important qualification. Because of the effects of the transient, the phase point that is mapped backwards along the each straight-line trajectory *is not* (x,v) *itself*, but the ''surrogate'' point $x^T \equiv x + \tilde{G}(x,v)$, $v^T \equiv v + \tilde{H}(x,v)$. Of course, the two functions $\tilde{G}$ and $\tilde{H}$ are not known explicitly, since that would require the complete solution of the nonlinear Vlasov equations for x_0^T and v_0^T, Eqs. (33) and (34). The essential point, here, is that these functions exist and do not depend explicitly on t. It is worth noting that there is a relationship between the functions $\tilde{G},\tilde{H}$ and the functions G,H introduced in Eqs. (27) and (30). Equations (56) imply that

$$x_0 \approx x(x_0,v_0,t) - v(x_0,v_0,t)t + \tilde{G}(x(x_0,v_0,t) - v(x_0,v_0,t)t, v), \tag{58a}$$

$$v_0 \approx v(x_0,v_0,t) + \tilde{H}(x(x_0,v_0,t) - v(x_0,v_0,t)t, v). \tag{58b}$$

Taking the limit $t \to \infty$ and using Eq. (31) yields

$$x_0 = x_\infty + \tilde{G}(x_\infty, v_\infty), \tag{59a}$$

$$v_0 = v_\infty + \tilde{H}(x_\infty, v_\infty), \tag{59b}$$

where (x_∞, v_∞) are the quantities defined in Eqs. (27) and (30).

Substituting Eqs. (32) into the initial condition $\mathcal{F}_\alpha$ yields the solution f_α^T to Eq. (24) in the form

$$f_\alpha^T(x,v,t) = \mathcal{F}_\alpha^T(x - vt, v) + g_\alpha^T(x,v,t), \tag{60}$$

where

$$\mathcal{F}_\alpha^T(x,v) \equiv \mathcal{F}_\alpha(x + \tilde{G}(x,v), v + \tilde{H}(x,v)), \tag{61}$$

and

$$g_\alpha^T(x,v,t) = \mathcal{F}_\alpha(x_0^T, v_0^T) - \mathcal{F}_\alpha^T(x - vt, v). \tag{62}$$

Clearly, $g_\alpha^T(x,v,t) \to 0$ uniformly as $t \to \infty$, since according to Eqs. (32) the two terms on the right-hand side of Eq. (62) become identical in this limit. This means that, as t goes to infinity, f_α^T can be obtained by advection along straight-line trajectories, as long as we replace the initial condition $\mathcal{F}_\alpha$ by the modified function $\mathcal{F}_\alpha^T$ that contains the effects of the transient field T.

It follows from Eq. (60) that f_α^T can be replaced in Eq. (23) by a spatially uniform equilibrium that yields the same values for macroscopic quantities in the time-asymptotic limit. To be explicit, let us consider any integral of the form

$$\int_{\mathbb{R}} du\, \mathcal{G}(v,u) f_\alpha^T(x,u,t), \tag{63}$$

which can be either a charge or current density ($\mathcal{G}(v,u) = 1, u$), or any higher moment ($\mathcal{G}(v,u) = u^n$), or a filtered distribution function.[19] Substituting Eq. (60) into Eq. (63) and introducing the spatial Fourier series of $\mathcal{F}_\alpha^T$, it is easy to see that for all $k \neq 0$,

$$\lim_{t\to\infty} \int_{\mathbb{R}} du\, \mathcal{G}(v,u) f_{\alpha,k}^T(u,t) = \lim_{t\to\infty} \int_{\mathbb{R}} du\, \mathcal{G}(v,u) \mathcal{F}_{\alpha,k}^T(u) e^{-ikut} = 0 \tag{64}$$

by the Riemann–Lebesgue Lemma. Hence,

$$\lim_{t\to\infty} \int_{\mathbb{R}} du\, \mathcal{G}(v,u) f_\alpha^T(x,u,t) = \int_{\mathbb{R}} du\, \mathcal{G}(v,u) F_\alpha^T(u), \tag{65}$$

where

$$F_\alpha^T(v)=\mathcal{F}_{\alpha,0}^T(v)=\frac{1}{2\pi}\int_{-\pi}^{+\pi}dx\mathcal{F}_\alpha(x+\tilde{G}(x,v),v+\tilde{H}(x,v)). \tag{66}$$

The charge and current densities associated with F_α^T are

$$\rho_T=\frac{1}{2\pi}\sum_\alpha q_\alpha\int_{\mathbb{R}}dv\int_{-\pi}^{+\pi}dx\mathcal{F}_\alpha(x+\tilde{G}(x,v),v+\tilde{H}(x,v)), \tag{67}$$

$$j_T=\frac{1}{2\pi}\sum_\alpha q_\alpha\int_{\mathbb{R}}dv\int_{-\pi}^{+\pi}dx\ v\mathcal{F}_\alpha(x+\tilde{G}(x,v),v+\tilde{H}(x,v)). \tag{68}$$

Changing the integration variables to $(x_0,v_0)=[x+\tilde{G}(x,v),v+\tilde{H}(x,v)]$ and noting that this transformation preserves the area in phase space (since it is the time-limit of a Hamiltonian flow), we get

$$\rho_T=\frac{1}{2\pi}\sum_\alpha q_\alpha\int_{\mathbb{R}}dv_0\int_{-\pi}^{+\pi}dx_0\mathcal{F}_\alpha(x_0,v_0), \tag{69}$$

$$j_T=\frac{1}{2\pi}\sum_\alpha q_\alpha\int_{\mathbb{R}}dv_0\int_{-\pi}^{+\pi}dx_0[v_0+H(x_0,v_0)]\mathcal{F}_\alpha(x_0,v_0). \tag{70}$$

Since the spatially uniform part of the initial condition $\mathcal{F}_\alpha(x,v)$ has been taken to be a Vlasov equilibrium, it is easy to see that Eqs. (69) and (70) reduce to

$$\rho_T=0,\quad j_T=j_\infty^T, \tag{71}$$

where

$$j_\infty^T\equiv\frac{1}{2\pi}\sum_\alpha q_\alpha\int_{\mathbb{R}}dv_0\int_{-\pi}^{+\pi}dx_0H(x_0,v_0)\mathcal{F}_\alpha(x_0,v_0). \tag{72}$$

Whenever T is a solution of the self-consistent VPA equation, Eq. (12), $j_\infty^T=0$, as follows directly from the time-and-space averaged Ampère equation, Eq. (21). In general, let us define the subspace $\mathcal{T}_L\subset\mathcal{T}$ of the transient fields that satisfy the hypotheses of Theorem 1 and for which $j_\infty^T=0$. Then, Eq. (71) leads directly to the main result of this section:

Theorem 2: *For any choice of $\mathcal{F}_\alpha$ and $T\in\mathcal{T}_L$ the time-asymptotic equation, Eq. (22a), has the solution $A\equiv0$.*

V. VANISHING ASYMPTOTIC STATES

Theorem 2 can be given the following interpretation: the asymptotic equation, Eq. (22a), possesses an infinite dimensional manifold of vanishing solutions with respect to the initial condition $\mathcal{F}_\alpha(x,v)$, viewed as an infinite dimensional parameter in the Banach space $\mathcal{S}$ of all the $\mathcal{C}^1(\mathbb{R}\times\mathbb{R})$ functions such that

$$\|\mathcal{F}_\alpha\|_{\mathcal{S}}\equiv\sup_x\int_{\mathbb{R}}dv(1+|v|)|\mathcal{F}_\alpha(x,v)|<\infty. \tag{73}$$

In Eq. (6) $\mathcal{F}_\alpha(x,v)$ was written as the sum of a spatially uniform Vlasov equilibrium $F_\alpha(v)$ and another function $h_\alpha(x,v)$; in many situations it is convenient to consider $F_\alpha\in\mathcal{S}$ as fixed and study how the solutions to the VPA problem depend on h_α. Physically, this corresponds to assuming a given ''background equilibrium'' and varying a spatially dependent perturbation. Theorem 2 tells us that $A\equiv0$ is a solution to Eq. (22a) for all $h_\alpha\in\mathcal{S}$. Schematically, if we consider the h_α-A plane (where each axis represents an infinite dimensional space) this ''basic solution branch'' $A\equiv 0$ can be drawn as the horizontal line along the h_α axis. We shall call these solutions *vanishing asymptotic states*.

It is important to emphasize that these vanishing asymptotic states do not necessarily correspond to solutions of the complete system, Eqs. (22). In fact, it is not at all certain that if $A=0$ is substituted into Eq. (22b) the equation that results for T from Eq. (23),

$$T=\mathcal{N}(T,F_\alpha,h_\alpha) \tag{74}$$

will possess a solution in $\mathcal{T}$. Whenever it does possess such a solution for an initial perturbation h_α, we shall say that the corresponding point on the zero solution branch for the asymptotic equation, Eq. (22a), is *accessible* to the system. Clearly, at least one vanishing asymptotic state is always accessible, namely the ''origin'' $A\equiv h_\alpha\equiv 0$. Any other accessible vanishing asymptotic state (a.v.a.s. from now on) will correspond, physically, to strongly Landau damped evolution,[6] with the field damping completely to zero *before* trapping effects are able to sustain traveling-wave propagation.

Let us consider a generic a.v.a.s. $A\equiv 0$, $h_\alpha=h_\alpha^0$, associated with a transient field $T=T_0$ such that the nonlinear Ampère equation, Eq. (74) is satisfied. The corresponding distribution function $f_\alpha^{T_0}$ will be the solution of the Vlasov equation, Eq. (24), with the initial condition $f_\alpha^{T_0}(x,v,0)\equiv\mathcal{F}_\alpha^0(x,v)\equiv F_\alpha(v)+h_\alpha^0(x,v)$, where $F_\alpha(v)$ is a given Vlasov equilibrium. We now introduce the following definition:

Definition 1: An a.v.a.s. $(A,T,h_\alpha)=(0,T_0,h_\alpha^0)$ will be called *critical* if in every (arbitrarily small) neighborhood of $(0,T_0,h_\alpha^0)$ in $\mathcal{AP}\times\mathcal{T}\times\mathcal{S}$ there is a point $(A,T_0+\delta T,h_\alpha^0+\delta h_\alpha)$ such that the initial perturbation $h_\alpha^0+\delta h_\alpha$ generates a solution to the VPA problem in which the electric field has transient part $T_0+\delta T$ and *nonzero* time-asymptotic part A.

Obviously, this situation is physically very interesting, especially in the class of problems that fall under the label of ''nonlinear Landau damping;''[20] in these cases, we expect an a.v.a.s. of this kind to mark the transition between solutions that Landau damp completely and solutions that contain a nonzero small-amplitude asymptotic part. Another important example of a critical a.v.a.s. is found whenever the background equilibrium F_α allows undamped traveling wave solutions for perturbations of arbitrarily small amplitude. Examples are the recently discovered small-amplitude BGK and BGK-type solutions.[10,12] In these cases, of course, the critical a.v.a.s. is given by the ''origin'' $(A,T,h_\alpha)=(0,0,0)$.

Now, the fact that $A\equiv 0$ provides a solution to the asymptotic equation *in isolation* for any choice of h_α, suggests that the asymptotic equation itself may be amenable to a *bifurcation analysis* at the critical states. In fact, if we consider the asymptotic equation alone, in an arbitrarily small neighborhood of a critical a.v.a.s. the solution to the equation cannot be unique, since from Definition 1 both $A\equiv 0$ *and* the nonzero asymptotic solutions to the VPA system corresponding to $h_\alpha^0+\delta h_\alpha$ satisfy the equation. Thus, according to a rather general definition in nonlinear analysis (e.g., Ref. 21, p. 151), $A\equiv 0$, $h_\alpha=h_\alpha^0$ is a *bifurcation point* for the asymptotic equation. Of course, this is not true in general for the complete VPA system, for example, it is not true whenever the solution to the initial value problem is known to be unique. In summary, we have the following theorem:

Theorem 3: *Every critical a.v.a.s for the VPA initial value problem corresponds to a bifurcation point for the asymptotic Ampère equation, Eq. (22a).*

VI. TIME-ASYMPTOTIC LINEAR ANALYSIS

Let us now consider a given critical a.v.a.s. $\eta_0\equiv(0,T_0,h_\alpha)$. Instead of carrying out the bifurcation analysis of Eq. (22a) at η_0 directly, we shall study the equivalent problem comprised of the countable set of equations for the Fourier–Bohr coefficients of A,

$$a_{k,\omega_i}=\mathcal{N}_{k,\omega_i}(A+T,F_\alpha,h_\alpha). \tag{75}$$

These equations correspond to all the k and ω_i such that $a_{k,\omega_i}\neq 0$; by definition, the Fourier–Bohr coefficients $\mathcal{N}_{k,\omega_i}$ of the nonlinear operator $\mathcal{N}$ coincide with those of $P_a\mathcal{N}$. The double-indexed sequences $\mathbf{y}\equiv\{y_{m,\eta}\}$ of Fourier–Bohr coefficients of functions that are periodic in space and a.p.c. in time are characterized by the Riesz-type condition[22] $\Sigma_{m,\eta}|y_{m,\eta}|^2<\infty$, where it is understood that the index-set for η can be any countable subset of the real axis $\mathbb{R}$, and not just the integers. With the supnorm,

$$\|\mathbf{y}\|=\sup_{m,\eta}|y_{m,\eta}|, \tag{76}$$

the set of all the $\mathbf{y}$ is a Banach space, which we shall denote l_b. The sequence of nonlinear time-asymptotic equations, Eq. (75), can be reformulated as a single functional equation in l_b by writing $\mathbf{a}=\{a_{k,\omega_i}\}$ for the double sequence of the Fourier–Bohr coefficients of A and defining $N(\mathbf{a})\equiv\{\mathcal{N}_{k,\omega_i}(A+T,F_\alpha,h_\alpha)\}$. Then, Eq. (75) becomes

$$\mathbf{a}=N(\mathbf{a}). \tag{77}$$

Here, we are not explicitly writing the dependence of N on the transient field T, which is being treated as a free parameter along with F_α and h_α.

In a local analysis, the natural question is whether there is any appropriate *linear* approximation to the nonlinear problem under consideration. We answer this question by first establishing the following result:

Lemma 1: For $E\in\mathcal{W}$, $\mathcal{F}_\alpha\in\mathcal{S}$ and $k\neq 0$,

$$\begin{aligned}\mathcal{N}_{k,\omega_i}(E,F_\alpha,h_\alpha)&=\frac{2}{k}\sum_\alpha\frac{q_\alpha^2}{m_\alpha}\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma dt\int_{-\pi}^{+\pi}dx\; e^{-i\omega_i t-ikx}E(x,t)P\int_{\mathbb{R}}dv\,\frac{\mathbf{f}'_\alpha(E,\mathcal{F}_\alpha)}{\omega_i+kv}\\&=\frac{2}{k}\sum_\alpha\frac{q_\alpha^2}{m_\alpha}\lim_{n\to\infty}\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma dt\int_{-\pi}^{+\pi}dx\; e^{-i\omega_i t-ikx}E(x,t)\int_{\Omega_{n,i}^c}dv\,\frac{\mathbf{f}'_\alpha(E,\mathcal{F}_\alpha)}{\omega_i+kv},\end{aligned} \tag{78}$$

where $\mathbf{f}'_\alpha(E,\mathcal{F}_\alpha)=\partial f_\alpha/\partial v(x,v,t)$ was defined in Eq. (11) and $\Omega_{n,i}^c$ represents the real axis $\mathbb{R}$ minus the one-dimensional sphere of radius $r_n<1/n$ centered at $-\omega_i/k$, $\Omega_{n,i}\equiv B[-(\omega_i/k),r_n]$.

Proof: The Fourier–Bohr coefficients $\mathcal{N}_{k,\omega_i}$ are given by

$$\mathcal{N}_{k,\omega_i}(E,F_\alpha,h_\alpha)=\frac{4\pi}{ik}\sum_\alpha q_\alpha\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma dt\; e^{-i\omega_i t}\int_{\mathbb{R}}dv\, f_{\alpha,k}(v,t), \tag{79}$$

The properties of f_α enable us to apply Fubini's theorem and rewrite this equation as

$$\mathcal{N}_{k,\omega_i}(E,F_\alpha,h_\alpha)=\frac{4\pi}{ik}\sum_\alpha q_\alpha\lim_{\sigma\to\infty}\lim_{n\to\infty}\int_{\Omega_{n,i}^c}dv\,\frac{1}{\sigma}\int_0^\sigma dt\; e^{-i\omega_i t}f_{\alpha,k}(v,t). \tag{80}$$

We substitute into the right-hand side of Eq. (80) the expression for f_α that is given by the Vlasov equation in integro-differential Fourier-transformed form

$$f_{\alpha,k}(v,t)=h_{\alpha,k}(v)e^{-ikvt}-\frac{q_\alpha}{m_\alpha}\int_0^t d\tau\frac{1}{2\pi}\int_{-\pi}^{+\pi}dx e^{-ik[x+v(t-\tau)]}E(x,\tau)\frac{\partial f_\alpha}{\partial v}(x,v,\tau). \tag{81}$$

The term proportional to $h_{\alpha,k}(v)$ does not contribute to Eq. (80) by the Riemann–Lebesgue Lemma. After an integration by parts, we are left with

$$-\frac{2}{k}\sum_{\alpha}\frac{q_{\alpha}^{2}}{m_{\alpha}}\lim_{\sigma\to\infty}\frac{1}{\sigma}P\int_{\mathbb{R}}dv\,\frac{e^{-i\omega_i\sigma}}{\omega_i+kv}\int_0^{\sigma}dte^{-ikv(\sigma-t)}\int_{-\pi}^{+\pi}dxe^{-ikx}E(x,t)\,\frac{\partial f_\alpha}{\partial v}(x,v,t)$$

$$+\frac{2}{k}\sum_{\alpha}\frac{q_{\alpha}^{2}}{m_{\alpha}}\lim_{\sigma\to\infty}\frac{1}{\sigma}\,P\int_{\mathbb{R}}dv\int_0^{\sigma}dt\,\frac{e^{-i\omega_i t}}{\omega_i+kv}\int_{-\pi}^{+\pi}dxe^{-ikx}E(x,t)\,\frac{\partial f_\alpha}{\partial v}(x,v,t). \tag{82}$$

Here, the two v-integrals, taken separately, would not be well-defined if the principal values had not been introduced via Eq. (80). The first term in Eq. (82) can be greatly simplified by noting that it contains the last term on the right-hand side of Eq. (81) evaluated at $t=\sigma$. Hence, that whole term vanishes in the limit $\sigma\to\infty$, since $1/\sigma$ multiplies functions that are bounded in σ. Finally, via another change in the order of integration (based on Fubini's theorem, and also on the Lebesgue dominated convergence theorem) in order to bring the limit $n\to\infty$ inside the integrals in dt and dx, we are left with the expression for the $\mathcal{N}_{k,\omega_i}$ on the first line in Eq. (78).

In order to obtain the second expression in Eq. (78), we need to change the order of the limits $\sigma\to\infty$ and $n\to\infty$. This can be done most conveniently in Eq. (79) in order to take advantage of the properties of f_α. Thus, we break the v integration in Eq. (79) into two parts, on $\Omega_{n,i}$ and $\Omega_{n,i}^c$, respectively. As $n\to\infty$, the part on $\Omega_{n,i}$ vanishes, since

$$\left|\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^{\sigma}dte^{-i\omega_i t}\int_{\Omega_{n,i}}dvf_{\alpha,k}(v,t)\right|\leqslant\frac{2M_\alpha}{n} \tag{83}$$

due to the *a priori* bound in Eq. (2). Taking the limit $n\to\infty$ we obtain the following expression for $\mathcal{N}_{k,\omega_i}(E,F_\alpha,h_\alpha)$:

$$\mathcal{N}_{k,\omega_i}(E,F_\alpha,h_\alpha)=\frac{4\pi}{ik}\sum_{\alpha}q_\alpha\lim_{n\to\infty}\lim_{\sigma\to\infty}\int_{\Omega_{n,i}^c}dv\,\frac{1}{\sigma}\int_0^{\sigma}dtf_{\alpha,k}(v,t). \tag{84}$$

Then, the same procedure that was applied to Eq. (80) (with minor adaptations) leads to the expression on the second line in Eq. (78). □

The fact that the order of the limits $n\to\infty$ and $\sigma\to\infty$ in Eq. (78) can be changed is crucial for the next Lemma, which shows that the explicit dependence on E in Eq. (78) actually involves only the asymptotic part A, while the term containing T vanishes. (Of course, the transient T still appears implicitly through the distribution function f_α, which is determined by *both* A and T via the Vlasov equation.)

Lemma 2: For any choice of $E\in\mathcal{W}$, $\mathcal{F}_\alpha\in\mathcal{S}$ *and* $k\neq 0$,

$$\frac{2}{k}\sum_{\alpha}\frac{q_{\alpha}^{2}}{m_{\alpha}}\lim_{n\to\infty}\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^{\sigma}dt\;e^{-i\omega_i t}\int_{-\pi}^{+\pi}dx\;e^{-ikx}T(x,t)\int_{\Omega_{n,i}^c}dv\,\frac{\mathbf{f}'_\alpha(E,\mathcal{F}_\alpha)}{\omega_i+kv}=0. \tag{85}$$

Proof: Integrating by parts gives

$$\left|\int_{\Omega_{n,i}^c}dv\,\frac{\dfrac{\partial f_\alpha}{\partial v}(x,v,t)}{\omega_i+kv}\right|=\left|-\left[\frac{f_\alpha(x,v,t)}{\omega_i+kv}\right]_{-(\omega_i/k)-r_n}^{-(\omega_i/k)+r_n}+k\int_{\Omega_{n,i}^c}dv\,\frac{f_\alpha(x,v,t)}{(\omega_i+kv)^2}\right|$$

$$\leqslant\frac{4\pi M_\alpha}{r_n}+\frac{k}{r_n^2}\int_{\mathbb{R}}dv|f_\alpha|\leqslant\frac{4\pi M_\alpha}{r_n}+\frac{k\tilde{M}}{r_n^2}\equiv M_{\alpha,n}, \tag{86}$$

where we have exploited the *a priori* bound on $f_\alpha(x,v,t)$ and the fact that the total number of particles $\tilde{M}$ must be conserved (as follows from integrating the Vlasov equation over the whole phase plane). Hence, in the second line of Eq. (78),

$$\left| \lim_{\sigma\to\infty} \frac{1}{\sigma} \int_0^{\sigma} dt\ e^{-i\omega_i t} \int_{-\pi}^{+\pi} dx\ e^{-ikx} T(x,t) \int_{\Omega_{n,i}^c} dv\, \frac{\mathbf{f}'_\alpha(E,\mathcal{F}_\alpha)}{\omega_i + kv} \right|$$

$$\leqslant \lim_{\sigma\to\infty} \frac{1}{\sigma} \int_0^{\sigma} dt |T(x,t)| 2\pi M_{\alpha,n} = 2\pi M_{\alpha,n} \mathbf{M}_t[|T(x,t)|] = 0, \tag{87}$$

where the mean value $\mathbf{M}_t$ was defined in Eq. (17) and $\mathbf{M}_t[|T(x,t)|]$ is zero since $T(x,t)$ is a transient. Thus Lemma 2 is proved. Clearly since $M_{\alpha,n}$ diverges in the limit $n\to\infty$, the order in which the two limits are taken in Eq. (85) is crucial. □

According to Eqs. (78) and (85), when $k\neq 0$ the Fourier–Bohr coefficients $\mathcal{N}_{k,\omega_i}$ can be rewritten as

$$\mathcal{N}_{k,\omega_i}(A+T,F_\alpha,h_\alpha) = \lim_{\sigma\to\infty} \frac{1}{\sigma} \int_0^{\sigma} dt \frac{1}{2\pi} \int_{-\pi}^{+\pi} dx e^{-i\omega_i t - ikx} A \mathcal{E}^{(k,\omega_i)}(A+T,F_\alpha,h_\alpha), \tag{88}$$

where we have introduced the notation

$$\mathcal{E}^{(k,\omega_i)}(E,F_\alpha,h_\alpha) \equiv \frac{4\pi}{k} \sum_\alpha \frac{q_\alpha^2}{m_\alpha} P \int_{\mathbb{R}} dv\, \frac{\mathbf{f}'_\alpha(E,\mathcal{F}_\alpha)}{\omega_i + kv}. \tag{89}$$

We next consider $k=0$ which we excluded from Lemmas 1 and 2. There are two possibilities, $\omega_i=0$ and $\omega_i\neq 0$. However, it is not necessary to study the former case since $a_{0,0}=0$ follows immediately from the fact that the VPA system conserves energy. Indeed, if $a_{0,0}$ were nonzero the corresponding time-asymptotic electric field, being uniform in both space and time, clearly would accelerate all the particles to infinite energies. Thus, we do not consider this case; for the same reason, we exclude the possibility of $\omega=0$ being an accumulation point for the ω_i. Then, for $k=0$ there must exist a $\lambda\in\mathbb{R}$ such that $k=0$, $\omega_i\leqslant\lambda \Rightarrow a_{k,\omega_i}=0$, which in fact is a well-known sufficient condition for the spatial average $A_0(t)$ to have an almost periodic primitive (Ref. 17, p. 74). Hence, we next consider the remaining case, $k=0$, $\omega_i\neq 0$, under this condition. The analogues of Lemmas 1 and 2 are developed immediately using the same arguments employed in these Lemmas, with Eq. (81) for the kth Fourier component replaced by

$$f_{\alpha,0}(v,t) = F_\alpha(v) - \frac{q_\alpha}{m_\alpha} \int_0^t d\tau \frac{1}{2\pi} \int_{-\pi}^{+\pi} dx\ E(x,\tau) \frac{\partial f_\alpha}{\partial v}(x,v,\tau), \tag{90}$$

which is obtained via the time integration of the $k=0$ component of the Vlasov equation. Substituting Eq. (90) into Eq. (4) and carrying out steps analogous to those implemented above for $k\neq 0$ leads to an expression identical to Eq. (88), with

$$\mathcal{E}^{(0,\omega_i)}(E,F_\alpha,h_\alpha) \equiv \frac{4\pi}{\omega_i^2} \sum_\alpha \frac{q_\alpha^2}{m_\alpha} \int_{\mathbb{R}} dv\, \mathbf{f}_\alpha(E,\mathcal{F}_\alpha). \tag{91}$$

Thus, Lemmas 1 and 2 have been extended to include the case $k=0$.

Equation (88) combined with Eq. (89) for $k\neq 0$ and Eq. (91) for $k=0$ leads directly to the following theorem;

Theorem 4: *Given a critical a.v.a.s.* $\eta_0\in\mathcal{AP}\times\mathcal{T}\times\mathcal{S}$, *let* $\mathcal{E}^{(k,\omega_i)}(E,F_\alpha,h_\alpha)$ *be continuous in* E *at* η_0. *Then,* $\mathcal{N}_{k,\omega_i}(A+T,F_\alpha,h_\alpha)$ *is Fréchet-differentiable with respect to* A *at* η_0, *the derivative being*

$$\mathcal{L}_{k,\omega_i}(T_0,F_\alpha,h_\alpha^0)A \equiv \lim_{\sigma\to\infty} \frac{1}{\sigma} \int_0^{\sigma} dt \frac{1}{2\pi} \int_{-\pi}^{+\pi} dx e^{-i\omega_i t - ikx} A \mathcal{E}^{(k,\omega_i)}(T_0,F_\alpha,h_\alpha^0). \tag{92}$$

Proof: Clearly $\mathcal{L}_{k,\omega_i}$ is a continuous linear operator, and due to Eq. (88) [Lemmas 1 and 2] and the continuity of $\mathcal{E}^{(k,\omega_i)}$ at η_0,

$$
\begin{aligned}
&|\mathcal{N}_{k,\omega_i}(T_0+A,F_\alpha,h_\alpha^0)-\mathcal{N}_{k,\omega_i}(T_0,F_\alpha,h_\alpha^0)-\mathcal{L}_{k,\omega_i}(T_0,F_\alpha,h_\alpha^0)A| \\
&\quad=\left|\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma dt\,\frac{1}{2\pi}\int_{-\pi}^{+\pi}dx e^{-i\omega_i t-ikx}A[\mathcal{E}^{(k,\omega_i)}(T_0+A,F_\alpha,h_\alpha^0)-\mathcal{E}^{(k,\omega_i)}(T_0,F_\alpha,h_\alpha^0)]\right| \\
&\quad\leqslant\|A\|\|\mathcal{E}^{(k,\omega_i)}(T_0+A,F_\alpha,h_\alpha^0)-\mathcal{E}^{(k,\omega_i)}(T_0,F_\alpha,h_\alpha^0)\|=o(\|A\|).
\end{aligned}
\tag{93}
$$

□

The sequence $\mathcal{L}_{k,\omega_i}A$ defines a linear operator $L\mathbf{a}\equiv\{\mathcal{L}_{k,\omega_i}(T_0,F_\alpha,h_\alpha^0)A\}$ on l_b. Introducing $M(\mathbf{a})\equiv N(\mathbf{a})-L\mathbf{a}$ enables us to write Eq. (77) in the form

$$(I-L)\mathbf{a}=M(\mathbf{a}). \tag{94}$$

The operator $(I-L)$ represents the linear approximation (at the given critical a.v.a.s.) to the original nonlinear equation, Eq. (77). Hence, the first step in the analysis of the nonlinear problem will be the study of the invertibility of this linear operator. Specifically, we must determine the null space of $(I-L)$; hence, we must study the *linearized time-asymptotic equation*

$$(I-L)\mathbf{a}=0. \tag{95}$$

Before proceeding with the general analysis of this equation, we consider a particularly simple case for which the analysis is straightforward. It is provided by the a.v.a.s. $h_\alpha\equiv 0$, i.e., the zero field solution

$$A(x,t)=T_0(x,t)\equiv 0,\quad f_\alpha^{T_0}(x,v,t)=\mathbf{f}_\alpha(0,F_\alpha)=F_\alpha(v) \tag{96}$$

for which $\mathcal{E}^{(k,\omega_i)}(T_0,F_\alpha,h_\alpha^0)$, which appears in Eq. (92), can be calculated immediately. From Eq. (89) and Eq. (91) it is

$$\mathcal{E}^{(k,\omega_i)}(0,F_\alpha,0)\equiv\frac{4\pi}{k}\sum_\alpha\frac{q_\alpha^2}{m_\alpha}P\int_{\mathbb{R}}dv\,\frac{F_\alpha'(v)}{\omega_i+kv} \tag{97}$$

for $k\neq 0$, and

$$\mathcal{E}^{(0,\omega_i)}(0,F_\alpha,0)\equiv\frac{4\pi}{\omega_i^2}\sum_\alpha\frac{q_\alpha^2}{m_\alpha}\int_{\mathbb{R}}dv F_\alpha(v)=\frac{\omega_p^2}{\omega_i^2} \tag{98}$$

for $k=0$, where ω_p is the *plasma frequency* that corresponds to the equilibrium $F_\alpha(v)$. The expressions given in Eqs. (97) and (98) are both constants and can be moved out of the integrals in Eq. (92). Thus, Eq. (95) takes the simple form

$$a_{k,\omega_i}\mathcal{D}_0(k,\omega_i)=0, \tag{99}$$

where

$$\mathcal{D}_0(k,\omega_i)\equiv 1-\frac{4\pi}{k}\sum_\alpha\frac{q_\alpha^2}{m_\alpha}P\int_{\mathbb{R}}dv\,\frac{F_\alpha'(v)}{\omega_i+kv} \tag{100}$$

for $k\neq 0$ and $\mathcal{D}_0(0,\omega_i)\equiv\omega_p^2/\omega_i^2$ (for $k=0$). $\mathcal{D}_0(k,\omega_i)$ is the well-known *Vlasov dielectric function*,[23] and Eq. (99) is trivial to solve; it implies that $a_{k,\omega_i}=0$ except for all the choices of k and ω_i that satisfy the *Vlasov dispersion relation*

$$\mathcal{D}_0(k,\omega_i)=0. \tag{101}$$

In general the linear operator in Eq. (95) is determined by $\mathcal{E}^{(k,\omega_i)}(T_0,F_\alpha,h_\alpha^0)$ through Eq. (92), and $\mathcal{E}^{(k,\omega_i)}(T_0,F_\alpha,h_\alpha^0)$ in turn is determined solely by the initial condition and the transient field T_0 via the distribution function $f_\alpha^{T_0}$ at the a.v.a.s. However, $f_\alpha^{T_0}$ generally will not be a simple Vlasov equilibrium because the transient T_0 is nonzero and affects the distribution function. Unfortunately, it is very difficult to solve exactly both the nonlinear Ampère equation for T_0, Eq. (74), and the Vlasov equation for $f_\alpha^{T_0}$, Eq. (24). However, it is possible, without doing this, to obtain important results on the asymptotic linear operator by exploiting the general properties of particle motion in a transient field. If we define a Vlasov equilibrium $F_\alpha^{T_0}$ according to Eq. (66), Eq. (65) shows that this $F_\alpha^{T_0}$ yields the same values for macroscopic quantities as the distribution functions $f_\alpha^{T_0}$ in the time asymptotic limit. We now raise the question of whether these Vlasov equilibria $F_\alpha^{T_0}$ yield the same values as $f_\alpha^{T_0}$ for the integrals in Eqs. (89) and (91) and therefore the same linearized operator $\mathcal{L}_{k,\omega_i}$. In fact, we have the following result:

Theorem 5: *At a given critical a.v.a.s. η_0,*

$$\mathcal{L}_{k,\omega_i}(T_0,F_\alpha,h_\alpha)A=[1-\mathcal{D}(k,\omega_i)]a_{k,\omega_i}, \tag{102}$$

where

$$\mathcal{D}(k,\omega_i)\equiv 1-\frac{4\pi}{k}\sum_\alpha \frac{q_\alpha^2}{m_\alpha}P\int_{\mathbb{R}}dv\,\frac{F_\alpha^{T_0\prime}(v)}{\omega_i+kv} \tag{103}$$

for $k\neq 0$ and $\mathcal{D}(0,\omega_i)\equiv 1-\omega_p^2(T_0)/\omega_i^2$ (for $k=0$). Here $F_\alpha^{T_0\prime}$ is the velocity derivative of the time-asymptotic Vlasov equilibrium defined in Eq. (66) and $\omega_p^2(T_0)$ is the corresponding plasma frequency.

Proof: We shall consider the case $k\neq 0$; the case $k=0$ is straightforward. After substituting Eq. (60) into Eq. (92) and explicitly writing out the expression for $\mathcal{E}^{(k,\omega_i)}$, Eq. (89), we can change the order of the limits $\sigma\to\infty$ and $n\to\infty$ according to Lemma 1. Then the term containing the functions $g_\alpha^{T_0}$ vanishes by Fréchet's Lemma[16] and we are left with

$$\frac{4\pi}{k}\sum_\alpha \frac{q_\alpha^2}{m_\alpha}\lim_{n\to\infty}\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma dt\,\frac{1}{2\pi}\int_{-\pi}^{+\pi}dx\; e^{-i\omega_i t-ikx}A(x,t)\int_{\Omega_{n,i}^c}dv\,\frac{\frac{\partial}{\partial v}[\mathcal{F}_\alpha^{T_0}(x-vt,v)]}{\omega_i+kv}. \tag{104}$$

By applying the convolution theorem to the integral in *dx* this becomes

$$\frac{4\pi}{k}\sum_\alpha \frac{q_\alpha^2}{m_\alpha}\lim_{n\to\infty}\lim_{\sigma\to\infty}\frac{1}{\sigma}\int_0^\sigma dt e^{-i\omega_i t}\sum_{k'}A_{k-k'}(t)\int_{\Omega_{n,i}^c}dv\,\frac{\frac{\partial}{\partial v}[\mathcal{F}_{\alpha,k'}^{T_0}(v)e^{-ik'vt}]}{\omega_i+kv}. \tag{105}$$

It can be easily seen that the term corresponding to $k'=0$ gives Eq. (102). Hence, it will be enough to show that all the terms with $k'\neq 0$ are equal to zero in order to establish the result. An integration by parts gives

$$\int_{\Omega_{n,i}^c} dv \frac{\frac{\partial}{\partial v}[\mathcal{F}_{\alpha,k'}^{T_0}(v)e^{-ik'vt}]}{\omega_i+kv} = k\int_{\Omega_{n,i}^c} dv \frac{\mathcal{F}_{\alpha,k'}^{T_0}(v)e^{-ik'vt}}{(\omega_i+kv)^2} - \frac{1}{r_n}\left[\mathcal{F}_{\alpha,k'}^{T_0}\left(-\frac{\omega_i}{k}-r_n\right)e^{i((k'/k)\omega_i+k'r_n)t} + \mathcal{F}_{\alpha,k'}^{T_0}\left(-\frac{\omega_i}{k}+r_n\right)e^{i((k'/k)\omega_i-k'r_n)t}\right]. \tag{106}$$

Now, the first term on the right-hand side vanishes in the time-asymptotic limit by the Riemann–Lebesgue Lemma. When the other terms are substituted into Eq. (105), they generate quantities proportional to the Fourier–Bohr coefficients $a_{k-k',\tilde{\omega}_i}$, where $\tilde{\omega}_i \equiv -(1-(k'/k))\omega_i \pm k'r_n$. Since A is almost periodic, all these coefficients will be equal to zero except those corresponding to some very special choices of the radii r_n such that $\tilde{\omega}_i$ belongs to the countable set of nonzero frequencies of A. However, since the set of these ''bad radii'' has zero measure in $\mathbb{R}$, we can always pick the sequence r_n (when introducing the family of spheres in Eq. (78)) in such a way that $a_{k-k',\tilde{\omega}_i}=0\,\forall n$; thus, the limit for $n\to\infty$ in Eq. (105) also will be zero, which proves the assertion. □

The conclusion of this analysis is that the operator $(I-L)$ can be written component-wise as

$$(I-L)\mathbf{a} \equiv \{a_{k,\omega_i}\mathcal{D}(k,\omega_i)\}. \tag{107}$$

Exploiting the regularity and integrability of $F_\alpha^{T_0}$, it is easy to see that the function $\mathcal{D}(k,\omega_i)$ is bounded, i.e., there is a constant $\hat{M}$ such that

$$|\mathcal{D}(k,\omega_i)| \leqslant \hat{M}. \tag{108}$$

This implies that $I-L$ maps l_b into l_b and is continuous, since

$$\sum_{k,\omega_i} |a_{k,\omega_i}\mathcal{D}(k,\omega_i)|^2 \leqslant \hat{M}^2 \sum_{k,\omega_i} |a_{k,\omega_i}|^2 < \infty \tag{109}$$

and

$$\|(I-L)\mathbf{a}\| = \sup_{k,\omega_i} |a_{k,\omega_i}\mathcal{D}(k,\omega_i)| \leqslant \hat{M}\|\mathbf{a}\|. \tag{110}$$

The null space of $I-L$ consists of those sequences whose only nonzero entries correspond to pairs of real indexes (k,ω_i) that satisfy the *time-asymptotic Vlasov dispersion relation*

$$\mathcal{D}(k,\omega_i)=0 \tag{111}$$

which is determined by the asymptotic Vlasov equilibria $F_\alpha^{T_0}$ corresponding to the transient field T_0 at the critical a.v.a.s. under consideration.

It is important to note that in most physically relevant situations there is only a *finite* number N of such pairs. In these cases the linear operator $I-L$ has a finite-dimensional null space $\mathcal{K} \equiv \mathrm{Ker}(I-L)$, with $\dim(\mathcal{K})=N$, and $I-L$ is a *Fredholm operator*. A good example is given by the Vlasov dispersion curve for the Maxwellian (Fig. 1), whose features are quite representative of what one encounters in most applications. There is a *cutoff wave number* k_d such that for $k>k_d$ Eq. (111) has no solution at all; hence, given any wave number $k \leqslant k_d$, Eq. (111) has a finite number N_k of simple real roots $\lambda_1(k)$, $\lambda_2(k)\cdots\lambda_{N_k}(k)$. Since the basic wave number is $k=1$, there will be $\varrho=[k_d]$ admissible wave numbers before cutoff, for a total of $N=\Sigma_{k=1}^{\varrho}N_k$ possible modes.

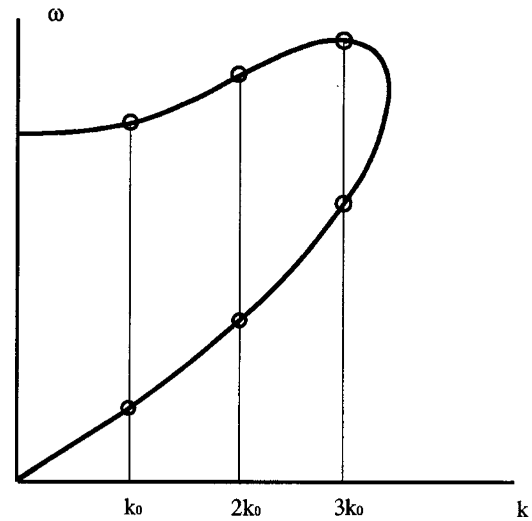

FIG. 1. The Vlasov dispersion curve for a Maxwellian $e-p$ plasma with $T_e = T_p$, and the roots $\omega(k)$ of the corresponding dispersion relation for a basic wave number k_0 and its harmonics; k is in units of the inverse Debye length $k_D \equiv 1/\lambda_D$ and ω is in units of the plasma frequency ω_p.

VII. TIME-ASYMPTOTIC FIELD SOLUTION

Now that we have carried out the analysis of the linearized time-asymptotic problem, we are ready for our final step, which is simply to show that the solution to the linearized problem provides (as should be expected) a leading order solution for A to the *nonlinear* time-asymptotic equation, Eq. (22a). Since the Fréchet derivative at the a.v.a.s. has a nontrivial null-space and is not invertible, the simplest thing to do is to apply the classical method of analysis known as the Alternative Method.[24] To do this, we decompose the space l_b into the direct sum of $\mathcal{K}$ and its complement $\mathcal{H}$. From Eq. (107) it follows immediately that $Rg(I-L) = [\text{Ker}(I-L)]^{\perp} = \mathcal{H}$. Let us consider the two projectors $Q_{\mathcal{K}}$ and $Q_{\mathcal{H}}$ associated, respectively, with $\mathcal{K}$ and $\mathcal{H}$. $Q_{\mathcal{K}}$ is the operator that starts from any element of l_b and annihilates all the entries except for those that have one of the N pairs of indexes that satisfy the Vlasov dispersion relation. Conversely, $Q_{\mathcal{H}}$ cancels only the entries that have such indexes. Let us define $\Psi \equiv Q_{\mathcal{K}} a$ and $\Phi \equiv Q_{\mathcal{H}} a$, so that $a = \Psi + \Phi$. Then, a standard procedure[24] leads from Eq. (94) to the two equations,

$$Q_{\mathcal{K}} M(\Psi + \Phi) = 0, \tag{112}$$

$$(I-L)\Phi = Q_{\mathcal{H}} M(\Psi + \Phi). \tag{113}$$

In the language of bifurcation theory, Eq. (112) is the *bifurcation equation*, while Eq. (113) is the *auxiliary equation*. It is easy to verify that when the operator $I-L$ is restricted to the subspace $\mathcal{H}$, it is invertible; from Eq. (107) it follows immediately that given $\mathbf{b} \in \mathcal{H}$,

$$(I-L)^{-1}\mathbf{b} = \left\{ \frac{b_{k,\omega_i}}{\mathcal{D}(k,\omega_i)} \right\}, \tag{114}$$

where $\mathcal{D}(k,\omega_i) \neq 0 \quad \forall b_{k,\omega_i}$ since $\mathbf{b} \in \mathcal{H}$. Then, the auxiliary equation, Eq. (113) can be written as a fixed point problem for Φ,

$$\Phi = (I-L)^{-1} Q_{\mathcal{H}} M(\Psi + \Phi). \tag{115}$$

Here, we are not interested in a detailed fixed point analysis of this equation, since we have been assuming from the beginning that an asymptotically almost periodic solution to the VPA problem does exist. In fact, the existence and uniqueness of the solution for A implies that the solutions for Φ and Ψ must also exist and be unique. What is important is that Eqs. (112) and (115) enable us to write the general form of the nonlinear solution for the time-asymptotic field A, as expressed in the following theorem:

Theorem 6: *Let $\eta \in \mathcal{AP}\times\mathcal{T}\times\mathcal{S}$ be a critical a.v.a.s., and let $\mathcal{E}^{(k,\omega_i)}(E,F_\alpha,h_\alpha)$ be continuous in E at η. Then, the general solution to the time-asymptotic equation, Eq. (22a), in a neighborhood of η in $\mathcal{AP}\times\mathcal{T}\times\mathcal{S}$ is given by*

$$A(x,t)=\sum_{k,\omega_i}\psi_{k,\omega_i}e^{ikx+i\omega_i t}+o(\|A\|), \tag{116}$$

where k and ω_i satisfy the time-asymptotic Vlasov dispersion relation, Eq. (111), and the amplitudes ψ_{k,ω_i} satisfy the bifurcation equation

$$Q_{\mathcal{K}}M(\Psi+\Phi(\Psi))=0, \tag{117}$$

where $\Phi(\Psi)=o(\|A\|)$ is determined by the auxiliary equation, Eq. (115).

Proof: In the light of the previous results, the proof reduces to establishing Eq. (116). For any pair (k,ω_i) such that $\mathcal{D}(k,\omega_i)\neq 0$, Eq. (115) gives

$$\phi_{k,\omega_i}=[\mathcal{D}(k,\omega_i)]^{-1}\mathcal{M}_{k,\omega_i}(A+T_0,F_\alpha,h_\alpha), \tag{118}$$

where

$$\mathcal{M}_{k,\omega_i}(A+T_0,F_\alpha,h_\alpha)=\mathcal{N}_{k,\omega_i}(A+T_0,F_\alpha,h_\alpha)-\mathcal{L}_{k,\omega_i}(T_0,F_\alpha,h_\alpha)A. \tag{119}$$

$\mathcal{M}_{k,\omega_i}(A+T_0,F_\alpha,h_\alpha)$ coincides with the expression on the first line in Eq. (93) (since $\mathcal{N}_{k,\omega_i}(T_0,F_\alpha,h_\alpha)=0$ at the a.v.a.s.) and

$$\begin{aligned}|\phi_{k,\omega_i}|&\leqslant|\mathcal{D}(k,\omega_i)^{-1}||\mathcal{M}_{k,\omega_i}(A+T_0,F_\alpha,h_\alpha)|\\&\leqslant|\mathcal{D}(k,\omega_i)^{-1}|\|A\|\|\mathcal{E}^{(k,\omega_i)}(T_0+A,F_\alpha,h_\alpha)-\mathcal{E}^{(k,\omega_i)}(T_0,F_\alpha,h_\alpha)\|=o(\|A\|).\end{aligned} \tag{120}$$

The decomposition $\mathbf{a}=\Psi+\Phi$ corresponds to the decomposition of the time-asymptotic field A in the form

$$A(x,t)=A_{\mathcal{K}}(x,t)+A_{\mathcal{H}}(x,t), \tag{121}$$

where

$$A_{\mathcal{K}}(x,t)\sim\sum_{k,\omega_i}\psi_{k,\omega_i}e^{ikx+i\omega_i t},\quad A_{\mathcal{H}}(x,t)\sim\sum_{k,\omega_i}\phi_{k,\omega_i}e^{ikx+i\omega_i t}. \tag{122}$$

Since these Fourier–Bohr series can be summed *uniformly* (e.g., via the Fejer–Bochner summation method) Eq. (120) implies that

$$\|A_{\mathcal{H}}\|=o(\|A\|) \tag{123}$$

and Eq. (116) follows immediately. □

As previously mentioned, in most cases of physical interest $\mathcal{K}$ is *finite-dimensional*, so that Eq. (117) reduces to a finite system of N nonlinear algebraic equations for the amplitudes ψ_{k,ω_i}.

VIII. CONCLUSION

The general solution for a small-amplitude time-asymptotic electric field near a critical a.v.a.s. is given by Eq. (116) as a superposition of undamped traveling-wave modes, whose frequencies and wave numbers satisfy the time-asymptotic Vlasov dispersion relation, Eq. (111). Remarkably, this result holds regardless of the details of the transient field T. Of course, T determines the time-asymptotic Vlasov equilibrium $F_\alpha^{T_0}$, and also the amplitudes a_{k,ω_i} via the nonlinear operator M in Eqs. (115) and (117). In this sense, the results reported here represent the first step in the study of the initial value problem for the VPA system near a critical a.v.a.s.; a complete quantitative analysis requires the study of the transient equation, Eq. (22b), and then of the bifurcation equation, Eq. (117). An example of how this can be done has recently been given by the authors in Ref. 20 (see also Ref. 14) where they derived an approximate solution for f_α from a ''transiently linearized'' Vlasov equation. That equation was solved analytically by applying Hamiltonian perturbation theory in order to determine the characteristics associated with a time-asymptotic field A of the form given by Eq. (116). Here, however, we have developed a much more general framework for the rigorous analysis of the long-time behavior of waves propagating in plasmas. This framework relies on two essential results of the analysis we have reported.

First, the fact that critical points for the VPA initial value problem are bifurcation points for the time-asymptotic equation opens an interesting new perspective on the study of nonlinear Landau damping, and in particular on the transition between initial conditions that Landau damp to a zero electric field and those that lead to a nonzero time-asymptotic field.[20]

Second, the determination of the time-asymptotic linearized operator $\mathcal{L}_{k,\omega_i}$ provides a new and more solid foundation for the nonlinear analysis of the VPA system. By computing the Fourier–Bohr coefficients $\mathcal{N}_{k,\omega_i}$ in the form of Eq. (88), we have been able, in effect, to take the limit $t \to \infty$ *before* linearizing the nonlinear Ampère equation. As a consequence, we have obtained a linear approximation to the time-asymptotic VPA problem which is very different from the traditional linear theory[6,25,26] and which is uniformly valid in the time-asymptotic limit. Indeed, the linearization is not about an initial unperturbed Vlasov equilibrium F_α, but about the *time-asymptotic* Vlasov equilibrium $F_\alpha^{T_0}$, that incorporates the cumulative effects of the transient field T_0. It is very significant that the corresponding *nonlinear* solution for the time-asymptotic electric field, Eq. (115), is formally so similar (especially in terms of the Vlasov dispersion relation) to the multiple-wave BGK-type nonlinear solutions[12] which have been traditionally regarded as the natural candidates to describe long-time plasma-wave propagation.

1. C. Bardos and P. Degond, Ann. Inst. Henri Poincarè, Analyse Nonlinéaire **2**, 101 (1985).
2. J. Cooper, Math. Methods Appl. Sci. **5**, 516 (1983).
3. R. Di Perna and P. L. Lions, C. R. Acad. Sci. Paris Sér. I Math. **307**, 655 (1988).
4. K. Pfaffelmoser, J. Diff. Eqns. **95**, 281 (1992).
5. T. M. O'Neil, Phys. Fluids **8**, 2255 (1965).
6. L. D. Landau, J. Phys. (U.S.S.R.), **10**, 25 (1946).
7. J. H. Malmberg and C. B. Wharton, Phys. Rev. Lett. **19**, 775 (1967).
8. L. Demeio and P. F. Zweifel, Phys. Fluids B **2**, 1252 (1990).
9. G. Manfredi, Phys. Rev. Lett. **79**, 2815 (1997).
10. J. P. Holloway and J. J. Dorning, Phys. Rev. A **44**, 3856 (1991).
11. M. L. Buchanan and J. J. Dorning, Phys. Rev. E **52**, 3015 (1995).
12. M. Buchanan and J. J. Dorning, Phys. Rev. Lett. **70**, 3732 (1993); Phys. Rev. E **50**, 1465 (1994).
13. I. B. Bernstein, J. M. Greene, and M. D. Kruskal, Phys. Rev. **108**, 546 (1957).
14. C. Lancellotti, Ph.D. dissertation, University of Virginia, 1998.
15. A. Wintner, Am. J. Math. **67**, 277 (1945).
16. M. Fréchet, Rev. Sci. **79**, 341 (1941).
17. A. M. Fink, *Almost Periodic Differential Equations* (Springer, New York, 1974).
18. A. B. Rechester and T. H. Stix, Phys. Rev. A **19**, 1656 (1979).
19. A. J. Klimas, J. Comput. Phys. **68**, 202 (1987).
20. C. Lancellotti and J. J. Dorning, Phys. Rev. Lett. **81**, 5137 (1998).
21. M. S. Berger, *Nonlinearity and Functional Analysis* (Academic, New York, 1977).
22. F. Riesz and B. Sz.-Nagy, *Functional Analysis* (Dover, New York, 1990).
23. A. Vlasov, J. Phys. (U.S.S.R.), **9**, 25 (1945).

[24] L. Cesari, in *Lecture Notes in Pure and Applied Mathematics*, Vol. 19 in *Nonlinear Functional Analysis and Differential Equations*, edited by L. Cesari, R. Kannan, and J. D. Schuur (Marcel Dekker, New York, 1976).
[25] N. G. Van Kampen, Physica (Amsterdam) **21**, 949 (1955).
[26] K. M. Case, Ann. Phys. **7**, 349 (1959).

Conformal behavior of the Lorentz–Dirac equation and Machian particle dynamics

Sebastiano Sonego[a)]
Università di Udine, DIC Via delle Scienze 208, 33100 Udine, Italy
and International School for Advanced Studies, Via Beirut 2-4, 34014 Trieste, Italy

(Received 25 November 1998; accepted for publication 5 March 1999)

It is shown that the self-interaction force on a pointlike electric charge in curved space-time has conformal weight -1. Motivated by this result, a conformally covariant version of the Lorentz–Dirac equation is presented, where the particle mass is treated as a position- and time-dependent quantity. This feature suggests that the underlying dynamics is Machian. © *1999 American Institute of Physics.* [S0022-2488(99)04607-1]

I. INTRODUCTION

Let us consider a point particle with electric charge e, tracing out a world line $x^\mu(\tau)$ in a four-dimensional space–time $(\mathcal{M},g_{ab})$, where τ is the particle proper time.[1] It is known, from the classic analysis of DeWitt–Brehme–Hobbs,[2,3] that the particle experiences an electromagnetic self-interaction force

$$F_a=\tfrac{2}{3}e^2d_a+\tfrac{1}{3}e^2k_a{}^bR_{bc}v^c+e^2v^b\int_{-\infty}^{\tau}\mathrm{d}\tau'\ f_{aba'}v^{a'}. \tag{1.1}$$

In this equation v^a is the four-velocity of the particle, the one-form d_a is defined as

$$d_a:=\ddot{v}_a-v_a\dot{v}_b\dot{v}^b, \tag{1.2}$$

where $(\cdots)^\bullet:=v^a\nabla_a(\cdots)$, $k_a{}^b=\delta_a{}^b+v_av^b$ is the projector onto the three-space orthogonal to v^a, R_{ab} is the Ricci tensor, and $f_{aba'}$ is a bi-tensor associated with the presence of ''tails'' in the electromagnetic field.[2,4] Thus, there are three contributions to F_a: One, the so-called von Laue force, is proportional to d_a and represents a simple covariant generalization of the usual force

$$\mathbf{F}=\tfrac{2}{3}e^2\ddot{\mathbf{v}} \tag{1.3}$$

due to the self-field of the charge.[5–7] The second term was discovered by Hobbs[3] and has no counterpart in special relativity, being proportional to the Ricci curvature of space–time. Remarkably, the Hobbs force is not associated with the acceleration $\dot{v}^a$ and vanishes in vacuum, if Einstein's field equation $G_{ab}=\kappa T_{ab}$ is assumed to hold. Finally, we have a third term that depends on the entire past history of the particle and on the property of waves of being backscattered by the space–time curvature,[2,4] summarized by the quantity $f_{aba'}$. This contribution is usually very small.[8]

The behavior of the von Laue term under a conformal transformation[9]

$$g_{ab}\rightarrow\tilde{g}_{ab}=\mathrm{e}^{-2\Phi}g_{ab} \tag{1.4}$$

has been investigated by Fulton, Rohrlich, and Witten.[10] They found that d_a changes in a rather messy way. This result seems to suggest that there is no simple conformally covariant generalization of the Lorentz–Dirac equation. In Sec. II of the present paper it is shown that, taking into

[a)]Electronic mail: sebastiano.sonego@dic.uniud.it account the Hobbs term, the ''bad behavior'' of d_a under a conformal transformation is neutralized by counterterms coming from the Ricci tensor, so that F_a transforms into $\tilde{F}_a = e^{\Phi} F_a$, i.e., F_a has conformal weight -1 (see Ref. 11, p. 447, for the definition of conformal weight). This conclusion makes the existence of a conformally covariant version of the Lorentz–Dirac equation much more plausible, because the self-interaction force no longer causes any problem, and the only obstruction is due to the presence of the particle mass in the equation of motion. This difficulty is circumvented in Sec. III by allowing the possibility that mass is position dependent. Finally, the implications of such an hypothesis on the formulation of a Machian dynamics, and the possibility that general relativity corresponds only to a particular gauge choice in a wider conformally covariant theory, are discussed in Sec. IV.

II. TRANSFORMATION OF THE SELF-INTERACTION FORCE

Let us treat separately the local contributions—the von Laue and Hobbs forces—and the nonlocal tail term.

A. Local terms

First of all, let us express the acceleration $a_a := v^b \nabla_b v_a$ in $(\mathcal{M}, g_{ab})$ in terms of the corresponding one $\tilde{a}_a = \tilde{v}^b \tilde{\nabla}_b \tilde{v}_a$ in $(\mathcal{M}, \tilde{g}_{ab})$. Requiring that $g_{ab} v^a v^b = \tilde{g}_{ab} \tilde{v}^a \tilde{v}^b = -1$, it must be $\tilde{v}^a = e^{\Phi} v^a$. Remembering that, for a generic X_a,

$$\tilde{\nabla}_b X_a = \nabla_b X_a - C^c{}_{ba} X_c, \tag{2.1}$$

where

$$C^c{}_{ba} = g_{ba} g^{cd} \nabla_d \Phi - \delta^c{}_b \nabla_a \Phi - \delta^c{}_a \nabla_b \Phi \tag{2.2}$$

(see Ref. 11, pp. 445–446), one obtains by straightforward calculations

$$\tilde{a}_a = a_a - k_a{}^b \nabla_b \Phi. \tag{2.3}$$

Next, let us compute $\ddot{v}_a$:

$$\ddot{v}_a = v^c \nabla_c a_a = v^c \nabla_c (k_a{}^b \nabla_b \Phi) + v^c \nabla_c \tilde{a}_a. \tag{2.4}$$

The two terms on the right-hand side of Eq. (2.4) are, separately,

$$\begin{aligned} v^c \nabla_c (k_a{}^b \nabla_b \Phi) &= k_a{}^b v^c \nabla_c \nabla_b \Phi + a_a v^b \nabla_b \Phi + v_a a^b \nabla_b \Phi \\ &= k_a{}^b v^c \nabla_c \nabla_b \Phi + k_a{}^c \nabla_c \Phi v^b \nabla_b \Phi \\ &\quad + v_a k^{bc} \nabla_b \Phi \nabla_c \Phi + \tilde{a}_a v^b \nabla_b \Phi + e^{-\Phi} \tilde{v}_a \tilde{a}^b \nabla_b \Phi \end{aligned} \tag{2.5}$$

and

$$v^c \nabla_c \tilde{a}_a = e^{-\Phi} \tilde{v}^b \tilde{\nabla}_b \tilde{a}_a + e^{-\Phi} \tilde{v}_a \tilde{a}^b \nabla_b \Phi - v^b \nabla_b \Phi \tilde{a}_a, \tag{2.6}$$

where Eqs. (2.1)–(2.3) have been used. Furthermore,

$$a_b a^b = k^{bc} \nabla_b \Phi \nabla_c \Phi + 2 k^{bc} \nabla_c \Phi \tilde{a}_b + k^{bc} \tilde{a}_b \tilde{a}_c, \tag{2.7}$$

where we have used the property $v^a \tilde{a}_a = \tilde{v}^a \tilde{a}_a = 0$. Placing Eqs. (2.4)–(2.7) into the definition of d_a, Eq. (1.2), we get

$$d_a = e^{-\Phi} \tilde{d}_a + v^c k_a{}^b \nabla_c \nabla_b \Phi + k_a{}^c \nabla_c \Phi v^b \nabla_b \Phi. \tag{2.8}$$

Thus, the von Laue force does not change by a simple rescaling under a conformal transformation. However, let us now consider the Hobbs term. We have

$$R_{bc}=\tilde{R}_{bc}-2\nabla_b\nabla_c\Phi-g_{bc}g^{de}\nabla_d\nabla_e\Phi-2\nabla_b\Phi\nabla_c\Phi+2g_{bc}g^{de}\nabla_d\Phi\nabla_e\Phi \tag{2.9}$$

(see Ref. 11, p. 446), and consequently

$$\tfrac{1}{2}k_a{}^bR_{bc}v^c=\tfrac{1}{2}k_a{}^b\tilde{R}_{bc}v^c-k_a{}^b\nabla_b\nabla_c\Phi v^c-k_a{}^b\nabla_b\Phi v^c\nabla_c\Phi. \tag{2.10}$$

Combining Eqs. (2.8) and (2.10) we obtain finally

$$d_a+\tfrac{1}{2}k_a{}^bR_{bc}v^c=\mathrm{e}^{-\Phi}(\tilde{d}_a+\tfrac{1}{2}\tilde{k}_a{}^b\tilde{R}_{bc}\tilde{v}^c). \tag{2.11}$$

Therefore, under a conformal transformation the local part of F_a, i.e., the *sum* of the von Laue and Hobbs forces, is just rescaled.

B. Nonlocal term

Let us first remember that $f_{aba'}=\nabla_b v_{aa'}-\nabla_a v_{ba'}$, where $v_{aa'}$ is a bi-tensor satisfying the wave equation[2]

$$g^{bc}\nabla_b\nabla_c v_{aa'}-R_a{}^b v_{ba'}=0, \tag{2.12}$$

together with the Lorentz gauge condition $\nabla^a v_{aa'}=0$. Equation (2.12) can be regarded as a particular case of the field equation for the vector potential A_a,

$$g^{bc}\nabla_b\nabla_c A_a-R_a{}^bA_b-\nabla_a\nabla^bA_b=-4\pi J_a, \tag{2.13}$$

when there are no sources (i.e., $J_a=0$) and $\nabla^aA_a=0$. By an easy but tedious calculation one can check that Eq. (2.13) is conformally invariant, i.e., it holds in $(\mathcal{M},g_{ab})$ iff the corresponding equation

$$\tilde{g}^{bc}\tilde{\nabla}_b\tilde{\nabla}_cA_a-\tilde{R}_a{}^bA_b-\tilde{\nabla}_a\tilde{\nabla}^bA_b=-4\pi\tilde{J}_a \tag{2.14}$$

holds in $(\mathcal{M},\tilde{g}_{ab})$, provided that one defines $\tilde{J}_a:=\mathrm{e}^{2\Phi}J_a$.[12] (This is, of course, related to the conformal invariance of the Maxwell equations for the electromagnetic field $F_{ab}=\nabla_aA_b-\nabla_bA_a$.[11,12]) For Eq. (2.12), then, one has that

$$\tilde{g}^{bc}\tilde{\nabla}_b\tilde{\nabla}_c v_{aa'}-\tilde{R}_a{}^b v_{ba'}-\tilde{\nabla}_a\tilde{\nabla}^b v_{ba'}=0 \tag{2.15}$$

and

$$\tilde{\nabla}^a v_{aa'}=-\tilde{g}^{bc}C^a{}_{bc}v_{aa'}=-2\tilde{\nabla}^a\Phi v_{aa'}. \tag{2.16}$$

Choosing $v_{aa'}=\tilde{v}_{aa'}+\nabla_a\Lambda_{a'}$, where $\Lambda_{a'}$ satisfies the condition

$$\tilde{g}^{ab}\tilde{\nabla}_a\tilde{\nabla}_b\Lambda_{a'}=-2\tilde{\nabla}^a\Phi v_{aa'}, \tag{2.17}$$

and replacing into Eq. (2.15), we find that

$$\tilde{g}^{bc}\tilde{\nabla}_b\tilde{\nabla}_c\tilde{v}_{aa'}-\tilde{R}_a{}^b\tilde{v}_{ba'}=0. \tag{2.18}$$

Since $\tilde{f}_{aba'}=\tilde{\nabla}_b\tilde{v}_{aa'}-\tilde{\nabla}_a\tilde{v}_{ba'}=f_{aba'}$, we have for the tail term,

$$e^2\tilde{v}^b\int_{-\infty}^{\tilde{\tau}}\mathrm{d}\tilde{\tau}'\,\tilde{f}_{aba'}\tilde{v}^{a'}=\mathrm{e}^{\Phi}e^2v^b\int_{-\infty}^{\tau}\mathrm{d}\tau'\,f_{aba'}v^{a'}, \tag{2.19}$$

where we have used the property $\mathrm{d}\tilde{\tau}=\mathrm{e}^{-\Phi}\,\mathrm{d}\tau$, which follows from (1.4).

III. CONFORMALLY COVARIANT LORENTZ–DIRAC EQUATION

Putting together Eqs. (2.11) and (2.19), one finds that $\tilde{F}_a=\mathrm{e}^{\Phi}F_a$. This result should not be surprising, because one can write $F_a=eF_{ab}^{(\mathrm{self})}v^b$, where $F_{ab}^{(\mathrm{self})}$ is the self-field of the particle evaluated at its position, after mass renormalization has been performed.[2] Since Maxwell equations are conformally covariant, one expects that $\tilde{F}_{ab}^{(\mathrm{self})}=F_{ab}^{(\mathrm{self})}$, so that $\tilde{F}_a=eF_{ab}^{(\mathrm{self})}\tilde{v}^b=\mathrm{e}^{\Phi}F_a$, as we have indeed just shown.

Far from representing a small general-relativistic effect, the Hobbs term is then conceptually very important, as it guarantees a ''good behavior'' of F_a under conformal transformations, which the von Laue term alone has not. In a conformally covariant perspective, it is improper to regard F_a as composed of three terms, because two of them mix under a conformal transformation. There are then just two contributions: The local one, given by the sum of the von Laue and the Hobbs forces, and the nonlocal integral.

In all the previous calculations it has been assumed that $\tilde{e}=e$. This hypothesis simultaneously guarantees the ''good behavior'' of the radiation reaction under conformal transformations, and the conformal covariance of the inhomogeneous Maxwell equation $\nabla^bF_{ba}=-4\pi J_a$. It seems to suggest a topological origin for charge (see, e.g., Ref. 13).

The properties of the Lorentz–Dirac equation[5,6]

$$ma_a=eF_{ab}v^b+F_a \tag{3.1}$$

under the conformal transformation (1.4) have been investigated in Ref. 10, where it was concluded that Eq. (3.1) is not conformally covariant, unless one rewrites it in terms of tensor densities. Our analysis suggests that such result, based on the exclusion of the Hobbs term, should be reconsidered. However, since

$$m\tilde{a}_a=-mk_a{}^b\nabla_b\Phi+\mathrm{e}^{-\Phi}(eF_{ab}\tilde{v}^b+\tilde{F}_a), \tag{3.2}$$

it seems that even taking into account the full expression of the radiation reaction, Eq. (3.1) still fails to be conformally covariant. This conclusion rests upon an implicit hypothesis, though—that the mass m do not change under a conformal transformation. If we drop this assumption,[10] and require instead that

$$\tilde{m}=\mathrm{e}^{\Phi}m, \tag{3.3}$$

then we must rewrite Eq. (3.2) in terms of the proper time derivative of momentum $\tilde{p}_a=\tilde{m}\tilde{v}_a$, namely,

$$\tilde{v}^b\tilde{\nabla}_b\tilde{p}_a=-\tilde{\nabla}_a\tilde{m}+eF_{ab}\tilde{v}^b+\tilde{F}_a, \tag{3.4}$$

where we have used the property $\tilde{k}_a{}^b=k_a{}^b$. Since, by Eq. (2.3),

$$\tilde{v}^b\tilde{\nabla}_b\tilde{p}_a+\tilde{\nabla}_a\tilde{m}=\mathrm{e}^{\Phi}(v^b\nabla_bp_a+\nabla_am), \tag{3.5}$$

we now have that the modified Lorentz–Dirac equation

$$v^b\nabla_bp_a=-\nabla_am+eF_{ab}v^b+F_a \tag{3.6}$$

is conformally covariant.

IV. DISCUSSION AND OUTLOOKS

The conformal covariance of Eq. (3.6) relies on the transformation law (3.3) for mass, that might seem odd at first sight. However, one may convincingly argue that Eq. (3.3) must be correct on heuristic grounds, assuming that at least part of the mass is of electromagnetic origin.[5–7] Of course, for a classical point charge the electromagnetic mass is actually divergent, but one can nevertheless express it formally as $m_{\rm em}=\lim_{\epsilon\to 0} e^2/2\epsilon$, where ϵ is the radius of a three-dimensional ball centered on the particle, measured in the comoving frame.[2] Then $\tilde{m}_{\rm em} =\lim_{\tilde{\epsilon}\to 0} e^2/2\tilde{\epsilon}={\rm e}^{\Phi} m_{\rm em}$, with Φ evaluated at the particle position. More generally, the energy density u of the electromagnetic field in any given reference frame transforms as $u\to\tilde{u}={\rm e}^{4\Phi}u$, as one can easily see from the expression $u=g^{ab}(E_aE_b+B_aB_b)/8\pi$, where E_a and B_a are the electric and magnetic fields, and recalling the transformation laws of E_a and B_a.[12] Since a three-dimensional volume element is rescaled by the factor ${\rm e}^{-3\Phi}$ under the conformal transformation (1.4), it follows that the mass–energy of the electromagnetic field contained into a small region changes according to Eq. (3.3). It is then natural to require that *any* mass, independent of its nature, transform in the same way. It should also be pointed out that the transformation law (3.3) guarantees that the Hamilton–Jacobi and Klein–Gordon equations for free massive particles are conformally covariant.[14]

The conformally covariant generalization (3.6) of the Lorentz–Dirac equation contains the force $-\nabla_a m$, which is however not observed in Nature. This problem is not as serious as it may look, because one can always ''gauge away'' such a term by a suitable choice of Φ. In order to have a better understanding of this point, let us consider a neutral particle (since the issue is clearly unrelated to the presence of a charge) with equation of motion

$$\tilde{v}^b\tilde{\nabla}_b\tilde{p}_a=-\tilde{\nabla}_a\tilde{m} \tag{4.1}$$

in the space–time $(\mathcal{M},\tilde{g}_{ab})$. If we choose Φ such that ${\rm e}^{-\Phi}\tilde{m}=$const we have, in a space–time $(\mathcal{M},g_{ab})$ with $g_{ab}={\rm e}^{2\Phi}\tilde{g}_{ab}$, that $g^{ab}p_ap_b=$const and $v^b\nabla_b p_a=0$, which is the usual description of a free neutral particle given by general relativity. The latter can thus be regarded as a particular gauge choice within a wider conformally covariant framework.

One can also take a different standpoint. Suppose that the geometry of space–time is $\tilde{g}_{ab}$ and that the particle mass $\tilde{m}$ is point dependent. A natural choice of units in which to express space and time measurements is then, e.g., $\tilde{L}=e^2/\tilde{m}c^2$ (the so-called classical radius of the particle) and $\tilde{T}=e^2/\tilde{m}c^3$, where c is the speed of light. Other choices will lead to the same dependence of $\tilde{L}$ and $\tilde{T}$ on $\tilde{m}$. But then, the *measured* space and time intervals will not agree with those computed on the basis of the metric $\tilde{g}_{ab}$, essentially because the units $\tilde{L}$ and $\tilde{T}$ are not constant with respect to $\tilde{g}_{ab}$, due to their dependence on $\tilde{m}$. However, one can introduce a *phenomenological* metric g_{ab}, which is so defined that the units $\tilde{L}$ and $\tilde{T}$ are constant with respect to it. Of course, such a metric is just $g_{ab}={\rm e}^{2\Phi}\tilde{g}_{ab}$, with en${}^{-\Phi}=$const$/\tilde{m}$. It is not hard to realize that this argument is completely general, in the sense that it can be extended to units based, e.g., on the atomic size and lifetimes, and can be used to argue that the metric adopted in general relativity might not have a fundamental status.[15]

The physical meaning of the force $-\tilde{\nabla}_a\tilde{m}$ in $(\mathcal{M},\tilde{g}_{ab})$ can be found by rewriting the equation of motion (4.1) in the equivalent form

$$\tilde{m}\tilde{a}_a=-k_a{}^b\tilde{\nabla}_b\tilde{m}. \tag{4.2}$$

In the Newtonian limit, Eq. (4.2) gives $\tilde{m}\tilde{\mathbf{a}}=-\tilde{\boldsymbol{\nabla}}(\tilde{m}c^2)$, which shows that one can identify $-\tilde{\nabla}_a\tilde{m}$ with a covariant generalization of the Newtonian gravitational force, and $\Phi=c^2\ln(\tilde{m}/{\rm const})$ with the corresponding gravitational potential. Remarkably, an equation like (3.3), with Φ the gravitational potential, was suggested by Nordström already in 1913, within the context of a conformally flat theory of gravity.[16] The present discussion highlights the fact that Einstein's choice of fully incorporating gravity into the metrical structure g_{ab} of space–time is only one among infinitely many possible ''gauges'' that are selected by fixing the conformal factor.

In the space–time $(\mathcal{M},\tilde{g}_{ab})$ gravity is identified with the effect of a position- and time-dependent inertia. Actually, one might well dispense with the notion of gravity, and think only in terms of a mass field. This point of view leads to some conceptual advantages. For example, it allows one to give a unified description of dynamics for massive and massless particles. Multiplying both sides of Eq. (4.1) by $\tilde{m}$ we obtain $\tilde{p}^b\tilde{\nabla}_b\tilde{p}_a=-\tilde{\nabla}_a(\tilde{m}^2/2)$. Thus, massive particles are affected by the force $-\tilde{\nabla}_a\tilde{m}$, while massless ones simply move on null geodesics. In the present context, gravity appears as just one manifestation of mass, therefore massless particles are not affected by it. Furthermore, since Φ depends on the distribution of matter (as we know from the Newtonian limit), so does $\tilde{m}$, because of Eq. (3.3). In other words, the inertia of a particle in $(\mathcal{M},\tilde{g}_{ab})$ depends on the arrangement of the other bodies; thus, particle dynamics in $(\mathcal{M},\tilde{g}_{ab})$ is Machian.[17] From the knowledge of the field equation for Φ and using Eq. (3.3), one could write down a field equation for $\tilde{m}$ (as, for example, in the Hoyle–Narlikar theory[18,19]), but this investigation is beyond the scope of the present article, so let us limit ourselves to a few qualitative remarks.

There are essentially two contributions to the mass $\tilde{m}$ of a particle. One is due to local matter, such as, e.g., a nearby planet or a star. This is usually very small—a fraction of the order of Φ/c^2 of the total mass—but, since matter is not distributed homogeneously, it leads to a position-dependent $\tilde{m}$ and is responsible for gravitational forces. The other comes from the large-scale distribution of matter in the universe, which gives a constant Φ (at least, constant over scales on which particle motion is usually analyzed) that has a much larger magnitude than the one due to local matter, because of the enormous number of sources. This accounts for a bulk, constant part of $\tilde{m}$—what is commonly meant by ''inertia'' of the particle. Thus, one would expect that the contribution to Φ/c^2 due to cosmological matter be of order 1, i.e., that $GM/Rc^2\sim 1$, where G, M, and R are Newton's gravitational constant, the mass, and the radius of the visible universe, the latter two quantities being defined as $M\sim\rho R^3$ and $R\sim c/H$, with ρ and H the values of the mean density and of the Hubble parameter, respectively. This relation, when rewritten as $G\rho/H^2\sim 1$, expresses essentially the so-called ''flatness problem,''[19] that seems therefore to find a natural explanation in a Machian framework.

ACKNOWLEDGMENTS

It is pleasure to thank Professor M. A. Abramowicz for hospitality at Chalmers University, Göteborg, where this work was completed, and an anonymous referee for suggesting some improvements in the presentation.

[1] Latin letters a, b,... are used as abstract indices (see, e.g., Ref. 11, pp. 24–25), which just indicate the tensorial nature of an object without requiring the specification of a coordinate system; greek letters μ, ν,... run from 0 to 3 and denote components in some chart. I choose +2 as signature of the metric and work in units in which $c=1$, except in the last section, where c will occasionally be reinserted for clarity. The conventions for the curvature tensors are those of Ref. 11; notice that R_{ab} then has the opposite sign than in Refs. 2 and 3.

[2] B. S. DeWitt and R. W. Brehme, ''Radiation damping in a gravitational field,'' Ann. Phys. (N.Y.) **9**, 220–259 (1960).

[3] J. M. Hobbs, ''A vierbein formalism of radiation damping,'' Ann. Phys. (N.Y.) **47**, 141–165 (1968).

[4] S. Sonego and V. Faraoni, ''Huygens' principle and characteristic propagation property for waves in curved spacetimes,'' J. Math. Phys. **33**, 625–632 (1992); V. Faraoni and S. Sonego, ''On the tail problem in cosmology,'' Phys. Lett. A **170**, 413–420 (1992); L. Bombelli and S. Sonego, ''Relationships between various characterizations of wave tails,'' J. Phys. A **27**, 7177–7199 (1994); B. C. Nolan, ''A characterization of strong wave tails in curved spacetimes,'' Class. Quantum Grav. **14**, 1295–1308 (1997).

[5] F. Rohrlich, *Classical Charged Particles* (Addison–Wesley, Reading, MA, 1990).

[6] F. Rohrlich, ''The electron: Development of the first elementary particle theory,'' in *The Physicist's Conception of Nature*, edited by J. Mehra (Reidel, Dordrecht, 1973), pp. 331–369; F. Rohrlich, ''The dynamics of a charged sphere and the electron,'' Am. J. Phys. **65**, 1051–1056 (1997).

[7] J. D. Jackson, *Classical Electrodynamics*, 2nd ed. (Wiley, New York, 1975).

[8] C. Morette DeWitt and B. S. DeWitt, ''Falling charges,'' Physica (Amsterdam) **1**, 3–20 (1964).

[9] The conformal factor is written here in exponential form because this simplifies many expressions; the conversion to the more common notation $\tilde{g}_{ab}=\Omega^2 g_{ab}$ is, of course, $\Phi=-\ln\Omega$. Hereafter, indices of geometrical objects pertaining to the

space–time $(\mathcal{M},\tilde{g}_{ab})$ are lowered and raised using $\tilde{g}_{ab}$ and its inverse $\tilde{g}^{ab}$. Thus we have, for example, $\tilde{X}^a := \tilde{g}^{ab}\tilde{X}_b$.

[10] T. Fulton, F. Rohrlich, and L. Witten, ''Conformal invariance in physics,'' Rev. Mod. Phys. **34**, 442–457 (1962).

[11] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984).

[12] S. Sonego and M. A. Abramowicz, ''Maxwell equations and the optical geometry,'' J. Math. Phys. **39**, 3158–3166 (1998).

[13] C. W. Misner and J. A. Wheeler, ''Classical physics as geometry,'' Ann. Phys. (N.Y.) **2**, 525–603 (1957); J. A. Wheeler, ''Superspace and the nature of quantum geometrodynamics,'' in *Battelle Rencontres 1967*, edited by C. M. DeWitt and J. A. Wheeler (New York, Benjamin, 1968), pp. 242–307.

[14] Of course, the Klein–Gordon equation is conformally covariant only if it contains a suitable term, proportional to the scalar curvature (see, e.g., Ref. 11, pp. 447–448). The presence of such a term can be justified by theoretical arguments that are not directly related to the issue of conformal transformations; see S. Sonego and V. Faraoni, ''Coupling to the curvature for a scalar field from the equivalence principle,'' Class. Quantum Grav. **10**, 1185–1187 (1993); A. A. Grib and E. A. Poberii, ''On the difference between conformal and minimal couplings in general relativity,'' Helv. Phys. Acta **68**, 380–395 (1995).

[15] S. Sonego and M. A. Abramowicz, '' 'Hidden geometry' interpretation of general relativity'' (unpublished).

[16] G. Nordström, ''Träge und schwere Masse in der Relativitätsmechanik,'' Ann. Phys. (Leipzig) **40**, 856–878 (1913). For recent accounts of Nordström's theory, see: A. Pais, '*Subtle is the Lord...*' (Oxford University Press, Oxford, 1982), pp. 232–236; C. H. Brans, in *On Einstein's Path*, edited by A. Harvey (Springer, Berlin), preprint gr-qc/9705069.

[17] For updated formulations and discussions of Mach's principle, see *Mach's Principle—From Newton's Bucket to Quantum Gravity*, edited by J. Barbour and H. Pfister (Birkhauser, Boston, 1995); H. Bondi and J. Samuel, ''The Lense-Thirring effect and Mach's principle,'' Phys. Lett. A **228**, 121–126 (1997).

[18] F. Hoyle and J. V. Narlikar, ''A new theory of gravitation,'' Proc. R. Soc. London, Ser. A **282**, 191–207 (1964); F. Hoyle and J. V. Narlikar, ''A conformal theory of gravitation,'' *ibid.* **294**, 138–148 (1966).

[19] J. V. Narlikar, *Introduction to Cosmology*, 2nd ed. (Cambridge University Press, Cambridge, 1993).

Soliton stability in a $Z(2)$ field theory

J. J. P. Veerman
Laboratório de Física Teórica e Computacional, Departamento de Física, Universidade Federal de Pernambuco, 50670-901 Recife, Pernambuco, Brazil

D. Bazeia
Center for Theoretical Physics, Laboratory for Nuclear Science and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307 and Departamento de Física, Universidade Federal da Paraíba, Caixa Postal 5008, 58051-970 João Pessoa, Paraíba, Brazil

Fernando Moraes
Laboratório de Física Teórica e Cuputacional, Departamento de Física, Universidade Federal de Pernambuco, 50670-901 Recife, Pernambuco, Brazil

(Received 6 October 1998; accepted for publication 20 January 1999)

We investigate the stability of the coupled soliton solutions of a two-component $Z(2)$ vector field model, in contraposition to similar solutions of a $Z(2)\times Z(2)$ model recently introduced. We demonstrate that the coupled soliton solutions of the $Z(2)$ model are classically unstable. © *1999 American Institute of Physics.* [S0022-2488(99)01107-X]

$Z(2)$ field theoretical models play a very important role in condensed matter physics. They have been used to describe a wide range of physical systems exhibiting phase transitions involving break of $Z(2)$ symmetry. Examples[1] of such systems are uniaxial antiferromagnets like Rb_2NiF_4 or K_2MnF_4, systems presenting order–disorder transitions on bipartite lattices like in β-brass, or liquid–gas transitions, etc. On the other hand, topological defects may be generated in phase transitions involving broken symmetry. They are low-energy, spatially localized, solutions of the field equations, which are topologically stable. They are of fundamental importance for a variety of physical phenomena in the systems where they appear. As an example of their importance, we mention the quasi-one-dimensional organic system *trans*-polyacetylene. The relevant topological defect here, the soliton, is responsible for a tremendous increase in the conductivity to almost metallic levels of this insulator when charged solitons are introduced by doping.[2]

In this work we are interested in double soliton solutions for coupled scalar fields in two-dimensional space–time. Such solutions have been recently investigated in a class of systems defined by a very specific potential.[3–5] These works have shown that there are solutions of the second order equations of motion that are also solutions of some first order differential equations. Also, the important issue of the stability of the soliton solutions has been addressed.[4] It was found that the soliton solutions of those systems, if they exist, are intrinsically stable when they also satisfy the first order equations. This is also important for condensed matter systems. For instance, there is evidence that solitons in coupled scalar field theories may be important to describe ferroelectric crystals[5] and hydrogen-bonded chains.[6] And it is known that stable solutions play relevant role at the quantum level.

Other issues concerning stability of the soliton solutions for coupled scalar fields have recently been considered in Ref. 7, for the class of systems introduced in Refs. 3–5. As we know, however, in the past a $Z(2)$ coupled scalar field model was shown to present very similar coupled soliton solutions.[8] Furthermore, the motivations presented in that work are closely related to the basic motivations introduced in more recent work.[3–5] For this reason, because that $Z(2)$ model and its soliton solutions are closely related to the models introduced in Refs. 3–5, it seems important to investigate the classical or linear stability of the soliton solutions found in Ref. 8 in order to identify possible distinctions in these two approaches. This is our main motivation, and here we present a detailed analysis of the stability of the coupled soliton solitons found in Ref. 8. Although the problem seems to be hard to handle, here we offer a simple and general way of investigating the issue of stability. Unfortunately, however, we conclude that the solutions found in Ref. 8 are always unstable.

We start with the Lagrangian density[8]

$$\mathcal{L}=\tfrac{1}{2}\partial_\alpha\phi\partial^\alpha\phi+\tfrac{1}{2}\partial_\alpha\chi\partial^\alpha\chi-U(\phi,\chi), \tag{1}$$

where the potential is

$$U(\phi,\chi)=\lambda\phi^4+\lambda\chi^4+2\lambda\phi^2\chi^2-(\mu+\nu)\phi^2-(\mu-\nu)\chi^2-\gamma. \tag{2}$$

The gradient of the potential with respect to the fields is given by

$$\nabla_{\phi,\chi}U=\begin{pmatrix}-2(\mu+\nu)\phi+4\lambda\phi\chi^2+4\lambda\phi^3\\ -2(\mu-\nu)\chi+4\lambda\phi^2\chi+4\lambda\chi^3\end{pmatrix}. \tag{3}$$

The potential has stationary points at $(\phi,\chi)=(\pm\sqrt{(\mu+\nu)/2\lambda},0)$. These points are (nondegenerate) minima when the Hessian of the potential is definite positive. The Hessian is given by

$$\text{Hess}\,U=\begin{pmatrix}12\lambda\phi^2+4\lambda\chi^2-2(\mu+\nu) & 8\lambda\phi\chi\\ 8\lambda\phi\chi & 12\lambda\chi^2+4\lambda\phi^2-2(\mu-\nu)\end{pmatrix}. \tag{4}$$

Substituting the value of the stationary points, we see that

$$\text{Hess}\,U=\begin{pmatrix}4(\mu+\nu) & 0\\ 0 & 4\nu\end{pmatrix}. \tag{5}$$

Thus, in order for the stationary points above to be minima of the potential, we require

$$\nu>0,\quad \mu>-\nu. \tag{6}$$

The equations of motion corresponding to the Lagrangian are obtained as the Euler–Lagrange equations of the functional $\int\mathcal{L}$. We are looking for static soliton solutions, and may thus neglect time. The functional becomes

$$E_c[\phi,\chi]=\int dx\left\{\frac{1}{2}\left(\frac{\partial\phi}{\partial x}\right)^2+\frac{1}{2}\left(\frac{\partial\chi}{\partial x}\right)^2+U(\phi,\chi)\right\}. \tag{7}$$

The Euler–Lagrange equations are [we write Δ_x for (d^2/dx^2)]

$$-\Delta_x\begin{pmatrix}\phi\\ \chi\end{pmatrix}-\nabla_{\phi,\chi}U=0. \tag{8}$$

The soliton solutions connect the two minima of the potential at $(\phi,\chi)=(\pm\sqrt{(\mu+\nu)/2\lambda},0)$. There are two sets of static soliton solutions,

$$\bar{\phi}=\pm\sqrt{\frac{\mu+\nu}{2\lambda}}\tanh\sqrt{\mu+\nu}\,x, \tag{9}$$

$$\bar{\chi}=0, \tag{10}$$

which together with Eq. (6) requires that

$$\nu>0,\quad \mu+\nu>0,\quad \lambda>0. \tag{11}$$

The second set of solutions is given by

$$\overline{\phi} = \pm \sqrt{\frac{\mu+\nu}{2\lambda}} \tanh \sqrt{4\nu} x, \tag{12}$$

$$\overline{\chi} = \pm \sqrt{\frac{\mu-3\nu}{2\lambda}} \operatorname{sech} \sqrt{4\nu} x, \tag{13}$$

implying in this case [with Eq. (6)]

$$\nu > 0, \quad \lambda > 0, \quad \mu - 3\nu > 0. \tag{14}$$

The first pair of solutions can be investigated easily. The calculations follow the same steps already introduced in Ref. 3 for the related pair of solutions. Therefore, here we will focus attention on the stability analysis of the coupled solitons of the second solution set.

Classical stability may be discussed in the following way: If we are to have stable solitons, the second variation of $E_c[\phi,\chi]$ evaluated at the solution should be a positive differential operator. We obtain thus

$$\mathrm{Hess}(E_c[\phi,\chi]) = \begin{pmatrix} -\Delta_x & 0 \\ 0 & -\Delta_x \end{pmatrix} + \mathrm{Hess}\, U, \tag{15}$$

as can be most easily seen from Eq. (8) and noting that Δ is linear. We will call this operator $\hat{S}$. Its lowest eigenvalue will be denoted by $E_0(\mu,\nu,\lambda)$. We will show that

$$E_0(\mu,\nu,\lambda) < 0. \tag{16}$$

This way we establish that the soliton solutions are always unstable.

For the second solution pair we get the Hessian,

$$\mathrm{Hess}(E_c[\phi,\chi]) = \begin{pmatrix} 12\lambda \overline{\phi}^2 + 4\lambda \overline{\chi}^2 - 2(\mu+\nu) & 8\lambda \overline{\phi}\,\overline{\chi} \\ 8\lambda \overline{\phi}\,\overline{\chi} & 12\lambda \overline{\chi}^2 + 4\lambda \overline{\phi}^2 - 2(\mu-\nu) \end{pmatrix}. \tag{17}$$

In order to decouple the corresponding eigenvalue equations we need to diagonalize the above matrix. After some algebra we find for its eigenvalues

$$\begin{aligned} V_{\pm} &= 2\mu - 12\nu + 16\nu f \pm 2\sqrt{16\nu^2 f^2 + 4\nu(\mu-5\nu) f + (\mu-2\nu)^2} \\ &= 2\nu\{\delta - 6 + 8f \pm \sqrt{16 f^2 + 4(\delta-5) f + (\delta-2)^2}\}, \end{aligned} \tag{18}$$

where $f = f(x)$ stands for $\tanh^2(\sqrt{4\nu x})$ and $\delta = \mu/\nu$. Notice that $f(x)$ varies in [0,1), and $\nu > 0$ and $\delta > 3$ are parameters [see Eq. (14)]. The operator $\hat{S}$ is now

$$\hat{S} = \begin{pmatrix} -\Delta_x + V_+ & 0 \\ 0 & -\Delta_x + V_- \end{pmatrix}. \tag{19}$$

Notice that $V_{\pm}$ now only depends on δ and ν. To eliminate the dependence on ν, write

$$V_{\pm}(x) = 4\nu U_{\pm}(\sqrt{4\nu} x),$$

and in the operator substitute $x = y/\sqrt{4\nu}$. It is easy to see that $\hat{S}$ now becomes

$$\hat{S} = 4\nu \begin{pmatrix} -\Delta_y + U_+(y) & 0 \\ 0 & -\Delta_y + U_-(y) \end{pmatrix}. \tag{20}$$

Now restrict attention to U_- and drop the subscript. We write

$$U_\delta = \tfrac{1}{2}\delta - 3 + 4f - \tfrac{1}{2}\sqrt{16f^2 + 4(\delta-5)f + (\delta-2)^2}, \tag{21}$$

where $f = \tanh^2(y)$ and $\delta > 3$.

We wish to derive an upper estimate for the lowest eigenvalue of the equation

$$(-\Delta_y + U_\delta(y))\psi(y) = \epsilon(\delta)\psi(y). \tag{22}$$

Notice that $U_\delta(-\infty) = U_\delta(\infty) = 1$ and that U_δ is well-shaped. It is known that in the one-dimensional case there is *always* at least a bound eigenstate.[9] That is, there is an eigenfunction ψ with associated eigenvalue less than 1, and with the property that $\int \psi^* \psi dy = 1$. Our estimate relies on the following observation. Let U_1 and U_2 be two potentials as above, but with the property that for all y: $U_1(y) \leqslant U_2(y)$. The associated eigenvalues, λ_1 and λ_2 then satisfy the same relation, $\lambda_1 \leqslant \lambda_2$.

For $\delta > 3$, the potential $U_\delta(y)$ is a (weakly) decreasing function of δ,

$$\frac{\partial U_\delta}{\partial \delta} \leqslant 0.$$

It then follows that if we denote by λ_δ the lowest eigenvalue associated with U_δ,

$$\lambda_\delta \leqslant \lambda_3, \tag{23}$$

where λ_3 stands for the case $\delta = 3$. It is an easy calculation to show that

$$U_3(y) = 1 - \frac{2}{\cosh^2(y)}. \tag{24}$$

The corresponding eigenvalue equation

$$(-\Delta_y + U_3(y))\psi(y) = \epsilon(3)\psi(y) \tag{25}$$

is easily solvable (see Refs. 3 and 10). The calculations lead to only one bound state at $\epsilon_0(3) = 0$ and a continuous spectrum $\epsilon_c(3) > 1$. Therefore $\lambda_3 = 0$ and by (23)

$$\lambda_\delta < 0. \tag{26}$$

In fact, we can show that we have here a strict inequality. Notice that $H_\delta = -\Delta_y + U_\delta$ implies that $H_\delta = H_3 + (U_\delta - U_3)$. Now take ψ_0 to be the groundstate eigenfunction of H_3. It follows that

$$\int_{-\infty}^{+\infty} \psi_0^*(y) H_\delta \psi_0(y) dy = \int_{-\infty}^{+\infty} \psi_0^*(y) H_3 \psi_0(y) dy + \int_{-\infty}^{+\infty} \psi_0^*(y)[U_\delta(y) - U_3(y)]\psi_0(y) dy.$$

Since $(U_\delta(y) - U_3(y)) \leqslant 0$ for $y \in (-\infty, +\infty)$, we have that $\int_{-\infty}^{+\infty} \psi_0^*(y)[U_\delta(y) - U_3(y)]\psi_0(y) dy < 0$. Also, since ψ_0 is not the ground state of H_δ for $\delta > 3$, then $\int_{-\infty}^{+\infty} \psi_0^*(y) H_\delta \psi_0(y) dy > \lambda_\delta$. With $\int_{-\infty}^{+\infty} \psi_0^*(y) H_3 \psi_0(y) dy = 0$ it follows that

$$\lambda_\delta < 0. \tag{27}$$

Alternatively, we see that the limit $\delta \to 3$ transforms the second pair of solutions (12) and (13) back to the first pair (9) and (10), for which we have $\lambda_3 = 0$. Thus, unicity of the ground state allows writing $\lambda_\delta \leqslant 0$ for $\delta \geqslant 3$, or better $\lambda_\delta < 0$ for $\delta > 3$, which is the region in parameter space where the second pair of solutions appears. This concludes our demonstration that $E_0(\mu, \nu, \lambda) < 0$ for all parameter values that respect Eq. (14), and thus that the soliton solutions discussed here are always unstable.

ACKNOWLEDGMENTS

This work is supported in part by funds provided by the U.S. Department of Energy (D.O.E.) under cooperative research agreement DE-FC02-94ER40818, and by CNPq and FINEP, Brazil.

[1] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, 1995), Chap. 3.
[2] T.-C. Chung, F. Moraes, J. D. Flood, and A. J. Heeger, Phys. Rev. B **29**, 2341 (1984).
[3] D. Bazeia, M. J. dos Santos, and R. F. Ribeiro, Phys. Lett. A **208**, 84 (1995).
[4] D. Bazeia and M. M. Santos, Phys. Lett. A **217**, 28 (1996).
[5] D. Bazeia, R. F. Ribeiro, and M. M. Santos, Phys. Rev. E **54**, 2943 (1996).
[6] D. Bazeia, J. R. S. Nascimento, and D. Toledo, Phys. Lett. A **228**, 357 (1997).
[7] D. Bazeia, J. R. S. Nascimento, R. F. Ribeiro, and D. Toledo, J. Phys. A **30**, 8157 (1997).
[8] H. M. Ruck, Nucl. Phys. B **167**, 320 (1980).
[9] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics, Part II* (McGraw–Hill, New York, 1953), p. 1654.
[10] L. Landau and E. Lifchitz, *Quantum Mechanics* (Pergamon, Oxford, 1977), p. 73.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

On the replica symmetric equations for the Hopfield model

L. Pastur
Institute for Low Temperature Physics, 310164, Kharkov, Ukraine and University Paris-7, 75251, Paris, France

M. Shcherbina
Institute for Low Temperature Physics, 310164, Kharkov, Ukraine

B. Tirozzi
Department of Physics of Rome University ''La Sapienza,'' 5, p-za A. Moro, Rome, Italy

(Received 18 February 1999; accepted for publication 17 March 1999)

We prove the central limit theorem for the cavity field in the case of the Hopfield model. © *1999 American Institute of Physics.* [S0022-2488(99)00908-1]

I. INTRODUCTION

During a long period of time the theory of disordered systems has been widely discussed in the physics literature. One of the first systems of this type which has been extensively investigated is spin glass. It is a disordered spin system with a zero average random interaction among spins. The mean field model of spin glass is analyzed and many interesting results are found in the physics literature.[1] The problem, as usual, is to study the thermodynamic behavior of the model, and this has been done, computing the limiting averaged free-energy. The most frequently applied method for solving such a problem is the replica calculation which until now has not found any rigorous support. For a long time there were only a few rigorous results (see Refs. 2 and 3) in this field, dealing only with the high-temperature region. Then in Refs. 4 and 5 we introduced an alternative rigorous method, the cavity method, which can work also in the low-temperature region. In this paper we solve some problems connected with the application of this approach to the Hopfield model.[6] The Hamiltonian of this model is

$$\mathrm{H}=-\frac{1}{2}\sum_{i,j=1}^{N} I_{ij}\sigma_i\sigma_j+\sum_{i=1}^{N} h_i\sigma_i\,, \tag{1}$$

where

$$I_{ij}=\frac{1}{N}\sum_{\mu=1}^{p} \xi_i^{\mu}\xi_j^{\mu}\,, \quad p=\alpha N \quad (\alpha>0) \tag{2}$$

with $\xi_i^{\mu}=\pm 1$ being independent random variables with zero mean.

One of the main points of the usual mean field theory (e.g., the Curie–Weiss model) was that the simple equality

$$\langle\sigma_1\rangle=\left\langle \tanh\left(\beta\sum_{i=2}^{N} I_{ij}\sigma_i+\beta h_1\right)\right\rangle \tag{3}$$

(valid for any interaction I_{ij}) can be rewritten as

$$\langle\sigma_1\rangle=\tanh\left(\beta\sum_{i=2}^{N} I_{ij}\langle\sigma_i\rangle+\beta h_1\right). \tag{4}$$

This relation follows from the factorization of the second correlation function in the thermodynamic limit

$$|\langle\sigma_i\sigma_j\rangle-\langle\sigma_i\rangle\langle\sigma_j\rangle|\to 0 \quad \text{as} \quad N\to\infty. \tag{5}$$

This property in the usual mean field theory is valid in the whole temperature region, if we add in the Hamiltonian some proper ''symmetry breaking field.''

It was rather natural to expect that in the mean-field-type model of spin glass (1) relations similar to (4) and (5) are also valid if we introduce some proper ''symmetry breaking field.''

In Ref. 4, using some special kind of infinitesimal field, it was shown that for the Sherrington–Kirkpatrick (SK) model[7] the relation (5) is equivalent to the self-averaging property of the Edwards–Anderson order parameter

$$E\{(q_N-E\{q_N\})^2\}\to 0 \quad \text{as} \quad N\to\infty, \tag{6}$$

where

$$q_N=\frac{1}{N}\sum_{i=1}^{N}\langle\sigma_i\rangle^2. \tag{7}$$

Deriving (5) from (6) and studying the moments of the random variable $\langle\sigma_1\rangle$, it was proved that if (6) is true, then in distribution

$$\langle\sigma_1\rangle=\tanh\left(\beta\sum_{i=2}^{N} I_{ij}\langle\sigma_j\rangle_0+\beta h_1\right). \tag{8}$$

Here and below the symbol $\langle\cdots\rangle_0$ means the average with respect to the Gibbs measure, corresponding to the Hamiltonian (1), if we set here $\sigma_1=0$.

This idea was developed in Ref. 8, where the infinitesimal field was replaced by an ordinary Gaussian one that, in particular, allows us to derive on the basis of Griffith's lemma an important relation, valid for almost all values of β and h_i,

$$\frac{1}{N}\sum_{i=1}^{N} h_i(\sigma_i-\langle\sigma_i\rangle)\to 0 \quad \text{as} \quad N\to\infty,$$

in the Gibbs measure and probability. This relation simplifies considerably the method of Ref. 4 and allows us to prove that (6) is equivalent to (8) in a more natural way. The further development of this method allows us to prove that (6) and (8) for the SK model are valid for a large region of parameters, including the low temperatures.[9]

The same method was used in Ref. 5, where we derived (8) from (6). Similar ideas were used later by M. Talagrand[10,11] to prove (8) for the SK model with a high temperature and for the Hopfield model with small α's. Analogous results were found in Ref. 12.

Now let us recall that in the case of the SK model

$$I_{ij}=\frac{1}{\sqrt{N}}J_{ij},$$

where $\{J_{ij}\}$ are independent (Gaussian) random variables with zero mean. Since $\langle\sigma_j\rangle_0$ does not depend on J_{1j} and the Lindeberg condition in the form

$$E\left\{N^{-2}\sum_{j=2}^{N}\langle\sigma_j\rangle_0^4\right\}\to 0 \quad \text{as} \quad N\to\infty \tag{9}$$

is trivial due to the bounded values of spins $\{\sigma_j\}$, the central limit theorem gives us immediately that the cavity field $\Sigma I_{1j}\langle\sigma_j\rangle_0$ converges in distribution to the Gaussian random variable. Thus, from (8) we obtain the relation

$$\overline{q_N}=E\{\langle\sigma_1\rangle_0^2\}=\int \tanh^2\beta(J\sqrt{\overline{q_N}}u+h_1)\frac{e^{-u^2/2}du}{\sqrt{2\pi}}d\mu(h_1)+o(1)\quad(\overline{q_N}=E\{q_N\}),$$

which gives us the so-called replica symmetric equation for $\overline{q}$.

However, in the case of the Hopfield interaction (2), the situation is more complicated. If we write in a natural way

$$\sum_{i=2}^{N} I_{1j}\langle\sigma_i\rangle_0=\sum_{\mu=2}^{p}\frac{\xi_1^\mu}{\sqrt{N}}\langle t_1^\mu\rangle_0$$

with

$$t_1^\mu=\frac{1}{\sqrt{N}}\sum_{i=2}^{N}\xi_i^\mu\sigma_i, \tag{10}$$

then $\langle t_1^\mu\rangle_0$, like in the SK case, are independent of $\{\xi_1^\mu\}$. However, to apply the central limit theorem now we have to check the condition of the type

$$E\left\{N^{-2}\sum_{\mu=2}^{p}\langle t_1^\mu\rangle_0^4\right\}\to 0\quad\text{as}\quad N\to\infty. \tag{11}$$

Since t_1^μ can be of the order $\sqrt{N}$, the property (11) is much less trivial then (9). The main aim of the present paper is to prove (11). In fact, we prove a stronger condition on $\langle t_1^\mu\rangle_0$ (see Lemma 1 below).

The paper is organized as follows. In Sec. II we describe exactly the model and the results, in Sec. III we prove the main results, and in the Appendix we prove some auxiliary results.

II. THE MODEL AND THE MAIN RESULTS

Define H as a sum of two Hamiltonians

$$\mathrm{H}=\mathrm{H}_0+\mathrm{H}_1, \tag{12}$$

where

$$\begin{aligned}\mathrm{H}_0&=-\frac{J}{2N}\sum_{\mu=s+1}^{p}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j-\varepsilon_1\sum_{\mu=s+1}^{p}\gamma^\mu t^\mu+\varepsilon_2\sum_{i=1}^{N}h_i\sigma_i,\\ \mathrm{H}_1&=-\frac{J(1+d_N\zeta)}{2N}\sum_{\nu=1}^{s}\sum_{i,j=1}^{N}\xi_i^\nu\xi_j^\nu\sigma_i\sigma_j-h^1\sum_{i=1}^{N}\xi_i^1\sigma_i-\sum_{\nu=2}^{s}\gamma^\nu t^\nu,\end{aligned} \tag{13}$$

$s=[\log^{1/2}N]$ is the number of the patterns which are expected to be condensed, J, h^1, ε_1, and ε_2 are positive parameters, $d_N=s^{-2/3}$, ζ is an independent random variable uniformly distributed in the interval (1,2), and variables γ^μ,h_i are independent Gaussian random variables with zero mean and variance 1.

The Hamiltonian H_0 contains the contribution of the noncondensed patterns and the Hamiltonian H_1 includes terms due to the condensed patterns. The random variables γ^μ,h_i play the role of ''symmetry breaking fields'' (we need them for technical reasons) and after thermodynamic limit $N\to\infty$ we send $\varepsilon_1,\varepsilon_2\to 0$.

The variable t^μ is just a convenient notation for the following linear combination of spins:

$$t^\mu \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^\mu \sigma_i, \quad \mu = 1,...,p. \tag{14}$$

We will use also notations

$$m^\mu \equiv \frac{1}{\sqrt{N}} t^\mu, \quad r_N = p^{-1} \sum_{\mu=s+1}^{p} \langle t^\mu \rangle^2, \quad U_N = N^{-1} \sum_{\mu=s+1}^{p} (t^\mu)^2,$$

$$q_N = N^{-1} \sum_{i=1}^{N} \langle \sigma_i \rangle^2, \quad \overline{r_N} = E\{r_N\} \quad \overline{q_N} = E\{q_N\}. \tag{15}$$

The main result of the paper is the lemma, which allows us to overcome one of the main technical difficulties, arising in the Hopfield model, if we try to generalize to it the methods proposed in Refs. 4 and 9 for the SK model. This difficulty is connected to the fact that the variables t^μ, which play here the role of ''spins'' of the SK model, are unbounded.

Lemma 1: Consider the set

$$\mathcal{M} \equiv \{\mathbf{m} = (m^1, \ldots, m^p) : \max_{\nu \geq s+1} |m^\nu| \geq 4\delta_N\},$$

where $\delta_N = s^{-1/3} = d_N^{1/3}$. *Let* $\chi_{\mathcal{M}}(m)$ *be the characteristic function of the set* $\mathcal{M}$:

$$\chi_{\mathcal{M}}(\mathbf{m}) = \begin{cases} 1, & \mathbf{m} \in \mathcal{M}, \\ 0, & \text{otherwise.} \end{cases}$$

Then for any $n > 0$ *there exists a constant* C_n *independent of N such that*

$$\text{Prob}\{\langle \chi_{\mathcal{M}}(\mathbf{m}) \rangle \leq e^{-\beta J N d_N^2/4}\} \geq 1 - C_n N^{-n}. \tag{16}$$

Moreover, if we add to the Hamiltonian H *any Hamiltonian* $\tilde{\text{H}}$ *which is symmetrical with respect to the variables* $\{\xi_1^\mu\}_{\mu>s}, \{\xi_2^\mu\}_{\mu>s}, \ldots, \{\xi_N^\mu\}_{\mu>s}$ *and the free energies of the sums* $\text{H}^a + \tilde{\text{H}}(\text{H} + \tilde{\text{H}})$ *satisfy the large deviation bounds of the type*

$$\text{Prob}\{|f(\text{H} + \tilde{\text{H}}) - E_{/\zeta} f(\text{H} + \tilde{\text{H}})\}| > \varepsilon\} \leq D_n N^{-n},$$

then the estimate of the probability (16) is valid for Gibbs averages with respect to $\text{H} + \tilde{\text{H}}$.

Here and below the symbol $E_{/\zeta}\{\cdots\}$ means the average with respect to all random variables of the problem except ζ.

Remarks:

(1) This lemma allows us to treat m^ν ($\nu > s$) like variables satisfying inequalities

$$|m^\nu| \leq 4 d_N^{1/2}, \tag{17}$$

because the measure of m^ν satisfying the opposite inequality decays exponentially with probability 1, starting from some large enough N, and so those m^ν can add only exponentially small contributions in our estimates. In particular,

$$E\left\{ \sum_{\nu=s+1}^{p} \langle (m^\nu)^4 \rangle \right\} \leq 16 d_N E\left\{ \sum_{\nu=s+1}^{p} \langle (m^\nu)^2 \rangle \right\} \leq 16 d_N E\{\|\mathcal{J}\|\} = \text{const}\, d_N. \tag{18}$$

(2) The other important corollary of Lemma 1 is the formula, analogous to the formula of integration by parts, which usually is used for Gaussian variables.

Proposition 1: Let t_1^μ *be defined by (10). Then for any* $\mu_1,\dots,\mu_k>s$, $\mu_i\neq\mu_j$, *and any bounded function* $F_1(\{\sigma\},\{\xi_i^\mu\}),\dots,F_l(\{\sigma\},\{\sigma_i^\mu\})$ *which do not depend on* ξ_1^μ,

$$N^{k/2}E\{\xi_1^{\mu_1}\cdots\xi_1^{\mu_k}\langle t_1^{\mu_1}\cdots t_1^{\mu_{k_1}}F_1(\{\sigma\},\{\xi_i^\mu\})\rangle\cdots\langle t_1^{\mu_{k_{l-1}+1}}\cdots t_r^{\mu_k}F_l(\{\sigma\},\{\xi_i^\mu\})\rangle\}$$

$$=N^{k/2}E\left\{\frac{\partial^k}{\partial\xi_1^{\mu_1}\cdots\partial\xi_1^{\mu_1}}\langle t_1^{\mu_1}\cdots t_1^{\mu_{k_1}}F_1(\{\sigma\},\{\xi_i^\mu\})\rangle\cdots\langle t_1^{\mu_{k_{l-1}+1}}\cdots t_1^{\mu_k}F_l(\{\sigma\},\{\xi_i^\mu\})\rangle\right\}+R, \tag{19}$$

with

$$|R|\leqslant d_N^{1/2}\cdot(2l)^{k+1}\max|F_1|\cdots\max|F_l|\cdot(2+\sqrt{\alpha})^k.$$

The same relations are valid for the Hamiltonian H^a *and any perturbed Hamiltonian* $\mathrm{H}^a+\tilde{\mathrm{H}}$ *with perturbation* $\tilde{\mathrm{H}}$, *if this perturbation is symmetrical with respect to* ξ_1^μ *in the sense of Lemma 1 and for any k, p,*

$$E^{1/p}\left\{\left\langle\left(\frac{\partial^k\tilde{\mathrm{H}}}{\partial\xi_1^{\mu_1}\cdots\partial\xi_1^{\mu_k}}\right)^{2p}\right\rangle\right\}\leqslant\frac{C_{k,p}}{N^k}. \tag{20}$$

with the constant $C_{k,p}$ *independent of N. Here and below the symbol* $\partial/\partial\xi_1^\mu$ *means the formal derivative, which can be obtained if we replace* ξ_1^μ *by a continuous variable.*

Lemma 1 and Proposition 1 allow us to prove a bit different variant of the main theorem of Ref. 5.

Theorem 1: *Consider the Hopfield model of the form (12) and (13). Set*

$$\Delta_N=E\left\{(Np)^{-1}\sum_{\mu>s,i\geqslant 1}\langle(t^\mu-\langle t^\mu\rangle)(\sigma_i-\langle\sigma_i\rangle)\rangle^2\right\}. \tag{21}$$

Then for almost all values of parameters J, h^1, ε_1, ε_2, α, *and* β, *parameters* $\overline{q}_N,\overline{r}_N$ *satisfy the system of equations*

$$\overline{r}_N=\frac{\{\overline{q}_N+\varepsilon_1^2J^2\beta^2(1-\overline{q}_N)^2\}}{(1-\beta J(1-\overline{q}_N))^2}+O(\Delta_N^{1/2})+o(1),$$

$$\overline{q}_N=E\left\{\int\frac{dv\,\exp(-v^2/2)}{\sqrt{2\pi}}\tanh^2\beta\left((\alpha\overline{r}_N(\varepsilon_1))^{1/2}v+h^1\xi_1^1+J\sum_{\nu=1}^{s}c^\nu\xi_1^\nu+\varepsilon_2h_1\right)\right\}$$
$$+O(\Delta_N^{1/2})+o(1), \tag{22}$$

where

$$\overline{r}_N(\varepsilon_1)=J^2\overline{r}_N+\frac{2J\beta\varepsilon_1^2}{1-\beta J(1-\overline{q}_N)}+\varepsilon_1^2$$

and parameters c^ν *satisfy the equations*

$$c^\nu=E\left\{\int\frac{dv\,\exp(-v^2/2)}{\sqrt{2\pi}}\xi_1^\mu\tanh\beta\left((\alpha\overline{r}_N(\varepsilon_1))^{1/2}v+h^1\xi_1^1+J\sum_{\nu=1}^{s}c^\nu\xi_1^\nu+\varepsilon_2h_1\right)\right\}$$
$$+O(\Delta_N^{1/2})+o(1).$$

To study the influence of the ‘‘condensed’’ patterns we use the following.

Lemma 2: Consider the ‘‘approximate’’ Hamiltonian of the form

$$\mathrm{H}^a(\mathbf{c})=\mathrm{H}_0-J(1+d_N\zeta)\sum_{\nu=1}^{s} c^\nu\sum_{i=1}^{N} \xi_i^\nu\sigma_i-\sum_{\nu=1}^{s} \gamma^\nu t^\nu-h^1\sum_{i=1}^{N} \xi_i^1\sigma_i+J(1+d_N\zeta)\frac{N}{2}\sum_{\nu=1}^{s} (c^\nu)^2, \tag{23}$$

where H_0 *is defined by formula (13) and* $\mathbf{c}\equiv(c^1,\ldots,c^s)$. *Then the free energies of the initial Hamiltonian* H *and the ‘‘approximate’’ Hamiltonian* H^a *satisfy the inequality*

$$0\leq\min_{\mathbf{c}} E_{/\zeta}\{f(\mathrm{H}^a(\mathbf{c}))\}-E_{/\zeta}\{f(\mathrm{H})\}\leq\frac{\mathrm{const}}{\log N}, \tag{24}$$

and for almost all J, h^1, ε_1, ε_2,

$$\overline{r_N^*}\equiv E\left\{p^{-1}\sum_{\mu=s+1}^{p} \langle t^\mu\rangle_*^2\right\}=\overline{r_N}+o(1),\quad \overline{q_N^*}\equiv E\left\{N^{-1}\sum_{i=1}^{N} \langle\sigma_i\rangle_*^2\right\}=\overline{q_N}+o(1). \tag{25}$$

Here and everywhere below we use notation

$$\mathrm{H}_*^a=\mathrm{H}^a(\mathbf{c}_*(\zeta))$$

for the Hamiltonian H^a computed at the point $\mathbf{c}_*(\zeta)\equiv(c_*^1(\zeta),\ldots,c_*^s(\zeta))$ which provides the minimum value of the mean free energy $E_{/\zeta}\{f(\mathrm{H}^a(c^1,\ldots,c^s))\}$, and the symbol $\langle\cdots\rangle_*$ for the respective Gibbs average. Thus the symbols $\overline{q_N^*}, \overline{r_N^*}$ are the values of order parameters computed by means of this Gibbs measure.

Remarks:

(1) Lemma 2 allows us to substitute the Hamiltonian H with H_*^a, which is linear with respect to the first s patterns.

(2) Since the Hamiltonian evidently has the form

$$\mathrm{H}_*^a=\mathrm{H}+\frac{J(1+\zeta d_N)N}{2}\sum_{\nu=1}^{s} (m^\nu-c_*^\nu)^2,$$

it is easy to see that it satisfies conditions of Lemma 1 and Proposition 1 for the perturbed Hamiltonians, and the estimates (19) are also valid for the $\langle\cdots\rangle_*$ averages.

Using Lemmas 1 and 2 and Proposition 1, one can prove Theorem 1 by the method proposed in Ref. 5. One of the main steps in the proof is given by the lemma about the properties of the cavity field, which needs some extra definitions for its statement.

Define the Hamiltonian $\Phi(\tau)$, interpolating between systems of $N-1$ and N spins,

$$\Phi(\tau)=\mathrm{H}_1^a-\frac{J\tau}{\sqrt{N}}\sum_{\mu=1}^{p} \xi_1^\mu t_1^\mu, \tag{26}$$

where

$$\mathrm{H}_1^a=\mathrm{H}_1^0-\frac{J}{2N}\sum_{\mu=s+1}^{p} (t_1^\mu)^2-J(1+\zeta d_N)\sum_{\nu=1}^{s} c_*^\nu\sum_{i=2}^{N} \xi_i^\nu\sigma_i-h^1\sum_{i=2}^{N} \xi_i^1\sigma_i$$
$$+\frac{JN(1+\zeta d_N)}{2}\sum_{\nu=1}^{s} (c_*^\nu)^2-\sum_{\nu=2}^{s} \gamma^\nu t_1^\nu. \tag{27}$$

It is easy to see that if $\tau=\sigma_1$, then $\Phi(\tau)$ coincides with H^a.

Consider also the corresponding partition function

$$Z(\tau)=\sum_{\sigma_2,\ldots,\sigma_N} e^{-\beta\Phi(\tau)}, \tag{28}$$

and define the relative free energy

$$u(\tau)=\ln\frac{Z(\tau)}{Z(0)}. \tag{29}$$

Lemma 3: The function $u(\tau)$ can be represented in the form

$$u(\tau)=\beta J\tau\sum_{\nu=s+1}^{p}\frac{\xi_1^\nu}{\sqrt{N}}\langle t_1^\nu\rangle_0+\beta J\tau\sum_{\mu=1}^{s}\xi_1^\mu c_*^\mu+(\langle U_N\rangle_0-\alpha\overline{r_N})\frac{\tau^2(\beta J)^2}{2}+R_N(\tau), \tag{30}$$

where $\langle\cdots\rangle_\tau$ is the Gibbs averaging, corresponding to the Hamiltonian $\Phi(\tau)$, U_N is defined by (15), and the remainder R_N can be estimated as

$$E\{R_N^2\}\leqslant\text{const}\,\Delta_N+o(1). \tag{31}$$

Remarks:

(1) Since $\langle t_1^\nu\rangle_0$ do not depend on ξ_1^ν, using Lemma 1, one can prove that the cavity field [the first sum on the rhs of (30)] converges in distribution to $\beta J_\tau\sqrt{\alpha r_N}\nu_1$, where ν_1 is a Gaussian random variable with zero mean and variance 1.

(2) At the present time there exists a proofs of Lemma 3 (see Refs. 5 and 10). Thus, in this paper we omit the proof of his lemma and the derivation of the equations (22).

III. PROOFS OF LEMMAS 1 AND 2 AND PROPOSITION 1

Proof of Lemma 1: For given $\mu\leqslant s$ and $\nu\geqslant s+1$ consider the set

$$\mathcal{A}^{\mu\nu}=\{\mathbf{m}=(m^1,\ldots,m^p):(1+d_N\zeta)|m^\mu|\leqslant|m^\nu|-3\delta_N\}. \tag{32}$$

Its Gibbs measure is

$$a^{\mu\nu}=\langle\theta(|m^\nu|-(1+d_N\zeta)|m^\mu|-3\delta_N)\rangle, \tag{33}$$

where $\theta(m)=\frac{1}{2}(1+\text{sign}\,m)$.

Let us assume that for any n we have proved the estimate

$$\text{Prob}\{a^{\mu\nu}\leqslant e^{-\beta JNd_N^2/5}\}\geqslant 1-C_n'N^{-n}. \tag{34}$$

Consider the set

$$\mathcal{A}\equiv\bigcup_{\mu=1}^{s}\bigcup_{\nu=s+1}^{p}\mathcal{A}^{\mu\nu}=\{\mathbf{m}=(m^1,\ldots,m^p):(1+d_N\zeta)\min_{\mu\leqslant s}|m^\mu|\leqslant\max_{\nu\geqslant s+1}|m^\nu|-3\delta_N\}. \tag{35}$$

Then it follows from (34) that the Gibbs measure $\langle\chi_{\mathcal{A}}(m)\rangle$ of the set $\mathcal{A}$ satisfies the estimate

$$\text{Prob}\{\langle\chi_{\mathcal{A}}(m)\rangle\leqslant s(p-s)e^{-\beta JNd_N^2/5}\leqslant e^{-\beta JNd_N^2/6}\}\geqslant 1-s(p-s)C_n'N^{-n}\geqslant 1-C_{n-2}N^{-(n-2)},$$

when N is large enough. Then we use the inequality

$$\langle\chi_{\mathcal{M}}(m)\rangle=\langle\chi_{\mathcal{M}\cap\mathcal{A}}(m)\rangle+\langle\chi_{\mathcal{M}\cap\overline{\mathcal{A}}}(m)\rangle\leqslant\langle\chi_{\mathcal{A}}(m)\rangle+\langle\chi_{\mathcal{M}\cap\overline{\mathcal{A}}}(m)\rangle. \tag{36}$$

However,

$$\mathcal{M}\cap\overline{\mathcal{A}}\equiv\{\mathbf{m}:(1+d_N\zeta)\min_{\mu\leqslant s}|m^\mu|\geqslant \max_{\nu\geqslant s+1}|m^\nu|-3\delta_N,\ \max_{\nu\geqslant s+1}|m^\nu|>4\delta_N\}$$

$$\subset\{\mathbf{m}:(1+d_N\zeta)\min_{\mu\leqslant s}|m^\mu|\geqslant\delta_N\}$$

$$\subset\left\{\mathbf{m}:(1+d_N\zeta)^2\sum_{\mu\leqslant s}(m^\mu)^2\geqslant s\,\delta_N^2\right\}. \tag{37}$$

By using the definition (15) of m^μ and then the result of Ref. 13, we get that for any $\tilde{d}>0$ [in our case $\tilde{d}=(\log N)^{1/6}/(1+d_N\zeta)-(1+\sqrt{\alpha})^2$] the probability of the event

$$\left\{\sum_{\mu\leqslant s}(m^\mu)^2\geqslant(1+\sqrt{\alpha})^2+\tilde{d}\right\}\subset\left\{\frac{1}{N}\sum_{i,j}J_{ij}\sigma_i\sigma_j\geqslant(1+\sqrt{\alpha})^2+\tilde{d}\right\}\subset\{\|\mathcal{J}\|\geqslant(1+\sqrt{\alpha})^2+\tilde{d}\}$$

is less than $e^{-N^{2/3}\tilde{d}^{4/3}\,\mathrm{const}}$. Therefore the probability to have the last set in (37) nonempty is also less then $e^{-N^{2/3}\,\mathrm{const}}$. Thus (36), (37), and (34) prove (16).

Now we are left to prove (34).

To this end we use the standard representation

$$\exp\left\{\beta JN\left((1+d_N\zeta)\frac{(m^\mu)^2}{2}+\frac{(m^\nu)^2}{2}\right)\right\}=\left((2\pi\beta JN)^{-1}(1+d_N\zeta)\right)^{1/2}\int dx\,dy$$
$$\times\exp\left\{\beta JN\left(xm^\mu+ym^\nu-\frac{x^2}{2(1+d_N\zeta)}-\frac{y^2}{2}\right)\right\} \tag{38}$$

and study

$$a_1^{\mu\nu}=\frac{\int\theta(|y|-|x|-2\delta_N)\exp\{\beta JNF_{N,\mu\nu}(x,y)\}dx\,dy}{\int\exp\{\beta JNF_{N,\mu\nu}(x,y)\}dx\,dy}, \tag{39}$$

where $F_{N,\mu\nu}(x,y)$ is a random function defined by the formulas

$$F_{N,\mu\nu}(x,y)\equiv f_{N,\mu\nu}(Jx,Jy)-\frac{x^2}{2(1+d_N\zeta)}-\frac{y^2}{2},$$

$$f_{N,\mu\nu}(x,y)\equiv\frac{1}{\beta JN}\log\sum_{\{\sigma\}}\exp\{-\beta\mathrm{H}_{\mu\nu}(\sigma;x,y),\} \tag{40}$$

with the Hamiltonian $\mathrm{H}_{\mu,\nu}(\sigma;x,y)$ of the form

$$\mathrm{H}_{\mu\nu}(\sigma;x,y)=-NJ\sum_{\mu'\neq\nu;\mu'>s}(m^{\mu'})^2-NJ(1+\zeta d_N)\sum_{\mu'\neq\mu;\mu'\leqslant s}(m^{\mu'})^2-\sum_{i=1}^N h_i\sigma_i$$
$$-\varepsilon_1\sqrt{N}\sum_{\nu'>s}\gamma^{\nu'}m^{\nu'}-\sqrt{N}\sum_{\mu'\leqslant s}\gamma^{\mu'}m^{\mu'}-Nxm^\mu-Nym^\nu.$$

Then we use the inequality which follows from the Laplace method:

$$\frac{\int \theta(|y|-|x|-2\delta_N)\exp\{\beta JN(m^\mu x+m^\nu y-x^2/2(1+d_N\zeta)-y^2/2)\}dxdy}{\int \exp\{\beta JN(m^\mu x+m^\nu y-x^2/2(1+d_N\zeta)-y^2/2)\}dxdy}$$
$$\geqslant(\theta(|m^\nu|-(1+d_N\zeta)|m^\mu|-3\delta_N)-e^{-N\beta J\delta_N^2/2(2+d_N\zeta)}). \qquad (41)$$

Indeed, if $|m^\nu|-(1+d_N\zeta)|m^\mu|-3\delta_N<0$, this inequality is trivial.

If $|m^\nu|-(1+d_N\zeta)|m^\mu|-3\delta_N\geqslant 0$, then the θ-function on the rhs of (41) is equal to 1 and the ratio on the lhs is equal to

$$1-\frac{\int_{\mathcal{D}}\exp\{\beta JN(-(x-x^*)^2/2(1+d_N\zeta)-(y-y^*)^2/2)\}dxdy}{\int \exp\{\beta JN(-(x-x^*)^2/2(1+d_N\zeta)-(y-y^*)^2/2)\}dxdy},$$

where $\mathcal{D}=\{(x,y):|y|-|x|-2\delta_N\leqslant 0\}$ and $x^*=(1+d_N\zeta)m^\mu$, $y^*=m^\nu$. In our case $(|m^\nu|-(1+d_N\zeta)|m^\mu|-3\delta_N\geqslant 0)$ and so $(x*,y^*)\notin\mathcal{D}$. According to the Laplace method, the last ratio in the above formula can be estimated from above:

$$\frac{\int_{\mathcal{D}}\exp\{\beta JN(-(x-x^*)^2/2(1+d_N\zeta)-(y-y^*)^2/2)\}dxdy}{\int \exp\{\beta JN(-(x-x^*)^2/2(1+d_N\zeta)-(y-y^*)^2/2)\}dxdy}$$
$$\leqslant\exp\left\{\beta JN\max_{\mathcal{D}}\left[-\frac{(x-x^*)^2}{2(1+d_N\zeta)}-\frac{(y-y^*)^2}{2}\right]\right\}\leqslant e^{-N\beta J\delta_N^2/2(+d_N\zeta)}.$$

Thus, we obtain that

$$a_1^{\mu\nu}=\frac{\int\theta(|y|-|x|-2\delta_N)\exp\{\beta JNF_{N,\mu\nu}(x,y)\}dxdy}{\int\exp\{\beta JNF_{N,\mu\nu}(x,y)\}dxdy}$$
$$=\frac{\Sigma_{\{\sigma\}}\exp\{-\beta H_{\mu\nu}(\sigma;0,0)\}\int\theta(|y|-|x|-2\delta_N)e^{\beta JN(m^\mu x+m^\nu y-x^2/2(1+d_N\zeta)-y^2/2)}\,dxdy}{2\pi\beta JN(1+\zeta d_N)^{-1/2}\Sigma_{\{\sigma\}}e^{-\beta H(\sigma)}}$$
$$\geqslant\frac{\Sigma_{\{\sigma\}}\exp\{-\beta H_{\mu\nu}(\sigma;0,0)\}\theta(|m^\nu|-(1+d_N\zeta)|m^\mu|-3\delta_N)\int e^{\beta JN(m^\mu x+m^\nu y-x^2/2(1+d_N\zeta)-y^2/2)}\,dxdy}{2\pi\beta JN(1+\zeta d_N)^{-1/2}\Sigma_{\{\sigma\}}e^{-\beta H(\sigma)}}$$
$$-e^{-N\beta J\delta_N^2/2(2+d_N\zeta)}$$
$$=a^{\mu,\nu}-e^{-N\beta J\delta_N^2/2(2+d_N\zeta)}.$$

So,

$$a^{\mu\nu}\leqslant a_1^{\mu\nu}+e^{-N\beta J\delta_N^2/2(2+d_N\zeta)}. \qquad (42)$$

Now we apply the Laplace method to the integral on the rhs of (39) [let us recall that evidently $|(\partial/\partial x)f_{N,\mu\nu}(x,y)|,|(\partial/\partial y)f_{N,\mu\nu}(x,y)|\leqslant 1$ and therefore $F_{N,\mu\nu}(x,y)$ has bounded derivatives in the field of interest]. We have

$$a_1^{\mu\nu}=\frac{\int_{|y|-|x|\geqslant 2\delta_N}\exp\{\beta JNF_{N,\mu\nu}(x,y)\}dxdy}{\int\exp\{\beta JNF_{N,\mu\nu}(x,y)\}dxdy}$$
$$\leqslant\exp\{\beta JN[\max_{|y|-|x|\geqslant 2\delta_N}F_{N,\mu\nu}(x,y)-\max_{x,y}F_{N,\mu\nu}(x,y)]\}. \qquad (43)$$

We will show below that

$$\max_{|y|-|x|\geq 2\delta_N} F_{N,\mu\nu}(x,y)-\max_{x,y} F_{N,\mu\nu}(x,y)\leq -\zeta\frac{d_N\delta_N^2}{2}. \tag{44}$$

From this inequality, using (43) and the fact that $1\leq\zeta\leq 2$, we get

$$a_1^{\mu\nu}\leq e^{-\beta JNd_N\delta_N^2/2}, \tag{45}$$

if N is large enough. This inequality together with (42) prove (34) and thus Lemma 1.

Now let us show (44). Denote by $E_{\mu\nu}\{\cdots\}$ the average with respect to all random parameters of the problem except γ^μ, γ^ν, and ζ and rewrite $F_{N,\mu\nu}(x,y)$ as

$$F_{N,\mu\nu}(x,y)=E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}+R_{N,\mu\nu}(x,y),\quad R_{N,\mu\nu}(x,y)\equiv F_{N,\mu\nu}(x,y)-E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}.$$

Then we have

$$\max_{|y|-|x|\geq 2\delta_N} F_{N,\mu\nu}(x,y)-\max_{x,y} F_{N,\mu\nu}(x,y)$$
$$\leq \max_{|y|-|x|\geq 2\delta_N} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}-\max_{x,y} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}+2\max_{x,y}|R_{N,\mu\nu}(x,y)|\}. \tag{46}$$

To proceed further, we use the following proposition.

Proposition 2: Let $f(x,y)$ be a smooth function, satisfying the symmetry conditions

$$f(x,y)=f(y,x),\quad f(x,y)=f(-x,y). \tag{47}$$

Consider the function $F(x,y)$ of the form

$$F(x,y)\equiv f(x,y)-\frac{x^2}{2(1+d)}-\frac{y^2}{2}\quad (d>0). \tag{48}$$

Then

$$\max F(x,y)-\max_{|x|\leq|y|-2\delta} F(x,y)\geq\frac{2d\delta^2}{(1+d)}. \tag{49}$$

We prove this proposition in the Appendix.

Let us remark that

$$\max_{|y|-|x|\geq 2\delta_N} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}-\max_{x,y} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}$$
$$\leq \max_{|y|-|x|\geq 2\delta_N} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}|_{\gamma^\mu=\gamma^\nu=0}$$
$$-\max_{x,y} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}|_{\gamma^\mu=\gamma^\nu=0}+2\left(\left|\frac{\gamma^\mu}{\sqrt{N}}\right|+\left|\frac{\varepsilon_1\gamma^\nu}{\sqrt{N}}\right|\right).$$

Now we apply Proposition 2 to the function $E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}|_{\gamma^\mu=\gamma^\nu=0}$ and use also the estimate

$$2\left(\left|\frac{\gamma^\mu}{\sqrt{N}}\right|+\left|\frac{\varepsilon_1\gamma^\nu}{\sqrt{N}}\right|\right)\leq\frac{1-d_N\zeta}{1+\zeta d_N}\zeta d_N\delta_N^2. \tag{50}$$

Since γ^μ and γ^ν are Gaussian random variables, this estimate is valid with probability larger than $1-e^{-N\,\text{const}\,d_N^2\delta_N^4}$. Thus,

$$\max_{|y|-|x|\geqslant 2\delta_N} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}-\max_{x,y} E_{\mu\nu}\{F_{N,\mu\nu}(x,y)\}$$

$$\leqslant -\frac{2d_N\zeta\delta_N^2}{1+\zeta d_N}+\frac{d_N\zeta\delta_N^2(1-\zeta d_N)}{1+\zeta d_N}$$

$$=-\zeta d_N\delta_N^2=-\zeta d_N^2 \tag{51}$$

with the same probability.

If also

$$2\max_{x,y}|R_{N,\mu\nu}(x,y)|\leqslant\zeta\frac{d_N^2}{2}, \tag{52}$$

then (46) and (51) imply (44). Thus, we obtain that the probability of (45) is bounded from above by the sum of probabilities of (50) and (52).

Now we are faced with the problem of estimating the probability of (52). To this aim let us remark, that since $|(\partial/\partial x)f_{N,\mu\nu}|,|(\partial/\partial x)f_{N,\mu\nu}|\leqslant 1$, all the extremal points of $F_{N,\mu,\nu}(x,y)$ are inside the square $\{(x,y):|x|,|y|\leqslant 1\}$. Besides, for any (x,y) from this square there exist i,j, $(|i|,|j|\leqslant M,\ M\equiv[4\sqrt{2}d_N^{-2}])$ such that

$$\begin{aligned}
f_{N,\mu\nu}(x,y)-E_{/\zeta}\{f_{N,\mu\nu}(x,y)\}&=f_{N,\mu\nu}\left(\frac{i}{M},\frac{j}{M}\right)-E_{/\zeta}\left\{f_{N,\mu\nu}\left(\frac{i}{M},\frac{j}{M}\right)\right\}+f_{N,\mu\nu}(x,y)\\
&\quad -f_{N,\mu\nu}\left(\frac{i}{M},\frac{j}{M}\right)+E_{/\zeta}\left\{f_{N,\mu\nu}\left(\frac{i}{M},\frac{j}{M}\right)\right\}-E_{/\zeta}\{f_{N,\mu\nu}(x,y)\}\\
&=f_{N,\mu\nu}\left(\frac{i}{M},\frac{j}{M}\right)-E_{/\zeta}\left\{f_{N,\mu\nu}\left(\frac{i}{M},\frac{j}{M}\right)\right\}+R_1+R_2,
\end{aligned} \tag{53}$$

where

$$|R_{1,2}|\leqslant\sqrt{\left(x-\frac{i}{M}\right)^2+\left(y-\frac{j}{M}\right)^2}\leqslant\frac{1}{\sqrt{2}M}\leqslant\frac{d_N^2}{8}.$$

Thus, for our goal it is enough to estimate the probability of the event

$$\left|f_{N,\mu,\nu}\left(\frac{i}{M},\frac{j}{M}\right)-E_{/\zeta}\left\{f_{N,\mu,\nu}\left(\frac{i}{M},\frac{j}{M}\right)\right\}\right|\leqslant\frac{d_N^2}{4}, \tag{54}$$

because this inequality and (53) imply (52).

Since, according to Ref. 14, the probability of the last event for fixed (i,j) can be estimated as

$$\text{Prob}\left\{|f_{N,\mu\nu}(x_k,y_k)-E_{/\zeta}\{f_{N,\mu\nu}(x_k,y_k)\}|\leqslant\frac{d_N^2}{4}\right\}\geqslant 1-D_nN^{-n},$$

and the number of the events of such a type is $4M^2$, the probability of the inequality (52) is more then $1-4M^2D_nN^{-n}\geqslant 1-C_n'N^{2-n}$. On the other hand, since γ^ν, γ^μ are Gaussian random variables, the probability of (50) is more then $1-e^{-N\,\text{const}\,d_N^2}$. Therefore the inequalities (42)–(54) and the last conclusions prove (34), that, as it was mentioned above, implies (16).

For the perturbed Hamiltonian the proof is the same.

Proof of Proposition 1: To simplify formulas we prove formula (19) in the case $l=1$. The general case for this formula and also its modification for the perturbed Hamiltonian can be proved similarly.

Let us introduce the vector $\overline{\theta}\in \mathbf{R^k}$, $\overline{\theta}\equiv(\theta_1,\dots,\theta_k)$, and define $\mathrm{H}(\overline{\theta})$ to be equal to the Hamiltonian H_*^a, if we substitute in the latter $\xi_1^{\mu_1},\dots,\xi_1^{\mu_k}$ by $\theta_1\xi_1^{\mu_1},\dots,\theta_k\xi_1^{\mu_k}$. Let

$$\phi(\theta_1,\dots,\theta_k)\equiv N^{k/2}\langle t_1^{\mu_1}\cdots t_1^{\mu_k}F_1(\{\sigma\},\{\xi_i^{\mu}\})\rangle_{\mathrm{H}(\overline{\theta})}.$$

Consider

$$E\left\{\int_0^1\cdots\int_0^1 d\theta_1\cdots d\theta_k\xi_1^{\mu_1}\cdots\xi_1^{\mu_k}\frac{\partial^k}{\partial\theta_1\cdots\partial\theta_k}\phi(\theta_1\cdots\theta_k)\right\}$$
$$=E\left\{\int_0^1\cdots\int_0^1\theta_2\cdots d\theta_k\,\xi_1^{\mu_1}\cdots\xi_1^{\mu_k}\frac{\partial^{(k-1)}}{\partial\theta_2\cdots\partial\theta_k}(\phi(1,\theta_2\cdots\theta_k)-\phi(0,\theta_2\cdots\theta_k))\right\}.$$

Since $\phi(0,\theta_2\dots\theta_k)$ does not depend on $\xi_1^{\mu_1}$, the second term on the rhs of the last relation after averaging with respect to $\xi_1^{\mu_1}$ becomes zero. Repeating this procedure $k-1$ times more, we get

$$E\left\{\int_0^1\cdots\int_0^1 d\theta_1\cdots d\theta_k\xi_1^{\mu_1}\cdots\xi_1^{\mu_k}\frac{\partial^k}{\partial\theta_1\cdots\partial\theta_k}\phi(\theta_1\cdots\theta_k)\right\}=E\{\xi_1^{\mu_1}\cdots\xi_1^{\mu_k}\phi(1,\dots,1)\}.$$

Therefore

$$E\{\xi_1^{\mu_1}\cdots\xi_1^{\mu_k}\phi(1,\dots,1)\}=E\left\{\xi_1^{\mu_1}\cdots\xi_1^{\mu_k}\frac{\partial^k}{\partial\theta_1\cdots\partial\theta_k}\phi(1\cdots1)\right\}+R_{\mu_1\cdots\mu_k}(\phi)$$
$$=N^{k/2}E\left\{\frac{\partial^k}{\partial\xi_1^{\mu_1}\cdots\partial\xi_1^{\mu_k}}\langle t_1^{\mu_1}\cdots t_1^{\mu_k}F_1(\{\sigma\},\{\xi_1^{\mu}\})\rangle_{\mathrm{H}(\overline{1})}\right\}+R_{\mu_1\cdots\mu_k}(\phi), \tag{55}$$

where

$$|R_{\mu_1\cdots\mu_k}(\phi)|\leqslant N^{k/2}E\left\{\max_{|\theta_1|,\dots,|\theta_k|\leqslant1}\frac{\partial^{k+1}}{\partial^2\theta_1\cdots\partial\theta_k}\phi(\theta_1\cdots\theta_k)\right\}+\cdots$$
$$+N^{k/2}E\left\{\max_{|\theta_1|,\dots|\theta_k|\leqslant1}\frac{\partial^{k+1}}{\partial\theta_1\cdots\partial^2\theta_k}\phi(\theta_1\cdots\theta_k)\right\}. \tag{56}$$

Now we should remark that since the Hamiltonian H_*^a contains ξ_1^{μ} only in the form $N^{-1/2}\xi_1^{\mu}\sigma_1(Jt_1^{\mu}+\varepsilon_1\gamma^{\mu})$, after differentiation with respect to ξ_1^{μ} this term will appear in a number of places. Thus, we obtain

$$R_{\mu_1,\dots,\mu_k}\leqslant\mathrm{const}\,E\{N^{-1/2}\langle|t_1^{\mu_1}|(|t_1^{\mu_1}|+\varepsilon_1|\gamma^{\mu_1}|)^2\cdot|t_1^{\mu_2}|(|(t_1^{\mu_2}|+\varepsilon_1|\gamma_{\mu_2}|)\cdots|t_1^{\mu_k}|(|(t_1^{\mu_k}|$$
$$+\varepsilon_1|\gamma_{\mu_k}|)|F_1(\{\sigma\},\{\xi_i^{\mu}\})|\rangle_{\mathrm{H}_*^a+S^1_{\mu_1,\dots,\mu_k}}\}\cdots+E\{N^{-1/2}\langle|t_1^{\mu_1}|(|(t_1^{\mu_1}|+\varepsilon_1|\gamma_{\mu_1}|)$$
$$\cdot|t_1^{\mu_2}|(|(t_1^{\mu_2}|+\varepsilon_1|\gamma_{\mu_2}|)\cdots|t_1^{\mu_k}|(|(t_1^{\mu_k}|+\varepsilon_1|\gamma_{\mu_k}|)^2|F_1(\{\sigma\},\{\xi_i^{\mu}\})|\rangle_{\mathrm{H}_*^a+S^k_{\mu_1,\dots,\mu_k}}\} \tag{57}$$

with $S^i_{\mu_1,\dots,\mu_k}$ of the form

$$S^i_{\mu_1,\ldots,\mu_k}=\frac{\theta_1(\mu_1,\ldots,\mu_k)}{\sqrt{N}}t_1^{\mu_1}\sigma_1+\cdots+\frac{\theta_k(\mu_1,\ldots,\mu_k)}{\sqrt{N}}t_1^{\mu_k}\sigma_1.$$

By using the inequalities $|S^i_{\mu_{1,\ldots,}\mu_k}|\leqslant k$ and then the bound (17), we obtain the estimate

$$\begin{aligned}|R_{\mu_1,\ldots,\mu_1}|&\leqslant \text{const}\, d_N E\{\langle((t_1^{\mu_1})^2+\varepsilon_1^2\gamma_{\mu_1}^2)((t_1^{\mu_2})^2+\varepsilon_1^2\gamma_{\mu_2}^2)\cdots((t_1^{\mu_k})^2\varepsilon_1^2\gamma_{\mu_k}^2)\rangle\}\\ &=\text{const}\, d_N^{1/2}E\left\{p^{-k}\sum_{\mu_1,\ldots,\mu_k=s+1}^{p}\langle((t_1^{\mu_1})^2+\varepsilon_1^2\gamma_{\mu_1}^2)((t_1^{\mu_2})^2+\varepsilon_1^2\gamma_{\mu_2}^2)\cdots((t_1^{\mu_2})^2\varepsilon_1^2\gamma_{\gamma_k}^2)\rangle\right\}\\ &\leqslant \text{const}\, d_N^{1/2}E\{(\alpha\|\mathcal{J}\|+\alpha\varepsilon_1^2)^k\}=O(d_N^{1/2}).\end{aligned}\tag{58}$$

Proposition 1 is proved.

Proof of Lemma 2: It is easy to see that for any **c**

$$\mathrm{H}^a(\mathbf{c})-\mathbf{H}=\frac{J(1+\zeta d_N)N}{2}\sum_{\nu=1}^{s}(m^\nu-c^\nu)^2.$$

Then, on the basis of the Bogolyubov inequality

$$\frac{1}{N}\langle \mathrm{H}_N^{(2)}-\mathrm{H}_N^{(1)}\rangle_{\mathrm{H}_2}\leqslant f_N(\mathrm{H}_2)-f_N(\mathrm{H}_1)\leqslant\frac{1}{N}\langle \mathrm{H}_N^{(2)}-\mathrm{H}_N^{(1)}\rangle_{\mathrm{H}_1},\tag{59}$$

we have for any c^ν

$$\frac{J(1+\zeta d_N)}{2}\sum_{\nu=1}^{s}\langle(m^\nu-c^\nu)^2\rangle_{\mathrm{H}^a(\mathbf{c})}\leqslant f_N(\mathrm{H}^a(\mathbf{c}))-f_N(\mathrm{H})\leqslant\frac{J(1+\zeta d_N)}{2}\sum_{\nu=1}^{s}\langle(m^\nu-c^\nu)^2\rangle.\tag{60}$$

Taking the minimum with respect to all **c** and averaging with respect to all random parameters of the problem, except ζ, we get

$$\begin{aligned}\frac{J(1+\zeta d_N)}{2}E_{/\zeta}\left\{\left\langle\sum_{\nu=1}^{s}(m^\nu-\tilde{c}^\nu)^2\right\rangle_{\mathrm{H}^a(\tilde{\mathbf{c}})}\right\}&\leqslant E_{/\zeta}\{\min_{c^\nu}f_N(\mathrm{H}^a(\mathbf{c}))\}-E_{/\zeta}\{f_N(\mathrm{H})\}\\ &\leqslant\frac{J(1+\zeta d_N)}{2}E_{/\zeta}\left\{\left\langle\sum_{\nu=1}^{s}(m^\nu-\langle m^\nu\rangle)^2\right\rangle\right\},\end{aligned}\tag{61}$$

where $\tilde{\mathbf{c}}$ is a random minimum point of the function $f_N(\mathrm{H}^a(\mathbf{c}))$. Integrating by parts with respect to the variables $\{\gamma^\nu\}_{\nu=1}^{s}$, it is easy to obtain

$$\begin{aligned}E_{/\zeta}\left\{\left\langle\frac{1}{2}\sum_{\nu=1}^{s}(m^\nu-\langle m^\nu\rangle)^2\right\rangle\right\}&=E_{/\zeta}\left\{\frac{1}{2\beta\sqrt{N}}\sum_{\nu=1}^{s}\gamma^\mu\langle m^\nu\rangle\right\}\\ &\leqslant E^{1/2}\left\{\frac{1}{4N}\sum_{\nu=1}^{s}(\gamma^\nu)^2\right\}E^{1/2}\left\{\frac{1}{N}\sum_{\nu=1}^{s}\langle t^\nu\rangle^2\right\}\\ &\leqslant\left(\frac{s}{p}\right)^{1/2}E^{1/2}\{\|\mathcal{J}\|\}.\end{aligned}$$

Substituting this estimate in (61), we get

$$\frac{J(1+\zeta d_N)}{2}E_{/\zeta}\left\{\left\langle \sum_{\nu=1}^{s}(m^\nu-\tilde{c}^\nu)^2\right\rangle_{\mathrm{H}^a(\tilde{\mathbf{c}})}\leqslant E_{/\zeta}\{\min_{c^\nu} f_N(\mathrm{H}^a(\mathbf{c}))\}-E_{/\zeta}\{f_N(\mathrm{H})\}\leqslant \mathrm{const}\left(\frac{s}{p}\right)^{1/2}.\tag{62}$$

Now to prove Lemma 2 we are left to prove that

$$|E_{/\zeta}\{\min_{c^\nu} f_N(\mathrm{H}^a(\mathbf{c}))\}-\min_{c^\nu} E_{/\zeta}\{f_N(\mathrm{H}^a(\mathbf{c}))\}|\leqslant o(1)\quad \text{as}\ \ N\to\infty.\tag{63}$$

Since evidently for fixed N $f_N(\mathrm{H}^a(\mathbf{c}))$ and $E_{/\zeta}\{f_N(\mathrm{H}^a(\mathbf{c}))\}$ tend to infinity, as $\mathbf{c}\to\infty$, these functions take their minimal values at the finite extremal points. But the extremal conditions for these functions have the form $c^\nu=\langle m^\nu\rangle_{\mathrm{H}^a(\mathbf{c})}$ and $c^\nu=E_{/\zeta}\{\langle m^\nu\rangle_{\mathrm{H}^a(\mathbf{c})}\}$ and since $|m^\nu|\leqslant 1$ to prove (63) it is enough to prove that

$$\mathrm{Prob}\{X\}\equiv\mathrm{Prob}\left\{\sup_{|c^\nu|\leqslant 1}|f_N(\mathrm{H}^a(\mathbf{c}))-E_{/\zeta}\{f_N(\mathrm{H}^a(\mathbf{c}))\}|\leqslant\frac{2J}{\log N}\right\}\leqslant o(1).\tag{64}$$

But, using once more the fact that the derivatives of the functions $f_N(\mathrm{H}^a(\mathbf{c}))$ and $E_{/\zeta}\{f_N(\mathrm{H}^a(\mathbf{c}))\}$ are bounded, we obtain that

$$|f_N(\mathrm{H}^a(\mathbf{c}_1))-f_N(\mathrm{H}^a(\mathbf{c}_2))|\leqslant J\sqrt{s}\left(\sum (c_1^\nu-c_2^\nu)^2\right)^{1/2}$$

and so, if $|c_1^\nu-c_2^\nu|\leqslant(2k)^{-1}(k\equiv[s\log N])$, then

$$|f_N(\mathrm{H}^a(\mathbf{c}_1))-f_N(\mathrm{H}^a(\mathbf{c}_2))|\leqslant\frac{Js}{2k}\leqslant\frac{J}{2\log N}.$$

Therefore

$$\sup_{|c^\nu|\leqslant 1}|f_N(\mathrm{H}^a(\mathbf{c}))-E_{/\zeta}\{f_N(\mathrm{H}^a(\mathbf{c}))\}|\leqslant\sup_{|j_1|,\dots,|j_s|\leqslant k}\left|f_N\left(\mathrm{H}^a\left(\frac{j_1}{k},\dots,\frac{j_s}{k}\right)\right)-E_{/\zeta}\left\{f_N\left(\mathrm{H}^a\left(\frac{j_1}{k},\dots,\frac{j_s}{k}\right)\right)\right\}\right|+\frac{J}{\log N}.$$

Thus,

$$\mathrm{Prob}\{X\}\leqslant\sum_{j_1,\dots,j_s}\mathrm{Prob}\{X_{j_1,\dots,j_s}\},\tag{65}$$

where $j_1,\dots.j_s$ are integer numbers from the interval $(-k,k)$ and $X_{j_1,\dots,j_s}$ is the notation of the event

$$\left|f_N\left(\mathrm{H}^a\left(\frac{j_\perp}{k},\dots,\frac{j_s}{k}\right)\right)-E_{/\zeta}\left\{f_N\left(\left(\frac{j_1}{k},\dots,\frac{j_s}{k}\right)\right)\right\}\right|\geqslant\frac{J}{\log N}.$$

According to the result of Ref. 14, the probability of the last event can be estimated by $D_1N^{-1}\log^2 N$. Thus, on the basis of (65),

$$\begin{aligned}\mathrm{Prob}\{X\}\leqslant D_1N^{-1}(2k+1)^s&\leqslant D_1\exp\{s\log(2k+1)+2\log\log N-\log N\}\\&=D_1\exp\{\mathrm{const}[\log N]^{1/2}(\log\log N)-\log N\}\\&\leqslant N^{-1/2},\end{aligned}$$

and we obtain (63) which joined with (62), proves (24).

Now let us prove relations (25). To this end we use

Lemma 4: Consider the sequence of convex random functions $\{f_n(t)\}_{n=1}^{\infty}(f_n''(t)\geqslant 0)$ *in the interval* (a,b). *If functions* f_n *are self-averaging, i.e., uniformly in t*

$$\lim_{n\to\infty} E\{(f_n(t)-E\{f_n(t)\})^2\}=0,$$

and bounded $(|E\{f_n(t)\}|\leqslant C$ *uniformly in* $n,t\in(a,b))$, *then for almost all t,*

$$\lim_{n\to\infty} E\{[f_n'(t)-E\{f_n'(t)\}]^2\}=0, \tag{66}$$

i.e., the derivatives $f_n'(t)$ *are also self-averaging ones for almost all t.*

In addition, if we consider another sequence of convex functions $\{g_n(t)\}_{n=1}^{\infty}(g_n''\geqslant 0)$ *which are also self-averaging uniformly in t,*

$$\lim_{n\to\infty} E\{(g_n(t)-E\{g_n(t)\})^2\}=0$$

and

$$\lim_{n\to\infty} |E\{f_n(t)\}-E\{g_n(t)\}|=0, \tag{67}$$

uniformly in t, then for all t which satisfy (66),

$$\lim_{n\to\infty} |E\{f_n'(t)\}-E\{g_n'(t)\}|=0, \quad \lim_{n\to\infty} E\{[g_n'(t)-E\{g'(t)\}]^2\}=0. \tag{68}$$

We prove Lemma 4 in the Appendix. Now let us note that, using this lemma and (24), we obtain

$$\beta E_{/\zeta}\{(1-q_N^*)\}=-E_{/\zeta}\left\{\frac{\partial f_N(\mathrm{H}^a)}{\partial\varepsilon_2}\right\}=-E_{/\zeta}\left\{\frac{\partial f_N(\mathrm{H})}{\partial\varepsilon_2}\right\}+o(1)=\beta E_{/\zeta}\{(1-q_N)\}+o(1),$$

$$\beta E_{/\zeta}\{(\langle U\rangle_*-\alpha r_N^*)\}=-E_{/\zeta}\left\{\frac{\partial f_N(\mathrm{H}^a)}{\partial\varepsilon_1}\right\}=-E\left\{\frac{\partial f_N(\mathrm{H})}{\partial\varepsilon_1}\right\}+o(1)=\beta E_{/\zeta}\{(\langle U\rangle-\alpha r_N)\}+o(1), \tag{69}$$

$$E_{/\zeta}\{\langle U_N\rangle_*\}+\sum_{\mu=1}^{s}\int_1^2 d\zeta(c_*^{\nu}(\zeta))^2$$
$$=-2E_{/\zeta}\left\{\frac{\partial f_N(\mathrm{H}^a)}{\partial J}\right\}=-2E_{/\zeta}\left\{\frac{\partial f_N(\mathrm{H})}{\partial J}\right\}+o(1)=E\left\{\langle U_N\rangle+\sum_{\mu=1}^{s}\langle(m^{\mu})^2\rangle\right\}+o(1),$$

Remark also that

$$\sum_{\mu=1}^{s}\int_1^2 d\zeta(_*^{\nu}(\zeta)^2)=-2\,\frac{1}{d_N}\int_1^2 d\zeta\, E_{/\zeta}\left\{\frac{\partial f_N(\mathrm{H}^a)}{\partial\zeta}\right\}$$
$$=-2\,\frac{1}{d_N}E_{/\zeta}\{(f_N(\mathrm{H}^a)|_{\zeta=2}-f_N(\mathrm{H}^a)|_{\zeta=1})\}$$
$$=-2\,\frac{1}{d_N}E_{/\zeta}\{(f_N(\mathrm{H})|_{\zeta=2}-f_N(\mathrm{H})|_{\zeta=1})\}+o(1)$$
$$=-2\,\frac{1}{d_N}\int_1^2 d\zeta\, E_{/\zeta}\left\{\frac{\partial f_N(\mathrm{H})}{\partial\zeta}\right\}+O\left(\frac{1}{\log Nd_N}\right)=E\left\{\sum_{\mu=1}^{s}\langle(m^{\mu})^2\rangle\right\}+o(1).$$

These relations, joined to (69), prove (25).

APPENDIX: PROOF OF PROPOSITION 2 AND LEMMA 4

Proof of Proposition 2: Due to the symmetry of the problem we can restrict ourselves to the case when x and y are positive.

If $x>0$ and $y>x+2\delta$, consider $x'=y$ and $y'=x$. Then

$$F(x',y')-F(x,y)=f(y,x)-\frac{y^2}{2(1+d)}-\frac{x^2}{2}-f(x,y)+\frac{x^2}{2(1+d)}+\frac{y^2}{2}$$
$$=\frac{d}{2(1+d)}(y^2-x^2)$$
$$\geqslant 4\delta^2\frac{d}{2(1+d)}. \tag{A1}$$

Proposition 2 is proved.

Proof of Lemma 4: Denote

$$d_n^{(1)}\equiv \max_{a\leqslant t\leqslant b} E\{[f_n(t)-E\{f_n(t)\}]^2\},\quad d_n^{(2)}\equiv \max_{a\leqslant t\leqslant b} E\{[g_n(t)-E\{g_n(t)\}]^2\},$$

$$d_n^{(3)}\equiv \max_{a\leqslant t\leqslant b} |E\{f_n(t)\}-E\{g_n(t)\}|,\quad \varepsilon_n=[\max\{d_n^{(1)},d_n^{(2)},d_n^{(3)}\}]^{1/3}.$$

Then, using the convexity of $f_n(t)$, we have

$$f_n'(t)-E\{f_n'(t)\}\leqslant\frac{f_n(t+\varepsilon_n)-f_n(t)}{\varepsilon_n}-E\{f_n'(t)\}$$
$$=-\frac{f_n(t)-E\{f_n(t)\}}{\varepsilon_n}+\frac{f_n(t+\varepsilon_n)-E\{f_n(t+\varepsilon_n)\}}{\varepsilon_n}$$
$$+\left[\frac{E\{f_n(t+\varepsilon_n)\}-E\{f_n(t)\}}{\varepsilon_n}-E\{f_n'(t)\}\right]. \tag{A2}$$

Denote also

$$R_n^+(t)\equiv\frac{E\{f_n(t+\varepsilon_n)\}-E\{f_n(t)\}}{\varepsilon_n}-E\{f_n'(t)\}$$

and prove that $R_n^+(t)\to 0$ for almost all t. To this end we study

$$\int_{a_1}^{b_1}R_n^+(t)dt=\left[\frac{F_n(b_1+\varepsilon_n)-F_n(b_1)}{\varepsilon_n}-F_n'(b_1)\right]-\left[\frac{F_n(a_1+\varepsilon_n)-F_n(a_1)}{\varepsilon_n}-F_n'(a_1)\right], \tag{A3}$$

where $F_n(t)\equiv\int_a^t E\{f_n(\tau)\}\,d\tau$ and (a_1,b_1) is some subinterval of (a,b). It is evident that

$$0\leqslant F_n''(t)=E\{f_n'(t)\}\leqslant 2(b-b_1)^{-1}\left[E\{f_n(b)\}-E\left\{f_n\left(\frac{b+b_1}{2}\right)\right\}\right].$$

Therefore if n is large enough to provide $\varepsilon_n\leqslant(b-b_1)/2$, then, according to the Taylor formula, the rhs of (A3) is of order $O(\varepsilon_n)$. Thus, since $R_n^+(t)\geqslant 0$, it follows from (A3) that $R_n^+(t)\to 0$ for almost all t.

On the other hand, similarly to (A2) we get that

$$f_n'(t)-E\{f_n'(t)\}\geqslant-\frac{f_n(t-\varepsilon_n)-E\{f_n(t-\varepsilon_n)\}}{\varepsilon_n}+\frac{f_n(t)-E\{f_n(t)\}}{\varepsilon_n}+R_n^-(t)$$

and $R_n^-(t)\equiv E\{f_n(t)\}-E\{f_n(t-\varepsilon_n)\}/\varepsilon_n-E\{f_n'(t)\}\to 0$ for almost all t. Then

$$\begin{aligned}E\{[f_n'(t)-E\{f'(t)\}]^2\}\leqslant 2E\left\{\left(\frac{f_n(t+\varepsilon_n)-E\{f_n(t+\varepsilon_n)\}}{\varepsilon_n}\right)^2\right\}\\+2E\left\{\left(\frac{f_n(t-\varepsilon_n)-E\{f_n(t-\varepsilon_n)\}}{\varepsilon_n}\right)^2\right\}+4E\left\{\left(\frac{f_n(t)-E\{f_n(t)\}}{\varepsilon_n}\right)^2\right\}\\+(R_n^-(t))^2+(R_n^+(t))^2\leqslant 8\varepsilon_n+(R_n^-(t))+(R_n^+(t))^2\to 0\end{aligned}$$

for almost all t.

By the same way,

$$\begin{aligned}E\{[g_n'(t)-E\{f_n'(t)\}]^2\}&\leqslant 2E\left\{\left(\frac{g_n(t+\varepsilon_n)-E\{f_n(t+\varepsilon_n)\}}{\varepsilon_n}\right)^2\right\}\\&\quad+2E\left\{\left(\frac{g_n(t-\varepsilon_n)-E\{f_n(t-\varepsilon_n)\}}{\varepsilon_n}\right)^2\right\}+4E\left\{\left(\frac{g_n(t)-E\{f_n(t)\}}{\varepsilon_n}\right)^2\right\}\\&\quad+(R_n^-(t))^2+(R_n^+(t))^2\\&=2E\left\{\left(\frac{g_n(t+\varepsilon_n)-E\{g_n(t+\varepsilon_n)\}}{\varepsilon_n}\right)^2\right\}\\&\quad+2E\left\{\left(\frac{g_n(t-\varepsilon_n)-E\{g_n(t-\varepsilon_n)\}}{\varepsilon_n}\right)^2\right\}+4E\left\{\left(\frac{g_n(t)-E\{g_n(t)\}}{\varepsilon_n}\right)^2\right\}\\&\quad+2\left(\frac{E\{g_n(t+\varepsilon_n)\}-E\{f_n(t+\varepsilon_n)\}}{\varepsilon_n}\right)^2\\&\quad+2\left(\frac{E\{g_n(t-\varepsilon_n)\}-E\{f_n(t-\varepsilon_n)\}}{\varepsilon_n}\right)^2+4\left(\frac{E\{g_n(t)\}-E\{f_n(t)\}}{\varepsilon_n}\right)^2\\&\quad+(R_n^-(t))^2+(R_n^+(t))^2\\&\leqslant 16\varepsilon_n+(R_n^-(t))^2+(R_n^+(t))^2.\end{aligned}\qquad\text{(A4)}$$

On the other hand,

$$E\{[g_n'(t)-E\{f_n'(t)\}]^2\}=E\{[g_n'(t)-E\{g_n'(t)\}]^2\}+[E\{g_n'(t)\}-E\{f_n'(t)\}]^2.$$

Therefore (A4) proves (67) and (68).

[1] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[2] M. Aizemann, J. L. Lebowitz, and D. Ruelle, ''Some rigorous results on the Sherrington–Kirkpatrick spin glass model,'' Commun. Math. Phys. **112**, 3–20 (1987).
[3] E. Scacciatelli and B. Tirozzi, ''Fluctuation of the free energy in the Hopfield model,'' J. Stat. Phys. **67**, 981–1008 (1991).
[4] L. Pastur and M. Shcherbina, ''Absence of self-averaging of the order parameter in the Sherrington-Kirkpatrick model,'' J. Stat. Phys. **62**, 1–26 (1991).
[5] L. Pastur, M. Shcherbina, and B. Tirozzi, ''The replica-symmetric solution without replica trick for the Hopfield model,'' J. Stat. Phys. **74(5/6)**, 1161–1183 (1994).
[6] J. Hopfield, ''Neural networks and physical systems with emergent collective computational abilities,'' Physica A **79**, 2554–2558 (1982).
[7] D. Sherrington and S. Kirkpatrick, ''Infinite models of spin glasses,'' Phys. Rev. B **17**, 4384–4403 (1978).

[8] M. Shcherbina, ''More About the Absence of Selfaveraging of the Order Parameter for the Sherrington-Kirkpatrick Model,'' CARR Preprint 3/91, Rome (1991).
[9] M. Shcherbina, ''On the replica symmetric solution for the Sherrington-Kirkpatrick model,'' Helv. Phys. Acta **70**, 772–797 (1997).
[10] M. Talagrand, ''Rigorous results for the Hopfield models with many patterns,'' Prob. Theor. Rel. Fields **110**, 109–176 (1998).
[11] M. Talagrand, ''The Sherrington Kirkpatrick model: A challenge to mathematicians,'' Prob. Theor. Rel. Fields **110**, 177–286 (1998).
[12] A. Bovier and V. Gayrard, ''An almost sure large deviation principle for the Hopfield model,'' Ann. Prob. **24**, 1444–1475 (1997).
[13] M. Shcherbina and B. Tirozzi, ''The free energy of a class of Hopfield models,'' J. Stat. Phys. **72**, 113–125 (1993).
[14] J. Feng and B. Tirozzi, ''The SLLN for free-energy of a class of neural works,'' Helv. Phys. Acta **68**, 367–379 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Relation between the Kadometsev–Petviashvili equation and the confocal involutive system

Cewen Cao
Department of Mathematics, Zhengzhou University, Zhengzhou, Henan 450052, People's Republic of China

Yongtang Wu
Department of Computer Science, Hong Kong Baptist University, 224 Waterloo Road, Kowloon, Hong Kong

Xianguo Geng
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China and Department of Mathematics, Zhengzhou University, Zhengzhou, Henan 450052, People's Republic of China

(Received 23 December 1998; accepted for publication 1 April 1999)

The special quasiperiodic solution of the $(2+1)$-dimensional Kadometsev–Petviashvili equation is separated into three systems of ordinary differential equations, which are the second, third, and fourth members in the well-known confocal involutive hierarchy associated with the nonlinearized Zakharov–Shabat eigenvalue problem. The explicit theta function solution of the Kadometsev–Petviashvili equation is obtained with the help of this separation technique. A generating function approach is introduced to prove the involutivity and the functional independence of the conserved integrals which are essential for the Liouville integrability.
 [S0022-2488(99)04207-3]

I. INTRODUCTION

The confocal involutive system was first used systematically by J. Moser[1] in the investigation of soliton equations such as the Korteweg–de Vries (KdV) equation. Various new kinds of confocal involutive systems are obtained by the approach of nonlinearization of eigenvalue problems[2–4] or constrained flows,[5,6] which are useful tools in the study of $(1+1)$-dimensional soliton equations. The mechanism of the technique is that the finite-parametric solution (such as the finite-band solution) to a soliton equation could be separated into two systems of ordinary differential equations (ODEs), which are two members in the associated confocal involutive system. There are at least two useful applications. First, it provides an effective way in the numerical analysis and the graphic presentation of solutions, where only numerical solution of ODEs and algebraic calculations are concerned. As examples on the Dirac-Nonlinear Schrödinger (NS) system and the Toda lattice, see Refs. 7 and 8. Second, it provides another way to get the explicit expressions of quasiperiodic solutions by means of the Riemann theta functions resorting to the elliptic variables and the Abel–Jacobi coordinates, in contrast with that of the Baker function approach.[9,10]

In the present paper, we are going to extend the $(1+1)$ framework to the Kadometsev–Petviashvili (KP) equation, one of the typical $(2+1)$-dimensional soliton equations:

$$w_t=\tfrac{1}{4}(w_{xx}+3w^2)_x+\tfrac{3}{4}\partial_x^{-1}w_{yy}\,. \tag{1.1}$$

It is well-known that Eq. (1.1) is the compatible condition of two overdetermined linear equations, the Lax pair. Together with its conjugate version, the Lax pair is nonlinearized into the coupled nonlinear Schrödinger (NS) equation and the coupled modified Korteweg-de Vries (mKdV) equation, which are the second and the third members, respectively (the X_2 flow and the X_3 flow) in the Ablowitz–Kaup–Newell–Segur (AKNS) hierarchy.[11–13]

The AKNS hierarchy is the isospectral class of the Zhakharov–Shabat (ZS) eigenvalue problem. In order to separate the X_2 and X_3 flows, consider the nonlinearized ZS eigenvalue problem, which is completely integrable with a series of involutive integrals $\{F_k\}$[3]. According to the terminology suggested by Moser,[1] the integrals are called confocal since their generating function has the form:

$$\sum_{k=0}^{\infty} \frac{F_k}{\lambda^{k+1}} = \sum_{j=1}^{N} \frac{E_j}{\lambda - \alpha_j},$$

which gives confocal quadrics in $\mathbb{R}^N$. The coupled NS equation (the X_2 flow) has a Lax pair, which is nonlinearized into the H_1 flow and H_2 flow; while the Lax pair of the coupled mKdV equation (the X_3 flow) is nonlinearized into the H_1 flow and H_3 flow. It turns out that $\{H_k\}$ is essentially the confocal involutive system $\{F_k\}$ with some modifications (see Sec. V below). Thus the special solution of the KP equation is separated into three members in the confocal involutive system:

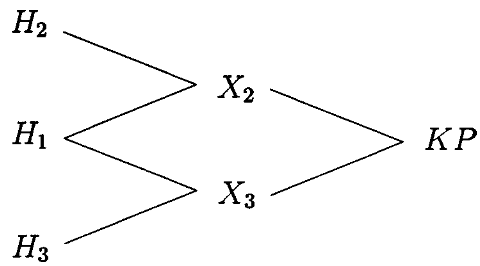

Specifically, let *x, y, t* be the variables of the H_1, H_2, and H_3 flow, and $(p(x,y,t),q(x,y,t))$ be a compatible solution of the flows. Then

$$w(x,y,t) = 2\langle p,p\rangle\langle q,q\rangle = 2\langle Ap,q\rangle + 4H_1 \tag{1.2}$$

solves the KP equation (1.1) (see Sec. VII below).

A deeper understanding of the nonlinearized ZS eigenvalue problem is necessary in realizing the above framework. The Liouville integrability plays a central role in the whole structure. We have introduced an effective way, the generating function flow method, to prove:

(i) the involutivity of $\{F_k\}$;
(ii) the functional independence of $F_0,\ldots,F_{N-1}$;

which are essential for the Liouville integrability. The functional independence means a sufficient number of integrals, which is important and usually difficult to verify. Some results are obtained by Ma *et al.*[14,15] It should be noted that the well-known three-body problem was proved nonintegrable by Bruns[16,17] since it has an insufficient number of integrals. The generating function flow method is also effective in calculating the evolution equations of the Abel–Jacobi coordinates, which give a clear dynamic picture of the KP flow, X_k flow, and H_k flow (see Sec. VIII below). The explicit expression of the finite-band potential and the solution to the KP equation by means of the Riemann theta function are obtained as a consequence (see Sec. X and XI).

II. PRELIMINARIES

The AKNS hierarchy of soliton equations is the isospectral class of the Zakharov–Shabat eigevalue problem:

$$\chi_x = U\chi, \quad U = \begin{pmatrix} \frac{\lambda}{2} & u \\ v & -\frac{\lambda}{2} \end{pmatrix}. \tag{2.1}$$

As the basis of investigating the zero-curvature form of the AKNS equation:

$$V_x-[U,V]=U_t\equiv U_*\begin{pmatrix}u_t\\ v_t\end{pmatrix},$$

we have the fundamental identity:

$$V_x-[U,V]=U_*\{P(K-\lambda J)\gamma\},\tag{2.2}$$

where U_* is the differential of the map $(u,v)^T\to U(u,v)$, P is the projective map $\gamma=(\gamma^1,\gamma^2,\gamma^3)^T\to(\gamma^1,\gamma^2)^T$, and $(\partial=\partial_x)$

$$V=\begin{pmatrix}\gamma_3 & \gamma_2\\ \gamma_1 & -\gamma_3\end{pmatrix},\quad K=\begin{pmatrix}0 & \partial & 2u\\ \partial & 0 & -2v\\ -u & v & \partial\end{pmatrix},\quad J=\begin{pmatrix}0 & 1 & 0\\ -1 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}.\tag{2.3}$$

The Lenard gradients $\{g_j\}$ and the AKNS vector fields $\{X_j\}$ are defined recursively by

$$\begin{aligned}&Kg_j=Jg_{j+1},\quad Jg_{-1}=0,\\ &X_j=PJg_j=(g_j^2,-g_j^1)^T.\end{aligned}\tag{2.4}$$

The explicit recursive formulas for the first two components g_j^1, g_j^2 are evident. In order to eliminate the ambiguity in determining the third component g_j^3 since the existence of ∂^{-1}, consider

$$g_\lambda=\sum_{j=0}^{\infty}g_{j-1}\lambda^{-j},\quad g_{-1}=(0,0,\tfrac{1}{2})^T,\tag{2.5}$$

which satisfies $(K-\lambda J)g_\lambda=0$. By Eq. (2.2), $V=\sigma(g_\lambda)$ solves the stationary Lax equation $V_x-[U,V]=0$, thus $\det\sigma(g_\lambda)$ is a constant along the x flow. We adopt:

$$\det\sigma(g_\lambda)=-g_\lambda^1g_\lambda^2-(g_\lambda^3)^2=-\tfrac{1}{4}.\tag{2.6}$$

This implies the explicit recursive formula for g_j^3 by comparing the coefficients of λ^{-s}. Thus we have

$$\begin{aligned}&g_{m+1}^1=-\partial g_m^1+2vg_m^3,\\ &g_{m+1}^2=\partial g_m^2+2ug_m^3,\\ &g_{m+1}^3=-\sum_{\substack{j+k=m\\ j,k\geqslant 0}}g_j^1g_k^2+g_j^3g_k^3.\end{aligned}\tag{2.7}$$

The first few members are:

$$g_{-1}=\begin{pmatrix}0\\ 0\\ \frac{1}{2}\end{pmatrix},\quad g_0=\begin{pmatrix}v\\ u\\ 0\end{pmatrix},\quad g_0=\begin{pmatrix}-v_x\\ u_x\\ -uv\end{pmatrix},$$

$$g_2=\begin{pmatrix} v_{xx}-2uv^2 \\ u_{xx}-2u^2v \\ uv_x-u_xv \end{pmatrix}, \quad g_3=\begin{pmatrix} -v_{xxx}+6uvv_x \\ u_{xxx}-6uvu_x \\ -u_{xx}v-uv_{xx}+u_xv_x+3u^2v^2 \end{pmatrix}, \tag{2.8}$$

$$X_0=\begin{pmatrix} u \\ -v \end{pmatrix}, \quad X_1=\begin{pmatrix} u_x \\ v_x \end{pmatrix}, \quad X_2=\begin{pmatrix} u_{xx}-2u^2v \\ -v_{xx}+2uv^2 \end{pmatrix}, \quad X_3=\begin{pmatrix} u_{xxx}-6uvu_x \\ v_{xxx}-6uvv_x \end{pmatrix}.$$

Using the bi-Hamiltonian structure $Jg_4=Kg_3$, we have

$$X_4=\begin{pmatrix} u_{xxxx}-8uvu_{xx}-6u_x^2v-4uu_xv_x-2u^2v_{xx}+6u^3v^2 \\ -v_{xxxx}+8uvv_{xx}+6uv_x^2+4vu_xv_x+2v^2u_{xx}-6u^2v^3 \end{pmatrix}.$$

III. THE ZS–BARGMANN SYSTEM

Consider N copies of the linear ZS equation (2.1):

$$\partial_x\begin{pmatrix} p_j \\ q_j \end{pmatrix}=\begin{pmatrix} \frac{1}{2}\alpha_j & u \\ v & -\frac{1}{2}\alpha_j \end{pmatrix}\begin{pmatrix} p_j \\ q_j \end{pmatrix}, \quad j=1,...,N, \tag{3.1}$$

with distinct eigenvalues $\lambda=\alpha_1,\dots\alpha_N$. Let $A=\mathrm{diag}(\alpha_1,...,\alpha_N)$. It is proved in Ref. 3 that the reflectionless potential is expressed by the squared sum of eigenfunctions:

$$\begin{aligned} u(x)&=-\langle p,p\rangle=-\sum_{j=1}^{N}\{p_j(x)\}^2, \\ v(x)&=\langle q,q\rangle=\sum_{j=1}^{N}\{q_j(x)\}^2. \end{aligned} \tag{3.2}$$

Thus in the reflectionless case, what the solution $(p_j,q_j)^T$ of the linear Eq. (3.1) satisfies, is actually a system of nonlinear equations:

$$\begin{aligned} p_x&=\frac{1}{2}Ap-\langle p,p\rangle q=-\frac{\partial H_1}{\partial q}, \\ q_x&=-\frac{1}{2}Aq+\langle q,q\rangle p=\frac{\partial H_1}{\partial p}, \\ H_1&=-\tfrac{1}{2}\langle Ap,q\rangle+\tfrac{1}{2}\langle p,p\rangle\langle q,q\rangle. \end{aligned} \tag{3.3}$$

This procedure is called nonlinearization of the eigenvalue problem, which is developed to a general approach to produce new finite-dimensional integrable systems from infinite-dimensional ones (soliton hierarchies).[2–4]

Now consider the problem of integrability of the ZS–Bargmann system (3.3). The functional gradient

$$\nabla\alpha_j=\begin{pmatrix} \delta\alpha_j/\delta u \\ \delta\alpha_j/\delta v \end{pmatrix}=\begin{pmatrix} q_j^2 \\ -p_j^2 \end{pmatrix}$$

is extended to $\nabla\alpha_j=(q_j^2,-p_j^2,p_jq_j)^T$, which satisfies the Lenard eigenvalue problem $(K-\alpha_jJ)\nabla\alpha_j=0$. The condition (3.2) is put in the general form (the Bargmann constraint):

$$g_0 + cg_{-1} = \sum_{j=1}^{N} \nabla \alpha_j, \tag{3.4}$$

where the third component $\langle p,q\rangle = c/2$ gives nothing essential since $\langle p,q\rangle$ is a conserved integral of the x flow. The solution of the Lenard eigenvalue problem with general parameter λ is constructed as

$$G_\lambda = g_{-1} + \sum_{j=1}^{N} \frac{1}{\lambda - \alpha_j} \nabla \alpha_j, \tag{3.5}$$

$$(K - \lambda J) G_\lambda = 0. \tag{3.6}$$

Actually,

$$(K - \lambda J) G_\lambda = J\left(g_0 - \sum_{j=1}^{N} \nabla \alpha_j \right) = J(cg_{-1}) = 0.$$

By the fundamental identity (2.2), the Lax equation along the x flow

$$V_x - [U,V] = 0 \tag{3.7}$$

has a solution $V_\lambda = \sigma(G_\lambda)$:

$$V_\lambda = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix} + \sum_{j=1}^{N} \frac{\epsilon_j}{\lambda - \alpha_j}, \quad \sigma(\nabla \alpha_j) = \epsilon_j = \begin{pmatrix} p_j q_j & -p_j^2 \\ q_j^2 & -p_j q_j \end{pmatrix}, \tag{3.8}$$

which is called the Lax matrix of the ZS–Bargmann system (3.3). Equation (3.7) implies that $F_\lambda = \det V_\lambda$ is invariant along the x flow. Therefore, we have the generating function of integrals of Eq. (3.3),

$$F_\lambda = -\frac{1}{4} - Q_\lambda(p,q) + Q_\lambda(p,p) Q_\lambda(q,q) - Q_\lambda^2(p,q) = -\frac{1}{4} + \sum_{k=0}^{\infty} \lambda^{-k-1} F_k, \tag{3.9}$$

where $Q_\lambda(\xi,\eta) = \langle (\lambda E - A)^{-1} \xi, \eta \rangle$, and

$$F_0 = -\langle p,q\rangle,$$

$$F_1 = -\langle Ap,q\rangle + \langle p,p\rangle\langle q,q\rangle - \langle p,q\rangle^2, \tag{3.10}$$

$$F_k = -\langle A^k p,q\rangle + \sum_{i+j=k-1} \langle A^i p,p\rangle\langle A^j q.q\rangle - \langle A^i p,q\rangle\langle A^j p,q\rangle.$$

Specifically,

$$H_1 = \tfrac{1}{2} F_1 + \tfrac{1}{2} F_0^2.$$

Regard the generating function F_λ as a Hamiltonian in the symplectic space $(\mathbb{R}^{2N}, dp \wedge dq)$. The canonical equation is

$$\frac{d}{dt_\lambda} \begin{pmatrix} p_k \\ q_k \end{pmatrix} = \frac{2}{\lambda - \alpha_k} V_\lambda \begin{pmatrix} p_k \\ q_k \end{pmatrix}, \quad k = 1,\ldots,N. \tag{3.11}$$

Proposition 3.1: The Lax matrix V_μ satisfies the Lax equation along the F_λ flow:

$$\frac{d}{dt_\lambda} V_\mu = \frac{2}{\lambda-\mu}[V_\lambda, V_\mu], \quad \forall \lambda, \mu \in \mathbb{C}. \tag{3.12}$$

Proof: Since

$$\epsilon_k = \begin{pmatrix} p_k \\ q_k \end{pmatrix} (p_k, q_k) \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

by Eq. (3.11) we have

$$\frac{d}{dt_\lambda}\epsilon_k = \frac{2}{\lambda-\alpha_k}[V_\lambda, \epsilon_k].$$

$$\begin{aligned}\frac{d}{dt_\lambda} V_\mu &= \sum_{k=1}^{N} \frac{1}{\mu-\alpha_k} \frac{2}{\lambda-\alpha_k}[V_\lambda, \epsilon_k] \\ &= \frac{2}{\lambda-\mu}\sum_{k=1}^{N}\left(\frac{1}{\mu-\alpha_k} - \frac{1}{\lambda-\alpha_k}\right)[V_\lambda, \epsilon_k] \\ &= \frac{2}{\lambda-\mu}[V_\lambda, V_\mu - V_\lambda] = \mathrm{RHS}.\end{aligned}$$

Corollary 3.2: $(F_\mu, F_\lambda) = 0, \quad \forall \lambda, \mu \in \mathbb{C};$

$$(F_j, F_k) = 0, \quad \forall j,k = 0,1,2,\ldots .$$

Proof: Equation (3.12) implies that $F_\mu = \det V_\mu$ is invariant along the t_λ flow. And the derivative of the function F_μ along the F_λ flow is exactly the Poisson bracket.

IV. ELLIPTIC COORDINATES AND FUNCTIONAL INDEPENDENCE

It is easy to see that each one of F_λ, V_λ^{12}, and V_λ^{21}, as a rational function of λ, has simple poles at α_j's, since the coefficient of $(\lambda-\alpha_j)^{-2}$ is zero in F_λ. We have

$$\begin{aligned} F_\lambda &= Q_\lambda(p,p)Q_\lambda(q,q) - \left\{Q_\lambda(p,q) + \frac{1}{2}\right\}^2 = -\frac{b(\lambda)}{4a(\lambda)}, \\ V_\lambda^{12} &= -Q_\lambda(p,p) = -\langle p,p\rangle \frac{m(\lambda)}{a(\lambda)}, \\ V_\lambda^{21} &= Q_\lambda(q,q) = \langle q,q\rangle \frac{n(\lambda)}{a(\lambda)}, \end{aligned} \tag{4.1}$$

where

$$a(\lambda) = \prod_{k=1}^{N} (\lambda-\alpha_k), \quad b(\lambda) = \prod_{k=1}^{N} (\lambda-\beta_k),$$

$$m(\lambda) = \prod_{k=1}^{N-1} (\lambda-\mu_k), \quad n(\lambda) = \prod_{k=1}^{N-1} (\lambda-\nu_k).$$

The roots $\{\mu_k\}$ and $\{\nu_k\}$ are defined as elliptic variables.

Proposition 4.1: The elliptic coordinates satisfy the evolution equations along the t_λ flow:

$$\frac{1}{2\sqrt{R(\mu_k)}}\frac{d\mu_k}{dt_\lambda}=\frac{m(\lambda)}{a(\lambda)(\lambda-\mu_k)m'(\mu_k)},$$
$$\frac{1}{2\sqrt{R(\nu_k)}}\frac{d\nu_k}{dt_\lambda}=\frac{-n(\lambda)}{a(\lambda)(\lambda-\nu_k)n'(\nu_k)}, \tag{4.2}$$

where

$$R(\lambda)=a(\lambda)b(\lambda)=\prod_{k=1}^{2N}(\lambda-\lambda_k), \tag{4.3}$$

with $\lambda_k=\alpha_k$, $\lambda_{N+k}=\beta_k$, $(k=1,...,N)$

Proof: Substitute $\lambda=\mu_k$, ν_k, respectively, in Eq. (4.1). We have

$$V^{11}_{\mu_k}=\frac{\sqrt{R(\mu_k)}}{2a(\mu_k)},\quad V^{11}_{\nu_k}=\frac{\sqrt{R(\nu_k)}}{2a(\nu_k)}.$$

In the second and third components of Eq. (3.12),

$$\frac{d}{dt_\lambda}V^{12}_\mu=\frac{-4}{\lambda-\mu}(V^{12}_\lambda V^{11}_\mu-V^{11}_\lambda V^{12}_\mu),$$

$$\frac{d}{dt_\lambda}V^{21}_\mu=\frac{4}{\lambda-\mu}(V^{21}_\lambda V^{11}_\mu-V^{11}_\lambda V^{21}_\mu),$$

let $\mu=\mu_k$ and $\mu=\nu_k$, respectively. After some calculations we have Eq. (4.2).

By means of the interpolation formula for polynomials with degree not more than $g-1=N-2$, we have $(s=1,...,g)$

$$\sum_{k=1}^{g}\frac{\mu_k^{g-s}}{2\sqrt{R(\mu_k)}}\frac{d\mu_k}{dt_\lambda}=\frac{\lambda^{g-s}}{a(\lambda)},$$

$$\sum_{k=1}^{g}\frac{\nu_k^{g-s}}{2\sqrt{R(\nu_k)}}\frac{d\nu_k}{dt_\lambda}=\frac{-\lambda^{g-s}}{a(\lambda)}.$$

For fixed λ_0, introduce the quasi-Abel–Jacobi coordinates:

$$\tilde{\phi}_s=\sum_{k=1}^{g}\int_{\lambda_0}^{\mu_k}\frac{\mu^{g-s}\,d\mu}{2\sqrt{R(\mu)}},\quad \tilde{\psi}_s=\sum_{k=1}^{g}\int_{\lambda_0}^{\nu_k}\frac{\nu^{g-s}\,d\nu}{2\sqrt{R(\nu)}},\quad (s=1,...,g). \tag{4.4}$$

Proposition 4.2: (Straightening of the F_λ flow)

$$\frac{d\tilde{\phi}_s}{dt_\lambda}=\frac{\lambda^{g-s}}{a(\lambda)},\quad \frac{d\tilde{\psi}_s}{dt_\lambda}=\frac{-\lambda^{g-s}}{a(\lambda)}. \tag{4.5}$$

Corollary 4.3: (Straightening of the F_k flow). Let t_k be the variable of the F_k flow. Then

$$\frac{d\tilde{\phi}}{dt_0}=0, \quad \left(\frac{d\tilde{\phi}}{dt_1},\frac{d\tilde{\phi}}{dt_2},\ldots,\frac{d\tilde{\phi}}{dt_{N-1}}\right)=\begin{pmatrix} 1 & A_1 & A_2 & \cdots & A_{N-2} \\ & 1 & A_1 & \cdots & A_{N-3} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & A_1 \\ & & & & 1 \end{pmatrix}, \tag{4.6}$$

and $d\tilde{\phi}/dt_k=-d\tilde{\psi}/dt_k$, where A_j's are the coefficient in the expansion

$$\frac{\lambda^N}{a(\lambda)}=\frac{1}{(1-\alpha_1\lambda^{-1})\cdots(1-\alpha_N\lambda^{-1})}=1+\sum_{j=1}^{\infty} A_j\lambda^{-j}, \tag{4.7}$$

which could be represented through the power sums of α_k, $s_l=\Sigma_{k=1}^{N}\alpha_k^l$:

$$A_1=s_1, \quad A_2=\tfrac{1}{2}(s_2+s_1^2), \quad A_3=\tfrac{1}{6}(2s_3+3s_2s_1+s_1^3),$$

with the recursive formula:

$$A_k=\frac{1}{k}\left(s_k+\sum_{\substack{i+j=k\\ i,j\geqslant 1}} s_iA_j\right). \tag{4.8}$$

Proof: According to the definition of the Poisson bracket:

$$\frac{d\tilde{\phi}}{dt_\lambda}=(\tilde{\phi},F_\lambda)=\sum_{k=0}^{\infty}\frac{1}{\lambda^{k+1}}(\tilde{\phi},F_k)=\sum_{k=0}^{\infty}\frac{1}{\lambda^{k+1}}\frac{d\tilde{\phi}}{dt_k}.$$

With the supplementary definition $A_0=1$, $A_{-j}=0$, $(j=1,2,3,...)$, the comparison of the coefficients of λ^{-k-1} in Eq. (4.5) yields $d\tilde{\phi}_s/dt_k=A_{k-s}$, and

$$\frac{d\tilde{\phi}}{dt_k}=(A_{k-1},A_{k-2},\ldots,A_{k-g})^T. \tag{4.9}$$

The proof of Eq. (4.8) is elementary, resorting to the expansion $\Sigma_1^\infty k^{-1}s_k\lambda^{-k}$ of the RHS of Eq. (4.7).

Proposition 4.4: $F_0,F_1,\ldots,F_{N-1}$ given in Eq. (3.10) are functionally independent.

Proof: We need only prove the linear independence of the differentials $dF_0,dF_1,\ldots,dF_{N-1}$. Recall the expression of the Poisson bracket by means of the symplectic structure $\omega^2=dp\wedge dq$:[18]

$$(H,F)=\omega^2(IdF,IdH).$$

Suppose $\Sigma_{k=0}^{N-1}c_kdF_k=0$. Let $H=\tilde{\phi}_s$, we have

$$0=\sum_{k=0}^{N-1}c_k\omega^2(IdF_k,Id\tilde{\phi}_s)=\sum_{k=0}^{N-1}c_k(\tilde{\phi}_s,F_k)=\sum_{k=1}^{N-1}c_k\frac{d\tilde{\phi}_s}{dt_k}.$$

Hence $c_1=\cdots=c_{N-1}=0$ since the coefficient determinant is equal to 1 by Eq. (4.6). Thus $c_0dF_0=0$. And $c_0=0$ since

$$dF_0=-\sum(q_kdp_k+p_kdq_k)\neq 0.$$

Note: Corollary 3.2 and the present Proposition complete the proof of the Liouville integrability of the ZS–Bargmann system (3.3) with the Hamiltonian $H_1=\frac{1}{2}F_1+\frac{1}{2}F_0^2$ and N integrals $F_0,F_1,\ldots,F_{N-1}$, which are involutive in pairs and functionally independent.

V. THE POLYNOMIAL INTEGRALS $\{H_k\}$

Introduce another set of polynomial integrals $\{H_k\}$ for the ZS–Bargmann system (3.3) recursively by:

$$\begin{gathered} H_0=\tfrac{1}{2}F_0, \\ H_{k+1}=\tfrac{1}{2}F_{k+1}+2\sum_{i+j=k} H_iH_j. \end{gathered} \tag{5.1}$$

The first few members are

$$\begin{gathered} H_0=\tfrac{1}{2}F_0, \quad H_1=\tfrac{1}{2}F_1+\tfrac{1}{2}F_0^2, \\ H_2=\tfrac{1}{2}F_2+F_1F_0+F_0^3, \\ H_3=\tfrac{1}{2}F_3+F_2F_0+\tfrac{1}{2}F_1^2+3F_1F_0^2+\tfrac{5}{2}F_0^4, \\ H_4=\tfrac{1}{2}F_4+F_3F_0+F_2F_1+3F_2F_0^2+3F_1^2F_0+10F_1F_0^3+7F_0^5. \end{gathered} \tag{5.2}$$

Here H_1 is exactly the Hamiltonian for the ZS–Bargmann system (3.3). Equation (5.1) is equivalent to

$$H_\lambda=\tfrac{1}{2}(F_\lambda+\tfrac{1}{4})+2H_\lambda^2,$$

or

$$-4F_\lambda=(1-4H_\lambda)^2, \tag{5.3}$$

where $H_\lambda=\Sigma_{k=0}^{\infty}H_k\lambda^{-k-1}$ is the generating function of $\{H_k\}$.

Proposition 5.1: $\{H_k\}$ satisfies the Liouville conditions of completely integrability:

$$\begin{gathered} \text{(i)} \quad (H_\lambda,H_\mu)=0, \quad \forall \lambda,\mu\in\mathbb{C}, \\ (H_j,H_k)=0, \quad \forall j,k=0,1,2,\ldots; \end{gathered} \tag{5.4}$$

$$\text{(ii)} \quad H_0,H_1,\ldots,H_{N-1} \text{ are functionally independent.}$$

Proof: Since $\{\alpha_j\}$ are the only poles of F_λ, the expansion

$$-4F_\lambda=1-4\sum_{k=0}^{\infty}F_k\lambda^{-k-1}=1+O\left(\frac{1}{\lambda}\right)$$

convergent absolutely outside the circle:

$$|\lambda|>\max(|\alpha_1|,\ldots,|\alpha_1|).$$

Thus for sufficiently large $|\lambda|$ we solve

$$1-4H_\lambda=\sqrt{-4F_\lambda}. \tag{5.5}$$

According to the Leibniz rule for the Poisson bracket, we obtain

$$(H_\lambda, H_\mu) = \frac{1}{16\sqrt{F_\lambda F_\mu}} (F_\lambda, F_\mu) = 0,$$

for sufficiently large $|\lambda|$ and $|\mu|$. By the principle of analytic continuation, Eq. (5.4) is valid for any λ and μ. From Eq. (5.3) we have

$$\tfrac{1}{2} dF_\lambda = (1 - 4H_\lambda) dH_\lambda,$$

$$\frac{1}{2}\begin{pmatrix} dF_0 \\ dF_1 \\ \vdots \\ dF_{N-1} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ -4H_0 & 1 & & \\ \cdots & \cdots & \cdots & \\ -4H_{N-2} & \cdots & -4H_0 & 1 \end{pmatrix} \begin{pmatrix} dH_0 \\ dH_1 \\ \vdots \\ dH_{N-1} \end{pmatrix}.$$

Thus the linear independence of $dH_0, \ldots, dH_{N-1}$ is equivalent to that of $dF_0, \ldots, dF_{N-1}$, which completes the proof of the functional independence of $H_0, \ldots, H_{N-1}$.

The explicit form of H_k by means of canonical coordinates p,q is obtained through substituting the expression (3.10) into Eq. (5.2). The first few are as follows:

$$\begin{aligned} H_0 &= -\tfrac{1}{2}\langle p,q\rangle, \\ H_1 &= -\tfrac{1}{2}\langle Ap,q\rangle + \tfrac{1}{2}\langle p,p\rangle\langle q,q\rangle, \\ H_2 &= -\tfrac{1}{2}\langle A^2p,q\rangle + \tfrac{1}{2}\langle Ap,p\rangle\langle q,q\rangle + \tfrac{1}{2}\langle p,p\rangle\langle Aq,q\rangle - \langle p,p\rangle\langle p,q\rangle\langle q,q\rangle, \\ H_3 &= -\tfrac{1}{2}\langle A^3p,q\rangle + \tfrac{1}{2}\langle A^2p,p\rangle\langle q,q\rangle + \tfrac{1}{2}\langle Ap,p\rangle\langle Aq,q\rangle + \tfrac{1}{2}\langle p,p\rangle\langle A^2q,q\rangle - \langle Ap,p\rangle\langle p,q\rangle\langle q,q\rangle \\ &\quad - \langle p,p\rangle\langle Ap,q\rangle\langle q,q\rangle - \langle p,p\rangle\langle p,q\rangle\langle Aq,q\rangle + \tfrac{1}{2}\langle p,p\rangle^2\langle q,q\rangle^2 + 2\langle p,p\rangle\langle p,q\rangle^2\langle q,q\rangle. \end{aligned} \tag{5.6}$$

VI. RELATION OF H_k AND X_k

The kernel of J is of dimension 1 with the generator g_{-1}. Exerting $J^{-1}K$ on Eq. (3.4) k times gives

$$\sum_{j=1}^{N} \alpha_j^k \nabla \alpha_j = g_k + c_1 g_{k-1} + \cdots + c_k g_0 + c_{k+1} g_{-1}, \tag{6.1}$$

since each time there is an extra term $c_i g_{-1}$. Hence we have

$$0 = \sum_{j=1}^{N} a(\alpha_j)\nabla \alpha_j = g_N + c_{N1} g_{N-1} + \cdots + c_{NN} g_0 + c_{N,N+1} g_{-1}, \tag{6.2}$$

where

$$\begin{aligned} a(\lambda) &= \prod_{j=1}^{N} (\lambda - \alpha_j) = \sum_{k=0}^{N} a_{N-k}\lambda^k, \quad (a_0 = 1), \\ c_{N,N-i} &= \sum_{k=i}^{N} a_{N-k} c_{k-i}, \quad c_{N,N+1} = \sum_{k=0}^{N} a_{N-k} c_{k+1}. \end{aligned} \tag{6.3}$$

Acting with PJ on Eq. (6.2) yields:

Proposition 6.1: Let $(p(x), q(x))$ be a solution of the ZS–Bargmann system (3.3). Then

$$\begin{pmatrix} u \\ v \end{pmatrix} = f(p,q) = \begin{pmatrix} -\langle p,p\rangle \\ \langle q,q\rangle \end{pmatrix} \tag{6.4}$$

solves the Nth stationary AKNS equation

$$X_N + c_{N1}X_{N-1} + \cdots + c_{NN}X_0 = 0. \tag{6.5}$$

Proposition 6.2: The coefficients in Eq. (6.1) are given by:

$$c_k = -4H_{k-1}. \tag{6.6}$$

Furthermore,

$$g_\lambda = \frac{1}{1-4H_\lambda} G_\lambda . \tag{6.7}$$

Proof: Define the generating function of $\{c_k\}$ by

$$c_\lambda = 1 + \sum_{k=1}^{\infty} c_k \lambda^{-k}.$$

Multiplied by λ^{-k-1} and summed with respect to k from 0 to ∞, Eq. (6.1) becomes

$$G_\lambda = g_{-1} + \begin{pmatrix} Q_\lambda(q,q) \\ -Q_\lambda(p,p) \\ Q_\lambda(p,q) \end{pmatrix} = c_\lambda g_\lambda .$$

Thus $\sigma(G_\lambda) = c_\lambda \sigma(g_\lambda)$ and $F_\lambda = -\frac{1}{4}c_\lambda^2$ due to Eq. (2.6), which yields $c_\lambda = 1-4H_\lambda$ and Eq. (6.6) by comparing with Eq. (5.3).

Denote the solution variety of Eqs. (3.3) and (6.5) by $\mathcal{N}$ and $\mathcal{M}$, respectively. By Proposition 6.1 f maps $\mathcal{N}$ into $\mathcal{M}$. Consider the canonical equation of the H_k flow:

$$\frac{d}{d\tau_k}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} -\dfrac{\partial H_k}{\partial q} \\ \dfrac{\partial H_k}{\partial p} \end{pmatrix} = I\nabla H_k \tag{6.8}$$

and the solution of the initial value problem:

$$\begin{pmatrix} p(\tau_k) \\ q(\tau_k) \end{pmatrix} = h_k^{\tau_k}\begin{pmatrix} p_0 \\ q_0 \end{pmatrix}. \tag{6.9}$$

Specifically, $\tau_1 = x$. It is well-known[18] that the involutivity $(H_j, H_k) = 0$ implies the commutativity of $h_j^{\tau_j}$ and $h_k^{\tau_k}$. Hence $I\nabla H_k$ can be regarded as a vector field on $\mathcal{N}$. The importance of H_k is that the differential f_* maps $I\nabla H_k$ exactly into the AKNS vector field X_k restricted to $\mathcal{M}$, where

$$f_*(p,q)\begin{pmatrix} \delta p \\ \delta q \end{pmatrix} = \frac{d}{d\epsilon}\bigg|_{\epsilon=0} f(p+\epsilon\delta p, q+\epsilon\delta q) = \begin{pmatrix} -2\langle p,\delta p\rangle \\ 2\langle q,\delta q\rangle \end{pmatrix}.$$

Proposition 6.3:

$$f_*(I\nabla H_\lambda) = PJg_\lambda|_{\mathcal{M}}, \quad \lambda \in \mathbb{C},$$
$$f_*(I\nabla H_k) = X_k|_{\mathcal{M}}, \quad k = 0,1,2,\ldots . \tag{6.10}$$

Proof: From the canonical Eq. (3.11) of the t_λ flow we get

$$\begin{aligned}\frac{d}{dt_\lambda}\langle p,p\rangle &= 4V_\lambda^{11}Q_\lambda(p,p)+4V_\lambda^{12}Q_\lambda(p,q)=2Q_\lambda(p,p),\\ \frac{d}{dt_\lambda}\langle q,q\rangle &= 4V_\lambda^{21}Q_\lambda(p,q)-4V_\lambda^{11}Q_\lambda(q,q)=-2Q_\lambda(q,q).\end{aligned} \tag{6.11}$$

Thus

$$f_*(I\nabla F_\lambda)=f_*\left\{\frac{d}{dt_\lambda}\begin{pmatrix}p\\q\end{pmatrix}\right\}=\frac{d}{dt_\lambda}\begin{pmatrix}-\langle p,p\rangle\\ \langle q,q\rangle\end{pmatrix}=-2\begin{pmatrix}Q_\lambda(p,p)\\ Q_\lambda(q,q)\end{pmatrix}.$$

By Eqs. (5.3), (3.5), and (6.7):

$$\nabla H_\lambda=\tfrac{1}{2}\nabla F_\lambda+4H_\lambda\nabla H_\lambda\,,$$

$$f_*(I\nabla H_\lambda)=\frac{1}{2(1-4H_\lambda)}f_*(I\nabla F_\lambda)=\frac{-1}{1-4H_\lambda}\begin{pmatrix}Q_\lambda(p,p)\\ Q_\lambda(q,q)\end{pmatrix}=\frac{1}{1-4H_\lambda}PJG_\lambda=PJg_\lambda\,.$$

Hence we have the first part of Eq. (6.10). The second part is obtained by comparing the coefficients of the same power λ^{-k-1}.

Corollary 6.4: (i) In the case $k=2$, denote $\tau_1=x,\ \tau_2=y$. Let $(p(x,y),q(x,y))$ be a compatible solution of

$$\begin{pmatrix}p\\q\end{pmatrix}_x=I\nabla H_1,\quad \begin{pmatrix}p\\q\end{pmatrix}_y=I\nabla H_2.$$

Then $u(x,y)=-\langle p,p\rangle, v(x,y)=\langle q,q\rangle$ solves the coupled NS equation:

$$\begin{pmatrix}u\\v\end{pmatrix}_y=X_2=\begin{pmatrix}u_{xx}-2u^2v\\ -v_{xx}+2uv^2\end{pmatrix}. \tag{6.12}$$

(ii) In the case $k=3$, denote $\tau_1=x,\ \tau_3=t$. Let $(p(x,t),q(x,t))$ be a compatible solution of

$$\begin{pmatrix}p\\q\end{pmatrix}_x=I\nabla H_1,\quad \begin{pmatrix}p\\q\end{pmatrix}_t=I\nabla H_3.$$

Then $u(x,t)=-\langle p,p\rangle,\ v(x,t)=\langle q,q\rangle$ solves the coupled mKdV equation:

$$\begin{pmatrix}u\\v\end{pmatrix}_t=X_3=\begin{pmatrix}u_{xxx}-6uu_xv\\ v_{xxx}-6uvv_x\end{pmatrix}. \tag{6.13}$$

VII. THE KP EQUATION

In this section, the special solution of the KP equation is first separated into two AKNS flows X_2 and X_3, and then into three confocal flows H_1,H_2,H_3.

Proposition 7.1: Let $u(x,y,t),v(x,y,t)$ be a compatible solution of the coupled NS equation (6.12) and the coupled mKdV (6.13). Then

$$w(x,y,t)=-2u(x,y,t)v(x,y,t) \tag{7.1}$$

solves the KP equation:

$$w_t=\tfrac{1}{4}(w_{xx}+3w^2)_x+\tfrac{3}{4}\partial_x^{-1}w_{yy}\,. \tag{7.2}$$

Proof: A direct calculation gives:

$$w_{tx}=w_{xxxx}+\tfrac{3}{2}(w^2)_{xx}+6(u_{xxx}v_x+u_xv_{xxx}+2u_{xx}v_{xx}),$$

$$\tfrac{3}{4}w_{yy}=\tfrac{3}{4}w_{xxxx}+\tfrac{3}{4}(w^2)_{xx}+6(u_{xxx}v_x+u_xv_{xxx}+2u_{xx}v_{xx}).$$

The difference gives Eq. (7.2)

Theorem 7.2: Let $(p(x,y,t),q(x,y,t))$ be a compatible solution of three systems of ODEs:

$$\begin{pmatrix}p\\q\end{pmatrix}_x=I\nabla H_1,\quad \begin{pmatrix}p\\q\end{pmatrix}_y=I\nabla H_2,\quad \begin{pmatrix}p\\q\end{pmatrix}_t=I\nabla H_3, \tag{7.3}$$

where H_j's are given explicitly by Eq. (5.6). Then

$$w(x,y,t)=2\langle p,p\rangle\langle q,q\rangle=2\langle Ap,q\rangle+4H_1 \tag{7.4}$$

solves the KP equation (7.2).

Proof: Since the flow operators defined by Eq. (6.9) h_1^x,h_2^y,h_3^t commute, we write the compatible solution in two ways:

$$\begin{pmatrix}p\\q\end{pmatrix}=h_1^xh_2^y\left\{h_3^t\begin{pmatrix}p_0\\q_0\end{pmatrix}\right\}=h_1^xh_3^t\left\{h_2^y\begin{pmatrix}p_0\\q_0\end{pmatrix}\right\}, \tag{7.5}$$

and regard the quantities in the parentheses $\{\cdot\}$ as initial values. Then, as a function of (x,y) and (x,t), respectively,

$$(u(x,y,t),v(x,y,t))=(-\langle p,p\rangle,\langle q,q\rangle)$$

Eqs. (6.12) and (6.13) simultaneously by Corollary 6.4. According to Proposition 7.1,

$$w=-2uv=2\langle p,p\rangle\langle q,q\rangle$$

is a solution to Eq. (7.2).

VIII. EVOLUTION PICTURE VIA THE ABEL–JACOBI COORDINATES

The shape of Eq. (4.4) suggests the consideration of the holomorphic differential

$$\tilde{\omega}_s=\frac{\lambda^{g-s}d\lambda}{2\sqrt{R(\lambda)}},\quad s=1,\dots,g, \tag{8.1}$$

on the hyperelliptic curve Γ:

$$\xi^2-4R(\lambda)=0, \tag{8.2}$$

with genus $g=N-1$ since $\deg R=2N$ by Eq. (4.3). For the same λ, there are two points on different sheets of the Riemann surface Γ:

$$\rho(\lambda)=(\lambda,2\sqrt{R(\lambda)}),\quad \rho_-(\lambda)=(\lambda,-2\sqrt{R(\lambda)}).$$

At infinity, the affine equation (8.2) is transformed into ($z=\lambda^{-1}$, $\hat{\xi}=z^N\xi$):

$$\hat{\xi}^2-4R_*(z)=0, \tag{8.3}$$

with

$$R_*(z)=z^{2N}R(z^{-1})=\prod_{j=1}^{2N}(1-\lambda_j z).$$

The two infinities are represented as:

$$\infty_l=(z=0,\hat{\xi}=(-1)^l 2),\quad l=1,2. \tag{8.4}$$

Take the canonical basis of cycles on $\Gamma: a_1,\ldots,a_g; b_1,\ldots,b_g$. Let $C=(C_{js})$ be the inverse of the periodic matrix (A_{sk}):

$$C=(A_{sk})^{-1}_{g\times g},\quad A_{sk}=\int_{a_k}\tilde{\omega}_s. \tag{8.5}$$

Then for the normalized holomorphic differential

$$\omega_j=\sum_{s=1}^{g}C_{js}\tilde{\omega}_s,\quad \omega=(\omega_1,\ldots,\omega_g)^T=C\tilde{\omega},$$

we have

$$\int_{a_k}\omega_j=\delta_{jk},\quad \int_{b_k}\omega_j=B_{jk}. \tag{8.6}$$

According to the Riemannian bilinear relation,[19] the matrix $B=(B_{jk})$ is symmetric and has positive definite imaginary part, and is used to construct the Riemannian theta function of Γ:

$$\theta(\zeta)=\sum_{z\in\mathcal{Z}^g}\exp\pi\sqrt{-1}(\langle Bz,z\rangle+2\langle\zeta,z\rangle),\quad \zeta\in\mathbb{C}^g.$$

For fixed λ_0, the Abel–Jacobi coordinates are defined as

$$\phi=\sum_{k=1}^{g}\int_{\rho(\lambda_0)}^{\rho(\mu_k)}\omega,\quad \psi=\sum_{k=1}^{g}\int_{\rho(\lambda_0)}^{\rho(\nu_k)}\omega. \tag{8.7}$$

Lemma 8.1: Let $S_k=\lambda_1^k+\cdots+\lambda_{2N}^k$. Then the coefficients of

$$\frac{1}{\sqrt{R_*(z)}}=\sum_{k=0}^{\infty}\Lambda_k z^k \tag{8.8}$$

satisfy the recursive formula:

$$\Lambda_0=1,\quad \Lambda_1=\tfrac{1}{2}S_1,$$
$$\Lambda_k=\frac{1}{2k}\left(S_k+\sum_{\substack{i+j=k\\ i,j\geqslant 1}}S_i\Lambda_j\right). \tag{8.9}$$

Proof: Since $-\ln(1-t)=\Sigma_{k=1}^{\infty}k^{-1}t^k$, we have

$$\ln\frac{1}{\sqrt{R_*(z)}}=-\frac{1}{2}\sum_{j=1}^{2N}\ln(1-\lambda_j z)=\sum_{k=1}^{\infty}\frac{1}{2k}S_k z^k.$$

By differentiating with regard to z and comparing the coefficients of z^k, we get Eq. (8.9).

Let $C_1,\ldots,C_g$ be the column vectors of C defined by Eq. (8.5). Then by direct calculations, the coefficients in

$$\frac{1}{2\sqrt{R_*(z)}}(C_1z+C_2z+\cdots+C_gz^g)=\sum_{k=1}^{\infty}\Omega_k z^k \tag{8.10}$$

are

$$\Omega_k=\tfrac{1}{2}(\Lambda_{k-1}C_1+\cdots+\Lambda_{k-g}C_g) \tag{8.11}$$

with additional defined $\Lambda_{-s}=0$, $(s=1,2,...)$. Specifically,

$$\begin{gathered}\Omega_0=0,\quad \Omega_1=\tfrac{1}{2}C_1,\\ \Omega_k=\tfrac{1}{2}(\Lambda_{k-1}C_1+\cdots+\Lambda_1C_{k-1}+C_k),\quad (k=1,...,g).\end{gathered} \tag{8.12}$$

Proposition 8.2: The τ_k flow is straightened by the Abel–Jacobi coordinates:

$$\frac{d\phi}{d\tau_k}=\Omega_k,\quad \frac{d\psi}{d\tau_k}=-\Omega_k. \tag{8.13}$$

Proof: By Eqs. (4.1), (4.3), and (5.3) we have

$$\sqrt{R(\lambda)}=a(\lambda)(1-4H_\lambda). \tag{8.14}$$

Since the derivative of a function along a Hamiltonian flow is equal to its Poisson bracket with the Hamiltonian, Eq. (5.3) implies

$$\frac{d}{d\tau_\lambda}=\frac{1}{2(1-4H_\lambda)}\frac{d}{dt_\lambda}. \tag{8.15}$$

Therefore, from Eqs. (4.5), (8.14), and (8.10) we get

$$\frac{d\tilde{\phi}}{d\tau_\lambda}=\frac{\lambda^g}{2\sqrt{R(\lambda)}}(\lambda^{-1},\ldots,\lambda^{-g})^T,$$

$$\frac{d\phi}{d\tau_\lambda}=C\frac{d\tilde{\phi}}{d\tau_\lambda}=\frac{\lambda^g}{2\sqrt{R(\lambda)}}(C_1\lambda^{-1}+\cdots+C_g\lambda^{-g})=\sum_{k=1}^{\infty}\Omega_k\lambda^{-k-1}.$$

Hence we obtain the first part of Eq. (8.13) after comparing the coefficients of λ^{-k-1}, while the second part is obtained similarly.

The straightened equations (8.13) are easily integrated by quadratures: $\phi=\phi_0+\Sigma\Omega_k\tau_k$. And the evolution picture of the confocal flow and AKNS flow becomes very simple through the ''window'' of the Abel–Jacobi coordinates ϕ (as well as ψ):

$$\begin{gathered}\text{confocal}\ \ H_k:\quad \phi=\phi_0+\Omega_k\tau_k,\\ \text{AKNS}\ \ X_k:\quad \phi=\phi_0+\Omega_1x+\Omega_k\tau_k.\end{gathered} \tag{8.16}$$

Specifically,

$$\left.\begin{array}{c}\text{ZS–Bargmann}\\ \text{Stationary AKNS}\end{array}\right\}:\quad \phi=\phi_0+\Omega_1x,$$

$$\text{coupled NS:}\quad \phi=\phi_0+\Omega_1 x+\Omega_2 y,$$
$$\text{coupled mKdV:}\quad \phi=\phi_0+\Omega_1 x+\Omega_3 t, \tag{8.17}$$
$$\text{KP:}\quad \phi=\phi_0+\Omega_1 x+\Omega_2 y+\Omega_3 t.$$

The corresponding explicit solutions are obtained by some inversion procedures from ϕ,ψ to the coordinates u, v, and w via the elliptic coordinates $\{\mu_k\}$ and $\{\nu_k\}$.

IX. INVERSION FROM ϕ,ψ TO $\{\mu_k\},\{\nu_k\}$

The Abel map $A:\mathrm{Div}(\Gamma)\to\mathcal{J}(\Gamma)=\mathbb{C}^g/\mathcal{T}$ is defined by:

$$A(P)=\int_{P_0}^{P}\omega,\quad A\left(\sum n_k P_k\right)=\sum n_k A(P_k), \tag{9.1}$$

where $P_0=\rho(\lambda_0)$ is fixed, $\mathrm{Div}(\Gamma)$ is the divisor group, and the lattice $\mathcal{T}$ is spanned by the periodic vectors $\{\delta_j;B_j\}$, which are the column vectors of E and (B_{js}) defined by Eq. (8.6). The definition of Abel–Jacobi coordinates is rewritten as

$$\phi=A\left\{\sum_{j=1}^{g}\rho(\mu_j)\right\},\quad \psi=A\left\{\sum_{j=1}^{g}\rho(\nu_j)\right\}. \tag{9.2}$$

According to the Riemann theorem,[19] there exists a constant vector K such that

(i) $\theta(A(\rho(\lambda))-\phi-K)$ has exactly g zeros at $\lambda=\mu_1,\ldots,\mu_g$;

(ii) $\theta(A(\rho(\lambda))-\psi-K)$ has exactly g zeros at $\lambda=\nu_1,\ldots,\nu_g$.

And we have the inversion formula:

$$\sum_{j=1}^{g}\mu_j^s=I_s(\Gamma)-\sum_{l=1}^{2}\operatorname*{Res}_{\lambda=\infty_l}\lambda^s d\ln\theta(A(\rho(\lambda))-\phi-K),$$
$$\sum_{j=1}^{g}\nu_j^s=I_s(\Gamma)-\sum_{l=1}^{2}\operatorname*{Res}_{\lambda=\infty_l}\lambda^s d\ln\theta(A(\rho(\lambda))-\psi-K), \tag{9.3}$$

where

$$I_s(\Gamma)=\sum_{j=1}^{g}\int_{a_j}\lambda^s\omega_j.$$

In the neighborhood of ∞_l $(l=1,2)$, since the two-valued function $\lambda^{-N}\sqrt{R(\lambda)}$ tends to $(-1)^l$ due to Eq. (8.4), we have $(z=\lambda^{-1})$:

$$\lambda^{-N}\sqrt{R(\lambda)}=(-1)^l\sqrt{R_*(z)},$$

By Eq. (8.1) we get:

$$\omega=C\tilde{\omega}=(-1)^{l-1}\frac{z^{-1}\,dz}{z\sqrt{R_*(z)}}(C_1 z+\cdots+C_g z^g).$$

With the help of Eq. (8.10) we obtain:

$$\omega=(-1)^{l-1}\sum_{k=1}^{\infty}\Omega_k z^{k-1}\,dz, \tag{9.4}$$

$$A(\rho(z^{-1}))=-\eta_l+(-1)^{l-1}\sum_{k=1}^{\infty}\frac{1}{k}\Omega_k z^k, \tag{9.5}$$

where

$$\eta_l=\int_{\infty_l}^{P_0}\omega.$$

Hence we have the power series expansions near ∞_l in the local coordinate $z=\lambda^{-1}$:

$$\ln\theta(A(\rho(\lambda))-\phi-K)=\ln\theta\left(\phi+K+\eta_l+(-1)^l\sum_{k=1}^{\infty}\frac{1}{k}\Omega_k z^k\right)=\ln\theta(\phi+K+\eta_l)+\sum_{k=1}^{\infty}f_k^l z^k,$$

$$\begin{aligned}\ln\theta(A(\rho(\lambda))-\psi-K)&=\ln\theta\left(-\psi-K-\eta_l+(-1)^{l-1}\sum_{k=1}^{\infty}\frac{1}{k}\Omega_k z^k\right)=\ln\theta(-\psi-K-\eta_l)\\&\quad+\sum_{k=1}^{\infty}\hat{f}_k^l z^k.\end{aligned} \tag{9.6}$$

Here the fact $\theta(\zeta)=\theta(-\zeta)$ is used. The power sums are obtained after substituting Eq. (9.6) into Eq. (9.3):

$$\begin{aligned}\sum_{j=1}^{g}\mu_j^s&=I_s(\Gamma)-sf_s^1-sf_s^2,\\ \sum_{j=1}^{g}\nu_j^s&=I_s(\Gamma)-s\hat{f}_s^1-s\hat{f}_s^2.\end{aligned} \tag{9.7}$$

The coefficients $f_k^l,\hat{f}_k^l$ in the Taylor expansions (9.6) are calculated by the chain rule of differentiation. Denote $\partial_j=\partial/\partial\zeta_j$, $\partial_{jk}^2=\partial^2/\partial\zeta_j\partial\zeta_k$, etc. Adopting the Einstein summation convention, finally we get:

$$\begin{aligned}\sum_{j=1}^{g}\mu_j&=I_1(\Gamma)+\Omega_1^j\partial_j\ln\frac{\theta_1}{\theta_2},\\
\sum_{j=1}^{g}\mu_j^2&=I_2(\Gamma)+\Omega_2^j\partial_j\ln\frac{\theta_1}{\theta_2}-\Omega_1^j\Omega_1^k\partial_{jk}^2\ln\theta_1\theta_2,\\
\sum_{j=1}^{g}\mu_j^3&=I_3(\Gamma)+\Omega_3^j\partial_j\ln\frac{\theta_1}{\theta_2}-\frac{3}{2}\Omega_2^j\Omega_1^k\partial_{jk}^2\ln\theta_1\theta_2+\frac{1}{2}\Omega_1^j\Omega_1^k\Omega_1^r\partial_{jkr}^3\ln\frac{\theta_1}{\theta_2},\\
\sum_{j=1}^{g}\mu_j^4&=I_4(\Gamma)+\Omega_4^j\partial_j\ln\frac{\theta_1}{\theta_2}-\left(\frac{4}{3}\Omega_3^j\Omega_1^k+\frac{1}{2}\Omega_2^j\Omega_2^k\right)\partial_{jk}^2\ln\theta_1\theta_2+\Omega_2^j\Omega_1^k\Omega_1^r\partial_{jkr}^3\ln\frac{\theta_1}{\theta_2}\\
&\quad-\frac{1}{6}\Omega_1^j\Omega_1^k\Omega_1^r\Omega_1^i\partial_{jkri}^4\ln\theta_1\theta_2,\\
\sum_{j=1}^{g}\nu_j&=I_1(\Gamma)-\Omega_1^j\partial_j\ln\frac{\theta_1^*}{\theta_2^*},\end{aligned} \tag{9.8}$$

$$\sum_{j=1}^{g} \nu_j^2 = I_2(\Gamma) - \Omega_2^j \partial_j \ln \frac{\theta_1^*}{\theta_2^*} - \Omega_1^j \Omega_1^k \partial_{jk}^2 \ln \theta_1^* \theta_2^* ,$$

$$\sum_{j=1}^{g} \nu_j^3 = I_3(\Gamma) - \Omega_3^j \partial_j \ln \frac{\theta_1^*}{\theta_2^*} - \frac{3}{2} \Omega_2^j \Omega_1^k \partial_{jk}^2 \ln \theta_1^* \theta_2^* - \frac{1}{2} \Omega_1^j \Omega_1^k \Omega_1^r \partial_{jkr}^3 \ln \frac{\theta_1^*}{\theta_2^*},$$

$$\sum_{j=1}^{g} \nu_j^4 = I_4(\Gamma) - \Omega_4^j \partial_j \ln \frac{\theta_1^*}{\theta_2^*} - \left(\frac{4}{3} \Omega_3^j \Omega_1^k + \frac{1}{2} \Omega_2^j \Omega_2^k \right) \partial_{jk}^2 \ln \theta_1^* \theta_2^* - \Omega_2^j \Omega_1^k \Omega_1^r \partial_{jkr}^3 \ln \frac{\theta_1^*}{\theta_2^*}$$
$$- \frac{1}{6} \Omega_1^j \Omega_1^k \Omega_1^r \Omega_1^i \partial_{jkri}^4 \ln \theta_1^* \theta_2^* ,$$

where $(l=1,2)$

$$\theta_l = \theta(\phi + K + \eta_l) = \theta\left(\sum \Omega_k \tau_k + \phi_0 + K + \eta_l \right),$$
$$\theta_l^* = \theta(-\psi - K - \eta_l) = \theta\left(\sum \Omega_k \tau_k - \psi_0 - K - \eta_l \right). \tag{9.9}$$

Since $\{\mu_j\}$ and $\{\nu_k\}$ appear symmetrically in Eq. (9.2), what we can expect to get in the inversion formula is only the symmetric function of $\{\mu_j\}$ or $\{\nu_j\}$.

Denote $x=\tau_1$, $y=\tau_2$, $t=\tau_3$. By the chain rule of differentiation for composition functions, Eq. (9.8) is further simplified as:

$$\sum_{j=1}^{g} \mu_j = I_1(\Gamma) + \partial_x \ln \frac{\theta_1}{\theta_2},$$

$$\sum_{j=1}^{g} \mu_j^2 = I_2(\Gamma) + \partial_y \ln \frac{\theta_1}{\theta_2} - \partial_x^2 \ln \theta_1 \theta_2 ,$$

$$\sum_{j=1}^{g} \mu_j^3 = I_3(\Gamma) + \left(\partial_t + \frac{1}{2} \partial_x^3 \right) \ln \frac{\theta_1}{\theta_2} - \frac{3}{2} \partial_y \partial_x \ln \theta_1 \theta_2 ,$$

$$\sum_{j=1}^{g} \mu_j^4 = I_4(\Gamma) + (\partial_{\tau_4} + \partial_y \partial_x^2) \ln \frac{\theta_1}{\theta_2} - \left(\frac{4}{3} \partial_t \partial_x + \frac{1}{2} \partial_y^2 + \frac{1}{6} \partial_x^4 \right) \ln \theta_1 \theta_2 ,$$

$$\sum_{j=1}^{g} \nu_j = I_1(\Gamma) - \partial_x \ln \frac{\theta_1^*}{\theta_2^*}, \tag{9.10}$$

$$\sum_{j=1}^{g} \nu_j^2 = I_2(\Gamma) - \partial_y \ln \frac{\theta_1^*}{\theta_2^*} - \partial_x^2 \ln \theta_1^* \theta_2^* ,$$

$$\sum_{j=1}^{g} \nu_j^3 = I_3(\Gamma) - \left(\partial_t + \frac{1}{2} \partial_x^3 \right) \ln \frac{\theta_1^*}{\theta_2^*} - \frac{3}{2} \partial_y \partial_x \ln \theta_1^* \theta_2^* ,$$

$$\sum_{j=1}^{g} \nu_j^4 = I_4(\Gamma) - (\partial_{\tau_4} + \partial_y \partial_x^2) \ln \frac{\theta_1^*}{\theta_2^*} - \left(\frac{4}{3} \partial_t \partial_x + \frac{1}{2} \partial_y^2 + \frac{1}{6} \partial_x^4 \right) \ln \theta_1^* \theta_2^* .$$

X. RELATION BETWEEN $\{\mu_k\}$, $\{\nu_k\}$, AND $(u,v)=f(p,q)$

Proposition 10.1: Let $u=-\langle p,p\rangle$, $v=\langle q,q\rangle$. Then

$$\frac{1}{u}\frac{du}{d\tau_\lambda}=\frac{m(\lambda)}{\sqrt{R(\lambda)}},\quad \frac{1}{v}\frac{dv}{d\tau_\lambda}=\frac{-n(\lambda)}{\sqrt{R(\lambda)}}. \tag{10.1}$$

Proof: By Eqs. (8.15), (6.11), (4.1), and (8.14), we have

$$\frac{d}{d\tau_\lambda}\langle p,p\rangle=\frac{1}{2(1-4H_\lambda)}\frac{d}{dt_\lambda}\langle p,p\rangle=\frac{Q_\lambda(p,p)}{1-4H_\lambda}=\frac{\langle p,p\rangle m(\lambda)}{(1-4H_\lambda)a(\lambda)}=\frac{\langle p,p\rangle m(\lambda)}{\sqrt{R(\lambda)}}.$$

The calculation for v is similar.

Proposition 10.2: Let

$$\sigma_k=\sum_{j=1}^{g}\mu_j^k,\quad \hat{\sigma}_k=\sum_{j=1}^{g}\nu_j^k,\quad S_k=\sum_{j=1}^{2N}\lambda_j^k. \tag{10.2}$$

Then

$$\frac{1}{u}\frac{du}{d\tau_k}=T_k,\quad \frac{1}{v}\frac{dv}{d\tau_k}=-\hat{T}_k\quad (k=0,1,2,...), \tag{10.3}$$

where $T_k,\hat{T}_k$ are determined, respectively, by the recursive formulas:

$$\begin{gathered}\text{(i)}\quad T_0=1,\quad T_1=\tfrac{1}{2}S_1-\sigma_1,\\ T_k=\frac{1}{k}\left(\frac{1}{2}S_k-\sigma_k\right)+\frac{1}{k}\sum_{\substack{i+j=k\\ i,j\geqslant 1}}\left(\frac{1}{2}S_i-\sigma_i\right)T_j,\end{gathered} \tag{10.4}$$

$$\begin{gathered}\text{(ii)}\quad \hat{T}_0=1,\quad \hat{T}_1=\tfrac{1}{2}S_1-\hat{\sigma}_1,\\ \hat{T}_k=\frac{1}{k}\left(\frac{1}{2}S_i-\hat{\sigma}_i\right)+\frac{1}{k}\sum_{\substack{i+j=k\\ i,j\geqslant 1}}\left(\frac{1}{2}S_i-\hat{\sigma}_i\right)\hat{T}_{j,}.\end{gathered} \tag{10.5}$$

Proof: By a way quite similar to the proof of Eq. (8.9), we have Eq. (10.4) based on Eq. (10.1) and the expansion:

$$\frac{m(\lambda)}{\sqrt{R(\lambda)}}=\lambda^{-1}\exp\left\{\sum_{k=1}^{\infty}k^{-1}\left(\frac{1}{2}S_k-\sigma_k\right)\lambda^{-k}\right\}=\sum_{k=0}^{\infty}T_k\lambda^{-k-1}.$$

Equation (10.5) is proved with slight adjustment.

Substituting Eq. (9.10) into Eq. (10.3) in the case of $k=1$, we have

$$\frac{u_x}{u}=N_1-\partial_x\ln\frac{\theta_1}{\theta_2},\quad \frac{v_x}{v}=-N_1-\partial_x\ln\frac{\theta_1^*}{\theta_2^*}, \tag{10.6}$$

with

$$N_1=\tfrac{1}{2}S_1-I_1(\Gamma). \tag{10.7}$$

Thus

$$u=u_0e^{N_1x}\frac{\theta_2}{\theta_1}=u_0e^{N_1x}\frac{\theta(\Sigma\Omega_k\tau_k+\phi_0+K+\eta_2)}{\theta(\Sigma\Omega_k\tau_k+\phi_0+K+\eta_1)},$$
$$v=v_0e^{-N_1x}\frac{\theta_2^*}{\theta_1^*}=v_0e^{-N_1x}\frac{\theta(\Sigma\Omega_k\tau_k-\psi_0-K-\eta_2)}{\theta(\Sigma\Omega_k\tau_k-\psi_0-K-\eta_1)}, \tag{10.8}$$

where u_0,v_0 are independent of $x=\tau_1$, but dependent (if we consider the next flows) upon $y=\tau_2$, $t=\tau_3,\tau_4,\ldots$. Nevertheless, for the H_1 flow, except $x=\tau_1$, no other independent variables $\tau_2,\tau_3,\ldots$ appear. Thus we have:

Proposition 10.1: The stationary AKNS equation (6.5) has solution

$$u(x)=u(0)e^{N_1x}\frac{\theta(\Omega_1x+\phi_0+K+\eta_2)\,\theta(\phi_0+K+\eta_1)}{\theta(\Omega_1x+\phi_0+K+\eta_1)\,\theta(\phi_0+K+\eta_2)},$$
$$v(x)=v(0)e^{-N_1x}\frac{\theta(\Omega_1x-\psi_0-K-\eta_2)\,\theta(\psi_0+K+\eta_1)}{\theta(\Omega_1x-\psi_0-K-\eta_1)\,\theta(\psi_0+K+\eta_2)}, \tag{10.9}$$

which is called the finite-band potential of the ZS spectral problem (2.1).

Consider the H_1, H_2, and H_3 flow. Substitute the expressions of u,v of Eq. (10.8) into $w=-2uv$. By Theorem 7.2, we have the solution of KP:

$$w(x,y,t)=C(y,t)\,\frac{\theta_2\theta_2^*}{\theta_1\theta_1^*}.$$

The ‘‘coefficient’’ $C(y,t)$ is determined by letting $x=0$. Thus we obtain:

$$w(x,y,t)=w(0,y,t)\,\frac{\theta_2\theta_2^*}{\theta_1\theta_1^*}\left(\frac{\theta_1\theta_1^*}{\theta_2\theta_2^*}\right)_{x=0}.$$

Proposition 10.2: The KP equation has a special solution in the form:

$$w(x,y,t)=w(0,y,t)\,\frac{\theta(\Omega_1x+\Omega_2y+\Omega_3t+\phi_0+K+\eta_2)\,\theta(\Omega_1x+\Omega_2y+\Omega_3t-\psi_0-K-\eta_2)}{\theta(\Omega_1x+\Omega_2y+\Omega_3t+\phi_0+K+\eta_1)\,\theta(\Omega_1x+\Omega_2y+\Omega_3t-\psi_0-K-\eta_1)}$$
$$\cdot\,\frac{\theta(\Omega_2y+\Omega_3t+\phi_0+K+\eta_1)\,\theta(\Omega_2y+\Omega_3t-\psi_0-K-\eta_1)}{\theta(\Omega_2y+\Omega_3t+\phi_0+K+\eta_2)\,\theta(\Omega_2y+\Omega_3t-\psi_0-K-\eta_2)}. \tag{10.10}$$

We list the first members of Eq. (10.3) for later use:

$$\frac{u_x}{u}=\Sigma_1,\quad \frac{u_y}{u}=\frac{1}{2}\Sigma_2+\frac{1}{2}\Sigma_1^2,$$
$$\frac{u_t}{u}=\tfrac{1}{3}\Sigma_3+\tfrac{1}{2}\Sigma_2\Sigma_1+\tfrac{1}{6}\Sigma_1^3, \tag{10.11}$$
$$\frac{u_{\tau_4}}{u}=\tfrac{1}{4}\Sigma_4+\tfrac{1}{3}\Sigma_3\Sigma_1+\tfrac{1}{8}\Sigma_2^2+\tfrac{1}{4}\Sigma_2\Sigma_1^2+\tfrac{1}{24}\Sigma_1^4,$$

where $\Sigma_k=\tfrac{1}{2}S_k-\sigma_k$. It is more convenient to represent the remainders in terms of u and its derivatives:

$$\frac{u_x}{u}=\left(\frac{S_1}{2}-\sigma_1\right),\quad \frac{u_y}{u}=\frac{1}{2}\left(\frac{S_2}{2}-\sigma_2\right)+\frac{1}{2}\left(\frac{u_x}{u}\right)^2,$$

$$\frac{u_t}{u}=\frac{1}{3}\left(\frac{S_3}{2}-\sigma_3\right)+\frac{u_y u_x}{u^2}-\frac{1}{3}\left(\frac{u_x}{u}\right)^3, \tag{10.12}$$

$$\frac{u_{\tau_4}}{u}=\frac{1}{4}\left(\frac{S_4}{2}-\sigma_4\right)+\frac{u_t u_x}{2}+\frac{1}{2}\left(\frac{u_y}{u}\right)^2-\frac{u_y u_x^2}{u^3}+\frac{1}{4}\left(\frac{u_x}{u}\right)^4.$$

XI. ANOTHER FORM OF ALGEBRO-GEOMETRIC SOLUTION OF KP

In order to derive another form of the explicit solution of KP, we take the initial value as $h_4^{\tau}(p_0,q_0)^T$ and consider

$$\begin{pmatrix} p \\ q \end{pmatrix}=h_1^x h_2^y h_3^t h_4^{\tau_4}\begin{pmatrix} p_0 \\ q_0 \end{pmatrix},$$

$$u=-\langle p,p\rangle,\quad v=\langle q,q\rangle,\quad w=-2uv.$$

The introduction of τ_4 is convenient in the proof of some important facts. Since in the final results there is no τ_4 and the initial data can be chosen arbitrarily, we may let $\tau_4=0$ in the end. Substitute Eq. (9.10) into the second of Eq. (10.12):

$$\begin{aligned} 2\frac{u_y}{u}&=\frac{S_2}{2}-I_2(\Gamma)-\partial_y \ln\frac{\theta_1}{\theta_2}+\partial_x^2 \ln\theta_1\theta_2+\left(\frac{u_x}{u}\right)^2 \\ &=\frac{S_2}{2}-I_2(\Gamma)-(\partial_y+\partial_x^2)\ln\frac{\theta_1}{\theta_2}+2\partial_x^2\ln\theta_1+\left(\frac{u_x}{u}\right)^2. \end{aligned} \tag{11.1}$$

By Eqs. (10.8) and (10.6) we have

$$-(\partial_y+\partial_x^2)\ln\frac{\theta_1}{\theta_2}=(\partial_y+\partial_x^2)\ln u-N_2,\quad N_2=\partial_y\ln u_0. \tag{11.2}$$

Substitute this into Eq. (11.1). By using Eq. (6.12), we have:

$$\partial_x^2\ln\theta_1+uv+\frac{1}{2}\left(\frac{S_2}{2}-I_2(\Gamma)-N_2\right)=0. \tag{11.3}$$

Hence we obtain the solution of the KP equation: $w=-2uv$.

Theorem 11.1: The KP equation (7.2) has the solution:

$$w(x,y,t)=2\partial_x^2\ln\theta(\Omega_1 x+\Omega_2 y+\Omega_3 t+D)+w_0, \tag{11.4}$$

where the constants D and w_0 are

$$D=\phi_0+K+\int_{\infty_1}^{P_0}\omega,\quad w_0=\frac{S_2}{2}-I_2(\Gamma)-N_2.$$

Proof: Since u_0 (therefore N_2) in Eq. (10.8) is independent of x, the only thing required to prove is the independence of N_2 with regard to y and t. Substitute Eq. (9.10) into the third equation of Eq. (10.11):

$$3\frac{u_t}{u}=\frac{S_3}{2}-I_3(\Gamma)-\left(\partial_t+\frac{1}{2}\partial_x^3\right)\ln\frac{\theta_1}{\theta_2}+\frac{3}{2}\partial_y\partial_x\ln\theta_1\theta_2+3\frac{u_yu_x}{u^2}-\left(\frac{u_x}{u}\right)^3$$
$$=\frac{S_3}{2}-I_3(\Gamma)-\left(\partial_t+\frac{1}{2}\partial_x^3+\frac{3}{2}\partial_y\partial_x\right)\ln\frac{\theta_1}{\theta_2}+3\partial_y\partial_x\ln\theta_1+3\frac{u_yu_x}{u^2}-\left(\frac{u_x}{u}\right)^3. \qquad (11.5)$$

Again by Eqs. (10.8) and (10.6) we have

$$-\left(\partial_t+\frac{1}{2}\partial_x^3+\frac{3}{2}\partial_y\partial_x\right)\ln\frac{\theta_1}{\theta_2}=\left(\partial_t+\frac{1}{2}\partial_x^3+\frac{3}{2}\partial_y\partial_x\right)\ln u-N_3,$$
$$N_3=\partial_t\ln u_0. \qquad (11.6)$$

Put it into Eq. (11.5). Taking advantage of Eqs. (6.12) and (6.13), after some calculations we get

$$\partial_y\partial_x\ln\theta_1+(u_xv-uv_x)+\frac{1}{3}\left(\frac{S_3}{2}-I_3(\Gamma)-N_3\right)=0. \qquad (11.7)$$

Note that N_3 is independent of x. Acting with ∂_y,∂_x on Eqs. (11.3) and (11.7), respectively, the difference of the results gives $\partial_yN_2=0$, which means N_2 is independent of y.

To prove the independence of N_2 with regard to t, we treat the fourth member of Eq. (10.11) with the expression σ_4 in Eq. (9.10):

$$4\frac{u_{\tau_4}}{u}=\frac{S_4}{2}-I_4(\Gamma)-(\partial_{\tau_4}+\partial_y\partial_x^2)\ln\frac{\theta_1}{\theta_2}+\left(\frac{4}{3}\partial_t\partial_x+\frac{1}{2}\partial_y^2+\frac{1}{6}\partial_x^4\right)\ln\theta_1\theta_2$$
$$+4\frac{u_tu_x}{u^2}+2\left(\frac{u_y}{u}\right)^2-\frac{4u_yu_x^2}{u^3}+\left(\frac{u_x}{u}\right)^4.$$

Once more by Eqs. (10.8) and (10.6) we get

$$-\left(\partial_{\tau_4}+\partial_y\partial_x^2+\frac{4}{3}\partial_t\partial_x+\frac{1}{2}\partial_y^2+\frac{1}{6}\partial_x^4\right)\ln\frac{\theta_1}{\theta_2}=\left(\partial_{\tau_4}+\partial_y\partial_x^2+\frac{4}{3}\partial_t\partial_x+\frac{1}{2}\partial_y^2+\frac{1}{6}\partial_x^4\right)\ln u-N_4,$$
$$N_4=\partial_{\tau_4}\ln u_0. \qquad (11.8)$$

With the same technique, after tedious calculations we obtain

$$\frac{1}{12}(8\partial_t\partial_x+3\partial_y^2+\partial_x^4)\ln\theta_1+\left(u_{xx}v+uv_{xx}-u_xv_x-\frac{5}{2}u^2v^2\right)+\frac{1}{4}\left(\frac{S_4}{2}-I_4(\Gamma)-N_4\right)=0. \qquad (11.9)$$

Exert $12\partial_x,8\partial_t+\partial_x^3,3\partial_y$ on Eqs. (11.9), (11.3), (11.7), respectively, and cancel the terms containing the derivatives of $\ln\theta_1$. Finally we have

$$0=4\partial_tN_2+\partial_yN_3=5\partial_tN_2,$$

since $\partial_tN_2=\partial_yN_3=\partial_t\partial_y\ln u_0$. Thus N_2 is independent of t. This completes the proof of the theorem.

Note: The algebro-geometric solution of the coupled NS is suggested by Eqs. (10.8) and (11.2):

$$u(x,y)=u(0,0)e^{N_1x+N_2y}\frac{\theta(\Omega_1x+\Omega_2y+\phi_0+K+\eta_2)\,\theta(\phi_0+K+\eta_1)}{\theta(\Omega_1x+\Omega_2y+\phi_0+K+\eta_1)\,\theta(\phi_0+K+\eta_2)},$$
$$v(x,y)=v(0,0)e^{-N_1x-N_2y}\frac{\theta(\Omega_1x+\Omega_2y-\psi_0-K-\eta_2)\,\theta(\psi_0+K+\eta_1)}{\theta(\Omega_1x+\Omega_2y-\psi_0-K-\eta_1)\,\theta(\psi_0+K+\eta_2)}. \tag{11.10}$$

The only thing which should be discussed is the constant N_2 in the expression of $v(x,y)$.

The solution of KP in the form of Eq. (11.4) is obtained in Ref. 9. And the solution of NS in the form of Eq. (11.10) coincides with those found in Refs. 20–23.

ACKNOWLEDGMENTS

This work was partially supported by Hong Kong RGC/97-98/21. Project No. 19671074 was supported by National Natural Science Foundation of China. One of the authors (X.G.G.) would like to thank the Henan Science Foundation Committee of China for financial support.

[1] J. Moser, ''Integrable Hamiltonian system and spectral theory,'' *Proceedings of the 1983 Beijing Symposium on Differential Geometry and Differential Equations* (Science, Beijing, 1986), pp. 157–229.
[2] C. W. Cao and X. G. Geng, in *Nonlinear Physics, Research Reports in Physics*, edited by C. Gu *et al.* (Springer, Berlin, 1990), pp. 68–78.
[3] C. W. Cao, Sci. China, Ser. A **33**, 528 (1990).
[4] C. W. Cao and X. G. Geng, J. Phys. A **23**, 4117 (1990).
[5] M. Antonowicz and S. Rauch-Wojciechowski, J. Phys. A **24**, 5043 (1991).
[6] M. Antonowicz and S. Rauch-Wojciechowski, J. Math. Phys. **33**, 2115 (1992).
[7] Y. T. Wu and T. Mitsui, Jpn. J. Indust. Appl. Math. **13**, 333 (1996).
[8] O. Ragnisco, C. W. Cao, and Y. T. Wu, J. Phys. A **28**, 573 (1995).
[9] E. D. Belokolos, A. I. Bobenko, V. Z. Enol'skii, A. R. Its, and V. B. Matveev, *Algebro-Geometric Approach to Nonlinear Integrable Equations* (Springer, Berlin, 1994).
[10] R. G. Zhou, J. Math. Phys. **38**, 2535 (1997).
[11] B. Konopelchenko, J. Sidorenko, and W. Strampp, Phys. Lett. A **157**, 17 (1991).
[12] Y. Cheng and Y. S. Li, Phys. Lett. A **157**, 22 (1991).
[13] Y. Cheng and Y. S. Li, J. Phys. A **25**, 419 (1992).
[14] W. X. Ma, B. Fuchsteiner, and W. Oevel, Physica A **233**, 331 (1996).
[15] W. X. Ma, Q. Ding, W. G. Zhang, and B. Q. Lu, Nuovo Cimento B **111**, 1135 (1996).
[16] E. H. Bruns, Acta Math. **11**, 25 (1887).
[17] F. Diacu, The Math. Intelligencer **18**, 66 (1996).
[18] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, Berlin, 1978).
[19] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1978).
[20] E. Previato, Duke Math. J. **52**, 329 (1985).
[21] A. Its, Vestn. Leningr. Univ., Fiz., Khim. **7**, 37 (1996).
[22] M. Adams, J. Harnard, and E. Previato, Commun. Math. Phys. **117**, 451 (1988).
[23] S. J. Alber, ''Complex deformation of integrable Hamiltonians over generalized Jacobi varieties,'' in *Nonlinear Processes in Physics, Proceedings of the III Potsdam-V Kiev Workshop* (Clarkson University, Potsdam, NY, 1991), edited by A. S. Fokas, D. J. Kaup, A. S. Newell, and V. E. Zakharov (Springer-Verlag, Berlin, 1993), pp. 6–20.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

A new integrable Davey–Stewartson-type equation

Attilio Maccari
Technical Institute ''G. Cardano,'' Piazza della Resistenza 1, 00015 Monterotondo, Rome, Italy

(Received 9 December 1998; accepted for publication 12 March 1999)

A new integrable nonlinear partial differential equation (PDE) in $2+1$ dimensions is derived starting from the Konopelchenko–Dubrovsky equation. We use an asymptotically exact reduction method based on Fourier expansion and spatio-temporal rescaling and obtain a new integrable Davey–Stewartson-type equation. In order to demonstrate the integrability of the new equation by the inverse scattering method, we apply the reduction technique to the Lax pair of the Konopelchenko–Dubrovsky equation and find the corresponding Lax pair of the new equation. The new equation reduces to the Davey–Stewartson or the nonlinear Schrodinger equation by appropriate limits. © *1999 American Institute of Physics.* [S0022-2488(99)04407-2]

I. INTRODUCTION

As it has been known for some time, very large classes of nonlinear evolution partial differential equations (PDEs) in $1+1$ and $2+1$ dimensions, with a dispersive linear part, can be reduced, by a limiting procedure involving the wave modulation induced by weak nonlinear effects, to a very limited number of ''universal'' nonlinear evolution PDEs. These model equations [of which the nonlinear Schrodinger (NLS) equation in $1+1$ dimension is the most important] appear in many applicative fields (in plasma physics, nonlinear optics, hydrodynamics, etc.) because this reduction technique is able to take into account weakly nonlinear effects. The model equations are integrable, since it is sufficient that the very large class from which they are obtainable contain just one integrable equation, because it is clear from this method that the property of integrability is inherited through the limiting technique.[1–5] This last statement about the integrability is based on heuristic considerations and could not be characterized as rigorous theorem; indeed, no precise definition of integrability is available for nonlinear evolution PDEs.[5]

The reduction method provides a powerful tool to understand the integrability of known equations and to derive new integrable equations likely to be relevant in applicative contexts. It can be also applied to construct approximate solutions for weakly nonlinear ordinary differential equations.[6,7]

The first step of the reduction method is to consider the linearized equation, i.e., the equation obtained after neglecting all the nonlinear terms. The linear evolution is best described by Fourier modes, as they have a constant amplitude and a well defined group velocity (in $2+1$ dimensions) $\mathbf{V}=(V_1(K,K_2),V_2(K_1,K_2))$ depending on the wave numbers K_1,K_2. Subsequently, we introduce the slow variables

$$\xi=\epsilon^p(x-V_1t),\quad \eta=\epsilon^p(y-V_2t),\quad \tau=\epsilon^q t, \tag{1.1}$$

where $p,q>0$, and ϵ is the expansion parameter, supposed to be sufficiently small. If now we introduce the nonlinear terms, the amplitude of the Fourier mode becomes a slowly varying function of space and time and then the rescaled variables ξ,η,τ account for the need to look on larger space and time scales, in order to determine the effects of the nonlinear terms.

In precedent papers, this method has been applied to certain integrable equations in $2+1$ dimensions. The most interesting results are that the Davey–Stewarston equation[8,9] is the typical model equation in $2+1$ dimensions and new integrable nonlinear PDEs can be obtained together with their Lax pair.[10–12] Moreover, the reduction method has been used to derive an equation of applicative relevance in plasma physics.[13,14]

In this paper we consider the Konopelchenko–Dubrovsky (KD) equation[15,16] (integrable by means of the spectral transform[17,18])

$$U_t-U_{xxx}-6\beta UU_x+\tfrac{3}{2}\alpha^2U^2U_x-3W_y+3aU_xW=0, \tag{1.2a}$$

$$W_x=U_y, \tag{1.2b}$$

where $U=U(x,y,t)$, $W=W(x,y,t)$, the subscripts denote partial differentiation and α,β are real parameters.

Applying the reduction method, a new class of integrable equations depending on two real parameters (a,b) can be derived

$$i\Psi_\tau+L_1\Psi+\Psi\Phi+\Psi\chi=0, \tag{1.3a}$$

$$L_2\chi=L_3|\Psi|^2,\quad \Phi_\xi=\chi_\eta\mp(|\Psi|^2)_\eta, \tag{1.3b}$$

where the linear differential operators are given by

$$L_1=\left(\frac{b^2-a^2}{4}\right)\frac{\partial^2}{\partial\xi^2}-a\,\frac{\partial^2}{\partial\xi\partial\eta}-\frac{\partial^2}{\partial\eta^2}, \tag{1.3c}$$

$$L_2=\left(\frac{b^2+a^2}{4}\right)\frac{\partial^2}{\partial\xi^2}+a\,\frac{\partial^2}{\partial\xi\partial\eta}+\frac{\partial^2}{\partial\eta^2}, \tag{1.3d}$$

$$L_3=\pm\frac{1}{4}\left(b^2+a^2+\frac{8b^2(a-1)}{(a-2)^2-b^2}\right)\frac{\partial^2}{\partial\xi^2}\pm\left(a+\frac{2b^2}{(a-2)^2-b^2}\right)\frac{\partial^2}{\partial\xi\partial\eta}\pm\frac{\partial^2}{\partial\eta^2}, \tag{1.3e}$$

and $\Psi=\Psi(\xi,\eta,\tau)$ is complex while $\Phi=\Phi(\xi,\eta,\tau)$, $\chi=\chi(\xi,\eta,\tau)$ are real. If we take $\xi=\eta$, we recover the nonlinear Schrodinger (NLS) equation.

The paper is organized as follows. In the next section we apply the reduction method to the KD equation (1.2) and obtain the new Eq. (1.3). We demonstrate that the equation reduces to the Davey–Stewartson equation if $\alpha=0$, $\beta\neq0$ or $\alpha\neq0$, $\beta=0$ and to the NLS equation if $\xi=\eta$.

In Sec. III we discuss in some detail how the reduction method can be applied to the Lax pair of the KD equation. We demonstrate that the Lax pair of the new Eq. (1.3) is

$$L\hat{\varphi}=0,\quad \hat{\varphi}_\tau+A\hat{\varphi}=0, \tag{1.4a}$$

with

$$L=\begin{pmatrix} \partial_\eta+\dfrac{(a-ib)}{2}\partial_\xi & \dfrac{2-a-ib}{\sqrt{2|(a-2)^2-b^2|}}\Psi \\ \dfrac{2-a+ib}{\sqrt{2|(a-2)^2-b^2|}}\Psi^* & \partial_\eta+\dfrac{(a+ib)}{2}\partial_\xi \end{pmatrix},\quad \varphi=\begin{pmatrix}\varphi_+\\ \varphi_-\end{pmatrix}, \tag{1.4b}$$

and

$$A=\begin{pmatrix} a_1\partial_{\xi\xi}+a_2\chi+a_3|\Psi|^2+a_4\Phi & a_5\Psi_\xi+a_6\Psi\partial_\xi+a_7\Psi_\eta \\ b_5\Psi^*_\xi+b_6\Psi^*\partial_\xi+b_7\Psi^*_\eta & b_1\partial_{\xi\xi}+b_2\chi+b_3|\Psi|^2+b_4\Phi \end{pmatrix}, \tag{1.4c}$$

where

$$a_1=-\frac{ib^2}{2},\quad a_2=\mp\left(\frac{a}{2b}-\frac{a^2}{4b}-\frac{b}{4}+\frac{i}{2}\right),\quad a_3=\frac{a^2}{4b}-\frac{a}{2b}+\frac{b}{4}-\frac{b}{|(a-2)^2-b^2|},$$

$$a_4=\frac{a-2+ib}{2b},\quad a_5=\frac{(a-2+ib)(ia-b)}{2\sqrt{2|(a-2)^2-b^2|}},\quad a_6=\frac{(2-a-ib)b}{\sqrt{2|(a-2)^2-b^2|}},$$

$$a_7=i\frac{(a-2+ib)}{\sqrt{2|(a-2)^2-b^2|}},\quad b_1=\frac{ib^2}{2},\quad b_2=\mp\left(\frac{a}{2b}-\frac{a^2}{4b}-\frac{b}{4}-\frac{i}{2}\right),$$

$$b_3=a_3,\quad b_4=\frac{a-2-ib}{2b},\quad b_5=\frac{(2-a+ib)(ia+b)}{2\sqrt{2|(a-2)^2-b^2|}},$$

$$b_6=a_6,\quad b_7=i\frac{(2-a+ib)}{\sqrt{2|(a-2)^2-b^2|}}. \tag{1.4d}$$

II. A NEW INTEGRABLE NONLINEAR PDE IN 2+1 DIMENSIONS

The linear dispersive part of the Konopelchenko–Dubrovsky Eq. (1.2) admits as a solution a Fourier mode, with a group velocity $\mathbf{V(K)}=(V_1(K_1,K_2),V_2(K_1,K_2))$,

$$V_1(K_1,K_2)=3\left(K_1^2+\frac{K_2^2}{K_1^2}\right),\quad V_2(K_1,K_2)=-\frac{6K_2}{K_1}, \tag{2.1}$$

where

$$\mathbf{V(K)}=\frac{\partial\omega}{\partial\mathbf{K}}, \tag{2.2}$$

and $\omega=\omega(K_1,K_2)$ is the dispersion relation. The envelope of a wave packet concentrated around a value $\mathbf{K}$ travels with the constant group velocity (2.1) and gets slowly dispersed.

We introduce a formal asymptotic Fourier expansion

$$U(x,y,t)=\sum_{n=-\infty}^{n=+\infty}\epsilon^{\gamma_n}\psi_n(\xi,\eta,\tau;\epsilon)\exp\{in(K_1x+K_2y-\omega t)\}, \tag{2.3}$$

with $\gamma_n=|n|$, for $n\neq 0$, $\gamma_0=r$ rational number to be determined later on and $\psi_{-n}(\xi,\eta,\tau;\epsilon)=\psi_n^*(\xi,\eta,\tau;\epsilon)$ [recall that $U(x,y,t)$ is real]. An analogous treatment is assumed for $W(x,y,t)$,

$$W(x,y,t)=\sum_{n=-\infty}^{n=+\infty}\epsilon^{\gamma_n}\phi_n(\xi,\eta,\tau;\epsilon)\exp\{in(K_1x+K_2y-\omega t)\}. \tag{2.4a}$$

The functions $\psi_n(\xi,\eta,\tau,\epsilon)$'s depend on the parameter ϵ and we suppose that the ψ_n's remain finite in the limit $\epsilon\to 0$ and moreover they can be expanded in power series of ϵ, i.e.,

$$\psi_n(\xi,\eta,\tau;\epsilon)=\sum_{i=0}^{\infty}\epsilon^i\psi_n^{(i)}(\xi,\eta,\tau). \tag{2.4b}$$

In the following for simplicity we use the abbreviations $\psi_n^{(0)}=\psi_n$ for $n\neq 0,1$, $\psi_1^{(0)}=\Psi$, $\psi_0^{(0)}=\Psi_0$ (and $\phi_n^{(0)}=\phi_n$ for $n\neq 0,1$, $\phi_1^{(0)}=\phi$, $\phi_0^{(0)}=\Phi$).

The final goal is to obtain the evolution equation for the modulation amplitude $\Psi=\Psi(\xi,\eta,\tau)$ and to understand how it is modified by choosing different wave numbers. We insert the expansions (2.2)–(2.3) into the KD Eq. (1.2) and consider the different equations obtained considering the coefficients of the Fourier modes $\exp(in(K_1 x+K_2 y-\omega t))$. From Eq. (1.2b), we get

$$\phi=\frac{K_2}{K_1}\Psi, \quad \phi^{(p)}=\frac{i}{K_1}\left(\frac{K_2}{K_1}\Psi_{\xi}-\Psi_{\eta}\right), \quad \phi^{(2p)}=\frac{1}{K_1^2}\left(\Psi_{\xi\eta}-\frac{K_2}{K_1}\Psi_{\xi\xi}\right), \tag{2.5a}$$

$$\Phi_{\xi}=\Psi_{0,\eta}. \tag{2.5b}$$

In Eq. (1.2a), we separate the contributions from the linear and nonlinear parts, $\epsilon^{\gamma_n}D_n\psi_n=\epsilon^2F_n$, where D_n is a linear differential operator acting on $\psi_n(\xi,\eta,\tau)$ and F_n is the contribution of the nonlinear part. The operator D_n is

$$\begin{aligned} D_n=&(-in\omega+\epsilon^q\partial_{\tau}-V_1\epsilon^p\partial_{\xi}-V_2\epsilon^p\partial_{\eta})-(inK_1+\epsilon^p\partial_{\xi})^3 \\ &-3(inK_2+\epsilon^p\partial_{\eta})\left(\frac{K_2}{K_1}-\frac{i}{K_1}\epsilon^p\partial_{\eta}+\frac{iK_2}{K_1^2}\epsilon^p\partial_{\xi}+\frac{1}{K_1^2}\epsilon^{2p}\partial_{\xi\eta}-\frac{K_2}{K_1^3}\epsilon^{2p}\partial_{\xi\xi}\right)+o(\epsilon^{3p}). \end{aligned} \tag{2.6}$$

It is easily seen that the first F_n have the following explicit form:

$$F_0=6\left(\beta-\alpha\frac{K_2}{K_1}\right)(|\Psi|^2)_{\xi}\epsilon^p+3\alpha(|\Psi|^2)_{\eta}\epsilon^p+o(\epsilon^{p+2},\epsilon^{p+2r}), \tag{2.7a}$$

$$F_2=3i(\alpha K_2-2\beta K_1)\Psi^2+o(\epsilon^p), \tag{2.7b}$$

$$\begin{aligned} F_1=&-6i\beta K_1\Psi_0\Psi\epsilon^{r-1}+3i\alpha K_1\Phi\psi\epsilon^{r-1}+\tfrac{3}{2}\alpha^2 iK_1|\Psi|^2\Psi\epsilon \\ &+3i(\alpha K_2-2\beta K_1)\psi_2\Psi^*\epsilon+o(\epsilon^{p+r-1},\epsilon^{2p}). \end{aligned} \tag{2.7c}$$

After taking $q=2$, $p=1$, $r=2$ for the proper balance of terms, we obtain at the lowest order for $n=0$ and $n=2$,

$$V_1\Psi_{0,\xi}+V_2\Psi_{0,\eta}+3\Phi_{\eta}+3\alpha(|\Psi|^2)_{\eta}+6\left(\beta-\alpha\frac{K_2}{K_1}\right)(|\Psi|^2)_{\xi}=0, \tag{2.8a}$$

$$\psi_2^{(0)}=\left(\frac{\beta}{K_1^2}-\frac{\alpha K_2}{2K_1^3}\right)\Psi^2, \tag{2.8b}$$

and for $n=1$,

$$i\Psi_{\tau}+\left(3K_1-3\frac{K_2^2}{K_1^3}\right)\Psi_{\xi\xi}+6\frac{K_2}{K_1^2}\Psi_{\xi\eta}-\frac{3}{K_1}\Psi_{\eta\eta}-3\alpha K_1\Phi\Psi+\chi\Psi=0, \tag{2.8c}$$

where

$$\chi=6\beta K_1\Psi_0+N|\Psi|^2, \tag{2.9a}$$

and

$$N=-\frac{3}{2}\alpha^2K_1+\frac{3}{2}\alpha^2\frac{K_2^2}{K_1^3}-6\alpha\beta\frac{K_2}{K_1^2}+6\frac{\beta^2}{K_1}. \tag{2.9b}$$

From Eq. (2.5b), we get

$$6\beta K_1\Phi_{\xi}=\chi_{\eta}-N|\Psi|^2_{\eta}. \tag{2.10}$$

Substituting (2.9a) into (2.8a), we obtain

$$V_1\chi_{\xi\xi}+V_2\chi_{\xi\eta}+3\chi_{\eta\eta}=\left[36\beta K_1\left(\alpha\frac{K_2}{K_1}-\beta\right)+V_1N\right]\times(|\Psi|^2)_{\xi\xi}+[V_2N-18\alpha\beta K_1](|\Psi|^2)_{\xi\eta}+3N(|\Psi|^2)_{\eta\eta}. \quad (2.11)$$

If in Eqs. (2.8c)–(2.10)–(2.11) we set $\alpha=0$, $\beta\neq 0$ or $\alpha\neq 0$, $\beta=0$ then after trivial rescalings we obtain the Davey–Stewartson equation. Moreover, we get the unidimensional NLS equation if $\xi=\eta$.

If $\alpha\neq 0$, $\beta\neq 0$, after the cosmetic rescaling

$$\xi'=-\frac{\alpha}{2\sqrt{3}\beta}\xi,\quad \eta'=\frac{1}{\sqrt{3}}\eta,\quad \chi'=K_1\chi,\quad \Phi'=-3\alpha K_1^2\Phi, \quad (2.12a)$$

$$\tau'=\frac{\tau}{K_1},\quad \Psi'=\sqrt{|N|K_1}\Psi, \quad (2.12b)$$

and with the introduction of two real parameters,

$$a=\frac{\alpha K_2}{\beta K_1},\quad b=\frac{\alpha K_1}{\beta}, \quad (2.12c)$$

we arrive at the model equation of Davey–Stewartson-type (1.3).

Integrable Davey–Stewartson-type equations has been extensively investigated by many authors.[19–22] A very detailed list of Davey–Stewartson equations integrable by the inverse scattering method has been recently given.[23] Equation (1.3) does not appear in these papers. We expect that this new equation be integrable by the inverse scattering method, because it has been obtained from an integrable equation and the property of integrability is supposed to keep through the application of the reduction method. The integrability of Eq. (1.3) will be explicitly demonstrated in the next section.

III. THE LAX PAIR FOR THE MODEL NONLINEAR PDE

In order to demonstrate the integrability of the new Eq. (1.3) by the inverse scattering spectral transform, we apply the reduction method to the Lax pair of the KD equation. The Lax operators are

$$L=\frac{\partial}{\partial y}+\frac{\partial^2}{\partial x^2}+\alpha U(x,y,t)\frac{\partial}{\partial x}+\beta U(x,y,t),\quad L\varphi(x,y,t)=0, \quad (3.1)$$

$$A=-4\frac{\partial^3}{\partial x^3}-6\alpha U(x,y,t)\frac{\partial^2}{\partial x^2}-3\beta U_x(x,y,t)-\frac{3}{2}\alpha\beta U^2(x,y,t)+3\beta W(x,y,t)-3\alpha U_x(x,y,t)\frac{\partial}{\partial x}-\frac{3}{2}\alpha^2U^2(x,y,t)\frac{\partial}{\partial x}-6\beta U(x,y,t)\frac{\partial}{\partial x}+3\alpha W(x,y,t)\frac{\partial}{\partial x}, \quad (3.2)$$

$$\varphi_t(x,y,t)+A\varphi(x,y,t)=0. \quad (3.3)$$

We introduce an asymptotic Fourier expansion

$$\varphi(x,y,t)=\sum_{n=-\infty}^{n=+\infty}\epsilon^{\gamma_n}\varphi_n(\xi,\eta,\tau;\epsilon)\exp\left(i\frac{n}{2}z+(\lambda_1x+\lambda_2y+\lambda_3t)\right), \quad (3.4)$$

where n is odd, $z=K_1x+K_2y-\omega t$, $\gamma_{n+2}=1+\gamma_n$, $\gamma_n=\gamma_{-n}$ for $n>0$, the $\varphi_n(\xi,\eta,\tau;\epsilon)$'s are functions which parametrically depend on ϵ and remain finite if $\epsilon\rightarrow 0$ and $\lambda_1,\lambda_2,\lambda_3$ are real constants to be determined later on. Substituting the expression for $\varphi(x,y,t)$ in Eq. (3.1), the coefficients of the Fourier modes generate a series of relations. Obviously, each relation must be valid for a given order of approximation in ϵ. In particular, for the fundamental harmonics $\varphi_\pm(x,y,t)=\varphi_{\pm 1}(x,y,t)$, considering terms $O(\epsilon^0)$ in (3.1) and (3.3), the constants $\lambda_1,\lambda_2,\lambda_3$ can be fixed,

$$\left(\pm\frac{iK_1}{2}+\lambda_1\right)^2+\left(\pm\frac{iK_2}{2}+\lambda_2\right)=0, \tag{3.5a}$$

$$\left(\mp\frac{i\omega}{2}+\lambda_3\right)-4\left(\pm\frac{iK_1}{2}+\lambda_1\right)^3=0, \tag{3.5b}$$

and then

$$\lambda_1=-\frac{K_2}{2K_1},\quad \lambda_2=\frac{K_1^2}{4}-\frac{K_2^2}{4K_1^2},\quad \lambda_3=-\frac{K_2^3}{2K_1^3}+\frac{3}{2}K_2K_1. \tag{3.6}$$

The new spectral problem is obtained by means of the successive order ϵ for Eq. (3.1),

$$\varphi_{+,\eta}+\left(iK_1-\frac{K_2}{K_1}\right)\varphi_{+,\xi}+\left[\beta-\frac{\alpha}{2}\left(\frac{K_2}{K_1}+iK_1\right)\right]\Psi\varphi_-=0, \tag{3.7a}$$

$$\varphi_{-,\eta}-\left(iK_1+\frac{K_2}{K_1}\right)\varphi_{-,\xi}+\left[\beta-\frac{\alpha}{2}\left(\frac{K_2}{K_1}-iK_1\right)\right]\Psi^*\varphi_+=0. \tag{3.7b}$$

With the rescaling (2.12), we obtain

$$\varphi_{+,\eta}+\left(\frac{a-ib}{2}\right)\varphi_{+,\xi}+\frac{2-a-ib}{\sqrt{2|(a-2)^2-b^2|}}\Psi\varphi_-=0, \tag{3.8a}$$

$$\varphi_{-,\eta}+\frac{a+ib}{2}\varphi_{-,\xi}+\frac{2-a+ib}{\sqrt{2|(a-2)^2-b^2|}}\Psi^*\varphi_+=0. \tag{3.8b}$$

The spectral problem can be again obtained, when we consider the temporal evolution Eq. (3.3) at the same order ϵ. Only if we take into account the next orders of approximation of Eq. (3.3), i.e., the order ϵ^2, the temporal evolution can be determined. However, new quantities, the corrections $\tilde{\varphi}_\pm(\xi,\eta,\tau)$ of order ϵ to the fundamental harmonics $\varphi_\pm(\xi,\eta,\tau)$, appear. These unknown quantities can be eliminated in Eq. (3.3) taking advantage of the relation obtained from Eq. (3.1), considering terms of order ϵ^2. This elimination is possible only because Eqs. (3.1) and (3.3) are identical at the orders ϵ and ϵ^2. After a lengthy calculation we arrive at the final form for the operator A which is given in Eqs. (1.4c)–(1.4d). The compatibility condition between equations (1.4a),

$$L_\tau=[L,A]=LA-AL, \tag{3.9}$$

leads to the new integrable Eq. (1.3).

IV. CONCLUSION

The reduction method is a powerful tool to identify new nonlinear PDEs integrable that are also likely to be applicable in a regime of weak nonlinearity.

In this paper, we have derived a new integrable nonlinear evolution Davey–Stewartson-type equation from the KD equation, by means of a reduction method based on Fourier expansion and space–time rescalings. It reduces to the ordinary Davey–Stewartson equation for appropriate values of its parameters. Moreover, we have applied the reduction method to the Lax pair of the original equation and have demonstrated the integrability property of the new equation, by exhibiting the corresponding Lax pair.

[1] F. Calogero and W. Eckhaus, ''Necessary conditions for integrability of nonlinear PDEs,'' Inverse Probl. **3**, L27–L32 (1987).
[2] F. Calogero and W. Eckhaus, ''Nonlinear evolution equations, rescalings, model PDEs and their integrability. I,'' Inverse Probl. **3**, 229–262 (1987).
[3] F. Calogero and W. Eckhaus, ''Nonlinear evolution equations, rescalings, model PDEs, and their integrability. II,'' Inverse Probl. **4**, 11–33 (1988).
[4] F. Calogero and A. Maccari, ''Equations of nonlinear Schrodinger-type in $1+1$ and $2+1$ dimensions obtained from integrable PDE,'' ''Inverse problems: An interdisciplinary study,'' in Proceedings of the meeting on inverse problems, Montpellier, 1986, *Advances in Electronics and Electron Physics*, edited by P. C. Sabatier (Academic, New York, 1988), Vol. 19, pp. 463–480.
[5] F. Calogero, ''Why are certain nonlinear PDEs both widely applicable and integrable,'' in *What is Integrability (for Nonlinear PDEs)*, edited by V. E. Zakharov (Springer, Berlin, 1991).
[6] A. Maccari, ''Dissipative bidimensional systems and resonant excitations,'' Int. J. Nonlinear Mech. **33**, 713–726 (1998).
[7] A. Maccari, ''Approximate solution of a class of nonlinear oscillators in resonance with a periodic excitation,'' Nonlinear Dynamics **15**, 329–343 (1998).
[8] A. Davey and K. Stewartson, ''On three-dimensional packets of surface waves,'' Proc. R. Soc. London, Ser. A **338**, 101–110 (1974).
[9] D. Anker and N. C. Freeman, ''On the soliton solutions of the Davey–Stewartson equation for long waves,'' Proc. R. Soc. London, Ser. A **360**, 529–540 (1978).
[10] A. Maccari, ''The Kadomtsev–Petviashvili equation as a source of integrable model equations,'' J. Math. Phys. **37**, 6207–6212 (1996).
[11] A. Maccari, ''The Boussinesq equation as a source of equations integrable by the spectral transform,'' Nonlinearity **10**, 849–856 (1997).
[12] A. Maccari, ''A generalized Hirota equation in $2+1$ dimensions,'' J. Math. Phys. **39**, 6547–6551 (1998).
[13] A. Maccari, ''A model equation for nonresonant interacting ion acoustic plasma waves,'' J. Plasma Phys. **60**, 275–288 (1998).
[14] A. Maccari, ''Nonresonant interacting ion acoustic waves in a magnetized plasma,'' J. Phys. A **32**, 693–709 (1999).
[15] B. Konopelchenko and V. Dubrovsky, ''Some new integrable nonlinear evolution equations in $2+1$ dimensions,'' Phys. Lett. **102A**, 15–17 (1984).
[16] B. Konopelchenko, *Introduction to Multidimensional Integrable Equations* (Plenum, New York, 1992).
[17] Z. Jiang and R. K. Bullough, ''Combined $\bar{\partial}$ and Riemann–Hilbert inverse methods for integrable nonlinear evolution equations in $(2+1)$ dimensions,'' J. Phys. A **20**, L429–L435 (1987).
[18] Z. Jiang, ''Construction of scattering data for a class of multidimensional scattering operators,'' Inverse Probl. **5**, 349–374 (1989).
[19] A. S. Fokas, ''On the simplest integrable equation in $2+1$,'' Inverse Probl. **10**, L19–L22 (1994).
[20] E. I. Shul'man, ''On the integrability of equations of Davey–Stewartson-type,'' Theor. Math. Phys. **56**, 720–724 (1984).
[21] A. S. Fokas and P. M. Santini, ''Recursion operators and bi-Hamiltonian structures in multidimensions. I,'' Commun. Math. Phys. **115**, 375–419 (1988).
[22] F. Pempinelli, M. Boiti, and P. C. Sabatier, ''First and second order nonlinear evolution equations from an inverse spectral problem,'' Inverse Probl. **9**, 1–37 (1993).
[23] A. Folguera, ''Davey–Stewartson-type equations of the second and third order: Derivation and classification,'' J. Math. Phys. **39**, 2947–2968 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Equivalence classes of perturbations in cosmologies of Bianchi types I and V: Formulation

Zbigniew Banach
Centre of Mechanics, Institute of Fundamental Technological Research,
Department of Fluid Mechanics, Polish Academy of Sciences,
Swietokrzyska 21, 00-049 Warsaw, Poland

(Received 20 October 1998; accepted for publication 10 March 1999)

In this paper we deal for the first time with gauge-invariant perturbations of anisotropic cosmological models of Bianchi types I and V from a unified point of view. Motivated by Ehlers' pioneering concepts, the key idea is to identify the gauge-invariant perturbations with the equivalence classes of tangents to one-parameter families of exact solutions to Einstein's field equations. For cases where these models are filled with a nonbarotropic perfect fluid, we show that a set of 26 ''geometrically'' independent, not identically vanishing gauge-invariant variables, denoted collectively by **D** and referred to as the complete set of basic variables, can be used to extract the equivalence classes of tangents from **D** in a unique way. The set **D** is complete because it has the following property: any gauge-invariant quantity is obtainable linearly from the basic variables through purely algebraic and differential operations. Mathematically, this approach to the gauge problem is a nontrivial example of the general scheme that we have described in our two previous papers [Int. J. Theor. Phys. **36**, 1787, 1817 (1997)], and the new concepts developed were also applied to the construction of a complete set of basic gauge-invariant variables for the cases of a fixed background de Sitter space–time and an almost-Robertson–Walker universe model. Arguments are given that there are a number of advantages to be gained by replacing the coordinate-based method of Bardeen or the covariant formalism of Ellis and Bruni by the present one. © *1999 American Institute of Physics.* [S0022-2488(99)03608-7]

I. INTRODUCTION

Beginning from Einstein's theory of gravity, considerable efforts have been expended in attempts to define gauge-invariant perturbations of homogeneous and isotropic cosmological models. The coordinate-based method of Bardeen[1] and the covariant formalism of Ellis and Bruni[2] are a selection of the more important comprehensive treatments. In the pioneering paper by Bardeen,[1] *before introducing a set of gauge-invariant variables*, a separation was made of the metric and matter perturbations into scalar, vector, and tensor parts. However, as noted already by Ellis and Matravers,[3] ''that separation is non-locally defined[4] and is in effect dependent on the coordinate choice made (for example, the Bardeen formalism is gauge invariant only if one restricts coordinate changes to the vector kinds when vector modes are investigated, and to the scalar kind when the scalar modes are the theme of interest).'' Consequently, the coordinate-based analysis of cosmological perturbations, while it helps us understand some important problems,[1] is not a fully gauge-invariant activity.

As an alternative to this line of development, Ellis and Bruni[2] proposed a covariant treatment of perturbations in Robertson–Walker space–times. After a brief outline of the Stewart–Walker approach to gauge-invariant perturbations (see Lemma 2.2 in Ref. 5), the basic idea of Ellis and Bruni was to introduce geometrically defined exact variables A (i.e., these variables are meaningful in any space–time) such that their values A_0 in a Robertson–Walker universe vanish. Then, because of Lemma 2.2 of Stewart and Walker,[5] the quantity A itself is a gauge-invariant perturbation in an almost-Robertson–Walker universe, and its physical significance is apparent through the covariant definition. As in the work of Bruni *et al.*[6] and Dunsby *et al.*[7] (see also the method of Ellis *et al.*[8]), we shall refer to such quantities as *covariant gauge-invariant variables*. It was the aim of Dunsby[9] to further extend this approach to perturbations of Bianchi type-I cosmological models. In his paper he illustrates one of the main advantages of using the covariant approach. Since the Stewart–Walker lemma is valid for any background space–time, one can often consider the same gauge-invariant variables in perturbing different universe models; in particular, covariant gauge-invariant variables can be easily identified in perturbing homogeneous anisotropic space–times. Thus, it is easy to see from Dunsby's analysis of the Bianchi type-I model that the covariant extension to Bianchi type V is very straightforward.

Recently, Banach and Piekarski[10,11] introduced new geometric techniques to develop linear perturbation theory for an arbitrary system of diffeomorphism-invariant, covariant field equations.[12,13] Given the nonlinear field equations of Einstein's gravity theory, these techniques were applied to the study of infinitesimal perturbations in Robertson–Walker universe models.[14–19] In this and the companion paper,[20] we demonstrate how the Banach–Piekarski formalism can be used to address the issue of describing the equivalence classes of perturbations (tangents) in a Bianchi type-I or Bianchi type-V universe[21,22] dominated by a nonbarotropic perfect fluid.[23,24] The crucial point to observe is that, for the aforementioned matter source and background models, a satisfactory (i.e., unique) characterization of cosmological perturbations can be obtained if one defines in a suitable way 26 ''geometrically'' independent, not identically vanishing gauge-invariant variables. The set consisting of these *basic variables*, denoted by **D** and referred to as the *complete set*, is important because it enables one to divide the infinitesimal perturbations into physically relevant equivalence classes: two infinitesimal perturbations $\delta\mathcal{F}_0$ and $\delta\mathcal{F}_0'$ are said to be equivalent if there is a vector field v on the space–time manifold X such that $\delta\mathcal{F}_0'$ differs from $\delta\mathcal{F}_0$ by the action of the Lie derivative $\mathcal{L}_v$ on the background solution $\mathcal{F}_0$ to nonlinear field equations.[25] Thus, it appears that $\delta\mathcal{F}_0$ and $\delta\mathcal{F}_0+\mathcal{L}_v\mathcal{F}_0$ represent the same perturbation of $\mathcal{F}_0$, and clearly $\delta\mathcal{F}_0$ satisfies the linearized field equations if and only if $\delta\mathcal{F}_0+\mathcal{L}_v\mathcal{F}_0$ does.[26] The situation may therefore be summarized as follows. The object of most physical interest is not just one perturbation $\delta\mathcal{F}_0$, but a whole equivalence class of all perturbations $\delta\mathcal{F}_0'$ that are equivalent to $\delta\mathcal{F}_0$. The equivalence class of $\delta\mathcal{F}_0$ is denoted $[\delta\mathcal{F}_0]$ and is called the *gauge-invariant perturbation* associated with $\delta\mathcal{F}_0$. Here, we prove that the complete set of basic variables, namely **D**, provides a mathematically simplest representation of the gauge-invariant perturbation $[\delta\mathcal{F}_0]$. As a matter of fact, $[\delta\mathcal{F}_0]$ is uniquely determined from **D** and vice versa. This important property of **D** serves to illustrate one of the key features of the present formalism: that is, because we start with a complete set of basic gauge-invariant variables, valid for an almost-Bianchi type-I or almost-Bianchi type-V universe model, it is possible to identify $[\delta\mathcal{F}_0]$ with **D** and thereby exploit the theory based on **D** as a *fundamental concept* for the description of infinitesimal perturbations.

Since the perturbations of anisotropic cosmological models have been studied before by several authors,[9,27–31] the natural question is the following: why do we need yet another formalism for cosmological perturbations when there are already so many on the market? First, note that in Refs. 9 and 27–31 the equivalence classes of perturbations (tangents) are not explicitly described. This is, without a doubt, a very important problem to consider.[25] Second, the previous papers concentrated upon the Bianchi type-I background model and thus did not analyze linear perturbations in a Bianchi type-V cosmology. Third, a decisive motivation for our work appears in the following crucial fact:[20] a knowledge of **D** is all one needs in the sense that if A denotes any gauge-invariant tensor field defined on X, then A is *obtainable linearly* from the basic gauge-invariant variables **D** through *purely algebraic and differential operations*. By contrast, as discussed in a companion paper,[20] one may observe that any description that chooses, at the outset, to introduce covariant gauge-invariant variables or to start with Lemma 2.2 of Stewart and Walker[5] is *necessarily incomplete*. Finally, the technique we present here could also be used to resolve disagreements between the results of different existing formalisms. Specifically, we are able to explain the coordinate-based method of Tomita and Den,[27] Den,[28,29] and Noh and Hwang[30,31] in such a way that the objections raised to this method by Ellis and Matravers[3] do not apply. For more details, one should consult Ref. 20. To sum up, in our opinion, the present formulation of linear perturbation theory is *canonical* because it provides a complete framework both for constructing all gauge-invariant variables and for determining the equivalence classes of perturbations.

In a separate paper,[32] we will derive a full set of propagation equations that involves only **D**. These equations are deterministic, i.e., the local existence and uniqueness of solutions can be demonstrated for ''arbitrary'' initial data. As a consequence, if we introduce $[\delta\mathcal{F}_0]$ and describe $[\delta\mathcal{F}_0]$ in terms of **D**, then we will remove the necessity for finding a satisfactory specific gauge or for referring to spurious ''gauge mode'' solutions. However, one expected feature from such an approach is that since the unperturbed shear tensor does not vanish, the dynamics of linearized perturbations will be extremely complicated. Thus, except for some particularly simple situations [e.g., the background spatial curvature is negligible (Bianchi type-I space–times), the metric is axially symmetric in both the background and the perturbed model, the ratio of pressure to energy density in the background is independent of time, etc.], we may be no longer able to analyze the propagation equations analytically. The implications of this observation are considered in some detail in Sec. V.

Here we proceed as follows. Restricting attention to the case of a nonbarotropic perfect fluid, in Sec. II a number of tensor fields appropriate to the specification of infinitesimal perturbations are introduced. Within the covariant formalism, in Sec. III the background models are briefly reviewed: Bianchi types I and V. The aim in Sec. IV is to prove that the equivalence classes of perturbations (tangents) can be described in terms of a finite set of gauge-invariant variables. Section V is for final remarks.

It is assumed throughout that the space–time metric has signature $(-,+,+,+)$, and the speed of light is taken to be unity $(c=1)$. We choose units so that the Einstein gravitational constant equals one $(\kappa=8\pi G=1)$. Lower case Greek characters $\alpha,\beta,\gamma,\ldots$, refer to space–time indices. By definition, $T_{\alpha\beta}$ is the energy–momentum tensor that takes the perfect fluid form.

II. LINEARIZATION PROCEDURE

A. Field equations for nonbarotropic perfect fluids

In general relativity, the metric of space–time is assumed to obey Einstein's field equations, given by

$$R_{\alpha\beta}-\tfrac{1}{2}R^{\mu}{}_{\mu}g_{\alpha\beta}+\Lambda g_{\alpha\beta}=T_{\alpha\beta}, \tag{2.1}$$

in the standard notation. Consider the case of a perfect fluid. If u^α is the four-velocity of the fluid (normalized by $u^\alpha u_\alpha=-1$), we find that the energy–momentum tensor $T_{\alpha\beta}$ can be decomposed as

$$T_{\alpha\beta}=(e+p)u_\alpha u_\beta+pg_{\alpha\beta}, \tag{2.2}$$

where e is the energy density and p is the pressure. Additionally, as we are considering nonbarotropic perfect fluids,[23,24] it will be natural to introduce a number flux density N^α. This density is required to satisfy the continuity equation:

$$N^\alpha{}_{;\alpha}=0. \tag{2.3}$$

Decomposing N^α with respect to u^α, we immediately see that

$$N^\alpha=nu^\alpha. \tag{2.4}$$

In the rest frame of the fluid, n is the number density. Another useful quantity is the temperature of the fluid; we denote this temperature by T. For nonbarotropic perfect fluids, we regard n and T as independent quantities and postulate the following equations of state:

$$e = e(n,T), \quad p = p(n,T). \tag{2.5}$$

Because of the first law of thermodynamics,[24] the problem of explicitly determining these equations of state reduces to the problem of explicitly expressing the specific free energy in terms of n and T [see, e.g., Eqs. (5.17) and (5.18) in Ref. 17]. However, the present general form of Eqs. (2.5) suffices for our purposes here.

It will be convenient to take the contravariant form $g^{\alpha\beta}$ of the metric $g_{\alpha\beta}$ to be more fundamental and the covariant form $g_{\alpha\beta}$ as derived from it by the relations

$$g^{\alpha\mu} g_{\mu\beta} = \delta^{\alpha}{}_{\beta}, \tag{2.6}$$

where $\delta^{\alpha}{}_{\beta}$ is the Kronecker delta. As a result of this convention, a perfect fluid flow in a curved space–time is described by giving the fields $g^{\alpha\beta}$, u^{α}, n, and T. For brevity, we denote these primary fields by $\mathcal{F}$:

$$\mathcal{F} := \{g^{\alpha\beta}, u^{\alpha}, n, T\}. \tag{2.7}$$

Ultimately, therefore, the system of equations for the specification of $\mathcal{F}$ consists of Eqs. (2.1)–(2.6), the others being only consequences of them.

B. Infinitesimal perturbations

Quite often, the treatment of nonbarotropic perfect fluids by means of exact, nonlinear field equations is far too cumbersome and unnecessarily detailed. Fortunately, the most practical problems can be adequately treated in the much simpler framework of linear perturbation theory. The basic assumption of this theory, *which seems necessary in order to give a clear idea of what the perturbation method is to be* (see, e.g., the discussion in Sec. 7.1 of Ref. 25), may be formulated as follows: Consider an open interval $I := (-d, d)$ of $\mathbb{R}$, $d > 0$; then, for each $\epsilon \in I$ there exists a *classical* solution,

$$\mathcal{F}_{\epsilon}(x^{\mu}) := \{g^{\alpha\beta}(x^{\mu}, \epsilon), u^{\alpha}(x^{\mu}, \epsilon), n(x^{\mu}, \epsilon), T(x^{\mu}, \epsilon)\} \tag{2.8}$$

of the exact, nonlinear field equations. The parameter ϵ appearing in this definition measures the size of the perturbation, in the sense that $\mathcal{F}_{\epsilon}(x^{\mu})$ depends continuously and differentiably on $\epsilon \in I$ for each (x^{μ}), and

$$\mathcal{F}_0(x^{\mu}) := \mathcal{F}_{\epsilon}(x^{\mu})_{|\epsilon=0} \tag{2.9}$$

is a background solution. Now, putting Eqs. (2.8) and (2.9) together, we have

$$\mathcal{F}_0(x^{\mu}) := \{q^{\alpha\beta}(x^{\mu}), w^{\alpha}(x^{\mu}), n_0(x^{\mu}), T_0(x^{\mu})\}, \tag{2.10}$$

where

$$q^{\alpha\beta}(x^{\mu}) := g^{\alpha\beta}(x^{\mu}, \epsilon)_{|\epsilon=0}, \quad w^{\alpha}(x^{\mu}) := u^{\alpha}(x^{\mu}, \epsilon)_{|\epsilon=0}, \tag{2.11a}$$

$$n_0(x^{\mu}) := n(x^{\mu}, \epsilon)_{|\epsilon=0}, \quad T_0(x^{\mu}) := T(x^{\mu}, \epsilon)_{|\epsilon=0}, \quad T_0 > 0. \tag{2.11b}$$

Here we postulate that $\mathcal{F}_0$ is obtainable from a solution of the ''unperturbed'' cosmological equations and that the forms of $q^{\alpha\beta}$, w^{α}, n_0, and T_0 are consistent with the background space–time geometry, which is that of a Bianchi type-I or Bianchi type-V space–time.[21,22]

Since $\mathcal{F}_{\epsilon}$ depends differentiably on $\epsilon \in I$, it will be possible to define the perturbation of $\mathcal{F}_0$ by the formula

$$\delta\mathcal{F}_0 := \{G^{\alpha\beta}, U^{\alpha}, n_0 M, T_0 K\}, \tag{2.12}$$

in which

$$G^{\alpha\beta} := \left(\frac{\partial g^{\alpha\beta}}{\partial \epsilon}\right)_{\epsilon=0}, \quad U^{\alpha} := \left(\frac{\partial u^{\alpha}}{\partial \epsilon}\right)_{\epsilon=0}, \tag{2.13a}$$

$$M := \frac{1}{n_0}\left(\frac{\partial n}{\partial \epsilon}\right)_{\epsilon=0}, \quad K := \frac{1}{T_0}\left(\frac{\partial T}{\partial \epsilon}\right)_{\epsilon=0}. \tag{2.13b}$$

We call $\delta\mathcal{F}_0$ the *infinitesimal perturbation* of $\mathcal{F}_0$. It is important to stress that the infinitesimal perturbation so defined has the absolute geometrical meaning independent of any particular choice of the coordinate system (x^{μ}) in X.[10,11] So as to derive a closed set of governing equations for perturbation for the system, we must first differentiate Einstein's field equations (2.1) and the equation of balance of number density (2.3) with respect to ϵ at $\epsilon=0$ and then apply the definitions of various quantities involved. In this way, we obtain a linear equation for $\delta\mathcal{F}_0$, i.e., an equation that can be expressed in the form[26]

$$\mathcal{Y}(\delta\mathcal{F}_0) = 0, \tag{2.14}$$

where $\mathcal{Y}$ is a linear differential space–time operator acting on $\delta\mathcal{F}_0$. As to the explicit characterization of $\mathcal{Y}$, it will be considered in a separate paper.[32]

Given the background space–time metric $q_{\alpha\beta}$ $(q^{\alpha\mu}q_{\mu\beta} = \delta^{\alpha}{}_{\beta})$ and a geometrically preferred (unperturbed) four-velocity w^{α} $(w_{\alpha} := q_{\alpha\mu}w^{\mu})$, the tensor $\gamma_{\alpha\beta} := q_{\alpha\beta} + w_{\alpha}w_{\beta}$ projects into the local rest spaces of comoving observers:

$$\gamma^{\alpha\beta} := q^{\alpha\beta} + w^{\alpha}w^{\beta}, \quad \gamma^{\alpha}{}_{\beta} := \delta^{\alpha}{}_{\beta} + w^{\alpha}w_{\beta}. \tag{2.15}$$

With the help of these projection tensors, we find that the metric and velocity perturbations, namely $G^{\alpha\beta}$ and U^{α}, have the covariant irreducible decompositions,

$$G^{\alpha\beta} = w^{\alpha}w^{\beta}Q + w^{\alpha}Q^{\beta} + w^{\beta}Q^{\alpha} + 2D\gamma^{\alpha\beta} + 2F^{\alpha\beta}, \tag{2.16a}$$

$$U^{\alpha} = Vw^{\alpha} + V^{\alpha}, \tag{2.16b}$$

where the scalars (Q,D,V), the spatial vectors (Q^{α},V^{α}), and a spatial two-tensor $F^{\alpha\beta}$ (which is symmetric and trace-free) are related to $\mathcal{F}_0$ and $\delta\mathcal{F}_0$ by

$$Q := w_{\mu}w_{\nu}G^{\mu\nu}, \quad Q^{\alpha} := -w_{\mu}\gamma^{\alpha}{}_{\nu}G^{\mu\nu}, \tag{2.17a}$$

$$D := \tfrac{1}{6}\gamma_{\mu\nu}G^{\mu\nu}, \quad F^{\alpha\beta} := \tfrac{1}{2}\gamma^{\alpha}{}_{\mu}\gamma^{\beta}{}_{\nu}G^{\mu\nu} - D\gamma^{\alpha\beta}, \tag{2.17b}$$

$$V := -w_{\mu}U^{\mu}, \quad V^{\alpha} := \gamma^{\alpha}{}_{\mu}U^{\mu}. \tag{2.17c}$$

Associated with $\delta\mathcal{F}_0$ is the object

$$J(\delta\mathcal{F}_0) := \{Q,Q^{\alpha},D,F^{\alpha\beta},V,V^{\alpha},M,K\}, \tag{2.18}$$

which we also call the *infinitesimal perturbation* of $\mathcal{F}_0$. This object contains the same information as $\delta\mathcal{F}_0$, and we can use $\mathbf{W} := J(\delta\mathcal{F}_0)$ in place of $\delta\mathcal{F}_0$. Such is indeed the case because, after specifying the background solution (i.e., $\mathcal{F}_0$), $\delta\mathcal{F}_0$ is uniquely determined from $\mathbf{W}$, and conversely. Defining $Y(\mathbf{W})$ by $Y(\mathbf{W}) := \mathcal{Y}(\delta\mathcal{F}_0)$, it follows that if $\mathcal{Y}(\delta\mathcal{F}_0) = 0$, then

$$Y(\mathbf{W}) = 0. \tag{2.19}$$

In the present paper, however, we consider the infinitesimal perturbations of $\mathcal{F}_0$ from a purely geometric point of view. Consequently, $\delta\mathcal{F}_0$ and $\mathbf{W}$ *are not necessarily assumed* to satisfy Eqs.

(2.14) and (2.19). This observation serves as an alternative, and in certain respects a more general and enlightening, starting point from which one could introduce and describe the equivalence classes of perturbations.[10,11]

III. CHARACTERIZATION OF THE BACKGROUND MODELS: BIANCHI TYPES I AND V

In this section we recall the more relevant properties of anisotropic background models of Bianchi types I and V; the discussion will be confined only to what is necessary for our aim. However, before briefly describing these background models, we have to define a few mathematical quantities. Thus, it is assumed that $\boldsymbol{\nabla}_\alpha$ is the covariant derivative based upon the background metric $q_{\alpha\beta}$. We use the standard kinematical decomposition of $\boldsymbol{\nabla}_\alpha$. More explicitly, a *dot* denotes the covariant derivative along the unperturbed fluid flow lines and a *slash* corresponds to the three-dimensional covariant derivative defined by totally projecting the covariant derivative $\boldsymbol{\nabla}_\alpha$ orthogonal to w^α, so, for example, $\dot{A}_{\alpha\cdots}:=w^\beta\,\boldsymbol{\nabla}_\beta A_{\alpha\cdots}$ and $A_{\alpha\cdots|\beta}:=\gamma^\mu{}_\alpha\gamma^\nu{}_\beta\cdots\boldsymbol{\nabla}_\nu A_{\mu\cdots}$. For $A_{\alpha\cdots|\beta}$, a certain care is needed if the tensor $A_{\alpha\ldots}$ itself is not totally orthogonal to w^α, but this will not be a problem here. As usual, round brackets enclosing indices denote symmetrization and square brackets denote antisymmetrization.

Let $h_{\alpha\beta}$ be the projection tensor into the tangent three-spaces orthogonal to u^α $(h_{\alpha\beta}:=g_{\alpha\beta}+u_\alpha u_\beta \Rightarrow h^\alpha{}_\mu h^\mu{}_\beta=h^\alpha{}_\beta$ and $h_\alpha{}^\beta u_\beta=0)$. Then the first covariant derivative of the four-velocity vector can be written as

$$u_{\alpha;\beta}=\omega_{\alpha\beta}+\sigma_{\alpha\beta}+\tfrac{1}{3}\theta h_{\alpha\beta}-a_\alpha u_\beta, \tag{3.1}$$

where $\omega_{\alpha\beta}$ is the vorticity tensor $(\omega_{\alpha\beta}=\omega_{[\alpha\beta]},\, \omega_{\alpha\beta}u^\beta=0)$, $\sigma_{\alpha\beta}$ is the shear tensor $(\sigma_{\alpha\beta}=\sigma_{(\alpha\beta)},\, g^{\alpha\beta}\sigma_{\alpha\beta}=0,\, \sigma_{\alpha\beta}u^\beta=0)$, θ is the expansion, and a_α is the acceleration $(a^\alpha u_\alpha=0)$. In an exact Bianchi type-I or Bianchi type-V universe model, the following conditions are automatically satisfied:

$$\omega_{\alpha\beta}=0, \quad a_\alpha=0, \tag{3.2a}$$

$$h^{\alpha\beta}\theta_{;\beta}=0, \quad h^{\alpha\beta}n_{;\beta}=0, \quad h^{\alpha\beta}T_{;\beta}=0. \tag{3.2b}$$

Moreover, if $d_{\alpha\beta}$ is the background shear tensor $[d_{\alpha\beta}:=(\sigma_{\alpha\beta})_{\epsilon=0}]$ and H is the average rate of expansion $[H:=\frac{1}{3}(\theta)_{\epsilon=0}]$, it will be possible to verify that[9,22]

$$\dot{d}_{\alpha\beta}+3Hd_{\alpha\beta}=0. \tag{3.3}$$

Setting $d^\alpha{}_\beta:=d_\beta{}^\alpha:=q^{\alpha\mu}d_{\mu\beta}$ (and $d^{\alpha\beta}:=q^{\alpha\mu}d_\mu{}^\beta$), focus now on the eigenvalue equation,

$$d^\alpha{}_\mu l_p^\mu=(H_p-H)l_p^\alpha, \quad p=1,2,3, \tag{3.4}$$

with eigenvectors l_p^μ $(l_p^\mu w_\mu=0)$ and eigenvalues H_p-H. We shall refer to H_p as the expansion in the p direction. Evidently, since H_p is a scalar,[33] the spatial gradient of H_p vanishes:

$$H_{p|\alpha}=0, \quad p=1,2,3. \tag{3.5}$$

Note that[21,22]

$$k(2H_1-H_2-H_3)=0, \tag{3.6}$$

where k is the constant. By an appropriate choice of units, the value of k can be made to be 0 or -1. The corresponding solutions for the background metric $q_{\alpha\beta}$ represent, respectively, a Bianchi type-I space–time and a Bianchi type-V space–time. If the eigenframe $\{l_p^\mu\}$ of $d_{\alpha\beta}$ is known, the objects l^p_μ dual to l_p^μ are completely determined from $l^p_\mu l_q^\mu=\delta^q{}_p$ and $l^p_\mu w^\mu=0$. However, owing to the fact that the eigenvectors l^μ_p are not uniquely defined by Eq. (3.4), the additional question arises of how to normalize l^μ_p. This question can be answered most easily in a special coordinate system in which the line element of Bianchi models takes the form[21,22]

$$ds^2=-dt^2+(R_1(t))^2(dx^1)^2+e^{2kx^1}[(R_2(t))^2(dx^2)^2+(R_3(t))^2(dx^3)^2], \tag{3.7}$$

where R_p, $p=1,2,3$, are the expansion factors related to H_p, $p=1,2,3$, by

$$H_p=\frac{1}{R_p}\frac{dR_p}{dt}=\dot{R}_p/R_p, \tag{3.8}$$

and where $t:=x^0$ and k is the (constant) spatial curvature $(k=0,-1)$. In such a coordinate system, both $d^\alpha{}_0$ and $d^0{}_\alpha$ vanish, $d^p{}_q$ equals $(H_p-H)\delta^p{}_q$ (where $p=1,2,3$ and $q=1,2,3$), w^α is given by $w^\alpha=\delta^\alpha{}_0$, and we postulate that $l^\mu_p=\delta^\mu{}_p$ (hence $l^p_\mu=\delta^p{}_\mu$).

Writing $B:=(\frac{1}{6})(2H_1-H_2-H_3)$ and $C:=(1/2\sqrt{3})(H_2-H_3)$ and remembering that $H:=(\frac{1}{3})\times(\theta)_{\epsilon=0}$, one finds the expansions H_p, $p=1,2,3$, to be

$$H_1=H+2B, \tag{3.9a}$$

$$H_2=H-B+\sqrt{3}C, \quad H_3=H-B-\sqrt{3}C. \tag{3.9b}$$

Now, let $\mathcal{R}_{\alpha\beta}$ be the Ricci tensor of the connection defined by the metric $q_{\alpha\beta}$, and suppose that $\mathcal{R}$ is the curvature scalar $(\mathcal{R}:=q^{\mu\nu}\mathcal{R}_{\mu\nu})$. Since the extrinsic curvature $k_{\alpha\beta}:=w_{\alpha|\beta}$ has the properties $k_{(\alpha\beta)}=k_{\alpha\beta}$ and $k_{[\alpha\beta]}=0$, there exists a family of three-surfaces $\Sigma_\perp$ everywhere orthogonal to the unperturbed fluid flow vector w^α; these are instantaneous surfaces of simultaneity for all the fundamental observers.[2] Motivated by the above observations, it proves useful to introduce the following quantity:

$$\frac{6k}{(R_1)^2}:=-6H^2+d^{\mu\nu}d_{\mu\nu}+\mathcal{R}+2w^\mu w^\nu\mathcal{R}_{\mu\nu}=-6H^2+6(B^2+C^2)+\mathcal{R}+2w^\mu w^\nu\mathcal{R}_{\mu\nu}. \tag{3.10}$$

This quantity acquires a special significance:[2] it is the Ricci scalar of the three-dimensional surfaces $(k=0,-1)$.

With these preparations behind us, the coupled system of governing equations for the background may be written as[21,22]

$$H^2-B^2-C^2+\frac{k}{(R_1)^2}=\frac{1}{3}(e_0+\Lambda), \tag{3.11a}$$

$$2\dot{H}+3(H^2+B^2+C^2)+\frac{k}{(R_1)^2}=-p_0+\Lambda, \tag{3.11b}$$

$$\dot{B}+3HB=0, \quad \dot{C}+3HC=0, \tag{3.11c}$$

$$kB=0, \quad kH=k(\dot{R}_1/R_1), \tag{3.11d}$$

$$\dot{n}_0+3n_0H=0, \tag{3.11e}$$

where [see Eqs. (2.5) and (2.11)]

$$e_0=e(n_0,T_0), \quad p_0=p(n_0,T_0). \tag{3.12}$$

Equation (3.11e) is the standard equation of balance of n_0. In the case when $k=0$, the conditions (3.6) and (3.11d) are automatically satisfied, and thus they are not needed to determine the propagation of interesting quantities along the flow lines. Note that the result of differentiating the first of Eqs. (3.11) with respect to time and then using (3.11b)–(3.11d) is the equation of the conservation of energy:

$$\dot{e}_0+3H(e_0+p_0)=0. \tag{3.13}$$

Another remark is also in order. Instead of considering Eq. (3.3), we may equivalently consider the scalar equations (3.11c).

This remark brings to an end our description of the background models.

IV. ANALYSIS OF THE GAUGE PROBLEM

A. Equivalence classes of perturbations (tangents)

In Einstein's theory of gravity, the action of the diffeomorphism group on the space–time manifold X induces an action on the space of field configurations on X, and the only thing that has an immediate physical meaning is the quotient space of orbits,[34] i.e., two field configurations are regarded as equivalent if they are connected by a diffeomorphism transformation. Within the framework of a linear approximation, these facts imply that two infinitesimal perturbations $\delta\mathcal{F}_0$ and $\delta\mathcal{F}_0'$, *not necessarily satisfying Eq. (2.14)*, characterize the same perturbation of the background solution $\mathcal{F}_0$ if (and only if) there exists a vector field $v=v^\mu(\partial/\partial x^\mu)$ on the space–time manifold X, such that $\delta\mathcal{F}_0'-\delta\mathcal{F}_0$ is the *Lie derivative* $\mathcal{L}_v\mathcal{F}_0$ of $\mathcal{F}_0$ with respect to v.[4,5] As to the explicit specification of $\mathcal{L}_v\mathcal{F}_0$, putting $w:=w^\mu(\partial/\partial x^\mu)$ and $q:=q^{\mu\nu}[(\partial/\partial x^\mu)\otimes(\partial/\partial x^\nu)]$ and recalling the definition (2.10) of $\mathcal{F}_0$, we obtain[35]

$$\mathcal{L}_v\mathcal{F}_0:=\{(\mathcal{L}_vq)^{\alpha\beta},(\mathcal{L}_vw)^\alpha,\mathcal{L}_vn_0,\mathcal{L}_vT_0\}, \tag{4.1}$$

where $(\mathcal{L}_vq)^{\alpha\beta}$ and $(\mathcal{L}_vw)^\alpha$ are the components of $\mathcal{L}_vq$ and $\mathcal{L}_vw$.

Let $\mathcal{P}$ be the set whose elements are arbitrary, linear perturbations of $\mathcal{F}_0$; thus $\delta\mathcal{F}_0\in\mathcal{P}$. Obviously, $\mathcal{P}$ carries a natural structure of a vector space. The set $\mathcal{P}_L$ consisting of $\mathcal{L}_v\mathcal{F}_0$ for all vector fields v on X may be considered as the subspace of $\mathcal{P}$. Given the object $J(\delta\mathcal{F}_0)$ as in Eq. (2.18), we denote by $\mathcal{W}$ the collection of $J(\delta\mathcal{F}_0)$, where $\delta\mathcal{F}_0\in\mathcal{P}$, and by $\mathbf{W},\mathbf{W}'$, and similar symbols the elements of $\mathcal{W}$. Moreover, we introduce the following subspace of $\mathcal{W}$:

$$\mathcal{W}_L:=\{J(\delta\mathcal{F}_0);\delta\mathcal{F}_0\in\mathcal{P}_L\}. \tag{4.2}$$

Knowing $\mathcal{F}_0$, elementary inspection shows that the correspondence $\delta\mathcal{F}_0\mapsto J(\delta\mathcal{F}_0)$ is a linear map that assigns to each $\delta\mathcal{F}_0\in\mathcal{P}$ an element $J(\delta\mathcal{F}_0)\in\mathcal{W}$; we denote this map by $J:\mathcal{P}\to\mathcal{W}$. Since for every $\mathbf{W}\in\mathcal{W}$ there is just one $\delta\mathcal{F}_0\in\mathcal{P}$ such that $\mathbf{W}=J(\delta\mathcal{F}_0)$, the map $J:\mathcal{P}\to\mathcal{W}$ can be said to be *one-to-one* and *onto*. The similar remark concerns the restriction of J to $\mathcal{P}_L$, still denoted by J. As suggested by the definition (4.2), the subspace $\mathcal{W}_L$ of $\mathcal{W}$ is the image of $\mathcal{P}_L$ under J. Clearly, $\mathbf{W}$ belongs to $\mathcal{W}_L$ if and only if $\mathbf{W}$ equals $J(\mathcal{L}_v\mathcal{F}_0)$ for some v. *In order to simplify our notation, we abbreviate* $J(\mathcal{L}_v\mathcal{F}_0)$ *as* $L_v\mathcal{F}_0$.

The next stage in the discussion is to calculate $L_v\mathcal{F}_0$. First of all, projecting the vector field v^α on X along and orthogonal to the background fluid four-velocity w^α, we obtain the decomposition of v^α, namely,

$$v^\alpha=\vartheta w^\alpha+\vartheta^\alpha, \tag{4.3}$$

in which the scalar ϑ and the spatial vector ϑ^α are given by

$$\vartheta:=-w_\mu v^\mu,\quad \vartheta^\alpha:=\gamma^\alpha{}_\mu v^\mu. \tag{4.4}$$

With the help of the definition of $\mathcal{L}_v$,[35] we then find from Eqs. (2.17) and (2.18) that $L_v\mathcal{F}_0$ can be written as

$$L_v\mathcal{F}_0=\{Q_v,Q_v^\alpha,D_v,F_v^{\alpha\beta},V_v,V_v^\alpha,M_v,K_v\}, \tag{4.5}$$

where

$$Q_v:=2\dot{\vartheta},\quad Q_v^\alpha:=\dot{\vartheta}^\alpha-H\vartheta^\alpha-d^\alpha{}_\mu\vartheta^\mu-\gamma^{\alpha\mu}\vartheta_{|\mu}, \tag{4.6a}$$

$$D_v:=-H\vartheta-\tfrac{1}{3}\vartheta^\mu{}_{|\mu}, \tag{4.6b}$$

$$F_v^{\alpha\beta}:=-d^{\alpha\beta}\vartheta+\tfrac{1}{3}\gamma^{\alpha\beta}\vartheta^\mu{}_{|\mu}-\gamma^{\mu(\alpha}\vartheta^{\beta)}{}_{|\mu}, \tag{4.6c}$$

$$V_v:=-\dot{\vartheta},\quad V_v^\alpha:=-\dot{\vartheta}^\alpha+H\vartheta^\alpha+d^\alpha{}_\mu\vartheta^\mu, \tag{4.6d}$$

$$M_v:=-3H\vartheta,\quad K_v:=\frac{1}{T_0}\dot{T}_0\vartheta. \tag{4.6e}$$

The notation and conventions are based on those of Secs. II and III. In obtaining Eqs. (4.6), we have used the fact that the form of $\mathcal{F}_0$ is consistent with exact anisotropic cosmological solutions of Einstein's field equations, namely spatially homogeneous cosmologies of Bianchi types I and V.

The net upshot of these considerations may be stated very neatly. In linear perturbation theory, the object of most physical interest is not just one perturbation $\delta\mathcal{F}_0\in\mathcal{P}$, but a whole equivalence class of all perturbations $\delta\mathcal{F}'_0\in\mathcal{P}$ that are equivalent to $\delta\mathcal{F}_0$: two infinitesimal perturbations $\delta\mathcal{F}_0\in\mathcal{P}$ and $\delta\mathcal{F}'_0\in\mathcal{P}$ are said to be equivalent if $\delta\mathcal{F}'_0-\delta\mathcal{F}_0$ equals $\mathcal{L}_v\mathcal{F}_0$ for some v. The equivalence class of $\delta\mathcal{F}_0\in\mathcal{P}$ is denoted $[\delta\mathcal{F}_0]$ and is called the *gauge-invariant perturbation* associated with $\delta\mathcal{F}_0$. In this way, we verify that the gauge-invariant perturbations are elements of $\mathcal{P}/\mathcal{P}_L$, the *quotient space* of $\mathcal{P}$ by $\mathcal{P}_L$. Another route to discussing the gauge problem is to introduce the equivalence class $[\mathbf{W}]$ of $\mathbf{W}\in\mathcal{W}$: two infinitesimal perturbations $\mathbf{W}\in\mathcal{W}$ and $\mathbf{W}'\in\mathcal{W}$, *not necessarily satisfying Eq. (2.19)*, will be taken to be equivalent if there is a vector field $v^\alpha=\vartheta w^\alpha+\vartheta^\alpha$ on X such that $\mathbf{W}'=\mathbf{W}+L_v\mathcal{F}_0$. Then we have the *gauge-invariant perturbation* $[\mathbf{W}]$ associated with $\mathbf{W}$ and the *quotient space* $\mathcal{W}/\mathcal{W}_L$, which consists of $[\mathbf{W}]$ for all $\mathbf{W}\in\mathcal{W}$.

The essential point in the theory of gauge-invariant perturbations is to describe the elements of $\mathcal{P}/\mathcal{P}_L$ or $\mathcal{W}/\mathcal{W}_L$ explicitly. These issues will be considered in Secs. IV B and IV C.

B. Basic gauge-invariant variables

In an almost-Bianchi type-I or almost-Bianchi type-V universe model, we have "absolute" or "prior geometric" elements (e.g., the background solution $\mathcal{F}_0$, the average rate of expansion H, the background shear tensor $d_{\alpha\beta}$, the eigenframe $\{l_p^\mu\}$ of $d_{\alpha\beta}$, etc.), and we can use these absolute elements and the equivalence classes of perturbations to introduce *local* gauge-invariant variables. By analogy with the general discussion in Sec. 4.2 of Ref. 10, the gauge-invariant quantity $\overline{A}(x,\mathbf{W})$ is such a linear and local algebropartial differential operation on $\mathbf{W}$ that, for each space–time point x and each vector field v on X, the "gauge mode" perturbation $L_v\mathcal{F}_0$ of $\mathcal{F}_0$ can be added to $\mathbf{W}$ without the need of replacing $\overline{A}(x,\mathbf{W})$ by $\overline{A}(x,\mathbf{W}+L_v\mathcal{F}_0)$:

$$\overline{A}(x,\mathbf{W})=\overline{A}(x,\mathbf{W}+L_v\mathcal{F}_0). \tag{4.7}$$

Because of this, the value of $\overline{A}(\cdot,\mathbf{W}')$ at $x\in X$ is independent of the choice of $\mathbf{W}'\in[\mathbf{W}]$ and the objects $A(x,[\mathbf{W}]):=\overline{A}(x,\mathbf{W})$ define a "function" $[\mathbf{W}]\mapsto A(\cdot,[\mathbf{W}])$ on the quotient space $\mathcal{W}/\mathcal{W}_L$. Here the most important examples of $x\mapsto A(x,[\mathbf{W}])$, which actually work and which provide the mathematically simplest representation of $[\mathbf{W}]$, are given by

$$\chi := Q + 2V, \quad \Gamma := K + \frac{1}{3HT_0}\dot{T}_0 M, \tag{4.8a}$$

$$\Omega := -\frac{1}{2}Q + \frac{1}{3H^2}(\dot{H}M - H\dot{M}), \tag{4.8b}$$

$$\Omega^\alpha := -3H(V^\alpha + Q^\alpha) + \gamma^{\alpha\mu}M_{|\mu}, \tag{4.8c}$$

$$\Omega^{\alpha\beta} := \ddot{Z}^{\alpha\beta} + 3H\dot{Z}^{\alpha\beta} + 2d_\mu{}^\alpha d^\beta{}_\nu Z^{\mu\nu} - 2d_\mu{}^{(\alpha} Z^{\beta)\nu} d^\mu{}_\nu + 2[\gamma^{\mu(\alpha}\dot{V}^{\beta)}]_{|\mu}$$
$$+ 4H\gamma^{\mu(\alpha}V^{\beta)}{}_{|\mu} - 4d^{\mu(\alpha}V^{\beta)}{}_{|\mu} + 2d_\mu{}^{(\alpha}\gamma^{\beta)\nu}V^\mu{}_{|\nu} + 2(2d^{\alpha\beta}{}_{|\mu} - \gamma^{\nu(\alpha}d^{\beta)}{}_{\mu|\nu})V^\mu, \tag{4.8d}$$

$$S^{\alpha\beta\mu\nu} := \frac{k}{(R_1)^2} Z^{\sigma[\alpha}\gamma^{\beta][\mu}\gamma^{\nu]}{}_\sigma + [(\gamma^{\sigma[\alpha}Z^{\beta][\mu}\gamma^{\nu]\lambda})_{|\sigma}]_{|\lambda}, \tag{4.8e}$$

$$\Theta^p := l^p_\mu l^p_\nu (\dot{Z}^{\mu\nu} + 2\gamma^{\lambda\mu}V^\nu{}_{|\lambda}), \tag{4.8f}$$

$$\Theta^{pq}_q := -l^\nu_q (l^q_\lambda l^p_{\mu|\nu} + l^p_\lambda l^q_{\mu|\nu})(\dot{Z}^{\mu\lambda} + 2\gamma^{\tau(\mu}V^{\lambda)}{}_{|\tau}) + l^\nu_q l^q_\lambda l^p_\mu [(\dot{Z}^{\lambda\mu} + 2\gamma^{\tau(\mu}V^{\lambda)}{}_{|\tau})_{|\nu}$$
$$-2d_\tau{}^{[\mu}Z^{\lambda]\tau}{}_{|\nu} + (d^{\mu\tau}\gamma_{\nu\sigma} - \gamma^{\mu\tau}d_{\nu\sigma})Z^{\sigma\lambda}{}_{|\tau} - d^\mu{}_{\nu|\tau}Z^{\tau\lambda}] \quad (p \neq q), \tag{4.8g}$$

$$\Omega^{pq}_r := l^\nu_r l^p_{(\lambda} l^q_{\mu)} [(\dot{Z}^{\lambda\mu} + 2\gamma^{\tau(\lambda}V^{\mu)}{}_{|\tau})_{|\nu} + 2(\gamma^\lambda{}_\sigma d^{\mu\tau} - d^\lambda{}_\sigma \gamma^{\mu\tau})\gamma_{\nu\kappa}Z^{\sigma\kappa}{}_{|\tau} + 2d^\lambda{}_{\nu|\sigma}Z^{\mu\sigma}] + l^\nu_r (l^p_\mu l^q_{\nu|\lambda}$$
$$+ l^q_\mu l^p_{\nu|\lambda})(\dot{Z}^{\lambda\mu} + 2\gamma^{\tau(\lambda}V^{\mu)}{}_{|\tau}) + l^\nu_{r|\nu} l^p_\lambda l^q_\mu (\dot{Z}^{\lambda\mu} + 2\gamma^{\tau(\lambda}V^{\mu)}{}_{|\tau}) \quad (p \neq q, r \neq p, r \neq q), \tag{4.8h}$$

where

$$Z^{\alpha\beta} := \frac{2}{3}(M - 3D)\gamma^{\alpha\beta} + \frac{2}{3H}Md^{\alpha\beta} - 2F^{\alpha\beta}. \tag{4.9}$$

In keeping with the notation of Sec. III, the quantity $\ddot{Z}^{\alpha\beta}$ is the covariant derivative of $\dot{Z}^{\alpha\beta}$ along w^μ and the tensor fields $l^\mu_{p|\nu}$ and $l^p_{\mu|\nu}$ are the spatially totally projected covariant derivatives of l^μ_p and l^p_μ. Also, it appears from Eqs. (4.8f) and (4.8g) that *we do not adopt* the Einstein summation convention for repeated Latin indices.

According to Eqs. (4.8g) and (4.8h), we have the condition $p \neq q$ for Θ^{pq}_q and the conditions $p \neq q$, $r \neq p$, and $r \neq q$ for Ω^{pq}_r; moreover, $\Omega^{pq}_r = \Omega^{qp}_r$. It is clear, from the manner in which the objects Θ^p, Θ^{pq}_q, and Ω^{pq}_r are defined [see Eqs. (4.8f)–(4.8h)], that the sets $\{\Theta^p\}$, $\{\Theta^{pq}_q\}$, and $\{\Omega^{pq}_r\}$ consist of *scalar* gauge-invariant variables; these sets may be written as

$$\{\Theta^p\} := \{\Theta^1, \Theta^2, \Theta^3\}, \tag{4.10a}$$

$$\{\Theta^{pq}_q\} := \{\Theta^{12}_2, \Theta^{13}_3, \Theta^{21}_1, \Theta^{23}_3, \Theta^{31}_1, \Theta^{32}_2\}, \tag{4.10b}$$

$$\{\Omega^{pq}_r\} := \{\Omega^{12}_3, \Omega^{13}_2, \Omega^{23}_1\}. \tag{4.10c}$$

Now, if we make use of Eq. (4.8e), we verify that a complete set of symmetry conditions for $S^{\alpha\beta\mu\nu}$ is $S^{\alpha\beta\mu\nu} = S^{[\alpha\beta][\mu\nu]}$, $S^{\alpha[\beta\mu\nu]} = 0$, and $w_\alpha S^{\alpha\beta\mu\nu} = 0$; thus, there are six linearly independent, not identically vanishing, components in $\{S^{\alpha\beta\mu\nu}\}$. An alternative viewpoint is to express $S^{\alpha\beta\mu\nu}$ in terms of

$$S := \gamma_{\mu\nu}\gamma_{\lambda\sigma}S^{\mu\lambda\sigma\nu} \tag{4.11a}$$

and

$$S^{\alpha\beta} := \gamma_{\mu\nu}(S^{\mu\alpha\beta\nu} - \tfrac{1}{3}\gamma^{\alpha\beta}\gamma_{\lambda\sigma}S^{\mu\lambda\sigma\nu}). \tag{4.11b}$$

The symmetries of $S^{\alpha\beta\mu\nu}$ imply

$$S^{\alpha\beta} = S^{\beta\alpha}, \quad \gamma_{\mu\nu}S^{\mu\nu} = 0, \quad w_\mu S^{\mu\alpha} = 0. \tag{4.12}$$

Through Eqs. (4.11a) and (4.11b), we can completely describe the tensor field $S^{\alpha\beta\mu\nu}$, since this tensor field is a *totally* spatial quantity:

$$S^{\alpha\beta\mu\nu} = -\tfrac{1}{3}S\,\gamma^{\mu[\alpha}\gamma^{\beta]\nu} - 2\,\gamma^{\mu[\alpha}S^{\beta]\nu} + 2\,\gamma^{\nu[\alpha}S^{\beta]\mu}. \tag{4.13}$$

Next, returning to the definitions (4.8c) and (4.8d) of Ω^α and $\Omega^{\alpha\beta}$, and using Eq. (4.8f) for Θ^p, we easily recognize that

$$w_\mu \Omega^\mu = 0, \quad w_\mu \Omega^{\mu\alpha} = 0, \quad \Omega^{\alpha\beta} = \Omega^{\beta\alpha}, \tag{4.14a}$$

$$l^p_\mu l^p_\nu \Omega^{\mu\nu} = \dot{\Theta}^p + (3H + 2H_p)\Theta^p \quad (p \text{ not summed}). \tag{4.14b}$$

Consequently, the formulas (4.8c) really provide only three independent equations for the determination of $\{\Omega^\alpha\}$, and the set $\{\Omega^{\alpha\beta}\}$ is uniquely specified by giving $\{\Theta^p\}$ and $\{l^p_\mu l^q_\nu \Omega^{\mu\nu}; p < q\}$. As regards the definition (4.8a) of χ, an equation for χ can be obtained from the relation $\chi = -[\partial(u^\alpha u_\alpha)/\partial\epsilon]_{\epsilon=0}$ with the result $\chi = Q + 2V = 0$, which is a direct consequence of $u^\mu u_\mu = -1$. Thus the gauge-invariant quantity χ will not be significant to us in considering linearization about the universe models of Bianchi types I and V.

Let us summarize the situation. For each $[\mathbf{W}]$, consider the set

$$\varphi([\mathbf{W}]) := \{\chi, \Gamma, \Omega, S, \Omega^\alpha, \Omega^{\alpha\beta}, S^{\alpha\beta}, \Theta^p, \Theta^{pq}_q, \Omega^{pq}_r\}. \tag{4.15}$$

From the discussion presented above it clearly follows that this set consists of gauge-invariant variables. Besides, the total number of independent, not identically vanishing, ''scalar functions'' in $\varphi([\mathbf{W}])$ is 26 $[f_n : X \to \mathbb{R}\ (n = 1,2,...,26)]$. For reasons to become clear later,[20] we shall refer to $\varphi([\mathbf{W}])$ as the *complete* set of *basic* gauge-invariant variables associated with $[\mathbf{W}]$. In order to obtain $\varphi([\mathbf{W}])$, we have used one representative member of $[\mathbf{W}]$, namely, the infinitesimal perturbation $\mathbf{W} = J(\delta\mathcal{F}_0)$ of $\mathcal{F}_0$ given by Eq. (2.18). However, the value of $\varphi([\mathbf{W}])$ does not depend on the choice of $\mathbf{W}' \in [\mathbf{W}]$ and the objects $\varphi([\mathbf{W}]), \mathbf{W} \in \mathcal{W}$, indeed define a function on the quotient space $\mathcal{W}/\mathcal{W}_L$:

$$[\mathbf{W}] \mapsto \varphi([\mathbf{W}]). \tag{4.16}$$

The key steps on which the derivation of this result rests are, first, the observation that the individual quantities appearing on the right-hand sides of Eqs. (4.8) are obtained by means of *linear*, algebropartial differential operations on $\mathbf{W}$, and second, the fact that all these quantities *vanish* if $\mathbf{W}$ is a trivial, ''gauge mode'' perturbation of $\mathcal{F}_0$, i.e., if $\mathbf{W}$ equals $L_v\mathcal{F}_0 := J(\mathcal{L}_v\mathcal{F}_0)$ for some v. Under the foregoing statements, our construction of $\varphi([\mathbf{W}])$ is valid in the class of spatially homogeneous, nonaxisymmetric cosmologies of Bianchi types I and V ($H_1 \neq H_2$, $H_1 \neq H_3$, $H_2 \neq H_3$) with a nonbarotropic perfect fluid as the source. Its usefulness will appear indisputably in Sec. IV C, where we prove that Eqs. (4.8)–(4.15) support an interpretation of $\mathbf{W} \mapsto \varphi([\mathbf{W}])$ as a ''coordinate system'' on $\mathcal{W}/\mathcal{W}_L$.

We now give a necessary word concerning the special cases. When $d_{\alpha\beta} = 0$, we reduce immediately to the situation in almost-Robertson–Walker universe models.[2] Surprisingly enough, the definition of $\varphi([\mathbf{W}])$ for this situation (see Refs. 18 and 19) differs considerably from that presented here. A similar remark holds for the case when the metric $q_{\alpha\beta}$ is axially symmetric ($R_1 \neq R_2 = R_3$); this case must be treated separately as well. We finally mention the following. If we begin from the *general* Bianchi type-I background model,[9] some of the formulas (4.8) simplify enormously; we can then set

$$k=0, \quad l^{\mu}_{p|\nu}=0, \quad l^{p}_{\mu|\nu}=0, \tag{4.17a}$$

$$d^{\alpha\beta}{}_{|\mu}=0, \quad d^{\alpha}{}_{\beta|\mu}=d_{\beta}{}^{\alpha}{}_{|\mu}=0, \quad d_{\alpha\beta|\mu}=0. \tag{4.17b}$$

Moreover, with this choice of the model, we can introduce

$$\Omega^{\alpha\beta}_{\mu}:=(\dot{Z}^{\alpha\beta}+2\gamma^{\nu(\alpha}V^{\beta)}{}_{|\nu})_{|\mu}+2(\gamma_{\sigma}{}^{(\alpha}d^{\beta)\nu}-d_{\sigma}{}^{(\alpha}\gamma^{\beta)\nu})\gamma_{\mu\lambda}Z^{\sigma\lambda}{}_{|\nu}, \tag{4.18}$$

in place of Eqs. (4.8g) and (4.8h). For $k=0$, the essential five properties of the quantity $\Omega^{\alpha\beta}_{\mu}$ are that it (a) is gauge invariant, (b) contains the same information as Θ^{pq}_{q} and Ω^{pq}_{r}, (c) is defined without making any explicit or implicit reference to the eigenframe $\{l^{\mu}_{p}\}$ of $d_{\alpha\beta}$, (d) has the symmetry properties $\Omega^{\alpha\beta}_{\mu}=\Omega^{\beta\alpha}_{\mu}$ and $w^{\mu}\Omega^{\alpha\beta}_{\mu}=w_{\mu}\Omega^{\mu\alpha}_{\nu}q^{\nu\beta}=0$, and (e) satisfies the condition of the form

$$l^{p}_{\alpha}l^{p}_{\beta}l^{\mu}_{q}\Omega^{\alpha\beta}_{\mu}=l^{\mu}_{q}\Theta^{p}{}_{|\mu} \quad (p \;\text{ not summed}). \tag{4.19}$$

Despite these conclusions, our motive for not preferring to replace Θ^{pq}_{q} and Ω^{pq}_{r} by $\Omega^{\alpha\beta}_{\mu}$ when $k=0$ is that we have not found a way to do so generally; if $k=-1$, the construction of $\varphi([\mathbf{W}])$ based on the shear eigenframe appears to us to be canonical.

C. Completeness of the set $\varphi([\mathbf{W}])$ of basic gauge-invariant variables

We denote by $\mathcal{D}$ the set consisting of $\varphi([\mathbf{W}])$ for all $[\mathbf{W}]\in\mathcal{W}/\mathcal{W}_L$ and by $\mathbf{D}$, $\mathbf{D}'$, and similar symbols the elements of $\mathcal{D}$. A function φ from $\mathcal{W}/\mathcal{W}_L$ onto $\mathcal{D}$, defined by Eqs. (4.8)–(4.15), is a linear map that assigns to each $[\mathbf{W}]\in\mathcal{W}/\mathcal{W}_L$ an element $\varphi([\mathbf{W}])\in\mathcal{D}$; thus, $\mathcal{D}$ carries a canonical structure of a vector space induced by that of $\mathcal{W}/\mathcal{W}_L$. More precisely, $\mathcal{D}$ is a function space in which the usual operations of addition and scalar multiplication are introduced. A necessary and sufficient condition that φ be a one-to-one mapping of $\mathcal{W}/\mathcal{W}_L$ onto $\mathcal{D}$ is that $\varphi([\mathbf{W}])$ equals a zero-vector of $\mathcal{D}$ if and only if $[\mathbf{W}]$ equals a zero-vector of the quotient space $\mathcal{W}/\mathcal{W}_L$, i.e., if and only if $[\mathbf{W}]$ can be identified with $[L_v\mathcal{F}_0]$, where v is an arbitrary vector field on the space–time manifold X. For essentially obvious reasons, it will be convenient to call this condition a *natural condition* for the existence of a ''coordinate system'' on $\mathcal{W}/\mathcal{W}_L$. Now, after these preparations, we are in a position to formulate the following theorem.

Theorem: Let $\mathcal{F}_0$ be a nonaxisymmetric Bianchi type-I or Bianchi type-V solution of the nonlinear field equations (2.1)–(2.6); in other words, suppose that this solution is such that $H_1\neq H_2$, $H_1\neq H_3$, and $H_2\neq H_3$. Then the linear mapping $\varphi:\mathcal{W}/\mathcal{W}_L\rightarrow\mathcal{D}$ defined as in Sec. IV B satisfies a *natural condition* for the existence of a ''coordinate system'' on $\mathcal{W}/\mathcal{W}_L$. Thus, for each $\mathbf{D}\in\mathcal{D}$ there is just one $[\mathbf{W}]\in\mathcal{W}/\mathcal{W}_L$ such that $\mathbf{D}=\varphi([\mathbf{W}])$, and the mapping $\varphi:\mathcal{W}/\mathcal{W}_L\rightarrow\mathcal{D}$ is one-to-one and onto. Put somewhat differently, one can introduce the inverse of φ, namely, $\varphi^{-1}:\mathcal{D}\rightarrow\mathcal{W}/\mathcal{W}_L$, by setting $\varphi^{-1}(\varphi([\mathbf{W}]))=[\mathbf{W}]$.

Sketch of the proof: A straightforward but tedious application of Eqs. (4.5)–(4.15) yields the identity $\varphi([L_v\mathcal{F}_0])=0$, so the proof of the theorem reduces to showing that if $[\mathbf{W}]\in\mathcal{W}/\mathcal{W}_L$ belongs to the *kernel* of φ, denoted by ker φ, then the representative member of $[\mathbf{W}]$, namely $\mathbf{W}$, is an element of $\mathcal{W}_L$. Thus, assume that $\mathbf{W}\in[\mathbf{W}]$, the infinitesimal perturbation of $\mathcal{F}_0$ has the property that $\varphi([\mathbf{W}])=0$, i.e., satisfies the conditions of the form

$$\chi=0, \quad \Gamma=0, \quad \Omega=0, \quad S=0, \tag{4.20a}$$

$$\Omega^{\alpha}=0, \quad \Omega^{\alpha\beta}=0, \quad S^{\alpha\beta}=0, \tag{4.20b}$$

$$\Theta^{p}=0, \quad \Theta^{pq}_{q}=0, \quad \Omega^{pq}_{r}=0, \tag{4.20c}$$

where the objects χ through Ω^{pq}_{r} are given by Eqs. (4.8) and (4.11). Under the foregoing conditions to guide us, we wish to demonstrate that there exists a vector field $v=v^{\alpha}(\partial/\partial x^{\alpha})$ on the space–time manifold X such that

$$\mathbf{W}:=\{Q,Q^{\alpha},D,F^{\alpha\beta},V,V^{\alpha},M,K\},$$

equals $L_v\mathcal{F}_0:=\{Q_v,Q^{\alpha}_v,D_v,F^{\alpha\beta}_v,V_v,V^{\alpha}_v,M_v,K_v\}$, i.e.,

$$Q=Q_v:=2\dot{\vartheta}, \tag{4.21a}$$

$$Q^{\alpha}=Q^{\alpha}_v:=\dot{\vartheta}^{\alpha}-H\vartheta^{\alpha}-d^{\alpha}{}_{\mu}\vartheta^{\mu}-\gamma^{\alpha\mu}\vartheta_{|\mu}, \tag{4.21b}$$

$$D=D_v:=-H\vartheta-\tfrac{1}{3}\vartheta^{\mu}{}_{|\mu}, \tag{4.21c}$$

$$F^{\alpha\beta}=F^{\alpha\beta}_v:=-d^{\alpha\beta}\vartheta+\tfrac{1}{3}\gamma^{\alpha\beta}\vartheta^{\mu}{}_{|\mu}-\gamma^{\mu(\alpha}\vartheta^{\beta)}{}_{|\mu}, \tag{4.21d}$$

$$V=V_v:=-\dot{\vartheta}, \tag{4.21e}$$

$$V^{\alpha}=V^{\alpha}_v:=-\dot{\vartheta}^{\alpha}+H\vartheta^{\alpha}+d^{\alpha}{}_{\mu}\vartheta^{\mu}, \tag{4.21f}$$

$$M=M_v:=-3H\vartheta, \tag{4.21g}$$

$$K=K_v:=\frac{1}{T_0}\dot{T}_0\vartheta. \tag{4.21h}$$

For our purposes here, we have used the definitions (4.6) and the decomposition $v^{\alpha}=\vartheta w^{\alpha}+\vartheta^{\alpha}$ of v^{α} with respect to the background fluid four-velocity w^{α}.

(a) To start with, let

$$\vartheta:=-\frac{1}{3H}M \tag{4.22}$$

be a candidate for the timelike part of v^{α}, and define the spacelike part ϑ^{α} of v^{α} by saying that it satisfies the differential equations[36]

$$\dot{\vartheta}^{\alpha}-H\vartheta^{\alpha}-d^{\alpha}{}_{\mu}\vartheta^{\mu}=-V^{\alpha}\quad(\alpha=0,\ldots,3). \tag{4.23}$$

The definitions (4.22) and (4.23) trivially prove Eq. (4.21g) for M and Eq. (4.21f) for V^{α}. Because of $\Gamma=0$ and $\Omega=0$, K equals $T_0^{-1}\dot{T}_0\vartheta$, Q equals $2\dot{\vartheta}$, and Eqs. (4.21h) and (4.21a) hold. Then we may conclude from $\chi=0$ that Eq. (4.21e) is valid for V as well. By using the definition (4.8c) of Ω^{α} and substituting the formulas (4.21f) and (4.21g) onto the left-hand side of $\Omega^{\alpha}=0$, we immediately derive Eq. (4.21b) for Q^{α}.

(b) At this stage of the proof, since the objects Q, Q^{α}, V, V^{α}, M, and K may be written in the desired form, it remains to show that the conditions $S=0, \Omega^{\alpha\beta}=0, S^{\alpha\beta}=0, \Theta^{p}=0, \Theta^{pq}_{q}=0$, and $\Omega^{pq}_{r}=0$ yield Eqs. (4.21c) and (4.21d) for D and $F^{\alpha\beta}$. The key step in the derivation of these two equations is simply this. According to the considerations in Sec. IV B, an equivalent statement to $S=0$ and $S^{\alpha\beta}=0$ is $S^{\alpha\beta\mu\nu}=0$. Given the definitions (4.8e) and (4.9) of $S^{\alpha\beta\mu\nu}$ and $Z^{\alpha\beta}$, we recognize $S^{\alpha\beta\mu\nu}=0$ as the condition for

$$Z^{\alpha\beta}:=\frac{2}{3}(M-3D)\gamma^{\alpha\beta}+\frac{2}{3H}Md^{\alpha\beta}-2F^{\alpha\beta}=-2(H\vartheta+D)\gamma^{\alpha\beta}-2\vartheta d^{\alpha\beta}-2F^{\alpha\beta}. \tag{4.24}$$

This condition implies that there exists a *spacelike* vector field z^{α} on X $(w_{\alpha}z^{\alpha}=0)$, *not necessarily equal to* ϑ^{α}, such that

$$Z^{\alpha\beta}=2\gamma^{\mu(\alpha}z^{\beta)}{}_{|\mu}. \tag{4.25}$$

For further details, and for the more general situations and conditions that also lead to the relation like Eq. (4.25), see the discussion on pp. 351–353 in Ref. 37. Now, in order to emphasize that z^α may differ from ϑ^α, we introduce the following vector field on X:

$$c^\alpha := z^\alpha - \vartheta^\alpha. \tag{4.26}$$

Using the obvious identities $\gamma_{\mu\nu} d^{\mu\nu} = 0$ and $\gamma_{\mu\nu} F^{\mu\nu} = 0$ and exploiting Eqs. (4.24)–(4.26), we then obtain

$$D = -H\vartheta - \tfrac{1}{3}\vartheta^\mu{}_{|\mu} - \tfrac{1}{3}(\gamma_{\mu\nu}\Delta^{\mu\nu}) \tag{4.27a}$$

and

$$F^{\alpha\beta} = -d^{\alpha\beta}\vartheta + \tfrac{1}{3}\gamma^{\alpha\beta}\vartheta^\mu{}_{|\mu} - \gamma^{\mu(\alpha}\vartheta^{\beta)}{}_{|\mu} + \tfrac{1}{3}\gamma^{\alpha\beta}(\gamma_{\mu\nu}\Delta^{\mu\nu}) - \Delta^{\alpha\beta}, \tag{4.27b}$$

where $\Delta^{\alpha\beta}$ is the spatial symmetric two-tensor related to c^α by

$$\Delta^{\alpha\beta} := \gamma^{\mu(\alpha} c^{\beta)}{}_{|\mu}. \tag{4.28}$$

If, without in any way restricting the generality of the conclusion, we demonstrate that $\Delta^{\alpha\beta}$ can be set equal to zero, Eqs. (4.21c) and (4.21d) will be derived, and this derivation will complete the proof of the theorem.

(c) To see how the identity $\Delta^{\alpha\beta} = 0$ can result from $\varphi([\mathbf{W}]) = 0$, by means of Eqs. (4.8d) and (4.8f)–(4.8h) we must carefully analyze the consequences of substituting $V^\alpha = -\dot{\vartheta}^\alpha + H\vartheta^\alpha + d^\alpha{}_\mu \vartheta^\mu$ and $Z^{\alpha\beta} = 2[\gamma^{\mu(\alpha}\vartheta^{\beta)}{}_{|\mu} + \Delta^{\alpha\beta}]$ onto the left-hand sides of $\Omega^{\alpha\beta} = 0$, $\Theta^p = 0$, $\Theta^{pq}_q = 0$, and $\Omega^{pq}_r = 0$; we are thereby led to a set of differential constraints for $\Delta^{\alpha\beta}$. However, before exploiting these constraints, some comments are in order. Equations (4.23) and (4.25) leave a residual ''gauge'' freedom in defining the spatial vector fields ϑ^α and z^α on X, i.e., ϑ^α and z^α are not uniquely determined from these equations. Thus, in particular, given any choice of ϑ^α consistent with Eq. (4.23), we can replace ϑ^α by $\vartheta^\alpha + a^\alpha$, where a^α is such that $w_\mu a^\mu = 0$ and $\dot{a}^\alpha - Ha^\alpha - d^\alpha{}_\mu a^\mu = 0$. Also, a significant feature follows immediately from Eq. (4.25). Let y^α be a vector field on X. Then, provided $w_\mu y^\mu = 0$ and $\gamma^{\mu(\alpha} y^{\beta)}{}_{|\mu} = 0$, the spacelike vector field z^α can be replaced by $z^\alpha + y^\alpha$.

(d) The gauge-invariant variables have as simple a form as possible if expressed in terms of a coordinate system (x^α) for which ds^2 is given by Eq. (3.7). Under this choice of coordinates, the differential conditions $\Omega^{\alpha\beta} = 0$, $\Theta^p = 0$, $\Theta^{pq}_q = 0$, and $\Omega^{pq}_r = 0$ can be solved for

$$\Delta^{\alpha\beta} := \gamma^{\mu(\alpha} z^{\beta)}{}_{|\mu} - \gamma^{\mu(\alpha}\vartheta^{\beta)}{}_{|\mu}, \tag{4.29}$$

by using Eqs. (3.6) and (3.11d), and then the tensor field,

$$\overline{\Delta}^{\alpha\beta} := \gamma^{\mu(\alpha}(z^{\beta)} + y^{\beta)})_{|\mu} - \gamma^{\mu(\alpha}(\vartheta^{\beta)} + a^{\beta)})_{|\mu} = \Delta^{\alpha\beta} + \gamma^{\mu(\alpha} y^{\beta)}{}_{|\mu} - \gamma^{\mu(\alpha} a^{\beta)}{}_{|\mu}, \tag{4.30}$$

can be set equal to zero by suitably constructing a^α and y^α:

$$\Delta^{\alpha\beta} \to \overline{\Delta}^{\alpha\beta} = 0. \tag{4.31}$$

Taken together with the previous results, this completes the proof of the theorem. ■

Remark: The detailed analysis of the differential constraints for $\Delta^{\alpha\beta}$, which is purely technical and requires the introduction of many auxiliary formulas, is available on request.

In virtue of the implication $([\mathbf{W}] \in \ker \varphi) \Rightarrow ([\mathbf{W}] = [L_v \mathcal{F}_0]$ for some $v)$, the mapping $[\mathbf{W}] \mapsto \varphi([\mathbf{W}])$ is very important: it establishes one possible sense in which the present approach to linear perturbation theory determines potentially everything, namely, that one can extract $[\mathbf{W}] \in \mathcal{W}/\mathcal{W}_L$ from $\mathbf{D} := \varphi([\mathbf{W}]) \in \mathcal{D}$ in a unique way. There is, however, another sense in which one would like to have the basic gauge-invariant variables,

$$\varphi([\mathbf{W}]) := \{\chi, \Gamma, \Omega, S, \Omega^{\alpha}, \Omega^{\alpha\beta}, S^{\alpha\beta}, \Theta^{p}, \Theta_{q}^{pq}, \Omega_{r}^{pq}\},$$

determine everything. This would be to show that any gauge-invariant tensor field defined on X is obtainable linearly from the basic variables through purely algebraic and differential operations. Moreover, we should ask the following questions. What is the relationship between the notion of basic gauge-invariant variables introduced here and in Refs. 10 and 11 and the notion of covariant gauge-invariant variables arising from consideration of the standard theory of Ellis and Bruni?[2] Is Lemma 2.2 of Stewart and Walker[5] sufficient to define an almost-Bianchi type-I or almost-Bianchi type-V universe model, or do new principles remain to be found? How do our ideas relate to the coordinate-based method of Tomita and Den,[27] Den,[28,29] and Noh and Hwang?[30,31] Within the framework here set up, can we find a closed set of linear ''propagation'' equations that involves only **D**? If it does exist and if **D** is a classical solution to these propagation equations, will one be able to construct **W**, which satisfies the condition (2.19), and is such that $\mathbf{D} = \varphi([\mathbf{W}])$? These and similar questions will be formulated and answered in separate papers (see, e.g., Ref. 20).

V. FINAL REMARKS

In this paper, we have developed a theory of cosmological perturbations in anisotropic background models of Bianchi types I and V, based on the geometric approach of Banach and Piekarski.[10,11] For the case of a nonbarotropic perfect fluid, we derived a complete set of basic gauge-invariant variables and concluded with an analysis showing how this set can be used to determine the equivalence class of perturbations (tangents). When considering an arbitrary system of diffeomorphism-invariant, covariant field equations, the establishment of an analytic description of the equivalence classes of tangents must be regarded as the primary problem in any complete formulation of linear perturbation theory, which is why a construction of the set $\varphi([\mathbf{W}])$ of basic gauge-invariant variables is of fundamental importance.[10,11] An additional reason for the mathematical and physical relevance of this construction arises from the fact that, as it will be discussed in a companion paper,[20] the appropriate local algebropartial differential operations on $\varphi([\mathbf{W}])$ enable us to define all gauge-invariant variables.

In the covariant approach[2,9] usually taken to the development of linear perturbation theory, the key result that makes the study of covariantly defined gauge-invariant quantities possible is the Stewart–Walker lemma (see Lemma 2.2 in Ref. 5). This states that if $\{A(x^{\mu},\epsilon); \epsilon \in I\}$ is a C^1 curve of geometrical objects determinable tensor algebraically from the fields $\mathcal{F}_{\epsilon}(x^{\mu}), \epsilon \in I$, and their first-, second-, or higher-order covariant derivatives with respect to $g_{\alpha\beta}(x^{\mu},\epsilon)$, then the linear perturbation δA_0 of a quantity $A_0 := (A)_{\epsilon=0}$ on $(X, q_{\alpha\beta})$, namely, $\delta A_0 := (\partial A/\partial\epsilon)_{\epsilon=0}$, is gauge invariant if and only if one of the following three conditions holds: (i) A_0 vanishes, (ii) A_0 is a constant scalar, or (iii) A_0 is a constant linear combination of products of Kronecker deltas.

How our variables are related with the gauge-invariant variables of Stewart and Walker.[5] To see this (more details will be given in Ref. 20), we need to compute δA_0 by differentiating $A(x^{\mu},\epsilon)$ with respect to ϵ at $\epsilon = 0$. Proceeding in this way and using the definition of a complete set, we can show that any of gauge-invariant variables δA_0 is uniquely determined from a knowledge of $\mathcal{F}_0$ and $\varphi([\mathbf{W}])$ alone. However, in the case of interest to us—namely the construction of $\varphi([\mathbf{W}])$ and $\{\delta A_0\}$ for anisotropic background models of Bianchi types I and V—the converse is not true. More precisely, while the set $\{\delta A_0\}$ Stewart and Walker[5] uses as a basic set consists of gauge-invariant quantities, it does not provide a complete framework for defining $\varphi([\mathbf{W}])$ (i.e., describing the equivalence classes of perturbations), unless supplemented by the extra condition $d_{\alpha\beta} = 0$ that restricts the space–time geometry to Robertson–Walker universe models. By what we have said above, the conclusion is simply this: the construction of $\varphi([\mathbf{W}])$ is superior to the construction of $\{\delta A_0\}$.

Another remark is also in order. The result of Stewart and Walker,[5] that the linear perturbation of some quantity only makes sense if the corresponding unperturbed quantity either is a constant scalar or vanishes or is a constant linear combination of products of Kronecker deltas, emerges as an immediate consequence of their definition of δA_0. Of course, this result has the well-known corollary that since the existence of $\{\delta A_0\}$ is completely dependent on the background chosen, a gauge-invariant formulation of linear perturbation theory by means of $\{\delta A_0\}$ will ''almost always'' be impossible. Fortunately, in our previous papers,[10,11] we introduced the notion of a complete set of basic gauge-invariant variables in a way that eliminated any explicit or implicit reference to the properties of the background and thus presented a satisfactory method of overcoming this apparent difficulty. Consequently, for an arbitrary choice of $\mathcal{F}_0$, linear perturbation theory can be formulated via the introduction of $\varphi([\mathbf{W}])$ in a manner that does not require the existence of $\{\delta A_0\}$.

Next, we mention the following. The considerations of Sec. 4.2 in Ref. 10 allow us to prove that nondegenerate linear perturbation theory for which sufficiently many gauge-invariant variables do exist that the equivalence classes of perturbations are completely determined by them (at least in principle) is a theory with sufficiently many tensor fields on X. Thus, when introducing the notion of a ''coordinate system'' on $\mathcal{W}/\mathcal{W}_L$, it is not necessary to impose restrictions on $\mathcal{F}_0$. Nevertheless, there is some confusion in the literature on this point since it is known mathematically that in the case of, for instance, almost-Robertson–Walker universe models, the set $\varphi([\mathbf{W}])$ differs considerably from that obtained here.[18,19] This point is resolved by remarking that the precise form of $\varphi([\mathbf{W}])$ is dependent on the choice of $\mathcal{F}_0$, while the existence of $\varphi([\mathbf{W}])$ is not.

Finally, because of the general results in Sec. 4 of Ref. 11, we should understand clearly that a complete set of propagation equations can be obtained directly in terms of the basic gauge-invariant variables. The mathematical description of these equations, while elementary, is formally too elaborate to present here. However, we plan to return to it in a later paper.[32] Compared with the Robertson–Walker case, we have one new feature of the cosmological perturbations in a spatially anisotropic background. The nonvanishing background shear causes extremely complicated couplings between (Γ,Ω,S), Ω^{α}, $(\Omega^{\alpha\beta},S^{\alpha\beta})$, and $(\Theta^{p},\Theta_{q}^{pq},\Omega_{r}^{pq})$. Thus, except for some particularly simple situations, the analytical solutions of the propagation equations are very difficult to get, and we have to content ourselves with the numerical analysis. Certain assumptions greatly simplify the discussion, e.g., the requirement that the metric is axially symmetric in the background. Unfortunately, from the viewpoint of the present paper, the typical consequence of these assumptions is that the symmetry properties of $\mathcal{F}_0$ and the definition of $\varphi([\mathbf{W}])$ are *substantially* changed, so it is not a generic strategy. Our basic philosophy was not to derive and solve a simple system of differential equations for the specific gauge-invariant variables, but rather to formulate the mathematical principles underlying the theory of gauge-invariant perturbations $[\mathbf{W}]\in\mathcal{W}/\mathcal{W}_L$ in anisotropic cosmological models.

Perhaps the most interesting concrete application of our findings is to the problem of a *semiclassical description* in which the background geometry is taken in the classical framework and the gauge-invariant perturbations are considered as *quantum variables*. For more details, one should consult Ref. 20.

ACKNOWLEDGMENTS

I am grateful to Dr. S. Piekarski for his continuous interest in my work and many discussions. Also, I would like to thank the anonymous referee for a critical reading of the original manuscript and valuable suggestions on clarifying the exposition.

[1] J. M. Bardeen, Phys. Rev. D **22**, 1882 (1980).
[2] G. F. R. Ellis and M. Bruni, Phys. Rev. D **40**, 1804 (1989).
[3] G. F. R. Ellis and D. R. Matravers, Gen. Relativ. Gravit. **27**, 777 (1995).
[4] J. M. Stewart, Class. Quantum Grav. **7**, 1169 (1990).
[5] J. M. Stewart and M. Walker, Proc. R. Soc. London, Ser. A **341**, 49 (1974).
[6] M. Bruni, P. K. S. Dunsby, and G. F. R. Ellis, Astrophys. J. **395**, 34 (1992).
[7] P. K. S. Dunsby, M. Bruni, and G. F. R. Ellis, Astrophys. J. **395**, 54 (1992).
[8] G. F. R. Ellis, M. Bruni, and J. Hwang, Phys. Rev. D **42**, 1035 (1990).
[9] P. K. S. Dunsby, Phys. Rev. D **48**, 3562 (1993).
[10] Z. Banach and S. Piekarski, Int. J. Theor. Phys. **36**, 1787 (1997).
[11] Z. Banach and S. Piekarski, Int. J. Theor. Phys. **36**, 1817 (1997).
[12] J. Lee and R. M. Wald, J. Math. Phys. **31**, 725 (1990).
[13] V. Iyer and R. M. Wald, Phys. Rev. D **50**, 846 (1994).

[14] Z. Banach and S. Piekarski, J. Math. Phys. **35**, 4809 (1994).
[15] Z. Banach and S. Piekarski, J. Math. Phys. **35**, 5885 (1994).
[16] Z. Banach and S. Piekarski, Int. J. Theor. Phys. **35**, 633 (1996).
[17] Z. Banach and S. Piekarski, Int. J. Theor. Phys. **35**, 665 (1996).
[18] Z. Banach and S. Piekarski, Gen. Relativ. Gravit. **28**, 1335 (1996).
[19] Z. Banach and S. Piekarski, Ann. Inst. Henri Poincaré: Phys. Theor. **65**, 273 (1996).
[20] Z. Banach, J. Math. Phys. **40**, 3995 (1999).
[21] M. P. Ryan and L. C. Shepley, *Homogenous Relativistic Cosmologies* (Princeton University Press, Princeton, NJ, 1975).
[22] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (VEB Deutscher Verlag der Wissenschaften, Berlin, 1980).
[23] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[24] L. Smarr, C. Taubes, and J. R. Wilson, in *Essays in General Relativity*, edited by F. J. Tipler (Academic, New York, 1980), p. 157.
[25] J. Ehlers, in *Relativity, Astrophysics, and Cosmology*, edited by W. Israel (Reidel, Boston, 1973), Vol. 38, p. 1.
[26] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984).
[27] K. Tomita and M. Den, Phys. Rev. D **34**, 3570 (1986).
[28] M. Den, Prog. Theor. Phys. **77**, 653 (1987).
[29] M. Den, Prog. Theor. Phys. **79**, 1110 (1988).
[30] H. Noh and J. Hwang, Phys. Rev. D **52**, 1970 (1995).
[31] H. Noh and J. Hwang, Phys. Rev. D **52**, 5643 (1995).
[32] Z. Banach, in preparation.
[33] We stress that our analysis of the eigenvalue equation (3.4) is dealt with completely in a covariant manner. Thus, the definitions of H_p and l_p^μ have a transparent geometrical meaning, i.e., they do not depend on the choice of (x^μ).
[34] C. J. Isham, in *Canonical Gravity: From Classical to Quantum*, edited by J. Ehlers and H. Friedrich (Springer-Verlag, Berlin, 1994), p. 150.
[35] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds, and Physics. Part I: Basics* (North-Holland, Amsterdam, 1989).
[36] If we choose a coordinate system (x^α) so that the line element of a Bianchi type-I or Bianchi type-V background model takes the form (3.7), then we have $\vartheta^0=0$, $\{\vartheta^\alpha\}=\{0,\vartheta^p\}$, $V_\nu^0=0$, $V^0=0$, and $V_\nu^p=-\partial\vartheta^p/\partial t$. Obviously, with this choice of coordinates, Eqs. (4.23) are equivalent to the differential conditions $\partial\vartheta^p/\partial t=-V^p$; thus they may be integrated to give $\{\vartheta^\alpha\}=\{0,\vartheta^p\}$.
[37] C. Truesdell and R. Toupin, in *Handbuch der Physik*, edited by S. Flügge (Springer-Verlag, Berlin, 1960), p. 226.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Equivalence classes of perturbations in cosmologies of Bianchi types I and V: Interpretation

Zbigniew Banach
Centre of Mechanics, Institute of Fundamental Technological Research, Department of Fluid Mechanics, Polish Academy of Sciences, Swietokrzyska 21, 00-049 Warsaw, Poland

(Received 20 October 1998; accepted for publication 29 March 1999)

In the preceding paper, 26 gauge-invariant variables were defined that characterize an almost-Bianchi type I or almost-Bianchi type V universe filled with a nonbarotropic perfect fluid. One can think of these basic variables, denoted collectively by **D**, as having at least two aspects. First, **D** gives an explicit (i.e., analytical) representation of the equivalence class of perturbations. In fact, this equivalence class is uniquely determined from **D** and vice versa. Second, any gauge-invariant quantity with respect to a Bianchi type I or Bianchi type V background model is obtainable linearly from **D** through purely algebraic and differential operations. Among many other things, the above properties of **D** facilitate new insights into the question of why a standard and well-known formulation based on the Stewart–Walker lemma does not enable one to describe the equivalence classes of perturbations and to find gauge-invariant variables independent of **D**. If we pose an analogous question in regard to almost-Robertson–Walker universe models, a different but none-the-less instructive answer is obtained. In this case, the Stewart–Walker lemma provides a complete framework both for constructing all gauge-invariant variables and for determining the equivalence classes of perturbations. Because of the sometimes confusing statements in the literature, nontrivial comparisons with other work on linear perturbations in anisotropic background models are made. We also present what we believe are some interesting applications of our ideas to the subject of quantum field theory in curved space–time. © *1999 American Institute of Physics.* [S0022-2488(99)01607-2]

I. INTRODUCTION

In this paper, we continue the systematic investigation of an almost-Bianchi type I or type V universe dominated by a nonbarotropic perfect fluid, i.e., a general perfect fluid with two essential thermodynamic variables. Therefore, it should not be surprising that, to avoid the risk of appearing to be repetitive and even trite, *without further comment* we shall use here those symbols and notions which either appear for the first time in Ref. 1 or are reasonably standard, and the analysis proceeds in a way similar to that already made familiar.

Recalling the result of Banach,[1] one can show that a set of 26 ''geometrically'' independent, not identically vanishing gauge-invariant variables, denoted collectively by **D** and referred to as the complete set of basic variables, can be used to extract the equivalence classes of tangents (perturbations) from **D** in a unique way. This paper will focus upon the study of another important property of **D** which can be characterized by the following observation: any gauge-invariant quantity with respect to a Bianchi type I or type V background model is obtainable *linearly* from the basic variables **D** through *purely* local (i.e., algebraic and differential) operations. Among many other things, the above observation facilitates new insights into the question of why the Stewart–Walker lemma (see Lemma 2.2 in Ref. 2) *does not enable one* to find gauge-invariant variables independent of **D**. Denoting by δA_0 the linear perturbation of a geometric background quantity A_0, this lemma states that for δA_0 to be invariant under the action of an ''infinitesimal diffeomorphism''[3] A_0 must either be a constant scalar, vanish, or be a linear combination of products of Kronecker deltas with constant coefficients. However, if A_0 satisfies one of these three conditions, then the gauge-invariant quantity δA_0 can be expressed locally in terms of $\mathcal{F}_0$ and **D**, and a knowledge of δA_0 *is not necessary* to describe an almost-Bianchi type I or type V universe model.

For these models, the converse is not true. More precisely, a problem which begins: ''find a complete, finite set $\{\delta A_0\}$ of covariantly defined, gauge-invariant quantities δA_0 with a simple geometric and physical meaning, that code the information we need to determine the equivalence classes of perturbations, and extract **D** from $\mathcal{F}_0$ and $\{\delta A_0\}$ in a unique way'' *will be impossible to solve*. By what we have said above, the conclusion is simply this: the construction of **D** is superior to the construction of $\{\delta A_0\}$. However, if we pose an analogous question in regard to almost-Robertson–Walker universe models, a different but none-the-less instructive answer will be obtained.[4,5] In this case, the inverse problem *has a positive solution*, and we conclude that the Stewart–Walker lemma provides a complete framework both for constructing *all* ''geometrically'' independent, not identically vanishing gauge-invariant variables and for determining the equivalence classes of perturbations (see Ref. 5, pp. 289–292).

Various motivations lie behind such a careful analysis of the properties of **D**. To begin with, for an arbitrary background space–time $(X, q_{\alpha\beta})$, there does not appear to be any natural choice of δA_0, nor does there appear to be any unified treatment of the exact and the linearized theory. Thus in particular, the covariant formalism of Ellis and Bruni[6] cannot be extended to this case. It might seem that the lack of an acceptable method for obtaining sufficiently many covariant gauge-invariant variables[7] would pose an insurmountable obstacle to the formulation of linear perturbation theory in general. However, in our previous papers,[8,9] (see also Ref. 1), we introduced the notion of a gauge-invariant variable in a way that eliminated any explicit or implicit reference to Lemma 2.2 of Stewart and Walker[2] and thus presented a satisfactory means of overcoming these apparent difficulties. Moreover, we have found an intriguing result:[8,9] the construction of a ''coordinate system'' on $\mathcal{P}/\mathcal{P}_L$ or $\mathcal{W}/\mathcal{W}_L$ does not depend on the specific symmetry properties of the background space–time geometry chosen; in other words, the set **D** can be proven to exist for any possible choice of the background.

To define a set of gauge-invariant variables, most analysis of inhomogeneities in an expanding universe have categorized the metric and matter perturbations into three distinct types: scalar, vector, and tensor perturbations.[10–15] Such an approach to the gauge problem has been criticized by Ellis and Matravers[16] on the grounds that (i) they doubt the value of introducing the idea of ''gauge-invariant harmonic amplitudes'' into general relativity at all and (ii) if one is to introduce it, then one should accept a concept which is completely dependent on the coordinate choice made (for example, the Bardeen formalism[10] is gauge invariant only if one restricts coordinate changes to the vector kinds when vector modes are investigated, and to the scalar kind when the scalar modes are the theme of interest). The present formalism enables us to reformulate the coordinate-based method of Bardeen[10] in terms of properties of the complete set **D**. This reformulation is both *fully* covariant and gauge invariant; thus it sidesteps the usual problems and *resolves disagreements* between the results of different existing formalisms.

Are there at least possible concrete applications of our ideas? With regard to this question, it is an important—and conceptually and mathematically nontrivial—challenge to examine the effect of using a *semiclassical description*[17–21] in which the background geometry is taken in the classical framework and the gauge-invariant perturbations are considered as *quantum variables*. Progress towards this goal has been slow, mainly because of the fact that a ''presymplectic vector space structure''[19] $\overline{\omega}$ defined on the space $\mathcal{P}$ of infinitesimal perturbations $\delta\mathcal{F}_0$ is not necessarily nondegenerate. When it is degenerate, we must reduce $\mathcal{P}$ by taking the ''symplectic quotient'' of $(\mathcal{P}, \overline{\omega})$,[22] thereby producing a quotient space $\mathcal{P}/\mathcal{P}_L$ on which there is defined a nondegenerate symplectic vector space structure ω. The complete set **D** of basic gauge-invariant variables should be of great use in the study of these problems; up to the present, one has simply never succeeded in going beyond the self-evident observation that $\overline{\omega}$ always is gauge invariant, i.e., for any pair of linearized solutions $\delta\mathcal{F}_0$ and $\delta\mathcal{F}_0'$ of Eq. (2.14) in Ref. 1, the quantity $\overline{\omega}(\delta\mathcal{F}_0, \delta\mathcal{F}_0')$ depends only upon the gauge equivalence class of $\delta\mathcal{F}_0$ and $\delta\mathcal{F}_0'$. Indeed, it would appear that one is now in a position to formulate, for the first time, a sensible prescription—valid for an arbitrary Lagrangian field theory—for constructing ω from $\overline{\omega}$ and **D**.

The program of this paper is as follows. Section II introduces a number of geometric background quantities A_0 satisfying one of the three conditions in the Stewart–Walker lemma and then shows that from a knowledge of **D** alone, together with the information about $\mathcal{F}_0$, one can compute the linear perturbations of these quantities, namely $\{\delta A_0\}$. In Sec. III we compare our results with Bardeen's and with other treatments of gauge-invariant perturbations in anisotropic background models. Specifically, we relate our work to Tomita and Den,[11] Den,[12,13] and Noh and Hwang[14,15] on the Bardeen nonlocal approach,[10] and to Dunsby[23] on the Ellis–Bruni covariant method.[6] In Sec. IV we demonstrate how the entire formalism can be adapted naturally into a semiclassical description. Section V is for final remarks.

II. RELATIONSHIP WITH THE STEWART–WALKER LEMMA

Let $\{A(x^\mu,\epsilon);\ \epsilon\in I\}$ be a curve of geometrical objects determinable tensor-algebraically from the fields $\mathcal{F}_\epsilon(x^\mu)$, $\epsilon\in I$, and their first-, second-, or higher-order covariant derivatives with respect to $g_{\alpha\beta}(x^\mu,\epsilon)$, and suppose that $A(x^\mu,\epsilon)$ depends continuously and differentiably on ϵ. In this way of thinking, it will be convenient to regard $\{\mathcal{F}_\epsilon(\cdot);\ \epsilon\in I\}$ *not* as a one-parameter family of exact solutions to the nonlinear field equations, but rather as an arbitrary C^r curve (r sufficiently large) in the function space $\{\mathcal{F}(\cdot)\}$ passing through the background solution $\mathcal{F}_0(\cdot)$. With the notation $A_0 \coloneqq (A)_{\epsilon=0}$, we can introduce the quantity δA, denoted in the Introduction by δA_0, which represents the *linear perturbation* of A_0:

$$\delta A \coloneqq \left(\frac{\partial A}{\partial \epsilon}\right)_{\epsilon=0}. \tag{2.1}$$

As shown already by Ehlers,[24] this linear perturbation is gauge invariant (i.e., invariant under the action of an ''infinitesimal diffeomorphism''[3,25]) if and only if the following condition holds for each vector field v on X:[26]

$$\mathcal{L}_v A_0 = 0. \tag{2.2}$$

According to the Stewart–Walker lemma (see Lemma 2.2 in Ref. 2), in order to satisfy the above condition, it is necessary to use a scalar A that is constant in the unperturbed space–time $(X,q_{\alpha\beta})$, or any tensor $A_\alpha\ldots$ that vanishes in $(X,q_{\alpha\beta})$, or a tensor whose ''background value'' is a constant linear combination of products of Kronecker's deltas $\delta^\alpha{}_\beta$.

Having made these general remarks, we wish now to provide representative examples of the gauge-invariant quantities δA. Next, we will prove that, for the *general* Bianchi type I or type V background model $(k=0,-1)$, a knowledge of the complete set of basic variables, namely

$$\varphi([\mathbf{W}]) \coloneqq \{\chi,\Gamma,\Omega,S,\Omega^\alpha,\Omega^{\alpha\beta},S^{\alpha\beta},\Theta^p,\Theta^{pq}_q,\Omega^{pq}_r\}, \tag{2.3}$$

determines δA in a unique way. If the background spatial curvature k vanishes (Bianchi type I space–times), the set $\varphi([\mathbf{W}])$ can be replaced by

$$\varphi_1([\mathbf{W}]) \coloneqq \{\chi,\Gamma,\Omega,S,\Omega^\alpha,\Omega^{\alpha\beta},S^{\alpha\beta},\Omega^{\alpha\beta}_\mu,\Theta^p\}. \tag{2.4}$$

The mathematical details of our construction of $\varphi([\mathbf{W}])$ and $\varphi_1([\mathbf{W}])$ are given in Sec. IV B of Ref. 1 [see especially Eqs. (4.8), (4.11), and (4.18)]. For simplicity, we abbreviate $\varphi([\mathbf{W}])$ as **D** and $\varphi_1([\mathbf{W}])$ as $\mathbf{D}_1$.

A. Discussion of the case when $H_1\neq H_2\neq H_3\neq H_1$

From the above analysis plus the definition $h_{\alpha\beta}\coloneqq g_{\alpha\beta}+u_\alpha u_\beta$ and the decomposition of $u_{\alpha;\beta}$ [see Eq. (3.1) in Ref. 1] we conclude that the simplest physical objects $A(\cdot,\epsilon)$ satisfying for $\epsilon=0$ the condition (2.2) can be described as follows:

(1) The specific entropy:

$$s := \text{entropy per particle.} \tag{2.5}$$

(2) The normalized curvature scalars:

$$\mathcal{A} := \frac{1}{n^{2/3}} \left[-\frac{2}{3}(\theta)^2 + \sigma^{\mu\nu}\sigma_{\mu\nu} + R^{\mu}{}_{\mu} + 2u^{\mu}u^{\nu}R_{\mu\nu} \right], \tag{2.6a}$$

$$\mathcal{B} := \frac{1}{n^{2/3}} \left[-\frac{2}{3}(\theta)^2 + \sigma^{\mu\nu}\sigma_{\mu\nu} + 2(e+\Lambda) \right], \tag{2.6b}$$

$$\mathcal{C} := \frac{1}{n^{2/3}} [-2(\theta)^2 - 4u^{\mu}\theta_{;\mu} - 3\sigma^{\mu\nu}\sigma_{\mu\nu} - 6(p-\Lambda)]. \tag{2.6c}$$

(3) The orthogonal spatial gradients of n and $\theta := u^{\mu}{}_{;\mu}$:

$$X^{\alpha} := h^{\alpha\beta} n_{;\beta}, \quad Z^{\alpha} := h^{\alpha\beta}\theta_{;\beta}. \tag{2.7}$$

(4) The acceleration:

$$a^{\alpha} := u^{\beta} u^{\alpha}{}_{;\beta}. \tag{2.8}$$

(5) The spatial two-tensor associated with $\sigma_{\alpha\beta}$:

$$\mathcal{E}_{\alpha\beta} := h^{\mu}{}_{\alpha} h^{\nu}{}_{\beta} (u^{\lambda}\sigma_{\mu\nu;\lambda} + \theta\sigma_{\mu\nu}). \tag{2.9}$$

(6) The normalized trace-free curvature tensor:

$$\mathcal{A}^{\alpha\beta} := -\frac{1}{3}\mathcal{A}h^{\alpha\beta} + \frac{1}{n^{2/3}} \left[-\frac{2}{9}(\theta)^2 h^{\alpha\beta} - \frac{1}{3}\theta\sigma^{\alpha\beta} + \sigma^{\alpha}{}_{\mu}\sigma^{\mu\beta} + R_{\mu\nu}h^{\mu\alpha}h^{\nu\beta} + R^{\lambda}{}_{\mu\tau\nu}u_{\lambda}u^{\tau}h^{\mu\alpha}h^{\nu\beta} \right]. \tag{2.10}$$

Several points concerning these definitions should be noted. First, we assume that there exists the specific entropy, i.e., the entropy per particle. We denote this entropy by s. If $\mathcal{F}_{\epsilon}(\cdot)$ is a solution of the nonlinear field equations, and such is *certainly* the case when $\epsilon=0$, then $s:=s(\cdot,\epsilon)$ satisfies the property that $(nsu^{\alpha})_{;\alpha}=0$. After combining this property of s with the equation of balance of number density, we derive that the nonbarotropic perfect fluid is locally adiabatic:[27,28]

$$u^{\alpha}s_{;\alpha}=0. \tag{2.11}$$

That is, entropy is constant along the flow lines of the fluid. In this way, we arrive at the following conclusion: the specific entropy s is a scalar that is *constant* in the unperturbed space–time $(X,q_{\alpha\beta})$.

Second, it can be shown that the *same* conclusion concerns $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. More precisely, if $\epsilon=0$, we have $kn_0 \sim k(R_1)^{-3}$ and $\mathcal{A}=\mathcal{B}=\mathcal{C}=(n_0)^{-2/3}[6k/(R_1)^2]=\text{constant}$; thus the Lie derivatives[26] $\mathcal{L}_v\mathcal{A}_0$, $\mathcal{L}_v\mathcal{B}_0$, and $\mathcal{L}_v\mathcal{C}_0$ vanish. For $\epsilon\neq 0$, provided $\mathcal{F}_{\epsilon}(\cdot)$ satisfies the nonlinear field equations, $\mathcal{A}$ equals $\mathcal{B}$ and $\mathcal{C}$ is dynamically related to $\mathcal{B}$, a^{μ}, and $\omega_{\mu\nu}$ by $\mathcal{C}=\mathcal{B}-4n^{-2/3}(a^{\mu}_{;\mu}+\omega^{\mu\nu}\omega_{\mu\nu})$. Under these circumstances, in the special case of vanishing vorticity (and only then) the quantity $n^{2/3}\mathcal{B}$ acquires a special significance:[6] it is the Ricci scalar ${}^3R$ of the three-dimensional spaces $\Sigma_{\perp}$ everywhere orthogonal to the fluid flow vector u^{α}; that is, $\omega_{\mu\nu}=0 \Rightarrow {}^3R=n^{2/3}\mathcal{B}$. For this reason, we call $\mathcal{B}$ (as well as $\mathcal{A}$ and $\mathcal{C}$) the normalized curvature scalar. (By way of digression, the meaning of $n^{2/3}\mathcal{B}$ when $\omega^{\mu\nu}\omega_{\mu\nu}\neq 0$ is discussed in Appendix B of Ref. 6.)

Third, the important physical variable turns out to be the orthogonal density gradient X^α defined by one of Eqs. (2.7). Ellis and Bruni,[6] Bruni *et al.*,[29] Dunsby *et al.*,[30] and Dunsby[23] gave the first systematic treatment of the properties of this variable. From the viewpoint of the formalism developed by Ellis and Bruni,[6] both quantities in Eqs. (2.7) are gauge invariant, as they vanish in the Bianchi type I and type V universes $[X^\alpha \cong \epsilon(\delta X^\alpha)$ and $Z^\alpha \cong \epsilon(\delta Z^\alpha)]$. Interpreting Eq. (2.8), this equation defines the standard kinematic quantity which vanishes in the background, and so is *gauge invariant*. As regards $\mathcal{E}_{\alpha\beta}$ [see Eq. (2.9)], this quantity is *also* gauge invariant, because of the condition[1] $(\mathcal{E}_{\alpha\beta})_{\epsilon=0} = \dot{d}_{\alpha\beta} + 3Hd_{\alpha\beta} = 0$. Note that $\mathcal{E}_{\alpha\beta}$ is constructed from the shear tensor $\sigma_{\alpha\beta}$ and that $\mathcal{E}_{\alpha\beta} = \mathcal{E}_{\beta\alpha}$, $g^{\alpha\beta}\mathcal{E}_{\alpha\beta} = 0$, and $\mathcal{E}_{\alpha\beta}u^\beta = 0$.

Next, using the definition (2.10) of $\mathcal{A}^{\alpha\beta}(\mathcal{A}^{\alpha\beta} = \mathcal{A}^{\beta\alpha}, g_{\alpha\beta}\mathcal{A}^{\alpha\beta} = 0, \mathcal{A}^{\alpha\beta}u_\beta = 0)$, we obtain $(\mathcal{A}^{\alpha\beta})_{\epsilon=0} = 0$; thus $\delta\mathcal{A}^{\alpha\beta}$ is *gauge invariant*. If $\omega_{\alpha\beta} = 0$, the tensor $\mathcal{A}^{\alpha\beta}$, which we call the normalized trace-free curvature tensor, is related to the trace-free part of the Ricci tensor ${}^3R_{\alpha\beta}$ of $\Sigma_\perp$. In fact, given a unit timelike vector u^α, we can define at each point the three-curvature tensor ${}^3R^\mu{}_{\alpha\nu\beta}$ and its trace ${}^3R_{\alpha\beta} := {}^3R^\mu{}_{\alpha\mu\beta}$. When u^α is hypersurface-orthogonal, these are the Riemann and Ricci tensors of three-surfaces $\Sigma_\perp$, but in general $\omega_{\alpha\beta} \neq 0 \Rightarrow {}^3R_{[\alpha\beta]} \neq 0$, and these tensors do not have all the usual symmetries of Riemann and Ricci (more details in Refs. 29 and 31). Let u^α be the fluid four-velocity and suppose that $\omega_{\alpha\beta} = 0$. Then we can split ${}^3R_{\alpha\beta}$ into its trace ${}^3R = n^{2/3}\mathcal{A}$ and its symmetric trace-free part ${}^3\mathcal{R}_{\alpha\beta} = n^{2/3}(g_{\alpha\mu}g_{\beta\nu}\mathcal{A}^{\mu\nu})$. The tensor field $(\mathcal{A}^{\alpha\beta})_{\epsilon=0}$ vanishes because $({}^3R^\mu{}_{\alpha\nu\beta})_{\epsilon=0}$ equals $[k/(R_1)^2](\gamma^\mu{}_\nu\gamma_{\alpha\beta} - \gamma^\mu{}_\beta\gamma_{\alpha\nu})$.

Finally, it is easy to find many further gauge-invariant quantities by finding more complex *invariantly* defined quantities that vanish in the background, for example, the vorticity; the orthogonal spatial gradients of $T, e, p, s, n^{2/3}\mathcal{A}, n^{2/3}\mathcal{B}$, and $n^{2/3}\mathcal{C}$; the divergence of the acceleration and its spatial gradient; and so on:

$$\omega_{\alpha\beta} := h^\mu{}_\alpha h^\nu{}_\beta u_{[\mu;\nu]}, \tag{2.12a}$$

$$\Pi^\alpha := h^{\alpha\beta}T_{;\beta}, \quad \overline{X}^\alpha := h^{\alpha\beta}e_{;\beta}, \tag{2.12b}$$

$$Y^\alpha := h^{\alpha\beta}p_{;\beta}, \quad \mathcal{E}^\alpha := h^{\alpha\beta}s_{;\beta}, \tag{2.12c}$$

$$C^\alpha := h^{\alpha\beta}(n^{2/3}\mathcal{A})_{;\beta}, \quad \overline{C}^\alpha := h^{\alpha\beta}(n^{2/3}\mathcal{B})_{;\beta}, \tag{2.12d}$$

$$\tilde{C}^\alpha := h^{\alpha\beta}(n^{2/3}\mathcal{C})_{;\beta}, \tag{2.12e}$$

$$\Xi := a^\mu{}_\mu, \quad \Xi^\alpha := h^{\alpha\beta}\Xi_{;\beta}. \tag{2.12f}$$

Together with the variables $\delta\mathcal{C}$ and Z^α, these will not be significant to us in considering linearization around the Bianchi type I and type V universes, for there are *simpler* gauge-invariant quantities which can be used to determine them [see Eqs. (2.17a)–(2.17k)].

Before calculating δA for the quantities defined by Eqs. (2.5)–(2.10) and (2.12), let us introduce the following notation:[1]

$$e_0 := e(n_0, T_0), \quad p_0 := p(n_0, T_0), \tag{2.13a}$$

$$e_T := \frac{\partial e_0}{\partial T_0}, \quad p_T := \frac{\partial p_0}{\partial T_0}, \tag{2.13b}$$

$$N := u^\mu n_{;\mu} + nu^\mu{}_{;\mu}, \tag{2.13c}$$

$$\vec{\gamma} \circ \vec{\Theta} := \sum_{p=1}^{3} [(l_p^\mu l_p^\nu \gamma_{\mu\nu})\Theta^p], \tag{2.13d}$$

$$\vec{d}\circ\vec{\Theta}:=\sum_{p=1}^{3}\,[(l_p^{\mu}l_p^{\nu}d_{\mu\nu})\Theta^p]. \tag{2.13e}$$

From Eq. (2.13c) we obtain $N_0:=(N)_{\epsilon=0}=0$, so that the linear perturbation δN of N_0 is gauge invariant. However, if $\{\mathcal{F}_\epsilon(\cdot);\ \epsilon\in I\}$ *is* a one-parameter family of exact solutions to the nonlinear field equations, then $N(\cdot,\epsilon)=0$ and δN *vanishes identically*. With all these preparatory remarks in mind, it is only a matter of labor to prove that

$$\delta N=\tfrac{1}{2}n_0(\vec{\gamma}\circ\vec{\Theta}) \tag{2.14a}$$

and that the linear perturbations of s, $\mathcal{A}$, $\mathcal{B}$, X^α, a^α, $\mathcal{E}_{\alpha\beta}$, and $\mathcal{A}^{\alpha\beta}$ at $\epsilon=0$ are related to the gauge-invariant variables Γ, Ω, S, Ω^α, $\Omega^{\alpha\beta}$, $S^{\alpha\beta}$, and Θ^p in $\mathbf{D}:=\varphi([\mathbf{W}])$ [see Eq. (2.3)] by[32]

$$\delta s=\frac{e_T}{n_0}\Gamma, \tag{2.14b}$$

$$\delta\mathcal{A}=\frac{2}{(n_0)^{2/3}}\left(\frac{2}{3}\Omega^{\mu}{}_{|\mu}-\frac{1}{3H}d^{\mu}{}_{\nu}\Omega^{\nu}{}_{|\mu}-S\right), \tag{2.14c}$$

$$\delta\mathcal{B}=\frac{2}{(n_0)^{2/3}}\left\{2\left[\frac{3k}{(R_1)^2}-e_0-\Lambda\right]\Omega-H(\vec{\gamma}\circ\vec{\Theta})+\frac{1}{2}(\vec{d}\circ\vec{\Theta})+T_0e_T\Gamma\right\}, \tag{2.14d}$$

$$\delta X^\alpha=n_0\Omega^\alpha, \tag{2.14e}$$

$$\delta a^\alpha=-\frac{1}{3H}\dot{\Omega}^\alpha+\frac{1}{3}\left(\frac{\dot{H}}{H^2}-1\right)\Omega^\alpha-\frac{1}{3H}d^{\alpha}{}_{\mu}\Omega^\mu-\gamma^{\alpha\mu}\Omega_{|\mu}, \tag{2.14f}$$

$$\begin{aligned}\delta\mathcal{E}_{\alpha\beta}=&-\frac{1}{6}\gamma_{\alpha\beta}(\vec{\gamma}\circ\vec{\Theta})^{\cdot}-\frac{1}{2}H\gamma_{\alpha\beta}(\vec{\gamma}\circ\vec{\Theta})+\dot{\Omega}d_{\alpha\beta}+\frac{1}{2}d_{\alpha\beta}(\vec{\gamma}\circ\vec{\Theta})\\&+\frac{1}{2}\gamma_{\alpha\mu}\gamma_{\beta\nu}\Omega^{\mu\nu}+\frac{1}{3H}d_{\mu(\alpha}\Omega^{\mu}{}_{|\beta)}-\frac{1}{3H}\gamma_{\nu(\alpha}d^{\mu}{}_{\beta)}\Omega^{\nu}{}_{|\mu},\end{aligned} \tag{2.14g}$$

$$\begin{aligned}\delta\mathcal{A}^{\alpha\beta}=-\frac{1}{(n_0)^{2/3}}\Bigg[&2S^{\alpha\beta}+\frac{1}{9}\gamma^{\alpha\beta}\Omega^{\mu}{}_{|\mu}-\frac{1}{3}\gamma^{\mu(\alpha}\Omega^{\beta)}{}_{|\mu}-\frac{1}{3H}d^{\alpha\beta}\Omega^{\mu}{}_{|\mu}\\&+\frac{1}{3H}\left(\gamma^{\mu(\alpha}d^{\beta)}{}_{\nu}-\frac{2}{3}\gamma^{\alpha\beta}d^{\mu}{}_{\nu}\right)\Omega^{\nu}{}_{|\mu}+\frac{1}{3H}d^{\mu(\alpha}\Omega^{\beta)}{}_{|\mu}+\frac{2}{3H}(\gamma^{\nu(\alpha}d^{\beta)}{}_{\mu|\nu}-d^{\alpha\beta}{}_{|\mu})\Omega^{\mu}\Bigg].\end{aligned} \tag{2.14h}$$

Equations (2.14g) and (2.14h) yield the conditions

$$\gamma^{\mu\nu}(\delta\mathcal{E}_{\mu\nu})=0,\quad \gamma_{\mu\nu}(\delta\mathcal{A}^{\mu\nu})=0 \tag{2.15a}$$

and

$$d^{\mu\nu}(\delta\mathcal{E}_{\mu\nu})=(d^{\mu\nu}d_{\mu\nu})\dot{\Omega}+\tfrac{1}{2}(d^{\mu\nu}d_{\mu\nu})(\vec{\gamma}\circ\vec{\Theta})+\tfrac{1}{2}(\vec{d}\circ\vec{\Theta})^{\cdot}+3H(\vec{d}\circ\vec{\Theta}), \tag{2.15b}$$

where we have used the identity (4.14b) of Ref. 1.

Given all these results, especially Eqs. (2.14a)–(2.14h), *it is now obvious mathematically* how to extract δN, δs, $\delta\mathcal{A}$, $\delta\mathcal{B}$, δX^α, δa^α, $\delta\mathcal{E}_{\alpha\beta}$, and $\delta\mathcal{A}^{\alpha\beta}$ from a knowledge of $\mathcal{F}_0$ and $\mathbf{D}$ *alone*. At the same time, we may interpret some of the above formulas in another way by noticing that Γ, Ω, S, Ω^{α}, $\Omega^{\alpha\beta}$, and $S^{\alpha\beta}$ are *uniquely* determined from $\mathcal{F}_0$, δs, $\delta\mathcal{A}$, $\delta\mathcal{B}$, δX^{α}, $\delta\mathcal{E}_{\alpha\beta}$, $\delta\mathcal{A}^{\alpha\beta}$, and Θ^p $(p=1,2,3)$. Thus, the set

$$\varphi_2([\mathbf{W}]) := \{\chi, \delta s, \delta\mathcal{B}, \delta\mathcal{A}, \delta X^{\alpha}, \delta\mathcal{E}_{\alpha\beta}, \delta\mathcal{A}^{\alpha\beta}, \Theta^p, \Theta^{pq}_q, \Omega^{pq}_r\} \tag{2.16}$$

constitutes a *new* complete set of basic gauge-invariant variables; this set contains all significant information about $[\mathbf{W}]$. Since the specification of $\varphi_2([\mathbf{W}])$ for each $[\mathbf{W}]$ is mathematically equivalent to the specification of $\mathbf{W} \mapsto \varphi([\mathbf{W}])$, it can also be interpreted as a construction of the *alternative* ''coordinate system''[1] on $\mathcal{W}/\mathcal{W}_L$. Nevertheless, it should be stressed that if $d^{\mu\nu}d_{\mu\nu} \neq 0$ and $\vec{d}\circ\vec{\Theta} \neq 0$, then Eqs. (2.14a), (2.14b), (2.14d), (2.14e), (2.14g), and (2.15b) *are not sufficient* to express Ω and $\Omega^{\alpha\beta}$ in terms of *only* $\mathcal{F}_0$, δN, δs, $\delta\mathcal{B}$, δX^{α}, and $\delta\mathcal{E}_{\alpha\beta}$. Similarly, in our model, *serious difficulties* would arise if we attempt to define a meaningful set of gauge-invariant variables δA for the determination of Θ^p, Θ^{pq}_q, and Ω^{pq}_r. Put somewhat differently, knowledge of the set $\mathbf{D}$ allows one to compute $\{\delta A\}$, but knowledge of the set $\{\delta A\}$ *does not permit one* inversely to compute $\mathbf{D}$. Moreover, except for some ''particularly simple'' situations ($d_{\alpha\beta}=0$; see Refs. 4 and 5), *we may be no longer able* to represent the equivalence classes of perturbations by means of $\{\delta A\}$.

Turning now to the formulas (2.12) and the quantities $\mathcal{C}$ and Z^{α} in Eqs. (2.6c) and (2.7), the corresponding linear perturbations δA are determined by the relations

$$\delta\omega_{\alpha\beta} = -\frac{1}{3Hn_0}\gamma_{\mu[\alpha}(\delta X^{\mu})_{|\beta]}, \tag{2.17a}$$

$$\delta\Pi^{\alpha} = -\frac{1}{3Hn_0}\dot{T}_0(\delta X^{\alpha}) + \frac{1}{e_T}n_0T_0\gamma^{\alpha\beta}(\delta s)_{|\beta}, \tag{2.17b}$$

$$\delta\overline{X}^{\alpha} = \frac{1}{n_0}(e_0+p_0)(\delta X^{\alpha}) + n_0T_0\gamma^{\alpha\beta}(\delta s)_{|\beta}, \tag{2.17c}$$

$$\delta Y^{\alpha} = -\frac{1}{3Hn_0}\dot{p}_0(\delta X^{\alpha}) + \frac{1}{e_T}n_0T_0p_T\gamma^{\alpha\beta}(\delta s)_{|\beta}, \tag{2.17d}$$

$$\delta Z^{\alpha} = -3H(\delta a^{\alpha}) - \frac{1}{n_0}[(\delta X^{\alpha})^{\cdot} + 4H(\delta X^{\alpha}) + d^{\alpha}{}_{\mu}(\delta X^{\mu}) - \gamma^{\alpha\mu}(\delta N)_{|\mu}], \tag{2.17e}$$

$$\delta\mathcal{E}^{\alpha} = \gamma^{\alpha\beta}(\delta s)_{|\beta}, \tag{2.17f}$$

$$\begin{aligned}\delta\mathcal{C} = \frac{1}{H}(\delta\mathcal{B})^{\cdot} + \delta\mathcal{B} - \frac{2}{H(n_0)^{2/3}}\Bigg\{ & d^{\mu\nu}(\delta\mathcal{E}_{\mu\nu}) - \frac{2}{n_0}\left[\frac{k}{(R_1)^2} - \frac{1}{2}(e_0+p_0)\right](\delta N) \\ & + n_0T_0(\delta s)^{\cdot} + \left(n_0\dot{T}_0 - \frac{1}{e_T}T_0p_T\dot{n}_0\right)(\delta s)\Bigg\},\end{aligned} \tag{2.17g}$$

$$\delta C^{\alpha} = \frac{4k}{n_0(R_1)^2}(\delta X^{\alpha}) + (n_0)^{2/3}\gamma^{\alpha\beta}(\delta\mathcal{A})_{|\beta}, \tag{2.17h}$$

$$\delta\overline{C}^{\alpha} = \frac{4k}{n_0(R_1)^2}(\delta X^{\alpha}) + (n_0)^{2/3}\gamma^{\alpha\beta}(\delta\mathcal{B})_{|\beta}, \tag{2.17i}$$

$$\delta\tilde{C}^{\alpha} = \frac{4k}{n_0(R_1)^2}(\delta X^{\alpha}) + (n_0)^{2/3}\gamma^{\alpha\beta}(\delta\mathcal{C})_{|\beta}, \tag{2.17j}$$

$$\delta\Xi=(\delta a^{\mu})_{|\mu}, \quad \delta\Xi^{\alpha}=\gamma^{\alpha\beta}[(\delta a^{\mu})_{|\mu}]_{|\beta}. \tag{2.17k}$$

Given these linear perturbations, we might now say justifiably that *they all can be generated from* $\mathcal{F}_0$, δN, δs, $\delta\mathcal{A}$, $\delta\mathcal{B}$, δX^{α}, δa^{α}, and $\delta\mathcal{E}_{\alpha\beta}$ via a local algebropartial differential transformation, so that, in linear perturbation theory, Eqs. (2.14) are *superior* to those just obtained.

The analysis so far has been for an almost-Bianchi type I or type V universe model ($k=0$, -1). However, in the case of a Bianchi type I cosmology ($k=0$), it proves useful, as we already mentioned,[1] both to replace Θ^{pq}_{q} and Ω^{pq}_{r} by a new gauge-invariant quantity $\Omega^{\alpha\beta}_{\mu}$ [see Eq. (4.18) in Ref. 1] and to write Eq. (2.16) in the more convenient form

$$\varphi_3([\mathbf{W}]):=\{\chi,\delta s,\delta\mathcal{B},\delta\mathcal{A},\delta X^{\alpha},\delta\mathcal{E}_{\alpha\beta},\delta\mathcal{A}^{\alpha\beta},\Omega^{\alpha\beta}_{\mu},\Theta^{p}\}. \tag{2.18}$$

Now, consider a tensor field $\mathcal{E}_{\alpha\beta\mu}$ ($\mathcal{E}_{[\alpha\beta]\mu}=0, g^{\alpha\beta}\mathcal{E}_{\alpha\beta\mu}=0, u^{\alpha}\mathcal{E}_{\alpha\beta\mu}=u^{\alpha}\mathcal{E}_{\beta\mu\alpha}=0$) defined in terms of the shear tensor $\sigma_{\alpha\beta}$ by

$$\mathcal{E}_{\alpha\beta\mu}:=h^{\tau}{}_{\alpha}h^{\lambda}{}_{\beta}h^{\nu}{}_{\mu}\sigma_{\tau\lambda;\nu}. \tag{2.19}$$

If $k=0$, this tensor field vanishes in the background, and we obtain the result that $\delta\mathcal{E}_{\alpha\beta\mu}$ is a *gauge-invariant* quantity. Proceeding through the same steps as lead to the formulas (2.14) and (2.17), we verify that a knowledge of $\mathcal{F}_0$ and $\{\Omega,\Omega^{\alpha},\Omega^{\alpha\beta}_{\mu},\Theta^{p}\}\subset\mathbf{D}_1$ *is all we need* to determine $\delta\mathcal{E}_{\alpha\beta\mu}$:

$$\delta\mathcal{E}_{\alpha\beta\mu}=\gamma_{\mu\nu}\Omega^{\nu}d_{\alpha\beta}+\frac{2}{3}[d_{\mu(\alpha}\gamma_{\beta)\nu}-\gamma_{\mu(\alpha}d_{\beta)\nu}]\Omega^{\nu}+\frac{2}{3H}[d^{\sigma}{}_{\mu}d_{\sigma(\alpha}\gamma_{\beta)\nu}-d_{\mu(\alpha}d_{\beta)\nu}]\Omega^{\nu}$$
$$-\frac{1}{6}\gamma_{\alpha\beta}(\vec{\gamma}\circ\vec{\Theta})_{|\mu}+d_{\alpha\beta}\Omega_{|\mu}+\frac{1}{2}\gamma_{\alpha\nu}\gamma_{\beta\sigma}\Omega^{\nu\sigma}_{\mu}. \tag{2.20}$$

Then, from Eq. (4.19) in Ref. 1, we get

$$\gamma^{\alpha\beta}(\delta\mathcal{E}_{\alpha\beta\mu})=0, \tag{2.21}$$

as it should be. Alternatively, within the framework here set up, $\Omega^{\alpha\beta}_{\mu}$ is obtainable from $\mathcal{F}_0$, δs, $\delta\mathcal{B}$, δX^{α}, $\delta\mathcal{E}_{\alpha\beta\mu}$, and Θ^{p} ($p=1,2,3$) through purely algebraic and differential operations, so that $[\mathbf{W}]$ can be described by means of the gauge-invariant variables Θ^{p} and χ ($\chi=-[\partial(u^{\alpha}u_{\alpha})/\partial\epsilon]_{\epsilon=0}=0$) and the following system of linear perturbations:

$$\mathbf{E}:=\{\delta s,\delta\mathcal{B},\delta\mathcal{A},\delta X^{\alpha},\delta\mathcal{E}_{\alpha\beta},\delta\mathcal{A}^{\alpha\beta},\delta\mathcal{E}_{\alpha\beta\mu}\}. \tag{2.22}$$

Thus, as in an almost-Bianchi type V universe model discussed before, the net upshot of these considerations is simply this. Since the *scalar* fields Θ^{p} ($p=1,2,3$) *are not expressible in terms of* $\{\delta A\}$ unless the background shear takes a simple form $[d_{\alpha\beta}:=(\sigma_{\alpha\beta})_{\epsilon=0}=0]$,

$$\Theta^{p}=2l^{p}_{\alpha}l^{p}_{\beta}\left[\frac{1}{3n_0}(\delta N)\gamma^{\alpha\beta}+\gamma^{\alpha\mu}\gamma^{\beta\nu}(\delta\sigma_{\mu\nu})\right], \tag{2.23}$$

the theory based on $\mathbf{E}$ or $\{\delta A\}=\mathbf{E}\cup\{\delta A\}$ alone seems to be *somewhat less complete* than that based on χ, $\mathbf{E}$, and Θ^{p} ($p=1,2,3$):

$$\varphi_4([\mathbf{W}]):=\{\chi,\mathbf{E},\Theta^{p}\}. \tag{2.24}$$

Focus now, for specifity, on the magnetic part $H_{\alpha\beta}$ of the Weyl tensor $C_{\alpha\beta\mu\nu}$. Restricting attention to the Bianchi type I background model, this part of $C_{\alpha\beta\mu\nu}$ vanishes in $(X,q_{\alpha\beta})$ since $(\omega_{\alpha\beta})_{\epsilon=0}=0$ and $(\mathcal{E}_{\alpha\beta\mu})_{\epsilon=0}=0$. By writing $H_{\alpha\beta}$ in the form [see, e.g., Eq. (4.19) in Ref. 33]

$$H_{\alpha\beta}=2a_{(\alpha}\omega_{\beta)}+u^{\nu}\eta^{\lambda\tau}{}_{\nu(\alpha}\mathcal{E}_{\beta)\lambda\tau}+h_{\alpha}{}^{\mu}h_{\beta}{}^{\sigma}u^{\nu}\eta^{\lambda\tau}{}_{\nu(\mu}\omega_{\sigma)\lambda;\tau}, \tag{2.25}$$

where $\eta_{\alpha\beta\mu\nu}$ is the space–time permutation tensor and $\omega^{\alpha}=g^{\alpha\beta}\omega_{\beta}$ is the vorticity vector ($\omega_{\alpha\beta}=\eta_{\alpha\beta\mu\nu}\omega^{\mu}u^{\nu}$), we then see that, *to linear order*, $H_{\alpha\beta}$ is *completely* determined by giving $\delta\omega_{\alpha\beta}$ and $\delta\mathcal{E}_{\alpha\beta\mu}$:

$$\delta H_{\alpha\beta}=\epsilon^{\lambda\tau}{}_{(\alpha}[\delta\mathcal{E}_{\beta)\lambda\tau}+(\delta\omega_{\beta)\lambda})_{|\tau}]. \tag{2.26}$$

Here $\epsilon^{\lambda\tau}{}_{\alpha}$ stands for the background value of $u^{\nu}\eta^{\lambda\tau}{}_{\nu\alpha}$.

To sum up, the notion of a gauge-invariant variable introduced by Banach[1] and Banach and Piekarski[8,9] *generalizes* the notion of a gauge-invariant variable arising from the approach of Stewart and Walker.[2] For the case of nonaxisymmetric cosmologies of Bianchi types I and V ($H_1\neq H_2, H_1\neq H_3, H_2\neq H_3$), it is this generalization which enables us to define satisfactorily the concept of a ''coordinate system'' on $\mathcal{W}/\mathcal{W}_L$ and to replace $\{\delta A\}$ by **D**. Our reasons for presenting the detailed construction of **D** are twofold. First, [**W**] is uniquely determined from **D** and conversely. Second, if $A_0:=(A)_{\epsilon=0}$ satisfies one of the three conditions in the Stewart–Walker lemma, then a knowledge of **D** consequently permits us to calculate δA. The natural question arises as to the possibility of recasting the mathematical formalism in such a way that **D** is obtainable from $\{\delta A\}$. However, we have seen that such a reformulation *is not possible*, i.e., the acceptance of $\{\delta A\}$ as fundamental gauge-invariant quantities in place of **D** involves the sacrifice of some information originally contained in our *complete* set **D**. Thus, despite the desirability of having only $\{\delta A\}$ as gauge-invariant variables, this sacrifice of the informational content of the formalism would be unacceptable, as far as the basic structure of the theory and the equivalence classes of perturbations are concerned.

B. Almost-Robertson–Walker universe models

In exact Robertson–Walker space–times ($k=0,\pm 1$), the shear tensor $d_{\alpha\beta}:=(\sigma_{\alpha\beta})_{\epsilon=0}$ vanishes identically, so that $\delta\sigma_{\alpha\beta}$ is a gauge-invariant quantity. If $d_{\alpha\beta}=0$, the formula (2.23) may be used to express Θ^{p} ($p=1,2,3$) in terms of δN and $\delta\sigma_{\alpha\beta}$, and from Eqs. (4.8d), (4.8g), (4.8h), and (4.18h) of Ref. 1 we obtain directly

$$\Omega^{\alpha\beta}=\frac{2}{n_0}\left[\frac{1}{3}(\delta N)^{\cdot}+2H(\delta N)\right]\gamma^{\alpha\beta}+2[(\delta\sigma_{\mu\nu})^{\cdot}+3H(\delta\sigma_{\mu\nu})]\gamma^{\mu\alpha}\gamma^{\nu\beta}, \tag{2.27a}$$

$$\Theta^{pq}_{q}=2l^{\nu}_{q}[-(l^{q}_{\lambda}l^{p}_{\mu})_{|\nu}(\delta\sigma_{\epsilon\tau})+l^{q}_{\lambda}l^{p}_{\mu}(\delta\sigma_{\epsilon\tau})_{|\nu}]\gamma^{\epsilon\mu}\gamma^{\tau\lambda}, \tag{2.27b}$$

$$\Omega^{pq}_{r}=2l^{p}_{\lambda}l^{q}_{\mu}\gamma^{\mu\epsilon}\gamma^{\lambda\tau}[l^{\nu}_{r}(\delta\sigma_{\epsilon\tau})]_{|\nu}-4l^{\nu}_{r|\lambda}l^{p}_{(\mu}l^{q}_{\nu)}\left[\frac{1}{3n_0}(\delta N)\gamma^{\mu\lambda}+\gamma^{\mu\epsilon}\gamma^{\lambda\tau}(\delta\sigma_{\epsilon\tau})\right], \tag{2.27c}$$

$$\Omega^{\alpha\beta}_{\mu}=2\left[\frac{1}{3n_0}(\delta N)\gamma^{\alpha\beta}+\gamma^{\alpha\lambda}\gamma^{\beta\tau}(\delta\sigma_{\lambda\tau})\right]_{|\mu}. \tag{2.27d}$$

These results are the statement that a knowledge of gauge-invariant variables δN and $\delta\sigma_{\alpha\beta}$ determines $\Omega^{\alpha\beta}$, Θ^{p}, Θ^{pq}_{q}, Ω^{pq}_{r}, and $\Omega^{\alpha\beta}_{\mu}$. Thus, for almost-Robertson–Walker universe models, we need not introduce the gauge-invariant objects $\Omega^{\alpha\beta}$, Θ^{p}, Θ^{pq}_{q}, Ω^{pq}_{r}, and $\Omega^{\alpha\beta}_{\mu}$ to represent [**W**], since the following set, much simpler than that appropriate for the nonaxisymmetric case [see Eqs. (2.16) and (2.18)], is found to give such a representation:[4,5]

$$\varphi_{\mathrm{RW}}([\mathbf{W}]):=\{\chi,\delta N,\delta s,\delta\mathcal{B},\delta\mathcal{A},\delta X^{\alpha},\delta\sigma_{\alpha\beta},\delta\mathcal{A}^{\alpha\beta}\}. \tag{2.28}$$

To the extent that $\{\mathcal{F}_{\epsilon}(\cdot);\epsilon\in I\}$ is an arbitrary curve in the function space $\{\mathcal{F}(\cdot)\}$ passing through the background solution $\mathcal{F}_0(\cdot)$ (i.e., that $\chi=0$ and $\delta N\neq 0$), this set consists of 17 ''geometrically'' independent, not identically vanishing gauge-invariant components.

As demonstrated in Sec. 4.2 of Ref. 5, any gauge-invariant quantity is obtainable linearly from $\varphi_{\mathrm{RW}}([\mathbf{W}])$ through purely algebraic and differential operations. To illustrate this property of $\varphi_{\mathrm{RW}}([\mathbf{W}])$, let $E_{\alpha\beta}$ be the electric part of the Weyl tensor. Then $(E_{\alpha\beta})_{\epsilon=0}=0$ and the linear perturbation $\delta E_{\alpha\beta}$ is gauge invariant. Appeal to the identity

$$E_{\alpha\beta}=\tfrac{1}{3}\theta\sigma_{\alpha\beta}-\sigma_{\alpha\mu}\sigma^{\mu}{}_{\beta}+\tfrac{1}{3}h_{\alpha\beta}(\sigma^{\mu\nu}\sigma_{\mu\nu})-\tfrac{1}{2}[R_{\mu\nu}h^{\mu}{}_{\alpha}h^{\nu}{}_{\beta}-\tfrac{1}{3}h_{\alpha\beta}(R_{\mu\nu}h^{\mu\nu})]+n^{2/3}h_{\alpha\mu}h_{\beta\nu}\mathcal{A}^{\mu\nu} \tag{2.29}$$

shows that

$$\delta E_{\alpha\beta}=H(\delta\sigma_{\alpha\beta})+(n_0)^{2/3}\gamma_{\alpha\mu}\gamma_{\beta\nu}(\delta\mathcal{A}^{\mu\nu})-\tfrac{1}{2}\,\delta[R_{\mu\nu}h^{\mu}{}_{\alpha}h^{\nu}{}_{\beta}-\tfrac{1}{3}h_{\alpha\beta}(R_{\mu\nu}h^{\mu\nu})], \tag{2.30}$$

where

$$\begin{aligned}\delta\left[R_{\mu\nu}h^{\mu}{}_{\alpha}h^{\nu}{}_{\beta}-\frac{1}{3}h_{\alpha\beta}(R_{\mu\nu}h^{\mu\nu})\right]=&(n_0)^{2/3}[\gamma_{\alpha\mu}\gamma_{\beta\nu}(\delta\mathcal{A}^{\mu\nu})]+(\delta\sigma_{\alpha\beta})^{\cdot}+3H(\delta\sigma_{\alpha\beta})\\&+\frac{1}{3Hn_0}\gamma_{\mu(\alpha}[(\delta X^{\mu})^{\cdot}]_{|\beta)}-\frac{1}{9Hn_0}\gamma_{\alpha\beta}[(\delta X^{\mu})^{\cdot}]_{|\mu}\\&+\frac{1}{3n_0}\left(4-\frac{\dot H}{H^2}\right)\gamma_{\mu(\alpha}(\delta X^{\mu})_{|\beta)}-\frac{1}{9n_0}\left(4-\frac{\dot H}{H^2}\right)\gamma_{\alpha\beta}(\delta X^{\mu})_{|\mu}\\&-\frac{1}{12H^2}(n_0)^{2/3}\left\{[(\delta\mathcal{B})_{|\alpha}]_{|\beta}-\frac{1}{3}\gamma_{\alpha\beta}\gamma^{\mu\nu}[(\delta\mathcal{B})_{|\mu}]_{|\nu}\right\}\\&-\frac{1}{3Hn_0}\left\{[(\delta N)_{|\alpha}]_{|\beta}-\frac{1}{3}\gamma_{\alpha\beta}\gamma^{\mu\nu}[(\delta N)_{|\mu}]_{|\nu}\right\}\\&+\frac{1}{6H^2}n_0T_0\left\{[(\delta s)_{|\alpha}]_{|\beta}-\frac{1}{3}\gamma_{\alpha\beta}\gamma^{\mu\nu}[(\delta s)_{|\mu}]_{|\nu}\right\}.\end{aligned} \tag{2.31}$$

Obviously, Eqs. (2.30) and (2.31) give a relationship for $\delta E_{\alpha\beta}$ in terms of $\varphi_{\mathrm{RW}}([\mathbf{W}])$. The extra assumption which can be made here is that $\delta\mathcal{F}_0$ satisfies the condition (2.14) of Ref. 1. Under these circumstances, the left-hand side of Eq. (2.31) equals zero and $\delta E_{\alpha\beta}$ is given by

$$\delta E_{\alpha\beta}=H(\delta\sigma_{\alpha\beta})+(n_0)^{2/3}\gamma_{\alpha\mu}\gamma_{\beta\nu}(\delta\mathcal{A}^{\mu\nu}). \tag{2.32}$$

The treatment of other gauge-invariant variables follows similar lines.

Since $\varphi_{\mathrm{RW}}([\mathbf{W}])$ is a subset of $\{\delta A\}$, it should be clear from our discussion above that the only essential structure of linear perturbation theory used in the definition of almost-Robertson–Walker universe models ($d_{\alpha\beta}=0$) was Lemma 2.2 of Stewart and Walker.[2] However, in more general cases—such as infinitesimal perturbations of anisotropic cosmological models ($d_{\alpha\beta}\neq 0$)—this lemma does not provide a complete framework for describing the equivalence classes of perturbations or for constructing all gauge-invariant quantities, and the basic concepts, notions, and ideas are qualitatively different (see, e.g., Refs. 8 and 9).

III. ANALYSIS OF OTHER APPROACHES

A. Covariant methods

Within the framework of a covariant formalism, the perturbations of homogeneous and isotropic cosmological models have been studied by several authors (Ellis and Bruni,[6] Ellis *et al.*,[31] Bruni *et al.*,[29] and Dunsby *et al.*[30]). Beginning from Eq. (2.2) and the Stewart–Walker lemma, the basic idea of these authors was to introduce a set $\{A\}$ of geometrically defined exact variables A such that $\mathcal{L}_v A_0=0$ for each vector field v on X. This definition of $\{A\}$ is relatively broad and includes, for example, studies of the meaning of $\mathcal{L}_v A_0 = 0$ when $A_0 \neq 0$. However, in practice, references to $\mathcal{L}_0 A_0$ usually include the assumption that A_0 *vanishes* in a Robertson–Walker universe model. Understood in this more limited sense, the quantity A itself is a gauge-invariant perturbation in an almost-Robertson–Walker universe, and its physical significance is apparent through the covariant definition. As in the work of Bruni *et al.*,[29] we shall refer to such quantities as *covariant gauge-invariant variables*.

The purpose of Dunsby[23] was to further extend this approach to perturbations of Bianchi type I cosmological models. His conclusion that it should be possible to give a more general argument for the utility of a covariant formalism relies on a number of observations: (i) that the Stewart–Walker lemma is valid for any background space-time; (ii) that one can often consider the same gauge-invariant variables in perturbing different universe models; (iii) that covariant gauge-invariant variables can be easily identified in perturbing homogeneous anisotropic space–times; (iv) that the differential equations governing these variables provide a unified treatment for the exact and the linearized theory; and (v) that the covariant extension to Bianchi type V is very straightforward. There may also be other important observations that we have failed to notice.

How do our concepts relate to the covariant method? If A is a gauge-invariant quantity with respect to homogeneous background space-times (i.e., if A has the property that $A_0 = 0$), then from a knowledge of $\varphi_{\mathrm{RW}}([\mathbf{W}])$ or $\varphi([\mathbf{W}])$ alone, together with a specification of $\mathcal{F}_0$, one can compute A approximately; the approximation takes place by neglecting higher-order terms in the exact expression for A $[A \cong \epsilon(\delta A)]$. As regards the details, the considerations of Secs. II A and II B were partly devoted to a systematic first-order expansion of the covariant gauge-invariant variables A $(A_0 = 0)$.

We see two main disadvantages in following this covariant approach. First, even in the case of Robertson–Walker background models, the covariant gauge-invariant variables do not provide a complete framework for determining the equivalence classes of perturbations and for identifying all gauge-invariant quantities. Specifically, by means of the definition (2.28) of $\varphi_{\mathrm{RW}}([\mathbf{W}])$, we obtain $s_0 \neq 0$ and $(n_0)^{2/3}\mathcal{A}_0 = (n_0)^{2/3}\mathcal{B}_0 = 6k/R^2$ $(k = 0, \pm 1)$. Second, the construction of Refs. 23 and 29–31 has the feature that it appears to depend on our choice of $(X, q_{\alpha\beta})$, and—as already remarked—in the absence of three-dimensional homogeneous hypersurfaces, simply no natural, ''preferred'' choice of covariant gauge-invariant variables is available. This poses a potential serious difficulty for the formulation of linear perturbation theory in general. Fortunately, given the alternative construction of Banach and Piekarski,[8,9] this difficulty can be resolved by modifying the definition of a gauge-invariant variable and by formulating the theory via the introduction of **D**.

B. Harmonic decompositions for an almost-Bianchi type I universe model

In a spatially flat anisotropic universe, a completely general perturbation of the gravitational field can be written as a linear combination of perturbations associated with individual spatial harmonics (for review see Refs. 10–15). The metric perturbations and the matter perturbations are classified into three types, i.e., scalar, vector, and tensor according to the properties of infinitesimal gauge transformations. For example, a scalar representation of the perturbation in the contravariant metric tensor is given by[1]

$$Q = Q_0 \mathcal{H}, \quad Q^\alpha = Q_L \mathcal{H}^\alpha, \quad D = D_L \mathcal{H}, \tag{3.1a}$$

$$F^{\alpha\beta} = \frac{1}{3} F_T \gamma^{\alpha\beta} \mathcal{H} - \frac{1}{n} F_T \gamma^{\mu(\alpha} \mathcal{H}^{\beta)}{}_{|\mu}, \tag{3.1b}$$

where Q_0, Q_L, D_L, and F_T are the expansion coefficients (amplitudes) whose covariant spatial gradients vanish. Explaining Eqs. (3.1) still further, n is a wave number (not to be confused with the number density, also denoted by n) which sets the spatial scale of the perturbation ($\dot{n} \neq 0, n_{|\alpha}=0$), $\mathcal{H}$ is a scalar harmonic independent of time ($\dot{\mathcal{H}}=0$), and $\mathcal{H}^\alpha$ is a spatial vector related to $\mathcal{H}$ by

$$\mathcal{H}^\alpha := -\frac{1}{n}\gamma^{\alpha\beta}\mathcal{H}_{|\beta}. \tag{3.2}$$

We can take the scalar harmonics as being eigenfunctions of the covariantly defined Laplace–Beltrami operator:[34]

$$\gamma^{\alpha\beta}[(\mathcal{H}_{|\alpha})_{|\beta}] = -n^2\mathcal{H}. \tag{3.3}$$

For zero background curvature ($k=0$), $\mathcal{H}$ is the usual plane wave.[12] Another remark is also in order. Since we study only the linear perturbations, there is no coupling between different wave numbers. Consequently, we drop the eigenvalue index n from the $\mathcal{H}$'s and omit the symbol Σ_n that could actually be a summation over a discrete set or an integral over a continuously varying index.[29]

The variations of expansion coefficients in Eqs. (3.1) under the gauge transformations $\mathbf{W} \to \mathbf{W}' := \mathbf{W} + L_v\mathcal{F}_0$ are obtained if we decompose the vector field v and the infinitesimal perturbation $\mathbf{W}'$ harmonically. Writing $\vartheta := -w_\alpha v^\alpha$ in the form $\vartheta = \vartheta_0\mathcal{H}$ ($\vartheta_{0|\alpha}=0$, $\dot{\vartheta}_0 \neq 0$) and denoting by $\vartheta_L\mathcal{H}^\alpha$ the scalar part of $\vartheta^\alpha := \gamma^\alpha{}_\beta v^\beta$ ($\vartheta_{L|\alpha}=0$, $\dot{\vartheta}_L \neq 0$), we conclude that the final result for the changes in the amplitudes of the metric perturbations is

$$Q_0' = Q_0 + 2\dot{\vartheta}_0, \tag{3.4a}$$

$$Q_L' = Q_L + n\vartheta_0 + \dot{\vartheta}_L + \frac{\dot{\Delta}}{2\Delta}\vartheta_L, \tag{3.4b}$$

$$D_L' = D_L - H\vartheta_0 - \frac{n}{3}\vartheta_L, \tag{3.4c}$$

$$F_T' = F_T - \frac{3}{2}\left(H + \frac{\dot{\Delta}}{2\Delta}\right)\vartheta_0 + n\vartheta_L, \tag{3.4d}$$

where $\Delta := n^2$.

For the scalar type, only two independent gauge-invariant quantities can be constructed from the metric tensor amplitudes alone, since there are two gauge functions and four metric tensor amplitudes. In our covariant notation, by inspection of Eqs. (3.4), these are conveniently taken as

$$\Phi_1 := \frac{1}{2}Q_0 + 2\left[\left(3H + \frac{\dot{\Delta}}{2\Delta}\right)^{-1}\left(D_L + \frac{1}{3}F_T\right)\right]^{\cdot}, \tag{3.5a}$$

$$\Phi_2 := -Q_L + \frac{1}{n}\dot{F}_T - 2n\left(3H + \frac{\dot{\Delta}}{2\Delta}\right)^{-1}\left(D_L + \frac{1}{3}F_T\right) - \frac{3}{n}\left[\left(3H + \frac{\dot{\Delta}}{2\Delta}\right)^{-1}\left(D_L + \frac{1}{3}F_T\right)\left(H + \frac{\dot{\Delta}}{2\Delta}\right)\right]^{\cdot}. \tag{3.5b}$$

The recognition that Φ_1 and Φ_2 are important in the context of an almost-Bianchi type I universe model goes back at least to Den,[12] and Eqs. (3.5a) and (3.5b) are equivalent to the central formulas (3.7) and (3.8) of his paper. We also define Φ by

$$\Phi := \frac{1}{n} H \Phi_2 . \tag{3.6}$$

This quantity is directly associated with the perturbation of the intrinsic curvature of a zero-shear hypersurface (i.e., Φ_H in Bardeen's paper[10]) in isotropic models:

$$d_{\alpha\beta} = 0 \Rightarrow \dot{\Delta} = -2H\Delta . \tag{3.7}$$

As emphasized in the Introduction (see also Ref. 1), there are a number of important *prima facie* questions that can be asked of any gauge-invariant formulation of linear perturbation theory. In the context of a harmonic decomposition of $\mathbf{W}$, the following question is particularly relevant: At what stage in the procedure is it ''necessary'' to make a harmonic decomposition corresponding to that made by Bardeen[10] and Den?[12] In particular, must this be done before defining a set $\mathbf{A}$ of gauge-invariant amplitudes, or can the objects[1] $\overline{A}(x,\mathbf{W})$ with a property $\overline{A}(x,\mathbf{W}) = \overline{A}(x,\mathbf{W} + L_v \mathcal{F}_0)$ be constructed first and then a notion of gauge-invariant amplitudes extracted afterwards? This is closely related to the question of whether the key quantities Φ_1 and Φ_2 are to be introduced before, or after, effectively describing the equivalence classes of perturbations. None of the previous approaches to this problem has yielded an argument which entirely carries conviction, although Ellis and Matravers[16] claim that the splitting of gauge-dependent quantities into scalar, vector, and tensor parts is of heuristic value only.

With the foregoing observations to guide us, we now seek an interpretation of how Den's variables Φ_1 and Φ_2 adapt to our formalism, and to this end introduce the following scalar representation of Ω, Ω^α, S, and $\vec{\gamma}\circ\vec{\Theta}$:

$$\Omega = \Omega_0 \mathcal{H}, \quad \Omega^\alpha = \Omega_1 \mathcal{H}^\alpha, \tag{3.8a}$$

$$S = S_0 \mathcal{H}, \quad \vec{\gamma}\circ\vec{\Theta} = \Theta_0 \mathcal{H}. \tag{3.8b}$$

It is important to stress that since Ω, Ω^α, S, and $\vec{\gamma}\circ\vec{\Theta}$ are gauge-invariant quantities, the amplitudes Ω_0, Ω_1, S_0, and Θ_0 are *automatically* gauge invariant. Thus the objections raised to harmonic decompositions by Ellis and Matravers,[16] that ''there is considerable effort going on at present into distinguishing if anisotropies are due to scalar, vector, or tensor modes; but this is not a fully covariant and gauge-invariant activity,'' do not apply here. Using the definitions of Ω, Ω^α, S, and $\vec{\gamma}\circ\vec{\Theta}$ [see Eq. (2.13d) and Eqs. (4.8) and (4.11a) of Ref. 1] and analyzing the gauge-dependent objects[1] Q, Q^α, D, $F^{\alpha\beta}$, V^α, and M harmonically, it is illustrative to express Φ_1 and Φ_2 in terms of Ω_0, Ω_1, S_0, and Θ_0. After a fair amount of algebraic calculation, we then obtain

$$\Phi_1 = -\Omega_0 + \left[\frac{1}{\Delta} \left(3H + \frac{\dot{\Delta}}{2\Delta} \right)^{-1} S_0 \right]^{\cdot} , \tag{3.9a}$$

$$\Phi_2 = \frac{1}{3H} \Omega_1 + \frac{1}{2n} \Theta_0 + \frac{3}{n} \left[\frac{H}{\Delta} \left(3H + \frac{\dot{\Delta}}{2\Delta} \right)^{-1} S_0 \right]^{\cdot} - \frac{1}{n} \left(3H + \frac{\dot{\Delta}}{2\Delta} \right)^{-1} S_0 . \tag{3.9b}$$

At first sight, these results seem highly mysterious, for how could one extract Φ_1 and Φ_2 from Ω_0, Ω_1, S_0, and Θ_0 in a unique way if Φ_1 and Φ_2 depend on the metric tensor amplitudes alone? The key is the general observation of Sec. II, showing that $\varphi([\mathbf{W}])$ *determines everything*.

In the case of other gauge-invariant amplitudes [see, e.g., Eqs. (3.9) and (3.10) in Ref. 12 or Eqs. (30)–(39) in Ref. 14], the procedure will be the same. First, separate $\varphi([\mathbf{W}])$ into scalar, vector, and tensor parts: this will define a complete set $\mathbf{A}_C$ of gauge-invariant amplitudes. Second, following the method outlined in Refs. 10–15, construct another set of gauge-invariant amplitudes that describes what one is interested in (e.g., energy density perturbation amplitudes, velocity perturbation amplitudes, etc.). Finally, show that this set is obtainable linearly from $\mathbf{A}_C$ through purely algebraic and differential operations.

It is clear from the above analysis that the existence of a set $\mathbf{A}$ of gauge-invariant amplitudes ($\mathbf{A}_C \subset \mathbf{A}$) is an immediate consequence of the existence of $\varphi([\mathbf{W}])$. Because of this, the issue of how to interpret decompositions like (3.1) is irrelevant to the geometric (i.e., intrinsic) definition of an almost-Bianchi type I universe. Put somewhat differently, the general theory of gauge-invariant perturbations[8,9] is a theory of *fields*, not amplitudes. Although in appropriate circumstances harmonic functions may be available and useful,[35] they play no fundamental role in either the formulation or interpretation of the theory.

IV. APPLICATION OF *D* TO QUANTUM FIELD THEORY IN CURVED SPACE–TIME

Among the issues that can be studied systematically *only* with this sort of approach, the examination of the effect of using a *semiclassical description* in which the background geometry is taken in the classical framework and the gauge-invariant perturbations are considered as *quantum variables* presents a most intriguing challenge. Clearly, there are many ways of performing this task, but a very natural way consists in applying the methods of symplectic geometry and geometric quantization.[18] For a Lagrangian formulation of covariant field theories,[17] the important object is a ''presymplectic form'' $\mathcal{F}_0 \mapsto \overline{\omega}(\mathcal{F}_0|\cdot,\cdot)$ defined on the space $\mathcal{G}$ of field configurations $\mathcal{F}_0$ on X. Such a presymplectic form can be used to construct a real-valued, bilinear functional of two ''infinitesimal perturbations'' $\delta\mathcal{F}_0 \in \mathcal{P}$ and $\delta\mathcal{F}_0' \in \mathcal{P}$ of $\mathcal{F}_0$, denoted $\overline{\omega}(\mathcal{F}_0|\delta\mathcal{F}_0,\delta\mathcal{F}_0')$. This functional satisfies the property that when $\mathcal{F}_0 \in \mathcal{G}$ is a solution to the nonlinear field equations and $\delta\mathcal{F}_0$ and $\delta\mathcal{F}_0'$ solve the linearized field equations, then $\overline{\omega}(\mathcal{F}_0|\delta\mathcal{F}_0,\delta\mathcal{F}_0')$ is gauge invariant, i.e., we have

$$\overline{\omega}(\mathcal{F}_0|\delta\mathcal{F}_0+\mathcal{L}_v\mathcal{F}_0,\delta\mathcal{F}_0'+\mathcal{L}_{v'}\mathcal{F}_0)=\overline{\omega}(\mathcal{F}_0|\delta\mathcal{F}_0,\delta\mathcal{F}_0'), \tag{4.1}$$

where v and v' are two arbitrary[36] vector fields on X. In other words, the two-form $\overline{\omega}(\cdot|\cdot,\cdot)$ fails to be a symplectic form on $\mathcal{G}_S \subset \mathcal{G}$, the space of solutions to the nonlinear field equations, because it is *degenerate*; equivalently, for each $\mathcal{F}_0 \in \mathcal{G}_S$, the set $\mathcal{P}_S \subset \mathcal{P}$ consisting of classical solutions to Eq. (2.14) of Ref. 1 is unsuitable to serve as phase space of linear perturbation theory because it is ''too large.''

However, with the help of a complete set $\mathbf{D}$ of basic gauge-invariant variables, we can try to prove that $\overline{\omega}(\mathcal{F}_0|\delta\mathcal{F}_0,\delta\mathcal{F}_0')$ depends only on[1] $\mathbf{D} := \varphi([J(\delta\mathcal{F}_0)])$ and $\mathbf{D}' := \varphi([J(\delta\mathcal{F}_0')])$, i.e., that there exists a bilinear functional $\omega(\mathcal{F}_0|\mathbf{D},\mathbf{D}')$ of $\mathbf{D} \in \mathcal{D}_S \subset \mathcal{D}$ and $\mathbf{D}' \in \mathcal{D}_S \subset \mathcal{D}$ related to $\overline{\omega}(\mathcal{F}_0|\delta\mathcal{F}_0,\delta\mathcal{F}_0')$ by

$$\omega(\mathcal{F}_0|\mathbf{D},\mathbf{D}')=\overline{\omega}(\mathcal{F}_0|\delta\mathcal{F}_0,\delta\mathcal{F}_0'). \tag{4.2}$$

If this reduction process gives rise to a *symplectic structure* $\omega(\mathcal{F}_0|\cdot,\cdot): \mathcal{D}_S \times \mathcal{D}_S \to \mathbb{R}$ via $(\mathbf{D},\mathbf{D}') \mapsto \omega(\mathcal{F}_0|\mathbf{D},\mathbf{D}')$, it will be possible to find a quantum theory in which the functions $\omega(\mathcal{F}_0|\mathbf{D},\cdot)$ on the ''classical phase space''

$$\mathcal{D}_S := \{\varphi([J(\delta\mathcal{F}_0)]); \quad \delta\mathcal{F}_0 \in \mathcal{P}_S\} \tag{4.3}$$

are represented (irreducibly) by operators $\hat{\omega}(\mathcal{F}_0|\mathbf{D},\cdot)$ satisfying the following commutation relations (see Ref. 19, p. 37):

$$[\hat{\omega}(\mathcal{F}_0|\mathbf{D},\cdot),\hat{\omega}(\mathcal{F}_0|\mathbf{D}',\cdot)]=-i\,\omega(\mathcal{F}_0|\mathbf{D},\mathbf{D}')\hat{I}, \tag{4.4}$$

where $\hat{I}$ denotes the identity operator and where we choose units so that $\hbar = 1$.

The above discussion has laid out the basic mathematical framework of the geometric formulation of quantum field theory on a background space–time. In attempting to find concrete examples of ω, our idea is to begin with the usual Hilbert Lagrangian density $L_H := \sqrt{-g}(R-2\Lambda)$

and the background metrics $q_{\alpha\beta}$ of constant curvature (de Sitter and anti-de Sitter space–times). However, since the construction of ω for these cases is not immediate, it will be presented in a separate paper.

V. FINAL REMARKS

The notion of a complete set of basic gauge-invariant variables has a number of independent connotations. One such connotation is that of being a representative of $[\mathbf{W}]$, the equivalence class of perturbations. In Sec. IV B of Ref. 1, we denoted this representative of $[\mathbf{W}]$ by $\varphi([\mathbf{W}])$ [see Eq. (4.15)]. Another connotation of the term ''complete set'' is one where the appropriate local algebropartial differential operations on $\varphi([\mathbf{W}])$ enable us to obtain all gauge-invariant quantities. There is yet a third connotation of this term arising from consideration of the dynamic equations governing linearized perturbations. As a matter of fact, if $\mathbf{W}$ satisfies the condition (2.19) of Ref. 1, a deterministic system of propagation equations can be derived that involves only $\varphi([\mathbf{W}])$. Conversely, if $\mathbf{D}$ is a classical solution to these propagation equations, one will be able to construct $\mathbf{W}$ which satisfies the condition (2.19) of Ref. 1 and is such that $\mathbf{D}=\varphi([\mathbf{W}])$.

Mathematically, this approach to the gauge problem is a nontrivial example of the general scheme that we have described in our two previous papers,[8,9] and the new concepts developed were also applied to the *explicit* definition of a complete set $\mathbf{D}$ of basic gauge-invariant variables for the case of an almost-Robertson–Walker universe model.[4,5] Details of the description of a gas of massive collisionless particles are given in Ref. 4, where we study the gauge problem in a broader context, i.e., for different general-relativistic models such as the Einstein–Liouville system. We can also obtain (see Sec. 5 in Ref. 9) an analytical form of $\mathbf{D}$ for the infinitesimal perturbation of the metric tensor itself (pure gravity) defined on a fixed background de Sitter space–time and obeying the linearized empty-space Einstein equations with non-negative cosmological constant Λ; the case $\Lambda=0$ corresponds to linear perturbation theory in Minkowski space–time.

On these grounds, it was possible to show that when the Stewart–Walker lemma[2] does not provide a completely satisfactory algorithm to describe $[\mathbf{W}]$ in terms of $\{\delta A\}$, this difficulty can be cured by working instead with $\mathbf{D}$. The construction of Ref. 1 gives rise to a significant enlargement of the class of gauge-invariant quantities admitted by the theory. Because of this, our method not only *generalizes* the covariant formalism of Ellis and Bruni,[6] but also *explains* in a fully covariant and gauge-invariant manner what Bardeen's major paper[10] is about. Moreover, the existence of $\mathbf{D}$ fits in well with the Hájíček–Isham view[21] of quantum field theory that *only* gauge-invariant objects should be quantized. Finally, it appears that our main ideas are broadly modifiable. For example, we can introduce the notion of an ''infinitesimal gauge-invariant variable'' for all Lagrangian field theories with local symmetries. Roughly speaking,[17] a local symmetry is a field variation on space–time about the field configuration $\mathcal{F}$—such as the gauge transformations of Yang–Mills theory or the diffeomorphisms of general relativity—that keeps the action $S[\mathcal{F}]$ invariant [see Eq. (2.1) in Ref. 17] and that is ''local'' in a suitable sense. This definition is certainly required for the deepest understanding of the gauge problem, and would be expected to be capable of yielding a natural generalization of the notion of a complete set $\mathbf{D}$ previously obtained ''only'' for the case of diffeomorphism-invariant, covariant field theories.[8,9]

To sum, the importance of $\mathbf{D}$ cannot be overemphasized. The information contained in a complete set is all-inclusive—it is equivalent to specifying all equivalence classes of perturbations, to constructing all gauge-invariant variables, and to obtaining all imaginable types of presentations of linear perturbation theory. If the set of basic gauge-invariant variables is known, there remains not a single perturbational attribute that is not completely and precisely determined.

ACKNOWLEDGMENTS

I am grateful to Dr. S. Piekarski for his continuous interest in my work and many discussions. Also, I would like to thank the anonymous referee for a critical reading of the original manuscript and valuable suggestions on clarifying the exposition.

[1] Z. Banach, J. Math. Phys. **40**, 3978 (1999).
[2] J. M. Stewart and M. Walker, Proc. R. Soc. London, Ser. A **341**, 49 (1974).
[3] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984).
[4] Z. Banach and S. Piekarski, Gen. Relativ. Gravit. **28**, 1335 (1996).
[5] Z. Banach and S. Piekarski, Ann. Inst. Henri Poincaré: Phys. Theor. **65**, 273 (1996).
[6] G. F. R. Ellis and M. Bruni, Phys. Rev. D **40**, 1804 (1989).
[7] We define these variables in Sec. III A and in the Introduction of Ref. 1.
[8] Z. Banach and S. Piekarski, Int. J. Theor. Phys. **36**, 1787 (1997).
[9] Z. Banach and S. Piekarski, Int. J. Theor. Phys. **36**, 1817 (1997).
[10] J. M. Bardeen, Phys. Rev. D **22**, 1882 (1980).
[11] K. Tomita and M. Den, Phys. Rev. D **34**, 3570 (1986).
[12] M. Den, Prog. Theor. Phys. **77**, 653 (1987).
[13] M. Den, Prog. Theor. Phys. **79**, 1110 (1988).
[14] H. Noh and J. Hwang, Phys. Rev. D **52**, 1970 (1995).
[15] H. Noh and J. Hwang, Phys. Rev. D **52**, 5643 (1995).
[16] G. F. R. Ellis and D. R. Matravers, Gen. Relativ. Gravit. **27**, 777 (1995).
[17] J. Lee and R. M. Wald, J. Math. Phys. **31**, 725 (1990).
[18] N. M. J. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1991), 2nd ed.
[19] R. M. Wald, *Quantum Field Theory in Curved Space-Time and Black Hole Thermodynamics* (University of Chicago Press, Chicago, 1994).
[20] P. Hájíček and C. J. Isham, J. Math. Phys. **37**, 3505 (1996).
[21] P. Hájíček and C. J. Isham, J. Math. Phys. **37**, 3522 (1996).
[22] This reduction process generally involves the removal of gauge freedom.
[23] P. K. S. Dunsby, Phys. Rev. D **48**, 3562 (1993).
[24] J. Ehlers, in *Relativity, Astrophysics, and Cosmology*, edited by W. Israel (Reidel, Boston, 1973), Vol. 38, p. 1.
[25] J. M. Stewart, Class. Quantum Grav. **7**, 1169 (1990).
[26] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds, and Physics. Part I: Basics* (North-Holland, Amsterdam, 1989).
[27] L. Smarr, C. Taubes, and J. R. Wilson, in *Essays in General Relativity*, edited by F. J. Tipler (Academic, New York, 1980), p. 157.
[28] Z. Banach and S. Piekarski, Int. J. Theor. Phys. **35**, 665 (1996).
[29] M. Bruni, P. K. S. Dunsby, and G. F. R. Ellis, Astrophys. J. **395**, 34 (1992).
[30] P. K. S. Dunsby, M. Bruni, and G. F. R. Ellis, Astrophys. J. **395**, 54 (1992).
[31] G. F. R. Ellis, M. Bruni, and J. Hwang, Phys. Rev. D **42**, 1035 (1990).
[32] As noted already in Sec. 5.2 of Ref. 28, the expression (2.14b) for δs is a consequence of the first law of thermodynamics.
[33] G. F. R. Ellis, in *General Relativity and Cosmology*, Proceedings of XLVII Enrico Fermi Summer School, edited by R. K. Sachs (Academic, New York, 1971), p. 104.
[34] S. W. Hawking, Astrophys. J. **145**, 544 (1966).
[35] If the separation into individual harmonics can be made, then the quasi-linear, partial differential equations for $\mathbf{D}$ become ordinary differential equations in time for $\mathbf{A}_C$.
[36] Precisely speaking, v and v' must be of class C^r (r sufficiently large); otherwise $\mathcal{L}_v\mathcal{F}_0$ and $\mathcal{L}_{v'}\mathcal{F}_0$ cannot be classical solutions of Eq. (2.14) in Ref. 1.

Gödel-type space–times in induced matter gravity theory

H. L. Carrion,[a] M. J. Rebouças,[b] and A. F. F. Teixeira[c]
Centro Brasileiro de Pesquisas Físicas, Departamento de Relatividade e Partículas, Rua Dr. Xavier Sigaud 150, 22290-180 Rio de Janeiro—RJ, Brazil

(Received 2 March 1999; accepted for publication 26 April 1999)

Five-dimensional (5D) generalized Gödel-type manifolds are examined in the light of the equivalence problem techniques, as formulated by Cartan. The necessary and sufficient conditions for local homogeneity of these 5D manifolds are derived. The local equivalence of these homogeneous Riemannian manifolds is studied. It is found that they are characterized by three essential parameters k, m^2, and ω: identical triads (k, m^2, ω) correspond to locally equivalent 5D manifolds. An irreducible set of isometrically nonequivalent 5D locally homogeneous Riemannian generalized Gödel-type metrics are exhibited. A classification of these manifolds based on the essential parameters is presented, and the Killing vector fields as well as the corresponding Lie algebra of each class are determined. It is shown that the generalized Gödel-type 5D manifolds admit maximal group of isometry G_r with $r=7$, $r=9$, or $r=15$ depending on the essential parameters k, m^2, and ω. The breakdown of causality in all these classes of homogeneous Gödel-type manifolds are also examined. It is found that in three out of the six irreducible classes the causality can be violated. The unique generalized Gödel-type solution of the induced matter (IM) field equations is found. The question as to whether the induced matter version of general relativity is an effective therapy for these types of causal anomalies of general relativity is also discussed in connection with a recent work by Romero, Tavakol, and Zalaletdinov.
[S0022-2488(99)00108-5]

I. INTRODUCTION

The field equations of the general relativity theory, which in the usual notation are written in the form

$$G_{\alpha\beta}=\kappa T_{\alpha\beta}, \tag{1.1}$$

relate the geometry of the space–time to its source. The general relativity theory, however, does not prescribe the various forms of matter, and takes over the energy-momentum tensor $T_{\alpha\beta}$ from other branches of physics. In this sense, general relativity (GR) is not a closed theory. The separation between the gravitational field and its source has been often considered as one undesirable feature of GR.[1–3]

Recently, Wesson and co-workers[4,5] have introduced a new approach to GR in which the matter and its role in the determination of the space–time geometry are given from a purely five-dimensional geometrical point of view. In their five-dimensional (5D) version of general relativity the field equations are given by

$$\hat{G}_{AB}=0. \tag{1.2}$$

[a]Electronic mail: lenyj@cbpf.br
[b]Electronic mail: reboucas@cbpf.br
[c]Electronic mail: teixeira@cbpf.br

Henceforth, the five-dimensional geometrical objects are denoted by overhats, and Latin letters are 5D indices and run from 0 to 4. In this new approach to GR the 5D vacuum field equations (1.2) give rise to both curvature and matter in four dimensions. Indeed, it can be shown[5] that it is always possible to rewrite the 15 field equations (1.2) as a set of equations such that 10 of which are precisely Einstein's field equations (1.1) in four dimensions with an *induced* energy-momentum

$$\kappa T_{\alpha\beta}=\frac{\phi_{\alpha;\beta}}{\phi}-\frac{\varepsilon}{2\phi^2}\left\{\frac{\phi^* g^*_{\alpha\beta}}{\phi}-g^{**}_{\alpha\beta}+g^{\gamma\delta}g^*_{\alpha\gamma}g^*_{\beta\delta}-\frac{g^{\gamma\delta}g^*_{\gamma\delta}g^*_{\alpha\beta}}{2}+\frac{g_{\alpha\beta}}{4}[g^{*\gamma\delta}g^*_{\gamma\delta}+(g^{\gamma\delta}g^*_{\gamma\delta})^2]\right\}, \tag{1.3}$$

where the Greek letters denote 4D indices and run from 0 to 3, $g_{44}\equiv\varepsilon\phi^2$ with $\varepsilon=\pm 1$, $\phi_\alpha \equiv\partial\phi/\partial x^\alpha$, a star denotes $\partial/\partial x^4$, and a semicolon denotes the usual 4D covariant derivative. Obviously, the remaining five equations (a wave equation and four conservation laws) are automatically satisfied by any solution of the 5D vacuum equations (1.2). Thus, not only the matter but also its role in the determination of the geometry of the 4D space–time can be considered to have a five-dimensional geometrical origin. This approach unifies the gravitational field with its source (not just with a particular field) within a purely 5D geometrical framework. This 5D version of general relativity is often referred to as induced matter gravity theory (IM gravity theory, for short). The IM theory has become a focus of a recent research field.[6] The basic features of the theory have been explored by Wesson and others,[7–11] whereas the implications for cosmology and astrophysics have been investigated by a number of researchers.[12–32] For a fairly updated list of references on IM gravity theory and related issues we refer the reader to Ref. 6.

In general relativity, the causal structure of 4D space–time has locally the same qualitative nature as the flat space–time of special relativity—causality holds locally. The global question, however, is left open and significant differences can occur. On large scale, the violation of causality is not excluded. Actually, it has long been known that there are solutions to the Einstein field equations which possess causal anomalies in the form of closed timelike curves. The famous solution found by Gödel[33] in 1949 might not be the first, but it certainly is the best known example of a cosmological model which makes it apparent that general relativity, as it is normally formulated, does not exclude the existence of closed timelike world lines, despite its Lorentzian character which leads to the local validity of the causality principle. Owing to its striking properties, Gödel's model has a well-recognized importance and has to a certain extent motivated the investigations on rotating cosmological Gödel-type models and on causal anomalies in the framework of general relativity[34–52] and other theories of gravitation.[53–63]

Two recent articles have been concerned with *five-dimensional* Gödel-type space–times. First in Ref. 64 the main geometrical properties of five-dimensional Riemannian manifolds endowed with a 5D counterpart of the 4D Gödel-type metric of general relativity were investigated. Among several results, an irreducible set of isometrically nonequivalent 5D homogeneous (locally) Gödel-type metrics were exhibited. Therein it was also shown that, apart from the degenerated Gödel-type metric, in all classes of homogeneous Gödel-type geometries there is breakdown of causality. As no use of any particular field equations was made in this first paper, its results hold for any 5D Gödel-type manifolds regardless of the underlying 5D Kaluza–Klein gravity theory. In the second article[65] the classes of 5D Gödel-type space–times discussed in Ref. 64 were investigated from a more physical viewpoint. Particularly, the question was examined as to whether the induced matter theory of gravitation permits the family of noncausal solutions of Gödel-type metrics studied in Ref. 64. It was shown that the IM gravity excludes this class of 5D Gödel-type noncausal geometries as solution to its field equations.

In both articles[64,65] the 5D Gödel-type family of metrics discussed is the simplest 5D class of geometries for which the section $u=$const (u is the extra coordinate) is the 4D Gödel-type metric of general relativity. Actually, the 5D Gödel-type line element of both papers does not depend on the fifth coordinate u, and therefore, as regards the IM theory, a radiationlike equation of state is an underlying assumption of both articles. However, it is well known[6] that the dependence of the

5D metric on the extra coordinate is necessary to ensure that the 5D IM theory permits the induction of matter of a very general type in four dimensions.

In this work, on the one hand, we shall examine the main geometrical properties of a class of *generalized* Gödel-type geometries in which the 5D metric depends on the fifth coordinate, generalizing therefore the results found in Ref. 64. On the other hand, we shall also investigate the question as to whether the induced matter gravity theory, as formulated by Wesson and co-workers,[4,5] admits these generalized Gödel-type metrics as solutions to its field equations, thus also extending the investigations of Ref. 65.

The outline of this article is as follows. In the next section we present a summary of some important prerequisites for Sec. III, where using the equivalence problem techniques as formulated by Cartan[66] we derive the necessary and sufficient conditions for local homogeneity of this class of 5D generalized Gödel-type manifolds. In Sec. III we also exhibit an irreducible set of isometrically nonequivalent homogeneous generalized Gödel-type metrics. In Sec. IV we discuss the integration of the Killing equations and present the Killing vector fields as well as the corresponding Lie algebra for all homogeneous generalized Gödel-type metrics. In the last section we examine whether the IM field equations permit solutions of this generalized Gödel-type class of geometries. The unique solution of this type is found therein. The question as to whether the IM version of general relativity rules out the existence of closed timelike curves of Gödel type is also discussed (Sec. V) in connection with a recent paper by Romero *et al.*[67]

II. PREREQUISITES

The arbitrariness in the choice of coordinates in the metric theories of gravitation gives rise to the problem of deciding whether or not two manifolds whose metrics g and $\tilde{g}$ are given explicitly in terms of coordinates, viz.,

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \quad \text{and} \quad d\tilde{s}^2 = \tilde{g}_{\mu\nu} d\tilde{x}^\mu d\tilde{x}^\nu, \tag{2.1}$$

are locally isometric. This is the so-called equivalence problem (see Ref. 66 for the local equivalence of n-dimensional Riemannian manifolds, and Refs. 68 and 69 for the special case $n=4$ of general relativity).

The Cartan solution[66] to the equivalence problem for Riemannian manifolds can be summarized as follows. Two n-dimensional Lorentzian Riemannian manifolds $\mathcal{M}_n$ and $\tilde{\mathcal{M}}_n$ are locally equivalent if there exist coordinate and generalized n-dimensional Lorentz transformations such that the following *algebraic* equations relating the frame components of the curvature tensor and their covariant derivatives:

$$\begin{aligned} R^A{}_{BCD} &= \tilde{R}^A{}_{BCD}\,, \\ R^A{}_{BCD;M_1} &= \tilde{R}^A{}_{BCD;M_1}, \\ R^A{}_{BCD;M_1 M_2} &= \tilde{R}^A{}_{BCD;M_1 M_2}, \\ &\vdots \\ R^A{}_{BCD;M_1 \ldots M_{p+1}} &= \tilde{R}^A{}_{BCD;M_1 \ldots M_{p+1}} \end{aligned} \tag{2.2}$$

are compatible as *algebraic* equations in (x^μ, ξ^A). Here and in what follows we use a semicolon to denote covariant derivatives. Note that x^μ are coordinates on the manifold $\mathcal{M}_n$ while ξ^A parametrize the group of allowed frame transformations [n-dimensional generalized Lorentz group usually denoted[70] by $O(n-1,1)$]. Reciprocally, Eqs. (2.2) imply local equivalence between the n-dimensional manifolds $\mathcal{M}_n$ and $\tilde{\mathcal{M}}_n$.

In practice, a fixed frame is chosen to perform the calculations so that only coordinates appear in the components of the curvature tensor, i.e., there is no explicit dependence on the parameters ξ^A of the generalized Lorentz group.

Another important practical point to be considered, once one wishes to test the local equivalence of two Riemannian manifolds, is that before attempting to solve Eqs. (2.2) one can extract and compare partial pieces of information at each step of differentiation as, for example, the number $\{t_0, t_1, \ldots, t_p\}$ of functionally independent functions of the coordinates x^μ contained in the corresponding set

$$I_p = \{R^A{}_{BCD}, R^A{}_{BCD;M_1}, R^A{}_{BCD;M_1M_2}, \ldots, R^A{}_{BCD;M_1M_2\ldots M_p}\}, \tag{2.3}$$

and the isotropy subgroup $\{H_0, H_1, \ldots, H_p\}$ of the symmetry group G_r under which the set corresponding I_p is invariant. They must be the same for each step $q = 0,1,...,p$ if the manifolds are locally equivalent.

In practice it is also important to note that in calculating the curvature and its covariant derivatives, in a chosen frame, one can stop as soon as one reaches a step at which the pth derivatives (say) are algebraically expressible in terms of the previous ones, and the residual isotropy group (residual frame freedom) at that step is the same isotropy group of the previous step, i.e., $H_p = H_{(p-1)}$. In this case further differentiation will not yield any new piece of information. Actually, if $H_p = H_{(p-1)}$ and, in a given frame, the pth derivative is expressible in terms of its predecessors, for any $q > p$ the qth derivatives can all be expressed in terms of the 0th,1st,$\ldots$,$(p-1)$th derivatives.[66,69]

Since there are t_p essential coordinates, in five dimensions clearly $5 - t_p$ are ignorable, so the isotropy group will have dimension $s = \dim(H_p)$, and the group of isometries of the metric will have dimension r given by (see Ref. 66)

$$r = s + 5 - t_p, \tag{2.4}$$

acting on an orbit with dimension

$$d = r - s = 5 - t_p. \tag{2.5}$$

III. HOMOGENEITY AND NONEQUIVALENT METRICS

The line element of the five-dimensional *generalized* Gödel-type manifolds $\mathcal{M}_5$ we are concerned with is given by

$$d\hat{s}^2 = dt^2 + 2H(x)dt\,dy - dx^2 - G(x)dy^2 - \tilde{F}^2(\tilde{u}) - (d\tilde{z}^2 + d\tilde{u}^2), \tag{3.1}$$

where $H(x)$, $G(x)$, and $\tilde{F}(\tilde{u})$ are arbitrary real functions. By a suitable choice of coordinates the line element (3.1) can be brought into the form

$$d\hat{s}^2 = [dt + H(x)dy]^2 - dx^2 - D^2(x)dy^2 - F^2(u)dz^2 - du^2, \tag{3.2}$$

where $D^2(x) = G + H^2$ and u clearly is a new fifth coordinate.

At an arbitrary point of $\mathcal{M}_5$ one can choose the following set of linearly independent one-forms $\hat{\Theta}^A$:

$$\hat{\Theta}^0 = dt + H(x)dy, \quad \hat{\Theta}^1 = dx, \quad \hat{\Theta}^2 = D(x)dy, \quad \hat{\Theta}^3 = F(u)dz, \quad \hat{\Theta}^4 = du, \tag{3.3}$$

such that the Gödel-type line element (3.2) can be written as

$$d\hat{s}^2 = \hat{\eta}_{AB}\hat{\Theta}^A\hat{\Theta}^B = (\hat{\Theta}^0)^2 - (\hat{\Theta}^1)^2 - (\hat{\Theta}^2)^2 - (\hat{\Theta}^3)^2 - (\hat{\Theta}^4)^2. \tag{3.4}$$

Here and in what follows capital letters are 5D Lorentz frame indices and run from 0 to 4; they are raised and lowered with Lorentz matrices $\hat{\eta}^{AB}=\hat{\eta}_{AB}=\text{diag}(+1,-1,-1,-1,-1)$, respectively.

Using as input the one-forms (3.3) and the Lorentz frame (3.4), the computer algebra package CLASSI,[6,71] e.g., gives the following nonvanishing Lorentz frame components $\hat{R}_{ABCD}$ of the curvature:

$$\hat{R}_{0101}=\hat{R}_{0202}=-\frac{1}{4}\left(\frac{H'}{D}\right)^2, \tag{3.5}$$

$$\hat{R}_{0112}=\frac{1}{2}\left(\frac{H'}{D}\right)', \tag{3.6}$$

$$\hat{R}_{1212}=\frac{D''}{D}-\frac{3}{4}\left(\frac{H'}{D}\right)^2, \tag{3.7}$$

$$\hat{R}_{3434}=\frac{\ddot{F}}{F}, \tag{3.8}$$

where the prime and the dot denote, respectively, derivative with respect to x and u.

For 5D (local) homogeneity from Eq. (2.5) one must have $t_q=0$ for $q=0,1,...,p$, that is, the number of functionally independent functions of the coordinates x^μ in the set I_p must be zero. Therefore, from Eqs. (3.5–3.8) we conclude that for 5D homogeneity it is necessary that

$$\frac{H'}{D}=\text{const}\equiv-2\omega, \tag{3.9}$$

$$\frac{D''}{D}=\text{const}\equiv m^2, \tag{3.10}$$

$$\frac{\ddot{F}}{F}=\text{const}\equiv k. \tag{3.11}$$

The above necessary conditions are also sufficient for 5D local homogeneity. Indeed, under these conditions the nonvanishing frame components of the curvature reduce to

$$\hat{R}_{0101}=\hat{R}_{0202}=-\omega^2, \tag{3.12}$$

$$\hat{R}_{1212}=m^2-3\omega^2, \tag{3.13}$$

$$\hat{R}_{3434}=k. \tag{3.14}$$

Following Cartan's method for the local equivalence, we calculate the first covariant derivative of the Riemann tensor. One obtains the following non-null covariant derivatives of the curvature:

$$\hat{R}_{0112;1}=\hat{R}_{0212;2}=\omega(m^2-4\omega^2). \tag{3.15}$$

Clearly, regardless of the value of the constant k, the first covariant derivative of the curvature is algebraically expressible in terms of the Riemann tensor. Moreover, the number of functionally independent functions of the coordinates x^μ among the components of the curvature and its first covariant derivative is zero ($t_0=t_1=0$). As far as the dimension of the residual isotropy group is concerned, we distinguish three different classes of locally homogeneous 5D generalized Gödel-type curved manifolds, according to the relevant parameters m^2, ω, and k, namely,[72]

(1) $\dim(H_0)=\dim(H_1)=2$ when

(a) $\omega\neq 0$, any real k, $m^2\neq 4\omega^2$;
(b) $\omega=0$, $k\neq 0$, $m^2\neq 0$;

(2) $\dim(H_0)=\dim(H_1)=4$ when

(a) $\omega\neq 0$, any real k, $m^2=4\omega^2$;
(b) $\omega=0$, $k=0$, $m^2\neq 0$;
(c) $\omega=0$, $k\neq 0$, $m^2=0$;

(3) $\dim(H_0)=\dim(H_1)=10$ when $\omega=k=m^2=0$.

Thus, from Eqs. (2.4) and (2.5) one finds that the locally homogeneous 5D generalized Gödel-type manifolds admit a (local) G_r, with either $r=7$, $r=9$, or $r=15$ acting on an orbit of dimension $d=5$, that is, on the manifold $\mathcal{M}_5$.

The above results can be collected together in the following theorems:

Theorem 1: *The necessary and sufficient conditions for a five-dimensional generalized Gödel-type manifold to be locally homogeneous are those given by Eqs. (3.9–3.11).*

Theorem 2: *The five-dimensional homogeneous generalized Gödel-type manifolds are locally characterized by three independent real parameters, ω, k, and m^2: identical triads (ω,k,m^2) specify locally equivalent manifolds.*

Theorem 3: *The five-dimensional locally homogeneous generalized Gödel-type manifolds admit group of isometry G_r with*

(i) $r=7$ *if either of the above conditions (1a) and (1b) is fulfilled;*
(ii) $r=9$ *if one of the above set of conditions (2a), (2b), and (2c) is fulfilled;* and
(iii) $r=15$ *if the above condition (3) is satisfied.*

We shall now focus our attention on the irreducible set of isometrically nonequivalent homogeneous generalized Gödel-type metrics. These nonequivalent classes of metrics can be obtained by a similar procedure to that used by Rebouças and Tiomno,[41] namely by integrating Eqs. (3.9–3.11), and eliminating through coordinate transformations the nonessential integration constants taking into account the relevant parameters according to the above Theorem 2. For the sake of brevity, however, we shall only present the irreducible classes without going into details of calculations. It turns out that one ought to distinguish six classes of metrics according to the following.

Class I: $m^2>0$, any real $k,\omega\neq 0$. The line element for this class of homogeneous generalized Gödel-type manifolds can always be brought [in cylindrical coordinates (r,ϕ,z)] into the form

$$d\hat{s}^2=[dt+H(r)d\phi]^2-D^2(r)d\phi^2-dr^2-F^2(u)dz^2-du^2 \tag{3.16}$$

with the metric functions given by

$$H(r)=\frac{2\omega}{m^2}[1-\cosh(mr)], \tag{3.17}$$

$$D(r)=m^{-1}\sinh(mr), \tag{3.18}$$

$$F(u)=\begin{cases}\alpha^{-1}\sin(\alpha u) & \text{if } k=-\alpha^2<0,\\ u & \text{if } k=0,\\ \alpha^{-1}\sinh(\alpha u) & \text{if } k=\alpha^2>0.\end{cases} \tag{3.19}$$

According to Theorem 3, the possible isometry groups for this class are either G_7 (for $m^2\neq 4\omega^2$) or G_9 (when $m^2=4\omega^2$), irrespective of the value of k.

Class II: $m^2=0$, any real $k,\omega\neq 0$. The line element for this class can be brought into the form (3.16), with the metric function $F(u)$ given by (3.19), but now the functions $H(r)$ and $D(r)$ are given by

$$H(r)=-\omega r^2 \quad \text{and} \quad D(r)=r. \tag{3.20}$$

For this class from Theorem 3 there is a group G_7 of isometries, regardless of the value of k.

Class III: $m^2\equiv-\mu^2<0$, any real $k,\omega\neq 0$. Similarly for this class the line element reduces to (3.16) with $F(u)$ given by (3.19) and

$$H(r)=\frac{2\omega}{\mu^2}[\cos(\mu r)-1], \tag{3.21}$$

$$D(r)=\mu^{-1}\sin(\mu r). \tag{3.22}$$

From Theorem 3, regardless of the value of k for this class there is a group G_7 of isometries.

Class IV: $m^2\neq 0$, any real k, and $\omega=0$. We shall refer to this class as degenerated Gödel-type manifolds, since the cross term in the line element, related to the rotation ω in 4D Gödel model, vanishes. By a trivial coordinate transformation one can make $H=0$ with $D(r)$ given, respectively, by (3.18) or (3.22) depending on whether $m^2>0$ or $m^2\equiv-\mu^2<0$. The function $F(u)$ depends on the sign of k and is again given by (3.19). For this class according to Theorem 3 one may have either a G_7 for $k\neq 0$, or a G_9 for $k=0$.

Class V: $m^2=0$, $k\neq 0$, and $\omega=0$. By a trivial coordinate transformation one can make $H=0$, $D=r$, and $F(u)=\alpha^{-1}\sin(\alpha u)$ or $F(u)=\alpha^{-1}\sinh(\alpha u)$ depending on whether $k<0$ or $k>0$, respectively. From Theorem 3 there is a group G_9 of isometries.

Class VI: $m^2=0$, $k=0$, and $\omega=0$. From (3.12–3.14) this corresponds to the 5D flat manifold. Therefore, one can make $H=0$, $D(r)=r$, and $F(u)=u$. Theorem 3 ensures that there is a group G_{15} of isometries.

IV. KILLING VECTOR FIELDS

In this section we shall present the infinitesimal generators of isometries of the 5D homogeneous generalized Gödel-type manifolds, whose line element (3.16) can be brought into the Lorentzian form (3.4) with $\hat{\Theta}^A$ given by

$$\hat{\Theta}^0=dt+H(r)d\phi, \quad \hat{\Theta}^1=dr, \quad \hat{\Theta}^2=D(r)d\phi, \quad \hat{\Theta}^3=F(u)dz, \quad \hat{\Theta}^4=du, \tag{4.1}$$

where the functions $H(r)$, $D(r)$, and $F(u)$ depend upon the essential parameters m^2, k, and ω according to the above classes of locally homogeneous manifolds.

Denoting the coordinate components of a generic Killing vector field $\hat{K}$ by $\hat{K}^u\equiv(Q,R,S,\overline{Z},U)$, where Q, R, S, $\overline{Z}$, and U are functions of all coordinates t,r,ϕ,z,u, then the 15 Killing equations

$$\hat{K}_{(A;B)}\equiv\hat{K}_{A;B}+\hat{K}_{B;A}=0 \tag{4.2}$$

can be written in the Lorentz frame (3.4)–(4.1) as

$$T_t=0, \quad T_u-U_t=0, \tag{4.3}$$

$$R_r=0, \quad U_r+R_u=0, \tag{4.4}$$

$$U_u=0, \tag{4.5}$$

$$D(T_r-R_t)-H_rP=0, \tag{4.6}$$

$$DP_u + U_\phi - HU_t = 0, \tag{4.7}$$

$$T_\phi + H_r R - DP_t = 0, \tag{4.8}$$

$$R_\phi - HR_t - D_r P + DP_r = 0, \tag{4.9}$$

$$P_\phi - HP_t + D_r R = 0, \tag{4.10}$$

$$T_z - FZ_t = 0, \tag{4.11}$$

$$FZ_r + R_z = 0, \tag{4.12}$$

$$Z_z + UF_u = 0, \tag{4.13}$$

$$U_z + FZ_u - ZF_u = 0, \tag{4.14}$$

$$DP_z + F(Z_\phi - HZ_t) = 0, \tag{4.15}$$

where the subscripts denote partial derivatives, and where we have made

$$T \equiv HS + Q, \quad P \equiv DS, \quad \text{and} \quad Z \equiv F\overline{Z} \tag{4.16}$$

to make easier the comparison and the use of the results obtained in Ref. 42. To this end we note that with the changes $u \to z$ and $U \to Z$ the above equations (4.3–4.10) are formally identical to the Killing equations (4)–(11) of Ref. 42. However, in the equations (4.3–4.10) the functions T,R,P,U depend additionally on the fifth coordinate u. Taking into account this similitude, the integration of the Killing equations (4.3–4.15) can be obtained in two steps as follows. First, by analogy with (4)–(11) of Ref. 42 one integrates (4.3–4.10), but at this step instead of the integration constants one has integration functions of the fifth coordinate u. Second, one uses the remaining equations (4.11–4.15) to achieve explicit forms for these integration functions and to obtain the last component U of the generic Killing vector K.

We have used the above two-step procedure to integrate the Killing equations (4.3–4.15) for all classes of homogeneous generalized Gödel-type manifolds. However, for the sake of brevity, we shall only present the Killing vector fields and the corresponding Lie algebras without going into details of calculations, which can be verified by using, for example, the computer algebra program KILLNF, written in CLASSI by Åman.[71]

Class I: $m^2 > 0$, any real $k, \omega \neq 0$. In the integration of the Killing equation for this general class one is led to distinguish two different subclasses of solutions depending on whether $m^2 \neq 4\omega^2$ or $m^2 = 4\omega^2$. We shall refer to these subclasses as classes Ia and Ib, respectively.

Class Ia: $m^2 > 0$, any real $k, m^2 \neq 4\omega^2$. In the coordinate basis in which (3.16) is given, a set of linearly independent Killing vector fields K_N (N is an enumerating index) is given by

$$K_1 = \partial_t, \quad K_2 = \frac{2\omega}{m}\partial_t - m\partial_\phi, \tag{4.17}$$

$$K_3 = -\frac{H}{D}\sin\phi\,\partial_t + \cos\phi\,\partial_r - \frac{D_r}{D}\sin\phi\,\partial_\phi, \tag{4.18}$$

$$K_4 = -\frac{H}{D}\cos\phi\,\partial_t - \sin\phi\,\partial_r - \frac{D_r}{D}\cos\phi\,\partial_\phi, \tag{4.19}$$

$$K_5 = \sin z\,\partial_u + \frac{F_u}{F}\cos z\,\partial_z, \tag{4.20}$$

$$K_6 = \cos z \partial_u - \frac{F_u}{F} \sin z \partial_z , \tag{4.21}$$

$$K_7 = \partial_z . \tag{4.22}$$

The Lie algebra has the following nonvanishing commutators:

$$[K_2, K_3] = -mK_4, \quad [K_2, K_4] = mK_3, \quad [K_3, K_4] = mK_2, \tag{4.23}$$

$$[K_5, K_6] = -kK_7, \quad [K_5, K_7] = -K_6, \quad [K_6, K_7] = K_5. \tag{4.24}$$

Therefore the corresponding algebra is $\mathcal{L}_{\mathrm{Ia}} = \mathcal{L}_k \oplus \tau \oplus \mathrm{so}(2,1)$. Here and in what follows the symbols $\oplus$ and $\in\!\!\!\!+$ denote and direct and semidirect sum of subalgebras, and the subalgebra $\mathcal{L}_k$ is so(3) for $k<0$, so(2,1) for $k>0$, and $t^2 \in\!\!\!\!+ \mathrm{so}(2)$ for $k=0$. For the present class $\mathcal{L}_k$ is generated by K_5, K_6, and K_7, the symbol τ is associated to the time translation K_1, and finally the infinitesimal generators of subalgebra so(2,1) are K_2, K_3, and K_4.

Class Ib: $m^2 = 4\omega^2$, any real $k, \omega \neq 0$. For this class the Killing vector fields are

$$K_1 = \partial_t, \quad K_2 = \partial_t - m\partial_\phi, \tag{4.25}$$

$$K_3 = -\frac{H}{D} \sin \phi \partial_t + \cos \phi \partial_r - \frac{D_r}{D} \sin \phi \partial_\phi , \tag{4.26}$$

$$K_4 = -\frac{H}{D} \cos \phi \partial_t - \sin \phi \partial_r - \frac{D_r}{D} \cos \phi \partial_\phi , \tag{4.27}$$

$$K_5 = -\frac{H}{D} \cos(mt + \phi) \partial_t + \sin(mt + \phi) \partial_r + \frac{1}{D} \cos(mt + \phi) \partial_\phi , \tag{4.28}$$

$$K_6 = -\frac{H}{D} \sin(mt + \phi) \partial_t - \cos(mt + \phi) \partial_r + \frac{1}{D} \sin(mt + \phi) \partial_\phi , \tag{4.29}$$

$$K_7 = \sin z \partial_u + \frac{F_u}{F} \cos z \partial_z , \tag{4.30}$$

$$K_8 = \cos z \partial_u - \frac{F_u}{F} \sin z \partial_z , \tag{4.31}$$

$$K_9 = \partial_z , \tag{4.32}$$

whose Lie algebra is given by

$$[K_1, K_5] = -mK_6, \quad [K_1, K_6] = mK_5, \quad [K_2, K_3] = -mK_4, \tag{4.33}$$

$$[K_2, K_4] = mK_3, \quad [K_3, K_4] = mK_2, \quad [K_5, K_6] = mK_1, \tag{4.34}$$

$$[K_7, K_8] = -kK_9, \quad [K_7, K_9] = -K_8, \quad [K_8, K_9] = K_7. \tag{4.35}$$

So, the corresponding algebra for this case is $\mathcal{L}_{\mathrm{Ib}} = \mathcal{L}_k \oplus \mathrm{so}(2,1) \oplus \mathrm{so}(2,1)$. As in the previous class, the subalgebra $\mathcal{L}_k$ depends on the sign of k, and here is generated by K_7, K_8, and K_9. The two subalgebras so(2,1) are generated by the Killing vector fields K_1, K_5, K_6 and K_2, K_3, K_4.

Class II: $m^2 = 0$, any real $k, \omega \neq 0$. For this class the Killing vector fields turn out to be the following:

$$K_1=\partial_t, \quad K_2=\partial_\phi, \tag{4.36}$$

$$K_3=-\omega r \sin\phi\partial_t-\cos\phi\partial_r+\frac{1}{r}\sin\phi\partial_\phi, \tag{4.37}$$

$$K_4=-\omega r \cos\phi\partial_t+\sin\phi\partial_r+\frac{1}{r}\cos\phi\partial_\phi, \tag{4.38}$$

$$K_5=\sin z\partial_u+\frac{F_u}{F}\cos z\partial_z, \tag{4.39}$$

$$K_6=\cos z\partial_u-\frac{F_u}{F}\sin z\partial_z, \tag{4.40}$$

$$K_7=\partial_z. \tag{4.41}$$

The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_3]=K_4, \quad [K_2,K_4]=-K_3, \quad [K_3,K_4]=2\omega K_1, \tag{4.42}$$

$$[K_5,K_6]=-kK_7, \quad [K_5,K_7]=-K_6, \quad [K_6,K_7]=K_5. \tag{4.43}$$

Therefore, the corresponding algebra for this case is $\mathcal{L}_{\rm II}=\mathcal{L}_k\oplus\mathcal{L}_4$. The subalgebra $\mathcal{L}_4$ is generated by K_1, K_2, K_3, and K_4. This algebra $\mathcal{L}_4$ is soluble and does not contain Abelian 3D subalgebras; it is classified as type III with $q=0$ by Petrov.[73] The subalgebra $\mathcal{L}_k$ is the same as the previous classes and is generated by K_5, K_6, and K_7.

Class III: $m^2\equiv-\mu^2<0$, any real $k,\omega\neq0$. For this class the set of linearly independent Killing vector fields we have found is given by

$$K_1=\partial_t, \quad K_2=\frac{2\omega}{\mu}\partial_t+\mu\partial_\phi, \tag{4.44}$$

$$K_3=-\frac{H}{D}\sin\phi\partial_t+\cos\phi\partial_r-\frac{D_r}{D}\sin\phi\partial_\phi, \tag{4.45}$$

$$K_4=-\frac{H}{D}\cos\phi\partial_t-\sin\phi\partial_r-\frac{D_r}{D}\cos\phi\partial_\phi, \tag{4.46}$$

$$K_5=\sin z\partial_u+\frac{F_u}{F}\cos z\partial_z, \tag{4.47}$$

$$K_6=\cos z\partial_u-\frac{F_u}{F}\sin z\partial_z, \tag{4.48}$$

$$K_7=\partial_z. \tag{4.49}$$

The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_3]=\mu K_4, \quad [K_2,K_4]=-\mu K_3, \quad [K_3,K_4]=\mu K_2, \tag{4.50}$$

$$[K_5,K_6]=-kK_7, \quad [K_5,K_7]=-K_6, \quad [K_6,K_7]=K_5. \tag{4.51}$$

Thus, the corresponding algebra for this case is $\mathcal{L}_{\rm III}=\mathcal{L}_k\oplus\tau\oplus{\rm so}(3)$. Here τ is associated to the Killing vector field K_1, whereas the subalgebra so(3) corresponds to K_2, K_3, and K_4. Again $\mathcal{L}_k$ is generated by K_5, K_6, and K_7.

Class IV: $m^2\neq 0$, any real $k,\omega=0$. In the integration of the Killing equation for this general class one is led to distinguish two different subclasses according to $k\neq 0$ or $k=0$. We shall denote these subclasses as classes IVa and IVb, respectively.

Class IVa: $m^2\neq 0$, $k\neq 0$, $\omega=0$. This class corresponds to the so-called degenerated Gödel-type manifolds. One obtains for this class the following Killing vector fields:

$$K_1=\partial_t, \quad K_2=\partial_\phi, \tag{4.52}$$

$$K_3=\cos\phi\,\partial_r-\frac{D_r}{D}\sin\phi\,\partial_\phi, \tag{4.53}$$

$$K_4=-\sin\phi\,\partial_r-\frac{D_r}{D}\cos\phi\,\partial_\phi, \tag{4.54}$$

$$K_5=\sin z\,\partial_u+\frac{F_u}{F}\cos z\,\partial_z, \tag{4.55}$$

$$K_6=\cos z\,\partial_u-\frac{F_u}{F}\sin z\,\partial_z, \tag{4.56}$$

$$K_7=\partial_z, \tag{4.57}$$

where $D(r)=(1/m)\sinh mr$ for $m^2>0$, or $D(r)=(1/\mu)\sin\mu r$ for $m^2\equiv-\mu^2<0$, and the function $F(u)$ for $k\neq 0$ is given by (3.19). The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_3]=K_4, \quad [K_2,K_4]=-K_3, \quad [K_3,K_4]=-m^2K_2, \tag{4.58}$$

$$[K_5,K_6]=-kK_7, \quad [K_5,K_7]=-K_6, \quad [K_6,K_7]=K_5, \tag{4.59}$$

where one should substitute $-m^2$ by μ^2 if $m^2<0$. So, the corresponding Lie algebra is $\mathcal{L}_{\rm IVa}=\mathcal{L}_k\oplus\tau\oplus\mathcal{L}_m$, where $\mathcal{L}_m$ is so(2,1) for $m^2>0$, and so(3) for $m^2=-\mu^2<0$. The subalgebra $\mathcal{L}_k$ (generated by K_5, K_6, and K_7) is so(3) for $k<0$, and so(2,1) for $k>0$. Again τ is associated to the Killing vector field K_1.

Class IVb: $m^2\neq 0$, $k=0$, $\omega=0$. We shall refer to this class as doubly degenerated Gödel-type manifolds. One obtains for this class the following Killing vector fields:

$$K_1=\partial_t, \quad K_2=\partial_\phi, \tag{4.60}$$

$$K_3=\cos\phi\,\partial_r-\frac{D_r}{D}\sin\phi\,\partial_\phi, \tag{4.61}$$

$$K_4=-\sin\phi\,\partial_r-\frac{D_r}{D}\cos\phi\,\partial_\phi, \tag{4.62}$$

$$K_5=\sin z\,\partial_u+\frac{1}{u}\cos z\,\partial_z, \tag{4.63}$$

$$K_6=\cos z\,\partial_u-\frac{1}{u}\sin z\,\partial_z, \tag{4.64}$$

$$K_7=\partial_z\,, \tag{4.65}$$

$$K_8=u\sin z\,\partial_t+t\sin z\,\partial_u+\frac{1}{u}t\cos z\,\partial_z\,, \tag{4.66}$$

$$K_9=u\cos z\,\partial_t+t\cos z\,\partial_u-\frac{1}{u}t\sin z\,\partial_z\,, \tag{4.67}$$

where again $D(r)=(1/m)\sinh mr$ for $m^2>0$, or $D(r)=(1/\mu)\sin\mu r$ for $m^2\equiv-\mu^2<0$.

The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_3]=K_4,\quad [K_2,K_4]=-K_3,\quad [K_3,K_4]=-m^2K_2, \tag{4.68}$$

$$[K_5,K_7]=-K_6,\quad [K_6,K_7]=K_5,\quad [K_1,K_8]=K_5, \tag{4.69}$$

$$[K_1,K_9]=K_6,\quad [K_5,K_8]=K_1,\quad [K_6,K_9]=K_1, \tag{4.70}$$

$$[K_7,K_8]=K_9,\quad [K_7,K_9]=-K_8,\quad [K_8,K_9]=-K_7, \tag{4.71}$$

where one should substitute $-m^2$ by μ^2 if $m^2<0$. So, the corresponding Lie algebra is $\mathcal{L}_{\mathrm{IVb}}=t^3 \mathbin{⋲} \mathrm{so}(2,1)\oplus\mathcal{L}_m$, where $\mathcal{L}_m$ is generated by K_2, K_3, K_4, and is either so(2,1) or so(3) depending on whether $m^2>0$ or $m^2=-\mu^2<0$. The subalgebra $t^3 \mathbin{⋲} \mathrm{so}(2,1)$ is generated by K_1,K_5,K_6,K_7,K_8,K_9.

Class V: $m^2=0$, $k\neq0$, $\omega=0$. A set of linearly independent Killing vector field for this class is

$$K_1=\partial_t\,,\quad K_2=\partial_\phi\,, \tag{4.72}$$

$$K_3=\cos\phi\,\partial_r-\frac{1}{r}\sin\phi\,\partial_\phi\,, \tag{4.73}$$

$$K_4=-\sin\phi\,\partial_r-\frac{1}{r}\cos\phi\,\partial_\phi\,, \tag{4.74}$$

$$K_5=\sin z\,\partial_u+\frac{F_u}{F}\cos z\,\partial_z\,, \tag{4.75}$$

$$K_6=\cos z\,\partial_u-\frac{F_u}{F}\sin z\,\partial_z\,, \tag{4.76}$$

$$K_7=\partial_z\,, \tag{4.77}$$

$$K_8=r\sin\phi\,\partial_t+t\sin\phi\,\partial_r+\frac{1}{r}t\cos\phi\,\partial_\phi\,, \tag{4.78}$$

$$K_9=r\cos\phi\,\partial_t+t\cos\phi\,\partial_r-\frac{1}{r}t\sin\phi\,\partial_\phi\,, \tag{4.79}$$

where $F(u)$ depends upon the sign of k and is given by Eq. (3.19).

The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_3]=K_4,\quad [K_2,K_4]=-K_3,\quad [K_5,K_6]=-kK_7, \tag{4.80}$$

$$[K_5,K_7]=-K_6,\quad [K_6,K_7]=K_5,\quad [K_1,K_8]=-K_4, \tag{4.81}$$

$$[K_1, K_9]=K_3, \quad [K_4, K_8]=-K_1, \quad [K_3, K_9]=K_1, \tag{4.82}$$

$$[K_2, K_8]=K_9, \quad [K_2, K_9]=-K_8, \quad [K_8, K_9]=-K_2. \tag{4.83}$$

So, the corresponding Lie algebra is $\mathcal{L}_{\rm V}=t^3 \oplus\!\!\!\!\!\supset \mathrm{so}(2,1)\oplus\mathcal{L}_k$, where $\mathcal{L}_k$ is generated by K_5, K_6, K_7, and is either so(2,1) or so(3) depending on whether $k>0$ or $k<0$. The subalgebra $t^3 \oplus\!\!\!\!\!\supset \mathrm{so}(2,1)$ is generated by $K_1, K_2, K_3, K_4, K_8, K_9$.

Class VI: $m^2=0$, $k=0$, $\omega=0$. From (3.12)–(3.14) this case corresponds to the 5D flat manifold whose Lie algebra is $\mathcal{L}_{\rm VI}=\mathrm{so}(4,1)$ since it clearly has the well-known 15 Killing vector fields, namely five translations, four space–time rotations, and six space rotations.

It is worth noting that none of the above Lie algebras is semi-simple, but some of their subalgebras are. Besides, most of the simple subalgebras are noncompact. The 3D subalgebra so (3) present in all classes is compact, though.

The number of Killing vector fields we have found for each of the above six classes makes explicit that the 5D locally homogeneous generalized Gödel-type manifolds admit a group of isometry G_7 when (1a): $m^2\neq 4\omega^2$, any real $k,\omega\neq 0$, or when (1b): $m^2\neq 0$, $k\neq 0$, $\omega=0$. Groups G_9 of isometry occur when (2a): $m^2=4\omega^2$, any real $k,\omega\neq 0$, or (2b): $m^2\neq 0$, $k=0$, $\omega=0$, or when (2c): $m^2=0$, $k\neq 0$, $\omega=0$. Clearly when $m^2=\omega=k=0$ there is G_{15}. These possible groups are in agreement with Theorem 3 of the previous section. Actually the integration of the Killing equations constitutes a different way of deriving that theorem. Furthermore, these equations also show that the isotropy subgroup H of G_r is such that $\dim(H)=2$ when the above conditions (1a) and (1b) are satisfied, while the conditions (2a)–(2c) lead to $\dim(H)=4$, also in agreement with the previous section. Clearly $\dim(H)=10$ when $m^2=\omega=k=0$.

V. CAUSAL ANOMALIES AND FINAL REMARKS

In this section we shall initially be concerned with the problem of causal anomalies in the generalized Gödel-type manifolds. Then we proceed by examining whether the IM gravity allows solutions of generalized Gödel-type metrics (3.16). Finally, we conclude by addressing to the general question as to whether the IM gravity theory rules out the 4D noncausal Gödel-type solutions to Einstein's equations of general relativity.

In the first three of the six classes of homogeneous generalized Gödel-type manifolds we have discussed in Sec. III, there are closed timelike curves. Indeed, the analysis made in a previous paper[64] can be easily extended to the generalized 5D Gödel-type manifolds of the present article. To this end, we write the line element (3.16) in the form

$$ds^2=dt^2+2H(r)dt\,d\phi-dr^2-G(r)d\phi^2-F^2(u)dz^2-du^2, \tag{5.1}$$

where $G(r)=D^2-H^2$ and (r,ϕ,z) are cylindrical coordinates. Now, the existence of closed timelike curves of the Gödel-type depends on the behavior of $G(r)$. Indeed, if $G(r)<0$ for a certain range of r ($r_1<r<r_2$, say), Gödel's circles[74] $u,t,z,r=$const are closed timelike curves.

Since one can always make $H=0$ for the generalized Gödel-type manifolds of classes IV, V, and VI, then $G(r)>0$ for all $r>0$. Thus there are no closed timelike Gödel's circles in these classes of manifolds.

On the other hand, following the above-outlined reasoning it is easy to show (see Ref. 64 for details) that for each of the remaining three classes (classes I–III) one can always find a critical radius r_c such that for all $r>r_c$ one has $G(r)<0$, making clear that there are closed timelike curves in these families of homogeneous generalized Gödel-type manifolds. However, in what follows we shall show that these types of noncausal *curved* manifolds are not permitted in the context of the induced matter theory.

In the Lorentz frame $\hat{\Theta}^A$ given by (4.1) the nonvanishing frame components of the Einstein tensor $\hat{G}_{AB}=\hat{R}_{AB}-\frac{1}{2}R\,\hat{\eta}_{AB}$ are

$$\hat{G}_{00} = -\frac{D''}{D} + \frac{3}{4}\left(\frac{H'}{D}\right)^2 - \frac{\ddot{F}}{F}, \tag{5.2}$$

$$\hat{G}_{02} = \frac{1}{2}\left(\frac{H'}{D}\right)', \tag{5.3}$$

$$\hat{G}_{11} = \hat{G}_{22} = \frac{1}{4}\left(\frac{H'}{D}\right)^2 + \frac{\ddot{F}}{F}, \tag{5.4}$$

$$\hat{G}_{33} = \hat{G}_{44} = \frac{D''}{D} - \frac{1}{4}\left(\frac{H'}{D}\right)^2, \tag{5.5}$$

where the prime and dot denote derivative with respect to r and u, respectively.

The field equations (1.2) require that $\hat{G}_{02}=0$, which in turn implies that

$$\frac{H'}{D} = \text{const} \equiv -2\omega. \tag{5.6}$$

Inserting (5.6) into (5.4), (5.5), and (5.2) one easily finds that the IM field equations are fulfilled if and only if the independent parameters ω, k, and m^2 [see Eqs.(3.9) and (3.10)] vanish identically, which leads to the only solution given by

$$H = a, \quad D = br + c, \quad \text{and} \quad F = \beta u + \gamma, \tag{5.7}$$

where, a, b, c, β, and γ are arbitrary real constants. However, these constants have no physical meaning, and can be taken to be $a=c=\gamma=0$ and $b=\beta=1$ by a suitable choice of coordinates. Indeed, if one performs the coordinate transformations

$$t = \overline{t} - \frac{a}{b}\overline{\phi}, \quad r = \overline{r} - \frac{c}{b}, \tag{5.8}$$

$$\phi = \frac{\overline{\phi}}{b}, \quad z = \frac{\overline{z}}{\beta}, \quad u = \overline{u} - \frac{\gamma}{\beta}, \tag{5.9}$$

the line element (5.1) becomes

$$d\hat{s}^2 = d\overline{t}^2 - d\overline{r}^2 - \overline{r}^2 d\overline{\phi}^2 - d\overline{z}^2 - d\overline{u}^2, \tag{5.10}$$

in which we obviously have $G(\overline{r}) = \overline{r}^2 > 0$ for $\overline{r} \neq 0$. The line element (5.10) corresponds to a manifestly flat 5D manifold, making it clear that the underlying manifold can be taken to be the simply connected Euclidean manifold $\mathbb{R}^5$, and therefore as $G(\overline{r}) > 0$ no closed timelike circles are permitted. Furthermore, the above results clearly show that the IM theory does not admit any *curved* 5D Gödel-type metric (3.16) as solution to its field equations (1.2).

However, in a recent work McManus[17] has shown that a one-parameter family of solutions of the field equations (1.2) previously found by Ponce de Leon[75] was in fact flat in five dimensions. And yet, the corresponding 4D induced models were shown to be a perfect fluid family of Friedmann–Robertson–Walker curved models (see Refs. 11, 13, 19 and also 76–78, where other Riemann-flat solutions are also discussed).

Therefore a question which naturally arises here is whether the above 5D flat metric, which is the only solution to the IM field equations, can similarly give rise to any 4D *curved* space–time. However, from (5.10) one obviously has that the corresponding 4D space–time is nothing but the Minkowski flat space (this result can also be derived by using a computer algebra package as, e.g., CLASSI[71,69] to calculate the 4D curvature tensor for $m^2=\omega=0$). In brief, the only solution of the

IM field equations (1.2) of generalized Gödel-type is the 5D flat space (5.7), which gives rise only to the 4D Minkowski (flat) space–time, whose topology can be taken to be the simply connected Euclidean $\mathbb{R}^5$, in which no closed timelike curves are permitted.

Although the above results can be looked upon as if the induced matter theory works as an effective therapy for the causal anomalies which arise when one starts from the specific generalized 5D Gödel-type family of metrics (5.1), this does not ensure that the induced matter version of general relativity is an efficient treatment for the causal anomalies (solutions with closed timelike curves) in general relativity as it has been conjectured in Ref. 65. Actually, in a recent paper (which unfortunately was not initially noticed by Rebouças and Teixeira[65]) Romero *et al.*[67] (see also Ref. 79) have shown that the induced matter 5D scheme is indeed general enough to locally generate all solutions to 4D Einstein field equations. This is ensured by a theorem due to Campbell[80] which states that any analytic *n*-dimensional Riemannian space can be locally embedded in a $(n+1)$-dimensional Ricci-flat space. In our context this amounts to saying that there must exist a five-dimensional Ricci-flat space which locally gives rise to the 4D Gödel noncausal solution of Einstein equations of general relativity. Thus, what still remains to be done regarding Gödel-type spaces is to find out this 5D Ricci-flat space which gives rise (locally) to the 4D Gödel-type space–times of general relativity.

To conclude, it is worth stressing some features of the local underlying embedding of the induced matter theory. Any Riemann-flat manifold obviously is also Ricci-flat. The reverse, however, does not necessarily hold, and one can have Ricci-flat spaces which are not Riemann-flat. For the generalized 5D Gödel-type geometries we have discussed in this paper the condition for Ricci-flatness ($\hat{R}_{AB}=0$) necessarily leads to Riemann-flat spaces. Remarkably many solutions of the field equations (1.2) are indeed Riemann-flat (see Refs. 11, 17, 19, and 75–78). From a purely mathematical 5D point of view all Riemann-flat spaces are locally equivalent (locally isometric). However, from the viewpoint of the 5D induced matter gravity all the above-referred 5D Riemann-flat solutions give rise to physically (and geometrically) distinct 4D space–times.[11,17,19,75–78] On the other hand, in the light of the equivalence problem techniques we have discussed in Sec. II, these 5D Riemann-flat examples also show that all 5D Cartan scalars (2.3) can vanish identically, with or without the vanishing of the corresponding (induced) 4D Cartan scalars.

ACKNOWLEDGMENT

The authors are grateful to the scientific agency CNPq for the grants under which this work was carried out. M.J.R. thanks C. Romero, V. B. Bezerra, and J. B. Fonseca-Neto for motivating and fruitful discussions.

[1] A. Einstein, *The Meaning of Relativity* (Princeton University Press, Princeton, 1956).
[2] J. A. Wheeler, *Einstein's Vision* (Springer-Verlag, Berlin, 1968).
[3] A. Salam, Rev. Mod. Phys. **52**, 525 (1980).
[4] P. S. Wesson, Gen. Relativ. Gravit. **22**, 707 (1990).
[5] P. S. Wesson and J. Ponce de Leon, J. Math. Phys. **33**, 3883 (1992).
[6] J. M. Overduin and P. S. Wesson, Phys. Rep. **283**, 303 (1997).
[7] P. Ponce de Leon and P. S. Wesson, J. Math. Phys. **34**, 4080 (1993).
[8] P. S. Wesson, J. Ponce de Leon, P. H. Lim, and H. Liu, Int. J. Mod. Phys. D **2**, 163 (1993).
[9] P. S. Wesson, Mod. Phys. Lett. A **10**, 15 (1995).
[10] S. Rippl, C. Romero, and R. Tavakol, Class. Quant. Grav. **12**, 2411 (1995).
[11] P. S. Wesson, J. Ponce de Leon, H. Liu, B. Mashhoon, D. Kalligas, C. W. F. Everitt, A. Billyard, P. H. Lim, and J. M. Overduin, Int. J. Mod. Phys. A **11**, 3247 (1996).
[12] P. S. Wesson, Mod. Phys. Lett. A **7**, 921 (1992).
[13] P. S. Wesson, Astrophys. J. **394**, 19 (1992).
[14] P. S. Wesson and H. Liu, Astrophys. J. **440**, 1 (1995).
[15] P. S. Wesson and J. Ponce de Leon, Astron. Astrophys. **294**, 1 (1995).
[16] H. Liu and P. S. Wesson, Int. J. Mod. Phys. D **3**, 627 (1994).
[17] D. J. Mc Manus, J. Math. Phys. **35**, 4889 (1994).
[18] A. A. Coley, Astrophys. J. **427**, 585 (1994).
[19] A. A. Coley and D. J. Mc Manus, J. Math. Phys. **36**, 335 (1995).
[20] P. S. Wesson and J. Ponce de Leon, Gen. Relativ. Gravit. **26**, 555 (1994).

[21] B. Mashhoon, H. Liu, and P. S. Wesson, Phys. Lett. B **331**, 305 (1994).
[22] P. S. Wesson, Phys. Lett. B **276**, 299 (1992).
[23] H. Liu and P. S. Wesson, J. Math. Phys. **33**, 3888 (1992).
[24] P. S. Wesson and J. Ponce de Leon, Class. Quantum Grav. **11**, 1341 (1994).
[25] H. Liu and P. S. Wesson, Class. Quantum Grav. **13**, 2311 (1996).
[26] A. Billyard and P. S. Wesson, Phys. Rev. D **53**, 731 (1996).
[27] P. H. Lim and P. S. Wesson, Astrophys. J. **397**, L91 (1992).
[28] D. Kalligas, P. S. Wesson, and C. W. F. Everitt, Astrophys. J. **439**, 548 (1995).
[29] P. H. Lim, J. M. Overduin, and P. S. Wesson, J. Math. Phys. **36**, 6907 (1995).
[30] P. S. Wesson, H. Liu, and P. H. Lim, Phys. Lett. B **298**, 69 (1993).
[31] H. Liu, P. S. Wesson, and J. Ponce de Leon, J. Math. Phys. **34**, 4070 (1993).
[32] H. Liu and P. S. Wesson, Phys. Lett. B **381**, 420 (1996).
[33] K. Gödel, Rev. Mod. Phys. **21**, 447 (1949).
[34] M. M. Som and A. K. Raychaudhuri, Proc. R. Soc. London, Ser. A **304**, 81 (1968).
[35] A. Banerjee and S. Banerji, J. Phys. A **1**, 188 (1968).
[36] F. Bampi and C. Zordan, Gen. Relativ. Gravit. **9**, 393 (1978).
[37] M. Novello and M. J. Rebouças, Phys. Rev. D **19**, 2850 (1979).
[38] M. J. Rebouças, Phys. Lett. A **70**, 161 (1979).
[39] A. K. Raychaudhuri and S. N. G. Thakurta, Phys. Rev. D **22**, 802 (1980).
[40] S. K. Chakraborty, Gen. Relativ. Gravit. **12**, 925 (1980).
[41] M. J. Rebouças and J. Tiomno, Phys. Rev. D **28**, 1251 (1983).
[42] A. F. F. Teixeira, M. J. Rebouças, and J. E. Åman, Phys. Rev. D **32**, 3309 (1985). Note that in this reference there are two misprints, namely: in eq. (5) the term $-R_z$ should read $+R_z$, and in eq. (38) the term $\kappa_3 mD$ should read $\kappa_3 m^2 D$.
[43] M. J. Rebouças and A. F. F. Teixeira, Phys. Rev. D **34**, 2985 (1986).
[44] M. J. Rebouças, J. E. Åman, and A. F. F. Teixeira, J. Math. Phys. **27**, 1370 (1986).
[45] M. J. Rebouças and J. E. Åman, J. Math. Phys. **28**, 888 (1987).
[46] F. M. Paiva, M. J. Rebouças, and A. F. F. Teixeira, Phys. Lett. A **126**, 168 (1987).
[47] K. Dunn, Gen. Relativ. Gravit. **21**, 137 (1989).
[48] R. X. Saibatalov, Gen. Relativ. Gravit. **7**, 697 (1995).
[49] M. Rooman and P. Spindel, Class. Quantum Grav. **15**, 3241 (1998).
[50] M. Tsamparlis, D. Nikolopoulos, and P. S. Apostolopoulos, Class. Quantum Grav. **15**, 2909 (1998).
[51] A. M. Candela and M. Sánchez, ''Geodesic Connectedness in Gödel-type Space-times'' Report 17/98, Dipartimento Interuniversitario di Matematica, Università degli Studi—Politecnico di Bari (1998).
[52] A. Krasiński, J. Math. Phys. **39**, 2148 (1998). This reference contains a quite good overview of the literature on rotating models in general relativity.
[53] E. P. V. Vaidya, M. L. Bedran, and M. M. Som, Prog. Theor. Phys. **72**, 857 (1984).
[54] L. L. Smalley, Phys. Rev. D **32**, 3124 (1985).
[55] J. D. Oliveira, A. F. F. Teixeira, and J. Tiomno, Phys. Rev. D **34**, 3661 (1986).
[56] L. L. Smalley, Phys. Lett. A **113**, 463 (1986).
[57] W. M. Silva, Jr., J. Math. Phys. **32**, 3223 (1991).
[58] A. J. Accioly and A. T. Gonçalves, J. Math. Phys. **28**, 1547 (1987).
[59] A. J. Accioly and G. E. A. Matsas, Phys. Rev. D **38**, 1083 (1988).
[60] T. Singh and A. K. Agrawal, Fortschr. Phys. **42**, 71 (1994).
[61] J. D. Barrow and M. P. Dabrowski, Phys. Rev. D **58**, 103502 (1998).
[62] J. E. Åman, J. B. Fonseca-Neto, M. A. H. MacCallum, and M. J. Rebouças, Class. Quantum Grav. **15**, 1089 (1998).
[63] J. B. Fonseca-Neto and M. J. Rebouças, Gen. Relativ. Gravit. **30**, 1301 (1998).
[64] M. J. Rebouças and A. F. F. Teixeira, J. Math. Phys. **39**, 2180 (1998).
[65] M. J. Rebouças and A. F. F. Teixeira, Int. J. Mod. Phys. A **13**, 3181 (1998).
[66] E. Cartan, *Leçons sur la Géométrie des Éspaces de Riemann* (Gauthier-Villars, Paris, 1928, 2nd ed. 1946, reprinted 1951). English translation by J. Glazebrook (Math Sci Press, Brookline, MA, 1983).
[67] C. Romero, R. Tavakol, and R. Zalaletdinov, Gen. Relativ. Gravit. **3**, 365 (1996).
[68] A. Karlhede, Gen. Relativ. Gravit. **12**, 693 (1980).
[69] M. A. H. MacCallum and J. E. F. Skea, ''SHEEP: A Computer Algebra System for General Relativity,'' in *Algebraic Computing in General Relativity, Lecture Notes from the First Brazilian School on Computer Algebra*, edited by M. J. Rebouças and W. L. Roque (Oxford U.P., Oxford 1994), Vol. II. See also references therein.
[70] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge U.P., Cambridge, 1973).
[71] J. E. Åman, ''Manual for CLASSI: Classification Programs for Geometries in General Relativity,'' Institute of Theoretical Physics Technical Report, 1987. Third provisional edition. Distributed with SHEEP sources.
[72] The integration of the Killing equations for these classes of generalized Gödel-type manifolds, which will be performed in the next section, shows that the isotropy group indeed depends on these special relations between the essential parameters ω, k, and m^2.
[73] A. Z. Petrov, *Einstein Spaces*, first English edition (Pergamon, New York, 1969). See p. 63.
[74] M. O. Calvão, M. J. Rebouças, A. F. F. Teixeira, and W. M. Silva, Jr., J. Math. Phys. **29**, 1127 (1988).
[75] J. Ponce de Leon, Gen. Relativ. Gravit. **20**, 539 (1988).
[76] G. Abolghasem, A. A. Coley, and D. J. Mc Manus, J. Math. Phys. **37**, 361 (1996).

[77] A. Billyard and P. S. Wesson, Gen. Relativ. Gravit. **28**, 137 (1996).
[78] H. Liu and P. S. Wesson, Gen. Relativ. Gravit. **30**, 509 (1998).
[79] J. E. Lidsey, C. Romero, R. Tavakol, and S. Rippl, Class. Quantum Grav. **14**, 865 (1997).
[80] J. E. Campbell *A Course of Differential Geometry* (Clarendon, Oxford, 1926).

Geodesic completeness of orthogonally transitive cylindrical space–times

L. Fernández-Jambrina[a)]
Departamento de Enseñanzas Básicas de la Ingeniería Naval, E.T.S.I. Navales, Arco de la Victoria s/n, E-28040 Madrid, Spain

(Received 20 January 1999; accepted for publication 4 May 1999)

In this paper a theorem is derived in order to provide a wide sufficient condition for an orthogonally transitive cylindrical space–time to be singularity free. The applicability of the theorem is tested on examples provided by the literature that are known to have regular curvature invariants. © *1999 American Institute of Physics.* [S0022-2488(99)02708-5]

I. INTRODUCTION

The issue of establishing whether a Lorentzian manifold is geodesically complete is not in principle a simple one since there is no Hopf–Rinow theorem that could settle the matter as it happens in the Riemannian case. One could think that regularity of the curvature invariants might be helpful, but there are known examples of space–times with regular invariants, such as Taub-NUT,[1] that enclose geodesics that are not complete in their affine parametrization due to a phenomenon called imprisoned incompleteness.

Taking completeness of causal geodesics (g completeness) as a definition of the absence of singularities (no observer in free fall leaves the space–time in a finite proper time), one can resort to many theorems in the literature (cf. Refs. 1 and 2, and references therein) in order to determine whether a space–time is singular. But on the contrary, theorems that provide large families of nonsingular space–times are not very usual[3,4] and in principle the proof of geodesic completeness involves cumbersome calculations.[5]

Instead of dealing with general Lorentzian manifolds, we shall approach orthogonally transitive cylindrical space–times[6] since they have provided many examples of regular manifolds (cf. Refs. 7 and 8) in inhomogeneous cosmology.[9]

Our aim will be the generalization of the theorem on diagonal orthogonally transitive cylindrical space–times in Ref. 4 to nondiagonal models and thereby comprise all known nonsingular cylindrical perfect fluid space–times in the literature.

First we shall show that the second-order system of geodesic equations can be reduced by the use of constants of motion to three first-order equations plus two quadratures. This fact will simplify the analysis of the prolongability of the geodesics and will enable us to write a sufficient condition for completeness of orthogonally transitive cylindrical space–times in a theorem.

II. GEODESIC EQUATIONS

We shall write the metric of an orthogonally transitive cylindrical space–time in a chart using isotropic coordinates t, r for the subspace orthogonal to the orbits of the isometry group and coordinates ϕ, z adapted to the commuting generators of the group of isometries. The metric shall be determined by four functions, g, f, A, ρ, of the coordinates t, r,

$$ds^2=e^{2g(t,r)}\{-dt^2+dr^2\}+\rho^2(t,r)e^{2f(t,r)}d\phi^2+e^{-2f(t,r)}\{dz+A(r,t)d\phi\}^2 \tag{1}$$

and we shall assume that these functions are C^2 in their range,

[a)]Electronic mail: lfernandez@etsin.upm.es

$$-\infty<t,z<\infty, \quad 0<r<\infty, \quad 0<\phi<2\pi. \tag{2}$$

The axis will be located where the norm of the axial Killing field vanishes,

$$0=\Delta=g(\xi,\xi)=\rho^2(t,r)e^{2f(t,r)}+e^{-2f(t,r)}A^2(r,t), \tag{3}$$

which means that both A and ρ must vanish on the axis, since f is a smooth function.

Since the choice of isotropic coordinates is not unique, we can take advantage of this freedom to have $r=0$ as the equation for the axis. In order to avoid conical singularities, the usual requirement[10] will be imposed in order to have a well-defined axis,

$$\lim_{r\to 0}\frac{g(\mathrm{grad}\,\Delta,\mathrm{grad}\,\Delta)}{4\Delta}=1. \tag{4}$$

Denoting by a dot differentiation with respect to the affine parameter, two of the four second-order geodesic equations,

$$\ddot{x}^i+\Gamma^i_{jk}\dot{x}^j\dot{x}^k=0, \tag{5}$$

can be integrated, taking into account that there are two first integrals of the geodesic motion associated with the generators of the isometries. These are the angular momentum around the axis, L, and the linear momentum along the axis, P, of a test particle of unit mass,

$$L=e^{2f(t,r)}\rho^2(t,r)\dot{\phi}+e^{-2f(t,r)}A(t,r)\{\dot{z}+A(t,r)\dot{\phi}\}, \tag{6}$$

$$P=e^{-2f(t,r)}\{\dot{z}+A(t,r)\dot{\phi}\}. \tag{7}$$

The affine parametrization is determined, up to an affinity of the real line, by the prescription

$$\delta=e^{2g(t,r)}\{\dot{t}^2-\dot{r}^2\}-\{L-PA(t,r)\}^2\rho^{-2}(t,r)e^{-2f(t,r)}-P^2e^{2f(t,r)}, \tag{8}$$

where δ is one for timelike, zero for null, and minus one for spacelike geodesics. Since we are dealing just with causal geodesics, for our purposes δ will always be positive. After writing $\dot{z},\dot{\phi}$ as functions of L and P, the second-order equations in t and r,

$$\begin{aligned}\ddot{t}&+g_t(t,r)\dot{t}^2+2g_r(t,r)\dot{t}\dot{r}+g_t(t,r)\dot{r}^2-P^2e^{2\{f(t,r)-g(t,r)\}}f_t(t,r)\\&+e^{-2\{f(t,r)+g(t,r)\}}\frac{\{L-PA(t,r)\}^2}{\rho^2(t,r)}\left\{\frac{\rho_t(t,r)}{\rho(t,r)}+f_t(t,r)+\frac{PA_t(t,r)}{L-PA(t,r)}\right\},\end{aligned} \tag{9}$$

$$\begin{aligned}\ddot{r}&+g_r(t,r)\dot{t}^2+2g_t(t,r)\dot{t}\dot{r}+g_r(t,r)\dot{r}^2+P^2e^{2\{f(t,r)-g(t,r)\}}f_r(t,r)\\&-e^{-2\{f(t,r)+g(t,r)\}}\frac{\{L-PA(t,r)\}^2}{\rho^2(t,r)}\left\{\frac{\rho_r(t,r)}{\rho(t,r)}+f_r(t,r)+\frac{PA_r(t,r)}{L-PA(t,r)}\right\},\end{aligned} \tag{10}$$

can be rearranged in a form that will be useful afterwards,

$$\{e^{2g(t,r)}\dot{t}\}^{\cdot}-e^{-2g(t,r)}F(t,r)F_t(t,r)=0, \tag{11}$$

$$\{e^{2g(t,r)}\dot{r}\}^{\cdot}+e^{-2g(t,r)}F(t,r)F_r(t,r)=0, \tag{12}$$

$$F(t,r)=e^{g(t,r)}\sqrt{\delta+P^2e^{2f(t,r)}+\{L-PA(t,r)\}^2\frac{e^{-2f(t,r)}}{\rho^2(t,r)}}, \tag{13}$$

which have the same structure as in the diagonal case.

If at least one of the constants L, P, δ is different from zero, the system is equivalent to three equations of first order for future-pointing (past-pointing) geodesics,

$$\dot{t}=\pm e^{-2g(t,r)}F(t,r)\cosh\xi(t,r), \tag{14}$$

$$\dot{r}=e^{-2g(t,r)}F(t,r)\sinh\xi(t,r), \tag{15}$$

$$\dot{\xi}(t,r)=-e^{-2g(t,r)}\{\pm F_t(t,r)\sinh\xi(t,r)+F_r(t,r)\cosh\xi(t,r)\}, \tag{16}$$

parametrizing (8) by a function $\xi(t,r)$. More explicitly, the last equation takes the form,

$$\dot{\xi}=-\frac{e^{-g}}{\sqrt{\delta+\Lambda^2\rho^{-2}e^{-2f}+P^2e^{2f}}}\left\{\cosh\xi\left(\delta g_r+\Lambda^2\frac{e^{-2f}}{\rho^2}h_r+P^2e^{2f}q_r\right)\right.$$
$$\left.\pm\sinh\xi\left(\delta g_t+\Lambda^2\frac{e^{-2f}}{\rho^2}h_t+P^2e^{2f}q_t\right)\right\}, \tag{17}$$

$$h=g-f-\ln\rho+\ln|\Lambda|,\quad \Lambda=L-PA,\quad q=g+f, \tag{18}$$

which will be useful for deriving prolongability conditions for causal geodesics. The minus (plus) sign corresponds to past-pointing (future-pointing) geodesics.

Note that the general equations are obtained from those of the diagonal case just replacing L by Λ, which can be therefore considered as a sort of ''effective angular momentum'' in the case where the Killing fields are not orthogonal.

III. PROLONGABILITY OF THE GEODESICS

In this section we shall introduce two theorems on causal geodesic completeness of orthogonally transitive cylindrical space–times. Null coordinates,

$$u=\frac{t+r}{2},\quad v=\frac{t-r}{2}, \tag{19}$$

will play an important role in the results.

Theorem 1: An orthogonally transitive cylindrical space–time endowed with a metric whose local expression in terms of C^2 metric functions f, g, A, ρ is given by (1) such that the axis located at $r=0$ has complete future causal geodesics if the following set of conditions is fulfilled:

(1) For large values of t and increasing r,

$$\text{(a)}\quad \begin{cases} g_u\geqslant 0\\ h_u\geqslant 0\\ q_u\geqslant 0,\end{cases}$$

$$\text{(b) either}\quad \begin{cases} g_r\geqslant 0 \ \text{ or } \ |g_r|\lesssim g_u\\ h_r\geqslant 0 \ \text{ or } \ |h_r|\lesssim h_u\\ q_r\geqslant 0 \ \text{ or } \ |q_r|\lesssim q_u\,.\end{cases}$$

(2) For $L\neq 0$ and large values of t and decreasing r,

$$\text{(a)}\quad \delta g_v+P^2e^{2f}q_v+\Lambda^2\frac{e^{-2f}}{\rho^2}h_v\geqslant 0,$$

(b) either $\delta g_r + P^2 e^{2f} q_r + \Lambda^2 \frac{e^{-2f}}{\rho^2} h_r \leqslant 0$ or $\delta g_r + P^2 e^{2f} q_r$

$$+ \Lambda^2 \frac{e^{-2f}}{\rho^2} h_r \lesssim \delta g_v + P^2 e^{2f} q_v + \Lambda^2 \frac{e^{-2f}}{\rho^2} h_v .$$

(3) For large values of the time coordinate t, constants a, b exist such that

$$\text{that} \quad \left.\begin{matrix} 2g(t,r) \\ g(t,r)+f(t,r)+\ln\rho-\ln|\Lambda| \\ g(t,r)-f(t,r) \end{matrix}\right\} \geqslant -\ln|t+a|+b.$$

(4) The limit $\lim_{r\to 0}(A/\rho)$ exists.

A theorem can be obtained for past-pointing geodesics just changing the sign of the time derivatives in the previous one.

Theorem 2: An orthogonally transitive cylindrical space–time endowed with a metric whose local expression in terms of C^2 metric functions f, g, A, ρ is given by (1) such that the axis is located at $r=0$ has complete past causal geodesics if the following set of conditions is fulfilled:

(1) For small values of t and increasing r,

$$(\mathrm{a}) \quad \begin{cases} g_v \leqslant 0 \\ h_v \leqslant 0 \\ q_v \leqslant 0, \end{cases}$$

$$(\mathrm{b}) \text{ either } \begin{cases} g_r \geqslant 0 \text{ or } |g_r| \lesssim -g_v \\ h_r \geqslant 0 \text{ or } |h_r| \lesssim -h_v \\ q_r \geqslant 0 \text{ or } |q_r| \lesssim -q_v . \end{cases}$$

(2) For $L \neq 0$ and small values of t and decreasing r,

$$(\mathrm{a}) \quad \delta g_u + P^2 e^{2f} q_u + \Lambda^2 \frac{e^{-2f}}{\rho^2} h_u \leqslant 0$$

(b) either $\delta g_r + P^2 e^{2f} q_r + \Lambda \frac{2e^{-2f}}{\rho^2} h_r \leqslant 0$ or $\delta g_r + P^2 e^{2f} q_r$

$$+ \Lambda^2 \frac{e^{-2f}}{\rho^2} h_r \lesssim \left| \delta g_u + P^2 e^{2f} q_u + \Lambda^2 \frac{e^{-2f}}{\rho^2} h_u \right| .$$

(3) For small values of the time coordinate t, constants a, b exist such that

$$\text{that} \quad \left.\begin{matrix} 2g(t,r) \\ g(t,r)+f(t,r)+\ln\rho-\ln|\Lambda| \\ g(t,r)-f(t,r) \end{matrix}\right\} \geqslant -\ln|t+a|+b.$$

(4) The limit $\lim_{r\to 0}(A/\rho)$ exists.

The theorems in Ref. 4 can be seen to be subcases of the ones introduced here.

IV. PROOF OF THE THEOREMS

In order to achieve prolongability of causal geodesics we just have to impose that $\dot{t}$ remains finite for finite values of the affine parameter. The radial velocity, $\dot{r}$, need not be considered since it cannot become singular if $\dot{t}$ is not singular too. The other derivatives, $\dot{z}$ and $\dot{\phi}$, are quadratures of smooth functions of t and r and therefore they only may turn singular if t or r become so. We shall focus on future-pointing geodesics. The analysis for past-pointing geodesics is entirely similar.

A way of preventing the hyperbolic functions of ξ from becoming singular is to require that ξ does not grow indefinitely. Therefore for large values of ξ, $\dot{\xi}$ must eventually become negative. Since the constants, L, P, δ may vanish independently, one is not to expect compensations between them. Hence their respective terms in (17) have to become negative for large values of positive ξ as it is stated in conditions (1a) and (1b), taking into account that $\cosh\xi = \sinh\xi + e^{-\xi}$ and that therefore the terms in the negative exponential of ξ (1b) need not be negative but just of the same order as those in (1a).

By imposing condition (4) in the theorem, we require that the geometry of the space–time in the vicinity of the axis is determined by ρ and not by A. Hence for negative decreasing ξ the axis could be singular only for geodesics with $L \neq 0$, since the terms A/ρ are finite at the axis. Geodesics with zero angular momentum just cross $r=0$ and reappear with positive ξ and polar angle $\phi+\pi$ and hence need not be taken into account. In this case we can therefore allow compensations between the nonzero L term and the other ones. Splitting $\cosh\xi$ as $e^{\xi}-\sinh\xi$, the condition for $\dot{\xi}$ to become positive for large values of t and negative ξ is stated in (2a) for the terms in $\sinh\xi$ and in (2b) for the terms in e^{ξ}, that can be at most of the same order as the former ones since they are exponentially damped.

No further conditions need be imposed on $\dot{\xi}$. But $\dot{t}$ could turn singular also for the $e^{-2g}F$ term. A way of preventing it is to impose a growth slower than linear for $\dot{t}$ due to each of the three terms (δ,Λ,P) for large values of t. This is done in condition (3).

The condition on g must be refined since we have not yet considered the geodesics that cannot be parametrized by ξ. These are those with zero F, that is, null geodesics with zero $\dot{z}$ and $\dot{\phi}$. Since $\dot{t}=|\dot{r}|$ for such geodesics, the equations of geodesic motion,

$$\ddot{t}+g_t\dot{t}^2+2g_r\dot{t}\dot{r}+g_t\dot{r}^2=0, \tag{20}$$

$$\ddot{r}+g_r\dot{t}^2+2g_t\dot{t}\dot{r}+g_r\dot{r}^2=0, \tag{21}$$

can be reduced to a single one that can be integrated,

$$\ddot{t}+2g_t\dot{t}^2+2g_r\dot{t}\dot{r}=0 \Rightarrow (e^{2g}\dot{t})^{\cdot}=K, \tag{22}$$

which, in order to have t extendible to arbitrary values of the affine parameter, can be controlled by imposing at most linear growth for $\dot{t}$ as it is done in condition (3).

V. COMPLETENESS OF SEVERAL CYLINDRICAL MODELS

Since all known diagonal cylindrical perfect fluid models (Refs. 11–14) with regular curvature invariants have already been shown to be geodesically complete,[4] we shall only be concerned about the nondiagonal ones. To our knowledge there are just two, and both can be derived from Einstein space–times using the Wainwright–Ince–Marshman generation algorithm for stiff perfect fluids.[15]

(1) Mars:[16] It is the first known nonsingular nondiagonal cylindrical cosmological model in the literature. In another context it was previously published by Letelier.[17] In isotropic coordinates the metric functions can be written as

$$g(t,r)=\tfrac{1}{2}\ln\cosh(2at)+\tfrac{1}{2}\alpha a^2r^2, \quad f(t,r)=\tfrac{1}{2}\ln\cosh(2at),$$
$$\rho(t,r)=r, \quad A(t,r)=ar^2, \tag{23}$$

where a is a constant and $\alpha>1$. If $\alpha=1$ the pressure and the density of the fluid vanish.

All functions are even in t and therefore the analysis of past-pointing geodesics is equivalent to that of future-pointing ones and shall be omitted. The derivatives g_u, q_u in condition (1a) are positive for positive t whereas h_u also needs large radial coordinate.

The derivatives g_r, q_r in condition (1b) are positive regardless of t and h_r is positive for large r.

Conditions (2a), (2b) are satisfied when r is small and t is positive since the L term is dominant for small r.

The functions in condition (3) are all positive except for the $\ln r$ term in h when r decreases. But this can be bounded by a logarithm of t and therefore the condition is fulfilled. The ratio A/ρ tends to zero for decreasing r and hence condition (4) in theorem 1 is satisfied. Hence this space–time is causally g complete.

(2) Griffiths–Bičák:[18] The previous model is comprised in this one for $c=0$ after a redefinition of constants. The metric functions can be written as

$$g(t,r)=\tfrac{1}{2}\ln\cosh(2at)+\tfrac{1}{2}a^2r^2+\tfrac{1}{2}\Omega(t,r),\quad f(t,r)=\tfrac{1}{2}\ln\cosh(2at),$$
$$\rho(t,r)=r,\quad A(t,r)=ar^2, \tag{24}$$

where Ω is a function that is obtained from a solution, σ, of the wave equation

$$\Omega_r=r(\sigma_t^2+\sigma_r^2),\quad \Omega_t=2r\sigma_t\sigma_r,$$
$$\sigma(t,r)=bt+\sqrt{2}c\sqrt{\frac{\sqrt{(\alpha^2+r^2-t^2)^2+4\alpha^2t^2}+\alpha^2+r^2-t^2}{(\alpha^2+r^2-t^2)^2+4\alpha^2t^2}}. \tag{25}$$

The analysis of the geodesics of this space–time can be dealt with easily since

$$\Omega_u=r(\sigma_t+\sigma_r)^2,\quad \Omega_v=-r(\sigma_t-\sigma_r)^2, \tag{26}$$

and therefore Ω contributes with an additional positive term to conditions (1a), (1b) in theorem 1, that were already checked for Mars space–time. On the contrary, Ω adds a negative term in conditions (2a), (2b) but it is negligible for small r.

Finally, the contribution of Ω to g grows at most as $\sqrt{t}$ for large t and is therefore negligible compared to the other terms in condition (3).
A similar reasoning is valid for theorem 2, although the metric function Ω is not even in time.

Hence these space–times are geodesically complete.

ACKNOWLEDGMENTS

The present work has been supported by Dirección General de Enseñanza Superior Project No. PB95-0371. The author wishes to thank F. J. Chinea, L. M. González-Romero, and M. J. Pareja for valuable discussions.

[1] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge University Press, Cambridge, 1973).
[2] J. Beem, P. Ehrlich, and K. Easley, *Global Lorentzian Geometry* (Dekker, New York, 1996).
[3] A. Romero and M. Sánchez, Geom. Dedicata **53**, 103 (1994).
[4] L. Fernández-Jambrina and L. M. González-Romero, Class. Quantum Grav. **16**, 953 (1999).
[5] F. J. Chinea, L. Fernández-Jambrina, and J. M. M. Senovilla, Phys. Rev. D **45**, 481 (1992).
[6] J. Wainwright, J. Phys. A **14**, 1131 (1981).
[7] *El Escorial Summer School on Gravitation and General Relativity 1992: Rotating Objects and Other Topics*, edited by F. J. Chinea and L. M. González-Romero (Springer, Berlin, 1993).
[8] J. M. M. Senovilla, Gen. Relativ. Gravit. **30**, 701 (1998).
[9] A. Krasiński, *Inhomogeneous Cosmological Models* (Cambridge University Press, Cambridge, 1997).
[10] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field equations* (Cambridge University Press, Cambridge, 1980).
[11] J. M. M. Senovilla, Phys. Rev. Lett. **64**, 2219 (1990).

[12] E. Ruiz and J. M. M. Senovilla, Phys. Rev. D **45**, 1995 (1992).
[13] M. Mars, Class. Quantum Grav. **12**, 2831 (1995).
[14] L. Fernández-Jambrina, Class. Quantum Grav. **14**, 3407 (1997).
[15] J. Wainwright, W. C. W. Ince, and B. J. Marshman, Gen. Relativ. Gravit. **10**, 259 (1979).
[16] M. Mars, Phys. Rev. D **51**, R3989 (1995).
[17] P. S. Letelier, J. Math. Phys. **20**, 2078 (1979).
[18] J. B. Griffiths and J. Bičak, Class. Quantum Grav. **12**, L81 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Canonical angular supermomentum tensors in general relativity

Janusz Garecki[a)]
Institute of Physics, University of Szczecin, Wielkopolska 15, 70-451 Szczecin, Poland

(Received 9 April 1998; accepted for publication 9 December 1998)

We introduce *the canonical supermomentum tensors* ${}_mS^{ikl}(P;v^t)$ and ${}_gS^{ikl}(P;v^t)$ for matter and gravitation, respectively. These tensors give kinds of the *quasilocal quantities* which have recently been considered in the framework of general relativity though usually constructed in some special cases and for the mass only. Our method to construct the quasilocal quantities (i.e., superenergy and supermomentum tensors) is very general and does not require introducing any supplementary elements into general relativity. It merely *extracts covariant data* from the suitable canonical objects which exist in the framework of the standard general relativity. Having given a constructive definition of the canonical supermomentum tensors for gravitation and matter, we apply the canonical gravitational supermomentum tensor ${}_gS^{ikl}(P;v^t)$ for the local analysis of the *plane and plane-fronted* gravitational waves. Next, we apply the canonical supermomentum tensors ${}_gS^{ikl}(P;v^t)$ and ${}_mS^{ikl}(P;v^t)$, gravitation and matter, to local and global analysis of the Friedman, Schwarzschild, and Kerr space–times. We show that the canonical angular supermomentum tensors and integral angular supermomenta *have better geometric and physical properties* than the canonical objects which describe *densities* of the angular momenta and integral quantities determined by these densities. © *1999 American Institute of Physics.* [S0022-2488(99)02907-2]

I. INTRODUCTION

It is well known that in the framework of the *standard*[1] general relativity (**GR**) the gravitational field has *nontensorial strength* $\Gamma^i_{kl}=\{{}^{\,i}_{kl}\}$ as well as it has *no energy-momentum tensor*. One can only attribute to this field the so-called *gravitational energy-momentum pseudotensors*. The best of the possible gravitational energy-momentum pseudotensors seems to be the *canonical gravitational energy-momentum pseudotensor* ${}_Et_i^{\ k}$ proposed by Einstein.[2,3]

Some years ago[4,5] we introduced the *canonical gravitational superenergy tensor* into **GR** to avoid difficulties connected with a lack of any gravitational energy-momentum tensor in the theory. This tensor was generalized to matter fields, too.[4,5]

The superenergy tensors of gravitation and matter give a kind of so-called ''quasilocal quantities'' (see, e.g., Refs. 6–10) usually constructed in some particular cases and for mass (or energy) only. However, these ''quasilocal quantities'' usually demand introduction of some supplementary structures, apart from the metric, into **GR**.

Our method of construction of the ''quasilocal quantities''—the superenergy (and supermomentum) tensors—is very general and does not require introducing of any supplementary elements into **GR**. It is a generalization of an early idea by Pirani[11] and it only *extracts covariant data* from the suitable canonical objects which exist in the framework of the standard[1] **GR**.

In Refs. 12 and 13 the gravitational, canonical superenergy tensor ${}_gS_i^{\ k}(P;v^l)$ has been used to the local analysis of the *cylindrical, plane, and plane-fronted gravitational waves* and the canonical superenergy tensors ${}_mS_a^{\ b}(P;v^l)$ and ${}_gS_a^{\ b}(P;v^l)$ of matter and gravitation were used in Refs.

a)Electronic mail: garecki@wmf.univ.szczecin.pl

4,5,14,15 for local and global superenergetic analysis of the Friedman cosmological models and the Schwarzschild space–time.

In this paper we want to generalize the idea of the canonical superenergy tensors onto *angular momentum* and introduce *the canonical angular supermomentum tensors*.

As it is known,[16] the situation with the angular momentum in **GR** is much more difficult and obscure than with the energy-momentum.[17] For example, even matter field *does not possess any angular momentum tensor density* because coordinates $\{x^i\}$ *do not form* any radius vector in some general components.

The canonical angular supermomentum tensors which we introduce in this paper are *our proposal* for the nonexistent in **GR** angular momentum tensors, matter and gravitation.

Our convention is the same as in our previous paper,[15] i.e., we use the metric signature $(+---)$ and the Latin indices run over the values 0, 1, 2, 3; ∇_i means covariant derivative and ∂_i or $_{,i}$ denote partial derivative $\partial/\partial x^i$; and $:=$ means ''by definition.'' Round brackets mean symmetrization and square brackets mean alternation, i.e., $(a|c|b)=\frac{1}{2}(acb+bca)$ and $[a|c|b]=\frac{1}{2}[acb-bca]$. Our convention for the Riemann curvature and Ricci tensors and for the Einstein equations,

$$G_{ik}:=R_{ik}-\frac{1}{2}g_{ik}R=\frac{8\pi G}{c^4}T_{ik}=:\beta T_{ik},$$

follow the standard textbooks.[3]

In Sec. II we present our constructive definition of the canonical angular supermomentum tensors. This definition uses normal coordinates[18–20] and it is a straightforward generalization of our definition of the canonical superenergy tensors given in final form in Ref. 15. We calculate the analytic form of the canonical supermomentum tensors of gravitation ${}_gS^{abc}(P;v^l)$ and matter ${}_mS^{abc}(P;v^l)$.

In Sec. III we give the most natural physical interpretation of tensors ${}_gS^{abc}(P;v^l)$ and ${}_mS^{abc}(P;v^l)$. The physical interpretation follows the line developed in Ref. 15.

In Sec. IV we present some applications of the canonical angular supermomentum tensors ${}_gS^{abc}(P;v^l)$ and ${}_mS^{abc}(P;v^l)$. At first, we apply the canonical, gravitational angular supermomentum tensor ${}_gS^{abc}(P;v^l)$ to local analysis of the plane and plane-fronted gravitational waves. Then, we apply the both canonical angular supermomentum tensors ${}_gS^{abc}(P;v^l)$ and ${}_mS^{abc}(P;v^l)$, gravitation and matter, to analysis of the Friedman and Schwarzschild space–times. Finally, we apply the gravitational angular supermomentum tensor ${}_gS^{abc}(P;v^l)$ to the analysis of the stationary Kerr space–time.

In Sec. V we give conclusions.

In Appendix A we review the Bergmann–Thomson expression for the angular momentum in **GR**. In Appendix B we present the results of our calculations concerning plane and plane-fronted gravitational waves. Appendix C is devoted to the angular supermomentum in special relativity.

II. CANONICAL ANGULAR SUPERMOMENTUM TENSORS IN GENERAL RELATIVITY

In analogy with the definition of the canonical superenergy tensors given in Ref. 15 in the normal coordinates **NCS(P)** we define the *angular supermomentum tensor*

$$S^{(a)(b)(c)}(P):=\lim_{\Omega\to P}\frac{\int_\Omega[M^{(a)(b)(c)}(y)-M^{(a)(b)(c)}(P)]\,d\Omega}{1/2M}$$

$$=(-)\lim_{\Omega\to P}\frac{\int_\Omega[M^{(a)(b)(c)}(y)-M^{(a)(b)(c)}(P)]\,d\Omega}{1/2\int_\Omega\sigma(P;y)\,d\Omega},\qquad(1)$$

where

$$M^{(a)(b)(c)}(y) := M^{ikl}(y) e_i^{(a)}(y) e_k^{(b)}(y) e_l^{(c)}(y), \tag{2}$$

and

$$M^{(a)(b)(c)}(P) := M^{ikl}(P) e_i^{(a)}(P) e_k^{(b)}(P) e_l^{(c)}(P) = M^{ikl}(P)\, \delta_i^a \delta_k^b \delta_l^c = M^{abc}(P) \tag{3}$$

are *physical or tetrad components* of the field $M^{ikl} = (-) M^{kil}$ which describes angular momentum. Here $e^i_{(a)}(y)$, $e_k^{(b)}(y)$ are the components of the orthonormal tetrad $e^i_{(a)}(P) = \delta^i_a$ and its dual $e_k^{(b)}(P) = \delta_k^b$, respectively, such that $e^i_{(a)}(P) e_k^{(b)}(P) = \delta_a^b$ and the tetrads are parallelly propagated along geodesics through **P**. In (1) Ω is a sufficiently small domain which surrounds **P** and it has the following properties:

$$\int_\Omega y^i \, d\Omega = 0, \quad \int_\Omega y^i y^k \, d\Omega = \delta^{ik} M, \tag{4}$$

where

$$M = \int_\Omega (y^0)^2 \, d\Omega = \int_\Omega (y^1)^2 \, d\Omega = \int_\Omega (y^2)^2 \, d\Omega = \int_\Omega (y^3)^2 \, d\Omega \tag{5}$$

is a common value of the moments of inertia of the domain Ω with respect to the subspaces $y^i = 0$ $(i = 0,1,2,3)$.

Of course, we can write M in a covariant form as

$$M = (-) \int_\Omega \sigma(P;y) \, d\Omega, \tag{6}$$

where

$$\sigma(P;y) \doteq \tfrac{1}{2}(y^{0^2} - y^{1^2} - y^{2^2} - y^{3^2})$$

denotes the two-point *world function* introduced by J. L. Synge.[21] We have used this fact in Eq. (1) already. The symbol $\doteq$ means that an equation is valid only in some special coordinates.

For all quantities attached to the point **P** there is *an equality between tetrad and normal components*. We use this and skip tetrad brackets for indices of any quantities attached to the point **P**; for example, we will write $S^{abc}(P)$ instead of $S^{(a)(b)(c)}(P)$ and so on.

For matter, as ${}_mM^{ikl}(y) = (-)\, {}_mM^{ikl}(y)$, we take

$${}_mM^{ikl}(y) = \sqrt{|g|}(y^i T^{kl} - y^k T^{il}), \tag{7}$$

where $T^{ik} = T^{ki}$ are the components of the symmetric energy-momentum tensor of matter (the source terms in the Einstein equations) and $\{y^i\}$ denote normal coordinates. Equation (7) gives the *total matter angular momentum density*, spinorial and orbital.[22]

It is interesting to note that the normal coordinates $\{y^i\}$ form the components of the *local radius vector* originating from **P** (a global radius vector does not exist in **GR**). So, the components of the ${}_mM^{ikl}(y)$ form a tensor density.

For the gravitational field we prefer the formula given in Ref. 22, which is closely connected with the canonical energy-momentum complex (see, e.g., Ref. 15 and Appendix A)

$${}_gM^{ikl}(y) = {}_{\mathrm{F}}U^{k[li]}(y) - {}_{\mathrm{F}}U^{i[lk]}(y) + y^i \sqrt{|g|}\, {}_{\mathrm{BT}}t^{kl} - y^k \sqrt{|g|}\, {}_{\mathrm{BT}}t^{il}, \tag{8}$$

where

$${}_{\mathrm{F}}U^{k[li]} := g^{km}\, {}_{\mathrm{F}}U_m{}^{[li]} \tag{9}$$

and

$$ {}_{\mathrm{BT}}t^{kl} := g^{ki}_{\mathrm{E}} t_i{}^{l} + \frac{g^{mk}_{,p}}{\sqrt{|g|}}\, {}_{\mathrm{F}}U_m{}^{[lp]}. \tag{10} $$

Here

$$ {}_{\mathrm{F}}U_i{}^{[kl]} = \alpha \frac{g_{ia}}{\sqrt{|g|}} [(-g)(g^{ka}g^{lb} - g^{la}g^{kb})]_{,b} \tag{11} $$

are the so-called *von Freud superpotentials* and

$$ {}_{\mathrm{E}}t_i{}^{k} = \alpha\{\delta^k_i g^{ms}(\Gamma^l_{mr}\Gamma^r_{sl} - \Gamma^r_{ms}\Gamma^l_{rl}) + g^{ms}_{,i}[\Gamma^k_{ms} - \tfrac{1}{2}(\Gamma^k_{tp}g^{tp} - \Gamma^l_{tl}g^{kt})g_{ms} - \tfrac{1}{2}(\delta^k_s\Gamma^l_{ml} + \delta^k_m\Gamma^s_{sl})]\} \tag{12} $$

are the components of the *Einstein canonical energy-momentum pseudotensor* of the gravitational field. Here $\alpha = c^4/16\pi G = 1/2\beta$.

One can interpret (8) as a sum of the *spinorial and orbital angular momentum density* of the gravitational field (see Appendix A).

Remark: The Landau–Lifschitz formula

$$ {}_gM^{ikl} = (-g)(y^i{}_{\mathrm{LL}}t^{kl} - y^k{}_{\mathrm{LL}}t^{il}), \tag{13} $$

where ${}_{\mathrm{LL}}t^{kl} = {}_{\mathrm{LL}}t^{lk}$ denote the components of the so-called ''Landau–Lifschitz gravitational energy-momentum pseudotensor''[3] *is not useful for our purposes*. It is because, when applied to (1), it gives trivial result

$$ {}_gS^{ikl}(P;v^l) = 0, \tag{14} $$

which is *not satisfactory* from the physical point of view. This is a consequence of the following: the formula (13) *does not include any spinorial angular momentum of the gravitational field*.

Substituting the following expansions up to the third order,[23]

$$ M^{ikl}(y) = y^i\hat{T}^{kl} - y^k\hat{T}^{il} + \tfrac{1}{2}(2\hat{T}^{kl}{}_{,p}y^iy^p - 2\hat{T}^{il}{}_{,p}y^ky^p) + R_3, \tag{15} $$

$$ e^{(a)}_i(y) = \hat{e}^{(a)}_i - \tfrac{1}{6}\hat{R}^p{}_{lim}\hat{e}^{(a)}_p y^l y^m + R_3, \tag{16} $$

and so on into (1), we get the following components ${}_mS^{abc}(P)$ of the *matter angular supermomentum tensor*

$$ {}_mS^{abc}(P) = 2(\delta^{ap}\hat{T}^{bc}{}_{,p} - \delta^{bp}\hat{T}^{ac}{}_{,p}) = 2(\delta^{ap}\nabla_p\hat{T}^{bc} - \delta^{bp}\nabla_p\hat{T}^{ac}), \tag{17} $$

or, in a covariant form,

$$ {}_mS^{abc}(P;v^l) = 2[(2\hat{v}^a\hat{v}^p - \hat{g}^{ap})\nabla_p\hat{T}^{bc} - (2\hat{v}^b\hat{v}^p - \hat{g}^{bp})\nabla_p\hat{T}^{ac}]. \tag{18} $$

Here v^i is the four velocity $(v^iv_i = 1)$ of an observer **O** which is at rest at the beginning **P** of the normal coordinates **NCS(P)** and $g^{ab} \doteq \eta^{ab}$ denotes the components of the local metric. A hat over a quantity means the value of the quantity at the point **P**, for example, $\hat{T}^{ab} = T^{ab}(P)$.

After substitution expansion (8) (up to the third order) and (16) into (1), we get, for gravitational field,

$$ {}_gS^{abc}(P;v^l) = \delta^{lt}{}_g\hat{M}^{abc}{}_{,lt}. \tag{19} $$

Equation (19) can be covariantly written in the following form:

$$ {}_gS^{abc}(P;v^l)=\alpha(2\hat{v}^p\hat{v}^t-\hat{g}^{pt})[\beta(\hat{g}^{ca}\hat{g}^{br}-\hat{g}^{cb}\hat{g}^{ar})\nabla_{(t}\hat{E}_{pr)}+2\hat{g}^{ar}\nabla_{(t}\hat{R}^{(c}{}_r{}^{b)}{}_{p)}-2\hat{g}^{br}\nabla_{(t}\hat{R}^{(c}{}_r{}^{a)}{}_{p)} + \tfrac{2}{3}\hat{g}^{bc}(\nabla_r\hat{R}^r{}_{(t}{}^a{}_{p)}-\beta\nabla_{(p}\hat{E}^a_{t)})-\tfrac{2}{3}\hat{g}^{ac}(\nabla_r\hat{R}^r{}_{(t}{}^b{}_{p)}-\beta\nabla_{(p}\hat{E}^b_{t)})]. \quad (20) $$

We will call tensors (18) and (20) *the canonical, angular supermomentum tensors* of matter and gravitation, respectively.

It is very interesting to note that *only spinorial part* $S^{ikl}={}_{\mathrm{F}}U^{i[kl]}-{}_{\mathrm{F}}U^{k[il]}$ of the canonical angular momentum pseudotensor ${}_gM^{ikl}(y)$ gives nonzero contribution to the tensor ${}_gS^{abc}(P;v^l)$. The *orbital part* $O^{ikl}=y^i\sqrt{|g|}_{\mathrm{BT}}t^{kl}-y^k\sqrt{|g|}_{\mathrm{BT}}t^{il}$ gives *no contribution* to the tensor ${}_gS^{abc}(P;v^l)$.

In vacuum, when $T_{ik}=0\Rightarrow E_{ik}:=T_{ik}-\frac{1}{2}g_{ik}T=0$, the canonical angular supermomentum tensor ${}_gS^{abc}(P;v^l)$ given by (20) simplifies to give

$$ {}_gS^{abc}(P;v^l)=2\alpha(2\hat{v}^p\hat{v}^t-\hat{g}^{pt})[\hat{g}^{ar}\nabla_{(t}\hat{R}^{(b}{}_r{}^{c)}{}_{p)}-\hat{g}^{br}\nabla_{(t}\hat{R}^{(a}{}_r{}^{c)}{}_{p)}]. \quad (21) $$

Remarks:

(1) It follows from the formulas (18), (20), and (21) that the canonical angular supermomentum tensors ${}_mS^{abc}(P;v^l)$ and ${}_gS^{abc}(P;v^l)$ *do not require any radius vector* to be constructed.

(2) It also follows from the formulas (7) and (8)–(12) that

$$ {}_mM^{abc}(P)={}_m\hat{M}^{abc}=0, \quad {}_gM^{abc}(P)={}_g\hat{M}^{abc}=0. $$

Despite that the above quantities vanish, we have manifestly introduced them into the definition (1) of the canonical angular supermomentum tensors to emphasize their relative character.

III. PHYSICAL INTERPRETATION OF THE CANONICAL ANGULAR SUPERMOMENTUM TENSORS

Following the line developed in Ref. 15, we give, in our opinion, the most natural physical interpretation of the canonical angular supermomentum tensors ${}_gS^{abc}(P;v^l)=(-){}_gS^{bac}(P;v^l)$ and ${}_mS^{abc}(P;v^l)=(-){}_mS^{bac}(P;v^l)$.

A. Matter canonical angular supermomentum tensor ${}_mS^{abc}(P;v^l)$

In sufficiently small surroundings of the point **P**, we consider the differences

$$ \begin{aligned} {}_mM^{(a)(b)(c)}(y)-{}_mM^{(a)(b)(c)}(P) &= {}_mM^{ikl}(y)e_i^{(a)}(y)e_k^{(b)}(y)e_l^{(c)}(y)-{}_m\hat{M}^{ikl}\hat{e}_i^{(a)}\hat{e}_k^{(b)}\hat{e}_l^{(c)} \\ &= [y^i\hat{T}^{kl}-y^k\hat{T}^{il}+(y^i\hat{T}^{kl}{}_{,p}-y^k\hat{T}^{il}{}_{,p})y^p+R_3] \\ &\quad \times(\hat{e}_i^{(a)}-\tfrac{1}{6}\hat{R}^t{}_{ris}\hat{e}_t^{(a)}y^ry^s+R_3) \\ &\quad \times(\hat{e}_k^{(b)}-\tfrac{1}{6}\hat{R}^u{}_{ekd}\hat{e}_u^{(b)}y^ey^d+R_3)(\hat{e}_l^{(c)}-\tfrac{1}{6}\hat{R}^v{}_{flg}\hat{e}_v^{(c)}y^fy^g+R_3) \\ &\quad -{}_mM^{abc}(P) \\ &= y^a\hat{T}^{bc}-y^b\hat{T}^{ac}+(y^a\hat{T}^{bc}{}_{,p}-y^b\hat{T}^{ac}{}_{,p})y^p+R_3. \end{aligned} \quad (22) $$

In every previous expansion R_3 denotes the remainder of the third order.

Integrating the differences (22) over a sufficiently small domain Ω with the same properties as the domain Ω which occurs in the formulas (1), we obtain

$$ \begin{aligned} \int_\Omega[M^{(a)(b)(c)}(y)-M^{(a)(b)(c)}(P)]\,d\Omega &\approx {}_mS^{abc}(P;v^l)\cdot\tfrac{1}{2}\int_\Omega(y^d)^2\,d\Omega \\ &= {}_mS^{abc}(P;v^l)\cdot\tfrac{1}{2}M=:{}_mS^{abc}(P;v^l;\Omega). \end{aligned} \quad (23) $$

Here ${}_mS^{abc}(P;v^l;\Omega)=(-)_mS^{abc}(P;v^l;\Omega)$ is *a tensor* whose components give *the approximate values of the relative angular momentum* inside the domain Ω.

We can easily notice that this tensor ${}_mS^{abc}(P;v^l;\Omega)$ is the product of:

(1) the tensor ${}_mS^{abc}(P;v^l)$, which depends only on the matter and gravitational fields and four-velocity of the observer **O**, and,
(2) the term $\frac{1}{2}\int_\Omega (y^d)^2\, d\Omega=:\frac{1}{2}M$, which depends on the common value M of the moments of inertia of the domain Ω only.

The supermomentum tensor of matter ${}_mS^{abc}(P;v^l)$ is exactly the first term in this product.

B. Gravitational canonical angular supermomentum tensor ${}_g\boldsymbol{S}^{abc}(\boldsymbol{P};\boldsymbol{v}^l)$

For a gravitational field, the canonical angular momentum pseudotensor ${}_gM^{nmk}$ has the form given by (8)–(12), i.e.,

$$ {}_gM^{nmk}=\sqrt{|g|}(x^n g^{mi}{}_{\mathrm{E}}t_i{}^k - x^m g^{ni}{}_{\mathrm{E}}t_i{}^k)+x^n{}_{\mathrm{F}}U_i{}^{[kl]}g^{im}{}_{,l}-x^m{}_{\mathrm{F}}U_i{}^{[kl]}g^{in}{}_{,l}+g^{im}{}_{\mathrm{F}}U_i{}^{[kn]}-g^{in}{}_{\mathrm{F}}U_i{}^{[km]}. \tag{24}$$

As for the case of matter fields, let us consider the differences

$$\begin{aligned} {}_gM^{(a)(b)(c)}(y)-{}_gM^{(a)(b)(c)}(P)&={}_gM^{nmk}(y)e_n^{(a)}(y)e_m^{(b)}(y)e_k^{(c)}(y)-{}_g\hat{M}^{nmk}\hat{e}_n^{(a)}\hat{e}_m^{(b)}\hat{e}_k^{(c)}\\ &=[({}_{\mathrm{F}}\hat{U}^{m[kn]}{}_{,r}-{}_{\mathrm{F}}\hat{U}^{n[km]}{}_{,r})y^r+\tfrac{1}{2}({}_{\mathrm{F}}\hat{U}^{m[kn]}{}_{,rs}-{}_{\mathrm{F}}\hat{U}^{n[km]}{}_{,rs})y^r y^s+R_3]\\ &\quad\times(\hat{e}_n^{(a)}-\tfrac{1}{6}\hat{R}^e{}_{ind}\hat{e}_e^{(a)}y^i y^d+R_3)(\hat{e}_m^{(b)}-\tfrac{1}{6}\hat{R}^n{}_{fmg}\hat{e}_n^{(b)}y^f y^g+R_3)\\ &\quad\times(\hat{e}_k^{(c)}-\tfrac{1}{6}\hat{R}^j{}_{pkt}\hat{e}_j^{(c)}y^p y^t+R_3)-{}_gM^{abc}(P)\\ &=({}_{\mathrm{F}}\hat{U}^{b[ca]}{}_{,r}-{}_{\mathrm{F}}\hat{U}^{a[cb]}{}_{,r})y^r+\frac{1}{2!}({}_{\mathrm{F}}\hat{U}^{b[ca]}{}_{,pt}-{}_{\mathrm{F}}\hat{U}^{a[cb]}{}_{,pt})y^p y^t+R_3. \end{aligned}\tag{25}$$

Integrating these differences over the domain Ω with the same properties as the domain Ω used in Sec. III A, we have

$$\begin{aligned} \int_\Omega [{}_gM^{(a)(b)(c)}(y)-{}_gM^{(a)(b)(c)}(P)]\, d\Omega &\approx {}_gS^{abc}(P;v^l)\tfrac{1}{2}\int_\Omega (y^d)^2\, d\Omega\\ &={}_gS^{abc}(P;v^l)\tfrac{1}{2}M=:{}_gS^{abc}(P;v^l;\Omega). \end{aligned}\tag{26}$$

Here ${}_gS^{abc}(P;v^l;\Omega)=(-)_gS^{abc}(P;v^l;\Omega)$ is a tensor whose components give the *approximate values* of the (relative) *gravitational canonical angular momentum* inside the domain Ω.

The tensor ${}_gS^{abc}(P;v^l;\Omega)$ also factorizes onto a product of the two terms: the gravitational, canonical supermomentum tensor ${}_gS^{abc}(P;v^l)$ and the term $\frac{1}{2}M$ which depends on the (common) value of the moments of inertia of the domain Ω with respect to the subspaces $y^i=0$ $(i=0,1,2,3)$.

IV. APPLICATION OF THE CANONICAL, ANGULAR SUPERMOMENTUM TENSORS IN GR

A. Plane and plane-fronted gravitational waves (pp-waves)

By linearly polarized, plane gravitational wave we understand the solution to the Einstein vacuum equations which, in coordinates (t,x,y,z), has the line element of the form[24]

$$ds^2=dt^2-L^2(e^{2\beta}dx^2+e^{-2\beta}dy^2)-dz^2=du\,dv-L^2(e^{2\beta}dx^2+e^{-2\beta}dy^2), \tag{27}$$

where

$$L = L(u), \quad \beta = \beta(u), \quad u = t - z, \quad v = t + z.$$

The Einstein equations for (27) reduce just to one equation,

$$L'' + (\beta')^2 L = 0, \tag{28}$$

for the two *metric functions* $L = L(u)$, $\beta = \beta(u)$. Here $\beta' := d\beta/du$, $L' := dL/du$, and so on. A plane wave described by (27) and (28) propagates in the positive direction of the z axis.

By a suitable coordinate transformation (see, e.g., Ref. 14) one can bring the line element given by (27) and (28) to the following form:

$$ds^2 = 2(Y^2 - X^2)\frac{F(U)}{2}dU^2 + 2dUdV - dX^2 - dY^2 \tag{29}$$

with an arbitrary function $F = F(U) = F(t - z)$.

The *plane-fronted gravitational wave with parallel rays* (pp-wave; see, e.g., Ref. 25) is a generalization of the plane wave and has the following line element in coordinates ($U = t - z$, V, X, Y)

$$ds^2 = 2H(X,Y,U)dU^2 + 2dUdV - dX^2 - dY^2, \tag{30}$$

where

$$\Delta H := \left(\frac{\partial^2}{\partial X^2} + \frac{\partial^2}{\partial Y^2}\right)H = 0. \tag{31}$$

The vector field tangent to the V-lines is covariantly constant and null.

Let us calculate the components of the canonical, gravitational angular supermomentum tensor ${}_gS^{abc}(P;v^l)$ given by (21) for the line element (30) and (31). To simplify calculations we use the *null coreper*

$$\vartheta^0 = HdU + dV, \quad \vartheta^1 = dU, \quad \vartheta^2 = dX, \quad \vartheta^3 = dY. \tag{32}$$

After some tedious calculations we obtain that the components of the tensor

$${}_gS^{abc}(P;v^l) = (-)\,{}_gS^{bac}(P;v^l)$$

do not vanish in the null coreper and they do not vanish in *any other coreper, too* (for results of the performed calculations, see Appendix B). It means that the plane-fronted and plane gravitational waves *possess and carry* angular supermomentum. For the sake of comparison, we calculate the components of the canonical angular momentum pseudotensor (24) for pp and plane waves, which, in the null coreper (32), give

$${}_gM^{nmk} = 0, \tag{33}$$

i.e., they give *no angular momentum*. However, the components of the pseudotensor ${}_gM^{nmk}$ in the coordinates (t,x,y,z), for example, *do not vanish*. This is not surprising because the components ${}_gM^{nmk}$ form *neither a tensor nor any other geometric object*. On the contrary, the components of ${}_gS^{abc}(P;v^l)$, as being the components of a tensor, *do not vanish in any coordinates* and, therefore, they give *a coordinate independent, covariant characteristic* of the plane and plane-fronted gravitational waves. This characteristic depends on the gravitational field and on the four-velocity v^i of an observer **O** only.

B. Friedman universes

Friedman universes ($\equiv$ Friedman cosmological models) are the solutions to the Einstein equations with the Friedman–Lemaître–Robertson–Walker line element (**FLRW** line element) which, in the comoving coordinates $x^0=ct$, $x^1=r$, $x^2=\vartheta$, $x^3=\varphi$, reads

$$ds^2=c^2dt^2-R^2(t)\left[\frac{dr^2}{(1-kr^2)}+r^2(d\vartheta^2+\sin^2\vartheta d\varphi^2)\right], \tag{34}$$

where the *curvature index* $k=0,\pm1$ and $R=R(t)$ is the *scale factor*. The parameter t represents universal time called the *cosmic time*.

The scale factor $R(t)$ satisfies Einstein field equations (sometimes called *Friedman equations*)

$$\frac{3\dot{R}^2}{c^2R^2}+\frac{3k}{R^2}=\frac{\epsilon}{2\alpha}, \tag{35}$$

$$\frac{\dot{R}^2}{c^2R^2}+\frac{2\ddot{R}}{c^2R^2}+\frac{k}{R^2}=(-)\frac{p}{2\alpha}. \tag{36}$$

In (35) and (36) ϵ is the energy density of the cosmological fluid and p is the pressure; $\dot{R}:=dR/dt$, $\ddot{R}:=d^2R/dt^2$.

The canonical angular supermomentum tensors ${}_mS^{abc}(P;v^l)$ and ${}_gS^{abc}(P;v^l)$, similarly as with canonical superenergy tensors ${}_gS_i{}^k(P;v^l)$, ${}_mS_i{}^k(P;v^l)$, are *defined locally* and they depend on the four-velocity v^i ($v^iv_i=1$) of an observer **O** being at rest at the point **P**.

If the vector field v^i ($v^iv_i=1$) is given over a certain domain Ω of space–time or over the whole space–time, then one can uniquely determine tensor fields ${}_gS^{abc}(x;v^l)$ and ${}_mS^{abc}(x;v^l)$ [this is by performing the ''averaging'' given by (1) at every point of the domain Ω or at every point of the entire space–time and by using the field v^i to present the obtained results in a covariant way].

In the framework of Friedman cosmology *there always exists* such a vector field: it is just the four-velocity v^i of the *isotropic or fundamental observers* being at rest in comoving coordinates. This vector field is *geometrically and physically distinguished* in the case of Friedman cosmology: it forms timelike, geodesic, and hypersurface-orthogonal congruence.

So, in Friedman universes, one can introduce *unique* tensor fields ${}_gS^{abc}(x;v^l)$, ${}_mS^{abc}(x;v^l)$ and, in analogy to global superenergetic quantities considered in Refs. 4, 13, and 15, one can define *global angular supermomentum*

$$S^{ab}:=\frac{1}{c}\int_{t=\mathrm{const}}({}_gS^{ab0}+{}_mS^{ab0})\sqrt{|g|}\,dr\,d\vartheta\,d\varphi \tag{37}$$

of matter and gravitation.

As we restrict ourselves to the analysis of the global angular supermomentum components given by (37) in this paper, we just need to give the components

$${}_mS^{ab0}(x;v^l)=(-){}_mS^{ba0}(x;v^l) \quad \text{and} \quad {}_gS^{ab0}(x;v^l)=(-){}_gS^{ba0}(x;v^l) \tag{38}$$

of the tensors ${}_mS^{abc}(x;v^l)$ and ${}_gS^{abc}(x;v^l)$ which determine these global quantities.

After very simple but something lengthly calculations, we obtain, for Friedman universes

$${}_mS^{ab0}(x;v^l)=0={}_gS^{ab0}(x;v^l). \tag{39}$$

It follows from (39) that the components

$$S^{ab0}(x;v^l):={}_mS^{ab0}(x;v^l)+{}_gS^{ab0}(x;v^l) \tag{40}$$

and the components $S^{ab}=(-)S^{ba}$ of the global, canonical angular supermomentum of the Friedman universes *trivially vanish*. It can also be seen from (39) that the *global, canonical angular supermomenta* of gravitation

$$ {}_gS^{ab}:=\frac{1}{c}\int_{t=\mathrm{const}} {}_gS^{ab0}(x;v^l)\sqrt{|g|}\,dr\,d\vartheta\,d\varphi \tag{41} $$

and matter

$$ {}_mS^{ab}:=\frac{1}{c}\int_{t=\mathrm{const}} {}_mS^{ab0}(x;v^l)\sqrt{|g|}\,dr\,d\vartheta\,d\varphi \tag{42} $$

also *trivially vanish* for these universes since their integrands vanish. These results seem to be quite reasonable and somehow expected.

For the sake of comparison, we give here also the components ${}_gM^{ik0}$ and ${}_mM^{ik0}=\sqrt{|g|}(x^iT^{k0}-x^kT^{i0})$ of the canonical quantities which give the angular momentum of gravitation and matter, respectively.

The set $\{x^i\}$ $(i=0,1,2,3)$ here is the global, comoving coordinates (ct,x,y,z) in which the **FLRW** line element is given by

$$ ds^2=c^2dt^2-R^2(t)\,\frac{(dx^2+dy^2+dz^2)}{[1+k(x^2+y^2+z^2)/4]^2}. \tag{43} $$

Global coordinates (ct,x,y,z) are closely related to the Cartesian coordinates in a flat space–time. They have to be applied here if one wants to obtain any reasonable global results by using canonical angular momentum pseudotensor ${}_gM^{ikl}=(-)_gM^{kil}$ given by (8)–(12) or by (24) and by using matter angular momentum geometric object[26] ${}_mM^{ikl}=\sqrt{|g|}(x^iT^{kl}-x^kT^{il})$.

After some calculations [especially lengthy for components ${}_gM^{ik0}=(-)_gM^{ki0}$] we obtain

$$ \begin{gathered} {}_mM^{120}=(-)_mM^{210}=0,\quad {}_mM^{130}=(-)_mM^{310}=0,\quad {}_mM^{230}=(-)_mM^{320}=0,\\ {}_mM^{010}=(-)_mM^{100}=(-)\,\frac{xR^3\epsilon}{u^3}=(-)\,\frac{6\alpha Rx}{u^3}(\dot{R}^2+k),\\ {}_mM^{020}=(-)_mM^{200}=(-)\,\frac{yR^3\epsilon}{u^3}=(-)\,\frac{6\alpha Ry}{u^3}(\dot{R}^2+k),\\ {}_mM^{030}=(-)_mM^{300}=(-)\,\frac{zR^3\epsilon}{u^3}=(-)\,\frac{6\alpha Rz}{u^3}(\dot{R}^2+k), \end{gathered} \tag{44} $$

where

$$ u:=1+\frac{k(x^2+y^2+z^2)}{4}=:1+\frac{kr^2}{4}. \tag{45} $$

$$ \begin{gathered} {}_gM^{120}=(-)_gM^{210}=0,\quad {}_gM^{130}=(-)_gM^{310}=0,\quad {}_gM^{230}=(-)_gM^{320}=0,\\ {}_gM^{010}=(-)_gM^{100}=(-)\,\frac{2\alpha x}{u^2}\left[\frac{R}{u}\left(6\dot{R}^2-\frac{k^2r^2}{2}\right)+\dot{R}kt+kR\right],\\ {}_gM^{020}=(-)_gM^{200}=(-)\,\frac{2\alpha y}{u^2}\left[\frac{R}{u}\left(6\dot{R}^2-\frac{k^2r^2}{2}\right)+\dot{R}kt+kR\right], \end{gathered} \tag{46} $$

$$ {}_gM^{030}=(-)\,{}_gM^{300}=(-)\,\frac{2\alpha z}{u^2}\left[\frac{R}{u}\left(6\dot{R}^2-\frac{k^2r^2}{2}\right)+\dot{R}kt+kR\right]. $$

From (44)–(46) it follows that the components $M^{ik0}=(-)M^{ki0}$ of the canonical angular momentum complex of matter and gravitation $M^{ikl}:={}_MM^{ikl}+{}_gM^{ikl}$, which satisfies *local conservation laws*

$$ M^{ikl}{}_{,l}=0, $$

in coordinates (ct,x,y,z) are

$$ \begin{aligned} &M^{120}=M^{130}=M^{230}=0,\\ &M^{010}=(-)\,\frac{2\alpha x}{u^2}\left[\frac{R}{u}\left(9\dot{R}^2-\frac{k^2r^2}{2}\right)+\dot{R}kt+\frac{3kR}{u}\right],\\ &M^{020}=(-)\,\frac{2\alpha y}{u^2}\left[\frac{R}{u}\left(9\dot{R}^2-\frac{k^2r^2}{2}\right)+\dot{R}kt+kR=\frac{3kR}{u}\right],\\ &M^{030}=(-)\,\frac{2\alpha z}{u^2}\left[\frac{R}{u}\left(9\dot{R}^2-\frac{k^2r^2}{2}\right)+\dot{R}kt+kR+\frac{3kR}{u}\right]. \end{aligned} \tag{47} $$

This leads to the conclusion that the global, canonical angular momentum of the Friedman universes

$$ M^{ik}=\frac{1}{c}\int_{t=\mathrm{const}}({}_mM^{ik0}+{}_gM^{ik0})\,dxdydz \tag{48} $$

vanishes. The same results hold for the global *matter canonical angular momentum*

$$ {}_mM^{ik}=\frac{1}{c}\int_{t=\mathrm{const}}{}_mM^{ik0}\,dxdydz \tag{49} $$

and for the global *gravitational canonical angular momentum*

$$ {}_gM^{ik}=\frac{1}{c}\int_{t=\mathrm{const}}{}_gM^{ik0}\,dxdydz \tag{50} $$

separately: they all vanish in coordinates (ct,x,y,z), too. So, we have the same global results [in coordinates (ct,x,y,z)] for canonical angular momentum and canonical angular supermomentum in the case of the Friedman universes.

However, the canonical angular supermomenta possess much better geometric properties than the canonical angular momenta. Namely, we have the following.

(1) The components of ${}_gS^{ikl}(x;v^t)$ and ${}_mS^{ikl}(x;v^t)$ and the components of $S^{ikl}(x;v^t)={}_gS^{ikl}(x;v^t)+{}_mS^{ikl}(x;v^t)$ *form tensors*. Because of this the canonical angular supermomenta ${}_gS^{ikl}(P;v^t)$, ${}_mS^{ikl}(P;v^t)$, and $S^{ikl}(P;v^t)={}_gS^{ikl}(P;v^t)+{}_mS^{ikl}(P;v^t)$ *are localized* and

$$ {}_gS^{ik0}(x;v^t)=0,\quad {}_mS^{ik0}(x;v^t)=0,\quad S^{ik0}(x;v^t)={}_gS^{ik0}(x;v^t)+{}_mS^{ik0}(x;v^t)=0, \tag{51} $$

independent of the coordinates (x^1,x^2,x^3) which have been used for the parametrization of spaces $t=\mathrm{const}$. So, the global angular supermomenta vanish, i.e.,

$$ S^{ik}=0,\quad {}_gS^{ik}=0,\quad {}_mS^{ik}=0, \tag{52} $$

independent of the coordinates used within the spaces $t=\text{const}$.

(2) The components of the canonical angular supermomentum tensors *do not depend* on the so-called ''radius vector'' which *does not exist globally* in the framework of **GR**. On the contrary, the canonical angular momenta ${}_gM^{ikl}$ and $M^{ikl}={}_gM^{ikl}+{}_mM^{ikl}$ explicitly depend on ''radius vector'' and *form neither a tensor nor any other geometric objects*. So, they must be used only in *special coordinates* in order to give physically meaningful results. In the case of the Friedman universes these are the coordinates (ct,x,y,z) in which the **FLRW** line element is in the form given by (43). Moreover, even the matter canonical angular momentum object

$${}_mM^{ikl}=\sqrt{|g|}(x^iT^{kl}-x^kT^{il})$$

cannot be *any tensor density defined globally* because it needs a global *radius vector* $\vec{x}=x^i\partial_i$ to be a tensor density. However, radius vector *does not exist globally* in the framework of **GR**.

C. Schwarzschild space–time

The static, spherically symmetric space–time in the coordinates $(x^0=ct,\ x^1=r,\ x^2=\vartheta,\ x^3=\varphi)$ has the line element given by[3]

$$ds^2=e^{\nu(r)}c^2dt^2-e^{\lambda(r)}dr^2-r^2(d\vartheta^2+\sin^2\vartheta d\varphi^2). \tag{53}$$

In the exterior region $r\geqslant r_*$, the Einstein equations demand

$$e^{\nu}=e^{-\lambda}=1-\frac{2Gm}{c^2r}=:1-\frac{r_g}{r}, \tag{54}$$

where m is the mass of a spherical star (source of the gravitational field) having ''radius'' $r=r_*>r_g$. Here $r_g:=2Gm/c^2$ is the *Schwarzschild radius.*

As a source of the gravitational field we have taken a normal static and spherical star which consists of perfect fluid with the energy-momentum tensor $T^{ik}=T^{ki}$ in the form

$$T^{ik}=(\epsilon+p)u^iu^k-g^{ik}p. \tag{55}$$

It is known that one can join smoothly the interior and the exterior Schwarzschild solutions on the star surface and obtain *complete Schwarzschild space–time*. In the complete Schwarzschild space–time there exists a distinguished timelike unit vector field $v^i:v^iv_i=1$, orthogonal to the spatial slices $x^0=ct=\text{const}$. So, similarly as in the case of Friedman cosmological models, one can obtain unique tensor fields ${}_gS^{abc}(x)$ and ${}_mS^{abc}(x)$.

After some calculations, from (53), (54), and (21), outside of the star, i.e., in the domain $r>r_*$, where $T^{ik}=0$, we get

$${}_mS^{abc}(x)=0,\quad {}_gS^{010}(x)=(-){}_gS^{100}(x)=(-)\alpha\frac{r_g^2}{6r^4(r-r_g)}. \tag{56}$$

The other components of the ${}_gS^{ikl}(x)$ vanish in the domain.

On the other hand, inside the star, where $T^{ik}\neq 0$, for the line element (1) and for the energy-momentum tensor (55) we get from (18) and (20)

$$\begin{gathered}{}_mS^{010}=(-){}_mS^{100}\neq 0,\quad {}_mS^{122}=(-){}_mS^{212}={}_mS^{133}\sin^2\vartheta=(-){}_mS^{313}\sin^2\vartheta\neq 0,\\ {}_gS^{010}=(-){}_gS^{100}\neq 0,\quad {}_gS^{122}=(-){}_gS^{212}\neq 0,\quad {}_gS^{133}=(-){}_gS^{313}\neq 0,\end{gathered} \tag{57}$$

which is independent of any specific form of the interior solutions. The remaining components of the tensors ${}_mS^{abc}(x)$ and ${}_gS^{abc}(x)$ *vanish identically* in the case. It can be seen from (56) and (57) that physically most important, spatial components

$$S^{12}=(-)S^{21},\quad S^{13}=(-)S^{31},\quad S^{23}=(-)S^{32}$$

of the global angular supermomentum $S^{ik}=(-)S^{ki}$ (matter and gravitation) in the complete Schwarzschild space–time

$$S^{ik}:=\int_{x^0=ct=\mathrm{const}}({}_mS^{ik0}+{}_gS^{ik0})\sqrt{|g|}\,dr\,d\vartheta\,d\varphi \tag{58}$$

are trivially equal to zero. By *trivially* we mean by means of the fact that their integrands

$$S^{120}:={}_mS^{120}+{}_gS^{120},\quad S^{130}:={}_mS^{130}+{}_gS^{130},\quad S^{230}:={}_mS^{230}+{}_gS^{230}$$

identically vanish. We find this result quite reasonable. The same is correct for the *exterior* and *interior* Schwarzschild solutions separately. The analogical results can be obtained for the global *angular supermomentum of matter*

$${}_mS^{ik}:=\int_{x^0=\mathrm{const}}{}_mS^{ik0}\sqrt{|g|}\,dr\,d\vartheta\,d\varphi, \tag{59}$$

and *gravitation*

$${}_gS^{ik}:=\int_{x^0=ct=\mathrm{const}}{}_gS^{ik0}\sqrt{|g|}\,dr\,d\vartheta\,d\varphi. \tag{60}$$

The final results

$$S^{12}=S^{13}=S^{23}=0,\quad {}_gS^{12}={}_gS^{13}={}_gS^{23}=0,\quad {}_mS^{12}={}_mS^{13}={}_mS^{23}=0$$

are independent of the spatial coordinates used for parametrization of the slices $x^0=ct=\mathrm{const}$.

One can compare the above results with the results of calculations of the global angular momentum of the Schwarzschild space–time by using canonical, angular-momentum complex given by Bergmann and Thomson (see Appendix A)

$$M^{ijk}=(-)M^{jik}=x^i_{\mathrm{BT}}K^{jk}-x^j_{\mathrm{BT}}K^{ik}+{}_{\mathrm{F}}U^{i[jk]}-{}_{\mathrm{F}}U^{j[ik]}=(x^i_{\mathrm{F}}U^{j[kl]}-x^j_{\mathrm{F}}U^{i[kl]})_{,l}=:M^{[ij][kl]}{}_{,l}\,. \tag{61}$$

However, while using the complex (61), one should take the isotropic Cartesian coordinates ($x^0=ct, x^1=x,\ x^2=y, x^3=z$) (because the canonical angular momentum complex forms neither a tensor nor any other geometric object). In these coordinates the exterior Schwarzschild line element is of the form[3]

$$ds^2=\left[\frac{1-r_g/4\rho}{1+r_g/4\rho}\right]^2c^2dt^2-\left(1+\frac{r_g}{4\rho}\right)^4(dx^2+dy^2+dz^2), \tag{62}$$

where

$$\begin{aligned} r&=\rho\left(1+\frac{r_g}{4\rho}\right)^2,\\ \rho^2&=x^2+y^2+z^2, \end{aligned} \tag{63}$$

or, asymptotically, when $\rho\to\infty$ ($\Leftrightarrow r\to\infty$)

$$ds^2=\left(1-\frac{r_g}{\rho}\right)c^2dt^2-\left(1+\frac{r_g}{\rho}\right)(dx^2+dy^2+dz^2). \tag{64}$$

The asymptotic form (64) of the exterior Schwarzschild line element is sufficient to calculate the components $M^{ik}=(-)M^{ki}$ of the global angular momentum of the complete Schwarzschild space–time because they are determined by surface integrals over boundary of a spatial slice $x^0=ct=\text{const}$ (see Appendix A and Refs. 3 and 22):

$$M^{ik}=\oint_{\text{over a sphere } S^2 \text{ with } R\to\infty} M^{[ik][0\alpha]} n_\alpha r^2\, d\Omega. \tag{65}$$

From (64) and (65) we get, after some calculations,

$$M^{12}=M^{13}=M^{23}=0. \tag{66}$$

So, at least in asymptotically Lorentzian coordinates, we have equivalent global results concerning angular momentum and angular supermomentum. However, one should emphasize that the expressions (18), (20), and (21) of Sec. II *are tensorial* and are defined without *any use of radius vector*. This means they can be applied for any other spatial coordinates used for parametrization of the space $x^0=ct=\text{const}$ and give the same results, i.e.,

$$S^{12}=S^{13}=S^{23}=0, \quad {}_gS^{12}={}_gS^{13}={}_gS^{23}=0, \quad {}_mS^{12}+{}_mS^{13}={}_mS^{23}=0.$$

D. Stationary Kerr space–time

We use the Kerr space–time in *Boyer–Lindquist coordinates* $x^0=t$, $x^1=r$, $x^2=\theta$, $x^3=\varphi$ (see, e.g., Refs. 3 and 27) in which the line element reads as

$$ds^2=\left(1-\frac{r_g r}{\rho^2}\right)dt^2-\frac{\rho^2}{\Delta}dr^2-\rho^2 d\theta^2-\left(r^2+a^2+\frac{r_g r a^2}{\rho^2}\sin^2\theta\right)\sin^2\theta d\varphi^2+\frac{2r_g r a}{\rho^2}\sin^2\theta d\varphi dt, \tag{67}$$

where

$$\Delta:=r^2-r_g r+a^2, \quad \rho^2:=r^2+a^2\cos^2\theta, \quad r_g:=2mG. \tag{68}$$

Here m is the mass of the source and $M=ma$, its angular momentum.

The nonvanishing components[28] ${}_gS^{abc}(P;v^l)$ of the canonical gravitational angular supermomentum tensor (21) for the line element (67) and (68) are

$$\begin{aligned}
&{}_gS^{130}=(-)_gS^{310}, \quad {}_gS^{133}=(-)_gS^{313}, \quad {}_gS^{121}=(-)_gS^{211}\ {}_gS^{122}=(-)_gS^{212},\\
&{}_gS^{230}=(-)S^{320}, \quad {}_gS^{233}=(-)_gS^{323}, \quad {}_gS^{010}=(-)_gS^{100}, \quad {}_gS^{013}=(-)_gS^{103},\\
&{}_gS^{020}=(-)_gS^{200}, \quad {}_gS^{023}=(-)_gS^{203}, \quad {}_gS^{031}=(-)_gS^{301}, \quad {}_gS^{032}=(-)_gS^{302}.
\end{aligned} \tag{69}$$

One can get the components of ${}_gS^{abc}(P;v^l)$ in other coordinates by using the tensor transformation rule.

The analysis of the asymptotic behavior of the nonzero components of the tensor field ${}_gS^{abc}(x;v^l)$ determined by (69) shows that they are at least of the $0(r^{-4})$ order. So, the integrals[29]

$$S^{ik}=\frac{1}{c}\int_{x^0=t=\text{const}} {}_gS^{ik0}\sqrt{|g|}\,dr d\theta d\varphi, \quad \sqrt{|g|}=\rho^2\sin\theta, \tag{70}$$

which give the components of the global angular supermomentum for Kerr space–time *are convergent*. These components are also independent of time.

Of all the integrals (70), which give global angular supermomentum the *spatial components*

$$S^{12}=(-)S^{21}(=0), \quad S^{13}=(-)S^{31}\neq 0, \quad S^{23}=(-)S^{32}\neq 0,$$

are most physically important.

It is interesting to compare (69) with the components of the canonical angular momentum object (8) for Kerr space–time in Boyer–Lindquist coordinates. After some calculations[30] we get that the following components of the canonical angular momentum object ${}_gM^{ikl}=(-)_gM^{kil}$ given by (8) do not vanish:[31]

$$\begin{gathered}
{}_gM^{010}=(-)_gM^{100},\quad {}_gM^{011}=(-)_gM^{101},\quad {}_gM^{012}=(-)_gM^{102},\quad {}_gM^{013}=(-)_gM^{103},\\
{}_gM^{020}=(-)_gM^{200},\quad {}_gM^{021}=(-)_gM^{201},\quad {}_gM^{022}=(-)_gM^{202},\quad {}_gM^{023}=(-)_gM^{203},\\
{}_gM^{030}=(-)_gM^{300},\quad {}_gM^{033}=(-)_gM^{303},\\
{}_gM^{121}=(-)_gM^{211},\quad {}_gM^{122}=(-)_gM^{212},\\
{}_gM^{230}=(-)_gM^{320},\quad {}_gM^{231}=(-)_gM^{321},\quad {}_gM^{232}=(-)_gM^{322},\quad {}_gM^{233}=(-)_gM^{323},\\
{}_gM^{130}=(-)_gM^{310},\quad {}_gM^{131}=(-)_gM^{311},\quad {}_gM^{132}=(-)_gM^{312},\\
{}_gM^{133}=(-)_gM^{313}.
\end{gathered}\tag{71}$$

Here more compoments are different from zero than in (69). Moreover, these components are *nontensorial*. Therefore, the components of ${}_gM^{ikl}$ in any other system of coordinates must be calculated separately.

Having analyzed the asymptotic behavior of the components (71) at spatial infinity we conclude that the integrals[32]

$$M^{ik}=\frac{1}{c}\int_{x^0=t=\mathrm{const}} {}_gM^{ik0}\,dr\,d\theta d\varphi,\tag{72}$$

which represent the components of the global angular momentum are convergent *iff* $i,k=1,2,3$; $(M^{12}=0)$, i.e., only the integrals which represent *spatial components* of the global angular momentum are convergent. The convergent integrals are also independent of time.

So, for the most important spatial components M^{ik} we have similiar result as for the spatial components S^{ik}—the integrals which represent these components are convergent and independent of time. However, the values and dimension of the corresponding convergent integrals, except $S^{12}=M^{12}=0$, are, of course, different.

In order to conclude, we notice that the components of the ${}_gS^{abc}(P;v^l)$ have *better geometrical properties* (form a tensor) than the components of ${}_gM^{ikl}$ and *give better convergence* of the corresponding integrals which represent global quantities for Kerr space–time (all S^{ik} are convergent, but not all M^{ik}).

The superiority of the superenergetic quantities over canonical energetic quantities for Kerr geometry is more explicit if we compare the canonical superenergy and linear supermomentum with canonical energy and momentum, all calculated in the Boyer–Lindquist coordinates. Namely, we can see that the superenergy density ${}_g\epsilon_s := {}_gS_{ik}v^iv^k$ is a *positive-definite scalar* and the integrals

$$S_i=\frac{1}{c}\int_{x^0=\mathrm{const}} {}_gS_i{}^0(P;v^l)\sqrt{|g|}\,dr\,d\theta d\varphi,\tag{73}$$

which represent global superenegetic quantities are convergent $(S_0>0, S_1=S_2=0, S_3\neq 0)$. In consequence, they have very good physical sense.

On the other hand, the integrals

$$P_i=\frac{1}{c}\int_{x^0=t=\mathrm{const}} {}_EK_i{}^0\,dr\,d\theta d\varphi,\tag{74}$$

which give the global energetic quantities for Kerr space–time, *are divergent* in Boyer–Linquist coordinates (except $P_1 = P_2 = 0$). So, the integrals (74) *have no physical sense* in the Boyer–Lindquist coordinates. It is not surprising because canonical energy-momentum complex ${}_{\mathrm{E}}K_i^{\ k} = \sqrt{|g|}(T_i^k + {}_{\mathrm{E}}t_i^{\ k})$ can only be used in asymptotically flat coordinates.

V. CONCLUSION

In this paper we have generalized the idea of the canonical superenergy tensors and introduced *the canonical angular supermomentum tensors* for matter and for gravitation. Similarly as it was in the case of the canonical superenergy tensors, the canonical supermomentum tensors ${}_gS^{ikl}(P;v^t)$ and ${}_mS^{ikl}(P;v^t)$ which were introduced in the paper have much better geometrical properties than the canonical angular momentum pseudotensor ${}_gM^{ikl}$ and the canonical object ${}_mM^{ikl}$. Moreover, the canonical angular supermomentum tensors ${}_mS^{ikl}(P;v^t)$ and ${}_gS^{ikl}(P;v^t)$ of matter and gravitation, and their sum $S^{ikl} = {}_mS^{ikl} + {}_gS^{ikl}$ *do not require any radius vector* to be defined. The canonical angular supermomentum tensors, as well as canonical superenergy tensors considered in our previous papers, are *our proposal* as *substitutes* for the angular momentum tensors of matter and gravitation in **GR** (which actually do not exist). We want to emphasize that the canonical angular supermomentum tensors introduced and considered in the paper are *constructed locally*, defined pointwise, and explicitly depend on the four-velocity v^i of an observer **O** being at rest at the beginning **P** of the applied normal coordinates **NCS(P)**. This means they are the most appropriate for the *local analysis* of the gravitational and matter fields. However, if there exists a distinguished, unit, timelike vector field $\vec{v}$, then these tensors can also be used for the *global analysis* of the gravitational and matter fields. Such a situation appears to be valid for nearly all models of space–time (Friedman, Schwarzschild, and Kerr) we dealt with in this paper.

In the paper we applied the canonical angular supermomentum tensor ${}_gS^{ikl}(P;v^t) = (-){}_gS^{kil}(P;v^l)$ to the local analysis of the vacuum solutions to the Einstein equations which correspond to the *plane and plane-fronted* gravitational waves. We obtained *fully covariant results* in this case. Then, as already mentioned, we applied the canonical supermomentum tensor fields

$$ {}_mS^{ikl}(x;v^t), \quad {}_gS^{ikl}(x;v^t) $$

and their sum

$$ S^{ikl}(x;v^t) = {}_mS^{ikl}(x;v^t) + {}_gS^{ikl}(x;v^t) $$

for the global analysis of the Friedman universes and Schwarzschild space–time. These were extremely important applications since the Friedman universes are *a cornerstone* of relativistic cosmology and the Schwarzschild solution is of *primary astrophysical interest*. As a result of calculations, we obtained that the global, canonical angular supermomenta *trivially vanish* for both Friedman universes in all comoving coordinates and for Schwarzschild space–time. It seems to be a very reasonable result.

Comparing our results with the results obtained by using the canonical angular momentum pseudotensor ${}_gM^{ikl}$ and canonical angular momentum complex $M^{ikl} = {}_gM^{ikl} + {}_mM^{ikl}$, matter and gravitation, we conclude that in **GR**, the canonical angular supermomentum tensors are *much better tools* for the analysis (in particular, for the local analysis) of the gravitational and matter fields than the canonical angular momentum objects. The same is true when we compare the energetic and superenergetic quantities for Friedman universes[5,13,15] and for gravitational waves.[12,14]

Finally, we used the tensor ${}_gS^{abc}(P;v^l)$ for the analysis of the stationary Kerr space–time which is of great importance for relativistic astrophysics. This example can also be considered as a proof of the superiority of the canonical angular supermomentum tensor ${}_gS^{abc}(P;v^l)$ to canonical, angular momentum object ${}_gM^{ikl}$.

ACKNOWLEDGMENTS

The author wishes to thank the referee for his constructive comments. Especially, the author would like to thank Dr. M. P. Dąbrowski for his help in preparation of the English version of this paper.

APPENDIX A: BERGMANN–THOMSON ANGULAR MOMENTUM IN GR

The canonical Bergmann–Thomson total angular momentum density, matter and gravitation, can most easily be obtained in the following way. At first, let us transform the Einstein equations written in mixed form and multiplied by $\sqrt{|g|}$,

$$\sqrt{|g|}G_i^k=\beta\sqrt{|g|}T_i^k, \tag{A1}$$

to the so-called *superpotential form*

$$_EK_i{}^k={}_FU_i{}^{[kl]}{}_{,l}, \tag{A2}$$

where

$$_EK_i{}^k:=\sqrt{|g|}(T_i{}^k+{}_Et_i{}^k)$$

is the *canonical, Einstein energy-momentum complex*, matter and gravitation, and $_FU_i{}^{[kl]}$ are *von Freud superpotentials*.

From (A2), after series of operations, we get the *Bergmann–Thomson energy-momentum complex* $_{\mathrm{BT}}K^{jk}$, matter and gravitation. At first, we form the following quantity:

$$g^{ij}{}_EK_i{}^k=g^{ij}{}_FU_i{}^{[kl]}{}_{,l}=(g^{ij}{}_FU_i{}^{[kl]})_{,l}-{}_FU_i{}^{[kl]}g^{ij}{}_{,l} \tag{A3}$$

or

$$g^{ij}{}_EK_i{}^k+{}_FU_i{}^{[kl]}g^l{}_{,l}={}_FU^{j[kl]}{}_{,l}. \tag{A4}$$

Then, we write (A4) in the form

$$_{\mathrm{BT}}K^{jk}={}_FU^{j[kl]}{}_{,l}, \tag{A5}$$

where

$$_{\mathrm{BT}}K^{jk}:={}_EK^{jk}+{}_FU_i{}^{[kl]}g^{ij}{}_{,l}=:\sqrt{|g|}(T^{jk}+{}_{\mathrm{BT}}t^{jk}) \tag{A6}$$

is the *Bergmann–Thomson energy-momentum complex*, matter and gravitation, which satisfies local conservation laws

$$_{\mathrm{BT}}K^{jk}{}_{,k}=0. \tag{A7}$$

Here $_{\mathrm{BT}}t^{jk}\neq{}_{\mathrm{BT}}t^{kj}$ are the components of the so-called *Bergmann–Thomson energy-momentum pseudotensor* of the gravitational field.[22]

Finally, from (A5) we get

$$x^i{}_{\mathrm{BT}}K^{jk}-x^j{}_{\mathrm{BT}}K^{ik}=x^i{}_FU^{j[kl]}{}_{,l}-x^j{}_FU^{i[kl]}{}_{,l}=(x^i{}_FU^{j[kl]}-x^j{}_FU^{j[kl]})_{,l}-{}_FU^{j[kl]}+{}_FU^{i[kj]}, \tag{A8}$$

or

$$x^i{}_{\mathrm{BT}}K^{jk}-x^j{}_{\mathrm{BT}}K^{ik}+{}_{\mathrm{BT}}S^{ijk}=M^{[ij][kl]}{}_{,l}, \tag{A9}$$

where

$$ {}_{\mathrm{BT}}S^{ijk} := {}_F U^{i[jk]} - {}_F U^{j[ik]} = \frac{\alpha}{\sqrt{|g|}}[(-g)(g^{kj}g^{il} - g^{ki}g^{jl})]_{,l} =: \frac{\alpha}{\sqrt{|g|}} g^{[ij][kl]}{}_{,l} \tag{A10} $$

and

$$ M^{[ij][kl]} := x^i{}_F U^{j[kl]} - x^j{}_F U^{i[kl]}. \tag{A11} $$

The quantity

$$ x^i{}_{\mathrm{BT}}K^{jk} - x^j{}_{\mathrm{BT}}K^{ik} + {}_{\mathrm{BT}}S^{ijk} =: {}_{\mathrm{BT}}M^{ijk} \tag{A12} $$

is the *Bergmann–Thomson angular momentum complex* (matter and gravitation) and the quantities $M^{[ij][kl]}$ are *superpotentials*.[22]

One can interpret physically the angular momentum complex (A12) as a sum of the *orbital part*

$$ x^i{}_{\mathrm{BT}}K^{jk} - x^j{}_{\mathrm{BT}}K^{ik} = \sqrt{|g|}(x^i T^{jk} - x^j T^{ik}) + \sqrt{|g|}(x^i{}_{\mathrm{BT}}t^{jk} - x^j{}_{\mathrm{BT}}T^{ik}) $$

of the angular momentum density of matter and gravitation (matter part includes also spin density[22]) and a *spinorial part*

$$ {}_{\mathrm{BT}}S^{ijk} = {}_F U^{i[jk]} - {}_F U^{j[ik]} = \frac{\alpha}{\sqrt{|g|}} g^{[ij][kl]}{}_{,l} \tag{A13} $$

of the gravitational angular momentum density given by (8) of Sec. II. From (A5) we have

$$ 2\,{}_{\mathrm{BT}}K^{[ij]} = \sqrt{|g|}({}_{\mathrm{BT}}t^{ij} - {}_{\mathrm{BT}}t^{ji}) = {}_{\mathrm{BT}}S^{ijk}{}_{,k} = \left(\frac{\alpha}{\sqrt{|g|}}\right)_{,k} g^{[ij][kl]}{}_{,l}, \tag{A14} $$

because the dynamical energy-momentum tensor of matter T^{ik} is symmetric: $T^{ik} = T^{ki}$. Equation (A14) justifies the above physical interpretation of the pseudotensor ${}_{\mathrm{BT}}S^{ijk} = (-)_{\mathrm{BT}}S^{jik}$ as a quantity which describes *canonical spin density* of the gravitational field.[33–35]

From (A9) we have the following expression for the components $M^{ik} = (-)M^{ki}$ of the global angular momentum (matter and gravitation) of an isolated system equipped with asymtotically Lorentzian coordinates

$$ {}_{\mathrm{BT}}M^{ik} = \frac{1}{c} \oint_{\text{over sphere having } R\to\infty} (x^i{}_F U^{j[0\alpha]} - x^j{}_F U^{i[0\alpha]}) n_\alpha r^2\, d\Omega, \tag{A15} $$

where $r^2 = x^2 + y^2 + z^2$, n_a are the components of the unit (exterior) normal to the sphere, and $d\Omega = \sin\theta d\theta d\varphi$.

APPENDIX B: COMPONENTS OF ${}_gS^{abc}(P;v^l)$ FOR GRAVITATIONAL WAVES

Here we give the components

$$ {}_gS^{abc}(P;v^l) = (-)_gS^{bac}(P;v^l) \tag{B1} $$

of the canonical, gravitational supermomentum tensor for the line element (30) and (31) and for the line element (29). In order to simplify the calculations we used the *null coreper*

$$ \vartheta^0 = HdU + dV, \quad \vartheta^1 = dU, \quad \vartheta^2 = dX, \quad \vartheta^3 = dY. \tag{B2} $$

After long (but rather simple) calculations, we obtain the following nonzero components ${}_gS^{abc}(P;v^l)$ in this coreper:

$$
\begin{aligned}
&{}_gS^{120}=(-)\,{}_gS^{210}=(-)\,\frac{4\alpha v^1}{3}(v^2H_{xxu}+v^3H_{xyu}),\\
&{}_gS^{121}=(-)\,{}_gS^{211}=0,\quad {}_gS^{122}=(-)\,{}_gS^{212}=(-)\,\frac{4}{3}\alpha(v^1)^2H_{xxu},\\
&{}_gS^{123}=(-)\,{}_gS^{213}=(-)\,\frac{4\alpha}{3}(v^1)^2H_{xyu},\\
&{}_gS^{130}=(-)\,{}_gS^{310}=(-)\,\frac{4\alpha}{3}v^1(v^2H_{xyu}+v^3H_{yyu}),\\
&{}_gS^{131}=(-)\,{}_gS^{311}=0,\quad {}_gS^{132}=(-)\,{}_gS^{312}=(-)\,\frac{4}{3}\alpha(v^1)^2H_{xyu},\\
&{}_gS^{133}=(-)\,{}_gS^{313}=(-)\,\frac{4\alpha}{3}(v^1)^2H_{yyu},\quad {}_gS^{230}=(-)\,{}_gS^{320}=0,\\
&{}_gS^{232}=(-)\,{}_gS^{322}=0,\quad {}_gS^{231}=(-)\,{}_gS^{321}=0,\quad {}_gS^{233}=(-)\,{}_gS^{323}=0,\\
&{}_gS^{010}=(-)\,{}_gS^{100}=\frac{4\alpha}{3}(v^2)^2H_{xxu}+\frac{4\alpha}{3}(v^3)^2H_{yyu}+\frac{8\alpha}{3}v^2v^3H_{xyu}+\frac{2\alpha}{3}(H_{xxu}+H_{yyu}),
\end{aligned}
\tag{B3}
$$

$$
\begin{aligned}
&{}_gS^{011}=(-)\,{}_gS^{101}=0,\quad {}_gS^{012}=(-)\,{}_gS^{102}=\frac{4\alpha}{3}v^1(v^2H_{xxu}+v^3H_{xyu}),\\
&{}_gS^{020}=(-)\,{}_gS^{200}=(-)\,\frac{16\alpha}{3}(v^2)^2H_{xxx}-\frac{16\alpha}{3}(v^3)^2H_{yyx}-4\alpha v^0v^2H_{xxu}-4\alpha v^0v^3H_{xyu}\\
&\qquad -\frac{32\alpha}{3}v^2v^3H_{xxy}-\frac{8\alpha}{3}(H_{xxx}+H_{yyx}),\\
&{}_gS^{021}=(-)\,{}_gS^{201}=0,\\
&{}_gS^{022}=(-)\,{}_gS^{202}=(-)4\alpha v^0v^1H_{xxu}-\frac{16\alpha}{3}v^1v^2H_{xxx}-\frac{16\alpha}{3}v^1v^3H_{xxy}+2\alpha H_{xxu},\\
&{}_gS^{023}=(-)\,{}_gS^{203}=(-)4\alpha v^0v^1H_{xyu}-\frac{16\alpha}{3}v^1v^2H_{yxx}-\frac{16\alpha}{3}v^1v^3H_{yyx}+2\alpha H_{xyu},\\
&{}_gS^{030}=(-)\,{}_gS^{300}=(-)4\alpha v^0v^2H_{xyu}-4\alpha v^0v^3H_{yyu}-\frac{32\alpha}{3}v^2v^3H_{yyx}-\frac{16\alpha}{3}(v^2)^2H_{xxy}\\
&\qquad -\frac{16\alpha}{3}(v^3)^2H_{yyy}-\frac{8\alpha}{3}(H_{xxy}+H_{yyy}),\\
&{}_gS^{031}=(-)\,{}_gS^{301}=0,
\end{aligned}
$$

$$ {}_gS^{032}=(-)_gS^{302}=(-)4\alpha v^0v^1H_{xyu}-\frac{16\alpha}{3}v^1v^2H_{xyx}-\frac{16\alpha}{3}v^1v^3H_{xyy}+2\alpha H_{xyu}, $$

$$ {}_gS^{033}=(-)_gS^{303}=(-)4\alpha v^0v^1H_{yyu}-\frac{16\alpha}{3}v^1v^2H_{yyx}-\frac{16\alpha}{3}v^1v^3H_{yyy}+2\alpha H_{yyu}. $$

In the above formulas (B3),

$$ H_x:=\frac{\partial H}{\partial X},\quad H_{xx}:=\frac{\partial^2 H}{\partial X^2},\quad H_{xy}:=\frac{\partial^2 H}{\partial X\partial Y}, $$

and so on; v^i, $(i=0,1,2,3)$ is the four-velocity of an observer **O** being at rest at the origin **P** of the **NCS(P)** adapted to the coordinates (t,x,y,z).

For the plane wave

$$ H(X,Y,U)=(Y^2-X^2)\frac{F(U)}{2} $$

in the null coreper we have

$$ {}_gS^{120}=(-)_gS^{210}=\frac{4\alpha}{3}v^1v^2F',\quad {}_gS^{121}=(-)_gS^{211}=0, $$

$$ {}_gS^{122}=(-)_gS^{212}=\frac{4\alpha}{3}(v^1)^2F',\quad {}_gS^{123}=(-)_gS^{213}=0, $$

$$ {}_gS^{130}=(-)_gS^{310}=(-)\frac{4\alpha}{3}v^1v^3F',\quad {}_gS^{131}=(-)_gS^{311}=0, $$

$$ {}_gS^{132}=(-)_gS^{312}=0,\quad {}_gS^{133}=(-)_gS^{313}=(-)\frac{4\alpha}{3}(v^1)^2F', $$

$$ {}_gS^{23k}=(-)_gS^{32k}=0,\quad k=0,1,2,3; $$

$$ {}_gS^{010}=(-)_gS^{100}=\frac{4\alpha}{3}F'(v^{3^2}-v^{2^2}),\quad {}_gS^{011}=(-)_gS^{101}=0, \tag{B4} $$

$$ {}_gS^{012}=(-)_gS^{102}=(-)\frac{4\alpha}{3}v^1v^2F',\quad {}_gS^{013}=(-)_gS^{103}=\frac{4\alpha}{3}v^1v^3F', $$

$$ {}_gS^{020}=(-)_gS^{200}=4\alpha v^0v^2F',\quad {}_gS^{021}=(-)_gS^{201}=0, $$

$$ {}_gS^{022}=(-)_gS^{202}=2\alpha F'(2v^0v^1-1),\quad {}_gS^{023}=(-)_gS^{203}=0, $$

$$ {}_gS^{030}=(-)_gS^{300}=(-)4\alpha v^0v^3F',\quad {}_gS^{031}=(-)_gS^{301}=0, $$

$$ {}_gS^{032}=(-)_gS^{302}=0,\quad {}_gS^{033}=(-)_gS^{303}=2\alpha F'(1-2v^0v^1), $$

$F':=dF/dU$.

It is seen from the above results that the plane-fronted and plane gravitational waves *possess and carry* angular supermomentum.

APPENDIX C: ANGULAR SUPERMOMENTUM IN SPECIAL RELATIVITY

In Minkowski space–time equipped with Lorentzian coordinates ($x^0=ct, x^1=x, x^2=y, x^3=z$) we have from (1)

$$
\begin{aligned}
{}_mS^{ikl}(P;v^t) &= 2(2v^iv^p-\eta^{ip})\partial_p\hat{T}^{kl}-2(2v^kv^p-\eta^{kp})\partial_p\hat{T}^{il}\\
&= 2(\delta^{ip}\partial_p\hat{T}^{kl}-\delta^{kp}\partial_p\hat{T}^{il})\\
&= 2(\partial^i\hat{T}^{kl}-\partial^k\hat{T}^{il}) = 2(\partial_i\hat{T}^{kl}-\partial_k\hat{T}^{il}) = 2(\nabla_i\hat{T}^{kl}-\nabla_k\hat{T}^{il}), \qquad \text{(C1)}
\end{aligned}
$$

where $v^i \doteq \delta^i_0$ and $T^{ik}=T^{ki}$ are the components of the energy-momentum tensor and $\delta^{ik} := 2v^iv^k - \eta^{ik}$ is an auxilliary, positive-definite metric.

By comparison of the structure of the angular supermomentum tensor ${}_mS^{ikl}(P;v^t) = (-)_mS^{kil}(P;v^t)$ given by (C1) with the structure of the angular momentum (Carthesian) tensor of matter

$$
{}_mM^{ikl} = (-)_mM^{kil} = x^iT^{kl} - x^kT^{il}, \qquad \text{(C2)}
$$

one can easily derive a simple rule of construction of the tensor ${}_mS^{ikl}(P;v^t)$ from the angular momentum (Cartesian) tensor of matter ${}_mM^{ikl}$:

$$
x^i \Rightarrow 2\partial_i = 2\nabla_i. \qquad \text{(C3)}
$$

Of course, one can consider the tensor field

$$
{}_mS^{ikl}(x) \doteq 2[\partial_iT^{kl}(x) - \partial_kT^{il}(x)] = 2[\nabla_iT^{kl}(x) - \nabla_kT^{il}(x)]
$$

and define the global angular supermomentum ${}_mS^{ik} = (-)_mS^{ki}$ of matter by means of the following integrals:

$$
{}_mS^{ik} := \frac{1}{c}\int_{x^0=ct=\text{const}} {}_mS^{ik0}(x)\, dxdydz. \qquad \text{(C4)}
$$

If the matter energy-momentum tensor $T^{ik}=T^{ki}$ is diagonal (for example, for a perfect fluid), then we have from (C4) (as a trivial result)

$$
{}_mS^{12} = {}_mS^{13} = {}_mS^{23} = 0. \qquad \text{(C5)}
$$

Similar results can be obtained in this case for the spatial components ${}_mM^{12}$, ${}_mM^{13}$, ${}_mM^{23}$ of the global (orbital) angular momentum of matter

$$
{}_mM^{ik} := \frac{1}{c}\int_{x^0=ct=\text{const}} {}_mM^{ik0}\, dxdydz, \qquad \text{(C6)}
$$

namely,

$$
{}_mM^{12} = {}_mM^{13} = {}_mM^{23} = 0. \qquad \text{(C7)}
$$

[1] By ''standard general relativity'' we mean general relativity without any supplementary element like tetrads, second metric, or an arbitrary vector field.
[2] Called *canonical* because it can be obtained as canonical one from a suitable Lagrangian (see Ref. 3).

[3] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, Oxford, 1975).
[4] J. Garecki, in *Proceedings: Einstein Centenary Symposium*, Vol. II (Duhita Publishers c/o Einstein Foundation International, Nagpur, 1981).
[5] J. Garecki, Int. J. Theor. Phys. **35**, 2195 (1996).
[6] O. B. Zaslavskii, Class. Quantum Grav. **13**, L63 (1996).
[7] Z. D. Christodoulou and S. T. Yau, Contemp. Math. **71**, 9 (1988).

[8] W. T. Shaw, Contemp. Math. **71**, 15 (1988).
[9] R. Penrose, Proc. R. Soc. London, Ser. A **381**, 53 (1982).
[10] S. A. Hayward, Phys. Rev. D **53**, 1938 (1996).
[11] F. A. E. Pirani, Phys. Rev. **105**, 1089 (1957).
[12] J. Garecki, Acta Phys. Pol. B **11**, 255 (1980).
[13] J. Garecki, Rep. Math. Phys. (Toruń) **40**, 485 (1997).
[14] J. Garecki, Rep. Math. Phys. (Toruń) **33**, 57 (1993).
[15] J. Garecki, ''Canonical Energetic and Canonical Superenergetic Quantities for Friedman Universes,'' paper accepted for publication in Rep. Math. Phys. (Toruń) **43**, 377 (1999).
[16] J. Winicour, ''Angular Momentum in General Relativity'' in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980).
[17] J. Goldberg, ''Invariant Transformations, Conservation Laws and Energy-Momentum,'' in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980).
[18] J. A. Schouten, *Ricci Calculus* (Springer-Verlag, Berlin, 1954).
[19] A. Z. Petrov, *New Methods in General Relativity* (Nauka, Moscow, 1966) (in Russian) [English edition of Petrov's book: *Einstein Spaces* (Pergamon, New York, 1969)].
[20] C. Mǿller, Mat. Fys. Medd. K. Dan. Vidensk. Selsk. **31**, 14 (1959).
[21] J. L. Synge, *Relativity: the General Theory* (North-Holland, Amsterdam, 1960). We have used the world function $\sigma(P;y)$ in the formula (1) already.
[22] P. G. Bergmann and R. Thomson, Phys. Rev. **89**, 401 (1953).
[23] J. Garecki, *Zeszyty Naukowe WSP Szczecin*, **26**, 25 (1978) (in Polish; summary in English).
[24] Ch. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[25] J. Ehlers and W. Kundt, in *Gravitation: an introduction to current research*, edited by L. Witten (Wiley, New York, 1962).
[26] The components ${}_mM^{ikl}=\sqrt{|g|}(x^iT^{kl}-x^kT^{il})$ *do not form* any tensor density unless $\{x^i=y^i\}$, where $\{y^i\}$ mean normal coordinates. But they *always form* components of a geometric object (see, e.g., Ref. 18).
[27] S. Chandrasekhar, *The Mathematical Theory of Black Holes* (Clarendon, Oxford, 1983).
[28] Calculations of these components are very simple but tedious. The analytic form of the calculated nonzero components is too complicated to be presented here.
[29] Integrating over radial coordinate r we have to restrict ourselves[3] to $r_{\mathrm{hor}}\leqslant r<\infty$ where $r_{\mathrm{hor}}=r_g/2+\sqrt{(r_g/2)^2-a^2}$.
[30] The calculations are more complicated than calculations of the components ${}_gS^{abc}(P;v^l)$.
[31] The analytic form of these components is much more complicated than the form of the components of ${}_gS^{abc}(P;v^l)$.
[32] See Ref. 29.
[33] In special relativity the antisymmetric part of a conserved energy-momentum tensor is proportional to the ordinary divergence of the three-index tensor which describes intrinsic angular momentum density (see, e.g., Refs. 34 and 35). Here we follow this line.
[34] V. Sabatta and M. Gasperini, *Introduction to Gravitation* (World Scientific, Singapore, 1985).
[35] W. Kopczyński and A. Trautman, *Space–time and Gravitation* (Wiley, New York, 1992).

Scale symmetries of spherical string fluids

E. N. Glass[a] and J. P. Krisch
Department of Physics, University of Michigan, Ann Arbor, Michigan 48109

(Received 8 March 1999; accepted for publication 18 May 1999)

We consider homothetic maps in a family of spherical relativistic star models. A generalization of Vaidya's radiating metric provides a fluid atmosphere of radiation and strings. The similarity structure of the string fluid is investigated. © *1999 American Institute of Physics.* [S0022-2488(99)03308-3]

I. INTRODUCTION

Metric symmetries have always played a large role in the development of exact solutions to the Einstein field equations. Often a choice of metric symmetry is made based on an assumed symmetry of the matter distribution, i.e., spherical symmetry for astrophysical objects or cylindrical symmetry for a simple string.[1] A homothetic motion (homothety) describes the symmetry of scale transformations, and homothetic symmetry has been called ''similarity of the first kind'' by Cahill and Taub.[2] One must distinguish between *geometrical* and *physical* self-similarity. Geometrical similarity is a property of the space–time metric, whereas physical similarity is a property of the matter fields. These need not be equivalent and the relationship between them also depends on the nature of the matter. Yavuz and Yilmaz[3] recently investigated inheritance symmetries wherein the stress energy inherits metric symmetries of the type

$$\mathcal{L}_\xi g_{ab} = 2\Psi g_{ab},$$

where $\mathcal{L}_\xi$ is the Lie derivative along the vector ξ. Some of the possibilities are

$\Psi = \Psi(x^a)$, ξ is a conformal Killing vector,

$\Psi = 1$, ξ is a homothetic vector,

$\Psi = 0$, ξ is a Killing vector.

Carter and Henriksen[4] have introduced the idea of *kinematic self-similarity* in the context of relativistic fluid mechanics and an extended analysis has been given by Coley.[5] A kinematic self-similarity vector satisfies the conditions

$$\mathcal{L}_\xi u_a = \text{const}\, u_a,$$

$$\mathcal{L}_\xi h_{ab} = 2h_{ab},$$

where $h_{ab} = g_{ab} - u_a u_b$ is the first fundamental form of the three-spaces orthogonal to u^a. The case const$\neq 1$ is called ''similarity of the second kind.''

In this work we apply the ideas of scaling and homothety to a string fluid atmosphere. Since our primary interest is in the extended Schwarzschild mass function $m(u,r)$ and the related string atmosphere, we apply scaling to the mass in two different ways. First, we assume diffusive mass transport and investigate the symmetries of the diffusion equation and second, we investigate the scaling properties of the metric and from those derive mass transport equations.

In Sec. II we briefly describe the Schwarzschild string fluid atmosphere. Section III studies the symmetry map which takes the diffusion equation to an ordinary differential equation. New

[a]Permanent address: Physics Department, University of Windsor, Ontario N9B 3P4, Canada.

diffusion solutions are found. Geometric symmetries, homothetic and conformal, are developed in Sec. IV. Mass transport is discussed in Sec. V. One of the results of the homothetic analysis are new self-similar solutions to the Einstein equations.

Our sign conventions are $2A_{c;[ab]}=A_e R^e{}_{cab}$, and $R_{ab}=R^e{}_{abe}$. Latin indices range over $(0,1,2,3)=(u,r,\vartheta,\varphi)$. Overdots abbreviate $\partial/\partial u$, and primes abbreviate $\partial/\partial r$. Overhead carets denote unit vectors. We use units where $G=c=1$. Einstein's field equations are $G_{ab}=-8\pi T_{ab}$, and the metric signature is $(+,-,-,-)$.

II. STRING FLUID ATMOSPHERE

Recently, Glass and Krisch[6,7] showed that there can be a spherically symmetric string fluid atmosphere outside a Schwarzschild horizon. The space–time metric is

$$ds^2_{\mathrm{GK}}=A\,du^2+2\,du\,dr-r^2(d\vartheta^2+\sin^2\vartheta\,d\varphi^2), \tag{1}$$

where $A=1-2m(u,r)/r$. Initially $m(u,r)=m_0$ provides the vacuum Schwarzschild solution in the region $r>2m_0$. The metric can be written in a natural basis as

$$g^{\mathrm{GK}}_{ab}=\hat{v}_a\hat{v}_b-\hat{r}_a\hat{r}_b-\hat{\vartheta}_a\hat{\vartheta}_b-\hat{\varphi}_a\hat{\varphi}_b, \tag{2}$$

where the unit vectors are defined by

$$\hat{v}_a dx^a=A^{1/2}\,du+A^{-1/2}\,dr,\quad \hat{v}^a\partial_a=A^{-1/2}\partial_u, \tag{3a}$$

$$\hat{r}_a\,dx^a=A^{-1/2}\,dr,\quad \hat{r}^a\partial_a=A^{-1/2}\partial_u-A^{1/2}\partial_r, \tag{3b}$$

$$\hat{\vartheta}_a\,dx^a=r\,d\vartheta,\quad \hat{\vartheta}^a\partial_a=-r^{-1}\partial_\vartheta, \tag{3c}$$

$$\hat{\varphi}_a\,dx^a=r\sin\vartheta\,d\varphi,\quad \hat{\varphi}^a\partial_a=-(r\sin\vartheta)^{-1}\partial_\varphi. \tag{3d}$$

$\hat{v}^a$ is hypersurface-orthogonal with h_{ab} the first fundamental form of the hypersurface,

$$h_{ab}dx^a dx^b=(g^{\mathrm{GK}}_{ab}-\hat{v}_a\hat{v}_b)dx^a dx^b=-A^{-1}dr^2-r^2(d\vartheta^2+\sin^2\vartheta\,d\varphi^2). \tag{4}$$

The kinematics of the $\hat{v}^a$ flow are described by

$$\hat{v}^a_{;b}=a^a\hat{v}_b+\sigma^a{}_b-(\Theta/3)(\hat{r}^a\hat{r}_b+\hat{\vartheta}^a\hat{\vartheta}_b+\hat{\varphi}^a\hat{\varphi}_b), \tag{5}$$

where

$$a^a=[\dot{m}/r+A\partial_r(m/r)]A^{-3/2}\hat{r}^a, \tag{6a}$$

$$\sigma^a{}_b=(\Theta/3)(-2\hat{r}^a\hat{r}_b+\hat{\vartheta}^a\hat{\vartheta}_b+\hat{\varphi}^a\hat{\varphi}_b), \tag{6b}$$

$$\Theta=(\dot{m}/r)A^{-3/2}. \tag{6c}$$

The string distribution is described by a string bivector Σ_{ab}. Spherical symmetry demands that the averaged string bivector will describe a world-sheet in either the (u,r) or the (ϑ,φ) plane. The string bivector is timelike and given by

$$\Sigma^{ac}=\hat{r}^a\hat{v}^c-\hat{r}^c\hat{v}^a, \tag{7}$$

where $\Sigma^{ac}\Sigma_c{}^b=\hat{v}^a\hat{v}^b-\hat{r}^a\hat{r}^b$. The two-surfaces spanned by Σ_{ab} are orthogonally transitive to the two-surfaces spanned by the dual bivector

$$\Sigma^*_{ab}=\hat{\vartheta}_a\hat{\varphi}_b-\hat{\vartheta}_b\hat{\varphi}_a, \tag{8}$$

which follows from the Frobenius surface-forming condition satisfied by Σ_{ab}. It is also true that $\Sigma_a^{*c}\Sigma_{cb}^{*}=\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b$.

The Einstein tensor computed from (1) can be written as a two-fluid system $G_{ab}^{\text{null}}+G_{ab}^{\text{matter}}$.

$$G_{ab}=(2\dot{m}/r^2)l_a l_b-(2m'/r^2)(\hat{v}_a\hat{v}_b-\hat{r}_a\hat{r}_b)+(m''/r)(\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b), \tag{9}$$

where $l_a dx^a=du$. The Einstein field equations $G^a{}_{b;a}=0$ are satisfied for arbitrary $m(u,r)$.

In Glass and Krisch[6,7] mass transport was modeled by diffusion, and the diffusion equation used is given by

$$\dot{\rho}=\mathcal{D}r^{-2}\partial_r(r^2\partial_r\rho), \tag{10}$$

where $\mathcal{D}$ is the positive coefficient of self-diffusion (taken to be constant).

III. SIMILARITY MAP OF THE DIFFUSION EQUATION

There is a similarity technique explained by Bluman and Cole[8] that maps the diffusion equation into an ordinary differential equation. Our primary interest is in the Schwarzschild mass function $m(u,r)$. New functional solutions for $m(u,r)$ are new solutions to the field equations for the parameter extended radiating atmosphere. The behavior of $m(u,r)$ describes the string fluid atmosphere beyond the Schwarzschild horizon through the relations $\dot{m}=4\pi\mathcal{D}r^2\rho'$ and $4\pi\rho=m'/r^2$. The mass function obeys a diffusion equation

$$\dot{m}=\mathcal{D}r^2\partial_r(r^{-2}\partial_r m) \tag{11}$$

with homogeneous solution $m_{\text{hom}}(r)=m_0+\frac{4}{3}\pi r^3\rho_0$ which can be added to each time-dependent solution.

The similarity technique (for a fully general analysis see Bluman and Kumei[9]) requires one to introduce an independent dimensionless variable. A standard choice in diffusion problems is the Boltzmann transformation:[10]

$$\eta=r(4\mathcal{D}u)^{-1/2}. \tag{12}$$

[Note that as a mapping from (u,r) to $(u^{-1/2},\eta)$ the Jacobian is singular implying a breakdown of the $1-1$ mapping along r.] The argument of the equation, $m(u,r)$, is replaced by a dimensionless function $F(\eta)$: We look for a general solution of the form

$$m(u,r):=c_0 r^\alpha u^\beta F(\eta). \tag{13}$$

The constant c_0 is intended to map the dimensions of $r^\alpha u^\beta$ to mass for arbitrary constants α and β. Upon substituting Eq. (13) into the diffusion equation (11) we obtain the ordinary differential equation

$$F_{\eta\eta}+2[(\alpha-1)\eta^{-1}+\eta]F_\eta+[\alpha(\alpha-3)\eta^{-2}-4\beta]F=0, \tag{14}$$

where F_η abbreviates $dF/d\eta$.

There are many analytic solutions of Eq. (14) which depend on the values of α and β. The choice $\alpha=\beta=0$ has the differential equation $F_{\eta\eta}+2(\eta-1/\eta)F_\eta=0$ with solution

$$F(\eta)=k_0+k_1[-\eta e^{-\eta^2}+(\sqrt{\pi}/2)\text{erf}(\eta)], \tag{15}$$

where $\text{erf}(\eta):=(2/\sqrt{\pi})\int_0^\eta \exp(-s^2)ds$, $\lim_{\eta\to 0}\text{erf}(\eta)=2\eta/\sqrt{\pi}$. This is the mass solution given in Eq. (40) of Glass and Krisch[7] (with $k_0=0$ and with the homogeneous solution m_{hom} added). At fixed time u, it describes a mass with value $m_{\text{hom}}+c_0k_1$ as $\eta\to\infty$. At late times c_0k_1 is radiated away. There is no length scale in this description so the m_{hom} atmosphere is unbounded.

Other choices can be made, for example $\alpha=n$, $\beta=-n/2$. This choice has $\text{const}\times r^n u^{-n/2}=\eta^n$ and one can solve Eq. (14) or see directly from Eq. (13) that

$$F(\eta)=\eta^{-n}.$$

If we write $F(\eta)=\eta^{-n}H(\eta)$ then $H(\eta)$ satisfies the case $\alpha=\beta=0$ and we have a new family of solutions parametrized by n:

$$F(\eta)=\eta^{-n}[k_0+k_1[-\eta e^{-\eta^2}+(\sqrt{\pi}/2)\text{erf}(\eta)]]. \tag{16}$$

Solution (15) is included here when $n=0$.

IV. SYMMETRIES

Because the string fluid naturally lives on a two-dimensional world sheet, the question of the symmetries of these two-dimensional subspaces is interesting. We examine how the mass distribution and stress energy content reflect the separate two-surface symmetries

$$\begin{aligned} \mathcal{L}_\xi(\hat{v}_a\hat{v}_b-\hat{r}_a\hat{r}_b)&=2\mu(\hat{v}_a\hat{v}_b-\hat{r}_a\hat{r}_b),\\ \mathcal{L}_\xi(\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b)&=2\nu(\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b). \end{aligned} \tag{17}$$

For similarity of the second kind, the map action must be

$$\begin{aligned} \mathcal{L}_\xi\hat{v}_a&=\gamma\hat{v}_a, \quad \gamma\neq 1,\\ \mathcal{L}_\xi\hat{r}_a&=\hat{r}_a,\\ \mathcal{L}_\xi(\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b)&=2(\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b). \end{aligned} \tag{18}$$

A. Homothetic map

The similarity vector which preserves the distinct two-surfaces of the matter distribution in Eq. (17) is

$$\xi^a\partial_a=[\nu u_0+(2\mu-\nu)u)]\partial_u+\nu r\partial_r, \tag{19}$$

with kinematic transformations

$$\mathcal{L}_\xi\hat{v}_a=\mu\hat{v}_a, \quad \mathcal{L}_\xi\hat{r}_a=\mu\hat{r}_a, \tag{20a}$$

$$\mathcal{L}_\xi\hat{\vartheta}_a=\nu\hat{\vartheta}_a, \quad \mathcal{L}_\xi\hat{\varphi}_a=\nu\hat{\varphi}_a, \tag{20b}$$

when the metric function A satisfies $(\kappa:=2\mu/\nu-1)$

$$\psi\dot{A}/A+rA'/A+\kappa-1=0, \tag{21}$$

with $\psi(u):=u_0+\kappa u$. The constraint (21) requires the mass function to have the form

$$r-2m(u,r)=\psi^{(2-\kappa)/\kappa}f(\psi/r^\kappa), \tag{22}$$

where f is an arbitrary function.

If $\mu=\nu=\kappa=1$ then the map is homothetic with $\mathcal{L}_\xi g_{ab}=2g_{ab}$.

B. Another homothetic map

The case $\kappa=0$ requires a separate solution. The metric function A satisfies

$$u_0\dot{A}/A+rA'/A=1. \tag{23}$$

Constraint (23) has the integral

$$r-2m(u,r)=r_1 e^{2u/u_0}\tilde{f}(e^{u/u_0}r_0/r) \tag{24}$$

with $\tilde{f}$ an arbitrary function. When $\nu=1$ and $\mu=1/2$ the u dependence is eliminated from ξ^a and the transformation acts on the (ϑ,φ) two-surfaces homothetically

$$\mathcal{L}_\xi(\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b)=2(\hat{\vartheta}_a\hat{\vartheta}_b+\hat{\varphi}_a\hat{\varphi}_b)$$

but preserves the scale of the string two-surfaces,

$$\mathcal{L}_\xi(\hat{v}_a\hat{v}_b-\hat{r}_a\hat{r}_b)=\hat{v}_a\hat{v}_b-\hat{r}_a\hat{r}_b\,.$$

C. Interpreting the scale parameter

Under the action of the homothety $\xi^a\partial_a=(u_0+u)\partial_u+r\partial_r$ the acceleration of $\hat{v}^a$, given in Eq. (6a), has the following Lie derivative: with $a^b=a\hat{r}^b$, $a:=[\dot{m}/r+A\partial_r(m/r)]A^{-3/2}$,

$$\mathcal{L}_\xi a^b=\left(\frac{a_\xi}{a}-1\right)a^b, \tag{25}$$

where $a_\xi:=[(u_0+u)\partial_u+r\partial_r]a$. Similarly the rate-of-shear given in Eqs. (6b) and (6c) obeys

$$\mathcal{L}_\xi\sigma^a{}_b=\left(\frac{\Theta_\xi}{\Theta}-1\right)\sigma^a{}_b\,, \tag{26}$$

where $\Theta_\xi=[(u_0+u)\partial_u+r\partial_r]\Theta$.

There is no information to be gained by analyzing the scaling properties of the Raychaudhuri equation

$$a^b_{;b}-\sigma_{ab}\sigma^{ab}-\Theta^2/3-\Theta_{,a}\hat{v}^a=-R_{ab}\hat{v}^a\hat{v}^b,$$

since it is identically satisfied by g^{GK}_{ab}.

D. Conformal map

The case $\mu=\nu=\kappa=1$, with $\psi=u_0+u$, has an interesting conformal symmetry. We see from Eq. (22) that $A=(\psi/r)f(\psi/r)=F(\psi/r)$. Metric (1) is written as

$$ds^2_{\mathrm{GK}}=F(\psi/r)du^2+2\,du\,dr-r^2\,d\Omega^2. \tag{27}$$

We define a new coordinate $y:=r/\psi$ and rewrite (27) as

$$ds^2_{\mathrm{GK}}=[F(1/y)+2y]du^2+2\psi\,du\,dy-y^2\psi^2\,d\Omega^2. \tag{28}$$

Now we factor out ψ^2 and introduce a new time coordinate $dw:=du/\psi$ to obtain

$$ds^2_{\mathrm{GK}}=\psi^2[(F+2y)dw^2+2\,dw\,dy-y^2\,d\Omega^2].$$

Upon choosing $F(1/y)=1-2M(y)/y-2y$, $M(y)$ arbitrary, we have

$$ds^2_{\rm GK}=e^{2w}[(1-2M/y)dw^2+2\,dw\,dy-y^2d\Omega^2]. \tag{29}$$

The argument above shows that the similarity transformation generated by vector $\xi^a\partial_a=(u_0+u)\partial_u+r\partial_r$ conformally relates the radiating string atmosphere of metric (1) to a previously identified family of static string atmospheres,[7] i.e.,

$$\mathcal{L}_\xi g^{\rm GK}_{ab}=2e^{2w}g^{\rm static}_{ab}. \tag{30}$$

E. Similarity of the second kind

For similarity vector $\xi^a\partial_a=(u+u_0)\partial_u+r\partial_r$ the metric function A must satisfy

$$(u_0+u)\dot{A}/A+rA'/A=\gamma-1. \tag{31}$$

Equation (31) has solution

$$r-2m(u,r)=r_2(u_0+u)^\gamma h[(u_0+u)/r], \tag{32}$$

where h is an arbitrary function and $\gamma\neq 1$.

V. MASS TRANSPORT

The mass functions found by similarity analysis obey certain transport equations. Most of the transport equations have the form of the ''telegrapher'' equation. This can describe dispersive and lossy electromagnetic wave motion.[11] Some forms have been interpreted by Kac[12] as a random Poisson process. Mass transport through the atmosphere is affected by the homothetic symmetries. The transport equations can be constructed from the similarity solutions of Sec. IV.

A. $\kappa=0$ homothety

Differentiation of Eq. (23), a constraint on the mass function, yields an inhomogeneous wave equation

$$\ddot{A}-3\dot{A}/u_0-(r/u_0)^2\nabla^2A=-2A/u_0^2. \tag{33}$$

B. $\kappa=1$ homothety

Recall metric function $A=1-2m(u,r)/r$. One can see directly from Eq. (22) that $A=(\psi/r)f(\psi/r)=F(\psi/r)$ with $\psi=u_0+u$. A, and thus m/r, satisfies a wave equation on the flat tangents to the $\hat{v}_a\hat{v}_b-\hat{r}_a\hat{r}_b$ two-spaces. Defining $\tau=\ln(\psi)$ and $z=\ln(r)$ we have

$$A=F(\tau-z). \tag{34}$$

It is clear that A, generated by homothety $\xi^a\partial_a=(u_0+u)\partial_u+r\partial_r$, satisfies the wave equation

$$\frac{\partial^2A}{\partial\tau^2}-\frac{\partial^2A}{\partial z^2}=0. \tag{35}$$

Alternatively, we can find a wave equation on the curved manifold by writing

$$\partial F/\partial u=(1/r)\hat{F},\quad \partial F/\partial r=-(\psi/r^2)\hat{F}$$

where $\hat{F}$ is the derivative of F with respect to its argument. It follows that

$$\ddot{A}=(1/r^2)\hat{\hat{F}},\quad (r^2A')'=(\psi^2/r^2)\hat{\hat{F}}.$$

Thus

$$\ddot{A}-v_s^2\nabla^2 A=0 \tag{36}$$

where $\nabla^2=r^{-2}(\partial/\partial r)r^2(\partial/\partial r)$ and $v_s=r/\psi$. The wave speed varies with u and r.

If v_s were constant v, then Eq. (36) would have the general solution

$$A(u,r)=\frac{f(r-vu)}{r}+\frac{g(r+vu)}{r} \tag{37}$$

in terms of two arbitrary functions f and g. Substituting $A=f/r$ into (36) one finds

$$(v^2-v_s^2)\hat{\hat{f}}=0.$$

This reflects ''damped, yet relatively undistorted, progressing wave solutions,'' [13] a special case of the telegrapher's equation.

For new time coordinate $t\mapsto u+u_0=e^{t/t_0}$ and with $A_t:=\partial A/\partial t$, Eq. (36) transforms to

$$A_{tt}-A_t/t_0-(r/t_0)^2\nabla^2 A=0. \tag{38}$$

C. $\kappa \geqslant 2$ two-surface symmetry

With $A=1-2m(u,r)/r=\psi^{(2-\kappa)/\kappa}r^{-1}f(\psi/r^\kappa)$ we can write

$$A=r^{1-\kappa}H(\psi/r^\kappa),\quad H:=(\psi/r^\kappa)^{(2-\kappa)/\kappa}f.$$

Differentiation yields

$$\ddot{A}=\kappa^2 r^{1-3\kappa}\hat{H}$$

and

$$(r^2A')'=(1-\kappa)(2-\kappa)r^{1-\kappa}H-3\kappa(1-\kappa)\psi r^{1-2\kappa}\hat{H}+\kappa^2\psi^2\hat{H}.$$

It follows that

$$(r^2A')'=(1-\kappa)(2-\kappa)A-3(1-\kappa)\psi\dot{A}+\psi^2\ddot{A}.$$

Transforming to a new time coordinate $e^{t/t_0}=\psi=u_0+\kappa u$ yields the inhomogeneous wave equation

$$A_{tt}+(2-3/\kappa)A_t/t_0-(r/t_0)^2\nabla^2 A=(1-1/\kappa)(2/\kappa-1)A/t_0^2, \tag{39}$$

where $A_t:=\partial A/\partial t$.

D. Similarity of the second kind

Differentiation of the constraint on metric function A, Eq. (31), yields the homogeneous wave equation

$$\ddot{A}+v_s\left(\frac{\gamma-1}{r}\right)\dot{A}-v_s^2\nabla^2 A+v_s^2\left(\frac{\gamma-1}{r^2}\right)(rA')=0, \tag{40}$$

where $v_s=r/(u_0+u)$ and $\gamma\neq 1$. As above, the wave speed varies with u and r.

VI. DISCUSSION

Similarity is physically important since scaling behavior may offer clues about possible relationships between macroscopic and microscopic physics (i.e., Ehrenfest's classical adiabatic invariants and quantization rules). Using scaling, one can model long term behaviors with single solutions to the field equations in which only the scaling variable changes as a function of time. Self-similar behavior is an important aspect of many evolutionary processes both linear and nonlinear.[14] The simplifications of the nonlinear field equations of general relativity are a good example of the value of similarity methods. In addition, we have seen that the special homothety of fluid two-surfaces can be associated with self-similar behavior in the fluid parameters.

In this paper our primary interest is in the extended Schwarzschild mass function $m(u,r)$ and the related string atmosphere. We applied scaling to the mass in two different ways. First, we assumed a mass transport and investigated the scaling properties and second, we investigated the scaling properties of the metric and from those derived mass transport equations. In the first case, assuming diffusive mass transport with a Boltzmann scaling variable, we developed a new family of diffusive mass functions and the associated family of string atmospheres. In the second case, we examined the scaling symmetries of the orthogonal two-surfaces (u,r) and (ϑ,φ). A two-parameter similarity generator acted separately on the (u,r) two-surface containing the string fluid and the orthogonal (ϑ,φ) two-surface subject to the mass parameter obeying a constraining first order differential equation. The similarity map affects all metric components equally when the parameters are both equal to 1. For this case, where the transformation is a homothety for the entire space–time, the mass constraint conformally relates a radiating string atmosphere and a static atmosphere. Other parameter choices could be made, for example, the choices which remove time dependence from the generator. This time independent mapping acts on the (ϑ,φ) two-surface homothetically while preserving the scale of the (u,r) string two-surface. For all the parameter choices associated with the scaling action of the generator, a mass transport equation is implied. This equation is, in general, the telegrapher's equation. The telegrapher's equation and the diffusion equation have both macroscopic and microscopic interpretations.[12,15,16] The appearance of both of these mass transport equations in conjunction with the description of a macroscopic string fluid atmosphere is suggestive of the quantum nature of the fundamental string fluid bits. The classical continuum fluid describes only the averaged fluid behavior, with the mass transport equations suggesting the underlying quantum nature of the fluid.

[1] W. A. Hiscock, Phys. Rev. D **31**, 3288 (1985).
[2] A. H. Cahill and M. E. Taub, Commun. Math. Phys. **21**, 1 (1971).
[3] I. Yavuz and I. Yilmaz, Gen. Relativ. Gravit. **29**, 1295 (1997).
[4] B. Carter and R. N. Henriksen, J. Math. Phys. **32**, 2580 (1991).
[5] A. A. Coley, Class. Quantum Grav. **14**, 87 (1997).
[6] E. N. Glass and J. P. Krisch, Phys. Rev. D **57**, R5945 (1998).
[7] E. N. Glass and J. P. Krisch, Class. Quantum Grav. **16**, 1175 (1999).
[8] G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations* (Springer, Berlin, 1974).
[9] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations* (Springer, New York, 1989).
[10] R. Ghez, *A Primer of Diffusion Problems* (Wiley, New York, 1988), p. 140.
[11] J. A. Stratton, *Electromagnetic Theory* (McGraw–Hill, New York, 1941), p. 346.
[12] B. Gaveau, T. Jacobson, M. Kac, and L. S. Schulman, Phys. Rev. Lett. **53**, 419 (1984).
[13] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Wiley, New York, 1962), Vol. II, p. 193.
[14] P. M. Benoit and A. A. Coley, Class. Quantum Grav. **15**, 2397 (1998).
[15] E. Nelson, Phys. Rev. **150**, 1079 (1966).
[16] E. Nelson, *Quantum Fluctuations* (Princeton University Press, Princeton, 1985).

The Levi-Civita space–time as a limiting case of the γ space–time

L. Herrera[a)]
Escuela de Física, Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela and Centro de Astrofísica Teórica, Merida, Venezuela

Filipe M. Paiva[b)]
Departamento de Física Teórica, Universidade do Estado do Rio de Janeiro, Rua São Francisco Xavier 524, 20550-013 Rio de Janeiro, RJ, Brazil

N. O. Santos[c)]
Laboratório de Astrofísica e Radioastronomia, Centro Regional Sul de Pesquisas Espaciais INPE/MCT, Cidade Universitária, 97105-900 Santa Maria RS, Brazil

(Received 26 January 1999; accepted for publication 7 April 1999)

It is shown that the Levi-Civita metric can be obtained from a family of the Weyl metric, the γ metric, by taking the limit when the length of its Newtonian image source tends to infinity. In this process a relationship appears between two fundamental parameters of both metrics.
[S0022-2488(99)02507-4]

I. INTRODUCTION

One of the most interesting metrics of the family of Weyl solutions[1] is the so-called γ metric.[2,3] This metric, which is also known as the Zipoy–Voorhees metric,[4] is continuously linked to the Schwarzschild space–time through one of its parameters and corresponds to a solution of the Laplace equation in cylindrical coordinates. Its Newtonian image source[5] is given by a finite rod of matter. For a particular value of the mass density of the rod, the metric becomes spherically symmetric (Schwarzschild metric).

In this article we show that by extending the length of the rod to infinity we obtain the Levi-Civita space–time. At the same time a link is established between the parameter $\gamma/2$, measuring the mass density of the rod in the γ metric, and the parameter σ, which is thought to be related to the linear energy density of the source of the Levi-Civita space–time.[5] Since σ is the real source, not the Newtonian image source, and $\gamma/2$ measures the line mass density of the Newtonian image source, not of the real source, our result illustrates further the difficulties appearing in the interpretation of the Levi-Civita metric as representing an infinite line mass of density σ.[6]

In Sec. II we describe the γ metric. In Sec. III we show that it has a limit on the Levi-Civita space–time. In Sec. IV some other limits are studied in order to build a limiting diagram for the γ metric. Finally Sec. V presents our conclusions.

II. THE γ METRIC

In cylindrical coordinates, static axisymmetric solutions to Einstein's equations are given by the Weyl metric[1]

$$ds^2=e^{2\lambda}\,dt^2-e^{-2\lambda}[e^{2\mu}(d\rho^2+dz^2)+\rho^2\,d\phi^2], \tag{2.1}$$

[a)]Postal address: Apartado 80793, Caracas 1080A, Venezuela; electronic mail: laherrera@telcel.net.ve
[b)]Electronic mail: fmpaiva@symbcomp.uerj.br
[c)]Electronic mail: nos@lacesm.ufsm.br

with

$$\lambda_{,\rho\rho}+\rho^{-1}\lambda_{,\rho}+\lambda_{,zz}=0, \tag{2.2}$$

and

$$\mu_{,\rho}=\rho(\lambda_{,\rho}^2-\lambda_{,z}^2), \tag{2.3}$$

$$\mu_{,z}=2\rho\lambda_{,\rho}\lambda_{,z}, \tag{2.4}$$

where a comma denotes partial derivation. Observe the most amazing fact, as Synge writes,[1] that (2.2) is just the Laplace equation for λ in Euclidean space. The γ metric is defined by[2]

$$e^{2\lambda}=\left[\frac{R_1+R_2-2m}{R_1+R_2+2m}\right]^{\gamma}, \tag{2.5}$$

$$e^{2\mu}=\left[\frac{(R_1+R_2+2m)(R_1+R_2-2m)}{4R_1R_2}\right]^{\gamma^2}, \tag{2.6}$$

where

$$R_1^2=\rho^2+(z-m)^2, \quad R_2^2=\rho^2+(z+m)^2 \tag{2.7}$$

and γ is a constant. It is worth noticing that λ as given by (2.5) corresponds to the Newtonian potential of a line segment of mass density $\gamma/2$ and length $2m$, symmetrically distributed along the z axis. The particular case $\gamma=1$ corresponds to the Schwarzschild metric. This is more easily seen using Erez–Rosen coordinates,[4] given by

$$\rho^2=(r^2-2mr)\sin^2\theta, \quad z=(r-m)\cos\theta, \tag{2.8}$$

which yields the line element[2]

$$ds^2=F\,dt^2-F^{-1}[G\,dr^2+H\,d\theta^2+(r^2-2mr)\sin^2\theta\,d\phi^2], \tag{2.9}$$

where

$$F=\left(1-\frac{2m}{r}\right)^{\gamma}, \tag{2.10}$$

$$G=\left(\frac{r^2-2mr}{r^2-2mr+m^2\sin^2\theta}\right)^{\gamma^2-1}, \tag{2.11}$$

$$H=\frac{(r^2-2mr)^{\gamma^2}}{(r^2-2mr+m^2\sin^2\theta)^{\gamma^2-1}}. \tag{2.12}$$

Now, it is easy to check that $\gamma=1$ corresponds to the Schwarzschild metric. The total mass of the source is $M=\gamma m$,[2,3] and its quadrupole moment Q is given by

$$Q=\frac{\gamma}{3}M^3(1-\gamma^2), \tag{2.13}$$

so that $\gamma>1(\gamma<1)$ corresponds to an oblate (prolate) spheroid. We shall now show that *elongating* the Newtonian image source to infinity we obtain the Levi-Civita space–time. To achieve that, use will be made of the Cartan scalars. In Sec. III these scalars are obtained for the γ metric, and are compared to the corresponding quantities of the Levi-Civita metric in the limit $m\rightarrow\infty$.

III. THE LEVI-CIVITA LIMIT

Since the limit $m\rightarrow\infty$ taken on the γ metric in the form (2.1) diverges, we use the Cartan scalar approach to obtain a finite limit.[7,8]

It is known[9] that the so-called 14 algebraic invariants (and even all the polynomial invariants of any order) are not sufficient for locally characterizing a space–time, in the sense that two metrics may have the same set of invariants and not be equivalent. As an example, all these invariants vanish for both Minkowski and plane-wave[9] space–times and they are not the same. A complete local characterization of space–times may be done by the Cartan scalars. Briefly, the Cartan scalars are the components of the Riemann tensor and its covariant derivatives (up to possibly the tenth order) calculated in a constant frame.[10,9,11,12]

Therefore it is possible to establish unambiguously the local equivalence between two given metrics by comparing their respective Cartan scalars, in other words: Two metrics are equivalent if and only if there exist coordinate and Lorentz transformations which transform the Cartan scalars of one of the metrics into the Cartan scalars of the other. It should be stressed that, although the Cartan scalars provide a local characterization of the space–time, global properties such as topological defects do not probably appear in them.

In practice, the Cartan scalars are calculated with the Karlhede algorithm,[11] using the spinorial formalism. For the purpose here, the relevant quantities are the Weyl spinor Ψ_A, and its first covariant symmetrized derivative $\nabla\Psi_{AB'}$, which represent the Weyl tensor and its covariant derivative. Due to the amount of calculations, the computer algebra systems SHEEP/CLASSI[10,9] and MAPLE were used throughout this section.

In order to calculate the Cartan scalars for the γ metric, we take the line element in spherical coordinates [Eq. (2.9)] written in the same tetrad basis used in Ref. 2. In the zeroth order we find that the Ricci spinor and curvature scalar vanish and the Weyl spinor satisfies the relation: $\Psi_0=\Psi_4$, $\Psi_1=-\Psi_3$, $\Psi_2\neq 0$. It can be easily shown that this corresponds to a Petrov type I metric, which therefore has no isotropies. For putting these Cartan scalars in a canonical form, two tetrad transformations are done, which in the spinorial formalism are given by

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}, \quad \begin{bmatrix} A & 0 \\ 0 & 1/A \end{bmatrix}. \tag{3.1}$$

The first transformation puts the zeroth-order Cartan scalars in the form: $\Psi_0'\neq 0$, $\Psi_1'=0$, $\Psi_2'\neq 0$, $\Psi_3'=0$, $\Psi_4'\neq 0$. The second transformation with $A=(\Psi_4'/\Psi_0')^{1/8}$ gives finally: $\Psi_0''=\Psi_4''$, $\Psi_1''=0$, $\Psi_2''\neq 0$, $\Psi_3''=0$, which is the canonical form for Petrov type I metrics.

We come out with two independent functions of the coordinates r and θ [Eqs. (3.2) and (3.3)]. So, up to zeroth order, the isometry group is of dimension $4-2=2$ (where 4 is the dimension of the space–time). Since the metric is independent of the coordinates t and ϕ, its isometry group is of dimension 2. Therefore, the first-order Cartan scalars will present no new information about isometries and the Karlhede algorithm ends in the first order, with the Ψ_A, and $\nabla\Psi_{AB'}$ in the canonical basis defined by the transformation (3.1).

Instead of calculating the first-order Cartan scalars in the canonical basis, for computational reasons they were calculated in the initial basis and afterwards transformed to the canonical basis. Finally, to have the Cartan scalars in the cylindrical coordinate system one has to invert the coordinate transformation from cylindrical to spherical given by Eq. (2.8) (this leads to $r=[\sqrt{(z-m)^2+\rho^2}+\sqrt{(z+m)^2+\rho^2}+2m]/2$ and $\cos\theta=[\sqrt{(z+m)^2+\rho^2}-\sqrt{(z-m)^2+\rho^2}]/2m$) and apply it to the Cartan scalars, remembering that they transform like scalars.

In the canonical basis, in cylindrical coordinates, the zeroth-order Cartan scalars of the γ metric [Eqs. (2.1) and (2.5)–(2.7)] are (dropping the primes):

$$\Psi_2 = \frac{e^{2\lambda}}{e^{2\mu}} \frac{m\gamma(R_1+R_2-2\gamma m)}{(R_1+R_2+2m)(R_1+R_2-2m)R_1R_2}, \tag{3.2}$$

$$\Psi_0 = \Psi_4 = -\Psi_2 \frac{\sqrt{f^2+g^2}}{2R_1R_2(R_1+R_2-2\gamma m)}, \tag{3.3}$$

where

$$\begin{aligned} f^2 = \{ & [(R_1-R_2-2m)(R_1-R_2+2m)\gamma^2 \\ & -(R_1+R_2+2m)(R_1+R_2-2m)](R_1+R_2) \\ & -2(R_1+R_2-6\gamma m)R_1R_2\}^2, \end{aligned} \tag{3.4}$$

$$g^2 = (\gamma^2-1)^2(R_1-R_2)^2(R_1+R_2+2m)(R_1+R_2-2m)(R_1-R_2+2m)(R_1-R_2-2m) \tag{3.5}$$

and R_1 and R_2 are given by Eq. (2.7). The first-order Cartan scalars are too long and will not be shown.

Note that the Cartan scalars depend only on the coordinates ρ and z, while the line element depends on these coordinates and the differentials of the four coordinates. Also, they transform like scalars while the metric components transform like tensor components. Finally, the metric usually has features that are due to the nonessential coordinates (like the singularity on the Schwarzschild horizon) while, on the other hand, since only the essential coordinates appear on the Cartan scalars, they do not present such problems. So, in principle, a coordinate system can be found which provides a well-behaved limit for the Cartan scalars while the metric still diverges. Due to this fact it is easier to investigate limits using the Cartan scalars rather than using the line element.

The first step is to investigate the limits using Ψ_2; later we shall investigate whether the other Cartan scalars share the same limits. After a lengthy but straightforward calculation the leading term in the series expansion of Ψ_2 as $m\to\infty$ is:

$$2^{-(\gamma-1)} m^{2(\gamma^2-\gamma)} \rho^{-2(\gamma^2-\gamma+1)} \gamma(1-\gamma), \tag{3.6}$$

which is either divergent or finite depending on the value of γ. Nevertheless, this expression suggests that a finite limit may arise for all values of γ if we define a new radial coordinate $\overline{\rho}$ by $m^{2(\gamma^2-\gamma)}\rho^{-2(\gamma^2-\gamma+1)} = \overline{\rho}^{-2(\gamma^2-\gamma+1)}$ that is,

$$\rho = 2^{\beta} m^{\alpha} \overline{\rho}, \tag{3.7}$$

where

$$\alpha = \frac{\gamma^2-\gamma}{\gamma^2-\gamma+1} \tag{3.8}$$

and

$$\beta = \frac{-\gamma}{\gamma^2-\gamma+1}. \tag{3.9}$$

The constant β was introduced to provide the correct power of 2 in the limiting Cartan scalar.

Indeed, noting that $-\frac{1}{3} \leqslant \alpha < 1$ and using Eq. (3.7) in Eq. (3.2), a lengthy but straightforward calculation shows that in the $\{t, \overline{\rho}, z, \phi\}$ coordinate system one has:

$$\lim_{m\to\infty} \Psi_2 = \tfrac{1}{2}\bar{\rho}^{-2(\gamma^2-\gamma+1)}\gamma(1-\gamma), \tag{3.10}$$

which is finite. Similarly, one finds that all Cartan scalars have a finite limit in this new coordinate system. The question now is to find out to which metric this set of Cartan scalars belongs. This is not a straightforward task, but fortunately, calling

$$\gamma = 2\sigma \tag{3.11}$$

and $\bar{\rho} = r$ we are led to following set of Cartan scalars:

$$\psi_2 = (1-2\sigma)\sigma r^{-2(4\sigma^2-2\sigma+1)}, \tag{3.12}$$

$$\psi_0 = \psi_4 = (4\sigma-1)\psi_2, \tag{3.13}$$

$$\nabla\psi_{01'} = \nabla\psi_{50'} = \sqrt{2}(8\sigma^2-4\sigma+1)(4\sigma-1)(2\sigma-1)\sigma r^{-3(4\sigma^2-2\sigma+1)}, \tag{3.14}$$

$$\nabla\psi_{10'} = \nabla\psi_{41'} = \sqrt{2}(4\sigma-1)(2\sigma-1)\sigma r^{-3(4\sigma^2-2\sigma+1)}, \tag{3.15}$$

$$\nabla\psi_{21'} = \nabla\psi_{30'} = \sqrt{2}(4\sigma^2-2\sigma+1)(2\sigma-1)\sigma r^{-3(4\sigma^2-2\sigma+1)}, \tag{3.16}$$

which are the Cartan scalars of the Levi-Civita space–time[13]

$$ds^2 = r^{4\sigma}\,dt^2 - r^{8\sigma^2-4\sigma}(dr^2+dz^2) - \frac{r^{2-4\sigma}}{a}\,d\phi^2. \tag{3.17}$$

This shows that in the $\{t,\bar{\rho},z,\phi\}$ coordinate system, the limit of the γ metric as $m\to\infty$ is locally the Levi-Civita metric, provided the radial coordinates $\bar{\rho}$ and r are identified and the parameter γ divided by 2 of the γ metric is identified with the density parameter σ of the Levi-Civita metric, i.e., Eq. (3.11) holds.

We use the word *locally* since the Cartan scalars provide a local characterization of the metric. Furthermore, there is a parameter a in the Levi-Civita metric which does not appear in its Cartan scalars since it is a topological defect and can be eliminated by a coordinate transformation. For studying the global properties of the limit one has to investigate the metric (or the line element) directly. In fact one may ask whether, using the $\{t,\bar{\rho},z,\phi\}$ coordinate system, the Levi-Civita limit can be obtained directly from the line element of the γ metric.

In the $\{t,\bar{\rho},z,\phi\}$ coordinate system, the γ metric may be written as

$$ds^2 = e^{2\lambda}\,dt^2 - e^{-2\lambda}e^{2\mu}2^{2\beta}m^{2\alpha}\,d\bar{\rho}^2 - e^{-2\lambda}e^{2\mu}\,dz^2 - e^{-2\lambda}2^{2\beta}m^{2\alpha}\bar{\rho}^2\,d\phi^2, \tag{3.18}$$

where λ and μ are expressed with the $\bar{\rho}$ coordinate. The limit of the component $g_{\bar{\rho}\bar{\rho}}$ is precisely the g_{rr} of the Levi-Civita metric but the other metric components diverge. Now, the divergences can be easily removed by similar transformations on the coordinates t, z, and ϕ, given by

$$t = 2^{-\beta(\gamma^2+1)}m^{-\beta}\bar{t}, \tag{3.19}$$

$$z = 2^{\beta}m^{\alpha}\bar{z}, \tag{3.20}$$

$$\phi = 2^{\beta\gamma^2}m^{\beta\gamma}\frac{1}{\sqrt{a}}\bar{\phi}. \tag{3.21}$$

In the $\{\bar{t},\bar{\rho},\bar{z},\bar{\phi}\}$ coordinate system the γ metric becomes:

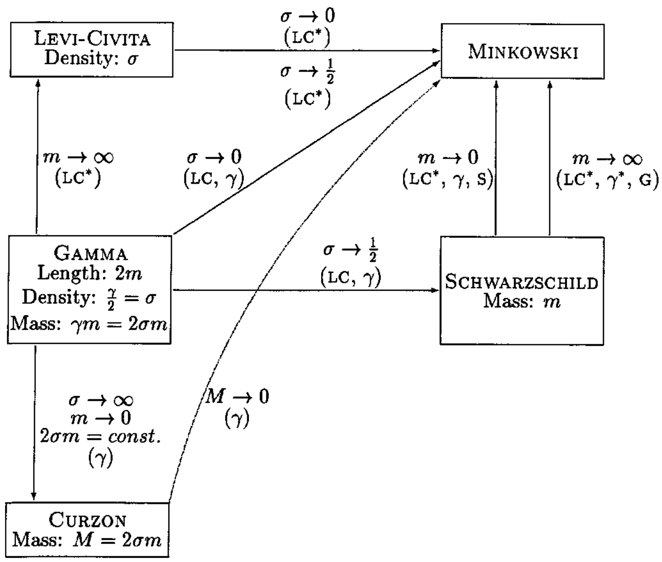

FIG. 1. Limiting diagram for the γ metric. In brackets are the coordinate systems where each limit works. γ means the original $\{t,\rho,z,\phi\}$ cylindrical coordinate system for the γ metric; LC means the Levi-Civita coordinate system $\{\bar{t},\bar{\rho},\bar{z},\bar{\phi}\}$, i.e., the γ coordinate system plus the coordinate transformation given by Eqs. (3.7) and (3.19)–(3.21); S means the usual Schwarszchild coordinates and G the Geroch coordinates first used to find the Minkowskian limit of Schwarzschild. The asterisk means that the limit is local, i.e., it was taken with the Cartan scalars and/or on the line element but a topological defect was found.

$$ds^2=e^{2\lambda}2^{-2(\gamma^2+1)\beta}m^{-2\beta}\,d\bar{t}^2-e^{-2\lambda}e^{2\mu}2^{2\beta}m^{2\alpha}\,d\bar{\rho}^2-e^{-2\lambda}e^{2\mu}2^{-2\beta}m^{-2\alpha}\,d\bar{z}^2$$
$$-e^{-2\lambda}2^{2(\beta+\gamma^2\beta)}m^{2(\alpha+\gamma\beta)}\bar{\rho}^2\frac{1}{a}\,d\bar{\phi}^2 \tag{3.22}$$

and its limit as $m\to\infty$ is precisely the Levi-Civita metric (3.17). The only drawback of this limit is the introduction of an infinite topological defect. In other words, the limit of the γ metric in the $\{\bar{t},\bar{\rho},\bar{z},\bar{\phi}\}$ coordinate system is the Levi-Civita metric only locally. This confirms the result we found previously with the Cartan scalars. Whether a coordinate system for the γ metric exists which provides a global limit into the Levi-Civita metric, i.e., with a finite topological defect, is still an open question.

IV. A LIMITING DIAGRAM FOR THE γ METRIC

We shall now study the limit we have just found, find other limits in the coordinate systems of Sec. III, and discuss some limits known in the literature in order to build the limiting diagram for the γ metric shown in Fig. 1.

A. Limits in the Schwarzschild coordinates

In the usual Schwarzschild coordinates, in the limit $m\to0$, the Schwarzschild line element tends to Minkowski. The limit $m\to\infty$ diverges. This can be easily checked by hand or from the Cartan scalars.[7]

B. The Geroch limits

In 1969 Geroch[14] showed that in the coordinate system (Geroch coordinates) defined by

$$x=r+m^{4/3},\quad \rho=m^{4/3}\theta,\quad t'=t,\quad \varphi'=\varphi, \tag{4.1}$$

the limit of the Schwarzschild metric as $m\rightarrow\infty$ is the Minkowski space–time. He also presented a coordinate system where the limit is a Kasner space–time. These results show that the limit of a space–time as some parameter goes to infinity is a coordinate-dependent process.

Later, Paiva, Rebouças, and MacCallum[7] reobtained these limits by using the Cartan scalar technique, and extended the results presenting new limits of the Schwarzschild metric and developing an approach to find all limits of a given space–time (see also Refs. 8, 15, and 16)

C. Limits in the γ coordinates

We shall call γ coordinates the original $\{t,\rho,z,\phi\}$ cylindrical coordinates used for the γ metric [Eqs. (2.1) and (2.5)–(2.7)]. Its is known[17] that in this coordinate system the limit $\gamma\rightarrow\infty$, $m\rightarrow 0$ with $\gamma m=\text{const}$ leads to the Curzon metric and, as shown in Sec. II, the limit $\gamma\rightarrow 1$ leads to Schwarzschild. Besides one can easily see that as $\gamma\rightarrow 0$ the γ metric tends to Minkowski. The coordinate systems in which the Curzon and Schwarzschild metrics are expressed when obtained as limit of the γ metric will also be called γ coordinates.

In the γ-coordinate system, the line elements of Curzon (see Ref. 17) and Schwarzschild tend to Minkowski as $m\rightarrow 0$ (this arises directly from the line element). Although the Schwarzschild line element in γ coordinates diverges as $m\rightarrow\infty$ it can be shown that its Cartan scalars (those of the γ metric with $\gamma=1$) tend to zero, which is locally Minkowski.

D. Limits in the LC coordinates

We have shown that in the coordinate system $\{\bar{t},\bar{\rho},\bar{z},\bar{\phi}\}$ the limit, as $m\rightarrow\infty$, of the γ metric is locally the Levi-Civita metric. This new coordinate system will therefore be called Levi-Civita coordinates or LC coordinates for short, for both the γ metric and the Levi-Civita metric (although the metric equivalence is only local). Coincidentally, this is the usual coordinate system for the Levi-Civita metric.

As $\gamma\rightarrow 0$ or $\gamma\rightarrow 1$, the γ metric in LC coordinates tends to Minkowski or Schwarzschild, as can be seen directly from the line element (3.22). The last one giving Schwarzschild in LC coordinates.

The limits of the Levi-Civita metric as $\sigma\rightarrow 0$ and $\sigma\rightarrow 1/2$ giving locally Minkowski can be directly found from the line element in LC coordinates.

As $m\rightarrow 0$, the Schwarzschild line element in LC coordinates diverge but its Cartan scalars tend to 0, i.e., locally Minkowski.

Finally, the LC-coordinate system turns out to be a new coordinate system (Geroch coordinates is the old one) where the Schwarzschild line element tends to Minkowski as $m\rightarrow\infty$. The equivalence is local since an infinite topological defect appears. This limit can also be done with the Cartan scalars.

One could point out that it is not proved that all limits found with the line element can be found with the Cartan scalars and vice-versa. Nevertheless, for our purposes, i.e., finding some specific limits, this proof is not necessary. In fact, both the line element and Cartan scalars are equivalent local representations of the space–time according to Cartan's equivalence theorem.[12] Therefore, limits found with one or the other are limits of the space–time. As a matter of fact, the Levi-Civita limit of the γ metric was found here using both approaches.

V. CONCLUSION

We have seen so far that by extending the length of the Newtonian image source of the γ metric to infinity we arrive at the Levi-Civita space–time. The amazing fact is that the finite rod does not represent the *real* source of the γ metric (it is just its Newtonian image source), whereas the infinite line singularity is thought to be the real source of the LC space–time. The link between the parameters γ and $\sigma(\gamma=2\sigma)$ appearing in the limiting process is quite consistent with previous results,[14,7] in the sense that the Schwarzschild metric ($\gamma=1$) leads (locally) to Minkowski space–

time as $m \to \infty$ and the Levi-Civita metric ($m \to \infty$) leads (locally) also to Minkowski if $\sigma = 1/2$. It should be interesting to find out if restrictions on σ, based on the existence of timelike circular geodesics[18] in LC ($\sigma < 1/4$), appear in the γ metric.

We shall now proceed to the interpretation of the limiting diagram of the γ metric (Fig. 1). In order to build this diagram, we introduced a new coordinate system for this metric (the LC coordinates) and found two new limits as $m \to \infty$: Schwarzschild $\to$ Minkowski in γ coordinates and LC coordinates and γ metric $\to$ Levi-Civita in LC coordinates.

One notices, that as it is presented, the diagram is quite consistent. It supports the current interpretations of σ as being the density in the Levi-Civita metric; $\gamma/2$ and $2m$, respectively, as the density and the length in the γ metric; m as the mass in the Schwarzschild solution, and M as the mass in the Curzon solution.

Note that from the γ metric one can reach Minkowski either through Levi-Civita by making $m \to \infty$ and then $\sigma \to 1/2$ or through Schwarzschild by making $\sigma \to 1/2$ and then $m \to \infty$. The limit $\sigma \to 0$ is similar; the difference being that, since the mass in the γ metric is $2\sigma m$, the limit $\sigma \to 0$ leads to Schwarzschild with zero mass, which is Minkowski.

To extend this work one should use the Cartan scalar techniques developed in Refs. 7 and 8 to find all limits of the γ metric. The main difficulty is computational, since in the present case too many Cartan scalars are different from zero and they depend on more than one coordinate and parameter. Afterwards one should do the same with the other metrics appearing on the diagram. Finally, one should find a single coordinate system which provides all the limits on the diagram and does not present an infinite topological defect. This would help the understanding of the topological defects in the Levi-Civita metric.

ACKNOWLEDGMENTS

F.M.P. gratefully acknowledges financial assistance from FAPERJ and the Laboratório de Astrofísica e Radioastronomia—Centro Regional Sul de Pesquisas Espaciais (INPE/MCT—Santa Maria RS) for the kind hospitality during the final stage of this work.

[1] H. Weyl, Ann. Phys. (Leipzig) **54**, 117 (1918); **59**, 185 (1919); T. Levi-Civita, Atti Accad. Naz. Lincei, Cl. Sci. Fis. Mat. Nat. Rend. **28**, 101 (1919); J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1960), p. 313.
[2] F. Esposito and L. Witten, Phys. Lett. **58B**, 357 (1975).
[3] K. L. Virbhadra, preprint grqc/9606004, 1996.
[4] R. Bach and H. Weyl, Math. Z. **13**, 134 (1920); G. Darmois, *Les Équations de la Gravitation Einsteinienne* (Gauthier-Villars, Paris, 1927), p. 36; G. Erez and N. Rosen, Bull. Res. Counc. Isr., Sect. F **8**, 47 (1959); D. M. Zipoy, J. Math. Phys. **7**, 1137 (1966); R. Gautreau and J. L. Anderson, Phys. Lett. **25A**, 291 (1967); F. I. Cooperstock and G. J. Junevicus, Int. J. Theor. Phys. **9**, 59 (1968); B. H. Voorhees, Phys. Rev. D **2**, 2119 (1970).
[5] W. B. Bonnor, in *Proceedings of the Third Canadian Conference on General Relativity and Relativistic Astrophysics*, edited by A. O. Coley, F. Cooperstock, and B. Tupper (World Scientific, Singapore, 1990), p. 216.
[6] W. B. Bonnor and W. Davidson, Class. Quantum Grav. **9**, 2065 (1992); W. B. Bonnor and M. A. P. Martins, *ibid.* **8**, 727 (1991).
[7] F. M. Paiva, M. J. Rebouças, and M. A. H. MacCallum, Class. Quantum Grav. **10**, 1165 (1993).
[8] F. M. Paiva and C. Romero, Gen. Relativ. Gravit. **25**, 1305 (1993).
[9] M. A. H. MacCallum and J. E. F. Skea, SHEEP: *A Computer Algebra System for General Relativity*, in Algebraic Computing in General Relativity: Lecture Notes from the First Brazilian School on Computer Algebra Vol. 2, edited by M. J. Rebouças and W. L. Roque (Oxford University Press, Oxford, 1992).
[10] F. M. Paiva, Publicação Especial do Observatório Nacional, No. 16/95, ON-Rio de Janeiro, 1995.
[11] A. Karlhede, Gen. Relativ. Gravit. **12**, 693 (1980).
[12] É. Cartan, *Leçons sur la Géométrie des Espaces de Riemann* (Gauthier-Villars, Paris, 1951).
[13] M. F. A. da Silva, L. Herrera, F. M. Paiva, and N. O. Santos, J. Math. Phys. **36**, 3625 (1995).
[14] R. Geroch, Commun. Math. Phys. **13**, 180 (1969).
[15] F. M. Paiva, M. J. Rebouças, G. S. Hall, and M. A. H. MacCallum, Class. Quantum Grav. **15**, 1031 (1998).
[16] F. M. Paiva, M. J. Rebouças, and A. F. F. Teixeira, J. Math. Phys. **38**, 4228 (1997).
[17] W. B. Bonnor, Gen. Relativ. Gravit. **24**, 551 (1990).
[18] R. Gautreau and R. Hoffman, Nuovo Cimento B **61**, 411 (1969).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Hyperbolic Kac–Moody algebra from intersecting *p*-branes

V. D. Ivashchuk,[a)]
Center for Gravitation and Fundamental Metrology, VNIIMS, 3/1 M. Ulyanovoy Str., Moscow 117313, Russia

S.-W. Kim[b)]
Department of Science Education and Basic Science Research Institute, Ewha Woman's University, Seoul 120-750, Korea

V. N. Melnikov[c)]
Center for Gravitation and Fundamental Metrology, VNIIMS, 3/1 M. Ulyanovoy Str., Moscow 117313, Russia

(Received 23 March 1998; accepted for publication 4 November 1998)

A subclass of a recently discovered class of solutions in multidimensional gravity with intersecting *p*-branes related to Lie algebras and governed by a set of harmonic functions is considered. This subclass in case of three Euclidean *p*-branes (one electric and two magnetic) contains a cosmological solution to $D=11$ supergravity related to hyperbolic Kac–Moody algebra $\mathcal{F}_3$ (of rank 3). This solution describes the non-Kasner power-law inflation. © *1999 American Institute of Physics.* [S0022-2488(99)02007-1]

I. INTRODUCTION

Here we consider a recently discovered class of solutions with intersecting *p*-branes.[1] These solutions are governed by a set of harmonic functions. The number of harmonic functions in general is less than the number of *p*-branes (as it takes place in orthogonal case[2–16]). The solutions correspond to a block-orthogonal set of *p*-brane vectors U^s [see (2.17) below] and may be considered as a Majumdar–Papapetrou type extension for the extremal limit of ''block-orthogonal'' black holes recently found in Ref. 17 (for ''orthogonal'' black holes see Refs. 18–23 and references therein.)

For the one-block case the solution is governed by one harmonic function and for a special configuration may be related to some simple finite dimensional Lie algebra or infinite dimensional hyperbolic Kac–Moody (KM) algebra.[24,25] The affine KM algebras do not appear among the solutions from Ref. 1.

Let us consider the simplest example of $D=11$ supergravity[26] (corresponding to M-theory[27]). It is known[3–5] that the orthogonal (or A_1+A_1) intersection rules for the M-theory read

$$3\cap 3=1, \quad 3\cap 6=2, \quad 6\cap 6=4 \tag{1.1}$$

(here we are counting dimensions of world sheets and their intersections). For the simplest $A_2=sl(3)$ Lie algebra the intersection rules are modified as follows:[1]

$$3\cap 3=0, \quad 3\cap 6=1, \quad 6\cap 6=3. \tag{1.2}$$

The rules (1.2) are obtained from (1.1) by a shift of one unit. (For $3\cap 3=0$ the ''truncated'' theory or without the Chern–Simons term should be considered). These modified rules may be written for a wide class of models and Lie algebras (finite or hyperbolic) and are defined by Dynkin diagrams.[1,23]

[a)]Electronic mail: ivas@rgs.phys.msu.su
[b)]Electronic mail: sungwon@mm.ewha.ac.kr
[c)]Electronic mail: rgs@com2com.ru

Hyperbolic algebras appeared in different areas of mathematical physics, e.g. in ordinary gravity[28] ($\mathcal{F}_3$ hyperbolic algebra), supergravity:[29,30] (E_{10} hyperbolic algebra),[31] ($\mathcal{F}_3$ hyperbolic algebra), strings, etc. (see also Ref. 32, and references therein). In Ref. 31 it was shown that the chiral reduction of a simple ($N=1$) supergravity from four dimensions to one dimension gives rise to the hyperbolic algebra of rank 3 (namely $\mathcal{F}_3$).

In Ref. 1 we considered some examples of hyperbolic intersection rules for the hyperbolic KM algebras of rank 2. These examples were suggested for so-called B_D models with $D\geqslant 14$,[23] containing $D-11$ scalar fields with negative kinetic terms. (B_{11} is the truncated bosonic sector of $D=11$ supergravity. B_{12} is the 12-dimensional model[33] corresponding to the low energy limit of F-theory.[34])

Here an example of cosmological solution in $D=11$ supergravity with three p-branes (two magnetic and one electric) that have intersection rules corresponding to the hyperbolic KM algebra $\mathcal{F}_3$ is constructed.

II. THE MODEL

We consider a model governed by the action[11]

$$S=\int d^Dz\sqrt{|g|}\left\{R[g]-h_{\alpha\beta}g^{MN}\partial_M\varphi^\alpha\partial_N\varphi^\beta-\sum_{a\in\Delta}\frac{\theta_a}{n_a!}\exp[2\lambda_a(\varphi)](F^a)^2\right\}, \tag{2.1}$$

where $g=g_{MN}\,dz^M\otimes dz^N$ is a metric, $\varphi=(\varphi^\alpha)\in\mathbf{R}^l$ is a vector of scalar fields, $(h_{\alpha\beta})$ is a symmetric nondegenerate $l\times l$ matrix $(l\in\mathbf{N})$, $\theta_a=\pm1$, $F^a=dA^a$ is a n_a-form $(n_a\geqslant1)$, λ_a is a one-form on $\mathbf{R}^l$: $\lambda_a(\varphi)=\lambda_{\alpha a}\varphi^\alpha$, $a\in\Delta$, $\alpha=1,\ldots,l$. Here Δ is some finite set.

We consider a manifold

$$M=M_0\times M_1\times\cdots\times M_n, \tag{2.2}$$

with a metric

$$g=e^{2\gamma(x)}g^0+\sum_{i=1}^n e^{2\phi^i(x)}g^i, \tag{2.3}$$

where $g^0=g^0_{\mu\nu}(x)dx^\mu\otimes dx^\nu$ is a metric on the manifold M_0, and $g^i=g^i_{m_in_i}(y_i)dy_i^{m_i}\otimes dy_i^{n_i}$ is a Ricci-flat metric on the manifold M_i ($\mathrm{Ric}[g^i]=0$), $i=1,\ldots,n$. Any manifold M_ν is oriented and connected and $d_\nu\equiv\dim M_\nu$, $\nu=0,\ldots,n$. Let

$$\tau_i\equiv\sqrt{|g^i(y_i)|}\,dy_i^1\wedge\ldots\wedge dy_i^{d_i},\quad \varepsilon(i)\equiv\mathrm{sign}(\det(g^i_{m_in_i}))=\pm1 \tag{2.4}$$

denote the volume d_i-form and signature parameter, respectively, $i=1,\ldots,n$. Let $\Omega=\Omega_n$ be a set of all subsets of $\{1,\ldots,n\}$, $|\Omega|=2^n$. For any $I=\{i_1,\ldots,i_k\}\in\Omega$, $i_1<\ldots<i_k$, we denote

$$\tau(I)\equiv\tau_{i_1}\wedge\ldots\wedge\tau_{i_k},\quad d(I)\equiv\sum_{i\in I}d_i,\quad \varepsilon(I)\equiv\prod_{i\in I}\varepsilon(i). \tag{2.5}$$

We also put $\tau(\emptyset)=\varepsilon(\emptyset)=1$ and $d(\emptyset)=0$.

For fields of forms we consider the following composite electromagnetic ansatz:

$$F^a=\sum_{I\in\Omega_{a,e}}\mathcal{F}^{(a,e,I)}+\sum_{J\in\Omega_{a,m}}\mathcal{F}^{(a,m,J)}, \tag{2.6}$$

where

$$\mathcal{F}^{(a,e,I)}=d\Phi^{(a,e,I)}\wedge\tau(I), \tag{2.7}$$

$$\mathcal{F}^{(a,m,J)} = e^{-2\lambda_a(\varphi)} * (d\Phi^{(a,m,J)} \wedge \tau(J)) \tag{2.8}$$

are elementary forms of electric and magnetic types, respectively, $a \in \Delta$, $I \in \Omega_{a,e}$, $J \in \Omega_{a,m}$, and $\Omega_{a,e} \subset \Omega$, $\Omega_{a,m} \subset \Omega$. In (2.8) $* = *[g]$ is the Hodge operator on (M,g). For scalar functions we put

$$\varphi^\alpha = \varphi^\alpha(x), \quad \Phi^s = \Phi^s(x), \tag{2.9}$$

$s \in S$.

Here and below

$$S = S_e \sqcup S_m, \quad S_v = \cup_{a \in \Delta} \{a\} \times \{v\} \times \Omega_{a,v}, \tag{2.10}$$

$v = e, m$.

Due to (2.7) and (2.8)

$$d(I) = n_a - 1, \quad d(J) = D - n_a - 1, \tag{2.11}$$

for $I \in \Omega_{a,e}$, $J \in \Omega_{a,m}$.

A. The sigma model

Let $d_0 \neq 2$ and

$$\gamma = \gamma_0(\phi) \equiv \frac{1}{2 - d_0} \sum_{j=1}^{n} d_j \phi^j, \tag{2.12}$$

i.e., the generalized harmonic gauge is used.

We impose the restriction on sets $\Omega_{a,v}$. These restrictions guarantee the block-diagonal structure of a stress-energy tensor (like for the metric) and the existence of σ-model representation.[11]

We denote $w_1 \equiv \{i | i \in \{1, \ldots, n\}, \quad d_i = 1\}$, and $n_1 = |w_1|$ (i.e., n_1 is the number of one-dimensional spaces among M_i, $i = 1, \ldots, n$).

Restriction 1: Let 1a) $n_1 \leqslant 1$ or (1b) $n_1 \geqslant 2$ and for any $a \in \Delta$, $v \in \{e, m\}$, $i, j \in w_1$, $i < j$, there are no $I, J \in \Omega_{a,v}$ such that $i \in I$, $j \in J$ and $I \setminus \{i\} = J \setminus \{j\}$.

Restriction 2 (only for $d_0 = 1,3$): Let (2a) $n_1 = 0$ or (2b) $n_1 \geqslant 1$ and for any $a \in \Delta$, $i \in w_1$ there are no $I \in \Omega_{a,m}$, $J \in \Omega_{a,e}$ such that $\bar{I} = \{i\} \sqcup J$ for $d_0 = 1$ and $J = \{i\} \sqcup \bar{I}$ for $d_0 = 3$. Here and in what follows

$$\bar{I} \equiv \{1, \ldots, n\} \setminus I. \tag{2.13}$$

It was proved in Ref. 11 that equations of motion for the model (2.1) and the Bianchi identities: $d\mathcal{F}^s = 0$, $s \in S_m$, for fields from (2.3) to (2.12), when Restrictions 1 and 2 are imposed, are equivalent to equations of motion for the σ-model governed by the action

$$S_\sigma = \int d^{d_0}x \sqrt{|g^0|} \left\{ R[g^0] - \hat{G}_{AB} g^{0\mu\nu} \partial_\mu \sigma^A \partial_\nu \sigma^B - \sum_{s \in S} \varepsilon_s e^{-2U_A^s \sigma^A} g^{0\mu\nu} \partial_\mu \Phi^s \partial_\nu \Phi^s \right\}, \tag{2.14}$$

where $(\sigma^A) = (\phi^i, \varphi^\alpha)$, the index set S from (2.10), target space metric

$$(\hat{G}_{AB}) = \begin{pmatrix} G_{ij} & 0 \\ 0 & h_{\alpha\beta} \end{pmatrix}, \tag{2.15}$$

with

$$G_{ij}=d_i\delta_{ij}+\frac{d_id_j}{d_0-2}, \tag{2.16}$$

vectors

$$(U^s_A)=(d_i\delta_{iI_s},-\chi_s\lambda_{\alpha a_s}), \tag{2.17}$$

where $s=(a_s,v_s,I_s)$, $\chi_e=+1$, $\chi_m=-1$;

$$\delta_{iI}=\sum_{j\in I}\delta_{ij} \tag{2.18}$$

is the indicator of i belonging to I: $\delta_{iI}=1$ for $i\in I$ and $\delta_{iI}=0$ otherwise; and

$$\varepsilon_s=(-\varepsilon[g])^{(1-\chi_s)/2}\varepsilon(I_s)\theta_{a_s}, \tag{2.19}$$

$s\in S$, $\varepsilon[g]\equiv\text{sign}\det(g_{MN})$. More explicitly (2.19) reads $\varepsilon_s=\varepsilon(I_s)\theta_{a_s}$ for $v_s=e$ and $\varepsilon_s=-\varepsilon[g]\varepsilon(I_s)\theta_{a_s}$ for $v_s=m$.

B. Exact solutions in a block-orthogonal case

Let us define the scalar product as follows:

$$(U,U')=\hat{G}^{AB}U_AU'_B, \tag{2.20}$$

for $U,U'\in\mathbf{R}^N$, where $(\hat{G}^{AB})=(\hat{G}_{AB})^{-1}$. The scalar products (2.20) for vectors U^s were calculated in 11

$$(U^s,U^{s'})=d(I_s\cap I_{s'})+\frac{d(I_s)d(I_{s'})}{2-D}+\chi_s\chi_{s'}\lambda_{\alpha a_s}\lambda_{\beta a_{s'}}h^{\alpha\beta}\equiv B^{ss'}, \tag{2.21}$$

where $(h^{\alpha\beta})=(h_{\alpha\beta})^{-1}$; $s=(a_s,v_s,I_s)$ and $s'=(a_{s'},v_{s'},I_{s'})$ belongs to S.

Let

$$S=S_1\sqcup\ldots\sqcup S_k, \tag{2.22}$$

$S_i\neq\emptyset$, $i=1,\ldots,k$, and

$$(U^s,U^{s'})=0 \tag{2.23}$$

for all $s\in S_i$, $s'\in S_j$, $i\neq j$; $i,j=1,\ldots,k$. Relation (2.22) means that the set S is a union of k nonintersecting (nonempty) subsets $S_1,\ldots,S_k$. According to (2.23) the set of vectors $(U^s,s\in S)$ has a block-orthogonal structure with respect to the scalar product (2.20), i.e., it splits into k mutually orthogonal blocks $(U^s,s\in S_i)$, $i=1,\ldots,k$.

Here we consider exact solutions in the model (2.1), when vectors $(U^s,s\in S)$ obey the block-orthogonal decomposition (2.22), (2.23) with scalar products defined in (2.21).[1] These solutions may be obtained from the corresponding solutions of the σ-model,[1] that are presented in the Appendix.

The solution reads:

$$g=U\left\{g^0+\sum_{i=1}^{n}U_ig^i\right\}, \tag{2.24}$$

$$U=\left(\prod_{s\in S} H_s^{2d(I_s)\varepsilon_s\nu_s^2}\right)^{1/(2-D)}, \tag{2.25}$$

$$U_i=\prod_{s\in S} H_s^{2\varepsilon_s\nu_s^2\delta_{iI_s}}, \tag{2.26}$$

$$\varphi^\alpha=-\sum_{s\in S}\lambda_{a_s}^{\alpha}\chi_s\varepsilon_s\nu_s^2\ln H_s, \tag{2.27}$$

$$F^a=\sum_{s\in S}\mathcal{F}^s\delta_{a_s}^a, \tag{2.28}$$

where $\mathrm{Ric}[g^0]=\mathrm{Ric}[g^i]=0$,

$$\mathcal{F}^s=\nu_s dH_s^{-1}\wedge\tau(I_s), \quad \text{for } v_s=e, \tag{2.29}$$

$$\mathcal{F}^s=\nu_s(*_0 dH_s)\wedge\tau(\bar{I}_s), \quad \text{for } v_s=m, \tag{2.30}$$

H_s are harmonic functions on (M_0,g^0) coinciding inside blocks of matrix $(B^{ss'})$ from (2.21) $(H_s=H_{s'},\ s,s'\in S_j,\ j=1,\ldots,k)$ and relations

$$\sum_{s'\in S} B^{ss'}\varepsilon_{s'}\nu_{s'}^2=-1 \tag{2.31}$$

for the matrix $(B^{ss'})$ (2.21), parameters ε_s (2.19) and ν_s are imposed, $s\in S$, $i=1,\ldots,n$; $\alpha=1,\ldots,l$. Here $\lambda_a^\alpha=h^{\alpha\beta}\lambda_{\beta a}$, $*_0=*[g^0]$ is the Hodge operator on (M_0,g^0) and $\bar{I}$ is defined in (2.13).

In deriving the solutions the following relations for contravariant components of U^s-vectors were used:[11]

$$U^{si}=\delta_{iI_s}-\frac{d(I_s)}{D-2}, \quad U^{s\alpha}=-\chi_s\lambda_{a_s}^{\alpha}, \tag{2.32}$$

$s=(a_s,v_s,I_s)$.

Thus, we obtained the generalization of the solutions from Ref. 11 to the block-orthogonal case [here we eliminate the misprint with sign in Eq. (5.19) in Ref. 11].

Remark 1: The solution is also valid for $d_0=2$, if Restriction 2 is replaced by Restriction 2*.

Restriction 2 (for $d_0=2$):* For any $a\in\Delta$ there are no $I\in\Omega_{a,m}$, $J\in\Omega_{a,e}$ such that $\bar{I}=J$ and for $n_1\geqslant 2$, $i,j\in w_1$, $i\neq j$, there are no $I\in\Omega_{a,m}$, $J\in\Omega_{a,e}$ such that $i\in I$, $j\in\bar{J}$, $I\setminus\{i\}=\bar{J}\setminus\{j\}$.

It may be proved using a more general form of the sigma-model representation (see Remark 2 in Ref. 11).

III. SOLUTIONS RELATED TO LIE ALGEBRAS AND INTERSECTION RULES

Here we put

$$(U^s,U^s)\neq 0 \tag{3.1}$$

for all $s\in S$ and introduce the quasi-Cartan matrix $A=(A^{ss'})$

$$A^{ss'}\equiv\frac{2(U^s,U^{s'})}{(U^{s'},U^{s'})}, \tag{3.2}$$

$s,s' \in S$. From (2.23) we get a block-orthogonal structure of A:

$$A = \begin{pmatrix} A_{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_{(k)} \end{pmatrix}, \tag{3.3}$$

where $A_{(i)} = (A^{ss'}, s,s' \in S_i)$, $i=1,\dots,k$. Here we tacitly assume that the set S is ordered, $S_1 < \dots < S_k$ and the order in S_i is inherited by the order in S.

We note that due to (2.23) the relation (A5) may be rewritten as

$$\sum_{s' \in S_i} (U^s, U^{s'}) \varepsilon_{s'} \nu_{s'}^2 = -1, \tag{3.4}$$

$s \in S_i$, $i=1,\dots,k$. Hence, parameters $(\nu_s, s \in S_i)$ depend upon vectors $(U^s, s \in S_i)$, $i=1,\dots,k$.

For $\det A_{(i)} \neq 0$ relation (3.4) may be rewritten in the equivalent form

$$\varepsilon_s \nu_s^2 (U^s, U^s) = -2 \sum_{s' \in S} A_{ss'}^{(i)}, \tag{3.5}$$

$s \in S_i$, where $(A_{ss'}^{(i)}) = A_{(i)}^{-1}$. Thus, Eq. (3.4) may be resolved in terms of ν_s for certain $\varepsilon_s = \pm 1$, $s \in S_i$.

In what follows we consider the block-orthogonal decomposition to be irreducible, i.e., for any i the block $(U^s, s \in S_i)$ does not split into two mutually orthogonal subblocks. In this case any matrix $A_{(i)}$ is indecomposable (or irreducible) in the sense that there is no renumbering of vectors which would bring $A_{(i)}$ to the block diagonal form $A_i = \mathrm{diag}(A'_{(i)}, A''_{(i)})$.

Let A be the generalized Cartan matrix.[24,25] In this case

$$A^{ss'} \in -\mathbf{Z}_+ \equiv \{0,-1,-2, \dots\} \tag{3.6}$$

for $s \neq s'$ and A generates generalized symmetrizable Kac–Moody algebra.[24,25]

Now we fix $i \in \{1,\dots,k\}$. From (3.3) and (3.6) we get

$$A_{(i)}^{ss'} \in -\mathbf{Z}_+, \tag{3.7}$$

$s,s' \in S_i$, $s \neq s'$. There are three possibilities for $A_{(i)}$: (a) $\det A_{(i)} > 0$, (b) $\det A_{(i)} < 0$ and (c) $\det A_{(i)} = 0$. For $\det A_{(i)} \neq 0$ the corresponding Kac–Moody algebra is simple, since $A_{(i)}$ is indecomposable.[25]

A. Finite dimensional Lie algebras

Let $\det A_{(i)} > 0$. In this case $A_{(i)}$ is the Cartan matrix of a simple finite-dimensional Lie algebra and $A_{(i)}^{ss'} \in \{0,-1,-2,-3\}$, $s \neq s'$. The elements of inverse matrix $A_{(i)}^{-1}$ are positive (see Chap. 7 in Ref. 25) and hence we get from (3.5)

$$\varepsilon_s (U^s, U^s) < 0, \tag{3.8}$$

$s \in S_i$.

B. Hyperbolic Kac–Moody algebras

Let $\det A_{(i)} < 0$. Among irreducible symmetrizable matrices satisfying (3.7) there exists a large subclass of Cartan matrices, corresponding to infinite-dimensional simple hyperbolic generalized Kac–Moody (KM) algebras of ranks $r=2,\dots,10$.[24,25]

For the hyperbolic algebras the following relations are satisfied:

$$\varepsilon_s(U^s,U^s)>0, \tag{3.9}$$

$s \in S_i$. This relation is valid, since $(A^{-1}_{(i)})_{ss'} \leqslant 0$, $s,s' \in S$, for any hyperbolic algebra.[35]

It was shown in Ref. 1 that affine KM algebras with $\det A_{(i)}=0$ do not appear in the solutions.[1]

C. Intersection rules

From the orthogonality relation (2.23) and (2.21) we get

$$d(I_s \cap I_{s'}) = \Delta(s,s') \tag{3.10}$$

where $s \in S_i$, $s' \in S_j$, $i \neq j$, and

$$\Delta(s,s') \equiv \frac{d(I_s)d(I_{s'})}{D-2} - \chi_s \chi_{s'} \lambda_{a_s} \cdot \lambda_{a_{s'}}. \tag{3.11}$$

Here $\lambda \cdot \lambda' \equiv h^{\alpha\beta}\lambda_\alpha \lambda'_\beta$. Let

$$N(a,b) \equiv \frac{(n_a-1)(n_b-1)}{D-2} - \lambda_a \cdot \lambda_b, \tag{3.12}$$

$a,b \in \Delta$. The matrix (3.12) is called the fundamental matrix of the model (2.1).[23] For $s_1,s_2 \in S$, $s_1 \neq s_2$, the symbol of orthogonal intersection (3.11) may be expressed by means of the fundamental matrix[23]

$$\Delta(s_1,s_2) = \bar{D}\bar{\chi}_{s_1}\bar{\chi}_{s_2} + \bar{n}_{a_{s_1}}\chi_{s_1}\bar{\chi}_{s_2} + \bar{n}_{a_{s_2}}\chi_{s_2}\bar{\chi}_{s_1} + N(a_{s_1},a_{s_2})\chi_{s_1}\chi_{s_2}, \tag{3.13}$$

where $\bar{D}=D-2$, $\bar{n}_a = n_a - 1$, $\bar{\chi}_s = \frac{1}{2}(1-\chi_s)$. More explicitly (3.13) reads

$$\Delta(s_1,s_2) = N(a_{s_1},a_{s_2}), \quad v_{s_1}=v_{s_2}=e, \tag{3.14}$$

$$\Delta(s_1,s_2) = \bar{n}_{a_{s_1}} - N(a_{s_1},a_{s_2}), \quad v_{s_1}=e, \quad v_{s_2}=m, \tag{3.15}$$

$$\Delta(s_1,s_2) = \bar{D} - \bar{n}_{a_{s_1}} - \bar{n}_{a_{s_2}} + N(a_{s_1},a_{s_2}), \quad v_{s_1}=v_{s_2}=m. \tag{3.16}$$

This follows from the relations

$$d(I_s) = \bar{D}\bar{\chi}_s + \bar{n}_{a_s}\chi_s, \tag{3.17}$$

equivalent to (2.11). Let

$$K(a) \equiv n_a - 1 - N(a,a) = \frac{(n_a-1)(D-n_a-1)}{D-2} + \lambda_a \cdot \lambda_a, \tag{3.18}$$

$a \in \Delta$.

The parameters (3.18) play a rather important role in supergravitational theories, since they are preserved under Kaluza–Klein reduction[2] and define the norms of U^s vectors:

$$(U^s,U^s) = K(a_s), \tag{3.19}$$

$s \in S$.

Here we put $K(a) \neq 0$, $a \in \Delta$. Then, we obtain the general intersection rule formulas

$$d(I_{s_1}\cap I_{s_2})=\Delta(s_1,s_2)+\tfrac{1}{2}K(a_{s_2})A^{s_1s_2} \tag{3.20}$$

$s_1\neq s_2$, where $(A^{s_1s_2})$ is the quasi-Cartan matrix (3.2) (see also (6.32) from Ref. 23).

In most models including $D=11$ supergravity, $D=12$ theory,[33] $D<11$ supergravities,[2] $K(a)=2$ and (3.20) has the following form

$$d(I_{s_1}\cap I_{s_2})=\Delta(s_1,s_2)+A^{s_1s_2}, \tag{3.21}$$

$s_1\neq s_2$, and get $A^{s_1s_2}=A^{s_2s_1}$, i.e., the Cartan matrix is symmetric.

IV. EXAMPLES

A. Hyperbolic algebra of rank three

Now we consider the example of the solution corresponding to the hyperbolic KM algebra $\mathcal{F}_3$ with the Cartan matrix

$$A=\begin{pmatrix} 2 & -2 & 0 \\ -2 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}. \tag{4.1}$$

The hyperbolic algebra $\mathcal{F}_3$ corresponding to (4.1), is an infinite dimensional Lie algebra generated by the (Serre) relations[24,25]

$$[h_i,h_j]=0, \quad [e_i,f_j]=\delta_{ij}h_j, \tag{4.2}$$

$$[h_i,e_j]=A_{ij}e_j, \quad [h_i,f_j]=-A_{ij}f_j, \tag{4.3}$$

$$(\mathrm{ad}e_i)^{1-A_{ij}}(e_j)=0 \quad (i\neq j), \tag{4.4}$$

$$(\mathrm{ad}f_i)^{1-A_{ij}}(f_j)=0 \quad (i\neq j). \tag{4.5}$$

$\mathcal{F}_3$ contains $A_1^{(1)}$ affine Kac–Moody subalgebra (it corresponds to the Geroch group) and A_2 subalgebra.

The calculation of inverse matrix gives us

$$A^{-1}=-\begin{pmatrix} \frac{3}{2} & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \tag{4.6}$$

and, hence,

$$\sum_{j=1}^{3} A_{ij}^{-1}=-\frac{9}{2},-5,-2, \tag{4.7}$$

for $i=1,2,3$, respectively.

Now we construct an example of the solution with the A-matrix (4.1) for $D=11$ supergravity governed by the action (in the bosonic sector)

$$S=\int d^{11}z\sqrt{|g|}\left\{R[g]-\frac{1}{4!}(F^4)^2\right\}+c\int A^3\wedge F^4\wedge F^4, \tag{4.8}$$

where $c=\mathrm{const}$, $F^4=dA^3$.

We consider a configuration with three p-branes, one electric and two magnetic. We denote $S=\{s_1,s_2,s_3\}$, $\nu_{s_1}=\nu_{s_3}=m$, $\nu_{s_2}=e$ and get $d(I_{s_1})=d(I_{s_3})=6$, $d(I_{s_2})=3$. From intersection rules (3.21) we obtain

$$d(I_{s_1}\cap I_{s_2})=0, \quad d(I_{s_2}\cap I_{s_3})=1, \quad d(I_{s_1}\cap I_{s_3})=4. \tag{4.9}$$

For the manifold (2.2) we put $n=5$ and $d_1=2$, $d_2=4$, $d_3=d_4=1$, $d_5=2$. The corresponding sets for p-branes are the following: $I_{s_1}=\{1,2\}$, $I_{s_2}=\{4,5\}$, $I_{s_3}=\{2,3,4\}$.

The corresponding solution reads

$$g=H^{-12}\{-dt\otimes dt+H^9g^1+H^{13}g^2+H^4g^3+H^{14}g^4+H^{10}g^5\}, \tag{4.10}$$

$$F^4=\frac{dH}{dt}\left\{\nu_{s_1}\tau_3\wedge\tau_4\wedge\tau_5+\frac{\nu_{s_2}}{H^2}dt\wedge\tau_4\wedge\tau_5+\nu_{s_3}\tau_1\wedge\tau_5\right\}, \tag{4.11}$$

where

$$\nu_{s_1}^2=\tfrac{9}{2}, \quad \nu_{s_2}^2=5, \quad \nu_{s_3}^2=2 \tag{4.12}$$

[see relations (3.5) and (4.7)], all metric g^i are Ricci-flat ($i=1,\ldots,5$) with the Euclidean signature [this agrees with relations (3.9) and (2.19)], and

$$H=ht+h_0>0, \tag{4.13}$$

h,h_0 are constants. [We remind that here $(U^s,U^s)=2$.]

The solution (4.10)–(4.13) satisfies not only equations of motion for the truncated model (without the Chern–Simons term), but also the equations of motion for the ''total'' model (4.8), since the only modification related to ''Maxwell's'' equations

$$d*F^4=\text{const}\,F^4\wedge F^4, \tag{4.14}$$

is trivial due to $F^4\wedge F^4=0$ (since $\tau_i\wedge\tau_i=0$).

The metric (4.10) may be also rewritten using the synchronous time variable t_s,

$$g=-dt_s\otimes dt_s+f^{3/5}g^1+f^{-1/5}g^2+f^{8/5}g^3+f^{-2/5}g^4+f^{2/5}g^5, \tag{4.15}$$

where $f=5ht_s=H^{-5}>0$, $h>0$, $t_s>0$. The metric describes the power-law ''inflation'' in $D=11$. It is singular for $t_s\to+0$. It is interesting to note that the powers in scale-factors $f^{2\alpha_i}$ do not satisfy Kasner-like relations:[36] $\Sigma_{i=1}^5 d_i\alpha_i=\Sigma_{i=1}^5 d_i(\alpha_i)^2=1$. For flat g^i the calculation of the Riemann tensor squared gives us (see Refs. 37, 38)

$$R_{MNPQ}[g]R^{MNPQ}[g]=At_s^{-4}, \tag{4.16}$$

where $A=2\times1.0714$.

B. A_3 Lie algebra

Here we present for comparison the solution of $D=11$ supergravity corresponding to A_3 Lie algebra with the Cartan matrix

$$A=\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}. \tag{4.17}$$

The calculation of inverse matrix gives in this case

$$\sum_{j=1}^{3} A_{ij}^{-1} = \frac{3}{2}, 2, \frac{3}{2}, \tag{4.18}$$

for $i=1,2,3$, respectively.

Like in the example mentioned above we consider three p-branes, one electric and two magnetic, i.e., in this case $S=\{s_1, s_2, s_3\}$, $v_{s_1}=v_{s_3}=m$, $v_{s_2}=e$, $d(I_{s_1})=d(I_{s_3})=6$, $d(I_{s_2})=3$. From intersection rules (3.21) we obtain

$$d(I_{s_1}\cap I_{s_2})=1, \quad d(I_{s_2}\cap I_{s_3})=1, \quad d(I_{s_1}\cap I_{s_3})=4. \tag{4.19}$$

For the manifold (2.2) we put $n=5$ and $d_1=2$, $d_2=3$, $d_3=1$, $d_4=2$, $d_5=2$. The corresponding sets for p-branes are the following: $I_{s_1}=\{1,2,3\}$, $I_{s_2}=\{3,5\}$, $I_{s_3}=\{2,3,4\}$.

The corresponding solution reads

$$g = H^{16/3}\{d\rho\otimes d\rho + H^{-3}g^1 + H^{-6}g^2 + H^{-10}(-dt\otimes dt) + H^{-3}g^4 + H^{-4}g^5\} \tag{4.20}$$

$$F^4 = \frac{dH}{d\rho}\left\{ \nu_{s_1}\tau_4\wedge\tau_5 + \frac{\nu_{s_2}}{H^2} d\rho\wedge dt\wedge\tau_5 + \nu_{s_3}\tau_1\wedge\tau_5 \right\}, \tag{4.21}$$

where

$$\nu_{s_1}^2 = \tfrac{3}{2}, \quad \nu_{s_2}^2 = 2, \quad \nu_{s_3}^2 = \tfrac{3}{2}. \tag{4.22}$$

Here the metrics g^i are Ricci-flat $(i=1,2,4,5)$ with the Euclidean signature, and

$$H = c\rho + c_0 > 0, \tag{4.23}$$

c, c_0 are constants. So, we obtained the multidimensional ''cosmological'' solution with the Euclidean ''time'' ρ.

V. DISCUSSION

Here we obtained the example of the cosmological solution with three Euclidean intersecting p-branes (one electric and two magnetic) satisfying intersection rules for the hyperbolic Kac–Moody Lie algebra $\mathcal{F}_3$ [see (3.21) and (4.1)]. The corresponding A_3 solution contains three pseudo-Euclidean p-branes. The difference in sign rules (restriction on ε_s) for finite and hyperbolic algebras is a consequence of inequalities for elements of the inverse Cartan matrix: $A_{ij}^{-1} > 0$ for simple (or semisimple) finite dimensional Lie algebra and $A_{ij}^{-1} \leqslant 0$ (for simple hyperbolic KM algebra). In this paper the hyperbolic KM algebra $\mathcal{F}_3$ appeared only on the simplest level of the Cartan matrix (governing the intersection rules) but the full structure of the algebra, including Serre relations (4.4) and (4.5), was not used. We may suppose that at the second step a more deep penetrating into a ''structure'' of infinite dimensional hyperbolic algebras will be achieved when general cosmological solutions related to hyperbolic Toda-lattices will be considered.

ACKNOWLEDGMENTS

This work was supported in part by the Russian Ministry for Science and Technology, Russian Foundation for Basic Research, Grant No. 98-02-16414, Korea Research Foundation made in the program year of 1997 and KOSEF No. 95-0702-04-01-3. The authors are grateful to M.A. Grebeniuk for useful discussions and V.V. Nikulin for useful information. V.D.I and V.N.M also thank the members of the Department of Science Education and Basic Science Research Institute of Ewha Womans University (Seoul, Korea) for kind hospitality during their visits in February of

1998. The authors are grateful to the organizers and participants of the Second Winter School on Branes, Fields and Mathematical Physics (APCTP, Seoul), where the results of this paper were reported.

APPENDIX: BLOCK-ORTHOGONAL SOLUTIONS IN THE σ-MODEL

Equations of motion corresponding to (2.14) have the following form:

$$R_{\mu\nu}[g^0]=\hat{G}_{AB}\partial_\mu\sigma^A\partial_\nu\sigma^B+\sum_{s\in S}\varepsilon_s e^{-2U_A^s\sigma^A}\partial_\mu\Phi^s\partial_\nu\Phi^s, \tag{A1}$$

$$\hat{G}_{AB}\Delta[g^0]\sigma^B+\sum_{s\in S}\varepsilon_s U_A^s e^{-2U_C^s\sigma^C}g^{0\mu\nu}\partial_\mu\Phi^s\partial_\nu\Phi^s=0, \tag{A2}$$

$$\partial_\mu(\sqrt{|g^0|}g^{0\mu\nu}e^{-2U_A^s\sigma^A}\partial_\nu\Phi^s)=0, \tag{A3}$$

$s\in S$. Here $\Delta[g^0]$ is the Laplace-Beltrami operator corresponding to g^0.

Proposition 1: Let (M_0,g^0) be Ricci-flat $R_{\mu\nu}[g^0]=0$. Then the field configuration

$$g^0, \quad \sigma^A=\sum_{s\in S}\varepsilon_s U^{sA}\nu_s^2\ln H_s, \quad \Phi^s=\frac{\nu_s}{H_s}, \tag{A4}$$

$s\in S$, satisfies the field equations (A1)–(A3) with $V=0$ if (real) numbers ν_s obey the relations

$$\sum_{s'\in S}(U^s,U^{s'})\varepsilon_{s'}\nu_{s'}^2=-1, \tag{A5}$$

$s\in S$, functions $H_s>0$ are harmonic, i.e.,

$$\Delta[g^0]H_s=0, \tag{A6}$$

$s\in S$ and H_s are coinciding inside blocks:

$$H_s=H_{s'} \tag{A7}$$

for $s,s'\in S_i$, $i=1,\ldots,k$.

The Proposition can be readily verified by a straightforward substitution of (A4)–(A7) into equations of motion (A1)–(A3). In the special (orthogonal) case, when any block contains only one vector (i.e., all $|S_i|=1$) the Proposition coincides with Proposition 1 of Ref. 11. In the general case vectors inside each block S_i are not orthogonal. The solution under consideration depends on k independent harmonic functions. For a given set of vectors $(U^s,s\in S)$ the maximal number k arises for the irreducible block-orthogonal decomposition (2.22), (2.23), when any block $(U^s,s\in S_i)$ does not split into two mutually-orthogonal subblocks.

[1] V. D. Ivashchuk and V. N. Melnikov, hep-th/9802121; Class. Quantum Grav. **16**, 849 (1999).
[2] K. S. Stelle, hep-th/9701088.
[3] A. Strominger, Phys. Lett. B **383**, 44 (1996); hep-th/9512059.
[4] P. K. Townsend, Phys. Lett. B **373**, 68 (1996); hep-th/9512062.
[5] A. A. Tseytlin, Phys. Lett. B **487**, 141 (1997); hep-th/9609212.
[6] G. Papadopoulos and P. K. Townsend, Phys. Lett. B **380**, 273 (1996); hep-th/9603087.
[7] A. A. Tseytlin, Nucl. Phys. B **475**, 149 (1996); hep-th/9604035.
[8] J. P. Gauntlett, D. A. Kastor, and J. Traschen, Nucl. Phys. B **478**, 544 (1996); hep-th/9604179.
[9] V. D. Ivashchuk and V. N. Melnikov, hep-th/9612089; Gravitation and Cosmology **2**, 204 (1996).
[10] V. D. Ivashchuk and V. N. Melnikov, Phys. Lett. B **403**, 23 (1997).
[11] V. D. Ivashchuk and V. N. Melnikov, hep-th/9705036; Class. Quantum Grav. **14**, 3001 (1997); Corrigenda **15**, 3941 (1998).
[12] V. D. Ivashchuk, V. N. Melnikov, and M. Rainer, gr-qc/9705005 ; Gravitation & Cosmology **4**, 73 (1998).

[13] I. Ya. Aref'eva and A. I. Volovich, Class. Quantum Grav. **14**, 2990 (1997); hep-th/9611026.
[14] I. Ya. Aref'eva and O. A. Rytchkov, hep-th/9612236.
[15] R. Argurio, F. Englert, and L. Hourant, Phys. Lett. B **398**, 2991 (1997); hep-th/9701042.
[16] I. Ya. Aref'eva, M. G. Ivanov, and O. A. Rytchkov, hep-th/9702077.
[17] K. A. Bronnikov, hep-th/9710207; Gravitation & Cosmology **4**, 49 (1998).
[18] M. Cvetic and A. A. Tseytlin, Nucl. Phys. B **478**, 181 (1996).
[19] N. Ohta and T. Shimizu, hep-th/9701095.
[20] I. Ya. Aref'eva, M. G. Ivanov, and I. V. Volovich, hep-th/9702079; Phys. Lett. B **406**, 44 (1997).
[21] N. Ohta, hep-th/9702164.
[22] K. A. Bronnikov, V. D. Ivashchuk, and V. N. Melnikov, gr-qc/9710054; Gravitation & Cosmology **3**, 203 (1997).
[23] V. D. Ivashchuk and V. N. Melnikov, hep-th/9708157; J. Math. Phys. **39**, 2866 (1998).
[24] V. G. Kac, *Infinite-dimensional Lie Algebras* (Cambridge University Press, Cambridge, 1990).
[25] J. Fuchs and C. Schweigert, *Symmetries, Lie algebras and Representations. A Graduate Course for Physicists* (Cambridge University Press, Cambridge, 1997).
[26] E. Cremmer, B. Julia, and J. Scherk, Phys. Lett. B **76**, 409 (1978).
[27] E. Witten, Nucl. Phys. B **443**, 85 (1995); hep-th/9503124; P. Townsend, Phys. Lett. B **350**, 184 (1995); hep-th/9612121.
[28] A. Feingold and I. B. Frenkel, Math. Ann. **263**, 87 (1983).
[29] B. Julia, in Lectures in Applied Mathematics **21**, 355 (1985).
[30] S. Mizoguchi, hep-th/9703160.
[31] H. Nicolai, Phys. Lett. B **276**, 333 (1992).
[32] V. V. Nikulin, alg-geom/9711032.
[33] N. Khvengia, Z. Khvengia, H. Lü, and C. N. Pope, hep-th/9703012.
[34] C. M. Hull, Nucl. Phys. B **468**, 113 (1996); C. Vafa, *ibid.* **469**, 403 (1996); hep-th/9602022.
[35] V. V. Nikulin (private communication).
[36] V. D. Ivashchuk, Phys. Lett. A **170**, 16 (1992).
[37] V. D. Ivashchuk and V. N. Melnikov, Class. Quantum Grav. **11**, 1793 (1994).
[38] V. D. Ivashchuk and V. N. Melnikov, hep-th/9612089; Gravitation & Cosmology **1**, 204 (1996).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Connection with torsion from Ashtekar–Barbero connection

Merced Montesinos[a)]
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, and Departamento de Física, Centro de Investigación y de Estudios Avanzados del I.P.N., Av. I.P.N. No. 2508, 07000 Ciudad de México, México

(Received 11 December 1998; accepted for publication 12 March 1999)

A natural three-dimensional covariant derivative with torsion emerges from the Ashtekar–Barbero connection in the canonical approach to general relativity. The torsion is related to the extrinsic curvature when the Barbero–Immirzi parameter β does not vanish. Alternatively, torsion can be avoided, but at the price of a nonvanishing covariant derivative of the triad field. The properties of both cases are discussed. © *1999 American Institute of Physics.* [S0022-2488(99)01108-1]

I. INTRODUCTION

Nonperturbative quantum gravity, or loop quantum gravity,[1] emerges from mixing general relativity and quantum mechanics. More precisely, it comes from implementing and extending the Dirac method[2] for constrained systems to the Hamiltonian formulation of general relativity in terms of the Ashtekar–Barbero variables[3,4] (for instance, the Dirac approach does not provide a mechanism for defining the measure of the Hilbert space). This theory has two non trivial aspects as a quantum theory of the gravitational field. First, it is a nonperturbative quantum approach[5] for describing the gravitational field. Second it shows that the quantum nature of the gravitational field is associated to loop-like excitations, as opposed to point-like ones from usual quantum field theory. When these nontrivial facts are put together, several striking results follow from the loop quantum gravity approach to the gravitational field (see Refs. 6, 7, and 8 for a complete review). All of them show a discrete nature of *space* at Planck length $l_{\mathrm{P}}=\sqrt{G\hbar/c^3}$.

The success of the quantum theory motivates us to investigate in detail the classical formulation on which the quantization is based and, in particular, the geometry defined by this formalism. In this paper therefore, we focus on some aspects of the geometry of general relativity in terms of real Ashtekar–Barbero variables.[3,4] In Sec. II, it is shown that a natural three-dimensional covariant derivative with torsion on Σ comes from the Ashtekar–Barbero connection, the torsion is related to the extrinsic curvature K^i_a. This covariant derivative kills (the inverse of) the triad field e^i_a. Alternatively, we introduce a mechanism for avoiding torsion related to the definition of a different covariant derivative. The relationship with the Cartan framework is clarified in both cases.

II. THE CONNECTION

We begin by recalling a few facts on affine connections (see for instance Ref. 9). Let $e^i = e^i_a\, dx^a$ $(i=1,2,3)$ be the set of 1-forms dual to the orthonormal frame $e_i = e^a_i(\partial/\partial x^a) = e^a_i \partial_a$ of the tangent space of a three-dimensional manifold Σ, where x^a $(a=1,2,3)$ are local coordinates on the manifold. If ∇ is an affine connection we have, in the orthonormal frame e_i,

$$\nabla_{e_i} e_j =: -\alpha_{ij}{}^k e_k; \tag{1}$$

[a)]Electronic mail: merced@fis.cinvestav.mx therefore the connection is completely determined by the $3\times3\times3=27$ coefficients of the affine connection $\alpha_{ij}{}^{k}$. By plugging the definition of e_i in Eq. (1) and using the properties of the affine connection, we get

$$e^k_c e^b_i(\partial_b e^c_j - A_{ba}{}^{c} e^a_j) = -\alpha_{ij}{}^{k},$$
$$\nabla_{\partial_b}\partial_a = : -A_{ba}{}^{c}\partial_c, \tag{2}$$

where the orthonormality condition $e^a_j e^i_a = \delta^i_j$ has been used. $A_{ba}{}^{c}$ are the coefficients of the affine connection with respect to the coordinate frame. The following step is to compute the antisymmetric and symmetric parts of $\alpha_{ij}{}^{k}$:

$$\alpha_{[ij]}{}^{k} := \tfrac{1}{2}(\alpha_{ij}{}^{k} - \alpha_{ji}{}^{k}) = e^a_{[i}e^b_{j]}\,\partial_a e^k_b + e^k_c e^a_i e^b_j T_{ab}{}^{c},$$
$$\alpha_{(ij)}{}^{k} := \tfrac{1}{2}(\alpha_{ij}{}^{k} + \alpha_{ji}{}^{k}) = e^a_{(i}e^b_{j)}\,\partial_a e^k_b + e^k_c e^a_i e^b_j K_{ab}{}^{c}. \tag{3}$$

$T_{ab}{}^{c} := \frac{1}{2}(A_{ab}{}^{c} - A_{ba}{}^{c})$ and $K_{ab}{}^{c} := \frac{1}{2}(A_{ab}{}^{c} + A_{ba}{}^{c})$ are the torsion tensor and the symmetric part of the coefficients $A_{ab}{}^{c}$, respectively. The set $\alpha_{[ij]}{}^{k}$ and $\alpha_{(ij)}{}^{k}$ determine $T_{ab}{}^{c}$ and $K_{ab}{}^{c}$ (and vice versa) through Eq. (3). Up to now we have been concerned with general aspects of affine connections. We now specialize to the Levi-Civita connection D. This connection is given in a coordinate frame by

$$T_{ab}{}^{c} = 0,$$
$$K_{ab}{}^{c} \equiv \Gamma_{ab}{}^{c} = -\tfrac{1}{2}q^{ce}(\partial_a q_{be} + \partial_b q_{ea} - \partial_e q_{ab}), \tag{4}$$

where $q_{ab} = e^i_a e^j_b \delta_{ij}$ is the three-dimensional metric on Σ. Therefore, in the case of the Levi-Civita connection Eq. (3) reduces to

$$\omega_{[ij]}{}^{k} := e^a_{[i}e^b_{j]}\,\partial_a e^k_b,$$
$$\omega_{(ij)}{}^{k} := e^a_{(i}e^b_{j)}\,\partial_a e^k_b + e^k_c e^a_i e^b_j \Gamma_{ab}{}^{c}. \tag{5}$$

Nevertheless, frequently it is necessary to compute covariant derivatives acting on both internal and space indices. The most general covariant derivative involving $SO(3)$ and space indices is given by

$$\mathcal{D}_a \lambda^i_b := \partial_a \lambda^i_b + \epsilon^i{}_{jk} A^j_a \lambda^k_b + A_{ab}{}^{c}\lambda^i_c, \tag{6}$$

so this expression can be computed once we have specified the rule to get the expressions of A^i_a and $A_{ab}{}^{c}$ $(9+27$ quantities). The usual way to specify them is through the *intrinsic* geometry of Σ. In this case we are forced to look at the action of the covariant derivative on the triad field e^a_i (or the inverse e^i_a). The most general allowed possibility is given by

$$\mathcal{D}_a e^i_b = \partial_a e^i_b + e^i{}_{jk} A^j_a e^k_b + A_{ab}{}^{c} e^i_c = t_{ab}{}^{i}, \tag{7}$$

where $t_{ab}{}^{i}$ is the *triadity* of the covariant derivative, i.e., the action of the covariant derivative on (the inverse of) the triad field e^i_b. Therefore, A^i_a and $A_{ab}{}^{c}$ are completely determined once the triadity $t_{ab}{}^{i}$ is given together with the other nine quantities, for instance, the value of the torsion tensor $T_{ab}{}^{c}$. One option is to choose $t_{ab}{}^{i} = 0$ and $T_{ab}{}^{c} = 0$. This election determines $A^i_a \equiv \Gamma^i_a = \Gamma^i_a[e]$ via

$$\partial_{[a} e^i_{b]} + \epsilon^i{}_{jk}\Gamma^j_{[a} e^k_{b]} = 0, \tag{8}$$

and $A_{ab}{}^{c}$ is given by

$$A_{ab}{}^{c}=\Gamma_{ab}{}^{c}=-\tfrac{1}{2}q^{ce}(\partial_a q_{be}+\partial_b q_{ea}-\partial_e q_{ab}), \tag{9}$$

with $q_{ab}=e^i_a e^j_b \delta_{ij}$.

Before continuing, it is important to emphasize some properties of the framework we are dealing with. In general, the covariant derivative (6) is well defined once we have A^i_a and $A_{ab}{}^c$ [which are obtained through (7)]. What are the differences between vanishing and nonvanishing triadity $t_{ab}{}^i$ in (7)? The difference is that when the covariant derivative kills (the inverse of) the triad, $\mathcal{D}_a e^i_b=0$ (and therefore we have vanishing triadity), we are in the Cartan framework, while when $t_{ab}{}^i\neq 0$ we are out the Cartan framework. This fact can be seen from (7). When $t_{ab}{}^i$ is vanishing, the antisymmetric part of (7) reads as

$$\partial_{[a}e^i_{b]}+\epsilon^i{}_{jk}A^j_{[a}e^k_{b]}+T_{ab}{}^c e^i_c=0, \tag{10}$$

which is one of Cartan's structure equations. When $t_{ab}{}^i\neq 0$ this last equation does not hold. We now relate these notions to the fields of canonical gravity.

Non perturbative quantum gravity is based on the following canonical pair of phase space variables,[3,4]

$$\begin{aligned} A^i_a &:= \Gamma^i{}_a+\beta K^i_a, \\ \tilde{E}^a_i &:= (\det(e^i_b))e^a_i, \end{aligned} \tag{11}$$

where A^i_a is the Ashtekar–Barbero connection [in fact, when $\beta=i$, the initial formulation due to Ashtekar, A^i_a comes from the ADM formalism of the (self-dual sector) of the Plebanski action[10]], $\tilde{E}^a_i$ is the densitized triad e^a_i, and β is the Barbero–Immirzi parameter.[4,11,12] In the Ashtekar formulation of general relativity the covariant derivative is defined to act on internal indices only. The extension of the covariant derivative to act on space indices is usually defined by choosing a torsion-free extension.[4,5] In this paper, we explore the different ways in which this extension can be made. As we have mentioned, we have the possibility of working within the Cartan framework or not. Therefore, we can define two covariant derivatives that act on space indices as well. Here we explore both alternatives. Let us consider first the Cartan framework. Here, the triadity must vanish and we need to specify the torsion in order to determine completely the full covariant derivative given by (6). It is clear that if we choose

$$\begin{aligned} t_{ab}{}^i &= 0, \\ T_{ab}{}^c &= A_{[ab]}{}^c=-\beta\epsilon^i{}_{jk}K^j_{[a}e^k_{b]}e^c_i, \end{aligned} \tag{12}$$

then [by using (7)] the connections A^i_a, $A_{ab}{}^c$ are

$$\begin{aligned} A^i_a &= \Gamma^i{}_a+\beta K^i_a, \\ A_{ab}{}^c &= K_{ab}{}^c+T_{ab}{}^c=\Gamma_{ab}{}^c-2T_{(a}{}^c{}_{b)}-\beta\epsilon^i{}_{jk}K^j_{[a}e^k_{b]}e^c_i. \end{aligned} \tag{13}$$

The covariant derivative $\mathcal{D}$ kills the metric q_{ab}, $\mathcal{D}_a q_{bc}=0$. Of course, we have also $\mathcal{D}_a\delta_{ij}=0$. Note that the torsion has here a geometric origin: it depends on the extrinsic curvature K^i_a. The fact that the Ashtekar–Barbero connection has torsion and that this is related to the extrinsic curvature means that the foliation of spacetime plays a double role: defining the new variables, through the connection A^i_a, and inducing geometric nonvanishing torsion on the three-dimensional connection. Of course, the torsion appears only in the case of a nonvanishing Barbero-Immirzi parameter.

There is a second possibility for defining the covariant derivative (6), taking into account the Ashtekar–Barbero connection A^i_a. This second covariant derivative has *no torsion*, but has nonvanishing triadity. In this sense there is a mechanism for avoiding torsion, but it should be clear that this is a different covariant derivative than the one considered before, and most important, we are *not* in the Cartan framework in this last case. Concretely,

$$t_{ab}{}^{i} = \beta \epsilon^{i}{}_{jk} K_{a}^{j} e_{b}^{k}, \qquad T_{ab}{}^{c} = A_{[ab]}{}^{c} = 0, \tag{14}$$

with A_a^i and $A_{ab}{}^c$ given [using (7)] by

$$A_a^i = \Gamma^i{}_a + \beta K^i{}_a, \qquad A_{ab}{}^c = \Gamma_{ab}{}^c. \tag{15}$$

Even though $\mathcal{D}$ has nonvanishing triadity, the covariant derivative kills the metric q_{ab}, $\mathcal{D}_a q_{bc} = 0$. Of course, also $\mathcal{D}_a \delta_{ij} = 0$ holds. On the other hand, as we have mentioned, the covariant derivative based in the Ashtekar–Barbero connection A_a^i acts only on internal indices. It is usually stated in the literature[4,5] that the covariant derivative can be extended to act on space indices as well, and that such an extension can be chosen free of torsion. The last covariant derivative we have defined is precisely this extension, but in this case the covariant derivative does not kill the triad field, i.e., the covariant derivative has *non*vanishing triadity.

III. CONCLUDING REMARKS

We have shown that two natural covariant derivatives arise from the Ashtekar–Barbero connection. One of these covariant derivatives kills the inverse triad field e_a^i (and of course the three-dimensional metric q_{ab}) and has torsion. The torsion is related to the extrinsic curvature of Σ. The Barbero–Immirzi parameter enters in the value of the torsion. Therefore, the torsion has a pure *geometric* origin as opposed to models where the torsion is put by hand. This structure sits in the Cartan framework. Alternatively, there is a way for avoiding torsion, which is related to the definition of a different covariant derivative. The covariant derivative is well defined, but this case sits outside the Cartan framework. A deeper analysis of the implications of these remarks was not made here, but it would be interesting to develop it in a future work. For instance, the role of the torsion in the $3+1$ decomposition of the spacetime needs to be clarified (the torsion might change the expressions of the constraints of the Hamiltonian formalism as well as the equations of motion for the canonical fields).

ACKNOWLEDGMENTS

M.M. thanks Abhay Ashtekar the comments on this work and also to the referee for his valuable comments. M.M.'s postdoctoral fellowship at the Department of Physics and Astronomy of the University of Pittsburgh is funded through the CONACyT of Mexico, Fellow No. 91825. Also, M.M. thanks financial support provided by the *Sistema Nacional de Investigadores* of the Secretaría de Educación Pública (Mexico).

[1] T. Jacobson and L. Smolin, ''Nonperturbative quantum geometries,'' Nucl. Phys. B **299**, 295 (1988); C. Rovelli and L. Smolin, ''Knot theory and quantum gravity,'' Phys. Rev. Lett. **61**, 1155 (1988); C. Rovelli and L. Smolin, ''Loop space representation of quantum general relativity,'' Nucl. Phys. B **331**, 80 (1990).
[2] P. A. M. Dirac, *Lectures on Quantum Mechanics* (Belfer Graduate School of Science, New York, 1964).
[3] A. Ashtekar, ''New variables for classical and quantum gravity,'' Phys. Rev. Lett. **57**, 2244 (1986); A. Ashtekar, ''New Hamiltonian formulation of general relativity,'' Phys. Rev. D **36**, 1587 (1987).
[4] J. F. Barbero, ''Real Ashtekar variables for Lorentzian signature space-times,'' Phys. Rev. D **51**, 5507 (1995).
[5] A. Ashtekar, *Non-Perturbative Canonical Gravity (Notes Prepared in Collaboration With R. S. Tate)* (World-Scientific, Singapore, 1991).
[6] L. Smolin, ''Recent developments in nonperturbative quantum gravity,'' *Proceedings of the XXII G.I.F.T. International Seminar on Theoretical, Quantum Gravity and Cosmology*, edited by J. Pérez-Mercader, J. Sola, and E. Verdaguer (World Scientific, Singapore, 1992).

[7] A. Ashtekar, ''Mathematical problems of non-perturbative quantum general relativity,'' in *Gravitation and Quantization (Les Houches)*, edited by B. Julia and J. Zinn-Justin (Elsevier, Paris, 1992).
[8] C. Rovelli, ''Loop quantum gravity,'' Living Reviews in Relativity, http://www.livingreviews.org/Articles/Volume 1/1998-1rovelli, 1998; C. Rovelli, ''Strings, loops and others: a critical survey of the present approaches to quantum gravity,'' in *Gravitation and Relativity*: *At the Turn of the Millenium*, edited by N. Dadhich and J. Narlikar (Inter-University Centre for Astronomy and Astrophysics, India, 1998), gr-qc/9803024.
[9] Y. Choquet-Bruhat, C. De Witt-Morette, and C. Dillarck-Bleick, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1991), Part I.
[10] J. F. Plebanski, ''On the separation of Einsteinian substructures,'' J. Math. Phys. **18**, 2511 (1977).
[11] C. Rovelli and T. Thiemann, ''The Immirzi parameter in quantum general relativity,'' Phys. Rev. D **57**, 1009 (1998).
[12] G. Immirzi, ''Quantum gravity and Regge calculus,'' Nucl. Phys. B (Proc. Suppl.) **57**, 65 (1997).

One-instanton prepotentials from WDVV equations in $N=2$ supersymmetric SU(4) Yang–Mills theory

Yűji Ohta
Research Institute for Mathematical Sciences, Kyoto University, Sakyoku, Kyoto 606, Japan

(Received 28 April 1999; accepted for publication 10 May 1999)

Prepotentials in $N=2$ supersymmetric Yang–Mills theories are known to obey nonlinear partial differential equations called Witten–Dijkgraaf–Verlinde–Verlinde (WDVV) equations. In this paper, the prepotentials at the one-instanton level in $N=2$ supersymmetric SU(4) Yang–Mills theory are studied from the standpoint of WDVV equations. Especially, it is shown that the one-instanton prepotentials are obtained from WDVV equations by assuming the perturbative prepotential and by using the scaling relation as a subsidiary condition but are determined without introducing the Seiberg–Witten curve. In this way, various one-instanton prepotentials which satisfy both WDVV equations and the scaling relation can be derived, but it turns out that among them there exist one-instanton prepotentials which coincide with the instanton calculus. © *1999 American Institute of Physics.* [S0022-2488(99)01708-9]

I. INTRODUCTION

A class of developments of quantum field theory in the 1990s may be represented by two keywords: $N=2$ supersymmetry and duality. For example, the mirror symmetry[1–4] established in the beginning of the nineties was based on the (trivial) isomorphism between left and right U(1) currents of (2,2)-superconformal field theory,[5,6] and this isomorphism predicted the existence of a pair of Calabi–Yau manifolds whose axes of the Hodge diamond were exchanged. Candelas *et al.*[3,4] skillfully used the consequence expected from this duality of the Hodge structure and showed that the number of rational curves on the Calabi–Yau quintic three-fold could be determined from mirror symmetry. The coincidence of their result with mathematically rigorous results[1,7,8] was a great surprise!

On the other hand, also in the recent studies of low-energy effective dynamics of $N=2$ supersymmetric Yang–Mills theory, $N=2$ supersymmetry and duality play a crucial role. Before the arrival of Seiberg and Witten's proposal of using electro-magnetic duality for the description of the low-energy effective action of SU(2) gauge theory,[9,10] though it has been known that the prepotential which is a generating function of the low-energy effective action is not renormalized beyond one-loop in perturbative calculation due to $N=2$ supersymmetry,[11–13] actually the prepotential was expected to receive instanton corrections.[14] Unfortunately, such corrections were not so extensively discussed, but thanks to their proposal, it made it possible to extract information on instanton effects from a Riemann surface and periods of meromorphic differential on it. Namely, the low-energy effective theory turned out to be parametrized by a Riemann surface. The validity of their proposal was discussed by Klemm *et al.*[15] with the aid of the Picard–Fuchs equation and the instanton corrections to the prepotential were revealed. The instanton corrections obtained in this way showed extremely good agreement with the prediction of instanton calculus.[16–21]

However, deeper and striking features of prepotentials of $N=2$ supersymmetric Yang–Mills theories may be nicely interpreted in terms of differential equations satisfied by prepotentials. For instance, it is well known that the prepotentials satisfy a Euler equation called a scaling relation,[22–26] and, in fact, this simple equation simplified and accelerated the study of prepotentials. As for another characteristic equation, we can mention that there is a nonlinear system of partial differential equations called Witten–Dijkgraaf–Verlinde–Verlinde (WDVV) equations[27–31] (rigorously speaking, the WDVV equations in $N=2$ Yang–Mills theory are not equivalent to those arising in two-dimensional topological field theory[32–35]). Actually, these equations hold not only in four-dimensional gauge theories but also in higher dimensions even if hypermultiplets are included.[28,29] Accordingly, it becomes possible to regard the prepotentials in various gauge theories as a member of solutions to WDVV equations. Then, what is the most general solution (function form of prepotentials) to the WDVV equations? Unfortunately, we cannot precisely know the answer to this question, but Braden *et al.*[36] partially found the answer. They assumed the function form of the prepotential which is expected from known examples and found a new prepotential which is considered as that in five-dimensional gauge theory, although their study was restricted to the perturbative part. Of course, among the solutions found by them we can see the existence of the prepotential in four-dimensional Yang–Mills theory. This seems to indicate that the prepotentials can be constructed without introducing the Riemann surface, provided the WDVV equations are used. Finding whether nonperturbative prepotentials are available from WDVV equations without using the Riemann surface is the subject of this paper.

The paper is organized as follows. In Sec. II, the construction of a perturbative solution to WDVV equations for SU(4) gauge theory in four dimensions discussed by Braden *et al.*[36] is summarized. We can see that the perturbative prepotential is, in fact, obtained from WDVV equations. In Sec. III, we add the nonperturbative part for this perturbative prepotential and try to solve the WDVV equations. Though the nonperturbative part satisfies a nonlinear differential equation, restricting it at the one-instanton level, we can reduce it to a linear differential equation satisfied by the one-instanton prepotential. For this reason, the one-instanton prepotential is investigated in this paper. To solve this equation, the scaling relation is used as a subsidiary condition, but it turns out that there are miscellaneous solutions which do not contradict both WDVV equations and the scaling relation. In Sec. IV we compare our result with the prediction of the one-instanton calculus. It is shown that among our one-instanton prepotentials obtained from WDVV equations there are one-instanton prepotentials which agree with the prediction of instanton calculus. In this way, we conclude that it is possible to obtain a nonperturbative prepotential from WDVV equations without relying on the Riemann surface. Section V is a brief summary.

II. PERTURBATIVE PREPOTENTIAL FROM WDVV EQUATIONS

In this section, we briefly outline the construction to get perturbative prepotential from WDVV equations in the SU(4) gauge theory presented by Braden *et al.*[36] Note that the SU(4) model is the simplest and nontrivial example for a study of WDVV equations.

In this case, the WDVV equations for the prepotential $\mathcal{F}$ take the form

$$(\mathcal{F}_i)(\mathcal{F}_k)^{-1}(\mathcal{F}_j)=(\mathcal{F}_j)(\mathcal{F}_k)^{-1}(\mathcal{F}_i), \quad i,j,k=1,...,3, \tag{2.1}$$

where

$$(\mathcal{F}_i)\equiv(\mathcal{F}_i)_{jk}=\frac{\partial^3\mathcal{F}}{\partial a_i\partial a_j\partial a_k} \tag{2.2}$$

are the matrix notations and in this paper the brackets are always added as $(\mathcal{F}_i)$ when "$\mathcal{F}_i$" mean matrices. The coordinates a_i are the periods of the SU(4) gauge theory (see the Appendix).

Braden *et al.*[36] considered the perturbative prepotential in the form

$$\mathcal{F}_{\text{per}}(a_1,a_2,a_3)=\sum_{i<j=1}^{4} f(a_{ij}), \quad a_{ij}=a_i-a_j, \quad \sum_{i=1}^{4} a_i=0. \tag{2.3}$$

Of course $\mathcal{F}_{\text{per}}$ may depend on the mass scale $\Lambda\equiv\Lambda_{SU(4)}^8$ of the theory, but we can ignore its dependence for the moment because Λ-differentiation is not included in the WDVV equations (2.1).

Under the assumption (2.3), we can find that when $\mathcal{F}_{\rm per}$ satisfies (2.1) there is a functional relation

$$g(a_{12})g(a_{34})-g(a_{13})g(a_{24})+g(a_{14})g(a_{23})=0, \tag{2.4}$$

where

$$g(a)\equiv\left(\frac{\partial^3 f}{\partial a^3}\right)^{-1}. \tag{2.5}$$

With the aid of several conditions, we can conclude that g is an odd function with

$$g(0)=g''(0)=0, \tag{2.6}$$

where the prime means the differentiation over the argument.[36]

There are several functions enjoying the properties (2.4) and (2.6), but a function which is necessary for us among them is the function of the form $g(a)=a$. Namely,

$$f(a)=\frac{a^2}{2}\ln a+O(a^2). \tag{2.7}$$

Note that $O(a^2)$ term cannot be fixed from the WDVV equations because they are third-order differential equations. Namely, the classical part is not fixed. It is easy to see that substituting (2.7) back to (2.3) in fact yields the perturbative part of the SU(4) prepotential (in a suitable normalization).

Marshakov *et al.*[29] give a general proof that the perturbative prepotentials in various gauge theories satisfy WDVV equations, but this is also confirmed by Ito and Yang[30] in their study of these equations.

III. ONE-INSTANTON PREPOTENTIALS

A. Differential equations for one-instanton prepotentials

Next, let us consider whether nonperturbative prepotential $\mathcal{F}$ is available from the WDVV equations by assuming the form

$$\mathcal{F}(a_1,a_2,a_3,\Lambda)=\mathcal{F}_{\rm per}(a_1,a_2,a_3)+\mathcal{F}_{\rm ins}(a_1,a_2,a_3,\Lambda), \tag{3.1}$$

where

$$\mathcal{F}_{\rm ins}(a_1,a_2,a_3,\Lambda)=\sum_{k=1}^{\infty}\mathcal{F}_k(a_1,a_2,a_3)\Lambda^k. \tag{3.2}$$

In order to derive differential equations for $\mathcal{F}_{\rm ins}$, we assume that $\mathcal{F}_{\rm per}$ is already given by (2.3) with (2.7).

Then substituting $\mathcal{F}$ into (2.1) we can obtain a single nonlinear differential equation for $\mathcal{F}_{\rm ins}$, but if we restrict only the case $k=1$ (one-instanton level), the equation reduces to

$$\begin{aligned}&\frac{\partial_1^3\mathcal{F}_1}{a_{12}a_{13}a_{14}}-\frac{\partial_2^3\mathcal{F}_1}{a_{12}a_{23}a_{24}}+\frac{\partial_3^3\mathcal{F}_1}{a_{13}a_{23}a_{34}}+\frac{6\partial_1\partial_2\partial_3\mathcal{F}_1}{a_{14}a_{24}a_{34}}-\frac{A_{012}\partial_2\partial_3^2\mathcal{F}_1+A_{021}\partial_2^2\partial_3\mathcal{F}_1}{a_{12}a_{13}a_{23}a_{24}a_{34}}\\&-\frac{A_{102}\partial_1\partial_3^2\mathcal{F}_1-A_{201}\partial_1^2\partial_3\mathcal{F}_1}{a_{12}a_{13}a_{14}a_{23}a_{34}}+\frac{A_{120}\partial_1\partial_2^2\mathcal{F}_1+A_{210}\partial_1^2\partial_2\mathcal{F}_1}{a_{12}a_{13}a_{14}a_{23}a_{24}}=0,\end{aligned} \tag{3.3}$$

where $\partial_i\equiv\partial/\partial a_i$ and

$$
\begin{aligned}
A_{012}&=a_1a_2-3a_2^2-2a_1a_3+4a_2a_3+a_1a_4+a_2a_4-2a_3a_4,\\
A_{021}&=2a_1a_2-a_1a_3-4a_2a_3+3a_3^2-a_1a_4+2a_2a_4-a_3a_4,\\
A_{102}&=3a_1^2-a_1a_2-4a_1a_3+2a_2a_3-a_1a_4-a_2a_4+2a_3a_4,\\
A_{201}&=2a_1a_2-4a_1a_3-a_2a_3+3a_3^2+2a_1a_4-a_2a_4-a_3a_4,\\
A_{120}&=3a_1^2-4a_1a_2-a_1a_3+2a_2a_3-a_1a_4+2a_2a_4-a_3a_4,\\
A_{210}&=4a_1a_2-3a_2^2-2a_1a_3+a_2a_3-2a_1a_4+a_2a_4+a_3a_4.
\end{aligned} \tag{3.4}
$$

B. The solutions

In order to solve (3.3), let us introduce the new variables

$$x=a_{12},\quad y=a_{13},\quad z=a_{14}. \tag{3.5}$$

In addition, using Euler derivatives $\theta_x=x\partial/\partial x$, etc., we can rewrite (3.3) as

$$L(\theta_x,\theta_y,\theta_z)\mathcal{F}_1=0, \tag{3.6}$$

where

$$
\begin{aligned}
L(\theta_x,\theta_y,\theta_z)=&\,yz(y-z)\theta_x(\theta_x-1)(\theta_x-2)+z(4xy-3y^2-2xz+yz)(\theta_x-1)\theta_x\theta_y\\
&+z(3x^2-4xy-xz+2yz)(\theta_y-1)\theta_x\theta_y-xz(x-z)\theta_y(\theta_y-1)(\theta_y-2)\\
&-y(3x^2-xy-4xz+2yz)(\theta_z-1)\theta_x\theta_y-y(4xz+yz-3z^2-2xy)(\theta_x-1)\theta_x\theta_z\\
&+6(x-y)(x-z)(y-z)\theta_x\theta_y\theta_z+x(xz+4yz-3z^2-2xy)(\theta_y-1)\theta_y\theta_z\\
&+x(3y^2+2xz-4yz-xy)(\theta_z-1)\theta_y\theta_z+xy(x-y)\theta_z(\theta_z-1)(\theta_z-2).
\end{aligned} \tag{3.7}
$$

Here, suppose that $\mathcal{F}_1$ is given by

$$\mathcal{F}_1=x^{\nu_1}y^{\nu_2}z^{\nu_3}F(x,y,z), \tag{3.8}$$

where

$$F(x,y,z)=\sum_{i,j,k=0}^{\infty} B_{\epsilon_x i,\epsilon_y j,\epsilon_z k}\, x^{\epsilon_x i}y^{\epsilon_y j}z^{\epsilon_z k} \tag{3.9}$$

and the expansion coefficients are assumed to be independent of x, y, and z. In (3.9), ϵ_x, etc. are signature symbols (the choice of signatures depends on where the convergence region is), e.g., $\epsilon_x=\pm$. Then from (3.3) we get the differential equation for F,

$$L(\theta_x+\nu_1,\theta_y+\nu_2,\theta_z+\nu_3)F=0, \tag{3.10}$$

and the indicial equations for ν_i:

$$
\begin{gathered}
\nu_1(\nu_1-2\nu_2-1)(\nu_1+\nu_2-3\nu_3-2)=0,\quad \nu_2(2\nu_1-\nu_2+1)(\nu_1+\nu_2-3\nu_3-2)=0,\\
\nu_2(\nu_2-2\nu_3-1)(3\nu_1-\nu_2-\nu_3+2)=0,\quad (\nu_2-\nu_3)(\nu_1^2-\nu_1-\nu_1\nu_2+\nu_2\nu_3)=0,\\
\nu_1(\nu_1^2-3\nu_1+5\nu_2-3\nu_1\nu_2+\nu_3-\nu_1\nu_3+4\nu_2\nu_3+2)=0,
\end{gathered} \tag{3.11}
$$

$$\nu_3^2(\nu_3-3\nu_2-3)+5\nu_1\nu_2\nu_3+3\nu_2\nu_3-2\nu_1^2\nu_3+2\nu_1\nu_3+2\nu_3+\nu_1\nu_2=0,$$

$$\nu_3^2(\nu_3-\nu_2-3)-2\nu_2^2\nu_3+3\nu_1\nu_2\nu_3+3\nu_2\nu_3+3\nu_1\nu_2+2\nu_3=0.$$

The sets of possible ν_i are thus given by

$$\boldsymbol{\nu}\equiv(\nu_1,\nu_2,\nu_3)=(-2,-2,-2),(-1,-1,-1),(0,-1,-1)^4,(0,0,0)^2,(0,0,1),(0,0,2)^2,$$
$$(0,1,0,),(0,1,1),(0,2,0)^2,(1,0,0)^2,(1,0,1),(1,0,2),(2,0,0)^2, \tag{3.12}$$

where the superscript means degeneracy, e.g., $(0,0,0)^2$ is composed of two (0, 0, 0), but we do not discuss the consequence of degeneracy in this paper. For the indices (3.12), it is straightforward to obtain F. Though we have expressed F in (3.9) as an infinite series, actually we can restrict possible terms in (3.9) by considering the degree counting of the one-instanton prepotential.

C. Scaling relation for one-instanton prepotential

If the WDVV equations can, in fact, yield a physically acceptable prepotential, the prepotential obtained from those equations must also satisfy the fundamental homogeneity condition called the scaling relation.[22–26] Therefore, we may use it as a subsidiary condition for the problem how to solve WDVV equations in gauge theory.

To see this, first, let us recall the scaling relation[22–26]

$$\sum_{i=1}^{3} a_i\frac{\partial\mathcal{F}}{\partial a_i}+\Lambda_{\mathrm{SU}(4)}\frac{\partial\mathcal{F}}{\partial\Lambda_{\mathrm{SU}(4)}}=2\mathcal{F}. \tag{3.13}$$

We need a scaling relation for $\mathcal{F}_1$, not for $\mathcal{F}$ itself, but in order to extract it from (3.13), the Λ-dependence of perturbative prepotential which cannot be fixed from WDVV equations must be included appropriately. For this, our choice here is

$$\mathcal{F}_{\mathrm{per}}=\sum_{i<j=1}^{4}\frac{a_{ij}^2}{2}\ln\frac{a_{ij}}{\Lambda_{\mathrm{SU}(4)}}. \tag{3.14}$$

Then from (3.13), $\mathcal{F}_1$ is found to satisfy

$$\sum_{i=1}^{3} a_i\frac{\partial\mathcal{F}_1}{\partial a_i}+6\mathcal{F}_1=0, \tag{3.15}$$

which indicates that $\mathcal{F}_1$ is a homogeneous function of degree -6.

In the variables (3.5), (3.15) becomes

$$x\partial_x\mathcal{F}_1+y\partial_y\mathcal{F}_1+z\partial_z\mathcal{F}_1+6\mathcal{F}_1=0. \tag{3.16}$$

Accordingly, from (3.8), (3.9), and (3.16) it must be always true that

$$\nu_1+\nu_2+\nu_3+\epsilon_x i+\epsilon_y j+\epsilon_z k=-6. \tag{3.17}$$

D. Examples of one-instanton prepotentials

We now have enough information to construct explicit one-instanton prepotentials which do not contradict WDVV equations and the scaling relation.

To begin with, let us consider the case $\boldsymbol{\epsilon}\equiv(\epsilon_x,\epsilon_y,\epsilon_z)=(+,+,+)$. In this case, we can easily find that there exists only one solution which satisfies (3.17). It is the solution with $\boldsymbol{\nu}=(-2,-2,-2)$, and thus

$$\mathcal{F}_1=\frac{B_{0,0,0}}{x^2y^2z^2}. \tag{3.18}$$

However, when $\boldsymbol{\epsilon}=(-,-,-)$ or one entry of $\boldsymbol{\epsilon}$ differs from the others, e.g., $\boldsymbol{\epsilon}=(-,-,+)$, the situation changes, in particular, drastically in the latter case. In the former case, it is easy to see that F consists of a finite number of terms for all indices in (3.12), but in the latter case F is generally represented by an infinite number of terms as long as (3.17) is satisfied. We do not know whether it is possible to find any physical meaning for this type of one-instanton prepotential, but it may be interesting to recall that a similar one was observed in the one-instanton prepotential in the five-dimensional gauge theory.[26]

Since the latter case mentioned above is slightly intractable, let us consider an example of the former case instead. In the case of $\boldsymbol{\nu}=(-1,-1,-1)$ with $\boldsymbol{\epsilon}=(-,-,-)$, for instance, we have

$$\mathcal{F}_1=\frac{1}{xyz}\left[B_{0,-1,-2}\left(\frac{1}{yz^2}+\frac{1}{y^2z}\right)+B_{-1,0,-2}\left(\frac{1}{xz^2}+\frac{1}{x^2z}\right)\right.$$
$$\left.+B_{-1,-2,0}\left(\frac{1}{x^2y}+\frac{1}{xy^2}\right)+B_{-1,-1,-1}\frac{1}{xyz}\right]. \tag{3.19}$$

Note that (3.19) includes the one-instanton prepotential of the form (3.18).

In a similar manner, we can construct one-instanton prepotentials for all other possible values of ν_i, which do not contradict both WDVV equations and the scaling relation, but it would not be necessary to explicitly show them here. However, we should point out that since (3.3) is a partial differential system, we can expect that there exist more and more various solutions. In fact, this observation is right, and we can show that also in the variables

$$(x,y,z)=(a_{12},a_{23},a_{24}),(a_{13},a_{23},a_{34}) \tag{3.20}$$

we can construct miscellaneous one-instanton prepotentials. Among them one-instanton prepotentials of the form (3.18) are included.

Remark: The function form of the one-instanton prepotentials can be determined by solving the WDVV equations, but its numerical factors, i.e., instanton expansion coefficients, are not obtained because they correspond to integration constants. In order to determine them, it is necessary to rewrite the scaling relation as a relation between prepotential and moduli. Then substituting the one-instanton prepotential obtained from WDVV equations into this scaling relation, we will be able to get the expansion coefficients. Of course, in this case the moduli must be represented as a function of periods and its expansion coefficients must be determined. However, since knowing moduli is equivalent to introducing a Seiberg-Witten curve, this method based on scaling relation represented by using moduli is not preferable in the formalism of WDVV equations because prepotentials available from WDVV equations should be determined without introduction of Seiberg–Witten curves. Accordingly, when the determination of instanton expansion coefficients is required, they should be determined from the result of instanton calculus.

IV. ONE-INSTANTON PREPOTENTIAL FROM INSTANTON CALCULUS

We have derived one-instanton prepotentials by solving the WDVV equations in the previous section. Though these one-instanton prepotentials satisfy the WDVV equations and the scaling relation, unfortunately, in view of WDVV equations, we cannot determine which ones are physically acceptable. For this reason, in order to extract physically meaningful one-instanton prepotentials among them, we must compare our result with the one-instanton prediction of instanton calculus.

In the case of SU(4) gauge theory, a one-instanton contribution for prepotential is given by[17,18]

$$\mathcal{F}_1=\frac{\Delta_4'}{\Delta_4}, \tag{4.1}$$

where we have omitted the numerical normalization factor and

$$\Delta_4'=\sum_{i=1}^{4}\prod_{\substack{k<l=1\\k,l\neq i}}^{4}(a_k-a_l)^2,\quad \Delta_4=\prod_{k<l=1}^{4}(a_k-a_l)^2. \tag{4.2}$$

The closed form of the one-instanton prepotential for SU(N_c) gauge theory is also obtained by solving Picard–Fuchs equations[37] and direct calculation of period integrals.[38,39]

Note that (4.1) is a sum of (3.18) and those for (3.20):

$$\mathcal{F}_1=\frac{1}{(a_{12}a_{13}a_{14})^2}+\frac{1}{(a_{12}a_{23}a_{24})^2}+\frac{1}{(a_{13}a_{23}a_{34})^2} \tag{4.3}$$

up to constant factors. Accordingly, we can conclude that the WDVV equations can yield physical prepotential without introducing a Riemann surface.

V. SUMMARY

In this paper, we have discussed the nonperturbative prepotential of $N=2$ supersymmetric SU(4) Yang–Mills theory from the standpoint of WDVV equations. Especially, we have found a differential equation for a one-instanton prepotential and constructed its solutions. The method to get prepotentials based on WDVV equations is fascinating in the point that the prepotentials can be obtained without introducing Seiberg–Witten curves, but it has been shown that unfortunately too many prepotentials exist in contrast with the approach based on Seiberg–Witten curves which uniquely determines a prepotential. Nevertheless, we have succeeded in showing that one-instanton prepotentials which coincide with the one-instanton calculus can be obtained from WDVV equations.

As for another aspect of WDVV equations, we should mention a connection to topological field theory in two dimensions. From the appearance of WDVV equations, it may be natural to think that the low-energy effective theory is actually a kind of topological field theory, but we must notice that *a priori* there is no reason that the effective theory must be a topological field theory. Therefore, topological or not topological: that is the question.

An approach to argue this implication more explicitly is to regard the Seiberg–Witten curves often identified with spectral curves of integrable system as if they were superpotentials of topological $\mathbf{CP}^1$ model.[31] Although this observation strongly relies on the existence of a Riemann surface (Seiberg–Witten curve), it enables us to find a connection to topological field theory, specifically, topological string theory at genus zero level.[31] Then, even if we do not assume a Riemann surface, can we find a topological nature of the effective theory? Probably WDVV equations give the answer, but the study is a subject in the future.

ACKNOWLEDGMENT

The author acknowledges Professor H. Kanno for discussions about WDVV equations.

APPENDIX: THE SU(4) SEIBERG–WITTEN SOLUTION

In this Appendix, we briefly summarize the SU(4) Seiberg–Witten solution. The SU(4) Seiberg–Witten curve is given by the hyperelliptic curve of genus three[40–43]

$$y^2=(x^4-ux^2-vx-w)^2-\Lambda_{\mathrm{SU}(4)}^8, \tag{A1}$$

where $(x,y)\in\mathbf{C}^2$ is the local coordinate, and u, v, and w are moduli of the theory. Then the Seiberg–Witten differential and its periods are given by

$$\lambda_{\mathrm{SW}}=\frac{x\partial_x W}{y}dx, \quad W=x^4-ux^2-vx-w \tag{A2}$$

and

$$a_i=\oint_{\alpha_i}\lambda_{\mathrm{SW}}, \quad a_{D_i}=\oint_{\beta_i}\lambda_{\mathrm{SW}}, \quad i=1,2,3, \tag{A3}$$

respectively, where α_i and β_i are the canonical bases of the one-cycles on the curve and the numerical normalization factor of the Seiberg–Witten differential is ignored. It is convenient to use the period vector

$$\Pi=\begin{pmatrix} a_{D_i} \\ a_i \end{pmatrix}. \tag{A4}$$

In general, these periods satisfy Fuchsian differential equations and in the case at hand they are given by[15,18,44,45]

$$\begin{aligned}
&\mathcal{L}_1\Pi\equiv[\partial_v^2-\partial_u\partial_w]\Pi=0,\\
&\mathcal{L}_2\Pi\equiv[4\partial_u^2-2u\partial_u\partial_w-v\partial_v\partial_w-\partial_w]\Pi=0,\\
&\mathcal{L}_3\Pi\equiv[v\partial_w^2+2u\partial_v\partial_w-4\partial_u\partial_v]\Pi=0,\\
&[4(u^2+24w)\partial_u^2+9v^2\partial_v^2-16(\Lambda_{\mathrm{SU}(4)}^8-w^2)\partial_w^2+12uv\partial_u\partial_v-32uw\partial_u\partial_w+3v\partial_v-16w\partial_w+1]\Pi\\
&\quad=0,
\end{aligned} \tag{A5}$$

which can be summarized as

$$\begin{aligned}
&\left[\theta_v(\theta_v-1)-\frac{v^2}{uw}\theta_u\theta_w\right]\Pi=0,\\
&\left[(2\theta_u+\theta_v+1)\theta_w-\frac{4w}{u^2}\theta_u(\theta_u-1)\right]\Pi=0,\\
&\left[(2\theta_u+3\theta_v+4\theta_w-1)^2-\frac{16\Lambda_{\mathrm{SU}(4)}^8}{w^2}\theta_w(\theta_w-1)\right]\Pi=0,
\end{aligned} \tag{A6}$$

where $\theta_w=w\partial_w$, etc., are Euler derivatives, provided the first, second, and last equations in (A5) are chosen as independent equations. Note that the third one in (A5) is not an independent equation because $(v\partial_w\mathcal{L}_1+\partial_v\mathcal{L}_2+\partial_u\mathcal{L}_3)\Pi=0$.

Introducing new variables x, y, and z by

$$x=\frac{\Lambda_{\mathrm{SU}(4)}^8}{4w^2}, \quad y=\frac{v^2}{4uw}, \quad z=\frac{w}{u^2}, \tag{A7}$$

we find that (A6) is converted into

$$[(8\theta_x+1)^2-64x(2\theta_x+\theta_y-\theta_z)(2\theta_x+\theta_y-\theta_z+1)]\Pi=0,$$

$$[\theta_y(2\theta_y-1)-2y(\theta_y+2\theta_z)(2\theta_x+\theta_y-\theta_z)]\Pi=0, \tag{A8}$$

$$[(2\theta_x+\theta_y-\theta_z)(4\theta_z-1)-4z(\theta_y+2\theta_z)(\theta_y+2\theta_z+1)]\Pi=0.$$

This system (A8) further simplifies to

$$[\theta_x^2-x(2\theta_x+\theta_y-\theta_z-1/2)(2\theta_x+\theta_y-\theta_z+1/2)]\tilde{\Pi}=0,$$

$$[\theta_y(\theta_y-1/2)-y(\theta_y+2\theta_z+1/2)(2\theta_x+\theta_y-\theta_z-1/2)]\tilde{\Pi}=0,$$

$$[(2\theta_x+\theta_y-\theta_z-1/2)\theta_z-z(\theta_y+2\theta_z+1/2)(\theta_y+2\theta_z+3/2)]\tilde{\Pi}=0 \tag{A9}$$

by $\Pi=x^{-1/8}z^{1/4}\tilde{\Pi}$. An analytic solution around $(x,y,z)=(0,0,0)$ is given by

$$\tilde{\Pi}=\sum_{m,n,p=0}^{\infty}\frac{(1/2)_{n+2p}(-1/2)_{2m+n-p}}{(1)_m(1/2)_n}\frac{x^m}{m!}\frac{y^n}{n!}\frac{z^p}{p!}, \tag{A10}$$

which is known as the type 54*b* Srivastava and Karlsson's (Gaussian) hypergeometric function in three variables,[46] and we denote it by

$$G_{54b}[\alpha,\beta;\gamma,\delta;x,y,z]=\sum_{m,n,p=0}^{\infty}\frac{(\alpha)_{n+2p}(\beta)_{2m+n-p}}{(\gamma)_m(\delta)_n}\frac{x^m}{m!}\frac{y^n}{n!}\frac{z^p}{p!}, \tag{A11}$$

which recovers Horn's $\mathcal{H}_4$ for $p=0$. Note that G_{54b} satisfies

$$[(\theta_x+\gamma-1)\theta_x-x(2\theta_x+\theta_y-\theta_z+\beta)(2\theta_x+\theta_y-\theta_z+\beta+1)]G_{54b}=0,$$

$$[(\theta_y+\delta-1)\theta_y-y(\theta_y+2\theta_z+\alpha)(2\theta_x+\theta_y-\theta_z+\beta)]G_{54b}=0, \tag{A12}$$

$$[(2\theta_x+\theta_y-\theta_z+\beta)\theta_z-z(\theta_y+2\theta_z+\alpha)(\theta_y+2\theta_z+\alpha+1)]G_{54b}=0.$$

[1] S. T. Yau, *Essays on Mirror Manifolds* (International, Hong Kong, 1992).
[2] D. R. Morrison, J. Am. Math. Soc. **6**, 223 (1993).
[3] P. Candelas, X. C. de la Ossa, P. S. Green, and L. Parkes, Nucl. Phys. B **359**, 21 (1991).
[4] P. Candelas, X. C. de la Ossa, P. S. Green, and L. Parkes, Phys. Lett. B **258**, 118 (1991).
[5] L. J. Dixson, in *Superstrings, Unified Theories and Cosmology, 1987*, Proceedings of the summer workshop in high energy physics and cosmology, Trieste, The ICTP series in theoretical physics, Vol. 4, edited by G. Furlan, J. C. Pati, D. W. Sciama, E. Sezgin, and Q. Shafi (World Scientific, Singapore, 1988).
[6] B. R. Greene and M. R. Plesser, Nucl. Phys. B **338**, 15 (1990).
[7] D. R. Morrison, ''Mathematical aspects of mirror symmetry,'' alg-geom/9609021.
[8] S. Katz, ''Rational curves on Calabi-Yau manifolds: verifying predictions of mirror symmetry,'' hep-th/9202017.
[9] N. Seiberg and E. Witten, Nucl. Phys. B **431**, 484 (1994).
[10] N. Seiberg and E. Witten, Nucl. Phys. B **435**, 129 (1994).
[11] P. S. Howe, K. S. Stelle, and P. K. Townsend, Nucl. Phys. B **214**, 519 (1983).
[12] P. S. Howe, K. S. Stelle, and P. K. Townsend, Nucl. Phys. B **236**, 125 (1984).
[13] P. S. Howe, K. S. Stelle, and P. C. West, Phys. Lett. B **124**, 55 (1983).
[14] N. Seiberg, Phys. Lett. B **206**, 75 (1988).
[15] A. Klemm, W. Lerche, and S. Theisen, Int. J. Mod. Phys. A **11**, 1929 (1996).
[16] D. Finnell and P. Pouliot, Nucl. Phys. B **453**, 225 (1995).
[17] K. Ito and N. Sasakura, Phys. Lett. B **382**, 95 (1996).
[18] K. Ito and N. Sasakura, Nucl. Phys. B **484**, 141 (1997).
[19] M. J. Slater, Phys. Lett. B **403**, 57 (1997).
[20] N. Dorey, V. V. Khoze, and M. P. Mattis, Phys. Lett. B **388**, 324 (1996); Phys. Rev. D **54**, 2921 (1996).
[21] H. Aoyama, T. Harano, M. Sato, and S. Wada, Phys. Lett. B **388**, 331 (1996).
[22] M. Matone, Phys. Lett. B **357**, 342 (1995).
[23] J. Sonnenschein, S. Theisen, and S. Yankielowicz, Phys. Lett. B **367**, 145 (1996).
[24] T. Eguchi and S.-K. Yang, Mod. Phys. Lett. A **11**, 131 (1996).
[25] P. S. Howe and P. C. West, Nucl. Phys. B **486**, 425 (1997).
[26] H. Kanno and Y. Ohta, Nucl. Phys. B **530**, 73 (1998).

[27] A. Marshakov, A. Mironov, and A. Morozov, Phys. Lett. B **389**, 43 (1996).
[28] A. Marshakov, A. Mironov, and A. Morozov, Mod. Phys. Lett. A **12**, 773 (1997).
[29] A. Marshakov, A. Mironov, and A. Morozov, ''More evidence for the WDVV equations in $N=2$ SUSY Yang-Mills theories,'' hep-th/9701123.
[30] K. Ito and S.-K. Yang, Phys. Lett. B **433**, 56 (1998).
[31] K. Ito, C.-S. Xiong, and S.-K. Yang, Phys. Lett. B **441**, 155 (1998).
[32] E. Witten, Nucl. Phys. B **340**, 281 (1990).
[33] R. Dijkgraaf and E. Witten, Nucl. Phys. B **342**, 486 (1990).
[34] R. Dijkgraaf, E. Verlinde, and H. Verlinde, Nucl. Phys. B **352**, 59 (1991).
[35] B. Dubrovin, ''Geometry of 2d topological field theories,'' hep-th/9407018.
[36] H. W. Braden, A. Marshakov, A. Mironov, and A. Morozov, Phys. Lett. B **448**, 195 (1999).
[37] K. Ito and S.-K. Yang, Int. J. Mod. Phys. A **13**, 5373 (1998).
[38] E. D'Hoker, I. M. Krichever, and D. H. Phong, Nucl. Phys. B **489**, 179 (1997).
[39] T. Masuda and H. Suzuki, Int. J. Mod. Phys. A **13**, 1495 (1998).
[40] A. Klemm, W. Lerche, S. Yankielowicz, and S. Theisen, Phys. Lett. B **344**, 169 (1995).
[41] P. C. Argyres and A. E. Faraggi, Phys. Rev. Lett. **74**, 3931 (1995).
[42] A. Hanany and Y. Oz, Nucl. Phys. B **452**, 283 (1995).
[43] M. R. Abolhasani, M. Alishahiha, and A. M. Ghezelbash, Nucl. Phys. B **480**, 279 (1996).
[44] J. M. Isidro, A. Mukherjee, J. P. Nunes, and H. J. Schnitzer, Nucl. Phys. B **492**, 647 (1997).
[45] M. Alishahiha, Phys. Lett. B **398**, 100 (1997).
[46] H. M. Srivastava and P. W. Karlsson, *Multiple Gaussian Hypergeometric Series* (Ellis Horwood Limited, West Sussex, 1985).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Self-duality and gravitational perturbations

G. F. Torres del Castillo
Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, 72570 Puebla, Puebla, México

H. G. Solís-Rodríguez
Escuela Superior de Física y Matemáticas, Instituto Politécnico Nacional, 07738 México, D.F., México

(Received 2 February 1999; accepted for publication 21 April 1999)

It is shown that there exists a one-to-one correspondence between the real solutions of the Einstein vacuum field equations linearized about the Minkowski metric and the (complex) metric perturbations whose curvature to first order in the metric perturbation is self-dual. It is also shown that the self-duality condition of the curvature to first order in the metric perturbation is equivalent to a set of first-order equations for the metric perturbation whose solution is given by a scalar potential that obeys the wave equation.
[S0022-2488(99)00808-7]

I. INTRODUCTION

Every real solution of the source-free Maxwell equations, in flat or curved space–time, is the real part of a self-dual (or an anti-self-dual) Maxwell field. This fact allows us to find all the solutions of the source-free Maxwell equations (which constitute a set of second-order partial differential equations for the vector potential) by solving the self-duality conditions (which are a set of first-order partial differential equations for the vector potential) and in this manner one can demonstrate that in flat space–time or in a curved space–time that admits a geodetic and shear-free null congruence defined by a repeated principal null direction of the curvature, every solution of the source-free Maxwell equations can be locally expressed in terms of a single complex scalar potential.[1] The resulting expression coincides with one previously obtained by means of other approaches (see, e.g., Refs. 2–4).

The linearized Einstein vacuum field equations are similar to the source-free Maxwell equations in several ways. In fact, if the background space–time is flat, then the curvature tensor to first order in the metric perturbation, $K_{\alpha\beta\gamma\delta}$, obeys equations analogous to the source-free Maxwell equations. Moreover, in the case of a flat background, the tensor field $K_{\alpha\beta\gamma\delta}$ is invariant under the gauge transformations $h_{\alpha\beta}\mapsto h_{\alpha\beta}+2\nabla_{(\alpha}\xi_{\beta)}$ and, making use of spherical or cylindrical coordinates, it can be expressed in terms of a single complex scalar potential that satisfies the wave equation.[5,6] On the other hand, an expression for the metric perturbations of flat space–time or of an algebraically special solution of the Einstein vacuum field equations in terms of a complex scalar potential has been obtained by various approaches[7,3,8] which, however, do not demonstrate that every solution of the Einstein vacuum field equations linearized about these backgrounds can be written in that form. It may be recalled that, by contrast, in the standard treatment of the Einstein vacuum field equations linearized about the Minkowski metric, by imposing an appropriate gauge condition, one finds that each Cartesian component of the metric perturbation satisfies the wave equation (see, e.g., Refs. 9 and 10).

In this paper we show that, by analogy with the electromagnetic case, every real solution of the Einstein vacuum field equations linearized about the Minkowski metric is the real part of a metric perturbation whose curvature tensor to first order in the metric perturbation, $K_{\alpha\beta\gamma\delta}$, is self-dual (or anti-self-dual) and that if $K_{\alpha\beta\gamma\delta}$ is self-dual, then the metric perturbation can be expressed in terms of a complex scalar potential that satisfies the wave equation.

II. SELF-DUALITY AND THE LINEARIZED EINSTEIN VACUUM FIELD EQUATIONS

In the linearized Einstein theory it is assumed that, in a suitable coordinate system, the metric of the space–time can be written in the form

$$g_{\alpha\beta}=\eta_{\alpha\beta}+h_{\alpha\beta}, \tag{1}$$

where $h_{\alpha\beta}$, the metric perturbation, represents a small deviation of the metric $g_{\alpha\beta}$ from the Minkowski metric $(\eta_{\alpha\beta})=\mathrm{diag}(1,-1,-1,-1)$. The tensor field

$$K_{\alpha\beta\gamma\delta}\equiv\tfrac{1}{2}\{\partial_\alpha\partial_\gamma h_{\beta\delta}-\partial_\beta\partial_\gamma h_{\alpha\delta}+\partial_\beta\partial_\delta h_{\alpha\gamma}-\partial_\alpha\partial_\delta h_{\beta\gamma}\} \tag{2}$$

is the first-order contribution in $h_{\alpha\beta}$ to the curvature tensor corresponding to the metric $g_{\alpha\beta}$ and is invariant under the gauge transformations

$$h_{\alpha\beta}\mapsto h_{\alpha\beta}+\partial_\alpha\xi_\beta+\partial_\beta\xi_\alpha, \tag{3}$$

where ξ_α is an arbitrary vector field. From Eq. (2) it follows that $K_{\alpha\beta\gamma\delta}$ possesses the symmetries of the Riemann curvature tensor

$$K_{\alpha\beta\gamma\delta}=-K_{\beta\alpha\gamma\delta}=-K_{\alpha\beta\delta\gamma}=K_{\gamma\delta\alpha\beta}, \tag{4}$$

$$K_{\alpha\beta\gamma\delta}+K_{\alpha\delta\beta\gamma}+K_{\alpha\gamma\delta\beta}=0, \tag{5}$$

and

$$\partial_\alpha K_{\beta\gamma\delta\epsilon}+\partial_\beta K_{\gamma\alpha\delta\epsilon}+\partial_\gamma K_{\alpha\beta\delta\epsilon}=0. \tag{6}$$

Conversely, if $K_{\alpha\beta\gamma\delta}$ satisfies Eqs. (4)–(6), then there exists locally a symmetric tensor field $h_{\alpha\beta}$, defined up to the gauge transformations (3), such that Eq. (2) holds.

Defining the right dual $K^*_{\alpha\beta\gamma\delta}$ of $K_{\alpha\beta\gamma\delta}$ by means of

$$K^*_{\alpha\beta\gamma\delta}\equiv\tfrac{1}{2}K_{\alpha\beta}{}^{\rho\sigma}\epsilon_{\rho\sigma\gamma\delta}, \tag{7}$$

where $\epsilon_{\alpha\beta\gamma\delta}$ is completely antisymmetric with $\epsilon_{0123}=1$ and the indices are raised and lowered by means of $\eta^{\alpha\beta}$ and $\eta_{\alpha\beta}$, one finds that

$$K^*_{\alpha\beta\gamma\delta}=-K^*_{\beta\alpha\gamma\delta}=-K^*_{\alpha\beta\delta\gamma} \tag{8}$$

and

$$K^*_{\alpha\beta\gamma\delta}-K^*_{\gamma\delta\alpha\beta}=\tfrac{1}{2}(\epsilon_{\alpha\beta\delta\rho}K_\gamma{}^\rho+\epsilon_{\beta\alpha\gamma\rho}K_\delta{}^\rho+\epsilon_{\gamma\beta\delta\rho}K_\alpha{}^\rho+\epsilon_{\delta\alpha\gamma\rho}K_\beta{}^\rho), \tag{9}$$

where we have made use of the symmetric tensor field

$$K_{\alpha\beta}\equiv K_{\alpha\gamma\beta}{}^\gamma. \tag{10}$$

Moreover,

$$K^*_{\alpha\beta\gamma\delta}+K^*_{\alpha\delta\beta\gamma}+K^*_{\alpha\gamma\delta\beta}=-\epsilon_{\beta\gamma\delta\rho}K_\alpha{}^\rho \tag{11}$$

and

$$\partial_\alpha K^*_{\beta\gamma\delta\epsilon}+\partial_\beta K^*_{\gamma\alpha\delta\epsilon}+\partial_\gamma K^*_{\alpha\beta\delta\epsilon}=0. \tag{12}$$

Thus, $K^*_{\alpha\beta\gamma\delta}$ possesses all symmetries of $K_{\alpha\beta\gamma\delta}$ if and only if $K_{\alpha\beta}=0$, i.e., if the linearized Einstein vacuum field equations are satisfied.

Let $h_{\alpha\beta}$ be a real metric perturbation such that $K_{\alpha\beta}=0$. Then, the complex tensor field

$$P_{\alpha\beta\gamma\delta} \equiv K_{\alpha\beta\gamma\delta} - iK^{*}_{\alpha\beta\gamma\delta} \tag{13}$$

is self-dual,

$$P^{*}_{\alpha\beta\gamma\delta} = iP_{\alpha\beta\gamma\delta}, \tag{14}$$

and, according to the preceding paragraph, satisfies Eqs. (4)–(6). Hence, there exists a complex symmetric tensor field, $h^{c}_{\alpha\beta}$, such that

$$P_{\alpha\beta\gamma\delta} \equiv \tfrac{1}{2}\{\partial_\alpha\partial_\gamma h^{c}_{\beta\delta} - \partial_\beta\partial_\gamma h^{c}_{\alpha\delta} + \partial_\beta\partial_\delta h^{c}_{\alpha\gamma} - \partial_\alpha\partial_\delta h^{c}_{\beta\gamma}\} \tag{15}$$

and the original metric perturbation $h_{\alpha\beta}$ coincides with the real part of $h^{c}_{\alpha\beta}$ up to a gauge transformation (3).

Conversely, given a (complex) symmetric tensor field $h^{c}_{\alpha\beta}$ such that the tensor field (15) satisfies the self-duality condition [Eq. (14)], then the real and imaginary parts of $P_{\alpha\beta\gamma\delta}$, being linear combinations of $P_{\alpha\beta\gamma\delta}$ and its complex conjugate, satisfy Eqs. (4)–(6). If $K_{\alpha\beta\gamma\delta}$ denotes the real part of $P_{\alpha\beta\gamma\delta}$, then denoting the complex conjugation by a bar, from Eq. (14) we see that

$$K^{*}_{\alpha\beta\gamma\delta} = \frac{1}{2}(P^{*}_{\alpha\beta\gamma\delta} + \overline{P^{*}_{\alpha\beta\gamma\delta}}) = -\frac{1}{2i}(P_{\alpha\beta\gamma\delta} - \overline{P_{\alpha\beta\gamma\delta}}),$$

i.e., the right dual of the real part of $P_{\alpha\beta\gamma\delta}$ is minus the imaginary part of $P_{\alpha\beta\gamma\delta}$ and, therefore, $K_{\alpha\beta\gamma\delta}$ and $K^{*}_{\alpha\beta\gamma\delta}$ satisfy Eqs. (4)–(6), which implies that $K_{\alpha\beta} = 0$. Thus, the real part of $h^{c}_{\alpha\beta}$ satisfies the linearized Einstein field equations. The tensor $K_{\alpha\beta\gamma\delta}$ is related to the real part of $h^{c}_{\alpha\beta}$ according to Eq. (2).

III. INTEGRATION OF THE SELF-DUALITY CONDITIONS

In what follows, it will be convenient to make use of the spinor formalism. Letting

$$\partial_{AB'} \equiv \frac{\partial}{\partial x^{AB'}} \tag{16}$$

$(A,B,\ldots = 0,1;\ A',B',\ldots = 0',1')$ where the $x^{AB'}$ are (complex) coordinates such that the Minkowski metric is given by

$$ds^2 = \eta_{\alpha\beta}dx^\alpha dx^\beta = \epsilon_{AB}\epsilon_{C'D'}dx^{AC'}dx^{BD'}, \tag{17}$$

the spinor equivalent of Eq. (2) is given by

$$K_{AA'BB'}{}^{CC'DD'} = \tfrac{1}{2}\epsilon_{AB}\epsilon^{CD}\partial^{S}{}_{(A'}\partial^{R(C'}h_{|SR|B')}{}^{D')} + \tfrac{1}{2}\epsilon_{A'B'}\epsilon^{CD}\partial_{(A}{}^{S'}\partial^{R(C'}h_{B)RS'}{}^{D')}$$
$$+ \tfrac{1}{2}\epsilon_{AB}\epsilon^{C'D'}\partial^{S}{}_{(A'}\partial^{(C|R'|}h_{|S|}{}^{D)}{}_{B')R'} + \tfrac{1}{2}\epsilon_{A'B'}\epsilon^{C'D'}\partial_{(A}{}^{S'}\partial^{(C|R'|}h_{B)}{}^{D)}{}_{S'R'},$$

where $h_{ABA'B'} = h_{BAB'A'}$ is the spinor equivalent of $h_{\alpha\beta}$, the parentheses denote symmetrization on the indices enclosed, and the indices between bars are excluded from the symmetrization. Therefore, the self-duality conditions (14) amount to

$$\partial^{S}{}_{(A'}\partial^{(C|R'|}h_{|S|}{}^{D)}{}_{B')R'} = 0, \quad \partial_{(A}{}^{S'}\partial^{(C|R'|}h_{B)}{}^{D)}{}_{S'R'} = 0, \tag{18}$$

or, equivalently,

$$\partial^{S}{}_{(A'}H^{CD}{}_{|S|B')} = 0, \quad \partial_{(A}{}^{S'}H^{CD}{}_{B)S'} = 0 \tag{19}$$

with

$$H^{CD}{}_{AB'} \equiv \partial^{(C|R'|}h_A{}^{D)}{}_{B'R'} = \partial^{(C|R'|}h^{D)}{}_{AR'B'}\,. \tag{20}$$

Since the background metric is flat, Eqs. (19) are locally equivalent to the existence of a symmetric spinor field Λ_{AB} such that

$$H_{ABCD'} = \partial_{CD'}\Lambda_{AB}\,. \tag{21}$$

There always exists a (possibly complex) vector field $\xi_{AB'}$ such that

$$\Lambda_{AB} = -\partial_{(A}{}^{C'}\xi_{B)C'} \tag{22}$$

(actually, there are an infinite number of such vector fields). Then making use of Eqs. (20)–(22) one finds that under the gauge transformation $h_{ABA'B'} \mapsto h_{ABA'B'} + \partial_{AA'}\xi_{BB'} + \partial_{BB'}\xi_{AA'}$ [see Eq. (3)], $H_{ABCD'}$ is transformed into

$$\partial_{(A}{}^{R'}[h_{B)CR'D'} + \partial_{B)R'}\xi_{CD'} + \partial_{|CD'|}\xi_{B)R'}] = H_{ABCD'} + \partial_{CD'}\partial_{(A}{}^{R'}\xi_{B)R'} = H_{ABCD'} - \partial_{CD'}\Lambda_{AB} = 0. \tag{23}$$

Thus, by means of the gauge transformations (3), the self-duality conditions (18) can be reduced to the first-order differential equations

$$\partial_{(A}{}^{R'}h_{B)CR'D'} = 0. \tag{24}$$

Conditions (24) are similar to the (gauge-independent) equations for the vector potential of a self-dual electromagnetic field, $\nabla_{(A}{}^{R'}\Phi_{B)R'} = 0$ (see, e.g., Ref. 1), with an extra tensor index and, as in the case of the latter equations in a flat background, the solution of Eq. (24) is given, modulo gauge transformations, by a scalar potential that satisfies the wave equation. In order to prove this assertion, we start by considering Eq. (24) with $A = B = 0$, which yields $\partial_0{}^{R'}h_{0CR'D'} = 0$. These equations are locally equivalent to the existence of a set of functions, $M_{CD'}$, such that

$$h_{0CR'D'} = \partial_{0R'}M_{CD'}\,. \tag{25}$$

The symmetry of $h_{00R'D'}$ in the primed indices and Eq. (25) give $0 = h_{00}{}^{R'}{}_{R'} = \partial_0{}^{R'}M_{0R'}$, hence, there exists locally a function α such that

$$M_{0R'} = \partial_{0R'}\alpha. \tag{26}$$

Equation (24) with $A = 0$, $B = 1$, and $C = 0$ gives $\partial_0{}^{R'}h_{10R'D'} + \partial_1{}^{R'}h_{00R'D'} = 0$, or, making use of Eqs. (25) and (26), $0 = \partial_0{}^{R'}h_{10R'D'} + \partial_1{}^{R'}\partial_{0R'}\partial_{0D'}\alpha = \partial_0{}^{R'}(h_{10R'D'} - \partial_{1R'}\partial_{0D'}\alpha)$, which is locally equivalent to the existence of two functions $\mu_{A'}$ such that

$$h_{10R'D'} = \partial_{1R'}\partial_{0D'}\alpha + \partial_{0R'}\mu_{D'}\,. \tag{27}$$

Since $h_{10R'D'}$ must coincide with $h_{01D'R'}$, from Eqs. (25) and (27) it follows that $\partial_{1R'}\partial_{0D'}\alpha + \partial_{0R'}\mu_{D'} = \partial_{0D'}M_{1R'}$. Applying $\partial_0{}^{D'}$ to both sides of the last equation one finds that $\partial_{0R'}\partial_0{}^{D'}\mu_{D'} = 0$; hence,

$$\partial_0{}^{D'}\mu_{D'} = f(x^{1A'}), \tag{28}$$

where f is some function of two variables. By expressing Eq. (28) in the form $\partial_{0D'}(\mu^{D'} + \frac{1}{2}f(x^{1A'})x^{0D'}) = 0$, one finds that there exists locally a function β such that $\mu^{D'} + \frac{1}{2}f(x^{1A'})x^{0D'} = \partial_0{}^{D'}\beta$ and substituting this expression into Eq. (27) we obtain

$$h_{10R'D'} = \partial_{1R'}\partial_{0D'}\alpha + \partial_{0R'}(\partial_{0D'}\beta - \tfrac{1}{2}f(x^{1A'})x^0{}_{D'}). \tag{29}$$

Equation (24) with $A=0$ and $B=1=C$ yields $0=\partial_0{}^{R'}h_{11R'D'}+\partial_1{}^{R'}h_{01R'D'}$ or, equivalently, making use of Eq. (29), $0=\partial_0{}^{R'}h_{11R'D'}+\partial_1{}^{R'}h_{10D'R'}=\partial_0{}^{R'}(h_{11R'D'}-\partial_{1R'}\partial_{1D'}\alpha-\partial_{1R'}\partial_{0D'}\beta-\frac{1}{2}\partial_{1R'}f(x^{1A'})x^0{}_{D'})$, hence

$$h_{11R'D'}=\partial_{1R'}\partial_{1D'}\alpha+\partial_{1R'}\partial_{0D'}\beta+\tfrac{1}{2}\partial_{1R'}f(x^{1A'})x^0{}_{D'}+\partial_{0R'}\nu_{D'}\,, \tag{30}$$

where the $\nu_{D'}$ are some functions. From the symmetry of $h_{11R'D'}$ on the primed indices we obtain the condition $0=h_{11R'}{}^{R'}=\partial_{0R'}(\nu^{R'}-\partial_1{}^{R'}\beta+\frac{1}{6}\partial_{1S'}f(x^{1A'})x^{0S'}x^{0R'})$, therefore

$$\nu_{R'}=\partial_{1R'}\beta-\tfrac{1}{6}\partial_{1S'}f(x^{1A'})x^{0S'}x^0{}_{R'}+\partial_{0R'}\gamma, \tag{31}$$

where γ is some function. Substituting Eq. (31) into Eq. (30) we find that

$$\begin{aligned} h_{11R'D'}&=\partial_{1R'}\partial_{1D'}\alpha+\partial_{1R'}\partial_{0D'}\beta+\partial_{0R'}\partial_{1D'}\beta+\partial_{0R'}\partial_{0D'}\gamma\\ &\quad+\tfrac{1}{6}\partial_{1R'}f(x^{1A'})x^0{}_{D'}+\tfrac{1}{6}\partial_{1D'}f(x^{1A'})x^0{}_{R'}\,. \end{aligned} \tag{32}$$

Considering now Eq. (24) with $A=B=1$ and $C=0$, using Eq. (29) we obtain $0=\partial_1{}^{R'}h_{10R'D'}=\partial_{0D'}\partial_1{}^{R'}\partial_{0R'}\beta-\frac{1}{2}\partial_{1D'}f(x^{1A'})$, which implies that

$$\partial_1{}^{R'}\partial_{0R'}\beta=\tfrac{1}{2}\partial_{1D'}f(x^{1A'})x^{0D'}+g(x^{1A'}), \tag{33}$$

where g is some function of two variables, and from Eq. (24) with $A=B=C=1$ and Eqs. (32) and (33) it follows that $0=\partial_1{}^{R'}h_{11R'D'}=\partial_{0D'}(\partial_1{}^{R'}\partial_{0R'}\gamma+\frac{1}{6}\partial_{1R'}\partial_{1S'}f(x^{1A'})x^{0R'}x^{0S'}+\partial_{1S'}g(x^{1A'})x^{0S'})$, which implies that

$$\partial_1{}^{R'}\partial_{0R'}\gamma=-\tfrac{1}{6}\partial_{1R'}\partial_{1S'}f(x^{1A'})x^{0R'}x^{0S'}-\partial_{1S'}g(x^{1A'})x^{0S'}-b(x^{1A'}), \tag{34}$$

where b is some function of two variables.

Making use of the definitions

$$\tilde{\xi}_{0A'}\equiv-\tfrac{1}{2}\partial_{0A'}\alpha-\tfrac{1}{2}f_{A'}\,,\quad \tilde{\xi}_{1A'}\equiv-\tfrac{1}{2}\partial_{1A'}\alpha-\partial_{0A'}\beta+\tfrac{1}{2}\partial_{1S'}f_{A'}x^{0S'}+g_{A'}\,, \tag{35}$$

where $f_{A'}$ and $g_{A'}$ are functions of $x^{1A'}$ only such that

$$\partial_1{}^{A'}f_{A'}=f,\quad \partial_1{}^{A'}g_{A'}=g, \tag{36}$$

from Eqs. (25), (26), (29), (32), (35), and (36) we find that

$$\begin{aligned} h_{00R'D'}&=-\partial_{0R'}\tilde{\xi}_{0D'}-\partial_{0D'}\tilde{\xi}_{0R'}\,,\\ h_{10R'D'}&=-\partial_{1R'}\tilde{\xi}_{0D'}-\partial_{0D'}\tilde{\xi}_{1R'}\,,\\ h_{11R'D'}&=-\partial_{1R'}\tilde{\xi}_{1D'}-\partial_{1D'}\tilde{\xi}_{1R'}+\partial_{0R'}\partial_{0D'}\gamma+\partial_{1R'}\partial_{1D'}f_{S'}x^{0S'}\\ &\quad-\tfrac{1}{3}\partial_{1R'}fx^0{}_{D'}-\tfrac{1}{3}\partial_{1D'}fx^0{}_{R'}+\partial_{1R'}g_{D'}+\partial_{1D'}g_{R'}\\ &=-\partial_{1R'}\tilde{\xi}_{1D'}-\partial_{1D'}\tilde{\xi}_{1R'}+\partial_{0R'}\partial_{0D'}\psi, \end{aligned} \tag{37}$$

where

$$\psi\equiv\gamma+\tfrac{1}{6}\partial_{1A'}\partial_{1B'}f_{C'}x^{0A'}x^{0B'}x^{0C'}+\partial_{1A'}g_{B'}x^{0A'}x^{0B'}+b_{A'}x^{0A'}, \tag{38}$$

and $b_{A'}$ are functions of $x^{1A'}$ only such that

$$\partial_1{}^{A'} b_{A'} = b. \tag{39}$$

Then, from Eqs. (34), (36), (38), and (39) it follows that the scalar potential ψ satisfies the wave equation

$$\partial_1{}^{R'} \partial_{0R'} \psi = 0. \tag{40}$$

Thus, Eqs. (37) show that, in effect, the solution of Eqs. (24) and, therefore, of the self-duality conditions (18) is locally given, modulo gauge transformations, by a single scalar potential that obeys the wave equation.

It may be noticed that the complex vector field (35) satisfies the conditions $\partial_{(A}{}^{C'} \tilde{\xi}_{B)C'} = 0$, which means that $\tilde{\xi}_{AB'}$ is the vector potential of a self-dual electromagnetic field and that the condition $H_{ABCD'} = 0$ is preserved under the gauge transformation given by $\tilde{\xi}_{AB'}$ [cf. Eqs. (22) and (23)]; in other words, the metric perturbation $h_{ABA'B'} + \partial_{AA'} \tilde{\xi}_{BB'} + \partial_{BB'} \tilde{\xi}_{AA'}$ also satisfies Eq. (24), as can be directly verified.

The preceding result can also be expressed in a covariant form. If o_A is any *constant* spinor, then the solution of the self-duality conditions (14) and (15) is locally given, modulo gauge transformations, by

$$h_{ABC'D'} = \nabla^C{}_{C'} \nabla^D{}_{D'} (\psi o_A o_B o_C o_D), \tag{41}$$

where ψ is a solution of the wave equation

$$\nabla_{AA'} \nabla^{AA'} \psi = 0, \tag{42}$$

and any real solution of the Einstein vacuum field equations linearized about the Minkowski metric is locally given, modulo gauge transformations, by the real part of the perturbation given by Eq. (41). The covariant expression for the metric perturbations of an algebraically special solution of the Einstein vacuum field equations given in Ref. 8 (which is equivalent to the expressions found in Refs. 7 and 3) reduces to Eq. (41) when the multiple principal spinor of the background conformal curvature is constant. As pointed out in the Introduction, the derivations presented in Refs. 3, 7, and 8 do not demonstrate that all the metric perturbations of the algebraically special vacuum space–times are given by the formula obtained in those works. If the constant spinor o_A in Eq. (41) corresponds to the null direction $\partial_t + \partial_z$, then the first-order curvature $K_{\alpha\beta\gamma\delta}$ corresponding to the real part of the perturbation (41) agrees with the expression obtained in Ref. 6 by direct integration in circular cylindrical coordinates of the equations for the curvature in the linearized Einstein theory.

From Eq. (41) it also follows that $\nabla_{(M}{}^{D'} h_{AB)C'D'} = \nabla_{(M}{}^{D'} \nabla^C{}_{C'} \nabla^D{}_{D'} (\psi o_A o_{B)} o_C o_D)$ is (the spinor equivalent of) a Lanczos potential for the first-order curvature (see, e.g., Ref. 11). The metric perturbation, the Lanczos potential, and the curvature are given by second, third, and fourth derivatives, respectively, of the scalar potential ψ.

As a simple example, we can consider a potential ψ independent of $x^{00'}$. Then Eq. (40) reduces to $\partial_{01'} \partial_{10'} \psi = 0$, or, taking $u \equiv x^{11'}$, $v \equiv x^{00'}$, $\zeta \equiv x^{01'}$, and $\overline{\zeta} \equiv x^{10'}$,

$$\partial_\zeta \partial_{\overline{\zeta}} \psi(u, \zeta, \overline{\zeta}) = 0. \tag{43}$$

The only nonvanishing component of $h_{11R'D'} = \partial_{0R'} \partial_{0D'} \psi$ [cf. Eq. (37)] is

$$h_{111'1'} = \partial_\zeta \partial_\zeta \psi \equiv F(u, \zeta, \overline{\zeta}). \tag{44}$$

By virtue of Eq. (43), $\partial_{\overline{\zeta}} F = 0$; therefore, $F = F(u, \zeta)$ and

$$(\eta_{\alpha\beta}+h^{c}_{\alpha\beta})dx^{\alpha}dx^{\beta}=2du\,dv-2d\zeta d\overline{\zeta}+h_{111'1'}du^{2}$$
$$=2du\,dv-2d\zeta d\overline{\zeta}+F(u,\zeta)du^{2} \tag{45}$$

is a (possibly complex) solution of the linearized Einstein vacuum field equations, where $F(u,\zeta)$ is an arbitrary function. The real part of the metric (45),

$$2du\,dv-2d\zeta d\overline{\zeta}+\tfrac{1}{2}[F(u,\zeta)+\overline{F(u,\zeta)}]du^{2}, \tag{46}$$

is also a solution of the linearized Einstein vacuum field equations, which turns out to be an exact solution of the Einstein vacuum field equations, as can be easily seen by applying the Xanthopoulos theorem,[12] taking into account that the metric (46) is of the form $(\eta_{\alpha\beta}+l_{\alpha}l_{\beta})dx^{\alpha}dx^{\beta}$, with l_{α} being a null vector field with respect to the metric $\eta_{\alpha\beta}$. The metric (46) corresponds to the well-known pp waves.

IV. CONCLUDING REMARKS

As we have shown, the Einstein vacuum field equations linearized about the Minkowski metric, which are a set of second-order partial differential equations for the metric perturbation, can be reduced to a set of first-order partial differential equations whose solution is locally determined, modulo gauge transformations, by a scalar potential. It would be desirable to have analogous results in the case of curved backgrounds, but then the perturbation of the curvature is no longer gauge-invariant and, therefore, the self-duality of the curvature perturbation has no gauge-invariant meaning. Moreover, when the background is not flat, the self-duality of the curvature perturbation is a very strong condition that may not be satisfied, modulo gauge transformations, by all the solutions of the linearized Einstein vacuum field equations (see, e.g., Ref. 13).

ACKNOWLEDGMENTS

One of the authors (HGSR) thanks the Departamento de Física Mathemática of the Universidad Autónoma de Puebla for its hospitality and the Academia Mexicana de Ciencias for the financial support. The authors would like to thank the referee for helpful comments.

[1] G. F. Torres del Castillo, Gen. Relativ. Gravit. **31**, 205 (1999).
[2] J. M. Cohen and L. S. Kegeles, Phys. Rev. D **10**, 1070 (1974).
[3] R. M. Wald, Phys. Rev. Lett. **41**, 203 (1978).
[4] G. F. Torres del Castillo, J. Math. Phys. **25**, 342 (1984).
[5] W. B. Campbell and T. Morgan, Physica **53**, 264 (1971).
[6] G. F. Torres del Castillo and J. M. Rojas-Marcial, Rev. Mex. Fís. **39**, 32 (1993).
[7] L. S. Kegeles and J. M. Cohen, Phys. Rev. D **19**, 1641 (1976).
[8] G. F. Torres del Castillo, J. Math. Phys. **27**, 1586 (1986).
[9] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 4th ed. (Pergamon, Oxford, 1975).
[10] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[11] F. Bampi and G. Caviglia, Gen. Relativ. Gravit. **15**, 375 (1983).
[12] B. C. Xanthopoulos, J. Math. Phys. **19**, 1607 (1978).
[13] R. M. Wald, Class. Quantum Grav. **3**, 55 (1986).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

The spectrum of the Liouville–von Neumann operator in the Hilbert–Schmidt space

I. Antoniou
International Solvay Institutes for Physics and Chemistry, C.P. 231, Campus Plaine ULB, Bd. du Triomphe, 1050 Brussels, Belgium and Theoretische Natuurkunde, Free University of Brussels and Moscow State University, Department of Mechanics and Mathematics, 119899 Moscow, Vorobjovy Gory, Russia

S. A. Shkarin
International Solvay Institutes for Physics and Chemistry, C.P. 231, Campus Plaine ULB, Bd. du Triomphe, 1050 Brussels, Belgium and Moscow State University, Department of Mechanics and Mathematics, 119899 Moscow, Vorobjovy Gory, Russia

Z. Suchanecki
International Solvay Institutes for Physics and Chemistry, C.P. 231, Campus Plaine ULB, Bd. du Triomphe, 1050 Brussels, Belgium and Theoretische Natuurkunde, Free University of Brussels and Hugo Steinhaus Center and Institute of Mathematics, Wroclaw Technical University, ul. Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

(Received 5 February 1999; accepted for publication 17 March 1999)

The singular continuous spectrum of the Liouville operator of quantum statistical physics is, in general, properly included in the difference of the spectral values of the singular continuous spectrum of the associated Hamiltonian. The absolutely continuous spectrum of the Liouvillian may arise from a purely singular continuous Hamiltonian. We provide the correct formulas for the spectrum of the Liouville operator and show that the decaying states of the singular continuous subspace of the Hamiltonian do not necessarily contribute to the absolutely continuous subspace of the Liouvillian.
[S0022-2488(99)00208-X]

I. INTRODUCTION

The Liouville operator is the starting point of both classical and quantum statistical physics,[1,2] since it generates the evolution of states ρ:

$$\partial_t \rho = -i\mathbb{L}\rho.$$

The formal relation between the Liouville operator $\mathbb{L}$ and its corresponding Hamiltonian H is given by the Poisson brackets $\mathbb{L}\rho = i\{H,\rho\}$ for classical systems and by the commutator $\mathbb{L}\rho = [H, \rho]$ for quantum systems. Here ρ denotes the density function or the density operator.

It is well known that the Liouville operator $\mathbb{L}$ (also called the Liouvillian) corresponding to an essentially self-adjoint Hamiltonian H is also essentially self-adjoint.[3] Moreover, the spectrum $\sigma(H)$ of H determines the spectrum $\sigma(\mathbb{L})$ of $\mathbb{L}$ which consists of the differences between the spectral values of H:

$$\sigma(\mathbb{L}) = \{\lambda - \lambda' : \lambda, \lambda' \in \sigma(H)\}. \tag{1}$$

It may therefore be natural to expect a similar relation to hold separately for each part of the spectrum of the Hamiltonian and the Liouvillian, namely for the point spectrum, the absolutely continuous spectrum, and the singular continuous spectrum:

$$\sigma_{\mathrm{p}}(\mathbb{L}) = \{\lambda - \lambda' : \lambda, \lambda' \in \sigma_{\mathrm{p}}(H)\} \tag{2}$$

$$\sigma_{ac}(\mathbb{L})=\{\lambda-\lambda':\lambda,\lambda'\in\sigma_{ac}(H)\} \tag{3}$$

$$\sigma_{sc}(\mathbb{L})=\{\lambda-\lambda':\lambda,\lambda'\in\sigma_{sc}(H)\}. \tag{4}$$

Twenty years ago, Spohn[4] argued that formulas (2)–(4) hold true. The objective of this paper is to demonstrate that formulas (3) and (4) are not true. Spohn's proof concerning the singular spectrum (4) is based on the argument that the set $\{\lambda-\lambda':\lambda,\lambda'\in\sigma_{sc}(H)\}$ has Lebesgue measure zero, because the set $\sigma_{sc}(H)$ has Lebesgue measure zero. This is, however, not true, as can be shown on a simple example of an operator with Cantor spectrum. Another example contradicting explicitly Spohn's assertion has been found by L. Bos and B. S. Pavlov[5] who constructed an example of a Hamiltonian with purely singular spectrum for which the spectrum of its Liouvillian has an absolutely continuous component. Bos and Pavlov's example has been generalized further by M. Gadella and two of us (IA and ZS).[6]

Let us remark here that the problem of finding a Hamiltonian with pure singular spectrum such that the corresponding Liouvillian has nonempty absolutely continuous spectrum reduces to the problem of finding a singular measure on the real line with absolutely continuous convolution square. The answer to this problem is known since 1938 when N. Wiener and A. Wintner[7] constructed a singular continuous measure μ for which the convolution $\mu*\mu$ is absolutely continuous. However, the Wiener–Wintner paper[7] as well as many other refinements and generalizations of their construction[8] remained unnoticed by the physics community.

We show that it is always possible to find a Hamiltonian with purely continuous singular spectrum concentrated on an arbitrary thin set, precisely on the set of Hausdorff dimension zero, for which the spectrum of the Liouvillian has nontrivial absolutely continuous component. In fact, the Liouvillian corresponding to a Hamiltonian with purely singular continuous spectrum may have even a purely absolutely continuous spectrum.

The interest of this discussion goes beyond the result of Spohn. A possible consequence is the nonequivalence between the Hilbert space and the Liouville space description of quantum mechanics hinted by I. Prigogine.[9] For example, time and entropy operators do exist as superoperators on the Liouville space,[10,11] while they cannot exist as operators on the Hilbert space of wave functions.[12] This may also have profound consequences from the point of view of scattering theory and resonance behavior that are under present investigation.

In Sec. II we provide basic notion, facts, and notations from the spectral theory of self-adjoint Hamiltonian operators in Hilbert space and the associated Liouville operators in the Hilbert–Schmidt space, and establish the relation between the spectral measure of the Liouville operator and the spectral measure of the corresponding Hamiltonian operator. In Sec. III we provide a counterexample which shows that the singular continuous spectrum of the Liouvillian is properly included in the differences of the spectral values of the singular continuous spectrum of the Hamiltonian and correct formula (4). In Sec. IV we show that the absolutely continuous spectrum of the Liouvillian may arise from a pure singular continuous Hamiltonian and correct formula (3). Finally, we show that the decaying states of the singular spectrum of the Hamiltonian do not necessarily contribute to the possible absolutely continuous subspace of the associated Liouvillian.

II. BASIC FACTS FROM THE SPECTRAL THEORY OF THE HAMILTONIAN AND LIOUVILLIAN OPERATORS

We collect here the basic facts and notions and define the notation we employ.

Let $\mathcal{H}$ be a Hilbert space and H the self-adjoint Hamiltonian operator on $\mathcal{H}$. For a given Hamiltonian H we define the corresponding Liouvillian $\mathbb{L}$ as the essentially self-adjoint operator $\mathbb{L}=H\otimes I-I\otimes H$ defined on the Hilbert tensor product space $\mathcal{H}\otimes\mathcal{H}^{\times}$, where $\mathcal{H}^{\times}$ denotes the dual space of $\mathcal{H}$ (usually identified with $\mathcal{H}$) and I is the identity operator on $\mathcal{H}$. Recall that the space $\mathcal{H}\otimes\mathcal{H}^{\times}$ is isometrically isomorphic with the space $\mathcal{B}^2_{\mathcal{H}}$ of all Hilbert–Schmidt operators on $\mathcal{H}$ with the scalar product $(A,B)=\mathrm{tr}(A^{\dagger}B)$. In the sequel we shall use Dirac's notation $|f\rangle\langle g|$ for the tensor product of elements $f\in\mathcal{H}$, $g\in\mathcal{H}^{\times}$.

Let us denote by $\{E_\lambda\}$ the spectral family of H, i.e., $H=\int_{-\infty}^{\infty}\lambda\, dE_\lambda$, and let $\{\mathbb{E}_\lambda\}$ denote the the spectral family of the Liouvillian. Then we have

$$\mathbb{E}_\lambda|f\rangle\langle g|=\int_{-\infty}^{\infty}|E_{\lambda+\lambda'}f\rangle d\langle E_{\lambda'}g|. \tag{5}$$

Denote by $\mathcal{H}_{\rm p}$ the closed linear hull of all eigenvectors of H. The continuous subspace of H is $\mathcal{H}_{\rm c}=\mathcal{H}\ominus\mathcal{H}_{\rm p}$. Recall that the singular continuous subspace $\mathcal{H}_{\rm sc}$ of $\mathcal{H}_{\rm c}$ consists of all $f\in\mathcal{H}_{\rm c}$ for which there exists a Borel set B_0 of Lebesgue measure zero such that $\int_{B_0}dE_\lambda f=f$. By $\mathcal{H}_{\rm ac}=\mathcal{H}_{\rm c}\ominus\mathcal{H}_{\rm sc}$ we shall denote the absolutely continuous subspace of $\mathcal{H}_{\rm c}$. Recall also that $\mathcal{H}_{\rm p}$, $\mathcal{H}_{\rm c}$, $\mathcal{H}_{\rm sc}$, and $\mathcal{H}_{\rm ac}$ are closed linear subspaces of $\mathcal{H}$ which reduce the operator H. The spectra of the corresponding reductions of H will be called respectively point, continuous, singular continuous, and absolutely continuous spectrum of H, and will be denoted by $\sigma_{\rm p}(H)$, $\sigma_{\rm c}(H)$, $\sigma_{\rm sc}(H)$, and $\sigma_{\rm ac}(H)$ correspondingly.[13]

Let $\mu=\mu_f$ denote, for a given $f\in\mathcal{H}$, the spectral measure on $\sigma(H)$ determined by the nondecreasing function $\lambda\mapsto\langle f,E_\lambda f\rangle$. Let $f=f_{\rm p}+f_{\rm sc}+f_{\rm ac}$ be the decomposition of f corresponding to the direct sum $\mathcal{H}_{\rm p}\oplus\mathcal{H}_{\rm sc}\oplus\mathcal{H}_{\rm ac}$. Putting $\mu_{\rm p}=\mu_{f_{\rm p}}$, $\mu_{\rm sc}=\mu_{f_{\rm sc}}$, and $\mu_{\rm ac}=\mu_{f_{\rm ac}}$ we obtain the Jordan decomposition of μ,

$$\mu=\mu_{\rm p}+\mu_{\rm sc}+\mu_{\rm ac}, \tag{6}$$

onto the point, singular continuous, and absolutely continuous component. Conversely, given any three finite Borel measures $\mu_{\rm p}$, $\mu_{\rm sc}$, and $\mu_{\rm ac}$, where $\mu_{\rm p}$ is concentrated on a countable set of points and the other two measures are respectively singular and absolutely continuous, one can always construct a Hilbert space $\mathcal{H}$ and a self-adjoint operator H such that these measures are spectral measures associated with some $f\in\mathcal{H}$. Moreover, the point, singular, and absolutely continuous spectrum of H coincide with $\mu_{\rm p}$, $\mu_{\rm sc}$, and $\mu_{\rm ac}$, respectively. This can be proved by taking $\mathcal{H}$ as the direct sum $L^2(\mathbf{R},\mu_{\rm p})\oplus L^2(\mathbf{R},\mu_{\rm sc})\oplus L^2(\mathbf{R},\mu_{\rm ac})$ and H as the operator of multiplication by λ.[13]

The spectral measure of a self-adjoint operator $H=\int_{\sigma(H)}\lambda\, dE_\lambda$ on a Hilbert space $\mathcal{H}$ corresponding to an element $h\in\mathcal{H}$ is

$$F(\lambda)=\langle h,E_\lambda h\rangle \quad \text{for } \lambda\in\mathbf{R}. \tag{7}$$

Since the Liouville operator corresponding to this self-adjoint operator H is defined on $\mathcal{H}\otimes\mathcal{H}^\times$ by $\mathbb{L}=H\otimes I-I\otimes H$, it follows from (5) that

$$(|h\rangle\langle h|,\mathbb{E}_\lambda|h\rangle\langle h|)=\int_{-\infty}^{\infty}\langle h,E_{\lambda+\lambda'}h\rangle d\langle h,E_{\lambda'}h\rangle. \tag{8}$$

Formula (8) can be written as

$$\mathbb{F}(\lambda)=\int_{-\infty}^{\infty}F(\lambda+\lambda')dF(\lambda'), \tag{9}$$

where

$$\mathbb{F}(\lambda)=(|h\rangle\langle h|,\mathbb{E}_\lambda|h\rangle\langle h|) \tag{10}$$

is the distribution function associated with the Liouville operator $\mathbb{L}$ for the operator $|h\rangle\langle h|$.

The relation between the spectral measure $\mu_{\mathbb{L}}$ of the Liouville operator and the spectral measure of the Hamiltonian operator H is given by the following.

Lemma 1: The measure μ_{L} *associated to the distribution function (10) of the Liouville operator is the convolution of the measure* μ *associated to the distribution function (7) of the Hamiltonian operator with the reflection* $\overline{\mu}$ *of* μ *with respect to 0, i.e.,*

$$\mu_{\mathrm{L}} = \mu * \overline{\mu},$$

where $\overline{\mu}(\Delta) = \mu(-\Delta) = \mu(\{-x : x \in \Delta\})$, *for each Borel set* Δ.

We remind the reader that the convolution $\mu_1 * \mu_2$ of the Borel measure μ_1 with the Borel measure μ_2 is defined as the composition of the product $\mu_1 \times \mu_2$ with the measurable map $s:\mathbf{R} \times \mathbf{R} \to \mathbf{R}$ given by

$$s(x,y) = x + y, \tag{11}$$

i.e., $(\mu_1 * \mu_2)(\Delta) = \mu_1 \times \mu_2(s^{-1}\Delta)$, for every Borel subset Δ of the real line.

Proof: Let us denote by $\mu_1 \overline{*} \mu_2$ (μ_1 and μ_2 are Borel measures on $\mathbf{R}$) the composition of the measure $\mu_1 \times \mu_2$ with the measurable map $d:\mathbf{R} \times \mathbf{R} \to \mathbf{R}$ given by

$$d(x,y) = x - y. \tag{12}$$

Since we have

$$\mu_{\mathrm{L}} = \mu \overline{*} \mu,$$

the assertion of the lemma follows from a more general property

$$\mu_1 \overline{*} \mu_2 = \mu_1 * \overline{\mu_2}. \tag{13}$$

To prove (13) let us define (measurable) transformation $t:\mathbf{R} \times \mathbf{R} \to \mathbf{R} \times \mathbf{R}$:

$$t(x,y) = (x, -y). \tag{14}$$

We want to show first that

$$\mu_1 \times \overline{\mu_2} = (\mu_1 \times \mu_2) \circ t. \tag{15}$$

For that, we need only to show this identity for sets of the form $A \times B$, where A and B are measurable sets. Indeed, we have

$$\begin{aligned} ((\mu_1 \times \mu_2) \circ t)(A \times B) &= (\mu_1 \times \mu_2)[t^{-1}(A \times B)] \\ &= (\mu_1 \times \mu_2)(A \times (-B)) \\ &= \mu_1(A)\mu_2(-B) \\ &= \mu_1(A)\overline{\mu_2}(B) = (\mu_1 \times \overline{\mu_2})(A \times B), \end{aligned} \tag{16}$$

which proves (15). Therefore, taking any measurable set C we have

$$\begin{aligned} \mu_1 \overline{*} \mu_2(C) &= (\mu_1 \times \mu_2) d^{-1}(C) \\ &= \int \int_{\{(x,y): x - y \in C\}} d\mu_1 \times \mu_2 \\ &= \int \int_{t^{-1}\{(x,y): x + y \in C\}} d\mu_1 \times \mu_2 \\ &= \int \int_{\{(x,y): x + y \in C\}} d\mu_1 \times \overline{\mu_2} = \mu_1 \times \overline{\mu_2}(s^{-1}(C)) = \mu_1 * \overline{\mu_2}(C). \end{aligned} \tag{17}$$

■

III. THE SINGULAR CONTINUOUS SPECTRUM OF THE LIOUVILLIAN IS PROPERLY INCLUDED IN THE DIFFERENCES OF THE SPECTRAL VALUES OF THE SINGULAR CONTINUOUS SPECTRUM OF THE HAMILTONIAN

We provide a counterexample to formula (4) for the singular continuous spectrum of the Liouville operator, which shows that the singular continuous spectrum $\sigma_{sc}(\mathbb{L})$ of the Liouville operator is properly included in the differences of the spectral values of a Hamiltonian with purely singular continuous spectrum.

Consider a Hamiltonian operator H for which $\sigma(H)$ is a Cantor set. The existence of such operators is guaranteed by the above construction of the Hamiltonian with given spectral measure. Let us note, however, that there are more ''physical'' examples of Hamiltonians with Cantor-like spectrum. For example, there is an absolutely summable sequence $\{a_n\}$ such that the spectrum of the Hamiltonian

$$H=-\frac{d^2}{dx^2}+\sum_{n=0}^{\infty} a_n \cos(x2^{-n}) \tag{18}$$

is the Cantor set (see Ref. 14 and references therein). Suppose that $\sigma(H)=C$, where C is the Cantor set on the interval [0, 1]. According to the general property of spectra of functions of the Hamiltonian (see Ref. 3) we have (1).

However, according to a theorem of Steinhaus,[15] we have

$$\{\lambda-\lambda':\lambda,\lambda'\in C\}=[-1\ 1], \tag{19}$$

as $[-1, 1]$ is a set of nonzero Lebesgue measure. Formula (19) contradicts (4). This property of the Cantor set does not, however, imply that the spectrum of the Liouvillian is absolutely continuous. To show this let us consider the well-known ''devil's staircase'' distribution function $F(\lambda)$ on the Cantor set C.[16] We remind the reader that $F(\lambda)$ has constant value equal $k/2^n$ on each interval which is removed in the nth step of the iterative construction of the Cantor set[16] (k are such that the fraction $k/2^n$ is nonreducible). [The ''devil's staircase'' is defined as follows: it is $\frac{1}{2}$ on the interval $(\frac{1}{3}, \frac{2}{3})$, $\frac{1}{4}$ on $(\frac{1}{9}, \frac{2}{9})$, $\frac{3}{4}$ on $(\frac{7}{9}, \frac{8}{9})$, $\frac{1}{8}$ on $(\frac{1}{27}, \frac{2}{27})$, $\frac{3}{8}$ on $(\frac{7}{27}, \frac{8}{27})$, $\frac{5}{8}$ on $(\frac{19}{27}, \frac{20}{27})$, $\frac{7}{8}$ on $(\frac{25}{27}, \frac{26}{27})$, and so on.] It is well known and not difficult to check that $F(\lambda)$ is a nondecreasing continuous function such that $F(\lambda)=0$ for $\lambda\leqslant 0$ and $F(\lambda)=1$ for $\lambda\geqslant 1$.

The Hamiltonian with spectral measure $dF(\lambda)$ is the multiplication operator

$$Hf(\lambda)=\lambda f(\lambda) \tag{20}$$

on the Hilbert space $L([0,1],dF)$. The spectral projectors E_λ of H are

$$H=\int_0^1 \lambda\, dE_\lambda,$$

$$E_\lambda f(\lambda')=\mathbb{1}_{[0,\lambda)} f(\lambda'),$$

where

$$\mathbb{1}_{[0,\lambda)}(\lambda')=\begin{cases}1, & 0\leqslant\lambda'<\lambda\\ 0, & \text{otherwise.}\end{cases}$$

The spectral measure dF corresponds to the cyclic vector $e=1$:

$$F(\lambda)=\langle e,E_\lambda e\rangle.$$

The spectral measure of the Liouvillian which corresponds to H (20) is (10):

$$\mathbb{F}(\lambda)=(|e\rangle\langle e|,\mathbb{E}_\lambda|e\rangle\langle e|).$$

From the general properties of the convolution and Lemma 1, the function $\mathbb{F}(\lambda)$ is continuous. We shall show that it is, however, not absolutely continuous. In order to show this, let us consider the Fourier–Stieltjes transform

$$\hat{F}(t)=\int_0^1 e^{it\lambda}\,dF(\lambda). \tag{21}$$

Let us find the explicit expression for $\hat{F}(t)$:

$$\begin{aligned}\hat{F}(t)=\int_0^1 e^{it\lambda}\,dF(\lambda)&=\frac{1}{2}\int_0^{1/3} e^{it\lambda}\,dF(3\lambda)+\frac{1}{2}\int_{2/3}^1 e^{it\lambda}\,dF(3\lambda-2)\\&=\frac{1}{2}\int_0^1 e^{it\lambda/3}\,dF(\lambda)+\frac{1}{2}\int_0^1 e^{it\lambda/3}e^{it2/3}\,dF(\lambda)\\&=\frac{1}{2}[1+e^{2it/3}]\hat{F}\left(\frac{t}{3}\right)\\&=\frac{1}{2}[1+e^{2it/3}]\frac{1}{2}[1+e^{2it/3^2}]\hat{F}\left(\frac{t}{3^2}\right)\\&=\left(\prod_{k=1}^{\infty}\frac{1}{2}[1+e^{2it/3}]\right)\hat{F}(0).\end{aligned} \tag{22}$$

In the first identity in the second row, we have used the fact that $F(\lambda)$ satisfies functional equation of the De Rham type (compare with Ref. 17):

$$F(\lambda)=\begin{cases}\frac{1}{2}F(3\lambda), & 0\leqslant\lambda<\frac{1}{3},\\ \frac{1}{2}, & \frac{1}{3}\leqslant\lambda<\frac{2}{3},\\ \frac{1}{2}F(3\lambda-2)+\frac{1}{2}, & \frac{2}{3}\leqslant\lambda\leqslant 1.\end{cases} \tag{23}$$

Since $\hat{F}(0)=1$, we obtain

$$\hat{F}(t)=\prod_{k=1}^{\infty}[1+e^{2it/3}]. \tag{24}$$

Choosing $t_n=\pi 3^n$ we see that

$$\hat{F}(t_n)=\prod_{k=n+1}^{\infty}\frac{1}{2}[1+e^{2\pi i/3^{k-n}}]=\prod_{k=1}^{\infty}\frac{1}{2}[1+e^{2\pi i/3^k}] \tag{25}$$

does not depend on n. Moreover, $\hat{F}(t_n)$ is nonzero. Indeed, taking n_0 such that $\pi/3^n<1/2^n$, for $n>n_0$, and using the equality $\Pi_{n=1}^{\infty}\cos(t/2^n)=(\sin t)/t$ we have

$$
\begin{aligned}
|\hat{F}(t_n)| &= \prod_{k=1}^{\infty} \cos\frac{\pi}{3^k} > \prod_{k=1}^{n_0} \cos\frac{\pi}{3^k} \prod_{k=n_0+1}^{\infty} \cos\frac{1}{2^k} \\
&= \prod_{k=1}^{n_0} \cos\frac{\pi}{3^k} \prod_{k=1}^{\infty} \cos\frac{2^{-n_0}}{2^k} \\
&= 2^{n_0} \sin\frac{1}{2^{n_0}} \prod_{k=1}^{n_0} \cos\frac{\pi}{3^k} > 0.
\end{aligned} \tag{26}
$$

This means that the Fourier transform $\hat{F}(t)$ does not converge to zero as $t\to\infty$. The same is true for the Fourier transform $\hat{\mathbb{F}}(t)$ of $\mathbb{F}(\lambda)$ because $|\hat{\mathbb{F}}(t)| = |\hat{F}(t)|^2$. Therefore, in view of the Riemann–Lebesgue lemma, $F(\lambda)$ as well as $\mathbb{F}(\lambda)$ cannot be purely absolutely continuous. In fact, one can show, moreover, that $\mathbb{F}(\lambda)$ is purely singular continuous. The proof of this fact, which has been communicated to us by Professor O. G. Smolyanov, is given in the Appendix. Therefore, the singular continuous spectrum has Lebesgue measure zero.

Therefore, the singular continuous spectrum $\sigma_{\mathrm{sc}}(\mathbb{L})$ of the Liouville operator being a set with Lebesgue measure zero is properly included in the set of differences

$$\sigma_{\mathrm{sc}}(H) - \sigma_{\mathrm{sc}}(H) = [-1, 1].$$

Therefore formula (4) for the singular continuous spectrum of $\mathbb{L}$ should be corrected to

$$\sigma_{\mathrm{sc}}(\mathbb{L}) \subset \{\sigma_{\mathrm{sc}}(H) - \sigma_{\mathrm{sc}}(H)\}. \tag{4'}$$

IV. THE ABSOLUTELY CONTINUOUS SPECTRUM OF THE LIOUVILLIAN MAY ARISE FROM A PURE SINGULAR CONTINUOUS HAMILTONIAN

Using the results from Ref. 18 on convolutions of singular measures we shall show that it is possible to find Hamiltonians with purely singular spectral measures concentrated on a set of the Hausdorff dimension zero. Therefore, the associated Liouvillian has nontrivial absolutely continuous subspace.

Let us recall that the p-dimensional Hausdorff measure, $p \geqslant 0$, on subsets E of $\mathbf{R}$ is defined to be the number

$$\chi_p(E) = \lim_{\varepsilon\to 0} \inf\left\{ \sum_{n=1}^{\infty} |I_n|^p : E \subset \bigcup_{n=1}^{\infty} I_n, \ |I_n| < \varepsilon, \quad n = 1,2,... \right\}, \tag{27}$$

where the infimum is taken over the coverings of E by intervals of the length not greater than ε ($|I_n|$ denotes the length of the interval I_n).

An important property of the Hausdorff measure is that for each set E there is a value p_0 such that $\chi_p(E) = \infty$ for $p < p_0$ and $\chi_p(E) = 0$ for $p > p_0$. This value p_0 is called the *Hausdorff dimension* of the set E.

The Hausdorff dimension of any open subset of $\mathbf{R}$ is equal 1. Hausdorff dimensions smaller than one characterize fractal sets. For example, the Hausdorff dimension of the Cantor set is equal to $\log 2/\log 3$.

The key point in the further construction is a result on Ref. 18 which shows that one can construct two Cantor-type subsets E_1 and E_2 of the unit interval $I = [0, 1]$ such that

$$E_1 + E_2 = \{\lambda + \lambda' : \lambda \in E_1, \lambda' \in E_2\} = I \tag{28}$$

and

$$\chi_p(E_1) = \chi_p(E_2) = 0, \tag{29}$$

for any $p>0$. Moreover, the *natural* probability measures, i.e., constructed in the same way as the Cantor measure, μ_i, supported on E_i, $i=1,2$, have the property that the convolution $\mu_1*\mu_2$ is Lebesgue on I.

Let us take two such singular measures μ_1 and μ_2 on I and define

$$\mu=\mu_1+\overline{\mu_2},$$

which is a singular measure with support $E_1\cup E_2$.

We define, as before, a Hamiltonian with spectrum $E_1\cup E_2$ and μ as the spectral measure given by the multiplication operator $Hf(\lambda)=\lambda f(\lambda)$ on the Hilbert space $L^2(I,\mu)$. Its spectral resolution $H=\int_{\sigma(H)}\lambda\, dE_\lambda$ can be given explicitly if we define projectors E_λ as the operators of multiplication by the indicators $\mathbb{1}_{[0,\lambda]}$ {projectors on the spaces $L^2([0,\lambda],\mu)$}. Then $e\equiv 1$ is the corresponding cyclic vector and $\mu([0,\lambda])=\langle e,E_\lambda e\rangle$.

The Liouvillian corresponding to H has the spectral projectors $\mathbb{E}_\lambda$ and the measure generated by $(|e\rangle\langle e|,\mathbb{E}_\lambda|e\rangle\langle e|)$ is given by the convolution of μ and $\overline{\mu}$. We have

$$\mu*\overline{\mu}=(\mu_1+\overline{\mu_2})*\overline{(\mu_1+\overline{\mu_2})}=(\mu_1+\overline{\mu_2})*(\overline{\mu_1}+\mu_2)=\mu_1*\overline{\mu_2}+\mu_1*\mu_2+\overline{\mu_2}*\overline{\mu_1}+\overline{\mu_2}*\mu_2.$$

Here, $\mu_1*\mu_2$ is an absolutely continuous measure by the hypothesis. Because $\overline{\mu_2}*\overline{\mu_1}=\overline{\mu_1}*\overline{\mu_2}$ and $\overline{\mu_1},\overline{\mu_2}$ are reflections of μ_1,μ_2, then $\overline{\mu_2}*\overline{\mu_1}$ is also absolutely continuous. These two convolutions form the absolutely continuous part of the spectral measure of the Liouvillian.

In view of the result of the previous section, formula (3) for the absolutely continuous spectrum of $\mathbb{L}$ should be corrected to

$$\sigma_{\rm ac}(\mathbf{L})=\{\sigma_{\rm ac}(H)-\sigma_{\rm ac}(H)\}\cup\{[\sigma_{\rm sc}(H)-\sigma_{\rm sc}(H)]\setminus\sigma_{\rm sc}(\mathbb{L})\}. \qquad (3')$$

V. DECAYING STATES IN THE SINGULAR CONTINUOUS SUBSPACE OF THE HAMILTONIAN

The survival amplitude of a wave function ψ in the Hilbert space $\mathcal{H}$ is

$$\langle\psi,U_t\psi\rangle=\left\langle \psi,\int_{-\infty}^{\infty}e^{-i\lambda t}\,dE_\lambda\psi\right\rangle=\int_{-\infty}^{\infty}e^{-i\lambda t}\,d\|E_\lambda\psi\|^2=\hat{\mu}_\psi(t).$$

This is the Fourier transform $\hat{\mu}_\psi$ of the measure μ_ψ corresponding to the distribution function $\|E_\lambda\psi\|^2$. From the Riemann–Lebesgue lemma, if ψ is in the absolutely continuous subspace of H it decays:

$$\langle\psi,U_t\psi\rangle\to 0 \quad \text{as } t\to\infty.$$

This is well known in quantum mechanics. See, for example, Ref. 19.

There may exist, however, decaying states ψ, $\hat{\mu}_\psi\to 0$, as $t\to\infty$, in the singular continuous subspace of $\mathcal{H}$ as well. Do the decaying states in $\mathcal{H}_{\rm sc}$ give rise necessarily to absolutely continuous subspace of the associated Liouville operator? The following counterexample shows that there is no such necessity. We shall construct a Borel probability measure μ on $\mathbf{R}$ such that $\overline{\mu}=\mu$ (μ is symmetric with respect to zero), the measures μ and $\mu*\mu$ are singular continuous, and $\lim_{t\to\infty}\hat{\mu}(t)=0$, where $\hat{\mu}$ is the Fourier transform of μ.

Let μ_n be the Borel probability measure on $\mathbf{R}$ defined by the formula

$$\mu_n=\tfrac{1}{2}\delta_{(2/9)^n}+\tfrac{1}{2}\delta_{-(2/9)^n},$$

where δ_a denotes the normalized measure concentrated at the point a. The function $\cos(t(\frac{2}{9})^n)$ is the Fourier transform of the measure μ_n and the function

$$\mathcal{F}_N(t)=\prod_{n=0}^{N}\cos\left(t\left(\frac{2}{9}\right)^n\right)$$

is the Fourier transform of the convolution $\mu_0*\mu_1*\cdots*\mu_N$. Evidently, the sequence $\mathcal{F}_N(t)$ converges uniformly to the function

$$\mathcal{F}(t)=\prod_{n=0}^{\infty}\cos\left(t\left(\frac{2}{9}\right)^n\right) \tag{30}$$

on any segment in **R**. Therefore, the sequence $\mu_0*\mu_1*\cdots*\mu_N$ weakly converges to some Borel probability measure μ on **R** and the Fourier transform $\hat{\mu}$ of this measure coincides with $\mathcal{F}$.

Now we shall check that μ and $\mu*\mu$ are singular and that

$$\lim_{t\to\infty}\mathcal{F}(t)=0. \tag{31}$$

The continuity of the measures μ and $\mu*\mu$ follows from condition (31), which we shall prove below.

The support of the measure $\mu_0*\mu_1*\cdots*\mu_N$ is the set

$$\left\{\sum_{j=0}^{N}\frac{\varepsilon_j}{(9/2)^j}:\varepsilon_j\in\{0,1\}\right\}.$$

Hence, the compact set

$$K=\left\{\sum_{j=0}^{\infty}\frac{\varepsilon_j}{(\frac{9}{2})^j}:\varepsilon_j\in\{0,1\}\right\}$$

is the support of μ. From the last formula it is easy to see that for any $n\in\mathbf{Z}_+$ the set K is a subset of the union of 2^n segments of the length $\frac{18}{7}(\frac{2}{9})^n$ ($K\subset[-\frac{9}{7},\frac{9}{7}]$, $K\subset[-\frac{9}{7},-\frac{5}{7}]\cap[\frac{5}{7},\frac{9}{7}]$, $K\subset[-\frac{9}{7},-\frac{73}{63}]\cap[-\frac{53}{63},-\frac{5}{7}]\cap[\frac{5}{7},\frac{53}{63}]\cap[\frac{73}{63},\frac{9}{7}],\ldots$). Therefore, the set $K+K$ is contained in the union of 4^n segments of the length $\frac{36}{7}(\frac{2}{9})^n$. So, for any n, the Lebesgue measure of $K+K$ does not exceed $\frac{36}{7}(\frac{8}{9})^n$. Hence, the Lebesgue measure of $K+K$ is zero and, consequently, the measures μ and $\mu*\mu$ are singular ($K+K$ is the support of $\mu*\mu$). It remains to prove (31). This property was proved by Erdös[20] (see also Ref. 21) for a wide class of functions $\mathcal{F}(t)$ of the form (30), where the number $\frac{2}{9}$ can be any rational value which is not reciprocal of an integer.

Remark: The number 9 in the definitions of $\mathcal{F}$ and μ_n can be replaced by an arbitrary odd integer $n\geqslant 9$. The Hausdorff dimension of the support K of measure μ in this case is equal to $(\ln 2)/(\ln n)$ [the Hausdorff dimension of the support of the measure $\mu*\mu$ does not exceed $2(\ln 2)/(\ln n)$]. So there exists a measure with desired properties and with arbitrary small Hausdorff dimension of the support.

VI. CONCLUDING REMARKS

Our main result is to prove the correct relations between the spectra of the Liouville and Hamiltonian operator:

$$\sigma_{\mathrm{p}}(\mathbb{L})=\{\lambda-\lambda':\lambda,\lambda'\in\sigma_{\mathrm{p}}(H)\} \tag{2}$$

$$\sigma_{\mathrm{ac}}(\mathbb{L})=\{\sigma_{\mathrm{ac}}(H)-\sigma_{\mathrm{ac}}(H)\}\cup\{[\sigma_{\mathrm{sc}}(H)-\sigma_{\mathrm{sc}}(H)]\setminus\sigma_{\mathrm{sc}}(\mathbb{L})\}. \tag{3'}$$

$$\sigma_{\mathrm{sc}}(\mathbb{L})\subset[\sigma_{\mathrm{sc}}(H)-\sigma_{\mathrm{sc}}(H)]. \tag{4'}$$

The appearance of absolutely continuous spectrum of the Liouvillian from Hamiltonians with purely singular continuous spectra is very surprising. It shows that even in the conventional formulation the statistical description in Liouville space is not entirely equivalent to the wave function description in the original Hilbert space. This nonequivalence was mentioned to us by I. Prigogine several times.[9] In addition to the time and entropy operators[10,11] mentioned in the Introduction, our results show that it is always possible to construct Hamiltonians for which no scattering states may exist, because of the absence of absolutely continuous spectrum, while their corresponding Liouvillians may have scattering states because of the absolutely continuous spectrum, contrary to the general belief that scattering theory is equivalent in both Hilbert and Liouville space.[2]

These facts show that the Liouville space formulation of quantum mechanics contains more possibilities than the formulation in the Hilbert space of wave functions corresponding to a Hamiltonian operator. This is in fact true as the Liouville operator admits extensions beyond the Hilbert–Schmidt space which are nonreducible to wave function evolutions.[23]

ACKNOWLEDGMENTS

We want to thank Professor M. Gadella and Professor B. S. Pavlov for fruitful discussions and Professor I. Prigogine for his constant encouragement and support. We are grateful to Professor O. Smolyanov who kindly gave us the proof of the theorem in the Appendix. The support of the Belgian Government, Interuniversity Attraction Poles, and the European Commission Esprit Project No. 21042 CTIAC is acknowledged.

APPENDIX: SINGULARITY OF THE CONVOLUTION SQUARE OF THE CANTOR MEASURE

Theorem (O. G. Smolyanov): *Let μ be the Cantor measure on the interval $[0,1]$ and $\nu=\mu*\mu$. Then there exists a Borel set $A\subset[0,2]$ of the Lebesgue measure zero such that $\nu(A)=1$.*

Proof: Let $A_0=[0,1]$,

$$A_n=\tfrac{1}{3}A_{n-1}\cup(\{\tfrac{2}{3}\}+\tfrac{1}{3}A_{n-1}),\quad \text{for } n=1,2,...,$$

and put

$$f_n=(\tfrac{2}{3})^n \mathbb{1}_{A_n},\quad \text{for } n=0,1,2,... .$$

Denote by μ_n the measure with the density f_n, i.e.,

$$\mu_n(B)=\int_B f_n(x)\,dx,$$

where B is a Borel subset of $[0,1]$. It is easy to see that $g_n\overset{\text{df}}{=}f_n*f_n$ is the density of the convolution $\nu_n\overset{\text{df}}{=}\mu_n*\mu_n$.

The sequence $\{\mu_n\}$ converges weakly (see Ref. 22) to the Cantor measure μ and consequently $\nu_n\to\nu$ weakly.

To prove that ν is singular with respect to the Lebesgue measure it is enough to show that the sequence of densities $\{g_n\}$ converges to 0 in Lebesgue measure $|\cdot|$, i.e., for each $\varepsilon>0$,

$$|\{x:g_n(x)\geqslant\varepsilon\}|\to 0 \quad \text{as } n\to\infty.$$

Note that

$$g_0(x)=\begin{cases} x, & x\in[0,1],\\ 2-x, & x\in[1,2],\\ 0, & \text{otherwise,}\end{cases}$$

and

$$g_n(x)=\begin{cases} \tfrac{3}{4}g_{n-1}(3x), & x\in[0,\tfrac{2}{3}]\\ \tfrac{3}{2}g_{n-1}(3x-2), & x\in[\tfrac{2}{3},\tfrac{4}{3}]\\ \tfrac{3}{4}g_{n-1}(3x-4), & x\in[\tfrac{4}{3},2]\\ 0, & \text{otherwise.}\end{cases}$$

Observe that

$$|\{x:g_n(x)\geqslant\varepsilon\}|\leqslant 2\cdot 3^{-n}\,\mathrm{card}\{k:c_k^n\geqslant\varepsilon\},$$

where $c_k^n\overset{\mathrm{df}}{=}\max\{g_n(x):x\in[2(k-1)/3^n,2k/3^n]\}$, for $n=0,1,2,...,$ $k=1,...,3^n$. Therefore, in order to prove that $g_n\to 0$ in measure, it is sufficient to show that for each $r\in\mathbf{N}$

$$3^{-n}\,\mathrm{card}\{k:c_k^n\geqslant 2^{-r}\}\to 0 \quad \text{as} \quad n\to\infty.$$

Next observe that, for each n, c_k^n can only assume values $2^j(\tfrac{3}{4})^n$, $j=0,...,n$, and denote

$$\alpha_j^n=\mathrm{card}\{k:c_k^n=2^j(\tfrac{3}{4})^n\}.$$

Then

$$\mathrm{card}\{k:c_k^n\geqslant 2^{-r}\}=\sum_{\{j:2^j(\frac{3}{4})^n\geqslant 2^{-r}\}}\alpha_j^n.$$

Note that

$$2^j(\tfrac{3}{4})^n\geqslant 2^{-r}\Leftrightarrow j/n\geqslant 2-\log_2 3-r/n.$$

Since $2-\log_2 3>\tfrac{2}{5}$, then for each r there is $n_0(r)$ such that $2-\log_2 3-r/n>\tfrac{2}{5}$, for each $n>n_0(r)$. Therefore

$$\mathrm{card}\{k:c_k^n\geqslant 2^{-r}\}\leqslant\sum_{j>2n/5}\alpha_j^n,$$

for $n>n_0(r)$. Consequently, we need only to prove that

$$3^{-n}\sum_{j>2n/5}\alpha_j^n\to 0 \quad \text{as} \quad n\to\infty. \tag{A1}$$

Notice first that α_j^n are the coefficients in the binomial expansion of

$$\psi_n(x)\overset{\mathrm{df}}{=}(2+x)^n=\alpha_0^n+\alpha_1^n x+\cdots+\alpha_n^n x^n, \tag{A2}$$

i.e., $\alpha_j^m=2^{n-j}n!/j!(n-j)!$, which can be proved by induction. Indeed, (A2) is true for $n=0$. Now, assume that (A2) is true for some n. It is obvious that $\alpha_{n+1}^{n+1}=1$ and $\alpha_0^{n+1}=2\alpha_0^n=2^{n+1}$. If $j\in\{1,...,n\}$, then

$$\alpha_j^{n+1}=2\alpha_j^n+\alpha_{j-1}^n=2\cdot 2^{n-j}\frac{n!}{j!(n-j)!}+2^{n-j+1}\frac{n!}{(j-1)!(n-j+1)!}$$
$$=2^{n+1-j}\frac{(n+1)!}{j!(n+1-j)!},$$

which implies that (A2) is also true for $n+1$.

Let us now assume that $n/5\in\mathbf{N}$ and evaluate (A1) using the Cauchy form of the Taylor formula

$$3^{-n}\sum_{j>2n/5}\alpha_j^n=3^{-n}\frac{\psi_n^{(2n/5)}(\theta)(1-\theta)^{2n/5-1}}{(2n/5-1)!}$$
$$\leqslant 3^{-n}\frac{n!(2+\theta)^{3n/5}(1-\theta)^{2n/5-1}}{(3n/5)!(2n/5-1)!}$$
$$\leqslant 3^{-n}\frac{n!2^{3n/5}}{(3n/5)!(2n/5-1)!},$$

where $0<\theta<1$. Hence, using Stirling formula we get

$$3^{-n}\sum_{j>2n/5}\alpha_j^n=\frac{2n/5\sqrt{2\pi n}}{\sqrt{2\pi 3n/5}\sqrt{2n/5}}3^{-n}\frac{n^n2^{3n/5}}{(3n/5)^{3n/5}(2n/5)^{2n/5}}a(n)$$
$$=c(n)\left(3^{-1}\frac{2^{3/5}}{(3/5)^{3/5}(2/5)^{2/5}}\right)^n<c(n)p^n,$$

where $a(n)\to 1$ as $n\to\infty$, $c(n)\leqslant\text{const}\sqrt{n}$, and $p<1$. Since the right-hand side goes to 0, as $n\to\infty$, this ends the proof of the theorem.

[1] I. Prigogine, *Non-Equilibrium Statistical Mechanics* (Wiley, New York, 1962).
[2] E. E. Prugovecki, *Quantum Mechanics in Hilbert Space* (Academic, New York, 1981).
[3] M. Reed and B. Simon, *Methods of Modern Mathematical Physics Vol. I: Functional Analysis* (Academic, New York, 1980).
[4] H. Spohn, ''The spectrum of the Liouville-von Neumann operator,'' J. Math. Phys. **17**, 57–60 (1976).
[5] L. Bos and B. S. Pavlov, ''Absolute Continuity of Convolution of Singular Measures and Localization Problems,'' in *Chapman and Hill/CRC Research Notes in Mathematics Vol. 399: ''Generalized Functions, Operator Theory and Dynamical Systems,''* edited by I. Antoniou and G. Lumer (Chapman and Hill, London, 1999).
[6] I. Antoniou, M. Gadella, and Z. Suchanecki, ''General properties of the Liouville operator,'' Int. J. Theor. Phys. **37**, 1641–1654 (1998).
[7] N. Wiener and A. Wintner, ''Fourier-Stieltjes transforms and singular infinite convolution,'' Am. J. Math. **60**, 513–522 (1938).
[8] C. Karanikas and S. Koumandos, ''Continuous singular measures with absolutely continuous convolution squares on locally compact groups,'' Ill. J. Math. **35**, 490–495 (1991).
[9] I. Prigogine, *From Being to Becoming* (Freeman, New York, 1980); ''Dissipative processes in quantum theory,'' Phys. Rep. **219**, 93–120 (1992), and personal communication.
[10] M. Courbage, ''On necessary and sufficient conditions for the existence of time and entropy operators in quantum mechanics,'' Lett. Math. Phys. **4**, 425–432 (1980).
[11] B. Misra, I. Prigogine, and M. Courbage, ''Liapunov variable: entropy and measurements in quantum mechanics,'' Proc. Natl. Acad. Sci. USA **76**, 4768–4772 (1979).
[12] W. Pauli, *General Principles of Quantum Mechanics* (Springer-Verlag, New York, 1980).
[13] H. Weidmann, *Linear Operators in Hilbert Space* (Springer Verlag, New York, 1980).
[14] J. Bellisard, ''Schroedinger Operators with Almost Periodic Potential: an Overview,'' Lect. Notes Phys. **153**, 356–363 (1982).
[15] H. Steinhaus, ''A new property of the Cantor set,'' (in Polish) Wektor **7**, 1–3 (1917); English translation: H. Steinhaus, *Selected Papers* (PWN, Warsaw, 1985).
[16] P. Billingsley, *Probability and Measure* (Wiley, New York, 1985).
[17] S. Tasaki, Z. Suchanecki, and I. Antoniou, ''Ergodic properties of piecewise linear maps on fractal repellers,'' Phys. Lett. A **179**, 103–110 (1993).

[18] N. Levenberg, Jr., G. J. Martin, A. L. Shields, and S. Zdravkovska, ''Factorizations of Lebesgue measure via convolution,'' Proc. Am. Math. Soc. **104**, 419–430 (1988).
[19] P. Exner, *Open Quantum Systems and Feynman Integrals* (Reidel, The Netherlands, 1985).
[20] P. Erdös, ''On a family of symmetric Bernoulli convolutions,'' Am. J. Math. **61**, 974–976 (1939).
[21] A. Zygmund, *Trigonometric Series* (Cambridge, U.P., Cambridge, 1968), Vols. I, II.
[22] V. S. Varadarajan, ''Measures on Topological Spaces,'' Mat. Sb. **55**, 35–103 (1961).
[23] I. Antoniou and Y. Melnikov, ''Non-Factorizable Extensions of the Liouville Operator to Rigged Hilbert-Schmidt Space,'' in preparation.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Anomalies and analytic torsion on hyperbolic manifolds

A. A. Bytsenko,[a] A. E. Gonçalves,[b] and M. Simões[c]
Departamento de Fisica, Universidade Estadual de Londrina, Caixa Postal 6001, Londrina–Parana, Brazil

F. L. Williams[d]
Department of Mathematics, University of Massachusetts, Amherst, Massachusetts 01003

(Received 1 February 1999; accepted for publication 2 April 1999)

The global additive and multiplicative properties of the Laplacian on j-forms and related zeta functions are analyzed. The explicit form of zeta functions on a product of closed oriented hyperbolic manifolds $\Gamma\backslash\mathbb{H}^d$ and of the multiplicative anomaly are derived. We also calculate in an explicit form the analytic torsion associated with a connected sum of such manifolds. © *1999 American Institute of Physics.* [S0022-2488(99)01407-3]

I. INTRODUCTION

The additive and multiplicative properties of (pseudo-) differential operators as well as properties of their determinants have been studied actively during recent years in the mathematical and physical literature. The anomaly associated with product of regularized determinants of operators can be expressed by means of the noncommutative residue, the Wodzicki residue[1] (see also Refs. 2 and 3). The Wodzicki residue, which is the unique extension of the Dixmier trace to the wider class of (pseudo-) differential operators,[4,5] has been considered within the noncommutative geometrical approach to the standard model of the electroweak interactions[6–8] and the Yang–Mills action functional. Some recent papers along these lines can be found in Refs. 9–12.

The product of two (or more) differential operators of Laplace type can arise in higher derivative field theories (for example, in higher derivative quantum gravity[13]). The zeta function associated to the product of Laplace-type operators acting in irreducible rank 1 symmetric spaces and the explicit form of the multiplicative anomaly have been derived in Ref. 11.

Under such circumstances we should note that the conformal deformation of a metric and the corresponding conformal anomaly can also play an important role in quantum theories with higher derivatives. It is well known that evaluation of the conformal anomaly is actually possible only for even-dimensional spaces and up to now its computation is extremely involved. The general structure of such an anomaly in curved d-dimensional spaces (d even) has been studied in Ref. 14. We briefly mention here analysis related to this phenomenon for constant curvature spaces. The conformal anomaly calculation for the d-dimensional sphere can be found, for example, in Ref. 15. The explicit computation of the anomaly (of the stress-energy tensor) in irreducible rank 1 symmetric spaces has been carried out in Refs. 16–18 using the zeta-function regularization and the Selberg trace formula.

Recently the topology of manifolds have been studied by means of quantum field theory methods. New invariants related to three-manifolds[19] can be constructed within the framework of Chern–Simons gauge theory and can be specified in terms of the axioms of topological quantum field theory. Also, an important role is played by semiclassical approximations for Chern–Simons theory associated with partition functions involving quadratic functionals. It has been shown, in fact, that the analytic or Ray–Singer torsion (a topological invariant)[20] occurs within quantum

[a]Electronic mail: abyts@fisica.uel.br
[b]Electronic mail: goncalve@fisica.uel.br
[c]Electronic mail: simoes@npd.uel.br
[d]Electronic mail: williams@math.umass.edu

field theory as the partition function of a certain quadratic functional.[21,22] Recall that Ray–Singer torsion $T_{\rm an}(X)$ is defined for every closed Riemannian manifold X and orthogonal representation χ of $\pi_1(X)$. The definition of the torsion involves the spectrum of the Laplacian on twisted j-forms. It has been proved in Refs. 23 and 24 that when χ is acyclic and orthogonal the value $T_{\rm an}(X)$ coincides with the so-called Reidemeister torsion, which can be computed from a twisted cochain complex of a finite complex by taking a suitable alternating product of determinants.[25]

The purpose of the present paper is to investigate the spectral zeta functions associated with a product and Krönecker sum of Laplacians on j-forms and to calculate in an explicit form the analytic torsion on closed oriented hyperbolic manifolds $\Gamma\backslash\mathbb{H}^d$ and on a connected sum of such manifolds.

II. THE SPECTRAL ZETA FUNCTION AND THE TRACE FORMULA

We shall be working with irreducible rank 1 symmetric spaces $X=G/K$ of noncompact type. Thus G will be a connected noncompact simple split rank 1 Lie group with finite center and $K\subset G$ will be a maximal compact subgroup. Up to local isomorphism we choose $X=\mathrm{SO}_1(d,1)/\mathrm{SO}(d)$. Thus the isotropy group K of the base point $(1,0,...,0)$ is $\mathrm{SO}(d)$; X can be identified with hyperbolic d-space $\mathbb{H}^d$, $d=\dim X$. It is possible to view $\mathbb{H}^d$, for example, as one sheet of the hyperboloid of two sheets in $\mathbb{R}^{d+1}$ given by $q(x)=-x_0^2+x_1^2+\cdots+x_d^2=-1$, $x_0>0$, with the metric induced by the quadratic form $q(x)$. Let $\Gamma\subset G$ be a discrete, co-compact, torsion free subgroup, and let $\chi(\gamma)=\mathrm{trace}(\chi(\gamma))$ be the character of a finite-dimensional unitary representation χ of Γ for $\gamma\in\Gamma$. Let $L^{(j)}\equiv\Delta_\Gamma^{(j)}$ be the Laplacian on j-forms acting on the vector bundle $V(X_\Gamma)$ over $X_\Gamma=\Gamma\backslash G/K$ induced by χ. Note that the nontwisted j-forms on X_Γ are obtained by taking $\chi=1$. One can define the heat kernel of the elliptic operator $\mathcal{L}^{(j)}=L^{(j)}+b^{(j)}$ by

$$\mathrm{Tr}(e^{-t\mathcal{L}^{(j)}})=\frac{-1}{2\pi i}\mathrm{Tr}\int_{\mathcal{C}_0}e^{-zt}(z-\mathcal{L}^{(j)})^{-1}\,dz, \tag{2.1}$$

where $\mathcal{C}_0$ is an arc in the complex plane $\mathbb{C}$; the $b^{(j)}$ are endomorphisms of the vector bundle $V(X_\Gamma)$. By standard results in operator theory there exist $\varepsilon,\delta>0$ such that for $0<t<\delta$ the heat kernel expansion holds

$$\omega_\Gamma^{(j)}(t,b^{(j)})=\sum_{l=0}^{\infty}n_l(\chi)e^{-(\lambda_l^{(j)}+b^{(j)})t}=\sum_{0\leqslant l\leqslant l_0}a_l(\mathcal{L}^{(j)})t^{-l}+\mathcal{O}(t^{\varepsilon}), \tag{2.2}$$

where $\{\lambda_l^{(j)}\}_{l=0}^{\infty}$ is the set of eigenvalues of operator $L^{(j)}$ and $n_l(\chi)$ denote the multiplicity of $\lambda_l^{(j)}$. Eventually we would also like to take $b^{(j)}=0$, but for now we consider only nonzero modes: $b^{(j)}+\lambda_l^{(j)}>0$, $\forall l:\lambda_0^{(j)}=0$, $b^{(j)}>0$.

Let a_0,n_0 denote the Lie algebras of A, N in an Iwasawa decomposition $G=KAN$. Since the rank of G is 1, $\dim a_0=1$ by definition, say $a_0=\mathbb{R}H_0$ for a suitable basis vector H_0. One can normalize the choice of H_0 by $\beta(H_0)=1$, where $\beta:a_0\rightarrow\mathbb{R}$ is the positive root which defines n_0; for more detail see Ref. 26. Since Γ is torsion free, each $\gamma\in\Gamma-\{1\}$ can be represented uniquely as some power of a primitive element $\delta:\gamma=\delta^{j(\gamma)}$, where $j(\gamma)\geqslant 1$ is an integer and δ cannot be written as γ_1^j for $\gamma_1\in\Gamma$, $j>1$ an integer. Taking $\gamma\in\Gamma$, $\gamma\neq 1$, one can find $t_\gamma>0$ and $m_\gamma\in M\overset{\mathrm{def}}{=}\{m_\gamma\in K|m_\gamma a=am_\gamma,\forall a\in A\}$ such that γ is G conjugate to $m_\gamma\exp(t_\gamma H_0)$, namely for some $g\in G$, $g\gamma g^{-1}=m_\gamma\exp(t_\gamma H_0)$. Additionally, let $\chi_\sigma(m)=\mathrm{trace}(\sigma(m))$ be the character of σ, for σ a finite-dimensional representation of M.

Theorem 2.1: *(Fried's trace formula[27]) For* $0\leqslant j\leqslant d-1$,

$$\mathrm{Tr}(e^{-t\mathcal{L}^{(j)}})=I^{(j)}(t,b^{(j)})+I^{(j-1)}(t,b^{(j-1)})+H^{(j)}(t,b^{(j)})+H^{(j-1)}(t,b^{(j-1)}), \tag{2.3}$$

where

$$I^{(j)}(t,b^{(j)})\overset{\text{def}}{=}\frac{\chi(1)\mathrm{Vol}(\Gamma\backslash G)}{4\pi}\int_{\mathbb{R}}\mu_{\sigma_j}(r)e^{-t[r^2+b^{(j)}+(\rho_0-j)^2]}dr, \tag{2.4}$$

$$H^{(j)}(t,b^{(j)})\overset{\text{def}}{=}\frac{1}{\sqrt{4\pi t}}\sum_{\gamma\in C_\Gamma-\{1\}}\chi(\gamma)t_\gamma j(\gamma)^{-1}C(\gamma)\chi_{\sigma_j}(m_\gamma)\exp\left\{-\left[b^{(j)}t+(\rho_0-j)^2t+\frac{t_\gamma^2}{4t}\right]\right\}, \tag{2.5}$$

$\rho_0=(d-1)/2$, *and the function* $C(\gamma),\ \gamma\in\Gamma$, *defined on* $\Gamma-\{1\}$ *by*

$$C(\gamma)\overset{\text{def}}{=}e^{-\rho_0 t_\gamma}|\det_{n_0}(\mathrm{Ad}(m_\gamma e^{t_\gamma H_0})^{-1}-1)|^{-1}. \tag{2.6}$$

For Ad *denoting the adjoint representation of G on its complexified Lie algebra, one can compute* t_γ *as follows:*[28]

$$e^{t_\gamma}=\max\{|c|\,|\,c=\text{an eigenvalue of }\mathrm{Ad}(\gamma)\}. \tag{2.7}$$

Here C_Γ is a complete set of representatives in Γ of its conjugacy classes; Haar measure on G is suitably normalized. In our case $K\simeq\mathrm{SO}(d)$, $M\simeq\mathrm{SO}(d-1)$. For $j=0$ (i.e., for smooth functions or smooth vector bundle sections) the measure $\mu_0(r)$ corresponds to the trivial representation of M. For $j\geqslant 1$ there is a measure $\mu_\sigma(r)$ corresponding to a general irreducible representation σ of M. Let σ_j be the standard representation of $M=\mathrm{SO}(d-1)$ on $\Lambda^j\mathbb{C}^{(d-1)}$. If $d=2n$ is even, then $\sigma_j(0\leqslant j\leqslant d-1)$ is always irreducible; if $d=2n+1$, then every σ_j is irreducible except for $j=(d-1)/2=n$, in which case σ_n is the direct sum of two (1/2)-spin representations $\sigma^\pm:\sigma_n=\sigma^+\oplus\sigma^-$. For $j=n$ the representation τ_n of $K=\mathrm{SO}(2n)$ on $\Lambda^n\mathbb{C}^{2n}$ is not irreducible, $\tau_n=\tau_n^+\oplus\tau_n^-$ is the direct sum of (1/2)-spin representations. The Harish-Chandra Plancherel measures $\mu_{\sigma_j}(r)$ are given by the following theorem.

Theorem 2.2: *Let the group* $G=\mathrm{SO}_1(2n,1)$. *Then*

$$\mu_{\sigma_j}(r)=\binom{2n-1}{j}\frac{\pi r}{2^{4n-4}\Gamma(n)^2}\prod_{i=2}^{j+1}\left[r^2+\left(n+\frac{3}{2}-i\right)^2\right]\times\prod_{i=j+2}^{n}\left[r^2+\left(n+\frac{1}{2}-i\right)^2\right]\tanh(\pi r)\quad\text{for }0\leqslant j\leqslant n-1, \tag{2.8}$$

$$\mu_{\sigma_j}(r)=\binom{2n-1}{j}\frac{\pi r}{2^{4n-4}\Gamma(n)^2}\prod_{i=2}^{2n-j}\left[r^2+\left(n+\frac{3}{2}-i\right)^2\right]\times\prod_{i=2n-j+1}^{n}\left[r^2+\left(n+\frac{1}{2}-i\right)^2\right]\tanh(\pi r)\quad\text{for }n\leqslant j\leqslant 2n-1, \tag{2.9}$$

and $\mu_{\sigma_j}(r)=\mu_{\sigma_{2n-j-1}}(r)$.

For the group $G=\mathrm{SO}_1(2n+1,1)$ *one has*

$$\mu_{\sigma_j}(r)=\binom{2n}{j}\frac{\pi}{2^{4n-2}\Gamma\left(n+\frac{1}{2}\right)^2}\prod_{i=1}^{j+1}[r^2+(n+1-i)^2]\prod_{i=j+2}^{n}[r^2+(n-i)^2]\quad\text{for }0\leqslant j<n, \tag{2.10}$$

$$\mu_{\sigma_j}(r)=\binom{2n}{j}\frac{\pi}{2^{4n-2}\Gamma\left(n+\frac{1}{2}\right)^2}\prod_{i=1}^{2n-j+1}[r^2+(n+1-i)^2]$$

$$\times\prod_{i=2n-j+2}^{n}[r^2+(n-i)^2]\quad \text{for } n+1\leqslant j\leqslant 2n-1. \tag{2.11}$$

We should note that the reason for the pair of terms $\{I^{(j)},I^{(j-1)}\}$, $\{H^{(j)},H^{(j-1)}\}$ in the trace formula Eq. (2.3) is that τ_j satisfies $\tau_j|_M=\sigma_j\oplus\sigma_{j-1}$.

Finally, using the result of Theorem 2.2, we have

$$\mu_{\sigma_j}(r)=C^{(j)}(d)P(r,d)\times\begin{cases}\tanh(\pi r) & \text{for } d=2n,\\ 1 & \text{for } d=2n+1,\end{cases}$$

$$=C^{(j)}(d)\times\begin{cases}\displaystyle\sum_{l=0}^{d/2-1} a_{2l}^{(j)}(d)r^{2l+1}\tanh(\pi r) & \text{for } d=2n,\\ \displaystyle\sum_{l=0}^{(d-1)/2} a_{2l}^{(j)}(d)r^{2l} & \text{for } d=2n+1,\end{cases} \tag{2.12}$$

$$C^{(j)}(d)=\binom{d-1}{j}\frac{\pi}{2^{2d-4}\Gamma(d/2)^2}, \tag{2.13}$$

where the $P(r,d)$ are even polynomials [with suitable coefficients $a_{2l}^{(j)}(d)$] of degree $d-1$ for $G\neq \mathrm{SO}(2n+1,1)$, and of degree $d=2n+1$ for $G=\mathrm{SO}_1(2n+1,1)$.[29,26]

A. Case of the trivial representation

For $j=0$ we take $I^{(-1)}=H^{(-1)}=0$. Since σ_0 is the trivial representation $\chi_{\sigma_0}(m_\gamma)=1$. In this case, Fried's formula (2.3) reduces exactly to the trace formula for $j=0$:[28,30]

$$\omega_\Gamma^{(0)}(t,b^{(0)})=\frac{\chi(1)\mathrm{vol}(\Gamma\backslash G)}{4\pi}\int_{\mathbb{R}}\mu_{\sigma_0}(r)e^{-(r^2+b^{(0)}+\rho_0^2)t}dr+H^{(0)}(t,b^{(0)}), \tag{2.14}$$

where ρ_0 is associated with the positive restricted (real) roots of G (with multiplicity) with respect to a nilpotent factor N of G in an Iwasawa decomposition $G=KAN$. The function $H^{(0)}(t,b^{(0)})$ has the form

$$H^{(0)}(t,b^{(0)})=\frac{1}{\sqrt{4\pi t}}\sum_{\gamma\in C_\Gamma-\{1\}}\chi(\gamma)t_\gamma j(\gamma)^{-1}C(\gamma)e^{-[b^{(0)}t+\rho_0^2t+t_\gamma^2/(4t)]}. \tag{2.15}$$

B. Case of zero modes

It can be shown[31] that the Mellin transform of $H^{(0)}(t,0)$ ($b^{(0)}=0$, i.e., the zero modes case),

$$\mathfrak{h}^{(0)}(s)\overset{\text{def}}{=}\int_0^\infty H^{(0)}(t,0)t^{s-1}dt, \tag{2.16}$$

is a holomorphic function on the domain $\mathrm{Re}\,s<0$. Then, using the result of Refs. 29 and 26 one can obtain on $\mathrm{Re}\,s<0$

$$\mathfrak{h}^{(0)}(s)=\sum_{\gamma\in C_\Gamma-\{1\}}\chi(\gamma)t_\gamma j(\gamma)^{-1}C(\gamma)\int_0^\infty \frac{e^{-(\rho_0^2 t+t_\gamma^2/(4t))}}{\sqrt{4\pi t}}t^{s-1}\,dt$$
$$=\frac{(2\rho_0)^{1/2-s}}{\sqrt{\pi}}\sum_{\gamma\in C_\Gamma-\{1\}}\chi(\gamma)t_\gamma j(\gamma)^{-1}C(\gamma)t_\gamma^{s+1/2}K_{1/2-s}(t_\gamma\rho_0), \tag{2.17}$$

where $K_\nu(s)$ is the modified Bessel function, and finally

$$\mathfrak{h}^{(0)}(s)=\frac{\sin(\pi s)}{\pi}\Gamma(s)\int_0^\infty \psi_\Gamma(t+2\rho_0;\chi)(2\rho_0 t+t^2)^{-s}\,dt. \tag{2.18}$$

Here $\psi_\Gamma(s;\chi)\equiv d(\log Z_\Gamma(s;\chi))/ds$, and $Z_\Gamma(s;\chi)$ is a meromorphic suitably normalized Selberg zeta function.[32–37,30,38,29]

III. THE MULTIPLICATIVE ANOMALY

In this section the product of the operators on j-forms $\otimes\mathcal{L}_p^{(j)}$, $\mathcal{L}_p^{(j)}=L^{(j)}+b_p^{(j)}$, $p=1,2$, will be considered. We are interested in multiplicative properties of determinants, the multiplicative anomaly.[39,2,3] The multiplicative anomaly $F(\mathcal{L}_1^{(j)},\mathcal{L}_2^{(j)})$ reads

$$F(\mathcal{L}_1^{(j)},\mathcal{L}_2^{(j)})=\det{}_\zeta[\otimes_p\mathcal{L}_p^{(j)}][\det{}_\zeta(\mathcal{L}_1^{(j)})\det{}_\zeta(\mathcal{L}_2^{(j)})]^{-1}, \tag{3.1}$$

where we assume a zeta-regularization of determinants, i.e.,

$$\det{}_\zeta(\mathcal{L}_p^{(j)})\overset{\text{def}}{=}\exp\left(-\frac{\partial}{\partial s}\zeta(s|\mathcal{L}_p^{(j)})\bigg|_{s=0}\right). \tag{3.2}$$

Generally speaking, if the anomaly related to elliptic operators is nonvanishing, then the relation $\log\det(\otimes\mathcal{L}_p^{(j)})=\mathrm{Tr}\log(\otimes\mathcal{L}_p^{(j)})$ does not hold.

A. The zeta function of the product of Laplacians

The spectral zeta function associated with the product $\otimes\mathcal{L}_p^{(j)}$ has the form

$$\zeta(s|\otimes_p\mathcal{L}_p^{(j)})=\sum_{l\geqslant 0}n_l\prod_p^2(\lambda_l^{(j)}+b_p^{(j)})^{-s}. \tag{3.3}$$

We shall always assume that $b_1^{(j)}\neq b_2^{(j)}$, say $b_1^{(j)}>b_2^{(j)}$. If $b_1^{(j)}=b_2^{(j)}$, then $\zeta(s|\otimes\mathcal{L}_p^{(j)})=\zeta(2s|\mathcal{L}^{(j)})$ is a well-known function. For $b_1^{(j)},b_2^{(j)}\in\mathbb{R}$, set $b_+\overset{\text{def}}{=}(b_1^{(j)}+b_2^{(j)})/2$, $b_-\overset{\text{def}}{=}(b_1^{(j)}-b_2^{(j)})/2$, thus $b_1^{(j)}=b_++b_-$ and $b_2^{(j)}=b_+-b_-$.

Theorem 3.1:[11] *The spectral zeta function can be written as follows:*

$$\zeta(s|\otimes_p\mathcal{L}_p^{(j)})=(2b_-)^{1/2-s}\frac{\sqrt{\pi}}{\Gamma(s)}\int_0^\infty \omega_\Gamma^{(j)}(t,b_+)I_{s-1/2}(b_-t)dt, \tag{3.4}$$

where the integral converges absolutely for $\mathrm{Re}\,s>d/4$.

This formula is a main starting point to study the zeta function. It expresses $\zeta(s|\otimes\mathcal{L}_p^{(j)})$ in terms of the Bessel function $I_{s-1/2}(b_-t)$ and $\omega_\Gamma^{(j)}(t,b_+)$, where the trace formula applies to $\omega_\Gamma^{(j)}(t,b_+)$. Let $B_p(j)=(\rho_0(p)-j)^2+b_p^{(j)}$ and $A\overset{\text{def}}{=}\chi(1)\mathrm{vol}(\Gamma\backslash G)C^{(j)}(d)/4$.

Theorem 3.2: *For* $\mathrm{Re}\,s>d/4$ *the explicit meromorphic continuation holds:*

$$\zeta(s|\otimes_p \mathcal{L}_p^{(j)}) = A\sum_{l=0}^{d/2-1} [a_{2l}^{(j)}(d)(\mathcal{F}_l^{(j)}(s)-E_l^{(j)}(s)) + a_{2l}^{(j-1)}(d)(\mathcal{F}_l^{(j-1)}(s)-E_l^{(j-1)}(s))] + \mathcal{I}^{(j)}(s)+\mathcal{I}^{(j-1)}(s), \tag{3.5}$$

where

$$E_l^{(j)}(s) \stackrel{\text{def}}{=} 4\int_0^\infty \frac{dr\, r^{2j+1}}{1+e^{2\pi r}} \prod_p (r^2+B_p(j))^{-s}, \tag{3.6}$$

which is an entire function of s,

$$\mathcal{F}_l^{(j)}(s) \stackrel{\text{def}}{=} (B_1(j)B_2(j))^{-s} \frac{l!(2B_1(j)B_2(j)/(B_1(j)+B_2(j)))^{l+1}}{(2s-1)(2s-2)\cdots(2s-(l+1))} \times F\left(\frac{l+1}{2},\frac{l+2}{2};s+\frac{1}{2};\left(\frac{B_1(j)-B_2(j)}{B_1(j)+B_2(j)}\right)^2\right), \tag{3.7}$$

$$\mathcal{I}^{(j)}(s) \stackrel{\text{def}}{=} (2b_-)^{1/2-s}\frac{\sqrt{\pi}}{\Gamma(s)}\int_0^\infty H^{(j)}(t,b_+)I_{s-1/2}(b_-t)t^{s-1/2}\,dt, \tag{3.8}$$

and $F(\alpha,\beta;\gamma;z)$ is the hypergeometric function.

The goal now is to compute the zeta function and its derivative at $s=0$. Thus we have

$$\mathcal{F}_l^{(j)}(0) = \frac{(-1)^{l+1}}{l+1}\left(\frac{2B_1(j)}{B_1(j)+B_2(j)}\right)^{l+1} F\left(\frac{l+1}{2},\frac{l+2}{2};\frac{1}{2};\left(\frac{B_1(j)-B_2(j)}{B_1(j)+B_2(j)}\right)^2\right) = \frac{(-1)^{l+1}}{2(l+1)}\sum_p^2 B_p(j)^{l+1}, \tag{3.9}$$

$$E_l^{(j)}(0) = 4\int_0^\infty \frac{dr\, r^{2l+1}}{1+e^{2\pi r}} = \frac{(-1)^l}{l+1}(1-2^{-2l-1})\mathcal{B}_{2l+2}, \tag{3.10}$$

$$\mathcal{I}^{(j)}(0)=0, \tag{3.11}$$

where $\mathcal{B}_{2n}$ are the Bernoulli numbers.

Proposition 3.3: A preliminary form of the zeta function $\zeta(s|\otimes_p\mathcal{L}_p^{(j)})$ at $s=0$ is

$$\zeta(0|\otimes_p\mathcal{L}_p^{(j)}) = A\sum_{l=0}^{(d/2)-1}\frac{(-1)^{l+1}}{2(l+1)}\Bigg[\sum_p (a_{2l}^{(j)}(d)B_p(j)^{l+1} + a_{2l}^{(j-1)}(d)B_p(j-1)^{l+1}) + (2-2^{-2l})\mathcal{B}_{2l+2}(a_{2l}^{(j)}(d)+a_{2l}^{(j-1)}(d))\Bigg]. \tag{3.12}$$

Proposition 3.4: The derivative of the zeta function at $s=0$ has the form

$$\zeta'(0|\otimes_p\mathcal{L}_p^{(j)}) = A\sum_{l=0}^{(d/2)-1}\left[\sum_m^4 (a_{2l}^{(j)}(d)\mathcal{E}_m^{(j)} + a_{2l}^{(j-1)}(d)\mathcal{E}_m^{(j-1)})\right], \tag{3.13}$$

where

$$\mathcal{E}_1^{(j)} = l!(B_1(j)^{l+1} + B_2(j)^{l+1}) \sum_{k=0}^{l} \frac{(-1)^{k+1}}{k!(l-k)!(j+1-k)!}, \tag{3.14}$$

$$\mathcal{E}_2^{(j)} = B_2(j)^{l+1} \left(\frac{B_1(j)-B_2(j)}{2B_1(j)} \right) \frac{(-1)^l}{(l+1)!} \sum_{k=1}^{\infty} \frac{(l+k+1)!}{(k+1)!} \sigma_n \left(\frac{B_1(j)-B_2(j)}{B_1(j)} \right)^k, \tag{3.15}$$

$$\mathcal{E}_3^{(j)} = \log(B_1(j)B_2(j)) \frac{(-1)^l}{2(l+1)} (B_1(j)^{l+1} + B_2(j)^{l+1}) - 4 \int_0^\infty \frac{r^{2l+1} \log((r^2+B_1(j))/(r^2+B_2(j)))dr}{1+e^{2\pi r}}, \tag{3.16}$$

$$\mathcal{E}_4^{(j)} \equiv \frac{d}{ds} \mathcal{I}^{(j)}(s) \bigg|_{s=0} = \int_0^\infty [H^{(j)}(t,b_1^{(j)}) + H^{(j)}(t,b_2^{(j)})] t^{-1}\, dt, \tag{3.17}$$

and $\sigma_n \overset{\text{def}}{=} \Sigma_{k=1}^n k^{-1}$.

B. The residue formula and the multiplicative anomaly

The value of $F(\mathcal{L}_1,\mathcal{L}_2)$ can be expressed by means of the noncommutative Wodzicki residue.[1] Let $\mathcal{O}_p$, $p=1$, 2, be invertible elliptic (pseudo-) differential operators of real nonzero orders α and β such that $\alpha+\beta\neq 0$. Even if the zeta functions for operators $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_1 \otimes \mathcal{O}_2$ are well defined and if their principal symbols satisfy the Agmon–Nirenberg condition (with appropriate spectra cuts), one has in general that $F(\mathcal{O}_1,\mathcal{O}_2)\neq 1$. For such invertible elliptic operators the formula for the anomaly of commuting operators holds:

$$\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2) = \mathcal{A}(\mathcal{O}_2,\mathcal{O}_1) = \log(F(\mathcal{O}_1,\mathcal{O}_2)) = \frac{\text{res}[(\log(\mathcal{O}_1^\beta \otimes \mathcal{O}_2^{-\alpha}))^2]}{2\alpha\beta(\alpha+\beta)}. \tag{3.18}$$

More general formulas have been derived in Refs. 2 and 3. Furthermore, the anomaly can be iterated consistently. Indeed, using Eq. (3.18) we have

$$\begin{aligned}
\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2) &= \zeta'(0|\mathcal{O}_1\mathcal{O}_2) - \zeta'(0|\mathcal{O}_1) - \zeta'(0|\mathcal{O}_2),\\
\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\mathcal{O}_3) &= \zeta'(0|\overset{3}{\underset{j}{\otimes}}\mathcal{O}_j) - \sum_j^3 \zeta'(0|\mathcal{O}_j) - \mathcal{A}(\mathcal{O}_1,\mathcal{O}_2),\\
&\cdots\cdots\cdots\cdots \\
\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\ldots,\mathcal{O}_n) &= \zeta'(0|\overset{n}{\underset{j}{\otimes}}\mathcal{O}_j) - \sum_j^n \zeta'(0|\mathcal{O}_j) - \mathcal{A}(\mathcal{O}_1,\mathcal{O}_2)\\
&\quad - \mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\mathcal{O}_3)\ldots - \mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\ldots,\mathcal{O}_{n-1}).
\end{aligned} \tag{3.19}$$

In particular, for $n=2$ and $\mathcal{O}_p \equiv \mathcal{L}_p^{(j)}$ the anomaly is given by the following theorem.

Theorem 3.5: *The explicit formula for the multiplicative anomaly is*

$$\mathcal{A}(\mathcal{L}_1^{(j)},\mathcal{L}_2^{(j)}) = A \sum_{l=0}^{(d/2)-1} [\Omega_l^{(j)} + \Omega_l^{(j-1)}], \tag{3.20}$$

where

$$\Omega_l^{(j)} = \frac{a_{2l}^{(j)}(d)(-1)^l}{2}\left[\frac{l}{2}(B_1(j)-B_2(j))^2 B_2(j)^{l-1} + \frac{l(l-1)}{4}(B_1(j)-B_2(j))^3 B_2(j)^{l-2} + \sum_{p=3}^{l} \frac{l!}{(p+1)p!(l-p)!}\left(\frac{1}{p}+\frac{1}{p-1}+\sum_{q=1}^{p-2}\frac{1}{p-q-1}\right)(B_1(j)-B_2(j))^{p+1}B_2(j)^{l-p}\right]. \tag{3.21}$$

We note that for the four-dimensional space with $G=\mathrm{SO}_1(4,1)$, one derives from Theorem 3.5 the result

$$\mathcal{A}(\mathcal{L}_1^{(j)},\mathcal{L}_2^{(j)} = -A_G^{(j)}(b_1^{(j)}-b_1^{(j)})^2 - A_G^{(j-1)}(b_1^{(j-1)}-b_1^{(j-1)})^2, \tag{3.22}$$

which also follows from Wodzicki's formula (3.18), where we should set $A_G^{(j)} = A a_{21}^{(j)}(4)/4$.

IV. THE CONFORMAL ANOMALY AND ASSOCIATED OPERATOR PRODUCTS

In this section we start with a conformal deformation of a metric and the conformal anomaly of the energy stress tensor. It is well known that (pseudo-) Riemannian metrics $\mathrm{g}_{\mu\nu}(x)$ and $\tilde{\mathrm{g}}_{\mu\nu}(x)$ on a manifold X are (pointwise) conformal if $\tilde{\mathrm{g}}_{\mu\nu}(x)=\exp(2f)\mathrm{g}_{\mu\nu}(x)$, $f\in C^\infty(\mathbb{R})$. For constant conformal deformations the variation of the connected vacuum functional (the effective action) can be expressed in terms of the generalized zeta function related to an elliptic self-adjoint operator $\mathcal{O}$:[13]

$$\delta W = -\zeta(0|\mathcal{O})\log\mu^2 = \int_{X_\Gamma}\langle T_{\mu\nu}(x)\rangle \delta \mathrm{g}^{\mu\nu}(x)dx, \tag{4.1}$$

where $\langle T_{\mu\nu}(x)\rangle$ means that all connected vacuum graphs of the stress-energy tensor $T_{\mu\nu}(x)$ are to be included. Therefore Eq. (4.1) leads to

$$\langle T_\mu^\mu(x)\rangle = (\mathrm{Vol}(X_\Gamma))^{-1}\zeta(0|\mathcal{O}). \tag{4.2}$$

The formulas (3.5) and (3.9)–(3.11) give an explicit result for the conformal anomaly, namely,

$$\langle T_\mu^\mu(x)\rangle_{(\mathcal{O}=\otimes\mathcal{L}_p^{(j)})} = \frac{1}{(4\pi)^{d/2}\Gamma(d/2)}\sum_{l=0}^{(d/2)-1}\frac{(-1)^{l+1}}{2(l+1)}\left\{\sum_p \left[a_{2l}^{(j)}(d)B_p(j)^{l+1} + a_{2l}^{(j-1)}(d)B_p(j-1)^{l+1}\right] + (2-2^{-2l})\mathcal{B}_{2l+2}(a_{2l}^{(j)}(d)+a_{2l}^{(j-1)}(d))\right\}, \tag{4.3}$$

where d is even. For $B_{1,2}(j)=B(j)$ and $B_{1,2}(j-1)=B(j-1)$, the anomaly (4.3) has the form

$$\langle T_\mu^\mu(x)\rangle_{(\mathcal{L}^{(j)}\otimes\mathcal{L}^{(j)})} = \frac{1}{(4\pi)^{d/2}\Gamma(d/2)}\sum_{l=0}^{(d/2)-1}\frac{(-1)^{l+1}}{2(l+1)}\{[a_{2l}^{(j)}(d)B(j)^{l+1} + a_{2l}^{(j-1)}(d)B(j-1)^{l+1}] + (2-2^{-2l})\mathcal{B}_{2l+2}(a_{2l}^{(j)}(d)+a_{2l}^{(j-1)}(d))\}. \tag{4.4}$$

Note that for a minimally coupled scalar field of mass m, $B(0)=\rho_0^2+m^2$. The simplest case is, for example, $G=\mathrm{SO}_1(2,1)\simeq \mathrm{SL}(2,\mathbb{R})$; besides $X=\mathbb{H}^2$ is a two-dimensional real hyperbolic space. Then we have $\rho_0^2=1/4$, $a_{20}^{(0)}=1$, and finally

$$\langle T_\mu^\nu(x\in\Gamma\backslash\mathbb{H}^2)\rangle_{(\mathcal{L}^{(0)}\otimes\mathcal{L}^{(0)})}=-\frac{1}{4\pi}\left(b+\frac{1}{3}\right). \tag{4.5}$$

For real d-dimensional hyperbolic space the scalar curvature is $R(x)=-d(d-1)$. In the case of the conformally invariant scalar field we have $B(0)=\rho_0^2+R(x)(d-2)/[4(d-1)]$. As a consequence, $B(0)=\frac{1}{4}$ and

$$\langle T_\mu^\mu(x\in\Gamma\backslash\mathbb{H}^d)\rangle_{(\mathcal{L}^{(0)}\otimes\mathcal{L}^{(0)})}=\frac{1}{(4\pi)^{d/2}\Gamma(d/2)}\times\sum_{l=0}^{(d/2)-1}\frac{(-1)^{l+1}}{l+1}a_{2l}^{(0)}(d)\{2^{-2l-2}+(1-2^{-2l-1})\mathcal{B}_{2l+2}\}. \tag{4.6}$$

Thus in conformally invariant scalar theory the anomaly of the stress tensor coincides with one associated with operator product. This statement holds not only for hyperbolic spaces considered above, but for all constant curvature manifolds as well.[17]

V. PRODUCT OF EINSTEIN MANIFOLDS

In this section we consider the problem of the global existence of zeta functions on (pseudo-) Riemannian product manifolds, a product of two Einstein manifolds[40,41]

$$(X,\mathbf{g},\mathcal{P})=(X_1,\mathbf{g}_1,\mathcal{P}_1)\otimes(X_2,\mathbf{g}_2,\mathcal{P}_2), \tag{5.1}$$

where $\mathbf{g}=\mathbf{g}_1\otimes\mathbf{g}_2$ and the metric $\mathbf{g}$ separates the variables, i.e.,

$$ds^2=\mathbf{g}_{\alpha\beta}(x)dx^\alpha\otimes dx^\beta+\mathbf{g}_{\mu\nu}(y)dy^\mu\otimes dy^\nu. \tag{5.2}$$

The tangent bundle splits as $TX=TX_1\oplus TX_2$ and $\mathcal{P}=\mathcal{P}_1+\mathcal{P}_2$, where $\mathcal{P}_p(p=1,2)$ are the corresponding projections on TX_p,

$$\mathcal{P}^2=\mathrm{Id},\quad \mathbf{g}(\mathcal{P}\mathcal{X},\mathcal{P}\mathcal{Y})=\mathbf{g}(\mathcal{X},\mathcal{Y}),\quad \mathcal{X},\mathcal{Y}\in\mathcal{G}(X), \tag{5.3}$$

$\mathcal{G}(X)$ being the Lie algebra of vector fields $\mathcal{X}$ and $\mathcal{Y}$ on X. The trivial examples of an almost-product structure are given by the choices $\mathcal{P}=\pm\mathrm{Id}$ ($\pm$ identity).

We recall some facts about Einstein manifolds. An almost-product (pseudo-) Riemannian structure $(X,\mathbf{g},\mathcal{P})$ is integrable iff $\Delta\mathcal{P}=0$ for the Levi–Civita connection Δ of $\mathbf{g}$. The two integrable complementary subbundles, i.e., both foliations, are totally geodesic.[40,41] Let X be a pseudo-Kähler manifold. Such a manifold is an Einstein manifold iff in any adapted coördinates (x^α,y^α) both metrics $\mathbf{g}_1$ and $\mathbf{g}_2$ are Einstein metrics for the same constant λ,[40–42] i.e.,

$$\mathrm{Ric}(\mathbf{g})=\lambda\mathbf{g}. \tag{5.4}$$

Our consideration will be restricted to only locally decomposable manifolds. A wide class of (pseudo-) Riemannian manifolds includes nonlocally decomposable manifolds as well, which are given by warped product space–times.[43–45] Note that many exact solutions of Einstein equations (associated with Schwarzschild, Robertson–Walker, Reissner–Nordström, de Sitter space–times) and p-brane solutions are, in fact, warped product space–times.

A. The explicit form of the zeta function

We study the zeta function

$$\zeta(s|\mathcal{L}^{(j)}\oplus\mathcal{L}^{(k)})=\zeta_{\Gamma_1\backslash X\otimes\Gamma_2\backslash X}(s)=\frac{1}{\Gamma(s)}\int_0^\infty\omega_{\Gamma_1}^{(j)}(t)\omega_{\Gamma_2}^{(k)}(t)t^{s-1}\,dt,\quad \mathrm{Re}\,s>d. \tag{5.5}$$

Let $B=B_1(j)+B_2(k)$, $A_p\overset{\text{def}}{=}\chi(1)\ \text{vol}\ (\Gamma_p\backslash G)C^{(j)}(d)/4$, $y_p(s;z)\overset{\text{def}}{=}s/2+(-1)^{p-1}z$, $p=1,2(s,z\in\mathbb{C})$. The explicit construction gives more, namely the following.

Theorem 5.1: *The function $\zeta(s|\mathcal{L}^{(j)}\otimes\mathcal{L}^{(k)})$ admits an explicit meromorphic continuation to $\mathbb{C}$ with at most a simple pole at $s=1,2,...,d$. In particular on the domain $\text{Re}\, s<1$,*

$$
\begin{aligned}
\zeta(s|\mathcal{L}^{(j)}\oplus\mathcal{L}^{(k)})=&\frac{\pi^2}{2}A_1A_2[C^{(j)}(d)+C^{(j-1)}(d)][C^{(k)}(d)+C^{(k-1)}(d)]\\
&\times\sum_{m=0}^{n-1}\sum_{l=0}^{m}\sum_{\mu=0}^{n-1}\sum_{\nu=0}^{\mu}\frac{[a_{2m}^{(j)}(d)+a_{2m}^{(j-1)}(d)][a_{2m}^{(k)}(d)+a_{2m}^{(k-1)}(d)]}{(m-l)!(\mu-\nu)!}\\
&\times\frac{\int_0^\infty r^{2(m-l)}\,\text{sech}^2(\pi r)\mathcal{K}_{\mu-\nu}(s-l-\nu-1;r^2+B,\pi)dr}{(s-1)(s-2)\cdots(s-(l+1))(s-(l+2))\cdots(s-(l+1+\nu+1))}\\
&+C^{(j)}(d)V_1\sin(\pi s)\sum_{m=0}^{n-1}[a_{2m}^{(j)}(d)+a_{2m}^{(j-1)}(d)]\int_{\mathbb{R}}r^{2m+1}\tanh(\pi r)\\
&\times\Bigg[\int_0^\infty\Psi_{\Gamma_2}(\rho_0-k+t+\sqrt{r^2+B};\chi_2)(2t\sqrt{r^2+B}+t^2)^{-s}\,dt+\int_0^\infty\Psi_{\Gamma_2}(\rho_0-k+1\\
&+t+\sqrt{r^2+B};\chi_2)(2t\sqrt{r^2+B}+t^2)^{-s}dt\Bigg]dr+C^{(j)}(d)V_2\sin(\pi s)\\
&\times\sum_{m=0}^{n-1}[a_{2m}^{(k)}(d)+a_{2m}^{(k-1)}(d)]\int_{\mathbb{R}}r^{2m+1}\tanh(\pi r)\Bigg[\int_0^\infty\Psi_{\Gamma_1}(\rho_0-j+t\\
&+\sqrt{r^2+B};\chi_1)(2t\sqrt{r^2+B}+t^2)^{-s}\,dt+\int_0^\infty\Psi_{\Gamma_1}(\rho_0-j+1+t+\sqrt{r^2+B};\chi_1)\\
&\times(2t\sqrt{r^2+B}+t^2)^{-s}dt\Bigg]dr+\frac{1}{2\pi^3 i\Gamma(s)}\int_{\text{Re}_{z=\varepsilon}}dz\left[\sin\pi\left(z+\frac{s}{2}\right)\right]\\
&\times\left[\sin\pi\left(\frac{s}{2}-z\right)\right]\Gamma\left(z+\frac{s}{2}\right)\Gamma\left(\frac{s}{2}-z\right)\Bigg[\int_0^\infty(\Psi_{\Gamma_1}(\rho_0-j+t+B_1^{1/2};\chi_1)(2tB_1^{1/2}\\
&+t^2)^{-y_1(s,z)}+\Psi_{\Gamma_1}(\rho_0-j+1+t+B_1^{1/2};\chi_1)(2tB_1^{1/2}+t^2)^{-y_1(s,z)})dt\Bigg]\\
&\times\Bigg[\int_0^\infty(\Psi_{\Gamma_2}(\rho_0-k+t+B_2^{1/2};\chi_2)(2tB_2^{1/2}+t^2)^{-y_2(s,z)}+\Psi_{\Gamma_2}(\rho_0-k+1+t\\
&+B_2^{1/2};\chi_2)(2tB_2^{1/2}+t^2)^{-y_2(s,z)})dt\Bigg],
\end{aligned}
\tag{5.6}
$$

for any $-\frac{1}{2}\leqslant\varepsilon\leqslant\frac{1}{2}$. For a, $\delta>0$, $\nu\in\mathbb{N}$, the entire function $\mathcal{K}_\nu(s;\delta,a)$ of s is defined by

$$
\mathcal{K}_\nu(s;\delta,a)\overset{\text{def}}{=}\int_{\mathbb{R}}\frac{r^{2\nu}\,\text{sech}^2(ar)dr}{(\delta+r^2)^s}.
\tag{5.7}
$$

All of the integrals are entire functions of s.

The simplest case is, for example, $G=SO_1(2,1)\simeq SL(2,\mathbb{R})$; besides $X=\mathbb{H}^2$ is a two-dimensional real hyperbolic space. Then for $j=k=0$, $\Gamma_1=\Gamma_2=\Gamma$, we have $a_{20}^{(0)}(2)=1$ and $\mu_{\sigma 0}(r)=\pi r\tanh(\pi r)$ and for $\text{Re}\, s<1$ we have

$$\zeta(s|\mathcal{L}^{(0)}\oplus\mathcal{L}^{(0)})=\frac{\pi A_1^2}{2(s-1)(s-2)}\int_0^\infty \mathrm{sech}^2(\pi r)\mathcal{K}_0(s-2;r^2+B,\pi)dr+\frac{2}{\pi}A_1\sin(\pi s)$$
$$\times\int_{\mathbb{R}} r\tanh(\pi r)dr\int_0^\infty \psi_\Gamma\left(\frac{1}{2}+t+\sqrt{r^2+B};\chi\right)(2t\sqrt{r^2+B}+t^2)^{-s}\,dt$$
$$+\frac{1}{2\pi^3 i\Gamma(s)}\int_{\mathrm{Re}_{z=\varepsilon}}\left[\sin\pi\left(z+\frac{s}{2}\right)\right]\left[\sin\pi\left(\frac{s}{2}-z\right)\right]\Gamma\left(z+\frac{s}{2}\right)$$
$$\times\Gamma\left(\frac{s}{2}-z\right)dz\int_0^\infty\psi_\Gamma\left(\frac{1}{2}+t+B_0^{1/2};\chi\right)(2tB_0^{1/2}+t^2)^{-y_1(s;z)}dt$$
$$\times\int_0^\infty\psi_\Gamma\left(\frac{1}{2}+t+B_0^{1/2};\chi\right)(2tB_0^{1/2}+t^2)^{-y_2(s;z)}\,dt, \tag{5.8}$$

where $B=\frac{1}{2}+2b^{(0)}$ and $B(0)=\frac{1}{2}+b^{(0)}$.

VI. QUADRATIC FUNCTIONAL WITH ELLIPTIC RESOLVENT AND ANALYTIC TORSION

Let $\chi:\pi_1(X_\Gamma)\mapsto O(V,\langle\cdot,\cdot\rangle v)$ be a representation of $\pi_1(X_\Gamma)$ on a real vector space V. The mapping χ determines (on the basis of a standard construction in differential geometry) a real flat vector bundle ξ over X_Γ and a flat connection map D on the space $\Omega(X_\Gamma,\xi)$ of differential forms on X_Γ with values in ξ. One can say that χ determines the space of smooth sections in the vector bundle $\Lambda(TX_\Gamma)^*\otimes\xi$.

Let D_j denote the restriction of D to the space $\Omega^j(X_\Gamma,\xi)$ of j- forms and let

$$H^j(D)=\ker(D_j)[\mathrm{Im}(D_{j-1})]^{-1} \tag{6.1}$$

be the corresponding cohomology spaces. There exists a canonical Hermitian structure on the bundle χ which we denote by $\langle\cdot,\cdot\rangle_V$. The above-mentioned Hermitian structure determines for each $x\in X_\Gamma$ a linear map $\langle\cdot,\cdot\rangle_x:\xi_x\otimes\xi_x\mapsto\mathbb{R}$, and the diagram for linear maps hold (see Ref. 46 for details)

$$(\Lambda^p(T_xX_\Gamma)^*\otimes\xi_x)\otimes(\Lambda^q(T_xX_\Gamma)^*\otimes\xi_x)\overset{\wedge}{\mapsto}\Lambda^{p+q}(T_xX_\Gamma)^*\otimes(\xi_x\otimes\xi_x)\overset{\langle\cdot,\cdot\rangle_x}{\mapsto}\Lambda^{p+q}(T_xX_\Gamma)^*, \tag{6.2}$$

where the image of $\omega_x\otimes\tau_x$ under the first map has been denote by $\langle\omega_x\wedge\tau_x\rangle_x$.

We define the quadratic functional S_D on $\Omega^j(X_\Gamma,\xi)$ by

$$S_D(\omega)=\int_{X_\Gamma}\langle\omega(x)\wedge(D_j\omega)(x)\rangle_x. \tag{6.3}$$

One can construct from the metric on X_Γ and Hermitian structure in ξ a Hermitian structure in $\Lambda(T_xX_\Gamma)^*\otimes\xi$ and the inner products $\langle\cdot,\cdot\rangle_j$ in the space $\Omega^j(X_\Gamma,\xi)$. Thus

$$S_D(\omega)=\langle\omega,T\omega\rangle_j,\quad T=*D_j, \tag{6.4}$$

where (*) is the Hodge-star map. Recall that the map T is formally self-adjoint with the property $T^2=D_j^*D_j$. Let $(\mathfrak{J}_p,D_p)$ be a complex, i.e., a sequence of vector space $\mathfrak{J}_p$ and linear operators D_p acting from the space $\mathfrak{J}_p$ to the space $\mathfrak{J}_{p+1}$ $(\mathfrak{J}_{-1}=\mathfrak{J}_{d+1})$ and satisfying $D_{p+1}D_p=0$ for all $p=0,1,...,d$. Let us define the adjoint operators $D_p^*:\mathfrak{J}_{p+1}\mapsto\mathfrak{J}_p$ by $\langle a,D_pb\rangle_{p+1}=\langle D_p^*a,b\rangle_p$. For the functional (6.3) there is a canonical topological elliptic resolvent $R(S_D)$ (a chain of linear maps)

$$0\overset{0}{\mapsto}\Omega^0(X_\Gamma,\xi)\overset{D_0}{\mapsto}\cdots\overset{D_{d-2}}{\mapsto}\Omega^{d-1}(X_\Gamma,\xi)\overset{D_{d-1}}{\mapsto}\ker(S_D)\overset{0}{\mapsto}0. \tag{6.5}$$

From Eq. (6.5) it follows that for $R(S_D)$ we have $H^p(R(S_D))=H^{d-p}(D)$ and $\ker(S_D)\equiv\ker(T)=\ker(D_j)$.

Let us choose an inner product $\langle\cdot,\cdot\rangle_{H^p}$ in each space $H^p(R(S_D))$. The partition function of S_D associated to points in the moduli space of flat gauge fields $\omega(x)$ on X_Γ with the resolvent (6.5) can be written in the form (see, for example, Ref. 46)

$$\mathcal{Z}(\beta)\equiv\mathcal{Z}(\beta;R(S_D),\langle\cdot,\cdot\rangle_H,\langle\cdot,\cdot\rangle)=\mathfrak{I}(\beta,\zeta,\eta)\,\tau(X_\Gamma,\chi,\langle\cdot,\cdot\rangle_H)^{1/2}, \tag{6.6}$$

where $\beta=i\lambda$, $\lambda\in\mathbb{R}$, $\mathfrak{I}(\beta,\zeta,\eta)$ is known function of β. The function ζ appearing in the partition function above can be expressed in terms of the dimensions of the cohomology spaces of D,

$$\zeta\equiv\zeta(0||T|)=-\sum_{p=0}^{d}(-1)^p\dim H^p(R(S))=(-1)^{n+1}\sum_{q=0}^{n}(-1)^q\dim H^q(D). \tag{6.7}$$

The dependence of $\eta=\eta(0|T_D)$ on the connection map D can be expressed with the help of formulas for the index of the twisted signature operator for a certain vector bundle over $X_\Gamma\otimes[0,1]$.[47] It can be shown that[46]

$$\eta(s|B^{(l)})=2\,\eta(s|T_{D^{(l)}}), \tag{6.8}$$

where the $B^{(l)}$ are elliptic self-adjoint maps on $\Omega(X_\Gamma,\xi)$ defined on j-forms by

$$B_j^{(l)}=(-i)^{\lambda(j)}(*D^{(l)}+(-1)^{j+1}D^{(l)}*). \tag{6.9}$$

In this formula $\lambda(j)=(j+1)(j+2)+n+1$ and for the Hodge star-map we have used that $*\alpha\wedge\beta=\langle\alpha,\beta\rangle_{\text{vol}}$. From the Hodge theory

$$\dim\ker B^{(l)}=\sum_{q=0}^{d}\dim H^q(D^{(l)}). \tag{6.10}$$

The metric dependence of η enters through $L^r(TX_\Gamma)$ and $\eta(0|T_{D^{(0)}})$, where $L^r(TX_\Gamma)$ is the r'th term in Hirzebruch's L-polynomial (see Ref. 47 for details) and $D^{(0)}$ is an arbitrary flat connection map on $\Omega(X_\Gamma,\xi)$. For $d=3$ the only contribution of the L-polynomial is $L_0=1$ and the metric dependence of η is determined alone by $\eta(0|T_{D^{(0)}})$.

The factor $\tau(X_\Gamma,\chi,\langle\cdot,\cdot\rangle_H)$ is independent of the choice of metric g on X.[46] In fact, this quantity is associated with the analytic (Ray–Singer) torsion $T_{\text{an}}(X_\Gamma)$ (Ref. 20) of the representation χ of $\pi_1(X_\Gamma)$ constructed using the metric g. If $H^0(D)\neq 0$ and $H^q(D)=0$ for $q=1,...,n$, $d=2n+1$ the dimension of X, then the product

$$\tau(X_\Gamma,\chi,\langle\cdot,\cdot\rangle_H)=T_{\text{an}}(X_\Gamma)\cdot\text{Vol}(X_\Gamma)^{-\dim H^0(D)}, \tag{6.11}$$

is independent of the choice of metric g, i.e., the metric dependence of the Ray–Singer torsion $T_{\text{an}}(X_\Gamma)$ factors out as $\text{Vol}(X_\Gamma)^{-\dim H^0(D)}$.

A. Connected sum of three-manifolds

Recall that an embedding of the cohomology $H(X_\Gamma;\xi)$ into $\Omega(X_\Gamma;\xi)$ as the space of harmonic forms induces a norm $|\cdot|^{\text{RS}}$ on the determinant line $\det H(X_\Gamma;\xi)$. The Ray–Singer norm $\|\cdot\|^{\text{RS}}$ on $\det H(X_\Gamma;\xi)$ is defined by[20]

$$\|\cdot\|^{\text{RS}}\overset{\text{def}}{=}|\cdot|\prod_{j=0}^{\dim X}\left[\exp\left(-\frac{d}{ds}\zeta(s|\mathcal{L}^{(j)})\bigg|_{s=0}\right)\right]^{(-1)^j j/2}, \tag{6.12}$$

where the zeta function $\zeta(s|\mathcal{L}^{(j)})$ of the Laplacian acting on the space of j-forms orthogonal to the harmonic forms has been used. For a closed connected orientable smooth manifold of odd dimension and for any Euler structure $\eta\in \mathrm{Eul}(X)$ the Ray–Singer norm of its cohomological torsion $T_{\rm an}(X_\Gamma,\eta)\in \det H(X_\Gamma;\xi)$ is equal to the positive square root of the absolute value of the monodromy of ξ along the characteristic class $c(\eta)\in H^1(X_\Gamma)$:[48] $\|T_{\rm an}(X_\Gamma)\|^{\rm RS}=|\det_\xi c(\eta)|^{1/2}$. In the special case where the flat bundle ξ is acyclic $[H^q(X_\Gamma;\xi)=0$ for all $q]$ we have

$$[T_{\rm an}(X_\Gamma,\eta)]^2=|\det_\xi c(\eta)|\prod_{j=0}^{\dim X}\left[\exp\left(-\frac{d}{ds}\zeta(s|\mathcal{L}^{(j)})\bigg|_{s=0}\right)\right]^{(-1)^{j+1}j}. \tag{6.13}$$

If ξ is unimodular, then $|\det_\xi c(\eta)|=1$ and the torsion $T_{\rm an}(X_\Gamma)$ does not depend on the choice of η. For odd-dimensional manifold the Ray–Singer norm is a topological invariant: it does not depend on the choice of metric on X and ξ, used in the construction. But for even-dimensional X this is not the case.[49]

Suppose the flat bundle ξ is acyclic. The analytic torsion $T_{\rm an}(\Gamma\backslash\mathbb{H}^3)$ can be expressed in terms of the Selberg zeta functions $Z_j(s,\chi_j)$. Indeed the Ruell zeta function in three dimensions associated with the closed oriented hyperbolic manifold $X_\Gamma=\Gamma\backslash\mathbb{H}^3$ has the form

$$\mathcal{R}_\chi(s)=\prod_{j=0}^{2}Z_j(j+s,\chi_j)^{(-1)^j}=\frac{Z_0(s,\chi_0)Z_2(2+s,\chi_2)}{Z_1(1+s,\chi_1)}. \tag{6.14}$$

The function $\mathcal{R}_\chi(s)$ extends meromorphically to the entire complex plane $\mathbb{C}$.[50] For the Ray–Singer torsion one gets[51]

$$[T_{\rm an}(\Gamma\backslash\mathbb{H}^3)]^2=\mathcal{R}_\chi(0)=\frac{[Z_0(2,\chi_0)]^2}{Z_1(1,\chi_1)}\exp\left(-\frac{\mathrm{Vol}(\mathcal{F})}{3\pi}\right), \tag{6.15}$$

where $\mathrm{Vol}(\mathcal{F})$ is a volume of the fundamental domain $\mathcal{F}$ of $\Gamma\backslash\mathbb{H}^3$. In the presence of nonvanishing Betti numbers $b_j=b_j(X_\Gamma)$ we have[51,52]

$$[T_{\rm an}(\Gamma\backslash\mathbb{H}^3)]^2=\frac{(b_1-b_0)![Z_0^{b_0}(2,\chi_0)]^2}{[b_0!]^2Z_1^{(b_1-b_0)}(1,\chi_1)}\exp\left(-\frac{\mathrm{Vol}(\mathcal{F})}{3\pi}\right). \tag{6.16}$$

In Chern–Simons theory the partition function at level k $(\lambda=2\pi k)$ depends on a framing (i.e., on a trivialization of the normal bundle to the link) of twice the tangent bundle as a Spin (6) bundle, henceforth referred to as two-framing. In particular, for the SU(2) theory in the large k-limit the partition function for a connected sum $\mathcal{X}=X_{\Gamma,1}\#X_{\Gamma,2}\#\cdots\#X_{\Gamma,N}$ can be written as follows:[53]

$$\mathcal{Z}(\mathcal{X})=\frac{\otimes_{l=1}^{N}\mathcal{Z}(X_{\Gamma,l})}{[\mathcal{Z}(S^3)]^{N-1}}. \tag{6.17}$$

Equation (6.17) holds for any given two-framings among $X_{\Gamma,p}$ and $X_{\Gamma,q}$, $p,q\leqslant N$, the induced two-framing on $X_{\Gamma,p}\#X_{\Gamma,q}$, and a canonical two-framing on $\mathbb{S}^3$. Since the Ray–Singer torsion on $\mathbb{S}^3$ is to be equal to one, $\mathcal{Z}(\mathbb{S}^3)=\sqrt{2}\pi k^{-3/2}$, the partition function associated with the semiclassical approximation $(k\to\infty)$ takes the form

$$\mathcal{Z}_{\rm sc}(\mathcal{X})=\left(\frac{k^3}{2\pi^2}\right)^{(N-1)/2}\bigotimes_{l=1}^{N}\mathcal{Z}_{\rm sc}(X_{\Gamma,l})=\sqrt{2}\pi k^{-3/2}\bigotimes_{l=1}^{N}|\mathcal{R}_{\chi(l)}(0)|^{1/2}, \tag{6.18}$$

while in the presence of nonvanishing Betti numbers $b_{jl}=b_j(X_{\Gamma,l})$ one gets

$$\mathcal{Z}_{\rm sc}(\mathcal{X})=\sqrt{2}\pi k^{-3/2}\bigotimes_{l=1}^{N}\left[\frac{(b_{1l}-b_{0l})!(Z_0^{(b_{0l})}(2,\chi_0))^2}{(b_{0l}!)^2Z_1^{(b_{1l}-b_{0l})}(1,\chi_1)}\right]^{1/4}\exp\left[-\frac{1}{12\pi}\bigoplus_{l=1}^{N}\mathrm{Vol}(\mathcal{F}_l)\right]. \tag{6.19}$$

In the case of nontrivial characters, $b_0(X_{\Gamma,l})=0$. If $b_1=0$, then Eq. (6.15) holds.

For the trivial character one has $b_0=1$ (for any closed manifold) and $b_1=0$ for an infinite number of $X_\Gamma=\Gamma\backslash\mathbb{H}^3$. The function $\mathcal{R}(s)$ has a zero at $s=0$ of order 4.[27] However, there is a class of compact sufficiently large hyperbolic manifolds which admit arbitrarily large values of $b_1(X_\Gamma)$. Sufficiently large manifolds X_Γ contain a surface S whose fundamental group $\pi_1(S)$ is infinite and such that $\pi_1(S)\subset\pi_1(X_\Gamma)$.

It seems that the most important problem in three-topology is the classification problem. In general, hyperbolic manifolds have not been completely classified and therefore a systematic computation is not yet possible. However, this is not the case for certain sufficiently large manifolds, the Haken manifolds.[54] There exists an algorithm for the enumeration of all Haken manifolds and there exists an algorithm for recognizing homeomorphy of the Haken manifolds.[55] Both algorithms depend on normal surface theory in the manifold, developed primarily by Haken. These manifolds give an essential contribution to the partition function (6.19).

VII. CONCLUDING REMARKS

We have obtained an explicit formula for the multiplicative anomaly (Theorem 3.5). The anomaly is equal to zero for $d=2$ and for the odd-dimensional cases. We have preferred to limit discussion here on various particular cases in detail and emphasize the general picture. It seems to us that the explicit results for the anomaly (3.20) and (3.22) are not only interesting as mathematical results, but are very important in several physical examples where the determinant of a product of differential operators is not equal to the product of the corresponding functional determinants (see, for example, Ref. 56). Because a large amount of recent activity has involved the calculation of the conformal anomaly of dilaton coupled matter, it would also be of great interest to generalize our results to the dilaton-dependent trace anomaly.

We have also considered product structures on closed real hyperbolic manifolds. In fact, the explicit form of the zeta function on product spaces (Theorem 5.1) is derived. As an example the zeta function associated with the Krönecker sum of Laplacians on twisted forms is calculated in the two-dimensional case.

Finally, the explicit formulas for analytic torsion $T_{\rm an}(X_\Gamma)$ (a topological invariant) on manifolds of the form $X_\Gamma=\Gamma\backslash\mathbb{H}^d$ and on a connected sum of such manifolds are derived. Thus our results enable one to calculate the Chern–Simons partition function appearing in the semiclassical approximation for the Witten invariant. We have also discussed briefly the fact that for Haken manifolds a larger influence, due to their Betti numbers, is exerted on the form of the partition function. We hope that proposed discussion of this invariant will be interesting in view of future applications to concrete problems in quantum field theory.

ACKNOWLEDGMENTS

The first author was partially supported by a CNPq grant (Brail), RFFI grant (Russia) No. 98-02-18380-a, and by GRACENAS grant (Russia) No. 6-18-1997.

1. M. Wodzicki, *Lecture Notes in Mathematics* **1289**, 320 (1987).
2. M. Kontsevich and S. Vishik, ''Determinants of Elliptic Pseudo-Differential Operators,'' preprint MPI/94-30 (1994).
3. M. Kontsevich and S. Vishik, ''Geometry of Determinants of Elliptic Operators,'' hep-th/9406140 (1994).
4. A. Connes, Commun. Math. Phys. **117**, 673 (1988).
5. D. Kastler, Commun. Math. Phys. **166**, 633 (1995).
6. A. Connes and J. Lott, Nucl. Phys. B **18**, 29 (1990).
7. A. Connes, *Non-Commutative Geometry* (Academic, New York, 1994).
8. A. Connes, Commun. Math. Phys. **182**, 155 (1996).
9. W. Kalau and M. Walze, J. Geom. Phys. **16**, 327 (1995).
10. E. Elizalde, L. Vanzo, and S. Zerbini, Commun. Math. Phys. **194**, 613 (1998).
11. A. A. Bytsenko and F. L. Williams, J. Math. Phys. **39**, 1075 (1998).
12. E. Elizalde, G. Cognola, and S. Zerbini, ''Applications in Physics of the Multiplicative Anomaly Formula Involving Some Basic Differential Operators,'' Preprint Trento University 413 (1998); to be published in Nucl. Phys. B.
13. I. L. Buchbinder, S. D. Odintsov, and I. L. Shapiro, *Effective Action in Quantum Gravity* (Hilger, Bristol, 1992).
14. S. Deser and A. Schwimmer, Phys. Lett. B **309**, 279 (1993).

[15] E. Copeland and D. Toms, Class. Quantum Grav. **3**, 431 (1986).
[16] A. A. Bytsenko, E. Elizalde, and S. D. Odintsov, J. Math. Phys. **36**, 5084 (1995).
[17] A. A. Bytsenko, A. E. Gonçalves, and F. L. Williams, Mod. Phys. Lett. A **13**, 99 (1998).
[18] A. A. Bytsenko, A. E. Gonçalves, and F. L. Williams, JETP Lett. **67**, 176 (1998).
[19] E. Witten, Commun. Math. Phys. **121**, 351 (1989).
[20] D. B. Ray and I. M. Singer, Adv. Math. **7**, 145 (1971).
[21] A. S. Schwarz, Lett. Math. Phys. **2**, 247 (1978).
[22] A. S. Schwarz, Commun. Math. Phys. **67**, 1 (1979).
[23] W. Müller, Adv. Math. **28**, 233 (1978).
[24] J. Cheeger, Ann. Math. **109**, 259 (1979).
[25] J. Milnor, Bull. Am. Math. Soc. **72**, 358 (1966).
[26] F. L. Williams, J. Math. Phys. **38**, 796 (1997).
[27] D. Fried, Invent. Math. **84**, 523 (1986).
[28] N. Wallach, Bull. Am. Math. Soc. **82**, 171 (1976).
[29] A. A. Bytsenko, G. Cognola, L. Vanzo, and S. Zerbini, Phys. Rep. **266**, 1 (1996).
[30] F. Williams, ''Lectures on the Spectrum of $L^2(\Gamma\backslash G)$,'' in *Pitman Rearch Notes in Mathematics* (Longman House, 1990), Vol. 242.
[31] F. Williams, Pac. J. Math. **182**, 137 (1998).
[32] A. Selberg, J. India Math. Soc. **20**, 47 (1956).
[33] D. Fried, Ann. Sci. E'cole Norm. Sup. **10**, 133 (1977).
[34] R. Gangolli, Ill. J. Math. **21**, 1 (1977).
[35] R. Gangolli and G. Warner, Nagoya Math. J. **78**, 1 (1980).
[36] D. Scott, Ann. Math. **253**, 177 (1980).
[37] M. Wakayama, Hiroshima Math. J. **15**, 235 (1985).
[38] F. Williams, ''Some Zeta Functions Attached to $\Gamma\backslash G/K$,'' in *New Developments in Lie Theory and Their Applications*, edited by J. Tirao and N. Wallach, Birkhäuser Progress in Math. Ser. **105**, 163 (1992).
[39] C. Kassel, ''Le Residue non Commutatif (d'apres M. Wodzicki),'' Semin. Bourbaki, 41 eme ann. 1988–1989, Exp., 708 (1989).
[40] K. Yano, *Differential Geometry on Complex and Almost Complex Spaces* (Pergamon, Oxford, 1965).
[41] K. Yano and M. Kon, *Structures on Manifolds*, Series in Pure Math. (World Scientific, Singapore, 1984), Vol. 3.
[42] A. L. Besse, *Einstein Manifolds* (Springer-Verlag, New York, 1987).
[43] B. O'Neil, *Semi-Riemannian Geometry* (Academic, New York, 1983).
[44] R. Deszcz, L. Verstralen, and L. Vrancken, Gen. Relativ. Gravit. **23**, 671 (1991).
[45] J. Carot and J. da Costa, Class. Quant. Grav. **10**, 461 (1993).
[46] D. H. Adams and S. Sen, ''Partition Function of a Quadratic Functional and Semiclassical Approximation for Witten's 3-Manifold Invariant,'' hep-th/9503095 (1995).
[47] M. F. Atiyah, V. K. Patodi, and I. M. Singer, Math. Proc. Camb. Philos. Soc. **77**, 43 (1975).
[48] M. Farber and V. Turaev, ''Poincaré-Reidemeister Metric, Euler Structures, and Torsion,'' math.DG/9803137 (1998).
[49] J.-M. Bismut and W. Zhang, Asterisque **205**, (1992).
[50] A. Deitmar, Abh. Math. Sem. **59**, 101 (1989).
[51] A. A. Bytsenko, L. Vanzo, and S. Zerbini, Nucl. Phys. B **505**, 641 (1997).
[52] A. A. Bytsenko, A. E. Gonçalves, and W. da Cruz, Mod. Phys. Lett. A **13**, 2453 (1998).
[53] E. Witten, Commun. Math. Phys. **121**, 351 (1989).
[54] W. Haken, Acta Math. **105**, 245 (1961).
[55] S. V. Matveev, Russian Math. Surveys (Usp. Mat. Nauk) **52**, 1029 (1997).
[56] G. Cognola and S. Zerbini, Nucl. Phys. **532**, 407 (1998).

Matrix representation of octonions and generalizations

Jamil Daboul[a)] and Robert Delbourgo[b)]
School of Mathematics and Physics, University of Tasmania, Hobart, Australia

(Received 31 December 1998; accepted for publication 25 March 1999)

We define a special matrix multiplication among a special subset of $2N\times 2N$ matrices, and study the resulting (nonassociative) algebras and their subalgebras. We derive the conditions under which these algebras become alternative nonassociative, and when they become associative. In particular, these algebras yield special matrix representations of octonions and complex numbers; they naturally lead to the Cayley–Dickson doubling process. Our matrix representation of octonions also yields elegant insights into Dirac's equation for a free particle. A few other results and remarks arise as byproducts. © *1999 American Institute of Physics.* [S0022-2488(99)03108-4]

I. INTRODUCTION

Complex numbers and functions have played a pivotal role in physics for three centuries. On the other hand, their generalization to other Hurwitz algebras does not seem to have fired the interest of physicists to the same extent, because there is still no *compelling* application of them. Thus, despite the fascination of quaternions and octonions for over a century, it is fair to say that they still await universal acceptance. This is not to say that there have not been valiant attempts to find appropriate uses for them. One can point to their possible impact on quantum mechanics and Hilbert space,[1] relativity and the conformal group,[2] field theory and functional integrals,[3] internal symmetries in particle physics,[4] color field theories,[5] and formulations of wave equations.[6]

In all these cases, there is nothing that stands out and commands our attention; rather, the attempts to describe relativistic physics in terms of quaternions and octonions look rather contrived if not forced, especially for the case of octonions. In this paper we describe a generalization of octonions that allows for Lie algebras beyond the obvious SU(2) structure that is connected with quaternions. We do not presume that they will lead to new physics, but we do think they will at least provide a new avenue for investigation.

Since octonions are not associative, they cannot be represented by matrices with the usual multiplication rules. In this paper, we give representations of octonions and other nonassociative algebras by special matrices, which are endowed with very special multiplication rules; these rules can be regarded as an adaptation and generalization of Zorn's multiplication rule.[7] These matrix representations suggest generalizations of octonions to other nonassociative algebras, which, in turn, lead one almost automatically to a construction of new algebras from old ones, with double the number of elements; we have called these ''double algebras.'' Closer inspection reveals that our procedure can be made to correspond to the Cayley–Dickson construction method,[8] except that in our case the procedure seems rather natural, once one accepts the multiplication rule, whereas the Cayley–Dickson rule looks *ad hoc* at first sight.

II. DEFINITIONS, NOTATIONS, AND A REVIEW OF THE OCTONION ALGEBRA $\mathcal{O}$

The Cayley or the octonion algebra $\mathcal{O}$ is an eight-dimensional nonassociative algebra, which is defined in terms of the basis elements e_μ $(\mu=0,1,2,...,7)$ and their multiplication table. e_0

[a)]Permanent address: Department of Physics, Ben Gurion University of the Negev, 84105 Beer Sheva, Israel; electronic mail: daboul@bgumail.bgu.ac.il
[b)]Electronic mail: Bob.Delbourgo@utas.edu.au

stands for the unit element. We can efficiently summarize the table by introducing the following notation [in general, we shall use Greek indices ($\mu,\nu,...$) to include the 0 and latin indices $(i,j,k,\dots)$ when we exclude the 0]:

$$\hat{e}_k \equiv e_{4+k}, \quad \text{for } k=1,2,3. \tag{1}$$

The multiplication rules among the basis elements of octonions e_μ can be expressed in the form

$$-e_4 e_i = e_i e_4 = \hat{e}_i, \quad e_4 \hat{e}_i = -\hat{e}_i e_4 = e_i, \quad e_4 e_4 = -e_0, \tag{2}$$

$$e_i e_j = -\delta_{ij} e_0 + \epsilon_{ijk} e_k, \tag{3}$$

$$\hat{e}_i \hat{e}_j = -\delta_{ij} e_0 - \epsilon_{ijk} e_k \quad (i,j,k=1,2,3), \tag{4}$$

$$-\hat{e}_j e_i = e_i \hat{e}_j = -\delta_{ij} e_4 - \epsilon_{ijk} \hat{e}_k. \tag{5}$$

We can formally summarize the rules above by

$$e_\mu e_\nu = g_{\mu\nu} e_0 + \sum_{k=1}^{7} \gamma^k_{\mu\nu} e_k, \quad g_{\mu\nu} := \mathrm{diag}(1,-1,...,-1), \quad \gamma^k_{ij} = -\gamma^k_{ji}, \tag{6}$$

where $\mu,\nu=0,1,...,7$, and $i,j,k=1,...,7$. The multiplication properties are sometimes displayed graphically by a circle surrounded by a triangle, but we shall not bother to exhibit that.

The multiplication law (3) shows that the first four elements form a closed *associative* subalgebra of $\mathcal{O}$, which is known as the *quaternion algebra*,

$$\mathcal{Q} \equiv \langle e_0, e_1, e_2, e_3 \rangle_{\mathbb{R}}, \tag{7}$$

while the other rules (2), (4), and (5) show that $\mathcal{O}$ can be graded as follows:

$$\mathcal{O} = \mathcal{Q} \oplus \hat{\mathcal{Q}}, \quad \text{where } \hat{\mathcal{Q}} := e_4 \mathcal{Q}. \tag{8}$$

$\mathcal{O}$ is a nonassociative algebra. Now a measure of the nonassociativity in any algebra $\mathcal{A}$ is provided by the *associator*, which is defined for any three elements, as follows:

$$(x,y,z) := (xy)z - x(yz), \quad \text{for } x,y,z \in \mathcal{A}. \tag{9}$$

In particular, the associators for the octonion basis are

$$(e_i, e_j, e_k) = 2\epsilon_{ijkl} e_l, \tag{10}$$

where ϵ_{ijkl} are *totally* antisymmetric[9] and equal to unity for the following seven combinations:[5]

$$1247, \quad 1265, \quad 2345, \quad 2376, \quad 3146, \quad 3157, \quad \text{and} \quad 4576. \tag{11}$$

The quaternionic subalgebra $\mathcal{Q}$. It is very well known that the quaternions form an associative subalgebra $\mathcal{Q}$, which can be represented by the Pauli matrices:

$$e_0 \to \sigma_0 = 1, \quad \text{and} \quad e_j \to -i\sigma_j \quad (j=1,2,3), \tag{12}$$

where, as usual,

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{13}$$

It is trivial to check that the above map is an isomorphism:

$$e_i e_j \Leftrightarrow -\sigma_i\sigma_j = -(\delta_{ij} + i\epsilon_{ijk}\sigma_k) \Leftrightarrow -\delta_{ij} + \epsilon_{ijk} e_k . \tag{14}$$

III. NONASSOCIATIVE MULTIPLICATION

In contrast to $\mathcal{Q}$, *the Cayley algebra* $\mathcal{O}$ *cannot be represented by matrices with the usual multiplication rules*, because $\mathcal{O}$ is not associative. However, as we demonstrate below, it is possible to represent octonions by matrices, provided one defines a special multiplication rule among them.

A. Zorn's representation of octonions

Zorn[7] gave a representation of the octonions[8] in terms of 2×2 matrices M, whose diagonal elements are scalars and whose off-diagonal elements are three-dimensional vectors:

$$\mathcal{O} \ni x \to \begin{pmatrix} \alpha & \mathbf{a} \\ \mathbf{b} & \beta \end{pmatrix}, \tag{15}$$

and invoked a peculiar multiplication rule for these matrices.[7] With a slight modification of the rule adopted by Humphreys,[10] p. 105, our rule is

$$\begin{pmatrix} \alpha & \mathbf{a} \\ \mathbf{b} & \beta \end{pmatrix} * \begin{pmatrix} \alpha' & \mathbf{a}' \\ \mathbf{b}' & \beta' \end{pmatrix} = \begin{pmatrix} \alpha\alpha' + \mathbf{a}\cdot\mathbf{b}' & \alpha\mathbf{a}' + \beta'\mathbf{a} - \mathbf{b}\times\mathbf{b}' \\ \alpha'\mathbf{b} + \beta\mathbf{b}' + \mathbf{a}\times\mathbf{a}' & \beta\beta' + \mathbf{b}\cdot\mathbf{a}' \end{pmatrix}. \tag{16}$$

We propose to adapt this multiplication law to octonions and also replace the necessary three-dimensional basis vectors $\hat{v}_k$ by Pauli matrices σ_k $(k=1,2,3)$, so that the octonions can be represented by the following ordinary 4×4 matrices:

$$\begin{gathered} e_0 \Leftrightarrow \Omega_0 \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad e_k \Leftrightarrow \Omega_k \equiv \begin{pmatrix} 0 & -\sigma_k \\ \sigma_k & 0 \end{pmatrix} \quad (k=1,2,3), \\ e_4 \Leftrightarrow \Omega_4 \equiv \begin{pmatrix} i & 0 \\ 0 & -1 \end{pmatrix}, \quad \hat{e}_k \Leftrightarrow \hat{\Omega}_k \equiv \begin{pmatrix} 0 & i\sigma_k \\ i\sigma_k & 0 \end{pmatrix}. \end{gathered} \tag{17}$$

[Note the equality of Ω_k $(k=1,2,3)$ to the Dirac matrices γ_k, and Ω_4 to $i\gamma_0$ in the Pauli–Dirac representation.] It can be shown by explicit multiplication, that the above map (17) becomes an isomorphism, provided we define the modified product rule, which we denote by ♥:

$$\begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & A' \\ B' & \beta' \end{pmatrix} = \begin{pmatrix} \alpha\alpha' + \frac{1}{2}\mathrm{Tr}(AB') & \alpha A' + \beta' A + \dfrac{i}{2}[B,B'] \\ \alpha' B + \beta B' - \dfrac{i}{2}[A,A'] & \beta\beta' + \frac{1}{2}\mathrm{Tr}(BA') \end{pmatrix}, \tag{18}$$

where $[A,B] \equiv AB - BA$ is the commutator of A and B. Of course, $A = \mathbf{a}\cdot\boldsymbol{\sigma}$ and $B = \mathbf{b}\cdot\boldsymbol{\sigma}$ are traceless: $\mathrm{Tr}\,A = \mathrm{Tr}\,B = 0$.

In particular, the above multiplication rule yields the following relations:

$$\begin{aligned} \begin{pmatrix} 0 & \eta\sigma_i \\ \xi\sigma_i & 0 \end{pmatrix} \heartsuit \begin{pmatrix} 0 & \eta'\sigma_j \\ \xi'\sigma_j & 0 \end{pmatrix} &= \begin{pmatrix} \eta\xi'\delta_{ij} & \xi\xi'\dfrac{i}{2}[\sigma_i,\sigma_j] \\ -\eta\eta'\dfrac{i}{2}[\sigma_i,\sigma_j] & \xi\eta'\delta_{ij} \end{pmatrix} \\ &= \delta_{ij}\begin{pmatrix} \eta\xi' & 0 \\ 0 & \xi\eta' \end{pmatrix} + \epsilon_{ijk}\begin{pmatrix} 0 & -\xi\xi'\sigma_k \\ \eta\eta'\sigma_k & 0 \end{pmatrix}, \end{aligned} \tag{19}$$

which are helpful for checking the multiplication rules (2)–(5), by substituting the appropriate coefficients, η and ξ.

B. The standard conjugate of octonions

Usually, octonions are studied over the field of real numbers $\mathbb{R}$,

$$x=\sum_{\mu=0}^{7} x_\mu e_\mu \equiv x_0+\mathbf{x}, \quad \text{for } x_\mu \in \mathbb{R}, \tag{20}$$

although later we will find it interesting to deal with their complex extension. The standard *conjugate* $\overline{x}$ of an octonion over $\mathbb{R}$ is defined by

$$\overline{x}:=x_0e_0-\sum_{i=1}^{7} x_ie_i \equiv x_0-\mathbf{x}. \tag{21}$$

The reason for this definition is that the product of $\overline{x}$ with x yields a positive definite norm:

$$n(x)=x\overline{x}=\overline{x}x=\sum_{\mu=0}^{7} x_\mu^2 \geqslant 0. \tag{22}$$

Moreover, this norm obeys the decomposition law,

$$n(xy)=n(x)n(y). \tag{23}$$

However, with complex octonions [real $x \rightarrow$complex z in (20)], we shall still formally define the *conjugate* $\overline{z}$ of z, to be

$$\overline{z}:=z_0e_0-\sum_{i=1}^{7} z_ie_i, \quad \text{for } z_\mu \in \mathbb{C}. \tag{24}$$

It follows that the product $z\overline{z}$ is again proportional to unity:

$$n(z)=z\overline{z}=\overline{z}z=\sum_{\mu=0}^{7} z_\mu^2 \in \mathbb{C}, \tag{25}$$

but $n(z)$ ceases to be real, in general; therefore $n(z)$ should simply be regarded as a scalar function, but not a norm.

It is interesting to calculate $n(z)$ by using the matrix representation (26): First, we note that if z is mapped into the matrix Z, then $\overline{z}$ will be mapped into $\overline{Z}$, as follows:

$$z \rightarrow Z \equiv \sum_{\mu=0}^{7} z_\mu \Omega_\mu = \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix}, \quad \overline{z} \rightarrow \overline{Z} \equiv \begin{pmatrix} \beta & -A \\ -B & \alpha \end{pmatrix}, \tag{26}$$

where $A=\mathbf{a}\cdot\boldsymbol{\sigma}$ and $B=\mathbf{b}\cdot\boldsymbol{\sigma}$, with

$$\alpha=z_0+iz_4, \quad \beta=z_0-iz_4, \quad a_k=-z_k+iz_{4+k}, \quad b_k=z_k+iz_{4+k} \quad (k=1,2,3). \tag{27}$$

Second,

$$z\overline{z} \leftrightarrow Z \heartsuit \overline{Z} = \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix} \heartsuit \begin{pmatrix} \beta & -A \\ -B & \alpha \end{pmatrix} = \left(\alpha\beta - \frac{1}{2}\mathrm{Tr}(AB) \right) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = n(z) I_{4\times 4}. \tag{28}$$

Therefore, we reproduce the expression (25), as expected:

$$n(z):=\frac{1}{4}\operatorname{Tr}(Z\heartsuit\overline{Z})=\alpha\beta-\frac{1}{2}\operatorname{Tr}(AB)=\alpha\beta-\mathbf{a}\cdot\mathbf{b}=\sum_{\mu=0}^{7} z_\mu^2. \tag{29}$$

C. Hermitian conjugate of octonions

Since σ_i are Hermitian matrices, *all our representation matrices* Ω_k *are anti-Hermitian, with the exception of the identity* Ω_0 (which is Hermitian, of course):

$$\Omega_k^\dagger=-\Omega_k,\quad k=1,2,...,7. \tag{30}$$

This fact enables us to prove that the following ''hermiticity'' property also holds for the ♥ products:

$$(\Omega_\mu\heartsuit\Omega_\nu)^\dagger=\Omega_\nu^\dagger\heartsuit\Omega_\mu^\dagger,\quad\text{for }\ \mu,\nu=0,1,...,7. \tag{31}$$

First, we note that this equality holds trivially for $(\Omega_0\heartsuit\Omega_\mu)^\dagger=\Omega_\mu^\dagger=\Omega_\mu^\dagger\heartsuit\Omega_0^\dagger$. Second, we prove (31) for $j,k\neq 0$ by using (6) and noting that γ_{ij}^k are real and antisymmetric in *j*, *k*, so that

$$(\Omega_j\heartsuit\Omega_k)^\dagger=-\delta_{jk}\Omega_0+\sum_{i=1}^{7}\gamma_{jk}^i\Omega_i^\dagger=-\delta_{kj}\Omega_0+\sum_{i=1}^{7}\gamma_{kj}^i\Omega_i=\Omega_k\heartsuit\Omega_j=\Omega_k^\dagger\heartsuit\Omega_j^\dagger,\quad j,k=1,...,7. \tag{32}$$

The conjugation property (31) of Ω_μ suggests the following formal definition for the *Hermitian conjugate* of the octonionic basis:

$$e_0^\dagger=e_0,\quad e_j^\dagger=-e_j\quad(j=1,2,...,7), \tag{33}$$

whereupon the ''*number operators*'' become equal to the identity element:

$$N_\mu:=e_\mu^\dagger e_\mu=e_\mu e_\mu^\dagger=e_0=1\quad(\text{no summation})\quad(\mu=0,1,...,7). \tag{34}$$

We can now define the *Hermitian conjugate* of the complex octonions *z* in a natural way, by

$$z^\dagger:=\sum_{\mu=0}^{7}\overline{z_\mu}e_\mu^\dagger=\overline{z_0}e_0-\sum_{i=1}^{7}\overline{z_i}e_i\equiv\overline{z_0}-\overline{\mathbf{z}},\quad\text{where }\ z_i\in\mathbb{C}. \tag{35}$$

We then calculate

$$\begin{aligned}zz^\dagger&\equiv(z_0+\mathbf{z})(\overline{z_0}-\overline{\mathbf{z}})=|z_0|^2+(\overline{z_0}\mathbf{z}-z_0\overline{\mathbf{z}})-\mathbf{z}\overline{\mathbf{z}}\\&=\sum_{\mu=0}^{7}|z_\mu|^2-\sum_{k=1}^{7}(z_0\overline{z_k}-z_k\overline{z_0})e_k+\sum_{1\leqslant i<j}(z_i\overline{z_j}-z_j\overline{z_i})e_ie_j\\&=N(z)+2i\sum_{k=i}^{7}\mathfrak{I}\left(\overline{z_0}z_k+\sum_{1\leqslant i<j\leqslant 7}z_i\overline{z_j}\gamma_{ij}^k\right)e_k,\end{aligned} \tag{36}$$

where

$$N(z)=\sum_{\mu=0}^{7}|z_\mu|^2. \tag{37}$$

The definition $N(z)$ is perfectly reasonable for a norm, although the decomposition law (23) is *not* satisfied. We see that the ''space components'' $(zz^\dagger)_i$ of $zz^\dagger$ are pure imaginary. To understand why this is expected on general grounds, it is useful to introduce the concept of a *Hermitian octonion:* $y^\dagger = y$, which signifies that

$$\overline{y_0} = y_0, \quad \overline{y_i} = -y_i \quad (i=1,...,7), \tag{38}$$

so that y_0 must be real and all the space components must be pure imaginary.

Since $zz^\dagger$ is Hermitian by (31), we see that its space components can only be pure imaginary. If we wish to get rid of these components and retain only the zero component, we must add the standard conjugate. Thus, we may define the *Hermitian norm* by

$$N(z) = (zz^\dagger + \overline{zz^\dagger})/2. \tag{39}$$

Hence, if z is mapped into Z, then $z^\dagger$ will be mapped into $Z^\dagger$, which is obtained by the standard Hermitian conjugation of the matrix Z.

One of the main insights gained by using the matrix representation is when we calculate the Hermitian norm. If

$$z \to Z = z^\mu \Omega_\mu = \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix}, \quad \text{then} \quad z^\dagger \to Z^\dagger = \begin{pmatrix} \overline{\alpha} & B^\dagger \\ A^\dagger & \overline{\beta} \end{pmatrix}. \tag{40}$$

The product

$$Z \heartsuit Z^\dagger = \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix} \heartsuit \begin{pmatrix} \overline{\alpha} & B^\dagger \\ A^\dagger & \overline{\beta} \end{pmatrix} = \begin{pmatrix} \alpha\overline{\alpha} + \frac{1}{2}\mathrm{Tr}(AA^\dagger) & \alpha\beta^\dagger + \overline{\beta}A + \dfrac{i}{2}[B,A^\dagger] \\ \overline{\alpha}\beta + \beta A^\dagger - \dfrac{i}{2}[A,B^\dagger] & \beta\overline{\beta} + \frac{1}{2}\mathrm{Tr}(BB^\dagger) \end{pmatrix}. \tag{41}$$

The zero component of $zz^\dagger$ is proportional to the trace of $Z \heartsuit Z^\dagger$, so that the new *Hermitian norm* can be expressed in terms of the representation matrices Z as follows:

$$\begin{aligned} N(z) &= \tfrac{1}{4}\mathrm{Tr}(Z \heartsuit Z^\dagger) = \tfrac{1}{2}(|\alpha|^2 + |\beta|^2) + \tfrac{1}{4}(\mathrm{Tr}(AA^\dagger) + \mathrm{Tr}(BB^\dagger)) \\ &= \frac{1}{2}\left(|\alpha|^2 + |\beta|^2 + \sum_{k=1}^{3} (|a_k|^2 + |b_k|^2) \right) = \sum_{\mu=0}^{7} |z_\mu|^2, \end{aligned} \tag{42}$$

in accordance with (37). For real z_μ we get $\beta \to \overline{\alpha}$, $B \to -A^\dagger$ in (26). Therefore, $AB \to -AA^\dagger = -\mathbf{a}\cdot\boldsymbol{\sigma}\overline{\mathbf{a}}\cdot\boldsymbol{\sigma} = -\mathbf{a}\cdot\overline{\mathbf{a}}$. Thus, the formally defined scalar reduces to a conventional norm:

$$N(z) = \alpha\beta - A\cdot B = |\alpha|^2 + \mathbf{a}\cdot\overline{\mathbf{a}} \to |\alpha|^2 + \sum_{k=1}^{3} |a_k|^2 = \sum_{\mu=0}^{7} x_\mu^2 \equiv n(z) \geqslant 0. \tag{43}$$

D. Nonassociative algebras from Lie algebras

The main advantage of our matrix representation over the Zorn vector representation is that our multiplication rule can be generalized to *any number n of dimensions*, whereas the Zorn rule is restricted, since it is defined in terms of vector product $\mathbf{a}\times\mathbf{b}$, which only applies to 3-vectors!

In particular, given *any* representation of an n-dimensional Lie algebra $\mathbf{g}$ in terms of Hermitian $N\times N$ matrices $\lambda_k (k=1,2,...,n)$, we can then define $2n+2$ different $2N$-dimensional matrices,

$$e_0 \Leftrightarrow \Omega_0 \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad e_k \Leftrightarrow \Omega_k \equiv \begin{pmatrix} 0 & -\lambda_k \\ \lambda_k & 0 \end{pmatrix} \quad (k=1,...,n),$$

$$\hat{e}_0 \Leftrightarrow \Omega_{n+2} \equiv \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \hat{e}_k \Leftrightarrow \hat{\Omega}_k \equiv \begin{pmatrix} 0 & i\lambda_k \\ i\lambda_k & 0 \end{pmatrix}. \tag{44}$$

If we multiply these matrices using the ♥ rule, we end up with a closed algebra, which we shall call the *double algebra* g^D, with the following product rules for their basis elements ($\hat{\Omega}_0 \equiv \Omega_{n+2}$):

$$-\hat{\Omega}_0\Omega_k = \Omega_k\hat{\Omega}_0 = \hat{\Omega}_k, \quad \hat{\Omega}_0\hat{\Omega}_k = -\hat{\Omega}_k\hat{\Omega}_0 = \Omega_k, \quad \hat{\Omega}_0\hat{\Omega}_0 = -\Omega_0,$$

$$\Omega_i\Omega_j = -\delta_{ij}\Omega_0 + f_{ijk}\Omega_k, \quad \hat{\Omega}_i\hat{\Omega}_j = -\delta_{ij}\Omega_0 - f_{ijk}\Omega_k, \tag{45}$$

$$-\hat{\Omega}_j\Omega_i = \Omega_i\hat{\Omega}_j = -\delta_{ij}\hat{\Omega}_0 - f_{ij}\hat{\Omega}_k.$$

Above, the f_{ijk} are the structure constants of the Lie algebra **g**, defined as usual by

$$[L_i, L_j] = i f_{ijk} L_k. \tag{46}$$

The matrices (44) can be regarded as the ♥- matrix representation of the following (nonassociative) abstract algebra:

$$-\hat{e}_0 e_k = e_k \hat{e}_0 = \hat{e}_k, \quad \hat{e}_0\hat{e}_k = -\hat{e}_k\hat{e}_0 = e_k, \quad \hat{e}_0\hat{e}_0 = -e_0, \quad \text{where } \hat{e}_0 \equiv e_{n+2}, \tag{47}$$

$$e_i e_j = -\delta_{ij} e_0 + f_{ijk} e_k, \tag{48}$$

$$\hat{e}_i \hat{e}_j = -\delta_{ij} e_0 - f_{ijk} e_k, \tag{49}$$

$$-\hat{e}_j e_i = e_i \hat{e}_j = -\delta_{ij}\hat{e}_0 - f_{ijk}\hat{e}_k. \tag{50}$$

These rules (47)–(50) can all be summarized by $(\mu,\nu = 0,1,2,...,2n+2)$,

$$e_\mu e_\nu = g_{\mu\nu} e_0 + \sum_{k=1}^{2n+2} \gamma^k_{\mu\nu} e_k, \quad g_{\mu\nu} := \text{diag}(1,-1,...,-1), \quad \gamma^k_{ij} = -\gamma^k_{ji}. \tag{51}$$

We note from (44) that the e_μ, $\mu = 0,1,...,n$ correspond to a *subalgebra* $\mathbf{g}_+$ of $\mathbf{g}^D$. The rules (47) show that the double algebra $\mathbf{g}^D$ is obtained from $\mathbf{g}_+$ simply by adding a new element, called $\hat{e}_0$, and defining the other $\hat{e}_k$. This Lie algebra example then automatically leads us to a more general doubling procedure, which can be applied to *any algebra* and not just to those constructed from Lie algebras. In fact, this doubling idea is exactly the procedure that is known as the *Cayley–Dickson process*, as we shall see below.

IV. DEFORMED MULTIPLICATION AND THE $\mathcal{A}^♥$ ALGEBRA

Begin with the following subset of $2N\times 2N$-matrices:

$$\mathcal{A} := \left\{ \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix} \middle| A, B \in M_{N\times N} \right\}, \tag{52}$$

where the $N\times N$ matrices α and β in the first and fourth quadrants are proportional to unit matrices.

Among these matrices we may define a more general multiplication rule than that given in (18).[11] We shall still denote it by ♥ since it only introduces two complex *deformation parameters* λ_0 and λ (their values will be restricted as we impose further conditions on the subalgebras):

$$X \heartsuit X' \equiv \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & A' \\ B' & \beta' \end{pmatrix} := \begin{pmatrix} \alpha\alpha' + \lambda_0 A\cdot B' & \alpha A' + \beta' A - \lambda[B,B'] \\ \alpha' B + \beta B' + \lambda[A,A'] & \beta\beta' + \lambda_0 B\cdot A' \end{pmatrix}. \tag{53}$$

As before, $[A,B] \equiv AB - BA$ denotes the commutator, but $A\cdot B$ may now be chosen to be any suitable *bilinear map* into an appropriate field F. For example, one might define $A\cdot B$ by $A\cdot B \equiv \mathrm{Tr}(AB)/N$, or if A and B belong to a Lie algebra, then one could take $A\cdot B$ to be the adjoint trace: $A\cdot B := \mathrm{Tr}(\mathrm{ad}\,A\ \mathrm{ad}\,B)$, where ad denotes the adjoint representation.[12]

When $\lambda = 0$ and $\lambda_0 = 1$ the multiplication rule (53) looks *almost* like the usual one for matrices. However, it still yields nonassociativity, since we are replacing matrix products, such as AB, by $A\cdot B$ times the unit matrix. But in any case, it is evident that with the ♥ product the set $\mathcal{A}$ becomes a closed algebra, which we denote by $\mathcal{A}^\heartsuit$.[13]

A. Complex numbers from real

Before continuing, let us consider the simplest example of the above matrices, namely, the case $N=1$. In this circumstance, the matrices A and B become simple *commuting* numbers, a and b. If we specialize further, and choose $\beta = \alpha$ and $b = -a$ to be real, we end up with two-parameter matrices. Their products are

$$X \heartsuit X' \equiv \begin{pmatrix} \alpha & a \\ -a & \alpha \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & a' \\ -a' & \alpha' \end{pmatrix} := \begin{pmatrix} \alpha\alpha' - \lambda_0 a a' & \alpha a' + \alpha' a \\ \alpha' a + \alpha a' & \alpha\alpha' - \lambda_0 a a' \end{pmatrix}, \tag{54}$$

and this is nothing but the multiplication rule of two complex numbers z and z', provided that we set $\lambda_0 = 1$ and identify α and a with the real and imaginary parts of z. Thus, a *subalgebra* of $\mathcal{A}^\heartsuit$ for $N = \lambda_0 = 1$ becomes isomorphic to the complex numbers $\mathbb{C}$:

$$\mathcal{A}^\heartsuit \ni X = \begin{pmatrix} \alpha & a \\ -a & \alpha \end{pmatrix} \Leftrightarrow z \equiv \alpha + i a \in \mathbb{C}. \tag{55}$$

B. Simple and Hermitian conjugates

The attractive feature of the generalization (53) is that most results and definitions needed for octonions apply almost automatically to $\mathcal{A}^\heartsuit$. For example, for every element $X \in \mathcal{A}^\heartsuit$ we can define a conjugate element $\overline{X}$, as follows:

$$\overline{X} = \overline{\begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix}} := \begin{pmatrix} \beta & -A \\ -B & \alpha \end{pmatrix}. \tag{56}$$

By substituting $A' \to -A$, $B' \to -B$, $\alpha' \to \beta$, and $\beta' \to \alpha$ in (53), we get immediately

$$X \heartsuit \overline{X} = \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix} \heartsuit \begin{pmatrix} \beta & -A \\ -B & \alpha \end{pmatrix} = (\alpha\beta - \lambda_0 A\cdot B) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \equiv n(X) I_{2N\times 2N}, \tag{57}$$

where $n(X) \in \mathbb{C}$. In the meantime, we should again look upon $n(X)$ simply as a scalar function, defined by the map $\mathcal{A}^\heartsuit \to \mathbb{C}$ in (58). Later, we shall study the conditions on $\mathcal{A}^\heartsuit$ under which $n(X)$ becomes a norm.

V. SUBALGEBRAS OF $\mathcal{A}^\heartsuit$

The algebra $\mathcal{A}^\heartsuit$ has several interesting subalgebras.

(i) An obvious subalgebra is the one obtained by choosing both matrices A and B to be traceless:

$$\mathcal{A}_0^{\heartsuit} := \left\{ \begin{pmatrix} \alpha & A \\ B & \beta \end{pmatrix} \middle| \mathrm{Tr}\, A = \mathrm{Tr}\, B = 0 \right\}. \tag{58}$$

(ii) This subalgebra has, in turn, another subalgebra $\mathcal{A}_A^{\heartsuit} \subset \mathcal{A}_0^{\heartsuit}$, in which A and B become antisymmetric matrices.

(iii) A third subalgebra, which we denote by $\mathcal{A}_+^{\heartsuit}$, is obtained by choosing $\beta = \alpha$ and $B = -A$:

$$\mathcal{A}_+^{\heartsuit} := \left\{ \begin{pmatrix} \alpha & A \\ -A & \alpha \end{pmatrix} \right\}. \tag{59}$$

It is easily verified that products of such matrices stay in the same class:

$$\begin{aligned} X \heartsuit X' &= \begin{pmatrix} \alpha & A \\ -A & \alpha \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & A' \\ -A' & \alpha' \end{pmatrix} \\ &= \begin{pmatrix} \alpha\alpha' - \lambda_0 A \cdot A' & \alpha A' + \alpha' A - \lambda [A, A'] \\ -\alpha A' - \alpha' A + \lambda [A, A'] & \alpha\alpha' - \lambda_0 A \cdot A' \end{pmatrix} \in \mathcal{A}_+^{\heartsuit}. \end{aligned} \tag{60}$$

Moreover, $\mathcal{A}_+^{\heartsuit}$ has an interesting property.

Proposition 1: The subalgebra $\mathcal{A}_+^{\heartsuit}$ is flexible for all matrices A.

To put this result into perspective, we note that all Abelian or anticommutative algebras are flexible; thus, if $yx = \pm xy$, then $x(yx) = \pm (yx)x = (xy)x$, so that $(x,y,x) = 0$. Therefore, it is of interest to show that $\mathcal{A}_+^{\heartsuit}$, which is neither Abelian nor anticommutative, is also flexible.

Proof: We shall prove the above assertion by explicit multiplication. However, to simplify the calculations we first note that the multiples of unity added to each element do not affect the associators:

$$(X + \alpha 1, Y + \beta 1, Z + \gamma 1) = (X, Y, Z), \tag{61}$$

where 1 is the identity matrix. This follows immediately from the linearity of associators:

$$(X + \alpha 1, Y, Z) = (X, Y, Z) + \alpha (1, Y, Z) = (X, Y, Z). \tag{62}$$

The property (61) is helpful for calculating associators of the subalgebra $\mathcal{A}_+^{\heartsuit}$, since we can set the α's equal to zero, when calculating the associators.

We now calculate explicitly the associator (X_1, X_2, X_3) for general matrices from $\mathcal{A}_+^{\heartsuit}$, but using only those with $\alpha_i = 0$, i.e.,

$$X_i = \begin{pmatrix} 0 & A_i \\ -A_i & 0 \end{pmatrix} \in \mathcal{A}_+^{\heartsuit}, \quad \text{for } i = 1,2,3. \tag{63}$$

We get

$$(X_1, X_2, X_3) \equiv (X_1 X_2) X_3 - X_1 (X_2 X_3) = \begin{pmatrix} p & P \\ -P & p \end{pmatrix} - \begin{pmatrix} q & Q \\ -Q & q \end{pmatrix} = \begin{pmatrix} p - q & P - Q \\ Q - P & p - q \end{pmatrix}, \tag{64}$$

where

$$p = \lambda \lambda_0 ([A_1, A_2] \cdot A_3), \tag{65}$$

$$P=-\lambda_0(A_1\cdot A_2)A_3+\lambda^2[[A_1,A_2],A_3] \tag{66}$$

and

$$q=\lambda\lambda_0(A_1\cdot[A_2,A_3]), \tag{67}$$

$$Q=-\lambda_0(A_2\cdot A_3)A_1+\lambda^2[A_1,[A_2,A_3]]. \tag{68}$$

Therefore, the elements of the associator (X_1,X_2,X_3) are

$$p-q=\lambda\lambda_0([A_1,A_2]\cdot A_3-A_1\cdot[A_2,A_3])=0, \tag{69}$$

$$\begin{aligned}P-Q&=-\lambda_0((A_1\cdot A_2)A_3-(A_2\cdot A_3)A_1)+\lambda^2([[A_1,A_2],A_3]-[A_1,[A_2,A_3]])\\&=-\lambda_0((A_1\cdot A_2)A_3-(A_2\cdot A_3)A_1)+\lambda^2[A_2,[A_3,A_1]].\end{aligned} \tag{70}$$

In other words, *the associator* (X,Y,X) *vanishes identically, for any* λ, λ_0, $A_1=A_3$ and A_2,

$$(X,Y,X)=0,\quad \text{for } X,Y\in\mathcal{A}_+^{\heartsuit}. \tag{71}$$

□

(iv) As a fourth subalgebra, let $\mathbf{g}$ be a given Lie algebra of dimension n, and let $V_{\mathbf{g}}$ be the algebra spanned by the representation matrices of $\mathbf{g}$. Then, we can define a subalgebra of $\mathcal{A}^{\heartsuit}$ via

$$\mathbf{g}^D:=\left\{\begin{pmatrix}\alpha & A\\ B & \beta\end{pmatrix}\middle| A,B\in V_{\mathbf{g}}\right\}. \tag{72}$$

Clearly the off-diagonal elements, such as $\alpha A'+\beta' A+\lambda[B,B']$, of the products, $X\heartsuit X'$ belong to $V_{\mathbf{g}}$. Hence, $\mathbf{g}^D$ are subalgebras of $\mathcal{A}_0$. Moreover, half of $\mathbf{g}^D$, obtained by the intersection of $\mathbf{g}^D$ with $\mathcal{A}_+^{\heartsuit}$, will be a subalgebra of $\mathbf{g}^D$:

$$\mathbf{g}_0^D:=\left\{\begin{pmatrix}\alpha & A\\ -A & \alpha\end{pmatrix}\middle| A\in V_{\mathbf{g}}\right\}\subset\mathbf{g}^D\subset\mathcal{A}_+^{\heartsuit}. \tag{73}$$

The commutators of the elements of $\mathbf{g}_0^D$ constitute a Lie algebra, which is isomorphic to the original algebra $\mathbf{g}$.

VI. GRADING OF $\mathcal{A}^{\heartsuit}$

Proposition 2: The algebra $\mathcal{A}^{\heartsuit}$ *can be graded, as follows:*

$$\mathcal{A}^{\heartsuit}=\mathcal{A}_+^{\heartsuit}\oplus\mathcal{A}_-^{\heartsuit}=\mathcal{A}_+^{\heartsuit}\oplus K\mathcal{A}_+^{\heartsuit}=A_+^{\heartsuit}\oplus K\heartsuit\mathcal{A}_+^{\heartsuit}, \tag{74}$$

where the ''grading matrix'' is

$$K\equiv\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{75}$$

Observe that $K\heartsuit X=KX$ for any $X\in A^{\heartsuit}$. Also, of course,

$$\mathcal{A}_{\eta}^{\heartsuit}\heartsuit\mathcal{A}_{\eta'}^{\heartsuit}\subseteq\mathcal{A}_{\eta\eta'}^{\heartsuit}. \tag{76}$$

Proof: Every matrix $X\in\mathcal{A}^{\heartsuit}$ can be decomposed, as follows:

$$X\equiv\begin{pmatrix}\alpha & A\\ B & \beta\end{pmatrix}=\begin{pmatrix}\alpha_+ & A_+\\ -A_+ & \alpha_+\end{pmatrix}+\begin{pmatrix}\alpha_- & A_-\\ A_- & -\alpha_-\end{pmatrix}\equiv X_++\hat{X}_- \tag{77}$$

$$\equiv X_+ + KX_- , \tag{78}$$

where

$$\alpha_\pm \equiv \frac{1}{2}(\alpha \pm \beta), \quad A_\mp \equiv \frac{1}{2}(A \pm B), \quad K \equiv \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{79}$$

The first set of matrices (with $\beta = \alpha$ and $B = -A$) constitutes the subalgebra $\mathcal{A}_+^\heartsuit$, which we defined earlier in (59). The second set of matrices (with $\beta = -\alpha$ and $B = A$) will be called $\mathcal{A}_-^\heartsuit$. Since $\mathcal{A}_+^\heartsuit$ is a subalgebra of $\mathcal{A}^\heartsuit$, clearly $\mathcal{A}_+^\heartsuit \mathcal{A}_+^\heartsuit = \mathcal{A}_+^\heartsuit$. The rest of the inclusion relations (76), namely,

$$\hat{X} \heartsuit \hat{X}' \in \mathcal{A}_+^\heartsuit , \quad X \heartsuit \hat{X}' \in \mathcal{A}_-^\heartsuit , \quad \hat{X} \heartsuit X' \in \mathcal{A}_-^\heartsuit , \tag{80}$$

follow immediately from the equalities (85)–(87), which we shall prove below. □

Proposition 3: The following equalities hold for any $X, X' \in \mathcal{A}_+^\heartsuit$:

$$KXK = \overline{X}, \tag{81}$$

$$(KX) \heartsuit (KX') = X' \heartsuit \overline{X}, \tag{82}$$

$$X \heartsuit (KX') = K(\overline{X} \heartsuit X'), \tag{83}$$

$$(KX) \heartsuit X' = K(X' \heartsuit X). \tag{84}$$

Proof: The proof follows simply by explicit matrix multiplication, using (60):

$$KX \heartsuit KX' = \begin{pmatrix} \alpha & A \\ A & -\alpha \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & A' \\ A' & -\alpha' \end{pmatrix} = \begin{pmatrix} \alpha\alpha' + \lambda_0 A \cdot A' & \alpha A' - \alpha' A - \lambda[A,A'] \\ -\alpha A' + \alpha' A + \lambda[A,A'] & \alpha\alpha' + \lambda_0 A \cdot A' \end{pmatrix}$$

$$= \begin{pmatrix} \alpha' & A' \\ -A' & \alpha' \end{pmatrix} \heartsuit \begin{pmatrix} \alpha & -A \\ A & \alpha \end{pmatrix} \equiv X' \heartsuit \overline{X} \in \mathcal{A}_+^\heartsuit , \tag{85}$$

$$X \heartsuit (KX') = \begin{pmatrix} \alpha & A \\ -A & \alpha \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & A' \\ A' & -\alpha' \end{pmatrix} = \begin{pmatrix} \alpha\alpha' + \lambda_0 A \cdot A' & \alpha A' - \alpha' A + \lambda[A,A'] \\ \alpha A' - \alpha' A + \lambda[A,A'] & -\alpha\alpha' - \lambda_0 A \cdot A' \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \left\{ \begin{pmatrix} \alpha & -A \\ A & \alpha \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & A' \\ -A' & \alpha' \end{pmatrix} \right\} = K(\overline{X} \heartsuit X') \equiv \widehat{\overline{X} \heartsuit X'} \in \mathcal{A}_-^\heartsuit , \tag{86}$$

$$KX \heartsuit X' = \begin{pmatrix} \alpha & A \\ A & -\alpha \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & A' \\ -A' & \alpha' \end{pmatrix} = \begin{pmatrix} \alpha\alpha' - \lambda_0 A \cdot A' & \alpha A' - \alpha' A + \lambda[A,A'] \\ \alpha A' - \alpha' A + \lambda[A,A'] & -\alpha\alpha' - \lambda_0 A \cdot A' \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \left\{ \begin{pmatrix} \alpha' & A' \\ -A' & \alpha' \end{pmatrix} \heartsuit \begin{pmatrix} \alpha & A \\ -A & \alpha \end{pmatrix} \right\} = K(X' \heartsuit X) \in \mathcal{A}_-^\heartsuit . \tag{87}$$

A. Matrix representation of the Cayley–Dickson process

If we multiply the grading matrix K by a real or complex scalar v, and let $\mu \equiv v^2$, we get

$$v_{\text{op}} := vK, \quad v_{\text{op}} v_{\text{op}} := v^2 1 = \mu 1. \tag{88}$$

Therefore, using the relations (81)–(84), we get the multiplication rule

$$(X_1 + v_{\text{op}} X_2) \heartsuit (X_3 + v_{\text{op}} X_4) = (X_1 \heartsuit X_3 + \mu X_4 \heartsuit \overline{X_2}) + v_{\text{op}}(\overline{X_1} \heartsuit X_4 + X_3 \heartsuit X_2), \quad \forall X_i \in \mathcal{A}_+^{\heartsuit}. \tag{89}$$

This is exactly the multiplication rule given by Cayley and Dickson, where one starts with an abstract algebra $\mathcal{B}$ and defines an abstract operator v_{op}, and essentially *postulates* the following multiplication rule:[8]

$$(b_1 + v_{\text{op}} b_2)(b_3 + v_{\text{op}} b_4) = (b_1 b_3 + \mu b_4 \overline{b_2}) + v_{\text{op}}(\overline{b_1} b_4 + b_3 b_2), \quad b_i \in \mathcal{B}, \tag{90}$$

where $\overline{b_i} \in \mathcal{B}$ is the conjugate of b_i, and $v_{\text{op}} \notin \mathcal{B}$, such that $v_{\text{op}}^2 = \mu \cdot 1$.

Observe that the $\heartsuit$ multiplication rule provides an explicit matrix representation of the Cayley–Dickson process,[8] provided that the original algebra $\mathcal{B}$ can be represented by $\mathcal{A}_+^{\heartsuit}$.

B. Composition algebras from 2×2 matrices

One may wonder what happens if we allow the rudimentary 2×2 matrices to contain arbitrary complex elements. Since

$$X \heartsuit X' \equiv \begin{pmatrix} \alpha & a \\ b & \beta \end{pmatrix} \heartsuit \begin{pmatrix} \alpha' & a' \\ b' & \beta' \end{pmatrix} := \begin{pmatrix} \alpha\alpha' + \lambda_0 a b' & \alpha a' + \beta' a \\ \alpha' b + \beta b' & \beta b' + \lambda_0 b a' \end{pmatrix}, \tag{91}$$

when $X' = \overline{X}$ this product yields a ''norm,''

$$X * \overline{X} = n(X) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{where } n(x) = \alpha\beta - \lambda_0 ab. \tag{92}$$

It is easy to check, by explicit multiplication, that the following identity holds for any $\lambda_0 \in \mathbb{C}$:

$$\begin{aligned} n(x)n(x') &= (\alpha\beta - \lambda_0 ab)(\alpha'\beta' - \lambda_0 a'b') \\ &= (\alpha\alpha' + \lambda_0 ab')(\beta\beta' + \lambda_0 ba') - \lambda_0(\alpha a' + \beta' a)(\alpha' b + \beta b') = n(xx'), \end{aligned} \tag{93}$$

which informs us that the standard norm (92) for $N=1$ obeys the composition law.

Clearly, the norm (92) is degenerate for any $\lambda_0 \in \mathbb{C}$, if we allow X to be any 2×2 matrix. [For example, simply choosing $\beta = a = b = 1$ and $\alpha = \lambda_0$ will yield an $x \neq 0$ with $n(x) = \alpha\beta - \lambda_0 ab = 0$.] The question then arises, when is the norm $n(x)$ in (92) nondegenerate? We can certainly guarantee that $n(x)$ is nondegenerate, if we restrict X to have the following special form:

$$X = \begin{pmatrix} z & w \\ -\overline{w} & \overline{z} \end{pmatrix}, \quad \text{and} \quad \Re\lambda_0 > 0, \tag{94}$$

whence

$$n(x) = |z|^2 + \lambda_0 |w|^2 \neq 0, \quad \text{for } \Re\lambda_0 \neq 0. \tag{95}$$

[Of course, there exist a few equivalent variations of the conditions (94). For instance, we can replace $-\overline{w}$ by $\overline{w}$, but demand that $\Re\lambda_0 < 0$.]

Anyhow, this means that we are dealing with a division algebra, which must therefore be one of the four possibilities. Because $X \in M_{2\times 2}$, we may expand it in terms of Pauli matrices, getting

$$X=\begin{pmatrix} z & w \\ -\overline{w} & \overline{z} \end{pmatrix}=z_1\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}+iw_2\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}+iw_1\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}+iz_2\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$\equiv x_0\sigma_0 - i\mathbf{x}\cdot\boldsymbol{\sigma} \leftrightarrow x_0 e_0 + \sum_{k=1}^{3} x_k e_k, \quad \text{where } w_i, z_i \in \mathbb{R}. \tag{96}$$

Since $e_0 \rightarrow \sigma_0$ and $e_j \rightarrow -i\sigma_j$ $j=1,2,3$ are known representations of the quaternions, we conclude that this is the quaternion algebra over the *real* field $\mathbb{R}$, as expected. Indeed, the matrix (96) is the usual representation of $\mathcal{Q}$ in terms of standard matrices. Later on we shall describe another representation by nonstandard matrices.

VII. CONDITIONS ON DEFORMATION PARAMETERS

Previously we showed that $\mathcal{A}_+^{\heartsuit}$ is flexible. We now ask under what conditions $\mathcal{A}_+^{\heartsuit}$ can become alternative.

For this, we must have $(X_1, X_1, X_3)=0$. By setting $X_2 = X_1$ and noting Eq. (70), we get the condition

$$[A_1,[A_3,A_1]] = \frac{\lambda^2}{\lambda_0} = ((A_1\cdot A_1)A_3 - (A_1\cdot A_3)A_1). \tag{97}$$

This condition can be satisfied if $\lambda^2 = \lambda_0/4$ and $A_i = \mathbf{a}_i\cdot\boldsymbol{\sigma}$: Indeed, for such A_i, we get

$$\begin{aligned}[A_2,[A_3,A_1]] &= [\mathbf{a}_2\cdot\boldsymbol{\sigma}, 2i[\mathbf{a}_3\times\mathbf{a}_1]\cdot\boldsymbol{\sigma}] = -4[\mathbf{a}_2\times[\mathbf{a}_3\times\mathbf{a}_1]]\cdot\boldsymbol{\sigma} \\ &= 4((\mathbf{a}_2\cdot\mathbf{a}_3)\mathbf{a}_1 - (\mathbf{a}_2\cdot\mathbf{a}_1)\mathbf{a}_3)\cdot\boldsymbol{\sigma} \\ &= 4((A_1\cdot A_2)A_3 - (A_2\cdot A_3)A_1), \end{aligned} \tag{98}$$

where we used $(A\cdot B)=\frac{1}{2}\mathrm{Tr}(AB)$. Hence, we can have $\lambda = \pm\frac{1}{2}\sqrt{\lambda_0}$. For the special choice $\lambda_0 = -1$, we get $\lambda = \pm 1/2$.

This sign ambiguity is the origin of the nonuniqueness of the ♥ product.[14]

VIII. SUMMARY

Our algebras provide concrete the matrix representation of a big class of nonassociative algebras. They may suggest new constructions in the future. One such possibility, which leads to notions of triality, is described in Appendix A, but anyhow the formulation (18) and its deformations (53) permit generalizations that have an obvious affinity to higher symmetry groups, rather than the simple case of SU(2). We also believe that our treatment of hermiticity and norm for the complex case is reasonable; we illustrate their utility with reference to the Dirac equation in Appendix B.

ACKNOWLEDGMENTS

We are extremely grateful to Dr. G. Joshi for supplying us with a comprehensive list of references in this vast topic and for pointing us in the right direction, and to Dr. B. Gardner for useful comments.

APPENDIX A: THE ◇ PRODUCT

In this appendix we try another type of product, which we denote by ◇, where the commutators $[B,B']$ in the ♥ product are now replaced by the standard matrix products BB'.

Let us first consider the simplest case, $N=1$, where the matrices A and B become scalars, so that we shall first deal with 2×2 matrices:

$$X\equiv\begin{pmatrix}\alpha & a\\ b & \beta\end{pmatrix}. \tag{A1}$$

We define the new matrix product, as follows:

$$X\diamondsuit X'\equiv\begin{pmatrix}\alpha & a\\ b & \beta\end{pmatrix}\diamondsuit\begin{pmatrix}\alpha' & a'\\ b' & \beta'\end{pmatrix}:=\begin{pmatrix}\alpha\alpha'+\lambda_0 ab & \alpha a'+\beta' a+\lambda bb'\\ \alpha' b+\beta b'+\eta\lambda aa' & \beta\beta'+\lambda_0 ba'\end{pmatrix}, \tag{A2}$$

where η, λ, λ_0 are arbitrary complex numbers. We now ask the question, whether for such a product, we can define for every X a conjugate $\overline{X}$, such that $X\diamondsuit\overline{X}=n(X)\cdot 1$, where $n(X)$ is some quadratic form of X, i.e., $n(sX)=s^2n(X)$.

Let us try the following ansatz:

$$\overline{X}\equiv\begin{pmatrix}\beta+\gamma & -a\\ -b & \alpha+\delta\end{pmatrix}. \tag{A3}$$

We want to determine γ and δ, and derive conditions on α, β, a,b, by demanding that $X\diamondsuit\overline{X}\propto 1$:

$$X\diamondsuit\overline{X}=\begin{pmatrix}\alpha & a\\ b & \beta\end{pmatrix}\diamondsuit\begin{pmatrix}\beta+\gamma & -a\\ -b & \alpha+\delta\end{pmatrix}=\begin{pmatrix}\alpha(\beta+\gamma)-\lambda_0 ab & \delta a-\lambda b^2\\ \gamma b-\eta\lambda a^2 & \beta(\alpha+\delta)-\lambda_0 ab\end{pmatrix}$$
$$=n(X)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},\quad\text{where}\quad n(X)=\alpha(\beta+\gamma)-\lambda_0 ab. \tag{A4}$$

This condition is obeyed, if

$$\delta a-\lambda b^2=0,\quad \gamma b-\eta\lambda a^2=0,\quad\text{and}\quad \alpha\gamma=\beta\delta. \tag{A5}$$

We cannot satisfy these conditions for general a and b. (For example, if $a=0$ and $b\neq 0$, then δ would be infinite.) Thus, for $\eta\neq 0$ we must assume that either both a and b are zero or both are unequal to zero. But for $\eta=0$ we must demand $b=0$. With these restrictions, we get

$$\delta(X)=\lambda\frac{b^2}{a},\quad \gamma(X)=\eta\lambda\frac{a^2}{b},\quad\text{and}\quad \frac{\beta}{\alpha}=\frac{\gamma}{\delta}=\eta\frac{a^3}{b^3}. \tag{A6}$$

An additional condition can be obtained by demanding that the adjoint operation is an involution, so that $\bar{\bar{X}}=X$, whence

$$\bar{\bar{X}}=\begin{pmatrix}\alpha+\delta(X)+\gamma(\overline{X}) & a\\ b & \beta+\gamma(X)+\delta(\overline{X})\end{pmatrix}$$

$$=\begin{pmatrix}\alpha+\lambda\left(\frac{b^2}{a}-\eta\frac{a^2}{b}\right) & a\\ b & \beta+\lambda\left(\eta\frac{a^2}{b}+\frac{b^2}{-a}\right)\end{pmatrix}$$

$$=\begin{pmatrix}\alpha & a\\ b & \beta\end{pmatrix}=X. \tag{A7}$$

Thus, we get the new condition

$$b^3=\eta a^3, \quad \text{or} \quad b=\xi a, \quad \text{where} \quad \xi=\eta^{1/3}. \tag{A8}$$

Note that for each η we have three cubic roots ξ. Substituting (A8) into the third equality in (A6), we get

$$\alpha=\beta \quad \text{and} \quad \gamma=\delta. \tag{A9}$$

Hence, a can be any complex number as long as $b=\xi a$. Finally, by noting all the above conditions, we get for every $\eta\in\mathbb{C}$ three sets of 2×2 matrices, which are closed algebras under the product $\diamondsuit$:

$$X(\xi)=\left\{\begin{pmatrix}\alpha & a\\ \xi a & \alpha\end{pmatrix}\right\}, \quad \xi=\eta^{1/3}\in\mathbb{C}. \tag{A10}$$

For these matrices, the adjoint and the corresponding quadratic form are

$$\overline{X}\equiv\begin{pmatrix}\alpha+\xi^2\lambda\alpha & -a\\ -\xi a & \alpha+\xi^2\lambda a\end{pmatrix}, \tag{A11}$$

and

$$n(X)=\alpha(\beta+\gamma)-\lambda_0 ab=\alpha^2+\eta\lambda\alpha\frac{a^2}{b}-\lambda_0 ab=\alpha^2+\xi^2\lambda\alpha a-\xi\lambda_0 a^2. \tag{A12}$$

Note that $n(X)$ is quadratic in X, i.e., $N(sX)=s^2N(X)$, $s\in\mathbb{C}$.

For the special case $b=-a$, or $\xi=-1$, we obtain a known quadratic form,[11]

$$n(X)=\alpha^2+\lambda\alpha a+\lambda_0 a^2. \tag{A13}$$

Proceeding to larger matrices, let

$$X\diamondsuit X'\equiv\begin{pmatrix}\alpha & A\\ B & \beta\end{pmatrix}\diamondsuit\begin{pmatrix}\alpha' & A'\\ B' & \beta'\end{pmatrix}:=\begin{pmatrix}\alpha\alpha'+\lambda_0 A\cdot B' & \alpha A'+\beta' A+\lambda BB'\\ \alpha' B+\beta B'+\eta\lambda AA' & \beta\beta'+\lambda_0 B\cdot A'\end{pmatrix}, \quad \eta\in\mathbb{C}. \tag{A14}$$

One can readily check that such matrices yield closed algebras with respect to the above $\diamondsuit$ product. By restricting B to be ξA, we get the following subalgebra:

$$X(\xi)=\left\{\begin{pmatrix}\alpha & A\\ \xi A & \alpha\end{pmatrix}\right\}, \quad \xi=\eta^{1/3}\in\mathbb{C}. \tag{A15}$$

We can check, using (A14), that products of two such matrices yield a matrix of the same type:

$$X \diamondsuit X' \equiv \begin{pmatrix} \alpha & A \\ \xi A & \alpha \end{pmatrix} \diamondsuit \begin{pmatrix} \alpha' & A' \\ \xi A' & \alpha' \end{pmatrix} = \begin{pmatrix} \alpha\alpha' + \xi\lambda_0 A\cdot A' & \alpha A' + \alpha' A + \xi^2\lambda A A' \\ \xi(\alpha' A + \alpha A') + \eta\lambda A A' & \alpha\alpha' + \xi\lambda_0 A\cdot A' \end{pmatrix}. \tag{A16}$$

However, if we replace the *scalar a* in Eqs. (A11) and (A12) by a *matrix A*, we do not get an adjoint nor a bilinear form, since the appropriate items do not stay scalar, as they should.

Finally, we note that if we replace the simple products AA' in (A16) by anticommutators $\{A,A'\}/2$, and if we make the scalar products $A\cdot A'$ symmetric, i.e.,

$$\begin{pmatrix} \alpha & A \\ \xi A & \alpha \end{pmatrix} \diamondsuit_S \begin{pmatrix} \alpha' & A' \\ \xi A' & \alpha' \end{pmatrix} := \begin{pmatrix} \alpha\alpha' + \xi\lambda_0 A\cdot A' & \alpha A' + \alpha' A + \xi^2\lambda\{A,A'\}/2 \\ \xi(\alpha' A + \alpha A') + \eta\lambda\{A,A'\}/2 & \alpha\alpha' + \xi\lambda_0 A\cdot A' \end{pmatrix},$$

then the product becomes Abelian. Therefore the new algebra will automatically become *flexible*. One can similarly symmetrize the more general product (112) by defining, $X \diamondsuit_S X' := (X \diamondsuit X' + X' \diamondsuit X)/2$, and also get a flexible algebra.

APPENDIX B: APPLICATION TO THE DIRAC EQUATION

In momentum space, the free Dirac equation reads as

$$P\Psi \equiv (p_0 - \mathbf{p}\cdot\boldsymbol{\alpha} - m\beta)\Psi = \begin{pmatrix} p_0 - m & -\mathbf{p}\cdot\boldsymbol{\sigma} \\ -\mathbf{p}\cdot\boldsymbol{\sigma} & p_0 + m \end{pmatrix}\Psi = 0, \tag{B1}$$

and it can be rewritten in terms of octonions as follows:

$$p\psi \equiv (p_0 + i\mathbf{p}\cdot\hat{\mathbf{e}} + ime_4)\psi = 0, \tag{B2}$$

where we have used the correspondence

$$\alpha_k = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0 \end{pmatrix} = -i\hat{\Omega}_k \Leftrightarrow -i\hat{e}_k, \quad \beta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = -i\Omega_4 \Leftrightarrow -ie_4. \tag{B3}$$

Notice that we cannot write the Dirac equation in terms of quaternions alone, since we require five different basis elements: $(e_0, e_4, \hat{e}_k;\ k=1,2,3)$ or $(e_0, e_4, e_k;\ k=1,2,3)$.

The octonion p is nicely Hermitian $(p = p^\dagger)$ and has zero norm:

$$p\overline{p} = n(p) = p_0^2 - \mathbf{p}^2 - m^2 = 0. \tag{B4}$$

Therefore the solution ψ of the *Dirac octonionic equation* (B2) is elegantly given by the standard conjugate of p, namely,

$$\psi = \overline{p} \equiv p_0 - i\mathbf{p}\cdot\hat{\mathbf{e}} - ime_4 \Leftrightarrow \Psi = \overline{P} = \begin{pmatrix} p_0 + m & \mathbf{p}\cdot\boldsymbol{\sigma} \\ \mathbf{p}\cdot\boldsymbol{\sigma} & p_0 - m \end{pmatrix}. \tag{B5}$$

The first and second columns of Ψ are proportional to the positive energy solutions $u^1(p)$ and $u^2(p)$, while the third and fourth columns yield the negative energy solutions $v^1(p)$ and $v^2(p)$, if we replace m by $-m$, because

$$(p_0 - \mathbf{p}\cdot\boldsymbol{\alpha} - m\beta)u^i(p) = 0, \quad (p_0 - \mathbf{p}\cdot\boldsymbol{\alpha} + m\beta)v^i(p) = 0 \quad (i=1,2). \tag{B6}$$

Therefore the *physical* and *normalizable* solutions can be expressed as

$$\Psi_p \equiv (u^1|u^2|v^1|v^2) = \frac{1}{\sqrt{2m(p_0+m)}}\begin{pmatrix} p_0+m & \mathbf{p}\cdot\boldsymbol{\sigma} \\ \mathbf{p}\cdot\boldsymbol{\sigma} & p_0+m \end{pmatrix} \Leftrightarrow \frac{p_0+m-i\mathbf{p}\cdot\hat{\mathbf{e}}}{\sqrt{2m(p_0+m)}}, \tag{B7}$$

ensuring that

$$\overline{u}^{\dagger}(p)u^j(p) \equiv u^{i\dagger}(p)\beta u^j(p) = \delta_{ij} \quad \text{and} \quad \overline{v}^{\dagger}(p)v^j(p) \equiv v^{i\dagger}(p)\beta v^j(p) = -\delta_{ij} \quad (i,j=1,2). \tag{B8}$$

[1] D. Finkelstein, J. M. Jauch, S. Shiminovitch, and D. Speiser, J. Math. Phys. **3**, 207 (1962); **4**, 136 (1963); **4**, 788 (1963); G. G. Emch, Helv. Phys. Acta **36**, 739 (1963); L. P. Horwitz and L. C. Biedenharn, *ibid.* **38**, 385 (1965); Ann. Phys. (N.Y.) **157**, 432 (1984); M. Gunaydin, C. Piron, and H. Ruegg, Commun. Math. Phys. **61**, 69 (1978); C. G. Nash and G. C. Joshi, J. Math. Phys. **28**, 2883 (1987); **28**, 2886 (1987); S. de Leo and P. Rotelli, Prog. Theor. Phys. **92**, 917 (1994).

[2] F. Gursey, *Proceedings of ''Symmetries in Physics (1600–1980)''* Sant Feliu de Guixols, 1992; S. de Leo, J. Math. Phys. **37**, 2955 (1996). These articles contain reviews of this topic.

[3] F. Gursey, *Quaternion Methods in Field Theory* John Hopkins Workshop, QCD 161, J55, 1980; K. Morita, Prog. Theor. Phys. **67**, 1860 (1982); S. L. Adler, Phys. Rev. Lett. **55**, 783 (1985); Commun. Math. Phys. **104**, 611 (1986); C. Nash and G. C. Joshi, J. Math. Phys. **28**, 463 (1987); Int. J. Theor. Phys. **27**, 409 (1988); S. L. Adler, *Quaternionic Quantum Mechanics and Quantum Fields* (Oxford University Press, Oxford, 1995).

[4] F. Gursey, ''Algebraic methods and quark structure,'' in *Proceedings of the Kyoto Symposium*, January 1975; F. Gursey, ''Octonionic structure in particle physics,'' *Proceedings of ICGTMP78*, Berlin, 1979, p. 509; T. Kugo and P. Townsend, Nucl. Phys. B **221**, 357 (1983); A. Sudbery and J. Phys. A **12**, 939 (1984); P. S. Howe and P. C. West, Int. J. Mod. Phys. A **7**, 6639 (1992); S. de Leo, Int. J. Theor. Phys. **35**, 1821 (1996); F. Gursey and C. H. Tze, *On the Role of Division, Jordan and Related Algebras in Particle Physics* (World Scientific, Singapore, 1996).

[5] F. Gursey, ''Color, quarks and octonions,'' John Hopkins Workshop, January, 1974; S. L. Adler, Phys. Rev. D **21**, 2903 (1980); S. Catto and F. Gursey, Nuovo Cimento A **99**, 685 (1988); S. de Leo and P. Rotelli, J. Phys. G **22**, 1137 (1996).

[6] A. J. Davies and G. C. Joshi, J. Math. Phys. **27**, 3036 (1986); P. Rotelli, Mod. Phys. Lett. A **4**, 933 (1989); S. de Leo and P. Rotelli, *ibid.* **11**, 357 (1996); Int. J. Theor. Phys. **37**, 1511 (1998).

[7] M. Zorn, Abh. Mat. Sem. Hamburg **9**, 395 (1933); R. D. Schafer, Am. J. Math. **76**, 435 (1954); G. Moreno, Bol. Soc. Mat. Mexicana **4**, 13 (1998).

[8] R. D. Schafer, *An Introduction to Non-Associative Algebras* (Academic, New York, 1966); H. P. Peterson, Abh. Mat. Sem. Hamburg **9**, 35 (1971); S. Okubo, *Introduction to Octonions and Other Non-Associative Algebras in Physics* (Cambridge University Press, Cambridge, 1995).

[9] The complete antisymmetry of ϵ_{ijkl} in its first three indices is an immediate consequence of the fact that $\mathcal{O}$ is an alternative algebra, such algebras being defined by the two ''*alternative conditions*.'' $x^2y = x(xy)$, and $(yx)x = yx^2$, $\forall x,y \in \mathcal{A}$. These conditions can be expressed in terms of associators as $(x,x,y) = (y,x,x) = 0$, $\forall x,y \in \mathcal{A}$. They enable one to prove that the associator (x,y,z) of an alternative algebra is totally antisymmetric in its three arguments. The proof is based on the ''*linearization technique*,'' which consists of replacing one or more of the arguments, say x, by $x + \lambda w$. By these means we get $(x+\lambda w, x+\lambda w, y) = (x,x,y) + \lambda^2(w,w,y) + \lambda(w,x,y) + \lambda(x,w,y) = \lambda((w,x,y) + (x,w,y)) = 0$, so that $(w,x,y) = -(x,w,y)$. One can similarly prove the antisymmetry among the other arguments. In particular, the total antisymmetry yields $(x,y,x) = 0$, or $(z,y,x) + (x,y,z) = 0$, where the second equality follows the first, again by linearization. Algebras that obey these last conditions are called *flexible*. These conditions are much less restrictive than the alternative conditions. In fact, most of the interesting nonassociative algebras are flexible. For example, both the Lie and Jordan algebras are flexible, the *Jordan algebras* $\mathcal{T}$ being defined by the (Jordan) identities: $(x^2,y,x) = (x,y,x^2) = 0$, $\forall x,y \in \mathcal{T}$. See N. Jacobson, *Structure and Representations of Jordan Algebras*, (American Mathematical Society, 1968). We shall find another class of flexible algebras below.

[10] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer-Verlag New York, 1972).

[11] K. A. Zhevlakov, A. M. Slin'ko, I. P. Shestakov, and A. I. Shirshov, *Rings That Are Nearly Associative* (Academic, New York, 1982).

[12] Note, that although $X_i \cdot X_j = -2\delta_{ij}$ $(i,j=1,2,3)$ for $X_i \in so(3)$, the diagonal elements of $X_i \cdot X_j$ can have different signs, as in $so(2,1)$, where $(X_i \cdot X_j) = \text{diag}(-2,-2,2)$.

[13] Actually we can generalize $\mathcal{A}^{\heartsuit}$ further, by replacing the product of the complex numbers $\alpha\beta$ with a bilinear product $\alpha\beta \leftrightarrow \alpha\cdot\beta = \alpha_+^2 + \lambda_1\alpha_-^2$, where $\alpha_\pm = (\alpha \pm \beta)/\sqrt{2}$. We have refrained from doing so in the text.

[14] Thus, if we reverse the sign in front of the σ in (17), we only need to change the sign of the commutators in (18) to fix the multiplication rule.

On Lie algebras all nilpotent subalgebras of which are Abelian

Elke Dallmer
Ernst-Moritz-Arndt-Universität, Institut für Mathematik und Informatik, Fr.-L.-Jahn-Strasse 15a, D-17489 Greifswald, Germany

(Received 2 February 1998; accepted for publication 8 June 1998)

By Kirillov's theorem, every non-Abelian nilpotent Lie algebra contains the three-dimensional Heisenberg algebra $H(3)$ as a subalgebra. [R. Schimming, Arch. Math. **24**, 65–74 (1988)]. Here we are interested in a sufficient condition for a non-nilpotent Lie algebra to contain a subalgebra isomorphic to $H(3)$. Explicitly, we show: Every indecomposable non-nilpotent Lie algebra L of dimension dim $L \geqslant 3$ with a nonvanishing center $Z(L) \neq \{0\}$ contains a non-Abelian nilpotent subalgebra and hence a copy of $H(3)$. *© 1999 American Institute of Physics.* [S0022-2488(98)02410-4]

I. MOTIVATION BY THE YANG–MILLS EQUATIONS

The given problem stems from the study of constant Yang–Mills potentials. These give examples of Huygens's-type differential equations with constant coefficients and examples of harmonic differential operators.[1]

The Yang–Mills equations on a Riemannian manifold (E,g) for a potential $A = A_\alpha dx^\alpha$ with values in some Lie algebra L are given by

$$D_\alpha F^{\alpha\beta} \equiv \partial_\alpha F^{\alpha\beta} + [A_\alpha, F^{\alpha\beta}] = 0,$$

where $[\cdot,\cdot]$ denotes the commutator in L, $F_{\alpha\beta} = \partial_\alpha A_\beta - \partial_\beta A_\alpha + [A_\alpha, A_\beta]$ are the field strength components, $\partial_\alpha = \partial/\partial x^\alpha$ the partial derivatives with respect to local coordinates x^α of $x \in E$, $g = g_{\alpha\beta} dx^\alpha dx^\beta$ denotes a Riemannian metric, and D_α is the corresponding covariant derivative. If the Yang–Mills potential A has constant components A_α then the Yang–Mills equations $D_\alpha F^{\alpha\beta} = 0$ collapse to the algebraic equations

$$[A_\alpha, F^{\alpha\beta}] = [A_\alpha, [A^\alpha, A^\beta]] = 0.$$

A general solution of the Yang–Mills equations is not available. It depends on the structure of the Lie algebra L and the signature of the metric g whether or not there exists a nontrivial solution of these equations (i.e., potentials A with $F_{\alpha\beta} \equiv [A_\alpha, A_\beta] \neq 0$ and $D_\alpha F^{\alpha\beta} = 0$). One should consider the structure of Lie algebras in detail. For special types of Lie algebras, like Abelian, nilpotent, compact, and for all Lie algebras of dim $L \leqslant 5$ we can decide whether or not there exists a nontrivial solution for the constant Yang–Mills equations.[2–4] The following holds true:

(1) If a subalgebra M of a Lie algebra L admits a nontrivial solution of the Yang–Mills equations then so does L.
(2) Every Yang–Mills potential A with values in a two-step nilpotent Lie algebra L is a constant Yang–Mills potential.

Example: The Heisenberg algebra $H = H(2m+1)$ has as the only nonvanishing structure relations $[X_j, Y_j] = Z$ for $j = 1,...,m$ with respect to a base $\{X_1, \ldots, X_m, Y_1, \ldots, Y_m, Z\}$. Hence it is two-step nilpotent, i.e.,

$$[H,[H,H]] = 0,$$

and so every element of $H(2m+1)$ is a constant Yang–Mills potential.

Corollary: Every non-Abelian nilpotent Lie algebra L admits a nontrivial solution of the Yang–Mills equations.

It would be desirable to know if a given non-nilpotent Lie algebras contains a (non-Abelian) nilpotent subalgebra, because then it is easy to find out whether or not there exists a nontrivial solution of the constant Yang–Mills equations. The work on low-dimensional Lie algebras[5,6] leads to the conjecture that an indecomposable Lie algebra with a nonvanishing center contains a non-Abelian nilpotent subalgebra. This conjecture will be the subject of the present paper.

II. PRELIMINARIES ON THE STRUCTURE OF LIE ALGEBRAS

In this section we give a brief survey of the most important definitions and theorems which we need in the following. We refer the reader to the literature for the proofs of the lemmas and theorems.

We denote the direct sum of vector spaces by $\dot{+}$, the direct sum of Lie algebras by $\oplus$, and the semidirect sum of Lie algebras by $\niplus$.

Definition: A Lie algebra L is called *decomposable* if we can decompose it into the direct sum of two ideals. Otherwise L is called *indecomposable*.

Now we define solvable and nilpotent Lie algebras.

Definition: Let L be a Lie algebra. Let $L^{(0)}:=L$ and $L^{(n+1)}:=[L^{(n)},L^{(n)}]$ with $n\in\mathbb{N}$. Then $(L^{(n)})_{n\in\mathbb{N}}$ is called the *derived series* of L. The Lie algebra L is called *solvable* if there exists a $k\in\mathbb{N}$ such that $L^{(k)}=\{0\}$. Let $L^0:=L$ and $L^{n+1}:=[L,L^n]$ with $n\in\mathbb{N}$. Then $(L^n)_{n\in\mathbb{N}}$ is called the *lower central series* of L. The Lie algebra L is called *nilpotent* if there exists a $k\in\mathbb{N}$ such that $L^k=\{0\}$.

Remark: There holds $L^{(1)}=L^1=:L'$. L' is called the *commutator algebra of L*.

Lemma 2: Nilpotent and solvable Lie algebras satisfy the following:

(i) Let $L\neq\{0\}$ be a nilpotent Lie algebra. Then $Z(L)\neq\{0\}$ (Ref. 7, p. 111).
(ii) A Lie algebra L is solvable if and only if $[L,L]$ is nilpotent (Ref. 7, p. 117).
(iii) There exists a maximal solvable ideal in L, the so-called *radical* $R=R_L$. Similarly there exists a maximal nilpotent ideal in L, the so-called *nilradical* $N=N_L$ (Ref. 7, pp. 114, 175).

Lemma 3: The center $Z(L)=\{X\in L|(\forall Y\in L)[X,Y]=0\}$, the nilradical N_L and the radical R_L are *characteristic ideals* of a Lie algebra L, i.e., they are invariant under all derivations of L (Ref. 8, pp. 155, 210).

Beyond this we need the following.

Lemma 4: A characteristic ideal in an ideal of L is an ideal of L as well (Ref. 8, p. 155). Nilpotent and solvable Lie algebras contain—in contrast to semisimple or simple Lie algebras—many ideals.

Definition: A Lie algebra L is called *semisimple* if the radical is trivial, $R_L=\{0\}$, and it is called *simple* if it is not Abelian and contains no ideals except L and $\{0\}$.

Obviously, every simple Lie algebra is semisimple as well. One can show that every semisimple Lie algebra is the direct sum of simple ideals (Ref. 7, p. 126). Furthermore, we need the following.

Lemma 5: Let $L\neq\{0\}$ be a semisimple Lie algebra. Then there holds:

(i) $[L,L]=L$ (Ref. 7, p. 127).
(ii) $Z(L)=\{0\}$ (Ref. 8, p. 201).

All (finite-dimensional) Lie algebras are composed from semisimple and solvable Lie algebras.

Levi's theorem: Every Lie algebra L is isomorphic to a semidirect sum of its radical R with a semisimple subalgebra S, i.e., $L\cong R\niplus S$ (Ref. 7, p. 146).

Thus, Lie algebras can be classified by means of classification of the following objects:

(1) solvable Lie algebras R,
(2) semisimple Lie algebras S (or rather simple Lie algebras, since every semisimple Lie algebra can be decomposed into the direct sum of simple ideals),

(3) semidirect sums of solvable Lie algebras with a semisimple one.

Simple Lie algebras are completely classified nowadays. The solution of the other two problems is still lacking.[5,6,9]

One important result in the theory of representations of Lie algebras is the following.

Weyl's theorem: A finite-dimensional representation $\phi: S \rightarrow gl(V)$ of a semisimple Lie algebra S on a finite-dimensional vector space V is completely reducible (Ref. 10, p. 28).

Corollary: Let S be a semisimple Lie algebra, V a finite-dimensional S-module. The vector spaces

$$V_0 := \{ v \in V | S.v = 0 \}, \quad V_{\text{eff}} := \text{span } S \cdot V$$

are S-invariant subspaces and $V = V_0 \dot{+} V_{\text{eff}}$.

Proof: The action of S on V, $S \times V \rightarrow V$, $(X,v) \mapsto X \cdot v =: \phi(X)(v)$, defines a representation ϕ of S in V. By the Weyl's theorem this representation is completely reducible, i.e., every submodule has a complement. Obviously, V_0 and V_{eff} are S-submodules of V. We show that V is the direct sum $V = V_0 \dot{+} V_{\text{eff}}$: Choose a subspace $V_1 \subseteq V_{\text{eff}}$ such that $V_{\text{eff}} = (V_0 \cap V_{\text{eff}}) \dot{+} V_1$. Then $S \cdot V = S \cdot V_{\text{eff}} = S \cdot V_1 \subseteq V_1$, hence $V_{\text{eff}} \subseteq V_1$ and therefore $V_0 \cap V_{\text{eff}} = \{0\}$. On the other side $V = V_0 + V_{\text{eff}}$: Choose a submodule $W \subseteq V$ such that $V = (V_0 + V_{\text{eff}}) \dot{+} W$. Then $S \cdot W \subseteq W \cap V_{\text{eff}} = \{0\}$. Therefore $W \subseteq V_0$, hence $W = \{0\}$. □

We will also need the following:

Lemma 6: Let H be a nilpotent subalgebra of L and let $M \subseteq L$ be an $(\text{ad } H)$-invariant subspace. Then

$$M_0 := \{ Y \in M | (\forall X \in H) \exists n \in \mathbb{N}: \ (\text{ad } X)^n Y = 0 \}$$

is an $(\text{ad } H)$-invariant subspace of M with a complement in M, i.e., there exists an $(\text{ad } H)$-invariant subspace $M_1 \subseteq M$ such that $M = M_0 \dot{+} M_1$.

Proof: We use the weight-decomposition of a (real) Lie algebra with respect to a nilpotent subalgebra. For details see Ref. 11, Chap. VII and Ref. 13, Chap. I.10. First we consider the complexification $M^{\mathbb{C}} = M \otimes_{\mathbb{R}} \mathbb{C}$ of M and its generalized weight-decomposition with respect to $H^{\mathbb{C}} = H \otimes_{\mathbb{R}} \mathbb{C}$:

$$M^{\mathbb{C}} = \sum_{\lambda \in \Lambda} M_{\lambda}^{\mathbb{C}},$$

where

$$M_{\lambda}^{\mathbb{C}} = \{ X \in M^{\mathbb{C}} | (\forall Y \in H^{\mathbb{C}}) \exists n \in \mathbb{N}: \quad (\text{ad } Y - \lambda(Y) \text{Id}_{M^{\mathbb{C}}})^n X = 0 \},$$

and Λ is the set of complex weights with respect to $H^{\mathbb{C}}$. More precisely there holds:

$$M^{\mathbb{C}} = M_0^{\mathbb{C}}(H^{\mathbb{C}}) \dot{+} \sum_{j=1}^{m} M_{\lambda_j}^{\mathbb{C}}(H^{\mathbb{C}}), \quad \lambda_j \neq 0 \quad \text{for } 1 \leq j \leq m.$$

We also know that

$$M_0(H) \otimes_{\mathbb{R}} \mathbb{C} \cong M_0^{\mathbb{C}}(H) \cong M_0^{\mathbb{C}}(H^{\mathbb{C}}).$$

We define the conjugation c on the complexification $L^{\mathbb{C}} = L \otimes_{\mathbb{R}} \mathbb{C}$ of the Lie algebra L by $c(X + iY) = X - iY$ with $X, Y \in L$. Further, we use the following.

Lemma:

(i) If $\lambda \in \Lambda$ is a weight then so is $\overline{\lambda} \circ c$. Moreover, $c(L_{\lambda}^{\mathbb{C}}) = L_{\overline{\lambda} \circ c}^{\mathbb{C}}$.

(ii) $L_{\lambda}^{\mathbb{C}} \dot{+} L_{\overline{\lambda} \circ c}^{\mathbb{C}}$ is c-invariant.

(iii) Let $X \in L_\lambda^{\mathbb{C}}$, then the real and imaginary parts of X are in $L \cap (L_\lambda^{\mathbb{C}} \dot{+} L_{\overline{\lambda \circ c}}^{\mathbb{C}})$.

With λ also $\overline{\lambda \circ c}$ is in Λ and the two root spaces are conjugated to each other in the real case. We define the class $[\lambda] := \{\lambda, \overline{\lambda \circ c}\}$ and the corresponding real root spaces of M by $M_{[\lambda]} := M \cap (M_\lambda^{\mathbb{C}} \dot{+} M_{\overline{\lambda \circ c}}^{\mathbb{C}})$.

Hence in the real case we have the following generalized weight-decomposition:

$$M = M_0 \dot{+} \sum_{j=1}^{m} M_{[\lambda_j]} =: M_0 \dot{+} M_1, \quad \lambda_j \neq 0 \quad \text{for } 1 \leqslant j \leqslant m. \qquad \square$$

III. THE RESULT ON LIE ALGEBRAS WITH ONLY ABELIAN NILPOTENT SUBALGEBRAS

In this section we prove our main result:

Theorem: An indecomposable non-nilpotent Lie algebra L of dimension $\dim L \geqslant 3$ with a nontrivial center $Z(L) \neq \{0\}$ contains a non-Abelian nilpotent subalgebra.

The following is logically equivalent:

Theorem: A non-nilpotent Lie algebra L of dimension $\dim L \geqslant 3$ with a nontrivial center $Z(L) \neq \{0\}$ all of whose nilpotent subalgebras are Abelian is decomposable. Moreover, $Z(L)$ is a direct summand of L, i.e., there exists an ideal I in L with $L = Z(L) \oplus I$.

Proof: By Levi's theorem the following cases have to be considered:

(1) L is non-nilpotent and solvable.
(2) L is isomorphical to a semidirect sum of its radical $R \neq \{0\}$ and a semisimple subalgebra $S \neq \{0\}$, i.e., $L \cong R \oplus\!\!\!\!\supset S$.

Notice that the case of a semisimple Lie algebra L is canceled, because then the assumption $Z(L) \neq \{0\}$ is not satisfied.

Case 1: Let L be a non-nilpotent solvable Lie algebra and denote the nilradical of L by N. Generally, $[R,L] \subseteq N$ (Ref. 7, p. 175). In particular, for solvable L we have $L' \subseteq N$. Let us consider the factor algebra L/N. For X, $Y \in L$ there holds

$$[X+N, Y+N] = [X,Y] + N \subseteq L' + N \subseteq N + N = N,$$

i.e., the quotient L/N is Abelian. Next we consider the canonical projection $\pi: L \rightarrow L/N$. Let H be a Cartan subalgebra of L. The image of a Cartan subalgebra under a subjective Lie algebra homomorphism is a Cartan subalgebra and a nilpotent Lie algebra is its only Cartan subalgebra (Ref. 7, p. 131). Since L/N is an Abelian Lie algebra it is the only Cartan subalgebra in L/N and therefore $\pi(H) = L/N$. Thus $L = N + H$, on the one hand. On the other hand we have $L \cong N \dot{+} L/N$. So we find in H a subalgebra $\overline{H}$ that is isomorphic to $\pi(H)$, because by assumption H is Abelian. Hence we get a direct sum of vector spaces $L = N \dot{+} \overline{H}$. In particular, $[N, \overline{H}] \subseteq N$ and N is an ideal in L. Therefore $L = N \oplus\!\!\!\!\supset \overline{H}$ is a semidirect sum of Lie subalgebras of L.

Let us apply Lemma 6 to the case $M = N$. The vector space

$$N_0 = \{Y \in N \,|\, (\forall X \in \overline{H}) \exists n \in \mathbb{N} : (\text{ad } X)^n Y = 0\}$$

is an $(\text{ad } \overline{H})$-invariant subspace with a complement N_1 in N: $N = N_0 \dot{+} N_1$, where N_1 is $(\text{ad } \overline{H})$-invariant as well. We show: $L_1 := N_0 \oplus\!\!\!\!\supset \overline{H}$ is nilpotent, i.e., there exists an $n \in \mathbb{N}$ with $L_1^n = \{0\}$. Namely, $L_1^1 = [L_1, L_1] = [\overline{H}, N_0]$, because $\overline{H}$ and N_0 are Abelian. Further, $L_1^1 = [\overline{H}, N_0] \subseteq N_0$ because N_0 is invariant under the action of ad $\overline{H}$. Induction with respect to n gives $L_1^n = [L_1, L_1^{n-1}] = [\overline{H}, L_1^{n-1}] = (\text{ad } \overline{H})^n N_0$, since $L_1^{n-1} \subseteq N_0$ for $n > 1$. By definition of N_0 we know that for every $Y \in N_0$ there exists an $n \in \mathbb{N}$ such that $(\text{ad } \underline{X})^n Y = 0$ for all $X \in \overline{H}$. Thus $N_0 \oplus\!\!\!\!\supset \overline{H}$ is nilpotent, and so by assumption Abelian. Therefore, $[N_0, \overline{H}] = \{0\}$. Next we show: $N_0 = Z(L)$. By definition of N_0 we obtain the inclusion $Z(L) \subseteq N_0$. The other inclusion $N_0 \subseteq Z(L)$ is proved as follows:

$$[N_0,L]=[N_0,N\oplus\overline{H}]=[N_0,N]+[N_0,\overline{H}]=0,$$

since $[N_0,\overline{H}]=\{0\}$ and N, and therefore also N_0, is Abelian by assumption. Hence $N_0\subseteq Z(L)$.

Our next goal is to show: $N_1=L'$. On the one hand,

$$L'=[L,L]=[N\oplus\overline{H},N\oplus\overline{H}]=[N,N]+[\overline{H},\overline{H}]+[\overline{H},N]=[\overline{H},N],$$

$$[\overline{H},N]=[\overline{H},N_0\dotplus N_1]=[\overline{H},N_0]+[\overline{H},N_1]\subseteq N_1,$$

because $N_0\oplus\overline{H}$ is Abelian and N_1 is invariant under the action of $(\mathrm{ad}\ \overline{H})$. One the other hand, let us apply the construction of N_1 by the proof of Lemma 6. Namely,

$$N_\alpha^{\mathbb{C}}:=\{Y\in N^{\mathbb{C}}|(\forall X\in\overline{H}^{\mathbb{C}})\exists n\in\mathbb{N}:(\mathrm{ad}\ X-\alpha(X))^nY=0\}$$

with a complex weight $0\neq\alpha\in(\overline{H}^{\mathbb{C}})^*$. Then for a $Y\in N_\alpha^{\mathbb{C}}$ there exists a $X\in\overline{H}^{\mathbb{C}}$ with $\alpha(X)\neq 0$ and an $n\in\mathbb{N}$ with

$$0=(\alpha(X)-\mathrm{ad}\ X)^nY=\sum_{k=0}^{n}\binom{n}{k}\alpha(X)^{n-k}(-\mathrm{ad}\ X)^kY=\alpha(X)^nY$$

$$-\sum_{k=1}^{n}\binom{n}{k}\alpha(X)^{n-k}(-\mathrm{ad}\ X)^{k-1}[X,Y].$$

From this it follows

$$Y=\alpha(X)^{-n}\sum_{k=1}^{n}\binom{n}{k}\alpha(X)^{n-k}(-\mathrm{ad}\ X)^{k-1}[X,Y].$$

Therefore $N_\alpha^{\mathbb{C}}\subseteq L'^{\mathbb{C}}$ and hence

$$N_1^{\mathbb{C}}:=\sum_{0\neq\alpha\in(\overline{H}^{\mathbb{C}})^*}N_\alpha^{\mathbb{C}}\subseteq L'^{\mathbb{C}}\quad\text{and so}\quad N_1\subseteq L'.$$

Finally, we get the decomposition $N=Z(L)\oplus L'$ with $N_0=Z(L)$ and $N_1=L'$ as well as

$$L=N\oplus\overline{H}=(N_0\oplus N_1)\oplus\overline{H}=N_0\oplus(N_1\oplus\overline{H})=Z(L)\oplus(L'\oplus\overline{H}).$$

(Note: $[N_0,N_1\oplus\overline{H}]=0$ and $N_1\oplus\overline{H}$ is a subalgebra of L.) Therefore, here $Z(L)\neq\{0\}$ is direct summand in L.

Case 2: We consider the Levi decomposition $L=R\oplus S$ with a radical $R=R_L\neq\{0\}$ and a semisimple subalgebra $S\neq\{0\}$. There holds $[R,L]\subseteq R$ and $[S,S]=S$.

From case 1 we know $R=N_R\oplus\overline{H}_R$ with $N_R=Z(R)\oplus R'$, and therefore $R=(Z(R)\oplus R')\oplus\overline{H}_R$. In particular, $Z(R)\cap R'=\{0\}$. Thus,

$$\begin{aligned}L'=[L,L]=[R\oplus S,R\oplus S]=[R,R]+[S,S]+[S,R]\\ &=R'+S+[S,R]=R'+S+[S,N_R\oplus\overline{H}_R]=R'+S+[S,N_R]+[S,\overline{H}_R]\\ &=R'+S+[S,Z(R)\oplus R']+[S,\overline{H}_R]\\ &=R'+S+[S,Z(R)]+[S,R']+[S,\overline{H}_R].\end{aligned}$$

From the Jacobi identity it follows $[S,R']=[S,[R,R]]\subseteq[[S,R],R]\subseteq[R,R]=R'$. Therefore

$$L'=R'+S+[S,Z(R)]+[S,\overline{H}_R].$$

We can take R as an S-module with respect to the action of ad, because $[S,R]\subseteq[L,R]\subseteq R$. It follows that $[S,N_R]\subseteq N_R$, since N_R is a characteristic ideal in R, i.e., N_R is a S-submodule of R. By Weyl's theorem it follows that N_R has a S-module complement. As we know from case 1, we can assume that $\overline{H}_R$ is this module complement. Therefore, we obtain

$$R=N_R\dot{+}\overline{H}_R,\quad [S,\overline{H}_R]\subseteq\overline{H}_R.$$

There holds $[S,Z(R)]\subseteq Z(R)$ because $Z(R)$ is a characteristic ideal in R. By Weyl's theorem we conclude that $Z(R)=V_0\dot{+}V_{\text{eff}}$ with

$$\begin{aligned}V_0&=\{X\in Z(R)\,|\,(\text{ad } S)\cdot X=0\}\\&=\{X\in Z(R)\,|\,(\forall U\in S)[X,U]=0\}\\&=\{X\in R\,|\,(\forall V\in R)[X,V]=0\wedge(\forall U\in S)[S,U]=0\}\\&=\{X\in L\,|\,(\forall Y\in L)[X,Y]=0\}=Z(L)\end{aligned}$$

and

$$V_{\text{eff}}=\text{span}(\text{ad } S).Z(R)=[S,Z(R)],$$

i.e., $Z(R)=Z(L)\dot{+}[S,Z(R)]$ and, in particular, $Z(L)\cap[S,Z(R)]=\{0\}$. Since $Z(R)\cap R'=\{0\}$, $[S,Z(R)]\subseteq Z(R)$ and $[S,\overline{H}_R]\subseteq\overline{H}_R$, we obtain

$$L'=R'\dot{+}S\dot{+}[S,Z(R)]\dot{+}[S,\overline{H}_R],$$

i.e., the sum is direct. With $Z(R)=Z(L)+[S,Z(R)]$ then we conclude $L'\cap Z(L)=\{0\}$. So, we obtain L as the direct sum of vector spaces:

$$L=R\dot{+}S=Z(L)\dot{+}[S,Z(R)]\dot{+}R'\dot{+}\overline{H}_R\dot{+}S=Z(L)\dot{+}I.$$

If $L'\subseteq I$ with $L=Z(L)\dot{+}I$ then I is an ideal in L: $[I,L]\subseteq L'\subseteq I$, and hence $L=Z(L)\oplus I$ is a direct sum of ideals in L.

Actually, we have shown the assertion that $Z(L)$ is a direct summand in L, if every nilpotent subalgebra of L is abelian. The theorem is proven. □

ACKNOWLEDGMENTS

I am greatly indebted to Herrn Professor K.-H. Neeb from the University of Erlangen for many valuable comments. Also I want to thank my mentor Professor R. Schimming from the University of Greifswald for drawing my attention to the problem on constant Yang–Mills potentials and for helpful discussions.

[1] R. Schimming, Forum Math. **3**, 177–203 (1991).
[2] R. Schimming, Arch. Math. **24**, 65–74 (1988).
[3] R. Schimming and E. Mundt, J. Math. Phys. **33**, 1–5 (1992).
[4] E. Mundt, Semin. Sophus Lie **3**, 107–115 (1993).
[5] G. M. Mubarakzyanov, Izv. Vyss. Uchebn. Zaved. Mat. **32**, 114–123 (1963); **34**, 99–105 (1963); **35**, 104–116 (1963); **55**, 95–98 (1966) (Russian).
[6] P. Turkowski, J. Math. Phys. **29**, 2139–2144 (1988); **31**, 1344–1350 (1990).
[7] J. Hilgert and K.-H. Neeb, *Lie-Gruppen und Lie-Algebren* (Vieweg, Braunschweig, 1991).
[8] J. Tits, *Liesche Gruppen und Algebren* (Springer, Berlin, 1983).
[9] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (PWN, Warszawa, 1980).
[10] J. E. Humphreys, *Introduction to Lie Algebras and Representations Theory* (Springer, Berlin, 1972).
[11] N. Bourbaki, *Groupes et Algèbres de Lie VII* (Hermann, Paris, 1975).
[12] M. Wüstner, Dissertation, Darmstadt, 1995.

Hopf algebra extension of a Zamolochikov algebra and its double

Jintai Ding
Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221

(Received 20 April 1998; accepted for publication 29 January 1999)

The particles with a scattering matrix $R(x)$ are defined as operators $\Phi_i(z)$ satisfying the relation $\Sigma_{i',j'} R_{i,j}^{j',i'}(x_1/x_2)\Phi_{i'}(x_1)\Phi_{j'}(x_2)=\Phi_i(x_2)\Phi_j(x_1)$. The algebra generated by those operators is called a Zamolochikov algebra. We construct a new Hopf algebra by adding half of the Faddeev–Reshetikhin–Takhtajan–Semenov-Tian-Shansky (FRTS) construction of a quantum affine algebra with this $R(x)$. Then we double it to obtain a new Hopf algebra such that the full FRTS construction of a quantum affine algebra is a Hopf subalgebra inside. Drinfeld realization of quantum affine algebras is included as an example. © *1999 American Institute of Physics.* [S0022-2488(99)04107-9]

I. INTRODUCTION

In physics, the particles with a scattering matrix $R(x)$ in $\mathrm{End}(V)\otimes\mathrm{End}(V)$ are defined with the operators $\Phi_i(x)$ index by a linear independent basis of V such that

$$\sum_{i',j'} R_{i,j}^{j',i'}(x_1/x_2)\Phi_{i'}(x_1)\Phi_{j'}(x_2)=\Phi_i(x_2)\Phi_j(x_1),$$

where V is a vector space and x is a parameter in $\mathbb{C}$. This naturally gives an algebra with these current generators $\Phi_i(x)$, which we will call a Zamolochikov algebra. However this algebra is not given a Hopf algebra structure. We construct a Hopf algebra on this algebra by adding structures coming from the structures of the affine quantum groups.

The definition of quantum groups discovered by Drinfeld and Jimbo is presented as a deformation of the simple Lie algebra by the basic generators and the relations based on the data coming from the corresponding Cartan matrix. The extension of the realization of the affine Kac–Moody algebra $\hat{\mathfrak{g}}$ associated to a simple Lie algebra $\mathfrak{g}$ as a central extension of the corresponding loop algebra $\mathfrak{g}\otimes\mathbb{C}[t,t^{-1}]$ has two different approaches. The first approach was given by Faddeev, Reshetikhin, and Takhtajan,[1] who obtained a realization of the quantum loop algebra $U_q(\mathfrak{g}\otimes\mathbb{C}[t,t^{-1}])$ via a canonical solution of the Yang–Baxter equation depending on a parameter $z\in\mathbb{C}$. This approach was completed by Reshetikhin and Semenov-Tian-Shansky[2] by incorporating the central extension in the previous realization. We call this approach FRTS construction. The second approach was given by Drinfeld, who[3] gave a realization of the quantum affine algebra $U_q(\mathfrak{g})$ and its special degeneration called the Yangian. As an algebra, this realization is equivalent to the FRTS construction[4] through certain Gauss decomposition for the case of $U_q(\widehat{\mathfrak{gl}}(n))$. Although we cannot extend the conventional comultiplication to the current operators of Drinfeld to derive a closed comultiplication formula, Drinfeld also gave the Hopf algebra structure for such a formulation,[4] which[5,6] we used to study vertex operators and zeros and poles of the quantum current operators. In the Drinfeld realization of quantum affine algebras, the structure constants are certain rational functions $g_{ij}(z)$. In Ref. 6, we generalize this type of Hopf algebras by substituting $g_{ij}(z)$ by other functions that satisfy the functional property of $g_{ij}(z)$.

In this paper, we will use the idea of FRTS construction to define a Hopf algebra generated by a current operator valued matrix on $V,L(x)$, such that

$$R(x_1/x_2)L_1(x_1)L_2(x_2)=L_2(x_2)L_1(x_1)R(x_1/x_2),$$

where $L_1(x)=L(x)\otimes 1$ and $L_2(x)=1\otimes L(x)$; and the commutation relation between the particles and this new operator matrix $L(x)$ is presented as

$$\Phi(x_1)_1 L(x_2)_2 = R((x_1/x_2)q^{c/2})^{-1} L(x_2)_2 \Phi(x_1)_1 .$$

This relation can be interpreted as that $\Phi(x)$ is an intertwiner for the algebra generated by $L(x)$.[7] With this we can define a comultiplication on the algebra generated by $L(x)$ and $\Phi(x)$, where the comultiplication for $L(z)$ comes from FRTS construction and the comultiplication of $\Phi(x)$ is defined as

$$\Delta(\Phi(z)) = \Phi(x)\otimes 1 + L(xq^{c_1/2})\otimes \Phi(zq^{c_1}),$$

which is a generalization of Drinfeld construction. Then combining the idea of FRTS construction and Drinfeld realization, we give a double for such a construction, where the FRTS construction is a Hopf subalgebra and Drinfeld realization is a special case of our realization with certain diagonal $R(x)$. This paper is basically a result of the combination of the two approaches, the FRTS construction and the Drinfeld realization.

This paper contains two additional sections. In Sec. II we define the Zamolochikov algebra and present its Hopf algebra extension. Section III describes the double of such a construction and the related examples.

II. HOPF ALGEBRA EXTENSION OF A ZAMOLOCHIKOV ALGEBRA

Let V be the vector space $\mathbb{C}^n$. Let x be a parameter in $\mathbb{C}$. A function valued R-matrix $R(x)$ is a function valued operator in $\mathrm{End}(V)\otimes \mathrm{End}(V)$, which satisfies the so-called Yang–Baxter equation:

$$R_{12}(z)R_{13}(z/w)R_{23}(w) = R_{23}(w)R_{13}(z/w)R_{12}(z),$$

where $R_{12}(x)\Sigma_{ij} = f_{ij}(x)a_i\otimes b_j\otimes 1 = R(x)\otimes 1$, $R_{13}(x)=\Sigma_{ij} f_{ij}(x)a_i\otimes 1\otimes b_j$, $R_{23}(x)=\Sigma_{ij} f_{ij}(x)1\otimes a_i\otimes b_j = 1\otimes R$, and $R(x)=f_{ij}(x)\Sigma_{ij}\, a_i\otimes b_j$. We also require that $R(x)$ satisfies the unitary condition

$$R_{21}(z)^{-1} = R(z^{-1}),$$

where $R_{21}(z)=f_{ij}(x)\Sigma_{ij}\, b_j\otimes a_i$.

Definition 2.1: The associative algebra $P[R(x)]$ is an algebra generated by operators $\Phi_i(x)$ index by a linear independent basis e_i of V. Let $\Phi(x)=\Sigma \Phi_i(x)\otimes e_i$. The commutation relations are presented as

$$R(x_1/x_2)\Phi(x_1)_1\Phi(x_2)_2 = \Phi(x_2)_2\Phi(x_1)_1 ,$$

where

$$\Phi(x_1)_1\Phi(x_2)_2 = \sum \Phi_i(x_1)\Phi_j(x_2)e_i\otimes e_j , \quad \Phi(x_2)_2\Phi(x_1)_1 = \sum \Phi_j(x_2)\Phi_i(x_1)e_i\otimes e_j .$$

As explained in Sec. I, this system is used in the description particles $(\Phi_i(x))$ in physics with the scattering matrix $R(x)$ and in some other context. This relation is also satisfied by the vertex operators for quantum affine algebras,[7,8] etc., and this type of system also appeared in describing the elliptic type of algebras.[9] However, they are all described as algebras, not Hopf algebras. Following the idea in Refs. 6, 10–12 we would like to extend this algebra with additional current operators coming from the FRTS construction to give a Hopf algebra structure to such a system.

Definition 2.2: The algebra $EP[R(x)]$ is an associative algebra generated by $\Phi_i(x)$ indexed by a linear independent basis e_i of $V, l_{ij}(x)$ indexed by the linear independent basis e_{ij} of $\mathrm{End}(V)$

and a central element c. Let $\Phi(x)=\Phi_i(x)\otimes e_i$ and the operator valued matrix $L(x)=\Sigma l_{ij}(x)\otimes e_{ij}$, such that $L(x)$ is invertible. They satisfy the commutation relations:

$$R(x_1/x_2)\Phi(x_1)_1\Phi(x_2)_2=\Phi(x_2)_2\Phi(x_1)_1,$$

$$\Phi(x_1)_1L(x_2)_2=R(q^{c/2}x_1/x_2)^{-1}L(x_2)_2\Phi_i(x_1)_1,$$

$$R(x_1/x_2)L(x_1)_1L(x_2)_2=L(x_2)_2L(x_1)_1R(x_1/x_2).$$

Here

$$\Phi(x_1)_1L(x_2)_2=\sum \Phi_i(x_1)L_{kl}(x_2)e_i\otimes e_{kl}, \quad L(x_2)_2\Phi_i(x_1)_1=\sum L_{kl}(x_2)\Phi_i(x_1)e_i\otimes e_{kl},$$

$$L(x_1)_1L(x_2)_2=PL(x_1)_2L(x_2)_1P=\sum L_{ij}(x_1)L_i(x_2)e_{ij}\otimes e_{kl},$$

and P is the permutation operator.

Theorem 2.1: *The algebra $EP[R(x)]$ has a Hopf algebra structure, which is given by the following formulas.*

Coproduct Δ,

$$(0) \quad \Delta(q^c)=q^c\otimes q^c,$$

$$(1) \quad \Delta(\Phi_i(z))=\Phi_i(z)\otimes 1+\sum L_{ij}(zq^{c_1/2})\otimes\Phi_j(zq^{c_1}),$$

$$(2) \quad \Delta(L_{ij}(z))=\sum L_{ik}(zq^{-c_2/2})\otimes L_{kj}(zq^{c_1/2}),$$

where $c_1=c\otimes 1$ and $c_2=1\otimes c$.

Counit ϵ,

$$\epsilon(q^c)=1, \quad \epsilon(L_{ij}(z))=\delta_{ij},$$

$$\epsilon(\Phi_i^{\pm}(z))=0.$$

Antipode a,

$$(0) \quad a(q^c)=q^{-c},$$

$$(1) \quad a(\Phi_i(z))=\sum -(L(zq^{-c/2})^{-1})_{ij}\Phi_j(zq^{-c}),$$

$$(2) \quad a(L(z))=(L(z))^{-1}.$$

We will use the notation to denote the comultiplication,

$$\Delta\Phi(x_1)=\Phi(x_1)\overline{\otimes}1+L(x_1q^{c_1/2})\overline{\otimes}\Phi(x_1q^{c_1}),$$

$$\Delta(L(x_2)=(L(x_2q^{-c_2/2})\overline{\otimes}L(x_2q^{c_1/2}))$$

Proof: For the comultiplication above we have that

$$\Delta\Phi(x_1)_1\Delta L(x_2)_2=(\Phi(x_1)\overline{\otimes}1+L(x_1q^{c_1/2})\overline{\otimes}\Phi(x_1q^{c_1}))_1(L(x_2q^{-c_2/2})\overline{\otimes}L(x_2q^{c_1/2})_2$$
$$=R(x_1/x_2q^{(c_1+c_2)/2})^{-1}(L(x_2q^{-c_2}/2)\overline{\otimes}L(x_2q^{c_1/2}))_2(\Phi(z)\overline{\otimes}1)_1$$
$$+R(x_2/x_1q^{(c_1+c_2)/2})^{-1}(L(x_2q^{-c_2/2})\overline{\otimes}L(x_2q^{c_1/2}))_2(L(x_1q^{c_1/2})$$
$$\overline{\otimes}\Phi(x_1q^{c_1}))_1,$$

$$R(x_1/x_2)\Delta\Phi(x_1)_1\Delta\Phi(x_2)_2=R(x_1/x_2)(\Phi(x_1)\overline{\otimes}1+L(x_1q^{c_1/2})\overline{\otimes}\Phi(x_1q^{c_1}))_1(\Phi(x_2)\overline{\otimes}1$$
$$+L(x_2q^{c_1/2})\overline{\otimes}\Phi(x_2q^{c_1}))_2$$
$$=(\Phi(x_2)\overline{\otimes}1)_2(\Phi(x_1)\overline{\otimes}1)_1$$
$$+(L(x_2q^{c_1/2})\overline{\otimes}\Phi(x_2q^{c_1}))_2(\Phi(x_1)\overline{\otimes}1)_1+(L(x_1q^{c_1/2})$$
$$\overline{\otimes}\Phi(x_1q^{c_1}))_1(L(x_2q^{c_1/2})\overline{\otimes}\Phi(x_2q^{c_1}))_2+R_{21}(x_2/x_1)^{-1}L(x_1q^{c_1/2})$$
$$\overline{\otimes}\Phi(x_1q^{c_1}))_1(\Phi(x_2)\overline{\otimes}1)_2$$
$$=\Delta\Phi(x_2)_2\Phi(x_1)_1.$$

Similarly, we can prove the rest of the theorem by direct calculation.

This construction of comultiplication follows partially the idea of constructing comultiplications for the quantum Lie algebra,[10] where the cases without the parameter x are given. With our construction, we can extend the Hopf algebra structures to the special Zamolochikov algebra $Z_{n,k}(\xi,\tau)$, which is defined as the algebra generated $\Phi(z)$ with a Belavin elliptic R-matrix $R(z)$.[9] We expect that the new Hopf algebra structure should be very useful in the study of the representation theory of the elliptic Zamolochikov algebras and hopefully even the related Sklyanin elliptic algebras.

III. THE DOUBLE OF $EP[R(x)]$

In this section, we will present a double of the algebra $EP[R(x)]$ following the Drinfeld realization of the quantum affine algebra $U_q(\widehat{\mathfrak{sl}}(2))$.

Definition 3.1: The algebra $DEP[R(x)]$ is an associative algebra generated by $\Phi_i(x)$ indexed by a linear independent basis e_i of $V, l_{ij}(x)$ and $l^*_{ij}(x)$ indexed by the linear independent basis e_{ij} of End(V), $\Phi^*(x)$ indexed by a linear independent basis e^*_i of V^*, the dual space of V, and a central element c. Let $\Phi(x)=\Phi_i(x)\otimes e_i$, $\Phi^*(x)=\Phi^*_i(x)\otimes e^*_i$ the operator valued matrix $L(x)=\Sigma l_{ij}(x)\otimes e_{ij}$, $L^*(x)=\Sigma l^*_{ij}(x)\otimes e_{ij}$, such that $L(x)$ and $L^*(x)$ are invertible. They satisfy the commutation relations:

$$R(x_1/x_2)\Phi(x_1)_1\Phi(x_2)_2=\Phi(x_2)_2\Phi(x_1)_1,$$

$$\Phi(x_1)_1L(x_2)_2=R(q^{c/2}x_1/x_2)^{-1}L(x_2)_2\Phi_i(x_1)_1,$$

$$R(x_1/x_2)L(x_1)_1L(x_2)_2=L(x_2)_2L(x_1)_1R(x_1/x_2).$$

$$R(x_1/x_2)L^*(x_1)_1L^*(x_2)_2=L^*(x_2)_2L^*(x_1)_1R(x_1/x_2),$$

$$R(x_1/x_2q^{-c})L(x_1)_1L^*(x_2)_2=L^*(x_2)_2L^*(x_1)_1R(x_1/x_2q^{c}),$$

$$\Phi^*(x_2)_2\Phi^*(x_1)_1=\Phi^*(x_1)_1\Phi^*(x_2)_2R_{21}(x_2/x_1),$$

$$L^*(x_2)_2\Phi^*(x_1)_1=\Phi^*(x_1)_1L^*(x_2)_2R_{21}(q^{-c/2}x_2/x_1),$$

$$\Phi(x_1)_1\Phi(x_2)_2^* - \Phi^*(x_2)_2\Phi(x_1)_1 = 1/(q-q^{-1})(L^*(wq^{c/2})\delta(z/wq^{-c}) - \delta(z/wq^{c})L^*(zq^{c/2}),$$

$$L(x_1)_1\Phi^*(x_2)_2R_{21}(q^{-c/2}x_2/x_1) = \Phi^*(x_2)_2L(x_1)_1,$$

$$R(q^{c/2}x_1/x_2)L^*(x_1)_1\Phi(x_2)_2 = \Phi(x_2)_2L^*(x_1)_1.$$

Here $\Phi(x_1)_1L(x_2)_2 = \Sigma\Phi_i(x_1)L_{kl}(x_2)e_i\otimes e_{kl}$, $L(x_2)_2\Phi_i(x_1)_1 = \Sigma L_{kl}(x_2)\Phi_i(x_1)e_i\otimes e_{kl}$, $L(x_1)_1L(x_2)_2 = PL(x_1)_2L(x_2)_1P = \Sigma L_{ij}(x_1)L_i(x_2)e_{ij}\otimes e_{kl}$, and the others are defined in the same way. P is the permutation operator. $\delta(z)$ is the distribution with the support at 1.

Theorem 3.1: *$DEP[R(x)]$ has an Hopf algebra structure. The comultiplication Δ, the counit ϵ and the antipode a are given by the following formulas.*

Coproduct Δ,

$$(0)\quad \Delta(q^c) = q^c\otimes q^c,$$

$$(1)\quad \Delta(\Phi_i(z)) = \Phi_i(z)\otimes 1 + \sum L_{ij}(zq^{c_1/2})\otimes\Phi_j(zq^{c_1}),$$

$$(2)\quad \Delta(L_{ij}(z)) = \sum L_{ik}(zq^{-c_2/2})\otimes L_{kj}(zq^{c_1/2}),$$

$$(3)\quad \Delta(\Phi_i^*(z)) = 1\otimes\Phi_i^*(z) + \sum \Phi_j^*(zq^{c_2})\otimes L_{ij}^*(zq^{c_2/2}),$$

$$(2)\quad \Delta(L_{ij}^*(z)) = \sum L_{ik}^*(zq^{c_2/2})\otimes L_{kj}^*(zq^{-c_1/2}),$$

where $c_1 = c\otimes 1$ and $c_2 = 1\otimes c$.

Counit ϵ,

$$\epsilon(q^c) = 1\,\epsilon(L(z)) = \epsilon(L^*(z)) = I,$$

$$\epsilon(\Phi(z)) = 0 = \epsilon(\Phi^*(z)).$$

Antipode a,

$$(0)\quad a(q^c) = q^{-c},$$

$$(1)\quad a(\Phi(z)) = -L(zq^{-c/2})^{-1}\Phi(zq^{-c}),$$

$$(2)\quad a(\Phi^*(z)) = -\Phi^*(zq^{-c})L^*(zq^{-c/2})^{-1},$$

$$(3)\quad a(L(z)) = L(z)^{-1},$$

$$(4)\quad a(L^*(z)) = L^*(z)^{-1}.$$

Proof:

$$\begin{aligned}\Delta\Phi^*(z)_1\Delta\Phi(w)_2^*R_{21}(w/z) &= (1\otimes\Phi^*(z) + \Phi^*(zq^{c_2})\otimes L^*(zq^{c_2/2}))_1(1\otimes\Phi^*(w) + \Phi^*(wq^{c_2})\\ &\quad\otimes L^*(wq^{c_2/2}))_2R_{21}(x_1/x_2)\\ &= (1\otimes\Phi^*(w)_2(1\otimes\Phi^*(z)_1 + (\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2(1\\ &\quad\otimes\Phi^*(z))_1 + (\Phi^*(zq^{c_2})\otimes L^*(zq^{c_2/2}))_1(1\otimes\Phi^*(w))_2R(z/w)^{-1}\end{aligned}$$

$$+(1\otimes\Phi^*(w)+\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2(1\otimes\Phi^*(z)+\Phi^*(zq^{c_2})\otimes L^*(zq^{c_2/2}))_1+(\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2(\Phi^*(zq^{c_2})\otimes L^*(zq^{c_2/2}))_1,$$

$$\Delta\Phi^*(z)_1\Delta L^*(w)_2R_{21}(q^{-(c_1+c_2)/2}w/z)=(1\otimes\Phi^*(z)+\Phi^*(zq^{c_2})\otimes L^*(zq^{c_2/2}))_1(L^*(wq^{c_2/2})\otimes L^*(wq^{-c_1/2}))_2R_{21}(q^{-(c_1+c_2)/2}w/z)$$
$$=\Delta L^*(w)_2\Delta\Phi^*(z)_1,$$

$$\Delta L(z)_1\Delta\Phi^*(w)_2R_{21}(q^{-(c_1+c_2)/2}w/z)=(L(zq^{-c_2/2})\otimes L(zq^{c_1/2})_1(1\otimes\Phi^*(w)+\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2R_{21}(q^{-(c_1+c_2)/2}w/z)$$
$$=(1\otimes\Phi^*(w))_2(L(zq^{-c_2/2})\otimes L(zq^{c_1/2})_1+(L(zq^{-c_2/2})\otimes L(zq^{c_1/2})_1(\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2R_{21}(q^{-(c_1+c_2)/2}w/z)$$
$$=(1\otimes\Phi^*(w))_2(L(zq^{-c_2/2})\otimes L(zq^{c_1/2})_1+(\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2(L(zq^{-c_2/2})\otimes L(zq^{c_1/2})_1$$
$$=\Delta(\Phi^*(w))_2\Delta(L(x_1))_1,$$

$$R(q^{(c_1+c_2)/2}z/w)\Delta L^*(z)_1\Delta\Phi(w)_2=R(q^{(c_1+c_2)/2}z/w)(L^*(zq^{c_2/2})\otimes L^*(zq^{-c_1/2}))_1(\Phi(w)\otimes 1+L(wq^{c_1/2})\otimes\Phi(wq^{c_1}))_2$$
$$=(\Phi(w)\otimes 1)_2(L^*(zq^{c_2/2})\otimes L^*(zq^{-c_1/2}))_1+R_{21}(q^{-(c_1+c_2)/2}w/z)(L^*(zq^{c_2/2})\otimes L^*(zq^{-c_1/2}))_1L(wq^{c_1/2})\otimes\Phi(wq^{c_1}))_2$$
$$=\Delta\Phi(w)_2\Delta L^*(z)_1,$$

$$\Delta\Phi(z)_1\Delta\Phi(w)_2^*-\Delta\Phi^*(w)_2\Delta\Phi(z)_1=(\Phi(z)\otimes 1+L(zq^{c_1/2})\otimes\Phi(zq^{c_1}))_1(1\otimes\Phi^*(w)+\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2-(1\otimes\Phi^*(w)+\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2(\Phi(z)1+L(zq^{c_1/2})\otimes\Phi(zq^{c_1}))_1$$
$$=0+1/(q-q^{-1})(L^*(wq^{c_1/2+c_2})\delta(z/wq^{-c_1-c_2})\otimes L^*(wq^{c_2/2})-\delta(z/wq^{c_1-c_2})L(zq^{c_1/2}))\otimes L^*(wq^{c_2/2})+L(zq^{c_1/2})\otimes(1/(q-q^{-1})(L^*(wq^{c_2/2})\delta(z/wq^{-c_2+c_1})-\delta(z/wq^{c_1+c_2})L(zq^{c_1+c_2/2})+(L(zq^{c_1/2})\otimes\Phi(zq^{c_1}))_1(\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2-(\Phi^*(wq^{c_2})\otimes L^*(wq^{c_2/2}))_2(L(zq^{c_1/2})\otimes\Phi(zq^{c_1}))_1.$$

Because

$$(L(zq^{c_1/2}\otimes 1)_1(\Phi^*(wq^{c_2}))\otimes 1)_2R_{21}(q^{-c_1+c_2}w/z)=(\Phi^*(wq^{c_2}))\otimes 1)_2(L(zq^{c_1/2}\otimes 1)_1,$$

and

$$(R_{21}(q^{-c_1+c_2}w/z))^{-1}(1\otimes\Phi(zq^{c_1}))_1L^*(wq^{c_2/2}))_2=L^*(wq^{c_2/2}))_2(1\otimes\Phi(zq^{c_1}))_1,$$

we have that

$$\Delta\Phi(z)_1\Delta\Phi(w)_2^*-\Delta\Phi^*(w)_2\Delta\Phi(z)_1$$
$$=1/(q-1^{-1})(\Delta L(z/wq^{(c_1+c_2)/2}\delta(z/wq^{-(c_1+c_2)})-\delta(z/wq^c)\Delta L^*(zq^{(c_1+c_2)/2}).$$

Similar calculation gives the complete proof for the theorem.

In all the setting above, $l_{ij}(z)$, $l_{ij}^*(z)$, $\Phi_i(z)$, and $\Phi_i^*(z)$ are functional operators, namely the operator depending the variable z. On the other hand, we can assume that z is a formal variable and $l_{ij}(z)=\Sigma_{n\in\mathbb{Z}}l_{ij}(n)z^{-n}$, $l_{ij}^*(z)=\Sigma_{n\in\mathbb{Z}}l_{ij}^*(n)z^{-n}$, $\Phi_i(z)=\Sigma_{n\in\mathbb{Z}}\Phi_i(n)z^{-n}$, $\Phi_i^*(z)=\Sigma_{n\in\mathbb{Z}}\Phi_i^*(n)z^{-n}$. We can define an algebra $DZP[R(x)]$.

We assume that the entries of $R(z)$ are meromorphic functions. Let $R'(z)=R(z)f(z)$, where $f(z)$ is the common divisor of all the functions $F(z)$, such that $F(z)R(z)$ has no poles.

Definition 3.2: The algebra $DZP[R(x)]$ is an associative algebra generated by $\Phi_i(x)$ indexed by a linear independent basis e_i of $V,l_{ij}(x)$, and $l_{ij}^*(x)$ indexed by the linear independent basis e_{ij} of End(V), $\Phi^*(x)$ indexed by a linear independent basis e_i^* of V^*, the dual space of V, and a central element c. Let $\Phi(x)=\Phi_i(x)\otimes e_i$, $\Phi^*(x)=\Phi_i^*(x)\otimes e_i^*$ the operator valued matrix $L(x)=\Sigma l_{ij}(x)\otimes e_{ij}$, $L^*(x)=\Sigma l_{ij}^*(x)\otimes e_{ij}$, such that $L(x)$ and, $L^*(x)$ are invertible. They satisfy the commutation

$$R'(x_1/x_2)\Phi(x_1)_1\Phi(x_2)_2=f(x_1/x_2)\Phi(x_2)_2\Phi(x_1)_1,$$

$$\Phi(x_1)_1L(x_2)_2=R(q^{c/2}x_1/x_2)^{-1}L(x_2)_2\Phi_i(x_1)_1,$$

$$R(x_1/x_2)L(x_1)_1L(x_2)_2=L(x_2)_2L(x_1)_1R(x_1/x_2).$$

$$R(x_1/x_2)L^*(x_1)_1L^*(x_2)_2=L^*(x_2)_2L^*(x_1)_1R(x_1/x_2),$$

$$R(x_1/x_2q^{-c})L(x_1)_1L^*(x_2)_2=L^*(x_2)_2L^*(x_1)_1R(x_1/x_2q^c),$$

$$f(x_2/x_1)\Phi^*(x_2)_2\Phi^*(x_1)_1=\Phi^*(x_1)_1\Phi^*(x_2)_2R'_{21}(x_2/x_1),$$

$$L^*(x_2)_2\Phi^*(x_1)_1=\Phi^*(x_1)_1L^*(x_2)_2R_{21}(q^{-c/2}x_2/x_1),$$

$$\Phi(x_1)_1\Phi(x_2)_2^*-\Phi^*(x_2)_2\Phi(x_1)_1=1/(q-q^{-1})(L^*(wq^{c/2})\delta(z/wq^{-c})-\delta(z/wq^c)L^*(zq^{c/2}),$$

$$L(x_1)_1\Phi^*(x_2)_2R_{21}(q^{-c/2}x_2/x_1)=\Phi^*(x_2)_2L(x_1)_1,$$

$$R(q^{c/2}x_1/x_2)L^*(x_1)_1\Phi(x_2)_2=\Phi(x_2)_2L^*(x_1)_1.$$

Here $\Phi(x_1)_1L(x_2)_2=\Sigma\Phi_i(x_1)L_{kl}(x_2)e_i\otimes e_{kl}$, $L(x_2)_2\Phi_i(x_1)_1=\Sigma L_{kl}(x_2)\Phi_i(x_1)e_i\otimes e_{kl}$, $L(x_1)_1L(x_2)_2=PL(x_1)_2L(x_2)_1P=\Sigma L_{ij}(x_1)L_i(x_2)e_{ij}\otimes e_{kl}$, and the others are defined in the same way as above. P is the permutation operator. $\delta(z)=\Sigma_{n\in\mathbb{Z}}z^n$. The operator $R(z)$ and $R_21(z)$ are expanded in appropriate directions.

If the poles of the matrix of $R(z)$ are beyond a finite disk around zero, we can always impose the condition that $l_{kl}(n)=0=l_{kl}^*(-n)=l_{ij}(0)=l_{ji}^*(0)$, for $n<0$, $i<j$. Then the condition of the invertibility is equivalent and requires that $l_{ii}(0)$ and $l_{ii}^*(0)$ are invertible.

Example 3.1: Let V be one dimensional, and $R(z)=z-wq^2/zq^2-w$. Let $l_{11}(n)=0=l_{11}^*(-n)$, for $n<0$. Then the algebra $DZP[R(x)]$ is the quantum affine algebra $U_q(\widehat{\mathfrak{sl}}(2))$. If we choose $R(z)$ to be other functions with the property $R(z)=(R(z^{-1})^{-1}$, then it is an algebra defined in Ref. 6 as a generalization of the the quantum affine algebra $U_q(\widehat{\mathfrak{sl}}(2))$.

Example 3.2: Let $V=C^n$ and $R(z)=\Sigma(z-wq^2)/(zq^2-w)e_{ii}\otimes e_{ii}+\Sigma_{|i-j|=1}(z-wq^{-1})/(zq^{-1}-w)(e_{j,j}\otimes e_{i,i})+\Sigma_{|i-j|>1}e_{ii}\otimes e_{jj}$. Let $l_{kl}(n)=0=l^*_{kl}(-n)=l_{ij}(0)=l^*_{ji}(0)$, for $n<0$, $i<j$. Then the algebra $DZP[R(x)]$ is an algebra, whose quotient (modular the cubic relations) is $U_q(\widehat{\mathfrak{sl}}(n))$. If we substitute $z-wq^2/zq^2-w$ and $(z-wq^{-1})/(zq^{-1}-w)$ by other functions, it will be the generalization of $U_q(\widehat{\mathfrak{sl}}(n))$ without the cubic relations.[6]

Example 3.3: Let $V=C^n$ and $R(z)$ be the projection of the universal R-matrix $\mathfrak{R}$ of $U_q(\widehat{\mathfrak{sl}}(n))$. Let $l_{kl}(n)=0=l^*_{kl}(-n)=l_{ij}(0)=l^*_{ji}(0)$, for $n<0$, $i<j$. The operator $L(z)$ can be identified with the operator $(\mathrm{id}\otimes\pi_V)\mathfrak{R}_21(zq^{c/2})$ and $L^*(z)$ with the operator $(\mathrm{id}\otimes\pi_V)\mathfrak{R}^{-1}(z^{-1}q^{-c/2})$. The subalgebra generated by $L(z)$ and $L^*(z)$ is isomorphic to $U_q(\widehat{\mathfrak{sl}}(n))$. We conjecture that the algebra $DZP[R(x)]$ is isomorphic to $U_q(\widehat{\mathfrak{sl}}(n+1))$. This is because when q goes to 1, this algebra degenerates into $\widehat{\mathfrak{sl}}(n+1)$. Such a formulation should have application in construction of quantum W-algebras.[13] Also if we take $z=0$, it becomes that in Refs. 11 and 14.

From the definition, we can see both the subalgebra generated by $\Phi(z)$, $L(z)$, and $L^*(z)$ and the subalgebra generated by $\Phi^*(z)$, $L(z)$, and $L^*(z)$ are the Hopf algebras. If we take $R(z)$ to be the projection of the universal R-matrix $\mathfrak{R}$ of $U_q(\widehat{\mathfrak{sl}}(n))$ on finite dimensional representations, we will derive unconventional Hopf algebras from those subalgebras.

Our new algebras can be viewed as a simple generalization of the Drinfeld realization of the quantum affine algebra $U_q(\widehat{\mathfrak{sl}}(2))$, where the function $g(z)$ is substitute by a matrix $R(z)$, the operators are substituted by the vector valued operators and the relations looks the same. However such a generalization is highly nontrivial in the sense that all the Hopf algebra structures are preserved, in the other words, those new algebras are Hopf algebras, whose comultiplication, counit, and antipode symbolically are the same. These new Hopf algebras should be very useful in various applications in mathematics and physics, for example, the study of the representation theory of the elliptic Zamolochikov algebras.[15]

ACKNOWLEDGMENTS

We would like to thank M. Jimbo and B. Feigin for useful discussions. This project is supported by a grant from the Ministry of Education of Japan.

[1] L. D. Faddeev, N. Yu. Reshetikhin, and L. A. Takhtajan, *Quantization of Lie Groups and Lie Algebras, Yang–Baxter equation in Integrable Systems*, Advanced Series in Mathematical Physics Vol. 10 (World Scientific, Singapore, 1989), pp. 299–309.
[2] N. Yu. Reshetikhin, and M. A. Semenov-Tian-Shansky, ''Central Extensions of Quantum Current Groups,'' Lett. Math. Phys. **19**, 133–142 (1990).
[3] V. G. Drinfeld, ''New realization of Yangian and quantum affine algebra,'' Soviet Math. Doklady **36**, 212–216 (1988).
[4] J. Ding and I. B. Frenkel, ''Isomorphism of two realizations of quantum affine algebra $U_q(\widehat{\mathfrak{gl}}(n))$,'' Commun. Math. Phys. **156**, 277–300 (1993).
[5] J. Ding and T. Miwa, ''Zeros and poles of quantum current operators and the condition of quantum integrability,'' Publ. RIMS **33**, 277–284 (1997).
[6] J. Ding and K. Iohara, ''Generalization and deformation of the quantum affine algebras,'' Lett. Math. Phys. **41**, 181–193 (1997).
[7] I. B. Frenkel and N. Yu. Reshetikhin, ''Quantum affine algebras and holonomic difference equations,'' Commun. Math. Phys. **146**, 1–60 (1992).
[8] B. Davis, O. Foda, M. Jimbo, T. Miwa, and A. Nakayashiki, ''Diagonalization of the *XYZ* Hamiltonian by vertex operators,'' Commun. Math. Phys. **151**, 89–153 (1993).
[9] B. Feigin and V. Odesski, ''Sklyanin Elliptic algebras,'' Funkts. Anal. Prilozhen. **23**, 45–54 (1989).
[10] D. Bernard, ''Quantum Lie algebra and differential calculus on quantum groups,'' preprint, SPht-90-119 (1990).
[11] S. Majid, ''New Quantum Groups by Double-Bosonisation,'' q-alg/9610004 (1996).
[12] V. G. Drinfeld, ''Hopf algebra and the quantum Yang–Baxter Equation,'' Dokl. Akad. Nauk **283**, 1060–1064 (1985).
[13] E. Frenkel and B. Feigin, ''Quantum $\mathcal{W}$-algebra and elliptic algebras RIMS-1027,'' q-alg/9508009 (1995).
[14] M. Jimbo, ''A *q*-difference analogue of $U(\mathfrak{g})$ and Yang-Baxter equation,'' Lett. Math. Phys. **10**, 63–69 (1985).
[15] E. K. Sklyanin, ''On some algebraic structures related to the Yang–Baxter equation,'' Funkts. Anal. Prilozhen **16**, 22–34 (1982).

Conservation laws for linear equations on braided linear spaces

M. Klimek[a)]
Institute of Mathematics and Computer Science, Technical University of Częstochowa, ul.Dąbrowskiego 73, 42-200 Częstochowa, Poland

(Received 29 April 1999; accepted for publication 11 May 1999)

The properties of linear equations on braided linear spaces are investigated and conservation laws for them are derived. The conserved currents are given in the explicit form. The procedure is then applied to scalar wave equations on a quantum plane and on q-Minkowski space. © *1999 American Institute of Physics.* [S0022-2488(99)02208-2]

I. INTRODUCTION

In the classical field theory we can derive conservation laws for linear equations using the Takahashi–Umezawa method.[1] We have shown previously that it can be extended to models on lattices[2–4] and on quantum Minkowski spaces by Podleś and Woronowicz.[5–10] The aim in this paper is to develop a similar procedure for linear equations on braided linear spaces.

The class of braided linear spaces was introduced by Majid[11] as braided Hopf groups with a full structure of the Hopf group consistent with a braided tensor product:

$$(a\otimes c)(b\otimes d)=a\Psi(c\otimes b)d.$$

The important examples of braided linear spaces are q-Euclidean and q-Minkowski space endowed, respectively, with action of the $SO_q(4)$ and q-Poincaré algebra. For q-Minkowski space investigated equations include the scalar, spinor and vector wave equations.[12,13] In the last section we show how to obtain the conserved currents for a scalar wave equation on q-Minkowski space.

Let us notice that technically the procedure of the derivation of the conserved currents is very similar to the method introduced for discrete space and for quantum Minkowski spaces mentioned earlier.

The paper is organized as follows: In Secs. II and III we reformulate the Leibnitz rule for braided differential calculus. It appears that the deformation is given by a certain transformation operator. This operator differs from the ones investigated earlier in discrete and noncommutative models. Nevertheless, the properties of the transformation operator are analogous. There also exists the inverse transformation operator due to the bi-invertibility of the R matrix. This operator is then used in the modification of the Leibnitz rule. In Sec. IV we solve uniquely and explicitly a certain operator equation for the operator Γ_μ that enables us to construct the conserved currents in Sec. V. Section VI includes an application of the presented method to wave equations on the quantum plane and on q-Minkowski space.

II. BRAIDED LINEAR SPACES

Let us begin with a brief sketch of properties of braided covector space. This structure is a braided Hopf algebra defined by generators 1, x_i ($i=1,...,n$) and relations of the braided group structure:

$$x_i x_j=(R')^{kl}_{ij}x_l x_k, \tag{1}$$

[a)]Electronic mail: klimek@matinf.pcz.czest.pl

$$\Delta x_i = x_i \otimes 1 + 1 \otimes x_i \,, \tag{2}$$

$$\epsilon x_i = 0, \quad S x_i = -x_i \,, \tag{3}$$

$$\Psi(x_i \otimes x_j) = x_k \otimes x_l R^{lk}_{ij} \,, \tag{4}$$

where bi-invariant matrix R fulfills QYBE and besides the first inverse R^{-1} there also exists the second inverse $\widetilde{R}$:

$$(R^{-1})^{ij}_{kl} R^{kl}_{ab} = R^{ij}_{kl} (R^{-1})^{kl}_{ab} = \delta^i_a \delta^j_b \,, \tag{5}$$

$$\widetilde{R}^{ib}_{al} R^{ak}_{jb} = R^{ib}_{al} \widetilde{R}^{ak}_{jb} = \delta^i_j \delta^k_l \,. \tag{6}$$

The matrix R' defining the commutation rule for coordinates (1) is an invertible matrix given by the formula[11]

$$R' = P + P \Pi_{j \neq i} (PR - \lambda_j), \tag{7}$$

with P the permutation matrix and λ_i the nonzero eigenvalue of the minimal polynomial of PR:

$$\Pi_j (PR - \lambda_j) = 0.$$

Differentiation on braided linear space was defined by Majid[14,15] as an infinitesimal translation, and this leads to the following properties of partial derivatives:

$$\partial^i \partial^j = (R')^{ij}_{kl} \partial^l \partial^k, \tag{8}$$

$$\partial^i x_j = \delta^i_j + R^{ki}_{jl} x_k \partial^l. \tag{9}$$

We shall consider the functions of coordinates on braided linear spaces understood as a formal power series:

$$f(\vec{x}) = \sum_{l=0}^{\infty} a^{k_0 \cdots k_l} x_{k_0} \cdots x_{k_l}. \tag{10}$$

For functions, the commutation rule (9) transforms into the Leibnitz rule given by Majid[11,14,15] in the form

$$\partial^i (fg) = (\partial^i f) g + \cdot \Psi^{-1} (\partial^i \otimes f) g, \tag{11}$$

where the inverse braiding looks as follows on monomials of arbitrary order [e^i is a basic covector that means $(e^i)_j = \delta^i_j$]:

$$\Psi^{-1}(\partial^i \otimes x_1 \cdots x_k) = e^i_1 x_2 \cdots x_k x_{k+1} (PR)_{12} \cdots (PR)_{k,k+1} \otimes \partial^{k+1}. \tag{12}$$

It is easy to deduce that the Leibnitz rule (11) can be rewritten similarly to the ones appearing in the discrete calculus[4] and in the differential calculus on quantum Minkowski space by Podleś and Woronowicz.[5,10] In this formula the deformation of the classical Leibnitz rule is described by the transformation operator ζ^i_j:

$$\partial^i (fg) = (\partial^i f) g + (\zeta^i_j f) \partial^j g. \tag{13}$$

The transformation operator for braided differential calculus is defined by formula (12) and on an arbitrary monomial of the first order is given by

$$\zeta^i_j x_k = R^{li}_{kj} x_l . \tag{14}$$

We extend the ζ operator to arbitrary function (10) using its multiplicity property, namely,

$$\zeta^i_j(fg) = (\zeta^i_k f)(\zeta^k_j g). \tag{15}$$

Let us be reminded that in the classical differential calculus the transformation operator is the identity operator:

$$\zeta^i_j = \delta^i_j .$$

For the discrete calculus it is the shift operator on the lattice in the given direction,[4] while on the quantum Minkowski space it is defined by its multiplicity and action on monomials of the first order.[5]

Now we can modify the Leibnitz rule (13) in such a way as to obtain on the right-hand side operators acting only on one of the functions in the product. To this aim we use the inverse transformation operator ζ^-:

$$\zeta^k_j \zeta^{-\ i}_k = \zeta^{-\ k}_j \zeta^i_k = \delta^i_j . \tag{16}$$

The existence of the inverse operator is due to the bi-invertibility of the R matrix, and it is defined by its action on monomials of the first order:

$$\zeta^{-\ i}_j x_k = \tilde{R}^{li}_{kj} x_l , \tag{17}$$

and the multiplicity property:

$$\zeta^{-\ i}_j(fg) = (\zeta^{-\ k}_j f)(\zeta^{-\ i}_k g). \tag{18}$$

The properties of the transformation operators ζ and ζ^- imply the following modified Leibnitz rule:

$$\partial^k[(\zeta^{-\ i}_k f)g] = (-\partial^{\dagger\ i} f)g + f(\partial^i g) = f(-\overleftarrow{\partial}^{\dagger\ i} + \partial^i)g, \tag{19}$$

where we have denoted as a conjugated derivative,

$$\partial^{\dagger\ i} := -\partial^k \zeta^{-\ i}_k . \tag{20}$$

We see that after modification we deal with the Leibnitz rule, where the right-hand side is analogous to the classical differential calculus:

$$\partial^i(fg) = f(-\overleftarrow{\partial}^{\dagger\ i} + \partial^i)g,$$

with the exception that classically $\partial^{\dagger\ i} = -\partial^i$, while in our noncommutative case it is deformed by the inverse transformation operator (19), (20).

Let us notice that the form of the modified Leibnitz rule for braided differential calculus (19) is identical with the formulas derived for discrete and quantum models; the only difference is enclosed in the explicit formula for the transformations operators ζ and ζ^-.

III. LINEAR EQUATIONS ON BRAIDED LINEAR SPACES

We shall consider equations of the form

$$\Lambda(\partial)\Phi = 0, \tag{21}$$

$$\Lambda(\partial)=\Lambda_0+\sum_{l=1}^{N}\Lambda_{\mu_1\cdots\mu_l}\partial^{\mu_1}\cdots\partial^{\mu_l}. \tag{22}$$

They are linear equations with coefficients (which may be matrices) fulfilling the following conditions:

$$\sum_{\mu_k}\partial^{\mu_k}\Lambda_{\mu_1\cdots\mu_k\cdots\mu_l}=0, \tag{23}$$

$$\sum_{\mu_k}\zeta_j^{\mu_k}\Lambda_{\mu_1\cdots\mu_k\cdots\mu_l}=\Lambda_{\mu_1\cdots j\cdots\mu_l}, \tag{24}$$

where $l=1,...,N,\ k=1,...,l$.

The class of equations obeying (23), (24), in general, admits the coefficients depending on coordinates. It includes equations with constant coefficients such as, for example, a scalar, spinor, and vector wave equation on q-Minkowski space.[12,13] For constant coefficients the condition (23) can be replaced with stronger one:

$$\partial^i\Lambda_{\mu_1\cdots\mu_l}=0, \tag{25}$$

for $l=1,...,N$ and $i=1,...,N$.

Finally, due to the commutation relation for partial derivatives (8), the coefficients of Eq. (21) fulfill the symmetry condition, namely,

$$(R')^{\mu_k\mu_{k+1}}{}_{ij}\Lambda_{\mu_1\cdots\mu_k\mu_{k+1}\cdots\mu_l}=\Lambda_{\mu_1\cdots ji\cdots\mu_l},\quad l=1,...,N. \tag{26}$$

In earlier papers we investigated an analogous class of equations on discrete spaces and on quantum Minkowski space. The properties (23), (24) of coefficients together with the modified Leibnitz rule (19) allowed us to derive the conservation laws for arbitrary linear equation fulfilling these conditions. The crucial point in the construction, just as in the classical procedure of Takahashi and Umezawa,[1] is the solution of an operator equation for the operator Γ_μ:

$$\sum_{\mu}(-\overleftarrow{\partial}^{\dagger\,\mu}+\partial^\mu)\circ\Gamma_\mu(\partial,\overleftarrow{\partial}^\dagger)=\Lambda(\partial)-\Lambda(\overleftarrow{\partial}^\dagger), \tag{27}$$

where the operator $\Lambda(\overleftarrow{\partial}^\dagger)$ of the conjugated equation looks as follows:

$$\Lambda(\overleftarrow{\partial}^\dagger)=\Lambda_0+\sum_{l=1}^{N}\overleftarrow{\partial}^{\dagger\,\mu_1}\cdots\overleftarrow{\partial}^{\dagger\,\mu_l}\Lambda_{\mu_1\cdots\mu_l}. \tag{28}$$

We introduced the notation for the product "$\circ$" to underline the way it acts on monomials of derivatives:

$$\begin{aligned}(-\overleftarrow{\partial}^{\dagger\,\mu}+\partial^\mu)\circ\overline{[\nu_1,\ldots,\nu_l]}a(\vec{x})[\rho_1,\ldots,\rho_k]:=&-\overline{[\nu_1,\ldots,\nu_l,\mu]}a(\vec{x})[\rho_1,\ldots,\rho_k]\\&+\overline{[\nu_1,\ldots,\nu_l]}\partial^\mu a(\vec{x})[\rho_1,\ldots,\rho_k],\end{aligned} \tag{29}$$

where we have denoted the monomials of derivatives as follows:

$$[\rho_1,\ldots,\rho_k]:=\partial^{\rho_1}\cdots\partial^{\rho_k}, \tag{30}$$

$$\overline{[\nu_1,\ldots,\nu_l]}:=\overleftarrow{\partial}^{\dagger\,\nu_1}\cdots\overleftarrow{\partial}^{\dagger\,\nu_l}. \tag{31}$$

IV. THE CONSTRUCTION OF THE Γ_μ OPERATOR

The above equation (27) for the Γ_μ operator can be uniquely solved according to the following proposition.

Proposition 4.1: The unique solution of (27) in the class of polynomials of derivatives $\overleftarrow{\partial}^\dagger$ and ∂ for the equation operator Λ fulfilling (23), (24) is of the form

$$\Gamma_\mu(\partial,\overleftarrow{\partial}^\dagger)=\Lambda_\mu+\sum_{l=1}^{N-1}\sum_{k=0}^{l}\overleftarrow{\partial}^{\dagger\,\mu_1}\cdots\overleftarrow{\partial}^{\dagger\,\mu_k}\Lambda_{\mu_1\cdots\mu_k\mu\mu_{k+1}\cdots\mu_l}\partial^{\mu_{k+1}}\cdots\partial^{\mu_l}. \tag{32}$$

Proof: the proof for braided differential calculus is analogous to the one conducted in Refs. 4, 5 for discrete and quantum differential calculi so we enclose the shortened version in Appendix A.

Additionally, as the condition (23) is weaker than the condition for constant coefficients (25), we conclude that for this case the Γ_μ operator is given by the analogous proposition.

Proposition 4.2: The unique solution of (27) in the class of polynomials of derivatives $\overleftarrow{\partial}^\dagger$ and ∂ for the equation operator Λ with constant coefficients fulfilling (24), (25) is of the form

$$\Gamma_\mu(\partial,\overleftarrow{\partial}^\dagger)=\Lambda_\mu+\sum_{l=1}^{N-1}\sum_{k=0}^{l}\overleftarrow{\partial}^{\dagger\,\mu_1}\cdots\overleftarrow{\partial}^{\dagger\,\mu_k}\Lambda_{\mu_1\cdots\mu_k\mu\mu_{k+1}\cdots\mu_l}\partial^{\mu_{k+1}}\cdots\partial^{\mu_l}. \tag{33}$$

Proof: this is a corollary valid by Proposition 4.1.

The construction of the Γ_μ operator enables us to derive the conservation laws for linear equations in classical commutative models. This is not the case for equations on the discrete and noncommutative spaces due to the deformation of the Leibnitz rule.

We introduce the operator $\hat{\Gamma}_\mu$,

$$\hat{\Gamma}_\mu(\partial,\overleftarrow{\partial}^\dagger)=\overleftarrow{\zeta}_\mu^{-\,j}\Lambda_j+\sum_{l=1}^{N-1}\sum_{k=0}^{l}\overleftarrow{\partial}^{\dagger\,\mu_1}\cdots\overleftarrow{\partial}^{\dagger\,\mu_k}\overleftarrow{\zeta}_\mu^{-\,j}\Lambda_{\mu_1\cdots\mu_k j\mu_{k+1}\cdots\mu_l}\partial^{\mu_{k+1}}\cdots\partial^{\mu_l}, \tag{34}$$

which for a pair of arbitrary functions F and G is connected with the Γ_μ operator by the equality

$$\sum_\mu \partial^\mu F\,\hat{\Gamma}_\mu(\partial,\overleftarrow{\partial}^\dagger)G=\sum_\mu F(-\overleftarrow{\partial}^{\dagger\,\mu}+\partial^\mu)\circ\Gamma_\mu(\partial,\overleftarrow{\partial}^\dagger)G. \tag{35}$$

The operator $\hat{\Gamma}_\mu$ shall be used in an explicit construction of the conserved currents in the same way the operator Γ_μ was applied in the classical procedure for commutative differential calculus.

V. THE CONSERVATION LAWS FOR LINEAR EQUATIONS ON BRAIDED LINEAR SPACES

We use the properties of the $\hat{\Gamma}_\mu$ operator derived in the previous section to prove the following proposition which allows us to obtain the conservation law for arbitrary equation obeying conditions (23), (24).

Proposition 5.1: Let us assume that function Φ is an arbitrary solution of equation (21) with coefficients fulfilling (23), (24), which means

$$\Lambda(\partial)\Phi=0, \tag{36}$$

and function Φ' solves the conjugated equation

$$\Phi'\Lambda(\overleftarrow{\partial}^\dagger)=0. \tag{37}$$

Then

$$J_\mu = \Phi' \hat{\Gamma}_\mu(\partial, \bar{\partial}^\dagger)\Phi, \tag{38}$$

where the operator $\hat{\Gamma}_\mu$ *is defined by (34), is a current that obeys the conservation law on braided linear space:*

$$\sum_\mu \partial^\mu J_\mu = 0. \tag{39}$$

Proof: The conservation law results from the modified Leibnitz rule (19), and from the properties of the $\hat{\Gamma}$ and Γ operators:

$$\sum_\mu \partial^\mu J_\mu = \sum_\mu \partial^\mu \Phi' \hat{\Gamma}_\mu(\partial, \bar{\partial}^\dagger)\Phi = \Phi' \left(\sum_\mu (-\bar{\partial}^{\dagger\,\mu} + \partial^\mu)\circ\Gamma_\mu(\partial, \bar{\partial}^\dagger) \right)\Phi = \Phi'(\Lambda(\partial) - \Lambda(\bar{\partial}^\dagger))\Phi$$
$$= 0. \tag{40}$$

Thus, the conservation law for an arbitrary linear equation fulfilling (23), (24) is valid provided functions Φ' and Φ are solutions of the corresponding equations:

$$\Lambda(\partial)\Phi = 0, \quad \Phi'\Lambda(\bar{\partial}^\dagger) = 0. \tag{41}$$

Let us notice that for Eqs. (21) with constant coefficients fulfilling stronger conditions (24), (25) the Proposition 5.1 is also valid due to Proposition 4.2. The formula (38) shall be used in examples enclosed in the next section.

VI. APPLICATIONS

We shall now apply the general formulas of Propositions 4.2 and 5.1 to scalar wave equations on two braided linear spaces: on quantum plane and on the q-Minkowski space, which is the most interesting from the physical point of view.

A. The conserved currents for scalar wave equation on a quantum plane

The commutation relations for derivatives and coordinates on a quantum plane look as follows:[11]

$$yx = qxy, \quad \partial^y \partial^x = q^{-1}\partial^x \partial^y. \tag{42}$$

The Leibnitz rule (13) in differential calculus is defined by the following R matrix:

$$R = \begin{bmatrix} q^2 & 0 & 0 & 0 \\ 0 & q & q^2-1 & 0 \\ 0 & 0 & q & 0 \\ 0 & 0 & 0 & q^2 \end{bmatrix}, \quad R' = q^{-2}R. \tag{43}$$

It is easy to check the explicit form of the transformation operator for monomials of the first order:

$$\zeta_x^x x = q^2 x, \quad \zeta_x^x y = qy, \quad \zeta_y^x x = (q^2-1)y, \quad \zeta_y^x y = 0, \tag{44}$$

$$\zeta_y^y y = q^2 y, \quad \zeta_y^y x = qx, \quad \zeta_x^y y = 0, \quad \zeta_x^y x = 0. \tag{45}$$

These formulas together with the condition (16), yield the inverse transformation operator, which can be described explicitly on monomials of the first order as

$$\zeta_x^{-\,x} x = q^{-2}x, \quad \zeta_x^{-\,x} y = q^{-1}y, \quad \zeta_y^{-\,x} x = (q^{-2}-1)y, \quad \zeta_y^{-\,x} y = 0, \tag{46}$$

$$\zeta_y^{-\ y} y = q^{-2} y, \quad \zeta_y^{-\ y} x = q^{-1} x, \quad \zeta_x^{-\ y} y = 0, \quad \zeta_x^{-\ y} x = 0. \tag{47}$$

We shall investigate the wave equation on a quantum plane:

$$\Box \Phi = \partial^a \partial^b g_{ba} \Phi = (\partial^x \partial^y + q \partial^y \partial^x) \Phi = 0. \tag{48}$$

From the general formula for the Γ_μ operator (32), we obtain in our case the following operator of the first order:

$$\Gamma_x = q \overleftarrow{\partial}^{\dagger\, y} + \partial^y, \quad \Gamma_y = \overleftarrow{\partial}^{\dagger\, x} + q \partial^x. \tag{49}$$

The modified Γ operator looks as follows (34):

$$\hat{\Gamma}_x = \overleftarrow{\partial}^{\dagger\, x} \overleftarrow{\zeta}_x^{-\ y} + q\, \overleftarrow{\partial}^{\dagger\, y} \overleftarrow{\zeta}_x^{-\ x} + q \overleftarrow{\zeta}_x^{-\ y} \partial^x + \overleftarrow{\zeta}_x^{-\ x} \partial^y, \tag{50}$$

$$\hat{\Gamma}_y = \overleftarrow{\partial}^{\dagger\, x} \overleftarrow{\zeta}_y^{-\ y} + q\, \overleftarrow{\partial}^{\dagger\, y} \overleftarrow{\zeta}_y^{-\ x} + q \overleftarrow{\zeta}_y^{-\ y} \partial^x + \overleftarrow{\zeta}_y^{-\ x} \partial^y. \tag{51}$$

Having constructed the $\hat{\Gamma}_\mu$ operator we are able to derive the conservation laws and explicit conserved currents for the wave equation. To this aim we need the solutions of (48) and of its conjugation:

$$\Phi' \overleftarrow{\Box}^{\dagger} = \Phi' (\overleftarrow{\partial}^{\dagger\, y} \overleftarrow{\partial}^{\dagger\, x} + q \overleftarrow{\partial}^{\dagger\, x} \overleftarrow{\partial}^{\dagger\, y}) = 0. \tag{52}$$

Let us notice that the function

$$\Phi(x,y) = \phi_1(x) + \phi_2(y), \tag{53}$$

where ϕ_1 and ϕ_2 are arbitrary functions is always a solution for the equation (48). More solutions can be produced from the above solution Φ using the symmetry operators that transform solutions into solutions. The set of symmetry operators for (48) includes momenta $P^x = \partial^x$, $P^y = \partial^y$; rotation operator $M = x \partial^x - y \partial^y$; dilatation operator $D = x \partial^x + y \partial^y$; and conformal boosts $K^x = (xy) \partial^x - qyD$ and $K^y = (xy) \partial^y - xD$.

These operators are quantum deformations for conformal algebra on two-dimensional space. Additionally, one can show that the operators from infinite-dimensional set:

$$L^m = x^m \partial^x, \quad \tilde{L}^n = y^n \partial^y, \quad n,m > 1, \tag{54}$$

also produces the solution from the solution of the wave equation (48).

According to Proposition 5.1 we obtain the currents connected with deformed conformal algebra $\delta^\omega \in \{P^x, P^y, M, D, K^x, K^y\}$:

$$J_\mu^\omega = \Phi' \hat{\Gamma}_\mu (\partial, \overleftarrow{\partial}^{\dagger}) \delta^\omega \Phi, \tag{55}$$

which fulfill the conservation law:

$$\partial^x J_x^\omega + \partial^y J_y^\omega = 0. \tag{56}$$

The operators L^m and $\tilde{L}^n$ add to the set of conserved currents the following expressions:

$$J_\mu^m = \Phi' \hat{\Gamma}_\mu (\partial, \overleftarrow{\partial}^{\dagger}) L^m \Phi, \quad \tilde{J}_\mu^n = \Phi' \hat{\Gamma}_\mu (\partial, \overleftarrow{\partial}^{\dagger}) \tilde{L}^n \Phi, \quad n,m > 1. \tag{57}$$

B. The conserved currents for scalar wave equation on *q*-Minkowski space

Let us now pass to the scalar wave equation on q-Minkowski space. We have studied this equation in Ref. 16 within the framework given by Azcárraga, Kulish, and Rodenas.[17,18] Now we would like to derive conservation laws using the formalism by Majid, the more interesting that in Ref. 19 he introduced q-conformal algebra in the scalar and spinor representation and showed that it is also a braided Hopf algebra.

We shall follow the notations of Majid from the mentioned paper. The momenta fulfill the conditions

$$p^i p^j = (R')^{ij}_{kl} p^l p^k, \tag{58}$$

$$p^i x_j = -\delta^i_j + x_k (R^{-1})^{ik}_{lj} p^l. \tag{59}$$

The q-conformal algebra includes momenta $\mathbf{p}$, rotations $\mathbf{l}^+$ and $\mathbf{l}^-$, dilatation operator $\mathbf{s}$, and special conformal generators $\mathbf{c}$. They form braided Hopf algebra with algebraic relations (written in vector form):

$$\mathbf{l}_1^\pm \mathbf{l}_2^\pm \mathbf{R} = \mathbf{R} \mathbf{l}_2^\pm \mathbf{l}_1^\pm, \quad \mathbf{l}_1^+ \mathbf{l}_2^- \mathbf{R} = \mathbf{R} \mathbf{l}_2^- \mathbf{l}_1^+, \tag{60}$$

$$\mathbf{l}_1^+ \mathbf{p}_2 = \frac{1}{\lambda} \mathbf{R}_{21}^{-1} \mathbf{p}_2 \mathbf{l}_1^+, \quad \mathbf{l}_1^- \mathbf{p}_2 = \lambda \mathbf{R} \mathbf{p}_2 \mathbf{l}_1^-, \tag{61}$$

$$\mathbf{s}\mathbf{p} = \frac{\mathbf{1}}{\lambda} \mathbf{p}\mathbf{s}, \quad \mathbf{l}^\pm \mathbf{s} = \mathbf{s} \mathbf{l}^\pm, \tag{62}$$

and for special conformal generators we have

$$\mathbf{c}_2 \mathbf{c}_1 = \mathbf{c}_1 \mathbf{c}_2 \mathbf{R}', \quad \mathbf{s}\mathbf{c} = \lambda \mathbf{c}\mathbf{s}, \tag{63}$$

$$\mathbf{l}_1^+ \mathbf{c}_2 = \lambda \mathbf{c}_2 \mathbf{l}_1^+ \mathbf{R}_{21}, \quad \mathbf{l}_1^- \mathbf{c}_2 = \frac{1}{\lambda} \mathbf{c}_2 \mathbf{l}_1^{-1} \mathbf{R}^{-\mathbf{1}}, \tag{64}$$

with the following commutation relation between momenta and special conformal generators:

$$\mathbf{p}\mathbf{c} = \mathbf{c}\mathbf{p} + \frac{\mathbf{l}^+ \mathbf{s}^{-\mathbf{1}} - \mathbf{l}^- \mathbf{s}}{q - q^{-1}}. \tag{65}$$

We see from (59) that in our case the transformation operators ζ and ζ^- for $-\mathbf{p}$ look as follows:

$$\zeta^i_j x_m = (R^{-1})^{ik}_{jm} x_k, \tag{66}$$

$$\zeta^{-\ i}_j x_m = \hat{R}^{ik}_{jm} x_k, \tag{67}$$

where the matrix $\hat{R}$ is the second inverse of R^{-1} fulfilling the condition

$$\hat{R}^{ib}_{al} (R^{-1})^{ak}_{jb} = (R^{-1})^{ib}_{al} \hat{R}^{ak}_{jb} = \delta^i_j \delta^k_l. \tag{68}$$

Now, the scalar wave equation on q-Minkowski space is of the form

$$\Box \Phi = p^i p^j g_{ji} \Phi = 0. \tag{69}$$

The set of symmetry operators for the above equation includes momenta $\mathbf{p}$, rotation operators $\mathbf{l}^\pm$, dilatation $\mathbf{s}$, and conformal boosts $\mathbf{K} = \mathbf{c} + (\mathbf{x}_1 \mathbf{l}_1^+ \mathbf{s}^{-\mathbf{1}} - \mathbf{x}_1 \mathbf{l}_1^- \mathbf{s})/(q - q^{-1})$, where $\mathbf{x}_1 \mathbf{l}_1^\pm$ is a vector with components $(\mathbf{x}_1 \mathbf{l}_1^\pm)_k = x_m l^{\pm\ m}_k$.

The momenta and rotations commute with the d'Alembert operator (69) due to the following properties of R and R' matrices,[11]

$$g_{ka}R^{ia}_{jl}=\frac{1}{\lambda^2}(R^{-1})^{ia}_{jk}g_{al}, \quad g_{ka}\tilde{R}^{ia}_{jl}=\lambda^2 R^{ia}_{jk}g_{al}, \tag{70}$$

$$g_{ai}R^{ak}_{jl}=\frac{1}{\lambda^2}(R^{-1})^{ak}_{il}g_{ja}, \quad g_{ai}\tilde{R}^{ak}_{jl}=\lambda^2 R^{ak}_{il}g_{ja}, \tag{71}$$

$$R'^{ak}_{jl}g_{ai}=(R'^{-1})^{ak}_{il}g_{ja}. \tag{72}$$

The above formulas also yield the explicit expression for the matrix $\hat{R}$ used in the construction of the inverse operator (67):

$$\hat{R}^{ij}_{kl}=\frac{1}{\lambda^2}g^{ja}g_{lb}(R^{-1})^{ib}_{ka}. \tag{73}$$

The dilatation operator does not commute with the d'Alembert operator, but it belongs to the set of symmetry operators due to the relation

$$\Box \mathbf{s}=\lambda^2 \mathbf{s}\Box. \tag{74}$$

Let us notice that we have modified the special conformal generators so as to obtain the symmetry operators $\mathbf{K}$ without imposing any condition on rotation operators $\mathbf{l}^{\pm}$.

To derive the conservation law we construct the Γ_μ operator

$$\Gamma_\mu(\mathbf{p},\overleftarrow{\mathbf{p}}^\dagger)=\overleftarrow{p}^{\dagger j}g_{j\mu}+g_{\mu j}p^j, \tag{75}$$

where

$$p^{\dagger j}=-p^k\zeta^{-\,j}_k.$$

The modified operator $\hat{\Gamma}_\mu$ includes the ζ^- operator (67):

$$\hat{\Gamma}_\mu(\mathbf{p},\overleftarrow{\mathbf{p}}^\dagger)=\overleftarrow{p}^{\dagger k}\overleftarrow{\zeta}^{-\,j}_\mu g_{kj}+\overleftarrow{\zeta}^{-\,j}_\mu g_{jk}p^k. \tag{76}$$

Now we can write down the currents following Proposition 5.1:

$$J^\omega_\mu=\Phi'\hat{\Gamma}_\mu(\mathbf{p},\overleftarrow{\mathbf{p}}^\dagger)\,\delta^\omega\Phi, \tag{77}$$

with $\delta^\omega\in\{\mathbf{p},\mathbf{l}^+,\mathbf{l}^-,\mathbf{s},\mathbf{K}\}$.

They obey the conservation law:

$$\sum_\mu p^\mu J^\omega_\mu=0. \tag{78}$$

VII. FINAL REMARKS

Let us notice that noncommutative spaces were studied by Doplicher, Fredenhagen, Roberts[20,21] and Madore and Mourad in Ref. 22 in the context of quantization of Minkowski space–time.

In the paper we have considered the fundamental properties of free field-theoretic models on a class of noncommutative spaces including q-Minkowski space. We obtained the conservation laws for a wide class of linear equations on braided linear spaces. In this paper we have extended the results derived in Refs. 3–6 for discrete spaces and for quantum Minkowski spaces.

The next step after the construction of conserved currents is to investigate the integrals of motion. In classical field theory as well as for discrete spaces,[3,4] they are produced from the time component of conserved current by integrating over space-like dimensions. In the noncommutative case the rules of integration still need further development. For braided linear spaces the global integration was introduced in papers by Chryssomalakos[23,24] and Kempf and Majid.[25] In the mentioned construction the main feature of global integration is its invariance with respect to translations. Thus, integrals of the functions that have a proper asymptotic behavior yield the vanishing boundary terms. In the derivation of integrals of motion we should apply the global integral over space-like dimensions. To this aim we shall formulate the Fubini theorem in which we express the global integral over whole space as a sequence of the iterated integrals. Having constructed iterated integrals, we will be able to follow the classical method and use conserved currents in the derivation of an analog of the integral of motion for specific noncommutative spaces.

APPENDIX: THE UNIQUE SOLUTION OF THE OPERATOR EQUATION (27)

Proof of the Proposition 4.1: We denote the monomials of derivatives by formulas (30), (31). Due to the modified Leibnitz rule (19), we should consider the solution of the operator equation (27) in the form of the polynomial of order $N-1$ with arbitrary coefficients:

$$\Gamma_\mu(\partial,\overleftarrow{\partial}^\dagger)=a^0_\mu+\sum_{l=1}^{N-1}\sum_{k=0}^{l}\overline{[\mu_1,\ldots,\mu_k]}a^k_{\mu\mu_1\cdots\mu_l}[\mu_{k+1},\ldots,\mu_l], \tag{A1}$$

where the coefficients a^k can depend on coordinates of covectors $\vec{x}$.

The condition (27) applied to the above polynomial yields the equations for coefficients $a^k_{\mu\mu_1\cdots\mu_l}$:

$$\begin{aligned}\sum_\mu(-\overleftarrow{\partial}^{\dagger\,\mu}+\partial^\mu)\circ\Gamma_\mu(\partial,\overleftarrow{\partial}^\dagger)=&-\sum_{l=1}^{N-1}\sum_{k=0}^{l}\sum_\mu\overline{[\mu_1,\ldots,\mu_k,\mu]}a^k_{\mu\mu_1\cdots\mu_l}[\mu_{k+1},\ldots,\mu_l]\\&+\sum_{l=1}^{N-1}\sum_{k=0}^{l}\overline{[\mu_1,\ldots,\mu_k]}\sum_\mu(\zeta^\mu_\nu a^k_{\mu\mu_1\cdots\mu_l})[\nu,\mu_{k+1},\ldots,\mu_l]\\&+\sum_{l=1}^{N-1}\sum_{k=0}^{l}\overline{[\mu_1,\ldots,\mu_k]}\sum_\mu(\partial^\mu a^k_{\mu\mu_1\cdots\mu_l})[\mu_{k+1},\ldots,\mu_l]-\sum_\mu\overline{[\mu]}a^0_\mu\\&+\sum_\mu(\zeta^\mu_\nu a^0_\mu)[\nu]+\sum_\mu(\partial^\mu a^0_\mu)=\Lambda(\partial)-\Lambda(\overleftarrow{\partial}^\dagger).\end{aligned} \tag{A2}$$

The resulting set of equations for functions $a^k_{\mu\mu_1\cdots\mu_l}$ is of the form

$$a^0_\mu=\Lambda_\mu, \tag{A3}$$

$$\partial^\alpha a^0_{\alpha\mu}+\zeta^\nu_\mu a^0_\nu=\Lambda_\mu, \tag{A4}$$

$$\zeta^\alpha_\mu a^0_{\alpha\mu_1\ldots\mu_l}+\partial^\alpha a^0_{\alpha\mu\mu_1\ldots\mu_l}=\Lambda_{\mu\mu_1\cdots\mu_l}, \tag{A5}$$

$$-a^k_{\mu\mu_1\cdots\mu_l}+\zeta^\alpha_{\mu_{k+1}}a^{k+1}_{\alpha\mu_1\cdots\mu_k\mu\mu_{k+2}\cdots\mu_l}+\partial^\alpha a^{k+1}_{\alpha\mu_1\cdots\mu_k\mu\mu_{k+1}\cdots\mu_l}=0, \tag{A6}$$

$$a^l_{\mu\mu_1\cdots\mu_l}+\partial^\alpha a^l_{\alpha\mu_1\cdots\mu_l\mu}=\Lambda_{\mu_1\cdots\mu_l\mu}, \tag{A7}$$

with $l=1,\ldots,N-1$, $k=0,\ldots,l-1$.

We begin to solve this set of equations by deriving the coefficients $a^0_{\mu\mu_1\cdots\mu_l}$ from Eqs. (3)–(5). Namely, for $l=N-1$ we have

$$\zeta_\mu^\alpha a^0_{\alpha\mu_1\cdots\mu_{N-1}} = \Lambda_{\mu\mu_1\cdots\mu_{N-1}}. \tag{A8}$$

Applying the inverse operator ζ^- and using the property of coefficients (24), we obtain:

$$a^0_{\mu\mu_1\cdots\mu_{N-1}} = \Lambda_{\mu\mu_1\cdots\mu_{N-1}}. \tag{A9}$$

We insert this solution into (A5) for $l=N-2$ and solve the next equation:

$$\zeta_\mu^\alpha a^0_{\alpha\mu_1\cdots\mu_{N-2}} + \partial^\alpha \Lambda_{\alpha\mu\mu_1\cdots\mu_{N-2}} = \Lambda_{\mu\mu_1\cdots\mu_{N-2}}. \tag{A10}$$

By assumption (23) after using ζ^- operator and (24), we derive $a^0_{\mu\mu_1\cdots\mu_{N-2}}$ as

$$a^0_{\mu\mu_1\cdots\mu_{N-2}} = \Lambda_{\mu\mu_1\cdots\mu_{N-2}}. \tag{A11}$$

Passing to the next equation from the subset (A5) and solving them in the similar way, we obtain the unique solution for coefficients a^0:

$$a^0_\mu = \Lambda_\mu, \quad a^0_{\mu\mu_1\cdots\mu_l} = \Lambda_{\mu\mu_1\cdots\mu_l}, \quad l=1,...,N-1. \tag{A12}$$

This R'-symmetric solution for initial coefficients allows us to evaluate the remaining ones using (A6), (A7), namely, we obtain the a^1 coefficients after writing the subset (A6) for $k=0$ and solving it the way we solved the subset (A3)–(A5) for a^0. The result is unique and looks as follows:

$$a^1_{\mu\mu_1\mu_2\cdots\mu_l} = \Lambda_{\mu_1\mu\mu_2\cdots\mu_l}, \quad l=1,...,N-1. \tag{A13}$$

The same method applied to subsets of (A6), (A7) for $k=1,...,N-2$ produces the unique solution of the set of equations for coefficients in the form

$$a^k_{\mu\mu_1\cdots\mu_l} = \Lambda_{\mu_1\cdots\mu_k\mu\mu_{k+1}\cdots\mu_l}, \quad l=1,...,N-1. \tag{A14}$$

The derivation of the explicit formulas for a unique solution of the coefficients of the operator Γ_μ concludes the proof of Proposition 4.1.

Let us notice once more that in the derivation of coefficients for the Γ_μ operator the crucial factors were the properties of coefficients of the equation (21), which enabled us to solve the equation (27) in the explicit form.

[1] Y. Takahashi, *An Introduction to Field Quantization* (Pergamon, Oxford, 1969).

[2] M. Klimek, ''Extension of q-deformed analysis and q-deformed models of classical mechanics,'' J. Phys. A **26**, 955–967 (1993).

[3] M. Klimek, ''The conservation laws for deformed classical models,'' Czech. J. Phys. **44**, 1049–1057 (1994).

[4] M. Klimek, ''The conservation laws and integrals of motion for a certain class of equations in discrete models,'' J. Phys. A **29**, 1747–1758 (1996).

[5] M. Klimek, ''Conservation laws for linear equations of motion on quantum Minkowski spaces,'' Commun. Math. Phys. **192**, 29–45 (1998).

[6] M. Klimek, ''The symmetry algebra and conserved currents for Klein-Gordon equation on quantum Minkowski space,'' in *Quantum Groups and Quantum Spaces*, Banach Center Publications, edited by R. Budzyński, W. Pusz, and S. Zakrzewski, Warszawa: Institute of Math., Polish Acad. Sci., Warszawa 1997, Vol. 40.

[7] P. Podleś and S. L. Woronowicz, ''On the structure of inhomogeneous quantum groups,'' Commun. Math. Phys. **185**, 325–358 (1997).

[8] P. Podleś and S. L. Woronowicz, ''On the classification of quantum Poincaré groups,'' Commun. Math. Phys. **178**, 61–82 (1996).

[9] P. Podleś and S. L. Woronowicz, ''Quantum deformation of Lorentz group,'' Commun. Math. Phys. **130**, 381–431 (1990).

[10] P. Podleś, ''Solutions of Klein–Gordon and Dirac equations on quantum Minkowski spaces,'' Commun. Math. Phys. **181**, 569–585 (1996).

[11] S. Majid, ''Introduction to braided geometry and q-Minkowski space'' in *Quantum Groups and their Applications*,

Proceedings of the International School of Physics ''Enrico Fermi'' Varenna, edited by L. Castellani and J. Wess, 1994.
[12] U. Meyer, ''Wave equations on q-Minkowski space,'' Commun. Math. Phys. **174**, 457–475 (1996).
[13] M. Pillin, ''q-deformed relativistic wave equations,'' J. Math. Phys. **35**, 2804–2817 (1994).
[14] S. Majid, ''Free braided differential calculus, braided binomial theorem and the braided exponential map,'' J. Math. Phys. **34**, 4843–4856 (1993).
[15] S. Majid, ''Braided momentum in the q-Poincaré group,'' J. Math. Phys. **34**, 2045–2058 (1993).
[16] M. Klimek, ''Integrals of motion for some equations on q-Minkowski space,'' Czech. J. Phys. **47**, 1199–1206 (1997).
[17] J. A. Azcárraga, P. P. Kulish, and F. Rodenas, ''On the physical contents of q-deformed Minkowski spaces,'' Phys. Lett. B **351**, 123–130 (1995).
[18] J. A. Azcárraga, P. P. Kulish, and F. Rodenas, ''Quantum groups and deformed special relativity,'' Fortschr. Phys. **44**, 1–40 (1996).
[19] S. Majid, ''Braided geometry of the conformal algebra,'' J. Math. Phys. **37**, 6495–6509 (1996).
[20] S. Doplicher, K. Fredenhagen, and J. E. Roberts, ''The quantum structure of spacetime at the Planck scale and quantum fields,'' Commun. Math. Phys. **172**, 187–220 (1995).
[21] S. Doplicher, K. Fredenhagen, and J. E. Roberts, ''Spacetime quantization induced by classical gravity,'' Phys. Lett. B **331**, 39–44 (1994).
[22] J. Madore and J. Mourad, ''Quantum space–time and classical gravity,'' J. Math. Phys. **39**, 423–442 (1998).
[23] Ch. Chryssomalakos, ''Remarks on quantum integration,'' Commun. Math. Phys. **184**, 1–25 (1997).
[24] Ch. Chryssomalakos, ''Applications of quantum groups,'' Ph.D. thesis, UC Berkeley, 1994.
[25] A. Kempf and S. Majid, ''Algebraic q-integration and Fourier theory on quantum and braided spaces,'' J. Math. Phys. **35**, 6802–6836 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Comment on ''The complete Schwarzschild interior and exterior solution in the harmonic coordinate system'' [J. Math. Phys. 39, 6086 (1998)]

László Á. Gergely
Laboratoire de Physique Théorique, Université Louis Pasteur, 3-5 rue de l'Université 67084 Strasbourg Cedex, France[a)]
and KFKI Research Institute for Particle and Nuclear Physics, Budapest 114, P.O. Box 49, H-1525 Hungary

(Received 18 November 1998; accepted for publication 17 March 1999)

[S0022-2488(99)01408-5]

In a recent paper Liu[1] considered the complete Schwarzschild interior and exterior solution in harmonic coordinates. There he argued about the necessity to keep the integration constant C_1 in R_{ex}, in contrast with previous treatments (Refs. 1–5 and 8 of Ref. 1). The purpose of this comment is to show that the above conclusion cannot be traced from the matching conditions between the vacuum exterior and the uniform density interior perfect fluid, as claimed in Ref. 1. The reason for this is that the last condition in Eqs. (7) of Ref. 1, namely $R'_{in}(a) = R'_{ex}(a)$, is *not* required by the junction conditions at $r=a$, as will be shown in what follows.

The junction of two space–times along a timelike hypersurface Σ can be done applying the Darmois–Israel matching procedure,[2,3] which requires the continuity across the junction of both the first and second fundamental forms (induced metric and extrinsic curvature) *of the junction hypersurface.*

In the standard coordinates, the metric (1) of Ref. 1 induces the three-metric given by the line element

$$ds^2_\Sigma = E(r)dt^2 - r^2(d\theta^2 + \sin^2\ \theta d\phi^2) \tag{1}$$

on the junction hypersurface $r=a$, which has the normal $n=1/\sqrt{G}\partial/\partial r$ and the nonvanishing extrinsic curvature components

$$K_{00} = -\frac{E'}{2\sqrt{G}}, \quad K_{22} = \frac{r}{\sqrt{G}}, \quad K_{33} = \frac{r}{\sqrt{G}} \sin^2\theta. \tag{2}$$

The continuity of both (1) and (2) implies that the metric function G is continuous and E is $\mathcal{C}^1$ across the junction. It is not to be expected that starting from the metrics written in other coordinate systems the conditions on E and G will be weakened. However, constraints on the new function R introduced by the coordinate transformation (2)–(5) of Ref. 1 will also emerge.

In the harmonic coordinate system the normal vector to Σ has the components $n^\mu = (\ln R)'/\sqrt{G}(0,X_1,X_2,X_3)$. The metric (6) of Ref. 1 (with a missing square on the last bracket corrected) induces the first fundamental form

$$ds^2_\Sigma = E(r)dt^2 - \frac{r^2}{R^2(r)}\, d\mathbf{X}^2. \tag{3}$$

This is still expressed in terms of the four space–time coordinates. When written in terms of the coordinates t, θ, and ϕ, intrinsic to Σ, the continuity of the induced metric again implies the

[a)]Visiting position, supported by the Hungarian State Eötvös Fellowship.

continuity of the metric function E alone. The extrinsic curvature tensor in the new coordinate system is found either by direct computation or by transforming its components (2) from standard to harmonic coordinates. The nonvanishing components are K_{00} given in (2) and

$$K_{ii}=\frac{(X_j^2+X_k^2)r}{R^4\sqrt{G}}\,, \quad K_{ij}=-\frac{X_iX_jr}{R^4\sqrt{G}}\,, \tag{4}$$

where $i\neq j\neq k$ take the values 1,2,3. The junction condition on the extrinsic curvature at arbitrary radius implies that G, R, and E' should be continuous. Altogether we find that G and R are $\mathcal{C}^0$ and E is $\mathcal{C}^1$. Thus the last relation in (7) of Ref. 1 does not hold.

The condition

$$E'_{in}=E'_{ex}\,, \tag{5}$$

though not listed among the continuity conditions (7) of Ref. 1, was fulfilled when imposing that the pressure vanishes on the junction. Indeed, Eq. (5) is a substitute for the requirement that the radial pressures on the two sides of Σ are equal, which was demonstrated for generic spherically symmetric static space–times in another context.[4]

Our criticism does not affect the main result of Ref. 1, which is the solution of R_{in} of the second-order differential equation (22) of Ref. 1. The arguments about the integration constants, however, should be reviewed. Requiring only the continuity of R and nothing more, one of the constants C_1 and C_2 can be freely specified, in particular $C_1=0$ can be chosen, in accordance with Refs. 1–5 and 8 of Ref. 1. Of course, the continuity of R' across the junction of the interior and exterior Schwarzschild solutions can be imposed as an additional requirement for other purposes [e.g., for having a smooth function $R(r)$ as in Ref. 1], but it is not a consequence of the junction conditions.

[1] Q. H. Liu, J. Math. Phys. **39**, 6086 (1998).
[2] G. Darmois, in *Mémorial des Sciences Mathématiques* (Gauthier-Villars, Paris, 1927), Fascicule 25, Chap. V.
[3] W. Israel, Nuovo Cimento B **XLIV**, 4349 (1966).
[4] L. Á. Gergely, Phys. Rev. D **58**, 084030 (1998).

Response to ''Comment on 'The complete Schwarzschild interior and exterior solution in the harmonic coordinate system' '' [J. Math. Phys. 40, 4177 (1999)]

Quan-Hui Liu
Department of Physics, Hunan University, Changsha 410082, People's Republic of China and Institute of Theoretical Physics, Academia Sinica, P.O. Box 2735, Beijing 100080, People's Republic of China[a)]

(Received 10 March 1999; accepted for publication 17 March 1999)

[S0022-2488(99)01508-X]

The preceding comment[1] applied the Darmois–Israel matching condition to the Schwarzschild interior and exterior solution in the harmonic coordinate system, and correctly pointed out that from this matching condition, one of the constants C_1 and C_2 in Eqs. (13) and (23) of my paper[2] can be freely specified, in particular $C_1=0$ can be chosen.

If only imposing the Darmois-Israel matching condition on the complete harmonic coordinates,[2] there are in fact infinite choices. I agree with Weinberg[3] that the particular choice $C_1=0$ is a ''convenient'' one. But we cannot use only this choice while keeping silent on the other possibilities. I used one half of the Darmois–Israel matching condition, the extrinsic-curvature and an *ad hoc* condition, the metric components, are continuous. It is equivalent to impose $R'_{in}(a)=R'_{ex}(a)$ on the Darmois–Israel matching condition to uniquely determine C_1 and C_2. We are confident that it is a good and reasonable choice. Furthermore, even C_1 and C_2 can be freely specified; we must carefully specify them. For example, a bad choice $C_2=0$ is presented in the Darmois–Israel matching condition. For our purpose of treating a star of $r=a$ with $a=9M/4\neq 0$,[2] this choice means that all the interior coordinates vanish, the metric components are singular, the proper time undefinable, and the whole interior gravitational field appears singular.

[a)]Mailing address. Electronic mail: liuqh@itp.ac.cn

[1] L. Á. Gergely, J. Math. Phys. **40**, 4177 (1999).
[2] Q-H. Liu, J. Math. Phys. **39**, 6086 (1998).
[3] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972), p. 181.

Erratum: ''Hirota bilinear approach to a new integrable differential-difference system'' [J. Math. Phys. 40, 2001 (1999)]

Xing-Biao Hu
State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific Engineering Computing, Academia Sinica, P.O. Box 2719, Beijing 100080, China

Yong-Tang Wu
Department of Computing Studies, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

Xian-Guo Geng
Department of Mathematics, Zhengzhou University, Henan 450052, China

(Received 13 April 1999; accepted for publication 13 April 1999)

[S0022-2488(99)02308-7]

There are some mistakes in Eqs. (11), (12), (15) and the formula (A4). The correct ones should be

$$u_t(n+1)+u_t(n-1)+u_t(n)+\tfrac{3}{4}(u(n+1)-u(n-1))^2+\tfrac{1}{4}(u(n+1)+u(n-1)-2u(n))^2 +\tfrac{1}{4}(v(n)-1)=0, \tag{11}$$

$$[(3D_t^2e^{D_n}+3D_t^2-e^{D_n}+1)f(n)\cdot f(n)]f(n)^2-[(1-e^{D_n})f(n)\cdot f(n)](D_t^2 f(n)\cdot f(n)) +D_t^2[(e^{D_n}-1)f(n)\cdot f(n)]\cdot f^2(n)=0, \tag{12}$$

$$u_t(n)+(T_++T_-+1)^{-1}[\tfrac{3}{4}(u(n+1)-u(n-1))^2+\tfrac{1}{4}(u(n+1)+u(n-1)-2u(n))^2 +\tfrac{1}{4}(v(n)-1)]=0, \tag{15}$$

$$(D_t^3D_z a\cdot a)a^2-(D_zD_t a\cdot a)(D_t^2 a\cdot a)=D_t^2(D_zD_t a\cdot a)\cdot a^2. \tag{A4}$$

In this case, (10) and (11) can be reduced to the equation found in Ref. 1. However, we can derive the following z flow from (13) and (14) by viewing t as an auxiliary variable:

$$U_{zz}(n)+U_z(n)(U_z(n-1)-U_z(n+1))=4e^{U(n+2)-U(n-1)}-4e^{U(n+1)-U(n-2)},$$

where $U(n)=\ln(f(n+1)/f(n))$. A Lax pair for it could be easily obtained from the bilinear Bäcklund transformation (16)–(18):

$$4\lambda^2e^{U(n+3)-U(n)}\psi_{n+3}+\lambda U_z(n+1)\psi_{n+2}-\psi_{n+1}+(\omega+2\lambda^{-1}\gamma^3)\psi_n=0,$$

$$\psi_{nz}+U_z(n)\psi_n+\lambda^{-1}\psi_{n-1}+\mu\psi_n=0,$$

where we have set $U(n)=\ln(g(n+1)/g(n))$, $f(n)=\psi_n g(n+1)$.

The mistakes that appeared in (11), (12), (15), and (A4) do not affect all the other deductions and calculations in the paper.

[1] X. B. Hu and Y. T. Wu, Phys. Lett. A **246**, 523 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 8 AUGUST 1999

Erratum: ''Operator formulation of Wigner's *R*-matrix theories for the Schrödinger and Dirac equations'' [J. Math. Phys. 39, 5231 (1998)]

Radosław Szmytkowski
Technical University of Gdańsk, Atomic Physics Division, ul. Narutowicza 11/12, 80952 Gdańsk, Poland

(Received 12 March 1999; accepted for publication 12 March 1999)

[S0022-2488(99)00708-2]

Equations (141) and (157) should read

$$\hat{\mathscr{A}}_{\mathfrak{b}}^{(\pm)} = (\gamma^{(\mp)})^2 \alpha_n^{(\pm)} \hat{\mathfrak{b}}^{(\mp)} \hat{\mathscr{A}}_{\mathfrak{b}}^{(\mp)} \hat{\mathfrak{b}}^{(\mp)} \alpha_n^{(\mp)}, \tag{141}$$

$$\mathsf{R}_{\mathfrak{b}}^{\pm}(E) = c\hbar\gamma^{(\pm)} \sum_k \frac{\mathsf{P}_{\mathfrak{b}k}^{(\pm)} \mathsf{P}_{\mathfrak{b}k}^{(\pm)\dagger}}{E_{\mathfrak{b}k} - E} - \frac{\mathsf{b}^{(\pm)}}{(\gamma^{(\mp)})^2 \mathsf{I} + (\mathsf{b}^{(\pm)})^2}. \tag{157}$$

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Point and line boundaries in scalar Casimir theory

Alfred Actor[a)]
Department of Physics, The Pennsylvania State University, Fogelsville, Pennsylvania 18051

(Received 6 February 1998; accepted for publication 16 April 1999)

A simple image-charge construction enables one to insert point or line boundaries (or planar or hyperplanar boundaries when sufficiently many spatial dimensions are available) at or into the central point, line, plane etc., of a great range of spatial backgrounds in quantum field theory which have appropriate symmetry. This nontrivial construction (which provides among other things the exact vacuum stress tensor $T_{\mu\nu}$ of the quantum field if $T_{\mu\nu}$ can be computed for the original background prior to point, line,..., insertion) works if all directions x_i perpendicular to the inserted object are symmetric under $x_i \rightarrow -x_i$. In other respects the symmetric spatial background can be quite arbitrary. While the inserted object experiences (by symmetry) no net Casimir force from the background, it does exert Casimir forces throughout this background which were originally not present. In addition to general theory, detailed examples are given (which include exact field $T_{\mu\nu}$'s and exact Casimir force densities) for arbitrary spatial dimension. First: point and line boundaries in otherwise empty space; then a planar boundary with a semi-infinite line extending from one side; finally, parallel planar boundaries with a point boundary halfway between them. Only scalar quantum fields are analyzed here; however the extension to the electromagnetic Casimir effect is discussed qualitatively. [S0022-2488(99)03807-4]

I. INTRODUCTION

A simple image-charge construction enables one to insert a point boundary or a line boundary into existing appropriately symmetric spatial backgrounds in quantum field theory (QFT). This construction involves reflection through the point or line to be occupied by the inserted boundary and the required background symmetry is thereby determined: The spatial background must be invariant under $x_i \rightarrow -x_i$ in all directions x_i perpendicular to the inserted object. (All spatial directions are perpendicular to a point.) Otherwise the spatial background can be quite arbitrary. Image-charge constructions are, of course, very familiar for planar boundary surfaces. The method here is an extension, of what one routinely does for planar surfaces, to reflections in two and more dimensions. Hyperplanes can similarly be inserted into appropriate backgrounds when sufficiently many dimensions are available. However, our attention here will be limited to point and line boundaries in scalar QFT.

By symmetry the inserted object (point, line,...) experiences no *net* Casimir force from the symmetric background into which it has been inserted. This balanced situation keeps the mathematics manageable and enables our simple but general construction to work. (In a nonsymmetric background the inserted object will experience a net Casimir force and one is confronted with a far more difficult problem.) The inserted object will, of course, exert (symmetrically in directions perpendicular to itself) Casimir forces throughout the original background which were nonexistent prior to its insertion. These additional Casimir forces may be complicated. Nonetheless, as we shall see, if one can calculate the vacuum stress tensor $T_{\mu\nu}$ for the original background, then one can also calculate $T_{\mu\nu}$ for the substantially more complicated system with inserted point or line boundary. From $T_{\mu\nu}$ all Casimir forces can straightforwardly be obtained, including the new forces exerted by the point or line boundary.

[a)]Electronic mail: aaa2@psu.edu

The image-charge construction presented here at a far more general level was first noticed[1] for the symmetric background called empty space. Otherwise these point and line boundary constructions can (to the author's knowledge) be found nowhere in the literature. Because this method provides quantitative access to many Casimir systems that could not previously be solved exactly, its description in some detail should be of interest to other workers in this area of QFT.

A secondary topic in this article will be the use of point and line boundary insertions into a scalar quantum field already distorted by an existing background to test the following hypotheses: (1) Distinct Dirichlet boundary (or background) objects mutually *attract*. (2) Distinct Neumann boundary objects mutually *attract*. (3) Distinct Dirichlet and Neumann objects mutually *repel*. This behavior has been observed for a variety of boundary geometries (in Refs. 2 and 3 and related unpublished calculations[4]). A typical Dirichlet (Neumann)—hereafter we use $D(N)$—object is understood to be a spatial surface on which the scalar quantum field $\hat{\phi}$ must satisfy $D(N)$ boundary conditions. The characterization ''distinct'' used above is important and means the object is, in a geometrical sense, uniquely defined. For comparison, in this context a cylindrical or spherical surface should *not* be imagined as consisting of two or more (obviously quite nonunique) pieces glued together. However, a plane with a line projecting from one side can reasonably be thought of as consisting of two distinct pieces, even though these pieces touch.

As will be seen, in the explicit examples considered in this paper rules (1)–(3) above hold unambiguously. This gives us additional confidence in the robustness of these rules and their (qualitative) independence of the geometric details of the boundaries involved.

The author has a preference for D over N boundary conditions in scalar QFT. One reason is that the notion of a D object can be generalized in an apparently deep way by introducing[3,5] a potential $V=V(\vec{x},t)\geqslant 0$ into the wave equation $\Box\hat{\phi}=0$ for a massless scalar field;

$$[\Box+V(\vec{x},t)]\hat{\phi}=0. \tag{1.1}$$

This brings great flexibility into the subtle process of coupling a quantum field to background structures which (partially or completely) expel it from those regions of space they occupy. Such a natural and compelling generalization of N boundary conditions seems to be impossible.

Another point to mention concerning N boundary conditions is an occasional sensitivity of the Casimir forces associated with them to one's choice of stress tensor. Scalar fields do not have a unique stress tensor. Commonly employed are the canonical ($T_{\mu\nu}$) and improved ($\theta_{\mu\nu}$) tensors. In the past (see Ref. 2 and unpublished work) one has noticed the existence of (a few) boundary geometries for which $T_{\mu\nu}$ and $\theta_{\mu\nu}$ yield different Casimir force densities when N conditions are involved. However, this seems never to happen when only D conditions are involved. This line of inquiry will be continued here for backgrounds with point and line boundaries. Again we find no evidence for ambiguous Casimir forces in pure Dirichlet systems, but some indication of such ambiguity when one or more boundary objects are Neumann.

The contents of this paper are now described. Section II presents the construction of isolated point and line boundaries in free space (of arbitrary dimension d) for a massless scalar field. Nonzero field mass $M>0$ does not introduce any particular problems and we choose $M=0$ merely for simplicity. Reference 1 discussed this same problem for $M\geqslant 0$ but only in terms of heat kernels. Here we present a more complete discussion beginning with the field modes and ending with the vacuum stress tensor of a scalar field distorted by an isolated point or line boundary in d-dimensional (and otherwise empty) space. Periodic arrays of these objects are also easily constructed and this is briefly discussed.

In Sec. III we show how in d dimensions to insert a line boundary perpendicular to an existing arbitrary one-dimensional (1D) background—say a 1D potential $V=V(x_d)$ with the line boundary parallel to the x_d axis. This is illustrated by a simple example: an infinite planar boundary with a semi-infinite line boundary extending from one side. The $T_{\mu\nu}$ and $\theta_{\mu\nu}$ of the distorted field are calculated and the nonuniform Casimir force density exerted by the projecting line on the plane is determined.

Section IV discusses and solves the problem of inserting point and line boundaries into the symmetry midplane of a symmetric (but otherwise arbitrary) 1D background—say a 1D symmetric potential $V=f(x_1^2)$ with midplane $x_1=0$. This is illustrated by a fairly complicated example: parallel planar boundaries with a point boundary in the midplane. $T_{\mu\nu}$ and $\theta_{\mu\nu}$ for the distorted quantum field between the planes are calculated, and from these the nonuniform Casimir force densities exerted by the point boundary on the planes are found.

Section V outlines the general point and line boundary construction procedure for an arbitrary background potential $V=V(\vec{x})$ having the required symmetry. Interesting examples here would include the insertion of a point boundary at the center of a spherical or rectangular cavity, or a line boundary through the center of these cavities, or a line boundary along the axis of a circular or rectangular waveguide, and so on. All of these calculations and many others are straightforward by our method, but for lack of space no additional examples can be presented here.

Further discussion will be found in Sec. VI, including comments on the relevance of these scalar field considerations to the electromagnetic case.

II. POINT AND LINE BOUNDARIES IN FREE SPACE

To illustrate the construction procedure for point and line boundaries consider what can be done with the ordinary plane-wave modes of QFT in d spatial dimensions.

Free space:

$$\phi(\vec{x}|\vec{k})_{0d}=(2\pi)^{-d/2}e^{i\vec{k}\cdot\vec{x}},$$

$$\text{all } \vec{x}, \quad -\infty<k_{1,2,\ldots,d}<\infty. \tag{2.1}$$

Planar boundary at $x_1=0$: $\eta_{D,N}=\mp 1$,

$$\phi(\vec{x}|\vec{k})_{1,d-1}\equiv(2\pi)^{-d/2}[e^{ik_1x_1}+\eta e^{-ik_1x_1}]e^{i(k_2x_2+\cdots+k_dx_d)},$$

$$k_1\geqslant 0, \quad -\infty<k_{2,\ldots,d}<\infty,$$

$$x_1\geqslant 0 \text{ or } \leqslant 0, \quad -\infty<x_{2,\ldots,d}<\infty,$$

$$D: \ \phi(x_1=0)_{1,d-1}=0,$$

$$N: \ \partial_1\phi(x_1=0)_{1,d-1}=0, \quad \forall\vec{k}. \tag{2.2}$$

Line boundary along the x_d *axis* $\vec{u}=(x_1,x_2,\ldots,x_{d-1})=0$: $\eta_{D,N}=\mp 1$,

$$\phi(\vec{x}|\vec{k})_{d-1,1}\equiv(2\pi)^{-d/2}[e^{i\vec{k}\cdot\vec{u}}+\eta e^{-i\vec{k}\cdot\vec{u}}]e^{ik_dx_d},$$

$$k_1\geqslant 0, \quad -\infty<k_{2,\ldots,d}<\infty,$$

$$x_1\geqslant 0 \text{ or } \leqslant 0, \quad -\infty<x_{2,\ldots,d}<\infty,$$

$$D: \ \phi(\vec{u}=0)_{d-1,1}=0, \quad \forall\vec{k},$$

$$N: \ \partial_i\phi(\vec{u}=0)_{d-1,1}=0, \quad i=1,2,\ldots,d-1. \tag{2.3}$$

Point boundary at $\vec{x}=0$: $\eta_{D,N}=\mp 1$,

$$\phi(\vec{x}|\vec{k})_{d0}\equiv(2\pi)^{-d/2}[e^{i\vec{k}\cdot\vec{x}}+\eta e^{-i\vec{k}\cdot\vec{x}}],$$

$$k_1\geqslant 0, \quad -\infty<k_{2,\ldots,d}<\infty,$$

$$x_1 \geqslant 0 \quad \text{or} \quad \leqslant 0, \quad -\infty < x_{2,...,d} < \infty,$$

$$D: \quad \phi(\vec{x}=0)=0, \quad \forall \vec{k}$$

$$N: \quad \partial_i \phi(\vec{x}=0)=0, \quad i=1,...,d. \tag{2.4}$$

In Eqs. (2.2)–(2.4) and throughout this paper the symbol η distinguishes Dirichlet (D) from Neumann (N) boundary conditions: $\eta_D=-1$ and $\eta_N=+1$. Moreover a label $pq=(0,d),(1,d-1),(2,d-2),\ldots,(d-1,1)$, $(d,0)$ is attached to ϕ_{pq} to emphasize the systematics. p represents the number of dimensions $x_1,x_2,\ldots,x_p$ in which an image-charge construction is used, and $q=d-p$ the remaining dimensions $x_{p+1},\ldots,x_d$ unused for this purpose. Between Eqs. (2.2) and (2.3) a number of boundary (hyper) objects in $d>3$ dimensions have been left out: specifically those with $p=2,...,d-2$. We could work with these objects exactly as we do with plane, line, and point boundaries. However, having no interest in these objects at this time we shall leave them entirely out of our discussion. Note that the D modes (2.2)–(2.4) all vanish on the objects on which they are supposed to vanish, while the N modes have vanishing normal derivatives on these same objects. (All directions are normal to a point.) Moreover the choice $\eta=0$ in Eqs. (2.2)–(2.4) eliminates the boundary objects altogether.

The complete orthonormal sets of modes (2.1)–(2.4) are delta-function normalized:

$$\int d^d k \, \phi(\vec{x}|\vec{k}) \overline{\phi}(\vec{y}|\vec{k}) = \delta(\vec{x}-\vec{y}), \tag{2.5}$$

where $k_1,\ldots,k_d$ have the ranges given in Eqs. (2.1)–(2.4) for $\phi=\phi_{pq}(\vec{x}|\vec{k})$. It is understood in Eq. (2.5) that $\vec{x}$ and $\vec{y}$ lie in the same spatial hemisphere, i.e., both $x_1,y_1\geqslant 0$ or both $x_1,y_1\leqslant 0$. Thus $\delta(x_1+y_1)$ vanishes and Eq. (2.5) results. Similarly,

$$\int d^d k \, \phi(\vec{x}|\vec{k}) \overline{\phi}(\vec{x}|\vec{k}') = \delta(\vec{k}-\vec{k}'), \tag{2.6}$$

where terms containing $\delta(\vec{k}+\vec{k}')$ vanish because $\delta(k_1+k_1')$ vanishes. Completeness and orthonormality are the reasons why $\vec{x}$ and $\vec{k}$ in each of Eqs. (2.2)–(2.4) are restricted to a hemisphere: $x_1\geqslant 0$ or $x_1\leqslant 0$ and $k_1\geqslant 0$.

Some standard information on spatial heat kernels is restated briefly in the Appendix. The definition (A1) becomes here

$$K(t|\vec{x},\vec{y})_{pq} \equiv \int d^d k \, e^{-t\vec{k}^2} \phi(\vec{x}|\vec{k})_{pq} \overline{\phi}(\vec{y}|\vec{k})_{pq}.$$

With the help of the formula

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} dk \, e^{-tk^2} e^{ik(x-y)} = \frac{1}{\sqrt{4\pi t}} e^{-(x-y)^2/4t}, \tag{2.7}$$

one easily verifies from modes (2.1) to (2.4) the spatial heat kernels listed below.

Free space:

$$K(t|\vec{x},\vec{y})_{0d} \equiv k_d(t|\vec{x}-\vec{y}) = (2\pi)^{-d}\int_{-\infty}^{\infty} d^d k \, e^{i\vec{k}\cdot(\vec{x}-\vec{y})} e^{-t\vec{k}^2} = (4\pi t)^{-d/2} e^{-(\vec{x}-\vec{y})^2/4t}. \tag{2.8}$$

Planar boundary at $x_1=0$:

$$K(t|\vec{x},\vec{y})_{1,d-1} \equiv [k_1(t|x_1-y_1) + \eta k_1(t|x_1+y_1)] k_{d-1}(t|x_2-y_2,\ldots,x_d-y_d). \tag{2.9}$$

Line boundary along the x_d axis:

$$K(t|\vec{x},\vec{y})_{d-1,1}\equiv\{k_{d-1}(t|x_1-y_1,\ldots,x_{d-1}-y_{d-1}) + \eta k_{d-1}(t|x_1+y_1,\ldots,x_{d-1}+y_{d-1})\}k_1(t|x_d-y_d). \tag{2.10}$$

Point boundary at $\vec{x}=0$:

$$K(t|\vec{x},\vec{y})_{d0}\equiv k_d(t|\vec{x}-\vec{y})+\eta k_d(t|\vec{x}+\vec{y}). \tag{2.11}$$

Again, in these formulas $\eta=-1$ for D and $\eta=+1$ for N boundary conditions. Also k_n always means the free-space heat kernel (2.8) for n spatial dimensions. One verifies that k_1 satisfies the free 1D heat equation

$$(-\partial_x^2)k_1(t|x\pm y)=(-\partial_y^2)k_1(t|x\pm y)=\partial_t k_1(t|x\pm y). \tag{2.12}$$

Each of the heat kernels (2.8)–(2.11) therefore satisfies the free-space heat equation

$$(-\Delta_x)K_{pq}=(-\Delta_y)K_{pq}=\partial_t K_{pq}. \tag{2.13}$$

Moreover the K_{pq} all fulfill the boundary conditions they are supposed to satisfy. Thus they are the correct heat kernels for these systems.

The heat kernels (2.9)–(2.11) were deduced in Ref. 1 without any discussion of the modes ϕ_{pq}. The derivation here is more complete. In Ref. 1 it was observed that the heat kernels for infinite arrays of equally spaced points or lines can also be obtained quite simply. For example, in mode language the isolated-point modes (2.4) can be replaced by

$$\phi(\vec{x}|\vec{k})_{d0}\equiv(2\pi)^{-d/2}\sum_{b=-\infty}^{\infty}[e^{i\vec{k}\cdot(\vec{x}+b\vec{B})}+\eta e^{i\vec{k}\cdot(-\vec{x}+b\vec{B})}], \tag{2.14}$$

where $\vec{B}$ is any constant vector. Here $\phi(\vec{x}=m\vec{B}|\vec{k})_{d0}^D=0$ for any integer m; also $\partial_i\phi(\vec{x}=m\vec{B}|\vec{k})_{d0}^N=0$ for $i=1,\ldots,d$ and any integer m. Thus the $D(N)$ modes (2.14) represent infinite arrays of Dirichlet (Neumann) points at $\vec{x}=m\vec{B}$. More complicated arrays of point boundaries can be similarly introduced.[1] The heat kernel obtained from modes (2.14) and Eq. (2.7) is

$$K(t|\vec{x},\vec{y})_{d0}=(4\pi t)^{-d/2}\sum_b \{e^{-(\vec{x}-\vec{y}+b\vec{B})^2/4t}+\eta e^{-(\vec{x}+\vec{y}+b\vec{B})^2/4t}\}.$$

Actually one obtains here a double sum $\Sigma_{b,b'}$ rather than Σ_b and $(b-b')\vec{B}$ in the exponents rather than $b\vec{B}$. But the sum over b' is totally redundant and can simply be discarded, this step amounting to nothing more than a constant (infinite) renormalization of K_{d0} which has no effect on boundary conditions, the heat equation, or other essential properties of K_{d0}.

The same device with slight modification in Eqs. (2.3) and (2.10) leads to explicit modes and heat kernels for infinite arrays of line boundaries with equal spacing. Due to a lack of space no further discussion of arrays of point and line boundaries will be given in this article.

A. Vacuum stress tensor for an isolated point boundary

It is helpful to know the vacuum stress tensors of isolated point and line boundaries as these are, after all, prototypical boundary objects. Let us consider the point boundary at $\vec{x}=0$ in d dimensions, using some formulas collected in the Appendix for the canonical and improved stress tensors $T_{\mu\nu}$ and $\theta_{\mu\nu}$. The calculations are fairly similar to ones presented in Ref. 2. Ultraviolet regularization is performed by discarding the free-space term $k_d(t|\vec{x}-\vec{y})$ in Eq. (2.11). From the remaining (purely boundary) term

$$\eta K(t|\vec{x},\vec{y})_{\text{point}}\equiv\eta k_d(t|\vec{x}+\vec{y}) \tag{2.15}$$

in K_{d0} we obtain using Eq. (A3) and the convenient formula

$$(r^2)^s\Gamma(-s)=\int_0^\infty dt\, t^{s-1}e^{-r^2/t}, \tag{2.16}$$

the regularized mode sum

$$\sum_n (w_n^2)^{-s}\phi_n(\vec{x})\overline{\phi_n}(\vec{y})_{\text{reg}}=\frac{1}{\Gamma(s)}\int_0^\infty dt\, t^{s-1}\eta(4\pi t)^{-d/2}e^{-(\vec{x}+\vec{y})^2/4t}$$
$$=\eta\frac{\Gamma(-s+d/2)}{\Gamma(s)(4\pi)^{d/2}}\left[\frac{1}{4}(\vec{x}+\vec{y})^2\right]^{s-d/2}. \tag{2.17}$$

Henceforth we drop the label ''reg'' on the mode sum. From Eq. (2.17) one verifies

$$\sum_n \frac{1}{\omega_n}\partial_i\phi_n(\vec{x})\partial_j\overline{\phi_n}(\vec{x})=-\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}r)^{d+1}}\left[\delta_{ij}-(d+1)\frac{x_ix_j}{r^2}\right],\quad r=|\vec{x}|.$$

Here the right-hand side is unchanged when $\partial_i\phi_n(\vec{x})\partial_j\overline{\phi_n}(\vec{x})$ on the left is replaced by $(\partial_i\partial_j\phi_n(\vec{x}))\overline{\phi_n}(\vec{x})$ or by $\phi_n(\vec{x})\partial_i\partial_j\overline{\phi_n}(\vec{x})$. Then using mode-sum formulas for $T_{\mu\nu},\theta_{\mu\nu}$ in the Appendix we find

$$T_{00}=0,\quad \theta_{00}=-\eta\frac{1}{2d}(d-1)\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}r)^{d+1}}; \tag{2.18}$$

$$T_{ii}=d\,\theta_{ii}=-\eta\frac{1}{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}r)^{d+1}}\left[2-(d+1)\frac{x_i^2}{r^2}\right]; \tag{2.19}$$

$$T_{ij}=d\,\theta_{ij}=\eta\frac{\Gamma\left(\frac{d+3}{2}\right)}{(\sqrt{4\pi}r)^{d+1}}\frac{x_ix_j}{r^2};\quad i\neq j \tag{2.20}$$

with $T_{0i}=\theta_{0i}=0$. $\theta_{\mu\nu}$ is traceless as it should be. Both $T_{\mu\nu}$ and $\theta_{\mu\nu}$ are conserved as they should be: $\partial_iT_{ij}=0$.

The results (2.18)–(2.20) of course display the physical boundary divergences (see, e.g., Refs. 6–8) known to be inseparable from boundaries (unless these divergences cancel away). Here $T_{00}=0$ merely because of cancellation. $T_{00}=0$ does *not* mean the point boundary has no effect on the vacuum energy distribution in its vicinity. There is no unique stress tensor for a scalar field, and θ_{00} does not vanish. Dimensionally $T_{\mu\nu}$ and $\theta_{\mu\nu}$ must be constructed from $1/r^{d+1}$ and x_ix_j/r^{d+3} making the presence of these $r\to 0$ divergences in Eqs. (2.18)–(2.20) inevitable. Physically, boundary divergences represent the extreme distortion of the quantum field by the point boundary away from the spatial uniformity it would possess in the absence of this boundary object.

Before leaving the point boundary we mention the connection between the exact boundary term (2.15) in the heat kernel and the heat kernel expansion for boundary problems (see, e.g., Ref. 9)

$$K(t|\vec{x},\vec{x}) \sim \sum_{n\geqslant 0} (4\pi t)^{(n-d)/2}[a_n(\vec{x})+b_n(\vec{x})], \quad t\to 0+$$

where $a_0=1$, $b_0=0$. Here the $a_n(\vec{x})$ are *functions* of $\vec{x}$ characterizing bulk properties of the distorted quantum field. The $b_n(\vec{x})$ are *distributions* defined only on the boundary surfaces, each distribution characterizing the field distortion near that particular boundary surface (but less well than the exact heat kernel does this). From the exact heat kernel

$$K(t|\vec{x},\vec{x})_{d0} = (4\pi t)^{-d/2}[1+\eta e^{-\vec{x}^2/t}]$$

and the limit

$$\lim_{t\to 0+} \frac{1}{\sqrt{\pi t}} e^{-x^2/t} = \delta(x),$$

we see that for the isolated point boundary $a_0=1$ (as it must) while

$$b_d(\vec{x}) = \eta 2^{-d}\delta(x_1)\,\delta(x_2)\cdots\delta(x_d).$$

All other a_n, b_n vanish.

B. Vacuum stress tensor for an isolated line boundary

The heat kernel (2.10) for the infinite line boundary along the x_d axis is UV regularized by discarding the free-space term. What remains is

$$\eta K(t|\vec{x},\vec{y})_{\text{line}} \equiv \eta k_{d-1}(t|\vec{u}+\vec{v})k_1(t|x_d-y_d),$$
$$\vec{u}=(x_1,\ldots,x_{d-1}), \quad \vec{v}=(y_1,\ldots,y_{d-1}), \tag{2.21}$$

leading to the regularized mode sum for the isolated line boundary

$$\sum_n (\omega_n^2)^{-s}\phi_n(\vec{x})\overline{\phi}_n(\vec{y}) = \eta\frac{\Gamma(-s+d/2)}{\Gamma(s)(4\pi)^{d/2}} 2^{d-2s}[(\vec{u}+\vec{v})^2+(x_d-y_d)^2]^{s-d/2}. \tag{2.22}$$

Thus

$$\sum_n \frac{1}{\omega_n}\partial_i\phi_n(\vec{x})\partial_j\overline{\phi}_n(\vec{x}) = -\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}u)^{d+1}}\left[\delta_{ij}-(d+1)\frac{x_ix_j}{u^2}\right], \quad i,j\neq d,$$
$$u^2=x_1^2+x_2^2+\cdots+x_{d-1}^2, \tag{2.23}$$

$$\sum_n \frac{1}{\omega_n}\partial_d\phi_n(\vec{x})\partial_d\overline{\phi}_n(\vec{x}) = -\sum_n \omega_n|\phi_n|^2 = \eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}u)^{d+1}}, \tag{2.24}$$

$$\sum_n \frac{1}{\omega_n}\partial_i\phi_n(\vec{x})\partial_j\overline{\phi}_n(\vec{x}) = 0 \quad \text{for } i=d, \quad j\neq d \text{ or } i\neq d, \quad j=d. \tag{2.25}$$

The right-hand sides of Eqs. (2.23) and (2.25) are unchanged when on the left $\partial_i\phi_n\partial_j\overline{\phi_n}$ is replaced by $(\partial_i\partial_j\phi_n)\overline{\phi_n}$ or by $\phi_n(\partial_i\partial_j\overline{\phi_n})$. The right-hand side of Eq. (2.24) changes sign when $\partial_d\phi_n\partial_d\overline{\phi_n}$ is replaced by $(\partial_d\partial_d\phi_n)\overline{\phi_n}$ or by $\phi_n(\partial_d\partial_d\overline{\phi_n})$. One then verifies the vacuum stress tensors

$$T_{00}=-T_{dd}=\eta\frac{1}{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}u)^{d+1}}, \tag{2.26}$$

$$\theta_{00}=-\theta_{dd}=\eta\frac{2-d}{2d}\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}u)^{d+1}}, \tag{2.27}$$

$$T_{ii}=d\theta_{ii}=-\eta\frac{1}{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}u)^{d+1}}\left[3-(d+1)\frac{x_i^2}{u^2}\right],\quad i<d, \tag{2.28}$$

$$T_{ij}=d\theta_{ij}=\eta\frac{\Gamma\left(\frac{d+3}{2}\right)}{(\sqrt{4\pi}u)^{d+1}}\frac{x_ix_j}{u^2},\quad i\neq j,\quad i,j<d, \tag{2.29}$$

$$T_{id}=\theta_{id}=0,\quad i<d, \tag{2.30}$$

with $T_{0\mu}=\theta_{0\mu}=0$ for $\mu=1,2,...,d$. Here $\theta_{\mu\nu}$ is traceless as it should be and both $T_{\mu\nu}$ and $\theta_{\mu\nu}$ are conserved as they should be.

From the exact heat kernel (2.10) for the isolated line boundary

$$K(t|\vec{x},\vec{x})_{d-1,1}=(4\pi t)^{-d/2}[1+\eta e^{-\vec{u}^2/t}]\sim(4\pi t)^{-d/2}+\eta\frac{1}{2^d\sqrt{\pi t}}\delta(x_1)\cdots\delta(x_{d-1}),$$

$$t\to 0+$$

we can read off the coefficients a_n, b_n in the heat-kernel expansion for this system.

III. LINE BOUNDARY PERPENDICULAR TO EXISTING ONE-DIMENSIONAL BACKGROUND

Our line boundary construction works when, in all spatial directions perpendicular to the line, space is symmetric. If the line boundary is to be positioned along the x_d axis then we need symmetry under $x_i\to -x_i$ for $i=1,...,d-1$. There is no need for spatial symmetry along x_d. In fact we can insert the line boundary into an arbitrary background potential $V(x_d)$.

The modes for a line boundary parallel to x_d in an arbitrary 1D background potential $V=V(x_d)$ are

$$\phi(\vec{x}|\vec{q},n)=(2\pi)^{(1-d)/2}[e^{i\vec{q}\cdot\vec{u}}+\eta e^{-i\vec{q}\vec{u}}]\phi_n(x_d),$$

$$\vec{u}=(x_1,\ldots,x_{d-1}),\quad \vec{q}=(k_1,\ldots,k_{d-1}),$$

$$k_1\geqslant 0,\quad -\infty<k_i<\infty,$$

$$x_1\geqslant 0 \text{ or } \leqslant 0,\quad -\infty<x_i<\infty,\quad i=2,...,d-1, \tag{3.1}$$

where $\phi_n(x_d)$ satisfies

$$[-\partial_d^2+V(x_d)]\phi_n(x_d)=k_n^2\phi_n(x_d). \tag{3.2}$$

The set $\{\phi_n(x_d)\}$ is complete in x_d. With momentum allowed to have the range indicated the modes (3.1) are orthonormal and complete. The heat kernel is

$$K(t|\vec{x},\vec{y})_{d-1,1}=K_1(t|x_d,y_d)\{k_{d-1}(t|\vec{u}-\vec{v})+\eta k_{d-1}(t|\vec{u}+\vec{v})\}, \tag{3.3}$$

$$K_1(t|x_d,y_d)\equiv\sum_n e^{-tk_n^2}\phi_n(x_d)\overline{\phi_n}(y_d), \tag{3.4}$$

where $\vec{v}=(y_1,\ldots,y_{d-1})$ and k_{d-1} is the free-space heat kernel. If the factor (3.4) is known one can compute $T_{\mu\nu},\theta_{\mu\nu}$ directly from Eq. (A3).

As a simple but nontrivial example of this construction let us attach a perpendicular semi-infinite line boundary to one side of an infinite plane. The plane is positioned at $x_d=0$ and the line lies along the positive x_d axis. Our goal is to calculate $T_{\mu\nu}$ and $\theta_{\mu\nu}$ in the half space $x_d>0$. These tensors are known (see, e.g., Ref. 2) in $x_d<0$ where there are no boundary objects. From $T_{\mu\nu},\theta_{\mu\nu}$ we can find the Casimir force density on the plane exerted by the projecting semi-infinite line boundary.

The x_d mode factor $\phi_n(x_d)$ in Eq. (3.1) is, in $x_d\geqslant0$,

$$\phi_d(x_d)=\frac{1}{\sqrt{2\pi}}[e^{ik_dx_d}+\gamma e^{-ik_dx_d}] \tag{3.5}$$

where $\gamma=-1(+1)$ for a Dirichlet (Neumann) plane at $x_d=0$. The heat kernel in $x_d,y_d\geqslant0$ is

$$K(t|\vec{x},\vec{y})_{d-1,1}=[k_1(t|x_d-y_d)+\gamma k_1(t|x_d+y_d)]\{k_{d-1}(t|\vec{u}-\vec{v})+\eta k_{d-1}(t|\vec{u}+\vec{v})\}. \tag{3.6}$$

Discarding the free-space term we are left with

$$K(t|\vec{x},\vec{y})_{d-1,d}=\gamma K_{\text{plane}}+\eta K_{\text{line}}+\eta\gamma K_{\text{point}}, \tag{3.7}$$

where K_{point} and K_{line} are defined by Eqs. (2.15) and (2.21) and

$$\gamma K(t|\vec{x},\vec{y})_{\text{plane}}\equiv\gamma k_1(t|x_d+y_d)k_{d-1}(t|\vec{u}-\vec{v}) \tag{3.8}$$

is the regularized heat kernel for an isolated plane positioned at $x_d=0$. The vacuum stress tensors computed from Eq. (3.7) obviously have the form

$$T_{\mu\nu}=T^{\gamma}_{\mu\nu\ \text{plane}}+T^{\eta}_{\mu\nu\ \text{line}}+T^{\eta\gamma}_{\mu\nu\ \text{point}} \tag{3.9}$$

with a similar formula for $\theta_{\mu\nu}$. Each term in Eq. (3.9) is separately conserved as are the separate terms in $\theta_{\mu\nu}$. The latter are separately traceless as well.

Perhaps the most interesting quantity to look at is the perpendicular Casimir force/area

$$F(x_1,\ldots,x_{d-1})/A\equiv\lim_{\epsilon\to0}[T_{dd}(x_d=-\epsilon)-T_{dd}(x_d=\epsilon)] \tag{3.10}$$

on the boundary plane at $x_d=0$. Because $T_{dd\ \text{plane}}=0$ for any $x_d\neq0$,[2] this plane's distortion of the quantum field does not contribute to the Casimir force acting on either side of it (quite appropriately). This leaves the Casimir force acting on the plane due to the field distortion in $x_d>0$ caused by the semi-infinite line, and by the pointlike juncture of this line with the plane. As Eq. (3.9) displays very explicitly this juncture acts as an independent boundary object in addition to the plane and the line. We find

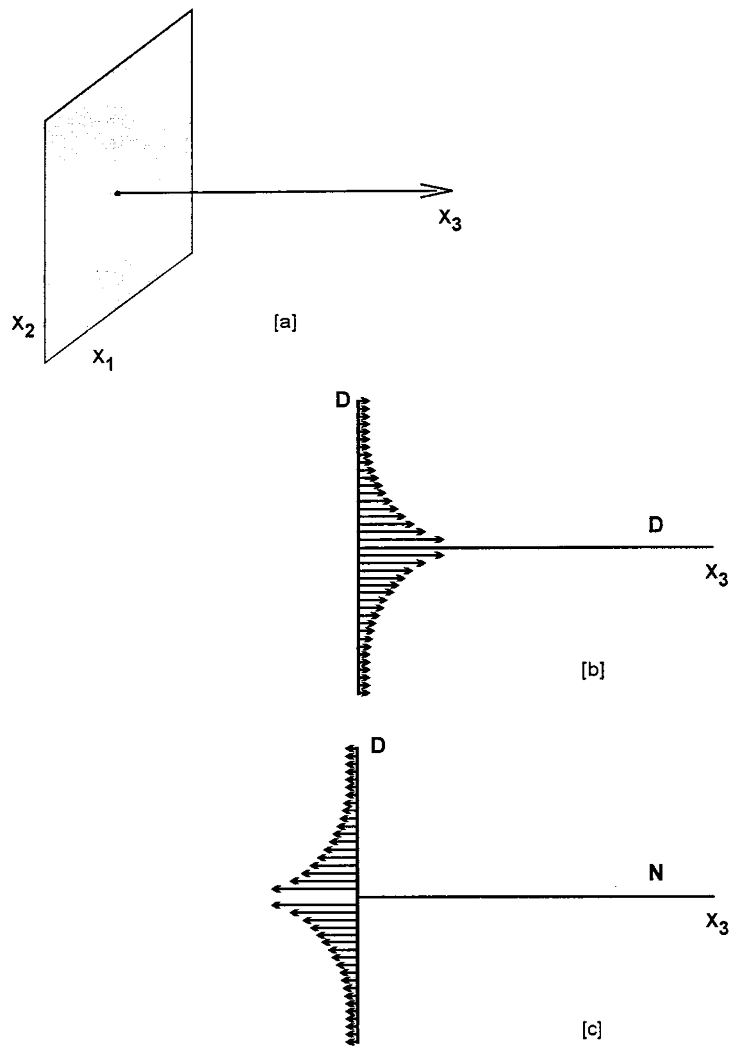

FIG. 1. The example in Sec. III for $d=3$ spatial dimensions. (a) A semi-infinite line boundary (along the positive x_3 axis) extends at right angles from one side of an infinite planar boundary (occupying the x_1x_2 plane). (b), (c) Side view: The semi-infinite $D(N)$ line attracts (repels) the D plane according to Eq. (3.11) (force density not to scale).

$$F(x_1,\ldots,x_{d-1})/A=\eta\frac{1}{2}(1+2\gamma)\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}u)^{d+1}},\tag{3.11}$$

where $u^2=x_1^2+\cdots+x_{d-1}^2$. Thus $T_{\mu\nu}$ predicts that a D line and a D plane attract, as do a N line and a N plane. However, a D line repels a N plane, and a N line repels a D plane. See Fig. 1.

Using Eq. (3.10) with $T_{dd}\rightarrow\theta_{dd}$ we obtain the Casimir force/area on the plane at $x_d=0$ predicted by the improved vacuum stress tensor,

$$F'(x_1,\ldots,x_{d-1})/A=\eta\frac{1}{2d}(2-d+2\gamma)\frac{\Gamma\left(\frac{d+1}{2}\right)}{(\sqrt{4\pi}u)^{d+1}}.\tag{3.12}$$

Using the symbol $F'_{\text{plane,line}}$ we see here that F'_{NN} is (attractive, 0, repulsive) for ($d<4$, $d=4$, $d>4$) and F'_{ND} is (repulsive, 0, attractive) for ($d<4$, $d=4$, $d>4$). The force densities F_{NN} and F_{ND} disagree with this. On the other hand, $F'_{DD}=F_{DD}$ and $F'_{DN}=F_{DN}$ for all spatial dimensions $d\geqslant 2$.

IV. POINT OR LINE BOUNDARY IN THE CENTRAL PLANE OF A SYMMETRIC ONE-DIMENSIONAL BACKGROUND

We now consider a background in d-dimensional space represented by an arbitrary potential of the form $V=f(x_1^2)$. The directions $x_2,\ldots,x_d$ are translation invariant and free. The non-negative potential function $f(x_1^2)\geqslant 0$ can be otherwise quite arbitrary, but it is symmetric under $x_1\to -x_1$. This symmetry enables one to easily insert a planar,..., line or point boundary into the central plane $x_1=0$. We show how to calculate $T_{\mu\nu},\theta_{\mu\nu}$ for the resulting system from the heat kernel for the original background. Then as a detailed example we perform these calculations for a point boundary midway between two parallel planes. This system and others like it have never been investigated to our knowledge. The exact stress-tensor results agree with our qualitative expectations concerning Casimir forces acting between distinct D and N objects.

A. Arbitrary background

The modes for the original spatial background *prior* to inserting the point or line boundary into the plane $x_1=0$ have the form

$$\phi(\vec{x}|n,\vec{q})=\phi_n(x_1)(2\pi)^{(1-d)/2}e^{i\vec{q}\cdot\vec{u}},$$

$$\vec{u}=(x_2,\ldots,x_d),\quad \vec{q}=(k_2,\ldots,k_d),\tag{4.1}$$

$$-\infty<k_i<\infty,\quad i=2,...,d$$

$$[-\partial_1^2+f(x_1^2)]\phi_n(x_1)=k_n^2\phi_n(x_1).\tag{4.2}$$

The corresponding background heat kernel is

$$K(t|\vec{x},\vec{y})=K_1(t|x_1,y_1)k_{d-1}(t|\vec{u}-\vec{v}),\tag{4.3}$$

$$K_1(t|x_1,y_1)\equiv\sum_n e^{-tk_n^2}\phi_n(x_1)\overline{\phi_n}(y_1),\tag{4.4}$$

where $\vec{v}=(y_2,\ldots,y_d)$ and k_{d-1} is the free-space heat kernel. The x_1 heat kernel (4.4) satisfies

$$[-\partial_1^2+f(x_1^2)]K_1(t|x_1,y_1)=\partial_t K_1(t|x_1,y_1)\tag{4.5}$$

and similarly in y_1. Because of the symmetry of the operator $-\partial_1^2+f(x_1^2)$ under $x_1\to -x_1$ it follows that $\phi_n(-x_1)$ is also a solution of Eq. (4.2), and $K_1(t|-x_1,y_1)$ is a solution of Eq. (4.5). Obvious properties of the free-space factor k_{d-1} in Eq. (4.3) then lead to the conclusion that both $K(t|\pm\vec{x},\vec{y})$ satisfy the full heat equation $[-\Delta_x+V]\ K=\partial_t K$ for this system, and similarly in $\vec{y}$.

1. Point boundary at $\vec{x}=0$

To insert a point boundary at $\vec{x}=0$ we define the new heat kernel

$$K(t|\vec{x},\vec{y})_{d0}\equiv K(t|\vec{x},\vec{y})+\eta K(t|\mp\vec{x},\pm\vec{y})\tag{4.6}$$

with $\eta_D=-1$, $\eta_N=+1$ as usual. This heat kernel satisfies the full heat equation as well as D or N boundary conditions at the point $x_1,\ldots,x_d=0$. Thus it represents the modified system with a point boundary at $\vec{x}=0$ in addition to the 1D background structure described by $V=f(x_1^2)$. The (orthonormal, complete) modes for the same system are

$$\phi(\vec{x}|n,\vec{q})_{d0}=(2\pi)^{(1-d)/2}[\phi_n(x_1)e^{i\vec{q}\cdot\vec{u}}+\eta\phi_n(-x_1)e^{-i\vec{q}\cdot\vec{u}}] \tag{4.7}$$

with now $k_2\geqslant 0$, $x_2\geqslant 0$ or $\leqslant 0$. K_{d0} can be obtained directly from these modes.

Using Eq. (A3) we can now compute $T_{\mu\nu},\theta_{\mu\nu}$ for the modified system if $K_1(t|x_1,y_1)$ for the original system without point boundary is known. Even if K_1 is unknown, Eq. (4.6) reveals a very fundamental property of Casimir forces: Inserted D and N point boundaries exert equal but opposite Casimir forces throughout the symmetric background into which they have been inserted. This is evident from the structure of Eq. (4.6). Consider the canonical vacuum stress tensor computed from Eq. (4.6),

$$T_{\mu\nu}=T_{\mu\nu\ \text{background}}+\eta T_{\mu\nu\ \text{point}}, \tag{4.8}$$

where the notation is hopefully obvious. $T_{\mu\nu\ \text{background}}$ is computed from $K(t|\vec{x},\vec{y})$ and $T_{\mu\nu\ \text{point}}$ from $K(t|-\vec{x},\vec{y})$. These tensors are separately conserved. Clearly the contribution from the point boundary to any Casimir force/area acting on a boundary surface, or to any Casimir force density acting on diffuse (Dirichlet-like) background structure represented by $V=f(x_1^2)$, will be equal but opposite for D,N points simply because $\eta_D=-1$ and $\eta_N=+1$. Exactly the same statements can be made about the improved vacuum stress tensor $\theta_{\mu\nu}$.

Another comment on Eq. (4.8) concerns UV renormalization. The free-space part of $T_{\mu\nu}$ will be found in $T_{\mu\nu\ \text{background}}$ and only this part of the tensor (4.8) needs UV renormalization. $T_{\mu\nu\ \text{point}}$ is UV finite by construction. This will be illustrated by the example to follow.

2. Line boundary along the x_d axis

To insert a line boundary into the symmetry plane $x_1=0$ (along the x_d axis) we define the heat kernel

$$K(t|\vec{x},\vec{y})_{d-1,1}=k_1(t|x_d,y_d)\{K_1(t|x_1,y_1)k_{d-2}(t|\vec{\alpha}-\vec{\beta})+\eta K_1(t|\mp x_1,\pm y_1)k_{d-2}(t|\vec{\alpha}+\vec{\beta})\},$$
$$\vec{\alpha}=(x_2,\ldots,x_{d-1}),\quad \vec{\beta}=(y_2,\ldots,y_{d-1}). \tag{4.9}$$

$K_{d-1,1}$ satisfies both the heat equation and D or N boundary conditions along the x_d axis. The corresponding modes are

$$\phi(\vec{x}|n,\vec{q})_{d-1,1}=(2\pi)^{(1-d)/2}e^{ik_dx_d}[\phi_n(x_1)e^{i\vec{q}\cdot\vec{\alpha}}+\eta\phi_n(-x_1)e^{-i\vec{q}\cdot\vec{\alpha}}]. \tag{4.10}$$

From $K_{d-1,1}$ we can calculate $T_{\mu\nu},\theta_{\mu\nu}$ for the new system with line boundary along the x_d axis if we know K_1 for the original background without this line. As for an inserted point it is obvious that D and N boundary lines exert equal but opposite Casimir forces throughout the original background.

B. Parallel-plane background

Let us now specialize to a familiar background: two static parallel planar boundaries at $x_1=\pm L/2$. The distortion of the quantum field between and outside these walls is such that each wall experiences an inward uniform force/area[10,2] (the same for D or N conditions)

$$F/A=\frac{d\Gamma\left(\frac{d+1}{2}\right)\zeta(d+1)}{(\sqrt{4\pi}L)^{d+1}}. \tag{4.11}$$

After point or line boundary insertion in the central plane $x_1=0$ the local Casimir force $F_1(x_2,\ldots,x_d)/A$ on the planes is nonuniform. Far from the point/line boundary F_1/A becomes of course the uniform force (4.11). However near the point/line boundary $F_1(x_2,\ldots,x_d)/A$ is significantly position dependent. As we shall see, if a $D(N)$ point is positioned between parallel D walls the attractive Casimir force/area between the latter is locally strengthened (weakened). If a $D(N)$ point is positioned between parallel N walls the attractive force/area between the latter is locally weakened (strengthened).

Note that we assume $d\geqslant 2$ henceforth, specifically *excluding* one spatial dimension $d=1$ from discussion. The obvious reason is that all boundaries are points in one dimension. Because one knows how to handle the static two-boundary problem for $d=1$, one can as easily deal with N static boundaries at arbitrary positions. There is no reason to apply the far more powerful but equally more specialized method of this paper.

The modes for the parallel-plane background are given by Eq. (4.1) with

$$\phi_n(x_1)=\frac{1}{\sqrt{2L}}[e^{in(\pi/L)(x_1+L/2)}+\gamma e^{-in(\pi/L)(x_1+L/2)}],$$

$$D:\quad \gamma=-1,\quad n=1,2,3,...,$$

$$N:\quad \gamma=1,\quad n=0,1,2,... . \tag{4.12}$$

All the D modes vanish at $x_1=\pm L/2$, and beyond these planes at $x_1=(p+1/2)L$ with p any integer where one might imagine other D planes if one so chooses. However we shall imagine empty space out beyond the planes at $x_1=\pm L/2$, so modes other than the ones (4.12) are appropriate in these regions. For the N modes $\partial_1\partial_n$ vanishes at $x_1=(p+1/2)L,\ p$ any integer.

The heat kernel for the parallel-plane background is (see, e.g., Ref. 2) given by Eq. (4.3) with

$$K_1(t|x_1,y_1)=(4\pi t)^{-1/2}\sum_{n=-\infty}^{\infty}[e^{-(n2L+x_1-y_1)^2/4t}+\gamma e^{-(n2L+L+x_1+y_1)^2/4t}]. \tag{4.13}$$

Here we discard a constant term for Neumann planes coming from the $n=0$ constant mode in Eq. (4.12), which seems to play no physical role. Note that $K_{1\mathrm{D}}$ with $\gamma=-1$ vanishes at $x_1=\pm L/2$ or $y_1=\pm L/2$. Also $\partial_1 K_{1N}$ vanishes at $x_1=\pm L/2$ or $y_1=\pm L/2$. Both of $K_1(t|\pm x_1,y_1)$ satisfy the 1D free heat equation in x_1, and similarly in y_1. Consequently $K(t|\pm\vec{x},\vec{y})$ satisfies the complete heat equation in $\vec{x}$, and likewise for $K(t|\vec{x},\pm\vec{y})$ in variable $\vec{y}$.

Inserting a point boundary between the planes at $\vec{x}=0$ by means of Eq. (4.6) we obtain a canonical vacuum stress tensor for the resulting system having the form (4.8);

$$T_{\mu\nu}=T_{\mu\nu\ \text{planes}}+\eta T_{\mu\nu\ \text{point}},\quad -L/2\leqslant x_1\leqslant L/2 \tag{4.14}$$

and similarly for $\theta_{\mu\nu}$. The field tensor $T_{\mu\nu\,\text{planes}}$ between parallel planes computed (with UV renormalization) from the background heat kernel (4.3) with factor (4.13) is well known (again see, e.g., Ref. 2). This tensor is conserved; therefore the remaining tensor $T_{\mu\nu\,\text{point}}$ describing the additional distortion of the quantum field by the inserted point boundary is separately conserved. $T_{\mu\nu\,\text{planes}}$ contains the only free-space contribution to $T_{\mu\nu}$ and only this part of $T_{\mu\nu}$ needs renormalization. The uniform force (4.11) is obtained from $T_{\mu\nu\,\text{planes}}$.

$T_{\mu\nu\,\text{point}}$ is computed without renormalization from

$$\eta K(t|-\vec{x},\vec{y})_{\text{point}}=\eta(4\pi t)^{-d/2}\sum_{n=-\infty}^{\infty}\{e^{-(n2L+x_1+y_1)^2/4t}+\gamma e^{-(n2L+L+x_1-y_1)^2/4t}\}e^{-(\vec{u}+\vec{v})^2/4t}. \tag{4.15}$$

From Eq. (A3) and this heat kernel we find

$$\eta\sum_n (\omega_n^2)^{-s}\phi_n(\vec{x})\overline{\phi_n}(\vec{y})_{\text{point}}=\eta\frac{\Gamma(-s+d/2)}{\Gamma(s)(4\pi)^{d/2}}2^{d-2s}\sum_{n=-\infty}^{\infty}\{[(n2L+x_1+y_1)^2+(\vec{u}+\vec{v})^2]^{s-d/2}+\gamma[(n2L+L+x_1-y_1)^2+(\vec{u}+\vec{v})^2]^{s-d/2}\}. \tag{4.16}$$

This basic mode sum yields all terms in $T_{\mu\nu\,\text{point}}$. UV renormalization is unnecessary because of the presence of $(\vec{u}+\vec{v})^2$ in every term on the right. Thus $[0]^{s-d/2}$ never occurs in the limit $\vec{x}\to\vec{y}$ (except at $\vec{x}=\vec{y}=0$, which produces the boundary divergences inseparable from the point boundary). In the corresponding mode sum for $T_{\mu\nu\,\text{planes}}$ one has $[(n2L+x_1-y_1)^2+(\vec{u}-\vec{v})^2]^{s-d/2}$ under the sum and the $n=0$ term needs renormalization. Because $[0]^{s-d/2}$ never occurs in Eq. (4.16) on the boundary planes $x_1=\pm L/2$, $T_{\mu\nu\,\text{point}}$ is finite on these planes. The boundary divergences in $T_{\mu\nu}$ inseparable from the boundary planes are found only in $T_{\mu\nu\,\text{planes}}$, just as the boundary divergences associated with the point boundary are found only in $T_{\mu\nu\,\text{point}}$. These comments can be repeated for the improved vacuum stress tensor $\theta_{\mu\nu}$.

Some helpful intermediate formulas in a moderately tedious calculation are

$$\eta\sum_n \frac{1}{\omega_n}\left\{\begin{matrix}|\partial_1\phi_n|^2\\ (\partial_1^2\phi_n)\overline{\phi_n}\ \text{ or }\ \phi_n(\partial_1^2\overline{\phi_n})\end{matrix}\right\}_{\text{point}}$$
$$=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\left\{[d(nL+x_1)^2-u^2][(nL+x_1)^2+u^2]^{-(d+3)/2}\mp\gamma\left[d\left(nL+\frac{L}{2}\right)^2-u^2\right]\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}\right\}, \tag{4.17}$$

$$\eta\sum_n \frac{1}{\omega_n}\left\{\begin{matrix}\partial_i\phi_n\partial_1\overline{\phi_n}\quad \text{or}\quad \phi_n\partial_i\partial_1\overline{\phi_n}\\ \partial_1\phi_n\partial_i\overline{\phi_n}\quad \text{or}\quad (\partial_1\partial_i\phi_n)\overline{\phi_n}\end{matrix}\right\}_{\text{point}}$$
$$=\eta\frac{\Gamma\left(\frac{d+3}{2}\right)}{(4\pi)^{(d+1)/2}}(2x_i)\sum_{n=-\infty}^{\infty}\left\{(nL+x_1)[(nL+x_1)^2+u^2]^{-(d+3)/2}\mp\gamma\left(nL+\frac{L}{2}\right)\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}\right\}\quad \text{for } i>1, \tag{4.18}$$

$$\eta\sum_n \frac{1}{\omega_n}\{\partial_i\phi_n\partial_j\overline{\phi_n}\ \text{ or }\ (\partial_i\partial_j\phi_n)\overline{\phi_n}\ \text{ or }\ \phi_n\partial_i\partial_j\overline{\phi_n}\}_{\text{point}}$$
$$=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\left\{-\delta_{ij}[(nL+x_1)^2+u^2]^{-(d+1)/2}-\gamma\delta_{ij}\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+1)/2}+(d+1)x_ix_j[(nL+x_1)^2+u^2]^{-(d+3)/2}+\gamma(d+1)x_ix_j\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}\right\}\quad i,j>1. \tag{4.19}$$

As a check on these expressions one can verify that each satisfies any D or N boundary condition on the planes it should satisfy. The identity (for $p=$integer)

$$\sum_{n=-\infty}^{\infty}\left(nL+\frac{L}{2}\right)^{2p+1}\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-s}=0 \tag{4.20}$$

is helpful. Another useful intermediate formula is

$$\eta\sum_n \frac{1}{\omega_n}|\vec{\nabla}\phi_n|^2_{\text{point}}=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\left\{[(nL+x_1)^2+u^2]^{-(d+1)/2}\right.$$
$$\left.+\gamma\left[(1-2d)\left(nL+\frac{L}{2}\right)^2+3u^2\right]\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}\right\}. \tag{4.21}$$

From the mode-sum formulas in the Appendix we now find relatively easily the components of $T_{\mu\nu\,\text{point}}$ in the region $-L/2\leqslant x_1\leqslant L/2$:

$$T_{00\ \text{point}}=\eta\gamma\frac{\Gamma\left(\frac{d+1}{2}\right)}{2(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\left[-d\left(nL+\frac{L}{2}\right)^2+u^2\right]\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}, \tag{4.22}$$

$$T_{11\ \text{point}}=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{2(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\{[(d-1)(nL+x_1)^2-2u^2]$$
$$\times[(nL+x_1)^2+u^2]^{-(d+3)/2}-\gamma[(nL+L/2)^2+u^2]^{-(d+1)/2}\}, \tag{4.23}$$

$$T_{ii\ \text{point}}=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{2(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\{[(nL+x_1)^2+u^2]^{-(d+3)/2}$$
$$\times[-2(nL+x_1)^2-2u^2+(d+1)x_i^2]+\gamma[(nL+L/2)^2+u^2]^{-(d+3)/2}$$
$$\times[(d-2)(nL+L/2)^2-3u^2+(d+1)x_i^2]\},\quad i>1,\ i\ \text{not summed}, \tag{4.24}$$

$$T_{1i\ \text{point}}=\eta x_i\frac{\Gamma\left(\frac{d+3}{2}\right)}{(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}(nL+x_1)[(nL+x_1)^2+u^2]^{-(d+3)/2},\quad i>1, \tag{4.25}$$

$$T_{ij\ \text{point}}=\eta\frac{\Gamma\left(\frac{d+3}{2}\right)}{(4\pi)^{(d+1)/2}}x_ix_j\sum_{n=-\infty}^{\infty}\left\{[(nL+x_1)^2+u^2]^{-(d+3)/2}+\gamma\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}\right\},$$
$$i\neq j\quad\text{and}\quad i,j>1, \tag{4.26}$$

$$T_{0i}=0,\quad i\geqslant 1. \tag{4.27}$$

In the limit $L\to\infty$ the tensor $T_{\mu\nu\,\text{point}}$ becomes the tensor (2.18)–(2.20) for an isolated point boundary at $\vec{x}=0$. $T_{\mu\nu\,\text{point}}$ is conserved as it should be.

The improved vacuum stress tensor $\theta_{\mu\nu\,\text{point}}$ is

$$\theta_{00\ \mathrm{point}}=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{2d(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\{[(nL+x_1)^2+u^2]^{-(d+1)/2}(1-d)$$

$$+\gamma[(nL+L/2)^2+u^2]^{-(d+3)/2}[(1-2d)(nL+L/2)^2-(d-2)u^2]\},\qquad(4.28)$$

$$\theta_{11\ \mathrm{point}}=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{2d(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\{[(nL+x_1)^2+u^2]^{-(d+3)/2}[(d-1)(nL+x_1)^2-2u^2]$$

$$+\gamma[(nL+L/2)^2+u^2]^{-(d+3)/2}[-(d^2-d+1)(nL+L/2)^2+(d-2)u^2]\},\qquad(4.29)$$

$$\theta_{ii\ \mathrm{point}}=\frac{1}{d}T_{ii\ \mathrm{point}},$$

$$i>1,\quad i\ \text{not summed},\qquad(4.30)$$

$$\theta_{i1\ \mathrm{point}}=\frac{1}{d}T_{i1\ \mathrm{point}},\quad i>1,\qquad(4.31)$$

$$\theta_{ij\ \mathrm{point}}=\frac{1}{d}T_{ij\ \mathrm{point}},$$

$$i\neq j\quad\text{and}\ i,j>1.\qquad(4.32)$$

$\theta_{\mu\nu\ \mathrm{point}}$ is conserved, and for $L\to\infty$ this tensor reduces to the isolated point boundary tensor given in Eqs. (2.18)–(2.20).

From the formula $f_i(\vec{x})=-\partial_j T_{ij}(\vec{x})$ for the local force density within the boundary-distorted quantum field we see that $f_1(\vec{x})=0$ (because $\partial^\mu T_{\mu\nu}=0$) everywhere away from the boundaries (i.e., in empty space). However $T_{\mu\nu}$ and $\theta_{\mu\nu}$ are discontinuous across the boundary planes and Casimir forces act on these planes. The Casimir force/area on the plane at $x_1=-L/2$ computed from $T_{\mu\nu}$ is

$$F_1(u)/A\equiv\lim_{\epsilon\to0}\left[T_{11}\left(x_1=-\frac{L}{2}-\epsilon\right)-T_{11}\left(x_1=-\frac{L}{2}+\epsilon\right)\right]$$

$$=d\,\frac{\Gamma\left(\frac{d+1}{2}\right)\zeta(d+1)}{(\sqrt{4\pi}L)^{d+1}}-T_{11\ \mathrm{point}}\left(x_1=-\frac{L}{2}\right)\qquad(4.33)$$

with

$$T_{11\ \mathrm{point}}(x_1=-L/2)=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{2(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}$$

$$\times\left\{(d-1-\gamma)\left(nL+\frac{L}{2}\right)^2-(2+\gamma)u^2\right\}.\qquad(4.34)$$

Here we remind the reader that $T_{11}=0$ outside the parallel planes, and the first term on the right in Eq. (4.33) (second equality), i.e., the uniform force (4.11) between parallel planes, comes from $T_{11\ \mathrm{planes}}$. The negative of Eq. (4.34) gives the additional nonuniform force/area on the left boundary plane due to the point boundary in the middle. At position $u=0$ on this plane ''beneath'' or nearest the point boundary the force/area (4.34) is strongest, and larger in magnitude than the first term in Eq. (4.33),

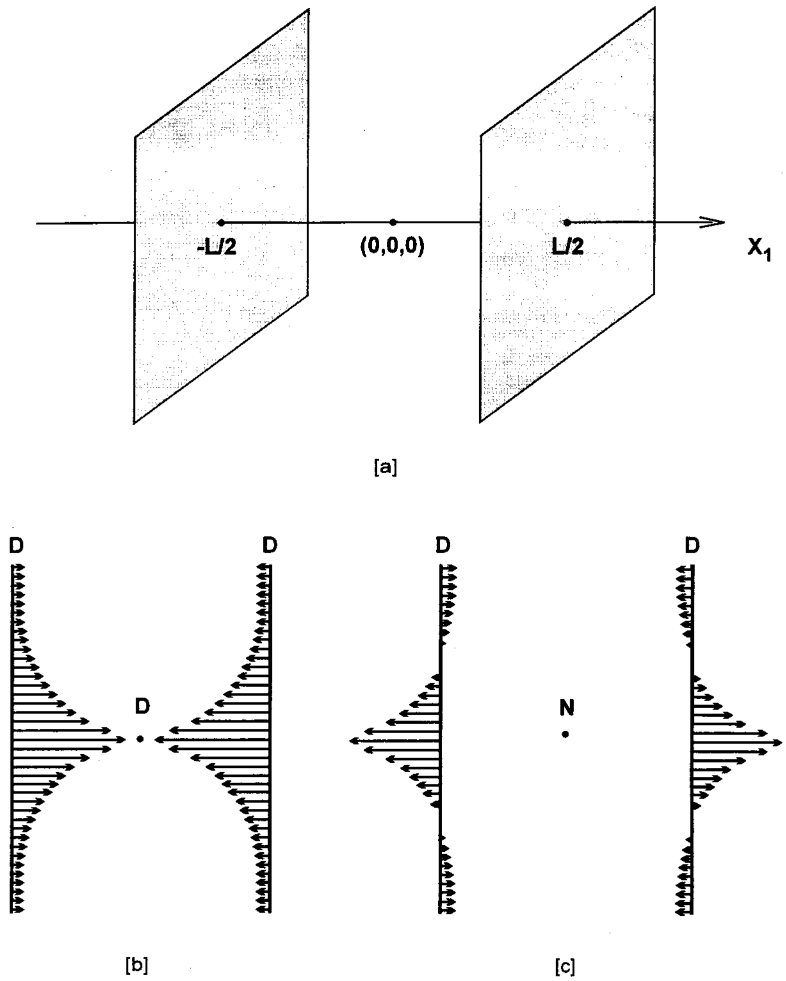

FIG. 2. The example in Sec. IV B for $d=3$ spatial dimensions. (a) A point boundary midway between parallel infinite boundary planes. (b), (c) Side view: A $D(N)$ point boundary attracts (repels) the parallel D planes enclosing it according to Eq. (4.33) (force density not to scale).

$$T_{11\ \mathrm{point}}(x_1=-L/2)|_{u=0}=\eta(d-1-\gamma)\frac{\Gamma\left(\frac{d+1}{2}\right)\zeta(d+1)}{(\sqrt{4\pi}L)^{d+1}}[2^{d+1}-1].$$

The identity

$$\sum_{n=-\infty}^{\infty}(n+1/2)^{-s}=2[2^s-1]\zeta(s)$$

is useful here. Figure 2 depicts the situation for parallel Dirichlet planes with a Dirichlet or Neumann point midway between them.

The Casimir force/area on the plane at $x_1=-L/2$ computed from $\theta_{\mu\nu}$ is given by

$$F_1'(u)/A=d\frac{\Gamma\left(\frac{d+1}{2}\right)\zeta(d+1)}{(\sqrt{4\pi}L)^{d+1}}-\theta_{11\ \mathrm{point}}\left(x_1=-\frac{L}{2}\right) \tag{4.35}$$

with

$$\theta_{11\ \text{point}}(x_1=-L/2)=\eta\frac{\Gamma\left(\frac{d+1}{2}\right)}{2(4\pi)^{(d+1)/2}}\sum_{n=-\infty}^{\infty}\left[\left(nL+\frac{L}{2}\right)^2+u^2\right]^{-(d+3)/2}$$
$$\times\left\{\frac{1}{d}[d-1-\gamma(d^2-d+1)]\left(nL+\frac{L}{2}\right)^2+\frac{1}{d}[-2+\gamma(d-2)]u^2\right\}. \tag{4.36}$$

The constant term in eq. (4.35)—again the uniform force (4.11)—comes from $\theta_{11\ \text{planes}}$. The additional nonuniform force/area on the left plane due to the point boundary at $\vec{x}=0$ is the negative of Eq. (4.36).

An important question is: For what backgrounds do $T_{\mu\nu}$ and $\theta_{\mu\nu}$ predict the same nonuniform force density on the boundary planes due to the point, and for which backgrounds are these force densities different? A comparison of Eqs. (4.34) and (4.36) shows that for $\gamma=-1$

$$T_{11\ \text{point}}\left(x_1=-\frac{L}{2}\right)=\theta_{11\ \text{point}}\left(x_1=-\frac{L}{2}\right),\quad \gamma=-1 \tag{4.37}$$

while for $\gamma=1$ these two functions of u^2 are somewhat different. Thus for parallel Dirichlet planes with a Dirichlet or Neumann point halfway between them, $T_{\mu\nu}$ and $\theta_{\mu\nu}$ predict identical Casimir force densities on the planes. For parallel Neumann planes the predicted force densities are not identical. The *total* force F_{point} on the $x_1=-L/2$ plane due to the point boundary is finite;

$$F_{\text{point}}\equiv\int_{-\infty}^{\infty}dx_2\cdots dx_d\left[-T_{11\ \text{point}}\left(x_1=-\frac{L}{2}\right)\right]=\eta\gamma\frac{\pi}{2^{d+2}L^2}.$$

Interestingly the force computed from $-\theta_{11\ \text{point}}(x_1=-L/2)$ is exactly the same. Note that F_{point} is an attractive inward force for $\eta\gamma=1$ (D planes and D point or N planes and N point) and a repulsive outward force for $\eta\gamma=-1$ (D planes and N point or N planes and D point).

V. GENERAL CASE

For completeness we now summarize the most general versions of the point boundary and line boundary constructions.

A. Point boundary insertion

If the spatial background is symmetric under $\vec{x}\to-\vec{x}$ [in particular if this background is described by a potential $V(\vec{x})$ with the property $V(-\vec{x})=V(\vec{x})$] then a point boundary can be inserted at $\vec{x}=0$. By now it should be quite obvious how. The heat kernel $K(t|\vec{x},\vec{y})$ for the original background satisfies by assumption

$$[-\Delta_x+V(\vec{x})]K(t|\pm\vec{x},\vec{y})=\partial_t K(t|\pm\vec{x},\vec{y}) \tag{5.1}$$

and similarly in $\vec{y}$. Moreover

$$K(t|-\vec{x},-\vec{y})=K(t|\vec{x},\vec{y}). \tag{5.2}$$

Thus the modified heat kernel

$$K(t|\vec{x},\vec{y})_{d0}\equiv K(t|\vec{x},\vec{y})+\eta K(t|\mp\vec{x},\pm\vec{y}) \tag{5.3}$$

satisfies the background heat equation (5.1). It also manifestly vanishes at $\vec{x}=0$ for $\eta_D=-1$, while $\vec{\nabla}K_{d0}$ vanishes at $\vec{x}=0$ for $\eta_N=+1$. Thus K_{d0} correctly represents the modified system of original background+point boundary at $\vec{x}=0$.

B. Line boundary insertion

If the spatial background is symmetric under $\vec{u}\rightarrow-\vec{u}$ where $\vec{u}=(x_1,x_2,\ldots,x_{d-1})$, then a line boundary can be inserted along the x_d axis. The background can be arranged in any fashion parallel to the x_d axis. The background heat kernel $K(t|\vec{u},x_d;\vec{v},y_d)$ has to satisfy

$$[-\Delta_x+V(\vec{u},x_d)]K(t|\pm\vec{u},x_d;\vec{v},y_d)=\partial_t K(t|\pm\vec{u},x_d;\vec{v},y_d) \tag{5.4}$$

and similarly in $\vec{y}=(\vec{v},y_d)$. The potential V satisfies $V(-\vec{u},x_d)=V(\vec{u},x_d)$ and consequently

$$K(t|-\vec{u},x_d;-\vec{v},y_d)=K(t|\vec{u},x_d;\vec{v},y_d). \tag{5.5}$$

Then the modified heat kernel

$$K(t|\vec{x},\vec{y})_{d-1,1}\equiv K(t|\vec{u},x_d;\vec{v},y_d)+\eta K(t|\mp\vec{u},x_d;\pm\vec{v},y_d) \tag{5.6}$$

satisfies the background heat equation (5.5) and moreover D or N boundary conditions along the x_d axis.

The heat kernels (5.3) and (5.6) obviously lead to vacuum stress tensors of the form

$$T_{\mu\nu}=T_{\mu\nu}^{\mathrm{BG}}+\eta\Delta T_{\mu\nu}. \tag{5.7}$$

Here $\eta\Delta T_{\mu\nu}$ is computed from the second term in Eqs. (5.3) and (5.6) and $T_{\mu\nu}^{\mathrm{BG}}$ from $K(t|\vec{x},\vec{y})$ for the original background. UV renormalization has to be performed on $T_{\mu\nu}^{\mathrm{BG}}$, not on $\Delta T_{\mu\nu}$. Point and line boundary insertion does not affect UV renormalization, and this is expressed by the UV-finite nature of $\Delta T_{\mu\nu}$. The additional Casimir force density $f_i(\vec{x})$ acting throughout the quantum field due to this field's additional distortion by the point or line boundary is given by

$$f_i(\vec{x})=-\eta\partial_j\Delta T_{ij}.$$

Equal but opposite force densities are exerted by D and N point or line boundaries.

VI. DISCUSSION

A. Quantum scalar field

In this article we have shown quite explicitly how to insert point and line boundaries into the central point, line, plane,..., of symmetric (but otherwise arbitrary) backgrounds in scalar QFT. The spatial background is required to be invariant under reflection through the inserted object along every direction x_i perpendicular to this object. Such a construction is familiar for planar boundaries but not for point and line boundaries. If one is able to compute the heat kernel $K(t|\vec{x},\vec{y})$ for the original background system—and therefore the vacuum stress tensor $T_{\mu\nu}$ and all Casimir force densities within this system—then one can also calculate $K,T_{\mu\nu}$, and all Casimir force densities for the substantially more complicated system with a point or line boundary at its center.

We used point and line boundary insertion to investigate the conjectured rules for scalar Casimir theory discussed in Sec. I: Distinct D objects attract; distinct N objects do the same; distinct D and N objects repel. These rules were found to be obeyed in the detailed examples of Secs. III and IV. Strong indications that these rules hold at a very general level could be discerned in the general mathematics.

Are global Casimir energy considerations at all useful for studying point and line boundaries? Let us return to the parallel planes (*DD* or *NN*) with a *D* or *N* point boundary between them. The four boundary configurations *DDD*,*DND*,*NDN*,*NNN* produce identical spectra in the quantum field. Consequently the (unrenormalized) global vacuum energy $E_{\text{vac}}=\Sigma\,\omega_n/2$ between the planes cannot distinguish among them. One soon realizes this is only to be expected. The point boundary between the planes causes a *finite* shift in the *infinite* (even after renormalization) vacuum energy between the infinite planes. (For finite planes one might wish to rethink this—see below.)

Another way to formulate the preceeding comment is as follows: Imagine displacing slowly the point boundary (keeping it in the midplane between the parallel planar boundaries) away to infinity. Done slowly this costs no energy because no field excitation occurs. Yet with the point boundary out of sight at infinity the system appears to be just two parallel planes distorting the quantum field everywhere. Evidently no change in vacuum energy is involved. Exactly the same argument applies to a line boundary inserted between parallel planes.

These remarks for the parallel-plane example have an obvious extension to more complicated backgrounds. If an inserted point or line boundary *is not* ''confined'' by its symmetric background in the sense that it *can* (at least in one direction) be displaced away to infinity with no expenditure of energy, then global vacuum energy considerations will be insensitive to the insertion of the point or line boundary.

On the other hand, if an inserted point or line boundary *is* confined by its symmetric background in the sense that it *cannot* be displaced to infinity without expenditure of energy, then global vacuum energy considerations should be helpful. Take for example a finite spherical cavity. The method of this paper enables one to insert a point boundary at the center of this cavity, or a line boundary diametrically through the center. Neither the point nor the line can be displaced without expending energy. The vacuum energy within the cavity is finite and therefore quite capable of detecting the insertion of either point or line. Very similar comments apply to finite cavities with other shapes.

The method of this paper can very straightforwardly be extended to scalar QFT at finite temperature. The notation the author would choose for this purpose is reviewed in Ref. 11.

B. Quantum electromagnetic field

What about fields other than scalar ones? Fermion boundary conditions on planes and other extended surfaces are not a clearly organized subject. Thus we leave these aside and mention gauge fields—especially the electromagnetic field in four space–time dimensions. Many Casimir calculations in electromagnetic theory have been done using Green functions. The latter are, of course, bilocal functions closely related to heat kernels and mode sums like (A3) with $s=1$. Thus the constructions presented in this paper can be applied to the electromagnetic field, much as one uses the image-charge constructions associated with planar metallic surfaces. This is, of course, important because the electromagnetic (EM) field is the one quantum field everyone believes in. The EM vacuum state is understood to pervade all of space, and the Casimir effect provides a wonderfully direct way to locally access and manipulate this vacuum state, and hence to further verify its existence. Naturally, when studying the static EM Casimir effect one chooses space to be three dimensional.

Technically it is clear that the methods of this paper straightforwardly modified to the EM problem enable one to insert metal point and line boundaries into preexisting appropriately symmetric metal and/or dielectric spatial backgrounds. We cannot undertake a technical discussion of this here; however we hope to present a relatively full account elsewhere. Still there are useful and general things which can be said about the EM problem.[4] The present article concludes with a brief account of these aspects.

A vacuum EM fluctuation can be any classical EM field, including ones which do not satisfy the vacuum Faraday's and Ampere's laws. [Of course, the further a given quantum fluctuation is from being a solution of these Maxwell equations, the more brief will be its existence, which reduces its influence.] Vacuum EM fluctuations of course have to satisfy charge conservation. In the presence of backgrounds one can think in terms of distinct electric and magnetic sectors, and discuss quantum fluctuations in these sectors separately as we shall do. It is known (although nowhere in the literature adequately discussed) that the electric and magnetic sectors in the EM Casimir effect *compete*, with electric quantum fluctuations giving rise to attractive Casimir forces and magnetic quantum fluctuations to repulsive Casimir forces between metal boundary objects. (See, e.g., Refs. 12 and 13 for early work suggesting this.) The seemingly unpredictable dependence of the static EM Casimir effect on boundary geometry[14–17] arises from this competition between electric and magnetic sectors.

There exists an interesting analogy between the scalar and EM Casimir effects. A scalar field interacting with a Dirichlet object is in some sense analogous to the electric vacuum sector interacting with a usual (i.e., electric charge) metal object. Just as distinct D objects have an attractive Casimir effect in scalar QFT, the electric vacuum sector always gives rise to attraction between distinct metal objects.[4,12,13]

Distinct N objects in scalar QFT also have an attractive Casimir effect. Such objects correspond to ''magnetic monopole metal'' objects in EM theory, with the magnetic vacuum sector in the role of the scalar field. Of course the magnetic vacuum sector interacting with monopole-metal objects is just the dual of the electric sector with metal objects, so it must have the same behavior.[18]

The Casimir repulsion[2,3] between distinct D and N objects in scalar QFT has an analog in electromagnetism. For example, parallel metal and monopole-metal planes repel.[18] This system is the only one discussed in the literature known to the author involving both metal and monopole metal boundaries. However, basic physical arguments[4] rather convincingly reveal that distinct metal and monopole metal objects must necessarily have a repulsive Casimir effect.

Returning to the insertion of metal point and line boundaries into symmetric backgrounds, the reader might ask how physically relevant such small boundaries can be. The answer to this question provides useful insight into the electromagnetic Casimir effect.

1. Metal point boundary

Idealized as a tiny metal sphere, such a boundary has little effect on the magnetic vacuum sector. The boundary condition on the magnetic field at the surface of metal is $B_\perp = 0$. Essentially all field lines of any vacuum fluctuation are already parallel to the ''surface'' of the metal point; thus $B_\perp = 0$ causes little distortion of any nearby quantum fluctuation of the magnetic field. Presumably there is no room on the surface of a metal point for currents to flow to cancel intruding $B_\perp$ from vacuum fluctuations—but neither are such induced currents needed. It is such induced surface currents on metal objects (which cancel intruding $B_\perp$ from magnetic vacuum fluctuations) which make the magnetic vacuum sector exert repulsive forces between distinct metal objects. For metal points we see this magnetic repulsion mechanism is strongly suppressed.

The effect of the metal point boundary on the electric vacuum sector is quite pronounced. The metal boundary condition $E_\parallel = 0$ at the metal point strongly distorts nearby vacuum fluctuations of the electric field, bending field lines so they radially approach the metal point. Induced electric surface charges are of course responsible for this. It seems physically reasonable for induced surface charges to find room to exist even where surface currents cannot. In general the closed electric field lines of vacuum fluctuations can be ''opened'' by induced charges on a given metal object, or by induced charges on two distinct metal objects. The latter mechanism is what makes the electric vacuum sector attractive. This mechanism still functions for metal point boundaries, while as we have seen magnetic vacuum repulsion is hardly present for metal points. Thus one expects that small metal objects will always experience attractive Casimir forces from other nearby metal objects. For example, two metal points attract,[12] and a metal point will be attracted to a metal plane.[13] Moreover, a small metal object between parallel metal planes should attract these planes much as in the scalar example of Sec. IV B.

Such considerations suggest an interesting modification of the standard experiment,[19,20] verifying the static Casimir force between parallel uncharged metal plates. Let these plates have area A and separation L. Insertion into the midplane of a small uncharged metal object should modify the usual Casimir force $F_{\mathrm{Cas}} = A[c/L^4]$ on one plane[19,20] as follows:

$$F_{\mathrm{Cas}} = A\,\frac{c}{L^4} + \frac{d}{L^2}. \tag{6.1}$$

Here the constant $d>0$ fixing the strength of the metal point's attraction to either metal plate can be approximately calculated. Even if the second term in Eq. (6.1) is small relative to the first, it may well be detectable because of the quite different power of L involved. Moreover, one could displace the small metal object nearer to one plane than the other and observe the resulting change in $F_{\rm Cas}$. Further, this object could be moved laterally through the space between the parallel planes, which should result in a transient increase in $F_{\rm Cas}$. Successful observation of any of these effects would make the already convincing experimental Casimir effect even more compelling.

2. Metal line boundary

Idealized as a thin long wire, a metal line boundary also has little effect on the magnetic vacuum sector, but a strong effect on the electric vacuum sector. Thus metal wires will experience attractive Casimir forces from other nearby metal objects. Additional modifications of the standard Casimir effect suggest themselves, e.g., positioning an uncharged metal wire between the parallel planes, or moving the wire laterally through the space between these planes.

A final comment recalls the different (for some boundary configurations) Casimir forces obtained from $T_{\mu\nu}$ and $\theta_{\mu\nu}$ for a scalar quantum field. A similar situation exists in EM theory. One way to view this situation is that the Casimir effect provides in principle a way to distinguish experimentally between different stress tensors.

An article discussing the EM Casimir effect from the perspective presented above is in preparation.

ACKNOWLEDGMENTS

The author thanks I. Bender for his advice and influence on these calculations. He also thanks Penn State for Sabbatical Leave in 1997/98 and the Institut für Theoretische Physik der Universität Heidelberg for its kind hospitality during this leave. Helpful comments from a referee are acknowledged.

APPENDIX: MODE SUMS FOR THE VACUUM STRESS TENSOR

In a convenient and general notation the spatial heat kernel for a scalar field interacting with a static background is defined by

$$K(t|\vec{x},\vec{y})\equiv\sum_n e^{-t\omega_n^2}\phi_n(\vec{x})\,\overline{\phi_n}(\vec{y}), \tag{A1}$$

$$[-\Delta_x+V(\vec{x})]\phi_n(\vec{x})=\omega_n^2\phi_n(\vec{x}). \tag{A2}$$

The spatial modes $\phi_n(\vec{x})$ are understood to comprise a complete orthonormal set, selected from other such sets by the background potential $V(\vec{x})$. This potential represents either softened Dirichlet boundary structure, or actual Dirichlet boundary conditions on one or more surfaces. [In a formal way one might also think of $V(\vec{x})$ as representing Neumann boundary conditions as well.] In any case $V(\vec{x})$ couples the quantum field to the classical background. The heat kernel (A1) by construction satisfies Eq. (A2) in $\vec{x}$ and in $\vec{y}$ separately.

From the heat kernel (A1) one can obtain quite directly a number of important local physical quantities, in particular the canonical and improved vacuum stress tensors $T_{\mu\nu}$ and $\theta_{\mu\nu}$. These tensors consist of a number of terms, each having a mode-sum representation which can be gotten directly from

$$\sum_n (\omega_n^2)^{-s}\phi_n(\vec{x})\,\overline{\phi_n}(\vec{y})=\frac{1}{\Gamma(s)}\int_0^\infty dt\,t^{s-1}K(t|\vec{x},\vec{y}). \tag{A3}$$

Here the Mellin transform of K introduces ultraviolet singularities at $t=0$ which must be removed from the mode sum on the left (UV regularization). Aside from this, an explicit calculation of the bilocal mode sum (A3) for a given system yields straightforwardly $T_{\mu\nu}$ and $\theta_{\mu\nu}$, Casimir force densities throughout the spatial background of this system, and so on.

Explicit mode-sum formulas for $T_{\mu\nu}$ and $\theta_{\mu\nu}$ are given in Ref. 2. We list these again here for a massless scalar field with $V(\vec{x})=0$:

$$T_{00}=\frac{1}{4}\sum_n \omega_n|\phi_n|^2+\frac{1}{4}\sum_n \frac{1}{\omega_n}|\vec{\nabla}\phi_n|^2,$$

$$T_{ii}=\frac{1}{2}\sum_n \frac{1}{\omega_n}|\partial_i\phi_n|^2-\frac{1}{4}\sum_n \frac{1}{\omega_n}|\vec{\nabla}\phi_n|^2+\frac{1}{4}\sum_n \omega_n|\phi_n|^2, \quad i \text{ not summed},$$

$$T_{0i}=\theta_{0i}=\frac{i}{4}\sum_n [(\partial_i\phi_n)\overline{\phi_n}-\phi_n(\partial_i\overline{\phi_n})],$$

$$T_{ij}=\frac{1}{4}\sum_n \frac{1}{\omega_n}[\partial_i\phi_n\partial_j\overline{\phi_n}+\partial_j\phi_n\partial_i\overline{\phi_n}], \quad i\neq j$$

$$\theta_{00}=\frac{2d-1}{4d}\sum_n \omega_n|\phi_n|^2+\frac{1}{4d}\sum_n \frac{1}{\omega_n}|\vec{\nabla}\phi_n|^2,$$

$$\theta_{ii}=\frac{d+1}{4d}\sum_n \frac{1}{\omega_n}|\partial_i\phi_n|^2+\frac{1}{4d}\sum_n \omega_n|\phi_n|^2-\frac{1}{4d}\sum_n \frac{1}{\omega_n}|\vec{\nabla}\phi_n|^2$$
$$-\frac{d-1}{8d}\sum_n \frac{1}{\omega_n}[\phi_n\partial_i^2\overline{\phi_n}+(\partial_i^2\phi_n)\overline{\phi_n}],$$
$$i \text{ not summed},$$

$$\theta_{ij}=\frac{d+1}{8d}\sum_n \frac{1}{\omega_n}[\partial_i\phi_n\partial_j\overline{\phi_n}+\partial_j\phi_n\partial_i\overline{\phi_n}]-\frac{d-1}{8d}\sum_n \frac{1}{\omega_n}[\phi_n\partial_i\partial_j\overline{\phi_n}+(\partial_i\partial_j\phi_n)\overline{\phi_n}],$$
$$i\neq j.$$

[1] A. Actor, J. Phys. A **28**, 5737 (1995).
[2] A. Actor and I. Bender, Fortschr. Phys. **44**, 281 (1996).
[3] A. Actor and I. Bender, Phys. Rev. D **52**, 3581 (1995).
[4] A. Actor and I. Bender (unpublished).
[5] M. Bordag and J. Lindig, J. Phys. A **29**, 4481 (1996).
[6] D. Deutsch and P. Candelas, Phys. Rev. D **20**, 3063 (1979).
[7] K. Symanzik, Nucl. Phys. B **190**, 1 (1981).
[8] A. Actor, Fortschr. Phys. **43**, 141 (1995).
[9] D. M. McAvity and H. Osborn, Class. Quantum Grav. **8**, 603 (1991).
[10] J. Ambjorn and S. Wolfram, Ann. Phys. (N.Y.) **147**, 1 (1983).
[11] A. Actor, Phys. Rev. D **50**, 6560 (1994).
[12] H. B. G. Casimir, J. Chim. Phys. Phys.-Chim. Biol. **46**, 407 (1949); E. M. Lifshitz, Zh. Eksp. Teor. Fiz. **29**, 94 (1955) [Sov. Phys. JETP **2**, 73 (1956)]; G. Feinberg and J. Sucher, J. Chem. Phys. **48**, 3333 (1968); Phys. Rev. A **2**, 2395 (1970).
[13] H. B. G. Casimir and D. Polder, Phys. Rev. **73**, 360 (1948); T. H. Boyer, *ibid.* **180**, 19 (1969).
[14] T. H. Boyer, Phys. Rev. **174**, 1764 (1968).
[15] L. L. DeRaad, Jr. and K. A. Milton, Ann. Phys. (N.Y.) **136**, 229 (1981).
[16] W. Lukosz, Physica (Amsterdam) **56**, 109 (1971).
[17] J. Ambjorn and S. Wolfram, Ann. Phys. (N.Y.) **147**, 1 (1983).
[18] T. H. Boyer, Phys. Rev. A **9**, 2078 (1974).
[19] H. B. G. Casimir, Proc. K. Ned. Akad. Wet. **51**, 793 (1948).
[20] M. J. Sparnaay, Physica (Amsterdam) **24**, 751 (1958).

Partially solvable quantum many-body problems in D-dimensional space ($D=1,2,3,...$)

F. Calogero
Dipartimento di Fisica, Università di Roma I "La Sapienza," 00185 Roma, Italy, and Instituto Nazionale di Fisica Nucleare, Sezione di Roma, Italy

(Received 2 September 1998; accepted for publication 14 June 1999)

A simple technique employed almost three decades ago to manufacture partially solvable quantum many-body problems is revisited. [A quantum problem is "partially solvable" if (only) some of its eigenvalues and eigenfunctions can be exhibited]. The models thereby generated are characterized by Hamiltonians of normal form, i.e., standard kinetic plus momentum-independent potential energy; in most cases the latter features three-body, in addition to two-body and one-body, interactions. The setting refers to D-dimensional space; the examples focus on $D=1$, $D=2$, and $D\geqslant 2$, and include generalizations of, and additional results on, cases recently discussed in the literature, as well as new models. [S0022-2488(99)02409-3]

I. INTRODUCTION

Several recent papers[1–8] have presented and discussed partially solvable quantum many-body problems. (We call "partially solvable" a quantum problem if some of, but not all, its eigenvalues and eigenfunctions can be exhibited; for simplicity attention is hereafter restricted to bound states and discrete eigenvalues.) This prompted us to revisit a simple technique employed almost three decades ago[9] to manufacture partially solvable quantum many-body problems. We thereby generate a variety of such models. They are characterized by Hamiltonians of normal form, with standard kinetic energy and momentum-independent potential energy; in most cases the latter features three-body, in addition to two-body and one-body (and possibly some special "N-body"), interactions. The setting refers to D-dimensional space, and attention is restricted to rotation-invariant Hamiltonians. The examples focus on $D=1$, $D=2$, and $D\geqslant 2$. We exhibit a fairly general class of models, which is then specialized to specific examples. These include generalized versions of cases discussed in the literature,[1–10] as well as new partially solvable quantum many-body problems. Moreover, in some cases, our treatment of known models goes beyond previous findings, inasmuch as the collection of eigenstates we exhibit is larger than that previously known.

In Sec. II we display for convenience a representative list of the partially solvable models treated in this paper. Section III explains the main idea and reports the basic formulas from which the various models are then easily obtained, and exhibited as specific examples, in Sec. IV. Section V contains some concluding remarks and hints at future developments.

The relations of the specific models exhibited herein with previous findings[1–11] are discussed below, in Sec. IV, on a case-by-case basis. The notation is defined at the beginning of Sec. III, but it is sufficiently self-evident to allow direct browsing through Sec. II.

II. A REPRESENTATIVE LIST OF PARTIALLY SOLVABLE QUANTUM MANY-BODY PROBLEMS IN D-DIMENSIONAL SPACE

In this section we present, for convenience, a representative list of the partially solvable quantum many-body problems treated in this paper. We only report here those with one-, two-, and three-body interactions, for whose definition the reader is referred, if need be, to (3.1) with (3.2) [and let us also recall that *partial solvability* entails the capability to exhibit at least one solution of the stationary Schrödinger equation (3.3)]. These results are derived, and explained in somewhat more detail, in Sec. IV, where we also report more general models than those displayed here, including some that include certain special ''N-body'' potentials.

The first model we report [see (4.1b), (4.3b), and (4.3c)] is characterized by translation-invariant interactions:

$$u_j^{(1)}(\vec{r})=0, \tag{2.1a}$$

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=u_{jk}^{(2)}(r_{12})=[(m_j+m_k)/(4m_jm_k)][\alpha_{jk}(\alpha_{jk}+D-2)/r_{12}{}^2-\beta_{jk}(2\alpha_{jk}+D-1)/r_{12}], \tag{2.1b}$$

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u_{j_1j_2j_3}^{(3)}(\vec{r}_{12},\vec{r}_{13})$$
$$=(m_{j_1})^{-1}[(\vec{r}_{12}\cdot\vec{r}_{13})/(r_{12}r_{13})^2][\alpha_{j_1j_2}-\beta_{j_1j_2}r_{12}][\alpha_{j_1j_3}-\beta_{j_1j_3}r_{13}]. \tag{2.1c}$$

Here and below D is the dimensionality of space and m_j are the particle masses; the $N(N-1)$ constants $\alpha_{jk}=\alpha_{kj}$, $\beta_{jk}=\beta_{kj}$ are largely arbitrary (say, nonnegative), and $\vec{r}_{jk}\equiv\vec{r}_j-\vec{r}_k$.

The second and third models [see (4.7c) with (4.10b) and (4.10d), and again (4.7c) but with (4.11b) and (4.11d)] read

$$u_j^{(1)}(r)=2m_j\alpha^2r^2, \tag{2.2a}$$

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=u_{jk}^{(2)}(r_1,r_2)$$
$$=g_{jk}a\{[(m_j+m_k)/(8m_jm_k)]\{(g_{jk}-1)a\cosh[2a(r_1-r_2)]/\sinh^2[a(r_1-r_2)]$$
$$-2(D-1)\{r_2-r_1\exp[a(r_1-r_2)]\}/[r_1r_2\{1-\exp[a(r_1r_2)]\}]\}$$
$$+\alpha\{r_1-r_2\exp[a(r_1-r_2)]\}/\{1-\exp[a(r_1-r_2)]\}\}, \tag{2.2b}$$

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)$$
$$=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}\{1-\exp[a(r_1-r_2)]\}^{-1}\{1-\exp[a(r_1-r_3)]\}^{-1}; \tag{2.2c}$$

and, with $u_j^{(1)}(r)$ given again by (2.2a),

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=u_{jk}^{(2)}(r_1,r_2)$$
$$=g_{jk}b[(m_j+m_k)/(4m_jm_k)]\{(g_{jk}-1)b[r_1{}^{2(b-1)}+r_2{}^{2(b-1)}]/(r_1{}^b-r_2{}^b)^2$$
$$+(b+D-2)[r_1{}^{b-2}-r_2{}^{b-2}]/(r_1{}^b-r_2{}^b)\}, \tag{2.3a}$$

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}r_1{}^{2(b-1)}/[(r_1{}^b-r_2{}^b)(r_1{}^b-r_3{}^b)]. \tag{2.3b}$$

Here the $N(N-1)/2$ constants $g_{jk}=g_{kj}$ are also largely arbitrary (nonnegative), as well as the two constants a and b.

The fourth model we report [see (4.14a)–(4.14c)] reads

$$u_j^{(1)}(\vec{r})=u_j^{(1)}(r)=v_j(r), \tag{2.4a}$$

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=[(m_j+m_k)/(4m_jm_k)]g_{jk}(g_{jk}-1)(r_1{}^2+r_2{}^2)/(\vec{r}_1\cdot\vec{r}_2)^2, \tag{2.4b}$$

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}(\vec{r}_2\cdot\vec{r}_3)/[(\vec{r}_1\cdot\vec{r}_2)(\vec{r}_1\cdot\vec{r}_3)]. \tag{2.4c}$$

In these formulas $v_j(r)$ are rotation-invariant one-body potentials (arbitrary, but solvable in D-dimensional space for arbitrary angular momentum), and the $N(N-1)/2$ constants $g_{jk}=g_{kj}$ are again arbitrary nonnegative numbers.

The fifth model we report [see (4.21a)–(4.21c)]

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=[(m_j+m_k)/(4m_jm_k)](r_1{}^2+r_2{}^2)[g_{jk}(g_{jk}-1)/(\vec{r}_1\cdot\vec{r}_2)^2+\tilde{g}_{jk}(\tilde{g}_{jk}-1)/(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_2)^2], \tag{2.5a}$$

$$\begin{aligned} u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=(m_{j_1})^{-1}[&(\vec{r}_2\cdot\vec{r}_3)\{g_{j_1j_2}g_{j_1j_3}/[(\vec{r}_1\cdot\vec{r}_2)(\vec{r}_1\cdot\vec{r}_3)]\\ &+\tilde{g}_{j_1j_2}\tilde{g}_{j_1j_3}/[(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_2)(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_3)]\}\\ &+(\hat{z}\cdot\vec{r}_2\wedge\vec{r}_3)\{g_{j_1j_2}\tilde{g}_{j_1j_3}/[(\vec{r}_1\cdot\vec{r}_2)(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_3)]\\ &-\tilde{g}_{j_1j_2}g_{j_1j_3}/[(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_2)(\vec{r}_1\cdot\vec{r}_3)]\}], \end{aligned} \tag{2.5b}$$

again with (2.4a), see above, and the $N(N-1)$ constants $g_{jk}=g_{kj}$, $\tilde{g}_{jk}=\tilde{g}_{kj}$ arbitrary nonnegative numbers. In this case $D=2$ (while in the other cases displayed above the dimensionality D of space is arbitrary), and we employ the definition $\hat{z}\cdot\vec{r}_j\wedge\vec{r}_k\equiv x_jy_k-y_jx_k$.

In the case of the first three models, exact formulas are given in Sec. IV for (at least) one eigenvalue and the corresponding eigenfunction; for the last two, large families of eigenvalues and eigenfunctions are given, depending on N quantum numbers (the *complete* energy spectrum and set of eigenfunctions depends generally on ND quantum numbers).

Finally let us re-emphasize that not all these models are quite new; for a discussion of this aspect on a case-by-case basis, including of course the attribution of due credits, see Sec. IV.

III. MAIN IDEA AND BASIC RESULTS

Notation: D-dimensional vectors are denoted by superimposed arrows, say $\vec{r}_j$, with the usual (Euclidean) definitions for the scalar product $\vec{r}_j\cdot\vec{r}_k$ and the modulus $r_j\equiv(\vec{r}_j\cdot\vec{r}_j)^{1/2}$. Indices generally run from 1 to N, for N-body problems. Unless otherwise indicated, both D and N are *arbitrary positive integers*. We denote by m_j the mass of the jth particle, and generally write the quantum many-body Hamiltonian as follows:

$$H=-\sum_{j=1}^{N}(2m_j)^{-1}\Delta_j+U(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.1}$$

with $\Delta_j\equiv\vec{\nabla}_j\cdot\vec{\nabla}_j$, $\vec{\nabla}_j\equiv\partial/\partial\vec{r}_j$ (hence we have chosen units so that $\hbar=1$). The potential energy $U(\vec{r}_1,\ldots,\vec{r}_N)$ will be conveniently split up into multibody contributions as follows:

$$U(\vec{r}_1,\ldots,\vec{r}_N)=\sum_{n=1}^{N}U^{(n)}(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.2a}$$

$$U^{(1)}(\vec{r}_1,\ldots,\vec{r}_N)=\sum_{j=1}^{N}u_j^{(1)}(\vec{r}_j), \tag{3.2b}$$

$$U^{(2)}(\vec{r}_1,\ldots,\vec{r}_N)=(1/2)\sum_{j,k=1;j\neq k}^{N}u_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)=\sum_{1\leqslant j<k\leqslant N}u_{jk}^{(2)}(\vec{r}_j,\vec{r}_k), \tag{3.2c}$$

$$U^{(3)}(\vec{r}_1,\ldots,\vec{r}_N)=(1/2)\sum_{j_1,j_2,j_3=1;j_1\neq j_2,j_2\neq j_3,j_3\neq j_1}^{N} u^{(3)}_{j_1j_2j_3}(\vec{r}_{j_1},\vec{r}_{j_2},\vec{r}_{j_3})$$
$$=\sum_{j_1,j_2,j_3=1;j_1\neq j_2,j_1\neq j_3,j_2<j_3}^{N} u^{(3)}_{j_1j_2j_3}(\vec{r}_{j_1},\vec{r}_{j_2},\vec{r}_{j_3}), \tag{3.2d}$$

and so on (but in fact consideration will be mainly focused on problems with up to three-body interactions, except possibly for some special ''N-body'' contributions, see below). Note that, as indicated by these formulas, $u^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)=u^{(2)}_{kj}(\vec{r}_2,\vec{r}_1)$, $u^{(3)}_{j_1j_2j_3}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u^{(3)}_{j_1j_3j_2}(\vec{r}_1,\vec{r}_3,\vec{r}_2)$. We do not assume *a priori* all particles to be equal, indeed we allow them to have different masses and interactions; and for simplicity we pay no attention hereafter to the statistics obeyed by identical particles (this gap can be easily filled by the diligent reader).

We seek solutions of the stationary Schrödinger equation

$$[H-E]\Psi(\vec{r}_1,\ldots,\vec{r}_N)=0, \tag{3.3}$$

using for the eigenfunction $\Psi(\vec{r}_1,\ldots,\vec{r}_N)$ the *ansatz*

$$\Psi(\vec{r}_1,\ldots,\vec{r}_N)=F(\vec{r}_1,\ldots,\vec{r}_N)\,\psi(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.4a}$$

$$F(\vec{r}_1,\ldots,\vec{r}_N)=\prod_{n=1}^{N} F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.4b}$$

$$F^{(1)}(\vec{r}_1,\ldots,\vec{r}_N)=\prod_{j=1}^{N} f^{(1)}_j(\vec{r}_j)=\exp\left[\sum_{j=1}^{N}\varphi^{(1)}_j(\vec{r}_j)\right], \tag{3.4c}$$

$$F^{(2)}(\vec{r}_1,\ldots,\vec{r}_N)=\prod_{j=1,k=1;j\neq k}^{N}[f^{(2)}_{jk}(\vec{r}_j,\vec{r}_k)]^{1/2}$$
$$=\prod_{1\leq j<k\leq N}[f^{(2)}_{jk}(\vec{r}_j,\vec{r}_k)]$$
$$=\exp\left[(1/2)\sum_{j,k=1;j\neq k}^{N}\varphi^{(2)}_{jk}(\vec{r}_j,\vec{r}_k)\right]$$
$$=\exp\left[\sum_{1\leq j<k\leq N}\varphi^{(2)}_{jk}(\vec{r}_j,\vec{r}_k)\right], \tag{3.4d}$$

and so on (in fact we shall hardly go beyond, see below). Note that these formulas imply $f^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)=f^{(2)}_{kj}(\vec{r}_2,\vec{r}_1)$ and likewise $\varphi^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)=\varphi^{(2)}_{kj}(\vec{r}_2,\vec{r}_1)$, as well as

$$f^{(1)}_j(\vec{r})=\exp[\varphi^{(1)}_j(\vec{r})],\quad f^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)=\exp[\varphi^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)], \tag{3.5a}$$

entailing

$$\vec{\nabla}f^{(1)}_j(\vec{r})=f^{(1)}_j(\vec{r})\vec{\nabla}\varphi^{(1)}_j(\vec{r}),\quad \Delta f^{(1)}_j(\vec{r})=f^{(1)}_j(\vec{r})\{\Delta\varphi^{(1)}_j(\vec{r})+[\vec{\nabla}\varphi^{(1)}_j(\vec{r})]^2\}, \tag{3.5b}$$

and analogous formulas for $f^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)$ and $\varphi^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)$. In the following we employ the f or φ notation according to convenience.

The *ansatz* (3.4) is of course highly redundant; we take advantage of this by making appropriate choices for the ''correlation factors'' $F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)$, see below. The idea is to impose that $\Psi(\vec{r}_1,\ldots,\vec{r}_N)$, see (3.4), satisfy the Schrödinger equation (3.3), and to identify the corresponding potential $U(\vec{r}_1,\ldots,\vec{r}_N)$, see (3.1) and (3.2), and the energy eigenvalue E; there will be cases when, to a specific potential $U(\vec{r}_1,\ldots,\vec{r}_N)$, there correspond different eigenfunctions $\Psi_{\mu_1,\mu_2,\ldots}(\vec{r}_1,\ldots,\vec{r}_N)$, labeled by one or more quantum numbers $\mu_1,\mu_2,\ldots$, and correspondingly energy eigenvalues $E_{\mu_1,\mu_2,\ldots}$ (possibly degenerate, see below).

Let us then insert the *ansatz* (3.4) into the Schrödinger equation (3.3). There are two approaches, which we like to set out explicitly so as to avoid misunderstandings. Firstly we obtain, via (3.1), the following formulas:

$$(\tilde{H}-\tilde{E})\psi=0, \tag{3.6}$$

with

$$\tilde{H}=-\sum_{j=1}^{N}(2m_j)^{-1}\Delta_j-i\sum_{j=1}^{N}[\vec{W}_j(\vec{r}_1,\ldots,\vec{r}_N)\cdot\vec{\nabla}_j+\vec{\nabla}_j\cdot\vec{W}_j(\vec{r}_1,\ldots,\vec{r}_N)]+\tilde{U}(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.7a}$$

$$\tilde{U}(\vec{r}_1,\ldots,\vec{r}_N)-\tilde{E}=U(\vec{r}_1,\ldots,\vec{r}_N)-E-\sum_{j=1}^{N}(2m_j)^{-1}[\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N)]^2, \tag{3.7b}$$

where

$$\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N)=[F(\vec{r}_1,\ldots,\vec{r}_N)]^{-1}\vec{\nabla}_jF(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.7c}$$

$$\vec{W}_j(\vec{r}_1,\ldots,\vec{r}_N)=-i(2m_j)^{-1}\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N). \tag{3.7d}$$

Note that in the right-hand-side of (3.7a) $\vec{\nabla}_j$ is an operator, so that

$$\vec{\nabla}_j\cdot\vec{W}_j\psi=\vec{W}_j\cdot(\vec{\nabla}_j\psi)+\psi(\vec{\nabla}_j\cdot\vec{W}_j). \tag{3.8}$$

These formulas entail the following (trivial) Remark: *If the Schrödinger equation (3.3) with (3.1) is solvable, also solvable is the Schrödinger equation (3.6) with (3.7).* Indeed the two problems are related by the unitary transformation $F^{-1}(H-E)F=(\tilde{H}-\tilde{E})$ [we allow for the possibility to chose $\tilde{E}\neq E$, so that neither $U(\vec{r}_1,\ldots,\vec{r}_N)$ nor $\tilde{U}(\vec{r}_1,\ldots,\vec{r}_N)$ contain an *additive constant*; a requirement which makes sense only if both potentials vanish asymptotically]. The (also trivial) classical counterpart of this *Remark* states that *the two classical problems characterized by the Hamiltonians*

$$H(\vec{p}_1,\ldots,\vec{p}_N;\vec{r}_1,\ldots,\vec{r}_N)=\sum_{j=1}^{N}(2m_j)^{-1}p_j^{\,2}+U(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.9a}$$

$$\tilde{H}(\vec{P}_1,\ldots,\vec{P}_N;\vec{r}_1,\ldots,\vec{r}_N)=\sum_{j=1}^{N}(2m_j)^{-1}P_j^{\,2}+2\sum_{j=1}^{N}\vec{P}_j\cdot\vec{W}_j(\vec{r}_1,\ldots,\vec{r}_N)+\tilde{U}(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.9b}$$

are equivalent [they entail the same evolution for the particle positions $\vec{r}_j(t)$]. Indeed they are related by the *canonical transformation* $\vec{p}_j=\vec{P}_j-i\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N)$, $\vec{r}_j=\vec{R}_j$, with $\vec{W}_j(\vec{r}_1,\ldots,\vec{r}_N)$ related to $\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N)$ by (3.7d), and of course $\tilde{U}(\vec{r}_1,\ldots,\vec{r}_N)$ related to $U(\vec{r}_1,\ldots,\vec{r}_N)$ by (3.7b). Note that these *Remarks* hold for an *arbitrary* choice of the N D-vector functions $\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N)$, provided they admit the representation (3.7c). They entail the possibility to introduce solvable many-body problems, by starting from a trivially solvable problem [say, from (3.1) with

$U(\vec{r}_1,\ldots,\vec{r}_N)$ given by (3.2a) where $U^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)=0$ for $n>1$ and $u_j^{(1)}(r)$ in (3.2b) are solvable one-body potentials] and getting a new, *apparently* nontrivial, Hamiltonian via (3.7) or (3.9). This is not the route we follow below.

The second approach is, perhaps, less trivial. It is based on another set of formulas [also equivalent, via (3.4), to (3.3) with (3.1)]:

$$-\sum_{j=1}^{N}(2m_j)^{-1}[\Delta_j+2\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N)\cdot\vec{\nabla}_j]\psi(\vec{r}_1,\ldots,\vec{r}_N)=0, \tag{3.10}$$

$$\vec{V}_j(\vec{r}_1,\ldots,\vec{r}_N)=\sum_{n=1}^{N}\vec{V}_j^{(n)}(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.11}$$

$$U(\vec{r}_1,\ldots,\vec{r}_N)-E=\sum_{j=1}^{N}(2m_j)^{-1}\sum_{n=1}^{N}[F(\vec{r}_1,\ldots,\vec{r}_N)]^{-1}\Delta_j F(\vec{r}_1,\ldots,\vec{r}_N), \tag{3.12a}$$

$$U(\vec{r}_1,\ldots,\vec{r}_N)-E=\sum_{j=1}^{N}(2m_j)^{-1}\left\{\sum_{n=1}^{N}[F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)]^{-1}\Delta_j F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)\right.$$
$$\left.+\sum_{m,n=1;m\neq n}^{N}\vec{V}_j^{(m)}(\vec{r}_1,\ldots,\vec{r}_N)\cdot\vec{V}_j^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)\right\}, \tag{3.12b}$$

$$U(\vec{r}_1,\ldots,\vec{r}_N)-E=\sum_{j=1}^{N}(2m_j)^{-1}\left\{\sum_{n=1}^{N}\vec{\nabla}_j\cdot\vec{V}_j^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)\right.$$
$$\left.+\sum_{m,n=1}^{N}\vec{V}_j^{(m)}(\vec{r}_1,\ldots,\vec{r}_N)\cdot\vec{V}_j^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)\right\}, \tag{3.12c}$$

where

$$\vec{V}_j^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)=[F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)]^{-1}\vec{\nabla}_j F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)=\vec{\nabla}_j\log[F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)]. \tag{3.13}$$

The equivalence among (3.12a) and (3.12b) is consistent with the definition (3.4b); likewise, the equivalence among (3.12b) and (3.12c) is entailed by the definition (3.13) [note that in the double sum in the right-hand-side of (3.12b) the ''diagonal'' term with $m=n$ is omitted while it is instead present in the double sum in the right-hand-side of (3.12c)].

The modified Schrödinger equation (3.10) provides a convenient starting point for the treatment of *completely solvable* models, including the well-known *one-dimensional* case with *equal particles* and *two-body inverse-square potentials* (possibly with an additional interaction, for instance a harmonic oscillator potential, or a ''Coulomb-type'' potential, depending only on the global coordinate $\Sigma_{j=1}^{n}x_j{}^2$).[11,7] Hereafter we set instead $\psi(\vec{r}_1,\ldots,\vec{r}_N)=1$, and thereby satisfy (3.10) trivially. We moreover set $F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)=1$ for $2<n<N$, namely we restrict only to $F^{(1)}(\vec{r}_1,\ldots,\vec{r}_N)$, $F^{(2)}(\vec{r}_1,\ldots,\vec{r}_N)$ and $F^{(N)}(\vec{r}_1,\ldots,\vec{r}_N)$ (with a special *ansatz* for this latter, see below) our freedom to chose, compatibly with the *ansatzen* (3.4c), (3.4d) etc., nontrivial ''correlation factors'' $F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)$ [which then determine the ''partially solvable'' potential $U(\vec{r}_1,\ldots,\vec{r}_N)$ via (3.12), see below]. To explain this restriction let us emphasize that, for $2n-1\leqslant N$, a nontrivial presence of the ''correlation factor'' $F^{(n)}(\vec{r}_1,\ldots,\vec{r}_N)$ yields up to $(2n-1)$-body interactions in the corresponding partially solvable potential [see (3.12) and below].

Hence [from (3.12b)]

$$U(\vec{r}_1,\ldots,\vec{r}_N)-E=\sum_{j=1}^{N}(2m_j)^{-1}\{[F^{(1)}]^{-1}\Delta_jF^{(1)}+[F^{(2)}]^{-1}\Delta_jF^{(2)}$$
$$+[F^{(N)}]^{-1}\Delta_jF^{(N)}+2\vec{V}_j^{(1)}\cdot\vec{V}_j^{(2)}+2\vec{V}_j^{(1)}\cdot\vec{V}_j^{(N)}+2\vec{V}_j^{(2)}\cdot\vec{V}_j^{(N)}\}. \tag{3.14}$$

To preserve rotation-invariance we moreover set [see (3.4c)]

$$f_j^{(1)}(\vec{r})=f_j^{(1)}(r)=\exp[\varphi_j^{(1)}(\vec{r})]=\exp[\varphi_j^{(1)}(r)] \tag{3.15a}$$

so that

$$[F^{(1)}(\vec{r}_1,\ldots,\vec{r}_N)]^{-1}\Delta_jF^{(1)}(\vec{r}_1,\ldots,\vec{r}_N)=[f_j^{(1)\prime\prime}(r_j)+(D-1)f_j^{(1)\prime}(r_j)/r_j]/f_j^{(1)}(r_j)$$
$$=\varphi_j^{(1)\prime\prime}(r_j)+[\varphi_j^{(1)\prime}(r_j)]^2+(D-1)\varphi_j^{(1)\prime}(r_j)/r_j, \tag{3.15b}$$

$$\vec{V}_j^{(1)}=\vec{r}_jf_j^{(1)\prime}(r_j)/[r_jf_j^{(1)}(r_j)]=\varphi_j^{(1)\prime}(r_j)\vec{r}_j/r_j. \tag{3.15c}$$

Of course appended primes denote differentiations with respect to the argument r_j.

We moreover set

$$F^{(N)}(\vec{r}_1,\ldots,\vec{r}_N)=F^{(N)}(\rho)=\exp[\Phi(\rho)], \tag{3.16a}$$

$$\rho^2=\sum_{j=1}^{N}2m_jr_j^2, \tag{3.16b}$$

so that

$$\sum_{j=1}^{N}(2m_j)^{-1}[F^{(N)}(\rho)]^{-1}\Delta_jF^{(N)}(\rho)=[\ddot{F}^{(N)}(\rho)+(ND-1)\dot{F}^{(N)}(\rho)/\rho]/F^{(N)}(\rho)$$
$$=\ddot{\Phi}(\rho)+[\dot{\Phi}(\rho)]^2+(ND-1)\dot{\Phi}(\rho)/\rho, \tag{3.16c}$$

$$\vec{V}_j^{(N)}=2m_j\vec{r}_j\dot{\Phi}(\rho)/\rho, \tag{3.16d}$$

where of course superimposed dots denote differentiation with respect to the argument ρ. In particular in the following we generally set (unless we state otherwise)

$$F^{(N)}(\rho)=\exp(-\alpha\rho^2/2),\quad \Phi(\rho)=-\alpha\rho^2/2,\quad \dot{\Phi}(\rho)=-\alpha\rho,\quad \ddot{\Phi}(\rho)=-\alpha, \tag{3.17a}$$

so that

$$\sum_{j=1}^{N}(2m_j)^{-1}[F^{(N)}(\rho)]^{-1}\Delta_jF^{(N)}(\rho)=\alpha^2\rho^2-\alpha ND, \tag{3.17b}$$

$$\vec{V}_j^{(N)}=-2m_j\alpha\vec{r}_j. \tag{3.17c}$$

To complete the construction of the potential $U(\vec{r}_1,\ldots,\vec{r}_N)$, see (3.14), we must still assign the two-body correlation functions $f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$ or $\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$, see (3.4d). The discussion of various specific choices for these functions, as well as for the functions $f_j^{(1)}(r)$ or $\varphi_j^{(1)}(r)$, is postponed to the following section.

We end this section by rewriting the expression for $U(\vec{r}_1,\ldots,\vec{r}_N)$ and E, see (3.14), entailed by the choices made so far, see (3.15) and (3.16) with (3.17):

$$U(\vec{r}_1,\ldots,\vec{r}_N)=\sum_{j=1}^{N} u_j^{(1)}(r_j)+\sum_{j,k=1;j<k}^{N} u_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)+\sum_{j_1,j_2,j_3=1;j_1\neq j_2,j_1\neq j_3,j_2<j_3}^{N} u_{j_1j_2j_3}^{(3)}(\vec{r}_{j_1},\vec{r}_{j_2},\vec{r}_{j_3}), \tag{3.18a}$$

$$E=\sum_{j=1}^{N} E_j^{(1)}+\sum_{j,k=1j:j<k}^{N} E_{jk}^{(2)}+\sum_{j_1,j_2,j_3=1;j_1\neq j_2,j_1\neq j_3,j_2<j_3}^{N} E_{j_1j_2j_3}^{(3)}, \tag{3.18b}$$

$$u_j^{(1)}(r)=2m_j\alpha^2r^2-\alpha D+\{(2m_j)^{-1}[f_j^{(1)\prime\prime}(r)+(D-1)f_j^{(1)\prime}(r)/r]$$
$$-2\alpha r f_j^{(1)\prime}(r)+E_j^{(1)}f_j^{(1)}\}/f_j^{(1)}(r), \tag{3.18c}$$

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=[f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)]^{-1}\{[(m_j+m_k)/(4m_jm_k)][(\Delta_1+\Delta_2)f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$$
$$+2\{[\varphi_j^{(1)\prime}(r_1)/r_1](\vec{r}_1\cdot\vec{\nabla}_1)+[\varphi_k^{(1)\prime}(r_2)/r_2](\vec{r}_2\cdot\vec{\nabla}_2)\}f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)]$$
$$-\alpha[(\vec{r}_1\cdot\vec{\nabla}_1)+(\vec{r}_2\cdot\vec{\nabla}_2)]f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)+E_{jk}^{(2)}f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)\}, \tag{3.18d}$$

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=E_{j_1j_2j_3}^{(3)}+(m_{j_1})^{-1}[\vec{\nabla}_1\varphi_{j_1j_2}^{(2)}(\vec{r}_1,\vec{r}_2)]\cdot[\vec{\nabla}_1\varphi_{j_1j_3}^{(2)}(\vec{r}_1,\vec{r}_3)]. \tag{3.18e}$$

Let us re-emphasize that we are still free to choose at our convenience the one-body functions $f_j^{(1)}(r)$ [or $\varphi_j^{(1)}(r)$, see (3.5a)] and the two-body functions $f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$ [or $\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$, see (3.5a)], as well as the energy constants $E_j^{(1)}$, $E_{jk}^{(2)}$, and $E_{j_1j_2j_3}^{(3)}$; we have on the other hand committed ourselves to the simple choice (3.17) for $\Phi(\rho)$. Cases in which a more general choice for $\Phi(\rho)$ is convenient will also be specially mentioned and considered below. It is easily seen that this shall entail the following modifications: in (3.18c) the term $2m_j\alpha^2r^2-\alpha D$ must be replaced by $\{\ddot{\Phi}(\rho)+[\dot{\Phi}(\rho)]^2+(ND-1)\dot{\Phi}(\rho)/\rho\}/N$ [see (3.17b) and (3.16c)], and in the other two places, in (3.18c) and (3.18d), where the constant α appears, it must be replaced by $-\dot{\Phi}(\rho)/\rho$ [see (3.17c) and (3.16d)], of course always with ρ defined by (3.16b). These modifications mar the interpretation of $u_j^{(1)}(r),u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$, respectively, $u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)$ as one-, two- respectively three-body potentials; this is why we prefer to write (3.18) as we did, rather than in the more general form entailed by these modifications. But we must also emphasize that, even though the formula (3.18) seems to provide a natural distinction between one-, two- and three-body interactions, there also are cases (see below) when parts, say, of the potential $u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$ are more naturally interpreted as one-body contributions: this is generally the case whenever $u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$, or parts of it, separate into additive contributions of two functions each depending only upon one of the two arguments $\vec{r}_1$ and $\vec{r}_2$. Hence a more complete analysis, including a proper separation in one-, two-, three- and possibly *N*-body potentials, can only be made after we specify, in the following section, various choices for the functions $f_j^{(1)}(r)$ and $f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$.

IV. EXAMPLES

In this section we derive various examples of partially solvable quantum many-body problems. Our main tool are the equations (3.18) displayed at the end of the preceding Sec. III, which are now implemented by making specific choices for the two-body ‘‘correlation functions’’ $f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$ and the ‘‘one-body’’ functions $f_j^{(1)}(r)$ [or equivalently $\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$ and $\varphi_j^{(1)}(r)$, see (3.5a)]. Let us re-emphasize the large freedom in the choice of these functions, and correspondingly of the ‘‘energies’’ $E_j^{(1)}$, $E_{jk}^{(2)}$, and $E_{j_1j_2j_3}^{(3)}$; the only restriction is that the wave function $\Psi(\vec{r}_1,\ldots,\vec{r}_N)$, see (3.4), be *normalizable* (possibly up to the free motion of the center of mass of the many-body system, see below).

We conveniently subdivide this section into four parts, treating thereby four different classes of partially solvable many-body problems.

A. Models of type A (translation invariant)

Let us set

$$E_j^{(1)}=E_{j_1j_2j_3}^{(3)}=0,\quad \alpha=0,\quad f_j^{(1)}(r)=1,\quad \varphi_j^{(1)}(r)=0, \tag{4.1a}$$

entailing [see (3.18c)]

$$u_j^{(1)}(r)=0, \tag{4.1b}$$

and let us moreover set

$$f_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)=f_{jk}^{(2)}(r_{jk})=f_{kj}^{(2)}(r_{jk}), \tag{4.1c}$$

where we introduce the convenient short-hand notation

$$\vec{r}_{jk}\equiv\vec{r}_j-\vec{r}_k,\quad r_{jk}\equiv(\vec{r}_{jk}\cdot\vec{r}_{jk})^{1/2}\equiv r_{kj}. \tag{4.1d}$$

Then [see (3.18d)]

$$\begin{aligned}u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)&=u_{jk}^{(2)}(r_{12})\\&=\{[(m_j+m_k)/(4m_jm_k)][f_{jk}^{(2)\prime\prime}(r_{12})+(D-1)f_{jk}^{(2)\prime}(r_{12})/r_{12}]\\&\quad+E_{jk}^{(2)}f_{jk}^{(2)}(r_{12})\}/f_{jk}^{(2)}(r_{12}),\end{aligned} \tag{4.1e}$$

and (see (3.18e))

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u_{j_1j_2j_3}^{(3)}(\vec{r}_{12},\vec{r}_{13})=(m_{j_1})^{-1}[(\vec{r}_{12}\cdot\vec{r}_{13})/(r_{12}r_{13})]\varphi_{j_1j_2}^{(2)\prime}(r_{12})\varphi_{j_1j_3}^{(2)\prime}(r_{13}). \tag{4.1f}$$

Hence, by choosing $f_{jk}^{(2)}(r)$ as an (appropriate; see examples below) solution of the equation

$$[(m_j+m_k)/(4m_jm_k)][f_{jk}^{(2)\prime\prime}(r)+(D-1)f_{jk}{}'(r)/r]+E_{jk}^{(2)}f_{jk}^{(2)}(r)=v_{jk}(r)f_{jk}^{(2)}(r), \tag{4.2a}$$

with $v_{jk}(r)$ arbitrarily given (and $E_{jk}^{(2)}$ appropriately chosen, see examples below), one obtains

$$u_{jk}^{(2)}(r)=v_{jk}(r). \tag{4.2b}$$

Note that (4.2a) is just the two-body D-dimensional Schrödinger equation (after separation of the two-body center-of-mass motion) for the jth and kth particles, with the pair potential $u_{jk}^{(2)}(r)=v_{jk}(r)$, r being the (scalar) interparticle distance. It is thus seen that solutions (i.e., eigenvalues and eigenfunctions) of *many-body problems* with arbitrary translation- and rotation-invariant pair potentials can be found, provided solutions of the corresponding problem are known *in the two-body context*, and additional, *appropriate* three-body interactions are also present. On the other hand *partially solvable two-body problems* can easily be manufactured by choosing appropriately a two-body wave function and computing the corresponding potential. In this manner it is easy to manufacture a large zoo of *partially solvable rotation- and translation-invariant quantum many-body problems*.

For instance by choosing

$$f_{jk}^{(2)}(r)=r^{\alpha_{jk}}\exp(-\beta_{jk}r), \tag{4.3a}$$

we conclude that the N-body problem without one-body interactions, with the long-range two-body interactions

$$u_{jk}^{(2)}(r)=[(m_j+m_k)/(4m_jm_k)][\alpha_{jk}(\alpha_{jk}+D-2)/r^2-\beta_{jk}(2\alpha_{jk}+D-1)/r], \tag{4.3b}$$

and the long-range three-body interactions

$$u^{(3)}_{j_1j_2j_3}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u^{(3)}_{j_1j_2j_3}(\vec{r}_{12},\vec{r}_{13})$$
$$=(m_{j_1})^{-1}[(\vec{r}_{12}\cdot\vec{r}_{13})/(r_{12}r_{13})^2][\alpha_{j_1j_2}-\beta_{j_1j_2}r_{12}][\alpha_{j_1j_3}-\beta_{j_1j_3}r_{13}], \quad (4.3c)$$

is *partially solvable*, possessing a ground state with energy

$$E=-\sum_{j,k=1;j<k}^{N}[(m_j+m_k)/(4m_jm_k)]\beta_{jk}{}^2 \quad (4.3d)$$

and eigenfunction

$$\Psi(\vec{r}_1,\ldots,\vec{r}_N)=\exp\left[-\sum_{j,k=1;j<k}^{N}\beta_{jk}r_{jk}\right]\prod_{j,k=1;j<k}^{N}(r_{jk})^{\alpha_{jk}}. \quad (4.3e)$$

Here the $N(N-1)$ constants $\alpha_{jk}=\alpha_{kj}$ and $\beta_{jk}=\beta_{kj}$ are *arbitrary*, except for the conditions

$$\alpha_{jk}>0, \quad \sum_{k=1,k\neq j}^{N}\beta_{jk}>0, \quad (4.3f)$$

which are sufficient to guarantee that the eigenfunction (4.3e) be normalizable (in the N-body center-of-mass frame).

Let us however emphasize that these results are not new: they essentially coincide with those of Ref. 9, except for the specific example given above, and the fact that here we kept arbitrary the dimensionality D of space (in Ref. 9 attention focused on the physical $D=3$ case, although it was emphasized there that the extension to different space dimensionality is easy). Let us also note that the results, as given here, hold only for $D\geqslant 2$; the one-dimensional case requires a (marginally) modified treatment because in that case the derivative of $r_{12}=|x_1-x_2|$ with respect to x_1 or x_2 has a discontinuity at $x_1=x_2$ (for an example belonging to this case see Ref. 10).

Let us finally focus on the special case with

$$f^{(2)}_{jk}(r)=r^{g_{jk}}, \quad \varphi^{(2)}_{jk}{}'(r)=g_{jk}/r, \quad \varphi^{(2)}_{jk}{}''(r)=-g_{jk}/r^2 \quad (4.4)$$

[namely, the case (4.3a) with $\beta_{jk}=0$ and, for notational convenience, $\alpha_{jk}=g_{jk}=g_{kj}$], but now without setting $\Phi(\rho)=0$ [as entailed above by the choice (3.17) with $\alpha=0$]. As noted above [see after (3.18e)], and using (4.4), this entails the replacement

$$[f^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)]^{-1}\{-\alpha[(\vec{r}_1\cdot\vec{\nabla}_1)+(\vec{r}_2\cdot\vec{\nabla}_2)]f^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)\}\Rightarrow g_{jk}\dot{\Phi}(\rho)/\rho \quad (4.5a)$$

on the right-hand side of (3.18d), as well as the replacement

$$2m_j\alpha^2r^2-\alpha D\Rightarrow\{\ddot{\Phi}(\rho)+[\dot{\Phi}(\rho)]^2+(ND-1)\dot{\Phi}(\rho)/\rho\}/N \quad (4.5b)$$

on the right-hand side of (3.18c). Hence in this special case the partially solvable potential can be written as follows:

$$U(\vec{r}_1,\ldots,\vec{r}_N)=\alpha^2\sum_{j=1}^{N}2m_jr_j{}^2+\sum_{j,k=1;j<k}^{N}[(m_j+m_k)/(2m_jm_k)]g_{jk}(g_{jk}+D-2)/r_{jk}{}^2$$
$$+\sum_{j_1,j_2,j_3=1;j_1\neq j_2,j_1\neq j_3,j_2<j_3}^{N}(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}[(\vec{r}_{j_1j_2}\cdot\vec{r}_{j_1j_3})/(r_{j_1j_2}r_{j_1j_3})^2]$$
$$+u^{(N)}(\rho), \quad (4.6a)$$

with ρ defined by (3.16b), and $u^{(N)}(\rho)$ a (largely arbitrary) function (''N-body potential''). The corresponding eigenfunction [see (3.3)] reads

$$\Psi(\vec{r}_1,\ldots,\vec{r}_N)=\chi_\mu(\rho)\prod_{j,k=1;j<k}^{N}(r_{jk})^{g_{jk}}, \tag{4.6b}$$

with energy, see (3.3),

$$E=E_\mu, \tag{4.6c}$$

the two quantities $\chi_\mu(\rho)=\log[\Phi(\rho)]$ and E_μ being related by the eigenvalue equation

$$-\ddot{\chi}_\mu(\rho)-(ND-1+G)\dot{\chi}_\mu(\rho)/\rho+[\alpha^2\rho^2+u^{(N)}(\rho)]\chi_\mu(\rho)=E_\mu\chi_\mu(\rho), \tag{4.6d}$$

or equivalently, via $\chi_\mu(\rho)=\rho^{(1-ND-G)/2}\psi_\mu(\rho)$, by the Schrödinger-type equation

$$-\ddot{\psi}_\mu(\rho)[(ND+G-1)(ND+G-3)/4+\alpha^2\rho^2+u^{(N)}(\rho)]\psi_\mu(\rho)=E_\mu\psi_\mu(\rho). \tag{4.6e}$$

Here

$$G=\sum_{j,k=1;j<k}^{N} g_{jk}, \tag{4.6f}$$

α is an arbitrary constant [which has been conveniently reintroduced on the right-hand side of (4.6a) as well as in (4.6d) and (4.6e)], the dots denote of course differentiation with respect to the global radial variable ρ, see (3.16b), μ is a quantum number, and the eigenvalue equation (4.6d) and (4.6e) is constrained by the requirements that (4.6b) be *normalizable*, sufficient conditions for the existence of a discrete infinity of such solutions, parametrized, say, by $\mu=0,1,2,\ldots$, are the requirement that all the ''coupling constants'' g_{jk} be non-negative, $g_{jk}\geqslant 0$, and that the potential $\alpha^2\rho^2+u^{(N)}(\rho)$ be confining. The two (well-known) cases in which the solutions $\chi_\mu(\rho)$ or $\psi_\mu(\rho)$ can be explicitly exhibited in terms of known functions (in fact, elementary functions and Laguerre polynomials), and the corresponding eigenvalues easily computed, for *all* values of the quantum number μ, are those characterized by the presence of centrifugal and harmonic or attractive Coulomb potentials (this latter not being confining, but possessing an infinite sequence of bound states due to its long range), corresponding, respectively, to $\alpha\neq 0$, $u^{(N)}(\rho)=\lambda^2/\rho^2$, or $\alpha=0$, $u^{(N)}(\rho)=\lambda^2/\rho^2-q^2/\rho$; but in these two cases the partial solvability of this N-body problem, as detailed above, is not a new result: it had already been demonstrated (in the equal-particle case: $m_j=m$, $g_{jk}=g$) by Khare.[8]

B. Models of type B

Let us set

$$f_j^{(1)}(r)=1,\quad \varphi_j^{(1)}(r)=0,\quad E_j^{(1)}=\alpha D,\quad E_{j_1j_2j_3}^{(3)}=0, \tag{4.7a}$$

and

$$f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=f_{jk}^{(2)}(r_1,r_2),\quad \varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=\varphi_{jk}^{(2)}(r_1,r_2). \tag{4.7b}$$

Then

$$u_j^{(1)}(r)=2m_j\alpha^2r^2, \tag{4.7c}$$

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=u_{jk}^{(2)}(r_1,r_2)=[(m_j+m_k)/(4m_jm_k)]\{\varphi_{jk,11}^{(2)}(r_1,r_2)+[\varphi_{jk,1}^{(2)}(r_1,r_2)]^2+\varphi_{jk,22}^{(2)}(r_1,r_2)+[\varphi_{jk,2}^{(2)}(r_1,r_2)]^2+(D-1)[\varphi_{jk,1}^{(2)}(r_1,r_2)/r_1+\varphi_{jk,2}^{(2)}(r_1,r_2)/r_2]\}-\alpha[r_1\varphi_{jk,1}^{(2)}(r_1,r_2)+r_2\varphi_{jk,2}^{(2)}(r_1,r_2)]+E_{jk}^{(2)}, \tag{4.7d}$$

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)=(m_{j_1})^{-1}\varphi_{j_1j_2,1}^{(2)}(r_1,r_2)\varphi_{j_1j_3,1}^{(2)}(r_1,r_3), \tag{4.7e}$$

where of course $\varphi_{jk,1}^{(2)}(r_1,r_2)\equiv\partial\varphi_{jk}^{(2)}(r_1,r_2)/\partial r_1$, $\varphi_{jk,2}^{(2)}(r_1,r_2)\equiv\partial\varphi_{jk}^{(2)}(r_1,r_2)/\partial r_2$, $\varphi_{jk,11}^{(2)}(r_1,r_2)\equiv\partial^2\varphi_{jk}^{(2)}(r_1,r_2)/\partial r_1{}^2$ and so on.

We now specialize further the choice of the functions $f_{jk}^{(2)}(r_1,r_2)$ [or $\varphi_{jk}^{(2)}(r_1,r_2)$, see (3.5a)] by setting

$$f_{jk}^{(2)}(r_1,r_2)=f_{jk}^{(2)}[\chi_j(r_1)-\chi_k(r_2)],\quad \varphi_{jk}^{(2)}(r_1,r_2)=\varphi_{jk}^{(2)}[\chi_j(r_1)-\chi_k(r_2)], \tag{4.8a}$$

so that

$$\varphi_{jk,1}^{(2)}(r_1,r_2)=\chi_j{}'(r_1)\dot{\varphi}_{jk}^{(2)}, \tag{4.8b}$$

$$\varphi_{jk,11}^{(2)}(r_1,r_2)=[\chi_j{}'(r_1)]^2\ddot{\varphi}_{jk}^{(2)}+\chi_j''(r_1)\dot{\varphi}_{jk}^{(2)}, \tag{4.8c}$$

$$\varphi_{jk,2}^{(2)}(r_1,r_2)=-\chi_k{}'(r_2)\dot{\varphi}_{jk}^{(2)}, \tag{4.8d}$$

$$\varphi_{jk,22}^{(2)}(r_1,r_2)=[\chi_k{}'(r_2)]^2\ddot{\varphi}_{jk}^{(2)}-\chi_k''(r_2)\dot{\varphi}_{jk}^{(2)}, \tag{4.8e}$$

where the superimposed dots denote differentiation with respect to the argument of the $\varphi_{jk}^{(2)}$ functions, which is always the quantity $[\chi_j(r_1)-\chi_k(r_2)]$, and appended primes denote differentiation with respect to r_1 or r_2, as the case may be.

With this choice

$$\begin{aligned}u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)&=u_{jk}^{(2)}(r_1,r_2)\\&=[(m_j+m_k)/(4m_jm_k)][\{[\chi_j{}'(r_1)]^2+[\chi_k{}'(r_2)]^2\}\{\ddot{\varphi}_{jk}^{(2)}+[\dot{\varphi}_{jk}^{(2)}]^2\}\\&\quad+\{\chi_j''(r_1)-\chi_k''(r_2)+(D-1)[\chi_j{}'(r_1)/r_1-\chi_k{}'(r_2)/r_2]\}\dot{\varphi}_{jk}^{(2)}]\\&\quad-\alpha[r_1\chi_j{}'(r_1)-r_2\chi_k{}'(r_2)]\dot{\varphi}_{jk}^{(2)}+E_{jk}^{(2)},\end{aligned} \tag{4.8f}$$

where the argument of the functions $\varphi_{jk}^{(2)}$ is always $[\chi_j(r_1)-\chi_k(r_2)]$, and

$$u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)=(m_{j_1})^{-1}[\chi_{j_1}{}'(r_1)]^2\dot{\varphi}_{j_1j_2}^{(2)}\dot{\varphi}_{j_1j_3}^{(2)}, \tag{4.8g}$$

where the argument of $\varphi_{j_1j_2}^{(2)}$ is $\chi_{j_1}(r_1)-\chi_{j_2}(r_2)$ and likewise the argument of $\dot{\varphi}_{j_1j_3}^{(2)}$ is $\chi_{j_1}(r_1)-\chi_{j_3}(r_3)$.

Let us now make the choice

$$f_{jk}^{(2)}(x)=x^{g_{jk}},\quad \dot{\varphi}_{jk}^{(2)}(x)=g_{jk}/x,\quad \ddot{\varphi}_{jk}^{(2)}(x)=-g_{jk}/x^2, \tag{4.9a}$$

where the $N(N-1)/2$ constants $g_{jk}=g_{kj}\geqslant 0$ are non-negative but otherwise arbitrary. Then

$$u_{jk}^{(2)}(r_1,r_2)=[(m_j+m_k)/(4m_jm_k)][g_{jk}(g_{jk}-1)[\chi_j(r_1)-\chi_k(r_2)]^{-2}\{[\chi_j{}'(r_1)]^2+[\chi_k{}'(r_2)]^2\}$$
$$+g_{jk}[\chi_j(r_1)-\chi_k(r_2)]^{-1}\{\chi_j''(r_1)-\chi_k''(r_2)+(D-1)[\chi_j{}'(r_1)/r_1-\chi_k{}'(r_2)/r_2]\}]$$
$$-\alpha g_{jk}[\chi_j(r_1)-\chi_k(r_2)]^{-1}[r_1\chi_j{}'(r_1)-r_2\chi_k{}'(r_2)]+E_{jk}^{(2)}, \tag{4.9b}$$

$$u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}[\chi_{j_1}{}'(r_1)]^2[\chi_{j_1}(r_1)-\chi_{j_2}(r_2)]^{-1}[\chi_{j_1}(r_1)-\chi_{j_3}(r_3)]^{-1}. \tag{4.9c}$$

Let us finally make two simple choices for the functions $\chi_j(r)$, namely

$$\chi_j(r)=\exp(-ar) \tag{4.10a}$$

or

$$\chi_j(r)=r^b \tag{4.11a}$$

[we leave as an exercise for the diligent reader to explore the more general choice $\chi_j(r)=c_jr^{b_j}\exp(-a_jr)$ involving $3N$ arbitrary constants]. The first choice, (4.10a), yields

$$u_{jk}^{(2)}(r_1,r_2)=g_{jk}\alpha\{[(m_j+m_k)/(8m_jm_k)]\{(g_{jk}-1)a\cosh[2a(r_1-r_2)]\sinh^2[a(r_1-r_2)]$$
$$-2(D-1)\{r_2-r_1\exp[a(r_1-r_2)]\}/[r_1r_2\{1-\exp[a(r_1-r_2)]\}]\}$$
$$+\alpha\{r_1-r_2\exp[a(r_1-r_2)]\}/\{1-\exp[a(r_1-r_2)]\}\}, \tag{4.10b}$$

$$E_{jk}^{(2)}=-[(m_j+m_k)/(4m_jm_k)]g_{jk}a^2, \tag{4.10c}$$

$$u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}\{1-\exp[a(r_1-r_2)]\}^{-1}\{1-\exp[a(r_1-r_3)]\}^{-1}. \tag{4.10d}$$

These potentials, which must be complemented by (4.7c), of course lack translation invariance, and they are singular whenever two particles are at the same distance from the origin (the two-body potential is however nonsingular in the special case $g_{jk}=1$). It is easily seen that the conditions $g_{jk}\geqslant 0$, $\alpha>0$, $a\geqslant 0$ are *sufficient* to guarantee normalizability of the (ground-state) eigenfunction $\Psi(\vec{r}_1,\ldots,\vec{r}_N)$, given by (3.4) with $\psi=1$, $F^{(n)}=1$ for all values of n except $n=2$ and $n=N$, $F^{(N)}$ given by (3.17a) with (3.16b) and $F^{(2)}$ given by (3.4d) with (4.8a), (4.9a), and (4.10a). (*Necessary and sufficient* conditions can also be easily obtained; this is left as an easy exercise for the diligent reader.) The corresponding (ground-state) energy is given by (3.18b) with $E_j^{(1)}=\alpha D$, $E_{j_1j_2j_3}^{(3)}=0$ [see (4.7a)] and $E_{jk}^{(2)}$ given by (4.10c).

The second choice, (4.11a), yields

$$u_{jk}^{(2)}(r_1,r_2)=g_{jk}b[(m_j+m_k)/(4m_jm_k)]\{(g_{jk}-1)b[r_1^{2(b-1)}+r_2^{2(b-1)}]/(r_1^b-r_2^b)^2$$
$$+(b+D-2)[r_1^{b-2}-r_2^{b-2}]/(r_1^b-r_2^b)\}, \tag{4.11b}$$

$$E_{jk}^{(2)}=\alpha bg_{jk}, \tag{4.11c}$$

$$u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}r_1^{2(b-1)}/[(r_1^b-r_2^b)(r_1^b-r_3^b)]. \tag{4.11d}$$

Likewise these potentials, which must again be complemented by (4.7c), lack translation invariance and are singular whenever two particles are at the same distance from the origin; again, the two-body potential becomes nonsingular (except at the origin) in the special case $g_{jk}=1$. And it is again easily seen that the conditions $g_{jk}\geqslant 0$, $\alpha>0$, $b\geqslant 0$ are *sufficient* to guarantee normalizability of the (ground-state) eigenfunction $\Psi(\vec{r}_1,\ldots,\vec{r}_N)$, given by (3.4) with $\psi=1$, $F^{(n)}=1$ for all values of n except $n=2$ and $n=N$, $F^{(N)}$ given by (3.17a) with (3.16b) and $F^{(2)}$ given by (3.4d) with

(4.8a), (4.9a) and (4.11a). (*Necessary and sufficient* conditions can also be easily obtained; this is left as an easy exercise for the diligent reader.) The corresponding (ground-state) energy is given by (3.18b) with $E_j^{(1)}=\alpha D$, $E_{j_1j_2j_3}^{(3)}=0$ [see (4.7a)] and $E_{jk}^{(2)}$ given by (4.11c).

The special case of equal particles, $m_j=m$, $g_{jk}=g$, is particularly interesting, but it is not new; it was introduced and discussed by Ghosh,[4] who moreover made the very interesting observation that the three-body contributions cancel out neatly for $b=1$ or $b=2$ [in these two cases $U^{(3)}(\vec{r}_1,\ldots,\vec{r}_N)$, as given by (3.2d) with (4.11d), vanishes]. Hence in these special cases the Hamiltonian (3.1) only features one-body and two-body interactions, see (4.7c) (with $m_j=m$) and (4.11b) (with $m_j=m$, $g_{jk}=g$).

Let us finally emphasize that the eigenstate constructed herein is again the ground state. It would be easy to also construct a sequence of other eigenstates for this ''Ghosh model,'' see (4.11), by using the more general *ansatz* (3.16) rather than the special choice (3.17); but since this has already been done by Ghosh,[4] we do not discuss this development here (the diligent reader may easily extend the Ghosh result to the case of unequal particles treated here). It would moreover be easy, in this case (4.11a), to introduce more general N-body potentials depending on the global radius ρ, see (3.16b), in analogy with the treatment given at the end of the preceding Sec. IV A [the main change would be the replacement in (4.6b) of G with bG, see (4.6f)]. But such an extension of the Ghosh treatment is not new; it has been given, in the equal particles case ($m_j=m$, $g_{jk}=g$), and for the two explicitly solvable potentials (centrifugal plus harmonic or plus Coulomb) by Khare.[8] Further extensions to the case of unequal particles, and of more general potentials—for which an analytic solution would however be generally available only for one discrete eigenvalue—are left for the diligent reader.

C. Models of type C

Let us set $\alpha=0$ [entailing $F^{(N)}=1$, see (3.17); the choice $\alpha\neq 0$ yields no additional generality, see below], $E_{jk}^{(2)}=E_{j_1j_2j_3}^{(3)}=0$, and

$$f_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)=(c_{jk})^{g_{jk}} \tag{4.12a}$$

with

$$c_{jk}=c_{kj}\equiv\vec{r}_j\cdot\vec{r}_k, \tag{4.12b}$$

so that

$$[f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)]^{-1}[\Delta_1+\Delta_2]f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=g_{jk}(g_{jk}-1)(r_1^2+r_2^2)/c_{12}^2, \tag{4.12c}$$

$$\vec{\nabla}_1\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=g_{jk}\vec{r}_2/c_{12},\quad \vec{\nabla}_2\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=g_{jk}\vec{r}_1/c_{12}, \tag{4.12d}$$

$$(\vec{r}_1\cdot\vec{\nabla}_1)\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=(\vec{r}_2\cdot\vec{\nabla}_2)\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=g_{jk}. \tag{4.12e}$$

[The diligent reader is invited to explore the results that obtain by setting, more generally,

$$f_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)=f_{jk}^{(2)}(c_{jk}), \tag{4.13}$$

rather than (4.12a).]

Hence from (3.18d) we get

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=[(m_j+m_k)/(4m_jm_k)]\{g_{jk}(g_{jk}-1)(r_1^2+r_2^2)/c_{12}^2+2g_{jk}[\varphi_j^{(1)}(r_1)/r_1+\varphi_k^{(1)}(r_2)/r_2]\}, \tag{4.12f}$$

and from (3.18e)

$$u^{(3)}_{j_1j_2j_3}(r_1,r_2,r_3)=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}c_{23}/(c_{12}c_{13}). \tag{4.12g}$$

Via (3.18a) it is clear that the second term on the right-hand side of (4.12f) should actually be included in the one-body potential. Hence we conclude that the partially solvable potential is in this case given by (3.18a) with

$$u^{(1)}_j(\vec{r})=u^{(1)}_j(r)=v_j(r), \tag{4.14a}$$

$$u^{(2)}_{jk}(\vec{r}_1,\vec{r}_2)=[(m_j+m_k)/(4m_jm_k)]g_{jk}(g_{jk}-1)(r_1^2+r_2^2)/(\vec{r}_1\cdot\vec{r}_2)^2, \tag{4.14b}$$

$$u^{(3)}_{j_1j_2j_3}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=(m_{j_1})^{-1}g_{j_1j_2}g_{j_1j_3}(\vec{r}_2\cdot\vec{r}_3)/[(\vec{r}_1\cdot\vec{r}_2)(\vec{r}_1\cdot\vec{r}_3)]. \tag{4.14c}$$

In these formulas $v_j(r)$ are arbitrary (rotation-invariant) one-body potentials, and the $N(N-1)/2$ constants g_{jk} are arbitrary non-negative numbers. The eigenvalues are given by (3.18b) with $E^{(2)}_{jk}=E^{(3)}_{j_1j_2j_3}=0$, and the corresponding eigenfunctions $\Psi(\vec{r}_1,\ldots,\vec{r}_N)$ are given by (3.4), with $\psi=1$, $F^{(n)}=1$ for $n>2$ and $f^{(2)}_{jk}(\vec{r}_j,\vec{r}_k)$ given by (4.12a) and (4.12b). The ''one-body eigenvalues and eigenfunctions'' $E^{(1)}_j$ and $f^{(1)}_j(r)$ are the solutions of the one-body problem that obtains from (4.14a), (4.12f), (3.18a), and (3.18c):

$$-(2m_j)^{-1}[f^{(1)\prime\prime}_j(r)+(D-1+G_j)f^{(1)\prime}_j(r)/r]+v_j(r)f^{(1)}_j(r)=E^{(1)}_jf^{(1)}_j(r), \tag{4.15a}$$

with

$$G_j=\sum_{k=1;k\neq j}^{N}[(m_j+m_k)/(2m_k)]g_{jk}, \tag{4.15b}$$

or equivalently, via $f^{(1)}_j(r)=r^{(1-D-G_j)/2}\psi_j(r)$,

$$-(2m_j)^{-1}\{\psi_j''(r)-[(D+G_j-1)(D+G_j-3)/4]r^{-2}\psi_j(r)\}+v_j(r)\psi_j(r)=E^{(1)}_j\psi_j(r). \tag{4.15c}$$

For given one-body potentials $v_j(r)$, these one-body Schrödinger equations possess generally an infinity of eigenvalues and correspondingly of eigenfunctions, parametrized by quantum numbers μ_j, that take discrete values, say $\mu_j=0,1,2,...$, if the potential $v_j(r)$ is confining. Hence we see that the partially solvable potential we have manufactured, (3.18a) with (4.14), has a known set of eigenvalues and eigenfunctions [assuming the one-dimensional Schrödinger equations (4.15) are solvable] which is parametrized by N quantum numbers; while the *complete* spectrum, and the corresponding *complete* set of eigenfunctions, is generally parametrized by ND quantum numbers. [The fact that one gets *all* eigenvalues and eigenstates for $D=1$ is trivial: it is clear from (4.14b) and (4.14c) that for $D=1$ the two-body and three-body potentials reduce to one-body potentials, hence the N-body problem separates into N decoupled one-body problems.]

For instance for

$$v_j(r)=2m_j\alpha_j^2r^2, \tag{4.16a}$$

the eigenvalues of (4.15) are

$$E^{(1)}_j=\alpha_j(D+G_j+2\mu_j), \tag{4.16b}$$

with $\mu_j=0,1,2,...$, and the corresponding eigenfunctions $\psi_j(r)$ are the well-known harmonic oscillator wave functions, see (4.15c) with (4.16a). Of course if some of the (non-negative) quantities α_j coincide, or are integer multiples of each other, the spectrum of the many-body problem, see (3.3) and (3.18b) with $E^{(2)}_{jk}=E^{(3)}_{j_1j_2j_3}=0$, is degenerate.

There is a variant of the approach of this Sec. IV C which is worth mentioning. It is obtained by setting

$$f_j^{(1)}(r)=1, \quad \varphi_j^{(1)}(r)=0, \quad E_j^{(1)}=0, \tag{4.17a}$$

namely eliminating the one-body contributions, and using [instead of (3.17) with $\alpha=0$, see above], the ''N-body'' *ansatz* (3.16). The corresponding partially solvable potential then reads

$$U(\vec{r}_1,\ldots,\vec{r}_N)=\alpha^2\sum_{j=1}^{N} 2m_j r_j^2+\sum_{j,k=1;j<k}^{N} u_{jk}^{(2)}(\vec{r}_j,\vec{r}_k) + \sum_{j_1,j_2,j_3=1;j_1\neq j_2,j_1\neq j_3,j_2<j_3}^{N} u_{j_1j_2j_3}^{(3)}(\vec{r}_{j_1},\vec{r}_{j_2},\vec{r}_{j_2})+u^{(N)}(\rho), \tag{4.17b}$$

with $u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)$, respectively, $u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)$ given by (4.14b), respectively, (4.14c), ρ defined by (3.16b), and $u^{(N)}(\rho)$ a (largely arbitrary) ''N-body potential.'' The corresponding eigenfunction [see (3.3)] reads

$$\Psi_\mu(\vec{r}_1,\ldots,\vec{r}_N)=\chi_\mu(\rho)\prod_{j,k=1;j<k}^{N}(\vec{r}_j\cdot\vec{r}_k)^{g_{jk}}, \tag{4.17c}$$

and the corresponding equations for the energy eigenvalue E, see (3.3), and for the functions $\varphi_\mu(\rho)=\log[\Phi(\rho)]$, are exactly the same as in Sec. IV A, see (4.6c)–(4.6f). Hence in this variant, as in the analogous model treated at the end of Sec. IV A, the set of eigenvalues and eigenfunctions is parametrized by only one quantum number μ [rather than by N of them, see for instance (4.16b)]; moreover, only for the choice $\alpha=0$, $u^{(N)}(\rho)=-q^2/\rho$ is it possible to obtain explicitly *all* the eigenvalues and eigenfunctions of (4.6d) and (4.6e) [the other solvable case, $\alpha\neq 0$, $u^{(N)}(\rho)=0$, is of no interest here, since it is a special case of the ''less partially'' solvable case discussed above, see (4.14) with (4.16a)].

The models considered in this Sec. IV C are, to the best of our knowledge, new.

D. Models of type D

In this section we restrict attention to *two-dimensional* space, $D=2$, and we set, as above,

$$c_{jk}=c_{kj}\equiv\vec{r}_j\cdot\vec{r}_k\equiv x_jx_k+y_jy_k, \tag{4.18a}$$

and, in addition,

$$s_{jk}=-s_{kj}\equiv\hat{z}\cdot\vec{r}_j\wedge\vec{r}_k\equiv x_jy_k-y_jx_k. \tag{4.18b}$$

The notation $\hat{z}\cdot\vec{r}_j\wedge\vec{r}_k$ is formally *three-dimensional*, with the unit vector $\hat{z}\equiv(0,0,1)$ orthogonal to the plane in which all the vectors $\vec{r}_j\equiv(x_j,y_j,0)$ lie, and the symbol $\wedge$ denoting the standard *three-dimensional* vector product; it entails $\hat{z}\wedge\vec{r}_j\equiv(-y_j,x_j,0)$ and it underlines the rotation-invariant character of the (pseudoscalar) quantity s_{jk}, see (4.18b).

We then set $\alpha=0$ (as in the preceding Sec. IV C, the choice of an arbitrary value for α entails no increase in generality, see below), and, as a natural generalization of (4.12a),

$$f_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)=(c_{jk})^{g_{jk}}|s_{jk}|^{\tilde{g}_{jk}}, \tag{4.18c}$$

so that

$$[f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)]^{-1}[\Delta_1+\Delta_2]f_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=(r_1^2+r_2^2)[g_{jk}(g_{jk}-1)/c_{12}^2+\tilde{g}_{jk}(\tilde{g}_{jk}-1)/s_{12}^2], \tag{4.18d}$$

$$\vec{\nabla}_1\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=g_{jk}\vec{r}_2/c_{12}-\tilde{g}_{jk}(\hat{z}\wedge\vec{r}_2)/s_{12}, \quad \vec{\nabla}_2\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=g_{jk}\vec{r}_1/c_{12}+\tilde{g}_{jk}(\hat{z}\wedge\vec{r}_1)/s_{12}, \tag{4.18e}$$

$$(\vec{r}_1\cdot\vec{\nabla}_1)\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=(\vec{r}_2\cdot\vec{\nabla}_2)\varphi_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=g_{jk}+\tilde{g}_{jk}. \tag{4.18f}$$

[Again, as in the preceding Sec. IV C, the diligent reader is invited to explore the results that obtain by setting, more generally,

$$f_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)=f_{jk}^{(2)}(c_{jk},s_{jk}), \tag{4.19}$$

rather than (4.18c).]

Hence from (3.18d) we get

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=[(m_j+m_k)/(4m_jm_k)]\{(r_1^2+r_2^2)[g_{jk}(g_{jk}-1)/c_{12}^2+\tilde{g}_{jk}(\tilde{g}_{jk}-1)/s_{12}^2] \\ +2(g_{jk}+\tilde{g}_{jk})[\varphi_j^{(1)}(r_1)/r_1+\varphi_k^{(1)}(r_2)/r_2]\}, \tag{4.20a}$$

and from (3.18e) we get

$$u_{j_1j_2j_3}^{(3)}(r_1,r_2,r_3)=(m_{j_1})^{-1}[g_{j_1j_2}g_{j_1j_3}c_{23}/(c_{12}c_{13})+\tilde{g}_{j_1j_2}\tilde{g}_{j_1j_3}c_{23}/(s_{12}s_{13}) \\ +g_{j_1j_2}\tilde{g}_{j_1j_3}s_{23}/(c_{12}s_{13})-\tilde{g}_{j_1j_2}g_{j_1j_3}s_{23}/(s_{12}c_{13})]. \tag{4.20b}$$

As in the preceding Sec. IV C, via (3.18a) it is again clear that the second term on the right-hand side of (4.20a) should be considered part of the one-body potential. Hence we conclude that the partially solvable potential is now given by (3.18a) with

$$u_j^{(1)}(\vec{r})=u_j^{(1)}(r)=v_j(r), \tag{4.21a}$$

$$u_{jk}^{(2)}(\vec{r}_1,\vec{r}_2)=[(m_j+m_k)/(4m_jm_k)](r_1^2+r_2^2)[g_{jk}(g_{jk}-1)/(\vec{r}_1\cdot\vec{r}_2)^2+\tilde{g}_{jk}(\tilde{g}_{jk}-1)/(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_2)^2], \tag{4.21b}$$

$$\begin{aligned} u_{j_1j_2j_3}^{(3)}(\vec{r}_1,\vec{r}_2,\vec{r}_3)=(m_{j_1})^{-1}[(\vec{r}_2\cdot\vec{r}_3)\{&g_{j_1j_2}g_{j_1j_3}/[(\vec{r}_1\cdot\vec{r}_2)(\vec{r}_1\cdot\vec{r}_3)] \\ &+\tilde{g}_{j_1j_2}\tilde{g}_{j_1j_3}/[(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_2)(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_3)]\} \\ &+(\hat{z}\cdot\vec{r}_2\wedge\vec{r}_3)\{g_{j_1j_2}\tilde{g}_{j_1j_3}/[(\vec{r}_1\cdot\vec{r}_2)(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_3)] \\ &-\tilde{g}_{j_1j_2}g_{j_1j_3}/[(\hat{z}\cdot\vec{r}_1\wedge\vec{r}_2)(\vec{r}_1\cdot\vec{r}_3)]\}]. \end{aligned} \tag{4.21c}$$

In these formulas the N functions $v_j(r)$ are arbitrary (rotation-invariant) one-body potentials, and the $N(N-1)$ constants g_{jk} and $\tilde{g}_{jk}$ are arbitrary nonnegative numbers. The eigenvalues are given by (3.18b) with $E_{jk}^{(2)}=E_{j_1j_2j_3}^{(3)}=0$, and the corresponding eigenfunctions $\Psi(\vec{r}_1,\ldots,\vec{r}_N)$ are given by (3.4), with $\psi=1$, $F^{(n)}=1$ for $n>2$ and $f_{jk}^{(2)}(\vec{r}_j,\vec{r}_k)$ given by (4.18a)–(4.18c). The ''one-body eigenvalues and eigenfunctions'' $E_j^{(1)}$ and $f_j^{(1)}(r)$ are the solutions of the one-body problem that obtains from (4.21a), (4.20a), and (3.18c):

$$-(2m_j)^{-1}[f_j^{(1)\prime\prime}(r)+(1+G_j)f_j^{(1)\prime}(r)/r]+v_j(r)f_j^{(1)}(r)=E_j^{(1)}f_j^{(1)}(r), \tag{4.22a}$$

with

$$G_j=\sum_{k=1;k\neq j}^{N}[(m_j+m_k)/(2m_k)](g_{jk}+\tilde{g}_{jk}), \tag{4.22b}$$

or equivalently, via $f_j^{(1)}(r)=r^{-(1+G_j)/2}\psi_j(r)$,

$$-(2m_j)^{-1}\{\psi_j''(r)-[(G_j^2-1)/4]r^{-2}\psi_j(r)\}+v_j(r)\psi_j(r)=E_j^{(1)}\psi_j(r). \tag{4.22c}$$

These last developments have followed closely the analogous treatment given in the preceding Sec. IV C. Also applicable, practically *verbatim*, are the subsequent observations given there, except that one must now set $D=2$. It is thus seen that for the potential we have now manufactured, see (4.21), *partial solvability* generally entails the possibility to exhibit eigenvalues and eigenfunctions depending on N quantum numbers, while *complete solvability* would require knowledge of eigenvalues and eigenfunctions depending on $2N$ quantum numbers. In particular, for the one-body potential (4.16a), the treatment of the preceding Sec. IV C remains applicable [including (4.16b)], up to trivial modifications.

The special case with equal particles, $m_j=m$, $\alpha_j=\alpha$ [see (4.16a)], $\tilde{g}_{jk}=g$, and moreover $g_{jk}=0$, is not new: the partially solvable potential corresponding to this case has been introduced and discussed by Murthy, Bhaduri, and Sen.[1,5] The set of eigenvalues and eigenfunctions exhibited in these papers is however, for large N, much smaller than the set exhibited herein: it depends on 3, rather than N, quantum numbers.

There also is, in close analogy to the elaboration reported at the end of the preceding Sec. IV C, the variant of this treatment that is obtained by adopting (4.17a) and using the *ansatz* (3.16) instead of (3.17). We leave the derivation of the corresponding formulas to the diligent reader: the changes relative to the treatment given at the end of Sec. IV C entail essentially the replacement of (4.14) with (4.21), the insertion of an additional factor $(s_{jk})^{\tilde{g}_{jk}}$ in (4.17c), and the replacement of g_{jk} with $g_{jk}+\tilde{g}_{jk}$ in the definition (4.6f) of G (and of course the specialization $D=2$). But the results so obtained are not quite new: they have been previously given by Khare,[8] in the case of equal particles ($m_j=m$, $\tilde{g}_{jk}=g$) with moreover $g_{jk}=0$, and with special attention focused on the completely solvable ''N-body'' Coulomb and centrifugal potentials.

V. FINAL REMARKS AND OUTLOOK

Clearly the technique described in this paper yields easily a much larger gamut of partially solvable quantum many-body problems than the examples presented above. In selecting these examples, we have been mainly motivated by ''aesthetic'' considerations, and/or the intent to demonstrate how to cover in a unified manner several cases recently treated in the literature.[1–8] Applications have not been discussed: this is of course the most interesting direction for future developments.

We did not treat ''few-body'' problems, such as those characterized by a relatively small number of particles, say $N=3$ or $N=4$. For some special variants, applicable to such problems, of the technique used herein the interested reader is referred to the original paper where this technique was introduced.[9]

Another technique to manufacture completely or partially solvable quantum many-body problems, so much investigated in the literature that we refrain from mentioning any reference here since any appropriate list should be too long, relies on the introduction of (collective) coordinates such that the many-body Schrödinger equation becomes *separable*. These models tend to be farther from physics, because they generally give rise to many-body forces that do not seem to have any counterpart in nature, or break important invariance properties such as physical (i.e., three-dimensional) rotation invariance. We have, by and large, avoided to present models of this type, except to the extent we used the overall radial coordinate ρ, see (3.16b), mainly to make contact with the recent literature.

Let us finally emphasize that several of the examples presented above are characterized by potentials which are homogeneous functions of degree -2 in the particle coordinates [see for instance (2.1) with $\beta_{jk}=0$, (2.3), as well as (2.4) and (2.5) with an appropriate choice—''centrifugal''—of the one-body potentials $v_j(r)$]. Obviously for such potentials the introduction of an overall radial coordinate such as ρ, see (3.16b), allows separation of the problem into a (generally easily solvable) ''radial'' part involving a single coordinate and a (generally much more difficult) ''angular'' part, for which sometimes one very simple (generally, just constant) solution can be found (in fact the potentials may be adjusted so that such a solution exist). Several classes of partially solvable potentials which have appeared in the literature are of this kind. In the examples provided above it has been generally possible to go somewhat beyond this simple approach.

[1] M. V. N. Murthy, R. K. Bhaduri, and Diptiman Sen, ''Novel correlations in two dimensions: some exact solutions,'' Phys. Rev. Lett. **76**, 4103–4106 (1996).
[2] N. Gurappa, C. Nagaraja Kumar, and Prasanta K. Panigrahi, ''New exactly and conditionally exactly solvable *N*-body problems on one dimension,'' Mod. Phys. Lett. A **11**, 1737–1744 (1996).
[3] A. Khare, ''A class of exact solutions for *N* anyons in a *N*-body potential,'' Phys. Lett. A **221**, 365–369 (1996).
[4] P. K. Ghosh, ''Calogero-Sutherland type models in higher dimensions,'' Phys. Lett. A **229**, 203–207 (1997). This very interesting paper is marred by two misprints: in Eq. (10), $g(g-1)$ should read g^2, and in Eq. (11) 4 should be replaced by 2 and r_j by $r_j - r_k$.
[5] R. K. Bhaduri, A. Khare, J. Law, M. V. N. Murthy, and D. Sen, ''Novel correlations in two dimensions: two-body problem,'' J. Phys. A **30**, 2557–2576 (1997).
[6] A. Khare and K. Ray, ''A quantum many-body problem in two dimensions: ground state,'' Phys. Lett. A **230**, 139–143 (1997).
[7] A. Khare, ''Exact solution of an *N*-body problem in one dimension,'' J. Phys. A **29**, L45–L49 (1996); F. Calogero, ''Exact solution of an *N*-body problem in one dimension: two comments,'' *ibid.* **29**, 6455–6457 (1996); A. Khare, ''Reply to comment 'Exact solution of an *N*-body problem in one dimension','' *ibid.* **29**, 6459–6460 (1996).
[8] A. Khare, ''Exact ground state of several *N*-body problems with an *N*-body potential,'' J. Math. Phys. **40**, 2640–2646 (1999).
[9] F. Calogero and C. Marchioro, ''Exact bound states of some *N*-body systems with two- and three-body forces,'' J. Math. Phys. **14**, 182–184 (1973).
[10] F. Calogero and C. Marchioro, ''Exact ground state of some one-dimensional *N*-body systems with inverse (''Coulomb-like'') and inverse-square (''centrifugal'') pair potentials,'' Phys. Rev. Lett. **27**, 86–88 (1971).
[11] F. Calogero, ''Solution of the one-dimensional *N*-body problem with quadratic and/or inversely quadratic pair potentials,'' J. Math. Phys. **12**, 419–436 (1971); *ibid.* **37**, 3646(E) (1996).

An estimation theoretical characterization of coherent states

Akio Fujiwara[a)]
Department of Mathematics, Osaka University, 1-16 Machikane-yama, Toyonaka, Osaka 560-0043, Japan

Hiroshi Nagaoka[b)]
Graduate School of Information Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

(Received 23 May 1997; accepted for publication 2 June 1999)

We introduce a class of quantum pure state models called the coherent models. A coherent model is an even-dimensional manifold of pure states whose tangent space is characterized by a symplectic structure. In a rigorous framework of noncommutative statistics, it is shown that a coherent model inherits and expands the original spirit of the minimum uncertainty property of coherent states. © *1999 American Institute of Physics.* [S0022-2488(99)02509-8]

I. INTRODUCTION

Quantum estimation theory, originated in optical communications, offers a rigorous approach toward the optimization of detection processes in quantum communication systems.[1,2] It aims to find, for a given smooth parametric family of density operators (a model) $\mathcal{P}=\{\rho_\theta;\theta=(\theta^1,\ldots,\theta^n)\in\Theta\subset\mathbf{R}^n\}$, the optimum measurement (positive operator-valued measure) $M=\{M(B);B$ is a Borel set in $\mathbf{R}^n\}$ for the parameter θ under the unbiasedness condition: For all $\theta\in\Theta$,

$$\int \hat{\theta}^j \,\mathrm{Tr}\,\rho_\theta M(d\hat{\theta})=\theta^j, \quad j=1,...,n.$$

Here Tr denotes the operator trace. Normally a more tractable (weaker) condition is adopted, called the local unbiasedness condition: A measurement M is called locally unbiased at a given point θ if M satisfies at θ the above equality and its formal differentiation,

$$\frac{\partial}{\partial\theta^i}\int \hat{\theta}^j \,\mathrm{Tr}\,\rho_\theta M(d\hat{\theta})=\delta_i^j, \quad i,j=1,...,n.$$

It is well known that when $n=1$, the quantum Cramér–Rao inequality with respect to the symmetric logarithmic derivative (SLD) offers the achievable lower bound (i.e., the bound attained by a certain measurement) of the variance of estimation. This is also regarded as a rigorous modification of the uncertainty relation. When $n\geqslant 2$, on the other hand, a matrix version of the SLD Cramér–Rao inequality itself does not always have an absolute significance because the lower bound cannot be attained, in general, unless the model has commutative SLDs. We therefore often deal with the minimization problem of the scalar quantity $\mathrm{tr}\,GV_\theta[M]$ with respect to M, where tr denotes the matrix trace on the parameter space Θ, G a real symmetric positive matrix representing the weight, and $V_\theta[M]$ the covariance matrix at θ with respect to a locally unbiased measurement M whose $(i,\,j)$ entry is

[a)]Electronic mail: fujiwara@math.wani.osaka-u.ac.jp
[b)]Electronic mail: nagaoka@is.uec.ac.jp

$$(V_\theta[M])^{ij}=\int(\hat\theta^i-\theta^i)(\hat\theta^j-\theta^j)\mathrm{Tr}\,\rho_\theta M(d\hat\theta).$$

If there is a number C such that $\mathrm{tr}\,GV_\theta[M]\geqslant C$ holds for all M, C is called a Cramér–Rao-type bound or simply a CR bound. The CR bound C may depend on both G and θ. The problem of finding the achievable CR bound is, in general, a hard one and has been solved only in a few special models such as the quantum Gaussian model[3,2] and the two-dimensional spin-$\frac{1}{2}$ model.[4,5]

Holevo showed that if a model having the right logarithmic derivative (RLD) exhibits a certain ''nice'' property of a tangent space, the CR bound based on the RLD is expressed only in terms of the SLDs (Ref. 2, p. 280). Moreover it was shown that this gives the achievable CR bound for the Gaussian model of quantum oscillators. Motivated by these facts and that the SLD Fisher information is well defined also for pure state models,[6] we will introduce a class of pure state models called the coherent models,[7] each having a ''nice'' tangent space, and will explore their parameter estimation theory.

The construction of the paper is as follows. In Sec. II, we explore some basic characteristics inherent in pure state models that are closely related with Holevo's commutation operator. In Sec. III, a special class of pure state models, called the coherent models, is introduced of which the SLD tangent space forms an invariant subspace with respect to the commutation operator. In Sec. IV, we derive a CR bound, called the generalized RLD bound, for a model that has an invariant SLD tangent space with respect to the commutation operator. Here the model is not assumed to be pure. In Sec. V, we show that for a coherent model, there exists a random measurement that attains the generalized RLD bound. In Sec. VI, the above results are demonstrated in two simple coherent models: a canonical squeezed state model and a spin coherent state model. In the final section we give conclusions.

II. COMMUTATION OPERATOR

In the study of noncommutative statistics, Holevo introduced useful mathematical tools called the square summable operators and the commutation operators associated with quantum states. We here give a brief summary: for details, consult Ref. 2. Let $\mathcal{H}$ be a separable complex Hilbert space that corresponds to a physical system of interest, and let ρ be a fixed density operator. We define a real Hilbert space $\mathcal{L}_h^2(\rho)$ associated with ρ by the completion of $\mathcal{B}_h(\mathcal{H})$, the set of bounded self-adjoint operators, with respect to the pre-inner product $\langle X,Y\rangle_\rho=\mathrm{Re}\,\mathrm{Tr}\,\rho XY$. Letting $\rho=\sum_j s_j|\psi_j\rangle\langle\psi_j|$ be the spectral representation, an element $X\in\mathcal{L}_h^2(\rho)$ can be regarded as an equivalence class of such self-adjoint operators (called square summable operators) satisfying $\sum_j s_j\|X\psi_j\|^2<\infty$ [so that $\psi_j\in\mathrm{Dom}(X)$ if $s_j\neq 0$] under the identification $X_1\sim X_2$ if $X_1\psi_j=X_2\psi_j$ for $s_j\neq 0$. The space $\mathcal{L}_h^2(\rho)$ thus provides a convenient tool to cope with unbounded observables. Let $\mathcal{L}^2(\rho)$ be the complexification of $\mathcal{L}_h^2(\rho)$. Note that $\mathcal{L}^2(\rho)$ is also regarded as the completion of $\mathcal{B}(\mathcal{H})$, the set of bounded operators, with respect to the pre-inner product,

$$\langle X,Y\rangle_\rho=\tfrac{1}{2}\mathrm{Tr}\,\rho(YX^*+X^*Y).$$

Thus $\mathcal{L}^2(\rho)$ is regarded as a complex Hilbert space with the inner product $\langle\cdot,\cdot\rangle_\rho$. We further introduce two sesquilinear forms on $\mathcal{B}(\mathcal{H})$ by

$$(X,Y)_\rho=\mathrm{Tr}\,\rho YX^*,\quad [X,Y]_\rho=\frac{1}{2i}\mathrm{Tr}\,\rho(YX^*-X^*Y),$$

and extend them to $\mathcal{L}^2(\rho)$ by continuity.

The *commutation operator* $\mathcal{D}_\rho:\mathcal{L}^2(\rho)\to\mathcal{L}^2(\rho)$ with respect to ρ is defined by $[X,Y]_\rho=\langle X,\mathcal{D}_\rho Y\rangle_\rho$, which is formally represented by the operator equation $\rho(\mathcal{D}_\rho X)+(\mathcal{D}_\rho X)\rho=(1/i)\times(\rho X-X\rho)$. (To be precise, this definition is different from Holevo's original definition by a factor of 2.) The operator $\mathcal{D}_\rho$ is a complex-linear bounded skew-adjoint operator. Moreover, since the forms $[\cdot,\cdot]_\rho$ and $\langle\cdot,\cdot\rangle_\rho$ are real on the real subspace $\mathcal{L}_h^2(\rho)$, this subspace is invariant under the operation of $\mathcal{D}_\rho$. Thus, $\mathcal{D}_\rho$ can also be regarded as a real-linear bounded skew-adjoint operator when restricted to $\mathcal{L}_h^2(\rho)$ as $\mathcal{D}_\rho:\mathcal{L}_h^2(\rho)\to\mathcal{L}_h^2(\rho)$.

Our main concern lies in the case where ρ is pure. In this case the above setting is considerably simplified as follows: Let $\rho=|\psi\rangle\langle\psi|$. Then, for $X,Y\in\mathcal{L}^2(\rho)$,

$$\langle X,Y\rangle_\rho=\tfrac{1}{2}\{\langle Y^*\psi|X^*\psi\rangle+\langle X\psi|Y\psi\rangle\},$$

$$[X,Y]_\rho=\frac{1}{2i}\{\langle Y^*\psi|X^*\psi\rangle-\langle X\psi|Y\psi\rangle\},$$

$$(X,Y)_\rho=\langle Y^*\psi|X^*\psi\rangle.$$

Here $X\psi$, for example, stands for the vector $X_1\psi$, where X_1 is an arbitrary representative of X. (It is independent of the choice of a representative.) In particular, if $X,Y\in\mathcal{L}_h^2(\rho)$, we have

$$\langle X,Y\rangle_\rho=\mathrm{Re}\langle Y\psi|X\psi\rangle=\mathrm{Re}\langle X\psi|Y\psi\rangle, \tag{1}$$

$$[X,Y]_\rho=\mathrm{Im}\langle Y\psi|X\psi\rangle=-\mathrm{Im}\langle X\psi|Y\psi\rangle, \tag{2}$$

$$(X,Y)_\rho=\langle Y\psi|X\psi\rangle=\overline{\langle X\psi|Y\psi\rangle}. \tag{3}$$

It should be noted that operators X and Y (whether bounded or not) are identified with each other in $\mathcal{L}^2(\rho)$ iff $X\psi=Y\psi$ and $X^*\psi=Y^*\psi$. In particular, self-adjoint operators X and Y are identified in $\mathcal{L}_h^2(\rho)$ iff $X\psi=Y\psi$.

Lemma 1: Let $\rho=|\psi\rangle\langle\psi|$. *Then for all* $X\in\mathcal{L}_h^2(\rho)$,

$$(\mathcal{D}_\rho X)\psi=i(X-\langle\psi|X\psi\rangle I)\psi,$$

where I denotes the identity in $\mathcal{L}_h^2(\rho)$.

Proof: For $X\in\mathcal{L}_h^2(\rho)$, let Z be the element in $\mathcal{L}_h^2(\rho)$ having a representative $Z_1=i(|X\psi\rangle\times\langle\psi|-|\psi\rangle\langle X\psi|)$. Then $Z\psi=i(X-\langle\psi|X\psi\rangle I)\psi$. On the other hand, for $Y\in\mathcal{L}_h^2(\rho)$, we have

$$\langle Y\psi|Z\psi\rangle=i\{\langle Y\psi|X\psi\rangle-\langle\psi|X\psi\rangle\langle\psi|Y\psi\rangle\},$$

and hence $\langle Y,Z\rangle_\rho=[Y,X]_\rho$ because of (1) and (2). Thus $Z=\mathcal{D}_\rho X$, which completes the proof.□

Note that Lemma 1 does not imply $\mathcal{D}_\rho X=i(X-\langle\psi|X\psi\rangle I)$, since the right-hand side is not a self-adjoint element in $\mathcal{L}^2(\rho)$ unless it equals 0.

Let us introduce a linear subspace,

$$\mathcal{T}_h(\rho)=\{X\in\mathcal{L}_h^2(\rho);\langle I,X\rangle_\rho=0\},$$

of $\mathcal{L}_h^2(\rho)$. Here ρ is not necessarily pure. This subspace is itself a real Hilbert space with the inner product $\langle\cdot,\cdot\rangle_\rho$. Now consider again the special case that ρ is pure: $\rho=|\psi\rangle\langle\psi|$. Then from Lemma 1, we obtain the important relation:

$$(\mathcal{D}_\rho X)\psi=(iX)\psi,\quad X\in\mathcal{T}_h(\rho). \tag{4}$$

This equation, combined with (1), implies that $\mathcal{D}_\rho$ is a unitary transformation on $(\mathcal{T}_h(\rho),\langle\cdot,\cdot\rangle_\rho)$. In particular, $\mathcal{D}_\rho$ is nondegenerate on $\mathcal{T}_h(\rho)$, and so is the skew-symmetric bilinear form $[\cdot,\cdot]_\rho$. In other words, the real linear space $\mathcal{T}_h(\rho)$ is regarded as a symplectic space[8] with the symplectic form $[\cdot,\cdot]_\rho$. We also note that $\mathcal{D}_\rho^2=-\mathcal{I}$ on $\mathcal{T}_h(\rho)$ [$\mathcal{I}$ denotes the identity operator acting on $\mathcal{T}_h(\rho)$], since $\mathcal{D}_\rho$ is unitary and skew adjoint. Indeed, Eq. (4) immediately leads to $(\mathcal{D}_\rho^2X)\psi=-X\psi$, and hence $\mathcal{D}_\rho^2X=-X$ for all $X\in\mathcal{T}_h(\rho)$, whereas $\mathcal{D}_\rho X\neq iX$, as mentioned earlier. In other words, $\mathcal{D}_\rho$ is an almost complex structure on $\mathcal{T}_h(\rho)$.

III. COHERENT MODEL

Let $\mathcal{P}=\{\rho_\theta;\theta=(\theta^1,\ldots,\theta^n)\in\Theta\}$ be an n-dimensional model, where ρ_θ are not necessarily pure for the present, and Θ is an open subset of $\mathbf{R}^n$. We assume the following regularity conditions.

(a) The parametrization $\theta\mapsto\rho_\theta$ is assumed to be appropriately smooth and nondegenerate so that the derivatives $\{\partial\rho_\theta/\partial\theta^j\}_{j=1}^n$ exist in trace class and form a linearly independent set at each point θ.

(b) There exists a constant c such that

$$\left|\frac{\partial}{\partial\theta^j}\mathrm{Tr}\,\rho_\theta X\right|^2\leqslant c\langle X,X\rangle_{\rho_\theta},$$

for all $X\in\mathcal{B}(\mathcal{H})$ and j.

From the condition (b), the linear functionals $X\mapsto(\partial/\partial\theta^j)\mathrm{Tr}\,\rho_\theta X$ can be extended to continuous linear functionals on $\mathcal{L}^2(\rho_\theta)$.

Given a model that satisfies (a) and (b), the *symmetric logarithmic derivative* (SLD) $L^S_{\theta,j}$ in the jth direction is defined by the requirement that

$$\frac{\partial}{\partial\theta^j}\mathrm{Tr}\,\rho_\theta X=\langle L^S_{\theta,j},X\rangle_{\rho_\theta},\quad L^S_{\theta,j}\in\mathcal{L}^2(\rho_\theta),$$

for all $X\in\mathcal{L}^2(\rho_\theta)$. It is easily verified that $L^S_{\theta,j}\in\mathcal{L}^2_h(\rho_\theta)$; so the definition is formally written as $\partial\rho_\theta/\partial\theta^j=1/2(L^S_{\theta,j}\rho_\theta+\rho_\theta L^S_{\theta,j})$. The SLDs belong to $\mathcal{T}_h(\rho_\theta)$, since $\langle I,L^S_{\theta,j}\rangle_{\rho_\theta}=(\partial/\partial\theta^j)\mathrm{Tr}\,\rho_\theta=0$, and the SLD Fisher information matrix defined by $J^S_\theta=[\langle L^S_{\theta,j},L^S_{\theta,k}\rangle_{\rho_\theta}]$ gives a Cramér–Rao inequality $V_\theta[M]\geqslant(J^S_\theta)^{-1}$, where M is an arbitrary locally unbiased measurement for the parameter θ; see Ref. 2, p. 276.

In the rest of this section, we restrict ourselves to pure state models. Some remarks are in order. First, by differentiating the identity $\rho_\theta^2=\rho_\theta$, we see that the element in $\mathcal{L}^2_h(\rho_\theta)$ having a representative $2\,\partial\rho_\theta/\partial\theta^j$ gives the SLD $L^S_{\theta,j}$. Thus, for a pure state model, the condition (a) implies (b). Second, associated with a pure state model $\{\rho_\theta;\theta\in\Theta\}$ is, at least locally, a smooth family $\{\psi_\theta;\theta\in\Theta\}$ of normalized vectors in $\mathcal{H}$ such that $\rho_\theta=|\psi_\theta\rangle\langle\psi_\theta|$. In what follows, we shall frequently use this representation.

A convenient way of finding SLDs for a pure state model ρ_θ is as follows: Let $L^A_{\theta,j}$ be the antisymmetric logarithmic derivative (ALD) satisfying

$$\frac{\partial}{\partial\theta^j}\mathrm{Tr}\,\rho_\theta X=[L^A_{\theta,j},X]_{\rho_\theta},\quad L^A_{\theta,j}\in\mathcal{T}_h(\rho_\theta),$$

for all $X\in\mathcal{L}^2(\rho_\theta)$, or formally $\partial\rho_\theta/\partial\theta^j=(L^A_{\theta,j}\rho_\theta-\rho_\theta L^A_{\theta,j})/2i$. (This definition is different from Ref. 6 by a factor of i.) Then the SLD is given by $L^S_{\theta,j}=-\mathcal{D}_\theta L^A_{\theta,j}$, where $\mathcal{D}_\theta=\mathcal{D}_{\rho_\theta}$, since $\langle L^S_{\theta,j},X\rangle_{\rho_\theta}=[L^A_{\theta,j},X]_{\rho_\theta}$. Note that since $\mathcal{D}_\theta^2=-\mathcal{I}$ on $\mathcal{T}_h(\rho_\theta)$, then $L^A_{\theta,j}=\mathcal{D}_\theta L^S_{\theta,j}$, which assures the existence and the uniqueness of the ALD for a pure state model. The advantage of the use of the ALD is this: Every pure state model can be expressed in the form $\rho_\theta=U_\theta\rho_0U_\theta^*$, where $\{U_\theta\}_\theta$ is a smooth family of unitary operators (which do not necessarily form a group representation), so that the ALD is explicitly given by

$$L^A_{\theta,j}=2i(A_{\theta,j}-\langle I,A_{\theta,j}\rangle_{\rho_\theta}),$$

where $A_{\theta,j}$ is the skew-adjoint element in $\mathcal{L}^2(\rho_\theta)$ having a representative $(\partial U_\theta/\partial\theta^j)U_\theta^*$, the local generator of U_θ. For a group covariant pure state model, the generator of the group is usually obvious.

Let $\mathcal{T}_\theta^S(\mathcal{P})$ be the real-linear subspace of $\mathcal{T}_h(\rho_\theta)$ spanned by the SLDs $\{L_{\theta,j}^S\}_j$. Since the tangent vectors of the manifold $\mathcal{P}$ at the point ρ_θ are faithfully represented by the elements of $\mathcal{T}_\theta^S(\mathcal{P})$ via the correspondence $(\partial/\partial\theta^j)_\theta \mapsto L_{\theta,j}^S$, we call $\mathcal{T}_\theta^S(\mathcal{P})$ the *SLD tangent space* of the model $\mathcal{P}$ at θ. A pure state model $\mathcal{P}=\{\rho_\theta;\theta\in\Theta\}$ is called *locally coherent* at θ if $\mathcal{T}_\theta^S(\mathcal{P})$ is $\mathcal{D}_\theta$ invariant. The model is called *coherent* if it is locally coherent for all $\theta\in\Theta$.

When the Hilbert space $\mathcal{H}$ is finite dimensional, the totality of pure states forms a complex projective space and is an example of coherent model. The Riemannian metric on the model induced by the SLD Fisher information matrix J_θ^S is identical to the Fubini–Study metric up to a constant factor[6] and hence is a Kähler metric. The associated fundamental 2-form[9] in this case is nothing but the symplectic structure $[\cdot,\cdot]_\rho$.

Theorem 2: *Consider a pure state model of the form $\rho_\theta=U_{g(\theta)}\rho_0 U_{g(\theta)}^*$, where $\{U_g;g\in\mathcal{G}\}$ is a projective unitary representation of a Lie group $\mathcal{G}$ and $g(\cdot):\theta\mapsto g(\theta)$ is the parametrization of the elements of $\mathcal{G}$ by a local coordinate system satisfying $g(0)=e$ (: the unit element). This model is coherent iff it is locally coherent at ρ_0.*

Proof: We only need to prove the if part. Let $\Lambda_\theta:\mathcal{G}\to\mathcal{G}$ be the left translation by $g(\theta)^{-1}$ that maps $h\mapsto g(\theta)^{-1}h$. Then its differential $(d\Lambda_\theta)_{g(\theta)}:T_{g(\theta)}(\mathcal{G})\to T_e(\mathcal{G})$ is represented by a nonsingular real matrix $a_j^k(\theta)$ such that $(d\Lambda_\theta)_{g(\theta)}([\partial g(\theta)/\partial\theta^j]_\theta)=\Sigma_k a_j^k(\theta)[\partial g(\theta)/\partial\theta^k]_{\theta=0}$. Now since $\rho_{\theta+\Delta\theta}=U_{g(\theta)}\rho_{\Delta\theta'}U_{g(\theta)}^*$, where $\Lambda_\theta(g(\theta+\Delta\theta))=g(\Delta\theta')$, we find that $\partial\rho_\theta/\partial\theta^j =\Sigma_k a_j^k(\theta)U_\theta[\partial\rho_\theta/\partial\theta^k]_{\theta=0}U_\theta^*$. This implies that the SLDs at θ are given by $L_{\theta,j}^S =\Sigma_k a_j^k(\theta)U_\theta L_{0,k}^S U_\theta^*$. As a consequence,

$$L_{\theta,j}^S\psi_\theta=\sum_k a_j^k(\theta)U_\theta L_{0,k}^S\psi_0. \tag{5}$$

Here we have set as $\rho_\theta=|\psi_\theta\rangle\langle\psi_\theta|$ with $\psi_\theta=U_\theta\psi_0$. Now suppose $\mathcal{P}$ is locally coherent at ρ_0. Then the vector $(\mathcal{D}_0 L_{0,k}^S)\psi_0=iL_{0,k}^S\psi_0$ [see (4)] belongs to the real linear span of $\{L_{0,k'}^S\psi_0\}_{k'=1}^n$; hence, the vector $(\mathcal{D}_\theta L_{\theta,j}^S)\psi_\theta=iL_{\theta,j}^S\psi_\theta$ belongs to the real linear span of $\{L_{\theta,j'}^S\psi_\theta\}_{j'=1}^n$ because of (5) and the nonsingularity of the matrix $a_j^k(\theta)$. This implies that $\mathcal{P}$ is locally coherent at every point θ. □

It is clear from the definition that if $\mathcal{P}$ is locally coherent at θ, then $\mathcal{T}_\theta^S(\mathcal{P})$ forms a symplectic space with the symplectic form being the restriction of $[\cdot,\cdot]_{\rho_\theta}$. In particular, the dimensionality of $\mathcal{T}_\theta^S(\mathcal{P})$ is necessarily even (say $n=2m$), and it has a symplectic basis $\{\tilde{L}_{\theta,j}^S\}_{j=1}^{2m}$, satisfying

$$[\tilde{L}_{\theta,j}^S,\tilde{L}_{\theta,k}^S]_{\rho_\theta}=\begin{cases}-1, & \text{if } j \text{ is odd and } k=j+1,\\ 1, & \text{if } j \text{ is even and } k=j-1,\\ 0, & \text{otherwise.}\end{cases}$$

Furthermore, since $\mathcal{D}_\theta$ is unitary on $\mathcal{T}_\theta^S(\mathcal{P})$ with respect to the inner product $\langle\cdot,\cdot\rangle_{\rho_\theta}$, we can take $\{\tilde{L}_{\theta,j}^S\}$ to be orthonormal. Such a basis, which we shall call a *normalized ρ_θ-symplectic basis*, satisfies

$$\mathcal{D}_\theta\begin{bmatrix}\tilde{L}_{\theta,1}^S\\ \tilde{L}_{\theta,2}^S\\ \tilde{L}_{\theta,3}^S\\ \tilde{L}_{\theta,4}^S\\ \vdots\\ \tilde{L}_{\theta,2m-1}^S\\ \tilde{L}_{\theta,2m}^S\end{bmatrix}=\begin{bmatrix}0 & 1 & & & & & \\ -1 & 0 & & & & & \\ & & 0 & 1 & & & \\ & & -1 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & 1\\ & & & & & -1 & 0\end{bmatrix}\begin{bmatrix}\tilde{L}_{\theta,1}^S\\ \tilde{L}_{\theta,2}^S\\ \tilde{L}_{\theta,3}^S\\ \tilde{L}_{\theta,4}^S\\ \vdots\\ \tilde{L}_{\theta,2m-1}^S\\ \tilde{L}_{\theta,2m}^S\end{bmatrix}. \tag{6}$$

Thus, the SLD tangent space of a coherent model is decomposed into two-dimensional $\mathcal{D}_\theta$-invariant subspaces. This suggests the importance of studying two-dimensional coherent models.

Now, let us characterize a two-dimensional coherent model.

Theorem 3: *For a two-dimensional pure state model* $\mathcal{P}=\{\rho_\theta=|\psi_\theta\rangle\langle\psi_\theta|;\theta\in\Theta\}$, *the following three conditions are equivalent.*

(α) $\mathcal{P}$ *is locally coherent at* θ.
(β) $L^S_{\theta,1}\psi_\theta$ *and* $L^S_{\theta,2}\psi_\theta$ *are linearly dependent.*
(γ) $L^A_{\theta,1}\psi_\theta$ *and* $L^A_{\theta,2}\psi_\theta$ *are linearly dependent.*

Before going to the proof, we should remark that the condition (β) does not conflict with the fact that $L^S_{\theta,1}$ and $L^S_{\theta,2}$ are linearly independent due to the nondegeneracy of the parametrization $\theta\mapsto\rho_\theta$. Indeed, the linear independence of $\{L^S_{\theta,1},L^S_{\theta,2}\}$ is concerned with the real linear structure of $\mathcal{L}^2_h(\rho_\theta)$ and is equivalent to the real linear independence of $\{L^S_{\theta,1}\psi_\theta,L^S_{\theta,2}\psi_\theta\}$. On the other hand, the condition (β) asserts the complex linear dependence of the same vectors.

Proof: The proof relies essentially on (4). We only need to show that $(\alpha)\Leftrightarrow(\beta)$, since $(\beta)\Leftrightarrow(\gamma)$ is obvious from the identity $L^S_{\theta,j}\psi_\theta=-(\mathcal{D}_\theta L^A_{\theta,j})\psi_\theta=-iL^A_{\theta,j}\psi_\theta$. Let $\varphi_j:=L^S_{\theta,j}\psi_\theta$, and assume ($\alpha$) first. Then there exist real numbers x, y such that $\mathcal{D}_\theta L^S_{\theta,1}=xL^S_{\theta,1}+yL^S_{\theta,2}$. This is equivalent to $i\varphi_1=x\varphi_1+y\varphi_2$ and leads to (β). Assume (β) in turn. Recalling the real linear independence of $\{\varphi_1,\varphi_2\}$, we see that there exist real numbers x, y satisfying $\varphi_2=(x+iy)\varphi_1$ with $y\neq0$. It then follows that $i\varphi_1=-(x/y)\varphi_1+(1/y)\varphi_2$ and $\mathcal{D}_\theta L^S_{\theta,1}=-(x/y)L^S_{\theta,1}+(1/y)L^S_{\theta,2}$. Similarly, $\mathcal{D}_\theta L^S_{\theta,2}$ is shown to be a real linear combination of $\{L^S_{\theta,1},L^S_{\theta,2}\}$, and thus ($\alpha$) is verified. □

The following corollary, whose proof is now straightforward from Theorem 3 and (4), offers a mostly useful method to treat group covariant coherent models as exemplifed in Sec. VI. Moreover, Eq. (7) in the corollary reveals a close connection with the conventional definition of coherent states.[10] Indeed, this fact gave a motive for the nomenclature of the coherent model.

Corollary 4: Let $\mathcal{P}=\{\rho_\theta=|\psi_\theta\rangle\langle\psi_\theta|;\theta\in\Theta\}$ *be a two-dimensional pure state model and let* $\mathcal{T}^A_\theta(\mathcal{P})$ *be the real linear span of ALDs* $\{L^A_{\theta,1},L^A_{\theta,2}\}$ *at* θ. *Then* $\mathcal{P}$ *is locally coherent at* θ *iff there exist nonzero elements* X_1,X_2 *in* $\mathcal{T}^A_\theta(\mathcal{P})$, *satisfying*

$$(X_1+iX_2)\psi_\theta=0. \tag{7}$$

Moreover, (7) is also necessary and sufficient for $\{cX_j\}_{j=1,2}$ *to form a normalized* ρ_θ*-symplectic basis of* $\mathcal{T}^S_\theta(\mathcal{P})[=\mathcal{T}^A_\theta(\mathcal{P})]$ *with a common normalizing constant c. Under the condition (7), the linear relations,*

$$L^A_{\theta,1}=c_{11}X_1+c_{12}X_2,\quad L^A_{\theta,2}=c_{21}X_1+c_{22}X_2,$$

imply

$$L^S_{\theta,1}=c_{12}X_1-c_{11}X_2,\quad L^S_{\theta,2}=c_{22}X_1-c_{21}X_2.$$

IV. GENERALIZED RLD BOUND

Throughout this section we consider an n-dimensional model $\mathcal{P}=\{\rho_\theta\}$ of general (i.e., not necessarily pure) states satisfying the regularity conditions (a) and (b) presented in Sec. III.

Let $\mathcal{L}^2_+(\rho)$ denote the completion of $\mathcal{B}(\mathcal{H})$ with respect to the preinner product $(\cdot,\cdot)_\rho$. Since $(X,X)_\rho\leqslant2\langle X,X\rangle_\rho$, then $\mathcal{L}^2(\rho)\subset\mathcal{L}^2_+(\rho)$. The *right logarithmic derivative* (RLD) $L^R_{\theta,j}$ in the jth direction of a model $\mathcal{P}=\{\rho_\theta\}$, when it exists, is defined by the requirement that

$$\frac{\partial}{\partial\theta^j}\mathrm{Tr}\,\rho_\theta X=(L^R_{\theta,j},X)_{\rho_\theta},\quad L^R_{\theta,j}\in\mathcal{L}^2_+(\rho_\theta),$$

for all $X\in\mathcal{L}_+^2(\rho_\theta)$, or, formally, $\partial\rho_\theta/\partial\theta^j=(L_{\theta,j}^R)^*\rho_\theta=\rho_\theta L_{\theta,j}^R$. The covariance matrix of an arbitrary locally unbiased estimator M is then bounded from below as

$$V_\theta[M]\geqslant(J_\theta^R)^{-1}, \tag{8}$$

where $J_\theta^R=[(L_{\theta,j}^R,L_{\theta,k}^R)_{\rho_\theta}]$ is the RLD Fisher information matrix (Ref. 2, p. 279). When a real positive definite matrix G is specified as the weight for the estimation accuracy, the total deviation is bounded from below as

$$\operatorname{tr} GV_\theta[M]\geqslant C^R:=\operatorname{tr} G\operatorname{Re}(J_\theta^R)^{-1}+\operatorname{tr}\operatorname{abs} G\operatorname{Im}(J_\theta^R)^{-1}, \tag{9}$$

where $\operatorname{tr}\operatorname{abs} A$ denotes the absolute sum of the eigenvalues of matrix A; see Ref. 2, p. 284. The RLD thus gives a CR bound and plays a crucial role in optical communication theory.[3,2]

The RLD exists iff there is a constant c such that

$$\left|\frac{\partial}{\partial\theta^j}\operatorname{Tr}\rho_\theta X\right|^2\leqslant c(X,X)_{\rho_\theta}, \tag{10}$$

for all $X\in\mathcal{B}(\mathcal{H})$. Thus, the RLD does not always exist for a model satisfying the weaker condition (b). In particular, it never exists for a pure state model $\rho_\theta=|\psi_\theta\rangle\langle\psi_\theta|$. To see this, let us fix a θ arbitrarily and take a vector $x\in\mathcal{H}$ such that $\langle\psi_\theta|x\rangle=0$ and $\langle\partial\psi_\theta/\partial\theta^j|x\rangle\neq 0$. [This is indeed possible because ψ_θ and $\partial\psi_\theta/\partial\theta^j$ are linearly independent, owing to $(\partial/\partial\theta^j)\langle\psi_\theta|\psi_\theta\rangle=0$ and $(\partial/\partial\theta^j)|\psi_\theta\rangle\langle\psi_\theta|\neq 0$.] Then $X=|x\rangle\langle\psi_\theta|$ satisfies $(X,X)_{\rho_\theta}=0$ and $(\partial/\partial\theta^j)\operatorname{Tr}\rho_\theta X\neq 0$. It is, however, important to notice that what is needed in estimation theory is not the RLD itself but the inverse of the RLD Fisher information matrix, as indicated by (8) and (9).

In his book (Ref. 2, p. 280), Holevo has shown that when a model satisfying the regularity conditions (a) and (10) has a $\mathcal{D}_\theta$-invariant SLD tangent space, the $(J_\theta^R)^{-1}$ is expressed only in terms of SLDs; so is the CR bound (9). We generalize this result to a wider class of models that satisfy only the weaker conditions (a) and (b).

Theorem 5: *Suppose we are given an n-dimensional model $\mathcal{P}=\{\rho_\theta\}$ having a $\mathcal{D}_\theta$-invariant SLD tangent space $\mathcal{T}_\theta^S(\mathcal{P})$. Then for all locally unbiased measurements M at θ,*

$$V_\theta[M]\geqslant(J_\theta^S)^{-1}+i(J_\theta^S)^{-1}D_\theta(J_\theta^S)^{-1},$$

where $D_\theta=[[L_{\theta,j}^S,L_{\theta,k}^S]_{\rho_\theta}]$.

Proof: Let us introduce a family of inner products on $\mathcal{L}^2(\rho_\theta)$ having a parameter $\epsilon\in(0,1]$:

$$(X,Y)_{\rho_\theta}^{(\epsilon)}=(1-\epsilon)(X,Y)_{\rho_\theta}+\epsilon\langle X,Y\rangle_{\rho_\theta}.$$

Since

$$\epsilon\langle X,X\rangle_{\rho_\theta}\leqslant(X,X)_{\rho_\theta}^{(\epsilon)}\leqslant(2-\epsilon)\langle X,X\rangle_{\rho_\theta},$$

there exists, for each ϵ, a unique operator $L_{\theta,j}^{(\epsilon)}\in\mathcal{L}^2(\rho_\theta)$ that satisfies

$$\frac{\partial}{\partial\theta^j}\operatorname{Tr}\rho_\theta X=(L_{\theta,j}^{(\epsilon)},X)_{\rho_\theta}^{(\epsilon)},$$

for all $X\in\mathcal{L}^2(\rho_\theta)$. Then, in a quite similar way to the derivation of the quantum Cramér–Rao inequality, we have

$$V_\theta[M]\geqslant(J_\theta^{(\epsilon)})^{-1},\quad J_\theta^{(\epsilon)}=[(L_{\theta,j}^{(\epsilon)},L_{\theta,k}^{(\epsilon)})_{\rho_\theta}^{(\epsilon)}]. \tag{11}$$

Now, observing the identity $(X,Y)^{(\epsilon)}_{\rho_\theta}=\langle X,Y\rangle_{\rho_\theta}+i(1-\epsilon)[X,Y]_{\rho_\theta}$, and using the definition of $\mathcal{D}_{\rho_\theta}=\mathcal{D}_\theta$, we see that, for all $Y\in\mathcal{L}^2(\rho_\theta)$,

$$\frac{\partial}{\partial\theta^j}\operatorname{Tr}\rho_\theta Y=\langle L^S_{\theta,j},Y\rangle_{\rho_\theta}=(L^{(\epsilon)}_{\theta,j},Y)^{(\epsilon)}_{\rho_\theta}=\langle\{\mathcal{I}+i(1-\epsilon)\mathcal{D}_\theta\}L^{(\epsilon)}_{\theta,j},Y\rangle_{\rho_\theta}.$$

Then $L^S_{\theta,j}=\{\mathcal{I}+i(1-\epsilon)\mathcal{D}_\theta\}L^{(\epsilon)}_{\theta,j}$; hence $(L^{(\epsilon)}_{\theta,j},L^{(\epsilon)}_{\theta,k})^{(\epsilon)}_{\rho\theta}=\langle L^S_{\theta,j},\{\mathcal{I}+i(1-\epsilon)\mathcal{D}_\theta\}^{-1}L^S_{\theta,k}\rangle_{\rho_\theta}$. Let us introduce Dirac's notation $|L^S_{\theta,j}\rangle$ for the Hilbert space $(\mathcal{L}^2(\rho_\theta),\langle\cdot,\cdot\rangle_{\rho_\theta})$, and let $\Gamma_\theta:=[|L^S_{\theta,1}\rangle,\ldots,|L^S_{\theta,n}\rangle]$. Then $\Gamma^*_\theta\Gamma_\theta=J^S_\theta$ and $\Gamma^*_\theta\mathcal{D}_\theta\Gamma_\theta=D_\theta$. And the matrix $J^{(\epsilon)}_\theta$ can be written in the form $J^{(\epsilon)}_\theta=\Gamma^*_\theta\{\mathcal{I}+i(1-\epsilon)\mathcal{D}_\theta\}^{-1}\Gamma_\theta$. Thus, from the assumption that $\mathcal{T}^S_\theta(\mathcal{P})$ is $\mathcal{D}_\theta$ invariant, the inverse of $J^{(\epsilon)}_\theta$ is explicitly given by

$$(J^{(\epsilon)}_\theta)^{-1}=(J^S_\theta)^{-1}\Gamma^*_\theta\{\mathcal{I}+i(1-\epsilon)\mathcal{D}_\theta\}\Gamma_\theta(J^S_\theta)^{-1}=(J^S_\theta)^{-1}+i(1-\epsilon)(J^S_\theta)^{-1}D_\theta(J^S_\theta)^{-1}. \quad (12)$$

Combining (11) and (12), and taking the limit $\epsilon\downarrow 0$, we have the theorem. □

Theorem 5 asserts that even for a model that does not have the RLDs, the $\lim_{\epsilon\downarrow 0}(J^{(\epsilon)}_\theta)^{-1}$ indeed gives a generalization of $(J^R_\theta)^{-1}$ as long as the SLD tangent space is $\mathcal{D}_\theta$ invariant. Then, by using Theorem 5 and an analogous argument to the derivation of (9), we obtain the CR bound,

$$C^R=\operatorname{tr}G(J^S_\theta)^{-1}+\operatorname{tr}\operatorname{abs}G(J^S_\theta)^{-1}D_\theta(J^S_\theta)^{-1}, \quad (13)$$

for models each having a $\mathcal{D}_\theta$-invariant SLD tangent space $\mathcal{T}^S_\theta(\mathcal{P})$. This may be called a generalized RLD bound. We will show in the next section that this bound is achievable in a coherent model.

V. OPTIMAL ESTIMATION FOR TWO-DIMENSIONAL COHERENT MODELS

We now proceed to a parameter estimation for a pure coherent model. In particular, taking into account the symplectic structure (6) of the SLD tangent space, we restrict ourselves to a two-dimensional case. We note that as long as we are concerned with the achievable CR bound at each point on the model $\{\rho_\theta\}$, we can take the weight as $G=I$ without loss of generality. In fact, let M be a locally unbiased measurement for the parameter $\theta=(\theta^1,\theta^2)$ and let $p(\hat\theta^1,\hat\theta^2)d\hat\theta=\operatorname{Tr}\rho_\theta M(d\hat\theta)$ be the corresponding joint distribution. The coordinate transformation $\eta^i=\Sigma_j h^i_j\theta^j$, where $H=[h^i_j]$ is a real regular matrix, then induces another measurement $N(d\hat\eta)$, which corresponds to the joint distribution $q(\hat\eta^1,\hat\eta^2)d\hat\eta=p(\hat\theta^1,\hat\theta^2)d\hat\theta$ and is locally unbiased for the parameter $\eta=(\eta^1,\eta^2)$. In this case, $\operatorname{tr}V_\eta[N]=\operatorname{tr}({}^tHH)V_\theta[M]$. Thus, the parameter estimation for θ with the weight $G={}^tHH$ is equivalent to that for η with the weight I.

Now suppose we are given a two-dimensional coherent model $\mathcal{P}=\{\rho_\theta;\theta=(\theta^1,\theta^2)\in\Theta\}$. Let $\{L^i\}$ be the dual basis of the SLDs: $L^i=\Sigma_j J^{ij}L^S_{\theta,j}$ with J^{ij} being the (i,j) entry of $(J^S_\theta)^{-1}$. Then

$$(J^S_\theta)^{-1}=\begin{bmatrix}\langle L^1,L^1\rangle_{\rho_\theta} & \langle L^1,L^2\rangle_{\rho_\theta}\\ \langle L^2,L^1\rangle_{\rho_\theta} & \langle L^2,L^2\rangle_{\rho_\theta}\end{bmatrix}$$

and

$$(J^S_\theta)^{-1}D_\theta(J^S_\theta)^{-1}=\begin{bmatrix}0 & [L^1,L^2]_{\rho_\theta}\\ [L^2,L^1]_{\rho_\theta} & 0\end{bmatrix}.$$

Thus the generalized RLD bound (13) for $G=I$ can be rewritten in the form

$$C^R=\langle L^1,L^1\rangle_{\rho_\theta}+\langle L^2,L^2\rangle_{\rho_\theta}+2|[L^1,L^2]_{\rho_\theta}|. \quad (14)$$

We will show that the bound C^R is achievable. In what follows, we fix a $\theta=(\theta^1,\theta^2)$ arbitrarily.

Let us consider a random measurement as follows. We first introduce a linear transformation ϕ: $\mathcal{T}_\theta^S(\mathcal{P})\to\mathcal{T}_\theta^S(\mathcal{P})$ by

$$\phi(X)=\langle L^1,X\rangle_{\rho_\theta}L^1+\langle L^2,X\rangle_{\rho_\theta}L^2.$$

Since ϕ is symmetric and positive definite, it has positive eigenvalues λ_1,λ_2, and mutually orthogonal unit eigenvectors A_1,A_2 satisfying $\phi(A_\nu)=\lambda_\nu A_\nu$, $\nu=1,2$. We next take positive numbers p_1,p_2 satisfying $p_1+p_2=1$. Now letting

$$\int \xi E_\nu(d\xi), \quad \nu=1,2$$

be the spectral decompositions of arbitrarily fixed representatives of A_ν, we define a generalized measurement,

$$M(\nu,d\xi)=p_\nu E_\nu(d\xi).$$

This has the following physical interpretation: Select one of the two ''observables'' A_1,A_2 according to the probability p_1,p_2, respectively, and measure it in a usual sense.

Now suppose we have selected A_ν and have obtained an outcome ξ. We identify this result with a pair of real quantities,

$$\hat\theta^i(\nu,\xi)=\theta^i+\frac{\xi}{p_\nu}\langle L^i,A_\nu\rangle_{\rho_\theta}, \quad i=1,2.$$

The pair $\{\hat\theta^i(\nu,\xi)\}_{i=1,2}$ satisfies the local unbiasedness condition at θ:

$$\sum_{\nu=1}^{2}\int \hat\theta^i(\nu,\xi)\mathrm{Tr}\,\rho_\theta M(\nu,d\xi)=\theta^i, \quad i=1,2, \tag{15}$$

$$\sum_{\nu=1}^{2}\int \hat\theta^i(\nu,\xi)\frac{\partial}{\partial\theta^j}\mathrm{Tr}\,\rho_\theta M(\nu,d\xi)=\delta_j^i, \quad i,j=1,2. \tag{16}$$

To prove (15), we used the fact that $A_\nu\in\mathcal{T}_\theta^S(\mathcal{P})$, i.e., $\langle I,A_\nu\rangle_{\rho_\theta}=0$. To prove (16), observe that

$$\int \xi\frac{\partial}{\partial\theta^j}\mathrm{Tr}\,\rho_\theta E_\nu(d\xi)=\langle L_{\theta,j}^S,A_\nu\rangle_{\rho_\theta},$$

so that the left-hand side of (16) becomes

$$\sum_{\nu=1}^{2}\langle L^i,A_\nu\rangle_{\rho_\theta}\langle L_{\theta,j}^S,A_\nu\rangle_{\rho_\theta}=\langle L^i,L_{\theta,j}^S\rangle_{\rho_\theta}=\delta_j^i.$$

With this measurement M,

$$\begin{aligned}\mathrm{tr}\,V_\theta[M]&=\sum_{\nu=1}^{2}\int[(\hat\theta^1(\nu,\xi)-\theta^1)^2+(\hat\theta^2(\nu,\xi)-\theta^2)^2]\mathrm{Tr}\,\rho_\theta M(\nu,d\xi)\\&=\sum_{\nu=1}^{2}\frac{1}{p_\nu}[\langle L^1,A_\nu\rangle_{\rho_\theta}^2+\langle L^2,A_\nu\rangle_{\rho_\theta}^2].\end{aligned}$$

In the second equality, we used the fact that

$$\int \xi^2 \,\mathrm{Tr}\,\rho_\theta E_\nu(d\xi)=\langle A_\nu ,A_\nu\rangle_{\rho_\theta}=1.$$

Since, for given $\mu_1,\mu_2>0$, $\mu_1/p_1+\mu_2/p_2$ takes the minimum $(\sqrt{\mu_1}+\sqrt{\mu_2})^2$ at $p_\nu=\sqrt{\mu_\nu}/(\sqrt{\mu_1}+\sqrt{\mu_2})$, we see

$$\begin{aligned}\min_{\{p_\nu\}} \mathrm{tr}\, V_\theta[M]&=\left[\sqrt{\langle L^1,A_1\rangle_{\rho_\theta}^2+\langle L^2,A_1\rangle_{\rho_\theta}^2}+\sqrt{\langle L^1,A_2\rangle_{\rho_\theta}^2+\langle L^2,A_2\rangle_{\rho_\theta}^2}\right]^2\\ &=\left[\sqrt{\langle A_1,\phi(A_1)\rangle_{\rho_\theta}}+\sqrt{\langle A_2,\phi(A_2)\rangle_{\rho_\theta}}\right]^2\\ &=\left[\sqrt{\lambda_1}+\sqrt{\lambda_2}\right]^2\\ &=\langle L^1,L^1\rangle_{\rho_\theta}+\langle L^2,L^2\rangle_{\rho_\theta}+2\sqrt{\langle L^1,L^1\rangle_{\rho_\theta}\langle L^2,L^2\rangle_{\rho_\theta}-\langle L^1,L^2\rangle_{\rho_\theta}^2}. \end{aligned}\tag{17}$$

The last equality follows from the fact that the trace $\lambda_1+\lambda_2$ and the determinant $\lambda_1\lambda_2$ of the linear transformation ϕ are independent of the choice of the basis that represents ϕ in a matrix form.

The random measurement presented above was first introduced in Ref. 5 by one of the present authors. In that paper, it was also shown that the problem of finding the achievable CR bound for an arbitrary two-parameter faithful spin-$\frac{1}{2}$ model can be reduced to an easy minimization problem. Interestingly, the explicit solution of the minimization problem, i.e., the achievable CR bound, turns out to be identical to the quantity (17), although the model treated there is not pure nor has, in general, a $\mathcal{D}_\rho$-invariant tangent space.

Now we establish the relation between (14) and (17) for a coherent model.

Theorem 6: *For a two-dimensional coherent model* $\{\rho_\theta=|\psi_\theta\rangle\langle\psi_\theta|\}$, *the lower bound (14) is identical to (17). In other words, the generalized RLD bound (14) is achievable.*

Proof: By Theorem 3, $L^1\psi_\theta$ and $L^2\psi_\theta$ are linearly dependent. Therefore

$$\det\begin{bmatrix}\langle L^1\psi_\theta|L^1\psi_\theta\rangle & \langle L^1\psi_\theta|L^2\psi_\theta\rangle\\ \langle L^2\psi_\theta|L^1\psi_\theta\rangle & \langle L^2\psi_\theta|L^2\psi_\theta\rangle\end{bmatrix}=0,$$

which leads to

$$(\mathrm{Im}\langle L^1\psi_\theta|L^2\psi_\theta\rangle)^2=\langle L^1\psi_\theta|L^1\psi_\theta\rangle\langle L^2\psi_\theta|L^2\psi_\theta\rangle-(\mathrm{Re}\langle L^1\psi_\theta|L^2\psi_\theta\rangle)^2.$$

By (1) and (2), this can be read as

$$|[L^1,L^2]_{\rho_\theta}|^2=\langle L^1,L^1\rangle_{\rho_\theta}\langle L^2,L^2\rangle_{\rho_\theta}-\langle L^1,L^2\rangle_{\rho_\theta}^2,$$

which proves the theorem.

It should be noted that a more convincing result has been obtained by Matsumoto.[11] He proved that the CR bound (13) is achievable for a $2m$-dimensional coherent model with an arbitrary weight G.

It is also worth noting that the achievability of (14) is closely related to the Heisenberg uncertainty relation. By a coordinate transformation, we can assume that the SLD Fisher information matrix is diagonal at a fixed $\rho_\theta=|\psi_\theta\rangle\langle\psi_\theta|$. Then there exist nonzero real numbers c_1,c_2 and a normalized ρ_θ-symplectic basis $\{\tilde{L}_1^S,\tilde{L}_2^S\}$ such that $L_j^S=c_j\tilde{L}_j^S$. Then $L^j=\tilde{L}_j^S/c_j$, and, by (7),

$$(c_1L^1+ic_2L^2)\psi_\theta=0.$$

This is nothing but the equality condition for the Heisenberg uncertainty relation. So we have

$$\langle L^1,L^1\rangle_{\rho_\theta}\langle L^2,L^2\rangle_{\rho_\theta}=|[L^1,L^2]_{\rho_\theta}|^2.$$

This equation, combined with the assumption that $\langle L^1,L^2\rangle_{\rho_\theta}=0$, gives another proof of Theorem 6 for an orthogonal parametrization at ρ_θ.

VI. EXAMPLES

In this section we calculate the achievable CR bounds for canonical and spin coherent models. Throughout this section, adjoint operators and complex conjugate numbers are denoted by † and *, respectively, according to the convention in physics. Also, we use the same letter for both a square summable operator and the corresponding element in $\mathcal{L}_h^2(\rho)$.

A. Canonical squeezed state model

The canonical squeezed state[12,13] is defined by

$$\rho_{q,p}=D(q,p)|0\rangle_{\xi\xi}\langle 0|D^\dagger(q,p) \quad (q,p\in\mathbf{R}),$$

where $D(q,p)=\exp(za^\dagger - z^* a)$ denotes the shift operator with $z=(q+ip)/\sqrt{2}$, and a and $a^\dagger$ are annihilation and creation operators, respectively, with $a=(Q+iP)/\sqrt{2}$. Further, $|0\rangle_\xi=\exp[(\xi a^{\dagger 2}-\xi^* a^2)/2]|0\rangle$ is the squeezed vacuum with $|0\rangle$ the Fock vacuum, and ξ is a complex number that represents the squeezing ratio: let $\xi=se^{i\theta}$.

Comparing the identity $b|z\rangle_\xi=\beta|z\rangle_\xi$ with Corollary 4, where $|z\rangle_\xi=D(q,p)|0\rangle_\xi$, $b=a\cosh s - a^\dagger e^{i\theta}\sinh s$, and $\beta=z\cosh s - z^* e^{i\theta}\sinh s$, we see that $\rho_{q,p}$ is a two-dimensional coherent model, and a normalized $\rho_{q,p}$-symplectic basis is given by

$$\tilde{L}_1^S=\sqrt{2}[(Q-qI)(\cosh s-\cos\theta\sinh s)-(P-pI)\sin\theta\sinh s],$$

$$\tilde{L}_2^S=\sqrt{2}[(P-pI)(\cosh s+\cos\theta\sinh s)-(Q-qI)\sin\theta\sinh s].$$

The SLDs at $\rho_{q,p}$ are calculated by operating $-\mathcal{D}_{q,p}$ to ALDs at $\rho_{q,p}$. By expanding ALDs, $L_q^A=2(P-pI)$, $L_p^A=-2(Q-qI)$ into linear combinations of $\tilde{L}_1^S$, $\tilde{L}_2^S$, and using the relations $\mathcal{D}_{q,p}\tilde{L}_1^S=\tilde{L}_2^S$, $\mathcal{D}_{q,p}\tilde{L}_2^S=-\tilde{L}_1^S$, we have

$$L_q^S=2[(Q-qI)(\cosh 2s-\cos\theta\sinh 2s)-(P-pI)\sin\theta\sinh 2s],$$

$$L_p^S=2[(P-pI)(\cosh 2s+\cos\theta\sinh 2s)-(Q-qI)\sin\theta\sinh 2s].$$

The corresponding SLD Fisher information matrix becomes

$$J_{q,p}^S=2\begin{bmatrix}\cosh 2s-\cos\theta\sinh 2s & -\sin\theta\sinh 2s\\ -\sin\theta\sinh 2s & \cosh 2s+\cos\theta\sinh 2s\end{bmatrix}.$$

Then from (17), we have

$$\min_M \operatorname{tr} V_{q,p}[M]=\cosh 2s+1.$$

B. Spin coherent state model

The spin coherent state[14,15] in the spin j representation is defined by

$$\rho_{\theta,\varphi}=R(\theta,\varphi)|j\rangle\langle j|R^\dagger(\theta,\varphi), \quad (0\le\theta\le\pi,\ 0\le\varphi<2\pi),$$

where (θ,φ) is the polar coordinate system (the north pole is $\theta=0$ and the x axis corresponds to $\varphi=0$), $R(\theta,\varphi)=\exp[i\theta(J_x\sin\varphi-J_y\cos\varphi)]$ the rotation through an angle $-\theta$ about an axis $(\sin\varphi,-\cos\varphi,0)$, and $|j\rangle$ the highest weight state with respect to J_z that corresponds to the north pole.

Since $J_+|j\rangle=(J_x+iJ_y)|j\rangle=0$, we find that $\rho_{\theta,\varphi}$ is a two-dimensional coherent model, and a normalized $\rho_{0,0}$-symplectic basis is $\tilde{L}_1^S(0,0)=\sqrt{2/j}J_x$, $\tilde{L}_2^S(0,0)=\sqrt{2/j}J_y$. A normalized $\rho_{\theta,\varphi}$-symplectic basis is then calculated as

$$\tilde{L}_k^S(\theta,\varphi)=R(\theta,\varphi)\tilde{L}_k^S(0,0)R^\dagger(\theta,\varphi),$$

where $k=1,2$.

On the other hand, the generators of rotations about axes $(\sin\varphi,-\cos\varphi,0)$ and $(\cos\varphi,\sin\varphi,0)$ at $\theta=0$ are $i(J_x\sin\varphi-J_y\cos\varphi)$ and $i(J_x\cos\varphi+J_y\sin\varphi)$, respectively. Therefore ALDs for the model at $\rho_{\theta,\varphi}$ are given by

$$\begin{aligned}L_\theta^A(\theta,\varphi)&=R(\theta,\varphi)\{-2(J_x\sin\varphi-J_y\cos\varphi)\}R^\dagger(\theta,\varphi)\\&=-\sqrt{2j}\{\tilde{L}_1^S(\theta,\varphi)\sin\varphi-\tilde{L}_2^S(\theta,\varphi)\cos\varphi\},\end{aligned}$$

$$\begin{aligned}L_\varphi^A(\theta,\varphi)&=R(\theta,\varphi)\{-2(J_x\cos\varphi+J_y\sin\varphi)\sin\theta\}R^\dagger(\theta,\varphi)\\&=-\sqrt{2j}\{\tilde{L}_1^S(\theta,\varphi)\sin\theta\cos\varphi+\tilde{L}_2^S(\theta,\varphi)\sin\theta\sin\varphi\}.\end{aligned}$$

The SLDs at $\rho_{\theta,\varphi}$ are calculated by operating $-\mathcal{D}_{\theta,\varphi}$ to ALDs, to obtain

$$L_\theta^S(\theta,\varphi)=\sqrt{2j}\{\tilde{L}_1^S(\theta,\varphi)\cos\varphi+\tilde{L}_2^S(\theta,\varphi)\sin\varphi\},$$

$$L_\varphi^S(\theta,\varphi)=-\sqrt{2j}\{\tilde{L}_1^S(\theta,\varphi)\sin\theta\sin\varphi-\tilde{L}_2^S(\theta,\varphi)\sin\theta\cos\varphi\}.$$

Since $\rho_{\theta,\varphi}$-symplectic basis $\{\tilde{L}_k^S(\theta,\varphi)\}_{k=1,2}$ is orthonormal, the SLD Fisher information matrix and the matrix D are easily calculated:

$$J_{\theta,\varphi}^S=\begin{bmatrix}2j & 0\\ 0 & 2j\sin^2\theta\end{bmatrix},\quad D_{\theta,\varphi}=\begin{bmatrix}0 & -2j\sin\theta\\ 2j\sin\theta & 0\end{bmatrix}.$$

We thus have

$$\min_M \operatorname{tr} V_{\theta,\varphi}[M]=\frac{1}{2j}\left(1+\frac{1}{\sin\theta}\right)^2.$$

VII. CONCLUSIONS

We introduced a class of quantum pure state models called the coherent models. They are characterized by a symplectic structure of the tangent space, and have a close connection with the conventional generalized coherent states in mathematical physics. A Cramér–Rao-type bound for a coherent model was derived by an analogous argument to the derivation of the right logarithmic derivative bound. Moreover, by an argument of random measurement, this lower bound was found to be achievable.

ACKNOWLEDGMENT

We thank Keiji Matsumoto for helpful suggestions, with which the manuscript has been improved as compared with the early version[16] of this paper.

[1] C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic, New York, 1976).
[2] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).
[3] H. P. Yuen and M. Lax, IEEE Trans. Inf. Theory **IT-19**, 740 (1973).
[4] H. Nagaoka, IEICE Technical Report **IT89-42**, 1989, p. 9.
[5] H. Nagaoka, Trans. Jpn. Soc. Indust. Appl. Math. **1**, 43 (1991) (in Japanese).

[6] A. Fujiwara and H. Nagaoka, Phys. Lett. A **201**, 119 (1995).
[7] A. Fujiwara and H. Nagaoka, in *Quantum Coherence and Decoherence*, edited by K. Fujikawa and Y. A. Ono (Elsevier, Amsterdam, 1996), p. 303.
[8] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics* (Cambridge University Press, Cambridge, 1984).
[9] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, II* (Wiley, New York, 1969).
[10] J. R. Klauder and B. Skagerstam, *Coherent States* (World Scientific, Singapore, 1985).
[11] K. Matsumoto, ''A geometrical approach to quantum estimation theory,'' Ph.D. thesis, University of Tokyo, 1998.
[12] H. P. Yuen, Phys. Rev. A **13**, 2226 (1976).
[13] C. M. Caves, K. S. Thorne, R. W. P. Drever, V. D. Sandberg, and M. Zimmermann, Rev. Mod. Phys. **52**, 341 (1980).
[14] J. M. Radcliffe, J. Phys. A **4**, 313 (1971).
[15] F. T. Arecchi, E. Courtens, R. Glimore, and H. Thomas, Phys. Rev. **6**, 2211 (1972).
[16] A. Fujiwara, Mathematical Engineering Technical Report 94-9, 94-10, University of Tokyo, 1994.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Exact solutions of Dirac equation for neutrinos in presence of external fields

M. N. Hounkonnou[a)] and J. E. B. Mendy
Unité de Recherche en Physique Théorique (URPT), Institut de Mathématiques et de Sciences Physiques (IMSP), B.P.613 Porto-Novo, Republic of Benin

(Received 20 July 1998; accepted for publication 11 March 1999)

Exact solutions of the Dirac equation in a chiral representation for neutrinos in the presence of external stresses are investigated in terms of special functions, using the algebraic method of the separation of variables in Cartesian, cylindrical, and spherical coordinates.
[S0022-2488(99)02408-1]

I. INTRODUCTION

It is in the neutrino physics domain that most inconsistencies with the standard model are found. From the results presented in the standard model, it seems clear that neutrinos have opened a window toward new physics. It is nowadays very interesting to study the neutrino beams by means of electromagnetic fields in order to understand their space distributions and to give exacts solutions of the Dirac equation for them.

The investigations of methods giving exact solutions of the Dirac equation with or without external interactions[1–30] have excited great interest in the past few years because of the usefulness of solvable problems to build a consistent and comprehensible theory, and to find an adequate mathematical framework for the interpretation of numerous experimental results.

The methods used in the most fundamental works on this Dirac equation in the presence of external interaction are essentially based on the quaternionic approach proposed by Hautot,[29] the Stäckel separation method developed by Bagrov *et al.*,[22] the shift operator method,[25] the algebraic method proposed by Komarov and Romanov,[30] and the algebraic method of the separation of variables adopted by Shishkin *et al.*[6] Our aim in this paper is to present the theory of neutrinos in external fields from this latter point of view.

Despite the remarkable results attained in these previous works, using one or the other of the mentioned methods, the large set of representations adapted to various situations does not permit us to elaborate on a unified mathematical framework for the exact solutions in the standard system of coordinates like Cartesian, cylindrical, and spherical.

Clearly, it is not often explained in these works why a matrix or geometric representation used to solve the Dirac equation in one system of coordinates is not quite adapted to solve the same problem in other systems of coordinates! What kind of mathematical difficulties occur and oblige us to choose specific representations for a specific system of coordinates! To discuss this equation, we opt here to work in a unique representation, whatever the coordinate system considered.

The algebraic method of the separation of variables[6] consists of reducing the Dirac operator to a sum of two commuting differential operators following the scheme

$$\{H_{\mathrm{D}}\}\Psi \Rightarrow \{H_{\mathrm{D}}\}\Gamma\Gamma^{-1}\Psi \Rightarrow \{\hat{K}_1+\hat{K}_2\}\Phi, \quad \text{with } \Psi=\Gamma\Phi, \tag{1}$$

$$\{\hat{K}_1+\hat{K}_2\}\Phi=0, \tag{2}$$

$$[\hat{K}_1,\hat{K}_2]_-=\hat{K}_1\hat{K}_2-\hat{K}_2\hat{K}_1=0, \tag{3}$$

[a)]Electronic mail: hounkon@syfed.bj.refer.org

where H_D is the initial Dirac operator; and $\hat{K}_1$ and $\hat{K}_2$, the searched commuting differential operators, depend on a set of variables.

Ψ and Φ are the initial spinor, solution of the Dirac equation, and the searched spinor, corresponding to the operator $\hat{K}_1+\hat{K}_2$, respectively.

Thus the initial problem is equivalently reduced to solve the following:

$$\hat{K}_1\Phi=\lambda\Phi=-\hat{K}_2\Phi, \tag{4}$$

where λ is a constant of separation.

Here, we have shown that, adapting this algebraic method of separation of variables in a chiral representation, it is possible to give exact solutions in terms of special functions to Dirac equation with various external fields in different systems of coordinates such as a Cartesian, cylindrical, and spherical one.

The paper is organized as follows. In Sec. II, considering the system of Cartesian coordinates, exact solutions to the Dirac equation in the presence of four types of external fields are given in terms of confluent hypergeometric functions. In Sec. III, the same problem is considered in cylindrical coordinates. In Sec. IV, we discuss the difficulties arising when the spherical coordinates are considered in a chiral representation.

II. CARTESIAN COORDINATES

In Cartesian coodinates, the Dirac equation for a neutrino of mass m in external electromagnetic field reads as follows in chiral representation:

$$\{\gamma_\chi^0\partial_t+\gamma_\chi^1\partial_x+\gamma_\chi^2\partial_y+\gamma_\chi^3\partial_z+m+g\epsilon_{ijk}\gamma_\chi^i\gamma_\chi^jH_k\}\Psi_\chi=0, \tag{5}$$

where

$$\gamma_\chi^0=\begin{pmatrix}0 & I\\ I & 0\end{pmatrix},\qquad \gamma_\chi^k=\begin{pmatrix}0 & -\sigma^k\\ \sigma^k & 0\end{pmatrix},\qquad (k=1,2,3); \tag{6}$$

g is the coupling constant; H_k the external electromagnetic field; and Ψ_χ the spinor solution of the Dirac equation;

$$\epsilon_{ijk}=\begin{cases}+1, & \text{if } (ijk) \text{ is an even permutation of } (123);\\ -1, & \text{if } (ijk) \text{ is an odd permutation of } (123);\\ 0, & \text{otherwise.}\end{cases}$$

Using the algebraic method of the separation of variables, we can write Eq. (5) as follows:

$$\{\hat{K}_{xy}+\hat{K}_{zt}\}\Phi_\chi=0,\quad \hat{K}_{xy}\Phi_\chi=\lambda\Phi_\chi=-\hat{K}_{zt}\Phi_\chi, \tag{7}$$

where

$$\hat{K}_{xy}=\gamma_\chi^1\partial_y-\gamma_\chi^2\partial_x+m\gamma_\chi^1\gamma_\chi^2-gH_z(x,y), \tag{8}$$

$$\hat{K}_{zt}=\gamma_\chi^0\gamma_\chi^1\gamma_\chi^2\partial_t+\gamma_\chi^3\gamma_\chi^1\gamma_\chi^2\partial_z-gH_z(z,t), \tag{9}$$

and

$$\Psi_\chi=\gamma_\chi^1\gamma_\chi^2\Phi_\chi. \tag{10}$$

The magnetic vector H_z reads as

$$H_z = H_z(x,y) + H_z(z,t). \tag{11}$$

Applying the condition $\nabla \underline{H}_z = 0$ on Eq. (11), we find that $H_z(z,t) = H_z(t)$, and if we fix our attention to time-independent fields, this term can be omitted.

$\lambda \Phi_\chi = -\hat{K}_{zt}\Phi_\chi$ establishes the relationship between the different spinor composents of Φ_χ and Ψ_χ, and the value of the constant of separation λ, which satisfies the relation $\lambda^2 = E^2 - k_z^2$.

From $\lambda \Phi_\chi = -\hat{K}_{zt}\Phi_\chi$, we find

$$\Phi_\chi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \dfrac{E+k_z}{\lambda}\eta_1 \\ \dfrac{-E+k_z}{\lambda}\eta_2 \\ \eta_1 \\ \eta_2 \end{pmatrix}. \tag{12}$$

Substituting Eqs. (6) and (11) in (7), we obtain

$$[\gamma_\chi^1 \partial_y - \gamma_\chi^2 \partial_x + m\gamma_\chi^1 \gamma_\chi^2 - gH_z(x,y) - \lambda]\Phi_\chi = 0. \tag{13}$$

This equation splits into a system of ordinary differential equations, mixing the two-component spinors η and ξ as follows:

$$(\sigma^2 d_x - ik_y\sigma^1)\eta - (gH_z(x) + \lambda + im\sigma^3)\xi = 0, \tag{14}$$

$$(-\sigma^2 d_x + ik_y\sigma^1)\xi - (gH_z(x) + \lambda + im\sigma^3)\eta = 0. \tag{15}$$

Now, we are going to solve these equations to obtain the explicit expressions of η_1, η_2, ξ_1, and ξ_2, for some relevant external fields H_z.

(i) H_z is a constant magnetic field: $H_z = \beta$.

In this case, Eqs. (14) and (15) become

$$(\sigma^2 d_x - ik_y\sigma^1)\eta - (g\beta + \lambda + im\sigma^3)\xi = 0, \tag{16}$$

$$(-\sigma^2 d_x + ik_y\sigma^1)\xi - (g\beta + \lambda + im\sigma^3)\eta = 0. \tag{17}$$

Inserting (12) into the spinor equation (16), and using the standard representation for the Pauli matrices, we obtain

$$(d_x + k_y)\eta_2 - i(g\beta + \lambda + im)\left(\frac{E+k_z}{\lambda}\right)\eta_1 = 0, \tag{18}$$

$$(d_x - k_y)\eta_1 - i(g\beta + \lambda + im)\left(\frac{E-k_z}{\lambda}\right)\eta_2 = 0, \tag{19}$$

where

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}. \tag{20}$$

From (19), we have

$$\eta_2=\frac{-i}{(g\beta+\lambda+im)\left(\frac{E-k_z}{\lambda}\right)}(d_x-k_y)\,\eta_1, \tag{21}$$

which we insert into (18) to obtain a second-order differential equation for η_1,

$$\left[\frac{d^2}{dx^2}-k_y^2+(g\beta+\lambda+im)^2\right]\eta_1=0. \tag{22}$$

The solution of this equation reads as

$$\eta_1=c_1\exp[\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]+c_2\exp[-\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]. \tag{23}$$

Substituting (23) into (21), and using (12), we have

$$\eta_2=\frac{-i}{(g\beta+\lambda+im)\left(\frac{E-k_z}{\lambda}\right)}\{c_1[\sqrt{k_y^2-(g\beta+\lambda+im)^2}-k_y]\exp[\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]$$
$$-c_2[\sqrt{k_y^2-(g\beta+\lambda+im)^2}+k_y]\exp[-\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]\}, \tag{24}$$

$$\xi_1=\frac{E+k_z}{\lambda}\{c_1\exp[\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]+c_2\exp[-\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]\}, \tag{25}$$

$$\xi_2=\frac{i}{(g\beta+\lambda+im)}\{c_1[\sqrt{k_y^2-(g\beta+\lambda+im)^2}-k_y]\exp[\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]$$
$$-c_2[\sqrt{k_y^2-(g\beta+\lambda+im)^2}+k_y]\exp[-\sqrt{k_y^2-(g\beta+\lambda+im)^2}x]\}. \tag{26}$$

(ii) H_z linearly depends on spatial coordinate x: $H_z=\beta x$.

Substituting the field expression into the spinor equation (14) with the standard representation for Pauli matrices, and taking into account (12), we obtain

$$(d_x+k_y)\,\eta_2-i(g\beta x+\lambda+im)\left(\frac{E+k_z}{\lambda}\right)\eta_1=0, \tag{27}$$

$$(d_x-k_y)\,\eta_1-i(g\beta x+\lambda+im)\left(\frac{E-k_z}{\lambda}\right)\eta_2=0. \tag{28}$$

From (28), we have

$$\eta_2=\frac{-i}{(g\beta x+\lambda+im)\left(\frac{E-k_z}{\lambda}\right)}(d_x-k_y)\,\eta_1, \tag{29}$$

which we insert into (27) to obtain a second-order ordinary differential equation for η_1,

$$\left[\frac{d^2}{dx^2}-k_y^2+(g\beta x+\lambda+im)^2\right]\eta_1=0. \tag{30}$$

Making the change of variables,

$$g\beta x+\lambda+im=\sqrt{\frac{g\beta}{2}}y, \tag{31}$$

we reduce (30) to a parabolic cylinder equation,[31]

$$\left[\frac{d^2}{dy^2}+\frac{y^2}{4}-\frac{k_y^2}{2g\beta}\right]\eta_1=0, \tag{32}$$

whose solution can be written in terms of confluent hypergeometric functions $M(a,b,z)$:[31]

$$\eta_1=c_1\exp\left(\frac{-iy^2}{4}\right)M\left(\frac{1}{4}-\frac{ik_y^2}{4g\beta},\frac{1}{2},\frac{i}{2}y^2\right)+y\exp\left(\frac{i\pi}{4}\right)c_2\exp\left(\frac{-iy^2}{4}\right)M\left(\frac{3}{4}-\frac{ik_y^2}{4g\beta},\frac{3}{2},\frac{i}{2}y^2\right). \tag{33}$$

Substituting (33) into (29), and using (12) and the recurrence relation

$$M'(a,b,z)=\frac{a}{b}M(a+1,b+1,z), \tag{34}$$

we obtain explicit expressions for other components of the 4-spinor Φ_χ:

$$\eta_2=\frac{i\exp\left(\frac{-iy^2}{4}\right)}{\sqrt{\frac{g\beta}{2}}y\left(\frac{E-k_z}{\lambda}\right)}\left[c_1\left(M\left(\frac{1}{4}-\frac{ik_y^2}{4g\beta},\frac{1}{2},\frac{i}{2}y^2\right)\left(i\sqrt{\frac{g\beta}{2}}y+k_y\right)-iy\sqrt{2g\beta}\left(\frac{1}{2}-\frac{ik_y^2}{2g\beta}\right)\right.\right.$$
$$\left.\times M\left(\frac{5}{4}-\frac{ik_y^2}{4g\beta},\frac{3}{2},\frac{i}{2}y^2\right)\right)-c_2\exp\left(\frac{i\pi}{4}\right)\left(M\left(\frac{3}{4}-\frac{ik_y^2}{4g\beta},\frac{3}{2},\frac{i}{2}y^2\right)\left(\sqrt{2g\beta}-i\sqrt{\frac{g\beta}{2}}y^2\right.\right.$$
$$\left.\left.\left.-yk_y\right)-i\sqrt{2g\beta}\left(\frac{3}{6}-\frac{ik_y^2}{6g\beta}\right)y^2M\left(\frac{7}{4}-\frac{ik_y^2}{4g\beta},\frac{5}{2},\frac{i}{2}y^2\right)\right)\right], \tag{35}$$

$$\xi_1=\frac{E+k_z}{\lambda}\left[c_1\exp\left(\frac{-iy^2}{4}\right)M\left(\frac{1}{4}-\frac{ik_y^2}{4g\beta},\frac{1}{2},\frac{i}{2}y^2\right)+y\exp\left(\frac{i\pi}{4}\right)c_2\exp\left(\frac{-iy^2}{4}\right)\right.$$
$$\left.\times M\left(\frac{3}{4}-\frac{ik_y^2}{4g\beta},\frac{3}{2},\frac{i}{2}y^2\right)\right], \tag{36}$$

$$\xi_2=-i\frac{\exp\left(\frac{-iy^2}{4}\right)}{\sqrt{\frac{g\beta}{2}}y}\left[c_1\left(M\left(\frac{1}{4}-\frac{ik_y^2}{4g\beta},\frac{1}{2},\frac{i}{2}y^2\right)\left(i\sqrt{\frac{g\beta}{2}}y+k_y\right)-iy\sqrt{2g\beta}\left(\frac{1}{2}-\frac{ik_y^2}{2g\beta}\right)\right.\right.$$
$$\left.\times M\left(\frac{5}{4}-\frac{ik_y^2}{4g\beta},\frac{3}{2},\frac{i}{2}y^2\right)\right)-c_2\exp\left(\frac{i\pi}{4}\right)\left(M\left(\frac{3}{4}-\frac{ik_y^2}{4g\beta},\frac{3}{2},\frac{i}{2}y^2\right)\right.$$
$$\left.\left.\times\left(\sqrt{2g\beta}-i\sqrt{\frac{g\beta}{2}}y^2-yk_y\right)-i\sqrt{2g\beta}\left(\frac{3}{6}-\frac{ik_y^2}{6g\beta}\right)y^2M\left(\frac{7}{4}-\frac{ik_y^2}{4g\beta},\frac{5}{2},\frac{i}{2}y^2\right)\right)\right]. \tag{37}$$

(iii) The magnetic field H_z is inversely proportional to the spatial coordinate x: $H_z(x)=\beta/x$. Substituting the field expression into (14) and taking into account (12), we obtain

$$(d_x+k_y)\eta_2-i\left(\frac{g\beta}{x}+\lambda+im\right)\left(\frac{E+k_z}{\lambda}\right)\eta_1=0, \tag{38}$$

$$(d_x-k_y)\eta_1-i\left(\frac{g\beta}{x}+\lambda+im\right)\left(\frac{E-k_z}{\lambda}\right)\eta_2=0. \tag{39}$$

If we eliminate η_2 from (39) and (38), we find a second-order differential equation as follows:

$$\left[\frac{d^2}{dx^2}-k_y^2+\left(\frac{g\beta}{x}+\lambda+im\right)^2\right]\eta_1=0, \tag{40}$$

whose solution can be expressed in terms of confluent hypergeometric functions as follows:

$$\eta_1=\exp\left(\frac{-y}{2}\right)\left[a_1y^{1/2+\mu}M\left(\frac{1}{2}+\mu-\tilde{k},1+2\mu,y\right)+a_2y^{1/2-\mu}M\left(\frac{1}{2}-\mu-\tilde{k},1-2\mu,y\right)\right], \tag{41}$$

where

$$y=2\sqrt{k_y^2-(\lambda+im)^2}x, \quad \tilde{k}=\frac{g\beta(\lambda+im)}{\sqrt{k_y^2-(\lambda+im)^2}},$$

$$\mu=\sqrt{\frac{1}{4}-g^2\beta^2}. \tag{42}$$

Inserting (41) into (39), and using (12) and the recurrence relation (34), the other components of the 4-spinor are explicitly determined as

$$\begin{aligned}\eta_2=&\frac{-i\lambda\tilde{k}y\exp\left(\frac{-y}{2}\right)}{(2g^2\beta^2+\tilde{k}y)(\lambda+im)(E-k_z)}\left[a_1\left(M\left(\frac{1}{2}+\mu-\tilde{k},1+2\mu,y\right)\right.\right.\\&\times\left(\frac{-g\beta(\lambda+im)}{\tilde{k}}y^{1/2+\mu}+\frac{2g\beta(\lambda+im)}{\tilde{k}}\left(\frac{1}{2}+\mu\right)y^{-1/2+\mu}-k_yy^{1/2+\mu}\right)\\&\left.+\frac{2g\beta(\lambda+im)}{\tilde{k}}\frac{(\frac{1}{2}+\mu-\tilde{k})}{1+2\mu}y^{1/2+\mu}M\left(\frac{3}{2}+\mu-\tilde{k},2+2\mu,y\right)\right)+a_2\left(M\left(\frac{1}{2}-\mu-\tilde{k},1-2\mu,y\right)\right.\\&\times\left(\frac{-g\beta(\lambda+im)}{\tilde{k}}y^{1/2-\mu}+\frac{2g\beta(\lambda+im)}{\tilde{k}}\left(\frac{1}{2}-\mu\right)y^{-1/2-\mu}-k_yy^{1/2-\mu}\right)\\&\left.\left.+\frac{2g\beta(\lambda+im)}{\tilde{k}}\frac{(\frac{1}{2}-\mu-\tilde{k})}{1-2\mu}y^{1/2-\mu}M\left(\frac{3}{2}-\mu-\tilde{k},2-2\mu,y\right)\right)\right],\end{aligned} \tag{43}$$

$$\xi_1=\frac{(E+k_z)\exp\left(\frac{-y}{2}\right)}{\lambda}\left[a_1y^{1/2+\mu}M\left(\frac{1}{2}+\mu-\tilde{k},1+2\mu,y\right)+a_2y^{1/2-\mu}M\left(\frac{1}{2}-\mu-\tilde{k},1-2\mu,y\right)\right], \tag{44}$$

$$\xi_2=\frac{i\tilde{k}y\exp\left(\frac{-y}{2}\right)}{(2g^2\beta^2+\tilde{k}y)(\lambda+im)}\left[a_1\left(M\left(\frac{1}{2}+\mu-\tilde{k},1+2\mu,y\right)\left(\frac{-g\beta(\lambda+im)}{\tilde{k}}y^{1/2+\mu}\right.\right.\right.$$

$$+\frac{2g\beta(\lambda+im)}{\tilde{k}}\left(\frac{1}{2}+\mu\right)y^{-1/2+\mu}-k_y y^{1/2+\mu}\Bigg)$$

$$+\frac{2g\beta(\lambda+im)}{\tilde{k}}\frac{(\frac{1}{2}+\mu-\tilde{k})}{1+2\mu}y^{1/2+\mu}M\left(\frac{3}{2}+\mu-\tilde{k},2+2\mu,y\right)\Bigg)$$

$$+a_2\left(M\left(\left(\frac{1}{2}-\mu-\tilde{k},1-2\mu,y\right)\left(\frac{-g\beta(\lambda+im)}{\tilde{k}}y^{1/2-\mu}\right.\right.\right.$$

$$+\frac{2g\beta(\lambda+im)}{\tilde{k}}\left(\frac{1}{2}-\mu\right)y^{-1/2-\mu}-k_y y^{1/2-\mu}\Bigg)$$

$$\left.+\frac{2g\beta(\lambda+im)}{\tilde{k}}\frac{(\frac{1}{2}-\mu-\tilde{k})}{1-2\mu}y^{1/2-\mu}M\left(\frac{3}{2}-\mu-\tilde{k},2-2\mu,y\right)\Bigg)\Bigg)\right]. \tag{45}$$

(iv) The magnetic field depends exponentially on the spatial coordinate x: $H_z(x)=\beta\exp\alpha x$. Here, substituting field expression into (14) and taking into account (12), we obtain

$$(d_x+k_y)\eta_2-i(g\beta\exp\alpha x+\lambda+im)\left(\frac{E+k_z}{\lambda}\right)\eta_1=0, \tag{46}$$

$$(d_x-k_y)\eta_1-i(g\beta\exp\alpha x+\lambda+im)\left(\frac{E-k_z}{\lambda}\right)\eta_2=0. \tag{47}$$

Making the change of variables $v=\exp\alpha x$, and eliminating η_2 from (47) and (46), leads to

$$\left[\frac{d^2}{dv^2}+\frac{1}{v}\frac{d}{dv}+\frac{(\lambda+im)^2-k_y^2}{\alpha^2v^2}+\frac{2g\beta(\lambda+im)}{\alpha^2v}+\frac{g^2\beta^2}{\alpha^2}\right]\eta_1=0. \tag{48}$$

Writing

$$\eta_1=c(v)\zeta, \tag{49}$$

reduces Eq. (48) to a canonic form:

$$\left[\frac{d^2}{dv^2}+q(v)\right]\zeta=0. \tag{50}$$

Then, the change of variables $z=(-2ig\beta/\alpha)v$ permits us to obtain the solution of the Whittaker equation[31] as follows:

$$\eta_1=a_1\exp\left(\frac{ig\beta}{\alpha}v\right)v^{\mu}M\left(\frac{1}{2}+\mu-k,1+2\mu,\frac{-2ig\beta}{\alpha}v\right)$$

$$+a_2\exp\left(\frac{ig\beta}{\alpha}v\right)v^{-\mu}M\left(\frac{1}{2}-\mu-k,1-2\mu,\frac{-2ig\beta}{\alpha}v\right), \tag{51}$$

where

$$\mu=\sqrt{\frac{k_y^2-(\lambda+im)^2}{\alpha^2}},\quad k=\frac{(i\lambda-m)}{\alpha}. \tag{52}$$

Putting (51) into (47), and using (12) and the recurrence relation (34), we explicitly determine the other components of Φ_χ:

$$\eta_2=\frac{-i\lambda\,\exp\left(\frac{ig\beta}{\alpha}v\right)}{(g\beta v+\lambda+im)(E-k_z)}\Bigg[a_1M\left(\frac{1}{2}+\mu-k,1+2\mu,\frac{-2ig\beta}{\alpha}v\right)(ig\beta v^{1+\mu}+\alpha\mu v^{\mu}-k_yv^{\mu})$$
$$+a_1\alpha' v^{1+\mu}\left(\frac{\frac{1}{2}+\mu-k}{1+2\mu}\right)M\left(\frac{3}{2}+\mu-k,2+2\mu,\frac{-2ig\beta}{\alpha}v\right)$$
$$+a_2M\left(\frac{1}{2}-\mu-k,1-2\mu,\frac{-2ig\beta}{\alpha}v\right)(ig\beta v^{1-\mu}-\alpha\mu v^{-\mu}-k_yv^{-\mu})$$
$$+a_2\alpha' v^{1-\mu}\left(\frac{\frac{1}{2}-\mu-k}{1-2\mu}\right)M\left(\frac{3}{2}-\mu-k,2-2\mu,\frac{-2ig\beta}{\alpha}v\right)\Bigg], \tag{53}$$

$$\xi_1=\frac{E+k_z}{\lambda}\Bigg[a_1\exp\left(\left(\frac{ig\beta}{\alpha}v\right)v^{\mu}M\left(\frac{1}{2}+\mu-k,1+2\mu,\frac{-2ig\beta}{\alpha}v\right)\right.$$
$$+a_2\exp\left(\frac{ig\beta}{\alpha}v\right)v^{-\mu}M\left(\frac{1}{2}-\mu-k,1-2\mu,\frac{-2ig\beta}{\alpha}v\right)\Bigg], \tag{54}$$

$$\xi_2=\frac{i\,\exp\left(\frac{ig\beta}{\alpha}v\right)}{(g\beta v+\lambda+im)}\Bigg[a_1M\left(\frac{1}{2}+\mu-k,1+2\mu,\frac{-2ig\beta}{\alpha}v\right)(ig\beta v^{1+\mu}+\alpha\mu v^{\mu}-k_yv^{\mu})$$
$$+a_1\alpha' v^{1+\mu}\left(\frac{\frac{1}{2}+\mu-k}{1+2\mu}\right)M\left(\frac{3}{2}+\mu-k,2+2\mu,\frac{-2ig\beta}{\alpha}v\right)$$
$$+a_2M\left(\frac{1}{2}-\mu-k,1-2\mu,\frac{-2ig\beta}{\alpha}v\right)(ig\beta v^{1-\mu}-\alpha\mu v^{-\mu}-k_yv^{-\mu})$$
$$+a_2\alpha' v^{1-\mu}\left(\frac{\frac{1}{2}-\mu-k}{1-2\mu}\right)M\left(\frac{3}{2}-\mu-k,2-2\mu,\frac{-2ig\beta}{\alpha}v\right)\Bigg], \tag{55}$$

where $\alpha'=-2ig\beta$.

III. CYLINDRICAL COORDINATES

The Dirac equation in cylindrical coordinates for the neutrino of mass m in an external magnetic field, expressed in chiral representation, reads as

$$\left\{\gamma_\chi^0\partial_t+\gamma_\chi^1\partial_r+\frac{\gamma_\chi^2}{r}\partial_\theta+\gamma_\chi^3\partial_z+m+g\gamma_\chi^1\gamma_\chi^2H\right\}\Psi_\chi=0. \tag{56}$$

Following the pairwise scheme of variable separation, we transform Eq. (56) to

$$\left[-\gamma_\chi^2\partial_r+\frac{\gamma_\chi^1}{r}\partial_\theta+m\gamma_\chi^1\gamma_\chi^2-gH+\lambda\right]\Phi_\chi=0, \tag{57}$$

$$[\gamma_\chi^0\gamma_\chi^1\gamma_\chi^2\partial_t+\gamma_\chi^3\gamma_\chi^1\gamma_\chi^2\partial_z-\lambda]\Phi_\chi=0, \tag{58}$$

where Φ_χ is related to Ψ_χ by

$$\Psi_\chi=\gamma_\chi^1\gamma_\chi^2\Phi_\chi. \tag{59}$$

From Eq. (58), we find the following relations:

$$\lambda^2=E^2-k_z^2, \quad \Phi_\chi=\begin{pmatrix}\xi_1\\ \xi_2\\ \eta_1\\ \eta_2\end{pmatrix}=\begin{pmatrix}-\dfrac{E+k_z}{\lambda}\eta_1\\ \dfrac{E-k_z}{\lambda}\eta_2\\ \eta_1\\ \eta_2\end{pmatrix}. \tag{60}$$

The Eq. (57) splits into a system of ordinary differential equations, mixing the two-component spinors η and ξ as follows:

$$\left(\sigma^2 d_r-\frac{ik_\theta}{r}\sigma^1\right)\eta-(gH-\lambda+im)\xi=0, \tag{61}$$

$$\left(-\sigma^2 d_r+\frac{ik_\theta}{r}\sigma^1\right)\xi-(gH-\lambda+im)\eta=0, \tag{62}$$

involving the new system of equations,

$$\left(d_r+\frac{k_\theta}{r}\right)\eta_2+i(gH-\lambda+im)\left(\frac{E+k_z}{\lambda}\right)\eta_1=0, \tag{63}$$

$$\left(d_r-\frac{k_\theta}{r}\right)\eta_1+i(gH-\lambda+im)\left(\frac{E-k_z}{\lambda}\right)\eta_2=0, \tag{64}$$

which are more adapted to our situation.

Let us now analyze its solutions in terms of special functions, providing a concrete prescription for the external magnetic field.

(i) H is a constant magnetic field: $H=\beta$.

From (64), we have

$$\eta_2=\frac{i\lambda}{(g\beta-\lambda+im)(E-k_z)}\left(d_r-\frac{k_\theta}{r}\right)\eta_1, \tag{65}$$

which we insert into (63) to obtain a second-order differential equation for η_2,

$$\left[\frac{d^2}{dr^2}-\frac{k_\theta(k_\theta-1)}{r^2}+(g\beta-\lambda+im)^2\right]\eta_1=0, \tag{66}$$

whose solution can be written as follows:

$$\eta_1=cr^{1/2}H^{(1)}_{k_\theta-1/2}[(g\beta-\lambda+im)r]; \tag{67}$$

$H_\alpha^{(1)}(r)$ is the Hankel function of the first kind. Substituting (67) into (65), and using (60) and the following recurrence relation:

$$H'^{(1)}_{\mu}(z)=H^{(1)}_{\mu-1}(z)-\frac{\mu}{z}H^{(1)}_{\mu}(z), \tag{68}$$

we completely defined the three other components of the 4-spinor Φ_χ:

$$\eta_2=\frac{i\lambda}{E-k_z}cr^{-1/2}H^{(1)}_{k_\theta+1/2}[(g\beta-\lambda+im)r], \tag{69}$$

$$\xi_1=-\frac{E+k_z}{\lambda}cr^{1/2}H^{(1)}_{k_\theta-1/2}[(g\beta-\lambda+im)r], \tag{70}$$

$$\xi_2=icr^{1/2}H^{(1)}_{k_\theta+1/2}[(g\beta-\lambda+im)r]. \tag{71}$$

(ii) The magnetic field H is inversely proportional to the spatial coordinate r: $H=\alpha/r$. Substituting the field expression into (63) and (64), and if eliminating η_2 from these equations, we find a second-order differential equation:

$$\left[\frac{d^2}{dr^2}-\frac{k_\theta(k_\theta-1)-g^2\alpha^2}{r^2}-2(\lambda-im)g\frac{\alpha}{r}+(\lambda-im)^2\right]\eta_1=0. \tag{72}$$

Then, the change of variables $\rho=2i(\lambda-im)r$ permits us to obtain the solution

$$\eta_1=c_0\exp\left(\frac{-\rho}{2}\right)\rho^{1/2+\mu}M\left(\frac{1}{2}+\mu-k,1+2\mu,\rho\right)+c_1\exp\left(\frac{-\rho}{2}\right)\rho^{1/2-\mu}M\left(\frac{1}{2}-\mu-k,1-2\mu,\rho\right), \tag{73}$$

where

$$\mu=\sqrt{k_\theta(k_\theta-1)-g^2\alpha^2+\tfrac{1}{4}},\quad k=ig\alpha. \tag{74}$$

Inserting (73) into (63), and using (60) and the recurrence relation (34), we have the three other components of the 4-spinor Φ_χ,

$$\begin{aligned}\eta_2=\frac{-2\lambda\rho\exp\left(\frac{-\rho}{2}\right)}{(2ig\alpha-\rho)(E-k_z)}\Bigg[&c_0\Bigg(M\left(\frac{1}{2}+\mu-k,1+2\mu,\rho\right)\left(\frac{-1}{2}\rho^{1/2+\mu}+\left(\frac{1}{2}+\mu\right)\rho^{\mu-1/2}-k_\theta\rho^{\mu-1/2}\right)\\&+\frac{\frac{1}{2}+\mu-k}{1+2\mu}\rho^{1/2+\mu}M\left(\frac{3}{2}+\mu-k,2+2\mu,\rho\right)\Bigg)\\&+c_1\Bigg(M\left(\frac{1}{2}-\mu-k,1-2\mu,\rho\right)\left(\frac{-1}{2}\rho^{1/2-\mu}+\left(\frac{1}{2}-\mu\right)\rho^{-\mu-1/2}-k_\theta\rho^{-\mu-1/2}\right)\\&+\frac{\frac{1}{2}-\mu-k}{1-2\mu}\rho^{1/2-\mu}M\left(\frac{3}{2}-\mu-k,2-2\mu,\rho\right)\Bigg)\Bigg],\end{aligned} \tag{75}$$

$$\xi_1 = -\frac{E+k_z}{\lambda}\left[c_0 \exp\left(\frac{-\rho}{2}\right)\rho^{1/2+\mu} M\left(\frac{1}{2}+\mu-k,1+2\mu,\rho\right) + c_1 \exp\left(\frac{-\rho}{2}\right)\rho^{1/2-\mu} M\left(\frac{1}{2}-\mu-k,1-2\mu,\rho\right)\right], \tag{76}$$

$$\xi_2 = \frac{-2\rho\exp\left(\frac{-\rho}{2}\right)}{2ig\alpha-\rho}\left[c_0\left(M\left(\frac{1}{2}+\mu-k,1+2\mu,\rho\right)\left(\frac{-1}{2}\rho^{1/2+\mu}+\left(\frac{1}{2}+\mu\right)\rho^{\mu-1/2}-k_\theta\rho^{\mu-1/2}\right) + \frac{\frac{1}{2}+\mu-k}{1+2\mu}\rho^{1/2+\mu}M\left(\frac{3}{2}+\mu-k,2+2\mu,\rho\right)\right) + c_1\left(M\left(\frac{1}{2}-\mu-k,1-2\mu,\rho\right)\left(\frac{-1}{2}\rho^{1/2-\mu}+\left(\frac{1}{2}-\mu\right)\rho^{-\mu-1/2}-k_\theta\rho^{-\mu-1/2}\right) + \frac{\frac{1}{2}-\mu-k}{1-2\mu}\rho^{1/2-\mu}M\left(\frac{3}{2}-\mu-k,2-2\mu,\rho\right)\right)\right]. \tag{77}$$

Now, we are going to solve the Dirac equation for a neutrino with anomalous electric interaction.

(iii) The electric field $\mathbf{E}=\epsilon$ is applied along the z axis.

The Dirac equation takes the form

$$\left\{ \gamma_\chi^0\partial_t + \gamma_\chi^1\partial_r + \frac{\gamma_\chi^2\partial_\theta}{r} + \gamma_\chi^3\partial_z + m + g\gamma_\chi^3\gamma_\chi^0\epsilon \right\}\Psi_\chi = 0. \tag{78}$$

Applying the pairwise scheme of variable separation, we can write Eq. (78) as follows:

$$\left(\gamma_\chi^1\gamma_\chi^3\gamma_\chi^0\partial_r + \frac{\gamma_\chi^2\gamma_\chi^3\gamma_\chi^0\partial_\theta}{r} + g\epsilon - \lambda \right)\Phi_\chi = 0, \tag{79}$$

$$(\gamma_\chi^3\partial_t + \gamma_\chi^0\partial_z - m\gamma_\chi^3\gamma_\chi^0 - \lambda)\Phi_\chi = 0, \tag{80}$$

where Φ_χ is related to Ψ_χ by

$$\Psi_\chi = \gamma_\chi^3\gamma_\chi^0\Phi_\chi. \tag{81}$$

From Eq. (80), we find the relations

$$\lambda^2 = E^2 - k_z^2 + m^2, \qquad \Phi_\chi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{-i(E+k_z)}{m-\lambda}\eta_1 \\ \frac{-i(E-k_z)}{m+\lambda}\eta_2 \\ \eta_1 \\ \eta_2 \end{pmatrix}. \tag{82}$$

The 4-spinor Eq. (79) splits into a system of ordinary differential equations, mixing the two-component spinors η and ξ as follows:

$$\left(\sigma^2\partial_r+\frac{\partial_\theta}{r}\sigma^1\right)\eta+i(g\epsilon-\lambda)\xi=0, \tag{83}$$

$$\left(\sigma^2\partial_r+\frac{\partial_\theta}{r}\sigma^1\right)\xi+i(g\epsilon-\lambda)\eta=0, \tag{84}$$

which can be transformed to a simple system,

$$\left(d_r-\frac{k_\theta}{r}\right)\eta_2+i(g\epsilon-\lambda)\left(\frac{E+k_z}{m-\lambda}\right)\eta_1=0, \tag{85}$$

$$\left(d_r+\frac{k_\theta}{r}\right)\eta_1-i(g\epsilon-\lambda)\left(\frac{E-k_z}{m+\lambda}\right)\eta_2=0. \tag{86}$$

If we eliminate η_2 from (86) and (85), we find a second-order ordinary differential equation,

$$\left[\frac{d^2}{dr^2}-\frac{k_\theta(k_\theta+1)}{r^2}+(g\epsilon-\lambda)^2\right]\eta_1=0, \tag{87}$$

whose solution is given by

$$\eta_1=cr^{1/2}H^{(1)}_{k_\theta+1/2}[(g\epsilon-\lambda)r]. \tag{88}$$

Substituting (88) into (86), and using (82) and the recurrence relation (68), we determine the other components of the 4-spinor Φ_χ:

$$\eta_2=\frac{i(m+\lambda)}{k_z-E}cr^{1/2}H^{(1)}_{k_\theta-1/2}[(g\epsilon-\lambda)r], \tag{89}$$

$$\xi_1=\frac{i(E+k_z)}{\lambda-m}cr^{1/2}H^{(1)}_{k_\theta+1/2}[(g\epsilon-\lambda)r], \tag{90}$$

$$\xi_2=cr^{1/2}H^{(1)}_{k_\theta-1/2}[(g\epsilon-\lambda)r]. \tag{91}$$

IV. DISCUSSIONS

In each case, the original Dirac spinor Ψ_χ in chiral representation should be faithfully recovered using relations (6) and (10) in Cartesian coordinates or (59), (81) in cylindrical coordinates and the explicit expressions for the 4-spinor Φ_χ components, obtained in the various system of coordinates for a specific applied field.

For instance, in the cylindrical coordinates, the explicit structure of the Dirac 4-spinor Ψ_χ for the neutrino with an anomalous electric interaction is given by

$$\Psi_\chi=\gamma_\chi^3\gamma_\chi^0\Phi_\chi=\begin{pmatrix}\xi_{1\chi}\\ \xi_{2\chi}\\ \eta_{1\chi}\\ \eta_{2\chi}\end{pmatrix}, \tag{92}$$

where

$$\xi_{1\chi}=\frac{i(E+k_z)}{m-\lambda}cr^{1/2}H^{(1)}_{k_\theta+1/2}[(g\epsilon-\lambda)r], \tag{93}$$

$$\xi_{2\chi}=-cr^{1/2}H^{(1)}_{k_\theta-1/2}[(g\epsilon-\lambda)r], \tag{94}$$

$$\eta_{1\chi}=cr^{1/2}H^{(1)}_{k_\theta+1/2}[(g\epsilon-\lambda)r], \tag{95}$$

$$\eta_{2\chi}=\frac{i(m+\lambda)}{E-k_z}cr^{1/2}H^{(1)}_{k_\theta-1/2}[(g\epsilon-\lambda)r]. \tag{96}$$

At the opposite of Cartesian and cylindrical coordinates that have been worked out to give standard ordinary differential equations whose solutions are obtained in terms of special mathematical functions for the set of magnetic field expressions, the system of spherical coordinates showed some difficulties in exhibiting exact finite physical analytical solutions. The introduction of spherical symmetry in a chiral representation has contributed to elevating the order of ordinary differential equations governing a single spinor component. From a second-order differential equation in the case of Cartesian and cylindrical coordinate systems, we return back to a more complicated greater order differential equation.

Indeed, in spherical coordinates, the Dirac equation for a neutrino with anomalous electric interaction takes the form

$$\left\{\gamma_\chi^0\partial_t+\gamma_\chi^1\partial_r+\frac{\gamma_\chi^2}{r}\partial_\theta+\frac{\gamma_\chi^3}{r\sin\theta}\partial_\varphi+m+g\gamma_\chi^1\gamma_\chi^0\mathbf{E}\right\}\Psi_\chi=0. \tag{97}$$

By the pairwise scheme of variable separation, we transform (97) to

$$\left[-\gamma_\chi^1\partial_t-\gamma_\chi^0\partial_r+m\gamma_\chi^1\gamma_\chi^0+g\mathbf{E}+\frac{k}{r}\right]\Phi_\chi=0, \tag{98}$$

$$\left[-\gamma_\chi^2\gamma_\chi^1\gamma_\chi^0\partial_\theta+\frac{\gamma_\chi^3\gamma_\chi^1\gamma_\chi^0}{\sin\theta}\partial_\varphi-k\right]\Phi_\chi=0, \tag{99}$$

and

$$\Psi_\chi=\gamma_\chi^1\gamma_\chi^0\Phi_\chi. \tag{100}$$

If we consider, for example, a field inversely varying with respect to the spatial coordinate r: $\mathbf{E}=\epsilon/r$, and substituting it into (98), we obtain

$$(i\sigma^1 E + d_r)\,\eta - \left(\frac{g\epsilon+k}{r} - m\sigma^1\right)\xi = 0, \tag{101}$$

$$(i\sigma^1 E - d_r)\xi + \left(\frac{g\epsilon+k}{r} + m\sigma^1\right)\eta = 0, \tag{102}$$

which can be reduced to

$$\left\{\frac{m}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\left[-\frac{d^2}{dr^2}+\frac{(g\epsilon+k)}{r^2}\left(iE+\frac{2m\frac{(g\epsilon+k)}{r}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\right)\frac{d}{dr}+\frac{\frac{2imE}{r}\frac{(g\epsilon+k)^3}{r^3}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}-E^2\right]+m\right\}$$
$$\times\eta_1+\left\{\frac{m}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\left[-\frac{(g\epsilon+k)}{r}\frac{d^2}{dr^2}+\frac{(g\epsilon+k)}{r^2}\left(1+\frac{2m\frac{(g\epsilon+k)^2}{r^2}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\right)\frac{d}{dr}\right.\right.$$
$$\left.\left.+\frac{g\epsilon+k}{r}\left(-E^2+\frac{\frac{2imE}{r}\frac{(g\epsilon+k)}{r}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\right)\right]+\frac{g\epsilon+k}{r}\right\}\eta_2=0, \tag{103}$$

and

$$\left\{\frac{m}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\left[-\frac{d^2}{dr^2}+\frac{(g\epsilon+k)}{r^2}\left(iE+\frac{2m\frac{(g\epsilon+k)}{r}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\right)\frac{d}{dr}+\frac{\frac{2imE}{r}\frac{(g\epsilon+k)^3}{r^3}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}-E^2\right]+m\right\}$$
$$\times\eta_2+\left\{\frac{m}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\left[-\frac{(g\epsilon+k)}{r}\frac{d^2}{dr^2}+\frac{(g\epsilon+k)}{r^2}\left(1+\frac{2m\frac{(g\epsilon+k)^2}{r^2}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\right)\frac{d}{dr}\right.\right.$$
$$\left.\left.+\frac{g\epsilon+k}{r}\left(-E^2+\frac{\frac{2imE}{r}\frac{(g\epsilon+k)}{r}}{\frac{(g\epsilon+k)^2}{r^2}-m^2}\right)\right]+\frac{g\epsilon+k}{r}\right\}\eta_1=0. \tag{104}$$

These equations can be solved using either numerical integration methods or infinite series solutions.

Thus, the spherical coordinates clearly show the limitation of using a chiral representation to find exact solutions in terms of special functions for a Dirac equation in the presence of external fields.

The results obtained in this paper prove again that the general algebraic method of the separation of variables developed by Shishkin *et al.*[7] constitutes one of the most appropriate and powerful techniques in searching for exact solutions to the Dirac equation.

ACKNOWLEDGMENTS

The authors are indebted to Professor J. Govaerts of the Unité de Physique Nucléaire (FYNU), the Université Catholique de Louvain (Belgium), for useful discussions. J. E. B. M. thanks the Belgian Cooperation Interuniversitaire (CIUF-CUD) for financial support of this work.

[1] G. V. Shishkin, and I. E. Andrushkevich, Theor. Math. Phys. **70**, 204 (1987).
[2] G. V. Shishkin, Trans. Inst. Phys. Estonian Acad. Sci. **65**, 144 (1989) (in Russian).
[3] G. V. Shishkin, *Gravity and Quantum Field Theory* (Nauka, Alma-Ata, 1989).
[4] G. V. Shishkin, in *Gravitation and Electromagnetism* (University Press, Minsk, 1990).
[5] G. V. Shishkin and W. D. Cabos, J. Math. Phys. **32**, 3184 (1991).
[6] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **30**, 2132 (1989).
[7] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **33**, 2093 (1992).
[8] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **33**, 4037 (1992).
[9] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **34**, 5037 (1993).
[10] V. M. Villalba and U. Percoco, J. Math. Phys. **31**, 715 (1990).
[11] V. M. Villalba, J. Phys. A **31**, 1483 (1990).
[12] V. M. Villalba, J. Phys. A **36**, 3332 (1995).
[13] D. Brill and J. Wheeler, Rev. Mod. Phys. **29**, 465 (1957).
[14] S. Chandrasekhar, Proc. R. Soc. London, Ser. A **349**, 571 (1976).
[15] A. H. Cook, Proc. R. Soc. London, Ser. A **383**, 247 (1982).
[16] B. Carter and R. G. Mc Lenaghan, Phys. Rev. D **19**, 1093 (1979).
[17] A. L. Dudley and J. D. Finlay, III, J. Math. Phys. **20**, 311 (1979).
[18] E. G. Kalnins, W. Miller, Jr., and G. C. Williams, J. Math. Phys. **27**, 1893 (1986).
[19] V. G. Bagrov, I. M. Ternov and Yu. I. Klimento, Izv. Vuz. Fiz. **2**, 50 (1968).
[20] V. G. Bagrov, I. M. Ternov, and P. V. Borizkov, Izv. Vuz. Fiz. **11**, 38 (1971).
[21] V. G. Bagrov, D. M. Gitman, M. Ternov, V. Khalikov, and V. N. Shapovalov, *Exact Solutions of Relativistic Wave Equations* (Nauka, Novosibirsk, 1982), (in Russian).
[22] V. G. Bagrov and D. M. Gitman, *Exact Solutions of Relativistic Wave Equations* (Kluwer, Dordrecht, 1990).
[23] A. O. Barut and I. H. Duru, Phys. Rev. D **36**, 3705 (1987).
[24] A. O. Barut and A. T. Alan, Europhys. Lett. **17**, 119 (1992).
[25] O. L. de Lange and R. E. Raab, *Operator Methods in Quantum Mechanics* (Clarendon, Oxford, 1991).
[26] F. Dominguez-Adame, J. Rev. A. **23**, 1993 (1990).
[27] F. Dominguez-Adame and M. A. González, Europhys. Lett. **23**, 193 (1990).
[28] F. Dominguez-Adame and B. Méndez, Nuovo Cimento B **107**, 489 (1992).
[29] A. Hautot, J. Math Phys. **14**, 2011 (1973).
[30] L. I. Komarov and T. S. Romanova, Izv. Acad. Nauk. BSSR **2**, 98 (1982).
[31] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1964).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Exactly solvable models of δ'-sphere interactions in nonrelativistic quantum mechanics

M. N. Hounkonnou[a)]
Institut de Mathématiques et de Sciences Physiques, Université Nationale du Bénin, B. P. 613 Porto-Novo, Bénin

M. Hounkpe
Institut de Mathématiques et de Sciences Physiques, Université Nationale du Bénin, B.P. 613 Porto-Novo, Bénin, and Department of Polical Science, Yale University, New Haven, Connecticut

J. Shabani[b)]
UNESCO, Regional office B. P. 3311, Dakar, Senegal

(Received 20 January 1999; accepted for publication 23 March 1999)

We introduce and perform a systematic study of a new exactly solvable model of sphere interactions in quantum mechanics : the δ' interaction, formally given by $-\Delta+\alpha\delta'(|\mathbf{x}|-R)$, $\mathbf{x}\in\mathbb{R}^3$, $R>0$, $\alpha\in\mathbb{R}$. We also consider the cases of a δ' plus a Coulomb interaction and finitely many δ'-sphere interactions with support on concentric spheres. For all these models, we provide basic properties and discuss the stationary scattering theory. We also briefly discuss the δ'-sphere interaction of the second type. [S0022-2488(99)02508-6]

I. INTRODUCTION

In recent years, there has been a lot of interest in the study of sphere interactions in quantum mechanics, both from the mathematical point of view and for their application in modelling physical phenomena.[1–11]

So far, these studies focused on the δ-sphere interactions and their various generalizations. In this paper we provide a new exactly solvable model of sphere interactions : the δ'-sphere interaction formally given in three dimensions by the Hamiltonian

$$H=-\Delta+\alpha\delta'(|\mathbf{x}|-R),\quad \mathbf{x}\in\mathbb{R}^3,\quad R>0,\quad \alpha\in\mathbb{R}.$$

As indicated in Ref. 12, this model is different from the δ-sphere interaction of the second type introduced in Refs. 1,3 and inadequately called the ''δ'-sphere interaction.''

The paper is organized as follows. In Sec. II, we give a rigorous mathematical definition of H using the theory of self-adjoint extensions of symmetric closed operators in Hilbert spaces. We obtain the basic properties of H, including the resolvent equation, the spectral properties, and the scattering data. In Secs. III and IV, we generalize the results of Sec. II to the case of a δ'-sphere plus a Coulomb interaction and finitely many δ'-sphere interactions, respectively. In Sec. V, we introduce and briefly discuss the δ'-sphere interaction of the second type.

II. THE δ'-SPHERE INTERACTION

A. Definition of the Hamiltonian

Consider the radial Schrödinger equation for a δ'-sphere interaction given by the formal expression

[a)]Electronic mail: hounkon@syfed.bj.refer.org
[b)]On leave of absence from University of Burundi, Faculty of Science, B. P. 2700 Bujumbura, Burundi.

$$h_{l,\alpha_l} f_l(k,r) \equiv \left[-\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + \alpha_l \delta'(r-R) \right] f_l(k,r) = k^2 f_l(k,r), \tag{1}$$

and assume that the function $f_l(k,r)$ is continuous at $r=R$, i.e.,

$$f_l(k,R_+) = f_l(k,R_-) \equiv f_l(k,R). \tag{2}$$

This assumption implies that the derivative $f_l'(k,r)$ is discontinuous at $r=R$, since otherwise the operator h_{l,α_l} would coincide with the free Hamiltonian corresponding to the partial wave l.

Therefore, at $r=R$, the function $f_l'(k,r)$ may be defined by

$$f_l'(k,R) \equiv \tfrac{1}{2}[f_l'(k,R_+) + f_l'(k,R_-)]. \tag{3}$$

Let us integrate Eq. (1) from $r=R-\varepsilon$ to $r=R+\varepsilon$ and take the limit when $\varepsilon \to 0_+$. We obtain the following boundary condition:

$$\left(1+\frac{\alpha_l}{2}\right) f_l'(k,R_+) - \left(1-\frac{\alpha_l}{2}\right) f_l'(k,R_-) = 0. \tag{4}$$

From the above discussion, it follows that a δ'-sphere interaction may be fully characterized by the boundary conditions (2) and (4).

Let us now provide a rigorous mathematical definition of a quantum Hamiltonian describing a δ'-sphere interaction.

Consider in $L^2(\mathbb{R}^3)$ the closed and non-negative operator,

$$\dot{H} = -\Delta,$$

$$\mathcal{D}(\dot{H}) = \{ f \in H^{2,2}(\mathbb{R}^3) | f(\partial \overline{K(O,R)) = f'}(\partial \overline{K(O,R)) = 0} \}, \tag{5}$$

where $H^{m,n}(\Omega)$ is the Sobolev space of indices (m,n) and $\overline{K(O,R)}$ is the closed ball of radius R centered at the origin of $\mathbb{R}^3$.

We decompose $L^2(\mathbb{R}^3)$ with respect to angular momenta,

$$L(\mathbb{R}^3) = L^2((0,\infty); r^2 dr) \otimes L^2(S^2), \tag{6}$$

and introduce the unitary transformation,

$$U: \begin{cases} L^2((0,\infty); r^2 dr) \to L^2((0,\infty); dr) \equiv L^2((0,\infty)), \\ f \mapsto (Uf)(r) = r f(r), \end{cases} \tag{7}$$

in order to get the following representation of $L^2(\mathbb{R}^3)$:

$$L^2(\mathbb{R}^3) = \bigoplus_{l=0}^{\infty} U^{-1} L^2((0,\infty)) \otimes [Y_l^{-l}, \dots, Y_l^l], \tag{8}$$

where $[\cdots]$ denotes the linear span of spherical harmonics.

With respect to the decomposition (8), $\dot{H}$ reads as

$$\dot{H} = \bigoplus_{l=0}^{\infty} U^{-1} \dot{h}_l U \otimes 1, \tag{9}$$

where

$$\dot{h}_l = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2},$$

$$\mathcal{D}(\dot{h}_l) = \{f \in L^2((0,\infty))/f, \quad f' \in AC_{\rm loc}((0,\infty)); \quad f(0_+)=0, \quad \text{if } l=0;$$
$$f(R_\pm) = f'(R_\pm) = 0; \quad -f'' + l(l+1)r^{-2}f \in L^2((0,\infty))\}; \quad l \in \mathbb{N}_0, \tag{10}$$

and $AC_{\rm loc}(\Omega)$ denotes the set of the locally absolutely continuous functions on $\Omega \subset \mathbb{R}$.

Thus, the adjoint $\dot{H}^\star$ of $\dot{H}$ is

$$\dot{H}^\star = \bigoplus_{l=0}^{\infty} U^{-1}\dot{h}_l^\star U \otimes 1, \tag{11}$$

where

$$\dot{h}_l^\star = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2},$$

$$\mathcal{D}(\dot{h}_l^\star) = \{f \in L^2((0,\infty))/f, f' \in AC_{\rm loc}((0,\infty)\setminus\{R\}); \quad f(0_+)=0, \quad \text{if } l=0;$$
$$-f'' + l(l+1)r^{-2}f \in L^2((0,\infty))\}; \; l \in \mathbb{N}_0. \tag{12}$$

In particular, the equation

$$\dot{h}_l^\star \psi_l(k) = k^2 \psi_l(k), \;\; \text{Im}(k) > 0, \;\; l \in \mathbb{N}_0, \tag{13}$$

has two linearly independent solutions,

$$\psi_l^{(1)}(k,r) = \begin{cases} G_l(k,R)F_l(k,r) & ; \quad r<R, \\ 0 & ; \quad r>R, \end{cases} \tag{14}$$

$$\psi_l^{(2)}(k,r) = \begin{cases} 0 & ; \quad r<R, \\ F_l(k,R)G_l(k,r) & ; \quad r>R, \end{cases} \tag{15}$$

where

$$F_l(k,r) = \Gamma\left(l+\frac{3}{2}\right)\left(\frac{k}{2}\right)^{-l-1/2} r^{1/2} J_{l+1/2}(kr), \tag{16}$$

$$G_l(k,r) = \frac{i\pi}{2}\Gamma\left(l+\frac{3}{2}\right)^{-1}\left(\frac{k}{2}\right)^{l+1/2} r^{1/2} H_{l+1/2}^{(1)}(kr), \tag{17}$$

and $J_\nu(.)$, [resp., $H_\nu(\cdot)$] denote the Bessel (resp., Hankel) functions of order ν.[13] Therefore $\dot{h}_l$ has deficiency indices (2,2), and consequently all self-adjoint (s.a.) extensions of $\dot{h}_l$ are given by a 4-parameter family of s.a. operators.[14]

In this paper, we consider a special one-parameter family h_{l,α_l} of s.a. extensions of $\dot{h}_l$ defined by

$$h_{l,\alpha_l} = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2},$$

$$\mathcal{D}(h_{l,\alpha_l})=\Big\{ f\in L^2((0,\infty))/f,f'\in AC_{\mathrm{loc}}((0,\infty)\backslash\{R\});$$
$$f(0_+)=0\,,\quad \text{if}\quad l=0\,;\qquad f(R_+)=f(R_-)=f(R);$$
$$\left(1+\frac{\alpha_l}{2}\right)f'(R_+)-\left(1-\frac{\alpha_l}{2}\right)f'(R_-)=0;$$
$$-f''+l(l+1)r^{-2}f\in L^2((0,\infty))\Big\}\,;\qquad -\infty<\alpha_l<\infty\,,\qquad l\in\mathbb{N}_0 \tag{18}$$

The case $\alpha_l=0$ coincides with the free kinetic energy Hamiltonian $h_{l,o}$ for a fixed angular momentum l.

Let $\alpha=\{\alpha_l\}_{l\in\mathbb{N}_0}$ and introduce in $L^2(\mathbb{R}^3)$ the operator

$$H_\alpha=\bigoplus_{l=0}^{\infty} U^{-1}h_{l,\alpha_l}U\otimes 1\,. \tag{19}$$

By definition, H_α is the rigorous mathematical formulation of the formal expression $H=-\Delta+\alpha\delta'(|x|-R)$. Actually, it provides a slight generalization of H, since α may depend on $l\in\mathbb{N}_0$.

The case $\alpha=0$ leads to the free Hamiltonian,

$$H_o=-\Delta,\qquad \mathcal{D}(H_o)=H^{2,2}(\mathbb{R}^3). \tag{20}$$

Next, we introduce the free resolvent,

$$g_{l,k}=(h_{l,o}-k^2)^{-1},\qquad \mathrm{Im}(k)>0, \tag{21}$$

with integral kernel

$$g_{l,k}(r,r')=\begin{cases} F_l(k,r')G_l(k,r) & ;\quad r'\leqslant r,\\ F_l(k,r)G_l(k,r') & ;\quad r'\geqslant r.\end{cases} \tag{22}$$

B. The resolvent equation

Theorem 2.1:

(i) The resolvent of h_{l,α_l} is given by

$$(h_{l,\alpha_l}-k^2)^{-1}=(h_{l,o}-k^2)^{-1}+\lambda_l(k)(\tilde{\phi}_l(-\overline{k}),.)\phi_l(k),$$
$$k^2\in\rho(h_{l,\alpha_l})\,,\qquad \mathrm{Im}(k)>0\,,\qquad l\in\mathbb{N}_0, \tag{23}$$

where

$$\lambda_l(k)=\alpha_l\left[1-\frac{\alpha_l}{2}g'_{l,k}(R,R)\right]^{-1}, \tag{24}$$

$$\phi_l(k,r)=\begin{cases} G_l(k,R)F_l(k,r) & ;\quad r\leqslant R,\\ F_l(k,R)G_l(k,r) & ;\quad r\geqslant R,\end{cases} \tag{25}$$

$$\tilde{\phi}_l(k,r)=\begin{cases} G'_l(k,R)F_l(k,r) & ;\quad r< R,\\ F'_l(k,R)G_l(k,r) & ;\quad r> R,\end{cases} \tag{26}$$

we note that $\phi_l(k,r)=g_{l,k}(R,r)$, $\mathrm{Im}(k)>0$.

(ii) The resolvent of H_α is given by

$$(H_\alpha-k^2)^{-1}=(H_o-k^2)^{-1}+\bigoplus_{l=0}^{\infty}\bigoplus_{m=-l}^{l}\lambda_l(k)\times(|\cdot|^{-1}\tilde{\phi}_l(-\overline{k})Y_l^m,\cdot)|\cdot|^{-1}\phi_l(k)Y_l^m$$

$$k^2\in\rho(H_\alpha),\quad \mathrm{Im}(k)>0. \tag{27}$$

Proof: Since $\dot{h}_l$ has deficiency indices (2,2), it follows from Krein's formula[14] that the resolvent of h_{l,α_l} is given by

$$(h_{l,\alpha_l}-k^2)^{-1}=(h_{l,o}-k^2)^{-1}+\sum_{i,j=1}^{2}\mu_{ij}(k)(\psi_l^{(j)}(-\overline{k}),.)\psi_l^{(i)}(k),$$

$$k^2\in\rho(h_{l,\alpha_l}),\quad \mathrm{Im}(k)>0,\quad l\in\mathbb{N}_0, \tag{28}$$

where $\psi_l^{(m)}$; $m=1,2$ are defined by (14) and (15), respectively.

For the determination of the μ_{ij}, we proceed as follows. Let $g\in L^2((0,\infty))$ and define the function

$$\chi_l(k,r)=((h_{l,\alpha_l}-k^2)^{-1}g)(r). \tag{29}$$

Since $\chi_l\in\mathcal{D}(h_{l,\alpha_l})$, it follows that χ_l should satisfy the boundary conditions in (18).

The implementation of these boundary conditions gives

$$\mu(k)=\frac{\alpha_l}{1-\dfrac{\alpha_l}{2}g'_{l,k}(R,R)}\begin{pmatrix}G_l(k,R)G'_l(k,R) & G_l(k,R)F'_l(k,R)\\ F_l(k,R)G'_l(k,R) & F_l(k,R)F'_l(k,R)\end{pmatrix}. \tag{30}$$

Inserting (30) into (28), we reduce the expression (28) to (23).

Equation (27) follows from (19) and (23).

We note that $\det[\mu(k)]=0$. This means that $\dot{h}_l$ is not the maximal commun part of h_{l,α_l} and $h_{l,o}$.

Next, we should like to provide some additional information on the domain of h_{l,α_l} and to show that the δ'-sphere interaction is in fact a local interaction.

Theorem 2.2: The domain $\mathcal{D}(h_{l,\alpha_l})$ consists of functions of the type

$$\varphi_l(k,r)=F_l(k,r)+\lambda_l(k)F'_l(k,R)g_{l,k}(R,r), \tag{31}$$

where $\lambda_l(k)$ is defined by (24), $F_l\in\mathcal{D}(h_{l,o})$, and $k^2\in\rho(h_{l,\alpha_l})$, $\mathrm{Im}(k)>0$.

The decomposition (31) is unique and with $\varphi_l\in\mathcal{D}(h_{l,\alpha_l})$ of this form, we obtain

$$(h_{l,\alpha_l}-k^2)\varphi_l=(h_{l,o}-k^2)F_l. \tag{32}$$

Furthermore, if $\varphi_l\in\mathcal{D}(h_{l,\alpha_l})$ and $\varphi_l=0$ in an open set $\mathcal{O}\subset(0,\infty)$, then $h_{l,\alpha_l}\varphi_l=0$ in $\mathcal{O}$ which means that the δ'-sphere interaction is a local interaction.

Proof: One may follow step by step[15] where a similar result was obtained for point interactions.

C. Spectral properties

Spectral properties of h_{l,α_l} are provided by the following theorem where $\sigma(.)$, $\sigma_{\mathrm{ess}}(.)$, $\sigma_{\mathrm{ac}}(.)$, $\sigma_{\mathrm{sc}}(.)$ and $\sigma_p(.)$ denote the spectrum, essential spectrum, absolutely continuous spectrum, singularly continuous spectrum, and point spectrum, respectively.

Theorem 2.3: For all $\alpha_l \in (-\infty,\infty)$, we have the following results:

- $$\sigma_{\mathrm{ess}}(h_{l,\alpha_l}) = \sigma_{\mathrm{ac}}(h_{l,\alpha_l}) = [0,\infty), \tag{33}$$
- $$\sigma_{\mathrm{sc}}(h_{l,\alpha_l}) = \emptyset, \tag{34}$$
- $$\sigma_p(h_{l,\alpha_l}) \cap [0,\infty) = \emptyset. \tag{35}$$

Negative eigenvalues of h_{l,α_l} are obtained from the equation

$$1 - \frac{\alpha_l}{2} g'_{l,i\sqrt{-E}}(R,R) = 1 - \frac{\alpha_l}{2} \frac{d}{dr}[rI_{l+1/2}(\sqrt{-E}r)K_{l+1/2}(\sqrt{-E}r)]_{r=R} = 0; \quad E<0, \tag{36}$$

which implies that

$$\sigma_p(h_{l,\alpha_l}) = \begin{cases} \emptyset, & \text{if} \quad \alpha_l \leqslant 2(2l+1), \\ \{E_o\}, & \text{if} \quad \alpha_l > 2(2l+1), \end{cases} \tag{37}$$

where E_o is a solution of (36) and $I_\nu(\cdot)$, $K_\nu(\cdot)$ are modified Bessel functions.[13]

Proof: Equations (33) and (34) follow from Weyl's theorem Ref. 16, p. 112 and Ref. 16, theorem XIII.20, respectively.

For $k \geqslant 0$, consider the function $f_l(k,r) \in \mathcal{D}(h_{l,\alpha_l})$. By inspection, we may show that the equation

$$-f_l''(k,r) + l(l+1)r^{-2} f_l(k,r) = k^2 f_l(k,r) \tag{38}$$

can be solved uniquely in terms of Bessel functions which do not belong to $L^2((0,\infty))$. This yields Eq. (35).

The bound state equation (36) obtained by using a Bessel function ansatz in (38) and the boundary conditions in (18) can be analyzed using monotonicity properties of the modified Bessel functions, and indeed one proves Eq. (37).

Next we briefly discuss resonances of h_{l,α_l}.

Following, e.g., Ref. 17 we define the resonances of h_{l,α_l} as the poles of the resolvent (23) in the unphysical sheet $\mathrm{Im}(k)<0$, i.e., as solutions of the equation

$$1 - \frac{\alpha_l}{2} g'_{l,k}(R,R) = 0, \quad \mathrm{Im}(k)<0. \tag{39}$$

First we discuss the solutions located on the negative imaginary k axis. Let $k=ix$, $x>0$. Then analytic continuation of Bessel functions in (39) yields

$$-\frac{2}{\alpha_l} = \frac{d}{dr}[rI_{l+1/2}(xr)K_{l+1/2}(xr)]_{r=R}. \tag{40}$$

This equation can be analyzed using monotonicity properties of the modified Bessel functions.

We prove that in each fixed partial wave, Eq. (40) has exactly one solution $x>0$ if $\alpha_l < 2(2l+1)$. The case $\alpha_l = 2(2l+1)$ gives a zero energy resonance, i.e., $x_o = 0$.

A systematic study of the solutions located off the imaginary axis can be carried out following, e.g., the techniques used in Ref. 1 in the case of δ-sphere interactions.

Here, we simply note that in each fixed partial wave, h_{l,α_l} has an infinite number of resonances off the imaginary axis.

D. Stationary scattering theory for the pair $(h_{l,\alpha_l};\ h_{l,o})$

For $k \geqslant 0$, let us define the functions

$$\mathcal{F}_{l,\alpha_l}(k,r) = F_l(k,r) + \lambda_l(k) F_l'(k,R) g_{l,k}(R,r), \tag{41}$$

where $F_l \in \mathcal{D}(h_{l,o})$ and $g_{l,k}$ and $\lambda_l(k)$ are defined by (22) and (24) respectively.

By inspection, we show that the function $\mathcal{F}_{l,\alpha_l}$ fulfills the following conditions:

$$\mathcal{F}_{l,\alpha_l}(k,R_+) = \mathcal{F}_{l,\alpha_l}(k,R_-) \equiv \mathcal{F}_{l,\alpha_l}(k,R), \tag{42}$$

$$\left(1+\frac{\alpha_l}{2}\right)\mathcal{F}'_{l,\alpha_l}(k,R_+) - \left(1-\frac{\alpha_l}{2}\right)\mathcal{F}'_{l,\alpha_l}(k,R_-) = 0, \tag{43}$$

$$-\mathcal{F}''_{l,\alpha_l}(k,r) + l(l+1)r^{-2}\mathcal{F}_{l,\alpha_l}(k,r) = k^2 \mathcal{F}_{l,\alpha_l}(k,r). \tag{44}$$

Consequently, the functions $\mathcal{F}_{l,\alpha_l}(k,r)$ constitute a set of generalized eigenfunctions associated with h_{l,α_l} or in other words, $\mathcal{F}_{l,\alpha_l}(k,r)$ are the scattering wave functions of h_{l,α_l}.

As usual, the phase shifts of h_{l,α_l} may be obtained through the asymptotic behavior of $\mathcal{F}_{l,\alpha_l}(k,r)$ as $r \to \infty$. Indeed, one has[18]

$$\begin{aligned}
\mathcal{F}_{l,\alpha_l}(k,r) \underset{r\to\infty}{\overset{k>0,}{\to}} & A_l(k)\sin\left(kr-\frac{l\pi}{2}\right) + \lambda_l(k)F_l(k,R)F_l'(k,R)B_l(k)\exp\left[-i\left(kr-\frac{l\pi}{2}\right)\right] \\
&= [A_l(k) - iB_l(k)\lambda_l(k)F_l(k,R)F_l'(k,R)]\sin\left(kr-\frac{l\pi}{2}\right) \\
&\quad + \lambda_l(k)B_l(k)F_l(k,R)F_l'(k,R)\cos\left(kr-\frac{l\pi}{2}\right) \\
&= [C_{1,l}^2(k) + C_{2,l}^2(k)]^{1/2} \sin\left(kr-\frac{l\pi}{2}+\delta_{l,\alpha_l}(k)\right) + o(1),
\end{aligned} \tag{45}$$

which defines the phase shifts by

$$\begin{aligned}
\delta_{l,\alpha_l}(k) &= -\arctan\frac{C_{2,l}(k)}{C_{1,l}(k)}, \\
&= -\arctan\frac{B_l(k)\lambda_l(k)F_l(k,R)F_l'(k,R)}{A_l(k) - iB_l(k)\lambda_l(k)F_l(k,R)F_l'(k,R)},
\end{aligned} \tag{46}$$

where we have used the notations[18]

$$A_l(k) = 2^{-l}k^{-l-1}\Gamma(2l+2)\Gamma(l+1)^{-1}, \tag{47}$$

$$B_l(k) = \frac{1}{kA_l(k)}. \tag{48}$$

The on-shell scattering matrix is defined by

$$\begin{aligned} S_{l,\alpha_l}(k) &= \exp[2i\delta_{l,\alpha_l}(k)] \\ &= 1-2ikB_l^2(k)\lambda_l(k)F_l(k,R)F_l'(k,R). \end{aligned} \tag{49}$$

The corresponding effective range expansion reads[19] as

$$[(2l+1)!!]^2k^{2l+1}\cot\delta_{l,\alpha_l}(k) = -a_{l,\alpha_l}^{-1}+\frac{1}{2}r_{l,\alpha_l}k^2+o(k^4), \tag{50}$$

where the coefficients a_{l,α_l} and r_{l,α_l} are called the partial wave scattering length and effective range parameters, respectively.

A straightforward computation yields

$$a_{l,\alpha_l}=\lambda_l(0)(l+1)R^{2l+1}\cdots \quad . \tag{51}$$

The on-shell scattering amplitude $f_\alpha(k,\omega,\omega')$ corresponding to H_α is given by

$$\begin{aligned} f_\alpha(k,\omega,\omega') &= 4\pi\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{S_{l,\alpha_l}(k)-1}{2ik}\overline{Y_l^m(\omega')Y_l^m(\omega)} \\ &= 4\pi\sum_{l=0}^{\infty}\sum_{m=-l}^{l} f_{l,\alpha_l}(k)\overline{Y_l^m(\omega')Y_l^m(\omega)} \\ & k\geqslant 0; \quad \omega,\ \omega'\in S^2, \end{aligned} \tag{52}$$

where the partial wave scattering amplitude $f_{l,\alpha_l}(k)$ is given by

$$f_{l,\alpha_l}(k) = -B_l^2(k)\lambda_l(k)F_l(k,R)F_l'(k,R). \tag{53}$$

The on-shell scattering operator $S_\alpha(k)$ in $L^2(S^2)$ corresponding to H_α is defined by

$$\begin{aligned} (S_\alpha(k)\phi)(\omega) &= \phi(\omega)-\frac{k}{2\pi i}\int_{S^2}d\omega' f_\alpha(k,\omega,\omega')\phi(\omega') \\ & k\geqslant 0; \quad \omega,\ \omega'\in S^2, \end{aligned} \tag{54}$$

which means that $S_\alpha(k)$ reads as

$$S_\alpha(k)=1+2ik\sum_{l=0}^{\infty}\sum_{m=-l}^{l} f_{l,\alpha_l}(k)(Y_l^m,.)Y_l^m(\omega). \tag{55}$$

III. THE δ'-SPHERE PLUS COULOMB INTERACTION

In this section we consider the system formally given by

$$-\Delta+\gamma|\mathbf{x}|^{-1}+\alpha\delta'(|\mathbf{x}|-R), \quad \gamma\in\mathbb{R}, \quad R>0, \quad \mathbf{x}\in\mathbb{R}^3. \tag{56}$$

A. Defintion of the Hamiltonian

Consider in $L^2(\mathbb{R}^3)$ the operator

$$\begin{aligned} \dot{H}_\gamma &= -\Delta+\gamma|x|^{-1} \\ \mathcal{D}(\dot{H}_\gamma) &= \{f\in H^{2,2}(\mathbb{R}^3)/f(\overline{\partial K(O,R))=f'}(\overline{\partial K(O,R))=0}\}. \end{aligned} \tag{57}$$

Using the decomposition (8) we can write $\dot{H}_\gamma$ in the form

$$\dot{H}_\gamma = \bigoplus_{l=0}^{\infty} U^{-1}\dot{h}_{l,\gamma}U\otimes 1, \tag{58}$$

where

$$\dot{h}_{l,\gamma} = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + \frac{\gamma}{r},$$

$$\mathcal{D}(\dot{h}_{l,\gamma}) = \{f\in L^2((0,\infty))/f,f'\in AC_{\mathrm{loc}}((0,\infty));$$
$$f(0_+)=0 \text{ if } l=0;\quad f(R_\pm)=f'(R_\pm)=0;$$
$$-f''+l(l+1)r^{-2}f+\gamma r^{-1}f\in L^2((0,\infty))\},\quad \gamma\in\mathbb{R}, l\in\mathbb{N}_0. \tag{59}$$

The adjoint $\dot{H}^\star_\gamma$ of $\dot{H}_\gamma$ is defined by

$$\dot{H}^\star_\gamma = \bigoplus_{l=0}^{\infty} U^{-1}\dot{h}^\star_{l,\gamma}U\otimes 1, \tag{60}$$

where

$$\dot{h}^\star_{l,\gamma} = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + \frac{\gamma}{r},$$

$$\mathcal{D}(\dot{h}^\star_{l,\gamma}) = \{f\in L^2((0,\infty))/f,f'\in AC_{\mathrm{loc}}((0,\infty)\setminus\{R\});\quad f(0_+)=0\quad \text{if}\ \ l=0;$$
$$-f''+l(l+1)r^{-2}f+\gamma r^{-1}f\in L^2((0,\infty))\};\ \ \gamma\in\mathbb{R},\ \ l\in\mathbb{N}_0. \tag{61}$$

A direct computation shows that the equation

$$\dot{h}^\star_{l,\gamma}\psi_{l,\gamma}(k) = k^2\psi_{l,\gamma}(k),\ \ \mathrm{Im}(k)>0,\quad l\in\mathbb{N}_0, \tag{62}$$

has two linearly independent solutions:

$$\psi^{(1)}_{l,\gamma}(k,r) = \begin{cases} G_{l,\gamma}(k,R)F_{l,\gamma}(k,r) & ;\quad r<R,\\ 0 & ;\quad r>R, \end{cases} \tag{63}$$

$$\psi^{(2)}_{l,\gamma}(k,r) = \begin{cases} 0 & ;\quad r<R,\\ F_{l,\gamma}(k,R)G_{l,\gamma}(k,r) & ;\quad r>R, \end{cases} \tag{64}$$

where

$$F_{l,\gamma}(k,r) = r^{l+1}\exp(ikr)\ {}_1F_1\left(l+1+\frac{i\gamma}{2k};2l+2;-2ikr\right), \tag{65}$$

$$G_{l,\gamma}(k,r) = \Gamma(2l+2)^{-1}\Gamma\left(l+1+\frac{i\gamma}{2k}\right)(-2ik)^{2l+1}\exp(ikr)\times U\left(l+1+\frac{i\gamma}{2k};\ \ 2l+2;\ \ 2ikr\right), \tag{66}$$

and ${}_1F_1(a;b;z)$, $U(a;b;z)$ denote (ir)regular confluent hypergeometric functions, respectively.[13]

As in the short range case ($\gamma=0$), $\dot{h}_{l,\gamma}$ has deficiency indices (2,2) and consequently all its self-adjoint (s.a.) extensions are given by a four-parameter family of s.a. operators.

In this section we will consider the following one-parameter family of s.a. extensions of $\dot{h}_{l,\gamma}$,

$$h_{l,\gamma,\alpha_l}=-\frac{d^2}{dr^2}+\frac{l(l+1)}{r^2}+\frac{\gamma}{r},$$

$$\begin{aligned}\mathcal{D}(h_{l,\gamma,\alpha_l})=\Big\{& f\in L^2((0,\infty))/f,f'\in AC_{\mathrm{loc}}((0,\infty)\backslash\{R\});\\ & f(0_+)=0,\quad \text{if } l=0;\quad f(R_+)=f(R_-)\equiv f(R);\\ & \left(1+\frac{\alpha_l}{2}\right)f'(R_+)-\left(1-\frac{\alpha_l}{2}\right)f'(R_-)=0;\\ & -f''+l(l+1)r^{-2}f+\gamma r^{-1}f\in L^2((0,\infty))\Big\};\\ & -\infty<\alpha_l<\infty,\quad \gamma\in\mathbb{R},\quad l\in\mathbb{N}_0.\end{aligned} \tag{67}$$

The case $\alpha_l=0$ yields the free Coulomb Hamiltonian $h_{l,\gamma,o}$ for a fixed angular momentum l. The formal expression (56) is therefore represented by the following Hamiltonian in $L^2(\mathbb{R}^3)$:

$$H_{\gamma,\alpha}=\bigoplus_{l=0}^{\infty} U^{-1}h_{l,\gamma,\alpha_l}U\otimes 1. \tag{68}$$

The case $\alpha=0$ leads to the Coulomb Hamiltonian

$$H_{\gamma,o}\equiv H_\gamma=-\Delta+\frac{\gamma}{|x|};\quad \mathcal{D}(H_\gamma)=H^{2,2}(\mathbb{R}^3). \tag{69}$$

Next we introduce the free resolvent,

$$g_{l,\gamma,k}=(h_{l,\gamma}-k^2)^{-1},\ k\neq-\frac{i\gamma}{2n},\ n\in\mathbb{N},\ \mathrm{Im}(k)>0,\ l\in\mathbb{N}_0, \tag{70}$$

with integral kernel

$$g_{l,\gamma,k}(r,r')=\begin{cases}F_{l,\gamma}(k,r')G_{l,\gamma}(k,r) & ;\quad r'\leqslant r,\\ F_{l,\gamma}(k,r)G_{l,\gamma}(k,r') & ;\quad r'\geqslant r.\end{cases} \tag{71}$$

B. The resolvent equation

Theorem 3.1: The resolvent of h_{l,γ,α_l} is given by

$$\begin{aligned}&(h_{l,\gamma,\alpha_l}-k^2)^{-1}=(h_{l,\gamma}-k^2)^{-1}+\lambda_{l,\gamma}(k)(\tilde{\phi}_{l,\gamma}(-\bar{k}),.)\phi_{l,\gamma}(k),\\ &k^2\in\rho(h_{l,\gamma,\alpha_l}),\quad \mathrm{Im}(k)>0,\quad l\in\mathbb{N}_0,\quad \gamma\in\mathbb{R},\end{aligned} \tag{72}$$

where

$$\lambda_{l,\gamma}(k)=\alpha_l\left[1-\frac{\alpha_l}{2}g'_{l,\gamma,k}(R,R)\right]^{-1}, \tag{73}$$

$$\phi_{l,\gamma}(k,r)=\begin{cases} G_{l,\gamma}(k,R)F_{l,\gamma}(k,r) & ; \quad r\leqslant R, \\ F_{l,\gamma}(k,R)G_{l,\gamma}(k,r) & ; \quad r\geqslant R, \end{cases} \tag{74}$$

$$\tilde{\phi}_{l,\gamma}(k,r)=\begin{cases} G'_{l,\gamma}(k,R)F_{l,\gamma}(k,r) & ; \quad r< R, \\ F'_{l,\gamma}(k,R)G_{l,\gamma}(k,r) & ; \quad r> R, \end{cases} \tag{75}$$

we note that $\phi_{l,\gamma}(k,r)=g_{l,\gamma,k}(R,r)$, $\mathrm{Im}(k)>0$.

Proof: One can follow step by step the proof of Theorem 2.1.

Theorem 3.2: The domain $\mathcal{D}(h_{l,\gamma,\alpha_l})$ consists of functions of the type

$$\varphi_{l,\gamma}(k,r)=F_{l,\gamma}(k,r)+\lambda_{l,\gamma}(k)F'_{l,\gamma}(k,R)g_{l,\gamma,k}(R,r); \tag{76}$$

$F_{l,\gamma}\in\mathcal{D}(h_{l,\gamma,\alpha_l})$ and $k^2\in\rho(h_{l,\gamma,\alpha_l})$, $\mathrm{Im}(k)>0$.

The decomposition (76) is unique and with $\varphi_{l,\gamma}\in\mathcal{D}(h_{l,\gamma,\alpha_l})$ of this form, we obtain

$$(h_{l,\gamma,\alpha_l}-k^2)\varphi_{l,\gamma}=(h_{l,\gamma}-k^2)F_{l,\gamma}. \tag{77}$$

Moreover, if $\varphi_{l,\gamma}\in\mathcal{D}(h_{l,\gamma,\alpha_l})$ and $\varphi_{l,\gamma}=0$ in an open set $\mathcal{O}\subset(0,\infty)$, then $h_{l,\gamma,\alpha_l}\varphi_l=0$ in $\mathcal{O}$.

C. Spectral properties

In a similar way to Sec. II, we obtain the following results.

Theorem 3.3: For all $\alpha_l\in(-\infty,\infty)$ and $\gamma\in\mathbb{R}$, we have the following:

- $\sigma_{\mathrm{ess}}(h_{l,\gamma,\alpha_l})=\sigma_{\mathrm{ac}}(h_{l,\gamma,\alpha_l})=[0,\infty)$, (78)
- $\sigma_{\mathrm{sc}}(h_{l,\gamma,\alpha_l})=\varnothing$, (79)
- $\sigma_p(h_{l,\gamma,\alpha_l})\cap[0,\infty)=\varnothing$. (80)

The negative eigenvalues of h_{l,γ,α_l} are obtained from the equation

$$1-\frac{\alpha_l}{2}g'_{l,\gamma,i\sqrt{-E}}(R,R)=0, \quad E<0, \tag{81}$$

which has at most one solution $E_o<0$ for $\gamma\geqslant 0$ and infinitely many for $\gamma<0$.

Proof: Similar to the proof of Theorem 2.3

D. Stationary scattering theory for the pair $(h_{l,\gamma,\alpha_l};h_{l,\gamma})$

For $k\geqslant 0$ let us define the function

$$\mathcal{F}_{l,\gamma,\alpha_l}(k,r)=F_{l,\gamma}(k,r)+\lambda_{l,\gamma}(k)F'_{l,\gamma}(k,R)g_{l,\gamma,k}(R,r), \tag{82}$$

where $F_{l,\gamma}$ and $\lambda_{l,\gamma}(k)$ are defined by (65) and (73), respectively.

A straightforward calculation shows that $\mathcal{F}_{l,\gamma,\alpha_l}(k,r)$ are the scattering wave functions of h_{l,γ,α_l}.

For the determination of the phase shifts of h_{l,γ,α_l}, we follow the strategy of Sec. II D and use the asymptotic behavior of hypergeometric functions given in Ref. 18.

The Coulomb modified phase shift $\delta^{(c)}_{l,\gamma,\alpha_l}(k)$ is given by

$$\delta^{(c)}_{l,\gamma,\alpha_l}(k)=-\arctan\frac{B_{l,\gamma}(k)\lambda_{l,\gamma}(k)F_{l,\gamma}(k,R)F'_{l,\gamma}(k,R)}{A_{l,\gamma}(k)-iB_{l,\gamma}(k)\lambda_{l,\gamma}(k)F_{l,\gamma}(k,R)F'_{l,\gamma}(k,R)}, \tag{83}$$

where we have used the notations[19]

$$A_{l,\gamma}(k)=2^{-l}\exp(\pi\gamma/4k)k^{-l-1}\Gamma(2l+2)\left|\Gamma\left(l+1+\frac{i\gamma}{2k}\right)\right|, \tag{84}$$

$$B_{l,\gamma}(k)=\frac{1}{kA_{l,\gamma}(k)}. \tag{85}$$

We note that the total phase shift $\delta_{l,\gamma,\alpha_l}(k)$ corresponding to h_{l,γ,α_l} splits up into

$$\delta_{l,\gamma,\alpha_l}(k)=\delta^{(c)}_{l,\gamma,\alpha_l}(k)+\delta_{l,\gamma}(k), \quad k>0, \quad \gamma\in\mathbb{R}, \tag{86}$$

where

$$\delta_{l,\gamma}(k)=\arg\Gamma\left(l+1+\frac{i\gamma}{2k}\right), \quad k>0, \quad \gamma\in\mathbb{R} \tag{87}$$

represents the pure Coulomb phase shift.

The Coulomb modified on-shell scattering matrix elements are given by

$$S^{(c)}_{l,\gamma,\alpha_l}(k)=\exp[2i\delta^{(c)}_{l,\gamma,\alpha_l}(k)]=1-2ikB^2_{l,\gamma}(k)\lambda_{l,\gamma}(k)F_{l,\gamma}(k,R)F'_{l,\gamma}(k,R). \tag{88}$$

The corresponding partial wave scattering amplitudes are given by

$$f^{(c)}_{l,\gamma,\alpha_l}(k)=(2ik)^{-1}(S^{(c)}_{l,\gamma,\alpha_l}(k)-1)=-B^2_{l,\gamma}(k)\lambda_{l,\gamma}(k)F_{l,\gamma}(k,R)F'_{l,\gamma}(k,R). \tag{89}$$

The Coulomb modified effective range expansion corresponding to h_{l,γ,α_l} reads[20] as

$$\Gamma(2l+2)^{-1}(2k)^{2l}\left|\Gamma\left(l+1+\frac{i\gamma}{2k}\right)\right|\exp(-\pi\gamma/2k)\times[k\cot\delta^{(c)}_{l,\gamma,\alpha_l}(k)-ik+\exp(\pi\gamma/2k)h_\gamma(k)]$$

$$=-\frac{1}{a^{(c)}_{l,\gamma,\alpha_l}}+o(k^2) \quad k>0, \quad \gamma\in\mathbb{R}, \tag{90}$$

where $a^{(c)}_{l,\gamma,\alpha_l}$ is the Coulomb modified scattering length and the function $h_\gamma(k)$ is defined by

$$h_\gamma(k)=\gamma\left|\Gamma\left(1+\frac{i\gamma}{2k}\right)\right|^2\left[\frac{ik}{\gamma}+\ln\left(\frac{2k}{i|\gamma|}\right)+\Psi\left(1+\frac{i\gamma}{2k}\right)\right]. \tag{91}$$

In Eq. (91), $\Psi(z)$ denotes a digamma function.[13]

In the short range case ($\gamma=0$), Eq. (90) simplifies to

$$\Gamma\left(l+\frac{3}{2}\right)\left(\frac{k}{2}\right)^{2l+1}\frac{\pi}{2}\cot\delta_{l,\alpha_l}(k)=-\frac{1}{a_{l,\alpha_l}}+o(k^2), \tag{92}$$

where a_{l,α_l} is given by (51).

Using the properties of the hypergeometric functions, one may obtain explicitly $a^{(c)}_{l,\gamma,\alpha_l}$ in the expansion (90).

Indeed, a tedious but straightforward calculation gives

$$-\frac{1}{a^{(c)}_{l,\gamma,\alpha_l}}=\begin{cases}\dfrac{1-\alpha_l[rI_\nu(y)K_\nu(y)]'_{r=R}}{\alpha_l\gamma^{-\nu}\Gamma(2l+2)^2[r^{1/2}I_\nu(y)]'_{r=R}[r^{1/2}I_\nu(y)]_{r=R}} & ;\quad \gamma\geqslant 0,\\[2ex] \dfrac{2+i\pi\alpha_l[rJ_\nu(z)H^{(2)}_\nu(z)]'_{r=R}}{2\alpha_l|\gamma|^{-\nu}\Gamma(2l+2)^2[r^{1/2}J_\nu(z)]'_{r=R}[r^{1/2}J_\nu(z)]_{r=R}} & ;\quad \gamma\leqslant 0\,,\end{cases}\tag{93}$$

where we have used the notations $\nu=2l+1$, $y=(4\gamma r)^{1/2}$ and $z=(4|\gamma|r)^{1/2}$.

In the short range case $\gamma=0$, we obtain

$$-a^{-1}_{l,\alpha_l}=\frac{2-\dfrac{\alpha_l}{2l+1}}{2\alpha_l(l+1)R^{2l+1}}=-[\lambda_l(0)(l+1)R^{2l+1}]^{-1}.\tag{94}$$

IV. FINITELY MANY δ'-INTERACTIONS WITH SUPPORT ON CONCENTRIC SPHERES

In this section, we study in dimensions $n=3$ the case of $N-\delta'$-interactions with support on concentric spheres of radii $0<R_1<\cdots<R_N$ formally given by

$$-\Delta+\sum_{j=1}^{N}\alpha_j\delta'(|\mathbf{x}|-R_j),\quad R_j>0,\ \mathbf{x}\in R^3.\tag{95}$$

A. Defintion of the Hamiltonian

Consider in $L^2(\mathbb{R}^3)$ the closed, symmetric and non-negative operator,

$$\tilde{H}=-\Delta,$$

$$\mathcal{D}(\tilde{H})=\{f\in H^{2,2}(\mathbb{R})/f(\overline{\partial K(O,R_j)})=f'(\overline{\partial K(O,R_j)})=0\,,\quad 1\leqslant j\leqslant N\}.\tag{96}$$

Using the decomposition (8), we can write $\tilde{H}$ in the form

$$\tilde{H}=\bigoplus_{l=0}^{\infty}U^{-1}h_{l,\{R\}}U\otimes 1,\tag{97}$$

where

$$h_{l,\{R\}}=-\frac{d^2}{dr^2}+\frac{l(l+1)}{r^2},$$

$$\mathcal{D}(h_{l,\{R\}})=\{f\in L^2((0,\infty))/f,f'\in AC_{\mathrm{loc}}((0,\infty));f(0_+)=0\quad\text{if}\quad l=0;$$
$$f(R_{j\pm})=f'(R_{j\pm})=0;\quad -f''+l(l+1)r^{-2}f\in L^2((0,\infty))\}.\tag{98}$$

The adjoint $\tilde{H}^\star$ of $\tilde{H}$ is given by

$$\tilde{H}^\star=\bigoplus_{l=0}^{\infty}U^{-1}h^\star_{l,\{R\}}U\otimes 1,\tag{99}$$

where

$$h^{\star}_{l,\{R\}} = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2},$$

$$\mathcal{D}(h^{\star}_l) = \{f \in L^2((0,\infty))/f, f' \in AC_{loc}((0,\infty)\backslash\{R\}); \quad f(0_+) = 0 \quad \text{if} \quad l = 0;$$
$$-f'' + l(l+1)r^{-2} f \in L^2((0,\infty))\}; \quad l \in \mathbb{N}_0,$$

$$\{R\} = \{R_1, \ldots, R_N\}, \tag{100}$$

A straightforward computation shows that the equation

$$h^{\star}_{l,\{R\}} \psi_l(k) = k^2 \psi_l(k), \quad \text{Im}(k) > 0, \tag{101}$$

has $2N$ linearly independent solutions,

$$\psi^{(1)}_{l,j}(k,r) = \begin{cases} G_l(k,R_j) F_l(k,r) & ; \quad r < R_j, \\ 0 & ; \quad r > R_j, \end{cases} \tag{102}$$

$$\psi^{(2)}_{l,j}(k,r) = \begin{cases} 0 & ; \quad r < R_j, \\ F_l(k,R_j) G_l(k,r) & ; \quad r > R_j, \end{cases} \tag{103}$$

Therefore $h_{l,\{R\}}$ has deficiency indices $(2N,2N)$ and consequently all self-adjoint (s.a.) extensions of $h_{l,\{R\}}$ are given by a $4N^2$-parameter family of s.a. operators.

In this paper, we consider a special N-parameter family of s.a extensions of $h_{l,\{R\}}$ corresponding to the formal expression (95).

We introduce in $L^2((0,\infty))$ the following family of closed extensions of $h_{l,\{R\}}$:

$$h_{l,\{\alpha_l\},\{R\}} = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2},$$

$$\mathcal{D}(h_{l,\{\alpha_l\},\{R\}}) = \{f \in L^2((0,\infty))/f, f' \in AC_{\text{loc}}((0,\infty)\backslash\{R\});$$

$$f(0_+) = 0 \quad \text{if} \quad l = 0\,; f(R_{j^+}) = f(R_{j^-}) \equiv f(R_j);$$

$$\left(1 + \frac{\alpha_{jl}}{2}\right) f'(R_{j^+}) - \left(1 - \frac{\alpha_{jl}}{2}\right) f'(R_{j^-}) = 0\,;$$

$$-f'' + l(l+1)r^{-2} f \in L^2((0,\infty))\}\,;$$

$$\{\alpha_l\} = \{\alpha_{1l}, \ldots, \alpha_{Nl}\}, \, -\infty < \alpha_{jl} < \infty, \, l \in \mathbb{N}_0. \tag{104}$$

Following, e.g., Ref. 15, Chap. II 3 we note that there exists an intermediate operator in $L^2((0,\infty))$ with deficiency indices (N,N) which is a proper extension of $h_{l,\{R\}}$.

A simple integration by parts shows that $h_{l,\{\alpha_l\},\{R\}}$ is symmetric. Moreover, since $h_{l,\{\alpha_l\},\{R\}}$ may be obtained as an extension of an operator with deficiency (N,N) and the N boundary conditions in (104) are symmetric and linearly independent, it follows from Ref. 21, Theorem XII, p. 4.30 that $h_{l,\{\alpha_l\},\{R\}}$ is self-adjoint.

The case $\alpha_{jl} = 0$ for all $j = 1, \ldots, N$ (i.e., $\{\alpha_l\} = 0$) coincides with the free kinetic energy Hamiltonian $h_{l,o,\{R\}} \equiv h_{l,o}$ for a fixed angular momentum l.

By definition the operator $H_{\{\alpha_l\},\{R\}}$ given in $L^2(\mathbb{R}^3)$ by

$$H_{\{\alpha_l\},\{R\}}=\bigoplus_{l=0}^{\infty} U^{-1}h_{l,\{\alpha_l\},\{R\}}U\otimes 1 \tag{105}$$

describes $N-\delta'$-interactions with support on concentric spheres of radii $0<R_1<\ldots<R_N$. The case $\{\alpha_l\}=0$ yields the free Hamiltonian

$$H_{o,\{R\}}\equiv H_o=-\Delta,\quad \mathcal{D}(H_o)=H^{2,2}(\mathbb{R}^3). \tag{106}$$

B. The resolvent equation

Theorem 4.1: The resolvent of $h_{l,\{\alpha_l\},\{R\}}$ is given by

$$(h_{l,\{\alpha_l\},\{R\}}-k^2)^{-1}=(h_{l,o}-k^2)^{-1}+\sum_{j,j'=1}^{N}\lambda_{jj'}(k)(\tilde{\phi}_{l,j'}(-\overline{k}),.)\phi_{l,j}(k)$$
$$k^2\in\rho(h_{l,\{\alpha_l\},\{R\}}),\quad \mathrm{Im}(k)>0,\quad l\in\mathbb{N}_0, \tag{107}$$

where

$$[\lambda(k)]^{-1}_{jj'}=\left[\alpha_{jl}^{-1}\delta_{jj'}-\frac{1}{2}g'_{l,k}(R_j,R_{j'})\right], \tag{108}$$

$$\phi_{l,j}(k,r)=\begin{cases}G_l(k,R_j)F_l(k,r) & ;\quad r\leqslant R_j,\\ F_l(k,R_j)G_l(k,r) & ;\quad r\geqslant R_j,\end{cases} \tag{109}$$

$$\tilde{\phi}_{l,j}(k,r)=\begin{cases}G'_l(k,R_j)F_l(k,r) & ;\quad r<R_j,\\ F'_l(k,R_j)G_l(k,r) & ;\quad r>R_j,\end{cases} \tag{110}$$

$$1\leqslant j\leqslant N.$$

Proof: Equation (107) follows from Krein's formula.[14]
In order to determine the factors $\lambda_{jj'}(k)$ we proceed as follows
Let $g\in L^2((0,\infty))$ and define the function

$$\tilde{\chi}_l(k,r)=((h_{l,\{\alpha_l\},\{R\}}-k^2)^{-1}g)(r), \tag{111}$$

where k is chosen in such a way that $\det[\lambda(k)]\neq 0$. Since $\tilde{\chi}_l\in\mathcal{D}(h_{l,\{\alpha_l\},\{R\}})$, it follows from Eq. (104) that $\tilde{\chi}$ satisfies the following conditions:

$$\tilde{\chi}_l\in AC_{\mathrm{loc}}((0,\infty)), \tag{112}$$

$$\left(1+\frac{\alpha_{jl}}{2}\right)\tilde{\chi}'_l(k,R_{j+})-\left(1-\frac{\alpha_{jl}}{2}\right)\tilde{\chi}'_l(k,R_{j-})=0, \tag{113}$$

$$((h_{l,\{\alpha_l\},\{R\}}-k^2)\tilde{\chi}_l)(r)=-\tilde{\chi}''_l(k,r)+\frac{l(l+1)}{r^2}\tilde{\chi}_l(k,r)-k^2\tilde{\chi}_l(k,r)=g(r),$$
$$r>0,\quad r\neq R_j,\quad 1\leqslant j\leqslant N. \tag{114}$$

The implementation of these conditions gives the factors $\lambda_{jj'}(k)$.

The resolvent of $H_{\{\alpha_l\},\{R\}}$ may be easily obtained using Eq. (105) and (107). We get

$$(H_{\{\alpha_l\},\{R\}}-k^2)^{-1}=(H_o-k^2)^{-1}+\bigoplus_{l=0}^{\infty}\bigoplus_{m=-l}^{l}\sum_{j,j'=1}^{N}\lambda_{jj'}(k)\times(|\cdot|^{-1}\widetilde{\phi}_{l,j'}(-\overline{k})Y_l^m,.)|\cdot|^{-1}\phi_{l,j}(k)Y_l^m$$

$$k\in\rho(H_{\{\alpha_l\},\{R\}}),\quad \mathrm{Im}(k)>0,\ l\in\mathbb{N}_0. \tag{115}$$

The following theorem provides additional information on the domain of $h_{l,\{\alpha_l\},\{R\}}$ and shows that $h_{l,\{\alpha_l\},\{R\}}$ describes a local interaction.

Theorem 4.2: The domain $\mathcal{D}(h_{l,\{\alpha_l\},\{R\}})$ consists of functions of the type

$$\widetilde{\varphi}_l(k,r)=F_l(k,r)+\sum_{j,j'=1}^{N}\lambda_{jj'}(k)F_l'(k,R_{j'})g_{l,k}(R_j,r), \tag{116}$$

where $F_l\in\mathcal{D}(h_{l,\{\alpha_l\},\{R\}})$ and $k^2\in\rho(h_{l,\{\alpha_l\},\{R\}})$, $\mathrm{Im}(k)>0$.

The decomposition (116) is unique and with $\widetilde{\varphi}_l\in\mathcal{D}(h_{l,\{\alpha_l\},\{R\}})$ of this form, we obtain

$$(h_{l,\{\alpha_l\},\{R\}}-k^2)\widetilde{\varphi}_l=(h_{l,o}-k^2)F_l. \tag{117}$$

Moreover if $\widetilde{\varphi}_l\in\mathcal{D}(h_{l,\{\alpha_l\},\{R\}})$ and $\widetilde{\varphi}_l=0$ in an open set $\mathcal{O}\subset(0,\infty)$, then $h_{l,\{\alpha_l\},\{R\}}\widetilde{\varphi}_l=0$ in $\mathcal{O}$, which means that $h_{l,\{\alpha_l\},\{R\}}$ describes a local interaction.

C. Spectral properties

Theorem 4.3: For all $j=1,\ldots,N$, let $\alpha_{jl}\in(-\infty,\infty)$ and assume that $\alpha_{jl}\neq 0$. Then $h_{l,\{\alpha_l\},\{R\}}$ has at most N eigenvalues which are all negative and simple. The remaining part of the spectrum is purely absolutely continuous and covers the non-negative real axis,

$$\sigma_{\mathrm{ess}}(h_{l,\{\alpha_l\},\{R\}})=\sigma_{\mathrm{ac}}(h_{l,\{\alpha_l\},\{R\}})=[0,\infty), \tag{118}$$

$$\sigma_{\mathrm{sc}}(h_{l,\{\alpha_l\},\{R\}})=\emptyset. \tag{119}$$

Proof: One may follow step by step[2] where a similar result was obtained for finitely many δ-interactions with support on concentric spheres.

D. Stationary scattering theory for the pair $(h_{l,\{\alpha_l\},\{R\}};h_{l,o})$

For $k\geqslant 0$, let us define the functions

$$\mathcal{F}_{l,\{\alpha_l\},\{R\}}(k,r)=F_l(k,r)+\sum_{j,j'=1}^{N}\lambda_{jj'}(k)F_l'(k,R_{j'})g_{l,k}(R_j,r). \tag{120}$$

One can easily show that $\mathcal{F}_{l,\{\alpha_l\},\{R\}}$ are the scattering wave functions of $h_{l,\{\alpha_l\},\{R\}}$.

As usual, the phase shifts of $h_{l,\{\alpha_l\},\{R\}}$ may be obtained through the asymptotic behavior of $\mathcal{F}_{l,\{\alpha_l\},\{R\}}(k,r)$ as $r\rightarrow\infty$. Indeed, using Ref. 18, we obtain

$$\mathcal{F}_{l,\{\alpha_l\},\{R\}}(k,r) \underset{r\rightarrow\infty}{\overset{k>0}{\rightarrow}} A_l(k)\sin\left(kr-\frac{l\pi}{2}\right)$$

$$+\sum_{j,j'=1}^{N} \lambda_{jj'}(k)F_l(k,R_j)F_l'(k,R_{j'})B_l(k)\exp\left[-i\left(kr-\frac{l\pi}{2}\right)\right]$$

$$=\left[A_l(k)-iB_l(k)\sum_{j,j'=1}^{N} \lambda_{jj'}(k)F_l(k,R_j)F_l'(k,R_{j'})\right]\sin\left(kr-\frac{l\pi}{2}\right)$$

$$+B_l(k)\sum_{j,j'=1}^{N} \lambda_{jj'}(k)F_l(k,R_j)F_l'(k,R_{j'})\cos\left(kr-\frac{l\pi}{2}\right)$$

$$=[\tilde{C}_{1,l}^2(k)+\tilde{C}_{2,l}^2(k)]^{1/2}\sin\left(kr-\frac{l\pi}{2}+\delta_{l,\{\alpha_l\},\{R\}}(k)\right)+o(1), \tag{121}$$

which defines the phase shifts by

$$\delta_{l,\{\alpha_l\},\{R\}}(k)=-\arctan\frac{\tilde{C}_{2,l}(k)}{\tilde{C}_{1,l}(k)}$$

$$=-\arctan\frac{B_l(k)\sum_{j,j'=1}^{N} \lambda_{jj'}(k)F_l(k,R_j)F_l'(k,R_{j'})}{A_l(k)-iB_l(k)\sum_{j,j'=1}^{N} \lambda_{jj'}(k)F_l(k,R_j)F_l'(k,R_{j'})}, \tag{122}$$

where $A_l(k)$ and $B_l(k)$ are defined by (47) and (48), respectively.

Using Eq. (122), we can now compute the other scattering data. Indeed, we obtain the following results.

(i) The on-shell scattering matrix,

$$S_{l,\{\alpha_l\},\{R\}}(k)=\exp[2i\delta_{l,\{\alpha_l\},\{R\}}(k)]$$

$$=1-2ikB_l^2(k)\sum_{j,j'=1}^{N} \lambda_{jj'}(k)F_l(k,R_j)F_l'(k,R_{j'}). \tag{123}$$

(ii) The partial wave scattering length,

$$a_{l,\{\alpha_l\},\{R\}}=\sum_{j,j'=1}^{N} \lambda_{jj'}(0)(l+1)R_j^{l+1}R_{j'}^{l}. \tag{124}$$

(iii) The partial wave scattering amplitude,

$$f_{l,\{\alpha_l\},\{R\}}(k)=-B_l^2(k)\sum_{j,j'=1}^{N} \lambda_{jj'}(k)F_l(k,R_j)F_l'(k,R_{j'}). \tag{125}$$

V. THE δ'-SPHERE INTERACTION OF THE SECOND TYPE

In this section, following Refs. 1,12, we provide another exactly solvable model that we call the δ'-sphere interaction of the second type.

This model is obtained by formally interchanging the role of f and f' in the boundary conditions at $r=R$ in (18).

The radial quantum Hamiltonian describing this model is therefore defined by

$$h_{l,\beta_l} = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2},$$

$$\begin{aligned}
\mathcal{D}(h_{l,\beta_l}) = \{ & f \in L^2((0,\infty))/f,f' \in AC_{\text{loc}}((0,\infty)\backslash\{R\}); \quad f(0_+)=0, \quad \text{if} \quad l=0; \\
& f'(R_+) = f'(R_-) \equiv f'(R); \\
& \left(1+\frac{\beta_l}{2}\right) f(R_+) - \left(1-\frac{\beta_l}{2}\right) f(R_-) = 0; \\
& -f'' + l(l+1)r^{-2} f \in L^2((0,\infty))\}; \; -\infty < \beta_l < \infty, \; l \in \mathbb{N}_0.
\end{aligned} \tag{126}$$

Now we can apply the techniques used in the previous sections in order to carry out a systematic study of this model.

In particular, we can show that the resolvent of h_{l,β_l} is given by

$$\begin{aligned}
(h_{l,\beta_l} - k^2)^{-1} = (h_{l,o} - k^2)^{-1} + \hat{\lambda}_l(k) (\phi_l(-\overline{k}), \cdot) \tilde{\phi}_l(k), \\
k^2 \in \rho(h_{l,\beta_l}), \quad \text{Im}(k) > 0, \quad l \in \mathbb{N}_0,
\end{aligned} \tag{127}$$

where

$$\hat{\lambda}_l(k) = -\beta_l \left[1 + \frac{\beta_l}{2} g'_{l,k}(R,R) \right]^{-1}, \tag{128}$$

and ϕ_l, $\tilde{\phi}_l$ are defined by (25) and (26), respectively.

We also note that the domain $\mathcal{D}(h_{l,\beta_l})$ consists of functions of the form

$$\hat{\varphi}_l(k,r) = F_l(k,r) + \hat{\lambda}_l(k) F_l(k,R) \tilde{\phi}_l(k,r), \tag{129}$$

and that h_{l,β_l} describes a local interaction.

ACKNOWLEDGMENTS

M.H. would like to thank the Swedish International Development Cooperation Agency (SIDA) and the Abdus Salam International Center for Theoretical Physics (ICTP), Trieste, Italy for supporting his visit to the ICTP where a part of this work was performed. M.N.H gratefully thanks Professor J. P. Antoine as well as the staff of the Institut de Physique Théorique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, for their kind hospitality. He also thanks the Administration Générale pour la Coopération au Développement (AGCD), Belgium, for the financial support provided to him during his stay in Belgium.

[1] J-P. Antoine, F. Gesztesy, and J. Shabani, J. Phys. A **20**, 3687–3712 (1987).
[2] J. Shabani, J. Math. Phys. **29**, 660–664 (1988).
[3] J. Shabani, Nuovo Cimento B **101**, 429–439 (1988).
[4] L. Dabrowski and J. Shabani, J. Math. Phys. **29**, 2241–2244 (1988).
[5] J. Shabani, Ann. Soc. Sci. Bruxelles **105**, 105–111 (1991).

[6] T. Ikebe and S. Shimada, J. Math. Kyoto Univ. **1–31**, 219–258 (1991).
[7] T. Ikebe, J. Anal. Math. **59**, 39–43 (1992).
[8] J.-P. Antoine, P. Exner, P. Sĕba, and J. Shabani, Ann. Phys. (N.Y.) **233**, 1–16 (1994).
[9] J.-P. Antoine, P. Exner, P. Sĕba, and J. Shabani, *Operator Theory: Advances and Applications*, 1994, Vol. 70, pp. 79–87.
[10] J. Dittrich, P. Exner, and P. Sĕba, J. Math. Phys. **30**, 2275–2282 (1989).
[11] J. Dittrich, P. Exner, and P. Sĕba, J. Math. Phys. **33**, 2207–2214 (1992).
[12] M. N. Hounkonnou, M. Hounkpe, and J. Shabani, J. Math. Phys. **38**, 2832–2850 (1997).
[13] M. Abramowitz and I. A. Stegun: *Handbook of Mathematical Functions* (Dover, New York, 1972).
[14] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space* (Pitman, Boston, 1981), Vol. 2.
[15] S. Albeverio, F. Gesztesy, R. Hoegh-Krohn, and H. Holden, *Solvable Models in Quantum Mechanics*, Texts and Monographs in Physics (Springer-Verlag, Berlin, 1988).
[16] M. Reed and B. Simon, *Methods of Modern Mathematical Physics Vol. 4; Analysis of Operators* (Academic, New York, 1978).
[17] H. Nussenveig, *Causality and Dispersion Relations* (Academic, New York, 1972).
[18] F. Gesztesy and B. Thaller, J. Phys. A **14**, 639–657 (1981).
[19] R. G. Newton, *Scattering Theory of Waves and Particles*, 2nd ed. (Springer-Verlag, New York, 1982).
[20] D. Bollé and F. Gesztesy, Phys. Rev. A **30**, 1279 (1984).
[21] N. Dunford and J. T. Schwarz, *Linear Operators Part II- Spectral Theory* (Interscience, New York, 1963).

Degeneracy and para-supersymmetry of Dirac Hamiltonian in (2+1)-space–time

M. A. Jafarizadeh[a)]
Faculty of Physics, Tabriz University, Tabriz 51664, Iran and Institute for Studies in Theoretical Physics and Mathematics, Tehran 19395-1795, Iran

S. K. Moayedi[b)]
Faculty of Physics, Tabriz University, Tabriz 51664, Iran

(Received 2 February 1999; accepted for publication 19 February 1999)

The quantum mechanics of a spin $\frac{1}{2}$ particle on a locally spatial constant curvature part of a $(2+1)$-space–time in the presence of a constant magnetic field of a magnetic monopole has been investigated. It has been shown that these two-dimensional Hamiltonians have the degeneracy group of $SL(2,c)$, and para-supersymmetry of arbitrary order or shape invariance. Using this symmetry we have obtained its spectrum algebraically. The Dirac's quantization condition has been obtained from the representation theory. Also, it is shown that the presence of angular deficit suppresses both the degeneracy and shape invariance. © *1999 American Institute of Physics.* [S0022-2488(99)00409-0]

I. INTRODUCTION

Quantum theories, particularly quantum gravity in $(2+1)$-dimensions provide us with a useful field of investigation not only for theoretical and mathematical issues, but also in some cases, for actual physical problems.[1,2] In the past decade many interesting physical problems in $(2+1)$-gravity have been solved, such as classical scattering, quantum scattering, bound states of a scalar and spinor point particle both in the presence and absence of a magnetic monopole and also magnetic vortex.[3–6] In Ref. 6 some interesting results have been obtained in studying quantum scattering and bound states of a scalar charged particle in the background metric corresponding to $(2+1)$-manifold both with local and global constant curvature in the presence of a magnetic monopole, which satisfy the coupled Einstein–Maxwell equations.

Here in this article we investigate the quantum mechanics of a charged spin $\frac{1}{2}$ point particle on a $(2+1)$-space–time with spatial part of local constant curvature in the presence of a constant magnetic field of a magnetic monopole. We show that these two-dimensional Hamiltonians have the degeneracy group of $SL(2,c)$ type and para-supersymmetry of arbitrary order or shape invariance. Using these symmetries we have obtained their spectra algebraically. Also, the Dirac quantization follows naturally from the representation theory. In the case of local constancy of the curvature, the presence of angular deficit suppresses both the degeneracy and shape invariance.

The paper is organized as follows: In Sec. II we briefly describe the $(2+1)$-space–time metric of Ref. 6 and assume that angular deficit is absent. Section III is devoted to the algorithm of the manipulation of the Dirac operator in these spacetimes. The Dirac operator has been given in terms of the generators of the $sl(2,c)$ Lie algebra, which reduces the familiar Dirac operator on the S^2 in special case.[7] In Sec. IV we obtain the left and right invariant generators of the $SL(2,c)$ Lie group in terms of Euler's angles,[8] where after eliminating the ψ coordinate we get the eigenspectrum of massless Dirac operator together with its degeneracy which is in agreement with those of Ref. 7 in the special case of S^2. Also, as a special case, we obtain the massless Dirac operator in the presence of a magnetic vortex.[5,9] In Sec. V we add the constant magnetic field of

[a)]Electronic mail: jafarzadeh@ark.tabrizu.ac.ir
[b)]Electronic mail: moayedi@ark.tabrizu.ac.ir

the magnetic monopole. Using again the representation of the $SL(2,c)$ Lie group we obtain the eigenspectrum of a charged spin $\frac{1}{2}$ particle algebraically together with its degeneracy group. As a special case we obtain the monopole harmonics.[10–12] Also we obtain the familiar Dirac quantization from the representation theory. In Sec. VI, using the right invariant generators and eliminating the coordinate ψ, we obtain the raising and lowering operators of the magnetic charge. Using them we show the presence of the para-supersymmetry of arbitrary order p, or equivalently, the shape invariance symmetry associated with the Dirac operator in the presence of a magnetic monopole. Finally in Sec. VII we add the angular deficit to the background space–time metric which leads to the suppression of both degeneracy and the shape invariance symmetry. Thus we obtain the eigenspectrum by solving the Dirac operator by usual method which is in agreement with the result of Ref. 4 for the special case of the cone.

II. (2+1)-SPACE–TIME WITH LOCAL SPATIAL CONSTANT CURVATURE

In $(2+1)$ dimensions the Einstein–Hilbert action of gravity coupled to matter and electromagnetic field, together with the cosmological term can be written as[6]

$$S=\int d^3x\sqrt{|\det g_{\mu\nu}|}\left\{\frac{1}{4\pi G}(R+2\Lambda)+\frac{1}{4}F_{\mu\nu}F^{\mu\nu}-\mathcal{L}_M\right\}, \tag{2.1}$$

where $\mathcal{L}_M$ is the matter Lagrangian corresponding to a very massive point particle with mass M located in the origin. We have rescaled G by a factor of 4. As it is shown in Ref. 6, the following $(2+1)$-space–time metric corresponds to a massive point mass M together with non-negative cosmological constant Λ and magnetic monopole field

$$ds^2=dt^2-\frac{1}{2\lambda}\left(d\theta^2+(1-GM)^2\frac{\sin^2\alpha\theta}{\alpha^2}d\phi^2\right), \tag{2.2}$$

where α and λ satisfy the following relation:

$$\Lambda=\alpha^2\lambda, \tag{2.3}$$

and GM satisfies the condition $GM<1$. The parameter α in Eq. (2.3) chooses one of the values 0,1, i. In the case of $\alpha=1$, λ is a positive real number and we have

$$0\leqslant\theta<\pi,\quad 0\leqslant\phi<2\pi.$$

For $\alpha=i$, λ is a negative real number and $(2+1)$-space–time metric (2.2) is Euclidean. In this case we have $0\leqslant\theta<\infty$ and the range of ϕ is the same as $\alpha=1$ case. Finally for $\alpha=0$, λ is a positive real number again and θ plays the role of the radial variable[6] and the range of the coordinate ϕ is the same as $\alpha=1$, i case. The magnetic monopole field corresponding to the system (2.1) is

$$\mathcal{B}=g(1-GM)\alpha\sin\alpha\theta, \tag{2.4}$$

with

$$g=\begin{cases}\dfrac{1}{2\sqrt{\pi G\lambda}} & \text{if } \alpha=0 \text{ and } 1\\[2ex] \dfrac{-1}{2\sqrt{\pi G|\lambda|}} & \text{if } \alpha=i\end{cases} \tag{2.5}$$

which extremizes the Einstein–Hilbert action given in (2.1). In other words they are the solution of the Einstein–Maxwell equation which extremizes this action. The magnetic field given in (2.4) is the magnetic field of a magnetic monopole located in the origin of the R^3 Euclidean space where the constant curvature spatial part of the space–time is embedded in it. The corresponding magnetic potential one-form A in the coordinates θ and ϕ is

$$A = -g(1-GM)\cos\alpha\theta d\phi. \tag{2.6}$$

In the next section after introducing the abstract Dirac operator we write it on the spatial part of metric (2.2) in the case $M=0$.

III. DIRAC OPERATOR ON TWO-DIMENSIONAL SPACES WITH GLOBAL CONSTANT CURVATURE

The massless Dirac operator on a given d-dimensional Riemannian manifold with metric $g_{\mu\nu}$ can be written as[7,8]

$$D = -i\gamma^a E_a^\mu \left(\partial_\mu + \tfrac{1}{8}\omega_{\mu ab}[\gamma^a,\gamma^b]\right), \tag{3.1}$$

where γ^a, s are the generators of the flat Clifford algebra which satisfy the following anticommutation relation,

$$\{\gamma^a,\gamma^b\} = 2\delta^{ab}, \quad a,b = 1,...,d. \tag{3.2}$$

Also $E_a^\mu, \omega_{\mu ab}$ are d-beins and spin connections, respectively, which satisfy the following relations:

$$\begin{gathered} E_a^\mu g_{\mu\nu} E_b^\nu = \delta_{ab}, \quad E_a^\mu e_\mu^b = \delta_a^b, \\ \partial_\mu e_\nu^a - \Gamma_{\mu\nu}^\lambda e_\lambda^a + \omega_{\mu ab} e_\nu^b = 0. \end{gathered} \tag{3.3}$$

Here in this article we are concerned with the manifolds described by metric (2.2) and gauge potential (2.6). We also assume that M vanishes in the rest of the article except in the last section. Then the spatial part of the metric (2.2) reads

$$ds_{(2)}^2 = \frac{1}{2\lambda}\left(d\theta^2 + \frac{\sin^2\alpha\theta}{\alpha^2} d\phi^2\right), \tag{3.4}$$

it is clear from the above metric that the spatial part consists of a two-dimensional manifold of global constant curvature. Using Eqs. (3.3) we obtain the following expression for the nonvanishing components of zwei-beins and spin connections associated with metric (3.4):

$$\begin{gathered} E_1^\theta = \sqrt{2\lambda}, \quad E_2^\phi = \sqrt{2\lambda}\frac{\alpha}{\sin\alpha\theta}, \\ \omega_{\phi 12} = -\cos\alpha\theta. \end{gathered} \tag{3.5}$$

By using Eq. (3.1), we obtain the Dirac operator on a manifold described by metric (3.4) as follows:

$$D_2 = -i\sqrt{2\lambda}\gamma^1\left(\partial_\theta + \frac{1}{2}\frac{\alpha}{\tan\alpha\theta}\right) - i\sqrt{2\lambda}\gamma^2\frac{\alpha}{\sin\alpha\theta}\partial_\phi. \tag{3.6}$$

For $\alpha=1$, D_2 becomes the Dirac operator on the two-dimensional sphere S^2.[7] It is more convenient to consider the two-dimensional manifold (3.4) as a submanifold of the three-dimensional manifold M_3 with the line element,

$$ds^2_{(3)}=dr^2+\alpha^2r^2\left(d\theta^2+\frac{\sin^2\alpha\theta}{\alpha^2}d\phi^2\right), \tag{3.7}$$

which is parametrized with the following local coordinates:

$$x_1=r\sin\alpha\theta\cos\phi,\quad x_2=r\sin\alpha\theta\sin\phi,\quad x_3=r\cos\alpha\theta. \tag{3.8}$$

Now we consider $r=\text{constant}$ submanifold of M_3 with the following metric:

$$ds^2_{(2)}=\alpha^2r^2\left(d\theta^2+\frac{\sin^2\alpha\theta}{\alpha^2}d\phi^2\right). \tag{3.9}$$

For $\alpha=1$ and $r=(1/\sqrt{2\lambda})>0$ this submanifold coincides with the special case of $\alpha=1$ of the two-dimensional manifold with metric (3.4), while for $\alpha=0$, with the assumption of $\lim_{\alpha\to 0,r\to\infty}\alpha r=\text{finite}=(1/\sqrt{2\lambda})$, it is the same as the $\alpha=0$ case of metric (3.4). Finally for $\alpha=i$, with the assumption $r=(1/\sqrt{2|\lambda|})>0$ it changes to the $\alpha=i$ case of metric (3.4). Now we try to define the Dirac operator on the manifold M_3. Using Eq. (3.3), for nonvanishing components of 3-beins and spin connections we get

$$E_1^\theta=\frac{1}{\alpha r},\quad E_2^\phi=\frac{1}{r\sin\alpha\theta},\quad E_3^r=1,$$

$$\omega_{\theta 13}=\alpha,\omega_{\phi 21}=\cos\alpha\theta,\omega_{\phi 23}=\sin\alpha\theta. \tag{3.10}$$

Finally, using the relations (3.1) and (3.10) the Dirac operator associated with metric (3.7) reads

$$D_3=-i\gamma^1\frac{1}{\alpha r}\left(\partial_\theta+\frac{1}{2}\frac{\alpha}{\tan\alpha\theta}\right)-i\gamma^2\frac{1}{\alpha r}\frac{\alpha}{\sin\alpha\theta}\partial_\phi-i\gamma^3\left(\partial_r+\frac{1}{r}\right). \tag{3.11}$$

It is straightforward to see that the Dirac operator D_3 yields

$$\left(-i\gamma^3D_3+\frac{1}{r}\right)\Bigg|_{r=\text{constant}}=-i\Gamma^1\frac{1}{\alpha r}\left(\partial_\theta+\frac{1}{2}\frac{\alpha}{\tan\alpha\theta}\right)-i\Gamma^2\frac{1}{\alpha r}\frac{\alpha}{\sin\alpha\theta}\partial_\phi, \tag{3.12}$$

with Γ^a defined as

$$\Gamma^a=-i\gamma^3\gamma^a,\quad a=1,2,$$

which satisfy the following Clifford algebra

$$\{\Gamma^a,\Gamma^b\}=2\delta^{ab}. \tag{3.13}$$

Assuming the equivalence of the metric of submanifold given in (3.9) with the two-dimensional metric (3.4) and also replacing αr with $1/\sqrt{2\lambda}$, we can deduce that the operator (3.12) is the same as Dirac operator D_2 given in (3.6). In terms of local coordinates (3.8) the operator D_3 can be written as

$$D_3=-i\sigma_i\partial_i, \tag{3.14}$$

where σ_i, s are Pauli matrices. Using the identity $(\sigma_ix_i/r)^2=I$, we have

$$D_3=\left(\frac{\sigma_ix_i}{r}\right)^2D_3=-i\frac{\sigma_ix_i}{r}\left(\partial_r+\frac{i}{r}\epsilon_{ijk}\sigma_ix_j\partial_k\right). \tag{3.15}$$

Now, comparing the operator (3.15) with (3.11), it follows that the operator γ^3 has the following form:

$$\gamma^3 = \frac{\sigma_i x_i}{r}. \tag{3.16}$$

Using the relations (3.12) and (3.15) together with the relation (3.11) the Dirac operator D_2 over two dimensional manifold with metric (3.4) can be represented as

$$D_2 = \frac{1}{r} - \frac{i}{r}\epsilon_{ijk}\sigma_i x_j \partial_k = \sqrt{2\lambda}(\sigma_1 I_1 + \sigma_2 I_2 + \alpha\sigma_3 I_3 + \alpha I), \tag{3.17}$$

where I is a 2×2 identity matrix and the differential operators I_i with $i=1,2,3$ in (3.17) have the following form:

$$\begin{aligned} I_1 &= i\left(\sin\phi\partial_\theta + \frac{\alpha}{\tan\alpha\theta}\cos\phi\partial_\phi\right),\\ I_2 &= i\left(-\cos\phi\partial_\theta + \frac{\alpha}{\tan\alpha\theta}\sin\phi\partial_\phi\right),\\ I_3 &= -i\partial_\phi, \end{aligned} \tag{3.18}$$

and satisfy the following $sl(2,c)$ Lie algebra:

$$[I_1,I_2] = i\alpha^2 I_3, \quad [I_2,I_3] = iI_1, \quad [I_3,I_1] = iI_2. \tag{3.19}$$

It is clear that for $\alpha=1$ this algebra becomes an $su(2)$ Lie algebra, for $\alpha=i$ we get $su(1,1)$ Lie algebra, and finally for $\alpha=0$ we get $iso(2)$ Lie algebra.[6,13]

IV. DEGENERACY GROUP OF THE DIRAC OPERATOR ON TWO-DIMENSIONAL MANIFOLDS WITH GLOBAL CONSTANT CURVATURE

In order to obtain the degeneracy group of the Dirac operator D_2 over the two-dimensional manifold with metric (3.4) we need to know the left and right invariant generators of $SL(2,c)$ group which have the following form in the Eulerean coordinates:[6]

$$\begin{aligned} L_1^{(L)} &= i\left(\sin\phi\partial_\theta + \frac{\alpha}{\tan\alpha\theta}\cos\phi\partial_\phi - \frac{\alpha}{\sin\alpha\theta}\cos\phi\partial_\psi\right),\\ L_2^{(L)} &= i\left(-\cos\phi\partial_\theta + \frac{\alpha}{\tan\alpha\theta}\sin\phi\partial_\phi - \frac{\alpha}{\sin\alpha\theta}\sin\phi\partial_\psi\right),\\ L_3^{(L)} &= -i\partial_\phi,\\ L_1^{(R)} &= i\left(\sin\psi\partial_\theta + \frac{\alpha}{\tan\alpha\theta}\cos\psi\partial_\psi - \frac{\alpha}{\sin\alpha\theta}\cos\psi\partial_\phi\right),\\ L_2^{(R)} &= i\left(-\cos\psi\partial_\theta + \frac{\alpha}{\tan\alpha\theta}\sin\psi\partial_\psi - \frac{\alpha}{\sin\alpha\theta}\sin\psi\partial_\phi\right),\\ L_3^{(R)} &= -i\partial_\psi, \end{aligned} \tag{4.1}$$

where $0\leqslant\phi<2\pi$, $0\leqslant\psi<4\pi$ and $0\leqslant\theta<\pi$ for $\alpha=1$, while $0\leqslant\theta<\infty$ when $\alpha=0,i$. It is rather well known that both left and right invariant generators satisfy $sl(2,c)$ Lie algebra denoted by $sl(2,c)_L$ and $sl(2,c)_R$, respectively and also they commute with each other. That is we have

$$\begin{gathered}[L_1^{(L)},L_2^{(L)}]=i\alpha^2L_3^{(L)},\quad [L_2^{(L)},L_3^{(L)}]=iL_1^{(L)},\quad [L_3^{(L)},L_1^{(L)}]=iL_2^{(L)},\\ [L_1^{(R)},L_2^{(R)}]=i\alpha^2L_3^{(R)},\quad [L_2^{(R)},L_3^{(R)}]=iL_1^{(R)},\quad [L_3^{(R)},L_1^{(R)}]=iL_2^{(R)},\\ [\mathbf{L^{(L)}},\mathbf{L^{(R)}}]=0.\end{gathered}\tag{4.2}$$

Now, using the generators (4.1) we define the following new bases:

$$\begin{gathered}K_1^{(L)}:=L_1^{(L)}\otimes I+\tfrac{1}{2}\alpha\sigma_1\quad K_2^{(L)}:=L_2^{(L)}\otimes I+\tfrac{1}{2}\alpha\sigma_2,\quad K_3^{(L)}:=L_3^{(L)}\otimes I+\tfrac{1}{2}\sigma_3,\\ K_1^{(R)}:=L_1^{(R)}\otimes I,\quad K_2^{(R)}:=L_2^{(R)}\otimes I,\quad K_3^{(R)}:=L_3^{(R)}\otimes I.\end{gathered}\tag{4.3}$$

Using the commutation relations (4.2) and the properties of Pauli matrices it is rather straightforward to show that the newly defined left and right invariant operators (4.3) also satisfy $sl(2,c)$ Lie algebra separately and commute with each other. Now, the operator F defined as

$$F:=\sqrt{2\lambda}(\sigma_1L_1^{(L)}+\sigma_2L_2^{(L)}+\alpha\sigma_3L_3^{(L)}+\alpha I),\tag{4.4}$$

commute with all the generators given in (4.3), that is

$$[F,\mathbf{K^{(L)}}]=0,\quad [F,\mathbf{K^{(R)}}]=0.\tag{4.5}$$

Therefore, in order to obtain the eigenspectrum of operator F, we need the set of commutative operators expressed in terms of operators (4.3), which are

$$\{K_3^{(R)},K_3^{(L)},K_1^{(L)2}+K_2^{(L)2}+\alpha^2K_3^{(L)2},K_1^{(R)2}+K_2^{(R)2}+\alpha^2K_3^{(R)2}\}.$$

Then, we have the following simultaneous eigenvalue equations:

$$\begin{gathered}K_3^{(R)}\Psi=q\Psi,\quad K_3^{(L)}\Psi=m\Psi,\\ (K_1^{(R)2}+K_2^{(R)2}+\alpha^2K_3^{(R)2})\Psi=\alpha^2l(l+1)\Psi,\\ (K_1^{(L)2}+K_2^{(L)2}+\alpha^2K_3^{(L)2})\Psi=\alpha^2j(j+1)\Psi.\end{gathered}\tag{4.6}$$

Obviously the operators $K_3^{(R)}$ and $\dot{K}_3^{(L)}$ have the following differential form;

$$\begin{gathered}K_3^{(R)}=\begin{pmatrix}-i\partial_\psi & 0\\ o & -i\partial_\psi\end{pmatrix},\\ K_3^{(L)}=\begin{pmatrix}-i\partial_\phi+\frac{1}{2} & 0\\ o & -i\partial_\phi-\frac{1}{2}\end{pmatrix}.\end{gathered}\tag{4.7}$$

Therefore, the two component spinor Ψ reads

$$\Psi=\begin{pmatrix}ae^{i(m-1/2)\phi+iq\psi}f_1(\theta)\\ be^{i(m+1/2)\phi+iq\psi}f_2(\theta)\end{pmatrix}.\tag{4.8}$$

In Eq. (4.8) a and b are constants and f_1 and f_2 are functions of variable θ. In order the spinor Ψ to become a periodic function of ϕ with period 2π, the quantum number m must be a half-integer number. Now, using the left and right invariant generators (4.1) we have

$$L_1^{(L)2}+L_2^{(L)2}+\alpha^2 L_3^{(L)2}=L_1^{(R)2}+L_2^{(R)2}+\alpha^2 L_3^{(R)2}$$
$$=-\left\{\partial_\theta^2+\frac{\alpha}{\tan\alpha\theta}\partial_\theta+\frac{\alpha^2}{\sin^2\alpha\theta}(\partial_\phi^2+\partial_\psi^2-2\cos\alpha\theta\partial_\phi\partial_\psi)\right\}. \tag{4.9}$$

The operator (4.9) yields the following eigenvalue equation:[14]

$$-\left\{\partial_\theta^2+\frac{\alpha}{\tan\alpha\theta}\partial_\theta+\frac{\alpha^2}{\sin^2\alpha\theta}(\partial_\phi^2+\partial_\psi^2-2\cos\alpha\theta\partial_\phi\partial_\psi)\right\}Y_{nq}^l(\theta,\phi,\psi)=\alpha^2 l(l+1)Y_{nq}^l(\theta,\phi,\psi), \tag{4.10}$$

where the eigenfunction (4.10), that is $Y_{nq}^l(\theta,\phi,\psi)$ reads

$$Y_{nq}^l(\theta,\phi,\psi)=e^{in\phi+iq\psi}P_{nq}^l(\cos\alpha\theta). \tag{4.11}$$

On the other hand, the operators $K_1^{(L)2}+K_2^{(L)2}+\alpha^2K_3^{(L)2}$ and $K_1^{(R)2}+K_2^{(R)2}+\alpha^2K_3^{(R)2}$ may be represented in the form

$$K_1^{(R)2}+K_2^{(R)2}+\alpha^2K_3^{(R)2}=\begin{pmatrix} L_1^{(R)2}+L_2^{(R)2}+\alpha^2L_3^{(R)2} & 0\\ 0 & L_1^{(R)2}+L_2^{(R)2}+\alpha^2L_3^{(R)2}\end{pmatrix},$$
$$K_1^{(L)2}+K_2^{(L)2}+\alpha^2K_3^{(L)2}$$
$$=\begin{pmatrix} L_1^{(L)2}+L_2^{(L)2}+\alpha^2L_3^{(L)2}+\frac{3}{4}\alpha^2+\alpha^2L_3^{(L)} & \alpha(L_1^{(L)}-iL_2^{(L)})\\ \alpha(L_1^{(L)}+iL_2^{(L)}) & L_1^{(L)2}+L_2^{(L)2}+\alpha^2L_3^{(L)2}+\frac{3}{4}\alpha^2-\alpha^2L_3^{(L)}\end{pmatrix}, \tag{4.12}$$

where the operators $L_1^{(L)}+iL_2^{(L)}$ and $L_1^{(L)}-iL_2^{(L)}$ are the raising and lowering operators of index n of eigenfunction (4.11), that is we have

$$(L_1^{(L)}+iL_2^{(L)})Y_{nq}^l(\theta,\phi,\psi)=\sqrt{\alpha^2(l+n+1)(l-n)}\,Y_{n+1q}^l(\theta,\phi,\psi),$$
$$(L_1^{(L)}-iL_2^{(L)})Y_{nq}^l(\theta,\phi,\psi)=\sqrt{\alpha^2(l-n+1)(l+n)}\,Y_{n-1q}^l(\theta,\phi,\psi).$$

Summarizing the above explanation the eigenfunction Ψ of eigenvalue Eq. (4.6) read

$$\Psi=\Psi_{l,j=l\pm\frac{1}{2},m,q}(\theta,\phi,\psi)=\Omega\begin{pmatrix}\pm\sqrt{l\pm m+\frac{1}{2}}\,Y_{m-(1/2)q}^l(\theta,\phi,\psi)\\ \sqrt{l\mp m+\frac{1}{2}}\,Y_{m+(1/2)q}^l(\theta,\phi,\psi)\end{pmatrix}, \tag{4.13}$$

where Ω is constant of normalization. Note that in Eq. (4.13), q takes integer values because m can take on only half-integer values as said before.[14] Now, by taking the operator F of Eq. (4.4) to the power 2 we arrive at

$$F^2=2\lambda(K_1^{(L)2}+K_2^{(L)2}+\alpha^2K_3^{(L)2}+\tfrac{1}{4}\alpha^2 I)=2\lambda(j+\tfrac{1}{2})^2\alpha^2 I.$$

Therefore, we have the following eigenvalue equation:

$$F\Psi_{l,j=l\pm1/2,m,q}(\theta,\phi,\psi)=\pm\sqrt{2\lambda(j+\tfrac{1}{2})^2\alpha^2}\,\Psi_{l,j=l\pm1/2,m,q}(\theta,\phi,\psi). \tag{4.14}$$

Now, transfering the factor $e^{iq\psi}$ which appears in the wave function (4.13) to the left of the operator F and eliminating it from both sides of Eq. (4.14), we get

$$F(q)\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)=\pm\sqrt{2\lambda(j+\tfrac{1}{2})^2\alpha^2}\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi), \tag{4.15}$$

where $F(q)$ and $\Psi_{l,j=l\pm\frac{1}{2},m,q}(\theta,\phi)$ are

$$F(q)=\sqrt{2\lambda}\begin{pmatrix} \alpha(1-i\partial_\phi) & e^{-i\phi}\left(-\partial_\theta+i\,\dfrac{\alpha}{\tan\alpha\theta}\,\partial_\phi+q\,\dfrac{\alpha}{\sin\alpha\theta}\right) \\ e^{i\phi}\left(\partial_\theta+i\,\dfrac{\alpha}{\tan\alpha\theta}\,\partial_\phi+q\,\dfrac{\alpha}{\sin\alpha\theta}\right) & \alpha(1+i\partial_\phi) \end{pmatrix},$$

$$\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)=\Omega\begin{pmatrix} \pm\sqrt{l\pm m+\frac{1}{2}}e^{i(m-1/2)\phi}P^l_{m-(1/2)\,q}(\cos\alpha\theta) \\ \sqrt{l\mp m+\frac{1}{2}}e^{i(m+1/2)\phi}P^l_{m+(1/2)\,q}(\cos\alpha\theta) \end{pmatrix}. \tag{4.16}$$

Now we consider the limiting case of $q\to 0$. In this limit the operator $F(q)$ becomes the same as the operator D_2 in (3.17), that is we have

$$\lim_{q\to 0} F(q)=D_2.$$

Specially for $\alpha=1$ we get the Dirac operator on S^2,[7] and the eigenfunction introduced in (4.16) becomes the wellknown spinor harmonics.[10]

There is another interesting limiting case: to let $\alpha\to 0$ and $l\to\infty$ but αl to remain finite, that is

$$\lim_{\alpha\to 0, l\to\infty} \alpha l=k.$$

In this limit θ plays the role of radial coordinate and we have[14]

$$\lim_{\alpha\to 0, l\to\infty} P^l_{nq}(\cos\alpha\theta)=J_{|n-q|}(kr),$$

where $J_{|n-q|}(kr)$ is the Bessel function with index $|n-q|$. In brief, we have

$$Z_{k,m,q}(r,\phi)=\lim_{\alpha\to 0, l\to\infty}\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)=\Omega'\begin{pmatrix} \pm e^{i(m-1/2)\phi}J_{|m-1/2-q|}(kr) \\ e^{i(m+1/2)\phi}J_{|m+1/2-q|}(kr) \end{pmatrix}, \tag{4.17}$$

where Ω' is the new constant of normalization. The operator $F(q)$, in the limit of $\alpha\to 0$ reads

$$\lim_{\alpha\to 0}F(q)=\sqrt{2\lambda}\begin{pmatrix} 0 & e^{-i\phi}\left(-\partial_r+\dfrac{i}{r}\,\partial\phi+\dfrac{q}{r}\right) \\ e^{i\phi}\left(\partial r+\dfrac{i}{r}\,\partial\phi+\dfrac{q}{r}\right) & 0 \end{pmatrix}. \tag{4.18}$$

In this limit the operator $F(q)$ has the following eigenvalue:

$$E=\pm\sqrt{2\lambda}k.$$

The operator (4.18) is exactly the Dirac operator of a very light spin $\frac{1}{2}$ particle in the presence of magnetic vortex with gauge potential $\mathbf{A}=e_\phi q/r$.[9] The wave function (4.17) is the eigenstate associated with the scattering of a massless fermion from a vortex. It is obvious that in the case $q=0$, (4.17) and (4.18) represent the wave function and the Dirac operator of a free massless fermion on two dimensional flat space, respectively. In the next section we obtain the eigenspectrum of the Dirac operator in the presence of the magnetic monopole (2.4) for the special case of $M=0$.

V. THE DIRAC OPERATOR ON A TWO-DIMENSIONAL MANIFOLD WITH GLOBAL CONSTANT CURVATURE IN THE PRESENCE OF THE MAGNETIC FIELD OF A MAGNETIC MONOPOLE

The massless Dirac operator on a Riemannian manifold with metric $g_{\mu\nu}$ in the presence of gauge field A_μ is[15]

$$D(A)=-i\gamma^a E_a^\mu\left(\partial_\mu+\tfrac{1}{8}\omega_{\mu ab}[\gamma^a,\gamma^b]+ieA_\mu\right). \tag{5.1}$$

Therefore, using the beins and spin connections given in (3.5) and considering the gauge potential $A=-g\cos\alpha\theta d\phi$ we obtain the following expression for the Dirac operator:

$$D_2(A)=-i\sqrt{2\lambda}\,\gamma^1\left(\partial_\theta+\frac{1}{2}\frac{\alpha}{\tan\alpha\theta}\right)-i\sqrt{2\lambda}\,\gamma^2\frac{\alpha}{\sin\alpha\theta}(\partial_\phi-ieg\cos\alpha\theta). \tag{5.2}$$

Now we try to obtain the Dirac operator $D_2(A)$ given in (5.2) from the Dirac operator on the manifold described by the metric (3.7) and by the gauge field with connection

$$A_r=0,\quad A_\theta=0,\quad A_\phi=-g\cos\alpha\theta. \tag{5.3}$$

Using the beins and spin connections given in (3.10) and gauge field connection (5.3), the Dirac operator on the manifold (3.7) in the presence of gauge field (5.3) reads

$$D_3(A)=-i\gamma^1\frac{1}{\alpha r}\left(\partial_\theta+\frac{1}{2}\frac{\alpha}{\tan\alpha\theta}\right)-i\gamma^2\frac{1}{\alpha r}\frac{\alpha}{\sin\alpha\theta}(\partial_\phi-ieg\cos\alpha\theta)-i\gamma^3\left(\partial_r+\frac{1}{r}\right). \tag{5.4}$$

It is straightforward to show that the following relation between the operators $D_2(A)$ and $D_3(A)$ holds,

$$D_2(A)=\left.\left(-i\gamma^3D_3(A)+\frac{1}{r}\right)\right|_{r=\text{constant}}. \tag{5.5}$$

The gauge field (5.3) has the following form in the Cartesian-type coordinates (3.8),

$$A_i=g\epsilon_{ij3}\frac{x_jx_3}{r(x_1^2+x_2^2)},\quad i,j=1,2,3, \tag{5.6}$$

where it satisfies the following gauge condition:

$$\mathbf{r}\cdot\mathbf{A}=0.$$

In the local coordinates (3.8) together with the gauge connection (5.6), the Dirac operator can be written as

$$D_3(A)=\sigma_i\left(\frac{1}{i}\partial_i+eA_i\right). \tag{5.7}$$

With further little algebra one can show that the Dirac operator given in (5.7) takes the following form:

$$D_3(A)=\left(\frac{\sigma_ix_i}{r}\right)^2D_3(A)=-i\gamma^3\left(\partial_r-\frac{1}{r}\boldsymbol{\sigma}.\mathbf{r}\times\left(\frac{1}{i}\boldsymbol{\nabla}+e\mathbf{A}\right)\right). \tag{5.8}$$

Finally using the relation (5.5) between the operators $D_2(A)$ and $D_3(A)$, the Dirac operator on the two-dimensional manifold (3.4) and in the presence of magnetic field of a magnetic monopole is

$$D_2(A)=F(eg)-\alpha\sqrt{2\lambda}eg\gamma^3. \tag{5.9}$$

Therefore, according to Sec. IV, we have the following eigenvalue equation:

$$D_2(A)\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)=\pm\sqrt{2\lambda\alpha^2[(j+\tfrac{1}{2})^2-q^2]}\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi), \tag{5.10}$$

where q is equal to the product of electric and magnetic charge, that is

$$q=eg.$$

Therefore, the Dirac quantization condition follows naturally from the finite representation of the $SL(2,c)$ Lie group. Also $j+\frac{1}{2}\geqslant q$ and for $j+\frac{1}{2}=q$ the operator (5.9) becomes noninvertible. It is clear that for $\alpha=1$, the operator $D_2(A)$ becomes the Dirac operator on S^2 in the presence of magnetic field of magnetic monopole[11,12] with monopole harmonics as its eigenfunctions.

VI. PARA-SUPERSYMMETRY AND SHAPE INVARIANCE OF THE DIRAC EQUATION

In this section using the left and right invariant generators introduced in Sec. IV, we try to investigate the shape invariance symmetry and para-supersymmetry of the two-dimensional Dirac operator. Here it is more convenient to work with bases $\{J_+^{(R)},J_-^{(R)},J_3^{(R)}\}$ rather than with $\{L_1^{(R)},L_2^{(R)},L_3^{(R)}\}$ which are defined as

$$\begin{aligned}J_\pm^{(R)}&=L_1^{(R)}\pm iL_2^{(R)}=e^{\pm i\psi}\left(\pm\partial_\theta+i\frac{\alpha}{\tan\alpha\theta}\partial_\psi-i\frac{\alpha}{\sin\alpha\theta}\partial_\phi\right),\\ J_3^{(R)}&=L_3^{(R)}=-i\partial_\psi.\end{aligned} \tag{6.1}$$

Clearly these new bases have the following commutation relations:

$$\begin{aligned}[][J_+^{(R)},J_-^{(R)}]&=2\alpha^2J_3^{(R)},\\ [J_3^{(R)},J_\pm^{(R)}]&=\pm J_\pm^{(R)}.\end{aligned} \tag{6.2}$$

Using the relations (4.9) and (4.10) we arrive at

$$\begin{aligned}(J_+^{(R)}\otimes I)(J_-^{(R)}\otimes I)\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi,\psi)&=\alpha^2(l-q+1)(l+q)\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi,\psi),\\ (J_-^{(R)}\otimes I)(J_+^{(R)}\otimes I)\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi,\psi)&=\alpha^2(l+q+1)(l-q)\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi,\psi).\end{aligned} \tag{6.3}$$

The above relations indicate that $J_+^{(R)}\otimes I$ and $J_-^{(R)}\otimes I$ are raising and lowering operators of index q, respectively, that is,

$$\begin{aligned}J_+^{(R)}\otimes I\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi,\psi)&=\sqrt{\alpha^2(l+q+1)(l-q)}\Psi_{l,j=l\pm 1/2,m,q+1}(\theta,\phi,\psi),\\ J_-^{(R)}\otimes I\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi,\psi)&=\sqrt{\alpha^2(l-q+1)(l+q)}\Psi_{l,j=l\pm 1/2,m,q-1}(\theta,\phi,\psi).\end{aligned} \tag{6.4}$$

Now by transfering $e^{iq\psi}$, which is only ψ dependent factor in the eigenspinor $\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi,\psi)$, to the left-hand sides of the lowering and raising operator in (6.4) we arrive at

$$J_+^{(R)}(q)\otimes I\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)=\sqrt{\alpha^2(l+q+1)(l-q)}\Psi_{l,j=l\pm 1/2,m,q+1}(\theta,\phi),$$
$$J_-^{(R)}(q)\otimes I\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)=\sqrt{\alpha^2(l-q+1)(l+q)}\Psi_{l,j=l\pm 1/2,m,q-1}(\theta,\phi), \tag{6.5}$$

where $J_\pm^{(R)}(q)$ read

$$J_\pm^{(R)}(q)=\pm\frac{\partial}{\partial\theta}-i\frac{\alpha}{\sin\alpha\theta}\frac{\partial}{\partial\phi}-q\frac{\alpha}{\tan\alpha\theta}. \tag{6.6}$$

But $J_+^{(R)}(q)\otimes I$ and $J_-^{(R)}(q)\otimes I$ are still raising and lowering operators of index q of the eigenspinors $\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)$ and the relations (6.5) indicate that $\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)$ can form the basis for a representation of para-supersymmetry of order p, where p is an arbitrary integer. According to Refs. 13 and 16, the nonunitary para-supersymmetric algebra of order p can be generated by parafermionic generators of order p, denoted by Q_1 and Q_2 and a bosonic generator H, which satisfy the following relations:

$$Q_1^pQ_2+Q_1^{p-1}Q_2Q_1+\cdots+Q_1Q_2Q_1^{p-1}+Q_2Q_1^p=2pQ_1^{p-1}H, \tag{6.7a}$$

$$Q_2^pQ_1+Q_2^{p-1}Q_1Q_2+\cdots+Q_2Q_1Q_2^{p-1}+Q_1Q_2^p=2pQ_2^{p-1}H, \tag{6.7b}$$

$$Q_1^{p+1}=Q_2^{p+1}=0, \tag{6.7c}$$

$$[H,Q_1]=[H,Q_2]=0. \tag{6.7d}$$

By introducing the operators,

$$X_+(q):=J_+^{(R)}(q)\otimes I,\quad X_-(q):=J_-^{(R)}(q)\otimes I,\quad \mathcal{H}_q:=H_q\otimes I, \tag{6.8}$$

we can represent the generators Q_1, Q_2, and H by the following $(p+1)\times(p+1)$ matrices of the form:

$$(Q_1)_{qq'}:=X_-(q)\delta_{q+1,q'},$$
$$(Q_2)_{qq'}:=X_+(q'-1)\delta_{q,q'+1}, \tag{6.9}$$
$$(H)_{qq'}:=\mathcal{H}_q\delta_{q,q'}\quad q,q'=1,\ldots,p+1,$$

where each element of these matrices is a 2×2 matrix. In (6.9) we need to choose the Hamiltonians $\mathcal{H}$, with $q=1,...,p+1$ so that the generators (6.9) satisfy the para-supersymmetric algebraic relations (6.7). The generators Q_1, Q_2, and H, as defined in (6.9), satisfy the Eq. (6.7c), but Eqs. (6.7a) and (6.7b) lead to the following equations:

$$X_+(p-2)\cdots X_+(1)X_+(0)X_-(1)+\cdots+X_+(p-2)X_-(p-1)X_+(p-2)X_-(p-3)\cdots X_+(0)$$
$$+X_-(p)X_+(p-1)X_+(p-2)\cdots X_+(0)=2pX_+(p-2)X_+(p-3)\cdots X_+(0)\mathcal{H}_1, \tag{6.10a}$$

$$X_+(p-1)\cdots X_+(1)X_+(0)X_-(1)+X_+(p-1)\cdots X_+(2)X_+(1)X_-(2)X_+(1)+\cdots$$
$$+X_+(p-1)X_-(p)X_+(p-1)X_+(p-2)\cdots X_+(1)=2pX_+(p-1)X_+(p-2)\cdots X_+(1)\mathcal{H}_2, \tag{6.10b}$$

$$X_-(1)\cdots X_-(p-1)X_-(p)X_+(p-1)+X_-(1)\cdots X_-(p-2)X_-(p-1)X_+(p-2)X_-(p-1)$$
$$+\cdots+X_-(1)X_+(0)X_-(1)X_-(2)\cdots X_-(p-1)=2pX_-(1)X_-(2)\cdots X_-(p-1)\mathcal{H}_p, \tag{6.10c}$$

$$X_-(2)\cdots X_-(p-1)X_-(p)X_+(p-1)X_-(p)+\cdots+X_-(2)X_+(1)X_-(2)X_-(3)\cdots X_-(p)$$
$$+X_+(0)X_-(1)X_-(2)\cdots X_-(p)=2pX_-(2)X_-(3)\cdots X_-(p)\mathcal{H}_{p+1}. \tag{6.10d}$$

Finally Eq. (6.7d) imply the following equations:

$$\mathcal{H}_q X_-(q)=X_-(q)\mathcal{H}_{q+1},$$
$$\mathcal{H}_{q+1}X_+(q-1)=X_+(q-1)\mathcal{H}_q. \tag{6.11}$$

Now, defining the Hamiltonians $\mathcal{H}_q$, $q=1,...,p$ as

$$\mathcal{H}_q=\tfrac{1}{2}X_-(q)X_+(q-1)+\tfrac{1}{2}C_qI \quad q=1,2,...,p,$$
$$\mathcal{H}_{p+1}=\tfrac{1}{2}X_+(p-1)X_-(p)+\tfrac{1}{2}C_pI. \tag{6.12}$$

By definitions (6.12) relations (6.11) are satisfied for the special case $q=p$. In order for the relations (6.11) to be satisfied for $q=1,...,p-1$, too, we need to choose the constants C_q as

$$E_{q+1}-E_q=C_q-C_{q+1}, \tag{6.13}$$

where

$$E_q:=\alpha^2[l(l+1)-q(q-1)]. \tag{6.14}$$

To obtain the relation (6.13) we have used the following shape invariance property between the operators $X_\pm(q)$:

$$X_+(q-1)X_-(q)-X_-(q+1)X_+(q)=E_q-E_{q+1}. \tag{6.15}$$

Substituting (6.12) in formula (6.10a), and also using the shape invariance property (6.15) we obtain

$$C_1=\frac{1}{p}[(1-p)E_1+E_2+E_3+\cdots+E_p]. \tag{6.16}$$

Finally combining (6.13) with (6.16) we obtain

$$C_q=\frac{1}{p}\sum_{q'=1}^{p}E_{q'}-E_q. \tag{6.17}$$

From (6.17) we can see that the following relation among the constants C_q holds

$$C_1+C_2+\cdots+C_p=0.$$

Using the substitution (6.12) and the shape invariance property (6.15) together with the constants C_q given in (6.17) one can straightforwardly show that Eqs. (6.10b), (6.10c), (6.10d) are satisfied, too. Also using the result given in (6.17) and the shape invariance relation (6.15) it follows that the Hamiltonians $\mathcal{H}_q$ are isospectral and we have

$$\mathcal{H}_q\Psi_{l,j=l\pm 1/2,m,q-1}(\theta,\phi)=E\Psi_{l,j=l\pm 1/2,m,q-1}(\theta,\phi), \quad q=1,...,p+1, \tag{6.18}$$

with

$$E=\frac{1}{2p}\sum_{q=1}^{p}E_q. \tag{6.19}$$

Substituting E_q in (6.19) and by using the relation (6.14) we get

$$E=\tfrac{1}{6}\alpha^2[3l(l+1)+1-p^2]. \tag{6.20}$$

In a similar manner by substituting (6.14) in (6.17) we have

$$C_q=\tfrac{1}{3}\alpha^2[3q(q-1)+1-p^2]. \tag{6.21}$$

Substituting the constants C_q in (6.21), and also using the relations (6.6) and (6.8) after the substituting (6.12) we obtain the explicit differential form of the Hamiltonian $\mathcal{H}_q$ as

$$\mathcal{H}_q=-\frac{1}{2}\left[\frac{\partial^2}{\partial\theta^2}+\frac{\alpha}{\tan\alpha\theta}\frac{\partial}{\partial\theta}+\frac{\alpha^2}{\sin^2\alpha\theta}\frac{\partial^2}{\partial\phi^2}-\frac{2i(q-1)\alpha^2}{\sin\alpha\theta\tan\alpha\theta}\frac{\partial}{\partial\phi}-\frac{(q-1)^2\alpha^2}{\sin^2\alpha\theta}+\frac{1}{3}\alpha^2(p^2-1)\right]\otimes I. \tag{6.22}$$

The bases of the representation of para-supersymmetric algebra of order p can be represented by column matrix with $(p+1)$ row, that is

$$(\Psi_{l,j=l\pm 1/2,m}(\theta,\phi))_q:=\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi),\quad q=0,1,...,p, \tag{6.23}$$

where $\Psi_{l,j=l\pm 1/2,m}(\theta,\phi)$ is the eigenvector of paraboson operator H with eigenvalue E, that is

$$H\Psi_{l,j=l\pm 1/2,m}(\theta,\phi)=E\Psi_{l,j=l\pm 1/2,m}(\theta,\phi). \tag{6.24}$$

It follows rather trivially from the commutation relations (6.7d) and also from the relations (6.7c) that $Q_1^q\Psi_{l,j=l\pm 1/2,m}(\theta,\phi)$ and $Q_2^q\Psi_{l,j=l\pm 1/2,m}(\theta,\phi)$ for $q=1,...,p$ are eigenstates of the bosonic generator H with the corresponding eigenvalue E. Hence, it follows that

$$\Psi_{l,j=l\pm 1/2,m,q}(\theta,\phi)=\frac{X_-(q+1)}{\sqrt{E_{q+1}}}\frac{X_-(q+2)}{\sqrt{E_{q+2}}}\cdots\frac{X_-(q'+q)}{\sqrt{E_{q'+q}}}\Psi_{l,j=l\pm 1/2,m,q'+q}(\theta,\phi),$$
$$q=0,1,...,p-q'. \tag{6.25}$$

The representation of para-supersymmetry algebra of order p is valid even for the special case of $\alpha=0$. Using the relation (4.17) in the limiting cases of $\alpha\to 0$ and $l\to\infty$ we arrive at the following shape invariance relations:

$$\begin{aligned}\left(\frac{\partial}{\partial r}-\frac{i}{r}\frac{\partial}{\partial\phi}-\frac{q}{r}\right)\otimes IZ_{k,m,q}(r,\phi)&=kZ_{k,m,q+1}(r,\phi),\\ \left(-\frac{\partial}{\partial r}-\frac{i}{r}\frac{\partial}{\partial\phi}-\frac{q}{r}\right)\otimes IZ_{k,m,q}(r,\phi)&=kZ_{k,m,q-1}(r,\phi).\end{aligned} \tag{6.26}$$

Note that in the limiting case of $\alpha=0$ there is neither any highest nor lowest state in the realization of para-supersymmetry, since there is no first order differential operator which can kill any of the Bessel functions. In this case we can have a para-supersymmetry of infinite order where the Bessel functions form its bases.

VII. SOLUTION OF THE DIRAC EQUATION ON THREE-DIMENSIONAL MANIFOLDS WITH A LOCAL CONSTANT CURVATURE

For nonvanishing M, the spatial part of the metric given in (2.2) describes a two-dimensional manifold with local constant curvature. In this section we try to solve the Dirac equation of a point mass $\mathcal{M}$ and charge e on the manifold with metric (2.2), in the presence of magnetic field with connection (2.6). The angular deficit due to the presence of a very heavy point mass M destroys the global constancy of the curvature, hence we lose both the degeneracy and para-supersymmetry. Therefore, we cannot solve the Dirac equation by the algebraic methods any longer. Thus we have to solve it by an ordinary method of solution of coupled first order differential equations. In a flat $(2+1)$-space–time one can represent the Dirac γ matrices as[4,5]

$$\gamma^0=\sigma_3, \quad \gamma^1=i\sigma_2, \quad \gamma^2=-i\sigma_1, \tag{7.1}$$

where γ^a, $a=0,1,2$ close the Clifford algebra as

$$\{\gamma^a,\gamma^b\}=2\eta^{ab}, \quad a,b=0,1,2. \tag{7.2}$$

The Minkowski metric η_{ab} has the following signature:

$$\eta_{ab}=\mathrm{diag}(+,-,-).$$

One can write the metric (2.2) in the following form:

$$ds^2=dt^2-\rho^2\left(\beta^{-2}d\theta^2+\frac{\sin^2\alpha\theta}{\alpha^2}d\phi^2\right), \tag{7.3}$$

where ρ^2 and β are defined as

$$\beta=1-GM,$$

$$\rho^2=\frac{(1-GM)^2}{2\lambda}.$$

Since λ is negative when $\alpha=i$ we have $\rho^2<0$, hence the metric (7.3) is Euclidean while for other values of α it is Minkowskian. We can choose the three-beins associated with the metric (7.3) as

$$e^a_\mu=\begin{pmatrix} 1 & 0 & 0 \\ 0 & \rho\dfrac{1}{\beta}\cos\phi & \rho\dfrac{1}{\beta}\sin\phi \\ 0 & -\rho\dfrac{\sin\alpha\theta}{\alpha}\sin\phi & \rho\dfrac{\sin\alpha\theta}{\alpha}\cos\phi \end{pmatrix}. \tag{7.4}$$

With the above choice of 3-beins we can compare our results in a special case with those of Ref. 4. The inverse of matrix (7.4) is

$$E^\mu_a=\begin{pmatrix} 1 & 0 & 0 \\ 0 & \dfrac{\beta}{\rho}\cos\phi & -\dfrac{1}{\rho}\dfrac{\alpha}{\sin\alpha\theta}\sin\phi \\ 0 & \dfrac{\beta}{\rho}\sin\phi & \dfrac{1}{\rho}\dfrac{\alpha}{\sin\alpha\theta}\cos\phi \end{pmatrix}. \tag{7.5}$$

According to Ref. 4, the Dirac equation in $(2+1)$-space–time for a fermion with mass $\mathcal{M}$ and electric charge e, in the presence of a gauge field with gauge connection A_μ is

$$\left[i\gamma^a E^\mu_a\left(\partial_\mu-\frac{i}{2}\omega^b_\mu\gamma_b+ieA_\mu\right)-\mathcal{M}\right]\Psi(t,\theta,\phi)=0, \tag{7.6}$$

where ω^a_μ is given by

$$\epsilon^{\kappa\mu\nu}\partial_\mu e^a_\nu=\epsilon^{\kappa\mu\nu}\epsilon^a_{bc}\omega^b_\mu e^c_\nu. \tag{7.7}$$

Hence

$$\omega_\mu^a = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta\cos\alpha\theta - 1 & 0 & 0 \end{pmatrix}. \tag{7.8}$$

Therefore, the Dirac Eq. (7.6) can be written as

$$\left\{ i\left[\gamma^0\partial_t + \frac{1}{\rho}\gamma^\theta\left(\beta\partial_\theta - \frac{\alpha}{2\sin\alpha\theta}(1-\beta\cos\alpha\theta)\right) + \frac{1}{\rho}\frac{\alpha}{\sin\alpha\theta}\gamma^\phi(\partial_\phi + ieA_\phi)\right] - \mathcal{M}\right\}\Psi(t,\theta,\phi) = 0, \tag{7.9}$$

where γ^θ and γ^ϕ are defined as

$$\gamma^\theta = \cos\phi\,\gamma^1 + \sin\phi\,\gamma^2, \quad \gamma^\phi = -\sin\phi\,\gamma^{1+}\cos\phi\,\gamma^2.$$

In the limiting case of $\alpha \to 0$, the coordinate θ becomes similar to a radial coordinate r and the Dirac equation (7.9) in the absence of gauge field A_ϕ becomes exactly the Dirac equation associated with a massive fermion on a $(2+1)$-space–time dimension with conical spatial part.[4] Now let the Dirac spinor have the following time dependence

$$\Psi(t,\theta,\phi) = e^{-iEt}\Psi(\theta,\phi),$$

together with the following ϕ dependence:

$$\Psi(\theta,\phi) = \begin{pmatrix} e^{i(m-1/2)\phi} f_1(\theta) \\ e^{i(m+1/2)\phi} f_2(\theta) \end{pmatrix}, \tag{7.10}$$

where m is a half-integer number. One can show that the functions $f_1(\theta)$ and $f_2(\theta)$ satisfy the following differential equations:

$$\left\{ (1-z^2)\frac{d^2}{dz^2} - 2z\frac{d}{dz} - \frac{1}{1-z^2}\left(\left(\frac{m}{\beta}\right)^2 + \left(eg \pm \frac{1}{2}\right)^2 - 2\frac{m}{\beta}\left(eg \pm \frac{1}{2}\right)z\right)\right\} f_{1,2}(z)$$
$$= -\left(\frac{\rho^2}{\alpha^2\beta^2}(E^2 - \mathcal{M}^2) + \left(e^2g^2 - \frac{1}{4}\right)\right) f_{1,2}(z), \tag{7.11}$$

with z defined as

$$z = \cos\alpha\theta.$$

Now defining

$$\frac{\rho^2}{\alpha^2\beta^2}(E^2 - \mathcal{M}^2) + (e^2g^2 - \tfrac{1}{4}) = c(c+1),$$

with c as a real number, one can give the solutions of Eq. (7.11) in terms of hypergeometric functions.[17] In the limiting cases of $\alpha, e \to 0$ and $c \to \infty$ such that the product αc remains constant we obtain[17]

$$\lim_{\alpha\to 0, c\to\infty, e\to 0} f_{1,2}(\theta) = J_{|m/\beta \mp \frac{1}{2}|}(\alpha c r). \tag{7.12}$$

Writing the half-integer number m as

$$m = n + \tfrac{1}{2},$$

with n as an arbitrary integer the Dirac Eq. (7.9) takes the following solution in the above mentioned limit:

$$\Psi(t,r,\phi)=e^{-iEt}\exp i\left(n+\tfrac{1}{2}-\tfrac{1}{2}\sigma_3\right)\phi\begin{pmatrix} AJ_{|(1/\beta)(n+(1-\beta)/2)|}(\alpha c r) \\ BJ_{|(1/\beta)(n+(1-\beta)/2)+1|}(\alpha c r)\end{pmatrix}, \tag{7.13}$$

where A and B are arbitrary constants. The solution (7.13) is eigenstate of Dirac Hamiltonian associated with a fermion on $(2+1)$ spacetime with conical spatial part[4] with corresponding eigenvalues,

$$E=\pm\sqrt{\mathcal{M}^2+\frac{1}{\rho^2}\beta^2\alpha^2c^2}. \tag{7.14}$$

[1] S. Deser, R. Jackiw, and G. 'T Hooft, Ann. Phys. (N.Y.) **152**, 220 (1984).
[2] S. Deser and R. Jackiw, Ann. Phys. (N.Y.) **153**, 405 (1985).
[3] S. Deser and R. Jackiw, Commun. Math. Phys. **118**, 495 (1988).
[4] P. Gerbert and R. Jackiw, Commun. Math. Phys. **124**, 229 (1989).
[5] P. Gerbert, Phys. Rev. D **40**, 1346 (1989).
[6] M. A. Jafarizadeh, H. Fakhari, and S. K. Moayedi, ''Solution of (2+1)-dimensional Einstein–Maxwell theory with parasupersymmetry,'' to appear in Phys. Rev. D.
[7] H. Grosse, C. Klimcik, and P. Presnajder, Commun. Math. Phys. **185**, 155 (1997).
[8] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).
[9] B. Thaller, *The Dirac Equation* (Springer, Berlin, 1992).
[10] T. T. Wu and C. N. Yang, Nucl. Phys. B **107**, 365 (1976).
[11] H. Grosse and P. Presnajder, Lett. Math. Phys. **33**, 171 (1995).
[12] H. Grosse, C. Klimcik, and P. Presnajder, ''Simple field theoretical models on noncommutative manifolds,'' Paper No. hep-th/9510177, 1995.
[13] M. A. Jafarizadeh and H. Fakhri, Ann. Phys. (N.Y.) **266**, 178 (1998).
[14] N. Ja Vilenkin and A. U. Klimyk, *Representation of Lie Groups and Special Functions* (Kluwer Academic, Dordrecht, 1991).
[15] M. Nakahara, *Geometry, Topology, and Physics* (Adam Hilger, New York, 1989).
[16] M. A. Jafarizadeh and H. Fakhri, Ann. Phys. (N.Y.) **262**, 260 (1998).
[17] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1965).

On a purported local extension of the quantum formalism

Joseph Melia
Department of Philosophy, University of York, York, Y01 5DD, United Kingdom

Michael Redhead
34 Coniger Road, London SW6 3TA, United Kingdom

(Received 22 April 1998; accepted for publication 5 March 1999)

It is widely believed that Bell has proved there can be no consistent local extension of the quantum formalism. Against this, Angelidis has presented a hidden variable theory which, he claims, makes precisely the same predictions as the quantum formalism and which also satisfies locality. In this note, we argue that Angelidis' theory does not live up to its inventor's claims. © *1999 American Institute of Physics.* [S0022-2488(99)02607-9]

I. INTRODUCTION

Since the early days of quantum mechanics, a number of physicists have doubted whether quantum mechanics was a complete theory and wondered whether it was possible to extend the quantum formalism by adjoining hidden variables.[1] In 1952, Bohm answered this question in the affirmative[2] and in doing so refuted von Neumann's influential yet flawed proof that no such extension was possible.[3] However, Bohm's hidden variable theory has not won wide support partly because the theory is *nonlocal*: there is instantaneous action at a distance. Since there is an obvious problem reconciling such nonlocal theories with Relativity, hidden variable theories would look much more promising if they also satisfied locality. Accordingly, the question as to whether or not *local* hidden variable theories are possible assumes great significance. In 1964 Bell appeared to prove that this question had a negative answer:[4] He showed that any local hidden variables theory is incompatible with certain quantum mechanical predictions. Since these predictions have been borne out by the experiments of Aspect and others[5] the prospects for hidden variable theories have looked grim.

Angelidis disagrees.[6] He claims to have done to Bell what Bohm did to von Neumann: He has found a theory which is local and which generates a family of probability functions converging uniformly to the probability function generated by quantum mechanics. If this were true, then Angelidis' theory would be a counterexample to Bell's theorem and a promising path would once again be open to hidden variable theorists.

Unfortunately, Angelidis' theory fails to live up to his claims: As formulated, the theory does not make the same predictions as quantum mechanics, and while there is a natural extension of his theory which does make the same predictions, the extension is *not* local. Bell's Theorem stands.

II. ANGELIDIS' THEORY

The disagreement between Angelidis and Bell can most easily be understood by considering the following thought experiment, due originally to Einstein, Podolsky, and Rosen and later simplified by Bohm.[7] In this experiment, photons γ_1 and γ_2, created by the spontaneous annihilation decay of the nonfactorizable singlet state $|\gamma_1\gamma_2\rangle$, are emitted in opposite directions and arrive at polarizers P_1 and P_2, respectively. Behind each polarizer lies a photon detector. If α and β represent the angles of polarization of P_1 and P_2 then, according to quantum mechanics, the probability that both detectors register a photon is $1/2\cos^2(\alpha-\beta)$. Could a local hidden variable theory assign the same probabilities to this experiment as quantum mechanics?

To answer this, we need to know just what locality entails. First, let us fix our terminology. Let QF stand for the classical quantum formalism. Let $p^T{}_{12}(\alpha,\beta)$ be the probability that a theory

T assigns to both detectors registering a photon given that the angles of polarization are α and β. Let λ represent our hidden variable and Λ the set of values the hidden variable could take. Let $p^*{}_1(\lambda,\alpha)(p_2^*(\lambda,\beta))$ be the chance that the photon passes through $P_1(P_2)$ given that the system is in state λ and the angle of polarization is $\alpha(\beta)$. Finally, let $\rho(\lambda)$ be a weight function which represents the chance that the hidden variable takes the value λ.

Bell and Angelidis agree that any local theory should meet the following constraints:

$$\text{(L1)}\ p^T{}_{12}(\alpha,\beta)=\int_\Lambda \rho(\lambda)p^*{}_1(\lambda,\alpha)p^*{}_2(\lambda,\beta),$$

where the function $p^*{}_1$ must not depend upon the variable β and the function $p^*{}_2$ must not depend upon the variable α.

(L2) The specified range Λ of the variable λ must depend upon neither the variable α nor the variable β.

(L3) The function ρ must depend upon neither the variable α nor the variable β.[8]

Bell's claim is that no hidden variable theory which meets constraints (L1)–(L3) can yield the same statistical predictions as QF. According to Bell, the QF probability function $p^{\mathrm{QF}}{}_{12}$ cannot be represented, either precisely or arbitrarily closely in the form

$$\forall\alpha,\beta\left[1/2\cos^2(\alpha-\beta)=\int_\Lambda \rho(\lambda)p^*{}_1(\lambda,\alpha)p^*{}_2(\lambda,\beta)d\lambda.\right.$$

According to Angelidis, you can. Consider the theory T which consists of the following four postulates:

$$(\Pi_1)\quad p^*{}_1(\lambda,\alpha)=\cos^2(\lambda-\alpha),$$

$$(\Pi_2)\quad p^*{}_2(\lambda,\beta)=\cos^2(\lambda-\beta),$$

$$(\Pi_3)\quad \rho(\lambda,\mu):=1/2\left[\delta(\lambda-\mu)+\delta\left(\lambda-\mu+\frac{\pi}{2}\right)\right],$$

$$(\Pi_4)\quad \Lambda:=\{\lambda|-\infty<\lambda<+\infty\}.$$

One can think of the hidden variable λ as a common plane of polarization of the two photons emitted when the atom decays. The functions $p^*{}_i(\lambda,\gamma)$ represent the probabilities that a photon will be detected at wing i ($i=1$ or $i=2$) given that the photons are plane polarized in the λ direction *or* in the $\lambda-\frac{1}{2}\pi$ direction, and the polarizer P_i is set in the γ direction.

The third postulate is the ''conditional probability distribution for the spherically symmetric singlet state $|\gamma_1,\gamma_2\rangle$ to spontaneously disintegrate into two back to back photons plane-polarized in a *specific* but randomly chosen direction, given by a variable μ, out of *all* the equally likely choices of directions...''[9] δ is simply the Dirac delta function and the final postulate does nothing more than specify the range of λ.

T generates a family of functions $p^\mu{}_{12}$ such that

$$p^\mu{}_{12}(\alpha,\beta)=\int_\Lambda \rho(\lambda,\mu)p^*{}_1(\lambda,\alpha)p^*{}_2(\lambda,\beta)d\lambda=1/4[1+\cos 2(\mu-\alpha)\cos 2(\mu-\beta)],$$

and families of functions $p^\mu{}_1$ and $p^\mu{}_2$ such that

$$p^\mu{}_1=\int_\Lambda \rho(\lambda,\mu)p^*{}_1(\lambda,\alpha)d\lambda=\tfrac{1}{2},$$

$$p^{\mu}{}_2=\int_{\Lambda}\rho(\lambda,\mu)p^{*}{}_2(\lambda,\beta)d\lambda=\tfrac{1}{2}.$$

Finally, theory T entails the following important sentence (Σ):

$$(\Sigma)\quad (\forall\epsilon>0)(\exists\eta>0)(\forall\mu\in M)(\forall\alpha,\beta\in D)[(|\mu-\alpha|<\eta)\vee(|\mu-\beta|<\eta)\rightarrow|p^{\mu}{}_{12}(\alpha,\beta)-p_{12}(\alpha,\beta)|<\epsilon].$$

A logically equivalent way of writing this sentence is

$$(\Sigma)\quad (\forall\epsilon>0)(\exists\eta>0)(\forall\mu\in M)(\forall\alpha,\beta\in D)[(\mu\in S_{\alpha}\cup S_{\beta})\rightarrow|p^{\mu}{}_{12}(\alpha,\beta)-p_{12}(\alpha,\beta)|<\epsilon]$$

where $S_{\alpha}=\{\mu|-\eta+\alpha<\mu<\alpha+\eta\}$ and $S_{\beta}=\{\mu|-\eta+\beta<\mu<\beta+\eta\}$.

According to Angelidis, (Σ) ''expresses the *formal definition* of the uniform convergence of the family of functions $\{p^{\mu}{}_{12}|\mu\in M\}$ to the function $p^{\mathrm{QF}}{}_{12}$.''

Angelidis bases his physical interpretation of this theory around (Σ): ''For *any* chosen values of α and β, whenever a value of μ, characterising the random direction of the common plane of polarization of a single pair of back to back photons, *happens by pure chance* to belong to subset S_{α} or S_{β}, this single pair of back to back photons gets through polarisers P_1 and P_2 *and causes* a coincidence count with probability given by a *value* of the QF probability function $p^{\mathrm{QF}}{}_{12}$.'' [10]

So if μ is close to either α or β, then the chance of a coincidence count is close to the chance predicted by QF. But what if μ is not close to α or β? Well, in that case, (Σ) is still true just because the antecedant is false. However, we cannot infer that the ''single pair of back to back photons with $\mu_1\in M$ causes a coincidence count with probability $\frac{1}{2}\cos^2(\alpha_1-\beta_1)$. But the single pair of back to back photons with $\mu_1\in M$ may fall inside another subset, say, $S_{\alpha 4}$ or $S_{\beta 4}$ of the set M... so that it causes a coincidence count with a different probability $\frac{1}{2}\cos^2(\alpha_4-\beta_4)$.'' Angelidis concludes that ''The universal quantifiers $(\forall\mu\in M)$ and $(\forall\alpha,\beta\in D)$ occurring in the prefix of the sentence Σ take into account the *whole* array of such possibilities... so that the detectors accordingly register coincidence (and single) counts with the same probabilities as those given by QF for *each* and *every* pair of back to back photons emitted by the source.'' [11]

This ends the summary of Angelidis' theory. I shall now argue that the paper contains two flaws: (1) Angelidis' family of functions does *not* converge uniformly to the QF probability function; (2) Angelidis' theory does not predict the same probability count as those given by QF for each and every pair of back to back photons emitted by the source.

III. UNIFORM CONVERGENCE

Let us examine a little more closely Angelidis' notion of uniform convergence.

We know when a countable sequence of functions $\{q^n|n\in N\}$ defined on some domain D uniformly converges to q: they converge uniformly if, for any small number ϵ we please, there is an n such that any $q^{n'}$ (with n' larger than n) is within an ϵ of q for *any* value of q and $q^{n'}$. More formally:

$$(\forall\epsilon>0)(\exists n\in N)(\forall n'\in N)(n'>n\rightarrow\forall\alpha\beta\{|q^n(\alpha,\beta)-q(\alpha,\beta)|<\epsilon\}).$$

However, since Angelidis' theory deals with the uniform convergence of an *uncountable* family of functions, the definition must be extended to cover this case. So when does the set $\{f^{\mu}|\mu\in M\}$, with M uncountable, converge to g?

Angelidis extends the definition of uniform convergence by introducing the notion of a direction: N is a *direction* in X precisely when (a) N is a set of subsets of X partially ordered by reverse inclusion; (b) for any $x,y\in N$ there is a $z\in N$ with $z\subseteq x$ and $z\subseteq y$. Example: if X is the set of real numbers, then the set of basic neighborhoods containing the number 0 is a direction in X.[12]

Then Angelidis' definition of *uniform convergence* is as follows: Let D^2 be a subset of R^2, and let N be a direction in M. The family of functions $\{f^{\mu}|\mu\in M\}$ is said to converge uniformly to g on D^2 if for every $\epsilon>0$ there exists an $\eta>0$ (with η depending only on ϵ) corresponding to a basic neighborhood N_η in N such that for any μ in M and any x in D^2 *whenever* the values of μ are in N_η then $|f^\mu(x)-g(x)|<\epsilon$ holds. In symbols this becomes,

$$(\forall \epsilon>0)(\exists \eta>0)(\forall \mu \in M)(\forall x \in D^2)(\mu \in N_\eta \rightarrow |p^\mu{}_{12}(\alpha,\beta)-p_{12}(\alpha,\beta)|<\epsilon].$$

Now, it isn't at all clear what the N_η are supposed to be here. Angelidis tells us that they are basic neighborhoods (unlike Angelidis' N_x) and it is natural to think that they are basic neighborhoods of η. But then, why quantify over the variable η? And indeed, it would be perfectly all right to say that $\{f^\mu|\mu \in M\}$ uniformly converges to f^η iff, for any ϵ there is some basic neighborhood of η such that any μ in N_η, $|f^\mu(x)-f^\eta(x)|<\epsilon$. But here η is a *name* for an element of R—it is not a free variable which can be quantified over; nor is there any reason why η has to be greater than zero.

The ambiguity of the N_η allows Angelidis to make a serious mistake in his formal definition of uniform convergence. Angelidis claims that sentence (Σ) expresses the formal definition of uniform convergence. Recall that this sentence is

$$(\Sigma) \quad (\forall \epsilon>0)(\exists \eta>0)(\forall \mu \in M)(\forall \alpha,\beta \in D)[(|\mu-\alpha| < \eta)\vee(|\mu-\beta|<\eta)\rightarrow|p^\mu{}_{12}(\alpha,\beta)-p_{12}(\alpha,\beta)|<\epsilon].$$

.

In this case $N_\eta=\{\mu|\alpha-\eta<\mu<\alpha+\eta\}$. Again, this significantly differs from Angelidis' own definition of S_α on p. 1645, where $S_\alpha=\{\mu|\alpha-2\epsilon<\mu<\alpha+2\epsilon\}$. For N_η the subscript is an index of the distance from α that the μ in N_η are allowed to be. For S_α the subscript tells us which value of D the μ in S_α are close to.

Worse still, (Σ) does *not* express the notion of uniform convergence. For (Σ) says that if μ is close to α or is close to β then $p^\mu{}_{12}$ is close to p_{12} *at* (α,β). We require something more of uniform convergence—we require that if μ be close to α or β then $p^\mu{}_{12}$ be close to p_{12} for *all* values of these functions. To see how short of uniform convergence Angelidis' definition falls, consider the family of functions $\{q^\mu(\alpha):=\alpha-\mu\}$. Let $q(\alpha)$ be the zero function (so $q(\alpha)=0$ for all α). Now, by letting $\eta=\epsilon$ it is easy to see that

$$(\forall \epsilon>0)(\exists \eta>0)(\forall \mu \in R)(\forall \alpha \in R)(|\mu-\alpha|<\eta\rightarrow|q^\mu(\alpha)-q(\alpha)|<\epsilon).$$

So, if μ is close to α then q^μ is close to q *at* α. But there is no reasonable sense of *uniform convergence* on which the family of functions can be said to converge to the zero function. True, for any μ and for any x, if μ is sufficiently close to x then the function q^μ is sufficiently close to the function q *at the point* α—but this is a far cry from implying that the function q^μ is close to q for all values of α.

It is clear that a family of functions f^μ will *not* uniformly converge to the function g if there is some ϵ such that, for every μ there is some α,β with $|f^\mu(\alpha,\beta)-g(\alpha,\beta)|\geqslant\epsilon$. For in such a case, the family is always at least an ϵ away from g at some point $\langle\alpha,\beta\rangle$. In Angelidis' theory, we can find an ϵ such that ϵ equals 1/4. For, for any μ let $\alpha=\mu+45$ and let $\beta=\mu-45$. Now,

$$p^{\mathrm{QF}}{}_{12}(\alpha,\beta)=1/2\cos^2(\alpha-\beta)=1/2\cos^2(90)=0$$

while

$$p^\mu{}_{12}(\alpha,\beta)=1/4[1+\cos 2(\mu-\alpha)\cos 2(\mu-\beta)]=1/4[1+\cos 2(-45)\cos 2(45)]=1/4.$$

Since every one of Angelidis' functions is at least 1/4 away from the QF function at some point $\langle\alpha,\beta\rangle$, the set does not uniformly converge to the QF function.

IV. ON THE STATISTICAL PREDICTIONS OF ANGELIDIS' THEORY

In this section we argue that Angelidis' theory does not make the same statistical predictions for the EPRB experiment as the quantum formalism.

Suppose we fix an α and a β and repeat the EPRB experiment many times. Then what proportion of coincidence counts does Angelidis' theory say we should expect? There has been a suspicious change of notation in Angelidis' paper which makes this question surprisingly difficult to answer. $p^{\mathrm{QF}}{}_{12}(\alpha,\beta)$, is the chance that both detectors fire given the polarizers are set at angles α and β, respectively, according to QF. We would expect any rival theory to QF to yield a similar probability function. But Angelidis' theory actually yields a set of probability functions $p^{\mu}{}_{12}(\alpha,\beta)$. Moreover, the superscript μ no longer represents a *theory* (as it does in ''$p^{\mathrm{QF}}{}_{12}(\alpha,\beta)$''). Rather, it has come to represent the direction of polarization of the two photons.

This is odd. We expected any competitor of QF to produce a function $p^{T}{}_{12}(\alpha,\beta)$ as close to $p^{\mathrm{QF}}{}_{12}(\alpha,\beta)$ as is compatible with experimental error. But $p^{\mu}{}_{12}(\alpha,\beta)$ tells us only the chance of a coincidence *given* that the common plane of polarization of the two photons is μ. In order to work out the chance of a coincidence full stop, we need a weight function $\rho^{*}(\mu)$ which tells us how likely it is that the atom will decay into two photons plane polarized in the μ direction. The chance of a coincidence will then be equal to $\int_{\mu}\rho^{*}(\mu)p^{\mu}{}_{12}(\alpha,\beta)d\mu$. But Angelidis never tells us what this weight function is. Accordingly, it is hard to see how his theory manages to make any statistical predictions at all for the EPRB experiment he is attempting to model.

Angelidis seems to think that there is no need for him to specify this weight function. He seems to think that sentence (Σ) contains all the information we need to know. Recall that (Σ) says that, when hidden variable μ happens by pure chance to be close to α or β, then the two photons get through their respective polarizers with a probability close to $p^{\mathrm{QF}}{}_{12}(\alpha,\beta)$. But, as Angelidis admits, the conditional sentence (Σ) tells us nothing about what happens when μ is *not* close to either α or β. However, ''the single pair of back-to-back photons with $\mu\in M$ may fall inside another subset, say, $S_{\alpha 4}$ or $S_{\beta 4}$ of the set M, that is, $\mu\in S_{\alpha 4}$ OR $\mu\in S_{\beta 4}$, so that it causes a coincidence count with a different probability $\frac{1}{2}\cos^2(\alpha_4-\beta_4)$, determined by the consequent in Σ deduced from Σ (by *modus ponens*) under another value assignment.''[13] He goes on to add ''The universal quantifiers $(\forall\mu\in M)$ and $(\forall\alpha,\beta\in D)$ occurring in the prefix of the sentence Σ take into account the *whole* array of such possibilities so that the detectors accordingly register coincidence (and single counts) with the same probabilities as those given by QF for *each* and *every* pair of back to back photons emitted by the source.''

This is not so. Angelidis' explanation of how to interpret the physical significance of Σ is not complete. Standard quantum mechanics tells us that if a particular α and β are chosen so that α and β are at right angles then we will never, no matter how many times we repeat the experiment, register photons at both polarizers. Now, it is true that, on those particular occasions when the back to back photons are emitted so that their common plane of polarization μ is very close to either α or β, then the chance of a correlation will be very small. But what happens on those occasions where μ is not close to the settings of either of the polarizers? It is true, as Angelidis says, that there *exists* an α^{*} such that α^{*} is close to $\mu\alpha\nu\delta$ that, *had been the case* that the polarizer had been placed at angle α^{*} then the probabilities ascribed by T to a coincidence count are the same as that ascribed by quantum mechanics. But this does not tell us what we wanted to know! The situations where polarizer 1 is set at angle α^{*} is a *different* physical situation from the one that was under consideration. We need to know what happens when polarizers are at the particular settings α and β and the hidden variable μ is not close to either. Angelidis' advice that we choose an α^{*} close to μ simply dodges the question. In effect, Angelidis is only considering experiments where μ *is* close to one of the two polarizer settings. This information is not sufficient to tell us what proportion of coincidences we should expect if the polarizers are set at α and β and the experiment repeated many times.

Perhaps, though, Angelidis could augment his theory so that μ always *is* close to one of the two polarizer settings. Should Angelidis accept the postulate the value of the hidden variable μ is always close to the angle of one of the two polarizers, then his theory would both ascribe a probability to a coincidence count and, since $p^{\mu}{}_{12}(\alpha,\beta)$ approaches arbitrarily close to

$p^{QF}{}_{12}(\alpha,\beta)$, this probability can be made arbitrarily close to the probabilities ascribed by QF. The trouble with this proposal is that it straightforwardly violates the postulates of locality. In particular, it violates (L3), which effectively forbids that the angle of polarization of the two back to back photons be a function of the settings of the polarizers themselves.

V. CONCLUSION

The conclusion of this paper is clear: Angelidis has failed to provide us with a theory which is both local and which makes the same predictions as the standard quantum formalism. As such, Angelidis' theory simply leaves Bell's theorems untouched and the prospects for a local extension of the quantum formalism look as slim as ever.

[1] The most famous attack on the completeness of quantum mechanics comes from A. Einstein, B. Podolsky, and N. Rosen, ''Can quantum-mechanical description of physical reality be considered complete?'' Phys. Rev. **47**, 777–780 (1935).
[2] D. Bohm, ''A suggested interpretation of the quantum theory in terms of 'hidden' variables. I,'' Phys. Rev. **85**, 166–179 (1952).
[3] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, 1955).
[4] J. S. Bell, ''On the Einstein–Podolsky–Rosen paradox'' Physics (Long Island City, NY) **1**, 195–200 (1964).
[5] See A. Aspect, P. Grangier, and G. Roger, ''Experimental tests of realistic local theories via Bell's Theorem,'' Phys. Rev. Lett. **47**, 460–463 (1981); ''Experimental realisation of Einstein–Podolsky–Rosen–Bohm Gedankenexperiment: A new violation of Bell's inequalities,'' **49**, 91–94 (1982).
[6] T. D. Angelidis, ''A local extension of the quantum formalism,'' J. Math. Phys. **34**, 1635–1653 (1993).
[7] D. Bohm, *Quantum Theory*. (Prentice–Hall, Englewood Cliffs, NJ, 1951).
[8] After all, if a theory did make ρ depend upon either of these variables, then the theory is effectively saying that the setting of the polarizers influences the way in which the singlet state decays. That these three conditions are entailed by locality is widely accepted. See, for example, J. P. Jarrett, ''On the physical significance of the locality conditions in the Bell arguments,'' Nous **18**, 569–589 (1984).
[9] T. D. Angelidis, ''A local extension of the quantum formalism,'' J. Math. Phys. **34**, 1642 (1993).
[10] T. D. Angelidis, ''A local extension of the quantum formalism,'' J. Math. Phys. **34**, 1646 (1993).
[11] T. D. Angelidis, ''A local extension of the quantum formalism,'' J. Math. Phys. **34**, 1646 (1993).
[12] Angelidis makes a mess of the definition of N. He lets each N_x be the *set* of basic neighborhoods containing x, and then defines direction N as $\{N_x | x \in X\}$ thus making a direction a set of sets of subsets of X. This conflicts with his own use of the symbol N_x in the very next paragraph. Moreover, it is not true that for any two N_x and N_y there is an N_z which is contained in both. N is not even a directed set.
[13] T. D. Angelidis, ''A local extension of the quantum formalism,'' J. Math. Phys. **34**, 1646 (1993).

On one generalization for the projection matrices method in the *S*-matrix factorization problem

Dmitry I. Muravyev
Samara State Architecture and Civil Engineering Academy, Department of Physics, Molodogvardeyskaya 194, Samara 443001, Russia

(Received 20 August 1998; accepted for publication 22 January 1999)

A boundary value problem for a 2×2 *S*-matrix is solved by the method of projection matrices. The *S*-matrix is assumed to be diagnoalized by orthogonal matrix $U(k^2)=\Omega(k^2)/\sqrt{\det \Omega(k^2)}$, where matrix elements of $\Omega(k^2)$ are polynomials. The equation $\det \Omega(k^2)=0$ has roots of any order. The example is considered for the roots with order 2.
[S0022-2488(99)00605-2]

I. INTRODUCTION

The factorization of the *S*-matrix into Jost matrices can be considered as the Riemann–Hilbert boundary value problem for a half-plane. Its solution can be expressed in terms of an infinite integral in the scalar case.[1] The boundary value problem is reduced to a system of singular integral equations in the matrix case.[2] It does not seem to be the most optimal way for physical applications.

Projection matrices were used for factorization of the 2×2 rational *S*-matrix in Ref. 3. Thus, the boundary value problem was solved without using a system of the singular integral equations. It follows from Ref. 4 that projection matrices can be applied in other cases.

In Ref. 5 the boundary value problem was solved with the following assumption. The 2×2 *S*-matrix can be diagonalized by orthogonal matrix $U(k^2)$ on the real axis:

$$U(k^2)=\begin{pmatrix} \cos\varepsilon(k^2) & \sin\varepsilon(k^2) \\ -\sin\varepsilon(k^2) & \cos\varepsilon(k^2) \end{pmatrix}=\frac{1}{\sqrt{\det \Omega(k^2)}}\Omega(k^2), \tag{1}$$

$$\Omega(k^2)=\begin{pmatrix} Q_N(k^2) & P_M(k^2) \\ -P_M(k^2) & Q_N(k^2) \end{pmatrix}, \tag{2}$$

where $P_M(k^2)$ and $Q_N(k^2)$ are the polynomials in k of the degree $2M$ and $2N$, respectively, and $\operatorname{Im} P_M(k^2)$ and $\operatorname{Im} Q_N(k^2)=0$ at $\operatorname{Im} k=0$. In addition, all roots of the equation

$$\det \Omega(k^2)=Q_N^2(k^2)+P_M^2(k^2)=0 \tag{3}$$

have order 1. However, the formulas for the projection matrices are not obtained in the case when the roots have any other order.

It should be pointed out that the minimum order of roots can be set in advance. For example, if $\tan[\varepsilon(k^2)/j]$ is approximated by rational function, then the minimum order of any root of Eq. (3) is j, $j=1,2,...,j<+\infty$.

In this paper generalization of the results of Ref. 5 is suggested. The boundary value problem is solved by method of projection matrices in the case when the roots of Eq. (3) have any order.

The formulation of the problem is presented in Sec. II. The general form of the solution is discussed in Sec. III. Projection matrices are calculated in Sec. IV. The example for the roots with order 2 is considered in Sec. V. Section VI contains conclusions.

II. THE FORMULATION OF THE PROBLEM

Let us assume that the matrix $S(k)$ is known on the real k axis:

$$S(k)=U(k^2)S^{(0)}(k)U^T(k^2), \quad \operatorname{Im} k=0, \tag{4}$$

where $U(k^2)$ is defined by (1) and (2). Without generality loss one can suppose that the polynomials $P_M(k^2)$, and $Q_N(k^2)$ have no zeros simultaneously. The matrix $S^{(0)}(k)$ has the form

$$S^{(0)}(k)=\begin{pmatrix} \exp[i2\delta_1(k)] & 0 \\ 0 & \exp[i2\delta_2(k)] \end{pmatrix}, \quad \operatorname{Im} k=0,$$

where

$$\delta_j(\pm\infty)=0, \quad \delta_j(-k)=-\delta_j(k), \quad \operatorname{Im}\delta_j(k)=0, \quad j=1,2, \quad \operatorname{Im} k=0.$$

The Hölder condition should be assumed to be valid for the diagonal elements $[S^{(0)}(k)]_{jj}$ on the real axis:

$$|[S^{(0)}(k_1)]_{jj}-[S^{(0)}(k_2)]_{jj}|\leqslant A|k_1-k_2|^{\mu}, \quad A>0, \; 0<\mu\leqslant 1, \quad j=1,2.$$

Let there be two numerical sets as well:

$$\Delta_j=\{\kappa_j^{(i)}\}_{i=1}^{m_j}, \quad 0<\kappa_j^{(1)}<\kappa_j^{(2)}<\cdots<\kappa_j^{(m_j)}<+\infty, \quad m_j<+\infty, \quad j=1,2.$$

The number of the elements in the set Δ_j and behavior of $\delta_j(k)$ at the point $k=0$ are related according to the Levinson's theorem:

$$\delta_j(+0)=\begin{cases} 0, & \Delta_j=\varnothing, \\ \pi m_j, & \Delta_j\neq\varnothing, \end{cases} \quad j=1,2.$$

Here, the case when $0\in\Delta_1$ or $0\in\Delta_2$ is not considered because it is not significant to the matrix problem.

The piecewise analytical matrix $F(k)$ (i.e., when matrix elements are the piecewise analytical functions) has to be constructed in the complex k-plane. The boundary values of $F(k)$ must be continuous on the real axis and they meet the condition

$$F_+(k)=S^{-1}(k)F_-(k), \quad \operatorname{Im} k=0, \tag{5}$$

where $F_+(k)$ is analytical at $\operatorname{Im} k>0$, $F_-(k)$ at $\operatorname{Im} k<0$. In addition,

$$F_+^*(k)=F_+(-k^*)=F_-(k^*), \tag{6}$$

$$\det F_+(i\kappa_j^{(\beta)})=0, \quad \beta=1,2,...,m_j, \quad j=1,2, \tag{7}$$

$$\lim_{k\to\pm\infty} F_+(k)=E, \tag{8}$$

where E is the matrix unit.

III. THE GENERAL FORM OF THE SOLUTION.

Let us consider properties of the roots of Eq. (3). Since the polynomials $P_M(k^2)$ and $Q_N(k^2)$ depend on k^2 and $\operatorname{Im} P_M(k^2)$ and $\operatorname{Im} Q_N(k^2)=0$ at $\operatorname{Im} k=0$, the following statement takes place.

If $i\lambda_1$ is any root of Eq. (3), $-i\lambda_1$, $i\lambda_1^*$, $-i\lambda_1^*$ are the roots of this equation.

It is easy to prove that

$$\mathrm{Re}\,(i\lambda_1)\neq 0 \tag{9}$$

and

$$\mathrm{Im}\,(i\lambda_1)\neq 0. \tag{10}$$

If (9) or (10) is false, then $P_M^2(-\lambda_1^2)$, $Q_N^2(-\lambda_1^2)\geqslant 0$. The polynomials $P_M(k^2)$ and $Q_N(k^2)$ have no common zeros, thus

$$Q_N^2(-\lambda_1^2)+P_M^2(-\lambda_1^2)>0,$$

i.e., $i\lambda_1$ is not a root of Eq. (3). Consequently, conditions (9) and (10) are true for any root of Eq. (3).

Let γ_1 be an order of the roots $i\lambda_1$, $-i\lambda_1$, $i\lambda_1^*$, $-i\lambda_1^*$. Based on Eq. (3),

$$\sum_{j=1}^{\alpha}\gamma_j=L, \quad L=\max\{M,N\}, \tag{11}$$

and α is of particular importance for the solution of the boundary value problem.

For convenience Eq. (3) is rewritten as two equations:

$$Q_N(k^2)+iP_M(k^2)=0, \tag{12}$$

$$Q_N(k^2)-iP_M(k^2)=0.$$

For definiteness, let $i\lambda_j$ be the root of (12) and $\mathrm{Im}(i\lambda_j)>0$, $j=1,2,...,\alpha$

Let us find the solution of the boundary value problem in the form

$$F_+(k)=\Omega(k^2)G_+(k)W_1(k^2)\times\cdots\times W_\alpha(k^2)C, \tag{13}$$

$$C=[\lim_{k\to\pm\infty}k^{-2L}\Omega(k^2)]^{-1}, \tag{14}$$

$$W_j(k^2)=W_j^{(1)}(k^2)\times\cdots\times W_j^{(\gamma_j)}(k^2), \quad j=1,2,...\alpha, \tag{15}$$

$$W_j^{(i)}(k^2)=P_j^{(i)}\frac{1}{k^2+\lambda_j^2}+(E-P_j^{(i)})\frac{1}{k^2+\lambda_j^{*2}}, \quad i=1,2,...,\gamma_j, \quad j=1,2,...,\alpha, \tag{16}$$

$$(P_j^{(i)})^2=P_j^{(i)}, \quad P_j^{(i)}\neq E,0, \ \ i=1,2,...,\gamma_j, \quad j=1,2,...,\alpha, \tag{17}$$

$$(P_j^{(i)})^*=E-P_j^{(i)}, \quad i=1,2,...,\gamma_j, \quad j=1,2,...,\alpha. \tag{18}$$

$G_+(k)$ is the diagonal matrix. It satisfies to conditions

$$G_+(k)=(S^{(0)}(k))^{-1}G_-(k), \quad \mathrm{Im}\,k=0, \tag{19}$$

$$G_+^*(k)=G_+(-k^*)=G_-(k^*), \tag{20}$$

$$\det G_+(i\kappa_j^{(\beta)})=0, \quad \beta=1,2,...,m_j, \quad j=1,2, \tag{21}$$

$$\lim_{k\to\pm\infty}G_+(k)=E. \tag{22}$$

The matrix elements $[G_+(k)]_{jj}$ can be calculated with the formulas of the scalar boundary value problem:

$$[G_+(k)]_{jj}=\Pi_\pm(j,k)\exp\left(\frac{1}{i2\pi}\int_{-\infty}^{+\infty}\ln[\exp(-i2\delta_j(k'))\Pi_-^2(j,k')]\frac{dk'}{k'-k\mp i0}\right),\quad j=1,2, \tag{23}$$

$$\Pi_\pm(j,k)=\begin{cases}1, & \Delta_j=\varnothing,\\ \prod_{\beta=1}^{m_j}(k\mp i\kappa_j^{(\beta)})/(k\pm i\kappa_j^{(\beta)}), & \Delta_j\neq\varnothing,\end{cases}\quad j=1,2. \tag{24}$$

Let us show that $F_+(k)$ in the form (13) satisfies to conditions (5)–(8). Really, taking into account (1), (4), and (19) it is easy to check (5). To convince oneself that (6) is correct, we must use (2), (14)–(16), (18), and (20). Some steps are required for the proof of (7). From (17),

$$\det P_j^{(i)}=0,\quad \mathrm{Sp}\,P_j^{(i)}=1,\quad i=1,2,\ldots,\gamma_j,\quad j=1,2,\ldots,\alpha. \tag{25}$$

Furthermore, from (16) and (25)

$$\det W_j^{(i)}(k^2)=(k^2+\lambda_j^2)^{-1}(k^2+\lambda_j^{*2})^{-1},\quad i=1,2,\ldots,\gamma_j,\quad j=1,2,\ldots,\alpha. \tag{26}$$

Property (7) is true since (2), (15), (16), (21), and (26) are valid. Condition (8) is evident because of (2), (11), (14)–(16), and (22).

The explicit form of the projection matrices $P_j^{(\beta)}$ must be chosen in such a way that $F_+(k)$ will have no poles in the points $k=i\lambda_j$, $i\lambda_j^*$, $\beta=1,2,\ldots,\gamma_j$, $j=1,2,\ldots,\alpha$.

IV. THE CONSTRUCTION OF THE PROJECTION MATRICES

Let us prove that

$$\Omega(k^2)\left[P_0\prod_{j=1}^{\alpha}\frac{1}{(k^2+\lambda_j^2)^{\gamma_j}}+(E-P_0)\prod_{j=1}^{\alpha}\frac{1}{(k^2+\lambda_j^{*2})^{\gamma_j}}\right]C=E, \tag{27}$$

where

$$P_0=\frac{1}{2}\begin{pmatrix}1 & -i\\ i & 1\end{pmatrix},\quad P_0^2=P_0,\quad P_0^*=E-P_0. \tag{28}$$

Since $i\lambda_j$ and $-i\lambda_j$, $j=1,2,\ldots,\alpha$, are the roots of Eq. (12) and taking into account (2), (14), and (28)

$$\Omega(k^2)P_0\prod_{j=1}^{\alpha}\frac{1}{(k^2+\lambda_j^2)^{\gamma_j}}C=[Q_N(k^2)+iP_M(k^2)]P_0\prod_{j=1}^{\alpha}\frac{1}{(k^2+\lambda_j^2)^{\gamma_j}}C=P_0. \tag{29}$$

The complex conjugation of (29) at $\mathrm{Im}\,k=0$ produces

$$\Omega(k^2)(E-P_0)\prod_{j=1}^{\alpha}\frac{1}{(k^2+\lambda_j^{*2})^{\gamma_j}}C=[Q_N(k^2)-iP_M(k^2)](E-P_0)\prod_{j=1}^{\alpha}\frac{1}{(k^2+\lambda_j^{*2})^{\gamma_j}}C=E-P_0. \tag{30}$$

Thus, (27) results from (29) and (30). Formula (27) gives an important presentation of the matrix $\Omega(k^2)$:

$$\Omega(k^2)=C^{-1}\left[P_0\prod_{j=1}^{\alpha}(k^2+\lambda_j^2)^{\gamma_j}+(E-P_0)\prod_{j=1}^{\alpha}(k^2+\lambda_j^{*2})^{\gamma_j}\right]. \tag{31}$$

Let us fix four roots $i\lambda_1$, $-i\lambda_1$, $i\lambda_1^*$, and $-i\lambda_1^*$. Their order is γ_1. Let us consider the matrix

$$X_{1+}^{(1)}(k)=\Omega(k^2)A_1^{(1)}(k), \tag{32}$$

$$A_1^{(1)}(k)=G_+(k)W_1^{(1)}(k^2)=G_+(k)\left[P_1^{(1)}\frac{1}{k^2+\lambda_1^2}+(E-P_1^{(1)})\frac{1}{k^2+\lambda_1^{*2}}\right], \tag{33}$$

where the projection matrix $P_1^{(1)}$ is computed from the formulas

$$P_1^{(1)}=\frac{1}{\mathrm{Sp}\,Y_1^{(1)}}Y_1^{(1)}, \tag{34}$$

$$Y_1^{(1)}=[G_+(i\lambda_1)]^{-1}P_0[G_+(i\lambda_1)]^*. \tag{35}$$

Formulas (34) and (35) are in complete accordance with condition (18).

Let us show that

$$(E-P_0)G_+(i\lambda_1)P_1^{(1)}=0. \tag{36}$$

Insertion of (34) and (35) into (36) proves the validity of this equation. The complex conjugation of (36) in terms of (18), (20), and (28) results in

$$P_0G_+(i\lambda_1^*)(E-P_1^{(1)})=0. \tag{37}$$

It will be obvious from (31)–(33), (36), and (37) that matrix $X_{1+}^{(1)}(k)$ has no poles at the points $k=i\lambda_1$, $i\lambda_1^*$. Thus $X_{1+}^{(1)}(k)$ is analytical at $\mathrm{Im}\,k>0$ except the cases when $k=\infty$ (in a general case).

Let us rewrite the matrix $X_{1+}^{(1)}(k)$ in the form

$$X_{1+}^{(1)}(k)=C^{-1}\left[P_0(k^2+\lambda_1^2)^{\gamma_1-1}\prod_{j=2}^{\alpha}(k^2+\lambda_1^2)^{\gamma_1}+(E-P_0)(k^2+\lambda_1^{*2})^{\gamma_1-1}\prod_{j=2}^{\alpha}(k^2+\lambda_1^{*2})^{\gamma_1}\right]Z_{1+}^{(1)}\times(k), \tag{38}$$

$$Z_{1+}^{(1)}(k)=[P_0(k^2+\lambda_1^2)+(E-P_0)(k^2+\lambda_1^{*2})]A_1^{(1)}(k), \tag{39}$$

where (31) should be taken into account. According to (33), (36), (37), and (39), the matrix $Z_{1+}^{(1)}(k)$ has no poles at the points $k=i\lambda_1$, $i\lambda_1^*$. In addition, it follows from (21), (26), (33), and (39) that

$$\det Z_{1+}^{(1)}(i\lambda_1)=\det G_+(i\lambda_1)\neq 0. \tag{40}$$

For all projection matrices $P_j^{(i)}$, $i=1,2,...,\gamma_j$, $j=1,2,...,\alpha$, the formulas can be written now based on (31)–(40).

Let us assume that the matrix $X_{n+}^{(j)}(k)$ is computed:

$$X_{n+}^{(j)}(k)=\Omega(k^2)A_n^{(j)}(k),\quad j=0,1,2,...,\gamma_n-1,$$

where

$$A_n^{(j)}(k)=G_+(k)W_1(k^2)\times\cdots\times W_{n-1}(k^2)\times W_n^{(1)}(k^2)\times\cdots\times W_n^{(j)}(k^2),\quad j=0,1,2,...,\gamma_n-1, \tag{41}$$

$$A_n^{(\gamma_n)}(k)\equiv A_{n+1}^{(0)}(k),\quad n<\alpha,\quad A_1^{(0)}(k)\equiv G_+(k),\quad W_0(k^2)\equiv E,\quad W_n^{(0)}(k^2)\equiv E. \tag{42}$$

The projection matrix $P_n^{(j+1)}$ can be computed from the formulas

$$P_n^{(j+1)}=\frac{1}{\mathrm{Sp}\,Y_n^{(j+1)}}Y_n^{(j+1)},\quad j=0,1,2,...,\gamma_n-1, \tag{43}$$

$$Y_n^{(j+1)}=[Z_{n+}^{(j)}(i\lambda_n)]^{-1}P_0[Z_{n+}^{(j)}(i\lambda_n)]^*,\quad j=0,1,2,...,\gamma_n-1, \tag{44}$$

$$Z_{n+}^{(j)}(k)=[P_0(k^2+\lambda_n^2)^j+(E-P_0)(k^2+\lambda_n^{*2})^j]A_n^{(j)}(k),\quad j=0,1,2,...,\gamma_n-1. \tag{45}$$

It should be pointed out that

$$\det Z_{n+}^{(j)}(i\lambda_j)\neq 0,\quad j=0,1,2,...,\gamma_n-1.$$

So far as $F_+(k)$ can be written in the form

$$F_+(k)=X_{a+}^{(\gamma_a)}(k)C,$$

the formulas (13)–(16), (23), (24), and (41)–(45) give the solution of the boundary value problem given in Sec. II.

V. THE EXAMPLE

Let us assume that

$$S^{(0)}(k)=\begin{pmatrix}\dfrac{(k+i\kappa)(k+i\varphi)}{(k-i\kappa)(k-i\varphi)} & 0\\ 0 & 1\end{pmatrix},\quad \varphi,\kappa>0,\quad \tan\frac{\varepsilon(k^2)}{2}=\frac{k^2}{ak^2+b},\quad \mathrm{Im}\,a,\mathrm{Im}\,b=0.$$

The function $\tan\varepsilon(k^2)$ has the form

$$\tan\varepsilon(k^2)=P_2(k^2)/Q_2(k^2),\quad P_2(k^2)=2k^2(ak^2+b),\quad Q_2(k^2)=(ak^2+b)^2-k^4.$$

Equation (12) is rewritten as

$$[k^2(a+i)+b]^2=0.$$

Let $i\lambda_1$ and $-i\lambda_1$ be the roots of this equation and $\mathrm{Im}(i\lambda_1)>0$. The order of $i\lambda_1$ is $\gamma_1=2$. Since $\gamma_1=L=\max\{2,2\}=2$, two projection matrices have to be constructed [see (11), (13), (15), and (16)].

The solution of the boundary value problem has the form

$$F_+(k)=\Omega(k^2)G_+(k)\left[P_1^{(1)}\frac{1}{k^2+\lambda_1^2}+(E-P_1^{(1)})\frac{1}{k^2+\lambda_1^{*2}}\right]\left[P_1^{(2)}\frac{1}{k^2+\lambda_1^2}+(E-P_1^{(2)})\frac{1}{k^2+\lambda_1^{*2}}\right]C,$$

$$G_+(k)=\begin{pmatrix}\dfrac{k-i\kappa}{k+i\varphi} & 0\\ 0 & 1\end{pmatrix},\quad C=\frac{1}{1+4a^2}\begin{pmatrix}-1 & -2a\\ 2a & -1\end{pmatrix}.$$

The matrix $P_1^{(1)}$ is computed from (34) and (35):

$$P_1^{(1)}=\frac{1}{(\lambda_1+\varphi)(\lambda_1^*-\kappa)+(\lambda_1^*+\varphi)(\lambda_1-\kappa)}\begin{pmatrix}(\lambda_1+\varphi)(\lambda_1^*-\kappa) & -i(\lambda_1+\varphi)(\lambda_1^*-\varphi)\\ i(\lambda_1-\kappa)(\lambda_1^*-\kappa) & (\lambda_1^*+\varphi)(\lambda_1-\kappa)\end{pmatrix}.$$

According to (41), (42), and (45):

$$Z_{1+}^{(1)}(k)=[P_0(k^2+\lambda_1^2)+(E-P_0)(k^2+\lambda_1^{*2})]G_+(k)\left[P_1^{(1)}\frac{1}{k^2+\lambda_1^2}+(E-P_1^{(1)})\frac{1}{k^2+\lambda_1^{*2}}\right].$$

L'Hospital's rule is convenient to be used to compute $Z_{1+}^{(1)}(i\lambda_1)$:

$$Z_{1+}^{(1)}(i\lambda_1)=P_0B+H,$$

$$B=G_+(i\lambda_1)P_1^{(1)},\quad H=\frac{\lambda_1^{*2}-\lambda_1^2}{i2\lambda_1}(E-P_0)\frac{d}{dk}G_+(k)\bigg|_{k=i\lambda_1}P_1^{(1)}+(E-P_0)G_+(i\lambda_1)(E-P_1^{(1)}).$$

The matrix H has the structure

$$H=\frac{1}{4\lambda_1(\lambda_1+\varphi)^2}\begin{pmatrix} h_{11} & ih_{22}\\ -ih_{11} & h_{22}\end{pmatrix},$$

and its elements can be computed using formulas

$$\begin{aligned}h_{11}&=-(\lambda_1^{*2}-\lambda_1^2)(\varphi+\kappa)[P_1^{(1)}]_{11}+2\lambda_1(\lambda_1+\varphi)([P_1^{(1)}]_{11}^*(\lambda_1-\kappa)-i[P_1^{(1)}]_{21}(\lambda_1+\varphi))\\ &=-(\lambda_1^{*2}-\lambda_1^2)(\varphi+\kappa)[P_1^{(1)}]_{11}+2\lambda_1(\lambda_1+\varphi)(\lambda_1-\kappa),\end{aligned}$$

$$\begin{aligned}h_{22}&=i(\lambda_1^{*2}-\lambda_1^2)(\varphi+\kappa)[P_1^{(1)}]_{12}+2\lambda_1(\lambda_1+\varphi)([P_1^{(1)}]_{11}(\lambda_1+\varphi)+i[P_1^{(1)}]_{12}(\lambda_1-\kappa))\\ &=i(\lambda_1^{*2}-\lambda_1^2)(\varphi+\kappa)[P_1^{(1)}]_{12}+2\lambda_1(\lambda_1+\varphi)^2.\end{aligned}$$

The matrix (P_0B) does not impact on the final result. However, it is required to compute $[Z_{1+}^{(1)}\times(i\lambda_1)]^{-1}$. The result is [see (43) and (44)]

$$P_1^{(2)}=\frac{1}{h_{11}h_{22}^*+h_{11}^*h_{22}}\begin{pmatrix} h_{22}h_{11}^* & -ih_{22}h_{22}^*\\ ih_{11}h_{11}^* & h_{22}^*h_{11}\end{pmatrix}.$$

VI. CONCLUSION

The discussion above showed that the projection matrices method appears to perform well in the case when the roots of Eq. (3) have any order. This is especially important because the minimum order of the roots can be set in advance. In addition, the class of the unitary matrices for which the boundary value problem can be solved without using the system of the singular integral equations is extended.

[1] F. D. Gakhov, *Boundary Value Problems* (Pergamon, Oxford, 1966).
[2] N. P. Vekua, *System of Singular Integral Equations* (Noordhoff, Groningen, 1967).
[3] T. Fulton and R. G. Newton, Nuovo Cimento **3**, 677 (1956).
[4] G. N. Chebotarev, Uch. Zapiski Kazan. Univ., Mat. **116**, 31 (1956) (in Russian).
[5] A. F. Krutov, D. I. Muravyev, and V. E. Troitsky, J. Math. Phys. **38**, 6 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Classical mechanics and geometric quantization on an infinite dimensional disc and Grassmannian

O. T. Turgut
Institut Mittag-Leffler, Auravägen 17, S-182 62, Djursholm, Sweden
and Department of Physics, Bogazici University, 80815 Bebek, Istanbul, Turkey[a]

(Received 7 December 1998; accepted for publication 2 February 1999)

We discuss the classical mechanics on the Grassmannian and the disc modeled on the ideal $\mathcal{L}^{(2,\infty)}$. We apply methods of geometric quantization to these systems. Their relation to a flat symplectic space is also discussed. © *1999 American Institute of Physics.* [S0022-2488(99)01808-3]

I. INTRODUCTION

We will analyze geometric quantization of a classical system which has as its phase space the infinite-dimensional Grassmannian or the disc modeled on the ideal $\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-)$. There are two motivations for our work. The classical dynamics studied should correspond to the large-N_c limit of a quantum system which requires a logarithmic renormalization. Its quantization should give us an understanding of this system in the Schrödinger picture. This picture has some advantages over the scattering matrix, as is well known in the physics literature. The second is to study and understand infinite-dimensional systems; their quantization should lead to some interesting mathematical questions. A good example is typical two-dimensional field theory models, which do not require a renormalization but only a normal ordering.[1,2] It will be interesting to develop the necessary tools for a more complicated systems, and perhaps give a more precise meaning to renormalized field theories. We should add that, in this article, we do not study any particular Hamiltonian or associated delicate domain problems. In some sense we have only made an attempt to study part of the kinematics. The full understanding will require studying a specific model.

II. THE DISC AND THE GRASSMANNIAN

Our approach is inspired from the discussion of the Grassmannian in the book by Pressley and Segal[3] and closely follows our previous work.[2] We will extend some of our previous ideas to this case.

Let $\mathcal{H}$ be a separable infinite-dimensional complex Hilbert space; $\mathcal{H}_-$ and $\mathcal{H}_+$ are two orthogonal isomorphic subspaces with $\mathcal{H}=\mathcal{H}_-\oplus\mathcal{H}_+$. Physically, one can think of these two spaces as the decomposition of the one-particle Hilbert space into positive and negative energy states.

Define the disc $D_{1+}(\mathcal{H}_-,\mathcal{H}_+)$ to be the set of all operators $Z:\mathcal{H}_+\rightarrow\mathcal{H}_-$ such that $1-Z^\dagger Z>0$ and Z is in $\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-)$. We refer to the Appendix for the ideals $\mathcal{L}^{(2,\infty)}$ and $\mathcal{L}^{(1,\infty)}$.

Since the space $\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-)$ is contractible, the set of Z for the disc can be taken as a coordinate system. It is an infinite-dimensional complex manifold, modeled on a Banach space.

In a similar spirit, we define the Grassmannian to be the set of closed subspaces W of $\mathcal{H}$, such that the projection $\mathrm{Pr}_+:W\rightarrow\mathcal{H}_+$ is Fredholm and the projection $\mathrm{Pr}_-:W\rightarrow\mathcal{H}_-$ is in $\mathcal{L}^{(2,\infty)}(W,\mathcal{H}_-)$. We first define natural group actions on these spaces. These group actions are used to prove that these are manifolds. (Generalizations of this kind have been pointed out in Ref. 3, and explored in Ref. 4 for the ideals $\mathcal{L}^p$. See also Ref. 5.)

We introduce the following pseudo-unitary group which is a subset of the invertible operators from $\mathcal{H}$ to $\mathcal{H}$:

[a]After March 1999.

$$\mathrm{U}_{1+}(\mathcal{H}_-,\mathcal{H}_+)=\{g|g\epsilon g^\dagger=\epsilon,\quad g^{-1}\ \text{exists and}\ [\epsilon,g]\in\mathcal{L}^{(2,\infty)}\}. \tag{1}$$

Here $\epsilon=\begin{pmatrix}-1&0\\0&1\end{pmatrix}$ with respect to the decomposition $\mathcal{H}=\mathcal{H}_-\oplus\mathcal{H}_+$. If we decompose the matrix into block forms,

$$g=\begin{pmatrix}a&b\\c&d\end{pmatrix}, \tag{2}$$

we have $a:\mathcal{H}_-\to\mathcal{H}_-$, $b:\mathcal{H}_+\to\mathcal{H}_-$, $c:\mathcal{H}_-\to\mathcal{H}_+$, and $d:\mathcal{H}_+\to\mathcal{H}_+$. Then, the off-diagonal elements b and c are in $\mathcal{L}^{(2,\infty)}$ and the diagonal elements a and d are bounded operators. In fact, they are invertible operators, since their spectrum does not contain zero. The conditions on the off-diagonal elements imply some control over how much $\mathcal{H}_+$ and $\mathcal{H}_-$ mix with each others.

We define an action of $\mathrm{U}_{1+}(\mathcal{H}_-,\mathcal{H}_+)$ on the disc D_{1+}:

$$Z\mapsto g\circ Z=(aZ+b)(cZ+d)^{-1}. \tag{3}$$

The condition $1-Z^\dagger Z>0$ implies that $cZ+d$ is invertible and bounded. Since the space of $\mathcal{L}^{(2,\infty)}$ is a two-sided ideal, $(aZ+b)(cZ+d)^{-1}$ is still in $\mathcal{L}^{(2,\infty)}$. Thus our action is well defined.

The stability subgroup of the point $Z=0$ is $\mathrm{U}(\mathcal{H}_-)\times\mathrm{U}(\mathcal{H}_+)$,$\mathrm{U}(\mathcal{H}_\pm)$ being the group of *all* unitary operators on $\mathcal{H}_\pm$. Moreover, any point Z is the image of 0 under the action of the group, $g\circ(Z=0)=bd^{-1}$. [Note that bd^{-1} is in $\mathcal{L}^{(2,\infty)}$ and $d^\dagger d=1+b^\dagger b$ implies that $1-(bd^{-1})^\dagger bd^{-1}>0$.] We therefore see that D_{1+} is a homogeneous space and given by the quotient

$$\mathrm{D}_{1+}=\mathrm{U}_{1+}(\mathcal{H}_-,\mathcal{H}_+)/\mathrm{U}(\mathcal{H}_-)\times\mathrm{U}(\mathcal{H}_+). \tag{4}$$

It is possible to view Gr_{1+} as a coset space of complex Lie groups. Incidentally this will define a complex structure on Gr_{1+} which will be useful for geometric quantization. Define a subset of the general linear group

$$\mathrm{GL}_{1+}=\{\gamma|\gamma\ \text{is invertible};\quad[\epsilon,\gamma]\in\mathcal{L}^{(2,\infty)}\}. \tag{5}$$

When we decompose g into 2×2 submatrices $\gamma_{12},\gamma_{21}\in\mathcal{L}^{(2,\infty)}$ while γ_{11} and γ_{22} are Fredholm. This is a Banach–Lie group modeled on $\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-)$. We take the space of all endomorphisms on $\mathcal{H}$ with the same condition on the off-diagonals, $\mathrm{End}_{\mathcal{L}^{(2,\infty)}}(\mathcal{H})$, and give it the natural topology under the norm $\|A\|_+=\|[\epsilon,A]_+\|+\|[\epsilon,A]\|_{\mathcal{L}^{(2,\infty)}}$. The invertible elements of $\mathrm{End}_{\mathcal{L}^{(2,\infty)}}(\mathcal{H})$ are a group, open under this topology, and it has a tangent space which comes from the natural imbedding. It is straightforward to define the "Borel subgroup"

$$\mathrm{B}_{1+}=\left\{\beta=\begin{pmatrix}\beta_{11}&\beta_{12}\\0&\beta_{22}\end{pmatrix}\middle|\beta\in\mathrm{GL}_{1+}\right\}. \tag{6}$$

This is the stability group of $\mathcal{H}_-$ under the action of GL_{1+} on $\mathcal{H}$. Thus the Grassmannian (which is the orbit of $\mathcal{H}_-$) is the complex coset space,

$$\mathrm{Gr}_{1+}=\mathrm{GL}_{1+}/\mathrm{B}_{1+}. \tag{7}$$

It will be convenient to use the following operators for the points on D_{1+},$\Phi:\mathcal{H}\to\mathcal{H}$,

$$\Phi=1-2\begin{pmatrix}(1-ZZ^\dagger)^{-1}&-(1-ZZ^\dagger)^{-1}Z\\Z^\dagger(1-ZZ^\dagger)^{-1}&-Z^\dagger(1-ZZ^\dagger)^{-1}Z\end{pmatrix}. \tag{8}$$

One can see that under the transformation $Z\mapsto g\circ Z,\Phi\mapsto g^{-1}\Phi g$. Here Φ satisfies $\epsilon\Phi^\dagger\epsilon=\Phi$ and $\Phi^2=1$. Also, $\Phi-\epsilon\in\mathcal{L}^{(2,\infty)}$, so that as an operator Φ does not differ from ϵ in an arbitrary way.

We can equivalently define the Grassmannian to be the following set of operators on $\mathcal{H}$:

$$\mathrm{Gr}_{1+}=\{\Phi|\Phi=\Phi^{\dagger};\Phi^{2}=1;\Phi-\epsilon\in\mathcal{L}^{(2,\infty)}\}. \tag{9}$$

Since $\Phi^2=1$ and it is self-adjoint, it can be diagonalized by the action of

$$\mathrm{U}_{1+}(\mathcal{H})=\{g|g^{\dagger}g=1;[\epsilon,g]\in\mathcal{L}^{(2,\infty)}\}. \tag{10}$$

Let us split g into 2×2 blocks:

$$g=\begin{pmatrix} g_{11} & g_{12}\\ g_{21} & g_{22}\end{pmatrix}. \tag{11}$$

The convergence condition on $[\epsilon,g]$ is the statement that the off-diagonal blocks g_{12} and g_{21} are in $\mathcal{L}^{(2,\infty)}$. It then follows that g_{11} and g_{22} are Fredholm operators. The Fredholm index of g_{11} is opposite to that of g_{22}; this integer is a homotopy invariant of g and we can decompose $\mathrm{U}_{1+}(\mathcal{H})$ into connected components labeled by this integer.

We can see that $\mathrm{U}_{1+}(\mathcal{H})$ is a real form of this group. $\mathrm{GL}_{1+}(\mathcal{H})$ is the topological product of $\mathrm{U}_{1+}(\mathcal{H})$ and the contractible space of positive definite elements by using the fact that $\mathrm{End}_{\mathcal{L}^{(2,\infty)}}(\mathcal{H})$ has the square root of positive elements well defined and continuous under its topology. (We can show that U_{1+} is a deformation retract of GL_{1+}, similar to the finite-dimensional case.)

With the projection $g\rightarrow g\epsilon g^{\dagger}$, we see that Gr_{1+} is a homogeneous space of $\mathrm{U}_{1+}(\mathcal{H})$:

$$\mathrm{Gr}_{1+}=\mathrm{U}_{1+}(\mathcal{H})/\mathrm{U}(\mathcal{H}_{-})\times\mathrm{U}(\mathcal{H}_{+}). \tag{12}$$

Any $\Phi\in\mathrm{Gr}_{1+}$ can be diagonalized by an element of $\mathrm{U}_{1+}(\mathcal{H})$, $\Phi=g\epsilon g^{\dagger}$; this g is ambiguous up to right multiplication by an element that commutes with ϵ. Such elements form the subgroup

$$\mathrm{U}(\mathcal{H}_{-})\times\mathrm{U}(\mathcal{H}_{+})=\left\{h\middle|h=\begin{pmatrix} h_{11} & 0\\ 0 & h_{22}\end{pmatrix};\quad h_{11}^{\dagger}h_{11}=1=h_{22}^{\dagger}h_{22}\right\}. \tag{13}$$

Each point $\Phi\in\mathrm{Gr}_{1+}$ corresponds to a subspace of $\mathcal{H}$: the eigenspace of Φ with eigenvalue -1. Thus Gr_{1+} consists of all subspaces obtained from $\mathcal{H}_{-}$ by an action of U_{1+}.

To define the tangent space at each point we can use the action of the group on itself. For our purposes it is better to take the group action on the left. Since the stability subgroup of ϵ is $\mathrm{U}(\mathcal{H}_{-})\times\mathrm{U}(\mathcal{H}_{+})$, in both cases the tangent space is isomorphic to the corresponding off-diagonal algebras. In each case this is equivalent to $\mathcal{L}^{(2,\infty)}(\mathcal{H}_{+},\mathcal{H}_{-})$ as a vector space, due to the Hermiticity (or pseudo-Hermiticity) condition.

Any given $u\in\mathrm{U}_{1+}$ defines a vector at a given point, and a vector field can be expanded in terms of the local set of vectors. The action of a vector field on Φ is given by $V_{u(\Phi)}(\Phi)=[u(\Phi),\Phi]=g[g^{-1}u(\Phi)g,\epsilon]g^{-1}$. The tangent space has a set of vectors which are given by the completion of the finite rank operators inside $\mathcal{L}^{(2,\infty)}$. This is also an ideal inside $\mathcal{B}=B(\mathcal{H})$ and is a separable Banach space under the same norm as $\mathcal{L}^{(2,\infty)}$; we will denote this set by $(\mathcal{L}^{(2,\infty)})^{(0)}$. The tangent space has a noncanonical decomposition at each point which is isomorphic to $(\mathcal{L}^{(2,\infty)})^{(0)}+\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$; the second part is a "transversal piece." As we will see, this quotient will be important for the dynamical system we have in mind.

We introduce the cotangent space as a formal expression $d\Phi$ via its contraction with the vector field at a given point; $d\Phi(V_{u(\Phi)})=V_{u(\Phi)}(\Phi)$. (The dual space requires more care in infinite dimensions. We can think of the norm dual of $\mathcal{L}^{(2,\infty)}$, yet this space does not have a simple characterization. If we assume that the tangent space is, in fact, the norm dual of the cotangent space, we have a simple description of the cotangent space. We refer to Gohberg and Krein for the details.[6] In this article we will leave the question of the dual open, and use one-forms only when we have an explicit formula.)

We would like to think of the D_{1+} and Gr_{1+} as classical phase spaces. To do this we need to introduce a Poisson bracket. We will search for a symplectic form on this space. It is tempting to generalize the finite-dimensional formula to this case. If we write down $\Omega=(i/4)\mathrm{Tr}\Phi d\Phi\wedge d\Phi$ we see that the trace, in general, does not exist. However, one can see that the divergence is logarithmic, in fact the formal expression $\Phi d\Phi\wedge d\Phi$ belongs to $\mathcal{L}^{(1,\infty)}$. Hence we can replace the ordinary trace by the Dixmier trace. The Dixmier trace is used in noncommutative geometry; for a masterful presentation of its properties we refer to the book and lecture notes by Connes.[7]

Each choice of the trace will give another symplectic form. They all agree on the ''measurable'' part of the ideal $\mathcal{L}^{(1,\infty)}$. Since the ''measurable'' elements do not form an ideal, we cannot assume that the symplectic form is independent of the choice of the limit point.

Another important point is to remember that the Dixmier trace vanishes on the ideal generated by the completion of finite rank operators, $(\mathcal{L}^{(1,\infty)})^{(0)}$. In the applications one expects that the operators we have to consider are pseudo-differential operators on manifolds. The physically relevant group of transformations is modeled on pseudo-differential operators which belong to the specific ideals that we have defined. In the case of classical pseudo-differential operators, the Dixmier trace is uniquely defined; it is equal to the Wodzicki residue of the pseudo-differential operator, as shown by Connes.[8] An interesting application of a new trace to the class of elliptic pseudodifferential operators is given in Ref. 9, and to chiral anomaly in Ref. 10. An interesting discussion of the central extensions and Schwinger terms are given in Ref. 11.

In this article we will consider the general case, and show the dependence of the symplectic form to this limiting process ω explicitly on our definition of the symplectic form:

$$\Omega_\omega=\frac{i}{4}\mathrm{Tr}_\omega\,\Phi d\Phi\wedge d\Phi. \tag{14}$$

The existence of such a trace is the reason for our choice $\mathcal{L}^{(2,\infty)}$. One can check that the above form is closed; it is not so obvious that it is nondegenerate. In fact, it vanishes whenever the result of the contractions with the vectors at a given point is in the completion of the finite rank operators inside $\mathcal{L}^{(2,\infty)},(\mathcal{L}^{(2,\infty)})^{(0)}\neq\mathcal{L}^{(2,\infty)}$. This completion is a separable Banach space, and an ideal inside $\mathcal{B}$ as well, whereas $\mathcal{L}^{(2,\infty)}$ is a nonseparable Banach space.

The above form is invariant under the action of $\mathrm{U}_{1+}(\mathcal{H})$ for the Gr_{1+} and invariant under the action of $\mathrm{U}_{1+}(\mathcal{H}_-,\mathcal{H}_+)$ for the D_{1+}. The formal expression is defined as

$$i_{V_u}i_{V_v}\Omega_\omega=\frac{i}{8}\mathrm{Tr}_\omega\,\Phi[[u,\Phi],[v,\Phi]]. \tag{15}$$

One can show the invariance using this expression immediately (see below). Thus, Gr_{1+} and D_{1+} are both homogeneous manifolds with an invariant closed two-form similar to the finite-dimensional case.

Unfortunately, this form is degenerate; it vanishes on the part of the tangent space which corresponds to $(\mathcal{L}^{(2,\infty)})^{(0)}$. (This means that we throw away a large part of the symplectic manifold. Perhaps a physically more appropriate choice is to use a combination, which keeps the information about the ''small'' directions. Although this can be done we will focus on the above symplectic form.) To see this let us calculate the contraction of Ω_ω at a point Φ, with a vector field which belongs to the part $(\mathcal{L}^{(2,\infty)})^{(0)}$. Let us assume that this vector is generated by u acting from the left. The condition for the vector of be in $(\mathcal{L}^{(2,\infty)})^{(0)}$ is simply $[\epsilon,g^{-1}ug]\in(\mathcal{L}^{(2,\infty)})^{(0)}$ at the point $\Phi=g\epsilon g^{-1}$:

$$i_{V_u}\Omega_\omega=\frac{i}{8}\mathrm{Tr}_\omega\,\Phi[[u,\Phi],d\Phi]. \tag{16}$$

We will show that the contraction of this one-form with an arbitrary vector on the tangent space at the same point Φ is zero, hence the form is zero. Any such vector on the tangent is again generated by the left action with a Lie algebra element v,

$$i_{V_v} i_{V_u} \Omega_\omega = \frac{i}{8} \mathrm{Tr}_\omega\, \epsilon [[\epsilon, g^{-1} u g], [\epsilon, g^{-1} v g]]. \tag{17}$$

Using

$$\left[\epsilon, \begin{pmatrix} \mathcal{B} & \mathcal{L}^{(2,\infty)} \\ \mathcal{L}^{(2,\infty)} & \mathcal{B} \end{pmatrix} \right] = \begin{pmatrix} 0 & \mathcal{L}^{(2,\infty)} \\ \mathcal{L}^{(2,\infty)} & 0 \end{pmatrix}$$

and similarly for the other part, we have

$$\begin{pmatrix} 0 & \mathcal{L}^{(2,\infty)} \\ \mathcal{L}^{(2,\infty)} & 0 \end{pmatrix} \begin{pmatrix} 0 & (\mathcal{L}^{(2,\infty)})^{(0)} \\ (\mathcal{L}^{(2,\infty)})^{(0)} & 0 \end{pmatrix} = \begin{pmatrix} (\mathcal{L}^{(1,\infty)})^{(0)} & 0 \\ 0 & (\mathcal{L}^{(1,\infty)})^{(0)} \end{pmatrix}, \tag{18}$$

where we use $(\mathcal{L}^{(2,\infty)})^{(0)} \mathcal{L}^{(2,\infty)} \in (\mathcal{L}^{(1,\infty)})^{(0)}$ (see the Appendix for a proof). The Dixmier trace vanishes on $(\mathcal{L}^{(1,\infty)})^{(0)}$ and this shows that the form is zero. In general, since the tangent space has the direction given by $(\mathcal{L}^{(2,\infty)})^{(0)}$, the cotangent space has one-forms which do not vanish on them. If Λ is a form such that $\Lambda(V_u) \neq 0$, then $\Lambda = i_Y \Omega_\omega$ has no solution for the vector Y. If we assume that the de Rham theory makes sense on these spaces, since D_{1+} is contractible, and $\pi_1(\mathrm{Gr}_{1+}) = 0$, we would expect that there is a function f such that $\Lambda = df$, and this will show that one cannot obtain a Hamiltonian vector field for any given function f in general. Nevertheless, as we will see, for the relevant part of the space, that is for ''large'' motions, directions which belong to $\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$, the form is nondegenerate. This will allow us to define classical dynamics for certain systems.

Before we continue, let us point out an important observation. For clarity let us concentrate on Gr_{1+}. There is an interesting leaf of Gr_{1+} which corresponds to the orbit of ϵ under the following subgroup:

$$\mathrm{U}_{1+}^{(0)} = \begin{pmatrix} \mathcal{B} & (\mathcal{L}^{(2,\infty)})^{(0)} \\ (\mathcal{L}^{(2,\infty)})^{(0)} & \mathcal{B} \end{pmatrix}. \tag{19}$$

We denote this orbit by $\mathrm{Gr}_{1+}^{(0)}$. Since the connected components are still labeled by the integers, each connected component of Gr_{1+} has the connected component of $\mathrm{Gr}_{1+}^{(0)}$ inside. If we take $\mathrm{Gr}_{1+}^{(0)}$'s own tangent space, generated by $\mathrm{U}_{1+}^{(0)}$, the symplectic form vanishes on this orbit using the fact that $(\mathcal{L}^{(2,\infty)})^{(0)} \mathcal{L}^{(2,\infty)} \in (\mathcal{L}^{(1,\infty)})^{(0)}$ (see the Appendix for a proof). The same remarks apply to the disc D_{1+}, and we spare the reader the details.

Since the group action preserves the two from Ω_ω the Lie derivative along the direction of any vector field generated by the group action gives us zero. This raises the possibility of finding the moment maps which would generate the infinitesimal action of $\mathrm{U}_{1+}(\mathcal{H}_-, \mathcal{H}_+)$ and $\mathrm{U}_{1+}(\mathcal{H})$, respectively. We can start with the finite-dimensional answer; one can check that to avoid the divergence the finite-dimensional answer has to be modified as $-\mathrm{Tr}_\omega^\epsilon u(\Phi - \epsilon)$, where u is a Hermitian matrix which is in the Lie algebra of $\mathrm{U}_+(\mathcal{H})$ for the Gr_{1+} and a pseudo-Hermitian ($u^\dagger = \epsilon u \epsilon$) operator which belongs to $\mathrm{U}_{1+}(\mathcal{H}_-, \mathcal{H}_-)$ for D_{1+}. We use a conditionally convergent trace for the variable $\Phi - \epsilon$ if we think of the group acting from the left. To see that the above expression makes sense consider $u \in \left(\begin{smallmatrix} \mathcal{B} & \mathcal{L}^{(2,\infty)} \\ \mathcal{L}^{(2,\infty)} & \mathcal{B} \end{smallmatrix}\right)$, and $\Phi - \epsilon \in \left(\begin{smallmatrix} \mathcal{L}^{(1,\infty)} & \mathcal{L}^{(2,\infty)} \\ \mathcal{L}^{(2,\infty)} & \mathcal{L}^{(1,\infty)} \end{smallmatrix}\right)$. (Here $\mathcal{B}$ is the space of bounded operators). As a result, $u(\Phi - \epsilon) \in \left(\begin{smallmatrix} \mathcal{L}^{(1,\infty)} & \mathcal{L}^{(2,\infty)} \\ \mathcal{L}^{(2,\infty)} & \mathcal{L}^{(1,\infty)} \end{smallmatrix}\right)$. Let us define $M = \Phi - \epsilon$. Formally the equation for the group action is satisfied,

$$-\mathrm{Tr}_\omega^\epsilon\, u\, d\Phi = \Omega_\omega(V_u, \ ), \quad V_u(\Phi) = [u, \Phi], \tag{20}$$

but we must be careful since if $g^{-1} u g \in (\mathcal{L}^{(2,\infty)})^{(0)}$, the right-hand side vanishes as we have shown. The only way to satisfy this equation is to show that whenever the right-hand side vanishes, the left vanishes as well. This can be checked as follows:

$$df_u(V_v)=\mathrm{Tr}^{\epsilon}_{\omega}\, u[v,\Phi]=\mathrm{Tr}^{\epsilon}_{\omega}\, g^{-1}ug[\epsilon,g^{-1}vg], \tag{21}$$

for any $v\in \mathrm{U}_{1+}$. By the same argument as above, we see that the resulting expression inside the trace is of class $(\mathcal{L}^{(1,\infty)})^{(0)}$. Hence the trace is also zero.

We notice that the form corresponding to the moment map $-\mathrm{Tr}^{\epsilon}_{\omega}(uM^{(0)})$ for any $u\in \mathrm{U}_{1+}$ when contracted with the elements of the tangent space of $D_1^{(0)}$ or $Gr_{1+}^{(0)}$ also gives zero. This is in some sense the orbit one can neglect; we can think of it as the null orbit. This shows that we cannot take the separable part in our definition of the group U_{1+}, denoted as $(\mathrm{U}_{1+})^{(0)}$, if we use the Dixmier trace. In a recent preprint,[12] it is argued that there is no positive trace on the ideal $(\mathcal{L}^{(2,\infty)})^{(0)}$. This implies that a similar construction cannot be achieved for the group $(\mathrm{U}_{1+})^{(0)}$.

However, it is not enough to show that the moment functions satisfy a consistent equation. Normally in the Hamiltonian formalism we are given a function and asked to find the vector field generated by this function. Our discussion shows that this vector field at every point is only determined in the equivalence class $\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$. It is easy to see that if we take, instead of u, a vector generated by $u+v$ such that $g^{-1}vg\in(\mathcal{L}^{(2,\infty)})^{(0)}$, $-df_u=i/8\,\mathrm{Tr}_{\omega}\Phi[[u+v,\Phi],d\Phi]$ $=i/8\,\mathrm{Tr}_{\omega}\Phi[[u,\Phi],d\Phi]+i/8\,\mathrm{Tr}_{\omega}[[v,\Phi],d\Phi]$ for the last piece is zero by the previous argument. This implies that the moment functions do not have unique vector fields; they generate the motions in the ''transversal direction'' $\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$ with an undetermined piece in $(\mathcal{L}^{(2,\infty)})^{(0)}$. Let us look at the infinitesimal part of this evolution for two different choices of the vector field in the equivalence class, $\Delta\Phi=t[u,\Phi]$ and $\Delta'\Phi=t[u',\Phi]$. The difference of these two infinitesimal evolutions is given by $\Delta\Phi-\Delta'\Phi=[u-u',\Phi]=g[g^{-1}(u-u')g,\epsilon]g^{-1}$. Since the ambiguity is a result of the difference, which satisfies $[g^{-1}(u-u')g,\epsilon]\in(\mathcal{L}^{(2,\infty)})^{(0)}$, this term is in the orbit $D_{1+}^{(0)}$, or in the same connected component of $\mathrm{Gr}_{1+}^{(0)}$. Thus the difference of the infinitesimal evolutions can be conjugated to the orbit $\mathrm{D}_{1+}^{(0)}$, or to the same connected component of $\mathrm{Gr}_{1+}^{(0)}$. This implies that the relevant space for the classical dynamics is not the original quotient we look for, but a smaller one, given by $\mathrm{U}_{1+}/\mathrm{U}_{1+}^{(0)}$. In fact as we will see later on, this reduction has an interesting consequence. But, for the moment, we will continue to use the ''unreduced phase space.''

To talk about classical evolution, we will make the proposal that this type of classical systems is defined through an equivalence relation. We will assume that two dynamical evolutions are equivalent if they could be conjugated to the ''null'' orbit, $\mathrm{D}_{1+}^{(0)}$, for D_{1+}, and to the same connected component of $\mathrm{Gr}_{1+}^{(0)}$, for Gr_{1+}. Another way to think about this is that the dynamics in the ''small'' directions cannot be determined. The moment functions we consider only determine the evolution under equivalence.

This feature will affect the dynamics generated by Hamiltonians of more complicated functions. The generic Hamiltonians we have in mind are quadratic functions of the variable $M=\Phi-\epsilon$ plus a moment map. We think of the moment map as the free part of the Hamiltonian since it generates the group action up to an equivalence, and the quadratic piece as an ''interaction.''

Formally we can write the Hamiltonians as

$$h=i\,\mathrm{Tr}^{\epsilon}_{\omega}\, uM+\mathrm{Tr}^{\epsilon}_{\omega}\,\hat{K}(M)M, \tag{22}$$

where $\hat{K}$ is a linear operator which acts on the variable M. If we specify a basis, it can be expressed as $(\hat{K}(M)M)^k_p=\Sigma_{ijk}K^{ik}_{jl}M^j_iM^l_p$. The summability properties of the kernel K^{ik}_{jl} should be such that the resulting operator is in the ideal $\mathcal{L}^{(1,\infty)}$. For example, this can be achieved if the map $\hat{K}$ produces an operator in the group U_{1+}. Notice that it is important to use the Dixmier trace again. As we mentioned in the moment maps, since the symplectic form Ω_{ω} vanishes upon contraction with elements of the tangent space in the ''small'' directions, the interaction Hamiltonian must have the same property. As one can check, this form of the Hamiltonian, when differentiated, gives a form which vanishes on the same subspace. This will impose certain conditions on the choice of kernels. Interactions which are ''too weak'' will not affect the equations of motion.

Let us give a typical interaction Hamiltonian. We can take two moment functions and take their products

$$\sum_{ij} K^{ij}\,\mathrm{Tr}_\omega^\epsilon\, u_i M\ \mathrm{Tr}_\omega^\epsilon\, u_j M. \tag{23}$$

If the sum is over a finite number of terms, then clearly this is a well-defined expression. In general, one should be able to choose K^{ij} such that the above expression is finite. For these Hamiltonians, one can directly calculate the equations of motion by using the Poisson bracket relations among the moment functions only. As we will see later on, they also have a simpler description in the quantum case.

A word of caution should be said here. Typically, the Hamiltonians are more singular than the symplectic form, and they require further renormalizations. This implies a choice of domain for the Hamiltonian. This should restrict the accessible regions of the phase space, or the true phase space of the theory, and it may change the formulation of the problem drastically. In real physical systems, we expect these problems to modify the precise formulation of the field theory. In this work, we do not discuss this more difficult problem. In some sense this is part of the kinematics, although interactions may even change this.

There will always be vector fields generating the equations of motion up to equivalence. In infinite dimensions determination of the integral curves and their completeness are highly nontrivial issues; the answer depends on the Hamiltonian as well. This is the classical version of the unitarity condition in quantum mechanics.

From general principles we expect that the Poisson brackets of the moment functions will provide a realization of the corresponding Lie algebra possibly with a central extension. We can calculate the Poisson bracket of two moment functions using a formal manipulation and it gives us

$$\{f_u, f_v\} = f_{-i[u,v]} - i\,\mathrm{Tr}_\omega^\epsilon[\epsilon, u]v. \tag{24}$$

This is a central extension of the full group, indexed by the choice of limit process ω. (Note that there is no ambiguity in this relation since the right-hand side is the same for equivalent choices of the vector fields.) One can explicitly check that $c_\omega(u,v) = \mathrm{Tr}_\omega^\epsilon[\epsilon, u]v$ satisfies $c_\omega(u,v) = -c_\omega(v,u)$ and the cocycle condition, $c_\omega(u,[v,w]) + c_\omega(v,[w,u]) + c_\omega(w,[u,v]) = 0$. Since the central term vanishes on the ideal $(\mathcal{L}^{(1,\infty)})^{(0)}$ it will not be there when u or $v \in \mathrm{U}_{1+}^{(0)}$. The actual computation should be done with care, due to infinite dimensionality. One can see that the both sides of the equation after explicit calculation give the same result, hence they are identical. We will leave the details to the reader.

This gives us a symplectic realization of the Lie algebra of $\mathrm{U}_{1+}(\mathcal{H})$ for Gr_{1+} and $\mathrm{U}_{1+}(\mathcal{H}_-, \mathcal{H}_+)$ for D_{1+} except a central term. Since the calculation of the both sides are actually zero whenever $g^{-1}ug$ or $g^{-1}vg \in \mathrm{U}_{1+}^{(0)}$, this expression should be thought of as a realization of the ''large'' part of the Lie algebra. [It is more natural to look into the spaces which are modeled on the quotients $\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$. Since $(\mathcal{L}^{(2,\infty)})^{(0)}$ is closed, the quotient is well defined. Later on we will consider this point of view.] Let us consider this central extension for the pseudodifferential operators (pointed out to me by Mickelsson); then we have $\mathrm{Tr}_\omega[\epsilon,u]v = \mathrm{Res}([\epsilon,u]v)$. We recall that the residue is actually defined for all pseudodifferential operators and it satisfies $\mathrm{Res}[A,B] = 0$.[13,14] This implies that, in fact $\mathrm{Tr}_\omega[\epsilon,u]v = \partial\phi(u,v)$, for $\phi(u) = \mathrm{Res}(\epsilon u)$, hence the central extension is trivial.

Because of the infinite dimensionality, an attempt to remove the central term, in the general case, will result in a divergent expression. This gives us the Lie algebra of the nontrivial central extension of U_{1+} corresponding to ω. We expect that these extensions are not equivalent, in general, for different choices of ω, but they all agree on the subset modeled on the ''measureable'' elements—this is not a subalgebra of $\underline{U_{1+}}$, so one cannot reduce it to this case.

III. QUANTIZATION

We continue to think of classical mechanics in geometric terms.[15,16] Let us assume that the phase space, Γ, is a smooth manifold. If we have a Poisson structure on the algebra of smooth functions $C^\infty(\Gamma)$, we can introduce classical dynamics. Quantization of this classical system is given by a representation of the Poisson algebra of smooth functions by self-adjoint operators on a Hilbert space. This is an overambitious program; in general, there is no way to find such a representation. The difficulties and various methods have been well explained in the literature.[17–19] In this article we will follow our point of view in Ref. 2. We will find a representation of the Poisson algebra of moment functions. Any composite function, which is related to the product of two moment maps, can be quantized by giving an ordering rule. We will not attempt to establish this idea in the present article.

Before we proceed further, we need to make a digression and introduce a generalized ''determinant.'' It does not satisfy all the properties of a determinant. As we will see, to think of it as a determinant simplifies the calculations. Let us define $\det_\omega(1+A)$ for $A\in\mathcal{L}^{(1,\infty)}$ as

$$\det_\omega(1+A)=\exp(\mathrm{Tr}_\omega(A)). \tag{25}$$

[One way to motivate this definition is the following. Let us take the determiant formula, $\log\det(1+A)=\mathrm{Tr}\log(1+A)$. We replace the trace by the Dixmier trace, and define $\log\det_\omega(1+A)=:\mathrm{Tr}_\omega\log(1+A)=\mathrm{Tr}_\omega(A-\frac{1}{2}A^2+\frac{1}{3}A^3-\cdots)=\mathrm{Tr}_\omega(A)$ since all the higher terms are in $\mathcal{L}^1$ and the Dixmier trace vanishes on them. This gives the above formula again.] One can see that it satisfies the multiplicative property of the determinants,

$$\det_\omega((1+A)(1+B))=\det_\omega(1+A)\det_\omega(1+B), \tag{26}$$

due to $\mathcal{L}^{(1,\infty)}\mathcal{L}^{(1,\infty)}\in\mathcal{L}^1$ and Tr_ω vanishes on the trace class operators $\mathcal{L}^1$. An interesting property is that $\det_\omega$ never vanishes.

We will use this to provide a representation of the Borel subroup on **C**, and attempt to follow the geometric quantization program.

We will introduce an *ad hoc* representation, which comes from the geometric quantization performed in Ref. 2, of the Lie algebra of the group $U_{1+}(\mathcal{H}_-,\mathcal{H}_+)$ on the space of ''holomorphic functions'' on D_{1+}:

$$\hat{f}_{-u}\Psi(Z)=-i\hbar\left[\mathcal{L}_{V_u}\Psi(Z)-\frac{1}{\hbar}\mathrm{Tr}_\omega(\gamma Z)\Psi(Z)\right], \tag{27}$$

where $u\in \underline{U}_1(\mathcal{H}_-,\mathcal{H}_+)$, and $V_u=V_u^Z\partial_Z+V_u^{Z^\dagger}\partial_{Z^\dagger}$ is the formal vector field generated by the action of $-u$. (The author is not aware of a well-established definition of holomorphicity for infinite dimensions. We assume that the algebraic operations on the coordinate Z after the application of a dual element, if it is finite, provides a holomorphic function. This will be used for the Grassmannian as well. It is easier to define the Lie derivative directly, by using the action of the Lie algebra on the disc. We define this as

$$\mathcal{L}_{V_u}\Psi(Z)=\lim_{t\to 0}\frac{\Psi(Z+t(\alpha Z+\beta-Z\gamma Z-Z\delta))-\Psi(Z)}{t}, \tag{28}$$

where

$$u=\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} \tag{29}$$

is the decomposition of u into block form; $\alpha^\dagger=-\alpha$, $\beta^\dagger=\gamma$, and $\delta^\dagger=-\delta$; and, further, $\gamma,\beta\in\mathcal{L}^{(2,\infty)}$. We notice that this differs from the finite-dimensional answer by a constant term, which is infinite in this case. The changes introduced in the functions are all ‘‘holomorphic,’’ so the action of the operators corresponding to the moment functions preserve the ‘‘holomorphicity condition.’’ In order to show that this is the correct representation, we need to prove that the commutation relations are satisfied acting on these set of functions.

Let us check that the commutation relations will give us a realization of the Poisson bracket relations satisfied by the moment maps:

$$[\hat{f}_{-u_1},\hat{f}_{-u_2}]\Psi(Z)=i\hbar\hat{f}_{[u_1,u_2]}\Psi(Z)+i\hbar\,\mathrm{Tr}_\omega^\epsilon([\epsilon,u_1]u_2)\Psi(Z). \tag{30}$$

A calculation shows that

$$[\hat{f}_{-u_1},\hat{f}_{-u_2}]\Psi(Z)=\hbar^2(\mathcal{L}_{u_1}\mathcal{L}_{u_2}-\mathcal{L}_{u_2}\mathcal{L}_{u_1})\Psi(Z)+\hbar[(\mathcal{L}_{u1}\,\mathrm{Tr}_\omega\,\gamma_2 Z)-(\mathcal{L}_{u_2}\,\mathrm{Tr}_\omega\,\gamma_1 Z)]\Psi(Z). \tag{31}$$

For simplicity we use $\mathcal{L}_u$ instead of $\mathcal{L}_{V_u}$. This is equal to

$$[\hat{f}_{-u_1},\hat{f}_{-u_2}]\Psi(Z)=\hbar^2\mathcal{L}_{[u_1,u_2]}\Psi(Z)+\hbar\,\mathrm{Tr}_\omega\{\gamma([u_1,u_2])Z\}\Psi(Z)+\hbar\,\mathrm{Tr}_\omega(\gamma_1\beta_2-\gamma_2\beta_1)\Psi(Z). \tag{32}$$

The last term, which is a constant multiple, can be rewritten as $\mathrm{Tr}_\omega^\epsilon[\epsilon,u_1]u_2$. Hence we see that it is a representation of the Poisson bracket relations (24).

This representation can, in fact, be integrated to a representation of the group action on the space of holomorphic functions:

$$\rho_\omega(g^{-1})\Psi=\mathrm{det}_\omega^{-1/\hbar}(d^{-1}cZ+1)\Psi((aZ+b)(cZ+d)^{-1}). \tag{33}$$

These representations are labeled by ω, the choice of limit point, and $\hbar$. Since the determinant never vanishes and is actually given by an exponential, $\hbar$ is any real number. (This point of view on quantization for finite-dimensional homogeneous spaces appeared in Ref. 20.) To justify this, we will compute the infinitesimal form of the representation and show that it is given by the operators corresponding to the moment maps. We write explicitly

$$\rho(g^{-1})\Psi(Z)=e^{-(1/\hbar)\mathrm{Tr}_\omega(d^{-1}cZ)}\Psi(g\circ Z) \tag{34}$$

and evaluate

$$\lim_{t\to 0}\left\{\frac{\rho_\omega(1+tu)-\rho_\omega(1)}{t}\right\}\Psi(Z)=\hbar\left[\mathcal{L}_{V_u}-\frac{1}{\hbar}\mathrm{Tr}_\omega(\gamma Z)\right]\Psi(Z). \tag{35}$$

So, we see that the infinitesimal form is given by the moment map operators. We still have to check that this is a representation:

$$\rho_\omega(g_1)\rho_\omega(g_2)\Psi(Z)=c_\omega(g_1,g_2)\rho_\omega(g_1g_2)\Psi(Z), \tag{36}$$

where $c_\omega(g_1,g_2)$ is a central term, which satisfies $c_\omega(g_1g_2,g_3)c_\omega(g_1,g_2)=c_\omega(g_1,g_2g_3)c_\omega(g_2,g_3)$. Since the disc is topologically trivial, the central extension can be described by specifying a function from the Cartesian product of the group to **C**. An explicit calculation, which is given in the Appendix, shows that the group property is satisfied with a central term,

$$c_\omega(g_1,g_2)=\mathrm{det}_\omega^{1/\hbar}[(d_1d_2)^{-1}c_1b_2+1]=\exp\left(\frac{1}{\hbar}\mathrm{Tr}_\omega[(d_1d_2)^{-1}c_1b_2]\right). \tag{37}$$

However, there is still one more point we need to consider. Recall that the orbit of ϵ corresponding to the subgroup $U_{1+}^{(0)}, D_{1+}^{(0)}$, when considered as a submanifold with its own tangent space, has no dynamics under our choice of the symplectic form. Hence the Poisson algebra of moment maps restricted on this submanifold is trivially true, being zero on both sides, whereas the representation space we choose, when restricted to the subspace $D_{1+}^{(0)}$, provides a nontrivial representation of the Poisson algebra of moment maps. This can be rectified by selecting a subspace of holomorphic functions which remain constant on the orbit $D_{1+}^{(0)}$. As a result, we define the quantum Hilbert space to be

$$\mathcal{H}_Q=\{\Psi(Z)|\Psi(Z) \quad \text{holomorphic on } \mathrm{D}_{1+} \quad \text{and} \quad \Psi|_{\mathrm{D}_{1+}^{(0)}}=\text{constant}\}. \tag{38}$$

This condition is consistent with the assumption that the dynamics along the directions $(\mathcal{L}^{(2,\infty)})^{(0)}$ is unimportant. If we consider ϵ to be the ''vacuum'' configuration, its orbit under $\mathrm{U}_{1+}^{(0)}$ is equivalent to this ''vacuum'' configuration.

For this choice of the quantum Hilbert space, we will exhibit a class of wave functions. We are unable to prove that these are the only possible ones. Since the wave functions should be constant on the orbit $\mathrm{D}_{1+}^{(0)}$, it suggests the use of the Dixmier trace again.

We can compose polynomials in the variable Z. The interesting thing is to note that we can go up to quadratic terms only inside the Dixmier trace. For any choice of $A_i \in \mathcal{L}^{(2,\infty)}$, and any two $B_j, B_k \in B(\mathcal{H})$, we can form

$$\mathrm{Tr}_\omega(A_i Z) \;\text{ and }\; \mathrm{Tr}_\omega(B_j Z B_k Z). \tag{39}$$

Using the generalized Hölder inequality,[21] we have the inequalities

$$|\mathrm{Tr}_\omega(A_i Z)| \le \|A_i\|_{\mathcal{L}^{(2,\infty)}}\|Z\|_{\mathcal{L}^{(2,\infty)}} \quad \text{and} \quad |\mathrm{Tr}_\omega(B_k Z B_j Z)| \le \|B_k\|\|B_j\|\|Z\|^2_{\mathcal{L}^{(2,\infty)}}. \tag{40}$$

These show the continuity in Z and with respect to A_i and B_j, B_k. We can compose products of these kinds of functions:

$$\Psi(Z)=\prod_{i,j,k} \mathrm{Tr}_\omega(B_j Z B_k Z)\mathrm{Tr}_\omega(A_i Z). \tag{41}$$

The reader can verify that any higher power of Z is irrelevant, so these are the only combinations we can make. Various superpositions of these functions will give us the set of wave functions. [One should perhaps compare this with the analysis given in Ref. 2. There, the holomorphic wave functions can be constructed using the analogy with the finite-dimensional case. We can use the fact that the dual of $\mathcal{L}^2$ is itself and the dual of $\mathcal{L}^1$ is $\mathrm{B}(\mathcal{H})$. We can write down a general holomorphic function as sums and products of the expressions of the form, $\mathrm{Tr}(AZ), \mathrm{Tr}(B_1 Z B_2 Z \cdots B_m Z)$, where $A \in \mathcal{L}^2$ and $B_k \in \mathrm{B}(\mathcal{H})$ for all $k=1,...,m$. Not all of them are linearly independent of course. With the appropriate inner product the quantum Hilbert space constructed out of these functions is isomorphic to the usual Fock space and it is a separable Hilbert space.]

We also would like to point out an interesting property of our wave functions. Since we have obtained them through the Dixmier trace, we effectively perform ''logarithmic wave function renormalization.''

The quantum Hilbert space may be very large, or maybe very small; it depends on the size of the space of ''holomorphic functions'' on D_{1+}, and the inner product. It is not clear how one should introduce a measure to define an inner product in this quantum Hilbert space. The term Hilbert space is only justified by thinking of this as a quantization of a classical system. It should be possible to extend the work in Ref. 22 to this case. Once this is done, the completion of the above set of wave functions with respect to this measure will be the quantum Hilbert space, $\mathcal{H}_Q$.

Next we will construct the quantum operators for the Gr_{1+}. Our approach will be somewhat *ad hoc* again. Since Gr_{1+} is topologically nontrivial, it may not be possible to represent wave functions as functions on the Grassmannian. It is natural to introduce them as sections of a complex line bundle. However, we will see that there are nonconstant holomorphic functions on this Grassmannian. (This can be contrasted with Refs. 3 and 2. There, the finite-dimensional Grassmannians are dense inside the full Grassmannian, and it is well known that on a compact complex manifold there are no nonconstant holomorphic functions. In our case, completion of finite rank objects will not be equal to the full space.) For the quantization of our classical system, we actually need the sections of a line bundle. For this it is better to think of Gr_{1+} as a quotient of another pair of groups as in the case discussed by Refs. 3 and 2. It is possible that the extension is nontrivial both topologically and algebraically. Essentially, we will enlarge ''numerator'' and ''denominator'' by the same amount. This will keep the same ratio. We define the group $\tilde{\mathrm{G}}_{1+}$:

$$\tilde{\mathrm{G}}_{1+}=\{(\gamma,q)|q\in \mathrm{GL}(\mathcal{H}_-);\quad \gamma\in \mathrm{GL}_{1+}(\mathcal{H}),\quad \gamma_{11}q^{-1}-1\in\mathcal{L}^{(1,\infty)}\}. \tag{42}$$

Here, γ_{11} denotes the mapping $\gamma_{11}:\mathcal{H}_-\to\mathcal{H}_-$ in the block form of the matrix $\gamma\in\mathrm{GL}_{1+}(\mathcal{H})$. One can prove that the set of q's which satisfies this condition is not empty using the definition of the group GL_{1+} following Pressley and Segal.[3] We can give a topology to this space using the two topologies inherited from the bounded and $\mathcal{L}^{(1,\infty)}$. (Notice that the extension for any two-sided ideal is mentioned in Ref. 3 on p. 98.)

Here $\tilde{\mathrm{G}}_{1+}$ is a complex Banach–Lie group under the multiplication $(\gamma,q)(\gamma',q')=(\gamma\gamma',qq')$. We introduce $\tilde{\mathrm{B}}_{1+}$, a closed complex subgroup of $\tilde{\mathrm{G}}_{1+}$;

$$\tilde{\mathrm{B}}_{1+}=\{(\beta,t)|\beta\in \mathrm{B}_{1+},\quad t\in\mathrm{GL}(\mathcal{H}_-),\quad \beta_{11}t^{-1}-1\in\mathcal{L}^{(1,\infty)}\}. \tag{43}$$

There is an action of $\tilde{\mathrm{B}}_{1+}$ on $\tilde{\mathrm{G}}_{1+}$. Since this action does not involve anything but multiplication in the group, it is holomorphic. We enlarged $\mathrm{GL}_{1+}(\mathcal{H})$ and B_{1+} with the same set of elements, thus the quotient is still the same:

$$\tilde{\mathrm{B}}_{1+}\to\tilde{\mathrm{G}}_{1+}\to\mathrm{Gr}_{1+}. \tag{44}$$

In this case as well, there are subgroups corresponding to completions of the finite rank elements. These subgroups now can be written as

$$\tilde{\mathrm{G}}_{1+}^{(0)}=\{(\gamma^{(0)},q)|q\in\mathrm{GL}(\mathcal{H}_-);\quad \gamma^{(0)}\in\mathrm{GL}_{1+}^{(0)}(\mathcal{H}),\quad \gamma_{11}^{(0)}q^{-1}-1\in(\mathcal{L}^{(1,\infty)})^{(0)}\}. \tag{45}$$

Existence can be proved along the same lines. The stability subgroup also changes to $\tilde{\mathrm{B}}_{1+}^{(0)}$, which is defined similarly. One can see that $\tilde{\mathrm{G}}_{1+}^{(0)}\subset\tilde{\mathrm{G}}_{1+}$ as a closed subset. This subgroup will correspond to the null orbit, and it will not have any extension.

Now, we can introduce the holomorphic line bundle corresponding to the representation $\rho(\beta,r)=\det_\omega^{1/\hbar}(\beta_{11}r^{-1})$. There is no condition for the number $\hbar$ to be an integer at this stage, since our definition for $\det_\omega$ has an exponential, it never vanishes, and the value of $\hbar$ could be any real number. We denote the line bundle as $(\tilde{\mathrm{G}}_{1+}\times_\rho\mathbf{C})/\tilde{\mathrm{B}}_{1+}$. A section of this line bundle can be identified with equivariant functions:

$$\psi:\tilde{\mathrm{G}}_{1+}\to\mathbf{C}\quad\text{such that}\quad\psi(\gamma\beta,qr)=\rho(\beta,r)\psi(\gamma,q). \tag{46}$$

Let us exhibit the functions which would satisfy this condition. They are given by generalized determinants very similar to the case discussed in Ref. 2. As an example we start with

$$\psi(\gamma,q)=\det_\omega^{1/\hbar}(\gamma_{11}q^{-1}). \tag{47}$$

One can see that this is an equivariant function on the space $\tilde{\mathrm{G}}_{1+}$, using the properties of the Dixmier trace. There is no restriction on the value of $\hbar$ due to holomorphicity, since the exponential is an entire function except possibly due to a topological obstruction. We can compose more functions of this type, if we allow ''mixing'' of the elements of γ_{11} with the elements of γ_{21}, in a controlled way. One can see that if we assume that the mixing is allowed by finite rank operators, the result will not change. More than that, it will not change if we use elements of $(\mathcal{L}^{(2,\infty)})^{(0)}$. To get different functions we need to mix them by elements of $\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$:

$$\det_\omega((1-A_iS)\gamma_{11}q^{-1}+A_i\gamma_{21}q^{-1}) \tag{48}$$

for $A_i\in\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-)$ and the mapping $S:\mathcal{H}_-\to\mathcal{H}_+$ is an isometric isomorphism which is given by mapping one set of orthonormal basis elements into the other (it can simply be taken as sending $e_{-i}\to e_i$). This form is guessed from the system studied in Refs. 2 and 3. However, we have two problems with this form. When we use $A_iS\gamma_{11}$, this expression is not convergent under the Dixmier trace. The second is that we want our wave functions to be constant when they are restricted to the null orbit. As we will see, this is not possible for the above form of the functions due to the term $A_iS\gamma_{11}$. One can see that dropping this term does not change the equivariance condition thanks to the Dixmier trace again. This means an infinite multiplicative renormalization. At the same time we see that it is simpler to just multiply with the function $\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})$, since it is invariant under the action of B_{1+}, hence it descends to a function on the quotient, Gr_{1+}. A similar argument shows that we can do better, we may add even a nonlinear term, which is still invariant. Any higher-order addition vanishes. We can write down a general expression,

$$\det_\omega^{1/\hbar}(\gamma_{11}q^{-1})\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})\mathrm{Tr}_\omega(B_j\gamma_{21}q^{-1}B_k\gamma_{21}q^{-1}), \tag{49}$$

where $B_j,B_k\in\mathrm{B}(\mathcal{H}_+,\mathcal{H}_-)$. For clarity, we will prove that this form is equivariant and satisfies all the requirements.

Let us look at the action by an element of B_{1+};

$$\begin{aligned}&\det_\omega^{1/\hbar}((\gamma\beta)_{11}(qr)^{-1})\mathrm{Tr}_\omega(A_i(\gamma\beta)_{21}(qr)^{-1})\mathrm{Tr}_\omega(B_j(\gamma\beta)_{21}(qr)^{-1}B_k(\gamma\beta)_{21}(qr)^{-1})\\&\quad=\det_\omega^{1/\hbar}(\gamma_{11}\beta_{11}r^{-1}q^{-1})\mathrm{Tr}_\omega(A_i\gamma_{21}\beta_{11}r^{-1}q^{-1})\mathrm{Tr}_\omega(B_j\gamma_{21}\beta_{11}r^{-1}q^{-1}B_k\gamma_{21}\beta_{11}r^{-1}q^{-1})\\&\quad=\det_\omega^{1/\hbar}(\gamma_{11}(1+I)q^{-1})\mathrm{Tr}_\omega(A_i\gamma_{21}(1+I)q^{-1})\mathrm{Tr}_\omega(B_j\gamma_{21}(1+I)q^{-1}B_k\gamma_{21}(1+I)q^{-1})\end{aligned}$$

using $\beta_{11}r^{-1}=1+I$ for $I\in\mathcal{L}^{(1,\infty)}$;

$$\begin{aligned}&\det_\omega^{1/\hbar}(\gamma_{11}q^{-1}(1+qIq^{-1}))\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1}+A_i\gamma_{21}Iq^{-1})\mathrm{Tr}_\omega((B_j\gamma_{21}q^{-1}+B_j\gamma_{21}Iq^{-1})\\&\qquad\times(B_k\gamma_{21}q^{-1}+B_k\gamma_{21}Iq^{-1}))\\&\qquad=\det_\omega(\gamma_{11}q^{-1}(1+qIq^{-1}))\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})\mathrm{Tr}_\omega(B_j\gamma_{21}q^{-1}B_k\gamma_{21}q^{-1})).\end{aligned}$$

In the last line we have used $\mathcal{L}^{(1,\infty)}\mathcal{L}^{(2,\infty)}\in\mathcal{L}^1$ and the Dixmier trace vanishes on them. As a result we get

$$\exp\frac{1}{\hbar}(\mathrm{Tr}_\omega(\gamma_{11}g^{-1}-1)+\mathrm{Tr}_\omega(\beta_{11}r^{-1}-1))\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})\mathrm{Tr}_\omega(B_j\gamma_{21}q^{-1}B_k\gamma_{21}q^{-1}), \tag{50}$$

where we use $(1+I)(1+qIq^{-1})=1+I+qJq^{-1}+IqJq^{-1}$ and the last term is zero inside the Dixmier trace for $\gamma_{11}q^{-1}=1+I$, $I\in\mathcal{L}^{(1,\infty)}$. This is the equivariance condition we want. We can take a product over the trace part only, and this will not change the result.

We need to check that when we reduce the wave function onto the ''small'' orbit, that is to the subgroup $\tilde{\mathrm{G}}_{1+}^{(0)}$, the resulting wave functions are just constant. This is necessary for consistency with the classical Poisson bracket calculation as we will see. Let us note the following: in fact, our system is invariant under a larger symmetry. This will be discussed in the next section, and we will see that the correct phase space is smaller.

One can see that we can compose a product wave function; in general, we define, as in the case of D_{1+},

$$\det_{\omega}^{1/\hbar}(\gamma_{11}q^{-1})\prod_{i,j,k}\ \mathrm{Tr}_{\omega}(A_i\gamma_{21}q^{-1})\mathrm{Tr}_{\omega}(B_j\gamma_{21}q^{-1}B_k\gamma_{21}q^{-1}). \tag{51}$$

These and their various superpositions are the most general wave functions we can construct. The next step is to complete these sets of wave functions with respect to an inner product to get the quantum Hilbert space. (Notice that in our previous work[2] we could have used the trace class operators for the mixing of γ_{11} and γ_{21}, not only the finite rank ones. Various superpositions of the wave functions described in that work will lead to a similar general form given here with a finite rank matrix used for the mixing. The finite rank operators are dense in the trace class and the determinant uses the ordinary trace. We can extend the wave functions in that case to the one with mixing elements in the trace class operators. This should actually be done by using the inner product in the quantum Hilbert space, but we expect some dominance property with respect to the parameters. In fact, the above claim is known to be true, hence there is no loss of generality in that case.) Notice that still there could be an integrality condition on $\hbar$, due to the fact that the first Chern class of the line bundle corresponding to the representation ρ should be in $H^2(\mathrm{Gr}_{1+},\mathbf{Z})$.

Let us note that the form of the wave function for the disc and the Grassmannian are quite the same—this is unlike the previous case studied in Ref. 2. We can, in fact, set up a one-to-one correspondence between the elements of the two quantum Hilbert spaces; for any choices of $A_i \in \mathcal{L}^{(2,\infty)}$ and B_j, B_k, which are operators in $\mathrm{B}(\mathcal{H}_-,\mathcal{H}_+)$ we have a wave function on the disc and on the Grassmannian. This suggests in some natural sense a boson–fermion correspondence, if we think of the disc corresponding to a bosonic and the Grassmannian to a fermionic system. Of course, this correspondence is only at a formal level, since we only have a set theoretical relation between the two quantum Hilbert spaces. One also has to check the inner products, to make sure that the linearly independent choices are mapped to each other in the same manner. This seems to be a reasonable expectation. Its physical meaning is not so obvious to the author; it maybe due to the fact that we have thrown away a large part of the phase space, which may have a lot of physical information about the system. We will see some further evidence of the equivalence of these two systems in this quantization scheme, when we look at these problems in a different way in the next section. (There is larger stability subgroup as we will see in the next section, and it seems to suggest a simpler solution.)

The above wave functions carry a representation of the group $\tilde{\mathrm{G}}_{1+}$ which comes from the left action;

$$\mathbf{r}(\gamma',q')\psi(\gamma,q)=\psi(\gamma'^{-1}\gamma,q^{-1}q'). \tag{52}$$

This group action is well defined. We give a proof for completeness: Let us denote the inverse element acting from the left by $(\lambda,r)\in \mathrm{G}_{1+}$. For simplicity we drop the products and compute the following expression:

$$\det_{\omega}^{1/\hbar}((\lambda\gamma)_{11}q^{-1}r^{-1})\mathrm{Tr}_{\omega}(A_i(\lambda\gamma)_{21}q^{-1}r^{-1})\mathrm{Tr}_{\omega}(B_j(\lambda\gamma)_{21}q^{-1}r^{-1}B_k(\lambda\gamma)_{21}q^{-1}r^{-1}). \tag{53}$$

We expand the products

$$\begin{aligned}
&\det_{\omega}^{1/\hbar}((\lambda_{11}\gamma_{11}+\lambda_{12}\gamma_{21})q^{-1}r^{-1})\mathrm{Tr}_{\omega}(A_i(\lambda_{21}\gamma_{11}+\lambda_{22}\gamma_{21})q^{-1}r^{-1})\\
&\quad\times\mathrm{Tr}_{\omega}(B_j(\lambda_{21}\gamma_{11}+\lambda_{22}\gamma_{21})q^{-1}r^{-1}B_k(\lambda_{21}\gamma_{11}+\lambda_{22}\gamma_{21})q^{-1}r^{-1}))\\
&\quad=\det_{\omega}^{1/\hbar}((1+I)(1+rJr^{-1})+\lambda_{12}\gamma_{21}q^{-1}r^{-1})\\
&\qquad\times\mathrm{Tr}_{\omega}(A_i\lambda_{21}(1+I)r^{-1}+A_i\lambda_{22}\gamma_{21}q^{-1}r^{-1})\\
&\qquad\times\mathrm{Tr}_{\omega}[B_j\lambda_{21}(1+I)r^{-1}+B_j\lambda_{22}\gamma_{21}q^{-1}r^{-1}][B_k\lambda_{21}(1+I)r^{-1}+B_k\lambda_{22}\gamma_{21}q^{-1}r^{-1}]\\
&\quad=\det_{\omega}^{1/\hbar}(1+I+J+\lambda_{12}\gamma_{21}q^{-1}r^{-1})\times\mathrm{Tr}_{\omega}(A_i(\lambda_{21}r^{-1}+\lambda_{22}\gamma_{21})q^{-1}r^{-1}\\
&\qquad\times\mathrm{Tr}_{\omega}[B_j\lambda_{21}r^{-1}+B_j\lambda_{22}\gamma_{21}q^{-1}r^{-1}][B_k\lambda_{21}r^{-1}+B_k\lambda_{22}\gamma_{21}q^{-1}r^{-1}],
\end{aligned}$$

where we use $\gamma_{11}q^{-1}=1+I$ and $\lambda_{11}r^{-1}=1+J$. Notice also that we used the vanishing of the Dixmier trace whenever the resulting multiplication is in the trace class operators—and this is the case, for example, for products of the kind $\lambda_{21}I$ and for other similar terms. All the terms in the above expression are actually in the ideal $\mathcal{L}^{(1,\infty)}$ except 1, hence we have the result well defined.

Clearly the action of a group element on a general wave function follows directly from this result. It follows immediately from the above formula that when we restrict to the subspace $\tilde{\mathrm{G}}_{1+}^{(0)}$ for the wave function elements and the group element multiplying from the left, the expression inside the determinant gives us 1. This shows that the representation is trivial.

For completeness we will prove that we have a true representation of the group $\tilde{\mathrm{G}}_{1+}$ on this space of wave functions. This is a technical step and it can be skipped. We will compare the action of the group elements and show that

$$\psi(\lambda(\sigma\gamma),r(sq))=\psi((\lambda\sigma)\gamma,(rs)q), \tag{54}$$

hence when we apply the left multiplication on the left this gives us a representation. We have the group action by (λ,r) given by the above formula that we have already calculated:

$$\det{}_{\omega}^{1/\hbar}(\lambda_{11}r^{-1})\det{}_{\omega}^{1/\hbar}(\gamma_{11}q^{-1}+\lambda_{12}\gamma_{21}q^{-1}r^{-1})\mathrm{Tr}_{\omega}A_i(\lambda_{21}r^{-1}+\lambda_{22}\gamma_{21}q^{-1}r^{-1})$$
$$\times\mathrm{Tr}_{\omega}(B_j(\lambda_{21}r^{-1}+\lambda_{22}\gamma_{21}q^{-1}r^{-1})B_k(\lambda_{21}r^{-1}+\lambda_{22}\gamma_{21}q^{-1}r^{-1})).$$

We can act with the element (σ,s) on this; it is better to break the terms into separate parts:

$$\det{}_{\omega}^{1/\hbar}(\lambda_{11}r^{-1})\det{}_{\omega}^{1/\hbar}((\sigma_{11}\gamma_{11}+\sigma_{12}\gamma_{21})q^{-1}s^{-1}+\lambda_{12}(\sigma_{21}\gamma_{11}+\sigma_{22}\gamma_{21})q^{-1}(rs)^{-1})$$
$$\times\det{}_{\omega}^{1/\hbar}(\lambda_{11}r^{-1})\det{}_{\omega}^{1/\hbar}(\sigma_{11}s^{-1})\det{}_{\omega}^{1/\hbar}(\gamma_{11}q^{-1}+\sigma_{12}\gamma_{21}q^{-1}s^{-1}$$
$$+\lambda_{12}\sigma_{21}(rs)^{-1}+\lambda_{12}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1}).$$

We compare this with the action of the group element $(\lambda\sigma,rs)$ on the same wave function:

$$\det{}_{\omega}^{1/\hbar}((\lambda\sigma)_{11}(rs)^{-1})\det{}_{\omega}^{1/\hbar}(\gamma_{11}q^{-1}+(\lambda\sigma)_{12}\gamma_{21}q^{-1}(rs)^{-1})$$
$$=\det{}_{\omega}^{1/\hbar}(\lambda_{11}\sigma_{11}s^{-1}r^{-1}+\lambda_{12}\sigma_{21}s^{-1}r^{-1})\det{}_{\omega}^{1/\hbar}(\gamma_{11}q^{-1}+\lambda_{11}\sigma_{12}\gamma_{21}q^{-1}s^{-1}r^{-1}$$
$$+\lambda_{12}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1})$$
$$=\det{}_{\omega}^{1/\hbar}(\lambda_{11}r^{-1})\det{}_{\omega}^{1/\hbar}(\sigma_{11}s^{-1})\det{}_{\omega}^{1/\hbar}(\gamma_{11}q^{-1}+\lambda_{12}\sigma_{21}(rs)^{-1}+\sigma_{12}\gamma_{21}q^{-1}s^{-1}$$
$$+\lambda_{12}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1}).$$

These two expressions are the same, hence we have a group action on this part of the wave function.

Let us check the next term:

$$\mathrm{Tr}_{\omega}(A_i(\lambda_{21}r^{-1}+\lambda_{22}(\sigma_{21}\gamma_{11}+\sigma_{22}\gamma_{21})q^{-1}(rs)^{-1}))$$
$$=\mathrm{Tr}_{\omega}(A_i(\lambda_{21}r^{-1}+\lambda_{22}\sigma_{21}(rs)^{-1}+\lambda_{22}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1})).$$

Let us compare this with the direct application of the product:

$$\mathrm{Tr}_{\omega}(A_i((\lambda\sigma)_{21}(rs)^{-1}+(\lambda\sigma)_{22}\gamma_{21}q^{-1}(rs)^{-1}))$$
$$=\mathrm{Tr}_{\omega}(A_i(\lambda_{21}\sigma_{11}s^{-1}r^{-1}++\lambda_{22}\sigma_{21}(rs)^{-1}+\lambda_{21}\sigma_{12}\gamma_{21}q^{-1}(rs)^{-1}+\lambda_{22}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1}))$$
$$=\mathrm{Tr}_{\omega}(A_i(\lambda_{21}r^{-1}+\lambda_{22}\sigma_{21}(rs)^{-1}+\lambda_{22}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1})).$$

These two expressions are the same. Let us look at the last type of term:

$$\begin{aligned}
&\mathrm{Tr}_\omega(B_j(\lambda_{21}r^{-1}+\lambda_{22}(\sigma_{21}\gamma_{11}+\sigma_{22}\gamma_{21})q^{-1}(rs)^{-1})B_k(\lambda_{21}r^{-1}+\lambda_{22}(\sigma_{21}\gamma_{11}+\sigma_{22}\gamma_{21})q^{-1}(rs)^{-1}))\\
&\quad=\mathrm{Tr}_\omega(B_j(\lambda_{21}r^{-1}+\lambda_{22}\sigma_{21}(rs)^{-1}+\lambda_{22}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1})B_k(\lambda_{21}r^{-1}+\lambda_{22}\sigma_{21}(rs)^{-1}\\
&\qquad+\lambda_{22}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1})).
\end{aligned}$$

If we look at the action of the product,

$$\begin{aligned}
&\mathrm{Tr}_\omega(B_j((\lambda\sigma)_{21}(rs)^{-1}+(\lambda\sigma)_{22}\gamma_{21}q^{-1}(rs)^{-1})B_k((\lambda\sigma)_{21}(rs)^{-1}+(\lambda\sigma)_{22}\gamma_{21}q^{-1}(rs)^{-1}))\\
&\quad=\mathrm{Tr}_\omega(B_j(\lambda_{21}r^{-1}+\lambda_{22}\sigma_{21}(rs)^{-1}+(\lambda_{21}\sigma_{12}+\lambda_{22}\sigma_{22})\gamma_{21}q^{-1}(rs)^{-1})B_k(\lambda_{21}r^{-1}\\
&\qquad+\lambda_{22}\sigma_{21}(rs)^{-1}+(\lambda_{21}\sigma_{12}+\lambda_{22}\sigma_{22})\gamma_{21}q^{-1}(rs)^{-1}))\\
&\quad=\mathrm{Tr}_\omega(B_j(\lambda_{21}r^{-1}+\lambda_{22}\sigma_{21}(rs)^{-1}+\lambda_{22}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1})\\
&\qquad\times B_k(\lambda_{21}r^{-1}+\lambda_{22}\sigma_{21}(rs)^{-1}+\lambda_{22}\sigma_{22}\gamma_{21}q^{-1}(rs)^{-1})),
\end{aligned}$$

where we used the vanishing of the Dixmier trace on $\mathcal{L}^1$ and some rearrangements on all the above calculations. We see that this term also respects the group action.

The above representation factors through the subgroup which corresponds to the elements s of $\mathrm{GL}(\mathcal{H}_-)$, with $\det_\omega(s)$ well defined. These are the elements in the subgroup, $s\in 1+\mathcal{L}^{(1,\infty)}$. Notice that for a fixed element γ, the freedom we have to choose different q's which satisfy the determinant condition is isomorphic to $\mathrm{GL}^{1+}=\{q\in \mathrm{GL}(\mathcal{H}_-)|q=1+\mathcal{L}^{(1,\infty)}\}$, although we could not find a way to reduce it to this group everywhere. Hence our representations can actually be reduced to the representations of another group, which is a central extension of the group GL_{1+}. The extension is trivial on the subgroup $\mathrm{GL}_{1+}^{(0)}$; this can be seen by noticing that $s\in 1+(\mathcal{L}^{(1,\infty)})^{(0)}$ will give us $\det_\omega(s)=1$. We can construct a commutative diagram, as we have done in Ref. 2, to show that this gives us a representation of a central extension of the group GL_{1+}. This is more transparent if we look in the Lie algebra level, and also this gives us a chance to compare the central term we have in the case of moment maps. The Lie algebra we need to consider is clearly $\{(u,r)|r\in \mathrm{End}(\mathcal{H}_-)\ \ u_{11}-r\in\mathcal{L}^{(1,\infty)}\}$. One can see that the Lie bracket $[(u,r),(v,s)]=([u,v],[r,s])$ is well defined, and this is the infinitesimal form of the group $\widetilde{\mathrm{G}}_{1+}$. One can see immediately that the set of possible r's is isomorphic to the set of $r\in \mathrm{GL}(\mathcal{H}_-)$ such that $r\in\mathcal{L}^{(1,\infty)}$. We can construct a central extension of the original Lie algebra by using, $(u,r)\mapsto(u,\mathrm{Tr}_\omega(u_{11}-r))$. We give the trivial Lie bracket to the complex numbers. Under this map, the commutator goes to

$$([u,v],\mathrm{Tr}_\omega([u,v]_{11}-[r,s]))=([u,v],\mathrm{Tr}_\omega^\epsilon([\epsilon,u]v)),\tag{55}$$

where we used

$$\begin{aligned}
&\mathrm{Tr}_\omega(u_{11}v_{11}-v_{11}u_{11}+u_{12}v_{21}-v_{12}u_{21}-[r,s])\\
&\quad=\mathrm{Tr}_\omega(u_{11}v_{11}-rs-v_{11}u_{11}+sr)+\mathrm{Tr}_\omega(u_{12}v_{21}-v_{12}u_{21})\\
&\quad=\mathrm{Tr}_\omega((u_{11}-r)v_{11}+r(v_{11}-s))-\mathrm{Tr}_\omega(v_{11}(u_{11}-r)+(v_{11}-s)r)+\mathrm{Tr}_\omega^\epsilon([\epsilon,u]v)\\
&\quad=\mathrm{Tr}_\omega^\epsilon([\epsilon,u]v)
\end{aligned}$$

since all the terms, as grouped, are in $\mathcal{L}^{(1,\infty)}$ and we use the Dixmier trace properties. Note that this is the same central extension in the case of moment maps that we have been using. We will show by an explicit calculation that our representations can be reduced to representations of this algebra. Let us first compute the infinitesimal action; for simplicity we drop the product signs:

$$\mathcal{L}_{(u,r)}\det_\omega^{1/\hbar}(\gamma_{11}q^{-1})\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})\mathrm{Tr}_\omega(B_j\gamma_{21}q^{-1}B_k\gamma_{21}q^{-1}).\tag{56}$$

It is again easier to use the derivation property of the Lie derivative and compute individual terms:

$$\mathcal{L}_{(u,r)}\,\det_\omega^{1/\hbar}(\gamma_{11}q^{-1})=\frac{-1}{\hbar}\mathrm{Tr}_\omega((u_{11}\gamma_{11}+u_{12}\gamma_{21})q^{-1}-\gamma_{11}q^{-1}r)\det_\omega^{1/\hbar}(\gamma_{11}q^{-1})$$

$$=\left(-\frac{1}{\hbar}\mathrm{Tr}_\omega(u_{11}-r)-\frac{1}{\hbar}\mathrm{Tr}_\omega(u_{12}\gamma_{21}q^{-1})\right)\det_\omega^{1/\hbar}(\gamma_{11}q^{-1}).$$

In a similar way we get

$$\mathcal{L}_{(u,r)}\,\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})=\mathrm{Tr}_\omega(A_i(u_{21}+u_{22}\gamma_{21}q^{-1}-\gamma_{21}q^{-1}r)) \tag{57}$$

and

$$\mathcal{L}_{(u,r)}\,\mathrm{Tr}_\omega(B_j\gamma_{21}q^{-1}B_k\gamma_{21}q^{-1})=\mathrm{Tr}_\omega(B_j(u_{21}+u_{22}\gamma_{21}q^{-1}-\gamma_{21}q^{-1}r)B_k(u_{21}+u_{22}\gamma_{21}q^{-1}-\gamma_{21}q^{-1}r)).$$

We define a new representation and, using the above expressions, check that it is, in fact, independent of the choice of the Lie algebra element r:

$$\hat{\mathrm{r}}[(u,r)]=\mathcal{L}_{(u,r)}+\frac{1}{\hbar}\mathrm{Tr}_\omega(u_{11}-r), \tag{58}$$

acting on the same set of wave functions. Using the above expression, we compute the action of the Lie algebra element $(u,r+s)$, where $s\in\mathcal{L}^{(1,\infty)}$ and this is the freedom we have. It is again simpler to check this on each basic piece:

$$(\hat{\mathrm{r}}[(u,r+s)]\det_\omega^{1/\hbar}(\gamma_{11}q^{-1}))\phi(A_i,B_j,B_k,\gamma,q)$$

$$=\left[\mathcal{L}_{(u,r+s)}+\frac{1}{\hbar}\mathrm{Tr}_\omega(u_{11}-(r+s))\right]\det_\omega^{1/\hbar}(\gamma_{11}q^{-1})\phi(A_i,B_j,B_k,\gamma,q)$$

$$=-\frac{1}{\hbar}\mathrm{Tr}_\omega(u_{12}\gamma_{21}q^{-1})\det_\omega^{1/\hbar}(\gamma_{11}q^{-1})\phi(A_i,B_j,B_k,\gamma,q).$$

For the other terms we only use the Lie derivative part since the scalar part is used in the above expression already:

$$\mathcal{L}_{(u,r+s)}\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})=\mathrm{Tr}_\omega(A_i(u_{21}+u_{22}\gamma_{21}q^{-1}-\gamma_{21}q^{-1}(r+s)))$$

$$=\mathrm{Tr}_\omega(A_i(u_{21}+u_{22}\gamma_{21}q^{-1})-A_i\gamma_{21}q^{-1}r)$$

$$=\mathcal{L}_{(u,r)}\,\mathrm{Tr}_\omega(A_i\gamma_{21}q^{-1})$$

by using the fact that, everytime s is multiplied with an element, the resulting term is in the ideal of trace class operators, and the Dixmier trace vanishes on them. The other term,

$$\mathcal{L}_{(u,r+s)}\,\mathrm{Tr}_\omega(B_j\gamma_{21}q^{-1}B_k\gamma_{21}q^{-1}), \tag{59}$$

can be shown to be independent of s by using the same reasoning as above. Hence, we can denote the representation we have as $\hat{\mathbf{r}}(u)$.

If we compute the commutator,

$$(\hat{\mathrm{r}}(u)\hat{\mathrm{r}}(v)-\hat{\mathrm{r}}(v)\hat{\mathrm{r}}(v))\psi(\gamma,q)=\left[\hat{\mathrm{r}}([v,u])-\frac{1}{\hbar}\mathrm{Tr}_\omega([u,v]_{11}-[r,p])\right]\psi(\gamma,q). \tag{60}$$

The last term is independent of the choices of r,p and equal to $-(1/\hbar)\mathrm{Tr}_\omega^\epsilon[\epsilon,u]v$ as we have seen before.

Hence the representations that we have obtained can be reduced to the representations of the central extension of the Lie group $U_{1+}(\mathcal{H})$. This is the quantization of our classical system; it may not be possible to express the central term corresponding to the group in this form, since the extension may have a topological twist in general. We are not able to answer this question, although the discussion in the last section hints that the correct phase space is topologically trivial. This will imply that the central extension actually comes from a central term, globably defined.

IV. FLAT GEOMETRY AND QUANTIZATION

In this section we will introduce a classical system which appears to be unrelated at first sight. This point of view was suggested by Rajeev in our discussions. We consider the set of elements $\overline{Z}$ such that they belong to the following quotient space $\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-)/(\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-))^{(0)}$ [equivalence classes of $Z:\mathcal{H}_+\rightarrow\mathcal{H}_-$ and $Z\in\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-)$ under $Z-Z'\in(\mathcal{L}^{(2,\infty)}(\mathcal{H}_+,\mathcal{H}_-))^{(0)}$]. There is a natural quotient norm,

$$\||\overline{Z}|\|=\inf_{Z_0\in(\mathcal{L}^{(2,\infty)})^{(0)}}\|Z+Z_0\|_{\mathcal{L}^{(2,\infty)}}, \tag{61}$$

where Z is a representative in the equivalence class of $\overline{Z}$.

There is a natural product from $\mathcal{L}^{(2,\infty)}\times\mathcal{L}^{(2,\infty)}\rightarrow\mathcal{L}^{(1,\infty)}$, and this reduces to the quotients $\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}\times\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}\rightarrow\mathcal{L}^{(1,\infty)}/(\mathcal{L}^{(1,\infty)})^{(0)}$ given by $\overline{ZZ'}=\overline{Z}\overline{Z}'$. The natural product $B(\mathcal{H})\times\mathcal{L}^{(2,\infty)}\rightarrow\mathcal{L}^{(2,\infty)}$ also descends to the quotient $B(\mathcal{H})\times(\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)})\rightarrow\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$.

The Dixmier trace is nondegenerate on this quotient space. We will use this important fact to introduce an obvious symplectic form,

$$\tilde{\Omega}_\omega=i\,\mathrm{Tr}_\omega\, d\overline{Z}\wedge d\overline{Z}^\dagger. \tag{62}$$

This flat geometry has a simple symmetry group given by rotations and translations. (This may not be the most general action, but it is the obvious one.) Due to the quotient we can allow for slight deviations from unitary transformations and write a general transformation as

$$\overline{Z}\mapsto\overline{eZf^{-1}+l}, \tag{63}$$

where $e\in GL(\mathcal{H}_-)$, $f\in GL(\mathcal{H}_+)$ such that $e^\dagger e-1$, $f^\dagger f-1\in K(\mathcal{H})$, and $l\in\mathcal{L}^{(2,\infty)}$. One can check that this is, in fact, a group under the obvious composition law, which we call as the affine group, $\mathcal{A}_{1+}$. This action is well defined and transitive. One can immediately check that the group action preserves the symplectic form due to the extra conditions we have:

$$\tilde{\Omega}_\omega=i\,\mathrm{Tr}_\omega\, ed\overline{Z}f^{-1}\wedge f^{-1\dagger}d\overline{Z}^\dagger e^\dagger=i\,\mathrm{Tr}_\omega(e^\dagger e)d\overline{Z}\wedge(f^\dagger f)^{-1}d\overline{Z}^\dagger=\tilde{\Omega}_\omega, \tag{64}$$

using the fact that the product $e^\dagger ed\overline{Z}\approx d\overline{Z}$ and the same for f (see the Appendix for a proof).

It is again natural to find the moment maps generating this action. We can find them using the infinitesimal form of the group action, $\tilde{V}_{(e,f,l)}(\overline{Z})=\overline{\alpha Z}-\overline{Z}\delta+\overline{\beta}$. [Here we denote the Lie algebra elements by the same letter α,β,δ, but they now satisfy $\alpha^\dagger+\alpha=1+K$, $\delta^\dagger+\delta=1+K$ where K is a compact operator, and $\overline{\beta}\in\mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}$, yet we denote the moment maps by $F_{(e,f,l)}$ to imply that they are coming from the affine action. We hope that this does not cause too much confusion.] Hence,

$$F_{(e,f,l)}=i\,\mathrm{Tr}_\omega(\alpha\overline{ZZ}^\dagger-\overline{Z}\delta\overline{Z}^\dagger+\overline{\beta Z}^\dagger+\overline{Z\beta}^\dagger). \tag{65}$$

If we compute $\{F_{(e_1,f_1,l_1)},F_{(e_2,f_2,l_2)}\}$, we will see that there is a central term. Since we will do this calculation below to make connection with the previous section, we postpone the result.

Geometric quantization gives us immediately the following general set of wave functions:

$$\Psi(\overline{Z},\overline{Z}^{\dagger})=e^{-(1/\hbar)\mathrm{Tr}_{\omega}\overline{ZZ^{\dagger}}}\prod_{i,j,k}\mathrm{Tr}_{\omega}(\overline{A_i}\overline{Z})\mathrm{Tr}_{\omega}(B_j\overline{Z}B_k\overline{Z}), \tag{66}$$

where $\overline{A_i}\in\mathcal{L}^{(2,\infty)}$ and $B_{j,k}$ are bounded. Naturally, this set of wave functions carries a representation of the central extension of the above group action, via the same type of operators we have found before:

$$\hat{F}\psi(\overline{Z})=\left(\mathcal{L}_{(\alpha,\delta,\beta)}-\frac{1}{\hbar}\mathrm{Tr}_{\omega}(\overline{\beta}^{\dagger}\overline{Z})\right)\psi(\overline{Z}). \tag{67}$$

(We skip a detailed derivation of this formula, but the reader can verify it by using standard geometric quantization).

This system has an interesting connection to our discussions on the previous section. Let us recall the disc case. One can recover the symplectic form for the disc using the following Kähler potential, $i\mathrm{Tr}_{\omega}\log(1-ZZ^{\dagger})$, just as in the finite-dimensional case. Let us expand the Kähler form, and use the properties of the Dixmier trace. We see that the result is a simple expression: $i\mathrm{Tr}_{\omega}ZZ^{\dagger}$. This is the result for a flat system, except for degeneracies. If we look at the quotient, as above, the result is the same as the Kähler potential of the above system.

We can also apply the quotient homomorphism to our pseudo-unitary group action; this gives

$$\overline{Z}\mapsto\overline{aZd^{-1}}+\overline{bd^{-1}}. \tag{68}$$

Let us show that the group property is preserved under this mapping:

$$\begin{aligned}\overline{g_2\circ(g_1\circ Z)}&=a_2(a_1\overline{Z}d_1^{-1}+\overline{b_1d_1^{-1}})d_2^{-1}+\overline{b_2d_2^{-1}}\\&=a_2a_1\overline{Z}(d_2d_1)^{-1}+a_2\overline{b_1(d_2d_1)^{-1}}\\&=\overline{(a_2a_1+b_2b_1)Z(d_2d_1+c_2c_1)^{-1}}+\overline{(a_2b_1+b_2d_1)(d_2d_1+c_2c_1)^{-1}}\\&=\overline{(g_2g_1)\circ Z}.\end{aligned}$$

This gives us an embedding of $\mathrm{U}_{1+}(\mathcal{H}_-,\mathcal{H}_+)$ into the affine group. Another interesting point is to look at the moment maps, and expand $\mathrm{Tr}_{\omega}u(\Phi-\epsilon)$ in the variable Z, by using the expression of Φ in terms of Z. The properties of the Dixmier trace can be used to see that most of the terms vanish; the result is the same as the moment maps of the flat system:

$$f_u=i\,\mathrm{Tr}_{\omega}(\alpha ZZ^{\dagger}-Z\delta Z^{\dagger}+\beta Z+Z^{\dagger}\beta^{\dagger}). \tag{69}$$

Of course, it is natural to go to the quotient again, and we get $F_u=F_{(a,d,bd^{-1})}$. We can now compute to see the Poisson bracket of these two moment maps, using the flat Poisson bracket:

$$\begin{aligned}\{F_u,F_v\}=i\,\mathrm{Tr}_{\omega}([\alpha_1,\alpha_2]\overline{ZZ^{\dagger}}-\overline{Z}[\delta_1,\delta_2]\overline{Z^{\dagger}}+(\alpha_2\beta_1-\alpha_1\beta_2+\beta_1\delta_2-\beta_2\delta_1)\overline{Z}^{\dagger}+\overline{Z}(\alpha_2\beta_1-\alpha_1\beta_2\\+\beta_1\delta_2-\beta_2\delta_1)^{\dagger})+i\,\mathrm{Tr}_{\omega}(\overline{\beta_1\beta_2^{\dagger}-\beta_2\beta_1^{\dagger}}).\end{aligned}$$

One can verify that the last term is a central term which is equal to the central term we have found before.

The above set of wave functions is equivalent to the wave functions on the disc and they carry the same representation of the central extension of the quotient group. This shows that the system we studied without the reduction can be put into a slightly bigger flat system.

The same question then arises for the Grassmannian. Its coordination will show that in each coordinate neighborhood, the symplectic form is given by the flat one, and similarly the moment functions will look like the flat geometry. Certainly, the quotient point of view, using $U_{1+}/(U_{1+})^{(0)}$, implies that there is a similar simplification. Now we will try to present an alternative point of view in the Grassmannian which keeps the complex structure. Consider the following subgroup:

$$\tilde{G}_{(1+,0)} = \left\{ (g,q) \middle| g \in \begin{pmatrix} \mathcal{B} & \mathcal{L}^{(2,\infty)} \\ (\mathcal{L}^{(2,\infty)})^{(0)} & \mathcal{B} \end{pmatrix} \right\}. \tag{70}$$

This is a closed subgroup, hence the quotient is a holomorphic manifold. Notice that the representation we have introduced for the subgroup $\tilde{B}_{1+}$ actually extends to a representation of this larger group:

$$\det{}_{\omega}^{1/\hbar}((\gamma\lambda)_{11}s^{-1}q^{-1}) = \det{}_{\omega}^{1/\hbar}(\gamma_{11}\lambda_{11}s^{-1}q^{-1} + \gamma_{12}\lambda_{21}s^{-1}q^{-1}). \tag{71}$$

Notice that the last term is actually zero under the Dixmier trace and the rest follows as before, showing that it is a one-dimensional holomorphic representation. The next thing is to check that the remaining part of the wave function is, in fact, invariant under the group $\tilde{G}_{(1+,0)}$. Thus, we will have a line bundle on this smaller quotient, which is the physically relevant phase space. Let us only check one of the terms,

$$\mathrm{Tr}_{\omega}(A_i(\gamma\lambda)_{21}s^{-1}q^{-1}) = \mathrm{Tr}_{\omega}(A_i(\gamma_{21}\lambda_{11}s^{-1}q^{-1} + \gamma_{22}\lambda_{21}s^{-1}q^{-1})) = \mathrm{Tr}_{\omega}(A_i\gamma_{21}q^{-1}),$$

by using the fact that the last term is in $(\mathcal{L}^{(1,\infty)})^{(0)}$ (and similarly for the other type of term). Hence we can consider the set of functions as the sections of a line bundle on the quotient, $\tilde{G}_{1+}/\tilde{G}_{(1+,0)}$. Since we already know that for B_{1+} the quotient cancels out the q parts, for the above form also we get

$$\tilde{G}_{1+}/\tilde{G}_{(1+,0)} \approx GL_{1+}/GL_{(1+,0)} = \begin{pmatrix} \mathcal{B} & \mathcal{L}^{(2,\infty)} \\ \mathcal{L}^{(2,\infty)} & \mathcal{B} \end{pmatrix} \bigg/ \begin{pmatrix} \mathcal{B} & \mathcal{L}^{(2,\infty)} \\ (\mathcal{L}^{(2,\infty)})^{(0)} & \mathcal{B} \end{pmatrix} \approx \mathcal{L}^{(2,\infty)}/(\mathcal{L}^{(2,\infty)})^{(0)}. \tag{72}$$

This shows that the relevant part of the phase spaces for the disc and the Grassmannian are of equal size. We are not able to provide a link with this and the coordinate description at the present moment. Our guess is that Grassmannian also has the same embedding into a flat system; this manifests the possible equivalence of the two systems. We hope to clarify some of these issues in a future publication.

ACKNOWLEDGMENTS

Most of this paper owes a lot to several conversations with S. G. Rajeev, and shows his guidance and influence. We thank J. Mickelsson for reading and suggesting corrections, M. Wodzicki for a series of marvelous lectures on traces, and E. Langmann for perusing through and several comments. We also would like to thank J. Brodzki, P. Bongaart, K. Gawedzki, G. Landi, and M. Walze, for many useful discussions. The author gratefully acknowledges the hospitality of S. Majid and his group, and the kind invitation of P. Goddard and D. Crighton to DAMTP, where most of this work was done during the author's stay in Cambridge. During this project the author was supported by the EPDI grant, and we thank IHES for the excellent help and working environment provided.

APPENDIX: MISCELLANEOUS RESULTS

The definitions of the operator ideals will be given. Let us start with the definition of $\mathcal{L}^{(2,\infty)}$. Operator ideals contain compact operators, thus they are given by the summability properties of the singular values of the operators. If $\mu_n(A) = n$th eigenvalue of $|A|$, then we define a new norm:

$$\|A\|_{\mathcal{L}^{(2,\infty)}}=\sup_N\frac{\Sigma_{n=1}^N\mu_n(A)}{\Sigma_{i=1}^N 1/n^{1/2}}. \tag{A1}$$

The set of all $A\in K(\mathcal{H})$ for which the above norm is finite is denoted by $\mathcal{L}^{(2,\infty)}$. It is a symmetrically normed ideal.[6] Since the sequence $1/n^{1/2}$ is regular, the same ideal can be defined through the asymptotic behavior of the singular values. In fact, the set of operators in $\mathcal{L}^{(2,\infty)}$ can be defined as $A\in K(\mathcal{H})$ such that $\mu_n(A)=O(n^{-1/2})$. This also gives a simple characterization of the completion of the finite rank operators inside $\mathcal{L}^{(2,\infty)}$, denoted as $(\mathcal{L}^{(2,\infty)})^{(0)}$; $A\in(\mathcal{L}^{(2,\infty)})^{(0)}$ iff $\mu_n(A)=o(n^{-1/2})$. [The symbol $\mu_n=O(\pi_n)$ means that, $\limsup_{n\to\infty}\mu_n/\pi_n<\infty$ and $\mu_n=o(\pi_n)$ iff $\lim_{n\to\infty}\mu_n/\pi_n=0$.] One can define the norm for $\mathcal{L}^{(1,\infty)}$ in the same way replacing the sequence $1/n^{1/2}$ by $1/n$. This is not a regular sequence, so the completions of finite rank operators are given by the behavior of the partial sums; $\sigma_N(A)=\Sigma_{n=1}^N\mu_n(A)$, $\sigma_N(A)=O(\log N)$. If $\mu_n(B)=o(1/n)$, then it implies that $\sigma_N(B)=o(\log N)$, hence $B\in(\mathcal{L}^{(1,\infty)})^{(0)}$. However, the converse is not true.

We give a proof that $A,B\in\mathcal{L}^{(1,\infty)}$, then $AB\in\mathcal{L}^1$. Let us assume that the hypothesis is true. It implies that $A,B\in\mathcal{L}^2$ as well. However, we know that $\mathcal{L}^2\mathcal{L}^2\in\mathcal{L}^1$, hence the result.

Next, we will prove that if $A\in\mathcal{L}^{(2,\infty)}$ and $B\in(\mathcal{L}^{(2,\infty)})^{(0)}$, then $AB\in(\mathcal{L}^{(1,\infty)})^{(0)}$.

We will use the inequalities satisfied by the singular values:

$$\mu_{n+m}(AB)\leq\mu_n(A)\mu_m(B). \tag{A2}$$

Choose $n+m=2N+j$, where $j=0,1$ and look at the following limit:

$$\limsup_{N\to\infty}(2N+j)\mu_{2N+j}(AB)\leq\limsup_{N\to\infty}(2N+j)^{1/2}\mu_N(A)(2N+j)^{1/2}\mu_{N+j}(B). \tag{A3}$$

Now we can use

$$\lim_{N\to\infty}(N^{1/2}\mu_N(A))=a<\infty \quad\text{and}\quad \lim_{N\to\infty}((N^{1/2}\mu_N(B))=0 \tag{A4}$$

in the above expression to get

$$\lim_{N\to\infty}(2N+j)\mu_{2N+j}(AB)=0, \tag{A5}$$

and this implies that the product is in the closure, $AB\in(\mathcal{L}^{(1,\infty)})^{(0)}$. One can imitate the above proof to show that $A,B\in\mathcal{L}^{(2,\infty)}$, then $AB\in\mathcal{L}^{(1,\infty)}$; we leave this to the reader.

Let us use the same idea to show that $K(\mathcal{H})\mathcal{L}^{(2,\infty)}\subset(\mathcal{L}^{(2,\infty)})^{(0)}$:

$$\limsup_{2N+j}(2N+j)^{1/2}\mu_{2N+j}(AK)\leq\limsup_{2N+j}(2N+j)^{1/2}\mu_N(A)\mu_{N+j}(K)=0 \tag{A6}$$

by using the fact that $\lim_N\mu_N(K)=0$ for a compact operator K.

In the second part we will prove some of the properties of the conditional Dixmier traces, which are identical to the usual trace conditions. We define the conditional trace as

$$\mathrm{Tr}_\omega^\epsilon A=\mathrm{Tr}_\omega^\epsilon\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}=\frac{1}{2}(\mathrm{Tr}_\omega(a_{11})+\mathrm{Tr}_\omega(a_{22}))=\frac{1}{4}\mathrm{Tr}_\omega[A+\epsilon A\epsilon] \tag{A7}$$

for $A=\left(\begin{smallmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{smallmatrix}\right)$. Notice that we have absorbed a factor of $\frac{1}{2}$ into the definition to make the formulas involving this trace look simpler. First property,

$$\mathrm{Tr}_\omega^\epsilon AB=\mathrm{Tr}_\omega^\epsilon BA \tag{A8}$$

if all the individual terms in the products $\Sigma_k a_{ik} b_{ki}$ are in the ideal $\mathcal{L}^{(1,\infty)}$. This is easy to see if we use the definition of the conditional trace, and

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}=\begin{pmatrix} a_{11}b_{11}+a_{12}b_{21} & * \\ * & a_{21}b_{12}+a_{22}b_{22} \end{pmatrix}. \tag{A9}$$

If each of the terms appearing in the diagonal parts is independently in $\mathcal{L}^{(1,\infty)}$, we can use $\mathrm{Tr}_\omega\, ab=\mathrm{Tr}_\omega\, ba$ and see that the result is the same when one takes the product in the opposite order.

We will show that the group representation is satisfied up to a central term. It is more convenient to use $\det_\omega$ instead of e^{Tr_ω} and the power $1/\hbar$ is not written since it is easy to put back. We compare the two ways of applying the representation, $\rho(g_2^{-1})\rho(g_1^{-1})\Psi(Z)$ and $\rho((g_1g_2)^{-1})\Psi(Z)$. The first expression gives

$$\det{}_\omega(d_1^{-1}c_1(a_2Z+b_2)(c_2Z+d_2)^{-1}+1)\det{}_\omega(d_2^{-1}c_2Z+1)\Psi((g_1g_2)\circ Z), \tag{A10}$$

and the second

$$\det{}_\omega((d_1d_2+c_1b_2)^{-1}(c_1a_2+d_1c_2)Z+1)\Psi((g_1g_2)^{-1}\circ Z). \tag{A11}$$

It is enough to compare the ''determinant'' pieces because the other parts are the same. Let us check the following:

$$\begin{aligned}
&\det{}_\omega(d_1^{-1}c_1(a_2Z+b_2)(c_2Z+d_2)^{-1}+1)\det{}_\omega(d_2^{-1}c_2Z+1)\\
&\quad=\det{}_\omega(d_2^{-1}d_1^{-1}c_1(a_2Z+b_2)(d_2^{-1}c_2Z+1)^{-1}+1)\det{}_\omega(d_2^{-1}c_2Z+1)\\
&\quad=\det{}_\omega((d_1d_2)^{-1}c_1(a_2Z+b_2)(d_2^{-1}c_2Z+1)^{-1}(d_2^{-1}c_2Z+1)+d_2^{-1}c_2Z+1)\\
&\quad=\det{}_\omega((d_1d_2)^{-1}c_1(a_2Z+b_2)+d_2^{-1}c_2Z+1)\\
&\quad=\det{}_\omega((d_1d_2)^{-1}(c_1a_2+d_1c_2)Z+(d_1d_2)^{-1}c_1b_2+1).
\end{aligned}$$

We used the multiplicative property of the $\det_\omega$, which comes from the properties of the Dixmier trace. The equalities are true by adding terms which give zero under the Dixmier trace; this is the advantage of using the symbol $\det_\omega$. Let us compare this with

$$\begin{aligned}
&\det{}_\omega((d_1d_2+c_1b_2)^{-1}(c_1a_2+d_1c_2)Z+1)\\
&\quad=\det{}_\omega[((d_1d_2)^{-1}c_1b_2+1)^{-1}(d_1d_2)^{-1}(c_1a_2+d_1c_2)Z+1]\\
&\quad=\det{}_\omega((d_1d_2)^{-1}c_1b_2+1)^{-1}\det{}_\omega((d_1d_2)^{-1}(c_1a_2+d_1c_2)Z+(d_1d_2)^{-1}c_1b_2+1),
\end{aligned}$$

hence they differ by a constant multiple which never vanishes,

$$c_\omega(g_1,g_2)=\det{}_\omega((d_1d_2)^{-1}c_1b_2+1). \tag{A12}$$

This trace is well defined as one can check, and, since it never vanishes, the two sides are equal. We need to further check that it obeys the cocyle condition,

$$\begin{aligned}
c_\omega(g_1g_2,g_3)c_\omega(g_1,g_2)&=c_\omega(g_1,g_2g_3)c_\omega(g_2,g_3)\\
&=\det{}^{1/\hbar}[(d_1d_2d_3)^{-1}c_1a_2b_3+(d_2d_3)^{-1}c_2b_3+(d_1d_2)^{-1}c_1b_2+1].
\end{aligned} \tag{A13}$$

The sum of all the terms except 1 inside the determinant sign are in the ideal $\mathcal{L}^{(1,\infty)}$, hence the Dixmier trace is well defined. The cocycle c_ω, in the finite-dimensional case, can be obtained from $\phi(g)=\det(d)$, as $c(g_1,g_2)=\phi(g_1)\phi(g_2)\phi(g_1g_2)^{-1}$. Clearly it is not well defined in infinite dimensions; in fact, the extension is nontrivial. Thus, we obtain a representation of a central extension $\hat{\mathrm{U}}_{1+}(\mathcal{H}_-,\mathcal{H}_+)$ in the quantum Hilbert space of holomorphic sections.

[1] S. G. Rajeev, Int. J. Mod. Phys. A **9**, 5583 (1994).
[2] S. G. Rajeev and O. T. Turgut, Commun. Math. Phys. **192**, 493 (1998).
[3] A. Pressley and G. Segal, *Loop Groups* (Oxford U.P., Oxford, 1988).
[4] J. Mickelsson and S. G. Rajeev, Commun. Math. Phys. **116**, 365 (1988); J. Mickelsson, *ibid.* **154**, 403 (1993).
[5] E. Langmann, Commun. Math. Phys. **162**, 1 (1994).
[6] I. C. Gohberg and M. G. Krein, *Introduction to The Theory of Nonself-adjoint Linear Operators*, Trans. Am. Math. Soc. 1968.
[7] A. Connes, *Non-commutative geometry; spectral aspects*, les houches lectures, 1995; *Non-commutative Geometry* (Academic, New York, 1994).
[8] A. Connes, Commun Math. Phys. **117**, 673 (1988).
[9] M. Kontsevich and S. Vishik, conference proceeding, hep-th/9405035.
[10] J. Mickelsson, Talk in baltic school in field theory, hep-th/9404093; E. Langmann and J. Mickelsson, Lett. Math. Phys. **36**, 45 (1996).
[11] M. Cederwall, G. Ferretti, B. E. W. Nilsson, and A. Westerberg, Commun. Math. Phys. **175**, 203 (1996).
[12] K. J. Dykema, T. Figiel, G. Weiss, and M. Wodzicki, ''The commutator structure of operator ideals,'' preprint.
[13] M. Wodzicki, Invent. Math. **75**, 143 (1984).
[14] B. Fedosov, F. Golse, E. Leichtnam, and E. Shrohe, J. Funct. Anal. **142**, 1 (1996).
[15] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed. (Benjamin/Cummings, New York, 1978).
[16] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed. (Springer Verlag, New York, 1986).
[17] S. Axelrod, S. D. Pietra, and E. Witten, J. Diff. Geom. **33**, 787 (1991).
[18] N. E. Hurt, *Geometric Quantization in Action* (Reidel, Dordrecht, 1983).
[19] A. Kirillov, ''Geometric Quantization,'' in *Dynamical Systems IV*, edited by V. I. Arnold (Springer-Verlag, New York, 1988).
[20] F. A. Berezin, Commun. Math. Phys. **40**, 153 (1975).
[21] B. Simon, *Trace Ideals and Their Applications* (Cambridge U.P., Cambridge, 1979).
[22] D. Pickrell, J. Funct. Anal. **70**, 323 (1987).

Gauge theory of *w* symmetry

Wei-zhong Zhao[a)]
Institute of Mathematics and Physics, Xinjiang University, Urumqi, Xinjiang 830046, People's Republic of China, and Department of Physics, Xinjiang University, Urumqi, Xinjiang 830046, People's Republic of China

(Received 4 February 1999; accepted for publication 13 May 1999)

The w_∞ algebra is a higher-spin extension of the Virasoro algebra. In this paper, we construct the gauge theory of *w* symmetry in terms of its representations. [S0022-2488(99)03508-2]

I. INTRODUCTION

The Virasoro algebra and its extensions play a fundamental role in the study of two-dimensional conformal field theories. The w_∞ algebra[1] is a higher-spin extension of the Virasoro algebra. It can be regarded as the algebra of smooth area-preserving diffeomorphisms of the cylinder. It is known that two-dimension gravity can be thought of as the result of gauging the Virasoro algebra. *W* gravity is the gauge theory of local *W* algebra symmetry. The study of *W* gravity has been guided by the analogy with two-dimension gravity. The w_3 gravity was first constructed in the *W* gravity.[2] It was begun from a free action realizing a chiral w_3 symmetry, and Noether-coupled it to spin-2 and spin-3 background gauge fields. Schoutens *et al.*[3] extended this work to a nonchiral gauged theory of w_3. In terms of these methods, chiral and nonchiral w_∞ gravity was constructed by Bergshoeff *et al.*[4] We know that it is difficult to construct a gauge theory of the Virasoro symmetry through the conventional method. The reason is that the Virasoro group is not only noncompact, but also does not admit any bi-invariant Cartan–Killing metric. For the Virasoro group, the representation that we know well is the highest-weight representation. But this representation is of no use to construct the gauge theory of the Virasoro group because the gauge field must form an adjoint representation. One must consider the other representation. In order to construct the wanted gauge theory, Cho generalized the Kaplansky–Feigin–Fuks (KFF) representation of the Virasoro algebra[5,6] and constructed the invariant tensors. With all of these elements, he gave a gauge theory of the Virasoro group.[7,8]

As a higher-spin extension of the Virasoro algebra, the generators of the w_∞ algebra can be expressed linearly in terms of the generators of the ''Virasoro-like'' algebras. From the representations of the ''Virasoro-like'' algebras, we can obtain the representations of the w_∞ algebra that we need for the gauge theory. In this paper, we will construct the gauge theory of *w* symmetry from its representations.

II. w_∞ ALGEBRA AND ITS REPRESENTATIONS

The w_∞ algebra is an infinite-dimensional algebra. Its commutation relation is given by[4]

$$[v_{m_1,m_2},v_{n_1,n_2}]\equiv[v_{\bar{m}},v_{\bar{n}}]=f_{\bar{m}\bar{n}}^{\bar{k}}v_{\bar{k}}=[(n_2+1)m_1-(m_2+1)n_1]v_{\bar{m}+\bar{n}}, \tag{1}$$

where

$$\bar{m}=(m_1,m_2). \tag{2}$$

The generators are given on a two-dimensional space (x,y) as follows:

[a)]Electronic mail: Zhaowz100@263.net

$$v_{\overline{m}}=e^{(im_1x)}[m_1y^{m_2+1}\partial_y+i(m_2+1)y^{m_2}\partial_x], \tag{3}$$

where m_2 is the index of the conformal spins satisfying $m_2+2\geqslant 2$ and $m_1\in Z$. The generators $v_{m,0}$ are just the Virasoro generators L_m. If we take $z=e^{ix}$, then the generators become

$$v_{\overline{m}}=(m_2+1)L^{1,0}_{m_1,m_2}-m_1L^{0,1}_{m_1,m_2}, \tag{4}$$

where

$$L^{1,0}_{m_1,m_2}=-z^{m_1+1}y^{m_2}\frac{d}{dz}, \tag{5}$$

and

$$L^{0,1}_{m_1,m_2}=-z^{m_1}y^{m_2+1}\frac{d}{dy}. \tag{6}$$

The generators $L^{1,0}_{m_1,m_2}$ and $L^{0,1}_{m_1,m_2}$ satisfy the ''Virasoro-like'' algebra, respectively, i.e.,

$$[L^{1,0}_{m_1,m_2},L^{1,0}_{n_1,n_2}]=(m_1-n_1)L^{1,0}_{m_1+n_1,m_2+n_2}, \tag{7}$$

$$[L^{0,1}_{m_1,m_2},L^{0,1}_{n_1,n_2}]=(m_2-n_2)L^{0,1}_{m_1+n_1,m_2+n_2}. \tag{8}$$

From the above commutation relations, we obtain the following adjoint representations of the ''Virasoro-like'' algebras:

$$(L^{1,0}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}=(m_1-n_1)\,\delta^{\overline{k}}_{\overline{m}+\overline{n}}, \tag{9}$$

$$(L^{0,1}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}=(m_2-n_2)\,\delta^{\overline{k}}_{\overline{m}+\overline{n}}, \tag{10}$$

where

$$\delta^{\overline{k}}_{\overline{m}+\overline{n}}=\delta^{k_1}_{m_1+n_1}\delta^{k_2}_{m_2+n_2}. \tag{11}$$

As the case of the Virasoro algebra,[7] we introduce the following KFF representations of the ''Virasoro-like'' algebras:

$$(L^{1,0}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}=[(\alpha+1)m_1+\beta-k_1]\,\delta^{\overline{k}}_{\overline{m}+\overline{n}}, \tag{12}$$

$$(L^{0,1}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}=[(\alpha+1)m_2-k_2]\,\delta^{\overline{k}}_{\overline{m}+\overline{n}}, \tag{13}$$

where α and β are arbitrary complex numbers. One can easily show that the KFF representations (12), (13) satisfy the following relations:

$$\begin{aligned}([L^{1,0}_{m_1,m_2},,L^{1,0}_{n_1,n_2}])^{k_1,k_2}_{l_1,l_2}&=(L^{1,0}_{m_1,m_2})^{k_1,k_2}_{h_1,h_2}(L^{1,0}_{n_1,n_2})^{h_1,h_2}_{l_1,l_2}-(L^{1,0}_{n_1,n_2})^{k_1,k_2}_{h_1,h_2}(L^{1,0}_{m_1,m_2})^{h_1,h_2}_{l_1,l_2}\\&=(m_1-n_1)(L^{1,0}_{m_1+n_1,m_2+n_2})^{k_1,k_2}_{l_1,l_2},\end{aligned} \tag{14}$$

$$([L^{0,1}_{m_1,m_2}, ,L^{0,1}_{n_1,n_2}])^{k_1,k_2}_{l_1,l_2}=(L^{0,1}_{m_1,m_2})^{k_1,k_2}_{h_1,h_2}(L^{0,1}_{n_1,n_2})^{h_1,h_2}_{l_1,l_2}-(L^{0,1}_{n_1,n_2})^{k_1,k_2}_{h_1,h_2}(L^{0,1}_{m_1,m_2})^{h_1,h_2}_{l_1,l_2}$$
$$=(m_2-n_2)(L^{0,1}_{m_1+n_1,m_2+n_2})^{k_1,k_2}_{l_1,l_2}. \tag{15}$$

In fact, the adjoint representations (9), (10) correspond to the $(\alpha=1,\ \beta=0)$ representation of the KFF representations (12), (13), respectively. From the adjoint representations of the ''Virasoro-like'' algebras, we have

$$(v_{\bar{m}})^{\bar{k}}_{\bar{n}}=(m_2+1)(L^{1,0}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}-m_1(L^{0,1}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}=[(n_2+1)m_1-(m_2+1)n_1]\,\delta^{\bar{k}}_{\bar{m}+\bar{n}}. \tag{16}$$

It is obvious that the above representation is the adjoint representation of the w_∞ algebra. It acts on an infinite-dimensional vector $\phi^{\bar{k}}$ as

$$(v_{\bar{m}}\phi)^{\bar{k}}=-f^{\bar{k}}_{\bar{m}\bar{n}}\phi^{\bar{n}}=-[(k_2+2)m_1-(m_2+1)k_1]\,\phi^{\bar{k}-\bar{m}}. \tag{17}$$

Using the KFF representations (12), (13), we have the KFF representation of the *w* symmetry,

$$(v_{\bar{m}})^{\bar{k}}_{\bar{n}}=\{(m_2+1)[(\alpha+1)m_1+\beta-k_1]-m_1[(\alpha+1)m_2-k_2]\}\,\delta^{\bar{k}}_{\bar{m}+\bar{n}}. \tag{18}$$

One can show that the above representation satisfies the relation

$$([v_{\bar{m}},v_{\bar{n}}])^{\bar{k}}_{\bar{l}}=(v_{\bar{m}})^{\bar{k}}_{\bar{h}}(v_{\bar{n}})^{\bar{h}}_{\bar{l}}-(v_{\bar{n}})^{\bar{k}}_{\bar{h}}(v_{\bar{m}})^{\bar{h}}_{\bar{l}}=[(n_2+1)m_1-(m_2+1)n_1](v_{\bar{m}+\bar{n}})^{\bar{k}}_{\bar{l}}. \tag{19}$$

If we take $m_2=n_2=k_2=0$, then the representation (18) gives the KFF representation of the Virasoro algebra.[7] This is the reason that we require $\beta=0$ in Eq. (13). The representation (18) acts on an infinite-dimensional vector space $V_{(\alpha,\beta)}$ as

$$(v_{\bar{m}}\phi)^{\bar{k}}=-(v_{\bar{m}})^{\bar{k}}_{\bar{n}}\phi^{\bar{n}}=-\{(m_2+1)[(\alpha+1)m_1+\beta-k_1]-m_1[(\alpha+1)m_2-k_2]\}\,\phi^{\bar{k}-\bar{m}}, \tag{20}$$

where $\phi^{\bar{k}}$ is an element of $V_{(\alpha,\beta)}$. We introduce the dual representation of the KFF representation of the *w* symmetry that acts on the vector space $V^*_{\alpha,\beta}$ dual to $V_{\alpha,\beta}$ as follows:

$$(v_{\bar{m}}\omega)_{\bar{k}}=(v_{\bar{m}})^{\bar{n}}_{\bar{k}}\omega_{\bar{n}}=[(m_2+1)(\alpha m_1+\beta-k_1)-m_1(\alpha m_2-k_2)]\,\omega_{\bar{k}+\bar{m}}, \tag{21}$$

where $w_{\bar{k}}$ is an element of $V^*_{\alpha,\beta}$. In terms of the dual representation, we have the following contraction between $\phi^{\bar{k}}$ and $w_{\bar{k}}$:

$$\phi\cdot\omega=\phi^{\bar{k}}\omega_{\bar{k}}, \tag{22}$$

which is invariant under the *w* transformation. From Eqs. (20) and (21), it is noted that there is an isomorphism $V^*_{\alpha,\beta}\cong V_{-1-\alpha,-\beta}$ if $\omega_{\bar{k}}=\phi^{-\bar{k}}$. We can generalize the KFF representation of the *w* symmetry to a $(\bar{p},\bar{q})$ tensor $T_{(\bar{p},\bar{q})}$ and define the tensor representation by

$$v_{\bar{m}}t^{\bar{i}\cdots\bar{j}}_{\bar{k}\cdots\bar{l}}=-(v_{\bar{m}})^{\bar{i}}_{\bar{n}}t^{\bar{n}\cdots\bar{j}}_{\bar{k}\cdots\bar{l}}-\cdots-(v_{\bar{m}})^{\bar{j}}_{\bar{n}}t^{\bar{i}\cdots\bar{n}}_{\bar{k}\cdots\bar{l}}+(v_{\bar{m}})^{\bar{n}}_{\bar{k}}t^{\bar{i}\cdots\bar{j}}_{\bar{n}\cdots\bar{l}}+\cdots+(v_{\bar{m}})^{\bar{n}}_{\bar{l}}t^{\bar{i}\cdots\bar{j}}_{\bar{k}\cdots\bar{n}}, \tag{23}$$

where $t^{\bar{i}\cdots\bar{j}}_{\bar{k}\cdots\bar{l}}$ is an infinite-dimensional $(\bar{p},\bar{q})$ tensor, and define a tensor module of mixed type $\hat{T}_{p,q}$ by allowing each of the $p+q$ indices to transform according to a different (α,β) representation. For example, we can define a (0,2) tensor of mixed type $\hat{t}_{\bar{n}\bar{k}}$, which transforms as

$$(v_{\bar{m}}\hat{t})_{\bar{n}\bar{k}}=[(m_2+1)(\alpha m_1+\beta-n_1)-m_1(\alpha m_2-n_2)]\hat{t}_{\bar{n}+\bar{m},\bar{k}}+[(m_2+1)(\alpha' m_1+\beta'-k_1)$$
$$-m_1(\alpha' m_2-k_2)]\hat{t}_{\bar{n},\bar{k}+\bar{m}}. \tag{24}$$

In terms of these generalizations, we have the invariant tensors $\hat{d}^{\bar{k}}_{\bar{i}\bar{j}\cdots\bar{l}}$, $\hat{f}^{\bar{k}}_{\bar{i}\bar{j}}$ and $g_{\bar{i}\bar{j}}$, where

$$\hat{d}^{\bar{k}}_{\bar{i}\bar{j}\cdots\bar{l}}=\delta^{\bar{k}}_{\bar{i}+\bar{j}+\cdots\bar{l}}. \tag{25}$$

The indices $\overline{i,j,\ldots,l,k}$ belong to (α_1,β_1), $(\alpha_2,\beta_2),\ldots$, (α_n,β_n) and $(\alpha_1+\alpha_2+\cdots+\alpha_n,\beta_1+\beta_2+\cdots+\beta_n)$ representations, respectively,

$$\hat{f}^{\bar{k}}_{\bar{i}\bar{j}}=(v_{\bar{i}})^{\bar{k}}_{\bar{j}}=\{(i_2+1)[(\alpha+1)i_1+\beta-k_1]-i_1[(\alpha+1)i_2+\beta-k_2]\}\delta^{\bar{k}}_{\bar{i}+\bar{j}}. \tag{26}$$

The indices $\overline{i,j,k}$ belong to $(\alpha=1,\ \beta=0)$, (α,β) and (α,β) representations, respectively. The metric $g_{\bar{i}\bar{j}}=\delta^0_{\bar{i}+\bar{j}}$ is an invariant tensor of the $(\alpha=-\frac{1}{2},\ \beta=0)$ representation. The invariant tensors $\hat{d}^{\bar{k}}_{\bar{i}\bar{j}},\hat{f}^{\bar{k}}_{\bar{i}\bar{j}}$ can be used to define the vector products. The invariant metric $g_{\bar{i}\bar{j}}$ allow us to define the index raising of the tensors and the scalar product. For example, with the invariant metric, one can obtain the index raising of the tensor,

$$\phi_{\bar{k}}=g_{\bar{k}\bar{l}}\phi^{\bar{l}}=\delta^0_{\bar{k}+\bar{l}}\phi^{\bar{l}}=\phi^{-\bar{k}}, \tag{27}$$

and the scalar product

$$\phi_{\bar{k}}\phi^{\bar{k}}=g_{\bar{k}\bar{l}}\phi^{\bar{l}}\phi^{\bar{k}}=\sum_{\bar{k}}\phi^{-\bar{k}}\phi^{\bar{k}}\equiv\phi^{-\bar{k}}\phi^{\bar{k}}. \tag{28}$$

With the invariant tensor $\hat{d}^{\bar{k}}_{\bar{i}\bar{j}\cdots\bar{l}}$, one can define the nth power $(\phi^n)^{\bar{k}}$ by

$$(\phi^n)^{\bar{k}}=\hat{d}^{\bar{k}}_{\bar{k}_1\bar{k}_2\cdots\bar{k}_n}\phi^{\bar{k}_1}\phi^{\bar{k}_2}\cdots\phi^{\bar{k}_n}=\sum_{\bar{k}_1+\cdots+\bar{k}_n=\bar{k}}\phi^{\bar{k}_1}\phi^{\bar{k}_2}\cdots\phi^{\bar{k}_n}. \tag{29}$$

These invariant tensors are very useful in our gauge theory. The concepts of hermiticity and the unitarity of the representation take the important roles in the gauge theory. We call a tensor field $t^{\bar{i}\cdots\bar{j}}_{\bar{k}\cdots\bar{l}}$ Hermitian[8] if

$$(t^{\bar{i}\cdots\bar{j}}_{\bar{k}\cdots\bar{l}})^*=t^{-\bar{i}\cdots-\bar{j}}_{-\bar{k}\cdots-\bar{l}}. \tag{30}$$

As the case of the Virasoro symmetry,[8] we can introduce the following complex conjugate KFF representations of the ‘‘Virasoro-like’’ algebras:

$$(\overline{L}^{1,0}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}=[(\alpha^*+1)m_1-\beta^*+k_1]\delta^{\bar{k}}_{\bar{n}-\bar{m}}, \tag{31}$$

and

$$(\overline{L}^{0,1}_{m_1,m_2})^{k_1,k_2}_{n_1,n_2}=[(\alpha^*+1)m_2+k_2]\delta^{\bar{k}}_{\bar{n}-\bar{m}}. \tag{32}$$

With the above complex conjugate representations, we have the conjugate KFF representation of w symmetry,

$$(\overline{v_{\bar m}})_{\bar n}^{\bar k} = (m_2+1)(\overline{L}_{m_1,m_2}^{1,0})_{n_1,n_2}^{k_1,k_2} - m_1(\overline{L}_{m_1,m_2}^{0,1})_{n_1,n_2}^{k_1,k_2}$$

$$= \{(m_2+1)[(\alpha^*+1)m_1 - \beta^* + k_1] - m_1[(\alpha^*+1)m_2+k_2]\}\delta_{\bar n - \bar m}^{\bar k}, \tag{33}$$

which acts on an infinite-dimensional complex conjugate vector space $\overline{V}_{\alpha,\beta}$ as

$$v_{\bar m}(\phi^{\bar k})^* = -(\overline{v_{\bar m}})_{\bar n}^{\bar k}(\phi^{\bar n})^*$$

$$= -\{(m_2+1)[(\alpha^*+1)m_1 - \beta^* + k_1] - m_1[(\alpha^*+1)m_2+k_2]\}(\phi^{\bar k + \bar m})^*, \tag{34}$$

where $(\phi^{\bar k})^*$ is an element of $\overline{V}_{\alpha,\beta}$.

Let us consider the following Hermitian product (φ,ϕ), defined by

$$(\varphi,\phi) = (\varphi^{\bar k})^*\phi^{\bar k}; \tag{35}$$

then we have

$$v_{\bar m}(\varphi,\phi) = -(\overline{v_{\bar m}})_{\bar n}^{\bar k}(\varphi^{\bar n})^*\phi^{\bar k} - (\varphi^{\bar k})^*(v_{\bar m})_{\bar n}^{\bar k}\phi^{\bar n}$$

$$= -[(\alpha+\alpha^*+1)m_1 + (m_2+1)(\beta-\beta^*)](\varphi^{\bar k})^*\phi^{\bar k - \bar m}. \tag{36}$$

So the Hermitian product becomes invariant under *w* symmetry if and only if

$$\alpha+\alpha^*+1=0, \tag{37}$$

and

$$\beta-\beta^*=0. \tag{38}$$

We call an (α,β) representation unitary if the above condition is satisfied. Clearly, the $(\alpha=-\frac{1}{2}, \beta=0)$ representation is unitary. This representation is important in the following gauge theory.

III. THE GAUGE THEORY OF *W* SYMMETRY

Let $A_\mu^{\bar k}$ be the Hermitian gauge potential that forms an adjoint representation, the matter field $\phi^{\bar k}$ is a Hermitian scalar multiplet that forms a $(\alpha=-\frac{1}{2}, \beta=0)$ representation, and the infinitesimal gauge parameter of *w* symmetry forms an adjoint representation. Under the infinitesimal gauge transformation, we require

$$\delta A_\mu^{\bar k} = -\frac{1}{g}[\partial_\mu\theta^{\bar k} + ig(A_\mu\times\theta)^{\bar k}] = -\frac{1}{g}\{\partial_\mu\theta^{\bar k} + ig[m_1(k_2+2) - k_1(m_2+1)]A_\mu^{\bar m}\theta^{\bar k - \bar m}\}, \tag{39}$$

$$\delta\phi^{\bar k} = i(\theta\times\phi)^{\bar k} = i\hat f_{\bar m \bar n}^{\bar k}\theta^{\bar m}\phi^{\bar n} = i[(m_2+1)(\tfrac{1}{2}m_1 - k_1) - m_1(\tfrac{1}{2}m_2 - k_2)]\theta^{\bar m}\phi^{\bar k - \bar m}. \tag{40}$$

The field strength $F_{\mu\nu}^{\bar k}$ and the covariant derivative D_μ are, respectively, defined by

$$F_{\mu\nu}^{\bar k} = \partial_\mu A_\nu^{\bar k} - \partial_\nu A_\mu^{\bar k} + ig(A_\mu\times A_\nu)^{\bar k} = \partial_\mu A_\nu^{\bar k} - \partial_\nu A_\mu^{\bar k} + ig[m_1(k_2+2) - k_1(m_2+1)]A_\mu^{\bar m}A_\nu^{\bar k - \bar m}, \tag{41}$$

$$D_\mu\phi^{\bar k} = \partial_\mu\phi^{\bar k} + ig(A_\mu\times\phi)^{\bar k} = \partial_\mu\phi^{\bar k} + ig[(m_2+1)(\tfrac{1}{2}m_1 - k_1) - m_1(\tfrac{1}{2}m_2 - k_2)]A_\mu^{\bar m}\phi^{\bar k - \bar m}, \tag{42}$$

where g is the coupling constant. For the Hermitian fields $D_\mu\phi^{\bar k}$ and $F_{\mu\nu}^{\bar k}$, we have

$$(D_\mu \phi^{\bar{k}})^* = D_\mu \phi^{-\bar{k}}, \tag{43}$$

and

$$(F_{\mu\nu}^{\bar{k}})^* = F_{\mu\nu}^{-\bar{k}}, \tag{44}$$

so here the coupling constant g is required to be real. Under the gauge transformations (39), (40), we have

$$\delta F_{\mu\nu}^{\bar{k}} = i[m_1(k_2+2) - k_1(m_2+1)]\theta^{\bar{m}} F_{\mu\nu}^{\bar{k}-\bar{m}}, \tag{45}$$

and

$$\delta(D_\mu \phi^{\bar{k}}) = i[(m_2+1)(\tfrac{1}{2}m_1 - k_1) - m_1(\tfrac{1}{2}m_2 - k_2)]\theta^{\bar{m}}(D_\mu \phi^{\bar{k}-\bar{m}}). \tag{46}$$

Clearly $F_{\mu\nu}^{\bar{k}}$ and $D_\mu \phi^{\bar{k}}$ transform covariantly as an adjoint representation and a $(\alpha=-1/2, \beta=0)$ representation, respectively. With the above definitions, we can construct the following Lagrangian:

$$\begin{aligned} L &= -\frac{1}{4}\kappa^6 d_{\bar{i}\bar{j}}^{\bar{k}} (\phi^6)_{\bar{k}} F_{\mu\nu}^{\bar{i}} F_{\mu\nu}^{\bar{j}} + \frac{1}{2}(D_\mu \phi_{\bar{k}})(D^\mu \phi^{\bar{k}}) + \frac{1}{2}\mu^2(\phi_{\bar{k}})\phi^{\bar{k}} - \frac{\lambda}{4}(\phi_{\bar{k}}\phi^{\bar{k}})^2 \\ &= -\frac{1}{4}\kappa^6 (\phi^6)^{-\bar{i}-\bar{j}} F_{\mu\nu}^{\bar{i}} F_{\mu\nu}^{\bar{j}} + \frac{1}{2}(D_\mu \phi^{\bar{k}})^*(D^\mu \phi^{\bar{k}}) + \frac{1}{2}\mu^2(\phi^{\bar{k}})^*\phi^{\bar{k}} - \frac{\lambda}{4}(\phi^{\bar{k}*}\phi^{\bar{k}})^2, \end{aligned} \tag{47}$$

where the scale parameter κ is to keep $\kappa\phi^{\bar{k}}$ dimensionless. One can easily show that this Lagrangian is invariant under the gauge transformations (39), (40). The form of this Lagrangian is the same as the case of the gauge theory of the Virasoro symmetry.[7] If we take the conformal spin indices $i_2 = j_2 = k_2 = 0$, then the Lagrangian (47) comes back to the case of the Virasoro symmetry.

IV. SUMMARY

It is not as the general method[4] of the gauge theory of w symmetry. In this paper, the gauge theory of w symmetry is constructed in terms of its representations. It is known that the gauge theory of the Virasoro symmetry from its representation[7] can be described as an effective field theory of the string theory. As a higher-spin extension of the Virasoro algebra, the w_∞ algebra possesses more symmetry. This gauge theory will be useful in further explorations for the string theory.

[1] C. N. Pope and X. Shen, Phys. Lett. B **236**, 21 (1990).
[2] C. M. Hull, Phys. Lett. B **240**, 110 (1989).
[3] K. Schoutens, A. Sevrin, and P. Van Nieuwenhuizen, Phys. Lett. B **243**, 245 (1990).
[4] E. Bergshoeff, C. N. Pope, L. J. Romans, E. Sezgin, X. Shen, and K. S. Stelle, Phys. Lett. B **243**, 350 (1990).
[5] I. Kaplansky, Commun. Math. Phys. **86**, 49 (1982).
[6] B. L. Feigin and D. S. Fuks, Funct. Anal. Appl. **16**, 47 (1982).
[7] Y. M. Cho, Phys. Rev. Lett. **67**, 3469 (1991).
[8] Y. M. Cho and S. W. Zoh, Phys. Rev. D **46**, 3483 (1992).

Method of handling the divergences in the radiation theory of sources that move faster than their waves

H. Ardavan
Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, England

(Received 16 April 1999; accepted for publication 19 June 1999)

The nonintegrable singularities that arise when the retarded potentials associated with supersonically or superluminally moving sources are differentiated are closely related to those encountered in the context of the Cauchy problem for hyperbolic equations over odd-dimensional space-times. The purpose of this paper is to point out that the field components are given in these cases by Hadamard's finite parts of the resulting divergent integrals, as in the case of the Cauchy problem, and to show that the procedure familiar from the subluminal regime—which is used by Hannay—is not applicable when there are source elements that approach the observer with the wave speed and zero acceleration at the retarded time. © *1999 American Institute of Physics.* [S0022-2488(99)01809-5]

I. INTRODUCTION

The emission of waves by a moving point source whose speed exceeds the wave speed is generally described by a Lienard–Wiechert potential that has extended singularities. These singularities occur on the envelope of the emitted wave fronts and its cusps, where the waves interfere constructively and so form caustics. A well-understood example is the emission of acoustic waves by a point source that moves along a straight line with a constant supersonic speed. In this case a simple caustic forms along a cone issuing from the source, the so-called Mach cone, and the Lienard–Wiechert potential describing the sound amplitude diverges algebraically as this cone is approached from inside.

Neither supersonically nor superluminally moving sources can ever be pointlike.[1] However, the retarded potential due to any physically realizable source of this kind consists of the superposition of the potentials of the moving volume elements that constitute it, so that the Lienard–Wiechert potentials in question act as Green's functions for the calculation of the potentials of viable extended sources.

Thus, the integral representing the superposition of the contributions of the various volume elements of a supersonically or superluminally moving extended source to its potential would generally entail an integrand with algebraic singularities, singularities that have extended loci but are as a rule integrable. If (in order to calculate the field) this integral representation of the potential is differentiated with respect to the space-time coordinates of the observer, therefore, there results a new integral with an integrand that has a higher order singularity and so is no longer integrable (see, e.g., Ref. 2). A question that thus arises, and has been a source of some confusion in the literature,[3] is how to handle such nonintegrable singularities.

II. RADIATION INTEGRALS INVOLVING SINGULAR KERNELS AND THEIR HADAMARD'S FINITE PARTS

Green's functions with algebraic singularities also arise when the wave equation is solved over a space-time that has an even number of spatial dimensions. For example, the solution of the two-dimensional wave equation

$$\partial^2 G_{2D}/\partial x^2+\partial^2 G_{2D}/\partial y^2-\partial^2 G_{2D}/\partial(ct)^2=-4\pi\delta(x-\xi)\delta(y-\eta)\delta(t-\tau), \qquad (1)$$

in the absence of boundaries, has the form

$$G_{2D}=2c[c^2(t-\tau)^2-R_{2D}^2]^{-1/2}\theta(t-\tau-R_{2D}/c), \tag{2}$$

where $R_{2D}\equiv[(x-\xi)^2+(y-\eta)^2]^{1/2}$, δ is the Dirac delta function and $\theta(x)$ is unity for $x>0$ and zero for $x<0$ (see Ref. 4, p. 842).

The square-root singularity of this Green's function over the space-time cone $t-\tau=R_{2D}/c$ is integrable. However, to solve the initial-boundary-value problem for the two-dimensional wave equation with Cauchy data that are prescribed on a space-like surface (such as $t=$const) in the (x,y,t)-space, one needs to evaluate an integral over this surface whose integrand consists of the product of the normal derivative of G_{2D} (e.g., $\partial G_{2D}/\partial t$) with the initial value of the wave amplitude (see Ref. 4, p. 893). The singularity of the kernel of the required integral at the intersection of the cone $t-\tau-R_{2D}/c=0$ with the space-like surface on which the Cauchy data is prescribed is consequently like that of $(t-t_0)^{-3/2}$ at $t-t_0=0$ and so is not integrable.

The first systematic discussion of such nonintegrable singularities was given by Hadamard in his general treatment of the Cauchy problem for hyperbolic partial differential equations that have an odd number of independent variables. From his work, and from a modern version of it featuring in the theory of generalized functions, it is well known that the way to handle the nonintegrable singularities of the derivatives of the Green's functions that appear in the solutions to the Cauchy problem is to take the so-called Hadamard finite part of the divergent integrals in question.[5,6]

There is a close analogy between the Lienard–Wiechert potentials for sources that move faster than their waves and the Green's functions for hyperbolic differential equations that govern the propagation of waves in two spatial dimensions. For example, the potential for a point source of unit strength that moves along a straight line ($x=$const, $y=$const) with a supersonic speed u ($>c$) is given, at an observation point P with the space-time coordinates $(x_P,y_P,z_P;t_P)$, by

$$G_M=2[(\tilde{z}-\tilde{z}_P)^2-(M^2-1)R_{2D}^2]^{-1/2}\theta[\tilde{z}-\tilde{z}_P-(M^2-1)^{1/2}R_{2D}], \tag{3}$$

where $\tilde{z}$ is defined as $z-ut$ and M stands for the Mach number u/c (see, e.g., Ref. 7). The Lienard–Wiechert potential G_M has precisely the same mathematical structure as the Green's function G_{2D} of the two-dimensional wave equation (1).

This analogy stems from the fact that the density ρ of a rectilinearly moving supersonic source with a fixed distribution pattern depends on z and t in only the combination $z-ut=\tilde{z}$, i.e., is of the form $\rho(x,y,z,t)=\rho(x,y,\tilde{z})$. In the absence of space-time boundaries, this symmetry of the source $(\partial\rho/\partial t=-u\,\partial\rho/\partial z)$ imposes a corresponding symmetry $(\partial\psi/\partial t=-u\,\partial\psi/\partial z)$ on the potential so that the wave equation governing the potential $\psi(x,y,\tilde{z})$ of such a source,

$$\partial^2\psi/\partial x^2+\partial^2\psi/\partial y^2-(M^2-1)\partial^2\psi/\partial\tilde{z}^2=-4\pi\rho(x,y,\tilde{z}), \tag{4}$$

has only three rather than four independent variables. Thus the functions G_{2D} and G_M are analogous because the differential operators in Eqs. (1) and (4) can be rendered identical by a mere rescaling of the coordinates.

Unlike some other symmetries (e.g., $\partial\psi/\partial t=0$) that alter the type of the differential equation governing the potential from hyperbolic to elliptic or parabolic at the same time as reducing the number of its independent variables, the symmetry imposed by the rigidity of a moving source distribution does not affect the type of the equation when the speed of the source exceeds the wave speed: for $M>1$, the coefficient of $\partial^2\psi/\partial\tilde{z}^2$ in (4) is negative and so the variable $\tilde{z}$ is time-like. Thus, the well-known singularity associated with the ray conoid of the two-dimensional wave equation[8] is mathematically identical to the caustic singularity that occurs on the envelope of the wave fronts emanating from a moving point source in three dimensions.

Inasmuch as the origins of the two types of singularity are mathematically the same, it is clear that the techniques for handling the analogous divergences that they cause in the contexts of radiation theory and of Cauchy problem should also be the same. The technique used in the context of the Cauchy problem has been known for some time,[5] but the applicability of this same technique to the corresponding problems in radiation theory does not seem to have been noted until recently.[2]

From the standpoint of the theory of generalized functions, there is a well-defined procedure for obtaining the physically relevant values of the divergent integrals that appear when the integral representations of the retarded potential are differentiated, a procedure involving integration by parts that extracts the so-called Hadamard finite part of the resulting divergent integrals (see, e.g., Ref. 6). Hadamard's finite part of the convolution of the density of a supersonically or superluminally moving extended source with the nonintegrable derivative of an associated Lienard–Wiechert potential yields the value that we would have obtained if we had first evaluated the original integral representing the retarded potential of the source as an explicit function of the spaec-time coordinates of the observer and then differentiated it.

What we have illustrated here with the aid of a simple source motion holds true also when the motion of the source is neither uniform nor rectilinear. Consider, for instance, an extended source with a rigidly rotating distribution pattern whose outer parts move with linear speeds exceeding the wave speed. The density of such a source depends on the azimuthal angle φ and the time t in only the combination $\varphi-\omega t \equiv \hat{\varphi}$, i.e., has the form $\rho(r,\varphi,z,t)=\rho(r,\hat{\varphi},z)$ where (r,φ,z) are the cylindrical polar coordinates based on the axis of rotation. The potential ψ that arises from the source in question is likewise subject to the symmetry $\partial\psi/\partial t=-\omega\partial\psi/\partial\varphi=-\omega\partial\psi/\partial\hat{\varphi}$ and so is governed by the following reduced version of the wave equation in cylindrical coordinates:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\psi}{\partial r}\right)+\frac{\partial^2\psi}{\partial z^2}+\left(\frac{1}{r^2}-\frac{\omega^2}{c^2}\right)\frac{\partial^2\psi}{\partial\hat{\varphi}^2}=-4\pi\rho(r,\hat{\varphi},z). \tag{5}$$

This is an equation of the mixed type. In the domain $r>c/\omega$, where the coefficient of $\partial^2\psi/\partial\hat{\varphi}^2$ is negative and the variable $\hat{\varphi}$ acts as a time-like coordinate, it is a hyperbolic differential equation in *two* spatial dimensions.

Just as the spherical wave fronts emanating from a rectilinearly moving supersonic point source form a Mach cone, so the envelope of the wave fronts from a corresponding circularly moving point source (Ref. 2, Figs. 1 and 4) constitutes a caustic that coincides with the ray conoid of (5) in its domain of hyperbolicity (Ref. 9, Sec. 6). The Lienard–Wiechert potential describing the amplitudes of these waves is identical to the Green's function G_0 of (5) in unbounded space (Ref. 10, Appendix). The divergences that arise in the context of the initial-boundary-value problem for the above two-dimensional equation, and from the differentiation of its solution

$$\psi(\mathbf{x}_P,t_P)=\int r\,dr\,d\hat{\varphi}\,dz\,\rho(r,\hat{\varphi},z)G_0(r,r_P,\hat{\varphi}-\hat{\varphi}_P,z-z_P) \tag{6}$$

in the radiation theory of rapidly rotating extended sources, once again have the same origins, therefore, and should be handled by the same technique.[2]

III. DERIVATIVES OF THE RETARDED POTENTIAL IN THE CASE OF A RAPIDLY ROTATING EXTENDED SOURCE

Improper handling of the divergences discussed above has led to erroneous conclusions in the analysis of certain radiation problems. Hannay[3] bases his analysis of the emission from rapidly rotating extended sources on the following form of the retarded potential

$$\psi(\mathbf{x}_P,t_P)=\int \rho(\mathbf{x},t_P-|\mathbf{x}-\mathbf{x}_P|/c)/|\mathbf{x}-\mathbf{x}_P|d^3\mathbf{x} \tag{7}$$

and contends that since the only singularity of the integrand in this expression is that at the point $\mathbf{x}=\mathbf{x}_P$, which is inoffensive, one can differentiate (7) under the integral sign to obtain

$$\frac{\partial \psi}{\partial \mathbf{x}_P} = \int \frac{\partial}{\partial \mathbf{x}_P}\left[\frac{\rho(\mathbf{x}, t_P - |\mathbf{x}-\mathbf{x}_P|/c)}{|\mathbf{x}-\mathbf{x}_P|}\right] d^3\mathbf{x} = \int \frac{\nabla \rho(\mathbf{x}, t_P - |\mathbf{x}-\mathbf{x}_P|/2)}{|\mathbf{x}-\mathbf{x}_P|} d^3\mathbf{x} \tag{8}$$

[see Ref. 3, Eqs. (1.2), (1.5), and (1.6)]. Having thus ''avoided'' the singularities that arise when the alternative form (6) of the retarded potential is differentiated, Hannay then uses (8) to derive an upper bound on the intensity of the resulting radiation.

The result claimed by Hannay[3] is erroneous because the steps, familiar from the subluminal regime, that are taken in Eq. (8) are not mathematically permissible when the moving source has volume elements that approach the observer with the wave speed and zero acceleration at the retarded time. To demonstrate this, here we shall render these steps explicit by taking them in the case of a specific source distribution, a source distribution that is bounded and smooth but entails motion at speeds exceeding the wave speed.

Let us consider a spherical source with the radius λ whose center moves on a circle of radius r_0 with the constant angular frequency ω and whose density smoothly reduces from a maximum ρ_0 at its center to zero at its boundary, e.g., has the form

$$\rho(r,\hat{\varphi},z) = \begin{cases} \rho_0 \cos^2[\pi X/(2\lambda)] & \text{if } X \leqslant \lambda, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where

$$X \equiv (z^2 + r^2 + r_0^2 - 2rr_0 \cos\hat{\varphi})^{1/2}$$

is the distance of a point $(r,\hat{\varphi},z) = (r,\varphi - \omega t, z)$ that is stationary in the rotating frame from the center $(r = r_0,\ \hat{\varphi} = 0,\ z = 0)$ of the sphere. The circle in broken lines in Fig. 1 shows the intersection, with the plane $z = 0$, of the boundary of the above source in the $(r,\hat{\varphi},z)$-space for $r_0 = \frac{3}{2}c/\omega$ and $\lambda = \frac{1}{2}c/\omega$. (The axes in this figure are marked in units of c/ω and the larger dotted circles designate the sonic or the light cylinder $r = c/\omega$ and the orbit $r = \frac{3}{2}c/\omega$ of the center of the source, respectively.)

Once the quantities $|\mathbf{x}-\mathbf{x}_P|$ and $d^3\mathbf{x}$ in (7) are expressed in terms of cylindrical coordinates and the above expression for the source density is inserted in this form of the retarded potential, there results an integral over the (r,φ,z)-space for which the domain of integration is automatically bounded.[11] Not only do we need to replace $\hat{\varphi}$ in the above expression for ρ by its retarded value

$$\hat{\varphi} = (\varphi - \omega t)|_{t = t_P - |\mathbf{x}-\mathbf{x}_P|/c} = \varphi - \omega t_P + [(z - z_P)^2 + r^2 + r_P^2 - rr_P \cos(\varphi - \varphi_P)]^{1/2}\omega/c \tag{10}$$

when substituting (9) in (7), but in addition we need to delineate the domain of integration in (7), by mapping the source boundary $X = \lambda$ from the $(r,\hat{\varphi},z)$-space onto the (r,φ,z)-space. The image of the source boundary under the mapping $\hat{\varphi} \rightarrow \varphi$ expressed in (10) is a surface whose shape is different for different observers, or at different observation times, and bears no direct relationship with the sphere $X = \lambda$ that appears in (9).

To specify the boundary of the domain of integration in (7), we need to solve the transcendental equation (10) for φ at every point $(r,\hat{\varphi},z)$ of the sphere $X = \lambda$. In the case of the source depicted in Fig. 1, and of an observer that is located at $(r_P, \varphi_P, z_P) = (\frac{5}{2}c/\omega, 0, 0)$ at the observation time $t_P = (2\pi - \arccos\frac{2}{5} + \sqrt{21}/2)\omega^{-1}$, the intersection of this domain of integration with the plane $z = 0$ has the shape shown by the solid curve in Fig. 1.

The boundary of the irregular volume occupied by the source in the (r,φ,z)-space intersects a circle $r = \text{const}$, $z = \text{const}$, (with $1 < r\omega/c < 2$ and $-\frac{1}{2} < z\omega/c < \frac{1}{2}$) at either two, four, or six values of φ. If we let $(\varphi_l^{(n)}, \varphi_u^{(n)})$, with $n = 1,2,\ldots$, denote the various intervals in φ that are occupied by the retarded distribution of the source at any given $(r,z;r_P,\varphi_P,z_P,t_P)$, then the volume integral in (7) may be written as a triple integral over the variables φ, z, and r, respectively, in which the functions $\varphi_l^{(n)}(r,z;r_P,\varphi_P,z_P,t_P)$ and $\varphi_u^{(n)}(r,z;r_P,\varphi_P,z_P,t_P)$ constitute the

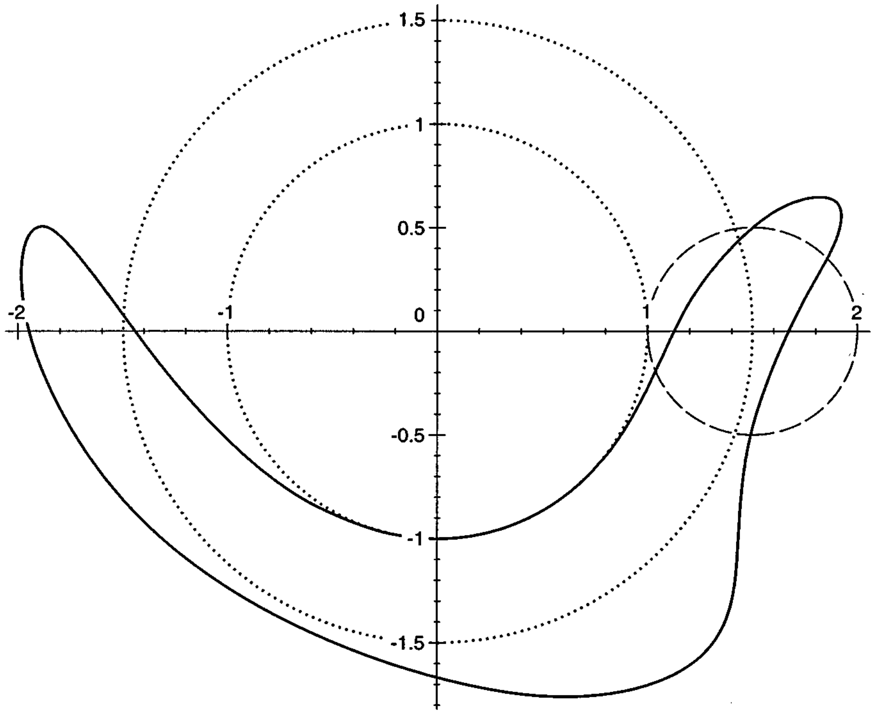

FIG. 1. The retarded shape of the source boundary in the (r,φ,z)-space (the solid curve) compared to its original shape in the $(r,\hat{\varphi},z)$-space (the smallest circle in broken lines).

various limits of integration with respect to φ. Differentiation of the integral in question entails the differentiation of these limits of integration, limits that are given by the solutions φ of (10) for a point $(r,\hat{\varphi},z)$ on the boundary of the source distribution.

Differentiating (10) with respect to $\mathbf{x}_P$ while holding $(r,\hat{\varphi},z)$ and the observation time $\hat{\varphi}_P$ constant, we find that the gradient of any of the $\varphi_l^{(n)}$ or $\varphi_u^{(n)}$ is given by an expression

$$\nabla_P\varphi=r_P^{-1}\hat{e}_{\varphi_P}-(\omega/c)\{[r_P-r\cos(\varphi-\varphi_P)]\hat{e}_{r_P}+(z_P-z)\hat{e}_{z_P}\}$$
$$\times\{[(z-z_P)^2+r^2+r_P^2-rr_P\cos(\varphi-\varphi_P)]^{1/2}-rr_P\omega\sin(\varphi_P-\varphi)/c\}^{-1}, \qquad (11)$$

whose denominator both vanishes and has a vanishing derivative at the boundary points that approach the observer with the wave speed and zero acceleration (see Ref. 2, Appendix B).

In Eq. (8), Hannay[3] applies Leibniz's formula for the differentiation of a definite integral assuming that there are no contributions from the limits of integration. Leibniz's formula, on the other hand, is not applicable if there are any points at which the limits of integration are not differentiable.[12] In the case considered here, where the derivatives of the limits of integration are singular, the gradient of the integral in question does not consist solely of the integral of the gradient of its kernel, as claimed by Hannay.[3] There is an additional contribution to the gradient of the potential: that which arises from the singularities of the gradients of the limits of integration in (7) and which comprises the boundary contribution to the Hadamard finite part of the gradient of the integral in (6).

The singularities of the gradients of the limits of integration in (7) are the images, under the mapping $\hat{\varphi}\rightarrow\varphi$, of the singularities of the integrand of the gradient of (6): They both arise from those source elements that approach the observer with the wave speed and zero acceleration at the retarded time.[2] By overlooking the contribution from the limits of integration in (8), Hannay[3] has discarded the boundary term in Hadamard's finite part of the divergent integral that results from the differentiation of the alternative form (6) of the retarded potential.

Note that it makes no difference if we extend the domain of integration in (7) over all (r,φ,z) by introducing a Heaviside step function that incorporates the vanishing of ρ outside $X=\lambda$ into the expression for ρ. Differentiation of such a step function results in a delta function, the gradient of whose argument vanishes—and so itself diverges—*algebraically* at those points of the boundary that approach the observer with the wave speed and zero acceleration. The product of this delta function and the vanishing value of ρ on the boundary is neither zero nor infinite; it is an *indeterminate* quantity that would have to be evaluated by means of a physically meaningful procedure. Far from ''avoiding'' the divergence that needs to be handled by Hadamard's technique, therefore, the adoption of this form of the retarded potential would merely replace the divergence in question by a closely related indeterminacy.

The contribution from the limits of integration to the right-hand side of (8) is zero, as assumed by Hannay,[3] only in the familiar subsonic or subluminal regimes where the derivatives of these limits are singularity free. In the case of a supersonically or superluminally moving accelerated source, this contribution is nonvanishing and has a value that may be calculated by means of Hadamard's method.[2] The upper bound derived by Hannay[3] applies only to the contribution to the derivative of the potential that arises from the derivative of its integrand, i.e., to the contribution that is retained by Hannay,[3] not to the contribution from the limits of integration that is overlooked by him.

ACKNOWLEDGMENT

The author is grateful to B. Mashhoon for helpful comments.

[1] B. M. Bolotovskii and V. L. Ginzburg, Sov. Phys. Usp. **15**, 184 (1972).
[2] H. Ardavan, Phys. Rev. E **58**, 6659 (1998).
[3] J. H. Hannay, Proc. R. Soc. London, Ser. A **452**, 2351 (1996).
[4] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Part 1 (McGraw-Hill, New York, 1953).
[5] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations* (Yale Univ. Press, New Haven, 1923).
[6] R. F. Hoskins, *Generalised Functions* (Horwood, London, 1979).
[7] A. P. Dowling and J. E. Ffowcs Williams, *Sound and Sources of Sound* (Horwood, London, 1983).
[8] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2 (Interscience, New York, 1962).
[9] H. Ardavan, Proc. R. Soc. London, Ser. A **424**, 113 (1989).
[10] H. Ardavan, Phys. Rev. D **29**, 207 (1984).
[11] The analysis in Ref. 3 is based on the assumption that the domain of integration in Eq. (7) extends over all space regardless of what the expression for the source density may be. That this is not so may be seen by writing out the explicit form of the integral in Eq. (7) for any source distribution—such as the one considered here—whose density vanishes outside a finite volume. For such source distributions, the integration in (7) does not extend over all space unless we introduce step functions that would incorporate the finiteness of their support into the mathematical expressions for their densities. The contributions of the required step functions toward the derivatives of the integral in (7), on the other hand, are the same as those of the original limits of integration.
[12] R. Courant, *Differential and Integral Calculus*, Vol. 2 (Blackie, New York, 1967).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Nonequilibrium dynamics of infinite particle systems with infinite range interactions

Changsoo Bahn[a)]
Department of Mathematics, Yonsei University, Seoul 120-749, Korea

Yong Moon Park[b)]
Department of Mathematics and Institute for Mathematical Sciences, Yonsei University, Seoul 120-749, Korea

Hyun Jae Yoo[c)]
Department of Mathematics, Kyungpook National University, Taegu 702-701, Korea

(Received 5 March 1999; accepted for publication 21 April 1999)

We discuss the existence and uniqueness of nonequilibrium dynamics of infinitely many particles interacting via superstable pair interactions in one and two dimensions. The interaction is allowed to be of infinite range and singular at the origin. Under suitable regularity conditions on the interaction potential, we show that if the potential decreases polynomially as the distance between interacting two particles increases, then the tempered solution to the system of Hamiltonian equations exists. Moreover, if the potential satisfies further that either it has a subexponential decreasing rate or it is everywhere two-times continuously differentiable, then we show that the tempered solution is unique. The results extend those of Dobrushin and Fritz obtained for finite range interactions. © *1999 American Institute of Physics.* [S0022-2488(99)01609-6]

I. INTRODUCTION

Our aim in this paper is to show the existence and uniqueness of nonequilibrium dynamics of infinitely many particles interacting via (infinite range) superstable interactions in dimensions one and two. In this paper we extend the results obtained by Dobrushin and Fritz in Refs. 1–3 where the interaction range is assumed to be finite. By suitably modifying and extending the methods given in Refs. 1–3, we remove the finite range assumption. The time evolution will be constructed in an explicitly defined set $\bar{\Omega}$ of allowed configurations characterized by a logarithmic order of energy fluctuations. The extension of the results for finite range interactions to infinite range interactions must be meaningful in the sense of practical and aesthetic points of view.

This study may be considered as being in the extension of the line of investigation on the dynamics of infinitely many particles that has been studied by various authors.[1–10] The main point in this study would be to show the existence of solutions to the infinite system of Hamiltonian equations [see (1.1)] and then to investigate the detailed properties of the solution such as the uniqueness of the solution and the essential self-adjointness of the generator of the evolution semigroup of the dynamics. While various techniques for showing the existence of the solutions have been developed in the papers including the above mentioned ones (see, e.g., Refs. 3 and 8 for further references), the method to show the energy bound of the solutions has not been fully developed for general dimensions. Only up to one and two dimensions and for finite range interactions, such an aim has been fulfilled in Refs. 1–3 and 5 for explicitly characterized initial conditions. The method of the energy bound has been also used to show the essential self-adjointness of the Liouville operator.[3,11]

[a)]Electronic mail: changsoo@math.yonsei.ac.kr
[b)]Electronic mail: ympark@bubble.yonsei.ac.kr
[c)]Electronic mail: yoohj@gauss.kyungpook.ac.kr

In Ref. 8, Siegmund–Schultze has shown that in any dimension, for a given translation invariant probability measure on the phase space such that the particle density and specific energy are finite, there is a class of configurations which carries a full measure such that the solutions with initial conditions from that class exist provided that the potential is not singular and has a finite interaction range. However, the method of Siegmund–Schultze do not yet permit us to prove the uniqueness of the solution and to control the properties of the solution. See also Refs. 9–10 and the references in Ref. 8 for earlier works.

We would like to mention the works related to the infinite system of Hamiltonian equations. In Ref. 12, Lang initiated to consider the system of first order differential equations for the system with additional white-noise terms.[12,13] Recently the system of stochastic differential equations related to the particle systems has been investigated using the Dirichlet form approach.[14–19]

Let us briefly describe the contents of this paper. We are going to consider the motion of a countable collection of identical particles of unit mass in the d-dimensional Euclidean space $\mathbb{R}^d$ with $d=1,2$. Let I be the index set for the particles. Configurations of the system are represented as infinite sequences $\omega=(q_k,\, p_k)_{k\in I}$, where $q_k\in\mathbb{R}^d$ and $p_k\in\mathbb{R}^d$ are the position and velocity of the particle labeled by $k\in I$, respectively. Sometimes we will use more informative notations $q_k=q_k(\omega)$, $p_k=p_k(\omega)$. Only locally finite configurations are allowed, i.e., the sequence $(q_k)_{k\in I}$ of positions can not have limit points at all, but some additional restrictions are necessary too. We assume that our particles interact by a symmetric pair potential $U:\mathbb{R}^d\to(-\infty,\infty]$ and $U(x)=U(-x)$. The potential U is allowed to be of an infinite range. Let $\operatorname{grad} U$ denote the gradient of U. The equations of motion read as

$$\frac{dp_k}{dt}=-\sum_{j\neq k}\operatorname{grad} U(q_k-q_j),\qquad \frac{dq_k}{dt}=p_k,\quad k\in I. \tag{1.1}$$

We will impose to the potential that it is superstable in the sense of Ruelle.[20,21] We allow that U is possibly singular at the origin and is of an infinite range. See Sec. II for the details.

The space $\overline{\Omega}$ of allowed configurations on which the evolution of the dynamics shall be constructed is defined in the following way. Let

$$g(u):=(1+\log(1+u))^{1/d},\quad u\geqslant 0, \tag{1.2}$$

and let $H(\omega,\mu,\sigma)$ denote the total energy plus a multiple of particle numbers in the sphere with center $\mu\in\mathbb{R}^d$ and radius $\sigma>0$, i.e.,

$$H(\omega,\mu,\sigma):=\frac{1}{2}\sum_{|q_k-\mu|\leqslant\sigma}\left[|p_k|^2+A+\sum_{j\neq k;|q_j-\mu|\leqslant\sigma}U(q_k-q_j)\right], \tag{1.3}$$

where $A\geqslant 0$ is a constant appearing in the superstable condition [see (2.1)] and makes $H(\omega,\mu,\sigma)$ non-negative numbers. Let us define

$$\overline{H}(\omega):=\sup_{\mu}\sup_{\sigma\geqslant g(|\mu|)}\sigma^{-d}H(\omega,\mu,\sigma). \tag{1.4}$$

$\overline{H}(\omega)$ is called the logarithmic fluctuation of energy of ω.[3] The allowed configuration space is defined by

$$\overline{\Omega}:=\{\omega:\omega \text{ is locally finite and } \overline{H}(\omega)<\infty\}. \tag{1.5}$$

It turns out that $\overline{\Omega}$ carries a large class of probability measures including Gibbs random fields with interaction U.[1,11,16]

We will show that in dimensions one and two, under proper conditions on the potential, there exists a tempered solution (see Sec. II for the definition) ω_t to (1.1) with any initial configuration $\omega\in\overline{\Omega}$. This solution can be obtained as the limit of solutions to finite subsystems. Furthermore, the tempered solution is unique when the potential U is either everywhere two-times continuously differentiable or subexponentially decreasing (see Theorem 2.2) when it is singular at the origin. The main idea used in this paper is to decompose the potential U into two parts:

$$U(x)=U^{(1)}(x)+U^{(2)}(x),$$

where $U^{(1)}$ is of finite range and $U^{(2)}$ is a smooth potential which decreases polynomially. See (2.2) and (2.6). For $U^{(1)}$, we use the method developed in Ref. 3 with suitable modifications and for $U^{(2)}$, we utilize the decay properties of $U^{(2)}$ stated in (2.6).

Employing the method developed in Sec. VI and using an appropriate modification of the method in Sec. 5 of Ref. 3, it may be possible to show the (anti) self-adjointness of the corresponding Liouville operator.[3,11] However we do not pursue the proof here and leave it to further study.

We organize this paper as follows. In Sec. II, we give necessary notations, impose conditions to the potential, and give main results of this paper. In Sec. III, we discuss the cutoff of total energy. In Sec. IV, we present the *a priori* bound which is a core in proving the results of this paper. We prove the existence and the uniqueness of the tempered solution in Sec. V and Sec. VI, respectively.

II. NOTATIONS AND MAIN RESULTS

In this section we give conditions to the interaction and state our main results. As mentioned in the Introduction, we will consider a symmetric pair potential U: $\mathbb{R}^d \to (-\infty,\infty]$, $d=1,2$, allowing the possibility of being of an infinite range and of singularity at the origin. We need, however, some regularity conditions on the potential in order to show the existence and uniqueness of the solutions to (1.1). First, we need superstability of the interaction to control the number of particles, just as in equilibrium theory.[20,21] There exist $A \geqslant 0$ and $B>0$ such that for any finite collection $q_1,q_2,\ldots,q_n$ of points of $\mathbb{R}^d$,

$$\sum_{k=1}^{n}\sum_{j\neq k} U(q_k-q_j) \geqslant -An+B\sum_{i\in\mathbb{Z}^d} n_i^2, \tag{2.1}$$

where $\mathbb{Z}^d$ is the d-dimensional integer lattice and for each site $i\in\mathbb{Z}^d$, n_i denotes the number of particles within the unit box with center i. This condition can be verified under natural assumptions, e.g., that U is not integrable near zero and integrable outside the origin.[21] We assume that $U(x)$ is continuously differentiable outside the origin and suppose that U is decomposed into two parts of singular and analytic potentials;

$$U(x)=U^{(1)}(x)+U^{(2)}(x). \tag{2.2}$$

Furthermore, we assume that $U^{(1)} \geqslant 0$ for all x and there exists a positive number r such that

$$U^{(1)}(x)=0, \quad \text{if } |x|\geqslant r. \tag{2.3}$$

It is important that singularity of U can not be too strong. Sometimes the velocity depends on the strength of the interaction and this velocity can become arbitrarily large in the extreme case. Thus, as in Ref. 3, we assume that there exist positive constants a and b such that

$$|x|\,|\mathrm{grad}\, U^{(1)}(x)| \leqslant a+b\, U^{(1)}(x). \tag{2.4}$$

From a mathematical point of view, (2.4) means that the singularity can not be stronger than that of $|x|^{-b}$. Also we need local Lipschitz condition. Let $L>1$ and suppose that

$$|\mathrm{grad}\, U^{(1)}(x)-\mathrm{grad}\, U^{(1)}(y)| \leqslant L[L+U^{(1)}(x)+U^{(1)}(y)]^c\,|x-y| \tag{2.5}$$

holds for all $x,y\neq 0$ with some constant $c>1$. (2.5) means that the singularity can not be too weak, logarithmic singularities are excluded. Finally we assume that the analytic potential $U^{(2)}(x)$ is everywhere two times continuously differentiable and decreases polynomially: There exist a decreasing function $\psi:\mathbb{R}^+\to\mathbb{R}^+$ and constants $\gamma>2d+3$ and $0<M$ such that

$$\psi(u)\leqslant M(1+u)^{-\gamma},\quad u\geqslant 0,$$

and

$$|U^{(2)}(x)|+|\mathrm{grad}\,U^{(2)}(x)|+\sum_{l,k=1}^{d}\left|\frac{\partial^2}{\partial x^{(l)}\,\partial x^{(k)}}U^{(2)}(x)\right|\leqslant\psi(|x|). \tag{2.6}$$

Notice that $U^{(2)}(x)\geqslant -M$, and hence $U(x)\geqslant -M$. We refer to Ref. 3 for the detailed accounting for the conditions on the potential.

We recall the notion of the logarithmic energy fluctuation $\bar{H}(\omega)$ of a given configuration ω and the possible configuration space given in (1.4)–(1.5). The configuration space $\bar{\Omega}$ is equipped with the product topology and with an associated Borel structure.[3] If $\omega_t^{(n)}$ is a sequence of trajectories in $\bar{\Omega}$, then convergence of $\omega_t^{(n)}$ means the uniform convergence on compact intervals of time of each of the components $q_k(\omega_t^{(n)})$ and $p_k(\omega_t^{(n)})$. This convergence need not be uniform in $k\in I$. A continuous trajectory $\omega_t:\mathbb{R}\to\bar{\Omega}$ is called a *tempered* solution to (1.1) with initial configuration $\omega\in\bar{\Omega}$ if $\omega_0=\omega$, $\bar{H}(\omega_t)$ is bounded in bounded intervals of time, and the components $q_k(t)=q_k(\omega_t)$, $p_k(t)=p_k(\omega_t)$ are continuously differentiable and satisfy (1.1) for all $k\in I$ and $t\in\mathbb{R}$. See Ref. 3 for the details.

As usual in the theory of infinite systems, solutions to (1.1) are constructed as limits of solutions to finite subsystems of (1.1). The passage to the infinite system is based on an *a priori* bound expressing a local version of the law of energy conservation. Just as in Refs. 1–3, energy flow will be controlled by means of a partial differential inequality formulated in terms of a spatial cutoff of total energy. Let us mention that the *a priori* bound can be obtained from this inequality only in dimensions one and two.

Our first main result concerns the existence of tempered solutions to (1.1).

Theorem 2.1: *Suppose that the hypotheses on the potential given in (2.1)–(2.6) are satisfied. Then, for any $\omega\in\bar{\Omega}$ there exists at least a tempered solution to (1.1) with initial configuration ω.*

The proof of the above theorem will be given in Sec. V. Next, we discuss the uniqueness of the solution. When the potential is singular at the origin, the method developed in this paper demands that the decreasing rate of the potential must be stronger than that given in (2.6), at least it must be of sub-exponential, in order that the tempered solution to (1.1) is unique. Recall the constant $c>1$ appearing in (2.5) for the Lipschitz condition of the potential.

Theorem 2.2: *Suppose that the conditions (2.1)–(2.5) for the potential hold. Suppose further that there exists a constant $\epsilon_0>0$ such that the function ψ in the right hand side of (2.6) satisfies $\psi(u)\leqslant\exp[-(\log u)^{c+\epsilon_0}]$ for large values of $u>0$. Then the tempered solution to (1.1) exists uniquely.*

When the potential is not singular at the origin, the polynomial decreasing rate is enough to insure the uniqueness.

Theorem 2.3: *Suppose that the potential $U(x)$ satisfies (2.1) and everywhere two-times continuously differentiable with a decreasing rate given in* (2.6) [*$U^{(2)}(x)$ being replaced by $U(x)$*]. Then the tempered solution to (1.1) exists uniquely.

The proofs of Theorem 2.2 and Theorem 2.3 will be given in Sec. VI.

III. CUTOFF OF TOTAL ENERGY

In this section, we introduce a smooth version of the energy functions $H(\omega,\mu,\sigma)$ defined in (1.3). In Ref. 3, exponentially decreasing functions have been used as a smooth version of the indicator functions. In order to deal with polynomially decreasing interaction potentials, we will use polynomially decreasing functions which satisfy all of the desired properties (Lemma 3.1).

Let $0<\lambda<1$. We will choose λ sufficiently small. Choose $\varphi:\mathbb{R}^+\to(0,1]$, continuously differentiable such that for $p>d+1$,

(i) $\varphi(u)=(1+\lambda u)^{-p}, \; u\geqslant 2$;
(ii) φ is concave for $u\leqslant 2$;
(iii) $\varphi(u)=\varphi(2)-\frac{1}{2}\varphi'(2)=(1/1+2\lambda)^p(1+(\lambda p/2(1+2\lambda)))$, for $u\leqslant 1$.

Notice that $\varphi(u)\leqslant(1+\lambda u)^{-p}$ and $0\leqslant-\varphi'(u)\leqslant p\lambda(1+\lambda u)^{-(p+1)}$.

Define

$$f(x,\sigma):=\int_{\mathbb{R}^d}\varphi\left(\frac{|x-y|}{\sigma}\right)\left(\frac{1}{1+\lambda|y|}\right)^{p+1}dy, \tag{3.1}$$

where $\sigma>0$. The definition (3.1) and the following lemma work well for $p>d$ but we need the condition $p>d+1$ for a later use (Lemma 3.3). In the proof of the *a priori* bound, $f(x-\mu,\sigma)$ will be used as a smooth version of the indicator function of the d-dimensional ball with center $\mu\in\mathbb{R}^d$ and radius σ. From now on we assume that $\sigma\geqslant 2$.

Lemma 3.1: There exist positive constants $c_1,\overline{c_1}$, and c_2 depending only on λ and p such that the following properties hold:

(a) $f(x-\mu,\sigma)\leqslant c_1(1+\lambda|x-\mu|/\sigma)^{-p}$ and $f(x,\sigma)\geqslant\overline{c_1}(1+|x|)^{-p}$ for all σ.
(b) $f(x-\mu,\sigma)\geqslant c_2>0$, if $|x-\mu|\leqslant\sigma$.
(c) $f(x,\sigma)\leqslant(1+\lambda|x-y|)^{p+1}f(y,\sigma)$.
(d) Denote by f' the derivative of f w.r.t. σ. Then $f'(x,\sigma)\leqslant(1+\lambda|x-y|)^{p+1}f'(y,\sigma)$.
(e) $|\mathrm{grad}\, f(x,\sigma)|\leqslant f'(x,\sigma)$.
(f) $g(|x|)\,|\mathrm{grad}\, f(x-\mu,\sigma)|\leqslant 4g(|\mu|+\sigma)f'(x-\mu,\sigma)$.

Proof: (a) We may assume $\mu=0$. Notice that

$$\left(1+\frac{\lambda|x|}{\sigma}\right)^p f(x,\sigma)\leqslant\int\left(\frac{\sigma+\lambda|x|}{\sigma+\lambda|x-y|}\right)^p\left(\frac{1}{1+\lambda|y|}\right)^{p+1}dy.$$

Using the inequality $|x|\leqslant|x-y|+|y|$, one can check that the right hand side of the inequality is finite uniformly in x and σ. This implies the first bound.

To obtain second bound, notice that $f(x,\sigma)\geqslant f(x,2)$. Since φ is concave for $u\leqslant 2$, there exist positive numbers $c>0$ and $\overline{c_1}$ such that the bound

$$\begin{aligned}(1+|x|)^p f(x,2)&\geqslant c\,(1+|x|)^p\int\left(\frac{1}{1+\lambda\,(|x-y|/2)}\right)^p\left(\frac{1}{1+\lambda|y|}\right)^{p+1}dy\\&\geqslant\overline{c_1}>0\end{aligned}$$

holds.

(b) Suppose $|x|\leqslant\sigma$. Then there exists a constant $c>0$ such that

$$\begin{aligned}f(x,\sigma)&\geqslant c\int_{|x-y|\leqslant\sigma}\left(\frac{1}{1+\lambda|y|}\right)^{p+1}dy\\&\geqslant c\int_{B_\sigma(0)\cap B_\sigma(x)}\left(\frac{1}{1+\lambda|y|}\right)^{p+1}dy,\\&\geqslant c\int_D\left(\frac{1}{1+\lambda|y|}\right)^{p+1}dy=:c_2>0,\end{aligned}$$

where $D=B_2(0)\cap B_2(z_1)$ and $z_1=2x/|x|$.

(c) The bound follows from the following relations:

$$f(x,\sigma)=\int_{\mathbb{R}^d}\varphi\left(\frac{|z|}{\sigma}\right)\left(\frac{1}{1+\lambda|x-z|}\right)^{p+1}dz$$

and

$$\frac{1+\lambda|y-z|}{1+\lambda|x-z|}\leqslant 1+\lambda|x-y|.$$

(d) Since

$$f'(x,\sigma)=-\int \varphi'\left(\frac{|x-y|}{\sigma}\right)\frac{|x-y|}{\sigma^2}\left(\frac{1}{1+\lambda|y|}\right)^{p+1}dy,$$

the result follows by the method same as that used in (c).

(e) This follows from a direct computation.

(f) We first show that

$$|x|\,|\mathrm{grad}\, f(x,\sigma)|\leqslant 4\sigma f'(x,\sigma). \tag{3.2}$$

In view of the fact (e), we may assume that $|x|\geqslant 4\sigma$. Let $D_1=\{y\in\mathbb{R}^d:|y|\leqslant|x-y|\}$ and $D_2=\mathbb{R}^d\setminus D_1$. Then $|y|\geqslant|x|/2$ if $y\in D_2$ and $-\varphi'(u)\leqslant p\lambda/(1+\lambda u)^{p+1}$.

We assert that for $y\in D_1$,

$$-\varphi'\left(\frac{|y|}{\sigma}\right)\left(\frac{1}{1+\lambda|x-y|}\right)^{p+1}\leqslant-\varphi'\left(\frac{|x-y|}{\sigma}\right)\left(\frac{1}{1+\lambda|y|}\right)^{p+1}. \tag{3.3}$$

Notice that $|y-x|\geqslant 2\sigma$ for $|y|\leqslant|x-y|$. One needs to show that for $y\in D_1$,

$$\left(\frac{1}{1+\lambda|y|/\sigma}\right)\left(\frac{1}{1+\lambda|x-y|}\right)\leqslant\left(\frac{1}{1+\lambda|x-y|/\sigma}\right)\left(\frac{1}{1+\lambda|y|}\right),$$

which is equivalent to

$$\frac{\sigma+\lambda|x-y|}{\sigma+\lambda|y|}\leqslant\frac{1+\lambda|x-y|}{1+\lambda|y|}.$$

Since $(\sigma+\lambda|x-y|)/(\sigma+\lambda|y|)$ is monotonic decreasing w.r.t. σ if $|y|\leqslant|x-y|$, this proves (3.3). Now (3.2) follows from the method used in Ref. 3, p. 545 and (f) follows from (3.2) by the method the same as that given in Ref. 3. □

From now on we will fix p such that $\gamma>2p+1>2(d+1)+1$ in the definition of φ in (3.1). The smooth version of total energy corresponding to $H(\omega,\mu,\sigma)$ given in (1.3) is defined as

$$W(\omega,\mu,\sigma)=\sum_{k\in I} f(q_k-\mu,\sigma)W_k(\omega), \tag{3.4}$$

where $q_k=q_k(\omega)$, $p_k=p_k(\omega)$, and

$$W_k(\omega)=2A+|p_k|^2+\sum_{j\neq k} U(q_k-q_j), \tag{3.5}$$

where A is the same as in (2.1). Define the logarithmic fluctuation of W by

$$\overline{W}(\omega)=\sup_{\mu}\sup_{\sigma\geqslant 2g(|\mu|)}\sigma^{-d}W(\omega,\mu,\sigma). \tag{3.6}$$

Recall that for any $i\in\mathbb{Z}^d$ and $\omega\in\overline{\Omega}$, $n_i(\omega)$ is the number of particles of ω in the unit box with a center at i. The superstability of U implies the following:

Lemma 3.2: There exists $0<\lambda<1$ *such that*

$$W(\omega,\mu,\sigma)\geqslant\frac{B}{4}\sum_{i\in\mathbb{Z}^d}f(i-\mu,\sigma)\,n_i(\omega)^2$$

and

$$\frac{\partial}{\partial\sigma}W(\omega,\mu,\sigma)\geqslant\frac{B}{4}\sum_{i\in\mathbb{Z}^d}f'(i-\mu,\sigma)\,n_i(\omega)^2.$$

Proof: Let $R>0$ be a sufficiently large natural number which will be fixed later and write the potential $U(x)$ as

$$U(x):=U^{(R,1)}(x)+U^{(R,2)}(x), \tag{3.7}$$

where $U^{(R,1)}(x):=U(x)1_{[0,R]}(|x|)$ and $U^{(R,2)}(x):=U(x)-U^{(R,1)}(x)$. We write

$$W(\omega,\mu,\sigma):=W^{(R,1)}(\omega,\mu,\sigma)+W^{(R,2)}(\omega,\mu,\sigma), \tag{3.8}$$

$$W^{(R,1)}(\omega,\mu,\sigma):=\sum_{k\in I}f(q_k-\mu,\sigma)\left[2A+|p_k|^2+\sum_{j\neq k}U^{(R,1)}(q_k-q_j)\right],$$

$$W^{(R,2)}(\omega,\mu,\sigma):=\sum_{k\in I}f(q_k-\mu,\sigma)\sum_{j\neq k}U^{(R,2)}(q_k-q_j).$$

It is easy to check that there is a positive constant c_3 such that

$$|W^{(R,2)}(\omega,\mu,\sigma)|\leqslant c_3\left(\sum_{k\in\mathbb{Z}^d}n_k(\omega)f(k-\mu,\sigma)\right)\left(\sum_{j\in\mathbb{Z}^d;|k-j|\geqslant R-2}n_j(\omega)(1+|k-j|)^{-\gamma}\right).$$

We use the fact that $n_k(\omega)n_j(\omega)\leqslant(1/2)(n_k(\omega)^2+n_j(\omega)^2)$, Lemma 3.1(c), $\gamma>2p+1$ and $p>d+1$ to obtain that

$$\begin{aligned}|W^{(R,2)}(\omega,\mu,\sigma)|&\leqslant c_3\left[\sum_{j\in\mathbb{Z}^d:|j|\geqslant R-2}(1+|j|)^{(-\gamma+p+1)}\right]\left[\sum_{k\in\mathbb{Z}^d}f(k-\mu,\sigma)n_k(\omega)^2\right]\\&\leqslant\frac{B}{4}\sum_{k\in\mathbb{Z}^d}f(k-\mu,\sigma)n_k(\omega)^2,\end{aligned} \tag{3.9}$$

where we have chosen R large enough.

On the other hand, by a similar method used above, it is not hard to see that if R is sufficiently large, then $U^{(R,1)}(x)$ is also superstable and satisfies that for any finite collection $q_1,\dots,q_n$ of points in $\mathbb{R}^d$,

$$\sum_{k=1}^{n}\sum_{j\neq k}U^{(R,1)}(q_k-q_j)\geqslant -A\sum_{k\in\mathbb{Z}^d}n_k+\frac{3}{4}B\sum_{k\in\mathbb{Z}^d}n_k^2, \tag{3.10}$$

where $A\geqslant 0$ and $B>0$ are the same constants that appeared in (2.1).

In order to control $W^{(R,1)}$, we modify the method used in the proof of Lemma 3.1 of Ref. 3. Introduce

$$\Lambda_u := \{x \in \mathbb{R}^d : u^{(i)} \leqslant x^{(i)} < u^{(i)} + mR, \quad 1 \leqslant i \leqslant d\},$$

where $x^{(i)}$ and $u^{(i)}$ are the coordinates of $x, u \in \mathbb{R}^d$, and m is a large natural number. Let P denote the set of pairs $[k,j]$ such that $|q_k - q_j| \leqslant R$ and let P_u be the set of $[k,j] \in P$ such that $k,j \in I_u$, where I_u is the set of particles in Λ_u. Put $\mathbb{Z}_u = \mathbb{Z}^d \cap \Lambda_u$. If ρ_u denotes the minimum of $f_k = f(q_k - \mu, \sigma)$ for $k \in I_u$, and $\lambda m R \sqrt{d} < \epsilon$, i.e., the diameter of Λ_u is less than ϵ/λ, then using Lemma 3.1 (c) and $U^{(R,1)}(x) \geqslant -M$, we obtain for each $[k,j] \in P_u$,

$$\begin{aligned}(f_k + f_j) U^{(R,1)}(q_k - q_j) &= 2\rho_u U^{(R,1)}(q_k - q_j) + (f_k + f_j - 2\rho_u) U^{(R,1)}(q_k - q_j)\\ &\geqslant 2\rho_u U^{(R,1)}(q_k - q_j) - 2M\rho_u((1+\epsilon)^{(p+1)} - 1),\end{aligned}$$

and so it follows from the above bound, (3.10) and Lemma 3.1 (c) that

$$\begin{aligned}\sum_{[k,j]\in P_u} (f_k + f_j) U^{(R,1)}(q_k - q_j) &\geqslant -A \sum_{k \in \mathbb{Z}_u} \rho_u n_k + \frac{3}{4} B \sum_{k\in\mathbb{Z}_u} \rho_u n_k^2\\ &\quad - \left(2M((1+\epsilon)^{(p+1)} - 1) \sum_{[k,j]\in P_u} \rho_u \right)\\ &\geqslant -A \sum_{k\in\mathbb{Z}_u} \rho_u n_k + \left(\frac{3}{4} B (1+\epsilon)^{-(p+1)} \sum_{k\in\mathbb{Z}_u} f(k-\mu,\sigma) n_k^2 \right)\\ &\quad - \left(M((1+\epsilon)^{(p+1)} - 1) \sum_{[k,j]\in P_u} (f_k + f_j) \right).\end{aligned}$$

Now let $z \in \Lambda_0 \cap R\ \mathbb{Z}^d$ be given. Then $\{\Lambda_u : u \in z + Rm\mathbb{Z}^d\}$ is a partition of $\mathbb{R}^d$. Let Q_z be the union of all P_u such that $u \in z + Rm\mathbb{Z}^d$. By the above estimate and $U^{(R,1)}(x) \geqslant -M$, we see that

$$\begin{aligned}W^{(R,1)}(\omega, \mu.\sigma) &\geqslant \left(\frac{3}{4} B(1+\epsilon)^{-(p+1)} \sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma) n_k^2 \right) - M \sum_{[k,j]\in P\setminus Q_z} (f_k + f_j)\\ &\quad - \left(M((1+\epsilon)^{(p+1)} - 1) \sum_{[k,j]\in Q_z} (f_k + f_j) \right)\\ &\geqslant \left(\frac{3}{4} B(1+\epsilon)^{-(p+1)} \sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma) n_k^2 \right)\\ &\quad - M \sum_{[k,j]\in P\setminus Q_z} (f_k + f_j) - \left(M((1+\epsilon)^{(p+1)} - 1) \sum_{[k,j]\in P} (f_k + f_j) \right). \end{aligned} \tag{3.11}$$

We notice that

$$\sum_{[k,j]\in P} (f_k + f_j) \leqslant c_4 \sum_{k\in\mathbb{Z}^d} \sum_{j\in\mathbb{Z}^d; |k-j| \leqslant R+2} n_k n_j f(k-\mu,\sigma).$$

Since $f(k-\mu,\sigma) \leqslant f(j-\mu,\sigma)\ (1+\lambda(R+2))^{(p+1)}$ if $|k-j| \leqslant R+2$, by Lemma 3.1 (c), and the number of j such that $|k-j| \leqslant R+2$ has an order of $(R+2)^d$, we use the Schwarz inequality to obtain

$$\sum_{[k,j]\in P}(f_k+f_j)\leqslant c_5\,(3+R)^{p+1+d}\sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma)\; n_k(\omega)^2. \tag{3.12}$$

Hence, inserting (3.12) into (3.11), we have

$$W^{(R,1)}(\omega,\mu.\sigma)\geqslant\left(\frac{3}{4}B(1+\epsilon)^{-(p+1)}-M((1+\epsilon)^{(p+1)}-1)\,c_5(3+R)^{p+1+d}\right)\sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma)n_k^2$$
$$-M\sum_{[k,j]\in P\setminus Q_z}(f_k+f_j). \tag{3.13}$$

We are going to sum both sides of the inequality (3.13) over $z\in\Lambda_0\cap R\mathbb{Z}^d$. The term in the left hand side and the first term in the right hand side of (3.13) are independent of z. Thus we consider the last term in the right hand side of (3.13). Given $[k,j]\in P$, the number of $z\in\Lambda_0\cap R\mathbb{Z}^d$ such that $[k,j]\in Q_z$ is larger than $(m-2)^d$, thus the number of z with $[k,j]\notin Q_z$ is less than $m^d-(m-2)^d\leqslant 2\,dm^{(d-1)}$. Consequently, summing both sides of (3.13) over $z\in\Lambda_0\cap R\;\mathbb{Z}^d$ we have

$$m^d\,W^{(R,1)}(\omega,\mu.\sigma)\geqslant\left[m^d\left(\frac{3}{4}B(1+\epsilon)^{-(p+1)}-M\big((1+\epsilon)^{(p+1)}-1\big)\,c_5(3+R)^{p+1+d}\right)\right.$$
$$\left.\times\sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma)n_k^2\right]-\left[M2\,dm^{(d-1)}\sum_{[k,j]\in P}(f_k+f_j)\right]. \tag{3.14}$$

Therefore, inserting (3.12) into (3.14) and dividing both sides of (3.14) by m^d, we get

$$W^{(R,1)}(\omega,\mu,\sigma)\geqslant C(R,\epsilon,m)\sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma)n_k(\omega)^2, \tag{3.15}$$

where

$$C(R,\epsilon,m)=\left(\frac{3}{4}B(1+\epsilon)^{-(p+1)}-M\big((1+\epsilon)^{(p+1)}-1\big)\,c_5(3+R)^{p+1+d}\right)$$
$$-c_5\frac{M2d}{m}(3+R)^{(p+1+d)}.$$

Now we choose and fix a sufficiently large R so that (3.9) and (3.10) hold. Then fix $\epsilon>0$ to be sufficiently small and fix m sufficiently large so that

$$C(R,\epsilon,m)\geqslant\tfrac{1}{2}B. \tag{3.16}$$

Finally we take and fix $0<\lambda<1$ so small that $\lambda mR\sqrt{d}<\epsilon$. By (3.8)–(3.9), (3.15)–(3.16), the first statement in the Lemma follows directly. Using Lemma 3.1 (d) instead of Lemma 3.1 (c), we obtain the second inequality in the same way. □

In the rest of this paper $\lambda>0$ will be fixed such that Lemma 3.2 holds. Finally, let us remark that H and W are equivalent in the following sense.

Lemma 3.3: There exist positive constants M_1 and M_2 such that

$$M_1\overline{H}(\omega)\leqslant\overline{W}(\omega)\leqslant M_2\overline{H}(\omega),$$

for all $\omega\in\overline{\Omega}$.

Proof: To prove the first part it is enough to show that

$$H(\omega,\mu,\sigma)\leqslant c\, W(\omega,\mu,\sigma), \tag{3.17}$$

for some constant $c>0$. By (2.6), $U(q_k-q_j)\geqslant -M(1+|q_k-q_j|)^{-\gamma}$ for all k,j. It follows from Lemma 3.1 (b) that

$$W(\omega,\mu,\sigma)\geqslant c_2 H(\omega,\mu,\sigma)-M\sum_k \sum_{j\neq k} f(q_k-\mu,\sigma)(1+|q_k-q_j|)^{-\gamma}.$$

By a simple computation and Lemma 3.2, we see that

$$\sum_k \sum_{j\neq k} f(q_k-\mu,\sigma)(1+|q_k-q_j|)^{-\gamma}\leqslant c_6 \sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma)n_k(\omega)^2$$
$$\leqslant c_6'\ W(\omega,\mu,\sigma).$$

Thus (3.17) follows from the above bounds.

Let us consider the second inequality. By the superstability of potential, one has that

$$\sum_{|j-\mu|\leqslant\sigma} n_j(\omega)^2\leqslant \frac{2}{B}H(\omega,\mu,\sigma+1). \tag{3.18}$$

We use the decomposition $U(x)=U^{(1)}(x)+U^{(2)}(x)$ in (2.2), so that $U^{(1)}(x)\geqslant 0$ and $U^{(1)}(x)=0$ if $|x|\geqslant r$, and $|U^{(2)}(x)|\leqslant M\ (1+|x|)^{-\gamma}$. We write

$$W(\omega,\mu,\sigma)=\sum_k f(q_k-\mu,\sigma)\left(2A+|p_k|^2+\sum_{j\neq k} U^{(1)}(q_k-q_j)\right)$$
$$+\sum_k f(q_k-\mu,\sigma)\sum_{j\neq k} U^{(2)}(q_k-q_j)$$
$$\equiv W_1+W_2.$$

Using Lemma 3.1 (a) and (c), and (3.18), we have that

$$|W_2|=\left|\sum_k f(q_k-\mu,\sigma)\sum_{j\neq k} U^{(2)}(q_k-q_j)\right|$$
$$\leqslant c_7 \sum_{k,j\in\mathbb{Z}^d;k\neq j} f(k-\mu,\sigma)\ n_k(\omega)\ n_j(\omega)\ (1+|k-j|)^{-\gamma}\leqslant c_7' \sum_{k\in\mathbb{Z}^d} f(k-\mu,\sigma)\ n_k(\omega)^2$$
$$\leqslant \tilde{c}\sum_{n=1}^{\infty} n^{-p} \sum_{k\in\mathbb{Z}^d;|k-\mu|\leqslant n\sigma} n_k(\omega)^2\leqslant \tilde{c}_1\sum_{n=1}^{\infty} n^{-p}H(\omega,\mu,n\sigma+1). \tag{3.19}$$

On the other hand, since $U^{(1)}(q_k-q_j)\geqslant 0$ and equals 0 for $|q_k-q_j|\geqslant r$, and since $f(q_k-\mu,\sigma)\leqslant f(q_j-\mu,\sigma)(1+\lambda|q_k-q_j|)^{(p+1)}$, we use Lemma 3.1 (a) to see that

$$W_1\leqslant c_8\sum_{n=1}^{\infty} n^{-p}\left(\sum_{|q_k-\mu|\leqslant n\sigma}\left(A+|p_k|^2+\sum_{|q_j-\mu|\leqslant n\sigma} U^{(1)}(q_k-q_j)\right)\right).$$

By adding and subtracting $U^{(2)}(q_k-q_j)$ to the last summand, we get that

$$W_1\leqslant c_8\sum_{n=1}^{\infty} n^{-p}\left(H(\omega,\mu,n\sigma)+\sum_{|q_k-\mu|\leqslant n\sigma,|q_j-\mu|\leqslant n\sigma} |U^{(2)}(q_k-q_j)|\right). \tag{3.20}$$

Observe that

$$\sum_{|q_k-\mu|\leqslant n\sigma,|q_j-\mu|\leqslant n\sigma} |U^{(2)}(q_k-q_j)|\leqslant c_9 \sum_{k,j\in\mathbb{Z}^d;|k-\mu|\leqslant n\sigma+1,|j-\mu|\leqslant n\sigma+1} n_k(\omega)\ n_j(\omega)(1+|k-j|)^{-\gamma}$$

$$\leqslant c_9' \sum_{|k-\mu|\leqslant n\sigma+1} n_k(\omega)^2\leqslant c_9''\ H(\omega,\mu,n\sigma+2). \tag{3.21}$$

It follows from (3.19)–(3.21) that

$$W(\omega,\mu,\sigma)\leqslant c_{10}\sum_{n=1}^{\infty} n^{-p}\ H(\omega,\mu,n\sigma+2)\leqslant c_{10}'\ \sigma^d\left(\sum_{n=1}^{\infty} n^{-(p-d)}\right)\ \overline{H}(\omega). \tag{3.22}$$

Dividing both sides of (3.22) by σ^d and using the fact that $p-d>1$, we get the desired result. □

IV. *A PRIORI* BOUND

Recall that a continuous trajectory $\omega_t:\mathbb{R}\to\overline{\Omega}$ is called a tempered solution to (1.1) with initial configuration ω if $\omega_0=\omega$ and the components $q_k(t)=q_k(w_t)$, $p_k(t)=p_k(w_t)$ are continuously differentiable and satisfy (1.1) for all $k\in I$ and $t\in\mathbb{R}$. We assume the dimension $d\leqslant 2$. The following is a local version of the law of energy conservation.

Proposition 4.1: There exists a constant $K>0$ *such that along any tempered solution* ω_t *to (1.1) we have*

$$\frac{\partial}{\partial t}W(\omega_t,\mu,\sigma)\leqslant K\ g(|\mu|+\sigma)[\overline{W}(\omega_t)]^{1/2}\frac{\partial}{\partial\sigma}W(\omega_t,\mu,\sigma),$$

for all $t\in\mathbb{R}$, $\mu\in\mathbb{R}^d$, *and* $\sigma\geqslant 2$.

Proof: In order to prove the proposition, we modify and extend the method used in the proof of Proposition 4.1 of Ref. 3. Recall the decomposition $U(x)=U^{(1)}(x)+U^{(2)}(x)$ in (2.2) and the inequality (2.4). A direct computation gives

$$\frac{\partial W}{\partial t}=\sum_{k\in I}\langle \operatorname{grad} f_k,p_k\rangle\ W_k+\frac{1}{2}\sum_k \sum_{j\neq k;|q_k-q_j|\leqslant r}(f_j-f_k)\langle \operatorname{grad} U^{(1)}(q_k-q_j),p_k+p_j\rangle$$

$$+\frac{1}{2}\sum_k \sum_{j\neq k;|q_k-q_j|\geqslant r}(f_j-f_k)\langle \operatorname{grad} U^{(2)}(q_k-q_j),p_k+p_j\rangle\equiv D_1+D_2+D_3, \tag{4.1}$$

where $f_k=f(q_k-\mu,\sigma)$.

For D_1, it is not hard to see that $|W_k|\leqslant W_k+M\ \Sigma_{j\neq k}(1+|q_k-q_j|)^{-\gamma}$. Thus

$$|D_1|\leqslant\sum_{k\in I}|\operatorname{grad} f_k|\ |p_k|\left[W_k+M\sum_{j\neq k}(1+|q_k-q_j|)^{-\gamma}\right].$$

Since $|p_k|\leqslant K_1\ g(|q_k|)\ [\overline{W}(\omega_t)]^{1/2}$, we use Lemma 3.1 (f) to get

$$|D_1|\leqslant 4K_1\ g(|\mu|+\sigma)[\overline{W}(\omega_t)]^{1/2}\sum_k f_k'\left[W_k(\omega_t)+M\sum_{j\neq k}(1+|q_k-q_j|)^{-\gamma}\right]. \tag{4.2}$$

By Lemma 3.1 (d), $f_k'\leqslant(1+\lambda|q_k-q_j|)^{p+1}\ f_j'$. Thus using Schwarz inequality and the fact that $\gamma>2p$, we get

$$\sum_{k\in I} f_k'\sum_{j\neq k}(1+|q_k-q_j|)^{-\gamma}\leqslant K_2\sum_{k\in\mathbb{Z}^d} f'(k-\mu,\sigma)\ n_k(\omega)^2. \tag{4.3}$$

From (4.2)–(4.3) and Lemma 3.2, one obtains

$$|D_1|\leqslant K_3\, g(|\mu|+\sigma)\,[\overline{W}(\omega_t)]^{1/2}\frac{\partial}{\partial\sigma}W(\omega_t,\mu,\sigma). \tag{4.4}$$

For D_3, since $|\text{grad}\, U^{(2)}(q_k-q_j)|\leqslant M\ (1+|q_k-q_j|)^{-\gamma}$, we have

$$|D_3|\leqslant K_4\sum_k\sum_{j\neq k}|\text{grad}\, f(z_{kj}-\mu,\sigma)|\,|q_k-q_j|\,(1+|q_k-q_j|)^{-\gamma}\ (g(|q_k|)+g(|q_j|))[\overline{W}(\omega_t)]^{1/2},$$

where z_{kj} is at some point between q_k and q_j. Note that, by Lemma 3.1 (f),

$$|\text{grad}\, f(z_{kj}-\mu,\sigma)|\leqslant\frac{1}{g(|z_{kj}|)}\ 4\ g(|\mu|+\sigma)\, f'(z_{kj}-\mu,\sigma).$$

We note that $g(x+y)\leqslant g(x)g(y)$ for any $x,y\geqslant 0$, and so $g(|q_k|)+g(|g_j|)\leqslant 2g(|z_{kj}|)g(|q_k-q_j|)$. Using Schwarz inequality, Lemma 3.1 (d), and Lemma 3.2, we obtain

$$\begin{aligned}|D_3|&\leqslant K_5\ g(|\mu|+\sigma)[\overline{W}(\omega_t)]^{1/2}\sum_{k\in\mathbb{Z}^d}\sum_{j\in\mathbb{Z}^d}[n_k(\omega)\ n_j(\omega)\ f'(l_{kj}-\mu,\sigma)|k-j|(1+|k-j|)^{-\gamma}\\&\quad\times(1+\log(1+|k-j|))]\\&\leqslant K_5'\ g(|\mu|+\sigma)[\overline{W}(\omega_t)]^{1/2}\\&\quad\times\sum_{k\in\mathbb{Z}^d}f_k'\ n_k^2(\omega)\left[\sum_{j\in\mathbb{Z}^d}|k-j|\,(1+|k-j|)^{-\gamma}\ (1+\log(1+|k-j|)\ (1+\lambda|k-j|)^{p+1}\right]\\&\leqslant K_5''\ g(|\mu|+\sigma)\,[\overline{W}(\omega_t)]^{1/2}\frac{\partial}{\partial\sigma}W(\omega_t,\mu,\sigma),\end{aligned}\tag{4.5}$$

where l_{kj} is a point between k and j. Here we have used $\gamma>2p+1$ to get the last inequality.

Finally for D_2, we use (2.4) to obtain the following bound:

$$\begin{aligned}|D_2|&\leqslant K_6\sum_k\sum_{j\neq k;|q_k-q_j|\leqslant r}|\text{grad}\, f(z_{kj}-\mu,\sigma)|\,(a+b\ U^{(1)}(q_k-q_j))\ g(|q_k|)\ [\overline{W}(\omega_t)]^{1/2}\\&\leqslant K_6'\ g(|\mu|+\sigma)[\overline{W}(\omega_t)]^{1/2}\sum_k\sum_{j\neq k;|q_k-q_j|\leqslant r}f'(z_{kj}-\mu,\sigma)\ (a+b\ U^{(1)}(q_k-q_j|))\frac{g(|q_k|)}{g(|z_{kj}|)}.\end{aligned}$$

Notice that if $|q_k-q_j|<r$, $g(|q_k|)\leqslant K_7\ g(|z_{kj}|)$. By Lemma 3.1 (d),

$$\sum_k\sum_{j\neq k;|q_k-q_j|\leqslant r}f'(z_{kj}-\mu,\sigma)\leqslant K_8\sum_{k\in\mathbb{Z}^d}\sum_{|j-k|\leqslant r}n_k\ n_j\ f_k'\leqslant K_8'\sum_{k\in\mathbb{Z}^d}f_k'\ n_k^2. \tag{4.6}$$

Moreover, a direct computation yields

$$
\begin{aligned}
&\sum_k \sum_{j\neq k;|q_k-q_j|\leqslant r} f'(z_{kj}-\mu,\sigma)U^{(1)}(q_k-q_j)\\
&\quad\leqslant K_9\sum_k f'_k \sum_{j\neq k;|q_k-q_j|\leqslant r} U^{(1)}(q_k-q_j)\\
&\quad\leqslant K_9\sum_k f'_k\left[2A+|p_k|^2+\sum_{j\neq k}U(q_k-q_j)\right]+K_{10}\sum_k f'_k \sum_{j\neq k;|q_k-q_j|\leqslant r}(1+|q_k-q_j|)^{-\gamma}\\
&\quad\leqslant K_9\sum_k f'_k\, W_k(\omega_t)+K_{11}\sum_{k\in\mathbb{Z}^d} f'(k-\mu,\sigma)\; n_k^2(\omega).
\end{aligned}\tag{4.7}
$$

Using (4.6)–(4.7) and Lemma 3.2, we see that

$$|D_2|\leqslant K_{12}\; g(|\mu|+\sigma)\;[\overline{W}(\omega_t)]^{1/2}\frac{\partial}{\partial\sigma}W(\omega_t,\mu,\sigma). \tag{4.8}$$

Thus the proof of the proposition follows from (4.4), (4.5), and (4.8). □

By the method in Proposition 4.2 of Ref. 3 and Lemma 3.3 one can prove the following proposition.

Proposition 4.2: For each $h>0$ *and* $T>0$ *there exists a finite* $\overline{h}=\overline{h}(h,T)$ *such that* $\overline{H}(\omega_0)\leqslant h$ *implies* $\overline{H}(\omega_t)\leqslant\overline{h}$ *for all* $|t|\leqslant T$ *provided that* ω_t *is a tempered solution to (1.1).*

As a consequence we have the following localization of particles.

Proposition 4.3: If ω_t *is a tempered solution to (1.1) and* $\overline{H}(\omega_0)<h$*, then there is a constant* $c(h,T)$ such that

$$|q_k(\omega_{t_1})-q_k(\omega_{t_2})|\leqslant c(h,T)\, g(|q_k(\omega_{t_1})|),$$

for all $k\in I,|t_1|\leqslant T,|t_2|\leqslant T$.

Proof: Let us simply write $q_k(\omega_t):=q_k(t)$. By Proposition 4.2, we have

$$\left|\frac{dq_k}{dt}\right|\leqslant g(|q_k|)\; c_1(h,T). \tag{4.9}$$

Define

$$q_{\max}:=\max_{|t|\leqslant T}|q_k(t)| \quad\text{and}\quad q_{\min}:=\min_{|t|\leqslant T}|q_k(t)|.$$

It follows from (4.9) that

$$|q_k(t_2)-q_k(t_1)|\leqslant 2\;T\,c_1(h,T)\;g(q_{\max}). \tag{4.10}$$

Similarly we have also

$$q_{\min}\geqslant q_{\max}-2\;T\,c_1(h,T)\;g(q_{\max}). \tag{4.11}$$

(4.11) implies the existence of $c_2(h,T)$ such that

$$\frac{g(q_{\max})}{g(q_{\min})}\leqslant c_2(h,T). \tag{4.12}$$

Inserting (4.12) into (4.10) and using the fact that $g(u)$ is an increasing function, we get the desired result. □

V. EXISTENCE

We are going to show the existence of tempered solutions to (1.1). For that purpose we have to investigate a kind of uniform boundedness. Let $(r_n)_{n=1}^{\infty}$ be any increasing sequence of positive numbers such that $r_n \to \infty$ as $n \to \infty$. Let $\omega_0 \in \overline{\Omega}$ be any initial configuration. For each $n \geqslant 1$, let $\omega_0^{(n)}$ be the configuration obtained by discarding the particles outside the r_n-ball with center the origin. Let $\omega^{(n)}(t) = (q_k^{(n)}(t), p_k^{(n)}(t))$ be the (finite system) solution of (1.1) with initial condition $\omega^{(n)}(0) = \omega_0^{(n)}$. We start with the following lemma saying that the energy densities of cutoff sub-systems are uniformly bounded.

Lemma 5.1: Let $\omega_0 \in \overline{\Omega}$ *and* $\omega_0^{(n)}$, $n \geqslant 1$, *be defined as above. There exists a constant* $c_0 > 0$ *independent of* ω_0 *such that the bounds*

$$\overline{H}(\omega_0^{(n)}) \leqslant c_0 \, \overline{H}(\omega_0)$$

hold uniformly in $n \in \mathbb{N}$.

Proof: For any $\mu \in \mathbb{R}^d$ and $\sigma > 0$, let us compare $H(\omega_0^{(n)}, \mu, \sigma)$ to $H(\omega_0, \mu, \sigma)$. For our purpose it is sufficient to consider the cases where the σ-ball $B_\sigma(\mu)$ touches the boundary of r_n-ball $B_{r_n}(0)$. In that case we see that

$$H(\omega_0^{(n)}, \mu, \sigma) \leqslant H(\omega_0, \mu, \sigma) - \sum^{*} U(q_k - q_l), \tag{5.1}$$

where $q_k = q_k(\omega_0)$, $p_k = p_k(\omega_0)$, and

$$\sum^{*} U(q_k - q_l) = \frac{1}{2}\left(\sum_{k \in \Lambda_1} \sum_{l \in \Lambda_2} U(q_k - q_l) + \sum_{k,l \in \Lambda_2} U(q_k - q_l) \right), \tag{5.2}$$

where $\Lambda_1 := B_{r_n}(0) \cap B_\sigma(\mu)$, $\Lambda_2 := B_\sigma(\mu) \setminus \Lambda_1$. By the superstability of the potential, we see that

$$\sum_{k,l \in \Lambda_2} U(q_k - q_l) \geqslant -A \sum_{l \in \Lambda_2 \cap \mathbb{Z}^d} n_l(\omega_0) \geqslant -A \sum_{l \in B_\sigma(\mu) \cap \mathbb{Z}^d} n_l(\omega_0)^2 \geqslant -\frac{2A}{B} H(\omega_0, \mu, \sigma + 1). \tag{5.3}$$

Here we have used (3.18) to get the last inequality. On the other hand, by using the decomposition of $U = U^{(1)} + U^{(2)}$ in (2.2), let us put

$$\sum_{k \in \Lambda_1} \sum_{l \in \Lambda_2} U(q_k - q_l) = \sum_{k \in \Lambda_1} \sum_{l \in \Lambda_2} U^{(1)}(q_k - q_l) + \sum_{k \in \Lambda_1} \sum_{l \in \Lambda_2} U^{(2)}(q_k - q_l). \tag{5.4}$$

Since $U^{(1)} \geqslant 0$ and $U^{(1)}(x) \geqslant -M(1 + |x|)^\gamma$, it is not hard to see that

$$\sum_{k \in \Lambda_1} \sum_{l \in \Lambda_2} U(q_k - q_l) \geqslant -c_{11} \sum_{k \in B_\sigma(\mu) \cap \mathbb{Z}^d} n_k(\omega_0)^2 \geqslant -c_{11}' H(\omega_0, \mu, \sigma + 1). \tag{5.5}$$

Inserting (5.3) and (5.5) into (5.1), we proved the lemma. □

When $q_k(0)$ is outside of the r_n-ball, define $q_k^{(n)}(t) \equiv q_k(0)$ and $p_k^{(n)}(t) \equiv 0$ for all $t \geqslant 0$.

Lemma 5.2: For each $k \in I$, *and* $T > 0$, *the sequence of continuous functions* $(q_k^{(n)}(\cdot))_{n=1}^{\infty}$ *and* $(p_k^{(n)}(\cdot))_{n=1}^{\infty}$ *on* $[0,T]$ *are equi-continuous.*

Proof: For any $-T \leqslant t_1 \leqslant t_2 \leqslant T$, it follows from Lemma 5.1, Proposition 4.2, and Proposition 4.3 that

$$|q_k^{(n)}(t_2)-q_k^{(n)}(t_1)|\leqslant\int_{t_1}^{t_2}|p_k^{(n)}(s)|\,ds\leqslant c_{12}\,g(|q_k(\omega_0)|)\ |t_2-t_1|. \tag{5.6}$$

On the other hand,

$$|p_k^{(n)}(t_2)-p_k^{(n)}(t_1)|\leqslant\sum_{j\neq k}\int_{t_1}^{t_2}|\mathrm{grad}U(q_j^{(n)}(s)-q_k^{(n)}(s))|\,ds.$$

We again decompose $U=U^{(1)}+U^{(2)}$ as in (2.2). By (2.5), we may assume that there exist $c>1$ and $L>1$ such that

$$|\mathrm{grad}U(x)|\leqslant L(L+U^{(1)}(x))^c. \tag{5.7}$$

Thus we have

$$\begin{aligned}\sum_{l\neq k}|\mathrm{grad}U(q_j^{(n)}-q_k^{(n)})|&\leqslant\sum_{j\neq k}L[L+U^{(1)}(q_j^{(n)}-q_k^{(n)})]^c\\&\leqslant L\left(\sum_{j\neq k:|q_j^{(n)}-q_k^{(n)}|\leqslant r}[L+U^{(1)}(q_j^{(n)}-q_k^{(n)})]\right)^c\\&\leqslant L\left(\sum_{j,l\in I_{k,n}}[L+U^{(1)}(q_j^{(n)}-q_l^{(n)})]\right)^c,\end{aligned} \tag{5.8}$$

where $I_{k,n}=\{j\in I:|q_j^{(n)}-q_k^{(n)}|\leqslant r\}$. By direct computations, we have that

$$\sum_{j,l\in I_{k,n}}L\leqslant L\ r^d\sum_{j\in\mathbb{Z}^d\cap B_r(q_k^{(n)})}(1+n_j^2)\leqslant c_{13}Lr^dH(\omega^{(n)},q_k^{(n)},r), \tag{5.9}$$

and

$$\begin{aligned}&\sum_{j,l\in I_{k,n}}U^{(1)}(q_j^{(n)}-q_l^{(n)})\\&\leqslant\sum_{j:|q_j^{(n)}-q_k^{(n)}|\leqslant r}\left(|p_j^{(n)}|^2+A+\sum_{l:|q_l^{(n)}-q_k^{(n)}|\leqslant r}U(q_j^{(n)}-q_l^{(n)})\right)+c_{14}\sum_{j,l\in I_{k,n}}(1+|q_j^{(n)}-q_l^{(n)}|)^{-\gamma}\\&\leqslant c_{14}'H(\omega^{(n)},q_k^{(n)},r).\end{aligned} \tag{5.10}$$

Inserting (5.9) and (5.10) into (5.8) and using Lemma 5.1, Proposition 4.2, and Proposition 4.3, we have that

$$|p_k^{(n)}(t_2)-p_k^{(n)}(t_1)|\leqslant c_{15}\,g(|q_k(\omega_0)|)^{cd}\ |t_2-t_1|. \tag{5.11}$$

Thus the lemma follows from (5.6) and (5.11). □

We are now ready to prove the existence of tempered solutions to (1.1). The method of the proof is given in Theorem 5.1 of Ref. 3, but we provide it for completeness.

Proof of Theorem 2.1: For each $n=1,2,\ldots$, let $\omega^{(n)}(t)$ be the tempered solution corresponding to the cutoff finite subsystem introduced above. By Lemma 5.2, for any finite $T>0$ and $k\in I$, the sequences $(q_k^{(n)}(t))_{n=1}^{\infty}$ and $(p_k^{(n)}(t))_{n=1}^{\infty}$ are equi-continuous. Thus by the Arzela–Ascoli theorem, after a diagonal procedure, we see that there exists a subsequence n_i, $i=1,2,\ldots$, such that for all $k\in I$, the sequences $(q_k^{(n_i)}(t))_{i=1}^{\infty}$ and $(p_k^{(n_i)}(t))_{i=1}^{\infty}$ uniformly converge on $[0,T]$ and define continuous functions $q_k(t)$ and $p_k(t)$. Since $T>0$ is arbitrary we can extend these functions to all $t\in\mathbb{R}$. Since $\overline{H}$ is lower semicontinuous, the *a priori* bounds remain in force for $\omega_t := (q_k, p_k)_{k\in I}$. Exploiting the continuity of the r.h.s. of the integral version of (1.1), we conclude that ω_t really satisfies (1.1). □

VI. UNIQUENESS

In order to prove uniqueness of tempered solution, we use an iteration method which turns out to be much more complicated compared to that used in Sec. 5 of Ref. 3. Consider a sequence $\delta(t,m)$, $m=0,1,\ldots,$ of non-negative and continuous functions on $[0,T]$ such that the bounds

$$\delta(t,m)\leqslant\delta(0,m)+L_1 a_m\int_0^t\delta(s,m+1)\,ds+L_2 b_m \tag{6.1}$$

hold for all $t\in[0,T]$ and m with some constants L_1 and L_2 (independent of m) and positive, increasing sequences $(a_m)_{m\geqslant 0}$ and $(b_m)_{m\geqslant 0}$. If

$$\delta(t,m)\leqslant Q_T\exp(mQ_T), \tag{6.2}$$

for $t\leqslant T$ and for all $m\geqslant 0$, and, furthermore, if the infinite series

$$\sum_{m=0}^{\infty}\frac{1}{m!}[(a_m)^m\ b_m]^{\alpha}<\infty \tag{6.3}$$

converges for any $\alpha\geqslant 1$, then (6.1) can be iterated infinitely many times to obtain

$$\delta(t,0)\leqslant\sum_{m=0}^{\infty}\ \delta(0,m)\frac{(L_1 t)^m}{m!}(a_m)^m+\sum_{m=0}^{\infty}\frac{(L_1 t)^m}{m!}(a_m)^m\ (L_2 b_m). \tag{6.4}$$

Recall the decomposition $U(x)=U^{(1)}(x)+U^{(2)}(x)$ in (2.2). $U^{(1)}(x)$ satisfies (2.3)–(2.5) and $U^{(2)}(x)$ satisfies (2.6).

Proof of Theorem 2.2: Let $f_0:\mathbb{R}\to[0,1]$ be a continuously differentiable nonincreasing function such that $f_0(u)=1$ if $u\leqslant 1$, $f_0(u)=0$ if $u\geqslant 2$, and $-f_0'(u)\leqslant f_0(u-1)$ for all $u\in\mathbb{R}$. Recall that the singular part of the interaction has finite range $r<\infty$. For each $R\geqslant r+1$ let $f_R(u) := f_0(u/R), u\in\mathbb{R}$. Then $f_R(u)$ also satisfies

$$f_R(u)=1\ \text{ if }\ u\leqslant R,\quad f_R(u)=0\ \text{ if }\ u\geqslant 2R,\quad\text{and }\ -f_R'(u)\leqslant f_R(u-R). \tag{6.5}$$

Suppose now that $\omega_t=(q_k(t),p_k(t))_{k\in I}$ and $\overline{\omega}_t=(\overline{q}_k(t),\overline{p}_k(t))_{k\in I}$ are two tempered solutions of (1.1) with the same initial condition $\omega_0=(q_k(0),p_k(0))\in\overline{\Omega}$. For each $R\geqslant r+1$, let us define

$$\begin{aligned} d_R(\omega_t,\overline{\omega}_t,m) &:= \sum_{k\in I} f_R(|q_k|-2\tilde{c}R(1+mg(m)))\ f_R(|\overline{q}_k|-2\tilde{c}R(1+mg(m)))\\ &\quad\times[|q_k-\overline{q}_k|_1+|p_k-\overline{p}_k|_1], \end{aligned} \tag{6.6}$$

where $\tilde{c}\geqslant 1$ is a sufficiently large fixed number that will be characterized later. We have used also the norm $|x|_1 := |x^{(1)}|+\cdots+|x^{(d)}|$ for each $x=(x^{(1)},\ldots,x^{(d)})\in\mathbb{R}^d$. We would like to show that for each $R\geqslant r+1$ the sequence of functions $\delta_R(t,m) := d_R(\omega_t,\overline{\omega}_t,m)$ satisfies (6.1)–(6.3) with suitably chosen sequences $(a_m)_{m\geqslant 0}$, $(b_m)_{m\geqslant 0}$ and constants $L_1(R)$, $L_2(R)$. Given $R\geqslant r+1$ and $T>0$, the r.h.s. of (6.6) is an absolutely continuous function of $t\in[0,T]$, thus we can differentiate it almost everywhere. By a notational convenience, let us drop the subscript R in $f_R(\cdot)$; $f(\cdot)\equiv f_R(\cdot)$ [cf. $f(x,\sigma)$ in (3.1)] and denote $f(t,k,m) := f(|q_k|-2\tilde{c}R(1+mg(m)))$, $\overline{f(t,k,m)} := f(|\overline{q}_k|-2\tilde{c}R(1+mg(m)))$. By differentiating the right hand side of (6.6), we have

$$\frac{d}{dt}(\delta_R(t,m)) = A^{(1)}(t,m) + A^{(2)}(t,m) + A^{(3)}(t,m) + A^{(4)}(t,m), \tag{6.7}$$

where

$$\begin{aligned} A^{(1)}(t,m) &= \sum_{k\in I} \left(\frac{d}{dt} f(t,k,m)\right)\overline{f}(t,k,m)(|q_k - \overline{q_k}|_1 + |p_k - \overline{p_k}|_1), \\ A^{(2)}(t,m) &= \sum_{k\in I} f(t,k,m)\left(\frac{d}{dt}\overline{f}(t,k,m)\right)(|q_k - \overline{q_k}|_1 + |p_k - \overline{p_k}|_1), \\ A^{(3)}(t,m) &= \sum_{k\in I} f(t,k,m)\overline{f}(t,k,m)\frac{d}{dt}|q_k - \overline{q_k}|_1, \\ A^{(4)}(t,m) &= \sum_{k\in I} f(t,k,m)\overline{f}(t,k,m)\frac{d}{dt}|p_k - \overline{p_k}|_1. \end{aligned} \tag{6.8}$$

Below, we will estimate each term in (6.8).

Since the bound

$$|p_k| \leqslant H(\omega, q_k, g(|q_k|))^{1/2} \leqslant c_{16}\, g(|q_k|)^{d/2} \tag{6.9}$$

holds for any $k \in I$, we use (6.5) to see that

$$\left|\frac{d}{dt} f(t,k,m)\right| \leqslant c_{17} f(t,k,m+1)\ (g(R)\, g(m))^{d/2}, \tag{6.10}$$

and

$$\left|\frac{d}{dt}\overline{f}(t,k,m)\right| \leqslant c_{17}\, \overline{f}(t,k,m+1)\ (g(R)\, g(m))^{d/2}. \tag{6.11}$$

Obviously,

$$\left|\frac{d}{dt}\right| q_k - \overline{q_k}|_1\, | \leqslant |p_k - \overline{p_k}|_1. \tag{6.12}$$

Thus, it follows from (6.9)–(6.12) that

$$|A^{(1)}(t,m) + A^{(2)}(t,m) + A^{(3)}(t,m)| \leqslant c_{18}(g(R) g(m))^{d/2} \delta_R(t,m+1). \tag{6.13}$$

Finally, let us consider the last term $A^{(4)}(t,m)$ in (6.8). Observe that

$$\begin{aligned} \left|\frac{d}{dt}\right| p_k - \overline{p_k}|_1\, | \leqslant & \sum_{l\neq k} |\operatorname{grad} U^{(1)}(q_l - q_k) - \operatorname{grad} U^{(1)}(\overline{q_l} - \overline{q_k})| \\ & + \sum_{l\neq k} |\operatorname{grad} U^{(2)}(q_l - q_k) - \operatorname{grad} U^{(2)}(\overline{q_l} - \overline{q_k})|. \end{aligned} \tag{6.14}$$

Thus

$$|A^{(4)}(t,m)| \leqslant Q_1(t,m) + Q_2(t,m), \tag{6.15}$$

where

$$Q_1(t,m)=\sum_{k\in I}\left(f(t,k,m)\overline{f}(t,k,m)\sum_{l\neq k}(|\operatorname{grad}U^{(1)}(q_l-q_k)-\operatorname{grad}U^{(1)}(\overline{q_l}-\overline{q_k})|)\right),\quad(6.16)$$

$$Q_2(t,m)=\sum_{k\in I}\left(f(t,k,m)\overline{f}(t,k,m)\sum_{l\neq k}(|\operatorname{grad}U^{(2)}(q_l-q_k)-\operatorname{grad}U^{(2)}(\overline{q_l}-\overline{q_k})|)\right).$$

In order to control the first term $Q_1(t,m)$, we use (2.5) to see that

$$\sum_{l\neq k}|\operatorname{grad}U^{(1)}(q_l-q_k)-\operatorname{grad}U^{(1)}(\overline{q_l}-\overline{q_k})|$$
$$\leqslant L\sum_{l\neq k}[L+U^{(1)}(q_l-q_k)+U^{(1)}(\overline{q_l}-\overline{q_k})]^c[|q_l-\overline{q_l}|_1+|q_k-\overline{q_k}|_1].\quad(6.17)$$

Following the method similar to that used to obtain (5.11), one can check that

$$\begin{aligned}&\sum_{l\neq k}[L+U^{(1)}(q_l-q_k)+U^{(1)}(\overline{q_l}-\overline{q_k})]^c\,|q_k-\overline{q_k}|_1\\&\leqslant c_{19}[H(\omega_t,q_k,r)+H(\overline{\omega_t},\overline{q_k},r)]^c\,|q_k-\overline{q_k}|_1\\&\leqslant c_{19}'[g(|q_k|)^{cd}+g(|\overline{q_k}|)^{cd}]\,|q_k-\overline{q_k}|_1\\&\leqslant c_{19}''\,g(R)^{cd}\,g(m)^{cd}\,|q_k-\overline{q_k}|_1.\end{aligned}\quad(6.18)$$

On the other hand, since $U^{(1)}$ has interaction range $r<R$, we see that

$$\begin{aligned}&\sum_{k\in I}f(|q_k|-2\tilde{c}R(1+mg(m)))\;f(|\overline{q_k}|-2\tilde{c}R(1+mg(m)))\\&\times\sum_{l\neq k}[L+U^{(1)}(q_l-q_k)+U^{(1)}(\overline{q_l}-\overline{q_k})]^c\,|q_l-\overline{q_l}|_1\\&\leqslant\sum_{l\in I}f(|q_l|-2\tilde{c}R(1+(m+1)g(m+1)))\;f(|\overline{q_l}|-2\tilde{c}R(1+(m+1)g(m+1)))\\&\quad\times|q_l-\overline{q_l}|_1\sum_{k\neq l}[L+U^{(1)}(q_l-q_k)+U^{(1)}(\overline{q_l}-\overline{q_k})]^c\\&\leqslant c_{20}\,g(R)^{cd}\,g(m)^{cd}\;\delta_R(t,m+1).\end{aligned}\quad(6.19)$$

Hence by (6.16)–(6.19), we obtain that

$$Q_1(t,m)\leqslant c_{21}\,g(R)^{cd}g(m)^{cd}\delta_R(t,m+1).\quad(6.20)$$

Let us estimate the second term $Q_2(t,m)$ in (6.16). We use (2.6) to obtain

$$\sum_{l\neq k} |\operatorname{grad} U^{(2)}(q_l-q_k)-\operatorname{grad} U^{(2)}(\overline{q_l}-\overline{q_k})|$$

$$\leqslant \sum_{l\neq k} (\psi(|q_l-q_k|)+\psi(|\overline{q_l}-\overline{q_k}|))\,[|q_l-\overline{q_l}|_1+|q_k-\overline{q_k}|_1]$$

$$=\left(\sum_{l\neq k}\psi(|q_l-q_k|)+\psi(|\overline{q_l}-\overline{q_k}|)\right)|q_k-\overline{q_k}|_1+\sum_{l\neq k}(\psi(|q_l-q_k|)+\psi(|\overline{q_l}-\overline{q_k}|))|q_l-\overline{q_l}|_1$$

$$\equiv S_1(t,k)+S_2(t,k). \tag{6.21}$$

Thus, by the definition of $Q_2(t,m)$ in (6.16), we have

$$Q_2(t,m)\leqslant\sum_{k\in I} f(t,k,m)\overline{f(}t,k,m)S_1(t,k)+\sum_{k\in I} f(t,k,m)\overline{f(}t,k,m)\,S_2(t,k)$$

$$\equiv Q_2^{(1)}(t,m)+Q_2^{(2)}(t,m). \tag{6.22}$$

In order to control $Q_2^{(1)}(t,m)$, we notice that by Lemma 3.1 (a) and (2.6), the bound

$$\psi(|q_l-q_k|)\leqslant c_{22}\,f(q_l-q_k,g(|q_k|)) \tag{6.23}$$

holds. We then use the above bound, the Schwarz inequality, and Lemma 3.2 to see that

$$\sum_{l\neq k}\psi(|q_l-q_k|)\leqslant c_{23}\,W(\omega_t,q_k,g(|q_k|))^{d/2}. \tag{6.24}$$

Repeating the same calculation after a change ω_t by $\overline{\omega_t}$, we conclude that

$$S_1(t,k)\leqslant c_{24}\,(g(R)\,g(m))^{d/2}\,|q_k-\overline{q_k}|_1,$$

and so

$$Q_2^{(1)}(t,m)\leqslant c_{24}\,(g(R)g(m))^{d/2}\delta_R(t,m+1). \tag{6.25}$$

Next, we consider $Q_2^{(2)}(t,m)$. We write

$$S_2(t,k)=\sum_{l\neq k}\psi(|q_l-q_k|)\;|q_l-\overline{q_l}|_1+\sum_{l\neq k}\psi(|\overline{q_l}-\overline{q_k}|)\;|q_l-\overline{q_l}|_1\equiv S_{2,1}(t,k)+S_{2.2}(t,k), \tag{6.26}$$

and

$$Q_2^{(2)}(t,m)=\sum_{k\in I} f(t,k,m)\overline{f(}t,k,m)\,S_{2,1}(t,k)+\sum_{k\in I} f(t,k,m)\overline{f(}t,k,m)\,S_{2,2}(t,k)$$

$$\equiv Q_2^{(2,1)}(t,m)+Q_2^{(2,2)}(t,m). \tag{6.27}$$

By symmetry, it is enough to control only $Q_2^{(2,1)}(t,m)$. For that purpose, let us further decompose $S_{2,1}$ as follows:

$$S_{2,1}(t,k)=\sum_{l\neq k:|q_l-q_k|\leqslant R}\psi(|q_l-q_k|)\;|q_l-\overline{q_l}|_1+\sum_{l:\,|q_l-q_k|>R}\psi(|q_l-q_k|)\;|q_l-\overline{q_l}|_1.$$

Substituting the above decomposition into the definition of $Q_2^{(2,1)}(t,m)$ in (6.27), one has that

$$Q_2^{(2,1)}(t,m)=\sum_{k\in I} f(t,k,m)\overline{f}(t,k,m)\sum_{l\neq k:|q_l-q_k|\leqslant R} \psi(|q_l-q_k|)\ |q_l-\overline{q_l}|_1$$
$$+\sum_{k\in I} f(t,k,m)\overline{f}(t,k,m)\sum_{l:|q_l-q_k|>R} \psi(|q_l-q_k|)\ |q_l-\overline{q_l}|_1\equiv D_1(t,m)+D_2(t,m). \tag{6.28}$$

By Proposition 4.3, it is not hard (and very important) to see that if $f(|q_k|-2\tilde{c}R(1+mg(m)))\neq 0$, $f(|\overline{q_k}|-2\tilde{c}R(1+mg(m)))\neq 0$, and $|q_l-q_k|\leqslant R$, then $f(|q_l|-2\tilde{c}R(1+(m+1)g(m+1)))=1$ and $f(|\overline{q_l}|-2\tilde{c}R(1+(m+1)g(m+1)))=1$ when $\tilde{c}\geqslant 1$ is a large enough constant. Thus, by interchanging the order of summation and using the argument used in deriving (6.24), we have the bound for $D_1(t,m)$ in (6.28):

$$D_1(t,m)\leqslant c_{25}\,(g(R)\,g(m))^{d/2}\ \delta_R(t,m+1). \tag{6.29}$$

Finally, let us consider $D_2(t,m)$. If $f(|q_k|-2\tilde{c}R(1+mg(m)))\neq 0$, it follows from Proposition 4.3 that

$$|q_l-\overline{q_l}|_1\leqslant c_{26}\,g(|q_l|)\leqslant c'_{26}\,g(|q_k|)\ g(|q_l-q_k|)\leqslant c''_{26}\,g(R)\ g(m)\ g(|q_l-q_k|).$$

Thus, using (6.23) and Lemma 3.2, we obtain that

$$\sum_{l:|q_l-q_k|>R} \psi(|q_l-q_k|)\ |q_l-\overline{q_l}|_1$$
$$\leqslant c_{27}\,g(R)g(m)\sum_{l\in\mathbb{Z}^d:|l-q_k|\geqslant R} n_l(\omega_t)\ \psi(|l-q_k|)\ g(|l-q_k|)$$
$$\leqslant c_{27}\,g(R)g(m)\left(\sum_{l:|l-q_k|\geqslant R}\psi(|l-q_k|)g(|l-q_k|)^2\right)^{1/2}\left(\sum_{l\in\mathbb{Z}^d} n_l(\omega_t)^2 f(l-q_k,g(|q_k|))\right)^{1/2}$$
$$\leqslant c'_{27}\,g(R)g(m)[\psi(R)R^{d+\epsilon_1}]^{1/2}[g(R)g(m)]^{d/2},$$

where $0<\epsilon_1<1$ is a small constant. Substituting the above bound into the expression of $D_2(t,m)$ in (6.28), we conclude that

$$D_2(t,m)\leqslant c'_{28}\,g(R)^2g(m)^2[\psi(R)R^{d+\epsilon_1}]^{1/2}\sum_{l\in\mathbb{Z}^d:|l|\leqslant 2\tilde{c}R(1+mg(m))+3R} n_l(\omega_t)$$
$$\leqslant c_{28}\,g(R)^2g(m)^2[\psi(R)R^{d+\epsilon_1}]^{1/2}[Rmg(m)]^d. \tag{6.30}$$

Here we have used superstability to get the last inequality [see (3.18)]. Thus it follows from (6.22), (6.25), and (6.28)–(6.30) that

$$Q_2(t,m)\leqslant c_{29}\,(g(R)g(m))^{d/2}\,\delta_R(t,m+1)\,+c_{28}\,g(R)^2g(m)^2\,[\psi(R)R^{d+\epsilon_1}]^{1/2}[Rmg(m)]^d. \tag{6.31}$$

Therefore from (6.13), (6.15), (6.20), and (6.31) we conclude that

$$\left|\frac{d}{dt}\delta_R(t,m)\right| \le \overline{C}_1[g(R)^{cd}g(m)^{cd}+(g(R)g(m))^{d/2}]\,\delta_R(t,m+1)$$
$$+\overline{C}_2\,g(R)^2g(m)^2\,[\psi(R)R^{d+\epsilon_1}]^{1/2}\,[Rmg(m)]^d$$
$$\le C_1(g(R)^{cd}\,g(m)^{cd})\;\delta_R(t,m+1)+C_2\,m^{2d}\,[\psi(R)]^{1/2}R^{d+(2/3)\,d}, \qquad (6.32)$$

where $\overline{C}_1, C_1, \overline{C}_2$, and C_2 are positive constants.

To use the iteration method, we see that for all $0\le t\le T$, the sequence $\delta_R(t,m) := d_R(\omega_t,\overline{\omega}_t,m)$ satisfies the relation (6.1) with $L_1=C_1\,g(R)^{cd}$, $L_2=T\;C_2\,[\psi(R)]^{1/2}R^{d+(2/3)d}$, $a_m=g(m)^{cd}$, and $b_m=m^{2d}$. It is easy to check that the conditions in (6.2)–(6.3) are fulfilled. Thus by (6.4) we have

$$\delta_R(\omega_t,\overline{\omega}_t,m)\le\sum_{m=0}^{\infty}\,\delta_R(\omega_0,\overline{\omega}_0,m)\,\frac{1}{m!}(C_1\,g(R)^{cd}t)^m\;g(m)^{cdm}$$
$$+C_3\,[\psi(R)]^{1/2}R^{d+(2/3)\,d}\sum_{m=0}^{\infty}\frac{1}{m!}(C_1\;g(R)^{cd}t)^m\;g(m)^{cdm}m^{2d}, \qquad (6.33)$$

where C_3 is a positive constant. Notice that $\delta_R(\omega_0,\overline{\omega}_0,m)=0$ for all $m=0,1,\ldots$. By using Hölder's inequality with conjugate pair $p=(1+\epsilon)$ and $p'=(1/\epsilon)\,p$ with small $\epsilon>0$, we obtain

$$\delta_R(\omega_t,\overline{\omega}_t,m)\le C_3\;[\psi(R)]^{1/2}R^{d+(2/3)\,d}(\exp[(C_1t)^{1+\epsilon}g(R)^{cd(1+\epsilon)}])^{1/(1+\epsilon)}$$
$$\times\left(\sum_{m=0}^{\infty}\frac{1}{m!}[g(m)^{cdm}m^{2d}]^{p'}\right)^{1/p'}. \qquad (6.34)$$

Notice that the last term converges to a constant. We check here that when R is large, $g(R)$ is similar to $(\log R)^{1/d}$ [see (1.2)]. On the other hand, when the potential is singular at the origin, then $c>1$ [see (2.5)]. Thus the exponential term in the r.h.s. of (6.34) increases faster than any polynomial order, but slower than any exponential order as $R\to\infty$. But, since we are assuming that $|U^{(2)}(x)|\le\psi(|x|)\le\exp[-(\log|x|)^{c+\epsilon_0}]$ for some $\epsilon_0>0$ and since the function $R\to\delta_R(\omega_t,\overline{\omega}_t,m)$ is increasing as R increases, by taking ϵ such that $c(1+\epsilon)<c+\epsilon_0$ and letting R go to infinity, we conclude from (6.34) that

$$\delta_{R_0}(\omega_t,\overline{\omega}_t,0)=0, \quad \text{for all} \quad R_0\ge r+1. \qquad (6.35)$$

Since $R_0\ge r+1$ can be taken arbitrarily large, (6.35) concludes the uniqueness. □

When the potential is not singular at the origin, the proof of uniqueness under the condition of the polynomial decreasing rate follows as a simple corollary to the method given in the above.

Proof of Theorem 2.3: Since the potential $U(x)$ is everywhere two-times continuously differentiable, i.e., $U^{(1)}(x)\equiv 0$, the terms concerning to $U^{(1)}$ disappear in the proof of Theorem 2.2. For example, we may discard the terms of (6.20). Thus, observing the derivation of (6.32), we have the bound

$$\left|\frac{d}{dt}\delta_R(t,m)\right|\le C_1\,(g(R)g(m))^{d/2}\delta_R(t,m+1)+C_2[\psi(R)]^{1/2}R^{d+(2/3)\,d}m^{2d}. \qquad (6.36)$$

That is, we may replace the constant cd by $d/2$ in (6.34). Notice again that $g(R)$ is similar to $(\log R)^{1/d}$ for large values of R. Since $\exp[(C_1T)^{1+\epsilon}(\log R)^{(1+\epsilon/2)}]$ increases slower than $(1+R)^{\alpha}$ for any $\alpha>0$ as $R\to\infty$, we see that the uniqueness holds with $\psi(R)=(1+R)^{-\gamma}$, the appropriate rate used to derive the *a priori* bound. □

ACKNOWLEDGMENTS

This work was supported in part by the Korea Research Foundation (98-015-D00043). H. J. Yoo was supported by a grant of the Post-Doc. Program from Kyungpook National University (1998).

[1] R. L. Dobrushin and J. Fritz, ''Non-equilibrium dynamics of one-dimensional infinite particle systems with a hard-core interaction,'' Commun. Math. Phys. **55**, 275–292 (1977).
[2] J. Fritz and R. L. Dobrushin, ''Non-equilibrium dynamics of two-dimensional infinite particle systems with a singular interaction,'' Commun. Math. Phys. **57**, 67–81 (1977).
[3] J. Fritz, ''Some remarks on non-equilibrium dynamics of infinite particle systems,'' J. Stat. Phys. **34**, 539–555 (1984).
[4] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai, *Ergodic Theory* (Springer-Verlag, New York, 1982).
[5] O. E. Lanford, III, ''The classical mechanics of one-dimensional systems of infinitely many particles,'' Commun. Math. Phys. **9**, 169–191 (1968).
[6] O. E. Lanford, III, ''Time evolution of large classical systems,'' in *Dynamical Systems. Theory and Applications*, Lecture Notes in Physics (Springer-Verlag, Berlin, 1975), Vol. 38, pp. 1–111.
[7] C. Marchioro, A. Pellegrinotti, and E. Presutti, ''Existence of time evolution of ν-dimensional statistical mechanics,'' Commun. Math. Phys. **40**, 175–185 (1975).
[8] Ra. Siegmund-Schultze, ''On non-equilibrium dynamics of multidimensional infinite particle systems in translation invariant case,'' Commun. Math. Phys. **100**, 245–265 (1985).
[9] Ya. G. Sinai, ''Construction of dynamics in one-dimensional systems of statistical mechanics,'' Theor. Math. Phys. **11**, 248–258 (1972).
[10] Ya. G. Sinai ''Construction of cluster dynamics for dynamical systems of statistical mechanics,'' Vestn. Mosk. Univ., Ser. 1: Mat., Mekh. **29**, 152–158 (1974).
[11] C. Marchioro, A. Pellegrinotti, and M. Pulvirenti, ''Selfadjointness of the Liouville operator for infinite classical systems,'' Commun. Math. Phys. **58**, 113–129 (1978).
[12] R. Lang, ''Unendlich-dimensionle Wienerprozesse mit Wechselwirkung. I,'' Z. Wahrscheinlichkeitstheor. Verwandte Geb. **38**, 55–72 (1977).
[13] J. Fritz, ''Gradient dynamics of infinite point systems,'' Ann. Prob. **15**, 478–514 (1987).
[14] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Analysis and geometry on configuration spaces,'' J. Funct. Anal. **154**, 444–500 (1998).
[15] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Analysis and geometry on configuration spaces: Gibbsian case,'' J. Funct. Anal. **157**, 242–271 (1998).
[16] V. Choi, Y. M. Park, and H. J. Yoo, ''Dirichlet forms and Dirichlet operators for infinite particle systems: Essential self-adjointness,'' J. Math. Phys. **39**, 6509–6536 (1998).
[17] H. Osada, ''Dirichlet form approach to infinite-dimensional Wiener processes with singular interaction,'' Commun. Math. Phys. **176**, 117–131 (1996).
[18] H. Spohn, ''Equilibrium fluctuations for interacting Brownian particles,'' Commun. Math. Phys. **103**, 1–33 (1985).
[19] M. W. Yoshida, ''Construction of infinite dimensional interacting diffusion processes through Dirichlet forms,'' Probab. Theory Relat. Fields **106**, 265–297 (1996).
[20] D. Ruelle, *Statistical Mechanics, Rigorous Results* (Benjamin, New York, 1969).
[21] D. Ruelle, ''Superstable interactions in classical mechanics,'' Commun. Math. Phys. **18**, 127–159 (1970).

Backscattering and inverse problem in random media

Boris M. Shevtsov[a)]
Radio Science Laboratory, Pacific Oceanological Institute,
43, Baltiyskay St., Vladivostok, 690041, Russia

(Received 1 June 1998; accepted for publication 11 December 1998)

With use of the invariant imbedding method and Markov process approximation, the statistical characteristics of the backscattering field from the impulse point source in a nonstationary multidimensional random medium are considered. The equations for the characteristic function and the density of the probability in the functional space of the backscattering are obtained. The equations for the statistical moments of the field are yielded and then solutions are presented in the matrix form. The additional averaging of the fast field variations is used for the simplification of the statistical equations. The procedures for the investigation of the direct and inverse statistical problem are proposed. The role of the phenomenological transfer theory is discussed from the statistical point of view. © *1999 American Institute of Physics.* [S0022-2488(99)00809-9]

I. INTRODUCTION

The invariant imbedding method[1–3] gives a good approach for the investigation of the wave problem in the determinant media and it allows applying the Markov process approximation[3–5] for the research of the statistical wave characteristics as well.

The recently developed method[4] of the wave problem inversion in a random medium is based on the assumption that the backscattering (reflected) field is approximately a Gaussian process. The direct solutions[5–7] of the wave statistical problem in the various cases show that this assumption takes place only in the single backscattering area, therefore it is interesting to extend this method on the multiple backscattering without the hypothesis of the Gaussian field and for the general case of a medium. This widening is needed for the remote sounding of the physical media with the large scattering depth as the dense plasma, the atmosphere clouds, and the light scattering layers of the ocean.

As the example, for the typical seawater,[8] the depth of the single backscattering is 35–50 m. At the same time, the active upper ocean layer, where the main dynamic processes take place, is 100–150 m. This layer is accessible for the modern light detection and ranging (lidar) systems but not for the known data lidar processing methods. The development of these methods is an actual problem now.

The hypothesis of the Gaussian field for the multiple backscattering leads to the noncorrect transport equations describing the transient signals with the mistake for the long time area.[9,10] The better alternative instead this hypothesis for the statistical inverse problem will be discussed below.

In the current paper, the scattering in the nonstationary multidimensional random medium of the scalar wave from the point impulsive source will be considered. We shall address the results[1] obtained for a determine inverse problem by using the invariant embedding method, which allows us to develop the analogous inverse procedure for the statistical case.

Following the work,[1] the next mathematical model accepted in statistical theory will be employed. The scalar wave propagates in the layer of the inhomogeneous nonstationary medium, which occupies the part of the space $L_0 \leqslant x \leqslant L$. The wave is created by the point impulsive source on the boundary of the layer $x=L$ and is described by the equation

[a)]Electronic mail: wave@online.vladivostok.ru

$$\left[\frac{\partial^2(\mathbf{1}+\epsilon(x,\rho,t))}{\partial t^2}-c^2\Delta\right]G(x,\rho,t;L,\rho_0,t_0)=\delta(x-L)\,\delta(\rho-\rho_0)\,\delta(t-t_0), \tag{1}$$

where $\epsilon(x,\rho,t)$ is the continuous and smooth characteristic of the scattering medium in the layer $L_0 \leqslant x \leqslant L$, ρ and ρ_0 are the vectors in the plane of constant x, c is the signal velocity in the homogeneous medium outside of the layer, where $\epsilon=0$, Δ is Laplacian; the product of the Dirac δ functions is in the right part of this equation.

Equation (1) has many applications in the different areas of physics, for example, in optics, acoustics, radio science, etc. It will be assumed that $-1<\epsilon(x,\rho,t)<\mu$ and all first partial derivatives of $\epsilon(x,\rho,t)$ exist and are bounded. When the statistical model of a medium will be discussed, the smallness of the fluctuations of the nonhomogeneity characteristic ($|\epsilon(x,\rho,t)| \ll 1$) will be supposed.

The solution of the mixed (initial and boundary values) problem for Eq. (1) named the Green's function will be used. The initial value is $G=0$ for $t<t_0$. The boundary conditions are the availability of the going away waves outside of the layer only, and the continuities of G and $\partial G/\partial x$ on two planes $x=L_0$ and $x=L$, where $\epsilon(x,\rho,t)$ breaks. The solution of Eq. (1) is the generalized function[11] and, for the homogeneous medium $[\epsilon(x,\rho,t)=0]$ has the form

$$\begin{aligned} g(x-L,\rho-\rho_0,t-t_0)&=\frac{\theta(t-t_0)}{4\pi c^2(t-t_0)}\,\delta\!\left(c(t-t_0)-\sqrt{(x-L)^2+(\rho-\rho_0)^2}\right)\\ &\equiv\frac{\theta(t-t_0)}{2\pi c}\,\delta(c^2(t-t_0)^2-(x-L)^2-(\rho-\rho_0)^2), \end{aligned} \tag{2}$$

where $\theta(t-t_0)$ is the Heaviside function.

In a nonhomogeneous medium, the solution $G(x,\rho,t;L,\rho_0,t_0)$ of the above problem is described by the system of the evolution equations[1] obtained in the frames of the imbedding method with the assumption that L is a variable parameter,

$$\begin{aligned} &\left[\frac{\partial}{\partial L}+\hat{M}(\rho_0,t_0)\right]G(x,\rho,t;L,\rho_0,t_0)\\ &\quad=-\int d\rho_1\,dt_1\,G(x,\rho,t;L,\rho_1,t_1)\,\frac{\partial^2}{\partial t_1^2}\,\epsilon(L,\rho_1,t_1)G_L(\rho_1,t_1;\rho_0,t_0), \end{aligned} \tag{3}$$

with the initial condition $G(x,\rho,t;L,\rho_0,t_0)|_{L=x}=G_x(\rho,t;\rho_0,t_0)$ and

$$\begin{aligned} &\left[\frac{\partial}{\partial L}+\hat{M}(\rho,t)+\hat{M}(\rho_0,t_0)\right]G_L(\rho,t;\rho_0,t_0)\\ &\quad=\frac{1}{c^2}\,\delta(\rho-\rho_0)\,\delta(t-t_0)-\int_1 d\rho\,dt_1\,G_L(\rho,t;\rho_1,t_1)\,\frac{\partial^2}{\partial t_1^2}\,\epsilon(L,\rho_1,t_1)G_L(\rho_1,t_1;\rho_0,t_0), \end{aligned} \tag{4}$$

with the initial data $G_L(\rho,t;\rho_0,t_0)|_{L=L_0}=g_0(\rho-\rho_0,t-t_0)$, where $g_0(\rho-\rho_0,t-t_0)=g(x-L,\rho-\rho_0,t-t_0)|_{x=L}$, the integral operators $\hat{M}(\rho,t)$ and $\hat{M}(\rho_0,t_0)$ are the inversions[1] of the Neumann[2] operator in a free space and have the kernels[1]

$$M(\rho-\rho_1,t-t_1)=-\frac{2}{(t-t_1)}\,\frac{\partial}{\partial t}\,g_0(\rho-\rho_1,t-t_1) \tag{5a}$$

and

$$M^{-1}(\rho-\rho_1,t-t_1)=2c^2 g_0(\rho-\rho_1,t-t_1). \tag{5b}$$

$G_L(\rho,t;\rho_0,t_0)$ is an impotent value in the inverse problem.[1] It may be presented in the composition $G_L(\rho,t;\rho_0,t_0)=g_0(\rho-\rho_0,t-t_0)+R_L(\rho,t;\rho_0,t_0)$, where the second term in the right part is the backscattering (reflected) field. The equation for $R_L(\rho,t;\rho_0,t_0)$ may be obtained from (4) with respect to the relation[1]

$$\hat{M}(\rho,t)g_0(\rho-\rho_0,t-t_0)=(1/2c^2)\,\delta(\rho-\rho_0)\,\delta(t-t_0). \tag{6}$$

Introducing the notations $\tau=\{q,q_0\}$, and $q=\{\rho,t\}$, and $q_0=\{\rho_0,t_0\}$ we can write the equation for $R_L(\rho,t;\rho_0,t_0)$ in the form

$$\frac{\partial}{\partial L}R_L(\tau)+[\hat{M}(q)+\hat{M}(q_0)]R_L(\tau)=-\int dq_1[g_0(q,q_1)+R_L(q,q_1)]\frac{\partial^2}{\partial t_1^2}\epsilon(L,q_1)[g_0(q_1,q_0)+R_L(q_1,q_0)], \tag{7}$$

with the initial condition $R_L(\tau)|_{L=L_0}=0$.

Amongst the four terms that took place under the integral in the right part of Eq. (7), the summand g_0g_0 describes the single scattering process, two cross-items g_0R_L and R_Lg_0 are associated with the multiple forward scattering when the backscattering is single and R_LR_L corresponds to the multiple backscattering. All the enumerated above items are the components of the inverse Neumann operator[2] in the nonhomogeneous medium. The nonlinear on R_L item in Eq. (7) creates the principal difficulties in the statistical analysis and leads to the non-Gaussian reflected field and to the nonclosed transport equation. At the same time, the linear on R_L cross-items do not create the problems with the closing of the statistical expressions. They correspond to the affect of the medium nonhomogeneities on the transport of the radiation and may be considered in the group with the linear items in the left part of Eq. (7). Having in mind the relation (5b), Eq. (7) may be rewritten as

$$\begin{aligned}&\frac{\partial}{\partial L}R_L(\tau)+\Bigg\{\hat{M}^{-1}(q)\left[\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\left(1+\frac{1}{2}\epsilon(L,q)\right)-\Delta_\rho\right]\\&\quad+\hat{M}^{-1}(q_0)\left[\left(1+\frac{1}{2}\epsilon(L,q_0)\right)\frac{1}{c^2}\frac{\partial^2}{\partial t_0^2}-\Delta_{\rho_0}\right]\Bigg\}R_L(\tau)\\&\quad=-\int dq_1\left[g_0(q,q_1)\frac{\partial^2}{\partial t_1^2}\epsilon(L,q_1)g_0(q_1,q_0)+R_L(q,q_1)\frac{\partial^2}{\partial t_1^2}\epsilon(L,q_1)R_L(q_1,q_0)\right].\end{aligned} \tag{8}$$

In a single backscattering area, the last item in Eq. (8) may be ignored and the generalized parabolic equation is obtained. In a small angle approximation, which gives the usual parabolic equation, which may be employed for the investigation of the radio or optic wave propagation in the turbulent atmosphere[3,12–14] with the limitation on a distance.

The roles in the statistical analysis of the enumerated above items under the integral in Eq. (7) depend from the size of the layer $L-L_0$. Two scales characterize the medium: the length of the single scattering l_s and the length of the single backscattering l_{SB}. For the anisotropic scattering medium, $l_s \ll l_{SB}$ and three scattering areas take place. Each of the enumerated above items in (7) is associated with one of these areas. The deference between forward and backward scattering is absent in the isotropic scattering medium ($l_s=l_{SB}$), and the linear and quadratic on R_L items under the integral in Eq. (7) contribute equally.

The structure of the solution singularities has to be considered for the discussion of the direct and inverse procedure. Performing the solution in the form

$$G_L(\rho,t;\rho_0,t_0)=\theta(t-t_0)\tilde{G}_L(\rho,t;\rho_0,t_0), \tag{9}$$

where $\theta(t-t_0)$ is the Heaviside function, substituting (9) in (4) and comparing the coefficients before two kinds of singularities $\theta(t-t_0)$ and $\delta(t-t_0)$ to zero, we obtain two equations:[1]

$$\frac{\partial}{\partial L}\tilde{G}_L(\rho;t;\rho_0,t_0)+\frac{1}{c}\frac{\partial}{\partial t}\tilde{G}_L(\rho,t;\rho_0,t_0)-\frac{1}{c}\frac{\partial}{\partial t_0}\tilde{G}_L(\rho,t;\rho_0,t_0)-\frac{2c}{\pi}\int_{t_0}^{t}dt_1\int d\rho_1\,\delta'(c^2(t-t_1)^2$$
$$-(\rho-\rho_1)^2)\tilde{G}_L(\rho_1,t_1;\rho_0,t_0)-\frac{2c}{\pi}\int_{t_0}^{t}dt_1\int d\rho\tilde{G}_L(\rho,t;\rho_1,t_1)\,\delta'(c^2(t_1-t_0)^2-(\rho_1-\rho_0)^2)$$
$$=h_L(\rho_0,t_0)\epsilon(L,\rho_0,t_0)\frac{\partial}{\partial t_0}\tilde{G}_L(\rho,t;\rho_0,t_0)-h_L(\rho,t)\frac{\partial}{\partial t}\epsilon(L,\rho,t)\tilde{G}_L(\rho,t;\rho_0,t_0)$$
$$+\int_{t_0}^{t}dt_1\int_1 d\rho_1\frac{\partial\tilde{G}_L(\rho,t;\rho_1,t_1)}{\partial t_1}\cdot\frac{\partial\epsilon(L,\rho_1,t_1)\tilde{G}_L(\rho_1,t_1;\rho_0,t_0)}{\partial t_1}, \tag{10}$$

where $\delta'(\cdots)$ is the derivative of the δ functions,

$$c^2\epsilon(L,\rho_0,t_0)h_L^2(\rho_0,t_0)+2c\cdot h_L(\rho_0,t_0)-1=0, \tag{11}$$

where $h_L(\rho_0,t_0)$ is the amplitude of the wave in the initial moment. This value is introduced with the use of the relationship $\tilde{G}_L(\rho,t;\rho_0,t_0)|_{t\to t_0}=h_L(\rho_0,t_0)\delta(\rho-\rho_0)$ and is determined by the characteristic of the medium in the point of the source in the initial moment through the solution of Eq. (11),

$$h_L(\rho_0,t_0)=\frac{\sqrt{1+\epsilon(L,\rho_0,t_0)}-1}{c\epsilon(L,\rho_0,t_0)}=\frac{n_L(\rho_0,t_0)-1}{c\epsilon(L,\rho_0,t_0)}, \tag{12}$$

where $n_L(\rho,t)=\sqrt{1+\epsilon(L,\rho,t)}$; it is a refractive index in electrodynamics.

In the homogeneous medium, $g_0(\rho-\rho_0,t-t_0)=\theta(t-t_0)\tilde{g}_0(\rho-\rho_0,t-t_0)$ and $\tilde{g}_0(\rho-\rho_0,t-t_0)|_{t\to t_0}=(1/2c)\delta(\rho-\rho_0)$ take off the sign ambiguity of (12).

The analogous equations for the backscattering can be obtained, if $\tilde{G}_L=\tilde{g}_0+\tilde{R}_L$ is substituted in (10) and $h_L=(1/2c)+r_L$ in (11), where $r_L(\rho_0,t_0)$ is the amplitude of the backscattering wave in the initial moment.

Equation (10) may be integrated along the characteristics,[2] which are defined for the partial differential equation, which is obtained from (10) in the first approximation on $t-t_0\to 0$,

$$\frac{\partial}{\partial L}\tilde{G}_L(\rho,t;\rho_0,t_0)+\frac{n_L(\rho,t)}{c}\frac{\partial}{\partial t}\tilde{G}_L(\rho,t;\rho_0,t_0)-\frac{n_L(\rho_0,t_0)}{c}\frac{\partial}{\partial t_0}\tilde{G}_L(\rho,t;\rho_0,t_0)$$
$$+h_L(\rho,t)\left[\frac{\partial}{\partial t}\epsilon(L,\rho,t)\right]\tilde{G}_L(\rho,t;\rho_0,t_0)\approx 0,$$

where $\tilde{G}_L(\rho,t;\rho_0,t_0)$ has the asymptotic behavior[2] for this time.

The integration process for Eq. (10) starts from $L=L_0$ and the point initial condition, which is given by the expression (12). The support of $\tilde{G}_L(\rho,t;\rho_0,t_0)$ expands, when $L-L_0$ increases. The characteristics multiply every step on L. In each point of the multiplying, the initial condition (12) is used and the definition of $\tilde{G}_L(\rho,t;\rho_0,t_0)$ for $t=t_0$ is completed through $\epsilon(L,\rho_0,t_0)$. This is a direct procedure. The inversion is obtained, if Eq. (10) is reintegrated (integrated in the inverse order) along the characteristics and $\epsilon(L,\rho_0,t_0)$ is defined through $\tilde{G}_L(\rho,t;\rho_0,t_0)$ for $t=t_0$. The inversion procedure is known as the ‘‘layer stripping’’ process.[2]

II. STATISTICAL MODEL

We suppose additionally the smallness ($|\epsilon(L,\rho,t)|\ll 1$) of the medium characteristic fluctuations, that will allow us to use the Markov apparatus[5–7] and to realize the diffusion approximation.[5–7] We assume that $\epsilon(L,\rho,t)$ is the Gaussian field with the statistical properties

$$\langle \epsilon(L,\rho,t)\rangle = 0 \quad \text{and} \quad \langle \epsilon(L,\rho,t)\epsilon(L',\rho',t')\rangle = B(L,\rho,t;L',\rho',t'). \tag{13}$$

For a characteristic functional of the backscattering $\Psi_L[\upsilon(\tau)] = \langle \exp[i\int d\tau R_L(\tau)\upsilon(\tau)]\rangle$, the equation is obtained by using (7)

$$\begin{aligned}\frac{\partial}{\partial L}\Psi_L[\upsilon(\tau)] &= \left\langle \left[i\int d\tau' \left[\frac{\partial}{\partial L}R_L(\tau')\right]\upsilon(\tau')\right]\exp\left[i\int d\tau R_L(\tau)\upsilon(\tau)\right]\right\rangle \\ &= -\int d\tau'\, \upsilon(\tau')[\hat{M}(q')+\hat{M}(q_0')]\frac{\delta}{\delta\upsilon(\tau')}\Psi_L[\upsilon(\tau)] \\ &\quad - i\int d\tau'\, \upsilon(\tau')\int dq_1\left[g_0(q',q_1)+\frac{\delta}{i\,\delta\upsilon(q',q_1)}\right] \\ &\quad \cdot \frac{\partial^2}{\partial t_1^2}\left[g_0(q_1,q_0')+\frac{\delta}{i\,\delta\upsilon(q_1,q_0')}\right]\left\langle \epsilon(L,q_1)\exp\left[i\int d\tau R_L(\tau)\upsilon(\tau)\right]\right\rangle,\end{aligned} \tag{14}$$

where the notation $\langle\cdots\rangle$ is the averaging on the realization set of the random value $\epsilon(L,q)$.

Note that the integral $\int d\tau R_L(\tau)\upsilon(\tau)$ is implied in the interval $[-\infty,\infty]$ of the variable t. We shall consider the divergent $R_L(\tau)$ and convergent $R_L^*(\tau)$ waves. The latter is found by Eq. (7) with the changes $\hat{M}(q)\to -\hat{M}(q)$ and $g_0(\tau)\to g_0^*(\tau)$, where $g_0^*(\tau)$ is the convergent wave in the free space. The convergent waves have the inversion course in time and may be placed formally on the negative semiaxis of t. The necessity of this expansion will be seen below.

In agreement with Furutsu–Novikov formula,[3,5]

$$\begin{aligned}\left\langle \epsilon(L,q_1)\exp\left[i\int d\tau R_L(\tau)\upsilon(\tau)\right]\right\rangle &= \int_{L_0}^{L}dx\int dq_2\, B(L,q_1;x,q_2)\left\langle \frac{\delta \exp[i\int d\tau R_L(\tau)\upsilon(\tau)]}{\delta\epsilon(x,q_2)}\right\rangle \\ &= \int_{L_0}^{L}dx\int dq_2\, B(L,q_1;x,q_2) \\ &\quad \times\left\langle i\int d\tau''\frac{\delta R_L(\tau'')}{\delta\epsilon(x,q_2)}\upsilon(\tau'')\exp\left[i\int d\tau R_L(\tau)\upsilon(\tau)\right]\right\rangle.\end{aligned} \tag{15}$$

Varying (7), we obtain the equation for the variational derivative,

$$\begin{aligned}\frac{\partial}{\partial L}\frac{\partial R_L(\tau'')}{\delta\epsilon(x,q_2)} &= -[\hat{M}(q'')+\hat{M}(q_0'')]\frac{\delta R_L(\tau'')}{\delta\epsilon(x,q_2)} - \int dq_3\frac{\delta R_L(q'',q_3)}{\delta\epsilon(x,q_2)}\frac{\partial^2}{\partial t_3^2}\epsilon(L,q_3)G_L(q_3,q_0'') \\ &\quad - \int dq_3\, G_L(q'',q_3)\frac{\partial^2}{\partial t_3^2}\epsilon(L,q_3)\frac{\delta R_L(q_3,q_0'')}{\delta\epsilon(x,q_2)},\end{aligned} \tag{16}$$

with the initial condition

$$\left.\frac{\delta R_L(\tau'')}{\delta\epsilon(x,q_2)}\right|_{L=x+0} = -\left[\frac{\partial^2}{\partial t_2^2}G_x(q'',q_2)\right]G_x(q_2,q_0'').$$

The comparison of Eq. (16) with (3) shows that the solution of (16) may be expressed though the solution of (3) as

$$\frac{\delta R_L(\tau'')}{\delta\epsilon(x,q_2)} = -\left[\frac{\partial^2}{\partial t_2^2} G(L,q'',x,q_2)\right] G(x,q_2;L,q_0'')$$

$$= -\hat{\Omega}_x^L(q'',\tilde{q})\left[\frac{\partial^2}{\partial t_2^2} G_x(\tilde{q},q_2)\right] G_x(q_2,\tilde{q}_0)\hat{\Omega}_x^L(\tilde{q}_0,q_0''), \tag{17}$$

where $\hat{\Omega}_x^L$ are the matriciants of the operators with the kernels, which may be defined from (16), a new value $G(L,q'';x,q_2)$ in the first line of (17) is the field on the right boundary of the layer, when the source takes place into the medium. The equation for this value may be produced like (3) by the invariant embedding method.[5] This equation differs from (3) by the operator $\hat{M}(\rho,t)$ instead of $\hat{M}(\rho_0,t_0)$. In the second of (17), the solution of (16) is presented in the symbolic operator form. The result (17) is well known for the one-dimensional problem.[5]

We see from (17) that the statistical problem is not local in general, and the closed equation for the backscattering field cannot be derived. The expression (17) demands that we consider two additional values (the field in the medium from the source on the right boundary and the field on the boundary from the source in the medium). Only for the case $B(L,q_1;x,q_2)=B(q_1;q_2)\delta(L-x)$, when $x=L$ in (17), the relation (17) yields the close equation for the backscattering. However, the δ-correlation model is not satisfactory for some real situations. The generalization of this model on the case of the nonzero correlation radius will be considered below.

If $|\epsilon(L,\rho,t)|\ll 1$, the internal field and the variational derivative vary not essentially on the size of the scatters, and we can take the solutions of Eqs. (16) and (7) in a zero approximation with respect to $\epsilon(L,q)$. Then we have for the variational derivative,

$$\frac{\delta R_L(\tau'')}{\delta\epsilon(x,q_2)} \approx -\exp\{-[\hat{M}(q'')+\hat{M}(q_0'')](L-x)\}$$

$$\cdot\left\{\frac{\partial^2}{\partial t_2^2}[g_0(q'',q_2)+\exp\{[\hat{M}(q'')+\hat{M}(q_2)](L-x)\}R_L(q'',q_2)]\right\}$$

$$\cdot[g_0(q_2,q_0'')+\exp\{[\hat{M}(q_2)+\hat{M}(q_0'')](L-x)\}R_L(q_2,q_0'')]. \tag{18}$$

This approach is associated with the perturbation theory[15] employed on the distance $L-x$ defined by the radius correlation of the medium, which is small in comparison with the size of the layer $L-L_0$. This approach is known as Chernov's local method[16] and allows us to obtain the close equation for the characteristic functional of the backscattering field,

$$\frac{\partial}{\partial L}\Psi_L[v(\tau)] = -\int d\tau'\, v(\tau')[\hat{M}(q')+\hat{M}(q_0')]\frac{\delta}{\delta v(\tau')}\Psi_L[v(\tau)]$$

$$-\int d\tau'\, v(\tau')\int dq_1\left\{\frac{\partial^2}{\partial t_1^2}\left[g_0(q',q_1)+\frac{\delta}{i\,\delta v(q',q_1)}\right]\right\}$$

$$\cdot\left[g_0(q_1,q_0')+\frac{\delta}{i\,\delta v(q_1,q_0')}\right]$$

$$\cdot\int_{L_0}^{L}dx\int dq_2\, B(L,q_1;x,q_2)\int d\tau''\, v(\tau'')\left\{\frac{\partial^2}{\partial t_2^2}\left[e^{-\hat{M}(q'')(L-x)}g_0(q'',q_2)\right.\right.$$

$$\left.\left.+e^{\hat{M}(,q_2)(L-x)}\frac{\delta}{i\,\delta v(q'',q_2)}\right]\right\}\cdot\left[e^{-\hat{M}(q_0'')(L-x)}g_0(q_2,q_0'')\right.$$

$$\left.+e^{\hat{M}(q_2,)(L-x)}\frac{\delta}{i\,\delta v(q_2,q_0'')}\right]\Psi_L[v(\tau)], \tag{19}$$

with the initial condition $\Psi_L[v(\tau)]|_{L=L_0}=1$. The operators $\hat{M}(\ ,q_2)$ and $\hat{M}(q_2,\ )$ in (19) are deferent by the action on the right and left arguments of the functions.

For the density of the probability in the backscattering functional space $P_L[\omega(\tau)]$, which is introduced here through the Fourier transformation,

$$\Psi_L[v(\tau)]=\int\ldots\int\prod_\tau d\omega(\tau)P_L[\omega(\tau)]\exp\left[i\int\tau\,\omega(\tau)v(\tau)\right], \tag{20}$$

the Fokker–Planck equation[3,5] may be obtained from (19),

$$\begin{aligned}\frac{\partial}{\partial L}P_L[\omega(\tau)]=&\int d\tau'\frac{\delta}{\delta\omega(\tau')}[\hat{M}(q')+\hat{M}(q_0')]\omega(\tau')P_L[\omega(\tau)]\\&+\int d\tau'\frac{\delta}{\delta\omega(\tau')}\int dq_1\left\{\frac{\partial^2}{\partial t_1^2}[g_0(q',q_1)+\omega(q',q_1)]\right\}\\&\cdot[g_0(q_1,q_0')+\omega(q_1,q_0')]\cdot\int_{L_0}^{L}dx\int dq_2\,B(L,q_1;x,q_2)\\&\times\int d\tau''\frac{\delta}{\delta\omega(\tau'')}\left\{\frac{\partial^2}{\partial t_2^2}[e^{-\hat{M}(q'')(L-x)}g_0(q'',q_2)+e^{\hat{M}(,q_2)(L-x)}\omega(q'',q_2)]\right\}\\&\cdot[e^{-\hat{M}(q_0'')(L-x)}g_0(q_2,q_0'')+e^{\hat{M}(q_2,)(L-x)}\omega(q_2,q_0'')]P_L[\omega(\tau)],\end{aligned} \tag{21}$$

with the obvious initial condition $P_L[\omega(\tau)]|_{L=L_0}=\Pi_\tau\delta(\omega(\tau))$.

With respect to the type of the latter equation, the considered above approach is known in statistical theory as the diffusion approximation.[5–7] The application condition of this method may be mounted, if we return to second and third lines of the expression (17), and notice that the nonaccounted for above two terms in (17) are the additional variation of the wave ''phase'' on the scatter. This variation is small in comparison with the unit that is really, if $|\epsilon(L,\rho,t)|\ll 1$.

Otherwise, Eq. (21) may be obtained from (7) through the Liouville equation[3,5] for the value $\varphi_L[\omega(\tau)]=\Pi_\tau\delta(R_L(\tau)-\omega(\tau))$. The functional $P_L[\omega(\tau)]=\langle\varphi_L[\omega(\tau)]\rangle$ will satisfy this equation after the averaging with respect to $\epsilon(L,\rho,t)$, if the approximation (18) will be used.

The statistical equation for $\tilde{R}_L(\rho,t;\rho_0,t_0)$ may be yielded by the change $\int dt'\to\int_{t_0}^{t}dt'+\int_{-t}^{t_0}dt'$ in (19) and (21) as Eqs. (4) and (10) are different only by the limits of the integral, when $|\epsilon(L,q)|\ll 1$.

III. THE STATISTICAL CHARACTERISTICS OF THE MEDIUM

Equation (21) for the value $P_L[\omega(\tau)]$ consists of 16 terms that are proportionally the correlation function of the medium. They may be collected in four groups with the deferent operator coefficients that are the Laplace transformation of the correlation function with the parameter of the transformation in the form of the operator,

$$\begin{aligned}\hat{K}_i=&\int_{L_0}^{L}dx\,B(L,q_1;x,q_2)\frac{\partial^2}{\partial t_2^2}e^{\hat{N}_i(L-x)},\quad i=1,2,3,4;\\\hat{N}_1=&-\hat{M}(q'')-\hat{M}(q_0''),\quad \hat{N}_2=-\hat{M}(q'')+\hat{M}(q_2,),\\\hat{N}_3=&\hat{M}(,q_2)-\hat{M}(q_0''),\quad \hat{N}_4=\hat{M}(,q_2)+\hat{M}(q_2,).\end{aligned} \tag{22}$$

If the Fourier transformation with respect to the difference argument q_1-q_2 is performed in (22), these four operator coefficients are the Born's expression[15] for the scattering characteristic of the medium (in first order of the perturbation theory). They are named in the suitable calibration as an indicatrix of the scattering or a phase function. The value $\hat{K}_1$ describes the scattering of the wave from the right hemisphere to back, $\hat{K}_4$—from the left hemisphere to back, $\hat{K}_2$ and $\hat{K}_3$ correspond to the wave scattering with the change of the hemisphere (from right to left and from left to right). The scattering takes place in the thin layer $L-x$, which is equal approximately to the correlation radius of the medium nonhomogeneities and is little more than $L-L_0$; therefore we can put $L_0=-\infty$ in (21) and (22). If the graphs for $g_0(q,q_0)$ and $\omega(q,q_0)$ are introduced, all 16 terms under the integral in (21) may be presented in the form of the Feynman four-tail diagram.[15]

For the isotropic medium, $\hat{K}_1=\hat{K}_4$ and $\hat{K}_2=\hat{K}_3$. For the medium with the isotropic scattering, all coefficients are equal. The medium with the point and δ correlation in time scatters has $B(L,q_1;x,q_2)=\sigma^2(L,q_1)\,\delta(L-x)\,\delta(q_1-q_2)$. For these scatters and the statistical layered medium, $\sigma^2(L,q_1)=\sigma^2(L)$, in the statistical homogeneous one, $\sigma^2(L,q_1)=\text{const}$ and so on, the classification of the media may be continued through the properties of the correlation function.

It is important for the statistical inverse problem. If $\hat{K}_1$ is defined by some way, the rest coefficients $\hat{K}_i$ may be known through the inverse and direct Laplace transformation too. For the determination of the correlation function through the statistical properties of the backscattering, Eq. (21) will be considered for the short time. The scales of the problem have to be discussed before.

The correlation function $B(L,q_1;x,q_2)$ has two sets of the scales l_x,l_y,l_z,τ and $\tilde{l}_x,\tilde{l}_y,\tilde{l}_z,\tilde{\tau}$. The first is that the set of the correlation radiuses measured the deference of the arguments $L-x$ and q_1-q_2. The second is the sizes of the variables L and q_1 (the spatial-time alterations of l_x,l_y,l_z,τ). The first set with the intensity of the medium fluctuations defines the third scale set of the problem, the lengths of the scattering l_S and l_{SB}, which is discussed above. We assume that the second scale is less than the third (the inverse case corresponds to the condition for the statistical quasihomogenous medium) and bigger than the first.

The term ''short time'' is associated with the value $t-t_0$ and the appropriate depth of the wave penetration into the medium, which are intermediate between the first and second scales of the problem. For this time, the same simplifications may be carried out in (21).

The short time corresponds to the single scattering; then the change $g_0+\omega\rightarrow g_0$ can be made in (21). The short time is less than the second scale of the problem, therefore the diffusion equation (21) has the constant coefficients in the chosen time interval. For the short time, this equation has the intermediate stationary ($\partial P_L/L\approx 0$) solution in the form of the Gaussian distribution defined by the next expression

$$\int d\tau'\frac{\delta}{\delta\omega(\tau')}[\hat{M}(q')+\hat{M}(q_0')]\omega(\tau')P_L[\omega(\tau)]+\int d\tau'\frac{\delta}{\delta\omega(\tau')}\int dq_1\left[\frac{\partial^2}{\partial t_1^2}g_0(q',q_1)\right]$$
$$\cdot[g_0(q_1,q_0')]\cdot\int_{-\infty}^{L}dx\int dq_2\,B(L,q_1;x,q_2)\int d\tau''\frac{\delta}{\delta\omega(\tau'')}\left[\frac{\partial^2}{\partial t_2^2}e^{-\hat{M}(q'')(L-x)}g_0(q'',q_2)\right]$$
$$\cdot[e^{-\hat{M}(q_0'')(L-x)}g_0(q_2,q_0'')]P_L[\omega(\tau)]\approx 0, \tag{23}$$

where the change $\int dt'\rightarrow\int_{t_0}^{t}dt'+\int_{-t}^{t_0}dt'$ is implied and the value $t-t_0$ corresponds to the short time interval.

The physical meaning of (23) may be explained. For the short time after the fast field variations, the solution of (21) takes the intermediate stationary form defined by (23).

This equation gives the locale relation between the correlation function and the statistical characteristics of the backscattering. The found relation (23) is very important for the inverse procedure in the statistical problem and plays the role of Eq. (11), which gives the point relation between the field and the medium characteristic in the determinant problem.

IV. THE ADDITIONAL AVERAGING OF THE FAST FIELD VARIATIONS

The solution of the evolution equation (21) contains the detailed statistical information that may be surplus for some cases, for example, if the average field in the point is not interested.

The more simple evolution equation for the statistical problem may be obtained from (21), if the additional averaging of the fast field variation on the first problem scale is carried out there. This operation is the passage to the slowly varying amplitude of the backscattering. It may be realized formally so.

For the simplification of the obtained above expressions, the normalized backscattering $\mathfrak{R}_L(\tau)=2c^2M(q)R_L(\tau)$ has to be introduced. This value satisfies Eq. (7) with the changes $g_0(q,q_0)\rightarrow\delta(q-q_0)$ and $\epsilon(L,q_1)\rightarrow\tilde{\epsilon}(L,q_1)=[\partial^2\epsilon(L,q_1)/\partial t_1^2]\hat{M}^{-1}(q_1)/2c^2$, where $\tilde{\epsilon}(L,q_1)$ is the effective characteristic of the medium. Presenting the normalized backscattering in the form $\mathfrak{R}_L(\tau)=\exp\{-[\hat{M}(q)+\hat{M}(q_0)](L-L_0)\}\overline{\mathfrak{R}}_L(\tau)$ and repeating the derivation of the Fokker–Planck equation, we come to the new expression instead (21),

$$\begin{aligned}\frac{\partial}{\partial L}P_L[\overline{\omega}(\tau)]=&\int d\tau'\frac{\delta}{\delta\overline{\omega}(\tau')}\int dq_1[e^{\hat{M}(q')(L-L_0)}\delta(q'-q_1)+e^{-\hat{M}(,q_1)(L-L_0)}\overline{\omega}(q',q_1)]\\&\cdot[e^{\hat{M}(q_0')(L-L_0)}\delta(q_1-q_0')+e^{-\hat{M}(q_1,)(L-L_0)}\overline{\omega}(q_1,q_0')]\\&\cdot\int_{L_0}^{L}dx\int dq_2\,\tilde{B}(L,q_1;x,q_2)\int d\tau''\frac{\delta}{\delta\overline{\omega}(\tau'')}\cdot[e^{-\hat{M}(q'')(L-x)}e^{\hat{M}(q'')(L-L_0)}\delta(q''-q_2)\\&+e^{\hat{M}(,q_2)(L-x)}e^{-\hat{M}(,q_2)(L-L_0)}\overline{\omega}(q'',q_2)]\cdot[e^{-\hat{M}(q_0'')(L-x)}e^{\hat{M}(q_0'')(L-L_0)}\delta(q_2-q_0'')\\&+e^{\hat{M}(q_2,)(L-x)}e^{-\hat{M}(q_2,)(L-L_0)}\overline{\omega}(q_2,q_0'')]P_L[\overline{\omega}(\tau)],\end{aligned}\tag{24}$$

where $\overline{\omega}(\tau)$ is the parameter of the distribution of the slowly vary normalized backscattering amplitude $\overline{\mathfrak{R}}_L(\tau)$ and $\tilde{B}(L,q_1;x,q_2)=\langle\tilde{\epsilon}(L,q_1)\tilde{\epsilon}(x,q_2)\rangle$.

Having in mind that $P_L[\overline{\omega}(\tau)]$ varies slowly on the first scale of the problem, and averaging additionally the expression (24) by the operation $(1/\Delta L)\int_L^{L+\Delta L}dL'$, where ΔL is bigger than the first problem scale, we obtain the equation

$$\begin{aligned}\frac{\partial}{\partial L}P_L[\overline{\omega}(\tau)]=&\int d\tau'\int d\tau''\int_{L_0}^{L}dx\int dq_1\int dq_2\,\hat{B}(L,q_1;x,q_2)\\&\times\frac{\delta}{\delta\overline{\omega}(\tau')}\Bigg(\frac{\overline{\omega}(q',q_1)\,\delta(q_1-q_0')}{\delta^*(q_0''-\check{q}_2)\,\delta^*(\hat{q}_2-q_0'')}+\frac{\overline{\omega}(q',q_1)\,\overline{\omega}(q_1,q_0')}{\delta^*(q_0''-\check{q}_2)\,\overline{\omega}^*(\check{q}_2,q_0'')}\\&+\frac{\delta(q'-q_1)\,\delta(q_1-q_0')}{\delta(q_0''-\hat{q}_2)\,\overline{\omega}(\hat{q}_2,q_0'')}+\frac{\delta(q'-q_1)\,\overline{\omega}(q_1,q_0')}{\delta(q_0''-\hat{q}_2)\,\delta(\check{q}_2-q_0'')}+\frac{\delta(q'-q_1)\,\overline{\omega}(q_1,q_0')}{\delta^*(q''-\hat{q}_2)\,\delta^*(\check{q}_2-q'')}\\&+\frac{\overline{\omega}(q',q_1)\,\overline{\omega}(q_1,q_0')}{\overline{\omega}^*(q'',\check{q}_2)\,\delta^*(\check{q}_2-q'')}+\frac{\delta(q'-q_1)\,\delta(q_1-q_0')}{\overline{\omega}(q'',\hat{q}_2)\,\delta(\hat{q}_2-q'')}\\&+\frac{\overline{\omega}(q',q_1)\,\delta(q_1-q_0')}{\delta(q''-\hat{q}_2)\,\delta(\check{q}_2-q'')}\Bigg)P_L[\overline{\omega}(\tau)]\\&+\int d\tau'\int d\tau''\int_{L_0}^{L}dx\int dq_1\int dq_2\,\tilde{B}(L,q_1;x,q_2)\\&\times\frac{\delta}{\delta\overline{\omega}(\tau')}\Bigg(\frac{\delta(q'-q_1)\,\delta(q_1-q_0')}{\delta^*(q''-\hat{q}_2)\,\delta^*(\hat{q}_2-q_0'')}+\frac{\delta(q'-q_1)\,\overline{\omega}(q_1,q_0')}{\delta^*(q''-\hat{q}_2)\,\overline{\omega}^*(\check{q}_2,q_0'')}\end{aligned}$$

$$+\frac{\overline{\omega}(q',q_1)\,\delta(q_1-q_0')}{\overline{\omega}^*(q'',\check{q}_2)\,\delta^*(\hat{q}_2-q_0'')}+\frac{\overline{\omega}(q',q_1)\,\overline{\omega}(q_1,q_0')}{\overline{\omega}^*(q'',\check{q}_2)\,\overline{\omega}^*(\hat{q}_2,q_0'')}+\frac{\delta(q'-q_1)\,\delta(q_1-q_0')}{\overline{\omega}(q'',\hat{q}_2)\,\overline{\omega}(\hat{q}_2,q_0'')}$$

$$+\frac{\overline{\omega}(q',q_1)\,\overline{\omega}(q_1,q_0')}{\delta(q''-\check{q}_2)\,\delta(\check{q}_2-q_0'')}+\frac{\delta(q'-q_1)\,\overline{\omega}(q_1,q_0')}{\overline{\omega}(q'',\hat{q}_2)\,\delta(\check{q}_2-q_0'')}$$

$$+\left.\frac{\overline{\omega}(q',q_1)\,\delta(q_1-q_0')}{\delta(q''-\check{q}_2)\,\overline{\omega}(\hat{q}_2,q_0'')}\right)\frac{\delta}{\delta\overline{\omega}(\tau'')}P_L[\overline{\omega}(\tau)], \tag{25}$$

where $f_1f_2/f_3f_4=f_1f_2f_3f_4$ is the compact form of the product notation, which shows the connection of (25) with (24) and the analogy of the equation coefficients with the Feynman diagrams, $f(q,\hat{q}_2)=e^{\hat{M}(,q_2)(L-x)}f(q,q_2)$ and $f(q,\check{q}_2)=e^{-\hat{M}(,q_2)(L-x)}f(q,q_2)$ are the actions of the displacement operator, f and f^* are the divergent and convergent waves.

The simplification of (25) depends from the model of the medium characteristic $\widetilde{B}(L,q_1;x,q_2)$. This function has the cofactor $\delta(\rho_1-\rho_2)$ for the layered medium and has the cofactor $\delta(t_1-t_2)$ for the time δ-correlation medium; then the some integrals in the coefficients of Eq. (25) may be calculated. The passage to the plane waves allows us to get rid of the dependence on ρ_1 and ρ_2.

For the three-dimensional medium, we have to average additionally Eq. (25) on ρ and t. The biquadratic coefficients in Eq. (25) are kept in the result (the others fall out)

$$\frac{\partial}{\partial L}P_L[\overline{\omega}(\tau)]=+\int d\tau'\int d\tau''\int_{L_0}^{L}dx\int dq_1\int dq_2\,\widetilde{B}(L,q_1;x,q_2)$$

$$\times\frac{\delta}{\delta\overline{\omega}(\tau')}\left(\frac{\delta(q'-q_1)\,\delta(q_1-q_0')}{\delta^*(q''-\hat{q}_2)\,\delta^*(\hat{q}_2-q_0'')}+\frac{\delta(q'-q_1)\,\overline{\omega}(q_1,q_0')}{\delta^*(q''-\hat{q}_2)\,\overline{\omega}^*(\check{q}_2,q_0'')}\right.$$

$$\left.+\frac{\overline{\omega}(q',q_1)\,\delta(q_1-q_0')}{\overline{\omega}^*(q'',\check{q}_2)\,\delta^*(\hat{q}_2-q_0'')}+\frac{\overline{\omega}(q',q_1)\,\overline{\omega}(q_1,q_0')}{\overline{\omega}^*(q'',\check{q}_2)\,\overline{\omega}^*(\check{q}_2,q_0'')}\right)\frac{\delta}{\delta\overline{\omega}(\tau'')}P_L[\overline{\omega}(\tau)]. \tag{26}$$

Note, for the three-dimensional medium, $q_1\approx q_2$ and $\tau'\approx\tau''$. It means that the expression in the right part of (26) is approximately diagonal and the consideration of the many-dimensional problem may be reduced to the analysis of the one-dimensional case. This fact was noticed earlier in the solution of the backscattering moment equations[14,15] (the solution of the many-dimensional problem was obtained as an analytical continuation of the one-dimensional result) and may be the base for the digital simulation of Eq. (26).

We have to remark that the passage to the layered medium in (26) is not possible, as this action is not permutable with the averaging on ρ.

V. STATISTICAL INVERSE PROBLEM

The local relation (23) allows integrating or reintegrating Eq. (21) along the characteristics of the transport operator, and how it is made for Eq. (10) with the use of (11). Equation (10) and the local relation (23) in a statistical problem play accordingly the roles of Eq. (11) and the point relation (11) in the determinant one. The direct and inverse procedures are realized the same in both cases. The difference is in the type of the evolution equations (10) and (21).

Equation (25) or (26) for the slowly varying statistical characteristic of the backscattering may be used instead of (21), if the weak effects of the fast field variations are fallen out. As the evolution equations for the direct and inverse problem investigation, the equations for the statistical moments of the backscattering field may be chosen, which will be considered below.

VI. THE MOMENT EQUATIONS

The characteristic functional of the backscattering may be decomposed in the Taylor series,

$$\Psi_L[\upsilon(\tau)]=\sum_{n=0}^{\infty}\frac{i^n}{n!}\int_{\dots}\int d\tau_1\cdots d\tau_n\, M_n(L;\tau_1,\dots,\tau_n)\upsilon(\tau_1)\cdots\upsilon(\tau_n), \tag{27}$$

with the coefficients represented through the density of the probability in the backscattering functional space as

$$M_n(L;\tau_1,\dots,\tau_n)=\left\langle \prod_{i=1}^{n} R_L(\tau_i)\right\rangle=\int_{\dots}\int \prod_{\tau} d\omega(\tau)\prod_{i=1}^{n}\omega(\tau_i)P_L[\omega(\tau)], \tag{28}$$

and named as the statistical moments of the backscattering. The equations are obtained from (21),

$$\begin{aligned}\frac{\partial}{\partial L}M_n(L;\tau_1,\dots,\tau_n)=&-\sum_{i=1}^{n}[\hat{M}(q^i)+\hat{M}(q_0^i)]M_n(L;\tau_1,\dots,\tau_n)\\&+\int_{\dots}\int\prod_{\tau} d\omega(\tau)\sum_{j=1}^{n}\int dq_1\left\{\frac{\partial^2}{\partial t_1^2}[g_0(q^j,q_1)+\omega(q^j,q_1)]\right\}\cdot[g_0(q_1,q_0^j)\\&+\omega(q_1,q_0^j)]\cdot\int_{L_0}^{L}dx\int dq_2\, B(L,q_1;x,q_2)\sum_{\substack{k=1\\k\neq j}}^{n}\left\{\frac{\partial^2}{\partial t_2^2}[e^{-\hat{M}(q^k)(L-x)}g_0(q^k,q_2)\right.\\&\left.+e^{\hat{M}(,q_2)(L-x)}\omega(q^k,q_2)]\right\}\cdot[e^{-\hat{M}(q_0^k)(L-x)}g_0(q_2,q_0^k)\\&+e^{\hat{M}(q_2,)(L-x)}\omega(q_2,q_0^k)]\frac{\Pi_{i=1}^{n}\omega(\tau_i)}{\omega(\tau_j)\omega(\tau_k)}P_L[\omega(\tau)].\end{aligned} \tag{29}$$

The expression under the integral in (29) gives 32 terms, which may be collected into five groups:

$$\begin{aligned}\frac{\partial}{\partial L}M_n(L;\tau_1,\dots,\tau_n)=&-\hat{Q}_n(\tau_1,\dots,\tau_n)M_n(L;\tau_1,\dots,\tau_n)\\&+\hat{D}_{n-2}(L;\tau_1,\dots,\tau_{n-2})M_{n-2}(L;\tau_1,\dots,\tau_{n-2})\\&+\hat{D}_{n-1}(L;\tau_1,\dots,\tau_{n-1})M_{n-1}(L;\tau_1,\dots,\tau_{n-1})\\&+\hat{D}_{n}(L;\tau_1,\dots,\tau_{n})M_{n}(L;\tau_1,\dots,\tau_{n})\\&+\hat{D}_{n+1}(L;\tau_1,\dots,\tau_{n+1})M_{n+1}(L;\tau_1,\dots,\tau_{n+1})\\&+\hat{D}_{n+2}(L;\tau_1,\dots,\tau_{n+2})M_{n+2}(L;\tau_1,\dots,\tau_{n+2}),\end{aligned} \tag{30}$$

where $\hat{Q}_n=\Sigma_{i=1}^{n}[\hat{M}(q^i)+\hat{M}(q_0^i)]$, $\hat{D}_n(L;\tau_1,\dots,\tau_n)$ is the operator coefficients, the explicit expressions of which may be yielded from (29). The initial conditions for (30) are $M_0=1$ and $M_n=0$, if $n>0$.

Eq. (30) can be rewritten in the matrix form

$$\frac{\partial}{\partial L}\mathbf{M}_L=\hat{\mathbf{A}}_L\mathbf{M}_L+\mathbf{B}_L,\quad \hat{\mathbf{A}}_L=\hat{\mathbf{Q}}+\hat{\mathbf{D}}_L,$$

$$\mathbf{M}_L=\begin{pmatrix} M_1\\ M_2\\ \cdots\\ M_n\\ \cdots \end{pmatrix},\quad \mathbf{B}_L=\begin{pmatrix} B_{1-1}\\ B_{2-2}\\ 0\\ 0\\ \cdots \end{pmatrix},\quad \hat{\mathbf{Q}}=\begin{pmatrix} \hat{Q}_1 & & \dots & 0 & \dots & & \\ 0 & \hat{Q}_2 & \dots & 0 & \dots & & \\ & & \dots & & & & \\ \dots & 0 & \dots & \hat{Q}_n & \dots & 0 & \dots \\ & & & & \dots & & \end{pmatrix},$$

$$\hat{\mathbf{D}}_L=\begin{pmatrix} \hat{D}_1 & \hat{D}_{1+1} & \hat{D}_{1+2} & \dots 0 \dots & & & \\ \hat{D}_{2-1} & \hat{D}_2 & \hat{D}_{2+1} & \hat{D}_{2+2} & \dots 0 \dots & & \\ \hat{D}_{3-2} & \hat{D}_{3-1} & \hat{D}_3 & \hat{D}_{3+1} & \hat{D}_{3+2} & \dots 0 \dots & \\ & & & \dots & & & \\ \dots 0 \dots & \hat{D}_{n-2} & \hat{D}_{n-1} & \hat{D}_n & \hat{D}_{n+1} & \hat{D}_{n+2} & \dots 0 \dots \\ & & & & \dots & & \end{pmatrix}, \tag{31}$$

where $\mathbf{B}_L$ is the vector function, $B_{1-1}=\hat{D}_{1-1}M_0$, $B_{2-2}=\hat{D}_{2-2}M_0$ are the known values expressed through the correlation function of the medium, $M_0=1$. The numeration for the elements of the five diagonal D matrix (31) is chosen as in Eqs. (30).

VII. EQUATIONS FOR THE SLOWLY EVOLVED MOMENTS

After the additional averaging of the fast field variations on L in (31), the equations for the even slowly evolved moments $\overline{M}_{2n}$ ($n=0,1,...$) are obtained. This result may be yielded from (25) as (31) was produced from (21). If the input is (31) we fall out the fast-evolved moments there. It means to put formally $\hat{\mathbf{Q}}=0$ and to delete the $n-1$ and $n+1$ diagonals of the D matrix and B_{1-1} in (31). For the many-dimensional medium, after the additional averaging of the fast field variations on ρ and t, the even coefficients of the D matrix are simplified essentially otherwise, compare (25) and (26). Equations for the slowly moments may be written as

$$\begin{aligned}\frac{\partial}{\partial L}\overline{M}_{2n}(L;\tau_2,\dots,\tau_{2n})=&\hat{D}_{2n-2}(L;\tau_1,\dots,\tau_{2n-2})\overline{M}_{2n-2}(L;\tau_1,\dots,\tau_{2n-2})\\ &+\hat{D}_{2n}(L;\tau_1,\dots,\tau_{2n})\overline{M}_{2n}(L;\tau_1,\dots,\tau_{2n})\\ &+\hat{D}_{2n+2}(L;\tau_1,\dots,\tau_{2n+2})\overline{M}_{2n+2}(L;\tau_1,\dots,\tau_{2n+2}).\end{aligned} \tag{32}$$

VIII. TRANSPORT EQUATION

The expression (32) for second moment of the backscattering is implied as the transport equation,

$$\begin{aligned}\frac{\partial}{\partial L}\overline{M}_2(L;\tau_1,\tau_2)=&B_{2-2}(L;\tau_1,\tau_2)+\hat{D}_2(L;\tau_1,\tau_2)\overline{M}_2(L;\tau_1,\tau_2)\\ &+\hat{D}_{2+2}(L;\tau_1,\dots,\tau_4)\overline{M}_4(L;\tau_1,\dots,\tau_4).\end{aligned} \tag{33}$$

The ''Gaussian hypothesis'' is the assumption that the relation between $\overline{M}_4$ and $\overline{M}_2$ is as for the Gaussian field. It is formally the noncorrect supposition about the absence of the correlation between the couples of the counterwaves, $\overline{M}_4=\langle RR^*RR^*\rangle\approx\langle RR^*\rangle\langle RR^*\rangle=\overline{M}_2\overline{M}_2$. If the approximate relation $\overline{M}_4\approx\overline{M}_2\overline{M}_2$ is put in (33), a closed equation for $\overline{M}_2$ is obtained,

$$\frac{\partial}{\partial L}\overline{M}_2(L;\tau_1,\tau_2)=B_{2-2}(L;\tau_1,\tau_2)+\hat{D}_2(L;\tau_1,\tau_2)\overline{M}_2(L;\tau_1,\tau_2)$$
$$+\hat{D}_{2+2}(L;\tau_1,\ldots,\tau_4)\overline{M}_2(L;\tau_1,\tau_2)\overline{M}_2(L;\tau_3,\tau_4), \tag{34}$$

where the operator $\hat{D}_{2+2}$ sorts out the couples of τ and sums; see (29).

Equation (34) is the statistical analog of the phenomenological transport equation[17] written for the backscattering wave. As an example, this equation was used for the layered media.[9] Its solution showed that the error of such an approach grows in comparison with the strong statistical theory as $\sqrt{t}$ in the multiple backscattering regimes, where the correlation between the couples of the counterwaves plays the essential role, and it cannot be thrown. However, Eq. (34) may be used efficiently for the single backscattering.

Note that Eq. (34) is quadratic on $\overline{M}_2$ and is formally the determinant equation (4), or (7) and (10). It mean that we can solve the statistical inverse problem as the determinant one by the use of Eq. (34) and the locale relation from (23) for the single backscattering regime.

For the further discussion of the moments, let us consider the solution for the equations.

IX. SOLUTION OF THE MOMENT EQUATIONS

The solution of (31) or (32) may be presented through a matriciant,

$$\hat{\Omega}_{L_0}^{L}=E+\int_{L_0}^{L}dL'\,\hat{\mathbf{A}}_{L'}+\int_{L_0}^{L}dL'\,\hat{\mathbf{A}}_{L'}\int_{L_0}^{L'}dL''\,\hat{\mathbf{A}}_{L''}+\cdots,$$

where E is the unit matrix, and $\hat{\mathbf{A}}_L$ is defined in (31). With the use of a Cauchy matrix $\hat{\mathbf{K}}(L,L')=\hat{\Omega}_{L_0}^{L}[\hat{\Omega}_{L_0}^{L'}]^{-1}$ and with respect to the zero initial condition for R_L, the solution of the moment equations (31) may be written as

$$\mathbf{M}_L=\int_{L_0}^{L}dL'\,\hat{\mathbf{K}}(L,L')\mathbf{B}_{L'}. \tag{35}$$

The solution of the slow moment equations (32) has the form of (35) with the simplification of $\hat{\mathbf{A}}_L$ and $\mathbf{B}_L$, as discussed above.

Every line of (35) with the component $M_n(L,\tau_1,\ldots,\tau_n)$ in the left part may be considered as the equation for the statistical characteristic of the medium $B(L,q_1;x,q_2)$, which is included in $\hat{\mathbf{A}}_L$ and $\mathbf{B}_L$. These equations are integrodifferential and nonline with respect to the unknown function $B(L,q_1;x,q_2)$. Any from them may be used for the solving of the inverse problem, for the determination of $B(L,q_1;x,q_2)$, in principal. It means that the any moment contains the full information about the random medium. And more, any moment may be expressed through any other by the use of (35), if the inverse operator to the corresponding line of the expression (35) will be yielded. The all moments are tied stringently between themselves, and this connection is defined by the properties of the medium. As an example, the connection between the slow moments for the statistic homogeneous media can be defined easy from (32) putting $\partial\overline{M}_{2n}/L=0$ (the stationary solution exists for these media) and obtaining the recurrent relation there. For the full system of the moments, it cannot be yielded since five, not three, moments are tied in (30).

We can yield the exact closed statistical transport equation expressing $\overline{M}_4$ through $\overline{M}_2$ by use of (35) and substituting it into (33). However, the relation between $\overline{M}_4$ and $\overline{M}_2$ is dependent on $B(L,q_1;x,q_2)$ in the all layers $[L_0,L]$, and the transport equation will be not local and will not be comfortable for the inverse procedure in the form of the ‘‘layer stripping’’ process[2] as in (10), (21), (25), (26), (30), and (32). The first approximation of the exact statistical transport equation is (34), which is comfortable for the ‘‘layer stripping’’ process,[2] but it is not correct for the multiple backscattering.

If the application for the lidar remote sounding is had in mind, the second line of (35) can be named as the lidar equation.[18]

As a result, if we know one of the moments, using (35) we can solve the inverse problem in principal and reconstruct the full statistic information of the backscattering (the characteristic functional). What the moment (first, second, or any more) we have to select for the solving of the inverse problem? It will be decided. In general, all the moments are not different in this meaning. The second slow moment $\langle RR^* \rangle$ can be chosen from the convenience, because it is the energy characteristic of the field and it can be measured directly, it is described in more simple equations than the full moment system.

In the nondissipative medium, which is the subject of this paper, one of the moments has to be known for the solving of the inverse problem. In the opposite case, it is not enough. Two moments and two equations (35) are necessary for the determination of the statistical and dissipative medium characteristics simultaneously. It was shown in the experimental work[19] with use of the statistical solution.[10]

X. THE FIRST MOMENT OF THE FIELD

It is of interest to consider $\langle R \rangle$ as one of the fast-varied statistical moments and to understand their role. Using (12) we can compare the first moment with the second in the initial time $t = t_0$. For $|\epsilon| \ll 1$, $r_L(\rho_0, t_0) \approx [-\epsilon(L,\rho_0,t_0) + \epsilon^2(L,\rho_0,t_0)/2]/8c$ from (12). Whence $\langle r_L \rangle \approx \langle \epsilon^2 \rangle /16c$ [the nonline form of (12) leads to the detection effect] and $\langle r_L^2 \rangle \approx \langle \epsilon^2 \rangle/(8c)^2$. The energy of the average field is less than the average energy of the field, $\langle r_L \rangle^2 \ll \langle r_L^2 \rangle$. In this sense, the fast statistical moments are associated with the weak random effects and the additional averaging of the fast variation of the field is justified.

XI. CONCLUSION

The statistical characteristics of the backscattering and the relation with the parameters of the random medium were considered above. Six ways of the inverse problem solving were shown (three with use of the full statistical characteristics system and three with the slow varied one). In every three, the inverse procedures are the ''layer stripping'' process[2] for the Fokker–Planck equation (21) or (25) and the same process for the moment equations (30) or (32) as the two equivalent approach and the inversion of the moment equations solution (35) as the alternative method.

The initial experimental data for the ''layer stripping'' process are the density of the probability in the backscattering functional space or the respective system of the statistical moments. Let us consider the case with the slow varied statistical characteristics $P_L[\overline{\omega(\tau)}]$ and $\overline{M}_{2n} = \langle \Pi_{i=1,j=1}^{n} \bar{\mathfrak{R}}_L(q_i, q_{0,i}), \bar{\mathfrak{R}}_L^*(q_j, q_{0,j}) \rangle$. It requires the measurement of the stochastic value $\bar{\mathfrak{R}}_L(q_i, q_{0,i}) \bar{\mathfrak{R}}_L^*(q_j, q_{0,j})$. The Fourier transformation of which with respect to $q_i - q_j$ and $q_{0,i} - q_{0,j}$ is the stochastic spectral transfer function from the point $(q_{0,i} + q_{0,j})/2$ to the point $(q_i + q_j)/2$, which may be measured[20] by the multiple canal analyzer. The realizations of $\bar{\mathfrak{R}}_L(q_i, q_{0,i}) \bar{\mathfrak{R}}_L^*(q_j, q_{0,j})$ are reproduced from the measured spectral data though the inverse Fourier transformation. The averaging on the set of the $\bar{\mathfrak{R}}_L(q_i, q_{0,i}) \bar{\mathfrak{R}}_L^*(q_j, q_{0,j})$ realizations gives $\overline{M}_{2n}$ and $P_L[\overline{\omega(\tau)}]$. In the result, the ''layer stripping'' process may be used.

As the alternative method, $\overline{M}_2$ is employed for the inversion of the moment equation solution (35).

ACKNOWLEDGMENTS

The author thanks Professor V. I. Klyatskin for the helpful discussion of the results and Professor O. A. Bukin for the interest to the experimental applications, which stimulated this research. This work was supported by Russian Fund of the Basic Research (Grant No. 97-02-17138).

[1] B. M. Shevtsov, J. Math. Phys. **37**, 6130 (1996).
[2] V. H. Weston, Inverse Probl. **8**, 919 (1992).
[3] V. I. Klyatskin, Prog. Opt. **33**, 1 (1994).
[4] M. Asch, W. Kohler, G. Papanicolaou, M. Postel, and B. White, Waves Random Media **6**, 293 (1996).
[5] V. I. Klyatskin, *Imbedding Method in the Wave Propagation Theory* (Nauka, Moscow, 1986).
[6] B. M. Shevtsov, Izv. Vyssh. Uchebn. Zaved. Radiofiz. **28**, 717 (1985).
[7] W. Kohler, G. Papanicolaou, and B. White, Lect. Appl. Math. **27**, 487 (1991).
[8] O. A. Bukin, A. Yu. Major, A. N. Pavlov, B. M. Shevtsov, and E. D. Kholodkevich, Int. J. Remote Sens. **19**, 707 (1998).
[9] B. M. Shevtsov, Izv. Vyssh. Uchebn. Zaved. Radiofiz. **25**, 1032 (1982).
[10] B. M. Shevtsov, Izv. Vyssh. Uchebn. Zaved. Radiofiz. **26**, 434 (1983).
[11] V. S. Vladimirov, *Generalized Functions in the Mathematical Physics* (Nauka, Moscow, 1979).
[12] V. I. Tatarskii, *The Effects of the Turbulent Atmosphere on Wave Propagation* (National Technical Information Service, Springfield, VA, 1971).
[13] Yu. A. Kravtsov and A. I. Saichev, Sov. Phys. Usp. **25**, 494 (1982).
[14] Yu. A. Kravtsov and A. I. Saichev, J. Opt. Soc. Am. A **2**, 2100 (1985).
[15] A. H. Nayfen, *Perturbation Method* (Wiley, New York, 1973).
[16] L. A. Chernov, *Waves in Random Media* (Nauka, Moscow, 1975).
[17] V. S. Vladimirov, *Equation of Mathematical Physics* (Dekker, New York, 1971).
[18] R. M. Measures, *Laser Remote Sensing, Fundamentals and Applications* (Wiley, New York, 1984).
[19] V. M. Fortus and B. M. Shevtsov, Opt. Spectrosc. **60**, 578 (1986).
[20] L. R. Bissonnette and L. H. Daniel, Appl. Opt. **34**, 6959 (1995).

Solution of a discrete inverse scattering problem and of the Cauchy problem of a class of discrete evolution equations

Henning Blohm[a)]
Fakultät für Mathematik und Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 1-4, D-07743 Jena, Germany

(Received 25 February 1999; accepted for publication 18 May 1999)

The Cauchy problem of a class of nonlinear evolution equations is solved by finding explicit solutions of a discrete inverse scattering problem that are not restricted to the pure soliton case and implementing appropriate time evolution of the scattering data. This yields operator-valued functions, which are shown to solve a hierarchy of operator evolution equations by applying methods similar to those in Marchenko's work. In addition the relation to canonical Lax constructions is investigated. Using methods introduced by Aden and Carl and Schiebold, one obtains scalar solutions to corresponding scalar equations, sometimes representable by determinants on operator ideals.
[S0022-2488(99)03008-X]

I. INTRODUCTION

The inverse scattering problem of the Schrödinger operator, investigated by Marchenko[1] and many others, lead to the so-called inverse scattering method for the Korteweg–de Vries equation

$$u_t = u_{xxx} + 6uu_x, \tag{1}$$

which provides a method for solving the Cauchy problem of (1). This was discovered by Gardner, *et al.*[2] in 1967. A discrete analog was developed by Flaschka[3] in 1974 and was used to solve the discrete nonlinear evolution equation

$$\partial_{tt} y_n = e^{-(y_n - y_{n-1})} - e^{-(y_{n+1} - y_n)} \quad (n \in \mathbb{Z}),$$

called the toda lattice equation.

As Flaschka did, we focus on reconstructing real sequences $a = (a_n)_{n\in\mathbb{Z}}$ and $b = (b_n)_{n\in\mathbb{Z}}$ from spectral information, the scattering data, of an operator $L \in \mathcal{L}(\ell_2(\mathbb{Z}))$ defined by

$$(Lx)_n = a_{n-1}x_{n-1} + b_n x_n + a_n x_{n+1} \quad (x \in \ell_2(\mathbb{Z}), n \in \mathbb{Z})$$

in the process of discrete inverse scattering. Here $\ell_2(\mathbb{Z})$ denotes the Hilbert space of square summable complex sequences over the integers $\mathbb{Z}$ and $\mathcal{L}(\ell_2(\mathbb{Z}))$ the space of all linear and continuous operators on $\ell_2(\mathbb{Z})$. Under the assumptions

$$a_n > 0 \quad (n \in \mathbb{Z}), \qquad \sum_{n=-\infty}^{\infty} |n|(|1-a_n^2| + |b_n|) < \infty, \tag{2}$$

we show how a and b can be represented as sequences of traces of one-dimensional operators, constructed from the scattering data, and therefore obtain an explicit solution of the described discrete inverse scattering problem. This is done in the first part of the paper.

[a)]Electronic mail: blohm@minet.uni-jena.de

The traditional way of construction of solutions of some evolution equation by solving an inverse scattering problem is to assume that the function under consideration already solves that equation and to determine the resulting time evolution of the scattering data. Equations that allow the effective application of this procedure are usually given by Lax pairs[4] or zero-curvature conditions (see, e.g., Faddeev[5] or Ablowitz[6]). Although that analysis is not carried out in this article, the method of Lax pairs was used to come up with the time evolution of the scattering data we implement in Sec. III. Employing ideas similar to those in Marchenko's book,[7] we find operator-valued functions that solve evolution equations easily accessible by an algorithm. Equivalent scalar equations are achieved by application of the trace operator defined on the finite-dimensional operators. Next we show that there is indeed a straight relation to a family of Lax pairs, as indicated above. Note that this suggests uniqueness assertions regarding the solution of the initial value problem solved, a question that is not subject to a rigorous investigation here. A short presentation of possible representations by determinants on operator ideals follows.

Finally, we demonstrate the solution of the Cauchy problem of the Toda lattice equation and the Langmuir lattice equation.

The method of representation of solutions of soliton equations by traces and determinants on operator ideals was introduced by Aden and Carl[8] and Schiebold.[9,10]

II. DISCRETE INVERSE SCATTERING

Our aim here is to derive a discrete version of the so-called Gelfand–Levitan–Marchenko equation found in the literature.

A. Derivation of the fundamental equation

There are many papers on discrete scattering. For a technically more detailed but somewhat different discussion we refer to Geronimo and Case.[11]

Under the hypothesis (2), L has essential spectrum

$$\sigma_e(L)=[-2,2]$$

and discrete spectrum in $\mathbb{R}\backslash[-2,2]$. This can be proved by observing that L is the sum of a right and left shift operator and an additional compact self-adjoint operator and by consulting, for example, Berthier.[12] We consider the linear differences equation

$$\lambda\varphi_n=a_{n-1}\varphi_{n-1}+b_n\varphi_n+a_n\varphi_{n+1}\quad(n\in\mathbb{Z}),\tag{3}$$

where $\lambda\in\mathbb{C}$ is fixed. Asymptotically (3) becomes

$$\lambda\varphi_n=\varphi_{n-1}+\varphi_{n+1}\quad(n\in\mathbb{Z}).\tag{4}$$

If $\lambda\neq0$, a fundamental system of solutions of (4) is given by $(z^n)_{n\in\mathbb{Z}}$ and $(z^{-n})_{n\in\mathbb{Z}}$ where $\lambda=z+z^{-1}$. Solutions of (3) that behave asymptotically similarly, the so-called Jost solutions, are of special interest to us. These can be obtained by solving the following linear summation equation:

$$\frac{1}{\beta_n}\Phi_n=z^n-\sum_{k=n}^{\infty}(1-a_k^2)\left(\frac{z^{n-k-1}-z^{-(n-k-1)}}{z-z^{-1}}\right)\frac{1}{\beta_k}\Phi_k+\sum_{k=n}^{\infty}b_k\left(\frac{z^{n-k}-z^{-(n-k)}}{z-z^{-1}}\right)\frac{1}{\beta_k}\Phi_k$$

using the method of successive approximation. Here we set $\beta_n:=\Pi_{j=n}^{\infty}a_j$ and assumed $z\in\overline{U}$, $U:=\{x\in\mathbb{C}\,|\,|z|<1\}$. This gives solutions

$$\Phi_n^+(z)=\sum_{m=n}^{\infty}k(n,m)z^m\quad(n\in\mathbb{Z})$$

of (3) for $\lambda := z + z^{-1}$ such that the series converges absolutely and $\lim_{n\to\infty} z^{-n}\Phi_n^+(z) = 1$ uniformly on $\overline{U}$. Similarly, or simply by a transformation of coordinates, we get another family of solutions

$$\Psi_n^-(z) = \sum_{m=-n}^{\infty} \ell(n,m) z^m \quad (n \in \mathbb{Z})$$

such that the series converges absolutely and $\lim_{n\to-\infty} z^n \Psi_n^-(z) = 1$ uniformly on $\overline{U}$.

Substituting the Jost solutions into (3), using our hypotheses (2), and comparing coefficients we get:

$$a_n = \frac{k(n+1,n+1)}{k(n,m)}, \quad k(n,n) = \prod_{j=n}^{\infty} \frac{1}{a_j} \neq 0,$$
$$b_n = \frac{k(n,n+1)}{k(n,n)} - \frac{k(n-1,n)}{k(n-1,n-1)} \, (n \in \mathbb{Z}). \tag{5}$$

In the particular case where $|z| = 1$, $z \neq \pm 1$ we define another set of solutions of (3) by

$$\Phi_n^-(z) = \Phi_n^+(z^{-1}) \quad (n \in \mathbb{Z}), \quad \Psi_n^+(z) = \Psi_n^-(z^{-1}) \quad (n \in \mathbb{Z}).$$

From the linear independence of $\Phi^+(z)$ and $\Phi^-(z)$ we infer the existence of coefficients $\alpha(z)$ and $\beta(z)$ such that

$$\Psi_n^-(z) = \beta(z)\Phi_n^+(z) + \alpha(z)\Phi_n^-(z) \quad (n \in \mathbb{Z}). \tag{6}$$

We call $r(z) := \beta(z)/\alpha(z)$ the reflection coefficient. By substitution of (6) into (3) and some computation, it is not hard to see that

$$\overline{\alpha(z)} = \alpha(\overline{z}), \quad \overline{\beta(z)} = \beta(\overline{z}), \quad \overline{r(z)} = r(\overline{z}), \tag{7}$$

and

$$|\alpha(z)|^2 - |\beta(z)|^2 = 1,$$

in particular

$$|r(z)|^2 = 1 - 1/|\alpha(z)|^2 < 1 \tag{8}$$

for all $|z| = 1$, $z \neq \pm 1$. Furthermore, α has the analytical continuation

$$\alpha(z) := \frac{z a_n (\Psi_n^-(z)\Phi_{n+1}^+(z) - \Psi_{n+1}^-(z)\Phi_n^+(z))}{z^2 - 1}$$

onto the open unit disk U. The zeros of α in U are simple and correspond via $\lambda = z + z^{-1}$ to the discrete eigenvalues of L. More precisely, if $\alpha(z) = 0$, then $\Phi^+(z)$ and $\Psi^-(z)$ are linearly dependent eigenvectors of L, and

$$\alpha'(z) = -\rho(z)/z \|\Phi^+(z)\|_2^2, \tag{9}$$

where $\Psi^-(z) = \rho(z)\Phi^+(z)$.

A nontrivial fact is that α has at most finitely many zeros in U. This can be seen by splitting L into two operators, one acting on $\ell_2(\{\dots,-2,-1\})$ and one acting on $\ell_2(\{0,1,2,...\})$. For either one we apply the results of Geronimo and Case and in turn, using the very same techniques, show that L can have only finitely many eigenvalues of magnitude greater than 2.

Let $z_1,\ldots,z_N$ denote the zeros of α in U, $\gamma_i := \|\Phi^+(z_i)\|_2^{-2}$ $(i=1,...,N)$. Integrating the expression $\Psi_n^-(z)z^{m-1}\alpha(z)^{-1}$ along the boundary of a ball of radius $0<\epsilon<1$ and applying the residue theorem to one side, using (9), while considering the limit as ϵ tends to one and using (6) on the other side, yields the following discrete version of the famous Gelfand–Levitan–Marchenko equation:

$$\frac{\delta_{nm}}{k(n,n)} = k(n,m) + \sum_{j=n}^{\infty} k(n,j) f(m+j) \quad (m \geqslant n \in \mathbb{Z}), \tag{10}$$

where

$$f(m) = \sum_{i=1}^{N} z_i^m \gamma_i + \frac{1}{2\pi i} \int_{\partial U} r(z) z^{m-1}\, dz \quad (m \in \mathbb{Z})$$

and δ is the Kronecker–Delta function. The numbers $z_1,\ldots,z_N$ together with the normalization constants $\gamma_1,\ldots,\gamma_N$ and the reflection coefficient r are called the scattering data. By setting $\kappa(n,m) := k(n,m)/k(n,n)$ for $m>n$, Eq. (10) can be reformulated as

$$0 = \kappa(n,m) + f(m+n) + \sum_{j=n+1}^{\infty} \kappa(n,j) f(m+j) \quad (m>n \in \mathbb{Z}). \tag{11}$$

We will focus on solving Eq. (11), that is, as we will see, uniquely solvable. Once we found κ, one can use

$$k(n,n)^{-2} = 1 + f(2n) + \sum_{j=n+1}^{\infty} \kappa(n,j) f(n+j)$$

to compute $k(n,n)$. By the definition of κ, all other values $k(n,m)$ for $m>n$ are accessible, and via (5) the values of a and b are thereby determined by κ. Finally, we state a consequence of (11) proved by Geronimo and Case:[11]

$$\sum_{m=0}^{\infty} |f(m)| < \infty$$

and in particular

$$\sum_{m=0}^{\infty} \left| \int_{\partial U} r(z) z^{m-1} dz \right| < \infty. \tag{12}$$

B. Solution of the inverse scattering problem

The following theorem shows how to solve Eq. (11) abstractly, using the trace operator tr, defined on the operator ideal of finite-dimensional operators.[13] We write $(\cdot|\cdot)$ for the inner product on a Hilbert space and $h \otimes g$ for the one-dimensional operator on a vector space V, defined for a linear functional h on V and an element g of V by

$$h \otimes g x := h(x) g \quad (x \in V).$$

An important property of the trace operator, that is used here and in later sections, is that it is multiplicative on operators of the form $T(h \otimes g)$, that is,

$$\mathrm{tr}(T(h\otimes g) S(h\otimes g)) = \mathrm{tr}(T(h\otimes g))\,\mathrm{tr}(S(h\otimes g))$$

for any two linear operators S and T on V. If H is a Hilbert space, its dual will be canonically identified with the space itself. By e_n we denote the vector of $\ell_2(\mathbb{Z})$ that has nth coordinate one and zero otherwise.

Theorem II.B.1: Let H be a Hilbert space, $V\in\mathcal{L}(H)$, $h,g\in H$, and let $B\in\mathcal{L}(H)$ be such that

$$(Bx|y)=-\sum_{k=0}^{\infty}(V^kx|h)(V^kg|y)\quad\text{for all}\quad x,y\in H.$$

Then:

(i) $VBV-B=h\otimes g$.

(ii) Assuming that V and $(1+V^{2n}B)$ are invertible for every $n\in\mathbb{Z}$ and setting

$$K(n,m):=V^{m-n}(1+V^{2n+1}BV)^{-1}V^{2n}(h\otimes g),\quad F(m):=-V^m(h\otimes g)\quad(n,m\in\mathbb{Z}),$$

it is

$$0=\mathrm{tr}(K(n,m))+\mathrm{tr}(F(m+n))+\sum_{j=n+1}^{\infty}\mathrm{tr}(F(m+j))\mathrm{tr}(K(n,j))\quad(n,m\in\mathbb{Z},m\geqslant n).$$

(iii) Under the hypotheses of *(ii)*, $\kappa(n,m):=\mathrm{tr}(K(n,m))$ $(n,m\in\mathbb{Z},m>n)$ defines a solution of Eq. (11) with kernel $f(m):=\mathrm{tr}(F(m))$ $(m\in\mathbb{Z})$.

Proof: For arbitrary $x,y\in H$ it is

$$(VBVx-Bx|y)=(VBVx|y)-(Bx|y)=-\sum_{k=0}^{\infty}(V^{k+1}x|h)(V^{k+1}g|y)-(Bx|y)$$
$$=(x|h)(g|y)=(h\otimes gx|y),$$

which implies (i). We set $L_n:=V^{2n}B$ and $W(n):=(1+VL_nV)^{-1}V^{2n}(h\otimes g)$ and compute

$$\sum_{j=n}^{\infty}\mathrm{tr}(F(m+j))\mathrm{tr}(K(n,j))=-\sum_{j=n}^{\infty}(V^{m+j}g|h)(V^{j-n}(1+VL_nV)^{-1}V^{2n}g|h)$$
$$=-\sum_{j=0}^{\infty}(V^j(1+VL_nV)^{-1}V^{2n}g|h)(V^{m+n+j}g|h)$$
$$=(V^{m+n}B(1+VL_nV)^{-1}V^{2n}g|h)=\mathrm{tr}(V^{m+n}BW(n))$$

and

$$F(m+n)+K(n,m)-F(m+n)K(n,n)+V^{m+n}BW(n)$$
$$=-V^{m+n}(h\otimes g)+V^{m-n}W(n)+V^{m+n}(h\otimes g)W(n)+V^{m+n}BW(n)$$
$$=-V^{m+n}(h\otimes g)+V^{m-n}(1+V^{2n}(h\otimes g)+V^{2n}B)W(n)$$
$$=-V^{m+n}(h\otimes g)+V^{m-n}(1+V^{2n+1}BV)W(n)=0.$$

Now, application of the trace operator to the last equation gives (ii), and statement (iii) is an immediate consequence of it. □

The theorem just proved suggests the following model for the solution of Eq. (11): We choose $H_1:=\ell_2^N$, $H_2:=L_2(T)$, $H:=H_1\oplus^2H_2$, where $T:=\partial U$ is the complex unit sphere equipped with the normalized Haar-measure μ. In addition let

(1) $V_1\in\mathcal{L}(H_1)$ such that $(V_1x)_i:=z_ix_i(x\in\ell_2^N)$;
(2) $V_2\in\mathcal{L}(H_2)$ multiplication with the identical mapping on T;

(3) $h_1, g_1 \in H_1$ such that $-\overline{h_1} g_1 = \gamma = (\gamma_1, \ldots, \gamma_N)$;
(4) $h_2, g_2 \in H_2 \cap L_\infty(T)$ such that $-\overline{h_2} g_2 = r$.

After setting $V := V_1 \oplus V_2$, $g := g_1 \oplus g_2$, $h := h_1 \oplus h_2$, we observe:

$$\begin{aligned}\operatorname{tr}(V^m(h \otimes g)) = (V^m g|h) = (V_1^m g_1|h_1) + (V_2^m g_2|h_2) \\ &= -\sum_{i=1}^{N} z_i^m \gamma_i - \int_T r(z) z^m \, d\mu(z) \\ &= -\left(\sum_{i=1}^{N} z_i^m \gamma_i + \frac{1}{2\pi i} \int_{\partial U} r(z) z^{m-1} \, dz \right) = -f(m).\end{aligned}$$

Consequently, we are going to show:

(1) The operator equation $VBV - B = h \otimes g$ can be solved as desired for Theorem II.B.1.
(2) The operator $1 + V^{2n} B$ is invertible for every $n \in \mathbb{Z}$.

We proceed with the solution of the first problem:

Lemma II.B.2: Suppose that $g, h \in L_2(T) \cap L_\infty(T)$ and that $V \in \mathcal{L}(L_2(T))$ is the multiplication with the identical mapping on T. Then there is exactly one operator $B \in \mathcal{L}(L_2(T))$ such that

$$(Bx|y) = -\sum_{k=0}^{\infty} (V^k x|h)(V^k g|y) \quad \text{for every } x, y \in L_2(T).$$

It follows that $\|B\| \leqslant \|g|L_\infty(T)\| \|h|L_\infty(T)\|$, and that B factors through $\ell_2(\mathbb{N}_0)$, $B = SR$ where $S \in \mathcal{L}(\ell_2(\mathbb{N}_0), L_2(T))$ is such that

$$(Sx|y) = -\sum_{n=0}^{\infty} (x|e_n)(V^n g|y) \quad \text{for every } x \in \ell_2(\mathbb{N}_0), y \in L_2(T)$$

and $\|S\| \leqslant \|g|L_\infty(T)\|$, as well as $R \in \mathcal{L}(L_2(T), \ell_2(\mathbb{N}_0))$ is such that

$$(Rx|y) = \sum_{k=0}^{\infty} (V^k x|h)(y|e_k) \quad \text{for every } x \in L_2(T), y \in \ell_2(\mathbb{N}_0)$$

and $\|R\| \leqslant \|h|L_\infty(T)\|$.

Proof: Since

$$\begin{aligned}\sum_{k=0}^{\infty} |(V^k x|h)(V^k g|y)| = \sum_{k=0}^{\infty} \left| \int_T (x\overline{h}) z^k \, d\mu(z) \right| \left| \int_T (g\overline{y}) z^k \, d\mu(z) \right| \\ \leqslant \|x\overline{h}|L_2(T)\| \|g\overline{y}|L_2(T)\| \leqslant \|x\| \|h|L_\infty(T)\| \|g|L_\infty(T)\| \|y\|\end{aligned}$$

for arbitrary $x, y \in L_2(T)$,

$$[x|y] := -\sum_{k=0}^{\infty} (V^k x|h)(V^k g|y)$$

defines a continuous sesquilinear form on $L_2(T)$. We conclude that for each $y \in L_2(T)$ there exists exactly one $Ay \in L_2(T)$ satisfying

$$[\cdot|y] = (\cdot|Ay). \tag{13}$$

Obviously $\|Ay\| \leqslant \|g|L_\infty(T)\|\|h|L_\infty(T)\|\|y\|$. Hence $B=A^*$ is the right choice. It is clear that B is uniquely determined through A by (13). Similarly S and R are given by their required properties. Finally, since

$$(SRx|y) = -\sum_{n=0}^{\infty} (Rx|e_n)(V^n g|y) = -\sum_{n=0}^{\infty} (V^n x|h)(V^n g|y)$$

for all $x,y \in L_2(T)$, $B=SR$. □

It is easier to solve the corresponding operator equation over $H_1 = \ell_2^N$ for V_1. Because of $\|V_1\| < 1$, we can simply choose

$$B_1 := -\sum_{n=0}^{\infty} V_1^n (h_1 \otimes g_1) V_1^n.$$

In this case we have an analogous factorization through $\ell_2(\mathbb{N}_0)$ by

$$S_1 := -\sum_{n=0}^{\infty} e_n \otimes (V_1^n g_1), \quad R_1 := \sum_{k=0}^{\infty} ((V_1^*)^k h_1) \otimes e_k .$$

The operators received from the application of Lemma II.B.2 to V_2, g_2, h_2 will be denoted by B_2, S_2, R_2 from now on. We define

$$K_1 := R_1 S_1, \quad K_2 := R_2 S_2$$

and write $I_i : H_i \hookrightarrow H$ and $P_i : H \to H_i$ for the canonical inclusion and projection operators, respectively. Let

$$R := R_1 P_1 + R_2 P_2 \in \mathcal{L}(H, \ell_2(\mathbb{N}_0)), \quad S := I_1 S_1 + I_2 S_2 \in \mathcal{L}(\ell_2(\mathbb{N}_0), H),$$

$$K := K_1 + K_2 \in \mathcal{L}(\ell_2(\mathbb{N}_0)), \quad B := SR \in \mathcal{L}(H).$$

It is easily checked that $K = RS$. For arbitrary $x = x_1 \oplus x_2 \in H$ and $y = y_1 \oplus y_2 \in H$ we compute

$$\begin{aligned}
(Bx|y) &= (I_1 S_1 R_1 x_1 + I_1 S_1 R_2 x_2 + I_2 S_2 R_1 x_1 + I_2 S_2 R_2 x_2 | y) \\
&= (S_1 R_1 x_1 | y_1) + (S_1 R_2 x_2 | y_1) + (S_2 R_1 x_1 | y_2) + (S_2 R_2 x_2 | y_2) \\
&= -\sum_{n=0}^{\infty} ((V_1^n x_1 | h_1)(V_1^n g_1 | y_1) + (V_2^n x_2 | h_2)(V_1^n g_1 | y_1) \\
&\quad + (V_1^n x_1 | h_1)(V_2^n g_2 | y_2) + (V_2^n x_2 | h_2)(V_2^n g_2 | y_2)) \\
&= -\sum_{n=0}^{\infty} (V^n x | h)(V^n g | y).
\end{aligned}$$

Thus B is the solution of the operator equation sought for.

Next we take care of the existence of the inverse operators. A by-product is the unique solvability of Eq. (11): Set $K^{(n)} := R V^{2n} S$. An easy computation shows that

$$(K^{(n+1)} x)_k = \sum_{j=0}^{\infty} f(2n+k+j+2) x_j ,$$

where $f(m) = -\mathrm{tr}(V^m (h \otimes g))$. Hence solving Eq. (11) is equivalent to solving

$$(1 + K^{(n+1)}) \kappa(n, n+1+\cdot) = -f(2n+1+\cdot)$$

for every $n\in\mathbb{Z}$. As we will see soon, $1+K^{(n)}$ is invertible and therefore (11) possesses at most one solution κ. The following lemma collects important properties of the operators $K^{(n)}$. Similar statements can be found in other settings in scattering theory.

Lemma II.B.3: Assuming the situation of our model, it is $\|K^{(n)}\|\leqslant\Sigma_{j=0}^{\infty}|f(2n+j)|$, $K^{(n)}$ is compact, and $-1\notin\sigma(K^{(n)})$ for every $n\in\mathbb{Z}$.

Proof: Let $N\in\mathbb{N}_0$ and $x\in\ell_2(\mathbb{N}_0)$, $\|x\|\leqslant 1$. It is

$$\left(\sum_{k=N}^{\infty}\left|\sum_{j=0}^{\infty}f(2n+k+j)x_j\right|^2\right)^{1/2}=\left(\sum_{k=N}^{\infty}\left|\sum_{j=k}^{\infty}f(2n+j)x_{j-k}\right|^2\right)^{1/2}$$
$$\leqslant\sum_{j=N}^{\infty}\left(\sum_{k=N}^{j}|f(2n+j)x_{j-k}|^2\right)^{1/2}$$
$$\leqslant\sum_{j=N}^{\infty}|f(2n+j)|.$$

Choosing $N=0$ gives the norm estimate. In addition we showed that for any $\epsilon>0$ there exists an $N\geqslant 0$ such that

$$\left(\sum_{j=N}^{\infty}|y_j|^2\right)^{1/2}<\epsilon$$

uniformly for all $y=K^{(n)}x$, $x\in\ell_2(\mathbb{N}_0)$, $\|x\|\leqslant 1$. We infer that $K^{(n)}$ is compact. Define

$$K_1^{(n)}:=R_1V_1^{2n}S_1,\quad K_2^{(n)}:=R_2V_2^{2n}S_2.$$

Then $K^{(n)}=K_1^{(n)}+K_2^{(n)}$, and for arbitrary $i,j\in\mathbb{N}_0$ we have

$$(K_1^{(n)}e_j|e_i)=(R_1V_1^{2n}S_1e_j|e_i)=(V_1^{2n+i}S_1e_j|h_1)=-(V_1^{2n+i+j}g_1|h_1).$$

If $x\in\ell_2(\mathbb{N}_0)$, this implies

$$(K_1^{(n)}x|x)=-\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}x_j(V_1^{n+j}g_1|(V_1^*)^{n+k}h_1)\overline{x_k}=-\sum_{i=1}^{N}g_{1,i}\overline{h_{1,i}}\left|\sum_{k=0}^{\infty}x_kz_i^{k+n}\right|^2\geqslant 0,$$

and therefore $K_1^{(n)}$ is a positive operator. Let us turn to the second operator: Since

$$(K_2^{(n)}e_j|e_i)=\cdots=-(V_2^{2n+j+i}g_2|h_2)=\int_T r(z)z^{2n+i+j}\,d\mu(z),$$

we see for arbitrary $x\in\ell_2(\mathbb{N}_0)$ that

$$(K_2^{(n)}x|x)=\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\int_T r(z)z^{2n+i+j}x_j\overline{x_k}\,d\mu(z)=\int_T r(z)q(z)\overline{p(z)}d\mu(z)$$

where $q(z)=\Sigma_{j=0}^{\infty}x_jz^{n+j}$, $p(z)=\Sigma_{k=0}^{\infty}x_kz^{-(n+k)}$ [in $L_2(T)$]. Thus:

$$|(K_2^{(n)}x|x)|\leqslant\|rp|L_2(T)\|\|q|L_2(T)\|\leqslant\|p|L_2(T)\|\|q|L_2(T)\|=\|x\|^2,$$

because $|r|\leqslant 1$ almost everywhere. Assuming $\|rp|L_2(T)\|=\|p|L_2(T)\|$ leads to

$$\int_T(1-|r(z)|^2)|p(z)|^2\,d\mu(z)=0,$$

which, taking Eq. (8) into account, implies that p vanishes μ almost everywhere and thus that $x=0$. Hence we have shown:

$$x\neq 0 \Rightarrow |(K_2^{(n)}x|x)|<\|x\|^2.$$

Now it is easy to see that $1+K^{(n)}$ is one-to-one. Indeed, if $x\neq 0$, we know that

$$((1+K^{(n)})x|x)=\|x\|^2+(K_1^{(n)}x|x)+(K_2^{(n)}x|x)>0,$$

and in particular it follows that $(1+K^{(n)})x\neq 0$. Now, as shown in Riesz theory of functional analysis, the fact that $K^{(n)}$ is compact implies that $1+K^{(n)}$ is invertible. □

Lemma II.B.4: In the situation of our model, the operators $1+V^nB$ are invertible for every $n\in\mathbb{Z}$ and

$$\lim_{n\to\infty}((1+V^nB)^{-1}-(1-V^nB))=0.$$

Proof: This is implied by the next lemma. □

Lemma II.B.5: Suppose $R\in\mathcal{L}(E,F)$ and $S\in\mathcal{L}(F,E)$ where E and F are Banach spaces. Then:

(i) $1+RS$ is invertible if and only if $1+SR$ is invertible, and in that case: $(1+RS)^{-1}=1-R(1+SR)^{-1}S$.

(ii) $\rho(SR)\backslash\{0\}=\rho(RS)\backslash\{0\}$, $\sigma(SR)\backslash\{0\}=\sigma(RS)\backslash\{0\}$.

Proof: Assertion *(i)* is easily proved using the stated formula, and *(ii)* is a simple consequence of it. □

Now we have essentially solved our discrete inverse scattering problem and close this section with a main result:

Theorem II.B.6: In the situation of our model the following assertions are true:

(i) Equation (11) possesses exactly one solution $\kappa(n,n+\cdot)$ in $\ell_2(\mathbb{N}_0)$ for each $n\in\mathbb{Z}$.

(ii) The operators $1+V^{2n}B$ and $1+V^{2n+1}BV$ are invertible for every $n\in\mathbb{Z}$.

(iii) $\kappa(n,m):=\mathrm{tr}(V^{m-n}(1+V^{2n+1}BV)^{-1}V^{2n}(h\otimes g))(m>n)$ gives the solution of Eq. (11) with kernel

$$f(m)=-\mathrm{tr}(V^m(h\otimes g))=\sum_{i=1}^{N} z_i^m\gamma_i+\frac{1}{2\pi i}\int_{\partial U} r(z)z^{m-1}\,dz \quad (m\in\mathbb{Z}).$$

(iv) Let $Q_n:=(1+V^{2n}B)^{-1}V^{2n}(h\otimes g)$, $q_n:=\mathrm{tr}(Q_n)$, and let k denote the function of the coefficients of the Jost solution Φ^+. Then:

$$0<k(n,n)=\prod_{j=n}^{\infty}\frac{1}{a_j}=\sqrt{1+q_n} \quad (n\in\mathbb{Z})$$

and $k(n,m)=\sqrt{1+q_n}\,\kappa(n,m)(m>n)$.

(v) For every $n\in\mathbb{Z}$:

$$a_n=\left(\frac{1+q_{n+1}}{1+q_n}\right)^{1/2}=\frac{k(n+1,n+1)}{k(n,n)}$$

and

$$b_n=\frac{k(n,n+1)}{k(n,n)}-\frac{k(n-1,n)}{k(n-1,n-1)}=\kappa(n,n+1)-\kappa(n-1,n)$$
$$=\mathrm{tr}(V(1+Q_n)^{-1}Q_n-V(1+Q_{n-1})^{-1}Q_{n-1}).$$

Proof: Almost everything has been proved before. Only assertion *(iv)* still calls for some consideration: Using the techniques of the proof of Theorem II.B.1, we see that

$$\begin{aligned}\frac{1}{k(n,n)^2} &= 1+f(2n)+\sum_{j=n+1}^{\infty}\kappa(n,j)f(n+j)\\ &= 1-\kappa(n,n)\\ &= 1-\mathrm{tr}((1+V^{2n+1}BV)^{-1}V^{2n}(h\otimes g))\\ &= 1-\mathrm{tr}((1+Q_n)^{-1}Q_n)\\ &= 1-\frac{q_n}{1+q_n}=\frac{1}{1+q_n}.\end{aligned}$$

We know that $k(n,n)>0$, and by the last computation, it is $1+q_n>0$. Hence:

$$k(n,n)=\sqrt{1+q_n}. \qquad \square$$

III. SOLUTION OF EVOLUTION EQUATIONS

In this section we implement a certain time development of the scattering data and show what evolution equation can be solved by the resulting functions a and b. From now on we will always use the constructions and notations of our model. For real numbers $\sigma_1,\ldots,\sigma_M$ set

$$A:=\sum_{k=1}^{M}\sigma_k(V^k-V^{-k}),\quad g(t):=e^{tA}g\quad (t\in\mathbb{R}). \tag{14}$$

We assume that g was given by initial values a and b that satisfy (2). All operators and functions, defined using g before, will from now on be defined for all $t\in\mathbb{R}$ by substitution of $g(t)$ for g. Consequently they will depend on an additional parameter t. The next Lemma shows that everything works out well. Note that, in the words of scattering theory, (14) just means that the discrete eigenvalues of L stay constant in time, while the reflection coefficient and the normalization constants γ_i have exponential time evolution.

Lemma III.1: For arbitrary $t\in\mathbb{R}$ the following statements are true:
(i) $B(t)=e^{tA}B(0)$;
(ii) $\forall n\in\mathbb{Z}$: $-1\notin\sigma(V^{2n}B(t))\cup\sigma(V^{2n+1}B(t)V)\cup\sigma(K^{(n)}(t))\cup\sigma(Q_n(t))$;
(iii) $\forall n\in\mathbb{Z}$: $\kappa(n,n+\cdot,t)\in\ell_2(\mathbb{N}_0)$;
(iv) $\forall m\in\mathbb{Z}$: $f(m,t):=\mathrm{tr}(F(m,t))\in\mathbb{R}$;
(iv) $\forall n\in\mathbb{Z}$: $1+q_n(t)>0$.
Proof: Let $x,y\in H$. Since

$$(e^{tA}Bx|y)=-\sum_{n=0}^{\infty}(V^nx|h)(V^ng|e^{tA^*}y)=-\sum_{n=0}^{\infty}(V^nx|h)(V^ne^{tA}g|y),$$

the uniqueness assertion of Lemma II.B.2 implies that $B(t)=e^{tA}B$. In order to prove *(ii)*, we return our attention to the proof of Lemma II.B.3: It is necessary that the negative Fourier coefficients of $\overline{h_2}g_2(t)$ are absolutely summable. This will be shown in the following Lemma III.2. Furthermore,

$$g_2(t)(z)=\exp\left(t2\sum_{k=1}^{M}\sigma_k(z^k-\overline{z}^k)\right)g_2(z)=\exp\left(t4i\sum_{k=1}^{M}\sigma_k\mathfrak{J}z^k\right)g_2(z)\quad (z\in T),$$

and so $|\overline{h_2}g_2(t)|$ is constant in t. From

$$g_1(t)_n = \exp\left(t2\sum_{k=1}^{M}\sigma_k(z_n^k - z_k^{-k})\right)g_{1,n} \quad (n=1,...,N),$$

we conclude that $\overline{h_1}g_1(t)$ has constant sign. These properties allow us to infer the assertions of Lemma II.B.3 for all $t\in\mathbb{R}$. Assertion *(iii)* follows easily from $\|V_1\|<1$ and noting that $h_2 \in L_\infty(T)$. Since

$$\overline{\mathrm{tr}(V^m(h\otimes g(t)))} = -\sum_{i=1}^{N}\exp\left(t\sum_{k=1}^{M}\sigma_k(z_i - z_i^{-1})\right)z_i^m\gamma_i - \int_T \exp\left(t\sum_{k=1}^{M}\sigma_k\overline{(z-z^{-1})}\right)\overline{r(z)z^m}\,d\mu(z),$$

Eq. (7) and the properties of the Haar measure μ give *(iv)*. In order to prove *(v)*, we observe that $-1\notin\sigma(Q_n(t))$ implies $1+q_n(t)\neq 0$ for all real t. By *(iii), (iv)* and by the construction of the operators $K^{(n)}$, not only κ gives solutions to the Eq. (11) but also its conjugate. Hence, by the uniqueness of solutions implied by *(ii)*, $\kappa(n,m,t)$ must be real for all t and $m>n$. As in the proof of Theorem II.B.6, it is

$$\frac{1}{1+q_n(t)} = 1-\kappa(n,n,t) = 1+f(2n,t)+\sum_{j=n+1}^{\infty}\kappa(n,j,t)f(n+j,t).$$

Therefore $q_n(t)$ is real. Now, since $1+q_n(0)>0$, $1+q_n(t)\neq 0$ for all $t\in\mathbb{R}$, assertion *(v)* follows from the continuity of q_n. □

We needed:

Lemma III.2: Suppose $g\in L_2(T)$ is such that $C:=\Sigma_{n=-\infty}^{0}|(g|f_n)|<\infty$, where $(f_n)_{n\in\mathbb{Z}}$ denotes the canonical orthonormal basis in $L_2(T)$, and let $c\in\mathbb{R}$, $m\in\mathbb{Z}$ be arbitrarily chosen. Furthermore define $h(z):=e^{cz^m}g(z)$ $(z\in T)$. Then: $\Sigma_{n=-\infty}^{0}|(h|f_n)|<\infty$.

Proof: Since

$$(h|f_n) = \int_T g(z)e^{cz^m}z^{-n}d\mu(z) = \sum_{k=0}^{\infty}\frac{c^k}{k!}\int_T g(z)z^{mk-n}\,d\mu(z),$$

$$\sum_{n=-\infty}^{0}|(h|f_n)| \leqslant \sum_{k=0}^{\infty}\frac{c^k}{k!}\sum_{n=-\infty}^{0}|(g|f_{n-mk})|.$$

For the case $m\geqslant 0$ we have $\Sigma_{n=-\infty}^{0}|(h|f_n)|\leqslant Ce^c$. Otherwise we use Hölder's inequality and Bessel's inequality to show

$$\sum_{n=-\infty}^{0}|(h|f_n)| \leqslant \sum_{k=0}^{\infty}\frac{c^k}{k!}\left(\sum_{n=1}^{-mk}|(g|f_n)|+C\right) \leqslant \sum_{k=0}^{\infty}\frac{c^k}{k!}(-mk)\left(\sum_{n=1}^{-mk}|(g|f_n)|^2\right)^{1/2}+Ce^c \leqslant -mce^c\|g\|+Ce^c.$$ □

Now we are in the position to define

$$a_n(t) := \left(\frac{1+q_{n+1}(t)}{1+q_n(t)}\right)^{1/2}$$

and

$$b_n(t) := \mathrm{tr}(V(1+Q_n(t))^{-1}Q_n(t) - V(1+Q_{n-1}(t))^{-1}Q_{n-1}(t))$$

for $n \in \mathbb{Z}$, $t \in \mathbb{R}$. These functions are differentiable in t, b_n and a_n^2 are even analytical in t. From now on we will frequently omit the parameter t in computations and definitions.

A. Derivation of evolution equations

Without necessary relation to the model introduced in Sec. II B, we show a derivation of a class of operator evolution equations that will—of course—later be shown to combine with our previous work.

Consider a family of operators $F_n(t)$ where $n \in \mathbb{Z}$ and t is real or complex and possibly confined to an open set. We require that $F_n(t)$ is invertible for all admissible t and that F_n is differentiable in t such that

$$\partial_t F_n = \sum_{k=1}^{M} \sigma_k F_{n+k} \quad (n \in \mathbb{Z}), \tag{15}$$

where $\sigma_1, \ldots, \sigma_M$ are some fixed scalars. Furthermore we demand that the expression

$$(F_{m+k} + F_{m-k}) F_m^{-1} \tag{16}$$

is independent of $m \in \mathbb{Z}$ for fixed k. Define

$$\Gamma_n := F_n^{-1} F_{n+1}, \quad B_n := \Gamma_{n+1} - \Gamma_n, \quad C_n := \Gamma_n^{-1} \Gamma_{n+1}$$

for $n \in \mathbb{Z}$. Under the two hypotheses above, we derive evolution equations for B_n and C_n. First note that trivially

$$B_n = \Gamma_{n+1} - \Gamma_n = \Gamma_n(\Gamma_n^{-1}\Gamma_{n+1} - 1) = \Gamma_n(C_n - 1), \tag{17}$$

and as a consequence of property (16):

$$B_n = \Gamma_n^{-1}(\Gamma_{n-1}^{-1} - \Gamma_n^{-1})\Gamma_n = \Gamma_n^{-1}(C_{n-1} - 1). \tag{18}$$

After setting

$$X_{j,n}^{(k)} = \begin{cases} 1 & \text{if } j = k \\ \Gamma_{n+1} - \Gamma_{n-k+1} & \text{if } j = k-1 \\ \Gamma_{n+1} \cdots \Gamma_{n+k-j-1}(\Gamma_{n+k-j} - \Gamma_{n-j}) & \text{if } 0 \leqslant j < k-1 \\ 0 & \text{otherwise;} \end{cases}$$

for $n, j \in \mathbb{Z}$, $1 \leqslant k \leqslant M$ and noting that $F_{n+k} = F_n \Gamma_n \cdots \Gamma_{n+k-1}$ for $k \geqslant 1$, we arrive at

$$\partial_t \Gamma_n = -F_n^{-1} \partial_t F_n \Gamma_n + F_n^{-1} \partial_t F_{n+1} = \sum_{k=1}^{M} \sigma_k (\Gamma_n \cdots \Gamma_{n+k-1}(\Gamma_{n+k} - \Gamma_n)) = \sum_{k=1}^{M} \sigma_k \Gamma_n X_{0,n}^{(k)} = \Gamma_n W_{0,n},$$

where $W_{j,n} := \sum_{k=1}^{M} \sigma_k X_{j,n}^{(k)}$. Therefore:

$$\partial_t C_n = -\Gamma_n^{-1} \partial_t \Gamma_n C_n + \Gamma_n^{-1} \partial_t \Gamma_{n+1} = C_n W_{0,n+1} - W_{0,n} C_n. \tag{19}$$

Similarly, since $\partial_t(\Gamma_n^{-1}) = -\Gamma_n^{-1} \partial_t \Gamma_n \Gamma_n^{-1} = -W_{0,n} \Gamma_n^{-1}$, it is

$$\Gamma_n^{-1}(\partial_t(\Gamma_{n-1}^{-1}) - \partial_t(\Gamma_n^{-1}))\Gamma_n = \Gamma_n^{-1} W_{0,n-1} \Gamma_{n-1}^{-1} \Gamma_n + \Gamma_n^{-1} W_{0,n} = -W_{1,n} C_{n-1} + C_n W_{1,n+1}.$$

Hence, by Eq. (18):

$$\partial_t B_n = C_n W_{1,n+1} - W_{1,n} C_{n-1} - W_{0,n} B_n + B_n W_{0,n} = C_n W_{1,n+1} - W_{1,n} C_{n-1} + [B_n, W_{0,n}]. \quad (20)$$

The point is that we can state an easy algorithm that lets us formulate the $X_{j,n}^{(k)}$ in terms of the $C_n - 1$ and B_n. Equation (18) together with Eq. (17) gives:

$$\Gamma_m B_m = (\Gamma_m - \Gamma_n) B_n + \Gamma_n B_n = \begin{cases} (B_{m-1} + \cdots + B_n) B_n + C_{n-1} - 1 & \text{if } m > n \\ C_{n-1} - 1 & \text{if } m = n \\ -(B_{n-1} + \cdots + B_m) B_n + C_{n-1} - 1 & \text{if } m < n; \end{cases}$$

and

$$\begin{aligned} \Gamma_m (C_n - 1) &= (\Gamma_m - \Gamma_n)(C_n - 1) + \Gamma_n (C_n - 1) \\ &= \begin{cases} (B_{m-1} + \cdots + B_n)(C_n - 1) + B_n & \text{if } m > n \\ B_n & \text{if } m = n \\ -(B_{n-1} + \cdots + B_m)(C_n - 1) + B_n & \text{if } m < n. \end{cases} \end{aligned}$$

All that remains to realize is, that after we start with $X_{k-1,n}^{(k)} = B_n + \cdots + B_{n-k+1}$, we can iterate over $X_{j,n}^{(k)} = \Gamma_{n+1} X_{j+1,n+1}^{(k)}$ $(0 \leqslant j < k)$, applying the equations above, to derive formulas of all $X_{j,n}^{(k)}$ in terms of the B_n, $C_n - 1$. We will do the computations for $k = 3$: Starting with $X_{2,n}^{(3)} = B_n + B_{n-1} + B_{n-2}$, we get

$$\begin{aligned} X_{1,n}^{(3)} = \Gamma_{n+1} X_{2,n+1}^{(3)} &= \Gamma_{n+1}(B_n + B_{n-1} + B_{n-2}) \\ &= B_n^2 + B_{n-1}^2 + B_n Bn - 1 + (C_n - 1) + (C_{n-1} - 1) + (C_{n-2} - 1). \end{aligned}$$

Similarly,

$$\begin{aligned} X_{0,n}^{(3)} = \Gamma_{n+1} X_{1,n+1}^{(3)} &= B_n^3 + (C_n - 1)(B_{n+1} + 2B_n) + B_{n+1} \\ &\quad + B_n + B_{n-1} + B_n (C_n - 1) + (B_n + B_{n-1})(C_{n-1} - 1). \end{aligned}$$

Let us return to the situation of our model: Consider

$$F_n(t) := V^{-n} e^{t\tilde{Z}} + V^n e^{tZ} B,$$

where $n \in \mathbb{Z}$, $t \in \mathbb{R}$ and $Z := \Sigma_{k=1}^M \sigma_k V^k$, $\tilde{Z} := \Sigma_{k=1}^M \sigma_k V^{-k}$, and define Γ_n, C_n, B_n, $X_{j,n}^{(k)}$ and $W_{j,n}$ just as before. It is readily seen that $\Gamma_n V = (1 + Q_n)$. Clearly, property (15) is satisfied, whereas property (16) is a consequence of the relation $F_{m+k} F_m^{-1} = V^{-m}(V^{-k} - V^k) e^{t\tilde{Z}} F_m^{-1} + V^k$, which in turn gives $F_{m+k} F_m^{-1} = -F_{m-k} F_m^{-1} + V^k + V^{-k}$ for all $m, k \in \mathbb{Z}$. The eventual goal is to derive evolution equations for the scalars a_n and b_n. The relation to our construction is as follows:

$$\begin{aligned} c_n := a_n^2 = \frac{1 + q_{n+1}}{1 + q_n} &= 1 + \operatorname{tr}((1 + Q_n)^{-1}(Q_{n+1} - Q_n)) \\ &= 1 + \operatorname{tr}(V^{-1} \Gamma_n^{-1} (\Gamma_{n+1} - \Gamma_n) V) \\ &= 1 + \operatorname{tr}(\Gamma_n^{-1} \Gamma_{n+1} - 1) = 1 + \operatorname{tr}(C_n - 1). \end{aligned} \quad (21)$$

From (17) we infer

$$(1 + Q_{n-1})^{-1} - (1 + Q_n)^{-1} = V^{-1}(\Gamma_{n-1}^{-1} - \Gamma_n^{-1}) = V^{-1} \Gamma_n (\Gamma_{n+1} - \Gamma_n) \Gamma_n^{-1}$$

and hence

$$b_n = \mathrm{tr}(V((1+Q_{n-1})^{-1} - (1+Q_n)^{-1})) = \mathrm{tr}(\Gamma_{n+1} - \Gamma_n) = \mathrm{tr}(B_n). \tag{22}$$

Note that the operators $X_{j,n}^{(k)}$ for $0 \leqslant j \leqslant k-1$ are one dimensional, of the form $T(((V^*)^{-1}h) \otimes g)$ for some operator T. The same holds for the operators $C_n - 1$, B_n. This allows us to obtain scalar evolution equations for c_n, b_n from Eqs. (19) and (20) by Eqs. (21) and (22), and by application of the trace operator.

For example, assuming $A = \Sigma_{j=1}^{3} \sigma_j (V^j - V^{-j})$ we arrive after some computation at the following scalar evolution equations:

$$\begin{aligned}\partial_t c_n = c_n(&\sigma_3(b_{n+1}^3 - b_n^3 - 3(b_{n+1} - b_n) + c_{n+1}(2b_{n+1} + b_{n+2}) \\ &+ c_n(b_{n+1} - b_n) + c_{n-1}(-2b_n - b_{n-1})) \\ &+ \sigma_2(b_{n+1}^2 - b_n^2 + c_{n+1} - c_{n-1}) + \sigma_1(b_{n+1} - b_n))\end{aligned} \tag{23}$$

and

$$\begin{aligned}\partial_t b_n = &c_n(\sigma_3(b_{n+1}^2 + b_n^2 + b_{n+1}b_n + c_{n+1} + c_n + c_{n-1} - 3) + \sigma_2(b_{n+1} + b_n) + \sigma_1) \\ &- c_{n-1}(\sigma_3(b_n^2 + b_{n-1}^2 + b_n b_{n-1} + c_n + c_{n-1} + c_{n-2} - 3) + \sigma_2(b_n + b_{n-1}) + \sigma_1).\end{aligned} \tag{24}$$

B. Relation to the Lax approach

In this section we show that there is a straight relation between the method presented in the previous parts of the paper and the method of Lax pairs.

Define

$$\partial_t L(t) e_n := \partial_t a_{n-1}(t) e_{n-1} + \partial_t b_n(t) e_n + \partial_t a_n(t) e_{n+1} \quad (n \in \mathbb{Z}).$$

This is a densely defined operator in $\ell_2(\mathbb{Z})$, which is indeed the strong derivative of L on $E := \mathrm{span}\{e_n | n \in \mathbb{Z}\}$. An additional operator $U(t)$ in $\ell_2(\mathbb{Z})$ shall be defined by

$$U(t) e_n := 1/2 \sum_{j=1}^{M} (d_{j,n}(t) e_{n-j} - d_{j,n+j}(t) e_{n+j}) \quad (n \in \mathbb{Z})$$

for some numbers $d_{j,n}$ that are to be determined later. We will see that the equation

$$[L,U] + \partial_t L = 0 \tag{25}$$

is equivalent to a system of evolution equations for a and b. The pair (L,U) is then called a Lax pair for these equations. This method of construction of evolution equations has been studied by many authors. A version for the Toda lattice equation can be found for example in Ref. 3. Define $d_{j,n} := 0$ if $j > M$ or $j \leqslant 0$, respectively. An easy computation shows that (25) is equivalent to the following system of equations:

$$a_{n-1} d_{j,n-1} + b_n d_{j+1,n} + a_n d_{j+2,n+1} - a_{n-j-1} d_{j,n} - b_{n-j-1} d_{j+1,n} - a_{n-j-2} d_{j+2,n} = 0, \tag{26}$$

for $j \geqslant 1$,

$$a_{n+1} d_{2,n+2} - a_{n-1} d_{2,n+1} + (b_{n+1} - b_n) d_{1,n+1} = 2 \partial_t a_n \tag{27}$$

and

$$(a_n d_{1,n+1} - a_{n-1} d_{1,n}) = \partial_t b_n. \tag{28}$$

This indeed determines a couple of evolution equations for a_n and b_n. A reformulation of (26) is

$$a_{n-j-1}d_{j,n}-a_{n-1}d_{j,n-1}=(b_n-b_{n-j-1})d_{j+1,n}+a_n d_{j+2,n+1}-a_{n-j-2}d_{j+2,n} \quad (j=1,...,M).$$

The ansatz $d_{j,n}=r_{j,n}(a_{n-1}\cdots a_{n-j})$ leads to:

$$r_{j,n}-r_{j,n-1}=(b_n-b_{n-j-1})r_{j+1,n}+a_n^2 r_{j+2,n+1}-a_{n-j-2}^2 r_{j+2,n} \quad (j=1,...,M), \tag{29}$$

$$r_{j,n}=0 \quad (j>M). \tag{30}$$

Obviously, the sequence $(r_{j,n})_{n\in\mathbb{Z}}$ is uniquely determined up to a constant for every $j=1,...,M$. Equations (27) and (28) now become

$$\partial_t a_n=1/2 a_n(a_{n+1}^2 r_{2,n+2}-a_{n-1}^2 r_{2,n+1}+(b_{n+1}-b_n)r_{1,n+1}) \tag{31}$$

and

$$\partial_t b_n=(a_n^2 r_{1,n+1}-a_{n-1}^2 r_{1,n}). \tag{32}$$

The upshot is that we can explicitly find a fundamental system of solutions of Eqs. (29) and (30).

Theorem III.B.1: For each $1\leqslant k\leqslant M$, the family $X_{j,n}^{(k)}$ of operators satisfies

$$X_{j,n}^{(k)}-X_{j,n-1}^{(k)}=B_n X_{j+1,n}^{(k)}-X_{j+1,n}^{(k)}B_{n-j-1}+C_n X_{j+2,n+1}^{(k)}-X_{j+2,n}^{(k)}C_{n-j-2},$$

where $n\in\mathbb{Z},\ j\geqslant 0$.

Proof: This is the result of a tedious but nevertheless easy computation after rewriting the equation in terms of the operators F_m and using the fact that $(F_{m+k}+F_{m-k})F_m^{-1}$ is independent of m. □

Let $x_{j,n}^{(k)}:=\mathrm{tr}(X_{j,n}^{(k)})$ if $j\neq k$ and $x_{k,n}^{(k)}:=1$. Then $r_{j,n}:=\Sigma_{k=1}^{M}\sigma_k x_{j,n}^{(k)}$ gives the general solution of Eqs. (29) and (30). In addition, Eqs. (31) and (32) are also satisfied as a consequence of Eqs. (19) and (20) and of Theorem III.B.1.

We have seen now that the evolution equations derived in Sec. III A correspond to Lax pairs constructed in this section. The Lax approach provides a second method of derivation of evolution equations by solving Eqs. (29) and (30) directly and substituting the results into Eqs. (31) and (32). Faced with a specific Lax pair, one can obtain the coefficients $\sigma_1,\ldots,\sigma_M$ by taking limits as n grows to infinity. This is implied by the following lemma:

Lemma III.B.2: If $0\leqslant j<k$, then $\lim_{n\to\infty}\mathrm{tr}(X_{j,n}^{(k)})=0$.

Proof: It is

$$BV^m f=(S_1(R_1V_1^m f_1+R_2V_2^m f_2))\oplus(S_2(R_2V_2^m f_2+R_1V_1^m f_1))$$

for arbitrary $f=f_1\oplus f_2\in H$. Clearly $\lim_{m\to\infty}V_1^m=0$. But also $(R_2V_2^m f_2|e_j)=(V_2^{m+j}f_2|h_2)$ and therefore

$$\|R_2V_2^m f_2\|=\left(\sum_{k=0}^{\infty}\left|\int_T \overline{h_2}f_2 z^{m+k}d\mu(z)\right|^2\right)^{1/2}\to 0 \quad (m\to\infty).$$

Thus $\lim_{m\to\infty}BV^m f=0$. Furthermore, we have that $(V^m g|h)\to 0$ as $n\to\infty$, since $g_2\overline{h_2}\in L_2(T)$. It requires a little computation to check that

$$\mathrm{tr}(X_{j,n}^{(k)})=\mathrm{tr}((1+V^{2n+1}BV)^{-1}V^j(V^k-V^{-k})(1+V^{2(n-j)})^{-1}V^{2(n-j)}(h\otimes g)).$$

Now, using the asymptotics assertion of Lemma II.B.4,

$$\lim_{n\to\infty}((1+V^{2n+1}BV)^{-1}V^j(V^k-V^{-k})(1+V^{2(n-j)}B)^{-1}V^{2(n-j)}$$
$$-(1-V^{2n+1}BV)V^j(V^k-V^{-k})(1-V^{2(n-j)}B)V^{2(n-j)})=0,$$

and we infer

$$\lim_{n\to\infty}\operatorname{tr}(X_{j,n}^{(k)})=\lim_{n\to\infty}\operatorname{tr}((1-V^{2n+1}BV)V^j(V^k-V^{-k})(1-V^{2(n-j)}B)V^{2(n-j)}(h\otimes g))=0.$$

□

Concluding this section, we show how to derive Eqs. (23) and (24) by a Lax ansatz for $M=3$: Obviously, $r_{3,n}=\sigma_3$. From

$$r_{2,n}-r_{2,n-1}=(b_n-b_{n-3})r_{3,n}=(b_n-b_{n-3})\sigma_3$$

it follows that $r_{2,n}=\sigma_3(b_n+b_{n-1}+b_{n-2})+\sigma_2$. It remains to solve

$$r_{1,n}-r_{1,n-1}=(b_n-b_{n-2})r_{2,n}+a_n^2r_{3,n+1}-a_{n-3}^2r_{3,n}.$$

Rewriting it as

$$r_{1,n}-r_{1,n-1}=\sigma_3(b_n^2-b_{n-2}^2+b_nb_{n-1}-b_{n-1}b_{n-2}+a_n^2-a_{n-3}^2)+\sigma_2(b_n-b_{n-2})$$

leads to

$$r_{1,n}=\sigma_3(b_n^2+b_{n-1}^2+b_nb_{n-1}+a_n^2+a_{n-1}^2+a_{n-2}^2)+\sigma_2(b_n+b_{n-1})+\tilde r$$

for some constant $\tilde r$. Since then $\lim_{n\to\infty}r_{1,n}=3\sigma_3+\tilde r$, it must be $\tilde r=\sigma_1-3\sigma_3$, and hence:

$$r_{1,n}=\sigma_3(b_n^2+b_{n-1}^2+b_nb_{n-1}+a_n^2+a_{n-1}^2+a_{n-2}^2-3)+\sigma_2(b_n+b_{n-1})+\sigma_1.$$

Now substitution into Eqs. (31) and (32) gives the evolution Eqs. (23) and (24).

At this point we note again the fact that the Lax equations are satisfied gives a plausible hint on the necessary time evolution of the scattering data of the operator L, as investigated in many books and articles on inverse scattering methods. This in turn suggests uniqueness assertions for solutions. However, a rigorous proof or investigation is omitted here.

C. Representation of solutions by determinants

In some cases, it is possible to represent the scalar functions a and b by determinants on operator ideals. Although, under the restrictions of the model we chose in this paper, we can restrict ourselves to ordinary determinants, we still give an outlook onto some generalizations. For the theory of traces and determinants on operator ideals we refer to Pietsch.[13]

Theorem III.C.1: Suppose $B\in\mathcal{A}$ where $\mathcal{A}$ is an quasi-Banach operator ideal admitting a continuous trace τ and a continuous determinant δ, and define $L_n:=V^{2n}B$ for $n\in\mathbb{Z}$. Then

$$c_n=\frac{\delta(1+L_{n+2})\,\delta(1+L_n)}{\delta(1+L_{n+1})^2}.$$

and

$$b_n:=\frac{\delta(1+L_{n+1}+VL_{n+1}-L_{n+1}V^{-1})}{\delta(1+L_{n+1})}-\frac{\delta(1+L_n+VL_n-L_nV^{-1})}{\delta(1+L_n)}.$$

Proof: By the one dimensionality of Q_n:

$$\begin{aligned}1+q_n=1+\mathrm{tr}(Q_n)=1+\tau(Q_n)&=\delta(1+Q_n)\\&=\delta(1+(1+L_n)^{-1}(VL_nV-L_n))\\&=\delta((1+L_n)^{-1}(1+VL_nV))\\&=\frac{\delta(1+VL_nV)}{\delta(1+L_n)}=\frac{\delta(1+L_{n+1})}{\delta(1+L_n)},\end{aligned}$$

which implies the first formula. It is $b_n=\mathrm{tr}(\Gamma_{n+1}-\Gamma_n)=\mathrm{tr}(Q_{n+1}V^{-1}-Q_nV^{-1})$. Similarly to above, we compute

$$\begin{aligned}1+\mathrm{tr}(Q_nV^{-1})=\delta(1+Q_nV^{-1})&=\delta((1+L_n)^{-1}(1+L_n+VL_n-L_nV^{-1}))\\&=\frac{\delta(1+L_n+VL_n-L_nV^{-1})}{\delta(1+L_n)}\end{aligned}$$

and the assertion follows. □

In the reflectionless case, that is if the reflection coefficient vanishes everywhere, the theorem above is applicable inserting the usual determinant on the finite-dimensional operators, viewed, for example, as a restriction of the continuous trace on the quasi-Banach operator ideal of the p-nuclear operators for $0<p\leqslant 2/3$. However, it should be pointed out that the methods of Sec. III A work under the stated requirements, and that one can construct solutions of the derived evolution equations that are not bound to the restrictions of our model.

D. Examples of solved equations

Finally, we use the presented methods to solve the Cauchy problem of two well-known evolution equations.

The Toda lattice equation as stated in Sec. I is given by

$$\partial_{tt}y_n=e^{-(y_n-y_{n-1})}-e^{-(y_{n+1}-y_n)}\quad(n\in\mathbb{Z}).\tag{33}$$

Choosing $\sigma_3:=\sigma_2:=0$ and $\sigma_1:=1$ in Eqs. (23) and (24) they become

$$\partial_tc_n=c_n(b_{n+1}-b_n)\tag{34}$$

and

$$\partial_tb_n=c_n-c_{n-1}.\tag{35}$$

Now, formally substituting $b_n=-\partial_ty_n$, $c_n=\exp(-(y_{n+1}-y_n))$ leads to Eq. (33). Rigorously: We intend to solve Eq. (33) for given initial values y and ∂_ty. For that purpose we assume that

$$b_n(0):=-\partial_ty_n,\quad a_n^2(0)=c_n(0):=e^{-(y_{n+1}-y_n)}$$

satisfy (2). Suppose now that z_n such that $\partial_tz_n=-b_n$ and $z_n(0)=y_n$. Then Eq. (34) implies that:

$$c_n=\gamma_ne^{-(z_{n+1}-z_n)}$$

for some constants γ_n. By the choice of $c_n(0)$, it must be $\gamma_n=1$. From Eq. (35) we finally conclude:

$$\partial_{tt}z_n=e^{-(z_n-z_{n-1})}-e^{-(z_{n+1}-z_n)},$$

i.e., that z is a solution of the initial value problem.

The Langmuir lattice equation can, without loss of generality, be written as:

$$\partial_t g_n = (1+g_n)(g_{n+1}-g_{n-1}) \quad (n\in\mathbb{Z}). \tag{36}$$

Formally: Choose $\sigma_3=0$, $\sigma_2=1$, and $\sigma_1=0$ to obtain the evolution equations

$$\partial_t c_n = c_n(c_{n+1}-c_{n-1}+b_{n+1}^2-b_n^2) \tag{37}$$

and

$$\partial_t b_n = c_n(b_{n+1}+b_n)-c_{n-1}(b_n+b_{n-1}). \tag{38}$$

Substituting $b_n=0$ and $c_n=1+g_n$ gives Eq. (36). Suppose now we are given initial values $1+g_n$. We set $a_n^2(0)=c_n(0):=1+g_n$, $b_n(0):=0$ and require that (2) is fulfilled. In order to have solved the initial value problem, we need to show that $b_n(t)=0$ also for $t\neq 0$:

Step 1. Suppose that $t_0\in\mathbb{R}$ is such that $b_n(t_0)=0$ for all $n\in\mathbb{Z}$. Repeated differentiation and application of Eq. (38) implies that:

$$\left(\frac{d}{dt}\right)^k b_n(t_0)=0 \quad \text{for all } k\in\mathbb{N}_0, n\in\mathbb{Z}.$$

Since b_n is analytical in t, there must be an open neighborhood of t_0 where b_n vanishes for each fixed n.

Step 2. Define

$$N:=\left\{ t\in\mathbb{R} \,\middle|\, \left(\frac{d}{dt}\right)^k b_n(t)=0 \quad \text{for all } k\in\mathbb{N}_0, n\in\mathbb{Z}\right\}.$$

Then N is a closed set, and since $0\in N$, it is not empty. Suppose $t_0\in N$. As we have seen above, there is an open neighborhood U of t_0 such that $b_0(t)=b_1(t)=0$ for all $t\in U$. But then all derivatives of b_0 and b_1 must vanish on U too. Since $c_n(t)\neq 0$ for all $n\in\mathbb{Z}$ and $t\neq 0$, Eq. (38) shows that b_2 as well as b_{-1} and all their derivatives must equal zero on U. Continuing this way we get $U\subset N$. Hence N is open. By the connectedness of $\mathbb{R}$, this means $N=\mathbb{R}$.

Hence the initial value problem of the Langmuir lattice equation is solved.

Note that here, unlike the Toda case, where arbitrary scattering data—unless inversion of the operators fails or $1+q_n(0)\leqslant 0$ for some n—leads to solutions of Eq. (33) via

$$y_n:=-\log(1+q_n),$$

the requirement that $b_n(0)=0$ for the Langmuir model poses a nontrivial restriction on the choice of the scattering data. One can indeed verify by computation, that in general the formula

$$g_n:=c_n-1=\frac{q_{n+1}-q_n}{1+q_n}=\operatorname{tr}((1+Q_n)^{-1}(Q_{n+1}-Q_n)),$$

even if all the inverse operators exist, does not yield solutions of Eq. (36).

IV. CONCLUSION

It has been shown that the methods presented here give solutions of a discrete inverse scattering problem and of the initial value problem of a class of evolution equations that, via the Lax approach, correspond in a rather direct manner to their linear problem. Similar work can be done for different inverse scattering problem and classes of nonlinear evolution equation and is on the way. It is possible, explicitly using the form of the solutions constructed here, to find solutions of the same evolution equations that are not subject to the restrictions of the chosen model. For example, inserting well-chosen diagonal operators and some C_0-semigroup theory in our construction, one can easily derive soliton solutions possessing an infinite number of solitons with interesting properties.

ACKNOWLEGMENT

Research supported by the Deutsche Forschungsgemeinschaft DFG.

[1] Z. S. Agranovich and V. A. Marchenko, *The Inverse Problem of Scattering Theory* (Gordon and Breach, New York, 1963).
[2] C. S. Gardner, J. M. Green, M. D. Kruskal, and R. M. Miura, Phys. Rev. Lett. **19**, 1095 (1967).
[3] H. Flaschka, Prog. Theor. Phys. **51**, 703 (1974).
[4] P. D. Lax, Commun. Pure Appl. Math. **21**, 467 (1968).
[5] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer, Berlin, 1987).
[6] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[7] V. A. Marchenko, *Nonlinear Equations and Operator Algebras* (Reidel, Dordrecht, 1988).
[8] H. Aden and B. Carl, J. Math. Phys. **37**, 1833 (1996).
[9] C. Schiebold, Physica D **2079**, 1 (1998).
[10] B. Carl and C. Schiebold, Nonlinearity **12**, 1 (1999).
[11] J. S. Geronimo and K. M. Case, Trans. Am. Math. Soc. **258**, 467 (1980).
[12] A. M. Berthier, *Spectral Theory and Wave Operators for the Schrödinger Equation*, Research Notes in Mathematics Vol. 71 (Pitman, Boston, 1982).
[13] A. Pietsch, *Eigenvalues and s-Numbers* (Cambridge University Press, Cambridge, 1987).

Generalized Gel'fand–Levitan integral equation for two block Ablowitz–Kaup–Newell–Segur systems

Gulmaro Corona Corona[a)]
Area de Análisis Matemático y sus Aplicaciones, UAM-A Edif. H-122, Depto. C.B., Av. Sn. Pablo 180, Col. Reynosa Tamaulipas, Atzcapotzalco, D.F., CP. 02200, Mexico

(Received 23 March 1998; accepted for publication 30 April 1999)

We derive a generalized Gel'fand–Levitan integral equation for two block Ablowitz–Kaup–Newell–Segur systems. This is possible if we suppose that the matrix coupling coefficients are invertible and come from simple zeroes of the determinant of the diagonal blocks of the scattering matrix. © *1999 American Institute of Physics.* [S0022-2488(99)02908-4]

I. INTRODUCTION

The coupled nonlinear Schrödinger (NLS) equations, a nonlinear evolution equation system, have been solved by means of a generalized Riemann–Hilbert problem[1] before it, in its classical form, were one of the more important ways of recovering the soliton solutions. ''Pseudopotentials'' (integrable coefficients of first-order differential form relation)[2–4] and prolongation structure of Lie groups[5,6] were essential ingredients in the beginnings of the inverse scattering method. By introducing the classical general principle of associating nonlinear evolution equations to linear operators,[7] the inverse scattering method was simplified, improved, and formulated via a Gel'fand–Levitan (GL) integral equation[8] (an approach different from the Riemann–Hilbert problem). Finding suitable linear operators become the focus of the inverse scattering method, which naturally is carried not only to classical but also to matrix GL integral equations;[9,10] in addition, the discovery of connections among matrix linear operators with nonlinear evolution equations contributed to exploring inverse scattering problems for nonlinear evolution equation systems. For instance, systems associated with the matrix operator

$$D_x(J,z)=\frac{d}{dx}-zJ-q, \tag{1}$$

with J an $n\times n$ constant diagonal matrix, have been called in the literature $n\times n$ Ablowitz–Kaup–Newell–Segur (AKNS) systems, which AKNS[11] used to systematically solve the NLS equation via a generalized GL approach with the 2×2 diagonal matrix $J=i\,\mathrm{diag}(1,-1)$.

The first-order linear operator in (1) with the 3×3 constant matrix

$$J=i\begin{pmatrix}2&0&0\\0&-1&0\\0&0&-1\end{pmatrix}$$

is connected with the coupled NLS equations. It has been naturally approached by Riemann–Hilbert problems.[12,13]

High dimensional nonlinear evolution equation systems are closely connected with multidimensional inverse scattering problems,[14] which eventually led to the $\overline{\partial}$ method.[15–18] In general, the GL approach has not been successfully established for dimensions higher than two,[15] generalized

[a)]Electronic mail: ccg@hp9000a1.uam.mx

Riemann–Hilbert problems have been the natural way to formulate the $\overline{\partial}$ approach. However, the relatively complicated structure of generalized GL equations makes them of some interest.

II. THE TWO BLOCK AKNS SYSTEMS

In this paper, generalizing the idea of deriving a GL equation for the NLS equation,[9,10] we shall derive a generalized GL equation for two block AKNS systems,[19] i.e., for the first order $n\times n$ linear operator in (1) with the constant diagonal $n\times n$ matrix

$$J=i\begin{pmatrix} p_-\mathbb{I}_{p_+} & 0_{p_+\times p_-} \\ 0_{p_-\times p_+} & -p_+\mathbb{I}_{p_-} \end{pmatrix}$$

being $p_\pm$ positive integers whose sum is n; $\mathbb{I}_{p_\pm}$ are the $p_\pm\times p_\pm$ identity matrices, respectively; and $0_{p_\pm\times p_\mp}$ are the $p_\pm\times p_\mp$ null matrices, respectively. The two block $n\times n$ AKNS potential q is a $n\times n$ matrix-valued function defined on the real line

$$q=\begin{pmatrix} 0_{p_+} & r_+ \\ r_- & 0_{p_-} \end{pmatrix}, \tag{2}$$

whose entries are in the Schwartz class where $0_{p_\pm}$ are the $p_\pm\times p_\pm$ null matrices and $r_\pm$ are $(p_\pm\times p_\mp)$ matrix-valued functions, respectively. In the sequel, we refer to them as the two block AKNS systems.

The method used here for deriving generalized GL equations cannot be generalized for $n\times n$ AKNS systems, $n\geqslant 2$, where the constant diagonal matrix J has all of its entries distinct, but we use it here to derive a GL equation for the present case provided the coupling matrix coefficients of the potential are invertible and come from simple zeroes of the determinant of the J-diagonal blocks of the scattering matrix.[19] We shall obtain a generalized GL equation by coupling two integral matrix equations into one equation system.

III. COUPLED MATRIX INTEGRAL EQUATIONS

We need to get suitable representations for the wave functions of the two block AKNS system. For this purpose, let us adopt the same notation as that used in Ref. 19 and recall some results established there: construct the wave functions $\psi_\pm$ with asymptotic behavior $\psi_\pm \sim e^{\pm ip_\pm xz}e_\pm$ as $x\to\infty$ for the two block AKNS systems.

With this at hand, we may get the representations for these wave functions,

$$\psi_\pm(x,z)=e_\pm e^{\pm ip_\pm xz}+\int_x^\infty G_\pm(x,\zeta)e^{\pm ip_\pm z\zeta}\,d\zeta, \tag{3}$$

where $e_\pm$ are $n\times p_\pm$ constant matrices such that the identity matrix $I_n=\|e_+e_-\|$.[19]

The kernels $G_\pm$ both have support in the region $\zeta>x$. For example, the matrix-valued function $\psi_-e^{ip_-xz}-e_-$ tends to zero, as z tends to infinity, in the upper half plane, so by the Paley–Wiener theorem,

$$\psi_-e^{ip_-xz}-e_-=\frac{1}{2\pi}\int_{-\infty}^0 K_-(x,\zeta)e^{iz\zeta}\,d\zeta,$$

that is,

$$\psi_- = e_- e^{-ip_- xz} + \int_{-\infty}^{0} K_-(x,\zeta) e^{iz\zeta - ip_- zx}\, d\zeta$$
$$= e_- e^{-ip_- xz} + \int_{-\infty}^{0} K_-(x,\zeta) e^{iz(\zeta - p_- x)} d\zeta$$
$$= e_- e^{-ip_- xz} + p_- \int_{x}^{\infty} K_-(x, p_-(x-\zeta)) e^{-izp_- \zeta}\, d\zeta.$$

So the result for ψ_- follows by taking

$$G_-(x,\zeta) = \frac{p_-}{2\pi} K_-(x, p_-(x-\zeta)).$$

The matrix functions $G_\pm$ are $n \times p_\pm$-matrix valued functions, respectively.

Write the relations[19] between the wave functions $\psi_\pm$ and those $\phi_\pm$ with asymptotic behavior $\phi_\pm \sim e^{\mp ip_\mp xz} e_\mp$ as $x \to -\infty$ in the following form:

$$\phi_+ d^{-1} = \psi_+ b d^{-1} + \psi_-,$$
$$\phi_- a^{-1} = \psi_+ + \psi_- c a^{-1}, \tag{4}$$

where a,b,c,d are the J-blocks of the scattering matrix[12]

$$s = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Combining the first relation with the representations for $\psi_\pm$, write the first equation as

$$\phi_+ d^{-1} - e_- e^{-ip_- xz} = \int_x^{\infty} G_-(x,\zeta) e^{-ip_- z\zeta}\, d\zeta + \left(e_+ e^{ip_+ xz} + \int_x^{\infty} G_+(x,\zeta) e^{ip_+ z\zeta}\, d\zeta \right) b d^{-1}. \tag{5}$$

Multiplying by $(2\pi)^{-1} e^{ip_- z\zeta}$ and integrating with respect to z over the real line, an integral matrix integral equation may be derived. To do this, we define the matrix valued function and the rational number, respectively, by

$$f_1^+(\zeta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} b d^{-1} e^{ip_+ \zeta z}\, dz, \quad p = \frac{p_+}{p_-}.$$

With this at hand, we get the following relations for each term in (4):

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ip_- z\zeta} \int_x^{\infty} G_-(x,\eta) e^{-ip_- z\eta}\, d\eta\, dz = G_-(x,\zeta),$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} b d^{-1} e^{ip_- \zeta z} e^{ip_+ xz}\, dz = f_1^+(x + p^{-1}\zeta),$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ip_- z\zeta} \int_x^{\infty} G_+(x,\eta) e^{ip_+ z\eta} b d^{-1}\, d\eta\, dz = \int_x^{\infty} G_+(x,\zeta) \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ip_+ z\eta} e^{ip_- z\zeta} b d^{-1} dz\, d\eta$$
$$= \int_x^{\infty} G_+(x,\eta) f_1^+(\eta + p^{-1}\zeta) d\eta.$$

The left-hand side can be evaluated by closing the contour in the upper half plane for $\zeta > x$ and using the residue theorem:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ip_-z\zeta}(\phi_+ d^{-1} - e_- e^{-ip_-xz})dz$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(\zeta-x)p_-z}(\phi_+ d^{-1}e^{ip_-xz} - e_-)dz = i\sum_{j=1}^{N} \phi_{+j}d'^{-1}(z_j)e^{ip_-z_j\zeta}$$
$$= -\sum_{j=1}^{N} \phi_{+j}(id')^{-1}(z_j)e^{ip_-z_j\zeta} = -\sum_{j=1}^{N} \psi_{+j}c_j^+(id')^{-1}(z_j)e^{ip_-z_j\zeta}$$
$$= -\sum_{j=1}^{N}\left(e_+e^{ip_+xz_j} + \int_x^\infty G_+(x,\eta)e^{ip_+z_j\eta}\,d\eta\right)c_j^+(id')^{-1}e^{ip_-z_j\zeta}$$
$$= -e_+\sum_{j=1}^{N} c_j^+(id')^{-1}e^{i(p_+x+p_-\zeta)z_j} - \int_x^\infty G_+(x,\eta)\left(\sum_{j=1}^{N} c_j^+(id')^{-1}e^{i(p_+\eta+p_-\zeta)z_j}\right)d\eta$$
$$= -f_2^+(x+p^{-1}\zeta) - \int_x^\infty G_+(x,\eta)f_2^+(\eta+p^{-1}\zeta)\,d\eta,$$

where the $p_+\times p_-$ matrix function

$$f_2^+(\zeta) = \sum_{j=1}^{N} c_j^+(id')^{-1}e^{ip_+z_j\zeta}.$$

Putting this together, we get

$$f_+(x+p^{-1}\zeta) + G_-(x,\zeta) + \int_x^\infty G_+(x,\eta)f_+(\eta+p^{-1}\zeta)\,d\eta = 0 \tag{6}$$

with $f_+ = f_1^+ + f_2^+$. Here, z_j^+ are the (simple) zeroes of d in the upper half plane and c_j^+ are the coupling coefficients[19]

$$\phi_+(x,z_j^+) = \psi_+(x,z_j^+)c_j^+ .$$

Now combining the second relation with the representations for $\psi_\pm$, we can write the second equation as

$$\phi_- a^{-1} = \psi_+ + \psi_- c a^{-1}.$$

Consequently,

$$\phi_- a^{-1} - e_+e^{ip_+xz} = \int_x^\infty G_+(x,\zeta)e^{ip_+z\zeta}\,d\zeta + \left(e_-e^{-ip_-xz} + \int_x^\infty G_-(x,\zeta)e^{-ip_-z\zeta}\,d\zeta\right)ca^{-1}.$$

Another integral equation system involving the matrix-valued functions $G_\pm$ is obtained by multiplying by $(2\pi)^{-1}e^{-ip_+z\zeta}$ and integrating with respect to z over the real line. Write

$$f_1^-(\zeta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} ca^{-1}e^{-ip_-\zeta z}\,dz.$$

Using this definition, we arrive at

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ip_+ z\zeta}\int_x^{\infty} G_+(x,\eta)e^{ip_+ z\eta}\,d\eta\,dz = G_+(x,\zeta),$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} ca^{-1}e^{-ip_+\zeta z}e^{-ip_- xz}\,dz = f_1^-(x+p\zeta),$$

and

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ip_+ z\zeta}\int_x^{\infty} G_+(x,\eta)e^{-ip_- z\eta}bd^{-1}\,d\eta\,dz$$
$$=\int_x^{\infty} G_+(x,\zeta)\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ip_+ z\eta}e^{-ip_-\zeta\eta}\,bd^{-1}\,dz\,d\eta$$
$$=\int_x^{\infty} G_+(x,\eta)f_1^-(\eta+p\zeta)d\eta.$$

The left-hand side can be evaluated by closing the contour in the lower half plane for $\zeta > x$ and using the residue theorem:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ip_+ z\zeta}(\phi_- a^{-1} - e_+ e^{ip_+ xz})dz$$
$$=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i(\zeta-x)p_+ z}(\phi_- a^{-1}e^{-ip_+ xz} - e_-)dz = i\sum_{j=1}^{N}\phi_{-j}a'^{-1}(z_j)e^{-ip_+ z_j\zeta}$$
$$=-\sum_{j=1}^{N}\phi_{-j}(ia')^{-1}(z_j)e^{-ip_+ z_j}\zeta$$
$$=-\sum_{j=1}^{N}\psi_{-j}c_j^-(ia')^{-1}(z_j)e^{-ip_+ z_j\zeta}$$
$$=-\sum_{j=1}^{N}\left(e_- e^{-ip_- xz_j} + \int_x^{\infty} G_-(x,\eta)e^{-ip_- z_j\eta}\,d\eta\right)c_j^-(ia')^{-1}e^{-ip_+ z_j\zeta}$$
$$=-e_+\sum_{j=1}^{N}c_j^-(ia')^{-1}e^{-i(p_- x+p_+\zeta)z_j} - \int_x^{\infty} G_-(x,\eta)\left(\sum_{j=1}^{N}c_j^-(ia')^{-1}e^{-i(p_-\eta+p_+\zeta)z_j}\right)d\eta$$
$$=-e_- f_2^-(x+p\zeta) - \int_x^{\infty} G_-(x,\eta)f_2^-(\eta+p\zeta)d\eta,$$

where the $p_-\times p_+$ matrix function $f_2^-(\zeta)=\Sigma_{j=1}^{N} c_j^+(ia')^{-1}e^{-ip_- z_j\zeta}$. Proceeding as before, we get

$$f_-(x+p\zeta)+G_-(x,\zeta)+\int_x^{\infty} G_+(x,\eta)f_-(\eta+p\zeta)d\eta = 0$$

with $f_- = f_1^- + f_2^-$. Here, z_j^- are the (simple) zeroes of d in the upper half plane and c_j^- are the coupling coefficients:[19]

$$\phi_-(x,z_j^-) = \psi_-(x,z_j^-)c_j^-.$$

IV. GENERALIZED GEL'FAND–LEVITAN INTEGRAL EQUATION

Coupling the integral equations, we get the generalized GL equation

$$f(x+p^{\pm}\zeta)+G(x,\zeta)\tau+\int_x^{\infty}G(x,\eta)f(\eta+p^{\pm}\zeta)d\eta=0, \tag{7}$$

where

$$f(x+p^{\pm}\zeta)=\begin{pmatrix} f_+(x+p^{-1}\zeta) & 0_{p_+\times p_+} \\ 0_{p_-\times p_-} & f_-(x+p\zeta) \end{pmatrix},$$
$$\tau=\begin{pmatrix} 0_{p_+\times p_-} & \mathbb{I}_{p_+\times p_+} \\ \mathbb{I}_{p_-\times p_-} & 0_{p_-\times p_+} \end{pmatrix}. \tag{8}$$

On the other hand, taking into account Eq. (3) and the definition in (8) for the matrix-valued function G, get

$$\Psi(x,z)=\|\psi_-(x,z)\psi_+(x,z)\|=e^{xzJ}+\int_x^{\infty}G(x,\zeta)e^{z\zeta J}\,d\zeta.$$

Evaluating $D_x(J,z)\Psi$ in two ways, one obtains after reducing terms

$$[G(x,x)+q]e^{xzJ}+\int_x^{\infty}[zJG(x,\zeta)+qG(x.\zeta)-G_x(x,\zeta)]e^{z\zeta J}d\zeta=0.$$

Now integrating by parts $\int_x^{\infty}zJG(x,\zeta)e^{z\zeta J}\,d\zeta$ and replacing it in the above equation, write

$$[G(x,x)+q-JG(x.x)J^{-1}]+\int_x^{\infty}[qG(x.\zeta)-G_x(x,\zeta)-JG_z(x,\zeta)J^{-1}]e^{z(\zeta-z)J}\,d\zeta=0.$$

If the potential has compact support, then the integrand is in $L_1(x,\infty)$. So by the Riemann–Lebesgue Lemma the integral term goes to zero as z goes to ∞. Since the first term is independent of z, we have that both terms must vanish. Hence

$$q(x)=J^{-1}[J,G(x,x,)],\quad q(x)G(x,z)=G_x(x,z)J+JG_z(x,z). \tag{9}$$

Since the potentials of compact support are dense and G, G_x, G_z depends continuously[19] on q, the equations in (9) are valid for other potentials.

To state our next result, we denote by $\mathbf{P}$[19] the set of the matrix-valued functions q defined on the real line given by (2).

Theorem: *If the potential q is in* **P** *then the entries of the matrix f of (5) are in the Schwartz class. In addition if the coupling coefficients*[19] *associated with q are invertible and come from simple zeroes of the determinant J-diagonal entries*[12] *of the scattering matrix, the matrix-valued function G satisfies the GL integral equation system (7) and q is related to G by the formulas in (9).*

Remark: Since the entries of the matrix f in (7) are rational functions of the entries of the scattering matrix then their denominators are lower bounded.[18] So the entries of f are in the Schwartz class because of belonging scattering matrix ones to the same class

Comment: Calling a type of integral equation as that given in (7) the ''Gel'fand–Levitan–Marchenko equation'' is common in the nonlinear evolution literature, however, Gel'fand–Levitan and Marchenko integral equations are different and describe distinct problems. For details see, e.g., Ref. 20.

V. CONCLUSIONS

It is possible to approach the two block AKNS systems by a generalized GL equation integral system under the hypothesis that the coupling matrix coefficients are invertible and come from simple zeroes of the determinant of $\det a$ and $\det d$. In fact, we can relate the potential q with the solution G of the GL equation system in (7) by the formulas in (9). The case $p_+ = 1$ and $p_- = 2$ is involved with the coupled nonlinear Schrödinger equations for which Riemann–Hilbert approaches have been formulated. In general, generalized GL equations are not successfully formulated for nonlinear evolution equation systems of dimension greater that two,[21] but for present $n \times n$ AKNS systems, this is possible with a relative simpleness and by a short calculation.

[1] R. G. Newton and R. Jost, Nuovo Cimento **1**, 590 (1955); see also the introduction to Ref. 14 and references therein.
[2] H. D. Wahlquist and F. B. Estabrook, ''*Korteweg–de Vries equation*,'' Phys. Rev. Lett. **31**, 1386 (1973).
[3] H. D. Wahlquist and F. B. Estabrook, ''*Prolongation structures of nonlinear evolution equations*,'' J. Math. Phys. **16**, 1–7 (1975).
[4] H. D. Wahlquist and F. B. Estabrook, ''*Prolongation structures of nonlinear evolution equations. II*,'' J. Math. Phys. **17**, 1293–1297 (1976).
[5] J. Corones, ''*Solitons and simple pseudopotentials*,'' J. Math. Phys. **17**, 756–759 (1976).
[6] J. Corones, ''*A Lie group framework for soliton equations. I. Path independence case*,'' J. Math. Phys. **11**, 2206–2213 (1977).
[7] P. D. Lax, ''*Integrals of nonlinear equations of evolution and solitary waves*,'' Physica D **1**, 467–490 (1980).
[8] D. J. Kaup, ''*The Estabrook–Wahlquist method with examples of applications*,'' Physica D **1**, 391–411 (1980).
[9] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, ''*Method for solving the Sine–Gordon equation*,'' Phys. Rev. Lett. **30**, 1262–1264 (1973).
[10] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, ''*Nonlinear-evolution equations of physical significance*,'' Phys. Rev. Lett. **30**, 125–127 (1974).
[11] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, ''*The inverse scattering transform-Fourier analysis for nonlinear analysis*,'' Stud. Appl. Math. **53**, 249–315 (1973).
[12] G. Corona-Corona, ''*Hamiltonian structure for the degenerate AKNS systems*,'' J. Math. Phys. **38**, 212–225 (1997).
[13] S. V. Manakov, ''*On the theory of two-dimensional stationary self-focusing of electromagnetic waves*,'' Sov. Phys. JETP **38**, 248–253 (1985).
[14] R. G. Newton, ''*On a generalized Hilbert problem*,'' J. Math. Phys. **23**, 2257–2264 (1982).
[15] R. Beals and R. R. Coifman, ''*The D-bar approach to inverse scattering and nonlinear evolutions*,'' Physica D **18**, 242–259 (1986).
[16] S. P. Novikov and G. M. Khenkin, ''*∂-equation in the multidimensional inverse problem*,'' Russian Math. Surveys **42**, 109–180 (1986).
[17] S. P. Novikov and G. M. Khenkin, ''*∂-equation in the multidimensional inverse problem*,'' Russian Math. Surveys **42**, 109–180 (1986).
[18] A. I. Nachman and M. J. Ablowitz, ''*A Muldimensional Inverse-Scattering Method*,'' Stud. Appl. Math. **71**, 243–250 (1984).
[19] G. Corona-Corona, ''*Couple matrix coefficients for the two block* AKNS *system*,'' J. Math. Phys. **39**, 2808–2812 (1998).
[20] R. G. Newton, ''*Inverse scattering.* I. *One dimension*,'' J. Math. Phys. **21**, 493–505 (1979).
[21] N. Sasa and J. Satsuma, ''*New type of soliton solutions for a higher order nonlinear Schrödinger equation*,'' J. Phys. Soc. Jpn. **60**, 409–417 (1987).

Irreversible weak limits of classical dynamical systems

F. Gentili
Dipartimento di Fisica dell'Università and Istituto Nazionale di Fisica Nucleare, Bologna, Italy

G. Morchio
Dipartimento di Fisica dell'Università and Istituto Nazionale di Fisica Nucleare, Pisa, Italy

(Received 2 March 1999; accepted for publication 9 April 1999)

A general discussion is given of weak limits of classical dynamical systems depending on a parameter. The resulting maps are shown to be invertible if and only if they define a group of measure preserving point transformations. The irreversible case automatically leads to positive bistochastic maps and is characterized in terms of convergence properties of the corresponding automorphisms of the observable algebra. Necessary and sufficient conditions are given for the limit to define a time-independent Markov process. Two models are discussed, for a particle in a periodic potential, and for a particle interacting with fixed configurations of external obstacles. © *1999 American Institute of Physics.* [S0022-2488(99)03208-9]

I. INTRODUCTION

Irreversible time evolution plays a major and generic role in the phenomenology of macroscopic systems. While its general description seems to be well established and understood in terms of stochastic maps and processes,[1–3] the derivation of irreversible behavior in dynamical systems usually makes reference to specific physical situations, and is under mathematical control only in special cases;[4,5] it is therefore not clear whether the mechanisms observed in such models may provide a general mathematical explanation of the origin of irreversibility on the same level as its general characterization.[6–9]

Many different structures are in fact present in models, i.e., a large (infinite) number of degrees of freedom, dynamical instabilities, restrictions on the states appearing ''at time zero,''[10–14] distinctions between ''macroscopic'' and ''microscopic'' observables, or between ''subsystem of interest'' and ''environment,''[15] moreover statistical arguments always play a critical role, starting with Boltzman's analysis.[16,12,13]

Since in all cases irreversible behavior is obtained as a limit description in the variation of some parameter, it is worthwhile to ask whether this ingredient is also sufficient, i.e., whether stochastic maps and processes *are the generic result of taking such limits*. The other ingredients can then be interpreted as characterizing specific classes of physical situations, where specific parameters and limits may play a significant physical role.

In models, the above-mentioned parameters appear in the description of the structure of the system (the number of particles in the Boltzmann–Grad case), of the states (the correlation functions at time zero), and of the dynamics (the rescaling of the particle radius); since dynamical instabilities seem to be an essential ingredient of irreversibility, it is reasonable to concentrate the attention on parameters which appear in the time evolution, leaving for the moment the system and its states invariant. Such a possibility is not even far from the situation in the Boltzmann–Grad limit since, in the spatially homogeneous case, that limit can be formulated in terms of an infinite system of hard spheres, in a fixed time zero configuration, in terms of a suitable rescaling of the radius ($r \to \lambda r$) and of the time scale ($t \to \lambda^{-2} t$ in three space dimensions).

One is therefore led to investigate limits of sequences of reversible dynamical systems. On the basis of the general interpretation of dynamical systems in terms of observables and states, the most general relevant limit is the limit for the result of measurements of fixed observables, given any state in some relevant class.

In the Hamiltonian case, observables may be identified with regular functions on the phase space, states with probability measures, and one has to take limits of mean values of observables over states evolved in time under groups of transformations defined by Hamiltonians depending on a parameter. More generally, i.e., for quantum systems and infinite systems, observables may be identified (see, e.g., Ref. 17) with the Hermitean elements of a C^* algebra, states with positive linear functionals, time evolution with a group of automorphisms of the observable algebra, and the relevant limit is the weak limit of the time evolution automorphisms, taken with respect to a subspace of the state space, interpreted as the set of states of interest.

In this form, the problem is similar to the one that appears in the framework of the algebraic formulation of quantum field theory and statistical mechanics,[17–18] in relation with the construction of the time evolution of infinite quantum systems in the presence of long-range interactions. In fact, a general analysis of limits of automorphisms arising in that context has been developed in Refs. 19–22. The relatively strong assumptions which had to be required in that analysis in order that the limit still give rise to a group of automorphisms, i.e., to a reversible dynamical system, leave open the possibility that the result of such limits may be interpreted, in some generality, in terms of irreversible time evolution.

The purpose of the present paper is to show that this is the case, and that in fact only two possibilities arise for weak limits of classical dynamical systems: either they still define groups of measure preserving transformations, or they give rise to positive ''bistochastic'' maps, and to stochastic processes if the limits exist for the results of observations at different times.

We restrict our attention in this paper to classical dynamical systems, under assumptions which typically hold for Hamiltonian models with a finite number of degrees of freedom. Our discussion covers the (realistic) case of noncompact phase space, with infinite invariant measure. In the case of a *finite* invariant measure, part of the results of Proposition 1 has been derived, in a rather different context, by Brown,[23] together with a density result which implies that, under suitable restrictions on the structure of the measure, *all* bistochastic maps arise as a weak limit of measure preserving invertible transformations. We also discuss the implications of the group structure in the time parameter, which is preserved if and only if the limit maps are invertible, the necessity of dynamical instability for an irreversible limit, and show that stochastic processes arise in general from taking weak limits at different times, providing a characterization of the Markov case.

In Sec. II we make our framework more precise and give the main results; Sec. III contains the proofs; in Sec. IV two models are discussed, the first exhibiting the role of weak limits in simple Hamiltonian systems, the second bearing some similarity to the Lorentz gas. In both cases, stochastic maps are obtained in explicit form, as weak limits of strictly deterministic and reversible dynamical systems, reproducing the general structures discussed in Sec. II; for the second model the result defines a time-independent Markov process.

II. ASSUMPTIONS AND RESULTS

We consider classical dynamical systems, defined as in Ref. 24 by groups of continuous transformations Γ_t, $t \in \mathbb{R}$ of a locally compact Hausdorff topological space X. We will discuss continuous transformations Γ_n^t depending on a parameter $n \in \mathbb{N}$, and their limits as $n \to \infty$.

The observables of the system are identified with the C^* algebra of the continuous functions on X vanishing at infinity, $C^0(X)$, i.e., the completion, in the sup norm, of $C_0^0(X)$, the space of continuous functions of compact support. The transformations Γ_n^t define as usual groups of automorphism α_n^t of $C^0(X)$, $\alpha_n^t f(x) = f(\Gamma_n^t(x))$.

Conversely, by Gelfand's isomorphism,[25] any group of automorphisms of a commutative C^* algebra defines a group of continuous transformation of its spectrum, a locally compact Hausdorff space (the set of multiplicative linear functionals being clearly left invariant by the transpose of an automorphism).

The states of the system are the positive normalized linear functional ω on $C^0(X)$, i.e., by the Riesz–Markov theorem, the regular Borel probability measures on X.

The only restrictions on α_n^t that we will assume are the existence of a common invariant measure, and a condition which keeps the time evolution ''far from infinity'' uniformly in n; more specifically, we will assume:

(M) a σ finite regular Borel measure μ is defined on X, and left invariant by all Γ_n^t, $n \in \mathbb{N}$;

(F) for all real t and compact set $K \subset X$ there is a compact set $M(K,t) \subset X$ such that

$$\Gamma_n^t(K) \subset M(K,t), \qquad \forall n \in \mathbb{N}.$$

Given the invariant measure μ, a distinguished class of states is given by the set of probability measures ν which are continuous with respect to μ: $\nu(x) = h(x)\mu(x)$, $h(x) \in L^1(X,d\mu)$. The set $\mathcal{F}$ of such states, which will be called ''regular,'' is the positive part of a closed subspace of the dual of $C^0(X)$, and defines a ''full folium'' of states in the sense of Ref. 26. Regular states will be identified in the following with functions in $L^1(X,d\mu)$.

The invariance of $d\mu$ implies that $\mathcal{F}$ is left invariant by the transpose α_n^{t*} of α_n^t, defined by $\alpha_n^{t*}\omega(f) = \omega(\alpha_n^t f)$. It also implies that time evolution can be formulated ''a la Koopman''[27] in $L^2(X,d\mu)$: $\forall g \in L^1 \cap L^2$, $U_n^t g$ is defined by $\alpha_n^{t*} g$, and extends by continuity to a unitary group in $L^2(X,d\mu)$.

The above general structure covers in particular the case of Hamiltonian systems, for which X is the phase space, Γ^t the Hamiltonian time evolution, $d\mu$ the (infinite) Liouville measure, and the regular states are the probability measures on the phase space with well-defined (L^1) density.

We do not assume that $d\mu$ is a finite measure; in fact, even in the Hamiltonian case, a sequence of groups of transformations does not in general leave invariant a common compact subspace, and there is no common invariant finite measure. For the same reason, X is assumed to be only locally compact.

Continuity of Γ_n^t is assumed only for simplicity [it allows for a discussion in terms of $C^0(X)$]. Propositions 1–5 still hold, with minor changes in the proofs, for groups Γ_n^t of *measurable* transformations; in this case, $\alpha_n^t(f) \equiv f(\Gamma_n^t(x))$ obviously define automorphisms of $L^\infty(X,d\mu)$.

The advantage of the above formulation is to make explicit the role of the observables and of the states of the system, so that the discussion of the limit $n \to \infty$ can be done with a clear physical interpretation.

The existence of a *limit description* of the dynamical systems (X,Γ_n^t) as $n \to \infty$, given a time zero state ω, amounts in fact to convergence of all the mean values $\omega(\alpha_n^t f)$, $f \in C^0(X)$. As we will see (Proposition 1), this is also equivalent to convergence of $\omega(\alpha_n^t \chi)$, for all characteristic functions χ of μ measurable subsets of X, i.e., convergence of the frequence of answers for all yes–no experiments based on μ measurable subsets of X, given the state ω.

Since we will admit as time zero states all regular states, we will investigate the limit of α_n^t in the weak topology $\tau_{\mathcal{F}}$ defined by $\mathcal{F}$ on $C^0(X)$. $\tau_{\mathcal{F}}$ can also be described as the ultraweak topology[28] associated with the representation of $C^0(X)$ as multiplication operators in $L^2(X,d\mu)$, and the $\tau_{\mathcal{F}}$ closure of $C^0(X)$ can be identified with $L^\infty(X,d\mu)$.

In the following, we will also use the strong and ultrastrong topologies defined on $C^0(X)$ and $L^\infty(X,d\mu)$ by the same representation; for bounded sequences f_n, the strong and ultrastrong topology coincide, and f_n converge to f iff

$$\int |f_n(x) - f(x)|^2 h(x) d\mu(x) \to 0, \quad \forall h(x) \in L^1(X,d\mu).$$

We will assume that the $\tau_{\mathcal{F}}$ weak limit of α_n^t exists for all t, and investigate under which condition it can be interpreted as an irreversible time evolution. More detailed results, which do not assume convergence for all times, are given in Ref. 29. The following result holds for any $\tau_{\mathcal{F}}$ convergent sequence of automorphisms of $C^0(X)$:

Proposition 1: If a sequence of automorphisms α_n^t of $C^0(X)$, satisfying M, converges in the weak topology $\tau_{\mathcal{F}}$, for all $t\in\mathbb{R}$, then

(i) the limit defines $\tau_{\mathcal{F}}$ continuous linear positive maps α^t of $C^0(X)$ into $L^\infty(X,d\mu)$, with norm at most one;

(ii) the transpose α^{t*} of α^t leaves $\mathcal{F}$ invariant;

(iii) α^t extends by continuity to a linear positive mapping of all $L^\infty(X,d\mu)$ into itself.

(iv) $\alpha^{t*}=\alpha^{-t}$ on $L^1\cap L^\infty$;

(v) $\forall p\in[1,\infty)$, $\forall h\in L^p\cap L^1$, $\|\alpha^{t*}h\|_p\leqslant\|h\|_p$.

If the sequence α_n^t also satisfies property F, then

(vi) α^{t*} preserve the L^1 norm on $(L^1\cap L^\infty)^+$, and α^t leaves the identity invariant;

(vii) the $\tau_{\mathcal{F}}$ continuous extension of α_n^t to $L^\infty(X,d\mu)$ also converges $\tau_{\mathcal{F}}$ for $n\to\infty$ to α^t on L^∞.

Positive maps are well-known candidates for the description of irreversible evolution. Moreover, property (vi) characterizes *bistochastic* maps, and it is generally assumed as fundamental[2,3] for an axiomatic description of irreversible processes.

Proposition 1 is however compatible with limits still defining groups of automorphisms of $L^\infty(X,d\mu)$. In order to characterize the irreversible case, one must ask whether the mapping α^t are invertible, and whether the group law $\alpha^t\alpha^s=\alpha^{t+s}$ is satisfied. *A priori*, as candidates for the description of irreversible time evolution arising in the above limit we have therefore:

(i) failure of the group law for α^t,

(ii) failure of invertibility of α^t.

It is also clear that

(iii) failure of the morphism property of α^t

gives a relevant notion of irreversible behavior. In fact, the transpose α^{t*} of a linear positive map α^t of any Abelian C^* algebra, with invariant identity, sends pure states into pure states if and only if α^t is a morphism, as is easily seen by using the Gelfand construction. Property (iii) therefore has, in general, the interpretation of a loss of information on the system, corresponding to a transformation of pure into mixed states. More concretely, a linear bounded transformation of $L^\infty(X,d\mu)$ is a morphism iff it maps characteristic functions of measurable sets into themselves, i.e., iff all yes–no questions are transformed into yes–no questions, equivalently, no information is lost in the time evolution.

Moreover, in the same generality, both *(ii)* and *(iii)* imply *(i)*. Only *(iii)* ⇒ *(i)* is in fact nontrivial, and follows from the fact that if the α^t satisfies the group law, then α^{-t} inverts α^t and is a positive map, so that α^{-t*} sends mixed states into mixed states, and therefore invertibility implies that α^{t*} sends pure states into pure states.

The following Proposition shows that in the present framework the notions (i)–(iii) coincide, so that *there is a unique notion of irreversibility for weak limits of classical dynamical systems*; it also provides necessary and sufficient conditions for the appearance of irreversibility.

Proposition 2: The following properties are equivalent for the weak limit α^t of a sequence of automorphisms α_n^t of $C^0(X)$, satisfying M:

(i) $\forall t\in\mathbb{R}$, α^t is an automorphism of $L^\infty(X,d\mu)$;

(ii) α^t, $t\in\mathbb{R}$, is a one parameter group of automorphisms of $L^\infty(X,d\mu)$;

(iii) $\forall t\in\mathbb{R}$, α_n^t converges in the ultrastrong topology associated to the representation of $C^0(X)$ as multiplication operators in $L^2(X,d\mu)$, and $\alpha^t(1)=1$.

(iv) For some $p\in(1,\infty)$ $\alpha^t:C_0^0(X)\mapsto L^\infty(X,d\mu)$ preserves the L^p norm, for all t,

$$\|\alpha^t f\|_p=\|f\|_p \quad \forall f\in C_0^0\,(X),\quad t\in\mathbb{R}.$$

(v) For some $p\in(1,\infty)$ α^{t*} preserves the L^p norm, for all t:

$$\|\alpha^{t*}h\|_p=\|h\|_p \quad \forall h\in L^1\cap L^p,\quad t\in\mathbb{R}.$$

(vi) For all $p\in[1,\infty]$, for all t, α^{t*} preserves the L^p norm.

Similar results hold for the convergence of the unitary group U_n^t which implement the time evolution in $L^2(X,d\mu)$:

Proposition 3: Assuming M, weak $\tau_{\mathcal{F}}$ convergence of α_n^t is equivalent to convergence of U_n^t in the weak operator topology in $L^2(X,d\mu)$. Assuming convergence and denoting by U^t the weak limit of U_n^t, conditions (i)–(vi) of Proposition 2 are equivalent to each of the following:

(i) $\forall t \in \mathbb{R}$ the operators U_n^t converge strongly in $L^2(X,d\mu)$;

(ii) $\forall t \in \mathbb{R}$ the operators U^t are unitary.

(iii) The operators U^t form a one-parameter unitary group.

The following result shows that dynamical instabilities growing with n (at fixed time) are necessary for irreversible behavior to take place in the limit $n \to \infty$:

Proposition 4: Assume that X is a metric space, with distance $d(x,y)$, and that the maps Γ_n^t, satisfying M and F, are stable, at fixed time, uniformly in n, except possibly on ''locally small subsets''; i.e., assume: $\forall$ compact $K \subset X$, $\epsilon > 0$, $t \in \mathbb{R}$, $\exists M(K,\epsilon,t) \in \mathbb{R}$ such that

$$d(\Gamma_n^t(x),\Gamma_n^t(y)) < M d(x,y) \quad \forall n \in \mathbb{N}, \quad \forall x,y \in K \backslash S,$$

$S = S(K,\epsilon,t)$ an open subset of K of μ measure smaller than ϵ.
Then, if the sequence α_n^t converges in the weak $\tau_{\mathcal{F}}$ topology, it also converges strongly and α^t is a group of automorphisms of $L^\infty(X,d\mu)$.

In general, the map α^t has an immediate probabilistic interpretation: Given any state $h \in L^1(X,d\mu)$ and denoting by χ_B the characteristic function of a Borel set $B \subset X$, the mean value

$$\int h(x) \alpha^t \chi_B(x) \, d\mu(x)$$

gives by definition the probability that a point distributed with $h(x) d\mu(x)$ at time 0 ends in B at time *t*. $\alpha^t \chi_B(x)$ is therefore interpreted as a probability $P_t(x,B)$ that a point starting at x at time 0 falls in B at time *t*.

Since $\alpha^t \chi_B$ is the characteristic function of a measurable set for all B if and only if α^t is an automorphism of $L^\infty(X,d\mu)$ (see Proposition 2), $P_t(x,B)$ takes values in $\{0, 1\}$ ($d\mu$ almost every) if and only if there is no irreversibility.

The situation is therefore very close to that of a stochastic process, and an examination of limits at different times is of interest. Consider in fact finite sequences of measurements at different times, $t_1 \cdots t_k$, of observables $f_1 \cdots f_k$, and assume that, given a regular state $\omega \in \mathcal{F}$, the mean values

$$\omega(\alpha_n^{t_1} f_1 \cdots \alpha_n^{t_k} f_k)$$

converge for $n \to \infty$.

The following Proposition shows that, if such a limit exists, it defines a stochastic process, i.e., a measure $d\rho_\omega(x(t))$ on trajectories in X, and gives necessary and sufficient condition for the process to be Markov.

Proposition 5: Let α_n^t satisfy M and F and converge weakly for $n \to \infty$; assume that for a fixed $\omega \in \mathcal{F}$, and for all finite sequences $f_1 \cdots f_k \in C^0(X)$ and $0 \leq t_1 < t_2 < \ldots < t_k \in \mathbb{R}$

$$\omega(\alpha_n^{t_1} f_1 \cdots \alpha_n^{t_k} f_k)$$

converge for $n \to \infty$.
Then, there exists a unique regular Borel probability measure $d\rho_\omega$ on $\Pi_{t \in \mathbb{R}^+} \dot{X}_t$, $\dot{X}$ the one point compactification of X, such that

$$\lim_n \omega(\alpha_n^{t_1} f_1 \cdots \alpha_n^{t_k} f_k) = \int d\rho_\omega(x(t)) \, f_1(x(t_1)) \cdots f_k(x(t_k)).$$

For $\omega(x)$ of compact support K, the measure $d\rho_\omega$ is supported on a product of compact sets K_t, $t>0$.
The process defined by $d\rho_\omega$ is Markov, with time-independent transition function $P_t(x,B) = \alpha^t \chi_B(x)$, if

$$\lim_n \omega(\alpha_n^{t_1} f_1 \cdots \alpha_n^{t_k} f_k) = \lim_{n_k} \cdots \lim_{n_2} \lim_{n_1} \omega(\alpha_{n_1}^{t_1}(f_1 \alpha_{n_2}^{t_2-t_1}(\cdots(f_{k-1}\alpha_{n_k}^{t_k-t_{k-1}} f_k)\cdots))).$$

This condition is also necessary for the time-independent Markov property if $\omega(x)$ is almost everywhere different from zero.

III. PROOFS

Identifying states ω in $\mathcal{F}$ with representative function h in $L^1(X,d\mu)$, we will write $\alpha_n^{t*}h$ for the representative of $\alpha_n^{t*}\omega$. The norm of the dual $C^0(X)^*$ of $C^0(X)$ coincides on $\mathcal{F}$ with the $L^1(X,d\mu)$ norm. Since $d\mu$ is sigma finite, the continuous functions with compact support are dense in $L^1(X,d\mu)$; $L^1 \cap C^0$ is therefore dense in $L^1 = \mathcal{F}$.

By weak and strong $\tau_{\mathcal{F}}$ topology we will denote, respectively, the weak topology defined on $C^0(X)$ by the linear functionals in $\mathcal{F}$ and the ultrastrong topology defined as above by the representation of $C^0(X)$ in $L^2(X,d\mu)$, identical to the strong topology for bounded sets. By $\tau_{\mathcal{F}}$ topology and $\tau_{\mathcal{F}}$ continuity we will refer to the weak one. The following facts are preliminary:

Lemma 6: (i) $\forall t \in \mathbb{R}$, α_n^{t*} maps $\mathcal{F}$ into $\mathcal{F}$ and

$$\alpha_n^{t*}h(x) = h(\Gamma_n^{-t}x), \tag{3.1}$$

(ii) α_n^t is continuous in the weak (and in the strong) $\tau_{\mathcal{F}}$ topology, and extends therefore by continuity to an automorphism of the Von Neumann algebra $L^\infty(X,d\mu)$;

(iii) $\forall f \in L^1 \cap L^\infty$ $\alpha_n^{t*} f = \alpha_n^{-t} f$;

(iv) $\forall p \in [1,\infty]$, $\forall f \in L^1 \cap L^p \|\alpha_n^{t*}f\|_p = \|f\|_p$.

Proof. Equation (3.1) follows immediately from the invariance of $d\mu$: $\forall h \in L^1(X,d\mu)$, $h(\Gamma_n^t x) \in L^1(X,d\mu)$ and

$$\int h(\Gamma_n^{-t}x) f(x)\, d\mu = \int h(x) f(\Gamma_n^t x)\, d\mu = \int h(x) \alpha_n^t f\, d\mu \quad \forall f \in C^0(X). \tag{3.2}$$

Invariance of $d\mu$ implies conservation of all L^p norms. Invariance of $\mathcal{F}$ implies (see, e.g., Ref. 22) weak $\tau_{\mathcal{F}}$ continuity of α_n^t, which extends by continuity to the weak closure L^∞ of $C^0(X)$.

Proof of Proposition 1: (i) For all f in $C^0(X)$, the weak $\tau_{\mathcal{F}}$ limit of $\alpha_n^t f$ defines a continuous linear functional on $L^1(X,d\mu)$, with norm bounded by $\sup|f|$, i.e., an element of $L^\infty(X,d\mu)$; the limit α^t of α_n^t therefore defines a linear map of $C^0(X)$ into $L^\infty(X,d\mu)$, of norm at most one; its transpose α^{t*} *a priori* maps $\mathcal{F}$ into $C^0(X)^*$, and is continuous in the norm of $C^0(X)^*$, denoted by $\|\cdot\|_*$: $\forall h \in L^1$,

$$\|\alpha^{t*}h\|_* = \sup_{f \in C^0(X)\, \|f\|_\infty = 1} \left|\lim_n \int h\alpha_n^t f\, d\mu\right| \leq \|h\|_1.$$

Weak $\tau_{\mathcal{F}}$ continuity of α^t is equivalent[22] to stability of $\mathcal{F}$ under α^{t*}, i.e., point *(ii);* for positivity see *(iii)* below.

(ii) Consider first states represented by functions $h \in L^1 \cap C^0$; by Lemma 1, $\alpha_n^{t*}h = \alpha_n^{-t}h$, which $\tau_{\mathcal{F}}$ converges for $n \to \infty$ to $\alpha^{-t}h$. This means that $\forall f \in C_0^0(X)$,

$$\int \alpha^{t*}h\, f\, d\mu = \lim_n \int \alpha_n^{t*}h\, f\, d\mu = \lim_n \int \alpha_n^{-t}h\, f\, d\mu = \int \alpha^{-t}h\, f\, d\mu, \tag{3.3}$$

so that

$$\alpha^{t*}h = \alpha^{-t}h \quad \forall h \in L^1 \cap C^0. \tag{3.4}$$

Moreover,

$$\|\alpha^{-t}h\|_1 = \|\alpha^{t*}h\|_1 = \|\alpha^{t*}h\|_* . \tag{3.5}$$

Equation (3.5), continuity in $C^0(X)^*$ norm of α^{t*} and density of $L^1 \cap C^0$ in $\mathcal{F}$ imply the stability of $\mathcal{F}$ under α^{t*}.

(*iii*) α^t extends by $\tau_{\mathcal{F}}$ continuity to the Von Neumann algebra L^∞. Positivity of α^t is equivalent to

$$\int h(x)\alpha^t f(x) d\mu \geqslant 0, \quad \forall f \geqslant 0, \quad f \in L^\infty, \quad h \geqslant 0, \quad h \in L^1. \tag{3.6}$$

For $f \in C^0(X)$, Eq. (3.6) follows from the positivity of the same expression with α_n^t replacing α^t, for $f \in L^\infty$, by definition of continuous extension of α^t.

(*iv*) Equation (3.5) extends to all $h \in L^1 \cap L^\infty$ since, $\forall h \in L^1 \cap L^\infty$, $f \in L^1 \cap C^0$,

$$\int \alpha^{t*}h \, f \, d\mu = \int h \alpha^t f \, d\mu = \int h \alpha^{-t*} f \, d\mu = \int \alpha^{-t} h \, f \, d\mu \tag{3.7}$$

and $L^1 \cap C^0$ is dense in $C^0(X)$;
(*v*) follows from Hölder inequality:

$$\int \alpha_n^{t*} h f \, d\mu \leqslant \|h\|_p \|f\|_q \quad \forall h \in L^1 \cap L^p, \quad f \in C^0 \cap L^q, \quad n \in \mathbb{N},$$

which implies

$$\int \alpha^{t*} h \, f \, d\mu \leqslant \|h\|_p \|f\|_q$$

so that, by density of $C^0 \cap L^q$ in L^q,

$$\|\alpha^{t*}h\|_p \leqslant \|h\|_p . \tag{3.8}$$

(*vi*) In order to prove

$$\|\alpha^{t*}h\|_1 = \|h\|_1 \quad \forall h \in L^{1+}, \tag{3.9}$$

consider first $h \in C_0^0(X)$. Denote by χ_K the characteristic function of a compact set $K \subset X$; using Lemma 1, condition F implies

$$\int \chi_K \, \alpha_n^{-t} h \, d\mu = \int \chi_K \, h \, d\mu \quad \forall n \tag{3.10}$$

if K is sufficiently large, depending only h and t. Given the support L of h, it is in fact enough to take $K \supset L$ such that $\Gamma_n^t(L) \subset K \forall n$. Using Eqs. (3.4) and (3.10) it follows

$$\int \alpha^{t*} h \, d\mu = \lim_K \int \chi_K \, \alpha^{-t} h \, d\mu = \lim_K \lim_n \int \chi_K \, \alpha_n^{-t} h \, d\mu = \lim_K \int \chi_K \, h \, d\mu = \int h \, d\mu.$$

Equation (3.9) then follows from density in L^1 of $C_0^0(X)$ and L^1 continuity of α^{t*}. Equation (3.9) also implies $\alpha^t 1 = 1$.

(vii) Using Lemma 1 and Eq. (3.4), $\forall h \in L^1 \cap L^\infty$, $f \in L^1 \cap C^0$,

$$\int \alpha_n^t h\ f\,d\mu = \int h\ \alpha_n^{t*} f\,d\mu = \int h\ \alpha_n^{-t} f\,d\mu \to \int h\ \alpha^{-t} f\,d\mu = \int h\ \alpha^{t*} f\,d\mu = \int \alpha^t h\ f\,d\mu \tag{3.11}$$

and $\tau_{\mathcal{F}}$ convergence of $\alpha_n^t h$ to $\alpha^t h$ follows from density of $L^1 \cap C^0$ in L^1 and uniform boundedness of $\alpha_n^t h$ in L^∞.

To obtain $\tau_{\mathcal{F}}$ convergence of $\alpha_n^t h$ to $\alpha^t h$ for all $h \in L^\infty$ consider a sequence $\{K_i\}$ of compact sets with χ_{K_i} converging $\tau_{\mathcal{F}}$ to 1 (it exists because μ is σ finite); then, $\forall h \in L^{\infty +}$, $\forall l \in L^{1+}$,

$$\int \alpha_n^t (h - h\chi_{K_i})\ l\ d\mu \leqslant \|h\|_\infty \int (1 - \alpha_n^t \chi_{K_i})\ l\ d\mu \tag{3.12}$$

but $\tau_{\mathcal{F}}$ convergence of α_n^t on $L^1 \cap L^\infty$, $\tau_{\mathcal{F}}$ continuity of α^t and $\alpha^t 1 = 1$ imply

$$\lim_i \lim_n \text{rhs of } (3.12) = 0;$$

on the other hand, by $\tau_{\mathcal{F}}$ continuity of α^t,

$$\lim_i \lim_n \int \alpha_n^t (h\chi_{K_i})\ l\ d\mu = \int \alpha^t h\ l\ d\mu$$

and therefore,

$$\lim_n \int \alpha_n^t h\ l\ d\mu = \int \alpha^t h\ l\ d\mu,$$

which extends by linearity to $h \in L^\infty$ and $l \in L^1$.

Proof of Proposition 2: $(vi) \Rightarrow (v)$ and $(ii) \Rightarrow (i)$ are obvious; $(iv) \Leftrightarrow (v)$ and $(i) \Rightarrow (vi)$ immediately follow from Proposition 1, (for the latter implication, see the proof of Lemma 6); it is therefore enough to prove $(v) \Rightarrow (i)$ and $(vi) \Rightarrow (iii) \Rightarrow (ii)$.

$(v) \Rightarrow (i)$: For any Borel set B of finite measure $d\mu$, $\alpha^{t*}\chi_B$ is real, non-negative, and

$$\|\alpha^{t*}\chi_B\|_\infty \leqslant \|\chi_B\|_\infty \leqslant 1,$$

so that

$$(\alpha^{t*}\chi_B)^p \leqslant \alpha^{t*}\chi_B; \tag{3.13}$$

(v) implies that for some $p \in (1,\infty)$,

$$\int (\alpha^{t*}\chi_B)^p d\mu = \int \chi_B^p\,d\mu = \int \chi_B\,d\mu \geqslant \int \alpha^{t*}\chi_B\,d\mu, \tag{3.14}$$

the last equality following from Proposition 1, *(v)*. Equations (3.13) and (3.14) imply

$$\alpha^{t*}\chi_B = (\alpha^{t*}\chi_B)^p$$

$d\mu$ almost every, so that $\alpha^{t^*}\chi_B$ is a measurable function taking values in {0, 1}, i.e., it is the characteristic function of a measurable set $B' \subset X$, and by Eqs. (3.13) and (3.14), $\mu(B') = \mu(B)$.

From linearity and the fact that L^∞ norms do not increase [Proposition 1, *(i)*], it follows that disjoint sets *B*, *C* have disjoint images B', C'; α^t, coinciding with α^{-t*} on $L^1 \cap L^\infty$, [Proposition 1, (iv)] is therefore a morphism of the algebra of simple functions based on sets of finite measure, and, by a density argument, of $L^\infty(X, d\mu)$. It also follows that α^{t*} preserves the L^1 norm on L^{1+}, which implies invertibility of α^t, actually $\alpha^{-t}\alpha^t = \alpha^t\alpha^{-t} = 1$; in fact, for all positive simple functions *f*, *g*,

$$\int f\ \alpha^{-t}\alpha^t g\, d\mu = \int \alpha^t f\ \alpha^t g\, d\mu = \int \alpha^t(fg)\ d\mu = \int f\ g\, d\mu.$$

The same holds for $\alpha^t\alpha^{-t}$, and invertibility of α^t follows by a density argument.

$(vi)\Rightarrow(iii)$: Convergence in the strong and ultrastrong topologies is equivalent for a bounded sequence,[28] and amounts to

$$\int |\alpha_n^t f - \alpha^t f|^2 h\, d\mu \to 0 \quad \forall h \in L^1(X, d\mu). \tag{3.15}$$

By a density argument, it is enough to obtain Eq. (3.15) for all f in $C_0^0(X) \subset L^1 \cap L^\infty$; in this case, $\alpha_n^t f = \alpha_n^{-t*} f$ converges to $\alpha^t f = \alpha^{-t*} f$ strongly in $L^2(X, d\mu)$, as a consequence of weak L^2 convergence and of conservation of the L^2 norms in the limit. $\alpha^t 1 = 1$ immediately follows from conservation of the L^1 norm.

$(iii)\Rightarrow(ii)$: Strong convergence implies in general that α^t is a morphism. The group property follows since for all f, g in $C_0^0(X)$,

$$\int f\ \alpha^{t+s} g\, d\mu = \lim_n \int f\ \alpha_n^t \alpha_n^s g\, d\mu = \lim_n \int \alpha_n^{-t} f\ \alpha_n^s g\, d\mu; \tag{3.16}$$

strong convergence in L^2 of $\alpha_n^{-t} f$ and $\alpha_n^s g$ imply in fact

$$\lim_n \int \alpha_n^{-t} f\ \alpha_n^s g\, d\mu = \int \alpha^{-t} f\ \alpha^s g\, d\mu = \int f\ \alpha^t \alpha^s g\, d\mu. \tag{3.17}$$

Proof of Proposition 3: The algebra $C_0^0(X)$ of continuous functions of compact support is dense in $C^0(X)$ and in $L^1(X, d\mu)$ in the respective norms. Moreover, the sequence $\alpha_n^t h$ is uniformly bounded in L^1 and in C^0. $\tau_{\mathcal{F}}$ convergence of α_n^t for all $f \in C^0$ is therefore equivalent to convergence on $C_0^0(X)$.

In the same way, since $C_0^0(X)$ is dense in L^2 and U_n^t preserves the L^2 norm, $\tau_{\mathcal{F}}$ convergence is equivalent to

$$\forall h_1,\ h_2 \in L^2 \quad \exists \lim_n \int h_1^* U_n^t h_2\, d\mu,$$

i.e., to convergence of U_n^t in the weak topology w_{op} for operators in L^2.
Denoting $U^t \equiv w_{\mathrm{op}} - \lim_n U_n^t$, $\forall h_1,\ h_2 \in C_0^0(X)$,

$$\int h_1^* U^t h_2\, d\mu = \lim_{n\to\infty} \int h_1^* \alpha_n^t h_2\, d\mu = \int h_1^* \alpha^t h_2\, d\mu, \tag{3.18}$$

which implies $U^t h = \alpha^t h$ for $h \in C_0^0(X)$ and, by L^2 continuity of both U^t and α^t (Proposition 1, v), for $h \in L^1 \cap L^\infty$. The rest of the proof of Proposition 3 is then reduced to the following simple fact: $\forall n \in \mathbb{N}$ let U_n^t be a one-parameter group of unitary operators on a Hilbert space $\mathcal{H}$, converging weakly for $n \to \infty$, $\forall t \in \mathbb{R}$, $w_{\mathrm{op}} - \lim_n U_n^t \equiv U^t$. Then the following are equivalent:
(a) $\forall t \in \mathbb{R}$, U_n^t converges strongly to U^t,
(b) $\forall t \in \mathbb{R}\ \forall h \in \mathcal{H} \|U^t h\| = \|h\|$,
(c) $\forall t \in \mathbb{R} U^t$ is unitary,
(d) U^t is a group of unitary operators.

In fact, $(a)\Leftrightarrow(b)$ is well known, and $(d)\Rightarrow(c)\Rightarrow(b)$ are trivial.
To prove $(a)\Rightarrow(d)$, we have, from Schwarz inequality,

$$\begin{aligned}|(h,(U_n^tU_n^s-U^tU^s)h)|&=|(h,(U_n^tU_n^s-U_n^tU^s+U_n^tU^s-U^tU^s)h)|\\&\leqslant|(h,U_n^t(U_n^s-U^s)h)|+|(h,(U_n^t-U^t)U^sh)|\\&\leqslant\|h\|\|(U_n^s-U^s)h\|+|(h,(U_n^t-U^t)U^sh)|,\end{aligned}$$

which implies the group property. In particular U^t is invertible, and unitarity follows since $(a)\Rightarrow(b)$.

Proof of Proposition 4: For fixed compact $K\subset X$, $\epsilon>0$, $t\in\mathbb{R}$, $f\in C_0^0(X)$ with support in K, by property F all the functions $\alpha_n^tf(x)=f(\Gamma_n^t(x))$ have support contained in a compact set K'. By assumption, there exists an open set S, of measure smaller than ϵ, such that

$$d(\Gamma_n^t(x),\Gamma_n^t(y))\;<\;M(K',\varepsilon,t)d(x,y)$$

for $x,y\in K'\backslash S$. Since f is of compact support and therefore uniformly continuous, it follows that the functions $\alpha_n^tf(x)$ are equicontinuous on $K'\backslash S$.

Since $K'\backslash S$ is closed and therefore compact, by the Ascoli–Arzelà theorem there exists a subsequence $\alpha_{n_k}^tf$ which converges there uniformly; its limit coincides, on $K'\backslash S$ with α^tf, as a consequence of weak convergence of $\alpha_n^tf(x)$ to α^tf. On compact sets, uniform convergence implies convergence in L^2; moreover, the functions $\alpha_n^tf(x)$ are bounded by $\sup|f|$, and therefore their L^2 norm coincides with the norm of their restriction to $K'\backslash S$ up to an error of order ϵ, and the same is true for $\alpha^tf(x)$; this implies

$$|\|\alpha^tf\|_{L^2}^2-\|f\|_{L^2}^2|\;\leqslant\;2\,\sup|f|^2\varepsilon\quad\forall\varepsilon>0.$$

It follows that α^t preserves the L^2 norm on $C_0^0(X)$, for all t, and therefore, by Proposition 2, (iv), it is a group of automorphisms of $L^\infty(X,d\mu)$.

Proof of Proposition 5: We will use the shorthand $\{t_j\}_{1,k}$ for $0\leqslant t_1<t_2<\cdots<t_k$, and $\{t_j\}_{1,]m[,k}$ for the same sequence without t_m; $X^{\{t_j\}_{1,k}}$ will denote the product of k copies of X, X_{t_j} the copy at time t_j; the state ω will be represented by the measure $\omega(x)d\mu$, $\omega(x)\in L^1(X,d\mu)$.

For fixed $\{t_j\}_{1,k}$, consider the algebra C^{sp} of functions $f\in C^0(X^{\{t_j\}_{1,k}})$ which are finite sums of products, i.e., $f(x_{1,k})=\Sigma_ic_if_1^i(x_1)\cdots f_k^i(x_k)$ with $f_j^i(x_j)\in C^0(X_{t_j})$. For fixed n the linear map of C^{sp} into itself

$$f(x_{1,k})\rightarrow\sum_i\;c_i\prod_{j=1}^k\;\alpha_n^{t_j}f_j^i(x_j)$$

is $\|\;\|_\infty$ preserving and extends to an automorphism of the completion $\overline{C^{sp}}$ of C^{sp} in the sup norm; in particular, it is positive.

It follows that $\rho_\omega^{\{t_j\}_{1,k}}$, defined by

$$\rho_\omega^{\{t_j\}_{1,k}}(f)\equiv\lim_n\int_X\sum_i\;c_i\prod_{j=1}^k\;\alpha_n^{t_j}\;f_j^i(x)\omega(x)\;d\mu(x)\tag{3.19}$$

is a bounded positive linear functional over C^{sp}, with $\|\rho_\omega\|\leqslant\|\omega\|$. Since by the Stone–Weierstrass theorem $\overline{C^{sp}}=C^0(X^{\{t_i\}_{1,k}})$, ρ_ω has a unique extension to C^0, with the same bound.

The functional so obtained defines, by the Riesz–Markov theorem, a finite regular Borel measure, which we indicate with $\rho_\omega^{\{t_j\}_{1,k}}$, over $X^{\{t_j\}_{1,k}}$. If X is not compact it is technically convenient to consider $\rho_\omega^{\{t_j\}_{1,k}}$ over $\dot{X}^{\{t_j\}_{1,k}}$, the product of the 1-point compactified spaces $\dot{X}_{t_j}$. However, if ω has compact support K, the construction also lives in $\Pi_{t\geqslant0}K_t$ with K_t the compact sets given by condition F, $\alpha_n^t\chi_K\leqslant\chi_{K_i}\forall n$. The measures $\rho_\omega^{\{t_j\}_{1,k}}$, for all k-tuples $0\leqslant t_1<t_2<\cdots<t_k$, define a (unique) measure over $\Pi_{t\geqslant0}\dot{X}_t$ iff they are compatible, i.e.,

$$\rho_\omega^{\{t_j\}_{1,k}}\,(f_1(x_1)\cdots f_{m-1}(x_{m-1})f_{m+1}(x_{m+1})\cdots f_k(x_k))$$
$$=\rho_\omega^{\{t_j\}_{1,]m[,k}}\,(f_1(x_1)\cdots f_{m-1}(x_{m-1})f_{m+1}(x_{m+1})\cdots f_k(x_k)).$$

Compatibility obviously holds for X compact, since then $1\in C^0(X)$, and $\alpha_n^t 1=1$. For noncompact X, compatibility follows from $\tau_{\mathcal{F}}$ continuity of α^t together with $\alpha^t 1=1$; take in fact an increasing sequence of positive functions $g_\nu\in C_0^0(X_{t_m})$ converging to 1 $d\mu$ almost everywhere. Clearly, for any positive $f_j\in C^0(X_{t_j}),\ j=1,...,k,\ j\neq m,$

$$\lim_\nu \lim_n \omega\left((1-\alpha_n^{t_m}g_\nu)\prod_j \alpha_n^{t_j}f_j\right)\leqslant \lim_\nu \lim_n \omega(1-\alpha_n^{t_m}g_\nu)\prod_j \|f_j\|_\infty . \tag{3.20}$$

By the monotone convergence theorem, $g_\nu\Pi_j f_j$ converge in $\|\ \|_{L^1(X^{\{t_j\}_{1,k}},\rho_\omega^{\{t_j\}_{1,k}})}$ to $\Pi_j f_j$, and therefore the left-hand side of Eq. (3.20) converges to

$$\rho_\omega^{\{t_j\}_{1,]m[,k}}\left(\prod_j f_j\right)-\rho_\omega^{\{t_j\}_{1,k}}\left(\prod_j f_j\right);$$

since g_ν converge in $\|\ \|_{L^1(X,\alpha^{t_m}*\omega)}$ to 1, the right-hand side converges to 0.
This implies the compatibility condition, since linear combinations of products of positive functions are dense in $C^0(X^{\{t_j\}_{1,]m[,k}})$.
From compatibility it also follows $\rho_\omega^{\{t_j\}_{1,k}}(1)=\omega(1)=1$.
For the Markov property, let us indicate by

$$p(t_1,x_1;\ldots;t_{k-1},x_{k-1}|t_k,f_k)\in L^\infty(\dot{X}^{\{t_j\}_{1,k-1}},\rho_\omega^{\{t_j\}_{1,k-1}})$$

the conditional expectations of the process, characterized by

$$\forall k\geqslant 2,\quad \forall j=1,...,k,\quad \forall f_j\in L^\infty(\dot{X}_{t_j},\rho_\omega^{t_j})$$

$$\int f_1(x_1)\cdots f_k(x_k)d\rho_\omega^{\{t_j\}_{1,k}}=\int f_1(x_1)\ \cdots\ f_{k-1}(x_{k-1})\ p(t_1,x_1;\ldots;t_{k-1},x_{k-1}|t_k,f_k)\ d\rho_\omega^{\{t_j\}_{1,k-1}}.$$

The process is Markov iff

$$p(t_1,x_1;\ldots;t_{k-1},x_{k-1}|t_k,f_k)=p(t_{k-1},x_{k-1}|t_k,f_k)$$

and it is also time independent iff

$$p(t_1,x_1|t_2,f_2)=p(0,x_1|(t_2-t_1)f_2).$$

Hence, for a time-independent Markov process, by iteration we obtain

$$\int f_1(x_1)\ \cdots\ f_k(x_k)\ d\rho_\omega^{\{t_j\}_{1,k}}$$
$$=\int p(0,x_0|t_1,f_1p(0,\cdot|t_2-t_1,f_2\ \cdots\ f_{k-1}p(0,\cdot|t_k-t_{k-1},f_k)\cdots))\omega(x_0)d\mu. \tag{3.21}$$

Notice now that, since the α_n^t are automorphisms,

$$\lim_{n_k} \cdots \lim_{n_2} \lim_{n_1} \omega\ (\alpha_{n_1}^{t_1} f_1 \cdots \alpha_{n_k}^{t_k} f_k) = \lim_{n_k} \cdots \lim_{n_2} \lim_{n_1} \omega\ (\alpha_{n_1}^{t_1}\ (f_1 \alpha_{n_2}^{t_2 - t_1}(\cdots(f_{k-1} \alpha_{n_k}^{t_k - t_{k-1}} f_k)\cdots))) \tag{3.22}$$

which, by definition and continuity of α^t, converges to

$$\omega(\alpha^{t_1}(f_1 \alpha^{t_2 - t_1}\ (\cdots\ (f_{k-1} \alpha^{t_k - t_{k-1}} f_k)\ \cdots))). \tag{3.23}$$

So, if the limit (3.19) coincides with (3.22), then the stochastic process ρ is Markov and time independent, with $p(0,x|t,f) = \alpha^t f(x)$.

The converse follows if $\omega(x)$ only vanishes on a set of zero measure; in fact, for $\omega(x) \neq 0$, $p(0,x|t,f) = \alpha^t f(x)$, and therefore, if the process is time-independent Markov, i.e., if (3.21) holds, then it is also given by the limit (3.22) and (3.23).

IV. MODELS

We consider two models for the general structures and results of Sec. II. For both models, we will prove weak convergence to positive maps, and for the second model the limit will be shown to define a time-independent Markov process.

The first model is a Hamiltonian system with one degree of freedom, namely a particle in a periodic potential, which is scaled so that its period converges to zero. The phase space is $\mathbb{R}^2$, with canonical variables p, q; the dynamics is defined by the Hamiltonian

$$H_n \equiv p^2/\ 2m + V(nq), \tag{4.1}$$

with $V(q) = V(q+1) \in C^2(\mathbb{R}^2)$, and $dV/dq = 0$ only for a finite number of points.

We will consider the maps defined on $\mathbb{R}^2$ by the solutions $p_t^{(n)}(p,q)$, $q_t^{(n)}(p,q)$ of the Hamiltonian equations of motion, with Hamiltonians (4.1) and initial conditions p, q, and denote by α_n^t the corresponding map on $C^0(\mathbb{R}^2)$, the space of continuous functions vanishing at infinity,

$$\alpha_n^t f(p,q) \equiv f(p_t^{(n)}(p,q), q_t^{(n)}(p,q)). \tag{4.2}$$

For $n = 1$, we denote

$$p_t(p,q) \equiv p_t^{(1)}(p,q), \quad q_t(p,q) \equiv q_t^{(1)}(p,q).$$

From the properties of $V(q)$ it follows that $p_t(p,q)$ and $q_t(p,q)$ mod $\mathbb{Z}$ are periodic in q, with period 1, and in t, with period $T(p,q)$ differentiable and periodic in q, for fixed p, except for q mod $\mathbb{Z}$ in a finite set $\{q_{i,\infty}(p)\}$.

We will assume nondegeneracy, i.e., for all p, $\partial/\partial q T(p,q) \neq 0$, except for a finite number of points $q_{i,0}(p)$ (in [0, 1)). It follows that, $\forall p \in \mathbb{R}$, $\forall \delta > 0$,

$$T(p,q)\ \leqslant\ 1/\delta, \quad |\partial/\partial q T(p,q)|\ \geqslant\ \delta \quad \forall q \in [0,1] \backslash I_\delta(p), \tag{4.3}$$

with $I_\delta(p)$ the union of a finite number of intervals, with total length converging to zero for $\delta \to 0$, and $\cap_\delta I_\delta(p) = \{q_{i,0}(p), q_{i,\infty}(p)\}$.

Proposition 7: Under the above assumptions, the automorphisms α_n^t*, Eq. (4.2), converge for all* $t \in \mathbb{R}$*, in the weak topology defined by Lebesgue absolutely continuous finite measures in* $\mathbb{R}^2$*, to positive bistochastic maps* α^t*, given by*

$$\alpha^t f(p,q) = \int_0^1 dx\ 1/T(p,x) \int_0^{T(p,x)} d\tau\ f(p_\tau(p,x), Q_t(p,x) + q) \tag{4.4}$$

with

$$Q_t(p,x)\ = \lim_{n\to\infty} 1/n\ q_{nt}(p,x)\ \equiv\ \beta(p,x)t. \tag{4.5}$$

The limit (4.5) exists pointwise for all p, x, and it is uniform, at fixed p, for $x \in [0,1] \backslash I_\delta(p)$, $\forall \delta > 0$.

Proof: It is easy to see that

$$p_t^{(n)}(p,q) = p_{nt}(p,nq), \quad q_t^{(n)}(p,q) = 1/n \; q_{nt}(p,nq). \tag{4.6}$$

It is enough to prove weak convergence of

$$\alpha_n^t f(p,q) = f(p_{nt}(p,nq), 1/n \, q_{nt}(p,nq)) \tag{4.7}$$

for f in $\mathcal{D}$, the space of C^∞ functions of compact support.
We introduce the convolutions

$$f_{k,n}^t(p,q) \equiv \int ds \, h_k(q-s) \; \alpha_n^t f(p,s), \tag{4.8}$$

with

$$0 \leqslant h_k \in \mathcal{D}, \quad \mathrm{Supp}(h_k) \subset [-1/k, 1/k], \quad \int h_k(s) ds = 1.$$

We will show that

$$\lim_k \lim_n f_{k,n}^t(p,q) \tag{4.9}$$

exists pointwise almost everywhere, and it is given by the right-hand side of Eq. (4.4).
Existence of the limit $n \to \infty$ in Eq. (4.9) implies convergence, for $n \to \infty$, of

$$\int dp \, ds \left(\int d \; q g(p,q) \; h_k(q-s) \right) \alpha_n^t f(p,s) \tag{4.10}$$

$\forall g \in \mathcal{D}$, i.e., weak convergence of $\alpha_n^t f$ with respect to a dense subspace of $L^1(\mathbb{R}^2)$ and therefore, by uniform boundedness of $\alpha_n^t f$, weak convergence. The weak limit is identified by the value taken on it by the measures with density in $\mathcal{D}$, and therefore by the limit, $k \to \infty$, of Eq. (4.10), which is given by the pointwise limit (4.9), $f_{k,n}^t$ being uniformly bounded.

In order to control the limit (4.9), for fixed $p \in \mathbb{R}$ and $h \in \mathcal{D}$, we use the periodicity properties of p_t, q_t,

$$p_t(p,q+n) = p_t(p,q), \quad q_t(p,q+n) = q_t(p,q) + n \quad \forall n \in \mathbb{Z} \tag{4.11}$$

to estimate, denoting $x(s) \equiv ns - [ns]$, $[\;]$ the integer part,

$$\begin{aligned} &\int ds \, h_k(q-s) f(p_{nt}(p,ns), 1/n \, q_{nt}(p,ns)) \\ &\quad = \int ds \, h_k(q-s) \; f(p_{nt}(p,x(s)), 1/n \; (q_{nt}(p,x(s)) + [ns])) \\ &\quad \simeq \int_0^1 dx \, f(p_{nt}(p,x), 1/n \, q_{nt}(p,x) + q), \end{aligned} \tag{4.12}$$

with an error bounded by $\epsilon_k + C(k)\epsilon_n$, ϵ_i denoting, here and in the following, a suitable infinitesimal sequence.

Periodicity in t of $q_t(p,q)$ mod $\mathbb{Z}$, more precisely, for $x \notin \{q_{i,\infty}(p)\}$,

$$q_{t+T(p,x)}(p,x) = q_t(p,x) + \sigma(p,x) \tag{4.13}$$

with

$$\sigma(p,x)=\pm 1 \quad \text{if } H(p,x) > \max V(q)$$

and otherwise vanishing, implies pointwise convergence of $1/n\, q_{nt}(p,x)$, uniform in $[0,1]\backslash I_\delta(p)$, $\forall \delta>0$, to

$$Q_t(p,x) \equiv \sigma(p,x)\ t/T(p,x). \tag{4.14}$$

Since $T(p,x)\to\infty$ for $x\to q_{i,\infty}(p)$, $Q_t(p,x)$ is continuous in x, at fixed p.
It follows that the right-hand side of Eq. (4.12) can be estimated as

$$\int_0^1 dx\, f(p_{nt}(p,x),\ Q_t(p,x)+q) \tag{4.15}$$

apart from an error ϵ_n.

By extending $Q_t(p,x)$ to a (continuous) periodic function of $x\in\mathbb{R}$, of period 1, the right-hand side of Eq. (4.15) can be written, with h_j as before,

$$\int_0^1 dx \int dy\, h_j(x-y)\ f(p_{nt}(p,y),Q_t(p,y)+q) \simeq \int_0^1 dx \int dy\, h_j(x-y)\ f(p_{nt}(p,y),Q_t(p,x)+q) \tag{4.16}$$

with an error ϵ_j.

By periodicity in time, $p_{nt}(p,y)=p_{\tau_n(p,y)}(p,y)$, with

$$\tau_n(p,y) \equiv nt \bmod T(p,y). \tag{4.17}$$

By continuity in x, uniform for t bounded, of $p_t(p,x)$, $p_{\tau_n(p,y)}(p,y)$ can be replaced, with an error $C(\delta)\epsilon_j$, by $p_{\tau_n(p,y)}(p,x)$, for $y\in[0,1]\backslash I_\delta(p)$, $\forall\delta>0$. Since $I_\delta(p)$ is the union of a finite number of intervals, restricting x to $[0,1]\backslash I_{\delta/2}(p)$ implies that, for large j, the domain of integration in y is contained in $[0,1]\backslash I_\delta(p)$; replacing in the right-hand side of (4.16) $p_{nt}(p,y)$ with $p_{\tau_n(p,y)}(p,x)$ therefore gives an error $\epsilon(\delta)+C(\delta)\epsilon_j$. Moreover, $\tau_n(p,y)$ is differentiable in y, apart from a finite number of points, with derivative

$$-[nt/T(p,y)]\ \partial/\partial y T(p,y)$$

large in modulus, uniformly in $y\in[0,1]\backslash I_\delta(p)$, $\forall\delta>0$. It follows that integration in y (in $[x-j^{-1},x+j^{-1})$ in the right-hand side of Eq. (4.16) can be replaced by a mean in τ, with an error $\epsilon_j+C(j,\delta)\epsilon_n+\epsilon(\delta)$; the result is then the right-hand side of Eq. (4.4).

From the fact that the sum of the errors made in the above estimates can be written as $\epsilon_k+C(\delta)\epsilon_j+(C(k)+C(j,\delta)\epsilon_n+\epsilon(\delta))$, it follows that the limit (4.9) exists and is given by the right-hand side of Eq. (4.4). ■

The second model is defined by a particle moving horizontally in $\mathbb{R}^2$ with speed 1, and interacting with fixed vertical ''rods,'' of equal height; the interaction is assumed only to change of the ''spin'' of the particle, taking values ± 1.

The space X of the configurations of the system is thus $\mathbb{R}^2\times\{-1,1\}$, on which we take the measure $\mu=d^2\mathbf{x}\otimes d\sigma$, with $\sigma(\{1\})=\sigma(\{-1\})=1$. As space of states we will take $L^1(\mathbb{R}^2\times\{-1,1\},\mu)$, i.e., the set of states defined by measures space which are absolutely continuous with respect to the Lebesgue measure. The dynamics will depend on a parameter n through a suitable scaling of the positions and of the height of the rods *in one given and fixed configuration* in a set to be specified in the following.

In general, a configuration ξ of rods is determined by their height h and by the set of their centers, $\xi\equiv\cup_{i\in\mathbb{N}}\{(x_i,y_i)\}$; we will consider only configurations with a finite number of rods in any compact subset of $\mathbb{R}^2$, and denote by Ξ the set of all such configurations. For fixed $\xi \in \Xi$, and $h > 0$, we will study the limit $n \to \infty$ of the dynamics α_n^t defined by the interaction with rods of height $n^{-2}h$ in the scaled configuration $n^{-1}\xi \equiv \cup_{i \in \mathbb{N}} \{(n^{-1}x_i, n^{-1}y_i)\}$. The limit will be taken in the weak topology defined by the state space $\mathcal{F} \equiv L^1(\mathbb{R}^2 \times \{-1,1\}, \mu)$. We will prove the following:

Proposition 8: Let $n_k \to \infty$, with $\Sigma 1/n_k < \infty$. For any denumerable set of positive times, for almost all choices of the configuration ξ with respect to the independent Poisson distribution (the measure ρ defined below), the dynamical systems defined as above by the configuration ξ converge weakly, for $k \to \infty$, with respect to the set $\mathcal{F}$ of states, to positive maps, and define, as in Proposition 5, a time-independent Markov process.

It is of course essential that *no integration* is performed over ρ; performing such an integration would in fact amount to *assuming* a stochastic process, indexed by the sample space of the distribution of the obstacles. On the contrary, Proposition 7 *implies* that *the same* stochastic process correctly describes the *result of all measurements, in the limit $n \to \infty$*, for ρ almost all *fixed* configurations of obstacles. The model is therefore relevant in order to clarify the basis for the use of stochastic maps and processes for the description of perfectly reversible dynamical systems, and it is worthwhile to control it with all mathematical details.

For completeness, we recall the construction of the probability measure ρ corresponding to the independent Poisson distribution of points in a plane, with unit mean density: following Ref. 30, Sec. 7.1.2, one considers the family of measures on Λ^j,

$$d\rho_\Lambda^j \equiv \frac{e^{-|\Lambda|}}{j!} \, d\mathbf{x}_1 \cdots d\mathbf{x}_j \tag{4.18}$$

with $j \in \mathbb{N}$ and Λ bounded Borel $\subset \mathbb{R}^2$; they define a positive linear functional

$$\rho(f) \equiv \sum_{j=0}^{\infty} \frac{e^{-|\Lambda|}}{j!} \int_{\Lambda^j} d\mathbf{x}_1 \cdots d\mathbf{x}_j \, f(\xi), \tag{4.19}$$

where $f: \Xi \to \mathbb{C}$ depends only on the points, $\mathbf{x}_1 \ldots \mathbf{x}_j$, of ξ inside Λ, over a suitable algebra $\mathcal{B}$ of functions on Ξ. The spectrum $\mathcal{E}$ of the $\| \|_\infty$ closure of $\mathcal{B}$ turns out be the $* - w$ closure of Ξ (the $*$ weak topology being the weakest one which makes the weak topology being the weakest one which makes the functions in $\mathcal{B}$ continuous). By the Riesz–Markov theorem, the positive functional ρ defines a Borel regular measure on $\mathcal{E}$, denoted again by ρ, and it is not difficult to see that ρ is in fact concentrated on Ξ, i.e., $\rho(\Xi) = \rho(\mathcal{E}) = 1$.[29] For functions depending only on a finite number of points, integration with the measure ρ is simply given by Eq. (4.19).

Proof of Proposition 8: The time evolution is given by

$$\Gamma_n^t(x,y,\sigma) \equiv (x+t,y,\sigma \cdot (-1)^{|n^{-1}\xi \cap B_{\mathbf{x},t,n^{-2}h}|}) = (x+t,y,\sigma \cdot (-1)^{|(\xi \cap B_{n\mathbf{x},nt,n^{-1}h}|}, \tag{4.20}$$

$|I|$ being the number of points in the finite set I, $\mathbf{x} \equiv (x,y)$, and

$$B_{\mathbf{x},t,h} \equiv \{(\alpha,\beta) \in \mathbb{R}^2 | \beta \in [y-h/2, y+h/2),$$

$$\alpha \in [x, x+t) \quad \text{for } t>0, \quad \alpha \in [x+t, x) \quad \text{for } t<0\}.$$

For the observables we have therefore:

$$\forall f \in C^0(X), \quad \alpha_n^t \, f(x,y,\sigma) = f(x+t,y,\sigma \cdot (-1)^{|\xi \cap B_{n\mathbf{x},nt,n^{-1}h}|}). \tag{4.21}$$

Clearly, μ is left invariant by the Γ_n^t, for all n and t. The lack of continuity of the maps Γ_n^t (a consequence of the discreteness of the ''spin'' variable) is not a problem since, as already remarked, measurability of Γ_n^t is enough for all the results of Sec. II. Here, α_n^t maps C^0 into $L^1 \cap L^\infty$, and defines automorphisms of the closure of this space, with the Sup norm.

We denote by $g_{\pm 1,\mathbf{x},t,h}(\xi)$ the characteristic functions of the sets

$$\left\{ \mathbf{x}\in\mathbb{R}^2 : |\xi\cap B_{\mathbf{x},t,h}| \text{ is } \begin{matrix}\text{even}\\ \text{odd}\end{matrix}\right\}.$$

From Eq. (4.21), using the density of the finite linear combinations of the characteristic functions of bounded Borel subsets of $\mathbb{R}^2$ in $L^1(\mathbb{R}^2,d^2\mathbf{x})$, it follows that the existence of the weak limit for α_n^t,

$$\forall f\in C^0(X), \quad \forall l\in L^1(X,d\mu) \quad \lim_n \int l\ \alpha_n^t f\, d\mu \equiv \int l\ \alpha^t f\, d\mu$$

is equivalent to the existence of the limit

$$\lim_n \int \chi_B(\mathbf{x})\ g_{\sigma,n\mathbf{x},nt,n^{-1}h}(\xi)\ d^2\mathbf{x}. \tag{4.22}$$

$\forall B$ bounded Borel subset of $\mathbb{R}^2$, and $\sigma=\pm 1$.

Moreover, calling $G_{\sigma,\mathbf{x},t,h}(\xi)$ the weak limit of $g_{\sigma,n\mathbf{x},nt,n^{-1}h}(\xi)$, α^t will be given by

$$\alpha^t f(\mathbf{x},\sigma) = \sum_{\tau=\pm 1} G_{\tau,\mathbf{x},t,h}(\xi)\ f(\mathbf{x}+t,\tau\cdot\sigma),$$

with $\mathbf{x}+t\equiv(x+t,y)$.

In the same way it turns out that the existence of the limit for the measurement of k observables at k successive times $0\leqslant t_1<t_2<\cdots<t_k$ is equivalent to weak convergence of the products $\Pi_{i=1}^k g_{\sigma_i,n\mathbf{x},nt_i,n^{-1}h}(\xi)$. Denoting $t_0\equiv 0$ and $\tau_i\equiv\Pi_{j\leqslant i}\sigma_j$, it follows

$$\prod_{i=1}^k g_{\sigma_i,n(\mathbf{x}+t_{i-1}),n(t_i-t_{i-1}),n^{-1}h}(\xi) = \prod_{i=1}^k g_{\tau_i,n\mathbf{x},nt_i,n^{-1}h}(\xi),$$

and we may write the request of convergence in a form that will be more suitable to demonstrate the Markov time-independent property for the limit stochastic process:

$\forall B$ bounded Borel set $\subset\mathbb{R}^2$, $\quad\forall k=1,2,...,\quad \forall\sigma_i=\pm 1,\quad i=1,...,k$

$$\exists \lim_n \int \chi_B(\mathbf{x}) \prod_{i=1}^k g_{\sigma_i,n(\mathbf{x}+t_{i-1}),n(t_i-t_{i-1}),n^{-1}h}(\xi)\ d^2\mathbf{x}. \tag{4.23}$$

Now, from the definition of ρ it follows that the characteristic functions of the sets $\{\mathbf{x}\in\mathbb{R}^2 : |\xi\cap B_{\mathbf{x},t,h}|=n\}$, denoted by $\tilde{g}_{n,\mathbf{x},t,h}(\xi)$, are in $L^1(\Xi,\rho)$, and their integral is given by

$$\int_\Xi \tilde{g}_{n,\mathbf{x},t,h}(\xi)\ d\rho(\xi) = \sum_{j=0}^\infty \frac{e^{-|\Lambda|}}{j!}\int_{\Lambda^j} d\mathbf{x}_1\ldots d\mathbf{x}_j\ \tilde{g}_{n,\mathbf{x},t,h}(\xi) = e^{-|t|h}\frac{|th|^n}{n!}.$$

Expressing the functions $g_{\sigma,\mathbf{x},t,h}$ in terms of the $\tilde{g}_{n,\mathbf{x},t,h}$, we obtain

$$\int_\Xi g_{\pm 1,\mathbf{x},t,h}(\xi)\ d\rho(\xi) = \frac{1\pm e^{-2|t|h}}{2}$$

and, for $B_{\mathbf{x}_i,t_i,h_i}\cap B_{\mathbf{x}_j,t_j,h_j}=\emptyset,\ \ i\neq j$,

$$\int_\Xi \prod_{i=1}^k g_{\sigma_i,\mathbf{x}_i,t_i,h_i}(\xi)\ d\rho(\xi) = \prod_{i=1}^k \int_\Xi g_{\sigma_i,\mathbf{x}_i,t_i,h_i}(\xi)\ d\rho(\xi) = \prod_{i=1}^k \frac{1+\sigma_i\cdot e^{-2|t_i|h_i}}{2}.$$

The aim is now to show that the conditions in Eqs. (4.22) and (4.23) are fulfilled for ρ—almost all $\xi \in \Xi$. We will show that those limits exist in the $L^1(\Xi,\rho)$ norm, and hence pointwise, ρ-almost everywhere on Ξ, for any subsequence n_k with $\Sigma_k 1/n_k < \infty$. Consider first the single time case: $\forall B$ bounded Borel $\subset \mathbb{R}^2$:

$$\int_\Xi d\rho(\xi) \left| \int_B d^2\mathbf{x}\, g_{\sigma,n\mathbf{x},nt,n^{-1}h}(\xi) - |B| \frac{1+\sigma\cdot e^{-2|t|h}}{2} \right|^2$$
$$= \int_\Xi d\rho(\xi) \int_B d^2\mathbf{x}\, g_{\sigma,n\mathbf{x},nt,n^{-1}h}(\xi) \int_B d^2\mathbf{x}'\, g_{\sigma,n\mathbf{x}',nt,n^{-1}h}(\xi) - \int_\Xi d\rho(\xi) \left(|B| \frac{1+\sigma\cdot e^{-2|t|h}}{2} \right)^2,$$

where we applied Fubini's theorem to g, which is measurable over $(\mathbb{R}^2\times\Xi, d^2\mathbf{x}\otimes d\rho)$, since it is pointwise approximated by a sequence of finite linear combinations of products $\phi_i(\mathbf{x})\chi_i(\xi)$, with $\phi_i(\mathbf{x})$ and $\chi_i(\xi)$ measurable.

Since the points of ξ are ρ uncorrelated, in the above integration no contribution comes from the set

$$Q \equiv \{(\mathbf{x},\mathbf{x}') \in B\times B : n|y-y'| \geq n^{-1}h\}.$$

As $n\to\infty$, Q becomes all of $B\times B$; more precisely, taken a rectangle in $\mathbb{R}^2$ of base Δ_1 and height Δ_2 which contains B, it is

$$|(B\times B)\backslash Q| \leq \int_B d^2\mathbf{x}\, \Delta_1 \int_{y-n^{-2}h}^{y+n^{-2}h} dy' \leq n^{-2} 2h \Delta_1^2 \Delta_2$$

and therefore, since $\|g\|_\infty = 1$, we conclude

$$\int_\Xi d\rho(\xi) \left| \int_B d^2\mathbf{x}\, g_{\sigma,n\mathbf{x},nt,n^{-1}h}(\xi) - |B| \frac{1+\sigma\cdot e^{-2|t|h}}{2} \right|^2 \leq n^{-2} 2h \Delta_1^2 \Delta_2. \tag{4.24}$$

The same reasoning and the same inequality apply for functions of the form

$$\prod_{i=1}^k g_{\sigma_i, n(\mathbf{x}+t_{i-1}), n(t_i - t_{i-1}), n^{-1}h}(\xi) \quad \text{with} \quad t_0 = 0 < t_1 < \cdots < t_k,$$

which describe all measurements at k successive times. Moreover, since the g's in the above product depend on ξ only through points in disjoint sets, and since the points of ξ are ρ uncorrelated, it follows that

$$\int_B d^2\mathbf{x} \prod_{i=1}^k g_{\sigma_i, n(\mathbf{x}+t_{i-1}), n(t_i-t_{i-1}), n^{-1}h}(\xi)$$

converges, for the moment in $\| \,\|_2$ over (Ξ,ρ), to the product of the limits, i.e., to

$$|B| \prod_{i=1}^k \frac{1+\sigma_i\cdot e^{-2(t_i-t_{i-1})h}}{2}.$$

Since $\rho(\Xi)=1$, convergence also takes place in the L^1 norm. Now, if a sequence $\{f_j\}$ converges in $\| \,\|_1$ to f and $\Sigma_{j=1}^\infty \|f_j - f_{j-1}\|_1 \leq \infty$, then f_j converge pointwise almost everywhere to f,[27] and in our case it is sufficient to take a subsequence with $\Sigma 1/n_j < \infty$ to assure ρ—almost everywhere pointwise convergence.

Therefore, for any denumerable family $\mathcal{R}$ of bounded Borel sets $B\subset\mathbb{R}^2$, $\forall k=1,2,\ldots$, $\forall \rho_i = \pm 1$, $i=1,\ldots,k$, $\forall t_0 = 0 < t_1 < \cdots < t_k$ in a denumerable set, for ρ-almost all $\xi\in\Xi$,

$$\lim_j \int_B d^2\mathbf{x} \prod_{i=1}^{k} g_{\sigma_i, n_j(\mathbf{x}+t_{i-1}), n_j(t_i-t_{i-1}), n_j^{-1}h}(\xi) = |B| \prod_{i=1}^{k} \frac{1+\sigma_i \cdot e^{-2(t_i-t_{i-1})h}}{2}.$$

By taking $\mathcal{R}$ such that the linear combinations of characteristic functions χ_B are dense in $L^1(\mathbb{R}^2, d^2\mathbf{x})$, and using the $\| \|_\infty$ boundedness of the integrand, uniform in j, convergence follows for any bounded Borel subset of $\mathbb{R}^2$; as discussed above, this implies weak convergence of α_n^t, and convergence of the means at different times to a stochastic process, for times in a denumerable set, for ρ—almost all (fixed) configurations of obstacles.
The weak limit is given by

$$\alpha^t f(\mathbf{x},\sigma) = \sum_{\tau=\pm 1} f(\mathbf{x}+t, \tau\cdot\sigma) \frac{1+\tau e^{-2|t|h}}{2}. \tag{4.25}$$

Convergence of $\Pi_{i=1}^k g$ to $\Pi_{i=1}^k G$ implies that limits of measurements at successive times define a time-independent Markov process, with transition function

$$P_t((\mathbf{x},\sigma),(B,\tau)) = \alpha^t \chi_{B,\tau}(\mathbf{x},\sigma) = \sum_{\rho=\pm 1} \chi_{B,\tau}(\mathbf{x}+t, \rho\cdot\sigma) \frac{1+\rho e^{-2|t|h}}{2}.$$

In fact,

$$\alpha_n^t f(\mathbf{x},\sigma) = \sum_{\tau=\pm 1} f(\mathbf{x}+t, \tau\cdot\sigma)\; g_{\tau, n\mathbf{x}, nt, n^{-1}h}(\xi),$$

so that

$$\begin{aligned}
&\alpha_n^{t_1}(f_1 \alpha_n^{t_2-t_1}(f_2 \alpha_n^{t_3-t_2} \cdots (f_{k-1}\alpha_n^{t_k-t_{k-1}} f_k) \cdots))(\mathbf{x},\sigma) \\
&= \prod_{i=1}^{k} \sum_{\tau_i=\pm 1} f_i\left(\mathbf{x}+t_i, \sigma\cdot \prod_{j=1}^{i} \tau_j\right) g_{\tau_i, n(\mathbf{x}+t_{i-1}), n_j(t_i-t_{i-1}), n^{-1}h}(\xi)
\end{aligned}$$

which $\tau_{\mathcal{F}}$ converges as $n\to\infty$ to

$$\begin{aligned}
&\prod_{i=1}^{k} \sum_{\tau_i=\pm 1} f_i\left(\mathbf{x}+t_i, \sigma\cdot\prod_{j=1}^{i} \tau_j\right) \frac{1+\tau_i e^{-2(t_i-t_{i-1})h}}{2} \\
&= \alpha^{t_1}(f_1 \alpha^{t_2-t_1}(f_2 \alpha^{t_3-t_2} \cdots (f_{k-1}\alpha^{t_k-t_{k-1}} f_k) \cdots))(\mathbf{x},\sigma),
\end{aligned} \tag{4.26}$$

which is the condition for the stochastic process to be Markov and time independent. The process is defined by the left-hand side of Eq. (4.26) for all positive times, and the restriction to a denumerable set of times is only involved in the construction of a fixed set of configurations of rods (of ρ-measure one) for which weak convergence holds.

[1] O. Penrose, *Foundations of Statistical Mechanics. A Deductive Treatment* (Pergamon, New York, 1970).
[2] P. M. Alberti and A. Uhlmann, *Stochasticity and Partial Order* (Reidel Dortrecht, 1982).
[3] R. F. Streater, *Statistical Dynamics* (Imperial College Press, 1995).
[4] P. A. Martin, *Modèles en Mécanique Statistique des Processus Irréversibles*, Lectures Notes in Physics Vol. 103 (Springer, New York, 1979).
[5] H. Spohn, *Large Scale Dynamics of Interacting Particles* (Springer, Berlin, 1991).
[6] J. L. Lebowitz, Prog. Theor. Phys. Suppl. **64**, 35 (1978).
[7] D. Ruelle, *Hasard et Chaos* (Jacob, Paris 1991).
[8] J. L. Lebowitz, *Microscopic Reversibility and Macroscopic Behaviour: Physical Explanation and Mathematical Derivations*, in *25 years of Non-Equilibrium Statistical Mechanics*, Lectures Notes in Physics Vol. 445 (Springer, New York, 1995), and references therein; Physica A **194**, 1 (1993).
[9] J. R. Dorfman, Lecture notes, Institute for Physical Science and Technology, University of Maryland.
[10] L. Boltzmann, Wien. Ber. **66**, 275 (1872).
[11] S. G. Brush, *The Kind of Motion We Call Heat* (North-Holland, Amsterdam, 1976), Vol. 2.

[12] H. Grad, in *Handbuch der Physik* (Springer, Berlin, 1964), Vol. 12.
[13] O. E. Lanford, *Time Evolution of Large Classical Systems*, Lectures Notes in Physics Vol. 38 (Springer, New York, 1975).
[14] C. Cercignani, R. Illner, and M. Pulvirenti, *The Mathematical Theory of Diluite Gases* (Springer, New York, 1994).
[15] E. B. Davies, Commun. Math. Phys. **33**, 171 (1973); **39**, 31 (1974).
[16] L. Boltzmann, Wied. Ann. **57**, 773 (1896).
[17] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics* (Springer, Berlin, 1979).
[18] R. Haag, *Local Quantum Physics* (Springer, New York, 1993).
[19] D. W. Robinson, Commun. Math. Phys. **7**, 337 (1968).
[20] D. A. Dubin and G. L. Sewell, J. Math. Phys. **11**, 2990 (1970).
[21] G. L. Sewell, Commun. Math. Phys. **33**, 43 (1973).
[22] G. Morchio and F. Strocchi, Commun. Math. Phys. **99**, 153 (1985); J. Math. Phys. **28**, 622 (1987).
[23] J. R. Brown, Pac. J. Math. **16**, 13 (1966).
[24] R. V. Kadison, Topology **3**, 177 (1965).
[25] M. Takesaki, *Theory of operator algebras* (Springer, New York, 1979), Vol. I.
[26] R. Haag, N. M. Hugenholz, and M. Winnink, Commun. Math. Phys. **5**, 215 (1967).
[27] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975).
[28] J. Dixmier, *Les algebres D'operateurs dans L'espace Hilbertien* (Gauthier-Villars, Paris, 1969).
[29] F. Gentili, thesis, Pisa, 1996.
[30] D. Ruelle, *Statistical Mechanics: Rigorous Results* (Benjamin, New York, 1969).

A coupled AKNS–Kaup–Newell soliton hierarchy

Wen-Xiu Ma[a)]
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong

Ruguang Zhou[b)]
Department of Mathematics, Xuzhou Normal University, Xuzhou 221009, People's Republic of China

(Received 29 July 1998; accepted for publication 30 April 1999)

A coupled AKNS–Kaup–Newell hierarchy of systems of soliton equations is proposed in terms of hereditary symmetry operators resulted from Hamiltonian pairs. Zero curvature representations and tri-Hamiltonian structures are established for all coupled AKNS–Kaup–Newell systems in the hierarchy. Therefore all systems have infinitely many commuting symmetries and conservation laws. Two reductions of the systems lead to the AKNS hierarchy and the Kaup–Newell hierarchy, and thus those two soliton hierarchies also possess tri-Hamiltonian structures.
 [S0022-2488(99)01209-8]

I. INTRODUCTION

Systems of soliton equations usually come in hierarchies. This kind of hierarchy possesses many nice properties, for instance, Lax representations or zero curvature representations, infinitely many commuting symmetries and conservation laws, hereditary recursion structures, bi-Hamiltonian formulations, and even multiple Hamiltonian formulations, etc., and they are often called soliton hierarchies. Well-known examples of such soliton hierarchies (for example, see Refs. 1, 2) contain the KdV hierarchy, the MKdV hierarchy, the AKNS hierarchy, the Kaup–Newell hierarchy, the Benjamin–Ono hierarchy,[3] the Tu hierarchy,[4] the Dirac hierarchy,[5] the coupled KdV hierarchies,[6,7] the coupled Harry–Dym hierarchies,[8] the coupled Burgers hierarchies,[9] and so on. It is very interesting to search for new soliton hierarchies, even hierarchies of systems that possess only infinitely many commuting symmetries.

An idea that allows us to achieve this is to construct soliton hierarchies of coupled systems of equations. It could be divided into two aspects in view of types of soliton equations. The one is to construct soliton hierarchies by coupling systems of the same type. Such examples are the coupled KdV hierarchies,[6,7] the coupled Harry–Dym hierarchies,[8] the coupled Burgers hierarchies,[9] and the perturbation systems of the KdV hierarchy,[10] etc. The other is to construct soliton hierarchies by coupling systems of different types. There are few examples in this aspect. A coupled AKNS–Kaup–Newell hierarchy of a complex form, recently introduced by Zhang,[11] is such an example.

In this paper, motivated by Zhang's coupled AKNS–Kaup–Newell hierarchy of a complex form, we would like to propose a hierarchy of coupled AKNS–Kaup–Newell evolution equations of real form. The hierarchy will be established in the second section, in terms of hereditary symmetry operators. The required hereditary symmetry operators can be generated by observing a set of Hamiltonian operators. Zero curvature representations will be computed in the third section for all systems in the hierarchy. Interestingly the discussion of the fourth section shows that all the systems have not only bi-Hamiltonian structures but also tri-Hamiltonian structures, although Zhang did not present Hamiltonian structures and consequent conservation laws due to a failure in determining Hamiltonian operators.[11] Therefore the resulting hierarchy has infinitely many commuting symmetries and conservation laws. Some concluding remarks are given in the last section.

[a)]Electronic mail: mawx@math.cityu.edu.hk
[b)]Electronic mail: rgzhou@public.xz.js.cn

II. HEREDITARY SYMMETRY OPERATORS

We want to establish a coupled AKNS–Kaup–Newell hierarchy in terms of hereditary symmetry operators resulted from Hamiltonian pairs. To this end, let us introduce a set of 2×2 matrix integrodifferential operators:

$$M=M(u)=\begin{pmatrix} \alpha_1 q\,\partial^{-1}q & \alpha_2+\alpha_3\partial-\alpha_1 q\,\partial^{-1}r \\ -\alpha_2+\alpha_3\partial-\alpha_1 r\,\partial^{-1}q & \alpha_1 r\,\partial^{-1}r \end{pmatrix},\quad u=\begin{pmatrix} q \\ r \end{pmatrix}, \tag{1}$$

where $\partial=\partial/\partial x$, $q=q(x,t)$, $r=r(x,t)$, and $\alpha_1,\alpha_2,\alpha_3$ are three arbitrary constants, and consider their Hamiltonian property. They are simple generalizations of the Hamiltonian operators in the AKNS case.[12] The following proposition shows that they are still Hamiltonian, indeed.

Proposition 1: The 2×2 *matrix integrodifferential operators defined by* (1) *are all Hamiltonian for any constants* $\alpha_1,\alpha_2,\alpha_3$.

Proof: Assume that

$$X=(X_1,X_2)^T,\quad Y=(Y_1,Y_2)^T,\quad Z=(Z_1,Z_2)^T,\quad W=(W_1,W_2)^T,$$

are two-dimensional vectors of functions. Since M is skew symmetric, we only need to prove that the Jacobi identity,

$$\langle Z,M'[MX]Y\rangle+\text{cycle}(X,Y,Z)\equiv 0(\text{mod}\,\partial), \tag{2}$$

holds for any X,Y,Z, where $\langle\cdot,\cdot\rangle$ denotes the standard inner product of $\mathbb{R}^2$. By (1), we immediately have

$$MX=\begin{pmatrix} \alpha_2X_2+\alpha_3X_{2x}+\alpha_1 qP(X) \\ -\alpha_2X_1+\alpha_3X_{1x}-\alpha_1 rP(X) \end{pmatrix}:=\begin{pmatrix} W_1(X) \\ W_2(X) \end{pmatrix},$$

where $P(X)=\partial^{-1}(qX_1-rX_2)$. Following the definition of the Gateaux derivative, two objects $M'[W]$ and $M'[W]Y$ are computed as follows:

$$M'[W]=\begin{pmatrix} \alpha_1 q\,\partial^{-1}W_1+\alpha_1W_1\,\partial^{-1}q & -\alpha_1 q\,\partial^{-1}W_2-\alpha_1W_1\,\partial^{-1}r \\ -\alpha_1 r\,\partial^{-1}W_1-\alpha_1W_2\,\partial^{-1}q & \alpha_1 r\,\partial^{-1}W_2+\alpha_1W_2\,\partial^{-1}r \end{pmatrix},$$

$$M'[W]Y=\begin{pmatrix} \alpha_1 q\partial^{-1}(W_1Y_1-W_2Y_2)+\alpha_1W_1\partial^{-1}(qY_1-rY_2) \\ -\alpha_1 r\partial^{-1}(W_1Y_1-W_2Y_2)-\alpha_1W_2\partial^{-1}(qY_1-rY_2) \end{pmatrix}.$$

Now we can have

$$\begin{aligned}\langle Z,M'[MX]Y\rangle&=\alpha_1(qZ_1-rZ_2)\partial^{-1}(W_1(X)Y_1-W_2(X)Y_2)\\&\quad+\alpha_1(W_1(X)Z_1-W_2(X)Z_2)\partial^{-1}(qY_1-rY_2).\end{aligned} \tag{3}$$

Upon observing that

$$\begin{aligned}W_1(X)Y_1-W_2(X)Y_2&=(\alpha_2X_2+\alpha_3X_{2x}+\alpha_1 qP(X))Y_1-(-\alpha_2X_1+\alpha_3X_{1x}-\alpha_1 rP(X))Y_2\\&=\alpha_1P(X)(qY_1+rY_2)+\alpha_2(X_2Y_1+X_1Y_2)+\alpha_3(X_{2x}Y_1-X_{1x}Y_2),\end{aligned}$$

we can make a decomposition,

$$\langle Z,M'[MX]Y\rangle=R(X,Y,Z)+S(X,Y,Z)+T(X,Y,Z), \tag{4}$$

where R,S,T are defined by

$$R(X,Y,Z)=\alpha_1^2(qZ_1-rZ_2)\partial^{-1}[P(X)(qY_1+rY_2)]+\alpha_1^2P(X)(qZ_1+rZ_2)\partial^{-1}(qY_1-rY_2), \tag{5}$$

$$S(X,Y,Z)=\alpha_1\alpha_2(qZ_1-rZ_2)\partial^{-1}(X_2Y_1+X_1Y_2)+\alpha_1\alpha_2(X_2Z_1+X_1Z_2)\partial^{-1}(qY_1-rY_2), \tag{6}$$

$$T(X,Y,Z)=\alpha_1\alpha_3(qZ_1-rZ_2)\partial^{-1}(X_{2x}Y_1-X_{1x}Y_2)+\alpha_1\alpha_3(X_{2x}Z_1-X_{1x}Z_2)\partial^{-1}(qY_1-rY_2). \tag{7}$$

For these three functions R,S,T, we can compute that

$$R(X,Y,Z)+\text{cycle}(X,Y,Z)=\alpha_1^2\partial\{P(Z)\partial^{-1}[P(X)(qY_1+rY_2)]\}+\text{cycle}(X,Y,Z), \tag{8}$$

$$S(X,Y,Z)+\text{cycle}(X,Y,Z)=\alpha_1\alpha_2\partial[P(Z)\partial^{-1}(X_2Y_1+X_1Y_2)]+\text{cycle}(X,Y,Z), \tag{9}$$

$$\begin{aligned}T(X,Y,Z)&+\text{cycle}(X,Y,Z)\\ &=\alpha_1\alpha_3(qZ_1-rZ_2)(X_2Y_1-X_1Y_2)+\alpha_1\alpha_3(qZ_1-rZ_2)\partial^{-1}(X_1Y_{2x}-X_2Y_{1x})\\ &\quad+\alpha_1\alpha_3(Z_1X_{2x}-Z_2X_{1x})\partial^{-1}(qY_1-rY_2)+\text{cycle}(X,Y,Z)\\ &=\alpha_1\alpha_3(qZ_1-rZ_2)(X_1Y_{2x}-X_2Y_{1x})+\alpha_1\alpha_3(Z_1X_{2x}-Z_2X_{1x})\partial^{-1}(qY_1-rY_2)\\ &\quad+\text{cycle}(X,Y,Z)\\ &=\alpha_1\alpha_3\partial[P(Z)\partial^{-1}(X_1Y_{2x}-X_2Y_{1x})]+\text{cycle}(X,Y,Z).\end{aligned} \tag{10}$$

They are all total derivatives and thus combining the decomposition (4) and the equalities (8), (9), (10) leads to the Jacobi identity (2). This completes the proof. ■

Now we would like to discuss some special cases of Hamiltonian pairs starting from the above Hamiltonian operators defined by (1), which allows us to generate hereditary symmetry operators and further soliton hierarchies. This idea has been successfully used to construct bi-Hamiltonian coupled KdV systems.[13,14] An important phenomenon we want to point out is that different soliton hierarchies can be derived from Hamiltonian operators of the same type. The following discussion shows an example of such phenomenon. It is also important to realize that not all Hamiltonian pairs may generate hereditary symmetry operators. Thus, care must be taken to restrict our attention to the cases where there exists at least one invertible Hamiltonian operator for each Hamiltonian pair. The required invertibility guarantees that Hamiltonian pairs can yield hereditary symmetry operators.[15]

Case 1: We make a choice of an invertible Hamiltonian operator,

$$J=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}, \tag{11}$$

which has an inverse operator,

$$J^{-1}=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}.$$

It follows from Proposition 1 that this operator J constitutes a Hamiltonian pair with M defined by (1). Therefore we can have a hereditary symmetry operator,

$$\Phi=MJ^{-1}=\begin{pmatrix}\alpha_2+\alpha_3\partial-\alpha_1q\,\partial^{-1}r & -\alpha_1q\,\partial^{-1}q\\ \alpha_1r\,\partial^{-1}r & \alpha_2-\alpha_3\partial+\alpha_1r\,\partial^{-1}q\end{pmatrix}, \tag{12}$$

where α_1, α_2, and α_3 are arbitrary. The reduction of $\alpha_1=-1$, $\alpha_2=0$ and $\alpha_3=\frac{1}{2}$ leads to the recursion operator for the normal AKNS hierarchy.[1,2,16]

Case 2: We make a choice of a Hamiltonian pair,

$$J=\begin{pmatrix} \beta_1 q\,\partial^{-1}q & 1-\beta_1 q\partial^{-1}r \\ -1-\beta_1 r\,\partial^{-1}q & \beta_1 r\,\partial^{-1}r \end{pmatrix},\quad M=\begin{pmatrix} 0 & \alpha_3\partial \\ \alpha_3\partial & 0 \end{pmatrix}. \tag{13}$$

The above proposition ensures that they constitute a Hamiltonian pair, indeed. Since the operator J has an invertible operator,

$$J^{-1}=\begin{pmatrix} \beta_1 r\,\partial^{-1}r & -1+\beta_1 r\,\partial^{-1}q \\ 1+\beta_1 q\,\partial^{-1}r & \beta_1 q\,\partial^{-1}q \end{pmatrix},$$

we can obtain the corresponding hereditary symmetry operator,

$$\Phi=MJ^{-1}=\begin{pmatrix} \alpha_3\partial+\alpha_3\beta_1\partial q\,\partial^{-1}r & \alpha_3\beta_1\partial q\,\partial^{-1}q \\ \alpha_3\beta_1\partial r\,\partial^{-1}r & -\alpha_3\partial+\alpha_3\beta_1 r\,\partial^{-1}r \end{pmatrix}, \tag{14}$$

where α_3 and β_1 are arbitrary. The reduction of $\alpha_3=\frac{1}{2}$ and $\beta_1=-1$ leads to the recursion operator for the normal Kaup–Newell hierarchy (for example, see Ref. 17).

More generally, we have the following case, which combines the above two cases.

Case 3: We make a choice of an invertible Hamiltonian operator,

$$J=\begin{pmatrix} \beta_1 q\,\partial^{-1}q & \beta_2-\beta_1 q\,\partial^{-1}r \\ -\beta_2-\beta_1 r\,\partial^{-1}q & \beta_1 r\,\partial^{-1}r \end{pmatrix}, \tag{15}$$

whose inverse operator can be given by

$$J^{-1}=\frac{1}{\beta_2^2}\begin{pmatrix} \beta_1 r\,\partial^{-1}r & -\beta_2+\beta_1 r\,\partial^{-1}q \\ \beta_2+\beta_1 q\,\partial^{-1}r & \beta_1 q\,\partial^{-1}q \end{pmatrix}.$$

It follows from Proposition 1 that the operator J constitutes a Hamiltonian pair with the Hamiltonian operator M defined by (1). In this case, we can generate the following corresponding hereditary symmetry operator:

$$\Phi=MJ^{-1}=\frac{1}{\beta_2^2}\begin{pmatrix} \alpha_2\beta_2+\alpha_3\beta_2\partial+(\alpha_2\beta_1-\alpha_1\beta_2)q\,\partial^{-1}r+\alpha_3\beta_1\,\partial q\,\partial^{-1}r, & (\alpha_2\beta_1-\alpha_1\beta_2)q\,\partial^{-1}q+\alpha_3\beta_1\,\partial q\,\partial^{-1}q \\ (\alpha_1\beta_2-\alpha_2\beta_1)r\,\partial^{-1}r+\alpha_3\beta_1\,\partial r\,\partial^{-1}r, & \alpha_2\beta_2-\alpha_3\beta_2\partial+(\alpha_1\beta_2-\alpha_2\beta_1)r\,\partial^{-1}q+\alpha_3\beta_1\,\partial r\,\partial^{-1}q \end{pmatrix}, \tag{16}$$

where five constants are arbitrary but $\beta_2\neq 0$.

Let us pick out a special subcase of $\alpha_2=0$ and $\beta_2=1$ from the third case. If $\alpha_3=0$, we just obtain a simple hereditary symmetry operator,

$$\Phi=\begin{pmatrix} -\alpha q\,\partial^{-1}r & -\alpha q\,\partial^{-1}q \\ \alpha r\,\partial^{-1}r & \alpha r\,\partial^{-1}q \end{pmatrix}, \tag{17}$$

where $\alpha=\alpha_1$ is arbitrary. This is equivalent to the first case with $\alpha_2=\alpha_3=0$.

If $\alpha_3\neq 0$, upon resetting $\alpha_3=\gamma$, $\alpha_1=\alpha$, $\alpha_3\beta_1=\beta$, we obtain a hereditary symmetry operator,

$$\Phi=\begin{pmatrix} \gamma\partial-\alpha q\,\partial^{-1}r+\beta\,\partial q\,\partial^{-1}r & -\alpha q\,\partial^{-1}q+\beta\,\partial q\,\partial^{-1}q \\ \alpha r\,\partial^{-1}r+\beta\,\partial r\,\partial^{-1}r & -\gamma\partial+\alpha r\,\partial^{-1}q+\beta\,\partial r\,\partial^{-1}q \end{pmatrix}, \tag{18}$$

where α,β,γ are arbitrary but $\gamma\neq 0$. Note that if we let the constant γ go to zero, the hereditary condition for Φ with a general constant γ becomes the one for Φ with $\gamma=0$. Therefore the constant γ can be chosen as zero, which does not affect the hereditary property of Φ. However, if $\gamma=0$, we do not know whether the operator Φ defined by (18) is decomposable, i.e, whether there exists any Hamiltonian pair J and M so that $\Phi=MJ^{-1}$.

We will focus on discussing a soliton hierarchy generated by the hereditary symmetry operator in (18) because of its generality. For $\gamma\neq 0$, we can rescale three constants to put a general case into a special case of the operator Φ defined by (18), and thus we pick out the following special case:

$$\Phi=MJ^{-1}=\begin{pmatrix} \frac{1}{2}\partial-\alpha q\,\partial^{-1}r-\frac{1}{2}\beta\,\partial q\,\partial^{-1}r & -\alpha q\,\partial^{-1}q-\frac{1}{2}\beta\,\partial q\,\partial^{-1}q \\ \alpha r\,\partial^{-1}r-\frac{1}{2}\beta\,\partial r\,\partial^{-1}r & -\frac{1}{2}\partial+\alpha r\,\partial^{-1}q-\frac{1}{2}\beta\,\partial r\,\partial^{-1}q \end{pmatrix}, \tag{19}$$

to discuss without loss of generality. To this special case, the corresponding hierarchy of evolution equations,

$$\begin{pmatrix} q_t \\ r_t \end{pmatrix}=\begin{pmatrix} \frac{1}{2}\partial-\alpha q\,\partial^{-1}r-\frac{1}{2}\beta\,\partial q\,\partial^{-1}r & -\alpha q\,\partial^{-1}q-\frac{1}{2}\beta\,\partial q\,\partial^{-1}q \\ \alpha r\,\partial^{-1}r-\frac{1}{2}\beta\,\partial r\,\partial^{-1}r & -\frac{1}{2}\partial+\alpha r\,\partial^{-1}q-\frac{1}{2}\beta\,\partial r\,\partial^{-1}q \end{pmatrix}^n \begin{pmatrix} q_x \\ r_x \end{pmatrix}, \quad n\geqslant 0, \tag{20}$$

contains two important reductions. If $\alpha\neq 0$ but $\beta=0$, the hierarchy reduces to the AKNS hierarchy. If $\alpha=0$ but $\beta\neq 0$ but $\beta=0$, the hierarchy reduces to the Kaup–Newell hierarchy. Thus, the hierarchy (20) generated by the hereditary symmetry operator (19) is called a coupled AKNS–Kaup–Newell hierarchy. All systems in the hierarchy (20) are real. Therefore the hierarchy (20) is a soliton hierarchy that we want to construct.

III. ZERO CURVATURE REPRESENTATIONS

In the previous section, we generated a coupled AKNS–Kaup–Newell hierarchy of real form by observing Hamiltonian operators. More importantly, the resulting hierarchy shares some common integrable properties. In this section, we want to show zero curvature representations for all systems in the hierarchy, and in the next section, we will establish tri-Hamiltonian structures.

To show zero curvature representations, let us impose a spectral problem,

$$\phi_x=U\phi, \quad U=U(u,\lambda)=\begin{pmatrix} \lambda & q \\ (\alpha+\beta\lambda)r & -\lambda \end{pmatrix}, \quad \phi=\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \tag{21}$$

where λ is a spectral parameter, and α and β are arbitrary constants. It is customary to solve the stationery zero curvature equation $V_x=[U,V]$ first. Suppose that

$$V=V(u,\lambda)=\begin{pmatrix} a & b \\ (\alpha+\beta\lambda)c & -a \end{pmatrix}=\sum_{i\geqslant 0}\begin{pmatrix} a_i & b_i \\ (\alpha+\beta\lambda)c_i & -a_i \end{pmatrix}\lambda^{-i}, \tag{22}$$

and then the stationary zero curvature equation becomes

$$a_x=(\alpha+\beta\lambda)(qc-rb), \quad b_x=2\lambda b-2qa, \quad c_x=2ra-2\lambda c. \tag{23}$$

Notice that a recursion relation to determine b and c may be found if we fix $a=(\alpha+\beta\lambda)\partial^{-1}(qc-rb)$. Actually, we have

$$b_x=2\lambda b-2(\alpha+\beta\lambda)q\partial^{-1}(qc-rb), \quad c_x=2(\alpha+\beta\lambda)r\partial^{-1}(qc-rb)-2\lambda c, \tag{24}$$

which equivalently leads to

$$\begin{pmatrix} -2\beta q\,\partial^{-1}q & 2+2\beta q\,\partial^{-1}r \\ -2+2\beta r\,\partial^{-1}q & -2\beta r\,\partial^{-1}r \end{pmatrix}\begin{pmatrix} c_{i+1} \\ b_{i+1} \end{pmatrix}=\begin{pmatrix} 2\alpha q\,\partial^{-1}q & \partial-2\alpha q\,\partial^{-1}r \\ \partial-2\alpha r\,\partial^{-1}q & 2\alpha r\,\partial^{-1}r \end{pmatrix}\begin{pmatrix} c_i \\ b_i \end{pmatrix}, \tag{25}$$

where $i \geqslant 0$. If we set

$$J = \begin{pmatrix} -2\beta q\,\partial^{-1} q & 2+2\beta q\,\partial^{-1} r \\ -2+2\beta r\,\partial^{-1} q & -2\beta r\,\partial^{-1} r \end{pmatrix}, \quad M = \begin{pmatrix} 2\alpha q\,\partial^{-1} q & \partial - 2\alpha q\,\partial^{-1} r \\ \partial - 2\alpha r\,\partial^{-1} q & 2\alpha r\,\partial^{-1} r \end{pmatrix}, \tag{26}$$

the operators J and M constitute a Hamiltonian pair, based on the result in the previous section. It is apparent that the corresponding hereditary symmetry operator $\Phi = MJ^{-1}$ is exactly the same as the one defined by (19), having noted that

$$J^{-1} = \frac{1}{2}\begin{pmatrix} -\beta\,\partial r\,\partial^{-1} r & -1-\beta\,\partial r\,\partial^{-1} q \\ 1-\beta\,\partial q\,\partial^{-1} r & -\beta\,\partial q\,\partial^{-1} q \end{pmatrix}.$$

The conjugate operator of Φ reads as

$$\Psi = \Phi^{\dagger} = J^{-1}M = \begin{pmatrix} -\frac{1}{2}\partial + \alpha r\,\partial^{-1} q - \frac{1}{2}\beta r\,\partial^{-1} q\,\partial & -\alpha r\,\partial^{-1} r - \frac{1}{2}\beta r\,\partial^{-1} r\,\partial \\ \alpha q\,\partial^{-1} q - \frac{1}{2}\beta q\,\partial^{-1} q\,\partial & \frac{1}{2}\partial - \alpha q\,\partial^{-1} r - \frac{1}{2}\beta q\,\partial^{-1} r\,\partial \end{pmatrix}. \tag{27}$$

Therefore upon noting (23) and choosing $a_0 = 1$, we obtain a solution to the stationary zero curvature equation $V_x = [U,V]$:

$$a_0 = 1, \quad b_0 = c_0 = 0; \quad b_1 = q, \quad c_1 = r;$$

$$\begin{pmatrix} c_{i+1} \\ b_{i+1} \end{pmatrix} = \Psi \begin{pmatrix} c_i \\ b_i \end{pmatrix}, \quad i \geqslant 1,$$

$$a_i = \alpha\partial^{-1}(qc_i - rb_i) + \beta\partial^{-1}(qc_{i+1} - rb_{i+1}), \quad i \geqslant 1; \tag{28}$$

from which we can get

$$\begin{pmatrix} c_2 \\ b_2 \end{pmatrix} = \Psi \begin{pmatrix} r \\ q \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -r_x - \beta q r^2 \\ q_x - \beta q^2 r \end{pmatrix},$$

and

$$a_1 = \beta\partial^{-1}(qc_2 - rb_2) = -\tfrac{1}{2}\beta q r.$$

It should be noted that we always need to select zero constants for integration in deriving a_i, b_i, c_i, $i \geqslant 1$. That is, we require that $a_i|_{[u]=0} = b_i|_{[u]=0} = c_i|_{[u]=0} = 0$, $i \geqslant 1$, where $[u] = (u, u_x, \ldots)$.

Now we can express the coupled AKNS–Kaup–Newell hierarchy (20) in another way. Let us define

$$u_t = K_n := J\begin{pmatrix} c_{n+1} \\ b_{n+1} \end{pmatrix} = \Phi^n \begin{pmatrix} 2q \\ -2r \end{pmatrix}, \quad n \geqslant 0, \tag{29}$$

where Φ is defined by (19). The first three systems of the hierarchy (29) are

$$u_t = \begin{pmatrix} q_t \\ r_t \end{pmatrix} = K_0 = J\begin{pmatrix} c_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} 2q \\ -2r \end{pmatrix},$$

$$u_t = \begin{pmatrix} q_t \\ r_t \end{pmatrix} = K_1 = J\begin{pmatrix} c_2 \\ b_2 \end{pmatrix} = \begin{pmatrix} q_x \\ r_x \end{pmatrix},$$

$$u_t=\begin{pmatrix} q_t \\ r_t \end{pmatrix}=K_2=J\begin{pmatrix} c_3 \\ b_3 \end{pmatrix}=\frac{1}{2}\begin{pmatrix} q_{xx}-2\alpha q^2r-\beta(q^2r)_x \\ -r_{xx}+2\alpha qr^2-\beta(qr^2)_x \end{pmatrix}. \tag{30}$$

Since $K_1=u_x$, all systems in the hierarchy (29), except the first system $u_t=K_0$, are exactly the coupled AKNS–Kaup–Newell systems in the hierarchy (20). Therefore (29) is another expression for the coupled AKNS–Kaup–Newell hierarchy (20).

Let us turn to the construction of zero curvature representations for all coupled AKNS–Kaup–Newell systems in the soliton hierarchy (29). We need a condition of $\alpha^2+\beta^2\neq 0$. With this condition, we have the injective property of the Gateaux derivative of U with respect to u, which is required in deriving systems of evolution equations from zero curvature equations. If the condition of $\alpha^2+\beta^2\neq 0$ is not satisfied, then the systems defined by (29) are linear and separated, and thus they are all trivial.

We choose Lax operators $V^{(n)}$ for $n\geqslant 0$ as

$$V^{(n)}=V^{(n)}(u,\lambda)=\overline{V}^{(n)}+\Delta_n, \quad \Delta_n=\begin{pmatrix} \delta_{1n} & 0 \\ 0 & \delta_{2n} \end{pmatrix}, \tag{31}$$

$$\overline{V}^{(n)}=\sum_{j=0}^{n}\begin{pmatrix} a_j & b_j \\ (\alpha+\beta\lambda)c_j & -a_j \end{pmatrix}\lambda^{n-j}=\begin{pmatrix} (\lambda^n a)_+ & (\lambda^n b)_+ \\ (\alpha+\beta\lambda)(\lambda^n c)_+ & -(\lambda^n a)_+ \end{pmatrix}, \tag{32}$$

where the subscript denotes choosing the polynomial part in λ, and δ_{1n} and δ_{2n} are two functions to be determined. At this moment, we can compute that

$$\overline{V}_x^{(n)}-[U,\overline{V}^{(n)}]=\begin{pmatrix} a_{nx}-\alpha(qc_n-rb_n) & b_{nx}+2qa_n \\ (\alpha+\beta\lambda)(c_{nx}-2ra_n) & -a_{nx}+\alpha(qc_n-rb_n) \end{pmatrix},$$

$$\Delta_{nx}-[U,\Delta_n]=\begin{pmatrix} \delta_{1n,x} & q(\delta_{1n}-\delta_{2n}) \\ -(\alpha+\beta\lambda)r(\delta_{1n}-\delta_{2n}) & \delta_{2n,x} \end{pmatrix}.$$

Therefore, if we take a choice,

$$\delta_{1n}=-\delta_{2n}=-a_n+\alpha\partial^{-1}(qc_n-rb_n), \quad n\geqslant 1, \tag{33}$$

then noting the injective property of U' under $\alpha^2+\beta^2\neq 0$, the zero curvature equation,

$$U_t-V_x^{(n)}+[U,V^{(n)}]=0, \tag{34}$$

equivalently yields the coupled AKNS–Kaup–Newell system,

$$u_t=K_n=M\begin{pmatrix} c_n \\ b_n \end{pmatrix}=J\begin{pmatrix} c_{n+1} \\ b_{n+1} \end{pmatrix},$$

for each $n\geqslant 1$. Moreover, it is easy to see that $u_t=K_0$ has a Lax pair U and $\overline{V}^{(0)}$. Therefore each coupled AKNS–Kaup–Newell system $u_t=K_n$ has a zero curvature representation with the Lax pair U and $V^{(n)}$ if we adopt $\delta_{10}=\delta_{20}=0$. We remark that for the systems $u_t=K_n$ with $\alpha=\beta=0$, $n\geqslant 0$, the above zero curvature representations still hold, but they are not sufficient, because we lose the injective property of U' in the case of $\alpha=\beta=0$.

IV. TRI-HAMILTONIAN STRUCTURES

To establish some kind of tri-Hamiltonian structures for the coupled AKNS–Kaup–Newell hierarchy, let us impose a third Hamiltonian operator,

$$N=M\Psi=\begin{pmatrix} -\alpha q\,\partial^{-1}q\,\partial+\alpha\,\partial q\,\partial^{-1}q-\frac{1}{2}\beta\,\partial q\,\partial^{-1}q\,\partial, & \frac{1}{2}\partial^2-\alpha q\,\partial^{-1}r\,\partial-\alpha\,\partial q\,\partial^{-1}r-\frac{1}{2}\beta\,\partial q\,\partial^{-1}r\,\partial \\ -\frac{1}{2}\partial^2+\alpha\,\partial r\,\partial^{-1}q+\alpha r^{-1}q\partial-\frac{1}{2}\beta\,\partial r\,\partial^{-1}q\,\partial, & -\alpha\,\partial r\,\partial^{-1}r+\alpha r\,\partial^{-1}r\,\partial-\frac{1}{2}\beta\,\partial r\,\partial^{-1}r\,\partial \end{pmatrix}. \tag{35}$$

It constitutes a Hamiltonian triple with J and M defined by (26), for any constants α, β, γ. That is, any linear combination of J, M, and N is still a Hamiltonian operator, which is automatically satisfied since J and M are a Hamiltonian pair.

Let us consider the first nonlinear system in the coupled AKNS–Kaup–Newell hierarchy (20):

$$u_t=K_2=M\begin{pmatrix} c_2\\ b_2\end{pmatrix}=\frac{1}{2}\begin{pmatrix} q_{xx}-2\alpha q^2r-\beta(q^2r)_x\\ -r_{xx}+2\alpha qr^2-\beta(qr^2)_x\end{pmatrix}. \tag{36}$$

It is apparent that this system could be written in three ways as

$$u_t=K_2=J\begin{pmatrix} c_3\\ b_3\end{pmatrix}=M\begin{pmatrix} c_2\\ b_2\end{pmatrix}=N\begin{pmatrix} c_1\\ b_1\end{pmatrix}.$$

Moreover, a direct calculation can show three gradient vectors,

$$\begin{pmatrix} c_1\\ b_1\end{pmatrix}=\begin{pmatrix} r\\ q\end{pmatrix}=\frac{\delta H_0}{\delta u},\quad H_0=qr; \tag{37}$$

$$\begin{pmatrix} c_2\\ b_2\end{pmatrix}=\Psi\begin{pmatrix} c_1\\ b_1\end{pmatrix}=\frac{1}{2}\begin{pmatrix} -r_x-\beta qr^2\\ q_x-\beta q^2r\end{pmatrix}=\frac{\delta H_1}{\delta u},$$

$$H_1=-\tfrac{1}{4}\beta q^2r^2-\tfrac{1}{4}qr_x+\tfrac{1}{4}q_xr; \tag{38}$$

$$\begin{pmatrix} c_3\\ b_3\end{pmatrix}=\Psi\begin{pmatrix} c_2\\ b_2\end{pmatrix}=\frac{1}{4}\begin{pmatrix} r_{xx}-2\alpha qr^2+3\beta qrr_x+\frac{3}{2}\beta^2q^2r^3\\ q_{xx}-2\alpha q^2r-3\beta qq_xr+\frac{3}{2}\beta^2q^3r^2\end{pmatrix}=\frac{\delta H_2}{\delta u},$$

$$H_2=\tfrac{1}{8}qr_{xx}+\tfrac{1}{8}q_{xx}r-\tfrac{1}{4}\alpha q^2r^2+\tfrac{3}{16}\beta q^2rr_x-\tfrac{3}{16}\beta qq_xr^2+\tfrac{1}{8}\beta^2q^3r^3. \tag{39}$$

Therefore a tri-Hamiltonian structure for the coupled AKNS–Kaup–Newell system (36) can be given by

$$u_t=K_2=J\,\frac{\delta H_2}{\delta u}=M\,\frac{\delta H_1}{\delta u}=N\,\frac{\delta H_0}{\delta u}, \tag{40}$$

where three Hamiltonian functions H_0, H_1, and H_2 are defined by (37), (38), and (39), respectively. Based on the recursion scheme in Refs. 18, 19, this leads to a tri-Hamiltonian structure for each nonlinear system in the coupled AKNS–Kaup–Newell hierarchy,

$$u_t=K_n=J\,\frac{\delta H_n}{\delta u}=M\,\frac{\delta H_{n-1}}{\delta u}=N\,\frac{\delta H_{n-2}}{\delta u},\quad n\geqslant 2. \tag{41}$$

The existence of all Hamiltonian functions H_n to satisfy $\delta H_n/\delta u=\Psi^n(\delta H_0/\delta u)$, $n\geqslant 0$, is guaranteed by the hereditary property of the hereditary symmetry operator Φ. They are all common conserved densities for the whole AKNS–Kaup–Newell hierarchy and commute with each other under three Poisson brackets associated with J, M, and N. This is because, for example, we can compute that

$$\{H_m,H_n\}_J:=\int\left\langle\frac{\delta H_m}{\delta u},J\frac{\delta H_n}{\delta u}\right\rangle dx=\int\left\langle\frac{\delta H_m}{\delta u},J\Psi\frac{\delta H_{n-1}}{\delta u}\right\rangle dx=\int\left\langle\frac{\delta H_m}{\delta u},\Phi J\frac{\delta H_{n-1}}{\delta u}\right\rangle dx$$
$$=\int\left\langle\Psi\frac{\delta H_m}{\delta u},J\frac{\delta H_{n-1}}{\delta u}\right\rangle dx=\{H_{m+1},H_{n-1}\}_J=\cdots=\{H_n,H_m\}_J,$$
$$m<n,\quad m,n\geqslant 0.$$

It gives rise to the commutativity of the conserved densities H_n, $n\geqslant 0$, by combining the skew-symmetric property of the Poisson brackets. Furthermore, we have

$$[K_m,K_n]=J\frac{\delta}{\delta u}\{H_m,H_n\}=0,\quad m,n\geqslant 0,\tag{42}$$

which implies that each coupled AKNS–Kaup–Newell system has infinitely many commuting symmetries. This may also be seen from a zero Lie derivative $L_{u_x}\Phi=0$. The property of $L_{u_x}\Phi=0$ also guarantees that the hereditary symmetry operator defined by (19) is a common recursion operator for all systems in the coupled AKNS–Kaup–Newell hierarchy (29).

V. CONCLUDING REMARKS

We have introduced a set of Hamiltonian operators and presented some corresponding hereditary symmetry operators. Therefore a coupled AKNS–Kaup–Newell hierarchy of systems of soliton equations of a real form is proposed. Zero curvature representations and tri-Hamiltonian structures are established for all systems in the hierarchy.

Interestingly, this coupled AKNS–Kaup–Newell hierarchy contains two different reductions of the AKNS hierarchy and the Kaup–Newell hierarchy. A natural problem we want to ask is what conditions could be found for the existence of similar coupled soliton hierarchies associated with two or more given soliton hierarchies and how one constructs such coupled soliton hierarchies if they exist.

Because our coupled AKNS–Kaup–Newell hierarchy includes the AKNS hierarchy and the Kaup–Newell hierarchy as two simple reductions, tri-Hamiltonian structures can be constructed for the AKNS hierarchy and the Kaup–Newell hierarchy, based on the obtained tri-Hamiltonian structures of the coupled AKNS–Kaup–Newell hierarchy. The corresponding tri-Hamiltonian structure for the Kaup–Newell system of nonlinear derivative Schrödinger equations has been raised recently in Ref. 20 and a nonlinearization problem has been manipulated for the associated spectral problem.[21] We believe that some other nice properties may also be achieved for the the coupled AKNS–Kaup–Newell hierarchy.

It is worthy pointing out that by using a similar deduction to the one in Section III, a general hereditary symmetry operator defined by (18) can be constructed from the following spectral problem:

$$\phi_x=U\phi,\quad U=U(u,\lambda)=\begin{pmatrix}\frac{1}{2\gamma}\lambda & q\\ \frac{1}{2\gamma}\left(\alpha-\frac{\beta}{\gamma}\lambda\right)r & -\frac{1}{2\gamma}\lambda\end{pmatrix},\tag{43}$$

with the same constants α, β, γ as ones in (18). It is apparent that the condition of $\gamma\neq 0$ is required, but α and β may be equal to zero. Only a condition of $\alpha^2+\beta^2\neq 0$ is needed for α and β, in order to guarantee the injective property of the Gateaux derivative U'. It also deserves to mention that the Hamiltonian operators defined by (1) can lead to other hierarchies of systems of evolution equations. For example, a hierarchy of bi-Hamiltonian systems $u_t=\Phi^n u_x$, $n\geqslant 0$, can be generated from a hereditary symmetry operator Φ defined by (17). What is more, we can make another choice of an invertible Hamiltonian operator,

$$J=\begin{pmatrix} 0 & \partial \\ \partial & 0 \end{pmatrix}, \tag{44}$$

which has an inverse operator

$$J^{-1}=\begin{pmatrix} 0 & \partial^{-1} \\ \partial^{-1} & 0 \end{pmatrix}.$$

It constitutes a Hamiltonian pair together with M defined by (1). Thus we can have the corresponding hereditary symmetry operator,

$$\Phi=MJ^{-1}=\begin{pmatrix} \alpha_2\partial^{-1}+\alpha_3-\alpha_1 q\partial^{-1}r\partial^{-1} & \alpha_1 q\partial^{-1}q\partial^{-1} \\ \alpha_1 r\partial^{-1}r\partial^{-1} & -\alpha_2\partial^{-1}+\alpha_3-\alpha_1 r\partial^{-1}q\partial^{-1} \end{pmatrix}. \tag{45}$$

This generates a new hierarchy $u_t=\Phi^n u_x$, $n\geqslant 0$, which is an inverse hierarchy of the Kaup–Newell hierarchy. In conclusion, Hamiltonian operators of the same type may lead to soliton hierarchies of different types.

ACKNOWLEDGMENTS

This work was supported by the City University of Hong Kong, the Research Grants Council of Hong Kong, the National Natural Science Foundation of China, the Chinese National Basic Research Project ‘‘Nonlinear Science,’’ and the Doctoral Program Foundation of Ministry of Higher Education, China.

[1] A. C. Newell, *Solitons in Mathematics and Physics* (SIAM, Philadelphia, 1985).
[2] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[3] A. S. Fokas and B. Fuchssteiner, Phys. Lett. A **86**, 341–345 (1981).
[4] G. Z. Tu, Sci. China, Ser. A **32**, 142–153 (1989).
[5] W. X. Ma, Ann. Math. **18**, 79–88 (1997).
[6] M. Boiti, P. J. Caudrey, and F. Pempinelli, Nuovo Cimento B **83**, 71–87 (1984).
[7] M. Antonowicz and A. P. Fordy, Physica D **28**, 345–357 (1987).
[8] M. Antonowicz and A. P. Fordy, J. Phys. A: Math. Gen. **21**, L269–L275 (1988).
[9] W. X. Ma, J. Phys. A: Math. Gen. **26**, L1169–L1174 (); W. X. Ma and Z. X. Zhou, Prog. Theor. Phys. **96**, 449–457 (1996).
[10] W. X. Ma and B. Fuchssteiner, Phys. Lett. A **213**, 49–55 (1996).
[11] B. C. Zhang, Ann. Diff. Eqns. **13**, 408–418 (1997).
[12] F. Magri, in *Nonlinear Evolution Equations and Dynamical Systems*, Lectures Notes in Physics, edited by M. Boiti, F. Pempinelli, and G. Soliani (Springer-Verlag, Berlin, 1980), Vol. 120, pp. 233–263.
[13] W. X. Ma, J. Phys. A: Math. Gen. **31**, 7585–7591 (1998).
[14] W. X. Ma and M. Pavlov, Phys. Lett. A **246**, 511–522 (1998).
[15] B. Fuchssteiner and A. S. Fokas, Physica D **4**, 47–66 (1981).
[16] W. X. Ma and W. Strampp, Phys. Lett. A **185**, 277–286 (1994).
[17] W. X. Ma, Q. Ding, W. G. Zhang, and B. Q. Lu, Nuovo Cimento B **111**, 1135–1149 (1996).
[18] F. Margi, J. Math. Phys. **19**, 1156–1162 (1978).
[19] I. M. Gel’fand and I. Y. Dorfman, Funct. Anal. Appl. **13**, 248–262 (1979).
[20] W. X. Ma and R. G. Zhou, J. Phys. A: Math. Gen. **32**, L239-L242 (1999).
[21] R. G. Zhou and W. X. Ma, preprint, 1998.

Instability and chaos in spatially homogeneous field theories

Luca Salasnich[a)]
Istituto Nazionale per la Fisica della Materia, Unità di Milano, Dipartimento di Fisica, Università di Milano, Via Celoria 16, 20133 Milano, Italy

(Received 28 July 1998; accepted for publication 2 April 1999)

Spatially homogeneous field theories are studied in the framework of dynamical system theory. In particular, we consider a model of inflationary cosmology and a Yang–Mills–Higgs system. We discuss also the role of quantum chaos and its application to field theories.
[S0022-2488(99)00209-1]

I. INTRODUCTION

Quantum field theory offers a wide variety of applications, in particular for condensed matter[1] and elementary particle physics.[2] Field theoretic ideas also reach for the cosmos through the development of the inflationary scenario—a speculative, but completely physical analysis of the early universe, which appears to be consistent with available observations.[3]

In the last years there has been much interest in the chaotic behavior of field theories.[4–8] In this paper we discuss and extend our recent results on instability and chaos[9–11] in classical and quantum field theory.[12–17] In Sec. II we show how spatially homogeneous field theories can be studied by using the dynamical system theory, and we introduce some basic definitions for the regular and chaotic dynamics of classical and quantum systems. In Sec. III we analyze the local stability of an inflationary scalar field minimally coupled to gravity and its point attractors in the phase space. The value of the scalar field in the vacuum is a bifurcation parameter, and we discuss the existence of a stable limit cycle. Finally, in Sec. IV we study the spatially homogeneous SU(2) Yang–Mills–Higgs system. We show that for this system a classical order–chaos transition occurs both in classical and quantum mechanics.

II. FIELD THEORIES AS DYNAMICAL SYSTEMS

In this section we introduce some basic ideas of the dynamical system theory. We clarify the concept of ergodic system giving a hierarchy of chaos.

Let us consider a classical relativistic scalar field theory with action

$$S[\phi]=\int d^4x\, L(\phi,\partial_\mu\phi), \tag{1}$$

where L is the Lagrangian density of the system, $\partial_\mu=(\partial/\partial t,\boldsymbol{\nabla})$ is the covariant derivative, and $\phi=\phi(x)$ is a real scalar field with $x_\mu=(t,\mathbf{x})$ the space–time position.[2] It is well known that by imposing the Hamilton's Least Action Principle,

$$\delta S[\phi]=0, \tag{2}$$

we obtain the Euler–Lagrange equation of motion of the system

[a)]Electronic mail: salasnich@mi.infm.it

$$\frac{\partial L}{\partial \phi} - \partial_\mu \frac{\partial L}{\partial(\partial_\mu \phi)} = 0. \tag{3}$$

The homogeneous space approximation means that we can neglect the spatial dependence of the field, thus we can perform the following substitution:

$$\phi(t,\mathbf{x}) \to \phi(t), \tag{4}$$

and the resulting equation of motion is given by

$$\frac{\partial L}{\partial \phi} - \frac{d}{dt}\frac{\partial L}{\partial \dot{\phi}} = 0. \tag{5}$$

By introducing the momentum

$$\chi = \frac{\partial L}{\partial \dot{\phi}}, \tag{6}$$

and the Hamiltonian

$$H(\chi,\phi) = \dot{\phi}\chi - L(\phi,\dot{\phi}), \tag{7}$$

the second-order equation of motion can be written as a system of two first-order Hamilton's equations,

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}), \tag{8}$$

where $\mathbf{z} = (\phi,\chi)$ is a point in a two-dimensional phase space and $\mathbf{f} = (f_1, f_2)$ is given by

$$f_1(\phi,\chi) = \frac{\partial H}{\partial \phi}, \quad f_2(\phi,\chi) = -\frac{\partial H}{\partial \chi}. \tag{9}$$

This is a general result: any homogenous field theory can be written as a system of N first-order differential equations, i.e., a dynamical system. In the next sections we shall consider nonconservative and non-Abelian field theories.

A. Dynamical system theory

A dynamical system is defined by N first-order differential equations,

$$\dot{\mathbf{z}}(t) = \mathbf{f}(\mathbf{z}(t),t), \tag{10}$$

where the variables $\mathbf{z} = (z_1,\ldots,z_N)$ are in the phase space Ω (the Euclidean space R^N, unless otherwise specified). These equations describe the time evolution of the variables and the system they represent.[9–11]

A solution of the dynamical system is a vector function $\mathbf{z}(\mathbf{z}_0,t)$, which satisfies (10) and the initial condition

$$\mathbf{z}(\mathbf{z}_0,0) = \mathbf{z}_0 . \tag{11}$$

Usually one writes simply $\mathbf{z}(t)$ without the initial condition dependence.

The time evolution of $\mathbf{z} \in \Omega$ is obtained with the one-parameter group of diffeomorphism $g^t : \Omega \to \Omega$, such that

$$\frac{d}{dt}(g^t\mathbf{z})|_{t=0}=\mathbf{f}(\mathbf{z},0). \tag{12}$$

The group g^t is called phase flux and the solution is called orbit. The system is called Hamiltonian, if the dimension of Ω is even and there exists a function $H(\mathbf{z},t)$, given by

$$\mathbf{f}(\mathbf{z}(t),t)=J\,\nabla H(\mathbf{z},t), \tag{13}$$

where

$$J=\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{14}$$

is the symplectic matrix and $H(\mathbf{z},t)$ is the Hamiltonian function.

On the phase space Ω one usually defines a probability measure $\mu:\Omega\rightarrow\Omega$, such that $\mu(\Omega)=1$. If we choose a subspace A of Ω, the system is measure preserving if

$$\mu(g^tA)=\mu(A). \tag{15}$$

We observe that for measure preserving dynamical systems one gets $\mathrm{div}(\mathbf{f})=0$. It is well known that Hamiltonian systems preserve their measure: the Liouville measure. Dynamical systems that do not preserve their measure are called dissipative, and usually have a measure contraction in time evolution.

The dynamics of a system is called regular if the orbits are stable to infinitesimal variations of initial conditions. It is called chaotic if the orbits are unstable to infinitesimal variations of initial conditions. Useful quantities to calculate this behavior are the Lyapunov exponents, which give the stability of a single orbit.

A vector of the tangent space $T\Omega_{\mathbf{z}}$ to the phase space Ω in the position $\mathbf{z}$ is given by

$$\omega(\mathbf{z})=\lim_{s\rightarrow 0}\frac{\mathbf{r}(s)-\mathbf{r}(0)}{s}, \tag{16}$$

where $\mathbf{r}(0)=\mathbf{z}$ and $\mathbf{r}(s)\in\Omega$. The tangent space vectors are the velocity vectors of the curves on M; there are obviously N independent vectors.

Now we can define the Lyapunov exponent,

$$\lambda(\mathbf{z})=\lim_{t\rightarrow\infty}\frac{1}{t}\ln|\omega(t)|, \tag{17}$$

where $\omega(t)$ is a tangent vector to $\mathbf{z}(t)$ with the condition that $|\omega(0)|=1$.

It can be demonstrated that the limit given by the previous equation exists for a compact phase space, and that it is metric independent. Fixing an orbit in the N-dimensional phase space, there are N distinct exponents $\lambda_1,\ldots,\lambda_N$, called first-order Lyapunov exponents. If the orbit has positive Lyapunov exponents, it is chaotic.

To characterize globally the chaoticity of a system, we can introduce the Kolmogorov–Sinai entropy, which is given by

$$h_{\mathrm{KS}}(\mu)=\int_A d\mu(\mathbf{z})\sum_{\lambda_i>0}\lambda_i(\mathbf{z}), \tag{18}$$

with an A subspace of Ω and λ_i Lyapunov exponents. The Kolmogorov–Sinai entropy is a very useful tool for showing chaotic behavior in the region A.

A system is called ergodic if the time average is equal to the phase space average,

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t dt\, f(g^t\mathbf{z}(t))=\int_\Omega d\mu(\mathbf{z})f(\mathbf{z}). \tag{19}$$

Incidentally, as is well known, Boltzmann started from the ''ergodic hypothesis'' to obtain statistical mechanics of equilibrium. But ergodicity is not sufficient to reach an equilibrium state: one must consider mixing systems.

In a mixing system, every finite element of the phase space occupies for $t\to\infty$ the entire phase space Ω; more precisely: $\forall A, B\subset\Omega$ with $\mu(A)$ and $\mu(B)\neq 0$,

$$\lim_{t\to\infty}\frac{\mu(B\cap g^t A)}{\mu(B)}=\mu(A). \tag{20}$$

To have quantitative information of orbit separations, we must introduce K systems (Kolmogorov), which are mixing systems with a positive metric entropy, i.e., $h_{\mathrm{KS}}>0$. Such systems are typical chaotic systems. Among the K systems, the most unpredictable ones are the B systems (Bernoulli), which have the Kolmogorov–Sinai entropy equal to the entropy of every partition, i.e., $h_{\mathrm{KS}}=h(A_i(0),\mu),\forall A_i(0)$.

B. Hamiltonian dynamics

Let us consider a Hamiltonian system with n degrees of freedom described by the Hamiltonian function $H(\mathbf{z})$, where $\mathbf{z}=(q_1,\ldots,q_n,p_1,\ldots,p_n)$ so that the phase space is $N=2n$ dimensional. The Hamiltonian system is called integrable if there are N functions $F_i=F_i(\mathbf{z})$ defined on Ω in involution:

$$[F_i,F_j]_{\mathrm{PB}}=\sum_{k=1}^{n}\frac{\partial F_i}{\partial q_k}\frac{\partial F_j}{\partial p_k}-\frac{\partial F_j}{\partial q_k}\frac{\partial F_i}{\partial p_k}=0, \quad \forall i,j, \tag{21}$$

and linearly independent. $[,]_{\mathrm{PB}}$ are the Poisson brackets.

For conservative systems we have $F_1=H(\mathbf{z})$, and also

$$\frac{dF_i}{dt}=[H,F_i]_{\mathrm{PB}}=0. \tag{22}$$

Because there are n constants of motion, every orbit can explore only the n-dimensional manifold $\Omega_f=\{\mathbf{z}:F_i(\mathbf{z})=f_i,\, i=1,\ldots,n\}$. If Ω_f is compact and connected, it is equivalent to an n-dimensional torus $T^n=\{(Q_1,\ldots,Q_n)\bmod 2\pi\}$. There are n irreducible and independent circuits γ_i on Ω_f and there exists a canonical transformation $(\mathbf{p},\mathbf{q})\to(\mathbf{P},\mathbf{Q})$, generated by the function $S(\mathbf{q},\mathbf{P})$, such that

$$P_i=\oint_{\gamma i} d\mathbf{q}\cdot\mathbf{p}, \quad Q_i=\frac{\partial S}{\partial P_i}. \tag{23}$$

The P_i are called action variables and the Q_i are called angle variables. The moments $\mathbf{p}$ and coordinates $\mathbf{q}$ are periodic functions of $\mathbf{Q}$ with period 2π. The Hamiltonian depends only on action variables, i.e., $H=H(\mathbf{P})$.

Adding a small perturbation $V(\mathbf{P},\mathbf{Q})$ to an integrable Hamiltonian $H_0(\mathbf{P})$, the total Hamiltonian can be written as

$$H(\mathbf{P},\mathbf{Q})=H_0(\mathbf{P})+gV(\mathbf{P},\mathbf{Q}), \tag{24}$$

and, generically, the integrability is destroyed. As a consequence, parts of phase space become filled with chaotic orbits, while in other parts the toroidal surfaces of the integrable system are deformed but not destroyed; thus we have a quasi-integrable system. By growing g, chaotic motion develops near the regions of phase space, where all the frequencies on the torus $\omega_i = \partial H(\mathbf{P})/\partial P$ are commensurate. Conversely, tori of the integrable system, on which the ω_i are incommensurate, are deformed, but not destroyed immediately [Kolmogorov–Arnold–Moser (KAM) theorem].[9,18] As g increases, the phase space generically develops a highly complex structure, with islands of regular motion (filled with quasiperiodic orbits) interspersed in regions of chaotic motion, but containing in turn more regions of chaos. As g grows further, the fraction of phase space filled with chaotic orbits grows until it reaches unity as the last KAM surface is destroyed. Then the motion is completely chaotic everywhere, except possibly for isolated periodic orbits.[9,18]

It is very useful to plot a $2n-1$ surface of section $\mathcal{P}\subset\Omega$, called a Poincaré section. For an integrable system with two degrees of freedom, the $q_1=0$ Poincaré section of a rational (resonant) torus is a finite number of points along a closed curve, while the section of an irrational (nonresonant) torus is a continuous closed curve. Adding a perturbation, in the section we present closed curves (KAM tori), whose points are stable (elliptic), and also curves formed by substructures, residua of resonant tori, whose points are unstable (hyperbolic). As the perturbation parameter increases, the closed curves are distorted and reduced in number.

C. Quantum chaos

We use the term quantum chaotic system in the precise and restricted sense of a quantum system whose classical analog is chaotic. In particular, we concentrate on energy levels of quantum systems (see, for example, Refs. 18 and 19).

Let us consider a classical regular Hamiltonian system. The short-range properties of the corresponding quantal spectrum tend to resemble those of a spectrum of randomly distributed numbers. This is because regular classical motion is associated with integrability or separability of the classical equations of motion. In quantum mechanics the separability corresponds to a number of independent conserved quantities (such as angular momentum), and each energy level can be characterized by the associated quantum numbers. Superimposing the terms arising from the various quantum numbers, a spectrum is generated like that of random numbers, at least over short intervals. In particular, the distribution $P(s)$ of nearest-neighbor spacings $s_i=(\epsilon_{i+1}-\epsilon_i)/d$, where d is the mean level spacing, is expected to follow the Poisson limit, i.e., $P(s)=\exp(-s)$.

Instead, when the classical dynamics of a physical system is chaotic, the system cannot be integrable, and there must be fewer constants of motion than degrees of freedom. Quantum mechanically, this means that once all good quantum numbers due to obvious symmetries, etc., are accounted for, the energy levels cannot simply be labeled by quantum numbers associated with certain constants of motion. The short-range properties of the energy spectrum then tend to resemble those of eigenvalue spectra of matrices with randomly chosen elements, and one gets a result very close to $P(s)=(\pi/2)s\exp[-(\pi/4)s^2]$, which is the so-called Wigner distribution.

The distribution $P(s)$ is the best spectral statistics to analyze a shorter series of energy levels and the intermediate regions between order and chaos. This distribution can be compared to the Brody distribution,

$$P(s,\omega)=\alpha(\omega+1)s^{\omega}\exp(-\alpha s^{\omega+1}), \tag{25}$$

with

$$\alpha=\left(\Gamma\left[\frac{\omega+2}{\omega+1}\right]\right)^{\omega+1}. \tag{26}$$

The Brody distribution interpolates between the Poisson distribution ($\omega=0$) of integrable systems and the Wigner distribution ($\omega=1$) of chaotic ones, and thus the parameter ω can be used as a simple quantitative measure of the degree of chaoticity.

III. A MODEL FOR INFLATIONARY COSMOLOGY

In this section we study the stability of a scaler inflation field[12,13] and analyze its bifurcation properties in the framework of the dynamical system theory.

It is generally believed that the universe, at a very early stage after the big bang, exhibited a short period of exponential expansion, the so-called inflationary phase. In fact, the assumption of an inflationary universe solves three major cosmological problems: the flatness problem, the homogeneity problem, and the formation of a structure problem.[3]

The Friedmann–Robertson–Walker metric of a homogeneous and isotropic expanding universe is given by

$$ds^2 = dt^2 - a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta\, d\varphi^2)\right], \tag{27}$$

where $k=1$, -1, or 0 for a closed, open, or flat universe, and $a(t)$ is the scale factor of the universe.

The evolution of the scale factor $a(t)$ is given by the Einstein equations

$$\ddot{a} = -\frac{4\pi}{3}G(\rho+3p)a, \quad \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi}{3}G\rho, \tag{28}$$

where ρ is the energy density of matter in the universe, and p its pressure. The gravitational constant $G=M_p^{-2}$ (with $\hbar=c=1$), where $M_p=1.2\times 10^{19}$ GeV is the Plank mass, and $H_u=\dot{a}/a$ is the Hubble "constant," which, in general, is a function of time.

The inflationary models postulate the existence of a scalar field ϕ, the so-called inflation field, with Lagrangian

$$L = \tfrac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - V(\phi), \tag{29}$$

where the potential $V(\phi)$ depends on the type of inflation model considered. The scalar field, if minimally coupled to gravity, satisfies the equation

$$\Box\phi = \ddot{\phi} + 3\left(\frac{\dot{a}}{a}\right)\dot{\phi} - \frac{1}{a^2}\nabla^2\phi = -\frac{\partial V}{\partial\phi}, \tag{30}$$

where $\Box$ is the covariant d'Alembertian operator. The Hubble "constant" H_u is related to the energy density of the field by

$$H_u^2 + \frac{k}{a^2} = \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\left[\frac{\dot{\phi}^2}{2} + \frac{(\nabla\phi)^2}{2} + V(\phi)\right]. \tag{31}$$

In a flat universe $k=0$ and, if the inflation field is sufficiently uniform [i.e., $\dot{\phi}^2, (\nabla\phi)^2 \ll V(\phi)$], we obtain a homogeneous field theory in one dimension,

$$\ddot{\phi} + 3H_u(\phi)\dot{\phi} + \frac{\partial V}{\partial\phi} = 0, \tag{32}$$

where the Hubble "constant" H_u is an explicit function of ϕ:

$$H_u^2 = \frac{8\pi G}{3}V(\phi). \tag{33}$$

A. Local instability for the inflationary self-energy

The second-order equation of motion of our cosmological model can be written as a system of two first-order differential equations,

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}), \tag{34}$$

where $\mathbf{z} = (\phi, \chi)$ is a point in the two-dimensional phase space and $\mathbf{f} = (f_1, f_2)$ is given by

$$f_1(\phi,\chi) = \chi, \quad f_2(\phi,\chi) = -3H_u(\phi)\chi - \frac{\partial V(\phi)}{\partial \phi}. \tag{35}$$

The system is nonconservative, because the function

$$\mathrm{div}(\mathbf{f}) = \frac{\partial g_1}{\partial \phi} + \frac{\partial g_2}{\partial \chi} = -3H_u(\phi), \tag{36}$$

is not identically zero. The fixed points of the system are those for which $f_1(\phi,\chi) = 0$ and $f_2(\phi,\chi) = 0$, i.e.,

$$\chi = 0, \quad \frac{\partial V(\phi)}{\partial \phi} = 0. \tag{37}$$

The deviation $\delta\mathbf{z}(t) = \hat{\mathbf{z}}(t) - \mathbf{z}(t)$ from the two initially neighboring trajectories $\mathbf{x}$ and $\hat{\mathbf{x}}$ in the phase space satisfies the linearized equations of motion,

$$\frac{d}{dt}\delta\mathbf{z}(t) = \Gamma(t)\,\delta\mathbf{z}, \tag{38}$$

where $\Gamma(t)$ is the stability matrix

$$\Gamma(t) = \begin{pmatrix} 0 & 1 \\ -\dfrac{\partial^2 V}{\partial \phi^2} - 3\chi \dfrac{\partial H}{\partial \phi} & -3H_u(\phi) \end{pmatrix}. \tag{39}$$

At least if an eigenvalue of $\Gamma(t)$ is real, the separation of the trajectories grows exponentially and the motion is unstable. Imaginary eigenvalues correspond to stable motion. In the limit of time that goes to infinity, from the eigenvalues of the stability matrix, we can obtain the Lyapunov exponents. For a two-dimensional dynamical system, the Lyapunov exponents cannot be positive[9] and so the system is not chaotic, i.e., there is not global instability. However, we can be assured that the universe is crowded with many interacting fields of which the inflation is but one. The nonlinear nature of these interactions can result in a complex chaotic evolution of the universe and the local instability of the inflation field is a precursor phenomenon of chaotic motion.

The eigenvalues of the stability matrix are given by

$$\sigma_{1,2} = -\frac{3}{2}H_u(\phi) \pm \frac{1}{2}\sqrt{9H_u^2(\phi) - 4\frac{\partial^2 V}{\partial \phi^2} - 12\chi\frac{\partial H_u}{\partial \phi}}. \tag{40}$$

The pair of eigenvalues become real and there is exponential separation of neighboring trajectories, i.e., unstable motion, if

$$\frac{\partial^2 V}{\partial \phi^2} + 3\chi\frac{\partial H_u}{\partial \phi} < 0. \tag{41}$$

Particularly when $\chi = 0$, e.g., the fixed points, we obtain local instability when

$$\frac{\partial^2 V}{\partial \phi^2} < 0, \tag{42}$$

i.e., for negative curvature of the potential energy. The fixed points are stable if they are a point of the local minimum of $V(\phi)$ and unstable if they are points of the local maximum.

The potential $V(\phi)$ depends on the type of inflation model considered, and it is usually some kind of double-well potential. We choose a symmetric double-well potential,

$$V(\phi)=\frac{\lambda}{4}(\phi^2-v^2)^2, \tag{43}$$

where $\pm v$ are the values of the inflaton field in the vacuum, i.e., the points of minimal energy of the system.

We observe that the inflation field value in the vacuum v is a bifurcation parameter. Bifurcation is used to indicate a qualitative change in the features of the system under the variation of one or more parameters on which the system depends. First of all we consider the case $v=0$, i.e., $V(\phi)=(\lambda/4)\phi^4$. In this situation there is only one fixed point $(\phi^*=0,\chi^*=0)$ which is a stable one being

$$\frac{\partial^2 V}{\partial \phi^2}=3\lambda\phi^2\geqslant 0. \tag{44}$$

The fixed point $(\phi^*=0,\chi^*=0)$ is a point attractor.

Instead, for $v\neq 0$, there are three fixed points:

$$(\phi^*=0,\chi^*=0),\quad (\phi^*=v,\chi^*=0),\quad (\phi^*=-v,\chi^*=0), \tag{45}$$

and the condition for the instability becomes

$$-\frac{v}{\sqrt{3}}<\phi<\frac{v}{\sqrt{3}}. \tag{46}$$

Obviously $(\phi^*=0,\chi^*=0)$ is an unstable fixed point, and, in particular, a saddle point because the stability matrix has real and opposite eigenvalues. On the other hand $(\phi^*=\pm v,\chi^*=0)$ are stable fixed points.

There are four possible functions for the Hubble ''constant:''

$$H_u(\phi)=\pm\gamma|\phi^2-v^2|, \tag{47}$$

but also

$$H_u(\phi)=\pm\gamma(\phi^2-v^2), \tag{48}$$

where $\gamma=\sqrt{2\pi G\lambda/3}$ is the dissipation parameter. The choice of the Hubble function is crucial for the dynamical evolution of the system.

In certain nonconservative systems, we could find closed trajectories or limit cycles toward which the neighboring trajectories spiral on both sides. It is sometimes possible to know that no limit cycle exists and the Bendixson criterion,[20] which establishes a condition for the nonexistence of closed trajectories, is useful in some cases. Bendixson criterion is as follows: if div(**f**) is not zero and does not change its sign within a domain D of the phase space, no closed trajectories can exist in that domain. In our case we have $\text{div}(\mathbf{f})=-3H_u(\phi)$, and so the presence of periodic orbit is related to the sign of $H_u(\phi)$.

If $H_u(\phi)=\gamma|\phi^2-v^2|$ we do not find periodic orbits and the inflation field goes to one of its two stable fixed points, which are points attractors (see Fig. 1). The vacuum is degenerate, but if we choose an initial condition around the saddle point, there is a dynamical symmetry breaking toward the positive v or negative $-v$ value of the inflation field in the vacuum. This symmetry breaking is unstable because neighbor initial conditions can go in different point attractors.

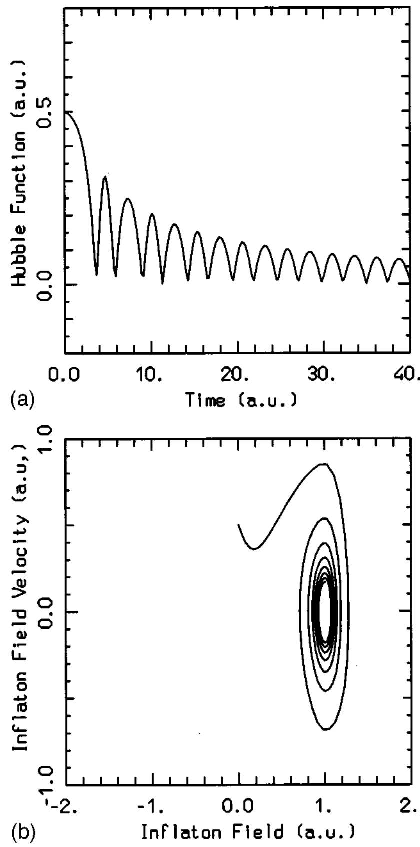

FIG. 1. The Hubble function versus time (top) and the phase space trajectory of the inflation field (bottom); for $H_u(\phi)=\gamma|\phi^2-v^2|$ with $\gamma=\frac{1}{2}$, $\lambda=3$, and $v=1$. Initial conditions: $\phi=0$ and $\dot{\phi}=\frac{1}{2}$.

Instead, if we choose $H_u(\phi)=\gamma(\phi^2-v^2)$, the numerical calculations of Fig. 2 show that a limit cycle exists, the two stable fixed points are not point attractors, and the inflaton field oscillates forever. Obviously the larger v is the larger the limit cycle.

B. A limit cycle in the cosmological model

The equation of motion of the inflation field with $H_u(\phi)=\gamma(\phi^2-v^2)$ reads as

$$\ddot{\phi}+3\gamma(\phi^2-v^2)\dot{\phi}+\lambda\phi(\phi^2-v^2)=0. \tag{49}$$

This equation can be written as

$$\frac{d}{dt}\left[\dot{\phi}+3\gamma\int_0^{\phi}(u^2-v^2)du\right]+\lambda\phi(\phi^2-v^2)=0, \tag{50}$$

and if we put

$$F(\phi)=3\int_0^{\phi}(u^2-v^2)du=\phi(\phi^2-3v^2), \quad G(\phi)=\phi(\phi^2-v^2), \tag{51}$$

and also $\omega=\dot{\phi}+\gamma F(\phi)$, we obtain the system

$$\dot{\phi}=\omega-\gamma F(\phi), \quad \dot{\omega}=-\lambda G(\phi). \tag{52}$$

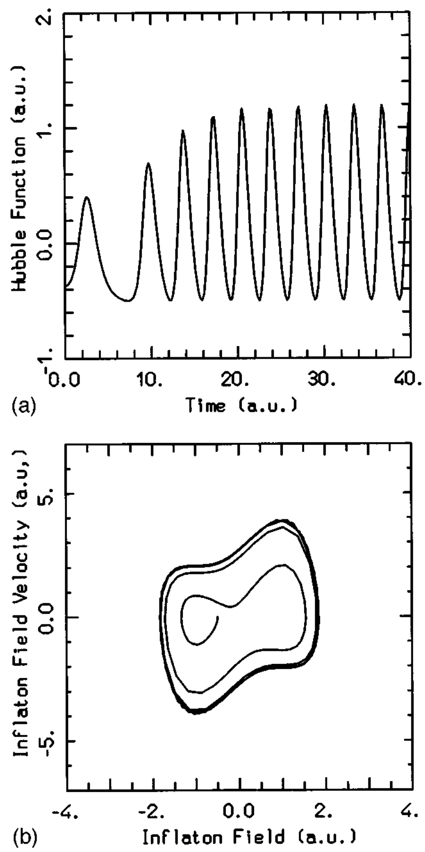

FIG. 2. The Hubble function versus time (top) and the phase space trajectory of the inflation field (bottom); for $H_u(\phi) = \gamma(\phi^2 - v^2)$ with $\gamma = \frac{1}{2}$, $\lambda = 3$, and $v = 1$. Initial conditions: $\phi = -\frac{1}{2}$ and $\dot{\phi} = 0$.

For systems of this kind, the Lienard theorem[21] states that there is a unique and stable limit cycle if the following conditions are satisfied: $F(\phi)$ is an odd function and $F(\phi) = 0$ only at $\phi = 0$ and $\phi = \pm\alpha$; $F(\phi) < 0$ for $0 < \phi < \alpha$, $F(\phi) > 0$ and is increasing for $\phi > \alpha$; $G(\phi)$ is an odd function and $\phi G(\phi) > 0$ for all $\phi > \alpha$. It is easy to check that the functions $F(\phi)$ and $G(\phi)$ satisfy all the conditions of the Lienard theorem with $\alpha = v$. The cubic force $G(\phi)$ tends to reduce any displacement for large $|\phi|$, whereas the damping $F(\phi)$ is negative at small $|\phi|$ and positive at large $|\phi|$. Since small oscillations are pumped up and large oscillations are damped down, it is not surprising that the system tends to settle into a self-sustained oscillation of some intermediate amplitude.

Let us consider a typical trajectory of the system. After the scaling $\psi = \lambda\omega$, we obtain

$$\dot{\phi} = \lambda\left[\psi - \frac{\gamma}{\lambda}F(\phi)\right], \quad \dot{\psi} = -G(\phi). \tag{53}$$

The cubic nullcline $\psi = (\gamma/\lambda)F(\phi)$ is the key to understanding the motion. Suppose that $\lambda \gg 1$ and the initial condition is far from the cubic nullcline, then we have $|\dot{\phi}| \sim O(\lambda) \gg 1$; hence, the velocity is enormous in the horizontal direction and tiny in the vertical direction, so trajectories move practically horizontally. If the initial condition is above the nullcline, then $\dot{\phi} > 0$; thus the trajectory moves sideways toward the nullcline. However, once the trajectory gets so close that $\psi \simeq (\lambda/\gamma)F(\phi)$ then the trajectory crosses the nullcline vertically and moves slowly along the backside of the branch until it reaches the knee and can jump sideways again. The period T of the limit cycle is essentially the time required to travel along the two slow branches, since the time spent in the jumps is negligible for large λ. By symmetry, the time spent on each branch is the same, so we have

$$T \simeq 2 \int_{t_A}^{t_B} dt, \tag{54}$$

where A and B are the initial and final points on the positive slow branch. To derive an expression for dt, we note that on the slow branches with a good approximation $\psi \simeq (\gamma/\lambda) F(\phi)$, and thus

$$\frac{d\psi}{dt} \simeq \frac{\gamma}{\lambda} F'(\phi) \frac{d\phi}{dt} = 3 \frac{\gamma}{\lambda} (\phi^2 - v^2) \frac{d\phi}{dt}. \tag{55}$$

Since $d\psi/dt = -\phi(\phi^2 - v^2)$, we obtain $dt \simeq -3(\gamma/\lambda)(d\phi/\phi)$, on the slow branches. The slow positive branch begins at $\phi_A = 2\gamma v/\lambda$ and ends at $\phi_B = \gamma v/\lambda$. Because $\gamma = \sqrt{2\pi G \lambda/3}$ we get $T \simeq 2 \ln 2 \sqrt{6\pi G/\lambda}$.

IV. THE HOMOGENEOUS SU(2) YMH SYSTEM

Now we study the suppression of classical chaos in the spatially homogeneous SU(2) Yang–Mills–Higgs (YMH) system induced by the Higgs field.[4,14,15] We analyze also the energy fluctuation properties of the system, which give a clear quantum signature of the classical chaos–order transition of the system.

The SU(2) YMH system describes the interaction between a scalar Higgs field ϕ and three non-Abelian Yang–Mills fields A_μ^a, $a = 1,2,3$. The Lagrangian density of the YMH system is given by

$$L = \tfrac{1}{2}(D_\mu \phi)^+ (D^\mu \phi) - V(\phi) - \tfrac{1}{4} F_{\mu\nu}^a F^{\mu\nu a}, \tag{56}$$

where

$$(D_\mu \phi) = \partial_\mu \phi - i g A_\mu^b T^b \phi, \tag{57}$$

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + g \epsilon^{abc} A_\mu^b A_\nu^c, \tag{58}$$

with $T^b = \sigma^b/2$, $b = 1,2,3$, generators of the SU(2) algebra, and where the potential of the scalar field (the Higgs field) is

$$V(\phi) = \mu^2 |\phi|^2 + \lambda |\phi|^4. \tag{59}$$

We work in the $(2+1)$-dimensional Minkowski space ($\mu = 0,1,2$) and choose spatially homogeneous Yang–Mills and the Higgs fields,

$$\partial_i A_\mu^a = \partial_i \phi = 0, \quad i = 1,2, \tag{60}$$

i.e., we consider the system in the region in which space fluctuations of fields are negligible compared to their time fluctuations.

In the gauge $A_0^a = 0$ and using the real triplet representation for the Higgs field, we obtain

$$L = \dot{\phi}^2 + \tfrac{1}{2}(\dot{A}_1^2 + \dot{A}_2^2) - g^2 [\tfrac{1}{2} A_1^2 A_2^2 - \tfrac{1}{2}(A_1 \cdot A_2)^2 + (A_1^2 + A_2^2)\phi^2 - (A_1 \cdot \phi)^2 - (A_2 \cdot \phi)^2] - V(\phi), \tag{61}$$

where $\phi = (\phi^1, \phi^2, \phi^3)$, $A_1 = (A_1^1, A_1^2, A_1^3)$, and $A_2 = (A_2^1, A_2^2, A_2^3)$.

When $\mu^2 > 0$, the potential V has a minimum at $|\phi| = 0$, but for $\mu^2 < 0$ the minimum is at

$$|\phi_0| = \sqrt{\frac{-\mu^2}{4\lambda}} = v,$$

which is the nonzero Higgs vacuum. This vacuum is degenerate and after spontaneous symmetry breaking the physical vacuum can be chosen as $\phi_0=(0,0,v)$. If $A_1^1=q_1$, $A_2^2=q_2$ and the other components of the Yang–Mills fields are zero, in the Higgs vacuum the Hamiltonian of the system reads as

$$H=\tfrac{1}{2}(p_1^2+p_2^2)+g^2v^2(q_1^2+q_2^2)+\tfrac{1}{2}g^2q_1^2q_2^2, \tag{62}$$

where $p_1=\dot{q}_1$ and $p_2=\dot{q}_2$. Here $w^2=2g^2v^2$ is the mass term of the Yang–Mills fields. This YMH Hamiltonian is a toy model for classical nonlinear dynamics, with the attractive feature that the model emerges from particle physics.

A. From chaos to order in the YMH system

The chaotic behavior of the YMH system can be studied by using the Toda criterion of the Gaussian curvature of the potential energy.[22,23] For our YMH system the potential energy is given by

$$V(q_1,q_2)=g^2v^2(q_1^2+q_2^2)+\tfrac{1}{2}g^2q_1^2q_2^2. \tag{63}$$

At low energy, the motion near the minimum of the potential, where the Gaussian curvature is positive, is periodic or quasiperiodic and is separated from the instability region by a line of zero curvature; if the energy is increased, the system will be, for some initial conditions, in a region of negative curvature, where the motion is chaotic. According to this scenario, the energy E_c of the chaos–order transition is equal to the minimum value of the line of zero Gaussian curvature $K_{\mathrm{G}}(q_1,q_2)$ on the potential-energy surface. For our potential the Gaussian curvature vanishes at the points that satisfy the equation,

$$(2g^2v^2+g^2q_2^2)(2g^2v^2+g^2q_1^2)-4g^4q_1^2q_2^2=0. \tag{64}$$

It is easy to show that the minimal energy on the zero-curvature line is given by

$$E_c=V_{\mathrm{min}}(K_{\mathrm{G}}=0,\overline{q_1})=6g^2v^4, \tag{65}$$

and by inverting this equation we obtain $v_c=(E/6g^2)^{1/4}$. Thus the curvature criterion suggest that there is a order–chaos transition by increasing the energy E of the system and a chaos–order transition by increasing the value v of the Higgs field in the vacuum. Thus, there is only one transition regulated by the parameter $E/(g^2v^4)$.

It is important to stress that the Toda criterion is not a fully reliable indicator of chaos.[23] In fact, the local instability of the Toda Criterion does not necessarily imply the global one, and the idea of an order–chaos transition with a critical energy is not strictly correct. The Toda curvature criterion should therefore be combined with the Poincaré sections, which are shown in Fig. 3. The numerical results confirm the analytical predictions of the curvature criterion: with $E=10$ and $g=1$ we get the critical value of the onset of chaos $v_c=(E/6g^2)^{1/4}\simeq 1.14$.

B. Spectral statistics of the YMH system

In quantum mechanics the generalized coordinates of the YMH system satisfy the usual commutation rules $[\hat{q}_k,\hat{p}_l]=i\delta_{kl}$, with $k,l=1,2$. Introducing the creation and destruction operators,

$$\hat{a}_k=\sqrt{\frac{\omega}{2}}\,\hat{q}_k+i\sqrt{\frac{1}{2\omega}}\,\hat{p}_k, \quad \hat{a}_k^+=\sqrt{\frac{\omega}{2}}\,\hat{q}_k-i\sqrt{\frac{1}{2\omega}}\,\hat{p}_k, \tag{66}$$

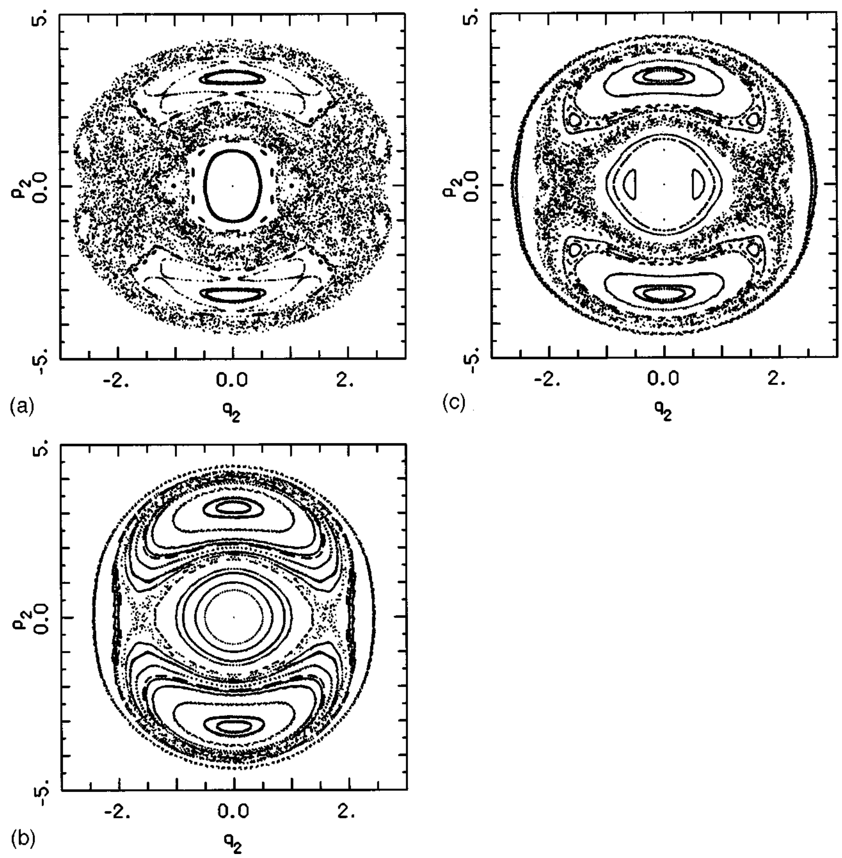

FIG. 3. The Poincarè sections of the YMH system. From the top: $v=1$, $v=1.1$, and $v=1.2$. Energy $E=10$ and interaction $g=1$.

the quantum YMH Hamiltonian can be written as

$$\hat{H}=\hat{H}_0+\tfrac{1}{2}g^2\hat{V}, \tag{67}$$

where

$$\hat{H}_0=\omega(\hat{a}_1^+\hat{a}_1+\hat{a}_2^+\hat{a}_2+1), \tag{68}$$

$$\hat{V}=\frac{1}{4\omega^2}(\hat{a}_1+\hat{a}_1^+)^2(\hat{a}_2+\hat{a}_2^+)^2, \tag{69}$$

with $\omega^2=2g^2v^2$ and $[\hat{a}_k,\hat{a}_l^+]=\delta_{kl}$, $k,l=1,2$.

We compute the energy levels of the YMH system with a numerical diagonalization of the truncated matrix of the quantum YMH Hamiltonian in the basis of the harmonic oscillators (see also Refs. 16 and 17). If $|n_1n_2\rangle$ is the basis of the occupation numbers of the two harmonic oscillators, the matrix elements are

$$\langle n_1'n_2'|\hat{H}_0|n_1n_2\rangle=\omega(n_1+n_2+1)\,\delta_{n_1'n_1}\delta_{n_2'n_2}, \tag{70}$$

and

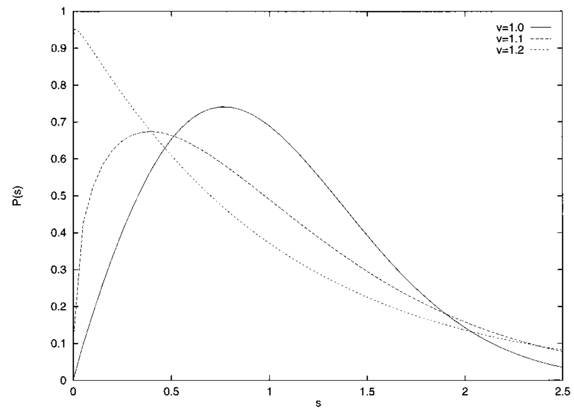

FIG. 4. The $P(s)$ distribution of Brody of the YMH system. The first 100 energy levels and $g=1$. The best fit Brody parameter is given by $\omega=0.92$ for $v=1.0$, $\omega=0.34$ for $v=1.1$, and $\omega=0.01$ for $v=1.2$.

$$\langle n_1' n_2' | \hat{V} | n_1 n_2 \rangle = \frac{1}{4\omega^2} [\sqrt{n_1(n_1-1)}\,\delta_{n_1' n_1 -2} + \sqrt{(n_1+1)(n_1+2)}\,\delta_{n_1' n_1+2} + (2n_1+1)\,\delta_{n_1' n_1}] \times [\sqrt{n_2(n_2-1)}\,\delta_{n_2' n_2-2} + \sqrt{(n_2+1)(n_2+2)}\,\delta_{n_2' n_2+2} + (2n_2+1)\,\delta_{n_2' n_2}]. \quad (71)$$

The symmetry of the potential enables us to split the Hamiltonian matrix into four submatrices reducing the computer storage required. These submatrices are related to the parity of the two occupation numbers n_1 and n_2: even–even, odd–odd, even–odd, and odd–even. The numerical energy levels depend on the dimension of the truncated matrix: we compute the numerical levels in double precision increasing the matrix dimension until the first 100 levels converge within eight digits (matrix dimension 1156×1156).

We have seen previously that the most used quantity to study the local fluctuations of the energy levels is the distribution $P(s)$ of nearest-neighbor spacings s_i of the energy levels. It is obtained by accumulating the number of spacings that lie within the bin $(s, s+\Delta s)$ and then normalizing $P(s)$ to unit.

We use the first 100 energy levels of the four submatrices to calculate the $P(s)$ distribution. In order to remove the secular variation of the level density as a function of the energy E, for each value of the coupling constant the corresponding spectrum is mapped into one that has a constant level density.

Figure 4 shows the $P(s)$ distribution of Brody for three different values of the Higgs vacuum v. The best fit Brody parameter ω is obtained by using the nearest-neighbor spacings of the first 100 unfolded energy levels of the YMH system. There is a Wigner–Poisson transition by increasing the value v of the Higgs field in the vacuum. Thus, by using the $P(s)$ distribution, it is possible to give a quantitative measure of the degree of quantal chaoticity of the system. Our numerical calculations show clearly the quantum chaos–order transition and its correspondence to the classical one.

V. CONCLUSIONS

We have seen that spatially homogeneous field theories can be studied as dynamical systems. After a brief review of the dynamical system theory, we have discussed two schematic models of field theory.

First, we have considered the stability of a nonconservative scalar inflation field. The value of the inflation field in the vacuum is a bifurcation parameter that changes dramatically the phase space structure. The main point is that for some functional solutions of the Hubble ''constant'' the system goes to a limit cycle, i.e., to a periodic orbit. The inflation field is not chaotic, but its local instability can give rise to a complex chaotic evolution of the universe due to its nonlinear interactions with other fields. In the future it will be very interesting to study these effects, which can perhaps lead to some observable implications like a fractal pattern in the spectrum of density fluctuations.

We have then analyzed the non-Abelian SU(2) Yang–Mills–Higgs system. We have given an analytical estimation (confirmed by numerical results of Poincaré sections) of the classical chaos–order transition as a function of the Higgs vacuum, the Yang–Mills coupling constant, and the energy of the system. A quantum signature of a chaos–order transition has been obtained by using the distribution $P(s)$ of nearest-neighbor spacings. The Wigner–Poisson transition of the $P(s)$ distribution follows very well the classical results of the Poincaré sections.

To conclude, we observe that there are yet many open problems about chaos in field theory. We make a list of some of them: (i) spatial chaos and space–time chaos; (ii) classical and quantum chaos in more realistic systems, for example, in QCD (some results can be found in Refs. 7 and 8); (iii) the connection between chaos and critical phenomena (finite temperature field theory).

[1] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971).
[2] C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1985).
[3] A. D. Linde, *Particle Physics and Inflationary Cosmology* (Harwood Academic, London, 1988).
[4] G. K. Savvidy, Phys. Lett. B **130**, 303 (1983); **159**, 325 (1985); Nucl. Phys. B **246**, 302 (1984).
[5] T. Kawabe and S. Ohta, Phys. Rev. D **44**, 1274 (1991); Phys. Lett. B **334**, 127 (1994); T. Kawabe, *ibid.* **343**, 254 (1995).
[6] M. S. Sriram, C. Mukku, S. Lakshmibala, and B. A. Bambah, Phys. Rev. D **49**, 4246 (1994); J. Segar and M. S. Sriram, *ibid.* **53**, 3976 (1996).
[7] M. A. Halasz and J. J. M. Verbaarschot, Phys. Rev. Lett. **74**, 3920 (1995).
[8] S. G. Matinyan and B. Muller, Phys. Rev. Lett. **78**, 2515 (1997).
[9] A. H. Nayfeh and B. Balachandran, *Applied Nonlinear Dynamics* (Wiley, New York, 1995).
[10] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer-Verlag, New York, 1983).
[11] D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations* (Oxford University Press, Oxford, 1987).
[12] L. Salasnich, Mod. Phys. Lett. A **10**, 3119 (1995).
[13] L. Salasnich, Nuovo Cimento B **112**, 873 (1997).
[14] L. Salasnich, Phys. Rev. D **52**, 6189 (1995).
[15] L. Salasnich, in *Perspectives on Theoretical Nuclear Physics*, edited by A. Fabrocini *et al.* (Edizioni ETS, Pisa, 1996), Vol. 6, pp. 261–268.
[16] L. Salasnich, Mod. Phys. Lett. A **12**, 1473 (1997).
[17] L. Salasnich, Phys. Atom. Nucl. **61**, 1878 (1998).
[18] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer-Verlag, Berlin, 1990).
[19] G. Casati and B. V. Chirikov, *Quantum Chaos* (Cambridge University Press, Cambridge, 1995).
[20] I. Bendixson, Acta Math. **24**, 1 (1901).
[21] A. Lienard, Rev. Gen. Electr. **23**, 901 (1928).
[22] M. Toda, Phys. Lett. A **48**, 335 (1974).
[23] G. Benettin, R. Brambilla, and L. Galgani, Physica A **87**, 381 (1997).

Dimension of the global attractor for damped semilinear wave equations with critical exponent

Zhou Shengfan[a)]
Department of Mathematics, Sichuan University, Chengdu 610064, People's Republic of China

(Received 28 January 1999; accepted for publication 21 April 1999)

We obtain an estimate of the upper bound of the Hausdorff dimension of the global attractor for damped semilinear wave equations with a critical exponent. The obtained Hausdorff dimension decreases as the damping grows for large damping. © *1999 American Institute of Physics.* [S0022-2488(99)01509-1]

I. INTRODUCTION

Our aim in this paper is to estimate the upper bound of the Hausdorff dimension of the attractor for damped semilinear wave equations with a homogeneous Dirichlet boundary condition when the nonlinearity satisfies the critical growth condition. Let Ω be an open bounded set of R^3 with a smooth boundary $\partial\Omega$, we consider the initial-boundary value problem of the equation

$$u_{tt}+\alpha u_t-\Delta u+f(u)=g, \quad x\in\Omega, \quad t>0, \tag{1.1}$$

$$u(x,t)|_{x\in\partial\Omega}=0, \quad t>0, \tag{1.2}$$

$$u(x,0)=u_0(x), \quad \frac{\partial u}{\partial t}(x,0)=u_1(x), \quad x\in\Omega, \tag{1.3}$$

where $u=u(x,t)$ is a real-valued function on $\Omega\times[0,+\infty)$, $g\in L^2(\Omega)$, $\alpha>0$, $f(u)=f_1(u)+f_2(u)\in C^1(R;R)$.

Let $G_i(s)=\int_0^s f_i(r)dr$, $i=1,2$. We make the following assumptions on functions $G_i(s)$, $f_i(s)$, $i=1,2$.

(i)

$$f_1(s)s\geqslant 0, \quad \lim_{|s|\to+\infty}\inf\frac{G_2(s)}{s^2}\geqslant 0, \quad \forall s\in R. \tag{1.4}$$

(ii) There exist constants $c_{1i}>0$, $c_{2i}>0$, $i=1,2$ such that

$$\lim_{|s|\to+\infty}\inf\frac{sf_i(s)-c_{1i}G(s)}{s^2}\geqslant 0, \quad i=1,2 \quad \forall s\in R, \tag{1.5}$$

$$|f_1'(s)|\leqslant c_{21}(1+|s|^2), \quad |f_2'(s)|\leqslant c_{22}(1+|s|^p), \quad \text{with } 0\leqslant p<2, \quad \forall s\in R. \tag{1.6}$$

(iii) For every $M>0$, there exists $c_3=c_3(M)$ such that

$$\|f'(u_1)-f'(u_2)\|_{L(H_0^1(\Omega),L^2(\Omega))}\leqslant c_3\|u_1-u_2\|^{\delta_1}, \quad \forall u_1,u_2\in H_0^1(\Omega), \quad \|u_1\|\leqslant M, \quad \|u_2\|\leqslant M, \tag{1.7}$$

[a)]Electronic mail: nic2601@scu.edu.cn

where $\delta_1>0$, $\|\cdot\|$ and $\|\cdot\|_{L(H_0^1(\Omega),L^2(\Omega))}$ denote the norms of $H_0^1(\Omega)$ and $L(H_0^1(\Omega),L^2(\Omega))$ [the space of linear continuous operators from $H_0^1(\Omega)$ into $L^2(\Omega)$], respectively.

The existence of the global attractor for (1.1)–(1.3) with the critical exponent has been studied by many authors,[1–7] of those, Yu[7] obtained an estimate of the upper bound of the Hausdorff dimension of attractors, but this upper bound of the Hausdorff dimension is directly proportional to the coefficient α of damping for $\alpha\geqslant\sqrt{2\lambda_1}$ and tends to infinity as $\alpha\to+\infty$, which does not conform to the physics. In the noncritical case, the author[8] showed the boundedness of the dimension of an attractor for large damping. In this article, we obtain an upper bound of the Hausdorff dimension of attractor of (1.1) for large damping when the functions $f(u)$ satisfies the conditions (1.4)–(1.7), which generates the result in Ref. 8. The main result is the following theorem.

Theorem 1.1: *If the function $f(u)$ satisfies the conditions (1.4)–(1.7), then for any $\alpha\geqslant\alpha_0>0$, the Hausdorff dimension d_{H} of the global attractor for system (1.1)–(1.3) satisfies*

$$d_{\mathrm{H}}\leqslant\min\left\{ m \,\middle|\, m\in N,\ \frac{1}{m}\sum_{j=1}^{m}\lambda_j^{-4\nu_0}\leqslant\frac{2\lambda_1\alpha^2}{k\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}\right\}$$
$$\leqslant\min\left\{ m \,\middle|\, m\in N,\ \frac{1}{m}\sum_{j=1}^{m}\lambda_j^{-4\nu_0}\leqslant\frac{2\lambda_1\alpha_0^2}{k\sqrt{\alpha_0^2+4\lambda_1}(\alpha_0+\sqrt{\alpha_0^2+4\lambda_1})}\right\},\qquad(1.8)$$

where $\{\lambda_j\}_{j\in N}:0<\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_m\leqslant\cdots$, are the eigenvalues of operator $-\Delta$ with the Dirichlet boundary condition on Ω and $k=k(\alpha_0)$ is a positive constant, $0<\nu_0<\min\{(2-p)/4,\frac{1}{4}\}$, $p\in[0,2)$ is as in (1.6).

Obviously, the upper bound of d_{H} in (1.8) is a decreasing function of α because

$$h(\alpha)=\frac{2\lambda_1\alpha^2}{k\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}$$

increases as α grows and

$$\lim_{m\to+\infty}\frac{1}{m}\sum_{j=1}^{m}\lambda_j^{-4\nu_0}=0,\qquad\lim_{\alpha\to+\infty}\frac{2\lambda_1\alpha^2}{k\sqrt{\alpha+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}=\frac{\lambda_1}{k^2}.$$

II. EXISTENCE OF THE GLOBAL ATTRACTOR

It is known that the operator $A=-\Delta:D(A)=H_0^1(\Omega)\cap H^2(\Omega)\to L^2(\Omega)$ is a self-adjoint positive linear, and its eigenvalues $\{\lambda_i\}_{i\in N}$ satisfy

$$0<\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_m\leqslant\cdots,\quad\text{and}\quad\lambda_m\to+\infty,\quad\text{as }m\to+\infty.$$

Let

$$E=H_0^1(\Omega)\times L^2(\Omega),$$

$$(u,v)=\int_\Omega uv\,dx,\quad|u|=(u,u)^{1/2},\quad\forall u,v\in L^2(\Omega),$$

$$((u,v))=\int_\Omega\nabla u\cdot\nabla v\,dx,\quad\|u\|=((u,u))^{1/2},\quad\forall u,v\in H_0^1(\Omega),$$

$$(y_1,y_2)_E=((u_1,u_2))+(v_1,v_2),\quad\forall y_i=(u_i,v_i)^T\in E,\quad i=1,2,$$

$$|y|_E=(y,y)_E^{1/2},\quad\forall y=(u,v)^T\in E$$

denote the usual inner products and norms in $L^2(\Omega)$, $H_0^1(\Omega)$, and E, respectively.

From the assumptions (1.4)–(1.7), it is easy to obtain the existence and uniqueness of the solution of (1.1)–(1.3), which defines a continuous semigroup of mapping,

$$S(t):E\to E, \quad \{u_0,u_1\}\mapsto\{u,u_t\}, \quad \text{for } t\geqslant 0 \tag{2.1}$$

(see, e.g., Ref. 4 or Ref. 6 for detail).

It is convenient to reduce (1.1) to an evolution equation of the first order in time. Let $\varphi=(u,v)^T$, $v=\dot{u}+\epsilon u$, where ϵ is chosen as

$$\epsilon=\frac{\lambda_1\alpha}{\alpha^2+4\lambda_1}, \tag{2.2}$$

then (1.1)–(1.3) can be written as

$$\dot{\varphi}+\Lambda\varphi=F(\varphi), \quad \varphi(0)=(u_0,u_1+\epsilon u_0), \tag{2.3}$$

where

$$F(\varphi)=\begin{pmatrix} 0 \\ -f(u)+g \end{pmatrix}, \quad \Lambda=\begin{pmatrix} \epsilon I & -I \\ A-\epsilon(\alpha-\epsilon)I & (\alpha-\epsilon)I \end{pmatrix}. \tag{2.4}$$

It is easy to see that the semigroup,

$$S_\epsilon(t):(u_0,u_1+\epsilon u_0)^T\to(u(t),u_t(t)+\epsilon u(t))^T, \quad E\to E, \tag{2.5}$$

defined by (2.3), has the following relation with $S(t)$:

$$S_\epsilon(t)=R_\epsilon S(t)R_{-\epsilon}, \tag{2.6}$$

where $R_\epsilon:\{u,v\}\to\{u,v+\epsilon u\}$ is an isomorphism of E. So, we need consider the equivalent system (2.3) only.

Lemma 2.1: For any $\varphi=(u,v)^T\in E$,

$$(\Lambda\varphi,\varphi)_E\geqslant\sigma|\varphi|_E^2+(\alpha/2)|v|^2, \tag{2.7}$$

where

$$\sigma=\frac{\lambda_1\alpha}{\sqrt{\alpha^2+4\lambda_1}(\alpha+\sqrt{\alpha^2+4\lambda_1})}. \tag{2.8}$$

Proof: See Lemma 1 in Ref. 8.

Lemma 2.2: Suppose $0<\alpha_0\leqslant\alpha<+\infty$ and $f(u)$ satisfies conditions (1.4)–(1.5). There exists a positive constant $r_0=r_0(\alpha_0)>0$ such that the bounded ball B_0 of E, $B_0=B_E(0,r_0)$, centered at 0 of radius r_0, is an absorbing set of the semigroup $S_\epsilon(t)$, $t\geqslant 0$ in E, that is, for any bounded set B of E, there exists $T_0(B)\geqslant 0$ such that the solution $\varphi(t)=(u(t),v(t))^T$ of (2.3) starting at B satisfies

$$|\varphi(t)|_E=(\|u(t)\|^2+|v(t)|^2)^{1/2}\leqslant r_0, \quad \forall t\geqslant T_0(B), \tag{2.9}$$

in which $v=u_t+\epsilon u$.

Proof: Let $G(s)=\int_0^s f(r)dr=G_1(s)+G_2(s)$. By (1.4) and (1.5), we have

$$G_1(s)\geqslant 0, \quad \liminf_{|s|\to+\infty}\frac{G(s)}{s^2}\geqslant 0, \quad \forall s\in R. \tag{2.10}$$

and there exists a constant $c_4>0$ such that

$$\lim_{|s|\to+\infty}\inf\frac{sf(s)-c_4G(s)}{s^2}\geqslant 0,\quad \forall s\in R. \tag{2.11}$$

Let $\overline{G}(u)=\int_\Omega G(u)dx$. Following the procedure leading to (23) in Ref. 8, we have

$$|\varphi|_E^2\leqslant 2[|\varphi(0)|_E^2+2\overline{G}(u_0)+2k_1]e^{-\beta t}+\frac{4}{\theta}\left(\frac{\alpha^2+4\lambda_1}{\lambda_1\alpha^2}|g|^2+2(c_1k_1+k_2)\right), \tag{2.12}$$

where k_1, $k_2\geqslant 0$, $\beta>0$, $\theta=\min\{1,2c_4\}$ are constants. Thus

$$\lim_{t\to+\infty}\sup|\varphi|_E^2\leqslant\frac{4}{\theta}\left(\frac{\alpha_0^2+4\lambda_1}{\lambda_1\alpha_0^2}|g|^2+2(c_1k_1+k_2)\right)=\frac{1}{2}r_0(\alpha_0). \tag{2.13}$$

The proof is completed.

Lemma 2.3: For any initial value $u_0\in H_0^1(\Omega)$, $u_1\in L^2(\Omega)$ with

$$\|u_0\|^2+|u_1+\epsilon u_0|^2\leqslant r_0^2, \tag{2.14}$$

there exists a constant $r_1=r_1(r_0)$ such that the solution of (2.3) $\varphi(t)=(u(t),v(t))^T$ in which $v=u_t+\epsilon u$ satisfies

$$|\varphi(t)|_E\leqslant r_1,\quad \forall t\geqslant 0. \tag{2.15}$$

Proof: By (2.11),

$$\overline{G}(u_0)\leqslant\frac{1}{c_4}\left[|f(u_0)|\cdot|u_0|+\frac{1}{8}\|u_0\|^2+k_3\right],\quad \forall u\in H_0^1(\Omega). \tag{2.16}$$

By (1.6),

$$|f'(s)|\leqslant c_5(1+|s|^2),\quad \forall s\in R,\quad c_5\geqslant 0. \tag{2.17}$$

By (2.17) and the Sobolev embedding theorem, $H_0^1(\Omega)\subset L^6(\Omega)$,

$$|f(u_0)|^2\leqslant c_6^2\int_\Omega(1+|u_0(x)|^3)^2\,dx\leqslant c_7(r_0). \tag{2.18}$$

By the inequality (2.12), we obtain (2.15). The proof is completed.

Let $u(t)$ be a solution of system (1.1)–(1.3) with the initial value u_0, u_1 satisfying (2.14). We decompose $u(t)$ into $u(t)=z(t)+w(t)$, where $z(t)$ and $w(t)$ satisfy, respectively,

$$z_{tt}+\alpha z_t-\Delta z+f_1(z)=0,$$

$$z(x,t)|_{x\in\partial\Omega}=0,\quad z(0)=u_0,\quad z_t(0)=u_1, \tag{2.19}$$

$$w_{tt}+\alpha w_t-\Delta w+f_1(u)-f_1(z)=g-f_2(u(t)),$$

$$w(x,t)|_{x\in\partial\Omega}=0,\quad w(0)=w_t(0)=0. \tag{2.20}$$

Lemma 2.4: There exist two positive constants $M_1(r_0)$ and $\sigma_1(r_0)$ such that

$$\|z(t)\|^2+|z_t(t)|^2\leqslant M_1(r_0)\exp(-\sigma_1(r_0)t),\quad \forall t\geqslant 0. \tag{2.21}$$

Proof: Let $z' = z_t + \epsilon z$, $\phi = (z,z')^T$, where ϵ is as in (2.2); then (2.19) can be written as

$$\phi_t + \Lambda\phi + F_1(\phi) = 0, \quad \phi(0) = (u_0, u_1 + \epsilon_1 u_0)^T, \tag{2.22}$$

where $F_1(\phi) = (0, f_1(z))^T$, Λ is as in (2.4). Write $\overline{G}_1(z) = \int_\Omega G_1(z)dx$. Taking the inner product $(\cdot,\cdot)_E$ of (2.22) with $\phi = (z,z')^T$, we find

$$\frac{d}{dt}[|\phi|_E^2 + 2\overline{G}_1(z)] + 2(\Lambda\phi,\phi)_E + 2\epsilon(f_1(z),z) = 0. \tag{2.23}$$

Similar to Lemma 2.3, there exists a constant $c_8 = c_8(r_0) \geqslant 0$ such that

$$|\phi(t)|_E^2 = \|z(t)\|^2 + |z_t(t) + \epsilon z(t)|^2 \leqslant c_8^2, \quad \forall t \geqslant 0. \tag{2.24}$$

From (1.4) and (1.6), we deduce that $f_1(0) = 0$ and $|f_1(s)| \leqslant c_9(|s|^3 + |s|)(c_9 > 0)$. Hence, for every $z \in H_0^1(\Omega)$, by the Sobolev embedding $H_0^1(\Omega) \subset L^4(\Omega)$ and (2.24), we have

$$0 \leqslant \overline{G}_1(z) \leqslant c_{10}(\|z\|_{L^4}^4 + \|z\|_{L^2}^2) \leqslant c_{11}\|z\|^2(\|z\|^2 + 1) \leqslant c_{12}(r_0)\|z\|^2, \tag{2.25}$$

i.e.,

$$\sigma\|z\|^2 \geqslant \frac{\sigma}{c_{12}(r_0)}\overline{G}_1(z). \tag{2.26}$$

By (1.4),

$$\epsilon(f_1(z),z) \geqslant 0. \tag{2.27}$$

Plugging (2.7), (2.26), and (2.27) into (2.23), we obtain

$$\frac{d}{dt}[|\phi|_E^2 + 2\overline{G}_1(z)] + \sigma|\phi|_E^2 + \frac{\sigma}{c_{12}(r_0)}\overline{G}_1(z) \leqslant 0. \tag{2.28}$$

By the Gronwall inequality, (2.24) and (2.28), we have

$$|\phi|_E^2 \leqslant (|\phi(0)|_E^2 + 2\overline{G}_1(z(0))\exp(-\sigma_1(r_0)t) \leqslant M_0(r_0)\exp(-\sigma_0(r_0)t), \quad \forall t \geqslant 0, \tag{2.29}$$

where $\sigma_0(r_0) = \min\{\sigma, \sigma/2c_{12}(r_0)\}$, $M_0(r_0) = r_0^2(1 + 2c_{12}(r_0))$. It follows that (2.21) holds. The proof is completed.

Lemma 2.5: There exist constants $M_2(r_0) > 0$ and $0 < \nu_0 < \min\{(2-p)/4, \frac{1}{4}\}$ such that $w(t)$ satisfies

$$|A^{\nu_0 + 1/2}w(t)|^2 + |A^{\nu_0}w_t(t)|^2 \leqslant M_2(r_0), \quad \forall t \geqslant 0, \tag{2.30}$$

where p is as in (1.6).

Proof: We make use of the space $D(A^{\nu/2}), \nu \geqslant 0$, which is a Hilbert space with the scalar product and the norm

$$((w_1, w_2))_\nu = (A^{\nu/2}w_1, A^{\nu/2}w_2), \quad \|w\|_\nu = ((w,w))_\nu; \tag{2.31}$$

then

$$H_0^\nu(\Omega) \subset D(A^{\nu/2}) \subset H^\nu(\Omega). \tag{2.32}$$

Set $0 \leqslant \nu \leqslant \frac{1}{4}$, we multiply the equation (2.20) by $A^{2\nu}w$ and integrate by parts to deduce

$$|A^{\nu}w_t(t)|^2+|A^{\nu+1/2}w(t)|^2+\alpha\int_0^t(A^{\nu}w_t(s),A^{\nu}w_t(s))ds$$
$$\leqslant\int_\Omega[|(f_1(u(t))-f_1(z(t)))|+|(g-f_2(u(t)))|]\cdot|A^{2\nu}w(t)|dx$$
$$+\int_0^t\int_\Omega[|(f_1'(u(s))-f_1'(z(s)))z_t(s)|+|f_1'(u(s))w_t(s)|$$
$$+|f_2'(u(s))u_t(s)|]\cdot|A^{2\nu}w(s)|dx\,ds. \tag{2.33}$$

From Lemma 2.3 and Lemma 2.4, we observe that $\|w(t)\|$ and $|w_t(t)|$ are uniformly bounded, i.e., $\|w(t)\|\leqslant c_{13}(r_0)$, $|w_t(t)|\leqslant c_{14}(r_0)$, $\forall t\geqslant 0$. By the embedding relation

$$H^{\nu_1}(\Omega)\subset H^{\nu_2}(\Omega),\quad \text{if}\quad \nu_1\geqslant\nu_2\quad\text{and}\quad H^{\nu}(\Omega)\subset L^q(\Omega),\quad\text{where}\quad \frac{1}{q}=\frac{1}{2}-\frac{\nu}{3}, \tag{2.34}$$

$$\int_\Omega|(f_1(u(t))-f_1(z(t)))A^{2\nu}w(t)|dx\leqslant|f_1(u(t))-f_1(z(t))||A^{2\nu}w(t)|\leqslant c_{15}(r_0),\quad\forall t\geqslant 0. \tag{2.35}$$

$$\int_\Omega|(g-f_2(u(t)))\cdot A^{2\nu}w(t)|dx\leqslant|g-f_2(u(t))||A^{2\nu}w(t)|\leqslant c_{16}(r_0),\quad\forall t\geqslant 0. \tag{2.36}$$

By the Holder inequality, (2.34), and (1.6), we have

$$\|f_2'(u(s))u_t(s)\|_{L^r}\leqslant c_{17}(r_0), \tag{2.37}$$

$$\|(f_1'(u(s))-f_1'(z(s)))z_t(s)\|_{L^\delta}\leqslant c_{18}(r_0)|A^{\nu+1/2}w(s)|, \tag{2.38}$$

$$\|f_1'(u(s))w_t(s)\|_{L^\delta}\leqslant c_{19}(r_0)|A^{\nu}w_t(s)|, \tag{2.39}$$

$$\|A^{2\nu}w(s)\|_{L^r}=\|A^{\nu-1/2}A^{\nu+1/2}w(s)\|_{L^r}\leqslant c_{20}|A^{\nu+1/2}w(s)|, \tag{2.40}$$

where $\tau=6/(3+p)$, $p\in[0,2)$ is as in (1.6), $1/\delta=\frac{5}{6}-2\nu/3$, and $1/r=1-1/\delta$. Setting $0<\nu<\min\{(2-p)/4,\frac{1}{4}\}$, plugging above estimates into (2.33), we have

$$|A^{\nu}w_t(t)|^2+|A^{\nu+1/2}w(t)|^2\leqslant c_{21}(r_0)+c_{22}(r_0)\int_0^t[|A^{\nu}w_t(s)|^2+|A^{\nu+1/2}w(s)|^2]ds. \tag{2.41}$$

By applying the Gronwall inequality and zero initial value conditions (2.20), we obtain (2.30). The proof is completed.

Lemma 2.6: If $f(u)$ satisfies (1.4)–(1.7), then the semigroup $S_\epsilon(t),t\geqslant 0$ possesses a global attractor β in E and β is included in the bounded ball B_0 of E.

Proof: It is a direct consequence of Lemmas 2.2, 2.4, 2.5, and Theorem I.1.1 in Ref. 9.

III. PROOF OF THEOREM 1.1

To estimate the Hausdorff dimension of the global attractor β for (2.3) in E, we consider the first variation equation of (2.3) with the initial value condition,

$$\Psi'=-\Lambda\Psi+F'(\varphi)\Psi,\quad \Psi(0)=(\xi,\eta)^T\in E, \tag{3.1}$$

where $\Psi=(U,V)^T\in E$, and $\varphi=(u,v)^T$ is a solution of (2.3) and

$$F'(\varphi)=\begin{pmatrix} 0 & 0 \\ -f'(u) & 0 \end{pmatrix}.$$

Lemma 3.1: The system (3.1) is a well-posed problem in E; the mapping $S_\epsilon(t)$ defined by (2.3) is Fréchet differentiable on E for any $t>0$; its differential at $\varphi=(u_0,u_1+\epsilon u_0)^T$ is the linear operator on E, $(\xi,\eta)^T\mapsto(U(t),V(t))^T$, where $(U,V)^T$ is the solution of (3.1).

Proof: It is a direct consequence of (1.7), (2.6), and Lemma VI. 6.1 in Ref. 9.

Lemma 3.2: Consider the system (2.3). Let Φ denote a set of m vectors $\{\Phi_1,\Phi_2,\ldots,\Phi_m\}$ that are orthonormal in E. If

$$q_m=\overline{\lim_{t\to+\infty}}\ \sup_{\Phi\subset E}\ \sup_{\varphi\in ß}\frac{1}{t}\int_0^t\sum_{j=1}^m\left((-\Lambda+F'(\varphi(\tau)))\Phi_j(\tau),\quad \Phi_j(\tau)\right)_E d\tau\leqslant 0, \tag{3.2}$$

then the Hausdorff dimension of the global attractor β is less than or equal to m.

Proof: This is a direct consequence of Theorem V. 3.3., Eqs. (V. 3.47)–(V. 3.49) of Ref. 9.

Lemma 3.3: For any orthonormal family of elements of E, $\{(\xi_j,\eta_j)^T\}_{j=1}^m$, we have

$$\sum_{j=1}^m |A^{(1/2)\nu}\xi_j|^2\leqslant\sum_{j=1}^m \lambda_j^{\nu-1},\quad \forall\nu\in[0,1). \tag{3.3}$$

Proof: See Lemma VI, 6.3 in Ref. 9.

Lemma 3.4: If the functions $f(u)$ satisfy assumptions (1.4)–(1.7), then for any $\alpha\geqslant\alpha_0>0$, the Hausdorff dimension $d_H(ß)$ of the global attractor ß for system (2.3) in E satisfies

$$d_H(ß)\leqslant\min\left\{m\,\middle|\, m\in N,\frac{1}{m}\sum_{j=1}^m \lambda_j^{-4\nu_0}\leqslant\frac{2\alpha\sigma}{k}\right\}, \tag{3.4}$$

where $k=k(\alpha_0)$ is a positive constant that is independent of α, σ is as in (2.8), $0<\nu_0<\min\{(2-p)/4,\frac{1}{4}\}$, and p is as in (1.6).

Proof: Let $m\in N$ be fixed. Consider m solutions $\Psi_1,\Psi_2,\ldots,\Psi_m$ of (5.1). At a given time τ, let $Q_m(\tau)$ denote the orthogonal projection in E onto the space span $\{\Psi_1(\tau),\Psi_2(\tau),\ldots,\Psi_m(\tau)\}$. Let $\Phi_j(\tau)=(\xi_j,\eta_j)^T\in E$, $j=1,2,...,m$, be an orthonormal basis of $Q_m(\tau)E$.

Suppose $\varphi(\tau)=(u(\tau),v(\tau))^T\in ß$; then $|\varphi(\tau)|_E\leqslant r_0$ (r_0 is defined by Lemma 2.2). By Lemma 2.1 and $|\Phi_j|_E=1$,

$$-(\Lambda\Phi_j,\Phi_j)_E\leqslant-\sigma-\frac{\alpha}{2}\|\eta_j\|^2, \tag{3.5}$$

$$(F'(\varphi(\tau))\Phi_j(\tau),\Phi_j(\tau))_E=(-f'(u(\tau))\xi_j(\tau),\eta_j(\tau))\leqslant|f'(u(\tau))\xi_j(\tau)|\cdot|\eta_j|. \tag{3.6}$$

By (2.17), (2.21), the Young inequality, and Hölder inequality,

$$\begin{aligned}
|f'(u(t)))\xi_j(t)|^2&\leqslant c_5^2\int_\Omega[1+(z(t)+w(t))^2]^2\xi_j^2(t)dx\\
&\leqslant c_{23}\int[1+z^4(t)+w^4(t)]\xi_j^2(t)dx\\
&\leqslant c_{24}[\|z(t)\|_{L^6}^4\cdot\|\xi_j(t)\|_{L^6}^2+|\xi_j(t)|^2+\|w(t)\|^4_{L^{6/(1-4\nu_0)}}\cdot\|\xi_j(t)\|L^{6/(1+8\nu_0)}]\\
&\leqslant c_{25}(r_0)(e^{-4\sigma_1(r_0)t}\|\xi_j(t)\|^2+|\xi_j(t)|^2+\|A^{\nu_0+1/2}w(t)\|^4\cdot\|A^{(1-4\nu_0)/2}\xi_j(t)\|^2)\\
&\leqslant c_{26}(r_0)(e^{-4\sigma_1(r_0)t}\|\xi_j(t)\|^2+\|A^{(1-4\nu_0)/2}\xi_j(t)\|^2),\quad\forall t\geqslant 0,
\end{aligned} \tag{3.7}$$

where $0<\nu_0<\min\{(2-p)/4,1/4\}$ is defined by Lemma 2.4. By (3.6) and (3.7), there exists a constant $k=k(\alpha_0)=c_{26}^2(r_0)>0$, which is independent of α, such that

$$(F'(\varphi(\tau))\Phi_j(\tau),\Phi_j(\tau))_E\leqslant\frac{k}{2\alpha}(e^{-4\sigma_1(r_0)\tau}\|\xi_j(\tau)\|^2+\|A^{(1-4\nu_0)/2}\xi_j(\tau)\|^2)+\frac{\alpha}{2}|\eta_j(t)|^2. \tag{3.8}$$

Thus, by $\|\xi_j(\tau)\|^2\leqslant|\Phi_j|_E^2=1$ and (3.2), (3.5) and (3.8),

$$\begin{aligned}q_m&=\overline{\lim_{t\to+\infty}}\sup_{\Phi\subset E}\sup_{\varphi\in ß}\frac{1}{t}\int_0^t\sum_{j=1}^m((-\Lambda+F'(\varphi(\tau)))\Phi_j(\tau),\Phi_j(\tau)_E\,d\tau\\&\leqslant\overline{\lim_{t\to+\infty}}\left(-m\sigma+\frac{km}{8\sigma_1(r_0)\alpha t}(1-e^{-4\sigma_1(r_0)t})+\frac{k}{2\alpha}\sum_{j=1}^m\lambda_j^{-4\nu_0}\right)\\&\leqslant-\frac{mk}{2\alpha}\left(\frac{2\alpha\sigma}{k}-\frac{1}{m}\sum_{j=1}^m\lambda_j^{-4\nu_0}\right).\end{aligned}$$

If

$$\frac{1}{m}\sum_{j=1}^m\lambda_j^{-4\nu_0}\leqslant\frac{2\alpha\sigma}{k},$$

then $q_m\leqslant0$; hence, (3.4) holds. The proof is completed.

Combining with Lemma 3.4 and (2.8), we complete the proof of Theorem 1.1.

ACKNOWLEDGMENT

Supported by the National Natural Science Foundation of China.

[1] S. S. Ceron and O. Lopes, ''α-contractions and attractors for dissipative semilinear hyperbolic equations and systems,'' Annali. Mat. Pura Appl. **140**, 193–206 (1991).
[2] J. K. Hale, *Asymptotic Behavior Of Dissipative Systems* (American Mathematical Society, Providence, RI, 1988).
[3] J. Arrieta, A. N. Carvalho, and J. K. Hale, ''A damping hyperbolic equation with critical exponent,'' Commun. Partial Diff. Eqns. **17**, 841–866 (1992).
[4] E. Feireisl and E. Zuazua, ''Global attractors for semilinear wave equations with locally distributed nonlinear damping and critical exponent,'' Commun. Partial Diff. Eqns. **18**, 1539–1555 (1993).
[5] Yu Huang, ''Global attractors for semilinear wave equations with nonlinear damping and critical exponent,'' Appl. Anal. **56**, 165–174 (1995).
[6] E. Feireisl, ''Global attractors for semilinear damped wave equations with supercritical exponent,'' J. Diff. Eqns. **116**, 431–447 (1995).
[7] Y. Yu, ''Estimate of the Hausdorff and fractal dimensions of the global attractors for nonlinear wave equations with damping,'' Acta Math. Appl. Sinica (in Chinese) **21**, 257–266 (1998).
[8] Z. Shengfan, ''On dimension of the global attractor for damped nonlinear wave equations,'' J. Math. Phys. **40**, 1432–1438 (1999).
[9] R. Témam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sciences 68 (Springer-Verlag, New York, 1988).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Families of quasi-bi-Hamiltonian systems and separability

Yunbo B. Zeng[a)]
Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China

Wen-Xiu Ma[b)]
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, People's Republic of China

(Received 2 October 1998; accepted for publication 14 June 1999)

It is shown how to construct an infinite number of families of quasi-bi-Hamiltonian (QBH) systems by means of the constrained flows of soliton equations. Three explicit QBH structures are presented for the first three families of the constrained flows. The Nijenhuis coordinates defined by the Nijenhuis tensor for the corresponding families of QBH systems are proved to be exactly the same as the separated variables introduced by mean of the Lax matrices for the constrained flows. © *1999 American Institute of Physics.* [S0022-2488(99)02909-6]

I. INTRODUCTION

As is known, some integrable systems possess bi-Hamiltonian structure. We recall some known results. Let M be a differential manifold, TM and T^*M its tangent and cotangent bundle, and θ_0 and θ_1: $T^*M \rightarrow TM$ two compatible Poisson tensors on M.[1] A vector field X is said to be bi-Hamiltonian (BH) with respect to θ_0 and θ_1, if two smooth functions, $H,F \in C^\infty(M)$, exist such that

$$X = \theta_0\, dH = \theta_1\, dF, \tag{1.1}$$

where dF denotes the differential of F (gradient ∇F for finite system and variation δF for field system). If θ_0 is invertible, the tensor $\Phi = \theta_1 \theta_0^{-1}$ is a Nijenhuis tensor or hereditary operator. The operator Φ maps a given BH vector field into another BH vector field. Hence having a Nijenhuis tensor, one can construct a hierarchy of Hamiltonian symmetries, and a related hierarchy of integrals of motion for the underlying system. The BH structure (1.1) ensures that the resulting integrals of motion are pairwise in involution with respect to both Poisson brackets. Thus the BH structure of a given system is important for its integrability.

Unfortunately, for a majority of the BH finite-dimensional systems, none of the θ_0 and θ_1 is invertible. In fact, all the known BH finite-dimensional systems arising from the constrained flows or stationary flows of soliton equations usually exist in an extended phase space and both θ_0 and θ_1 are degenerated (see, for example, Refs. 2–8). In their natural phase space these systems may satisfy a weaker condition than the BH one. The notion of a quasi-bi-Hamiltonian (QBH) system was introduced.[9,10] According to,[10] for $\dim M = 2n$, a vector field, X, is said to be a QBH vector field with respect to Poisson tensors, θ_0 and θ_1, if there exist three smooth functions H, F, ρ, such that

$$X = \theta_0 \nabla H = \frac{1}{\rho}\, \theta_1 \nabla F, \tag{1.2}$$

[a)]Electronic mail: yzeng@tsinghua.edu.cn
[b)]Electronic mail: mawx@cityu.edu.hk

where two Poisson tensors θ_0 and θ_1 are compatible and nondegenerated (invertible). The function ρ is called an integrating factor. On a $2n$-dimensional symplectic manifold M, let ($\mathbf{q}=(q_1,\ldots,q_n)$, $\mathbf{p}=(p_1,\ldots,p_n)$) be a set of canonical coordinates and θ_0 the canonical Poisson matrix $\theta_0=\begin{pmatrix} 0 & I \\ -I & 0\end{pmatrix}$ (I denoting the $n\times n$ identity matrix). As θ_0 and θ_1 are compatible and invertible, the Nijenhuis tensor $\Phi=\theta_1\theta_0^{-1}$ is maximal, i.e., it has n distinct eigenvalues $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)$. As is known,[11] in a neighborhood of a regular point, where the eigenvalues $\boldsymbol{\mu}$ are distinct, one can construct a canonical transformation $(\mathbf{q},\mathbf{p})\rightarrow(\boldsymbol{\mu},\boldsymbol{\nu})$(($\boldsymbol{\mu},\boldsymbol{\nu}$) referred to as the Nijenhuis coordinates) such that θ_1 and Φ take the Darboux form

$$\theta_1=\begin{pmatrix} 0 & \Lambda_1 \\ -\Lambda_1 & 0\end{pmatrix}, \quad \Phi=\begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_1\end{pmatrix}, \quad \Lambda_1=\mathrm{diag}(\mu_1,\ldots,\mu_n). \tag{1.3}$$

A QBH vector field is said to be Pfaffian[10] if, in the Nijenhuis coordinates, an integrating factor ρ in Eq. (1.2) is the product of the eigenvalues of Φ, i.e.,

$$\rho=\prod_{i=1}^{n}\mu_i\,. \tag{1.4}$$

In the Pfaffian case, the general solutions, H and F, of Eq. (1.2) are obtained and the Hamilton–Jacobi equation for H is shown to be separable by verifying the Levi–Civita conditions.[12] Some relationship between BH and QBH structure is discussed in Ref. 13. Several QBH systems are presented.[9,10,12–14] It is in general quite difficult to directly construct a BH or QBH structure for a given integrable Hamiltonian vector field. In recent years much work has been devoted to the constrained flows of soliton equations (see, for example, Refs. 2–8, 15–24). One of the aims of this paper is to show how to construct an infinite number of families of QBH systems from the constrained flows of soliton equations. We have presented some families of the constrained flows in order to study the dynamical r-matrices in Ref. 24. We now describe the explicit QBH structures for these families of the constrained flows.

The Lax representation for the constrained flows of soliton equations can always be deduced from the adjoint representation of the Lax pair for soliton equations.[16,17] There is an effective way for the separation of variables for some finite-dimensional integrable Hamiltonian systems with some kind of Lax matrices.[25,26] The separated variables for some constrained flows can be introduced and the Jacobi inversion problems for the constrained flows can be established by means of the Lax representation.[27,28] We are interested in the relationship between the two methods for the separability mentioned above. Another main aim of this paper is to prove that the Nijenhuis coordinates for the underlying families of QBH systems are usually the same as the separated variables introduced by the Lax matrices.

The paper is organized as follows.

In Sec. II we present a new QBH system. We directly construct the second compatible Poisson tensor by using a map relating this system to its modified version, and prove that the Nijenhuis coordinates for this system is equivalent to the separated variables defined by Lax matrix. We make some comparison of the two methods for separability. In Secs. III and IV, by using the constrained flows associated with the polynomial second-order spectral problems and the higher-order symmetry constraints, we propose a way to construct an infinite number of families of QBH systems. The explicit QBH structures of the first two families of constrained flows are given. The equivalence of the Nijenhuis coordinates and the separated variables is proved. In Sec. V we point out that the two compatible Poisson tensors θ_0, θ_1 and the integrating factor ρ given by the QBH structure (2.28) and (2.29a) are just that for the third family of QBH systems. Also some conclusions and a conjucture are given.

II. NEW QBH SYSTEM

In this section we present a new QBH system. By using a map relating this system to its modified version, the second compatible Poisson tensor is obtained from the image of the Poisson tensor for the modified version under the map. We use this system to illustrate how to prove the equivalence of the Nijenhuis coordinates and the separated variables introduced by the Lax matrix.

A. New finite-dimensional integrable Hamiltonian system

For Jaulent–Miodek (JM) spectral problem[29]

$$\psi_x = U(u,\lambda)\psi, \quad U(u,\lambda) = \begin{pmatrix} 0 & 1 \\ \lambda^2 - u_1\lambda - u_0 & 0 \end{pmatrix}, \quad \psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_0 \end{pmatrix}, \tag{2.1}$$

its adjoint representation is defined by[30]

$$V_x = [U,V] \equiv UV - VU, \tag{2.2}$$

where V is taken as

$$V = \sum_{i=0}^{\infty} V_i \lambda^{-i}, \quad V_i = \begin{pmatrix} a_i & b_i \\ c_i & -a_i \end{pmatrix}. \tag{2.3}$$

Then Eq. (2.2) and (2.3) yield

$$a_0 = a_1 = a_2 = b_0 = b_1 = 0, \quad b_2 = 1, \quad b_3 = \tfrac{1}{2}u_1,$$

$$a_3 = -\tfrac{1}{4}u_{1,x}, \quad c_0 = 1, \quad c_1 = -\tfrac{1}{2}u_1, \ldots,$$

and in general

$$\begin{pmatrix} b_{k+2} \\ b_{k+1} \end{pmatrix} = L \begin{pmatrix} b_{k+1} \\ b_k \end{pmatrix}, \quad k = 1,2,..., \tag{2.4a}$$

$$a_k = -\tfrac{1}{2}b_{k,x}, \quad c_k = a_{k,x} - u_0 b_k - u_1 b_{k+1} + b_{k+2}, \quad k = 1,2,..., \tag{2.4b}$$

where

$$L = \begin{pmatrix} u_1 - \tfrac{1}{2}D^{-1}u_{1,x} & \tfrac{1}{4}D^2 + u_0 - \tfrac{1}{2}D^{-1}u_{0,x} \\ 1 & 0 \end{pmatrix}, \quad D = \frac{\partial}{\partial x}, \quad DD^{-1} = D^{-1}D = 1.$$

The Jaulent–Miodek hierarchy associated with (2.1) can be written as an infinite-dimensional Hamiltonian system

$$u_{t_n} = \begin{pmatrix} u_1 \\ u_0 \end{pmatrix}_{t_n} = J \begin{pmatrix} b_{n+2} \\ b_{n+1} \end{pmatrix} = J \frac{\delta H_n}{\delta u}, \quad n = 1,2,..., \tag{2.5}$$

where the Hamiltonian H_n and the Hamiltonian operator J are given by

$$J = \begin{pmatrix} 0 & 2D \\ 2D & -u_{1x} - 2u_1 D \end{pmatrix}, \quad H_n = \frac{1}{n}(2b_{n+3} - u_1 b_{n+2}).$$

Under zero boundary condition we have

$$\frac{\delta\lambda}{\delta u} = \begin{pmatrix} \lambda\psi_1^2 \\ \psi_1^2 \end{pmatrix}, \quad L\frac{\delta\lambda}{\delta u} = \lambda\frac{\delta\lambda}{\delta u}. \tag{2.6}$$

The constrained flow of (2.5) consists of the equations obtained from the spectral problem (2.1) for N distinct λ_j and the restriction of the variational derivatives for conserved quantities H_l (for any fixed l) and λ_j:[15–17]

$$\Psi_{1,x}=\Psi_2, \quad \Psi_{2,x}=\Lambda^2\Psi_1-u_1\Lambda\Psi_1-u_0\Psi_1, \tag{2.7a}$$

$$\frac{\delta H_l}{\delta u}-\frac{1}{2}\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta u}=\begin{pmatrix} b_{l+2}\\ b_{l+1}\end{pmatrix}-\frac{1}{2}\begin{pmatrix}\langle\Lambda\Psi_1,\Psi_1\rangle\\ \langle\Psi_1,\Psi_1\rangle\end{pmatrix}=0, \tag{2.7b}$$

which has been recognized as a symmetry constraint.[18–20] Hereafter we denote the inner product in R^N by $\langle.,.\rangle$ and $\Psi_i=(\psi_{i1},\dots,\psi_{iN})^T$, $i=1,2$, $\Lambda=\mathrm{diag}(\lambda_1,\dots,\lambda_N)$.

For $l=4$, we have

$$H_4=\tfrac{7}{128}u_1^5+\tfrac{5}{16}u_1^3u_0-\tfrac{5}{32}u_{1x}^2u_1+\tfrac{3}{8}u_0^2u_1-\tfrac{1}{8}u_{1x}u_{0x}. \tag{2.8}$$

By introducing the Jacobi–Ostrogradsky coordinates

$$q_1=u_1, \quad q_2=u_0,$$

$$p_1=\frac{\delta H_4}{\delta u_{1x}}=-\frac{5}{16}u_1u_{1x}-\frac{1}{8}u_{0x}, \quad p_2=\frac{\delta H_4}{\delta u_{0x}}=-\frac{1}{8}u_{1x}, \tag{2.9}$$

the Eqs. (2.7) for $l=4$ are transformed into a finite-dimensional Hamiltonian system (FDHS)

$$\Psi_{1x}=\frac{\partial F_1}{\partial\Psi_2}=\Psi_2, \quad q_{1x}=\frac{\partial F_1}{\partial p_1}=-8p_2, \quad q_{2x}=\frac{\partial F_1}{\partial p_2}=-8p_1+20q_1p_2, \tag{2.10a}$$

$$\Psi_{2x}=-\frac{\partial F_1}{\partial\Psi_1}=\Lambda^2\Psi_1-q_1\Lambda\Psi_1-q_2\Psi_1, \tag{2.10b}$$

$$p_{1x}=-\frac{\partial F_1}{\partial q_1}=\frac{35}{128}q_1^4+\frac{15}{16}q_1^2q_2-10p_2^2+\frac{3}{8}q_2^2-\frac{1}{2}\langle\Lambda\Psi_1,\Psi_1\rangle, \tag{2.10c}$$

$$p_{2x}=-\frac{\partial F_1}{\partial q_2}=\frac{5}{16}q_1^3+\frac{3}{4}q_1q_2-\frac{1}{2}\langle\Psi_1,\Psi_1\rangle, \tag{2.10d}$$

or equivalently

$$P_x=\theta_0\nabla F_1,$$

where

$$P=(\Psi_1^T,q_1,q_2,\Psi_2^T,p_1,p_2)^T, \quad \theta_0=\begin{pmatrix}0 & I_{(N+2)\times(N+2)}\\ -I_{(N+2)\times(N+2)} & 0\end{pmatrix},$$

$$\begin{aligned}F_1={}&\tfrac{1}{2}\langle\Psi_2,\Psi_2\rangle-\tfrac{1}{2}\langle\Lambda^2\Psi_1,\Psi_1\rangle+\tfrac{1}{2}q_1\langle\Lambda\Psi_1,\Psi_1\rangle+\tfrac{1}{2}q_2\langle\Psi_1,\Psi_1\rangle\\ &-8p_1p_2+10q_1p_2^2-\tfrac{5}{16}q_1^3q_2-\tfrac{3}{8}q_1q_2^2-\tfrac{7}{128}q_1^5.\end{aligned} \tag{2.11}$$

The Lax representation for FDHS (2.10) can be deduced from the adjoint representation (2.2) by using the method in Refs. 16 and 17 which is sketched as follows. Due to (2.4a), (2.6), and (2.7b), we may define

$$\tilde{b}_m=\tfrac{1}{2}\langle\Lambda^{m-5}\Psi_1,\Psi_1\rangle, \quad m=5,6,\dots,$$

which together with (2.4b) and (2.10) yields

$$\tilde{a}_m = -\tfrac{1}{2}\langle \Lambda^{m-5}\Psi_1, \Psi_2\rangle, \quad \tilde{c}_m = -\tfrac{1}{2}\langle \Lambda^{m-5}\Psi_2, \Psi_2\rangle, \quad m=5,6,\ldots .$$

Set

$$\tilde{a}_m = a_m, \quad \tilde{b}_m = b_m, \quad \tilde{c}_m = c_m, \quad m=0,1,2,3,4.$$

Then the construction of $\tilde{a}_m, \tilde{b}_m, \tilde{c}_m$ ensures that under (2.10),

$$\tilde{V} = \sum_{i=0}^{\infty} \tilde{V}_i \lambda^{-i}, \quad \tilde{V}_i = \begin{pmatrix} \tilde{a}_i & \tilde{b}_i \\ \tilde{c}_i & -\tilde{a}_i \end{pmatrix},$$

also satisfies (2.2). Notice that

$$\sum_{m=5}^{\infty} \tilde{a}_m \lambda^{-m+4} = -\frac{1}{2}\sum_{m=0}^{\infty}\sum_{j=1}^{N}\left(\frac{\lambda_j}{\lambda}\right)^m \psi_{1j}\psi_{2j} = -\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\lambda-\lambda_j},$$

set

$$Q \equiv \begin{pmatrix} A(\lambda) & B(\lambda) \\ C(\lambda) & -A(\lambda) \end{pmatrix} = \lambda^4 \tilde{V}, \tag{2.12a}$$

we have

$$A(\lambda) = 2p_2\lambda + 2p_1 - 2q_1p_2 - \frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\lambda-\lambda_j}, \tag{2.12b}$$

$$B(\lambda) = \lambda^2 + \frac{1}{2}q_1\lambda + \frac{3}{8}q_1^2 + \frac{1}{2}q_2 + \frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}, \tag{2.12c}$$

$$\begin{aligned} C(\lambda) = {} & \lambda^4 - \frac{1}{2}q_1\lambda^3 - \left(\frac{1}{2}q_2 + \frac{1}{8}q_1^2\right)\lambda^2 + \left(\frac{1}{4}q_1^3 + \frac{1}{2}q_1q_2 - \frac{1}{2}\langle\Psi_1,\Psi_1\rangle\right)\lambda \\ & + \frac{1}{4}q_2^2 - \frac{5}{64}q_1^4 - 4p_2^2 - \frac{1}{2}\langle\Lambda\Psi_1,\Psi_1\rangle + \frac{1}{2}q_1\langle\Psi_1,\Psi_1\rangle - \frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}^2}{\lambda-\lambda_j}. \end{aligned} \tag{2.12d}$$

Since $\tilde{V}$ under (2.10) satisfies (2.2), then Q under (2.10) satisfies (2.2), too, namely

$$Q_x = [U, Q], \tag{2.13}$$

which presents the Lax representation for (2.10). This can also be verified by a direct calculation. Equation (2.13) implies that $\frac{1}{2}\operatorname{Tr} Q^2(\lambda) = A^2(\lambda) + B(\lambda)C(\lambda)$ is the generating function of the integrals of motion for (2.10). We have

$$A^2(\lambda) + B(\lambda)C(\lambda) = \lambda^6 - F_1\lambda + F_2 + \sum_{i=1}^{N}\frac{F^{(i)}}{\lambda-\lambda_i}, \tag{2.14}$$

$$\begin{aligned} F_2 = {} & -\tfrac{1}{2}\langle\Lambda\Psi_2,\Psi_2\rangle + \tfrac{1}{2}\langle\Lambda^3\Psi_1,\Psi_1\rangle - \tfrac{1}{4}q_1\langle\Lambda^2\Psi_1,\Psi_1\rangle + \left(\tfrac{1}{8}q_1^3 + \tfrac{1}{4}q_1q_2 - \tfrac{1}{4}\langle\Psi_1,\Psi_1\rangle\right) \\ & \times\langle\Psi_1,\Psi_1\rangle - \tfrac{1}{4}q_1\langle\Psi_2,\Psi_2\rangle + \left(\tfrac{3}{8}q_1^2 + \tfrac{1}{2}q_2\right)\left(-4p_2^2 - \tfrac{5}{64}q_1^4 + \tfrac{1}{4}q_2^2 - \tfrac{1}{2}\langle\Lambda\Psi_1,\Psi_1\rangle\right. \\ & \left. + \tfrac{1}{2}q_1\langle\Psi_1,\Psi_1\rangle\right) - \left(\tfrac{1}{4}q_2 + \tfrac{1}{16}q_1^2\right)\langle\Lambda\Psi_1,\Psi_1\rangle - 2p_2\langle\Psi_1,\Psi_2\rangle + 4(p_1 - q_1p_2)^2, \end{aligned} \tag{2.15}$$

$$F^{(i)}=(-2p_2\lambda_i+2q_1p_2-2p_1)\psi_{1i}\psi_{2i}-\tfrac{1}{2}(\lambda_i^2+\tfrac{1}{2}q_1\lambda_i+\tfrac{3}{8}q_1^2+\tfrac{1}{2}q_2)\psi_{2i}^2+\tfrac{1}{2}[\lambda_i^4-\tfrac{1}{2}q_1\lambda_i^3$$
$$-(\tfrac{1}{8}q_1^2+\tfrac{1}{2}q_2)\lambda_i^2+(\tfrac{1}{4}q_1^3+\tfrac{1}{2}q_1q_2-\tfrac{1}{2}\langle\Psi_1,\Psi_1\rangle)\lambda_i+\tfrac{1}{4}q_2^2-4p_2^2-\tfrac{5}{64}q_1^4-\tfrac{1}{2}\langle\Lambda\Psi_1,\Psi_1\rangle$$
$$+\tfrac{1}{2}q_1\langle\Psi_1,\Psi_1\rangle]\psi_{1i}^2+\tfrac{1}{4}\sum_{k\neq i}\frac{(\psi_{1i}\psi_{2k}-\psi_{1k}\psi_{2i})^2}{\lambda_k-\lambda_i},\quad i=1,...,N, \tag{2.16}$$

where $F^{(i)}$, $i=1,...,N$, F_1,F_2 are $N+2$ independent integrals of motion for (2.10). Notice that $\{A^2(\lambda)+B(\lambda)C(\lambda),\ A^2(\mu)+B(\mu)C(\mu)\}=0$, it can be shown that the equation (2.10) is a finite-dimensional integrable Hamiltonian system (FDIHS).

In order to find the QBH structure for (2.10), we need to use the modified system of (2.10). Let us consider the modified Jaulent–Miodek (MJM) spectral problem[31]

$$\phi_x=U(v,\lambda)\phi,\quad U(v,\lambda)=\begin{pmatrix} v_0 & \lambda \\ \lambda-v_1 & -v_0\end{pmatrix},\quad \phi=\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\quad v=\begin{pmatrix}v_0\\ v_1\end{pmatrix}. \tag{2.17}$$

Equations (2.2) and (2.3) yield

$$a_0=0,\quad b_0=1,\quad b_1=\tfrac{1}{2}v_1,\quad a_1=v_0,\quad c_0=1,\quad c_1=-\tfrac{1}{2}v_1,\ldots,$$
$$\begin{pmatrix}2a_{k+1}\\ -b_{k+1}\end{pmatrix}=L\begin{pmatrix}2a_k\\ -b_k\end{pmatrix},\quad k=1,2,...,$$
$$L=\begin{pmatrix}0 & -2v_0+D\\ \tfrac{1}{4}D+\tfrac{1}{2}D^{-1}v_0D & \tfrac{1}{2}v_1+\tfrac{1}{2}D^{-1}v_1D\end{pmatrix}. \tag{2.18}$$

The MJM hierarchy associated with (2.17) can also be written as a infinite-dimensional Hamiltonian system

$$v_{t_n}=\begin{pmatrix}v_0\\ v_1\end{pmatrix}_{t_n}=J\begin{pmatrix}2a_n\\ -b_n\end{pmatrix}=J\frac{\delta H_n}{\delta v},\quad n=1,2,..., \tag{2.19}$$

where the Hamiltonian H_n and the Hamiltonian operator J are given by

$$J=\begin{pmatrix}\tfrac{1}{2}D & 0\\ 0 & -2D\end{pmatrix},\quad H_n=\frac{-1}{n}[a_{n,x}-v_1b_n+2b_{n+1}].$$

Also we have

$$\frac{\delta\lambda}{\delta v}=\begin{pmatrix}2\phi_1\phi_2\\ \phi_1^2\end{pmatrix}. \tag{2.20}$$

In the similar way as for (2.7), the constrained flow of (2.19) is defined by

$$\Phi_{1,x}=v_0\Phi_1+\Lambda\Phi_2,\quad \Phi_{2,x}=(\Lambda-v_1)\Phi_1-v_0\Phi_2, \tag{2.21a}$$

$$\frac{\delta H_1}{\delta v}+\frac{1}{2}\begin{pmatrix}2\langle\Phi_1,\Phi_2\rangle\\ \langle\Phi_1,\Phi_1\rangle\end{pmatrix}=0, \tag{2.21b}$$

where $\Phi_i=(\phi_{i1},\ldots,\phi_{iN})^T$, $i=1,2$.

For $l=3$,

$$H_3=-(\tfrac{1}{4}v_{0x}^2-\tfrac{1}{16}v_{1x}^2+\tfrac{1}{4}v_0^4+\tfrac{5}{64}v_1^4-\tfrac{3}{8}v_{0x}v_1^2-\tfrac{3}{8}v_0^2v_1^2).$$

By introducing the Jacobi–Ostrogradsky coordinates

$$\tilde{q}_1 = v_1, \quad \tilde{q}_2 = v_0, \tag{2.22a}$$

$$\tilde{p}_1 = -\frac{\delta H_3}{\delta v_{1x}} = -\frac{1}{8} v_{1x}, \quad \tilde{p}_2 = -\frac{\delta H_3}{\delta v_{0x}} = \frac{1}{2} v_{0x} - \frac{3}{8} v_1^2, \tag{2.22b}$$

the equations (2.21) for $l=3$ are transformed into a FDHS

$$\Phi_{1,x} = \frac{\partial \tilde{F}_1}{\partial \Phi_2} = \tilde{q}_2 \Phi_1 + \Lambda \Phi_2, \quad \tilde{q}_{1x} = \frac{\partial \tilde{F}_1}{\partial \tilde{p}_1} = -8\tilde{p}_1, \quad \tilde{q}_{2x} = \frac{\partial \tilde{F}_1}{\partial \tilde{p}_2} = 2\tilde{p}_2 + \frac{3}{4}\tilde{q}_1^2, \tag{2.23a}$$

$$\Phi_{2,x} = -\frac{\partial \tilde{F}_1}{\partial \Phi_1} = \Lambda \Phi_1 - \tilde{q}_1 \Phi_1 - \tilde{q}_2 \Phi_2, \tag{2.23b}$$

$$\tilde{p}_{1x} = -\frac{\partial \tilde{F}_1}{\partial \tilde{q}_1} = -\frac{3}{2}\tilde{q}_1 \tilde{p}_2 - \frac{3}{4}\tilde{q}_1 \tilde{q}_2^2 - \frac{1}{4}\tilde{q}_1^3 - \frac{1}{2}\langle \Phi_1, \Phi_1 \rangle, \tag{2.23c}$$

$$\tilde{p}_{2x} = -\frac{\partial \tilde{F}_1}{\partial \tilde{q}_2} = \tilde{q}_2^3 - \frac{3}{4}\tilde{q}_1^2 \tilde{q}_2 - \langle \Phi_1, \Phi_2 \rangle, \tag{2.23d}$$

or

$$\tilde{P}_x = \theta_0 \nabla \tilde{F}_1,$$

where

$$\tilde{P} = (\Phi_1^T, \tilde{q}_1, \tilde{q}_2, \Phi_2^T, \tilde{p}_1, \tilde{p}_2)^T,$$

$$\tilde{F}_1 = -4\tilde{p}_1^2 + \tilde{p}_2^2 + \tfrac{3}{4}\tilde{q}_1^2 \tilde{p}_2 + \tfrac{3}{8}\tilde{q}_1^2 \tilde{q}_2^2 + \tfrac{1}{16}\tilde{q}_1^4 - \tfrac{1}{4}\tilde{q}_2^4 + \tilde{q}_2 \langle \Phi_1, \Phi_2 \rangle$$
$$+ \tfrac{1}{2}\langle \Lambda \Phi_2, \Phi_2 \rangle - \tfrac{1}{2}\langle \Lambda \Phi_1, \Phi_1 \rangle + \tfrac{1}{2}\tilde{q}_1 \langle \Phi_1, \Phi_1 \rangle.$$

B. The QBH structure for the FDIHS (2.10)

We now establish a map relating FDIHS (2.10) to (2.23), then use the map to construct the second compatible Poisson tensor for the FDIHS (2.10).

It is known[31] that a gauge transformation between the JM and MJM spectral problem is as follows

$$\psi_1 = \phi_1, \quad \psi_2 = \lambda \phi_2 + v_0 \phi_1, \quad u_1 = v_1, \quad u_0 = -v_{0x} - v_0^2, \tag{2.24}$$

which, together with (2.9) and (2.22), gives rise to the map relating (2.10) to (2.23), i.e., $P = M(\tilde{P})$:

$$\Psi_1 = \Phi_1, \quad \Psi_2 = \Lambda \Phi_2 + \tilde{q}_2 \Phi_1, \quad q_1 = \tilde{q}_1,$$
$$q_2 = -2\tilde{p}_2 - \tfrac{3}{4}\tilde{q}_1^2 - \tilde{q}_2^2, \quad p_1 = \tilde{q}_1 \tilde{p}_1 + \tfrac{1}{4}\tilde{q}_2^3 + \tfrac{1}{2}\tilde{q}_2 \tilde{p}_2 - \tfrac{1}{4}\langle \Phi_1, \Phi_2 \rangle, \quad p_2 = \tilde{p}_1. \tag{2.25}$$

The map M given by (2.25) transforms all equations in (2.10) except for (2.10c) into the corresponding equations in (2.23) except for (2.23c). In fact, the equation (2.10c) with an additive constant term $c = -\frac{1}{2}\tilde{F}_1$ is transformed into (2.23c) under the map (2.25). However, the second Poisson tensor constructed later by using the map (2.25) is valid for an arbitrary c, therefore we can take $c=0$. The Jacobi M' of the map M take the form

$$M'(\tilde{P})=\begin{pmatrix} I & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & -\frac{3}{2}\tilde{q}_1 & -2\tilde{q}_2 & 0 & 0 & -2\\ \tilde{q}_2 I & 0 & \Phi_1 & \Lambda & 0 & 0\\ -\frac{1}{4}\Phi_2^T & \tilde{p}_1 & \frac{3}{4}\tilde{q}_2^2+\frac{1}{2}\tilde{p}_2 & -\frac{1}{4}\Phi_1^T & \tilde{q}_1 & \frac{1}{2}\tilde{q}_2\\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{2.26}$$

According to the standard procedure,[32] the image of the Poisson tensor θ_0 for the FDIHS (2.23) under the map M gives rise to the second compatible Poisson tensor for the FDIHS (2.10). That is

$$\theta_1=M'\theta_0 M'^T|_{P=M(\tilde{P})}=\begin{pmatrix} 0_{(N+2)\times(N+2)} & A_1\\ -A_1^T & B_1 \end{pmatrix}, \tag{2.27a}$$

$$A_1=\begin{pmatrix} \Lambda & -\frac{1}{4}\Psi_1 & 0_{N\times 1}\\ 0_{1\times N} & q_1 & 1\\ 2\Psi_1^T & -\frac{1}{2}q_2-\frac{15}{8}q_1^2 & -\frac{3}{2}q_1 \end{pmatrix},\quad B_1=\begin{pmatrix} 0_{N\times N} & \frac{1}{4}\Psi_2 & 0_{N\times 1}\\ -\frac{1}{4}\Psi_2^T & 0 & p_2\\ 0_{1\times N} & -p_2 & 0 \end{pmatrix}. \tag{2.27b}$$

Furthermore, by a straightforward calculation, we can show the following proposition.

Proposition 1: The system (2.10) possesses the QBH *representation*

$$P_x=\theta_0\nabla F_1=\frac{1}{\rho}\,\theta_1\nabla E_1, \tag{2.28}$$

where

$$\rho=B(\lambda)|_{\lambda=0}=\tfrac{3}{8}q_1^2+\tfrac{1}{2}q_2-\tfrac{1}{2}\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle, \tag{2.29a}$$

$$\begin{aligned} E_1&=[A^2(\lambda)+B(\lambda)C(\lambda)]|_{\lambda=0}=F_2-\sum_{i=1}^{N}\lambda_i^{-1}F^{(i)}=(\tfrac{3}{16}q_1^2+\tfrac{1}{4}q_2)(\langle\Lambda^{-1}\Psi_2,\Psi_2\rangle-\langle\Lambda\Psi_1,\Psi_1\rangle)\\ &\times(\tfrac{3}{16}q_1^3+\tfrac{1}{4}q_1q_2)\langle\Psi_1,\Psi_1\rangle+2(p_1-q_1p_2)\langle\Lambda^{-1}\Psi_1,\Psi_2\rangle+(2p_2^2+\tfrac{5}{128}q_1^4-\tfrac{1}{8}q_2^2+\tfrac{1}{4}\langle\Lambda\Psi_1,\Psi_1\rangle\\ &-\tfrac{1}{4}q_1\langle\Psi_1,\Psi_1\rangle)\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle+(\tfrac{3}{8}q_1^2+\tfrac{1}{2}q_2)(\tfrac{1}{4}q_2^2-4p_2^2-\tfrac{5}{64}q_1^4)+4(p_1-q_1p_2)^2\\ &+\tfrac{1}{4}[\langle\Lambda^{-1}\Psi_1,\Psi_2\rangle^2-\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle\langle\Lambda^{-1}\Psi_2,\Psi_2\rangle]. \end{aligned} \tag{2.29b}$$

C. The Nijenhuis coordinates

We now prove that the Nijenhuis coordinates for QBH system (2.28) are the same as the separated variables defined by means of the Lax matrix (2.12b). As θ_0 and θ_1 are compatible and invertible, the matrix $\theta_1\theta_0^{-1}$ is maximal, it has $N+2$ distinct eigenvalues $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_{N+2})$. The explicit form of the canonical transformation from P to the Nijenhuis coordinates $(\boldsymbol{\mu},\boldsymbol{\nu})$ is given in what follows. The eigenvalues $\mu_1,\ldots,\mu_{N+2}$ are defined by the roots of the equation

$$f(\lambda)=|\lambda I-A_1|=0, \tag{2.30}$$

which, since A_1 depends only on (Ψ_1,q_1,q_2), gives rise to

$$\mu_j=f_j(\Psi_1,q_1,q_2),\quad j=1,\ldots,N+2, \tag{2.31}$$

$$\psi_{1j}=g_j(\boldsymbol{\mu}),\quad j=1,\ldots,N,\quad q_1=g_{N+1}(\boldsymbol{\mu}),\quad q_2=g_{N+2}(\boldsymbol{\mu}). \tag{2.32}$$

Then we introduce the generating function by

$$S=\sum_{j=1}^{N} \psi_{2j}g_j(\boldsymbol{\mu})+p_1 g_{N+1}(\boldsymbol{\mu})+p_2 g_{N+2}(\boldsymbol{\mu}), \tag{2.33a}$$

such that

$$\psi_{1j}=\frac{\partial S}{\partial \psi_{2j}}, \quad j=1,...,N, \quad q_1=\frac{\partial S}{\partial p_1}, \quad q_2=\frac{\partial S}{\partial p_2}, \tag{2.33b}$$

$$\nu_j=\frac{\partial S}{\partial \mu_j}=\sum_{j=1}^{N} \psi_{2j}\frac{\partial g_j}{\partial \mu_j}+p_1\frac{\partial g_{N+1}}{\partial \mu_j}+p_2\frac{\partial g_{N+2}}{\partial \mu_j}, \quad j=1,...,N+2. \tag{2.33c}$$

The equations (2.33b) reconstruct (2.31) or (2.32), the equations (2.33c) give the expression for ν_j. The system (2.10) written in terms of $(\boldsymbol{\mu},\boldsymbol{\nu})$ can be shown to be separable.

On the other hand, the separated variables $(\overline{\boldsymbol{\mu}},\overline{\boldsymbol{\nu}})$ for (2.10) can be constructed by means of the Lax matrix in the following way.[27,28] The coordinates $\overline{\mu}_1,\ldots,\overline{\mu}_{N+2}$ are introduced by the zeros of $B(\lambda)$:

$$B(\lambda)=\lambda^2+\frac{1}{2}q_1\lambda+\frac{3}{8}q_1^2+\frac{1}{2}q_2+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}=\frac{R(\lambda)}{K(\lambda)}, \tag{2.34}$$

with

$$R(\lambda)=\prod_{k=1}^{N+2}(\lambda-\overline{\mu}_k)=\sum_{k=0}^{N+2}\beta_k\lambda^{N+2-k}, \quad K(\lambda)=\prod_{k=1}^{N}(\lambda-\lambda_k)=\sum_{k=0}^{N}\alpha_k\lambda^{N-k}, \tag{2.35}$$

$$\alpha_0=1, \quad \alpha_1=-\sum_{j=1}^{N}\lambda_j, \quad \cdots, \quad \alpha_N=(-1)^N\prod_{j=1}^{N}\lambda_j,$$

$$\beta_0=1, \quad \beta_1=-\sum_{j=1}^{N+2}\overline{\mu}_j, \quad \cdots, \quad \beta_{N+2}=(-1)^N\prod_{j=1}^{N+2}\overline{\mu}_j,$$

and the canonically conjugate coordinates $\overline{\nu}_1,\ldots,\overline{\nu}_{N+2}$ are defined by

$$\overline{\nu}_k=-A(\overline{\mu}_k)=-2p_2\overline{\mu}_k-2p_1+2q_1p_2+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\overline{\mu}_k-\lambda_j}, \quad k=1,...,N+2. \tag{2.36}$$

The FDIHS (2.10) in terms of the coordinates $(\overline{\boldsymbol{\mu}},\overline{\boldsymbol{\nu}})$ will be shown to be separable later. We have the following proposition.

Proposition 2: The Nijenhuis coordinates $(\boldsymbol{\mu},\boldsymbol{\nu})$ *defined by (2.30) and (2.33) are exactly the same as the separated variables* $(\overline{\boldsymbol{\mu}},\overline{\boldsymbol{\nu}})$ *defined by (2.34) and (2.36). The* QBH *vector field (2.28) is Pfaffian in the Nijenhuis coordinates.*

Proof: We first show that

$$f(\lambda)=B(\lambda)K(\lambda)=R(\lambda). \tag{2.37}$$

We denote $f(\lambda)$ by $f_N(\lambda;\lambda_1,\ldots,\lambda_N)$ in order to prove (2.37) by induction. Obviously, (2.37) holds for $N=1$. Then we have by induction

$$f_N(\lambda;\lambda_1,\ldots,\lambda_N)=\begin{vmatrix} \lambda-\lambda_1 & 0 & \ldots & 0 & \frac{1}{4}\psi_{11} & 0\\ 0 & \lambda-\lambda_2 & \ldots & 0 & \frac{1}{4}\psi_{12} & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ 0 & 0 & \ldots & \lambda-\lambda_N & \frac{1}{4}\psi_{1N} & 0\\ 0 & 0 & \ldots & 0 & \lambda-q_1 & -1\\ -2\psi_{11} & -2\psi_{12} & \ldots & -2\psi_{1N} & \frac{1}{2}q_2+\frac{15}{8}q_1^2 & \lambda+\frac{3}{2}q_1 \end{vmatrix}$$

$$=(\lambda-\lambda_1)f_{N-1}(\lambda;\lambda_2,\ldots,\lambda_N)+\frac{1}{2}\psi_{11}^2\prod_{k=2}^{N}(\lambda-\lambda_k)$$

$$=\left(\lambda^2+\frac{1}{2}q_1\lambda+\frac{3}{8}q_1^2+\frac{1}{2}q_2+\frac{1}{2}\sum_{j=2}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}\right)K(\lambda)+\frac{1}{2}\frac{\psi_{11}^2}{\lambda-\lambda_1}K(\lambda)$$

$$=B(\lambda)K(\lambda). \tag{2.38}$$

Equation (2.38) implies that λ_1, similarly λ_k, $k=2,\ldots,N$, is not the zero of $f(\lambda)$. Thus (2.37) indicates that $f(\lambda)$ and $B(\lambda)$ have the same zeros, i.e., $\mu_k=\overline{\mu_k}$.

It follows from (2.34) that

$$\psi_{1j}^2=2\frac{R(\lambda_j)}{K'(\lambda_j)},\quad q_1=2(\beta_1-\alpha_1), \tag{2.39a}$$

$$\frac{1}{2}q_2+\frac{3}{8}q_1^2=\frac{1}{2}\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle+\frac{\beta_{N+2}}{\alpha_N}, \tag{2.39b}$$

where the prime denotes differentiation with respect to λ. The equations (2.39a) and (2.39b) yield

$$q_2=2\sum_{j=1}^{N}\frac{R(\lambda_j)}{\lambda_jK'(\lambda_j)}-3(\beta_1-\alpha_1)^2+2\frac{\beta_{N+2}}{\alpha_N}. \tag{2.39c}$$

According to (2.33a), one gets

$$S=\sum_{j=1}^{N}\psi_{2j}\sqrt{\frac{2R(\lambda_j)}{K'(\lambda_j)}}+2p_1(\beta_1-\alpha_1)+p_2\left[\sum_{j=1}^{N}\frac{2R(\lambda_j)}{\lambda_jK'(\lambda_j)}-3(\beta_1-\alpha_1)^2+2\frac{\beta_{N+2}}{\alpha_N}\right].$$

Notice that

$$\frac{\partial}{\partial\mu_k}\sum_{j=1}^{N}\psi_{2j}\sqrt{\frac{2R(\lambda_j)}{K'(\lambda_j)}}=\sum_{j=1}^{N}\frac{\psi_{2j}R(\lambda_j)}{\sqrt{2R(\lambda_j)K'(\lambda_j)}(\mu_k-\lambda_j)}=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\mu_k-\lambda_j},$$

$$\frac{\partial}{\partial\mu_k}\sum_{j=1}^{N}\frac{2R(\lambda_j)}{\lambda_jK'(\lambda_j)}=\sum_{j=1}^{N}\frac{2R(\lambda_j)}{\lambda_jK'(\lambda_j)(\mu_k-\lambda_j)}=\frac{1}{\mu_k}\sum_{j=1}^{N}\left[\frac{2R(\lambda_j)}{\lambda_jK'(\lambda_j)}+\frac{2R(\lambda_j)}{(\mu_k-\lambda_j)K'(\lambda_j)}\right]$$

$$=\frac{1}{\mu_k}\left(\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle+\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\mu_k-\lambda_j}\right),$$

$$\frac{\partial\beta_{N+2}}{\partial\mu_k}=\frac{\beta_{N+2}}{\mu_k},\quad \frac{\partial(\beta_1-\alpha_1)^2}{\partial\mu_k}=-q_1,\quad \frac{\partial(\beta_1-\alpha_1)}{\partial\mu_k}=-1,$$

and using $B(\mu_k)=0$, one finds from (2.33c) that ν_k,μ_k satisfy (2.36). Finally, it follows from (2.39b) that

$$\rho=B(\lambda)|_{\lambda=0}=\frac{1}{2}q_2+\frac{3}{8}q_1^2-\frac{1}{2}\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle=\frac{\beta_{N+2}}{\alpha_N}=\frac{(-1)^N}{\alpha_N}\prod_{j=1}^{N+2}\mu_j. \tag{2.40}$$

This completes the proof.

D. Comparison of the two methods for separability

For the FDIHS with QBH structure, the separated variables, i.e., the Nijenhuis coordinates, can be introduced by the Nijenhuis tensor. Then the separability of the Hamilton–Jacobi equation for the system can be shown by verifying the Levi-Civita conditions. For the FDIHS with some kind of Lax representation, the separated variables can be found and the separability of the Hamilton Jacobi equation for the system can be shown by means of the Lax representation. So far there is not an effective way to define separated variables for the FDIHSs with some kind of Lax matrices, such as the 3×3 Lax matrices[22] or the Lax matrices admitting dynamical r-matrix. However, if the separated variables can be introduced by the Lax matrix, one can further establish the Jacobi inversion problem for the system by means of the Lax representation. By using the standard Jacobi inversion technique, the solution to the system can be found.

We now use the Lax representation (2.12) to construct the Jacobi inversion problem for (2.10). Set

$$A^2(\lambda)+B(\lambda)C(\lambda)=\frac{W(\lambda)}{K(\lambda)},\quad W(\lambda)=\sum_{i=0}^{N+6}P_i\lambda^i, \tag{2.41}$$

then P_i are also the integrals of motion for (2.10). By substituting (2.13) and using (2.14), (2.41) leads to

$$P_{N+6}=1,\quad P_{N+6-i}=\alpha_i,\quad i=1,2,3,4,$$
$$F_1=-P_{N+1}+\alpha_5,\quad F_2=P_N-\alpha_1P_{N+1}+\alpha_1\alpha_5-\alpha_6,\dots. \tag{2.42}$$

The equations (2.34), (2.36), and (2.41) give rise to

$$\nu_k=\sqrt{\frac{W(\mu_k)}{K(\mu_k)}},\quad k=1,\dots,N+2, \tag{2.43}$$

which indicates that the Hamilton–Jacobi equation is separable. Replacing ν_k by $\partial S_k/\partial\mu_k$ and interpreting the P_i as integration constants, one gets the generating function S of the canonical transformation from (2.43)

$$S(\mu_1,\dots,\mu_{N+2};P_0,\dots,P_{N+1})=\sum_{k=1}^{N+2}\int^{\mu_k}\sqrt{\frac{W(\lambda)}{K(\lambda)}}\,d\lambda. \tag{2.44}$$

The linearizing coordinates are then

$$Q_i=\frac{\partial S}{\partial P_i}=\frac{1}{2}\sum_{k=1}^{N+2}\int^{\mu_k}\frac{\lambda^i}{\sqrt{W(\lambda)K(\lambda)}}\,d\lambda,\quad i=0,1,\dots,N+1. \tag{2.45}$$

The linear flow induced by (2.10) is then given by [using (2.42)]

$$Q_i=\gamma_i+x\frac{\partial F_1}{\partial P_i}=\gamma_i-x\,\delta_{i,N+1},\quad i=0,1,\dots,N+1, \tag{2.46}$$

where γ_i are arbitrary constants. Combining the equation (2.45) with the equation (2.46) leads to the Jacobi inversion problem for the FDIHS (2.10)

$$\frac{1}{2}\sum_{k=1}^{N+2}\int^{\mu_k}\frac{\lambda^i}{\sqrt{W(\lambda)K(\lambda)}}d\lambda=\gamma_i-x\,\delta_{i,N+1},\quad i=0,1,\ldots,N+1. \tag{2.47}$$

Since ψ_{1j},q_1,q_2 defined by (2.39) are the symmetric functions of μ_k, $k=1,\ldots,N+2$, by using the standard Jacobi inversion technique,[33] they can be solved in terms of Riemann theta functions from (2.47). After having ψ_{1j},q_1,q_2, the ψ_{2j},p_1,p_2 can be found by using (2.10a). In this way the solution to (2.10) is obtained.

III. THE FIRST FAMILY OF QBH SYSTEMS

In the following sections, by using the method described in the preceding section, we will present QBH representation for some families of FDIHSs given in Ref. 24, and prove the equivalence of two sets of separated variables.

A. The first family of FDIHSs

We first recall the constrained flows of the hierarchy of nonlinear evolution equations (NLEE) associated with the following polynomial second order spectral problem[31]

$$\psi_x=U(u,\lambda)\psi,\quad U(u,\lambda)=\begin{pmatrix}0 & 1\\ -\sum_{i=0}^{m}u_i\lambda^i & 0\end{pmatrix},\quad \psi=\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}, \tag{3.1}$$

where $u_m=-1$, $u=(u_{m-1},\ldots,u_0)^T$. The adjoint representation (2.2) of (3.1) yields

$$a_0=\cdots=a_m=b_0=\cdots=b_{m-1}=0,\quad b_m=1,\quad b_{m+1}=\tfrac{1}{2}u_{m-1},$$

$$a_{m+1}=-\tfrac{1}{4}u_{m-1,x},\quad c_0=1,\quad c_1=-\tfrac{1}{2}u_{m-1},\ldots,$$

and in general

$$\begin{pmatrix}b_{k+m}\\ \vdots\\ b_{k+1}\end{pmatrix}=L\begin{pmatrix}b_{k+m-1}\\ \vdots\\ b_k\end{pmatrix}, \tag{3.2a}$$

$$a_k=-\tfrac{1}{2}b_{k,x},\quad c_k=-\tfrac{1}{2}b_{k,xx}-\sum_{i=0}^{m}u_ib_{k+i},\quad k=1,2,\ldots, \tag{3.2b}$$

where

$$L=\begin{pmatrix}L_{m-1} & L_{m-2} & \ldots & L_1 & L_0\\ 1 & 0 & \ldots & 0 & 0\\ 0 & 1 & \ldots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \ldots & 1 & 0\end{pmatrix},$$

$$L_0=\tfrac{1}{4}D^2+u_0-\tfrac{1}{2}D^{-1}u_{0,x},\quad L_i=u_i-\tfrac{1}{2}D^{-1}u_{i,x},\quad i=1,\ldots,m-1.$$

The hierarchy of NLEEs associated with (3.1) can be written as an infinite-dimensional Hamiltonian system

$$u_{t_n}=\begin{pmatrix} u_{m-1}\\ \vdots \\ u_0\end{pmatrix}_{t_n}=J\begin{pmatrix} b_{n+m}\\ \vdots \\ b_{n+1}\end{pmatrix}=J\,\frac{\delta H_n}{\delta u},\quad n=1,2,..., \tag{3.3}$$

where the Hamiltonian H_n and the Hamiltonian operator J are defined by

$$J=\begin{pmatrix} 0 & 0 & \dots & 0 & 2D\\ 0 & 0 & \dots & 2D & J_{m-1}\\ 0 & 0 & \dots & J_{m-1} & J_{m-2}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 2D & J_{m-1} & \dots & J_1 & J_0\end{pmatrix},$$

$$J_i=-u_{i,x}-2u_iD,\quad i=0,1,...,m-1,\quad H_n=\frac{2}{m-2n-2}\sum_{i=1}^{m} iu_ib_{n+i+1}.$$

Under zero boundary condition we have

$$\frac{\delta\lambda}{\delta u}=(\lambda^{m-1}\psi_1^2,\lambda^{m-2}\psi_1^2,\dots,\psi_1^2)^T,\quad L\,\frac{\delta\lambda}{\delta u}=\lambda\,\frac{\delta\lambda}{\delta u}. \tag{3.4}$$

Similarly, the constrained flows of the NLEEs (3.3) are defined by[24]

$$\Psi_{1,x}=\Psi_2,\quad \Psi_{2,x}=\Lambda^m\Psi_1-\sum_{i=0}^{m-1} u_i\Lambda^i\Psi_1, \tag{3.5a}$$

$$\frac{\delta H_l}{\delta u}-\frac{1}{2}\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta u}=\begin{pmatrix} b_{m+l}\\ \vdots\\ b_{l+1}\end{pmatrix}-\frac{1}{2}\begin{pmatrix}\langle\Lambda^{m-1}\Psi_1,\Psi_1\rangle\\ \vdots\\ \langle\Psi_1,\Psi_1\rangle\end{pmatrix}=0. \tag{3.5b}$$

For $l=m$, (3.5b) leads to

$$u_{m-k}=\sum_{j=1}^{k}(-1)^{j-1}\frac{j+1}{2^j}\sum_{l_1+\dots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,\quad k=1,...,m, \tag{3.6}$$

where $l_1\geqslant 0,...,l_j\geqslant 0$. By substituting (3.6) into (3.5a), the first constrained flow of (3.3) can be written as a canonical FDHS

$$\Psi_{1,x}=\frac{\partial F_0}{\partial\Psi_2},\quad \Psi_{2,x}=-\frac{\partial F_0}{\partial\Psi_1}, \tag{3.7}$$

or

$$P_x=\theta_0\nabla F_0,$$

where

$$P=(\Psi_1^T,\Psi_2^T)^T,\quad \theta_0=\begin{pmatrix} 0 & I_{N\times N}\\ -I_{N\times N} & 0\end{pmatrix},$$

$$F_0=\frac{1}{2}\langle\Psi_2,\Psi_2\rangle+\sum_{j=0}^{m}\left(\frac{-1}{2}\right)^{j+1}\sum_{l_1+\cdots+l_{j+1}=m-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_{j+1}}\Psi_1,\Psi_1\rangle.$$

The entries of the Lax matrix for (3.7) are given by[24]

$$A(\lambda)=-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\lambda-\lambda_j},\quad B(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}, \tag{3.8a}$$

$$C(\lambda)=\lambda^m+\sum_{k=1}^{m}\lambda^{m-k}\sum_{j=1}^{k}\left(-\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}^2}{\lambda-\lambda_j}. \tag{3.8b}$$

We have

$$A(\lambda)^2+B(\lambda)C(\lambda)=\lambda^m+\sum_{i=1}^{N}\frac{F^{(i)}}{\lambda-\lambda_i},$$

$$F^{(i)}=\frac{1}{2}\left[\lambda_i^m+\sum_{k=1}^{m}\lambda_i^{m-k}\sum_{j=1}^{k}\left(-\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle\right]\psi_{1i}^2$$
$$-\frac{1}{2}\psi_{2i}^2+\frac{1}{4}\sum_{k\neq i}\frac{(\psi_{1i}\psi_{2k}-\psi_{1k}\psi_{2i})^2}{\lambda_k-\lambda_i},\quad i=1,\ldots,N, \tag{3.9}$$

where $F^{(i)}$, $i=1,\ldots,N$, are independent integrals of motion for (3.7) and $F_0=\Sigma_{i=0}^{N}F^{(i)}$. It can be shown that the system (3.7) is integrable in the Liouville's sense. The systems with $m=1,2,\ldots$ give rise to a family of FDIHSs which include the well-known Garnier system as the first member $(m=1)$. This family of FDIHSs was first given in Ref. 34.

In order to find the QBH structure for (3.7), we need to consider the following modified polynomial second-order spectral problem[31]

$$\phi_x=U(v,\lambda)\phi,\quad U(v,\lambda)=\begin{pmatrix} v_0 & \lambda\\ -\sum_{i=1}^{m}v_i\lambda^{i-1} & -v_0\end{pmatrix},\quad \phi=\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}, \tag{3.10}$$

where $v_m=-1$, $v=(v_0,\ldots,v_{m-1})^T$. The equations (2.2) and (2.3) yield

$$a_0=\cdots=a_{m-2}=b_0=\cdots=b_{m-3}=0,\quad b_{m-2}=1,\quad b_{m-1}=\tfrac{1}{2}v_{m-1},$$

$$a_{m-1}=v_0,\quad c_0=1,\quad c_1=-\tfrac{1}{2}v_{m-1},\ldots,$$

and in general

$$\begin{pmatrix}2a_{k+1}\\ -b_{k+1}\\ \vdots\\ -b_{k+m-1}\end{pmatrix}=L\begin{pmatrix}2a_k\\ -b_k\\ \vdots\\ -b_{k+m-2}\end{pmatrix}, \tag{3.11a}$$

$$c_{k+1}=a_{k,x}-\sum_{i=1}^{m}v_ib_{k+i-1},\quad k=1,2,\ldots, \tag{3.11b}$$

where

$$L=\begin{pmatrix} 0 & -2v_0+D & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ L_0 & L_1 & L_2 & \dots & L_{m-2} & L_{m-1} \end{pmatrix},$$

$$L_0=\tfrac{1}{4}D+\tfrac{1}{2}D^{-1}v_0D, \quad L_i=\tfrac{1}{2}v_i+\tfrac{1}{2}D^{-1}v_iD, \quad i=1,...,m-1.$$

The hierarchy of NLEEs associated with (3.10) is

$$v_{t_n}=\begin{pmatrix} v_0 \\ \vdots \\ v_{m-1} \end{pmatrix}_{t_n}=J\begin{pmatrix} 2a_n \\ -b_n \\ \vdots \\ -b_{n+m-2} \end{pmatrix}=J\frac{\delta H_n}{\delta u}, \quad n=1,2,..., \tag{3.12}$$

where the Hamiltonian H_n and the Hamiltonian operator J are given by

$$J=\begin{pmatrix} \frac{1}{2}D & 0 & 0 & \dots & 0 & 0 \\ 0 & J_2 & J_3 & \dots & J_{m-1} & -2D \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & J_{m-1} & -2D & \dots & 0 & 0 \\ 0 & -2D & 0 & \dots & 0 & 0 \end{pmatrix},$$

$$J_i=v_{i,x}+2v_iD, \quad i=0,1,...,m-1, \quad H_n=\frac{2}{m-2n-2}\left[a_{n,x}-\sum_{i=1}^{m} iv_ib_{n+i-1}\right].$$

Also we have

$$\frac{\delta\lambda}{\delta u}=(2\phi_1\phi_2,\phi_1^2,\lambda\phi_1^2,\dots,\lambda^{m-2}\phi_1^2)^T. \tag{3.13}$$

The constrained flows of (3.12) are defined by

$$\Phi_{1,x}=v_0\Phi_1+\Lambda\Phi_2, \quad \Phi_{2,x}=\left(\Lambda^{m-1}-\sum_{i=1}^{m-1} v_i\Lambda^{i-1}\right)\Phi_1-v_0\Phi_2, \tag{3.14a}$$

$$\frac{\delta H_l}{\delta v}+\frac{1}{2}\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta v}=\begin{pmatrix} 2a_l \\ -b_l \\ \vdots \\ -b_{l+m-2} \end{pmatrix}+\frac{1}{2}\begin{pmatrix} 2\langle\Phi_1,\Phi_2\rangle \\ \langle\Phi_1,\Phi_1\rangle \\ \vdots \\ \langle\Lambda^{m-2}\Phi_1,\Phi_1\rangle \end{pmatrix}=0. \tag{3.14b}$$

For $l=m-1$, (3.14b) leads to

$$v_0=-\tfrac{1}{2}\langle\Phi_1,\Phi_2\rangle, \tag{3.15a}$$

$$v_{m-k}=\sum_{j=1}^{k}(-1)^{j-1}\frac{j+1}{2^j}\sum_{l_1+\dots+l_j=k-j}\langle\Lambda^{l_1}\Phi_1,\Phi_1\rangle\dots\langle\Lambda^{l_j}\Phi_1,\Phi_1\rangle, \quad k=1,...,m-1. \tag{3.15b}$$

By substituting (3.15) into (3.14a), the first constrained flow of NLEE (3.12) can be written as a canonical FDHS

$$\Phi_{1,x}=\frac{\partial\widetilde{F}_0}{\partial\Phi_2},\quad \Phi_{2,x}=-\frac{\partial\widetilde{F}_0}{\partial\Phi_1}, \tag{3.16a}$$

or

$$\widetilde{P}_x=\theta_0\nabla\widetilde{F}_0,$$

where

$$\widetilde{P}=(\Phi_1^T,\Phi_2^T)^T,\quad \theta_0=\begin{pmatrix}0 & I_{N\times N}\\ -I_{N\times N} & 0\end{pmatrix},$$

$$\widetilde{F}_0=\tfrac{1}{2}\langle\Lambda\Phi_2,\Phi_2\rangle-\tfrac{1}{4}\langle\Phi_1,\Phi_2\rangle^2+\sum_{j=1}^{m}(-\tfrac{1}{2})^j\sum_{l_1+\ldots+l_j=m-j}\langle\Lambda^{l_1}\Phi_1,\Phi_1\rangle\cdots\langle\Lambda^{l_j}\Phi_1,\Phi_1\rangle. \tag{3.16b}$$

B. The QBH structure for the family of FDIHD (3.7)

It is known[31] that the gauge transformation between the spectral problems (3.1) and (3.10) is given by

$$\psi_1=\phi_1,\quad \psi_2=\lambda\phi_2+v_0\phi_1,$$
$$u_i=v_i,\quad i=1,...,m-1,\quad u_0=-v_{0x}-v_0^2, \tag{3.17}$$

which, together with (3.6) and (3.15), gives rise to the map relating (3.7) to (3.16), i.e., $P=M(\widetilde{P})$:

$$\Psi_1=\Phi_1,\quad \Psi_2=\Lambda\Phi_2-\tfrac{1}{2}\langle\Phi_1,\Phi_2\rangle\Phi_1. \tag{3.18}$$

In fact the map M transforms the first equation and the second equation with an additive term, $-c\Psi_1(c=\widetilde{F}_0)$, in (3.7) into the corresponding equations in (3.16). Since the θ_1 constructed in the following is valid for an arbitrary c, so we can take $c=0$. The Jacobi M' of the map M takes the form

$$M'(\widetilde{P})=\begin{pmatrix}I_{N\times N} & 0_{N\times N}\\ -\tfrac{1}{2}\langle\Phi_1,\Phi_2\rangle I_{N\times N}-\tfrac{1}{2}\Phi_1\Phi_2^T & \Lambda-\tfrac{1}{2}\Phi_1\Phi_1^T\end{pmatrix}. \tag{3.19}$$

Then the second compatible Poisson tensor for the vector field (3.7) is

$$\theta_1=M'\theta_0M'^T|_{P=M(\widetilde{P})}=\begin{pmatrix}0_{N\times N} & A_1\\ -A_1^T & B_1\end{pmatrix},$$
$$A_1=\Lambda-\tfrac{1}{2}\Psi_1\Psi_1^T,\quad B_1=\tfrac{1}{2}\Psi_2\Psi_1^T-\tfrac{1}{2}\Psi_1\Psi_2^T. \tag{3.20}$$

By a straightforward calculation, we have the following proposition.

Proposition 3: The system (3.7) possesses the QBH *representation*

$$P_x=\theta_0\nabla F_0=\frac{1}{\rho}\theta_1\nabla E_1, \tag{3.21a}$$

where

$$\rho = B(\lambda)|_{\lambda=0} = 1 - \tfrac{1}{2}\langle \Lambda^{-1}\Psi_1, \Psi_1\rangle, \tag{3.21b}$$

$$\begin{aligned} E_1 &= [A^2(\lambda) + B(\lambda)C(\lambda)]|_{\lambda=0} \\ &= -\sum_{i=1}^{N} \lambda_i^{-1} F^{(i)} \\ &= \tfrac{1}{2}\langle \Lambda^{-1}\Psi_2, \Psi_2\rangle + \tfrac{1}{4}[\langle \Lambda^{-1}\Psi_1, \Psi_2\rangle^2 - \langle \Lambda^{-1}\Psi_1, \Psi_1\rangle\langle \Lambda^{-1}\Psi_2, \Psi_2\rangle] \\ &\quad + \sum_{j=0}^{m} (-\tfrac{1}{2})^{j+1} \sum_{l_1+\ldots+l_{j+1}=m-j} \langle \Lambda^{l_1}\Psi_1, \Psi_1\rangle \ldots \langle \Lambda^{l_{j+1}-1}\Psi_1, \Psi_1\rangle. \end{aligned} \tag{3.21c}$$

C. The Nijenhuis coordinates

In the same way as for (2.30)–(2.33), the eigenvalues of the Nijenhuis tensor $\mu_1, \ldots, \mu_N$ are defined by the roots of the equation

$$f(\lambda) = |\lambda I - A_1| = 0, \tag{3.22a}$$

which gives

$$\psi_{1j} = g_j(\boldsymbol{\mu}), \quad j = 1,\ldots,N.$$

Then one defines

$$\nu_j = \frac{\partial S}{\partial \mu_j} = \sum_{j=1}^{N} \psi_{2j} \frac{\partial g_j}{\partial \mu_j}, \quad j = 1,\ldots,N. \tag{3.22b}$$

On the other hand, the generalized elliptic coordinates $(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\nu}})$ are defined by means of the Lax matrix in the following way.[24] The coordinates $\overline{\mu}_1, \ldots, \overline{\mu}_N$ are introduced by the zeros of $B(\lambda)$:

$$B(\lambda) = 1 + \frac{1}{2}\sum_{j=1}^{N} \frac{\psi_{1j}^2}{\lambda - \lambda_j} = \frac{R(\lambda)}{K(\lambda)}, \tag{3.23a}$$

where $K(\lambda)$ is defined by (2.35) and

$$R(\lambda) = \prod_{k=1}^{N} (\lambda - \overline{\mu}_k) = \sum_{k=0}^{N} \beta_k \lambda^{N-k},$$

$$\beta_0 = 1, \quad \beta_1 = -\sum_{j=1}^{N} \overline{\mu}_j, \ldots, \quad \beta_N = (-1)^N \prod_{j=1}^{N} \overline{\mu}_j, \tag{3.23b}$$

and the canonically conjugate coordinates $\overline{\nu}_1, \ldots, \overline{\nu}_N$ are defined by

$$\overline{\nu}_k = -A(\overline{\mu}_k) = \frac{1}{2}\sum_{j=1}^{N} \frac{\psi_{1j}\psi_{2j}}{\overline{\mu}_k - \lambda_j}, \quad k = 1,\ldots,N. \tag{3.23c}$$

We have the following proposition.

Proposition 4: The Nijenhuis coordinates $(\boldsymbol{\mu}, \boldsymbol{\nu})$ *defined by (3.22) are exactly the same as the generalized elliptic coordinates* $(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\nu}})$ *defined by (3.23). The* QBH *vector field (3.21) is Pfaffian in the Nijenhuis coordinates.*

Proof: Similarly, we have by induction

$$f_N(\lambda;\lambda_1,\ldots,\lambda_N)=|\lambda I-A_1|$$
$$=\begin{vmatrix} \lambda-\lambda_1+\frac{1}{2}\psi_{11}^2 & \frac{1}{2}\psi_{11}\psi_{12} & \cdots & \frac{1}{2}\psi_{11}\psi_{1N} \\ \frac{1}{2}\psi_{12}\psi_{11} & \lambda-\lambda_2+\frac{1}{2}\psi_{12}^2 & \cdots & \frac{1}{2}\psi_{12}\psi_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\psi_{1N}\psi_{11} & \frac{1}{2}\psi_{1N}\psi_{12} & \cdots & \lambda-\lambda_N+\frac{1}{2}\psi_{1N}^2 \end{vmatrix}$$
$$=(\lambda-\lambda_1)f_{N-1}(\lambda;\lambda_2,\ldots,\lambda_N)+\tfrac{1}{2}\psi_{11}^2\prod_{k=2}^{N}(\lambda-\lambda_k)$$
$$=B(\lambda)K(\lambda), \tag{3.24}$$

which shows that $\mu_k=\overline{\mu_k}$. It follows from (3.23a) that

$$\psi_{1j}^2=2\,\frac{R(\lambda_j)}{K'(\lambda_j)},\quad j=1,\ldots,N. \tag{3.25}$$

Thus we have

$$\nu_k=\sum_{j=1}^{N}\psi_{2j}\frac{\partial}{\partial\mu_k}\sqrt{\frac{2R(\lambda_j)}{K'(\lambda_j)}}=\sum_{j=1}^{N}\frac{\psi_{2j}R(\lambda_j)}{\sqrt{2R(\lambda_j)K'(\lambda_j)}(\mu_k-\lambda_j)}=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\mu_k-\lambda_j}, \tag{3.26}$$

which implies that $\nu_k=\overline{\nu_k}$, since $\mu_k=\overline{\mu_k}$. Finally, it is found from (3.23a) that

$$\rho=B(\lambda)|_{\lambda=0}=1-\frac{1}{2}\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle=\frac{\beta_N}{\alpha_N}.$$

This completes the proof.

IV. THE SECOND FAMILY OF QBH SYSTEMS

For $l=m+1$, it is found from (3.5b)[24] that

$$u_{m-k}=(-\tfrac{1}{2})^k u_{m-1}^k+\sum_{i=0}^{k-2}u_{m-1}^i\sum_{j=1}^{[(k-i)/2]}E_{i,j}\sum_{l_1+\ldots+l_j=k-i-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,$$
$$k=2,\ldots,m, \tag{4.1a}$$

$$L_0u_{m-1}=\langle\Lambda^{m-1}\Psi_1,\Psi_1\rangle-\sum_{i=1}^{m-1}L_i\langle\Lambda^{i-1}\Psi_1,\Psi_1\rangle, \tag{4.1b}$$

where

$$E_{i,j}=-(i+j+1)\beta_{i,j},\quad \beta_{i,j}=\left(-\frac{1}{2}\right)^{i+j}\frac{(i+j)!}{i!j!}.$$

Denote

$$q=u_{m-1},\quad p=-\tfrac{1}{8}u_{m-1,x}.$$

By substituting (4.1a), (3.5a) and (4.1b) become a canonical FDHS

$$P_x=\theta_0\nabla F_1, \tag{4.2a}$$

where

$$P=(\Psi_1^T,q,\Psi_2^T,p)^T, \quad \theta_0=\begin{pmatrix} 0 & I_{(N+1)\times(N+1)} \\ -I_{(N+1)\times(N+1)} & 0 \end{pmatrix},$$

$$F_1=\tfrac{1}{2}\langle\Psi_2,\Psi_2\rangle+(-\tfrac{1}{2}q)^{m+2}-4p^2 + \sum_{i=0}^{m} q^i \sum_{j=1}^{[(m+2-i)/2]} \beta_{i,j} \sum_{l_1+\cdots+l_j=m+2-i-2j} \langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle. \tag{4.2b}$$

The entries of the Lax matrix Q for (4.2) are of the form[24]

$$A(\lambda)=2p-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\lambda-\lambda_j}, \quad B(\lambda)=\lambda+\frac{1}{2}q+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}, \tag{4.3a}$$

$$C(\lambda)=\sum_{k=0}^{m+1}\lambda^{m+1-k}\tilde{c}_k-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}^2}{\lambda-\lambda_j}, \tag{4.3b}$$

where

$$\tilde{c}_k=\left(-\frac{1}{2}q\right)^k+\sum_{i=0}^{k-2}q^i\sum_{j=1}^{[(k-i)/2]}\beta_{i,j}\sum_{l_1+\ldots+l_j=k-i-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\ldots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,$$

$$k=1,\ldots,m+1, \tag{4.3c}$$

$$\tilde{c}_{m+2+k}=-\tfrac{1}{2}\langle\Lambda^k\Psi_2,\Psi_2\rangle, \quad k=0,1,\ldots. \tag{4.3d}$$

Similarly, the equality

$$A^2(\lambda)+B(\lambda)C(\lambda)=\lambda^{m+2}-F_1+\sum_{i=1}^{N}\frac{F^{(i)}}{\lambda-\lambda_i},$$

$$F^{(i)}=-2p\psi_{1i}\psi_{2i}-\frac{1}{2}\left(\lambda_i+\frac{1}{2}q\right)\psi_{2i}^2+\frac{1}{2}\sum_{k=0}^{m+1}\tilde{c}_k\lambda_i^{m+1-k}\psi_{1i}^2+\frac{1}{4}\sum_{k\neq i}\frac{(\psi_{1i}\psi_{2k}-\psi_{1k}\psi_{2i})^2}{\lambda_k-\lambda_i}, \tag{4.4}$$

$$i=1,\ldots,N,$$

determines $N+1$ independent integrals of motion F_0, $F^{(i)}$, $i=1,\ldots,N$, for the FDHS (4.2). The systems (4.2) for $m=1,2,\ldots$, give the second family of FDIHSs. By taking $m=1$ (4.2) gives rises to the multidimensional Henon–Heiles system. The system (4.2) was also studied by a recurrence relation in Ref. 35, however no explicit expressions like (4.2b) and (4.3) were given in that paper.

In the exactly the same way as we did in the preceding section, we can obtain another FDHS from (3.14) for $l=m$, find the map relating this FDHS to the FDHS (4.2) and finally, by using this map, obtain the second compatible Poisson tensor for the vector field for (4.2)

$$\theta_1=\begin{pmatrix} 0_{(N+1)\times(N+1)} & A_1 \\ -A_1^T & B_1 \end{pmatrix},$$

$$A_1=\begin{pmatrix} \Lambda & -\frac{1}{4}\Psi_1 \\ 2\Psi_1^T & -\frac{1}{2}q \end{pmatrix}, \quad B_1=\begin{pmatrix} 0_{N\times N} & \frac{1}{4}\Psi_2 \\ -\frac{1}{4}\Psi_2^T & 0 \end{pmatrix}. \tag{4.5}$$

By a straightforward calculation, we can show the following proposition.

Proposition 5: The system (4.2) possesses the QBH *representation*

$$P_x=\theta_0\nabla F_1=\frac{1}{\rho}\,\theta_1\nabla E_1, \tag{4.6}$$

where

$$\rho=B(\lambda)|_{\lambda=0}=\tfrac{1}{2}q-\tfrac{1}{2}\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle, \tag{4.7a}$$

$$\begin{aligned}E_1&=[A^2(\lambda)+B(\lambda)C(\lambda)]|_{\lambda=0}\\ &=-F_1-\sum_{i=1}^{N}\lambda_i^{-1}F^{(i)}=2p\langle\Lambda^{-1}\Psi_1,\Psi_2\rangle+\frac{1}{4}\,q\langle\Lambda^{-1}\Psi_2,\Psi_2\rangle+4p^2-\left(-\frac{1}{2}\,q\right)^{m+2}\\ &\quad-\sum_{i=0}^{m}q^i\sum_{j=1}^{[(m+2-i)/2]}\beta_{i,j}\sum_{l_1+\ldots+l_j=m+2-i-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle\\ &\quad+\frac{1}{4}[\langle\Lambda^{-1}\Psi_1,\Psi_2\rangle^2-\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle\langle\lambda^{-1}\Psi_2,\Psi_2\rangle]\\ &\quad-\frac{1}{2}\sum_{i=0}^{m+1}q^i\sum_{j=0}^{[(m+1-i)/2]}\beta_{i,j}\sum_{l_1+\ldots+l_{j+1}=m+1-i-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle\\ &\quad\times\langle\Lambda^{l_{j+1}-1}\Psi_1,\Psi_1\rangle.\end{aligned} \tag{4.7b}$$

In the same way, $\mu_1,\ldots,\mu_{N+1}$ in the Nijenhuis coordinates are defined by the roots of the equation

$$f(\lambda)=|\lambda I-A_1|=0, \tag{4.8a}$$

which gives

$$\psi_{1j}=g_j(\boldsymbol{\mu}),\quad j=1,\ldots N,\quad q=g_{N+1}(\boldsymbol{\mu}).$$

Then one defines

$$\nu_j=\frac{\partial S}{\partial\mu_j}=\sum_{j=1}^{N}\psi_{2j}\frac{\partial g_j}{\partial\mu_j}+p\,\frac{\partial g_{N+1}}{\partial\mu_j},\quad j=1,\ldots,N+1. \tag{4.8b}$$

On the other hand, the generalized parabolic coordinates $(\overline{\boldsymbol{\mu}},\overline{\boldsymbol{\nu}})$ are defined by means of the Lax matrix in the following way.[24] The coordinates $\overline{\mu}_1,\ldots,\overline{\mu}_{N+1}$ are introduced by the zeros of $B(\lambda)$:

$$B(\lambda)=\lambda+\frac{1}{2}q+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}=\frac{R(\lambda)}{K(\lambda)}, \tag{4.9a}$$

where $K(\lambda)$ is defined by (2.35) and $R(\lambda)$ by

$$R(\lambda)=\prod_{k=1}^{N+1}(\lambda-\overline{\mu}_k)=\sum_{k=0}^{N+1}\beta_k\lambda^{N+1-k},$$

$$\beta_0=1,\quad\beta_1=-\sum_{j=1}^{N}\overline{\mu}_j,\ldots,\quad\beta_{N+1}=(-1)^{N+1}\prod_{j=1}^{N+1}\overline{\mu}_j,$$

and the canonically conjugate coordinates $\overline{\nu}_1,\dots,\overline{\nu}_{N+1}$ are defined by

$$\overline{\nu}_k=-A(\overline{\mu}_k)=-2p+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\overline{\mu}_k-\lambda_j},\quad k=1,\dots,N+1. \tag{4.9b}$$

We have the following proposition.

Proposition 4: The Nijenhuis coordinates ($\boldsymbol{\mu},\boldsymbol{\nu}$) *defined by (4.8) are exactly the same as the generalized parabolic coordinates* ($\overline{\boldsymbol{\mu}},\overline{\boldsymbol{\nu}}$) *defined by (4.9). The* QBH *vector field (4.6) is Pfaffian in the Nijenhuis coordinates*.

Proof: In a similar way, we can show by induction that

$$f(\lambda)=B(\lambda)K(\lambda). \tag{4.10}$$

It follows from (4.9a) that

$$\psi_{1j}^2=2\frac{R(\lambda_j)}{K'(\lambda_j)},\quad j=1,\dots,N,$$

$$q=\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle+2\frac{\beta_{N+1}}{\alpha_N}=\sum_{j=1}^{N}\frac{2R(\lambda_j)}{\lambda_jK'(\lambda_j)}+2\frac{\beta_{N+1}}{\alpha_N}. \tag{4.11}$$

Then it is similar to find that

$$\nu_k=-A(\mu_k),$$

$$\rho=B(\lambda)|_{\lambda=0}=\frac{1}{2}q-\frac{1}{2}\langle\Lambda^{-1}\Psi_1,\Psi_1\rangle=\frac{\beta_{N+1}}{\alpha_N}.$$

This completes the proof.

V. CONCLUDING REMARKS

In the exact same way as we did in the preceding two sections, we can construct the third family of QBH systems from the constrained flows (3.5) for $l=m+2$, $m=1,2,\dots$. The QBH system (2.28) is just the second member ($m=2$) in the third family of QBH systems, and θ_1 and ρ given by (2.27) and (2.29a) are the second compatible Poisson tensor and the integrating factor for the third family of QBH systems.

In general, the constrained flow (3.5) for $l=m+k$ can be transformed into a FDIHS by introducing the Jacobi–Ostrogradsky coordinates. Under the map relating this FDIHS to that obtained from the modified constrained flow (3.14) for $l=m+k-1$, the image of the Poisson tensor θ_0 for the latter gives rise to the second compatible Poisson tensor θ_1 for the former. In this way, for each k we can obtain a family of QBH systems with $m=1,2,\dots$. The results obtained in the preceding sections suggest the following conjecture: each family of QBH systems ($l=m+k$, $m=1,2,\dots$) shares the same θ_1 and ρ for the QBH structure

$$\theta_0\nabla F_1=\frac{1}{\rho}\theta_1\nabla E_1,$$

and, in general, by means of the Lax matrix $Q=\begin{pmatrix}A(\lambda) & B(\lambda)\\ C(\lambda) & -A(\lambda)\end{pmatrix}$ and the expression

$$A^2(\lambda)+B(\lambda)C(\lambda)=\sum_{i=0}^{m+2k}\overline{F}_i\lambda^i+\sum_{i=1}^{N}\frac{F^{(i)}}{\lambda-\lambda_i},$$

we have

$$\rho = B(\lambda)|_{\lambda=0}, \quad E_1 = [A^2(\lambda) + B(\lambda)C(\lambda)]|_{\lambda=0} = \overline{F_0} - \sum_{i=1}^{N} F^{(i)} \lambda_i^{-1}.$$

For $k=1,2,...$, we find an infinite number of families of QBH systems. Furthermore we can show in a similar way that the Nijenhuis coordinates introduced by the Nijenhuis tensor are exactly the same as the separated variables defined by means of the Lax matrix for the QBH system in the family, and each QBH vector field is Pfaffian in the Nijenhuis coordinates.

ACKNOWLEDGMENTS

This work was supported by the Chinese Basic Research Project ''Nonlinear Science,'' the City University of Hong Kong and the Research Grants Council of Hong Kong. One of the authors (Y. B. Zeng) wishes to express his gratitude to Department of Mathematics of the City University of Hong Kong for warm hospitality.

[1] F. Magri, J. Math. Phys. **19**, 1156 (1978).
[2] M. Antonowicz, A. P. Fordy, and S. Rauch-Wojciechowski, Phys. Lett. A **124**, 143 (1987).
[3] M. Antonowicz and S. Rauch-Wojciechowski, J. Phys. A **24**, 5043 (1991).
[4] M. Antonowicz and S. Rauch-Wojciechowski, Phys. Lett. A **163**, 167 (1992).
[5] M. Antonowicz and S. Rauch-Wojciechowski, J. Math. Phys. **33**, 2115 (1992).
[6] Yunbo Zeng, J. Phys. A **24**, L11 (1993).
[7] Yunbo Zeng, J. Math. Phys. **34**, 4742 (1993).
[8] M. Blaszak, J. Phys. A **26**, 5985 (1993).
[9] R. Caboz, V. Ravoson, and L. Gavrilov, J. Phys. A **24**, L523 (1991).
[10] R. Brouzet, R. Caboz, J. Rabenivo, and V. Ravoson, J. Phys. A **29**, 2069 (1996).
[11] F. Magri and T. Marsico, *Electromagnetism and Geometrical Structures*, edited by G. Ferrarese (Bologna, Pitagora, 1996), p. 207.
[12] C. Morosi and G. Tondo, J. Phys. A **30**, 2799 (1997).
[13] M. Blaszak, J. Math. Phys. **39**, 3213 (1998).
[14] J. Rabenivo, J. Phys. A **34**, 7113 (1998).
[15] Yunbo Zeng, Phys. Lett. A **160**, 541 (1991).
[16] Yunbo Zeng and Yishen Li, J. Phys. A **26**, L273 (1993).
[17] Yunbo Zeng, Physica D **73**, 171 (1994).
[18] Yunbo Zeng, J. Phys. A **24**, L1065 (1991).
[19] W. X. Ma and W. Strampp, Phys. Lett. A **185**, 277 (1994).
[20] W. X. Ma, J. Phys. Soc. Jpn. **64**, 1085 (1995).
[21] W. X. Ma, B. Fuchssteiner, and W. Oevel, Physica A **233**, 331 (1996).
[22] W. X. Ma, Q. Ding, W. G. Zhang, and B. Q. Lu, Nuovo Cimento B **111**, 1135 (1996).
[23] O. Ragnisco and S. Rauch-Wojciechowski, Inverse Probl. **8**, 245 (1992).
[24] Yunbo Zeng and Runlian Lin, J. Math. Phys. **39**, 5964 (1998).
[25] E. K. Sklyanin, Prog. Theor. Phys. Suppl. **118**, 35 (1995).
[26] J. Harnad and P. Winternitz, Commun. Math. Phys. **172**, 263 (1995).
[27] Yunbo Zeng, J. Phys. A **30**, 3719 (1997).
[28] Yunbo Zeng, J. Phys. Soc. Jpn. **66**, 2277 (1997).
[29] M. Jaulent and K. Miodek, Lett. Math. Phys. **1**, 243 (1976).
[30] A. C. Newell, *Solitons in Mathematics and Physics* (SIAM, Philadelphia, 1985).
[31] M. Antonowicz and A. P. Fordy, Commun. Math. Phys. **124**, 465 (1989).
[32] B. A. Kupershmidt and J. Wilson, Invent. Math. **62**, 403 (1981).
[33] A. D. Dubrovin, Russian Math. Survey **36**, 11 (1981).
[34] Yunbo Zeng and Yishen Li, J. Math. Phys. **31**, 2835 (1990).
[35] J. C. Eilbeck, V. Z. Enol'skii, V. B. Kuznetsov, and A. V. Tsignov, J. Phys. A **27**, 567 (1994).

The generation of gravitational waves by the resonant interaction of sound waves

A. M. Anile
Department of Mathematics, University of Catania, Catania, Italy

John K. Hunter[a)]
Department of Mathematics and Institute of Theoretical Dynamics, University of California at Davis, Davis, California 95616

Binh Truong
Department of Mathematics, University of California at Davis, Davis, California 95616

(Received 23 February 1999; accepted for publication 5 May 1999)

We show that the resonant interaction of sound waves is a source of gravitational waves in general relativity. We derive an asymptotic equation that describes this process, and compute the interaction coefficient for the quadratically nonlinear coupling between sound waves and gravitational waves. © *1999 American Institute of Physics.* [S0022-2488(99)01709-0]

I. INTRODUCTION

We consider the propagation of two sound waves through a fluid that is contained in a relativistic space–time. If the interference pattern of the sound waves moves at the speed of light, then they resonate with a gravitational wave, and the resulting variations of the fluid energy–momentum tensor act as the source of a free gravitational wave. We will derive simplified, asymptotic equations that describe the generation of a gravitational wave by this resonant interaction.

Equations that describe the interaction between gravitational waves and waves in another field have been derived for scalar fields,[1] electromagnetic fields,[2] and Yang–Mills fields.[3] In these cases, the nongravitational waves propagate at the speed of light, so they resonate directly with a gravitational wave. The propagation speed of sound waves is less than the speed of light, so a single sound wave cannot generate a gravitational wave.[4] Thus, the resonant interaction of two sound waves is the simplest mechanism for the generation of gravitational waves by sound waves. In the Introduction, we summarize the equations that describe this process, and estimate the strength of the resulting gravitational wave.

We consider weak gravitational waves whose metric is a small perturbation of the flat Minkowski metric. Linearization of the Einstein field equations about the Minkowski metric leads to equations of the form

$$\mathcal{G}[\mathbf{h}] = \frac{8\pi G}{c^4}\mathbf{T}, \tag{1.1}$$

where $\mathbf{h}$ is the perturbation of the metric from the Minkowski metric, $\mathbf{T}$ is the energy–momentum tensor of the fluid, $\mathcal{G}$ is a linear partial differential operator acting on $\mathbf{h}$, c is the speed of light, and G is the gravitational constant. This linearization is valid provided that the background curvature of space–time caused by the unperturbed fluid is negligible over the region in which the waves interact.

[a)]Electronic mail: jkhunter@ucdavis.edu

First, we suppose that the sound waves are small-amplitude, harmonic acoustic wave packets. We denote the wave number four-vectors of the two acoustic waves by $k^{(J)}=(k_\alpha^{(J)})$, where $J=1,2$ and Greek indices run over 0, 1, 2, 3. We define associated phases θ_J by

$$\theta_J=k_\alpha^{(J)}x^\alpha. \tag{1.2}$$

We will not use the summation convention over the wave index J, and will indicate any sums with respect to J explicitly. According to linearized acoustics, the density perturbation ρ' in the acoustic wave field is given by

$$\rho'=(\overset{0}{\rho}+\overset{0}{p}/c^2)\sum_{J=1}^{2}\{A_J(x)e^{i\theta_J}+A_J^*(x)e^{-i\theta_J}\}+\cdots\ . \tag{1.3}$$

In (1.3), the dimensionless function A_J is a complex wave amplitude, and the quantities $\overset{0}{\rho}$ and $\overset{0}{p}$ are the unperturbed fluid density and pressure, repectively.

The energy–momentum tensor associated with the acoustic wave field (1.3) has the form

$$\begin{aligned}\mathbf{T}=\overset{0}{\mathbf{T}}&+\sum_{J=1}^{2}\{A_Je^{i\theta_J}\mathbf{P}_J+A_J^*e^{-i\theta_J}\mathbf{P}_J^*\}\\&+\sum_{J,K=1}^{2}\{A_JA_Ke^{i(\theta_J+\theta_K)}\mathbf{Q}_{JK}+A_J^*A_K^*e^{-i(\theta_J+\theta_K)}\mathbf{Q}_{JK}^*\\&+A_JA_K^*e^{i(\theta_J-\theta_K)}\mathbf{R}_{JK}+A_J^*A_Ke^{-i(\theta_J-\theta_K)}\mathbf{R}_{JK}^*\}+\cdots\ ,\end{aligned}$$

where $\mathbf{P}_J$, $\mathbf{Q}_{JK}$, and $\mathbf{R}_{JK}$ are constant, complex-valued tensors. This energy–momentum tensor is a source term for the gravitational field in Eq. (1.1), which, in essence, is a forced wave equation for the metric perturbation $\mathbf{h}$. When an acoustic source term is resonant, it generates a gravitational wave, provided that an appropriate interaction coefficient is nonzero. With a suitable choice of the signs of the wave number vectors, the condition for three-wave resonance to occur is that

$$k^{(1)}+k^{(2)}+k^{(3)}=0, \tag{1.4}$$

where $k^{(1)}$, $k^{(2)}$ satisfy the dispersion relation of acoustic waves, and $k^{(3)}$ is a null vector. The quantum mechanical interpretation of this triresonance condition is the conservation of energy and momentum in the resonant scattering of two phonons into a graviton.

It is convenient to use Lorentzian coordinates in which the fluid is at rest,

$$x=(t,\vec{x}),\quad \vec{x}=(x^1,x^2,x^3).$$

In these coordinates, the Minkowski metric components $\eta_{\alpha\beta}$ are given by

$$(\eta_{\alpha\beta})=\begin{pmatrix}-1&0&0&0\\0&1/c^2&0&0\\0&0&1/c^2&0\\0&0&0&1/c^2\end{pmatrix}.$$

We write a wave number four-vector k as

$$k=(-\omega,\vec{k}),\quad \vec{k}=(k_1,k_2,k_3),$$

where ω is the angular frequency and $\vec{k}$ is the spatial wave number vector. The dispersion relation of acoustic waves is

$$\omega^2 = c_s^2 \vec{k}^2, \tag{1.5}$$

where c_s is the sound speed of the unperturbed fluid, and the dispersion relation of gravitational waves is

$$\omega^2 = c^2 \vec{k}^2. \tag{1.6}$$

The triresonance condition (1.4) may be written as

$$\omega^{(1)} + \omega^{(2)} + \omega^{(3)} = 0, \quad \vec{k}^{(1)} + \vec{k}^{(2)} + \vec{k}^{(3)} = 0. \tag{1.7}$$

This condition implies that the spatial wave number vectors of the three waves are coplanar. We can therefore choose coordinates so that the propagation directions of the acoustic waves lie in the (x^1,x^2) plane and the propagation direction of the gravitational wave lies in the x^1 direction. As we will show, the gravitational wave generated by the resonant interaction is then plane polarized in the (x^2,x^3) directions. We may write the spatial wave number vectors explicitly as

$$\begin{aligned}\vec{k}^{(1)} &= \frac{\omega^{(1)}}{c_s}(\cos\varphi_1, \sin\varphi_1, 0),\\ \vec{k}^{(2)} &= \frac{\omega^{(2)}}{c_s}(\cos\varphi_2, \sin\varphi_2, 0),\\ \vec{k}^{(3)} &= \frac{\omega^{(3)}}{c}(1,0,0),\end{aligned} \tag{1.8}$$

where φ_1, φ_2 are the angles of the acoustic wave number vectors to the gravitational wave number vector, and $\varphi_1 \neq \varphi_2$. It follows from (1.7) and (1.8) that three-wave resonance occurs when the angles satisfy

$$c \sin(\varphi_1 - \varphi_2) = c_s(\sin\varphi_1 - \sin\varphi_2).$$

In a transverse, traceless gauge, the metric perturbation $h_{\alpha\beta}$ of the gravitational wave is given by

$$(h_{\alpha\beta}) = \{A_3(x)e^{i\theta_3} + A_3^*(x)e^{-i\theta_3}\}\begin{pmatrix} 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 1/c^2 & 0\\ 0 & 0 & 0 & -1/c^2\end{pmatrix}.$$

As we will show, the gravitational wave amplitude A_3 satisfies the equation

$$i\omega^{(3)}(\partial_t A_3 + c\,\partial_{x^1} A_3) + \Gamma A_1^* A_2^* = 0, \tag{1.9}$$

where the interaction coefficient Γ is given by

$$\Gamma = \frac{8\pi G c_s^2}{c^2}(\overset{0}{\rho} + \overset{0}{p}/c^2)\sin\varphi_1 \sin\varphi_2. \tag{1.10}$$

The sound wave amplitudes A_J in Eq. (1.3) satisfy

$$\partial_t A_J + C_J^j\,\partial_{x^j} A_J = 0, \quad J = 1,2, \tag{1.11}$$

where lower case latin indices run over $1,\ 2,\ 3$, and C_J^j is the spatial group velocity of the Jth wave,

$$C_J^j = \frac{c_s^2 k_j^{(J)}}{\omega^{(J)}}, \quad J = 1,2. \tag{1.12}$$

Equation (1.9) describes the forcing of the gravitational wave by the resonant interaction of the acoustic waves. If the acoustic waves collide head on in the rest frame of the fluid, then $\varphi_1 = 0$, $\varphi_2 = \pi$, and $\Gamma = 0$, so no gravitational wave is generated by the three-wave resonance. The interaction coefficient Γ is nonzero for the resonant interaction of oblique acoustic waves. The angular dependence of the interaction coefficient is explained by the fact that the component of the velocity perturbation of the Jth longitudinal sound wave in the x^2 direction transverse to the gravitational wave is proportional to $\sin \varphi_J$. The resonant component of the transverse momentum tensor, which forces the gravitational wave, is therefore proportional to $\sin \varphi_1 \sin \varphi_2$.

The form of the resonant interaction equations (1.9), (1.11) differs from the usual form of the three-wave resonant interaction equations for Hamiltonian systems,[5] because there is no forcing of an acoustic wave by the resonant interaction between the other acoustic wave and the gravitational wave. From the Lagrangian point of view, the asymmetry between the gravitational and acoustic equations is a consequence of the unusual structure of the variational principle for the Einstein field equations, which provides variational equations for the metric but no variational equations for the fluid. Instead, the fluid equations follow from the metric equations on account of the Bianchi identities. One consequence of this asymmetry is that the resonant interaction equations do not conserve energy exactly. The lack of exact energy conservation for the resonant interaction equations is not physically inconsistent, however. The amplitude of the gravitational wave is much smaller than the amplitudes of the acoustic waves, as can be seen from Eq. (1.18) below, so the depletion of energy of the acoustic waves due to the generation of the gravitational wave is negligible for the time scales over which the resonant interaction equations apply.

In the above discussion, we have described our results for the case of harmonic acoustic waves. In the following sections, we derive the resonant-interaction equations for the more general case of weakly nonlinear sound waves, where the nonlinear self-interaction of the waves leads to the distortion of the wave profiles. Harmonic waves are obtained as a limiting case of weakly nonlinear waves in which their wave amplitudes tends to zero.

The density perturbation for a sum of two weakly nonlinear sound waves is given by

$$\rho' = (\overset{0}{\rho} + \overset{0}{p}/c^2) \sum_{J=1}^{2} a_J(x, \theta_J) + \cdots , \tag{1.13}$$

where a_J is a 2π-periodic function of θ_J whose dependence on the phase variable θ_J describes the local profile of the sound wave. Expansion of the fluid equations implies that a_J satisfies an inviscid Burgers equation,

$$\partial_t a_J + C_J^j \, \partial_{x^j} a_J + \partial_{\theta_J}\left(\tfrac{1}{2}\,\omega^{(J)} \Lambda_J a_J^2\right) = 0, \tag{1.14}$$

where, for a barotropic fluid,

$$\Lambda_J = 1 - \frac{c_s^2}{c^2}. \tag{1.15}$$

The metric perturbation of the gravitational wave is given by

$$(h_{\alpha\beta}) = a_3(x,\theta_3)\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1/c^2 & 0 \\ 0 & 0 & 0 & -1/c^2 \end{pmatrix}, \tag{1.16}$$

where the gravitational wave amplitude a_3 satisfies

$$\omega^{(3)}\partial_{\theta_3}(\partial_t a_3 + c\,\partial_{x^1} a_3) + \Gamma\langle a_1 a_2\rangle = 0. \tag{1.17}$$

Here, Γ is defined in (1.10), and the bracket $\langle\cdot\rangle$ denotes a phase average with respect to θ_1 or θ_2 in which θ_3 is held fixed and $\theta_1+\theta_2+\theta_3=0$, in view of (1.2) and (1.4). That is,

$$\langle a_1 a_2\rangle(x,\theta_3) = \frac{1}{2\pi}\int_0^{2\pi} a_1(x,\xi)a_2(x,-\xi-\theta_3)\,d\xi.$$

For very small-amplitude sound waves, we can neglect the nonlinear term in (1.14), and consider harmonic waves with

$$a_J(x,\theta_J) = A_J(x)e^{i\theta_J} + A_J^*(x)e^{-i\theta_J}, \quad J=1,2,3.$$

In that case, (1.14) reduces to (1.11), and (1.17) reduces to (1.9).

The coupling between sound waves and gravitational waves is weak, but the volume of the source region in which the resonant interaction takes place can be large, so that it is possible for the sound waves to generate a significant flux of low-frequency gravitational waves. This mechanism contrasts with the rapid collapse of a massive object, which provides a strong, localized source of gravitational waves.[6]

To estimate the strength of a gravitational wave generated by such a resonant interaction, we suppose that the dimensionless amplitudes a_J of the sound waves are of the order ϵ, and that their interaction time is of the order T. Integration of (1.17) implies that the dimensionless gravitational wave amplitude a_3 is of the order $\epsilon^2 T\Gamma/\omega^{(3)}$. Using the expression in (1.10) for Γ, writing $\omega^{(3)}=2\pi f$, where f is the frequency of the gravitational wave, and neglecting the order one geometrical factor $\sin\varphi_1 \sin\varphi_2$, we find that the gravitational wave amplitude is of the order α, where

$$\alpha = \frac{4\epsilon^2 c_s^2 G T}{c^2 f}(\overset{0}{\rho} + \overset{0}{p}/c^2). \tag{1.18}$$

An upper limit on the interaction time T is the shock formation time of the sound waves, which is of the order

$$T = \frac{1}{\epsilon f}.$$

In that case, we have

$$\alpha = \frac{4\epsilon c_s^2 G}{c^2 f^2}(\overset{0}{\rho} + \overset{0}{p}/c^2). \tag{1.19}$$

In the present, matter-dominated era of the universe, this amplitude is too small for the resonant interaction of sound waves to be a significant source of gravitational waves. However, much stronger gravitational waves may have been generated in the radiation-dominated era. The equation of state for a radiation-dominated universe is

$$p = \tfrac{1}{3}\rho c^2.$$

Thus, from (1.19), the strength α of the gravitational wave is of the order

$$\alpha = \frac{16\epsilon G \overset{0}{\rho}}{9f^2}.$$

For example, in a flat Robertson–Walker universe, the density is given as a function of the age of the universe τ by[7]

$$\overset{0}{\rho} = \frac{3}{32\pi G\tau^2},$$

so that

$$\alpha = \frac{\epsilon}{6\pi\tau^2 f^2}.$$

The gravitational wave amplitude is therefore significant at sufficiently low frequencies.

II. THE FIELD EQUATIONS

We use units in which

$$8\pi G = 1, \quad c = 1. \tag{2.1}$$

The Einstein field equations are

$$G_{\alpha\beta} = T_{\alpha\beta}, \tag{2.2}$$

where $G_{\alpha\beta}$ is the Einstein tensor and $T_{\alpha\beta}$ is the energy–momentum tensor. The Einstein tensor $G_{\alpha\beta}$ is given in terms of the Ricci curvature tensor $R_{\alpha\beta}$, the scalar curvature $R = R^{\alpha}_{\alpha}$, the metric tensor $g_{\alpha\beta}$, and the associated connection coefficients $\Gamma^{\alpha}_{\beta\gamma}$ by

$$G_{\alpha\beta} = R_{\alpha\beta} - \tfrac{1}{2} R g_{\alpha\beta},$$

$$R_{\alpha\beta} = \partial_\lambda \Gamma^{\lambda}_{\alpha\beta} - \partial_\lambda \Gamma^{\lambda}_{\beta\lambda} + \Gamma^{\lambda}_{\alpha\beta}\Gamma^{\mu}_{\lambda\mu} - \Gamma^{\mu}_{\alpha\lambda}\Gamma^{\lambda}_{\beta\mu},$$

$$\Gamma^{\alpha}_{\beta\gamma} = \tfrac{1}{2} g^{\alpha\lambda}(\partial_\beta g_{\lambda\gamma} + \partial_\gamma g_{\lambda\beta} - \partial_\lambda g_{\beta\gamma}).$$

The Einstein equations (2.2) may be written in the equivalent form

$$R_{\alpha\beta} = T_{\alpha\beta} - \tfrac{1}{2} T g_{\alpha\beta}, \tag{2.3}$$

where T is the trace of the energy–momentum tensor.

We study weak gravitational waves for which the metric tensor $g_{\alpha\beta}$ is a small perturbation of the Minkowski metric $\eta_{\alpha\beta}$. We therefore write

$$g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}, \tag{2.4}$$

where the metric perturbation $h_{\alpha\beta}$ is small. Use of (2.4) in (2.3) and linearization of the result with respect to $h_{\alpha\beta}$ implies that

$$\mathcal{L}[h]_{\alpha\beta} = T_{\alpha\beta} - \tfrac{1}{2} T \eta_{\alpha\beta}, \tag{2.5}$$

where the linearized gravitational wave operator $\mathcal{L}$ is given by

$$\mathcal{L}[h]_{\alpha\beta}=\tfrac{1}{2}(-\partial^{\lambda}\partial_{\lambda}h_{\alpha\beta}-\partial_{\alpha}\partial_{\beta}h^{\lambda}_{\lambda}+\partial_{\beta}\partial_{\lambda}h^{\lambda}_{\alpha}+\partial_{\alpha}\partial_{\lambda}h^{\lambda}_{\beta}), \tag{2.6}$$

and

$$T=\eta^{\alpha\beta}T_{\alpha\beta}. \tag{2.7}$$

Throughout the paper, we use the Minkowski metric to raise and lower indices.

Computation of the divergence of (2.5) implies that the energy–momentum tensor satisfies the conservation laws,

$$\partial_{\alpha}T^{\alpha\beta}=0. \tag{2.8}$$

After evaluation at the Minkowski metric, the energy–momentum tensor of the fluid is given by

$$T^{\alpha\beta}=(\rho+p)u^{\alpha}u^{\beta}+p\,\eta^{\alpha\beta}, \tag{2.9}$$

where ρ is the mass–energy density, p is the pressure, and u^{α} is the four-velocity of the fluid. The four-velocity satisfies

$$u^{\alpha}u_{\alpha}=-1. \tag{2.10}$$

In a barotropic fluid, the pressure and density are related by

$$p=c_s^2\rho, \tag{2.11}$$

where the constant c_s is the sound speed. Fluids with more general equations of state could be treated in a similar way.

In Secs. III–IV, we derive an asymptotic solution of the fluid equations (2.8)–(2.11), which describes the sum of two weakly nonlinear sound waves. In Secs. V–VI, we solve the gravitational equations (2.5) for the metric perturbation.

In dimensioned variables, the length scale L of the variation of the background space–time metric caused by the presence of the fluid is given by

$$L=\left[\frac{8\pi G}{c^2}(\overset{0}{\rho}+\overset{0}{p}/c^2)\right]^{-1/2}.$$

The Einstein equations can be linearized about the Minkowski metric, provided that the length scale δ of the interaction region is small compared with L. To carry out the linearization in a systematic way, we use dimensionless variables in which $L=1$, and denote the corresponding space–time variables by y^{α}. The solution for the metric in a region of order δ about the origin is given by an expansion of the form

$$g_{\alpha\beta}(y;\delta)=\eta_{\alpha\beta}+\delta^2h_{\alpha\beta}(x)+O(\delta^3),\qquad x^{\alpha}=\frac{y^{\alpha}}{\delta}. \tag{2.12}$$

The corresponding expansion of the fluid's energy–momentum tensor $T_{\alpha\beta}(\mathbf{g})$ is

$$T_{\alpha\beta}(\mathbf{g})=T_{\alpha\beta}+O(\delta^2), \tag{2.13}$$

where the energy–momentum tensor $T_{\alpha\beta}$ on the right-hand side of (2.13) is evaluated at the Minkowski metric. Using (2.12) and (2.13) in (2.3), and setting $\delta=0$, we obtain the linearized equations (2.6)–(2.7).

III. EXPANSION OF THE FLUID EQUATIONS

The fluid variables are the pressure p, the density ρ, and the four-velocity u^α. We denote the constant fluid variables in the unperturbed state by

$$p=\overset{0}{p},\quad \rho=\overset{0}{\rho},\quad u^\alpha=\overset{0}{u}{}^\alpha.$$

We use a reference frame in which the unperturbed fluid is at rest, so that

$$(\overset{0}{u}{}^\alpha)=(1,0,0,0),\quad (\overset{0}{u}{}_\alpha)=(-1,0,0,0). \tag{3.1}$$

We look for small-amplitude, high-frequency asymptotic solutions of (2.8)–(2.11) of the form

$$\begin{aligned} p&=\overset{0}{p}+\epsilon\overset{1}{p}(x,\theta_1,\theta_2)+\epsilon^2\overset{2}{p}(x,\theta_1,\theta_2)+O(\epsilon^3),\\ \rho&=\overset{0}{\rho}+\epsilon\overset{1}{\rho}(x,\theta_1,\theta_2)+\epsilon^2\overset{2}{\rho}(x,\theta_1,\theta_2)+O(\epsilon^3),\\ u^\alpha&=\overset{0}{u}{}^\alpha+\epsilon\overset{1}{u}{}^\alpha(x,\theta_1,\theta_2)+\epsilon^2\overset{2}{u}{}^\alpha(x,\theta_1,\theta_2)+O(\epsilon^3), \end{aligned} \tag{3.2}$$

where the phase variable θ_J is given by

$$\theta_J=\frac{k_\alpha^{(J)}x^\alpha}{\epsilon},\quad J=1,2. \tag{3.3}$$

Here, $k^{(J)}$ is the scaled wave number vector of the Jth wave.

In (3.2)–(3.3), the small dimensionless parameter ϵ measures both the amplitude and the wavelength of the sound waves. This scaling is the appropriate one for the description of small-amplitude sound waves that propagate over a distance comparable with their shock-formation distance.[8,9] Once shocks form, the sound waves decay, or other physical effects become important, so the shock-formation distance effectively limits the length scale of the interaction region.

Because of the algebraic relations (2.10)–(2.11), there are four independent fluid variables, which we take to be the pressure p and the spatial velocity components u_j. Use of (3.1) and (3.2) in (2.10), and expansion of the result with respect to ϵ, implies that the timelike component of the four-velocity is given in terms of the spatial components by (3.2), with

$$\overset{1}{u}{}^0=0,\quad \overset{2}{u}{}^0=\tfrac{1}{2}\overset{1}{u}{}^j\overset{1}{u}{}_j. \tag{3.4}$$

From (2.11), the density perturbations are given in terms of the pressure perturbations by

$$\overset{1}{\rho}=\frac{\overset{1}{p}}{c_s^2},\quad \overset{2}{\rho}=\frac{\overset{2}{p}}{c_s^2}. \tag{3.5}$$

Use of (3.2) in the expression (2.9) for the energy–momentum tensor implies that

$$T^{\alpha\beta}=\overset{0}{T}{}^{\alpha\beta}+\epsilon\overset{1}{T}{}^{\alpha\beta}(x,\theta_1,\theta_2)+\epsilon^2\overset{2}{T}{}^{\alpha\beta}(x,\theta_1,\theta_2)+O(\epsilon^3), \tag{3.6}$$

where

$$\overset{0}{T}{}^{\alpha\beta}=\begin{pmatrix} \overset{0}{\rho} & 0 \\ 0 & \overset{0}{p}\,\delta^{jk} \end{pmatrix},$$

$$\overset{1}{T}{}^{\alpha\beta}=\begin{pmatrix} \overset{1}{\rho} & (\overset{0}{\rho}+\overset{0}{p})\overset{1}{u}{}^{k} \\ (\overset{0}{\rho}+\overset{0}{p})\overset{1}{u}{}^{j} & \overset{1}{p}\,\delta^{jk} \end{pmatrix}, \tag{3.7}$$

$$\overset{2}{T}{}^{\alpha\beta}=\begin{pmatrix} \overset{2}{\rho}+2(\overset{0}{\rho}+\overset{0}{p})\overset{2}{u}{}^{0} & (\overset{1}{\rho}+\overset{1}{p})\overset{1}{u}{}^{k}+(\overset{0}{\rho}+\overset{0}{p})\overset{2}{u}{}^{k} \\ (\overset{1}{\rho}+\overset{1}{p})\overset{1}{u}{}^{j}+(\overset{0}{\rho}+\overset{0}{p})\overset{2}{u}{}^{j} & \overset{2}{p}\,\delta^{jk}+(\overset{0}{\rho}+\overset{0}{p})\overset{1}{u}{}^{j}\overset{1}{u}{}^{k} \end{pmatrix}.$$

Use of (3.3) and (3.6) in the conservation law (2.8) implies that

$$\left(\frac{1}{\epsilon}\sum_{J=1}^{2} k_{\alpha}^{(J)}\partial_{\theta_J}+\partial_{\alpha}\right)(\overset{0}{T}{}^{\alpha\beta}+\epsilon\overset{1}{T}{}^{\alpha\beta}+\epsilon^{2}\overset{2}{T}{}^{\alpha\beta}+O(\epsilon^{3}))=0. \tag{3.8}$$

Expanding and equating the coefficients of ϵ^{-1} and ϵ^{0} to zero in (3.8), we obtain that

$$\sum_{J=1}^{2} k_{\alpha}^{(J)}\,\partial_{\theta_J}\overset{1}{T}{}^{\alpha\beta}=0, \tag{3.9}$$

$$\sum_{J=1}^{2} k_{\alpha}^{(J)}\,\partial_{\theta_J}\overset{2}{T}{}^{\alpha\beta}+\partial_{\alpha}\overset{1}{T}{}^{\alpha\beta}=0. \tag{3.10}$$

In order to write these equations in a more convenient form, we define a four-component vector field $\mathbf{U}=(U_{\alpha})$ of independent fluid variables by

$$\mathbf{U}=\begin{pmatrix} p \\ u_j \end{pmatrix}. \tag{3.11}$$

From (3.1), (3.2), and (3.11),

$$\mathbf{U}=\overset{0}{\mathbf{U}}+\epsilon\overset{1}{\mathbf{U}}(x,\theta_1,\theta_2)+\epsilon^{2}\overset{2}{\mathbf{U}}(x,\theta_1,\theta_2)+O(\epsilon^{3}), \tag{3.12}$$

$$\overset{0}{\mathbf{U}}=\begin{pmatrix} \overset{0}{p} \\ 0 \end{pmatrix}, \quad \overset{1}{\mathbf{U}}=\begin{pmatrix} \overset{1}{p} \\ \overset{1}{u}_j \end{pmatrix}, \quad \overset{2}{\mathbf{U}}=\begin{pmatrix} \overset{2}{p} \\ \overset{2}{u}_j \end{pmatrix}. \tag{3.13}$$

We define tensor-valued functions $B=(B^{\alpha\beta})$, $C=(C^{\alpha\beta})$ of $\mathbf{U}=(\rho,u_j)^{T}$ by

$$B(\mathbf{U})=\begin{pmatrix} p/c_s^{2} & (\overset{0}{\rho}+\overset{0}{p})u^{k} \\ (\overset{0}{\rho}+\overset{0}{p})u^{j} & p\,\delta^{jk} \end{pmatrix}, \tag{3.14}$$

$$C(\mathbf{U},\mathbf{U})=\begin{pmatrix} (\overset{0}{\rho}+\overset{0}{p})u^{j}u_{j} & (\rho+p)u^{k} \\ (\rho+p)u^{j} & (\overset{0}{\rho}+\overset{0}{p})u^{j}u^{k} \end{pmatrix}. \tag{3.15}$$

Since we are using a Lorentzian coordinate frame, we have $u_j = u^j$. From (3.14)–(3.15), $B(\mathbf{U})$ is a linear function of $\mathbf{U}$, and

$$C(\mathbf{U},\mathbf{V}) = \tfrac{1}{4}[C(\mathbf{U}+\mathbf{V},\mathbf{U}+\mathbf{V}) - C(\mathbf{U}-\mathbf{V},\mathbf{U}-\mathbf{V})]$$

is a bilinear function of $(\mathbf{U},\mathbf{V})$.

It follows from (3.4), (3.5), (3.7), (3.14), and (3.15) that

$$\overset{1}{T}{}^{\alpha\beta} = B^{\alpha\beta}(\overset{1}{\mathbf{U}}), \tag{3.16}$$

$$\overset{2}{T}{}^{\alpha\beta} = B^{\alpha\beta}(\overset{2}{\mathbf{U}}) + C^{\alpha\beta}(\overset{1}{\mathbf{U}},\overset{1}{\mathbf{U}}). \tag{3.17}$$

Therefore, Eqs. (3.9)–(3.10) can be written in the form

$$\sum_{J=1}^{2} k_\alpha^{(J)} \partial_{\theta_J} B^{\alpha\beta}(\overset{1}{\mathbf{U}}) = 0, \tag{3.18}$$

$$\sum_{J=1}^{2} k_\alpha^{(J)} \partial_{\theta_J} B^{\alpha\beta}(\overset{2}{\mathbf{U}}) + \partial_\alpha B^{\alpha\beta}(\overset{1}{\mathbf{U}}) + \sum_{J=1}^{2} k_\alpha^{(J)} \partial_{\theta_J} C^{\alpha\beta}(\overset{1}{\mathbf{U}},\overset{1}{\mathbf{U}}) = 0. \tag{3.19}$$

IV. SOLUTION OF THE FLUID EQUATIONS

We nondimensionalize the fluid variables so that

$$\overset{0}{\rho} + \overset{0}{p} = 1. \tag{4.1}$$

It follows from (3.11), (3.14), (3.15), and (4.1) that

$$k_\alpha^{(J)} B^{\alpha\beta}(\mathbf{U}) = A_J^{\beta\alpha} U_\alpha, \tag{4.2}$$

where the four-by-four matrices $A_J = (A_J^{\beta\alpha})$ are given by

$$A_J = \begin{pmatrix} k_0^{(J)}/c_s^2 & k^{(J)k} \\ k^{(J)j} & k_0^{(J)} \delta^{jk} \end{pmatrix}, \qquad J = 1,2. \tag{4.3}$$

Use of (4.2) in (3.18) implies that

$$A_1\, \partial_{\theta_1} \overset{1}{\mathbf{U}} + A_2\, \partial_{\theta_2} \overset{1}{\mathbf{U}} = 0. \tag{4.4}$$

Equation (4.4) is a symmetric hyperbolic system for $\overset{1}{\mathbf{U}}$ in two independent variables (θ_1, θ_2). The ''slow'' variables x occur in (4.4) as parameters.

The determinant of A_J is given by

$$\det A_J = \frac{(k_0^{(J)})^2}{c_s^2} \{(k_0^{(J)})^2 - c_s^2 k^{(J)j} k_j^{(J)}\}.$$

In order for θ_J to be the phase of a sound wave, we require that $k^{(J)}$ satisfies the acoustic dispersion relation,

$$(k_0^{(J)})^2 = c_s^2 k^{(J)j} k_j^{(J)}, \tag{4.5}$$

in which case A_J is singular. We denote a normalized null vector by $\mathbf{R}_J$, where

$$A_J\mathbf{R}_J=0, \qquad J=1,2. \tag{4.6}$$

From (4.3), (4.5), and (4.6), we find that

$$\mathbf{R}_J=\begin{pmatrix} k_0^{(J)} \\ -k_j^{(J)} \end{pmatrix}, \quad J=1,2. \tag{4.7}$$

From (4.6), a solution of (4.4) is given by

$$\overset{1}{\mathbf{U}}=b_1(x,\theta_1)\mathbf{R}_1+b_2(x,\theta_2)\mathbf{R}_2, \tag{4.8}$$

where $b_J(x,\theta_J)$ is an arbitrary scalar-valued function, which we assume is a 2π-periodic, zero-mean function of θ_J. The solution (4.8) describes a sum of two oscillatory sound waves. We will derive an equation for the amplitude function b_J from the solvability conditions for the second-order perturbation equation, (3.19).

Using (4.2), we can write (3.19) as

$$A_1\,\partial_{\theta_1}\overset{2}{\mathbf{U}}+A_2\,\partial_{\theta_2}\overset{2}{\mathbf{U}}=\mathbf{F}(\overset{1}{\mathbf{U}}), \tag{4.9}$$

where $\mathbf{F}=(F^\beta)$ with

$$F^\beta(\overset{1}{\mathbf{U}})=-\left\{\partial_\alpha B^{\alpha\beta}(\overset{1}{\mathbf{U}})+\sum_{J=1}^{2}k_\alpha^{(J)}\partial_{\theta_J}C^{\alpha\beta}(\overset{1}{\mathbf{U}},\overset{1}{\mathbf{U}})\right\}. \tag{4.10}$$

Equation (4.9) is a linear, nonhomogeneous hyperbolic equation for $\mathbf{U}2$ that we solve by the method of characteristics.

The line

$$\frac{d\theta_1}{ds}=\mu_1, \quad \frac{d\theta_2}{ds}=\mu_2,$$

is a characteristic of (4.9) if there is an eigenvector $\mathbf{R}$, such that

$$(\mu_2A_1-\mu_1A_2)\mathbf{R}=0.$$

Using (4.3), we find that

$$\det(\mu_2A_1-\mu_1A_2)=\frac{2(q-1)k_0^{(1)}k_0^{(2)}}{c_s^2}\mu_1\mu_2(\mu_2k_0^{(1)}-\mu_1k_0^{(2)})^2,$$

where

$$q=\frac{k_j^{(1)}k^{(2)j}c_s^2}{k_0^{(1)}k_0^{(2)}}. \tag{4.11}$$

Thus, Eq. (4.9) has two simple characteristics, given by $\mu_1=0$ and $\mu_2=0$, with associated eigenvectors $\mathbf{R}=\mathbf{R}_1$ and $\mathbf{R}=\mathbf{R}_2$, respectively. These characteristics correspond to acoustic waves. Equation (4.9) also has a double characteristic, given by

$$\mu_2k_0^{(1)}-\mu_1k_0^{(2)}=0,$$

which corresponds to vorticity waves. The associated eigenspace is two dimensional, but only one eigenvector is required to solve (4.9) because the acoustic wave interaction forces velocity perturbations that are coplanar with the acoustic wave number vectors. The relevant eigenvector $\mathbf{R} = \mathbf{R}_0$ is given by

$$\mathbf{R}_0 = \begin{pmatrix} 0 \\ k_0^{(1)} k_j^{(2)} + k_0^{(2)} k_j^{(1)} \end{pmatrix}. \tag{4.12}$$

We look for a solution of Eq. (4.9) of the form

$$\overset{2}{\mathbf{U}} = w_0(x,\theta_1,\theta_2)\mathbf{R}_0 + w_1(x,\theta_1,\theta_2)\mathbf{R}_1 + w_2(x,\theta_1,\theta_2)\mathbf{R}_2. \tag{4.13}$$

We use (4.13) in (4.9) and, after some algebra,[10] find that it is a solution, provided that the scalar-valued coefficients w_0, w_1, w_2 satisfy the equations

$$\sum_{J=1}^{2} k_0^{(J)} \partial_{\theta_J} w_0 = \sum_{J=1}^{2} k_0^{(J)} \{ V_J \partial_{\theta_J}(b_J)^2 + W \partial_{\theta_J}(b_1 b_2) + H^\alpha \partial_\alpha b_J \}, \tag{4.14}$$

$$\begin{aligned} \partial_{\theta_2} w_1 = {} & M_1(\partial_{\theta_1} b_1) b_2 + N_1 b_1 (\partial_{\theta_1} b_1) \\ & + P_1 b_1 (\partial_{\theta_2} b_2) + Q_2 b_2 (\partial_{\theta_2} b_2) + Y_1^\alpha \partial_\alpha b_1 + Z_1^\alpha \partial_\alpha b_2, \end{aligned} \tag{4.15}$$

$$\begin{aligned} \partial_{\theta_1} w_2 = {} & M_2 b_1 (\partial_{\theta_2} b_2) + N_2 b_2 (\partial_{\theta_2} b_2) \\ & + P_2 (\partial_{\theta_1} b_1) b_2 + Q_1 b_1 (\partial_{\theta_1} b_1) + Y_2^\alpha \partial_\alpha b_2 + Z_2^\alpha \partial_\alpha b_1. \end{aligned} \tag{4.16}$$

The coefficients in (4.14)–(4.16) are given by

$$\begin{gathered} V_J = \frac{(k_0^{(J)})^2}{2k_0^{(1)} k_0^{(2)}}, \quad W = \frac{2c_s^2 + 1 - q}{2c_s^2}, \quad H^\alpha = \frac{h^\alpha}{k_0^{(1)} k_0^{(2)} (1+q)}, \\ M_J = \frac{k_0^{(J)} q (1-c_s^2)}{c_s^2 (1-q)}, \quad N_J = \frac{(k_0^{(J)})^3 (1-c_s^2)}{c_s^2 k_0^{(1)} k_0^{(2)} (1-q)}, \quad P_J = \frac{k_0^{(1)} k_0^{(2)} (1 + q^2 - 2qc_s^2)}{2k_0^{(J)} c_s^2 (1-q)}, \\ Q_J = \frac{(k_0^{(J)})^3 (1 - qc_s^2)}{c_s^2 k_0^{(1)} k_0^{(2)} (1-q)}, \quad Y_J^\alpha = -\frac{k_0^{(J)} C_J^\alpha}{k_0^{(1)} k_0^{(2)} (1-q)}, \quad Z_J^\alpha = -\frac{h^\alpha}{2k_0^{(J)} (1-q)}. \end{gathered} \tag{4.17}$$

Here, q is defined in (4.11), and the four-vectors $h = (h^\alpha)$ and $C_J = (C_J^\alpha)$ are given by

$$h = \begin{pmatrix} 1+q \\ -c_s^2 k_j^{(1)}/k_0^{(1)} - c_s^2 k_j^{(2)}/k_0^{(2)} \end{pmatrix}, \tag{4.18}$$

$$C_J = \begin{pmatrix} 1 \\ -c_s^2 k_j^{(J)}/k_0^{(J)} \end{pmatrix}. \tag{4.19}$$

Integration of (4.14)–(4.16) implies that

$$w_0 = \sum_{J=1}^{2} \{ H^\alpha \partial_\alpha B_J + V_J b_J^2 \} + W b_1 b_2 + f_0, \tag{4.20}$$

$$w_1=(N_1 b_1\,\partial_{\theta_1} b_1+Y_1^\alpha\,\partial_\alpha b_1)\,\theta_2+M_1(\partial_{\theta_1} b_1)B_2$$
$$+P_1 b_1 b_2+\tfrac{1}{2}Q_2(b_2)^2+Z_1^\alpha\,\partial_\alpha B_2+f_1, \tag{4.21}$$

$$w_2=(N_2 b_2\,\partial_{\theta_2} b_2+Y_2^\alpha\,\partial_\alpha b_1)\,\theta_1+M_2(\partial_{\theta_2} b_2)B_1$$
$$+P_2 b_1 b_2+\tfrac{1}{2}Q_1(b_1)^2+Z_2^\alpha\,\partial_\alpha B_1+f_2. \tag{4.22}$$

Here, $f_0(x,k_0^{(2)}\theta_1-k_0^{(1)}\theta_2)$, $f_1(x,\theta_1)$, and $f_2(x,\theta_2)$ are arbitrary functions of integration, and

$$B_J(x,\theta_J)=\int_0^{\theta_J} b_J(x,\xi)\,d\xi-\frac{1}{2\pi}\int_0^{2\pi}\int_0^{\eta} b_J(x,\xi)d\xi\,d\eta \tag{4.23}$$

denotes the zero-mean integral of b_J with respect to θ_J.

Terms in w_J that grow linearly in θ_J are secular terms that invalidate the asymptotic solution. We therefore require that the coefficients of θ_2 and θ_1 in (4.21) and (4.22) are zero. It follows that

$$Y_J^\alpha\,\partial_\alpha b_J+N_J b_J\,\partial_{\theta_J} b_J=0,\quad J=1,2. \tag{4.24}$$

After using (4.11) and (4.17) in (4.24), we obtain an an inviscid Burgers equation for b_J,

$$C_J^\alpha\,\partial_\alpha b_J-\partial_{\theta_J}\left(\frac{(k_0^{(J)})^2}{2c_s^2}\Lambda_J b_J^2\right)=0, \tag{4.25}$$

where Λ_J is given in (1.15). The solutions for w_1, w_2 in (4.21)–(4.22) then reduce to

$$w_1=M_1(\partial_{\theta_1} b_1)B_2+P_1 b_1 b_2+\tfrac{1}{2}Q_2(b_2)^2+Z_1^\alpha\,\partial_\alpha B_2+f_1, \tag{4.26}$$

$$w_2=M_2(\partial_{\theta_2} b_2)B_1+P_2 b_1 b_2+\tfrac{1}{2}Q_1(b_1)^2+Z_2^\alpha\,\partial_\alpha B_1+f_2. \tag{4.27}$$

In summary, an asymptotic solution as $\epsilon\to 0$ of (2.8)–(2.11) for the fluid variables $\mathbf{U}$ in (3.11) is given by

$$\mathbf{U}=\overset{0}{\mathbf{U}}+\epsilon\{b_1(x,\theta_1)\mathbf{R}_1+b_2(x,\theta_2)\mathbf{R}_2\}+\epsilon^2\{w_0(x,\theta_1,\theta_2)\mathbf{R}_0$$
$$+w_1(x,\theta_1,\theta_2)\mathbf{R}_1+w_2(x,\theta_1,\theta_2)\mathbf{R}_2\}+O(\epsilon^3), \tag{4.28}$$

where the eigenvectors $\mathbf{R}_0$, $\mathbf{R}_1$, $\mathbf{R}_2$ are defined in (4.12) and (4.7), b_J satisfies (4.25), w_0, w_1, w_2 are given in (4.20) and (4.26)–(4.27), the phase variable θ_J is defined in (3.3), and the wave number vector $k^{(J)}$ satisfies the acoustic dispersion relation (4.5). From (3.5), (3.11), (4.7), and (4.8), we may write

$$\overset{1}{\rho}=\sum_{J=1}^{2} a_J,\quad a_J=\frac{k_0^{(J)}}{c_s^2}b_J. \tag{4.29}$$

Reintroducing dimensioned variables, and removing the expansion parameter ϵ by setting $\epsilon=1$, we find from (4.25) and (4.29) that the leading-order density perturbation ρ' is given by (1.13), where a_J satisfies (1.14).

The use of (4.8) and (4.13) in (3.16)–(3.17) implies that the terms in the expansion (3.6) of the energy–momentum tensor are given by

$$\overset{1}{T}{}^{\alpha\beta}(x,\theta_1,\theta_2)=\sum_{J=1}^{2} b_J(x,\theta_J)B^{\alpha\beta}(\mathbf{R}_J), \tag{4.30}$$

$$\overset{2}{T}{}^{\alpha\beta}(x,\theta_1,\theta_2)=\sum_{J=1}^{3} w_J(x,\theta_1,\theta_2)B^{\alpha\beta}(\mathbf{R}_J)+\sum_{J,K=1}^{2} b_J(x,\theta_J)b_K(x,\theta_K)C^{\alpha\beta}(\mathbf{R}_J,\mathbf{R}_K), \tag{4.31}$$

where B and C are defined in (3.14) and (3.15).

V. EXPANSION OF THE GRAVITATIONAL EQUATIONS

We suppose that the acoustic wave number vectors satisfy the triresonance condition (1.4). We define a gravitational phase variable by

$$\theta_3=\frac{k_\alpha^{(3)}x^\alpha}{\epsilon}, \tag{5.1}$$

where the gravitational wave number vector $k^{(3)}$ satisfies the dispersion relation

$$k_\alpha^{(3)}k^{(3)\alpha}=0. \tag{5.2}$$

The triresonance condition (1.4) implies that the phases are related by

$$\theta_1+\theta_2+\theta_3=0. \tag{5.3}$$

If $h(x;\epsilon)=h(x,\theta_1,\theta_2,\theta_3)$, where the phase variables θ_J are given by (3.3) and (5.1), then it follows from the chain rule that

$$\partial_\alpha h=\frac{1}{\epsilon}\sum_{J=1}^{3} k_\alpha^{(J)}\,\partial_{\theta_J}h+\partial_\alpha h. \tag{5.4}$$

The partial derivative ∂_α on the right-hand side of (5.4) is taken with respect to x^α holding θ_J fixed. Using (5.4) to expand partial derivatives in the linearized gravitational wave operator $\mathcal{L}$ given in (2.6), we find that

$$\mathcal{L}[h]_{\alpha\beta}=\frac{1}{\epsilon^2}\mathcal{A}[h]_{\alpha\beta}+\frac{1}{\epsilon}\mathcal{B}[h]_{\alpha\beta}+\mathcal{C}[h]_{\alpha\beta}, \tag{5.5}$$

where the linear operators $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are given by

$$\mathcal{A}[h]_{\alpha\beta}=\frac{1}{2}\sum_{J,K=1}^{3}\partial_{\theta_J}\partial_{\theta_K}\{-k_\lambda^{(J)}k^{(K)\lambda}h_{\alpha\beta}-k_\alpha^{(J)}k_\beta^{(K)}h_\lambda^\lambda+k_\lambda^{(J)}k_\beta^{(K)}h_\alpha^\lambda+k_\alpha^{(J)}k_\lambda^{(K)}h_\beta^\lambda\}, \tag{5.6}$$

$$\begin{aligned}\mathcal{B}[h]_{\alpha\beta}=\frac{1}{2}\sum_{J=1}^{3}\partial_{\theta_J}\{&-2k^{(J)\lambda}\partial_\lambda h_{\alpha\beta}-(k_\alpha^{(J)}\partial_\beta+k_\beta^{(J)}\partial_\alpha)h_\lambda^\lambda\\&+(k_\lambda^{(J)}\partial_\beta+k_\beta^{(J)}\partial_\lambda)h_\alpha^\lambda+(k_\alpha^{(J)}\partial_\lambda+k_\lambda^{(J)}\partial_\alpha)h_\beta^\lambda\},\end{aligned} \tag{5.7}$$

$$\mathcal{C}[h]_{\alpha\beta}=\tfrac{1}{2}(-\partial_\lambda\,\partial^\lambda h_{\alpha\beta}-\partial_\alpha\,\partial_\beta h_\lambda^\lambda+\partial_\alpha\,\partial_\lambda h_\alpha^\lambda+\partial_\beta\,\partial_\lambda h_\beta^\lambda). \tag{5.8}$$

We look for an asymptotic expansion of the metric perturbation h as $\epsilon\to 0$ of the form

$$h_{\alpha\beta}=\overset{0}{h}_{\alpha\beta}(x)+\epsilon^2\overset{2}{h}_{\alpha\beta}(x)+\epsilon^3\overset{3}{h}_{\alpha\beta}(x,\theta_1,\theta_2,\theta_3)+\epsilon^4\overset{4}{h}_{\alpha\beta}(x,\theta_1,\theta_2,\theta_3)+O(\epsilon^5). \tag{5.9}$$

We use (3.6), (5.5), and (5.9) in (2.5), expand, and equate coefficients of ϵ^0, ϵ^1, and ϵ^2 in the result. This yields the perturbation equations

$$\mathcal{C}[\overset{0}{h}]_{\alpha\beta}=\overset{0}{T}_{\alpha\beta}-\tfrac{1}{2}\overset{0}{T}\eta_{\alpha\beta}, \tag{5.10}$$

$$\mathcal{A}[\overset{3}{h}]_{\alpha\beta}=\overset{1}{T}_{\alpha\beta}-\tfrac{1}{2}\overset{1}{T}\eta_{\alpha\beta}, \tag{5.11}$$

$$\mathcal{A}[\overset{4}{h}]_{\alpha\beta}+\mathcal{B}[\overset{3}{h}]_{\alpha\beta}+\mathcal{C}[\overset{2}{h}]_{\alpha\beta}=\overset{2}{T}_{\alpha\beta}-\tfrac{1}{2}\overset{2}{T}\eta_{\alpha\beta}. \tag{5.12}$$

Equation (5.10) is a linearized Einstein field equation for the leading-order metric perturbation $\overset{0}{h}$ in which the source term is the constant energy–momentum tensor of the unperturbed fluid. The solution describes the small background space–time curvature caused by the presence of the fluid. Since our aim is to determine the propagating part of the metric, we will not consider this equation further. The source term in Eq. (5.11) is a linear sum of nonresonant acoustic terms, and it can be solved without the appearance of secular terms. The solution includes a gravitational wave that satisfies the homogeneous equation. The source term in Eq. (5.12) contains resonant terms that are quadratically nonlinear in the sound wave amplitudes. The imposition of solvability conditions leads to an equation for the amplitude of the gravitational wave included in the solution of (5.11). This equation describes the generation of a gravitational wave by the sound wave interaction, and is our main result. We carry out the detailed computations in the next section.

VI. SOLUTION OF THE GRAVITATIONAL EQUATIONS

A solution of the order ϵ perturbation equation (5.11) has the form

$$\overset{3}{h}_{\alpha\beta}=f_{\alpha\beta}(x,\theta_1,\theta_2)+h'_{\alpha\beta}(x,\theta_3)+\overline{h}_{\alpha\beta}(x), \tag{6.1}$$

where $f_{\alpha\beta}$ is a particular solution of (5.11) that corresponds to an oscillatory metric perturbation forced by the acoustic waves, $h'_{\alpha\beta}$ is a solution of the homogeneous equation that corresponds to the metric perturbation of a gravitational wave, and $\overline{h}_{\alpha\beta}$ is an arbitrary function of x. We assume that $h'_{\alpha\beta}$ is 2π periodic in θ_3 and, without loss of generality, we choose $\overline{h}_{\alpha\beta}$ so that $f_{\alpha\beta}$ and $h'_{\alpha\beta}$ have zero mean with respect to the phase variables.

From (4.30) and (5.6), a particular solution of (5.11) is given by

$$f_{\alpha\beta}(x,\theta_1,\theta_2)=\sum_{J=1}^{2} c_J(x,\theta_J)F^{(J)}_{\alpha\beta}, \tag{6.2}$$

where $\partial^2_{\theta_J}c_J=b_J$ and

$$\begin{aligned}&\tfrac{1}{2}(-k^{(J)}_\lambda k^{(J)\lambda}F^{(J)}_{\alpha\beta}-k^{(J)}_\alpha k^{(J)}_\beta F^{(J)\lambda}_\lambda+k^{(J)}_\lambda k^{(J)}_\beta F^{(J)\lambda}_\alpha+k^{(J)}_\alpha k^{(J)}_\lambda F^{(J)\lambda}_\beta)\\&\quad=B_{\alpha\beta}(\mathbf{R}_J)-\tfrac{1}{2}B^\lambda_\lambda(\mathbf{R}_J)\eta_{\alpha\beta}.\end{aligned}$$

This algebraic equation for $F^{(J)}_{\alpha\beta}$ is nonsingular whenever the wave number vector $k^{(J)}$ is not a null vector. The precise form of the solution is unimportant for our purposes, so we omit a detailed expression.

The term $h'_{\alpha\beta}(x,\theta_3)$ in (6.1) satisfies the homogeneous equation

$$\mathcal{A}[h']_{\alpha\beta}=0. \tag{6.3}$$

Using (5.6), we can write (6.3) as

$$\tfrac{1}{2}\partial^2_{\theta_3}(k^{(3)}_\alpha \Phi_\beta + k^{(3)}_\beta \Phi_\alpha)=0, \tag{6.4}$$

where

$$\Phi_\alpha = k^{(3)\lambda} h'_{\alpha\lambda} - \tfrac{1}{2} k^{(3)}_\alpha h'^{\lambda}_{\ \lambda}. \tag{6.5}$$

Since $h'_{\alpha\beta}$ is a zero-mean, 2π-periodic function of θ_3, it follows from (6.4), after two integrations with respect to θ_3, that

$$\Phi_\alpha = 0. \tag{6.6}$$

We use Lorentzian coordinates in which the gravitational wave propagates in the $(-x^1)$ direction. The covariant wave number four-vector of the wave is then

$$(k^{(3)}_\alpha) = \kappa_3(1,1,0,0), \tag{6.7}$$

where κ_3 is a constant scalar. It follows from (6.5), (6.6), and (6.7) that the metric $h'_{\alpha\beta}$ has the form

$$(h'_{\alpha\beta}) = \begin{pmatrix} h'_{00} & (h'_{00}+h'_{11})/2 & h'_{02} & h'_{03} \\ (h'_{00}+h'_{11})/2 & h'_{11} & h'_{02} & h'_{03} \\ h'_{02} & h'_{02} & h'_{22} & h'_{23} \\ h'_{03} & h'_{03} & h'_{23} & -h'_{22} \end{pmatrix}.$$

We can eliminate h'_{00}, h'_{11}, h'_{02}, and h'_{03} by means of an appropriate gauge transformation.[11] The metric perturbation associated with the gravitational wave is then given by

$$(h'_{\alpha\beta}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & h'_{22} & h'_{23} \\ 0 & 0 & h'_{23} & -h'_{22} \end{pmatrix}. \tag{6.8}$$

Equations for the second-order mean-field perturbation $\overset{2}{h}_{\alpha\beta}$ and the third-order gravitational wave perturbation $h'_{\alpha\beta}$ follow from the requirement that Eq. (5.12) has a solution for $\overset{4}{h}_{\alpha\beta}$ that is a periodic function of the phase variables. There are two solvability conditions for (5.12). The first condition is obtained by averaging the equation with respect to all the phase variables, and leads to an equation for the perturbation in the background gravitational field caused by the mean energy–momentum density of the sound waves. The second condition is obtained by averaging (5.12) with respect to θ_1 subject to the constraint (5.3) with θ_3 held fixed, and leads to an equation for the amplitude of the gravitational wave. Similar arguments have been used in the study of resonant wave interactions for hyperbolic conservation laws.[12,13] To derive these solvability conditions, we first introduce some notation.

Suppose that $f(x,\theta_1,\theta_2,\theta_3)$ is a 2π-periodic function of each of the phase variables θ_J $(J=1,2,3)$. We define

$$\overline{f}(x) = \frac{1}{(2\pi)^3}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} f(x,\xi_1,\xi_2,\xi_3)\, d\xi_1 d\xi_2 d\xi_3, \tag{6.9}$$

$$\langle f\rangle(x,\theta_3) = \frac{1}{2\pi}\int_0^{2\pi} f(x,\xi,-\xi-\theta_3,\theta_3)\, d\xi. \tag{6.10}$$

Thus, $\overline{f}$ is the mean of f with respect to the phase variables $(\theta_1,\theta_2,\theta_3)$, while $\langle f\rangle$ is the mean of f with respect to θ_1, or θ_2, in which θ_3 is held fixed and the three phase variables satisfy the constraint (5.3).

It follows from these definitions that

$$\overline{\partial_{\theta_J} f}=0, \quad J=1,2,3. \tag{6.11}$$

If $F(x)$ is independent of the phases, then $\overline{F}=F$, and if $f_J(x,\theta_J)$ is a periodic function of a single phase variable θ_J, then

$$\langle f_J\rangle=\overline{f_J}, \quad J=1,2, \quad \langle f_3\rangle=f_3. \tag{6.12}$$

Since $f(x,\theta_1,\theta_2,\theta_3)$ is periodic, we have

$$\begin{aligned}\langle \partial_{\theta_1} f-\partial_{\theta_2} f\rangle &= \frac{1}{2\pi}\int_0^{2\pi}\{\partial_{\theta_1} f(x,\xi,-\xi-\theta_3,\theta_3)-\partial_{\theta_2} f(x,\xi,-\xi-\theta_3,\theta_3)\}d\xi\\ &=\frac{1}{2\pi}\int_0^{2\pi}\partial_\xi f(x,\xi,-\xi-\theta_3,\theta_3)\,d\xi\\ &=0.\end{aligned}$$

Hence,

$$\langle \partial_{\theta_1} f\rangle=\langle \partial_{\theta_2} f\rangle. \tag{6.13}$$

We write (5.12) as

$$\mathcal{A}[\overset{4}{h}]_{\alpha\beta}+\mathcal{B}[\overset{3}{h}]_{\alpha\beta}+\mathcal{C}[\overset{2}{h}]_{\alpha\beta}=S_{\alpha\beta}, \tag{6.14}$$

where

$$S_{\alpha\beta}=\overset{2}{T}_{\alpha\beta}-\tfrac{1}{2}\overset{2}{T}\eta_{\alpha\beta}. \tag{6.15}$$

From (5.6) and (5.7), we see that each term in $\mathcal{A}[h]_{\alpha\beta}$ and $\mathcal{B}[h]_{\alpha\beta}$ is a derivative with respect to a phase variable, so (6.11) implies that

$$\overline{\mathcal{A}[h]_{\alpha\beta}}=0, \quad \overline{\mathcal{B}[h]_{\alpha\beta}}=0.$$

Therefore, averaging (6.14) with respect to all the phase variables, we obtain the first solvability condition,

$$\mathcal{C}[\overset{2}{h}]_{\alpha\beta}=\overline{S}_{\alpha\beta}. \tag{6.16}$$

To derive an expression for $\overline{S}_{\alpha\beta}$, we use (4.31) in (6.15), which gives

$$\begin{aligned}\overline{S}_{\alpha\beta}=&\sum_{J=1}^{3}\overline{w}_J\left\{B_{\alpha\beta}(\mathbf{R}_J)-\frac{1}{2}B^{\lambda}_{\lambda}(\mathbf{R}_J)\,\eta_{\alpha\beta}\right\}\\ &+\sum_{J=1}^{2}\overline{b_J^2}\left\{C_{\alpha\beta}(\mathbf{R}_J,\mathbf{R}_J)-\frac{1}{2}C^{\lambda}_{\lambda}(\mathbf{R}_J,\mathbf{R}_J)\,\eta_{\alpha\beta}\right\}.\end{aligned} \tag{6.17}$$

Averaging (4.20) and (4.26)–(4.27), and supposing for simplicity that the arbitrary functions of integration f_J have zero mean, we find that

$$\overline{w_0}=\sum_{J=1}^{2} V_J\overline{b_J^2}, \quad \overline{w_J}=\frac{1}{2}Q_J\overline{b_J^2}, \quad J=1,2.$$

Use of these expressions in (6.17) gives

$$\overline{S}_{\alpha\beta}=\sum_{J=1}^{2} \overline{b_J^2}M_{\alpha\beta}^{(J)},$$

$$M_{\alpha\beta}^{(J)}=C_{\alpha\beta}(\mathbf{R}_J,\mathbf{R}_J)-\tfrac{1}{2}C_\lambda^\lambda(\mathbf{R}_J,\mathbf{R}_J)\eta_{\alpha\beta} + V_J\{B_{\alpha\beta}(\mathbf{R}_0)-\tfrac{1}{2}B_\lambda^\lambda(\mathbf{R}_0)\eta_{\alpha\beta}\}+\tfrac{1}{2}Q_J\{B_{\alpha\beta}(\mathbf{R}_J)-\tfrac{1}{2}B_\lambda^\lambda(\mathbf{R}_J)\eta_{\alpha\beta}\}.$$

Equation (6.16) is a linearized Einstein equation for the second-order metric perturbation in (5.9) in which the source term is given by the mean energy–momentum tensor of the sound waves. This mean energy–momentum tensor leads to a small change in the background curvature of space time, in addition to the background curvature caused by the energy–momentum tensor of the unperturbed fluid.

To derive the second solvability condition, we average Eq. (6.14) with respect to θ_1, subject to the constraint (5.3) with θ_3 held fixed. After subtracting the mean equation (6.16) from the result, we obtain that

$$\langle \mathcal{A}[\overset{4}{h}]\rangle_{\alpha\beta}+\langle \mathcal{B}[\overset{3}{h}]\rangle_{\alpha\beta}=\langle S'\rangle_{\alpha\beta}, \tag{6.18}$$

where

$$S'_{\alpha\beta}=S_{\alpha\beta}-\overline{S}_{\alpha\beta}. \tag{6.19}$$

We will compute explicit expressions for each of the means in (6.18). The final result is given in (6.34) below.

First, we compute $\langle S'\rangle_{\alpha\beta}$. From (4.31), (6.15), (6.17), and (6.19) we have

$$\langle S'\rangle_{\alpha\beta}=\sum_{J=1}^{3}\langle w_J'\rangle\left\{B_{\alpha\beta}(\mathbf{R}_J)-\frac{1}{2}B_\lambda^\lambda(\mathbf{R}_J)\eta_{\alpha\beta}\right\} +2\langle b_1b_2\rangle\{C_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)-\tfrac{1}{2}C_\lambda^\lambda(\mathbf{R}_1,\mathbf{R}_2)\eta_{\alpha\beta}\}, \tag{6.20}$$

where

$$w_J'=w_J-\overline{w}_J. \tag{6.21}$$

Averaging (4.20) and (4.26)–(4.27), we find that

$$\langle w_0'\rangle=W\langle b_1b_2\rangle, \quad \langle w_J'\rangle=(M_J+P_J)\langle b_1b_2\rangle, \quad J=1,2, \tag{6.22}$$

where W, M_J, and P_J are defined in (4.17). In deriving (6.22), we have used the fact that

$$\langle b_1b_2\rangle=\langle(\partial_{\theta_1}b_1)B_2\rangle=\langle B_1(\partial_{\theta_2}b_2)\rangle,$$

which follows from (6.13), since Eq. (4.23) implies that $\partial_{\theta_J}B_J=b_J$. Substitution of (6.22) into (6.20) gives

$$\langle S'\rangle_{\alpha\beta}=\langle b_1b_2\rangle J_{\alpha\beta}, \tag{6.23}$$

where

$$J_{\alpha\beta}=W\{B_{\alpha\beta}(\mathbf{R}_0)-\tfrac{1}{2}B_\lambda^\lambda(\mathbf{R}_0)\,\eta_{\alpha\beta}\}+\sum_{J=1}^{2}(M_J+P_J)\left\{B_{\alpha\beta}(\mathbf{R}_J)-\frac{1}{2}B_\lambda^\lambda(\mathbf{R}_J)\,\eta_{\alpha\beta}\right\}+2\{C_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)-\tfrac{1}{2}C_\lambda^\lambda(\mathbf{R}_1,\mathbf{R}_2)\,\eta_{\alpha\beta}\}. \tag{6.24}$$

Using (3.14), (3.15), (4.7), (4.12), and (4.17) in (6.24), and simplifying the result with the aid of (1.4) and (6.7), we get[10]

$$(J_{\alpha\beta})=\begin{pmatrix} J_{00} & (J_{00}+J_{11})/2 & J_{02} & J_{03}\\ (J_{00}+J_{11})/2 & J_{11} & J_{02} & J_{03}\\ J_{02} & J_{02} & J_{22} & J_{23}\\ J_{03} & J_{03} & J_{23} & -J_{22}\end{pmatrix},$$

$$J_{00}=\left(\frac{1+3c_s^2+q+7qc_s^2}{2c_s^4}\right)k_0^{(1)}k_0^{(2)},\quad J_{11}=\left(\frac{3-3c_s^2+3q+qc_s^2}{2c_s^4}\right)k_0^{(1)}k_0^{(2)},$$

$$J_{01}=\left(\frac{1+q+2qc_s^2}{c_s^4}\right)k_0^{(1)}k_0^{(2)},\quad J_{02}=-\frac{1}{c_s^2}k_2^{(1)}(k_0^{(1)}-k_0^{(2)}), \tag{6.25}$$

$$J_{03}=-\frac{1}{c_s^2}k_3^{(1)}(k_0^{(1)}-k_0^{(2)}),\quad J_{22}=k_2^{(1)}k_2^{(2)}-k_3^{(1)}k_3^{(2)},\quad J_{23}=k_2^{(1)}k_3^{(2)}+k_3^{(1)}k_2^{(2)}.$$

Next, we consider the terms on the left-hand side of (6.18). From (5.6),

$$\langle\mathcal{A}[\overset{4}{h}]\rangle_{\alpha\beta}=\frac{1}{2}\sum_{J,K=1}^{3}\langle\partial_{\theta_J}\partial_{\theta_K}(k_\lambda^{(J)}k_\beta^{(K)}\overset{4}{h}{}_\alpha^\lambda+k_\alpha^{(J)}k_\lambda^{(K)}\overset{4}{h}{}_\beta^\lambda-k_\alpha^{(J)}k_\beta^{(K)}\overset{4}{h}{}_\lambda^\lambda)\rangle. \tag{6.26}$$

If f is any periodic function of the phases, then use of (6.13) and the triresonance condition (1.4) implies that

$$\begin{aligned}\left\langle\sum_{J,K=1}^{3}k_\alpha^{(J)}k_\beta^{(K)}\partial_{\theta_J}\partial_{\theta_K}f\right\rangle&=(k_\alpha^{(1)}+k_\alpha^{(2)})\left\langle\sum_{K=1}^{3}k_\beta^{(K)}\partial_{\theta_1}\partial_{\theta_K}f\right\rangle+k_\alpha^{(3)}\left\langle\sum_{K=1}^{3}k_\beta^{(K)}\partial_{\theta_3}\partial_{\theta_K}f\right\rangle\\&=k_\alpha^{(3)}\left\langle\sum_{K=1}^{3}k_\beta^{(K)}\partial_{\theta_K}(-\partial_{\theta_1}f+\partial_{\theta_3}f)\right\rangle\\&=k_\alpha^{(3)}(k_\beta^{(1)}+k_\beta^{(2)})\langle\partial_{\theta_1}(-\partial_{\theta_1}f+\partial_{\theta_3}f)\rangle+k_\alpha^{(3)}k_\beta^{(3)}\langle\partial_{\theta_3}(-\partial_{\theta_1}f+\partial_{\theta_3}f)\rangle\\&=k_\alpha^{(3)}k_\beta^{(3)}\langle\partial_{\theta_1}^2f-2\partial_{\theta_1}\partial_{\theta_3}f+\partial_{\theta_3}^2f\rangle.\end{aligned}$$

Using this result to compute the averages in (6.26), we find that

$$\langle\mathcal{A}[\overset{4}{h}]\rangle_{\alpha\beta}=\hat{A}_{\alpha\beta}, \tag{6.27}$$

where

$$\hat{A}_{\alpha\beta}=k_\lambda^{(3)}k_\beta^{(3)}\hat{h}_\alpha^\lambda+k_\alpha^{(3)}k_\lambda^{(3)}\hat{h}_\beta^\lambda-k_\alpha^{(3)}k_\beta^{(3)}\hat{h}_\lambda^\lambda, \tag{6.28}$$

$$\hat{h}_{\alpha\beta}=\tfrac{1}{2}\langle\partial^2_{\theta_1}\overset{4}{h}_{\alpha\beta}-2\partial_{\theta_1}\partial_{\theta_3}\overset{4}{h}_{\alpha\beta}+\partial^2_{\theta_3}\overset{4}{h}_{\alpha\beta}\rangle. \tag{6.29}$$

We can rewrite (6.28) as

$$\hat{A}_{\alpha\beta}=k^{(3)}_\alpha\Psi_\beta+k^{(3)}_\beta\Psi_\alpha,\quad \Psi_\alpha=k^{(3)}_\lambda\hat{h}^\lambda_\alpha-\tfrac{1}{2}k^{(3)}_\alpha\hat{h}^\lambda_\lambda. \tag{6.30}$$

The use of (6.7) in (6.30) gives

$$(\hat{A}_{\alpha\beta})=\begin{pmatrix}\hat{A}_{00} & (\hat{A}_{00}+\hat{A}_{11})/2 & \hat{A}_{02} & \hat{A}_{03}\\ (\hat{A}_{00}+\hat{A}_{11})/2 & \hat{A}_{11} & \hat{A}_{02} & \hat{A}_{03}\\ \hat{A}_{02} & \hat{A}_{02} & 0 & 0\\ \hat{A}_{03} & \hat{A}_{03} & 0 & 0\end{pmatrix}, \tag{6.31}$$

$$\hat{A}_{00}=2\kappa_3\Psi_0,\quad \hat{A}_{11}=2\kappa_3\Psi_1,\quad \hat{A}_{02}=\kappa_3\Psi_2,\quad \hat{A}_{03}=\kappa_3\Psi_3.$$

Finally, from (5.7), (6.1), (6.2), (6.11), and (6.12) we have

$$\langle\mathcal{B}[\overset{3}{h}]\rangle_{\alpha\beta}=\mathcal{B}[h']_{\alpha\beta}. \tag{6.32}$$

The use of (5.7), (6.7), and (6.8) implies that

$$\mathcal{B}[h']_{\alpha\beta}=\hat{B}_{\alpha\beta},$$

$$\hat{B}_{\alpha\beta}=\begin{pmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & \hat{B}_{22} & \hat{B}_{23}\\ 0 & 0 & \hat{B}_{23} & -\hat{B}_{22}\end{pmatrix}, \tag{6.33}$$

$$\hat{B}_{22}=\kappa_3\partial_{\theta_3}(\partial_0-\partial_1)h'_{22},\quad \hat{B}_{23}=\kappa_3\partial_{\theta_3}(\partial_0-\partial_1)h'_{23}.$$

Using (6.23), (6.25), (6.27), (6.31), (6.32), and (6.33), we may rewrite Eq. (6.18) as

$$\begin{pmatrix}\hat{A}_{00} & (\hat{A}_{00}+\hat{A}_{11})/2 & \hat{A}_{02} & \hat{A}_{03}\\ (\hat{A}_{00}+\hat{A}_{11})/2 & \hat{A}_{11} & \hat{A}_{02} & \hat{A}_{03}\\ \hat{A}_{02} & \hat{A}_{02} & 0 & 0\\ \hat{A}_{03} & \hat{A}_{03} & 0 & 0\end{pmatrix}+\begin{pmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & \hat{B}_{22} & \hat{B}_{23}\\ 0 & 0 & \hat{B}_{23} & -\hat{B}_{22}\end{pmatrix}$$

$$=\langle b_1b_2\rangle\begin{pmatrix}J_{00} & (J_{00}+J_{11})/2 & J_{02} & J_{03}\\ (J_{00}+J_{11})/2 & J_{11} & J_{02} & J_{03}\\ J_{02} & J_{02} & J_{22} & J_{23}\\ J_{03} & J_{03} & J_{23} & -J_{22}\end{pmatrix}. \tag{6.34}$$

Equating the (2,3)-components in (6.34), we obtain the second solvability condition for (6.14),

$$\begin{pmatrix}\hat{B}_{22} & \hat{B}_{23}\\ \hat{B}_{23} & -\hat{B}_{22}\end{pmatrix}=\langle b_1b_2\rangle\begin{pmatrix}J_{22} & J_{23}\\ J_{23} & -J_{22}\end{pmatrix}. \tag{6.35}$$

The remaining components of (6.34) can be satisfied by a suitable choice of the higher-order metric perturbation $\overset{4}{h}$. The use of (6.25) and (6.33) in (6.35) implies that the metric components of the gravitational wave satisfy

$$\kappa_3 \partial_{\theta_3}(\partial_0 - \partial_1)\begin{pmatrix} h'_{22} & h'_{23} \\ h'_{23} & -h'_{22} \end{pmatrix} = \langle b_1 b_2 \rangle \begin{pmatrix} k_2^{(1)}k_2^{(2)} - k_3^{(1)}k_3^{(2)} & k_2^{(1)}k_3^{(2)} + k_3^{(1)}k_2^{(2)} \\ k_2^{(1)}k_3^{(2)} + k_3^{(1)}k_2^{(2)} & -(k_2^{(1)}k_2^{(2)} - k_3^{(1)}k_3^{(2)}) \end{pmatrix}. \quad (6.36)$$

It follows from (6.8) and (6.36) that the gravitational wave generated by the resonant interaction is plane polarized with a metric perturbation given by

$$(h'_{\alpha\beta}) = b_3(x,\theta_3)\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & k_2^{(1)}k_2^{(2)} - k_3^{(1)}k_3^{(2)} & k_2^{(1)}k_3^{(2)} + k_3^{(1)}k_2^{(2)} \\ 0 & 0 & k_2^{(1)}k_3^{(2)} + k_3^{(1)}k_2^{(2)} & -(k_2^{(1)}k_2^{(2)} - k_3^{(1)}k_3^{(2)}) \end{pmatrix}, \quad (6.37)$$

where the gravitational wave amplitude b_3 satisfies the equation

$$\kappa_3 \partial_{\theta_3}(\partial_0 b_3 - \partial_1 b_3) = \langle b_1 b_2 \rangle. \quad (6.38)$$

When the spatial acoustic wave number vectors lie in the (x^1, x^2) plane, so that $k_3^{(1)} = k_3^{(2)} = 0$, the expression in (6.37) for the gravitational wave metric reduces to

$$(h'_{\alpha\beta}) = a_3(x,\theta_3)\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad a_3 = k_2^{(1)}k_2^{(2)} b_3. \quad (6.39)$$

From (6.39), the metric perturbation $h_{\alpha\beta} = \epsilon^3 h'_{\alpha\beta}$ associated with the gravitational wave is given by (1.16), after we remove the expansion parameter ϵ by setting $\epsilon = 1$, and restore dimensioned variables. Moreover, using (6.39) in (6.38), and changing $x^1 \to -x^1$ to obtain equations for a gravitational wave that propagates in the x^1 direction, we get Eq. (1.17) stated in the Introduction, with $\omega^{(3)} = -\kappa_3$.

ACKNOWLEDGMENTS

The work of J.K.H. was supported by the National Science Foundation under Grant No. DMS-9704155. We thank G. Alì for a number of helpful discussions.

[1] Y. Choquet-Bruhat and A. H. Taub, Gen. Relativ. Gravit. **8**, 561 (1977).
[2] Y. Choquet-Bruhat, in *Ondes et Radiation Gravitationelles*, edited by A. Lichnerowicz (Editions du CNRS, Paris, 1973).
[3] Y. Choquet-Bruhat and A. Greco, J. Math. Phys. **24**, 377 (1983).
[4] V. Pirronello, Astrophys. Space Sci. **102**, 11 (1984).
[5] V. E. Zakharov, V. S. L'vov, and G. Falkovich, *Kolmogorov Spectra of Turbulence I* (Springer-Verlag, Berlin, 1992).
[6] K. S. Thorne, ''Gravitational radiation,'' in *300 Years in Gravitation*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1987).
[7] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984).
[8] Y. Choquet-Bruhat, J. Math. Pures Appl. **48**, 117 (1969).
[9] J. K. Hunter and J. B. Keller, Commun. Pure Appl. Math. **36**, 547 (1983).
[10] B. Truong, ''Generation of gravitational waves by acoustic wave interactions,'' Ph.D. thesis, University of California at Davis, 1997.
[11] A. M. Anile, *Relativistic Fluids and Magneto-Fluids* (Cambridge University Press, Cambridge, 1989).
[12] A. J. Majda and R. R. Rosales, Stud. Appl. Math. **71**, 149 (1984).
[13] J. K. Hunter, A. J. Majda, and R. R. Rosales, Stud. Appl. Math. **75**, 187 (1986).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Gravitational waves in vacuum spacetimes with cosmological constant. I. Classification and geometrical properties of nontwisting type *N* solutions

Jiří Bičák[a] and Jiří Podolský[b]
Department of Theoretical Physics, Faculty of Mathematics and Physics, Charles University, V Holešovičkách 2, 180 00 Prague 8, Czech Republic

(Received 31 July 1998; accepted for publication 25 May 1999)

All nontwisting Petrov-type N solutions of vacuum Einstein field equations with cosmological constant Λ are summarized. They are shown to belong either to the nonexpanding Kundt class or to the expanding Robinson–Trautman class. Invariant subclasses of each class are defined and the corresponding metrics are given explicitly in suitable canonical coordinates. Relations between the subclasses and their geometrical properties are analyzed. In the subsequent paper these solutions are interpreted as exact gravitational waves propagating in de Sitter or anti-de Sitter spacetimes. © *1999 American Institute of Physics.* [S0022-2488(99)00509-5]

I. INTRODUCTION AND SUMMARY

Our purpose in this and the subsequent paper is to analyze all nontwisting type N solutions of Einstein's vacuum equations with Λ. There are several basic works on these solutions available in literature, in particular Refs. 1–3 (for pre-1980 works, see Ref. 4). None of them, however, discusses the physical interpretation of the solutions. Such an interpretation, based on the study of the deviation of geodesics, will be presented in the following paper. In this part we summarize, compare, classify, and generalize the mathematical results of Refs. 1–3.

We consider type N solutions in which the Debever–Penrose null vector field $\mathbf{k}$ is quadruple.[4] The vector field defines a congruence of null geodesics $x^\alpha(v)$ such that $dx^\alpha/dv=k^\alpha$, $k_\alpha k^\alpha=0$, $k_{\alpha;\beta}k^\beta=0$, v being an affine parameter. In general, a geodesic congruence is characterized by its expansion, $\Theta=\frac{1}{2}k^\alpha_{;\alpha}$, shear $|\sigma|=\sqrt{\frac{1}{2}k_{(\alpha;\beta)}k^{\alpha;\beta}-\Theta^2}$ and twist $\omega=\sqrt{\frac{1}{2}k_{[\alpha;\beta]}k^{\alpha;\beta}}$.[4] The Bianchi identities and the Kundt-Thompson theorem for type N solutions (see Ref. 5, and Theorem 7.5 in Ref. 4) imply $\sigma=0$ since $C^{\;;\delta}_{\alpha\beta\gamma\delta}=R^{\;;\delta}_{\alpha\beta\gamma\delta}=0$ for solutions of $R_{\alpha\beta}=\Lambda g_{\alpha\beta}$. In the following we assume $\omega=0$. Therefore, with (possibly) nonvanishing Λ we are left with two cases to consider: (i) the Kundt class of nonexpanding gravitational waves, $\Theta=0$ (cf. Ref. 6, Chap. 27 in Ref. 4); and (ii) the Robinson–Trautman class of expanding gravitational waves of type N, $\Theta\neq 0$ (cf. Ref. 7, Chap. 24 in Ref. 4). Hereafter we denote the Kundt class by $KN(\Lambda)$ and the Robinson–Trautman class by $RTN(\Lambda)$.

In Sec. II we analyze the class $KN(\Lambda)$ in detail. We do not present any new solution of this type, but we extend the results of Ref. 3 by giving the explicit forms of the transformations that leave the metrics of $KN(\Lambda)$ invariant. These enable us to give the *explicit* transformations to different canonical subclasses of $KN(\Lambda)$, which have not been given in the literature so far. We also introduce a convenient new notation for these subclasses and demonstrate how they are interconnected. In particular, we show that one of the subclasses is identical to the "Lobatchewski waves" studied by Siklos.[8] We also formulate the proposition (proven in the Appendix) that all vacuum solutions with Λ that are conformal to the "Kundt waves with $\Lambda=0$" belong to one specific subclass of $KN(\Lambda)$; we thus generalize the result of an analogous theorem for the *pp* waves.[8]

[a]Electronic mail: bicak@mbox.troja.mff.cuni.cz
[b]Electronic mail: podolsky@mbox.troja.mff.cuni.cz

In Sec. III the Robinson–Trautman solutions[7] of type N with Λ are discussed. The transformation between two standard forms of the metric (those of Refs. 4 and 1) is presented, and then transformations preserving the metric form given in Ref. 1 are generalized and used to demonstrate how the nonradiative part of the structural function of these solutions can be transformed away.

II. THE KUNDT CLASS OF SOLUTIONS $KN(\Lambda)$

This class has been investigated in detail by Ozsváth, Robinson, and Rózga.[3] They have shown that in this case the vector $\mathbf{k}$ can be normalized such that $\frac{1}{2}\mathcal{L}_{\mathbf{k}}g_{\alpha\beta}=\frac{1}{2}(k_{\alpha;\beta}+k_{\beta;\alpha})=Lk_\alpha k_\beta$, where $\mathcal{L}_{\mathbf{k}}$ is the Lie derivative and L is a scalar. Denoting $L'=\mathcal{L}_{\mathbf{k}}L$, we find $\mathcal{L}_{\mathbf{k}}L'=0$ so that L' is invariant under the renormalization of $\mathbf{k}$. A suitable coordinate system $(v,\xi,\bar{\xi},u)$, where $\xi,\bar{\xi}$ are space-like coordinates, v is a parameter along the null geodesics, and u is a retarded time with u=const. being a wavefront, $\mathbf{k}\equiv\psi(u)\partial_v$, can be introduced[3] in which the $KN(\Lambda)$ metrics have the form

$$ds^2=2\frac{1}{p^2}d\xi\, d\bar{\xi}-2\frac{q^2}{p^2}du\, dv+F\, du^2, \tag{1}$$

where

$$p=1+\frac{\Lambda}{6}\xi\bar{\xi},\quad q=\left(1-\frac{\Lambda}{6}\xi\bar{\xi}\right)\alpha+\bar{\beta}\xi+\beta\bar{\xi},$$

$$F=\kappa\frac{q^2}{p^2}v^2-\frac{(q^2)_{,u}}{p^2}v-\frac{q}{p}H,\quad \kappa=\frac{\Lambda}{3}\alpha^2+2\beta\bar{\beta}.$$

Here $\alpha(u)$ and $\beta(u)$ are *arbitrary* real and complex functions of u, respectively. These functions play the role of two arbitrary ''parameters,'' i.e., we can denote the Kundt class by $KN(\Lambda)\equiv KN(\Lambda)[\alpha,\beta]$. The parameter κ is related to the invariant L' by

$$L'=\kappa\frac{p^2}{q^2}. \tag{2}$$

The function $H=H(\xi,\bar{\xi},u)$ entering F is restricted by Einstein's equations, $H_{,\xi\bar{\xi}}+(\Lambda/3p^2)H=0$. There exists a general solution to this equation,

$$H(\xi,\bar{\xi},u)=(f_{,\xi}+\bar{f}_{,\bar{\xi}})-\frac{\Lambda}{3p}(\bar{\xi}f+\xi\bar{f}), \tag{3}$$

where $f(\xi,u)$ is an arbitrary function of ξ and u, analytic in ξ. The space–time is conformally flat if and only if the structural function H is of the form

$$H=H_c=\frac{1}{p}\left[\left(1-\frac{\Lambda}{6}\xi\bar{\xi}\right)\mathcal{A}+\bar{\mathcal{B}}\xi+\mathcal{B}\bar{\xi}\right], \tag{4}$$

with $\mathcal{A}(u)$ and $\mathcal{B}(u)$ being arbitrary real and complex functions, respectively. Since H_c of this form corresponds to (3) for f quadratic in ξ we easily infer the following.

Lemma 1: The $KN(\Lambda)$ solutions (1), (3) with $f=f_c=c_0(u)+c_1(u)\xi+c_2(u)\xi^2$, where $c_i(u)$ are arbitrary complex functions of u, are isometric to Minkowski (if $\Lambda=0$), de Sitter $(\Lambda>0)$ and anti-de Sitter spacetime $(\Lambda<0)$.

It can be proven by straightforward but tedious calculations that the following lemma (which is not formulated in Ref. 3 but is a consequence of results therein) is true.

Lemma 2: The metric of the $KN(\Lambda)[\alpha,\beta]$ class preserves its form (1) under the transformations $(v,\xi,\bar{\xi},u)\rightarrow(w,\eta,\bar{\eta},t)$, given by

$$v=a(t)\left[w+\frac{\left(1-\frac{\Lambda}{6}\eta\bar{\eta}\right)\gamma+\delta\bar{\eta}+\bar{\delta}\eta}{\left(1-\frac{\Lambda}{6}\eta\bar{\eta}\right)\alpha'+\beta'\bar{\eta}+\bar{\beta}'\eta}\right]+\Delta(t),$$

$$\xi=\frac{\bar{B}(t)+A(t)\eta}{\bar{A}(t)-\frac{\Lambda}{6}B(t)\eta},\quad u=u(t), \tag{5}$$

where $A(t),B(t)$ are arbitrary complex and $a(t),u(t)$ are real functions of t, respectively.

In the new coordinates $(w,\eta,\bar{\eta},t)$ the resulting metric $KN(\Lambda)[\alpha',\beta']$ has

$$\alpha'=\frac{\sqrt{a\dot{u}}}{\Phi}\left[\left(A\bar{A}-\frac{\Lambda}{6}B\bar{B}\right)\alpha+\bar{A}B\beta+A\overline{B\beta}\right],$$

$$\beta'=\frac{\sqrt{a\dot{u}}}{\Phi}\left[-\frac{\Lambda}{3}\overline{AB}\alpha+\bar{A}^2\beta-\frac{\Lambda}{6}\bar{B}^2\bar{\beta}\right], \tag{6}$$

with $\Phi=A\bar{A}+(\Lambda/6)B\bar{B}$ and the dot denotes d/dt. The remaining real functions $\gamma(t),\Delta(t)$ and a complex function $\delta(t)$ in (5) must satisfy the equations

$$\alpha'\delta-\beta'\gamma=\frac{1}{\Phi}(\bar{A}\dot{\bar{B}}-\dot{\bar{A}}\bar{B})\equiv C,$$

$$\bar{\beta}'\delta-\beta'\bar{\delta}=\frac{1}{\Phi}\left(\dot{A}\bar{A}-A\dot{\bar{A}}+\frac{\Lambda}{6}\dot{B}\bar{B}-\frac{\Lambda}{6}B\dot{\bar{B}}\right)\equiv D, \tag{7}$$

$$\frac{\kappa'}{a}\Delta=\frac{1}{2}\left(\frac{\dot{a}}{a}-\frac{\ddot{u}}{\dot{u}}\right)-\left(\frac{\Lambda}{3}\alpha'\gamma+\bar{\beta}'\delta+\beta'\bar{\delta}\right).$$

The structural function $H'(\eta,\bar{\eta},t)$ then takes the form

$$p'H'=\left(1-\frac{\Lambda}{6}\eta\bar{\eta}\right)E+\bar{F}\eta+F\bar{\eta}+\frac{\dot{u}}{\Phi}\sqrt{\frac{\dot{u}}{a}}\left(A-\frac{\Lambda}{6}\overline{B\eta}\right)\left(\bar{A}-\frac{\Lambda}{6}B\eta\right)pH, \tag{8}$$

where

$$E(t)\equiv2\dot{\gamma}+\frac{2}{a}(\alpha'\Delta)^{\cdot}-\alpha'\kappa'\frac{\Delta^2}{a^2}-2\frac{\Delta}{a}\left[\frac{1}{2}\alpha'\left(\frac{\dot{a}}{a}+\frac{\ddot{u}}{\dot{u}}\right)+C\bar{\beta}'+\bar{C}\beta'\right]$$
$$-\alpha'\left(2\delta\bar{\delta}-\frac{\Lambda}{3}\gamma^2\right)+2\gamma(\beta'\bar{\delta}+\bar{\beta}'\delta),$$

$$F(t)\equiv 2\dot{\delta}+\frac{2}{a}(\beta'\Delta)^{\cdot}-\beta'\kappa'\frac{\Delta^2}{a^2}-2\frac{\Delta}{a}\left[\frac{1}{2}\beta'\left(\frac{\dot a}{a}+\frac{\ddot u}{\dot u}\right)-\frac{\Lambda}{3}C\alpha'-D\beta'\right]$$
$$-\frac{\Lambda}{3}\gamma(\beta'\gamma-2\alpha'\delta)+2\overline{\beta'}\delta^2.$$

The transformation (5) is also important in connection with Lemma 1. Comparing Eq. (4) with Eq. (8), we see that we can 'generate' $H=H_c$ from $H=0$ by a coordinate transformation. Therefore, the conformally flat part, H_c, of H generated by f_c cannot represent a radiative field. Only if the function f is at least cubic in ξ, the resulting spacetime is of type N and can be interpreted as radiative (see the subsequent paper).

Another important application of the coordinate freedom (5) was suggested in Ref. 3: it is possible to use the transformation to get 'canonical' subclasses of $KN(\Lambda)[\alpha,\beta]$ corresponding to special values of parameters α and β. Without loss of generality we can assume $\alpha\geqslant 0$. In addition, the transformation (5) of a special form,

$$\xi=\eta,\quad v=a(t)w,\quad u=\int a(t)dt, \tag{9}$$

for $a(t)\neq 0$ results in scaling $\alpha(u)=\alpha'(t)/|a(t)|$, $\beta(u)=\beta'(t)/|a(t)|$ so that *we can always assume either* $\alpha=1$ *or* $\alpha=0$. Since the parameter L' (2) is invariant, the sign of κ is also an invariant. We can thus base the invariant canonical classification of all $KN(\Lambda)$ solutions on the sign of κ and the sign of Λ. There are nine possible cases (κ and Λ can both be positive, zero or negative). However, subclasses $\kappa<0,\Lambda>0$ and $\kappa=0,\Lambda>0$ and $\kappa<0,\Lambda=0$ are forbidden since they violate the relation $\kappa=(\Lambda/3)\alpha^2+2\beta\bar{\beta}$. The remaining possibilities are six subclasses which we shall now discuss.

A. Subclass $\kappa=0,\Lambda=0$

The equation $\kappa=0$ implies $\beta=0$ so that $\alpha=1$. A canonical representative of this subclass can be denoted as $PP\equiv KN(\Lambda=0)[\alpha=1,\beta=0]$. We are now using the notation PP since the corresponding metric,

$$PP:\ ds^2=2\,d\xi\,d\bar{\xi}-2\,du\,dv-(g+\bar{g})du^2, \tag{10}$$

with arbitrary $g(\xi,u)=f_{,\xi}$ analytic in ξ, describes well-known *pp* waves investigated by many authors (for details and references see Sec. 21.5 in Ref. 4).

B. Subclass $\kappa>0,\Lambda=0$

Since $0<\kappa=2\beta\bar{\beta}$, we have $\beta\neq 0$. Then Eqs. (5) with $A=\sqrt{\beta}$, $B=-(\alpha/2\beta)+i(J/\beta))A$, $a=1/\beta\bar{\beta}$, $\gamma=(1/2)(\alpha/\sqrt{\beta\bar{\beta}})^{\cdot}+(2J/\alpha)(J/\sqrt{\beta\bar{\beta}})^{\cdot}$, $\delta=\dot{A}/A$, $\Delta=\frac{1}{2}\dot{a}$, $u=t$, where J is a real function satisfying $(J/\sqrt{\beta\bar{\beta}})^{\cdot}=(i/4)(\alpha/\sqrt{\beta\bar{\beta}})(\bar{\beta}/\beta)(\beta/\bar{\beta})^{\cdot}$, transform any solution of this subclass to $KN\equiv KN(\Lambda=0)[\alpha=0,\ \beta=1]$. The metric,

$$KN:\ ds^2=2\,d\xi\,d\bar{\xi}+2(\xi+\bar{\xi})^2\left[-dv+\left(v^2-\frac{g+\bar{g}}{\xi+\bar{\xi}}\right)du\right]du, \tag{11}$$

with arbitrary $g(\xi,u)=\frac{1}{2}f_{,\xi}$, describes solutions discovered by Kundt.[6] This is a *special* ''Kundt solution'' in the ''Kundt class'' of nonexpanding waves. For details see Ref. 4, Chap. 27; the transformation between $\tilde{v}$ used there and v above is $v=\tilde{v}/(\xi+\bar{\xi})^2$.

C. Subclass $\kappa>0,\Lambda>0$

If $\alpha=1$ we can use (5) with $A=\exp(-(i/2)\chi)$, $B=cA$, $a=(\Lambda/6+\beta\bar{\beta})^{-1}$, $\gamma=-[\dot{W}+\frac{1}{2}(\dot{V}\exp(-i\chi)+\dot{\bar{V}}\exp(i\chi))]/(1+(\Lambda/6)c\bar{c})$, $\delta=\dot{A}/A+(\Lambda/6)\dot{c}\bar{c}/(1+(\Lambda/6)c\bar{c})$, $\Delta=\frac{1}{4}\dot{a}-(\Lambda/12)\times(c\bar{c})^{\cdot}(1+(\Lambda/6)c\bar{c})^{-1}a$, $u=t$, where $c\equiv V+W\bar{A}/A$, $V\equiv(6/\Lambda)\bar{\beta}$, $W\equiv-(6/\Lambda)\sqrt{\Lambda/6+\beta\bar{\beta}}$, χ is a real function satisfying $\dot{\chi}=(i/2)(\dot{V}\exp(-i\chi)-\dot{\bar{V}}\exp(i\chi))/W$ to obtain $\alpha'=0$, $\beta'=1$. If $\alpha=0$ the scaling (9) makes $|\beta|=1$ and then we put $\xi=(6/\Lambda)\beta/\eta$, $v=w+\frac{1}{2}(\dot{\beta}\bar{\beta}\eta+\beta\dot{\bar{\beta}}\bar{\eta})/(\eta+\bar{\eta})$, $u=t$. In both cases the representative of this subclass is $KN(\Lambda)I\equiv KN(\Lambda)[\alpha=0,\beta=1]$. The metric reads as

$$KN(\Lambda)I{:}\,ds^2=2\frac{d\xi\,d\bar{\xi}}{\left(1+\frac{\Lambda}{6}\xi\bar{\xi}\right)^2}-2\left(\frac{\xi+\bar{\xi}}{1+\frac{\Lambda}{6}\xi\bar{\xi}}\right)^2 du\,dv+\left[2\left(\frac{\xi+\bar{\xi}}{1+\frac{\Lambda}{6}\xi\bar{\xi}}\right)^2 v^2-\frac{\xi+\bar{\xi}}{1+\frac{\Lambda}{6}\xi\bar{\xi}}H\right]du^2. \tag{12}$$

$KN(\Lambda)I$ indicates that this solution is a generalization of the KN waves to the case $\Lambda\neq0$, ''I'' means ''of the first kind.'' The $KN(\Lambda)I$ solutions were first discovered by García Díaz and Plebański,[1,9] transformation between their coordinates and those used here is $\xi=\sqrt{6/\Lambda}\tanh\sqrt{\Lambda/6}\tilde{\xi}$, $v=r/\sinh\sqrt{\Lambda/6}(\tilde{\xi}+\bar{\tilde{\xi}})$, $u=-\sqrt{\Lambda/6}t$. However, in Refs. 1 and 9 not all $KN(\Lambda\neq0)$ solutions were found since (invariantly different) Subclasses E and F mentioned below were omitted.

D. Subclass $\kappa>0,\Lambda<0$

The same transformation as in the previous case leads to the metric (12) which has thus the same form for $\Lambda<0$. The transformation to the coordinates used in Ref. 1 is $\xi=\sqrt{-6/\Lambda}\tan\sqrt{-\Lambda/6}\tilde{\xi}$, $v=r/\sin\sqrt{-\Lambda/6}(\tilde{\xi}+\bar{\tilde{\xi}})$, $u=-\sqrt{-\Lambda/6}t$.

E. Subclass $\kappa<0,\Lambda<0$

Now $\kappa<0$ implies $\alpha=1$. Using (5) with $A=\exp(i\phi)$, $B=cA$, $a=(1+(6/\Lambda)\beta\bar{\beta})^{-1}$, $\gamma=0$, $\delta=\bar{A}^2\dot{\bar{c}}/(1+(\Lambda/6)c\bar{c})$, $\Delta=(3/2\Lambda)\dot{a}$, $u=t$, where $c\equiv(1/\beta)(\sqrt{1+(6/\Lambda)\beta\bar{\beta}}-1)$, ϕ is a real function satisfying $\dot{\phi}=i(\Lambda/12)(\dot{c}\bar{c}-c\dot{\bar{c}})/(1+(\Lambda/6)c\bar{c})$, we find $\beta'=0$ and the canonical representative, denoted by $KN(\Lambda^-)II\equiv KN(\Lambda<0)[\alpha=1,\beta=0]$, is

$$KN(\Lambda^-)II{:}\; ds^2=2\frac{d\xi\,d\bar{\xi}}{\left(1+\frac{\Lambda}{6}\xi\bar{\xi}\right)^2}-2\left(\frac{1-\frac{\Lambda}{6}\xi\bar{\xi}}{1+\frac{\Lambda}{6}\xi\bar{\xi}}\right)^2 du\,dv+\left[\frac{\Lambda}{3}\left(\frac{1-\frac{\Lambda}{6}\xi\bar{\xi}}{1+\frac{\Lambda}{6}\xi\bar{\xi}}\right)^2 v^2-\frac{1-\frac{\Lambda}{6}\xi\bar{\xi}}{1+\frac{\Lambda}{6}\xi\bar{\xi}}H\right]du^2. \tag{13}$$

Here $KN(\Lambda^-)II$ means generalized Kundt waves ''of the second kind'' with $\Lambda<0$. This class was first discovered by Ozsváth *et al.*[3] Observe that with $\kappa>0$, $\Lambda>0$, $KN(\Lambda^+)II\equiv KN(\Lambda^+)I$, since transformation (5) where $A=1/\sqrt{2}$, $B=\sqrt{3/\Lambda}$, $a=1$, $\gamma=\delta=\Delta=0$, $u=(\Lambda/6)t$ identifies $KN(\Lambda^+)[\alpha=0,\beta=1]$ with $KN(\Lambda^+)[\alpha'=1,\beta'=0]$.

F. Subclass $\kappa=0,\Lambda<0$

The relation $\kappa=0$ implies $\alpha=1$. Thus, $|\beta|=\sqrt{-\Lambda/6}$ and the representative can be denoted by $KN(\Lambda^-)III\equiv KN(\Lambda<0)[\alpha=1,\beta=\sqrt{-\Lambda/6}e^{i\omega(u)}]$, $\omega(u)$ being an arbitrary real function of u. The metric takes the form of (1) with $q=(1+\sqrt{-\Lambda/6}\xi e^{-i\omega(u)})(1+\sqrt{-\Lambda/6}\bar{\xi}e^{i\omega(u)})$. This class was also first discovered in Ref. 3. One can distinguish *two subsubclasses* of $KN(\Lambda^-)III$ according to the value of L: 1. $L\neq 0$ (**k** is not a Killing vector). 2. $L=0$ (**k** is a Killing vector). In case 2, β=const and the transformation $\xi=\sqrt{\beta/\bar{\beta}}\eta$ leads to $KN(\Lambda^-)III_K\equiv KN(\Lambda<0)[\alpha=1,\beta=\sqrt{-\Lambda/6}]$, where the suffix ''$K$'' stands for ''Killing:''

$$KN(\Lambda^-)III_{\rm K}:\ ds^2=2\frac{d\xi\, d\bar{\xi}}{\left(1+\frac{\Lambda}{6}\xi\bar{\xi}\right)^2}-2\left(\frac{\left(1+\sqrt{-\frac{\Lambda}{6}}\xi\right)\left(1+\sqrt{-\frac{\Lambda}{6}}\bar{\xi}\right)}{1+\frac{\Lambda}{6}\xi\bar{\xi}}\right)^2 du\, dv$$
$$-\frac{\left(1+\sqrt{-\frac{\Lambda}{6}}\xi\right)\left(1+\sqrt{-\frac{\Lambda}{6}}\bar{\xi}\right)}{1+\frac{\Lambda}{6}\xi\bar{\xi}}H\, du^2. \tag{14}$$

This subsubclass can be shown to be identical with the ''Lobatchevski waves'' studied by Siklos:[8] indeed, the transformation $\xi=-\sqrt{-6/\Lambda}(x+\frac{1}{2}+iy)/(x-\frac{1}{2}+iy)$, $v=(12/\Lambda)r$, brings the metric (14) into the form

$$ds^2=-\frac{3}{\Lambda}\cdot\frac{1}{x^2}(dx^2+dy^2+2\,du\,dr+\tilde{H}\,du^2), \tag{15}$$

where $\tilde{H}\equiv-(\Lambda/6)xH$. This is the Siklos metric.[8] Recently we analyzed in detail the behavior of test particles in these solutions and interpreted them as waves in the anti-de Sitter spacetime.[10] Impulsive waves of this type were investigated in Ref. 11.

G. Relations of the subclasses

Using the above results we can summarize the invariant canonical classification of the $KN(\Lambda)$ class of solutions in the following diagram:

$$KN(\Lambda)\begin{cases}\Lambda=0\begin{cases}\kappa=0: PP,\\ \kappa>0: KN,\end{cases}\\[2ex] \Lambda\neq 0\begin{cases}\kappa>0: KN(\Lambda)I,\\ \kappa<0: KN(\Lambda^-)II,\\ \kappa=0: KN(\Lambda^-)III\rightarrow KN(\Lambda^-)III_{\rm K}.\end{cases}\end{cases}$$

There is an asymmetry with respect to the sign of Λ: there are *three* distinct classes of nonexpanding waves for $\Lambda<0$ whereas there is only *one* such class for $\Lambda>0$. The reason is in the condition $\kappa=(\Lambda/3)\alpha^2+2\beta\bar{\beta}$ which for $\Lambda>0$ excludes the cases $\kappa<0$ and $\kappa=0$. Intuitively, fewer nonexpanding waves ''fit'' into the de Sitter universe which admits closed spacelike sections than into the anti-de Sitter space.

There exist natural relations between the $\Lambda=0$ and $\Lambda\neq 0$ subclasses. The metrics (12), (13) and (14) do not diverge as $\Lambda\rightarrow 0$, we can set $\Lambda=0$ and thus find

$$KN(\Lambda=0)I=KN,\quad KN(\Lambda^{-}=0)II=PP,\quad KN(\Lambda^{-}=0)III=PP. \tag{16}$$

Thus, it is natural to consider the $KN(\Lambda)I$ class as a generalization of the Kundt solution KN, and the classes $KN(\Lambda^{-})II$ and $KN(\Lambda^{-})III$ as generalizations of PP waves. There exists *no* generalization of PP waves to the case of $\Lambda>0$.

From metrics (10) and (15), with $\xi=(1/\sqrt{2})(x+iy)$, and from (11) and (12) we find $ds^2_{KN(\Lambda^{-})III_{\mathrm{K}}}=-(6/\Lambda)(\xi+\bar{\xi})^{-2}ds^2_{PP}$, and $ds^2_{KN(\Lambda)I}=(1+(\Lambda/6)\xi\bar{\xi})^{-2}ds^2_{KN}$. Therefore, the class $KN(\Lambda^{-})III_{\mathrm{K}}$ is conformal to the PP-class, and the class $KN(\Lambda)I$ is conformal to the KN-class. In fact, the solutions $KN(\Lambda^{-})III_{\mathrm{K}}$ and $KN(\Lambda)I$ are the only nontrivial spacetimes conformal to PP and KN, respectively. The theorem proven by Siklos[8] states that the only vacuum solutions (other than PP solutions themselves) which are properly (with a nonconstant factor) conformal to nonflat PP metrics are $KN(\Lambda^{-})III_{\mathrm{K}}$ metrics. However, we see from (16) that PP metrics are a special case of metrics $KN(\Lambda^{-})III_{\mathrm{K}}$ for $\Lambda=0$, and Siklos' theorem may just be formulated as follows.

Proposition 1: The only vacuum solutions conformal to nonflat PP metrics are $KN(\Lambda^{-})III_{\mathrm{K}}$ metrics.

In addition, the following analogous proposition can be proven for the KN solutions.

Proposition 2: The only vacuum solutions conformal to KN metrics are $KN(\Lambda)I$ metrics.

The proof is contained in the Appendix.

The conformal, homothetic and isometric symmetries of the $KN(\Lambda)$ solutions have been systematically investigated by Salazar, Garcı́a and Plebański[2] and by Siklos.[8] It is only in the subclasses PP and $KN(\Lambda^{-})III_{\mathrm{K}}$ that the vector $\mathbf{k}=\partial_v$ is a Killing vector; $k_{\alpha;\beta}=0$ only in the PP subclass. Let us finally summarize the classification of all $KN(\Lambda)$ solutions and compare our notation with notations used in the literature:

Notation in this paper	Notations in literature
$KN(\Lambda)$	$R(\Lambda,\alpha,\beta)$[3]
PP	pp, R[4,3] for Refs.
KN	K[6,4]
$KN(\Lambda)I$	$K(\Lambda)$[1,3]
$KN(\Lambda^{-})II$	$R(\Lambda)$[3]
$KN(\Lambda^{-})III$	$(IV)_1$[3]
$KN(\Lambda^{-})III_K$	$(IV)_0$[8,3]

Impulsive waves in the $KN(\Lambda)$ spacetimes were recently studied in Ref. 12.

III. THE ROBINSON–TRAUTMAN CLASS OF SOLUTIONS $RTN(\Lambda,\epsilon)$

The Robinson–Trautman solutions[7] satisfying the vacuum equations with Λ can be written as (see Ref. 4)

$$ds^2=2\frac{r^2}{P^2}d\zeta\,d\bar{\zeta}-2\,du\,dr-\left[\Delta\ln P-2r(\ln P)_{,u}-\frac{2m}{r}-\frac{\Lambda}{3}r^2\right]du^2, \tag{17}$$

where ζ is a complex spatial coordinate, r is an affine parameter along the rays generated by the vector field $\mathbf{k}$, u is a retarded time, m is a function of u which in some cases can be interpreted as mass, and $\Delta\equiv 2P^2\partial^2/\partial\zeta\partial\bar{\zeta}$. The function $P\equiv P(\zeta,\bar{\zeta},u)$ satisfies the equation $\Delta\Delta(\ln P)+12m(\ln P)_{,u}-4m_{,u}=0$. Here we restrict attention to the solutions of type N and denote these as $RTN(\Lambda)$. In this case $m=0$ and $\Delta\ln P=K(u)$. By a transformation $u=g(\tilde{u})$, $r=\tilde{r}/\dot{g}$, where $\dot{g}=dg/d\tilde{u}$, we can set the Gaussian curvature $K(u)$ of the 2-surfaces $2P^{-2}\,d\zeta\,d\bar{\zeta}$ to be $K=2\epsilon$, where $\epsilon=+1,0,-1$ (since $\tilde{P}=\dot{g}P$ and $\tilde{K}=\dot{g}^2K$, the sign of K is invariant). Thus, the different subclasses can be denoted as $RTN(\Lambda,\epsilon)$. The corresponding metrics can be written as

$$ds^2=2\frac{r^2}{P^2}d\zeta\, d\bar{\zeta}-2\,du\,dr-2\left[\epsilon-r(\ln P)_{,u}-\frac{\Lambda}{6}r^2\right]du^2. \tag{18}$$

Since $\epsilon=+1,0,-1$ and $\Lambda>0$, $\Lambda=0$, $\Lambda<0$, there are 9 invariant subclasses.

Another coordinate for the $RTN(\Lambda,\epsilon)$ class, suitable for physical interpretation, has been given by García Díaz and Plebański.[1] Their metric is expressed in terms of a function $f(\xi,u)$ which is an arbitrary function of u, analytic in spatial coordinate ξ,

$$ds^2=2v^2\,d\xi\,d\bar{\xi}+2v\bar{\mathrm{A}}\,d\xi\,du+2v\mathrm{A}\,d\bar{\xi}\,du+2\psi\,du\,dv+2(\mathrm{A}\bar{\mathrm{A}}+\psi\mathrm{B})du^2, \tag{19}$$

where

$$\mathrm{A}=\epsilon\xi-vf,\quad \mathrm{B}=-\epsilon+\frac{v}{2}(f_{,\xi}+\bar{f}_{,\bar{\xi}})+\frac{\Lambda}{6}v^2\psi,\quad \psi=1+\epsilon\xi\bar{\xi}.$$

It can be shown that the transformation relating (18) with (19) has the form

$$\xi=F(\zeta,u)=\int f(\xi(\zeta,u),u)du,$$

$$v=\frac{r}{1+\epsilon F\bar{F}}, \tag{20}$$

$$u\to -u,\quad P=(1+\epsilon F\bar{F})(F_{,\zeta}\bar{F}_{,\bar{\zeta}})^{-1/2}.$$

If f does not depend on ξ, we put $\xi=F(\zeta,u)=\zeta+\int f(u)du$.

The nonvanishing Weyl tensor components are proportional to $f_{,\xi\xi\xi}$ so that the solutions are conformally flat if f is quadratic in ξ. Thus, we can formulate

Lemma 3: The $RTN(\Lambda,\epsilon)$ solutions (19) with $f=f_c=c_0(u)+c_1(u)\xi+c_2(u)\xi^2$, where $c_i(u)$ are arbitrary complex functions of u, are isometric to Minkowski (if $\Lambda=0$), de Sitter ($\Lambda>0$) and anti-de Sitter spacetime ($\Lambda<0$).

Transformations preserving the form of (19) were studied in Ref. 2. However, more general transformations [Eq. (2.4) in Ref. 2 follows from Eqs. (21), (24) if we put $\Delta=0$ and $\alpha\bar{\alpha}+\epsilon\beta\bar{\beta}=1$] are given in the following.

Lemma 4: The coordinate transformations $(v,\xi,\bar{\xi},u)\to(w,\eta,\bar{\eta},t)$, which maintain invariant the form of the $RTN(\Lambda,\epsilon)$ metric (19), are

$$\xi=\frac{\alpha\eta+\beta}{\gamma\eta+\delta},$$

$$v=\frac{(\gamma\eta+\delta)(\overline{\gamma\eta}+\bar{\delta})}{\sqrt{(\alpha\delta-\beta\gamma)(\overline{\alpha\delta}-\overline{\beta\gamma})}}w, \tag{21}$$

$$u=\int\frac{\sqrt{(\alpha\delta-\beta\gamma)(\overline{\alpha\delta}-\overline{\beta\gamma})}}{\delta\bar{\delta}+\epsilon\beta\bar{\beta}}dt,$$

where $\alpha(t),\beta(t),\gamma(t),\delta(t)$ are arbitrary complex functions of t which satisfy the conditions

$$\bar{\gamma}\delta=-\epsilon\bar{\alpha}\beta,\quad \gamma\bar{\gamma}+\epsilon\alpha\bar{\alpha}=\epsilon(\delta\bar{\delta}+\epsilon\beta\bar{\beta}). \tag{22}$$

In the coordinates $(w,\eta,\bar{\eta},t)$, the new structural function $f'(\eta,t)$ is related to f as follows:

$$f' = \frac{\overline{\alpha\delta} - \overline{\beta\gamma}}{\delta\bar{\delta} + \epsilon\beta\bar{\beta}} \frac{(\gamma\eta+\delta)^2}{\sqrt{(\alpha\delta-\beta\gamma)(\overline{\alpha\delta}-\overline{\beta\gamma})}} f - \frac{1}{\alpha\delta-\beta\gamma}[(\dot{\beta}\delta-\beta\dot{\delta}) + (\dot{\alpha}\delta - \alpha\dot{\delta} + \dot{\beta}\gamma - \beta\dot{\gamma})\eta + (\dot{\alpha}\gamma - \alpha\dot{\gamma})\eta^2]. \tag{23}$$

From this relation and Lemma 3 we see that the conformally flat part f_c is indeed unimportant since it can be generated from $f=0$ by the coordinate transformation (21).

If $\epsilon \neq 0$, Eq. (22) implies that $|\alpha| = |\delta|$, $|\beta| = |\gamma|$ so that $\sqrt{(\alpha\delta-\beta\gamma)(\overline{\alpha\delta}-\overline{\beta\gamma})} = \delta\bar{\delta} + \epsilon\beta\bar{\beta}$. Equations (21)–(23) then simplify to

$$\xi = \frac{\alpha\eta+\beta}{\gamma\eta+\delta}, \quad v = \frac{(\gamma\eta+\delta)(\overline{\gamma\eta}+\bar{\delta})}{\alpha\bar{\alpha}+\epsilon\beta\bar{\beta}} w, \quad u=t, \tag{24}$$

where $\alpha(t)$ and $\beta(t)$ are arbitrary complex functions of t and $\gamma(t) = \bar{\beta}(t)e^{i\Gamma(t)}$, $\delta(t) = \bar{\alpha}(t)e^{i\Delta(t)}$, with $\Gamma(t)$, $\Delta(t)$ being arbitrary real functions of t satisfying $\Delta - \Gamma = (1+\epsilon)(\pi/2)$. The structural function is given by $f' = \{(\gamma\eta+\delta)^2 f - [(\dot{\beta}\delta-\beta\dot{\delta}) + (\dot{\alpha}\delta-\alpha\dot{\delta}+\dot{\beta}\gamma-\beta\dot{\gamma})\eta + (\dot{\alpha}\gamma-\alpha\dot{\gamma})\eta^2]\}e^{-i\Delta}/(\alpha\bar{\alpha}+\epsilon\beta\bar{\beta})$.

For $\epsilon=0$, Eqs. (22) imply $\gamma=0$ so that the transformations (21) yield

$$\xi = A(t)\eta + B(t), \quad v = \frac{w}{\sqrt{A(t)\bar{A}(t)}}, \quad u = \int \sqrt{A(t)\bar{A}(t)}\,dt, \tag{25}$$

where $A(t)$ and $B(t)$ are arbitrary complex functions of t. The relation (23) between the structural functions is now $f' = \sqrt{\bar{A}/A} f - (\dot{B}/A + \dot{A}/A\,\eta)$, so that, in contrast to the case $\epsilon \neq 0$, the term quadratic in η vanishes. Hence, Eq. (25) does not enable us to transform away the complete conformally flat part f_c. [We tried but without success to generalize (25) so that quadratic terms could be removed.]

Symmetries of the $RTN(\Lambda,\epsilon)$ solutions have been investigated in Ref. 13 (see also Ref. 4, Table 33.2) and, more systematically, in Ref. 2. The solutions which are not conformally flat allow the existence of at most two Killing vectors.

Note added in proof. We noticed that transformation (20) has been given in D. B. Singleton, ‘‘Homothetic motions and vacuum Robinson–Trautman solutions,’’ Gen. Relativ. Gravit. **22**, 1239 (1990).

ACKNOWLEDGMENTS

We thank Jerry Griffiths for reading the manuscript and useful suggestions. We also acknowledge the support of Grant No. GACR-202/99/0261 from the Czech Republic.

APPENDIX: PROOF OF PROPOSITION 2

For the conformally related metrics $g_{\alpha\beta}$ and $\hat{g}_{\alpha\beta}$, $\hat{g}_{\alpha\beta} = \Omega^{-2} g_{\alpha\beta}$, the trace-free Ricci tensors are related by[4]

$$S_{\alpha\beta} = R_{\alpha\beta} - \tfrac{1}{4} R g_{\alpha\beta},$$
$$\hat{S}_{\alpha\beta} = \hat{R}_{\alpha\beta} - \frac{1}{4}\hat{R}\hat{g}_{\alpha\beta} = S_{\alpha\beta} + \frac{2}{\Omega}\left(\Omega_{;\alpha\beta} - \frac{1}{4} g_{\alpha\beta}\Box\Omega\right), \tag{A1}$$

where $\Box\Omega = g^{\mu\nu}\Omega_{;\mu\nu}$, the covariant derivative is taken with respect to $g_{\alpha\beta}$ and the scalar curvature is

$$\hat{R}=\Omega^2 R+6\Omega\Box\Omega-12g^{\alpha\beta}\Omega_{;\alpha}\Omega_{;\beta}. \tag{A2}$$

Since $\hat{R}_{\alpha\beta}=\Lambda\hat{g}_{\alpha\beta}$, $\hat{R}=4\Lambda$, we have $\hat{S}_{\alpha\beta}=0$. In coordinates $(v,\xi,\bar{\xi},u)$,

$$g_{12}=1,\ g_{03}=-(\xi+\bar{\xi})^2,\ \ g_{33}=2(\xi+\bar{\xi})^2\left(v^2-\frac{1}{2}\frac{H}{\xi+\bar{\xi}}\right), \tag{A3}$$

$R_{33}=(\xi+\bar{\xi})H_{,\xi\bar{\xi}}$ and the other Ricci tensor components vanish. Therefore, $R=0$ and (A1) can be written as

$$\Omega_{,\alpha\beta}-\Gamma^{\gamma}_{\alpha\beta}\Omega_{,\gamma}-\tfrac{1}{4}g_{\alpha\beta}\Box\Omega+\tfrac{1}{2}\Omega R_{\alpha\beta}=0, \tag{A4}$$

which gives

$$\left(\frac{\Omega_{,v}}{\xi+\bar{\xi}}\right)_{,\xi}=0=\left(\frac{\Omega_{,v}}{\xi+\bar{\xi}}\right)_{,\bar{\xi}}, \tag{A5}$$

$$\Omega_{,vv}=0, \tag{A6}$$

$$\Omega_{,\xi\xi}=0=\Omega_{,\bar{\xi}\bar{\xi}}, \tag{A7}$$

$$\Omega_{,vu}+2v\Omega_{,v}-(\xi+\bar{\xi})(\Omega_{,\xi}+\Omega_{,\bar{\xi}})+\tfrac{1}{4}(\xi+\bar{\xi})^2\Box\Omega=0, \tag{A8}$$

$$\Omega_{,\xi\bar{\xi}}=\tfrac{1}{4}\Box\Omega, \tag{A9}$$

$$\Omega_{,\xi u}-\frac{\Omega_{,u}}{\xi+\bar{\xi}}-\frac{1}{2}\left(\frac{H}{\xi+\bar{\xi}}\right)_{,\xi}\Omega_{,v}=0=\Omega_{,\bar{\xi}u}-\frac{\Omega_{,u}}{\xi+\bar{\xi}}-\frac{1}{2}\left(\frac{H}{\xi+\bar{\xi}}\right)_{,\bar{\xi}}\Omega_{,v}, \tag{A10}$$

$$\begin{aligned}&\Omega_{,uu}-2v\Omega_{,u}-\left[2v\left(2v^2-\frac{H}{\xi+\bar{\xi}}\right)+\tfrac{1}{2}\left(\frac{H}{\xi+\bar{\xi}}\right)_{,u}\right]\Omega_{,v}\\&-\tfrac{1}{2}[(\xi+\bar{\xi})H]_{,\bar{\xi}}\Omega_{,\xi}-\tfrac{1}{2}[(\xi+\bar{\xi})H]_{,\xi}\Omega_{,\bar{\xi}}+2v^2(\xi+\bar{\xi})(\Omega_{,\xi}+\Omega_{,\bar{\xi}})\\&-\frac{1}{2}(\xi+\bar{\xi})^2\left(v^2-\frac{1}{2}\frac{H}{\xi+\bar{\xi}}\right)\Box\Omega+\frac{1}{2}(\xi+\bar{\xi})H_{,\xi\bar{\xi}}\Omega=0.\end{aligned} \tag{A11}$$

By using (A9), Eq. (A2) takes the form

$$\Omega\Omega_{,\xi\bar{\xi}}-\Omega_{,\xi}\Omega_{,\bar{\xi}}+\frac{\Omega_{,v}}{(\xi+\bar{\xi})^2}\left[\Omega_{,u}+\left(v^2-\frac{1}{2}\frac{H}{\xi+\bar{\xi}}\right)\Omega_{,v}\right]=\frac{\Lambda}{6}. \tag{A12}$$

Equations (A5) imply $\Omega_{,v}=A(u,v)(\xi+\bar{\xi})$. Equation (A6) gives $A_{,v}=0$ so that $\Omega=A(u)(\xi+\bar{\xi})v+B(\xi,\bar{\xi},u)$. By virtue of Eq. (A7) it must be of the form $\Omega=A(u)(\xi+\bar{\xi})v+C(u)\xi\bar{\xi}+D_1(u)\xi+D_2(u)\bar{\xi}+E(u)$. Equation (A8) combined with Eq. (A9) gives $dA/du=D_1+D_2$ so that $\Omega=A(u)(\xi+\bar{\xi})v+C(u)\xi\bar{\xi}+\tfrac{1}{2}dA/du(\xi+\bar{\xi})+D(u)(\xi-\bar{\xi})+E(u)$ with A,C,E being real functions of u, and $D(u)$ being pure imaginary. We distinguish the possibilities:

$$(1)\ \ A=0.$$

In this case Eq. (A10) implies $\Omega=C\xi\bar{\xi}+D(\xi-\bar{\xi})+E$, with C,E real constants, D a pure imaginary constant. Let (i) $C\neq 0$. Without loss of generality we can set $D=0$ by transformation ξ

$\rightarrow\xi'=\xi-D/C$. If $E\neq 0$, we can set $E=1$ by $\xi'=\xi/E$ which implies $\Omega=C\xi\bar{\xi}+1$. Equation (A12) gives $C=\Lambda/6$ so that $\Omega=1+(\Lambda/6)\xi\bar{\xi}$, and the metric $\hat{g}_{\alpha\beta}$ takes the canonical form (12) of the $KN(\Lambda)I$ metric. If $E=0$ we can make transformation $\xi'=1/(C\xi)$, after which $\hat{g}_{\alpha\beta}$ takes the canonical form (11) of the KN metric. Now (ii) $C=0$. If $D\neq 0$ we can set $E=0$ by $\xi'=\xi+E/(2D)$. The relation (A12) gives $\Omega=\sqrt{\Lambda/6}(\xi-\bar{\xi})$, with Λ necessarily being negative. It represents just another coordinate form of the $KN(\Lambda)I$ metric since by $\xi=(\xi'+i\sqrt{-6/\Lambda})/(i\xi'+\sqrt{-6/\Lambda})$ we get the $KN(\Lambda^-)I$ metric in the canonical form (12). If $D=0$ then $\Omega=E$ so that $\hat{g}_{\alpha\beta}$ is the KN metric by transformation $\xi'=|E|\xi$,

$$(2)\quad A\neq 0.$$

Without loss of generality we can assume $A=1$ by using transformation $w=A(u)v+\frac{1}{2}dA/du$, $t=\int A^{-1}(u)du$. Therefore, $\Omega=(\xi+\bar{\xi})v+C(u)\xi\bar{\xi}+D(u)(\xi-\bar{\xi})+E(u)$. In this case Eq. (A10) implies $H=2[(dC/du)\xi\bar{\xi}+\alpha(u)(\xi+\bar{\xi})+(dD/du)(\xi-\bar{\xi})+(dE/du)]$, with α being an arbitrary real function of u. Equation (A12) reduces to $\alpha(u)=C(u)E(u)+D(u)^2-\Lambda/6$, whereas Eq. (A11) is satisfied identically. The solution given by $\hat{g}_{\alpha\beta}$ is then conformally flat since $g_{\alpha\beta}$ for H of this form is conformally flat ($C_{\alpha\beta\gamma\delta}=0$). Therefore, the solution describes Minkowski, de Sitter or anti-de Sitter spacetime, according to the sign of Λ. It is interesting to notice that although $\hat{g}_{\alpha\beta}$ of this form describes a conformally flat vacuum solution, the KN solution to which it is conformal (given by $g_{\alpha\beta}$ with H of the same form) is conformally flat but *not* necessarily a vacuum solution. In general, it is a pure radiation solution, becoming a vacuum solution (Minkowski) only for $C=$const. In particular, if $\alpha=0$ and $D,E=$const, then the conformally flat KN pure radiation solution given by $H=2(dC/du)\,\xi\bar{\xi}$ is Wils' solution (3.10)[14] for $N=0$ [where $Q+\bar{Q}=dC/du$, $v\rightarrow v/(\xi+\bar{\xi})^2$]. The solution was used by Wils as an explicit counterexample of Theorem 32.17 in Ref. 4 according to which there are no other conformally flat pure radiation solutions besides the special PP wave of McLenaghan *et al.*[15]

We have thus shown that all vacuum solutions conformal to the KN class are (1) $KN(\Lambda)I$ solutions, (2) KN solutions themselves [$KN=KN(\Lambda=0)I$], (3) Minkowski, de Sitter and anti-de Sitter spacetimes (these are special cases of $KN(\Lambda)I$ given by $KN(\Lambda)I[H=H_c]$). Therefore, *all vacuum solutions conformal to the KN class belong to the $KN(\Lambda)I$ class*, which proves Proposition 2 in Sec. II.

[1] A. García Díaz and J. F. Plebański, ''All nontwisting *N*'s with cosmological constant,'' J. Math. Phys. **22**, 2655 (1981).
[2] H. I. Salazar, A. García Díaz, and J. F. Plebański, ''Symmetries of the nontwisting type-N solutions with cosmological constant,'' J. Math. Phys. **24**, 2191 (1983).
[3] I. Ozsváth, I. Robinson, and K. Rózga, ''Plane-fronted gravitational and electromagnetic waves in spaces with cosmological constant,'' J. Math. Phys. **26**, 1755 (1985).
[4] D. Kramer, H. Stephani, M. A. H. MacCallum, and H. Herlt, *Exact Solutions of the Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).
[5] W. Kundt and A. Thompson, ''Le tenseur de Weyl et une congruence associée de géodésiques isotropes sans distortion,'' C. R. Acad. Sci. (Paris), Ser. A **254**, 4257 (1962).
[6] W. Kundt, ''The plane-fronted gravitational waves,'' Z. Phys. **163**, 77 (1961).
[7] I. Robinson and A. Trautman, ''Spherical gravitational waves,'' Phys. Rev. Lett. **4**, 431 (1960).
[8] S. T. C. Siklos, ''Lobatchevski plane gravitational waves,'' in *Galaxies, Axisymmetric Systems and Relativity*, edited by M. A. H. MacCallum (Cambridge University Press, Cambridge, 1985).
[9] A. García Díaz, ''The generalizes Kundt solution with cosmological constant,'' Nuovo Cimento B **78**, 255 (1983).
[10] J. Podolský, ''Interpretation of the Siklos solutions as exact gravitational waves in the anti-de Sitter universe,'' Class. Quantum Grav. **15**, 719 (1998).
[11] J. Podolský and J. B Griffiths, ''Impulsive waves in de Sitter and anti-de Sitter space-times generated by null particles with an arbitrary multipole structure,'' Class. Quantum Grav. **15**, 453 (1998).
[12] J. Podolský, ''Non-expanding impulsive gravitational waves,'' Class. Quantum Grav. **15**, 3229 (1998).
[13] C. D. Collinson and D. C. French, ''Null tetrad approach to motions in empty space–time,'' J. Math. Phys. **8**, 701 (1967).
[14] P. Wils, ''Homogeneous and conformally Ricci flat pure radiation fields,'' Class. Quantum Grav. **6**, 1243 (1989).
[15] R. G. McLenaghan, N. Tariq, and B. J. O. Tupper, ''Conformally flat solutions of the Einstein–Maxwell equations for null electromagnetic fields,'' J. Math. Phys. **16**, 829 (1975).

Gravitational waves in vacuum spacetimes with cosmological constant. II. Deviation of geodesics and interpretation of nontwisting type *N* solutions

Jiří Bičák[a)] and Jiří Podolský[b)]
Department of Theoretical Physics, Faculty of Mathematics and Physics, Charles University, V Holešovičkách 2, 180 00 Prague 8, Czech Republic

(Received 31 July 1998; accepted for publication 25 May 1999)

In a suitably chosen essentially unique frame tied to a given observer in a general spacetime, the equation of geodesic deviation can be decomposed into a sum of terms describing specific effects: isotropic (background) motions associated with the cosmological constant, transverse motions corresponding to the effects of gravitational waves, longitudinal motions and Coulomb-type effects. Conditions under which the frame is parallelly transported along a geodesic are discussed. Suitable coordinates are introduced and an explicit coordinate form of the frame is determined for spacetimes admitting a nontwisting null congruence. Specific properties of all nontwisting type *N* vacuum solutions with cosmological constant Λ (nonexpanding Kundt class and expanding Robinson–Trautman class) are then analyzed. It is demonstrated that these spacetimes can be understood as exact transverse gravitational waves of two polarization modes ''+'' and ''×,'' shifted by $\pi/4$, which propagate ''on'' Minkowski, de Sitter or anti-de Sitter backgrounds. It is also shown that the solutions with $\Lambda>0$ may serve as exact demonstrations of the cosmic ''no-hair'' conjecture in radiative spacetimes with no symmetry. © *1999 American Institute of Physics.* [S0022-2488(99)00609-X]

I. INTRODUCTION AND SUMMARY

In the preceding paper[1] we classified nontwisting type *N* solutions of the vacuum Einstein's equations with a nonvanishing cosmological constant Λ and analyzed their geometrical properties. Here we wish to discuss their physical properties. We shall show that these solutions can be interpreted as gravitational waves propagating in spacetimes of constant curvature—in Minkowski, de Sitter or anti-de Sitter spaces. In our treatment we focus on the analysis of the equation of geodesic deviation.

We first discuss the equation of geodesic deviation in general spacetimes (Sec. II), briefly reviewing and extending[2–4] by using both a Newman–Penrose null tetrad and a physical frame of four independent vectors $\{\mathbf{e}_{(a)}\}$ tied to the geodesic with respect to which the relative motion is studied. In type *N* solutions only the Newman–Penrose scalar Ψ_4 is nonvanishing.

Starting from Sec. III we study nontwisting type *N* solutions with Λ. As shown in Ref. 1, they comprise the nonexpanding Kundt class $KN(\Lambda)$ and the expanding Robinson–Trautman class $RTN(\Lambda,\epsilon)$. By analyzing the geodesic deviation in these spacetimes we demonstrate that they can be interpreted as exact transverse gravitational waves with two polarization modes (shifted by $\pi/4$) propagating ''on'' Minkowski, de Sitter or anti-de Sitter space (depending on the values of Λ). In the Appendix we calculate the exact forms of wave amplitudes.

At the end of Sec. IV we discuss, for $\Lambda>0$, special timelike geodesics explicitly. We demonstrate that observers moving along these geodesics see waves decaying exponentially fast and the spacetimes to approach locally the de Sitter space—in agreement with the cosmic no-hair

[a)]Electronic mail: bicak@mbox.troja.mff.cuni.cz
[b)]Electronic mail: podolsky@mbox.troja.mff.cuni.cz

conjecture (see, e.g., Ref. 5, and references therein). As in our previous work,[6,7] this is an explicit demonstration of the conjecture under the presence of waves within exact theory.

II. THE RELATIVE MOTION OF FREE PARTICLES IN A GENERAL SPACETIME

It is natural to base the local characterization of radiative spacetimes on the equation of geodesic deviation,[2–4]

$$\frac{D^2Z^\mu}{d\tau^2}=-R^\mu_{\alpha\beta\gamma}u^\alpha Z^\beta u^\gamma, \tag{1}$$

where $\mathbf{u}=d\mathbf{x}/d\tau$, $\mathbf{u}\cdot\mathbf{u}=-1$ is the four-velocity of a free test particle (observer), τ is the proper time and $\mathbf{Z}(\tau)$ is the displacement vector. In order to obtain invariant results one sets up a frame $\{\mathbf{e}_{(a)}\}$ along the geodesic. The frame components $Z^{(a)}(\tau)$, $\mathbf{Z}=Z^{(a)}\mathbf{e}_{(a)}$, are invariant quantities. Choosing $\mathbf{e}_{(0)}=\mathbf{u}$ and perpendicular spacelike unit vectors $\{\mathbf{e}_{(1)},\mathbf{e}_{(2)},\mathbf{e}_{(3)}\}$ in the local hypersurface orthogonal to $\mathbf{u}$, we have $\mathbf{e}_{(a)}\cdot\mathbf{e}_{(b)}\equiv g_{\alpha\beta}e^\alpha_{(a)}e^\beta_{(b)}=\eta_{(a)(b)}=\mathrm{diag}(-1,1,1,1)$. The dual basis is $\mathbf{e}^{(0)}=-\mathbf{u}$ and $\mathbf{e}^{(i)}=\mathbf{e}_{(i)}$, $i=1,2,3$. By projecting (1) onto the frame we get

$$\ddot{Z}^{(i)}=-R^{(i)}_{(0)(j)(0)}Z^{(j)}, \tag{2}$$

where $Z^{(j)}=\mathbf{e}^{(j)}\cdot\mathbf{Z}=e^{(j)}_\mu Z^\mu$ determine directly the distance between close test particles,

$$\ddot{Z}^{(i)}\equiv\mathbf{e}^{(i)}\cdot\frac{D^2\mathbf{Z}}{d\tau^2}=e^{(i)}_\mu\frac{D^2Z^\mu}{d\tau^2}, \tag{3}$$

are physical relative accelerations, and $R_{(i)(0)(j)(0)}=e^\alpha_{(i)}u^\beta e^\gamma_{(j)}u^\delta R_{\alpha\beta\gamma\delta}$. Equation (1) also implies $d^2Z^{(0)}/d\tau^2=-u_\mu D^2Z^\mu/d\tau^2=R_{\mu\alpha\beta\gamma}u^\mu u^\alpha Z^\beta u^\gamma=0$ so that $Z^{(0)}=a_0\tau+b_0$, a_0,b_0 are constants. Setting $Z^{(0)}=0$, all test particles are "synchronized" by τ (they always stay in the same local hypersurface). From the definition of the Weyl tensor we get $R_{(i)(0)(j)(0)}=C_{(i)(0)(j)(0)}+\frac{1}{2}(\delta_{ij}R_{(0)(0)}-R_{(i)(j)})+\frac{1}{6}R\delta_{ij}$. Using Einstein's equations,

$$R_{(i)(0)(j)(0)}=C_{(i)(0)(j)(0)}-\frac{\Lambda}{3}\delta_{ij}-\frac{\kappa}{2}\left[T_{(i)(j)}-\delta_{ij}\left(T_{(0)(0)}+\frac{2}{3}T\right)\right], \tag{4}$$

$T=T^{(a)}_{(a)}$. Following Ref. 8 we introduce the null complex tetrad $\{\mathbf{e}_{\hat{a}}\}=\{\mathbf{e}_{\hat{1}},\mathbf{e}_{\hat{2}},\mathbf{e}_{\hat{3}},\mathbf{e}_{\hat{4}}\}=\{\mathbf{m},\bar{\mathbf{m}},\mathbf{l},\mathbf{k}\}$,

$$\mathbf{m}=\frac{1}{\sqrt{2}}(\mathbf{e}_{(1)}+i\mathbf{e}_{(2)}),\quad \bar{\mathbf{m}}=\frac{1}{\sqrt{2}}(\mathbf{e}_{(1)}-i\mathbf{e}_{(2)}),$$
$$\mathbf{l}=\frac{1}{\sqrt{2}}(\mathbf{u}-\mathbf{e}_{(3)}),\quad \mathbf{k}=\frac{1}{\sqrt{2}}(\mathbf{u}+\mathbf{e}_{(3)}). \tag{5}$$

Null tetrad components of the Weyl tensor are (see, e.g., Refs. 8 and 9)

$$\Psi_0=C_{\alpha\beta\gamma\delta}k^\alpha m^\beta k^\gamma m^\delta,\quad \Psi_1=C_{\alpha\beta\gamma\delta}k^\alpha l^\beta k^\gamma m^\delta,$$
$$\Psi_2=\tfrac{1}{2}C_{\alpha\beta\gamma\delta}k^\alpha l^\beta(k^\gamma l^\delta-m^\gamma\bar{m}^\delta), \tag{6}$$
$$\Psi_3=C_{\alpha\beta\gamma\delta}l^\alpha k^\beta l^\gamma\bar{m}^\delta,\quad \Psi_4=C_{\alpha\beta\gamma\delta}l^\alpha\bar{m}^\beta l^\gamma\bar{m}^\delta.$$

Regarding expressions (6) and inverting relations (5) we obtain

$$C_{(1)(0)(1)(0)}=\tfrac{1}{2}\mathcal{R}e\Psi_0+\tfrac{1}{2}\mathcal{R}e\Psi_4-\mathcal{R}e\Psi_2,\quad C_{(2)(0)(2)(0)}=-\tfrac{1}{2}\mathcal{R}e\Psi_0-\tfrac{1}{2}\mathcal{R}e\Psi_4-\mathcal{R}e\Psi_2,$$
$$C_{(1)(0)(2)(0)}=\tfrac{1}{2}\mathcal{I}m\Psi_0-\tfrac{1}{2}\mathcal{I}m\Psi_4,\quad C_{(3)(0)(3)(0)}=2\mathcal{R}e\Psi_2, \tag{7}$$
$$C_{(1)(0)(3)(0)}=-\mathcal{R}e\Psi_1+\mathcal{R}e\Psi_3,\quad C_{(2)(0)(3)(0)}=-\mathcal{I}m\Psi_1-\mathcal{I}m\Psi_3.$$

Substituting Eqs. (4) and (7) into Eq. (2) we arrive at

$$\ddot{Z}^{(1)}=\frac{\Lambda}{3}Z^{(1)}-\frac{\kappa}{2}\left(T_{(0)(0)}+\frac{2}{3}T\right)Z^{(1)}+\frac{\kappa}{2}T_{(1)(j)}Z^{(j)}+\mathcal{G}_1,$$
$$\ddot{Z}^{(2)}=\frac{\Lambda}{3}Z^{(2)}-\frac{\kappa}{2}\left(T_{(0)(0)}+\frac{2}{3}T\right)Z^{(2)}+\frac{\kappa}{2}T_{(2)(j)}Z^{(j)}+\mathcal{G}_2, \tag{8}$$
$$\ddot{Z}^{(3)}=\frac{\Lambda}{3}Z^{(3)}-\frac{\kappa}{2}\left(T_{(0)(0)}+\frac{2}{3}T\right)Z^{(3)}+\frac{\kappa}{2}T_{(3)(j)}Z^{(j)}+\mathcal{G}_3,$$

where

$$\mathcal{G}_1\equiv+\mathcal{C}Z^{(1)}-(\mathcal{L}_1-\mathcal{M}_1)Z^{(3)}-(\mathcal{A}_++\mathcal{B}_+)Z^{(1)}+(\mathcal{A}_\times-\mathcal{B}_\times)Z^{(2)},$$
$$\mathcal{G}_2\equiv+\mathcal{C}Z^{(2)}+(\mathcal{L}_2+\mathcal{M}_2)Z^{(3)}+(\mathcal{A}_++\mathcal{B}_+)Z^{(2)}+(\mathcal{A}_\times-\mathcal{B}_\times)Z^{(1)},$$
$$\mathcal{G}_3\equiv-2\mathcal{C}Z^{(3)}-(\mathcal{L}_1-\mathcal{M}_1)Z^{(1)}+(\mathcal{L}_2+\mathcal{M}_2)Z^{(2)},$$

and

$$\mathcal{C}=\mathcal{R}e\Psi_2,\quad \mathcal{L}_1=\mathcal{R}e\Psi_3,\quad \mathcal{L}_2=\mathcal{I}m\Psi_3,\quad \mathcal{M}_1=\mathcal{R}e\Psi_1,\quad \mathcal{M}_2=\mathcal{I}m\Psi_1,$$
$$\mathcal{A}_+=\tfrac{1}{2}\mathcal{R}e\Psi_4,\quad \mathcal{A}_\times=\tfrac{1}{2}\mathcal{I}m\Psi_4,\quad \mathcal{B}_+=\tfrac{1}{2}\mathcal{R}e\Psi_0,\quad \mathcal{B}_\times=\tfrac{1}{2}\mathcal{I}m\Psi_0. \tag{9}$$

Equations (8) are well suited for physical interpretation. The relative motions depend on 1. the cosmological constant Λ responsible for overall background isotropic motions; 2. the energy–momentum tensor $T_{(a)(b)}$ terms describing interaction with matter-content; 3. the terms depending on the local free gravitational field, and consisting of Coulomb, longitudinal and transverse (outgoing/ingoing) components with amplitudes given by Ψ_A's. In the following we put $T_{(a)(b)}=0$. Individual terms in Eq. (8) can be interpreted as follows.

Λ-term: Assuming $T_{(a)(b)}=0=\Psi_A$, Eq. (8) reduces to

$$\ddot{Z}^{(i)}=\frac{\Lambda}{3}Z^{(i)}. \tag{10}$$

Considering a sphere of test particles each having a position vector $\mathbf{Z}$, Eq. (10) implies that the acceleration of each particle is in the direction $\mathbf{Z}$ and has the same magnitude. Assume the frame $\{\mathbf{e}_{(i)}\}$ to be parallelly transported so that $\ddot{Z}^{(i)}=D^2(\mathbf{e}^{(i)}\cdot\mathbf{Z})/d\tau^2=d^2Z^{(i)}/d\tau^2$. Equations (10) have solutions

$$Z^{(i)}(\tau)=A_i\tau+B_i,\quad \text{for }\Lambda=0,$$
$$Z^{(i)}(\tau)=A_i\exp(\sqrt{\Lambda/3}\,\tau)+B_i\exp(-\sqrt{\Lambda/3}\,\tau),\quad \text{for }\Lambda>0, \tag{11}$$
$$Z^{(i)}(\tau)=A_i\cos(\sqrt{-\Lambda/3}\tau)+B_i\sin(\sqrt{-\Lambda/3}\tau),\quad \text{for }\Lambda<0,$$

where A_i, B_i are constants. As expected, conformally flat ($\Psi_A=0$) vacuum backgrounds (Minkowski, de Sitter or anti-de Sitter) are homogeneous and isotropic, so that the relative motion of test particles is isotropic.

Ψ_4-term: Assuming $\Lambda=0=T_{(a)(b)}$ and $\Psi_0=\Psi_1=\Psi_2=\Psi_3=0$, Eq. (8) reduces to

$$\ddot{Z}^{(1)}=-\mathcal{A}_+Z^{(1)}+\mathcal{A}_\times Z^{(2)},\quad \ddot{Z}^{(2)}=\mathcal{A}_+Z^{(2)}+\mathcal{A}_\times Z^{(1)},\quad \ddot{Z}^{(3)}=0, \tag{12}$$

which describe the influence of ''+'' and ''×'' polarization modes of a transverse gravitational wave with amplitudes $\mathcal{A}_+$ and $\mathcal{A}_\times$. If particles, initially at rest, lie in the $(\mathbf{e}_{(1)},\mathbf{e}_{(2)})$ plane, there is no motion in the longitudinal direction of $\mathbf{e}_{(3)}$. The ring of particles is deformed into an ellipse, the axes of different polarizations are shifted once with respect to the other by $\pi/4$ (such behavior is typical for linearized gravitational waves — cf., e.g., Ref. 10). Making a rotation in the transverse plane by an angle ϑ,

$$\mathbf{e}'_{(1)}=\cos\vartheta\,\mathbf{e}_{(1)}+\sin\vartheta\,\mathbf{e}_{(2)},\quad \mathbf{e}'_{(2)}=-\sin\vartheta\,\mathbf{e}_{(1)}+\cos\vartheta\,\mathbf{e}_{(2)} \tag{13}$$

— which corresponds to $\mathbf{m}'=\mathbf{e}^{-i\vartheta}\mathbf{m}$ — and using Eqs. (6) and (9), we find

$$\mathcal{A}'_+(\tau)=\cos 2\vartheta\mathcal{A}_+-\sin2\vartheta\mathcal{A}_\times,\quad \mathcal{A}'_\times(\tau)=\sin 2\vartheta\mathcal{A}_++\cos 2\vartheta\mathcal{A}_\times. \tag{14}$$

Taking $\vartheta=\vartheta_+(\tau)=-\frac{1}{2}\text{Arg}\,\Psi_4$, then $\mathcal{A}'_+=\frac{1}{2}|\Psi_4|$, $\mathcal{A}'_\times=0$—the wave is purely ''+'' polarized for an observer using $\mathbf{m}_+=e^{-i\vartheta_+}\mathbf{m}$; if $\vartheta=\vartheta_\times(\tau)=\vartheta_++\pi/4$, then $\mathcal{A}'_+=0$, $\mathcal{A}'_\times=\frac{1}{2}|\Psi_4|$—the wave is purely ''×'' polarized for an observer using $\mathbf{m}_\times=e^{-i\vartheta_\times}\mathbf{m}$. The amplitude $\mathcal{A}=\frac{1}{2}|\Psi_4|$ is invariant under the rotation. A general observer sees a superposition of the two polarization modes shifted by $\pi/4$.

One can similarly show[4,8] that Ψ_3 and Ψ_2 terms describe longitudinal modes and Coulomb-type effects; Ψ_1 and Ψ_0 terms are equivalent to Ψ_3 and Ψ_4 terms (if $\mathbf{k}\leftrightarrow\mathbf{l}$).

For given principal null vector $\mathbf{k}$ and observer's $\mathbf{u}$ we have chosen the frame vector $\mathbf{e}_{(3)}$ according to Eq. (5), which implies $k_{(1)}=0=k_{(2)}$, $k_{(3)}\neq0$, and makes the physical interpretation based on Eq. (8) simpler. This leads to essentially unique $\mathbf{k}$ and $\mathbf{l}$. More precisely, we easily show the following.

Proposition 1: Let $\mathbf{u}$ be the four-velocity ($\mathbf{u}\cdot\mathbf{u}=-1$) and $\mathbf{k}$ be the null vector. Then there exists a unit spacelike vector $\mathbf{e}_{(3)}$ which is the projection of the null direction given by $\mathbf{k}$ into the hypersurface orthogonal to $\mathbf{u}$. Such $\mathbf{e}_{(3)}$ is unique (up to reflections $\mathbf{e}_{(3)}\to-\mathbf{e}_{(3)}$) and is given by $\mathbf{e}_{(3)}=-\mathbf{u}+\sqrt{2}\mathbf{k}$, where $\mathbf{k}$ satisfies $\mathbf{k}\cdot\mathbf{u}=-1/\sqrt{2}$. Another null vector $\mathbf{l}$ in the plane $(\mathbf{u},\mathbf{e}_{(3)})$ such that $\mathbf{l}\cdot\mathbf{k}=-1$ is then given by $\mathbf{l}\equiv\sqrt{2}\mathbf{u}-\mathbf{k}$. The only remaining freedoms are rotations in the plane $(\mathbf{e}_{(1)},\mathbf{e}_{(2)})$ perpendicular to $\mathbf{e}_{(3)}$.

In what follows the orthonormal tetrad and the corresponding null frame (5) determined according to Proposition 1 will always be assumed.

Notice that Eqs. (8) represent possible motions seen by an observer with a given $\mathbf{u}$. By making the Lorentz boosts to other observers with $\mathbf{u}'$, Ψ_A change (see, e.g., Ref. 8). Thus, the ''strength of gravitational field'' is strongly observer dependent [cf. Refs. 2, 11, 12 and point 5 in the discussion following Eq. (34)].

III. THE CHOICE OF COORDINATES AND PARALLELLY PROPAGATED FRAMES

We shall now express our frames in coordinates suitable for spacetimes admitting a *nontwisting* null congruence and give the conditions for the frames to be parallelly transported. The field $\mathbf{k}$ is orthogonal to null hypersurfaces, say u=const., so that $k^\mu=g^{\mu\nu}u_{,\nu}$. It is convenient (cf. Refs. 9, 13) to choose as coordinates $u=x^3$, parameter $v=x^0$ along the null geodesics generated by k^μ, and two complex spacelike coordinates $\xi=x^1$ and $\bar{\xi}=x^2$ that label the geodesics on each surface u=const. The metric then takes the form

$$g_{\mu\nu}=\begin{pmatrix} 0 & 0 & 0 & g_{03} \\ 0 & g_{11} & g_{12} & g_{13} \\ 0 & g_{12} & g_{22} & g_{23} \\ g_{03} & g_{13} & g_{23} & g_{33} \end{pmatrix}, \tag{15}$$

where $g_{22}=\overline{g_{11}}$, $g_{23}=\overline{g_{13}}$ since $x^2=\overline{x^1}$; all other components are real, and

$$g_{12}>0,\ \ D=g_{12}^2-g_{11}g_{22}>0, \tag{16}$$

since the subspace $(\xi,\overline{\xi})$ is spacelike. The vector $\mathbf{k}$ is simply

$$k^\mu=(k^0,0,0,0), \tag{17}$$

and the four-velocity $\mathbf{u}$ of a particle moving along a geodesic $x^\mu(\tau)$, is given by $u^\mu=(\dot{v},\dot{\xi},\dot{\overline{\xi}},\dot{u})$, where the overdot is $d/d\tau$ and $\dot{u}\neq 0$ (otherwise the geodesic would not be timelike).

Proposition 2: In coordinates $(v,\xi,\overline{\xi},u)$ the interpretation null tetrad introduced in Proposition 1 has the form

$$\begin{aligned} m^\mu&=\Bigg(\frac{1}{g_{03}\dot{u}}[(g_{12}\dot{\xi}+g_{22}\dot{\overline{\xi}}+g_{23}\dot{u})g_+-(g_{11}\dot{\xi}+g_{12}\dot{\overline{\xi}}+g_{13}\dot{u})g_-\exp(-i\operatorname{Arg}g_{11})],\\ &\qquad g_-\exp(-i\operatorname{Arg}g_{11}),-g_+,0\Bigg),\\ \overline{m}^\mu&=\Bigg(\frac{1}{g_{03}\dot{u}}[(g_{11}\dot{\xi}+g_{12}\dot{\overline{\xi}}+g_{13}\dot{u})g_+-(g_{12}\dot{\xi}+g_{22}\dot{\overline{\xi}}+g_{23}\dot{u})g_-\exp(i\operatorname{Arg}g_{11})],\\ &\qquad -g_+,g_-\exp(i\operatorname{Arg}g_{11}),0\Bigg),\\ l^\mu&=\left(\sqrt{2}\dot{v}+\frac{1}{\sqrt{2}}\frac{1}{g_{03}\dot{u}},\sqrt{2}\dot{\xi},\sqrt{2}\dot{\overline{\xi}},\sqrt{2}\dot{u}\right),\\ k^\mu&=\left(-\frac{1}{\sqrt{2}}\frac{1}{g_{03}\dot{u}},0,0,0\right), \end{aligned} \tag{18}$$

where $g_\pm=\sqrt{(g_{12}\pm\sqrt{D})/(2D)}$. The tetrad is unique up to trivial reflections and rotations $m^\mu\to m^\mu e^{i\vartheta}$. The corresponding orthonormal frame obtained from Eqs. (18) using Eq. (5) is

$$\begin{aligned} e_{(0)}^\mu&=u^\mu=(\dot{v},\dot{\xi},\dot{\overline{\xi}},\dot{u}),\\ e_{(1)}^\mu&=\frac{1}{\sqrt{2}}\left(\frac{2}{g_{03}\dot{u}}\mathcal{R}e\{(g_{12}\dot{\xi}+g_{22}\dot{\overline{\xi}}+g_{23}\dot{u})G_-\},-\overline{G}_-,-G_-,0\right),\\ e_{(2)}^\mu&=\frac{1}{\sqrt{2}}\left(\frac{2}{g_{03}\dot{u}}\mathcal{I}m\{(g_{12}\dot{\xi}+g_{22}\dot{\overline{\xi}}+g_{23}\dot{u})G_+\},-i\overline{G}_+,iG_+,0\right),\\ e_{(3)}^\mu&=-\left(\dot{v}+\frac{1}{g_{03}\dot{u}},\dot{\xi},\dot{\overline{\xi}},\dot{u}\right), \end{aligned} \tag{19}$$

where $G_{\pm}=g_{+}\pm g_{-}\exp(i\mathrm{Arg}g_{11})$. The expressions (18) and (19) simplify considerably if $g_{11}=0$ since in this case $g_{-}=0$ and $g_{+}=1/\sqrt{g_{12}}=G_{+}=G_{-}$.

Proof: The last equation in (18) follows from Eq. (17) and $\mathbf{k}\cdot\mathbf{u}=-1/\sqrt{2}$, the equation for l^{μ} follows from $\mathbf{l}=\sqrt{2}\mathbf{u}-\mathbf{k}$. Vectors $\mathbf{m}$ and $\overline{\mathbf{m}}$ can then be determined from $\mathbf{e}_{\hat{a}}\cdot\mathbf{e}_{\hat{b}}=g_{\mu\nu}e^{\mu}_{\hat{a}}e^{\nu}_{\hat{b}}=g_{\hat{a}\hat{b}}$. The conditions $\mathbf{m}\cdot\mathbf{k}=g_{\hat{1}\hat{4}}=0=g_{\hat{2}\hat{4}}=\overline{\mathbf{m}}\cdot\mathbf{k}$ imply $m^3=0=\overline{m^3}$, $\mathbf{m}\cdot\mathbf{l}=g_{\hat{1}\hat{3}}=0$ implies $m^0=-(l_1m^1+l_2m^2)/l_0$. In given coordinates we have $\overline{m^1}=\overline{m^2}$ and $\overline{m^2}=\overline{m^1}$. Denoting $X=m^1$ and $Y=m^2$ we get $m^{\mu}=(-(l_1X+l_2Y)/l_0,X,Y,0)$ and $\overline{m}^{\mu}=(-(l_1\overline{Y}+l_2\overline{X})/l_0,\overline{Y},\overline{X},0)$. Functions *X, Y* can be determined as solutions of equations $\mathbf{m}\cdot\mathbf{m}=g_{\hat{1}\hat{1}}=0=g_{\hat{2}\hat{2}}=\overline{\mathbf{m}}\cdot\overline{\mathbf{m}}$ and $\mathbf{m}\cdot\overline{\mathbf{m}}=g_{\hat{1}\hat{2}}=1$,

$$g_{11}X^2+2g_{12}XY+g_{22}Y^2=0, \tag{20}$$

$$g_{11}X\overline{Y}+g_{12}(X\overline{X}+Y\overline{Y})+g_{22}\overline{X}Y=1. \tag{21}$$

(i) Assume $g_{11}\neq 0$. Then $X\neq 0$, and introducing a complex function C such that $Y=CX$, Eq. (20) implies $C=(-g_{12}\pm\sqrt{D})/g_{22}$, and Eq. (21) gives $X\overline{X}=|X|^2=(g_{12}\pm\sqrt{D})/(2D)$. Since $X=|X|e^{i\varphi}$, φ being a real function, we have $m^1=\sqrt{(g_{12}\pm\sqrt{D})/(2D)}\exp(i(\vartheta-\mathrm{Arg}g_{11}))$, $m^2=-\sqrt{(g_{12}\mp\sqrt{D})/(2D)}\exp(i\vartheta)$, where $\vartheta=\varphi+\mathrm{Arg}g_{11}$. The change from the upper to lower signs accompanied by $\vartheta\to-\vartheta+\pi+\mathrm{Arg}g_{11}$ results just in $\mathbf{m}\leftrightarrow\overline{\mathbf{m}}$, corresponding to a reflection $\mathbf{e}_{(2)}\leftrightarrow-\mathbf{e}_{(2)}$. By performing rotation $m^{\mu}\to m'^{\mu}=e^{-i\vartheta}m^{\mu}$ we can write the representatives of $\mathbf{m}$ and $\overline{\mathbf{m}}$ given by $\vartheta=0$ so that we arrive at Eq. (18).

(ii) If $g_{11}=0$, we simply find $m^1=0$ and $m^2=-1/\sqrt{g_{12}}$. Hence, the null tetrad has the form (18), and this implies the orthonormal frame (19).

In general, the frames $\{\mathbf{e}_a\}$ and $\{\mathbf{e}_{\hat{a}}\}$, related by Eq. (5), are not parallelly transported along the geodesic with tangent $\mathbf{u}=\mathbf{e}_{(0)}$. However, they are if $D\mathbf{k}/d\tau=0=D\mathbf{m}/d\tau$. Starting with an arbitrary $\mathbf{m}$, the second condition can always be satisfied by choosing $\mathbf{m}_{\parallel}=e^{i\vartheta_{\parallel}}\mathbf{m}$, where $\vartheta_{\parallel}=i\int_0^{\tau}\overline{\mathbf{m}}\cdot(D\mathbf{m}/d\tau)\,d\tau+\vartheta_0$, $\vartheta_0=\mathrm{const}$. We thus arrive at the following.

Proposition 3: Consider a geodesic $x^{\mu}(\tau)=(v,\xi,\overline{\xi},u)$ in spacetime with metric (15). Then the orthonormal frame $\{\mathbf{e}_a\}$ given by Eq. (19) and the null tetrad $\{\mathbf{e}_{\hat{a}}\}$ given by Eqs. (18) are parallelly transported along the geodesic if

$$g_{12,0}\dot{\xi}+g_{22,0}\dot{\overline{\xi}}+(g_{23,0}-g_{03,2})\dot{u}=0, \tag{22}$$

and

$$\begin{aligned}\dot{\vartheta}_{\parallel}(\tau)=\frac{i}{2D}[&(G_1\overline{G}_1-G_2\overline{G}_2)E\\&+G_1(2D\dot{m}^1+m^1(g_{12}\dot{g}_{12}-g_{22}\dot{g}_{11})+m^2(g_{12}\dot{g}_{22}-g_{22}\dot{g}_{12}))\\&+G_2(2D\dot{m}^2+m^1(g_{12}\dot{g}_{11}-g_{11}\dot{g}_{12})+m^2(g_{12}\dot{g}_{12}-g_{11}\dot{g}_{12}))],\end{aligned} \tag{23}$$

where $G_1=g_{12}g_{-}\exp(i\,\mathrm{Arg}\,g_{11})-g_{11}g_{+}$, $G_2=g_{22}g_{-}\exp(i\,\mathrm{Arg}\,g_{11})-g_{11}g_{+}$, $E=(g_{12,1}-g_{11,2})\dot{\xi}+(g_{22,1}-g_{12,2})\dot{\overline{\xi}}+(g_{23,1}-g_{13,2})\dot{u}=-\overline{E}$, $m^1=g_{-}\exp(-i\,\mathrm{Arg}\,g_{11})$, and $m^2=-g_{+}$. If, in addition, $g_{11}=0$, then $G_1=0$, $G_2=-\sqrt{g_{12}}$, and Eqs. (22), (23) reduce to

$$g_{12,0}\dot{\xi}+(g_{23,0}-g_{03,2})\dot{u}=0, \tag{24}$$

and

$$\dot{\vartheta}_{\parallel}(\tau)=-\frac{i}{2}\frac{1}{g_{12}}[g_{12,1}\dot{\xi}-g_{12,2}\dot{\overline{\xi}}+(g_{23,1}-g_{13,2})\dot{u}]. \tag{25}$$

Proof: Using $k_\mu k^\mu = 0$ and $k_\mu u^\mu = -1/\sqrt{2}$, it can be shown that **k** is parallelly transported if $\Gamma^1_{0\alpha} u^\alpha = 0$. Calculating the Christoffel symbols for the metric (15) we find (22). For proving Eq. (23) we use $m^3 = 0$ and $\overline{m}_0 = 0$; again the condition $\Gamma^1_{0\alpha} u^\alpha = 0$ and other Christoffel symbols for the metric (15).

IV. DEVIATION OF GEODESICS IN THE VACUUM NONTWISTING TYPE *N* SPACETIMES WITH COSMOLOGICAL CONSTANT

In this section we apply results given above to the nontwisting type N vacuum spacetimes with nonvanishing Λ. In the preceding paper[1] we showed that all such solutions belong either to the Kundt class of nonexpanding gravitational waves which we denoted by symbol $KN(\Lambda)$, or to the Robinson–Trautman class of expanding gravitational waves $RTN(\Lambda,\epsilon)$.

The class $KN(\Lambda)$ can be divided into six invariant canonical subclasses $KN(\Lambda)[\alpha,\beta]$, and the class $RTN(\Lambda,\epsilon)$ into nine invariant canonical subclasses, as analyzed in detail in Ref. 1. All $KN(\Lambda)$ metrics can be written in the form of Eq. (1, I), all $RTN(\Lambda,\epsilon)$ are described by Eq. (19, I). Both classes of metrics are of the form (15) in coordinates $x^\mu = (v,\xi,\overline{\xi},u)$. In the $KN(\Lambda)$ class we have

$$g_{12} = \frac{1}{p^2}, \quad g_{03} = -\frac{q^2}{p^2}, \quad g_{33} = F, \tag{26}$$

where $p = 1 + \Lambda/6\xi\overline{\xi}$, $q = (1-(\Lambda/6)\xi\overline{\xi})\alpha + \overline{\beta}\xi + \beta\overline{\xi}$, $F = \kappa(q^2/p^2)v^2 - ((q^2)_{,u}/p^2)v - (q/p)H$, $\kappa = (\Lambda/3)\alpha^2 + 2\beta\overline{\beta}$, $H = (f_{,\xi} + \overline{f}_{,\overline{\xi}}) - (\Lambda/3p)(\overline{\xi} f + \xi\overline{f})$; and in the $RTN(\Lambda,\epsilon)$ class,

$$g_{12} = v^2, \quad g_{13} = v\overline{\mathrm{A}}, \quad g_{23} = v\mathrm{A}, \quad g_{03} = \psi, \quad g_{33} = 2(\mathrm{A}\overline{\mathrm{A}} + \psi\mathrm{B}), \tag{27}$$

where $\mathrm{A} = \epsilon\xi - vf$, $\mathrm{B} = -\epsilon + (v/2)(f_{,\xi} + \overline{f}_{,\overline{\xi}}) + (\Lambda/6)v^2\psi$, $\psi = 1 + \epsilon\xi\overline{\xi}$, $\epsilon = -1,0,+1$, respectively.

Hence, for the $KN(\Lambda)$ solutions the orthonormal frame (19) is given by

$$\begin{aligned} e^\mu_{(0)} &= (\dot{v},\dot{\xi},\dot{\overline{\xi}},\dot{u}),\\ e^\mu_{(1)} &= -\frac{p}{\sqrt{2}}\left(\frac{2}{q^2}\,\frac{\mathcal{R}e\,\dot{\xi}}{\dot{u}},1,1,0\right),\\ e^\mu_{(2)} &= -\frac{p}{\sqrt{2}}\left(\frac{2}{q^2}\,\frac{\mathcal{I}m\,\dot{\xi}}{\dot{u}},i,-i,0\right),\\ e^\mu_{(3)} &= -\left(\dot{v} - \frac{p^2}{\dot{u}q^2},\dot{\xi},\dot{\overline{\xi}},\dot{u}\right), \end{aligned} \tag{28}$$

and for the $RTN(\Lambda,\epsilon)$ solutions we have

$$\begin{aligned} e^\mu_{(0)} &= (\dot{v},\dot{\xi},\dot{\overline{\xi}},\dot{u}),\\ e^\mu_{(1)} &= \frac{1}{\sqrt{2}}\,\frac{1}{v}\left(\frac{2v}{\psi\dot{u}}\mathcal{R}e\{v\dot{\xi} + \mathrm{A}\dot{u}\},-1,-1,0\right),\\ e^\mu_{(2)} &= \frac{1}{\sqrt{2}}\,\frac{1}{v}\left(\frac{2v}{\psi\dot{u}}\mathcal{I}m\{v\dot{\xi} + \mathrm{A}\dot{u}\},-i,i,0\right), \end{aligned} \tag{29}$$

$$e^{\mu}_{(3)}=-\left(\dot{v}+\frac{1}{\psi\dot{u}},\dot{\xi},\dot{\bar{\xi}},\dot{u}\right).$$

According to Proposition 3 these frames are parallelly transported along timelike geodesics $x^{\mu}(\tau)=(v,\xi,\bar{\xi},u)$ in the $KN(\Lambda)[\alpha,\beta]$ spacetimes if

$$\left(\frac{q}{p}\right)_{,\xi}=0=\left(\frac{q}{p}\right)_{,\bar{\xi}},\quad \dot{\vartheta}_{\parallel}(\tau)=i\left(\frac{p_{,\xi}}{p}\dot{\xi}-\frac{p_{,\bar{\xi}}}{p}\dot{\bar{\xi}}\right), \tag{30}$$

and in the case of $RTN(\Lambda,\epsilon)$ solutions if

$$\dot{\xi}=f\dot{u},\quad \dot{\bar{\xi}}=\bar{f}\dot{u},\quad \dot{\vartheta}_{\parallel}(\tau)=\frac{i}{2}(f_{,\xi}-\bar{f}_{,\bar{\xi}})\dot{u}. \tag{31}$$

The equation of geodesic deviation is now given by Eq. (8) with $T_{(a)(b)}=0$. The amplitudes (9) for both classes of spacetimes are calculated in the Appendix. We find that the invariant form of the equation of geodesic deviation with respect to the interpretation frame along any timelike geodesic in the $KN(\Lambda)$ and $RTN(\Lambda,\epsilon)$ spacetimes takes the form

$$\begin{aligned}\ddot{Z}^{(1)}&=\frac{\Lambda}{3}Z^{(1)}-\mathcal{A}_{+}Z^{(1)}+\mathcal{A}_{\times}Z^{(2)},\\ \ddot{Z}^{(2)}&=\frac{\Lambda}{3}Z^{(2)}+\mathcal{A}_{+}Z^{(2)}+\mathcal{A}_{\times}Z^{(1)},\\ \ddot{Z}^{(3)}&=\frac{\Lambda}{3}Z^{(3)},\end{aligned} \tag{32}$$

where the amplitudes of the transverse gravitational wave are given by

$$\mathcal{A}_{+}(\tau)=\tfrac{1}{2}pq\dot{u}^{2}\mathcal{R}e\{f_{,\xi\xi\xi}\},\quad \mathcal{A}_{\times}(\tau)=\tfrac{1}{2}pq\dot{u}^{2}\mathcal{I}m\{f_{,\xi\xi\xi}\}, \tag{33}$$

for the $KN(\Lambda)$ spacetimes, and by

$$\mathcal{A}_{+}(\tau)=-\frac{1}{2}\frac{\psi}{v}\dot{u}^{2}\mathcal{R}e\{f_{,\xi\xi\xi}\},\quad \mathcal{A}_{\times}(\tau)=-\frac{1}{2}\frac{\psi}{v}\dot{u}^{2}\mathcal{I}m\{f_{,\xi\xi\xi}\}, \tag{34}$$

in the $RTN(\Lambda,\epsilon)$ spacetimes [see Eqs. (A5) and (A10) in Appendix]. Equations (32)–(34) give relative accelerations of the free test particles in terms of their actual positions. They enable us to draw a number of simple conclusions.

(1) All particles move isotropically, one with respect to the other according to Eqs. (11) if no gravitational wave is present, i.e., if $f_{,\xi\xi\xi}=0$. In this case both the $KN(\Lambda)$ and $RTN(\Lambda,\epsilon)$ spacetimes are vacuum conformally flat [cf. (A3), (A8)], and therefore Minkowski ($\Lambda=0$), de Sitter ($\Lambda>0$) and anti-de Sitter ($\Lambda<0$) (see Lemma 1 and 3 in Ref. 1). Such spaces are maximally symmetric, homogeneous, isotropic, and they represent a natural background for other ''nontrivial'' $KN(\Lambda)$ and $RTN(\Lambda,\epsilon)$ type N solutions.

(2) If amplitudes $\mathcal{A}_{+}$ and $\mathcal{A}_{\times}$ do not vanish ($f_{,\xi\xi\xi}\neq 0$), the particles are influenced by the wave [see Eq. (12) and subsequent discussions] in a similar way as they are affected by a standard gravitational wave on Minkowski background (cf. Ref. 10). However, if $\Lambda\neq 0$, the influence of the wave adds with the (anti-) de Sitter isotropic expansion (contraction). This makes plausible our interpretation of the $KN(\Lambda)$ and $RTN(\Lambda,\epsilon)$ metrics as *exact gravitational waves propagating on the constant curvature backgrounds*.

(3) The wave propagates in the spacelike direction of $\mathbf{e}_{(3)}$ and has a *transverse character*, since only motions in the perpendicular directions of $\mathbf{e}_{(1)}$ and $\mathbf{e}_{(2)}$ are affected. The propagation direction given by $\mathbf{e}_{(3)}$ coincides with the projection of the Debever-Penrose vector $\mathbf{k}$ on the hypersurface orthogonal to the observer's velocity $\mathbf{u}$ (cf. Proposition 1).

(4) There are *two polarization modes* of the wave — ''+'' and ''×,'' $\mathcal{A}_+$ and $\mathcal{A}_\times$ being the amplitudes. Under rotation (13) in the transverse plane they transform according to Eq. (14) so that the helicity of the wave is 2, as with linearized waves on a Minkowski background. For the special choice of the frame given by $\vartheta(\tau)=\vartheta_+=-\frac{1}{2}pq\dot{u}^2\,\mathrm{Arg}\{f_{,\xi\xi\xi}\}$ for the $KN(\Lambda)$, and by $\vartheta_+=\frac{1}{4}(\psi/v)\dot{u}^2\,\mathrm{Arg}\{f_{,\xi\xi\xi}\}$ for the $RTN(\Lambda,\epsilon)$ spacetimes, the observer views pure ''+'' polarization, and for $\vartheta_\times=\vartheta_+ + \pi/4$ — pure ''×'' polarization.

(5) The waves have amplitude $\mathcal{A}=\frac{1}{2}pq\dot{u}^2|f_{,\xi\xi\xi}|$ for the $KN(\Lambda)$ class and $\mathcal{A}=\frac{1}{2}(\psi/v)\dot{u}^2|f_{,\xi\xi\xi}|$ for $RTN(\Lambda,\epsilon)$; this is invariant under rotations (13). However, the amplitude changes under Lorentz transformations to another observer $\mathbf{u}'$ with a spatial velocity $\vec{v}=(v_1,v_2,v_3)$ with respect to the original observer. For type N solutions we get $\mathcal{A}'=(1-v_3)^2/(1-v_1^2-v_2^2-v_3^2)\,\mathcal{A}$. By increasing speed in the wave-propagation direction $\mathbf{e}_{(3)}$ $(v_1=v_2=0,\ v_3>0)$, he experiences a weakening of the wave amplitude by factor $(1-v_3)/(1+v_3)$ $(\mathcal{A}'\to 0$ as $v_3\to 1)$, and by moving in the opposite direction, an increase of the amplitude $(\mathcal{A}'\to\infty$ as $v_3\to -1)$. By increasing speed in the transverse directions $\mathbf{e}_{(1)}$, $\mathbf{e}_{(2)}$, $(v_1^2+v_2^2=0,\ v_3=0)$, she experiences an increase by the factor $1/(1-v_1^2-v_2^2)$.

In general, all $KN(\Lambda)$ spacetimes contain singularities except for the homogeneous pp-waves[8] given by $p=1=q$ and $f_{,\xi\xi\xi}=6c_3(u)$, where $c_3(u)$ is a finite function of u. All other $KN(\Lambda)$ spacetimes are singular at $|\xi|=\infty$ where the amplitudes $\mathcal{A}_+$, $\mathcal{A}_\times$ diverge. Additional singularities in the amplitudes may occur if the coefficients $c_n(u)$, $n\geqslant 3$, of the analytic expansion of function $f(\xi,u)$ are badly behaved at some u.

$RTN(\Lambda,\epsilon)$ spacetimes also contain singularities. The character of the singularities depends on parameter ϵ and on the form of the function $f(\xi,u)$. As follows from Eq. (34), there is always a singularity at $v=0$. Another singularity is given by $\psi=\infty$ which occurs only for $\epsilon\neq 0$ at $|\xi|=\infty$. There may be singularities for special forms of f, namely, if $f_{,\xi\xi\xi}=\infty$. This occurs at $|\xi|=\infty$ if f contains the terms $c_n\xi^n$, $n\geqslant 4$. Another type of singularity may appear if some of the coefficients $c_n(u)$ diverge for some values of u. Singularities might be considered as ''sources'' of waves; however, it is far from certain whether nonsingular sources ''covering'' the regions in which singularities occur can be constructed. The singularities of the $RTN(\Lambda,\epsilon)$ spacetimes can invariantly be characterized by the nonvanishing invariant constructed recently[14] from the second derivatives of the Riemann tensor.

Finally, we shall discuss a special class of geodesics explicitly. Since for $f=f_c=c_0(u)+c_1(u)\xi+c_2(u)\xi^2$ the metrics represent Minkowski, de Sitter or anti-de Sitter space there always exists a transformation of coordinates which brings $g_{\mu\nu}[f=f_c]$ to $g_{\mu\nu}[f=0]$ (see Lemmas 2 and 4 in Ref. 1). It is thus sufficient to consider only the nontrivial part $f_w\equiv f-f_c$ of function $f(\xi,u)$. Moreover, one can always rearrange its analytic expansion so that $f=\Sigma_{n=0}^{\infty}c_n(u)\xi^n=\Sigma_{n=0}^{\infty}\tilde{c}_n(u)(\xi-\xi_0)^n=\Sigma_{n=3}^{\infty}\tilde{c}_n(u)(\xi-\xi_0)^n+\tilde{f}_c$, ξ_0 being an arbitrary complex constant. Therefore, it is natural to consider structural functions of the form

$$f_w=c_3(u)(\xi-\xi_0)^3+c_4(u)(\xi-\xi_0)^4+\cdots. \tag{35}$$

Consider a *special class of geodesics* characterized by $\xi=\xi_0=\text{const}$. For the $RTN(\Lambda,\epsilon)$ solutions these are geometrically privileged since the interpretation frame (29) is parallelly propagated along them [Eq. (31) is satisfied]. One can also find special geodesics for some subclasses of $KN(\Lambda)$: $\xi=\xi_0$ for the PP subclass, $\xi_0=\pm\sqrt{6/\Lambda}$ for $KN(\Lambda)I$, and $\xi_0=0$ for $KN(\Lambda^-)II$. The geodesics $\xi=\xi_0$ have the same forms as geodesics in the ''background'' since Christoffel symbols for $f=f_w$ and $\xi=\xi_0$ coincide with those for $f=0$. However, the test particles feel the tidal forces proportional to $\mathcal{A}_+$ and $\mathcal{A}_\times$ given by Eqs. (33), (34). The amplitudes do not vanish since $f_{w,\xi\xi\xi}=6c_3(u)$ is nonvanishing.

The timelike geodesics $\xi=\xi_0=\text{const}$ in the $RTN(\Lambda>0,\epsilon)[f_w]$ spacetimes are given by

$$v=\frac{\alpha}{1+\epsilon\xi_0\bar{\xi}_0}\left(C_1\cosh\frac{\tau}{\alpha}+C_2\sinh\frac{\tau}{\alpha}\right),\quad \dot{u}=-\left(C_1\sinh\frac{\tau}{\alpha}+C_2\cosh\frac{\tau}{\alpha}+C_3\right)^{-1},\tag{36}$$

where $\alpha=\sqrt{3/\Lambda}$, C_1, C_2, C_3 are real constants satisfying $C_1^2-C_2^2+C_3^2=2\epsilon$. The integration of Eq. (36) can be performed explicitly but we do not give it here since only $\dot{u}$ enters the amplitudes. The wave amplitudes (34) are $\mathcal{A}_+=\mathcal{R}e\,\mathcal{A}$ and $\mathcal{A}_\times=\mathcal{I}m\,\mathcal{A}$ where

$$\mathcal{A}(\tau)=-\frac{3}{\alpha}(1+\epsilon\xi_0\bar{\xi}_0)^2\left(C_1\cosh\frac{\tau}{\alpha}+C_2\sinh\frac{\tau}{\alpha}\right)^{-1}\left(C_1\sinh\frac{\tau}{\alpha}+C_2\cosh\frac{\tau}{\alpha}+C_3\right)^{-2}c_3(u(\tau)).\tag{37}$$

As proper time τ along geodesics increases, $\tau\rightarrow\infty$, particles recede from $v=0$ and amplitudes decay as $\mathcal{A}\sim\exp(-3\sqrt{\Lambda/3}\,\tau)$, i.e., *waves are damped exponentially*. The spacetime locally approaches the de Sitter universe. This is an explicit demonstration of the *cosmic no-hair conjecture* (see, e.g., Ref. 5) under the presence of waves within exact model spacetimes. (For a cosmic no-hair conjecture in the Robinson–Trautman spacetimes of Petrov type *II*, see Refs. 6, 7.)

Similarly, for the $KN(\Lambda>0)I[f_w]$ subclass [representing the only spacetimes of the $KN(\Lambda)$ type admitting $\Lambda>0$] the geodesics $\xi=\xi_0=\pm\sqrt{6/\Lambda}$ are given by

$$v=C_1\exp\left(\frac{\tau}{\alpha}\right),\quad u=-\frac{1}{2C_1}\exp\left(-\frac{\tau}{\alpha}\right)+C_2,\tag{38}$$

and

$$v=C_1\sinh\left(\frac{\tau}{\alpha}+2\tau_0\right),\quad u=\frac{1}{2C_1}\tanh\left(\frac{\tau}{2\alpha}+\tau_0\right)+C_2,\tag{39}$$

with C_1, C_2, τ_0 constants. For observers moving along these geodesics,

$$\mathcal{A}(\tau)=\pm12\sqrt{\frac{6}{\Lambda}}\dot{u}^2(\tau)c_3(u(\tau)).\tag{40}$$

After substitution of the explicit dependence of $u(\tau)$, we see that as $\tau\rightarrow+\infty$ the amplitudes behave like $\mathcal{A}\sim\exp(-2\sqrt{\Lambda/3}\,\tau)$. Again, gravitational waves are damped exponentially and the cosmic no-hair conjecture is confirmed.

ACKNOWLEDGMENTS

We thank Jerry Griffiths for reading the manuscript and for very helpful comments. We also acknowledge the support of Grant No. GACR-202/99/0261 from the Czech Republic.

APPENDIX: GRAVITATIONAL WAVE AMPLITUDES

We calculate amplitudes $\mathcal{A}_+=\frac{1}{2}\mathcal{R}e\Psi_4$ and $\mathcal{A}_\times=\frac{1}{2}\mathcal{I}m\Psi_4$ by using differential forms. Let $\{\mathbf{e}_{\hat{a}}\}=\{\mathbf{m},\bar{\mathbf{m}},\mathbf{l},\mathbf{k}\}$ be a null tetrad, $\mathbf{m}=\mathbf{e}_{\hat{1}}=m^\mu\partial_\mu$, $\bar{\mathbf{m}}=\mathbf{e}_{\hat{2}}=\bar{m}^\mu\partial_\mu$, $\mathbf{l}=\mathbf{e}_{\hat{3}}=l^\mu\partial_\mu$, $\mathbf{k}=\mathbf{e}_{\hat{4}}=k^\mu\partial_\mu$. The dual basis $\{\boldsymbol{\omega}^{\hat{a}}\}$ is given by one-forms $\boldsymbol{\omega}^{\hat{1}}=\bar{m}_\mu\,dx^\mu$, $\boldsymbol{\omega}^{\hat{2}}=m_\mu\,dx^\mu$, $\boldsymbol{\omega}^{\hat{3}}=-k_\mu\,dx^\mu$, $\boldsymbol{\omega}^{\hat{4}}=-l_\mu\,dx^\mu$; the metric is $ds^2=g_{\hat{a}\hat{b}}\boldsymbol{\omega}^{\hat{a}}\boldsymbol{\omega}^{\hat{b}}=2\boldsymbol{\omega}^{\hat{1}}\boldsymbol{\omega}^{\hat{2}}-2\boldsymbol{\omega}^{\hat{3}}\boldsymbol{\omega}^{\hat{4}}$ with $g_{\hat{1}\hat{2}}=\mathbf{e}_{\hat{1}}\cdot\mathbf{e}_{\hat{2}}=1$ and $g_{\hat{3}\hat{4}}=\mathbf{e}_{\hat{3}}\cdot\mathbf{e}_{\hat{4}}=-1$. The natural choice of the null basis for the metric $KN(\Lambda)$ is

$$\boldsymbol{\omega}^{\hat{1}}=\frac{d\bar{\xi}}{p},\quad \boldsymbol{\omega}^{\hat{2}}=\frac{d\xi}{p},\quad \boldsymbol{\omega}^{\hat{3}}=\frac{q^2}{p^2}du,\quad \boldsymbol{\omega}^{\hat{4}}=dv-\frac{1}{2}\frac{p^2}{q^2}F\,du;\tag{A1}$$

in coordinates $x^\mu=(v,\xi,\bar{\xi},u)$ we have

$$
\begin{aligned}
m^\mu&=(0,0,p,0),\quad \bar{m}^\mu=(0,p,0,0),\\
k^\mu&=(1,0,0,0),\quad l^\mu=\left(\frac{1}{2}\,\frac{p^4}{q^4}F,0,0,\frac{p^2}{q^2}\right).
\end{aligned}
\tag{A2}
$$

The nonvanishing components of the Weyl tensor in this null tetrad are

$$
C_{\hat{3}\hat{2}\hat{3}\hat{2}}\equiv\Psi_4=\frac{1}{2}\,\frac{p^5}{q^3}f_{,\xi\xi\xi}=\overline{C_{\hat{3}\hat{1}\hat{3}\hat{1}}}.
\tag{A3}
$$

The interpretation null tetrad for the $KN(\Lambda)$ spacetimes, given by Eq. (18), reads as

$$
\begin{aligned}
m^\mu&=\left(-\frac{p}{q^2}\,\frac{\dot{\xi}}{\dot{u}},0,-p,0\right),\quad \bar{m}^\mu=\left(-\frac{p}{q^2}\,\frac{\dot{\bar{\xi}}}{\dot{u}},-p,0,0\right),\\
k^\mu&=\left(\frac{1}{\sqrt{2}\dot{u}}\,\frac{p^2}{q^2},0,0,0\right),\quad l^\mu=\left(\sqrt{2}\dot{v}-\frac{1}{\sqrt{2}\dot{u}}\,\frac{p^2}{q^2},\sqrt{2}\dot{\xi},\sqrt{2}\dot{\bar{\xi}},\sqrt{2}\dot{u}\right).
\end{aligned}
\tag{A4}
$$

The relation between tetrads (A2) and (A4) is given by the Lorentz transformation, $\mathbf{k}_{\mathrm{natur}}=A\mathbf{k}_{\mathrm{interp}}$, $\mathbf{l}_{\mathrm{natur}}=(\mathbf{l}_{\mathrm{interp}}+Be^{i\vartheta}\bar{\mathbf{m}}_{\mathrm{interp}}+\bar{B}e^{-i\vartheta}\mathbf{m}_{\mathrm{interp}}+B\bar{B}\mathbf{k}_{\mathrm{interp}})/A$, $\mathbf{m}_{\mathrm{natur}}=e^{-i\vartheta}\mathbf{m}_{\mathrm{interp}}+B\mathbf{k}_{\mathrm{interp}}$, where $A=\sqrt{2}\dot{u}\,q^2/p^2$, $B=-\sqrt{2}\,\dot{\xi}/p$, $\vartheta=\pi$. The coefficients Ψ_A transform (see Ref. 8)

$$
\begin{aligned}
\Psi_4^{\mathrm{interp}}&=A^2\Psi_4^{\mathrm{natur}}=pq\dot{u}^2f_{,\xi\xi\xi},\\
\Psi_3^{\mathrm{interp}}&=\Psi_2^{\mathrm{interp}}=\Psi_1^{\mathrm{interp}}=\Psi_0^{\mathrm{interp}}=0.
\end{aligned}
\tag{A5}
$$

Similarly, the natural choice of null basis for the $RTN(\Lambda,\epsilon)$ metric is

$$
\boldsymbol{\omega}^{\hat{1}}=v\ d\bar{\xi}+\ \bar{\mathrm{A}}\ du,\quad \boldsymbol{\omega}^{\hat{2}}=v\,d\xi+\mathrm{A}\ du,\quad \boldsymbol{\omega}^{\hat{3}}=\psi\ du,\quad \boldsymbol{\omega}^{\hat{4}}=-dv-\mathrm{B}\ du,
\tag{A6}
$$

so that

$$
\begin{aligned}
m^\mu&=\left(0,0,\frac{1}{v},0\right),\quad \bar{m}^\mu=\left(0,\frac{1}{v},0,0\right),\\
k^\mu&=(-1,0,0,0),\quad l^\mu=\left(-\frac{\mathrm{B}}{\psi},-\frac{\mathrm{A}}{v\psi},-\frac{\bar{\mathrm{A}}}{v\psi},\frac{1}{\psi}\right).
\end{aligned}
\tag{A7}
$$

For this tetrad we obtain nonvanishing components,

$$
C_{\hat{3}\hat{2}\hat{3}\hat{2}}\equiv\Psi_4=-\frac{1}{2v\psi}f_{,\xi\xi\xi}=\overline{C_{\hat{3}\hat{1}\hat{3}\hat{1}}}.
\tag{A8}
$$

The interpretation null tetrad (18) reads as

$$
\begin{aligned}
m^\mu&=\left(\frac{1}{\psi\dot{u}}(v\dot{\xi}+\mathrm{A}\dot{u}),0,-\frac{1}{v},0\right),\quad \bar{m}^\mu=\left(\frac{1}{\psi\dot{u}}(v\dot{\bar{\xi}}+\bar{\mathrm{A}}\dot{u}),-\frac{1}{v},0,0\right),\\
k^\mu&=\left(-\frac{1}{\sqrt{2}}\,\frac{1}{\psi\dot{u}},0,0,0\right),\ l^\mu=\left(\sqrt{2}\dot{v}+\frac{1}{\sqrt{2}}\,\frac{1}{\psi\dot{u}},\sqrt{2}\dot{\xi},\sqrt{2}\dot{\bar{\xi}},\sqrt{2}\dot{u}\right).
\end{aligned}
\tag{A9}
$$

The relation between the tetrads (A7) and (A9) is again given by the Lorentz transformation with $A=\sqrt{2}\,\dot{u}\psi$, $B=-\sqrt{2}\,(v\dot{\xi}+\mathrm{A}\dot{u})$, $\vartheta=\pi$. We thus get

$$\Psi_4^{\text{interp}}=A^2\Psi_4^{\text{natur}}=-\frac{\psi}{v}\dot{u}^2 f_{,\xi\xi\xi},$$
$$\Psi_3^{\text{interp}}=\Psi_2^{\text{interp}}=\Psi_1^{\text{interp}}=\Psi_0^{\text{interp}}=0. \tag{A10}$$

[1] J. Bičák and J. Podolský, ''Gravitational waves in vacuum spacetimes with cosmological constant. I. Classification and geometrical properties of nontwisting type *N* solutions,'' J. Math. Phys. **40**, 4495 (1999).
[2] F. A. E. Pirani, ''On the physical significance of the Riemann tensor,'' Acta Phys. Pol. **15**, 389 (1956).
[3] F. A. E. Pirani, ''Invariant formulation of gravitational radiation theory,'' Phys. Rev. **150**, 1089 (1957).
[4] P. Szekeres, ''The gravitational compass,'' J. Math Phys. **6**, 1387 (1965).
[5] K. Maeda, ''Inflation and cosmic no hair conjecture,'' in *Proceedings of the 5th M. Grossman Meeting*, edited by D. G. Blair, M. J. Buckingham, and R. Ruffini (World Scientific, Singapore, 1989).
[6] J. Bičák and J. Podolský, ''Cosmic no-hair conjecture and black-hole formation: an exact model with gravitational radiation.'' Phys. Rev. D **52**, 887 (1995).
[7] J. Bičák and J. Podolský, ''Global structure of Robinson–Trautman radiative space–times with cosmological constant,'' Phys. Rev. D **55**, 1985 (1997).
[8] D. Kramer, H. Stephani, M. A. H. MacCallum, and H. Herlt, *Exact Solutions of the Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).
[9] E. Newman and R. Penrose, ''An approach to gravitational radiation by a method of spin coefficients,'' J. Math. Phys. **3**, 566 (1962).
[10] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[11] R. Penrose and W. Rindler, *Spinors and Space–time* (Cambridge University Press, Cambridge, 1984).
[12] B. Mashhoon, ''On the strength of a gravitational field,'' Phys. Lett. A **163**, 7 (1992).
[13] I. Robinson and A. Trautmann, ''Spherical gravitational waves,'' Phys. Rev. Lett. **4**, 431 (1960).
[14] J. Bičák and V. Pravda, ''Curvature invariants in type *N* spacetime,'' Class. Quantum Grav. **15**, 1539 (1998).

A construction of Killing spinors on S^n

H. Lü
Laboratoire de Physique Théorique de l'École Normale Supérieure, 24 Rue Lhomond-75231 Paris Cedex 05, France

C. N. Pope
Center for Theoretical Physics, Texas A&M University, College Station, Texas 77843

J. Rahmfeld[a)]
Department of Physics, Stanford University, Stanford, California 94305-4060

(Received 23 July 1998; accepted for publication 25 May 1999)

We derive simple general expressions for the explicit Killing spinors on the n-sphere, for arbitrary n. Using these results we also construct the Killing spinors on various AdS×Sphere supergravity backgrounds, including $AdS_5 \times S^5$, $AdS_4 \times S^7$, and $AdS_7 \times S^4$. In addition, we extend previous results to obtain the Killing spinors on the hyperbolic spaces H^n. *© 1999 American Institute of Physics.* [S0022-2488(99)00309-6]

I. INTRODUCTION

Finding the explicit form of Killing spinors on curved spaces can be an involved task. Often, one merely uses integrability conditions to establish their existence and to determine their multiplicities. In this way it is easy to show that spheres and anti-de Sitter spacetimes preserve all supersymmetries, i.e., they admit the maximum number of Killing spinors. However, one does not obtain explicit solutions by this method. Although establishing their existence is often sufficient, there are situations where it is necessary to know their explicit forms.

There exists a very simple explicit construction of the Killing spinors on n-dimensional anti-de Sitter spacetime AdS_n, for arbitrary n.[1] This exploits the fact that AdS_n can be written in horospherical coordinates, in terms of which the metric takes the simple form

$$ds^2 = dr^2 + e^{2r}\eta_{\alpha\beta}\,dx^\alpha\,dx^\beta, \tag{1}$$

where $\eta_{\alpha\beta}$ is the Minkowski metric in $(n-1)$ dimensions, and the Ricci tensor satisfies $R_{\mu\nu} = -(n-1)g_{\mu\nu}$. It was shown in Ref. 1 that the Killing spinors, satisfying $D_\mu \epsilon = \frac{1}{2}\Gamma_\mu \epsilon$, are then expressible as

$$\epsilon = e^{(1/2)r}\epsilon_0^+, \quad \text{or} \quad \epsilon = (e^{(-1/2)r} + e^{(1/2)r}x^\alpha \Gamma_\alpha)\epsilon_0^-, \tag{2}$$

where $\epsilon_\pm$ are arbitrary constant spinors satisfying $\Gamma_r \epsilon_0^\pm = \pm \epsilon_0^\pm$. One can alternatively write the two kinds of Killing spinors together in one equation, as

$$\epsilon = e^{(1/2)r\Gamma_r}(1 + \tfrac{1}{2}x^\alpha \Gamma_\alpha(1-\Gamma_r))\epsilon_0, \tag{3}$$

where ϵ_0 is an arbitrary constant spinor. It is therefore manifest that the number of independent Killing spinors is equal to the number of components in the spinors. (The Killing spinors for AdS_4, written in the standard AdS coordinate system, were obtained in Ref. 2.) It is worth remarking that the horospherical metric (1) can equally well have other spacetime signatures $(p, n-p)$, by taking other signatures $(p, n-p-1)$ for the metric $\eta_{\alpha\beta}$. The isometry group is $SO(p+1, n-p)$. The case $p=1$ gives AdS_n, with $SO(2, n-1)$, while $p=0$ gives the positive-

[a)]Electronic mail: rahmfeld@leland.stanford.edu definite hyperbolic metric on H^n, with $SO(1,n)$. [Expressions for the Killing spinors on H^2 and H^3, which are special cases of (3), were given in Ref. 3.] Thus, Eq. (3) gives the Killing spinors on all of the AdS_n spacetimes, hyperbolic spaces H^n, and the other maximally symmetric spacetimes with a $(p,n-p)$ signature.

There is an alternative Killing spinor equation that one can consider when n is even, namely, $D_\mu\epsilon=(i/2)\gamma\Gamma_\mu\epsilon$, where γ is the chirality operator, expressed as an appropriate product over the Γ_μ, with $\gamma^2=1$. We easily see that the solutions of this equation can be written as

$$\epsilon=e^{(i/2)r\gamma\Gamma_r}\left(1+\frac{i}{2}\gamma x^\alpha\Gamma_\alpha(1-i\gamma\Gamma_r)\right)\epsilon_0. \tag{4}$$

Note that in all the cases above, we considered a ''unit radius'' AdS_n, or H^n, etc., given by (1). It is trivial to extend the results to an arbitrary scale size, by replacing (1) by $ds^2=\lambda^{-2}(dr^2+e^{2r}\eta_{\alpha\beta}\,dx^\alpha\,dx^\beta)$, which has the Ricci tensor $R_{\mu\nu}=-(n-1)\lambda^2 g_{\mu\nu}$. The Killing spinor equations then become $D_\mu\epsilon=\frac{1}{2}\lambda\Gamma_\mu\epsilon$, etc. It is easily seen that the solutions are given by precisely the same expressions (3), etc., with no modifications whatsoever. (In Ref. 1 a different coordinatization of the general-radius AdS_n was used, in which the expressions for the Killing spinors *do* depend upon the scale-setting parameter.)

In this paper, we find an explicit construction of the Killing spinors on S^n. (Explicit results for $n=2$ and $n=3$ were obtained in Ref. 3.) One might think that since AdS_n can be related to S^n by appropriate complexifications of coordinates, it should be possible to obtain expressions for the Killing spinors on S^n that are analogous to those given above. However, things are not quite so simple, because the ability to write the metric on AdS_n in the simple form (1) depends rather crucially on the fact that its isometry group $SO(2,n-1)$ is noncompact. (One can easily see that (1) has $(n-1)$ commuting Killing vectors ∂_μ, which exceeds the rank $[(n+1)/2]$ of the isometry group when $n>3$. This is not possible for compact groups.) We shall thus present a different construction for the Killing spinors of S^n, which, although more complicated, is still explicit, and of an essentially simple structure. Our main result is contained in Eq. (7) in Sec. II, which also contains a detailed proof. In Sec. III we combine the results for AdS and spheres, to give the explicit expressions for Killing spinors in various $\mathrm{AdS}_m\times S^n$ supergravity backgrounds, with $(m,n)=(4,7)$, (7, 4), (5, 5), (3, 3), (3, 2), (2, 3), (2, 2). In Appendix A, we collect some useful expressions for the representation and decomposition of Dirac matrices.

II. KILLING SPINORS ON S^n

A. Results

We begin by writing the metric on a unit S^n in terms of that for a unit S^{n-1} as

$$ds_n^2=d\theta_n^2+\sin^2\theta_n\,ds_{n-1}^2, \tag{5}$$

with $ds_1^2=d\theta_1^2$. This has a Ricci tensor given by $R_{ij}=(n-1)g_{ij}$. We then consider the Killing spinor equation on the unit n-sphere, for arbitrary n, namely,

$$D_j\epsilon=\frac{i}{2}\Gamma_j\epsilon. \tag{6}$$

We shall first present our results for the solutions to this equation, and then present the proof later. We find that the Killing spinors can be written as

$$\epsilon=e^{(i/2)\theta_n\Gamma_n}\left(\prod_{j=1}^{n-1}e^{(-1/2)\theta_j\Gamma_{j,j+1}}\right)\epsilon_0, \tag{7}$$

where ϵ_0 is an arbitrary constant spinor, and the indices on the Dirac matrices are vielbein indices. We use the convention that the Γ matrices are Hermitian, satisfying the Clifford algebra

$\{\Gamma_i,\Gamma_j\}=2\delta_{ij}$. Note that here, and in all other analogous formulas in the paper, the factors in the product in (7) are ordered antilexigraphically i.e., starting with the θ_{n-1} term at the left. Note also that the exponential factors in (7) can be written as

$$e^{(i/2)\theta_n\Gamma_n}=\mathbb{1}\cos\tfrac{1}{2}\theta_n+i\Gamma_n\sin\tfrac{1}{2}\theta_n,\quad e^{(-1/2)\theta_j\Gamma_{j,j+1}}=\mathbb{1}\cos\tfrac{1}{2}\theta_j-\Gamma_{j,j+1}\sin\tfrac{1}{2}\theta_j. \tag{8}$$

One can also consider the Killing spinor equation with the opposite sign for the Γ_j term, namely,

$$D_j\epsilon_-=-\frac{i}{2}\Gamma_j\epsilon_-. \tag{9}$$

The previous solution (7) is easily modified to give solutions of this equation. One finds

$$\epsilon_-=e^{(-i/2)\theta_n\Gamma_n}\left(\prod_{j=1}^{n-1}e^{(-1/2)\theta_j\Gamma_{j,j+1}}\right)\epsilon_0. \tag{10}$$

This is immediately verified by noting that (9) is obtained from (6) by changing the sign of the gamma matrices.

The Killing spinors discussed above exist on S^n for any n. When n is even, there is an alternative equation that can also be considered, namely,

$$D_j\epsilon=\tfrac{1}{2}\gamma\Gamma_j\epsilon, \tag{11}$$

where γ is the chirality operator formed from the product of the Γ matrices, satisfying $\gamma^2=1$. In this case, we find that the corresponding Killing spinors can be written as

$$\epsilon=e^{(1/2)\theta_n\gamma\Gamma_n}\left(\prod_{j=1}^{n-1}e^{(-1/2)\theta_j\Gamma_{j,j+1}}\right)\epsilon_0. \tag{12}$$

We may again also consider the Killing spinors satisfying (11) with the sign of the Γ_j term reversed, namely,

$$D_j\epsilon=-\tfrac{1}{2}\gamma\Gamma_j\epsilon. \tag{13}$$

The solutions are again obtained by sending $\theta_n\to-\theta_n$, giving

$$\epsilon_-=e^{(-1/2)\theta_n\gamma\Gamma_n}\left(\prod_{j=1}^{n-1}e^{(-1/2)\theta_j\Gamma_{j,j+1}}\right)\epsilon_0. \tag{14}$$

As in the AdS and H^m cases discussed in Sec. I, we may again trivially extend the results to an n-sphere of arbitrary radius, with metric $ds_n^2=\lambda^{-2}(d\theta_n^2+\sin^2\theta_n\,ds_{n-1}^2)$ and Ricci tensor $R_{ij}=(n-1)\lambda^2 g_{ij}$. The Killing spinor equations are modified to $D_j\epsilon=(i/2)\lambda\Gamma_j\epsilon$, etc., but again the expressions (7), etc. for the Killing spinors receive no modification whatsoever.

B. Proofs

The proofs of these results proceed by substituting our expressions into the corresponding Killing spinor equations. We begin by showing that in the orthonormal basis $e^n=d\theta_n$, $e^a=\sin\theta_n e^a_{(n-1)}$, the spin connection for the metric (5) is given by

$$\omega^{ab}=\omega^{ab}_{(n-1)},\quad \omega^{a,n}=\cos\theta_n e^a_{(n-1)}, \tag{15}$$

where $a \leqslant n-1$, and $e^a_{(n-1)}$, and $\omega^{ab}_{(n-1)}$ are the vielbein and spin connection for S^{n-1}. (Note that the index n always denotes the specific value n of the dimension of the n-sphere.) Thus we can write the vielbein and spin connection on S^n as

$$e^j = \left(\prod_{k=j+1}^{n} \sin \theta_k \right) d\theta_j ,$$

$$\omega^{jk} = \cos \theta_k \left(\prod_{l=j+1}^{k-1} \sin \theta_l \right) d\theta_j , \quad 1 \leqslant j < k \leqslant n. \tag{16}$$

The Killing spinor equation (6) can be written as

$$\partial_j \epsilon + \frac{1}{4} \omega_j^{kl} \Gamma_{kl} \epsilon = \frac{i}{2} e_j^k \Gamma_k \epsilon, \tag{17}$$

where ω_j^{kl} and e_j^k are the coordinate-index components of ω^{kl} and e^k, i.e., $\omega^{kl} = \omega_j^{kl}\, d\theta_j$ and $e^k = e_j^k\, d\theta^j$. These can be read off from (16). Note that the indices on the Γ matrices in (17) are vielbein indices.

We now make the following two definitions:

$$U_j^k \equiv \left(\prod_{l=j+1}^{k} e^{(-1/2)\theta_l \Gamma_{l,l+1}} \right) \Gamma_{j,j+1} \left(\prod_{l=j+1}^{k} e^{(-1/2)\theta_l \Gamma_{l,l+1}} \right)^{-1}, \quad k \geqslant j, \tag{18}$$

$$V_j \equiv e^{(i/2)\theta_n \Gamma_n} U_j^{n-1} e^{(-i/2)\theta_n \Gamma_n}, \tag{19}$$

where, as usual, the factors with the larger l values in the product sit to the left of those with smaller l values. (Note that if the upper limit on the product is less than the lower limit, then it is defined to be 1.) It is now evident that verifying that the expression (7) gives a solution to the Killing spinor equation (6) amounts to proving that

$$V_j = -i e_j^j \Gamma_j + \sum_{k>j}^{n} \omega_j^{jk} \Gamma_{jk} . \tag{20}$$

We prove this by first establishing two lemmata. The first, whose proof is elementary, states that if X and Y are matrices such that $[X,Y] = 2Z$, and $[X,Z] = -2Y$, then

$$e^{(1/2)\theta X} Y e^{(-1/2)\theta X} = \cos \theta Y + \sin \theta Z. \tag{21}$$

The second lemma states that

$$U_j^k = \sec \theta_{k+1} \omega_j^{j,k+1} \Gamma_{j,k+1} + \sum_{l>j}^{k} \omega_j^{jl} \Gamma_{jl} , \quad k \geqslant j. \tag{22}$$

We prove this by induction. From the definition (18), we know that $U_j^j = \Gamma_{j,j+1}$, which clearly satisfies (22) since $\omega_j^{j,j+1} = \cos \theta_{j+1}$. Assuming then that (22) holds for a specific $k \geqslant j$, we will have that

$$U_j^{k+1} \equiv e^{(-1/2)\theta_{k+1}\Gamma_{k+1,k+2}} U_j^k e^{(+1/2)\theta_{k+1}\Gamma_{k+1,k+2}},$$

$$= \sec \theta_{k+1} \omega_j^{j,k+1} e^{(-1/2)\theta_{k+1}\Gamma_{k+1,k+2}} \Gamma_{j,k+1} e^{(1/2)\theta_{k+1}\Gamma_{k+1,k+2}} + \sum_{l>j}^{k} \omega_j^{jl} \Gamma_{jl} , \tag{23}$$

where we have made use of the fact that the Γ_{jl} in the last term all commute with $\Gamma_{k+1,k+2}$, since $l\leqslant k$. The first term can be evaluated using Lemma 1, giving

$$U_j^{k+1}=\sec\theta_{k+1}\omega_j^{j,k+1}(\cos\theta_{k+1}\Gamma_{j,k+1}+\sin\theta_{k+1}\Gamma_{j,k+2})+\sum_{l>j}^{k}\omega_j^{jl}\Gamma_{jl},$$
$$=\tan\theta_{k+1}\omega_j^{j,k+1}\Gamma_{j,k+2}+\sum_{l>j}^{k+1}\omega_j^{jl}\Gamma_{jl}. \tag{24}$$

Now, it follows from (16) that $\omega_j^{j,k+2}=\cos\theta_{k+2}\tan\theta_{k+1}\omega_j^{j,k+1}$. Using this, we then obtain (22) with k replaced by $k+1$, completing the inductive proof.

Having established the lemmata, we can substitute the expression U_j^{n-1} from (22) into the definition of V_j given in (18), giving

$$V_j=\sec\theta_n\omega_j^{jn}e^{(i/2)\theta_n\Gamma_n}\Gamma_{jn}e^{(-i/2)\theta_n\Gamma_n}+\sum_{l>j}^{n-1}\omega_j^{jl}\Gamma_{jl},$$
$$=-i\tan\theta_n\omega_j^{jn}\Gamma_j+\sum_{l>j}^{n}\omega_j^{jl}\Gamma_{jl}, \tag{25}$$

where we have used Lemma 1 to derive the second line. Since $e_j^j=\tan\theta_n\omega_j^{jn}$, as can be seen from (16), it follows that (25) gives (20). This completes the proof that (7) satisfies the Killing spinor equation (6). An essentially identical proof shows that (12) satisfies the alternative Killing spinor equation (11) in even dimensions.

III. KILLING SPINORS ON ADS×SPHERE

An application of the formulas obtained in this paper is to construct the explicit forms of the Killing spinors in the full D-dimensional spacetime of a supergravity theory that admits an $\mathrm{AdS}_m\times S^n$ solution, where $D=m+n$. Consider, for example, the $\mathrm{AdS}_4\times S^7$ solution of $D=11$ supergravity. This is obtained by taking $F_{\mu\nu\rho\sigma}=6m\epsilon_{\mu\nu\rho\sigma}$ with $\mu=0$, 8, 9, 10, implying that the Ricci tensors on AdS_4 and S^7 satisfy $R_{\mu\nu}=-12m^2g_{\mu\nu}$ and $R_{mn}=6m^2g_{mn}$, respectively.[4] The Killing spinors ϵ must satisfy

$$0=\delta\psi_M=D_M\epsilon-\tfrac{1}{288}(\hat\Gamma_{MNPQR}F^{NPQR}-8F_{MNPQ}\hat\Gamma^{NPQ})\epsilon. \tag{26}$$

Using the appropriate decomposition of Dirac matrices given in Appendix A, this implies that on AdS_4 and S^7 we must have

$$\mathrm{AdS}_4{:}D_\mu\epsilon_{AdS}=im\gamma\Gamma_\mu\epsilon_{\mathrm{AdS}},$$
$$S^7{:}D_j\eta=\frac{i}{2}m\Gamma_j\eta, \tag{27}$$

with $j=1,...,7$. From the results obtained in this paper, we find that the Killing spinors on $\mathrm{AdS}_4\times S^7$ can be written as

$$\mathrm{AdS}_4\times S^7{:}\epsilon=e^{(i/2)r\hat\gamma\hat\Gamma_r}\left(1+\frac{1}{2}x^\alpha(i\hat\gamma\hat\Gamma_\alpha+\hat\Gamma_r\hat\Gamma_\alpha)\right)e^{(i/2)\theta_7\hat\gamma\hat\Gamma_7}\left(\prod_{j=1}^{6}e^{(-1/2)\theta_j\hat\Gamma_{j,j+1}}\right)\epsilon_0, \tag{28}$$

where $\hat\gamma\equiv(-i/24)\epsilon^{\mu\nu\rho\sigma}\hat\Gamma_{\mu\nu\rho\sigma}=\gamma\otimes\mathbf{1}$ is a ‘‘pseudo chirality operator,’’ and ϵ_0 is an arbitrary 32-component constant spinor in $D=11$. Note that the explicit numerically assigned indices refer to the seven directions on the 7-sphere. Also, as implied by (27), the AdS_4 and S^7 have different radii, which are related by the 11-dimensional field equations. However, as noted before, in our coordinatization the Killing spinors are independent of the scale sizes.

In $D=11$ supergravity there is also a solution, $\mathrm{AdS}_7\times S^4$. An analogous calculation gives the result that the Killing spinors in this background can be written as

$$\mathrm{AdS}_7\times S^4\colon \epsilon=e^{(1/2)r\hat{\gamma}\hat{\Gamma}_r}(1+\tfrac{1}{2}x^{\alpha}(\hat{\gamma}\hat{\Gamma}_{\alpha}+\hat{\Gamma}_r\hat{\Gamma}_{\alpha}))e^{(1/2)\theta_4\hat{\gamma}\hat{\Gamma}_4}\left(\prod_{j=1}^{3} e^{(-1/2)\theta_j\hat{\Gamma}_{j,j+1}}\right)\epsilon_0, \tag{29}$$

where $\hat{\gamma}\equiv\hat{\Gamma}_{1234}=\mathbf{1}\otimes\gamma$, and all numerically assigned indices refer to the four directions on S^4. Again, ϵ_0 is an arbitrary 32-component constant spinor in $D=11$.

As another explicit example let us look at Type IIB supergravity on $\mathrm{AdS}_5\times S^5$. The gravitino transformation rules are

$$0=\delta\psi_M=D_M\epsilon+\frac{i}{1920}\hat{\Gamma}^{NPQRS}\hat{\Gamma}_M F_{NPQRS}\epsilon, \tag{30}$$

where ϵ is a ten-dimensional spinor of positive chirality, satisfying

$$\hat{\Gamma}_0\cdots\hat{\Gamma}_9\epsilon=\epsilon. \tag{31}$$

Choosing now $F_{\mu\nu\rho\lambda\sigma}=4m\epsilon_{\mu\nu\rho\lambda\sigma}$ and $F_{ijklm}=4m\epsilon_{ijklm}$, Eq. (30) reduces to

$$D_M\epsilon-\tfrac{1}{2}m(\sigma_1\times\mathbf{1}\times\mathbf{1})\hat{\Gamma}_M\epsilon=0, \tag{32}$$

where we are using the (odd, odd) decomposition of Dirac matrices given in Appendix A. With the ansatz

$$\epsilon=\begin{pmatrix}1\\0\end{pmatrix}\otimes\epsilon_{AdS}\otimes\eta, \tag{33}$$

for a spinor of positive chirality, we obtain the equations for the AdS_5 and S^5 subspaces:

$$\mathrm{AdS}_5: D_\mu\epsilon_{AdS}=\tfrac{1}{2}m\Gamma_\mu\epsilon_{AdS},$$

$$S^5: D_j\eta=\frac{i}{2}m\Gamma_j\eta, \tag{34}$$

which are the standard Killing spinor equations. Putting the AdS and S^n results together, we obtain the explicit expression for the Killing spinors on $\mathrm{AdS}_5\times S^5$,

$$\mathrm{AdS}_5\times S^5\colon \epsilon=e^{(i/2)r\hat{\gamma}\hat{\Gamma}_r}(1+\tfrac{1}{2}x^{\alpha}(i\hat{\gamma}\hat{\Gamma}_{\alpha}+\hat{\Gamma}_r\hat{\Gamma}_{\alpha}))e^{(-i/2)\theta_5\hat{\gamma}\hat{\Gamma}_5}\left(\prod_{j=1}^{4} e^{(-1/2)\theta_j\hat{\Gamma}_{j,j+1}}\right)\epsilon_0, \tag{35}$$

where ϵ_0 is an arbitrary 32-component constant spinor of positive chirality, and $\hat{\gamma}\equiv\hat{\Gamma}^{12345}=-\sigma_2\otimes\mathbf{1}\otimes\mathbf{1}$, where the numerical indices lie in S^5.

Four further analogous examples that arise in lower-dimensional supergravities are

$$\mathrm{AdS}_3\times S^3\colon \epsilon=e^{(-i/2)r\hat{\tilde{\gamma}}\hat{\Gamma}_r}(1+\tfrac{1}{2}x^{\alpha}(-i\hat{\tilde{\gamma}}\hat{\Gamma}_{\alpha}+\hat{\Gamma}_r\hat{\Gamma}_{\alpha}))e^{(-i/2)\theta_3\hat{\gamma}\hat{\Gamma}_3}\left(\prod_{j=1}^{2} e^{(-1/2)\theta_j\hat{\Gamma}_{j,j+1}}\right)\epsilon_0,$$

$$\mathrm{AdS}_3\times S^2\colon \epsilon=e^{(1/2)r\hat{\gamma}\hat{\Gamma}_r}(1+\tfrac{1}{2}x^{\alpha}(\hat{\gamma}\hat{\Gamma}_{\alpha}+\hat{\Gamma}_r\hat{\Gamma}_{\alpha}))e^{(1/2)\theta_2\hat{\gamma}\hat{\Gamma}_2}e^{(-1/2)\theta_1\hat{\Gamma}_{12}})\epsilon_0,$$

$$\mathrm{AdS}_2\times S^3: \epsilon = e^{(i/2)r\hat{\gamma}\hat{\Gamma}_r}(1+\tfrac{1}{2}x(i\hat{\gamma}\hat{\Gamma}_x+\hat{\Gamma}_r\hat{\Gamma}_x))e^{(i/2)\theta_3\hat{\gamma}\hat{\Gamma}_3}\left(\prod_{j=1}^{2} e^{(-1/2)\theta_j\hat{\Gamma}_{j,j+1}}\right)\epsilon_0,$$

$$\mathrm{AdS}_2\times S^2: \epsilon = e^{(i/2)r\hat{\gamma}\hat{\Gamma}_r}(1+\tfrac{1}{2}x(i\hat{\gamma}\hat{\Gamma}_x+\hat{\Gamma}_r\hat{\Gamma}_x))e^{(i/2)\theta_2\hat{\gamma}\hat{\Gamma}_2}e^{(-1/2)\theta_1\hat{\Gamma}_{12}}\epsilon_0,$$

or

$$\epsilon = e^{(1/2)r\hat{\gamma}\hat{\Gamma}_r}(1+\tfrac{1}{2}x(\hat{\gamma}\hat{\Gamma}_x+\hat{\Gamma}_r\hat{\Gamma}_x))e^{(1/2)\theta_2\hat{\gamma}\hat{\Gamma}_2}e^{(-1/2)\theta_1\hat{\Gamma}_{12}}\epsilon_0, \tag{36}$$

where the Dirac matrices are the ones appropriate to the total spacetime dimension in each case. In the case where one or another space in the factored product is even dimensional, $\hat{\gamma}$ is the pseudochirality operator given by the appropriate product of the overcareted Dirac matrices in the even-dimensional factor. For this reason, there are two possibilities in the $\mathrm{AdS}_2\times S^2$ example, reflecting the two possibilities for the Dirac matrix decomposition given in Appendix A. The first corresponds to taking $\hat{\gamma}$ to be the pseudochirality operator in AdS_2, and the second to taking it instead to be in S^2. In the case of $\mathrm{AdS}_3\times S^3$, $\hat{\gamma}\equiv\hat{\Gamma}^{123}=-i\sigma_2\otimes\mathbf{1}\otimes\mathbf{1}$, where the numerical indices lie in S^3, while $\hat{\hat{\gamma}}\equiv\frac{1}{6}\epsilon_{\mu\nu\rho}\hat{\Gamma}^{\mu\nu\rho}=\sigma_1\otimes\mathbf{1}\otimes\mathbf{1}$. In all the examples, ϵ_0 is an arbitrary constant spinor in the total space. It will be subject to a chirality condition in the $\mathrm{AdS}_3\times S^3$ example, if the $D=6$ supergravity is chosen to be the minimal chiral theory, and $\hat{\hat{\gamma}}$ can then be replaced by $\hat{\gamma}$ in the expression for the Killing spinors.

IV. DISCUSSION

In this paper, we have obtained explicit expressions for the Killing spinors on S^n for all n. We then used the results to obtain the full Killing spinors on various $\mathrm{AdS}_m\times S^n$ spacetimes that arise as solutions in supergravity theories. These are of considerable interest owing to the conjectured duality relation to conformal theories on the AdS boundaries. One further application of these results is to construct the Killing vectors, and conformal Killing vectors, from appropriate bilinear products $\epsilon'^{\dagger}\Gamma_i\epsilon$ of Killing spinors. As discussed in Ref. 3, products where the Killing spinors ϵ' and ϵ on S^n either both satisfy (6) or both satisfy (9) give Killing vectors, while products where one satisfies (6) and the other satisfies (9) give conformal Killing vectors. In general, it is necessary to use both of the Killing-vector constructions in order to obtain all the Killing vectors on S^n. At large n there is a considerable redundancy in the construction, since the number of Killing spinors grows exponentially with n, while the number of Killing vectors grows only quadratically with n. In certain low dimensions, there is a more elegant exact spanning of the Killing vectors using this construction, such as for S^7 where the antisymmetric products $\overline{\epsilon^{\alpha}}\Gamma_i\epsilon^{\beta}$ of the eight Killing spinors ϵ^{α} give the 28 Killing vectors of $SO(8)$.[4]

We shall present just one simple example here, for the case of S^2. From the matrix expression (B1) in the appendix, we find that from the Killing spinors $\epsilon=\Omega_2\binom{a}{b}$ and $\epsilon'=\Omega_2\binom{a'}{b'}$, we obtain the Killing vectors,

$$K=K^i\partial_i=E^i_j\epsilon'^{\dagger}\Gamma^j\epsilon=(b\overline{b'}-a\overline{a'})\frac{\partial}{\partial\theta_1}+i(a\overline{b'}e^{-i\theta_1}-\overline{a'}be^{i\theta_1})\frac{\partial}{\partial\theta_2}$$
$$+(a\overline{b'}e^{-i\theta_1}+\overline{a'}be^{i\theta_1})\cot\theta_2\frac{\partial}{\partial\theta_1}, \tag{37}$$

where E^i_j are the components of the inverse vielbein $E_j=E^i_j\partial_i$. Choosing different values for the constants a, b, a', b' spans the complete set of three Killing vectors of $SO(3)$.

ACKNOWLEDGMENTS

We are grateful to Renata Kallosh for posing the question of whether simple explicit expressions for the Killing spinors on S^n can be obtained. J. R. thanks Arvind Rajaraman for useful discussions. The Research was supported in part by Department of Energy Grant No. DE-FG03-95ER40917, National Science Foundation Grant No. PHY-9219345, and the Unité Propre du Centre National de la Recherche Scientifique, associée *à* l'École Normale Supérieure et *à* l'Université de Paris-Sud.

APPENDIX A: DIRAC MATRICES AND THEIR DECOMPOSITION ON PRODUCT SPACES

It is useful in general to represent the Dirac matrices in terms of the 2×2 Pauli matrices $\{\sigma_1,\sigma_2,\sigma_3\}$ as follows. In even dimensions $D=2n$, we have

$$\begin{aligned}
\Gamma_1 &= \sigma_1\otimes \mathbb{1}\otimes \mathbb{1}\otimes\cdots\otimes \mathbb{1},\\
\Gamma_2 &= \sigma_2\otimes \mathbb{1}\otimes \mathbb{1}\otimes\cdots\otimes \mathbb{1},\\
\Gamma_3 &= \sigma_3\otimes \sigma_1\otimes \mathbb{1}\otimes\cdots\otimes \mathbb{1},\\
\Gamma_4 &= \sigma_3\otimes \sigma_2\otimes \mathbb{1}\otimes\cdots\otimes \mathbb{1},\\
\Gamma_5 &= \sigma_3\otimes \sigma_3\otimes \sigma_1\otimes\cdots\otimes \mathbb{1},\\
&\ldots,\\
\Gamma_{2n-1} &= \sigma_3\otimes\sigma_3\otimes\cdots\otimes\sigma_3\otimes\sigma_1,\\
\Gamma_{2n} &= \sigma_3\otimes\sigma_3\otimes\cdots\otimes\sigma_3\otimes\sigma_2.
\end{aligned} \tag{A1}$$

In odd dimensions $D=2n+1$, we use the above construction for the Dirac matrices of $2n$ dimensions, and take

$$\Gamma_{2n+1} = \sigma_3\otimes\sigma_3\otimes\sigma_3\otimes\cdots\otimes\sigma_3. \tag{A2}$$

When performing Kaluza–Klein reductions, it is necessary to decompose the Dirac matrices of D dimensions in terms of those of the lower-dimensional spacetime M_m, and the internal space K_n, whose respective dimensions m and n add up to D. There are four cases that arise, namely, $(m,n)=$(even, odd), (odd, even), (even, even) and (odd, odd). If we denote the Dirac matrices of the spacetime M_m by Γ_μ, and those of the internal space K_n by Γ_i, then the Dirac matrices $\hat\Gamma_A$ of $M_m\times K_n$ can be written as

$$\begin{aligned}
&(\text{even, odd}): \hat\Gamma_\mu = \Gamma_\mu\otimes\mathbb{1}, \quad \hat\Gamma_i = \gamma\otimes\Gamma_i,\\
&(\text{odd, even}): \hat\Gamma_\mu = \Gamma_\mu\otimes\gamma, \quad \hat\Gamma_i = \mathbf{1}\otimes\Gamma_i,\\
&(\text{even, even}): \hat\Gamma_\mu = \Gamma_\mu\otimes\mathbb{1}, \quad \hat\Gamma_i = \gamma\otimes\Gamma_i,\\
&\qquad \text{or}\\
&\qquad \hat\Gamma_\mu = \Gamma_\mu\otimes\gamma, \quad \hat\Gamma_i = \mathbf{1}\otimes\Gamma_i,\\
&(\text{odd, odd}): \hat\Gamma_\mu = \sigma_1\otimes\Gamma_\mu\otimes\mathbf{1}, \quad \hat\Gamma_i = \sigma_2\otimes\mathbf{1}\otimes\Gamma_i.
\end{aligned} \tag{A3}$$

Note that in the final case the extra Pauli matrices σ_1 and σ_2 are needed in order to satisfy the Clifford algebra, in view of the fact that the Dirac matrices of D dimensions are twice the size of the simple tensor products of those in M_m and K_n. Note also in this case that the chirality operator in the total space is $\sigma_3 \otimes \mathbf{1} \times \mathbf{1}$.

APPENDIX B: SOME LOW-DIMENSIONAL EXAMPLES

In this appendix, we give explicit matrix expressions for the Killing spinors on the spheres S^2, S^3, S^4, and S^5. These examples arise in the near-horizon structures of Reißner–Nordstrøm black holes, dyonic strings, M5-branes, and D3-branes, respectively. In each case, we may write the expression (7) for the Killing spinors on S^n as $\epsilon = \Omega_n \epsilon_0$. For S^2, taking $\Gamma_i = \sigma_i$, where σ_i are the usual Pauli matrices, we find

$$\Omega_2 = \begin{pmatrix} e^{(-i/2)\theta_1} \cos\frac{1}{2}\theta_2 & e^{(i/2)\theta_1} \sin\frac{1}{2}\theta_2 \\ -e^{(-i/2)\theta_1} \sin\frac{1}{2}\theta_2 & e^{(i/2)\theta_1} \cos\frac{1}{2}\theta_2 \end{pmatrix}. \tag{B1}$$

To avoid clumsy expressions later, we may define $t_k = e^{(i/2)\theta_k}$, $\bar{t}_k = e^{(-i/2)\theta_k}$, $c_k = \cos\frac{1}{2}\theta_k$, $s_k = \sin\frac{1}{2}\theta_k$. The matrix Ω_2 thus becomes

$$\Omega_2 = \begin{pmatrix} \bar{t}_1 c_2 & t_1 s_2 \\ -\bar{t}_1 s_2 & t_1 c_2 \end{pmatrix}. \tag{B2}$$

For S^3, S^4, and S^5 we obtain

$$\Omega_3 = \begin{pmatrix} \bar{t}_1 t_3 c_2 & -i t_1 t_3 s_2 \\ -i \bar{t}_1 \bar{t}_3 s_2 & t_1 \bar{t}_3 c_2 \end{pmatrix}, \tag{B3}$$

$$\Omega_4 = \begin{pmatrix} \bar{t}_1 \bar{t}_3 c_2 c_4 & \bar{t}_1 \bar{t}_3 c_2 s_4 & -i t_1 \bar{t}_3 s_2 s_4 & -i t_1 \bar{t}_3 c_4 s_2 \\ -\bar{t}_1 \bar{t}_3 c_2 s_4 & \bar{t}_1 \bar{t}_3 c_2 c_4 & -i t_1 \bar{t}_3 c_4 s_2 & i t_1 \bar{t}_3 s_2 s_4 \\ i \bar{t}_1 t_3 s_2 s_4 & -i \bar{t}_1 t_3 c_4 s_2 & t_1 t_3 c_2 c_4 & -t_1 t_3 c_2 s_4 \\ -i \bar{t}_1 t_3 c_4 s_2 & -i \bar{t}_1 t_3 s_2 s_4 & t_1 t_3 c_2 s_4 & t_1 t_3 c_2 c_4 \end{pmatrix}, \tag{B4}$$

$$\Omega_5 = \begin{pmatrix} \bar{t}_1 \bar{t}_3 \bar{t}_5 c_2 c_4 & -i \bar{t}_1 \bar{t}_3 \bar{t}_5 c_2 s_4 & -t_1 \bar{t}_3 \bar{t}_5 s_2 s_4 & -i t_1 \bar{t}_3 \bar{t}_5 c_4 s_2 \\ -i \bar{t}_1 \bar{t}_3 \bar{t}_5 c_2 s_4 & \bar{t}_1 \bar{t}_3 \bar{t}_5 c_2 c_4 & -i t_1 \bar{t}_3 \bar{t}_5 c_4 s_2 & -t_1 \bar{t}_3 \bar{t}_5 s_2 s_4 \\ -\bar{t}_1 t_3 t_5 s_2 s_4 & -i \bar{t}_1 t_3 t_5 c_4 s_2 & t_1 t_3 t_5 c_2 c_4 & -i t_1 t_3 t_5 c_2 s_4 \\ -i \bar{t}_1 t_3 t_5 c_4 s_2 & -\bar{t}_1 t_3 t_5 s_2 s_4 & -i t_1 t_3 t_5 c_2 s_4 & t_1 t_3 t_5 c_2 c_4 \end{pmatrix}. \tag{B5}$$

In these examples we have used the representations of Dirac matrices given in Eqs. (A1) and (A2) of appendix A.

[1] H. Lü, C. N. Pope, and P. K. Townsend, ''Domain walls from anti-de Sitter spacetime,'' Phys. Lett. B **391**, 39 (1997), hep-th/9607164.
[2] P. Breitenlohner and D. Z. Freedman, ''Stability in gauged extended supergravity,'' Ann. Phys. (N.Y.) **144**, 249 (1982).
[3] Y. Fujii and K. Yamagishi, ''Killing spinors on spheres and hyperbolic manifolds,'' J. Math. Phys. **27**, 979 (1986).
[4] M. J. Duff, B. E. W. Nilsson, and C. N. Pope, ''Kaluza–Klein supergravity,'' Phys. Rep. **130**, 1 (1986).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Gauge fixation and global phase time for minisuperspaces

Claudio Simeone[a)]
Departamento de Física, Comisión Nacional de Energía Atómica, Av. del Libertador 8250, 1429 Buenos Aires, Argentina and Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón I, 1428 Buenos Aires, Argentina

(Received 22 March 1999; accepted for publication 16 April 1999)

Homogeneous and isotropic cosmological models whose Hamilton–Jacobi equation is separable are deparametrized by turning their action functional into that of an ordinary gauge system. Canonical gauge conditions imposed on the gauge system are used to define a global phase time in terms of the canonical coordinates and momenta of the minisuperspaces. The procedure clearly shows how the geometry of the constraint surface restricts the choice of time; the consequences that this has on the path integral quantization are discussed. © *1999 American Institute of Physics.* [S0022-2488(99)00709-4]

I. INTRODUCTION

While in ordinary mechanics the time is an absolute parameter, and this allows for the existence of a unitary quantum theory, in General Relativity the time is an arbitrary label of spacelike hypersurfaces, and physical quantities are invariant under diffeomorfisms. The gravitational field in General Relativity is a parametrized system, its evolution given in terms of a parameter τ which does not have physical significance.

A possible way to obtain a unitary quantum theory of gravitation is to consider that the time is hidden among the coordinates and momenta of the system, which then must be deparametrized by identifying the time as a first step before quantization. The identification of time is closely related to gauge fixation.[1] In the theory of gravitation the dynamical evolution is embodied in the motion of a spacelike hypersurface moving in space–time along the timelike direction; this motion includes arbitrary local deformations which yield a multiplicity of times. From a different point of view, the same motion can be generated by general gauge transformations. Hence, the gauge fixation is not only a way to select one path from each class of equivalent paths in phase space, but also a reduction procedure identifying a time for the system.

In the present work we exploit this fact to identify a global phase time[2] for minisuperspace models whose Hamilton–Jacobi (H–J) equation is solvable. We define a canonical transformation which turns the cosmological models into ordinary gauge systems by matching their Hamiltonian constraint $H\approx 0$ to one of the new momenta, namely P_0, of the gauge system.[3,4] Then we are able to avoid derivative gauges involving Lagrange multipliers,[5,6,7] and to use gauge conditions given in terms of only the coordinates and momenta (*canonical gauges*) to identify a time in terms of the original phase space variables of the cosmological models. The results of Ref. 8 are easily reproduced. We show how the geometry of the constraint surface determines restrictions on the existence of an intrinsic time;[9] we also discuss the consequences that these restrictions have for the path integral quantization of minisuperspaces, making more precise the analysis of Ref. 3.

II. PARAMETRIZED SYSTEMS AND ORDINARY GAUGE SYSTEMS

The action functional of a parametrized system described by the coordinates and momenta (q^i,p_i) has the form

[a)]Electronic mail: simeone@tandar.cnea.gov.ar

$$S[q^i,p_i,N]=\int_{\tau_1}^{\tau_2}\left(p_i\frac{dq^i}{d\tau}-NH\right)d\tau, \tag{1}$$

where N is a Lagrange multiplier enforcing the Hamiltonian constraint

$$H(q^i,p_i)\approx 0; \tag{2}$$

the constraint reflects the reparametrization invariance of the system, i.e., that its evolution is given in terms of the arbitrary parameter τ which does not have physical meaning.

The parametrized system described by (1) can be turned into an ordinary gauge system, that is, a system with a true Hamiltonian and a constraint which is linear and homogeneous in the momenta if the H–J equation is solvable.[3] Consider $W(q^i,\alpha_\mu,E)$ a complete solution of the τ-independent H–J equation,

$$H\left(q^i,\frac{\partial W}{\partial q^i}\right)=E, \tag{3}$$

which is obtained by matching the integration constants (α_μ,E) to $(\overline{P}_\mu,\overline{P}_0)$. The solution W generates a canonical transformation

$$p_i=\frac{\partial W}{\partial q^i},\quad \overline{Q}^i=\frac{\partial W}{\partial \overline{P}_i},\quad \overline{K}=N\overline{P}_0=NH, \tag{4}$$

which identifies the constraint H with the new momentum $\overline{P}_0$. The variables $(\overline{Q}^\mu,\overline{P}_\mu)$ are conserved observables because $[\overline{Q}^\mu,H]=[\overline{P}_\mu,H]=0$, so that they would not be appropriate to characterize the dynamical evolution. A second transformation generated by the function

$$F=P_0\overline{Q}^0+f(\overline{Q}^\mu,P_\mu,\tau) \tag{5}$$

gives

$$\begin{aligned}\overline{P}_0&=\frac{\partial F}{\partial \overline{Q}^0}=P_0,\quad \overline{P}_\mu=\frac{\partial f}{\partial \overline{Q}^\mu},\\ Q^0&=\frac{\partial F}{\partial P_0}=\overline{Q}^0,\quad Q^\mu=\frac{\partial f}{\partial P_\mu},\end{aligned} \tag{6}$$

and a new nonvanishing Hamiltonian

$$K=NP_0+\frac{\partial f}{\partial \tau}=NH+\frac{\partial f}{\partial \tau}, \tag{7}$$

so that (Q^μ,P_μ) are nonconserved observables. The two successive transformations $(q^i,p_i)\to(\overline{Q}^i,\overline{P}_i)\to(Q^i,P_i)$ lead to the action

$$\mathcal{S}=\int_{\tau_1}^{\tau_2}\left(P_i\frac{dQ^i}{d\tau}-NP_0-\frac{\partial f}{\partial \tau}\right)d\tau \tag{8}$$

which in terms of the original variables reads[3]

$$\mathcal{S}=\int_{\tau_1}^{\tau_2}\left(p_i\frac{dq^i}{d\tau}-NH\right)d\tau+[\overline{Q}^i\overline{P}_i-W+Q^\mu P_\mu-f]_{\tau_1}^{\tau_2}, \tag{9}$$

so that $\mathcal{S}$ and S differ only in surface terms and then yield the same dynamics. The action (8) contains a linear and homogeneous constraint $P_0 \approx 0$ and a nonzero Hamiltonian $(\partial f/\partial \tau)$ and is then that of an ordinary gauge system.

III. GAUGE FIXATION AND GLOBAL PHASE TIME

The constraint $P_0 \approx 0$ in Eq. (8) acts as a generator of gauge transformations yielding an infinite number of physically equivalent paths in the (Q^i, P_i) phase space. To select one path from each class of equivalent paths we must impose a gauge condition $\chi = 0$, the choice being restricted by

(1) The gauge condition must can be reached from any path by means of gauge transformations leaving the action unchanged.
(2) Only one point of each orbit (that is, each set of points on the constraint surface connected by gauge transformations) must be on the manifold defined by $\chi = 0$.

To accomplish with (1) the symmetries of the action must be examined; under a gauge transformation generated by a constraint G,

$$\delta_\epsilon Q^i = \epsilon(\tau)[Q^i, G], \quad \delta_\epsilon P_i = \epsilon(\tau)[P_i, G], \quad \delta_\epsilon N = \frac{\partial \epsilon(\tau)}{\partial \tau}, \tag{10}$$

the variation of the action $\mathcal{S}$ is

$$\delta_\epsilon \mathcal{S} = \left[\epsilon(\tau) \left(P_i \frac{\partial G}{\partial P_i} - G \right) \right]_{\tau_1}^{\tau_2} \tag{11}$$

and we have $\delta_\epsilon \mathcal{S} = 0$ for $G = P_0$. Therefore the action $\mathcal{S}$ is gauge invariant over the whole trajectorie and canonical gauge conditions $\chi(Q^i, P_i, \tau) = 0$ are admissible.[3] We should emphasize that if we worked with the original action S, as the constraint in S is $H \approx 0$ and H is not linear and homogeneous in the momenta for a parametrized system, then we should fix the gauge by means of a noncanonical condition involving a derivative of the multiplier N;[5,6] this is clearly not a good choice if we want to define a global phase time in terms of the phase space variables.

Condition (2) requires that a gauge transformation moves a point of an orbit off the surface $\chi = 0$, so that[10]

$$\delta_\epsilon \chi = \epsilon(\tau)[\chi, G] \neq 0$$

unless $\epsilon = 0$; this holds if

$$[\chi, G] \neq 0. \tag{12}$$

As Q^0 and P_0 are conjugate variables,

$$[Q^0, P_0] = 1, \tag{13}$$

so that a gauge condition of the form

$$\chi \equiv Q^0 - T(\tau) = 0 \tag{14}$$

with T a monotonic function is a good choice. Strictly speaking, Eq. (12) only ensures that the orbits are not tangent to the surface $\chi = 0$; however, as (14) defines a plane $Q^0 =$ constant for each τ, if at any τ any orbit was intersected more than once (then yielding Gribov copies[10]) at another τ it should be $[\chi, P_0] = 0$. Therefore our gauge fixation procedure avoids the Gribov problem.

Given a parametrized system with coordinates and momenta (q^i, p_i) a smooth function $t(q^i, p_i)$ fulfilling

$$[t,H]>0 \tag{15}$$

is a global phase time[2] for the system, and its values along any classical trajectory can parametrize its evolution. As the Poisson bracket is invariant under a canonical transformation, from (13) and (15) it follows that a globally good gauge choice given in terms of the coordinate Q^0 of the gauge system can be used to define a global phase time t for the parametrized system in terms of the coordinates and momenta (q^i,p_i). In other words, a gauge choice for the gauge system defines a particular foliation of space–time for the parametrized system. We shall see that for certain minisuperspace models a τ-dependent gauge condition of the form $\chi\equiv Q^0-T(\tau)=0$ defines an extrinsic time, that is, a time which is a function not only of the coordinates q^i but also of the momenta p_i, while an intrinsic time, i.e., a function of the coordinates q^i only, can be defined by means of a gauge condition like $\chi\equiv\eta Q^0P-T(\tau)$ with $\eta=\pm1$ if the potential of the model under consideration has a definite sign; in this situation, the constraint surface splits into two disjoint surfaces, and η is determined by the sheet on which the system evolves.

IV. MINISUPERSPACES

The action of an homogeneous and isotropic Friedmann–Robertson–Walker (FRW) universe is

$$S=\int_{\tau_1}^{\tau_2}(\pi_\phi\dot{\phi}+\pi_\Omega\dot{\Omega}-NH)d\tau \tag{16}$$

where ϕ is the matter field, $\Omega=\sqrt{(4/3\pi\mathcal{G})}\ln a(\tau)$ with $a(\tau)$ the scale factor in the FRW metric, and π_ϕ and π_Ω are their conjugate momenta; N is a Lagrange multiplier enforcing the Hamiltonian constraint[11]

$$H=G(\Omega)(\pi_\phi^2-\pi_\Omega^2)+v(\phi,\Omega)\approx0, \tag{17}$$

where $G(\Omega)>0$ and $v(\phi,\Omega)$ is the potential. Our aim is not to study the separability of the H–J equation in general, but to get a clear understanding of the details and also of the restrictions of deparametrizing minisuperspaces by imposing canonical gauge conditions; then we shall limit our analysis to easily solvable models.

A. A toy model

Consider the Hamiltonian constraint

$$H=-\tfrac{1}{4}e^{-3\Omega}\pi_\Omega^2+e^\Omega\approx0 \tag{18}$$

which corresponds to an open ''universe'' with null cosmological constant. For this model the authors of Ref. 8 found a time of the form

$$t\sim-e^{-4\Omega/3}\pi_\Omega \tag{19}$$

by matching the model to the parametrized system called ''ideal clock,'' whose Hamiltonian is $\tilde{H}=p_t-t^2\approx0$. They did it by performing a canonical transformation $(t,p_t)\rightarrow(\Omega,\pi_\Omega)$ and multiplying $\tilde{H}$ by a positive definite function of the form $\sim e^{-\Omega/3}$ to obtain H. Then we shall apply our procedure to the constraint $H'=e^{\Omega/3}H$,

$$H'=-\tfrac{1}{4}e^{-8\Omega/3}\pi_\Omega^2+e^{4\Omega/3}\approx0. \tag{20}$$

The constraint H' is equivalent to H because they differ only in a positive definite factor (see below). The τ-independent H–J equation associated with the Hamiltonian H' is

$$-\left(\frac{\partial W}{\partial \Omega}\right)^2 + 4e^{4\Omega} = 4e^{8\Omega/3}E, \tag{21}$$

and then matching $E=\overline{P}_0$ we have

$$W(\Omega,\overline{P}_0) = \pm \int 2\sqrt{e^{4\Omega} - \overline{P}_0 e^{8\Omega/3}}\, d\Omega, \tag{22}$$

with $+$ for $\pi_\Omega > 0$ and $-$ for $\pi_\Omega < 0$. According to Eq. (6), on the constraint surface

$$Q^0 = \overline{Q}^0 = \left[\frac{\partial W}{\partial \overline{P}_0}\right]_{\overline{P}_0=0} = \mp e^{2\Omega/3}. \tag{23}$$

The system described by Q^0 and P_0 has a constraint which is linear and homogeneous in the momenta. Its action functional is then invariant under general gauge transformations, so that there is gauge freedom at the end points and canonical gauges are admissible. If we choose $\chi \equiv Q^0 - T(\tau) = 0$ with T a monotonic function of τ then we have a global phase time that can be written in terms of the coordinate Ω only, the expression given by the sheet of the constraint surface on which the system evolves,

$$\begin{aligned} t(\Omega) &= -e^{2\Omega/3} \quad \text{if} \quad \pi_\Omega > 0, \\ t(\Omega) &= +e^{2\Omega/3} \quad \text{if} \quad \pi_\Omega < 0. \end{aligned} \tag{24}$$

As on the constraint surface we have

$$\pi_\Omega = \pm 2e^{2\Omega}, \tag{25}$$

then we can write

$$t(\Omega,\pi_\Omega) = -\tfrac{1}{2} e^{-4\Omega/3} \pi_\Omega, \tag{26}$$

which clearly agrees with (19).

B. True degrees of freedom

Let us go back to the general constraint (17). We shall restrict our analysis to the cases in which the potential $v(\phi,\Omega)$ has a definite sign. As the cases $v>0$ and $v<0$ are formally analogous, to simplify the notation we shall consider only $v>0$. Define the coordinates

$$x = x(\phi+\Omega), \quad y = y(\phi-\Omega), \tag{27}$$

so that $(\partial x/\partial\phi) = (\partial x/\partial\Omega)$, $(\partial y/\partial\phi) = -(\partial y/\partial\Omega)$. The momenta π_x, π_y are given by

$$\pi_\phi = \frac{\partial x}{\partial \phi}\pi_x + \frac{\partial y}{\partial \phi}\pi_y, \quad \pi_\Omega = \frac{\partial x}{\partial \Omega}\pi_x + \frac{\partial y}{\partial \Omega}\pi_y, \tag{28}$$

and then $\pi_\phi^2 - \pi_\Omega^2 = 4(\partial x/\partial\phi)(\partial y/\partial\phi)\pi_x\pi_y = -4(\partial x/\partial\Omega)(\partial y/\partial\Omega)\pi_x\pi_y$. If it is possible to choose the coordinates x and y so that $4(\partial x/\partial\phi)(\partial y/\partial\phi) = (v/G)$, as $(v/G)>0$ then we can multiply the constraint H by $(4G(\partial x/\partial\phi)(\partial y/\partial\phi))^{-1}$ and obtain a constraint H' which is equivalent to H because it differs only in a positive definite factor,

$$H' = \pi_x\pi_y + 1 \approx 0. \tag{29}$$

We shall turn the system described by (x,y,π_x,π_y) into an ordinary gauge system. The τ-independent H–J equation for the constraint (29) is

$$\frac{\partial W}{\partial x}\frac{\partial W}{\partial y}+1=E'$$

and matching the integration constants α, E' to the new momenta $\overline{P},\overline{P}_0$ it has the solution

$$W(x,y,\overline{P}_0,\overline{P})=\overline{P}x+y\left(\frac{\overline{P}_0-1}{\overline{P}}\right); \tag{30}$$

then

$$\begin{gathered}\pi_x=\frac{\partial W}{\partial x}=\overline{P},\quad \pi_y=\frac{\partial W}{\partial y}=\frac{\overline{P}_0-1}{\overline{P}},\\ \overline{Q}^0=\frac{\partial W}{\partial \overline{P}_0}=\frac{y}{\overline{P}},\quad \overline{Q}=\frac{\partial W}{\partial \overline{P}}=x+y\left(\frac{1-\overline{P}_0}{\overline{P}^2}\right).\end{gathered} \tag{31}$$

To go from the set $(\overline{Q}^i,\overline{P}_i)$ to (Q^i,P_i) we define

$$F=\overline{Q}^0P_0+\overline{Q}P+\frac{T(\tau)}{P} \tag{32}$$

with $T(\tau)$ a monotonic function (see the next section for a discussion about this choice). Then we have the canonical variables of the gauge system in terms of those of the minisuperspace,

$$\begin{gathered}P_0=\pi_x\pi_y+1,\quad P=\pi_x,\\ Q^0=\frac{y}{P},\quad Q=x+\left(\frac{y(1-P_0)-T(\tau)}{P^2}\right).\end{gathered} \tag{33}$$

There is no problem with P as a denominator because $P=\pi_x$ cannot be zero on the constraint surface.

As $[Q^0,P_0]=1$ we have $[y/\pi_x,H']=1$; H' differs from H in a positive definite factor, namely a, so that $1=[y/\pi_x,H']=[y/\pi_x,aH]=[y/\pi_x,a]H+[y/\pi_x,H]a\approx[y/\pi_x,H]a$; hence

$$[y/\pi_x,H]>0 \tag{34}$$

and a canonical gauge condition of the form $\chi\equiv Q^0-T(\tau)=0$ with T a monotonic function of τ, when imposed on the gauge system described by Q^i and P_i defines a global phase time $t\equiv y/\pi_x$ for the minisuperspace described by ϕ, Ω, π_ϕ, π_Ω. From (28) we have $\pi_x=(\pi_\phi+\pi_\Omega)(2(\partial x/\partial\phi))^{-1}$ and therefore

$$t(\phi,\Omega,\pi_\phi,\pi_\Omega)=2\,\frac{y(\phi-\Omega)}{\pi_\phi+\pi_\Omega}\,\frac{\partial x(\phi+\Omega)}{\partial\phi}. \tag{35}$$

The monotonic function of τ given by (35) depends on the coordinates and also on the momenta of the cosmological model, and is then an extrinsic time.

We can also identify a time in terms of the coordinates only, but, as we shall see, the definition depends on the sheet of the constraint surface on which the system evolves. The gauge choice

$$\chi \equiv \eta Q^0 P - T(\tau) = 0 \tag{36}$$

with $\eta = \pm 1$ gives

$$[\chi, P_0] = \eta P \tag{37}$$

and as $\eta Q^0 P = \eta y$ and $P = \pi_x$ we have

$$[\eta y, H'] = \eta \pi_x . \tag{38}$$

As before, as H' and H differ in a positive definite factor, if we can define η so that $[\eta y, H'] > 0$ then $[\eta y, H] > 0$ and ηy is a global phase time. We can chose $(\partial x/\partial \phi)$ as a positive definite function (and appropriately adjust the sign of $(\partial y/\partial \phi)$) to yield sign$(\pi_x) =$ sign$(\pi_\phi + \pi_\Omega)$. From the constraint equation we have

$$\pi_\Omega = \pm \sqrt{\frac{v(\phi,\Omega)}{G(\Omega)} + \pi_\phi^2} \tag{39}$$

and because v/G is positive definite, $\pi_\Omega \neq 0$ and the evolution of the system is restricted to one of the two disjoint surfaces (39), each one topologically equivalent to half a plane. Moreover, from (39) we have $|\pi_\Omega| > |\pi_\phi|$, yielding sign$(\pi_x) > 0$ for $\pi_\Omega > 0$ and sign$(\pi_x) < 0$ for $\pi_\Omega < 0$. Hence we can have a good definition of time on each sheet of the constraint surface by appropriately choosing η, the choice dictated by the sign of the momentum π_Ω,

$$\begin{aligned} t(\phi,\Omega) &= +y(\phi - \Omega) \quad \text{if} \quad \pi_\Omega > 0, \\ t(\phi,\Omega) &= -y(\phi - \Omega) \quad \text{if} \quad \pi_\Omega < 0. \end{aligned} \tag{40}$$

Therefore, even though we can not write a single expression which holds for both sheets of the constraint surface, if v has a definite sign, once we have on which sheet the system evolves we can identify a time in terms of the coordinates (intrinsic time). If, instead, we want an expression which holds automatically, that is, which does not depend on the sign of π_Ω, we must choose a time like that given in (35).

Examples:

(1) Consider a flat model with massless scalar field ϕ and a cosmological ''constant'' which decays with ϕ as $\Lambda = \Lambda_0 e^{-6\phi}$,

$$H = \tfrac{1}{4} e^{-3\Omega}(\pi_\phi^2 - \pi_\Omega^2) + \Lambda_0 e^{-6\phi} e^{3\Omega} \approx 0. \tag{41}$$

This constraint is equivalent to

$$H' = \pi_\phi^2 - \pi_\Omega^2 + 4\Lambda_0 e^{-6(\phi-\Omega)} \approx 0,$$

making the choice of variables $x = \phi + \Omega$, $y = -(\Lambda_0/6) e^{-6(\phi-\Omega)}$ obvious; by turning the system into an ordinary gauge system with coordinates and momenta (Q^0, Q, P_0, P) and fixing the gauge with the canonical condition $\chi \equiv Q^0 - T(\tau) = 0$ with T monotonic function we obtain the time

$$t(\phi,\Omega,\pi_\phi,\pi_\Omega) = -\frac{1}{3}\frac{\Lambda_0 e^{-6(\phi-\Omega)}}{\pi_\phi + \pi_\Omega}, \tag{42}$$

which on the constraint surface is equivalent to

$$t(\pi_\phi, \pi_\Omega) = \tfrac{1}{2}(\pi_\phi - \pi_\Omega).$$

The system also has an intrinsic time, which according to (40) can be written as

$$t=\mp\frac{\Lambda_0}{6}e^{-6(\phi-\Omega)},$$

with $-$ if the system is on the sheet $\pi_\Omega>0$ and $+$ if it is on the sheet $\pi_\Omega<0$.

(2) A closed ($k=1$) model with cosmological constant $\Lambda>0$ and massless scalar field ϕ, whose Hamiltonian constraint is

$$H=\tfrac{1}{4}e^{-3\Omega}(\pi_\phi^2-\pi_\Omega^2)-e^{\Omega}+\Lambda e^{3\Omega}\approx 0 \tag{43}$$

is not separable in terms of the variables $x(\phi+\Omega)$, $y(\phi-\Omega)$; moreover, its potential has not a definite sign. However, it is easy to show that the time obtained for the case $k=0$ (flat model) is also a global phase time for the case $k=1$. Then consider the constraint

$$H_0=\tfrac{1}{4}e^{-3\Omega}(\pi_\phi^2-\pi_\Omega^2)+\Lambda e^{3\Omega}\approx 0 \tag{44}$$

which is equivalent to

$$H_0'=\pi_\phi^2-\pi_\Omega^2+4\Lambda e^{6\Omega}\approx 0.$$

By choosing $y=-(1/3)e^{3(\Omega-\phi)}$, $x=(1/3)e^{3(\Omega+\phi)}$ the same procedure used in the preceding example gives the extrinsic time

$$t=-2/3\,\frac{\Lambda e^{6\Omega}}{\pi_\phi+\pi_\Omega}\approx\frac{1}{6}(\pi_\phi-\pi_\Omega). \tag{45}$$

Note that if we want to verify that this function is a global phase time also for the case $k=1$ we should not write it as $\tfrac{1}{6}(\pi_\phi-\pi_\Omega)$ because the last equality holds only on the surface $H_0\approx 0$. If we calculate the Poisson bracket of $t=-2/3[\Lambda e^{6\Omega}/(\pi_\phi+\pi_\Omega)]$ with the constraint $H'=4e^{3\Omega}H$ we obtain $[t,H']=[t,H_0']+[t,-4e^{4\Omega}]$, which, as it is easy to check, is the sum of two positive terms. As the constraints H and H' are equivalent, then we have

$$[t,H]>0,$$

and t is a global phase time also for the model given by (43).

C. Geometry of the constraint surface

Our deparametrization procedure gives a simple way to examine how the geometrical properties of the constraint surface impose restrictions on the definition of a global phase time. Consider the Hamiltonian constraint of the most general case of a FRW empty cosmological model,

$$H=-\tfrac{1}{4}e^{-3\Omega}\pi_\Omega^2-ke^{\Omega}+\Lambda e^{3\Omega}\approx 0 \tag{46}$$

with $k=\pm 1$ and $\Lambda>0$. For this model the authors of Ref. 8 found an extrinsic time

$$t\sim -e^{-2\Omega}\pi_\Omega \tag{47}$$

after performing a canonical transformation on the ideal clock and multiplying the constraint by a positive definite function of the form $\sim e^{\Omega}$. Then we shall apply our procedure to the constraint $H'=e^{-\Omega}H$,

$$H'=-\tfrac{1}{4}e^{-4\Omega}\pi_\Omega^2-k+\Lambda e^{2\Omega}\approx 0. \tag{48}$$

The constraints H and H' are equivalent because they differ only in a positive definite factor. The τ-independent H–J equation for the Hamiltonian H' is

$$-\left(\frac{\partial W}{\partial \Omega}\right)^2 - 4ke^{4\Omega} + 4\Lambda e^{6\Omega} = 4e^{4\Omega}E \tag{49}$$

and matching $E=\overline{P_0}$ we obtain the solution

$$W(\Omega,\overline{P_0}) = \pm \int 2e^{2\Omega}\sqrt{\Lambda e^{2\Omega} - k - \overline{P_0}}\, d\Omega, \tag{50}$$

with + for $\pi_\Omega > 0$ and − for $\pi_\Omega < 0$. According to Eq. (6), on the constraint surface we have

$$Q^0 = \overline{Q}^0 = \left[\frac{\partial W}{\partial \overline{P_0}}\right]_{\overline{P_0}=0} = \mp \Lambda^{-1}\sqrt{\Lambda e^{2\Omega} - k}. \tag{51}$$

If we fix the gauge by means of the canonical condition $\chi \equiv Q^0 - T(\tau) = 0$ with T a monotonic function of τ then we have that

$$t = \theta(-\pi_\Omega)\Lambda^{-1}\sqrt{\Lambda e^{2\Omega} - k} - \theta(\pi_\Omega)\Lambda^{-1}\sqrt{\Lambda e^{2\Omega} - k} \tag{52}$$

is a global phase time for the system. As on the constraint surface we have

$$\pi_\Omega = \pm 2e^{2\Omega}\sqrt{\Lambda e^{2\Omega} - k} \tag{53}$$

[so that in the case $k=1$ the natural size of the configuration space is given by $\Omega \geqslant -\ln(\sqrt{\Lambda})$ (Ref. 2)] then we can write

$$t(\Omega,\pi_\Omega) = -\tfrac{1}{2}\Lambda^{-1}e^{-2\Omega}\pi_\Omega, \tag{54}$$

which is in agreement with (47). Now an important difference between the cases $k=-1$ and $k=1$ arises; for $k=-1$ the potential has a definite sign, and the constraint surface splits into two disjoint sheets given by (53). In this case the evolution can be parametrized by a function of the coordinate Ω only, the choice given by the sheet on which the system remains, and we then say that it has an intrinsic time; if the system is on the sheet $\pi_\Omega > 0$ the time is $t = -\Lambda^{-1}\sqrt{\Lambda e^{2\Omega} - k}$, and if it is on the sheet $\pi_\Omega < 0$ we have $t = \Lambda^{-1}\sqrt{\Lambda e^{2\Omega} - k}$. For $k=1$, instead, the potential can be zero and the topology of the constraint surface is no more analogous to that of two disjoint planes. Although for $\Omega = -\ln(\sqrt{\Lambda})$ we have $v(\Omega)=0$ and $\pi_\Omega = 0$, it is easy to verify that $\dot{\pi}_\Omega \neq 0$ at this point. Hence, in this case the coordinate Ω does not suffice to parametrize the evolution, because the system can go from (Ω,π_Ω) to $(\Omega,-\pi_\Omega)$; therefore we must necessarily define a global phase time as a function of the coordinate and the momentum (extrinsic time); $t=t(\Omega,\pi_\Omega)$. This, of course, generalizes to the case of models with true degrees of freedom.

A remark should be made, and it is that we have multiplied H by different positive functions to make calculations simpler, or to obtain times that we could compare with previous results; different rescalings of the Hamiltonian constraint would lead to different times, but—at least at the classical level—they would be equivalent.

D. Path integral quantization

Suppose that we want to quantize a cosmological model described by (q^i,p_i) by means of a path integral in terms of the variables (Q^i,P_i) given by (33). As we showed in a previous paper,[4] this has practical advantages, for example, when trying to avoid the Gribov problem. If we pretend the quantum amplitude $\langle Q_2^i | Q_1^i \rangle$ to be equivalent to $\langle q_2^i | q_1^i \rangle$ we should verify that the paths in the integral are weighted by the action $\mathcal{S}$ in the same way that they are weighted by S, and that the quantum states $|Q^i\rangle$ are equivalent to $|q^i\rangle$. As the path integral in the variables (Q^i,P_i) is gauge invariant, this requirement is fulfilled if it is possible to impose a—globally good—gauge condition $\overline{\chi}=0$ such that $\tau=\tau(q^i)$ is defined, and such that the boundary terms in (9) vanish. This is the reason why we chose the generating function for $\overline{X^i}\to X^i$ as in (32); with this choice the boundary terms in (9) vanish if we fix the gauge by means of $\chi\equiv\eta PQ^0-T(\tau)=0;$ when written in terms of the original variables, this gauge condition involves only the coordinates q^i, and is associated to the identification of an intrinsic time.

An intrinsic time, however, can be defined only if the constraint surface splits into two disjoint sheets, that is, if the potential has a definite sign. In the most general case the definition of a global phase time must necessarily involve also the momenta, and then we cannot fix the gauge in the path integral in such a way that $\tau=\tau(q^i)$ [see the last example, where $t=-1/2\Lambda^{-1}e^{-2\Omega}\pi_\Omega \approx T(\tau)$, so that $\tau=\tau(\Omega,\pi_\Omega)$]. Hence, if we want to quantize the system by imposing canonical gauges in the path integral, in the most general case of a potential with a nondefinite sign we must admit the possibility of identifying the quantum states in the original phase space not by q^i but by a complete set of functions of the coordinates and momenta q^i and p_i.

V. CONCLUSIONS

Although gauge fixation and the identification of a global phase time are closely related, as the action of parametrized systems—like the gravitational field—is not gauge invariant at the boundaries, we could not, in principle, use this fact to obtain a direct procedure to deparametrize minisuperspaces; while ordinary gauge systems admit gauge conditions of the type $\chi(q^i,p_i,\tau)=0$, only derivative gauges would be admissible for parametrized systems. Then we would not be able to identify a time for a cosmological model as a function of its canonical variables by imposing on the system a gauge condition which is compatible with the symmetries of the action.

However, if the H–J equation is separable, a parametrized system described by (q^i,p_i) can be turned into an ordinary gauge system described by (Q^i,P_i) by matching H with P_0, and canonical gauges are therefore admissible. Then we are able to identify a global phase time for cosmological models in terms of their coordinates and momenta by imposing τ-dependent canonical gauge conditions on the ordinary gauge system. We have illustrated our procedure with simple models whose H–J equation is easily solvable. We have been able to show that sometimes a global phase time for a quite trivial model is also a good time for a more physical system (example 2); however, we believe that this is not the best way to proceed, because in a general case it would only work if we impose restrictions on the parameters of the model (as it happens when we consider a massive scalar field, when a relation between Λ and m should exist). Of course, a complete solution of the H–J equation is, in general, difficult to obtain; an example of a more interesting model to be studied could be the Bianchi type-IX universe, which is the anisotropic generalization of the closed FRW model, and whose H–J equation is solvable.[12]

Our procedure clearly shows the restrictions arising from the geometry of the constraint surface; a global phase time in terms of the coordinates q^i can be defined only if the potential of the model has a definite sign; in this case, the choice is determined by the sheet of the constraint surface on which the system evolves. In the most general case, a global phase time must be a function of the coordinates and the momenta; at the quantum level, our method completes the analysis of Ref. 3, clearly showing the relation existing between the geometrical properties of the constraint surface and the possibility of identifying the quantum states in the path integral by means of only the original coordinates.

[1] A. O. Barvinsky, Phys. Rep. **230**, 237 (1993).
[2] P. Hájícek, Phys. Rev. D **34**, 1040 (1986).
[3] R. Ferraro and C. Simeone, J. Math. Phys. **38**, 599 (1997).
[4] C. Simeone, J. Math. Phys. **39**, 3131 (1998).
[5] C. Teitelboim, Phys. Rev. D **25**, 3159 (1982).
[6] J. J. Halliwell, Phys. Rev. D **38**, 2468 (1988).
[7] M. Henneaux, C. Teitelboim, and J. D. Vergara, Nucl. Phys. B **387**, 391 (1992).
[8] S. C. Beluardi and R. Ferraro, Phys. Rev. D **52**, 1963 (1995).

[9] K. V. Kuchař, in *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics*, edited by G. Kunstatter, D. Vincent, and J. Williams (World Scientific, Singapore, 1992).
[10] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, New Jersey, 1992).
[11] J. J. Halliwell, in *Introductory Lectures on Quantum Cosmology*, in *Proceedings of the Jerusalem Winter School on Quantum Cosmology and Baby Universes*, edited by T. Piran (World Scientific, Singapore, 1990).
[12] V. Moncrief and M. P. Ryan, Jr., Phys. Rev. D **44**, 2375 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Exact solutions to sourceless charged massive scalar field equation on Kerr–Newman background

S. Q. Wu[a)] and X. Cai[b)]
Institute of Particle Physics, Hua-Zhong Normal University, Wuhan 430079, People's Republic of China

(Received 18 February 1999; accepted for publication 30 April 1999)

The covariant Klein–Gordon equation in the Kerr–Newman black hole geometry is separated into a radial part and an angular part. It is discovered that in the nonextreme case, these two equations belong to a generalized spin-weighted spheroidal wave equation. Then general exact solutions in integral forms and several special solutions with physical interest are given. While in the extreme case, the radial equation can be transformed into a generalized Whittaker–Hill equation. In both cases, five-term recurrence relations between coefficients in power series expansion of general solutions are presented. Finally, the connection between the radial equations in both cases is discussed.
[S0022-2488(99)01608-4]

I. INTRODUCTION

Since the Hawking effect[1] on a black hole was found, the evaporation of a black hole has been investigated in several coordinates by miscellaneous methods such as the path integral approach,[2] tortoise coordinate (r_*) transformation,[3,4] and r_*-coordinate analytical extension,[4] etc. Among these methods, the generalized tortoise transformation method has been used widely in the discussions, not only on the evaporation of a static black hole and a stationary black hole, but also on that of nonstationary ones.[3] Much more progress has been made. But this method cannot give an exact solution of the radial (r_*) equation, the radial wave function, which can be analyzed only in an asymptotic expression.

Couch[5] obtained a series of exact solutions by transforming the separated radial equation into a modified Whittaker–Hill equation under some special conditions. But these solutions seem to have nothing to do with the discussion on black hole evaporation.

Solutions to the generalized spheroidal wave equation have been studied to some extent[6,7] by using power series expansions around regular singular points, so that three-term recurrence relations between coefficients can be manipulated in terms of the continued fraction method. Leaver[6] has shown that Teukolsky master equations in Kerr geometry are, in fact, spin-weighted generalized spheroidal wave equations.

It appears to be more important to obtain an exact solution to the radial equation, for this is crucial in discussing the Hawking effect of a black hole. However, it is very difficult to do so. It is this motivation that stimulates our present research. Our main aim in this paper is to show that the separated radial part of a massive covariant Klein–Gordon equation on the Kerr–Newman black hole (KNBH) background is a generalized spin-weighted spheroidal wave equation of imaginary number order. In this paper, we shall discuss the solutions to a massive complex scalar field in the KNBH geometry with three parameters. In the nonextreme case, its general solutions of the separated parts are spin-weighted generalized spheroidal wave functions,[6–8] and some special solutions to the radial equation with physical interest are given. General solutions to the radial equation in the extreme case shall be briefly discussed. Finally, we show that the radial

[a)]Electronic mail: emu@iopp.ccnu.edu.cn
[b)]Electronic mail: xcai@wuhan.cngb.com

equation in the extreme case is a confluent equation of that in the nonextreme case.

In Sec. II we deal with the variable separation of a sourceless complex scalar field on KNBH and solutions to the angular part. In Sec. III and IV, the radial equation is solved in both nonextreme and extreme cases, respectively. In Sec. III A we reduce the radial equation to standard form, and in Secs. III B and III C we obtain general solutions and special ones including case $(\omega=\mu=0)$, respectively. Five-term recurrences between coefficients of solutions in power series forms are given in both cases. In addition, we give solutions in integral forms and some special solutions of physical interest in the nonextreme case. Conditions for general solutions exist and are given in these two cases. Section V is devoted to discussing the connection between the radial equation in the extreme case and that in the nonextreme case. Finally, we point out some probable applications and generalization of exact solutions in Sec. VI.

In the Appendix, three-term recurrence relations between coefficients in power series expansions around regular singular points for a generalized spheroidal wave equation are presented.

II. SEPARATION OF KLEIN–GORDON EQUATION AND SOLUTION TO THE ANGULAR EQUATION

The Kerr–Newman line element and electromagnetic one-form are given in the Boyer-Lindquist coordinates as follows:[4,9]

$$\begin{aligned} ds^2 &= g_{\mu\nu}\,dx^\mu\,dx^\nu \\ &= -\frac{\Delta}{\Sigma}(dt-a\sin^2\theta\,d\varphi)^2+\frac{\sin^2\theta}{\Sigma}[a\,dt-(r^2+a^2)d\varphi]^2+\Sigma\left(\frac{dr^2}{\Delta}+d\theta^2\right), \end{aligned} \tag{1}$$

$$\mathcal{A}=A_\mu\,dx^\mu=\frac{-er}{\Sigma}(dt-a\sin^2\theta\,d\varphi), \tag{2}$$

with event horizon function $\Delta=r^2-2Mr+a^2+e^2$, and $\Sigma=r^2+a^2\cos^2\theta$, where mass M, charge e, specific angular momentum $a=J/M$ being three parameters to describe KNBH. (Use Planck units system $G=\hbar=c=1$, and denote $\partial_\mu=\partial/\partial x^\mu$).

The determinant of the KNBH metric tensor is $g=\det(g_{\mu\nu})=-\Sigma^2\sin^2\theta$, while the electromagnetic four-vector potential A_μ apparently satisfies the following covariant Lorentz gauge condition:

$$\nabla_\mu A^\mu=\frac{1}{\sqrt{-g}}\partial_\mu(\sqrt{-g}g^{\mu\nu}A_\nu)=0. \tag{3}$$

In curved spacetime, a sourceless scalar field Φ with mass μ and charge q obeys the covariant Klein–Gordon equation (KGE):

$$(\Box_c-\mu^2)\Phi=0, \tag{4}$$

where the d'Alembert operator $\Box_c$ on the KNBH background is given by

$$\begin{aligned} \Box_c &\equiv \frac{1}{\sqrt{-g}}D_\mu(\sqrt{-g}g^{\mu\nu}D_\nu) \\ &= \frac{1}{\Sigma}\left\{\frac{-1}{\Delta}[(r^2+a^2)\partial_t+a\partial_\varphi+iqer]^2+\partial_r(\Delta\partial_r)\right. \\ &\quad \left.+\left(a\sin\theta\partial_t+\frac{1}{\sin\theta}\partial_\varphi\right)^2+\frac{1}{\sin\theta}\partial_\theta(\sin\theta\partial_\theta)\right\}, \end{aligned} \tag{5}$$

the covariant gauge differential operator here being $D_\mu=\partial_\mu-iqA_\mu$.

The scalar wave function Φ for the KGE of Eq. (4) has a solution of variables separable form $\Phi(t,r,\theta,\varphi)=R(r)S(\theta)e^{i(m\varphi-\omega t)}$:[9]

$$\frac{1}{\Delta}[\omega(r^2+a^2)-qer-ma]^2\Phi+\partial_r(\Delta\,\partial_r\Phi)-\mu^2\Sigma\Phi -\left(a\omega\sin\theta-\frac{m}{\sin\theta}\right)^2\Phi+\frac{1}{\sin\theta}\partial_\theta(\sin\theta\,\partial_\theta\Phi)=0. \tag{6}$$

The separated results of the above equation are

$$\partial_r[\Delta\,\partial_r R(r)]+\left\{\frac{[\omega(r^2+a^2)-qer-ma]^2}{\Delta}-\mu^2(r^2+a^2)-\lambda+2ma\omega\right\}R(r)=0 \tag{7}$$

and

$$\frac{1}{\sin\theta}\partial_\theta[\sin\theta\,\partial_\theta S(\theta)]+\left[\lambda-\frac{m^2}{\sin^2\theta}+(\mu^2-\omega^2)a^2\sin^2\theta\right]S(\theta)=0, \tag{8}$$

where λ is a separation constant.

The general solutions to the angular part are ordinary spheroidal angular wave functions[8,10] with spin weight $s=0$. When $a^2(\omega^2-\mu^2)=0$, these solutions degenerate to Legendre spherical functions.

Let $x=\cos\theta$, $S(\theta)=S(x)=(1-x^2)^{m/2}\Theta(x)$, Eq. (8) should take the following forms:

$$(1-x^2)S''(x)-2xS'(x)+\left[\lambda-\frac{m^2}{1-x^2}+a^2(\omega^2-\mu^2)(x^2-1)\right]S(x)=0 \tag{9}$$

and

$$(1-x^2)\Theta''(x)-2(1+m)x\Theta'(x)+[\lambda-m(m+1)+a^2(\omega^2-\mu^2)(x^2-1)]\Theta(x)=0. \tag{10}$$

Here and after, $S'(x)=\partial S(x)/\partial x$, etc.

The eigenfunctions to Eqs. (9) and (10) are generalized spheroidal wave functions[6–8] $S(x)=S_l^{m,0}(c,x)$ with eigenvalue $\lambda=\lambda_{ml}+c^2$, $c^2=a^2(\mu^2-\omega^2)$. When $\mu=0$, Eqs. (9) and (10) are special cases ($s=0$) of the following spin-weighted spheroidal wave equations:[6–8,10]

$$(1-x^2)P''(x)-2xP'(x)+\left[a^2\omega^2x^2-2a\omega sx-\frac{(m+sx)^2}{1-x^2}+s+\lambda'\right]P(x)=0 \tag{11}$$

and

$$(1-x^2)Q''(x)-2[s+(1+m)x]Q'(x) +[\lambda'-(m-s)(m+s+1)-2a\omega sx+a^2\omega^2x^2]Q(x)=0, \tag{12}$$

where $P(x)=(1-x)^{|m+s|/2}(1+x)^{|m-s|/2}Q(x)$ and $x=\cos\theta$.

When $a\omega=0$, the solutions to the above equations are Jacobi ultrasphere D functions[10] $D_{m,s}^l(x)$ or spin-weighted spherical harmonic functions[11] with eigenvalue $\lambda'=l(l+1)-s(s+1)$, $l=\max(|m|,|s|)$. In the general case, the solutions should be the generalized spin-weighted spheroidal wave functions[6–8] $P(x)=P_{m,s}^l(c,x)$, $c^2=-a^2\omega^2$. By taking account of some reasonable boundary conditions, these solutions could be a set of orthogonal polynomials.

In the following, we shall assume that all parameters, M,e,a,μ,q,m, are nonzero, and discuss the radial equation of Eq. (7) according to two cases, namely, the nonextreme case $(M^2\neq a^2+e^2)$ and the extreme case $(M^2=a^2+e^2)$. The special case $(\omega=\mu=0)$ will be included in Sec. III C.

III. SOLUTIONS TO THE RADIAL EQUATION IN THE NONEXTREME CASE $(M^2\neq a^2+e^2)$

A. Simplification of the radial equation in the case $(\epsilon\neq 0)$

In this case, we put $\epsilon=\sqrt{M^2-a^2-e^2}$ $(0<\epsilon<M)$. After making substitutions of $r=M+\epsilon z$ and $R(r)=R(z)=(z-1)^{i|B+A|/2}(z+1)^{i|B-A|/2}F(z)$, the exterior horizon and interior horizon are located at points $r_\pm=M\pm\epsilon$, $(z=\pm 1)$, respectively, the radial equation of Eq. (7) can be reduced to the following standard forms:

$$(z^2-1)R''+2zR'+\left[\epsilon^2(\omega^2-\mu^2)(z^2-1)+2\epsilon(A\omega-M\mu^2)z+\frac{(Az+B)^2}{z^2-1}\right.$$
$$\left.+(2\omega^2-\mu^2)(2M^2-e^2)-2qeM\omega-\lambda\right]R=0 \quad (1<z<\infty) \tag{13}$$

and

$$(z^2-1)F''+2[iA+(1+iB)z]F'+[\epsilon^2(\omega^2-\mu^2)(z^2-1)+2\epsilon(A\omega-M\mu^2)z$$
$$+(2\omega^2-\mu^2)(2M^2-e^2)-2qeM\omega-\lambda+A^2-B^2+iB]F=0 \quad (1<z<\infty), \tag{14}$$

where $A=2M\omega-qe$, $\epsilon B=\omega(2M^2-e^2)-qeM-ma$.

In order to study behaviors of solutions to Eqs. (13) and (14) in the interval $(-1<z<1)$, we rotate first T from the real axis to the imaginary axis $T=i\tau$ after making substitution of $z=\cosh T=\cosh(i\tau)=\cos\tau$, then return to the real z axis $z=\cos\tau$. Therefore, Eqs. (13) and (14) have corresponding forms in the interval $(|z|<1)$ as follows:

$$(1-z^2)R''-2zR'+\left[\epsilon^2(\omega^2-\mu^2)(z^2-1)+2\epsilon(A\omega-M\mu^2)z-\frac{(Az+B)^2}{1-z^2}\right.$$
$$\left.+(2\omega^2-\mu^2)(2M^2-e^2)-2qeM\omega-\lambda\right]R=0 \quad (-1<z<1) \tag{15}$$

and

$$(1-z^2)G''-2[A+(1+B)z]G'+[\epsilon^2(\omega^2-\mu^2)(z^2-1)+2\epsilon(A\omega-M\mu^2)z$$
$$+(2\omega^2-\mu^2)(2M^2-e^2)-2qeM\omega-\lambda+A^2-B^2-B]G=0 \quad (|z|<1), \tag{16}$$

where we have made a function transformation,

$$R(z)=(1-z)^{|B+A|/2}(1+z)^{|B-A|/2}G(z).$$

The essence of our manipulation is that we extend the domain of z from the real axis to the complex z plane $z=x+iy$ at first and then make an analytical extension on the complex z plane from the region outside the unit circle $(|z|>1)$ to that inside it $(|z|<1)$. The handled issue is that only derivative terms change to a negative sign, while the nonderivative terms, namely, terms in square brackets, make no change in symbol. This method is equivalent to that Eqs. (13)–(16) are solved initially and then the solutions are made an analytical extension on the complex z plane.

Both Eqs. (13) and (14) are the generalized spin-weighted spheroidal wave equation[6–8] with an imaginary number order, while both Eqs. (15) and (16) with a real number order. The formers are suitable, especially to study problems about the scattering state, whereas the latters are more convenient to investigate energy levels of bound states. Furthermore, we can apparently find out the connection between poles of scattering amplitudes and energy levels of bound states. Actually, the domain in which z takes values in Eqs. (13) and (14) is on the x axis, while those in Eqs. (15) and (16) on the y axis. So these equations can be thought of as equivalence. However, to be convenient, we have made a restriction on intervals that z takes values of $1<z<\infty$ in Eqs. (13) and (14), while of $|z|<1$ in Eqs. (15) and (16).

When $\mu=0$ or $M=0$, if taking $R_1(z_1)$ as the first solution to Eq. (15) in the interval of $|z|<1$, then $R_2(z_2)$, the second one to the same equation in that of $|z|>1$, might be

$$R_2(z_2)=(z_2-1)^{|B+A|/2}(z_2+1)^{|B-A|/2}\int_{-1}^{+1}e^{i\epsilon\omega z_2 z_1}(1-z_1)^{|B+A|/2}(1+z_1)^{|B-A|/2}R(z_1)dz_1. \tag{17}$$

Here, we have assumed that $\omega>0$. The integral equation of Eq. (17) connects the irregular solution $R_2(z_2)$ with the regular solution $R_1(z_1)$.

Comparing Eqs. (9)–(12) with Eqs. (13)–(16), especially Eqs. (11) and (12) with Eqs. (15) and (16), we can draw a conclusion that the separated angular and radial equations are ordinary differential equations of the same type, generalized spin-weighted spheroidal wave equations.[6–8] Furthermore, we discover that ϵ, A, B correspond to $-a$, $-s$, $-m$, respectively, when $\mu=0$. There may exist three pairs of power series solutions to generalized spheroidal wave equations around singular points $z=\pm1,\infty$, respectively. Added with some proper boundary conditions, these power series expansions of spheroidal wave functions can be cut off to be polynomials.

Therefore, in the following section, we shall only study the generalized spheroidal wave equation. The reader who has more interest in this equation can find more information in Refs. 6 and 7 (and references cited therein).

B. General solutions to the radial equation ($\epsilon\neq 0$)

The standard generalized spin-weighted spheroidal wave equation that we reduce to study is as follows:

$$(1-z^2)W''(z)-2[\alpha+(\beta+1)z]W'(z)+[\gamma^2(z^2-1)+2\,\delta z+\bar{\lambda}-\beta]W(z)=0, \tag{18}$$

where $\bar{\lambda}$ is a redefined eigenvalue that could make $W(z)$ finite at $z=\pm1$, and the region of z' taking values could be the whole complex z plane.

(i) For the radial equation of Eq. (16), we have

$$\alpha=A, \quad \beta=B, \quad \gamma^2=\epsilon^2(\omega^2-\mu^2), \quad \delta=\epsilon(A\omega-M\mu^2).$$

(ii) For the angular equation of Eq. (12), we have

$$\alpha=s, \quad \beta=m, \quad \gamma^2=a^2\omega^2, \quad \delta=-as\omega.$$

(iii) For the angular equation of Eq. (10), we have

$$\alpha=0, \quad \beta=m, \quad \gamma^2=a^2(\omega^2-\mu^2), \quad \delta=0.$$

The form of Eq. (18) is invariant both under the Laplace transformation and by changing parameters $\alpha,\beta,\gamma^2,\delta,z$ into $-\alpha,\beta,\gamma^2,-\delta,-z$, respectively. Namely, $W(z)=W(\alpha,\beta,\gamma,\delta;z)$ satisfies the following integral equation:

$$\int_0^{+\infty}e^{-tz}W(\alpha,\beta,\gamma,\delta;z)dz=W\left(\frac{\delta}{\gamma},-\beta,\gamma,-\alpha\gamma;\frac{t}{\gamma}\right)=W\left(\frac{-\delta}{\gamma},-\beta,\gamma,\alpha\gamma;\frac{-t}{\gamma}\right), \tag{19}$$

$$W(\alpha,\beta,\gamma,\delta;z)=W(-\alpha,\beta,\gamma,-\delta;-z)=\int_0^{+\infty}e^{-\gamma zt}W\left(\frac{-\delta}{\gamma},-\beta,\gamma,\alpha\gamma;t\right)dt. \tag{20}$$

The above formulas are integral solutions to Eq. (18). If one knows a solution, then he can obtain another by integral transformations of Eqs. (19) and (20). It is obvious that solutions are symmetry under the following condition:

$$\alpha=\delta=0 \quad (\gamma\neq 0).$$

This just is the case (iii). At this moment, the symmetric solutions are ordinary spheroidal angular wave functions.[10]

Now, we consider a solution to the generalized spheroidal equation of Eq. (18), which is in power series form in the interval of $-1<z<1$. According to the knowledge of an ordinary differential equation, one can know that Eq. (18) has two regular singularities $(z=\pm 1)$ and one confluently irregular singular point $(z=\infty)$. As $z=0$ is its ordinary point, we can make a Taylor expansion of $W(z)$ in the vicinity of ordinary point $(z=0)$:

$$W(z)=W_n(z)=\sum_{n=0}^{\infty}a_nz^n \quad (|z|<1). \tag{21}$$

Substituting the power series of Eq. (21) into Eq. (18), we obtain five-term recurrence relations between coefficients as follows:

$$(\bar{\lambda}-\gamma^2-\beta)a_0-2\alpha a_1+2a_2=0,$$

$$2\delta a_0+[\bar{\lambda}-\gamma^2-\beta-(2+2\beta)]a_1-4\alpha a_2+a_3=0,$$

$$\gamma^2a_0+2\delta a_1+[\bar{\lambda}-\gamma^2-\beta-2(3+2\beta)]a_2-6\alpha a_3+12a_4=0,$$

$$\cdots\cdots$$

$$\gamma^2a_{n-2}+2\delta a_{n-1}+[\bar{\lambda}-\gamma^2-\beta-n(n+1+2\beta)]a_n-2(n+1)\alpha a_{n+1}+(n+2)(n+1)a_{n+2}=0.$$

Redefine coefficients:

$$A_n=\frac{\gamma^2}{\bar{\lambda}-\gamma^2-\beta-n(n+1+2\beta)},$$

$$B_n=\frac{2\delta}{\bar{\lambda}-\gamma^2-\beta-n(n+1+2\beta)},$$

$$C_n=\frac{-2(n+1)\alpha}{\bar{\lambda}-\gamma^2-\beta-n(n+1+2\beta)},$$

$$D_n=\frac{(n+2)(n+1)}{\bar{\lambda}-\gamma^2-\beta-n(n+1+2\beta)}.$$

When taking limits $n\to\infty$, we have A_n, B_n, $C_n\to 0$, and $D_n\to -1$.

Then, five-term recurrence relations become

$$A_na_{n-2}+B_na_{n-1}+a_n+C_na_{n+1}+D_na_{n+2}=0. \tag{22}$$

After arranging coefficients A_n, B_n, C_n, D_n and making up them into a quasidiagonal band matrix Λ and $a_0, a_1, \ldots, a_n, \ldots,$ into a column vector $\vec{a} = (a_0, a_1, \ldots, a_n, \ldots)$, the above recurrence relations become an infinite tridiagonal matrix equation:

$$\Lambda \vec{a} = h \vec{a}. \tag{23}$$

The condition for solutions of Eq. (23) exist is that determinant of Λ is zero,

$$\det(\Lambda) = \begin{vmatrix} 1 & C_0 & 0 & 0 & 0 & 0 & 0 & 0\cdots \\ B_1 & 1 & C_1 & D_1 & 0 & 0 & 0 & 0\cdots \\ A_2 & B_2 & 1 & C_2 & D_2 & 0 & 0 & 0\cdots \\ 0 & A_3 & B_3 & 1 & C_3 & D_3 & 0 & 0\cdots \\ & & & \cdots & \cdots & & & \\ 0 & 0 & A_n & B_n & 1 & C_n & D_n & 0\cdots \\ & & & \cdots & \cdots & & & \end{vmatrix} = 0. \tag{24}$$

In fact, this condition could be satisfied, and we have $\det(\Lambda) \to 0$ when $n \to \infty$.

Matrix equation of Eq. (23), together with the determinant equation of Eq. (24) determines coefficients $a_0, a_1, \ldots, a_n, \ldots,$ and eigenvalue $\bar{\lambda}$, hence, eigenvalue $\bar{\lambda}$ will be a complicated function of $\alpha, \beta, \gamma, \delta$, as well as n. The second power series solution around the same point $z = 0$ can be obtained by Frobenius' method. To be finite at $z = \pm 1$, power series $W(z)$ could be truncated to be polynomial, and $\alpha, \beta, \gamma, \delta$ could be integers or half-integers. While in the general case, solutions to spin-weighted generalized spheroidal equation of Eq. (18) are transcendental functions.[6,7]

Absolutely, solution $W_n(z) = W_n(\alpha, \beta, \gamma, \delta; z)$ of Eq. (18) can be orthonormalized to constitute a set of complete functions:

$$\int_{-1}^{1} (1-z)^{\beta+\alpha} (1+z)^{\beta-\alpha} W_n(z) W_{n'}(z) dz = \delta_{n,n'}. \tag{25}$$

Solutions $W_n(z)$ at infinity can have asymptotic forms $W_n(z) \to e^{\mp \gamma z}$, $(z \to \pm \infty,\ \gamma > 0)$. This is consistent with that the Minkowski spacetime is an asymptotic spacetime of the Kerr–Newmann black hole. Thus, an ingoing wave and an outgoing wave at infinity can take the form of plane waves.

C. Special solutions to the radial equation in the case ($\epsilon \neq 0$)

In this section, we will base our discussion upon Eq. (15), namely,

$$(1-z^2) R''(z) - 2z R'(z) + \left[\gamma^2 (z^2 - 1) + 2\delta z - \frac{(\beta + \alpha z)^2}{z^2 - 1} \right.$$
$$\left. + (2\omega^2 - \mu^2)(2M^2 - e^2) - 2qeM\omega - \lambda \right] R(z) = 0 \quad (|z| < 1), \tag{26}$$

where

$$\gamma^2 = \epsilon^2 (\omega^2 - \mu^2), \quad \delta = \epsilon (A\omega - M\mu^2),$$

$$\alpha = A = 2M\omega - qe, \quad \beta = B = \frac{\omega(2M^2 - e^2) - qeM - ma}{\epsilon}.$$

Case 1: when $\gamma=\delta=0$, there exist three situations:
(i) $\omega=\pm\mu=qe/M\neq 0$ $(\alpha\neq 0)$; (ii) $\omega=\mu=qe/M=0$ $(\alpha=0)$ (This case can be thought of as a special one in the above-head case); (iii) $\omega=\mu=0$, $qeM\neq 0$ $(\alpha\neq 0)$.

Solutions in situations (i) and (iii) are Jacobi ultrasphere functions $R(z)=P_n^{(\beta+\alpha,\beta-\alpha)}(z)$,[10] whereas solutions in situation (ii) degenerate to be Legendre functions, $R(z)=P_n^{\beta}(z)$, or $Q_n^{\beta}(z)$.

Case 2: when $M\mu^2=0$, $\omega\neq 0$, $\delta/(\epsilon\omega)=\alpha$, this case has been considered in detail by Leaver.[6]

Case 3: when $\alpha=\delta=0$, $\gamma\neq 0$, Eqs. (26) is an ordinary spheroidal wave equation,[10] and its solutions are Prolate spheroidal angular wave functions $R(z)=S_n^{\beta,0}(\gamma,z)$.

Obviously, all these solutions are special cases of general solutions $R_n^{(\beta+\alpha,\beta-\alpha)}(\gamma,\delta;z)=(1-z)^{(\beta+\alpha)/2}(1+z)^{(\beta-\alpha)/2}W_n(\alpha,\beta,\gamma,\delta;z)$.

Solutions in Case 1 will be particularly important in physics to black hole evaporation, as a scattering cross section, stationary state energy levels, emission coefficients of black hole radiation, etc., could be analytically computed at an exact theoretical level by the use of Jacobi polynomials. Furthermore, there maybe exist special symmetry in such a case.

IV. SOLUTIONS TO THE RADIAL EQUATION IN THE EXTREME CASE $(M^2=a^2+e^2)$

In the extreme KNBH case $(\epsilon=0)$, we make the substitution $r=M(1+x)$; then the event horizon is located at a single point $(r_h=M)$, namely, $x=0$; hence the radial equation of Eq. (7) can be transformed into the following confluent equation:

$$x^2R''(x)+2xR'(x)+\Big[(\omega^2-\mu^2)M^2x^2+2(A\omega-M\mu^2)Mx + \left(A+\frac{\mathcal{B}}{x}\right)^2+(2\omega^2-\mu^2)(2M^2-e^2)-2qeM\omega-\lambda\Big]R(x)=0, \tag{27}$$

where $A=2M\omega-qe$, $M\mathcal{B}=B\epsilon=\omega(2M^2-e^2)-qeM-ma$.

Defining

$$C^2=M^2(\omega^2-\mu^2),\quad D=M(A\omega-M\mu^2),$$

$$\lambda_e=(2\omega^2-\mu^2)(2M^2-e^2)-2qeM\omega-\lambda-\tfrac{1}{4}+A^2,$$

and making substitutions

$$x=e^{i\nu\xi},\quad R(x)=R(\xi)=e^{-i\nu\xi/2}H(\xi);$$

then Eq. (27) is transformed into the generalized Whittaker–Hill equation (GWHE):

$$-\nu^{-2}H''(\xi)+[C^2e^{2i\nu\xi}+2De^{i\nu\xi}+2A\mathcal{B}e^{-i\nu\xi}+\mathcal{B}^2e^{-2i\nu\xi}+\lambda_e]H(\xi)=0. \tag{28}$$

Solutions of GWHE of Eq. (28) can be regarded formally as

$$H(\xi)=\sum_{n=-\infty}^{+\infty}g_ne^{in\nu\xi},\quad n=0,\pm 1,\pm 2,\ldots. \tag{29}$$

Substituting Eq. (29) into Eq. (28), we obtain five-term recurrence relations between coefficients:

$$C^2g_{n-2}+2Dg_{n-1}+(\lambda_e+n^2)g_n+2A\mathcal{B}g_{n+1}+\mathcal{B}^2g_{n+2}=0,$$

$$E\vec{g}=h_e\vec{g}=\sum_{n=-\infty}^{+\infty}\sum_{m=n-2}^{n+2}E_{m,n}g_n, \tag{30}$$

where we have recast recurrence relations in matrix form in Eq. (30), and defined matrix elements,

$$E_{n,n-2}=\frac{C^2}{\lambda_e+n^2},\quad E_{n,n-1}=\frac{2D}{\lambda_e+n^2},$$

$$E_{n,n}=1,\quad E_{n,n+1}=\frac{2A\mathcal{B}}{\lambda_e+n^2},\quad E_{n,n+2}=\frac{\mathcal{B}^2}{\lambda_e+n^2}.$$

The condition that solutions of simultaneous equations in Eq. (30) exist is that determinant det(E) must be zero, that is,

$$\det(E)=0, \tag{31}$$

$$\det|E-h_eI|=0. \tag{32}$$

The secular equation of Eq. (32) is a characteristic equation that determines the existence of periodic solutions of Eq. (29). Solutions $R(\xi)$ could be functions with period $4\pi/\xi$. There exist four series of periodic functions according to the period being odd or even. Equation (27) has two confluently irregular singular points $x=0,\infty$. Behaviors of its solutions at event horizon $r_h=0$ $(x=0)$ depend upon that of $R(\xi)$ at $\xi\rightarrow\pm i\infty$ (according to ν being a negative number or a positive number).

V. CONNECTION BETWEEN THE RADIAL EQUATION IN NONEXTREME CASE AND THAT IN EXTREME CASE

In this section, we illustrate that the radial equation of Eq. (27) in the extreme case is a confluent form of Eq. (13) in the nonextreme case, and give an expression to the first thermodynamic law in the extreme KNBH case.

After making substitutions of $\epsilon=M\varepsilon$, $\varepsilon z=x$, $\epsilon z=Mx$, $\mathcal{B}=\varepsilon B$, $(0<\varepsilon<1)$ in Eq. (13), we have

$$r=M+\epsilon z=M(1+x),\quad \Delta=\epsilon^2(z^2-1)=M^2(x^2-\varepsilon^2),$$

$$A=2M\omega-qe,\quad \epsilon B=M\mathcal{B}=\omega(2M^2-e^2)-qeM-ma.$$

Then Eq. (13) is equivalent to the following one in the nonextreme case:

$$\frac{\partial}{\partial x}\left[(x^2-\varepsilon^2)\frac{\partial R(x)}{\partial x}\right]+\left[M^2(\omega^2-\mu^2)(x^2-\varepsilon^2)+2M(A\omega-M\mu^2)x\right.$$
$$\left.+\frac{(Ax+\mathcal{B})^2}{x^2-\varepsilon^2}+(2\omega^2-\mu^2)(2M^2-e^2)-2qeM\omega-\lambda\right]R(x)=0. \tag{33}$$

Equation (33) has two regular singular points $x=\pm\varepsilon$ $(z=\pm1)$ that are located at the exterior horizon and the interior horizon (Cauchy surface) $r_\pm=M\pm\epsilon=M(1\pm\varepsilon)$, respectively, along with another irregular singular point $x=\infty$. After taking limits $\varepsilon\rightarrow0$, $x^2-\varepsilon^2\rightarrow x^2$, Eq. (33) in the nonextreme case tends to Eq. (27) in the extreme case. The latter has two confluently irregular singular points $x=0,\infty$. The irregular singular point $x=0$ that is located at event horizon $r_h=M$ in the extreme case is just one to which two irregular singular points $x=\varepsilon$ and $x=-\varepsilon$ in the nonextreme case concur when ε or $\epsilon\rightarrow0$.

In the extreme KNBH case $(M^2=a^2+e^2)$, surface gravity $\kappa_h=0$, event horizon $r_h=M$, reduced event horizon area $A_h=M^2+a^2=2M^2-e^2$, the first thermodynamic law of the extreme Kerr–Newman black hole is expressed as follows:

$$dM=\Omega_h\,dJ+\Phi_h de, \tag{34}$$

where $\Phi_h=(er_h)/A_h$, and $\Omega_h=a/A_h$ are the electric potential and angular velocity at the event horizon $(r_h=M)$, respectively.

VI. CONCLUSION

In this paper, a sourceless charged massive scalar Klein–Gordon field equation has been separated into the angular and radial parts. The separated equations are all generalized spin-weighted spheroidal wave equations. In the nonextreme case, we present general solutions in power series expansion and that of integral forms, as well as several special solutions with physical interest for the radial equation. These solutions can be orthonormalized to a set of complete functions. In addition, they have asymptotic behaviors of plane waves at infinity. On the base of these orthogonal functions or polynomials, we can expand the wave function of a complex scalar field to a quantized Klein–Gordon field on the Kerr–Newman background. In the extreme case, the radial equation can be reduced to a modified Whittaker–Hill equation. In both cases, we obtain five-term recurrence relations between coefficients in power series expansions.

At the base of this work, the quantum conservation laws about the Hawking process and the probable generalization to black hole thermodynamic laws can be discussed further. It is anticipated that the separated parts of the Dirac equation in the Kerr–Newman geometry could be reduced to the forms of a generalized spheroidal wave equation.

ACKNOWLEDGMENTS

S.Q.W. is grateful to W. Li for his helps on computer. This work is supported partly by the NNSF and the Hubei-NSF in China.

APPENDIX

In this appendix, we present three-term recurrence relations between coefficients in power series expansions around regular singular points $(z=\pm 1)$ for spin-weighted spheroidal wave equation of Eq. (18), namely,

$$(1-z^2)W_n''(z)-2[a+(b+1)z]W_n'(z)+[c^2(z^2-1)+2\,dz+\lambda_n-b]W_n=0 \quad (-1<z<1). \tag{A1}$$

Equation (A1) has two regular singular points $z=1$ and $z=-1$, with indices $\rho_-=0,-a-b$ and $\rho_+=0,a-b$, respectively. When $c,d\to 0$, $W_n(z)$ must tend to Jacobi polynomials.

Introducing a symbol $\epsilon=\mp 1$, we denote these two regular singular points $z=\pm 1=-\epsilon$. Then, we make power series expansions around regular singular points $z=-\epsilon$, respectively, where we have written them in a united manner:

$$W_n(z)=e^{-cz}\sum_{n=0}^{\infty} f_n(1+\epsilon z)^n. \tag{A2}$$

Substituting the above regular solutions of Eq. (A2) into Eq. (A1), we obtain three-term recurrence relations between coefficients as follows:

$$(1+b-\epsilon a)f_1+[\lambda_0+2ac-b-2\epsilon(bc+c+d)]f_0=0,$$

$$\cdots$$

$$(n+1)(n+1+b-\epsilon a)f_{n+1}+[\lambda_n+2ac-b-2\epsilon(bc+c+d)$$
$$-n(n+1+2b+4\epsilon c)]f_n+4\epsilon(nc+bc+d)f_{n-1}=0.$$

After defining coefficients,

$$A_n=\lambda_n+2ac-b-2\epsilon(bc+c+d)-n(n+1+2b+4\epsilon c),$$

$$B_n=(n+1)(n+1+b-\epsilon a),$$

$$C_n=4\epsilon[(n+b)c+d],$$

recurrence relations for the first term and the nth term can be written as

$$B_0 f_1+A_0 f_0=0,$$

$$B_n f_{n+1}+A_n f_n+C_n f_{n-1}=0. \tag{A3}$$

Three-term recurrence relations of Eq. (A3) can be handled by the continued fraction method,[6,7] or by the matrix method (see Liu's paper in Ref. 7) as we can array A_n, B_n, C_n to make up a generalized Jacobi tridiagonal band matrix. Similar three-term recurrence relations can also be obtained by expansions in the light of Jacobi polynomials, but the coefficients A_n, B_n, C_n will be more complicated than those presented here.

The second regular solutions around the same points can be easily obtained by Frobenius' method, and we have not presented them here. Irregular solutions are connected with these regular ones by integrals similar to those in Eqs. (19) and (20).

In order to make $W_n(z)$ finite at $z=\pm 1$, $W_n(z)$ must be truncated to be polynomials, then $W_n(z)$ is orthonormalized with eigenvalue λ_n and weight $(1-z)^{b+a}(1+z)^{b-a}$. Hence we have

$$\int_{-1}^{+1}(1-z)^{b+a}(1+z)^{b-a}W_m(z)W_n(z)dz=\delta_{m,n}. \tag{A4}$$

The Battle–Lemarié wavelet or Daubechies' compact support wavelets[12] can be used in numerical computation for the matrix equation of Eq. (A3) and to prove the convergence of polynomials $W_n(z)$, but we do not pursue this goal here.

[1] S. W. Hawking, Nature (London) **248**, 30 (1974); Commun. Math. Phys. **43**, 199 (1975); Phys. Rev. D **13**, 191 (1976); D. Lee, Nucl. Phys. B **264**, 437 (1986); S. Hawking and R. Penrose, *The Nature of Space And Time* (Princeton University Press, Princeton, NJ, 1996).
[2] B. Hartle and S. W. Hawking, Phys. Rev. D **13**, 2188 (1976).
[3] Z. H. Li and L. Liu, Acta Phys. Sin. **46**, 1273 (1997).
[4] T. Damour and R. Ruffini, Phys. Rev. D **14**, 332 (1976).
[5] W. E. Couch, J. Math. Phys. **26**, 2286 (1985); **22**, 1457 (1981).
[6] E. W. Leaver, J. Math. Phys. **27**, 1238 (1986); E. D. Fackerell and R. G. Crossman, *ibid.* **18**, 1849 (1977).
[7] J. W. Liu, J. Math. Phys. **33**, 4026 (1992); B. D. B. Figneiredo and M. Novello, *ibid.* **34**, 3121 (1993).
[8] D. R. Brill, P. L. Chrzanowski, C. M. Pereira, E. D. Fackerell, and J. R. Ipser, Phys. Rev. D **5**, 1913 (1972); B. P. Jensen, J. G. Mc Laughlin, and A. C. Ottewill, *ibid.* **51**, 5676 (1995); M. Carmeli, *Classical Fields: General Relativity and Gauge Theory* (Wiley, New York, 1982).
[9] B. Carter, Phys. Rev. **174**, 1559 (1968); L. Liu, *General Relativity* (Adv. Edu. Pub., Beijing, 1987).
[10] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953); *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun, 9th version (Dover, New York, 1972).
[11] J. N. Goldberg, A. J. Macfarlane, E. T. Newman, F. Rohrlich, and E. C. G. Sudarshan, J. Math. Phys. **8**, 2155 (1967).
[12] G. Battle, Commun. Math. Phys. **110**, 601 (1987); **114**, 93 (1988); I. Daubechies, A. Grossmann, and Y. Meyer, J. Math. Phys. **27**, 1271 (1986).

Commuting difference operators arising from the elliptic $C_2^{(1)}$-face model

Koji Hasegawa[a)]
Mathematical Institute, Tohoku University, Sendai 980-8578, Japan

Takeshi Ikeda[b)]
Department of Applied Mathematics, Okayama University of Science, Okayama 700-0005, Japan

Tetsuya Kikuchi[c)]
Mathematical Institute, Tohoku University, Sendai 980-8578, Japan

(Received 16 October 1998; accepted for publication 16 June 1999)

We study a pair of commuting difference operators arising from the elliptic $C_2^{(1)}$-face model. The operators, whose coefficients are expressed in terms of the Jacobi's elliptic theta function, act on the space of meromorphic functions on the weight space of the C_2-type simple Lie algebra. We show that the space of functions spanned by the level one characters of the affine Lie algebra $\widehat{\mathfrak{sp}}(4,\mathbb{C})$ is invariant under the action of the difference operators. © *1999 American Institute of Physics.* [S0022-2488(99)03109-6]

I. INTRODUCTION

In Ref. 1, one of the authors constructed an L operator for Belavin's elliptic quantum R matrix[2] acting on the space of meromorphic functions on the weight space of the A_n-type simple Lie algebra. The traces of the L operator, the transfer matrices, give rise to a family of commuting difference operators with an elliptic theta function coefficient. In Ref. 3, they are actually equivalent to Rnijsenaars' operators,[4] which are elliptic extensions of Macdonald's q-difference operators.[5] Our aim in the present paper is to take a step toward a generalization of the above construction to the root systems other than the type A. In this paper, we construct a pair of commuting difference operators acting on the space of functions on the C_2-type weight space.

In the construction of Refs. 1 and 3, a relation between Belavin's elliptic quantum R matrix and the face-type solution of the Yang–Baxter equation (YBE),[6] especially the *intertwining vectors*,[7,8] played the central role. For the root systems other than type A, it is known that no vertex-type R-matrices nor the intertwining vectors. Nevertheless, the face-type solutions of the YBE are known for all classical Lie algebras and their vector representations.[6] We will utilize this type of solution to introduce the difference operators. We take traces (see Sec. V) of the fused Boltzmann weights to obtain a pair of difference operators (Theorem 1).

We also show that the space that is spanned by the level one characters of the affine Lie algebra $\widehat{\mathfrak{sp}}(4,\mathbb{C})$ is invariant under the action of the difference operators (Theorem 2).

The plan of this paper is as follows. In Sec. II, we prepare the notation used in the text and state the main results. In Sec. III, we review the $C_n^{(1)}$-face model[6] in the vector representation, which was given by a set of functions called Boltzmann weights. In Sec. IV, we introduce the *path space*, on which the set of Boltzmann weights act naturally as linear maps and thereby explain the notion of the so-called *fusion procedure* (see, for example, Ref. 3 and references therein). We also give a set of formulas for *fused* Boltzmann weights, which leads to the explicit formula of our difference operators [Theorem 1,(ii)]. In Sec. V, we prove the commutativity of the difference operators. In Sec. VI, we prove a property that the difference operators preserve a three-

[a)]Electronic mail: kojihas@math.tohoku.ac.jp
[b)]Electronic: mail: ike@xmath.ous.ac.jp
[c)]Electronic mail: tkikuchi@math.tohoku.ac.jp

dimensional subspace spanned by the level one characters of the affine Lie algebra $\widehat{\mathfrak{sp}}(4,\mathbb{C})$.[9] In the Appendix, we give a formula of a similarity transformation of the Boltzmann weights.

Our result can be seen as a type C generalization of the Felder and Varchenko work,[10] where they showed that the Ruijsenaars system of difference operators can be recovered from the dynamical R matrices, which is nothing but the face-type solution of the YBE.

On the other hand, a BC_n generalization of the Macdonald polynomial theory is studied by Koornwinder.[11] In Ref. 12, van Diejen constructed the corresponding family of q-difference operators and he studied its elliptic extension in Ref. 13. He succeeded in constructing two elliptic commuting operators: one is of the first order and the other is of the nth order, so that they give rise to an elliptic extension of difference quantum Calogero–Moser system of type BC_2[12] in $n=2$. It is likely that our operators can be identified with his system with a special choice of parameters. We hope to report on this issue in the near future.

Extending this work by van Diejen, Hikami and Komori recently obtained a general family of n-commuting difference operators with elliptic function coefficients.[14,15] Besides the step parameter of difference operators and the modulus of elliptic functions, the family contains ten arbitrary parameters. Their construction uses the Shibukawa–Ueno elliptic R operator,[16] together with the elliptic K operators,[17,18] the elliptic solution to the reflection equation, and can be regarded as an elliptic generalization of the Dunkl-type operator approach to those systems, which have been extensively used by Cherednik[19] (see Ref. 20 for the BC_n case). It would be interesting if one can find an explicit relationship between their approach and ours.

II. NOTATION AND RESULTS

Let $\mathfrak{h}$ be a fixed Cartan subalgebra of the simple Lie algebra $\mathfrak{g}:=\mathfrak{sp}(4,\mathbb{C})$ and denote by $\mathfrak{h}^*$ the dual space of $\mathfrak{h}$. We realize the root system R for $(\mathfrak{g},\mathfrak{h})$ as $R:=\{\pm(\epsilon_1\pm\epsilon_2),\pm 2\epsilon_1,\pm 2\epsilon_2\}\subset\mathfrak{h}^*$. A normalized Killing form (,) is given by $(\epsilon_j,\epsilon_k)=\frac{1}{2}\delta_{jk}$. We will often identify the space $\mathfrak{h}$ and its dual $\mathfrak{h}^*$ via the form (,). The fundamental weights are given by $\varpi_1=\epsilon_1$, $\varpi_2=\epsilon_1+\epsilon_2$. Let $\mathcal{P}_d$ be the set of weights for the fundamental representation $L(\varpi_d)$. We have

$$\mathcal{P}_1=\{\pm\epsilon_1,\pm\epsilon_2\},\quad \mathcal{P}_2=\{\pm(\epsilon_1\pm\epsilon_2),0\}. \tag{2.1}$$

Note that, in these cases, the multiplicity of the weights are all one.

Fix an elliptic modulus τ in the upper half-plane $\Im\tau>0$ and a generic nonzero complex number $\hbar$. Let $[u]$ denote the Jacobi theta function with elliptic nome $p:=e^{2\pi i r}(\Im\tau>0)$, defined by

$$[u]:=ip^{1/8}\sin\pi u\prod_{m=1}^{\infty}(1-2p^m\cos 2\pi u+p^{2m})(1-p^m).$$

This is an odd function and has the following quasiperiodicity:

$$[u+m]=(-1)^m[u],\quad [u+m\tau]=(-1)^m e^{-\pi i m^2\tau-2\pi i m u}[u]\quad (m\in\mathbb{Z}). \tag{2.2}$$

Let d,d' be 1 or 2. Then the $C_2^{(1)}$-type Boltzmann weights of the type (d,d') are given as follows. For any square

$$\begin{pmatrix}\lambda & \mu\\ \kappa & \nu\end{pmatrix}\quad (\lambda,\mu,\nu,\kappa\in\mathfrak{h}^*)$$

of weights, the Boltzmann weight

$$W_{dd'}\left(\begin{matrix}\lambda & \mu\\ \kappa & \nu\end{matrix}\middle| u\right)$$

is given as a function of the spectral parameter $u \in \mathbb{C}$. See the next section for the explicit formula for W_{11}, which are expressed by the Jacobi theta function.

They satisfy the condition:

$$W_{dd'}\left(\begin{matrix} \lambda & \mu \\ \kappa & \nu \end{matrix}\middle| u\right)=0, \quad \text{unless} \quad \mu-\lambda, \nu-\kappa \in 2\hbar\mathcal{P}_d, \quad \kappa-\lambda, \nu-\mu \in 2\hbar\mathcal{P}_{d'},$$

and solve the YBE,

$$\sum_\eta W_{dd'}\left(\begin{matrix} \rho & \eta \\ \sigma & \kappa \end{matrix}\middle| u-v\right) W_{dd''}\left(\begin{matrix} \lambda & \mu \\ \rho & \eta \end{matrix}\middle| u-w\right) W_{d'd''}\left(\begin{matrix} \mu & \nu \\ \eta & \kappa \end{matrix}\middle| v-w\right)$$
$$=\sum_\eta W_{d'd''}\left(\begin{matrix} \lambda & \eta \\ \rho & \sigma \end{matrix}\middle| v-w\right) W_{dd''}\left(\begin{matrix} \eta & \nu \\ \sigma & \kappa \end{matrix}\middle| u-w\right) W_{dd'}\left(\begin{matrix} \lambda & \mu \\ \eta & \nu \end{matrix}\middle| u-v\right). \tag{2.3}$$

The original Boltzmann weights in Ref. 6 are of the type (1,1) in the above terminology. We generalized it by the fusion procedure (see Sec. IV) for the present purpose.

For $\lambda \in \mathfrak{h}^*$ and $p \in \mathcal{P}_d (d=1,2)$, we put

$$\lambda_p := (\lambda, p).$$

Theorem 1: *Let* $M_d(u)$ $(u \in \mathbb{C}, d=1,2)$ *be the following difference operators acting on the space of functions on* $\mathfrak{h}^*$,

$$(M_d(u)f)(\lambda) := \sum_{p \in \mathcal{P}_d} W_{d2}\left(\begin{matrix} \lambda & \lambda+2\hbar p \\ \lambda & \lambda+2\hbar p \end{matrix}\middle| u\right) T_{\hat{p}} f(\lambda),$$

where $T_{\hat{p}} f(\lambda) := f(\lambda+2\hbar p)$.

(i) *We have* $M_d(u)M_{d'}(v)=M_{d'}(v)M_d(u)$ $(u,v \in \mathbb{C}, d,d'=1,2)$.

(ii) *Let us define the following difference operators independent of the spectral parameter* u:

$$\tilde{M}_1 := \sum_{p \in \mathcal{P}_1} \prod_{\substack{q \in \mathcal{P}_1 \\ q \neq \pm p}} \frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]} T_{\hat{p}},$$

$$\tilde{M}_2 := \sum_{\substack{p=\pm\epsilon_1 \\ q=\pm\epsilon_2}} \left(\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}+\hbar]} T_{\hat{p}} T_{\hat{q}} + \frac{[2\hbar]}{[6\hbar]} \frac{[2\lambda_p+2\hbar]}{[2\lambda_p]} \frac{[2\lambda_q+2\hbar]}{[2\lambda_q]} \frac{[\lambda_{p+q}-5\hbar]}{[\lambda_{p+q}+\hbar]} \frac{[\lambda_{p+q}+2\hbar]}{[\lambda_{p+q}]} \right).$$

Then we have $M_1(u)=F(u)\tilde{M}_1$, $M_2(u)=G(u)(\tilde{M}_2-H(u))$, *where*

$$F(u) := \frac{[u][u+2\hbar]^2[u+4\hbar]}{[-3\hbar]^2[\hbar]^2},$$

$$G(u) := \frac{[u-\hbar][u]^2[u+\hbar][u+2\hbar][u+3\hbar]^2[u+4\hbar]}{[-3\hbar]^4[\hbar]^4}, \tag{2.4}$$

and

$$H(u) := \frac{[u+6\hbar][u-3\hbar][2\hbar]}{[u][u+3\hbar][6\hbar]}.$$

In Sec. VI, we introduce a space of Weyl group-invariant theta functions, which are preserved by the actions of the difference operators. For $\beta\in\mathfrak{h}^*$, we introduce the following operators $S_{\tau\beta}$, S_β acting on the functions on $\mathfrak{h}^*$:

$$(S_{\tau\beta}f)(\lambda):=\exp[2\pi i((\lambda,\beta)+\tau(\beta,\beta)/2)]f(\lambda+\tau\beta),$$

$$(S_\beta f)(\lambda):=f(\lambda+\beta).$$

They satisfy Heisenberg's relations,

$$S_\beta S_\gamma=S_\gamma S_\beta,\quad S_{\tau\beta}S_{\tau\gamma}=S_{\tau\gamma}S_{\tau\beta},\quad S_\gamma S_{\tau\beta}=e^{2\pi i(\gamma,\beta)}S_{\tau\beta}S_\gamma \tag{2.5}$$

$(\gamma,\beta,\in\mathfrak{h}^*)$.

Let $Q^\vee$, $P^\vee$ be the coroot and coweight lattice, respectively. Let $W\subset GL(\mathfrak{h}^*)$ denote the Weyl group for $(\mathfrak{g},\mathfrak{h})$. Let Th^W be a space of W-invariant theta functions, defined by

$$Th^W:=\left\{ f \text{ is a holomorphic function over } \mathfrak{h}^* \left| \begin{array}{ll} S_{\tau\alpha}f=S_\alpha f=f & (\forall\alpha\in Q^\vee) \\ f(w\lambda)=f(\lambda) & (\forall w\in W) \end{array}\right. \right\}.$$

It is well known that the space is spanned by the level one characters of the affine Lie algebra $\widehat{\mathfrak{sp}}(4,\mathbb{C})$, and the dimension of this space is three.

Theorem 2: *We have*

$$\tilde{M}_d(Th^W)\subset Th^W\quad(d=1,2).$$

The corresponding facts in the case of the A type are proved in Refs. 21 and 3.

III. THE $C_n^{(1)}$-FACE MODEL

Fix an integer $n\geqslant 2$. We review the definition of the $C_n^{(1)}$-face model given in Ref. 6. We realize the root system R of the type C_n as

$$R:=\{\pm(\epsilon_j\pm\epsilon_k),\pm 2\epsilon_l | 1\leqslant j<k\leqslant n, 1\leqslant l\leqslant n\},$$

where $\{\epsilon_j\}_{j=1}^n$ is a basis of a complex vector space denoted by $\mathfrak{h}^*$ with a bilinear form (,), defined by

$$(\epsilon_j,\epsilon_k):=\tfrac{1}{2}\delta_{jk}.$$

The vector space $\mathfrak{h}^*$ can be identified with the dual space of a Cartan subalgebra $\mathfrak{h}$ of the simple Lie algebra $\mathfrak{sp}(2n,\mathbb{C})$. The fundamental weights $\varpi_j(1\leqslant j\leqslant n)$ are given by $\varpi_j=\epsilon_1+\epsilon_2+\cdots+\epsilon_j$. Let $\mathcal{P}$ denote the set of weights that belongs to the vector representation $L(\varpi_1)$ of $\mathfrak{sp}(2n,\mathbb{C})$. We have

$$\mathcal{P}=\{\pm\epsilon_1,\pm\epsilon_2,\dots,\pm\epsilon_n\}.$$

Note that the multiplicity of the weights in $\mathcal{P}$ are all one.

We shall use the following notation frequently:

$$\hat{\mathcal{P}}:=2\hbar\mathcal{P}\quad\text{and}\quad\hat{p}:=2\hbar p\quad(p\in\mathcal{P}).$$

The Boltzmann weights are given by a set of functions of spectral parameter $u\in\mathbb{C}$ defined for any square

$$\begin{pmatrix} \lambda & \mu \\ \kappa & \nu \end{pmatrix}$$

of elements of $\mathfrak{h}^*$. Let us denote the functions by

$$W\left(\begin{matrix} \lambda & \mu \\ \kappa & \nu \end{matrix}\middle| u\right).$$

They satisfy the condition

$$W\left(\begin{matrix} \lambda & \mu \\ \kappa & \nu \end{matrix}\middle| u\right)=0, \quad \text{unless} \quad \mu-\lambda, \nu-\mu, \kappa-\lambda, \nu-\kappa \in \hat{\mathcal{P}}.$$

For p, q, r, $s \in \mathcal{P}$ such that $p+q=r+s$, we will write

$$s \overset{p}{\underset{r}{\boxed{u}}} q = W\left(\begin{matrix} \lambda & \lambda+\hat{p} \\ \lambda+\hat{s} & \lambda+\hat{p}+\hat{q} \end{matrix}\middle| u\right).$$

They are explicitly given as follows:

$$p \overset{p}{\underset{p}{\boxed{u}}} p = \frac{[c-u][u+\hbar]}{[c][\hbar]}, \tag{3.1}$$

$$p \overset{p}{\underset{q}{\boxed{u}}} q = \frac{[c-u][\lambda_{p-q}-u]}{[c][\lambda_{p-q}]} \quad (p \neq \pm q), \tag{3.2}$$

$$p \overset{q}{\underset{q}{\boxed{u}}} p = \frac{[c-u][u][\lambda_{p-q}+\hbar]}{[c][\hbar][\lambda_{p-q}]} \quad (p \neq \pm q), \tag{3.3}$$

$$p \overset{q}{\underset{-p}{\boxed{u}}} -q = -\frac{[u][\lambda_{p+q}+\hbar+c-u]}{[c][\lambda_{p+q}+\hbar]} \frac{[2\lambda_p+2\hbar]}{[2\lambda_q]} \frac{\Pi_{r\neq \pm p}[\lambda_{p+r}+h]}{\Pi_{r\neq \pm q}[\lambda_{q+r}]} \quad (p \neq q), \tag{3.4}$$

$$p \overset{p}{\underset{-p}{\boxed{u}}} -p = \frac{[c-u][2\lambda_p+\hbar-u]}{[c][2\lambda_p+\hbar]} - \frac{[u][2\lambda_p+\hbar+c-u]}{[c][2\lambda_p+\hbar]} \frac{[2\lambda_p+2\hbar]}{[2\lambda_p]} \prod_{q\neq \pm p} \frac{[\lambda_{p+q}+\hbar]}{[\lambda_{p+q}]}. \tag{3.5}$$

The *crossing parameter* c in the above formulas are fixed to be

$$c := -(n+1)\hbar. \tag{3.6}$$

Proposition 1: The Boltzmann weights (3.1),(3.2),(3.3),(3.4),(3.5) enjoy the following properties.

Initial condition:

$$\sum_{\eta} W\left(\begin{matrix} \lambda & \mu \\ \kappa & \nu \end{matrix}\middle| 0\right) = \delta_{\mu\kappa}. \tag{3.7}$$

Inversion relation:

$$\sum_{\eta} W\left(\begin{matrix}\lambda & \eta \\ \kappa & \nu\end{matrix}\middle| -u\right) W\left(\begin{matrix}\lambda & \mu \\ \eta & \nu\end{matrix}\middle| -u\right) = \delta_{\mu\kappa}\frac{[c+u][c-u][\hbar+u][\hbar-u]}{[c]^2[\hbar]^2}. \tag{3.8}$$

Crossing symmetry:

$$W\left(\begin{matrix}\lambda & \mu \\ \kappa & \nu\end{matrix}\middle| u\right) = \frac{g(\lambda,\kappa)}{g(\mu,\nu)} W\left(\begin{matrix}\kappa & \lambda \\ \nu & \mu\end{matrix}\middle| c-u\right), \tag{3.9}$$

where we put

$$g(\lambda,\mu) := [2\mu_p] \prod_{\substack{q\in\mathcal{P} \\ q\neq \pm p}} [\mu_{p+q}] \quad (\mu=\lambda+\hat{p}, p\in\mathcal{P}).$$

Reflection symmetry:

$$W\left(\begin{matrix}\lambda & \mu \\ \kappa & \nu\end{matrix}\middle| u\right) = \frac{g(\lambda,\kappa)g(\kappa,\nu)}{g(\lambda,\mu)g(\mu,\nu)} W\left(\begin{matrix}\lambda & \kappa \\ \mu & \nu\end{matrix}\middle| u\right). \tag{3.10}$$

Proof: The equation (3.7) is trivial. The two types of symmetries (3.9), (3.10) are easily checked by the explicit form. In the case of $\lambda=\nu$ the equation (3.8) is reduced to the following:

$$\begin{aligned}&\sum_{r\in\mathcal{P}} \frac{[\lambda_p+\lambda_r+\hbar+c-u][\lambda_q+\lambda_r+\hbar+c+u]}{[\lambda_p+\lambda_r+\hbar][\lambda_q+\lambda_r+\hbar]} G_{\lambda r} \\ &\quad = \delta_{p,q}\frac{[c-u][c+u][2\lambda_p][2\lambda_q+2\hbar]}{[\hbar]^2[2\lambda_p+\hbar]^2} G_{\lambda p}^{-1} + \frac{[c+u][2\lambda_p+\hbar+u][\lambda_p+\lambda_q+\hbar+c-u]}{[u][2\lambda_p+\hbar][\lambda_p+\lambda_q+\hbar]} \\ &\qquad - \frac{[c-u][2\lambda_q+\hbar-u][\lambda_p+\lambda_q+\hbar+c+u]}{[u][2\lambda_q+\hbar][\lambda_p+\lambda_q+\hbar]}.\end{aligned} \tag{3.11}$$

Here we denote by $G_{\lambda p}$ the following function:

$$G_{\lambda p} := -\frac{[2\lambda_p+2\hbar]}{[2\lambda_p]} \prod_{\substack{r\in\mathcal{P} \\ r\neq \pm p}} \frac{[\lambda_{p+r}+\hbar]}{[\lambda_{p+r}]} \quad (p\in\mathcal{P}). \tag{3.12}$$

One can find a proof of the equation (3.11) in Ref. 6 [see (3.5) and Lemma 3]. The cases $\nu=\lambda+2\hat{p}(p\in\mathcal{P})$ are trivial. The remaining cases are easily checked by using the following *three-term identity*:

$$[u+x][u-x][v+y][v-y]-[u+y][u-y][v+x][v-x]=[x+y][x-y][u+v][u-v] \tag{3.13}$$

$(u,v,x,y\in\mathbb{C})$. □

We adopted a slightly different formulas (3.3),(3.4) from the original ones [see (A1),(A2)] in Ref. 6. In the Appendix, we will give a similarity transformation (A3),(A4) which transforms our Boltzmann weights into the original ones. Thus, one has a way to prove the YBE for our Boltzmann weights, since such a transformation does not destroy the varidity of the YBE. If we follow this track, however, we must specify the arguments of the square roots contained in the expressions of the original formulas and the transformation. This way of proof may require a rather complicated discussion. In this paper, we will give a proof of the YBE for our Boltzmann weights directly without using the similarity transformation. In fact, our proof here goes quite parallel to the proof given in Ref. 6.

Theorem 3: *The Boltzmann weights*

$$W\begin{pmatrix} \lambda & \mu \\ \kappa & \nu \end{pmatrix} u \Bigg)$$

(3.1),(3.2),(3.3),(3.4),(3.5) solve the YBE *(2.3) for* $d=d'=d''=1$.

Proof: Set

$$X(\lambda,\mu,\nu,\kappa,\sigma,\rho|u,v) := \sum_\eta W\left(\begin{matrix}\rho & \eta\\ \sigma & \kappa\end{matrix}\middle| u\right) W\left(\begin{matrix}\lambda & \mu\\ \rho & \eta\end{matrix}\middle| u+v\right) W\left(\begin{matrix}\mu & \nu\\ \eta & \kappa\end{matrix}\middle| v\right), \tag{3.14}$$

$$Y(\lambda,\mu,\nu,\kappa,\sigma,\rho|u,v) := \sum_\eta W\left(\begin{matrix}\lambda & \eta\\ \rho & \sigma\end{matrix}\middle| v\right) W\left(\begin{matrix}\eta & \nu\\ \sigma & \kappa\end{matrix}\middle| u+v\right) W\left(\begin{matrix}\lambda & \mu\\ \eta & \nu\end{matrix}\middle| u\right), \tag{3.15}$$

and

$$Z(\lambda,\mu,\nu,\kappa,\sigma,\rho,|u,v) := X(\lambda,\mu,\nu,\kappa,\sigma,\rho|u,v) - Y(\lambda,\mu,\nu,\kappa,\sigma,\rho|u,v). \tag{3.16}$$

Regarding $Z(\lambda,\mu,\nu,\kappa,\sigma,\rho|u,v)$ as a function of u, we denote it by $Z(u)$.

The equations (3.7) and (3.8) imply $Z(0)=Z(-v)=0$. Since we have

$$Z(\lambda,\mu,\nu,\kappa,\sigma,\rho|u,v) = -\frac{g(\lambda,\rho)}{g(\nu,\kappa)} Z(\rho,\lambda,\mu,\nu,\kappa,\sigma|c-u-v,u) \tag{3.17}$$

by (3.9), this shows $Z(c-v)=Z(c)=0$ also. Thus, we have found the four zeros at $u=0, -v, c, c-v$ of $Z(u)$. By the exactly same argument in Ref. 6 using the quasiperiodicity property of $Z(u)$, (3.17), and the following symmetry [this follows from (3.10)]:

$$Z(\lambda,\mu,\nu,\kappa,\sigma,\rho|u,v) = \frac{g(\lambda,\rho)g(\rho,\sigma)g(\sigma,\kappa)}{g(\lambda,\mu)g(\mu,\nu)g(\nu,\kappa)} Z(\lambda,\rho,\sigma,\kappa,\nu,\mu|v,u),$$

we can reduce the proof of the YBE to the following two special cases:

$$Z(\lambda,\lambda+\hat{p},\lambda+\hat{p}+\hat{q},\lambda+\hat{p}+\hat{q}+\hat{r},\lambda+\hat{q}+\hat{r},\lambda+\hat{r}|u,v)=0, \tag{3.18}$$

where $r\neq\pm p,\pm q, p\neq\pm q$ and

$$Z(\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p}|u,v)=0. \tag{3.19}$$

In the case of the equation (3.18), each side of the YBE contains only one term, and they are manifestly the same. A proof of the last case (3.19) can be found in the original literature.[6] However, since the proof is brief and seems to contain some typographical errors, we will describe details of it in the following for the readers' convenience.

We will prove $Z(\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p}|u,v)=0$. Regarding $Y(\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p}|u,v)$ as a function of λ_p, we denote it by $f(\lambda_p)$. It reads as

$$f(\lambda_p)=G_{\lambda p}\frac{[u][v][w]}{[c]^3}\sum_{q\in\mathcal{P}}\frac{[\lambda_q+\lambda_p+\hbar+\tilde{u}][\lambda_q+\lambda_p+\hbar+\tilde{v}][\lambda_q+\lambda_p+\hbar+\tilde{w}]}{[\lambda_q+\lambda_p+\hbar]^3}G_{\lambda q}$$
$$+G_{\lambda p}^{-1}\frac{[\tilde{u}][\tilde{v}][\tilde{w}]}{[c]^3}\frac{[2\lambda_p+\hbar-u][2\lambda_p+\hbar-v][2\lambda_p+\hbar-w]}{[2\lambda_p+\hbar]^3}$$
$$+\sum_{\text{cyclic}}\frac{[u][\tilde{v}][\tilde{w}]}{[c]^3}\frac{[2\lambda_p+\hbar+\tilde{u}][2\lambda_p+\hbar-v][2\lambda_p+\hbar-w]}{[2\lambda_p+\hbar]^3}$$
$$+G_{\lambda p}\sum_{\text{cyclic}}\frac{[\tilde{u}][v][w]}{[c]^3}\frac{[2\lambda_p+\hbar-u][2\lambda_p+\hbar+\tilde{v}][2\lambda_p+\hbar+\tilde{w}]}{[2\lambda_p+\hbar]^3},$$

where we put $w=c-u-v$, $\tilde{u}=c-u$, $\tilde{v}=c-v$, $\tilde{w}=c-w$ and the summation Σ_{cyclic} is over the cyclic permutations of the three variables (u,v,w). From the explicit form, one can see that $X(\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p},\lambda,\lambda+\hat{p}|u,v)=f(-\lambda_p-\hbar)$. We will prove $f(\lambda_p)=f(-\lambda_p-\hbar)$.

Now consider a function,

$$\Phi(z):=\frac{[z+\lambda_p+\hbar+\tilde{u}][z+\lambda_p+\hbar+\tilde{v}][z+\lambda_p+\hbar+\tilde{w}]}{[z+\lambda_p+\hbar]^3}\frac{[0]'}{[\hbar]}\frac{[2z+2\hbar]}{[2z+\hbar]}\prod_{q\in\mathcal{P}}\frac{[z+\lambda_q+\hbar]}{[z+\lambda_q]}.$$

One sees that $\Phi(z)$ is a doubly periodic function of the periods 1 and τ. Its poles are located at $z=-\lambda_p-h$, $\lambda_q(q\in\mathcal{P})$, $-\hbar/2+\omega(\omega=0,1/2,\tau/2,(1+\tau)/2)$. The pole at $z=-\lambda_p-\hbar$ is of the second order, and the others are simple.

Let $f_i(\lambda_p)(i=1,2,3,4)$ denote the ith term of the above function $f(\lambda_p)$. Since we have

$$\operatorname*{Res}_{z=\lambda_q}\Phi(z)dz=-\frac{[\lambda_q+\lambda_p+\hbar+\tilde{u}][\lambda_q+\lambda_p+\hbar+\tilde{v}][\lambda_q+\lambda_p+\hbar+\tilde{w}]}{[\lambda_q+\lambda_p+\hbar]^3}G_{\lambda_q},$$

the relation $\Sigma\,\mathrm{Res}\,\Phi(z)dz=0$ implies $f_1(\lambda_p)=a(\lambda_p)+b(\lambda_p)$, where we set

$$a(\lambda_p):=G_{\lambda p}\frac{[u][v][w]}{[c]^3}\sum_{\omega}\operatorname*{Res}_{z=-\hbar/2+\omega}\Phi(z)dz, \tag{3.20}$$

$$b(\lambda_p):=G_{\lambda p}\frac{[u][v][w]}{[c]^3}\operatorname*{Res}_{z=-\lambda_p-\hbar}\Phi(z)dz. \tag{3.21}$$

Here the summation Σ_ω is over the half-periods $\omega=0,1/2,\tau/2,(1+\tau)/2$.

From (2.2) and (3.6), we have, for $\omega=0,1/2,\tau/2,(1+\tau)/2$.

$$\operatorname*{Res}_{z=-\hbar/2+\omega}\Phi(z)dz=\frac{1}{2}\frac{\left[\lambda_p+\frac{\hbar}{2}+\omega+\tilde{u}\right]\left[\lambda_p+\frac{\hbar}{2}+\omega+\tilde{v}\right]\left[\lambda_p+\frac{\hbar}{2}+\omega+\tilde{w}\right]}{\left[\lambda_p+\frac{\hbar}{2}+\omega\right]^3}e^{2\pi i\xi(\omega)}, \tag{3.22}$$

where we put $\xi(0)=\xi(\frac{1}{2})=0$, $\xi(\tau/2)=\xi((1+\tau)/2)=c$. Combining (3.20), (3.22), and Lemma 3 in Ref. 6, we can verify

$$a(\lambda_p)+f_4(\lambda_p)-f_2(-\lambda_p-\hbar)=-a(-\lambda_p-\hbar)+f_2(\lambda_p)-f_4(-\lambda_p-\hbar)=0. \tag{3.23}$$

Set $\phi(u)=(d/du)\log[u]$, then the residue $\mathrm{Res}_{z=-\lambda_p-\hbar}\Phi(z)dz$ can be expressed as

$$\operatorname*{Res}_{z=-\lambda_p-\hbar} \Phi(z)dz = G_{\lambda p}^{-1} \frac{[\tilde{u}][\tilde{v}][\tilde{w}]}{[0]'[\hbar]^2} \frac{[2\lambda_p+2\hbar][2\lambda_p]}{[2\lambda_p+\hbar]^2}$$
$$\times \left(\sum_{\text{cyclic}} \phi(\tilde{u}) - 3\phi(2\lambda_p) + 3\phi(2\lambda_p+\hbar) + \phi(\hbar) \right.$$
$$\left. + \sum_{\substack{q\in\mathcal{P} \\ q\neq\pm p}} \{\phi(-\lambda_p+\lambda_q) - \phi(-\lambda_p+\lambda_q-\hbar)\} \right). \tag{3.24}$$

Since $\phi(u)$ is an odd function, we have, from (3.21) and (3.24),

$$b(\lambda_p) - b(-\lambda_p-\hbar) = -3 \frac{[u][v][w]}{[c]^3} \frac{[\tilde{u}][\tilde{v}][\tilde{w}]}{[0]'[\hbar]^2} \frac{[2\lambda_p+2\hbar][2\lambda_p]}{[2\lambda_p+\hbar]^2}$$
$$\times \{\phi(2\lambda_p) + \phi(2\lambda_p+2\hbar) - 2\phi(2\lambda_p+\hbar)\}. \tag{3.25}$$

On the other hand, using the identity [see (3.13)]

$$[2\lambda_p+\hbar+\tilde{u}][2\lambda_p+\hbar-v][2\lambda_p+\hbar-w] - [2\lambda_p+\hbar-\tilde{u}][2\lambda_p+\hbar+v][2\lambda_p+\hbar+w]$$
$$= [\tilde{u}][v][w] \frac{[4\lambda_p+2\hbar]}{[2\lambda_p+\hbar]},$$

and its cyclic permutations of (u,v,w), we have

$$f_3(\lambda_p) - f_3(-\lambda_p-\hbar) = 3 \frac{[u][v][w][\tilde{u}][\tilde{v}][\tilde{w}]}{[c]^3} \frac{[4\lambda_p+2\hbar]}{[2\lambda_p+\hbar]^4}. \tag{3.26}$$

Now from (3.25) and (3.26), we have

$$b(\lambda_p) + f_3(\lambda_p) = b(-\lambda_p-\hbar) + f_3(-\lambda_p-\hbar), \tag{3.27}$$

where we used the following identity (Lemma 4 in Ref. 6):

$$\phi(u+\hbar) + \phi(u-\hbar) - 2\phi(u) = \frac{[\hbar]^2[2u][0]'}{[u]^2[u-\hbar][u+\hbar]}.$$

Combining (3.23) and (3.27), we obtained $f(\lambda_p) = f(-\lambda_p-h)$. □

IV. PATH SPACE AND FUSION PROCEDURE

In the previous section we introduced the Boltzmann weights $W(u)$ of the type (1,1) and proved that they satisfy the YBE. In what follows, we treat only the case of $n=2$. To construct commuting difference operators, we need the general types of the Boltzmann weights $W_{dd'}(u)$, which we call the fused Boltzmann weights.

First let us introduce the notion of the path space. Let $d=1,2$. For any $u\in\mathbb{C}$ and $\lambda,\mu\in\mathfrak{h}^*$ such that $\mu-\lambda\in 2\hbar\mathcal{P}_d$, we introduce a formal symbol,

$$g_\lambda^\mu(u) := \begin{cases} e_\lambda^\mu(u) & :d=1, \\ f_\lambda^\mu(u) & :d=2. \end{cases}$$

See (2.1) for the notation $\mathcal{P}_1$ and $\mathcal{P}_2$. We define the complex vector space,

$$\hat{\mathcal{P}}(\varpi_d^u)_\lambda^\mu := \begin{cases} \mathbb{C} g_\lambda^\mu(u) & : \mu-\lambda \in 2\hbar \mathcal{P}_d, \\ 0 & : \text{otherwise}, \end{cases}$$

for each $u \in \mathbb{C}$, and the *space of paths* from λ to μ of the type $(d_1,\dots,d_k;u_1,\dots,u_k)$,

$$\hat{\mathcal{P}}(\varpi_{d_1}^{u_1}\otimes\cdots\otimes\varpi_{d_k}^{u_k})_\lambda^\nu := \bigoplus_{\mu_1,\dots,\mu_{k-1}\in\mathfrak{h}^*} \hat{\mathcal{P}}(\varpi_{d_1}^{u_1})_\lambda^{\mu_1}\otimes\hat{\mathcal{P}}(\varpi_{d_2}^{u_2})_{\mu_1}^{\mu_2}\otimes\cdots\otimes\hat{\mathcal{P}}(\varpi_{d_k}^{u_k})_{\mu_{k-1}}^\nu. \tag{4.1}$$

The following set:

$$\{g_\lambda^{\mu_1}(\mu_1)\otimes g_{\mu_1}^{\mu_2}(u_2)\otimes\cdots\otimes g_{\mu_{k-1}}^\nu(u_k) \,|\, \mu_i-\mu_{i-1}\in 2\hbar\mathcal{P}_d(1\leqslant i\leqslant k), \mu_0=\lambda, \mu_k=\nu\},$$

of *paths* forms a basis of the space (4.1). Set also

$$\hat{\mathcal{P}}(\varpi_{d_1}^{u_1}\otimes\cdots\otimes\varpi_{d_k}^{u_k})_\lambda := \bigoplus_{\nu\in\mathfrak{h}^*}\hat{\mathcal{P}}(\varpi_{d_1}^{u_1}\otimes\cdots\otimes\varpi_{d_k}^{u_k})_\lambda^\nu$$

and

$$\hat{\mathcal{P}}(\varpi_{d_1}^{u_1}\otimes\cdots\otimes\varpi_{d_k}^{u_k}) := \bigoplus_{\lambda\in\mathfrak{h}^*}\hat{\mathcal{P}}(\varpi_{d_1}^{u_1}\otimes\cdots\varpi_{d_k}^{u_k})_\lambda.$$

In the following, we will construct the linear operators:

$$W_{dd'}(u-v): \hat{\mathcal{P}}(\varpi_d^u\otimes\varpi_{d'}^v)\to\hat{\mathcal{P}}(\varpi_{d'}^v,\otimes\varpi_d^u),$$

which satisfy the following YBE $(d,d',d''=1,2)$:

$$\begin{aligned}&(\mathrm{id}\otimes W_{dd'}(u-v))(W_{dd''}(u-w)\otimes\mathrm{id})(\mathrm{id}\otimes W_{d'd''}(v-w))\\ &\quad=(W_{d'd''}(v-w)\otimes\mathrm{id})(\mathrm{id}\otimes W_{dd''}(u-w))(W_{dd'}(u-v)\otimes\mathrm{id})\\ &\quad:\hat{\mathcal{P}}(\varpi_d^u\otimes\varpi_{d'}^v\otimes\varpi_{d''}^w)\to\hat{\mathcal{P}}(\varpi_{d''}^w\otimes\varpi_{d'}^v\otimes\varpi_d^u).\end{aligned} \tag{4.2}$$

First we define a linear operator $W(\varpi_1^u,\varpi_1^v):\hat{\mathcal{P}}(\varpi_1^u\otimes\varpi_1^v)\to\hat{\mathcal{P}}(\varpi_1^v\otimes\varpi_1^u)$ by

$$W(\varpi_1^u,\varpi_1^v)e_\lambda^\mu(u)\otimes e_\mu^\nu(v) := \sum_{\kappa\in\mathfrak{h}^*} W\left(\begin{matrix}\lambda & \mu\\ \kappa & \nu\end{matrix}\middle| u-v\right) e_\lambda^\kappa(v)\otimes e_\kappa^\nu(u).$$

Put $W_{11}(u-v) := W(\varpi_1^u,\varpi_1^v)$, then the YBE (4.2) for $d=d'=d''=1$ is nothing but (2.3).

To construct $W_{dd'}(u-v)$ other than $W_{11}(u-v)$, we will formulate the fusion procedure. Put

$$\begin{aligned}W(\varpi_1^{u_1}\otimes\varpi_1^{u_2}\otimes\cdots\otimes\varpi_1^{u_k},\varpi_1^v) &:= W^{1,2}(\varpi_1^{u_1},\varpi_1^v)W^{2,3}(\varpi_1^{u_2},\varpi_1^v)\cdots W^{k,k+1}(\varpi_1^{u_k},\varpi_1^v)\\ &:\hat{\mathcal{P}}(\varpi_1^{u_1}\otimes\varpi_1^{u_2}\otimes\cdots\otimes\varpi_1^{u_k}\otimes\varpi_1^v)\to\hat{\mathcal{P}}(\varpi_1^v\otimes\varpi_1^{u_1}\otimes\varpi_1^{u_2}\cdots\otimes\varpi_1^{u_k}),\end{aligned}$$

where

$$\begin{aligned} W(\varpi_1^{u_j},\varpi_1^{v})^{j,j+1} &:= \mathrm{id}^{\otimes(j-1)}\otimes W(\varpi_1^{u_j},\varpi_1^{v})\otimes \mathrm{id}^{\otimes(k-j)} \\ &: \hat{\mathcal{P}}(\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_{j-1}}\otimes \underbrace{\varpi_1^{u_j}\otimes\varpi_1^{v}}\otimes\varpi_1^{u_{j+1}}\otimes\cdots\otimes\varpi_1^{u_k}) \\ &\to \hat{\mathcal{P}}(\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_{j-1}}\otimes \underbrace{\varpi_1^{v}\otimes\varpi_1^{u_j}}\otimes\varpi_1^{u_{j+1}}\otimes\cdots\otimes\varpi_1^{u_k}). \end{aligned}$$

We also put

$$\begin{aligned} &W(\varpi_1^{u_1}\otimes\varpi_1^{u_2}\otimes\cdots\otimes\varpi_1^{u_k},\varpi_1^{v_1}\otimes\varpi_1^{v_2}\otimes\cdots\otimes\varpi^{v_l}) \\ &:= \overleftarrow{\prod_{1\leqslant j\leqslant l}} W(\varpi_1^{u_1}\otimes\varpi_1^{u_2}\otimes\cdots\otimes\varpi_1^{u_k},\varpi_1^{v_j})^{[j,k+j]} \\ &: \hat{\mathcal{P}}(\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_k}\otimes\varpi_1^{v_1}\otimes\cdots\otimes\varpi_1^{v_l}) \to \hat{\mathcal{P}}(\varpi_1^{v_1}\otimes\cdots\otimes\varpi_1^{v_l}\otimes\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_k}), \end{aligned}$$

where

$$\begin{aligned} &W(\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_k},\varpi_1^{v_j})^{[j,k+j]} \\ &:= \mathrm{id}^{\otimes(j-1)}\otimes W(\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_k},\varpi_1^{v_j})\otimes\mathrm{id}^{\otimes(l-j)} \\ &: \hat{\mathcal{P}}(\varpi_1^{v_1}\otimes\cdots\otimes\varpi_1^{v_{j-1}}\otimes\underbrace{\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_k}\otimes\varpi_1^{v_j}}\otimes\varpi_1^{v_{j+1}}\otimes\cdots\otimes\varpi_1^{v_l}) \\ &\to \hat{\mathcal{P}}(\varpi_1^{v_1}\otimes\cdots\otimes\varpi_1^{v_{j-1}}\otimes\underbrace{\varpi_1^{v_j}\otimes\varpi_1^{u_1}\otimes\cdots\otimes\varpi_1^{u_k}}\otimes\varpi_1^{v_{j+1}}\otimes\cdots\otimes\varpi_1^{v_l}). \end{aligned}$$

We will realize the space $\hat{\mathcal{P}}(\varpi_2^u)$ as a subspace of $\hat{\mathcal{P}}(\varpi_1^u\otimes\varpi_1^{u-\hbar})$. For this purpose, let us introduce the *fusion projector* $\pi_{\varpi_2^u}$ by specializing the parameter in $W(\varpi_1^u,\varpi_1^v)$:

$$\pi_{\varpi_2^u} := W(\varpi_1^{u-\hbar},\varpi_1^u) : \hat{\mathcal{P}}(\varpi_1^{u-\hbar}\otimes\varpi_1^u)\to\hat{\mathcal{P}}(\varpi_1^u\otimes\varpi_1^{u-\hbar}). \tag{4.3}$$

Lemma 1: The space $\pi_{\varpi_2^u}(\hat{\mathcal{P}}(\varpi_1^{u-\hbar}\otimes\varpi_1^u)_\lambda)$ has a basis $\{\overline{f_\lambda^{\lambda+\hat{r}}}(u)\}_{r\in\mathcal{P}_2}$, given by

$$\overline{f_\lambda^{\lambda+\hat{p}+\hat{q}}}(u) := [\lambda_{p-q}+\hbar]e_\lambda^{\lambda+\hat{p}}(u)\otimes e_{\lambda+\hat{p}}^{\lambda+\hat{p}+\hat{q}}(u-\hbar)+[\lambda_{q-p}+\hbar]e_\lambda^{\lambda+\hat{q}}(u)\otimes e_{\lambda+\hat{q}}^{\lambda+\hat{p}+\hat{q}}(u-\hbar), \tag{4.4}$$

where $p=\pm\epsilon_1$, $q=\pm\epsilon_2$, and

$$\overline{f_\lambda^{\lambda}}(u) := \sum_{p\in\mathcal{P}_1}[2\lambda_p+2\hbar]e_\lambda^{\lambda+\hat{p}}(u)\otimes e_{\lambda+\hat{p}}^{\lambda}(u-\hbar). \tag{4.5}$$

Proof: For $p,q\in\mathcal{P}_1$, $q\neq\pm p$, we have

$$\pi_{\varpi_2^u}(e_\lambda^{\lambda+\hat{p}}(u-\hbar)\otimes e_{\lambda+\hat{p}}^{\lambda+2\hat{p}}(u)) = \begin{pmatrix} & p & \\ p & \boxed{-\hbar} & p \\ & p & \end{pmatrix} e_\lambda^{\lambda+\hat{p}}(u)\otimes e_{\lambda+\hat{p}}^{\lambda+2\hat{p}}(u-h)=0,$$

$$\pi_{\varpi_2^u}(e_\lambda^{\lambda+\hat{p}}(u-\hbar)\otimes e_{\lambda+\hat{p}}^{\lambda+\hat{p}+\hat{q}}(u))=\left(\begin{matrix} & p & \\ p & \boxed{-\hbar} & q \\ & q & \end{matrix}\right) e_\lambda^{\lambda+\hat{p}}(u)\otimes e_{\lambda+\hat{p}}^{\lambda+\hat{p}+\hat{q}}(u-\hbar)$$

$$+\left(\begin{matrix} & p & \\ q & \boxed{-\hbar} & q \\ & p & \end{matrix}\right) e_\lambda^{\lambda+\hat{q}}(u)\otimes e_{\lambda+\hat{q}}^{\lambda+\hat{p}+\hat{q}}(u-\hbar)$$

$$=\frac{[-2\hbar]}{[-3\hbar][\lambda_{p-q}]}([\lambda_{p-q}+\hbar]e_\lambda^{\lambda+\hat{p}}(u)\otimes e_{\lambda+\hat{p}}^{\lambda+\hat{p}+\hat{q}}(u-\hbar)$$

$$+[\lambda_{q-p}+\hbar]e_\lambda^{\lambda+\hat{q}}(u)\otimes e_{\lambda+\hat{q}}^{\lambda+\hat{p}+\hat{q}}(u-\hbar)),$$

and

$$\pi_{\varpi_2^u}(e_\lambda^{\lambda+\hat{p}}(u-\hbar)\otimes e_{\lambda+\hat{p}}^{\lambda}(u))=\sum_{r\in\mathcal{P}_1}\left(\begin{matrix} & p & \\ r & \boxed{-\hbar} & -p \\ & -r & \end{matrix}\right) e_\lambda^{\lambda+\hat{r}}(u)\otimes e_{\lambda+\hat{r}}^{\lambda}(u-\hbar)$$

$$=\frac{[-\hbar][\lambda_{p+q}-\hbar][\lambda_{p-q}-\hbar]}{[-3\hbar][\lambda_{p+q}][\lambda_{p-q}][2\lambda_p]}$$

$$\times\left(\sum_{r\in\mathcal{P}_1}[2\lambda_r+2\hbar]e_\lambda^{\lambda+\hat{r}}(u)\otimes e_{\lambda+\hat{r}}^{\lambda}(u-\hbar)\right).$$

Here we have used the three-term identity (3.13). □

Thus, we know the subspace $\pi_{\varpi_2^u}(\hat{\mathcal{P}}(\varpi_1^{u-\hbar}\otimes\varpi_1^u)_\lambda)$ is naturally isomorphic to the space $\hat{\mathcal{P}}(\varpi_2^u)_\lambda$. In the following, we will identify the image $\mathrm{Im}(\pi_{\varpi_2^u})\subset\hat{\mathcal{P}}(\varpi_1^u\otimes\varpi_1^{u-\hbar})$ with the space $\hat{\mathcal{P}}(\varpi_2^u)$ via $\overline{f_\lambda^\mu}(u)\leftrightarrow f_\lambda^\mu(u)$.

Proposition 2: Define the operators $\tilde{W}_{dd'}(u-v)$ by

$$\tilde{W}_{21}(u-v):=W(\varpi_1^u\otimes\varpi_1^{u-\hbar},\varpi_1^v),\quad \tilde{W}_{12}(u-v):=W(\varpi_1^u,\varpi_1^v\otimes\varpi_1^{v-\hbar}) \tag{4.6}$$

and

$$\tilde{W}_{22}(u-v):=W(\varpi_1^u\otimes\varpi_1^{u-\hbar},\varpi_1^v\otimes\varpi_1^{v-h}).$$

We have

$$\tilde{W}_{dd'}(u-v)(\hat{\mathcal{P}}(\varpi_d^u\otimes\varpi_{d'}^v)_\lambda^\mu)\subset\hat{\mathcal{P}}(\varpi_{d'}^v\otimes\varpi_d^v)_\lambda^\mu.$$

Proof: From the definition of $\pi_{\varpi_2^u}$ (4.3) and the YBE (2.3),

$$W^{1,2}(u-v)W^{2,3}(u-v-\hbar)(\pi_{\varpi_2^u}\otimes\mathrm{id})=(\mathrm{id}\otimes\pi_{\varpi_2^u})W^{1,2}(u-v-\hbar)W^{2,3}(u-v). \tag{4.7}$$

Applying this to the definition of $\tilde{W}_{21}(u-v)$, we get

$$\tilde{W}_{21}(u-v)(\hat{\mathcal{P}}(\varpi_2^u\otimes\varpi_1^v)_\lambda^\mu)\subset\hat{\mathcal{P}}(\varpi_1^v\otimes\varpi_2^u)_\lambda^\mu.$$

By a same argument, we have

$$W^{2,3}(u-v+\hbar)W^{1,2}(u-v)(\mathrm{id}\otimes\pi_{\varpi_2^u})=(\pi_{\varpi_2^u}\otimes\mathrm{id})W^{2,3}(u-v)W^{1,2}(u-v+\hbar), \tag{4.8}$$

and

$$\widetilde{W}_{12}(u-v)(\hat{\mathcal{P}}(\varpi_1^u \otimes \varpi_2^v)_\lambda^\mu) \subset \hat{\mathcal{P}}(\varpi_2^v \otimes \varpi_1^u)_\lambda^\mu.$$

Together with the equations (4.7), (4.8) and the definition of $\widetilde{W}_{22}(u-v)$, we obtain

$$\widetilde{W}_{22}(u-v)(\hat{\mathcal{P}}(\varpi_2^u \otimes \varpi_2^v)_\lambda^\mu) \subset \hat{\mathcal{P}}(\varpi_2^v \otimes \varpi_2^u)_\lambda^\mu.$$

□

We denote by $W_{dd'}(u-v)$ the restricted operators $\widetilde{W}_{dd'}(u-v)|_{\hat{\mathcal{P}}(\varpi_d^u \otimes \varpi_{d'}^v)}$ and introduce their matrix coefficients by the following equation:

$$W_{dd'}(u-v) g_\lambda^\mu(u) \otimes g_\mu^\nu(v) = \sum_{\kappa \in \mathfrak{h}^*} W_{dd'}\left(\begin{matrix} \lambda & \mu \\ \kappa & \nu \end{matrix} \middle| u-v \right) g_\lambda^\kappa(v) \otimes g_\kappa^\nu(u).$$

By the construction, the operators $W_{dd'}(u-v)$ clearly satisfies the YBE (4.2) in operator form, and their coefficients

$$W_{dd'}\left(\begin{matrix} \lambda & \mu \\ \kappa & \nu \end{matrix} \middle| u-v \right)$$

satisfies the YBE (2.3). For $p,r \in \mathcal{P}_d$ and $s,q \in \mathcal{P}_{d'}$ $(d,d'=1,2)$ such that $p+q=r+s$, we write for brevity (as long as confusion does not arise)

$$\begin{matrix} & p & \\ s & \boxed{u} & q \\ & r & \end{matrix} = W_{dd'}\left(\begin{matrix} \lambda & \lambda+\hat{p} \\ \lambda+\hat{s} & \lambda+\hat{p}+\hat{q} \end{matrix} \middle| u \right). \tag{4.9}$$

We calculate the coefficients of the operator $W_{21}(u)$ as an example. In what follows, we will often omit the dependence of $g_\lambda^\mu(u) \in \hat{\mathcal{P}}(\varpi_d^u)$ on u (the spectral parameter) for brevity. Let $p \in \mathcal{P}_1$. From the definitions of f_λ^λ (4.5) and $\widetilde{W}_{21}$ (4.6), we have

$$\begin{aligned} W_{21}(u) f_\lambda^\lambda \otimes e_\lambda^{\lambda+\hat{p}} &= \widetilde{W}_{21}(u) f_\lambda^\lambda \otimes e_\lambda^{\lambda+\hat{p}} \\ &= \widetilde{W}_{21}(u)\left(\sum_{r \in \mathcal{P}_1} [2\lambda_r + 2\hbar] e_\lambda^{\lambda+\hat{r}} \otimes e_{\lambda+\hat{r}}^\lambda \otimes e_\lambda^{\lambda+\hat{p}} \right) \\ &= \sum_{q \in \mathcal{P}_1} e_\lambda^{\lambda+\hat{q}} \otimes \left(\sum_{\substack{s,t \in \mathcal{P}_1 \\ s+t=p-q}} V_q(\lambda;s,t;u) e_{\lambda+\hat{q}}^{\lambda+\hat{q}+\hat{s}} \otimes e_{\lambda+\hat{q}+\hat{s}}^{\lambda+\hat{p}} \right), \end{aligned}$$

where we denote by $V_q(\lambda;s,t;u)$ the following function:

$$\sum_{r \in \mathcal{P}_1} [2\lambda_r + 2\hbar] W_{11}\left(\begin{matrix} \lambda & \lambda+\hat{r} \\ \lambda+\hat{q} & \lambda+\hat{q}+\hat{s} \end{matrix} \middle| u \right) W_{11}\left(\begin{matrix} \lambda+\hat{r} & \lambda \\ \lambda+\hat{q}+\hat{s} & \lambda+\hat{p} \end{matrix} \middle| u-\hbar \right).$$

If $q \in \mathcal{P}_1$ such that $q \neq \pm p$, then the functions $V_q(\lambda;s,t;u)$ vanish except for $(s,t)=(p,-q)$ or $(-q,p)$, and one can easily show that

$$\frac{V_q(\lambda;p,-q;u)}{[(\lambda+\hat{q})_{p+q}+\hbar]} = \frac{V_q(\lambda;-q,p;u)}{[(\lambda+\hat{q})_{-q-p}+\hbar]}. \tag{4.10}$$

This equation implies that the vector

$$V_q(\lambda;p,-q;u) e_{\lambda+\hat{q}}^{\lambda+\hat{q}+\hat{p}} \otimes e_{\lambda+\hat{q}+\hat{p}}^{\lambda+\hat{p}} + V_q(\lambda;-q,p;u) e_{\lambda+\hat{q}}^{\lambda} \otimes e_\lambda^{\lambda+\hat{p}}$$

is proportional to $f^{\lambda+\hat{p}}_{\lambda+\hat{q}}$ and its coefficient [the both-hand sides of (4.10)] is calculated as

$$\frac{[u-\hbar][u+\hbar][u+3\hbar][2\hbar]}{[-3\hbar]^2[\hbar]^2}\,\frac{[\lambda_{q-p}-\hbar-u][2\lambda_q+2\hbar]}{[\lambda_{q-p}-\hbar][\lambda_{q+p}+\hbar]},$$

by using the three-term identity (3.13). This function is labeled by [see (4.9)]

$$\begin{array}{ccc} & 0 & \\ q & \boxed{u} & p \\ & p-q & \end{array} \quad (q\neq \pm p).$$

Let us consider the term for $q=p$. For all $s\in\mathcal{P}_1$ we have from the three-term identity,

$$\frac{V_p(\lambda;s,-s;u)}{[2(\lambda+\hat{p})_s+2\hbar]}=\frac{[u-\hbar][u+\hbar][u+3\hbar]}{[-3\hbar]^2[\hbar]}\,\frac{[u+\hbar]}{[\hbar]}\prod_{\substack{r\in\mathcal{P}_1\\ r\neq\pm p}}\frac{[\lambda_{p+r}+2\hbar]}{[\lambda_{p+r}+\hbar]}. \tag{4.11}$$

The right-hand side of this equation is independent of $s\in\mathcal{P}_1$. Thus, we see that the vector,

$$\sum_{s\in\mathcal{P}_1} V_p(\lambda;s,-s;u)e^{\lambda+\hat{p}+\hat{s}}_{\lambda+\hat{p}}\otimes e^{\lambda+\hat{p}}_{\lambda+\hat{p}+\hat{s}},$$

is proportional to $f^{\lambda+\hat{p}}_{\lambda+\hat{p}}$ and its coefficient is equal to the right-hand side of (4.11), which is labeled by

$$\begin{array}{ccc} & 0 & \\ p & \boxed{u} & p. \\ & 0 & \end{array}$$

Here we write all fused Boltzmann weights [the coefficients of the operator $W_{21}(u)$]. They are obtained by the three-term identity (3.13). We assume that $p,q\in\mathcal{P}_1$ satisfy $p\neq\pm q$. The common factor $[u-\hbar][u+\hbar][u+3\hbar][-3\hbar]^{-2}[\hbar]^{-1}$ is dropped:

$$\begin{array}{ccc} & p+q & \\ q & \boxed{u} & q \\ & p+q & \end{array}=\frac{[u+2\hbar]}{[\hbar]},$$

$$\begin{array}{ccc} & p-q & \\ q & \boxed{u} & q \\ & p-q & \end{array}=\frac{[u]}{[\hbar]}\,\frac{[2\lambda_q+2\hbar]}{[2\lambda_q]}\,\frac{[\lambda_{p-q}-\hbar]}{[\lambda_{p-q}+\hbar]},$$

$$\begin{array}{ccc} & 0 & \\ q & \boxed{u} & q \\ & 0 & \end{array}=\frac{[u+\hbar]}{[\hbar]}\prod_{\substack{r\in\mathcal{P}_1\\ r\neq\pm q}}\frac{[\lambda_{q+r}+2\hbar]}{[\lambda_{q+r}+\hbar]}, \tag{4.12}$$

$$\begin{array}{ccc} & q-p & \\ q & \boxed{u} & p \\ & 0 & \end{array}=\frac{[\lambda_{q-p}-u][\lambda_{q+p}+2\hbar]}{[2\lambda_p][\lambda_{q-p}+\hbar]}, \tag{4.13}$$

$$\begin{array}{ccc} & 0 & \\ q & \boxed{u} & p \\ & p-q & \end{array}=\frac{[2\hbar]}{[\hbar]}\,\frac{[2\lambda_{q-p}-\hbar-u][2\lambda_q+2\hbar]}{[2\lambda_{q-p}-\hbar][\lambda_{q+p}+\hbar]}, \tag{4.14}$$

$$\begin{array}{ccc} & p+q & \\ q & \boxed{u} & -q \\ & p-q & \end{array} = \frac{[2\hbar]}{[\hbar]}\frac{[2\lambda_q-u][\lambda_{p-q}-\hbar]}{[2\lambda_q][\lambda_{p+q}+\hbar]}.$$

Next, we give the example of W_{12}. In this case, the common factor $[u][u+2\hbar][u+4\hbar][-3\hbar]^{-2}[\hbar]^{-1}$ is dropped. To obtain them, we use only the three-term identity (3.13):

$$\begin{array}{ccc} & p & \\ p+q & \boxed{u} & p+q \\ & p & \end{array} = \frac{[u+3\hbar]}{[\hbar]},$$

$$\begin{array}{ccc} & p & \\ q-p & \boxed{u} & q-p \\ & p & \end{array} = \frac{[u+\hbar]}{[\hbar]}\frac{[2\lambda_p-2\hbar]}{[2\lambda_p]}\frac{[\lambda_{q-p}+2\hbar]}{[\lambda_{q-p}]},$$

$$\begin{array}{ccc} & p & \\ 0 & \boxed{u} & 0 \\ & p & \end{array} = \frac{[u+2\hbar]}{[\hbar]}\prod_{\substack{r\in\mathcal{P}_1\\ r\neq\pm p}}\frac{[\lambda_{p+r}-\hbar]}{[\lambda_{p+r}]}, \tag{4.15}$$

$$\begin{array}{ccc} & p & \\ 0 & \boxed{u} & q-p \\ & q & \end{array} = \frac{[\lambda_{p-q}-2\hbar-u][\lambda_{p+q}-\hbar]}{[2\lambda_p][\lambda_{q-p}]},$$

$$\begin{array}{ccc} & p & \\ p-q & \boxed{u} & 0 \\ & q & \end{array} = \frac{[2\hbar]}{[\hbar]}\frac{[\lambda_{p-q}-\hbar-u][2\lambda_q-2\hbar]}{[\lambda_{q-p}][\lambda_{p+q}]},$$

$$\begin{array}{ccc} & p & \\ p+q & \boxed{u} & q-p \\ & -p & \end{array} = \frac{[2\hbar]}{[\hbar]}\frac{[2\lambda_p-\hbar-u][\lambda_{p+q}+2\hbar]}{[2\lambda_p][\lambda_{q-p}]}.$$

Finally, we give the example of W_{22}. They are equivalent to the Boltzmam weights associated to the vector representation of the type B_2 Lie algebra (see Ref. 6). We write only two cases as an example, which is used to define the difference operator $M_2(u)$. We will drop the common factor $G(u)$ (2.4) here:

$$\begin{array}{ccc} & p+q & \\ 0 & \boxed{u} & 0 \\ & p+q & \end{array} = \frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}+\hbar]}, \tag{4.16}$$

$$\begin{array}{ccc} & 0 & \\ 0 & \boxed{u} & 0 \\ & 0 & \end{array} = \frac{[2\hbar]}{[6\hbar]}\left(\sum_{\substack{r=\pm\epsilon_1\\ s=\pm\epsilon_2}}\frac{[2\lambda_r+2\hbar][2\lambda_s+2\hbar]}{[2\lambda_r][2\lambda_s]}\frac{[\lambda_{r+s}-5\hbar][\lambda_{r+s}+2\hbar]}{[\lambda_{r+s}][\lambda_{r+s}+\hbar]}-\frac{[u+6\hbar][u-3\hbar]}{[u][u+3\hbar]}\right). \tag{4.17}$$

The formulas (4.15), (4.16), and (4.17) together give the explicit form of $\widetilde{M}_d$ [Theorem 1 (ii)].

We explain how to calculate the fused Boltzmann weight

$$\begin{array}{ccc} & 0 & \\ 0 & \boxed{u} & 0. \\ & 0 & \end{array}$$

According to the definition of the operator $W_{22}(u)$ and the vector f_λ^λ (4.5), the coefficient of $W_{22}(u)f_\lambda^\lambda\otimes f_\lambda^\lambda$ with respect to $f_\lambda^\lambda\otimes f_\lambda^\lambda$ is equal to

$$\frac{1}{[2\lambda_p+2\hbar]}\sum_{r\in\mathcal{P}_1}[2\lambda_r+2\hbar]W_{21}\left(\begin{matrix}\lambda & \lambda\\ \lambda+\hat{p} & \lambda+\hat{r}\end{matrix}\middle| u\right)W_{21}\left(\begin{matrix}\lambda+\hat{p} & \lambda+\hat{r}\\ \lambda & \lambda\end{matrix}\middle| u+\hbar\right). \tag{4.18}$$

In this summation, if r is equal to $-p$, then

$$W_{21}\left(\begin{matrix}\lambda & \lambda\\ \lambda+\hat{p} & \lambda-\hat{p}\end{matrix}\middle| u\right)=0,$$

so that (4.18) can be rewritten as

$$W_{21}\left(\begin{matrix}\lambda & \lambda\\ \lambda+\hat{p} & \lambda+\hat{p}\end{matrix}\middle| u\right)W_{21}\left(\begin{matrix}\lambda+\hat{p} & \lambda+\hat{p}\\ \lambda & \lambda\end{matrix}\middle| u+\hbar\right)$$
$$+\sum_{\substack{q\in\mathcal{P}_1\\ q\neq\pm p}}\frac{[2\lambda_q+2\hbar]}{[2\lambda_p+2\hbar]}W_{21}\left(\begin{matrix}\lambda & \lambda\\ \lambda+\hat{p} & \lambda+\hat{q}\end{matrix}\middle| u\right)W_{21}\left(\begin{matrix}\lambda+\hat{p} & \lambda+\hat{q}\\ \lambda & \lambda\end{matrix}\middle| u+\hbar\right).$$

By means of (4.12), (4.13), and (4.14), this function is equal to

$$\frac{[u-\hbar][u][u+\hbar][u+2\hbar][u+3\hbar][u+4\hbar][2\hbar]}{[-3\hbar]^3[\hbar]^4}$$
$$\times\left(\frac{[u+\hbar][u+2\hbar]}{[2\hbar][-3\hbar]}\prod_{\substack{q\in\mathcal{P}_1\\ q\neq\pm p}}\frac{[\lambda_{p+q}-\hbar][\lambda_{p+q}+2\hbar]}{[\lambda_{p+q}][\lambda_{p+q}+\hbar]}\right.$$
$$\left.+\frac{[\hbar]}{[-3\hbar]}\sum_{\substack{q\in\mathcal{P}_1\\ q\neq\pm p}}\frac{[2\lambda_q-2\hbar]}{[2\lambda_q]}\frac{[\lambda_{p+q}+2\hbar+u][\lambda_{p+q}-\hbar-u][\lambda_{p-q}-\hbar]}{[\lambda_{p+q}][\lambda_{p+q}-\hbar][\lambda_{p-q}+\hbar]}\right).$$

To obtain the formula (4.17), we use the following lemma.

Lemma 2: For any $p\in\mathcal{P}_1$, we have

$$\frac{[u+\hbar][u+2\hbar]}{[2\hbar][-3\hbar]}\prod_{\substack{q\in\mathcal{P}_1\\ q\neq\pm p}}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}\frac{[\lambda_{p+q}+2\hbar]}{[\lambda_{p+q}+\hbar]}$$
$$+\frac{[\hbar]}{[-3\hbar]}\sum_{\substack{q\in\mathcal{P}_1\\ q\neq\pm p}}\frac{[2\lambda_q-2\hbar]}{[2\lambda_q]}\frac{[\lambda_{p+q}+2\hbar+u][\lambda_{p+q}-\hbar-u][\lambda_{p-q}-\hbar]}{[\lambda_{p+q}][\lambda_{p+q}-\hbar][\lambda_{p-q}+\hbar]}$$
$$=\frac{[u][u+3\hbar]}{[6\hbar][-3\hbar]}\sum_{\substack{r=\pm\epsilon_1\\ s=\pm\epsilon_2}}\frac{[2\lambda_r+2\hbar][2\lambda_s+2\hbar]}{[2\lambda_r][2\lambda_s]}\frac{[\lambda_{r+s}-5\hbar][\lambda_{r+s}+2\hbar]}{[\lambda_{r+s}][\lambda_{r+s}+\hbar]}+\frac{[u+6\hbar][u-3\hbar]}{[6\hbar][u+3\hbar]}. \tag{4.19}$$

Proof: Let $f(\lambda_p)$ be (the left-hand side)−(the right-hand side) of (4.19), regarded as a function of λ_p. It is doubly periodic function of the periods $1,\tau$. Let us show that it is entire. The apparent poles of $f(\lambda_p)$ are located at

$$\lambda_p=\lambda_q,\quad \lambda_p=\lambda_q\pm h(p,q\in\mathcal{P}_1,p+q\neq 0),\quad \lambda_p=0(p\in\mathcal{P}_1).$$

Note that the left-hand side of (4.19) is clearly invariant under $\lambda_q \mapsto -\lambda_q$, and the right-hand side is W invariant. In view of the symmetry, it suffices to check the regularity at $\lambda_p=\lambda_q$, $\lambda_p=\lambda_q-\hbar$, and $\lambda_p=0$. By the three-term identity (3.13), it is easy to see that the residue of $f(\lambda_p)$ at $\lambda_p=\lambda_q-\hbar$ vanishes. Manifestly, the point $\lambda_p=\lambda_q$ and $\lambda_p=0$ is regular.

Now we have proved that $f(\lambda_p)$ is independent of λ_p. We will show $f(-\lambda_q-2\hbar)=0$. This can be directly checked by using the identity (3.13) twice, and the proof completes. □

V. COMMUTATIVITY OF THE DIFFERENCE OPERATORS

This section is devoted to the proof of commutativity of the difference operators [Theorem 1(i)]. For $t\in\mathcal{P}_d+\mathcal{P}_{d'}$ we will introduce the matrices $A_t(\lambda|u,v)$, $B_t(\lambda|v,u)$ whose index set is $I_t:=\{(p,q)\in\mathcal{P}_d\times\mathcal{P}_{d'}|p+q=t\}$:

$$A_t(\lambda|u,v)^{(p,q)}_{(r,s)}:=W_{d2}\left(\begin{matrix}\lambda & \lambda+\hat{p}\\ \lambda & \lambda+\hat{r}\end{matrix}\middle| u\right)W_{d'2}\left(\begin{matrix}\lambda+\hat{p} & \lambda+\hat{t}\\ \lambda+\hat{r} & \lambda+\hat{t}\end{matrix}\middle| v\right),$$

$$B_t(\lambda|v,u)^{(p,q)}_{(r,s)}:=W_{d'2}\left(\begin{matrix}\lambda & \lambda+\hat{q}\\ \lambda & \lambda+\hat{s}\end{matrix}\middle| v\right)W_{d2}\left(\begin{matrix}\lambda+\hat{q} & \lambda+\hat{t}\\ \lambda+\hat{s} & \lambda+\hat{t}\end{matrix}\middle| u\right).$$

With these matrices, we can write down both the left- and right-hand sides as

$$M_d(u)M_{d'}(v)=\sum_{t\in\mathcal{P}_d+\mathcal{P}_{d'}}\operatorname{tr}A_t(\lambda|u,v)T_{\hat{t}},\quad M_{d'}(v)M_d(u)=\sum_{t\in\mathcal{P}_d+\mathcal{P}_{d'}}\operatorname{tr}B_t(\lambda|v,u)T_{\hat{t}}.$$

Let us also define the matrix $W_t(\lambda|u-v)$ with the same index set:

$$W_t(\lambda|u-v)^{(p,q)}_{(r,s)}:=W_{dd'}\left(\begin{matrix}\lambda & \lambda+\hat{p}\\ \lambda+\hat{s} & \lambda+\hat{t}\end{matrix}\middle| u-v\right).$$

The YBE (2.3) implies

$$W_t(\lambda|u-v)A_t(\lambda|u,v)=B_t(\lambda|v,u)W_t(\lambda|u-v).$$

By the inversion relation (3.8), it can be seen that $W_t(\lambda|u-v)$ is invertible for generic $u,v\in\mathbb{C}$. It follows that $\operatorname{tr}A_t(\lambda|u,v)=\operatorname{tr}B_t(\lambda|v,u)$ for all $u,v\in\mathbb{C}$. Hence, we have $M_d(u)M_{d'}(v)=M_{d'}(v)M_d(u)$ for all $u,v\in\mathbb{C}$.

VI. SPACE OF WEYL GROUP-INVARIANT THETA FUNCTIONS

This section is devoted to the proof of Theorem 2. Let $Q^\vee$, $P^\vee$ be the coroot and coweight lattice, respectively. Under the identification $\mathfrak{h}=\mathfrak{h}^*$ via the form (,), these are given by

$$Q^\vee=\mathbb{Z}2\epsilon_1\oplus\mathbb{Z}2\epsilon_2,\quad P^\vee=Q^\vee+\mathbb{Z}(\epsilon_1+\epsilon_2).$$

Lemma 3: For all $\beta\in P^\vee$ and d=1,2, we have

$$[S_{\tau\beta},M_d(u)]=[S_\beta,M_d(u)]=0. \tag{6.1}$$

Proof: Note that if $p,q\in\mathcal{P}_1\,(q\neq\pm p)$ and $\beta\in P^\vee$ then $\beta_{p+q}\in\mathbb{Z}$. By the quasiperiodicity (2.2), we have

$$\frac{[(\lambda+\tau\beta)_{p+q}-\hbar]}{[(\lambda+\tau\beta)_{p+q}]}=e^{2\pi i\beta_{p+q}\hbar}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]},\quad \frac{[(\lambda+\beta)_{p+q}-\hbar]}{[(\lambda+\beta)_{p+q}]}=\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}.$$

Using these equations, we have, for all $p\in\mathcal{P}_1$,

$$S_{\tau\beta}\prod_{q\neq\pm p}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}T_{\hat{p}}f(\lambda)=e^{2\pi i((\lambda,\beta)+\tau(\beta,\beta)/2)}\prod_{q\neq\pm p}\frac{[(\lambda+\tau\beta)_{p+q}-\hbar]}{[(\lambda+\tau\beta)_{p+q}]}f(\lambda+\tau\beta+\hat{p})$$
$$=e^{2\pi i((\lambda,\beta)+\tau(\beta,\beta)/2+2\beta_p\hbar)}\prod_{q\neq\pm p}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}f(\lambda+\tau\beta+\hat{p})$$
$$=\prod_{q\neq\pm p}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}T_{\hat{p}}S_{\tau\beta}f(\lambda),$$

and

$$S_{\beta}\prod_{q\neq\pm p}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}T_{\hat{p}}f(\lambda)=\prod_{q\neq\pm p}\frac{[(\lambda+\beta)_{p+q}-\hbar]}{[(\lambda+\beta)_{p+q}]}f(\lambda+\beta+\hat{p})$$
$$=\prod_{q\neq\pm p}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}f(\lambda+\beta+\hat{p})$$
$$=\prod_{q\neq\pm p}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}]}T_{\hat{p}}S_{\beta}f(\lambda).$$

Note that $2\beta_p\hbar=(\hat{p},\beta)$, etc. Hence, we have $[S_{\tau\beta},M_1(u)]=[S_\beta,M_1(u)]=0$. In the same way, we can see that the principal part of $\widetilde{M}_2$ commutes with $S_{\tau\beta}$ and S_β, using the equations

$$\frac{[(\lambda+\tau\beta)_{p+q}-\hbar]}{[(\lambda+\tau\beta)_{p+q}+\hbar]}=e^{2\pi i(2\beta_{p+q}\hbar)}\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}+\hbar]},\quad \frac{[(\lambda+\beta)_{p+q}-\hbar]}{[(\lambda+\beta)_{p+q}+\hbar]}=\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}+\hbar]}.$$

Using (2.2) it is easy to see that the function,

$$C_{p,q}(\lambda):=\frac{[2\hbar]}{[6\hbar]}\frac{[2\lambda_p+2\hbar]}{[2\lambda_p]}\frac{[2\lambda_q+2\hbar]}{[2\lambda_q]}\frac{[\lambda_{p+q}-5\hbar]}{[\lambda_{p+q}+\hbar]}\frac{[\lambda_{p+q}+2\hbar]}{[\lambda_{p+q}]}\quad (p,q\in\mathcal{P}_1,p+q\neq 0),$$

satisfies $C_{p,q}(\lambda+\beta)=C_{p,q}(\lambda+\tau\beta)=C_{p,q}(\lambda)(\forall\beta\in P^\vee)$. This means that $S_{\tau\beta}$, $S_\beta\,(\beta\in P^\vee)$ commute with a multiplication by $C_{p,q}(\lambda)$. □

Lemma 4: For all $\gamma\in P^\vee$, *we have*

$$S_{\tau\gamma}Th^W\subset Th^W,\quad S_\gamma Th^W\subset Th^W. \tag{6.2}$$

Proof: Let $f\in Th^W$ and $\gamma\in P^\vee$. Since the bilinear form (,) is W invariant, we have $(S_{\tau\gamma}f)(w\lambda)=(S_{\tau w^{-1}(\gamma)}f)(\lambda)$. Using (2.5), we can write this as $(S_{\tau\gamma}S_{\tau(w^{-1}(\gamma)-\gamma)}f)(\lambda)$, which is equal to $S_{\tau\gamma}f(\lambda)$ in view of $w^{-1}(\gamma)-\gamma\in Q^\vee$. In the same way, we can show that $(S_\gamma f)(w\lambda)=(S_\gamma f)(\lambda)$.

Evidently, $S_{\tau\gamma}f$ and $S_\gamma f$ are holomorphic. For all $\alpha\in Q^\vee$, using (2.5) and $(\gamma,\alpha)\in\mathbb{Z}$, it can be seen that the operators S_α, $S_{\tau\alpha}$ commute with S_γ, $S_{\tau\gamma}$. Hence, $S_{\tau\gamma}f$ or $S_\gamma f$ are fixed by $S_{\tau\alpha}$ and S_α. □

Here we prove Theorem 2.

Proof of Theorem 2: Let f be any function in Th^W. In view of (6.1), we have $S_\alpha\widetilde{M}_d f=S_{\tau\alpha}\widetilde{M}_d f=\widetilde{M}_d f$ for all $\alpha\in Q^\vee\subset P^\vee$. It is clear from the explicit form of $\widetilde{M}_d$ that $\widetilde{M}_d f(w\lambda)=\widetilde{M}_d f(\lambda)$ for all $w\in W$.

Let us show that the function $\widetilde{M}_d f$ is holomorphic on $\mathfrak{h}^*$. For $\mu\in\mathfrak{h}^*$ and $z\in\mathbb{C}$, we denote by D_μ^z the line in $\mathfrak{h}^*$, defined by

$$D_\mu^z:=\{\lambda\in\mathfrak{h}^*|(\lambda,\mu)+z=0\}.$$

The coefficients of the difference operators $\tilde{M}_d$ have their possible simple poles along $D+P^\vee+\tau P^\vee$, where we put

$$D:=\bigcup_{p\in R_+} D_p^0 \ \cup \bigcup_{q\in\mathcal{P}_2-\{0\}} D_q^\hbar,$$

and R_+ is a fixed set of positive roots.

Next we will show that for any function f in Th^W, $\tilde{M}_d f$ is regular along D. Let us consider the meromorphic function $g:=(\Pi_{p\in R_+}[\lambda_p])\,\tilde{M}_d f$, which is regular along $D^0:=\cup_{p\in R_+}D_p^0$. Since $\tilde{M}_d f$ is W invariant, it is clear that g is W anti-invariant. This implies that g has zero along D^0 and hence $\tilde{M}_d f$ is regular along D^0.

The holomorphy along $\cup_{q\in\mathcal{P}_2-\{0\}}D_q^\hbar$ is somewhat nontrivial. Let $p=\pm\epsilon_1$, $q=\pm\epsilon_2$. Clearly, $\tilde{M}_1 f$ is regular along $D_{p+q}^\hbar$. Let us consider the function $\tilde{M}_2 f$. It suffices to show that the following function is regular along $D_{p+q}^\hbar$:

$$\frac{[\lambda_{p+q}-\hbar]}{[\lambda_{p+q}+\hbar]}T_{\hat{p}}T_{\hat{q}}f(\lambda)+\frac{[2\hbar]}{[6\hbar]}\frac{[2\lambda_p+2\hbar]}{[2\lambda_p]}\frac{[2\lambda_q+2\hbar]}{[2\lambda_q]}\frac{[\lambda_{p+q}-5\hbar]}{[\lambda_{p+q}+\hbar]}\frac{[\lambda_{p+q}+2\hbar]}{[\lambda_{p+q}]}f(\lambda).$$

We note that, for any W-invariant function f, we have $(T_{\hat{p}}T_{\hat{q}}f-f)|_{D_{p+q}^\hbar}=0$. In view of this, the residue of the above function along $D_{p+q}^\hbar$ is easily seen to vanish. Thus, we have proved that for any function f in Th^W, the functions $\tilde{M}_d f(d=1,2)$ are regular along D.

For $\beta,\gamma\in P^\vee$, we have, by the definitions of $S_{\tau\beta}$, S_γ, and (6.1),

$$\tilde{M}_d f(\lambda+\beta\tau+\gamma)=e^{-2\pi i((\lambda,\beta)+\tau(\beta,\beta)/2)}S_{\tau\beta}S_\gamma\tilde{M}_d f(\lambda)=e^{-2\pi i((\lambda,\beta)+\tau(\beta,\beta)/2)}\tilde{M}_d S_{\tau\beta}S_\gamma f(\lambda). \tag{6.3}$$

Since $S_{\tau\beta}S_\gamma f$ belongs to Th^W by (6.2), $\tilde{M}_d S_{\tau\beta}S_\gamma f$ is regular along D. Then (6.3) implies that $\tilde{M}_d f$ is regular along $D+\beta\tau+\gamma$. The proof is completed. □

ACKNOWLEDGMENTS

The authors are grateful to Masato Okado and Michio Jimbo who kindly let them know detailed points in the proof of the YBE of the Boltzmann weight. They also thank Toshiki Nakashima for useful information and Gen Kuroki, Masatoshi Noumi, Hiroyuki Ochiai, Yasuhiko Yamada, Yasushi Komori, and Kazuhiro Hikami for fruitful discussions and kind interest. Much of the work was done when T. I. was a postdoctral student at the Mathematical Institute, Tohokn University. He would like to thank the staff of the Institute for their support.

APPENDIX: SIMILARITY TRANSFORMATION

Our Boltzmann weights in Sec. III and the original form in Ref. 6 are slightly different. The original form of type (3.3) and (3.4) are given as follows:

$$p\underset{q}{\overset{q}{\boxed{u}}}p=\frac{[c-u][u]}{[c][\hbar]}\left(\frac{[\lambda_{p-q}+\hbar][\lambda_{p-q}-\hbar]}{[\lambda_{p-q}]^2}\right)^{1/2}\quad(p\neq\pm q), \tag{A1}$$

$$p\underset{-p}{\overset{q}{\boxed{u}}}-q=\frac{[u][\lambda_{p+q}+\hbar+c-u]}{[c][\lambda_{p+q}+\hbar]}(G_{\lambda p}G_{\lambda q})^{1/2}\quad(p\neq q). \tag{A2}$$

All the other Boltzmann weights [(3.1),(3.2),(3.5)] are the same as the ones we adopted in Sec. III. We denote these weights by

$$W_{JMO}\left(\begin{matrix}\lambda & \mu \\ \kappa & \nu\end{matrix}\middle| u\right).$$

Our Boltzmann weights are obtained from those by the following way. We introduce an ordering on the set $\mathcal{P}$ as

$$\epsilon_1 \prec \epsilon_2 \prec \cdots \prec \epsilon_n \prec -\epsilon_n \prec \cdots \prec -\epsilon_2 \prec -\epsilon_1 .$$

For $\lambda, \mu \in \mathfrak{h}^*$, such that $\mu - \lambda = \hat{q} \in 2\hbar\mathcal{P}$, we define the function $s(\lambda,\mu)$ by

$$s(\lambda,\mu) := \prod_{\substack{p\in\mathcal{P} \\ p\prec q}} [\lambda_{p-q}]^{-1/2}[\mu_{p-q}]^{-1/2}. \tag{A3}$$

The relation between the Boltzmann weights W in Sec. III and the ones in Ref. 6 is as follows:

$$W\left(\begin{matrix}\lambda & \mu \\ \kappa & \nu\end{matrix}\middle| u\right) = \frac{s(\lambda,\mu)s(\mu,\nu)}{s(\lambda,\kappa)s(\kappa,\nu)} W_{JMO}\left(\begin{matrix}\lambda & \mu \\ \kappa & \nu\end{matrix}\middle| u\right). \tag{A4}$$

[1] K. Hasegawa, ''On the crossing symmetry of the elliptic solution of the Yang–Baxter equation and a new *L* operator for Belavin's solution,'' J. Phys. A **26**, 3211–3228 (1993).

[2] A. A. Belavin, ''Dynamical symmetry of integrable quantum systems,'' Nucl. Phys. B **180**[FS2], 189–200 (1981).

[3] K. Hasegawa, ''Ruijsenaars' Commuting difference operators as commuting transfer matrices,'' Commun. Math. Phys. **187**, 289–325 (1997).

[4] S. N. M. Ruijsenaars, ''Complete integrability of relativistic Calogero–Moser systems and elliptic function identities,'' Commun. Math. Phys. **110**, 191–213 (1987).

[5] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed. (Oxford University Press, Oxford, 1995).

[6] M. Jimbo, T. Miwa, and M. Okado, ''Solvable lattice models related to the vector representation of classical simple Lie algebras,'' Commun. Math. Phys. **116**, 507–525 (1988).

[7] R. J. Baxter, ''Eight-vertex model in lattice statistics and one-dimensional anisotropic Heisenberg chain,'' Ann. Phys. (N.Y.) **76**, 1–24 (1973); **76**, 25–47 (1973); **76**, 48–71 (1973).

[8] M. Jimbo, T. Miwa, and M. Okado, ''Local state probabilities of solvable lattice models: An $A_n^{(1)}$ family,'' Nucl. Phys. B **300**[FS22], 74–108 (1988).

[9] V. G. Kac and D. H. Peterson, ''Infinite-dimensional Lie algebras, theta functions and modular forms,'' Adv. Math. **53**, 125–264 (1984).

[10] G. Felder and A. Varchenko, ''Elliptic quantum groups and Ruijsenaars models,'' J. Stat. Phys. **89**, 963–980 (1997).

[11] T. H. Koornwinder, ''Askey–Wilson polynomials for root systems of type *BC*,'' Contemp. Math. **138**, 189–204 (1992).

[12] J. F. van Diejen, ''Commuting difference operators with polynomial eigenfunctions,'' Compositio Mathematica **95**, 183–233 (1995).

[13] J. F. van Diejen, ''Integrability of difference Calogero–Moser systems,'' J. Math. Phys. **35**, 2983–3004 (1994).

[14] Y. Komori and K. Hikami, ''Quantum integrability of the generalized elliptic Ruijsenaars models,'' J. Phys. A **30**, 4341–4364 (1997).

[15] Y. Komori and K. Hikami, ''Conserved operators of the generalized elliptic Ruijsenaars models,'' J. Math. Phys. **39**, 6175–6190 (1998).

[16] Y. Shibukawa and K. Ueno, ''Completely $\mathbb{Z}$ symmetric *R*-matrix,'' Lett. Math. Phys. **25**, 239–248 (1992).

[17] Y. Komori and K. Hikami, ''Notes on operator-valued solutions of the Yang–Baxter equation and the reflection equation,'' Mod. Phys. Lett. A **11**, 2861–2870 (1996).

[18] Y. Komori and K. Hikami, ''Elliptic *K*-matrix associated with Belavin's symmetric *R*-matrix,'' Nucl. Phys. B **494**, 687–701 (1997).

[19] I. Cherednik, ''Double affine Hecke algebras and Macdonald's conjectures,'' Ann. Math. **141**, 191–216 (1995).

[20] M. Noumi, ''Macdonald–Koornwinder polynomials and affine Hecke rings (Japanese),'' *Various Aspects of Hypergeometric Functions (Japanese)*, Kyoto, 1994. Sūrikaisekikenkyūsho Kōkyūroku No. 919, 1995, pp. 44–55.

[21] K. Hasegawa, ''*L*-operator for Belavin's *R*-matrix acting on the space of theta functions,'' J. Math. Phys. **35**, 6158–6171 (1994).

Extended Jordanian twists for Lie algebras

P. P. Kulish
St. Petersburg Department of the Steklov Mathematical Institute, Fontanka 27, St. Petersburg 191011, Russia

V. D. Lyakhovsky
Department of Physics, St. Petersburg State University, Ulianovskaya 1, Petrodvorets, St. Petersburg 198904, Russia

A. I. Mudrov
Institute of Physics, St. Petersburg State University, Ulianovskaya 1, Petrodvorets, St. Petersburg 198904, Russia

(Received 24 June 1998; accepted for publication 9 April 1999)

Jordanian quantizations of Lie algebras are studied using the factorizable twists. For a restricted Borel subalgebra $\mathbf{B}^\vee$ of $\mathrm{sl}(N)$ the explicit expressions are obtained for the twist element $\mathcal{F}$, universal $\mathcal{R}$-matrix and the corresponding canonical element $\mathcal{T}$. It is shown that the twisted Hopf algebra $\mathcal{U}_{\mathcal{F}}(\mathbf{B}^\vee)$ is self-dual. The cohomological properties of the involved Lie bialgebras are studied to justify the existence of a contraction from the Dinfeld–Jimbo quantization to the Jordanian one. The construction of the twist is generalized to a certain type of inhomogenious Lie algebras. © *1999 American Institute of Physics.* [S0022-2488(99)02707-3]

I. INTRODUCTION

The thorough formulation of the theory of quantum groups by Drinfeld[1] includes two types of Hopf algebras: triangular (with the universal $\mathcal{R}$-matrix satisfying the relation $\mathcal{R}_{21}\mathcal{R}=1$) and quasitriangular (with $\mathcal{R}_{21}\mathcal{R}\neq 1$). Deformations of universal enveloping of simple Lie algebras initiated by the quantum inverse scattering method and discovered by Drinfeld and Jimbo[1,2] belong to the latter class. In the framework of the deformation quantization theory[3] these quantum algebras correspond to Lie bialgebras with classical r-matrix

$$r_{\mathrm{DJ}}=\sum_{i=1}^{k} t_{ij}H_i\otimes H_j+\sum_{\alpha\in\Phi_+} E_\alpha\otimes E_{-\alpha},$$

where k is the rank, t_{ij} is the inverse Cartan matrix, and Φ_+ is the set of positive roots. This r_{DJ} is one of the multitude of solutions to the classical Yang–Baxter equation. The detailed classification of solutions was performed for simple Lie algebras in Ref. 4, only for some of these classical r-matrices the corresponding quantum R-matrices are known explicitly.

Although the existence of quantization for any Lie bialgebra is now proved,[5] the explicit knowledge of the R-matrix as an algebraic element $\mathcal{R}$ or a matrix in some irreducible representations is required in the FRT approach[6] and in a variety of applications of quantum groups. One can mention the universal R-matrix of the quantum algebra $\mathcal{U}_q(\mathrm{sl}(2))$ (Ref. 1) which is a building block for the universal R-matrices for other simple Lie and Kac–Moody algebras. As about triangular quantum groups and twistings,[7,8] the well-known example is the Jordanian quantization of sl(2) or, more exactly, of its Borel subalgebra $\mathbf{B}_+(\{h,x|[h,x]=2x\})$ with $r=h\otimes x-x\otimes h=h\wedge x$ (Ref. 1) and the triangular R-matrix $\mathcal{R}=\mathcal{F}_{21}\mathcal{F}^{-1}$ defined by the twisting element[9,10]

$$\mathcal{F}=\exp\{\tfrac{1}{2}h\otimes\ln(1+2\xi x)\}. \tag{1}$$

This quantum algebra $\mathcal{U}_\xi(\mathrm{sl}(2))$ also found numerous applications from the deformed Heisenberg XXX-spin chain to the quantum Minkowski space (see e.g., Ref. 11) and in a few other cases.[12,13]

In the present paper we propose various extensions of this twist element. The suggested construction implies the existence (in the universal enveloping algebra to be deformed) of a subalgebra **L** with special properties of multiplication. This is a solvable subalgebra with at least four generators. All simple Lie algebras except sl(2) contain such **L** and in any of them a deformation induced by twist of **L** can be performed. In particular we study a Jordanian deformation of $\mathcal{U}(\mathrm{sl}(N))$, reaching a closed form of deformed compositions lacking in Ref. 9. Using the notion of factorizible twist,[14] we prove that the element $\mathcal{F}\in\mathcal{U}(\mathrm{sl}(N))^{\otimes 2}$,

$$\mathcal{F}=\exp\left\{2\xi\sum_{i=2}^{N-1} E_{1i}\otimes E_{iN}e^{-\sigma}\right\}\exp\{H\otimes\sigma\},\tag{2}$$

where $x=E_{1N}$, $H=E_{11}-E_{NN}$, $\sigma=\frac{1}{2}\ln(1+2\xi x)$, satisfies the twist equation. Hence, it defines a triangular deformation of $\mathcal{U}(\mathrm{sl}(N))$. In such Hopf algebras deformed by Jordanian twist the subset of Cartan generators $\{E_{ii}-E_{jj}\}$ with $i<j$, $i,j\neq 1,N$, remains untouched. Hence there is a possibility to perform additional multiparametric deformation using the Reshetikhin twist.[15]

The main ingredients of the quantum group theory[1] are constructed: the universal R-matrix, the dual Hopf algebra [quantized function algebra on $\mathrm{SL}(N)$], the universal $\mathcal{T}$-matrix (canonical element) for the subalgebra which induces the twist of $\mathcal{U}(\mathrm{sl}(N))$ and the self-duality of **L**. Cohomological interpretation of the interrelation between the Drinfeld–Jimbo (or standard) quantum algebra $\mathcal{U}_q(\mathrm{sl}(N))$ and the Jordanian (or nonstandard) one $\mathcal{U}_\xi(\mathrm{sl}(N))$ is discussed. The real form and the corresponding quantum linear space are given. We present also further generalization in which the subalgebra **L** is substituted by a certain type of inhomogeneous Lie algebras.

The connection of the Drinfeld–Jimbo deformation[1,2] with the Jordanian deformation was already pointed out in Ref. 9. The similarity transformation of the classical matrix r_{DJ} performed by the operator $\exp(\xi\mathrm{ad}E_{1N})$ (with the highest root generator E_{1N}) turns r_{DJ} into the sum $r_{\mathrm{DJ}}+\xi r_j$.[9] Hence,

$$r_j=-\xi\left(H_{1N}\wedge E_{1N}+2\sum_{k=2}^{N-1} E_{1k}\wedge E_{kN}\right)\tag{3}$$

is a classical r-matrix too, which defines the corresponding deformation. A contraction of the quantum plane $xy=qyx$ of $\mathcal{U}_q(\mathrm{sl}(2))$ with the above-mentioned transformation in the fundamental representation $M=1+\theta\rho(E_{1N})$, $\theta=\xi(1-q)^{-1}$, results in the Jordanian plane $x'y'=y'x'+\xi y'^2$ of $\mathcal{U}_\xi(\mathrm{sl}(2))$.[10] Later, this contraction in the fundamental representation of sl(3) and sl(N) was used in many papers (cf. Refs. 16 and 17 and references therein). Let us point out that in our formulas we do not refer to any particular representation of deformed algebras.

The paper is organized as follows. After recalling briefly the basic material on twisting of Hopf algebras (Sec. II), we construct an extended Jordanian twist $\mathcal{F}$ for a four-generator Lie algebra and apply it to twist the universal enveloping algebra $\mathcal{U}(\mathrm{sl}(N))$ (Sec. III). The next section contains a cohomological explanation of the connection between the Drinfeld–Jimbo and Jordanian quantization. The main objects of the theory of quantum groups are constructed in Sec. V. Further generalization of the extended Jordanian twist to a special class of inhomogenious Lie algebras and possible research topics are given in Sec. VI and in the Conclusion.

II. TWISTING OF HOPF ALGEBRAS

A Hopf algebra $\mathcal{A}(m,\Delta,\epsilon,S)$ with multiplication $m:\mathcal{A}\otimes\mathcal{A}\rightarrow\mathcal{A}$, coproduct $\Delta:\mathcal{A}\rightarrow\mathcal{A}\otimes\mathcal{A}$, counit $\epsilon:\mathcal{A}\rightarrow C$, and antipode $S:\mathcal{A}\rightarrow\mathcal{A}$ (see definitions in Refs. 1, 6, and 18) can be transformed[7] with an invertible element $\mathcal{F}\in\mathcal{A}\otimes\mathcal{A}$, $\mathcal{F}=\sum f_i^{(1)}\otimes f_i^{(2)}$ into a twisted one $\mathcal{A}_t(m,\Delta_t,\epsilon,S_t)$. This Hopf algebra $\mathcal{A}_t$ has the same multiplication and counit maps but the twisted coproduct and antipode:

$$\Delta_t(a)=\mathcal{F}\Delta(a)\mathcal{F}^{-1},\quad S_t(a)=vS(a)v^{-1},\quad v=\sum f_i^{(1)}S(f_i^{(2)}),\quad a\in\mathcal{A}.$$

The twisting element has to satisfy the equations

$$(\epsilon\otimes\mathrm{id})(\mathcal{F})=(\mathrm{id}\otimes\epsilon)(\mathcal{F})=1, \tag{4}$$

$$\mathcal{F}_{12}(\Delta\otimes\mathrm{id})(\mathcal{F})=\mathcal{F}_{23}(\mathrm{id}\otimes\Delta)(\mathcal{F}), \tag{5}$$

where the first one is just a normalizing condition and follows from the second relation modulo a nonzero scalar factor.

A quasitriangular Hopf algebra $\mathcal{A}(m,\Delta,\epsilon,S,\mathcal{R})$ has additionally an element $\mathcal{R}\in\mathcal{A}\otimes\mathcal{A}$ (a universal R-matrix) satisfying[1]

$$(\Delta\otimes\mathrm{id})(\mathcal{R})=\mathcal{R}_{13}\mathcal{R}_{23},\quad (\mathrm{id}\otimes\Delta)(\mathcal{R})=\mathcal{R}_{13}\mathcal{R}_{12}. \tag{6}$$

The coproduct Δ and its opposite Δ^{op} are related by the similarity transformation (twisting) with $\mathcal{R}$

$$\Delta^{\mathrm{op}}(a)=\mathcal{R}\Delta(a)\mathcal{R}^{-1},\quad a\in\mathcal{A},$$

and in this case the relation (5) is just the Yang–Baxter equation.

A twisted quasitriangular quantum algebra $\mathcal{A}_t(m,\Delta_t,\epsilon,S_t,\mathcal{R}_t)$ has the twisted universal R-matrix

$$\mathcal{R}_t=\tau(\mathcal{F})\mathcal{R}\mathcal{F}^{-1}, \tag{7}$$

where τ means the permutation of the tensor factors: $\tau(f\otimes g)=(g\otimes f)$, $\tau(\mathcal{F})=\mathcal{F}_{21}$.

Although, in principle, the possibility to quantize an arbitrary Lie bialgebra has been proved,[5] an explicit formulation of Hopf operations remains a nontrivial task. In particular, the knowledge of the explicit form of the twisting cocycle is a rare case even for classical universal enveloping algebras, despite the advanced Drinfeld theory.[8] Most of the such explicitly known twisting elements have the factorization property with respect to comultiplication [cf. (6)],

$$(\Delta\otimes\mathrm{id})(\mathcal{F})=\mathcal{F}_{23}\mathcal{F}_{13}\quad\text{or}\quad(\Delta\otimes\mathrm{id})(\mathcal{F})=\mathcal{F}_{13}\mathcal{F}_{23},$$

and a similar property involving $(\mathrm{id}\otimes\Delta)$. To satisfy the twist equation, these identities are to be combined with additional requirement $\mathcal{F}_{12}\mathcal{F}_{23}=\mathcal{F}_{23}\mathcal{F}_{12}$ or the Yang–Baxter equation on $\mathcal{F}$.[14,15]

An important subclass of factorizable twists consists of elements satisfying the following equations:

$$(\Delta\otimes\mathrm{id}(\mathcal{F})=\mathcal{F}_{13}\mathcal{F}_{23}, \tag{8}$$

$$(\mathrm{id}\otimes\Delta_t)(\mathcal{F})=\mathcal{F}_{12}\mathcal{F}_{13}. \tag{9}$$

It is easy to see that the universal R-matrix $\mathcal{R}$ satisfies these equations for $\Delta_t=\Delta^{\mathrm{op}}$. Another well-developed case is the Jordanian twist of sl(2) with $\mathcal{F}$ described by (1).[10] Due to the fact that the Cartan element h is primitive in $\mathrm{sl}(2):\Delta(h)=h\otimes 1+1\otimes h$, and σ is primitive in the Jordanian $\mathcal{U}_\xi(\mathrm{sl}(2)):\Delta_t(\sigma)=\sigma\otimes 1+1\otimes\sigma$, one gets

$$(\Delta\otimes\mathrm{id})e^{h\otimes\sigma}=e^{h\otimes 1\otimes\sigma}e^{1\otimes h\otimes\sigma},$$

$$(\mathrm{id}\otimes\Delta_t)e^{h\otimes\sigma}=e^{h\otimes\sigma\otimes 1}e^{h\otimes 1\otimes\sigma}.$$

It will be shown in the next section that the element $\mathcal{F}(2)$ also satisfies the factorization equations (8) and (9) and can be used to twist the universal enveloping algebra of $\mathrm{sl}(N)$.

Let us mention that the composition of appropriate twists can be defined $\mathcal{F}=\mathcal{F}_2\mathcal{F}_1$. The element $\mathcal{F}_1$ has to satisfy the twist equation with the coproduct of the original Hopf algebra, while $\mathcal{F}_2$ must be its solution for Δ_{t_1} of the intermediate Hopf algebra twisted by $\mathcal{F}_1$. In particular, if $\mathcal{F}$ is a solution to the twist equation (5), then $\mathcal{F}^{-1}$ satisfies this equation with $\Delta\rightarrow\Delta_t$.

III. FACTORIZABLE TWISTS

Now we shall propose a new factorizable twist similar to (1) and defined on the abstract set of generators.

Let $\mathbf{L}$ be a four-dimensional Lie algebra with generators $\{H,A,B,E\}$ containing $\mathbf{B}_+$ and representable in a form of semidirect sum of one-dimensional space V_H with basic element H and a Heisenberg subalgebra $\mathcal{H}(A,B,E)$:$\mathbf{L}=V_H\triangleright\mathcal{H}$:

$$\begin{gathered}[H,E]=2E,\\ [H,A]=\alpha A,\quad [H,B]=\beta B,\quad \alpha+\beta=2,\\ [E,A]=[E,B]=0,\\ [A,B]=\gamma E.\end{gathered}\tag{10}$$

Extending the twist deformation $\mathcal{U}_t(\mathbf{B}_+)$ performed by

$$\Phi=\exp\left(\tfrac{1}{2}H\otimes\ln(1+\gamma E)\right)=e^{H\otimes\sigma}$$

to the universal enveloping $\mathcal{U}(\mathbf{L})$, one gets the twisted algebra $\mathcal{U}_\Phi(\mathbf{L})$. It retains the initial multiplication defined by (10) while its coproduct $\Delta_\Phi=\Phi\Delta\Phi^{-1}$ becomes noncocommutative:

$$\begin{gathered}\Delta_\Phi(H)=H\otimes e^{-2\sigma}+1\otimes H,\\ \Delta_\Phi(A)=A\otimes e^{\alpha\sigma}+1\otimes A,\\ \Delta_\Phi(B)=B\otimes e^{\beta\sigma}+1\otimes B,\\ \Delta_\Phi(E)=E\otimes e^{2\sigma}+1\otimes E,\end{gathered}\tag{11}$$

We shall show that the algebra $\mathcal{U}(\mathbf{L})$ allows a more complicated twist deformation containing Φ as a factor.

Theorem 1: *The element*

$$\mathcal{F}=\Phi\Phi_1=\exp(H\otimes\sigma)\exp(A\otimes Be^{-2\sigma})\tag{12}$$

is a twist for $\mathcal{U}(\mathbf{L})$.

Proof: We shall show that $\mathcal{F}=\Phi\Phi_1$ belongs to the subclass defined by the equations (8) and (9). The equation (8) is obviously true: H and A are the primitive elements and B commutes with σ in $\mathcal{U}(\mathbf{L})$. To check the second equation (9) let us consider the coproducts $\Delta_{\mathcal{F}}(\sigma)$ and $\Delta_{\mathcal{F}}(B)$. It is known that in twisted (by Φ) universal enveloping of a Borel subalgebra the element σ is primitive.[10] The element σ commutes not only with B, but also with A, so σ remains primitive with respect to $\Delta_{\mathcal{F}}$. Using the properties of ''roots'' $\alpha-2=-\beta$, the twisted coproduct of B can be written in the following form:

$$\begin{aligned}\Delta_{\mathcal{F}}(B)&=\exp\left(\mathrm{ad}(A\otimes Be^{-\beta\sigma})\right)\circ\exp\left(\mathrm{ad}(H\otimes\sigma)\right)\circ(B\otimes 1+1\otimes B)\\ &=\exp\left(\mathrm{ad}(A\otimes Be^{-\beta\sigma})\right)\circ(B\otimes e^{\beta\sigma}+1\otimes B)\\ &=\exp\left(\mathrm{ad}(A\otimes Be^{-\beta\sigma})\right)\circ(B\otimes e^{\beta\sigma})+1\otimes B.\end{aligned}$$

From (10) one can see that $(\mathrm{ad}_A)^2 \circ B = 0$. So the obtained expression can be simplified,

$$\Delta_{\mathcal{F}}(B) = B \otimes e^{\beta\sigma} + (1 + [A,B]) \otimes B = B \otimes e^{\beta\sigma} + e^{2\sigma} \otimes B.$$

Now using the coproduct

$$\Delta_{\mathcal{F}}(Be^{-2\sigma}) = Be^{-2\sigma} \otimes e^{-\alpha\sigma} + 1 \otimes Be^{-2\sigma},$$

one can easily see that

$$\exp(\mathrm{ad}\,(H \otimes 1 \otimes \sigma)) \circ (A \otimes Be^{-2\sigma} \otimes e^{-\alpha\sigma}) = A \otimes Be^{-2\sigma} \otimes 1.$$

The latter guarantees the validity of the equation (9) for the twisting element $\mathcal{F}$. ●

The deformed algebra $\mathcal{U}_{\mathcal{F}}(\mathbf{L})$ has initial commutation relations generated by (10) and twisted coproducts:

$$\begin{aligned} \Delta_{\mathcal{F}}(H) &= H \otimes e^{-2\sigma} + 1 \otimes H - 2A \otimes Be^{(\alpha-4)\sigma}, \\ \Delta_{\mathcal{F}}(A) &= A \otimes e^{-\beta\sigma} + 1 \otimes A, \\ \Delta_{\mathcal{F}}(B) &= B \otimes e^{\beta\sigma} + e^{2\sigma} \otimes B, \\ \Delta_{\mathcal{F}}(E) &= E \otimes e^{2\sigma} + 1 \otimes E. \end{aligned} \tag{13}$$

Let us rewrite the twisting element $\mathcal{F}$ in the reverse order:

$$\mathcal{F} = \widetilde{\Phi_1}\Phi = \exp(A \otimes Be^{-\beta\sigma}) \exp(H \otimes \sigma) \tag{14}$$

Now we know that both $\mathcal{F}$ and Φ are twists for $\mathcal{U}(\mathbf{L})$ and both satisfy the equations (8) and (9). Hence $\widetilde{\Phi_1}$ is also a twist element with respect to the algebra $\mathcal{U}_{\Phi}(\mathbf{L})$. Using the coalgebra relations (11), it is easy to check that $\widetilde{\Phi_1}$ satisfies the general twist equation (5),

$$(\widetilde{\Phi_1})_{12}(\Delta_{\Phi} \otimes \mathrm{id})\widetilde{\Phi_1} = (\widetilde{\Phi_1})_{23}(\mathrm{id} \otimes \Delta_{\Phi})\widetilde{\Phi_1}.$$

Note that, contrary to the properties of $\mathcal{F}$ and Φ, this twist $(\widetilde{\Phi_1})$ does not belong to the subclass of factorizable twists defined by the equations (8) and (9).

Subalgebras of the type $\mathbf{L}$ exist in a large class of Lie algebras. They can also be found in any simple Lie algebra of rank greater than 1. Such simple algebras contain at least one pair of roots λ_1 and λ_2 such that $\lambda_3 = \lambda_1 + \lambda_2$ is also a root. The corresponding generators X_1, X_2, X_3 together with the Cartan element H_3 dual to the root λ_3 form the subalgebra equivalent to $\mathbf{L}$. As we have shown above such a subalgebra can be twisted with the element $\mathcal{F}$ and the corresponding deformation can be extended to the whole algebra $\mathcal{U}$ and its twisted version $\mathcal{U}_{\mathcal{F}}$ can be thus constructed.

We shall demonstrate the deformations generated by these twists in the case of simple algebras of series A_{N-1}. For our purposes it will be convenient to use the canonical basis of $\mathrm{gl}(N)$ for the compositions of $\mathcal{U}(\mathrm{sl}(N))$

$$[E_{ik}, E_{lm}] = \delta_{kl}E_{im} - \delta_{im}E_{lk}, \quad i,k,l,m = 1,\ldots N. \tag{15}$$

The Cartan elements of $\mathcal{U}(\mathrm{sl}(N))$ will be fixed as $H_{ik} = E_{ii} - E_{kk}$.

Let $H \in \mathbf{L}$ be identified with the Cartan element dual to the highest root of $\mathrm{sl}(N)$. This root will be denoted by λ_H. Collecting all the pairs of roots with the property $\lambda_H = \lambda_1 + \lambda_2$ one can get the multiparametric twist of the type $\mathcal{F}$ with

$$H=H_{1N},\quad E=E_{1N},$$
$$A=\sum_k \,(a^{1k}E_{1k}+a^{kN}E_{kN}), \tag{16}$$

$$B=\sum_k \,(b^{1k}E_{1k}+b^{kN}E_{kN}),$$
$$\{k=2,...,N-1;a^{mn},b^{nm}\in \mathbb{C}\}. \tag{17}$$

Here it is convenient to put $\gamma=2\xi$,

$$\sigma(E)=\tfrac{1}{2}\ln(1+2\xi E).$$

In these terms the consistency condition would take the form

$$[A,B]=e^{2\sigma}-1=2\xi E, \tag{18}$$

and the only nontrivial commutator of σ with the basic elements of **L** is

$$[H,\sigma]=1-e^{-2\sigma}. \tag{19}$$

According to Theorem 1, the element

$$\mathcal{F}=\Phi\Phi_1=\exp(H\otimes\sigma)\exp(A\otimes Be^{-2\sigma}) \tag{20}$$

is twisting for $\mathcal{U}(\mathrm{sl}(N))$. Using the particular properties of **L** one can apply the Cambell–Hausdorff formula to rewrite this twisting element in the following form:

$$\mathcal{F}=\exp(A\otimes Be^{-\sigma})\exp(H\otimes\sigma)=\exp(H\otimes\sigma+A\otimes B\sigma e^{-2\sigma}(1-e^{-\sigma})^{-1}). \tag{21}$$

Note: Any number of factors of the type Φ_1 can appear in the expression (20):

$$\mathcal{F}=\Phi\prod_j \Phi_j=\exp(H\otimes\sigma)\prod_j \exp(A_j\otimes B_j e^{-2\sigma}) \tag{22}$$

with A_j and B_j as in (17) and (16) and subject to the additional conditions

$$[A_{j1},A_{j2}]=[B_{j1},B_{j2}]=0,$$

while the correlation equation (18) takes the form

$$[A_j,B_k]=\delta_{jk}(e^{2\sigma}-1). \quad \bullet \tag{23}$$

Using the twist (20) with the sole factor Φ_1 one gets the maximal number of free parameters—the relation (18) imposes the only condition on the coefficients a's and b's,

$$\sum_{k=2}^{N-1}(a^{1k}b^{kN}-a^{kN}b^{1k})=2\xi. \tag{24}$$

On the contrary, supplying $\mathcal{F}$ with the maximal number $(N-2)$ of factors Φ_j one gets the $(N-2)^2$ conditions (23). In particular, one can satisfy $(N-2)(N-3)$ of these conditions using the basic relations (15) and the specific choice of A_j and B_j (one root λ_j for each factor Φ_j):

$$A_j=a^{1j}E_{1j}+a^{jN}E_{jN}\quad B_j=(b^{1j}E_{1j}+b^{jN}E_{jN}),\quad (\text{with no summation on } j). \tag{25}$$

Here the essential relations rest

$$a^{1j}b^{jN}-a^{jN}b^{1j}=2\xi \quad \{j=2,...N-1\}. \tag{26}$$

The equation (24) [as well as (26)] shows that it is natural to renormalize the element A (or the elements A_j) putting

$$A=2\xi\tilde{A}$$

so that

$$[\tilde{A},B]=E.$$

In these notations the twisting elements

$$\mathcal{F}=\exp(H\otimes\sigma)\exp(2\xi\tilde{A}\otimes Be^{-2\sigma}), \tag{27}$$

$$\mathcal{F}=\exp(H\otimes\sigma)\prod_{j}\exp(2\xi\widetilde{A_j}\otimes B_j e^{-2\sigma}) \tag{28}$$

have the trivial limit $\lim_{\xi\to 0}\mathcal{F}=1$. So does the universal R-matrix ($\mathcal{R}=\mathcal{F}_{21}\mathcal{F}^{-1}$) and one can easily write down the corresponding classical r-matrices

$$r=-(H\wedge E+2\tilde{A}\wedge B) \tag{29}$$

or

$$r=-\left(H\wedge E+2\sum \tilde{A}_j\wedge B_j\right). \tag{30}$$

Their form clearly indicates that twisting by $\mathcal{F}$ corresponds to the quantization of the self-dual Lie bialgebra ($\mathbf{L},\mathbf{L}^*\approx\mathbf{L}$) just as in the case of the Jordanian twist of $\mathbf{B}(1)$.[10,12] The same is true for the twisted Hopf algebra $\mathcal{U}_{\mathcal{F}}(\mathbf{L})$, it is self-dual. We shall discuss this property in the next section and prove it in Sec. V where the canonical element will be constructed.

For the special case of $\mathcal{U}(\mathrm{sl}(N))$, according to Theorem 1, the following form of twisting element $\mathcal{F}$ can be chosen:

$$\mathcal{F}=\exp(H_{1N}\otimes\sigma)\prod_{j=2}^{N-1}\exp(2\xi E_{1j}\otimes E_{jN}e^{-2\sigma}).$$

This twist of $\mathcal{U}(\mathrm{sl}(N))$ is generated by the twist of $\mathcal{U}(\mathbf{L})$ [here $\mathbf{L}$ is the restricted Borel subalgebra $\mathbf{B}^\vee$ of $\mathrm{sl}(N)$ with the basic elements $\{H_{1N},E_{1N},E_{1j},E_{jN}\}_{j=2,\dots,N-1}$] leading to the Hopf algebra $\mathcal{U}_\xi(\mathbf{B}^\vee)$ with the initial commutation relations [as in (15)], the twisted coproducts

$$\begin{aligned}
\Delta_{\mathcal{F}}H_{1N}&=H_{1N}\otimes e^{-2\sigma}+1\otimes H_{1N}-4\xi\sum_{j=2}^{N-1}E_{1j}\otimes E_{jN}e^{-3\sigma},\\
\Delta_{\mathcal{F}}E_{1i}&=E_{1i}\otimes e^{-\sigma}+1\otimes E_{1i},\\
\Delta_{\mathcal{F}}E_{iN}&=E_{iN}\otimes e^{\sigma}+e^{2\sigma}\otimes E_{iN},\\
\Delta_{\mathcal{F}}E_{1N}&=E_{1N}\otimes e^{2\sigma}+1\otimes E_{1N};
\end{aligned} \tag{31}$$

antipodes

$$S_{\mathcal{F}}(\sigma) = -\sigma, \quad S_{\mathcal{F}}(E_{1i}) = -E_{1i}e^{\sigma},$$
$$S_{\mathcal{F}}(E_{iN}) = -E_{iN}e^{-3\sigma}, \quad S_{\mathcal{F}}(E_{1N}) = -E_{1N}e^{-2\sigma}, \tag{32}$$
$$S_{\mathcal{F}}(H_{1N}) = -H_{1N}e^{2\sigma} - 4\xi \sum_{j=2}^{N-1} E_{1j}E_{jN};$$

and the universal R-matrix of the form

$$\mathcal{R} = \mathcal{F}_{21}\mathcal{F}^{-1} = \prod_j \exp(2\xi E_{jN}e^{-\sigma} \otimes E_{1j}) \exp(\sigma \otimes H_{1N})$$
$$\times \exp(-H_{1N} \otimes \sigma) \prod_j \exp(-2\xi E_{1j} \otimes E_{jN}e^{-\sigma}). \tag{33}$$

The coproducts and antipodes for other elements of $\mathcal{U}_{\xi}(\mathrm{sl}(N))$ can be calculated using the standard formulas. The obtained expressions are rather cumbersome. Thus, for example, in the case of $\mathcal{U}_{\xi}(\mathrm{sl}(3))$, the coproduct of E_{32} looks like

$$\Delta_{\mathcal{F}}E_{32} = E_{32} \otimes e^{-\sigma} + 1 \otimes E_{32} + \xi H_{13} \otimes E_{12}e^{-2\sigma} + 2\xi E_{12} \otimes H_{23}e^{-\sigma}$$
$$-\xi H_{13}E_{12} \otimes (e^{-\sigma} - e^{-3\sigma}) - 4\xi^2 E_{12} \otimes E_{23}E_{12}e^{-3\sigma} - 4\xi^2 E_{12}^2 \otimes E_{23}e^{-4\sigma}.$$

Twisting the coproducts is acting by the exponential of the adjoint operator defined on the tensor product $\mathcal{U}(\mathrm{sl}(N)) \otimes \mathcal{U}(\mathrm{sl}(N))$. One can check that this operator is nilpotent and all the twisted coproducts can be expressed through the finite number of its powers.

IV. CONNECTIONS BETWEEN STANDARD AND JORDANIAN DEFORMATIONS

It is well known that some sorts of Jordanian deformations can be treated as limiting structures for certain sequences of standard quantizations.[9,10,16,17] As will be shown below, this is due to the specific properties of Lie bialgebras involved in the quantizations. These properties are more transparent when formulated for quantum groups rather than for quantum algebras. For this reason, in the current section we use the dual picture to treat Lie bialgebraic characteristics.

The generators of the standard (FRT deformed) quantum group $\mathrm{Fun}_h(\mathrm{SL}(N))$ ($h = \ln q$) will be described by the entries of the $N \times N$ matrix T. Let T be subject to the similarity transformation with the matrix

$$M = 1 + \frac{\xi}{q-1}\rho(E_{1N}) \tag{34}$$

[for the generators the canonical coproduct ($\Delta T = T \otimes T$) is conserved]. For $q \neq 1$ the transformed quantum group $\mathrm{Fun}_{h;\xi}(\mathrm{SL}(N))$ is equivalent to the original one. Compare the corresponding Lie bialgebras: $(g, g^*_{h;0}) = (\mathrm{sl}(N), (\mathrm{sl}(N))^*)$ and $(g, g^*_{h;\xi})$. Here the Lie algebra $g = \mathrm{sl}(N)$ is not changed, the transformation $T \to MTM^{-1}$ does not touch the canonical coproduct for the generators of the Hopf algebra $\mathrm{Fun}_h(\mathrm{SL}(N))$. Only the second Lie multiplication ($\mu^*_{h;0}: V_{g*} \wedge V_{g*} \to V_{g*}$) changes:

$$\mu^*_{h;0} \to \mu^*_{h;\xi}.$$

The structure of the similarity transformation shows that the new Lie product decomposes as

$$\mu^*_{h;\xi} = \mu^*_{h;0} + \xi\mu'. \tag{35}$$

The component μ' is fixed by the commutation relations that can be extracted from the transformed $RTT=TTR$ equations. For this purpose one has to change the coordinate functions of SL(N) arranged in matrix T for the exponential ones $T=\exp(\epsilon Y)$ and also change the parameters $h\mapsto\epsilon h$, $\xi\mapsto\epsilon\xi$. Tending ϵ to zero one gets both summands in (35). The second one of them looks as follows:

$$\begin{aligned}
\mu'(Y_{1k},Y_{ij})&=2\delta_{ik}Y_{Nj}, \quad \text{for } k,j<N, \quad i>1;\\
\mu'(Y_{ij},Y_{lN})&=-2\delta_{jl}Y_{Nj}, \quad \text{for } j<N, \quad i,l>1;\\
\mu'(Y_{ij},Y_{1N})&=-\delta_{j1}Y_{i1}-\delta_{iN}Y_{Nj}, \quad \text{for } j<N, \quad i>1;
\end{aligned}\tag{36}$$

$$\begin{aligned}
\mu'(Y_{1i},Y_{1N})&=-Y_{1i}, \quad \text{for } N>i>1;\\
\mu'(Y_{1N},Y_{kN})&=Y_{kN}, \quad \text{for } 1<k<N;\\
\mu'(Y_{11},Y_{1N})=\mu'(Y_{1N},Y_{NN})&=-(Y_{11}-Y_{NN});\\
\mu'(Y_{1i},Y_{1k})&=\delta_{i1}Y_{Nk}, \quad \text{for } k,i<N,k>1;\\
\mu'(Y_{iN},Y_{kN})&=-\delta_{kN}Y_{i1}, \quad \text{for } k,i>1,i<N;\\
\mu'(Y_{1i},Y_{kN})&=\delta_{i1}Y_{k1}-\delta_{kN}Y_{Ni}-2\delta_{ik}(Y_{11}-Y_{NN}), \quad \text{for } i<N, \quad k>1.
\end{aligned}\tag{37}$$

Here for simplicity of exposition we use the canonical gl(N) basis. One can check that this deforming function μ' not only defines the infinitesimal deformation of $\mu^*_{h;0}$, but is itself a Lie multiplication.

Consider the decomposition (35) as a deformation equation for the initial dual Lie algebra $g^*_{h;0}$. Its main property is that μ' does not depend on h or ξ. So the transformed law has the form

$$\mu^*_{h;\xi}=\mu^*_{h;0}+\mu^*_{0;\xi}.\tag{38}$$

This means that $\mu^*_{h;\xi}$ is a Lie multiplication deformed in the first order. Both summands are Lie maps and at the same time can be considered as deforming functions of each other. As a result, both deforming functions are two-cocycles for the corresponding Lie algebras ($g^*_{0;\xi}$ with the multiplication $\mu^*_{0;\xi}$ and $g^*_{h;0}$ defined by $\mu^*_{h;0}$)

$$\mu^*_{h;0}\in Z^2(g^*_{0;\xi},g^*_{0;\xi}),$$

$$\mu^*_{0;\xi}\in Z^2(g^*_{h;0},g^*_{h;0}).$$

The equivalence of the algebraic structures in $\mathrm{Fun}_{h;\xi}(\mathrm{SL}(N))$ and $\mathrm{Fun}_h(\mathrm{SL}(N))$ (for $h\neq 0$) signifies that $\mu^*_{0;\xi}$ is, in fact, a coboundary,

$$\mu^*_{0;\xi}\in B^2(g^*_{h;0},g^*_{h;0}).$$

On the contrary, the composition $\mu^*_{h;0}$ corresponds to a nontrivial cohomology class

$$\mu^*_{h;0}\in H^2(g^*_{0;\xi},g^*_{0;\xi});$$

the deformation of $\mu^*_{0;\xi}$ by $\mu^*_{h;0}$ is essential.[19]

Notice that the multiplication maps here have certain cohomological properties also with respect to cochain complex C of maps $C^n:\wedge^n V_g\rightarrow V_g\wedge V_g$, where the g-module is chosen to be $\wedge^2 V_g$ with the canonically extended adjoint action on it. The dualization of spaces $V_g\Leftrightarrow V_{g*}$ converts the map μ^* into the chain $\mu^*\in C^1(g,g\wedge g)$. As it was mentioned above, the initial coproduct for the generators of $\mathrm{Fun}_h(\mathrm{SL}(N))$ rests unchanged under the transformation. All the Lie algebras $g^*_{h;\xi}$ are dual to one and the same $g=\mathrm{sl}(N)$. Thus both $\mu^*_{h;0}$ and $\mu^*_{0;\xi}$ are one-cocycles for the complex C.

This set of characteristics necessarily indicates that the classical r-matrix of $\mathcal{U}_{h;\xi}(\mathrm{sl}(N)) \approx (\mathrm{Fun}_{h;\xi}(\mathrm{SL}(N)))^*$ must also exhibit this decomposition property:

$$\begin{aligned} r_{h;\xi} &= r_{h;0} + r_{0;\xi} \\ &= \frac{h}{N}\left(\sum_{k=1}^{N-1} k(N-k) H_{k,k+1}\otimes H_{k,k+1} + \sum_{k<l}^{N-1} (N-l)k(H_{k,k+1}\otimes H_{l,l+1} + H_{l,l+1}\otimes H_{k,k+1}) \right) \\ &\quad + 2h\sum_{k<l} (E_{lk}\otimes E_{kl}) - \xi H_{1N}\wedge E_{1N} - 2\xi \sum_{k=2}^{N-1} E_{1k}\wedge E_{kN}. \end{aligned} \tag{39}$$

In the limit $h \to 0$ one gets the element

$$\lim_{h\to 0} r_{h;\xi} = r_{0;\xi} = -\xi\left(H_{1N}\wedge E_{1N} + 2\sum_{k=2}^{N-1} E_{1k}\wedge E_{kN}. \right) \tag{40}$$

that coincides with r-matrix that can be obtained from $\mathcal{R}$ presented above (33). So the Jordanian quantum group $\mathrm{Fun}_{0;\xi}(\mathrm{SL}(N))$ corresponds to the same r-matrix as the twisted algebra $\mathcal{U}_{\mathcal{F}}(\mathrm{sl}(N))$ [with $\mathcal{F}$ as in (28) and $\tilde{A}_j$, B_j as in (25) and (26)].

The r-matrices (39) and (40) are known for a long time. In Ref. 9 $r_{0;\xi}$ was obtained by applying $\mathrm{ad}_{E_{1N}}$ to the canonical antisymmetric $r_\wedge = \Sigma_{i<j} E_{ij}\wedge E_{ji}$. It was stressed that $r_{0;\xi}$ lay in the boundary of the dense set of orbits of standard quantizations induced by $r_\wedge$. The r-matrix (40) was also obtained in the discussion of conformal algebra deformations.[20]

The r-matrix $r_{0;\xi}$ is the element of the space $\mathbf{B}^\vee \wedge \mathbf{B}^\vee$. Its structure suggests the renumbering of the basic elements of $\mathbf{B}^\vee$; we shall describe the corresponding basis as the set

$$\{P_\alpha, X_\beta\}_{\alpha,\beta=1,\dots,N-1} \quad \text{with } \begin{cases} P_1 = E_{1N}, & P_i = E_{iN}; \\ X_1 = H_{1N}, & X_j = 2E_{1j}; \end{cases} \quad i,j=2,\dots,N-1.$$

In these notations $r_{0;\xi}$ takes the form

$$r_{0;\xi} = -\xi X_\alpha \wedge P_\alpha .$$

The basic exponential coordinate functions $\{Y_{1N}, Y_{iN}, Y_{1i}, Y_{11}-Y_{NN}\}$ are chosen so that they are canonically dual to those of $\{P_\alpha, X_\beta\}$. Let us apply the homomorphism

$$r_{0;\xi}: Y \to -\xi X_\alpha \wedge \langle P_\alpha, Y\rangle \tag{41}$$

to the Lie algebra $(\mathbf{B}^\vee)^*$ described by the last six compositions μ' [see (37)]. As a result we shall get the Lie algebra $\mathbf{B}^\vee$. The significant fact is that (41) is an isomorphism, that is, $\mathbf{B}^\vee \approx (\mathbf{B}^\vee)^*$. The twist $\mathcal{F}$ induces the self-dual Lie bialgebra $(\mathbf{B}^\vee, \mathbf{B}^\vee)$.

It is useful to compare this situation with that of a classical double of dual Lie algebras (g,g^*). There, the composition law of the double can also be presented as a sum of two multiplications with independent linear parameters. But in that case both summands are cohomologically nontrivial. What is more important, such parametrization (and subdivision) cannot be performed in only one algebra of a Lie bialgebra (g,g^*) corresponding to a classical double. In fact, these are the Lie bialgebras that can be parametrized in that case so that their arbitrary linear combination is again a Lie bialgebra.[21] When a Lie bialgebra is nontrivially decomposed (that is, the decomposition goes parallel in both dual algebras), the r-matrix for a linear combination of Lie bialgebras does not inherit the decomposition property.

To clarify the contraction properties of $\mathrm{Fun}_{h;\xi}(\mathrm{SL}(N))$, let us consider the one-parameter subvariety $\{g^*_{h;1-h}\}$ of Lie algebras $g^*_{h;\xi}$ [putting $\xi=1-h$ in (38)]. Each dual pair $(\mathrm{sl}(N),g^*_{h;1-h})$ is a Lie bialgebra and thus is quantizable.[5] The result is the set $\mathcal{A}_{s;h}$ of deformation quantizations parametrized by h and the deformation parameter s. This set can be considered smooth in the sense compatible with the formal series topology[22]—close Lie bialgebras give rise to close deformations. The one-dimensional boundaries $\mathcal{A}_{0;h}$ and $\mathcal{A}_{s;0}$ of $\mathcal{A}_{s;h}$ are formed respectively by the quantizations of $(\mathrm{sl}(N),g^*_{1;0})$ (the standard Lie bialgebra) and $(\mathrm{sl}(N),g^*_{0;1})$ (the Jordanian one). Each internal point in $\mathcal{A}_{s;h}$ can be connected with a boundary by a smooth parametric curve $a(u)$. One can choose it so that it starts in $\mathcal{A}_{0;h}$ and ends in $\mathcal{A}_{s;0}$. So a Jordanian Hopf algebra obtained by twisting deformation can be also treated as a limit point of a smooth one-dimensional subvariety $a(u)$. This does not necessarily mean that this limit is a faithful contraction—it may be impossible to attribute the curve $a(u)$ to an orbit of some Hopf algebra in $\mathcal{A}$. This is just what happens when the transformation M is applied to $\mathrm{Fun}_{h;0}(\mathrm{SL}(N))$. For every positive h fixed, the subset $\{\mathrm{Fun}_{h;\xi}(\mathrm{SL}(N))\}$ is in the $\mathrm{GL}(N^2)$-suborbit of the corresponding $\mathrm{Fun}_{h;0}(\mathrm{SL}(N))$. To attain the points $\mathrm{Fun}_{0;\xi}(\mathrm{SL}(N))$ one must tend h to zero and this can be done only by crossing the set of orbits refering to inequivalent Hopf algebras. These specific interrelations of different types of quantizations where noted in Ref. 9. It was demonstrated for the case of $\mathrm{sl}(N)$ that the standard deformation $\mathrm{Fun}_{h;0}(\mathrm{SL}(N))$ can be accompanied by a smooth transformation of a Jordanian deformation so that the latter reaches the orbit of $\mathrm{Fun}_{h;0}(\mathrm{SL}(N))$. Applying the operator M to an element of the set $\{\mathrm{Fun}_{h;0}(\mathrm{SL}(N))\}$ one gets an intersection point of an orbit and of a curve parametrized by ξ.

One of the principle conclusions is that the possibility of obtaining the Jordanian deformation $\mathrm{Fun}_{0;\xi}(\mathrm{SL}(N))$ as a limiting transformation of the standard quantum group $-\mathrm{Fun}_{h;0}(\mathrm{SL}(N))$ is provided by the fact that the one-cocycle $\mu^*_{0;\xi}\in Z^1(\mathrm{sl}(N),\ \mathrm{sl}(N)\wedge\mathrm{sl}(N))$ [that characterizes the Lie bialgebra for $\mathcal{U}_{\mathcal{F}}(\mathrm{sl}(N))$] is at the same time the two-coboundary $\mu^*_{0;\xi}\in B^2(g^*_{h;0},g^*_{h;0})$, the Lie algebra $g^*_{h;0}$ being the standard dual of $\mathrm{sl}(N)$. The same scenario can be performed on the dual list to get the twisted q-algebra $\mathcal{U}_{\mathcal{F}}(\mathrm{sl}(N))$ as a limit of the variety of standardly quantized algebras $\mathcal{U}_q(\mathrm{sl}(N))$.

V. CANONICAL ELEMENT AND JORDANIAN QUANTUM SPACE

The set $\{P_\alpha,X_\beta\}$ forms the basis appropriate to deal with the Lie bialgebras $(\mathbf{L},\mathbf{L}^*)$. To study the properties of R-matrix $\mathcal{R}$ and the canonical element $\mathcal{T}$ it is reasonable to perform the corresponding rearrangement of basis for the whole Hopf algebra $\mathcal{U}_\xi(\mathbf{B}^\vee)$. We shall consider the set

$$\{z_k\}_{k=1,\ldots,2(N-1)}=\{x_\alpha,\pi_\beta\}_{\alpha,\beta=1,\ldots,N-1} \tag{42}$$

as the generators of $\mathcal{U}_\xi(\mathbf{B}^\vee)$ with

$$x_1=H_{1N},\quad x_i=2E_{1i},$$

$$\pi_1=\frac{1}{\xi}\sigma=\frac{1}{2\xi}\ln(1+2\xi E_{1N}),\quad \pi_i=E_{iN}e^{-2\sigma}.$$

The basis will be formed by the set of ordered monomials:

$$\{z_{\vec{k}}\}_{\vec{k}=\{\vec{m},\vec{n}\}=\{m_1,\ldots,m_{N-1},n_1,\ldots,n_{N-1}\}}=\{x_1^{m_1}\cdots x_{N-1}^{m_{N-1}}\pi_1^{n_1}\cdots\pi_{N-1}^{n_{N-1}}\}. \tag{43}$$

In these terms the R-matrix (33) can be rewritten as

$$\mathcal{R}=\prod_{\alpha=1,\ldots,N-1}^{<}\exp(\pi_\alpha\otimes\xi x_\alpha)\prod_{\alpha=1,\ldots,N-1}^{>}\exp(-\xi x_\alpha\otimes\pi_\alpha), \tag{44}$$

where sybmols $<$ and $>$ mean that the factors in the products are ordered correspondingly. We shall use the standard Hopf algebra homomorphism $\mathcal{R}:\mathcal{A}^*\to\mathcal{A}_-$ where, in our case, $\mathcal{A}$ is the twisted algebra $\mathcal{U}_\xi(\mathbf{B}^\vee)$ and ''$-$'' indicates the opposite multiplication. It would be appropriate to consider $\mathcal{R}$ as belonging to $\mathcal{A}_-\otimes\mathcal{A}$ with the decomposition

$$\mathcal{R}=\sum R^{\vec{k}\vec{l}}y_{\vec{k}}\otimes z_{\vec{l}}. \tag{45}$$

It is implied that the basic monomials $y_{\vec{k}}\in\mathcal{A}_-$ contain the same sequences of generators z_k as the corresponding basic elements $z_{\vec{k}}\in\mathcal{A}$ [see (43)], but the multiplication that glue them is opposite to that of $\mathcal{A}$. Let $\{z^{\vec{k}}\}$ be the canonical dual basis of $\mathcal{A}^*$, $\langle z^{\vec{k}},z_{\vec{l}}\rangle=\delta^{\vec{k}}_{\vec{l}}$. The morphism $\mathcal{R}$ can be defined by its values on the basic elements:

$$\mathcal{R}(z^{\vec{k}})=\sum R^{\vec{l}\vec{k}}y_{\vec{l}}. \tag{46}$$

Let us extract the first terms of the decomposition (45) for the R-matrix (44):

$$\mathcal{R}=1\otimes 1+R^{kl}z_k\otimes z_l+\cdots. \tag{47}$$

(Note that in such a presentation the second term is not proportional to the classical r-matrix; the generators z_l do not form a Lie algebra.) The terms written explicitly in (47) are the only ones containing the first powers of generators. Thus the images $\mathcal{R}(z^k)$ are the linear combinations of the generators z_l. In our case the matrix $\{R^{kl}\}$ is invertible,

$$R=\xi\begin{pmatrix}0 & -I\\ I & 0\end{pmatrix}\Rightarrow\begin{cases}\mathcal{R}(x^\alpha)=\xi\pi_\alpha,\\ \mathcal{R}(\pi^\alpha)=-\xi x_\alpha,\end{cases}$$

$$R^{-1}=-\frac{1}{\xi^2}R\Rightarrow\begin{cases}\mathcal{R}^{-1}(x_\alpha)=-\dfrac{1}{\xi}\pi^\alpha,\\ \mathcal{R}^{-1}(\pi_\alpha)=\dfrac{1}{\xi}x^\alpha.\end{cases}$$

Reversing the formula (46) we get the expression for the elements of the dual basis in terms of generators $z^k{}_{|k=1,\ldots,2(N-1)}$:

$$z^{\vec{k}}=\sum R^{\vec{l}\vec{k}}\mathcal{R}^{-1}(y_{\vec{l}})=\sum R^{\vec{l}\vec{k}}((R^{-1})_{k_1 1}z^{k_1})^{l_1}\cdots((R^{-1})_{k_{2(N-1)}2(N-1)}z^{k_{2(N-1)}})^{l_{2(N-1)}}. \tag{48}$$

The basic decomposition for the R-matrix (44) can be written explicitly:

$$\mathcal{R}=\sum\frac{(-\xi)^{|\vec{n}|}\xi^{|\vec{m}|}}{c_{\vec{m}}c_{\vec{n}}}x_1^{n_1}\cdots x_{N-1}^{n_{N-1}}(\pi_1)^{m_1}\cdots(\pi_{N-1})^{m_{N-1}}\otimes x_1^{m_1}\cdots x_{N-1}^{m_{N-1}}(\pi_1)^{n_1}\cdots(\pi_{N-1})^{n_{N-1}},$$
$$|\vec{n}|=n_1+\cdots+n_{N-1}. \tag{49}$$

Here we used the inclusion $\mathcal{R}\in\mathcal{A}_-\otimes\mathcal{A}$ and the fact that all the generators π_α commute. The structure of $\mathcal{R}$-morphism is clearly seen here. It states the one-to-one correspondence between the basic monomials of $\mathcal{A}^*$ and $\mathcal{A}_-$. This evidently signifies that the Hopf algebras $\mathcal{A}^*$ and $\mathcal{A}_-$ are equivalent. One must also take into account that in our case $\mathcal{A}_-$ is the twisted universal enveloping algebra $\mathcal{U}_\xi(\mathbf{B}^\vee)$ with the opposite product. The result is

$$\mathcal{A}^*\approx\mathcal{A}_-\approx\mathcal{A}.$$

The Hopf algebra $\mathcal{U}_\xi(\mathbf{B}^\vee)$ is self-dual.

The structure constants $R^{\vec{l}\vec{k}}$ presented in the decomposition (49) can be substituted in the expression (48) to fix explicitly the form of the dual basis. Hence the canonical element $\mathcal{T}$ is completely defined

$$\mathcal{T}=\sum_{\vec{k}=(\vec{m},\vec{n})} z^{\vec{k}}\otimes z_{\vec{k}}$$
$$=\sum \frac{1}{c_{\vec{m}}c_{\vec{n}}}(\pi^1)^{n_1}\cdots(\pi^{N-1})^{n_{N-1}}(x^1)^{m_1}\cdots(x^{N-1})^{mN-1}$$
$$\otimes x_1^{m_1}\cdots x_{N-1}^{m_{N-1}}\pi_1^{n_1}\cdots\pi_{N-1}^{n_{N-1}}.$$

We can recollect this expansion into the ordered product using the following property of the $\mathcal{T}$-matrix, $(id\otimes S)(\mathcal{T})=\mathcal{T}^{-1}$:

$$\mathcal{T}^{-1}=\sum_{(\vec{m},\vec{n})}\frac{1}{c_{\vec{m}}c_{\vec{n}}}(\pi^{1n_1}\cdots(\pi^{N-1})^{n_{N-1}}(x^1)^{m_1}\cdots(x^{N-1})^{m_{N-1}}$$
$$\otimes(S(\pi_{N-1}))^{n_{N-1}}\cdots(S(\pi_1))^{n_1}(S(x_1))^{m_1}\cdots(S(x_{N-1}))^{m_{N-1}}.$$

The antipodes used here can be easily found using the expressions (32) given in Sec. III:

$$S(\pi_1)=-\pi_1,\quad S(\pi_i)=-\pi_i e^{\xi\pi_1},$$

$$S(x_1)=-x_1e^{2\xi\pi_1}-4\xi\sum x_i\pi_ie^{2\xi\pi_1},\quad S(x_i)=-x_ie^{\xi\pi_1}.$$

Note that the homomorphic image in $\mathcal{A}^*$ of the Abelian subalgebra generated by elements $\{\pi_\alpha\}$ is itself a commutative subalgebra. This enables us to write the final formula for the canonical element

$$\mathcal{T}=\prod_\alpha^{<}\exp(-x^\alpha\otimes S(x_\alpha))\prod_\alpha^{>}\exp(-\pi^\alpha\otimes S(\pi_\alpha)).\tag{50}$$

The corresponding constructions for Jordanian deformations of the Lie superalgebras of the type $\mathrm{sl}(M|N)$ can be easily performed.

Let us present a real form for $\mathcal{U}_\mathcal{F}(\mathrm{sl}(N))$. We focus first on the subalgebra $\mathbf{B}^\vee$ in the general setting of the previous section and with the basis $\{z_k\}$ [see (42)]. The antialgebraic antilinear transformation given on the generators by

$$\theta(x_\alpha)=-x_\alpha,\quad \theta(\pi_\alpha)=\pi_\alpha$$

respects the classical comultiplication and defines a real form on $\mathcal{U}(\mathbf{L})$. At the same time, the twisting element $\mathcal{F}$ turns into $\mathcal{F}^{-1}$. Henceforth, θ is a real form (cohomomorphic and antihomomorphic) for the twisted algebra $\mathcal{U}_\mathcal{F}(\mathbf{L})$ as well. Turning to the specific case of $\mathrm{sl}(N)$, the question is whether θ can be extended from the subalgebra $\mathbf{B}^\vee$ to the entire $\mathrm{sl}(N)$. This is possible, and the corresponding transformation is

$$\theta(E_{ij})=-E_{ij},\quad i,j<N \text{ or } i,j=N;\quad \theta(E_{kN})=E_{kN},\quad \theta(E_{Nk})=E_{Nk},\quad k<N.$$

It is evident that θ is a Lie algebra antiautomorphism. The real form for the $N=2$ case of the Jordanian $\mathcal{U}_\xi(\mathrm{sl}(2))$ was given in Ref. 23.

Twisting of a symmetry Hopf algebra $\mathcal{A}$ of a manifold $\mathcal{M}$ induces a deformation of its whole geometry, so that the notion of symmetry is conserved in the framework of the noncommutative geometry. Such deformation includes that of function algebras (vector bundles, $*$-structure, and so on) expressing new objects in terms of the nontwisted ones by explicit formulas involving twisting two-cocycle $\mathcal{F}$. Here we present, as an application of the developed Jordanian-type quantization of $\mathrm{sl}(N)$, the corresponding noncommutative space $\mathcal{M}_{\mathcal{F}}$. We deduce commutation relations for generators of $\mathcal{M}_{\mathcal{F}}$, and the differential calculus. The basic formula connecting multiplications in $\mathcal{A}$-modules $\mathcal{M}_{\mathcal{F}}$ and $\mathcal{M}$ (the twisted and the nontwisted ones) is[8]

$$f*g=\mathcal{F}^{-1}_{(1)}(f)\cdot\mathcal{F}^{-1}_{(2)}(g),\quad f,g\in\mathcal{M}. \tag{51}$$

The star stands for the new product on $\mathcal{M}_{\mathcal{F}}$ defined through the old one ''$\cdot$'' and the element $\mathcal{F}$. If $\mathcal{M}$ is classical, the twisting cocycle is represented by a bidifferential operator according to the corresponding representation of $\mathcal{F}$. Thus $\mathcal{M}_{\mathcal{F}}$ and $\mathcal{M}$ coincide as linear spaces, but they are endowed with different algebraic structures. The transformation is performed in such a way that the symmetry property $h(f\cdot g)=h_{(1)}(f)\cdot h_{(2)}(g)$, $h\in\mathcal{A}$, f, $g\in\mathcal{M}$, is inherited by the twisted algebra $\mathcal{A}_{\mathcal{F}}$.

Let x^α, $\alpha=1,...N$, be the generators of $\mathcal{M}_{\mathcal{F}}$. The representation of $\mathrm{sl}(N)$ is given by $E^m_n\to x^m\partial_n$. To evaluate commutation relations among the generators, it is sufficient to retain only the following terms:

$$\mathcal{F}=1\otimes 1+\xi(x^1\partial_1-x^N\partial_N)\otimes x^1\partial_N+2\xi\sum_{k=2}^{N-1}x^1\partial_k\otimes x^k\partial_N+\cdots,$$

with the rest of the series vanishing on the products $x^\mu\cdot x^\nu$. Resolving formula (51) (twisting is an invertible operation) we come to

$$\mathcal{F}_{(1)}(x^\mu)*\mathcal{F}_{(2)}(x^\nu)=x^\mu\cdot x^\nu=x^\nu\cdot x^\mu=\mathcal{F}_{(1)}(x^\nu)*\mathcal{F}_{(2)}(x^\mu).$$

This gives (commutators are understood in terms of the twisted product ''$*$'')[9]

$$[x^1,x^N]=-\xi x^1*x^1,\quad [x^i,x^k]=0,$$

$$[x^k,x^N]=-2\xi x^1*x^k,\quad [x^k,x^1]=0.$$

We use the convention that the small Latin indices run from 2 to $N-1$. Similarly, for the contravariant entities p_μ we have

$$[p_1,p_N]=-\xi p_N*p_N,\quad [p_i,p_k]=0,$$

$$[p_1,p_k]=-2\xi p_k*p_N,\quad [p_k,p_N]=0.$$

Let us note that after the quantization the bases p_μ and x^μ are no longer dual. The invariant canonical element turns out to be $p_\mu\cdot x^\mu=p_\mu*x^\mu+\xi p_1*x^N$. Nontrivial cross relations between coordinates and momenta are expressed by

$$[p_1,x^N]=\xi\left(x^N*p_N+2\sum_{k=2}^{N-1}p_k*x^k+p_1*x^1\right)$$

$$[p_1,x^1]=-\xi x^1*p_N,\quad [p_k,x^k]=-2\xi x^1*p_N,\quad [p_N,x^N]=-\xi x^1*p_N.$$

Partial derivatives ∂_μ satisfies the same identities as p_μ, whereas the cross relations are modified accordingly:

$$[\partial_1, x^N] = \xi\left(x^N * \partial_N + 2\sum_{k=2}^{N-1} \partial_k * x^k + \partial_1 * x^1 \right)$$

$$[\partial_1, x^1] = 1 - \xi x^1 * \partial_N, \quad [\partial_k, x^k] = 1 - 2\xi x^1 * \partial_N, \quad [\partial_N, x^N] = 1 - \xi x^1 * \partial_N.$$

VI. GROUP COCYCLES AND TWISTING

To generalize the construction of Sec. III let us arrange the generators of $\mathbf{B}^\vee$ into the two sets (H, A_j) and (E, B_j) spanning two mutually complementary Lie subalgebras. We denote them $\mathbf{H}$ and $\mathbf{H}^*$, respectively, regarding as dual linear spaces. Subalgebra $\mathbf{H}$ acts on $\mathbf{H}^*$, thus endowing $\mathbf{B}^\vee$ with the semidirect sum $\mathbf{L} = \mathbf{H} \triangleright \mathbf{H}^*$ structure. In this section we explain the twist quantization of interest in terms of a one-cocycle on the Lie algebra $\mathbf{H}$ and its formal Lie group $\mathbf{G} = \exp \mathbf{H}$.

Let H_μ be basic elements of a Lie algebra $\mathbf{H}$ and X^ν be their duals. Commutation relations in $\mathbf{H}$ are specified by the structure constants $C^\sigma_{\mu\nu}$:

$$[H_\mu, H_\nu] = C^\sigma_{\mu\nu} H_\sigma. \tag{52}$$

Suppose a left action of $\mathbf{H}$ on $\mathbf{H}^*$.

$$[H_\mu, X^\nu] = -L^{\nu}_{\mu\sigma} X^\sigma, \tag{53}$$

which enables us to build the semidirect sum $\mathbf{L} = \mathbf{H} \triangleright \mathbf{H}^*$ where $\mathbf{H}^*$ is assumed to be an abelian subalgebra. The element

$$r = X^\nu \otimes H_\nu - H_\nu \otimes X^\nu \in \mathbf{L} \wedge \mathbf{L} \tag{54}$$

is a solution to the classical Yang–Baxter equation if and only if

$$C^\sigma_{\mu\nu} = L^\sigma_{\mu\nu} - L^\sigma_{\nu\mu}. \tag{55}$$

The structure constants $L^\sigma_{\mu\nu}$ define also a left action of $\mathbf{H}$ on itself according to the rule

$$H_\mu \triangleright H_\nu = L^\sigma_{\mu\nu} H_\sigma.$$

The equality (55) implies that the following quasi-associativity property holds:

$$(H_\mu \triangleright H_\nu) \triangleright H_\sigma - (H_\nu \triangleright H_\mu) \triangleright H_\sigma = H_\mu \triangleright (H_\nu \triangleright H_\sigma) - H_\nu \triangleright (H_\mu \triangleright H_\sigma), \tag{56}$$

which is the pre-Lie structure due to Gerstenhaber.[24] Conversely, if a bilinear pairing $\triangleright$ on $\mathbf{H}$ satisfies this condition, the skew-symmetric operation

$$[H_\mu, H_\nu] = (H_\mu \triangleright H_\nu) - (H_\nu \triangleright H_\mu) \tag{57}$$

fulfills the Jacobi identity, and $\triangleright$ becomes a left representation of the Lie algebra $\mathbf{H}$ [equipped with the Lie bracket (57)] on itself.

Lie algebra action $\triangleright$ induces an action of the Lie group $\mathbf{G}$ turning $\mathbf{H}$ into the left $\mathbf{G}$-module. Consider now a one-cocycle φ on the group $\mathbf{G}$ with values in $\mathbf{H}$.[25] This means that φ obeys the equation

$$\varphi(xy) = \varphi(y) + y^{-1} \triangleright \varphi(x), \quad x, y \in \mathbf{G}. \tag{58}$$

The Lie algebra one-cocycle $\partial\varphi$ is in one-to-one correspondence with φ, being its derivative taken at the group identity.[25] It satisfies the equation

$$\partial\varphi([H_\mu, H_\nu]) = H_\mu \triangleright \partial\varphi(H_\nu) - H_\nu \triangleright \partial\varphi(H_\mu)$$

[cf. (57)]. Suppose the linear map $\delta\varphi$ to be nondegenerate. Then the identity map id:$\mathbf{H}\rightarrow\mathbf{H}$ is a one-cocycle with respect to the new action defined as $(\partial\varphi)^{-1}\circ\triangleright\circ\partial\varphi$. Thus nondegenerate one-cocycles of Lie algebras are in one-to-one correspondence with bilinear quasi-associative, in the sense of (56), operations on $\mathbf{H}$. Only nondegenerate cocycles are suitable for our purposes, so we will think of them as of identity maps, and all the freedom will be encoded in the choice of action $\triangleright$. Note that a one-coboundary normalized to id implies the existence of the right unity H_e, that is,

$$H_\mu \triangleright H_e = H_\mu .$$

The group cocycle in terms of Lie algebra coordinates ξ^μ in a neighborhood of the origin reads

$$\varphi^\mu(\xi) = \left(\frac{e^{-L(\xi)}-1}{-L(\xi)} \right)^\mu_\nu \xi^\nu,$$

and the coboundary can be written as

$$\varphi^\mu(\xi) = (1-e^{-L(\xi)})^\mu_\nu \xi^\nu_e, \quad H_\nu \xi^\nu_e = H_e .$$

Consider the semidirect sum $\mathbf{L}=\mathbf{H}\triangleright\mathbf{H}^*$ with the Lie bracket given by (52) and (53) such that the condition (55) holds. Since $\partial\varphi = id$ is nondegenerate, the function φ is invertible in a neighborhood of the origin in $\mathbf{H}$. Its inverse ψ as well as φ itself are treated as columns whose components are formal series in coordinate functions generating $\mathcal{U}(\mathbf{H}^*)$.

Theorem 2: *The element*

$$\mathcal{F} = \exp\big(H_\nu \otimes \psi^\nu(X)\big) \tag{59}$$

is a twist for the semidirect sum $\mathbf{L}=\mathbf{H}\triangleright\mathbf{H}^*$

Proof: The element $\exp\big(H_\nu\otimes\psi^\nu(X)\big)$ satisfies the identity (8). If we prove the second identity (9), the theorem will be stated. Denote $\tilde{X}^\mu = \psi^\mu(X)$ and evaluate $\Delta_{\mathcal{F}}(X)$:

$$\Delta_{\mathcal{F}}(X^\mu) = \exp(H\otimes\tilde{X})(X^\mu\otimes 1 + 1\otimes X^\mu)\exp(-H\otimes\tilde{X}) = X^\nu\otimes(e^{-L(\tilde{X})})^\mu_\nu + 1\otimes X^\mu. \tag{60}$$

The map $\Delta_{\mathcal{F}}(h)=\mathcal{F}\Delta(h)\mathcal{F}^{-1}$ is an algebra homomorphism $\mathcal{U}(\mathbf{L})\rightarrow\mathcal{U}(\mathbf{L})^{\otimes 2}$. Henceforth, the relation (60) entails the equation

$$\varphi\big(\Delta_{\mathcal{F}}(\tilde{X})\big) = e^{-L(1\otimes\tilde{X})}\varphi(\tilde{X}\otimes 1) + \varphi(1\otimes\tilde{X}). \tag{61}$$

Since φ is invertible as a map of $\mathbf{H}$ on $\mathbf{H}$, we find $\Delta_{\mathcal{F}}(\tilde{X}^\mu) = \mathcal{D}^\mu(\tilde{X}\otimes 1, 1\otimes\tilde{X})$ where $\mathcal{D}^\mu(\xi_1,\xi_2)$ is the Campbell—Hausdorff series. This yields (9) and therefore the twist equation (5) for $\exp\big(H_\nu \otimes \psi^\nu(X)\big)$ is valid. ●

Now we can evaluate the twisted coproducts in terms of new generators $\tilde{X}^\mu$. Straightforward calculations show that

$$\Delta_{\mathcal{F}}(H_\mu) = H_\nu \otimes g(\tilde{X})^\nu_\mu + 1\otimes H_\mu , \tag{62}$$

where $g(\xi):\mathbf{H}\rightarrow\mathbf{H}$ is a map to be found. Imposing coassociativity conditions we find that the function g realizes a left group action on $\mathbf{H}$ which is generated by a Lie algebra representation. To evaluate this action let us perform the following Lie algebra isomorphism $H_\mu \rightarrow H_\mu$, $X^\mu\rightarrow\xi X^\mu$. The specific form of the classical r-matrix allows us to consider ξ as a deformation parameter. Taking into account $(d/d\xi)\tilde{X}^\mu(0)=X^\mu$, $\tilde{X}^\mu(0)=0$, and calculating $(d/d\xi)\mathcal{F}\Delta(H_\nu)\mathcal{F}^{-1}|_{\xi=0}$ we come to

$$\frac{d}{d\xi}\Delta_{\mathcal{F}}(H_\nu)|_{\xi=0}=[H_\sigma\otimes X^\sigma,H_\nu\otimes 1+1\otimes H_\nu]=C^\mu_{\sigma\nu}H_\mu\otimes X^\sigma+L^\mu_{\nu\sigma}H_\mu\otimes X^\sigma=L^\mu_{\sigma\nu}H_\mu\otimes X^\sigma.$$

Performing this for the coproduct (62) and comparing the results we obtain $g(\tilde{X})=e^{L(\tilde{X})}$. Thus the coproduct on generators H_μ, $\tilde{X}_\nu$ reads

$$\Delta_{\mathcal{F}}(\tilde{X}^\mu)=\mathcal{D}^\mu(\tilde{X}\otimes 1,1\otimes\tilde{X}),\quad \Delta_{\mathcal{F}}(H_\mu)=H_\nu\otimes(e^{L(\tilde{X})})^\nu_\mu+1\otimes H_\mu. \tag{63}$$

Using these relations it is easy to find also the antipodes,

$$S_{\mathcal{F}}(\tilde{X}^\mu)=-\tilde{X}^\mu,\quad S_{\mathcal{F}}(H_\mu)=-H_\nu(e^{-L(\tilde{X})})^\nu_\mu. \tag{64}$$

Writing $\tilde{X}^\mu$ in terms of X^ν we can evaluate the twisted antipode on the classical generators as well.

VII. CONCLUSION

The triangular deformation of the universal enveloping algebra of sl(N), discussed already by Gerstenhaber *et al.*,[9] was realized in this paper as a twisted algebra with the explicit form (2) of the twisting element $\mathcal{F}$ (extended Jordanian twist). The Hopf subalgebra of the type $\mathcal{U}_{\mathcal{F}}(\mathbf{B}^\vee)$ generated by the twist is self-dual. The twisted coproduct of the sl(N) generators can be expressed as finite sums of classical generators and such a function σ of the highest root vector that is primitive with respect to the twisted coproduct. The commutation relations of the quantum space generators were defined using the twisting element action on the commutative coordinates. The cohomological properties of the involved Lie bialgebra permit the explanation of the connection of the Drinfeld–Jimbo (standard) quantization with this twisting.

The explicit expression of the twisting element $\mathcal{F}$ gives rise to a possibility to evaluate the Clebsch–Gordan coefficients (CGC) of the twisted sl(N) in terms of the original CGC and the entries of the matrix $F=(\rho_1\otimes\rho_2)\mathcal{F}$,[11] as well as to get the relations among the FTR-approach generators $L^{(\pm)}$ of the twisted algebra and the generators of the original algebras. It can be used also to construct the quantum double.[1,12]

The construction of the extended Jordanian twist was generalized to a certain class of inhomogenious Lie algebras, using properties of the Campbell–Hausdorff series. Further generalizations, in particular to Lie superalgebras, twisting of the corresponding Yangians, and new integrable models, twisting elements for other boundary solutions to the classical Yang–Baxter equation[26] are under study.

ACKNOWLEDGMENTS

PPK was supported by the RFFI Grant Nos. NN 96-01-00851 and 98-01-00310. VDL was supported by the RFFI Grant No. N 97-01-01152.

[1] V. G. Drinfeld, ''Quantum groups,'' in *Proceedings of the International Congress of Mathematicians, Berkeley, 1986*, edited by A. V. Gleason (AMS, Providence, 1987), Vol. 1, pp. 798–820.
[2] M. Jimbo, Lett. Math. Phys. **10**, 63 (1985); **11**, 247 (1986).
[3] F. Bayen, M. Flato, A Lichnerowicz, and D. Stenheimer, Ann. Phys. **111**, 61 (1978); M. Flato and D. Sternheimer, Lett. Math. Phys. **22**, 155 (1991).
[4] A. A. Belavin and V. G. Drinfeld, ''Triangle equations and simple Lie algebras,'' in *Mathematics Physics Reviews*, edited by S. P. Novikov (Harwood, New York, 1984), pp. 93–166.
[5] P. Etingof and D. Kazhdan, Selecta Math. 2, N **1**, 1-41 (1996); q-alg/9510020.
[6] L. D. Faddeev, N. Yu. Reshetikhin, and L. A. Takhtajan, Algebra i Analiz **1**, 178 (1989), [English translation: Leningrad Math. J. **1**, 193 (1990)].
[7] V. G. Drinfeld, Leningrad Math. J. **1**, 1419 (1990).
[8] V. G. Drinfeld, Dokl. Akad. Nauk SSSR **273(3)**, 531 (1983).
[9] M. Gerstenhaber, A. Giaquinto, and S. D. Schack, in *Quantum groups. Proceedings in EIMI 1990, Lecture Notes in Mathematics* edited by P. P. Kulish (Springer-Verlag, Berlin, 1992), Vol. 1510, pp. 9–46.
[10] O. V. Ogievetsky, in *Proceedings Winter School Geometry and Physics*, Zidkov, Suppl. Rendiconti cir. Math. Palermo, Serie II-N 37 (1993), p. 185; preprint MPI-Ph/92-99, Munich (1992).

[11] P. P. Kulish and A. A. Stolin, Czech. J. Phys. **47**, 123 (1997); preprint q-alg/9708024; J. A. de Azecarraga, P. P. Kulish, and F. Rodenas, Z. Phys. C **76**, 567 (1997).
[12] A. A. Vladimirov, Mod. Phys. Lett. A **8**(23), 2573 (1993); hep-th/9401101.
[13] A. Ballesteros, F. J. Herranz, M. A. del Olmo, C. M. Perena, and M. Santander, J. Phys. A **28**, 7113 (1995).
[14] N. Yu. Reshetikhin and M. A. Semenov-Tian-Shansky, J. Geom. Phys. **5**(4), 533 (1988).
[15] N. Yu. Reshetikhin, Lett. Math. Phys. **20**, 331 (1990).
[16] M. Alishahiha, J. Phys. A **28**, 6187 (1995).
[17] B. Abdesselam, A. Chakrabarti, and R. Chakrabarti, ''General Construction of Nonstandard R_h-matrices as Contraction Limits of R_q-matrices,'' preprint q-alg/9706033.
[18] V. Chari and A. Pressley, *A Guide to Quantum Groups* (CUP, 1994).
[19] A. Nijenhuis and R. Richardson, Bull. Am. Math. Soc. **72**, 1 (1966).
[20] J. Lukierski, P. Minnaert, and M. Mozrzymas, Phys. Lett. B **371**, 251 (1996).
[21] V. D. Lyakhovsky and A. M. Mirolyubov, Int. J. Mod. Phys. A **12**, 225 (1997).
[22] N. Burbaki, *Elements de Mathematique Algebras* (Hermann, Paris, 1970), L II, Chap. 4, Sec. 5.10.
[23] S. L. Woronowicz, Rep. Math. Phys. **30**, 259 (1991).
[24] M. Gerstenhaber, Ann. Math. **78**, 267 (1963).
[25] A. Guichardet, *Cohomologie des groupes topologiques et des algebres de Lie* (Cedic/Fernand Nathan, Paris 1980).
[26] M. Gerstenhaber and A. Giaquinto, Lett. Math. Phys. **40**, 337 (1997).

On the root mean square quantitative chirality and quantitative symmetry measures

Michel Petitjean[a)]
ITODYS (CNRS, ESA 7086), 1 rue Guy de la Brosse, 75005 Paris, France

(Received 22 February 1999; accepted for publication 22 March 1999)

The properties of the root mean square chiral index of a d-dimensional set of n points, previously investigated for planar sets, are examined for spatial sets. The properties of the root mean squares direct symmetry index, defined as the normalized minimized sum of the n squared distances between the vertices of the d-set and the permuted d-set, are compared to the properties of the chiral index. Some most dissymetric figures are analytically computed. They differ from the most chiral figures, but the most dissymetric 3-tuples and the most chiral 3-tuples have a common remarkable geometric property: the squared lengths of the sides are each equal to three times a squared distance vertex to the mean point. © *1999 American Institute of Physics.* [S0022-2488(99)01009-9]

I. INTRODUCTION

Chirality and symmetry properties of a solid body can be viewed as a continuous varying quantity taking values over [0;1] rather than a logical property, i.e., the body is or is not symmetric or chiral. The use of a chirality measure seems to be introduced by Rassat.[1] Then, various quantitative chirality or symmetry measures have been used.[2–12] This concept has received applications in physics, proposed mostly by the Avnir group.[2–4]

The root mean square chiral index CHI of a d-dimensional set of n points was defined[12] as the sum of the n squared distances between the vertices of the set and those of its inverted image, normalized to $4T/d$, T being the inertia of the set. This index is computed after minimization of the sum of the squared distances in respect to all rotations and translations and all permutations between equivalent vertices. It was shown to be a second kind of continuous chirality measure taking values over [0;1], the zero value corresponding to an achiral compound perfectly superposed to its inverted image. Similarly, the direct symmetry index DSI of a d-dimensional set of n points is defined here as follows. When all vertices are unequivalent, DSI is undefined. When there are at least two equivalent vertices, the sum of the n squared distances between the vertices and those of the permuted set is minimized for all rotations and translations and permutations (excluding the identity permutation) between equivalent vertices. DSI is the ratio of this minimized sum to twice the inertia T of the set.

The quantitative symmetry and chirality concepts used here are fully different from those of Avnir *et al.* for the following reasons: no achiral reference is needed to compute CHI, no symmetry assumptions are needed to compute CHI and DSI, no folding and unfolding process[5] are needed here, the normalization are different, and the farthest point from the centroid is not needed here, and, of course, the extremal figures are different.

The properties of CHI were examined for monodimensional sets and planar sets.[12] They are now examined for spatial sets. Hyperspatial sets (i.e., d is any positive integer) are examined when all vertices are unequivalent. The major difference between planar, spatial, and hyperspatial sets lies in the expression of the optimal rotation. The properties of DSI are also examined. For clarity, a set of $n=3$ points will be called a triangle. The most dissymetric triangles, i.e., those maximizing DSI, are here analytically computed when there are two or three equivalent vertices.

[a)]Phone: 33 (0)1 4427 4857; fax: 33 (0)1 4427 6814; electronic mail: Petitjean@itodys.jussieu.fr

II. NOTATIONS

The notations are those used in Ref. 12. X0 and X1 are the n rows and d columns arrays of coordinates. X0 is the fixed set and X1 is to move. The quote denotes the transposition operator. All vectors are written as one-column matrices. $\langle x|y\rangle$ is the scalar product of the vectors x and y, and when $d=3$, $x\wedge y$ is their cross product. The trace and the determinant operators are denoted, respectively, Tr and Det. Y1 is the rotated and translated image of X1, and $D^2=\mathrm{Tr}((\mathrm{X0}-\mathrm{Y1})\cdot(\mathrm{X0}-\mathrm{Y1})')$ is the sum of the squared distances. D^2 is minimized for rotation plus translation when X0 and X1 are centered before computing the optimal rotation. Translations will be no longer considered, and the centering condition will not be assumed unless otherwise mentioned. The following matrices are used: $\mathrm{V00}=\mathrm{X0}'\cdot\mathrm{X0}$, $\mathrm{V11}=\mathrm{X1}'\cdot\mathrm{X1}$, $\mathrm{V10}=\mathrm{X1}'\cdot\mathrm{X0}$, $\mathrm{V01}=\mathrm{V10}'$, and $T=(\mathrm{T0}+\mathrm{T1})/2$, $\mathrm{T0}=\mathrm{Tr(V00)}$ and $\mathrm{T1}=\mathrm{Tr(V11)}$ being the respective inertia of X0 and X1, reducing to the usual inertia when the arrays are centered. The identity matrix is **I**, and R is a rotation matrix, such that $\mathrm{Y1}=\mathrm{X1}\cdot R'$.

The correspondence between X0 and X1 is handled via an n-dimensional square permutation matrix P. Let be $\mathrm{Z1}=P\cdot\mathrm{Y1}$. When X1 is the inverted image of X0 and when the centering condition is satisfied, the chiral index of a spatial set is $\mathrm{CHI}=D^2/(4T/d)$, with $D^2=\mathrm{Tr}((\mathrm{X0}-P\cdot\mathrm{X1}\cdot R')\cdot(\mathrm{X0}-P\cdot\mathrm{X1}\cdot R')')$ being minimized over all rotations R and allowed permutations P. When X1 is a rotated and translated image of X0 and when the centering condition is satisfied, $\mathrm{DSI}=D^2/2T$, D being minimized over all rotations and allowed nonidentity permutations.

The computation of either CHI or DSI requires the optimal rotation superimposing two sets. When $d=3$, the analytical expression of the optimal rotation superposing X1 on X0, X0 and X1 being any n rows and 3 columns arrays of coordinates, is given in the Appendix.

III. THE OPTIMAL ROTATION FOR 3D ENANTIOMERS

In this section, the centering condition is not assumed and three-dimensional enantiomers are considered. For clarity, X0 is noted $\sim X$ and its inverted image is $\mathrm{X1}=-P\cdot X$, and we define $V=X'\cdot P\cdot X=-\mathrm{V01}$. From Appendix A, we have

$$D^2=D0^2-2\langle q|Bq\rangle, \tag{1}$$

the optimal quaternion q being the eigenvector associated to L1, the highest eigenvalue of B:

$$B=\begin{pmatrix} 0 & c' \\ c & A \end{pmatrix}, \tag{2}$$

$$A=\mathrm{Tr}(V+V')\cdot\mathbf{I}-(V+V'), \tag{3}$$

$$c=\begin{pmatrix} \mathrm{V(2,3)-V(3,2)} \\ \mathrm{V(3,1)-V(1,3)} \\ \mathrm{V(1,2)-V(2,1)} \end{pmatrix}. \tag{4}$$

When P is a symmetric permutation, c is null, and the eigenvalues of B are the three eigenvalues of A and zero.

IV. ENANTIOMERS WITH ALL VERTICES UNEQUIVALENT

All the conditions of the preceding section are assumed to stand, and the vertices are all unequivalent, i.e., the only allowed permutation is $P=\mathbf{I}$. $V=X'\cdot X$ is symmetric and c is therefore null. The sum of squares prior rotation is $\mathrm{D0}^2=4\,\mathrm{Tr}(V)$, which is the maximized D^2 value because zero is the smallest eigenvalue of B. We have $A=2(\mathrm{Tr}(V)\cdot\mathbf{I}-V)$. Let v1, v2, v3 be the eigenvalues of V arranged in decreasing order. The largest eigenvalue of B is $\mathrm{L1}=\mathrm{d1}=2(\mathrm{v1}+\mathrm{v2})$ and the optimal rotation of $-X$ is 180 degrees around the principal axis associated to the smallest eigenvalue of V. Now we have $D^2=4\,\mathrm{Tr}(V)-4(\mathrm{v1}+\mathrm{v2})$, i.e.,

$$D^2 = 4\mathrm{v}3. \tag{5}$$

We assume now that X is centered, i.e., V is n times its variance matrix. The chiral index of the set of n vertices is therefore:

$$\mathrm{CHI} = 3\mathrm{v}3/(\mathrm{v}1+\mathrm{v}2+\mathrm{v}3). \tag{6}$$

CHI is d times the percentage of inertia associated to the smallest eigenvalue of V. Looking at Eqs. (3) and (7) in Ref. 12 and Appendix 1 in Ref. 12, we can see that this is also true for planar sets $(d=2)$ and unidimensional sets $(d=1)$.

The eigenvalues of V being positive and in decreasing order, CHI is maximized for $\mathrm{v}1=\mathrm{v}2=\mathrm{v}3=v$, i.e., $\mathrm{CHI}=1$ and $V=v\cdot\mathbf{I}$. When X has only 4 vertices, it is therefore a regular tetrahedron (see Appendix 2 in Ref. 12).

V. HYPERSPATIAL SETS WITH ALL UNEQUIVALENT VERTICES

The optimal rotation superimposing two d-dimensional sets is unknown when $d>3$, except for enantiomers with all unequivalent vertices, as shown hereafter. The sum of squares to be minimized is $D^2=\mathrm{Tr}((X-X\cdot Q')\cdot(X-X\cdot Q')')=2(\mathrm{Tr}(X'\cdot X)-\mathrm{Tr}(Q\cdot X'\cdot X))$, X being the (n,d) array of coordinates and Q being an orthogonal matrix with $\det(Q)=-1$. Thus, $\mathrm{Tr}(Q\cdot X'\cdot X)$ has to be maximized. Assuming that X is in its principal components axis (i.e., $V=X'\cdot X$ is diagonal), we have to find the maximum of $E=\mathrm{v}(1)\cdot\mathrm{Q}(1,1)+\mathrm{v}(2)\cdot\mathrm{Q}(2,2)+...+\mathrm{v}(d)\cdot\mathrm{Q}(d,d)$, $\mathrm{v}(1),...,\mathrm{v}(d)$ being the eigenvalues of V in decreasing order.

$$E=[(\mathrm{v}(1)-\mathrm{v}(d))\cdot\mathrm{Q}(1,1)+(\mathrm{v}(2)-\mathrm{v}(d))\cdot\mathrm{Q}(2,2)+\cdots+(\mathrm{v}(d-1)-\mathrm{v}(d))\cdot\mathrm{Q}(d-1,d-1)]+\mathrm{v}(d)\cdot\mathrm{Tr}(Q).$$

The eigenvalues of Q can be either $+1$, or -1, or pairs of conjugate complex roots of 1. It follows that $\mathrm{Tr}(Q)$ is maximized when $d-1$ eigenvalues are $+1$ and one is -1. Obviously, the sum of the $d-1$ terms $(\mathrm{v}(i)-\mathrm{v}(d))\cdot Q(i,i)$ is also maximized for $\mathrm{Q}(i,i)=1$ when $i<d$. Thus E is maximized and D^2 is minimized when X and its enantiomer have opposite coordinates on the principal axis with smallest inertia. Thus, Eqs. (5) and (6) are generalized:

$$D^2=4\mathrm{v}(d), \tag{7}$$

and assuming X centered:

$$\mathrm{CHI}=d\cdot\mathrm{v}(d)/(\mathrm{v}(1)+\mathrm{v}(2)+\cdots+\mathrm{v}(d)). \tag{8}$$

As previously, CHI is maximized when all eigenvalues of V are equal. When $n=d+1$, CHI is therefore maximized when X is a regular d-simplex. (See Appendix 2 in Ref. 12.)

From Eq. (8), it is possible to compare practical CHI values with the distribution of CHI when X is an isotropic multinormal sample. V is a Wishart matrix,[13] from which the joint density of the percentages of inertia can be derived,[14] leading to the distribution[15] of CHI/d. Unfortunately, the final expression is not trivial when $d>2$.

VI. THE DIRECT SYMMETRY INDEX

In this section, d-dimensional sets are considered and the centering condition is not assumed. The situation where all vertices are unequivalent precludes the existence of direct symmetry in the set. This situation should not be confused with the purely geometric situation where all vertices are equivalent (i.e., undistinct), for which symmetry properties are potentially observable. Thus, we consider now only sets with at least two equivalent vertices. As for the chiral index, the sum D^2

of the n squared distances between the vertices and those of the permuted set is minimized for all rotations and all authorized permutations, excluding of course the identity permutation $P=\mathbf{I}$. When the set is centered DSI$=D^2/(2T)$.

P being fixed, the sum of the squared distances to minimize is, as previously, $D^2=\mathrm{Tr}((\mathrm{X0}-P\cdot\mathrm{X1}\cdot R')\cdot(\mathrm{X0}-P\cdot\mathrm{X1}\cdot R')')$. Setting X0=X1=$X$ and $V=X'\cdot P\cdot X$, we get

$$D^2=2(T-\mathrm{Tr}(V\cdot R')). \tag{9}$$

There are at least two equivalent points x and y. Thus the minimum of D^2 for all rotations and permutations cannot exceed the minimum of D^2 for all rotations and for the permutation exchanging x and y, i.e., P is such that $V=x\cdot y'+y\cdot x'+Z'\cdot Z$, with Z being the $(n-2,d)$ block extracted from X by elimination of x and y. For this permutation,

$$\mathrm{Tr}(V\cdot R')=y'\cdot R'\cdot x+x'\cdot R'\cdot y+\mathrm{Tr}(Z'\cdot Z\cdot R'). \tag{10}$$

Assuming $d>1$, a rotation exists R which rotates from $+90$ degrees the first axis toward the second axis, i.e., $R(2,1)=-R(1,2)=1$, $R(i,i)=1$ for $i>2$, all other elements of R being null. Thus, $R+R'=0$, $y'\cdot R'\cdot x+x'\cdot R'\cdot y=0$ and $-\mathrm{Tr}(Z'\cdot Z\cdot R')=\mathrm{Tr}(Z'\cdot Z\cdot R)=\mathrm{Tr}(R'\cdot Z'\cdot Z)=\mathrm{Tr}(Z'\cdot Z\cdot R')=0$, which means that $\mathrm{Tr}(V\cdot R')$ is null. Because a permutation and a rotation exist such that $\mathrm{Tr}(V\cdot R')=0$, it follows from (9) that the minimum of D^2 is upper bounded by $2T$, and then DSI pertains to $[0;1]$ when $d>1$.

The following centered set containing three points is such that DSI$=1$ for all $d>1$: $x=e1\cdot(-1-\sqrt{3})/2$, $y=e1\cdot(-1+\sqrt{3})/2$, $z=e1$, $e1$ being the first base vector, x and y being equivalent and z being not.

When $d=1$, x and y are numbers, Z is a vector, $T=x^2+y^2+Z'\cdot Z$, $R=1$, and $\mathrm{Tr}(V\cdot R')=2x\cdot y+Z'\cdot Z$. Thus, $4T-D^2=2\cdot(T+\mathrm{Tr}(V\cdot R'))=2(x+y)^2+4\cdot Z'\cdot Z$, which cannot be negative. Thus, for $d=1$, D^2 varies from 0 to $4T$ and the direct symmetry index pertains to $[0;2]$, the extremal value DSI$=2$ being reached for a centered set containing two opposite values. But of course, direct rotational symmetry has little interest for $d=1$.

Computing simultaneously CHI and DSI for spatial sets is easy, since they both lead to the same quadratic form defined by Eqs. (1)–(4), except that the quadratic form associated to DSI now takes the opposite sign, because X1 was set to X rather than to $-X$. It means that the smallest eigenvalue L4 should be used to compute DSI rather than L1 for CHI, the minimized sum of squared distances being now

$$D^2=\mathrm{D0}^2+2\mathrm{L4}. \tag{11}$$

As shown in the Appendix, L4 is always nonpositive. Another difference between CHI and DSI is that the normalizing coefficients are, respectively, $4T/d$ and $2T$, but this is not a crucial difference.

VII. THE DIRECT SYMMETRY INDEX OF PLANAR TRIANGLES

We assume that $d=2$. Let x be the column vector of the abscissas, and y the column vector of their ordinates: $x'=(x_1,x_2,\ldots,x_n)$ and $y'=(y_1,y_2,\ldots,y_n)$. The points will be $p_1,p_2,\ldots,p_n$. The image of (x,y) through the permutation P is (Px,Py). P being fixed, the distance D minimized for all rotations is known:[12]

$$D^2=2(T-E), \tag{12}$$

E being the non-negative number, such that

$$E^2=(x'P'x)^2+(x'P'y)^2+(y'P'x)^2+(y'P'y)^2+2(x'P'x)(y'P'y)-2(y'P'x)(x'P'y).$$

Thus,

$$E^2=(x'Px+y'Py)^2+(y'Px-x'Py)^2. \tag{13}$$

The minimization for rotations plus translations is reached when the set is centered. The inertia is thus $T=x'x+y'y$. We assume T non-null, i.e., there are at least two distinct points. Let $\mathbf{1}$ be the n-vector such that all its n components are 1. Centering means $\mathbf{1}'x=\mathbf{1}'y=0$. We define also $M=(P+P')/2$ and $N=(P-P')/2$, which implies that $x'Nx=y'Ny=0$.

We assume now that $n=3$, and that all vertices are equivalents. d_{12}, d_{23} and d_{31} are the respective lengths of the sides of the triangle. $\|p_1\|$, $\|p_2\|$ and $\|p_3\|$ are the lengths of the segments joining the barycenter to the vertices at the opposite of the sides with respective lengths d_{23}, d_{31} and d_{12}. The inertia can be also written as $T=\|p_1\|^2+\|p_2\|^2+\|p_3\|^2$, or $T=(d_{23}^2+d_{31}^2+d_{12}^2)/3$.

The surface S of the triangle is such that $16S^2=2(d_{12}^2d_{23}^2+d_{23}^2d_{31}^2+d_{31}^2d_{12}^2)-(d_{12}^4+d_{23}^4+d_{31}^4)$.

A. Extremal values for a given permutation

Using M and N, Eq. (13) becomes

$$E^2=(x'Mx+y'My)^2+4(x'Ny)^2. \tag{14}$$

The gradient of $(1-D^2/2T)^2=(E^2/T^2)$ if set to zero for x, then for y,

$$T(x'Mx+y'My)Mx+2T(x'Ny)Ny=E^2x, \tag{15}$$

$$T(x'Mx+y'My)My-2T(x'Ny)Nx=E^2y. \tag{16}$$

Multiplying on the left (15) by x' then (16) by y', and substracting,

$$T(x'Mx)^2-T(y'My)^2=E^2(x'x-y'y). \tag{17}$$

Then from (15) or (16),

$$T(x'Mx+y'My)(x'My)=E^2(x'y). \tag{18}$$

From (17) and (18), it comes

$$E^2(x'My)(x'x-y'y)=E^2(x'y)(x'Mx-y'My). \tag{19}$$

When $n=3$,5 permutations are possible: 3 are symmetric and 2 are circular.

When P is symmetric, $M=P$, $N=0$, and Eqs. (15) and (16) reduce to the same eigenvalues equations: $T^2Px=E^2x$ and $T^2Py=E^2y$. For $n=3$, the eigenvalues of P are $+1$, $+1$ and -1. Only the solution such that $E^2=T^2$ is possible, implying $D=0$, leading to a minimum for DSI, rather to a maximum.

When P is one of the 2 circular permutations (the other being its transposed), we have: $2\mathbf{M}=\mathbf{1}\cdot\mathbf{1}'-\mathbf{I}$, implying that $x'Mx=-x'x/2$ and $y'My=-y'y/2$, and then $E^2=4(x'Ny)^2+T^2/4$. Moreover, $2x'Ny$ is equal to the determinant of the matrix $[\mathbf{1}|x|y]$ or to the opposite of this determinant, depending on which circular permutation is used. That implies $E^2=4S^2+T^2/4$. The minimum is therefore reached by a null-area triangle: the points are aligned.

B. Maximizing DSI

Let us consider the symmetric permutation associating p_1 to itself. The following comes: $N=0$ and $E^2=2x_2x_3+x_1^2+2y_2y_3+y_1^2=(T-d_{23}^2)^2$. Similarly, the E values associated with the symmetric permutations associating p_2 with p_2 and p_3 with p_3 are such that $E^2=(T-d_{31}^2)^2$ and $E^2=(T-d_{12}^2)^2$. Both circular permutations lead to $E^2=4S^2+T^2/4$, S^2 and T being homogeneous polynomials of d_{12}^2, d_{23}^2 and d_{31}^2. The 4 expressions of E^2 are homogeneous polynomials of 3 variables, returning non-negative values.

For a given triangle, the optimal permutation is that which leads to the highest E value. Thus a maximum of DSI, or a minimum of $[\mathrm{Max}(E^2)/T^2]$, should be searched either among the extrema of E^2/T^2 associated with a permutation, or at the intersection of at least two of the 4 polynomials associated with the permutations, T being the same for all permutations.

It was shown above that only one extremum of E^2/T^2 is useful, and it is such that $E=T/2$ and $S=0$. This is possible only if the length of a side is equal to the sum of the two others. Assuming for example, that $d_{23}=d_{31}+d_{12}$, and reporting it in $E=|T-d_{23}^2|=|d_{12}^2+d_{31}^2-2d_{23}^2|/3$. The following comes: $E=(d_{23}^2+2d_{12}d_{31})/3$, or $E=(T/2)+d_{12}d_{31}$. The circular permutation could be optimal only if $d_{12}d_{31}=0$, which should imply that $\mathrm{DSI}=0$ (degenerate isocele triangle). Similar conclusions should be reached if d_{31} or d_{12} have been used: the extrema of E^2/T^2 associated to a given permutation are not adequate.

Thus, it is needed to look at the intersection of the polynomials. Noting that the E/T values depend only on the distances ratios, we can work with only two independent variables, and we search the minimum at the intersection of 3 among the 4 polynomials. There are at least 2 among 3 symmetric permutations which lead to the same value. Assuming, for example, that $E=|T-d_{12}^2|=|T-d_{31}^2|$, thus, $|d_{23}^2+d_{31}^2-2d_{12}^2|=|d_{12}^2+d_{23}^2-2d_{31}^2|$. Either we get $d_{12}^2=d_{31}^2$, which does not work because the triangle should be isocele ($\mathrm{DSI}=0$), or we get $2d_{23}^2=d_{31}^2+d_{12}^2$, and thus $E=|d_{12}^2-d_{31}^2|/2$. In this situation, the E value associated to the third symmetric permutation is null, and the E value associated to circular permutations is $E=d_{12}d_{31}$. The equality between the 3 nonzero E^2 values give the desired relation: $(d_{12}^2-d_{31}^2)^2=4d_{12}^2d_{31}^2$. Reusing $2d_{23}^2=d_{31}^2+d_{12}^2$, the ratios of the squared lengths of the sides are deduced: $(d_{12}^2/d_{23}^2)=1+\sqrt{2}/2$, $(d_{31}^2/d_{23}^2)=1-\sqrt{2}/2$, and $E^2/T^2=(1-\mathrm{DSI})^2=1/2$.

C. Remarkable geometric properties of the optimal triangles

Using the distances, we get the angles associated, respectively, to the points p_1, p_2 and p_3: $\pi/4$, $\pi/8$ and $5\pi/8$.

A possible set of coordinates of the most dissymetric triangle is

$$X=\begin{pmatrix} \sqrt{2}/3 & 1/3 \\ (-3-\sqrt{2})/6 & -1/6 \\ (3-\sqrt{2})/6 & -1/6 \end{pmatrix}. \tag{20}$$

It is easy to see that $d_{23}^2=3\|p_1\|^2$, $d_{12}^2=3\|p_2\|^2$, and $d_{31}^2=3\|p_3\|^2$. It should be pointed out that this relation is symmetric only for p_2 and p_3. This remarkable proportionality exists also for the degenerate triangle with only two equivalent vertices, which was cited in Sec. VI, and corresponding to the maximal value $\mathrm{DSI}=1$, for any dimension $d>1$. For $d=2$, the most chiral triangles also offer this remarkable proportionality, discarding which vertices are equivalent,[12] but none of them has the shape of the most dissymmetric triangle. The shape of the most dissymmetric triangle has been measured using random triangles, with vertices uniformly distributed over a square. The results (Table I) are in accordance with the theory.

VIII. DISCUSSION AND CONCLUSION

The properties of the RMS (root mean square) chiral index have been examined for spatial sets. As for planar sets, it is easily analytically computed, but the expression of the optimal 3D rotation is fully different from those of the 2D one. The optimal rotation is unknown for hyperspatial sets, except when X is superposed with its unpermuted enantiomer. When $d>3$, it is proposed to extend the iterative procedure[16] to compute the optimal rotation superposing two d-dimensional sets, and to use it for permuted enantiomers. Similarly, computing the RMS direct symmetry index is easy for 2D and 3D sets, but suffers from the same limitation than the chiral index when $d>3$.

Looking at Eq. (8), it is clear that the RMS chiral index is also extendible to continuous distributions with all distinct points, provided that the variance exists. When there are subsets of

TABLE I. Measure of the shape of the most dissymmetric triangle with three equivalent vertices. Ntr: number of random triangles. Popt: optimal permutation. DSI: direct symmetry index. The 3 angles are expressed as multiples of $\pi/8$.

Ntr	Popt	$(1\text{-DSI})^2$	Angles
1	321	0.724887	0.829182 1.305712 5.865106
11	312	0.632934	1.239900 2.710587 4.049513
13	231	0.546920	1.160093 1.859441 4.980466
37	213	0.539919	1.004153 1.865056 5.130791
85	321	0.537833	1.120757 1.822010 5.057233
179	321	0.519058	0.988815 2.067136 4.944049
363	321	0.513820	0.993675 2.054628 4.951697
751	231	0.503264	1.007408 2.009480 4.983112
13052	213	0.501541	0.999541 2.006778 4.993681
51783	231	0.500970	1.001485 2.007749 4.990766
161448	231	0.500631	1.001784 1.999454 4.998762
394890	231	0.500541	1.000720 2.005067 4.994213
1097067	231	0.500420	1.000829 2.002094 4.997077
1347455	312	0.500412	1.000807 2.002092 4.997100
1483751	132	0.500085	1.000198 1.999638 5.000164
62565625	132	0.500073	1.000213 1.999673 5.000114
90476880	132	0.500062	1.000012 2.000374 4.999613
143978185	321	0.500048	1.000032 2.000369 4.999599
178782085	312	0.500024	1.000037 2.000189 4.999775

undistinct points, handling continuous sets is more difficult, because the set of authorized permutations must be redefined. This latter remark applies to the direct symmetry index. Thus, the extension of CHI and DSI to continuous sets will be examined in a further work.

The most chiral triangles and the most dissymmetric triangles offer the same remarkable geometric property. Its extension to higher-dimensional simplices is an open problem.

The chiral index and the direct symmetry index provides a coherent quantification of rotational symmetries carrying more information than a boolean value indicating the presence or absence of such symmetries. Although a perfect symmetry can be destroyed when a small perturbation is applied, the ability to quantify proper and improper rotational symmetries provides a robust tool to overcome this problem. As a by-product of computing CHI or DSI, the axe and angle associated to the optimal quaternion locate nonambiguously the symmetry element. When computing either CHI or DSI, if more than one permutation leads to small values of the indices, the set of optimal quaternions provides informations about the existence of more than one symmetry element. Building an automated procedure returning all symmetry elements of a perturbated symmetric set is currently investigated.

ACKNOWLEDGMENT

The author thanks Professor D. Avnir for useful discussions and encouragement to publish this work.

APPENDIX: THE OPTIMAL ROTATION FOR SPATIAL SETS

In this section, the centering condition is not assumed, and $d=3$. X0 and X1 are any n rows and 3 columns arrays of coordinates. The identity permutation $P=\mathbf{I}$ is assumed, but the final result will be valid for any P with replacing X1 by $P\cdot\text{X1}$. The well-known Procrustes algorithm[16,17] used to find the optimal orthogonal transformation superposing two d-sets does not work for enantiomers, because it leads always to $D=0$ and the optimal orthogonal matrix has a negative determinant $(\text{Det}=-1)$. Some iterative procedures are available,[18,19] but the final expression of the optimal rotation was found by Diamond,[20] leading to express D^2 with a quadratic form of the quaternion associated to the rotation. This quadratic form is maximized by an orthonormal basis of four quaternions. For convenience, the expression of the optimal rotation is retrieved here following a different presentation.

A 3D-rotation R is associated to a 3D rotation axis u and to a rotation angle r. This is expressed with a quaternion $q=(p,u)$, with $p=\cos(r/2)$ and $\|u\|=\langle u|u\rangle^{1/2}=\sin(r/2)$. Thus $\langle q|q\rangle=1$, and the image of a point x through R is[21] $Rx=(1-2\langle u|u\rangle)x+2\langle u|x\rangle u+2p(u\wedge x)$. Because $(-p,-u)$ is the same rotation than (p,u), p is always taken non-negative, i.e., r takes values from 0 to 180 degrees.

Let c be the sum of the n vectors $x1_i\wedge x0_i$. Thus, we have $c(1)=\mathrm{V10}(2,3)-\mathrm{V10}(3,2)$, $c(2)=\mathrm{V10}(3,1)-\mathrm{V10}(1,3)$, and $c(3)=\mathrm{V10}(1,2)-\mathrm{V10}(2,1)$. The matrix A is defined as $A=(\mathrm{V10}+\mathrm{V01})-\mathrm{Tr}(\mathrm{V10}+\mathrm{V01})\cdot\mathbf{I}$. Let $\mathrm{D0}^2$ be the initial sum of squares, prior to rotating X1. Now, we have the following equalities: $\langle \mathrm{R}x1_i|x0_i\rangle=(1-2\langle u|u\rangle)\langle x1_i|x0_i\rangle+2\langle u|x1_i\rangle\langle u|x0_i\rangle+2p\langle u\wedge x1_i|x0_i\rangle=(1-2\langle u|u\rangle)\langle x1_i|x0_i\rangle+\langle u|(x1_i\cdot x0_i'+x0_i\cdot x1_i')u\rangle+2p\langle u|x1_i\wedge x0_i\rangle$. V10 is the sum of the n matrices $x1_i\cdot x0_i'$, and $\mathrm{Tr}(V10)$ is the sum of the n quantities $\langle x1_i|x0_i\rangle$. Thus we get $D^2=\mathrm{D0}^2-2\langle u|Au\rangle-4p\langle u|c\rangle$. Let us define the 4×4 matrix B:

$$B=\begin{pmatrix}0 & c'\\ c & A\end{pmatrix},$$

$q=(p,u)$ being the unknown quaternion; it follows that: $D^2=\mathrm{D0}^2-2\langle q|Bq\rangle$.

B is a constant symmetric matrix depending only on the input data, and the quadratic form $\langle q|Bq\rangle$ has to be maximized, q being a unit vector. This problem has a well-known solution:[17] the stationary points are an othonormal basis eigenvectors of B, and the associated eigenvalues are the optimal values of the quadratic form. The sense of each eigenvector is known because p must be non-negative. It is unimportant to get $+u$ or $-u$ when $p=0$. Let L1, L2, L3, L4 be the eigenvalues arranged in decreasing order.

B is the sum of two 4×4 symmetric matrices. One contains only A and zeros on the first row and column. Let B1 be this matrix. The other contains only c' on the right of the first row and c on the bottom of the first column, zero as a first diagonal element, and nine zeros in the remaining 3×3 block. Let B2 be this one-rank matrix, of which the four eigenvalues are obviously $\|c\|$ and zero with three as multiplicity. Let d1, d2, d3 be the eigenvalues of A arranged in decreasing order. Thus, the following inequalities stand:[22] the eigenvalues of A separate those of B: $\mathrm{L1}\geqslant\mathrm{d1}\geqslant\mathrm{L2}\geqslant\mathrm{d2}\geqslant\mathrm{L3}\geqslant\mathrm{d3}\geqslant\mathrm{L4}$, and $|\mathrm{Li}-\mathrm{di}'|\leqslant\|c\|$ for $i=1,2,3,4$, di' being the ith greatest value among $(0,\mathrm{d1},\mathrm{d2},\mathrm{d3})$.

Two situations may arise. If d1 and d3 have not the same sign, the first set of inequalities means that L1 and L4 have not the same sign. If d1 and d3 have the same sign, let us look at the determinant of B expressed after diagonalization of A. The components of c become c(1), c(2), c(3), and $\det(B)=-\mathrm{c}(1)^2\cdot\mathrm{d2}\cdot\mathrm{d3}-\mathrm{c}(2)^2\cdot\mathrm{d1}\cdot\mathrm{d3}-\mathrm{c}(3)^2\cdot\mathrm{d1}\cdot\mathrm{d2}$. This determinant cannot be positive, thus again L1 and L4 cannot have the same sign.

Thus L1 is always non-negative and L4 is always non-positive. The rotation minimizing D^2 is those associated to the quaternion q1, such that $D^2=\mathrm{D0}^2-2\mathrm{L1}$, and the rotation associated to q4 such that $D^2=\mathrm{D0}^2-2\mathrm{L4}$ is that which maximizes D^2. D^2 has one saddle point associated to q2 and one associated to q3.

Some minor properties of the four optimal quaternions are obtained from their othonormality. Considering the first row of the equation Bq=Lq, it comes that $\mathrm{D0}^2-D^2=2\mathrm{L}=2\langle v|c\rangle$, with $v=u/p$. It shows than only a positive L value leads to $D^2<\mathrm{D0}^2$. The three others equations may be rewritten: $(A-\langle v|n\rangle\mathbf{I})v+n=0$, but this is neither an eigenvector equation nor a linear system. Two distinct directions ui and uj are generally not orthogonal: $\cos(ui,uj)=-pi\cdot pj/((1-pi^2)\cdot(1-pj^2))^{1/2}=-\cot g(ri)\cdot\cot g(rj)$, ri and rj being the 2D-angles associated, respectively, to qi and qj.

[1] A. Rassat, ''Un critère de classement des systèmes chiraux de points à partir de la distance au sens de Haussdorf,'' Compt. Rend. Acad. Sci. Paris **299**, Sér. II, 53–55 (1984).

[2] Y. Pinto, P. W. Fowler, D. Mitchell, and D. Avnir, ''Continuous chirality analysis of model Stone–Wales rearrangements in fullerenes,'' J. Phys. Chem. B **102**, 5776–5784 (1998).
[3] D. R. Kanis, J. S. Wong, T. J. Marks, M. A. Ratner, H. Zabrodsky, S. Keinan, and D. Avnir, ''Continuous symmetry analysis of hyperpolarizabilities. Characterization of the second-order nonlinear optical response of distorted benzene,'' J. Phys. Chem. **99**, 11061–11066 (1995).
[4] V. Buch, E. Gershgoren, H. Zabrodsky Hel-Or, and D. Avnir, ''Symmetry loss as a criterion for cluster melting, with application to (D2)13,'' Chem. Phys. Lett. **247**, 149–153 (1995).
[5] H. Zabrodsky and D. Avnir, ''Continuous symmetry measures. 4. Chirality,'' J. Am. Chem. Soc. **117**, 462–473 (1995).
[6] G. Gilat and Y. Gordon, ''Geometric properties of chiral bodies,'' J. Math. Chem. **16**, 37–48 (1994).
[7] J. Maruani, G. Gilat, and R. Veysseyre, ''Chirality limits of convex bodies,'' Compt. Rend. Acad. Sci. Paris **319**, Sér. II, 1165–1172 (1994).
[8] N. Weinberg and K. Mislow, ''Distance functions as generators of chirality measures,'' J. Math. Chem. **14**, 427–450 (1993).
[9] Z. Zimpel, ''A geometrical approach to the degree of chirality and asymmetry,'' J. Math. Chem. **14**, 451–465 (1993).
[10] J. Maruani and P. G. Mezey, ''From symmetry to syntopy: a fuzzy-set approach to quasi-symmetric systems,'' J. Chim. Phys. **87**, 1025–1047 (1990).
[11] M. Petitjean, ''Calcul de chiralité quantitative par la méthode des moindres carrés,'' Compt. Rend. Acad. Sci. Paris, Sér. IIc **2**, 25–28 (1999).
[12] M. Petitjean, ''About second kind continuous chirality measures. 1. Planar sets,'' J. Math. Chem. **22**, 185–201 (1997).
[13] M. S. Srivastava and C. G. Khatri, *An Introduction to Multivariate Statistics* (Elsevier, New York, 1979), Chap. 3.
[14] M. S. Bartlett, ''The effect of standardization on a χ^2 approximation in factor analysis,'' Biometrika **38**, 337–344 (1951).
[15] F. J. Schuurmann, P. R. Krishnaiah, and A. K. Chattopadhyay, ''On the distributions of the ratios of the extreme roots to the trace of the Wishart matrix,'' J. Multivariate Anal. **3**, 445–453 (1973).
[16] J. R. Hurley and R. B. Cattell, ''The Procrustes Program: Producing direct rotation to test a hypothesized factor structure,'' Behav. Sci. **7**, 258–262 (1962).
[17] G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore, 1985), Secs. 8.1 and 12.4.
[18] A. D. McLachlan, ''Rapid comparison of protein structures,'' Acta Crystallogr. Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **38**, 871–873 (1982).
[19] M. J. Sippl and H. Stegbuchner, ''Superposition of three-dimensional objects: A fast and numerically stable algorithm for the calculation of the matrix of optimal rotation,'' Comput. Chem. **15**, 73–78 (1991).
[20] R. Diamond, ''A note on the rotational superposition problem,'' Acta Crystallogr., Sect. A: Found. Crystallogr. **44**, 211–216 (1988).
[21] R. Deheuvels, *Formes Quadratiques et Groupes Classiques* (Presses Universitaires de France, Paris, 1981), p. 375.
[22] J. H. Wilkinson, *The Algebraic Eigenvalue Problem* (Clarendon, Oxford, 1965), Secs. 39–41.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 9 SEPTEMBER 1999

Operator representations of a *q*-deformed Heisenberg algebra

Konrad Schmüdgen[a)]
Universität Leipzig, Fakultät für Mathematik und Informatik und NTZ, Augustusplatz 10/11, D-04109 Leipzig, Germany

(Received 3 June 1998; accepted for publication 9 July 1998)

A class of well-behaved $*$-representations of a q-deformed Heisenberg algebra previously introduced [Phys. Lett. B **291**, 273 (1992); Z. Phys. C **64**, 335 (1994)] is studied and classified.
[S0022-2488(99)02008-3]

I. INTRODUCTION

The idea to develop a q-deformed quantum mechanics by using quantum groups has been investigated in several papers.[1–5] Such approaches are usually based on a q-deformed phase space algebra which is derived from the noncommutative differential calculus of the q-deformed configuration space.[6,7] Following the standard procedure in quantum mechanics one has to represent the q-deformed position and momentum operators by essentially self-adjoint operators acting on a Hilbert space. More precisely, one has to find appropriate $*$-representations of the phase space $*$-algebra by unbounded operators in a Hilbert space. In the case of general Euclidean or Minykowski phase spaces the study and classification of these $*$-representations turns out to be technically complicated because of the many relations and also because of the various difficulties concerning unbounded operators.

The aim of this paper is to give a rigorous treatment of well-behaved operator representations for one of the simplest example—the one-dimensional q-deformed Heisenberg algebra which was invented in Refs. 4 and 2. Representations of this algebra have been investigated in Ref. 2. Since this $*$-algebra occurs as a subalgebra of other larger $*$-algebras, the study of general not necessarily irreducible $*$-representations seems to be important as well. We shall develop and analyze an operator-theoretic model for such general representations of the q-deformed Heisenberg algebra. This model might be used as a tool kit for the study of representations of larger $*$-algebras.

This paper is organized as follows. Section II contains the definition and some simple algebraic properties of the q-deformed Heisenberg algebra $\mathcal{A}(q)$. In Sec. III we develop a general operator-theoretic model for certain triples of operators which will lead in Sec. VI to representations of the $*$-algebra $\mathcal{A}(q)$. In Sec. IV the irreducibility and the unitary equivalence of these operator triples are investigated and a number of examples are treated. In Sec. V we give a characterization of these operator triples by a number of natural conditions. In Sec. VI we define the self-adjoint $*$-representations of the $*$-algebra $\mathcal{A}(q)$ obtained by means of these operator triples.

In a forthcoming paper we shall study the spectrum of the operator X. For this analysis the q-Fourier transform[8,9] will play a crucial role.

II. THE *q*-HEISENBERG ALGEBRA

For a positive real number $q \neq 1$, let $\mathcal{A}(q)$ denote the complex unital algebra with four generators $\mathbf{p},\mathbf{x},\mathbf{u},\mathbf{u}^{-1}$ subject to the defining relations

$$\mathbf{up} = q\mathbf{pu}, \quad \mathbf{ux} = q^{-1}\mathbf{xu}, \quad \mathbf{uu}^{-1} = \mathbf{u}^{-1}\mathbf{u} = 1, \tag{1}$$

[a)]Electronic mail: schmuedgen@mathematik.uni-leipzig.de

$$\mathbf{px}-q\mathbf{xp}=\mathrm{i}(q^{3/2}-q^{-1/2})\mathbf{u},\quad \mathbf{xp}-q\mathbf{px}=-\mathrm{i}(q^{3/2}-q^{-1/2})\mathbf{u}^{-1},\tag{2}$$

where i denotes the imaginary unit. An equivalent set of relations is obtained if (2) is replaced by

$$\mathbf{px}=\mathrm{i}q^{1/2}\mathbf{u}^{-1}-\mathrm{i}q^{-1/2}\mathbf{u},\quad \mathbf{xp}=\mathrm{i}q^{-1/2}\mathbf{u}^{-1}-\mathrm{i}q^{1/2}\mathbf{u}.\tag{2$'$}$$

From (1) and (2)$'$ it follows that the set of elements $\{\mathbf{p}^r\mathbf{u}^n,\mathbf{x}^s\mathbf{u}^n;\ r\in\mathbb{N}_0,\ s\in\mathbb{N},\ n\in\mathbb{Z}\}$ is a vector space basis of $\mathcal{A}(q)$.

The algebra $\mathcal{A}(q)$ becomes a $*$-algebra with involution defined on the generators by

$$\mathbf{p}=\mathbf{p}^*,\quad \mathbf{x}=\mathbf{x}^*,\quad \mathbf{u}^*=\mathbf{u}^{-1}.\tag{3}$$

Indeed, it suffices to check that the defining relations (1) and (2)$'$ of $\mathcal{A}(q)$ are invariant under the involution (3) which is easily done.

From (1), (2)$'$, and (3) we conclude that there are $*$-isomorphisms ρ_1 and ρ_2 of the $*$-algebras $\mathcal{A}(q)$ and $\mathcal{A}(q^{-1})$ such that

$$\rho_1(\mathbf{x})=\mathbf{p},\quad \rho_1(\mathbf{p})=\mathbf{x},\quad \rho_1(\mathbf{u})=\mathbf{u},\quad \text{and}\quad \rho_2(\mathbf{x})=\mathbf{x},\quad \rho_2(\mathbf{p})=\mathbf{p},\quad \rho_2(\mathbf{u})=-\mathbf{u}^*.$$

Because the $*$-algebras $\mathcal{A}(q)$ and $\mathcal{A}(q^{-1})$ are isomorphic, we shall assume in what follows that $0<q<1$.

III. AN OPERATOR-THEORETIC MODEL

(1) Let μ_1 be a finite positive Borel measure on the interval $[q,1)$. The measure μ_1 extends uniquely to a Borel measure μ on the half-axis $\mathbb{R}_+=(0,+\infty)$ by setting $\mu(q^n\mathcal{M}):=q^n\mu_1(\mathcal{M})$ for any Borel subset $\mathcal{M}$ of $[q,1)$. Then μ has the property that $\mu(q\mathcal{N})=q\mu(\mathcal{N})$ for an arbitrary Borel subset $\mathcal{N}$ of $\mathbb{R}_+$ or equivalently that $d\mu(qt)/qt=d\mu(t)/t$ for $t\in\mathbb{R}_+$. We shall work with the Hilbert spaces $\mathcal{H}:=L^2(\mathbb{R}_+,\mu)$ and $\mathfrak{H}:=L^2([q,1),\mu_1)$. First we define three linear operators *U, P*, and *X* on the Hilbert space $\mathcal{H}$:

(i) $(Uf)(t)=q^{1/2}f(qt)$ for $f\in\mathcal{H}$,
(ii) $(Pf)(t)=tf(t)$ for $f\in\mathcal{D}(P):=\{f\in\mathcal{H}:tf(t)\in\mathcal{H}\}$,
(iii) $(Xf)(t)=\mathrm{i}t^{-1}(f(q^{-1}t)-f(qt))$ for $f\in\mathcal{D}(X):=\{f\in\mathcal{H}:t^{-1}f(t)\in\mathcal{H}\}$.

These operators will play a crucial role throughout this paper. Roughly speaking and ignoring technical subtleties (domains, boundary conditions, etc.), we shall show that for all ''well-behaved'' $*$-representations of the q-deformed Heisenberg algebra $\mathcal{A}(q)$ the images of the generators **u, p**, and **x** act by the same formulas as the operators *U, P*, and *X*, respectively.

Obviously, P is an unbounded self-adjoint operator on $\mathcal{H}$. Using the relation $d\mu(qt)/qt=d\mu(t)/t$ one easily verifies that U is a unitary operator and that X is a symmetric operator on $\mathcal{H}$. Let $\mathcal{D}_0$ be the set of functions $f\in\mathcal{H}$ such that supp $f\in[a,b]$ for some $a>0$ and $b>0$. (Note that a and b may depend on f.) Clearly, $\mathcal{D}_0$ is dense linear subspace of $\mathcal{H}$ which is invariant under *U, P*, and *X*. It is straightforward to check that the operators *P, X*, and *U* applied to functions $f\in\mathcal{D}_0$ satisfy the defining relations (1), (2), and (3) of the $*$-algebra $\mathcal{A}(q)$. In turns out that the symmetric operator X is not essentially self-adjoint. Our next aim is to characterize the domain of the adjoint operator X^*.

For $f\in\mathfrak{H}=L^2([q,1),\mu_1)$ let f^e and f^o be the functions on $\mathbb{R}_+$ defined by

$$f^e(q^{2n}t)=f^o(q^{2n+1}t)=f(t)\quad \text{for } n\in\mathbb{N}_0,\quad t\in[q,1),\quad \text{and } f^e(t)=f^o(t)=0\ \text{ otherwise.}\tag{4}$$

Clearly, f^e and f^o are in $\mathcal{H}=L^2(\mathbb{R}_+,\mu)$ and we have $U(f^e)-q^{1/2}f\in\mathcal{D}(X)$ and $U(f^o)-q^{1/2}f^e\in\mathcal{D}(X)$. Let $\mathfrak{H}_e$ and $\mathfrak{H}_o$ denote the set of functions f^e and f^o, respectively, where $f\in\mathfrak{H}=L^2([q,1),\mu_1)$.

Lemma 1: The domain $\mathcal{D}(X^*)$ *is the direct sum of vector spaces* $\mathcal{D}(X)$, $\mathfrak{H}_e$ *and* $\mathfrak{H}_o$.

Proof: It is straightforward to check that $\mathcal{D}(X)+\mathfrak{H}_e+\mathfrak{H}_o\subseteq\mathcal{D}(X^*)$. In order to prove the converse, let $g\in\mathcal{D}(X^*)$. Then, by definition there is an $h\in\mathcal{H}$ such that $\langle Xf,g\rangle=\langle f,h\rangle$ for all $f\in\mathcal{D}(X)$. Inserting the definition of X and using once more the fact that $\mu(q\mathcal{N})=q\mu(\mathcal{N})$ for an arbitrary Borel subset $\mathcal{N}$ of $\mathbb{R}_+$ we easily conclude that $h(t)=it^{-1}(g(q^{-1}t)-g(qt))$. For a function $f\in\mathcal{H}$ let f_n denote the function in $L^2([q,1),\mu_1^+)$ given by $f_n(t)=f(q^nt)$. Then we get

$$\|h\|^2_{L^2(\mathbb{R}_+,\mu)}=\sum_{n=-\infty}^{\infty}\|h_n\|^2q^n\geqslant\sum_{n=0}^{\infty}\frac{\|g_{n+1}-g_{n-1}\|^2}{q^{2n}}q^n.$$

For $n\in\mathbb{N}$ we set $\alpha_n:=\|g_{n+1}-g_{n-1}\|q^{-n/2}$. Since $h\in L^2(\mathbb{R},\mu)$, the sequence (α_n) is in l_2. From the inequality

$$\|g_{2r}-g_{2s}\|\leqslant\alpha_{2r+1}q^{(2r+1)/2}+\cdots+\alpha_{2s+1}q^{(2s+1)/2}$$

we obtain

$$\|g_{2r}-g_{2s}\|^2\leqslant\left(\sum_{i=2s+1}^{\infty}|\alpha_i|^2\right)q^{2s+1}(1-q^2)^{-1},\quad r\geqslant s. \tag{5}$$

Since $(\alpha_n)\in l_2$, this implies that the sequence $(g_{2n})_{n\in\mathbb{N}}$ converges in the Hilbert space $L^2([q,1),\mu_1)$. Let us denote its limit by ξ. We extend ξ to a function ξ^e on $\mathbb{R}_+$ by setting $\xi^e(q^{2n}t):=\xi(t)$ and $\xi^e(q^{2n+1}t):=0$ for $n\in\mathbb{N}_0$, $t\in[q,1)$, and $\xi^e(t)=0$ for $t\geqslant1$. Replacing even indices by odd indices, a similar reasoning yields functions $\zeta\in L^2([q,1),\mu_1)$ and ζ^o on $\mathbb{R}_+$ such that $\zeta^o(q^{2n+1}t)=\zeta(t)$ and $\zeta^o(q^{2n}t)=0$ for $n\in\mathbb{N}$, $t\in[q,1)$, and $\zeta^o(t)=0$ for $t\geqslant1$. By construction, $\xi^e\in\mathfrak{H}_e$ and $\zeta^o\in\mathfrak{H}_o$. Our proof is complete once we have shown that $f:=g-\xi^e-\zeta^o$ belongs to the domain $\mathcal{D}(X)$ of the operator X.

Letting $r\to\infty$ in (5), we get

$$\|\xi-g_{2s}\|^2\leqslant q^{2s+1}(1-q^2)^{-1}\sum_{n=0}^{\infty}|\alpha_{2n}|^2. \tag{6}$$

From (6) and the corresponding estimation of $\|\zeta-g_{2s+1}\|^2$ we obtain

$$\sum_{n=0}^{\infty}\|t^{-1}f_n(t)\|^2q^n\leqslant\sum_{n=0}^{\infty}\frac{\|f_n\|^2}{q^{2n+2}}q^n=\sum_{r=0}^{\infty}\frac{\|\xi-g_{2r}\|^2}{q^{2r+2}}+\frac{\|\zeta-g_{2r+1}\|^2}{q^{2r+3}}=(q-q^3)^{-1}\sum_{n=0}^{\infty}|\alpha_n|^2<\infty.$$

Since $f(t)=g(t)$ for $t\geqslant1$, this inequality implies that the functions $t^{-1}f(t)$ and $f(t)$ are in $L^2(\mathbb{R}^+,\mu)$. Thus, $f\in\mathcal{D}(X)$. ∎

As shown in the preceding proof, for any function $g\in\mathcal{D}(X^*)$ the "even components" g_{2n} and the "odd components" g_{2n+1} both have "boundary limits" ξ and ζ in $L^2([q,1),\mu_1)$. By Lemma 1, any element $f\in\mathcal{D}(X^*)$ is of the form $f=f_X+f^e+f^o$ with uniquely determined functions $f_X\in\mathcal{D}(X)$, $f^e\in\mathcal{H}_e$ and $f^o\in\mathcal{H}_o$. By the definition of $\mathcal{H}_e$ and $\mathcal{H}_o$, there exist unique functions f_e, $f_o\in\mathfrak{H}=L^2([q,1),\mu_1)$ such that $(f_e)^e=f^e$ and $(f_o)^o=f^o$, where the function $(f_e)^e$ and $(f_o)^o$ on $\mathbb{R}$ are given by (4). This notation will be kept in the sequel.

Let $\langle\cdot,\cdot\rangle$ and $(\cdot,\cdot)$ denote the scalar products of the Hilbert spaces $L^2(\mathbb{R}_+,\mu)$ and $L^2([q,1),t^{-1}\mu_1)$, respectively.

Lemma 2: For arbitrary functions $f,g\in\mathcal{D}(X^*)$ we have

$$\langle X^*f,g\rangle-\langle f,X^*g\rangle=\frac{1}{2\mathrm{i}}\{(f_e+f_o,g_e+g_o)-(f_e-f_o,g_e-g_o)\}. \tag{7$_+$}$$

Proof: Let $h\in L^2(([q,1),\mu_1)$. From the definitions of the operator X and of the functions $h^e,h^o\in L^2(\mathbb{R},\mu)$ we easily derive that $(X^*h^e)(t)=-it^{-1}h(qt)$ for $t\in[1,q^{-1})$, $(X^*h^e)(t)=0$ for t

$\in \mathbb{R}_+ \backslash [1,q^{-1})$, $(X^* h^o)(t) = -it^{-1}h(t)$ for $t \in [q,1)$, and $(X^* h^o)(t) = 0$ for $t \in \mathbb{R}_+ \backslash [q,1)$. Inserting these expressions and using the symmetry of the operator X we compute

$$\langle X^* f,g\rangle - \langle f,X^* g\rangle = \langle X^* f_o, g_e\rangle - \langle f_e, X^* g_o\rangle$$
$$= -\mathrm{i}\int_q^1 (f_o(t)\overline{g_e(t)} + f_e(t)\overline{g_o(t)})t^{-1}\, d\mu(t)$$
$$= -\mathrm{i}\{(f_o, g_e) + (f_e, g_o)\} = \frac{1}{2\mathrm{i}}\{(f_e + f_o, g_e + g_o) - (f_e - f_o, g_e - g_o)\}.$$

■

Let us illustrate the preceding by the simplest example.

Example 1: Let μ_1 be the Delta measure δ_a, where a is a fixed number from the interval $[q,1)$. Then the measure μ is supported on the points $aq^n, n \in \mathbb{Z}$, and we have $\mu(\{aq^n\}) = q^n \mu(\{a\}) = q^n$. Hence the scalar product of the Hilbert space $\mathcal{H} = L^2(\mathbb{R}_+, \mu)$ is given by the Jackson integral

$$\langle f,g\rangle = \sum_{n=-\infty}^{+\infty} f(aq^n)\overline{g(aq^n)}q^n.$$

Let $e_n \in \mathcal{H}$ be the function $e_n(t) = q^{-n/2}\delta^t_{aq^n}$, where δ^t_s is the usual Kronecker symbol. Then the vectors $e_n, n \in \mathbb{Z}$, form an orthonormal basis of $\mathcal{H}$ and the actions of the operators U, P, and X on these vectors are given by

$$Ue_n = e_{n-1}, \quad Pe_n = aq^n e_n, \quad Xe_n = \frac{\mathrm{i}}{aq^n}(q^{-1/2}e_{n+1} - q^{1/2}e_{n-1}).$$

These equations are in accordance with formulas (5) in Ref. 2. If f is the function in $L^2([q,1),\mu_1) \cong \mathbb{C}$ with $f(a) = 1$, then by definition $f^e(aq^{2n}) = f^o(aq^{2n+1}) = 1$, $f^e(aq^{2n+1}) = f^o(aq^{2n}) = 0$ for $n \in \mathbb{N}_0$, and $f^e(t) = f^o(t) = 0$ for $t \geqslant 1$. Then we have $\mathcal{D}(X^*) = \mathcal{D}(X) + \mathbb{C}\cdot f^e + \mathbb{C}\cdot f^o$ by Lemma 1 and formula $(7)_+$ reads as

$$\langle X^*(\varphi + \alpha_1 f^e + \beta_1 f^o), \psi + \alpha_2 f^e + \beta_2 f^o\rangle - \langle \varphi + \alpha_1 f^e + \beta_1 f^o, X^*(\psi + \alpha_2 f^e + \beta_2 f^o)\rangle$$
$$= -\mathrm{i}a^{-1}\{\beta_1\overline{\alpha_2} + \alpha_1\overline{\beta_2}\}$$
$$= \frac{1}{2\mathrm{i}a}\{(\alpha_1 + \beta_1)\overline{(\alpha_2 + \beta_2)} - (\alpha_1 - \beta_1)\overline{(\alpha_2 - \beta_2)}\}$$

for $\alpha_1, \beta_1, \alpha_2, \beta_2 \in \mathbb{C}$. ■

(2) The above considerations carry over almost verbatim to the case where the positive half-axis $\mathbb{R}_+$ is replaced by the negative half-axis $\mathbb{R}_- = (-\infty,0)$. Any positive finite Borel measure μ_1 on the interval $[q,1)$ induces a positive Borel measure μ on $\mathbb{R}_-$ by defining $\mu(-q^n\mathcal{M}) := q^n\mu_1(\mathcal{M})$ for a Borel subset $\mathcal{M}$ of $[q,1)$. The operators U, P, and X on the Hilbert space $\mathcal{H}_- := L^2(\mathbb{R}_-, \mu)$ are defined by the same formulas as in the preceding subsection and Lemma 1 and its proof remain valid in this case as well. However, there is an essential difference which will be crucial in the sequel: Since in the proof of Lemma 2 the integration is over the interval $(-1,-q]$, the expression on the right-hand side of $(7)_+$ must be multiplied by -1. That is, instead of $(7)_+$ we now have

$$\langle X^* f,g\rangle - \langle f,X^* g\rangle = \frac{1}{2\mathrm{i}}\{(f_e + f_o, g_e + g_o) - (f_e - f_o, g_e - g_o)\} \qquad (7)_-$$

for $f,g \in \mathcal{D}(X^*)$.

(3) After the preceding preparations we are now able to develop the operator-theoretic model for the description of $*$-representations of the q-Heisenberg algebra $\mathcal{A}(q)$. For this let us fix two families $\{\mu_1^{j,+};j\in I_+\}$ and $\{\mu_1^{j,-};j\in I_-\}$ of finite positive Borel measures on the interval $[q,1)$.

As above, we define the Hilbert spaces $\mathcal{H}_{j,\pm}:=L^2(\mathbb{R}_\pm,\mu^{j,\pm})$, $j\in I_\pm$, and the operators $U_{j,\pm}$, $P_{j,\pm}$, and $X_{j,\pm}$ acting therein. We shall work with the representation Hilbert space $\mathcal{H}=\mathcal{H}_+\oplus\mathcal{H}_-$, where $\mathcal{H}_+:=\oplus_{j\in I_+}\mathcal{H}_{j,+}$ and $\mathcal{H}_-:=\oplus_{j\in I_-}\mathcal{H}_{j,-}$. The elements of $\mathcal{H}$ are pairs $\mathfrak{f}=(\mathfrak{f}^+,\mathfrak{f}^-)$, where $\mathfrak{f}^+=(f^{j,+};j\in I_+)\in\mathcal{H}_+$ and $\mathfrak{f}^-=(f^{j,-};j\in I)\in\mathcal{H}_-$. Let U, P, and X denote the operators on $\mathcal{H}$ which are defined as the direct sums of the operators $U_{j,+},U_{j,-}$; $P_{j,+},P_{j,-}$; and $X_{j,+},X_{j,-}$, respectively. Clearly, U is a unitary operator and P is a self-adjoint operator on $\mathcal{H}$. The operator X is only symmetric, but not self-adjoint. Our next aim is to describe all self-adjoint extensions $\tilde{X}$ of X on $\mathcal{H}$ which have the property that $U\tilde{X}U^{-1}=q\tilde{X}$.

Let V and W be two unitary linear transformations of the Hilbert space $\mathfrak{H}_-:=\oplus_{j\in I_-}L^2([q,1),t^{-1}\mu_1^{j,-})$ on the Hilbert space $\mathfrak{H}_+:=\oplus_{j\in I_+}L^2([q,1),t^{-1}\mu_1^{j,+})$. We define a linear operator $X_{V,W}$ as being the restriction of the adjoint operator X^* to the domain

$$\mathcal{D}(X_{V,W}):=\{\mathfrak{f}=\mathfrak{f}_X+\mathfrak{f}^e+\mathfrak{f}^o\in\mathcal{D}(X^*):\mathfrak{f}_X\in\mathcal{D}(X),$$
$$\mathfrak{f}_e^+=V(\mathfrak{f}_e^-+\mathfrak{f}_0^-)+W(\mathfrak{f}_e^--\mathfrak{f}_0^-),\ \ \mathfrak{f}_0^+=V(\mathfrak{f}_e^++\mathfrak{f}_0^-)-W(\mathfrak{f}_e^--\mathfrak{f}_0^-). \tag{8}$$

Proposition 1: $X_{V,W}$ is a self-adjoint operator on $\mathcal{H}$ such that $X\subseteq X_{V,W}$ and $UX_{V,W}U^*=qX_{V,W}$. In particular, we have $U\mathcal{D}(X_{V,W})=\mathcal{D}(X_{V,W})$. Conversely, for any self-adjoint extension $\tilde{X}$ of X satisfying $U\mathcal{D}(\tilde{X})\subseteq\mathcal{D}(\tilde{X})$ there exist unitary transformations V,W of $\mathfrak{H}_+$ onto $\mathfrak{H}_-$ such that $\tilde{X}=X_{V,W}$.

Proof: From $(7)_+$ and $(7)_-$ we obtain

$$-2\mathrm{i}(\langle X^*\mathfrak{f},\mathfrak{g}\rangle-\langle\mathfrak{f},X^*\mathfrak{g}\rangle)=(\mathfrak{f}_e^++\mathfrak{f}_o^+,\mathfrak{g}_e^++\mathfrak{g}_o^+)+(\mathfrak{f}_e^--\mathfrak{f}_o^-,\mathfrak{g}_e^--\mathfrak{g}_o^-)-(\mathfrak{f}_e^-+\mathfrak{f}_o^-,\mathfrak{g}_e^-+\mathfrak{g}_o^-)$$
$$-(\mathfrak{f}_e^+-\mathfrak{f}_o^+,\mathfrak{g}_e^+-\mathfrak{g}_o^+) \tag{9}$$

for arbitrary elements $\mathfrak{f}=\mathfrak{f}_X+\mathfrak{f}^e+\mathfrak{f}^o$ and $\mathfrak{g}=\mathfrak{g}_X+\mathfrak{g}^e+\mathfrak{g}^o$ of $\mathcal{D}(X^*)$. Here $\mathfrak{f}_e^+$ denotes the sequence $(f_e^{j,+};j\in I_+)\in\mathfrak{H}_+$ with $f_e^{j,+}\in L^2([q,1),\mu_1^j)$ such that the extension $(f_e^{j,+})^e$ of $f_e^{j,+}$ to $\mathbb{R}_+$ by means of formula (4) is just the $(j,+)$-component of the vector $\mathfrak{f}^e\in\mathcal{H}$. A similar meaning attached to the other symbols $f_e^-,f_o^+,f_o^-,\mathfrak{g}_e^+,\mathfrak{g}_o^-,\mathfrak{g}_o^+,\mathfrak{g}_o^-$ occuring in (9). If $\mathfrak{f},\mathfrak{g}\in\mathcal{D}(X_{V,W})$, then we have $f_e^++\mathfrak{f}_o^-=V(\mathfrak{f}_e^-+\mathfrak{f}_o^-)$, $\mathfrak{g}_e^++\mathfrak{g}_o^-=V(\mathfrak{g}_e^-+\mathfrak{g}_o^-)$, $\mathfrak{f}_e^+-\mathfrak{f}_o^+=W(\mathfrak{f}_e^--\mathfrak{f}_o^-)$, and $\mathfrak{g}_e^+-\mathfrak{g}_o^+=W(\mathfrak{g}_e^--\mathfrak{g}_o^-)$ by (8). Since $X_{V,W}\subseteq X^*$, we therefore obtain that $\langle X_{V,W}\mathfrak{f},\mathfrak{g}\rangle-\langle\mathfrak{f},X_{V,W}\mathfrak{g}\rangle=0$ by (9), that is, the operator $X_{V,W}$ is symmetric. Now let $\mathfrak{g}\in\mathcal{D}((X_{V,W})^*)$. Since $X\subseteq X_{V,W}\subseteq(X_{V,W})^*\subseteq X^*$, we then have $\langle X^*\mathfrak{f},\mathfrak{g}\rangle=\langle\mathfrak{f},X^*\mathfrak{g}\rangle$ and hence

$$(\mathfrak{f}_e^++\mathfrak{f}_o^+,\mathfrak{g}_e^++\mathfrak{g}_o^+)+(\mathfrak{f}_e^--\mathfrak{f}_o^-,\mathfrak{g}_e^--\mathfrak{g}_o^-)=(\mathfrak{f}_e^-+\mathfrak{f}_o^-,\mathfrak{g}_e^-+\mathfrak{g}_o^-)+(\mathfrak{f}_e^+-\mathfrak{f}_o^+,\mathfrak{g}_e^+-\mathfrak{g}_o^+) \tag{10}$$

for all $\mathfrak{f}\in\mathcal{D}(X_{V,W})$ by (9). Inserting (8) into (10), we get

$$(f_e^-+\mathfrak{f}_o^-,V^*(\mathfrak{g}_e^++\mathfrak{g}_o^-))+(\mathfrak{f}_e^--\mathfrak{f}_o^-,\mathfrak{g}_e^--\mathfrak{g}_o^-)=(\mathfrak{f}_e^-+\mathfrak{f}_o^-,\mathfrak{g}_e^-+\mathfrak{g}_o^-)+(\mathfrak{f}_e^--\mathfrak{f}_o^-,W^*(\mathfrak{g}_e^+-\mathfrak{g}_o^+)). \tag{11}$$

From the construction it is clear that for arbitrary $\mathfrak{h},\mathfrak{k}\in\mathfrak{H}_-$ there exists $\mathfrak{f}\in\mathcal{D}(X_{V,W})$ such that $\mathfrak{f}_e^-+\mathfrak{f}_o^-=\mathfrak{h}$ and $\mathfrak{f}_e^--\mathfrak{f}_o^-=\mathfrak{k}$. Therefore, it follows from (11) that $V^*(\mathfrak{g}_e^++\mathfrak{g}_o^+)=\mathfrak{g}_e^++\mathfrak{g}_o^-$ and $W^*(\mathfrak{g}_e^+-\mathfrak{g}_o^+)=\mathfrak{g}_e^--\mathfrak{g}_o^-$, which in turn implies that $\mathfrak{g}\in\mathcal{D}(X_{V,W})$. Thus we have shown that the operator $X_{V,W}$ is self-adjoint. From the relations $U(\mathfrak{f}_e)-q^{1/2}\mathfrak{f}_o\in\mathcal{D}(X)$ and $U(\mathfrak{f}_o)-q^{1/2}\mathfrak{f}_e\in\partial(X)$ we see that $U\mathcal{D}(X_{V,W})=\mathcal{D}(X_{V,W})$. Since $UXU^*=qX$ and hence $UX^*U^*=qX^*$ and $X_{V,W}$ is the restriction of X^* to $\mathcal{D}(X_{V,W})$, the latter yields $UX_{V,W}U^*=qX_{V,W}$.

Conversely, suppose that $\tilde{X}$ is a self-adjoint extension of X such that $U\mathcal{D}(\tilde{X}) \subseteq \mathcal{D}(\tilde{X})$. Since $\tilde{X}$ is symmetric, we have Eq. (10) for arbitrary elements $\mathfrak{f},\mathfrak{g} \in \mathcal{D}(\tilde{X})$. By assumption, $U\mathfrak{f} \in \mathcal{D}(\tilde{X})$ for all $\mathfrak{f} \in \mathcal{D}(\tilde{X})$. Replacing $\mathfrak{f}$ by $U\mathfrak{f}$ in (10) we get

$$(\mathfrak{f}_e^+ + \mathfrak{f}_o^+, \mathfrak{g}_e^+ + \mathfrak{g}_o^+) + (\mathfrak{f}_o^- - \mathfrak{f}_e^-, \mathfrak{g}_e^- - \mathfrak{g}_o^-) = (\mathfrak{f}_e^- + \mathfrak{f}_o^-, \mathfrak{g}_e^- + \mathfrak{g}_o^-) + (\mathfrak{f}_o^+ - \mathfrak{f}_e^+, \mathfrak{g}_e^+ - \mathfrak{g}_o^+). \tag{12}$$

Setting $\mathfrak{f} = \mathfrak{g}$ and combining formulas (10) and (12) we obtain

$$\|\mathfrak{f}_e^+ + \mathfrak{f}_o^+\| = \|\mathfrak{f}_e^- + \mathfrak{f}_o^-\| \quad \text{and} \quad \|\mathfrak{f}_e^+ - \mathfrak{f}_o^+\| = \|\mathfrak{f}_e^- - \mathfrak{f}_o^-\| \tag{13}$$

for all $\mathfrak{f} \in \mathcal{D}(\tilde{X})$.

For $\mathfrak{f} \in \mathcal{D}(X^*)$ we abbreviate $B_\pm(\mathfrak{f}) = (\mathfrak{f}_e^\pm + \mathfrak{f}_0^\pm, \mathfrak{f}_e^\pm - \mathfrak{f}_o^\pm)$. The vector space $B_\pm(\tilde{X}) = \{B_\pm(\mathfrak{f}) : \mathfrak{f} \in \mathcal{D}(\tilde{X})\}$ is called the ''boundary space'' of the operator $\tilde{X}$. We shall show that $B_+(\tilde{X}) = \mathfrak{H}_+ \oplus \mathfrak{H}_+$ and $B_-(\tilde{X}) = \mathfrak{H}_- \oplus \mathfrak{H}_-$. First let us note that the spaces $B_\pm(\tilde{X})$ are closed in $\mathfrak{H}_\pm \oplus \mathfrak{H}_\pm$. Otherwise, let $\tilde{\tilde{X}}$ denote the restriction of X^* to the domain $\mathcal{D}(\tilde{\tilde{X}}) = \{\mathfrak{f} \in \mathcal{D}(X^*) : B_\pm(\mathfrak{f}) \in \overline{B_\pm(\tilde{X})}\}$, where the bar means the closure in the Hilbert space $\mathfrak{H}_\pm \oplus \mathfrak{H}_\pm$. The symmetry of an operator Y such that $X \subseteq Y \subseteq X^*$ is equivalent to the validity of Eq. (10) for all $\mathfrak{f},\mathfrak{g} \in \mathcal{D}(Y)$. Hence $\tilde{\tilde{X}}$ is symmetric, because $\tilde{X}$ is so. Since a self-adjoint operator has no proper symmetric extension, we conclude that $\tilde{X} = \tilde{\tilde{X}}$ which means that $B_+(\tilde{X})$ and $B_-(\tilde{X})$ are closed. Next let us suppose that $(\xi,\zeta) \perp B_+(\tilde{X})$ in $\mathfrak{H}_+ \oplus \mathfrak{H}_+$. We then choose a vector $\mathfrak{g} \in \mathcal{D}(\tilde{X})$ such that $\xi = \mathfrak{g}_e^+ + \mathfrak{g}_o^+$, $\zeta = \mathfrak{g}_e^+ - \mathfrak{g}_o^+$ and $\mathfrak{g}_e^- = \mathfrak{g}_0^- = 0$. Then the right-hand side of (9) vanishes for all $\mathfrak{f} \in \mathcal{D}(\tilde{X})$, so that $\langle \tilde{X}\mathfrak{f}, \mathfrak{g} \rangle = \langle X^*\mathfrak{f}, \mathfrak{g} \rangle = \langle \mathfrak{f}, X^*\mathfrak{g} \rangle$ for all $\mathfrak{f} \in \mathcal{D}(\tilde{X})$ by (9). Consequently, $\mathfrak{g} \in \mathcal{D}(\tilde{X}^*)$. Since $\tilde{X}$ is self-adjoint, $\mathfrak{g}$ must be in $\mathcal{D}(\tilde{X})$. Because $(\xi,\zeta) \perp B_+(\tilde{X})$, this implies that $\xi = \zeta = 0$. This proves that $B_+(\tilde{X}) = \mathfrak{H}_+ \oplus \mathfrak{H}_+$. Similarly $B_-(\tilde{X}) = \mathfrak{H}_- \oplus \mathfrak{H}_-$.

Since $B_\pm(\tilde{X}) = \mathfrak{H}_\pm \oplus \mathfrak{H}_\pm$ as just shown, it follows from (13) that there are unitary operators V and W of $\mathfrak{H}_-$ onto $\mathfrak{H}_+$ such that $\mathfrak{f}_e^+ + \mathfrak{f}_o^+ = V(\mathfrak{f}_e^- + \mathfrak{f}_o^-)$ and $\mathfrak{f}_e^+ - \mathfrak{f}_o^+ = W(\mathfrak{f}_e^- - \mathfrak{f}_o^-)$ for all $\mathfrak{f} \in \mathcal{D}(\tilde{X})$. That is, $\mathcal{D}(\tilde{X}) \subseteq \mathcal{D}(X_{V,W})$. Since $\tilde{X}$ and $X_{V,W}$ are self-adjoint, we conclude that $\tilde{X} = X_{V,W}$. ■

IV. IRREDUCIBILITY AND UNITARY EQUIVALENCE

(1) The next two propositions decide when a triple of operators $\{P, X_{V,W}, U\}$ defined in the preceding section is irreducible and when two such triples are unitarily equivalent. Here we shall say that the triple $\{P, X_{V,W}, U\}$ on $\mathcal{H}$ is *irreducible* if any bounded operator A on $\mathcal{H}$ satisfying

$$PA \supseteq AP, \quad X_{V,W}A \supseteq AX_{V,W}, \quad \text{and} \quad AU = UA \tag{14}$$

is a scalar multiple of the identity operator on $\mathcal{H}$.

Recall that the operator triple $\{P, X_{V,W}, U\}$ depends on the two families $\{\mu_1^{j,\pm}; j \in I_\pm\}$ of measures on the interval $[q,1)$ and on the two unitary operators $V,W : \mathfrak{H}_- \to \mathfrak{H}_+$. In order to formulate the corresponding conditions it is convenient to work with the Hilbert spaces $\mathfrak{K}_\pm = \oplus_{j \in I_\pm} L^2([q,1), \mu_1^{j,\pm})$ rather than with $\mathfrak{H}_\pm = \oplus_{j \in I_\pm} L^2([q,1), t^{-1}\mu_1^{j,\pm})$. Further, let $P_\pm$ denote the self-adjoint operator on $\mathfrak{K}_\pm$ which acts componentwise as the multiplication by the variable t. Clearly, V and W are bounded linear operators of $\mathfrak{K}_-$ to $\mathfrak{K}_+$ such that

$$V' := P_+^{1/2} V P_-^{-1/2} \quad \text{and} \quad W' := P_+^{1/2} W P_-^{-1/2} \tag{15}$$

are unitary.

Proposition 2: The triple $\{P, X_{V,W}, U\}$ as defined above is irreducible if and only if any bounded self-adjoint operators A_+ on $\mathfrak{K}_+$ and A_- on $\mathfrak{K}_-$ satisfying

$$A_+P_+ = P_+A_+, \quad A_-P_- = P_-A_-, \quad A_+V' = V'A_-, \quad A_+W' = W'A_- \tag{16}$$

or equivalently

$$A_+P_+=P_+A_+\,,\ \ A_-P_-=P_-A_-\,,\ \ A_+V=VA_-\,,\ \ A_+W=WA_- \tag{17}$$

are scalar multiples of the identity.

Proof: We only show that the above condition implies the irreducibility of the triple. The proof of the converse implication is easier and will be omitted. Suppose that A is a bounded operator on $\mathcal{H}$ satisfying (14). Since the set of such A is invariant under the involution, we can assume that A is self-adjoint. Let $E(\cdot)$ denote the spectral projections of P. Since $PA\supseteq AP$, the subspace $\mathfrak{K}_+=E([q,1))\mathcal{H}$ of $\mathcal{H}$ reduces A and the restriction A_+ of A to $\mathfrak{K}_+$ commutes with the restriction P_+ of P to $\mathfrak{K}_+$. Similarly, the restrictions $\tilde{A}_-$ of A and $\tilde{P}_-$ of P to the reducing subspace $E((-1,q])\mathcal{H}$ commute. Changing the variable from t to $-t$, the Hilbert space $E((-1,q])\mathcal{H}$ and the operator $\tilde{P}_-$ become $\mathfrak{K}_-$ and P_-, respectively, and the operator $\tilde{A}_-$ goes into an operator, say A_-, on $\mathfrak{K}_-$. Thus, $A_-P_-=P_-A_-$. From the assumptions $AU=UA$ and $X_{V,W}A\subseteq AX_{V,W}$ it follows easily that $(A\mathfrak{f})_e^\pm=A_\pm\mathfrak{f}_e^\pm$ and $(A\mathfrak{f})_o^\pm=A_\pm\mathfrak{f}_o^\pm$ for $\mathfrak{f}\in\mathcal{D}(X_{V,W})$. Since $A\mathfrak{f}\in\mathcal{D}(X_{V,W})$ has to satisfy the relation (8), we obtain $A_+V=VA_-$ and $A_+W=WA_-$. Therefore, by the above condition, $A_\pm=\lambda_\pm I$ for some $\lambda_\pm\in\mathbb{C}$. Since $A_+V=VA_-$ and $AU=UA$, it follows that $\lambda_+=\lambda_-$ and $A=\lambda_+\cdot I$ on $\mathcal{H}$. ■

Using similar operator-theoretic arguments it is not difficult to prove the following.

Proposition 3: Two triples $\{P,X_{V,W},U\}$ *and* $\{\tilde{P},X_{\tilde{V},\tilde{W}},\tilde{U}\}$ *are unitarily equivalent if and only if there are unitary operators* A_+ *of* $\mathfrak{K}_+$ *to* $\tilde{\mathfrak{K}}_+$ *and* A_- *of* $\mathfrak{K}_-$ *to* $\tilde{\mathfrak{K}}_-$ *such that*

$$A_+P_+=\tilde{P}_+A_+\,,\ \ A_-P_-=\tilde{P}_-A_-\,,\ \ A_+V=\tilde{V}A_-\quad\text{and}\quad A_+W=\tilde{W}A_-\,, \tag{18}$$

where the tilde refers to the corresponding operators and spaces for the triple $\{\tilde{P},X_{\tilde{V},\tilde{W}},\tilde{U}\}$.

(2) We shall illustrate the preceding by describing a few examples of irreducible representations. We begin with the simplest possible case.

Example 2: Suppose that the Hilbert spaces $\mathfrak{K}_+$ and $\mathfrak{K}_-$ are one-dimensional. Then the families of measure $\{\mu_i^{j,+};\in I_+\}$ and $\{\mu_1^{j,-};j\in I_-\}$ consist only of single Dirac measures δ_a and δ_b, respectively, where $a,b\in[q,1)$. Then the triples $\{P,X_{V,W},U\}$ are parametrized by complex numbers $V=V'=e^{i\varphi}$ and $W=W'=e^{i\psi},\varphi,\psi\in\mathbb{R}$. The self-adjoint extension $X_{V,W}$ is then characterized by the boundary condition (8), that is,

$$\mathfrak{f}_e^+ + \mathfrak{f}_o^+ = e^{i\varphi}(\mathfrak{f}_e^- + \mathfrak{f}_o^-),\quad \mathfrak{f}_e^+ - \mathfrak{f}_o^+ = e^{i\psi}(\mathfrak{f}_e^- - \mathfrak{f}_o^-).$$

Each such triple is irreducible because the condition in Proposition 2 is trivially fulfilled. Two triples with different numbers VW^{-1} are not unitarily equivalent. The case where $e^{i\varphi}=e^{i\psi}=1$ and $a=b$ has been treated in detail in Ref. 3.

Example 3: Let P_+ be a self-adjoint operator and Z a unitary operator on a Hilbert space $\mathfrak{K}_+$ such that the commutant $\{P_+,Z\}'$ is equal to $\mathbb{C}\cdot I$. Such operators exist on any separable Hilbert space.[10] Upon scaling we can assume that the spectrum of P_+ is contained in $[q,1)$. By the spectral representation theorem (Ref. 11, chap. X, 5), we can represent P_+ up to unitary equivalence as the multiplication operator by the independent variable t on some direct sum Hilbert space $\mathfrak{K}_+=\oplus_{j\in I_+}L^2([q,1);\mu_1^{j,+})$. Let $\{\mu_1^{j,-};j\in I\}$ be an arbitrary family of measures on $[q,1)$ such that $\dim\mathfrak{K}_+=\dim\mathfrak{K}_-$, where $\mathfrak{K}_-:=\oplus_{j\in I_-}L^2([q,1);\mu_1^{j,-})$. Let W' be a unitary operator from $\mathfrak{K}_-$ to $\mathfrak{K}_+$. We set $V':=ZW'$ and define V and W by (15). Then the triple $\{P,X_{V,W},U\}$ is irreducible.

Indeed, if A_+ and A_- be bounded self-adjoint operators satisfying (17), then we have $A_+Z=A_+V'W'^*=V'A_-W'^*=V'W'^*A_+=ZA_+$ and $A_+P_+=P_+A_+$, so that $A_+=\lambda\cdot I$ for some $\lambda\in\mathbb{C}$ and hence $A_-=V'^*A_+V'=\lambda\cdot I$. By Proposition 2, the triple is irreducible. ■

Example 4: For this example we assume that there exist numbers $a,b\in[q,1)$ such that $\mu_1^{j,+}=\delta_a$ and $\mu_1^{k,-}=\delta_b$ for all $j\in I_+$ and $k\in I_-$. We shall show that in this case an irreducible triple $\{P,X_{V,W},U\}$ can be only obtained if both index sets I_+ and I_- are singletons or equivalently if $\dim\mathfrak{K}_+=\dim\mathfrak{K}_-=1$. Indeed, otherwise we take a self-adjoint operator A_+ on $\mathfrak{K}_+$ such that

$A_+V'W'^* = V'W'^*A_+$ and $A_+ \notin \mathbb{C}\cdot I$ and set $A_- := V'^*A_+V'$. Then the conditions (16) are fulfilled, hence the triple is not irreducible. ■

Example 5: If the spectra of the operators P_+ on $\mathfrak{K}_+$ and P_- on $\mathfrak{K}_-$ are singletons, then we have seen in Example 4 that irreducible triples exist only in the trivial case where I_+ and I_- are singletons. We now show that this is no longer true if both spectra consist of two points. To be more precise, we shall consider the following situation: The index sets $I_\pm$ are disjoint unions of two countable infinite sets $I_\pm^1$ and $I_\pm^2$ and there are numbers $a_1, a_2, b_1, b_2 \in [q,1)$, $a_1 \neq a_2$, such that $\mu_1^{j,+} = \delta_{a_1}$ for $j \in I_+^1$, $\mu_1^{j,+} = \delta_{a_2}$ for $j \in I_+^2$, $\mu_1^{j,-} = \delta_{b_1}$ for $j \in I_-^1$, and $\mu_1^{j,-} = \delta_{b_2}$ for $j \in I_-^2$. By identifying $I_\pm^j$ with the natural numbers the Hilbert spaces $\mathfrak{K}_+$ and $\mathfrak{K}_-$ become the direct sum $l_2(\mathbb{N}) \oplus l_2(\mathbb{N})$ of two l_2-spaces. We choose a bounded operator T on $l_2(\mathbb{N})$ such that $\{T,T^*\}' = \mathbb{C}\cdot I$ and $I \leqslant 3T^*T \leqslant 2\cdot I$. It is well known (see Ref. 12, Anhang, §4) that the operator matrix

$$Z = \begin{pmatrix} T & \sqrt{I-TT^*} \\ -\sqrt{I-T^*T} & T^* \end{pmatrix}$$

defines a unitary operator Z on $\mathfrak{K}_+ = \mathfrak{K}_- = l_2(\mathbb{N}) \oplus l_2(\mathbb{N})$. Let W' be an arbitrary unitary operator on $\mathfrak{K}_+ = \mathfrak{K}_-$ and set $V' := ZW'$. Then the triple $(P, X_{V,W}, U)$ is irreducible.

Indeed, let A_+ and A_- be self-adjoint bounded operators on $\mathfrak{K}_+ = \mathfrak{K}_-$ satisfying (17). Since $a_1 \neq a_2$, the relation $A_+P_+ = P_+A_+$ implies that A_+ is given by a diagonal operator matrix

$$A_+ = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix}.$$

From (17) we get $A_+Z = ZA_+$. Comparing the matrix entries of the first line yields $BT = TB$ and $B\sqrt{I-T^*T} = \sqrt{I-T^*T}C$. Since $B = B^*$, we have $BT^* = T^*B$. Therefore, B commutes with T and T^* and so with $\sqrt{I-T^*T}$ which in turn gives $\sqrt{I-T^*T}B = \sqrt{I-T^*T}C$. Because $\sqrt{I-T^*T}$ is invertible, we get $B = C$. Since $B \in \{T,T^*\}'$, we obtain $B = C = \lambda\cdot I$ for some $\lambda \in \mathbb{C}$. Thus, $A_+ = \lambda\cdot I$ and $A_- = V'^*A_+V' = \lambda\cdot I$, so that the triple is irreducible by Proposition 2. ■

V. A CHARACTERIZATION OF THE OPERATOR TRIPLES

Let $\{P, X_{V,W}, U\}$ be an operator triple as in Sec. II and let $\mathcal{D}_1$ be the set of all vectors $\mathfrak{f} = \mathfrak{f}_X + \mathfrak{f}^e + \mathfrak{f}^o \in \mathcal{D}(X_{V,W})$ with $\mathfrak{f}_X \in \mathcal{D}_o$, where $\mathcal{D}_o$ is as defined in Sec. II. Then $\mathcal{D}_1$ is a dense linear subspace of the Hilbert space $\mathcal{H}$ such that $\mathcal{D}_1$ is invariant under the operators P, $X_{V,W}$, and U and the restrictions of P and $X_{V,W}$ to $\mathcal{D}_1$ are essentially self-adjoint. Further, the three operators P, $X_{V,W}$, and U applied to vectors $\mathfrak{f} \in \mathcal{D}_1$ satisfy the relations (1) and (2). From the construction it is clear that the range $E([q,1))\mathcal{H} (\cong \mathfrak{K}_+)$ of the spectral projection $E([q,1))$ of the operator P is contained in $\mathcal{D}_1$. Our next proposition says that the operator triples $\{P, X_{V,W}, U\}$ can be characterized by some of the properties just mentioned.

Proposition 4: Let $\{P', X', U'\}$ be a triple of two self-adjoint operators P' and X' and a unitary operator U' on a Hilbert space $\mathcal{H}'$. Let $E(\cdot)$ denote the spectral measure of P'. Suppose that there exists a linear subspace $\mathcal{D}_1 \subseteq \mathcal{D}(P'X') \cap \mathcal{D}(X'P')$ of $\mathcal{H}$ such that

(i) $E([q,1))\mathcal{H} \subseteq \mathcal{D}_1$ *and* $E((-1,-q])\mathcal{H} \subseteq \mathcal{D}_1$.
(ii) *The operators P', X', U' satisfy the relations (1) and (2) for vectors in $\mathcal{D}_1$.*
(iii) *The restrictions $P' \lceil \mathcal{D}_1$ and $X' \lceil \mathcal{D}_1$ of P' and X' to $\mathcal{D}_1$ are essentially self-adjoint. Then $\{P', X', U'\}$ is unitarily equivalent to an operator triple $\{P, X_{V,W}, U\}$ defined in Sec. II.*

Sketch of proof: The restriction P_1' of P' to the invariant subspace $\mathcal{H}_1 := E([q,1))\tilde{\mathcal{H}}$ is obviously a bounded self-adjoint operator on the Hilbert space $\mathcal{H}_1$ with spectrum contained in the interval $[q,1]$. By the spectral representation theorem[11], there is a family $\{\mu_1^{j,+}; j \in I_+\}$ of finite positive Borel measures on $[q, 1]$ and a unitary isomorphism of $\mathcal{H}_1$ on $\mathfrak{K}_+ := \oplus_j L^2([q, 1], \mu_1^{j,+})$

such that P_1' is unitarily equivalent to the operator P_1 on $\mathfrak{K}_+$ which acts componentwise as the multiplication by the variable t. Since 1 is not an eigenvalue of P_1' by construction, we have $\mu_1^{j,+}(\{1\})=0$ for all $j\in I_+$. For simplicity let us identify $\mathcal{H}_1$ with $\mathfrak{K}_+$ and P_1' with P_1.

Next we show that $\ker P'=\{0\}$. Let $\mathfrak{f}\in\ker P'$. Since $P'\lceil\mathcal{D}_1$ is essentially self-adjoint by (iii), there exists a sequence $\{\mathfrak{f}_n\}$ of vectors $\mathfrak{f}_n\in\mathcal{D}_1$ such that $\mathfrak{f}_n\to\mathfrak{f}$ and $P'\mathfrak{f}_n\to P'\mathfrak{f}=0$ in $\mathcal{H}$. Since $X'P'\mathfrak{f}_n=\mathrm{i}(q^{1/2}U'^*+q^{1/2}U')\mathfrak{f}_n$ by (ii) and the operators U' and U'^* are bounded, we obtain $(q^{-1/2}U'^*+q^{1/2}U')\mathfrak{f}=0$ in the limit. This in turn yields that $q\|\mathfrak{f}\|=\|\mathfrak{f}\|$ and so $\mathfrak{f}=0$.

By (ii), we have $U'P'\mathfrak{f}=qP'U'\mathfrak{f}$ for all $\mathfrak{f}\in\mathcal{D}_1$. Since $p'\lceil\mathcal{D}_1$ is essentially self-adjoint, this remains valid for $\mathfrak{f}\in\mathcal{D}(P')$, so that $P'\subseteq qU'^*P'U'$. Since P' is self-adjoint, we conclude that $P'=qU'^*P'U'$. Hence we have $U'^nE(\mathfrak{N})=E(q^{-n}\mathfrak{N})$ for any Borel subset $\mathfrak{N}$ of $\mathbb{R}$ and arbitrary $n\in\mathbb{Z}$. Let $\mu^{j,+}$ be the extension of the measure $\mu_1^{j,+}$ to $\mathbb{R}_+$ as in Sec. II (1). From the preceding considerations it follows that $E(\mathbb{R}_+)\mathcal{H}=\oplus_j L^2(\mathbb{R}_+,\mu^{j,+})\equiv\mathcal{H}_+$ and that U' acts in each component by formula (i) in Sec. II (1). Proceeding in a similar manner, we obtain a family $\{\mu_1^{j,-};j\in I_-\}$ of measures on $[q,1]$ such that $\mu_1^{j,-}(\{1\})=0$ for $j\in I_-,E(\mathbb{R}_-)\mathcal{H}=\oplus_j L^2(\mathbb{R}_-,\mu^{j,-})\equiv\mathcal{H}_-$ in the notation of Sec. II and U' acts componentwise as given by formula (i) in Sec. II (1). Since $E(\{0\})\mathcal{H}=\ker P'=\{0\}$ as proved in the preceding paragraph, we conclude that $\mathcal{H}=\mathcal{H}_+\oplus\mathcal{H}_-$.

From the construction it is clear that P' and U' are the operators P and U, respectively, as in Sec. II. Let us finally turn to the operator X'. Recall that we have $X'P'\mathfrak{f}=\mathrm{i}(q^{-1/2}U'^*+q^{1/2}U')\mathfrak{f}$ for $\mathfrak{f}\in\mathcal{D}_1$. By arguing as in the paragraph before last, this relation remains valid for all $\mathfrak{f}\in\mathcal{D}(P')$. If f denotes a component of the vector $\mathfrak{f}$, then the preceding equation yields that $g:=tf\in\mathcal{H}$, $t^{-1}g=f\in\mathcal{H}$ and $(X'g)(t)=\mathrm{i}(q^{-1}f(q^{-1}t)-qf(qt))=\mathrm{i}t^{-1}(g(q^{-1}t)-g(qt))=(Xg)\times(t)$. Hence $X'\mathfrak{f}=X\mathfrak{f}$ for all $\mathfrak{f}\in\mathcal{D}(P')$. Since $X'\lceil\mathcal{D}_1$ is essentially self-adjoint, the relation $U'X'\mathfrak{f}=q^{-1}X'U'\mathfrak{f}$ for $\mathfrak{f}\in\mathcal{D}_1$ by (ii) extends to vectors $\mathfrak{f}\in\mathcal{D}(X')$, so that $U'X'U'^*=q^{-1}X'$. Thus, X' is a self-adjoint extension of the operator X such that $U\mathcal{D}(X')=\mathcal{D}(X')$. By Proposition 1, X' is of the form $X_{V,W}$. ■

VI. ∗-REPRESENTATIONS OF THE *q*-HEISENBERG ALGEBRA

(1) We have considered so far only operator triples and operator relations rather than representations of the algebra $\mathcal{A}(q)$. However, any operator triple $\{P,X_{V,W},U\}$ gives rise to a self-adjoint representation of the ∗-algebra as follows. Indeed, let $\mathcal{D}_1$ be the domain defined at the beginning of Sec. IV. For vectors in $\mathcal{D}_1$ the operators $P,X_{V,W},U$ satisfy the defining relations (1) and (2) of the algebra $\mathcal{A}(q)$. Hence there exists a unique ∗-representation π_1 of the ∗-algebra $\mathcal{A}(q)$ on the domain $\mathcal{D}_1$ such that

$$\pi_1(p)=P\lceil\mathcal{D}_1,\pi_1(x)=X_{V,W}\lceil\mathcal{D}_1,\pi_1(u)=U\lceil\mathcal{D}_1.$$

(For the notions on unbounded ∗-representations used in what follows we refer to Ref. 13. Recall that the symbol $T\lceil\mathcal{D}_1$ means the restriction of T to $\mathcal{D}_1$.)

The ∗-representation π_1 is not yet self-adjoint (see Ref. 13, Definition 8.1.10), because, roughly speaking, $\mathcal{D}_1$ is not the largest possible domain. However, since the operators $\pi_1(p)$ and $\pi_1(x)$ are essentially self-adjoint, it follows at once from Proposition 8.1.12(v) in Ref. 13 that the adjoint representation $\pi:=(\pi_1)^*$ is self-adjoint. It is not difficult to verify that the domain $\mathcal{D}$ of the ∗-representation π is just the intersection of domains of all possible products of the operators $P,X_{V,W},U$ (see Ref. 13, Proposition 8.1.17). From these facts it follows that the operator triple $\{P,X_{V,W},U\}$ is irreducible if and only if the ∗-representation π is so and that two triples are unitarily equivalent if and only if the corresponding ∗-representations are so. That is, Propositions 2 and 3 provide also the conditions for the irreducibility and the unitary equivalence of these ∗-representations of the ∗-algebra $\mathcal{A}(q)$.

(2) Finally, we briefly discuss how operator representations of the q-deformed Heisenberg algebra $\mathcal{A}(q)$ can be constructed by means of the Schrödinger representation $\mathcal{P}:=-\mathrm{i}(d/dt)$ and $\mathcal{Q}:=t$ of the ‘‘ordinary’’ momentum and position operators.

Let us write $q=e^{-\alpha}$ with $\alpha\in\mathbb{R}$. We define three operators U, P, and X on the Hilbert space $\mathcal{H}=L^2(\mathbb{R})$:

$$U=e^{\mathrm{i}\,Q},\quad P=e^{\alpha\mathcal{P}},\quad X=\mathrm{i}(q^{-1/2}e^{-\mathrm{i}\,Q}-q^{1/2}e^{\mathrm{i}\,Q})e^{-\alpha\mathcal{P}}. \tag{19}$$

The vector space $\mathcal{D}:=\mathrm{Lin}\{e^{\gamma t-t^2};\gamma\in\mathbb{C}\}$ is a dense linear subspace of $\mathcal{H}$. Since the operator $e^{\beta\mathcal{P}},\beta\in\mathbb{R}$, acts as $(e^{\beta\mathcal{P}}f)(t)=f(t-\beta\,\mathrm{i})$ on functions $f\in\mathbb{C}$ (see, for instance, Ref. 14 for a rigorous proof), the operators U, P, and X satisfy the relations (1) and (2)$'$ and the restrictions of these operators to the invariant dense domain $\mathcal{D}$ define a $*$-representation of the $*$-algebra $\mathcal{A}(q)$. This operator representation (19) appears already somewhat hidden in Ref. 1. Indeed, if we change the variable t to e^t, then the operator triple $\{U\oplus U,(-P)\oplus P,(-X)\oplus X\}$ on the direct sum Hilbert space $\mathcal{H}\oplus\mathcal{H}$ is easily seen to be unitarily equivalent to the triple in formula (2.2) in Ref. 1.

The operator representation (19) is irreducible on $\mathcal{H}$. Obviously, U is unitary and P is self-adjoint. However, an essential disadvantage of the representation (19) is that the operator X is only symmetric, but not essentially self-adjoint. The latter can be shown by the argument used in the proof of Proposition A.2 in Ref. 14. The reason for this failure is the fact the holomorphic function $h(z)=q^{-1/2}e^{\mathrm{i}\,z}-q^{1/2}e^{-\mathrm{i}\,z}$ admits the zero $z_o=i(\alpha/2)$ in the strip $\{z\in\mathbb{C}:\ 0<\mathrm{Im}\,z<\alpha\}$.

[1] M. Fichtmüller, A. Lorek, and J. Wess, ''q-Deformed phase space and its lattice structure,'' Z. Phys. C **71**, 533–538 (1996).
[2] A. Hebecker, S. Schreckenberg, J. Schwenk, W. Weich, and J. Wess, ''Representations of a q-deformed Heisenberg algebra,'' Z. Phys. C **64**, 335–359 (1994).
[3] A. Lorek, W. Weich, and J. Wess, ''Non-commutative Euclidean and Minkowski structures,'' preprint, MPI-PhI/96-124, Munich (1996).
[4] J. Schwenk and J. Wess, ''A q-deformed quantum mechanical toy model,'' Phys. Lett. B **291**, 273–277 (1992).
[5] W. Weich, ''The Hilbert space representations for $SO_q(3)$-symmetric quantum mechanics,'' preprint, Munich, hep-th/9404029 (1994).
[6] J. Wess and B. Zumino, ''Covariant differential calculus on the quantum hyperplane,'' Nucl. Phys. B, Proc. Suppl. **18B**, 302–312 (1991).
[7] W. Pusz and S. L. Woronowicz, ''Twisted second quantization,'' Rep. Math. Phys. **27**, 231–257 (1989).
[8] T. H. Koornwinder and R. F. Swartouw, ''On q-analogues of the Fourier and Hankel transforms,'' Trans. Am. Math. Soc. **333**, 445–461 (1992).
[9] H. T. Koelink and R. F. Swartouw, ''On the zeros of the Hahn-Exton q-Bessel function and associated q-Lommel polynomials,'' J. Math. Anal. Appl. **186**, 690–770 (1994).
[10] D. M. Topping, *Lectures on von Neumann Algebras* (Van Nostrand, New York, 1971).
[11] N. Dunford and J. Schwartz, *Linear Operators, Part II* (Interscience, New York, 1963).
[12] F. Riesz and B. Sz.-Nagy, *Vorlesungen überFunktionalanalysis* (DVW, Berlin, 1956).
[13] K. Schmüdgen, *Unbounded Operator Algebras and Representation Theory* (Birkhäuser, Basel, 1990).
[14] K. Schmüdgen, ''Integrable operator representations of $\mathbb{R}_q^2$, $X_{q,\gamma}$ and $SL_q(2,\mathbb{R})$,'' Commun. Math. Phys. **159**, 217–237 (1992).

Extended gauge theories starting from the matter field Lagrangian

R. Amorim[a)] and J. Barcelos-Neto[b)]
Instituto de Física, Universidade Federal do Rio de Janeiro, RJ 21945-970, Caixa Postal 68528, Brasil

(Received 26 January 1999; accepted for publication 19 June 1999)

We obtain an extended gauge theory by imposing that the matter field Lagrangian is invariant under a local gauge transformation that also contains a vector parameter besides the usual scalar one.
[S0022-2488(99)03810-4]

I. INTRODUCTION

It is well-known that usual gauge theories, described by vector fields, may arise from matter field Lagrangians by imposing their invariance under local phase transformation. In this procedure, the vector gauge field naturally appears as that one that compensates an extra term given by the derivative of the phase function.

Nowadays, there has been a great deal of interest in tensor gauge fields,[1] but these are never introduced as fields that compensate gauge transformations of matter fields. We mention that they are antisymmetric and their gauge transformations are taken as a direct extension of the vector case. Up to some multiplicative factor, these gauge transformations are given by (we concentrate on the Abelian case only)

$$\delta B_{\mu\nu}=\partial_\mu\xi_\nu-\partial_\nu\xi_\mu . \tag{1.1}$$

As we observe, if the vector gauge parameter ξ_μ is taken as the gradient of some scalar quantity, we obtain that $\delta B_{\mu\nu}=0$. This property gives us an interesting structure of constraints for gauge theories involving tensor fields. It means that their first-class constraints are not all independent (reducibility condition) and their quantization deserve some additional care comparing with the usual case of rank one gauge theories.[1,2] We also mention that antisymmetric tensor fields appear as massless solutions of modern string theories among other fields such as photons, gravitons, etc.[3]

When one takes into account the coupling between matter and tensor gauge fields, there are some discordances and inconsistencies in literature. For example, in the study of chiral anomaly, tensor gauge fields can be coupled to a tensor current $\bar\psi\Sigma^{\mu\nu}\psi$ (up to some coupling factor).[4] {We use the following convention and notation throughout this paper: $\{\gamma^\mu,\gamma^\nu\}=2\eta^{\mu\nu}$, $\eta^{\mu\nu}=\text{diag.}(+,-,-,-)$, $\gamma^0\gamma^{\mu\dagger}\gamma^0=\gamma^\mu$, $\epsilon^{0123}=1$, $\gamma_5=i\gamma_0\gamma_1\gamma_2\gamma_3$ and $\Sigma^{\mu\nu}=(i/4)[\gamma^\mu,\gamma^\nu]$.} In this situation, the gauge transformation for the matter field is considered to be the usual one, i.e., by means of a local phase exponential factor, what is compensated by the presence of a vector gauge field. The gauge transformation for the tensor gauge field, however, is not the same one as (1.1). In fact, it is taken as zero (Abelian case).

Another kind of coupling between matter and tensor gauge fields, compatible with the gauge transformation (1.1), has already been considered.[5] It has a topological nature, but the coupling is with a vector current, i.e., the interaction Lagrangian density is given by $(i/2m)\epsilon_{\mu\nu\lambda\rho}\partial^\nu B^{\lambda\rho}\bar\psi\gamma^\mu\psi$, where m is a mass parameter.[5] It is opportune to mention that topological interactions are also used

---
[a)]Electronic mail: amorim@if.ufrj.br
[b)]Electronic mail: barcelos@if.ufrj.br

to couple vector and tensor gauge fields[6] with the purpose to obtain an alternative mechanism of mass generation for gauge fields (vector or tensor), also compatible with the Salam–Weinberg theory when applied to the non-Abelian case.[7]

The purpose of the present paper is to show that tensor gauge fields can also be introduced as fields that may compensate extended gauge transformations for the matter field. In this case, we shall see that the transformation (1.1) is just part of a more general gauge transformation satisfying a non-Abelian algebra even though all fields are here taken as Abelian. Concerning the kinetical gauge sector, the problem appears to be more subtle, because we do not know how to write the gauge curvature tensor when fields of different ranks are put together. We also address our attention to this part and try to circumvent this problem by adopting an iterative process of looking for an invariant Lagrangian step by step. Our conclusion is that the Lagrangian we are able to obtain in this way is a trivial one, expressed in terms of collective fields.[8]

Our paper is organized as follows: In Sec. II we discuss the extended gauge transformation we are going to use for the matter field. We shall see that it is necessary to add a Lorentz-type gauge transformation in order to close the gauge algebra (which is also non-Abelian). Section III is devoted to the obtainment of the matter field Lagrangian that is invariant under this extended gauge transformation. In Sec. IV we discuss the problem of obtaining the Lagrangian for the gauge sector. We left Sec. V for some concluding remarks and introduce an Appendix where some details of the calculations are presented.

II. EXTENSION OF THE GAUGE TRANSFORMATION

Let us try an extension of the usual gauge transformation for the matter field by introducing a vector parameter. A natural way of doing this is to take

$$\delta\psi = i(\alpha(x) + \xi^\mu(x)\gamma_\mu)\psi. \tag{2.1}$$

Contrarily to the gauge transformations involving γ_5, the mass term $m\bar{\psi}\psi$ is invariant under the transformation (2.1),

$$\delta(\bar{\psi}\psi) = (i\xi_\mu\gamma^\mu\psi)^\dagger\gamma^0\psi + \bar{\psi}(i\xi_\mu\gamma^\mu\psi) = i\xi_\mu(-\psi^\dagger\gamma^{\mu\dagger}\gamma^0\psi + \bar{\psi}\gamma^\mu\psi) = i\xi_\mu(-\bar{\psi}\gamma^\mu\psi + \bar{\psi}\gamma^\mu\psi) = 0. \tag{2.2}$$

However, the kinetic term is not. Before trying to see what we have to do in order to get an invariant kinetic term, there is a problem that ought to be solved first. The transformation (2.1) does not close in an algebra. In fact,

$$[\delta_1, \delta_2]\psi = 4i\xi_1^\mu\xi_2^\nu\Sigma_{\mu\nu}\psi. \tag{2.3}$$

We observe that to close the algebra, what is an essential condition to have the theory consistently quantized, the transformation of the matter field must also include a Lorentz-type term,

$$\delta\psi = i(\alpha(x) + \xi^\mu(x)\gamma_\mu - \tfrac{1}{2}\omega(x)^{\mu\nu}\Sigma_{\mu\nu})\psi, \tag{2.4}$$

where the parameter $\omega^{\mu\nu}$ was generically taken as local. Now the transformation (2.4) closes in an algebra, whose general form reads

$$[\delta_1, \delta_2]\psi = i(\alpha_3 + \xi_3^\mu\gamma_\mu + \omega_3^{\mu\nu}\Sigma_{\mu\nu})\psi, \tag{2.5}$$

where

$$\alpha_3=0,$$

$$\xi_3^\mu=\xi_{1\nu}\omega_2^{\nu\mu}-\xi_{2\nu}\omega_1^{\nu\mu}, \tag{2.6}$$

$$\omega_3^{\mu\nu}=2(\xi_1^\mu\xi_2^\nu-\xi_2^\mu\xi_1^\nu)-\tfrac{1}{2}(\omega_1^{\mu\lambda}\omega_{2\lambda}{}^{\nu}-\omega_2^{\mu\lambda}\omega_{1\lambda}{}^{\nu}).$$

We observe that the tensor parameter $\omega_{\mu\nu}$ has actually to be local. So, to have a closed algebra, it is also necessary to include a local Lorentz-type gauge transformation. The presence of this term means that gravitation might be naturally embodied in the theory.

III. INVARIANT MATTER FIELD LAGRANGIAN

It is just a question of algebraic work to show that the transformation of the kinetic matter field Lagrangian $\mathcal{L}_0=i\bar{\psi}\partial\!\!\!/\psi$ under (2.4) is given by

$$\delta\mathcal{L}_0=i\,\delta\bar{\psi}\partial\!\!\!/\psi+i\bar{\psi}\partial\!\!\!/\delta\psi=-\bar{\psi}\gamma^\mu\psi\partial_\mu\alpha-4i\xi_\mu\bar{\psi}\Sigma^{\mu\nu}\partial_\nu\psi+i\omega_{\mu\nu}\bar{\psi}\gamma^\nu\partial^\mu\psi-\bar{\psi}\gamma^\mu\gamma^\nu\psi\partial_\mu\xi_\nu$$
$$+\tfrac{1}{2}\bar{\psi}\gamma^\lambda\Sigma^{\mu\nu}\psi\partial_\lambda\omega_{\mu\nu}. \tag{3.1}$$

We see that to implement the gauge invariance by means of gauge fields is not a simple task for the tensor sector. In order to have a general view of the problem and avoid the presence of big equations, it is convenient to use a compact notation by redefining some terms. For example, for the general transformation (2.4), we simply take

$$\delta\psi=i\Gamma\psi, \tag{3.2}$$

where

$$\Gamma=\alpha(x)+\xi^\mu(x)\gamma_\mu-\tfrac{1}{2}\omega(x)^{\mu\nu}\Sigma_{\mu\nu}. \tag{3.3}$$

Instead of expression (3.1), we now just have

$$\delta\mathcal{L}_0=\bar{\psi}[\Gamma,\gamma^\mu]\partial_\mu\psi-\bar{\psi}\partial\!\!\!/\Gamma\psi. \tag{3.4}$$

If one introduces a general compensating field A that interacts with the fermionic one as

$$\mathcal{L}_1=\bar{\psi}A\psi, \tag{3.5}$$

we obtain

$$\delta(\mathcal{L}_0+\mathcal{L}_1)=\bar{\psi}[\Gamma,\gamma^\mu]\partial_\mu\psi+\bar{\psi}(\delta A-i[\Gamma,A]-\partial\!\!\!/\Gamma)\psi. \tag{3.6}$$

We note that if one takes $\delta A=i[\Gamma,A]+\partial\!\!\!/\Gamma$, the last term of (3.6) cancels, but the invariance is only attained when $[\Gamma,\gamma^\mu]=0$. This occurs of course for the particular case of the electromagnetic field where Γ is just the parameter $\alpha(x)$. Further, with a convenient change of coordinates, this equation is also solved for a constant $\omega^{\mu\nu}$ associated with global Poincaré transformations.

To obtain the invariance for the general case, let us consider that instead of $\mathcal{L}_0$ we have a more general kind of Lagrangian like

$$\mathcal{L}_2=i\bar{\psi}K^\mu\partial_\mu\psi, \tag{3.7}$$

where K^μ is some vectorial function of auxiliary fields in a suitable combination of gamma matrices (containing of course the particular case given by $\mathcal{L}_0$). For the resulting Lagrangian $\mathcal{L}_1+\mathcal{L}_2$, we have

$$\delta(\mathcal{L}_1+\mathcal{L}_2)=\bar{\psi}([\Gamma,K^\mu]+i\,\delta K^\mu)\partial_\mu\psi+\bar{\psi}(\delta A-i[\Gamma,A]-K^\mu\partial_\mu\Gamma)\psi. \tag{3.8}$$

The invariance is finally attained if

$$\delta K^{\mu}=i[\Gamma,K^{\mu}], \tag{3.9}$$

$$\delta A=i[\Gamma,A]+K^{\mu}\partial_{\mu}\Gamma. \tag{3.10}$$

One important point to be emphasized is that, for Γ given by (3.3), the transformations (3.9) and (3.10) close in an algebra, independently of the form of K^{μ} and A (see Appendix).

To obtain a solution for Eqs. (3.9) and (3.10), we have to use some identities expressing products of gamma matrices in terms of the sixteen independent terms I, γ^{μ}, $\Sigma^{\mu\nu}$, γ_5, and $\gamma_5\gamma^{\mu}$. Let us list them below,

$$\begin{aligned}
\gamma^{\mu}\gamma^{\nu}&=\eta^{\mu\nu}-2i\Sigma^{\mu\nu},\\
\gamma_5\gamma^{\mu}\gamma^{\nu}&=\eta^{\mu\nu}\gamma_5+\epsilon^{\mu\nu\lambda\rho}\Sigma_{\lambda\rho},\\
\gamma^{\mu}\gamma^{\nu}\gamma^{\lambda}&=\eta^{\mu\nu}\gamma^{\lambda}-\eta^{\mu\lambda}\gamma^{\nu}+\eta^{\nu\lambda}\gamma^{\mu}-i\epsilon^{\mu\nu\lambda\rho}\gamma_5\gamma_{\rho},
\end{aligned} \tag{3.11}$$

$$\Sigma^{\mu\nu}\Sigma^{\lambda\rho}=\frac{1}{4}(\eta^{\mu\lambda}\eta^{\nu\rho}-\eta^{\mu\rho}\eta^{\nu\lambda})+\frac{i}{4}\epsilon^{\mu\nu\lambda\rho}\gamma_5$$
$$-\frac{i}{2}(\eta^{\mu\lambda}\eta^{\nu\alpha}\eta^{\rho\beta}+\eta^{\nu\rho}\eta^{\mu\alpha}\eta^{\lambda\beta}-\eta^{\mu\rho}\eta^{\nu\alpha}\eta^{\lambda\beta}-\eta^{\nu\lambda}\eta^{\mu\alpha}\eta^{\rho\beta})\Sigma_{\alpha\beta}.$$

To solve Eq. (3.9), we conclude that K^{μ} should be expressed by

$$K^{\mu}=H^{\mu\nu}\gamma_{\nu}+I^{\mu\nu\lambda}\Sigma_{\nu\lambda}, \tag{3.12}$$

where the fields $H^{\mu\nu}$ and $I^{\mu\nu\lambda}$ play the role of a kind of Stuckelberg fields. It is opportune to say that if we had written a simpler expression for K^{μ} as, for example, $H\gamma^{\mu}+I_{\nu}\Sigma^{\nu\mu}$, it would not work because Eq. (3.9) would lead to an expression with different symmetries on left and right sides.

Using K^{μ} given by expression (3.12), as well as the expression of Γ given by (3.3), into (3.9), we obtain that the auxiliary fields $H^{\mu\nu}$ and $I^{\mu\nu\lambda}$ must transform as

$$\delta H^{\mu\nu}=2I^{\mu\nu\lambda}\xi_{\lambda}-\omega^{\mu\nu}-H^{\mu}{}_{\lambda}\omega^{\lambda\nu}, \tag{3.13}$$

$$\delta I^{\mu\nu\lambda}=2(\eta^{\mu\lambda}\xi^{\nu}-\eta^{\mu\nu}\xi^{\lambda})+2(H^{\mu\lambda}\xi^{\nu}-H^{\mu\nu}\xi^{\lambda})+I^{\mu\rho\nu}\omega_{\rho}{}^{\lambda}-I^{\mu\rho\lambda}\omega_{\rho}{}^{\nu}. \tag{3.14}$$

Concerning the solution of (3.10), we see that it is attained by using (3.3) and (3.12) and we may conclude that the general expression for A should be

$$A=S-A_{\mu}\gamma^{\mu}+\tfrac{1}{2}B_{\mu\nu}\Sigma^{\mu\nu}+P\gamma_5+V_{\mu}\gamma_5\gamma^{\mu}. \tag{3.15}$$

It is interesting to notice the presence of chiral fields, P and V_{μ} in the gauge sector even though there is no chiral gauge transformation in the matter Lagrangian.

The transformations for the fields contained in A are obtained after a long algebraic work. The result is

$$\delta S=\partial_{\mu}\xi^{\mu}+H^{\mu\nu}\partial_{\mu}\xi_{\nu}-\tfrac{1}{4}I^{\mu\nu\lambda}\partial_{\mu}\omega_{\nu\lambda}, \tag{3.16}$$

$$\delta A_{\mu}=-\partial_{\mu}\alpha+iB_{\mu\nu}\xi^{\nu}-i\omega_{\mu\nu}A^{\nu}+H_{\mu\nu}\partial^{\nu}\alpha \tag{3.17}$$
$$+\frac{i}{2}\partial^{\nu}\omega_{\nu\mu}+\frac{i}{2}H^{\lambda\nu}\partial_{\lambda}\omega_{\nu\mu}-iI_{\lambda\mu\nu}\partial^{\lambda}\xi^{\nu},$$

$$\delta B_{\mu\nu}=-2i(\partial_\mu\xi_\nu-\partial_\nu\xi_\mu)+4i(\xi_\mu A_\nu-\xi_\nu A_\mu)$$
$$+i(\omega^\lambda{}_\mu B_{\lambda\nu}-\omega^\lambda{}_\nu B_{\lambda\mu})-2i(H_{\lambda\mu}\partial^\lambda\xi_\nu-H_{\lambda\nu}\partial^\lambda\xi_\mu)$$
$$+2I_{\lambda\mu\nu}\partial^\lambda\alpha+i(I_{\lambda\rho\mu}\partial^\lambda\omega^\rho{}_\nu-I_{\lambda\rho\nu}\partial^\lambda\omega^\rho{}_\mu), \tag{3.18}$$

$$\delta P=-2\xi_\mu V^\mu-\frac{i}{8}\epsilon_{\mu\nu\lambda\rho}I^{\alpha\mu\nu}\partial_\alpha\omega^{\lambda\rho}, \tag{3.19}$$

$$\delta V_\mu=-2P\xi_\mu-i\omega_{\mu\nu}V^\nu+\tfrac{1}{4}\epsilon_{\mu\nu\lambda\rho}\partial^\nu\omega^{\lambda\rho}$$
$$+\tfrac{1}{4}\epsilon_{\mu\nu\lambda\rho}H^{\alpha\nu}\partial_\alpha\omega^{\lambda\rho}-\tfrac{1}{2}\epsilon_{\mu\nu\lambda\rho}I^{\alpha\nu\lambda}\partial_\alpha\xi^\rho. \tag{3.20}$$

The transformations of A_μ and $B_{\mu\nu}$ contain the usual terms, that is to say, $\partial_\mu\alpha$ and $\partial_\mu\xi_\nu-\partial_\nu\xi_\mu$, respectively, but also contain some other terms in order to render the gauge invariant condition to the kinetic Lagrangian. It is interesting to notice, in these transformations, the presence of the terms $iB_{\mu\nu}\xi^\nu$ and $4i(\xi_\mu A_\nu-\xi_\nu A_\mu)$ (terms involving the same initial gauge fields coupled with the vector parameter ξ_μ).

IV. ON THE INVARIANT LAGRANGIAN FOR THE KINETICAL GAUGE SECTOR

In the usual case of invariant gauge theories just involving vector fields, the corresponding Lagrangian for the kinetical sector can be directly obtained by introducing the curvature gauge tensor, i.e.,

$$F_{\mu\nu}=i[D_\mu,D_\nu]=\partial_\mu A_\nu-\partial_\nu A_\mu, \tag{4.1}$$

where $D_\mu=\partial_\mu-iA_\mu$ is the covariant derivative. In the case of theories involving gauge fields of rank higher than one, the concept of covariant derivative is here meaningless. However, the stress tensor for these theories can be introduced as a direct extension of the vector case. For example, for a tensor gauge potential $B_{\mu\nu}$, we have

$$H_{\mu\nu\rho}=\partial_\mu B_{\nu\rho}+\partial_\rho B_{\mu\nu}+\partial_\nu B_{\rho\mu}. \tag{4.2}$$

We mention that the non-Abelian extension of this problem is much more subtle and the obtainment of the invariant Lagrangian is, in some sense, still controversial.[9] A similar problem appears to occur here, since the gauge transformations given by (3.16)–(3.20) have a non-Abelian nature. Let us try to circumvent this problem by following an iterative process. We then start from the simplest gauge invariant Lagrangian involving vector and tensor gauge fields, i.e.,

$$\mathcal{L}_g=-\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}+\tfrac{1}{12}H_{\mu\nu\lambda}H^{\mu\nu\lambda}. \tag{4.3}$$

This Lagrangian is invariant under the well-known gauge transformations,

$$\delta A_\mu=-\partial_\mu\alpha,$$
$$\delta B_{\mu\nu}=-2i(\partial_\mu\xi_\nu-\partial_\nu\xi_\mu), \tag{4.4}$$

that are just part of the full gauge transformations given by (3.17) and (3.18).

The iterative process consists in increasing step-by-step these gauge transformations and looking for the corrections we have to do in the previous corresponding Lagrangian, in order to get invariance under these partial transformations. So, instead of the simplest gauge transformations (4.4), we look at (3.17) and (3.18) and take

$$\delta A_\mu = -\partial_\mu \alpha + i B_{\mu\nu} \xi^\nu,$$
$$\delta B_{\mu\nu} = -2i(\partial_\mu \xi_\nu - \partial_\nu \xi_\mu) + 4i(\xi_\mu A_\nu - \xi_\nu A_\mu). \tag{4.5}$$

Now, the Lagrangian (4.3) is not invariant under the transformations above. Actually,

$$\delta \mathcal{L}_g = i\partial^\mu F_{\mu\nu} B^{\nu\lambda} \xi_\lambda - 4i\partial^\mu H_{\mu\nu\lambda} \xi^\nu A^\lambda. \tag{4.6}$$

We try to obtain a gauge invariant Lagrangian by adding to $\mathcal{L}_g$ two new terms involving two auxiliary fields, X^μ and $Y^{\mu\nu}$ (antisymmetric in μ and ν). After some attempts, we conclude that these terms are

$$\mathcal{L}_X = X^\nu \partial^\mu F_{\mu\nu} + \tfrac{1}{2}\partial_\mu X^\mu \partial_\nu X^\nu - \tfrac{1}{2}\partial_\mu X^\nu \partial^\mu X_\nu,$$
$$\mathcal{L}_Y = -\tfrac{1}{2} Y^{\nu\lambda} \partial^\mu H_{\mu\nu\lambda} + 2\partial_\mu Y^{\mu\lambda} \partial^\nu Y_{\nu\lambda} - \tfrac{1}{2}\partial_\mu Y^{\nu\lambda} \partial^\mu Y_{\nu\lambda}. \tag{4.7}$$

The Lagrangian $\mathcal{L}_g + \mathcal{L}_X + \mathcal{L}_Y$ is invariant under the transformations (4.5) since

$$\delta X^\mu = -iB^{\mu\lambda} \xi_\lambda, \quad \delta Y^{\mu\nu} = -4i(\xi^\mu A^\nu - \xi^\nu A^\mu). \tag{4.8}$$

We then easily observe that X^μ and $Y^{\mu\nu}$ are an example of collective fields.[8] This means that the Lagrangian $\mathcal{L}_g + \mathcal{L}_X + \mathcal{L}_Y$ can be trivially rewritten as

$$\mathcal{L}_g + \mathcal{L}_X + \mathcal{L}_Y = -\tfrac{1}{4}\tilde{F}_{\mu\nu}\tilde{F}^{\mu\nu} + \tfrac{1}{12}\tilde{H}_{\mu\nu\lambda}\tilde{H}^{\mu\nu\lambda}, \tag{4.9}$$

where

$$\tilde{F}_{\mu\nu} = \partial_\mu (A_\nu - X_\nu) - \partial_\nu (A_\mu - X_\mu),$$
$$\tilde{H}_{\mu\nu\rho} = \partial_\mu (B_{\nu\rho} - Y_{\nu\rho}) + \partial_\rho (B_{\mu\nu} - Y_{\mu\nu}) + \partial_\nu (B_{\rho\mu} Y_{\rho\mu}). \tag{4.10}$$

We observe that this kind of procedure lead always to the introduction of collective fields. Although they can play important roles in the derivation of Ward identities[8] or in the implementation of quantization procedures such as the field antifield formalism,[8,10] they do not modify the physical content displayed by the original theory without using collective fields.

V. CONCLUSION

In this paper, we have extended the gauge transformation of the matter field Lagrangian by first introducing a vector gauge parameter, besides de scalar one. We have seen that the gauge transformation with these two parameters does not close in an algebra. This is only achieved if a local Lorentz-type gauge transformation is also included. We have seen that a tensor gauge field naturally arises in this procedure to compensate the extra terms that appear in the transformation of the kinetic Lagrangian. In addition, it also arises scalar, pseudo scalar and axial fields, besides the usual vector one. We have also show that the Lagrangians for the gauge sector are only obtained by means of collective fields that, even though may be useful for writing auxiliary Lagrangians in the quantization procedure due to Batalin–Vilkovisky,[10] they lead to a trivial result in the case of kinetic Lagrangians. This problem continue under study. We are trying to find out some alternative mechanism of getting invariant Lagrangians without using collective fields and also avoiding to define a curvature tensor for the theory. Possible results shall be reported elsewhere.[11]

ACKNOWLEDGMENTS

This work is supported in part by Conselho Nacional de Desenvolvimento Cientifico e Tecnológico-CNPq, Financiadora de Estudos e Projetos-FINEP, and Fundação Universitária José Bonifácio-FUJB (Brazilian Research Agencies).

APPENDIX: DEMONSTRATION THAT GAUGE TRANSFORMATION (3.9) AND (3.10) CLOSE IN AN ALGEBRA

In this Appendix, we are going to show that the transformations (3.9) and (3.10) close in an algebra, consistently with (2.5), independently of the form of K^μ and A, since Γ is given by (3.3).

We have already seen that the transformation (3.2) satisfies the following algebra

$$[\delta_2,\delta_1]\psi=i\Gamma_3\psi, \tag{A1}$$

where

$$\Gamma_3=i[\Gamma_1,\Gamma_2]=\alpha_3(x)+\xi_3^\mu(x)\gamma_\mu-\tfrac{1}{2}\omega(x)_3^{\mu\nu}\Sigma_{\mu\nu}, \tag{A2}$$

with α_3, ξ_3^μ, and $\omega_3^{\mu\nu}$ being given by (2.6).

For the transformation (3.9), we have

$$[\delta_2,\delta_1]K^\mu=-[\Gamma_2,[\Gamma_1,K^\mu]]+[\Gamma_1,[\Gamma_2,K^\mu]]=[K^\mu,[\Gamma_2,\Gamma_1]]=i[\Gamma_3,K^\mu], \tag{A3}$$

where we have used the Jacobi identity and the expression (A2). We actually see that the algebra closes independently of the form we have for K^μ.

For the next relation, we have

$$\begin{aligned}\delta_2\delta_1A&=i[\Gamma_2,\delta_1,A]+\delta_1K^\mu\partial_\mu\Gamma_2\\&=-[\Gamma_2,[\Gamma_1,A]]+i[\Gamma_2,K^\mu\partial_\mu\Gamma_1]+\delta_1K^\mu\partial_\mu\Gamma_2\\&=-[\Gamma_2,[\Gamma_1,A]]+\delta_2K^\mu\partial_\mu\Gamma_1+iK^\mu[\Gamma_2,\partial_\mu\Gamma_1]+\delta_1K^\mu\partial_\mu\Gamma_2.\end{aligned} \tag{A4}$$

Hence,

$$\begin{aligned}[\delta_2,\delta_1]A&=-[\Gamma_2,[\Gamma_1,A]]-[\Gamma_1,[A,\Gamma_2]]+iK^\mu[\Gamma_2,\partial_\mu\Gamma_1]+iK^\mu[\Gamma_1,\partial_\mu\Gamma_2]\\&=[A,[\Gamma_2,\Gamma_1]]+iK^\mu\partial_\mu\Gamma_3=i[\Gamma_3,A]+K^\mu\partial_\mu\Gamma_3,\end{aligned} \tag{A5}$$

where we have used again the Jacobi identity and (A2).

[1] M. Kalb and P. Ramond, Phys. Rev. D **9**, 2273 (1974); M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, New Jersey, 1992), and references therein.
[2] R. K. Kaul, Phys. Rev. D **18**, 1127 (1978); C. R. Hagen, *ibid.* **19**, 2367 (1979); R. Banerjee and E. C. Marino, Nucl. Phys. B **507**, 501 (1997); R. Banerjee and J. Barcelos-Neto, Ann. Phys. **265**, 134 (1998), and references therein.
[3] J. Maharana and J. H. Schwarz, Nucl. Phys. B **390**, 3 (1993).
[4] T. E. Clark and S. T. Love, Nucl. Phys. B **223**, 135 (1983); W. Bardeen and N. Deo, *ibid.* **264**, 364 (1986); A. Das and J. Barcelos-Neto, Mod. Phys. Lett. A **5**, 2573 (1990), and references therein.
[5] R. Amorim and J. Barcelos-Neto, Phys. Lett. B **426**, 329 (1998).
[6] E. Cremmer and J. Scherk, Nucl. Phys. B **72**, 117 (1974); T. J. Allen, M. J. Bowick, and A. Lahiri, Mod. Phys. Lett. A **6**, 559 (1991).
[7] J. Barcelos-Neto and S. Rabello, Z. Phys. C **74**, 715 (1997).
[8] J. Alfaro and P. H. Damgard, Nucl. Phys. **404**, 75 (1993).
[9] J. Barcelos-Neto, A. Cabo, and M. B. D. Silva, Z. Phys. C **72**, 34 (1996); A. Lahiri, Phys. Rev. D **55**, 5045 (1997).
[10] I. A. Batalin and G. A. Vilkovisky, Phys. Lett. B **102**, 27 (1981); Phys. Rev. D **28**, 2567 (1983). For a general review, see M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, New Jersey, 1992), and references therein.
[11] R. Amorim and J. Barcelos-Neto (in progress).

Large-order perturbation theory for a non-Hermitian $\mathcal{PT}$-symmetric Hamiltonian

Carl M. Bender
Department of Physics, Washington University, St. Louis, Missouri 63130

Gerald V. Dunne[a]
Department of Physics, University of Connecticut, Storrs, Connecticut 06269

(Received 30 November 1998; accepted for publication 21 June 1999)

A precise calculation of the ground-state energy of the complex $\mathcal{PT}$-symmetric Hamiltonian $H=p^2+\frac{1}{4}x^2+i\lambda x^3$, is performed using high-order Rayleigh–Schrödinger perturbation theory. The energy spectrum of this Hamiltonian has recently been shown to be real using numerical methods. Here we present convincing numerical evidence that the Rayleigh–Schrödinger perturbation series is Borel summable, and show that Padé summation provides excellent agreement with the real energy spectrum. Padé analysis provides strong numerical evidence that the once-subtracted ground-state energy considered as a function of λ^2 is a Stieltjes function. The analyticity properties of this Stieltjes function lead to a dispersion relation that can be used to compute the imaginary part of the energy for the related real but unstable Hamiltonian $H=p^2+\frac{1}{4}x^2-\epsilon x^3$. *© 1999 American Institute of Physics.* [S0022-2488(99)01810-1]

It has been conjectured[1] that the spectrum of the complex Hamiltonian,

$$H=p^2+\frac{1}{4}x^2+i\lambda x^3, \tag{1}$$

is real and positive. Although there is no rigorous proof of this conjecture, it has been argued[2] that the reality and positivity of the spectrum is a consequence of the $\mathcal{PT}$ symmetry of H. (Recall that the parity operation acts as $\mathcal{P}: p\to -p$ and $\mathcal{P}: x\to -x$ and that the antiunitary time reversal operation acts as $\mathcal{T}: p\to -p$, $\mathcal{T}: x\to -x$, and $\mathcal{T}: i\to -i$.) The notion that $\mathcal{PT}$ symmetry can replace the much more restrictive condition of hermiticity has been studied in the context of quasiexactly solvable quantum theories,[3] new kinds of symmetry breaking in quantum field theory,[4,5] and complex periodic potentials.[6] There have been many other instances of non-Hermitian $\mathcal{PT}$-invariant Hamiltonians in physics. Energies of solitons in Toda theories with imaginary coupling have been found to be real.[7] Hamiltonians rendered non-Hermitian by an imaginary external field have been used to study population biology[8] and to study delocalization transitions, such as vortex flux-line depinning in type-II superconductors.[9]

In this paper we study the large-order behavior of Rayleigh–Schrödinger perturbation theory for the ground-state energy of the complex $\mathcal{PT}$-symmetric Hamiltonian (1). Note that this Hamiltonian describes a $0+1$-dimensional ϕ^3 field theory, and recall that ϕ^3 theories were the first quantum field theories in which the divergences of perturbation theory were studied.[10] For the Hamiltonian (1) the perturbation series for the ground-state energy is divergent, and we give strong numerical evidence that it is Borel summable. Furthermore, by studying the numerical properties of the Padé approximants we infer that the (once-subtracted) ground-state energy considered as a function of λ^2 is a Stieltjes function. This is a very strong result because it implies analyticity in the cut-λ^2 plane and other properties. [It is surprising that this Stieltjes condition holds for a complex Hamiltonian such as (1); the proof that the once-subtracted ground-state

[a]Electronic mail: dunne@hep.phys.uconn.edu

TABLE I. The first 20 perturbation coefficients b_n in the expansion (2) of the ground-state energy for the complex $\mathcal{PT}$-symmetric Hamiltonian (1).

n	b_n
1	11
2	−930
3	158 836
4	−38 501 610
5	11 777 967 516
6	−4 300 048 271 460
7	1 815 215 203 378 344
8	−868 277 986 898 581 530
9	464 025 598 165 231 889 260
10	−274 145 574 452 876 905 074 540
11	177 549 419 941 607 942 489 064 216
12	−125 174 233 315 525 265 299 874 890 500
13	95 490 636 687 662 293 430 130 201 941 400
14	−78 410 748 996 991 270 671 939 611 723 389 320
15	68 982 408 758 305 101 330 092 396 215 438 198 608
16	−64 750 700 102 454 900 598 854 145 411 501 140 103 290
17	64 606 224 564 767 863 138 999 679 663 986 778 514 033 420
18	−68 291 871 149 169 980 983 310 351 232 642 663 615 057 109 020
19	76 244 729 314 392 095 958 565 433 992 857 306 551 429 203 990 968
20	−89 660 576 791 390 730 762 095 201 994 590 409 692 301 843 683 859 820

energy of the conventional λx^{2N} anharmonic oscillator is a Stieltjes function of λ makes use of hermiticity.] We then use these analyticity properties to establish a dispersion relation that yields the precise large-order behavior of the perturbation series.

Let us consider the conventional Rayleigh–Schrödinger perturbation series about the ground state ($E_0=\frac{1}{2}$) of the harmonic oscillator $H_0=p^2+\frac{1}{4}x^2$. The perturbed energy has an asymptotic series representation in powers of λ^2 because the perturbation x^3 is an odd function of x:

$$E(\lambda)-\frac{1}{2}\sim\sum_{n=1}^{\infty} b_n\lambda^{2n}. \tag{2}$$

[We have chosen the form of H_0 so that the perturbative expansion coefficients b_n in (2) are integers.]

Using recursion formulas, we can easily generate as many terms as desired in this expansion. The coefficients b_n alternate in sign, and their magnitude grows rapidly with n. The first 20 values are listed in Table I. We have computed enough of the coefficients b_n so that we can fit the leading large-n behavior as

$$b_n\sim(-1)^{n+1}\frac{60^{n+1/2}}{(2\pi)^{3/2}}\Gamma\left(n+\frac{1}{2}\right)\left[1-O\left(\frac{1}{n}\right)\right]. \tag{3}$$

Therefore, although divergent, the series in (2) is potentially Borel summable.[11,12] Observe that if the factor of i were absent from the Hamiltonian (1), then the perturbation coefficients b_n would not alternate in sign and the perturbation series would not be Borel summable.

We have performed a Padé analysis[11,12] on the divergent series for the once-subtracted ground-state energy $[E(\lambda)-\frac{1}{2}]/\lambda^2$. Using the first 46 perturbation coefficients b_n, we find that for all real positive λ^2 the diagonal Padé sequence $P_N^N(\lambda^2)$ is monotone decreasing with increasing N, and the off-diagonal Padé sequence $P_{M+1}^M(\lambda^2)$ is monotone increasing with increasing M:

$$P_1^0<P_2^1<P_3^2<\cdots<P_{M+1}^M<\cdots<P_N^N<\cdots<P_2^2<P_1^1<P_0^0. \tag{4}$$

TABLE II. The diagonal and off-diagonal Padé sequences $P_N^N(\lambda^2)$ and $P_{N+1}^N(\lambda^2)$ evaluated at $\lambda=0.125$. Observe the rapid convergence and note that the inequalities in (4) are satisfied.

N	P_N^N	P_{N+1}^N
0	11.000 000 000	4.739 290 085
1	7.039 037 169	5.696 806 799
2	6.347 866 015	5.947 600 655
3	6.168 265 727	6.026 389 220
4	6.110 857 028	6.054 574 069
5	6.089 906 566	6.065 678 176
6	6.081 499 968	6.070 392 205
7	6.077 873 385	6.072 516 805
8	6.076 216 002	6.073 522 627
9	6.075 421 823	6.074 018 882
10	6.075 025 816	6.074 272 525
11	6.074 821 510	6.074 406 195
12	6.074 712 942	6.074 478 558
13	6.074 653 729	6.074 518 675
14	6.074 620 680	6.074 541 394
15	6.074 601 848	6.074 554 510
16	6.074 590 917	6.074 562 214
17	6.074 584 462	6.074 566 813
18	6.074 580 592	6.074 569 597
19	6.074 578 237	6.074 571 306
20	6.074 576 787	6.074 572 368
21	6.074 575 882	6.074 573 036
22	6.074 575 311	6.074 573 460

The results for $\lambda=0.125$ are shown in Table II. If the inequalities in (4) hold for all N and M and for all real positive λ^2, then it is rigorously true that $[E(\lambda)-\frac{1}{2}]/\lambda^2$ is a Stieltjes function of λ^2.[12] This means that $[E(\lambda)-\frac{1}{2}]/\lambda^2$ is analytic in the cut-λ^2 plane, vanishes as $|\lambda^2|\to\infty$, and is a Herglotz function of λ^2. [A function $f(z)$ is said to be Herglotz if $\mathrm{Im}\, f(z)$ is positive (negative) when z is in the upper (lower) plane.] The fact that (4) holds for $0\leqslant M,\ N\leqslant 23$ provide strong numerical evidence that $[E(\lambda)-\frac{1}{2}]/\lambda^2$ is a Stieltjes function. We stress that this is a much stronger result than merely saying that the divergent series (2) is Borel summable.

Furthermore, in addition to the inequality in (4), the limits of the two Padé sequences appear to be identical. Therefore, we can extract values for the Padé summed energy from the two Padé sequences. The best estimate for the ground-state energy is obtained by averaging the last diagonal and off-diagonal Padé approximants. (To obtain an estimate of the ground-state energy from this average we multiply the average by λ^2 and add $\frac{1}{2}$.) The results are shown in Table III for various values of the coupling λ. Previous numerical calculations of the ground-state energy were obtained by direct numerical integration of the Schrödinger equation (see Ref. 2); this technique gave a typical accuracy of about five decimal places. The agreement between the method of numerical integration and the Padé summation is excellent. Moreover, for $\lambda<\frac{1}{10}$ the Padé technique provides an accuracy of more than ten decimal places. The agreement is better for smaller values of λ, as is expected, because of a faster convergence rate of the Padé sequence.

The above Padé analysis provides strong evidence that the once-subtracted ground-state energy is analytic in the cut-λ^2 plane. Thus, we can derive a dispersion relation in the expansion parameter λ^2 to deduce the leading behavior of the imaginary part of the energy for negative λ^2. Physically, this means that we can compute the imaginary part of the energy (and hence the decay width) of the unstable ground state of the *real* Hamiltonian,

$$H=p^2+\frac{1}{4}x^2-\epsilon x^3. \tag{5}$$

TABLE III. The ground-state energy for the Hamiltonian (1) for various values of the coupling λ; the ground-state energy was computed by Padé summation and by direct numerical integration. The Padé sequences were computed for the once-subtracted energy $[E(\lambda)-\frac{1}{2}]/\lambda^2$. The diagonal Padé energy refers to the energy extracted from the diagonal Padé sequence $P_N^N(\lambda^2)$, and the off-diagonal Padé energy refers to the energy extracted from the off-diagonal Padé sequence $P_{N+1}^N(\lambda^2)$. The best estimate for Padé energy is the average of the diagonal and off-diagonal values.

λ	Diagonal Padé energy	Off-diagonal Padé energy	Padé energy	Numerical energy
0.015 625	0.502 63	0.502 63	0.502 63	0.502 63
0.031 25	0.509 98	0.509 98	0.509 98	0.509 98
0.0625	0.533 93	0.533 93	0.533 93	0.533 93
0.125	0.594 92	0.594 92	0.594 92	0.594 92
0.25	0.713 05	0.712 84	0.712 95	0.712 94
0.5	0.914 45	0.890 35	0.902 40	0.900 26
1.0	1.400 07	1.058 17	1.229 12	1.167 46
2.0	3.160 75	1.140 32	2.150 53	1.530 78

Note that the ambiguity in the choice of the sign of the coupling ϵ corresponds to choosing the sign of i in (1). This has no effect on the decay width; the sign simply distinguishes the direction (left or right) in which the potential in (5) is unstable.

In the $t=\lambda^2$ plane there is a cut along the negative t axis, and in the standard way the[13–15] b_n coefficients are related to the discontinuity across the cut by the exact formula

$$b_n=\frac{1}{\pi}\int_0^\infty \frac{dt}{t}\,\frac{D(-t)}{t^n}, \tag{6}$$

where $D(-t)(t>0)$ is the imaginary part of $E(\lambda)-\frac{1}{2}$, evaluated with λ^2 negative. From the growth estimate (3) we deduce that

$$D(-t)\sim-\frac{e^{-1/(60t)}}{2\sqrt{2\pi t}}[1+O(t)] \quad (t\to 0^+). \tag{7}$$

Thus, the leading contribution (for small ϵ) to the imaginary part of the energy for the unstable ground state of the Hamiltonian (5) is

$$\mathrm{Im}[E(\epsilon)]\sim\frac{\exp\left(-\frac{1}{60\epsilon^2}\right)}{(2\pi)^{3/2}\epsilon} \quad (\epsilon\to 0^+). \tag{8}$$

There are several ways to check this result. First, it agrees with a direct leading-order WKB calculation[16] of the imaginary part of the energy of the unstable ground state of the real Hamiltonian (5). Second, applying the ''bounce'' method[17] to the real unstable Hamiltonian (5), we find that

$$\mathrm{Im}[E(\epsilon)]_{\mathrm{bounce}}\sim cS_0^{1/2}\exp(-S_0) \quad (\epsilon\to 0^+), \tag{9}$$

where the action S_0 of the bounce solution is given by

$$S_0=2\int_0^{1/4\epsilon} dx\,\sqrt{\frac{1}{4}x^2-\epsilon x^3}=\frac{1}{60\epsilon^2}, \tag{10}$$

and c is a constant (whose determination requires the computation of a fluctuation determinant).

Finally, the answer in (8) is in agreement with the variational perturbation theory analysis in Ref. 18. In fact, Ref. 18 contains a higher-order WKB expression for $\mathrm{Im}[E(\epsilon)]$. Inserting this higher-order WKB result into the dispersion relation (6), we obtain a WKB-based prediction for the corrections to the leading-order growth of the b_n coefficients given in (3):

$$b_n^{\mathrm{WKB}} \sim (-1)^{n+1} \frac{60^{n+1/2}}{(2\pi)^{3/2}} \Gamma\left(n+\frac{1}{2}\right)\left[1-\frac{169}{120(n-\frac{1}{2})}-\frac{44\,507}{28\,800(n-\frac{1}{2})(p-\frac{3}{2})}\right.$$
$$-\frac{9\,563\,539}{1\,920\,000(n-\frac{1}{2})(n-\frac{3}{2})(n-\frac{5}{2})}-\frac{189\,244\,716\,209}{8\,294\,400\,000(n-\frac{1}{2})(n-\frac{3}{2})(n-\frac{5}{2})(n-\frac{7}{2})}$$
$$-\frac{42\,943\,442\,679\,817}{331\,776\,000\,000(n-\frac{1}{2})(n-\frac{3}{2})(n-\frac{5}{2})(n-\frac{7}{2})(n-\frac{9}{2})}$$
$$-\frac{342\,541\,916\,236\,654\,541}{398\,131\,200\,000\,000(n-\frac{1}{2})(n-\frac{3}{2})(n-\frac{5}{2})(n-\frac{7}{2})(n-\frac{9}{2})(n-\frac{11}{2})}$$
$$\left.-\frac{933\,142\,404\,651\,555\,165\,943}{143\,327\,232\,000\,000\,000(n-\frac{1}{2})(n-\frac{3}{2})(n-\frac{5}{2})(n-\frac{7}{2})(n-\frac{9}{2})(n-\frac{11}{2})(n-\frac{13}{2})}-\cdots\right]. \tag{11}$$

With these higher-order corrections, this growth estimate of the b_n coefficients is spectacularly accurate. For example,

$$\frac{b_{46}^{\mathrm{WKB}}}{b_{46}} = 1.000\,000\,008\,07. \tag{12}$$

To conclude, we note that the strategy employed here to relate the large-order Rayleigh–Schrödinger perturbation theory coefficients of a stable (and Borel-summable) problem to the imaginary part of the energy of an unstable (and Borel-nonsummable) problem is familiar from the quartic double-well potential $H=p^2+\frac{1}{4}x^2+gx^4$, which is stable when $g>0$ and unstable when $g<0$.[19,13,15] The novelty in this paper is that we begin with a *complex* Hamiltonian $H=p^2+\frac{1}{4}x^2+i\lambda x^3$, which, despite being non-Hermitian, nevertheless appears to be stable in the sense that it has a real and positive (and discrete) energy spectrum and a Borel-summable perturbation expansion for the ground-state energy. We can then relate the large-order perturbation coefficients to the imaginary part of the energy of an unstable state of the real but unstable Hamiltonian $H=p^2+\frac{1}{4}x^2-\epsilon x^3$. It is interesting to note that the quartic case is relevant to the physics of instantons[20,17] while the cubic case is relevant to ''bounces'' in scalar field theories[17] and to string perturbation theory.[21]

ACKNOWLEDGMENT

We are grateful to the U.S. Department of Energy for financial support.

[1] D. Bessis, private discussion. This problem originated from discussions between Bessis and J. Zinn-Justin, who was studying Lee–Yang singularities using renormalization group methods.
[2] C. M. Bender and S. Boettcher, Phys. Rev. Lett. **80**, 5243 (1998).
[3] C. M. Bender and S. Boettcher, J. Phys. A **31**, L273 (1998).
[4] C. M. Bender and K. A. Milton, Phys. Rev. D **55**, R3255 (1997).
[5] C. M. Bender and K. A. Milton, Phys. Rev. D **57**, 3595 (1998).
[6] C. M. Bender, G. V. Dunne, and P. N. Meisinger, Phys. Lett. A **252**, 272 (1999).
[7] T. J. Hollowood, Nucl. Phys. B **386**, 166 (1992).
[8] D. R. Nelson and N. M. Shnerb, Phys. Rev. E **58**, 1383 (1998).
[9] N. Hatano and D. R. Nelson, Phys. Rev. Lett. **77**, 570 (1996); Phys. Rev. B **56**, 8651 (1997).
[10] C. A. Hurst, Proc. Cambridge Philos. Soc. **48**, 635 (1952); W. Thirring, Helv. Phys. Acta **26**, 33 (1953).
[11] G. Hardy, *Divergent Series* (Oxford University Press, Oxford, 1949).
[12] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw-Hill, New York, 1978), Chap. 8.

[13] C. M. Bender and T. T. Wu, Phys. Rev. Lett. **27**, 461 (1971); Phys. Rev. D **7**, 1620 (1973).
[14] B. Simon, Ann. Phys. (N.Y.) **45**, 76 (1970); I. Herbst and B. Simon, Phys. Rev. Lett. **41**, 67 (1978).
[15] *Large-Order Behavior of Perturbation Theory*, edited by J. C. Le Guillou and J. Zinn-Justin (North-Holland, Amsterdam, 1990).
[16] R. Yaris, J. Bendler, R. A. Lovett, C. M. Bender, and P. A. Fetters, Phys. Rev. A **18**, 1816 (1978).
[17] S. Coleman, ''The uses of instantons,'' in *Aspects of Symmetry* (Cambridge University Press, Cambridge, 1985).
[18] H. Kleinert and I. Mustapic, Int. J. Mod. Phys. A **11**, 4383 (1996).
[19] J. S. Langer, Ann. Phys. (N.Y.) **41**, 108 (1967).
[20] A. M. Polyakov, Nucl. Phys. B **120**, 429 (1977).
[21] D. J. Gross and V. Periwal, Phys. Rev. Lett. **60**, 2105 (1988).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

A conserved current for the field perturbations in the Einstein–Yang–Mills-dilaton–axion theory

R. Cartas-Fuentevilla[a)]
Instituto de Física, Universidad Autónoma de Puebla, Apartado postal J-48 72570 Puebla, Pue., Mexico and Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637

(Received 30 July 1998; accepted for publication 22 April 1999)

Using the self-adjoint character of the operators governing the field perturbations in the Einstein–Yang–Mills-dilaton–axion theory, we demonstrate that a covariantly (both gauge covariant and space–time covariant) conserved current associated with the coupled field perturbations arises, by contrast to other approaches, in a natural way. Our results cover, in particular, the bosonic low-energy limit of the string theory. We discuss how the present results can serve as a starting point for future investigations. © *1999 American Institute of Physics.* [S0022-2488(99)01409-7]

I. INTRODUCTION

The concepts of energy and momentum have been under investigation for a long time in the context of field theories involving gravity, in the general relativity itself, and in the modern unified theories. The main problem is to find a local expression that is physically meaningful and related to some form of continuity equation, which yields a conserved quantity. In the scheme of the unified theories such as supergravity and (super)string theory, the problem may become more complicated since they predict that gravity is mediated by one or several long-range scalar fields in addition to the ordinary tensor field of the Einstein theory, being the main feature of these scalar fields that they appear nonminimally coupled to the gravity and matter fields. However, although such conservation laws have not been established in the exact theories, some conservations laws are known in the context of perturbation theory, for example, in the frameworks of the ordinary Einstein–Maxwell theory,[1] Yang–Mills theory, and general relativity.[2,3] On the other hand, although the approach of Refs. 1, 2 may provide a general method for constructing a conserved current for the field perturbations—called the symplectic current—in various gravity theories, the main novelty that arises in the present approach is that the symplectic current does not need to be defined *a priori* (and subsequently it will have the wanted properties), but that it emerges naturally as a consequence of the self-adjointness of the equations for the field perturbations. The self-adjointness is, as will be seen below, an intrinsic property of the operators governing the field perturbations in the Einstein–Yang–Mills-dilaton–axion (EYMDA) theory.

We start with the (bosonic) four-dimensional effective action,

$$S=\int \sqrt{-g}d^4x\left\{R-2(\partial_\mu\phi)\partial^\mu\phi-\frac{1}{2}\zeta(\phi)(\partial_\mu\eta)\partial^\mu\eta+\xi(\phi)\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu})\right.$$
$$\left.+\omega(\eta)\mathrm{Tr}(F_{\mu\nu}\tilde{F}^{\mu\nu})+V(\phi,\eta)\right\}, \tag{1}$$

where R is the Ricci scalar, $F_{\mu\nu}$ is the Yang–Mills field given by the matrix-valued two-form $F_{\mu\nu}=2\nabla_{[\mu}A_{\nu]}+2A_{[\mu}A_{\nu]}$. Here we consider non-Abelian gauge fields with an arbitrary gauge group [then, the electromagnetic case is, of course, just that one corresponding to the gauge group $U(1)$]. $\tilde{F}^{\mu\nu}=(1/2\sqrt{-g})\,\epsilon^{\mu\nu\lambda\rho}F_{\lambda\rho}$ is the dual matrix of $F_{\mu\nu}$. Tr denotes, as usual, the trace.

[a)]Electronic mail: rcartas@sirio.ifuap.buap.mx

Furthermore, ϕ and η correspond to the dilaton and axion fields, respectively. The arbitrary coupling functions $\zeta(\phi)$, $\xi(\phi)$, and $\omega(\eta)$ depend only on ϕ and η and contain no derivatives of these fields. By suitable choice of these coupling functions, the low-energy effective action in string theory is covered. The peculiar nature of the dilaton and axion couplings to tensor fields is due to the presence of these functions. On the other hand, $V(\phi,\eta)$ is known as the dilaton–axion potential, and also depends on ϕ and η alone. V may be a Liouville-type dilaton potential (a cosmological constant term with dilaton coupling); V also may contain the possible mass terms for dilaton and axion fields, i.e., it can take the form $V=m_D\phi^2+m_A\eta^2$, where m_D and m_A are the masses of the dilaton and axion fields respectively, etc.

Variations of S with respect to the axion field, dilaton field, gauge field A_μ, and metric tensor $g_{\mu\nu}$ give, respectively, the following equations of motion:

$$\nabla_\mu(\zeta\ \partial^\mu\eta)+\frac{d\omega}{d\eta}\mathrm{Tr}(F_{\mu\nu}\tilde{F}^{\mu\nu})+\frac{\partial V}{\partial\eta}=0, \tag{2}$$

$$\boldsymbol{\nabla}_\mu\ \boldsymbol{\nabla}^\mu\phi+\frac{1}{4}\left[\frac{d\xi}{d\phi}\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu})-\frac{1}{2}\frac{d\zeta}{d\phi}(\partial_\mu\eta)\partial^\mu\eta+\frac{\partial V}{\partial\phi}\right]=0, \tag{3}$$

$$\mathbb{V}_\mu(\omega\tilde{F}^{\mu\nu}+\xi F^{\mu\nu})=0,\quad \mathbb{V}_\mu\tilde{F}^{\mu\nu}=0, \tag{4}$$

where $\mathbb{V}_\mu=\nabla_\mu+[A_\mu,\ ]$; and the Einstein field equations read as

$$R_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}R=T_{\mu\nu}, \tag{5}$$

with the energy–momentum tensor of matter given by

$$\begin{aligned}T_{\mu\nu}=&2(\partial_\mu\phi)\partial_\nu\phi+\tfrac{1}{2}\zeta(\partial_\mu\eta)\partial_\nu\eta-g_{\mu\nu}[(\partial^\alpha\phi)\partial_\alpha\phi+\tfrac{1}{4}\zeta(\partial^\alpha\eta)\partial_\alpha\eta]\\&-2\xi\mathrm{Tr}(F_{\mu\alpha}F_{\nu\beta}g^{\alpha\beta}-\tfrac{1}{4}g_{\mu\nu}F_{\rho\lambda}F^{\rho\lambda})+\tfrac{1}{2}g_{\mu\nu}V.\end{aligned} \tag{6}$$

The remainder of the paper is organized as follows. Section II is devoted to the linearization of the equations of motion (2)–(6) around a general curved background. In Sec. III we establish the general relationship between adjoint operators and conserved currents, which allows us in Sec. IV to find the corresponding symplectic current for the perturbed EYMDA theory. Finally, in Sec. V we conclude and discuss future directions that this research will take.

II. LINEARIZATION OF EYMDA THEORY

In the following expressions and what follows the superscript B denotes the corresponding linear perturbations. In particular, the metric, gauge potential, dilaton, and axion perturbations are represented by $h_{\mu\nu}$, b_μ, ϕ^{B}, and η^{B}, respectively.

Furthermore, we can find that

$$(g^{\mu\nu})^{\mathrm{B}}=-h^{\mu\nu},\quad g^{\mathrm{B}}=gg_{\mu\nu}h^{\mu\nu},\quad F^{\mathrm{B}}_{\mu\nu}=\mathbb{V}_\mu b_\nu-\mathbb{V}_\nu b_\mu,$$

$$(\tilde{F}^{\mu\nu})^{\mathrm{B}}=\frac{2}{\sqrt{-g}}\epsilon^{\mu\nu\alpha\beta}\mathbb{V}_\alpha b_\beta-\frac{1}{2}\tilde{F}^{\mu\nu}g_{\alpha\beta}h^{\alpha\beta},$$

$$\xi^{\mathrm{B}}=\frac{d\xi}{d\phi}\phi^{\mathrm{B}},\quad\left[\left(\frac{d\xi}{d\phi}\right)^{\mathrm{B}}=\frac{d^2\xi}{d\phi^2}\phi^{\mathrm{B}}\right],$$

$$\zeta^{\mathrm{B}}=\frac{d\zeta}{d\phi}\phi^{\mathrm{B}},\quad\omega^{\mathrm{B}}=\frac{d\omega}{d\eta}\eta^{\mathrm{B}},$$

$$
\begin{aligned}
V^{\mathrm{B}} &= \frac{\partial V}{\partial \phi}\phi^{\mathrm{B}} + \frac{\partial V}{\partial \eta}\eta^{\mathrm{B}},\\
\left(\frac{\partial V}{\partial \phi}\right)^{\mathrm{B}} &= \frac{\partial^2 V}{\partial \phi^2}\phi^{\mathrm{B}} + \frac{\partial^2 V}{\partial \eta \partial \phi}\eta^{\mathrm{B}},\\
\left(\frac{\partial V}{\partial \eta}\right)^{\mathrm{B}} &= \frac{\partial^2 V}{\partial \phi \partial \eta}\phi^{\mathrm{B}} + \frac{\partial^2 V}{\partial \eta^2}\eta^{\mathrm{B}},\\
(\Gamma^{\lambda}_{\mu\nu})^{\mathrm{B}} &= \tfrac{1}{2} g^{\lambda\rho}[\nabla_\mu h_{\nu\rho} + \nabla_\nu h_{\mu\rho} - \nabla_\rho h_{\mu\nu}],
\end{aligned}
\tag{7}
$$

in the above expressions and throughout, the covariant derivative ∇_μ is with respect to the background metric $g_{\mu\nu}$, and similarly $\not{\nabla}_\mu$ is with respect to both background metric and background gauge field A_μ. The indices are raised and lowered by means of the background metric, for example, $h^{\mu\nu} = g^{\mu\alpha} g^{\nu\beta} h_{\alpha\beta}$, that will be used implicitly below.

In addition, using the expressions (7), one easily finds that

$$
\begin{aligned}
[\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu})]^{\mathrm{B}} &= \mathrm{Tr}[4F^{\mu\nu}\not{\nabla}_\mu b_\nu + 2g_{\rho\nu}F^{\nu\alpha}F^{\mu\rho}h_{\mu\alpha}],\\
[\mathrm{Tr}(F_{\mu\nu}\tilde{F}^{\mu\nu})]^{\mathrm{B}} &= \mathrm{Tr}[4\tilde{F}^{\mu\nu}\not{\nabla}_\mu b_\nu - \tfrac{1}{2}F_{\rho\gamma}\tilde{F}^{\rho\gamma}g^{\mu\nu}h_{\mu\nu}].
\end{aligned}
\tag{8}
$$

In this manner, using the expressions (7) and (8) we can find that the linearized versions of the equations of motion (2)–(4) are given, respectively, by (after some arrangements)

$$
\begin{aligned}
&-\left\{\zeta\nabla_\mu\partial^\mu + (\partial^\mu\zeta)\partial_\mu + \frac{d^2\omega}{d\eta^2}\mathrm{Tr}(F_{\mu\nu}\tilde{F}^{\mu\nu}) + \frac{\partial^2 V}{\partial\eta^2}\right\}\eta^{\mathrm{B}} - \left\{\left[\nabla_\mu\left(\frac{d\zeta}{d\phi}\partial^\mu\eta\right)\right] + \frac{d\zeta}{d\phi}(\partial^\alpha\eta)\partial_\alpha\right.\\
&\left.+\frac{\partial^2 V}{\partial\eta\partial\phi}\right\}\phi^{\mathrm{B}} - 4\frac{d\omega}{d\eta}\mathrm{Tr}(\tilde{F}^{\mu\nu}\not{\nabla}_\mu b_\nu) + \{\zeta[(\nabla^\alpha\partial^\mu\eta) + (\partial^\alpha\eta)\nabla^\mu - \tfrac{1}{2}g^{\mu\alpha}(\partial^\rho\eta)\nabla_\rho]\\
&+(\partial^\mu\zeta)(\partial^\alpha\eta)\}h_{\mu\alpha} + \frac{1}{2}\frac{d\omega}{d\eta}\mathrm{Tr}(F_{\rho\lambda}\tilde{F}^{\rho\gamma}g^{\mu\alpha}h_{\mu\alpha}) = 0,
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
&\left\{\frac{d\zeta}{d\phi}(\partial^\mu\eta)\partial_\mu - \frac{\partial^2 V}{\partial\phi\,\partial\eta}\right\}\eta^{\mathrm{B}} - 4\left\{\frac{1}{4}\left[\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu})\frac{d^2\xi}{d\phi^2} - \frac{1}{2}\frac{d^2\zeta}{d\phi^2}(\partial^\mu\eta)(\partial_\mu\eta) + \frac{\partial^2 V}{\partial\phi^2}\right] + \nabla^\mu\nabla_\mu\right\}\phi^{\mathrm{B}}\\
&-4\frac{d\xi}{d\phi}\mathrm{Tr}(F^{\mu\nu}\not{\nabla}_\mu b_\nu) - 4\left\{-(\nabla^\alpha\nabla^\mu\phi) + \frac{1}{8}\frac{d\zeta}{d\phi}(\partial^\mu\eta)(\partial^\alpha\eta) - (\nabla^\alpha\phi)\nabla^\mu\right.\\
&\left.+\frac{1}{2}g^{\mu\alpha}(\nabla^\rho\phi)\nabla_\rho\right\}h_{\mu\alpha} - 2\frac{d\xi}{d\phi}\mathrm{Tr}(F^{\lambda\mu}F^{\alpha}{}_{\lambda}h_{\mu\alpha}) = 0,
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
&4g^{\lambda\mu}\tilde{F}_{\lambda\nu}\partial_\mu\left(\frac{d\omega}{d\eta}\eta^{\mathrm{B}}\right) + 4\not{\nabla}^\mu\left(\frac{d\xi}{d\phi}F_{\mu\nu}\phi^{\mathrm{B}}\right) + 4\left\{g_\nu{}^\mu\not{\nabla}^\rho(\xi\not{\nabla}_\rho) - \not{\nabla}^\mu(\xi\not{\nabla}_\nu)\right.\\
&\left.+\frac{1}{\sqrt{-g}}g_{\alpha\nu}\epsilon^{\rho\alpha\lambda\mu}(\partial_\rho\omega)\partial_\lambda\right\}b_\mu + 4\left\{\frac{1}{2}g^{\mu\alpha}[\xi F^\rho{}_\nu\nabla_\rho - (\partial_\rho\omega)\tilde{F}^\rho{}_\mu]\right.\\
&\left.-g^\mu{}_\nu[\xi F^{\rho\alpha}\nabla_\rho - (\partial_\rho\omega)\tilde{F}^{\rho\mu}]\right\}h_{\mu\alpha} - 4\not{\nabla}^\alpha(\xi F^\mu{}_\nu h_{\mu\alpha}) = 0,
\end{aligned}
\tag{11}
$$

where we have multiplied Eqs. (9)–(11) by the factors -1, -4, 4, respectively, for future convenience.[4] In order to linearize the Einstein equations (5) and (6), it is suitable to write the left-hand side of Eqs. (5) as $R^{\mathrm{B}}_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R^{\mathrm{B}}-\frac{1}{2}Rh_{\mu\nu}$, where

$$R^{\mathrm{B}}_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}R^{\mathrm{B}}=[\mathcal{E}_V(h_{\alpha\beta})]_{\mu\nu}=-\nabla_\mu\nabla_\nu g_{\alpha\beta}h^{\alpha\beta}-\nabla^\alpha\nabla_\alpha h_{\mu\nu}+\nabla^\rho\nabla_\nu h_{\rho\mu}+\nabla^\rho\nabla_\mu h_{\rho\nu}$$
$$+g_{\mu\nu}(\nabla^\rho\nabla_\rho g_{\alpha\beta}h^{\alpha\beta}-\nabla^\alpha\nabla^\beta h_{\alpha\beta}), \tag{12}$$

i.e., $\mathcal{E}_V$ is the operator describing gravitational perturbations of vacuum space–times.[4] The perturbed term $-\frac{1}{2}Rh_{\mu\nu}$ will be absorbed in the operator $\mathcal{E}_S$ in the next equation. In this manner, using Eqs. (7) and (8), the perturbed version of Einstein equations takes the form

$$-\left\{\zeta(\partial_{(\mu}\eta)\partial_{\alpha)}-\frac{1}{2}\zeta g_{\mu\alpha}(\partial^\rho\eta)\partial_\rho+\frac{1}{2}g_{\mu\alpha}\frac{\partial V}{\partial\eta}\right\}\eta^{\mathrm{B}}-\left\{4(\partial_{(\mu}\phi)\partial_{\alpha)}\right.$$
$$\left.-2g_{\mu\alpha}(\partial^\rho\phi)\partial_\rho+\frac{1}{2}\frac{d\zeta}{d\phi}\left[(\partial_\mu\eta)(\partial_\alpha\eta)-\frac{1}{2}g_{\mu\alpha}(\partial^\rho\eta)(\partial_\rho\eta)\right]+\frac{1}{2}g_{\mu\alpha}\frac{\partial V}{\partial\phi}-\frac{d\xi}{d\phi}\mathrm{Tr}(T^{\mathrm{YM}}_{\mu\alpha})\right\}\phi^{\mathrm{B}}$$
$$-2\xi\ \mathrm{Tr}(\{g_{\mu\alpha}F^{\lambda\gamma}\nabla_\lambda+2[F^\gamma{}_{(\mu}\nabla_{\alpha)}+g^\gamma{}_{(\mu}F_{\alpha)}{}^\lambda\nabla_\lambda]\}b_\gamma)+[(\mathcal{E}_V+\mathcal{E}_S)h_{\rho\gamma}]_{\alpha\mu}=0, \tag{13}$$

where

$$T^{\mathrm{YM}}_{\mu\nu}=2[F_{\mu\alpha}F_{\nu\beta}g^{\alpha\beta}-\tfrac{1}{4}g_{\mu\nu}F_{\rho\lambda}F^{\rho\lambda}],$$

is the usual energy–momentum tensor of the background Yang–Mills field; in addition, $\mathcal{E}_S$ represents the operator acting on the metric perturbations associated with the perturbed tensor of matter and only corresponds to a function, without containing differential operators.

The coupled equations for the field perturbations (9)–(11) and (13) can be identified with the first, second, third, and fourth rows, respectively, of the following matrix:

$$\begin{bmatrix} \mathcal{E}_A & \mathcal{E}_{AD} & \mathcal{E}_{A(\mathrm{YM})} & \mathcal{E}_{AG} \\ \mathcal{E}_{DA} & \mathcal{E}_D & \mathcal{E}_{D(\mathrm{YM})} & \mathcal{E}_{DG} \\ \mathcal{E}_{(\mathrm{YM})A} & \mathcal{E}_{(\mathrm{YM})D} & \mathcal{E}_{\mathrm{YM}} & \mathcal{E}_{(\mathrm{YM})G} \\ \mathcal{E}_{GA} & \mathcal{E}_{GD} & \mathcal{E}_{G(\mathrm{YM})} & \mathcal{E}_G \end{bmatrix}\begin{bmatrix}\eta^{\mathrm{B}} \\ \phi^{\mathrm{B}} \\ (b_\mu) \\ (h_{\mu\nu})\end{bmatrix}=0. \tag{14}$$

From Eqs. (9)–(11) and (13) we can note that the $\mathcal{E}$'s are linear partial differential operators involving only the background fields, which satisfy Eqs. (2)–(6); furthermore, the presence of the coupling functions and of the dilaton–axion potential makes that all field perturbations be present in each of the perturbation equations.

At first sight it might appear that, due to the very complicated appearance of Eqs. (9)–(11) and (13), they have not any special property; however, as will be seen in the next section, they have an intrinsic property, the self-adjointness, which will guarantee automatically the existence of a conserved current for the field perturbations.

III. ADJOINT OPERATORS AND CONSERVED CURRENTS

According to the definition introduced in Ref. 5, if $\mathcal{P}$ is a linear partial differential operator, which takes matrix-valued tensor fields into themselves, then, the adjoint operator of $\mathcal{P}$, is that operator $\mathcal{P}^\dagger$, such that

$$\mathrm{Tr}\{f^{\rho\sigma\cdots}[\mathcal{P}(g_{\mu\nu\ldots})]_{\rho\sigma\ldots}-[\mathcal{P}^\dagger(f^{\rho\sigma\cdots})]^{\mu\nu\cdots}g_{\mu\nu\ldots}\}=\nabla_\mu J^\mu, \tag{15}$$

where Tr denotes again the trace and J^μ is some vector field (this definition generalizes that given by Wald in Ref. 4). From this definition, if $\mathcal{Q}$ and $\mathcal{R}$ are any two linear operators, one easily finds that

$$(\mathcal{Q}\mathcal{R})^{\dagger}=\mathcal{R}^{\dagger}\mathcal{Q}^{\dagger},\quad (\mathcal{Q}+\mathcal{R})^{\dagger}=\mathcal{Q}^{\dagger}+\mathcal{R}^{\dagger}. \tag{16}$$

In the case of a function F, such as the operator $\mathcal{E}_S$ of Eq. (13),

$$F^{\dagger}=F. \tag{17}$$

Furthermore, from the definition (15) we can easily see that if f and g are two independent solutions admitted by any linear system $\mathcal{P}(f)=0=\mathcal{P}(g)$ and the operator $\mathcal{P}$ is self-adjoint, it means that $\mathcal{P}^{\dagger}=\mathcal{P}$ (or antiself-adjoint, $\mathcal{P}^{\dagger}=-\mathcal{P}$), then J^{μ} appearing on the right-hand side is a covariantly conserved quantity depending on the fields f and g.[6]

In this manner, the above discussion yields a completely general result, the existence of a conserved current for a self-adjoint system of the differential equations. The important point of this fact is the adjoint character of the system under study, and as we will see, this is effectively the case for the system of equations (14). For this purpose let us consider the element $\mathcal{E}_{A(\mathrm{YM})}$, which is taking the matrix-valued field b_{ν} into a scalar field in Eq. (9); using the properties of the Tr and the background equation $\nabla_{\mu}\tilde{F}^{\mu\nu}=0$ [see Eq. (2)], it is easy to demonstrate that

$$\mathrm{Tr}[\psi_1(\mathcal{E}_{A(\mathrm{YM})}B_{\nu})]=\mathrm{Tr}[B^{\nu}(\mathcal{E}_{(\mathrm{YM})A}\psi_1)_{\nu}]+\nabla_{\mu}\ \mathrm{Tr}\left[4\frac{d\omega}{d\eta}\tilde{F}^{\nu\mu}B_{\nu}\psi_1\right],$$

where ψ_1 is any scalar field, B_{μ} is any matrix-valued field, and $\mathcal{E}_{(\mathrm{YM})A}$ is that operator taking η^{B} (a scalar field) into a matrix-valued field in Eq. (11). The above expression has the form (15) and allows us to identify that

$$\mathcal{E}^{\dagger}_{A(\mathrm{YM})}=\mathcal{E}_{(\mathrm{YM})A}. \tag{18}$$

Similarly, using the definition (15), the properties (16) and (17), and assuming that the background fields satisfy Eqs. (2)–(6), we can demonstrate the following relations:

$$\psi_2\mathcal{E}_A\psi_1=\psi_1\mathcal{E}_A\psi_2+\nabla_{\mu}\zeta(\psi_1\ \partial^{\mu}\psi_2-\psi_2\ \partial^{\mu}\psi_1)(\mathcal{E}^{\dagger}_A=\mathcal{E}_A),$$

$$\psi_2\mathcal{E}_{DA}\psi_1=\psi_1\mathcal{E}_{AD}\psi_2+\nabla_{\mu}\left[\frac{d\zeta}{d\phi}\psi_1\psi_2\ \partial^{\mu}\eta\right](\mathcal{E}^{\dagger}_{DA}=\mathcal{E}_{AD}),$$

$$A^{\mu\alpha}(\mathcal{E}_{GA}\psi_1)_{\mu\alpha}=\psi_1(\mathcal{E}_{AG}A_{\mu\alpha})+\nabla_{\mu}\zeta\psi_1[\tfrac{1}{2}A_{\alpha}{}^{\alpha}\ \partial^{\mu}\eta-A^{\mu\nu}\ \partial_{\nu}\eta](\mathcal{E}^{\dagger}_{AG}=\mathcal{E}_{GA}),$$

$$\psi_2\mathcal{E}_D\psi_1=\psi_1\mathcal{E}_D\psi_2+\nabla_{\mu}4(\psi_1\ \partial^{\mu}\psi_2-\psi_2\ \partial^{\mu}\psi_1)(\mathcal{E}^{\dagger}_D=\mathcal{E}_D),$$

$$\mathrm{Tr}[\psi_1(\mathcal{E}_{D(\mathrm{YM})}B_{\nu})]=\mathrm{Tr}[B^{\nu}(\mathcal{E}_{(\mathrm{YM})D}\psi_1)_{\nu}]+\nabla_{\mu}\mathrm{Tr}\left(4\frac{d\xi}{d\phi}F^{\nu\mu}B_{\nu}\psi_1\right)(\mathcal{E}^{\dagger}_{D(\mathrm{YM})}=\mathcal{E}_{(\mathrm{YM})D}),$$

$$\psi_1(\mathcal{E}_{DG}A_{\mu\alpha})=A^{\mu\alpha}(\mathcal{E}_{GD}\psi_1)_{\mu\alpha}+\nabla_{\mu}2\psi_1[2A^{\mu}{}_{\alpha}\partial^{\alpha}\phi-A_{\alpha}{}^{\alpha}\partial^{\mu}\phi](\mathcal{E}^{\dagger}_{DG}=\mathcal{E}_{GD}),$$

$$\mathrm{Tr}[A^{\nu}(\mathcal{E}_{\mathrm{YM}}B_{\mu})_{\nu}]=\mathrm{Tr}(B^{\nu}(\mathcal{E}_{\mathrm{YM}}A_{\mu})_{\nu})+\nabla_{\mu}8\mathrm{Tr}\left\{\xi[A_{\rho}\nabla^{[\mu}B^{\rho]}+B_{\rho}\nabla^{[\rho}A^{\mu]}]\right.$$

$$\left.+\frac{1}{2\sqrt{-g}}\epsilon^{\rho\alpha\mu\lambda}A_{\alpha}B_{\lambda}\partial_{\rho}\eta\right\}(\mathcal{E}^{\dagger}_{\mathrm{YM}}=\mathcal{E}_{\mathrm{YM}}),$$

$$\mathrm{Tr}[A^{\mu\alpha}(\mathcal{E}_{G(\mathrm{YM})}B_{\nu})_{\mu\alpha}]=\mathrm{Tr}[B^{\nu}(\mathcal{E}_{(\mathrm{YM})G}A_{\mu\alpha})_{\nu}]+\nabla_{\mu}2\xi\ \mathrm{Tr}\{4B_{\nu}F_{\alpha}{}^{[\nu}A^{\mu]\alpha}-A_{\alpha}{}^{\alpha}B_{\nu}F^{\mu\nu}\}(\mathcal{E}^{\dagger}_{G(\mathrm{YM})}$$

$$=\mathcal{E}_{(\mathrm{YM})G}), \tag{19}$$

where ψ_2 is another scalar field, A_μ and B_μ are any two matrix fields, and $A_{\mu\nu}$ any 2-index (symmetric) tensor field (such as $h_{\mu\nu}$). On the other hand, since $\mathcal{E}_G=\mathcal{E}_V+\mathcal{E}_S$, using the expression for $\mathcal{E}_V$ given in Eq. (12), the property (17), we can find that

$$A^{\mu\nu}(\mathcal{E}_G B_{\alpha\rho})_{\mu\nu}=B^{\alpha\rho}(\mathcal{E}_G A_{\mu\nu})_{\alpha\rho}+\nabla_\mu S^{\mu\alpha\beta\lambda\rho\gamma}(A_{\alpha\beta}\nabla_\lambda B_{\rho\gamma}-B_{\alpha\beta}\nabla_\lambda A_{\rho\gamma}) \quad (\mathcal{E}_G^\dagger=\mathcal{E}_G), \tag{20}$$

where $B_{\mu\nu}$ is another 2-index (symmetric) tensor field and[1]

$$S^{\mu\alpha\beta\lambda\rho\gamma}=g^{\mu(\rho}g^{\gamma)(\alpha}g^{\beta)\lambda}-\tfrac{1}{2}g^{\mu\lambda}g^{\alpha(\rho}g^{\gamma)\beta}-\tfrac{1}{2}g^{\mu(\alpha}g^{\beta)\lambda}g^{\rho\gamma}-\tfrac{1}{2}g^{\alpha\beta}g^{\mu(\rho}g^{\gamma)\lambda}+\tfrac{1}{2}g^{\alpha\beta}g^{\mu\lambda}g^{\rho\gamma}. \tag{21}$$

We thereby have found from Eqs. (18)–(20) that the matrix operator in Eq. (14) governing the linear field perturbations for the EYMDA theory, is self-adjoint, that is[7]

$$\begin{bmatrix} \mathcal{E}_A & \mathcal{E}_{AD} & \mathcal{E}_{A(\mathrm{YM})} & \mathcal{E}_{AG} \\ \mathcal{E}_{DA} & \mathcal{E}_D & \mathcal{E}_{D(\mathrm{YM})} & \mathcal{E}_{DG} \\ \mathcal{E}_{(\mathrm{YM})A} & \mathcal{E}_{(\mathrm{YM})D} & \mathcal{E}_{\mathrm{YM}} & \mathcal{E}_{(\mathrm{YM})G} \\ \mathcal{E}_{GA} & \mathcal{E}_{GD} & \mathcal{E}_{G(\mathrm{YM})} & \mathcal{E}_G \end{bmatrix}^\dagger = \begin{bmatrix} \mathcal{E}_A & \mathcal{E}_{AD} & \mathcal{E}_{A(\mathrm{YM})} & \mathcal{E}_{AG} \\ \mathcal{E}_{DA} & \mathcal{E}_D & \mathcal{E}_{D(\mathrm{YM})} & \mathcal{E}_{DG} \\ \mathcal{E}_{(\mathrm{YM})A} & \mathcal{E}_{(\mathrm{YM})D} & \mathcal{E}_{\mathrm{YM}} & \mathcal{E}_{(\mathrm{YM})G} \\ \mathcal{E}_{GA} & \mathcal{E}_{GD} & \mathcal{E}_{G(\mathrm{YM})} & \mathcal{E}_G \end{bmatrix}. \tag{22}$$

It is important to remark that the only assumption that we have made in order to establish the intrinsic property (22) is that the background fields satisfy Eqs. (2)–(7); moreover, we would like point out that any operator appearing in physics is not necessarily self-adjoint.[8]

With the end of obtaining the explicit form of the symplectic current for the EYMDA field perturbations, let S_1 and S_2 be any two independent solutions admitted by the system (14), given by

$$(S_1)=\begin{pmatrix} \eta_1^{\mathrm{B}} \\ \phi_1^{\mathrm{B}} \\ (b_\mu) \\ (h_{\mu\nu}) \end{pmatrix}, \quad (S_2)=\begin{pmatrix} \eta_2^{\mathrm{B}} \\ \phi_2^{\mathrm{B}} \\ (B_\mu) \\ (H_{\mu\nu}) \end{pmatrix}, \tag{23}$$

and from definition (14),

$$S_2(\mathcal{E}S_1)-S_1(\mathcal{E}^\dagger S_2)=\nabla_\mu J^\mu, \tag{24}$$

where $\mathcal{E}$ is the matrix operator (22), and using Eqs. (19) we easily find that

$$J^\mu=J_A{}^\mu+J_{AD}{}^\mu+J_{A(\mathrm{YM})}{}^\mu+J_{AG}{}^\mu+J_D{}^\mu+J_{D(\mathrm{YM})}{}^\mu+J_{DG}{}^\mu+J_{\mathrm{YM}}{}^\mu+J_{(\mathrm{YM})G}{}^\mu+J_G{}^\mu, \tag{25}$$

where

$$J_A{}^\mu=\zeta(\eta_1^B\,\partial^\mu\eta_2^B-\eta_2^B\,\partial^\mu\eta_1^B), \quad J_{AD}{}^\mu=\frac{d\zeta}{d\phi}(\partial^\mu\eta)[\eta_1^B\phi_2^B-\phi_1^B\eta_2^B],$$

$$J_{A(\mathrm{YM})}{}^\mu=4\frac{d\omega}{d\eta}\mathrm{Tr}[\tilde{F}^{\mu\rho}[B_\rho\eta_1^B-b_\rho\eta_2^B]],$$

$$J_{\mathrm{AG}}{}^\mu=\zeta[(\partial_\rho\eta)[h^{\mu\rho}\eta_2^B-H^{\mu\rho}\eta_1^B]+\tfrac{1}{2}(\partial^\mu\eta)[H\eta_1^B-h\eta_2^B]],$$

$$J_D{}^\mu=4[\phi_1^B\partial^\mu\phi_2^B-\phi_2^B\,\partial^\mu\phi_1^B],$$

$$J_{D(\mathrm{YM})}{}^{\mu}=4\frac{d\xi}{d\phi}\mathrm{Tr}[F^{\mu\rho}[B_\rho\phi_1^B-b_\rho\phi_2^B]], \tag{26}$$

$$J_{DG}{}^{\mu}=2[2(\partial^\rho\phi)[h^\mu{}_\rho\phi_2^B-H^\mu{}_\rho\phi_1^B]+(\partial^\mu\phi)[H\phi_1^B-h\phi_2^B]],$$

$$J_{\mathrm{YM}}{}^{\mu}=8\,\mathrm{Tr}\left\{\xi[B_\rho\,\nabla^{[\mu}b^{\rho]}-b_\rho\,\nabla^{[\mu}B^{\rho]}]+\frac{1}{2\sqrt{-g}}\epsilon^{\rho\alpha\mu\lambda}(\partial_\rho\eta)B_\alpha b_\lambda\right\},$$

$$J_{(\mathrm{YM})G}{}^{\mu}=2\xi\ \mathrm{Tr}\{2F_\rho{}^\mu[h^{\rho\lambda}B_\lambda-H^{\rho\lambda}b_\lambda]+2F_{\rho\lambda}[H^{\mu\rho}b^\lambda-h^{\mu\rho}B^\lambda]+F^{\mu\rho}(hB_\rho-Hb_\rho)\},$$

$$J_G{}^{\mu}=S^{\mu\alpha\beta\lambda\rho\gamma}(H_{\alpha\beta}\,\nabla_\lambda h_{\rho\gamma}-h_{\alpha\beta}\,\nabla_\lambda H_{\rho\gamma}),$$

where $H=g^{\mu\nu}H_{\mu\nu}$, $h=g^{\mu\nu}h_{\mu\nu}$ and the subscripts only denote the types of field perturbations involved; for example, $J_{A(\mathrm{YM})}{}^{\mu}$ involves axion and Yang–Mills perturbations. Note that J^μ is, just like the symplectic currents constructed by means of the approach of Refs. 1, 2, a function of an unperturbed set of fields and is bilinear and antisymmetric in pairs of field perturbations. Moreover, the dilaton–axion potential does not contribute to the symplectic current (those possible mass terms, etc.).

Our expression for J^μ is not gauge invariant with respect to either ordinary gauge transformations $B_\mu\to B_\mu+\partial_\mu\chi+[A_\mu,\chi]$ and $b_\mu\to b_\mu+\partial_\mu\chi+[A_\mu,\chi]$ or infinitesimal diffeomorphisms. However, the integral of J^μ over a compact Cauchy surface is gauge invariant.[1,2] On the other hand, since the dilaton and axion are fundamental physical fields (i.e., their zero modes are physically meaningful), there are no gauge invariances associated with these fields.

We can compare the expression for $J_{\mathrm{YM}}{}^{\mu}$ in Eq. (26) with that for the symplectic current introduced in 3, and note that they have the same form in terms of the gauge field perturbations. Moreover, if we set $\phi=\eta=\omega=V=0$ and $\xi=-k$ (gravitational constant) without imposing any restriction on the remaining gauge field in the background according to Eqs. (2), (3), it is not difficult to show that the sum [in the case when the gauge group corresponds to $U(1)$] $J_{\mathrm{YM}}{}^{\mu}+J_{(\mathrm{YM})G}{}^{\mu}+J_G{}^{\mu}$ reduces exactly to the expression (3.12) in Ref. 1 for the symplectic current in the Einstein–Maxwell theory.

IV. CONCLUDING REMARKS

In conclusion, we have applied to a system of perturbation equations derived from a string-inspired action, a general result that connects the self-adjointness of any linear system of differential equations with the existence of a conserved current. The applicability of the present approach depends on the fact that the self-adjoint character of the system under study is established. However, if a given system is not self-adjoint, one can still be able to find a conserved current, which will be associated, formally, to a solution of the original system and to a solution of the adjoint system, in accordance with the definition (15); but, at present, we do not know in general what one must understand physically by ‘‘a solution of the adjoint system.’’ In this sense, the present approach can be, in principle, a general method to generate a conserved current for any gravity theory, such as the approaches of Refs. 1, 2.

In addition to that, our approach is a more direct procedure of obtaining the symplectic current than that employed in Refs. 1, 2, it has another advantage: the self-adjointness allows the generation of solutions for perturbation equations by means of potentials, provided the corresponding decoupled equations are obtained (originally, with this idea the concept of adjoint operators was introduced by Wald[4] and Torres del Castillo[5]); that will permit, in its turn, a detailed study of the corresponding symplectic current, since we do not have yet a general physical interpretation for such a current (however, a connection with the conservation of energy has been shown in a particular case in Ref. 1; see also Ref. 3). Reference 8 shows how our approach applies when the string fields are involved, and references cited therein correspond to some particular cases whose perturbation study may be very interesting.

On the other hand, the action considered in the present work contains only the (truncated) bosonic sector of a generic low-energy string effective theory; the question of whether the presence of the additional terms in the generic action yields a self-adjoint perturbation equation remains to be worked out. Moreover, the inclusion of fermionic fields (superstrings theory) is possible, since the fundamental definition of Ref. 5, which has been our starting point, extends for spinor fields. All these questions will be the subject of a future investigation.

ACKNOWLEDGMENTS

This work was supported in part by CONACYT (México) and the Sistema Nacional de Investigadores (México). The author wishes to thank Dr. G. F. Torres del Castillo for his comments on the manuscript and suggestions.

[1] G. A. Burnett and R. M. Wald, Proc. R. Soc. London, Ser. A **430**, 57 (1990).
[2] J. Lee and R. M. Wald, J. Math. Phys. **31**, 725 (1990).
[3] C. Crnković and E. Witten, in *Three Hundred Years of Gravitation*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1987).
[4] R. M. Wald, Phys. Rev. Lett. **41**, 203 (1978).
[5] G. F. Torres del Castillo, Gen. Relativ. Gravit. **22**, 1085 (1990).
[6] G. F. Torres del Castillo and J. C. Flores-Urbina, ''Self-adjoint operators and conserved currents,'' to appear in Gen. Relativ. Gravit.
[7] B. P. Jeffryes, Class. Quantum Grav. **4**, L17 (1987).
[8] R. Cartas-Fuentevilla, Phys. Rev. D **56**, 7700 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

On the number of particles that a curved quantum waveguide can bind

Pavel Exner[a)]
Nuclear Physics Institute, Academy of Sciences, 25068 Řež near Prague, Czech Republic and Doppler Institute, Czech Technical University, Břehová 7, 11519 Prague, Czech Republic

Simeon A. Vugalter[b)]
Nuclear Physics Institute, Academy of Sciences, 25068 Řež near Prague, Czech Republic

(Received 9 February 1998; accepted for publication 8 April 1999)

We discuss the discrete spectrum of N particles in a curved planar waveguide. If they are neutral fermions, the maximum number of particles that the waveguide can bind is given by a one-particle Birman–Schwinger bound in combination with the Pauli principle. On the other hand, if they are charged, e.g., electrons in a bent quantum wire, the Coulomb repulsion plays a crucial role. We prove a sufficient condition under which the discrete spectrum of such a system is empty. © *1999 American Institute of Physics.* [S0022-2488(99)01309-2]

I. INTRODUCTION

A rapid progress of mesoscopic physics brought, in particular, interesting new problems concerning relations between geometry and spectral properties of quantum Hamiltonians. They involve models of quantum wires, dots, and similar systems. While in reality these are rather complicated systems composed of different semiconductor materials, experience tells us that their basic features can be explained using simple models in which electrons (regarded as free particles with an effective mass) are supposed to be confined to an appropriate spatial region, either by a potential or by a hard wall. A brief description of this approximation with a guide to further reading is given in Ref. 1. In addition, such models apply not only to electrons in semiconductor microstructures; a different example is represented by atoms trapped in hollow optical fibers.[2]

It is natural that most theoretical results up to date refer to the case of a single particle in the confinement. On the other hand, from the practical point of view it is rather an exception than a rule that an experimentalist is able to isolate a single electron or atom, and therefore many-body problems in this setting are of interest. For instance, two-dimensional quantum dots that can be regarded as artificial atoms have been studied recently, usually in the presence of a magnetic field, either for a pair of electrons or in the semiclassical situation when a Thomas–Fermi-type approach is applicable—cf. Refs. 3–6 and references therein.

In these studies, however, geometry of the dot played a little role, because the confinement was realized by a harmonic potential or a circular hard wall. This is not the case for open systems modeling quantum wires where a deformation of a straight channel is needed to produce nontrivial spectral properties. In particular, a quantum waveguide exhibit bound states if it is bent,[1,7,8] protruded,[9–11] or allowing a leak to another duct,[12–14] and the discrete spectrum depends substantially on the shape of the channel. With few exceptions such as Ref. 15, however, the known results refer to the one-particle case.

It is our aim in the present paper to initiate a rigorous investigation of many-particle effects in quantum waveguides. We are going to discuss here a system of N particles in a bent planar Dirichlet tube, i.e., a hard-wall channel, and ask whether N-particle bound states exist for a given

[a)]Electronic mail: exner@ujf.cas.cz
[b)]Present address: Galgenbergerstr. 19, 92637 Weiden, Germany. Electronic mail: wugalter@new-wen.baynet.de

geometry. After collecting the necessary preliminaries in the next section, we shall derive first in Sec. III a simple bound for the neutral case that follows from the Birman–Schwinger estimate of the one-particle Hamiltonian in combination with the Pauli principle.

The main result of the paper is formulated and proved in Sec. IV. It concerns the physically interesting case of charged particles; the example we have in mind is, of course, electrons in a bent semiconductor quantum wire. The electrostatic repulsion makes spectral analysis of the corresponding Hamiltonian considerably more complicated. Using a variational technique borrowed from atomic physics, we derive here a sufficient condition under which the discrete spectrum is empty. The condition is satisfied for N large enough and represents an implicit equation for the maximum number of charged particles that a waveguide of a given curvature and width can bind. Some other aspects of the result and open questions are discussed briefly in the concluding section.

II. PRELIMINARIES

The waveguide in question will be modeled by a curved planar strip Σ in $\mathbb{R}^2$, of a constant width $d=2a$. It can be obtained by transporting the perpendicular interval $[-a,a]$ along the curve Γ, which is the axis of Σ. Up to Euclidean transformations, the strip is uniquely characterized by its halfwidth a and the (signed) curvature $s\mapsto\gamma(s)$ of Γ, where s denotes the arclength. We adopt the regularity assumptions of Refs. 1,7: (i) Ω is not self-intersecting, (ii) $a\|\gamma\|_\infty<1$, (iii) γ is piecewise C^2 with γ', γ'' bounded, and restrict our attention to the case when the tube is curved in a bounded region only: and (iv) there is $b>0$ such that $\gamma(s)=0$ for $|s|>b$; without loss of generality we may assume that $2b>a$.

As usual, we put $\hbar=2m=1$; then the one-particle Hamiltonian of such a waveguide is the Dirichlet Laplacian $-\Delta_{\mathrm{D}}^{\Sigma}$ defined in the conventional way—cf. Ref. 16, Sec. XIII.15. Using the natural locally orthogonal curvilinear coordinates s,u in Σ one can map $-\Delta_{\mathrm{D}}^{\Sigma}$ unitarily onto the operator,

$$H_1=-\partial_s(1+u\gamma)^{-2}\partial_s-\partial_u^2+V(s,u) \tag{2.1}$$

on $L^2(\mathbb{R}\times(-a,a))$ with the effective curvature–induced potential,

$$V(s,u):=-\frac{\gamma(s)^2}{4(1+u\gamma(s))^2}+\frac{u\gamma''(s)}{2(1+u\gamma(s))^3}-\frac{5}{4}\,\frac{u\gamma''(s)^2}{(1+u\gamma(s))^4}, \tag{2.2}$$

which is e.s.a. on the core $D(H)=\{\psi:\psi\in C^\infty,\ \ \psi(s,\pm a)=0,\ H\psi\in L^2\}$—cf. Refs. 1,7 for more details.

If the waveguide contains N particles, the state Hilbert space is $L^2((\Sigma))^N$; the Pauli principle will be taken into account later. We assume that each particle has the charge e; using the same ‘‘straightening’’ transformation we are then able to rewrite the Hamiltonian as

$$H_N\equiv H_N(\gamma,a,e)=\sum_{j=1}^{N}\{-\partial_{s_j}(1+u_j\gamma(s_j))^{-2}\partial_{s_j}-\partial_{u_j}^2+V(s_j,u_j)\}+e^2\sum_{1\le j<l\le N}|\mathbf{r}_j-\mathbf{r}_l|^{-1}, \tag{2.3}$$

with the domain $(\mathcal{H}^2(\mathbb{R})\otimes\mathcal{H}_0^2(-a,a))^N$, where $\mathbf{r}_j=\mathbf{r}_j(s_j,u_j)$ are the Cartesian coordinates of the Nth particle.

As we have said our main aim in this paper is to estimate the maximum number of particles that a curved waveguide with given γ,a can bind, i.e., to find conditions under which the discrete spectrum of H_N is empty. To this end, one has to determine first the bottom of the essential spectrum. In complete analogy with the usual HVZ theorem,[16] we find

$$\sigma_{\mathrm{ess}}(H_N)=\left[\mu_{N-1}+\left(\frac{\pi}{2a}\right)^2,\infty\right), \tag{2.4}$$

where $\mu_{N-1}:=\inf\sigma(H_{N-1})$. Obviously,

$$\inf\sigma_{\mathrm{ess}}(H_N)\le\mu_{N-k}+k\left(\frac{\pi}{2a}\right)^2$$

holds for $k=1,\dots,N-1$, so

$$\inf\sigma_{\mathrm{ess}}(H_N)\le N\left(\frac{\pi}{2a}\right)^2. \tag{2.5}$$

In a straight tube the two expressions equal each other, while for $\gamma\neq 0$ we have a sharp inequality because $\mu_1<(\pi/2a)^2$ holds in this case.

III. NEUTRAL FERMIONS

If the particles in question are neutral fermions, one can get a simple upper bound on the number of bound states using the one-particle Hamiltonian (2.1); it is sufficient to estimate the dimension of $\sigma_{\mathrm{disc}}(H_1)$ and to employ the Pauli principle. To this aim, one has to estimate H_1 from above by an operator with the transverse and longitudinal variables decoupled; its projections to transverse modes are then one-dimensional Schrödinger operators to which the modified Birman–Schwinger bound may be applied.[17–19] In Ref. 1 we used this argument in the situation where a is small so that only the lowest transverse mode and the leading term in (2.2) may be taken into account.

A modification to the more general case is straightforward. We introduce the function

$$\tilde{W}(s):=\frac{\gamma(s)^2}{4\delta_-^2}+\frac{a|\gamma''(s)|}{2\delta_-^3}+\frac{5a^2\gamma'(s)^2}{4\delta_-^4}, \tag{3.1}$$

where

$$\delta_\pm:=1\pm a\|\gamma\|_\infty, \tag{3.2}$$

which majorizes the effective potential, $V(s,u)\le\tilde{W}(s)$. Furthermore, we set

$$\tilde{W}_j(s):=\max\left\{0,\left(\frac{\pi}{2a}\right)^2(1-j^2)\right\}, \tag{3.3}$$

for $j=2,3,\dots$; in view of the assumptions (ii), (iii) only a finite number of them is different from zero.

Replacing V by $\tilde{W}$, and $(1+u\gamma)^{-2}$ by δ_+^{-2}, we get an estimating operator with separating variables, or, in other words, a family of shifted one-dimensional Schrödinger operators; we are looking for the number of their eigenvalues below $\inf\sigma_{\mathrm{ess}}(H_1)=(\pi/2a)^2$. The mentioned modification of the Birman–Schwinger bound is based on splitting the rank–one operator corresponding to the singularity of the resolvent kernel $(1/2\kappa)e^{-\kappa|s-s'|}$ at $\kappa=0$ and applying a Hilbert–Schmidt estimate to the rest. In analogy with Refs. 17–19 we employ this trick for the lowest–mode component of the estimating operator, while for the higher modes we use the full resolvent at the values $\kappa_j:=(\pi/2a)\sqrt{j^2-1}$. In this way we arrive at the following conclusion.

Proposition III.1: The number N of neutral particles of half-integer spin S that a curved quantum waveguide can bind satisfies the inequality

$$N\le(2S+1)\left\{1+\delta_+^2\frac{\int_{\mathbb{R}^2}\tilde{W}(s)|s-t|\tilde{W}(t)\,ds\,dt}{\int_{\mathbb{R}}\tilde{W}(s)\,ds}+\sum_{j=2}^{\infty}\frac{a\delta_+^2}{\pi\sqrt{j^2-1}}\int_{\mathbb{R}}\tilde{W}_j(s)\,ds\right\}. \tag{3.4}$$

Remarks III.2: (a) As we have said, the number of nonzero terms in the last sum is finite. More exactly, the index j runs up to the entire part of $\sqrt{1+(2a/\pi)^2\|\tilde{W}\|_\infty}$; hence if a is small enough this term is missing at all.

(b) The assumption (iv) is not needed here. It is sufficient, e.g., that the functions γ,γ', and $|\gamma''|^{1/2}$ decay as $|s|^{-1-\varepsilon}$ as $|s|\to\infty$.

IV. MAIN RESULT: *N* CHARGED PARTICLES

We have said in the Introduction that the present study is motivated mainly by the need to describe electrons in curved quantum wires. Unfortunately, the above simple estimate have no straightforward consequences for the situation when the particles are charged. While the electrostatic repulsion adds a positive term to the Hamiltonian (2.3), it may move at the same time the bottom of the essential spectrum since the energies of the bound ''clusters'' are, of course, sensitive to the interaction change.

We need therefore another approach that would allow us to take the repulsion term in (2.3) into account. An inspiration can be found in an analysis of atomic N-body Hamiltonians. To formulate the result we need some notation. Given a positive β we denote by $\{\lambda_m\}_{m=1}^\infty$ the ordered sequence of eigenvalues of a Dirichlet Laplacian at the rectangle,

$$R_\beta:=[-\tfrac{3}{2}\beta\delta_+,\tfrac{3}{2}\beta\delta_+]\times[-a,a], \tag{4.1}$$

and set

$$T_\beta(N):=\begin{cases} 2\sum_{m=1}^{n}\lambda_m & \dots\quad N=2n,\\ 2\sum_{m=1}^{n}\lambda_m+\lambda_{n+1} & \dots\quad N=2n+1.\end{cases} \tag{4.2}$$

We have in mind here electrons and assume that the spin is $\frac{1}{2}$, otherwise $T_\beta(N)$ has to be replaced by the sum of the first N eigenvalues of $2S+1$ identical copies of the Laplacian. Now we are able state our main result.

Theorem IV.1: Assume (i)–(iv). $\sigma_{\rm disc}(H_N(\gamma,a,e))=\emptyset$ for $N\geq 2$ if the condition

$$T_\beta(N)+\frac{e^2}{2\beta\sqrt{7}}N(N-1)\geq\|\tilde{W}\|_\infty N+\left(\frac{\pi}{2a}\right)^2 N+\frac{e^2}{18\beta\sqrt{2}} \tag{4.3}$$

is valid for some $\beta\geq\max\{2b,596e^{-2}\}$.

Proof: We use a variational argument that relies on a suitable decomposition of the configuration space. Consider a pair of smooth functions v,g from $\mathbb{R}_+$ to $[0,1]$ such that

$$v(t)=\begin{cases}0 & \dots\quad t\leq 1,\\ 1 & \dots\quad t\geq\frac{3}{2},\end{cases} \tag{4.4}$$

and

$$v(t)^2+g(t)^2=1. \tag{4.5}$$

Elements of the configuration space are (s,u) with $s=\{s_1,\dots,s_N\}$ and $u=\{u_1,\dots,u_N\}$. We denote $\|s\|_\infty:=\max\{s_1,\dots,s_N\}$ and employ the functions

$$s\mapsto v(\|s\|_\infty\beta^{-1}),\quad g(\|s\|_\infty\beta^{-1}),$$

where $\beta > 2b > a$ is a parameter to be specified later. By an abuse of notation, we use the symbols v, g again both for these functions and the corresponding operators of multiplication. It is straightforward to evaluate $([H_N, v]\psi, v\psi)$ and the analogous expression with v replaced by g for a vector $\psi \in D(H_N)$; in both cases it is only the longitudinal kinetic part in (2.3) that contributes. This yields the identity

$$(H_N\psi,\psi) = (H_N v\psi, v\psi) + (H_N g\psi, g\psi) + \sum_{j=1}^{N} \{\|(1+u_j\gamma_j)^{-1} v_j\psi\|^2 + \|(1+u_j\gamma_j)^{-1} g_j\psi\|^2\},$$

where we have used the shorthand $v_j := \partial v/\partial s_j$, $g_j := \partial g/\partial s_j$, and $\gamma_j := \gamma(s_j)$. Notice further that the factors $(1+u_j\gamma_j)^{-1}$ may be neglected, because $v_j g_j$ are nonzero only if $s_j \geqslant \beta > 2b$, in which case $\gamma_j = 0$. Furthermore, with the exception of the hyperplanes where two or more coordinates coincide (which is a zero measure set) the norm $\|s\|_\infty$ coincides with just one of the coordinates $s_1, \dots, s_n$, and therefore

$$\sum_{j=1}^{N} \{\|v_j\psi\|^2 + \|g_j\psi\|^2\} \leqslant \|\psi\|^2 \max_{1\leqslant j\leqslant N} \{\|v_j\|_\infty^2 + \|g_j\|_\infty^2\} \leqslant \beta^{-2} C_0 \|\psi\|^2, \tag{4.6}$$

where $C_0 := \|v'\|_\infty^2 + \|g'\|_\infty^2$. We arrive at the estimate

$$(H_N\psi,\psi) \geqslant L_1[v\psi] + L_1[g\psi] \tag{4.7}$$

with

$$L_1[\phi] := (H_N\phi,\phi) - \frac{C_0}{\beta^2}\|\phi\|^2_{\mathcal{N}_\beta}, \tag{4.8}$$

where the last index symbolizes the norm of the vector ϕ restricted to the subset $\mathcal{N}_\beta := \{s : \beta \leqslant \|s\|_\infty \leqslant 3\beta/2\}$ of the configuration space.

Next, one has to estimate separately the contributions from the inner and outer parts. Let us begin with the exterior. We introduce the following functions:

$$f_1(s) = v(2s_1\|s\|_\infty^{-1}),$$

$$f_j(s) = v(2s_j\|s\|_\infty^{-1}) \prod_{n=1}^{j-1} g(2s_n\|s\|_\infty^{-1}), \quad j = 2, \dots, N-1,$$

$$f_N(s) = \prod_{n=1}^{N-1} g(2s_n\|s\|_\infty^{-1}).$$

It is clear from the construction that

$$\sum_{j=1}^{N} f_j(s)^2 = 1. \tag{4.9}$$

Moreover, the functions

$$s_j \mapsto v(2s_j\|s\|_\infty), g(2s_j\|s\|_\infty)$$

have a nonzero derivative only if $|s_j| \geqslant \frac{1}{2}\|s\|_\infty^{-1}$. Hence, on the support of $s \mapsto v(\|s\|_\infty\beta^{-1})$ the derivative is nonzero if $|s_j| \geqslant \frac{1}{2}\beta > b$. In other words, the function $s \mapsto f_j(s)^2 v(\|s\|_\infty\beta^{-1})$ has a zero derivative in all the parts of the configuration space, where at least one of the electrons dwells in the curved part of the waveguide. Commuting the (longitudinal kinetic part of) H_N with f_j, we get in the same way as above the identity

$$L_1[v\psi]=\sum_{j=1}^{N}\{L_1[f_j v\psi]-\|(\nabla_s f_j)v\psi\|^2\}, \tag{4.10}$$

where $\nabla_s:=(\partial_{s_1},\dots,\partial_{s_1})$. Next, we need a pointwise upper bound on $\Sigma_{j=1}^N(\nabla_s f_j)^2$: denoting $\sigma_j:=2s_j\|s\|_\infty$, we can write

$$\sum_{j=1}^{N}|(\nabla_s f_j)(s)|^2=\frac{4}{\|s\|_\infty^2}\{v'(\sigma_1)^2+g'(\sigma_1)^2v(\sigma_2)^2+g(\sigma_1)^2v'(\sigma_2)^2+\cdots$$
$$+g'(\sigma_1)^2g(\sigma_2)^2\cdots g(\sigma_N)^2+\cdots+g(\sigma_1)^2\dots g(\sigma_{N-1})^2g'(\sigma_N)^2\},$$

which gives after a partial resummation,

$$=\frac{4}{\|s\|_\infty^2}\{v'(\sigma_1)^2+g'(\sigma_1)^2+g(\sigma_1)^2g'(\sigma_2)^2+\cdots+g(\sigma_1)^2\cdots g(\sigma_{N-1})^2g'(\sigma_N)^2\}$$
$$\leqslant\frac{4}{\|s\|_\infty^2}\left\{v'(\sigma_1)^2+\sum_{j=1}^{N}g'(\sigma_j)^2\right\}\leqslant\frac{4NC_0}{\|s\|_\infty^2};$$

recall that $C_0:=\|v'\|_\infty^2+\|g'\|_\infty^2$. Consequently,

$$L_1[v\psi]\geqslant\sum_{j=1}^{N}L_1[f_j v\psi]-4NC_0\|v\psi\|s\|_\infty^{-1}\|^2$$
$$=\sum_{j=1}^{N}\{L_1[f_j v\psi]-4NC_0\|f_j v\psi\|s\|_\infty^{-1}\|^2\}=\sum_{j=1}^{N}L_2[f_j v\psi], \tag{4.11}$$

where

$$L_2[\phi]:=L_1[\phi]-4NC_0\|\phi\|s\|_\infty^{-1}\|^2. \tag{4.12}$$

Hence, we have to find a lower bound to $L_2(\psi_j)$ with $\psi_j:=f_j v\psi)$. Since $s_j\geqslant\frac{1}{2}\|s\|_\infty\geqslant\frac{1}{2}\beta>b$ holds on the support of ψ_j, we have $V(s_j,u_j)=0$ there. This allows us to write

$$(H_N\psi_j,\psi_j)=(H_{N-1}\psi_j,\psi_j)+\|\partial_{s_j}\psi_j\|^2+\|\partial_{u_j}\psi_j\|^2+e^2\sum_{j\neq l=1}^{N}(|\mathbf{r}_j-\mathbf{r}_l|^{-1}\psi_j,\psi_j),$$

where H_{N-1} refers to the system with the jth electron excluded, and therefore

$$(H_N\psi_j,\psi_j)\geqslant\left(\mu_{N-1}+\left(\frac{\pi}{2a}\right)^2\right)\|\psi_j\|^2+e^2\sum_{j\neq l=1}^{N}(|\mathbf{r}_j-\mathbf{r}_l|^{-1}\psi_j,\psi_j).$$

Since $|\mathbf{r}_j-\mathbf{r}_l|\leqslant\sqrt{(s_j-s_l)^2+4a^2}\leqslant2\sqrt{\|s\|_\infty^2+a^2}$, we have

$$(H_N\psi_j,\psi_j)\geqslant\left(\mu_{N-1}+\left(\frac{\pi}{2a}\right)^2\right)\|\psi_j\|^2+\frac{e^2(N-1)}{2}((\|s\|^2+a^2)^{-1/2}\psi_j,\psi_j).$$

The sought lower bound then follows from (4.12) and (4.8):

$$L_2[\psi_j] \geqslant \left(\mu_{N-1}+\left(\frac{\pi}{2a}\right)^2\right)\|\psi_j\|^2 - 4NC_0\|\psi_j\|s\|_\infty^{-1}\|^2$$

$$-C_0\beta^{-2}\|\psi_j\|^2_{\mathcal{N}_\beta}+\frac{e^2(N-1)}{2}((\|s\|^2+a^2)^{-1/2}\psi_j,\psi_j);$$

recall that $\mathcal{N}_\beta := \{s:\ \ \beta \leqslant \|s\|_\infty \leqslant 3\beta/2\}$. The second and the third term at the rhs can be combined, using

$$4NC_0\|\psi_j\|s\|_\infty^{-1}\|^2 + C_0\beta^{-2}\|\psi_j\|^2_{\mathcal{N}_\beta} \leqslant (4N+1)C_0\|\psi_j\|s\|_\infty^{-1}\|^2.$$

Furthermore, $\|s\|_\infty \geqslant \beta > 2b > a$ yields $(\|s\|^2+a^2)^{1/2} \leqslant \sqrt{2}\|s\|_\infty$ and

$$L_2[\psi_j] \geqslant \left(\mu_{N-1}+\left(\frac{\pi}{2a}\right)^2\right)\|\psi_j\|^2 + \left(\frac{e^2(N-1)}{2\sqrt{2}} - \frac{C_0(4N+1)}{\beta}\right)\|\psi_j\|s\|_\infty^{-1}\|^2. \tag{4.13}$$

We are interested in the situation when the second term at the rhs is positive. This is achieved if

$$\frac{e^2(N-1)}{2\sqrt{2}} > \frac{C_0(4N+1)}{\beta},$$

which is ensured if we choose β in such a way that

$$\beta > \frac{18\sqrt{2}C_0}{e^2}; \tag{4.14}$$

recall that $N \geqslant 2$. Owing to the identity (4.11) we then have

$$L_1[v\psi] \geqslant \left(\mu_{N-1}+\left(\frac{\pi}{2a}\right)^2\right)\|v\psi\|^2, \tag{4.15}$$

which means in view of (2.4) that the external part of ψ does not contribute to the discrete spectrum.

Let us turn now to the inner part. The corresponding quadratic form in the decomposition (4.7) can be estimated with the help of (2.3) and (4.8) by

$$L_1[g\psi] \geqslant \delta_+^{-2}\|\nabla_s g\ \psi\|^2 + \|\nabla_u g\ \psi\|^2 + \sum_{j=1}^{N} (V(s_j,u_j)g\psi, g\psi)$$

$$+e^2 \sum_{1\leqslant < k \leqslant N} (\|\mathbf{r}_j - \mathbf{r}_k\|^{-1} g\psi, g\psi) - \frac{C_0}{\beta^2}\|g\psi\|^2; \tag{4.16}$$

recall that $\delta_+ := 1 + a\|\gamma\|_\infty$. Using the function $\tilde{W}$ defined by (3.1) we find $|V(s_j,u_j)| \leqslant \tilde{W}(s_j)$, so

$$\max\{V(s,u):(s,u)\in\mathbb{R}\times[-a,a]\} \leqslant \|\tilde{W}\|_\infty.$$

Consequently, the curvature–induced potential term can be estimated by

$$\sum_{j=1}^{N} (V(s_j,u_j)g\psi, g\psi) \leqslant \|\tilde{W}\|_\infty N\|g\psi\|^2.$$

Furthermore, on the support of g, we have

$$|\mathbf{r}_j-\mathbf{r}_k|\leqslant 2\sqrt{\|s\|_\infty^2+a^2}\leqslant\sqrt{3\beta^2+4a^2},$$

because $\|s\|_\infty\leqslant\frac{3}{2}\beta$ holds there. At the same time, $\beta>2b>a$, so we arrive at the estimate

$$|\mathbf{r}_j-\mathbf{r}_k|\leqslant\sqrt{7}\beta,$$

which yields

$$\sum_{1\leqslant<k\leqslant N}(\|\mathbf{r}_j-\mathbf{r}_k\|^{-1}g\psi,g\psi)\geqslant\frac{N(N-1)}{2\beta\sqrt{7}}\|g\psi\|^2.$$

Now we can combine the above estimates with the inequality $C_0/\beta<e^2/18\sqrt{2}$, which follows from (4.14) to get the bound

$$L_1[g\psi]\geqslant\delta_+^{-2}\|\nabla_s g\;\psi\|^2+\|\nabla_u g\;\psi\|^2+\left[-N\|\tilde{W}\|_\infty+\frac{e^2N(N-1)}{2\beta\sqrt{7}}-\frac{e^2}{18\beta\sqrt{2}}\right]\|g\psi\|^2. \quad (4.17)$$

Now we can put the above results together. In view of the inequality (4.15) and of (2.5), the last bound tells us that H_N has no discrete spectrum for $N\geqslant 2$, provided

$$\delta_+^{-2}\|\nabla_s g\;\psi\|^2+\|\nabla_u g\;\psi\|^2+\left[\frac{e^2N(N-1)}{2\beta\sqrt{7}}-\frac{e^2}{18\beta\sqrt{2}}-N\|\tilde{W}\|_\infty-N\left(\frac{\pi}{2a}\right)^2\right]\|g\psi\|^2\geqslant 0, \quad (4.18)$$

for some β, which satisfies the condition

$$\beta\geqslant\max\left\{2b,\frac{18\sqrt{2}C_0}{e^2}\right\}. \quad (4.19)$$

The first two terms in (4.18) are nothing else than the quadratic form of the $2N$-dimensional Laplacian on R_β^N—cf. (4.1). By the Pauli principle each eigenvalue may appear only twice, thus one has to take the orthogonal sum of two copies of the Laplacian on R_β and to sum the first N eigenvalues of such an operator. This is exactly the quantity that we have called $T_\beta(N)$.

To finish the proof, it remains to estimate C_0, which appears in the conditions (4.14) and (4.19). We will not attempt an optimal bound and, put simply,

$$v(\xi):=\sin(4\pi\xi^2(1-2\xi^2)),$$

for $t-1=:\xi\in(0,\frac{1}{2})$; then

$$v'(\xi)^2+g'(\xi)^2=(8\pi)^2\xi^2(1-4\xi^2)^2$$

has the maximum value $2\sqrt{2}(8\pi)^2/3\approx 595.5$. ■

V. CONCLUSIONS

Since the present study is rather a foray into an unchartered territory, the result is naturally far from optimal. Let us add a few remarks. First of all, it is clear that the overall size of the curved region affects substantially the number of particles that the waveguide can bind. We know that *any* curved tube has a one-particle bound state,[1,8] hence a tube with N slight bends that very far from each other (so far that the repulsion is much smaller that the gap between the bound state energy and the continuum) can certainly bind N particles for N arbitrarily large.

The method we use is borrowed from atomic physics, where it yields bounds on ionization of an atom. Of course, there are differences. The binding is due to the curved hard wall of the waveguide rather than by the electrostatic attraction to the nucleus, and the spectrum of our one-particle operator (2.1) is finite. Consequently, there is a maximum number of particles that a given curved tube can bind as long as the particles are fermions. Bosons can occupy naturally a single state, and the idea of a *Bose condensate* of neutral spin–zero atoms in a curved hollow optical fiber is rather appealing.

On the other hand, a nonzero particle charge changes the picture, and even the number of bosons bind by a curved tube is limited: notice that the condition (4.3) is satisfied for large enough N without respect to the Pauli-principle term $T_\beta(N)$. Of course, the fermionic nature reduces the maximum number N further, since $T_\beta(N)$ growth for large N is between $o(N^3)$ in the limit $a \to 0$ and $o(N^2)$ for $2b \sim a$. At the same time, the maximum number also depends on the value of the charge. Since $1/\sqrt{7} - 1/18\sqrt{2} > 0$ and the remaining terms in (4.3) are independent of e, we see that $\sigma_{\mathrm{disc}}(H_N) = \emptyset$ for any $N \geqslant 2$, provided e is large enough. Thus, our result confirms the natural expectation that for a given curved tube and sufficiently charged particles just one-particle bound states can survive.

We have not addressed in this paper the question about the minimum number of particles that a curved quantum waveguide can bind. The gap between the trivial result that follows from the one-particle theory[1,7,8] and the condition (4.3) leaves a lot of space for improvements. Moreover, it is a natural question whether strongly curved tubes that can bind many particles allow for some semiclassical description analogous to the case of the quantum dots.[6] This is a task for a future work.

ACKNOWLEDGMENTS

The research has been partially supported by the Grants No. 1048801, GAAS, and ME099, Ministry of Education of the Czech Republic.

[1] P. Duclos and P. Exner, ''Curvature-induced bound states in quantum waveguides in two and three dimensions,'' Rev. Math. Phys. **7**, 73–102 (1995).
[2] C. M. Savage, S. Markensteiner, and P. Zoller, ''Atomic waveguides and cavities from hollow optical fibres,'' in *Fundamentals of Quantum Optics III*, edited by S. Eklotzky (Springer-Verlag, Berlin, 1993).
[3] V. Halonen, ''Electronic states and addition spectra of parabolic quantum dots in a magnetic field,'' Solid State Commun. **92**, 703–705 (1994).
[4] M. El-Said, ''Energy states of two electrons in a parabolic quantum dot in a magnetic field,'' J. Phys. I **5**, 1027–1036 (1995).
[5] F. M. Peeters and V. A. Schweigert, ''Two-electron quantum disks,'' Phys. Rev. B **53**, 1468–1474 (1996).
[6] J. Yngvason, ''Asymptotics of natural and artificial atoms in strong magnetic fields,'' in *Proceedings of the XIth Congress of Mathematical Physics*, edited by D. Iagolnitzer (International Press, Boston, 1995), pp. 185–205.
[7] P. Exner and P. Šeba, ''Bound states in curved quantum waveguides'' J. Math. Phys. **30**, 2574–2580 (1989).
[8] J. Goldstone and R. L. Jaffe, ''Bound states in twisting tubes,'' Phys. Rev. B **45**, 14 100–14 107 (1992).
[9] M. Andrews and C. M. Savage, ''Bound states in two-dimensional nonuniform waveguides,'' Phys. Rev. A **50**, 4535–4537 (1994).
[10] W. Bulla, F. Gesztesy, W. Renger, and B. Simon, ''Weakly coupled bound states in quantum waveguides,'' Proc. Am. Math. Soc. **125**, 1487–1495 (1997).
[11] P. Exner and S. A. Vugalter, ''Bound states in a locally deformed waveguide: the critical case,'' Lett. Math. Phys. **39**, 57–69 (1997).
[12] P. Exner, P. Šeba, M. Tater, and D. Vaněk, ''Bound states and scattering in quantum waveguides coupled laterally through a boundary window,'' J. Math. Phys. **37**, 4867–4887 (1996).
[13] P. Exner and S. A. Vugalter, ''Asymptotic estimates for bound states in quantum waveguides coupled laterally through a narrow window,'' Ann. Inst. Henri Poincaré Phys. Theor. **65**, 109–123 (1996).
[14] P. Exner and S. A. Vugalter, ''Bound-state asymptotic estimates for window-coupled Dirichlet strips and layers,'' J. Phys. A **30**, 7863–7878 (1997).
[15] Y. Nogami, F. M. Toyama, and Y. P. Varshni, ''Stability of two-electron bound states in a model quantum wire,'' Phys. Lett. A **207**, 355–361 (1995).
[16] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, IV. Analysis of Operators* (Academic, New York, 1978).
[17] N. Seto, ''Bargmann's inequalities in spaces of arbitrary dimension,'' Publ. RIMS **9**, 429–461 (1974).
[18] M. Klaus, ''On the bound state of Schrödinger operators in one dimension,'' Ann. Phys. (Leipzig) **108**, 288–300 (1977).
[19] R. G. Newton, ''Bounds for the number of bound states for Schrödinger equation in one and two dimensions,'' J. Operator Theory **10**, 119–125 (1983).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

A general BRST approach to string theories with zeta function regularizations

Stephen Hwang[a)]
Karlstad University, S-65188 Karlstad, Sweden

Robert Marnelius[b)]
ITP, Chalmers University of Technology, Göteborg University, S-412 96 Göteborg, Sweden

Panagiotis Saltsidis[c)]
DAMTP, University of Cambridge, Silver St, Cambridge CB3 9EW, United Kingdom

(Received 30 March 1999; accepted for publication 22 April 1999)

We propose a new general BRST approach to string and string-like theories that have a wider range of applicability than, e.g., the conventional conformal field theory method. The method involves a simple general regularization of all basic commutators, which makes all divergent sums expressible in terms of zeta functions from which finite values then may be extracted in a rigorous manner. The method is particularly useful in order to investigate possible state space representations to a given model. The method is applied to three string models: The ordinary bosonic string, the tensionless string, and the conformal tensionless string. We also investigate different state spaces for these models. The tensionless string models are treated in detail. Although we mostly rederive known results, they appear in a new fashion that deepens our understanding of these models. Furthermore, we believe that our treatment is more rigorous than most of the previous ones. In the case of the conformal tensionless string we find a new solution for $d=4$. © *1999 American Institute of Physics.* [S0022-2488(99)02010-1]

I. INTRODUCTION AND PRESENTATION OF THE METHOD

The usual operator formulation of BRST quantization of string theories are based on the following ingredients: First, specify a BRST-invariant vacuum state and then normal order the BRST operator in order to finally check in which dimension the BRST operator is nilpotent.[1] Even the conformal field theory method is based on these ingredients.[2] These methods have been very successful when applied to conventional string theories. However, there are models to which these methods are not applicable. We have, e.g., the tensionless string models that do not have a conventional vacuum state.[3,4] In this paper we present a new rigorous and very general approach to the operator version of BRST quantization, which not only makes it possible to consistently treat the conventional models, but also the tensionless string models. The method is particularly useful in order to investigate possible state space representations to a given model. The method seems, furthermore, to cast new light on the BRST method as a whole.

What we advocate here is the following general procedure in the BRST quantization.

(1) Construct a Hermitian and formally nilpotent BRST operator Q.

(2) Find a state space such that the properties above are true as operator equations in this space and for which the equation $Q|\phi\rangle=0$ has nontrivial solutions.

(3) Use a general precise regularization of the basic commutators in the above analysis.

The idea to start with a nilpotent BRST operator and then look for possible solutions is, of

[a)]Electronic mail: Stephen.Hwang@hks.se
[b)]Electronic mail: tferm@fy.chalmers.se
[c)]Electronic mail: P.Saltsidis@damtp.cam.ac.uk

course, very natural. However, usually one prescribes the setting first, which is what mathematicians would tell us to do. Usually the state space governs how we construct the BRST charge. But here this is not required. We only require a formally nilpotent BRST charge to start with. Then we look for possible state spaces. This freedom seems quite large. However, it is what a physicist likes to play with since physical models should also allow for our physical intuition to act. The structure of possible solutions of such a general prescription for finite degrees of freedom was given in Ref. 5 and further developed in Ref. 6. One may note that a Hermitian, nilpotent BRST charge may be constructed for a very large class of models in terms of a power expansion in $\hbar$ and in the ghost fields.[7] A solution is naturally obtained in a Weyl ordered form. However, these properties are only formal, since we must also find a state space in which the BRST charge makes sense. In order to do this in the case of infinite degrees of freedom we must make use of a regularization procedure. For instance, in the string models we encounter divergent sums. Such sums must be regularized in some way. Here we consider a simple general regularization of the basic commutators, a regularization that will make all infinite sums expressible as zeta functions. This regularization, which is presented in Sec. II, is therefore such that when the regulator is removed it will give finite results through analytic continuation in all cases considered. This makes it possible to rigorously compute all operator equations. Furthermore, and which is important, it allows us to investigate possible state spaces both rigorously and efficiently. We emphasize that the zeta function regularization here is used in a much more general and precise form than what one usually finds since here it appears from one single regularization of the basic commutators. (For applications of zeta regularization and for literature on the subject, see, e.g., Ref. 8.)

Instead of giving a detailed description of how the method is supposed to be applied, we treat three string models in detail: the closed ordinary bosonic string, the closed tensionless string, and the closed conformal tensionless string. The closed ordinary bosonic string is mainly treated for pedagogical reasons. However, it should be interesting to see how we investigate alternative state space representations. Our main interest is in the tensionless strings. Here we give a rather exhaustive analysis from which we are able to give precise results for all proposed versions.

Critical dimensions in string theories appear in connection with a particular state space. In our method they appear through an inconsistency: The formally nilpotent BRST charge is *not* nilpotent or is not appropriate on the considered state space. In our investigation of three string models we first look for a BRST-invariant vacuum state. Then we check whether or not the formally nilpotent BRST charge is nilpotent on this vacuum state. The conventional closed bosonic string turns out to be the most intricate example here. The conventional vacuum state is investigated in Sec. III. The appropriate BRST charge turns out to be the formally nilpotent charge shifted by a regulator-dependent term [Eq. (3.22)]. This charge is, however, only nilpotent at $d=26$. Thus, the standard results are obtained but in a different way than usually. This treatment demonstrates how our method works and maybe it also deepens our understanding of this well-known model. The second model we treat is the tensionless string.[9–11] In Sec. IV we recover the known result of no critical dimension in Ref. 10 for one vacuum state, and in Sec. V we find the critical dimension 26, as was found in Ref. 11 for another vacuum state. Our final example is the conformal tensionless string. In Sec. VI we first investigate the state space considered in Ref. 4. Here we find that the BRST charge is nilpotent in any dimension. However, the vacuum state is only BRST invariant in two dimensions and we have not found any BRST-invariant states in other dimensions. In Sec. VII we also consider another state space, which to our knowledge has not been considered previously. In this state space the BRST charge is again nilpotent in arbitrary dimensions, but the vacuum state is only BRST invariant in $d=4$. For other values of d we have not found any BRST-invariant states in this state space. In Sec. VII we also investigate the alternative vacuum state considered in Ref. 3, in which case we find that the BRST charge is not nilpotent in any dimensions and, hence, that there is no consistent BRST treatment at all, which is in agreement with the result of Ref. 3. In Secs. VIII and IX we investigate alternative vacua also for the ordinary bosonic string. Note that unlike some previous calculations our regularization makes all results finite, which makes us believe that the treatment given here is, in general, more rigorous than previous ones.

Normally a consistent BRST quantization requires BRST-invariant states with positive norms.

In fact, such a condition should really be inserted into condition 2 above. However, this would then exclude all models we treat, except possibly the ordinary bosonic string. The reason is that all our calculations are performed in the minimal sector, which is the sector with no dynamical Lagrange multipliers and antighosts. An artifact of this sector is that one cannot, in general, work on a truly inner product space. Consistent BRST-invariant states $|\phi\rangle$ requires here the existence of a dual BRST-invariant state $|\overline{\phi}\rangle$, satisfying the condition[5]

$$\langle\overline{\phi}|\phi\rangle \text{ is finite and} \neq 0. \tag{1.1}$$

This condition restricts the possible solutions. To work on truly inner product spaces one has to consider a BRST quantization in the complete sector with dynamical Lagrange multipliers and antighosts.[12] However, inconsistent solutions in the minimal sector will remain inconsistent in the complete sector. Furthermore, solutions containing BRST-invariant negative norm states in the minimal sector will retain these in the complete sector. In Sec. IX we present a state space representation for the conventional bosonic string model, which yields a finite set of BRST-invariant states, which, however, are shown to contain negative norm states. The same is true for the special solutions of the tensionless string in Sec. V. Both these options have therefore to be excluded. The consistency of the tensionless string model in Sec. IV seems unclear, since it has a vacuum solution that is not associated with any oscillators. The BRST-invariant states are not inner product states at all in the minimal sector. Although we expect that there exists a positive normed solution in the complete sector, this remains to be investigated.

It is, of course, also possible to investigate the fermionic extensions of the models considered here.[11,13] We expect such calculations to be quite straightforward using the method we present here. One may also consider other theories like, e.g., brane theories[14] and conventional field theories. We believe that also here our method should be important for rigorous results.

II. OUR REGULARIZATION

In string theory the string coordinates $X^{\mu}(\tau,\sigma)$ and the corresponding conjugate momenta $P_{\mu}(\tau,\sigma)$ satisfy the basic nonzero equal-time commutator (we suppress τ in the following):

$$[X^{\mu}(\sigma),P_{\nu}(\sigma')]=i\delta^{\mu}_{\nu}\delta(\sigma-\sigma'). \tag{2.1}$$

Since we shall only consider closed strings, we let $X^{\mu}(\sigma)$ and $P_{\mu}(\sigma)$ be periodic functions with period π. Thus, we may set ($\Sigma_n \equiv \Sigma_{n=-\infty}^{\infty}$ in the following)

$$X^{\mu}(\sigma)=\frac{1}{\sqrt{\pi}}\sum_n x^{\mu}_n e^{-2in\sigma}, \quad P^{\mu}(\sigma)=\frac{1}{\sqrt{\pi}}\sum_n p^{\mu}_n e^{-2in\sigma}, \tag{2.2}$$

and replace (2.1) by

$$[x^{\mu}_n,p_{m\nu}]=i\delta^{\mu}_{\nu}\delta^{0}_{n+m}. \tag{2.3}$$

The delta function in (2.1) is then the periodic delta function,

$$\delta(\sigma-\sigma')=\frac{1}{\pi}\sum_n e^{2in(\sigma-\sigma')}. \tag{2.4}$$

It is the appearance of a delta function in the commutator (2.1) that causes infinities in the quantum string theory. In order to handle these infinities in a well-defined manner we have to regularize the delta function, i.e., we have to make it an ordinary well-defined function. We have therefore to consider a modified or regularized commutator. We choose it to be of the form

$$[x^{\mu}_n,p_{m\nu}](s)=i\delta^{\mu}_{\nu}\delta^{0}_{n+m}f(|n|,s), \tag{2.5}$$

where $f(|n|,s)$ is a real function that satisfies the condition

$$f(|n|,0)=1. \tag{2.6}$$

It can only depend on the absolute value of n in order for x_n and p_n to retain their hermiticity properties from (2.2), even in the regularized case, i.e., $(x_n)^\dagger = x_{-n}$ and $(p_n)^\dagger = p_{-n}$. The choice of regulator function $f(|n|,s)$ is dictated by two considerations. First, in the computations that will follow, e.g., a calculation of the BRST nilpotency, the regulator should lead to a finite result as the regulator is removed $(s\to 0)$. Second, we would like our regulator to be as general as possible. A precise choice of $f(|n|,s)$ that satisfies these criteria is

$$f^{(\alpha)}(|n|,s)=\begin{cases}(|n|+\alpha)^{-s}, & |n|\geqslant -A+1,\\ 1, & |n|\leqslant -A,\end{cases} \tag{2.7}$$

where α is a real constant, which may be chosen to have any value, and where A is the closest integer to α satisfying $A\geqslant\alpha$. This choice will make all infinite sums expressible in terms of zeta functions (see below). This in turn will allow us to rigorously treat all infinite sums. Other regularizations are, of course, possible to use and should yield equivalent results. However, we think that the choice (2.7) is the most general one.

In the following BRST treatment of the various string models, we have also the fermionic ghost variable $c^I(\sigma)$ and $b^I(\sigma)$. Their basic commutators must be regularized in exactly the same way as in (2.5). For ghosts c_n^I, b_n^I (where I labels different types of ghosts) we have

$$[b_m^I, c_n^J](s)=\delta^0_{m+n}\delta^{IJ} f^{(\alpha)}(|n|,s). \tag{2.8}$$

The same is true for all other canonical variables one introduces. (We use graded commutators throughout.)

When calculating commutators we often get infinite sums of the form $\Sigma_n f^{(\alpha)}(|n|,s)$ that converge for $s>1$. However, by analytic continuations they may yield a finite value for $s<1$ and, in particular, for $s=0$. The choice (2.7) makes these sums expressible in terms of zeta functions. We have

$$\sum_n f^{(\alpha)}(|n|,s)=\begin{cases}1-2A+2\zeta(s|1+\alpha-A), & \alpha\leqslant 0,\\ -\alpha^{-s}-2\displaystyle\sum_{n=0}^{A-2}(n+1+\alpha-A)^{-s}+2\zeta(s|1+\alpha-A), & \alpha>0,\end{cases} \tag{2.9}$$

where $\zeta(s|a)$ is the Hurwitz zeta function defined for $0<a\leqslant 1$, $s>1$ by[15]

$$\zeta(s|a)\equiv\sum_{n=0}^{\infty}\frac{1}{(n+a)^s}. \tag{2.10}$$

By analytic continuation one finds the following value at $s=0$ $(0<a\leqslant 1)$:[15]

$$\zeta(0|a)=\tfrac{1}{2}-a, \tag{2.11}$$

which implies that [notice that the finite sums for $\alpha\leqslant 0$ and $\alpha>0$ are equal for $s=0$ in (2.9)]

$$\sum_n f^{(\alpha)}(|n|,0)=-2\alpha. \tag{2.12}$$

We have similarly, e.g. (we set $s=0$ for the finite sums),

$$\sum_{n=1}^{\infty} n f^{(\alpha)}(|n|,s) = \zeta(s-1|1+\alpha-A) - \alpha\zeta(s|1+\alpha-A) + \frac{1}{2}A(A-1),$$
$$\sum_{n=1}^{\infty} n^2 f^{(\alpha)}(|n|,s) = \zeta(s-2|1+\alpha-A) - 2\alpha\zeta(s-1|1+\alpha-A) + \alpha^2\zeta(s|1+\alpha-A) - \frac{1}{6}A(A-1)(2A-1), \tag{2.13}$$

which implies that

$$\lim_{s\to 0}\sum_{n=1}^{\infty} n f^{(\alpha)}(|n|,s) = \frac{1}{2}\alpha^2 - \frac{1}{12}, \quad \lim_{s\to 0}\sum_{n=1}^{\infty} n^2 f^{(\alpha)}(|n|,s) = -\frac{1}{3}\alpha^3, \tag{2.14}$$

since $\zeta(-1|a) = (6a-6a^2-1)/12$ and $\zeta(-2|a) = a(3a-2a^2-1)/6$.

The regularization (2.5) and (2.7) correspond by (2.1) to the following regularized delta function:

$$\delta_s(\sigma-\sigma') = \frac{1}{\pi}\left\{\sum_n f^{(\alpha)}(|n|,s)e^{2in(\sigma-\sigma')}\right\}, \tag{2.15}$$

which is a well-defined function for $s>1$ and by Eq. (2.12) gives a regularization such that $\lim_{s\to 0}\delta_s(0) = -2\alpha/\pi$. Note here that depending on the value of α we can get any value in this limit. This is one argument supporting our belief that our regularization gives the most general result possible. Notice also that $\lim_{s\to 0}\delta_s^{(k)}(0)$ is finite for any order of derivative k.

III. THE BOSONIC STRING

As an illustration of our method we first treat the ordinary bosonic string. The bosonic string is classically characterized by the constraints

$$\frac{1}{4T}(P+TX')^2(\sigma) = 0, \quad \frac{1}{4T}(P-TX')^2(\sigma) = 0, \tag{3.1}$$

where T is the string tension. The Fourier modes of the corresponding Hermitian constraint operators are

$$L_n \equiv \frac{1}{2}\sum_k \alpha_{n-k}\cdot\alpha_k, \quad K_n \equiv \frac{1}{2}\sum_k \tilde{\alpha}_{n-k}\cdot\tilde{\alpha}_k, \tag{3.2}$$

where

$$\alpha_n^\mu \equiv \left(\frac{1}{2\sqrt{T}}p_n^\mu - i\sqrt{T}nx_n^\mu\right), \quad \tilde{\alpha}_n^\mu \equiv \left(\frac{1}{2\sqrt{T}}p_{-n}^\mu - i\sqrt{T}nx_{-n}^\mu\right), \tag{3.3}$$

in terms of the x_n and p_n modes in (2.2). The regularized commutator (2.5) with the regularized function (2.7) implies that

$$[\alpha_n^\mu,\alpha_m^\nu](s)=n\,\eta^{\mu\nu}f^{(\alpha)}(|n|,s)\,\delta_{n+m}^0,\quad [\alpha_n^\mu,\tilde{\alpha}_m^\nu](s)=0,$$
$$[\tilde{\alpha}_n^\mu,\tilde{\alpha}_m^\nu](s)=n\,\eta^{\mu\nu}f^{(\alpha)}(|n|,s)\,\delta_{n+m}^0, \tag{3.4}$$

where $\eta^{\mu\nu}$ is a Minkowski metric, diag$(\eta^{\mu\nu})=(-1,+1,+1,+1)$. We consider now a corresponding BRST theory in which the manifestly Hermitian BRST charge is given by

$$Q=\sum_k (L_{-k}c_k^L+K_{-k}c_k^K)+\frac{1}{4}\sum_{k,l}(k-l)$$
$$\times(b_{k+l}^Lc_{-l}^Lc_{-k}^L+b_{k+l}^Kc_{-l}^Kc_{-k}^K-c_{-k}^Lc_{-l}^Lb_{k+l}^L-c_{-k}^Kc_{-l}^Kb_{k+l}^K), \tag{3.5}$$

where c_k^L,b_k^L and c_k^K,b_k^K are fermionic ghost modes satisfying (2.8). In the $s\to 0$ limit this BRST charge is formally nilpotent. A consequence of this nilpotency is that the extended constraints, $[Q,b_k^I]$, satisfy a closed algebra without any central extensions. The extended constraints to L_n and K_n are

$$\tilde{L}_n\equiv[Q,b_n^L]=L_n-\frac{1}{2}\sum_k (k+n)(c_k^Lb_{n-k}^L-b_{n-k}^Lc_k^L), \tag{3.6}$$

$$\tilde{K}_n\equiv[Q,b_n^K]=K_n-\frac{1}{2}\sum_k (k+n)(c_k^Kb_{n-k}^K-b_{n-k}^Kc_k^K). \tag{3.7}$$

For nonzero s we have, e.g., the following crucial commutator:

$$[\tilde{L}_m,\tilde{L}_{-m}](s)=\frac{1}{2}m\sum_k (\alpha_{-k}\cdot\alpha_k+k(c_{-k}^Lb_k^L-b_k^Lc_{-k}^L))(f^{(\alpha)}(|m+k|,s)+f^{(\alpha)}(|m-k|,s))$$
$$+\frac{1}{2}\sum_k (k\alpha_{-k}\cdot\alpha_k+(2m^2-k^2)(b_k^Lc_{-k}^L-c_{-k}^Lb_k^L))(f^{(\alpha)}(|m+k|,s)$$
$$-f^{(\alpha)}(|m-k|,s)), \tag{3.8}$$

which in the limit $s\to 0$ becomes

$$[\tilde{L}_m,\tilde{L}_{-m}](0)=2m\tilde{L}_0, \tag{3.9}$$

which is consistent with the fact that the BRST charge (3.5) is nilpotent in the $s\to 0$ limit. Now the crucial point is that the $s\to 0$ limit has no meaning before we specify on which state space the operators act. Below we show that the conventional choice of state space imply the expected result that we have a nilpotent BRST charge only in spacetime dimensions $d=26$ in the $s\to 0$ limit.

The standard choice of a vacuum state, $|0\rangle$, satisfies

$$\alpha_m|0\rangle=\tilde{\alpha}_m|0\rangle=0,\quad \forall m>0. \tag{3.10}$$

In order for this vacuum to be BRST invariant, $Q|0\rangle=0$, we must require the consistency conditions

$$[Q,\alpha_m^\mu]|0\rangle=0\Rightarrow mf^{(\alpha)}(|m|,s)\sum_{k=m}^{\infty} c_k^L\alpha_{m-k}^\mu|0\rangle=0\Rightarrow c_m^L|0\rangle=0,\quad \forall m>0,$$
$$[Q,\tilde{\alpha}_m^\mu]|0\rangle=0\Rightarrow mf^{(\alpha)}(|m|,s)\sum_{k=m}^{\infty} c_k^K\tilde{\alpha}_{m+k}^\mu|0\rangle=0\Rightarrow c_m^K|0\rangle=0,\quad \forall m>0. \tag{3.11}$$

These conditions in turn allow for the additional conditions

$$b_m^L|0\rangle = b_m^K|0\rangle = 0, \quad \forall m \geqslant 0, \tag{3.12}$$

for which the consistency conditions are

$$\begin{aligned} [Q, b_m^L]|0\rangle &\equiv \tilde{L}_m|0\rangle = 0, \quad \forall m \geqslant 0, \\ [Q, b_m^K]|0\rangle &\equiv \tilde{K}_m|0\rangle = 0, \quad \forall m \geqslant 0. \end{aligned} \tag{3.13}$$

They are satisfied for $m > 0$, but for $m = 0$ the situation is unclear. For nonzero s we have

$$\tilde{L}_0(s)|0\rangle = \frac{1}{2}\left(\alpha_0^2 + (d-2)\sum_{k=1}^{\infty} k f^{(\alpha)}(|k|,s)\right)|0\rangle, \tag{3.14}$$

which in the $s \to 0$ limit leads to the following finite expression:

$$\tilde{L}_0|0\rangle = \frac{1}{2}\left(\alpha_0^2 + (d-2)\left(\frac{1}{2}\alpha^2 - \frac{1}{12}\right)\right)|0\rangle, \tag{3.15}$$

where we have made use of the relation (2.14). The conditions (3.13) with the property (3.15) and a similar one for $\tilde{K}_n$ require the vacuum state to be an eigenstate to the momentum operator p_0^μ, with an eigenvalue that depends on the parameter α in the regularization function (2.7). [Notice that (3.3) implies that $\alpha_0^2 = \tilde{\alpha}_0^2 = p_0^2/4T$.] This is an unsatisfactory result. It means that the conventional vacuum state is not BRST invariant under the formally nilpotent BRST charge above. In fact, Q is not even nilpotent on the conventional vacuum state.

That the BRST operator (3.5) is *not* nilpotent on the above vacuum state may be seen by calculating the commutator $[\tilde{L}_m, \tilde{L}_{-m}]$ on the vacuum state for nonzero s. We find, from Eq. (3.8), that

$$\begin{aligned} [\tilde{L}_m, \tilde{L}_{-m}](s)|0\rangle = \Bigg(& m f^{(\alpha)}(|m|,s)\alpha_0^2 + \frac{1}{2} d \sum_{k=1}^{m} k(m-k) f^{(\alpha)}(|k|,s) f^{(\alpha)}(|k-m|,s) \\ & - \sum_{k=1}^{m} (2m-k)(k+m) f^{(\alpha)}(|k|,s) f^{(\alpha)}(|k-m|,s)\Bigg)|0\rangle. \end{aligned} \tag{3.16}$$

In the $s \to 0$ limit the right-hand side becomes

$$\left(m\alpha_0^2 + \frac{1}{12}(d-26)m^3 - \frac{1}{12}(d-2)m\right)|0\rangle, \tag{3.17}$$

which only is zero if $d = 26$ and if the eigenvalue of α_0^2 is 2, which is a regulator-independent condition. The reason for the different results from (3.9) and (3.15) is due to the fact that on the right-hand side of (3.8) the factor $f^{(\alpha)}(|m+k|,s) - f^{(\alpha)}(|m-k|,s)$, which is zero in the $s \to 0$ limit, is multiplied by an operator that is infinite on the above vacuum state. This means that $\tilde{L}_0$ and $\tilde{K}_0$ are not zero on the vacuum state $|0\rangle$ when (3.17) is zero. The solution of this dilemma is found when we rewrite the right-hand side of the commutator (3.8) by means of the regularized commutators (3.4) as follows:

$$[\tilde{L}_m,\tilde{L}_{-m}](s)=m\sum_{k=1}^{\infty}(\alpha_{-k}\cdot\alpha_k+kb^L_{-k}c^L_k+kc^L_{-k}b^L_k)(f^{(\alpha)}(|m+k|,s)+f^{(\alpha)}(|m-k|,s))$$
$$+\sum_{k=1}^{\infty}(k\alpha_{-k}\cdot\alpha_k+(k^2-2m^2)(b^L_{-k}c^L_k+c^L_{-k}b^L_k))(f^{(\alpha)}(|m+k|,s)-f^{(\alpha)}(|m-k|,s))$$
$$+mf^{(\alpha)}(|m|,s)\alpha_0^2+\frac{1}{2}d\sum_{k=1}^{m}k(m-k)f^{(\alpha)}(|k|,s)f^{(\alpha)}(|k-m|,s)$$
$$-\sum_{k=1}^{m}(2m-k)(k+m)f^{(\alpha)}(|k|,s)f^{(\alpha)}(|k-m|,s). \tag{3.18}$$

This expression implies that

$$[\tilde{L}_m,\tilde{L}_{-m}]=2m\left(\frac{1}{2}\alpha_0^2+\sum_{k=1}^{\infty}(\alpha_{-k}\cdot\alpha_k+kb^L_{-k}c^L_k+kc^L_{-k}b^L_k)\right)$$
$$+\tfrac{1}{12}(d-26)m^3-\tfrac{1}{12}(d-2)m, \tag{3.19}$$

in the $s\to 0$ limit, in agreement with (3.16) and (3.17). This may be rewritten as follows:

$$[\tilde{L}_m,\tilde{L}_{-m}]=2m\tilde{L}_0+\tfrac{1}{12}(d-26)m^3-\tfrac{1}{2}m(d-2)\alpha^2, \tag{3.20}$$

where

$$\tilde{L}_0=\frac{1}{2}\alpha_0^2+\sum_{k=1}^{\infty}(\alpha_{-k}\cdot\alpha_k+kb^L_{-k}c^L_k+kc^L_{-k}b^L_k)+\frac{1}{2}(d-2)\left(\frac{1}{2}\alpha^2-\frac{1}{12}\right) \tag{3.21}$$

is the original $\tilde{L}_0$ rewritten for finite s and taking the limit $s\to 0$. Equation (3.20) demonstrates the inconsistency with a nilpotent BRST charge obtained above and the reason why the conventional vacuum state is not BRST invariant. The remedy is obvious and expected: first, we notice that we may get a closed algebra at $d=26$ if we redefine the extended constraint $\tilde{L}_0$. This in turn may be accomplished by a redefinition of the original BRST charge: Simply replace Q in (3.5) by

$$Q'=Q-\frac{d-2}{4}\alpha^2(c_0^L+c_0^K), \tag{3.22}$$

which is not formally nilpotent for $d\neq 2$. The corresponding extended constraints are

$$\tilde{L}'_m=[Q',b^L_m]=\begin{cases}\tilde{L}_0-\dfrac{1}{4}(d-2)\alpha^2, & m=0,\\ \tilde{L}_m, & m\neq 0.\end{cases} \tag{3.23}$$

Note that $\tilde{L}'_0$ in distinction to $\tilde{L}_0$ in (3.21) is independent of the regulator parameter α. We have

$$\tilde{L}'_0=\frac{1}{2}\alpha_0^2+\sum_{k=1}^{\infty}(\alpha_{-k}\cdot\alpha_k+kb^L_{-k}c^L_k+kc^L_{-k}b^L_k)-\frac{(d-2)}{24}. \tag{3.24}$$

The spectrum is therefore regulator independent, as it should be. For $d=26$ we find now from (3.20),

$$[\tilde{L}'_m,\tilde{L}'_{-m}]=2m\tilde{L}'_0, \tag{3.25}$$

which is consistent with a nilpotent Q' for $d=26$, and that the conventional BRST vacuum is BRST invariant under Q'.

In Secs. VIII and IX we investigate other vacuum states for the bosonic string theory.

IV. THE BOSONIC TENSIONLESS STRING

The bosonic tensionless string (see Ref. 9) is characterized by the constraints

$$P^{\mu}(\sigma)P_{\mu}(\sigma)=0, \quad P^{\mu}(\sigma)X'_{\mu}(\sigma)=0. \tag{4.1}$$

These constraints follow from the bosonic string by dropping the term $T^2(X'(\sigma))^2$, which is assumed to be negligible in the $T\rightarrow 0$ limit. The Fourier modes of the corresponding Hermitian constraint operators are

$$\phi_n^{-1}\equiv\frac{1}{2}\sum_k p_k\cdot p_{n-k}, \quad \phi_n^L\equiv -i\frac{1}{2}\sum_k k(x_k\cdot p_{n-k}+p_{n-k}\cdot x_k). \tag{4.2}$$

Here a formally nilpotent BRST charge operator is given by

$$Q=\sum_k (\phi_{-k}^{-1}c_k^{-1}+\phi_{-k}^L c_k^L)-\frac{1}{2}\sum_{k,l}(k-l) \times\left(c_{-k}^{-1}c_{-l}^L b_{k+l}^{-1}+b_{k+l}^{-1}c_{-k}^{-1}c_{-l}^L+\frac{1}{2}c_{-k}^L c_{-l}^L b_{k+l}^L+\frac{1}{2}b_{k+l}^L c_{-k}^L c_{-l}^L\right). \tag{4.3}$$

We can check the nilpotency of Q by calculating the algebra of the extended constraints, given by

$$\tilde{\phi}_n^{-1}\equiv[Q,b_n^{-1}]=\phi_n^{-1}-\sum_k (n+k)c_k^L b_{n-k}^{-1},$$

$$\tilde{\phi}_n^L\equiv[Q,b_n^L]=\phi_n^L-\frac{1}{2}\sum_k (k+n)(c_k^{-1}b_{n-k}^{-1}-b_{n-k}^{-1}c_k^{-1}+c_k^L b_{n-k}^L-b_{n-k}^L c_k^L). \tag{4.4}$$

A straightforward calculation of the commutators yields

$$[\tilde{\phi}_m^{-1},\tilde{\phi}_n^L]=(m-n)\tilde{\phi}_{m+n}^{-1}, \quad [\tilde{\phi}_m^L,\tilde{\phi}_n^L]=(m-n)\tilde{\phi}_{m+n}^L. \tag{4.5}$$

Hence, we conclude that the BRST charge (4.3) is nilpotent.

We now look for a possible BRST-invariant vacuum state. Following Refs. 4 and 10, we consider a vacuum state defined by

$$p_n^{\mu}|0\rangle=b_n^{-1}|0\rangle=b_n^L|0\rangle=0, \quad \forall n. \tag{4.6}$$

[The crucial part is the first conditions. They may be viewed as the $T\rightarrow 0$ limit of (3.10) using (3.3).] In order for this vacuum state to be BRST invariant it has to satisfy the consistency conditions

$$[Q,p_n^{\mu}]|0\rangle=[Q,b_n^{-1}]|0\rangle=[Q,b_n^L]|0\rangle=0, \quad \forall n, \tag{4.7}$$

where

$$[Q,p_n^{\mu}]=-n\sum_k p_{n-k}^{\mu}c_k^L. \tag{4.8}$$

The first two conditions and the last one for $n\neq 0$ are easily seen to be satisfied due to (4.6). The only nontrivial condition is the last one for $n=0$. However, we have

$$\tilde{\phi}_0^L|0\rangle \equiv [Q,b_0^L]|0\rangle = \left(\phi_0^L - \frac{1}{2}\sum_k k[c_k^{-1}b_{-k}^{-1} - b_{-k}^{-1}c_k^{-1} + c_k^L b_{-k}^L - b_{-k}^L c_k^L]\right)|0\rangle$$

$$= -\frac{1}{2}\sum_k kp(i[p_{-k}^{\mu},x_{k\mu}] - [b_{-k}^{-1},c_k^{-1}] - [b_{-k}^L,c_k^L])|0\rangle$$

$$= -\frac{1}{2}(d-2)\sum_k kf^{(\alpha)}(|k|,s)|0\rangle = 0. \tag{4.9}$$

[It is zero for any choice of regulator function $f(|k|,s)!.$] It is then easily seen that

$$Q|0\rangle = 0, \tag{4.10}$$

which means that $|0\rangle$ defined by (4.6) is a BRST-invariant vacuum state for any dimension d. We must finally check that the BRST charge is nilpotent on the vacuum state. This may be accomplished by checking the algebra of the extended constraints. The most nontrivial one is $[\tilde{\phi}_m^L,\tilde{\phi}_{-m}^L]$, for which we find

$$[\tilde{\phi}_m^L,\tilde{\phi}_{-m}^L](s) = \frac{i}{2}\sum_k (k+m)kf^{(\alpha)}(|k+m|,s)(x_{-k}\cdot p_k - x_k\cdot p_{-k} + p_k\cdot x_{-k} - p_{-k}\cdot x_k)$$

$$+\frac{1}{2}\sum_k (k+2m)(k-m)f^{(\alpha)}(|k+m|,s)(b_{-k}^{-1}c_k^{-1} - b_k^{-1}c_{-k}^{-1}$$

$$+c_{-k}^{-1}b_k^{-1} - c_k^{-1}b_{-k}^{-1} + b_{-k}^L c_k^L - b_k^L c_{-k}^L + c_{-k}^L b_k^L - c_k^L b_{-k}^L). \tag{4.11}$$

A straightforward calculation yields that $[\tilde{\phi}_m^L,\tilde{\phi}_{-m}^L](0)|0\rangle$ is zero, which is consistent with Eq. (4.9). To conclude, within our regularization scheme we have shown that there exists a BRST-invariant vacuum state in any dimension. Furthermore, the BRST charge is nilpotent in the state space containing this vacuum state. This result is in agreement with the results of Ref. 10. Note that there exists a dual vacuum state $|\bar{0}\rangle$ satisfying $\langle\bar{0}|0\rangle$ finite and different from zero. $|\bar{0}\rangle$ satisfies $x_n^{\mu}|\bar{0}\rangle = c_n^{-1}|\bar{0}\rangle = c_n^L|\bar{0}\rangle = 0$ for all n together with their consistency conditions.

V. ALTERNATIVE QUANTIZATION OF THE BOSONIC TENSIONLESS STRING

Instead of a vacuum state satisfying (4.6) we follow Ref. 11 and consider

$$p_m^{\mu}|0\rangle = x_m^{\mu}|0\rangle = 0, \quad m>0. \tag{5.1}$$

The consistency conditions are

$$[Q,p_m^{\mu}]|0\rangle = [Q,x_m^{\mu}]|0\rangle = 0, \quad m>0, \tag{5.2}$$

and they require

$$c_m^{-1}|0\rangle = c_m^L|0\rangle = 0, \quad m>0, \tag{5.3}$$

for which

$$[Q,c_m^{-1}]|0\rangle = [Q,c_m^L]|0\rangle = 0, \quad m>0, \tag{5.4}$$

are automatically satisfied. This vacuum is then ghost fixed by

$$b_m^{-1}|0\rangle = b_m^L|0\rangle = 0, \quad m \geqslant 0. \tag{5.5}$$

The corresponding consistency conditions,

$$\tilde{\phi}_m^{-1}|0\rangle = \tilde{\phi}_m^L|0\rangle = 0, \quad m \geqslant 0, \tag{5.6}$$

are automatically satisfied for $m > 0$. For $m = 0$ we have

$$\tilde{\phi}_0^{-1}|0\rangle = 0 \Leftrightarrow p_0^2|0\rangle = 0. \tag{5.7}$$

For such a condition to be meaningful the vacuum state should be an eigenstate to p_0^μ, i.e., $p_0^\mu|p\rangle = p^\mu|p\rangle$. Equation (5.7) requires then the vacuum to be massless ($p^2 = 0$). For $m = 0$ we also have

$$\tilde{\phi}_0^L|p\rangle = (d-2)\left(\tfrac{1}{2}\alpha^2 - \tfrac{1}{12}\right)|p\rangle, \tag{5.8}$$

which satisfies (5.6) only for $d = 2$. For $s \neq 0$ we have

$$\tilde{\phi}_0^L(s) = i\sum_{k=1}^{\infty} k(x_{-k}\cdot p_k - p_{-k}\cdot x_k) + \sum_{k=1}^{\infty} k(b_{-k}^{-1}\cdot c_k^{-1} + c_{-k}^{-1}\cdot b_k^{-1} + b_{-k}^L\cdot c_k^L + c_{-k}^L\cdot b_k^L)$$
$$+ (d-2)\sum_{k=1}^{\infty} k f^{(\alpha)}(|k|,s). \tag{5.9}$$

From the commutator (4.11) we find

$$\lim_{s\to 0}[\tilde{\phi}_m^L, \tilde{\phi}_{-m}^L](s)|p\rangle = \tfrac{1}{6}\big((d-26)m^3 - (d-2)m\big)|p\rangle. \tag{5.10}$$

In fact, in the $s \to 0$ limit we have

$$[\tilde{\phi}_m^L, \tilde{\phi}_{-m}^L] = 2m\tilde{\phi}_0^L - \tfrac{1}{2}\alpha^2(d-2)m + \tfrac{1}{6}(d-26)m^3, \tag{5.11}$$

where $\tilde{\phi}_0^L$ on the right-hand side is the $s \to 0$ limit of (5.9). If we redefine Q by

$$Q' \equiv Q - \tfrac{1}{4}\alpha^2(d-2)c_0^L, \tag{5.12}$$

then the extended constraint operator $\tilde{\phi}_0^L$ is replaced by

$$\tilde{\phi}'^L_0 = \tilde{\phi}_0^L - \tfrac{1}{4}\alpha^2(d-2). \tag{5.13}$$

This operator together with $\tilde{\phi}_m^L$, $m \neq 0$ satisfy then an anomaly free algebra for $d = 26$, from which we conclude that Q' is nilpotent for $d = 26$. However, from (5.8) we have

$$\tilde{\phi}_0'^L|p\rangle = -\tfrac{1}{12}(d-2)|p\rangle. \tag{5.14}$$

It follows that we have no BRST-invariant vacuum of the above form for $d = 26$. However, there are BRST-invariant states in $d = 26$. In Ref. 16 it was shown that there are massless states with spin 0, 1 and 2. In Sec. IX we show that these states do not have positive norms. (From the results of Ref. 11 it seems that a BRST-invariant vacuum state only exists in the Ramond sector of the spinning tensionless string, in which case the critical dimension is 10.)

VI. THE BOSONIC CONFORMAL STRING

The conformal string is a tensionless string that is made manifestly conformally invariant.[3,4] By adding two extra dimensions, one timelike and one spacelike, one forms new coordinates that transform as $SO(d,2)$ vectors. By means of $SO(d,2)$-invariant constraints, one obtains then an $SO(d,2)$ conformally invariant formulation by construction. Let $X^M=(X^\mu,X^+,X^-)$ be the new coordinate vector, where the metric of the new coordinates is $\eta_{++}=\eta_{--}=0$, $\eta_{+-}=\eta_{-+}=1$. Classically the constraints are

$$\Phi^{-1}(\sigma)\equiv P^M(\sigma)P_M(\sigma)=0, \quad \Phi^0(\sigma)\equiv P^M(\sigma)X_M(\sigma)=0,$$
$$\Phi^1(\sigma)\equiv X^M(\sigma)X_M(\sigma)=0, \quad \Phi^L(\sigma)\equiv P^M(\sigma)X'_M(\sigma)=0, \tag{6.1}$$

and they reduce to the constraints (4.1) of the tensionless string by means of the gauge fixing conditions,

$$P^+(\sigma)=0, \quad X^+(\sigma)-1=0. \tag{6.2}$$

(A corresponding construction for particles were given in Ref. 17.) The Hermitian BRST charge operator is given by

$$Q=\sum_k (\phi^1_{-k}c^1_k+\phi^0_{-k}c^0_k+\phi^{-1}_{-k}c^{-1}_k+\phi^L_{-k}c^L_k)-\frac{1}{2}\sum_{k,l}\Big(2ic^1_{-k}c^{-1}_{-l}b^0_{k+l}+ic^1_{-k}c^0_{-l}b^1_{k+l}$$
$$-ic^{-1}_{-k}c^0_{-l}b^{-1}_{k+l}+(k+l)c^1_{-k}c^L_{-l}b^1_{k+l}+(k-l)c^{-1}_{-k}c^L_{-l}b^{-1}_{k+l}+kc^0_{-k}c^L_{-l}b^0_{k+l}$$
$$+\frac{1}{2}(k-l)c^L_{-k}c^L_{-l}b^L_{k+l}+\text{h.c.}\Big), \tag{6.3}$$

where h.c. are Hermitian conjugate terms and

$$\phi^{-1}_n\equiv\frac{1}{2}\sum_k p_k\cdot p_{n-k}, \quad \phi^0_n\equiv\frac{1}{4}\sum_k (x_k\cdot p_{n-k}+p_{n-k}\cdot x_k),$$
$$\phi^1_n\equiv\frac{1}{2}\sum_k x_k\cdot x_{n-k}, \quad \phi^L_n\equiv -i\frac{1}{2}\sum_k k(x_k\cdot p_{n-k}+p_{n-k}\cdot x_k), \tag{6.4}$$

which are the Fourier modes of the Hermitian operator constraints corresponding to (6.1). The BRST charge (6.3) is formally nilpotent and a consistent BRST quantization is possible if there exists a BRST-invariant vacuum state on which we have a nilpotent BRST operator. In order to investigate the existence of such a vacuum state, we need the extended constraint operators defined by

$$\tilde{\phi}^{-1}_n\equiv[Q,b^{-1}_n]=\phi^{-1}_n+\sum_k (2ic^1_kb^0_{n-k}+ic^0_kb^{-1}_{n-k}-(n+k)c^L_kb^{-1}_{n-k}),$$

$$\tilde{\phi}^0_n\equiv[Q,b^0_n]=\phi^0_n+\frac{1}{2}\sum_k (ic^1_kb^1_{n-k}-ib^1_{n-k}c^1_k-2nc^L_kb^0_{n-k}-ic^{-1}_kb^{-1}_{n-k}+ib^{-1}_{n-k}c^{-1}_k),$$
$$\tag{6.5}$$
$$\tilde{\phi}^1_n\equiv[Q,b^1_n]=\phi^1_n-\sum_k (2ic^{-1}_kb^0_{n-k}+ic^0_kb^1_{n-k}+(n-k)c^L_kb^1_{n-k}),$$

$$\tilde{\phi}_n^L \equiv [\mathcal{Q}, b_n^L] = \phi_n^L - \frac{1}{2}\sum_k ((k+n)(c_k^{-1}b_{n-k}^{-1} - b_{n-k}^{-1}c_k^{-1} + c_k^L b_{n-k}^L - b_{n-k}^L c_k^L)$$
$$+(k-n)(c_k^1 b_{n-k}^1 - b_{n-k}^1 c_k^1) + k(c_k^0 b_{n-k}^0 - b_{n-k}^0 c_k^0)]).$$

These operators are shown to satisfy the following commutator algebra:

$$[\tilde{\phi}_m^1, \tilde{\phi}_n^{-1}] = 2i\tilde{\phi}_{m+n}^0, \quad [\tilde{\phi}_m^L, \tilde{\phi}_n^L] = (m-n)\tilde{\phi}_{m+n}^L, \quad [\tilde{\phi}_m^0, \tilde{\phi}_n^L] = m\tilde{\phi}_{m+n}^0,$$
$$[\tilde{\phi}_m^1, \tilde{\phi}_n^0] = i\tilde{\phi}_{m+n}^1, \quad [\tilde{\phi}_m^{-1}, \tilde{\phi}_n^0] = -i\tilde{\phi}_{m+n}^{-1}, \tag{6.6}$$
$$[\tilde{\phi}_m^1, \tilde{\phi}_n^L] = (m+n)\tilde{\phi}_{m+n}^1, \quad [\tilde{\phi}_m^{-1}, \tilde{\phi}_n^L] = (m-n)\tilde{\phi}_{m+n}^{-1}.$$

These commutators are nonanomalous, as required by the formal nilpotence of Q.

Following Ref. 4, we consider now a vacuum state that satisfies the conditions

$$p_n^M|0\rangle = b_n^{-1}|0\rangle = c_n^1|0\rangle = 0, \quad \forall n. \tag{6.7}$$

These conditions are consistent with a BRST-invariant vacuum state, since

$$[Q, p_n^M]|0\rangle = \tilde{\phi}_n^{-1}|0\rangle = [Q, c_n^1]|0\rangle = 0, \quad \forall n, \tag{6.8}$$

are satisfied due to (6.7). Now these conditions do not specify a unique vacuum. We need further conditions for that. We may ghost fix the vacuum by the conditions

$$b_n^0|0\rangle = b_n^L|0\rangle = 0, \quad \forall n. \tag{6.9}$$

Their consistency conditions are

$$\tilde{\phi}_n^0|0\rangle = \tilde{\phi}_n^L|0\rangle = 0, \quad \forall n, \tag{6.10}$$

and are satisfied for $n \neq 0$ due to (6.7). By means of the regularization (2.7) we find, furthermore, that

$$\tilde{\phi}_0^L|0\rangle = 0, \tag{6.11}$$

for any s due to the symmetry properties we had in (4.9). Using (2.12) we have by a direct calculation,

$$\tilde{\phi}_0^0|0\rangle = \frac{i}{2}(d-2)\alpha|0\rangle, \tag{6.12}$$

in the limit $s \to 0$. Thus, $\tilde{\phi}_0^0$ has an imaginary continuous spectrum. The above vacuum state is therefore only BRST invariant for $d=2$. (There might exist BRST-invariant states built from the above vacuum state. However, we have not found any.)

If we define a new charge like in the bosonic string by

$$Q' = Q - \frac{i}{2}(d-2)\alpha c_0^0, \tag{6.13}$$

then the above vacuum state is BRST invariant under the new charge Q' for any dimension d. However, Q' is then neither Hermitian nor nilpotent for $d \neq 2$.

In fact, Q is nilpotent on $|0\rangle$ for any dimension d. This may be checked as in the previous models, by computing the algebra of the extended constraints acting on the vacuum state. The most nontrivial of these equations read as

$$[\tilde{\phi}_m^{-1},\tilde{\phi}_{-m}^1](s)|0\rangle=\frac{2-d}{2}\sum_k f^{(\alpha)}(|k|,s)f^{(\alpha)}(|k+m|,s)|0\rangle. \tag{6.14}$$

Here

$$\sum_k f^{(\alpha)}(|k|,s)f^{(\alpha)}(|k+m|,s)=\zeta\left(s,\alpha+\frac{1}{2}|m|,-\frac{m^2}{4}\right)+\zeta\left(s,\alpha-\frac{1}{2}|m|,-\frac{m^2}{4}\right)+g(m,s), \tag{6.15}$$

where we have introduced the zeta function,[8]

$$\zeta(s,a,b)\equiv\sum_k \frac{1}{[(k+a)^2+b]^s}, \tag{6.16}$$

and $g(m,s)$ involves only finite sums and $g(m,0)=-1$. Now since (we are indebted to Per Salomonson for this simple argument)

$$\left.\frac{d\zeta}{db}\right|_{s=0}=s\sum_k \left.\frac{1}{[(k+a)^2+b]^{s+1}}\right|_{s=0}=0, \tag{6.17}$$

we have

$$\zeta(0,a,b)=\zeta(0,a,0)=\zeta(0|a)=\tfrac{1}{2}-a, \tag{6.18}$$

which implies that

$$\sum_k f^{(\alpha)}(|k|,s)f^{(\alpha)}(|k+m|,s)\Big|_{s=0}=-2\alpha. \tag{6.19}$$

Hence

$$[\tilde{\phi}_m^{-1},\tilde{\phi}_{-m}^1]|0\rangle=(d-2)\alpha|0\rangle, \tag{6.20}$$

which is consistent with (6.6) and (6.12). Commutators of other constraint operators may similarly be shown to consistently act on the vacuum state, in accordance with Eq. (6.6). Thus, the nilpotency of the BRST operator holds as a true operator equation in the chosen state space. For $d\neq 2$ the considered vacuum state is not BRST invariant. Furthermore, we have not been able to find any BRST-invariant state in the state space. If there does not exist BRST-invariant states, the theory is nontrivial only in $d=2$.

Now instead of the conditions (6.9) we may also ghost fix the vacuum state by the conditions

$$c_n^0|0\rangle=b_n^L|0\rangle=0,\quad \forall n. \tag{6.21}$$

In this case all consistency conditions are satisfied, which means that this vacuum is BRST invariant under the original formally nilpotent BRST charge (6.3) for any dimension d. Hence, we have found two vacua: one that is BRST invariant only for $d=2$ and another that is BRST invariant for all dimensions. However, since we work in the minimal sector, we must make sure that there exists a dual vacuum state $|\overline{0}\rangle$. It is straightforward to show that $|\overline{0}\rangle$ exists for any dimensions in the first case above but only for $d=2$ in the second case. Thus, both solutions yield equivalent results, and we have found a BRST-invariant vacuum only for $d=2$, in agreement with the result of Ref. 4. That we have two solutions requires a selection condition. (What we have here is a noncanonical situation in the language of Ref. 5.)

VII. ALTERNATIVE QUANTIZATIONS OF THE BOSONIC CONFORMAL STRING

Here we will consider two alternative sets of state spaces to the one treated in the previous section.

A. A consistent solution at $d=4$

First we consider a state space with a vacuum state defined through the conditions

$$p_n^\mu|0\rangle=p_n^+|0\rangle=x_n^+|0\rangle=b_n^{-1}|0\rangle=c_n^1|0\rangle=b_n^0|0\rangle=b_n^L|0\rangle=0, \quad \forall n. \tag{7.1}$$

It is easily checked that this vacuum state satisfies

$$\tilde{\phi}_n^{-1}|0\rangle=\tilde{\phi}_n^L|0\rangle=0, \quad \forall m. \tag{7.2}$$

Furthermore,

$$\tilde{\phi}_0^0|0\rangle=\frac{i}{2}(d-4)\alpha|0\rangle, \tag{7.3}$$

in the limit $s\rightarrow 0$. By the same reasoning as in the previous section, this leads to the conclusion that the above vacuum state is only BRST invariant for $d=4$. We must also check the closure of the extended constraints when acting on the vacuum state. In precisely the same way that leads to Eq. (6.20), one finds in this case,

$$[\tilde{\phi}_m^{-1},\tilde{\phi}_{-m}^1]|0\rangle=(d-4)\alpha|0\rangle, \tag{7.4}$$

which is consistent with (6.6) and (7.3). Other commutators of constraint operators may similarly be shown to consistently act on the vacuum state. Concluding, we have a nonanomalous theory for any d, and the vacuum state is BRST invariant for $d=4$. We have not been able to find BRST-invariant states for other d. The vacuum dual to $|0\rangle$ is defined by

$$x_n^\mu|\overline{0}\rangle=x_n^-|\overline{0}\rangle=p_n^-|\overline{0}\rangle=c_n^{-1}|\overline{0}\rangle=b_n^1|\overline{0}\rangle=c_n^0|\overline{0}\rangle=c_n^L|\overline{0}\rangle=0, \quad \forall n. \tag{7.5}$$

It is easily shown that this vacuum is BRST invariant for any d.

The state space that we have defined here treats the $x^\pm$ coordinates differently than the x^μ coordinates. Thus, we have lost manifest $d+2$-dimensional $SO(d,2)$ covariance. However, as the x^μ coordinates are regarded as the physical ones [in the sense of the gauge fixing conditions (6.2)], we still have manifest Lorentz covariance in this physical subspace. (If we do not insist on manifest Lorentz covariance, we may treat x^0 and one of the space coordinates in a similar fashion as $x^\pm$ and get a BRST-invariant vacuum in $d=6$.)

B. An inconsistent solution

We now consider another choice of state space. In analogy with the alternative treatment of the tensionless string in Sec. V, we may for the conformal string try a vacuum state satisfying

$$p_m^M|0\rangle=x_m^M|0\rangle=0, \quad m>0. \tag{7.6}$$

This alternative quantization was investigated in Ref. 3. The consistency conditions to (7.6) requires

$$c_m^{-1}|0\rangle=c_m^0|0\rangle=c_m^1|0\rangle=c_m^L|0\rangle=0, \quad m>0. \tag{7.7}$$

The conditions (7.6) and (7.7) allow for a ghost fixing of the form

$$c_0^1|0\rangle=0, \quad b_m^1|0\rangle=0, \quad m>0, \tag{7.8}$$

$$b_m^{-1}|0\rangle=b_m^0|0\rangle=b_m^L|0\rangle=0, \quad m\geqslant 0. \tag{7.9}$$

The consistency conditions for (7.8) are automatically satisfied as well as those of (7.9) for $m>0$. However,

$$\tilde{\phi}_0^{-1}|0\rangle=0 \tag{7.10}$$

and

$$\tilde{\phi}_0^0|0\rangle=0 \tag{7.11}$$

yields further conditions on the vacuum state. They may be satisfied. [Equation (7.11) fixes the conformal dimension of the vacuum state.] The problem is the last one, which yields

$$\tilde{\phi}_0^L|0\rangle=(d-2)\left(\frac{1}{2}\alpha^2-\frac{1}{12}\right)|0\rangle. \tag{7.12}$$

Checking the commutation relations, we find

$$\lim_{s\to 0}[\tilde{\phi}_m^L,\tilde{\phi}_{-m}^L](s)=2m\tilde{\phi}_0^L+\tfrac{1}{6}(d-26)m^3-m(d-2)\alpha^2, \tag{7.13}$$

where $\tilde{\phi}_0^L$ on the right-hand side is the normal ordered operator in (7.12). Notice that (7.13) coincides almost exactly with (5.11), although we here have $d+2$ coordinates and four ghost fields. It looks now as if the situation is the same as for the tensionless string in Sec. V, i.e., it looks as if we may follow the arguments after (5.11), leading to a consistent BRST quantization in $d=26$. However, for the conformal string there is an additional nontrivial commutator, for which we find

$$\lim_{s\to 0}[\tilde{\phi}_m^1,\tilde{\phi}_{-m}^{-1}](s)|0\rangle=\left(2i\tilde{\phi}_0^0-\frac{1}{2}(d-6)m\right)|0\rangle, \tag{7.14}$$

where $\tilde{\phi}_0^0$ on the right-hand side is the normal ordered operator that follows from (6.5) through our regularization, i.e.,

$$\tilde{\phi}_0^0=\frac{1}{4}(x_0p_0+p_0x_0)+\frac{1}{2}\sum_{k=1}^{\infty}(p_{-k}\cdot x_k+x_{-k}\cdot p_k)+i(c_0^1b_0^1-c_0^{-1}b_0^{-1})$$
$$+i\sum_{k=1}^{\infty}(b_{-k}^{-1}\cdot c_k^{-1}-c_{-k}^{-1}\cdot b_k^{-1}-b_{-k}^1\cdot c_k^1+c_{-k}^1\cdot b_k^1). \tag{7.15}$$

The relation (7.14) says that Q can only be nilpotent in $d=6$, which contradicts the above result that required $d=26$. It follows that there is no consistent BRST quantization of the conformal string on the vacuum considered here. This result agrees with Ref. 3.

VIII. ALTERNATIVE STATE SPACE FOR THE BOSONIC STRING

In view of the treatments in Secs. IV and V, one may wonder if one may not have a corresponding vacuum state also in the ordinary bosonic string case. Below we demonstrate that this is not the case. The classical constraints (3.1) may also be written as

$$P^{\mu}(\sigma)P_{\mu}(\sigma)+T^2X^{\mu\prime}(\sigma)X'_{\mu}(\sigma)=0, \quad P^{\mu}(\sigma)X'_{\mu}(\sigma)=0. \tag{8.1}$$

The Fourier modes of the corresponding Hermitian constraint operators are

$$\begin{aligned}\Phi_n^{-1}&\equiv\frac{1}{2}\sum_k p_k\cdot p_{n-k}-2T^2\sum_k k(n-k)x_k\cdot x_{n-k},\\ \phi_n^L&\equiv -i\frac{1}{2}\sum_k k(x_k\cdot p_{n-k}+p_{n-k}\cdot x_k).\end{aligned} \tag{8.2}$$

The Hermitian and formally nilpotent BRST charge (3.5) may be rewritten as

$$\begin{aligned}Q=\sum_k (\Phi_{-k}^{-1}c_k^{-1}+\phi_{-k}^L c_k^L)-\frac{1}{2}\sum_{k,l}(k-l)\bigg(&c_{-k}^{-1}c_{-l}^L b_{k+l}^{-1}+\frac{1}{2}c_{-k}^L c_{-l}^L b_{k+l}^L\\ &+4T^2c_{-k}^{-1}c_{-l}^{-1}b_{k+l}^L+b_{k+l}^{-1}c_{-k}^{-1}c_{-l}^L+\frac{1}{2}b_{k+l}^L c_{-k}^L c_{-l}^L\bigg).\end{aligned} \tag{8.3}$$

[Notice that the ghost variables c_k^L, b_k^L here are not the same as in (3.5).] Here the extended constraints are given by

$$\begin{aligned}\tilde{\phi}_n^L&\equiv[Q,b_n^L]=\phi_n^L-\frac{1}{2}\sum_k (n+k)(c_k^{-1}b_{n-k}^{-1}-b_{n-k}^{-1}c_k^{-1}+c_k^L b_{n-k}^L-b_{n-k}^L c_k^L),\\ \tilde{\Phi}_n^{-1}&\equiv[Q,b_n^{-1}]=\Phi_n^{-1}-\sum_k (n+k)(c_k^L b_{n-k}^{-1}+4T^2c_k^{-1}b_{n-k}^L).\end{aligned} \tag{8.4}$$

A direct calculation of the commutator $[\tilde{\phi}_m^L,\tilde{\phi}_{-m}^L]$ gives

$$\begin{aligned}[\tilde{\phi}_m^L,\tilde{\phi}_{-m}^L](s)=m\sum_k &(-ikx_k\cdot p_{-k}-kc_k^L b_{-k}^L-kc_k^{-1}b_{-k}^{-1})(f^{(\alpha)}(|m+k|,s)+f^{(\alpha)}(|m-k|,s))\\ &+\sum_k (-ik^2x_k\cdot p_{-k}+(2m^2-k^2)(c_k^L b_{-k}^L+c_k^{-1}b_{-k}^{-1}))\\ &\times(f^{(\alpha)}(|m+k|,s)-f^{(\alpha)}(|m-k|,s)).\end{aligned}$$

The $s\to 0$ limit yields

$$[\tilde{\phi}_m^L,\tilde{\phi}_{-m}^L]=2m\tilde{\phi}_0^L, \tag{8.5}$$

which is consistent with the fact that the BRST charge (8.3) is nilpotent. One may note here that the second line in the last relation goes to zero as s goes to zero, if the operator

$$\Psi_{k,m}\equiv -ik^2x_k\cdot p_{-k}+(2m^2-k^2)(c_k^L b_{-k}^L+c_k^{-1}b_{-k}^{-1}), \tag{8.6}$$

is well defined. This is true for a state space where the vacuum has the form of the tensionless string vacuum defined by the condition

$$p_n^{\mu}|0\rangle=0, \quad \forall n. \tag{8.7}$$

However, this vacuum state is not appropriate for the tensile bosonic string. Consistency conditions of the form

$$[Q,p_n^{\mu}]|0\rangle=0, \quad \forall n \tag{8.8}$$

would require the ghost part of such a state to satisfy the relations

$$c_n^{-1}|0\rangle=0, \quad \forall n, \tag{8.9}$$

for which the consistency conditions,

$$[Q,c_n^{-1}]|0\rangle=0, \quad \forall n, \tag{8.10}$$

are satisfied. In fact, it is possible to ghost fix this vacuum state in a BRST-invariant way using the original BRST charge, which then also is nilpotent for any dimension d. However, there exists no dual vacuum state, $|\overline{0}\rangle$, to this solution. Note that $|\overline{0}\rangle$ must satisfy

$$x_n^{\mu}|\overline{0}\rangle=b_n^{-1}|\overline{0}\rangle=0, \tag{8.11}$$

for which the consistency condition,

$$[Q,b_n^{-1}]|\overline{0}\rangle=0, \tag{8.12}$$

cannot be satisfied.

IX. ALTERNATIVES WITH NEGATIVE NORM STATES

In the alternative treatments in Secs. V and VII, we considered a vacuum state of the type

$$p_m^{\mu}|0\rangle=x_m^{\mu}|0\rangle=0, \quad m>0. \tag{9.1}$$

It leads to a ''consistent'' BRST quantization for the bosonic tensionless string in $d=26$. In fact, even the ordinary bosonic string has a vacuum state satisfying (9.1). To see this one may notice that (9.1) is equivalent to (we let $|0\rangle$ be an eigenstate of p_0^{μ}, i.e., we set $|0\rangle=|p\rangle$),

$$\alpha_m^{\mu}|p\rangle=0, \quad \tilde{\alpha}_{-m}^{\mu}|p\rangle=0, \quad m>0, \tag{9.2}$$

due to the relations (3.3). Now (9.2) is different from the standard vacuum (3.10). The consistency conditions of the latter condition requires

$$c_{-m}^{K}|p\rangle=0, \quad m>0, \tag{9.3}$$

which means that the vacuum may be ghost fixed by

$$b_{-m}^{K}|p\rangle=0, \quad m\geqslant 0, \tag{9.4}$$

together with

$$c_m^{L}|p\rangle=0, \quad m>0; \quad b_m^{L}|p\rangle=0, \quad m\geqslant 0. \tag{9.5}$$

Calculating commutators for finite s and then taking the limit $s\to 0$, we find

$$[\tilde{K}_m,\tilde{K}_{-m}]=2m\tilde{K}_0-\tfrac{1}{12}(d-26)m^3+\tfrac{1}{2}m(d-2)\alpha^2, \tag{9.6}$$

where

$$\tilde{K}_0=\frac{1}{2}\alpha_0^2+\sum_{k=1}^{\infty}(\tilde{\alpha}_k\cdot\tilde{\alpha}_{-k}-kb_k^{K}c_{-k}^{K}-kc_k^{K}b_{-k}^{K})-\frac{1}{2}(d-2)\left(\frac{1}{2}\alpha^2-\frac{1}{12}\right). \tag{9.7}$$

We may then define a new BRST charge by

$$Q' = Q - \frac{d-2}{4}\alpha^2 (c_0^L - c_0^K), \tag{9.8}$$

which then is nilpotent for $d=26$. In $d=26$ the new extended constraints are

$$\begin{aligned} \tilde{L}_0' &= \frac{1}{2}\alpha_0^2 + \sum_{k=1}^{\infty} (\alpha_{-k}\cdot\alpha_k + kb_{-k}^L c_k^L + kc_{-k}^L b_k^L) - 1, \\ \tilde{K}_0' &= \frac{1}{2}\alpha_0^2 + \sum_{k=1}^{\infty} (\tilde{\alpha}_k \cdot \tilde{\alpha}_{-k} - kb_k^K c_{-k}^K - kc_k^K b_{-k}^K) + 1. \end{aligned} \tag{9.9}$$

However, $\tilde{L}_0'|p\rangle = 0$ requires $p^2 = 8T$ while $\tilde{K}_0'|p\rangle = 0$ requires $p^2 = -8T$. It follows that neither of these vacua are BRST invariant, or, in other words, we have no BRST-invariant vacuum satisfying (9.1). Still there are BRST-invariant states and they are massless with spin 0, 1, and 2, and spinless with $p^2 = \pm 8T$. Note, however, that since $\tilde{\alpha}_m^\mu$, $m>0$ act as creation operators, the space components $\tilde{\alpha}_m^i$ yield negative normed states. Thus, the BRST-invariant state spaces have negatively normed states. The same feature was also found for the alternative treatment of the tensionless string in Sec. V. (This alternative treatment has only massless states.[16]) It is clear that the existence of indefinite metric states in the BRST-invariant sector makes both these models inconsistent.

ACKNOWLEDGMENTS

P. S. wishes to thank R. vonUnge and U. Lindström for useful comments and suggestions. The work of P. S. was supported by the European Commission TMR Programme under the Grant No. ERBFMBI-CT97-2823.

[1] M. Kato and K. Ogawa, Nucl. Phys. B **212**, 443 (1983); S. Hwang, Phys. Rev. D **28**, 2614 (1983).
[2] D. Friedan, E. Martinec, and S. Shenker, Nucl. Phys. B **271**, 93 (1986).
[3] J. Isberg, ''Tensionless strings with manifest space–time conformal invariance,'' USITP-92-10 (1992).
[4] H. Gustafsson, U. Lindström, P. Saltsidis, B. Sundborg, and R. v. Unge, Nucl. Phys. B **440**, 495 (1995).
[5] S. Hwang and R. Marnelius, Nucl. Phys. B **315**, 638 (1989); **320**, 476 (1989).
[6] S. Hwang, Nucl. Phys. B **322**, 107 (1989); R. Marnelius, *ibid.* **372**, 218 (1992); **384**, 318 (1992).
[7] I. A. Batalin and E. S. Fradkin, Ann. Inst. Henri Poincaré **49**, 145 (1988).
[8] E. Elizalde *et al.*, *Zeta Regularization Techniques with Applications* (World Scientific, Singapore, 1994).
[9] A. Schild, Phys. Rev. D **16**, 1722 (1977); A. Karlhede and U. Lindström, Class. Quantum Grav. **3**, L73 (1986); F. Lizzi, B. Rai, G. Sparano, and A. Srivastava, Phys. Lett. B **182**, 326 (1986); R. Amorim and J. Barcelos-Neto, Z. Phys. C **38**, 643 (1988); U. Lindström, B. Sundborg, and G. Theodoridis, Phys. Lett. B **253**, 319 (1991); **258**, 331 (1991).
[10] F. Lizzi, B. Rai, G. Sparano, and A. Srivastava, Phys. Lett. B **182**, 326 (1986); J. Isberg, U. Lindström, B. Sundborg, and G. Theodoridis, Nucl. Phys. B **411**, 122 (1994).
[11] J. Gamboa, C. Ramirez, and M. Ruiz-Atlaba, Phys. Lett. B **225**, 335 (1989).
[12] R. Marnelius, Nucl. Phys. B **412**, 817 (1994); I. A. Batalin and R. Marnelius, *ibid.* **442**, 669 (1995).
[13] P. Saltsidis, Nucl. Phys. B **446**, 286 (1995).
[14] P. Saltsidis, Phys. Lett. B **401**, 21 (1997).
[15] E. T. Whittaker and G. N. Watson, *Modern Analysis*, 4th ed. (Cambridge University Press, Cambridge, 1952).
[16] J. Gamboa, C. Ramirez, and M. Ruiz-Atlaba, Nucl. Phys. B **338**, 143 (1990).
[17] R. Marnelius, Phys. Rev. D **20**, 2091 (1979); R. Marnelius and B. Nilsson, *ibid.* **22**, 830 (1979).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Time dependence of operators in anharmonic quantum oscillators: Explicit perturbative analysis

Peter B. Kahn
Department of Physics, State University of New York, Stony Brook, New York 11794

Yair Zarmi
Center for Environmental Physics, The Jacob Blaustein Institute for Desert Research, Ben-Gurion University of the Negev, Sede Boqer Campus, 84990, Israel

(Received 10 August 1998; accepted for publication 7 July 1999)

An explicit, order-by-order perturbative solution, valid over extended time scales, for the time dependence of operators of anharmonic oscillators, is developed within the framework of the method of normal forms. The freedom of choice of the zeroth-order term and, concurrently in the higher-order corrections, is exploited to develop a minimal normal form (MNF). The expansion for the eigenvalues of the perturbed Hamiltonian in a standard form is independent of the choice. However, the simple form obtained for the time dependence of the perturbative solution is more suitable than any other choice for application to high-lying excited states, as it offers a renormalized form for the propagator. © *1999 American Institute of Physics.* [S0022-2488(99)02310-5]

The issue of finding perturbative approximations, that are valid for long times, for the time dependence of quantum systems had been raised years ago in Ref. 1. Recently, it has been addressed again in the context of anharmonic oscillators. References 2 and 3 presented a detailed analysis of the quantum Duffing oscillator within the framework of the method of multiple time scales,[4,5] and in Ref. 6 this oscillator was analyzed within the framework of the method of normal forms.[7–9]

In the case of classical oscillator systems with nonlinear couplings, the freedom, inherent in the expansion, of the choice of the zeroth-order term and the concurrent modifications in the higher-order corrections, has been exploited for the derivation of *minimal normal forms* (MNF).[6,10] The errors incurred in approximations to the full solutions in the MNF choice grow appreciably more slowly, making the approximations valid for longer time spans, than in commonly employed expansion choices. This is demonstrated by several examples in Refs. 10 and 11.

One may naturally ask whether a MNF expansion can be employed advantageously also in the quantum mechanical case. Using the normal form expansion method, we obtain for the time dependence of the operators of the anharmonic oscillator an explicit, order-by-order, perturbative solution, valid over extended time scales. We show that the expansion for the eigenvalues of the perturbed Hamiltonian in a standard form is independent of the choice of the zeroth-order term. However, the simplicity and usefulness of the functional form of the time dependence found for the approximate solution is affected by the choice. The functional form obtained in the MNF choice is simpler and is more suitable than all other choices for application to high-lying excited states. It provides an order-by-order procedure for renormalizing the propagator.

Consider the quantum mechanical Hamiltonian of an anharmonic oscillator (in units of $\hbar\omega$):

$$H = a^\dagger a + \frac{1}{2} + \frac{1}{2^{p/2}p}\,\epsilon(a^\dagger + a)^p,$$

$$x = \frac{(a+a^\dagger)}{\sqrt{2}}, \quad \dot{x} = \frac{(a-a^\dagger)}{\sqrt{2}i}, \quad [a,a^\dagger] = 1.$$

The perturbation is turned on at $t=0$, so that the initial condition for the operator is

$$a(t=0)=a_0,$$

where a_0 is the unperturbed annihilation operator.

The normal form expansion entails the order-by-order parallel computation of two infinite series: The near-identity transformation (NIT), that relates the full solution for a, to the zeroth-order term, b,

$$a=b+\epsilon T_1(b^\dagger,b)+\epsilon^2 T_2(b^\dagger,b)+\cdots, \tag{1}$$

and the equation for the time dependence of the zeroth-order term (the normal form),

$$\frac{db}{dt}=U_0+\epsilon U_1+\epsilon^2 U_2+\cdots. \tag{2}$$

To be specific, we study in detail the case $p=4$, corresponding to the quantum Duffing oscillator. (The method clearly applies to any integer $p\geqslant 0$.) The time dependence of a is determined by

$$\frac{da}{dt}=ia+\frac{1}{4}i\epsilon(a^\dagger+a)^3=ia+\frac{1}{4}i\epsilon\begin{pmatrix} a^{\dagger^3}+3a^{\dagger^2}a+a^3+3a^\dagger \\ \underline{+3a^\dagger a^2+3a} \end{pmatrix}. \tag{3}$$

Inserting Eqs. (1) and (2) in Eq. (3), one obtains in each order, n, a relation between T_n and U_n that does not enable their unique determination. This nonuniqueness is partially reduced by choosing T_n not to depend explicitly on time (a permissible choice, as the system under study is autonomous). The remaining ambiguity is further reduced by requiring that the approximate solution constitutes a valid approximation over extended time scales. This requires that secular errors (errors that evolve linearly in time and limit the validity of the approximation to short times only) are eliminated in the approximate solution in each order. In the method of normal forms, one exploits the freedom in the expansion to meet this requirement by assigning all *resonant* terms [of the form $(b^\dagger\cdot b)^k\cdot b$, similar in form to the underlined terms in Eq. (3)], that arise in the nth-order dynamical equation, to U_n. The functional form of U_n is, thus,

$$U_n(b^\dagger,b)=g_n(b^\dagger\cdot b)b. \tag{4}$$

This accounts for the full effect of the perturbation on the zeroth-order term, b [Eq. (2)] in $O(\epsilon^n)$. All nonresonant terns are used for constructing T_n.

This division of terms between T_n and U_n enables the unique determination of U_n, but, still, not of T_n. The latter is determined up to a *free resonant term* [similar in form to Eq. (4)], which affects U_k and T_k in all orders $k\geqslant n+1$. For instance, in $O(\epsilon)$, U_1 is determined uniquely, while T_1 is not:

$$U_1=\tfrac{3}{4}i(b^\dagger\cdot b+1)b, \tag{5a}$$

$$T_1=-\tfrac{1}{16}b^{\dagger^3}-\tfrac{3}{8}b^{\dagger^2}b+\tfrac{1}{8}b^3-\tfrac{3}{8}b^\dagger+F_1(b^\dagger\cdot b)\cdot b; \tag{5b}$$

$F_1\cdot b$, the free resonant term in T_1, does not enter the $O(\epsilon)$ relation between T_1 and U_1. However, it affects all higher-order equations. In the following, we show its effect on U_2.

As all known terms in $O(\epsilon)$ are either cubic or linear in b and $b^\dagger$, one may choose for the free term in Eq. (5b):

$$F_1(b^\dagger\cdot b)\cdot b=\gamma b^\dagger\cdot b^2+\nu b. \tag{6}$$

With this choice, the expansion in powers of ϵ yields for U_2 the following expression (an asterisk denotes complex conjugation):

$$U_2=\tfrac{3}{4}i\{[(\gamma+\gamma^*)-\tfrac{17}{16}](b^\dagger b)^2+[(\nu+\nu^*)-(\gamma+\gamma^*)-\tfrac{17}{8}](b^\dagger b)-\tfrac{3}{2}\}b. \tag{7}$$

The commutator of b and $b^\dagger$ is

$$[b,b^\dagger]=1-\epsilon\{[a,T_1^\dagger]+[T_1,a^\dagger]\}+O(\epsilon^2)=1-\epsilon\{2(\gamma+\gamma^*)b^\dagger b+(\nu+\nu^*)\}+O(\epsilon^2). \tag{8}$$

In general, the NIT is not unitary, and the commutator is not equal 1, as b and $b^\dagger$ do not represent a wave function that is normalized to unity. Unitarity amounts to a specific choice of the free terms in T_n, in every order. For instance, in $O(\epsilon)$, it requires that γ and ν are purely imaginary.

One can show by induction that, for all n, U_n have an overall multiplicative i, with all remaining coefficients real. Thus, the structure of the normal form is

$$\frac{db}{dt}=i\left\{1+\sum_{n\geqslant 1}\epsilon^n f_n(b^\dagger\cdot b)\right\}b, \tag{9}$$

where f_n are real functions of $(b^\dagger\cdot b)$. Equation (9) implies that, within the framework of perturbation theory, $(b^\dagger\cdot b)$ is constant, and the solution for b is given explicitly by

$$b=\exp\left[i\left\{1+\sum_{n\geqslant 1}\epsilon^n f_n(b_0^\dagger\cdot b_0)\right\}t\right]b_0. \tag{10}$$

For a general NIT, this solution is cumbersome. In particular, the powers of $(b_0^\dagger\cdot b_0)$ in $f_n(b_0^\dagger\cdot b_0)$ increase with the order, n, in the sum in Eq. (10). This implies that the convergence (or asymptotic) characteristics of the phase factor in Eq. (10) become worse as one goes to high-lying excited states, for which the eigenvalues of $(b^\dagger\cdot b)$ (approximately, the occupation numbers of the unperturbed states) are large. For example, the ϵ term in the phase factor is uniquely determined and contains the monomials $(b^\dagger\cdot b)^k$, with $k=0,1$. In general, the ϵ^2 term will contain monomials $(b^\dagger\cdot b)^k$, with $k=0,1,2$, but is not unique. It depends on the choice of the free resonant term in T_1. In a similar manner, for $n\geqslant 2$, the functional form of the ϵ^n term in Eq. (10) depends on the choice of the free functions in T_k in the near identity transformation, Eq. (1), in all orders $k<n$.

The nonuniqueness of the expansion can be exploited to greatly simplify the solution by invoking the MNF choice. The latter corresponds to a specific choice of the free functions in the NIT, in all orders. We show the procedure in detail for the $O(\epsilon^2)$ contribution and sketch the main points for the analysis of the $O(\epsilon^3)$ contribution. Choosing the free parameters in F_1 to have the values

$$\gamma+\gamma^*=\tfrac{17}{16},\qquad \nu+\nu^*=\tfrac{27}{16},$$

one eliminates the quintic monomial in the expression for U_2 [Eq. (7)] and obtains

$$U_2=-\tfrac{3}{2}U_1. \tag{11}$$

Thus, U_2, now also contains only cubic and linear terms in b and $b^\dagger$.

To show how the MNF idea is extended to U_3, we note that, in $O(\epsilon^2)$, T_2 may include the following nonresonant terms:

$$b^{\dagger^5},b^{\dagger^4}b,b^{\dagger^3}b^2,b^\dagger b^4,b^5,b^{\dagger^3},b^{\dagger^2}b,b^3,b^\dagger.$$

The coefficients of these terms are determined uniquely by the $O(\epsilon^2)$ dynamical equation (which relates T_2 and U_2). In addition, T_2 may contain the following free resonant terms:

$$b^{\dagger^2}b^3,b^\dagger b^2,b,$$

which do not affect the $O(\epsilon^2)$ dynamical equation [just as the resonant terms in T_1 did not affect the $O(\epsilon)$ dynamical equation]. However, these free terms do affect the form of U_3. The coefficients of these three resonant terms are arbitrary. In general, U_3 may include the following resonant terms:

$$(b^\dagger b)^3 \cdot b, \quad (b^\dagger b)^2 \cdot b, \quad (b^\dagger b) \cdot b, \quad b.$$

One can choose the free parameters in T_2, to make U_3 also proportional to U_1, as in Eq. (11), so that it will also contain only cubic and linear terms in b and $b^\dagger$. The procedure can be repeated in higher orders, and the normal form becomes

$$\frac{db}{dt} = ib + \left(\epsilon - \frac{3}{2}\epsilon^2 + \cdots \right) U_1 = ib + \tfrac{3}{4} i(\epsilon - \tfrac{3}{2}\epsilon^2 + \cdots)(b^\dagger b + 1)\cdot b. \tag{12}$$

Equation (12) is solved by

$$b = \exp\{it\}\exp\{\tfrac{3}{4} i(\epsilon - \tfrac{3}{2}\epsilon^2 + \cdots)(b_0^\dagger b_0 + 1)t\}\cdot b_0. \tag{13}$$

Once the solution for b is known, a may be computed to the desired accuracy, using the NIT, Eq. (1). We note that the power of $(b_0^\dagger \cdot b_0)$ is the same for *all orders* in the exponent in Eq. (13), providing an approximation scheme that is more useful, even for high-lying excited states, than the general solution, Eq. (10), where the powers of $(b_0^\dagger \cdot b_0)$ in f_n increase with the order n. Thus, the MNF choice provides an order-by-order procedure for the derivation of a renormalized propagator.

As an example for the for the resulting time dependence, we show the action of b on the vacuum of the unperturbed a states. We denote the nth excited state of the oscillator by $|n,t\rangle_a$, so that $|n,t=0\rangle_a$ is the unperturbed state at $t=0$. Inverting Eq. (1) through $O(\epsilon)$ and using Eqs. (5b), (6), and (13), one finds

$$b(t)|0\rangle_a = \epsilon \exp(it)\left\{ \frac{\sqrt{6}}{16}\exp(3\epsilon it)|3,t=0\rangle_a + \frac{3}{8}\exp(3/2\epsilon it)|1,t=0\rangle_a \right\} + O(\epsilon^2), \quad t \leqslant O(1/\epsilon) \tag{14}$$

To compute the energy levels, we note that, through $O(\epsilon)$, the diagonalized Hamiltonian becomes

$$H = b^\dagger b + \tfrac{1}{2} + \epsilon[(\tfrac{3}{8} + (\gamma+\gamma^*))(b^\dagger b)^2 + (\tfrac{1}{2} + (\nu+\nu^*) - (\gamma+\gamma^*))(b^\dagger b) + \tfrac{1}{4}] + O(\epsilon^2). \tag{15}$$

To compute the eigenvalues of H, we need to find eigenstates of the operator $B = (b^\dagger \cdot b)$, which, as pointed out earlier, is time independent. We denote an eigenstate of B with eigenvalue β, by $|\beta\rangle$ and expand it in eigenfunctions of the unperturbed oscillator, using the notation of Eq. (13):

$$|\beta\rangle = |n,t=0\rangle_a + \epsilon \sum_{k\neq n} \alpha_k |k,t=0\rangle_a. \tag{16}$$

Inverting the NIT [Eq. (1)], and employing Eqs. (5b) and (6), one finds through $O(\epsilon)$,

$$\beta = n\{1 - \epsilon[(\gamma+\gamma^*)(n-1) + (\nu+\nu^*)]\} + O(\epsilon^2). \tag{17}$$

Only a states with population $n\pm 4$ and $n\pm 2$ are admixed into $|\beta\rangle$ through $O(\epsilon)$, with coefficients

$$\alpha_{n+4} = \frac{1}{16}\,\frac{\sqrt{(n+1)(n+2)(n+3)(n+4)}}{n},$$

$$\alpha_{n+2}=\frac{1}{8}\frac{(2n+3)\sqrt{(n+1)(n+2)}}{n},$$

$$\alpha_{n-2}=\frac{1}{8}\frac{(2n-1)\sqrt{n(n-1)}}{n}\quad(n\geqslant 2),$$

$$\alpha_{n-4}=\frac{1}{16}\frac{\sqrt{n(n-1)(n-2)(n-3)}}{n}\quad(n\geqslant 4).$$

We note that each eigenvalue of $B=(b^\dagger\cdot b)$ is close to an eigenvalue n of the unperturbed number operator, $(a^\dagger\cdot a)$. Also, the deviation of β from n depends on $(\gamma+\gamma^*)$ and on $(\nu+\nu^*)$.

Inserting Eq. (17) for β in Eq. (15), we find the energy of a $|\beta\rangle$ eigenstate to be

$$E_\beta=n+\tfrac{1}{2}+\epsilon(\tfrac{3}{8}n^2+\tfrac{1}{2}n+\tfrac{1}{4})+O(\epsilon^2). \tag{18}$$

This result is *independent* of γ and ν, namely, of the free resonant term in T_1. Clearly, the energy of an eigenstate of the perturbed Hamiltonian, when expressed in terms of the quantum number of the reference adjacent unperturbed state, is independent of the choice of the free functions in any order. Thus, for computing the eigenvalues through a given order in ϵ, there is no advantage in a specific choice (e.g., a unitary NIT) of the free resonant terms.

The choice of the free parameters depends on what one wishes to achieve. For a unitary NIT, b and $b^\dagger$ represent wave functions that are normalized to unity, obeying $[b,b^\dagger]=1$, so that $B=(b^\dagger\cdot b)$ is a true number operator. In $O(\epsilon)$, unitarity of the NIT implies: $\gamma+\gamma^*=\nu+\nu^*=0$. One then has

$$\beta=n.$$

Thus, the number eigenvalue for a b state is equal to the number eigenvalue of the adjacent unperturbed a state. H then becomes

$$H=b^\dagger b+\tfrac{1}{2}+\epsilon[\tfrac{3}{8}(b^\dagger b)^2+\tfrac{1}{2}(b^\dagger b)+\tfrac{1}{4}]+O(\epsilon^2).$$

In this case, computation of the time dependence of a is cumbersome, because the functional form of higher-order terms will not be the same as that of the first-order one. For a nonunitary NIT, one has $[b,b^\dagger]=1+O(\epsilon)$ [see Eq. (8)]. In the particular case of the MNF choice, one obtains

$$H=b^\dagger b+\tfrac{1}{2}+\epsilon[\tfrac{23}{16}(b^\dagger b)^2+\tfrac{9}{8}(b^\dagger b)+\tfrac{1}{4}]+O(\epsilon^2).$$

Here the time dependence of a is simpler, and is easily found to any desired order [Eq. (13)] and the result applies more readily also to high-lying excited states.

We mention in passing that when the perturbation has two terms,

$$H=a^\dagger a+\frac{1}{2}+\frac{1}{2^{p/2}p}\,\epsilon(a^\dagger a)^p+\frac{1}{2^{q/2}q}\,\epsilon^2(a^\dagger+a)^q,$$

with $p\leqslant q$, then a reduction of the normal form to a MNF, similar in structure to Eq. (12), can be affected, starting from $O(\epsilon^2)$. If $p>q$, so that (in the corresponding classical system) a competition of scales occurs, the normal form can be reduced to a MNF, starting only from $O(\epsilon^3)$.

In summary, the freedom in the choice of the zeroth-order approximation, inherent in any perturbation expansion scheme, may be exploited to construct a unitary NIT that yields a simple form for the diagonalized Hamiltonian, or a MNF, to obtain a simple form, that may be useful also for high-lying excited states, for the time dependence of the operators. The latter alternative provides an order-by-order algorithm for obtaining a renormalized propagator. The MNF choice may be also relevant in the study of coherent states, where nonunitary transformations have been considered,[12] and in non-Hermitian Hamiltonian problems that have recently received renewed attention,[13–17] and where unitarity of the NIT is not required. Finally, while the method of normal forms has been used in the present analysis, a parallel procedure can be easily formulated for the method of multiple time scales. Denoting the time scales by $\tau_n = \epsilon^n t$, the MNF constraints on the free resonant terms in T_1, $T_2, \ldots,$ correspond to specific requirements on the $\tau_2, \tau_3, \ldots,$ dependence of the solution in the latter method.

1 M. Kummer, Nuovo Cimento **B1**, 123 (1971).
2 C. M. Bender and L. Bettencourt, Phys. Rev. D **54**, 7710 (1996).
3 C. M. Bender and L. Bettencourt, Phys. Rev. Lett. **77**, 4114 (1996).
4 A. H. Nayfeh, *Perturbation Methods* (Wiley, New York, 1973).
5 C. M. Bender and S. A. Orszag, *Advanced Methods for Scientists and Engineers* (McGraw-Hill, New York, 1978).
6 P. B. Kahn and Y. Zarmi, Physica D **54**, 65 (1991).
7 H. Poincaré, *New Methods of Celestial Mechanics* [originally published as Les Methodes Nouvelles de la Mécanique Céleste (Gauthier-Villars, Paris, 1892)] (American Institute of Physics, Woodbury, NY, 1993).
8 V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations* (Springer-Verlag, New York, 1988).
9 J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields* (Springer-Verlag, New York, 1983).
10 P. B. Kahn and Y. Zarmi, *Nonlinear Dynamics: Exploration Through Normal Forms* (Wiley, New York, 1998).
11 P. B. Kahn and Y. Zarmi, Physica D **118**, 221 (1998).
12 A. Janussis, G. Kilros, and E. Vlahos, Commun. Theor. Phys. **5**, 33 (1996).
13 B. M. McCoy and T. T. Wu, Nuovo Cimento B **56**, 311 (1968).
14 L.-H. Gwa and H. Spohn, Phys. Rev. A **46**, 844 (1992).
15 H. C. Fogedby, A. B. Eriksson, and L. V. Mikheev, Phys. Rev. Lett. **75**, 1883 (1995).
16 N. Hatano and D. R. Nelson, Phys. Rev. Lett. **77**, 570 (1996).
17 N. Shnerb, Phys. Rev. B **55**, R3382 (1997).

Nonassociative structure of quantum mechanics in curved space–time

Edward H. Kerner
Sharp Physics Laboratory, University of Delaware, Newark, Delaware 19716

(Received 19 May 1999; accepted for publication 16 July 1999)

de Sitter space gives a local, osculating (and therefore prototypical) representation of a curved space–time. The de Sitter geodesics allow a description as uniform straight-line motion with respect to a family of preferred frames that are inertial frames interconnected by a projective (fractional-linear) transformation. Use of these frames overcomes the ambiguities of general covariance in stating commutation rules and Hamiltonian structure. But the Hamiltonian for a particle running on a geodesic then necessitates nonassociative elements. The latter are worked out as elements of a Cayley–Dickson algebra of 16 dimensions (doubled octonions, or decahexions), wherein a standard Hilbert space type of format for quantum mechanics is ruled out. A formal Schrödinger wave equation is devised, and exhibits antilinear as well as linear terms in the decahexion components of the wave function. All the same, the Heisenberg laws of motion for a dynamical variable are neatly and unambiguously formulated, giving a full account of the quantum time evolution of the dynamical variable (although the Heisenberg program of diagonalizing the Hamiltonian cannot be executed).
[S0022-2488(99)04510-7]

I. INTRODUCTION

The structure of spinors in a curved space was broached by Schrödinger[1] and by Bargmann[2] in the early days of quantum theory, and their formulation has since continued to hold sway.[3] The idea was simply to uplift the flat-space Dirac equation,

$$(\gamma^{\mu}\partial_{\mu}+m)\psi=0,$$

to a curved-space version,

$$(\gamma^{\mu}(x)\boldsymbol{\nabla}_{\mu}+m)\psi=0,$$

with covariant derivatives $\boldsymbol{\nabla}_{\mu}$ replacing ∂_{μ}, under general covariance guidelines, while requiring, among other things

$$[\gamma^{\mu}(x),\gamma^{\nu}(x)]=2g^{\mu\nu}(x),$$

that generalize the flat-space anticommutators.

The principal point is that, by explicit design, curvature is assumed to leave the fundamental gross format inherited from flat space untouched. The design is in itself clearly a purely formal recipe, failing to connect with underlying commutation rules or Hamiltonian structure, which is where the force of Planck's constant is ordinarily brought into full physical play. The failure, of course, is due at bottom to the amorphousness and vagueness of general covariance upon which the recipe insists.

In contrast, in the present note we aim to show that, from an *a priori* and unambiguous reckoning of the Hamiltonian H of a particle running along a geodesic path in curved space, and from the simplest physically based *a priori* commutation rules relating to that path, the spinor structure in the curved space is qualitatively, inevitably altered from the familiar flat-space Dirac format. This owes to the necessity for extracting a square root of H^2, as in ordinary Dirac theory, which, however, for geodesic motion in a curved space (typified here as a space of constant curvature) simply cannot be done at all *unless elements of a nonassociative algebra are admitted. It is these elements, constituting some sort of novel internal degrees of freedom, that necessarily enter the description of curved-space spinors*, and distinguish them at a basic structural level as something quite different from embellished flat-space spinors.

The present preliminary discussion is limited to geodesic motion in the simplest possible curved space. de Sitter space of symmetry $\mathcal{O}(3,2)$ of constant positive curvature, whose quantum theory has been appreciably discussed.[4] This restriction is less severe than appears at first glance, since it has been shown[5] that de Sitter spaces are the local osculating spaces to more generally curved Riemann spaces, seizing the local characterization of curvature, via higher order of contact, that is beyond reach from the low-order contact of mere tangent (Minkowski) spaces. In this sense, de Sitter space, so far from being highly specialized, comes close to being prototypical. The bubble of de Sitter space far overmatches the flake of Minkowski space in depicting the local geometric situation.

II. de SITTER SPACE HAMILTONIAN

A remarkable, if not widely familiar, feature of de Sitter space[6] is that its symmetry group of five-dimensional rotations $\mathcal{O}(3,2)$, can also and equivalently be read in terms of specialized projective (fractional-linear) transformations,

$$x_i' = \frac{A_{i\alpha}x_\alpha + B_i}{R_\alpha x_\alpha + 1}, \tag{1}$$

of space (x_1, x_2, x_3) and time (x_0 or t) coordinates (sum on repeated Greek indices understood). These projective transformations send unaccelerated free-particle motions into other free-particle motions, as with the familiar linear transformations of Galilean and special relativity. For example, if

$$x' = x_0' + v't',$$

then

$$x' = \frac{A_{11}x + A_{10}t + B_1}{R_1 x + R_0 t + 1}, \quad t' = \frac{A_{01}x + A_{00}t + B_0}{R_1 x + R_0 t + 1},$$

produces, owing to the common denominator,

$$x = x_0 + vt.$$

The projective transformations are *bona fide* inertial-frame transformations, in fact, the most general ones.

The tracing of the connection between projective and de Sitter groups is worked out in detail in Sec. III. Suffice it to say, the connection is encapsulated by the invariance under linear homogeneous transformations of the quadratic form,

$$X_1^2 + X_2^2 + X_3^2 + X_4^2 + f^2(X_4 - X_5)^2 = \text{invariant}, \tag{2}$$

where x_i has been written X_i/U in homogeneous coordinates, and where $X_4 \equiv icX_0$ and $X_5 \equiv ibU$ (b is a universal fundamental length and f a dimensionless constant). The homogeneous transformations in X_1, X_2, X_3, X_4, X_5 are essentially five-dimensional rotations that convert back to fractional-linear transformations in physical coordinates x_1, x_2, x_3, x_0. Based on Eq. (2), the infinitesimal invariant line element ds^2 in dt and $d\mathbf{r} \equiv (dx, dy, dz)$ is readily produced, and readily identifiable as one characterizing a space–time of constant curvature $a \equiv fb$.

Thus, notwithstanding the constant curvature of space–time, *the geodesics of de Sitter space one and all may be brought forth as the global free-particle motions*,

$$\frac{d^2\mathbf{r}}{dt^2}=0,$$

with the de Sitter group giving in projective language the description of motion from the viewpoints of equivalent inertial frames. This projective rephrasing of curvature works because the de Sitter free particle does not travel off to a remote place and stay there, as in flat space: it rides *through* a projective-geometric ''point at infinity,'' corresponding to the vanishing of the denominator $1+R_\alpha X_\alpha$ in Eq. (1), and reappears on the local scene (in fact, there is no distinction between the ''local'' and the ''infinitely remote'' scenes).

The inertial frames that here are interconnected in a projective account of the de Sitter group are clearly as distinguished for de Sitter space, as are the familiar linearly connected inertial frames of the Poincaré group for Minkowski space. All physical experience of the latter now leads to the basic physical hypothesis of the present work: *the inertial frame coordinates of a free particle, whether in de Sitter or Minkowski space, are the distinguished ones for prescribing commutation rules of the usual type* $(x,p_x)=i\hbar$, *etc., and building quantum theory along Hamiltonian lines*.

This abandonment of general covariance in favor of physically identifiable preferred inertial frames overcomes a longstanding difficulty of general covariance *vìs-a-vìs* quantum theory. Namely, if every kind of transformation is admitted between coordinate systems, all taken to be on the same footing, the formulation of commutation rules—showing at the deepest level just how and where Planck's constant enters physics—becomes completely ambiguous. This omnipresent ambiguity of general covariance is set aside once *distinguished* coordinates are recognized; and an approach is afforded to a basic background problem of general relativity, viz., the quantum theory of geodesics in a prescribed Riemannian geometric setting.

Of course, *after* preferred frames and clear commutation rules have been identified, any sort of coordinate transformation may be invoked according to convenience. An appreciably convenient transformation is, in fact,

$$\tan\frac{c\tau}{a}=\frac{ct/a}{1-ct/b},\qquad \boldsymbol{\rho}=\frac{\mathbf{r}}{\left[\left(\frac{ct}{a}\right)^2+\left(1-\frac{ct}{b}\right)^2\right]^{1/2}},$$

from $\mathbf{r},t$ to $\boldsymbol{\rho},\tau$, which carries the geodesics from $d^2\mathbf{r}/dt^2=0$ to the harmonic oscillator form

$$\frac{d^2\boldsymbol{\rho}}{d\tau^2}+\frac{c^2}{a^2}\boldsymbol{\rho}=0,$$

otherwise familiar in de Sitter space. Since τ depends on t alone, equal-t commutation rules go over simply to equal-τ ones. This gives a ladder spectrum of Klein–Gordon energy eigenvalues E with the correct[7] zero-point frequency c/a.

One last transformation,

$$\boldsymbol{\rho}=\frac{\mathbf{R}}{\left(1+\frac{R^2}{a^2}\right)^{1/2}}\quad(\tau\ \text{unchanged}),$$

brings a finally useful representation of the square of the Hamiltonian of a particle of mass m, running along a geodesic of de Sitter space, that springs out of the above inertially framed view of that space:

$$\frac{H^2}{c^2}=\left[\mathbf{P}^2+\frac{\mathbf{L}^2}{a^2}+\frac{\hbar^2}{a^2}\right]+\kappa^2\left[\frac{\hbar^2}{a^2}\left(1+\frac{\mathbf{R}^2}{a^2}\right)\right]=\left[\frac{H_1^2}{c^2}\right]+\kappa^2\left[\frac{H_2^2}{c^2}\right], \tag{3}$$

where

$$\mathbf{P}\equiv\frac{1}{2}\left(I+\frac{\mathbf{RR}}{a^2}\right)\cdot\mathbf{P}_c+\text{h.c.},\quad \mathbf{L}=\mathbf{R}\times\mathbf{P},$$

$$\kappa^2=\frac{m^2c^2a^2}{\hbar^2}-\frac{1}{4},$$

and $\mathbf{P}_c$ is the canonical mate $-i\hbar\boldsymbol{\nabla}$ to $\mathbf{R}$ stemming from the primitive commutation rules on the primitive geodesics $d^2\mathbf{r}/dt^2=0$. This places the floor $m^2\geqslant\frac{1}{4}\hbar^2/c^2a^2$ on masses so as to guarantee positive definite H^2.

III. EVOLUTION OF NONASSOCIATIVE ELEMENTS

The problem now is the same as that brought forth by Dirac over half a century ago in flat space: what is H when one knows the structure of H^2 from the fundamental commutation rules? The issue is unavoidable once the path of commutation rules and Hamiltonian structure is taken as fundamental in quantum theory $-H$, and not any function of it, is the master dynamical variable, and its character needs to be established if quantum physics is to be understood.

The central point is that *no Dirac square root of* H^2 *above exists* (except for $\kappa=0$). To see this, notice first that H_1 and H_2 separately are certainly Dirac calculable, for example,

$$H_1=\boldsymbol{\alpha}\cdot\mathbf{P}+\hat{\boldsymbol{\sigma}}\cdot\mathbf{L}-1,$$

$$H_2=\beta+\boldsymbol{\alpha}\cdot\mathbf{R}, \tag{4}$$

where, from here on, all of $\hbar,c,a=1$ and β, $\boldsymbol{\alpha}$ are standard Dirac matrices, with $\hat{\sigma}_x=i\alpha_y\alpha_z$, etc. These are not unique, merely illustrative of a large family.

Now H clearly must be linear in κ, and therefore must be of the form

$$H=AH_1+\kappa BH_2,$$

where A and B are suitable quantities placed into a Kronecker product (understood) with H_1 and H_2. Squaring produces ($\kappa\neq0$)

$$H^2=A^2H_1^2+\kappa(ABH_1H_2+BAH_2H_1)+\kappa^2B^2H_2^2,$$

so that, of course, A and B must be square roots of unity, while the cross terms must vanish.

It is easy to see that for H_1 and H_2 as in the example of Eq. (4), the cross-terms cannot cancel one another. Nor indeed can they cancel for any H_1 and H_2 whatever. For their cancellation, it is required that $H_1H_2=\pm H_2H_1$ or say kH_2H_1 and then it is enough that $AB=-kBA$ (i.e., A and B anticommuting or commuting). Multiplying on the left by H_1 and on the right by H_2 then gives

$$H_1^2H_2^2=k(H_1H_2)^2=k^3(H_2H_1)^2=k^3H_2(H_1H_2)H_1=k^3H_2(kH_2H_1)H_1=H_2^2H_1^2.$$

The separate squares H_1^2 and H_2^2 are unambiguous as given in Eq. (3), and a direct computation shows that H_1^2 and H_2^2 do not commute as required. In short, as the example, Eq. (4), clearly foretells, H_1H_2 and H_2H_1 are always so different as to disallow their cancellation: the cross-terms must *separately* vanish.

It is required in all, then, that

$$A^2=1=B^2, \tag{5}$$

$$AB=0=BA, \tag{6}$$

in order that H be extractable from H^2. Clearly, if A and B are elements of an associative algebra, conditions (5) and (6) are impossible. For associativity tells by (6) that $A(AB)=0$ means $(A^2)B=0$ or that $B=0$, and this contradicts $B^2=1$. Similarly, $(AB)B=0$ brings the contradiction that $A=0$ while $A^2=1$ is demanded. This argument rules out not only associative algebras $X(YZ)=(XY)Z$ but also those nonassociative algebras (such as the octonions) that are alternative, viz., $X(XY)=(X^2)Y,\ (XY)Y=X(Y^2)$.

This concludes the central point of the present paper: *The curvature of space–time necessarily induces a strictly nonassociative structure into the Hamiltonian for a particle going along a geodesic.* ($\kappa=0$ *excepted*).

IV. REALIZATION OF NONASSOCIATIVE ELEMENTS *A,B*

None of the familiar algebras—real numbers, complex numbers, quaternions,[8] octonions[9]—will do for *A,B*, as the first three are associative and the last alternative. But all after the first have a common root in the Cayley–Dickson doubling process,[10] wherein

- complex numbers are doublets of reals,
- quaternions are doublets of complexes,
- octonions are doublets of quaternions.

And now one further doubling, where the novel

- *decahexions* are doublets of octonions,

will be shown to answer to the requirements for *A* and *B*. The decahexions, which evidently have seen negligible development heretofore, probably for want of any applications, are the lowest in the Cayley–Dickson (infinite) hierarchy admitting explicit formation of *A* and *B*.

The members of a Cayley–Dickson algebra are of the form of ''hypercomplex'' numbers,

$$X\equiv x_0e_0+x_1e_1+x_1e_2+\cdots+x_ne_n\,,$$

with e_0 being an idemfactor (usually replaced by 1), and $e_i^2=-1$ $(i=1,2,3,...,n)$, and a suitable multiplication table for the basic elements $e_1,e_2,\ldots$ featuring anticommutativity $e_ie_j=-e_je_i$ (but $e_ie_0=e_0e_i=e_i$). The Cayley–Dickson hypercomplex numbers enjoy physically congenial properties, such as

$$\text{a conjugate}\quad \overline{X}=x_0-x_1e_1-x_2e_2-\cdots-x_ne_n,$$

$$\text{a trace}\quad \tfrac{1}{2}(X+\overline{X})=x_0,$$

$$\text{a norm}\quad X\overline{X}=\overline{X}X=x_0^2+x_1^2+x_2^2+\cdots+x_n^2,\qquad \text{and}$$

$$\text{a product conjugation}\quad \overline{XY}=\overline{Y}\,\overline{X}.$$

Usually the coefficients x_i are taken to be real, but occasionally may be taken to be imaginary or complex.

The doubling rule is as follows: If $X_1,\ X_2,\ Y_1,\ Y_2$ are members of one algebra, the doubled algebra consists of pairs such as $(X_1|X_2),\ (Y_1|Y_2)$, etc., with the product of doublets following the rule

$$(X_1|X_2)(Y_1|Y_2)\equiv(X_1Y_1-Y_2\overline{X}_2|\overline{X}_1Y_2+Y_1X_2)\equiv(Z_1|Z_2). \tag{7}$$

For example, for X,Y themselves simple two-element doublets,

$$X_1=(x_0,x_1)=x_0+e_1x_1,$$

$$X_2=(x_3,x_2)=x_3+e_1x_2,$$

$$Y_1=(y_0,y_1)=y_0+e_1y_1,$$

$$Y_2=(y_3,y_2)=y_3+e_1y_2,$$

one finds

$$Z_1=x_0y_0-x_1y_1-x_2y_2-x_3y_3+e_1(x_1y_0+x_0y_1+x_2y_3-x_3y_2),$$

$$Z_2=x_0y_3+x_3y_0+x_1y_2-x_2y_1+e_1(x_0y_2+x_2y_0+x_3y_1-x_1y_3). \tag{8}$$

In short, one gets a representation of quaternions whose standard format (with $e_3e_1=e_2$ cyclically),

$$Q_x=x_0+x_1e_1+x_2e_2+x_3e_3, \quad \text{or} \quad x_0+[\mathbf{x}],$$

$$Q_y=y_0+y_1e_1+y_2e_2+y_3e_3, \quad \text{or} \quad y_0+[\mathbf{y}],$$

with $\mathbf{x}$, $\mathbf{y}$ three-dimensional vectors (x_1,x_2,x_3), (y_1,y_2,y_3), may equally be written as

$$Q_x=(x_0+x_1e_1)+e_3(x_3+x_2e_1),$$

$$Q_y=(y_0+y_1e_1)+e_3(y_3+y_2e_1),$$

so as to show exact correspondence with the above Cayley–Dickson doublet of doublets. The convenient condensed multiplication rule,

$$Q_z=Q_xQ_y=x_0y_0-\mathbf{x}\cdot\mathbf{y}+[x_0\mathbf{y}+\mathbf{x}y_0+\mathbf{x}\times\mathbf{y}], \tag{9}$$

and the Cayley–Dickson rule (8) simply rephrase each other. The quaternion units e_1, e_2, e_3 can be represented through Pauli spin matrices σ_1, σ_2, σ_3 according to $e_k=-i\sigma_k$.

Next come doublets of quaternions, that is, octonions. In the Cayley–Dickson scheme,

$$\mathcal{O}_x=(x_0,x_1,x_2,x_3|x_7,x_4,x_5,x_6),$$

or, introducing the element e_7 (paralleling e_3 for quaternions),

$$\mathcal{O}_x=x_0+e_1x_1+e_2x_2+e_3x_3+e_7(x_7+x_4e_1+x_5e_2+x_6e_3).$$

This is, under the (partial) multiplication rules $e_7e_1=e_4$, $e_7e_2=e_5$, $e_7e_3=e_6$, the conventional representation,

$$\mathcal{O}_x=x_0+e_1x_1+e_2x_2+e_3x_3+x_4e_4+x_5e_5+x_6e_6+x_7e_7$$

(paralleling the conventional Q_x), with a product rule $\mathcal{O}_x\mathcal{O}_y$ that is easily worked out from the Cayley–Dickson fundament, Eq. (7).

Finally, the decahexions as Cayley–Dickson doublets of octonions are

$$\mathcal{D}=(\mathcal{O}_1|\mathcal{O}_2).$$

It is more convenient, however, to use the simple and associative quaternions as a subunit, representing each octonion in $\mathcal{D}$ as a pair of quaternions, so that $\mathcal{D}$ is equivalently a quartet of quaternions (or quatroquaternion):

$$\mathcal{D}_U=(U_1|U_2|U_3|U_4).$$

To obtain on this basis the multiplication rule for decahexions, we need only apply the Cayley–Dickson doubling rule twice in succession—once using $\mathcal{D}$ as an octonion pair, and then, in the resulting products of octonions, using each octonion as a quaternion pair. This gives the quadrupling rule, quite generally, as

$$\begin{gathered}\mathcal{D}_U\mathcal{D}_V=(U_1|U_2|U_3|U_4)(V_1|V_2|V_3|V_4)=(W_1|W_2|W_3|W_4),\\ W_1=U_1V_1-V_2\overline{U}_2-V_3\overline{U}_3-U_4\overline{V}_4,\\ W_2=\overline{U}_1V_2+V_1U_2+\overline{V}_3U_4-\overline{U}_3V_4,\\ W_3=\overline{U}_1V_3+V_4\overline{U}_2+V_1U_3-U_4\overline{V}_2,\\ W_4=U_1V_4-V_3U_2+\overline{V}_1U_4+U_3V_2.\end{gathered}\tag{10}$$

There are many ways now to attend to the algebraic requirements of Eqs. (5) and (6). One of the simplest is as follows. With A,B written as $(A_1|A_2|A_3|A_4)$ and $(B_1|B_2|B_3|B_4)$, take all the A_i quaternions to be traceless and to have vector parts that are collinear, say, along the unit three-dimensional sector $\hat{\mathbf{m}}$,

$$A_i=0+[\alpha_i\hat{\mathbf{m}}],\quad i=1,2,3,4.$$

Similarly, choose

$$B_i=0+[\beta_i\hat{\mathbf{n}}],$$

and, further, choose $\hat{\mathbf{m}}$ and $\hat{\mathbf{n}}$ to be orthogonal, $\hat{\mathbf{m}}\cdot\hat{\mathbf{n}}=0$. Then in the quaternion products occurring in AB according to Eq. (10), each product is of the form $0+[(\alpha\beta)\hat{\mathbf{m}}\times\hat{\mathbf{n}}]$, i.e., each resultant quaternion W_i, like A_i and B_i, is traceless and has vector parts all collinear to $\hat{\mathbf{m}}\times\hat{\mathbf{n}}$. One reads off at once,

$$AB=\begin{pmatrix}W_1\\W_2\\W_3\\W_4\end{pmatrix}=0+\begin{pmatrix}\alpha_1&-\alpha_2&-\alpha_3&\alpha_4\\-\alpha_2&-\alpha_1&\alpha_4&\alpha_3\\-\alpha_3&\alpha_4&-\alpha_1&\alpha_2\\\alpha_4&\alpha_3&\alpha_2&\alpha_1\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\end{pmatrix}[\hat{\mathbf{m}}\times\hat{\mathbf{n}}].$$

Now, choose $\alpha_1,\alpha_2,\alpha_3,\alpha_4=(\xi,\eta,-\eta,\xi)$, so that $AB=0$ means

$$\begin{pmatrix}\xi&-\eta&\eta&\xi\\-\eta&-\xi&\xi&-\eta\\\eta&\xi&-\xi&\eta\\\xi&-\eta&\eta&\xi\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\end{pmatrix}=(M)\beta=0.$$

The matrix M is of rank two and possesses two null vectors, $\mathrm{col}(x,0,0,-x)$ and $\mathrm{col}(0,y,y,0)$. Consequently, $\beta=\mathrm{col}(\beta_1,\beta_2,\beta_3,\beta_4)$ can be written as $\mathrm{col}(x,y,y,-x)$. And now $BA=0$ means similarly

$$\begin{pmatrix} x & -y & -y & -x \\ -y & -x & -x & y \\ -y & -x & -x & y \\ -x & y & y & x \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = 0,$$

and this is satisfied by $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (\xi, \eta, -\eta, \xi)$.

The squares of A and B are now simply the scalars

$$A^2 = -2(\xi^2 + \eta^2) e_0,$$

$$B^2 = -2(x^2 + y^2) e_0,$$

so it suffices to take all of ξ, η, x, y to be purely imaginary to achieve $A^2 = 1 = B^2$:

$$\xi = i\gamma, \quad \eta = i\delta, \quad x = iu, \quad y = iv,$$

$$\gamma^2 + \delta^2 = \tfrac{1}{2}, \quad u^2 + v^2 = \tfrac{1}{2}.$$

That is, we obtain two one-parameter families of decahexions,

$$\gamma = \frac{1}{\sqrt{2}} \cos\theta, \quad \delta = \frac{1}{\sqrt{2}} \sin\theta \quad \text{and} \quad u = \frac{1}{\sqrt{2}} \cos\varphi, \quad v = \frac{1}{\sqrt{2}} \sin\varphi.$$

In summary, the decahexions, with orthogonal $\hat{\mathbf{m}}$ and $\hat{\mathbf{n}}$,

$$A = (A_1 | A_2 | A_3 | A_4) = (i\gamma[\hat{\mathbf{m}}] | i\delta[\hat{\mathbf{m}}] | -i\delta[\hat{\mathbf{m}}] | i\gamma[\hat{\mathbf{m}}]),$$

$$B = (B_1 | B_2 | B_3 | B_4) = (iu[\hat{\mathbf{n}}] | iv[\hat{\mathbf{n}}] | iv[\hat{\mathbf{n}}] | -iu[\hat{\mathbf{n}}]), \tag{11}$$

and normalization as above, meet the fundamental algebraic requirements of Eqs. (5) and (6). The occurrence of imaginary elements in the quaternionic components can be interpreted as a shift from the quaternion basis e_1, e_2, e_3 to the basis ie_1, ie_2, ie_3, that is, to Pauli spin matrices σ_1, σ_2, σ_3, as noted previously.

V. ISSUES OF QUANTIZATION

Since Albert's theorem,[11] it has been known that a probability interpretation, conforming to standard inherited precepts for quantum theory, has to rest on the algebra of real numbers, or complex numbers, or quaternions, or octonions. The first three can form bases for a Hilbert space formulation of quantum mechanics in which state vectors obey a superposition principle with, respectively, real, complex, or quaternionic coefficients, as in ordinary standard (complex) quantum theory. But the octonions, together with all other nonassociative systems, are ruled out, since, for example, scalar products like

$$\int \bar{\theta} N \phi \, d\mathbf{R},$$

are ambiguous, owing to the distinction between $(\overline{\theta} N)\phi$ and $\overline{\theta}(N\phi)$ with nonassociative θ, N, ϕ, an omnipresent ambiguity for all triadic products.

The decahexions are ruled out, not only because of their nonassociativity, cutting even more deeply than for octonions, but also because they do not form a division algebra (one for which $XY \neq 0$ when $X \neq 0$ and $Y \neq 0$—the above A,B by construction admit $AB = 0$ for nonzero A and B).

In short, *there appears to be an inevitable clash between the rudiments of space–time curvature* (as locally embraced by de Sitter space) *and the rudiments of quantum theory, as ordinarily understood*. This clash has not to do with general relativity in itself, but simply with curvature as such as reflected in the Hamiltonian structure for the geodesics.

Clash notwithstanding, it remains of interest to see what shape a Schrödinger wave equation takes,

$$i\hbar \frac{\partial \psi}{\partial \tau} = H\psi, \tag{12}$$

where ψ lies in the same algebra as $H = AH_1 + \kappa BH_2$, while $i\hbar\,\partial/\partial\tau$ signifies still that H generates time translations. The simplest construction is over the miniature algebra A,B,O (null), and I (idemfactor), irrespective of any explicit representation of A,B. First, iteration, in general, brings

$$-\hbar^2 \frac{\partial^2}{\partial \tau^2}\psi = H(H\psi),$$

but, of course,

$$H(H\psi) \neq (H^2)\psi = (H_1^2 + \kappa^2 H_2^2)\psi.$$

Next, place

$$\psi = I\psi_0 + A\psi_1 + B\psi_2,$$

in Eq. (12) and equate separate I,A,B components:

$$i\hbar\dot{\psi}_0 = H_1\psi_1 + \kappa H_2\psi_2,$$

$$i\hbar\dot{\psi}_1 = H_1\psi_0,$$

$$i\hbar\dot{\psi}_2 = \kappa H_2\psi_0.$$

Using a spin-tower col(ψ_0,ψ_1,ψ_2), this is

$$i\hbar\begin{pmatrix}\dot{\psi}_0\\ \dot{\psi}_1\\ \dot{\psi}_2\end{pmatrix} = \begin{pmatrix} 0 & H_1 & \kappa H_2\\ H_1 & 0 & 0\\ \kappa H_2 & 0 & 0\end{pmatrix}\begin{pmatrix}\psi_0\\ \psi_1\\ \psi_2\end{pmatrix}.$$

Iteration gives

$$-\hbar^2\begin{pmatrix}\ddot{\psi}_0\\ \ddot{\psi}_1\\ \ddot{\psi}_2\end{pmatrix} = \begin{pmatrix} H_1^2+\kappa^2 H_2^2 & 0 & 0\\ 0 & H_1^2 & \kappa H_1 H_2\\ 0 & \kappa H_2 H_1 & H_2\end{pmatrix}\begin{pmatrix}\psi_0\\ \psi_1\\ \psi_2\end{pmatrix},$$

so that, as expected, the scalar element ψ_0 satisfies the Klein–Gordon equation. There is a differential conservation law here [using row $(\psi_0^\dagger,\psi_1^\dagger,\psi_2^\dagger)$, and H_1 from Eq. (4)],

$$\frac{\partial\rho}{\partial\tau} + \nabla_{\mathbf{R}}\cdot\mathbf{j} = 0,$$

$$\rho \equiv \psi_0^\dagger\psi_0 + \psi_1^\dagger\psi_1 + \psi_2^\dagger\psi_2,$$

$$\mathbf{j} \equiv \psi_0^\dagger[\boldsymbol{\alpha}\cdot(I+\mathbf{RR}) + \hat{\boldsymbol{\sigma}}\times\mathbf{R}]\psi_1 + \psi_1^\dagger[\boldsymbol{\alpha}\cdot(I+\mathbf{RR}) + \hat{\boldsymbol{\sigma}}\times\mathbf{R}]\psi_0.$$

While the spin towers thus appear to admit a touch of associativity, this falsifies the fundamental difficulties of the underlying and irremediable nonassociativity, for which $\theta(N\varphi)$ is distinct from $(\theta N)\varphi$ (operator N being the triplet $N_0+AN_1+BN_2$ with operator N_i, and vectors θ and φ being triplets $\theta_0+A\theta_1+B\theta_2$ and $\varphi_0+A\varphi,+B\varphi_2$). In fact, direct calculation gives

$$\theta(N\varphi)-(\theta N)\varphi=A(\theta_1N_2\varphi_2-\theta_2N_2\varphi_1)+B(\theta_2N_1\varphi_1-\theta_1N_1\varphi_2),$$

which, of course, is out of sight from any 3×3 matrix operator in N_0,N_1,N_2 and spin towers in $\theta_0,\theta_1,\theta_2$ and $\theta_0,\theta_1,\theta_2$.

Going now to the decahexion representation for A and B, a decahexion format for ψ must also be introduced, or in quatroquaternion terms,

$$\psi=(\psi_1|\psi_2|\psi_3|\psi_4).$$

Here each ψ_k is a quaternion, with coefficients that are Dirac spinors that are to be acted upon finally by H_1 and H_2:

$$\psi_k=\psi_{k0}e_0+\psi_{k1}e_1+\psi_{k2}e_2+\psi_{k3}e_3,$$

$$\psi_{ku}=\begin{pmatrix}\psi_{ku1}\\ \psi_{ku2}\\ \psi_{ku3}\\ \psi_{ku4}\end{pmatrix}.$$

In all, the components ψ_{kuv} are $4\times4\times4=64$ in number (quartets of quartets of quartets). The wave equation reads

$$\overline{i\hbar}(\dot\psi_1|\dot\psi_2|\dot\psi_3|\dot\psi_4)=H_1(A_1|A_2|A_3|A_4)(\psi_1|\psi_2|\psi_3|\psi_4)+\kappa H_2(B_1|B_2|B_3|B_4)(\psi_1|\psi_2|\psi_3|\psi_4),$$

where, using the fundamental multiplication rule, Eq. (10),

$$(A_1|A_2|A_3|A_4)(\psi_1|\psi_2|\psi_3|\psi_4)=(A_1\psi_1-\psi_2\overline{A_2}-\psi_3\overline{A_3}-A_4\overline{\psi_4}|\overline{A_1}\psi_2+\psi_1A_2+\overline{\psi_3}A_4-\overline{A_3}\psi_4|\overline{A_1}\psi_3$$
$$+\psi_4\overline{A_2}+\psi_1A_3-A_4\overline{\psi_2}|A_1\psi_4-\psi_3A_2+\overline{\psi_1}A_4+A_3\psi_2)$$

(similarly for $B\psi$ by replacing A_k by B_k).

The scheme may be viewed as an ensemble of interlocking quaternionic wave equations. They contain not only the quaternions $\psi_1,\psi_2,\psi_3,\psi_4$ but also terms in their (quaternionic) conjugates $\overline{\psi_1},\overline{\psi_2},\overline{\psi_3},\overline{\psi_4}$. These terms, mandated by the decahexion algebra, are of the nature of antilinear elements akin to complex conjugates in ordinary quantum mechanics.

We may finally turn away from Schrödinger models of quantum theory toward Heisenberg models. The power and primitiveness of the latter far surpass the former, since the central proposition is at its core simply the fundamental Heisenberg equations of motion, without a glance at a superposition principle, or Hilbert space, or even a probability interpretation. The dynamical law for any quantity F is the celebrated

$$\frac{dF}{d\tau}=\frac{1}{i\hbar}(F,H)=\frac{1}{i\hbar}(FH-HF).$$

This is compatible with $d(FG)/d\tau=(dF/d\tau)G+F(dG/d\tau)$ only if the underlying algebra is flexible Lie admissable,[9] which the present decahexion algebra is not. This does not prevent the sequence of derivatives $dF/d\tau,d^2F/d\tau^2,\ldots,$ from being computed and yielding a Maclaurin series expansion for F, as below. H cannot be diagonalized since H_1,H_2 do not commute.

For simplicity, we may limit the discussion to the miniature algebra A,B,O,I instead of the full decahexion algebra. The general (time-independent) dynamical variable is

$$F = F_0 + AF_1 + BF_2,$$

where F_i are functions of coordinates **R** and momenta **P** as well as Dirac matrices. Then F is a conserved quantity when (F,H) vanishes,

$$(F,H) = (F_1,H_1) + (F_2,H_2) + A(F_0,H_1) + \kappa B(F_0,H_2) = 0,$$

that is, when

$$(F_1,H_1) + (F_2,H_2) = 0,$$

$$(F_0,H_1) = 0,$$

$$(F_0,H_2) = 0.$$

It is quite possible now to have conserved quantities that involve arbitrary functions. An example is $F = BF_2(\mathbf{R})$ with arbitrary $F_2(\mathbf{R})$, since H_2, being independent of **P**, commutes with any $F_2(\mathbf{R})$. Thus, also $AH_1 + BF_2(\mathbf{R})$ is conserved.

The conventional Heisenberg rendition for $F(\tau)$,

$$F(\tau) = e^{(i/\hbar)H\tau} F e^{(-i/\hbar)H\tau},$$

will not work for nonassociative $H = AH_1 + \kappa BH_2$, because H is not power associative, e.g.,

$$H(H^2) = (AH_1 + \kappa BH_2)(H_1^2 + \kappa^2 H_2^2) = AH_1^3 + \kappa^2 AH_1H_2^2 + \kappa BH_2H_1^2 + \kappa^3 BH_2^3,$$

while $(H^2)H$ is different,

$$(H^2)H = AH_1^3 + \kappa^2 AH_2^2H_1 + \kappa BH_1^2H_2 + \kappa^3 BH_2^3.$$

Consequently the power series for $\exp((i/\hbar)H\tau)$ is ill defined.

Instead, computing successive derivatives of, say, $F_0(\mathbf{R},\mathbf{P})$, yields

$$\frac{dF_0}{d\tau} = \frac{1}{i\hbar}(F_0, AH_1 + \kappa BH_2) = \frac{1}{i\hbar}(A(F_0,H_1) + \kappa B(F_0,H_2)) = \frac{1}{i\hbar}(AD_1F_0 + BD_2F_0), \quad (13)$$

where the separate commutators are written as

$$D_1F_0 \equiv F_0H_1 - H_1F_0, \quad D_2F_0 \equiv \kappa(F_0H_2 - H_2F_0).$$

Then

$$\frac{d^2F_0}{d\tau^2} = \frac{1}{(i\hbar)^2}[D_1^2F_0 + D_2^2F_0],$$

$$\frac{d^3F_0}{d\tau^3} = \frac{1}{(i\hbar)^3}[AD_1(D_1^2 + D_2^2)F_0 + BD_2(D_1^2 + D_2^2)F_0],$$

$$\frac{d^4F_0}{d\tau^4} = \frac{1}{(i\hbar)^4}[D_1^2(D_1^2 + D_2^2)F_0 + D_2^2(D_1^2 + D_2^2)F_0],$$

$$= \frac{1}{(i\hbar)^4}[(D_1^2 + D_2^2)^2F_0], \quad \text{etc.}$$

More simply, writing the final result for $F_0(\tau)$ as

$$F_0(\tau)=f_0+Af_1+Bf_2,$$

gives, on the one hand,

$$\frac{dF_0(\tau)}{d\tau}=\frac{df_0}{d\tau}+A\frac{df_1}{d\tau}+B\frac{df_2}{d\tau},$$

and on the other hand,

$$\frac{dF_0}{d\tau}=\frac{1}{i\hbar}(f_0+Af_1+Bf_2,AH_1+\kappa BH_2)=\frac{1}{i\tau}(D_1f_1+D_2f_2+AD_1f_0+BD_2f_0).$$

Thereupon follows the triplet:

$$\frac{df_0}{d\tau}=\frac{1}{i\hbar}(D_1f_1+D_2f_2)$$

$$\frac{df_1}{d\tau}=\frac{1}{i\hbar}D_1f_0$$

$$\frac{df_2}{d\tau}=\frac{1}{i\hbar}D_2f_0.$$

By differentiating the first and using the other two,

$$\frac{d^2f_0}{d\tau^2}=\frac{1}{(i\hbar)^2}(D_1^2+D_2^2)f_0=-\left(\frac{D_1^2+D_2^2}{\hbar^2}\right)f_0,$$

so that f_0 can be written symbolically as

$$f_0(\tau)=\cos\left(\frac{\sqrt{D_1^2+D_2^2}\tau}{\hbar}\right)f_0=\cos(\Omega\tau)f_0,$$

and then

$$f_1(\tau)=\frac{1}{i\hbar}\left(D_1\frac{1}{\Omega}\sin\Omega\tau\right)f_0,$$

$$f_2(\tau)=\frac{1}{i\hbar}\left(D_2\frac{1}{\Omega}\sin\Omega\tau\right)f_0.$$

The meaning is that in the power series expansions of cosine and sine the individual terms recover the Maclaurin series,

$$F_0(\tau)=F_0+\frac{dF_0}{d\tau}\tau+\frac{d^2F_0}{d\tau^2}\frac{\tau^2}{2!}+\ldots,$$

built directly from the derivatives of F_0 as, given above.

In a similar fashion, the time evolution of AF_1 and of BF_2 may be obtained. The Heisenberg equations of motion, in short, are equal to describing in full the quantal time-evolution of dynamical variables using multiple commutators, just as Hamilton's equations of motion describe classical time evolution using multiple Poisson brackets.

Finally, it should be remarked that the nonassociative elements A and B do not appear to be observationally directly accessible, though their guiding hand may be revealed covertly, for instance, in the even-ordered derivatives $d^2F_0/d\tau^2$, $d^4F_0/d\tau^4$, etc., above, which are independent of A and B: or in squares like

$$\left(\frac{dF_0}{d\tau}\right)^2 = -\frac{1}{\hbar^2}[(D_1F_0)^2+(D_2F_0)^2]$$

[similarly to how the velocity $c\alpha$ in Dirac theory is not open to direct observation while $(c\alpha)^2 = c^2 = (\text{velocity})^2$ is].

VI. CONCLUSION

We may summarize as follows. The geodesics in a curved space—locally an osculating de Sitter space, which then is prototypical—admit description as uniform straight-line motion with respect to preferred frames of reference that are inertial frames interconnected by projective (fractional-linear) transformation. The preferred frames overcome the ambiguity of commutation rules and Hamiltonian structure inherent under general covariance. But the construction of the Hamiltonian then requires nonassociative elements, which can be realized as those of a Cayley–Dickson algebra having 16 basis elements (doubled octonions, or decahexions). The nonassociativity produces a serious conflict with a Hilbert space and probabilistic rendition of quantum mechanics like the standard type. The formal statement of a Schrödinger wave equation necessitates antilinear as well as linear terms. Nevertheless, the Heisenberg laws of motion can be unambiguously formulated, leading to a full rendition of the quantal time evolution of any dynamical variable (while the Heisenberg program of diagonalizing the Hamiltonian cannot be carried out). The nonassociative elements appear not to be directly observable, but they show themselves indirectly in terms in the time evolution of a dynamical variable.

[1] E. Schrödinger, Sitzungsber. K. Preuss. Akad. Wiss. 105 (1932).
[2] V. Bargmann, Sitzungsber. K. Preuss. Akad. Wiss. 336 (1932).
[3] N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Space* (Cambridge University Press, Cambridge, 1982).
[4] C. J. Isham, in *Proceedings of the 8th Texas Symposium on Relativistic Astrophysics*, edited by M. D. Papagrannis (New York Academy of Science, New York, 1977); A. Salam and J. Strathdee, Phys. Rev. D **18**, 4596 (1978); S. J. Avis, C. J. Isham, and D. Storey, *ibid.* **18**, 3565 (1978); C. Fronsdal, *ibid.* **12**, 3819 (1975).
[5] M. D. Maia, J. Math. Phys. **22**, 538 (1981).
[6] E. H. Kerner, Phys. Rev. D **22**, 280 (1980) (referred to as III).
[7] See Ref. 4 (Salam and Strathdee, and Avis, Isham, and Storey) and Ref. 6 (III).
[8] S. L. Adler, *Quaternionic Quantum Mechanics* (Oxford University Press, New York, 1995).
[9] S. Okubo, *Introduction to Octonion and Other Non-Associative Algebras in Physics* (Cambridge University Press, New York, 1995).
[10] R. D. Schafer, *An Introduction to Nonassociative Algebras* (Academic, New York, 1966).
[11] A. A. Albert, Ann. Math. **48**, 495 (1947).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Absence of confinement in the absence of vortices

Tamás G. Kovács[a]
Department of Physics, University of Colorado, Boulder Colorado 80309-0390

E. T. Tomboulis[b]
Department of Physics, UCLA, Los Angeles, California 90095-1547

(Received 9 February 1999; accepted for publication 14 June 1999)

We consider the Wilson loop expectation in $SU(2)$ lattice gauge theory in the presence of constraints. The constraints eliminate gauge field configurations, which, in physical terms, allow the presence of thick center vortices linking with the loop. We prove that, for dimension $d \geqslant 3$, the so-constrained Wilson loop follows perimeter law, i.e., nonconfining behavior, at weak coupling (low temperature). © *1999 American Institute of Physics.* [S0022-2488(99)02709-7]

I. INTRODUCTION

The precise physical mechanism(s) by which $SU(N)$ gauge theories in their critical dimension apparently avoid a phase transition and thus remain in a confining phase for arbitrarily weak coupling is currently receiving renewed attention (see, e.g., Ref. 1). The proposal that extended thick vortices are the configurations responsible for this behavior has been strongly supported by recent numerical simulations in the case of the $SU(2)$[2–4] and $SU(3)$[5] gauge groups. The string tension of the full Wilson loop at weak coupling is found in these simulations to be fully reproduced solely by the contribution of thick vortices. In this paper we give a proof of the converse statement, also suggested by the simulations. We consider the Wilson loop in the presence of constraints eliminating field configurations, which, in physical terms, allow thick vortices linking with the loop. We then show that the so-constrained Wilson loop exhibits nonconfining (perimeter law) behavior at weak coupling.

Rigorous consideration of the vortex mechanism of confinement was initiated in Ref. 6. There, in addition to developing a precise formulation on the lattice, a sufficiency criterion for confinement was derived in terms of the behavior of vortices enclosed in ''vortex containers'' linking with the Wilson loop. Furthermore, it was proven that in the presence of constraints eliminating all $Z(2)$ monopoles in the theory, the t'Hooft disorder operator exhibits nonconfining (i.e., area law) behavior at weak coupling. This operator provides an external color magnetic source, which, in the absence of such monopoles, can no longer be screened. As already pointed out in Ref. 6, however, elimination of the $Z(2)$ monopoles does not eliminate the presence of closed thick magnetic vortices, which may still link with and give confining behavior to the Wilson loop operator. A result pertinent to this question was subsequently obtained in Ref. 7. There it was proven that, in the presence of constraints eliminating vortices winding around the (periodic) lattice, the electric-flux-free energy order parameter exhibits nonconfining behavior at large β. This observable has been proven rigorously[8] only to form an upper bound on the Wilson loop. Thus, it only provides a sufficient criterion for confinement: confining behavior of the electric-flux-free energy implies confining behavior for the Wilson loop, but not the converse. At any rate, the electric flux is an order parameter that refers to the entire lattice. It is clearly important to obtain a statement concerning the effect of the constraints eliminating vortices on the Wilson loop itself, since it directly represents the actual potential between two external sources (quarks), as the lattice is taken to the thermodynamic limit. This is the question addressed in this paper.

[a]Electronic mail: kovacs@eotvos.Colorado.EDU
[b]Electronic mail: tombouli@physics.ucla.edu

The distinction between ''thick,'' ''thin,'' and ''hybrid'' vortices is crucial to the physical picture of the action of vortices at small gauge coupling (large β).[2] We cannot adequately review the underlying physical picture here; we refer to Ref. 2 for a recent physical discussion, as well as earlier references to these ideas. This paper is devoted to the proof of nonconfining behavior for the Wilson loop in the absence of linked vortices. An appropriate framework for these considerations is obtained by expressing the $SU(N)$ lattice gauge theory in terms of new separate $Z(N)$ and $SU(N)/Z(N)$ variables.[6,9] We employ this formulation in Sec. II, where we also consider a variety of alternative lattice actions for which we subsequently obtain our result. We also introduce the appropriate constraints for eliminating linking thick and hybrid vortices. The proof that the so-constrained Wilson loop follows perimeter law is given in Sec. III. We explicitly consider only the case of $SU(2)$, which exhibits all the relevant physical features of the general N case.

Notation: We work on a simple hypercubic lattice $\Lambda \subset \mathbf{Z}^d$ of size $L_\mu = 2^{N_\mu}$, integer N_μ, and lattice coordinates integer n^μ, $\mu = 1,...,d$. Elementary cells c^r, $(r = 0,1,...,d)$, in Λ will be denoted more explicitly as s (sites), b (bonds), p (plaquettes), c (3-cubes), etc. Each r-cell is assigned an orientation; $-c^r$ then denotes the oppositely oriented cell. The number of cells in a collection S of r-cells will be denoted by $|S|$.

We employ the standard formalism of lattice gauge theory. With each bond we associate a copy G_b of the gauge group G. The gauge field U_b is an element of G_b, with the assignment $U_{-b} = U_b^{-1}$. The configuration space is then $\Omega_\Lambda = \otimes_{b \in \Lambda} G_b$, and a field configuration $U_\Lambda = \{U_b\}_{b \in \Lambda}$ is an element of Ω_Λ. In the following we do not distinguish explicitly between the abstract group element U_b and its matrix representation, which, unless otherwise indicated, will always be the fundamental representation.

We will introduce various lattice variables, taking values in the center of the group G. If K is an Abelian group, multiplicatively written, a K-valued r-form α is the map α: $c^r \to \alpha[c^r] \in K$, with $\alpha[-c^r] = \alpha[c^r]^{-1}$. The exterior difference operator is then defined by

$$(d\alpha)[c^{r+1}] = \prod_{c^r \in \partial c^{r+1}} \alpha[c^r]. \tag{1.1}$$

Given any set Q of c^k cells, we generally employ the shorthand notation

$$\alpha[Q] = \alpha\left[\prod_{k \in Q} c^k\right] = \prod_{k \in Q} \alpha[c^k] \equiv \alpha_Q. \tag{1.2}$$

Thus, e.g., for a 1-form γ we write γ_b, $(d\gamma)_p$, and so on. In this paper $G = SU(2)$, $K = Z(2)$.

II. ACTIONS AND CONSTRAINTS

The Euclidean functional measure of the standard lattice $SU(2)$ theory is given by

$$d\mu_\Lambda(U) = Z_\Lambda^{-1} \prod_{b \in \Lambda} dU_b \exp\left(\sum_p A_p(U_p)\right), \tag{2.1}$$

where dU_b denotes normalized Haar measure on $SU(2)$, and A_p the plaquette action, which is a function of $U_p = \Pi_{b \in \partial p} U_b$, the product of bond variables U_b around the plaquette p. The partition function Z_Λ is defined by $\int d\mu_\Lambda(U) = 1$, and for any observable $F(U)$, i.e. (complex-valued) function on Ω_Λ,

$$\langle F(U) \rangle = \int d\mu_\Lambda(U) F(U). \tag{2.2}$$

The usual minimal plaquette action is the Wilson action,

$$A_p(U_p) = \beta \operatorname{tr} U_p, \tag{2.3}$$

which is a special case of the action,

$$A_p(U_p) = \beta \operatorname{tr} U_p + \lambda \operatorname{sign} \operatorname{tr} U_p . \tag{2.4}$$

(2.4) extrapolates between the Wilson action ($\lambda=0$) and the ''positive plaquette model'' action ($\lambda\to\infty$). Another action we consider in this paper is

$$A_p(U_p) = \beta \operatorname{tr} U_p + \ln(\theta(|\operatorname{tr} U_p| - k)), \tag{2.5}$$

where $\theta(x)=1$ if $x>0$, 0 if $x<0$, and $0<k<2$; in particular, k a constant, or any function of β such that $k\beta\to\infty$ as $\beta\to\infty$, e.g., $k(\beta)=\beta^{-1/2}$. All these actions have, of course, the same naive continuum limit. More generally, they are expected to be physically equivalent for sufficiently large β, where each plaquette becomes highly peaked around the perturbative vacuum $\operatorname{tr} U_p \to \operatorname{tr} 1$. Use of such alternative actions provides a check of the requirement that long distance physics should not depend on short distance details, such as the precise form of the latticized action. We will first prove our result working with the action (2.5) since, as it turns out, it allows a rather simpler proof of the result. We will then obtain the result for the more standard, but in fact physically equivalent, action (2.4).

To formulate our argument we rewrite the $SU(2)$ theory (2.1) in the $SO(3)\times Z(2)$ form.[6,9] Consider the configuration space Ω_Λ split into equivalence classes, each class the coset bond variable configuration $\hat{U}_\Lambda = \{\hat{U}_b\}_{b\in\Lambda}$, $\hat{U}_b \in SU(2)/Z(2) \sim SO(3)$. Thus, two configurations U_Λ, $U'_\Lambda \in \Omega_\Lambda$ are representatives of the same coset configuration $\hat{U}_\Lambda$ iff one has $U'_b = U_b \gamma_b$, for some $\gamma \in Z(2)$, $\forall b \in \Lambda$. Let $\eta_p(U) \equiv \operatorname{sign} \operatorname{tr} U_p$. Then

$$(d\eta)_c(\hat{U}) = \prod_{p\in\partial c} \eta_p \tag{2.6}$$

depends, as indicated, only on the coset configuration, since it is invariant under $U_b \to U_b \gamma_b$, $\gamma \in Z(2)$. Let σ be a $Z(2)$-valued 2-form on Λ. Adopt periodic boundary conditions (b.c.), and let $\{S^\alpha\}$, $\alpha=1,...,\binom{d}{2}$, be a set of 2-dimensional nonbounding closed surfaces forming a 2-cycle basis on $\Lambda = T^d(N_1,\ldots,N_d)$. Then the $SU(2)$ theory (2.1)–(2.2) can be expressed in the $SO(3) \times Z(2)$ form given by:

$$d\mu_\Lambda(U,\sigma) = Z_\Lambda^{-1} \prod_{b\in\Lambda} dU_b \prod_{p\in\Lambda} d\sigma_p \prod_{c\in\Lambda} \delta((d\eta)_c(d\sigma)_c) \prod_\alpha \delta(\eta[S^\alpha]\sigma[S^\alpha])$$
$$\times \exp\left(\sum_{p\in\Lambda} A_p(|\operatorname{tr} U_p|, \sigma_p) \right). \tag{2.7}$$

In (2.7), $\int d\sigma_p(\cdots) \equiv \frac{1}{2}\Sigma_{\sigma_p=\pm 1}(\cdots)$ denotes Haar measure over $Z(2)$, and

$$\delta(\alpha) \equiv \int d\tau\, \chi_\tau(\alpha) = \frac{1}{2}[1+\alpha] \tag{2.8}$$

defines the delta function on the group $Z(2)$. Here $\chi_\tau(\alpha)=\alpha$ if $\tau=-1$, and 1 otherwise, are the characters of $Z(2)$. The partition function Z_Λ is defined by $\int d\mu_\Lambda(\hat{U},\sigma)=1$. The plaquette action is given by

$$A_p(|\operatorname{tr} U_p|, \sigma_p) = \beta |\operatorname{tr} U_p| \sigma_p + \lambda \sigma_p \tag{2.9}$$

and

$$A_p(|\operatorname{tr} U_p|, \sigma_p) = \beta |\operatorname{tr} U_p| \sigma_p + \ln(\theta(|\operatorname{tr} U_p| - k)), \tag{2.10}$$

corresponding to (2.4) and (2.5), respectively. (2.7) is easily seen to be independent of the choice of basis $\{S^\alpha\}$. Note that (2.7) is indeed a measure on the coset configuration space since it is invariant under $U_b \to U_b \gamma_b$, for any $Z(2)$-valued 1-form γ. If $\hat{F}(\hat{U},\sigma)$ expresses a gauge-invariant observable $F(U)$ in the new variables, the expectation (2.2) then satisfies

$$\langle F(U)\rangle = \int d\mu_\Lambda(U,\sigma)\hat{F}(\hat{U},\sigma). \tag{2.11}$$

Free b.c. on $U_{\partial\Lambda}$ results in (2.7) without the product of δ functions over 2-cycles, and free b.c. on $\hat{U}_{\partial\Lambda}$, $\sigma_{\partial\Lambda}$. By selective omission of factors in this product, one may consider any mixture of free and periodic b.c. (2.7) may be straightforwardly generalized to accommodate other useful types of b.c.,[10] and indeed applied to Λ of any torsionless homology.

We are interested in the expectation of a Wilson loop subject to constraints whose physical effect is the elimination of its interaction with those vortices that are *not* energetically directly suppressed by the plaquette action at large β. These are extended vortices in the $\hat{U}_\Lambda$ configurations, ''thick'' vortices.[11] Thin vortices, vortices in the σ_Λ configurations, are directly suppressed by action cost proportional to their length. There are also hybrid vortices formed by the joining of ''punctured'' thick and thin vortices along their common coboundary, which represents a magnetic current ''loop,'' i.e., a coclosed set of 3-cubes $\mathcal{L}$ with $(d\eta)_c = (d\sigma)_c = -1$, $\forall c \in \mathcal{L}$. Such hybrids with a minimal short thin section but extended thick section can also affect long distance behavior at large β. We refer to Refs. 2, and 6, 7, 12, for a discussion of the physical interpretation of the constraints we now introduce.

We consider the constrained Wilson loop expectation:

$$W[C] = \langle \mathrm{tr}\, U[C]\,\theta[\mathrm{tr}\, U[C]\,\eta_S]\rangle, \tag{2.12}$$

where now $\langle - \rangle$ denotes expectations in the restriction of the measure (2.1) to

$$d\mu_\Lambda(U) = Z_\Lambda^{-1}\, d\nu_\Lambda(U) \exp\left(\sum_{p\in\Lambda} A_p(U_p)\right), \tag{2.13}$$

with

$$d\nu_\Lambda(U) = \prod_{b\in\Lambda} dU_b \prod_{c\in\Lambda} \delta((d\eta)_c) \prod_\alpha \delta(\eta[S^\alpha]), \tag{2.14}$$

and periodic b.c. (Z_Λ defined by $\langle 1\rangle = 1$). The constraint[6]

$$\prod_c \delta((d\eta)_c) \tag{2.15}$$

in the measure (2.13) eliminates all monopoles, and hence all hybrid vortices in the theory. The constraint[7]

$$\prod_\alpha \delta(\eta[S^\alpha]) \tag{2.16}$$

in (2.13) eliminates all thick vortices completely winding around the lattice, i.e., vortices rendered topologically stable by the nontrivial lattice topology (torus T^d) due to the periodic b.c. The factor[2,7]

$$\theta[\mathrm{tr}\, U[C]\,\eta_S], \tag{2.17}$$

where $\eta_S \equiv \Pi_{p\in S}\eta_p$, constraining the Wilson loop observable $\mathrm{tr}\, U[C]$ in (2.12), forbids the linking of a thick vortex (any odd number of thick vortices) with the loop (while still allowing thin vortices). Here S is any surface bounded by the loop. The expectation (2.12) is easily seen to be independent of the choice of S.

It should be noted that this last constraint may be included in the measure (2.13)–(2.14), as the constraints (2.15), (2.16) are, instead of as part of the observable, i.e., consider the expectation of $\mathrm{tr}\, U[C]$ with the factor (2.17) included in (2.14). Since, however, we are interested in a lower bound on the expectation, and, trivially, (2.17) is bounded from above by unity, its presence in the denominator in the normalized measure (2.13) may be removed, and one may as well consider the expectation (2.12) as defined above.

We pass then to $SO(3)\times Z(2)$ variables in which (2.12) assumes the form

$$W[C]=Z_\Lambda^{-1}\int d\nu_\Lambda(U)\prod_{p\in\Lambda} d\sigma_p \prod_{c\in\Lambda}\delta((d\eta)_c(d\sigma)_c)\prod_\alpha \delta(\eta[S^\alpha]\sigma[S^\alpha])$$
$$\cdot \exp\left(\sum_p A_p(|\mathrm{tr}\, U_p|,\sigma_p)\right)\mathrm{tr}\, U[C]\eta_S\sigma_S\ \theta[\mathrm{tr}\, U[C]\eta_S] \tag{2.18}$$

$$=Z_\Lambda^{-1}\int d\nu_\Lambda(U)\prod_{b\in\Lambda} d\gamma_b \exp\left(\sum_p A_p(|\mathrm{tr}\, U_p|,(d\gamma)_p)\right)$$
$$\times\theta[\mathrm{tr}\, U[C]\eta_S]\ \mathrm{tr}\, U[C]\eta_S\gamma[C], \tag{2.19}$$

where γ is a $Z(2)$-valued 1-form. The second equality is obtained by solving the constraints on σ_Λ that result from the restriction to (2.14):

$$(d\sigma)_c=1,\quad \forall c,\quad \delta(\sigma[S^\alpha])=1,\quad \forall\alpha\ \Rightarrow\ \sigma_p=\prod_{b\in\partial p}\gamma_b=(d\gamma)_p,$$
$$\sigma_S\equiv\prod_{p\in S}\sigma_p=\prod_{b\in C}\gamma_b=\gamma[C]. \tag{2.20}$$

Now,

$$\theta[\mathrm{tr}\, U[C]\eta_S]\ \mathrm{tr}\, U[C]\eta_S=\theta[\mathrm{tr}\, U[C]\eta_S]\ |\mathrm{tr}\, U[C]|, \tag{2.21}$$

so we can write (2.12) in the form

$$W[C]=\langle\theta[\mathrm{tr}\, U[C]\eta_S]\ |\mathrm{tr}\, U[C]|\ \gamma[C]\rangle_{SO(3)\otimes Z(2)}, \tag{2.22}$$

where

$$\langle - \rangle_{SO(3)\otimes Z(2)}=\int d\mu_\Lambda(U,\gamma)(-), \tag{2.23}$$

with

$$d\mu_\Lambda(U,\gamma)\equiv Z_\Lambda^{-1}\, d\hat\nu_\Lambda(U)\prod_{b\in\Lambda} d\gamma_b \exp\left(\sum_{p\in\Lambda} K_p(U)(d\gamma)_p\right) \tag{2.24}$$

and periodic b.c. $d\hat\nu_\Lambda(U)$ and $K_p(U)=K_p(\hat U)$ are defined as

$$d\hat\nu_\Lambda(U)\equiv d\nu_\Lambda(U),$$
$$K_p(U)\equiv\beta|\mathrm{tr}\, U_p|+\lambda \tag{2.25}$$

for the action (2.9), and

$$d\hat{\nu}_\Lambda(U) \equiv d\nu_\Lambda(U) \prod_{p\in\Lambda} \theta(|\mathrm{tr}\, U_p| - k),$$
$$K_p(U) \equiv \beta |\mathrm{tr}\, U_p| \tag{2.26}$$

for the action (2.10). It is easily verified that (2.24) is a reflection positive measure in $(d-1)$-dimensional planes with sites.

Our task in the following is to bound the constrained Wilson loop expectation (2.22) from below.

III. ABSENCE OF CONFINEMENT

We formulate our main result as the following.

Theorem III.1: *For sufficiently large* β*, and dimension* $d\geqslant 3$ *the constrained Wilson loop expectation given by (2.12), or, equivalently, (2.22), exhibits perimeter law, i.e., there exist constants* α, $\alpha_1(d)$, $\alpha_2(d)$ *such that*

$$W[C] \geqslant \alpha \exp(-\alpha_2 e^{-\alpha_1\beta}|C|). \tag{3.1}$$

From now on all expectation signs are meant in the measure (2.24), and, for brevity, we drop the label $SO(3)\otimes Z(2)$.

Choose the loop C so that it is bisected into two equal pieces by a $(d-1)$-dim plane with sites. Then reflection positivity (RP) of the measure (2.24) implies

$$\langle \mathrm{tr}\, U[C]\,\eta_S \gamma[C]\rangle \geqslant 0. \tag{3.2}$$

(Without loss of generality, the surface S may always be assumed to be also reflection symmetric in the hyperplane bisecting C.) Inserting

$$1 = \theta[\mathrm{tr}\, U[C]\,\eta_S] + \theta[-\mathrm{tr}\, U[C]\,\eta_S]$$

in (3.2), we then have

$$\begin{aligned} 0 &\leqslant \langle \theta[\mathrm{tr}\, U[C]\,\eta_S]\,\mathrm{tr}\, U[C]\,\eta_S\gamma[C]\rangle + \langle \theta[-\mathrm{tr}\, U[C]\,\eta_S]\,\mathrm{tr}\, U[C]\,\eta_S\gamma[C]\rangle \\ &= \langle \theta[\mathrm{tr}\, U[C]\,\eta_S]|\,\mathrm{tr}\, U[C]|\,\gamma[C]\rangle - \langle \theta[-\mathrm{tr}\, U[C]\,\eta_S]|\,\mathrm{tr}\, U[C]|\,\gamma[C]\rangle. \end{aligned} \tag{3.3}$$

On the other hand,

$$\langle |\mathrm{tr}\, U[C]|\,\gamma[C]\rangle = \langle \theta[\mathrm{tr}\, U[C]\,\eta_S]|\,\mathrm{tr}\, U[C]|\,\gamma[C]\rangle + \langle \theta[-\mathrm{tr}\, U[C]\,\eta_S]|\,\mathrm{tr}\, U[C]|\,\gamma[C]\rangle. \tag{3.4}$$

Adding (3.3) and (3.4), we have

$$\langle \theta[\mathrm{tr}\, U[C]\,\eta_S]|\,\mathrm{tr}\, U[C]|\,\gamma[C]\rangle \geqslant \tfrac{1}{2}\langle |\mathrm{tr}\, U[C]|\,\gamma[C]\rangle. \tag{3.5}$$

Hence from (2.22) we obtain

$$W[C] \geqslant \tfrac{1}{2}\langle |\mathrm{tr}\, U[C]|\,\gamma[C]\rangle. \tag{3.6}$$

Now

$$\frac{1}{2}\langle|\mathrm{tr}\, U[C]|\,\gamma[C]\rangle = Z_\Lambda^{-1}\int d\hat{\nu}_\Lambda(U)\,\frac{|\mathrm{tr}\, U[C]|}{2}\cdot\int\prod_{b\in\Lambda} d\gamma_b\,\gamma[C]\exp\left(\sum_{p\in\Lambda} K_p(U)(d\gamma)_p\right)$$

$$= \int d\hat{\mu}_\Lambda(U)\,\frac{|\mathrm{tr}\, U[C]|}{2}\langle\gamma[C]\rangle_{Z(2)}(\{K_p(U)\}), \tag{3.7}$$

where

$$d\hat{\mu}_\Lambda(U) \equiv Z_\Lambda^{-1}\, d\hat{\nu}(U) Z_{Z(2),\Lambda}(\{K_p(U)\}), \tag{3.8}$$

and

$$\langle - \rangle_{Z(2)}(\{K_p\}) = Z_{Z(2),\Lambda}^{-1}(\{K_p\})\int\prod_{b\in\Lambda} d\gamma_b(-)\exp\left(\sum_{p\in\Lambda} K_p\gamma_p\right) \tag{3.9}$$

denotes expectations in the $Z(2)$ lattice gauge theory with plaquette couplings K_p, and partition function $Z_{Z(2),\Lambda}$ defined by $\langle 1\rangle_{Z(2)}(\{K_p\})=1$. In (3.7) we have the U-dependent couplings $K_p(U)\geqslant 0$ given by (2.25), (2.26). By the Griffiths inequalities[13] applied to (3.9):

$$\langle\gamma[C]\rangle_{Z(2)}(\{K_p(U)\})\geqslant 0, \tag{3.10}$$

whereas

$$\tfrac{1}{2}|\mathrm{tr}\, U[C]| \geqslant \tfrac{1}{4}|\mathrm{tr}\, U[C]|^2 = \tfrac{1}{4}|\chi_F(U[C])|^2 = \tfrac{1}{4}[\chi_A(U[C])+1], \tag{3.11}$$

where χ_F and χ_A denote the $SU(2)$ characters in the fundamental and adjoint representation, respectively. Combining (3.10), (3.11) with (3.7) then gives

$$\frac{1}{2}\langle|\mathrm{tr}\, U[C]|\,\gamma[C]\rangle \geqslant \frac{1}{4}\int d\hat{\mu}_\Lambda(U)\langle\gamma[C]\rangle_{Z(2)}(\{K_p(U)\}) + \frac{1}{4}\int d\mu_\Lambda(U,\gamma)\chi_A(U[C])\,\gamma[C]. \tag{3.12}$$

Now by RP the second term on the rhs in (3.12) is non-negative. So, from (3.6), we obtain

$$W[C] \geqslant \frac{1}{4}\int d\hat{\mu}_\Lambda(U)\langle\gamma[C]\rangle_{Z(2)}(\{K_p(U)\}). \tag{3.13}$$

Action (2.10): Completing the proof of the result in the case of the action (2.10) is now straightforward. Applying again Griffiths inequalities, one has

$$\langle\gamma[C]\rangle_{Z(2)}(\{K_p\}) \geqslant \langle\gamma[C]\rangle_{Z(2)}(\{K_p'\}), \quad K_p\geqslant K_p', \quad \forall p. \tag{3.14}$$

With the measure given by (2.26), (3.8), use of (3.14) in (3.13) gives

$$W[C] \geqslant \tfrac{1}{4}\langle\gamma[C]\rangle_{Z(2)}(k\beta). \tag{3.15}$$

The rhs is the loop expectation in the $Z(2)$ gauge theory with $K_p = k\beta$, $\forall p$.

It is a well-known result that, for sufficiently large β and $d\geqslant 3$:

$$\langle\gamma[C]\rangle_{Z(2)}(\beta) \geqslant \mathrm{const}\,\exp(-\rho(\beta)|C|), \tag{3.16}$$

i.e., the expectation exhibits perimeter law. A proof[14] is by standard polymer expansion either as low-temperature expansion in ‘‘contours’’[15] or, after a duality transformation, as high-temperature expansion. Hence, (3.15) gives perimeter-law lower bound on $W[C]$ for all β such that $k\beta$ large enough. This concludes the proof of the theorem in the case of the action (2.10).

Action (2.9): A coclosed set of plaquettes that cannot be decomposed into two disjoint coclosed sets will be called a contour. (Two plaquettes are defined to be disjoint if they share no link.) Given a configuration $\{\gamma_b\}_{b\in\Lambda}$, the set of plaquettes on which $(d\gamma)_p=-1$ is a coclosed set ($(d-2)$-dim closed set on the dual lattice Λ^*). Such a coclosed set Γ can be uniquely decomposed into disjoint contours $\zeta_1,\ldots,\zeta_{|\Gamma|}$. Each such contour then is the site of a thin vortex. We expand the expectation $\langle\gamma[C]\rangle_{Z(2)}(\{K_p(U)\})$ in (3.13) in a contour expansion. The partition function in the denominator is expanded in the form

$$Z_{Z(2),\Lambda}(\{K_p(U)\})=\left(2^{-|\Lambda|}\prod_{p\in\Lambda}e^{K_p(U)}\right)\sum_{\Gamma\subset\Lambda}z_\Gamma(U). \tag{3.17}$$

The sum is over all sets Γ of disjoint (compatible) contours $\{\zeta_1,\ldots,\zeta_{|\Gamma|}\}$ and

$$z_\Gamma(U)\equiv\prod_{\zeta\in\Gamma}z_\zeta(U), \tag{3.18}$$

with

$$z_\zeta(U)\equiv\prod_{p\in\zeta}\exp(-2K_p(U)) \tag{3.19}$$

the activity of contour ζ, and $K_p(U)$ given by (2.25). Noting that $\langle\gamma[C]\rangle_{Z(2)}(\{K_p(U)\})=\langle\gamma[S]\rangle_{Z(2)}(\{K_p(U)\})$, where $\partial S=C$, it is easily seen that the expansion of the numerator

$$Z^C_{Z(2),\Lambda}(U)\equiv Z_{Z(2),\Lambda}(U)\langle\gamma[C]\rangle_{Z(2)}(\{K_p(U)\})$$

is given by the expansion (3.17) after replacement of the contour activities $z_\zeta(U)$, Eq. (3.19), by

$$z^C_\zeta(U)=(-1)^{l(C,\zeta)}z_\zeta(U),\quad l(C,\zeta)=|S\cap\zeta|\quad(\mathrm{mod}\,2). \tag{3.20}$$

(Here $l(C,\zeta)$ is the (mod 2) linking number of the contour ζ with the loop C.)

Letting

$$\mathcal{Z}_{Z(2),\Lambda}=\sum_{\Gamma\subset\Lambda}z_\Gamma(U),\quad \mathcal{Z}^C_{Z(2),\Lambda}=\sum_{\Gamma\subset\Lambda}z^C_\Gamma(U), \tag{3.21}$$

and applying Jensen's inequality in (3.13) gives

$$\begin{aligned}W[C]&\geqslant\frac{1}{4}\int d\hat\mu_\Lambda(U)\exp(\ln\mathcal{Z}^C_{Z(2),\Lambda}(U)-\ln\mathcal{Z}_{Z(2),\Lambda}(U))\\&\geqslant\frac{1}{4}\exp\left(\int d\hat\mu_\Lambda(U)(\ln\mathcal{Z}^C_{Z(2),\Lambda}(U)-\ln\mathcal{Z}_{Z(2),\Lambda}(U))\right)\equiv\frac{1}{4}\exp\mathcal{F}(\beta,\lambda).\end{aligned} \tag{3.22}$$

Noting that $\mathcal{Z}_{Z(2),\Lambda}(U)=1$ at $z_\zeta(U)=0$, we may define $\ln\mathcal{Z}_{Z(2),\Lambda}(U)$ as that continuous branch of the logarithm that vanishes at vanishing contour activity, and similarly for $\ln\mathcal{Z}^C_{Z(2),\Lambda}(U)$. A closed form representation of $\mathcal{F}(\beta,\lambda)$ is given by the Möbius inversion representation of $\ln\mathcal{Z}_{Z(2),\Lambda}(U)$ and of $\ln\mathcal{Z}^C_{Z(2),\Lambda}(U)$ as a finite series (for finite Λ) of logarithms of partition functions of linked clusters of single multiplicity.[16] Expanding these logarithms leads to the standard expansion of linked clusters of repeated multiplicities:

$$\mathcal{F}(\beta,\lambda)=\int d\hat\mu_\Lambda(U)\left(\sum_{Q\subset\Lambda}a(Q)(z^C_Q(U)-z_Q(U))\right). \tag{3.23}$$

The sum is over all linked clusters $Q=\{\zeta_1,\ldots,\zeta_{n_Q}\}$ of contours with the property that at least one contour in each Q winds around C. Multiple copies of a contour are allowed to appear as distinct members in a cluster, and

$$a(Q)=\sum_{G(Q)}(-1)^{|G(Q)|}. \tag{3.24}$$

Here the sum is over all connected graphs $G(Q)$ on the vertex set $\{\zeta,\ldots,\zeta_{n_Q}\}$ with a line connecting two vertices if they represent incompatible (not disjoint) contours. $|G(Q)|$ denotes the number of lines in $G(Q)$.

The expansion in (3.23) converges for sufficiently small activities $z_\zeta(U)$. Let $|z_\zeta(U)|\leqslant\exp(-b|\zeta|)$. The number of contours of size q with one plaquette fixed is bounded by $[10(d-2)]^{10(d-2)q}$. Applying well-known results on the convergence of the polymer-type cluster expansion,[17] it follows that the expansion on the rhs of (3.23) is absolutely convergent, uniformly in $|\Lambda|$ and over Ω_Λ, provided

$$b\geqslant 10(d-2)\ln 10(d-2)+\ln 5; \tag{3.25}$$

hence, by (2.25), λ sufficiently large. Uniform convergence allows integration term by term, so we have

$$\mathcal{F}(\beta,\lambda)=\sum_{Q\subset\Lambda}a(Q)(\langle z_Q^C(U)\rangle-\langle z_Q(U)\rangle). \tag{3.26}$$

We now observe that, for β sufficiently large, the series (3.26) converges for all $\lambda\geqslant 0$. To show this, we bound $\langle z_Q(U)\rangle$ uniformly in $|\Lambda|$ by chessboard estimates.[18] One finds (Appendix)

$$\langle z_Q(U)\rangle\leqslant\prod_{\zeta\in Q}\hat{z}(\beta,\lambda)^{|\zeta|}, \tag{3.27}$$

where

$$\hat{z}(\beta,\lambda)=c_1e^{-c_2\beta-2\lambda} \tag{3.28}$$

with positive constants $c_1(d)$, $c_2(d)$ depending only on the dimension d. It follows that (3.26) converges absolutely and uniformly in $|\Lambda|$, provided $|\hat{z}(\beta,\lambda)|\leqslant e^{-b}$ with b satisfying (3.25); hence for β large enough, and all $\lambda\geqslant 0$. Uniqueness of analytic continuation then implies that the representation of $\mathcal{F}$ by the cluster expansion in (3.26), originally obtained in the domain of large Re λ, is valid in this extended convergence domain Re $\lambda\geqslant 0$.

The leading contribution to (3.26) comes from the shortest possible contours, each consisting of $2(d-1)$ plaquettes forming the coboundary of a bond on the loop C, and there are $|C|$ of them. A bound on the remainder by the same as leading $\mathcal{O}(|C|)$-type behavior is a standard corollary of the polymer expansion convergence proof (e.g., Ref. 17). This concludes the proof of Theorem III.1 for the action (2.9).

ACKNOWLEDGMENTS

This research was supported by Department of Energy Grant No. DE-FG02-92ER-40672 and National Science Foundation Grant No. NSF-PHY 9531023.

APPENDIX: CHESSBOARD ESTIMATE

Let Π_e denote the set of ''even'' two-dimensional $[\kappa\lambda]$ planes: $x^\mu=2n^\mu$, integer n^μ, $\mu\neq\kappa$, λ. If $|\Pi_e|$ and $|\Lambda|$ denote the number of plaquettes in Π_e and Λ, respectively, we have $|\Pi_e|=|\Lambda|/s$, where $s=2^{d-3}d(d-1)$. Consider now a cluster Q, and let $\zeta_e\equiv\zeta\cap\Pi_e$. Using RP to reflect repeatedly in $(d-1)$-dimensional planes with sites $x^{\mu}=(2n^{\mu}+1)$, $\mu\neq\kappa$, λ; $x^{\mu}=n^{\mu}$, $\mu=\kappa$, λ, and the fact that $\exp(-2\beta|\mathrm{tr}\,U_p|)\leqslant 1$, one straightforwardly obtains the chessboard estimate

$$\begin{aligned}\langle z_Q(U)\rangle &= e^{-2\lambda|Q|}\left\langle \prod_{\zeta\in Q}\prod_{p\in\zeta} e^{-2\beta|\mathrm{tr}\,U_p|}\right\rangle\\ &\leqslant e^{-2\lambda|Q|}\left\langle \prod_{\zeta\in Q}\prod_{p\in\zeta_e} e^{-2\beta|\mathrm{tr}\,U_p|}\right\rangle\\ &\leqslant e^{-2\lambda|Q|}\left\langle \prod_{p\in\Pi_e} e^{-2\beta|\mathrm{tr}\,U_p|}\right\rangle^{|Q_e|/|\Pi_e|},\end{aligned} \tag{A1}$$

where $|Q|=\Sigma_{\zeta\in Q}|\zeta|$, and $|Q_e|=\Sigma_{\zeta\in Q}|\zeta_e|$. To estimate the last expectation on the rhs in (A1), it now suffices to bound the numerator from above by its maximum:

$$\left\langle \prod_{p\in\Pi_e} e^{-2\beta|\mathrm{tr}\,U_p|}\right\rangle \leqslant Z_\Lambda^{-1} e^{\lambda|\Lambda|} e^{2\beta(|\Lambda|-|\Pi_e|)}. \tag{A2}$$

The partition function in the denominator is bounded from below by restricting γ_b to 1, and each U_b integration to a small neighborhood of the identity such that $|\mathrm{tr}\,U_p|\geqslant 2e^{-\delta}$ for all p. Let τ_δ be the volume of this neighborhood. The constraints in $d\hat{\nu}_\Lambda(U)$ are then automatically satisfied, and we have

$$Z_\Lambda \geqslant \left(\left(\frac{\tau_\delta}{2}\right)^{2/(d-1)} e^{2\beta e^{-\delta}+\lambda}\right)^{|\Lambda|}. \tag{A3}$$

Now, given a cluster Q, we are free to pick the definition[19] of the "even" planes Π_e so that $|\zeta_e|\geqslant|\zeta|/s$, i.e., $|Q_e|\geqslant|Q|/s=|Q|/2^{(d-3)}d(d-1)$. Combining this with (A2), (A3), we find

$$\left\langle \prod_{p\in\Pi_e} e^{-2\beta|\mathrm{tr}\,U_p|}\right\rangle^{|Q_e|/|\Pi_e|} \leqslant \hat{a}(\beta)^{|Q|}, \tag{A4}$$

where we defined

$$\hat{a}\equiv\min_{\delta}\left(\left(\frac{2}{\tau_\delta}\right)^{2/(d-1)}\exp\left(-\frac{2\beta}{s}[1-s(1-e^{-\delta})]\right)\right). \tag{A5}$$

Thus $\hat{a}(\beta)\to 0$ exponentially as $\beta\to\infty$. (A1) with (A4) and

$$\hat{z}(\beta,\lambda)=e^{-2\lambda}\hat{a}(\beta) \tag{A6}$$

then produces the bound (3.27).

A more refined chessboard estimate, utilizing reflections onto every d-cell (dual site) of the lattice is possible (cf. Ref. 7), and results into elimination of the factor $1/s$ multiplying β in the exponent in (A5), a substantial numerical improvement. The simpler estimate given here, however, suffices for our purposes.

[1] *Lattice 97, Proceedings of the XVth International Symposium on Lattice Field Theory (Edinburgh)*, edited by C. T. H. Davies *et al.*, Nucl. Phys. B (Proc. Suppl.) **63**, (1998).
[2] T. G. Kovács and E. T. Tomboulis, Phys. Rev. D **57**, 4054 (1998); Nucl. Phys. B (Proc. Suppl.) **63**, 534 (1998).
[3] L. Del Debbio *et al.*, Phys. Rev. D **55**, 2298 (1997); Nucl. Phys. B (Proc. Suppl.) **63**, 552 (1998); L. Del Debbio *et al.*, hep-lat/9801027, 1998.
[4] M. Engelhardt *et al.*, Phys. Lett. B **431**, 141 (1998).
[5] T. G. Kovács and E. T. Tomboulis, Phys. Lett. B **443**, 239 (1998).
[6] G. Mack and V. B. Petkova, Ann. Phys. (N.Y.) **123**, 442 (1979); **125**, 117 (1980); Z. Phys. C **12**, 177 (1982).
[7] L. G. Yaffe, Phys. Rev. D **21**, 1574 (1980).

[8] E. T. Tomboulis and L. G. Yaffe, Commun. Math. Phys. **100**, 313 (1985).
[9] E. T. Tomboulis, Phys. Rev. D **23**, 2371 (1981).
[10] In particular, ''twisted'' periodic b.c. on the torus given by action $\tau^{\alpha} A_p(U_p)$, $\tau_\alpha \in Z(2)$, for $p \in S^*_\alpha$, where $\{S^*_\alpha\}$ a 2-cocycle basis, $(S^*_\alpha, S^\beta) = \delta^\beta_\alpha$, result into (2.7) with periodic b.c. and the substitution $\delta(\tau_\alpha \eta[S^\alpha]\sigma[S^\alpha])$ for each factor in the product over 2-cycles. Note that b.c. defined by summation over the twist τ_α eliminate the corresponding δ-function factor. Such b.c. enter in the definition of the magnetic- and electric-flux-free energy order parameters.
[11] In the continuum extrapolation they are configurations topologically characterized by nontrivial elements of $\pi_1(SU(2)/Z(2)) = Z(2)$ [$\pi_1(SU(N)/Z(N)) = Z(N)$ for general N].
[12] E. T. Tomboulis, Phys. Lett. B **303**, 103 (1993).
[13] R. B. Griffiths, J. Math. Phys. **8**, 978 (1967); D. G. Kelly and S. Sherman, *ibid.* **9**, 466 (1968); J. Ginibre, Commun. Math. Phys. **16**, 310 (1970).
[14] R. Marra and S. Miracle-Sole, Commun. Math. Phys. **67**, 233 (1979).
[15] This is a special case of the expansion we use below in treating the action (2.9).
[16] R. Kotecky and D. Preiss, Commun. Math. Phys. **103**, 491 (1986); G. Münster, Nucl. Phys. B **180** [FS 2], 23 (1981).
[17] C. Cammarota, Commun. Math. Phys. **85**, 517 (1982).
[18] J. Frölich, R. Israel, E. Lieb, and B. Simon, Commun. Math. Phys. **62**, 1 (1978).
[19] We are free to choose ''even'' coordinates (choice of lattice coordinate origin) and directions [$\kappa\lambda$].

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Covariant phase observables in quantum mechanics

Pekka Lahti[a)] and Juha-Pekka Pellonpää[b)]
Department of Physics, University of Turku, FIN-20014 Turku, Finland

(Received 12 February 1999; accepted for publication 14 July 1999)

In this paper we characterize all the phase shift covariant normalized positive operator measures, i.e., phase observables, and we investigate some of their examples. We also characterize those phase observables which arise from the phase space observables as their polar coordinate angle margins. © *1999 American Institute of Physics.* [S0022-2488(99)04810-0]

I. INTRODUCTION

The definition of the phase of an electromagnetic field has long been a problem in quantum physics. Many different approaches have been used to characterize the phase as a quantum observable. The traditional way of representing observables as self-adjoint operators in quantum mechanics has led to the search for self-adjoint phase operators. Apart from some partial success, all the attempts in this line of research have more or less failed; for an overview, see, e.g., Refs. 1–3. The reason for the failure of these approaches must be seen in the fact that the concept of an observable as a self-adjoint operator is unnecessarily restrictive; there is no spectral measure which is covariant under the shifts generated by the number observable of a single-mode field. The concept of an observable as a normalized positive operator measure (semispectral measure) has emerged from detailed investigations of the conceptual and mathematical foundations of quantum mechanics, and it has found ample applications in many important areas of quantum physics, as documented, for instance, in monographs.[4–7] In this approach a phase observable can be characterized in terms of its natural covariance condition together with the choice of the range of the phase variable, see, e.g., Ref. 5 or 7.

In this paper we adopt a formulation of a phase observable as a covariant normalized positive operator measure (Definition 2.1), and we prove a structure theorem, the phase theorem 2.2, for such observables. This result could be obtained as a special case of Theorem 4 of Ref. 8 which characterizes generalized imprimitivity systems for commutative groups. Here we give a direct constructive proof of the theorem without using any group theoretical arguments. The phase theorem allows us to exhibit classes of examples of phase observables (Sec. III), and it leads to a characterization of those phase space observables whose polar coordinate angle margins are phase observables (Sec. IV). Finally, we exhibit the conditions for the first moment operator of a phase observable to form a Heisenberg pair together with the number operator (Section V).

II. COVARIANT PHASE OBSERVABLES

In this section we characterize phase observables as a particular class of phase shift covariant operator measures defined on the Borel subsets of the real interval $[0,2\pi)$. To introduce these notions and to characterize them we need a few notations.

Let $\mathcal{H}$ denote a complex separable infinite-dimensional Hilbert space, with the inner product $\langle\cdot|\cdot\rangle$. For any two unit vectors $\psi,\varphi\in\mathcal{H}$, we let $|\psi\rangle\langle\varphi|$ denote the rank-one linear operator $\mathcal{H}\ni\eta\mapsto\langle\varphi|\eta\rangle\psi\in\mathcal{H}$. Let $\mathcal{K}:=\{\varphi_n\}_{n=0}^{\infty}$ be an orthonormal basis of $\mathcal{H}$, and let $N=\sum_{n=0}^{\infty}n|\varphi_n\rangle\langle\varphi_n|$ denote the self-adjoint operator with the domain $\mathcal{D}(N)=\{\psi\in\mathcal{H}|\sum_{n=0}^{\infty}n^2|\langle\varphi_n|\psi\rangle|^2<\infty\}$. We call N the *number operator* associated with $\mathcal{K}$. Let $\mathcal{B}[0,2\pi)$ denote the Borel subsets of the interval

a)Electronic mail: pekka.lahti@utu.fi
b)Electronic mail: juhpello@utu.fi

$[0,2\pi)$, and let $\mathcal{L}(\mathcal{H})$ denote the set of bounded operators on $\mathcal{H}$. We say that a positive normalized operator measure $E:\mathcal{B}[0,2\pi)\to\mathcal{L}(\mathcal{H})$ is a phase observable if it is covariant under the shifts generated by N. In other words, we adopt the following definition.

Definition 2.1: A map $E:\mathcal{B}[0,2\pi)\to\mathcal{L}(\mathcal{H})$ is a *phase observable* if

(i) $E(X)\geqslant O$ for all $X\in\mathcal{B}[0,2\pi)$ (positivity),
(ii) $E[0,2\pi)=I$ (normalization),
(iii) if $\{X_n\}_{n=1}^{\infty}\subset\mathcal{B}[0,2\pi)$ is a disjoint sequence, i.e., $X_n\cap X_m=\emptyset$, for $n\neq m$, then $E(\cup_{n=1}^{\infty}X_n)=\Sigma_{n=1}^{\infty}E(X_n)$, where the series converges in the weak operator topology (σ-additivity),
(iv) $e^{i\theta N}E(X)e^{-i\theta N}=E(X+\theta)$ for all $X\in\mathcal{B}[0,2\pi)$ and for all $\theta\in[0,2\pi)$, where $X+\theta:=\{x\in[0,2\pi)|(x-\theta)(\mathrm{mod}\,2\pi)\in X\}$ (covariance).

We recall that for positive operator measures the σ-additivity with respect to the weak operator topology is sufficient for the σ-additivity in the strong operator topology (Ref. 9, Proposition 1, p. 6). We have chosen the interval $[0,2\pi)$ as the range of a phase variable $\theta\in[0,2\pi)$, but any other interval $[a,a+2\pi)$, $a\in\mathbf{R}$, would do equally well.

For any bounded operator $A\in\mathcal{L}(\mathcal{H})$, let $A_{n,m}:=\langle\varphi_n|A\varphi_m\rangle$ denote its matrix elements with respect to $\mathcal{K}$. To shorten the notations we shall from now on write $\langle n|A|m\rangle$ and $|n\rangle\langle m|$ instead of $\langle\varphi_n|A\varphi_m\rangle$ and $|\varphi_n\rangle\langle\varphi_m|$, respectively. We let $\mathbf{C}$, $\mathbf{R}$, $\mathbf{Z}$, and $\mathbf{Z}^+$ denote the set of complex numbers, real numbers, integers, and positive integers, respectively, and we put $\mathbf{N}=\mathbf{Z}^+\cup\{0\}$. Let $I_k:=\Sigma_{n=0}^{k}|n\rangle\langle n|\in\mathcal{L}(\mathcal{H})$, $k\in\mathbf{N}$, so that $\{I_k\}_{k=0}^{\infty}$ is an increasing sequence of projection operators, with the unit operator I as the least upper bound. Therefore, $s\text{-}\lim_{k\to\infty}I_k=I$, and $s\text{-}\lim_{k\to\infty}AI_k=s\text{-}\lim_{k\to\infty}(I_kA)=A$, $A\in\mathcal{L}(\mathcal{H})$, as well as

$$A=s\text{-}\lim_{k\to\infty}(s\text{-}\lim_{l\to\infty}I_lAI_k)=s\text{-}\lim_{l\to\infty}(s\text{-}\lim_{k\to\infty}I_lAI_k),$$

so that we may write

$$A=\sum_{n,m=0}^{\infty}A_{n,m}|n\rangle\langle m|,$$

with the understanding that in the double summation the summation order is irrelevant and the series converge in the weak operator topology. We may now formulate the basic theorem of the paper.

Phase Theorem 2.2: Let $E:\mathcal{B}[0,2\pi)\to\mathcal{L}(\mathcal{H})$ be a phase observable. Then for any $X\in\mathcal{B}[0,2\pi)$,

(a) $E(X)=\Sigma_{n,m=0}^{\infty}c_{n,m}(2\pi)^{-1}\int_X e^{i(n-m)\theta}d\theta|n\rangle\langle m|$, where the series converge in the weak operator topology, and where
(b) $\{c_{n,m}|n,m\in\mathbf{N}\}\subset\mathbf{C}$, with $c_{n,n}=1$, for all $n\in\mathbf{N}$, and
(c) $\Sigma_{n,m=0}^{k}c_{n,m}|n\rangle\langle m|\geqslant O$, for all $k\in\mathbf{N}$.

Conversely, any family of complex numbers $\{c_{n,m}\in\mathbf{C}|n,m\in\mathbf{N}\}$ which has the properties (b) and (c), defines a unique phase observable of the form (a).

Proof: Let $\{c_{n,m}|n,m\in\mathbf{N}\}$ be a family of complex numbers such that $c_{n,n}=1$, for all $n\in\mathbf{N}$, and $\Sigma_{n,m=0}^{k}c_{n,m}|n\rangle\langle m|\geqslant O$, for all $k\in\mathbf{N}$. Then, for each $\theta\in[0,2\pi)$,

$$e^{iN\theta}\sum_{n,m=0}^{k}c_{n,m}|n\rangle\langle m|e^{-iN\theta}=\sum_{n,m=0}^{k}c_{n,m}e^{i(n-m)\theta}|n\rangle\langle m|\geqslant O,$$

which implies that

$$O \leqslant \frac{1}{2\pi}\int_X \sum_{n,m=0}^{k} c_{n,m} e^{i(n-m)\theta} |n\rangle\langle m| d\theta \leqslant \frac{1}{2\pi}\int_0^{2\pi} \sum_{n,m=0}^{k} c_{n,m} e^{i(n-m)\theta}|n\rangle\langle m| d\theta$$

$$= \sum_{n,m=0}^{k} c_{n,m}\left(\frac{1}{2\pi}\int_0^{2\pi} e^{i(n-m)\theta} d\theta\right)|n\rangle\langle m| = \sum_{n=0}^{k} |n\rangle\langle n| \leqslant I, \tag{1}$$

for all $k \in \mathbf{N}$, and for all $X \in \mathcal{B}[0,2\pi)$. Let $\mathrm{lin}\,\mathcal{K}$ be the linear hull of $\mathcal{K}$, and define, for each $X \in \mathcal{B}[0,2\pi)$, the sesquilinear function $B_X : \mathrm{lin}\,\mathcal{K} \times \mathrm{lin}\,\mathcal{K} \to \mathbf{C}$ as follows:

$$B_X(\varphi,\psi) := \sum_{n=0}^{k}\sum_{m=0}^{l} c_{n,m}\left(\frac{1}{2\pi}\int_X e^{i(n-m)\theta} d\theta\right)\langle\varphi|n\rangle\langle m|\psi\rangle,$$

where $\varphi = \Sigma_{n=0}^{k}\langle n|\varphi\rangle|n\rangle$, $\psi = \Sigma_{m=0}^{l}\langle m|\psi\rangle|m\rangle \in \mathrm{lin}\,\mathcal{K}$, $k,l \in \mathbf{N}$. From (1) one gets $|B_X(\varphi,\psi)| \leqslant \|\varphi\|\|\psi\|$, and $B_X(\varphi,\varphi) \geqslant 0$, for all $\varphi,\psi \in \mathrm{lin}\,\mathcal{K}$. Since $\overline{\mathrm{lin}\,\mathcal{K}} = \mathcal{H}$, the form B_X has a unique sesquilinear extension $\tilde{B}_X$ to $\mathcal{H}\times\mathcal{H}$ which has the following properties: $|\tilde{B}_X(\varphi,\psi)| \leqslant \|\varphi\|\|\psi\|$, and $\tilde{B}_X(\varphi,\varphi) \geqslant 0$, for all $\varphi,\psi \in \mathcal{H}$. Hence, there is a unique operator $E(X) \in \mathcal{L}(\mathcal{H})$ such as $\tilde{B}_X(\varphi,\psi) = \langle E(X)\varphi|\psi\rangle$, for all $\varphi,\psi \in \mathcal{H}$, and $O \leqslant E(X) \leqslant I$. The matrix elements of $E(X)$ with respect to $\mathcal{K}$ are

$$\langle n|E(X)|m\rangle = B_X(|n\rangle,|m\rangle) = c_{n,m}\frac{1}{2\pi}\int_X e^{i(n-m)\theta} d\theta, \quad n,m \in \mathbf{N}. \tag{2}$$

Therefore,

$$E(X) = \sum_{n,m=0}^{\infty} c_{n,m}\frac{1}{2\pi}\int_X e^{i(n-m)\theta} d\theta |n\rangle\langle m|,$$

for all $X \in \mathcal{B}[0,2\pi)$. Clearly, $E[0,2\pi) = I$.

Let $\{X_n\}_{n=1}^{\infty}$ be a family of disjoint sets of $\mathcal{B}[0,2\pi)$, and define the operators $E_k := E(\cup_{n=1}^{k} X_n)$, for all $k \in \mathbf{Z}^+$. From (2) one gets $E_k = \Sigma_{n=1}^{k} E(X_n)$, $k \in \mathbf{Z}^+$. Since $E_{k+1} - E_k = E(X_{k+1}) \geqslant O$, it follows that $O \leqslant E_k \leqslant E_{k+1} \leqslant I$, $k \in \mathbf{Z}^+$. Hence, the least upper bound $S \in \mathcal{L}(\mathcal{H})$ of the sequence $\{E_k\}_{k=1}^{\infty}$ exists, and $s\text{-}\lim_{k\to\infty} E_k = w\text{-}\lim_{k\to\infty} E_k = S$. Due to σ-additivity of the Lebesgue integral, the matrix elements of S with respect to $\mathcal{K}$ are

$$\begin{aligned}\langle n|S|m\rangle &= \lim_{k\to\infty}\langle n|E_k|m\rangle \\ &= \lim_{k\to\infty} c_{n,m}\frac{1}{2\pi}\int_{\cup_{n=1}^{k}X_n} e^{i(n-m)\theta} d\theta \\ &= c_{n,m}\frac{1}{2\pi}\int_{\cup_{n=1}^{\infty}X_n} e^{i(n-m)\theta} d\theta \\ &= \langle n|E(\cup_{n=1}^{\infty}X_n)|m\rangle.\end{aligned}$$

This concludes the proof that the map $\mathcal{B}[0,2\pi) \ni X \mapsto E(X) \in \mathcal{L}(\mathcal{H})$ is a positive normalized operator measure. The covariance condition (iv) is an immediate consequence of the structure of E. Hence E is a phase observable.

Assume now that $E : \mathcal{B}[0,2\pi) \to \mathcal{L}(\mathcal{H})$ is a phase observable. To show that it has the structure of (a), we determine the matrix elements of $E(X)$, $X \in \mathcal{B}[0,2\pi)$, with respect to $\mathcal{K}$. The covariance condition (iv) implies that

$$E_{n,m}(X+\theta) = e^{i(n-m)\theta}E_{n,m}(X), \tag{3}$$

for all $X\in\mathcal{B}[0,2\pi)$, $\theta\in[0,2\pi)$, and for all $n,m\in\mathbf{N}$. Therefore, for all $n,m\in\mathbf{N}$, and $k\in\mathbf{Z}^+$, we have

$$\begin{aligned}E_{n,m}[0,2\pi)&=E_{n,m}\left(\bigcup_{l=0}^{k-1}[l2\pi k^{-1},(l+1)2\pi k^{-1})\right)\\&=\sum_{l=0}^{k-1}E_{n,m}([0,2\pi k^{-1})+l2\pi k^{-1})\\&=\left[\sum_{l=0}^{k-1}e^{i2\pi(n-m)k^{-1}l}\right]E_{n,m}[0,2\pi k^{-1})\\&=\begin{cases}kE_{n,m}[0,2\pi k^{-1}) & \text{when } (n-m)k^{-1}\in\mathbf{Z},\\ 0 & \text{when } (n-m)k^{-1}\notin\mathbf{Z}.\end{cases}\end{aligned}\tag{4}$$

Suppose that $n,m\in\mathbf{N}$ and $k\in\mathbf{Z}^+$ are such that $(n-m)k^{-1}\notin\mathbf{Z}$. Then $\int_0^{2\pi k^{-1}}e^{i(n-m)\theta}d\theta\neq 0$, and we can define

$$c_{n,m}(k):=\frac{E_{n,m}[0,2\pi k^{-1})}{(2\pi)^{-1}\int_0^{2\pi k^{-1}}e^{i(n-m)\theta}d\theta},$$

so that

$$E_{n,m}[0,2\pi k^{-1})=c_{n,m}(k)\,\frac{1}{2\pi}\int_0^{2\pi k^{-1}}e^{i(n-m)\theta}d\theta=c_{n,m}(k)\,\frac{e^{i(n-m)2\pi k^{-1}}-1}{i(n-m)2\pi}.$$

On the other hand, for all $r\in\mathbf{Z}^+$, $(n-m)(rk)^{-1}\notin\mathbf{Z}$, and

$$\begin{aligned}E_{n,m}[0,2\pi k^{-1})&=E_{n,m}\left(\bigcup_{l=0}^{r-1}[l2\pi(rk)^{-1},(l+1)2\pi(rk)^{-1})\right)\\&=\left[\sum_{l=0}^{r-1}e^{i2\pi(n-m)(rk)^{-1}l}\right]E_{n,m}[0,2\pi(rk)^{-1})=c_{n,m}(rk)\,\frac{e^{i(n-m)2\pi k^{-1}}-1}{i(n-m)2\pi}.\end{aligned}$$

This shows that $c_{n,m}(k)=c_{n,m}(rk)$, $r\in\mathbf{Z}^+$. Since $(n-m)(|n-m|+1)^{-1}\notin\mathbf{Z}$, one has $c_{n,m}(k)=c_{n,m}((|n-m|+1)k)=c_{n,m}(|n-m|+1)$. Thus, for all $k\in\mathbf{Z}^+$, for which $(n-m)k^{-1}\notin\mathbf{Z}$, the number $c_{n,m}(k)$ is the same, and we may define $c_{n,m}:=c_{n,m}(|n-m|+1)$ for all $n,m\in\mathbf{N}$ and $n\neq m$.

If $(n-m)k^{-1}\in\mathbf{Z}$, $n,m\in\mathbf{N}$, $k\in\mathbf{Z}^+$, Eq. (4) gives

$$E_{n,m}[0,2\pi k^{-1})=\frac{\delta_{n,m}}{k}=\frac{1}{2\pi}\int_0^{2\pi k^{-1}}e^{i(n-m)\theta}\mathrm{d}\theta.\tag{5}$$

Thus, if we define $c_{n,n}:=1$, $n\in\mathbf{N}$, we get

$$E_{n,m}[0,2\pi k^{-1})=c_{n,m}\frac{1}{2\pi}\int_0^{2\pi k^{-1}}e^{i(n-m)\theta}\mathrm{d}\theta,\tag{6}$$

for all $k\in\mathbf{Z}^+$, and $n,m\in\mathbf{N}$.

Let $n,m\in\mathbf{N}$. From (3) one gets

$$E_{n,m}\left(\bigcup_{p=1}^{\infty}\{p^{-1}\}\right)=E_{n,m}(\{0\})\sum_{p=1}^{\infty}e^{i(n-m)p^{-1}},$$

which implies that $E_{n,m}(\{0\})=0$. Thus the measure $E_{n,m}$ is nonatomic, i.e., $E_{n,m}(\{x\})=e^{ix(n-m)}E_{n,m}(\{0\})=0$, $x\in[0,2\pi)$, which implies that the function $x\mapsto E_{n,m}[0,x)$ is continuous. Since, from (6) it follows that for all $k\in\mathbf{Z}^{+}$, $p\in\{1,2,...,k\}$,

$$E_{n,m}[0,2\pi pk^{-1})=E_{n,m}\left(\bigcup_{l=0}^{p-1}[l2\pi k^{-1},(l+1)2\pi k^{-1})\right)=c_{n,m}\frac{1}{2\pi}\int_{0}^{2\pi pk^{-1}}e^{i(n-m)\theta}\mathrm{d}\theta, \quad (7)$$

and since the function $x\mapsto E_{n,m}[0,x)$ is continuous, and the set $\{2\pi pk^{-1}\in[0,2\pi)|k\in\mathbf{Z}^{+},\ p\in\{1,2,...,k\}\}$ is dense in $[0,2\pi)$ it follows that for all $x\in(0,2\pi]$

$$E_{n,m}[0,x)=c_{n,m}\frac{1}{2\pi}\int_{0}^{x}e^{i(n-m)\theta}\mathrm{d}\theta.$$

Hence, by Hahn extension theorem

$$E_{n,m}(X)=c_{n,m}\frac{1}{2\pi}\int_{X}e^{i(n-m)\theta}\mathrm{d}\theta \quad (8)$$

for all $X\in\mathcal{B}[0,2\pi)$ and $n,m\in\mathbf{N}$.

Finally, we prove that

$$\sum_{n,m=0}^{k}c_{n,m}|n\rangle\langle m|\geqslant O, \quad (9)$$

for all $k\in\mathbf{N}$. Since $E(X)=E(X)^{*}$, $X\in\mathcal{B}[0,2\pi)$, we get

$$c_{n,m}=\overline{c_{m,n}}, \quad (10)$$

for all $n,m\in\mathbf{N}$. It follows from (10) that $\Sigma_{n,m=0}^{k}c_{n,m}\langle\psi|n\rangle\langle m|\psi\rangle\in\mathbf{R}$, for all $k\in\mathbf{N}$, $\psi\in\mathcal{H}$. Hence, if (9) does not hold, one may choose a $\varphi\in\mathcal{H}$ and an $l\in\mathbf{Z}^{+}$ such that $\Sigma_{n,m=0}^{l}c_{n,m}\langle\varphi|n\rangle\times\langle m|\varphi\rangle<0$, and define a function

$$f:[0,2\pi)\to\mathbf{R},\theta\mapsto f(\theta):=\sum_{n,m=0}^{l}c_{n,m}e^{i(n-m)\theta}\langle\varphi|n\rangle\langle m|\varphi\rangle.$$

Due to the continuity of f one can choose an $\epsilon\in(0,2\pi)$ such that $\int_{0}^{\epsilon}f(\theta)\mathrm{d}\theta<0$. Thus

$$\langle I_{l}\varphi|E[0,\epsilon)I_{l}\varphi\rangle=\frac{1}{2\pi}\int_{0}^{\epsilon}f(\theta)\mathrm{d}\theta<0,$$

which contradicts with the positivity of E. □

To close this section we note the following two points. If $c_{n,m}\neq 0$, choose $\psi=|n\rangle-|c_{n,m}|c_{n,m}^{-1}|m\rangle$, and substitute it to (c) to deduce that $c_{n,n}-|c_{n,m}|-|c_{n,m}|+c_{m,m}\geqslant 0$. This implies that

$$|c_{n,m}|\leqslant 1,$$

for all $n,m\in\mathbf{N}$. If E is a phase observable it can be seen directly from the phase theorem that

$$c_{n,m} = 2\pi \lim_{\epsilon\to 0} \frac{E_{n,m}[0,\epsilon)}{\epsilon},$$

for all $n,m \in \mathbf{N}$.

III. SOME EXAMPLES OF PHASE OBSERVABLES

The phase theorem fixes the structure of a phase observable E in terms of a set of complex numbers $c_{n,m}$, $n,m \in \mathbf{N}$:

$$E(X) = \sum_{n,m=0}^{\infty} c_{n,m} \left(\frac{1}{2\pi} \int_X e^{i(n-m)\theta} \mathrm{d}\theta \right) |n\rangle\langle m|, \quad X \in \mathcal{B}[0,2\pi). \tag{11}$$

Exhibiting various sets $\{c_{n,m} \in \mathbf{C} | n,m \in \mathbf{N}\}$, which fulfill conditions (b) and (c) of the phase theorem, one gets various examples of phase observables. In this section we shall give a few examples. Before doing that we shall notice the following obvious facts concerning any phase observable E. First, the phase probabilities in any number state $|n\rangle$, $n \in \mathbf{N}$, are uniform, that is,

$$\langle n|E(X)|n\rangle = \frac{1}{2\pi} \int_X \mathrm{d}\theta, \quad X \in \mathcal{B}[0,2\pi),$$

showing that phase is totally random in the number states. Second, since

$$\langle n|E(X)^2|n\rangle = \sum_{s=0}^{\infty} |c_{n,s}|^2 \left| \frac{1}{2\pi} \int_X e^{i(n-s)\theta} \mathrm{d}\theta \right|^2$$

and $|c_{n,s}| \leq 1$, choosing, for instance, $X = [0,\pi)$ and $n=0$, one gets $\langle 0|E[0,\pi)^2|0\rangle \leq \frac{3}{8}$, which, when compared with $\langle 0|E[0,\pi)|0\rangle = \frac{1}{2}$, confirms the well-known fact that there is no spectral measure having the structure (11). Finally, since the probabilities $\langle n|E(X)|n\rangle$, $X \in \mathcal{B}[0,2\pi)$, do not depend on $|n\rangle$, there is no informationally complete phase observable.[7] We now turn to the examples.

Example 3.1: Consider a set $\{c_{n,m} | n,m \in \mathbf{N}\} \subset \mathbf{C}$ such that $c_{n,n} = 1$, $n \in \mathbf{N}$, and $\Sigma_{n>m=0}^{\infty} |c_{n,m}| \leq 1$. Then the formula (11) determines a phase observable. To check this we need only to confirm that the operators $\Sigma_{n,m=0}^{k} c_{n,m} |n\rangle\langle m|$, $k \in \mathbf{N}$, are positive. Let $\psi = \Sigma_{n=0}^{\infty} d_n |n\rangle \in \mathcal{H}$, $d_n = |d_n| e^{i\theta_n} \in \mathbf{C}$, and write $c_{n,m} = |c_{n,m}| e^{i\theta_{n,m}}$, $n,m \in \mathbf{N}$, so that, for all $k \in \mathbf{N}$,

$$\sum_{n,m=0}^{k} c_{n,m} \langle \psi|n\rangle\langle m|\psi\rangle = \|\psi\|^2 + 2 \sum_{n>m=0}^{k} |c_{n,m}| |d_n| |d_m| \cos(\theta_{n,m} - \theta_n + \theta_m).$$

To prove that the right-hand side of the above equation is non-negative we observe that for any $0 \leq a$, $b \leq 1$ such that $a^2 + b^2 \leq 1$, one has $ab \leq 2^{-1}$. Therefore,

$$|d_n| |d_m| \leq 2^{-1} \|\psi\|^2$$

for all $n,m \in \mathbf{N}$. Since $\cos(x) \geq -1$, $x \in \mathbf{R}$, we then have

$$\|\psi\|^2 + 2 \sum_{n>m=0}^{k} |c_{n,m}| |d_n| |d_m| \cos(\theta_{n,m} - \theta_n + \theta_m) \geq \|\psi\|^2 \left(1 - \sum_{n>m=0}^{k} |c_{n,m}| \right) \geq 0,$$

which concludes the proof of the positivity of the operator $\Sigma_{n,m=0}^{k} c_{n,m} |n\rangle\langle m|$.

Example 3.2: Example 3.1 shows that if $s \neq t \in \mathbf{N}$, $z \in \mathbf{C}$, $|z| \leq 1$, then for all $X \in \mathcal{B}[0,2\pi)$,

$$E(X)=\frac{1}{2\pi}\int_X d\theta\, I+z\,\frac{1}{2\pi}\int_X e^{i(s-t)\theta}\mathrm{d}\theta|s\rangle\langle t|+\bar{z}\frac{1}{2\pi}\int_X e^{i(t-s)\theta}\mathrm{d}\theta|t\rangle\langle s|$$

is a phase observable. This example exhibits the fact that the only commutative phase observable is the trivial one $X\mapsto((1/2\pi)\int_X \mathrm{d}\theta)I$, and the degree of noncommutativity of E is the stronger, the more there are nonzero nondiagonal structure constants $c_{n,m}$, $n\neq m$, in (11).

Example 3.3: If we choose $c_{n,m}:=a_n\overline{a_m}=|a_n||a_m|e^{i(\nu_n-\nu_m)}$ for all $n,m\in\mathbf{N}$, where $a_n=|a_n|e^{i\nu_n}\in\mathbf{C}$ is arbitrary for all $n\in\mathbf{N}$, then $|a_n|^2=c_{n,n}=1$, $n\in\mathbf{N}$. Hence $c_{n,m}=e^{i(\nu_n-\nu_m)}$, $n,m\in\mathbf{N}$, where $\nu_n\in\mathbf{R}$, $n\in\mathbf{N}$. Since

$$\sum_{n,m=0}^{k} c_{n,m}\langle\psi|n\rangle\langle m|\psi\rangle=\sum_{n=0}^{k}\overline{d_n e^{-i\nu_n}}\sum_{m=0}^{k} d_m e^{-i\nu_m}=\left|\sum_{n=0}^{k} d_n e^{-i\nu_n}\right|^2\geqslant 0,$$

for all $\psi=\sum_{n=0}^{\infty} d_n|n\rangle\in\mathcal{H}$, and $k\in\mathbf{N}$, the set $\{e^{i(\nu_n-\nu_m)}|\nu_n,\nu_m\in\mathbf{R},n,m\in\mathbf{N}\}$ defines a phase observable

$$E(X)=\sum_{n,m=0}^{\infty}\frac{1}{2\pi}\int_X e^{i[(n-m)\theta+\nu_n-\nu_m]}\mathrm{d}\theta|n\rangle\langle m|,\quad X\in\mathcal{B}[0,2\pi).$$

Example 3.4: Put $\nu_n=0$, $n\in\mathbf{N}$, in the above example. Then

$$E(X)=\sum_{n,m=0}^{\infty}\frac{1}{2\pi}\int_X e^{i(n-m)\theta}\mathrm{d}\theta|n\rangle\langle m|,\quad X\in\mathcal{B}[0,2\pi).$$

This phase observable is the unique positive operator measure in the polar decomposition $a=(\int_0^{2\pi}e^{i\theta}\mathrm{d}E(\theta))\sqrt{N}$ of the lowering operator $a=\sum_{n=0}^{\infty}\sqrt{n+1}|n\rangle\langle n+1|$ associated with the number operator N, see, e.g., Ref. 7.

IV. THE PHASE OBSERVABLES ARISING FROM THE PHASE SPACE OBSERVABLES

The polar coordinate angle margins of the phase space observables constitute a physically relevant class of potential phase observables. In this section we characterize those phase space observables which give rise to phase observables. We introduce first the relevant phase space observables. Let $a=\sum_{n=0}^{\infty}\sqrt{n+1}|n\rangle\langle n+1|$ and $a^*=\sum_{n=0}^{\infty}\sqrt{n+1}|n+1\rangle\langle n|$ be the lowering and raising operators associated with the number operator $N=a^*a$, with the appropriate domains. Let $\lambda:\mathcal{B}(\mathbf{C})\to[0,\infty]$ be the two-dimensional Lebesgue measure. Let $D(z)=e^{za^*-\bar{z}a}$, $z\in\mathbf{C}$, be the unitary shift operator, and let T be any state, that is, a positive trace-one operator on $\mathcal{H}$. The set function

$$\mathcal{B}(\mathbf{C})\ni Z\mapsto A^T(Z):=\frac{1}{\pi}\int_Z D(z)TD(z)^*\mathrm{d}\lambda(z)\in\mathcal{L}(\mathcal{H}),$$

is then a positive normalized operator measure, the phase space observable defined by the state T.[4,10] We recall that the set of states is a closed σ-convex subset of the Banach space of the set of trace class operators on $\mathcal{H}$ (in trace norm). In particular, if $\{T_i\}_{i=1}^{\infty}$ is a sequence of states and if $\{\lambda_i\}_{i=1}^{\infty}$ is a sequence of non-negative numbers such that $\sum_{i=1}^{\infty}\lambda_i=1$, then the series $\sum_{i=1}^{\infty}\lambda_i T_i$ converges in trace norm to a state T, and we write $T=\sum_{i=1}^{\infty}\lambda_i T_i$, and we say that T is a mixture of the states T_i with the weights λ_i. Assume now that $T=\sum_{i=1}^{\infty}\lambda_i T_i$. By the dominated convergence theorem the phase space observables A^T and A^{T_i} are then related as follows: $A^T(Z)=\sum_{i=1}^{\infty}\lambda_i A^{T_i}(Z)$ for each $Z\in\mathcal{B}(\mathbf{C})$ (with the convergence in the weak operator topology). In that case, we write $A^T=\sum_{i=1}^{\infty}\lambda_i A^{T_i}$.

Writing $\mathbf{C}\ni z=|z|e^{i\arg z}\equiv re^{i\theta}$, $(r,\theta)\in[0,\infty)\times[0,2\pi)$, we may define the polar coordinate marginal measures of A^T:

$$\mathcal{B}[0,\infty) \ni R \mapsto A^T(R \times [0,2\pi)) =: F^T(R) \in \mathcal{L}(\mathcal{H}),$$

$$\mathcal{B}[0,2\pi) \ni X \mapsto A^T([0,\infty) \times X) =: E^T(X) \in \mathcal{L}(\mathcal{H}).$$

If T is a number state $|s\rangle$, $s \in \mathbf{N}$, that is $T = |s\rangle\langle s|$, or a mixture of number states, that is, $T = \Sigma_{s=0}^{\infty} \lambda_s |s\rangle\langle s|$, for some $0 \leqslant \lambda_s \leqslant 1$, $\Sigma_{s=0}^{\infty} \lambda_s = 1$, then the operator measures F^T and E^T are known to represent an unsharp number observable and the conjugate unsharp phase observable, respectively.[5,7] The next theorem identifies the phase space observables A^T whose angle margins E^T are phase observables, that is, for which

$$e^{i\theta N} E^T(X) e^{-i\theta N} = E^T(X + \theta),$$

for all $X \in \mathcal{B}[0,2\pi), \theta \in [0,2\pi)$.

Theorem 4.1: The angle margin E^T of the phase space observable A^T is a phase observable if and only if T is of the form $\Sigma_{s=0}^{\infty} \lambda_s |s\rangle\langle s|$, for some non-negative numbers $\lambda_s \in \mathbf{R}$, for which $\Sigma_{s=0}^{\infty} \lambda_s = 1$.

Proof: If T is of the form $T = \Sigma_{s=0}^{\infty} \lambda_s |s\rangle\langle s|$ for some $\lambda_s \geqslant 0$, $\Sigma_{s=0}^{\infty} \lambda_s = 1$, then the facts that $E^T = \Sigma_{s=0}^{\infty} \lambda_s E^{|s\rangle\langle s|}$ and $e^{i\theta N} D(z) = D(z e^{i\theta}) e^{i\theta N}$, $z \in \mathbf{C}$, imply that

$$e^{i\theta N} E^T(X) e^{-i\theta N} = E^T(X + \theta), \quad X \in \mathcal{B}[0,2\pi), \quad \theta \in [0,2\pi).$$

Assume now, that E^T is a phase observable. The structure constants $c_{n,m} \in \mathbf{C}$, $n,m \in \mathbf{N}$, of E^T can then be deduced from the following equations:

$$\langle n|E^T(X)|m\rangle = c_{n,m} \frac{1}{2\pi} \int_X e^{i(n-m)\theta} \mathrm{d}\theta = \frac{1}{2\pi} \int_X \int_0^{\infty} \langle n|D(re^{i\theta}) T D(re^{i\theta})^* |m\rangle \mathrm{d}r^2 \, \mathrm{d}\theta \tag{12}$$

for all $X \in \mathcal{B}[0,2\pi)$. Let $\phi_n^s(z) := \langle s|D(z)^*|n\rangle$ for all $s,n \in \mathbf{N}$ and $z \in \mathbf{C}$. These functions ϕ_n^s are easily seen to be of the form

$$\phi_n^s(re^{i\theta}) = e^{i(s-n)\theta} \phi_n^s(r) = e^{i(s-n)\theta} (-1)^{\max\{0,s-n\}} \sqrt{\frac{(\min\{n,s\})!}{(\max\{n,s\})!}} e^{-r^2/2} r^{|s-n|} L_{\min\{n,s\}}^{|s-n|}(r^2), \tag{13}$$

for all $r \in [0,\infty)$, $\theta \in [0,2\pi)$, where L_m^k, $m,k \in \mathbf{N}$, is the associated Laguerre polynomial. From (12) we get for all $n,m \in \mathbf{N}$ and $\mathrm{d}\theta$-almost every $\theta \in [0,2\pi)$,

$$c_{n,m} e^{i(n-m)\theta} = \int_0^{\infty} \sum_{s,t=0}^{\infty} T_{s,t} \overline{\phi_n^s(re^{i\theta})} \phi_m^t(re^{i\theta}) \mathrm{d}r^2 = \int_0^{\infty} \sum_{s,t=0}^{\infty} T_{s,t} \phi_n^s(r) \phi_m^t(r) e^{i(n-s+t-m)\theta} \mathrm{d}r^2.$$

Multiplying the above equation by $(2\pi)^{-1} e^{-il\theta}$, $l \in \mathbf{N}$, and integrating it over the interval $[0,2\pi)$ yields

$$\begin{aligned} c_{n,m} \delta_{n-m,l} &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} \sum_{s,t=0}^{\infty} T_{s,t} \phi_n^s(r) \phi_m^t(r) e^{i(n-s+t-m-l)\theta} dr^2 \, \dot{d}\theta \\ &= \sum_{s=0}^{\infty} T_{s,s} \int_0^{\infty} \phi_n^s(r) \phi_m^s(r) dr^2 \, \delta_{n-m,l} + \sum_{s \neq t=0}^{\infty} T_{s,t} \int_0^{\infty} \phi_n^s(r) \phi_m^t(r) dr^2 \, \delta_{n-m+t-s,l} \\ &= \sum_{s=0}^{\infty} T_{s,s} \int_0^{\infty} \phi_n^s(r) \phi_m^s(r) dr^2 \, \delta_{n-m,l} + \sum_{k=1}^{\infty} \sum_{s=0}^{\infty} T_{s,s+k} \int_0^{\infty} \phi_n^s(r) \phi_m^{s+k}(r) dr^2 \, \delta_{n-m+k,l} \\ &\quad + \sum_{k=1}^{\infty} \sum_{t=0}^{\infty} T_{t+k,t} \int_0^{\infty} \phi_n^{t+k}(r) \phi_m^t(r) dr^2 \, \delta_{n-m-k,l}. \end{aligned}$$

If we choose $l=n-m$, then

$$c_{n,m}=\sum_{s=0}^{\infty} T_{s,s}\int_0^\infty \phi_n^s(r)\phi_m^s(r)dr^2, \tag{14}$$

and, by choosing $l=n-m+k$ (notice that k is fixed now)

$$\sum_{s=0}^{\infty} T_{s,s+k}\int_0^\infty \phi_n^s(r)\phi_m^{s+k}(r)dr^2=0, \tag{15}$$

for all $n,m\in\mathbf{N}$ and $k\in\mathbf{Z}^+$. This equation allows us to show that $T_{s,s+k}=0$ for all $s\in\mathbf{N}$ and $k\in\mathbf{Z}^+$. Indeed, using the equations [Ref. 11, Eqs. 8.973(1), and 7.414(11) or Ref. 12, Eq. (4β), p. 119]

$$\begin{gathered}L_0^s(r^2)=1, \quad r\geqslant 0, \quad s\in\mathbf{N},\\ \int_0^\infty e^{-r^2}r^{2(\gamma-1)}L_n^\mu(r^2)dr^2=\frac{\Gamma(\gamma)\Gamma(1-\gamma+\mu+n)}{n!\Gamma(1-\gamma+\mu)}, \quad \gamma>0, \quad \mu,n\in\mathbf{N},\end{gathered} \tag{16}$$

it follows from (13) that

$$\int_0^\infty \phi_0^s(r)\phi_{2l+k}^{s+k}(r)dr^2=\begin{cases}0 & \text{when } s\geqslant l,\\ \dfrac{(-1)^s l!(l+k-1)!}{(l-s-1)!\sqrt{s!(s+k)!(2l+k)!}} & \text{when } 0\leqslant l\end{cases}$$

for all $l,k\in\mathbf{Z}^+$, since for each $m\in\mathbf{N}$, $|\Gamma(-m+\epsilon)|\to\infty$, $\epsilon\to 0$, and the right-hand side of (16) is zero when $1-\gamma+\mu=-m$. If one chooses $n=0$, and $m=2l+k$, $l\in\mathbf{Z}^+$, in (15), one gets

$$\sum_{s=0}^{l-1} T_{s,s+k}\int_0^\infty \phi_0^s(r)\phi_{2l+k}^{s+k}(r)dr^2=0. \tag{17}$$

We next fix $k\in\mathbf{Z}^+$. When one substitutes $l=1$ in (17), it follows that $T_{0,k}=0$. If $T_{s,s+k}=0$ for all $s\leqslant h-1$, $h\in\mathbf{Z}^+$, we substitute $l=h+1$ to (17) to get $T_{h,h+k}=0$. We conclude that if E^T is a phase observable, then T is diagonal in the number basis, and thus of the form $T=\sum_{s=0}^\infty \lambda_s|s\rangle\langle s|$, $\lambda_s\geqslant 0$, $s\in\mathbf{N}$ and $\sum_{s=0}^\infty \lambda_s=1$. This concludes the proof of the theorem. □

Remark 4.2: The proof of the above theorem shows that if $E^T(X)=A^T([0,\infty)\times X)$, $X\in\mathcal{B}[0,2\pi)$, is a phase observable, then the structure constants $c_{n,m}$, $n,m\in\mathbf{N}$, of E^T are

$$c_{n,m}=\sum\nolimits_{s=0}^{\infty} T_{s,s}\int_0^\infty \phi_n^s(r)\phi_m^s(r)dr^2. \tag{18}$$

One may thus easily confirm that $c_{n,n}=1$, $n\in\mathbf{N}$. Indeed, since the associated Laguerre polynomials satisfy the relation [Ref. 11, Eq. 7.414(3)]

$$\int_0^\infty e^{-r^2}r^{2\alpha}L_n^\alpha(r^2)L_m^\alpha(r^2)dr^2=\frac{(\alpha+n)!}{n!}\delta_{n,m}, \quad \alpha,n,m\in\mathbf{N}, \tag{19}$$

we have

$$\int_0^\infty \phi_n^s(r)\phi_n^s(r)dr^2=1, \quad n,s\in\mathbf{N}.$$

Therefore, it follows from (18) that

$$c_{n,n}=\sum_{s=0}^{\infty} T_{s,s}\int_0^{\infty}\phi_n^s(r)\phi_n^s(r)dr^2=\sum_{s=0}^{\infty} T_{s,s}=\mathrm{tr}(T)=1, \quad n\in\mathbf{N}.$$

Since $T\geqslant O$, we also observe that $c_{n,m}\in\mathbf{R}$, $n,m\in\mathbf{N}$. For all $\psi\in\mathcal{H}$, $k\in\mathbf{N}$, one gets

$$\sum_{n,m=0}^{k} c_{n,m}\langle\psi|n\rangle\langle m|\psi\rangle=\sum_{s=0}^{\infty} T_{s,s}\int_0^{\infty}\left|\sum_{m=0}^{k}\phi_m^s(r)\langle m|\psi\rangle\right|^2 dr^2\geqslant 0,$$

which confirms that the operators $\Sigma_{n,m=0}^{k} c_{n,m}|n\rangle\langle m|$, $k\in\mathbf{N}$, are positive. It is also easy to verify the condition $|c_{n,m}|\leqslant 1$, $n,m\in\mathbf{N}$, directly by using the Cauchy–Schwarz inequality:

$$\begin{aligned}|c_{n,m}|&\leqslant\sum_{s=0}^{\infty} T_{s,s}\left|\int_0^{\infty}\phi_n^s(r)\phi_m^s(r)dr^2\right|\\ &\leqslant\sum_{s=0}^{\infty} T_{s,s}\sqrt{\int_0^{\infty}\phi_n^s(r)\phi_n^s(r)dr^2}\sqrt{\int_0^{\infty}\phi_m^s(r)\phi_m^s(r)dr^2}\\ &=\mathrm{tr}(T)=1,\end{aligned}$$

for all $n,m\in\mathbf{N}$.

We note finally that there are phase observables which are not angle margins of some phase space observables. For instance, if we assume that there is a state T such that the phase observable E of Example 3.4 is of the form $E=E^T$, then (18) and (16) give

$$1=c_{0,2}=T_{0,0}\frac{1}{\sqrt{2}}.$$

But this means that $T_{0,0}=\sqrt{2}>1$, which is impossible.

V. THE FIRST MOMENT OPERATOR OF A PHASE OBSERVABLE AND THE NUMBER OPERATOR AS A HEISENBERG PAIR

The phase observables $E:\mathcal{B}[0,2\pi)\rightarrow\mathcal{L}(\mathcal{H})$ are compactly supported real operator measures. Therefore, their moment operators of all order $k\in\mathbf{N}$, $E^{(k)}:=\int_0^{2\pi}\theta^k\,\mathrm{d}E(\theta)$, are bounded self-adjoint operators. In particular,

$$E^{(1)}=\sum_{n\neq m=0}^{\infty}\frac{ic_{n,m}}{m-n}|n\rangle\langle m|+\pi I,$$

and one may calculate the formal commutator of the number operator N and $E^{(1)}$:

$$[N,E^{(1)}]=iI-i\sum_{n,m=0}^{\infty} c_{n,m}|n\rangle\langle m|$$

This shows that the operators $E^{(1)}$ and N form a Heisenberg pair[13] exactly when the series $\Sigma_{n,m=0}^{\infty} c_{n,m}|n\rangle\langle m|$ vanishes in a dense subspace of $\mathcal{H}$ which is contained in the domain of the commutator $[N,E^{(1)}]$.

The first moment operator of the phase observable of Example 3.4 and the number operator N form a Heisenberg pair.[14] In Example 3.3 we have $c_{n,m}=e^{i(\nu_n-\nu_m)}$, $\nu_n\in\mathbf{R}$, $n,m\in\mathbf{N}$. The related phase observables are obtained from Example 3.4 by the simple unitary transformation $|n\rangle\mapsto e^{i\nu_n}|n\rangle$. Thus the first moment operators of the observables of Example 3.4 form also Heisenberg pairs with N. The domains of commutators are then

$$\mathcal{D}_E := \left\{ \psi \in \mathcal{D}(N) \,\middle|\, \sum_{m=0}^{\infty} e^{-i\nu_m} \langle m|\psi\rangle = 0 \right\},$$

in which $[N,E^{(1)}]=iI$. Apart from that it is to be stressed that the moment operator $E^{(1)}$ is of a very limited use. First, since the phase observable E is not a projection valued measure, its first moment operator $E^{(1)}$ does not suffice to determine E. Second, it carries an essential nonuniqueness, since $e^{i\theta N}E^{(1)}e^{-i\theta N}=E^{(1)}-\theta I+2\pi E[0,\theta),\ \theta\in[0,2\pi)$. Third, though $\mathcal{D}_E$ is a dense subset of $\mathcal{H}$, it does not contain physically relevant vector states, as entailed by the well-known no-go theorem for the existence of a canonical phase operator.[3]

The phase observable of Example 3.2 gives an example of a pair $(E^{(1)},N)$ which is not a Heisenberg pair. Indeed, in that case we have

$$[N,E^{(1)}]=-i(z|s\rangle\langle t|+\overline{z}|t\rangle\langle s|)\neq iI$$

in $\mathcal{H}\setminus\{0\}$.

[1] W. P. Schleich and S. M. Barnett (eds.), *Quantum Phase and Phase Dependent Measurements*, Phys. Scr. **T48**, (1993).
[2] R. Lynch, ''The quantum phase problem: A critical review,'' Phys. Rep. **256**, 367–436 (1995).
[3] D. A. Dubin, M. A. Hennings, and T. B. Smith, ''Mathematical aspects of quantum phase,'' Int. J. Mod. Phys. B **9**, 2597–2687 (1995).
[4] E. B. Davies, *Quantum Theory of Open Systems* (Academic, London, 1976).
[5] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).
[6] G. Ludwig, *Foundations of Quantum Mechanics* (Springer, Berlin, 1983), Vol. I.
[7] P. Busch, M. Grabowski, and P. Lahti, *Operational Quantum Physics* (Springer, Berlin, 1997).
[8] A. S. Holevo, ''Generalized imprimitivity systems for Abelian groups,'' Sov. Math. (Iz. VUZ) **27**, 53–80 (1983).
[9] S. K. Berberian, *Notes on Spectral Theory* (Van Nostrand, Princeton, 1966).
[10] W. Stulpe, *Classical Representations of Quantum Mechanics* (Wissenschaft & Technik, Berlin, 1997).
[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products. Corrected and Enlarged Edition* (Academic, New York, 1980).
[12] H. Buchholz, *The Confluent Hypergeometric Function* (Springer, Berlin, 1969).
[13] J. C. Garrison and J. Wong, ''Canonically conjugate pairs, uncertainty relations, and phase operators,'' J. Math. Phys. **11**, 2242–2249 (1970).
[14] A. Galindo, ''Phase and number,'' Lett. Math. Phys. **8**, 495–500 (1984).

The semispin groups in string theory

Brett McInnes[a]
Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Republic of Singapore

(Received 30 March 1999; accepted for publication 12 May 1999)

In string theory, an important role is played by certain Lie groups which are locally isomorphic to $\mathrm{SO}(4m)$, $m \leqslant 8$. It has long been known that these groups are actually isomorphic not to $\mathrm{SO}(4m)$ but rather to the groups for which the half-spin representations are faithful, which we propose to call $\mathrm{Semispin}(4m)$. (They are known in the physics literature by the ambiguous name of ''$\mathrm{Spin}(4m)/Z_2$.'') Recent work on string duality has shown that the distinction between $\mathrm{SO}(4m)$ and $\mathrm{Semispin}(4m)$ can have a definite physical significance. This work is a survey of the relevant properties of $\mathrm{Semispin}(4m)$ and its subgroups. © *1999 American Institute of Physics.* [S0022-2488(99)03510-0]

I. INTRODUCTION

From a physical point of view, gauge theories have a serious drawback: the construction works equally well for all compact Lie groups. Even when the Lie algebra of the gauge group is known, the global structure of the group itself is not fixed by any fundamental principle. For example, it can be argued that the gauge group of the standard model[1] is ''really'' $[\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)]/Z_6$ rather than $\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)$; but since all known particles fall into multiplets which can be regarded as representations of *both* of these groups, there is (at present) no way of deciding the issue other than by an appeal to parsimony.

It is one of the many virtues of string theory that it puts an end to all uncertainty on this score.[2] The theory not only specifies the dimension of the gauge group (at 496 in the Type I and heterotic theories) but also its global structure within each version. In the heterotic ''$E_8\times E_8$'' theory, this is quite straightforward. First, there is in any case *only one* connected group with the Lie algebra of $E_8\times E_8$, namely $E_8\times E_8$ itself. Second, there is *only one* nontrivial disconnected Lie group with $E_8\times E_8$ as an identity component, namely the semidirect product $(E_8\times E_8)\lhd Z_2$, where Z_2 acts by exchanging the E_8 factors. As the corresponding string theory (initially) treats the two factors symmetrically, we conclude that the global version of the gauge group is $(E_8\times E_8)\lhd Z_2$. (The significance of disconnected gauge groups is discussed in Refs. [3–6]. By ''nontrivial'' we mean to exclude, for example, direct products of finite groups with $E_8\times E_8$, which are of little or no physical interest.)

The ''SO(32)'' cases (Type I and heterotic) are much less straightforward, because there are many nontrivial groups with the same Lie algebra as $\mathrm{SO}(n)$; in fact, there are eight nontrivial groups with the ''SO(32)'' algebra. It has been known from the beginning[7] that string theory selects from these eight a group known in the physics literature as $\mathrm{Spin}(32)/Z_2$, a most unfortunate convention, which only exacerbates the tendency to confuse this group with SO(32). This group is closely associated with (and is essentially defined by) the *half-spin* representations of Spin(32), and so we propose the name Semispin(32) for it. As we shall see, there is no non-trivial disconnected Lie group with Semispin(32) as identity component, so the gauge group is connected in this case.

To summarize, the heterotic string theories fix the global structures of their gauge groups. One theory uses the disconnected but simply connected group $(E_8\times E_8)\lhd Z_2$, while the other uses the connected but not simply connected group Semispin(32).

[a]Electronic mail: matmcinn@nus.edu.sg

The semispin groups are perhaps the least familiar of the compact simple Lie groups, and there is a venerable tradition of treating Semispin$(4m)$ as if it were the same as SO$(4m)$. We wish to argue that this tradition has outlived its usefulness, that string theory forces us to be fully aware of the differences between Semispin$(4m)$ and the other groups with the same algebra. There are two physically significant kinds of distinction, one representation theoretic, the other topological.

First, note that while such ambiguities have often arisen in the past, one of the groups in question has always been a *cover* of the other. For example, ''SO(10)'' grand unification[8] uses a certain 16-dimensional multiplet which does *not* correspond to any representation of SO(10). It is, of course, a representation of Spin(10). One can solve this ''problem'' by simply reading Spin(10) for SO(10); no harm is done, but only because every representation of SO(10) is automatically a representation of Spin(10), the latter being a *cover* of the former. In the opposite direction, one normally writes SU(3)×SU(2)×U(1) for (SU(3)×SU(2)×U(1)/Z_6, the ''true'' group,[1] with no ill effects because every representation of the latter is a representation of the former. The novelty in string theory is that *neither* SO(32) *nor* Semispin(32) is a cover of the other. Consequently, *both* have representations which cannot be regarded as representations of the other. The situation here is quite different to the superficially analogous ambiguities arising in earlier gauge theories.

Second, there exist Semispin(32) gauge configurations (over topologically nontrivial space–times) which are of considerable physical importance, but which *cannot* be interpreted as SO(32) gauge configurations.[9,10] The reader might argue that one can likewise construct SO(10) configurations which do not lift to Spin(10). The point, however, is just that ordinary gauge theory does not provide any fundamental justification for thinking that SO(10) is important. String theory, by contrast, does favor Semispin(32) over SO(32). The analysis of Semispin(32) bundles which cannot be regarded as Spin(32) or SO(32) bundles is therefore physically significant.

Finally, the study of duality[11] brings both points together in a potentially very confusing way. The *T*-duality between the two heterotic theories relies on relating ''$E_8 \times E_8$'' and ''SO(32)'' through their supposed common subgroup, ''SO(16)×SO(16).'' A global investigation shows that *no such common subgroup exists*; worse still, neither of the actual respective subgroups covers the other; worse yet again, each has representations which are *not* representations of the other, but which *are* crucial in establishing duality. Solving this problem leads to further topological obstructions, and, in the background, one has ''Wilson loops'' behaving in a way that depends very delicately on the global structure of various subgroups of Semispin(32) and Semispin(16). In short, the local simplicity of the duality argument conceals considerable complexity at the global level.

The purpose of this work is not to solve all of these problems, but rather to give a useful survey of those aspects of the Semispin groups (and their subgroups) which are most directly relevant to string theory. The main emphasis is on the structure of the groups themselves rather than their representations, since the latter are well understood and since it is the former which is needed for dealing with topological obstructions and for analyzing the effect of Wilson loops.

We begin with a brief survey of the family of nontrivial Lie groups with the algebra of SO(n).

II. GROUPS WITH THE ALGEBRA OF SO(n)

In order to understand the ways in which Semispin$(4m)$ differs from the other groups with the same Lie algebra, it is useful to begin with a complete classification. We refer the reader to Ref. 12 for the basic techniques, or to Ref. 13 for a much simpler account.

We shall not assume that the gauge group is connected; we have already seen that this would not be justifiable in one heterotic theory. On the other hand, it is true that most disconnected Lie groups are of little physical interest. Every compact Lie group can be expressed as a finite union of connected components,

$$G = G_0 \cup \gamma_1 \bullet G_0 \cup \gamma_2 \bullet G_0 \cup \cdots ,$$

where G_0 contains the identity and the γ_i are not elements of G_0. The nonidentity components of a gauge group are particularly important if space–time is not simply connected, since in that case parallel transport of particles around noncontractible paths (''Wilson loops'') can affect conserved charges.[3–5] However, a given component, $\gamma_i \bullet G_0$, can only give rise to such effects if γ_i *cannot* be chosen so as to commute with every element of G_0. The physically interesting disconnected groups are those such that none of the γ_i can be chosen to commute with every element of G_0. Such a group is called a *natural extension* of its identity component. For example, $(E_8 \times E_8) \lhd Z_2$ is a natural extension of $E_8 \times E_8$, and it is in fact the only other natural extension. (It is convenient to adopt the convention that a connected group is a natural extension of itself.) Henceforth, we confine attention to disconnected groups which are natural extensions of their identity components.

Next, some definitions. Let $\mathrm{Pin}(n), n \geqslant 2$, be defined as usual[14] in terms of a Clifford algebra with a basis $\{e_i\}$. We can write $\mathrm{Pin}(n)$ as a natural extension of $\mathrm{Spin}(n)$, when n is even,

$$\mathrm{Pin}(2m) = \mathrm{Spin}(2m) \cup e_1 \bullet \mathrm{Spin}(2m).$$

Notice that this is not necessarily a semidirect product, since $(e_1)^2 = -1 \in \mathrm{Spin}(2m)$. However, $\mathrm{Spin}(2m) \lhd Z_2$ can be defined (with the generator of Z_2 acting in the same way as conjugation by e_1); it is actually isomorphic to $\mathrm{Pin}(2m)$ if m is even, but not if m is odd. It, too, is a natural extension of $\mathrm{Spin}(2m)$. (There is no natural extension of $\mathrm{Spin}(n)$, other than itself, when n is odd.)

Let $\hat{K}_{m,n}$ be defined by

$$\hat{K}_{m,n} = \prod_{i=m}^{n} e_i$$

and set $\hat{K}_m = \hat{K}_{1,m}$. Then the center of $\mathrm{Spin}(n)$ is $\{\pm 1\}$ if n is odd, while the center of $\mathrm{Spin}(2m)$ is $\{\pm 1, \pm \hat{K}_{2m}\}$. Since $(\hat{K}_{2m})^2 = (-1)^m$, the center is Z_4 if m is odd, but $Z_2 \times Z_2$ if m is even.[15] Here we think of $\{1, \hat{K}_{2m}\}$ as the first Z_2, $\{1, -\hat{K}_{2m}\}$ as the second, and $\{\pm 1\}$ as the diagonal. Of course, we have

$$\mathrm{Spin}(n)/\{\pm 1\} = \mathrm{SO}(n) \quad \text{for all} \quad n \geqslant 2.$$

When n is odd, $\mathrm{SO}(n)$ has no natural extension other than itself, but when n is even it has two others. The first is $\mathrm{O}(2m)$, which may be expressed as

$$\mathrm{O}(2m) = \mathrm{SO}(2m) \cup A_{2m} \bullet \mathrm{SO}(2m),$$

where A_{2m} is a $(2m) \times (2m)$ orthogonal matrix satisfying $A_{2m}^2 = I_{2m}, \det A_{2m} = -1$. Thinking of $\mathrm{O}(2m)$ as the real subgroup of $\mathrm{U}(2m)$, we can also define

$$\mathrm{Oi}(2m) = \mathrm{SO}(2m) \cup i A_{2m} \bullet \mathrm{SO}(2m);$$

this group is also a natural extension of $\mathrm{SO}(2m)$, and it is not isomorphic to $\mathrm{O}(2m)$.

When n is even, $\mathrm{SO}(n)$ has a nontrivial quotient,

$$\mathrm{PSO}(2m) = \mathrm{SO}(2m)/\{\pm I_{2m}\},$$

the projective special orthogonal group. We can define $\mathrm{PO}(2m)$ as the same quotient of $\mathrm{O}(2m)$, and it is a natural extension of $\mathrm{PSO}(2m)$. Notice that $\mathrm{PSO}(2m)$ can be obtained directly from $\mathrm{Spin}(2m)$ by factoring out the entire center.

When n is a multiple of 4, we can also consider the quotients $\mathrm{Spin}(4m)/\{1, \hat{K}_{4m}\}$ and $\mathrm{Spin}(4m)/\{1, -\hat{K}_{4m}\}$. Let $\mathrm{Ad}(e_1)$ denote conjugation by e_1 in $\mathrm{Pin}(4m)$; then $\mathrm{Ad}(e_1)$ is an automorphism of $\mathrm{Spin}(4m)$, and

$$\mathrm{Ad}(e_1)\hat{K}_{4m} = -\hat{K}_{4m}.$$

It follows that $\mathrm{Spin}(4m)/\{1,\hat{K}_{4m}\}$ and $\mathrm{Spin}(4m)/\{1,-\hat{K}_{4m}\}$ are mutually isomorphic. Thus we obtain only one group in this way, not two. We define

$$\mathrm{Semispin}(4m) = \mathrm{Spin}(4m)/\{1,\hat{K}_{4m}\}.$$

This group is isomorphic to $\mathrm{SO}(4m)$ only if $\mathrm{Spin}(4m)$ admits an automorphism which maps $\hat{K}_{4m}$ to -1; but no such automorphism exists, except when $m=2$. Leaving that case to one side, $\mathrm{Ad}(e_1)$ is, up to inner automorphisms, the only outer automorphism of $\mathrm{Spin}(4m)$. Since $\mathrm{Ad}(e_1)$ does not map $\{1,\hat{K}_{4m}\}$ into itself, we see that, unlike $\mathrm{Spin}(4m)$, $\mathrm{SO}(4m)$, and $\mathrm{PSO}(4m)$, $\mathrm{Semispin}(4m)$ has no outer automorphism if $m \neq 2$. If, therefore, G is a compact disconnected group with $\mathrm{Semispin}(4m)$ as its identity component,

$$G = \mathrm{Semispin}(4m) \cup \gamma_1 \bullet \mathrm{Semispin}(4m) \cup \cdots,$$

then $\mathrm{Ad}(\gamma_i)$ must, for all i, be inner; $\mathrm{Ad}(\gamma_i) = \mathrm{Ad}(s_i)$ for some s_i in $\mathrm{Semispin}(4m)$. Thus $\gamma_i s_i^{-1}$ commutes with every element of $\mathrm{Semispin}(4m)$, and so we see that, when $m \neq 2$, $\mathrm{Semispin}(4m)$ has no natural extension other than itself.

When $m=2$, we have Spin(8), which has the triality map,[14] an automorphism of order three. This combines with $\mathrm{Ad}(e_1)$ to give D_6, the dihedral group of order six. Triality maps $\hat{K}_8$ to -1, so in fact

$$\mathrm{Semispin}(8) = \mathrm{SO}(8).$$

This is the only dimension in which the Semispin construction gives nothing new. Triality does not descend to SO(8) (because it does not preserve $\{\pm 1\}$) but it does descend to PSO(8).

We are now in a position to state the following theorem, the proof of which is an application of techniques given in Refs. 12 and 13.

Theorem 1: Let G be a compact Lie group which is a natural extension of its identity component. If the Lie algebra of G is isomorphic to that of $\mathrm{SO}(n)$, $n \geq 2$, then G is globally isomorphic to a group in the following list:

(1) $n=2$: SO(2), O(2), Pin(2).
(2) $n=$ odd: $\mathrm{SO}(n)$, $\mathrm{Spin}(n)$.
(3) $n=4m+2$, $m \geq 1$: $\mathrm{Spin}(n)$, $\mathrm{Pin}(n)$, $\mathrm{Spin}(n) \lhd Z_2$, $\mathrm{SO}(n)$, $\mathrm{O}(n)$, $\mathrm{Oi}(n)$, $\mathrm{PSO}(n)$, $\mathrm{PO}(n)$.
(4) $n=4m$, $m \neq 2$: $\mathrm{Spin}(n)$, $\mathrm{Pin}(n)$, $\mathrm{SO}(n)$, $\mathrm{O}(n)$, $\mathrm{Oi}(n)$, $\mathrm{PSO}(n)$, $\mathrm{PO}(n)$, $\mathrm{Semispin}(n)$.
(5) $n=8$: Spin(8), Pin(8), $\mathrm{Spin}(8) \lhd Z_3$, $\mathrm{Spin}(8) \lhd D_6$, SO(8), O(8), Oi(8), PSO(8), PO(8), $\mathrm{PSO}(8) \lhd Z_3$, $\mathrm{PSO}(8) \lhd D_6$.

Note that $\mathrm{Spin}(4) = \mathrm{SU}(2) \times \mathrm{SU}(2)$, $\mathrm{PSO}(4) = \mathrm{SO}(3) \times \mathrm{SO}(3)$, and $\mathrm{Semispin}(4) = \mathrm{SU}(2) \times \mathrm{SO}(3)$.

These, then, are the nontrivial distinct groups corresponding to the $\mathrm{SO}(n)$ algebra. When n is 32, there are no fewer than eight candidates. String theory selects a particular group from among these eight in the following extraordinary way. In the heterotic theories, gauge fields arise in connection with the lattice of momenta on a 16-dimensional torus. The lattice must be even and self-dual. The crucial point is that these requirements impose conditions not merely on the root system of the gauge group, *but also on its integral lattice*.[16] However, there is a deep connection between the integral lattice and the global structure of a compact, connected Lie group. Thus string theory provides a route from strictly physical conditions directly to the global structure of the (identity component of the) gauge group. As is well known, Semispin(32) satisfies these conditions, while SO(32), Spin(32), and PSO(32) do not. The argument is now completed by a glance at Theorem 1; we see that Spin(32) and PSO(32) each have a nontrivial disconnected version, and SO(32) has two, but Semispin(32) has none. The precise global structure of the gauge group is thereby fixed; it is Semispin(32).

We close this section with some remarks on the representation theory of Spin$(4m)$, Semispin$(4m)$, SO$(4m)$, and PSO$(4m)$. Recall that the basic faithful representation of Spin$(4m)$, obtained[14] by suitably restricting an irreducible representation of the Clifford algebra, has a canonical decomposition

$$\Delta_{4m}=\Delta_{4m}^{+}\oplus\Delta_{4m}^{-},$$

where Δ_{4m}^{+} is a representation with kernel $\{1,\hat{K}_{4m}\}$ and Δ_{4m}^{-} has kernel $\{1,-\hat{K}_{4m}\}$. Thus, neither Δ_{4m}^{+} nor Δ_{4m}^{-} is faithful; one must take their sum. Hence Δ_{4m}^{+} and Δ_{4m}^{-} are called the *half-spin representations*. Evidently they are faithful not on Spin$(4m)$ but on the group we (accordingly) call Semispin$(4m)$. The half-spin representations are of dimension 2^{2m-1}. Thus the so-called ''128-dimensional representation of SO(16)'' which plays a prominent role in string theory is in fact the defining representation of Semispin(16). Again, the defining representation of Semispin(32) is 32 768-dimensional, a decidedly inconvenient value. Fortunately, we have

$$\mathrm{PSO}(32)=\mathrm{Semispin}(32)/Z_2,$$

and so every representation of PSO(32) is automatically a representation of Semispin(32); thus, the latter has a more manageable (but unfaithful) 496-dimensional representation, which is also an unfaithful representation of SO(32) and Spin(32), namely, the adjoint. Similarly PSO(16) yields a 120-dimensional representation of Semispin(16), and so the latter has a *faithful* 248-dimensional representation defined by the direct sum, $\mathbf{120}\oplus\mathbf{128}$. As the representation is faithful, and as the (likewise faithful) adjoint of E_8 decomposes as $\mathbf{248}=\mathbf{120}\oplus\mathbf{128}$, this immediately shows that E_8 contains Semispin(16) and *not*, as is so often said, SO(16). Thus $(E_8\times E_8)\lhd Z_2$ has a maximal subgroup of the form $(\mathrm{Semispin}(16)\times\mathrm{Semispin}(16))\lhd Z_2$, and so we see that *the Semispin groups appear in both heterotic string theories*. In fact, Witten[17] has recently argued that the same is true of the Type I theory. The gauge group of Type I at the perturbative level is PO(32) (see Theorem 1). As this group is disconnected, while Semispin(32) has no nontrivial disconnected version, this appears to obstruct the supposed *S*-duality between the Type I and the ''SO(32)'' heterotic string theories.[11] However, Witten shows that a subtle nonperturbative effect breaks PO(32) to $\mathrm{PSO}(32)=\mathrm{Semispin}(32)/Z_2$; furthermore, there appear to be Type I nonperturbative states transforming ''spinorially'' under the gauge group. (Note that, like a spinor, a ''semispinor'' is odd under a 2π rotation; the nontrivial element in the center of Semispin(32) is the projection of -1 in Spin(32).) In short, the gauge group of Type I string theory is undoubtedly Semispin(32) precisely, not SO(32). The Semispin groups appear in all three string theories with nontrivial gauge groups.

All this appears to bode well for duality: in particular, since $(E_8\times E_8)\lhd Z_2$ contains $(\mathrm{Semispin}(16)\times\mathrm{Semispin}(16))\lhd Z_2$, one would expect this same group to appear on the Semispin(32) side. In fact, this is *not* the case, as we now show.

III. SUBGROUPS OF SEMISPIN$(4m)$ CORRESPONDING TO SO$(k)\times$SO$(4m-k)$

Evidently SO(32) contains SO(16)×SO(16) block-diagonally; more generally, SO$(4m)$ contains SO$(k)\times$SO$(4m-k)$, $k\geqslant 2$. The product is indeed direct, since SO(k) and SO$(4m-k)$ intersect trivially, in $\{I_{4m}\}$. However, Spin$(4m)$ *does not* contain Spin$(k)\times$Spin$(4m-k)$, because both factors contain $\{\pm 1\}$. In fact, the subgroup is Spin$(k)\bullet$Spin$(4m-k)$, a local direct product, where

$$\mathrm{Spin}(k)\bullet\mathrm{Spin}(4m-k)=(\mathrm{Spin}(k)\times\mathrm{Spin}(4m-k))/Z_2,$$

with Z_2 generated by $(-1,-1)$.

There is another important difference between SO$(4m)$ and Spin$(4m)$ in this area. It is clear that, when $k=2j$ is even, Spin$(k)\bullet$Spin$(4m-k)$ can be characterized as the group of all Spin$(4m)$ elements which commute with (that is, as the *centralizer* of) $\hat{K}_{2j}$. Now $\hat{K}_{2j}$ projects to the SO$(4m)$ matrix diag$(-I_{2j},I_{4m-2j})=K_{2j}$, but the SO$(4m)$ centralizer of K_{2j} is not SO$(2j)\times$SO$(4m-2j)$; rather, it is the disconnected subgroup $S(\mathrm{O}(2j)\times\mathrm{O}(4m-2j))$, the set of all pairs ($A$, B) in O$(2j)\times$O$(4m-2j)$ such that det A=det B. That is, a Wilson loop that breaks Spin$(4m)$ to a *connected* subgroup will break SO$(4m)$ to a *disconnected* subgroup. (Recall[2] that a Wilson loop in a gauge theory is a closed curve in space–time which has a nontrivial holonomy element even in the vacuum. The gauge group is broken to the centraliser of the (usually finite) subgroup generated by the holonomy element.)

Now we turn to the case of the Semispin$(4m)$. Suppose first that k is odd. Then Spin$(k)\bullet$Spin$(4m-k)$ does not contain $\hat{K}_{4m}$, and so it is unaffected by the projection from Spin$(4m)$ to Semispin$(4m)$. Thus, when k is odd, the subgroup of Semispin$(4m)$ corresponding to SO$(k)\times$SO$(4m-k)$ is globally isomorphic to Spin$(k)\bullet$Spin$(4m-k)$. Next, suppose that $k=2j$ is even but not a multiple of 4. Then $4m-2j$ is likewise even but not a multiple of 4, and so Spin$(2j)$ and Spin$(4m-2j)$ have Z_4 centers generated, respectively, by $\hat{K}_{2j}$ and $\hat{K}_{2j+1,4m}$. We have

$$\hat{K}_{2j}\hat{K}_{4m}=-\hat{K}_{2j+1,4m}$$

and

$$\hat{K}_{2j+1,4m}\hat{K}_{4m}=-\hat{K}_{2j},$$

and so the effect of factoring by $\hat{K}_{4m}$ is to identify the *entire* center of Spin$(4m-2j)$ with that of Spin$(2j)$. We have, when j is odd,

$$\mathrm{Spin}(2j)\mathbf{:}\mathrm{Spin}(4m-2j)=(\mathrm{Spin}(2j)\times\mathrm{Spin}(4m-2j))/Z_4$$

as the subgroup of Semispin$(4m)$ corresponding to SO$(2j)\times$SO$(4m-2j)$.

Finally, if $k=4j$ is a multiple of 4, then so is $4m-4j$ and both Spin$(4j)$ and Spin$(4m-4j)$ have centers isomorphic to $Z_2\times Z_2$. These centres are $\{\pm 1,\pm\hat{K}_{4j}\}$ and $\{\pm 1,\hat{K}_{4j+1,4m}\}$, respectively, and since we have

$$\pm\hat{K}_{4j}\hat{K}_{4m}=\pm\hat{K}_{4j+1,4m}$$

and

$$\pm\hat{K}_{4j+1,4m}\hat{K}_{4m}=\pm\hat{K}_{4j},$$

we see that, once again, the effect of the projection Spin$(4m)\to$Semispin$(4m)$ is to identify the entire center of Spin$(4m-4j)$ with that of Spin$(4j)$. We use the notation

$$\mathrm{Spin}(4j)\mathbf{:}\mathrm{Spin}(4m-4j)=(\mathrm{Spin}(4j)\times\mathrm{Spin}(4m-4j))/(Z_2\times Z_2).$$

Next, recall that, from a physical point of view, we are interested in obtaining all these groups as centralizers of some element in Semispin$(4m)$. We saw earlier that the centralizer of $\hat{K}_{2j}$ in Spin$(4m)$ is connected, but that of K_{2j} in SO$(4m)$ is not. Let K^*_{2j} be the projection of $\hat{K}_{2j}$ to Semispin$(4m)$. The centralizer of K^*_{2j} will include Spin$(2j)$**:**Spin$(4m-2j)$, but it will also include any Semispin$(4m)$ element L^* such that $\hat{L}$, a lift of L^* to Spin$(4m)$, satisfies $\hat{L}\hat{K}_{2j}=\hat{K}_{4m}\hat{K}_{2j}\hat{L}$. Projecting this to SO$(4m)$, we find that the corresponding matrices satisfy

$$LK_{2j}L^{-1}=-K_{2j},$$

whence Trace $K_{2j}=4(m-j)=0$. Thus if $j\neq m$, L^* does not exist, and so the centraliser of K^*_{2j} in Semispin$(4m)$ is precisely Spin$(2j)$**:**Spin$(4m-2j)$. If $j=m$, then we have

$$J_{2m}K_{2m}J_{2m}^{-1}=-K_{2m},$$

where

$$J_{2m}=\begin{pmatrix} 0 & -I_{2m} \\ I_{2m} & 0 \end{pmatrix}.$$

This solution is essentially unique. The corresponding element of Spin$(4m)$ is (see Ref. 16, p. 174, and modify suitably)

$$\hat{J}_{2m}=2^{-m}(1-e_1e_{1+2m})(1-e_2e_{2+2m})\cdots(1-e_{2m}e_{4m}).$$

Now,

$$\begin{aligned}\hat{K}_{2m}\hat{J}_{2m}&=2^{-m}(e_1+e_{1+2m})(e_2+e_{2+2m})\cdots(e_{2m}+e_{4m})\\ &=2^{-m}(e_{1+2m}+e_1)(e_{2+2m}+e_2)\cdots(e_{4m}+e_{2m})=\hat{J}_{2m}\hat{K}_{1+2m,4m}\\ &=(-1)^m\hat{K}_{4m}\hat{J}_{2m}\hat{K}_{2m},\end{aligned}$$

since $(\hat{K}_{2m})^2=(-1)^m$ and $\hat{K}_{2m}\hat{K}_{1+2m,4m}=\hat{K}_{4m}$. Thus if m is even, the projections to Semispin$(4m)$ satisfy $K_{2m}^*J_{2m}^*=J_{2m}^*K_{2m}^*$ as required. If m is odd, we project instead to Spin$(4m)/\{1,-\hat{K}_{4m}\}$ and recall that this is isomorphic to Semispin$(4m)$. A further exercise in Clifford algebra shows that

$$(\hat{J}_{2m})^2=(-1)^m\hat{K}_{4m},$$

and so the appropriate projections are of order two. The effect on Spin$(2m)$**:**Spin$(2m)$ of conjugation by J_{2m}^* is to exchange the two factors. We conclude that the *centralizer* of K_{2j}^* in Semispin$(4m)$ is

$$\text{Spin}(2j)\mathbf{:}\text{Spin}(4m-2j)\quad\text{if}\quad j\neq m,$$

$$(\text{Spin}(2m)\mathbf{:}\text{Spin}(2m))\lhd Z_2\quad\text{if}\quad j=m.$$

Finally, let us consider the specific case of Semispin(32). Its ''SO(16)×SO(16)'' subgroup is actually (Spin(16)**:**$Spin(16)$)$\lhd Z_2$ where the full center of each Spin(16) is identified with that of the other, and where Z_2 exchanges the two factors. Compare this with the ''SO(16)×SO(16)'' subgroup of $(E_8\times E_8)\lhd Z_2$, which is (Semispin(16)×Semispin(16))$\lhd Z_2$. Both groups have the ''exchange'' Z_2, which is welcome from the point of view of T-duality. But Spin(16)**:**Spin(16) *is not isomorphic to* Semispin(16)×Semispin(16). Both are $Z_2\times Z_2$ quotients of Spin(16)×Spin(16), but $Z_2\times Z_2$ acts differently in each case. This implies that neither is a cover of the other, and so they each have representations which cannot be regarded as representations of the other. For example, by factoring out $\{\pm 1\}$ in Spin(16)**:**Spin(16), we obtain SO(16)·SO(16), where the dot means that the two factors have a nontrivial intersection, $\{\pm I_{16}\}$. The tensor product of the vector with itself, (**16**, **16**), is faithful for this group, and so it is a representation of Spin(16)**:**Spin(16). This representation contains faithful copies of SO(16), something which is impossible for any representation of Semispin(16)×Semispin(16). On the other hand, let $(\mathbf{128},\mathbf{1})\oplus(\mathbf{1},\mathbf{128})$ be the defining representation of Semispin(16)×Semispin(16); this representation distinguishes the centers of the two Semispin(16) factors, which cannot happen in any representation of Spin(16)**:**Spin(16). (If we take the quotient of Spin(16)**:**Spin(16) by $\{1,\hat{K}_{16}\}$, then we obtain Semispin(16)•Semispin(16), in which the two factors intersect in $\{\pm 1\}$, where we use -1 to denote the projection of -1 in Spin(16); but $(\mathbf{128},\mathbf{1})\oplus(\mathbf{1},\mathbf{128})$ does not descend to a representation of this group.) The two groups do have some representations in common, such as (**128**,**128**) and $(\mathbf{120},\mathbf{1})\oplus(\mathbf{1},\mathbf{120})$, the latter being the defining representation for PSO(16)×PSO(16), which is a $Z_2\times Z_2$ quotient of both Semispin(16)×Semispin(16) and Spin(16)**:**Spin(16). However, while $(\mathbf{120},\mathbf{1})\oplus(\mathbf{1},\mathbf{120})$ is important for duality, so also are $(\mathbf{16},\mathbf{16})$ and $(\mathbf{128},\mathbf{1})\oplus(\mathbf{1},\mathbf{128})$.

We see, then, that the appearance of Semispin groups in both heterotic theories was somewhat deceptive; for Semispin$(4m)$ is strangely unlike Spin$(4m)$ and SO$(4m)$. While these last contain subgroups of the same kind as themselves, Spin$(2j)\bullet$Spin$(4m-2j)$ and SO$(2j)\times$SO$(4m-2j)$ respectively, Semispin$(4m)$ *does not contain* Semispin$(2j)\bullet$Semispin$(4m-2j)$. Instead, it contains Spin$(2j)$**:**Spin$(4m-2j)$. Thus we arrive at the disconcerting fact that while $E_8\times E_8$ contains Semispin groups, Semispin(32) itself does not. The ''common SO(16)×SO(16) subgroup'' which appears in the duality literature not only fails to be isomorphic to SO(16)×SO(16); it simply *does not exist*.

One way to approach this problem is to find a group which covers both Semispin(16)×Semispin(16) and Spin(16)**:**Spin(16), since all of the representations of *both* groups will then be representations of that group. One obvious choice is Spin(16)×Spin(16), but there is a better alternative, constructed as follows. Write Spin(16)×Spin(16) as $\text{Spin}(16)^L \times \text{Spin}(16)^R$ and let $\{\pm 1^L, \pm \hat{K}_{16}^L\}$, $\{\pm 1^R, \pm \hat{K}_{16}^R\}$ be the respective centers. We define

$$\text{Spin}(16)*\text{Spin}(16)=(\text{Spin}(16)\times\text{Spin}(16))/\{(1^L,1^R),(\hat{K}_{16}^L,\hat{K}_{16}^R)\}.$$

That is, we identify $\hat{K}_{16}^R$ with $\hat{K}_{16}^L$. Further factoring by this element produces Semispin(16)×Semispin(16), while factoring by $(-1^L,-1^R)$ produces Spin(16)**:**Spin(16). That is, Spin(16)∗Spin(16) is a double cover of *both* Semispin(16)×Semispin(16) *and* Spin(16)**:**Spin(16). Hence $(\mathbf{16},\mathbf{16})$, $(\mathbf{128},\mathbf{1})\oplus(\mathbf{1},\mathbf{128})$, and $(\mathbf{120},\mathbf{1})\oplus(\mathbf{1},\mathbf{120})$ are all (unfaithful) representations of Spin(16)∗Spin(16), and $(\mathbf{16},\mathbf{16})\oplus(\mathbf{128},\mathbf{1})\oplus(\mathbf{1},\mathbf{128})$ is a *faithful* representation of Spin(16)∗Spin(16), though it is not a representation of Semispin(16)×Semispin(16) or Spin(16)**:**Spin(16), much less SO(16)×SO(16). (It is faithful because $\hat{K}_{16}$, the one nontrivial element of Spin(16)∗Spin(16) mapped to the identity by $(\mathbf{128},\mathbf{1})\oplus(\mathbf{1},\mathbf{128})$, acts as -1 in $(\mathbf{16},\mathbf{16})$.)

In fact, Spin(16)∗Spin(16) is the gauge group of the unique tachyon-free ten-dimensional nonsupersymmetric heterotic string theory,[18,19] which plays a central role in recent investigations of strong-coupling duality.[20] The massless spectrum of this theory consists of a gravity multiplet, space–time vectors assigned to $(\mathbf{120},\mathbf{1})\oplus(\mathbf{1},\mathbf{120})$, and space–time spinors assigned to the ''SO(16)×SO(16)'' representation $(\mathbf{16},\mathbf{16})\oplus(\mathbf{128},\mathbf{1})\oplus(\mathbf{1},\mathbf{128})$, which, as we have seen, is a faithful representation of Spin(16)∗Spin(16).

We claim, then, that the string theorist's ''SO(16)×SO(16)'' is actually Spin(16)∗Spin(16). The strange feature of this conclusion is that Spin(16)∗Spin(16) *is not a subgroup of either* $E_8 \times E_8$ or Semispin(32). (Nor can it be embedded in Spin(32), SO(32), or PSO(32).) Thus *it does not make sense* to speak of breaking $E_8\times E_8$ or Semispin(32) to Spin(16)∗Spin(16) by a Wilson loop or in any other way. We believe that the way to solve this problem is through a study of ''generalized Stiefel–Whitney classes.''[9,10] For example, to establish the duality of a certain Semispin(32) configuration with an $(E_8\times E_8)\lhd Z_2$ configuration, one breaks Semispin(32) to (Spin(16)**:**Spin(16))$\lhd Z_2$, lifts this to a (Spin(16)∗Spin(16))$\lhd Z_2$ structure (checking that the appropriate generalized Stiefel–Whitney class vanishes), projects this to a (Semispin(16)×Semispin(16))$\lhd Z_2$ structure, and then extends to $(E_8\times E_8)\lhd Z_2$ The details of this process will be described elsewhere.

Let us summarize as follows: SO$(4m)$, $m>2$, has important subgroups of the form SO(k) ×SO$(4m-k)$, though in fact this is just the identity component of $S(\text{O}(k)\times \text{O}(4m-k))$. The other three connected groups locally isomorphic to SO$(4m)$, namely, Spin$(4m)$, Semispin$(4m)$, and PSO$(4m)$, have analogous subgroups given by the following Theorem.

Theorem 2: The Lie algebra inclusion $\mathcal{SO}(k)\oplus\mathcal{SO}(4m-k)\rightarrow\mathcal{SO}(4m)$ has the following counterparts at the Lie group level:

$$S(\text{O}(k)\times\text{O}(4m-k))\rightarrow\text{SO}(4m),$$

$$\mathrm{Spin}(k)\bullet\mathrm{Spin}(4m-k)\rightarrow\mathrm{Spin}(4m),$$

$$\mathrm{PS}(\mathrm{O}(k)\times\mathrm{O}(4m-k))\rightarrow\mathrm{PSO}(4m),$$

$$\mathrm{Spin}(k)\bullet\mathrm{Spin}(4m-k)\rightarrow\mathrm{Semispin}(4m)(k\ \ \text{odd}),$$

$$\mathrm{Spin}(2j)\mathbf{:}\mathrm{Spin}(4m-2j)\rightarrow\mathrm{Semispin}(4m)(k=2j,j\neq m),$$

$$(\mathrm{Spin}(2m)\mathbf{:}\mathrm{Spin}(2m))\lhd Z_2\rightarrow\mathrm{Semispin}(4m).$$

Here a single dot denotes a factoring by a diagonal Z_2, as also does the prefix P, while the double dot denotes a factoring by a diagonal Z_4 or by $Z_2\times Z_2$ as the case may be.

For the particularly important case of ''SO(16)×SO(16),'' we have the following diagram.

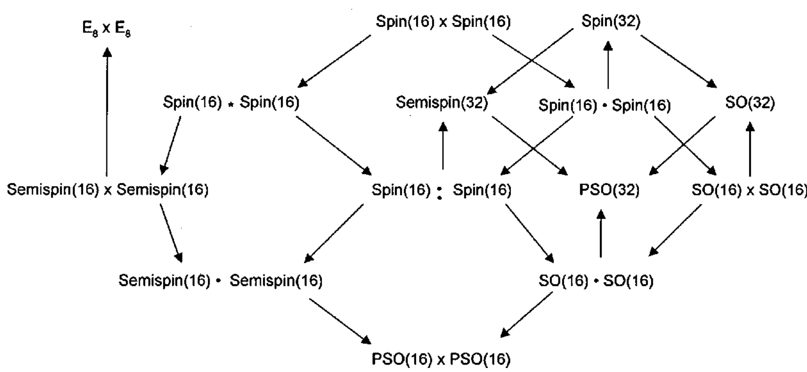

Here, an upward arrow indicates a subgroup, a downward arrow corresponds to a Z_2 factoring. The rows in the diagram therefore reflect the size of the fundamental group; the top row groups are simply connected, the next row has Z_2 as a fundamental group, and so down to PSO(16)×PSO(16), with fundamental group isomorphic to $Z_2\times Z_2\times Z_2\times Z_2$. From the diagram we see at once that Spin(16)**:**Spin(16) is a subgroup of Semispin(32), but there is no chain of arrows leading from Spin(16)**:**Spin(16) to $E_8\times E_8$. Finally, note that, with the sole exception of Semispin(32), every group G in the diagram has a natural extension of the form $G\lhd Z_2$.

IV. SUBGROUPS OF SEMISPIN $(4m)$ CORRESPONDING TO U$(2m)$

Another subgroup of SO$(4m)$ which plays an important role in the string literature (see, for example, Refs. 9, 20, 21) is the unitary group U$(2m)$. If $A+iB$ is any element of U$(2m)$, where A and B are real, then $\begin{pmatrix} A & -B \\ B & A \end{pmatrix}$ is an element of SO$(4m)$. The unitary subgroup can also be characterized as the centraliser of the matrix J_{2m} defined in the preceding section. (Notice that J_{2m} is SO$(4m)$-conjugate to diag $\left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\cdots\right)$, and the reader can take J_{2m} to be defined in this way if that is convenient.)

Now U$(2m)$ is *not* isomorphic to U(1)×SU$(2m)$, because U(1) and SU$(2m)$ intersect non-trivially. Let z be a primitive $(2m)$th root of unity, and let Z_{2m} act on U(1)×SU$(2m)$ by

$$(u,s)\rightarrow(uz^{-1},zs).$$

Then (U(1)×SU$(2m)$)/Z_{2m} is isomorphic to U$(2m)$. Elements of U$(2m)$ may therefore be represented as equivalence classes, $[u,s]_{2m}$.

Now we ask, what is the subgroup of Spin$(4m)$ which projects onto U$(2m)$? It is useful to notice that the answer cannot be isomorphic to U$(2m)$, for U$(2m)$ would be of maximal rank in Spin$(4m)$, and so the center of Spin$(4m)$ would be contained in the center of U$(2m)$; that is, we would have $Z_2\times Z_2$ contained in U(1), which is impossible. (This argument would not work for the $U(2m+1)$ subgroup of $SO(4m+2)$, and indeed the cover of $U(2m+1)$ in $Spin(4m+2)$ is again isomorphic to $U(2m+1)$.) In fact, it is not difficult to see that $Spin(4m)$ contains $(U(1)\times SU(2m))/Z_m$, which consists of pairs $[u,s]_m$. In this group, $[z^{-1},zI_{2m}]_m$ is not the identity, though $[z^{-1},zI_{2m}]_{2m}=1$; therefore $[z^{-1},zI_{2m}]_m$ corresponds to -1 in $Spin(4m)$. There is another important element of order two in this group, $[1,-I_{2m}]_m$, but of course there are others, such as $[z^{-1},-zI_{2m}]_m$. In order to determine the structure of the subgroup of $Semispin(4m)$ corresponding to $U(2m)$, we must determine which of these corresponds to $\hat{K}_{4m}$.

Theorem 3: Let $SemiU(2m)$ denote the projection of $(U(1)\times SU(2m))/Z_m$ to $Semispin(4m)$. Then the global structure of $SemiU(2m)$ is given as follows:

$$SemiU(2m)=[U(1)\times(SU(2m)/Z_2)]/Z_{m/2},\quad m \text{ even}$$
$$=[U(1)\times SU(2m)]/Z_m,\quad m \text{ odd}.$$

Proof: Under the embedding of $U(2m)$ in $SO(4m)$, the matrix J_{2m} arises from the $U(2m)$ matrix iI_{2m}, which is $[i,I_{2m}]_{2m}$. Thus we see that the $Spin(4m)$ element $\hat{J}_{2m}$ defined in the preceding section must be either $[i,I_{2m}]_m$ or $[iz^{-1},zI_{2m}]_m$. In either case we have $(\hat{J}_{2m})^2=[-1,I_{2m}]_m$. Now recall that $(\hat{J}_{2m})^2=(-1)^m\hat{K}_{4m}$ and that we have agreed to define $Semispin(4m)$ by $Spin(4m)/\{1,(-1)^m\hat{K}_{4m}\}$ for convenience, so that J^*_{2m} is always of order two. (See the remarks at the end of this section.) Thus when m is odd, we must factor by

$$-\hat{K}_{4m}=[-1,I_{2m}]_m\neq[1,-I_{2m}]_m.$$

Clearly, the factoring will affect U(1) but not $SU(2m)$. However, $U(1)/Z_2=U(1)$, since the map $u\rightarrow u^2$ is a group epimorphism for this infinite Abelian group. Thus we obtain $[U(1)\times SU(2m)]/Z_m$ when m is odd. When m is even, $\hat{K}_{4m}$ is $[-1,I_{2m}]_m$, which is equal to $[1,-I_{2m}]_m$. The factoring will affect both U(1) and $SU(2m)$ in this case, and, after it, the Z_2 in Z_m will act trivially; so we obtain $(U(1)/Z_2)\times(SU(2m)/Z_2)$, with an effective action by $Z_{m/2}$. Hence the group is $[(U(1)\times(SU(2m)/Z_2)]/Z_{m/2}$, and this completes the proof.

Notice that, according to this theorem,

$$SemiU(2)=[U(1)\times SU(2)]/Z_1=SO(2)\times SU(2),$$

which is indeed a subgroup of $Semispin(4)=SO(3)\times SU(2)$. Again,

$$SemiU(4)=[U(1)\times(SU(4)/Z_2)]/Z_1=SO(2)\times SO(6),$$

which is contained in $SO(8)=Semispin(8)$.

Clearly $SemiU(2m)$ is the identity component of the centralizer, in $Semispin(4m)$, of J^*_{2m}, the projection of $\hat{J}_{2m}$. Recall that, unlike J_{2m} in $SO(4m)$ and $\hat{J}_{2m}$ in $Spin(4m)$ (which are both of order 4), $\hat{J}^*_{2m}$ is of order 2; this is important for applications.[9,21] For example, consider a Semispin(32) heterotic theory compactified on a $K3$ surface which is a Kummer surface at an orbifold limit, with a pointlike instanton at the singular point.[21] Excising this point, we obtain a neighborhood which retracts to the projective sphere, S^3/Z_2. If J^*_{16} were of order four, then it could not be realized as a holonomy element over S^3/Z_2, and so the gauge group would not break. But J^*_{16} is of order two, and so it *can* be realized as a holonomy over S^3/Z_2. (In the literature it is always *assumed* that a finite group F can always be realised as a holonomy group over manifolds of the form M/F. This is true—with very mild conditions—but not at all obvious.[6]) Then (unless one arranges to avoid it[21]) Semispin(32) will break to the centralizer of J^*_{16} in Semispin(32). This includes SemiU(16), but it also includes K^*_{16}, as we saw in the preceding section. The matrix $K_{16}=\mathrm{diag}(-I_{16},I_{16})$ acts by complex conjugation on U(16), that is,

$$K_{16}\begin{pmatrix}A & -B\\ B & A\end{pmatrix}K_{16}^{-1}=\begin{pmatrix}A & B\\ -B & A\end{pmatrix}.$$

Similarly, conjugation by K_{16}^* maps elements of SemiU(16) to their complex conjugates. (A typical element of SemiU(16) $=[\mathrm{U}(1)\times(\mathrm{SU}(16)/Z_2)]/Z_4$ has the form $[u^2,[s]_2]_4$, with complex conjugate $[(\overline{u}^2,[\overline{s}]_2]_4$; bear in mind that Z_4 acts on $\mathrm{U}(1)\times(\mathrm{SU}(16)/Z_2)$ by $(u^2,[s]_2)\rightarrow(iu^2,[zs]_2)$, with $z^8=1$). As $(K_{16}^*)^2=1$, we see that J_{16}^* breaks Semispin(32) to SemiU(16)$\lhd Z_2$ and *not*[9,21] to U(16) or U(16)/Z_2, which are quite different to SemiU(16). Notice the differences with SO(32); J_{16} is of order four, and its centralizer in SO(32) is the connected group U(16), while J_{16}^* is of order two, with Semispin(32) centraliser isomorphic to the *disconnected* group SemiU(16)$\lhd Z_2$. On the other hand, J_{16} cannot break SO(32) over S^3/Z_2, but J_{16}^* can break Semispin(32). (Note also that there do exist bundles over S^3/Z_2 having the full disconnected group SemiU(16)$\lhd Z_2$ as holonomy group.[6])

In the same way, J_8^* is of order two, and it breaks Semispin(16) to SemiU(8)$\lhd Z_2$, with Z_2 generated by K_8^*, and with SemiU(8)$=[\mathrm{U}(1)\times(\mathrm{SU}(8)/Z_2)]/Z_2$. However, Semispin(16) is mainly of interest because it is a maximal subgroup of E_8. If J_8^* is embedded in E_8 through Semispin(16), then of course its centraliser must contain SemiU(8)$\lhd Z_2$; however, this cannot be the full centraliser, since the centralizer of any element of a simply connected compact Lie group (such as E_8, but not Semispin(16)) must be connected. Hence the centralizer of J_8^* must be a connected subgroup of E_8 containing SemiU(8)$\lhd Z_2$. Of course, Semispin(16) is such a subgroup, but there is another. The exceptional Lie group E_7 has a maximal rank subgroup[22] isomorphic to SU(8)/Z_2, and in fact one can prove that E_7 contains a disconnected subgroup with two connected components, one being SU(8)/Z_2. Combining this with a Pin(2) subgroup of SU(2), we obtain, after suitable identifications, SemiU(8)$\lhd Z_2$ as a subgroup of SU(2)$\bullet E_7$, which is a maximal subgroup of E_8. In fact, the centralizer of J_8^* in E_8 is SU(2)$\bullet E_7$, while that of $-J_8^*$ turns out to be just Semispin(16); this is important in applications.[9]

One of the most interesting and important applications where the distinction between SO($4m$) and Semispin($4m$) is crucial concerns $K3$ compactifications of the $(E_8\times E_8)\lhd Z_2$ heterotic theory. When the instanton numbers are assigned symmetrically to the two factors, one finds[9] that the corresponding (T-dual) ''SO(32)'' configuration corresponds to a Semispin(32) bundle which *does not* lift to a Spin(32) bundle. This is the Semispin analogue of the failure of the orthonormal frame bundles over certain Riemannian manifolds[14] to lift to spin bundles. If a Semispin(32) bundle does lift to a Spin(32) bundle, then it will automatically define (by projection) an SO(32) bundle; and so a Semispin(32) bundle which fails to lift to a Spin(32) bundle is said to lack a ''vector structure.''

Examples of such Semispin($4m$) bundles can be given by once again exploiting the fact that J_{2m}^* is of order two, whereas $\hat{J}_{2m}$, its counterpart in Spin($4m$), satisfies $(\hat{J}_{2m})^2=(-1)^m\hat{K}_{4m}$ and so is of order four (like J_{2m}). This makes it possible to construct a nontrivial U(1) bundle over a two-cycle in the base, such that connections on this bundle satisfy the usual (''Dirac'') integrality conditions, but their pullbacks to a covering bundle would not. When this U(1) bundle is extended to a Semispin(32) bundle, therefore, the latter cannot be lifted to a double cover. It is in precisely this way that the dual partner of the above[9] $(E_8\times E_8)\lhd Z_2$ compactification is constructed. One could not wish for a more striking confirmation of the importance of the distinction between SO($4m$) and Semispin($4m$).

In this spirit, we ask whether U(1) is indeed the precise global form of the gauge group in question. This U(1) may be identified as the explicit U(1) in $[\mathrm{U}(1)\times(\mathrm{SU}(16)/Z_2)]/Z_4$, but we know that this group is most naturally regarded as the identity component of SemiU(16)$\lhd Z_2$. Therefore one should really regard the canonical U(1) in Semispin(32) as the identity component of U(1)$\lhd Z_2$ or O(2) in the notation of Theorem 1. The corresponding subgroup of SO(32) consists of all 32×32 matrices of the form

$$\begin{pmatrix} I_{16}\cos\theta & -I_{16}\sin\theta \\ I_{16}\sin\theta & I_{16}\cos\theta \end{pmatrix},\quad \begin{pmatrix} -I_{16}\cos\theta & I_{16}\sin\theta \\ I_{16}\sin\theta & I_{16}\cos\theta \end{pmatrix}.$$

More generally, Semispin($4m$) has a canonical subgroup of the form

$$\mathrm{U}(1)\cup K^*_{2m}\bullet \mathrm{U}(1),$$

where U(1) corresponds to the Lie algebra element

$$\begin{pmatrix} 0 & -I_{2m} \\ I_{2m} & 0 \end{pmatrix},$$

or to its conjugate

$$\mathrm{diag}\left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},\ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},\dots\right)$$

if one prefers. Recalling that $(K^*_{2m})^2=(-1)^m$, we see that the global structure is O(2) if m is even, but Pin(2) if m is odd (see Theorem 1). One can actually prove that Semispin$(4m)$ has no Pin(2) subgroup containing U(1) when m is even, and no O(2) subgroup containing U(1) when m is odd.

In constructing Semispin$(4m)$ bundles without ''vector structure,'' then, one should really begin with nontrivial O(2) or Pin(2) bundles. (Of course, if the base manifold is simply connected, such a bundle will always reduce to a U(1) bundle, but realistic string compactifications are not likely to be simply connected.) Now we know that a U(1) instanton breaks Semispin$(4m)$ to SemiU$(2m)$; what is the corresponding subgroup for O(2) or Pin(2)? Since K^*_{2m} acts by complex conjugation on all of SemiU$(2m)$, the answer is the real subgroup of SemiU$(2m)$. The real subgroup of SU$(2m)$ is SO$(2m)$, while that of SU$(2m)/Z_2$ is PSO$(2m)$, and U(1) contributes J^*_{2m}; finally, -1, the central element of Semispin$(4m)$, must of course also be included. Thus an O(2) or Pin(2) instanton will break Semispin$(4m)$ to

$$Z_2\times Z_2\times \mathrm{PSO}(2m),\quad m\ \text{even},$$

$$Z_2\times Z_2\times \mathrm{SO}(2m),\quad m\ \text{odd},$$

with one Z_2 generated by -1 and the other by J^*_{2m}. In particular, then, an O(2) instanton in a Semispin(32) theory will reveal itself by the presence of PSO(16), where SemiU(16) might be expected. (Note that this same PSO(16) arises in the Spin(16)**:**Spin(16) subgroup of Semispin(32), as the diagonal subgroup.)

Before concluding this section, we draw the reader's attention to the following point. While it is true that Spin$(4m)/\{1,-\hat{K}_{4m}\}$ is isomorphic to Spin$(4m)/\{1,\hat{K}_{4m}\}$, the isomorphism is through an outer automorphism of Spin$(4m)$ which can change the way in which a given subalgebra is embedded in the algebra of Spin$(4m)$, and this in turn can affect the global structure of the subgroup to which that subalgebra exponentiates. A simple example is provided by Spin(4)=SU(2)×SU(2). Obviously SU(2)×SO(3) is isomorphic to SO(3)×SU(2), but it is true that a given, fixed SU(2) algebra exponentiates either to SU(2) or to SO(3), depending on whether one factors by $\{1,\hat{K}_4\}$ or $\{1,-\hat{K}_4\}$. We have chosen to define Semispin$(4m)$ by factoring $\{1,-\hat{K}_{4m}\}$ when m is odd, but one could decide to factor by $\{1,\hat{K}_{4m}\}$, though in that case J^*_{2m} will not commute with K^*_{2m} and it will not be of order two. If one does this, then $[\mathrm{U}(1)\times \mathrm{SU}(2m)]/Z_m$ no longer projects to $[(\mathrm{U}(1)/Z_2)\times \mathrm{SU}(2m)]/Z_m$. Instead we have

$$\hat{K}_{4m}=-(\hat{J}_{2m})^2=[-z^{-1},zI_{2m}]_m,$$

where z is a primitive $(2m)$th root of unity. This gives us $\hat{K}_{4m}=[z^{m-1},zI_{2m}]_m=[1,z^mI_{2m}]_m$, because, since m is odd, $m-1$ is even. Thus in fact $\hat{K}_{4m}=[1,-I_{2m}]_m$, and so the quotient of $(\mathrm{U}(1)\times \mathrm{SU}(2m))/Z_m$ by $\{1,\hat{K}_{4m}\}$ is isomorphic to $[\mathrm{U}(1)\times(\mathrm{SU}(2m)/Z_2)]/Z_m$, which is not isomorphic to $[\mathrm{U}(1)/Z_2\times \mathrm{SU}(2m)]/Z_m$. Thus there is no unique subgroup of Semispin$(4m)$ corresponding to U$(2m)$ unless one specifies precisely which projection from Spin$(4m)$ to Semispin$(4m)$ one proposes to use.

Our point of view is that for physical applications it is important that J^*_{2m} should be of order two rather than, like J_{2m} and $\hat{J}_{2m}$, of order four. This is stressed repeatedly, for example, in Ref. 9. This fixes the projections; we must factor out $\{1,\hat{K}_{4m}\}$ when m is even, and $\{1,-\hat{K}_{4m}\}$ when m is odd.

V. SUBGROUPS OF SEMISPIN(4*m*) CORRESPONDING TO Sp(1)•Sp(*m*)

Another subgroup of $\mathrm{SO}(4m)$ which is important in various applications (see, for example, Ref. 21) is the symplectic group $\mathrm{Sp}(m)$, which embeds through $\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)$. The latter has the global structure $[\mathrm{Sp}(1)\times\mathrm{Sp}(m)]/Z_2$. The corresponding subgroup of $\mathrm{Semispin}(4m)$ is given as follows.

Theorem 4: The Lie algebra inclusion $\mathcal{S}p(1)\oplus\mathcal{S}p(m)\rightarrow\mathcal{SO}(4m)$ has the following counterparts at the Lie group level:

$$\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)\rightarrow\mathrm{SO}(4m),$$

$$\mathrm{Sp}(1)\times\mathrm{Sp}(m)\rightarrow\mathrm{Spin}(4m)\quad m\ \text{odd},$$

$$\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)\rightarrow\mathrm{Spin}(4m)\quad m\ \text{even},$$

$$\mathrm{SO}(3)\times\mathrm{PSp}(m)\rightarrow\mathrm{PSO}(4m),$$

$$\mathrm{SO}(3)\times\mathrm{Sp}(m)\rightarrow\mathrm{Semispin}(4m)\quad m\ \text{odd},$$

$$\mathrm{SO}(3)\times\mathrm{PSp}(m)\rightarrow\mathrm{Semispin}(4m)\quad m\ \text{even}.$$

Here a dot denotes a factoring by a diagonal Z_2, and $\mathrm{PSp}(m)=\mathrm{Sp}(m)/Z_2$.

Proof: Consider first the case of $\mathrm{Spin}(4m)$. We know that $\mathrm{Spin}(4m)$ contains a subgroup of the form $[\mathrm{U}(1)\times\mathrm{SU}(2m)]/Z_m$, consisting of pairs $[u,s]_m$. Evidently we have

$$[-1,I_{2m}]_m=[1,-I_{2m}]_m,$$

when m is even, but not when m is odd. That is, the Z_2 in U(1) is identified with the Z_2 in $\mathrm{SU}(2m)$ when m is even, but not when m is odd. However, this U(1) is contained in Sp(1), and $\mathrm{SU}(2m)$ contains $\mathrm{Sp}(m)$; furthermore the central Z_2 in Sp(1) is the Z_2 in U(1), and the central Z_2 in $\mathrm{Sp}(m)$ is identical to the Z_2 in $\mathrm{SU}(2m)$. Thus we see that the central Z_2 in Sp(1) is identified, in $\mathrm{Spin}(4m)$, with the central Z_2 in $\mathrm{Sp}(m)$, if and only if m is even. Hence the group is $\mathrm{Sp}(1)\times\mathrm{Sp}(m)$ if m is odd, but $\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)$ if m is even.

We saw, in the proof of Theorem 3, that $(-1)^m\hat{K}_{4m}=[-1,I_{2m}]_m$, the generator of the Z_2 in Sp(1). Hence $\mathrm{Sp}(1)\times\mathrm{Sp}(m)$ projects to $(\mathrm{Sp}(1)/Z_2)\times\mathrm{Sp}(m)$ when m is odd, while $\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)$ projects to $(\mathrm{Sp}(1)/Z_2)\times(\mathrm{Sp}(m)/Z_2)$ when m is even. Recalling that $\mathrm{Sp}(1)/Z_2=\mathrm{SO}(3)$ and $\mathrm{Sp}(m)/Z_2=\mathrm{PSp}(m)$, we have the stated results. Similarly, in $\mathrm{SO}(4m)$, the central Z_2 coincides with the Z_2 in Sp(1) and $\mathrm{Sp}(m)$, so taking the quotient throughout $\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)\rightarrow\mathrm{SO}(4m)$, we obtain $\mathrm{SO}(3)\times\mathrm{PSp}(m)\rightarrow\mathrm{PSO}(4m)$. This completes the proof.

Notice that the theorem asserts that SO(3)×Sp(1) is contained in Semispin(4), which is correct since the latter is SO(3)×SU(2) and SU(2)=Sp(1). It also asserts that Semispin(8)=SO(8) contains SO(3)×PSp(2), which is correct since Sp(2)=Spin(5) and so PSp(2)=SO(5). The theorem gives us SO(3)×PSp(8) as the subgroup of Semispin(32) corresponding to Sp(1)•Sp(8) in SO(32); this agrees with Ref. 21, where the importance of the SO(3) factor, appearing unexpectedly as a subgroup of Semispin(32), is explained. (Note that the full cover of $\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)$ in $\mathrm{Spin}(4m)$, m even, is actually $Z_2\times\mathrm{Sp}(1)\bullet\mathrm{Sp}(m)$, where $Z_2=\{\pm1\}$; this projects to $Z_2\times\mathrm{SO}(3)\times\mathrm{PSp}(m)$, so one might give this as the correct subgroup of $\mathrm{Semispin}(4m)$.)

VI. CONCLUSION

The Semispin groups are of fundamental importance in string theory. The gauge groups of Type I and one of the heterotic theories are both Semispin(32) precisely, while E_8 contains Semispin(16); the latter in turn contains (see Theorem 2) Spin(6)**:**Spin(10)$=[SU(4)\times$Spin(10)$]/Z_4$, and so it provides a possible route to Spin(10) grand unification.

These facts alone warrant a detailed study of the Semispin groups and their remarkable subgroups. Theorem 1 places Semispin($4m$) in the context of the entire family of nontrivial groups locally isomorphic to SO(n), while Theorems 2, 3, and 4 list the most important subgroups. We hope that these theorems will be a useful reference for string theorists.

The most surprising finding of this investigation is no doubt the fact that Semispin groups do not contain smaller Semispin groups. This implies that the ''common SO(16)$\times$SO(16) subgroup of $E_8\times E_8$ and Semispin(32)'' simply does not exist, which is obviously a problem for duality. This problem can be overcome by going to a common double cover, but only if certain topological obstructions vanish. In some circumstances, therefore, duality can be obstructed topologically. We shall study this phenomenon elsewhere.

[1] L. O'Raifeartaigh, *Group Structure of Gauge Theories* (Cambridge University Press, Cambridge, 1987).
[2] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, 1987).
[3] J. Kiskis, Phys. Rev. D **17**, 3196 (1978).
[4] J. Preskill and L. M. Krauss, Nucl. Phys. B **341**, 50 (1990).
[5] B. McInnes, Class. Quantum Grav. **14**, 2527 (1997).
[6] B. McInnes, J. Phys. A **31**, 3607 (1998).
[7] D. J. Gross, J. A. Harvey, E. Martinec, and R. Rohm, Nucl. Phys. B **256**, 253 (1985).
[8] G. G. Ross, *Grand Unified Theories* (Addison–Wesley, Reading, 1984).
[9] M. Berkooz, R. G. Leigh, J. Polchinski, J. H. Schwarz, N. Seiberg, and E. Witten, Nucl. Phys. B **475**, 115 (1996).
[10] E. Witten, J. High Energy Phys. **9802**, 006 (1998).
[11] J. H. Schwarz, Nucl. Phys. B, Proc. Suppl. **55B**, 1 (1997).
[12] J. de Siebenthal, Comment. Math. Helv. **31**, 41 (1956).
[13] B. McInnes, J. Math. Phys. **38**, 4354 (1997).
[14] H. B. Lawson and M. L. Michelsohn, *Spin Geometry* (Princeton University Press, Princeton, 1989).
[15] I. R. Porteous, *Clifford Algebras and the Classical Groups* (Cambridge University Press, Cambridge, 1995).
[16] T. Bröcker and T. tom Dieck, *Representations of Compact Lie Groups* (Springer-Verlag, Berlin, 1985).
[17] E. Witten, J. High Energy Phys. **9812**, 019 (1998).
[18] L. Alvarez-Gaumé, P. Ginsparg, G. Moore, and C. Vafa, Phys. Lett. B **171**, 155 (1986).
[19] L. J. Dixon and J. A. Harvey, Nucl. Phys. B **274**, 93 (1986).
[20] J. D. Blum and K. R. Dienes, Nucl. Phys. B **516**, 83 (1998).
[21] P. S. Aspinwall, Nucl. Phys. B **496**, 149 (1997).
[22] J. A. Wolf, *Spaces of Constant Curvature* (Publish or Perish, Wilmington, 1984).

Quantum superstring field theory in the $\mathcal{B}_0$-gauge and the physical scattering amplitudes

Seichi Naito[a)]
Department of Physics, Osaka City University, Sumiyoshiku, Osaka, Japan

(Received 21 September 1998; accepted for publication 22 February 1999)

We propose the (BRST-*invariant*) quantum open superstring field theory in the ''$\mathcal{B}_0$-gauge,'' based on Neveu–Schwarz (NS) strings in 1 picture and Ramond (R) strings in $\frac{1}{2}$ picture. We give the propagators of these open NS and R superstrings. In order to obtain the BRST-invariant interaction terms among these superstrings, we *modify* the interaction terms among *three* superstrings (i.e., among NS–NS–NS and R–R–NS) by subtracting the infinite number of counter terms, each of which involves interaction terms among ''*more than four* superstrings.'' The modified action can be obtained successively, so that resulting amplitudes in **g**-loops should become BRST invariant. Thus obtained amplitudes are referred to as the ''*amputated scatts*,'' with the help of which the *physical* scattering amplitudes can be expressed. These physical scattering amplitudes among N_B bosonic (N_F fermionic) particles are *calculated* by using the *analytic inlint gluing operator* (which has already been proposed and used in the quantum bosonic string field theory ''in the $\mathcal{B}_0=0$ gauge'').
[S0022-2488(99)02410-X]

I. INTRODUCTION AND PRELIMINARIES

There exist just five superstring theories:[1] type I and the two heterotic theories (all of which have $N=1$ supersymmetry in ten-dimensional space–time), type IIA (having two supercharges with the opposite chirality), and type IIB (having two supercharges with the same chirality). Type I theory is based on unoriented open and closed superstrings, while the other four theories are based on oriented closed superstrings. (These different superstring theories might be related with each other nonperturbatively.[2]) These superstring theories are very attractive ones, which are expected to unify interactions in nature, including the Yang–Mills theory as well as the gravitational theory. In order to make sure if these superstring theories are the realistic ones, we must clarify nonperturbative properties of these theories, which can be compared with experimental results. For this purpose, quantum superstring field theory (QSFT) might be expected to be helpful.

Since heterotic *super*string theories involve oriented *closed* bosonic string, Witten's quantum *open bosonic* string field theory[3] (open QBFT) in the ''$\mathcal{B}_0=0$ gauge'' (which does not include any *closed* string) is not realistic one, but it will be still useful. Unoriented open strings could be obtained by truncating the formulas in oriented open strings to those invariant under the ''twisting.'' Manifest factorization (in closed string channels) which is absent in Witten's open QBFT can be restored by including bosonic closed strings. In closed QBFT, there appear various complications, since we must include infinite numbers of the interaction terms among closed bosonic strings. These complications in *closed* QBFT have been analyzed[4] in detail. Based on these analyses, we think that various techniques[5] (which are useful in Witten's *open* QBFT) would be useful also in *closed* QBFT. In other words, Witten's open QBFT is useful as a mathematical model, in clarifying mathematical structures of physically realistic theories. On the other hand, all five superstring theories involve closed *super*strings, while type I *super*string theory involves the

[a)]Electronic mail: naito-@mx6.mesh.ne.jp

additional (unoriented) open *bosonic* string. Witten[6] has proposed the quantum (oriented) open superstring field theory (to be referred to as open QSFT) based on the Ramond[7]–Neveu–Schwarz[8] (RNS) superstrings, being suggested by Friedan–Martinec–Schenker's (FMS) work.[9] Compared with Witten's open QBFT, Witten's open QSFT involves many additional complications: We have fermionic particles as well as bosonic particles. Furthermore, there occur problems related with the (*so-called*) pictures. Unfortunately, Witten's open QSFT (based on open NS states in 0 picture and open R states in $\frac{1}{2}$ picture) is plagued with (Wendt's[10]) singularities, since *local* picture-changing operators can happen to collide with each other. Thus, Witten's open QSFT is not satisfactory, even at the perturbative level. In order to avoid this difficulty, another *open* QSFT (different from the Witten's one) has been proposed by Preitschopf–Thorn–Yost (PTY),[11] based on open NS superstring in 1 picture and open R superstring in $\frac{1}{2}$ picture. However, PTY have extensively considered open NS superstrings under the two special gauge-fixing conditions $\mathcal{B}_1+\mathcal{B}_{-1}=0$ or $\mathcal{B}_0-\frac{1}{2}\sqrt{-1}(\mathcal{B}_1-\mathcal{B}_{-1})=0$, where propagators in the open NS superstring involve respectively the factor $(L_1+L_{-1})^{-1}$ or $(L_0-\frac{1}{2}\sqrt{-1}(L_1-L_{-1}))^{-1}$, both of which are so much different from the ordinary $1/L_0$ (in QBFT[5]). We doubt if their formalism (with these propagators) would be useful in calculating physical scattering amplitudes.

In this paper, we exclusively analyze the (oriented) *open* QSFT in the ''$\mathcal{B}_0$-gauge,'' the possibility of which has only briefly been remarked on in Ref. 11. As preliminaries, we hereafter introduce ''fieldinos,'' etc. (which are to be used in QSFT). In defining them, we use various external operators which are given in appendices. In the case when the rth open superstring is in the $\pi(r)$ sector $[\pi(r)$ being either NS or R], we have in the open QSFT the following rth fieldino $|\Psi^{\pi(r)}\rangle_r$ in the $\pi(r)$ sector:

$$|\Psi^{\rm NS}\rangle_r \equiv P_r^{\rm NS}({\rm GSO})\left(\sum_B |B\rangle_r\cdot\phi_B\right)\quad \text{for}\ \ \pi(r)={\rm NS}, \tag{1.1}$$

and

$$|\Psi^{R}\rangle_r \equiv P_r^{R}({\rm GSO})\left(\sum_F |F\rangle_r\cdot\psi_F\right)\quad \text{for}\ \ \pi(r)={\rm R}. \tag{1.2}$$

Operators ϕ_B's (ψ_F's) in Eq. (1.1) [(1.2)] will be referred to as the *bose* fields [*fermi* fields] in the ten-dimensional space–time. The rth external GSO-projection operator $P_r^{\rm NS}$ (GSO) [P_r^R(GSO)] used in Eqs. (1.1) [(1.2)] will be defined soon by Eq. (1.11), and it GSO *projects out* NS states $|B\rangle_r$'s (R states $|F\rangle_r$'s), some of which might be *physical states* $|b\rangle_r$'s ($|f\rangle_r$'s). As we shall see later, all of these physical states $|b\rangle_r$'s ($|f\rangle_r$'s) can be *made* Grassman *odd* (*even*), by using *proper* cocycle factors. Operators ϕ_b's (ψ_f's) [which will be referred to as *physical* bose fields (fermi fields)] become subsequently Grassman *even* (*odd*). On the other hand, *unphysical* ϕ_B's (ψ_F's) might be Grassman *even* in some case and Grassman *odd* in another case. In this paper, $|\Psi^{\rm NS}\rangle_r$ in (1.1) and $|\Psi^{R}\rangle_r$ in (1.2) will be projected into the following fieldino in the ''$\mathcal{B}_0$-gauge,''[11] i.e., into the rth NS fieldino $|\Psi(1)\rangle_r$ ''in picture 1'' and R fieldino $|\Psi(\frac{1}{2})\rangle_r$ ''in picture $\frac{1}{2}$,'' respectively:

$$\begin{aligned}|\Psi(1)\rangle_r &\equiv \mathcal{P}_r^{\rm NS}|\Psi^{\rm NS}\rangle_r \quad \text{for}\ \ \pi(r)={\rm NS},\\ |\Psi(\tfrac{1}{2})\rangle_r &\equiv \mathcal{P}_r^{\rm R}|\Psi^{\rm R}\rangle_r \quad \text{for}\ \ \pi(r)={\rm R},\end{aligned} \tag{1.3}$$

where $\mathcal{P}_r^{\pi(r)}$ is the projection operator defined by

$$\begin{aligned}\mathcal{P}_r^{\rm NS} &\equiv P_r(1)\cdot P_r(\mathcal{B}_0)\cdot P_r(1) \quad \text{for}\ \ \pi(r)={\rm NS},\\ \mathcal{P}_r^{R} &\equiv P_r(\tfrac{1}{2})\cdot P_r(\mathcal{B}_0)\cdot P_r(\tfrac{1}{2}) \quad \text{for}\ \ \pi(r)={\rm R}.\end{aligned} \tag{1.4}$$

In Eq. (1.4), we have used $P_r(\mathcal{B}_0)$ defined by

$$P_r(\mathcal{B}_0) \equiv 1 - \mathcal{Q}_r^{\pi(r)} \cdot \frac{\mathcal{B}_{0,r}}{L_{0,r}} = \frac{\mathcal{B}_{0,r}}{L_{0,r}} \cdot \mathcal{Q}_r^{\pi(r)}, \tag{1.5}$$

which is the projection operator since it satisfies

$$(P_r(\mathcal{B}_0))^2 = P_r(\mathcal{B}_0). \tag{1.6}$$

[In Eq. (1.5), we have used operators $\mathcal{B}_{0,r}$ in (A25), $L_{0,r}$ in (B9b), and the BRST chargino $\mathcal{Q}_r^{\pi(r)}$ in (B27), together with the ''contacting formula'' (B38b)]. On the other hand, $P_r(1)$ and $P_r(\frac{1}{2})$ in Eq. (1.4) are respectively defined[11] by

$$P_r(1) \equiv \left(\sum_{\pm} \frac{1}{2} \cdot X_{\pm,r}^{1/2} \cdot X_{\mp,r}^{1/2} \right) Y[\overline{r}] \cdot Y[r] \cdot P_r^{\mathrm{NS}}(\mathrm{GSO}), \tag{1.7a}$$

and

$$P_r(\tfrac{1}{2}) \equiv X_r^0 \cdot Y[\overline{r}] \cdot P_r^{\mathrm{R}}(\mathrm{GSO}). \tag{1.7b}$$

In Eqs. (1.7a) and (1.7b), we have used the *nonlocal* picture-changing operators $X_{\pm,r}^{1/2}$ in (C25) (which has been introduced by Preitschopf–Thorn–Jost[11]), as well as X_r^0 in (C14).[12] On the other hand, the rth external operator $Y[\overline{r}]$ and $Y[r]$ are respectively the *lower* and *upper* inverse picture-changing operator defined [see $Y_r(w_r)$ in (C8)] by

$$Y[\overline{r}] \equiv Y_r(-\sqrt{-1}) \quad \text{and} \quad Y[r] \equiv Y_r(+\sqrt{-1}). \tag{1.8}$$

The rth external operator $P_r(1)$ in (1.7a) and $P_r(\frac{1}{2})$ in (1.7b) are respectively the projection operators (into 1 picture and $\frac{1}{2}$ picture), since we find

$$(P_r(p))^2 = P_r(p) \quad \text{for } p = 1 \quad \text{and } \tfrac{1}{2}, \tag{1.9}$$

by using Eqs. (C22)–(C24) and (C32)–(C36). With the help of the ''commutability'' (C38) and (C39), it can be proved[11] that

$$(\mathcal{P}_r^{\pi(r)})^2 = \mathcal{P}_r^{\pi(r)}. \tag{1.10}$$

Hereafter, $\mathcal{P}_r^{\pi(r)}$ will be referred to as the ''projection operator into the $\mathcal{B}_0$-gauge.'' We impose that both fieldino's $|\Psi^{\mathrm{NS}}\rangle_r$ in (1.1) and $|\Psi^{\mathrm{R}}\rangle_r$ in (1.2) are Grassman *odd*, and we should use in Eq. (1.1) [(1.2)] the following (rth external) GSO-projection operator[13] $P_r^{\pi(r)}$ (GSO) [for $\pi(r) = \mathrm{R}$, NS]:

$$P_r^{\pi(r)}(\mathrm{GSO}) \equiv \left(\frac{1}{2} + \frac{1}{2} \varepsilon^{\pi(r)} \cdot \exp\left(\sqrt{-1} \cdot \pi \left(p_r(\phi) + \sum_{j=0}^{4} p_r(\phi^j) \right) \right) \right)$$
$$\times \left(\prod_{\varphi = \varphi^j, \phi} \left(\frac{1}{2} + \frac{1}{2} \varepsilon^{\pi(r)} \cdot \exp(\sqrt{-1} \cdot 2\pi p_r(\varphi)) \right) \right), \tag{1.11}$$

where the constant $\varepsilon^{\pi(r)}$ is defined by

$$\varepsilon^{\pi(r)} \equiv \begin{cases} -1 & \text{for } \pi(r) = \mathrm{R}, \\ +1 & \text{for } \pi(r) = \mathrm{NS}. \end{cases} \tag{1.12}$$

In Eq. (1.11), $p_r(\phi)$ $[p_r(\phi^j)]$ is the zero-mode's operator in the fracting operator ϕ (the jth spining operator ϕ^j for $j=0-4$) defined by Eq. (A3). [In this paper, eigenvalues of $p_r(\phi)$ and $p_r(\phi^j)$ will be called respectively the fracting number and the jth spining number, and all of them should be integers (half-integers) in the NS (R) sector.] The last factor on the right-hand side of Eq. (1.11) shows that any eigenvalue $p_r(\varphi)$ (for $\varphi=\phi^j,\phi$) should be some (half-)integer in the NS (R) sector, while the first factor on the right-hand side of Eq. (1.11) shows that the eigenvalue of $p_r(\phi)+\Sigma_{j=0}^4 p_r(\phi^j)$ should be some even (odd) integer in the (rth external) GSO-allowed NS (R) sector.

At this stage, we summarize briefly our previously obtained[14] results on *physical* [rth external] NS (R) states, especially from the point of the projection operators $P_r(\mathcal{B}_0)$ in (1.5) and $P_r^{\rm NS}(1)$ in (1.7a)] ($P_r^R(1/2)$ in (1.7b)); (the rth external) Grassman *odd* physical NS states $|b(0)\rangle_r$ (in 0 picture) and $|b(1)\rangle_r$ (in 1 picture) have been constructed[14] respectively by Eqs. (D1) and (D5). In the on-shell limit (D4), these (rth external) *physical* NS states can be proved to have the following properties:

$$\mathcal{Q}_r^{\rm NS}|b(p)\rangle_r=\mathcal{B}_{0,r}|b(p)\rangle_r=L_{0,r}|b(p)\rangle_r=0 \quad (\text{for } \ p=0,1) \tag{1.13}$$

and

$$P_r^{\rm NS}(1)|b(1)\rangle_r=|b(1)\rangle_r. \tag{1.14}$$

Equations (1.13) and (1.14) show that physical states $|b(1)\rangle_r$ in (D5) are the *simultaneous* eigenstates of both $P_r(\mathcal{B}_0)$ and $P_r^{\rm NS}(1)$. As for the (rth external) Grassman *even* physical R states $|f(\pm\frac{1}{2})\rangle_r$ in $\pm\frac{1}{2}$ picture, they can be constructed[14] by Eqs. (D12) and (D13). [Incidentally, $\exp(\frac{1}{2}\cdot\phi^j)$ for $j=0-4$ will always be assumed to be Grassman even, by choosing proper cocycle factors.[15] In this paper, quantities named by ''-on'' (''-ino'') will be Grassman *even* (*odd*). Then, the last part on the right-hand side of the FMS spinor[9] in (A24) has been called the ''octon'' (''octino'')[14] in the case when $\Sigma_{j=1}^4\varepsilon_{h(r)}^j=$even (odd), while $\exp(\frac{1}{2}\cdot\phi^0)$ has been called the ''spinon.'' [14] On the other hand, we *can* consider *two* cases when the operator $\exp(-\frac{1}{2}\cdot\phi)$ is *either* Grassman even *or* odd, by choosing the different cocycle factors. In the *even* (*odd*) case, it will be referred to as the ''fracton'' (''fractino''). Therefore, the Grassman *even physical* state (D13) can be obtained in the ''fracton'' (''fractino'') case, by using the ''octon'' (''octino'') in Eq. (D13). This fact is important in constructing type IIA and type IIB closed superstring theories. Although we exclusively consider the ''fracton'' case in this paper, the ''fractino'' case can be similarly analyzed. Incidentally, GSO-projection operator (1.11) in the R sector is the one to be used in the ''fracton'' case. In the on-shell limit (D14), the [rth external] *physical* R states (D12) and (D13) are shown to have the following properties:

$$\mathcal{Q}_r^{\rm R}|f(\pm\tfrac{1}{2})\rangle_r=\mathcal{B}_{0,r}|f(\pm\tfrac{1}{2})\rangle_r=L_{0,r}|f(\pm\tfrac{1}{2})\rangle_r=0, \tag{1.15}$$

as well as

$$P_r^{\rm R}(\tfrac{1}{2})|f(+\tfrac{1}{2})\rangle_r=|f(+\tfrac{1}{2})\rangle_r. \tag{1.16}$$

The ''rth external BRST chargino in the π-representation'' is denoted by $\mathcal{Q}_r^\pi$ (for $\pi=$NS,R), which will be defined by Eq. (B27) and has been used in Eqs. (1.13) and (1.15). Incidentally, $\mathcal{Q}_r$ in (B28) will be referred to as the ''rth external BRST-chargino in the bo-(sonized) representation.'' These BRST charginos in *three* representations can be proved to be just equal to each other;

$$\mathcal{Q}_r^{\rm NS}=\mathcal{Q}_r^{\rm R}=\mathcal{Q}_r \quad [\text{see Eq. (B29)}]. \tag{1.17}$$

With the help of $|\Psi(1)\rangle_r$ in (1.3) and $|\Psi(\frac{1}{2})\rangle_r$ in (1.4), we investigate the open QSFT in the ''$\mathcal{B}_0$-gauge,'' which is described by the following gauge-fixed action $S_{\rm GF}(\Psi)$:

$$S_{\mathrm{GF}}(\Psi)=S_{\mathrm{KIN}}(\Psi)+S_{\mathrm{INT}}(\Psi), \tag{1.18}$$

where the kinetic term $S_{\mathrm{KIN}}(\Psi)$ is given by

$$S_{\mathrm{KIN}}(\Psi)\equiv\tfrac{1}{2}\cdot\langle\nu_{\mathcal{S}}(r,s)|Y[\overline{r}]\cdot Y[r]\cdot\mathcal{Q}_r^{\mathrm{NS}}(:|\Psi(1)\rangle_r\cdot|\Psi(1)\rangle_s:)$$
$$+\tfrac{1}{2}\cdot\langle\nu_{\mathcal{S}}(r,s)|Y[\overline{r}]\cdot\mathcal{Q}_r^{\mathrm{R}}(:|\Psi(\tfrac{1}{2})\rangle_r\cdot|\Psi(\tfrac{1}{2})\rangle_s:), \tag{1.19}$$

while the interaction term $S_{\mathrm{INT}}(\Psi)$ is given by

$$S_{\mathrm{INT}}(\Psi)\equiv\tfrac{1}{3}G\cdot\langle\nu_{\mathcal{S}}(r,s,t)|Y[\overline{r}]\cdot Y[r](:|\Psi(1)\rangle_r|\Psi(1)\rangle_s|\Psi(1)\rangle_t:)$$
$$+G\cdot\langle\nu_{\mathcal{S}}(r,s,t)|Y[\overline{r}](:|\Psi(1)\rangle_r|\Psi(\tfrac{1}{2})\rangle_s|\Psi(\tfrac{1}{2})\rangle_t:), \tag{1.20}$$

G being some (dimensionless) coupling constant. The (*cycle-symmetric*) *elementary* N-vertex functino $\langle\nu_{\mathcal{S}}(1,2,...,N)|$ used in Eq. (1.19) [(1.20)] is defined later by Eq. (2.14) as the function of the rth external operators φ_r's (for $r=1-N$) (where φ_r's represent $X^{\pm j}$, ϕ^j, σ, ϕ, χ for $j=0-4$). The expectation value over *all* external operators is taken in Eqs. (1.19) and (1.20), so that the gauge-fixed action $S_{\mathrm{GF}}(\Psi)$ in (1.18) is the function of the *quantized* bose fields ϕ_B's and the *quantized* fermi fields ψ_F's. We notice that the kinetic term $S_{\mathrm{KIN}}(\Psi)$ in (1.19) is the linear sum over *all* normal-ordered products: $\phi_B\phi_{B'}$:'s and $:\psi_F\psi_{F'}:$'s, while the interaction term $S_{\mathrm{INT}}(\Psi)$ in (1.20) is the linear sum over *all* normal-ordered products $:\phi_B\phi_{B'}\phi_{B''}:$'s and $:\phi_B\psi_F\psi_{F'}:$'s, coefficients of these products being just c numbers. [Normal ordering : : will be defined later by the operator product expansions (1.25)–(1.27) among fieldinos.]

We consider the scattering amplitude among N_B *physical* bosonic particles and N_F *physical* fermionic particles, each of which are excited modes of open NS (R) superstring. Each of these *physical* bosonic [fermionic] particles are specified by the quantum number b_r for $r=1-N_B$ [f_r for $r=(N_B+1)-(N_B+N_F)$]. Then we construct (for $r=1$–N) the operators $B_{b_r}\{\phi\}$ and $F_{f_r}\{\psi\}$ by the following formulas:

$$B_{b_r}\{\phi\}=\langle\nu_{\mathcal{S}}(r',r)|\mathcal{Q}_{r'}^{\mathrm{NS}}\cdot Y[\overline{r'}]\cdot Y[r']|\Psi(1)\rangle_{r'}\cdot|b_r(1)\rangle_r\quad\text{for } r=1-N_B \tag{1.21a}$$

and

$$F_{f_r}\{\psi\}=\langle\nu_{\mathcal{S}}(r',r)|\mathcal{Q}_{r'}^{\mathrm{R}}\cdot Y[\overline{r'}]|\Psi(1/2)\rangle_{r'}\cdot|f_r(1/2)\rangle_r\quad\text{for } r=(N_B+1)-(N_B+N_F). \tag{1.21b}$$

Expectation values over external operators are taken in Eq. (1.21a) [(1.21b)], so that $B_b\{\phi\}$ ($F_f\{\psi\}$) is the *linear* sum of *quantized* bose fields ϕ_B's (fermi fields ψ_F's), and it will be called the *physical* operator of the b (f) particle. [Quantization of ϕ_B's and ψ_F's will be imposed later by the ''operator product expansion among fieldinos,'' contractions among NS and R fieldinos being given by the formulas (1.26) and (1.27), respectively. Furthermore, we shall modify $S_{\mathrm{INT}}(\Psi)$ in (1.20) into $S_{\mathrm{INT}}^{\mathcal{M}}(\Psi)$ in (3.55).] We propose that the scattering amplitude among these *physical* NS and R particles (the number of which being respectively equal to N_B and N_F) is given by the following formula:

$$(N_B,N_F)\equiv\langle\phi=\psi=0|(\vec{T}_{X^0}\cdot\exp(S_{\mathrm{INT}}^{\mathcal{M}}(\Psi)))_{\text{connected}}$$
$$\times\left(\prod_{r=1}^{N_B}B_{b_r}\{\phi\}\right)\cdot\left(\prod_{r=N_B+1}^{N_B+N_F}F_{f_r}\{\Psi\}\right)|\phi=\psi=0\rangle, \tag{1.22}$$

where the expectation value over *all quantized* bose fields ϕ_B's and fermi fields ψ_F's should be taken. [The suffix ''*connected*'' in Eq. (1.22) shows the *connected* part of $\vec{T}_{X^0}\cdot\exp(S_{\mathrm{INT}}^{\mathcal{M}}(\Psi))$. See

Eq. (3.56). We remark that the formula (1.22) in the open QSFT corresponds to the LSZ reduction formula in the quantum field theory (QFT).[16]] In Eq. (1.22), the bra state $\langle \phi = \psi = 0|$ and the ket state $|\phi = \psi = 0\rangle$ satisfy respectively that

$$\langle \phi = \psi = 0|\phi_B(-) = 0 = \langle \phi = \psi = 0|\phi_F(-) \quad \text{for any negative frequency parts} \quad \phi_B(-) \quad \text{and} \quad \psi_F(-) \tag{1.23a}$$

and

$$\phi_B(+)|\phi = \psi = 0\rangle = 0 = \phi_F(+)|\phi = \psi = 0\rangle \quad \text{for any positive frequency parts} \quad \phi_B(+) \quad \text{and} \quad \psi_F(+), \tag{1.23b}$$

together with the following normalization condition:

$$\langle \phi = \psi = 0|\phi = \psi = 0\rangle = 1. \tag{1.24}$$

Furthermore, the time-ordered products among *quantized* bose fields ϕ_B's (*quantized* fermi fields ψ_F's) in Eq. (1.22) are given in terms of the ''time-ordered product among fieldinos'' as follows;

$$T_{X^0}(|\Psi(p(\gamma))\rangle_\gamma \cdot |\Psi(p(\delta))\rangle_\delta) \equiv :|\Psi(p(\gamma))\rangle_\gamma \cdot |\Psi(p(\delta))\rangle_\delta: + \overbrace{|\Psi(p(\gamma))\rangle_\gamma \cdot |\Psi(p(\delta))\rangle_\delta}, \tag{1.25a}$$

to be referred to as the ''operator product expansion'' (OPE) among fieldinos. In OPE (1.25a), we have used the picture number $p(r)$ [for $r = \gamma,\ \delta$] defined by

$$p(r) = \begin{cases} 1 & \text{for } \pi(r) = \text{NS}, \\ \frac{1}{2} & \text{for } \pi(r) = \text{R}. \end{cases} \tag{1.25b}$$

The first term on the right-hand side of the OPE (1.25) represents the normal-ordered product among fieldinos, which is to be obtained by moving any negative frequency parts $\phi_B(-)$ and $\psi_F(-)$ to the left of any positive frequency part $\phi_B(+)$ and $\psi_F(+)$. On the other hand, by taking account of Eq. (3.35) in Ref. 11, the *contraction* among the open NS fieldinos is given by

$$\overbrace{|\Psi(1)\rangle_\gamma \cdot |\Psi(1)\rangle_\delta} = \mathcal{P}_\gamma^{\text{NS}} \cdot \mathcal{P}_\delta^{\text{NS}} \cdot \left(\sum_{\pm} \frac{1}{2} \cdot X_{\pm,\gamma}^{1/2} \cdot X_{\mp,\gamma}^{1/2} \right) \frac{\mathcal{B}_{0,\gamma}}{L_{0,\gamma}} |\nu_{\mathcal{S}}(\gamma,\delta)\rangle, \qquad \text{for } \pi(\gamma) = \pi(\delta) = \text{NS}, \tag{1.26}$$

while the contraction among the open R fieldinos is given by

$$\overbrace{|\Psi(1/2)\rangle_\gamma \cdot |\Psi(1/2)\rangle_\delta} = \mathcal{P}_\gamma^{\text{R}} \cdot \mathcal{P}_\delta^{\text{R}} \cdot \left(\frac{X_\gamma^0 \cdot \mathcal{B}_{0,\gamma}}{L_{0,\gamma}} \right) |\nu_{\mathcal{S}}(\gamma,\delta)\rangle, \quad \text{for } \pi(\gamma) = \pi(\delta) = \text{R}, \tag{1.27}$$

which just corresponds to Eq. (3.17) in Ref. 11. In Eqs. (1.26) and (1.27) we have used $\mathcal{P}_r^{\pi(r)}$ in (1.4).

Comment: Our propagator (1.26) of the open NS superstring involves the symmetrization among $X_{+,\gamma}^{1/2} \cdot X_{-,\gamma}^{1/2}$ and $X_{-,\gamma}^{1/2} \cdot X_{+,\gamma}^{1/2}$, so that it can be *rewritten* [with the help of the ''commutability'' (C39)] into the expression (3.1). The explicit formula of the *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ [which is antisymmetric under $\gamma \leftrightarrow \delta$ and to be used in the contractions (1.26) and

(1.27)] will be given later by the formulas (2.16)–(2.19). It is to be noticed that the contraction among fieldinos (1.26) [(1.27)] represents the contraction among *quantized* bose (fermi) fields, i.e., $\phi_B\cdot\phi_{B'}$ $[\psi_F\cdot\psi_{F'}]$.

In this paper, we will show in detail how we can calculate the physical scattering amplitude (1.22), which will be found to satisfy the definite conservation laws of the following quantum numbers G, F, and H (in each **g**-loop). The operator $\mathcal{G}_r$ given by

$$\mathcal{G}_r\equiv:\exp(g_{\mathcal{G}}\cdot\sigma_r(y_r))\cdot\exp(f_{\mathcal{G}}\cdot\phi_r(y_r))\cdot\exp(h_{\mathcal{G}}\cdot\chi_r(y_r)) \tag{1.28}$$

is assigned with the ghosting number G, the fracting number F, and the hilberting number H, which are defined respectively by

$$(G,F,H)\equiv(g_{\mathcal{G}},f_{\mathcal{G}},h_{\mathcal{G}}). \tag{1.29}$$

These quantum numbers $K(\mathcal{G}_r)$ $[K(\mathcal{G}_r')]$ for $\mathcal{G}_r$ $[\mathcal{G}_r']$ are assumed to be additive, so that we have for $K=G,F,H$ that

$$K(\mathcal{G}_r\cdot\mathcal{G}_r')\equiv K(\mathcal{G}_r)+K(\mathcal{G}_r'). \tag{1.30}$$

Later in this paper, we shall find it convenient to define the picturing number P by

$$P\equiv G+F, \tag{1.31}$$

since any component $\mathcal{Q}_r^{(g)}$ [see Eq. (C2)] of the BRST chargino has the equal picturing number $P=1$. Incidentally, we will find later [in Eq. (2.12)] that $\mathcal{G}_r$ in (1.28) is the primary operator of the conformal weight given by

$$\tfrac{1}{2}\cdot g_{\mathcal{G}}(g_{\mathcal{G}}-3)-\tfrac{1}{2}\cdot f_{\mathcal{G}}(f_{\mathcal{G}}+2)+\tfrac{1}{2}\cdot h_{\mathcal{G}}(h_{\mathcal{G}}-1). \tag{1.32}$$

In Sec. II, we define the ''inlayed coordinate system s(*phere*)'' (ICS s) by using Gidding–Martinec's fundamental equation (GM-$\mathcal{FE}$). We introduce inlaying operators $W^{\varphi}[\cdots]$'s (for $\varphi=\vec{X},\ \phi^j,\ \sigma,\ \phi,\ \chi$). For each φ-mode, we construct the inlaying vertex function ${}_s\langle IV^{\varphi}(1,...,N)|$ in the ''ICS s,'' which inlay the rth *external* primary operators (at the rth disk coordinate w_r) into the *inlint* operators [at the rth inlayed coordinate $z_{rs}(w_r)$]. Then we propose formulas, which give the *small* gluing functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ [*large* gluing function $|V_L(\gamma,\delta)\rangle$]. Then we can prove the ''gluing theorem,'' which exhibits the $\gamma\delta$-gluing effect caused by the *small* vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$. In Sec. III, we first point out that our original $S_{\mathrm{INT}}(\Psi)$ in (1.20) does *not* lead to the BRST-invariant theory. Therefore, we are forced to introduce the counter term $\Delta S_{\mathrm{INT}}(\Psi)$, in such way that our *modified* action $S_{\mathrm{INT}}^{\mathcal{M}}(\Psi)$ given by

$$S_{\mathrm{INT}}(\Psi)\rightarrow S_{\mathrm{INT}}^{\mathcal{M}}(\Psi)\equiv S_{\mathrm{INT}}(\Psi)-\Delta S_{\mathrm{INT}}(\Psi) \tag{1.33}$$

leads to the (BRST-*invariant*) ''*amputated N-scatts*'' [denoted by ${}^{\mathcal{M}}\langle S^{\mathcal{C}}[\dots]|$], which satisfy the ''general conservation of the *total* BRST chargino'' (GCTC). The ''amputated N-scatts'' as well as $\Delta S_{\mathrm{INT}}(\Psi)$ can be calculated successively. (It should be noticed that there does not exist any colliding Wendt's singularity in our amplitudes, since they involve only *nonlocal* picture-changing operators.) In Sec. IV, we calculate the physical scattering amplitudes in **g**-loops, applying the method developed in our previous paper. With the help of Samuel's fundamental equation (S-$\mathcal{FE}$), we introduce the ''inlayed coordinate system (*genus*) **g**'' (ICS **g**). Then we propose the explicit formula for the *analytic inlint gluing operator* $\langle\Omega_{\mathbf{g}}(g_B,g_F)\rangle$ (in QSFT), with the help of which we can derive various trace-formulas in **g**-loops. In Sec. V, we summarize and conclude our results. In Appendix A, we define various (external and inlint) operators φ's (i.e., $\varphi=X^{\pm j},\ \phi^j,\ \sigma,\ \phi,\ \chi$) used in this paper. We also define the normal orderings and contractions in the NS-, R-, Σ-, and bo-representations. In Appendix B, we define the stress operators (T's) (BRST charginos $\mathcal{Q}$'s) in the NS-, R-, and bo-representations. These operators in *three* representations are shown to be just equal to each other in the case when $D=10$, where D is the space–time dimension. [See Eq. (1.17) for $\mathcal{Q}$'s.] We also derive inlaying identities for these operators T's ($\mathcal{Q}$'s). Then, we prove that any inlaying N-vertex function in the ''ICS s'' (denoted by $_s\langle IV(1,...,N)|$) conserves the total BRST chargino, and this result will be referred to as the ''special conservation of the total chargino'' (SCTC). In Appendix C, we give various operators of *local* (inverse) picture-changing operators X's (Y's). We also derive *nonlocal* picture-changing operators [which are to be used in Eqs. (1.7a) and (1.7b)]. In Appendix D, we summarize various results of *physical vertex operators* in the NS and R sectors, which are used in describing external physical NS and R particles, respectively. In Appendix E, we construct the *small* gluing vertex functino (*large* gluing vertex function) and we derive gluing identities and gluing relations satisfied by this gluing vertex functino (function). It is to be noticed that the gluing theorem can be obtained by using the *small* gluing vertex functino. Finally we remark how *elementary* N-vertex functinos $\langle \nu_{\mathcal{S}}(1,...,N)|$ [to be used in $S_{\mathrm{KIN}}(\Psi)$ in (1.19), $S_{\mathrm{INT}}(\Psi)$ in (1.20), etc.] can be constructed by using the ''inlayed coordinate system m(*idpoint*)'' (ICS m), which is defined by using the Gross–Jevicki's fundamental equation (GJ-$\mathcal{FE}$). Then, applying the gluing theorem to these elementary vertex functinos, we derive various useful formulas among them.

II. INLAYING VERTEX FUNCTIONS, INLAYING IDENTITIES, AND GLUING THEOREM

In our previous paper[5] on quantum open bosonic string field theory (open QBFT) in the ''$\mathcal{B}_0=0$ gauge,'' we have introduced various techniques which can be generalized into those which make us possible to calculate the physical scattering amplitudes (1.22) in open QSFT. In this section, we shall briefly explain how *elementary* N-vertex functino $\langle \nu_{\mathcal{S}}(1,2,...,N)|$ [which is *cycle-symmetric* under $1\rightarrow 2\rightarrow\cdots\rightarrow N\rightarrow 1$ and to be used in defining $S_{\mathrm{KIN}}(\Psi)$ in (1.19), $S_{\mathrm{INT}}(\Psi)$ in (1.20), $B_b\{\phi\}$ in (1.21a), and $F_f\{\psi\}$ in (1.21b)] can be explicitly given in terms of the rth external operators φ's (for $r=1-N$).

In open QSFT, the *tree* physical scattering amplitudes can be calculated by using the Riemann *surface* $\mathcal{R}$ with *one* boundary (which is located on the real axis in the complex z plane). All punctures Z_{rs}'s (for $r=1-N$) are located on the real axis, and all interaction points $\mathcal{Y}^s_{+\iota}$'s [for $\iota=1-(N-2)$] among open strings exist in the upper-half complex z plane [to be denoted by $\mathbb{H}^+$]. Relabeling N punctures (except for the relabeling which is reduced to some cyclic-permutation) produces various different amplitudes. We construct the Schottky double $\mathcal{D}$ of $\mathcal{R}$, by taking *two copies of* $\mathcal{R}$ and gluing together identical parts of the boundary. Thus we can enlarge the Riemann surface $\mathcal{R}$ even into the lower-half complex z plane (to be denoted by $\mathbb{H}^-$).

In the following, we briefly explain how the Schottky double $\mathcal{D}$ can be described by using the following ''Gidding–Martinec's fundamental equation'' (GM-$\mathcal{FE}$), which holds within the *whole* complex z plane $\mathbb{C}(=\mathbb{H}^+ + \mathbb{H}^-)$:

$$\frac{dz_s}{d\rho_s}=\nu_s(z_s)=R_s\cdot\frac{\Pi_{r=1}^{N}(z_s-Z_{rs})}{(\Pi_{\iota=1}^{N-2}(z_s-\mathcal{Y}^s_{+\iota})\cdot(z_s-\mathcal{Y}^s_{-\iota}))^{1/2}}. \tag{2.1}$$

We notice that GM-$\mathcal{FE}$ (2.1) gives the strip coordinate $\rho_s=\tau+\sqrt{-1}\sigma$ in terms of the inlayed coordinate z_s [say, by $\rho_s=\rho_s(z_s)$]. With the help of $\rho_s(z_s)$ (for any $z_s\in\mathbb{C}$), the Schottky double $\mathcal{D}$ can be drawn with the strip coordinate ρ_s. Inversely solving $\rho_s=\rho_s(z_s)$, we find $z_s=z_s(\rho_s)$ which means the strip coordinate ρ_s is inlayed into the inlayed coordinate $z_s(\in\mathbb{C})$. In order for GM-$\mathcal{FE}$ (2.1) to be able to describe the Schottky double $\mathcal{D}$, the parameters in GM-$\mathcal{FE}$ (2.1) should be chosen so as to satisfy following conditions (1) and (2): (1) The change of the variable σ should be $\pm 2\pi$ when z goes along *any* small circle around the rth puncture Z_{rs} (for $r=1-N$). (2) The imaginary part of $\rho_s(\mathcal{Y}^s_{\pm\iota})$ at *any* interacting point $\mathcal{Y}^s_{\pm\iota}$ should be equal to $\pm\pi/2$ for any $\iota=1-(N-2)$. With the help of ''equal-τ curves'' passing through any interacting point $\mathcal{Y}^s_{\pm\iota}$, we can divide $\mathbb{C}$ into N punctured ring domains *plus* $(N-3)$ unpunctured ring domains (which do not include any Z_{rs}). When condition (1) is satisfied, the rth punctured ring domain (which includes the rth puncture Z_{rs}) is located within the unit disk $|w_r|<1$, where the rth disk coordinate[5] w_r is defined by the following Eq. (2.2):

$$w_r \equiv \sqrt{-1}\cdot \exp(\mp(\rho_s(z_s)-\rho_s(\mathcal{Y}^s_{+\iota}))) \quad \text{for } \operatorname{Re}\rho_s(Z_{rs})=\pm\infty. \tag{2.2}$$

On the boundary $|w_r|=1$ of the rth punctured ring domain, two interacting points (say, $\mathcal{Y}^s_{\pm\iota}$) are located at the rth disk coordinate $w_r=\pm\sqrt{-1}$. On the other hand, the Ith *un*punctured ring domain [for $I=1-(N-3)$] describes the open string, which is freely propagating between two upper interacting points (say, from $\mathcal{Y}^s_{+\iota}$ and to $\mathcal{Y}^s_{+\kappa}$). Then, $T_I[\equiv|\rho_s(\mathcal{Y}^s_{+\iota})-\rho_s(\mathcal{Y}^s_{+\kappa})|]$ will be referred to as the the Ith propagating strip-time (of the Ith unpunctured ring domain). We have seen[5] that T_I's [for $I=1-(N-3)$] can be used as the modular parameters, which specify the Riemann surface $\mathcal{R}$ (having one boundary) with N punctures. The inlayed coordinate system s(*phere*) [''ICS s''] is defined to be the system with the inlayed coordinate $z_s(\rho_s)$, which is obtained by using GM-$\mathcal{FE}$ (2.1). In particular, the ''ICS m(*idpoint*)'' [5] is the special case of the ''ICS s(*phere*)'' where there does *not* exist any *un*punctured ring domain.

Comment: Gross–Jevicki[17] have investigated the ''inlayed coordinate system m(*idpoint*)'' (ICS m) which is defined by Eq. (E36). On the other hand, Gidding–Martinec and Samuel–Bluhm have used[18] Eq. (2.1) in calculating the tree amplitudes. Therefore, Eq. (2.1) [(E36)] will be referred to as ''Gidding–Martinec's fundamental equation'' (GM-$\mathcal{FE}$) [''Gross-Jevicki's fundamental equation'' (GJ-$\mathcal{FE}$)].

In open QSFT, we use various modes' operators $X^{\pm j}$, ϕ^j, σ, ϕ, and χ, (introduced by Friedan–Martinec–Schenker[9]), i.e., the string coordinate $X^{\pm j}$, the jth spinning operator ϕ^j (for $j=0-4$), the ghosting operator σ, the fracting operator ϕ, and the hilberting operator χ, all of which are defined by Eqs. (A3)–(A5b). The argument of the rth *external* string's operators is taken to be the rth disk coordinate w_r, so that the rth *external* operators live only within the *unit* disk $|w_r|<1$. On the other hand, the argument of the *inlint* string's operators within the rth punctured ring domain is the rth inlayed coordinate $z_{rs}(w_r)$, which is determined as the function of w_r by solving $w_r\to\rho_s\to z_s$. [See Eq. (1.10) in Ref. 5.] For each rth punctured ring domain, we introduce the *following* rth inlaying operator $W^{\varphi}_r[z_{rs}(w_r)]$ (for each φ-mode), which is defined by the following formula as the functional of the inlint operator $\varphi(z_{rs}(w_r))$ as well as $\varphi_r(w_r;+)$ [which is the positive-frequency part of the rth external operator $\varphi_r(w_r)$], w_r being integrated along a *small* closed path enclosing 0 in the anti-clockwise direction]: As for $\vec{X}$-modes, we have[5]

$$\begin{aligned} W^{\vec{X}}_r[z_{rs}(w_r)] \equiv \exp\Bigg(& \sum_{j=0}^{4} \oint_0 \frac{dw_r}{2\pi\sqrt{-1}} \oint_0 \frac{dw'_r}{2\pi\sqrt{-1}} \partial_{w_r} X^{+j}_r(w_r;+)\partial_{w'_r} X^{-j}_r(w'_r;+) \\ & \cdot \log\frac{z_{rs}(w_r)-z_{rs}(w'_r)}{w_r-w'_r}\Bigg) \\ & \times :\exp\Bigg(\sum_{\pm}\sum_{j=0}^{4} \oint_0 \frac{dw_r}{2\pi\sqrt{-1}} X^{\pm j}(z_{rs}(w_r))\cdot \partial_{w_r} X^{\mp j}_r(w_r;+)\Bigg):, \end{aligned} \tag{2.3}$$

while we have for $\varphi(=\phi^j,\sigma,\phi,\chi)$-mode that

$$\begin{aligned} & W^{\varphi}_r[z_{rs}(w_r)] \\ & \equiv \exp\left(\frac{\varepsilon^{\phi}_{\varphi}}{2}\oint_0 \frac{dw_r}{2\pi\sqrt{-1}}\oint_0 \frac{dw'_r}{2\pi\sqrt{-1}}\partial_{w_r}\varphi_r(w_r;+)\partial_{w'_r}\varphi_r(w'_r;+)\log\frac{z_{rs}(w_r)-z_{rs}(w'_r)}{w_r-w'_r}\right) \\ & \times :\exp\left(\varepsilon^{\phi}_{\varphi}\cdot \oint_0 \frac{dw_r}{2\pi\sqrt{-1}}\left(\varphi(z_{rs}(w_r))+\varepsilon^{\phi}_{\varphi}\cdot\frac{Q(\varphi)}{2}\log(z^{(1)}_{rs}(w_r))\right)\cdot\partial_{w_r}\varphi_r(w_r;+)\right):. \end{aligned} \tag{2.4}$$

{The inlaying operator $W^{\sigma}[\cdots]$ given by Eq. (2.4) has been denoted by $W^{\varphi}[\cdots]$ in Ref. 5.} In the formulas (2.3) and (2.4), we have used $\varepsilon_{\varphi}^{\phi}$ in (A5b) and the following background charges $Q(\varphi)$'s:

$$Q(X^{\pm j})=Q(\phi^j)=0,\quad Q(\sigma)=-3,\quad Q(\phi)=+2,\quad Q(\chi)=-1. \tag{2.5}$$

With the help of the formulas (2.3)–(2.5), we define that

$$W_r[z_{rs}(w_r)]\equiv\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi,\chi}W_r^{\varphi}[z_{rs}(w_r)]\quad\text{for } r=1-N. \tag{2.6}$$

For each $\varphi(=\vec{X},\phi^j,\sigma,\phi,\chi)$-mode, the ''standard ket state'' $|p_r(\varphi)=0\rangle_r$ and the ''standard bra state'' ${}_r\langle q_r(\varphi)=0|$ are defined respectively as states satisfying conditions (A6) and (A7). On the other hand, the ''*dual* standard bra state'' ${}_s\langle p_s(\varphi)=0|$ is defined by Eq. (A9). These states are normalized so as to satisfy normalization conditions (A8) and (A10). Then, with the help of the techniques used in Ref. 5, we introduce in this paper the following inlaying *N*-vertex function ${}_s\langle IV^{\varphi}(1,2,...,N)|$ in the ''ICS *s*'' [for $\varphi(=\vec{X},\phi^j,\sigma,\phi,\chi)$-mode]:

$${}_s\langle IV^{\varphi}(1,2,...,N)|\equiv\vec{R}\cdot\left(\prod_{r=1}^{N}{}_r\langle q_r(\varphi)=0|W_r^{\varphi}[z_{rs}(w_r)]\right)\quad\text{for }\varphi=\vec{X},\phi^j,\sigma,\phi,\chi, \tag{2.7}$$

$\vec{R}$ being the *radial ordering* defined by Eq. (A12). With the help of the formulas (2.6) and (2.7), we shall introduce the following (Grassman even) inlaying *N*-vertex function in the ''ICS *s*:''

$$\begin{aligned}{}_s\langle IV(1,2,...,N)|&\equiv\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi,\chi}{}_s\langle IV^{\varphi}(1,2,...,N)|\\&\equiv\langle q_{\text{ext}}=0|\vec{R}\cdot\left(\prod_{r=1}^{N}W_r[z_{rs}(w_r)]\right),\end{aligned} \tag{2.8}$$

where $\langle q_{\text{ext}}=0|$ is the ''standard bra state'' given by

$$\langle q_{\text{ext}}=0|\equiv\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi,\chi}\left(\prod_{r=1}^{N}{}_r\langle q_r(\varphi)=0|\right). \tag{2.9}$$

In much the same way as in our previous paper,[5] we can prove the following *inlaying identity* in the ''ICS *s*:'' For any *external* primary operator $\mathcal{G}_r(w')$ [of the conformal weight $d(\mathcal{G})$], we have that

$${}_s\langle IV^{\varphi}(1,2,...,N)|\mathcal{G}_r(w')={}_s\langle IV^{\varphi}(1,2,...,N)|(z_{rs}^{(1)}(w'))^{d(\mathcal{G})}\cdot\mathcal{G}(z_{rs}(w')), \tag{2.10}$$

which shows that the *r*th external primary operator $\mathcal{G}_r(w')$ at the arbitrarily given w' (for $|w'|<1$) is *inlayed* into the *inlint* primary operator $\mathcal{G}(z_{rs}(w'))$ at the *r*th inlayed coordinate $z_{rs}(w')$, which has already been *radial ordered* with respect to any inlint operator in the inlaying vertex function ${}_s\langle IV^{\varphi}(\ldots)|$.

Comment: Although the inlint operator $\mathcal{G}(z_{rs}(w'))$ is located on the right of $(\vec{R}\cdot\Pi W_r^{\varphi}[\cdots])$ [in Eq. (2.10)], $\mathcal{G}(z_{rs}(w'))$ is *actually* located at the *radial-ordered* place *amid* $(\vec{R}\cdot\Pi W_r^{\varphi}[\cdots])$. This is because the *external* $\mathcal{G}_r(w')$ is exchanged with (i.e., *inlayed* into) the *inlint* $\mathcal{G}(z_{rs}(w'))$, $\mathcal{G}$ being created *out* (at the *radial-ordered* place) from $(\vec{R}\cdot\Pi W_r^{\varphi}[\cdots])$. [See the proof of Eqs. (2.23)–(2.25) in Ref. 5 for more details.]

Hereafter, the inlaying identity (2.10) will simply be expressed by

$$\mathcal{G}_r(w') \overset{\mathcal{I}}{\Rightarrow} (z_{rs}^{(1)}(w'))^{d(\mathcal{G})} \cdot \mathcal{G}(z_{rs}(w')) \quad \text{(to be referred to as the ``inlaying identity'').} \tag{2.11}$$

As we have proved ''Eq. (2.25) in Ref. 5,'' we can similarly prove that

$$:\exp(p\cdot\varphi_r(w')): \overset{\mathcal{I}}{\Rightarrow} (z_{rs}^{(1)}(w'))^{\varepsilon_\varphi^\phi\cdot(1/2)p(p+Q(\varphi))} :\exp(p\cdot\varphi(z_{rs}(w'))): \quad \text{for } \varphi=\phi^j,\sigma,\phi,\chi, \tag{2.12}$$

where we have used the constant $\varepsilon_\varphi^\phi\,[Q(\varphi)]$ defined by Eq. (A5b) [(2.5)]. Incidentally, $\varepsilon_\varphi^\phi \cdot \frac{1}{2}p(p+Q(\varphi))$ is the conformal weight of the primary operator: $\exp(p\cdot\varphi(*))$:. [See the conformal weight (1.32).]

The N-vertex funct*ino* ${}_s\langle V(1,2,...,N)|$ in the ''ICS s'' is constructed from the inlaying vertex function (2.8) by

$${}_s\langle V(1,2,...,N)| \equiv (\langle \mathcal{Q}=0|\xi_0)\cdot {}_s\langle IV(1,2,...,N)|(|\mathcal{Q}=0\rangle), \tag{2.13}$$

ξ_0 being the *zero-mode* of the *inlint* hilbertino (A27). Then the *elementary* N-vertex functino $\langle \nu_{\mathcal{S}}(1,2,...,N)|$ [used in Eqs. (1.19)–(1.21b)] is constructed by using the inlayed coordinates $z_{rm}(w_r)$'s in the ''ICS m'' [see Eq. (E34)] and it is given by

$$\langle \nu_{\mathcal{S}}(1,2,...,N)| \equiv {}_m\langle V(1,2,...,N)|, \tag{2.14}$$

which is *cycle-symmetric* under the cyclic permutation $1\to2\to\cdots\to N\to1$. [See cycle-symmetric coordinates $z_r(w_r)$'s in (E35).] In Eq. (2.13), $\langle\mathcal{Q}=0|$ is the ''*inlint* dual standard bra state'' defined by

$$\langle\mathcal{Q}=0| = (|\mathcal{Q}=0\rangle)^\dagger. \quad \text{[See Eq. (D2) for } r=0.] \tag{2.15}$$

At this stage, we define the *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$, which is to be used in defining the propagator (1.26) of the open NS superstring [the propagator (1.27) of the open R superstring]. The *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ (which is antisymmetric under $\gamma\leftrightarrow\delta$) is constructed by

$$|\nu_{\mathcal{S}}(\gamma,\delta)\rangle = \eta_{0,\gamma}|V_L(\gamma,\delta)\rangle, \tag{2.16}$$

where the *large* gluing vertex function $|V_L(\gamma,\delta)\rangle$ is given by

$$|V_L(\gamma,\delta)\rangle \equiv \prod_{\varphi=\vec{X},\phi^j,\sigma,\phi,\chi} |V^\varphi(\gamma,\delta)\rangle. \tag{2.17}$$

Furthermore, each $|V^\varphi(\gamma,\delta)\rangle$ (for $\varphi=\vec{X},\ \phi^j,\ \sigma,\ \phi,\ \chi$) in Eq. (2.17) is constructed by the following formulas: For $\vec{X}$-modes, we have that

$$\begin{aligned}|V^{\vec{X}}(\gamma,\delta)\rangle \equiv \exp\left(-\sum_{n=1}^\infty\sum_\pm\sum_{j=0}^4 \frac{(-)^n}{n} J_{-n,\gamma}(X^{\pm j})\cdot J_{-n,\delta}(X^{\mp j})\right)\\ \times\left(\sum_{\{p^{\pm j}\}}\exp\left(\sum_\pm\sum_{j=0}^4 \sqrt{-1}\cdot p^{\pm j}\cdot q_\gamma(X^{\mp j})\right)\right.\\ \left.\times\exp\left(-\sum_\pm\sum_{j=0}^4\sqrt{-1}\cdot p^{\pm j}\cdot q_\delta(X^{\mp j})\right)\right)|p_\gamma(\vec{X})=0\rangle_\gamma\cdot|p_\delta(\vec{X})=0\rangle_\delta,\end{aligned} \tag{2.18}$$

while we have for $\varphi[=\phi^j,\sigma,\phi(,\chi)]$ that

$$|V^{\varphi}(\gamma,\delta)\rangle \equiv \exp\left(-\sum_{n=1}^{\infty} \varepsilon_{\varphi}^{\phi}\cdot\frac{(-)^n}{n}\cdot J_{-n,\gamma}(\varphi)\cdot J_{-n,\delta}(\varphi)\right)$$
$$\times\left(\sum_{p\in Z(+1/2)} \exp(p\cdot q_{\gamma}(\varphi))\cdot\exp(-(p+Q(\varphi))q_{\delta}(\varphi))\right)|p_{\gamma}(\varphi)=0\rangle_{\gamma}\cdot|p_{\delta}(\varphi)=0\rangle_{\delta}$$
$$\text{for } \varphi=\phi^j,\sigma,\phi(,\chi), \tag{2.19}$$

where background charges $Q(\varphi)$'s are those given by Eq. (2.5). Incidentally, $|V^{\varphi}(\gamma,\delta)\rangle$ (for $\varphi=\vec{X},\ \phi^j,\ \phi$) is Grassman *even*, while $|V^{\varphi}(\gamma,\delta)\rangle$ (for $\varphi=\sigma,\chi$) is Grassman *odd*. [See Appendix E for more details.]

Using the techniques in Ref. 5, we can prove the following gluing theorem (which will be useful in calculating tree amplitudes[5]).

Gluing theorem: For any primary operator $\mathcal{G}_{\gamma}(w_{\gamma})$ of conformal weight $d(\mathcal{G})$, we have that

$$((\langle\mathcal{Q}=0|\xi_0)\cdot{}_l\langle IV(1,...,N,\gamma)|(|\mathcal{Q}=0\rangle))\cdot\mathcal{G}_{\gamma}(w_{\gamma})$$
$$\times((\langle\tilde{\mathcal{Q}}=0|\tilde{\xi}_0)\cdot{}_r\langle\tilde{I}V(\delta,N+1,...,N+M)|(|\tilde{\mathcal{Q}}=0\rangle))\cdot\left(\frac{1}{L_{0,\gamma}}\cdot|\nu_{\mathcal{S}}(\gamma,\delta)\rangle\right)$$
$$=\int_0^{\infty}dT_{\gamma}(\langle\mathcal{Q}=0|\xi_0)({}_g\langle IV(1,...,N,N+1,...,N+M)|(z_{\gamma g}^{(1)}(w_{\gamma}))^{d(\mathcal{G})}\cdot\mathcal{G}(z_{\gamma g}(w_{\gamma})))(|\mathcal{Q}=0\rangle), \tag{2.20}$$

where the *small* vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ is the one given by the formulas (2.16)–(2.19).

Proof: The gluing theorem (2.20) can be proved similarly to ''Eq. (5.10) in Ref. 5,'' provided that we make the following observations. First, the inlaying $(N+1)$-vertex function ${}_l\langle IV(1,...,N,\gamma)|$ in the ''ICS $l(eft)$'' inlays the γth external operator $\mathcal{G}_{\gamma}(w_{\gamma})$ into the inlint operator $(z_{\gamma l}^{(1)}(w_{\gamma}))^{d(\mathcal{G})}\cdot\mathcal{G}(z_{\gamma l}(w_{\gamma}))$, which is subsequently conformally mapped into $(z_{\gamma g}^{(1)}(w_{\gamma}))^{d(\mathcal{G})}\cdot\mathcal{G}(z_{\gamma g}(w_{\gamma}))$ by the conformal mapping operator $U_{-\gamma\delta}$ [given by Eq. (5.17) in Ref. 5]. We should notice that $z_{\gamma g}(w_{\gamma})$ is the inlayed coordinate within the unpunctured ring domain in the ''ICS $g(lued)$,'' while $z_{\gamma l}(w_{\gamma})$ is the γth inlayed coordinate within the γth punctured ring domain in the ''ICS $l(eft)$.'' The argument w_{γ} in $z_{\gamma g}(w_{\gamma})$ has been used to show that the strip-coordinate ''ρ'' for $z_{\gamma g}(w_{\gamma})$ is just equal to the strip-coordinate ''ρ'' for $z_{\gamma l}(w_{\gamma})$. [See ''Eqs. (5.25) and (5.26) in Ref. 5.''] (Q.E.D.)

Comment: On the left-hand side of Eq. (2.20), we have used two inlaying vertex functions; ${}_l\langle IV(1,...,N,\gamma)|$ in the ''ICS $s=l(eft)$'' is given in terms of *inlint* operators φ's, while ${}_r\langle\tilde{I}V(\delta,N+1,...,N+M)|$ in the ''ICS $s=r(ight)$'' is given in terms of inlint operators $\tilde{\varphi}$'s. Inlint operators φ's and $\tilde{\varphi}$'s are independent of each other. On the other hand, ${}_g\langle IV(1,...,N+M)|$ on the right-hand side of Eq. (2.20) is the inlaying vertex function in the ''ICS $s=g(lued)$.'' The ''ICS $s=g$'' is the ICS obtained by $\gamma\delta$-gluing the ''ICS $s=l$'' to the ''ICS $s=r$'' as follows. *All three* ''ICS $s=l,\ r,\ g$'' can be described by three GM-$\mathcal{FE}$'s (2.1) (with *different* modular parameters), provided that the numbers of punctured ring domains of three GM-$\mathcal{FE}$'s are respectively just equal to $N+1$, $M+1$, and $N+M$, respectively. Furthermore, the propagating strip-times of the $(N+M-4)$ unpunctured ring domains in the ''ICS $s=g$'' are *just equal to* those of the $(N-2)$ unpunctured ring domains in the ''ICS $s=l$'' and the $(M-2)$ unpunctured ring domains in the ''ICS $s=r$.'' Finally, ''T_{γ}'' (in the ''ICS g'') on the right-hand side of Eq. (2.20) is the *new* propagating strip-time of the ''$\gamma\delta$th unpunctured ring domain in the ICS $s=g$,'' which has been *created* by gluing the ''γth punctured ring domain in the ICS $s=l$'' to the ''δth punctured ring domain in the ICS $s=r$.'' This propagating strip-time T_{γ} on the right-hand side of Eq. (2.16) has been obtained through

$$\frac{1}{L_{0,\gamma}}=\int_0^\infty dT_\gamma\, x^{L_{0,\gamma}} \quad \text{and} \quad T_\gamma \equiv \log\left(\frac{1}{x}\right). \tag{2.21}$$

Incidentally, $z_{\gamma g}(w_\gamma)$ in Eq. (2.20) is the inlayed coordinate within the $\gamma\delta$th unpunctured ring domain. (See our previous paper[5] for more details.)

III. ‘‘AMPUTATED SCATTS’’ AND THE ‘‘GENERAL CONSERVATION OF THE TOTAL BRST CHARGINO’’

By using the ‘‘commutability’’ (C39) and formulas (C32)–(C36), the formula (1.26) is reduced to

$$\overbrace{|\Psi(1)\rangle_I\cdot|\Psi(1)\rangle_J}=\left(\sum_{\pm}\frac{1}{2}\cdot X^{1/2}_{\pm,I}\cdot X^{1/2}_{\mp,I}\right)\frac{\mathcal{B}_{0,I}}{L_{0,I}}\cdot P_I^{\mathrm{NS}}(\mathrm{GSO})|\nu_{\mathcal{S}}(I,J)\rangle \quad \text{in the case} \quad \pi(I)=\mathrm{NS}, \tag{3.1}$$

which is referred to as the propagator of the open NS superstring. On the other hand, by using the ‘‘commutability’’ (C38) and formulas (C22)–(C24), the formula (1.27) is also reduced to

$$\overbrace{|\Psi(\tfrac{1}{2})\rangle_I\cdot|\Psi(\tfrac{1}{2})\rangle_J}=X_I^0\cdot\frac{\mathcal{B}_{0,I}}{L_{0,I}}\cdot P_I^{\mathrm{R}}(\mathrm{GSO})|\nu_{\mathcal{S}}(I,J)\rangle \quad \text{in the case} \quad \pi(I)=\mathrm{R}, \tag{3.2}$$

which is referred to as the propagator of the open R superstring.

In general, the operator $\vec{T}_{X^0}\cdot\exp(S^{(\mathcal{M})}_{\mathrm{INT}}(\Psi))$ [with $S^{(\mathcal{M})}_{\mathrm{INT}}(\Psi)$ in (1.33)] can be calculated with the help of Wick's theorem,[16] by using the propagator (3.1) [(3.2)] of the open NS (R) superstring. The result obtained by applying Wick's theorem can be expressed by

$$\vec{T}_{X^0}\cdot\exp(s^{(\mathcal{M})}_{\mathrm{INT}}(\Psi))\equiv 1+\sum_{N=0}^{\infty}\sum_{\{p(r)\}}{}^{(\mathcal{M})}\langle S[p(1),p(2),\ldots,p(N)]|\cdot\frac{1}{N!}:\left(\prod_{r=1}^{N}|\Psi(p(r))\rangle_r\right): \tag{3.3a}$$

$$\equiv:\exp\left(\sum_{N=0}^{\infty}\sum_{\{p(r)\}}{}^{(\mathcal{M})}\langle S^{\mathcal{C}}[p(1),p(2),\ldots,p(N)]|\cdot\frac{1}{N!}:\left(\prod_{r=1}^{N}|\Psi(p(r))\rangle_r\right):\right):. \tag{3.3b}$$

Hereafter, the quantities with(out) the superscript ‘‘$\mathcal{M}$’’ will be referred to as the *(un)modified* ones. In the expansions (3.3a) and (3.3b), any ${}^{(\mathcal{M})}\langle S^{(\mathcal{C})}[p(1),p(2),\ldots,p(N)]|$ (which does *not* involve any operator ϕ_B and ψ_F) is obtained by *contractions* and it is assumed to be completely antisymmetric under any permutation of arguments, and we have used the following shorthand notation for the summation:

$$\sum_{\{p(r)\}}\equiv\prod_{r=1}^{N}\sum_{p(r)=1/2,1}. \tag{3.4}$$

Furthermore, ${}^{(\mathcal{M})}\langle S^{\mathcal{C}}[p(1),p(2),\ldots,p(N)]|$ is the *connected* part in ${}^{(\mathcal{M})}\langle S[p(1),p(2),\ldots,p(N)]|$. In Eqs. (3.3a) and (3.3b) and hereafter, the ordering in the product among external operators will be defined by

$$\prod_{r=1}^{N}F_r\equiv F_1\cdot F_2\cdots F_N, \tag{3.5}$$

for any Grassman even or odd operators F_r's. In Eqs. (3.3a) and (3.3b), any *quantized* bose field ϕ_B and *quantized* fermi field ψ_F is included in the *normal-ordered* product of fieldinos, where the normal ordering: : should be taken with respect to any quantized bose field ϕ_B and fermi field ψ_F.

We first notice that

$$\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\frac{1}{N!}\cdot\langle S[p(1),p(2),\ldots,p(N)]|\left(1+\varepsilon\left(\sum_{r=1}^{N}\mathcal{Q}_r\right)\right):\left(\prod_{r=1}^{N}|\Psi(p(r))\rangle_r\right):$$
$$=\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\frac{1}{N!}\cdot\langle S[p(1),p(2),\ldots,p(N)]|:\left(\prod_{r=1}^{N}(1+\varepsilon\cdot\mathcal{Q}_r)|\Psi(p(r))\rangle_r\right):+O(\varepsilon^2),\tag{3.6}$$

ε being an arbitrary Grassmann *odd* constant. [Actually, $\varepsilon^2\equiv 0$ in Eq. (3.6) and hereafter.] Hereafter, we investigate the interaction term $S_{\mathrm{INT}}(\mathring{\Psi})$ which is obtained from $S_{\mathrm{INT}}(\Psi)$ by the following replacements:

$$|\Psi(p(r))\rangle_r\to(1+\varepsilon\cdot\mathcal{Q}_r)|\Psi(p(r))\rangle_r(\equiv|\mathring{\Psi}(p(r))\rangle_r)\quad\text{for } r=1-N.\tag{3.7}$$

Since the elementary three-vertex function (2.14) satisfies that

$$\langle\nu_{\mathcal{S}}(r,s,t)|(\mathcal{Q}_r+\mathcal{Q}_s+\mathcal{Q}_t)=0\quad[\text{see Eq. (B34)}],\tag{3.8}$$

we find that

$$S_{\mathrm{INT}}(\mathring{\Psi})=S_{\mathrm{INT}}(\Psi).\tag{3.9}$$

Thus, we are lead to investigate

$$\vec{T}_{X^0}\cdot\exp(S_{\mathrm{INT}}(\mathring{\Psi}))(\equiv\vec{T}_{X^0}\cdot\exp(S_{\mathrm{INT}}(\Psi)))$$
$$=\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\frac{1}{N!}\cdot\langle\mathring{S}[p(1),p(2),\ldots,p(N)]|:\left(\prod_{r=1}^{N}|\mathring{\Psi}(p(r))\rangle_r\right):\tag{3.10a}$$
$$=\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\frac{1}{N!}\cdot\langle S[p(1),p(2),\ldots,p(N)]|:\left(\prod_{r=1}^{N}|\Psi(p(r))\rangle_r\right):,\tag{3.10b}$$

which is found to give that

$$\langle\mathring{S}[p(1),p(2),\ldots,p(N)]|-\langle S[p(1),p(2),\ldots,p(N)]|=\langle S[p(1),p(2),\ldots,p(N)]|\left(-\varepsilon\cdot\sum_{r=1}^{N}\mathcal{Q}_r\right).\tag{3.10c}$$

Furthermore, the difference $(\langle\mathring{S}[\cdots]|-\langle S[\cdots]|)$ on the left-hand side of Eq. (3.10c) can be found by calculating $\overbrace{S_{\mathrm{INT}}(\mathring{\Psi})\cdot S_{\mathrm{INT}}(\mathring{\Psi})}$ up to the "first order of ε." First, we find in the NS sector that

$$|\overbrace{\Psi(1)\rangle_I\cdot|\Psi(1)}\rangle_J=\left(\sum_{\pm}\frac{1}{2}\cdot X^{1/2}_{\pm,I}\cdot X^{1/2}_{\mp,I}\right)\frac{\mathcal{B}_{0,I}}{L_{0,I}}\cdot P^{\mathrm{NS}}_I(\mathrm{GSO})|\nu_{\mathcal{S}}(I,J)\rangle\to|\overbrace{\mathring{\Psi}(1)\rangle_I\cdot|\mathring{\Psi}(1)}\rangle_J$$
$$=|\overbrace{\Psi(1)\rangle_I\cdot|\Psi(1)}\rangle_J+\varepsilon\cdot\left(\sum_{\pm}\frac{1}{2}\cdot X^{1/2}_{\pm,J}\cdot X^{1/2}_{\mp,J}\right)\cdot P^{\mathrm{NS}}_J(\mathrm{GSO})|\nu_{\mathcal{S}}(I,J)\rangle,\tag{3.11}$$

while in the R sector we find that

$$\overbrace{|\Psi(\tfrac{1}{2})\rangle_I\cdot|\Psi(\tfrac{1}{2})\rangle_J}=X_I^0\cdot\frac{\mathcal{B}_{0,I}}{L_{0,I}}\cdot P_I^{\mathrm{R}}(\mathrm{GSO})|\nu_{\mathcal{S}}(I,J)\rangle\to\overbrace{|\mathring{\Psi}(\tfrac{1}{2})\rangle_I\cdot|\mathring{\Psi}(\tfrac{1}{2})\rangle_J}$$
$$=\overbrace{|\Psi(\tfrac{1}{2})\rangle_I\cdot|\Psi(\tfrac{1}{2})\rangle_J}+\varepsilon\cdot X_J^0\cdot P_J^{\mathrm{R}}(\mathrm{GSO})|\nu_{\mathcal{S}}(I,J)\rangle. \tag{3.12}$$

With the help of the formulas (3.11) and (3.12), we can calculate the following contractions:

$$:\overbrace{S_{\mathrm{INT}}(\mathring{\Psi})\cdot S_{\mathrm{INT}}(\mathring{\Psi})}:\equiv\sum_{\{p\}}\langle\mathring{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(b),p(C),p(D),p(E)]|$$
$$\times\frac{1}{4!}:|\mathring{\Psi}(p(B))\rangle_B|\mathring{\Psi}(p(C))\rangle_C|\mathring{\Psi}(p(D))\rangle_D|\mathring{\Psi}(p(E))\rangle_E: \tag{3.13}$$

and

$$:\overbrace{S_{\mathrm{INT}}(\Psi)\cdot S_{\mathrm{INT}}(\Psi)}:\equiv\sum_{\{p\}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|$$
$$\times\frac{1}{4!}:|\Psi(p(B))\rangle_B|\Psi((C))\rangle_C|\Psi(p(D))\rangle_D|\Psi(p(E))\rangle_E:. \tag{3.14}$$

Thus it is concluded that

$$\sum_{\{p\}}\frac{1}{4!}(\langle\mathring{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|-\langle S^{C}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|)$$
$$\times:|\Psi(p(B))\rangle_B|\Psi(p(C))\rangle_C|\Psi(p(D))\rangle_D|\Psi(p(E))\rangle_E:$$
$$=-\varepsilon\cdot G^2\langle\nu_{\mathcal{S}}(B,C,D,E)|(Y[\overline{I}]\cdot Y[I]:(|\Psi(1)\rangle_B|\Psi(1)\rangle_C|\Psi(1)\rangle_D|\Psi(1)\rangle_E):$$
$$+Y[\overline{I}]:(|\Psi(\tfrac{1}{2})\rangle_B|\Psi(\tfrac{1}{2})\rangle_C)X^{1/2}_{+BC}(|\Psi(\tfrac{1}{2})\rangle_D|\Psi(\tfrac{1}{2})\rangle_E):)$$
$$-\varepsilon\cdot G^2\langle\nu_{\mathcal{S}}(B,C,D,E)|(:(|\Psi(1)\rangle_B|\Psi(\tfrac{1}{2})\rangle_C)Y[\overline{I}](|\Psi(1)\rangle_D|\Psi(\tfrac{1}{2})\rangle_E):$$
$$+:(|\Psi(1)\rangle_B|\Psi(\tfrac{1}{2})\rangle_C)Y[\overline{I}](|\Psi(\tfrac{1}{2})\rangle_D|\Psi(1)\rangle_E):$$
$$+:(|\Psi(\tfrac{1}{2})\rangle_B|\Psi(1)\rangle_C)Y[\overline{I}](|\Psi(1)\rangle_D|\Psi(\tfrac{1}{2})\rangle_E):$$
$$+:(|\Psi(\tfrac{1}{2})\rangle_B|\Psi(1)\rangle_C)Y[\overline{I}](|\Psi(\tfrac{1}{2})\rangle_D|\Psi(1)\rangle_E):$$
$$+:(|\Psi(1)\rangle_B|\Psi(1)\rangle_C)Y[\overline{I}]Y[I]X^{1/2}_{+,DE}(|\Psi(\tfrac{1}{2})\rangle_D|\Psi(\tfrac{1}{2})\rangle_E):$$
$$+:(|\Psi(\tfrac{1}{2})\rangle_B|\Psi(\tfrac{1}{2})\rangle_C)Y[\overline{I}]Y[I]X^{1/2}_{+,DE}(|\Psi(1)\rangle_D|\Psi(1)\rangle_E):), \tag{3.15}$$

where we have used the *cycle-symmetric* elementary vertex functino $\langle\nu_{\mathcal{S}}(A,B,C,D)|$, which is obtained by applying the gluing theorem (E33b). Unfortunately, Eq. (3.15) does not vanish identically, and it is just equal to

$$-\varepsilon\cdot G^2\langle \nu_{\mathcal{S}}(B,C,D,E)|Y[\overline{I}](X^{1/2}_{+BC}-X[\overline{I}]):(|\Psi(\tfrac{1}{2})\rangle_B|\Psi(\tfrac{1}{2})\rangle_C|\Psi(\tfrac{1}{2})\rangle_D|\Psi(\tfrac{1}{2})\rangle_E):$$
$$-\varepsilon\cdot G^2\langle \nu_{\mathcal{S}}(B,C,D,E)|Y[\overline{I}]Y[I](X^{1/2}_{+,BC}-X[I]):(|\Psi(1)\rangle_B|\Psi(1)\rangle_C|\Psi(\tfrac{1}{2})\rangle_D|\Psi(\tfrac{1}{2})\rangle_E$$
$$+|\Psi(\tfrac{1}{2})\rangle_B|\Psi(\tfrac{1}{2})\rangle_C|\Psi(1)\rangle_D|\Psi(1)\rangle_E):. \tag{3.16}$$

With the help of

$$X^{1/2}_{+BC}-X[\overset{(-)}{I}]\equiv\{\mathcal{Q}_B+\mathcal{Q}_C,\overset{(-)}{\Delta}_{+BC}\},\ \text{[see Eqs. (C3) and (C25)]}, \tag{3.17}$$

the term (3.16) can be rewritten into

$$-\varepsilon\cdot G^2\langle \nu_{\mathcal{S}}(B,C,D,E)|Y[\overline{I}]\cdot\overline{\Delta}^{1/2}_{+BC}\cdot(\tilde{\mathcal{Q}}_B+\tilde{\mathcal{Q}}_C+\tilde{\mathcal{Q}}_D+\tilde{\mathcal{Q}}_E):(|\Psi(\tfrac{1}{2})\rangle_B|\Psi(\tfrac{1}{2})\rangle_C|\Psi(\tfrac{1}{2})\rangle_D|\Psi(\tfrac{1}{2})\rangle_E):$$
$$-\varepsilon\cdot G^2\langle \nu_{\mathcal{S}}(B,C,D,E)|Y[\overline{I}]Y[I]\cdot\Delta^{1/2}_{+BC}\cdot(\tilde{\mathcal{Q}}_B+\tilde{\mathcal{Q}}_C+\tilde{\mathcal{Q}}_D+\tilde{\mathcal{Q}}_E)$$
$$\times:(|\Psi(1)\rangle_B|\Psi(1)\rangle_C|\Psi(\tfrac{1}{2})\rangle_D|\Psi(\tfrac{1}{2})\rangle_E+|\Psi(\tfrac{1}{2})\rangle_B|\Psi(\tfrac{1}{2})\rangle_C|\Psi(1)\rangle_D|\Psi(1)\rangle_E): \tag{3.18a}$$

$$\equiv\sum_{\{p\}}\frac{1}{4!}\cdot\langle\Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|(-\varepsilon)(\tilde{\mathcal{Q}}_B+\tilde{\mathcal{Q}}_C+\tilde{\mathcal{Q}}_D+\tilde{\mathcal{Q}}_E)$$
$$\times:|\Psi(p(B))\rangle_B|\Psi(p(C))\rangle_C|\Psi(p(D))\rangle_D|\Psi(p(E))\rangle_E:. \tag{3.18b}$$

With the help of the counter term $\langle\Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|$ determined by Eq. (3.18b), the (BRST-invariant) ''amputated 4-scatt'' ${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|$ in 0-loop can be constructed by

$${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|\equiv\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(C),p(D),p(E)]|$$
$$-\langle\Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|. \tag{3.19}$$

The amputated 4-scatt in (3.19) has been constructed so as to satisfy the following ''general conservation of the total BRST chargino,''

$${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|(\tilde{\mathcal{Q}}_B+\tilde{\mathcal{Q}}_C+\tilde{\mathcal{Q}}_D+\tilde{\mathcal{Q}}_E)=0$$
$$\text{(to be referred to as the ''GCTC'')}. \tag{3.20}$$

It should be noticed that the *counter term* $\langle\Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\ldots,p(4)]|$ in Eq. (3.19) does not contain *any* propagator. Thus the interaction is partly modified from $S_{\mathrm{INT}}(\Psi)$ in (1.20) into $S^{\mathcal{M}}_{\mathrm{INT}}(4,0)$, which is given by

$$S^{\mathcal{M}}_{\mathrm{INT}}(4,0)\equiv S_{\mathrm{INT}}(\Psi)-\sum_{\{p\}}\frac{1}{4!}\cdot\langle\Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(B),p(C),p(D),p(E)]|$$
$$\times:|\Psi(p(B))\rangle_B|\Psi(p(C))\rangle_C|\Psi(p(D))\rangle_D|\Psi(p(E))\rangle_E:. \tag{3.21}$$

Similarly, we can calculate ${}^{\mathcal{M}}\langle\tilde{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\ldots,p(5)]|$ in the following formula, $S^{\mathcal{M}}_{\mathrm{INT}}(4,0)$ being given by Eq. (3.21).

$$\vec{T}_{X^0}\cdot \exp(S^{\mathcal{M}}_{\rm INT}(4,0)) \equiv 1+\sum_{N=3}^{4}\sum_{\{p(r)\}} {}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),p(2),\dots,p(N)]| \cdot \frac{1}{N!} : \left(\prod_{r=1}^{N} |\Psi(p(r))\rangle_r\right): + \sum_{\{p(r)\}} {}^{\mathcal{M}}\langle \tilde{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(1),p(2),\dots,p(5)]| \cdot \frac{1}{5!} : \left(\prod_{r=1}^{5} |\Psi(p(r))\rangle_r\right): + \text{other terms}. \tag{3.22}$$

[We notice that thus obtained ${}^{\mathcal{M}}\langle \tilde{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|$ is *partly* modified from $\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|$ by the effect due to the counter term in ''$S^{\mathcal{M}}_{\rm INT}(4.0)$.''] Even if we take account of the ''GCTC'' in (3.20), we shall still find that

$${}^{\mathcal{M}}\langle \tilde{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|\left(\sum_{r=1}^{5} \mathcal{Q}_r\right) \neq 0. \tag{3.23}$$

However, the term (3.23) can be expressed by

$$\langle \Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|\left(\sum_{r=1}^{5} \mathcal{Q}_r\right), \tag{3.24}$$

where the number of propagators in (the counter term) $\langle \Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|$ is *less* than that in ${}^{\mathcal{M}}\langle \tilde{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|$. {This situation will be imagined by comparing the counter term $\langle \Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(4)]|$ with $\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(4)]|$.} Then new modified action is (up to this approximation) constructed by

$$S^{\mathcal{M}}_{\rm INT}(4,0) \to S^{\mathcal{M}}_{\rm INT}(5,0) = S_{\rm INT}(\Psi) - \sum_{\{p\}}\sum_{n=4}^{5}\frac{1}{n!}\cdot \langle \Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),p(2),\dots,p(n)]| \times :|\Psi(p(1))\rangle_1 |\Psi(p(2))\rangle_2 \cdots |\Psi(p(n))\rangle_n :. \tag{3.25}$$

The (BRST-invariant) ''amputated 5-scatt'' ${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|$ in 0-loop is also constructed by

$${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)] \equiv {}^{\mathcal{M}}\langle \tilde{S}^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]| - \langle \Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|, \tag{3.26}$$

so as to satisfy the following ''GCTC.''

$${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\dots,p(5)]|\left(\sum_{r=1}^{5} \mathcal{Q}_r\right) = 0. \tag{3.27}$$

In the following, we shall show how to derive the *completely modified action* $S^{\mathcal{M}}_{\rm INT}(\Psi)$, which together with the formula (3.3b) gives the following (BRST-invariant) ''*amputated N-scatt*'' ${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}}[p(1),\dots,p(N)]|$ in **g**-loops;

$${}^{\mathcal{M}}\langle S^{\mathcal{C}}[p(1),\dots,p(N)]| = \sum_{\mathbf{g}=0}^{\infty} {}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}}[p(1),\dots,p(N)]|, \tag{3.28}$$

which satisfies the following ''GCTC:''

$${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}}[p(1),\dots,p(N)]|\left(\sum_{r=1}^{N} \mathcal{Q}_r\right) = 0. \tag{3.29}$$

Suggested by Eqs. (3.11), (3.12), and (3.16)–(3.21), we find that BRST-invariant amputated N-scatt ${}^{\mathcal{M}}\langle S^{\mathcal{C}}[p(1),\ldots,p(N)]|$ in (3.28) can be obtained as follows: Instead of $S_{\mathrm{INT}}(\Psi)$ in (1.20), we consider that

$$\begin{aligned}S_{\mathrm{INT}}^{\mathcal{T}}(\Psi+\Delta\widetilde{\Psi})\equiv\tfrac{1}{3}G\cdot\langle\nu_{\mathcal{S}}(r,s,t)|Y[\overline{r}]\cdot Y[r]{:}(|\Psi(1)\rangle_r+|\Delta\widetilde{\Psi}(1)\rangle_r)(|\Psi(1)\rangle_s\\ +|\Delta\widetilde{\Psi}(1)\rangle_s)(|\Psi(1)\rangle_t+|\Delta\widetilde{\Psi}(1)\rangle_t){:}+G\cdot\langle\nu_{\mathcal{S}}(r,s,t)|Y[\overline{r}]{:}(|\Psi(1)\rangle_r\\ +|\Delta\widetilde{\Psi}(1)\rangle_r)(|\Psi(\tfrac{1}{2})\rangle_s+|\Delta\widetilde{\Psi}(\tfrac{1}{2})\rangle_s)(|\Psi(\tfrac{1}{2})\rangle_t+|\Delta\widetilde{\Psi}(\tfrac{1}{2})\rangle_t){:}.\end{aligned}\tag{3.30}$$

The rth operators $|\Delta\widetilde{\Psi}(1)\rangle_r$ and $|\Delta\widetilde{\Psi}(\tfrac{1}{2})\rangle_r$ used in Eq. (3.30) will be referred to as the *counter* fieldinos, and they are defined respectively to have following properties. Correspondingly to Eqs. (1.25)–(1.27), we have

$$\begin{aligned}\vec{T}_{X^0}(|\Delta\widetilde{\Psi}(p(\gamma))\rangle_\gamma\cdot|\Delta\widetilde{\Psi}(p(\delta))\rangle_\delta)\\ \equiv{:}|\Delta\widetilde{\Psi}(p(\gamma))\rangle_\gamma\cdot|\Delta\widetilde{\Psi}(p(\delta))\rangle_\delta{:}+|\Delta\widetilde{\Psi}\overline{(p(\gamma))\rangle_\gamma\cdot|\Delta\widetilde{\Psi}(}p(\delta))\rangle_\delta,\end{aligned}\tag{3.31}$$

where contractions are given by

$$|\Delta\widetilde{\Psi}\overline{(1)\rangle_\gamma\cdot|\Delta\widetilde{\Psi}(1)}\rangle_\delta=-\Theta_\gamma^{\mathrm{NS}}\cdot P^{\mathrm{NS}}(\mathrm{GSO})|\nu_{\mathcal{S}}(\gamma,\delta)\rangle\tag{3.32}$$

and

$$|\Delta\widetilde{\Psi}\overline{(\tfrac{1}{2})\rangle_\gamma\cdot|\Delta\widetilde{\Psi}}(\tfrac{1}{2})\rangle_\delta=-\Theta_\gamma^{\mathrm{R}}\cdot P^{\mathrm{R}}(\mathrm{GSO})|\nu_{\mathcal{S}}(\gamma,\delta)\rangle.\tag{3.33}$$

The operators $\Theta_\gamma^{\mathrm{NS}}$ in Eq. (3.32) and $\Theta_\gamma^{\mathrm{R}}$ in Eq. (3.33) are those satisfying respectively that

$$\{\mathcal{Q}_r,\Theta_\gamma^{\pi(\gamma)}\}\equiv\mathcal{X}_\gamma^{\pi(\gamma)}\quad\text{for}\quad\pi(\gamma)=\mathrm{NS,R},\tag{3.34}$$

where $\mathcal{X}_\gamma^{\pi(\gamma)}$ is defined by

$$\mathcal{X}_I^{\pi(I)}\equiv\begin{cases}\displaystyle\sum_\pm\frac{1}{2}\cdot X_{\pm,I}^{1/2}\cdot X_{\mp,I}^{1/2} & \text{for}\quad\pi(I)=\mathrm{NS},\\ X_I^0\quad\text{for}\quad\pi(I)=\mathrm{R}.\end{cases}\tag{3.35}$$

On the other hand, the contractions among any $|\Delta\widetilde{\Psi}(\cdot)\rangle$ and any $|\Psi(\cdot)\rangle$ are just equal to zero, so that we have that

$$\begin{aligned}\vec{T}_{X^0}(|\Psi(p(\gamma))\rangle_\gamma+|\Delta\widetilde{\Psi}(p(\gamma))\rangle_\gamma)\cdot(|\Psi(p(\delta))\rangle_\delta+|\Delta\widetilde{\Psi}(p(\delta))\rangle_\delta)\\ \equiv{:}(|\Psi(p(\gamma))\rangle_\gamma+|\Delta\widetilde{\Psi}(p(\gamma))\rangle_\gamma)\cdot(|\Psi(p(\delta))\rangle_\delta+|\Delta\widetilde{\Psi}(p(\delta))\rangle_\delta){:}\\ +(|\overline{\Psi(p(\gamma))\rangle_\gamma+|\Delta\widetilde{\Psi}(p(\gamma))\rangle_\gamma)\cdot(|\Psi(p(\delta))\rangle_\delta+|\Delta\widetilde{\Psi}}(p(\delta))\rangle_\delta).\end{aligned}\tag{3.36}$$

The last contraction term on the right-hand side of Eq. (3.36) is found to be given by

$$(|\Psi(p(\gamma))\rangle_\delta+|\Delta\tilde{\Psi}(p(\gamma))\rangle_\gamma)\cdot(|\Psi(p(\delta))\rangle_\delta+|\Delta\tilde{\Psi}(p(\delta))\rangle_\delta)$$
$$=\left(\mathcal{X}_\gamma^{\pi(\gamma)}\cdot\frac{\mathcal{B}_{0,\gamma}}{L_{0,\gamma}}-\Theta_\gamma^{\pi(\gamma)}\right)\cdot P^{\pi(\gamma)}(\mathrm{GSO})|\nu_{\mathcal{S}}(\gamma,\delta)\rangle. \tag{3.37}$$

Similarly to Eqs. (3.13) and (3.14), we easily find from Eqs. (3.34) and (3.37) the following important identities:

$$(\mathcal{Q}_\gamma+\mathcal{Q}_\delta)\cdot(|\Psi(p(\gamma))\rangle_\gamma+|\Delta\tilde{\Psi}(p(\gamma))\rangle_\gamma)\cdot(|\Psi(p(\delta))\rangle_\delta+|\Delta\tilde{\Psi}(p(\delta))\rangle_\delta)=0. \tag{3.38}$$

Correspondingly to Eqs. (1.23) and (1.24), we introduce the *counter* bra state $\langle\tilde{\phi}=\tilde{\psi}=0|$ and the *counter* ket state $|\tilde{\phi}=\tilde{\psi}=0\rangle$, on which operates any *counter* fieldino $|\Delta\tilde{\Psi}(1)\rangle_r$ and $|\Delta\tilde{\Psi}(\frac{1}{2})\rangle_s$. Furthermore, the normal product among any *counter* fieldino satisfies

$$\langle\tilde{\phi}=\tilde{\psi}=0|:(|\Delta\tilde{\Psi}(1)\rangle_r\cdot|\Delta\tilde{\Psi}(\tfrac{1}{2})\rangle_s\cdots):|\tilde{\phi}=\tilde{\psi}=0\rangle=0 \tag{3.39}$$

and

$$\langle\tilde{\phi}=\tilde{\psi}=0|\tilde{\phi}=\tilde{\psi}=0\rangle=1. \tag{3.40}$$

In order to analyze

$$\vec{T}_{X^0}\cdot\exp(S_{\mathrm{INT}}^{\mathcal{T}}(\Psi+\Delta\tilde{\Psi})), \tag{3.41}$$

we consider the following transformation:

$$(|\Psi(p(r))\rangle_r+|\Delta\tilde{\Psi}(p(r))\rangle_r)\to(1+\varepsilon\cdot\mathcal{Q}_r)(|\Psi(p(r))\rangle_r+|\Delta\tilde{\Psi}(p(r))\rangle_r)$$
$$\equiv(|\mathring{\Psi}(p(r))\rangle_r+|\Delta\mathring{\tilde{\Psi}}(p(r))\rangle_r). \tag{3.42}$$

Similarly to Eq. (3.9), we find that

$$S_{\mathrm{INT}}^{\mathcal{T}}(\Psi+\Delta\tilde{\Psi})=S_{\mathrm{INT}}^{\mathcal{T}}(\mathring{\Psi}+\Delta\mathring{\tilde{\Psi}}). \tag{3.43}$$

On account of Eq. (3.38), we obtain that

$$(|\mathring{\Psi}(p(\gamma))\rangle_\gamma+|\Delta\mathring{\tilde{\Psi}}(p(\gamma))\rangle_\gamma)\cdot(|\mathring{\Psi}(p(\delta))\rangle_\delta+|\Delta\mathring{\tilde{\Psi}}(p(\delta))\rangle_\delta)$$
$$=(|\Psi(p(\gamma))\rangle_\gamma+|\Delta\tilde{\Psi}(p(\gamma))\rangle_\gamma)\cdot(|\Psi(p(\delta))\rangle_\delta+|\Delta\tilde{\Psi}(p(\delta))\rangle_\delta). \tag{3.44}$$

Therefore, we find in

$$:S^{\mathcal{T}}_{\mathrm{INT}}(\mathring{\Psi}+\Delta\mathring{\tilde{\Psi}})\cdot S^{\mathcal{T}}_{\mathrm{INT}}(\mathring{\Psi}+\Delta\mathring{\tilde{\Psi}}):\equiv\sum_{\{p\}}\ {}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),p(2),p(3),p(4)]|\frac{1}{4!}\times:\left(\prod_{r=1}^{4}(|\mathring{\Psi}(p(r))\rangle_r+|\Delta\mathring{\tilde{\Psi}}(p(r))\rangle_r\right): \tag{3.45}$$

that ${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[\ldots]|$ is just equal to the one obtained by

$$:S^{\mathcal{T}}_{\mathrm{INT}}(\Psi+\Delta\tilde{\Psi})\cdot S^{\mathcal{T}}_{\mathrm{INT}}(\Psi+\Delta\tilde{\Psi}):\equiv\sum_{\{p\}}\ {}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),p(2),p(3),p(4)]|\frac{1}{4!}\times:\left(\prod_{r=1}^{4}(|\Psi(p(r))\rangle_r+|\Delta\tilde{\Psi}(p(r))\rangle_r\right):, \tag{3.46}$$

which can be calculated by using contractions given by Eq. (3.37). Thus, we find from Eqs. (3.43), (3.45), and (3.46) that

$$\vec{T}_{X^0}\cdot\exp(S^{\mathcal{T}}_{\mathrm{INT}}(\mathring{\Psi}+\Delta\mathring{\tilde{\Psi}}))(\equiv\vec{T}_{X^0}\cdot\exp(S^{\mathcal{T}}_{\mathrm{INT}}(\Psi+\Delta\tilde{\Psi})))$$

$$=\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\ {}^{\mathcal{M}}\langle S[p(1),p(2),\ldots,p(N)]|\frac{1}{N!}:\left(\prod_{r=1}^{N}(|\mathring{\Psi}(p(r))\rangle_r+|\Delta\mathring{\tilde{\Psi}}(p(r))\rangle_r\right): \tag{3.47}$$

$$=\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\ {}^{\mathcal{M}}\langle S[p(1),p(2),\ldots,p(N)]|\frac{1}{N!}:\left(\prod_{r=1}^{N}(|\Psi(p(r))\rangle_r+|\Delta\tilde{\Psi}(p(r))s\rangle_r\right):. \tag{3.48}$$

Equating Eq. (3.47) to Eq. (3.48), we finally have proved the following ''general conservation of total chargino'' (GCTC):

$${}^{\mathcal{M}}\langle S[p(1),p(2),\ldots,p(N)]|\left(\sum_{r=1}^{N}\mathcal{Q}_r\right)=0, \tag{3.49}$$

(to be referred to as ''GCTC'')

Furthermore, we find from Eq. (3.48) that

$$\langle\tilde{\phi}=\tilde{\psi}=0|\vec{T}_{X^0}\cdot\exp(S^{\mathcal{T}}_{\mathrm{INT}}(\Psi+\Delta\tilde{\Psi}))|\tilde{\phi}=\tilde{\psi}=0\rangle$$

$$=\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\ {}^{\mathcal{M}}\langle S[p(1),p(2),\ldots,p(N)]|\cdot\frac{1}{N!}$$

$$\times\langle\tilde{\phi}=\tilde{\psi}=0|:\left(\prod_{r=1}^{N}(|\Psi(p(r))\rangle_r+|\Delta\tilde{\Psi}(p(r))\rangle_r\right):|\tilde{\phi}=\tilde{\psi}=0\rangle$$

$$=\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\ {}^{\mathcal{M}}\langle S[p(1),p(2),\ldots,p(N)]|\cdot\frac{1}{N!}:\left(\prod_{r=1}^{N}(|\Psi(p(r))\rangle_r\right):$$

$$\equiv:\exp\left(\sum_{N=0}^{\infty}\sum_{\{p(r)\}}\ {}^{\mathcal{M}}\langle S^{\mathcal{C}}[p(1),p(2),\ldots,p(N)]|\cdot\frac{1}{N!}:\left(\prod_{r=1}^{N}|\Psi(p(r))\rangle_r\right):\right):, \tag{3.50}$$

in the second step of which we have used Eqs. (3.39) and (3.40). Therefore, we can use the identification

$$\vec{T}_{X^0}\cdot\exp(S^{\mathcal{M}}_{\rm INT}(\Psi))\equiv\langle\tilde{\phi}=\tilde{\psi}=0|\vec{T}_{X^0}\cdot\exp(S^{\mathcal{T}}_{\rm INT}(\Psi+\Delta\tilde{\Psi}))|\tilde{\phi}=\tilde{\psi}=0\rangle, \tag{3.51}$$

which gives ${}^{\mathcal{M}}\langle S[p(1),\ldots,p(N)]|$ satisfying the ''GCTC'' (3.49). In the following, we explain how $S^{\mathcal{M}}_{\rm INT}(\Psi)$ on the left-hand side of Eq. (3.51) can be determined by terms on the right-hand side of Eq. (3.51). For this purpose, we first rewrite *counter* fieldinos within the time-ordered product into the normal-ordered form, with the help of Wick's theorem[16] by using the operator product expansion Eqs. (3.32)–(3.34) (among *counter* fieldinos). At this stage, fieldinos are still kept time ordered. For an example, we use

$$\begin{aligned}
&\vec{T}_{X^0}(|\Psi(p(\gamma))\rangle_\gamma+|\Delta\tilde{\Psi}(p(\gamma))\rangle_\gamma)\cdot(|\Psi(p(\delta))\rangle_\delta+|\Delta\tilde{\Psi}(p(\delta))\rangle_\delta)\\
&=\vec{T}_{X^0}(|\Psi(p(\gamma))\rangle_\gamma\cdot|\Psi(p(\delta))\rangle_\delta+|\Psi(p(\gamma))\rangle_\gamma|\Delta\tilde{\Psi}(p(\delta))\rangle_\delta\\
&\quad+|\Delta\tilde{\Psi}(p(\gamma))\rangle_\gamma\cdot|\Psi(p(\delta))\rangle_\delta)+:|\Delta\tilde{\Psi}(p(\gamma))\rangle_\gamma\cdot|\Delta\tilde{\Psi}(p(\delta))\rangle_\delta:\\
&\quad+\overbrace{|\Delta\tilde{\Psi}(p(\gamma))\rangle_\gamma\cdot|\Delta\tilde{\Psi}}(p(\delta))\rangle_\delta.
\end{aligned} \tag{3.52}$$

Carrying out these procedures, we are lead to

$$\vec{T}_{X^0}\cdot\exp(S^{\mathcal{T}}_{\rm INT}(\Psi+\Delta\tilde{\Psi}))\equiv\vec{T}_{X^0}:\exp(S^{\mathcal{M}}_{\rm INT}(\Psi)+\Sigma_{\rm INT}(\Psi;\Delta\tilde{\Psi})):. \tag{3.53}$$

The normal ordering : : on the right-hand side of Eq. (3.53) has been taken with respect *only* to *counter* fieldinos, and $\Sigma_{\rm INT}\{\Psi;\Delta\tilde{\Psi}\}$ is composed of various terms involving (*more than one*) counter fieldinos $\Delta\tilde{\Psi}$'s in the *normal-ordered form*. Therefore, Eq. (3.39) gives

$$\langle\tilde{\phi}=\tilde{\psi}=0|:(\Sigma_{\rm INT}(\Psi;\Delta\tilde{\Psi})\cdots\Sigma_{\rm INT}(\Psi;\Delta\tilde{\Psi})):|\tilde{\phi}=\tilde{\psi}=0\rangle=0. \tag{3.54}$$

On the other hand, the modified action $S^{\mathcal{M}}_{\rm INT}(\Psi)$ *determined* by Eq. (3.53) would be obtained in the following form:

$$S^{\mathcal{M}}_{\rm INT}(\Psi)=S_{\rm INT}(\Psi)-\sum_{n=1}^{\infty}\Delta S_{\rm INT}(\Psi;n), \tag{3.55}$$

where $\Delta S_{\rm INT}(\Psi;n)$ is the term resulting from contractions (of n times) among *counter* fieldinos, and it should *not* involve any *counter* fieldino. Furthermore, $\Delta S_{\rm INT}(\Psi;n)$ involves only fieldinos which are given in the *time-ordered* form. Thus we have *proved* the following *theorem*.

Theorem: With the help of $S^{\mathcal{M}}_{\rm INT}(\Psi)$ in (3.55) determined by Eq. (3.54),

$$\vec{T}_{X^0}\cdot\exp(S^{\mathcal{M}}_{\rm INT}(\Psi))\equiv:\exp\left(\sum_{N=0}^{\infty}\sum_{\{p(r)\}}{}^{\mathcal{M}}\langle S^{\mathcal{C}}[p(1),\ldots,p(N)]|\cdot\frac{1}{N!}:\left(\prod_{r=1}^{N}|\Psi(p(r))\rangle_r\right):\right), \tag{3.56}$$

where the (BRST-invariant) ''amputated N-scatt'' ${}^{\mathcal{M}}\langle S^{\mathcal{C}}[p(1),\ldots,p(N)]|$ contains any contribution in **g**-loops [which satisfies the ''general conservation of the total chargino'' (3.29)]. Our final result (or rule) in calculating the ''amputated N-scatt'' ${}^{\mathcal{M}}\langle S^{\mathcal{C}}[p(1),\ldots,p(N)]|$ is quite simple. Amputated scatts can be calculated by using the *un*modified interaction term $S_{\rm INT}(\Psi)$ in (1.20), provided that *all* propagators are *effectively* modified as follows:

$$\mathcal{X}_{\gamma}^{\pi(\gamma)}\cdot\frac{\mathcal{B}_{0,\gamma}}{L_{0,\gamma}}\cdot P^{\pi(\gamma)}(\mathrm{GSO})|\nu_{\mathcal{S}}(\gamma,\delta)\rangle\rightarrow\left(\mathcal{X}_{\gamma}^{\pi(\gamma)}\cdot\frac{\mathcal{B}_{0,\gamma}}{L_{0,\gamma}}-\Theta_{\gamma}^{\pi(\gamma)}\right)\cdot P^{\pi(\gamma)}(\mathrm{GSO})|\nu_{\mathcal{S}}(\gamma,\delta)\rangle. \tag{3.57}$$

Finally, we prove that the physical scattering amplitude (N_B,N_F) in (1.22) is given by

$$(N_B,N_F)=\sum_{\mathbf{g}=0}^{\infty}(N_B,N_F)_{\mathbf{g}}, \tag{3.58}$$

$(N_B,N_F)_{\mathbf{g}}$ being given by

$$\begin{aligned}(N_B,N_F)_{\mathbf{g}}\equiv{}^{\mathcal{M}}\langle S_{\mathbf{g}}^{\mathcal{C}}[p(1),p(2),\ldots,p(N_B+N_F)]|\langle\phi=\psi=0|\cdot\frac{1}{N_B!}\\ \times:\left(\prod_{r''=1}^{N_B}|\Psi(1)\rangle_{r''}\right)\cdot\frac{1}{N_F!}\left(\prod_{r''=N_B+1}^{N_B+N_F}|\Psi(\tfrac{1}{2})\rangle_{r''}\right):\left(\prod_{r=1}^{N_B}B_{b_r}\{\phi\}\right)\\ \times\left(\prod_{r=N_B+1}^{N_B+N_F}F_{f_r}\{\psi\}\right)|\phi=\psi=0\rangle\quad\text{with }N=N_B+N_F\\ [\text{see Eqs. (1.22) and (3.56)}]\end{aligned} \tag{3.59}$$

$$\begin{aligned}=\pm{}^{\mathcal{M}}\langle S_{\mathbf{g}}^{\mathcal{C}}[p(1),p(2),\ldots,p(r),\ldots,p(N_B+N_F)]|\\ \times\left(\prod_{r=1}^{N_B}|b_r(1)\rangle_r\right)\cdot\left(\prod_{r=N_B+1}^{N_B+N_F}|f_r(\tfrac{1}{2})\rangle_r\right).\end{aligned} \tag{3.60}$$

The signature $\pm$ in Eq. (3.60) is necessary, since we reorder Grassman odd operators for the purpose of using the following formulas [in the process of deriving Eq. (3.60)]:

$$\begin{aligned}|\Psi\overbrace{(1)\rangle_{r''}\cdot B_{b_r}}\{\phi\}\quad\text{for }r=1-N_B\quad[\text{use Eq. (1.21a)}]\\ =\langle\nu_{\mathcal{S}}(r',r)|\mathcal{Q}_{r'}\cdot Y[\bar{r}']Y[r']\cdot|\Psi(1)\rangle_{r''}\cdot|\Psi(1)\rangle_{r'}\\ \cdot|b_r(1)\rangle_r\quad[\text{use Eq. (3.1)}]\\ =\langle\nu_{\mathcal{S}}(r',r)|\mathcal{Q}_{r'}\cdot\frac{\mathcal{B}_{0,r'}}{L_{0,r'}}|\nu_{\mathcal{S}}(r'',r')\rangle(P_r(1)|b_r(1)\rangle_r)\quad[\text{use Eq. (D6)}]\\ =(\langle\nu_{\mathcal{S}}(r',r)|\nu_{\mathcal{S}}(r'',r')\rangle)|b_r(1)\rangle_r\quad[\text{use Eq. (E40)}]\\ -\langle\nu_{\mathcal{S}}(r',r)|\frac{\mathcal{Q}_{r'}}{L_{0,r'}}|\nu_{\mathcal{S}}(r'',r')\rangle(\mathcal{B}_{0,r}|b_r(1)\rangle_r)\quad[\text{use Eq. (D7)}]\\ =\delta_{rr''}\cdot|b_r(1)\rangle_r\end{aligned} \tag{3.61}$$

and

$$\langle\Psi\overbrace{(\tfrac{1}{2})\rangle_{r''}\cdot F_{f_r}}\{\psi\}=\delta_{rr''}\cdot|f_r(\tfrac{1}{2})\rangle_r\quad\text{for }r=(N_B+1)-(N_B+N_F), \tag{3.62}$$

which can be obtained similarly to Eq. (3.61) by using Eqs. (1.21b), (3.2), (D16), and (D19).

IV. PHYSICAL SCATTERING AMPLITUDES IN g-LOOPS

We have seen in Sec. III that the formula (3.60) gives the physical scattering amplitude $(N_B,N_F)_{\mathbf{g}}$ [in **g**-loops] among N_B *physical* bosonic and N_F *physical* fermionic particles, by using the (BRST-invariant) ''amputated N-scatt'' $\langle S_{\mathbf{g}}^{\mathcal{C}}[p(1),\ldots,p(N)]|$, which can be given by the formulas (3.50) and (3.28). [We have found that amputated scatts can be calculated by using unmodified interaction term $S_{\mathrm{INT}}(\Psi)$ in (1.20), provided that any propagator is *effectively* modified accordingly to Eq. (3.57).] As we have seen in our previous paper[5] on QBFT, physical scattering amplitudes (3.60) can be calculated by using the ''inlayed coordinate system *genus* **g**'' (ICS **g**), which describes the Riemann *surface* $\mathcal{R}$ with *B-boundaries* and *H-handles*, with $\mathbf{g}+1=B+H$. In this paper, we only consider Riemann surface $\mathcal{R}$ with $B=\mathbf{g}+1$ and $H=0$, i.e., we do *not* calculate the contribution obtained by gluing two *open* string punctures on different boundaries, since each gluing of this kind adds one handle to the Riemann surface $\mathcal{R}$. We notice that H handles in Riemann surface $\mathcal{R}$ represent contributions from H closed channels. Therefore, in order to restore the manifest factorization in the H channels (of closed string), we must include the closed string from the beginning, based on the open-closed string field theory. Then we could obtain the factorizable amplitudes in genus $\mathbf{g}=H+B-1$, where $H(B)$ is the number of handles (boundaries). We point out that *closed* QBFT has been already proposed.[4]

At this stage, we summarize properties of the Riemann surface $\mathcal{R}$ which is to be used in calculating $(N_B,N_F)_{\mathbf{g}}$ in (3.60) in **g**-loops. The Riemann surface $\mathcal{R}$ is punctured at $(N+2\mathbf{g})$ points on the real axis. The Schottky double $\mathcal{D}$ of the Riemann surface $\mathcal{R}$ has the Ith unpunctured ring domain, which has the Ith propagating strip-time T_I [for $1\leqslant I\leqslant(N+2\mathbf{g}-3)$]. This Schottky double $\mathcal{D}$ has been described by GM-$\mathcal{FE}$ in (2.1) [with $(N+2\mathbf{g})$ punctures]. Next, in constructing the Riemann surface $\mathcal{R}'$ with $(\mathbf{g}+1)$-boundaries, $2\mathbf{g}$-punctured ring domains in the Riemann surface $\mathcal{R}$ should be glued into **g**-*un*punctured ring domains. As we see from the (Wick's) expansion (3.3), this can be realized by the following **g**-contractions among fieldinos:

$$\prod_{h=1}^{\mathbf{g}} |\Psi(\overbrace{p(\gamma^h))\rangle_{\gamma^h}|\Psi(p}(\delta^h))\rangle_{\delta^h}$$
$$=\left(\prod_{\pi(\gamma^h)=NS}\left(\sum \frac{1}{2}\cdot X^{1/2}_{\pm\gamma^h}\cdot X^{1/2}_{\mp\delta^h}\right)\right)\left(\prod_{\pi(\gamma^h)=R} X^0_{\gamma^h}\right)$$
$$\times\prod_{h=1}^{\mathbf{g}}\left(\frac{\mathcal{B}_{0,\gamma^h}\cdot\eta_{0,\gamma^h}}{L_{0,\gamma^h}}\cdot P^{\pi(\gamma^h)}_{\gamma^h}(\mathrm{GSO})|V_L(\gamma^h,\delta^h)\rangle\right)\quad\text{with}\quad \mathbf{g}\equiv g_B+g_F, \tag{4.1}$$

where we have used the formulas (E31) and (E32). Thus we need the following N-vertex functino ${}_s\langle V(1,2,...,N)|_{(g_B,g_F)}$ ''in $\mathbf{g}(=g_B+g_F)$-loops'' in the *large* Hilbert space:

$${}_s\langle V(1,2,...,N)|_{(g_B,g_F)}\quad\text{with}\quad \mathbf{g}=g_B+g_F$$
$$\equiv(\langle\mathcal{Q}=0|\xi_0)\cdot{}_s\langle IV(1,2,...,N+2\mathbf{g})|\cdot(|\mathcal{Q}=0\rangle)$$
$$\times\prod_{h=1}^{\mathbf{g}}(\exp(-T_{\gamma^h}\cdot L_{0,\gamma^h}\cdot P^{\pi(\gamma^h)}_{\gamma^h}(\mathrm{GSO})|V_L(\gamma^h,\delta^h)\rangle) \tag{4.2}$$

$$\equiv\mathrm{Tr}_{\varphi}\,\Omega^N_s((g_B,g_F)), \tag{4.3}$$

$\Omega^N_s((g_B,g_F))$ being defined by

$$\Omega^N_s((g_B,g_F))\equiv\{{}_sN^{2\mathbf{g}}\}\cdot\prod_{h=1}^{\mathbf{g}}(P^{\pi(\gamma^h)}_{\gamma^h}(\mathrm{GSO})|V_L(\gamma^h,\delta^h)\rangle). \tag{4.4}$$

The operator $\Omega_s^N((g_B, g_F))$ in (4.4) is the function of N *external* operators as well as the *inlint* operators in the ''ICS s'' (representing the Riemann surface $\mathcal{R}$ with $N+2\mathbf{g}$ punctures). In Eqs. (4.1) and (4.2), the number of h satisfying $\pi(\gamma^h)=\mathrm{NS}$ (R) is denoted by g_B (g_F), and the operator $\{_s N^{2\mathbf{g}}\}$ in Eq. (4.4) is defined by

$$\{_s N^{2\mathbf{g}}\} \equiv \vec{R}\cdot(|\mathcal{Q}=0\rangle\cdot\langle\mathcal{Q}=0|\xi_0)\cdot\left(\prod_{h=1}^{\mathbf{g}}\prod_{\varphi}\left(_{\delta^h}\langle q_{\delta^h}(\varphi)=0|W^{\varphi}_{\delta^h}[z_{\delta^h s}(w_{\delta^h})]\right)\right)$$

$$\times\left(\prod_{r=1}^{N}\prod_{\varphi}\left(_r\langle q_r(\varphi)=0|W_r^{\varphi}[z_{rs}(w_r)]\right)\right)$$

$$\times\left(\prod_{h=1}^{\mathbf{g}}\prod_{\varphi}\left(_{\gamma^h}\langle q_{\gamma^h}(\varphi)=0|W^{\varphi}_{\gamma^h}[z_{\gamma^h s}(x_{\gamma^h}\cdot w_{\gamma^h})]\right)\right). \tag{4.5}$$

$W^{\varphi}[\cdots]$'s being inlaying operators given by Eqs. (2.4)–(2.6). Incidentally, the γ^h-th propagating strip-time T_{γ^h} in the formula (4.2) has been introduced through

$$\frac{1}{L_{0,\gamma^h}}=\int_0^{\infty} dT_{\gamma^h}\exp(-T_{\gamma^h}\cdot L_{0,\gamma^h}) \quad \text{for } h=1-\mathbf{g} \tag{4.6a}$$

with

$$x_{\gamma^h}\equiv\exp(-T_{\gamma^h}) \quad [\text{in Eq. (4.5)}]. \tag{4.6b}$$

Tr_{φ} in Eq. (4.3) [Π_{φ} in Eq. (4.5)] is the trace (product) over all *inlint* φ-modes (for $\varphi=\vec{X},\phi^j,\sigma,\phi,\chi$).

We notice that contractions (4.1) induce the hth $\gamma^h\delta^h$-gluing (for $h=1-\mathbf{g}$) which glues the γ^hth ring domain (punctured at $Z_{\gamma^h s}$) to the δ^hth ring domain (punctured at $Z_{\delta^h s}$) in the Schottky double $\mathcal{D}$, resulting to create the new hth *un*punctured ring domain which has the hth propagating strip-time $T_{\gamma^h s}\equiv\log(1/x_{\gamma^h})$ (for $h=1-\mathbf{g}$). Thus, the new Riemann surface $\mathcal{R}'$ (which has been obtained from the original Riemann surface $\mathcal{R}$ by the $\gamma\delta$-gluings) has N punctures on the real axis, having $(\mathbf{g}+1)$-*boundaries* on the real axis. Furthermore, the Schottky double $\mathcal{D}'$ [of this Riemann surface $\mathcal{R}'$] has the Ith unpunctured ring domain which has the propagating strip-time T_I [for $I=1-(N+2\mathbf{g}-3)$], as well as the hth *un*punctured ring domain having the hth propagating strip-time $\log(1/x_{\gamma^h})$ (for $h=1-\mathbf{g}$). Therefore, when the number of the punctures (boundaries) is N $(\mathbf{g}+1)$, the total number of modular parameters is just equal to $(N+3\mathbf{g}-3)$. In our previous paper,[5] we have proved that thus obtained Riemann surface $\mathcal{R}'$ can be conformally mapped [see Eqs. (6.16)–(6.18) in Ref. 5] to the other Riemann surface described by the ''ICS **g**,'' which will be explained in the following.

In order to describe the ''ICS **g**,'' we have used (in open QBFT[5]) the following ''Samuel's fundamental equation'' (S-$\mathcal{FE}$) in (4.7).[19] We briefly review our previous results, hereafter. (See Ref. 5 for more details.)

$$\frac{dz_{\mathbf{g}}}{d\rho_{\mathbf{g}}}=\nu_{\mathbf{g}}(z_{\mathbf{g}}) \quad \text{with} \quad \mathbf{g}=g_B+g_F. \tag{4.7}$$

[The coordinate $z_{\mathbf{g}}(\rho_{\mathbf{g}})$ satisfying S-$\mathcal{FE}$ (4.7) has been referred to as the inlayed coordinate $z_{\mathbf{g}}(\rho_{\mathbf{g}})$ in the ''inlayed coordinate system in (*genus*) **g**'' (ICS **g**).] We impose Witten's midpoint interactions in Eq. (4.7). Samuel[19] has imposed that

$$d\rho_{\mathbf{g}}=\frac{dz_{\mathbf{g}}}{\nu_{\mathbf{g}}(z_{\mathbf{g}})}=\frac{dS_h(z_{\mathbf{g}})}{\nu_{\mathbf{g}}(S_h(z_{\mathbf{g}}))} \quad \text{for each } h=1-\mathbf{g}, \tag{4.8}$$

where $S_h(z_{\mathbf{g}})$ is the SL$(2R)$ mapping defined by

$$S_h(z_{\mathbf{g}})\equiv\frac{\alpha_h\cdot z_{\mathbf{g}}+\beta_h}{\gamma_h\cdot z_{\mathbf{g}}+\delta_h}. \tag{4.9}$$

Furthermore, Samuel has found[19] that the function $\nu_{\mathbf{g}}(z_{\mathbf{g}})$ given by

$$(\nu_{\mathbf{g}}(z))^{-2}\equiv R_{\mathbf{g}}^{-2}\cdot\Theta\left(\vec{\Delta}^{z_0}-\sum_{\iota=1}^{\mathbf{g}}\int_{z_0}^{y_\iota^{\mathbf{g}}}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}+\int_{z_0}^{z}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}\middle|\tau\right)$$
$$\times\frac{(\Pi_{\iota=g+1}^{2N+4\mathbf{g}-4}E(z,y_\iota^{\mathbf{g}}|\tau))}{(\Pi_{r=1}^{N}(E(z,Z_{r\mathbf{g}}|\tau))^2)}\cdot(\sigma(z))^3\quad\text{with}\quad N=N_B+N_F \tag{4.10}$$

can satisfy the **g**-*conditions* (4.8), *provided* that the following **g**-*constraints* are imposed on modular parameters in $\nu_{\mathbf{g}}(z_{\mathbf{g}})$ in (4.10):

$$4\cdot\Delta_h^{z_0}-2=\sum_{\iota=1}^{2N+4\mathbf{g}-4}\int_{z_0}^{y_\iota^{\mathbf{g}}}dz'\frac{\omega_h(z')}{2\pi\sqrt{-1}}-\sum_{r=1}^{N}2\int_{z_0}^{Z_{r\mathbf{g}}}dz'\frac{\omega_h(z')}{2\pi\sqrt{-1}}\quad(h=1-\mathbf{g}). \tag{4.11}$$

[Unfortunately, there exist misprints in ''Eqs. (6.8) and Eq. (6.11) in Ref. 5,'' which are corrected in Eqs. (4.10) and (4.11).] In the following, we explain various functions used in Eqs. (4.10) and (4.11). Functions τ, ω, $\Delta_h^{z_0}$, E, σ, and Θ are the same as those given in VPFHLS's paper,[20] corresponding equations of which being given as follows: The period matrix (3.11), the holomorphic differential $\vec{\omega}(z)dz\equiv(\omega_1,\omega_2,\ldots,\omega_{\mathbf{g}})$ [with the *one-form* $\omega_h\equiv\omega_h(z)\cdot dz$ in (3.12) for $h=1-\mathbf{g}$] being bases for the holomorphic differentials of the Riemann surface, the **g**-*dimensional vector* of Riemann constants $\vec{\Delta}^{z_0}\equiv(\Delta_1^{z_0},\Delta_2^{z_0},\ldots,\Delta_{\mathbf{g}}^{z_0})$ in (3.13), the prime form (3.14), the sigma function ''σ'' (3.15), and the Riemann theta function (3.18). [Hereafter, the **g**-*dimensional vector space* is denoted by $V_{\mathbf{g}}$, so that $\vec{\omega}(z)$, $\vec{\Delta}^{z_0}\in V_{\mathbf{g}}$ in Eq. (4.10) and hereafter.] We notice that the *one-form* $\omega_h\equiv dz\cdot\omega_h(z)$ in (3.12) in VPFHLS[20] is written [in Samuel's paper] by the *one-form* $2\pi\sqrt{-1}\cdot\omega_h$ in (3.6).[19] The point $z_0\in F_{D\mathbf{g}}$ is arbitrary. [The boundary of the *fundamental region* $F_{D\mathbf{g}}$ in the ''ICS **g**'' has been given by Eq. (6.12) in Ref. 5.] Incidentally, the **g**-constraints in (4.11) [which should be imposed in order for $\nu_{\mathbf{g}}(z_{\mathbf{g}})$ in (4.10) to satisfy **g**-conditions in (4.8)] have been derived[19] by using the following formulas in VPFHLS's notations.[20] [See Table I and Eq. (3.18) of Ref. 19.]

$$E(S_h(z),z')\quad\text{for each}\quad S_h\quad(h=1-\mathbf{g})$$
$$=-\exp\left(-\tau_{hh}\sqrt{-1}\cdot\pi-\int_{z'}^{z}dz''\omega_h(z'')\right)\cdot\left(\frac{dS_h(z)}{dz}\right)^{1/2}\cdot E(z,z'), \tag{4.12}$$

$$\Theta\begin{pmatrix}\vec{a}\\\vec{b}\end{pmatrix}\left(\vec{\xi}+\int_{z_0}^{S_h(z)}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}\middle|\tau\right)\quad\text{for each}\quad S_h\quad[h=1-\mathbf{g}]$$

$$=\exp\left(-\tau_{hh}\sqrt{-1}\cdot\pi-\int_{z_0}^{z}dz'\omega_h(z')-2\pi\sqrt{-1}(\xi_h+b_h)\right)\Theta\begin{pmatrix}\vec{a}\\\vec{b}\end{pmatrix}$$
$$\times\left(\vec{\xi}+\int_{z_0}^{z}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}\middle|\tau\right), \tag{4.13}$$

$$\sigma(S_h(z))=(-)^{\mathbf{g}}\cdot\exp\left(\tau_{hh}(\mathbf{g}-1)\sqrt{-1}\cdot\pi-2\pi\sqrt{-1}\Delta_h^{z_0}+(\mathbf{g}-1)\int_{z_0}^{z}dz'\,\omega_h(z')\right)$$

$$\times\left(\frac{dS_h(z)}{dz}\right)^{-\mathbf{g}/2}\cdot\sigma(z)\quad\text{for each } S_h\quad[h=1-\mathbf{g}],\tag{4.14}$$

and

$$\frac{\sigma(\tilde{z})}{\sigma(z)}=\frac{\Theta\left(\Sigma_{i=1}^{\mathbf{g}}\int_{z_0}^{z_i}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}-\int_{z_0}^{\tilde{z}}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}-\vec{\Delta}^{z_0}\middle|\tau\right)}{\Theta\left(\Sigma_{i=1}^{\mathbf{g}}\int_{z_0}^{z_i}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}-\int_{z_0}^{z}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}-\vec{\Delta}^{z_0}\middle|\tau\right)}\cdot\frac{(\Pi_{i=1}^{\mathbf{g}}E(z_i,z))}{(\Pi_{i=1}^{\mathbf{g}}E(z_i,\tilde{z}))}.\tag{4.15}$$

Furthermore, we have used [in Eqs. (4.13) and (4.15)] the following Riemann theta function Θ in *genus* $\mathbf{g}$:

$$\Theta\begin{pmatrix}\vec{a}\\\vec{b}\end{pmatrix}(\vec{\xi}|\tau),\tag{4.16}$$

where and hereafter $\vec{a},\vec{b},\vec{\xi}(\in V_{\mathbf{g}})$ are shorthand notations for the $\mathbf{g}$-set of variables given by

$$\vec{a}\equiv(a_1,a_2,\dots a_{\mathbf{g}})\quad\text{and}\quad\vec{b}\equiv(b_1,b_2,\dots,b_{\mathbf{g}})\tag{4.17}$$

and

$$\vec{\xi}\equiv(\xi_1,\xi_2,\dots,\xi_{\mathbf{g}}).\tag{4.18}$$

The Riemann theta function (4.16) is defined by

$$\Theta\begin{pmatrix}\vec{a}\\\vec{b}\end{pmatrix}(\vec{\xi}|\tau)\equiv\sum_{\{n_h\}\in Z}\exp\left(2\pi\sqrt{-1}\left(\sum_{i,h=1}^{\mathbf{g}}\frac{1}{2}(n_i+a_i)\tau_{ih}(n_h+a_h)+\sum_{h=1}^{\mathbf{g}}(n_h+a_h)(\xi_h+b_h)\right)\right),\tag{4.19}$$

τ_{ih} being the period matrix elements. Furthermore, we shall use the following shorthand notation:

$$\Theta(\vec{\xi}|\tau)\equiv\Theta\begin{pmatrix}\vec{0}\\\vec{0}\end{pmatrix}(\vec{\xi}|\tau).\tag{4.20}$$

There exist various parameters in $\nu_{\mathbf{g}}(z)$ in (4.10): Constant $R_{\mathbf{g}}$ is some real normalization constant. The rth puncture $Z_{r\mathbf{g}}$ (for $r=1-N$) is located somewhere on the real axis (in the fundamental region $F_{D\mathbf{g}}$ in the ''ICS $\mathbf{g}$''), while the interacting points $\mathcal{Y}^{\mathbf{g}}_{\pm\iota}$ [for $\iota=1-(N-2)$], and $\mathcal{Y}^{\mathbf{g}}_{\pm\gamma^h}$, $\mathcal{Y}^{\mathbf{g}}_{\pm\delta^h}$ (for $h=1-\mathbf{g}$) are complex conjugates of each other among $\pm$ and all of them exist within $F_{D\mathbf{g}}$. Interacting points $\mathcal{Y}^{\mathbf{g}}_{\iota}$'s used in $\nu_{\mathbf{g}}(z)$ in (4.10) are differently numbered $(2N+4\mathbf{g}-4)$ coordinates of interacting points in $F_{D\mathbf{g}}$. In addition to these parameters, there exist $3\mathbf{g}$ real parameters characterizing hyperbolic $\mathrm{SL}(2R)$ matrices S_h (for $h=1-\mathbf{g}$), which are used in defining the functions Θ, E, τ, $\vec{\Delta}^{z_0}$, and $\vec{\omega}(z)dz$ in the formula $\nu_{\mathbf{g}}(z)$ in (4.10). Thus, we have $3N+7\mathbf{g}-3[=1+N+(2N+4\mathbf{g}-4)+3\mathbf{g}]$ *real* parameters *in total*. On the other hand, any one among $(N+3\mathbf{g}-3)$ unpunctured ring domains and N punctured ring domains in the ''ICS $\mathbf{g}$'' should have the same width 2π. Furthermore, propagating strip-times [i.e., T_I's for $I=1-(N+2\mathbf{g}-3)$ and T_{γ^h}'s for $h=1-\mathbf{g}$ in Eq. (4.6a)] can be arbitrary non-negative real numbers, and they can be used as *modular parameters*. Together with the $\mathbf{g}$-constraints in (4.11), we have $3N+7\mathbf{g}-6[=(N+3\mathbf{g}-3)+N+(N+3\mathbf{g}-3)+\mathbf{g}]$ constraints *in total*. Thus, *all* parameters in $\nu_{\mathbf{g}}(z)$

in (4.10) are determined in terms of $(N+3\mathbf{g}-3)$ modular parameters (i.e., propagating strip-times T_I's and T_{γ^h}'s), *except for* three free (real) parameters, existence of which reflects the fact that there remains still *one* SL$(2R)$ ambiguity in defining the ''ICS **g**.''[5]

With the help of the ''ICS **g**,'' $\Omega_{\mathbf{g}}^N((g_B,g_F))$ (in the ''ICS **g**'') is defined by

$$\Omega_{\mathbf{g}}^N(g_B,g_F)\equiv\{_{\mathbf{g}}N^{2\mathbf{g}}\}\cdot\left(\prod_{h=1}^{\mathbf{g}} P_{\gamma^h}^{\pi(\gamma^h)}(\mathrm{GSO})|V_L(\gamma^h,\delta^h)\rangle\right)$$

$$[\text{cf. Eqs. (6.4) and (6.19) in Ref. 5}], \tag{4.21}$$

where $\{_{\mathbf{g}}N^{2\mathbf{g}}\}$ (in the ''ICS **g**'') is given by

$$\{_{\mathbf{g}}N^{2\mathbf{g}}\}\equiv\vec{R}\cdot(U_s|\mathcal{Q}=0\rangle\cdot\langle\mathcal{Q}=0|\xi_0\cdot U_s^{-1})\cdot\left(\prod_{r=1}^{N}\prod_{\varphi}\left(_r\langle q_r^{\varphi}=0|W_r^{\varphi}[z_{r\mathbf{g}}]\right)\right),$$

$$\times\left(\prod_{h=1}^{\mathbf{g}}\prod_{\varphi}\left(_{\gamma^h}\langle q_{\gamma^h}^{\varphi}=0|W_{\gamma^h}^{\varphi}[z_{\gamma^h\mathbf{g}}]\right)\cdot\left(_{\delta^h}\langle q_{\delta^h}^{\varphi}=0|W_{\delta^h}^{\varphi}[z_{\delta^h\mathbf{g}}]\right)\right)$$

$$[\text{see Eq. (6.19) in Ref. 5}], \tag{4.22}$$

where the *inlint* conformal mapping operator U_s is the one given by Eqs. (6.16) and (6.17) in Ref. 5. [In Eq. (4.22), the existence of the inlint zero-mode ξ_0 is essential.]

The (rth external) physical vertex operators $\mathcal{G}_r(0)$'s [with the conformal weight $d(\mathcal{G})$] are *expressed* to be inlayed into inlint operators $\mathcal{G}$ (r; ''ICS **g**'') in the ''ICS **g**'' as follows:

$$\mathcal{G}_r(0)\overset{\mathcal{I}}{\Rightarrow}\left(\frac{dz_{r\mathbf{g}}(w)}{dw}\right)^{d(\mathcal{G})}\cdot\mathcal{G}(z_{r\mathbf{g}}(w))\bigg|_{w=0}\equiv\mathcal{G}(r;\text{``ICS }\mathbf{g}\text{''}). \tag{4.23}$$

As for various *nonlocal* operators, we also use the following (similar) shorthand notations for representing inlaying identities:

$$\Sigma\tfrac{1}{2}\cdot X_{\pm\gamma^h}^{1/2}\cdot X_{\mp\gamma^h}^{1/2}\overset{\mathcal{I}}{\Rightarrow}\Sigma\tfrac{1}{2}\cdot X_{\pm}^{1/2}(\gamma^h;\text{``ICS }\mathbf{g}\text{''})\cdot X_{\mp}^{1/2}(\gamma^h;\text{``ICS }\mathbf{g}\text{''})\quad\text{for }\pi(\gamma^h)=\mathrm{NS}, \tag{4.24}$$

$$X_{\gamma^h}^0\overset{\mathcal{I}}{\Rightarrow}X^0(\gamma^h;\text{``ICS }\mathbf{g}\text{''})\quad\text{for }\pi(\gamma^h)=\mathrm{R}, \tag{4.25}$$

$$\begin{aligned}\mathcal{B}_{0,I}&\equiv\oint_0\frac{w_I dw_I}{2\pi\sqrt{-1}}\mathcal{B}_I(w_I)\overset{\mathcal{I}}{\Rightarrow}\mathcal{B}_0(I;\text{``ICS }\mathbf{g}\text{''})\\&\equiv\oint_0\frac{w_I dw_I}{2\pi\sqrt{-1}}(z_{I\mathbf{g}}^{(1)}(w_I))^2\cdot\mathcal{B}(z_{I\mathbf{g}}(w_I))\\&=\oint_{C_{I\mathbf{g}}}\frac{dz_{\mathbf{g}}}{2\pi\sqrt{-1}}\nu_{\mathbf{g}}(z_{\mathbf{g}})\cdot\mathcal{B}(z_{\mathbf{g}})\quad\text{for }I=1-(N+2\mathbf{g}-3),\end{aligned} \tag{4.26}$$

$$\begin{aligned}\mathcal{B}_{0,\gamma^h} &\equiv \oint_0 \frac{w_{\gamma^h} dw_{\gamma^h}}{2\pi\sqrt{-1}} \mathcal{B}_{\gamma^h}(w_{\gamma^h}) \overset{\mathcal{I}}{\Rightarrow} \mathcal{B}_0(\gamma^h;\text{``ICS }\mathbf{g}\text{''}) \\ &\equiv \oint_0 \frac{w_{\gamma^h} dw_{\gamma^h}}{2\pi\sqrt{-1}} (z^{(1)}_{\gamma^h\mathbf{g}}(w_{\gamma^h}))^2 \cdot \mathcal{B}(z_{\gamma^h\mathbf{g}}(w_{\gamma^h})) \\ &= \oint_{C_{\gamma^h\mathbf{g}}} \frac{dz_{\mathbf{g}}}{2\pi\sqrt{-1}} \nu_{\mathbf{g}}(z_{\mathbf{g}}) \cdot \mathcal{B}(z_{\mathbf{g}}) \quad \text{for } h=1-\mathbf{g} \end{aligned} \tag{4.27}$$

and

$$\eta_{0,\gamma^h} \equiv \oint_0 \frac{dw_{\gamma^h}}{2\pi\sqrt{-1}} \eta_{\gamma^h}(w_{\gamma^h}) \overset{\mathcal{I}}{\Rightarrow} \eta_0(\gamma^h;\text{``ICS }\mathbf{g}\text{''}) \equiv \oint_{C_{\gamma^h\mathbf{g}}} \frac{dz_{\mathbf{g}}}{2\pi\sqrt{-1}} \eta(z_{\mathbf{g}}) \quad \text{for } h=1-\mathbf{g}. \tag{4.28}$$

Thus, $\mathcal{X}^{\pi(\gamma^h)}_{\gamma^h}$ defined by Eq. (3.35) is inlayed as follows:

$$\begin{aligned}&\mathcal{X}^{\pi(\gamma^h)}_{\gamma^h} \overset{\mathcal{I}}{\Rightarrow} \mathcal{X}^{\pi(\gamma^h)}(\gamma^h;\text{``ICS }\mathbf{g}\text{''}) \\ &\equiv \begin{cases} \Sigma\ \frac{1}{2} \cdot X^{1/2}_{\pm}(\gamma^h;\text{``ICS }\mathbf{g}\text{''}) \cdot X^{1/2}_{\mp}(\gamma^h;\text{``ICS }\mathbf{g}\text{''}) & \text{for } \pi(\gamma^h)=\text{NS}, \\ X^0(\gamma^h;\text{``ICS }\mathbf{g}\text{''}) \quad \text{for } \pi(\gamma^h)=\text{R}. \end{cases}\end{aligned} \tag{4.29}$$

In the case when three superstrings (say, the rth, sth, and tth superstring) interact with each other at the (upper and lower) ιth interacting points $[\text{for } \iota=1-(N+2\mathbf{g}-2)]$, the inverse picture-changing (external) operator $Y_{\pm\iota}$ is defined by

$$Y_{\pm\iota} \equiv Y_r(\pm\sqrt{-1}) \equiv Y_s(\pm\sqrt{-1}) \equiv Y_t(\pm\sqrt{-1}) \quad \text{for } \iota=1-(N+2\mathbf{g}-2). \tag{4.30}$$

At the ιth interacting point, there exists either $Y_{-\iota}\cdot Y_{+\iota}$ (in the case of interacting three strings NS–NS–NS) or $Y_{-\iota}$ (in the case of interacting three strings R–R–NS). This result will be expressed by

$$Y_{\text{``}\iota\text{''}} \equiv \begin{cases} Y_{-\iota}\cdot Y_{+\iota} & \text{for NS–NS–NS interacting point}, \\ Y_{-\iota} & \text{for R–R–S interacting point}. \end{cases} \tag{4.31}$$

The inverse picture-changing operator $Y_{\text{``}\iota\text{''}}$ would be inlayed as

$$\begin{aligned}&Y_{\text{``}\iota\text{''}} \overset{\mathcal{I}}{\Rightarrow} Y_{\text{``}\iota\text{''}}(\text{``ICS } g\text{''}) \quad \text{for } \iota=1-(N+2\mathbf{g}-2) \\ &\equiv \begin{cases} Y(\mathcal{Y}^*_{\iota\mathbf{g}})\cdot Y(\mathcal{Y}_{\iota\mathbf{g}}) & \text{for NS–NS–NS interacting point} \\ Y(\mathcal{Y}^*_{\iota\mathbf{g}}) & \text{for R–R–NS interacting point}, \end{cases}\end{aligned} \tag{4.32}$$

$\mathcal{Y}_{\iota\mathbf{g}}$ ($\mathcal{Y}^*_{\iota\mathbf{g}}$) being the inlayed coordinate [in the ''ICS **g**''] of the ιth upper (lower) interacting point. In much the same way as we have shown Eq. (6.21) in Ref. 5, we can show that

$$
\begin{aligned}
&(\mathcal{Q}=0|\xi_0)\cdot {}_s\langle IV(1,2,\dots,N+2\mathbf{g})|\cdot\left(\prod_{\iota=1}^{N+2\mathbf{g}-2} Y_{``\iota''}\right)\cdot\left(\prod_{I=1}^{N+2\mathbf{g}-3} \mathcal{X}_I^{\pi(I)}\cdot\mathcal{B}_{0,I}\right)\\
&\quad\times\left(\prod_{h=1}^{\mathbf{g}} \mathcal{X}_{\gamma^h}^{\pi(\gamma^h)}\cdot\mathcal{B}_{0,\gamma^h}\cdot\eta_{0,\gamma^h}\right)\cdot\left(\prod_{r=1}^{N}\mathcal{G}_r(0)\right)\cdot\left(\prod_{r=1}^{N}|\mathcal{Q}_r=0\rangle_r\right)\\
&\quad\times\prod_{h=1}^{\mathbf{g}}(\exp(-T_{\gamma^h}\cdot L_{0,\gamma^h})\cdot P_{\gamma^h}^{\pi(\gamma^h)}(\mathrm{GSO})|V_L(\gamma^h,\delta^h)\rangle)\cdot(|\mathcal{Q}\\
&=0\rangle)\\
&\quad=\mathop{\mathrm{Tr}}_{\varphi}\Bigg(\langle\Omega_{\mathbf{g}}((g_B,g_F))\rangle\cdot\left(\prod_{r=1}^{N}\mathcal{G}(r;``\mathrm{ICS}\ \mathbf{g}'')\right)\cdot\left(\prod_{\iota=1}^{N+2\mathbf{g}-2}Y_{``\iota''}(``\mathrm{ICS}\ \mathbf{g}'')\right)\\
&\quad\times\left(\prod_{I=1}^{N+2\mathbf{g}-3}\mathcal{X}^{\pi(I)}(I;``\mathrm{ICS}\ \mathbf{g}'')\cdot\mathcal{B}_0(I,``\mathrm{ICS}\ \mathbf{g}'')\right)\\
&\quad\times\left(\prod_{h=1}^{\mathbf{g}}\mathcal{X}^{\pi(\gamma^h)}(\gamma^h;``\mathrm{ICS}\ \mathbf{g}'')\cdot\mathcal{B}_0(\gamma^h;``\mathrm{ICS}\ \mathbf{g}'')\cdot\eta_0(\gamma^h;``\mathrm{ICS}\ \mathbf{g}'')\right)\Bigg),
\end{aligned}
\tag{4.33}
$$

where we have used the ''*analytic inlint gluing operator*'' $\langle\Omega_{\mathbf{g}}((g_B,g_F))\rangle$ defined by

$$
\begin{aligned}
\langle\Omega_{\mathbf{g}}((g_B,g_F))\rangle&\equiv\Omega_{\mathbf{g}}^N((g_B,g_F))\left(\prod_{r=1}^{N}|\mathcal{Q}_r=0\rangle_r\right)\\
&=(U_s|\mathcal{Q}=0\rangle\cdot\langle\mathcal{Q}=0|\xi_0\cdot U_s^{-1})\cdot\vec{R}\cdot\left(\prod_{h=1}^{\mathbf{g}}\prod_{\varphi}({}_{\delta^h}\langle q_{\delta^h}(\varphi)=0|W_{\delta^h}^{\varphi}[z_{\delta^h g}(w_{\delta^h})])\right)\\
&\quad\times\left(\prod_{h=1}^{\mathbf{g}}\prod_{\varphi}({}_{\gamma^h}\langle q_{\gamma^h}(\varphi)=0|W_{\gamma^h}^{\varphi}[z_{\gamma^h\mathbf{g}}(w_{\gamma^h})])\right)\\
&\quad\times\left(\prod_{h=1}^{\mathbf{g}}P_{\gamma^h}^{\pi(\gamma^h)}(\mathrm{GSO})|V_L(\gamma^h,\gamma^h)\rangle\right)\quad\text{with }\ \mathbf{g}=g_B+g_F
\end{aligned}
\tag{4.34}
$$

[cf. Eq. (6.22) in Ref. 5]. Similarly to Eq. (6.23) in Ref. 5, we can show that $\langle\Omega_{\mathbf{g}}((g_B,g_F))\rangle$ in (4.34) in the ''ICS $\mathbf{g}$'' satisfies the following *boundary conditions* for any *inlint* primary operator $\mathcal{G}(z')$ [of conformal weight $d(\mathcal{G})$]:

$$
\vec{R}\cdot\langle\Omega_{\mathbf{g}}(g_B,g_F)\rangle\cdot\mathcal{G}(z')=\left(\frac{dS_h(z')}{dz'}\right)^{d(\mathcal{G})}\cdot\vec{R}\cdot\langle\Omega_{\mathbf{g}}(g_B,g_F)\rangle\cdot\mathcal{G}(S_h(z'))
$$

$$
\text{for }\ h=1-\mathbf{g}\quad\text{and}\quad z'\in\partial F_{D\mathbf{g}}.
\tag{4.35}
$$

Comment: Equation (4.34) shows that $\langle\Omega_{\mathbf{g}}(g_B,g_F)\rangle$ is the function of only *inlint* operators [on the boundary $\partial F_{D\mathbf{g}}$ in the ''ICS $\mathbf{g}$'']. [5] It is important that the term on the left-hand side of (4.33) can be calculated by the term on the right-hand side of Eq. (4.33), the latter of which does *not* involve any *external* operator *but* involves only *inlint* operators having the argument in the ''ICS $\mathbf{g}$.'' We have proved[5] that the ''*analytic inlint gluing operator*'' $\langle\Omega_{\mathbf{g}}((g_B,g_F))\rangle$ can be expressed as the functional on inlint operators only on the boundary $\partial F_{D\mathbf{g}}$, so that the boundary conditions (4.35) are expected to determine $\langle\Omega_{\mathbf{g}}((g_B,g_F))\rangle$ up to the constant factor.

Using the method used in proposing Eqs. (6.27)–(6.32) in Ref. 5, we *propose* that

$$\langle \Omega_{\mathbf{g}}(g_B,g_F)\rangle = \left(\prod_{\varphi=\vec{X},\sigma,\chi} \langle \Omega_{\mathbf{g}}^{\varphi}(g_B,g_F)\rangle\right)\left(\prod_{\varphi=\phi^j,\phi} \langle \Omega_{\mathbf{g}}^{\varphi\{@\}}(g_B,g_F)\rangle\right) \times \prod_{h=1}^{\mathbf{g}} \left(\frac{1}{2}+\frac{1}{2}\varepsilon^{\pi(h)}\exp\left(\sqrt{-1}\,\pi\left(p_h(\phi)+\sum_{j=0}^{4} p_h(\phi^j)\right)\right)\right), \tag{4.36}$$

where the constant $\varepsilon^{\pi(h)}$ is the one given by Eq. (2.15). In Eq. (4.36), we have used the following formulas for the ''*analytic inlint gluing operator*'' $\langle \Omega_{\mathbf{g}}^{\varphi}((g_B,g_F))\rangle$ in each φ-mode. As for $\vec{X}$-modes, we propose that

$$\langle \Omega_{\mathbf{g}}^{\vec{X}}(g_B,g_F)\rangle = Z_1^{-5}(|p(\vec{X})=0\rangle\cdot\langle p(\vec{X})=0|)\cdot\left(\left(\prod_{h=1}^{\mathbf{g}}\int \frac{d^{10}k_h}{(2\pi)^{10}}\right) \times :\exp(C^{(2)}(\vec{X}))\cdot\exp\left(\sum_{i,h=1}^{\mathbf{g}}\frac{1}{2}\vec{k}_i\cdot C_{ih}^{(0)}\cdot\vec{k}_h+\sum_{h=1}^{\mathbf{g}}\vec{k}_h\cdot\vec{C}_h^{(1)}(\vec{X})\right):\right), \tag{4.37a}$$

with the inner products $\vec{k}_h\cdot\vec{C}_h$ and $\vec{k}_i\cdot\vec{k}_h$ defined respectively by

$$\begin{aligned}\vec{k}_h\cdot\vec{C}_h &\equiv -\sum_{\pm}\sum_{j=0}^{4} k_h^{\pm j}\cdot C_h^{\mp j} = -\sum_{\pm} k_h^{\pm 0}\cdot C_h^{\mp 0}+\sum_{j=1}^{8} k_h^{j}\cdot C_h^{j},\\ \vec{k}_i\cdot\vec{k}_h &\equiv -\sum_{\pm}\sum_{j=0}^{4} k_i^{\pm j}\cdot k_h^{\mp j} = -\sum_{\pm} k_i^{\pm 0}\cdot k_h^{\mp 0}+\sum_{j=1}^{8} k_i^{j}\cdot k_h^{j}.\end{aligned} \tag{4.37b}$$

[Hereafter, the ten-dimensional (space–time) vector space is denoted by V_{10}, so that $\vec{X}$, $\vec{k}_i$, $\vec{k}_h$, $\vec{C}_h\in V_{10}$ in Eqs. (4.37a) and (4.37b) and hereafter.] As for the $\vec{\phi}$-modes we propose that

$$\begin{aligned}\langle \Omega_{\mathbf{g}}^{\vec{\phi}\{@\}}(g_B,g_F)\rangle &\equiv \prod_{j=0}^{4}\langle \Omega_{\mathbf{g}}^{\phi^j\{@\}}(g_B,g_F)\rangle = Z_1^{-5/2}\left(\prod_{j=0}^{4}|p(\phi^j)=0\rangle\cdot\langle p(\phi^j)=0|\right)\\ &\times\left(\left(\prod_{h=1}^{\mathbf{g}}\prod_{j=0}^{4}\sum_{n_h(\phi^j)\in Z}\right):\exp\left(\sum_{j=0}^{4}C^{(2)}(\phi^j)\right)\cdot\exp\left(\sum_{j=0}^{4}\sum_{h=1}^{\mathbf{g}}\sqrt{-1}\,\pi\cdot n_h(\phi^j)\right)\right.\\ &\times\exp\left(\sum_{j=0}^{4}\sum_{i,h=1}^{\mathbf{g}}\frac{1}{2}\left(n_i(\phi^j)+\frac{1}{2}\delta_{\pi(i)}^{R}\right)C_{ih}^{(0)}\left(n_h(\phi^j)+\frac{1}{2}\delta_{\pi(h)}^{R}\right)\right)\\ &\times\left.\exp\left(\sum_{j=0}^{4}\sum_{h=1}^{\mathbf{g}}\left(n_h(\phi^j)+\frac{1}{2}\delta_{\pi(h)}^{R}\right)C_h^{(1)}(\phi^j)\right):\right),\end{aligned} \tag{4.38}$$

where the constant $\delta_{\pi(h)}^{R}$ is the one defined by

$$\delta_{\pi(h)}^{R}\equiv\begin{cases}1 & \text{for } \pi(h)=\mathrm{R},\\ 0 & \text{for } \pi(h)=\mathrm{NS}.\end{cases} \tag{4.39}$$

Furthermore, we propose the following formulas for $\varphi[=\sigma,\phi,\chi]$-modes;

$$\langle \Omega_{\mathbf{g}}^{\sigma}(g_B, g_F)\rangle = Z_1^{-1/2}(|p(\sigma)=0\rangle\cdot\langle p(\sigma)=0|\exp(+3\mathbf{g}\cdot q(\sigma))) \times\left(\left(\prod_{h=1}^{\mathbf{g}}\sum_{n_h(\sigma)\in Z}\right):\exp(C^{(2)}(\sigma))\cdot\exp\left(\sum_{h=1}^{\mathbf{g}}\sqrt{-1}\pi\cdot n_h(\sigma)\right)\right. \left.\times\exp\left(\sum_{i,h=1}^{\mathbf{g}}\frac{1}{2}n_i(\sigma)\cdot C_{ih}^{(0)}\cdot n_h(\sigma)+\sum_{h=1}^{\mathbf{g}}n_h(\sigma)\cdot C_h^{(1)}(\sigma)\right):\right), \tag{4.40}$$

$$\langle \Omega_{\mathbf{g}}^{\phi\{@\}}(g_B, g_F)\rangle = Z_1^{-1/2}(|p(\phi)=0\rangle\cdot\langle p(\phi)=0|\exp(-2\mathbf{g}\cdot q(\phi))) \times\left(\left(\prod_{h=1}^{\mathbf{g}}\sum_{n_h(\phi)\in Z}\right):\exp(C^{(2)}(\phi))\cdot\exp\left(\sum_{h=1}^{\mathbf{g}}\sqrt{-1}\pi\cdot n_h(\phi)\right)\right. \times\exp\left(\sum_{i,h=1}^{\mathbf{g}}\frac{1}{2}\left(n_i(\phi)-\frac{1}{2}\delta_{\pi(i)}^R\right)(-C_{ih}^{(0)})\left(n_h(\phi)-\frac{1}{2}\delta_{\pi(h)}^R\right)\right) \left.\times\exp\left(\sum_{h=1}^{\mathbf{g}}\left(n_h(\phi)-\frac{1}{2}\delta_{\pi(h)}^R\right)C_h^{(1)}(\phi)\right):\right) \text{with } \delta_{\pi(h)}^R \text{ being given by Eq. (4.39)}, \tag{4.41}$$

and

$$\langle \Omega_{\mathbf{g}}^{\chi}(g_B, g_F)\rangle = Z_1^{-1/2}(|p(\chi)=0\rangle\cdot\langle p(\chi)=0|\exp((\mathbf{g}+1)\cdot q(\chi))) \times\left(\left(\prod_{h=1}^{\mathbf{g}}\sum_{n_h(\chi)\in Z}\right):\exp(C^{(2)}(\chi))\cdot\exp\left(\sum_{h=1}^{\mathbf{g}}\sqrt{-1}\pi\cdot n_h(\chi)\right)\right. \left.\times\exp\left(\sum_{i,h=1}^{\mathbf{g}}\frac{1}{2}n_i(\chi)\cdot C_{ih}^{(0)}\cdot n_h(\chi)\right)\exp\left(\sum_{h=1}^{\mathbf{g}}n_h(\chi)\cdot C_h^{(1)}(\chi)\right):\right), \tag{4.42}$$

where we have used the bra state $\langle p(\chi)=0|$ satisfying Eqs. (A34)–(A35b). [Incidentally, $\exp((\mathbf{g}+1)\cdot q(\chi))$ in Eq. (4.42) has been obtained by $\exp(\mathbf{q}\cdot q(\chi))\cdot\exp(q(\chi))$, the latter of which comes from ξ_0 in Eq. (4.22).] In Eqs. (4.38) and (4.40)–(4.42), sums of $n_h(\varphi)$ [for $h=1-\mathbf{g}$ and $\varphi=\phi^j,\ \sigma,\ \phi,\ \chi$] should be taken over all *integers*. Provided that we choose cocycle factors[15] properly, we can consider in Eq. (4.38) [(4.41)] the case when $\exp(+\frac{1}{2}\phi^j)$ for $j=0-4$ [exp $(-\frac{1}{2}\phi)$] is Grassman even (even). Therefore, we have assured in Eqs. (4.38) and (4.40)–(4.42) that the *fermionic* loops have the sign opposite to the bosonic ones by inserting a sign factor $\exp(\Sigma_{h=1}^{\mathbf{g}}\sqrt{-1}\pi\cdot n_h(\varphi))$ (for $\varphi=\phi^j,\ \sigma,\ \phi,\ \chi$). [Incidentally, the proper cocycle factor gives Grassman even (odd) $\exp(-\frac{1}{2}\phi)$, which has been called the fracton (fractino).[3] Type II A superstring theory uses the fracton (fractino) in the left (right) moving modes of closed superstrings.]

In the formulas (4.37)–(4.42), $C_{ih}^{(0)}$ is given by the periodic matrix τ_{ih} as

$$C_{ih}^{(0)} \equiv 2\pi\sqrt{-1}\cdot\tau_{ih} \quad (i,h=1-\mathbf{g}), \tag{4.43}$$

while $C_h^{(1)}$ ($C^{(2)}$) for each φ-mode is expressed as the function of linear (quadratic) inlint operators on the boundary $\partial F_{D\mathbf{g}}$ [which is defined by Eqs. (6.12)–(6.14) in Ref. 5] as follows. As for $C_h^{(1)}(\varphi)$'s, we have that

$$\frac{\vec{C}_h^{(1)}(\vec{X})}{2\pi\sqrt{-1}} \equiv \oint_{\partial F_{D\mathbf{g}}}\frac{dz}{2\pi\sqrt{-1}}\sqrt{-1}\partial_z\vec{X}(z)\int_{z_0}^{z}dz'\frac{\omega_h(z')}{2\pi\sqrt{-1}} \quad (\in V_{10}) \tag{4.44}$$

and

$$\frac{C_h^{(1)}(\varphi)}{2\pi\sqrt{-1}} \equiv \oint_{\partial F_{D\mathbf{g}}} \frac{dz}{2\pi\sqrt{-1}} \partial_z \varphi(z) \int_{z_0}^{z} dz' \frac{\omega_h(z')}{2\pi\sqrt{-1}} - \varepsilon_\varphi^\phi Q(\varphi)\left(\Delta_h^{z_0} + \frac{1}{2}\right)$$

$$\text{for } \varphi = \phi^j, \sigma, \phi, \chi, \quad \text{with } \varepsilon_\varphi^\phi \text{ in (A5b)} \quad \text{and } Q(\varphi) \text{ in (2.5)}, \tag{4.45}$$

while we obtain for $C^{(2)}(\varphi)$'s that

$$C^{(2)}(\vec{X}) \equiv \frac{1}{2} \oint_{\partial F_{D\mathbf{g}}} \frac{dz}{2\pi\sqrt{-1}} \oint_{\partial F_{D\mathbf{g}}} \frac{dz'}{2\pi\sqrt{-1}} \sqrt{-1}\,\partial_z \vec{X}(z) \cdot \sqrt{-1}\,\partial_{z'} \vec{X}(z') \cdot \log\left(\frac{E(z,z')}{z-z'}\right), \tag{4.46}$$

and

$$C^{(2)}(\varphi) \equiv \varepsilon_\varphi^\phi \frac{1}{2} \oint_{\partial F_{D\mathbf{g}}} \frac{dz}{2\pi\sqrt{-1}} \partial_z \varphi(z) \oint_{\partial F_{D\mathbf{g}}} \frac{dz' 1}{2\pi\sqrt{-1}} \partial_{z'} \varphi(z') \log\left(\frac{E(z,z')}{z-z'}\right)$$
$$+ Q(\varphi) \oint_{\partial F_{D\mathbf{g}}} \frac{dz}{2\pi\sqrt{-1}} \partial_z \varphi(z) \log(\sigma(z))$$

$$\text{for } \varphi = \phi^j, \sigma, \phi, \chi, \quad \text{with } \varepsilon_\varphi^\phi \text{ in (A5b)} \quad \text{and } Q(\varphi) \text{ in (2.5)}. \tag{4.47}$$

It should be noticed that the integrations in Eqs. (4.44)–(4.47) have been carried out along the boundaries $\partial F_{D\mathbf{g}}$. [In Eq. (4.47), the sigma function $\sigma(z)$ in $\log(\sigma(z))$ is the one used in $\nu_{\mathbf{g}}(z)$ in (4.10), and it should *not* be confused with the ghosting operator σ in (A3).] The determinantal function Z_1 in the formulas (4.37)–(4.42) does *not* depend on inlint operators *but* depends only on modular parameters S_h's (for $h=1-\mathbf{g}$) used in Eq. (4.8),[19] and it has been calculated by Verlinde–Verlinde [and their result[21] will be given later by Eq. (4.66)].

For the general $z_r(\in \mathbb{C})$ *within* the fundamental region $F_{D\mathbf{g}}$ in the ''ICS **g**'' (i.e., for $z \in F_{D\mathbf{g}}$), the analytic inlint gluing operator $\langle \Omega_{\mathbf{g}}(g_B, g_F)\rangle$ in (4.36)–(4.42) gives following trace-formulas in **g**-loops. For ϕ^j's modes, we have that

$$\mathrm{Tr}_{\phi^j} \vec{R} \cdot \left(\prod_{r=1}^{N} :\exp(s_r^j \cdot \phi^j(z_r)): \langle \Omega_{\mathbf{g}}^{\phi^j\{@\}}(g_B, g_F)\rangle \right) \quad \text{for } j=0-4$$
$$= Z_1^{-1/2} \delta\left(\sum_{r=1}^{N} s_r^j\right) \exp\left(-\sum_{h=1}^{\mathbf{g}} \frac{\pi}{2}\sqrt{-1}\cdot \delta_{\pi(h)}^R\right)$$
$$\times \Theta_+^{\{@\}}\left(\sum_{r=1}^{N} s_r^j \int_{z_0}^{z_r} dz' \frac{\vec{\omega}(z')}{2\pi\sqrt{-1}} + \frac{1}{2}\cdot \vec{1} \,\middle|\, + \tau\right) \cdot \left(\prod_{\substack{r,t=1\\ r(<)t}}^{N} E(z_r, z_t)_R^{s_r^j \cdot s_t^j}\right), \tag{4.48}$$

where $\vec{\omega}$, $\vec{1} \in V_{\mathbf{g}}$. Furthermore, we also have the following trace-formulas for $\varphi[=\sigma,\phi,\chi]$-modes:

$$\mathrm{Tr}_\sigma \vec{R}\cdot\left(\prod_{r=1}^{N} :\exp(g_r\cdot\sigma(z_r)):\langle\Omega_{\mathbf{g}}^{\sigma}(g_B,g_F)\rangle\right)$$
$$=Z_1^{-1/2}\delta\left(\sum_{r=1}^{N} g_r+3\mathbf{g}-3\right)\cdot\Theta\left(\sum_{r=1}^{N} g_r\cdot\int_{z_0}^{z_r}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}+3\vec{\Delta}^{z_0}\middle|+\tau\right)$$
$$\times\left(\prod_{\substack{r,t=1\\ r(<)t}}^{N} E(z_r,z_t)_R^{g_r\cdot g_t}\right)\left(\prod_{r=1}^{N}\sigma(z_r)^{-3g_r}\right), \tag{4.49}$$

$$\mathrm{Tr}_\phi \vec{R}\cdot\left(\prod_{r=1}^{N} :\exp(f_r\cdot\phi(z_r)):\langle\Omega_{\mathbf{g}}^{\phi\{@\}}(g_B,g_F)\rangle\right)$$
$$=Z_1^{-1/2}\delta\left(\sum_{r=1}^{N} f_r-2\mathbf{g}+2\right)\cdot\exp\left(\sum_{h=1}^{\mathbf{g}}\frac{\pi}{2}\sqrt{-1}\cdot\delta^R_{\pi(h)}\right)$$
$$\times\Theta_-^{\{@\}}\left(-\sum_{r=1}^{N} f_r\int_{z_0}^{z_r}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}+2\vec{\Delta}^{z_0}+\frac{3}{2}\cdot\vec{1}\middle|-\tau\right)$$
$$\times\left(\prod_{\substack{r,t=1\\ r(<)t}}^{N} E(z_r,z_t)_R^{-f_r\cdot f_t}\right)\left(\prod_{r=1}^{N}\sigma(z_r)^{-2f_r}\right), \tag{4.50}$$

and

$$\mathrm{Tr}_\chi \vec{R}\cdot\left(\prod_{r=1}^{N} :\exp(h_r\cdot\chi(z_r)):\langle\Omega_{\mathbf{g}}^{\chi}(g_B,g_F)\rangle\right)$$
$$=Z_1^{-1/2}\delta\left(\sum_{r=1}^{N} h_r+\mathbf{g}\right)\cdot\Theta\left(\sum_{r=1}^{N} h_r\int_{z_0}^{z_r}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}+\vec{\Delta}^{z_0}\middle|+\tau\right)$$
$$\times\left(\prod_{\substack{r,t=1\\ r(<)t}}^{N} E(z_r,z_t)_R^{h_r\cdot h_t}\right)\left(\prod_{r=1}^{N}\sigma(z_r)^{-h_r}\right). \tag{4.51}$$

[Incidentally, we have $\vec{\omega}$, $\vec{\Delta}$, $\vec{1}\in V_{\mathbf{g}}$ in Eqs. (4.49)–(4.51) and hereafter.] In the trace formulas (4.48)–(4.51) [in $\mathbf{g}$-loops], we have used ''radial-ordered arguments'' [denoted by $(\dots)_R$] defined by Eq. (A12). In Eqs. (4.48) and (4.50), $\Theta_\pm^{\{@\}}(\vec{\xi}|\pm\tau)$ is defined by

$$\Theta_\pm^{\{@\}}(\vec{\xi}|\pm\tau)\equiv\Theta\begin{pmatrix}\pm\frac{1}{2}\vec{\delta}^{\{@\}}\\ \vec{0}\end{pmatrix}(\vec{\xi}|\pm\tau), \tag{4.52a}$$

where and hereafter we use the following shorthand notations:

$$\pm\frac{1}{2}\vec{\delta}^{@}=\left(\pm\frac{1}{2}\delta^R_{\pi(1)},\pm\frac{1}{2}\delta^R_{\pi(2)},\dots,\pm\frac{1}{2}\delta^R_{\pi(\mathbf{g})}\right)\quad(\in V_{\mathbf{g}})$$
$$\vec{0}=(0,0,...,0)\quad(\in V_{\mathbf{g}}),\quad \vec{1}=(1,1,...,1)\quad(\in V_{\mathbf{g}}). \tag{4.52b}$$

In order to take account of the last GSO-projection factor in Eq. (4.36), we notice that $\exp(\sqrt{-1}\cdot p_\rho)$ operates as

$$\exp(\sqrt{-1}\pi\cdot p_\rho)\cdot\Theta\begin{pmatrix}\vec{a}\\ \vec{0}\end{pmatrix}(\vec{\xi}|\pm\tau)\quad(\text{for}\ \ \rho=1-\mathbf{g})$$

$$\equiv\sum_{\{n_h\}\in Z}\exp(\sqrt{-1}\pi(n_\rho+a_\rho))$$

$$\times\exp\left(2\pi\sqrt{-1}\left(\sum_{i,h=1}^{\mathbf{g}}\frac{1}{2}(n_i+a_i)(\pm\tau_{ih})(n_h+a_h)+\sum_{h=1}^{\mathbf{g}}(n_h+a_h)\xi_h\right)\right),$$

$$=\Theta\begin{pmatrix}\vec{a}\\ \vec{0}\end{pmatrix}(\vec{\hat{\xi}}(\rho)|\tau),\tag{4.53}$$

where the following shorthand notation has been used:

$$\vec{\hat{\xi}}(\rho)\equiv(\xi_1,\ldots,\xi_{\rho-1},\xi_\rho+\tfrac{1}{2},\xi_{\rho+1},\ldots,\xi_{\mathbf{g}})\quad(\in V_{\mathbf{g}}).\tag{4.54}$$

Then, we find from Eq. (4.53) that

$$\exp\left(\sqrt{-1}\pi\left(p_\rho(\phi)+\sum_{j=0}^{4}p_\rho(\phi^j)\right)\right)\Theta_-^{\{@\}}(\vec{\xi}|-\tau)\left(\prod_{j=0}^{4}\Theta_+^{\{@\}}(\vec{\xi}^j|+\tau)\right)$$

$$=\Theta_-^{\{@\}}(\vec{\hat{\xi}}(\rho)|-\tau)\cdot\left(\prod_{j=0}^{4}\Theta_+^{\{@\}}(\vec{\hat{\xi}}^j(\rho)|+\tau)\right)\quad\text{for}\ \ \rho=1-\mathbf{g}.\tag{4.55}$$

For any $z_r\in F_{D\mathbf{g}}$, we can similarly derive the following trace-formula [for $\vec{X}$-modes]:

$$\mathrm{Tr}_{\vec{X}}\vec{R}\cdot\left(:\exp\left(\sum_{r=1}^{N}\sqrt{-1}\cdot\vec{p}_r\cdot\vec{X}(z_r)\right):\langle\Omega_{\mathbf{g}}^{\vec{X}}(g_B,g_F)\rangle\right)$$

$$=Z_1^{-5}(2\pi)^{10}\cdot\delta^{10}\left(\sum_{r=1}^{N}\vec{p}_r\right)\left(\left(\prod_{h=1}^{\mathbf{g}}\int\frac{d^{10}k_h}{(2\pi)^{10}}\right)\right.$$

$$\left.\times\exp\left(\sum_{i,h=1}^{\mathbf{g}}\pi\sqrt{-1}\cdot\vec{k}^i\cdot\tau_{ih}\cdot\vec{k}^h\right)\cdot\langle\mathrm{mod};\vec{k}|:\exp\left(\sum_{r=1}^{N}\sqrt{-1}\cdot\vec{p}_r\cdot\vec{X}(z_r)\right):|\mathrm{mod};\vec{k}\rangle\right).\tag{4.56}$$

In Eq. (4.56), both i and h (for $i,h=1-\mathbf{g}$) denote the labels of the ''glued lines'' in Eq. (4.1). Furthermore, we have used in Eq. (4.56) the following *function* $\langle\mathrm{mod};\vec{k}|\ldots|\mathrm{mod};\vec{k}\rangle$ defined by

$$\langle\mathrm{mod};\vec{k}|:\exp\left(\sum_{r=1}^{N}\sqrt{-1}\cdot\vec{p}_r\cdot\vec{X}(z_r)\right):|\mathrm{mod};\vec{k}\rangle$$

$$\equiv\exp\left(\sum_{r=1}^{N}\sum_{h=1}^{\mathbf{g}}\vec{p}_r\cdot\vec{k}_h\int_{z_0}^{z_r}dz'\ \omega_h(z')+\sum_{\substack{r,t=1\\ r(<)t}}^{N}\vec{p}_r\cdot\vec{p}_t\log\left(\frac{E(z_r,z_t)}{z_r-z_t}\right)\right),\tag{4.57}$$

which leads to the following expectation value of the *overally radial-ordered product*:

$$\langle \mathrm{mod};\vec{k}|\vec{R}\cdot\left(\prod_{r=1}^{N} :\exp(\sqrt{-1}\cdot\vec{p}_r\cdot\vec{X}(z_r)):\right)|\mathrm{mod};\vec{k}\rangle$$
$$\equiv\exp\left(\sum_{r=1}^{N}\sum_{h=1}^{\mathbf{g}}\vec{p}_r\cdot\vec{k}_h\int_{z_0}^{z_r}dz'\,\omega_h(z')\right)\times\left(\prod_{\substack{r,t=1\\ r(<)t}}^{N}E(z_r,z_t)_R^{\vec{p}_r\cdot\vec{p}_t}\right). \tag{4.58}$$

[We notice that $\vec{X},\vec{p}_r,\vec{p}_t,\vec{k}_h\in V_{10}$ in Eqs. (4.56)–(4.58) and hereafter.] Similarly, the trace-formula for $\vec{\phi}$-modes are given as follows:

$$\mathrm{Tr}_{\vec{\phi}}\vec{R}\cdot\left(:\exp\left(\sum_{r=1}^{N}\sum_{j=0}^{4}s_r^j\cdot\phi^j(z_r)\right):\langle\Omega_{\mathbf{g}}^{\vec{\phi}\{@\}}(g_B,g_F)\rangle\right)$$
$$=Z_1^{-5/2}\cdot\delta_K^5\left(\sum_{r=1}^{N}\vec{s}_r\right)\left(\prod_{h=1}^{\mathbf{g}}\sum_{\vec{n}_h}\right)\left(\exp\left(\sum_{i,h=1}^{\mathbf{g}}\pi\sqrt{-1}\cdot\vec{n}_i^{\pi(i)}\cdot\tau_{ih}\cdot\vec{n}_h^{\pi(h)}\right)\right.$$
$$\left.\times\exp\left(\sum_{j=0}^{4}\sum_{h=1}^{\mathbf{g}}\sqrt{-1}\pi\cdot n_h(\phi^j)\right)\cdot\langle\mathrm{mod};\vec{n}|:\exp\left(\sum_{r=1}^{N}\sum_{j=0}^{4}s_r^j\cdot\phi^j(z_r)\right):|\mathrm{mod};\vec{n}\rangle\right), \tag{4.59}$$

where we have introduced the following *function* $\langle\mathrm{mod};\vec{n}|\ldots|\mathrm{mod};\vec{n}\rangle$ defined by

$$\langle\mathrm{mod};\vec{n}|:\exp\left(\sum_{r=1}^{N}\sum_{j=0}^{4}s_r^j\cdot\phi^j(z_r)\right):|\mathrm{mod};\vec{n}\rangle$$
$$\equiv\exp\left(\sum_{r=1}^{N}\sum_{h=1}^{\mathbf{g}}\vec{n}_h^{\pi(h)}\cdot\vec{s}_r\int_{z_0}^{z_r}dz'\,\omega_h(z')+\sum_{\substack{r,t=1\\ r(<)t}}^{N}\vec{s}_r\cdot\vec{s}_t\cdot\log\left(\frac{E(z_r,z_t)}{z_r-z_t}\right)\right), \tag{4.60}$$

inner products $\vec{n}_h^{\pi(h)}\cdot\vec{s}_r$ and $\vec{s}_r\cdot\vec{s}_t$ being defined respectively by

$$\vec{n}_h^{\pi(h)}\cdot\vec{s}_r\equiv\sum_{j=0}^{4}(\vec{n}_h^{\pi(h)})^j\cdot s_r^j=\sum_{j=0}^{4}\left(n_h(\phi^j)+\frac{1}{2}\delta_{\pi(h)}^R\right)\cdot s_r^j\quad\text{for }\ n_h(\phi^j)\in Z,$$

and

$$\vec{s}_r\cdot\vec{s}_t\equiv\sum_{j=0}^{4}s_r^j\cdot s_t^j. \tag{4.61}$$

[The sum over $\vec{n}_h$ in Eq. (4.59) means the summation over all integers $n_h(\phi_j)$'s.] Therefore, we find from Eq. (4.60) that the expectation value of the *overally radial-ordered product* is given by the following formula:

$$\langle\mathrm{mod};\vec{n}|\vec{R}\cdot\left(\sum_{r=1}^{N}:\exp\left(\sum_{j=0}^{4}s_r^j\cdot\phi^j(z_r)\right):\right)|\mathrm{mod};\vec{n}\rangle$$
$$\equiv\exp\left(\sum_{r=1}^{N}\sum_{h=1}^{\mathbf{g}}\vec{n}_h^{\pi(h)}\cdot\vec{s}_r\int_{z_0}^{z_r}dz'\,\omega_h(z')\right)\cdot\left(\prod_{\substack{r,t=1\\ r(<)t}}^{N}E(z_r,z_t)_R^{\vec{s}_r\cdot\vec{s}_t}\right). \tag{4.62}$$

[The five-dimensional vector space will be denoted by V_5, so that we have $\vec{\phi},\vec{n}_h,\vec{s}_r,\vec{s}_t\in V_5$ in Eqs. (4.59)–(4.62) and hereafter.]

The total factor $Z_1^{-9}=Z_1^{-5}\cdot Z_1^{-5/2}\cdot(Z_1^{-1/2})^3$ [contributed from Eqs. (4.56), (4.59), and (4.49)–(4.51)] is quite important, since it gives the complicate integration measure factor in the following expression:

$$(N_B,N_F)_{\mathbf{g}}\equiv {}^{\mathcal{M}}\langle S_{\mathbf{g}}^{\mathcal{C}}[p(1),\ldots,p(N)]|\cdot\left(\prod_{r=1}^{N}\mathcal{G}_r(0)\right)\cdot\left(\prod_{r=1}^{N}|\mathcal{Q}_r=0\rangle_r\right)$$
$$=\sum_{\mathrm{SSD}}C_{\mathrm{SSD}}\left(\int_0^\infty\prod_{I=1}^{N+2\mathbf{g}-3}dT_I\right)\left(\int_0^\infty\prod_{h=1}^{\mathbf{g}}dT_{\gamma^h}\right)F(\{T_I\},\{T_{\gamma^h}\}), \tag{4.63}$$

where the summation is over all superstring (Feynman) diagrams (SSDs), C_{SSD}'s being constants determined from the Wick's expansion (3.50). On the other hand, the integrand $F(\{T_I\},\{T_{\gamma^h}\})$ is the one given by the term on the right-hand side of Eq. (4.33), $\mathcal{G}(r;\text{``ICS}\,\mathbf{g}\text{''})$'s being given as follows. In calculating $(N_B,N_F)_{\mathbf{g}}$ in (4.63) with the help of the trace-formulas (4.48)–(4.62), we use the *physical* NS state in 1 picture which is given by Eq. (D5) with (D1), as well as the *physical* R state in $\frac{1}{2}$ picture which is given by Eq. (D12). Therefore, $\mathcal{G}_r(0)$'s and $\mathcal{G}(r;\text{``ICS}\,\mathbf{g}\text{''})$'s are those given as follows. In Eq. (4.33), we use that

$$\mathcal{G}_r(0)\equiv\begin{cases}X_{\pm r}^{1/2}:\exp(+\sigma_r(0))::\exp(-\phi_r(0)):V_r^{0b}(0) & \text{for}\ \ \pi(r)=\mathrm{NS},\\ :\exp(+\sigma_r(0))::\exp(-\tfrac{1}{2}\phi_r(0)):V_r^{1f}(0)\cdot S_{\{h\},r}(0) & \text{for}\ \ \pi(r)=\mathrm{R}\end{cases} \tag{4.64}$$

and

$$\mathcal{G}(r;\text{``ICS}\ \ \mathbf{g}\text{''})$$
$$\equiv\begin{cases}X_{\pm}^{1/2}(r;\text{``ICS}\ \ \mathbf{g}\text{''}):\exp(+\sigma(Z_{r\mathbf{g}}))::\exp(-\phi(Z_{r\mathbf{g}})):V^{0b}(Z_{r\mathbf{g}}) & \text{for}\ \ \pi(r)=\mathrm{NS}\\ :\exp(+\sigma(Z_{r\mathbf{g}}))::\exp(-\tfrac{1}{2}\phi(Z_{r\mathbf{g}})):V^{1f}(Z_{r\mathbf{g}})\cdot S_{\{h\}}(Z_{r\mathbf{g}}) & \text{for}\ \ \pi(r)=\mathrm{R}.\end{cases} \tag{4.65}$$

[All operators in Eq. (4.65) are *inlint* operators, $Z_{r\mathbf{g}}$'s being punctures in Samuel $\mathcal{FE}$ (4.7).] Verlinde and Verlinde[21] has calculated the determinatal function Z_1 (up to a pure numerical constant C) and their result is given by

$$Z_1^{3/2}=C\cdot\frac{\Theta\left(\sum_{t=1}^{\mathbf{g}}\int_{z_0}^{z_t}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}-\int_{z_0}^{Z}dz'\frac{\vec{\omega}(z')}{2\pi\sqrt{-1}}-\vec{\Delta}^{z_0}\middle|\tau\right)}{(\det_{\mathbf{g}\times\mathbf{g}}\omega_s(z_t))\left(\prod_{t=1}^{\mathbf{g}}E(z_t,Z)\right)\cdot\sigma(Z)}\left(\prod_{1\leqslant t<s\leqslant g}E(z_t,z_s)\right)\left(\prod_{t=1}^{\mathbf{g}}\sigma(z_t)\right). \tag{4.66}$$

Samuel has explicitly shown[19] that the determinantal function (4.66) *actually* depends *only* on $(3\mathbf{g}-3)$ modular parameters [in $\mathrm{SL}(2R)$ group elements $(S_1,S_2,\ldots,S_{\mathbf{g}})$], which are those used in S-$\mathcal{FE}$ (4.7).

V. CONCLUSIONS AND DISCUSSIONS

In this paper, we have formulated (BRST-*invariant*) open quantum superstring field theory in the "$\mathcal{B}_0$-gauge." Since we use only *nonlocal* (i.e., we do *not* use *local*) picture-changing operators in our theory, there does not exist any singularity caused by colliding[10] of *local* picture-changing operators. Our QSFT has been analyzed by applying the method[5] which has been successfully used in calculating physical scattering amplitudes in open quantum bosonic string field theory (open QBFT) in the "$\mathcal{B}_0=0$ gauge." Hereafter, we summarize our main results obtained in this paper. In Sec. I, open QSFT in the "$\mathcal{B}_0$-gauge" is described by using the open NS (R) fieldino (1.3) in picture 1 ($\frac{1}{2}$). In the beginning, open QSFT is proposed to be given by the gauge-fixed action $S_{\mathrm{GF}}(\Psi)$" in (1.18) with the kinetic term $S_{\mathrm{KIN}}(\Psi)$ in (1.19) and the interaction term

$S_{\rm INT}(\Psi)$ in (1.20), which has been given by using NS (R) fieldinos $|\Psi(1)\rangle$'s in picture 1 [$|\Psi(\frac{1}{2})\rangle$'s in picture$\frac{1}{2}$] defined by Eq. (1.3). Bose fields ϕ_B's [in the fieldinos $|\Psi^{\rm NS}\rangle$'s in (1.1)] and fermi fields ψ_F's [in the fieldinos $|\Psi^R\rangle$'s in (1.2)] are *quantized* by imposing the operator product expansion (1.25) among fieldinos, where the contraction among open NS (R) fieldinos is given by the propagator (1.26) [(1.27)] of open NS (R) superstring. The propagator (1.26) involves the symmetrization among $X^{1/2}_{+,\gamma}\cdot X^{1/2}_{-,\gamma}$ and $X^{1/2}_{-,\gamma}\cdot X^{1/2}_{+,\gamma}$, so that it can be *rewritten* [with the help of the "commutability" (C39)] into the expression (3.1). Physical scattering amplitudes among N_B bosonic (N_F fermionic) particles are proposed to be given by the formula (1.22), which corresponds to the LSZ reduction formula in QFT.[16] [Throughout our analyses on open QSFT in the "$\mathcal{B}_0$-gauge," we have used the ghosting number G, the fracting number F, the hilberting number H, and the picturing number P defined by Eqs. (1.28)−(1.31).] In Sec. II, GM-$\mathcal{FE}$ (2.1) is used for the purpose of defining the "inlayed coordinate system s(*phere*) (ICS s) (with the help of which we can calculate physical scattering amplitudes in 0-loop). We have defined the rth disk coordinate w_r in (2.2) and used the rth inlayed coordinate[5] $z_{rs}(w_r)$ (to be used in the r_{th} punctured ring domain in the "ICS s"). We construct the r_{th} inlaying operator $W_r[z_{rs}(w_r)]$ in (2.6) (for $r=1-N$) by the formulas (2.3)–(2.5), which is the functional of the r_{th} *external* operators $\varphi_r(w_r;+)$'s, as well as the *inlint* operators $\varphi(z_{rs}(w_r))$ (within the r_{th} punctured ring domain). Then, the inlaying N-vertex functions ${}_s\langle IV(1,2,...,N)|$ (in the "ICS s") is constructed by the formulas (2.7)–(2.9). Furthermore, we find the inlaying identities (2.10) [which are simply expressed by Eq. (2.11)]. Finally, the N-vertex functino ${}_s\langle V(1,2,...,N)|$ (in the "ICS s") is constructed by the formula (2.13), where we have used the *inlint* zero-mode ξ_0. In the special case when the "ICS s" is just equal to the "inlayed coordinate system m(*idpoint*)" (ICS m) (where there does not exist any *un*punctured ring domain), the *elementary* N-vertex functino $\langle \nu_{\mathcal{S}}(1,2,...,N)|$ is given by the *cycle-symmetric* ${}_m\langle V(1,2,...,N)|$ in (2.14). [*Elementary* vertex functinos are to be used in the action (1.18)–(1.20), and it can be explicitly constructed by using the inlayed coordinates given by Eq. (E34).] The formula (2.16) shows that the *small* gluing vertex function $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ in (2.16) (which is *antisymmetric* under $\gamma\leftrightarrow\delta$) can be obtained from the *large* gluing vertex function $|V_L(\gamma,\delta)\rangle$ in (2.17)–(2.19). (The word "*small*" means "being without ξ_0-mode.") Then, we can prove the gluing theorem (2.20) [which is useful in calculating tree amplitudes], with the help of the similar techniques[5] to those used in open QBFT. In Sec III, we use that the contraction (1.26) [(1.27)] among open NS (R) fieldinos is just equal to Eq. (3.1) [(3.2)], which is referred to as the propagator of the open NS (R) superstring. With the help of Wick's theorem,[16] we have given the operator product expansion (3.3) among fieldinos, where (un)modified ${}^{(\mathcal{M})}\langle S\cdots]|$'s are terms obtained by various contractions which can be represented by *superstring* (Feynman) *diagrams*, so that they do *not* involve any quantized bose field ϕ_B as well as any fermi field ψ_F in ten-dimensional space–time. The *connected* part of ${}^{(\mathcal{M})}\langle S_{\mathbf{g}}[\cdots]|$ is denoted by ${}^{(\mathcal{M})}\langle S^{\mathcal{C}}_{\mathbf{g}}[\cdots]|$. [See Eqs. (3.3a) and (3.3b).] Unfortunately, the nonvanishing term (3.18) shows that the interaction term $S_{\rm INT}(\Psi)$ in (1.20) does not give the BRST-invariant amplitude, since the term (3.18) is nonvanishing. In order to obtain the BRST-invariant part out from $\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\ldots,p(4)]|$, we must subtract the proper counter term $\langle \Delta S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\ldots,p(4)]|$ which can be determined by Eq. (3.18b). Thus obtained result ${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\ldots,p(4)]|$ in (3.19) is shown to be BRST invariant. [See Eq. (3.20).] The subtracting process in Eq. (3.19) can be induced by the *effect* due to the counter term in the modified action $S^{\mathcal{M}}_{\rm INT}](4,0)$ in (3.21). We have proved that these procedures in obtaining the BRST-*invariant* amplitude ${}^{(\mathcal{M})}\times\langle S^{\mathcal{C}}_{\mathbf{g}=0}[p(1),\ldots,p(4)]|$ can be carried out to any order of perturbation. Our analyses and results on these procedures have been finally summarized in the *Theorem* [i.e., Eq. (3.56)]. In particular, the counter term $\Delta S_{\rm INT}(\Psi;n)$ in $S^{\mathcal{M}}_{\rm INT}(\Psi)$ in (3.55) is determined by solving Eq. (3.53). We have found that the "*amputated N-scatts* in **g**-loops" ${}^{\mathcal{M}}\langle S^{\mathcal{C}}_{\mathbf{g}}[p(1),\ldots,p(N)]|$ in Eq. (3.50) can be calculated by using unmodified interaction term $S_{\rm INT}(\Psi)$ in (1.20), provided that any propagator is *effectively* modified accordingly to Eq. (3.57). In Sec. IV, the *physical* scattering amplitude $(N_B,N_F)_{\mathbf{g}}$ in (3.60) among N_B bosonic (N_F fermionic) particles have been calculated by using the method developed in our previous paper on QBFT. First we use the S-$\mathcal{FE}$ (4.7) to describe the "inlayed coordinate system in *genus* **g**" (ICS **g**). The function $\nu_{\mathbf{g}}(z_{\mathbf{g}})$ in (4.10) can be chosen to satisfy the **g**-*conditions* (4.8) by imposing **g**-*constraints* (4.11) among modular parameters in the function $\nu_{\mathbf{g}}(z_{\mathbf{g}})$ in $\mathcal{FE}$ (4.7). Then S-$\mathcal{FE}$ (4.7) in the ''ICS **g**'' can describe the Riemann surface $\mathcal{R}'$ having $B=\mathbf{g}+1$ boundaries. Then, we have introduced the ''*analytic inlint gluing operator*'' $\langle\Omega_{\mathbf{g}}(g_B,g_F)\rangle$ in (4.34), with the help of which the term on the left-hand side of Eq. (4.33) can be calculated by the *trace* on the right-hand side of Eq. (4.33). Thus $(N_B,N_F)_{\mathbf{g}}$ in (3.62) can be calculated, since the term on the right-hand side of Eq. (4.33) [substituted with inlint operators (4.65)] can be calculated by using formulas (4.36)–(4.47) for $\langle\Omega_{\mathbf{g}}(g_B,g_F)\rangle$ in (4.34). [The trace in these calculations can be found by using trace-formulas (4.48)–(4.62) in **g**-loops]. In Appendix A, we define[9] various (*external* as well as *inlint*) operators used in the present paper, i.e., the string coordinate $X^{\pm j}$, the jth spinning operator ϕ^j, the ghosting operator σ, the fracting operator ϕ, the hilberting operator χ [given by Eq. (A3)], the anti-ghostino $\mathcal{B}$ in (A16), the ghostino $\mathcal{C}$ in (A17), the hilbertino ξ in (A18), the anti-hilbertino η in (A19), the anti-ghoston β in (A20), the ghoston γ in (A21), the string coordino $\Psi^{\pm j}$ in (A22), and the *FMS* spinor $S_{\{h\}}$ in (A24). The normal ordering : [the contraction ⌐¬] in the bo-representation is defined by Eq. (A14) [(A15)], while the normal ordering $: :_\pi$ [the contraction ⌐ $\boldsymbol{\pi}$ ¬] in the $\pi(=\mathrm{NS,R})$-representation is defined by Eq. (A47) [(A43)–(A46)]. The standard ket states $|p(\varphi)=0\rangle$'s, the standard bra states $\langle q(\varphi)=0|$'s, and the dual standard bra states $\langle p(\varphi)=0|$'s are those states satisfying conditions (A6), (A7), and (A9), respectively. In Appendix B, we define the stress operator T_r^π in (B4) [T_r in (B8)] and the BRST currentino $\mathcal{J}_r^\pi$ in (B17) [$\mathcal{J}_r$ in (B21)] in the π [bo]-representation. In the case when the dimension D of space–time is just equal to 10, these three operators T_r^{NS}, T_r^{R}, and T_r are shown to be equal to each other. [See Eq. (B9).] On the other hand, *three* BRST currentinos $\mathcal{J}_r^{\mathrm{NS}}$, $\mathcal{J}_r^{\mathrm{R}}$, and $\mathcal{J}_r$ satisfy Eqs. (B25) and (B26), so that they are *not* equal to each other even for $D=10$. *However*, in the case $D=10$, the ''BRST chargino in the π-representation'' [denoted by $\mathcal{Q}_r^{\mathrm{NS}}$ and $\mathcal{Q}_r^{\mathrm{R}}$] and the ''BRST chargino in the bo-representation'' [denoted by $\mathcal{Q}$] are just equal to each other, on account of the integration in the definitions (B27) and (B28). [See Eq. (B29).] We also prove in the case $D=10$ the inlaying identity (B11) [(B33)] for stress operators T's (BRST charginos $\mathcal{Q}$'s). Thus, we can prove the ''special conservation of the total BRST chargino'' (B34), the ''nilpotency'' (B38), and the ''contacting formula'' (B38b). In Appendix C, we define the *local* (inverse) picture-changing operator $X(Y)$ by Eqs. (C3)–(C7) [Eq. (C8)]. Furthermore, (*non-local*) picture-changing operators X^p (C14) (for any $p\in Z$, $Z+\frac{1}{2}$) and $X^p_\pm$ in (C25) (for any $p\in Z+\frac{1}{2}$) are explicitly given respectively by formulas (C15) and (C27). In particular, *nonlocal* operators $X^0[X^{1/2}_\pm]$ have been used for defining the projection operator $P_r(\frac{1}{2})[P_r(1)]$ in (1.7). We have also given various formulas (C22)–(C24) [(C32)–(C40)] satisfied by $X^0[X^{1/2}_\pm]$. In Appendix D, external *physical* NS (R) particles are given[14] by Eqs. (D1) and (D5) [(D12) and (D13)], which involve *physical* vertex operators $V^{nb}(V^{nf})$ for $n=0,1$ in the NS (R) sector. These results are based on the operator product expansions (OPEs)[14] (D24)–(D26) [(D28)–(D31)] among physical vertex operators in the NS (R) sector and the stress (stressino) operator. [Since we essentially use these OPEs, we have to use the *external* stress operator $T_r^{\mathrm{NS}}(T_r^{\mathrm{R}})$ in (B4) and *external* BRST charginos $\mathcal{Q}_r^{\mathrm{NS}}(\mathcal{Q}_r^{\mathrm{R}})$ in (B27) in the NS (R)-representation. On the other hand, the stress operator T_r in (B8) and the BRST chargino $\mathcal{Q}_r$ in (B28) in the bo-representation are also useful in open QSFT, since they are inlayed into *inlint* operators T and $\mathcal{Q}$ in the ''ICS *s*'' as well as the ''ICS **g**.'' In Appendix E, we construct the *small* gluing vertex function $|\mathcal{V}_{\mathcal{S}}^{\chi}(\gamma,\delta)\rangle$ in (E12) in $\xi\eta$-mode [*large* gluing vertex functino $|\nu_{\mathcal{L}}^{\chi}(\ldots)\rangle$ in (E18) in $\xi\eta$-mode]. Thus we find that the *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ in (E30) is related with the *large* gluing vertex function $|V_L(\gamma,\delta)\rangle$ in (E32) by the formula (E31). This $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ in (E30) satisfies the gluing *identity* (E29) and the gluing *relation* (E41) in the *small* Hilbert space, while $|V_L(\gamma,\delta)\rangle$ in (E32) satisfies the gluing relation (E44) in the *large* Hilbert space. It should be noticed that we have used the *small* gluing vertex function $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ in the gluing theorem (2.20). Finally, elementary vertex functinos have been constructed by using the inlayed coordinates given by Eq. (E34) in the ''ICS *m*,'' which satisfy the GJ-$\mathcal{FE}$ (E36). Then, applying the gluing theorem to these elementary vertex functinos, we have derived various useful formulas (E37)–(E40).

In our previous paper,[5] we have derived the generating functional (7.28) which generates *all* physical scattering amplitudes in QBFT. We should make *two* following corrections in ''Eqs. (7.22) and (7.29) of Ref. 5,'' respectively; Using notations in Ref. 5, we have that

$$
\begin{aligned}
&{}_k\langle \mathrm{mod;tr}|:\exp\left(\sum_{r=1}^{N}\sum_{\mu=1}^{24} i\cdot \oint \frac{dw_r}{2\pi i}P_r^{\mu\dagger}(w_r)\cdot X^{\mu}(z_r(w_r))\right):|\mathrm{mod;tr}\rangle_k \\
&\equiv \exp\left(\sum_{\mu=1}^{24}\sum_{r=1}^{N}\sum_{i=1}^{g} k_i^{\mu}\cdot \oint \frac{dw_r}{2\pi\cdot\sqrt{-1}}P_r^{\mu\dagger}(w_r)\int_{z_0}^{z_r(w_r)}\omega_i\right. \\
&\quad +\sum_{\mu=1}^{24}\sum_{r=1}^{N}\frac{1}{2}\oint\frac{dw_r}{2\pi i}\oint\frac{dw_r'}{2\pi i}P_r^{\mu\dagger}(w_r)\cdot P_r^{\mu\dagger}(w_r')\cdot\log\frac{E(z_r(w_r),z_r(w_r'))}{z_r(w_r)-z_r(w_r')} \\
&\quad \left.+\sum_{\mu=1}^{24}\sum_{\substack{r,s=1\\ r(<)s}}^{N}\oint\frac{dw_r}{2\pi i}\oint\frac{dw_s}{2\pi i}P_r^{\mu\dagger}(w_r)\cdot P_s^{\mu\dagger}(w_s)\cdot\log\frac{E(z_r(w_r),z_s(w_s))}{z_r(w_r)-z_s(w_s)}\right)
\end{aligned}
\tag{5.1}
$$

and

$$
\begin{aligned}
&{}_k\langle \mathrm{mod}|\vec{R}\cdot\left(\prod_{r=1}^{N}:\Psi(Z_r;P_r^{\dagger}):\right)|\mathrm{mod}\rangle_k \\
&=\prod_{r=1}^{N}\exp(-p_r^{+}\cdot X_{\mathrm{eff}}^{-}(Z_r;k^{-})-p_r^{-}\cdot X_{\mathrm{Eff}}^{+}(Z_r:k^{+})) \\
&\times\exp\left(\sum_{\mu=1}^{24}\sum_{r=1}^{N}\sum_{i=1}^{g}k_i^{\mu}\cdot\oint\frac{dw_r}{2\pi i}P_r^{\mu\dagger}(w_r)\int_{z_0}^{\tilde{z}_r(w_r;k^{+})}\omega_i\right. \\
&+\sum_{\mu=1}^{24}\sum_{\substack{r,s=1\\ r(<)s}}^{N}\oint\frac{dw_r}{2\pi i}\oint\frac{dw_s}{2\pi i}P_r^{\mu\dagger}(w_r)\cdot P_s^{\mu\dagger}(w_s)\cdot\log E(\tilde{z}_r(w_r;k^{+}),\tilde{z}_s(w_s;k^{+}))_R \\
&\left.+\sum_{\mu=1}^{24}\sum_{r=1}^{N}\frac{1}{2}\oint\frac{dw_r}{2\pi i}\oint\frac{dw_r'}{2\pi i}P_r^{\mu\dagger}(w_r)\cdot P_r^{\mu\dagger}(w_r')\cdot\log\frac{E(\tilde{z}_r(w_r;k^{+}),\tilde{z}_r(w_r';k^{+}))}{w_r-w_r'}\right).
\end{aligned}
\tag{5.2}
$$

Although we have not given in this paper the results in QSFT (to be obtained by applying the method in Sec. VII of Ref. 5), these results might be straightforwardly (but tediously) obtained, if we use the generating functional of physical vertex operators in each NS and R sector.[14] In the following, we shall briefly comment how the *physical* scattering amplitudes $(N_B,N_F)_{\mathbf{g}}$ in (4.63) can be calculated. First we consider the case when the integrand $F(\{T_I\},\{T_{\gamma^h}\})$ in Eq. (4.63) (in **g**-loops) is given by the *superstring* (Feynman) *diagram* composed of $I_B(I_F)$ propagators and $V_B(V_F)$ interaction vertices among NS–NS–NS (R–R–NS). These numbers are related to each others as follows:

$$2I_B+N_B=3V_B+V_F \tag{5.3}$$

and

$$2I_F+N_F=2V_F. \tag{5.4}$$

In any *connected* superstring (Feynman) diagrams in **g**-loops, we have the following relation among I's and V's:

$$V_B+V_F-1=I_B+I_F-\mathbf{g}. \tag{5.5}$$

In order for $(N_B,N_F)_{\mathbf{g}}$ to give nonvanishing physical amplitudes, the *total* quantum number K [being defined by Eq. (1.30)] of $F(\{T_I\},\{T_{\gamma^h}\})$ should satisfy the following *three* conditions (''conservation laws in **g**-loops''):

$$(G,F,H)=(3-3\mathbf{g},-2+2\mathbf{g},-\mathbf{g}). \tag{5.6}$$

For simplicity of notations, $F(\{T_I\},\{T_{\gamma^h}\})$ in (4.63) will be *shortly* denoted by

$$F(\{T_I\},\{T_{\gamma^h}\})\cong \mathrm{Tr}(\langle \Omega_{\mathbf{g}}((g_B,g_F))\rangle \cdot (X\cdot \exp(\sigma-\phi)\cdot V^{0b})^{N_B} \times(\exp(\sigma-1/2\cdot\phi)\cdot V^{1f}\cdot S_h)^{N_F}(\overline{Y}\cdot Y)^{V_B}\cdot(\overline{Y})^{V_F}(\mathcal{B}_0\cdot X\cdot X)^{I_B}\cdot(\mathcal{B}_0\cdot X)^{I_F}\cdot(\eta_0)^{\mathbf{g}}), \tag{5.7}$$

where $(X\cdot\exp(\sigma-\phi)\cdot V^{0b})^{N_B}$ *symbolically* represents N_B physical NS states in 1 picture, while $(\exp(\sigma-\frac{1}{2}\cdot\phi)\cdot V^{1f}\cdot S_h)^{N_F}$ *symbolically* represents N_F physical R states. [See Eq. (4.65).] In the following, we can count the total quantum numbers K's of $F(\{T_I\},\{T_{\gamma^h}\})$ in (5.7). We first notice that there always appears one factor $e^{+\chi}$ from each $\overline{Y}(Y)$ [see Eqs. (4.30) and (C8)], so that $(\overline{Y}\cdot Y)^{V_B}\cdot(\overline{Y})^{V_F}$ gives the factor $\exp((2V_B+V_F)\chi)$ in total, while $(\eta_0)^{\mathbf{g}}$ gives the factor $e^{-\mathbf{g}\chi}$. Then, since the condition $H=-\mathbf{g}$ must be satisfied for the term (5.7) to be nonvanishing, there should appear the factor $(e^{-\chi})^{2V_B+V_F}$ from $(X\cdot X)^{I_B}\cdot(X)^{I_F}$. On the other hand, each X involves the component having the factor $\mathcal{F}^{(\mathcal{B}\gamma)}\propto e^{-\chi}$, as seen from Eqs. (C15) and (C27). By noticing that this component of X has the same quantum numbers as those of $X^{(-1)}$ in (C7), this component of X will hereafter be *symbolically* expressed by $X^{(-1)}$. Thus, the nonvanishing contribution in the term (5.7) comes from

$$\begin{aligned}F(\{T_I\},\{T_{\gamma^h}\})&\\ \propto X^{N_B+2I_B+I_F-2V_B-V_F}&\cdot\exp((-N_B-\tfrac{1}{2}\cdot N_F)\phi)\cdot\exp((N_B+N_F-I_B-I_F)\sigma)\\ &\cdot\exp(-\mathbf{g}\cdot\chi)(\overline{Y}\cdot Y\cdot X^{(-1)}\cdot X^{(-1)})^{V_B}\cdot(\overline{Y}\cdot X^{(-1)})^{V_F}.\end{aligned} \tag{5.8}$$

It is remarkable that the last two terms on the right-hand side of Eq. (5.8) do *not* have any quantum number (G,F,H). Therefore, the last four terms on the right-hand side of Eq. (5.8) satisfy the *two* conditions $G=3-3\mathbf{g}$ and $H=-\mathbf{g}$ by themselves, because of the relation

$$N_B+N_F-I_B-I_F=3-3\mathbf{g}, \tag{5.9}$$

which can be derived by substituting the relations (5.3) and (5.4) into (5.5). Thus, the first term on the right-hand side of Eq. (5.8) should have *only* the nonvanishing fracting number F (G and H being equal to zeros). This is possible only in the case when we use the X's component involving the factor $\mathcal{F}^{(\vec{X}\cdot\vec{\Psi})}$, as seen from Eqs. (C15) and (C27). [Since this component of the *picture-changing operator* X has the same quantum number as those of $X^{(0)}$ in (C6), this component will hereafter be *symbolically* expressed by $X^{(0)}$.] Our conclusion at this stage is that we should use only the component $X^{(0)}$ in the first term of Eq. (5.8), so that the first *two* terms on the right-hand side of Eq. (5.8) lead [with the help of Eqs. (5.3)–(5.5)] to

$$\begin{aligned}(X^{(0)})^{N_B+2I_B+I_F-2V_B-V_F}&\cdot\exp((-N_B-\tfrac{1}{2}\cdot N_F)\phi)\\ &\propto\exp((2I_B+I_F-2V_B-V_F-1/2\cdot N_F)\phi)\\ &=\exp(\tfrac{2}{3}(I_B+I_F-N_B-N_F)\phi)\\ &=\exp((-2+2\mathbf{g})\phi).\end{aligned} \tag{5.10}$$

Equation (5.10) shows that the condition $F=-2+2\mathbf{g}$ can be satisfied. Thus *all three* conditions (5.6) can be satisfied by some terms in $F(\{T_I\},\{T_{\gamma^h}\})$ something like

$$F(\{T_I\},\{T_{\gamma^h}\})\cong \mathrm{Tr}(\langle\Omega_{\mathbf{g}}((g_B,g_F))\rangle\cdot(\exp(\sigma-\phi)\cdot V^{0b})^{N_B}$$
$$\times(\exp(\sigma-\tfrac{1}{2}\cdot\phi)\cdot V^{1f}\cdot S_h)^{N_F}\cdot(X^{(0)})^{N_B+2I_B+I_F-2V_B-V_F}(\overline{Y}\cdot Y\cdot X^{(-1)}\cdot X^{(-1)})^{V_B}$$
$$\times(\overline{Y}\cdot X^{(-1)})^{V_F}(\mathcal{B}_0)^{I_B+I_F}\cdot(\eta_0)^{\mathbf{g}}), \tag{5.11}$$

where the factor $\exp(\sigma-\phi)\cdot V^{0b}[\exp(\sigma-\frac{1}{2}\cdot\phi)\cdot V^{1f}\cdot S_h]$ is the one in the physical NS state (D1) in 0 picture [R state (D12) in $\frac{1}{2}$ picture], while we have *symbolically* used the abovementioned $X^{(-1)}[X^{(0)}]$. Since any Θ^{NS} and Θ^{R} [in the *modified* propagator (3.37)] give *effectively* one $X^{(+1)}$ (instead of $X^{(0)}$ or $X^{(-1)}$), they do not contribute to the amplitude (5.11).

APPENDIX A: DEFINITIONS OF VARIOUS EXTERNAL AND INLINT OPERATORS

In this Appendix, we shall give the definition of various operators φ $(=X^{\pm j},\ \phi^j,\ \sigma,\ \phi,\ \chi,\ \text{for}\ j=0-4)$ introduced by Friedan–Martinec–Schenker[9] and used in this paper. In calculating physical scattering amplitudes (N_B,N_F) in (1.22) among N_B bosonic and N_F fermionic physical particles (with $N\equiv N_B+N_F$), we use the rth *external* operator $\varphi_r(w_r)$ (for $r=1-N$) which lives within the *unit* disk $|w_r|<1$, and the *inlint* operator $\varphi_0(z)$ which lives within the *whole* complex z plane. In order to express both inlint and external operators altogether, we shall introduce notations $\varphi_r(y_r)$'s (for $r=0-N$) as follow. The argument y_r (for $r=1-N$) is equal to the rth disk coordinate w_r in (2.2), while y_0 is equal to the inlayed coordinate $z(\in\mathbb{C})$. Furthermore, the *inlint* operator $\varphi_0(z)$ is simply denoted by $\varphi(z)$, while $\varphi_r(w_r)$'s (for $r=1-N$) are *external* operators.

First, any ten-dimensional vector A^μ (for $\mu=0-9$) is transformed into the component $A^{\pm j}$ (for $j=0-4$) as follows:

$$A^{\pm 0}\equiv\frac{1}{\sqrt{2}}(A^0\pm A^9),$$
$$A^{\pm j}\equiv\frac{1}{\sqrt{2}}(-\sqrt{-1}\cdot A^j\pm A^{j+4})\quad\text{for}\ \ j=1-4, \tag{A1}$$

so that the Lorentz scalar product among A^μ and B^ν is expressed by

$$\vec{A}\cdot\vec{B}\equiv-\sum_{\pm}\sum_{j=0}^{4}A^{\pm j}\cdot B^{\mp j}\left(\equiv-\sum_{\pm}A^{\pm 0}\cdot B^{\mp 0}+\sum_{j=1}^{8}A^j\cdot B^j\right). \tag{A2}$$

In open QSFT, we have the string coordinate $X^{\pm j}$ (for $j=0-4$), the jth spining operator ϕ^j (for $j=0-4$), the ghosting operator σ, the fracting operator ϕ, and the Hilberting operator χ. These operators (represented by φ's) are coupled with background charges $Q(\varphi)$'s in (2.5). For any $r=0-N$, the operator $\varphi_r(y_r)$ is defined by

$$\varepsilon_\varphi^\phi\cdot\varphi_r(y_r)\equiv\left(p_r(\varphi)\cdot\log(y_r)-\sum_{n=1}^{\infty}\frac{1}{n}\cdot J_{n,r}(\varphi)\cdot y_r^{-n}\right)+\left(\varepsilon_\varphi^\phi\cdot q_r(\varphi)+\sum_{n=1}^{\infty}\frac{1}{n}\cdot J_{-n,r}(\varphi)\cdot y_r^{n}\right)$$
$$\equiv\varphi_r(y_r;+)+\varphi_r(y_r;-)\quad\text{for}\ \ \varphi=X^{\pm j},\phi^j,\sigma,\phi,\chi,\quad\text{and}\ \ j=0-4. \tag{A3a}$$

Incidentally, the string coordinate $X^{\pm j}$ in (A3) is the same as the one in Ref. 13, since we have the following relations among them:

$$p_r(X^{\pm j}) = -\sqrt{-1}\cdot p_r^{\pm j}, \quad q_r(X^{\pm j}) = q_r^{\pm j},$$
$$J_{n,r}(X^{\pm j}) = -\sqrt{-1}\cdot\sqrt{n}\cdot a_{n,r}^{\pm j}, \quad J_{-n,r}(X^{\pm j}) = -\sqrt{-1}\cdot\sqrt{n}\cdot a_{-n,r}^{\pm j}. \tag{A3b}$$

Operators in Eq. (A3) satisfy the following commutation relations (for r, $t = 0-N$ and $n,m = 1,2,3,\ldots$]:

$$[J_{n,r}(X^{\pm j}), J_{-m,t}(X^{\mp k})] = n\cdot\delta_{nm}\cdot\delta_{rt}\cdot\delta_{jk} \quad \text{for } j,k = 0-4,$$
$$[p_r(X^{\pm j}), q_t(X^{\mp k})] = \delta_{rt}\cdot\delta_{jk} \quad \text{for } j,k = 0-4, \tag{A4}$$

and

$$[J_{n,r}(\varphi), J_{-m,t}(\varphi')] = n\cdot\delta_{nm}\cdot\delta_{rt}\cdot\delta_{\varphi\varphi'}\cdot\varepsilon_\varphi^\phi \quad \text{for } \varphi,\varphi' = \phi^j,\sigma,\phi,\chi, \quad \text{and } j = 0-4,$$
$$[p_r(\varphi), q_t(\varphi')] = \delta_{rt}\cdot\delta_{\varphi\varphi'} \quad \text{for } \varphi,\varphi' = \phi^j,\sigma,\phi,\chi, \quad \text{and } j = 0-4, \tag{A5a}$$

all other commutators being equal to zero. In Eq. (A5a), we have used the constant ε_φ^ϕ defined by

$$\varepsilon_\varphi^\phi \equiv \begin{cases} -1 & \text{for } \varphi = \phi, \\ +1 & \text{for } \varphi(\neq\phi). \end{cases} \tag{A5b}$$

The (Grassman even) standard ket states[5] $|p_r(\vec{X}) = 0\rangle_r$ and $|p_r(\varphi) = 0\rangle_r$ (for $r = 0-N$) are states satisfying respectively that

$$X_r^{\pm j}(y_r;+)|p_r(\vec{X}) = 0\rangle_r = 0 \quad \text{for } j = 0-4,$$
$$\varphi_r(y_r;+)|p_r(\varphi) = 0\rangle_r = 0 \quad \text{for } \varphi = \phi^j,\sigma,\phi,\chi, \quad \text{and } j = 0-4, \tag{A6}$$

while the (Grassman even) standard bra states[5] ${}_r\langle q_r(\vec{X}) = 0|$ and ${}_r\langle q_r(\varphi) = 0|$ (for $r = 0-N$) are states satisfying respectively that

$${}_r\langle q_r(\vec{X}) = 0|X_r^{\pm j}(y_r;-) = 0 \quad \text{for } j = 0-4,$$
$${}_r\langle q_r(\varphi) = 0|\varphi_r(y_r;-) = 0 \quad \text{for } \varphi = \phi^j,\sigma,\phi,\chi, \quad \text{and } j = 0-4. \tag{A7}$$

Normalizations of these states are respectively fixed by

$${}_r\langle q_r(\vec{X}) = 0|p_t(\vec{X}) = 0\rangle_t = \delta_{rt},$$
$${}_r\langle q_r(\varphi) = 0|p_t(\varphi') = 0\rangle_t = \delta_{rt}\cdot\delta_{\varphi\varphi'} \quad \text{for } \varphi = \phi^j,\sigma,\phi,\chi, \quad \text{and } j = 0-4. \tag{A8}$$

Furthermore, the "dual standard bra states" ${}_r\langle p_r(\vec{X}) = 0|$ and ${}_r\langle p_r(\varphi) = 0|$ (for $r = 0-N$) are defined respectively by

$${}_r\langle p_r(\vec{X}) = 0| \equiv (|p_r(\vec{X}) = 0)_r)^\dagger,$$
$${}_r\langle p_r(\varphi) = 0| \equiv (|p_r(\varphi) = 0\rangle_r)^\dagger \quad \text{for } \varphi = \phi^j,\sigma,\phi,\chi, \quad \text{and } j = 0-4. \tag{A9}$$

We also impose the following normalization conditions:

$$ {}_r\langle p_r(\vec{X})=0|\exp\left(\sum_{\pm}\sum_{j=0}^{4} -\sqrt{-1}p^{\pm j}\cdot X_r^{\mp j}(y_r)\right)|p_t(\vec{X})=0\rangle_t=\delta_{rt}\left(\prod_{\mu=0}^{9} 2\pi\delta(p^\mu)\right), \tag{A10} $$

$$ {}_r\langle p_r(\varphi)=0|\exp(-Q(\varphi)\varphi_r(y_r))|p_t(\varphi')=0\rangle_t=\delta_{rt}\delta_{\varphi\varphi'} \quad \text{for } \varphi,\varphi'=\phi^j,\sigma,\phi,\chi, \quad \text{and } j=0-4, $$

where the background charges $Q(\varphi)$'s are given by Eq. (2.5). [Real numbers p^μ's are related with $p^{\pm j}$'s by Eq. (A1).] We finally remark that the Hermitian conjugation, ''†'' is given by

$$ (\varphi_r(1/y_r^*))^\dagger=\varphi_r(y_r)+\varepsilon_\varphi^\phi\cdot Q(\varphi)\cdot\log(y_r) \quad \text{for } \varphi=\sigma,\phi,\chi, \tag{A11a} $$

so that we find from Eq. (A3) that

$$ -(p_r(\varphi))^\dagger=p_r(\varphi)+Q(\varphi). \tag{A11b} $$

We find that ${}_r\langle p_r(\varphi)=0|$ is Grassman even (odd) for $\varphi=\vec{X},\phi^j,\ \phi$ (for $\varphi=\sigma,\ \chi$). The ''normal ordering in the bo-(sonized) representation'' is denoted by : :, which is defined to be placing any negative frequency part $\varphi_r(y_r;-)$ always to the left of any positive frequency part $\varphi_t'(y_t';+)$ [for any $\varphi,\varphi'=X^{\pm j},\phi^j,\sigma,\phi,\chi$]. Furthermore, the radial ordering operator $\vec{R}$ is defined to be rearranging any product [of external or inlint operators] according to the absolute values of operators' arguments:

$$ \vec{R}\cdot(\varphi_r(y_r)\cdot\varphi_t'(y_t'))\equiv\begin{cases}\varphi_r(y_r)\cdot\varphi_t'(y_t') & \text{for } |y_r|>|y_t'|,\\ \varphi_t'(y_t')\cdot\varphi_r(y_r) & \text{for } |y_t'|>|y_r|,\end{cases} $$
$$ \text{for } \varphi,\varphi'=X^{\pm j},\phi^j,\sigma,\phi,\chi, \quad \text{and } j=0-4. \tag{A12} $$

If we define the *radial-ordered* argument $(y_r-y_t')_R$ by

$$ (y_r-y_t')_R\equiv\begin{cases}(y_r-y_t') & \text{for } |y_r|>|y_t'|,\\ (y_t'-y_r) & \text{for } |y_t'|>|y_r|.\end{cases} \tag{A13} $$

Commutation relations (A3)–(A5) are found to lead to the following operator product expansions (OPEs):

$$ \vec{R}\cdot(\varphi_r(y_r)\cdot\varphi_t'(y_t'))=\overbrace{\varphi_r(y_r)\cdot\varphi_t'(y_t')}+:(\varphi_r(y_r)\cdot\varphi_t'(y_t')): $$
$$ \text{for } \varphi,\varphi'=X^{\pm j},\phi^j,\sigma,\phi,\chi, \quad \text{and } j=0-4. \tag{A14} $$

The first (second) term on the right-hand side of Eq. (A14) will be called the ''contraction (the normal-ordered product) in the bo-representation.'' All of nonvanishing contractions in Eqs. (A14) are those given by

$$ \overbrace{X_r^{\pm j}(y_r)\cdot X_t^{\mp k}(y_t')}=\delta_{rt}\cdot\delta_{jk}\cdot\log(y_r-y_t')_R \quad \text{for } j,k=0-4 $$
$$ \overbrace{\varphi_r(y_s)\cdot\varphi_t(y_t')}=\varepsilon_\varphi^\phi\cdot\delta_{rt}\cdot\log(y_r-y_t')_R \quad [\text{see } \varepsilon_\varphi^\phi \text{ (A5a)}] $$
$$ \text{for } \varphi=\phi^j,\sigma,\phi,\chi, \quad \text{and } j=0-4. \tag{A15} $$

(Throughout this paper, we will explicitly write out the radial ordering operator $\vec{R}$ in various formulas.)

[Hereafter, the conformal weight of the primary operator $\mathcal{G}$ will be denoted by $d(\mathcal{G})$.] The anti-ghostino $\mathcal{B}_r$ and the ghostino $\mathcal{C}_r$ are defined respectively [in terms of the ghosting operator σ_r in (A3)] by

$$\mathcal{B}_r(y_r)\equiv:\exp(-\sigma_r(y_r)): \quad [d(\mathcal{B})=+2] \tag{A16}$$

and

$$\mathcal{C}_r(y_r)\equiv:\exp(+\sigma_r(y_r)): \quad [d(\mathcal{C})=-1], \tag{A17}$$

cocycle factors[15] being abbreviated here and hereafter.

On the other hand, Friedan–Martinec–Schenker (FMS) have introduced[9] the hilbertino ξ_r, the anti-hilbertino η_r, the anti-ghoston β_r, the ghoston γ_r, and the string coordino $\Psi_r^{\pm j}$ (for $j=0-4$), which can respectively be constructed [in terms of the hilberting operator χ_r, the fracting operator ϕ_r, and the jth spining operator ϕ_r^j given by Eq. (A3)] as follows:

$$\xi_r(y_r)\equiv:\exp(+\chi_r(y_r)): \quad [d(\xi)=0], \tag{A18}$$

$$\eta_r(y_r)\equiv:\exp(-\chi_r(y_r)): \quad [d(\eta)=+1], \tag{A19}$$

$$\beta_r(y_r)\equiv\partial_{y_r}\xi_r(y_r):\exp(-\phi_r(y_r)): \quad [d(\beta)=+\tfrac{3}{2}], \tag{A20}$$

$$\gamma_r(y_r)\equiv:\exp(+\phi_r(y_r)):\eta_r(y_r) \quad [d(\gamma)=-\tfrac{1}{2}], \tag{A21}$$

and

$$\Psi_r^{\pm j}(y_r)\equiv:\exp(\pm\phi_r^j(y_r)): \quad \text{for } j=0-4 \quad [d(\Psi)=+\tfrac{1}{2}]. \tag{A22}$$

The rth punctured ring domain in the $\pi(r)=\mathrm{R}$ sector is characterized by the presence of the following external (inlint) operator at $w_r=0$ [at the rth puncture $Z_{r_\flat}$]:

$$:\exp(+\sigma_r(y_{0r}))::\exp(-\tfrac{1}{2}\phi_r(y_{0r})):S_{\{h(r)\},r}(y_{0r}). \tag{A23}$$

[See Eq. (D12).] In particular, the FMS spinor[9] $S_{\{h(r)\},r}(y_{0r})$ represents the fermionic spinor at y_{0r} and it is defined by

$$S_{\{h(r)\},r}(y_{0r})\equiv:\exp(+\tfrac{1}{2}\cdot\phi_r^0(y_{0r})):\prod_{j=1}^{4}:\exp(\varepsilon_{h(r)}^j\cdot\phi_r^j(y_{0r})):, \tag{A24a}$$

with

$$\varepsilon_{h(r)}^j\equiv\pm\tfrac{1}{2} \quad \text{for } h(r)=\pm. \tag{A24b}$$

[In Eqs. (A18)–(A24) and hereafter, cocycle factors[15] will be abbreviated.] We notice that $:\exp(\pm\varphi_r(y_r)):$ involves the factor $(y_r)^{\mp p_r(\varphi)}$ (for $\varphi=\sigma,\chi$), where eigenvalues $p_r(\varphi)$ are integer valued in *both* the Neveu–Schwarz (NS) and Ramond (R) sectors. Since any of $\mathcal{B}_r(y_r)$ in (A16), $\mathcal{C}_r(y_r)$ in (A17), $\xi_r(y_r)$ in (A18), and $\eta_r(y_r)$ in (A19) is single valued, each can respectively be Laurent expanded at $y_r=0$ as

$$\mathcal{B}_r(y_r)=\left(\sum_{n=-1}^{\infty}\mathcal{B}_{n,r}\cdot y_r^{-n-2}\right)+\left(\sum_{n=2}^{\infty}\mathcal{B}_{-n,r}\cdot y_r^{n-2}\right)\equiv\mathcal{B}_r(y_r;+)+\mathcal{B}_r(y_r;-) \quad \text{with } \mathcal{B}_{n,r}^{+}=\mathcal{B}_{-n,r}, \tag{A25}$$

$$\mathcal{C}_r(y_r)=\left(\sum_{n=2}^{\infty}\mathcal{C}_{n,r}\cdot y_r^{-n+1}\right)+\left(\sum_{n=-1}^{\infty}\mathcal{C}_{-n,r}\cdot y_r^{n+1}\right)\equiv\mathcal{C}_r(y_r;+)+\mathcal{C}_r(y_r;-)\quad\text{with }\mathcal{C}_{n,r}^{\dagger}=\mathcal{C}_{-n,r},\tag{A26}$$

$$\xi_r(y_r)=\left(\sum_{n=1}^{\infty}\xi_{n,r}\cdot y_r^{-n}\right)+\left(\sum_{n=0}^{\infty}\xi_{-n,r}\cdot y_r^{n}\right)\equiv\xi_r(y_r;+)+\xi_r(y_r;-)\quad\text{with }\xi_{n,r}^{\dagger}=\xi_{-n,r},\tag{A27}$$

and

$$\eta_r(y_r)=\left(\sum_{n=0}^{\infty}\eta_{n,r}\cdot y_r^{-n-1}\right)+\left(\sum_{n=1}^{\infty}\eta_{-n,r}\cdot y_r^{n-1}\right)\equiv\eta_r(y_r;+)+\eta_r(y_r;-)\quad\text{with }\eta_{n,r}^{\dagger}=\eta_{-n,r}.\tag{A28}$$

Finally, the ''normal ordering in the Σ-representation'' (denoted by $: :_\Sigma$) is defined by

$$:\mathcal{G}_r(y_r)\cdot\mathcal{G}'_t(y'_t):_\Sigma\equiv\vec{R}\cdot\mathcal{G}_r(y_r)\cdot\mathcal{G}'_t(y'_t)-\overbrace{\mathcal{G}_r(y_r)\cdot\mathcal{G}'_t(y'_t)}^{\Sigma}\quad\text{for }\mathcal{G},\mathcal{G}'=\mathcal{B},\mathcal{C},\xi,\eta,\vec{X},\tag{A29}$$

where the ''contractions in the Σ-representation''

$$\overbrace{\quad}^{\Sigma}$$

are given by

$$\overbrace{\mathcal{B}_r(y_r)\cdot\mathcal{C}_t(y'_t)}^{\Sigma}=-\overbrace{\mathcal{C}_t(y'_t)\cdot\mathcal{B}_r(y_r)}^{\Sigma}=\overbrace{\xi_r(y_r)\cdot\eta_t(y'_t)}^{\Sigma}=-\overbrace{\eta_t(y'_t)\cdot\xi_r(y_r)}^{\Sigma}=\frac{\delta_{rt}}{y_r-y'_t}\tag{A30a}$$

and

$$\overbrace{X_r^{\pm j}(y_r)\cdot X_t^{\mp k}(y'_t)}^{\Sigma}=\delta_{rt}\cdot\delta_{jk}\cdot\log(y_r-y'_t)_R,\tag{A30b}$$

all other contractions in the Σ-representation among $\mathcal{G}$ and $\mathcal{G}'$ (for $\mathcal{G},\mathcal{G}'=\mathcal{B},\mathcal{C},\xi,\eta,\vec{X}$) being equal to zero. Incidentally, $: :_\Sigma$ defined by Eq. (A29) is equivalent to placing any negative frequency part to the left of any positive frequency part (of operators $\mathcal{B}$, $\mathcal{C}$, ξ, η, $\vec{X}$).

Noticing that

$$:\exp(\pm\varphi_r(y_r)):|p_r(\varphi)=0\rangle_r=:\exp(\pm\varphi_r(y_r;-)):|p_r(\varphi)=0\rangle_r\quad\text{for }\varphi=\phi,\chi\tag{A31}$$

is analytic at $y_r=0$, we find from Eqs. (A25)–(A28) that

$$\begin{aligned}\mathcal{B}_r(y_r;+)|p_r(\sigma)=0\rangle_r&=\mathcal{C}_r(y_r;+)|p_r(\sigma)=0\rangle_r\\&=\xi_r(y_r;+)|p_r(\chi)=0\rangle_r\\&=\eta_r(y_r;+)|p_r(\chi)=\rangle_r\\&=0.\end{aligned}\tag{A32}$$

Similarly, we find from Eq. (A7) that

$$\begin{aligned}{}_r\langle q_r(\sigma)=0|\mathcal{B}_r(y_r;-)={}_r\langle q_r(\sigma)=0|\mathcal{C}_r(y_r;-)\\ &={}_r\langle q_r(\chi)=0|\xi_r(y_r;-)\\ &={}_r\langle q_r(\chi)=0|\eta_r(y_r;-)=0.\end{aligned} \tag{A33}$$

Since we have that

$$_r\langle p_r(\chi)=0|\begin{pmatrix}\xi_{0,r}\\ 1\end{pmatrix}|p_t(\chi)=0\rangle_t=\begin{pmatrix}\delta_{rt}\\ 0\end{pmatrix}, \tag{A34}$$

we find that ${}_r\langle p_r(\chi)=0|$ is Grassman *odd*. Furthermore, Eqs. (A8) and (A34) lead to

$$\begin{aligned}{}_r\langle q_r(\chi)=0| &= {}_r\langle p_r(\chi)=0|\xi_{0,r}\\ &={}_r\langle p_r(\chi)=0|\xi_r(+\infty) \quad [\text{from Eqs. (A9), (A27), and (A33)}],\end{aligned} \tag{A35a}$$

$$\begin{aligned}&={}_r\langle p_r(\chi)=0|{:}\exp(+\chi_r(+\infty)){:}, \quad [\text{from Eq. (A18)}],\\ &={}_r\langle p_r(\chi)=0|\exp(+q_r(\chi))\\ &\quad [\text{from Eqs. (A3), (A7), (A9), and (A11)}].\end{aligned} \tag{A35b}$$

On the other hand, the anti-ghoston β in (A20) and the ghoston γ in (A21) and the string coordino $\Psi^{\pm j}$ in (A22) in the R sector should be treated differently from those in the NS sector. Especially in the case $\pi(r)=\mathrm{R}$, the operator (A23) exists at y_{0r}. Incidentally, the presence of the operator (A23) (of conformal weight 0) reflects the presence of the rth external physical R-state in (D12) [in $\frac{1}{2}$ picture] at the puncture y_{0r} within the rth punctured ring domain. The presence of the operator (A23) induces in β in (A20), γ in (A21), and $\Psi^{\pm j}$ in (A22) the *branch point* at y_{0r}, which should be taken into account in Laurent expanding these operators. The branch point in the string coordino $\Psi^{\pm j}$ has been completely analyzed in our previous papers,[14] so that we have only to apply previous techniques to the analyses of β and γ. [On the other hand, we point out that there is no essential need to discriminate $\mathcal{B}$ in (A25), $\mathcal{C}$ in (A26), ξ in (A27), and η in (A28) in the NS sector from those in the R sector.] In the following, formulas with (without) the branch point are those in the R (NS) sector: In the special case when $y_{0r}=0$, the anti-ghoston β in (A20) [of conformal weight $d(\beta)=\frac{3}{2}$], the ghoston γ in (A21) [of $d(\gamma)=-\frac{1}{2}$], and the string coordino $\Psi^{\pm j}$ in (A22) [of $d(\Psi^{\pm j})=\frac{1}{2}$] can be respectively Laurent expanded at $y_r=0$ in the following forms. For $\pi(r)=\mathrm{R}$ (NS), we find that

$$\begin{aligned}\beta_r(y_r)&=\frac{1}{(\sqrt{y_r})^3}\left(\sum_{n=0}^{\infty}\beta_{n(-1/2),r}\cdot y_r^{-n(+1/2)}\right)\\ &\quad+\frac{1}{(\sqrt{y_r})^3}\left(\sum_{n=1}^{\infty}\beta_{-n(-1/2),r}\cdot y_r^{n(+1/2)}\right)\\ &=\beta_r(y_r;+)+\beta_r(y_r;-) \quad \text{with} \quad \beta^{\dagger}_{n(-1/2),r}=\beta_{-n(+1/2),r},\end{aligned} \tag{A36}$$

$$\begin{aligned}\gamma_r(y_r)&=\sqrt{y_r}\left(\sum_{n=1}^{\infty}\gamma_{n(+1/2),r}\cdot y_r^{-n(-1/2)}\right)+\sqrt{y_r}\left(\sum_{n=0}^{\infty}\gamma_{-n(+1/2),r}\cdot y_r^{n(-1/2)}\right)\\ &=\gamma_r(y_r;+)+\gamma_r(y_r;-) \quad \text{with} \quad \gamma^{\dagger}_{n(+1/2),r}=-\gamma_{-n(-1/2),r},\end{aligned} \tag{A37}$$

and

$$\Psi_r^{\pm j}(y_r)=\frac{1}{\sqrt{y_r}}\left(\sum_{n\in Z}\psi_{n(+1/2),r}^{\pm j}\cdot y_r^{-n(-1/2)}\right). \tag{A38}$$

[The suffix $-n(+1/2)$ means the suffix $-n$ $(-n+1/2)$ should be used in the R (NS) sector.] The expansion in the general case $y_{0r}\neq 0$ will be obtained in much the same way as we have obtained the expansions (2.24) and (2.28) in the first paper of Ref. 14.

At this stage, we introduce the following super-coordinate $X_r^{\pm j}(y_r,\theta_r)$, the super-ghostino $\mathcal{C}_r(y_r,\theta_r)$, and the super-antighoston $\beta_r(y_r,\theta_r)$:

$$X_r^{\pm j}(y_r,\theta_r)\equiv X_r^{\pm j}(y_r)+\theta_r\cdot\Psi_r^{\pm j}(y_r), \tag{A39}$$

$$\mathcal{C}_r(y_r,\theta_r)\equiv\mathcal{C}_r(y_r)+\theta_r\cdot\gamma_r(y_r), \tag{A40}$$

and

$$\beta_r(y_r,\theta_r)\equiv\beta_r(y_r)+\theta_r\cdot\mathcal{B}_r(y_r). \tag{A41}$$

We shall later use the following partial differentiation [of operators (A39)–(A41)]:

$$\mathcal{D}_{\theta_r}\equiv\frac{\partial}{\partial\theta_r}+\theta_r\cdot\frac{\partial}{\partial y_r}. \tag{A42}$$

The "contractions in the π-representation"

$$\overset{\sqcap\,\pi\,\sqcap}{}$$

[for $\pi=$R,NS] are those respectively defined by

$$\overbrace{X_r^{\pm j}(y_r,\theta_r)\cdot X_t^{\mp k}}^{\text{NS}}(y_t',\theta_t')\equiv\delta_{rt}\cdot\delta_{jk}\cdot\log(y_r-y_t'-\theta_r\theta_t')_R, \tag{A43}$$

$$\overbrace{\beta_r(y_r,\theta_r)\cdot\mathcal{C}_t}^{\text{NS}}(y_t',\theta_t')\equiv\delta_{rt}\cdot\frac{\theta_r-\theta_t'}{y_r-y_t'-\theta_r\theta_t'}, \tag{A44}$$

and

$$\overbrace{X_r^{\pm j}(y_r,\theta_r)\cdot X_t^{\mp k}}^{\text{R}}(y_t',\theta_t')\equiv\delta_{rt}\cdot\delta_{jk}\cdot\log\left(y_r-y_t'-\left(\frac{y_r-y_{0r}}{y_t'-y_{0t}}\right)^{\pm\varepsilon_h^j}\theta_r\theta_t'\right)_R, \tag{A45}$$

$$\overbrace{\beta_r(y_r,\theta_r)\cdot\mathcal{C}_t}^{\text{R}}(y_t',\theta_t')\equiv\frac{\delta_{rt}}{y_r-y_t'-\theta_r\theta_t'}\left(\theta_r\left(\frac{y_t'-y_{0t}}{y_r-y_{0r}}\right)-\theta_t'\left(\frac{y_t'-y_{0t}}{y_r-y_{0r}}\right)^{1/2}\right). \tag{A46}$$

In the formulas (A43) and (A45), we have used the radial-ordered argument $(\dots)_{\mathrm{R}}$ defined by Eq. (A13). Then, we can *define* the "normal-ordered product in the π-representation" (denoted by $:\,:_\pi$) as follows:

$$:G_r(y_r,\theta_r)\cdot H_t(y_t',\theta_t'):_\pi\equiv\vec{R}\cdot G_r(y_r,\theta_r)\cdot H_t(y_t',\theta_t')-\overbrace{G_r(y_r,\theta_r)\cdot H_t}^{\pi}(y_t',\theta_t')\quad\text{for }\ \pi=\text{NS,R}. \tag{A47}$$

We should notice that the formula (A46) has been chosen so as to involve the factor $(y_t' - y_{0t})/(y_r - y_{0r})$, reflecting the effect due to the existence of $:\exp(\sigma_r(y_{0r})):$ $[:\exp(\sigma_t(y_{0t}))]$, which is needed to make the conformal weight of the $r(t)$th external primary operators (A23) equal to zero. [See Eqs. (D12), (D14), and (D28).]

APPENDIX B: IDENTITIES $T_r^{\mathrm{R}} = T_r^{\mathrm{NS}} = T_r$ AND $\mathcal{Q}_r^{\mathrm{R}} = \mathcal{Q}_r^{\mathrm{NS}} = \mathcal{Q}_r$

In this appendix, $: :_\pi(: \;:)$ always means the ''normal ordering in the π- (bo-)representation'' defined by Eq. (A47) [(A14)], where we should use contractions (A43)–(A46) [(A15)]. Then, $\mathcal{F}_r^\pi(y_r, \theta_r)$ is the ''super-stressino operator in the π-representation'' defined by

$$\mathcal{F}_r^\pi(y_r,\theta_r)\left(\equiv \frac{1}{2}\cdot \mathcal{F}_r^\pi(y_r) + \theta_r \cdot T_r^\pi(y_r)\right) \quad \text{for } r = 0 - N \tag{B1}$$

$$\equiv \sum_{\pm} \sum_{j=0}^{4} \frac{1}{2} :\partial_{y_r} X_r^{\pm j}(y_r,\theta_r)\cdot \mathcal{D}_{\theta_r} X_r^{\mp j}(y_r,\theta_r):_\pi$$
$$+ :(-\mathcal{C}_r(y_r,\theta_r)\cdot \partial_{y_r}\beta_r(y_r,\theta_r) + \tfrac{1}{2}\cdot \mathcal{D}_{\theta_r}\mathcal{C}_r(y_r,\theta_r)\cdot \mathcal{D}_{\theta_r}\beta_r(y_r,\theta_r)$$
$$- \tfrac{3}{2}\cdot \partial_{y_r}\mathcal{C}_r(y_r,\theta_r)\cdot \beta_r(y_r,\theta_r)):_\pi \tag{B2}$$

$$\equiv :\mathcal{F}_r^{(\vec{X}\cdot\vec{\Psi})}(y_r,\theta_r):_\pi + :\mathcal{F}_r^{(gh)}(y_r,\theta_r):_\pi \quad \text{for } \pi = \mathrm{R,NS}, \tag{B3}$$

where $\mathcal{D}_{\theta_r}$ is defined by Eq. (A42). By substituting Eqs. (A39)–(A42) into Eq. (B2), $T_r^\pi(y_r)$ in (B1) (to be referred to as the ''stress operator in the π-representation'') is found to be given by

$$T_r^\pi(y_r) \equiv :T_r^{(\vec{X})}(y_r):_\pi + :(T_r^{(\vec{\Psi})}(y_r) + T_r^{(\mathcal{BC})}(y_r) + T_r^{(\beta\gamma)}(y_r)):_\pi \quad \text{for } \pi = \mathrm{R,NS} \text{ and } r = 0 - N, \tag{B4a}$$

where [and in Eq. (B3)] the operator $T^{(G)}$ $[\mathcal{F}^{(G)}]$ represents the component given in terms of G-modes. For examples, we have that

$$:T_r^{(\mathcal{BC})}(y_r):_\pi \equiv :\mathcal{C}_r(y_r)\cdot \partial_{y_r}\mathcal{B}_r(y_r):_\pi + 2:\partial_{y_r}\mathcal{C}_r(y_r)\cdot \mathcal{B}_r(y_r):_\pi \tag{B4b}$$

and

$$:T_r^{(\beta\gamma)}(y_r):_\pi \equiv -\tfrac{1}{2}:\gamma_r(y_r)\cdot \partial_{y_r}\beta_r(y_r):_\pi - \tfrac{3}{2}:\partial_{y_r}\gamma_r(y_r)\cdot \beta_r(y_r):_\pi . \tag{B4c}$$

{See also Eqs. (C16)–(C18) for $\mathcal{F}^{(G)}$'s [with $(G) = (\vec{X}\cdot\vec{\Psi}),(\beta\mathcal{C}),(\mathcal{B}\gamma)$] and Eq. (D22) for $T^{(G)}$ [with $(G) = (\vec{X}),(\vec{\Psi})$], which are to be given later.} On the other hand, $T_r^{(\varphi)}(y_r)$ (to be referred to as the ''φ-mode's stress operator in the bo-representation'') is defined by[9]

$$T_r^{(\varphi)}(y_r) \equiv \varepsilon_\varphi^\phi \cdot \frac{1}{2}:(\partial_{y_r}\varphi_r(y_r))^2: - \frac{Q(\varphi)}{2}\cdot \partial_{y_r}^2 \varphi_r(y_r) \quad \text{for } \varphi = \phi^j, \sigma, \phi, \chi, \tag{B5}$$

where we have used the background charges $Q(\varphi)$'s in (2.5) and the constant ε_φ^ϕ in (A5b). Using techniques in Ref. 5, we find after straightforward calculations (in the general space–time dimension D) that $T_r^\pi(y_r)$ in (B4) is given in terms of $T_r^{(\varphi)}(y_r)$'s in (B5) as follows:

$$T_r^\pi(y_r) = T_r(y_r) + \left(-\frac{D}{16} + \frac{5}{8}\right)\frac{\delta_\pi^{\mathrm{R}}}{(y_r - y_{0r})^2} \quad \text{for } \pi = \mathrm{R,NS}, \tag{B6}$$

where the constant δ_π^{R} is defined by

$$\delta_\pi^{\mathrm{R}} \equiv \begin{cases} 1 & \text{for } \pi = \mathrm{R}, \\ 0 & \text{for } \pi = \mathrm{NS}, \end{cases} \tag{B7}$$

while $T_r(y_r)$ (to be referred to as the ''stress operator in the bo-representation'') is defined by

$$T_r(y_r) \equiv :T_r^{(\vec{X})}(y_r):_\pi + \sum_{\varphi = \phi^j, \sigma, \phi, \chi} T_r^{(\varphi)}(y_r), \tag{B8}$$

where $T_r^{(\varphi)}(y_r)$'s are defined by Eq. (B5), while $T_r^{(\vec{X})}(y_r)$ is given by the first term on the right-hand side of Eq. (D22). Since we have $:T_r^{(\vec{X})}(y_r):_\pi \equiv :T_r^{(\vec{X})}(y_r):_\Sigma$, we find that

$$T_r^{\mathrm{R}}(y_r) = T_r^{\mathrm{NS}}(y_r) = T_r(y_r) \quad \text{for } D = 10 \quad \text{and } r = 0 - \mathrm{N} \tag{B9a}$$

$$\equiv \sum_{n=-\infty}^{\infty} L_{n,r} \cdot y_r^{-n-2} \quad [d(T) = 2]. \tag{B9b}$$

Furthermore, straightforward calculations lead to the following inlaying identity among the stress operators in the NS-representation:

$$T_r^{\mathrm{NS}}(w_r) \overset{\mathcal{I}}{\Rightarrow} (z_{r\mathfrak{z}}^{(1)}(w_r))^2 \cdot T^{\mathrm{NS}}(z_{r\mathfrak{z}}(w_r)) + \left(\frac{D}{8} - \frac{5}{4}\right)\left(\frac{z_{r\mathfrak{z}}^{(3)}(w_r)}{z_{r\mathfrak{z}}^{(1)}(w_r)} - \frac{3}{2}\left(\frac{z_{r\mathfrak{z}}^{(2)}(w_r)}{z_{r\mathfrak{z}}^{(1)}(w_r)}\right)^2\right). \tag{B10}$$

With the help of Eq. (B9), the inlaying identity (B10) is reduced (in the case $D=10$) to

$$T_r^{\pi(r)}(w_r) \overset{\mathcal{I}}{\Rightarrow} (z_{r\mathfrak{z}}^{(1)}(w_r))^2 \cdot T(z_{r\mathfrak{z}}(w_r)) \quad \text{for } r = 1 - N \quad \text{and in the case } D = 10. \tag{B11}$$

On the right-hand side of Eq. (B11), we have used the ''*inlint* stress operator in the bo-representation.'' Finally, the stressino operators in *three* representations are just equal to each other:

$$\mathcal{F}_r^{\mathrm{R}}(y_r) = \mathcal{F}_r^{\mathrm{NS}}(y_r) = \mathcal{F}_r(y_r) \quad \text{for } r = 0 - N. \tag{B12}$$

We can easily see that the stressino operator $\mathcal{F}_r^{\pi(r)}(w_r)$ is inlayed into the *inlint* stressino operator $\mathcal{F}(z_{r\mathfrak{z}}(w_r))$ as follows:

$$\mathcal{F}_r^{\pi(r)}(w_r) \overset{\mathcal{I}}{\Rightarrow} (z_{r\mathfrak{z}}^{(1)}(w_r))^{3/2} \cdot \mathcal{F}(z_{r\mathfrak{z}}(w_r)) \quad \text{for } r = 1 - N. \tag{B13}$$

It is tedious but straightforward to derive the following OPE among superstressino operators $\mathcal{F}_r^{\mathrm{NS}}(y_r, \theta_r)$'s in (B1) in the NS-representation:

$$\begin{aligned} &\vec{R} \cdot \mathcal{F}_r^{\mathrm{NS}}(y_r, \theta_r) \cdot \mathcal{F}_t^{\mathrm{NS}}(y_t', \theta_t') \quad \text{for } r, t = 0 - N \\ &= \frac{(D/4 - \frac{10}{4})\,\delta_{rt}}{(y_r - y_t' - \theta_r\theta_t')^3} + \frac{\frac{3}{2}(\theta_r - \theta_t')\,\delta_{rt}}{(y_r - y_t' - \theta_r\theta_t')^2} \cdot \mathcal{F}_t^{\mathrm{NS}}(y_t', \theta_t') \\ &\quad + \frac{(\theta_r - \theta_t')\,\delta_{rt}}{y_r - y_t' - \theta_r\theta_t'} \cdot \partial_{y_t'}\mathcal{F}_t^{\mathrm{NS}}(y_t', \theta_t') + \frac{\frac{1}{2}\delta_{rt}}{y_r - y_t' - \theta_r\theta_t'} \cdot \mathcal{D}_{\theta_t'}\mathcal{F}_t^{\mathrm{NS}}(y_t', \theta_t') \\ &\quad + (\text{terms regular at } y_r = y_t'). \end{aligned} \tag{B14}$$

Based on Friedan–Martinec–Schenker's work,[9] we introduce the following ''BRST supercurrent operator in the π-representation'' [to be denoted by $J_r^\pi(y_r, \theta_r)$] defined by

$$J_r^{\pi}(y_r,\theta_r)\equiv -:\mathcal{C}_r(y_r,\theta_r)(\mathcal{F}_r^{(\vec{X}\cdot\vec{\Psi})}(y_r,\theta_r)+\tfrac{1}{2}\cdot\mathcal{F}_r^{(gh)}(y_r,\theta_r)):_\pi + \tfrac{3}{4}:\mathcal{D}_{\theta_r}(\mathcal{C}_r(y_r,\theta_r)\cdot\mathcal{D}_{\theta_r}\mathcal{C}_r(y_r,\theta_r)\cdot\beta_r(y_r,\theta_r)):_\pi \tag{B15}$$

$$[\equiv j_r^{\pi}(y_r)+\theta_r\cdot\mathcal{J}_r^{\pi}(y_r)], \tag{B16}$$

where we have used Eqs. (A42), (B2), and (B3). Therefore, $\mathcal{J}_r^{\pi}(y_r)$ in (B16) (to be referred to as the BRST currentino in the π-representation) is given by

$$\mathcal{J}_r^{\pi}(y_r)\equiv\sum_{m=0}^{2}\mathcal{J}_r^{\pi}(m;y_r)\quad\text{for } r=0-N, \tag{B17}$$

where each component $\mathcal{J}_r^{\pi}(m;y_r)$ is given by the following term (involving the mth power of γ):

$$-2\cdot\mathcal{J}_r^{\pi}(0;y_r)\equiv -2\cdot\mathcal{C}_r(y_r):T_r^{(\vec{X},\vec{\Psi})}(y_r):_\pi - :\mathcal{C}_r(y_r)\cdot T_r^{(\mathcal{BC})}(y_r):_\pi, \tag{B18}$$

$$-2\cdot\mathcal{J}_r^{\pi}(1;y_r)= -\mathcal{C}_r(y_r):T_r^{(\beta\gamma)}(y_r):_\pi+\gamma_r(y_r)\cdot\mathcal{F}_r^{(\vec{X},\vec{\Psi})}(y_r)+\tfrac{1}{2}:\gamma_r(y_r)\cdot\mathcal{F}_r^{(\beta\mathcal{C})}(y_r):_\pi, \tag{B19}$$

and

$$-2\cdot\mathcal{J}_r^{\pi}(2;y_r)=\tfrac{1}{2}:\gamma_r^2(y_r):_\pi\cdot\mathcal{B}_r(y_r). \tag{B20}$$

On the other hand, $\mathcal{J}_r(y_r)$ (to be referred to as the ''BRST currentino in the bo-representation'') is defined by

$$\mathcal{J}_r(y_r)\equiv\sum_{g=-1}^{1}\mathcal{J}_r^{(g)}(y_r)\quad\text{for } r=0-N, \tag{B21}$$

where each component $\mathcal{J}_r^{(g)}(y_r)$ [involving the factor $:\exp(g\sigma):$] is given by

$$\mathcal{J}_r^{(1)}(y_r)\equiv :\exp(+\sigma_r(y_r)):\left(T_r^{(\vec{X})}(y_r)+\sum_{j=0}^{4}T_r^{(\phi^j)}(y_r)+T_r^{(\phi)}(y_r)+T_r^{(\chi)}(y_r)\right): + :\exp(\sigma_r(y_r))(-\tfrac{1}{2}(\partial_{y_r}\sigma_r(y_r))^2+\tfrac{1}{2}\partial_{y_r}^2\sigma_r(y_r)):, \tag{B22}$$

$$\mathcal{J}_r^{(0)}(y_r)\equiv -\frac{1}{2}\left(\sum_{\pm}\sum_{j=0}^{4}\partial_{y_r}X_r^{\pm j}(y_r):\exp(\mp\phi_r^j(y_r)):\right):\exp(+\phi_r(y_r))::\exp(-\chi_r(y_r)): = -\tfrac{1}{2}\mathcal{F}_r^{(\vec{X}\cdot\vec{\Psi})}(y_r)\cdot\gamma_r(y_r), \tag{B23}$$

and

$$\mathcal{J}_r^{(-1)}(y_r)\equiv\tfrac{1}{4}:\exp(-\sigma_r(y_r))::\exp(+2\phi_r(y_r))::\exp(-2\chi_r(y_r)):. \tag{B24}$$

After carrying out extremely tedious but straightforward calculations (by using techniques in Ref. 5), we find that these *local* BRST currentinos in *three* representations are *not* equal to each other. Actually, they are related to each other by

$$\mathcal{J}_r^{\mathrm{R}}(y_r)-\mathcal{J}_r^{\mathrm{NS}}(y_r)=\frac{-D+6}{16}\cdot\frac{\mathcal{C}_r(y_r)}{(y_r-y_{0r})^2}+\frac{1}{4}\cdot\frac{\partial_{y_r}\mathcal{C}_r(y_r)}{y_r-y_{0r}}\quad \text{for } r=0-N$$
$$=\frac{1}{4}\cdot\partial_{y_r}\left(\frac{\mathcal{C}_r(y_r)}{y_r-y_{0r}}\right)\quad \text{in the special case } D=10, \tag{B25}$$

and

$$\mathcal{J}_r^{\mathrm{NS}}(y_r)=\mathcal{J}_r(y_r)+\tfrac{3}{2}\cdot\partial_{y_r}(\mathcal{C}_r(y_r)\partial_{y_r}\phi_r(y_r))\quad \text{for } r=0-N. \tag{B26}$$

[Incidentally, y_{0r} in Eq. (B25) is the one in the formulas (A45) and (A46) and it is the (disk or inlayed) coordinate of the operator (A23) in the R sector.]

At this stage, $\mathcal{Q}_r^{\pi}$ (to be referred to as the ''BRST chargino in the π-representation'') is defined by

$$\mathcal{Q}_r^{\pi}\equiv\oint_{y_{0r}}\frac{dy_r}{2\pi\sqrt{-1}}\mathcal{J}_r^{\pi}(y_r)\quad \text{for } r=0-N\quad [\text{see the formulas (B17)–(B20)}], \tag{B27}$$

while $\mathcal{Q}_r$ (to be referred to as the ''BRST chargino in the bo-representation'') is defined by

$$\mathcal{Q}_r\equiv\oint_{y_{0r}}\frac{dy_r}{2\pi\sqrt{-1}}\mathcal{J}_r(y_r)\quad \text{for } r=0-N\quad [\text{see the formulas (B21)–(B24)}). \tag{B28}$$

Fortunately, since BRST charginos [(B27) and (B28)] are obtained by integrating the BRST currentinos, we find from Eqs. (B25) (*in the case* $D=10$) and (B26) that

$$\mathcal{Q}_r^{\mathrm{NS}}=\mathcal{Q}_r^{\mathrm{R}}=\mathcal{Q}_r\quad \text{for } r=0-N. \tag{B29}$$

Thus BRST charginos in *three* representations are shown to be just *equal* to each other.

Comment: Equivalence of three BRST charginos is *not at all* the trivial result, as seen from the nonequalities of three BRST currentino's [i.e., Eqs. (B25) and (B26)]. Fortunately, we can *explicitly* prove the equality (B29) in the special case $D=10$. Incidentally, we give the following formula which will be useful in the actual calculations:

$$\mathcal{Q}_r^{(1)}\equiv\oint_{y_{0r}}\frac{dy_r}{2\pi\sqrt{-1}}\mathcal{J}_r^{(1)}(y_r)\quad (\text{in the bo-representation})$$
$$=\oint_{y_{0r}}\frac{dy_r}{2\pi\sqrt{-1}}\mathcal{C}_r(y_r):\left(T_r^{(\vec{X})}(y_r)+T_r^{(\vec{\Psi})}(y_r)+\frac{\frac{5}{8}\cdot\delta_\pi^R}{(y_r-y_{0r})^2}\right):_\pi\quad (\text{in the } \pi\text{-representation})$$
$$+\oint_{y_{0r}}\frac{dy_r}{2\pi\sqrt{-1}}\mathcal{C}_r(y_r)(T_r^{(\xi\eta)}(y_r)+T_r^{(\phi)}(y_r))$$
$$+\oint_{y_{0r}}\frac{dy_r}{2\pi\sqrt{-1}}:\mathcal{C}_r(y_r)\cdot\partial_{y_r}\mathcal{C}_r(y_r)\cdot\mathcal{B}_r(y_r):_\Sigma\quad (\text{in the } \Sigma\text{-representation}), \tag{B30}$$

where we have used the constant δ_π^R in (B7) and the operator $T_r^{(\xi\eta)}(y_r)$ defined by

$$T_r^{(\xi\eta)}(y_r)\equiv:\partial_{y_r}\xi_r(y_r)\cdot\eta_r(y_r):_\Sigma=T_r^{(\chi)}(y_r)\quad \text{for } r=0-N. \tag{B31}$$

[See Eq. (B5)].

In the following, we shall prove the inlaying identity relevant to the BRST chargino. In proving the inlaying identity of the BRST currentino, it is useful to use $\mathcal{J}_r^{\rm NS}(y_r)$, i.e., the ''BRST currentino in the NS-representation.'' After extremely tedious calculations using techniques in Ref. 5, we can prove the following inlaying identity:

$$\mathcal{J}_r^{\rm NS}(w_r)\overset{\mathcal{I}}{\Rightarrow} z_{r_s}^{(1)}(w_r):\mathcal{J}^{\rm NS}(z_{r_s}(w_r))$$
$$+\left(\frac{D}{8}-\frac{5}{4}\right)\cdot\frac{z_{r_s}^{(3)}(w_r)}{z_{r_s}^{(1)}(w_r)}+\left(-\frac{3D}{16}+\frac{15}{8}\right)\cdot\left(\frac{z_{r_s}^{(2)}(w_r)}{r_{r_s}^{(1)}(w_r)}\right)^2\cdot\frac{\mathcal{C}(z_{r_s}(w_r))}{z_{r_s}^{(1)}(w_r)}, \tag{B32a}$$

which is reduced (when $D=10$) to

$$\mathcal{J}_r^{\rm NS}(w_r)\overset{\mathcal{I}}{\Rightarrow} z_{r_s}^{(1)}(w_r):\mathcal{J}^{\rm NS}(z_{r_s}(w_r)) \quad \text{for } r=1-N \quad \text{and } D=10. \tag{B32b}$$

Therefore, the inlaying identity (B32b) [together with the equality (B29)] leads to

$$\mathcal{Q}_r^{\pi(r)}\equiv \oint_0 \frac{dw_r}{2\pi\sqrt{-1}}\mathcal{J}_r^{\pi(r)}(w_r) \quad \text{for } r=1-N$$
$$\overset{\mathcal{I}}{\Rightarrow}\mathcal{Q}[r]\equiv \oint_0 \frac{dw_r}{2\pi\sqrt{-1}} z_{r_s}^{(1)}(w_r)\cdot\mathcal{J}(z_{r_s}(w_r))$$
$$=\oint_{Z_{r_s}}\frac{dz_r}{2\pi\sqrt{-1}}\mathcal{J}(z_r)$$
$$\text{for } r=1-N \quad \text{and in the case } D=10, \tag{B33}$$

where $z_{r_s}(0)=Z_{r_s}$ is the rth puncture in GM-$\mathcal{FE}$ in (2.1). Inlaying identity (B33) shows that the rth *external* BRST chargino $\mathcal{Q}_r^{\pi(r)}$ [in the $\pi(r)$-representation] is inlayed into the *inlint* BRST chargino $\mathcal{Q}[r]$ in the rth punctured ring domain.

Comment: Although $\mathcal{Q}[r]$ was originally given by the integration over the rth disk coordinate w_r, it could subsequently be rewritten (*in the case* $D=10$) into the *integration over the complex variable* z (in the complex plane) along the path enclosing each puncture Z_{r_s} in the rth punctured ring domain. This result will play very important roles in proving the following theorem (B34).

Theorem [''Special conservation of the total BRST chargino'' (SCTC)]:

$${}_s\langle IV(1,2,...,N)|\left(\sum_{r=1}^{N}\mathcal{Q}_r^{\pi(r)}\right)=0 \quad [\text{to be referred to as the ''SCTC''}]. \tag{B34}$$

Proof: With the help of the inlaying identity (B33), the left-hand side of Eq. (B34) is just inlayed into

$${}_s\langle IV(1,2,...,N)|\left(\sum_{r=1}^{N}\mathcal{Q}[r]\right), \tag{B35a}$$

where we have that

$$\sum_{r=1}^{N}\mathcal{Q}[r]=\sum_{r=1}^{N}\oint_{Z_{r_s}}\frac{dz_r}{2\pi\sqrt{-1}}\mathcal{J}(z_r). \tag{B35b}$$

Since the inlint BRST currentino $\mathcal{J}(z)$ in Eq. (B35b) is *analytic* throughout various ring domains in the ''ICS *s*,'' we can *deform* the integration contours (in the complex *z* plane) of various *inlint* $\mathcal{Q}[r]$'s in Eq. (B35b). Thus the sum on the right-hand side of Eq. (B35b) is furthermore reduced to the integration enclosing each interacting point $\mathcal{Y}^s_{\pm\iota}$. Finally, the latter integration is found to be vanishing, since the integrand $\mathcal{J}(z)$ is *analytic* (as the function of *z*) *even* at any interacting point $\mathcal{Y}^s_{\pm\iota}$ [in GM-$\mathcal{FE}$ in (2.1)]. Thus we obtain that

$$\sum_{r=1}^{N} \mathcal{Q}[r]=0, \tag{B36}$$

which together with Eq. (B34b) leads to Eq. (B34). (Q.E.D.)

Comment: It should be noticed that the rth inlayed coordinate $z_{r\mathcal{J}}(w_r)$ (as the function of w_r) is *singular* at $w_r=\pm\sqrt{-1}$, i.e., at any interacting point, so that $D=10$ is essential in proving Eq. (B34).

At this stage, we can explain the *reason* why it is useful to introduce *three* $\mathcal{Q}_r^{\mathrm{NS}}$, $\mathcal{Q}_r^{\mathrm{R}}$, and $\mathcal{Q}_r$, which are fortunately just equal to each other in the special case $D=10$. [See Eq. (B29).] In proving (in Appendix D) that the rth *external* physical NS- or R-states (for $r=1-N$) are invariant under the *external* BRST chargino $\mathcal{Q}_r^{\pi(r)}$, we can use the OPEs[14] (D24) and (D28) among $:T_r^{(\vec{X},\vec{\Psi})}(y'_r):_{\pi(r)}$ in (D22) (which are given in the ''$\pi(r)$-representation'') and ''external physical vertex operators in the $\pi(r)$ sector'' {i.e., $V_r^{nb}(y_{0r})$ [for $\pi(r)=NS$] and $V_r^{nf}(y_{0r})\cdot S_{\{h\},r}(y_{0r})$ [for $\pi(r)=R$]}. On the other hand, the ''*inlint* BRST chargino $\mathcal{Q}[r]$ in the bo-representation'' is used in Eq. (B35b) in proving that the ''SCTC'' in (B34) holds for the inlaying vertex function $_{\mathcal{J}}\langle IV(1,...,N)|$ in the ''ICS *s*.'' [See Eqs. (B35) and (B36).]

Friedan–Martinec–Schenker[9] have proved that the BRST chargino is nilpotent when $D=10$. Therefore, their result together with Eq. (1.58) leads to

$$(\mathcal{Q}_r^{\pi(r)})^2=(\mathcal{Q}[r])^2=0 \quad \text{for } D=10 \quad \text{and } r=1-N$$
$$\text{(to be referred to as the ``nilpotency''),} \tag{B37}$$

and we have the following anticommutation relation:

$$\{\mathcal{Q}_r,\mathcal{B}_r(y_r)\}=T_r(y_r) \quad \text{for } r=0-N. \tag{B38a}$$

The anticommutation relation (B38) leads to

$$\left\{\mathcal{Q}_r,\frac{\mathcal{B}_{0,r}}{L_{0,r}}\right\}=1 \quad \text{for } r=0-N \quad \text{(to be referred to as the ``contacting formula'').} \tag{B38b}$$

APPENDIX C: FORMULAS OF (INVERSE) PICTURE-CHANGING OPERATORS

We give explicit formulas of the *local* (inverse-) picture-changing operator $X_r(y_r)$ [$Y_r(y_r)$] used in this paper. The ''BRST chargino $\mathcal{Q}_r$ in the bo-representation'' is defined by Eq. (B28), i.e.,

$$\mathcal{Q}_r\equiv\sum_{g=-1}^{1}\mathcal{Q}_r^{(g)}, \tag{C1}$$

where the component $\mathcal{Q}_r^{(g)}$ is given by

$$\mathcal{Q}_r^{(g)}\equiv\oint_{y_r}\frac{dy'_r}{2\pi\sqrt{-1}}\mathcal{J}_r^{(g)}(y'_r), \tag{C2}$$

$\mathcal{J}_r^{(g)}(y_r')$ (for each $g=\pm 1,0$) being given by the formulas (B22)–(B24). With the help of the BRST chargino $\mathcal{Q}_r$ in (C1), the *local* picture-changing operator $X_r(y_r)$ is defined by[9]

$$X_r(y_r)\equiv\{-2\mathcal{Q}_r,\xi_r(y_r)\}\equiv\{-2\mathcal{Q}_r,\Theta(\beta_r(y_r))\}=\sum_{g=-1}^{1}\{-2\mathcal{Q}_r^{(g)},\xi_r(y_r)\}\equiv\sum_{g=-1}^{1}X_r^{(g)}(y_r), \tag{C3}$$

so that $X_r(y_r)$ in (C3) is BRST invariant, i.e.,

$$[\mathcal{Q}_r,X_r(y_r)]=0. \tag{C4}$$

The component $X_r^{(g)}(y_r)$ in Eq. (C3) (for each $g=\pm 1,0$) is given by

$$X_r^{(1)}(y_r)=-2{:}\exp(+\sigma_r(y_r)){:}\,\partial_{y_r}\xi_r(y_r), \tag{C5}$$

$$X_r^{(0)}(y_r)=-{:}\exp(+\phi_r(y_r)){:}\left(\sum_{\pm}\sum_{j=0}^{4}\partial_{y_r}X_r^{\pm j}(y_r)\cdot{:}\exp(\mp\phi_r^j(y_r)){:}\right), \tag{C6}$$

and

$$\begin{aligned}X_r^{(-1)}(y_r)=&-\tfrac{1}{2}{:}\exp(+2\phi_r(y_r)){:}{:}\exp(-\sigma_r(y_r)){:}\,\partial_{y_r}\eta_r(y_r)\\&-\tfrac{1}{2}\cdot\partial_{y_r}({:}\exp(+2\phi_r(y_r)){:}{:}\exp(-\sigma_r(y_r)){:}\,\eta_r(y_r)).\end{aligned} \tag{C7}$$

On the other hand, the *local* inverse picture-changing operator $Y_r(y_r)$ (of conformal weight 0) is given by[6]

$$Y_r(y_r)\equiv-2{:}\exp(-2\phi_r(y_r)){:}{:}\exp(+\sigma_r(y_r)){:}\,\partial_{y_r}\xi_r(y_r)[=2{:}\exp(+\sigma_r(y_r)){:}\,\delta^{(1)}(\gamma_r(y_r))], \tag{C8a}$$

with

$$\delta(\gamma_r(y_r))\equiv{:}\exp(-\phi_r(y_r)){:}, \tag{C8b}$$

and $Y_r(y_r)$ in (C8) has explicitly been shown to be BRST invariant, i.e.,

$$[\mathcal{Q}_r,Y_r(y_r)]=0. \tag{C9}$$

Furthermore, we can easily find that

$$\vec{R}\cdot(Y_r(y_r')\cdot X_r(y_r))=1+O(y_r'-y_r), \tag{C10a}$$

since we have that

$$Y_r(y_r)\cdot X_r^{(g)}(y_r)=\delta_{g,-1}\quad\text{for }\ g=\pm 1,0. \tag{C10b}$$

We find from the definitions (1.28)–(1.31) that $X_r^{(g)}$'s (for $g=\pm 1,0$) and Y_r have the following quantum numbers:

$$\begin{aligned}P(X_r^{(g)})&=G(X_r^{(g)})+F(X_r^{(g)})=g+(1-g)=1,\\P(Y_r)&\equiv G(Y_r)+F(Y_r)=1-2=-1,\end{aligned} \tag{C11}$$

and

$$G(X_r^{(g)})-H(X_r^{(g)})=g-g=0,$$
$$G(Y_r)-H(Y_r)=1-1=0. \tag{C12}$$

[Incidentally, Eq. (C11) shows that any component $X^{(g)}$ has the *same* picturing number $P=1$. This is the reason why we have defined P by Eq. (1.31). It is to be noticed that many other authors have called our ''F'' the ''picture number.'']

Hereafter, we give the nonlocal picture-changing operators X_r^p [for any $p\in Z$ $(Z+\frac{1}{2})$ in the $\pi(r)=\mathrm{R}$ (NS) sector] *and* $X_{\pm,r}^p$ [for any $p\in Z+\frac{1}{2}$ in the case $\pi(r)=\mathrm{NS}$] of the rth *external* operator. We expand any primary operator $G_r(y_r)$ [of conformal weight $d(G)$] by

$$G_r(Y_r)=\sum_{n\in Z(Z+1/2)} G_{n,r}\cdot y_r^{-n-d(G)}, \tag{C13a}$$

where mode operators $G_{n,r}$ are given by

$$G_{n,r}=\oint_0\frac{dy_r}{2\pi\sqrt{-1}}y_r^{d(G)+n-1}\cdot G_r(y_r). \tag{C13b}$$

Then, the nonlocal picture-changing operator's X_r^p [$p\in Z$ $(Z+\frac{1}{2})$ in the $\pi(r)=\mathrm{R}$ (NS) sector] are defined respectively by

$$X_r^p\equiv\{-2\cdot\mathcal{Q}_r^{\pi(r)},\Theta(\beta_{p,r})\}\quad\text{for}\;\;p\in Z\quad(Z+\tfrac{1}{2}). \tag{C14}$$

Then, we find that

$$X_r^p=(\mathcal{F}_{p,r}^{(\vec{X}\cdot\vec{\Psi})}+\mathcal{F}_{p,r}^{(\beta\mathcal{C})}+\mathcal{F}_{p,r}^{(\mathcal{B}\gamma)})\cdot\delta(\beta_{p,r})-\tfrac{1}{2}\cdot\mathcal{B}_{2p,r}\cdot\delta^{(1)}(\beta_{p,r})$$
$$\text{for}\;\;p\in Z\quad(Z+\tfrac{1}{2})\;\;\text{in the}\;\;\pi(r)=\mathrm{R}\;\;(\mathrm{NS})\;\;\text{sector}, \tag{C15}$$

where we have used that

$$\mathcal{F}_{p,r}^{(\vec{X}\cdot\vec{\Psi})}=\oint_0\frac{dw_r}{2\pi\sqrt{-1}}w_r^{p+1/2}\left(\sum_{\pm}\sum_{j=0}^{4}\partial_{w_r}X_r^{\pm j}(w_r)\cdot\Psi_r^{\mp j}(w_r)\right), \tag{C16}$$

$$\mathcal{F}_{p,r}^{(\beta\mathcal{C})}=\oint_0\frac{dw_r}{2\pi\sqrt{-1}}w_r^{p+1/2}(-2\cdot\mathcal{C}_r(w_r)\partial_{w_r}\beta_r(w_r)-3\cdot\partial_{w_r}\mathcal{C}_r(w_r)\cdot\beta_r(w_r)), \tag{C17}$$

$$\mathcal{F}_{p,r}^{(\mathcal{B}\gamma)}=\oint_0\frac{dw_r}{2\pi\sqrt{-1}}w_r^{p+1/2}\cdot\gamma_r(w_r)\cdot\mathcal{B}_r(w_r), \tag{C18}$$

and

$$\mathcal{B}_{2p,r}=\oint_0\frac{dw_r}{2\pi\sqrt{-1}}w_r^{1+2p}\cdot\mathcal{B}_r(w_r). \tag{C19}$$

Operators $\delta^{(n)}(\beta_{p,r})$ (for $n=0,1$) in Eq. (C15) can be calculated by

$$\delta^{(n)}(\beta_{-f-3/2,r})=\sum_{m=0}^{\infty}\sum_{k=0}^{m}\frac{(n+m)!}{(m-k)!k!(n+k)!}$$

$$\times(\beta_{-f-3/2,r})^k\oint_0\frac{dw_r}{2\pi\sqrt{-1}}w_r^{-1+(n+k+1)f}\cdot\delta^{(n+k)}(\beta_r(w_r)), \tag{C20}$$

with

$$\beta_{-f-3/2,r}=\oint_0\frac{dw_r'}{2\pi\sqrt{-1}}w_r'^{-1-f}\cdot\beta_r(w_r') \tag{C21a}$$

and

$$\delta(\beta_r(y_r))\equiv:\exp(+\phi_r(y_r)):. \tag{C21b}$$

Furthermore, it can be shown [for $p\in Z$ $(Z+\frac{1}{2})$] that

$$Y_r(\pm\sqrt{-1})\cdot X_r^p\cdot Y_r(\pm\sqrt{-1})=Y_r(\pm\sqrt{-1}), \tag{C22}$$

$$X_r^p\cdot Y_r(\pm\sqrt{-1})\cdot X_r^p=X_r^p \tag{C23}$$

and

$$Y_r(+\sqrt{-1})\cdot Y_r(-\sqrt{-1})=Y_r(-\sqrt{-1})\cdot Y_r(+\sqrt{-1}). \tag{C24}$$

Exclusively in the $\pi(r)=\mathrm{NS}$ sector, we construct the nonlocal picture-changing operators $X_{\pm,r}^p$ (for any $p\in Z+\frac{1}{2}$) as follows:

$$X_{\pm,r}^p\equiv\{-2\cdot\mathcal{Q}_r,\Theta(\beta_{\pm,r}^p)\}\quad(\text{for }\ p\in Z+\tfrac{1}{2}), \tag{C25}$$

with

$$\beta_{\pm,r}^p\equiv\tfrac{1}{2}((\pm\sqrt{-1})^p\cdot\beta_{-p,r}+(\pm\sqrt{-1})^{-p}\cdot\beta_{p,r}). \tag{C26}$$

Then, we find that

$$X_{\pm,r}^p=((\mathcal{F}^{(\vec{X}\cdot\vec{\Psi})})_{\pm,r}^p+(\mathcal{F}^{(\beta\mathcal{C})})_{\pm,r}^p+(\mathcal{F}^{(\mathcal{B}\gamma)})_{\pm,r}^p)\cdot\delta(\beta_{\pm,r}^p)-\tfrac{1}{2}\cdot\mathcal{B}_{\pm,r}^{\sharp 2p}\cdot\delta^{(1)}(\beta_{\pm,r}^p), \tag{C27}$$

with

$$(\mathcal{F}^{(M)})_{\pm,r}^p\equiv\tfrac{1}{2}((\pm\sqrt{-1})^p\cdot\mathcal{F}_{-p,r}^{(M)}+(\pm\sqrt{-1})^{-p}\cdot\mathcal{F}_{+p,r}^{(M)})\quad\text{for }(M)=(\vec{X}\cdot\vec{\Psi}),(\beta\mathcal{C}),(\mathcal{B}\gamma). \tag{C28a}$$

[Various operators in Eq. (C28) have been defined by Eqs. (C16)–(C18).] In particular, we find that

$$(\mathcal{F}^{(\mathcal{B}\gamma)})_{\pm,r}^p=\gamma_{\pm,r}^p\cdot\mathcal{B}_{\pm,r}^{\sharp 2p}+\gamma_{\mp,r}^p\cdot\mathcal{B}_{\oplus,r}^{2p}+(\text{terms which do not involve }\ \gamma_{\pm,r}^p), \tag{C28b}$$

where we have used that

$$\gamma_{\pm,r}^p\equiv(\pm\sqrt{-1})^p\cdot\gamma_{-p,r}+(\pm\sqrt{-1})^{-p}\cdot\gamma_{p,r}, \tag{C29}$$

$$\mathcal{B}_{\pm,r}^{\sharp 2p}\equiv\tfrac{1}{2}(\mathcal{B}_{0,r}\pm\mathcal{B}_{\natural,r}^{2p})=\tfrac{1}{2}\mathcal{B}_{0,r}+\tfrac{1}{4}((\pm\sqrt{-1})^{2p}\mathcal{B}_{-2p,r}+(\pm\sqrt{-1})^{-2p}\mathcal{B}_{+2p,r}), \tag{C30}$$

and

$$\mathcal{B}^{2p}_{\oplus,r} \equiv \tfrac{1}{4}(\mathcal{B}_{+2p,r} + \mathcal{B}_{-2p,r}). \tag{C31}$$

With the help of $X^p_{\pm,r}$ in (C27), we can prove (for any $p \in Z + \frac{1}{2}$) that

$$X^p_{\pm,r} \cdot Y_r(\pm\sqrt{-1}) \cdot X^p_{\pm,r} = X^p_{\pm,r}, \tag{C32}$$

$$Y_r(\pm\sqrt{-1}) \cdot X^p_{\pm,r} \cdot Y_r(\pm\sqrt{-1}) = Y_r(\pm\sqrt{-1}), \tag{C33}$$

$$X^p_{\pm,r} \cdot Y_r(\mp\sqrt{-1}) = Y_r(\mp\sqrt{-1}) \cdot X^p_{\pm,r}, \tag{C34}$$

$$\begin{aligned} Y_r(+\sqrt{-1}) \cdot Y_r(-\sqrt{-1}) &= Y_r(-\sqrt{-1}) \cdot Y_r(+\sqrt{-1}) \\ &= Y_r(+\sqrt{-1}) \cdot Y_r(-\sqrt{-1})(X^p_{+,r} \cdot X^p_{-,r}) Y_r(+\sqrt{-1}) \cdot Y_r(-\sqrt{-1}), \end{aligned} \tag{C35}$$

and

$$Y_r(\pm\sqrt{-1})[X^p_{+,r}, X^p_{-,r}] Y_r(\pm\sqrt{-1}) = 0. \tag{C36}$$

We remark that X^p_r in (C14) [for $p \in Z$ $(Z + \frac{1}{2})$] and $X^p_{\pm,r}$ in (C25) (for $p \in Z + \frac{1}{2}$) satisfy respectively that

$$[\mathcal{Q}_r, X^p_r] = 0 \quad \text{for any} \quad p \in Z, Z + \tfrac{1}{2}, \tag{C37a}$$

and

$$[\mathcal{Q}^{\mathrm{NS}}_r, X^p_{\pm,r}] = 0 \quad \text{for any} \quad p \in Z + \tfrac{1}{2}. \tag{C37b}$$

Furthermore, we have found the following "commutability:"

$$[\mathcal{B}_{0,r}, X^p_r] = [L_{0,r}, X^p_r] = 0 \quad \text{for any} \quad p \in Z, Z + \tfrac{1}{2}, \quad \text{(to be referred to as the "commutability")} \tag{C38}$$

and

$$\left[\mathcal{B}_{0,r}, \sum_{\pm} \frac{1}{2} \cdot X^p_{\pm,r} \cdot X^p_{\mp,r}\right] = \left[L_{0,r}, \sum_{\pm} \frac{1}{2} \cdot X^p_{\pm,r} \cdot X^p_{\mp,r}\right] = 0 \quad \text{for any} \quad p \in Z + \tfrac{1}{2} \quad \text{(to be referred to as the "commutability")}, \tag{C39}$$

together with

$$[L_{0,r}, \delta(\beta^p_{+,r}) \cdot \delta(\beta^p_{-,r})] = 0 \quad \text{for any} \quad p \in Z + \tfrac{1}{2}. \tag{C40}$$

APPENDIX D: PHYSICAL NEVEU–SCHWARZ AND RAMOND STATES

When the rth (external) open NS superstring is in some physical state (specified by the quantum number, say, b), this (Grassman *odd*) *physical* NS state in 0 picture can be described by[14]

$$|b(0)\rangle_r \equiv :\exp(+\sigma_r(0))::\exp(-\phi_r(0)):V^{0b}_r(0)|\mathcal{Q}_r = 0\rangle_r \quad \text{for} \quad \pi(r) = \mathrm{NS}, \tag{D1}$$

where any (Grassman *odd*) *physical* vertex operator $V_r^{0b}(0)$ [having the conformal weight $(-\Sigma_{j=0}^4 p_b^{+j}\cdot p_b^{-j}+N_b)$] has been explicitly constructed (in our papers in Ref. 14) as the functions of $\vec{X}$ and $\vec{\Psi}$. In Eq. (D1), the standard ket state $|\mathcal{Q}_r=0\rangle_r$ is given by

$$|\mathcal{Q}_r=0\rangle_r \equiv \prod_{\varphi=\vec{X},\vec{\phi},\sigma,\phi,\chi} |p_r(\varphi)=0\rangle_r, \tag{D2a}$$

with

$$|p_r(\vec{\phi})=0\rangle \equiv \prod_{j=0}^{4} |p_r(\phi^j)=0\rangle_r. \tag{D2b}$$

With the help of the ''operator product expansions'' (OPEs) [i.e., (D24)–(D26) to be given later,] we can prove that (Grassman *odd*) physical state $|b(0)\rangle_r$ in (D1) has the following properties:

$$\mathcal{Q}_r^{\rm NS}|b(0)\rangle_r=\mathcal{B}_{0,r}|b(0)\rangle_r=0 \quad [\text{see } \mathcal{Q}_r^{\rm NS} \text{ in (B27)}] \tag{D3}$$

in the *on-shell* limit, i.e.,

$$-\sum_{j=0}^{4} p_b^{+j}\cdot p_b^{-j}+N_b \to \frac{1}{2}. \tag{D4}$$

We notice that the *physical* (Grassman *odd*) NS states in 1 picture [which are denoted by $|b(1)\rangle_r$] can be constructed by using $|b(0)\rangle_r$'s in (D1) as[14]

$$|b(1)\rangle_r = X_{\pm,r}^{1/2}|b(0)\rangle_r = X_r^{-1/2}|b(0)\rangle_r \quad \text{for } \pi(r)=\text{NS}$$

$$=:\exp(+\sigma_r(0)):V_r^{1b}(0)|\mathcal{Q}_r=0\rangle_r-\frac{1}{2}:\exp(+\phi_r(0)):\eta_r(0):V_r^{0b}(0)|\mathcal{Q}_r=0\rangle_r, \tag{D5}$$

which has[11] the following properties:

$$P_r(1)|b(1)\rangle_r=|b(1)\rangle_r \quad [\text{see } P_r(1) \text{ in (1.7)}] \tag{D6}$$

and

$$\mathcal{Q}_r^{\rm NS}|b(1)\rangle_r=\mathcal{B}_{0,r}|b(1)\rangle_r=0 \quad [\text{see } \mathcal{Q}_r^{\rm NS} \text{ in (B27)}]. \tag{D7}$$

Incidentally, the conformal weights ''*d*'' of each factor in Eq. (D5) are given respectively by

$$d(e^{\sigma})+d(V^{1b})=-1+2\left(-\sum_{j=0}^{4} p_{b+}^{j}\cdot p_b^{-j}+N_b\right)=0 \tag{D8}$$

and

$$d(\gamma)+d(V^{0b})=-\frac{1}{2}+\left(-\sum_{j=0}^{4} p_b^{+j}\cdot p_b^{-j}+N_b\right)=0, \tag{D9}$$

where we have used the *on-shell limit* (D4). Finally, we remark that the *physical* (Grassman *odd*) NS states $|b(0)\rangle_r$ in (D1) and $|b(1)\rangle_r$ in (D5) are related to each other by the *local* (inverse) picture-changing operator by

$$\lim_{w\to 0} \vec{R}\cdot X_r(w)\cdot|b(0)\rangle_r=|b(1)\rangle_r \tag{D10}$$

and

$$\lim_{w\to 0} \vec{R}\cdot Y_r(w)\cdot|b(1)\rangle_r=|b(0)\rangle_r . \tag{D11}$$

On the other hand, when the rth open R superstring is in some *physical* state (specified by the quantum number, say, f), this (Grassman *even*) *physical* R state in $\pm\frac{1}{2}$ picture can be described by[14]

$$|f(+\tfrac{1}{2})\rangle_r \equiv :\exp(+\sigma_r(0))::\exp(-\tfrac{1}{2}\phi_r(0)):V_r^{1f}(0)\cdot S_{\{h\},r}(0)|\mathcal{Q}_r=0\rangle_r , \tag{D12}$$

$$|f(-\tfrac{1}{2})\rangle_r \equiv :\exp(+\sigma_r(0))::\exp(-\tfrac{3}{2}\phi_r(0)):V_r^{0f}(0)\cdot S_{\{h\},r}(0)|\mathcal{Q}_r=0\rangle_r , \quad \text{for} \;\; \pi(r)=R, \tag{D13}$$

where Grassman *odd* (*even*) $V_r^{1f}(0)\cdot S_{\{h\},r}(0)$'s $[V_r^{0f}(0)\cdot S_{\{h\},r}(0)$'s$]$ have been explicitly constructed (in our papers in Ref. 14) as the functions of $\vec{X}$ and $\vec{\Psi}$. [We have also proved the OPEs given later by Eqs. (D28)–(D31).] Both of Eqs. (D12) and (D13) have zero conformal weight in the following *on-shell* limit;

$$-\sum_{j=0}^{4} p_f^{+j}\cdot p_f^{-j}+N_f\rightarrow 0. \tag{D14}$$

Furthermore, we find that the *physical* (Grassman *even*) $|f(+1/2)\rangle_r$ in (D12) is related with the *physical* $|f(-1/2)\rangle_r$ in (D13) by

$$|f(+1/2)\rangle_r \equiv X_r^0|f(-1/2)\rangle_r . \tag{D15}$$

Therefore, we find that

$$\begin{aligned} P_r(\tfrac{1}{2})|f(+\tfrac{1}{2})\rangle_r &= X_r^0\cdot Y_r(-\sqrt{-1})|f(+\tfrac{1}{2})_r \quad [\text{from Eq. (1.7)}] \\ &= X_r^0\cdot Y_r(-\sqrt{-1})\cdot X_r^0|f(-\tfrac{1}{2})\rangle_r \quad [\text{from Eq. (D15)}] \\ &= X_r^0|f(-\tfrac{1}{2})\rangle_r \quad [\text{from Eq. (C23)}] \\ &= |f(+\tfrac{1}{2})\rangle_r \quad [\text{from Eq. (D15)}]. \end{aligned} \tag{D16}$$

Incidentally, with respect to the *local* operators $X_r(0)$ and $Y_r(0)$, we find that

$$|f(+\tfrac{1}{2})\rangle_r=\lim_{w\to 0}\vec{R}\cdot X_r(w)|f(-\tfrac{1}{2})\rangle_r \tag{D17}$$

and

$$\begin{aligned} \lim_{w\to 0}\vec{R}\cdot Y_r(w)|f(+1/2)\rangle_r &= :\exp(+2\cdot\sigma_r(0))::\exp(-\tfrac{5}{2}\cdot\phi_r(0)) \\ &\quad \times :\partial\xi_r(0)\cdot V_r^{1f}(0)\cdot S_{\{h\},r}(0)|\mathcal{Q}_r=0\rangle_r . \end{aligned} \tag{D18}$$

Physical R states (D12) and (D13) can be shown to have the following properties:

$$\mathcal{Q}_r^{\mathrm{R}}|f(\pm\tfrac{1}{2})\rangle_r=\mathcal{B}_{0,r}|f(\pm\tfrac{1}{2})\rangle_r=0 \quad [\text{see } \mathcal{Q}_r^{\mathrm{R}} \text{ in (B27)}]. \tag{D19}$$

In Eqs. (D12) and (D13), we have used the *external* FMS spinor $S_{\{h\},r}(0)$ in (A24) (for $r=1-N$). On the other hand, the rth *physical* vertex operators V_r^{0f} (V_r^{1f}) and V_r^{0b} (V_r^{1b}) [given as functions of $(X_r^{\pm j},\Psi_r^{\pm j})$'s] can be obtained from previously obtained[14] *physical* vertex operators V^{0f} (V^{1f}) and V^{0b} (V^{1b}) [given as functions of $(X^{\pm j},\Psi^{\pm j})$'s] by

$$X^{\pm j}(z)\to X_r^{\pm j}(y_r) \tag{D20}$$

and

$$\Psi^{\pm j}(z)\equiv :\exp(\pm\phi^j(z)):\to\Psi_r^{\pm j}(y_r)\equiv :\exp(\pm\phi_r^j(y_r)):. \tag{D21}$$

(See our papers in Ref. 14 for more details.)

In the following, we give various OPEs which have been derived in our previous papers.[14] The ''stress operator $:T_r^{(\vec{X},\vec{\Psi})}(y_r'):_\pi$ in the π-representation'' and the ''stressino operator $\mathcal{F}_r^{(\vec{X}\cdot\vec{\Psi})}\times(y_r')$'' are defined respectively by

$$\begin{aligned} :T_r^{(\vec{X},\vec{\Psi})}(y_r'):_\pi &\equiv :T_r^{(\vec{X})}(y_r'):_\pi + :T_r^{(\vec{\Psi})}(y_r'):_\pi \quad [\text{see Eq. (B4)}] \\ &= \sum_{j=0}^{4} :\partial_{y_r'}X_r^{+j}(y_r')\cdot\partial_{y_r'}X_r^{-j}(y_r'):_\pi \\ &\quad + \sum_{\pm}\sum_{j=0}^{4}\frac{1}{2}:\partial_{y_r'}\Psi_r^{\pm j}(y_r')\cdot\Psi_r^{\mp j}(y_r'):_\pi \quad \text{for } \pi=\text{NS,R} \end{aligned} \tag{D22}$$

and

$$\mathcal{F}_r^{(\vec{X}\cdot\vec{\Psi})}(y_r')\equiv\sum_{\pm}\sum_{j=0}^{4}\partial_{y_r'}X_r^{\pm j}(y_r')\cdot\Psi_r^{\mp j}(y_r') \tag{D23a}$$

$$\equiv\begin{cases}\displaystyle\sum_{n=-\infty}^{\infty}\mathcal{F}_{n,r}^{(\vec{X}\cdot\vec{\Psi})}\cdot y_r'^{\,-n-3/2} & \text{for } \pi(r)=R \\ \displaystyle\sum_{n=-\infty}^{\infty}\mathcal{F}_{n+1/2,r}^{(\vec{X}\cdot\vec{\Psi})}\cdot y_r'^{\,-n-2} & \text{for } \pi(r)=\text{NS}.\end{cases} \tag{D23b}$$

In Ref. 12, we have derived OPEs of the physical vertex operator V_r^{nf} (which is normal ordered in the R-representation) as well as of the *physical* vertex operator V_s^{nb} (which is normal ordered in the NS-representation). In the case $\pi(r)=\text{NS}$, we find the following OPEs with respect to the stress operator (which is normal ordered in the NS-representation):

$$\begin{aligned} &\vec{R}\cdot(:T_r^{(\vec{X},\vec{\Psi})}(y_r'):_{\text{NS}}\cdot V_r^{nb}(y_{0r})) \quad \text{for } n=0,1 \\ &= \frac{1}{(y_r'-y_{0r})^2}\left(\frac{1}{2}\cdot\delta_{n,0}-\sum_{j=0}^{4}p_b^{+j}\cdot p_b^{-j}+N_b\right)V_r^{nb}(y_{0r}) \\ &\quad + \frac{1}{y_r'-y_{0r}}\cdot\partial_{y_{0r}}V_r^{nb}(y_{0r})+O(1), \end{aligned} \tag{D24a}$$

where we have derived[14] that

$$:T_r^{(\vec{X},\vec{\Psi})}(y_r'):_{\text{NS}}=:T_r^{(\vec{X})}(y_r'):_\Sigma+\sum_{j=0}^{4}T_r^{(\phi^j)}(y_r'). \tag{D24b}$$

On the other hand, OPEs with respect to the stressino operator are given by

$$\vec{R}\cdot(\mathcal{F}_r^{(\vec{X},\vec{\Psi})}(y_r')\cdot V_r^{0b}(y_{0r}))=\frac{1}{y_r'-y_{0r}}\cdot V_r^{1b}(y_{0r})+O(1) \tag{D25}$$

and

$$\vec{R}\cdot(\mathcal{F}_r^{(\vec{X},\vec{\Psi})}(y_r')\cdot V_r^{1b}(y_{0r}))$$
$$=\frac{2}{(y_r'-y_{0r})^2}\left(\sum_{j=0}^{4} p_b^{+j}\cdot p_b^{-j}+N_b\right)V_r^{0b}(y_{0r})+\frac{1}{y_r'-y_{0r}}\cdot\partial_{y_{0r}}V_r^{0b}(y_{0r})+O(1). \tag{D26}$$

In OPEs (D24) and (D26), the number operator N_b (in the NS sector) is defined by

$$N_b=\sum_{\pm}\sum_{j=1}^{4}\sum_{n=1}^{\infty}\left(n\cdot e_b(n,\pm j)+\left(n-\frac{1}{2}\right)\cdot\tilde{e}_b\left(n-\frac{1}{2},\pm j\right)\right). \tag{D27}$$

On the other hand, we find in the case $\pi(r)=\mathrm{R}$ the following OPEs with respect to the stress operator (which is normal ordered in the R-representation):

$$\vec{R}\cdot(:T_r^{(\vec{X},\vec{\Psi})}(y_r'):_R\cdot(V_r^{nf}(y_{0r})\cdot S_{\{h\},r}(y_{0r})))\quad\text{for } n=0,1$$
$$=\frac{1}{(y_r'-y_{0r})^2}\left(-\sum_{j=0}^{4} p_f^{+j}\cdot p_f^{-j}+N_f\right)(V_r^{nf}(y_{0r})\cdot S_{\{h\},r}(y_{0r}))$$
$$+\frac{1}{y_r'-y_{0r}}\cdot\partial_{y_{0r}}(V_r^{nf}(y_{0r})\cdot S_{\{h\},r}(y_{0r}))+O(1), \tag{D28a}$$

where we have derived[14] that

$$:T_r^{(\vec{X},\vec{\Psi})}(y_r'):_{\mathrm{R}}=:T_r^{(\vec{X})}(y_r'):_{\Sigma}+\sum_{j=0}^{4} T_r^{(\phi^j)}(y_r')-\frac{\frac{5}{8}}{(y_r'-y_{0r})^2}, \tag{D28b}$$

while the number operator N_f (in the R sector) is defined by

$$N_f\equiv\sum_{n=1}^{\infty}\sum_{\pm}\sum_{j=1}^{4} n(e_f(n,\pm j)+\tilde{e}_f(n,\pm j)). \tag{D29}$$

Furthermore, we have found following OPEs with respect to the stressino operator:

$$\vec{R}\cdot(\mathcal{F}_r^{(\vec{X}\cdot\vec{\Psi})}(y_r')\cdot(V_r^{0f}(y_{0r})\cdot S_{\{h\},r}(y_{0r})))$$
$$=\frac{1}{(\sqrt{y_r'-y_{0r}})^3}(V_r^{1f}(y_{0r})\cdot S_{\{h\},r}(y_{0r}))$$
$$+\frac{1}{\sqrt{y_r'-y_{0r}}}\cdot(\hat{V}_r^{3f}(y_{0r})\cdot S_{\{h\},r}(y_{0r}))+O(\sqrt{y_r'-y_{0r}}) \tag{D30}$$

and

$$\vec{R}\cdot(\mathcal{F}_r^{(\vec{X}\cdot\vec{\Psi})}(y_r')(V_r^{1f}(y_{0r})\cdot S_{\{h\},r}(y_{0r})))$$
$$=\frac{-1}{(\sqrt{y_r'-y_{0r}})^3}\left(-\sum_{j=0}^{4}p_f^{+j}\cdot p_f^{-j}+N_f\right)(V_r^{0f}(y_{0r})\cdot S_{\{h\},r}(y_{0r}))$$
$$-\frac{1}{\sqrt{y_r'-y_{0r}}}\cdot\partial_{y_{0r}}(V_r^{0f}(y_{0r})\cdot S_{\{h\},r}(y_{0r}))+\frac{1}{\sqrt{y_r'-y_{0r}}}(\hat{V}_r^{4f}(y_{0r})\cdot S_{\{h\},r}(y_{0r}))$$
$$+O(\sqrt{y_r'-y_{0r}}). \tag{D31}$$

At this stage, we remark that Eq. (B6) has been derived from Eqs. (D24b) and (D28b) (in the case $D=10$), together with

$$:T_r^{(\mathcal{BC})}(y_r'):_\pi=T_r^{(\sigma)}(y_r')+\frac{\delta_\pi^R}{(y_r'-y_{0r})^2} \tag{D32}$$

and

$$:T_r^{(\beta\gamma)}(y_r'):_\pi=T_r^{(\phi)}(y_r')+T_r^{(\chi)}(y_r')-\frac{\frac{3}{8}\delta_\pi^R}{(y_r'-y_{0r})^2}, \tag{D33}$$

where we have used Eq. (B7). The presence of the last term of Eq. (D32) is induced by the following contraction in the R sector:

$$\overset{\sqcap\,\mathrm{R}\,\sqcap}{\mathcal{B}_r(y_r)\mathcal{C}_t(y_t')}\equiv\frac{\delta_{rt}}{y_r-y_t'}\cdot\left(\frac{y_t'-y_{0t}}{y_r-y_{0r}}\right)\quad[\text{see Eq. (A46)}]. \tag{D34}$$

APPENDIX E: SMALL GLUING VERTEX FUNCTINO, LARGE GLUING VERTEX FUNCTION, GLUING IDENTITIES, AND GLUING RELATIONS

In this appendix, we derive the *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ (which is useful in tree calculations), the *large* gluing vertex function $|V_L(\gamma,\delta)\rangle$ (which is useful in loop calculations), *gluing identities* and *gluing relations* (which are valid within the *small* or *large* Hilbert space). First, the *large* (*small*) Hilbert space[9] is the space with [without] ξ_0-mode, so that the difference among the *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ and the *large* gluing vertex function $|V_L(\gamma,\delta)\rangle$ exists only in the hilberting modes $[\chi,\ \xi,\ \eta]$.

In order to discuss ξ_0-mode, we use the following inlaying operator $\mathcal{W}_r^\chi[z_{r_\mathcal{J}}(w_r)]$ ''in $\xi\eta$-mode,'' rather than $W_r^\chi[z_{r_\mathcal{J}}(w_r)]$ (2.5) ''in χ-mode;''

$$\mathcal{W}_r^\chi[z_{r_\mathcal{J}}(w_r)]\equiv\exp\left(-\oint_0\frac{dw_r}{2\pi\sqrt{-1}}\oint_0\frac{dw_r'}{2\pi\sqrt{-1}}\xi_r(w_r;+)\,\eta_r(w_r';+)\right.$$
$$\left.\times\left(\frac{z_{r_\mathcal{J}}^{(1)}(w_r)}{z_{r_\mathcal{J}}(w_r)-z_{r_\mathcal{J}}(w_r')}-\frac{1}{w_r-w_r'}\right)\right):\exp\left(\oint_0\frac{dw_r}{2\pi\sqrt{-1}}(z_{r_\mathcal{J}}^{(1)}(w_r)\cdot\eta(z_{r_\mathcal{J}}(w_r))\right.$$
$$\left.\cdot\,\xi_r(w_r';+)+\xi(z_{r_\mathcal{J}}(w_r))\cdot\eta_r(w_r';+))\right):_\Sigma. \tag{E1}$$

Replacing $W_r^\chi[\cdots]$ in (2.7) with $\mathcal{W}_r^\chi[\cdots]$ in (E1) (''in $\xi\eta$-mode''), we can construct *another* inlaying N-vertex function ${}_\mathcal{J}\langle IV^\chi(1,2,...,N)|$ ''in $\xi\eta$-mode,'' which can be shown to satisfy the following inlaying identities for the hilbertino ξ and anti-hilbertino η:

$$_s\langle IV^\chi(1,2,...,N)|\mathcal{G}_r(w')={}_s\langle IV^\chi(1,2,...,N)|(z_{rs}^{(1)}(w'))^{d(\mathcal{G})}\mathcal{G}(z_{rs}(w'))$$
$$\text{for } \mathcal{G}=\xi,\eta \quad \text{with } d(\xi)=0,\ d(\eta)=1. \tag{E2}$$

Furthermore, we can derive from Eq. (E1) the following results:

$$\mathcal{W}_\gamma^\chi[w_\gamma]|p(\chi)=0\rangle=\exp\left(\sum_{n=1}^{\infty}\eta_{-n}\cdot\xi_{n,\gamma}+\sum_{n=0}^{\infty}\xi_{-n}\cdot\eta_{n,\gamma}\right)|p(\chi)=0\rangle \tag{E3}$$

and

$$(\langle\tilde{p}(\chi)=0|\tilde{\xi}_0)\tilde{\mathcal{W}}_\delta^\chi\left[\frac{-1}{w_\delta}\right]=(\langle\tilde{p}(\chi)=0|\tilde{\xi}_0)\exp\left(-\sum_{n=0}^{\infty}(-)^n\cdot\tilde{\eta}_n\cdot\xi_{n\delta}+\sum_{n=1}^{\infty}(-)^n\cdot\tilde{\xi}_n\cdot\eta_{n\delta}\right). \tag{E4}$$

It should be noticed that there exist zero-modes $\xi_0\cdot\eta_{0,\gamma}$ and $\tilde{\eta}_0\cdot\xi_{0,\delta}$ within the exponential factor on the right-hand side of Eqs. (E3) and (E4), respectively. Then, $\langle\mathcal{V}_S^\chi(\gamma,\delta)|$ is defined by

$$\langle\mathcal{V}_S^\chi(\gamma,\delta)|=\langle\mathcal{V}_S^\chi(\delta,\gamma)|\equiv({}_\gamma\langle p_\gamma(\chi)=0|\xi_{0,\gamma})({}_\delta\langle p_\delta(\chi)=0|\xi_{0,\delta})$$
$$\times(\langle p(\chi)=0|\xi_0)\mathcal{W}_\delta^\chi\left[\frac{-1}{w_\delta}\right]\cdot\mathcal{W}_\gamma^\chi[w_\gamma](|p(\chi)=0\rangle). \tag{E5}$$

Substituting Eqs. (E3) and (E4) into Eq. (E5), we find that

$$\langle\mathcal{V}_S^\chi(\gamma,\delta)|=({}_\gamma\langle p_\gamma(\chi)=0|\xi_{0,\gamma})({}_\delta\langle p_\delta(\chi)=0|\xi_{0,\delta})\exp\left(\sum_{n=1}^{\infty}(-)^n\cdot\xi_{n\delta}\cdot\eta_{n\gamma}-\sum_{n=1}^{\infty}(-)^n\cdot\eta_{n\delta}\cdot\xi_{n\gamma}\right), \tag{E6}$$

which does *not* have any zero-mode contribution within the exponential factor. [Hereafter, $\langle\mathcal{V}_S^\chi(\gamma,\delta)|$ will be referred to as the *small* two-point vertex function ''in $\xi\eta$-mode.''] On the other hand, with the help of

$$({}_\gamma\langle p_\gamma(\chi)=0|\xi_{0,\gamma})({}_\delta\langle p_\delta(\chi)=0|\xi_{0,\delta})=({}_\gamma\langle p_\gamma(\chi)=0|)({}_\delta\langle p_\delta(\chi)=0|\xi_{0,\delta})\exp(-\eta_{0,\delta}\cdot\xi_{0,\gamma})\cdot\xi_{0,\gamma}$$
$$=({}_\gamma\langle p_\gamma(\chi)=0|\xi_{0,\gamma})({}_\delta\langle p_\delta(\chi)=0|)\exp(+\xi_{0,\delta}\cdot\eta_{0,\gamma})\cdot\xi_{0,\delta}, \tag{E7}$$

$\langle\mathcal{V}_S^\chi(\gamma,\delta)|$ in (E6) ''in $\xi\eta$-mode'' can be rewritten into

$$\langle\mathcal{V}_S^\chi(\gamma,\delta)|=\langle\nu_\mathcal{L}^\chi(\gamma,\delta)|\xi_{0,\gamma}=\langle\nu_\mathcal{L}^\chi(\gamma,\delta)|\xi_{0,\delta}, \tag{E8}$$

$\langle\nu_\mathcal{L}^\chi(\gamma,\delta)|$ being given by

$$\langle\nu_\mathcal{L}^\chi(\gamma,\delta)|=\langle\nu_\mathcal{L}^\chi(\delta,\gamma)|\equiv({}_\gamma\langle p_\gamma(\chi)=0|)({}_\delta\langle p_\delta(\chi)=0|\xi_{0,\delta})$$
$$\times\exp\left(\sum_{n=1}^{\infty}(-)^n\cdot\xi_{n\delta}\cdot\eta_{n\gamma}-\sum_{n=0}^{\infty}(-)^n\cdot\eta_{n\delta}\cdot\xi_{n\gamma}\right), \tag{E9}$$

which has the zero-mode contribution within the exponential factor. [Hereafter, $\langle\nu_\mathcal{L}^\chi(\gamma,\delta)|$ will be referred to as the *large* two-point vertex functino ''in $\xi\eta$-mode.''] Furthermore, with the help of

$$({}_\gamma\langle p_\gamma(\chi)=0|)({}_\delta\langle p_\delta(\chi)=0|)\xi_{0,\delta}(1-\eta_{0,\delta}\cdot\xi_{0,\gamma})\,\eta_{0,\delta}$$
$$=({}_\gamma\langle p_\gamma(\chi)=0|)({}_\delta\langle p_\delta(\chi)=0|)$$
$$=-({}_\gamma\langle p_\gamma(\chi)=0|)({}_\delta\langle p_\delta(\chi)=0|)\xi_{0,\delta}(1-\eta_{0,\delta}\cdot\xi_{0,\gamma})\,\eta_{0,\gamma}, \tag{E10}$$

we find that

$$\langle \nu_{\mathcal{L}}^{\chi}(\gamma,\delta)|\eta_{0,\delta} = -\langle \nu_{\mathcal{L}}^{\chi}(\gamma,\delta)|\eta_{0,\gamma}. \tag{E11}$$

At this stage, $|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle$ is introduced by

$$|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle \equiv \exp\left(\sum_{n=1}^{\infty}(-)^{n}\cdot\eta_{-n,\gamma}\cdot\xi_{-n,\delta} - \sum_{n=1}^{\infty}(-)^{n}\cdot\xi_{-n,\gamma}\cdot\eta_{-n,\delta}\right)|p_{\gamma}(\chi)=0\rangle_{\gamma}\cdot|p_{\delta}(\chi)=0\rangle_{\delta}, \tag{E12}$$

which does not have any zero-mode component. Hereafter, $|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle$ will be referred to as the *small* gluing vertex function ''in $\xi\eta$-mode.'' It is to be noticed that $|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle$ in (E12) is related with $\langle\mathcal{V}_{S}^{\chi}(\gamma,\delta')|$ in (E6) ''in $\xi\eta$-mode'' by

$$\langle\mathcal{V}_{S}^{\chi}(\gamma,\delta')|\mathcal{V}_{S}^{\chi}(\delta',\delta)\rangle = ({}_{\gamma}\langle p_{\gamma}(\chi)=0|\xi_{0,\gamma})\cdot\exp\left(\sum_{n=1}^{\infty}\xi_{-n,\delta}\cdot\eta_{n,\gamma} + \sum_{n=1}^{\infty}\eta_{-n,\delta}\cdot\xi_{n,\gamma}\right)\cdot(|p_{\delta}(\chi)=0\rangle_{\delta}), \tag{E13}$$

which does not have any zero-mode component within the exponential factor. Incidentally, the operator (E13) maps the γth external $\xi_{-n,\gamma}$ ($\eta_{-n,\gamma}$) into the δth external $\xi_{-n,\delta}$ ($\eta_{-n,\delta}$) for nonzero modes (i.e., for $n\neq 0$). Since we have that

$$0 = \eta_{0,\gamma}|p_{\gamma}(\chi)=0\rangle_{\gamma}|p_{\delta}(\chi)=0\rangle_{\delta} = \eta_{0,\delta}|p_{\gamma}(\chi)=0\rangle_{\gamma}|p_{\delta}(\chi)=0\rangle_{\delta}, \tag{E14}$$

we can find that $|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle$ in (E12) ''in $\xi\eta$-mode'' satisfies the following *gluing relations* in the *small* Hilbert space:

$$\partial_{w}\xi_{\gamma}(w)|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle = \partial_{w}\xi_{\delta}(-1/w)|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle \tag{E15}$$

and

$$w^{2}\cdot\eta_{\gamma}(w)|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle = \eta_{\delta}(-1/w)|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle. \tag{E16}$$

Furthermore, $|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle$ in (E12) ''in $\xi\eta$-mode'' can be expressed by

$$|\mathcal{V}_{S}^{\chi}(\gamma,\delta)\rangle = \eta_{0,\gamma}|\nu_{\mathcal{L}}^{\chi}(\gamma,\delta)\rangle = -\eta_{0,\delta}|\nu_{\mathcal{L}}^{\chi}(\gamma,\delta)\rangle, \tag{E17}$$

$|\nu_{\mathcal{L}}^{\chi}(\gamma,\delta)\rangle$ being defined by

$$\begin{aligned}|\nu_{\mathcal{L}}^{\chi}(\gamma,\delta)\rangle &= -|\nu_{\mathcal{L}}^{\chi}(\delta,\gamma)\rangle \\ &\equiv \exp\left(\sum_{n=0}^{\infty}(-)^{n}\cdot\eta_{-n,\gamma}\cdot\xi_{-n,\delta} - \sum_{n=1}^{\infty}(-)^{n}\cdot\xi_{-n,\gamma}\cdot\eta_{-n,\delta}\right) \\ &\quad\times(\xi_{0,\gamma}|p_{\gamma}(\chi)=0\rangle_{\gamma})(|p_{\delta}(\chi)=0\rangle_{\delta}),\end{aligned} \tag{E18}$$

which has zero modes within the exponential factor, so that we find the following *gluing* relations even for the zero-mode $\xi_{0,\gamma}$:

$$\xi_{0,\gamma}|\nu_{\mathcal{L}}^{\chi}(\gamma,\delta)\rangle = \xi_{0,\delta}|\nu_{\mathcal{L}}^{\chi}(\gamma,\delta)\rangle. \tag{E19a}$$

Therefore, we find from Eqs. (E15)–(E17) and (E19) that $|\nu_{\mathcal{L}}^{\chi}(\gamma,\delta)\rangle$ in (E18) ''in $\xi\eta$-mode'' satisfies the following gluing relations within the *large* Hilbert space:

$$\xi_\gamma(w)|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle=\xi_\delta(-1/w)|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle,$$
$$w^2\cdot\eta_\gamma(w)|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle=\eta_\delta(-1/w)|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle. \tag{E19b}$$

[Hereafter, $|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle$ in (E18) will be referred to as the *large* gluing vertex functino ''in $\xi\eta$-mode.''] On the other hand, the gluing vertex functino $|V^\chi(\gamma,\delta)\rangle$ in (2.19) (''in χ-mode'') has the same gluing relations as those (E19b) satisfied by $|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle$ in (E18) ''in $\xi\eta$-mode.'' Therefore, $|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle$ in (E18) ''in $\xi\eta$-mode'' is just equal to $|V^\chi(\gamma,\delta)\rangle$ in (2.20) ''in χ-mode,'' i.e.,

$$|\nu_{\mathcal{L}}^\chi(\gamma,\delta)\rangle=|V^\chi(\gamma,\delta)\rangle. \tag{E20}$$

As for $|V^\varphi(\gamma,\delta)\rangle$'s in (2.19) and (2.20) [for $\varphi=\vec{X},\phi^j,\sigma,\phi(,\chi)$], they are found to satisfy the following *gluing identities*: By using the inlint standard ket states $|p(\varphi)=0\rangle$ [satisfying Eq. (A6)] and the inlint dual standard bra states $\langle p(\varphi)=0|$ in (A9) [for $\varphi=\vec{X},\phi^j,\sigma,\phi(,\chi)$, and $j=0-4$], we find for $\vec{X}$-modes that

$$_\gamma\langle q_\gamma(\vec{X})=0|\cdot{}_\delta\langle q_\delta(\vec{X})=0|(W_\gamma^{\vec{X}}[w_\gamma]|p(\vec{X})=0\rangle)\left(\langle\tilde{p}(\vec{X})=0|\tilde{W}_\delta^{\vec{X}}\left[\frac{-1}{w_\delta}\right]\right)|V^{\vec{X}}(\gamma,\delta)\rangle \tag{E21a}$$

$$=\left(:\exp\left(\sum_\pm\sum_{j=0}^4\sum_{n=1}^\infty\frac{1}{n}\cdot J_{-n}(X^{\pm j})\cdot\tilde{J}_n(X^{\mp j})\right)\right)$$
$$\times\left(\sum_{\{p^{\pm j}\}}\exp\left(\sum_\pm\sum_{j=0}^4 p^{\pm j}\cdot q(X^{\mp j})\right)|p(\vec{X})=0\rangle\right.$$
$$\left.\times\langle\tilde{p}(\vec{X})=0|\exp\left(-\sum_\pm\sum_{j=0}^4 p^{\pm j}\cdot\tilde{q}(X^{\mp j})\right)\right): \tag{E21b}$$

$$\equiv E_{\vec{X}}(J(\vec{X}),\tilde{J}(\vec{X})), \tag{E21c}$$

while we find for φ-modes [$\varphi=\phi^j,\sigma,\phi(,\chi)$] that

$$_\gamma\langle q_\gamma(\varphi)=0|\cdot{}_\delta\langle q_\delta(\varphi)=0|(W_\gamma^\varphi[w_\gamma]|p(\varphi)=0\rangle)\left(\langle\tilde{p}(\varphi)=0|\tilde{W}_\delta^\varphi\left[\frac{-1}{w_\delta}\right]\right)|V^\varphi(\gamma,\delta)\rangle \tag{E22a}$$

$$=:\exp\left(\sum_{n=1}^\infty\varepsilon_\varphi^\phi\cdot\frac{1}{n}\cdot J_{-n}(\varphi)\cdot\tilde{J}_n(\varphi)\right)$$
$$\times\left(\sum_{p\in Z(+1/2)}\exp(p\cdot q(\varphi))|p(\varphi)=0\rangle\langle\tilde{p}(\varphi)=0|\exp(-(p+Q(\varphi))\tilde{q}(\varphi))\right): \tag{E22b}$$

$$\equiv E_\varphi(J(\varphi),\tilde{J}(\varphi))\quad\text{for}\ \ \varphi=\phi^j,\sigma,\phi(,\chi). \tag{E22c}$$

In Eqs. (E21) and (E22), we have used inlaying operators $W^\varphi[\dots]$'s (2.4) and (2.5) ''in φ-modes.'' Incidentally, for any state $H(\tilde{J}(\vec{X}))|\tilde{p}(\vec{X})=0\rangle$ and $G(\tilde{J}(\varphi))|\tilde{p}(\tilde{\varphi})=0\rangle$, we find that

$$E_{\vec{X}}(J(\vec{X}),\tilde{J}(\vec{X}))\cdot H(\tilde{J}(\vec{X}))|\tilde{p}(\vec{X})=0\rangle=H(J(\vec{X}))|p(\vec{X})=0\rangle \tag{E23}$$

and

$$E_\varphi(J(\varphi),\tilde{J}(\varphi))\cdot G(\tilde{J}(\varphi))|\tilde{p}(\varphi)=0\rangle=G(J(\varphi))|p(\varphi)=0\rangle\quad\text{for}\ \ \varphi=\phi^j,\sigma,\phi(,\chi). \tag{E24}$$

[Incidentally, Eq. (E24) for $\varphi=\chi$ is the gluing identity within the *large* Hilbert space.] On the other hand, with the help of the inlaying operator $\mathcal{W}^{\chi}[\ldots]$ in (E1) ''in $\xi\eta$-mode'' and the *small* gluing vertex function $|\mathcal{V}_S^{\chi}(\gamma,\delta)\rangle$ in (E17) ''in $\xi\eta$-mode,'' we find from Eqs. (E3) and (E4) the following gluing identity within the *small* Hilbert space:

$$ {}_{\gamma}\langle q_{\gamma}(\chi)=0|\cdot{}_{\delta}\langle q_{\delta}(\chi)=0|(\mathcal{W}_{\gamma}^{\chi}[w_{\gamma}]|p(\chi)=0\rangle)\left((\langle\tilde{p}(\chi)=0|\tilde{\xi}_0)\cdot\tilde{\mathcal{W}}_{\delta}^{\chi}\left[\frac{-1}{w_{\delta}}\right]\right)|\mathcal{V}_S^{\chi}(\gamma,\delta)\rangle \tag{E25a}$$

$$ =:\exp\left(\sum_{n=1}^{\infty}\eta_{-n}\cdot\tilde{\xi}_n+\sum_{n=1}^{\infty}\xi_{-n}\cdot\tilde{\eta}_n\right)(|p(\chi)=0\rangle\cdot\langle\tilde{p}(\chi)=0|\tilde{\xi}_0):_{\Sigma} \tag{E25b}$$

$$ =\mathcal{E}_{\chi}(\mathcal{J}_{\chi},\tilde{\mathcal{J}}_{\chi}). \tag{E25c}$$

It should be noticed that zero-mode terms $\xi_0\cdot\eta_{0,\gamma}$ and $\tilde{\eta}_0\cdot\xi_{0,\delta}$ in Eqs. (E3) and (E4) are *absent* in the exponential factor of Eq. (E25b), since we have that

$$ {}_{\delta}\langle q_{\delta}(\chi)=0|\xi_{0,\delta}=0=\eta_{0,\gamma}|p_{\gamma}(\chi)=0\rangle_{\gamma}=0. \tag{E26}$$

Thus, $\mathcal{J}_{\chi}$ $[\tilde{\mathcal{J}}_{\chi}]$ in Eq. (E25c) represents inlint operators *without* ξ_0-mode (those *without* $\tilde{\xi}_0$-mode). Furthermore, we obtain for any state $K(\tilde{\mathcal{J}}_{\chi})|\tilde{p}(\chi)=0\rangle$ (without zero modes) that

$$ \mathcal{E}_{\chi}(\mathcal{J}_{\chi},\vec{\mathcal{J}}_{\chi})(K(\tilde{\mathcal{J}}_{\chi})|\tilde{p}(\chi)=0\rangle)=(K(\mathcal{J}_{\chi})|p(\chi)=0\rangle). \tag{E27}$$

The normal ordering operations : : in Eqs. (E21b), (E22b), and (E25b) mean that any operator with the superscript ''$\sim$'' should be moved and placed to the right of $\langle p(\varphi)=0|$ (for $\varphi=\vec{X},\varphi^j,\sigma,\phi$) or $\langle\tilde{p}(\chi)=0|\tilde{\xi}_0$.

Identities (E21)=(E21c), (E22)=(E22c), and (E25)=(E25c) will be hereafter referred to as the ''*gluing identities*,'' on the left-hand sides of which the γth punctured ring domain and the δth punctured ring domain exist, since we have both the inlaying operators $W_{\gamma}[w_{\gamma}]$ and $\tilde{W}_{\delta}[-1/w_{\delta}]$. On the other hand, we have not any punctured ring domains on the right-hand sides of the ''gluing identities.'' This fact suggests that $|V^{\varphi}(\gamma,\delta)\rangle$ in (2.18) and (2.19) (for $\varphi\neq\chi$) and $|\mathcal{V}_S^{\chi}(\gamma,\delta)\rangle$ in (E12) ''in $\xi\eta$-mode'' have the the following *gluing effect*: They *glue* the γth punctured ring domain and the δth punctured ring domain into one *un*punctured ring domain (described by *common* inlint operators, within the *small* Hilbert space). Combining all these ''gluing identities'' in various modes, we finally obtain one (GSO projected) ''gluing identity'' (in the *small* Hilbert space) which is symbolically expressed by

$$\begin{aligned}
&\left(\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi}{}_{\gamma}\langle q_{\gamma}(\varphi)=0|W_{\gamma}^{\varphi}[w_{\gamma}]\right)(({}_{\gamma}\langle p_{\gamma}(\chi)=0|\xi_{0,\gamma})\mathcal{W}_{\gamma}^{\chi}[w_{\gamma}])\\
&\quad\times\left(\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi,\chi}|p(\varphi)=0\rangle\right)\cdot\left(\prod_{\tilde{\varphi}=\tilde{\vec{X}},\tilde{\phi}^j,\tilde{\sigma},\tilde{\phi},\tilde{\chi}}\langle p(\tilde{\varphi})=0\right)\tilde{\xi}_0\\
&\quad\times\left(\prod_{\tilde{\varphi}=\tilde{\vec{X}},\tilde{\phi}^j,\tilde{\sigma},\tilde{\phi}}{}_{\delta}\langle q_{\delta}(\tilde{\varphi})=0|W_{\delta}^{\tilde{\varphi}}\left[\frac{-1}{w_{\delta}}\right]\right)\left(({}_{\delta}\langle p_{\delta}(\tilde{\chi})=0|\tilde{\xi}_{0,\delta})\cdot\mathcal{W}_{\delta}^{\tilde{\chi}}\left[\frac{-1}{w_{\delta}}\right]\right)\\
&\quad\times P_{\gamma}^{\pi(\gamma)}(\mathrm{GSO})|\mathcal{V}_S^{\chi}(\gamma,\delta)\rangle\left(\prod_{\varphi=\vec{X},\phi^J,\sigma,\phi}|V^{\varphi}(\gamma,\delta)\rangle\right)\\
&\quad=P^{\pi(\gamma)}(\mathrm{GSO})\cdot E(J,\tilde{J}),
\end{aligned}\tag{E28a}$$

with

$$E(J,\tilde{J}) \equiv E_{\vec{X}}(J(\vec{X}),\tilde{J}(\vec{X}))\cdot \mathcal{E}_{\chi}(\mathcal{J}_{\chi},\tilde{\mathcal{J}}_{\chi})\left(\prod_{\varphi=\phi^j,\sigma,\phi} E_{\varphi}(J(\varphi),\tilde{J}(\varphi))\right). \tag{E28b}$$

In Eq. (E28), $P_{\gamma}^{\pi(\gamma)}$ (GSO) [for $\pi(\gamma)=$NS,R] is the γth external GSO-projection operator and $P^{\pi(\gamma)}$ (GSO) is the inlint GSO-projection operator, both of which are defined by Eq. (1.11). The gluing identity (E28) is valid for any ket (bra) state in the *small* Hilbert space. Hereafter, the gluing identity (E25) will be simply expressed by

$$({}_{\gamma}\langle q_{\gamma}=0|W_{\gamma}[w_{\gamma}]|\mathcal{Q}=0\rangle)\cdot\left((\langle\tilde{\mathcal{Q}}=0|\tilde{\xi}_0)\times{}_{\delta}\langle q_{\delta}=0|\tilde{W}_{\delta}\left[\frac{-1}{w_{\delta}}\right]\right)(P_{\gamma}^{\pi}(\text{GSO})|\nu_{\mathcal{S}}(\gamma,\delta)\rangle)$$

$$=P^{\pi}(\text{GSO})\cdot E(J,\tilde{J})\quad(\text{to be referred to as the gluing identity}). \tag{E29}$$

In Equation (E29), we have used the *small* gluing vertex funct*ino* $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$, which has been constructed by

$$|\nu_{\mathcal{S}}(\gamma,\delta)\rangle\equiv|\mathcal{V}_{\mathcal{S}}^{\chi}(\gamma,\delta)\rangle\left(\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi}|V^{\varphi}(\gamma,\delta)\rangle\right). \tag{E30a}$$

Incidentally, we have also that

$$\langle\nu_{\mathcal{S}}(\gamma,\delta)|\equiv\langle\mathcal{V}_{\mathcal{S}}^{\chi}(\gamma,\delta)|\left(\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi}\langle V^{\varphi}(\gamma,\delta)|\right)\quad[\text{see Eq. (2.14)}]. \tag{E30b}$$

It is remarkable that the *small* gluing vertex funct*ino* $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ constructed by Eq. (E30) can be alternatively given by

$$|\nu_{\mathcal{S}}(\gamma,\delta)\rangle=\eta_{0,\gamma}|V_L(\gamma,\delta)\rangle\quad[\text{see Eq. (E17)}], \tag{E31}$$

where the *large* gluing vertex function $|V_L(\gamma,\delta)\rangle$ *can* be constructed in terms of φ-modes (i.e., without using $\xi\eta$-modes) by

$$|V_L(\gamma,\delta)\rangle\equiv\prod_{\varphi=\vec{X},\phi^j,\sigma,\phi,\chi}|V^{\varphi}(\gamma,\delta)\rangle, \tag{E32}$$

$|V^{\vec{X}}(\gamma,\delta)\rangle$ ($|V^{\varphi}(\gamma,\delta)\rangle$) being given by Eq. (2.18) [(2.19)]. [As we shall see in Sec. IV, $|V_L(\gamma,\delta)\rangle$ is useful especially in loop calculations, since it gives gluing identities within the *large* Hilbert space.]

With the help of the gluing identity (E29) in the *small* Hilbert space, we can prove the gluing theorem in the tree calculations [i.e., Eq. (2.20)], using the techniques used in Ref. 5. In particular, the elementary vertex functino $\langle\nu_{\mathcal{S}}(\ldots)|$'s satisfy the following gluing theorem:

$$\langle\nu_{\mathcal{S}}(1,2,...,n,\gamma)|\cdot\langle\nu_{\mathcal{S}}(\delta,n+1,n+2,...,n+m)|\nu_{\mathcal{S}}(\gamma,\delta)\rangle=\langle\nu_{\mathcal{S}}(1,2,...,n,n+1,n+2,...,n+m)| \tag{E33a}$$

$$[=\langle\nu_{\mathcal{S}}(2,...,n+m,1)|,\quad\text{being cycle symmetric}], \tag{E33b}$$

so that $\langle\nu_{\mathcal{S}}(1,2,...,N)|$ in (2.14) is found to be constructed by using the inlayed coordinate $z_{rm}(w_r)$ (for $r=1-N$) in the ''inlayed coordinate system m(*idpoint*)'' (ICS m), which is given by

$$z_{rm}(w_r)=\exp\left(\frac{\pi\sqrt{-1}}{N}\right)\cdot\frac{z_r(w_r)-1}{z_r(w_r)-\exp(2\pi\sqrt{-1}/N)}\quad\text{for } r=1-N, \tag{E34}$$

with

$$z_r(w_r)\equiv\left(\exp\frac{2\pi\sqrt{-1}}{N}(r-1)\right)\left(\frac{1+\sqrt{-1}\cdot w_r}{1-\sqrt{-1}\cdot w_r}\right)^{2/N}. \tag{E35}$$

Incidentally, there exist *only* N punctured ring domains and *not any* unpunctured ring domain in ''ICS m,'' and the inlayed coordinate ''z_m'' satisfies the following Gross–Jevicki fundamental equation (GJ-$\mathcal{FE}$):[17]

$$\frac{dz_m}{d\rho_m}=(-)^N\cdot\frac{\Pi_{r(\neq 2)}(z_m-z_{rm}(0))}{((z_m-\mathcal{Y}_+^m)\cdot(z_m-\mathcal{Y}_-^m))^{(N-2)/2}}, \tag{E36a}$$

with

$$\mathcal{Y}_\pm^m\equiv\exp(\pm\pi\sqrt{-1}/N). \tag{E36b}$$

It is to be noticed that GJ-$\mathcal{FE}$ (E36) leads to the rth inlayed coordinate $z_{rm}(w_r)$'s in (E34). [Unfortunately there exist misprints in Eq. (1.15) of Ref. 5, which are corrected in Eq. (E36).] We notice that the elementary vertex functino on the right-hand side of Eq. (E33) is *symmetric* under the *cyclic* permutation. [See the cycle-symmetric coordinate $z_r(w_r)$'s in (E35).] Therefore, we also have that

$$\langle\nu_{\mathcal{S}}(n+1,n+2,...,n+m,\delta)|\cdot\langle\nu_{\mathcal{S}}(\gamma,1,2,...,n)|\nu_{\mathcal{S}}(\delta,\gamma)\rangle=\langle\nu_{\mathcal{S}}(n+1,n+2,...,n+m,1,2,...,n)|. \tag{E37}$$

Since any elementary vertex functino $\langle\nu_{\mathcal{S}}(\dots)|$ is Grassman odd, comparing Eq. (E37) with Eq. (E33) leads to the fact that the *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ is odd under $\gamma\leftrightarrow\delta$, i.e.,

$$|\nu_{\mathcal{S}}(\gamma,\delta)\rangle=-|\nu_{\mathcal{S}}(\delta,\gamma)\rangle. \tag{E38}$$

Since we have from Eqs. (E38) that

$$\begin{aligned}\langle\nu_{\mathcal{S}}(1,2,...,n)|&=\langle\nu_{\mathcal{S}}(1,\gamma)|\cdot\langle\nu_{\mathcal{S}}(\delta,2,3,...,n)|\nu_{\mathcal{S}}(\gamma,\delta)\rangle\\&=-\langle\nu_{\mathcal{S}}(\delta,2,3,...,n)|(\langle\nu_{\mathcal{S}}(1,\gamma)|\nu_{\mathcal{S}}(\gamma,\delta)\rangle),\end{aligned} \tag{E39}$$

we also find that

$$\langle\nu_{\mathcal{S}}(\alpha,\gamma)|\nu_{\mathcal{S}}(\gamma,\delta)\rangle=-E_{\alpha,\delta}, \tag{E40}$$

where $E_{\alpha,\delta}$ is the operator changing the αth external operator $\mathcal{G}_\alpha$ into the δth external operator $\mathcal{G}_\delta$ (i.e., $E_{\alpha,\delta}|\mathcal{G}\rangle_\alpha=|\mathcal{G}\rangle_\delta$).

For any primary operator $\mathcal{G}$ (*without* ξ_0-mode) of conformal weight $d(\mathcal{G})$, we have the following *gluing relation* in the *small* Hilbert space satisfied by the total *small* gluing vertex functino $|\nu_{\mathcal{S}}(\gamma,\delta)\rangle$ in (E31):

$$\left((w^{2d(\mathcal{G})})\mathcal{G}_\gamma(w)-\mathcal{G}_\delta\left(\frac{-1}{w}\right)\right)|\nu_{\mathcal{S}}(\gamma,\delta)\rangle=0. \tag{E41}$$

Therefore, using

$$\mathcal{G}_r(w)=\frac{1}{w^{d(\mathcal{G})}}\left(\sum_{n=-\infty}^{\infty}g_{n,r}\cdot w^{-n}\right)\quad(\text{for}\ \ r=\gamma,\delta), \tag{E42}$$

we find from Eq. (E41) the following *gluing relations* for various modes (in the *small* Hilbert space):

$$g_{n,\gamma}|\nu_{\mathcal{S}}(\gamma,\delta)\rangle=(-)^{d(\mathcal{G})+n}\cdot g_{-n,\delta}|\nu_{\mathcal{S}}(\gamma,\delta)\rangle. \tag{E43}$$

It is quite important in loop calculations that the total *large* gluing vertex function $|V_L(\gamma,\delta)\rangle$ in (E32) "in φ-modes" satisfies the following gluing relations:

$$\left((w^{2d(\mathcal{G})})\mathcal{G}_\gamma(w)-\mathcal{G}_\delta\left(\frac{-1}{w}\right)\right)|V_L(\gamma,\delta)\rangle=0, \tag{E44}$$

which holds for any primary operator (with or without the ξ_0-mode).

[1] For reviews of Ramond–Neveu–Schwarz (as well as Green–Schwarz) superstring theory, see M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge U.P., New York, 1987), Vols. 1 and 2; M. Kaku, *Introduction to Superstring* (Springer-Verlag, New York, 1988).

[2] J. Polchinski, *String Theory I (Introduction to the bosonic string)* and *String Theory II (Superstring theory and beyond)* (Cambridge U.P., New York, 1998).

[3] E. Witten, "Non-commutative geometry and string field theory," Nucl. Phys. B **268**, 253–294 (1986). Witten's quantum open bosonic string field theory has been discussed by many authors. See C. B. Thorn, "Perturbation theory for quantized string fields," *ibid.* **287**, 61–92 (1987); A. LeClair, M. E. Peskin, and C. R. Preitschopf, "String field theory on the conformal plane (I), Kinematical principles," *ibid.* **317**, 411–463 (1988); "String field theory on the conformal plane (II), Generalized gluing," *ibid.* **317**, 464–508 (1989) and various papers cited therein.

[4] T. Kugo, H. Kunitomo, and K. Suehiro, "Non-polynomial closed string field theory," Phys. Lett. B **226**, 48–54 (1989); T. Kugo and K. Suehiro, "Non-polynomial closed string field theory: Action and its gauge invariance," Nucl. Phys. B **337**, 434–466 (1990). B. Zwiebach, "Closed string field theory: Quantum action and the Batalin-Vilkovsky master equation," *ibid.* **390**, 33–152 (1993); B. Zwiebach, "Oriented open-closed string field theory revisited," Annals Phys. **267**, 193–248 (1998), and various papers cited therein.

[5] S. Naito, "Inlaying vertex function and scattering amplitude," J. Math. Phys. **38**, 1413–1453 (1997).

[6] E. Witten, "Interacting field theory of open superstring," Nucl. Phys. B **276**, 291–324 (1986).

[7] P. Ramond, "Dual theory for free fermions," Phys. Rev. D **3**, 2415–2418 (1971).

[8] A. Neveu and J. H. Schwarz, "Factorizable dual model of pions," Nucl. Phys. B **31**, 86–112 (1971); "Quark model of dual pions," Phys. Rev. D **4**, 1109–1111 (1971).

[9] D. Friedan, E. Martinec, and S. Schenker, "Conformal Invariance, supersymmetry and string theory," Nucl. Phys. B **271**, 93–165 (1986).

[10] C. Wendt, "Scattering amplitudes and contact interactions in Witten's superstring field theory," Nucl. Phys. B **314**, 209–237 (1989).

[11] C. R. Preitschopf, C. B. Thorn, and S. Yost, "Superstring field theory," Nucl. Phys. B **337**, 363–433 (1990).

[12] J. P. Yamron, "A gauge invariant action for the free Ramond string," Phys. Lett. B **174**, 69–74 (1986); B. Sazdovic, "Equivalence of different formulations of the free Ramond string field theory," *ibid.* **195**, 536–540 (1987); I. Ya. Aref'eva and P. B. Mdvedev, "Truncation, picture-changing operation and spacetime super-symmetry in Neveu–Schwarz–Ramond string field theory," *ibid.* **202**, 510–514 (1988); "Anomalies in Witten's field theory of the NSR string," *ibid.* **212**, 299–308 (1988); T. Kugo and H. Terao, "New gauge symmetries in Witten's Ramond string field theory," *ibid.* **208**, 416–420 (1988).

[13] F. Gliozzi, J. Scherk, and D. Olive, "Supersymmetry, supergravity theories and the dual spinor model," Nucl. Phys. B **122**, 253–290 (1977).

[14] S. Naito, "Generating functionals of physical vertex operators in superstring," J. Math. Phys. **38**, 4980–5009 (1997); "Physical vertex operators in the Ramond sector," *ibid.* **39**, 1402–1440 (1998); "Operator product expansions and stress operators," *ibid.* **39**, 1848–1886 (1998).

[15] V. A. Kostelecky, O. Lechtenfeld, W. Lerche, S. Samuel, and S. Watamura, "Conformal techniques, bosonization and tree-level string amplitudes," Nucl. Phys. B **288**, 173–232 (1987).

[16] As a standard texbook on quantum field theory, we give J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw–Hill, New York, 1965). LSZ reduction formulas are given in Chap. 16, while Wick's theorem is given in Chap. 17.

[17] D. J. Gross and A. Jevicki, "Operator formulation of interacting string field theory (I)," Nucl. Phys. B **283**, 1–49 (1987); "Operator formulation of interacting string field theory (II)," *ibid.* **287**, 225–250 (1987); "Operator formulation of interacting string field theory (III), NSR superstring," *ibid.* **293**, 29–82 (1987).

[18] S. B. Giddings, "The Veneziano amplitude from interacting string field theory," Nucl. Phys. B **278**, 242–255 (1986); S. B. Giddings and E. Martinec, "Conformal geometry and string field theory," *ibid.* **278**, 91–120 (1986); S. Samuel, "Covariant off-shell string amplitudes," *ibid.* **308**, 285–316 (1988); R. Bluhm and S. Samuel, "The off-shell Koba-Nielsen formula," *ibid.* **323**, 337–373 (1989) and various papers cited therein.

[19] S. Samuel, "Solving the open bosonic string in perturbation theory," Nucl. Phys. B **341**, 513–610 (1990).

[20] P. D. Vecchia, F. Pezella, M. Frau, K. Hornfeck, A. Lerda, and S. Sciuto, "*N*-point *g*-loop vertex for a free bosonic theory with vacuum charge *Q*," Nucl. Phys. B **322**, 317–372 (1989).

[21] E. Verlinde and H. Verlinde, "Chiral bosonization, determinants and the string partition function," Nucl. Phys. B **288**, 357–396 (1987).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Asymptotics of Clebsch–Gordan coefficients

Matthias W. Reinsch and James J. Morehead
Department of Physics, University of California, Berkeley, California 94720

(Received 4 April 1999; accepted for publication 30 June 1999)

Asymptotic expressions for Clebsch–Gordan coefficients are derived from an exact integral representation. Both the classically allowed and forbidden regions are analyzed. Higher-order approximations are calculated. These give, for example, six digit accuracy when the quantum numbers are in the hundreds. © *1999 American Institute of Physics.* [S0022-2488(99)01210-4]

I. INTRODUCTION

This paper contains a detailed study of the asymptotics of Clebsch–Gordan coefficients and includes the derivation of new results. We use the term ''Clebsch–Gordan coefficient'' in its colloquial sense, i.e., the vector addition coefficients of SU(2). Thus our results also give the asymptotics of the $3j$-symbols. We consider the case in which all of the quantum numbers get large together. What this means is multiplying all of the quantum numbers by a number and studying the asymptotic behavior of the Clebsch–Gordan coefficient as this multiplier gets large. Such a multiplier is often called $1/\hbar$, so that the limit of large quantum numbers is the limit of small $\hbar$.

The history of this subject dates back to the early days of quantum mechanics and the study of the classical limit of quantum mechanical quantities. Numerous papers have been written in this area. We summarize the literature briefly here. In 1959, Wigner[1] discussed the physical interpretation and classical limits of Clebsch–Gordan coefficients. He described a certain average behavior, and did not analyze the oscillatory nature of the Clebsch–Gordan coefficients. There are references in this work to Edmonds[2] and Brussaard and Tolhoek.[3] In 1968, Ponzano and Regge[4] presented asymptotic expressions that included the oscillations. Their work included an interpretation of certain angles that occur in their results and in ours. Additionally, they discussed the allowed and forbidden regions. However, their derivation is, in their words, ''rather heuristic.'' It was borne out in their comparisons with the exact values. William Miller[5] derived similar expressions using semiclassical methods in 1974, but did not treat the forbidden region. Another work that relates to the present paper is that of Srinivasa Rao and V. Rajeswari.[6] It contains exact expressions for Clebsch–Gordan coefficients and their relationship to certain hypergeometric series. There is more information in the work of Biedenharn and Louck.[7]

In this paper, we start by deriving an exact integral representation for the Clebsch–Gordan coefficients. Then the methods of stationary phase are used to approximate this integral. The allowed and forbidden regions are treated separately, and the resulting expressions are related to the literature. These methods are then used to derive higher-order results, that is, the next order in an expansion in $\hbar$. These formulas are accurate to five or six digits when the quantum numbers are in the hundreds.

Possible applications of this work include high-angular momentum calculations and theoretical investigations which contain sums over large numbers of Clebsch–Gordan coefficients.[8,9]

II. EXACT EXPRESSIONS FOR THE CLEBSCH–GORDAN COEFFICIENT

Our starting point is an exact expression for the Clebsch–Gordan (vector-addition) coefficient, due to Wigner [see, for example, Eq. (3.6.11) of Ref. 2],

$$\begin{aligned}\langle j_1 m_1 j_2 m_2 | jm\rangle &= [(2j+1)(j_1+j_2-j)!(j_1-j_2+j)!(-j_1+j_2+j)!/(j_1+j_2+j+1)!]^{1/2}\\ &\quad\times[(j_1+m_1)!(j_1-m_1)!(j_2+m_2)!(j_2-m_2)!(j+m)!(j-m)!]^{1/2}\\ &\quad\times\sum_z \frac{(-1)^z}{z!(j_1+j_2-j-z)!(j_1-m_1-z)!(j_2+m_2-z)!(j-j_2+m_1+z)!(j-j_1-m_2+z)!}.\end{aligned} \tag{2.1}$$

A factor of δ_{m,m_1+m_2} has been omitted; throughout this paper we will assume that m is equal to m_1+m_2. Also, unless otherwise specified, sums over an index are sums over all integers. It will turn out, though, that the summand is nonzero for only finitely many values of the index.

We begin by deriving the following exact expression for the Clebsch–Gordan coefficient,

$$\begin{aligned}\langle j_1 m_1 j_2 m_2 | jm\rangle &= (-1)^{j+m} N_{j_1 m_1 j_2 m_2 jm} \frac{1}{(j_1-m_1)!(j_2-m_2)!}\left(\frac{d}{du}\right)^{j_1-m_1}\left(\frac{d}{dt}\right)^{j_2-m_2}\\ &\quad\times[(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}]\big|_{u=0,t=0},\end{aligned} \tag{2.2}$$

where $N_{j_1 m_1 j_2 m_2 jm}$ is defined to be

$$N_{j_1 m_1 j_2 m_2 jm} = \left[\frac{(2j+1)(j_1+m_1)!(j_1-m_1)!(j_2+m_2)!(j_2-m_2)!(j+m)!(j-m)!}{(j_1+j_2+j+1)!(j_1+j_2-j)!(j_1-j_2+j)!(-j_1+j_2+j)!}\right]^{1/2}. \tag{2.3}$$

Because the quantity being differentiated in Eq. (2.2) is a polynomial in the variables u and t, the operation of differentiating this quantity and then evaluating the result at $u=0$ and $t=0$ simply selects a particular coefficient in the polynomial. Thus, Eq. (2.2) expresses the Clebsch–Gordan coefficient as a certain coefficient in a polynomial that can be written in closed form. This equation can be derived from results in the literature,[10] but we give here an independent derivation of Eq. (2.2) from Eq. (2.1) to verify that all of the conventions involved are consistent.

In order to prove Eq. (2.2), we start by finding the coefficient of $u^{j_1-m_1}t^{j_2-m_2}$ in the polynomial $(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$. This is equal to the coefficient of $u^{j_1-m_1}$ in the polynomial that is given by $(u-1)^{j+j_1+j_2}$ times the u-dependent coefficient of $t^{j_2-m_2}$ in the polynomial $(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}$. Using the binomial theorem, we get

$$(t-1)^{j+j_2-j_1} = \sum_k \binom{j+j_2-j_1}{k} t^k (-1)^{j+j_2-j_1-k}, \tag{2.4}$$

and

$$(t-u)^{j_1+j_2-j} = \sum_l \binom{j_1+j_2-j}{l} t^l (-u)^{j_1+j_2-j-l}. \tag{2.5}$$

The coefficient of $t^{j_2-m_2}$ in the product of these is

$$\begin{aligned}&\sum_k \binom{j+j_2-j_1}{k}(-1)^{j+j_2-j_1-k}\binom{j_1+j_2-j}{j_2-m_2-k}(-u)^{j_1+j_2-j-(j_2-m_2-k)}\\ &\quad=(-1)^{j_2+m_2}u^{j_1-j+m_2}\sum_k \binom{j+j_2-j_1}{k}\binom{j_1+j_2-j}{j_2+m_2-k}u^k.\end{aligned} \tag{2.6}$$

As explained above, we need to multiply this polynomial by

$$(u-1)^{j+j_1-j_2}=\sum_l \binom{j+j_1-j_2}{l} u^l(-1)^{j+j_1-j_2-l}, \tag{2.7}$$

and find the coefficient of $u^{j_1-m_1}$. The result is

$$\begin{aligned}(-1)^{j_2+m_2}\sum_k &\binom{j+j_2-j_1}{k}\binom{j_1+j_2-j}{j_2-m_2-k}\binom{j+j_1-j_2}{j-m-k}(-1)^{j_1-j_2+m+k}\\ &=(-1)^{j+m}\sum_z (-1)^z\binom{j+j_2-j_1}{j_2+m_2-z}\binom{j_1+j_2-j}{z}\binom{j+j_1-j_2}{j_1-m_1-z},\end{aligned} \tag{2.8}$$

where we have redefined the index of summation according to $k=j-j_1-m_2+z$ in the final line of this equation and made use of the identity $\binom{a}{b}=\binom{a}{a-b}$. Since $j+m$ is always an integer, $(-1)^{2(j+m)}$ is equal to one, and we have shown that the right-hand side of Eq. (2.2) is equal to

$$\begin{aligned}N_{j_1m_1j_2m_2jm}&\sum_z (-1)^z\binom{j+j_2-j_1}{j_2+m_2-z}\binom{j_1+j_2-j}{z}\binom{j+j_1-j_2}{j_1-m_1-z}\\ &=\sum_z \frac{(-1)^z N_{j_1m_1j_2m_2jm}(j+j_2-j_1)!(j_1+j_2-j)!(j+j_1-j_2)!}{(j-j_1-m_2+z)!(j_2+m_2-z)!(j_1+j_2-j-z)!z!(j_1-m_1-z)!(j-j_2+m_1+z)!}.\end{aligned} \tag{2.9}$$

This is the same as the right-hand side of Eq. (2.1) and completes the proof of Eq. (2.2). An alternative proof begins by introducing a factor of x^z into the sum in Eq. (2.1) and deriving a third-order differential equation for the resulting function of x. This differential equation can be solved using hypergeometric functions, and the result eventually leads to the expression shown in Eq. (2.2).

Equation (2.2) can be used to obtain an exact expression for the Clebsch–Gordan coefficient as an integral. One uses the orthogonality of the functions $\exp(in\theta)$ on the interval $[-\pi, \pi]$ to select the desired coefficients in the polynomials. Thus, we substitute $\exp(i\theta)$ for t and $\exp(i\phi)$ for u in the polynomial $(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$ in Eq. (2.2), multiply by $\exp[-i(j_1-m_1)\phi-i(j_2-m_2)\theta]$, and integrate the two variable from $-\pi$ to π. The resulting expression for the Clebsch–Gordan coefficient is

$$\begin{aligned}\langle j_1m_1j_2m_2|jm\rangle&=(-1)^{j+m}N_{j_1m_1j_2m_2jm}\,\frac{1}{(2\pi)^2}\\ &\quad\times\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-i(j_1-m_1)\phi-i(j_2-m_2)\theta}(e^{i\theta}-1)^{j+j_2-j_1}(e^{i\theta}-e^{i\phi})^{j_1+j_2-j}\\ &\quad\times(e^{i\phi}-1)^{j+j_1-j_2}\,d\theta\,d\phi.\end{aligned} \tag{2.10}$$

This may be rewritten using the definition of the sin function, whereupon it becomes natural to redefine the angles by a factor of 2. The resulting form is

$$\begin{aligned}\langle j_1m_1j_2m_2|jm\rangle&=(-1)^{j+m}(2i)^{j+j_1+j_2}\pi^{-2}N_{j_1m_1j_2m_2jm}\\ &\quad\times\int_{-\pi/2}^{\pi/2}\int_{-\pi/2}^{\pi/2} e^{2im_1\phi+2im_2\theta}\sin^{j+j_2-j_1}\theta\,\sin^{j_1+j_2-j}(\theta-\phi)\\ &\quad\times\sin^{j+j_1-j_2}\phi\,d\theta\,d\phi.\end{aligned} \tag{2.11}$$

It is this integral expression for the Clebsch–Gordan coefficient that we use in the following sections to derive formulas for the asymptotic behavior of these coefficients.

It is also possible to express the Clebsch–Gordan coefficient as a coefficient of a term in a polynomial in one variable, and thus as a one-dimensional integral. Equation (2.2) shows how the Clebsch–Gordan coefficient is related to the coefficient of $u^{j_1-m_1}t^{j_2-m_2}$ in the polynomial $(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$. This is the same as the coefficient of $u^{j_1-m_1}(u^M)^{j_2-m_2}$ in the polynomial $(u^M-1)^{j+j_2-j_1}(u^M-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$ (that is, t has been replaced by u^M) for sufficiently large integers M. This can be seen as follows. We start by imagining the polynomial $(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$ expanded out into a sum of monomials. If t is replaced by u^M, each of the monomials is now just a coefficient times a power of u. We do not want any of these terms to have the same power of u, otherwise they would combine and the coefficients would change. Thus, we look at the original polynomial $(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$ and ask what the highest power of u is. This is $(j_1+j_2-j)+(j+j_1-j_2)=2j_1$. We therefore select M to be $2j_1+1$. The result is that the coefficient of $u^{j_1-m_1}t^{j_2-m_2}$ in the polynomial $(t-1)^{j+j_2-j_1}(t-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$ is the same as the coefficient of $u^{j_1-m_1+(2j_1+1)(j_2-m_2)}$ in the polynomial $(u^{2j_1+1}-1)^{j+j_2-j_1}(u^{2j_1+1}-u)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$. We may drop an overall factor of $u^{j_1+j_2-j}$, so this coefficient is the same as the coefficient of $u^{j-m+2j_1(j_2-m_2)}$ in the polynomial $(u^{2j_1+1}-1)^{j+j_2-j_1}(u^{2j_1}-1)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}$. The resulting expression for the Clebsch–Gordan coefficient as a coefficient in a polynomial in one variable is

$$\langle j_1m_1j_2m_2|jm\rangle=(-1)^{j+m}N_{j_1m_1j_2m_2jm}\frac{1}{[j-m+2j_1(j_2-m_2)]!}\left(\frac{d}{du}\right)^{j-m+2j_1(j_2-m_2)}$$
$$\times[(u^{2j_1+1}-1)^{j+j_2-j_1}(u^{2j_1}-1)^{j_1+j_2-j}(u-1)^{j+j_1-j_2}]|_{u=0}. \tag{2.12}$$

As above, the selection of the coefficient in the polynomial can also be carried out with an integral,

$$\langle j_1m_1j_2m_2|jm\rangle=(-1)^{j+m}N_{j_1m_1j_2m_2jm}\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{-i[j-m+2j_1(j_2-m_2)]\phi}$$
$$\times(e^{i(2j_1+1)\phi}-1)^{j+j_2-j_1}(e^{i2j_1\phi}-1)^{j_1+j_2-j}(e^{i\phi}-1)^{j+j_1-j_2}\,d\phi. \tag{2.13}$$

This may be rewritten as

$$\langle j_1m_1j_2m_2|jm\rangle=(-1)^{j+m}N_{j_1m_1j_2m_2jm}\frac{(2i)^{j+j_1+j_2}}{2\pi}$$
$$\times\int_{-\pi}^{\pi}e^{i(2j_1m_2+m)\phi}\sin^{j+j_2-j_1}[(j_1+1/2)\phi]$$
$$\times\sin^{j_1+j_2-j}(j_1\phi)\sin^{j+j_1-j_2}(\phi/2)\,d\phi, \tag{2.14}$$

and this may be simplified to the form

$$\langle j_1m_1j_2m_2|jm\rangle=(-1)^{j+m}N_{j_1m_1j_2m_2jm}\pi^{-1}2^{j+j_1+j_2}\int_0^{\pi}\cos\left[(2j_1m_2+m)\phi+\frac{\pi}{2}(j+j_1+j_2)\right]$$
$$\times\sin^{j+j_2-j_1}[(j_1+1/2)\phi]\sin^{j_1+j_2-j}(j_1\phi)\sin^{j+j_1-j_2}(\phi/2)\,d\phi. \tag{2.15}$$

Although this is a one-dimensional integral (as opposed to the two-dimensional integral presented above), it seems to be not as useful for the study of asymptotics because of the presence of the magnetic quantum numbers in the argument of the cosine function.

III. STATIONARY-PHASE APPROXIMATION OF THE INTEGRAL EXPRESSION FOR THE CLEBSCH–GORDAN COEFFICIENT

In order to carry out a stationary-phase approximation of the integral expression for the Clebsch–Gordan coefficient presented in the previous section, we begin by writing the expression in Eq. (2.11) in the form

$$\langle j_1 m_1 j_2 m_2 | jm \rangle = (-1)^{j+m}(2i)^{j+j_1+j_2}\pi^{-2}N_{j_1 m_1 j_2 m_2 jm}\int_{-\pi/2}^{\pi/2}\int_{-\pi/2}^{\pi/2} e^{g(\theta,\phi)}\,d\theta\,d\phi, \tag{3.1}$$

where the function $g(\theta,\phi)$ is defined to be

$$g(\theta,\phi) = 2im_1\phi + 2im_2\theta + (j+j_2-j_1)\ln(\sin\theta) + (j_1+j_2-j)\ln[\sin(\theta-\phi)] + (j+j_1-j_2)\ln(\sin\phi). \tag{3.2}$$

Note that $g(\theta,\phi)$ has singularities where it goes to $-\infty$, but the integral is still well defined because the integrand is $\exp(g)$.

To find the stationary-phase points (as explained in Appendix A), we must first compute the first derivatives of the function g,

$$\frac{\partial g}{\partial\theta} = 2im_2 + (j+j_2-j_1)\cot\theta + (j_1+j_2-j)\cot(\theta-\phi), \tag{3.3}$$

$$\frac{\partial g}{\partial\phi} = 2im_1 - (j_1+j_2-j)\cot(\theta-\phi) + (j+j_1-j_2)\cot\phi. \tag{3.4}$$

Setting these first derivatives equal to zero results in a system of two equations in two variables. The identity

$$\cot(\theta-\phi) = \frac{1+\cot\theta\cot\phi}{\cot\phi-\cot\theta} \tag{3.5}$$

may be used to transform this system to an equivalent system,

$$2im + (j+j_2-j_1)\cot\theta + (j+j_1-j_2)\cot\phi = 0, \tag{3.6}$$

$$2im_2 + (j+j_2-j_1)\cot\theta + (j_1+j_2-j)\,\frac{1+\cot\theta\cot\phi}{\cot\phi-\cot\theta} = 0. \tag{3.7}$$

In order to be clear on phase conventions, choices of signs, and branch cuts, we write out the steps involved in solving this system of two equations for $\cot\theta$ and $\cot\phi$. We start by multiplying the second equation by $(j+j_1-j_2)(\cot\phi-\cot\theta)$ and substituting in first one:

$$[2im_2 + (j+j_2-j_1)\cot\theta]\{-[2im + (j+j_2-j_1)\cot\theta] - (j+j_1-j_2)\cot\theta\} + (j_1+j_2-j)\{(j+j_1-j_2) - \cot\theta[2im + (j+j_2-j_1)\cot\theta]\} = 0. \tag{3.8}$$

This is a quadratic equation in $\cot\theta$:

$$\cot^2\theta\ [(j+j_2-j_1)(-2j) - (j_1+j_2-j)(j+j_2-j_1)] + \cot\theta\ [(j+j_2-j_1)(-2im) + 2im_2(-2j) + (j_1+j_2-j)(-2im)] + 2im_2(-2im) + (j_1+j_2-j)(j+j_1-j_2) = 0. \tag{3.9}$$

Simplifying this results in

$$-\cot^2\theta\ (j+j_2-j_1)(j_1+j_2+j)-4i\cot\theta\ (j_2m+m_2j)$$

$$+4m_2m+(j_1+j_2-j)(j+j_1-j_2)=0. \tag{3.10}$$

The two solutions for the quantities $\cot\theta$ and $\cot\phi$ are (the upper choice of sign is one solution and the lower choice of sign is the other)

$$\cot\theta=\frac{-2i(j_2m+m_2j)\mp\beta}{(j_1+j_2+j)(j+j_2-j_1)},$$
$$\cot\phi=\frac{-2i(j_1m+m_1j)\pm\beta}{(j_1+j_2+j)(j+j_1-j_2)}, \tag{3.11}$$

where β is defined to be

$$\beta=\sqrt{4m_1m_2j^2-4mm_1j_2^2-4mm_2j_1^2+(j_1+j_2-j)(j+j_2-j_1)(j+j_1-j_2)(j_1+j_2+j)}. \tag{3.12}$$

In this equation, we use the usual choice of branch cut for the square-root function: if the argument is negative, then the result is a positive number times the imaginary unit. As discussed in Sec. III A, the quantity β is real for classically allowed sets of quantum numbers, and it is pure imaginary for classically forbidden sets of quantum numbers. It should be noted that this is the same definition for the symbol β as in Ref. 5.

The stationary-phase approximation of the integral $\int_{-\pi/2}^{\pi/2}\int_{-\pi/2}^{\pi/2}e^{g(\theta,\phi)}\,d\theta\,d\phi$ that appears in the expression for the Clebsch–Gordan coefficient in Eq. (3.1) is given by a sum of terms of the form

$$\frac{2\pi}{\sqrt{\det\partial^2 g/\partial(\theta,\phi)^2}}e^{g(\theta,\phi)}, \tag{3.13}$$

summed over stationary-phase points. The branch cut for the square root function is just below the negative imaginary axis, as is usual. The symbol $\partial^2 g/\partial(\theta,\phi)^2$ denotes the 2×2 Hessian matrix of second-order derivatives of the function $g(\theta,\phi)$, whose entries are given by

$$\frac{\partial^2 g}{\partial\theta^2}=-(j+j_2-j_1)\csc^2\theta-(j_1+j_2-j)\csc^2(\theta-\phi),$$

$$\frac{\partial^2 g}{\partial\theta\partial\phi}=(j_1+j_2-j)\csc^2(\theta-\phi), \tag{3.14}$$

$$\frac{\partial^2 g}{\partial\phi^2}=-(j_1+j_2-j)\csc^2(\theta-\phi)-(j+j_1-j_2)\csc^2\phi.$$

Using the identity $\csc^2\theta=1+\cot^2\theta$ these quantities can be expressed in terms of the cotangents in Eq. (3.11) without addressing the issue of branch cuts of the arc-cotangent function. The value of $\csc^2(\theta-\phi)$ can be determined from the quantities in Eq. (3.11) using the identity $\sin(\theta-\phi)=\sin\theta\cos\phi-\sin\phi\cos\theta=\sin\theta\sin\phi(\cot\phi-\cot\theta)$. The determinant becomes

$$\det\frac{\partial^2 g}{\partial(\theta,\phi)^2}=(j+j_2-j_1)\csc^2\theta(j+j_1-j_2)\csc^2\phi$$
$$+(j_1+j_2-j)\csc^2(\theta-\phi)[(j+j_2-j_1)\csc^2\theta+(j+j_1-j_2)\csc^2\phi],$$
$$=(1+\cot^2\theta)(1+\cot^2\phi)\Bigg\{(j+j_2-j_1)(j+j_1-j_2)$$
$$+\frac{(j_1+j_2-j)}{(\cot\phi-\cot\theta)^2}[(j+j_2-j_1)(1+\cot^2\theta)+(j+j_1-j_2)(1+\cot^2\phi)]\Bigg\}. \tag{3.15}$$

In this form the determinant is expressed entirely in terms of the cotangents of θ and ϕ. The quantity $e^{g(\theta,\phi)}$ can also be expressed in this way. Choices of branch cuts are not necessary when expressing $\sin\theta$ and $\sin\phi$ in terms of the cotangents because only even powers of the sine functions appear. Using the identity

$$e^{i\theta}=\cos\theta+i\sin\theta=\sin\theta\ \ (\cot\theta+i), \tag{3.16}$$

we obtain for the factor $e^{g(\theta,\phi)}$ in Eq. (3.13),

$$e^{2im_1\phi+2im_2\theta}\sin^{j+j_2-j_1}\theta\sin^{j_1+j_2-j}(\theta-\phi)\sin^{j+j_1-j_2}\phi$$
$$=(i+\cot\phi)^{2m_1}(i+\cot\theta)^{2m_2}\sin^{j+j_2-j_1+2m_2}\theta\sin^{j_1+j_2-j}(\theta-\phi)\sin^{j+j_1-j_2+2m_1}\phi$$
$$=(i+\cot\phi)^{2m_1}(i+\cot\theta)^{2m_2}\sin^{2j_2+2m_2}\theta(\cot\phi-\cot\theta)^{j_1+j_2-j}\sin^{2j_1+2m_1}\phi$$
$$=(i+\cot\phi)^{2m_1}(i+\cot\theta)^{2m_2}(1+\cot^2\theta)^{-j_2-m_2}(\cot\phi-\cot\theta)^{j_1+j_2-j}(1+\cot^2\phi)^{-j_1-m_1}$$
$$=\frac{(i+\cot\phi)^{m_1-j_1}}{(-i+\cot\phi)^{j_1+m_1}}\frac{(i+\cot\theta)^{m_2-j_2}}{(-i+\cot\theta)^{j_2+m_2}}(\cot\phi-\cot\theta)^{j_1+j_2-j}. \tag{3.17}$$

Using this equation and Eq. (3.15), all of the quantities in the expression in Eq. (3.13) can be expressed in term of the cotangents of θ and ϕ, given in Eq. (3.11). It should be noted that all of the exponents in Eq. (3.17) are integers, so choices of branch cuts are not necessary.

A. Allowed region

It is useful to introduce the concepts of a triangle-allowed region and a classically allowed region of the space of values for the quantum numbers. We define the triangle-allowed region to be the set of quantum numbers for which j_1, j_2, and j satisfy the triangle inequalities and for which the inequalities $|m|\leqslant j$ and $\{|m_i|\leqslant j_i,\ i=1,2\}$ hold. The Clebsch–Gordan coefficient is zero outside of this region, so it is only within this region that asymptotic expressions are desired. The triangle-allowed region is divided into a classically allowed region and a classically forbidden region. As is usual, we call these the allowed and forbidden regions for brevity. The allowed region is defined to be the set of quantum numbers for which it is possible to define **j**-vectors in a three-dimensional space in such a way that their lengths are equal to the *j*-values and their *z*-components are equal to the *m*-values (and, of course, such that $\mathbf{j}=\mathbf{j}_1+\mathbf{j}_2$). An example of such a construction for a set of allowed quantum numbers is shown in Fig. 1. It follows from the definition that the allowed region is contained in the triangle-allowed region. Examples of classically forbidden points are easily found in extreme cases, such as $m_1=j_1$. In this case, there is only one classically allowed value for m_2 (assuming a set of triangle-allowed *j*-values have been given), because the $\mathbf{j}_1$-vector must point in the *z* direction, and thus the *j*-triangle lies in a vertical plane.

The allowed region is the same as the region in which the three λ-values defined in Eq. (3.25) satisfy the triangle inequalities. This is because the λ-values are the lengths of the projections of

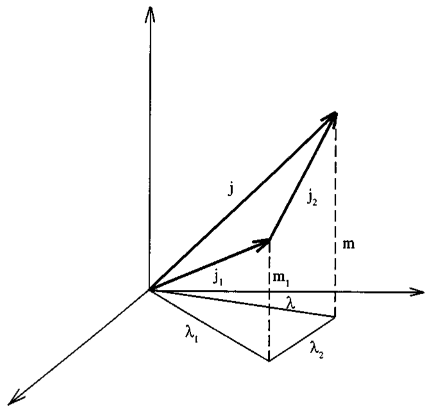

FIG. 1. An example of a choice of three **j**-vectors, demonstrating that a set of quantum numbers is classically allowed.

the **j**-vectors into the *xy* plane. If the λ-values satisfy the triangle inequalities, then it is possible to draw a triangle in the *xy* plane with sides equal to the λ-values. From this, one can construct the **j**-vectors by simply including the *m*-values as *z*-components. Conversely, if the **j**-vectors can be constructed, then their projections into the *xy* plane form a triangle (with the tail of $\mathbf{j}_2$ at the tip of $\mathbf{j}_1$), and the λ-values satisfy the triangle inequalities.

It is explained later in this paper that the λ-values satisfy the triangle inequalities if and only if the quantity $(-\lambda_1+\lambda_2+\lambda_3)(\lambda_1-\lambda_2+\lambda_3)(\lambda_1+\lambda_2-\lambda_3)$ is non-negative (that is, it is not possible for two of the factors to be negative). This observation together with the fact that the quantity β defined in Eq. (3.12) may be written as

$$\beta=\sqrt{(\lambda_1+\lambda_2+\lambda_3)(-\lambda_1+\lambda_2+\lambda_3)(\lambda_1-\lambda_2+\lambda_3)(\lambda_1+\lambda_2-\lambda_3)} \tag{3.18}$$

leads us to the result that the sign of β^2 distinguishes the allowed and forbidden regions: it is positive in the allowed region, and it is negative in the forbidden region. In the allowed region, β is four times the area of the triangle whose sides are the λ-values. This triangle is the projection of the *j*-triangle into the *xy* plane (see Fig. 1). From Eq. (3.12) it is apparent that for fixed values of the *j* quantum numbers, β^2 is a quadratic polynomial in the *m* quantum numbers. Thus, in the (m_1,m_2) plane the boundary between the allowed and forbidden regions is an ellipse. This is shown in Fig. 2 for one choice of values for j_1, j_2, and *j*. The boundary of the triangle-allowed region is the irregular hexagon. The forbidden region is composed of six subregions. The points that separate them are indicated in Fig. 2. These are the points where the ellipse that separates the allowed and forbidden regions is tangent to the hexagon that defines the triangle-allowed region. The coordinates of these points can be calculated from the expression for β and the equations for the straight-line sections of the boundary of the triangle-allowed region. The resulting coordinates of these points are indicated in the figure.

The calculations involved in the stationary-phase approximation of the integral expression for the Clebsch–Gordan coefficient are different in the allowed and forbidden regions. We will treat the allowed region first. The sum over stationary-phase points for the case where the set of quantum numbers is in the allowed region is a sum over both of the solutions for the cotangents of θ and ϕ given in Eq. (3.11). This is analogous to the behavior demonstrated in Appendix A, and it is also the same as in the calculation of the stationary-phase approximation of the Airy integral, which is the canonical example of a stationary-phase calculation. In the case of the Airy integral, there are allowed and forbidden regions in position space, and in the allowed region the contour of integration is deformed to run over both of the stationary-phase points. We define c_θ and c_ϕ to be the first solution for the cotangents in Eq. (3.11):

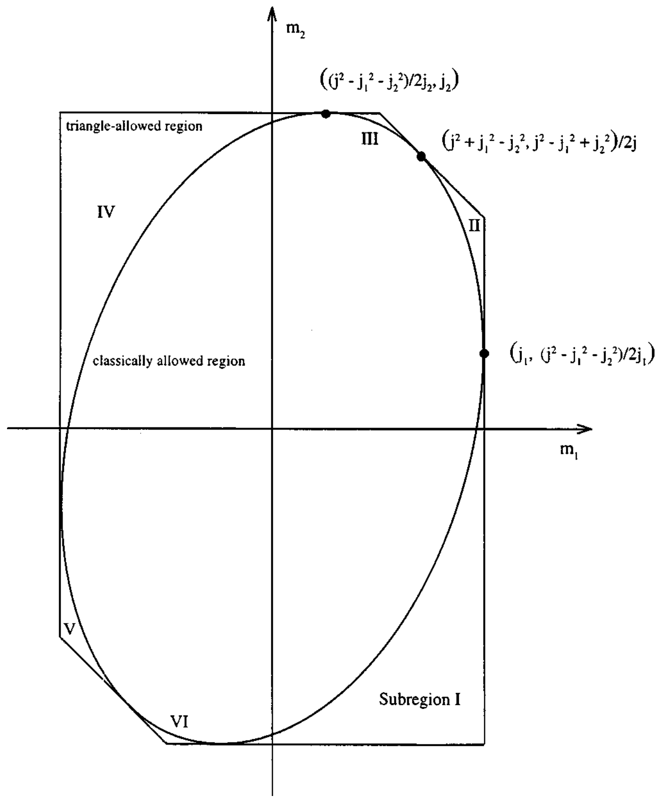

FIG. 2. The triangle-allowed region and the classically allowed region, shown in the m_1-m_2 plane, for the case of j-values in the ratio $j:j_1:j_2=4:2:3$. The six forbidden subregions are labeled with Roman numerals.

$$c_\theta=\frac{-2i(j_2m+m_2j)-\beta}{(j+j_2-j_1)(j_1+j_2+j)},$$
$$c_\phi=\frac{-2i(j_1m+m_1j)+\beta}{(j_1+j_2+j)(j+j_1-j_2)}. \tag{3.19}$$

Because of the form of the two solutions given in Eq. (3.11), the second solution is obtained from this one by multiplying by -1 and complex conjugating. Note this is only valid in the allowed region, where the quantity β is real. The first term in the sum over stationary-phase points is given by plugging the expression for $\det[\partial^2 g/\partial(\theta,\phi)^2]$, given in Eq. (3.15), and the expression for $e^{g(\theta,\phi)}$, given in Eq. (3.17), into the quantity in Eq. (3.13), using c_θ and c_ϕ for the cotangents. The second term in the sum over stationary-phase points is the same, except $-c_\theta^*$ and $-c_\phi^*$ are used for the cotangents. The result for the determinant in the second term is obtained by simply complex conjugating the first value, since all of the cotangents in this expression are squared. As for the $e^{g(\theta,\phi)}$ factor in the second term, we start by considering the expression for this factor in the first term:

$$\frac{(i+c_\phi)^{m_1-j_1}}{(-i+c_\phi)^{j_1+m_1}}\frac{(i+c_\theta)^{m_2-j_2}}{(-i+c_\theta)^{j_2+m_2}}(c_\phi-c_\theta)^{j_1+j_2-j}. \tag{3.20}$$

The complex conjugate of this is

$$\frac{(-i+c_\phi^*)^{m_1-j_1}}{(i+c_\phi^*)^{j_1+m_1}}\frac{(-i+c_\theta^*)^{m_2-j_2}}{(i+c_\theta^*)^{j_2+m_2}}(c_\phi^*-c_\theta^*)^{j_1+j_2-j}$$
$$=(-1)^{j_1+j_2+j}\frac{(i-c_\phi^*)^{m_1-j_1}}{(-i-c_\phi^*)^{j_1+m_1}}\frac{(i-c_\theta^*)^{m_2-j_2}}{(-i-c_\theta^*)^{j_2+m_2}}(-c_\phi^*+c_\theta^*)^{j_1+j_2-j}. \qquad (3.21)$$

This shows that the $e^{g(\theta,\phi)}$ factor in the second term [which appears after the $(-1)^{j_1+j_2+j}$ in the last line of Eq. (3.21)] is $(-1)^{j_1+j_2+j}$ times the complex conjugate of the $e^{g(\theta,\phi)}$ factor in the first term. Thus the second term is $(-1)^{j_1+j_2+j}$ times the complex conjugate of the first term. It is therefore convenient to obtain the sum over stationary-phase points by taking $i^{j_1+j_2+j}$ times the first term, adding the complex conjugate of this product, and then dividing by $i^{j_1+j_2+j}$. Using the fact that the real part of a quantity x is given by $\mathcal{R}[x]=(x+x^*)/2$, our stationary-phase approximation for the integral expression for the Clebsch–Gordan coefficient in Eq. (3.1) can be written as

$$\langle j_1m_1j_2m_2|jm\rangle\approx(-1)^{j+m}(2i)^{j+j_1+j_2}\pi^{-2}N_{j_1m_1j_2m_2jm}\frac{2}{i^{j_1+j_2+j}}\mathcal{R}\left[\frac{2\pi i^{j_1+j_2+j}}{\sqrt{\det[\partial^2 g/\partial(\theta,\phi)^2]}}e^{g(\theta,\phi)}\right]$$
$$=(-1)^{j+m}2^{j+j_1+j_2+2}\pi^{-1}N_{j_1m_1j_2m_2jm}\mathcal{R}\left[\frac{i^{j_1+j_2+j}}{\sqrt{\det[\partial^2 g/\partial(\theta,\phi)^2]}}e^{g(\theta,\phi)}\right], \qquad (3.22)$$

where the quantities $\det[\partial^2 g/\partial(\theta,\phi)^2]$ and $e^{g(\theta,\phi)}$ are obtained from Eqs. (3.15) and (3.17) using the c_θ and c_ϕ given in Eq. (3.19).

Although the expression in Eq. (3.22) gives a value that is a real number, it involves intermediate quantities that are complex. It is possible to transform this expression so that only real quantities are involved. This transformation is very lengthy, and it is not practical to describe it in detail here. Instead, we present an expression that is exactly equal to the expression in Eq. (3.22) in the allowed region. This equality can be verified most convincingly by substituting numerical values into the expressions and evaluating the results to high numerical precision (much higher than the level at which discrepancies would occur if order $\hbar$ terms were dropped). A brief description of the transformation is the following. Every complex quantity $x+iy$ that occurs in Eq. (3.22) is written as the product of a modulus and a phase, $\sqrt{x^2+y^2}\exp[i\tan^{-1}(y/x)]$, where care must be taken that correct branches are used for each $x+iy$, that is, one must examine the quantities x and y to determine the range of phase factors $(x+iy)/\sqrt{x^2+y^2}$ that can occur in the allowed region, and make branch choices accordingly. At some stages in the calculation, large polynomials are involved, and computer-aided symbol manipulation becomes useful in working with these. Our result may be put in the form

$$\langle j_1m_1j_2m_2|jm\rangle\approx 2I_{j_1m_1j_2m_2jm}\sqrt{\frac{j}{\pi\beta}}\cos\left[\chi+\frac{\pi}{4}-\pi(j+1)\right], \qquad (3.23)$$

where χ is defined to be

$$\chi=\left(j_1+\frac{1}{2}\right)\cos^{-1}\left[\frac{(-m)(j_1^2+j_2^2-j^2)-m_2(j_1^2+j^2-j_2^2)}{\alpha\lambda_1}\right]$$
$$+\left(j_2+\frac{1}{2}\right)\cos^{-1}\left[\frac{m_1(j^2+j_2^2-j_1^2)-(-m)(j_2^2+j_1^2-j^2)}{\alpha\lambda_2}\right]$$
$$+\left(j+\frac{1}{2}\right)\cos^{-1}\left[\frac{m_2(j_1^2+j^2-j_2^2)-m_1(j^2+j_2^2-j_1^2)}{\alpha\lambda_3}\right]$$
$$-m_1\cos^{-1}\left[\frac{\lambda_1^2+\lambda_3^2-\lambda_2^2}{2\lambda_1\lambda_3}\right]+m_2\cos^{-1}\left[\frac{\lambda_3^2+\lambda_2^2-\lambda_1^2}{2\lambda_2\lambda_3}\right], \tag{3.24}$$

and

$$\alpha=\sqrt{(j+j_1+j_2)(-j+j_1+j_2)(j-j_1+j_2)(j+j_1-j_2)},$$
$$\lambda_i=\sqrt{j_i^2-m_i^2}\quad i=1,2, \tag{3.25}$$
$$\lambda_3=\sqrt{j^2-m^2}.$$

(The quantity α is four times the area of the j-triangle shown in Fig. 1.) Note that the $\cos^{-1}$ functions in Eq. (3.24) are the usual principal branch, whose range is the interval from zero to π. The quantity $I_{j_1m_1j_2m_2jm}$ is defined to be

$$I_{j_1m_1j_2m_2jm}=\sqrt{\frac{(j+1/2)(j+j_1+j_2)}{j(j+j_1+j_2+1)}}$$
$$\times\frac{f(j_1+m_1)f(j_1-m_1)f(j_2+m_2)f(j_2-m_2)f(j+m)f(j-m)}{f(j_1+j_2+j)f(j_1+j_2-j)f(j_1-j_2+j)f(-j_1+j_2+j)}, \tag{3.26}$$

where the function f is defined to be

$$f(n)=\sqrt{\frac{n!}{\sqrt{2\pi n}n^ne^{-n}}}, \tag{3.27}$$

that is, $f(n)$ is the square root of the ratio of $n!$ to the Stirling approximation of $n!$ Note that for large n, $f(n)$ approaches one. Thus, for large quantum numbers, $I_{j_1m_1j_2m_2jm}$ approaches one. It differs from unity by a correction that is order $\hbar$, as can be deduced from the discussion of the Stirling approximation in Appendix A. As mentioned above, we present our approximation in the form given in Eq. (3.23) so that the exact equality of this expression and the complex expression given in Eq. (3.22) can be verified numerically. The factor $I_{j_1m_1j_2m_2jm}$ may be dropped without reducing the quality of the approximation, that is, the ratio of our approximation to the exact value differs from unity by a quantity that is order $\hbar$. Thus, we may write our approximation in the form

$$\langle j_1m_1j_2m_2|jm\rangle\approx 2\sqrt{\frac{j}{\pi\beta}}\cos\left[\chi-\pi\left(j+\frac{3}{4}\right)\right]. \tag{3.28}$$

Ponzano and Regge[4] give a geometrical interpretation of the five angles that occur in the expression for χ in Eq. (3.24). An equation similar to Eq. (3.28) also appears in Ref. 5, but the $(j+1/2)$ factors in χ are included at the end of the calculation to improve the accuracy, and the $\pi(j+1)$ in Eq. (3.23) is missing so that the formula gives the wrong sign for even j-values and does not give the right magnitude for half-integer j values.

TABLE I. For each forbidden subregion, the choice of root in Eq. (3.11), the sign function as in Eq. (3.30), and the largest λ [which determines the form of χ, as in Eq. (3.31)] are given.

Forbidden subregion	Choice of root	Sign function	Largest λ
I	lower	1	λ_3
II	upper	$(-1)^{j_1-m_1}$	λ_2
III	lower	$(-1)^{j_1-j+m_2}$	λ_1
IV	upper	$(-1)^{j_1+j_2-j}$	λ_3
V	lower	$(-1)^{j_2-j-m_1}$	λ_2
VI	upper	$(-1)^{j_2+m_2}$	λ_1

B. Forbidden region

In the forbidden region, only one of the stationary-phase points is used in the approximation. This is analogous to the situation in the Airy function problem mentioned above, where there are dominant and subdominant branches, and in the forbidden region only the subdominant branch exists. Similarly, the model problem in Appendix A shows how for the case of $m>n$, two stationary-phase points are used, while for the case $m<n$ only one stationary-phase point is involved. The choice of which of the two roots in Eq. (3.11) is to be used for our approximation of the Clebsch–Gordan coefficient is indicated in Table I. Given the m-values of a point in the forbidden region in Fig. 2, it is inconvenient to determine which subregion it is in by using nested if–then statements, because the relative ordering of, say, the m_2-coordinates of the points on the boundaries between the forbidden subregions changes as the j-values are changed. A much simpler way to determine which branch to use is to find the sign of a certain polynomial which we describe here. As can be seen from Table I, the choice of branch alternates as one goes around the diagram in Fig. 2. Thus we use the sign of the product of three expressions that flip signs in the right way. Given the coordinates of one of the boundary points in the (m_1,m_2)-plane, a vector perpendicular to it can be constructed by exchanging the coordinates and changing the sign of one of them. The dot-product of this vector and (m_1,m_2) is a function on the (m_1,m_2)-plane that changes sign at the boundary between the two subregions in question. Thus we are led to consider the sign of the function

$$[(m_1,m_2)\cdot(-2j_2^2,j^2-j_1^2-j_2^2)][(m_1,m_2)\cdot(-j^2+j_1^2-j_2^2,j^2+j_1^2-j_2^2)]$$
$$\times[(m_1,m_2)\cdot(-j^2+j_1^2+j_2^2,2j_1^2)].$$

If this quantity is positive (negative), then the upper (lower) choice of root in Eq. (3.11) is used. Once the cotangents of the angles at the stationary-phase point are determined, the approximation of the Clebsch–Gordan coefficient can be evaluated from the expression

$$\langle j_1m_1j_2m_2|jm\rangle\approx(-1)^{j+m}(2i)^{j+j_1+j_2}\pi^{-2}N_{j_1m_1j_2m_2jm}\frac{2\pi e^{g(\theta,\phi)}}{\sqrt{\det[\partial^2 g/\partial(\theta,\phi)^2]}}. \tag{3.29}$$

All of the quantities needed to evaluate this expression were expressed in terms of the cotangents of the angles in Eqs. (3.15) and (3.17). It may be noted that in the forbidden region the cotangents become pure imaginary, as can be seen from Eq. (3.11). This behavior is similar to that in the model problem in Appendix A, where the angle suddenly jumps in terms of its real part (but the analogy is not perfect because in the model problem the cotangent is pure imaginary in both the region $m>n$ and the region $m<n$).

Evaluating Eq. (3.29) results in a real value, although complex numbers are involved at intermediate steps. As in the case of our analysis in the allowed region, the expression may be transformed to a form that involves only operations with real numbers. This can be done in a way that parallels the previous calculation, with hyperbolic functions playing the role of trigonometric functions. The transformation involves choices of branch cuts and depends on which of the six subregions of the forbidden region one is working in. Thus there are six different all-real expressions for the forbidden region. In the interest of brevity, we will present only one of these here. In subregion VI, the expression in Eq. (3.29) is exactly equal to

$$(-1)^{j_2+m_2} 2 I_{j_1 m_1 j_2 m_2 j m} \sqrt{\frac{j}{\pi|\beta|}} \exp(-\chi^{(\mathrm{vi})}), \tag{3.30}$$

where $\chi^{(\mathrm{vi})}$ is defined to be

$$\begin{aligned}\chi^{(\mathrm{vi})}=&\left(j_1+\frac{1}{2}\right)\cosh^{-1}\left[\frac{-m(j_1^2+j_2^2-j^2)-m_2(j_1^2+j^2-j_2^2)}{\alpha\lambda_1}\right]\\&-\left(j_2+\frac{1}{2}\right)\cosh^{-1}\left[\frac{-m_1(j^2+j_2^2-j_1^2)-m(j_2^2+j_1^2-j^2)}{\alpha\lambda_2}\right]\\&-\left(j+\frac{1}{2}\right)\cosh^{-1}\left[\frac{-m_2(j_1^2+j^2-j_2^2)+m_1(j^2+j_2^2-j_1^2)}{\alpha\lambda_3}\right]\\&-m\cosh^{-1}\left[\frac{\lambda_1^2+\lambda_3^2-\lambda_2^2}{2\lambda_1\lambda_3}\right]-m_2\cosh^{-1}\left[\frac{\lambda_1^2+\lambda_2^2-\lambda_3^2}{2\lambda_2\lambda_1}\right].\end{aligned} \tag{3.31}$$

This all-real expression was derived by a very lengthy calculation, as in the case of the analysis in the allowed region. Again, an exact equality such as the one above can be checked easily by substituting in test numbers and evaluating to sufficient precision. As before, to actually use the approximation, one would drop the factor of $I_{j_1 m_1 j_2 m_2 j m}$ since it can be approximated by unity, to the order that we are working in this section. All-real expressions for the other subregions of the forbidden region can most easily be obtained by using the symmetries of the Clebsch–Gordan coefficients to related the expressions for the different subregions. If one prefers not to work with six different expressions for the forbidden region, one can use the polynomial discussed above to select the required stationary-phase point and then plug this into the approximation given in Eq. (3.29). This requires operations with complex numbers, but is easier to implement in a computer program. Alternatively, to obtain an approximate value for the Clebsch–Gordan coefficient for a given point in the forbidden region, one could work with only one all-real expression for a particular forbidden subregion and use the symmetries of the Clebsch–Gordan coefficients to map the given point to a point that is within the subregion for which the expression is valid.

It is interesting to compare the all-real expressions obtained in the allowed region, Eq. (3.23), and in the forbidden region, Eq. (3.30). They are similar in form, but the behavior is oscillatory in the allowed region and exponentially decaying in the forbidden region. This is the behavior expected in quantum mechanical problems that have an allowed region and a forbidden region.

In the forbidden region, writing the approximation in an all-real form is illuminating because it makes it apparent that sign functions exist. We call the factor $(-1)^{j_2+m_2}$ in Eq. (3.30) a sign function. The remaining factors in that equation are all positive, so the sign function gives the sign of the result. However, since the result is an approximation of a Clebsch–Gordan coefficient, the sign function also gives the sign of the Clebsch–Gordan coefficient, at least in the asymptotic regime. Thus, the sign functions are actually properties of the Clebsch–Gordan coefficients themselves, for a given choice of phase conventions. We are using the conventions defined by Eq. (2.1). The existence of sign functions was not clear from Eq. (2.1), which was our starting point. The sign functions for each of the six forbidden subregions are given in Table I.

The angle χ, given in Eq. (3.24), that appears in our approximation in the allowed region can be rewritten in several different ways. The reason is that the angles that multiply the m's in the equation for χ are two of the interior angles in the triangle formed by the three λ-values. If we call these angles α_1, α_2, and α_3 (where α_i is the angle opposite the side of length λ_i), then we have

$$\begin{aligned} m_1+m_2&=m, \\ \alpha_1+\alpha_2+\alpha_3&=\pi. \end{aligned} \tag{3.32}$$

Thus, the vectors $(m_1, m_2, -m)$ and $(\alpha_1, \alpha_2, \alpha_3-\pi)$ are both perpendicular to (1, 1, 1), and their cross product is parallel to (1, 1, 1). Each of the three components of their cross product are thus equal, and each one could be used as part of χ in the allowed region,

$$m_1\alpha_2-m_2\alpha_1=m_2(\alpha_3-\pi)+m\alpha_2=-m\alpha_1-m_1(\alpha_3-\pi). \tag{3.33}$$

In the forbidden region there is no such flexibility in how to write the corresponding terms, because there do not exist three angles corresponding to α_1, α_2, and α_3. This is because it is not possible to form a triangle using the three λ-values. Three different quantities like

$$\frac{\lambda_3^2+\lambda_1^2-\lambda_2^2}{2\lambda_3\lambda_1} \tag{3.34}$$

can be written down by cyclically permuting the indices, but only two of these can be used as arguments of the $\cosh^{-1}$ function in an all-real expression. This can be seen in the following way. The λ's are non-negative and if three non-negative numbers fail to satisfy the triangle inequalities, exactly one triangle inequality is violated. [Proof: Let $\lambda_{\max}$ be the largest value, λ_{mid} be the middle value, and $\lambda_{\min}$ be the smallest. Then $-\lambda_{\min}+\lambda_{\text{mid}}+\lambda_{\max}\geqslant 0$ and $\lambda_{\min}-\lambda_{\text{mid}}+\lambda_{\max}\geqslant 0$, so we must have $\lambda_{\min}+\lambda_{\text{mid}}-\lambda_{\max}<0$.] Now we consider rewriting the expression

$$\frac{\lambda_3^2+\lambda_1^2-\lambda_2^2}{2\lambda_3\lambda_1}=1+\frac{(\lambda_3-\lambda_1)^2-\lambda_2^2}{2\lambda_3\lambda_1}=1-\frac{(\lambda_1+\lambda_2-\lambda_3)(\lambda_2+\lambda_2-\lambda_1)}{2\lambda_3\lambda_1}. \tag{3.35}$$

This shows that of the three permutations of the expression in Eq. (3.34), exactly two will be greater than unity. It is these two that must be used as arguments of the $\cosh^{-1}$ function in an all-real expression. Thus there is no flexibility in ways to write the m-terms in χ as in the allowed region. Throughout each one of the six subregions of the forbidden region, a single triangle inequality for the λ's is violated. It is not possible that one triangle inequality is violated in one part of a subregion and another triangle inequality is violated in another part of the same subregion because at the boundary between these two parts β would be zero, as can be seen from Eq. (3.18). However, β^2 is a quadratic polynomial in m_1 and m_2 [see Eq. (3.12)], the zero-contour of which is the ellipse in Fig. 2, so it is not possible for it to be zero along a curve in the forbidden region. The λ that is largest in each subregion is indicated in Table I. The forms of all-real expressions in each of the forbidden subregions will reflect the fact that in each one of the subregions one of the λ's is larger than the sum of the other two.

It remains to discuss the case of points that are on the boundary between the allowed and forbidden regions. The quantity β in Eq. (3.12) is zero on this boundary, and since β is invariant under the full 72-element symmetry group of the 3-j symbol,[10] the Clebsch–Gordan coefficient cannot be approximated on the boundary with the formulas presented in this paper. The reason is that β occurs in the denominator in Eqs. (3.28) and (3.30). Since β^2 is a homogeneous polynomial in the quantum numbers, it will be zero for sets of quantum numbers equal to any multiple of a set of quantum numbers for which β is zero. The behavior of the Clebsch–Gordan coefficients in the direction transverse to the boundary should be similar to that of the Airy function (see Ref. 4).

The invariance of β under the 72-element symmetry group of the 3-j symbols may be shown as follows. We begin by constructing the 3×3 Regge array of linear combinations of quantum numbers, given in Ref. 10. For any integer n, we define the polynomial p_n to be the sum of the nth powers of the nine elements of this matrix. These polynomials are invariant under the symmetry group, because if two sets of quantum numbers are related by a Regge symmetry, we can construct the 3×3 Regge array for each set and compute p_n. The results are the same because of the commutativity of addition. It is possible to write β^2 in terms of the p_n:

$$\beta^2=(p_1^4-6p_2p_1^2-27p_2^2+108p_4)/324. \tag{3.36}$$

The coefficients in this equation may be simplified slightly by using the relation $p_1=3(j_1+j_2+j)$. This equation proves the invariance of β under the symmetry group.

An example of quantum numbers for which β is zero is

$$(j_1,m_1,j_2,m_2,j,m)=(3,-2,6,4,7,2). \tag{3.37}$$

This point is not on the edge of the triangle-allowed region. Points for which β is zero and which are on the edge of the triangle-allowed region are easier to find. For example, one can choose $(j_1,m_1)=(j_2,m_2)=\frac{1}{2}(j,m)$. For such a point, j_1+j_2-j is zero.

The reason we are unable to approximate the Clebsch–Gordan coefficient for cases in which β is zero is that the determinant of the 2×2 matrix of second derivatives of $g(\theta,\phi)$ is zero at the stationary-phase points. This can be shown by plugging the solutions for the cotangents of θ and ϕ at a stationary-phase point into Eq. (3.15) for the determinant; the result has an overall factor of β after being simplified [see Eq. (B7)]. This determinant appears in the denominator of Eq. (3.13), so our method cannot be applied. Note that when β is zero, the two solutions for the cotangents at the stationary-phase point are the same [see Eq. (3.11)]. Also, it should be noted that if any of the m-values has its absolute value close to the corresponding j, then the corresponding λ will be small [see Eq. (3.25)], and the area of the λ-triangle will be small. Thus, β will be small, and the set of quantum numbers is close to the boundary. In contrast to this, there is no difficulty with the approximation if the m values are close to zero. These considerations are mirrored in the approximation (using Stirling's formula) of $N_{j_1m_1j_2m_2jm}$, defined in Eq. (2.3); no factorials of m-values appear, only factorials of $j-m, j+m$, etc.

IV. HIGHER-ORDER APPROXIMATION

The methods used in the previous sections can be extended to higher order. In this section, we derive the next correction to the previous results. The approximation that is obtained in this way gives results that are accurate to six digits, for example, when the quantum numbers are in the hundreds.

Let (θ_0,ϕ_0) be a stationary-phase point, i.e., a point at which $\partial g/\partial\theta=\partial g/\partial\phi=0$. We write the Taylor expansion of the function $g(\theta,\phi)$ about the point (θ_0,ϕ_0) as a sum of homogeneous polynomials,

$$g(\theta_0+x,\phi_0+y)=g_0+g_2+g_3+g_4+\cdots, \tag{4.1}$$

where

$$\begin{aligned} g_0&=g(\theta_0,\phi_0),\\ g_2&=g_{\theta\theta}x^2/2+g_{\theta\phi}xy+g_{\phi\phi}y^2/2,\\ g_3&=g_{\theta\theta\theta}x^3/6+g_{\theta\theta\phi}x^2y/2+g_{\theta\phi\phi}xy^2/2+g_{\phi\phi\phi}y^3/6,\\ g_4&=g_{\theta\theta\theta\theta}x^4/24+g_{\theta\theta\theta\phi}x^3y/6+g_{\theta\theta\phi\phi}x^2y^2/4+g_{\theta\phi\phi\phi}xy^3/6+g_{\phi\phi\phi\phi}y^4/24, \end{aligned} \tag{4.2}$$

where, for example, $g_{\theta\theta\phi}$ is defined to be $\partial^3g/\partial\theta^2\partial\phi$ at the stationary-phase point.

To obtain the next higher stationary-phase approximation for the Clebsch–Gordan coefficient, we terminate the series in Eq. (4.1) at the fourth-order term. The reason for this is explained below. Thus, the approximation of the function g has derivatives at the stationary-phase point (θ_0,ϕ_0) that agree with those of g through fourth order. Our approximation of the integrand $\exp(g)$ is

$$\begin{aligned}\exp[g(\theta_0+x,\phi_0+y)]&\approx\exp(g_0)\exp(g_2)\exp(g_3)\exp(g_4)\\&=\exp(g_0)\exp(g_2)(1+g_3+g_3^2/2\cdots)(1+g_4+g_4^2/2+\cdots).\end{aligned}\tag{4.3}$$

When this is multiplied out, each of the terms may be integrated over the xy plane in closed form. We are interested in the asymptotic behavior of the resulting terms. The question is how the terms behave when all of the quantum numbers (j_1,m_1,j_2,m_2,j,m) are multiplied by the same factor (such a factor is called $1/\hbar$, as explained in the Introduction). The stationary-phase point (θ_0,ϕ_0) is independent of the factor, i.e., (θ_0,ϕ_0) is order $\hbar^0$, as can be seen from Eq. (3.11). The second derivatives of g at (θ_0,ϕ_0) are order $1/\hbar$, as can be seen from Eq. (3.14), so to see how the integral of $\exp(g_2)$ depends on $\hbar$, we define new variables of integration to be the old variables times $\hbar^{-1/2}$. From this it follows that the integral of $\exp(g_2)$ is order $\hbar$. By the same reasoning, the integral of a homogeneous quartic polynomial times $\exp(g_2)$ is order $\hbar^3$, and the integral of a homogeneous sixth-order polynomial times $\exp(g_2)$ is order $\hbar^4$. The polynomials g_4 and g_3^2 have coefficients that are order $1/\hbar$ and $1/\hbar^2$, respectively, so the integrals of these times $\exp(g_2)$ are both order $\hbar^2$. This is one order of $\hbar$ smaller than the integral of $\exp(g_2)$. The integral of a homogeneous polynomial of odd degree times $\exp(g_2)$ vanishes due to antisymmetry. Thus, the next-higher-order approximation of the integral of $\exp(g)$ is obtained by integrating

$$\exp(g_0)\,\exp(g_2)(1+g_4+g_3^2/2).\tag{4.4}$$

Terms coming from g_5, etc., contribute at higher orders.

To find the ratio of the integral of $g_4\exp(g_2)$ to the integral of $\exp(g_2)$ the following integrals are necessary:

$$\begin{aligned}i_1&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(g_2)\,dx\,dy,\\i_2&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^4\exp(g_2)\,dx\,dy,\\i_3&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^3y\exp(g_2)\,dx\,dy,\\i_4&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^2y^2\exp(g_2)\,dx\,dy.\end{aligned}\tag{4.5}$$

We will need the ratios i_2/i_1, i_3/i_1, and i_4/i_1. The integrals are tabulated and these ratios can be worked out without the use of any information about relationships between the various derivatives of the function g at the stationary-phase point. The results are

$$\begin{aligned}i_2/i_1&=\frac{3g_{\phi\phi}^2}{(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^2},\\i_3/i_1&=\frac{-3g_{\theta\phi}g_{\phi\phi}}{(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^2},\\i_4/i_1&=\frac{2g_{\theta\phi}^2+g_{\theta\theta}g_{\phi\phi}}{(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^2}.\end{aligned}\tag{4.6}$$

It may be noted that $g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2$ is the determinant of the 2×2 matrix of second partial derivatives of the function g.

The ratio, which we denote by δ_4, of the integral of $g_4 \exp(g_2)$ to the integral of $\exp(g_2)$ works out to be

$$\delta_4=\frac{g_{\theta\theta\theta\theta}g_{\phi\phi}^2-4g_{\theta\theta\theta\phi}g_{\theta\phi}g_{\phi\phi}+2g_{\theta\theta\phi\phi}(2g_{\theta\phi}^2+g_{\theta\theta}g_{\phi\phi})-4g_{\theta\phi\phi\phi}g_{\theta\theta}g_{\theta\phi}+g_{\phi\phi\phi\phi}g_{\theta\theta}^2}{8(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^2}. \tag{4.7}$$

Next, we move on to the g_3^2 term in Eq. (4.4). To find the ratio of the integral of $\frac{1}{2}g_3^2 \exp(g_2)$ to the integral of $\exp(g_2)$ the following integrals are necessary:

$$\begin{aligned}
i_5&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^6 \exp(g_2)\,dx\,dy,\\
i_6&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^5 y \exp(g_2)\,dx\,dy,\\
i_7&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^4 y^2 \exp(g_2)\,dx\,dy,\\
i_8&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^3 y^3 \exp(g_2)\,dx\,dy.
\end{aligned} \tag{4.8}$$

We will need the ratios i_5/i_1, i_6/i_1, i_7/i_1, and i_8/i_1. As in the case of the g_4 calculation, the integrals are tabulated and these ratios can be worked out without the use of any information about relationships between the various derivatives of the function g at the stationary-phase point. The results are

$$\begin{aligned}
i_5/i_1&=\frac{-15g_{\phi\phi}^3}{(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^3},\\
i_6/i_1&=\frac{15g_{\theta\phi}g_{\phi\phi}^2}{(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^3},\\
i_7/i_1&=\frac{-3g_{\phi\phi}(4g_{\theta\phi}^2+g_{\theta\theta}g_{\phi\phi})}{(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^3},\\
i_8/i_1&=\frac{3g_{\theta\phi}(2g_{\theta\phi}^2+3g_{\theta\theta}g_{\phi\phi})}{(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^3}.
\end{aligned} \tag{4.9}$$

The ratio, which we denote by δ_6, of the integral of $\frac{1}{2}g_3^2 \exp(g_2)$ to the integral of $\exp(g_2)$ works out to be

$$\begin{aligned}
\delta_6=[&2g_{\theta\phi}(3g_{\theta\theta}g_{\phi\phi}+2g_{\theta\phi}^2)(g_{\theta\theta\theta}g_{\phi\phi\phi}+9g_{\theta\theta\phi}g_{\theta\phi\phi})\\
&-3(g_{\theta\theta}g_{\phi\phi}+4g_{\theta\phi}^2)(2g_{\theta\theta\theta}g_{\theta\phi\phi}g_{\phi\phi}+2g_{\phi\phi\phi}g_{\phi\theta\theta}g_{\theta\theta}+3g_{\theta\theta\phi}^2g_{\phi\phi}+3g_{\phi\phi\theta}^2g_{\theta\theta})\\
&+30g_{\theta\phi}(g_{\theta\theta\theta}g_{\theta\theta\phi}g_{\phi\phi}^2+g_{\phi\phi\phi}g_{\phi\phi\theta}g_{\theta\theta}^2)-5(g_{\theta\theta\theta}^2g_{\phi\phi}^3+g_{\phi\phi\phi}^2g_{\theta\theta}^3)]/[24(g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2)^3].
\end{aligned} \tag{4.10}$$

The only remaining matter is to find the necessary values of the higher derivatives of the function g at the stationary-phase point. The definition of g, given in Eq. (3.3), can be used to find its second derivatives, given in Eq. (3.14), its third derivatives,

$$\frac{\partial^3 g}{\partial\theta^3}=2(j+j_2-j_1)\csc^2\theta\ \cot\theta+2(j_1+j_2-j)\ \csc^2(\theta-\phi)\cot(\theta-\phi),$$

$$\frac{\partial^3 g}{\partial\theta^2\partial\phi}=-2(j_1+j_2-j)\ \csc^2(\theta-\phi)\ \cot(\theta-\phi),$$

$$\frac{\partial^3 g}{\partial\theta\partial\phi^2}=2(j_1+j_2-j)\ \csc^2(\theta-\phi)\ \cot(\theta-\phi), \tag{4.11}$$

$$\frac{\partial^3 g}{\partial\phi^3}=2(j_1+j_2-j)\csc^2(\theta-\phi)\ \cot(\phi-\theta)+2(j+j_1-j_2)\ \csc^2\phi\cot\phi,$$

and its fourth derivatives,

$$\frac{\partial^4 g}{\partial\theta^4}=-2(j+j_2-j_1)\ \csc^2\theta\ \ (3\cot^2\theta+1)-2(j_1+j_2-j)\csc^2(\theta-\phi)[3\cot^2(\theta-\phi)+1],$$

$$\frac{\partial^4 g}{\partial\theta^3\partial\phi}=2(j_1+j_2-j)\ \csc^2(\theta-\phi)[3\cot^2(\theta-\phi)+1],$$

$$\frac{\partial^4 g}{\partial\theta^2\partial\phi^2}=-2(j_1+j_2-j)\ \csc^2(\theta-\phi)[3\cot^2(\theta-\phi)+1], \tag{4.12}$$

$$\frac{\partial^4 g}{\partial\theta\partial\phi^3}=2(j_1+j_2-j)\ \csc^2(\theta-\phi)[3\cot^2(\theta-\phi)+1],$$

$$\frac{\partial^4 g}{\partial\phi^4}=-2(j_1+j_2-j)\ \csc^2(\theta-\phi)[3\cot^2(\theta-\phi)+1]-2(j+j_1-j_2)\ \csc^2\phi\ \ (3\cot^2\phi+1).$$

Given the values of the cotangents of θ and ϕ at a stationary-phase point, these derivatives can be evaluated without having to find the angles, i.e., without having to make any choices of branch cuts. One way to do this is to use Eq. (3.5) to evaluate $\cot(\theta-\phi)$ and the identity $\csc^2\theta=1+\cot^2\theta$ to evaluate the squared cosecants.

The relationship between the Clebsch–Gordan coefficient and the integral of exp (g) is given in Eq. (3.1). Combining this with our higher-order approximation for the integral results in the following higher-order approximation for the Clebsch–Gordan coefficient. Each stationary-phase point contributes

$$(-1)^{j+m}(2i)^{j+j_1+j_2}\pi^{-2}N_{j_1m_1j_2m_2jm}\frac{2\pi e^{g(\theta_0,\phi_0)}}{\sqrt{g_{\theta\theta}g_{\phi\phi}-g_{\theta\phi}^2}}(1+\delta_4+\delta_6). \tag{4.13}$$

As explained in the previous sections, for points in the allowed region the sum over stationary-phase points is a sum over both of the solutions for the cotangents of θ and ϕ given in Eq. (3.11), and for points in the forbidden region only one of these solutions contributes. Given values for $\cot\theta$ and $\cot\phi$, Eqs. (3.14), (4.11), and (4.12) are used to evaluate the higher derivatives of the function g at the stationary-phase point. Then Eqs. (4.7) and (4.10) are used to obtain δ_4 and δ_6.

As shown in Eq. (3.17), the quantity $e^{g(\theta_0,\phi_0)}$ can also be evaluated using the values of the cotangents of θ and ϕ. The quantity $N_{j_1 m_1 j_2 m_2 j m}$ can be approximated to sufficient accuracy using the next correction to Stirling's approximation for the factorials.

Finally, we present some numerical examples. We begin with the allowed region.

For $(j_1,m_1,j_2,m_2,j,m)=(200,\ 100,\ 300,\ 150,\ 400,\ 250)$, the values are

$$\begin{aligned} \text{exact} &= 0.070\,349\,9, \\ \text{approx} &= 0.070\,349\,6. \end{aligned} \tag{4.14}$$

For $(j_1,m_1,j_2,m_2,j,m)=(200,\ 100,\ 300+\frac{1}{2},\ 150+\frac{1}{2},\ 400+\frac{1}{2},\ 250+\frac{1}{2})$, the values are

$$\begin{aligned} \text{exact} &= 0.073\,063\,6, \\ \text{approx} &= 0.073\,063\,3. \end{aligned} \tag{4.15}$$

In the forbidden region, the Clebsch–Gordan coefficients are much smaller. The following examples are from subregion I.

For $(j_1,m_1,j_2,m_2,j,m)=(200,\ 150,\ 300,-250,\ 400,-100)$, the values are

$$\begin{aligned} \text{exact} &= 3.089\,61\times 10^{-19}, \\ \text{approx} &= 3.089\,58\times 10^{-19}. \end{aligned} \tag{4.16}$$

For $(j_1,m_1,j_2,m_2,j,m)=(200,\ 150,\ 300+\frac{1}{2},-250+\frac{1}{2},\ 400+\frac{1}{2},-100+\frac{1}{2})$, the values are

$$\begin{aligned} \text{exact} &= 5.327\,18\times 10^{-19}, \\ \text{approx} &= 5.327\,12\times 10^{-19}. \end{aligned} \tag{4.17}$$

Further examples of results from the higher-order approximation are discussed in Appendix B.

V. CONCLUSION

The methods presented in this paper provide simple formulas for calculating first-order approximations to Clebsch–Gordan coefficients in the allowed region and in all of the forbidden subregions. Additionally, a higher-order approximation is derived, although the expressions are more complicated. We do not know if the quantity $\delta_4+\delta_6$ in Eq. (4.13) can be simplified when expressed in terms of the quantum numbers (see Appendix B for a special case). It appears to be complicated, as is often the case for higher-order approximations. The geometrical structure is not as clear.

Our higher-order approximation provides the only known way to compute certain digits of some Clebsch–Gordan coefficients. By this we mean that given any computer, we can always find quantum numbers large enough so that the exact calculation is not feasible. The beginning digits may be calculated using first-order approximations; the higher-order approximation makes it possible to compute further digits.

The methods of this paper could also be used to derive asymptotic expressions for the $6j$-symbols, etc. The starting point would again be an exact expression for the quantity of interest. One would then have to construct a polynomial with the property that the coefficient of one of its terms is this exact expression. Then an integral expression would be obtained, and finally this integral would be approximated using the stationary-phase method.

As mentioned in the Introduction, this work could have applications in high-angular momentum calculations and theoretical investigations which contain sums over large numbers of Clebsch–Gordan coefficients.[8,9]

Our analysis in the forbidden region led us to the realization that simple sign functions exist there that give the sign of the exact Clebsch–Gordan coefficients. These are summarized in Table I.

A subject for future work is the approximation of Clebsch–Gordan coefficients and $6j$-symbols near the boundary between the allowed and forbidden regions. Ponzano and Regge[4] have conjectured and supplied numerical evidence for a typical Airy function caustic behavior. Also, of course, it should be possible to extend the present calculations to even higher orders.

APPENDIX A: A ONE-DIMENSIONAL EXAMPLE

In this appendix we consider a one-dimensional example of an integral that gives a Fourier coefficient of a function which is an integer power of a fixed function. We are interested in the asymptotics of the result for large values of the two integers involved.

We define the function $F(m,n)$ for positive integers m and n by

$$F(m,n)=\int_{-\pi/2}^{\pi/2}\cos^n x\, e^{imx}\, dx. \tag{A1}$$

It is possible to evaluate this integral exactly in closed form:

$$F(m,n)=\begin{cases} 2^{-n}\pi\binom{n}{(n-m)/2}, & n-m \text{ even} \\ (-1)^{(n+1-m)/2}\, 2^{n+2} n!\dfrac{[(m+n+1)/2]!\,(m-n-1)!}{(m+n+1)!\,[(m-n-1)/2]!}, & n-m \text{ odd}, n<m \\ 2^{n+2} n!\dfrac{[(n+1-m)/2]!\,[(n+1+m)/2]!}{(n+1-m)!\,(n+1+m)!}, & n-m \text{ odd}, n>m\,. \end{cases} \tag{A2}$$

In deriving these results, one uses the definition of the beta function, $B(z+1,w+1)=\int_0^1 t^z(1-t)^w\,dt$, and the relation between the beta function and the gamma function, $B(z,w)=\Gamma(z)\Gamma(w)/\Gamma(z+w)$. One also uses the results that for integers $k\geqslant 0$,

$$\begin{aligned}\Gamma\left(k+\frac{1}{2}\right)&=\frac{\sqrt{\pi}(2k)!}{2^{2k}k!},\\ \Gamma\left(-k+\frac{1}{2}\right)&=(-1)^k\frac{2^{2k}\sqrt{\pi}k!}{(2k)!}.\end{aligned} \tag{A3}$$

1. Asymptotics of the exact expressions

In order to compare the stationary-phase approximations derived in the following subsection with the exact value of $F(m,n)$, we will use Stirling's approximation for the factorials in the exact expressions in Eq. (A2). The accuracy to which we will work is that the ratio of the exact value to the approximation should go to unity as n and m go to infinity, holding the ratio of n to m fixed. The difference between the logarithm of the exact expression and the logarithm of the approximation thus goes to zero as the two integers get large (the errors are of order $1/n$).

Stirling's approximation, through order unity (for the logarithms), is

$$x!\approx\sqrt{2\pi x}\,x^x e^{-x}. \tag{A4}$$

The next correction to this is a multiplicative factor of $e^{1/(12x)}$. Thus, the ratio of $x!$ to the approximation given in Eq. (A4) approaches unity as x goes to infinity.

Our approximation of the exact expression for $F(m,n)$ works out to be

$$F(m,n)\approx\begin{cases}\sqrt{\dfrac{2\pi}{n}}\left(\dfrac{1-m/n}{1+m/n}\right)^{m/2}\left[1-\left(\dfrac{m}{n}\right)^2\right]^{-(n+1)/2}, & n>m;\\ 0, \quad n-m \text{ even}, n<m;\\ (-1)^{(n+1-m)/2}\,2\sqrt{\dfrac{2\pi}{n}}\left(\dfrac{m/n-1}{m/n+1}\right)^{m/2}\left[\left(\dfrac{m}{n}\right)^2-1\right]^{-(n+1)/2}, & n-m \text{ odd}, n<m.\end{cases} \tag{A5}$$

In deriving this result, we have used the fact that the inequality $n>m$ implies $n-m\gg 1$. This is true because we are holding the ratio of the two integers fixed while letting them become large. In other words, errors of order $1/n$ are the same order as errors of order $1/(n-m)$. Thus the Stirling approximation is used for quantities such as $(n-m)!$. Similar remarks apply to the inequality $n<m$.

In Eq. (A5), only positive quantities are raised to powers that could be noninteger. Thus there are no phase ambiguities. If one is sloppy about phases, the last expression appears to be the same as the first, differing only be a factor of 2. The origin of this factor of two has a simple interpretation in the stationary-phase approximation, described in the next subsection.

It is remarkable that the first expression (for the case $n>m$, $n-m$ even) and the third expression (for the case $n>m$, $n-m$ odd) in Eq. (A2) have the same asymptotics to the order at which we are working. A calculation is involved in showing this. The result that comes from applying the Stirling approximation to the third expression is

$$\sqrt{\frac{2\pi}{n}}\left(\frac{1-m/(n+1)}{1+m/(n+1)}\right)^{m/2}\left[1-\left(\frac{m}{n+1}\right)^2\right]^{-(n+1)/2}.$$

To the accuracy to which we are working, this turns out to be the same as the first expression in Eq. (A5), although some work is required to show this.

2. Stationary-phase approximation

To do a stationary-phase approximation for the function $F(m,n)$, we write the function as

$$F(m,n)=\int_{-\pi/2}^{\pi/2} e^{ng(x)}\,dx, \tag{A6}$$

where the function g is defined by

$$g(z)=\ln\cos z+i\frac{m}{n}z. \tag{A7}$$

With the usual choice of branch cut for the logarithm function, the function $g(z)$ is analytic everywhere in the complex plane except for vertical lines that intersect the real axis at odd multiples of π, and at the intervals on the real axis where $\cos z$ is nonpositive. The identity

$$\cos(x+iy)=\cos x\cosh y-i\sin x\sinh y \tag{A8}$$

is useful in showing this. Knowledge of the region of analyticity of g allows us to deform the contour of integration in Eq. (A6) without changing the value of the integral. We would like to deform the contour so that the phase of the integrand $e^{ng(z)}$ is constant. To do this, we need to know the imaginary part of $g(z)$. With the help of Eq. (A8) we find that this is

$$\Im[g(x+iy)]=-\tan^{-1}(\tan x\tanh y)+\frac{m}{n}x. \tag{A9}$$

If a contour is selected in such a way that this function is a constant, then $g(x+iy)$ will equal an imaginary constant plus a real-valued function along the contour. The integrand $e^{ng(z)}$ will then equal a constant phase factor raised to the nth power times a fixed real-valued function raised to the nth power. This fixed real-valued function may be approximated by a Gaussian, and the integral may then be evaluated. Stationary-phase points z_s satisfy the condition

$$g'(z_s)=0. \tag{A10}$$

This is equivalent to the condition

$$\tan z_s = i\,\frac{m}{n}. \tag{A11}$$

We note that the value of g'' at a stationary-phase point is

$$g''(z_s) = -\sec^2 z_s = \left(\frac{m}{n}\right)^2 - 1. \tag{A12}$$

It is necessary to distinguish two cases, the case $m<n$ and the case $m>n$ (recall that m and n are both positive by assumption). We first consider the case $m<n$. In this case, it follows from a study of Eq. (A9) that a constant phase contour exists that connects the endpoints of the integral and passes through the stationary-phase point

$$z_s = i\,\tanh^{-1}\frac{m}{n} \tag{A13}$$

in a direction that is parallel to the real axis. The integrand is approximated by

$$e^{ng(z_s)+ng''(z_s)(z-z_s)^2/2},$$

and the result for the integral is

$$\sqrt{\frac{2\pi}{-ng''(z_s)}}\,e^{ng(z_s)} = \sqrt{\frac{2\pi}{n}}\left(\frac{1-m/n}{1+m/n}\right)^{m/2}\left[1-\left(\frac{m}{n}\right)^2\right]^{-(n+1)/2}, \tag{A14}$$

which agrees with the result in Eq. (A5).

We now move on to the case $m>n$. In this case, no single contour exists with the properties that it connect the endpoints of the integral and that the quantity in Eq. (A9) be constant. Instead, we choose a contour consisting of three straight-line pieces. The first part, C_1, is defined to start at $-\pi/2$ and go vertically upwards to $-\pi/2+iY$, Y being a large positive real number. The second part, C_2, is defined to go from $-\pi/2+iY$ to $\pi/2+iY$, and the third part, C_3, goes straight down to the $\pi/2$ endpoint of the integral. The parts C_1 and C_3 contain stationary-phase points, which we call z_{s-} and z_{s+}, and which are given by

$$z_{s\pm} = \pm\frac{\pi}{2} + i\,\tanh^{-1}\frac{n}{m}. \tag{A15}$$

For the integral along C_1 the integrand is approximated by

$$e^{ng(z_{s-})+ng''(z_{s-})(z-z_{s-})^2/2}.$$

We parametrize the curve C_1 by $z=z_{s-}+it$, where t is a real parameter. Then dz is $i\,dt$ and the resulting approximation for the integral along C_1 is

$$i\sqrt{\frac{2\pi}{+ng''(z_{s-})}}\,e^{ng(z_{s-})}. \tag{A16}$$

Similarly, the approximation for the integral along C_3 is

$$-i\sqrt{\frac{2\pi}{+ng''(z_{s+})}}\,e^{ng(z_{s+})}. \tag{A17}$$

Because of the condition $m>n$ the integral along the curve C_2 goes to zero as Y goes to infinity. The approximation for $F(m,n)$ is thus the sum of the expressions given in Eqs. (A16) and (A17). If $m-n$ is odd the sum is zero, in agreement with the exact value. If $m-n$ is even, the sum agrees with the approximation of the exact result, given in Eq. (A5).

Thus we see that depending on the ratio m/n, different numbers of stationary-phase points must be considered due to fundamental changes in the form of the stationary-phase contours as m/n goes from one side of the critical value of 1 to the other side. On either side of the critical value of 1, a stationary-phase approximation is possible. The behavior near the critical value is discussed briefly in the main part of this paper.

APPENDIX B: THE CASE OF VANISHING MAGNETIC QUANTUM NUMBERS

Clebsch–Gordan coefficients for the case of vanishing magnetic quantum numbers ($m_1 = m_2 = m = 0$) are of interest in atomic and nuclear physics. Many of the expressions derived in this paper simplify in this case. Also, comparisons with the asymptotics of the exact closed-form expression, given in Eq. (B11), are possible.

The vanishing of the magnetic quantum numbers implies that β, defined in Eq. (3.12), is real. The case $\beta=0$ is simple because one of the j quantum numbers is then equal to the sum of the other two, and the integral in Eq. (2.11) may be evaluated exactly with a small amount of effort. Thus, we will consider the case $\beta>0$. Two other facts that will be used throughout this appendix are that the set of quantum numbers is in the allowed region (since β is real) and that the j quantum numbers are integers (since $j_i - m_i$ is always an integer).

First, we work out the simplifications that occur in the all-real expression in Eq. (3.28). Equation (3.24) becomes

$$\chi = \frac{\pi}{2}\left(j + j_1 + j_2 + \frac{3}{2}\right), \tag{B1}$$

and Eq. (3.28) becomes

$$\langle j_1 0 j_2 0 | j 0\rangle \approx 2\sqrt{\frac{j}{\pi\beta}}\cos\left[\frac{\pi}{2}(j_1 + j_2 - j)\right]. \tag{B2}$$

This agrees with the first-order approximation of the exact expression, which can be obtained from the higher-order approximation [Eq. (B12)] of the exact result, given in Eq. (B11). If $j_1 + j_2 - j$ is odd, then both of the expressions are zero. (In this case, $j_1 + j_2 + j$ is also odd since $2j$ is even.) If $j_1 + j_2 - j$ is even, then both have a sign of $(-1)^{(j_1+j_2-j)/2}$.

Next, we move on to the higher-order approximation. From Eq. (3.11) it is apparent that the two solutions for the contangents of θ and ϕ are related by simply reversing the signs. It follows from Eqs. (3.14), (4.11), and (4.12) that the values of the second- and fourth-order derivatives of $g(\theta,\phi)$ are unchanged, while the third-order derivatives have their signs flipped. Equations (4.7) and (4.10) imply that the quantities δ_4 and δ_6 are unchanged. Finally, Eq. (3.17) implies that the quantity $e^{g(\theta_0,\phi_0)}$ gets multiplied by $(-1)^{j_1+j_2-j}$ for the second root. Because j is an integer, this phase factor is the same as $(-1)^{j_1+j_2+j}$. Since we are in the allowed region, we must sum over both stationary-phase points. We see that for odd values of j_1+j_2+j the result is zero, while for even values of j_1+j_2+j the result is twice the contribution obtained from one of the stationary-phase points.

The values of the second derivatives of $g(\theta,\phi)$ at the stationary-phase points are obtained from Eqs. (3.11) and (3.14), and the results simplify quite a bit:

$$g_{\theta\theta}=-\frac{4j_2(j_1+j)}{j_1+j_2+j}, \tag{B3}$$

$$g_{\theta\phi}=\frac{4j_1j_2}{j_1+j_2+j}, \tag{B4}$$

$$g_{\phi\phi}=-\frac{4j_1(j_2+j)}{j_1+j_2+j}. \tag{B5}$$

From these equations results the following expression for the determinant of the 2×2 Hessian matrix of second derivatives of $g(\theta,\phi)$:

$$\det\frac{\partial^2 g}{\partial(\theta,\phi)^2}=\frac{16j_1j_2j}{j_1+j_2+j}. \tag{B6}$$

We note that this result is nonzero, and it does not vanish in any special cases that have nonzero j quantum numbers, which seems to contradict the statement made at the end of Sec. III about the vanishing of the determinant when β vanishes. The resolution of this apparent contradiction has to do with the fact that in the stationary-phase analysis of this paper we do not simultaneously consider the cases $m_1=m_2=m=0$ and $\beta=0$. These two conditions together would imply α is zero. The quantity α is defined in Eq. (3.25). It vanishes when $j=j_1+j_2$, $j_1=j_2+j$ or $j_2=j+j_1$. In general, as long as α is nonzero, the expression for the determinant can be put (after some work) in the form

$$\det\frac{\partial^2 g}{\partial(\theta,\phi)^2}=\frac{\beta P_1+\beta^2P_2}{\alpha^2(j+j_1+j_2)^2}, \tag{B7}$$

where P_1 and P_2 are (large) polynomials in the quantum numbers. This equation justifies the statement that the determinant is zero in cases where β is zero. On the other hand, in cases where $m_1=m_2=m=0$, P_1 vanishes and α and β are equal, and the result simplifies to that shown in Eq. (B6). The case of $m_1=m_2=m=0$ and $\beta=0$ requires a separate treatment. It is necessary to go back to the original integral representation for the Clebsch–Gordan coefficient. The integral may be approximated by the methods of stationary phase, but it is simpler just to evaluate it or Eq. (2.15) exactly, which is possible at that point.

Higher-order derivatives of $g(\theta,\phi)$ at the stationary-phase points simplify as well. Equations (4.7) and (4.10) for δ_4 and δ_6 yield results that are much simpler than for the general case of nonzero magnetic quantum numbers:

$$\begin{aligned}\delta_4+\delta_6=(&j_1^5j_2-2j_1^3j_2^3+j_1j_2^5+j_1^5j-j_1^3j_2^2j-j_1^2j_2^3j+j_2^5j-j_1^3j_2j^2-10j_1^3j_2^2j^2-j_1j_2^3j^2-2j_1^3j^3\\&-j_1^2j_2j^3-j_1j_2^2j^3-2j_2^3j^3+j_1j^5+j_2j^5)/(12jj_1j_2\beta^2).\end{aligned} \tag{B8}$$

This expression may be rewritten in a more compact form, as explained after Eq. (B12). As discussed above, for odd values of $j+j_1+j_2$ the higher-order approximation of the Clebsch–Gordan coefficient vanishes identically. For even values of $j+j_1+j_2$, the result reduces to

$$\langle j_1 0 j_2 0 | j 0\rangle \approx 2(-1)^{(j_1+j_2-j)/2}\sqrt{\frac{2j+1}{2\pi\beta}}\sqrt{\frac{j+j_1+j_2}{j+j_1+j_2+1}}(1+\delta_4+\delta_6) \times\left[1+\frac{1}{24}\left(\frac{2}{j}+\frac{2}{j_1}+\frac{2}{j_2}-\frac{1}{j+j_1+j_2}-\frac{1}{-j+j_1+j_2}-\frac{1}{j-j_1+j_2}-\frac{1}{j+j_1-j_2}\right)\right], \tag{B9}$$

where we have approximated the factorials in $N_{j_1m_1j_2m_2jm}$ using the form of Stirling's approximation that is appropriate for this order,

$$x! \approx \sqrt{2\pi x}\, x^x e^{-x}\left(1+\frac{1}{12x}\right). \tag{B10}$$

The exact value of the Clebsch–Gordan coefficient is (Ref. 7, p. 87, and Ref. 11)

$$\langle j_1 0 j_2 0 | j 0\rangle = \begin{cases} 0 & j+j_1+j_2 \text{ odd}, \\ (-1)^{(j_1+j_2-j)/2}\sqrt{2j+1}\dfrac{\sqrt{(-j_1+j_2+j)!(j_1-j_2+j)!(j_1+j_2-j)!/(j_1+j_2+j+1)!}}{[((-j_1+j_2+j)/2)!((j_1-j_2+j)/2)!((j_1+j_2-j)/2)!]/((j_1+j_2+j)/2)!}, & j+j_1+j_2 \text{ even}, \end{cases} \tag{B11}$$

and this may be approximated using Eq. (B10). The result is

$$\langle j_1 0 j_2 0 | j 0\rangle \approx \begin{cases} 0, & j+j_1+j_2 \text{ odd} \\ 2(-1)^{(j_1+j_2-j)/2}\sqrt{\dfrac{2j+1}{2\pi\beta}}\sqrt{\dfrac{j+j_1+j_2}{j+j_1+j_2+1}}\left(1-\dfrac{jj_1j_2}{\beta^2}\right), & j+j_1+j_2 \text{ even} \end{cases}. \tag{B12}$$

To the order that we are working, the higher-order stationary-phase result, given in Eq. (B9), and the corresponding approximation of the exact result, given in Eq. (B12), agree. Equation (B9) contains two factors that have the form of unity plus a small correction. If these are multiplied out and only the first-order terms are kept, the result is the factor of $(1-jj_1j_2/\beta^2)$ in Eq. (B12). This shows that Eqs. (B9) and (B12) are equivalent, and it also provides an alternative way of writing the expression in Eq. (B8).

[1] E. P. Wigner, *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959).
[2] A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U.P. Princeton, NJ, 1968).
[3] P. J. Brussaard and H. A. Tolhoek, Physica **23**, 955 (1957).
[4] G. Ponzano and T. Regge, in *Spectroscopic and Group Theoretical Methods in Physics* (North-Holland, Amsterdam, 1968).
[5] W. H. Miller, in *Advances in Chemical Physics, Vol. 25*, edited by I. Prigonine and S. A. Rice (Wiley, New York, 1974).
[6] K. Srinivasa Rao and V. Rajeswari, *Quantum Theory of Angular Momentum, Selected Topics* (Springer/Narosa, Berlin, 1993).
[7] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics: Theory and Application* (Addison–Wesley, Reading, MA, 1981).
[8] O. Schnetz, ''Generating Functions for Multi-*j*-Symbols'' LANL preprint math-ph/9805027 (1998).
[9] J.-J. Labarthe, J. Phys. A **8**, 1543 (1975).
[10] T. Regge, Nuovo Cimento **10**, 544 (1958).
[11] G. Racah, Phys. Rev. **62**, 438 (1942); **63**, 367 (1943).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Exact solution of the Herrera equation of motion in classical electrodynamics

G. Ares de Parga and R. Mares
Departamento de Física, Escuela Superior de Física y Matemáticas, I.P.N., Edif.9 U.P. ''Adolfo López Mateos,'' C.P. 07738, México, D.F.

(Received 29 March 1999; accepted 22 June 1999)

The Landau–Lifshitz equation is derived as a first-order iteration of the Lorentz–Dirac equation for the charged particle. In those cases with null electromagnetic field's gradient, the Landau-Lifshitz gives the so named Herrera equation. A general method for the solution of the latter equation is presented and applied to the motion of a particle in a uniform electromagnetic field. © *1999 American Institute of Physics.* [S0022-2488(99)01610-2]

I. INTRODUCTION

Classical electrodynamics has two main parts; the theory of Maxwell's equations and the dynamics of the charged particles. The first one is universally accepted, but the same is not true for the dynamics of the particle when one considers the effect of radiation. A charged accelerated particle in an applied field radiates energy. In return, the radiation affects the motion of the particle. Accordingly, the mathematical theory of classical electrodynamics should comprise two basic sets of equations; one, which describes the resulting fields, is represented by the Maxwell's equation; and the other, which includes the effect of both the applied field and the radiative reaction, specifies the motion of the particle. General equations of the motion are obtained from very fundamental principles; relativistic covariance and conservation of the four-momentum and we can add another two, simplicity and mass renormalization. In addition, the equation should also reduce to the Lorentz force equation when radiation can be neglected. A partial differential equation which satisfies all four conditions was derived some time ago by Dirac.[1] But as it is well known,[2] the so named Lorentz–Dirac equation is one of the most controversial equations in the history of physics. Indeed, the appearance of the third time derivative brings the equation outside of the dynamical equation which uniquely specifies the trayectory of a particle once the initial conditions of position and velocity are given. A natural solution of the Lorentz–Dirac equation leads to runaway acceleration which can be eliminated by imposing asymptotic conditions, but then the solutions give preacceleration violating the physical causality.[2]

In the following, we consider the approach of Landau and Lifshitz.[3] They start from the Lorentz–Dirac equation and by a first order iteration of the same, they arrive to an equation considered as exact. Indeed Spohn[4] has recently demostrated that the Landau–Lifshitz equation is the effective second order equation that restricts the solution of the Lorentz–Dirac equation to the critical surface on which all solutions are guaranteed not to run to infinity. The Landau procedure reduces to the Herrera equation[5,6] when the gradient of the field is null. Though quadratic in the fields, the differential equation involves only first derivatives in the velocity, and so, does not have the undesirable properties of the Lorentz–Dirac equation. In the paper, we have utilized the Ansatz of Shen[7] for solving the Lorentz–Dirac equation in his approximation. So doing one separates the contribution of the reaction force.

The paper is organized as follows: In Sec. II, we present the basic steps to the Landau–Lifshitz equation and as a special case of the last equation, the Herrera equation. In Sec. III, is presented our Ansätz to the solving of the Herrera equation. In Sec. IV we apply the method for two cases. In Sec. V, we summarize the main ideas and results.

II. EQUATION OF MOTION

The Lorentz–Dirac equation for a charged particle of rest mass m and charge e in a field given by the antisymmetric Faraday's tensor $F_{\mu\nu}$, is

$$\dot{v}^{\mu}(\tau)=\frac{e}{mc}F^{\mu}{}_{\nu}v^{\nu}(\tau)+\tau_0\ddot{v}^{\mu}(\tau)-\frac{\tau_0}{c^2}v^{\mu}(\tau)\dot{v}_{\nu}\dot{v}^{\nu}. \tag{1}$$

In the equation, we have introduced the time-parameter $\tau_0=\frac{2}{3}(e^2/mc^3)$, and the units are Gaussian. The dot-point denotes differentiation with respect to the particle proper time τ. The four-velocity has the components

$$v^{\mu}(\tau)=\gamma(c,v^k(\tau)=\dot{x}^k(\tau)), \tag{2}$$

where the Latin subscript k takes on values 1–3, the Greek subscripts assume values 0–3.

In the iteration of Eq. (1) one assumes that the coefficients of the second and third terms on the right-hand side of Eq. (1) are small and that the applied field does not depend explicitly on the time τ, i.e., $F_{\mu\nu}=F_{\mu}(x(\tau))$. So, for the first order in τ_0, we have,

$$\dot{v}^{\mu}(\tau)=aF^{\mu}_{\nu}v^{\nu}+bK^{\mu}_{\nu\sigma}v^{\nu}v^{\sigma}+\varepsilon F^{\mu}_{\nu}F^{\nu}_{\sigma}\dot{v}^{\sigma}-\frac{\varepsilon v^{\mu}(\tau)}{c^2}F^{\lambda}_{\nu}F^{\nu}_{\sigma}v_{\lambda}v^{\sigma}, \tag{3}$$

in which $a=(e/mc)$, $b=(\tau_0 e/mc)$, $K^{\mu}_{\nu\sigma}=(\partial F^{\mu}_{\nu}/\partial x^{\sigma})$, and $\varepsilon=(\tau_0 e^2/m^2c^2)$.

Equation (3) is the Landau–Lifshitz equation and in the case when the gradient $K^{\mu}_{\nu\sigma}=0$, one obtains the Herrera equation,

$$\dot{v}^{\mu}(\tau)=aF^{\mu}_{\nu}v^{\nu}(\tau)+\varepsilon F^{\mu}_{\nu}F^{\nu}_{\sigma}v^{\sigma}(\tau)-\frac{\varepsilon v^{\mu}(\tau)}{c^2}F^{\lambda}_{\nu}F^{\nu}_{\sigma}v_{\lambda}v^{\sigma}. \tag{4}$$

Herrera notes that Eq. (4) is of the Newtonian class in which the right-hand side is the total applied four-force made up of the Lorentz force linear in the applied field and two other terms quadratic in the field. Furthermore, the equation has none of the problems of the Lorentz–Dirac equation, there are no runaway solutions and no preaccelerations. It has to be pointed out that Herrera claimed unsubstaintiated that his equation is exact because it does not possess the dificulties of the Lorentz–Dirac one. Nevertheless as we mentioned above, the equation which is substantially supported is the Landau–Lifshitz equation as Spohn[4] showed.

It is interesting to compare the Herrera equation with equation of motion of M_0-Papas,

$$\dot{v}^{\mu}(\tau)=\frac{e}{mc}F^{\mu}_{\nu}(v^{\nu}(\tau)+\tau_0\dot{v}^{\mu}(\tau))+\frac{\tau_0}{m}F^{\lambda}_{\sigma}v^{\sigma}_{\lambda}\dot{v}_{\lambda}v^{\mu}.$$

The Herrera equation, derived as a first-order iteration of the Lorentz–Dirac equation has the added force term which is quadratic in the applied field and the term is equivalent to a Poynting-type momentum being transfered to the particle. The Mo–Papas equation assumes new intuitive ideas: radiation reaction should be expressible by the external fields and the charge's kinematics, a charge experiences, in addition to the Lorentz force, another external force proportional to its acceleration and finally, inertia plus radiation is balanced by these two external forces. Nevertheless since this last equation possesses a third derivative with respect to the time, it will present similar difficulties as the Lorentz–Dirac equation and consequently our method will not apply for this case.

Even if our main point it is to solve the Landau–Lifshitz equation, we will consider as a first step, just the case when $K_{\nu\tau}=0$. That is the Herrera equation case.

Using the fundamental identity,

$$F_{\mu\nu}=(\delta_{\mu o}\delta_{\nu j}-\delta_{\mu j}\delta_{\nu o})E_j-\delta_{\mu j}\delta_{\nu k}\varepsilon_{jkl}B_l\,. \tag{5}$$

Herrera has succeeded in to expressing Eq. (3) in the form, explicitly in the fields,

$$\dot{\mathbf{v}}(\tau)=\frac{e}{m}\left(\gamma\mathbf{E}+\frac{\mathbf{v}\times\mathbf{B}}{c}\right)-\varepsilon\mathbf{v}\left[\gamma^2(\mathbf{E}^2+\mathbf{B}^2)-\left(\frac{\mathbf{v}\cdot\mathbf{E}}{c}\right)^2-\left(\frac{\mathbf{v}\cdot\mathbf{B}}{c}\right)^2-2\gamma\mathbf{v}\cdot\left(\frac{\mathbf{E}\times\mathbf{B}}{c}\right)\right]$$
$$+\varepsilon(\mathbf{E}(\mathbf{v}\cdot\mathbf{E})+\mathbf{B}(\mathbf{v}\cdot\mathbf{B})+c\gamma\mathbf{E}\times\mathbf{B}), \tag{6a}$$

and

$$\dot{\gamma}(t)=\frac{e}{mc^2}\mathbf{v}\cdot\mathbf{E}-\varepsilon\gamma\left[(\gamma^2-1)(\mathbf{E}^2+\mathbf{B}^2)-\left(\frac{\mathbf{v}\cdot\mathbf{E}}{c}\right)^2-\left(\frac{\mathbf{v}\cdot\mathbf{B}}{c}\right)^2\right]+\varepsilon(2\gamma^2-1)\mathbf{v}\cdot(\mathbf{E}\times\mathbf{B}). \tag{6b}$$

Equations (6a) and (6b) are the point of depart for the two applications of Herrera[4] to the motion of the charged particle in a uniform magnetic field and uniform electric field.

III. EXACT SOLUTION OF THE HERRERA EQUATION OF MOTION

It can be easily verified that if the four-vector $f^\mu(\tau)$ is a solution of equation

$$\dot{f}^\mu(\tau)=T^\mu_\lambda f^\lambda(\tau), \tag{7}$$

where

$$T^\mu_\lambda=aF^\mu_\lambda+\varepsilon F^\mu_\nu F^\nu_\lambda, \tag{8}$$

then

$$v^\mu(\tau)=\eta(\tau)f^\mu(\tau), \tag{9}$$

is a solution of the Herrera equation in the form of Eq. (4), with

$$\eta(\tau)=\left[1+\frac{2\varepsilon}{c^2}\int_0^\tau F_{\nu\varepsilon}F^\nu_\lambda f^\varepsilon f^\lambda d\tau\right]^{-1/2}. \tag{10}$$

$\eta(\tau)$ represents the damping effect due to the radiation reaction. Therefore, once the solution to Eq. (7) is found at least a formal expression can be immediately written down for the solution of the Herrera equation.

We can solve Eq. (7) formally, by means of an exponential expression, if the applied field $F_{\mu\nu}$ is constant in space and time,

$$f^\mu(\tau)=\exp[(aF^\mu_\nu+\varepsilon F^\lambda_\nu F^\mu_\lambda)\tau]f^\nu(0). \tag{11}$$

Also, the method can be applied to obtain solutions to the equations of motion in cases of physical interest if the fields $F_{\mu\nu}$ are just dependent on the time since we can always solve Eq. (7).

Introducing Eqs. (11) and (10) into Eq. (9), we find a formal solution for the Herrera equation,

$$v^\mu(\tau)=\frac{\exp[(aF^\mu_\nu+\varepsilon F^\mu_\lambda F^\lambda_\nu)\tau]f^\nu(0)}{\left[1+\frac{2\varepsilon}{c^2}\int_0^\tau F_{\nu\varepsilon}F^\nu_\lambda f^\varepsilon f^\lambda d\tau\right]^{1/2}} \tag{12}$$

for the case of uniform electromagnetic fields.

For further developement of Eq. (12) we can apply to it the methods of Kumar[8] on the expansion of a function of noncommunting operators and on expanding the exponential. Nevertheless, we prefer to solve Eq. (7) more directly and simultaneously to give a physical insight into the force-term $\varepsilon F^{\mu}_{\nu} F^{\mu}_{\sigma} v^{\sigma}$. To this end, we made the Ansatz,

$$f^{\mu}(\tau)=e^{\alpha(\tau)} f^{\mu}_{L}(\tau), \tag{13}$$

where $f^{\mu}_{L}(\tau)$ is any solution of the Lorentz equation,[7,8] without radiation. One can readly show that $\alpha(\tau)$ is of the form,

$$\alpha(\tau)=\frac{\varepsilon}{c^2}\int_0^{\tau} F^{\mu}_{\lambda} F^{\lambda}_{\nu} f^{\nu}_{L}(\tau') f_{\mu L}(\tau') d\tau' \tag{14}$$

with the normalization condition, $f^{\mu}_{L}(\tau') f_{\mu L}(\tau')=c^2$.

Introducing Eqs. (14), (13), and (10) in Eq. (9), the solution of the Herrera equation reads

$$v^{\mu}(\tau)=\frac{\exp\left[\frac{\varepsilon}{c^2}\int_0^{\tau} F^{\mu}_{\lambda} F^{\lambda}_{\nu} f^{\nu}_{L}(\tau') f_{\mu L}(\tau') d\tau'\right]}{\left[1+\frac{2\varepsilon}{c^2}\int_0^{\tau} F_{\nu\varepsilon} F^{\nu}_{\lambda} f^{\varepsilon} f^{\lambda} d\tau'\right]^{-1/2}} f^{\mu}_{L}(\tau). \tag{15}$$

Let us consider the case of a uniform electromagnetic field. Using the fact that $F_{\mu\nu}$ is antisymmetric and the tensor $\dot{f}^{\mu}_{L} \dot{f}_{\nu L}$ is symmetric and the product of a symmetric tensor with an antisymmetric tensor gives identically zero, we find

$$\frac{d}{d\tau}(F^{\mu}_{\lambda} F^{\lambda}_{\nu} f^{\nu}_{L} f_{\mu L})=2F_{\mu\lambda} f^{\lambda}_{L} F^{\mu\nu} \dot{f}_{\nu L}=2F^{\mu\nu} \dot{f}_{\mu L} \dot{f}_{\nu L}=0, \tag{16}$$

i.e., $F^{\mu}_{\lambda} F^{\lambda}_{\nu} f^{\nu}_{L} f_{\mu L}$ is a constant of time.

Therefore, in a constant field

$$\alpha(\tau)=\frac{\varepsilon}{c^2} F^{\mu}_{\lambda} F^{\lambda}_{\nu} f^{\nu}_{L} f_{\mu L} \tau. \tag{17}$$

To compute $v^{\mu}(\tau)$, one needs to find $f^{\mu}_{L}(\tau)$ first.

For a charged particle in a constant electric and magnetic field this has become a standard textbook exercise. Piña and recently Muñoz and Hyman[9–11] present a fully covariant solution to the problem of a charged particle in a spatially uniform electromagnetic field $F_{\alpha\beta}$. The integration method naturally leads to a solution in manifestly covariant form. Shen,[7] applies a more elegant method for obtaining formal solutions to the Lorentz equation using the projection operators, first given by Rosen[12] for finite transformation in SU(3) space. Nevertheless as we will see in the next section, our method has the advantage of simplicity.

IV. APPLICATIONS

A. Motion in a uniform magnetic field

The relativistic motion of a particle in a field specified by $B_k=(0,0,B_0)$ it is a good testing ground for applying an equation of motion. Writing $\omega_0=(eB_0/mc)$ and $B_0^2=\lambda\omega_0$, where $\lambda=(m^2c^2/e^2)\,\omega_0$ we find that Eqs. (10), (14) are then expressed as

$$\alpha(\tau)=-\varepsilon\lambda\,\omega_0\tau, \tag{18}$$

and

$$\eta(\tau)=(\gamma^2(0)-(\gamma^2(0)-1)e^{-2\lambda w_0\tau})^{-1/2}, \tag{19}$$

where we have chosen $v_3(0)=0$.

For a particle in a uniform magnetic field, Eq. (15) now yields the solution in a closed form

$$v^\mu(\tau)=\frac{e^{-\lambda w_0\tau}}{(\gamma^2(0)-(\gamma^2(0)-1)e^{-2\lambda w_0\tau})^{1/2}}f_L^\mu(\tau), \tag{20}$$

with $f_L^\mu(\tau)$ any solution of the Lorentz equation. Equation (20) coincides with the results obtained by a parametrization of the Larmor formula and averaging it.[13]

B. Motion in a uniform electric field

Another example is the motion in a uniform electric field given by $E_k=(E_0,0,0)$; the general solution is also in this case, obtainable in closed form.

In writing the relations, we have introduced, following Herrera, the abbreviation $\Im(t)=(v_2^2+v_3^2)/c^2$ corresponding to the square of the transverse four-velocity. The exact solution with the proper time τ as the independent variable is

$$v^\mu(\tau)=\frac{e^{2\varepsilon E_0^2\tau}f_L^\mu(\tau)}{[\gamma^2(o)\Im(0)-(1+\gamma^2(o)\Im(0))e^{2\varepsilon E_0^2\tau}]^{1/2}}, \tag{21}$$

where the Lorentz solutions are given by the expressions

$$\begin{aligned} f_L^0(\tau)&=f_L'(0)\sin h(\kappa\tau)+\gamma_L(0)\cos h(\kappa\tau)\\ f_L'(\tau)&=\gamma_L(0)\sin h(\kappa\tau)+f_L'(0)\cos h(\kappa\tau), \end{aligned} \tag{22}$$

where $\kappa=e\tau_0/mc$ and $f_L^\mu=\gamma(c,\mathbf{v})$.

This is the same result that Herrera[6] obtained.

V. CONCLUSION

The Landau–Lifshitz equation, derived as a first-order iteration of the Lorentz–Dirac equation, is considered as the exact equation of motion for classical electrodynamics of a charged particle. The equation is of the Newtonian class since the total applied force is made up of the Lorentz force and two other terms quadratic in external fields. Therefore, there are no runaway solutions and no preacceleration. We have solved exactly the equation in two cases of physical importance: uniform magnetic and electric fields, and the results are quite similar to those obtained by Herrera in both situations. The method can be extended to the domain of oscillatory field. In those circumstances the Herrera equation is not valid and the Landau–Lifshitz equation has to be used. Similar methods can be used to solve it in a future paper. But the fundamental question of whether the Landau–Lifshitz equation of motion or for that matter, the Lorentz–Dirac equation and other proposed equations are in agreement with the experiment, is an open question. Finally, in accordance with Spohn[4] and the limits of classical electrodynamics,[13] that is where quantum effects are not important, the equation of Landau–Lifshitz is a good candidate to be tested experimentally.[13]

ACKNOWLEDGMENTS

The authors acknowledge the referee's comments and partial financial support from COFAA-EDD, IPN.

[1] P.A.M. Dirac, Proc. R. Soc. London, Ser. A **167**, 148 (1938).
[2] F. Rohrlich, *Classical Charged Particles* (Addison–Wesley, Reading, 1965).
[3] L. D. Landau and E. M. Lifshhitz, *The Classical Theory of Fields*, 4th ed. (Pergamon, Oxford, 1975), Chap. IX, Sec. 76.
[4] H. Spohn, ''The critical manifold of the Lorentz-Dirac equation,'' preprint, May, 1999.
[5] J.C. Herrera, Phys. Rev. D **7**, 1567 (1973); **10**, 4297 (1974).
[6] J.C. Herrera, Nuovo Cimento **7013**, 12 (1970).
[7] C.S. Shen, Phys. Rev. D **17**, 434 (1978).
[8] K. Kumar, J. Math. Phys. **12**, 1928 (1965).
[9] E. Piña, Rev. Mex. Fis. **16**, 233 (1967).
[10] A.T. Hyman, Am. J. Phys. **65**, 195 (1977).
[11] G. Muñoz, Am. J. Phys. **65**, 429 (1977).
[12] P. Rosen, J. Math. Phys. **12**, 673 (1971).
[13] G. Ares de Parga and R. Mares, ''Drift of the center of motion of a charged particle due to the damping effect,'' Nuovo Cimento (submitted).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Diffusive energy scattering from weakly random surfaces

Guillaume Bal[a] and George Papanicolaou[b]
Department of Mathematics, Stanford University, Stanford, California 94305

Leonid Ryzhik[c]
Department of Mathematics, University of Chicago, Chicago, Illinois 60637

(Received 17 March 1999; accepted for publication 19 June 1999)

We derive transport theoretic boundary conditions for acoustic wave reflection at a weakly rough boundary in an inhomogeneous half space. We use the Wigner distribution to go from waves to energy transport in the high frequency limit. We generalize known results on the reflection of acoustic plane waves in a homogeneous medium. We analyze higher order corrections, which include a enhanced backscattering effect in the back direction. © *1999 American Institute of Physics.* [S0022-2488(99)02009-5]

I. INTRODUCTION

Wave propagation in weakly fluctuating random media over distances large compared to the wavelength can be described by incoherent energy transport. This is the radiative transport regime.[1,2] Near boundaries and interfaces, waves undergo coherent or partially coherent reflection. Angularly resolved energy reflection and transmission in homogeneous media in average, has been studied extensively in the past.[3–6] Recently, the problem has been revisited in the transport theoretic context using a plane wave decomposition; see Ref. 7. There, the boundary conditions for the transport equations in a domain with rough boundaries of small amplitude and homogeneous background are derived. In this paper, we derive boundary conditions in the case of inhomogeneous domains. We systematically use the Wigner transform to study the reflection of angularly resolved acoustic energy density from a Dirichlet surface. The scale of the volume inhomogeneities is large compared to the wave length. The boundary conditions are a direct generalization of those obtained in the case of a homogeneous medium, with a reflection operator depending upon the position at the boundary. Our main ingredient is a perturbation analysis around the flat boundary case studied in Ref. 8.

A. Transport equations for the energy density

As we recall in Sec. II, the phase space acoustic energy density $\mu(\mathbf{x},\mathbf{k})$ satisfies the transport equation

$$\nabla_{\mathbf{k}}\omega_+ \cdot \nabla_{\mathbf{x}}\mu - \nabla_{\mathbf{x}}\omega_+ \cdot \nabla_{\mathbf{k}}\mu = 0. \tag{1}$$

Here $\omega_+(\mathbf{x},\mathbf{k})$ is the eigenfrequency of the acoustic waves given by

$$\omega_+ = c|\mathbf{k}|, \quad c = \frac{1}{\sqrt{\kappa\rho}}, \tag{2}$$

where $\rho(\mathbf{x})$ is the density and $\kappa(\mathbf{x})$ is the compressibility of the background medium. These equations hold in the high frequency regime for monochromatic waves, when the wavelength is

[a]bal@math.stanford.edu
[b]papanico@math.stanford.edu
[c]ryzhik@math.uchicago.edu

much smaller than the variations of the density and compressibility. They were derived in domains without boundaries or interfaces by several authors in the context of geometrical optics (see for instance Ref. 9).

B. Scattering from a rough boundary in a homogeneous medium

In the preceding section, the free transport equation (1) is posed in the whole space, with no boundary. In realistic applications, it is important to consider problems in bounded domains with correct boundary conditions.

We assume that the domain is given by $H_\epsilon = \{\mathbf{x} \in R^n : x_n > \epsilon \eta h(\mathbf{x}'/\epsilon)\}$. The function $h(\mathbf{x}')$ is a mean zero stationary random process with covariance function $R(\mathbf{y}')$ defined by

$$\langle h(\mathbf{x}'+\mathbf{y}')h(\mathbf{x}')\rangle = R(\mathbf{y}'). \tag{3}$$

The power spectrum $\hat{R}(\mathbf{k}')$ is the Fourier transform of $R(\mathbf{y}')$:

$$R(\mathbf{y}') = \int \frac{d\mathbf{p}}{(2\pi)^{d-1}} e^{i\mathbf{p}'\cdot\mathbf{y}'} \hat{R}(\mathbf{p}'). \tag{4}$$

The boundary of our domain is varying on the scale of the wavelength ϵ, which gives rise to scattering of incoherent, or diffuse, energy from the boundary. The small parameter $\eta \ll 1$ is measuring the height of the surface relative to the wavelength.

One finds by a direct computation using a plane wave decomposition that the average energy of reflected waves going in the direction $\mathbf{k}^+(\mathbf{k}')$ is given, for Dirichlet boundary conditions, by

$$\mu_{\text{out}}(\mathbf{x}',\mathbf{k}') = \mu_{\text{out}}^{\text{spec}} + \mu_{\text{out}}^{\text{diff}} = \left(1 - \eta^2 \int \frac{d\mathbf{p}'}{(2\pi)^{n-1}} \hat{R}(\mathbf{k}'-\mathbf{p}') p_n^+ k_n^+ \right) \mu_{\text{in}}(\mathbf{x}',\mathbf{k}')$$
$$+ \eta^2 \int \frac{d\mathbf{p}'}{(2\pi)^{n-1}} \hat{R}(\mathbf{k}'-\mathbf{p}') (p_n^+)^2 \mu_{\text{in}}(\mathbf{x}',\mathbf{p}'). \tag{5}$$

Here we have defined for every horizontal wave vector $\mathbf{k}'$:

$$\mathbf{k}^\pm(\mathbf{k}') = (\mathbf{k}', k_n^\pm), \quad |\mathbf{k}'|^2 + (k_n^\pm)^2 = K^2 = \frac{\omega_+^2}{c^2}. \tag{6}$$

Notice that $\mathbf{k}^+$ corresponds to outgoing waves, and $\mathbf{k}^-$ corresponds to incoming waves.

A detailed computation both for the Dirichlet and Neumann problems can be found, for instance, in Refs. 3–6. The first term in (5) represents the specular reflection including a correction due to surface roughness. The second term is produced by the diffuse scattering from the rough boundary.

The plane wave decomposition used in this calculation is not available in non homogeneous media. Thus one cannot directly generalize the above calculations to variable media. We have recently derived in Ref. 8 transport boundary conditions for the energy in inhomogeneous media with boundaries which vary on a large scale compared to the wavelength. We treat here the Dirichlet problem in an inhomogeneous domain with rough boundary for $\eta \ll 1$ as a perturbation of the smooth boundary considered in Ref. 8. We show that the results of Ref. 8 allow us to compute the higher corrections in η of the reflected energy, which coincide with (5) in the case of a homogeneous medium.

It is known that the Born expansion we use in this paper diverges for the Neumann and impedance problems at grazing angles and we do not consider them here. However, our method can be adapted to incorporate the smoothing method (Refs. 7, 10, and 11). Then, we can reproduce the results obtained in Refs. 12 and 13 for uniform media including the coherent backscattering effect, now in the setup of a variable media.

The paper is organized as follows. In Sec. II, we review the results of Ref. 8 and some basic facts from the Wigner distribution theory. We state our results for rough boundaries in Sec. III. The perturbation expansion is treated in Sec. IV up to second order in η. Finally, we investigate higher order corrections in Sec. V.

II. ENERGY PROPAGATION IN A HALF SPACE

A. The Wigner distribution

One way to describe scattering of phase space resolved energy is to consider the Wigner distribution matrix of the family $\mathbf{w}_\epsilon$ of the solutions of (12). A detailed exposition to the theory of Wigner distributions can be found in Refs. 14 and 8. Given a family of vector-valued functions $\mathbf{u}_\epsilon(\mathbf{x})$ bounded in $L^2(R^n)$, its Wigner transform matrix is defined by

$$\mathbf{W}_\epsilon(\mathbf{x},\mathbf{k}) = \int \frac{d\mathbf{y}}{(2\pi)^n} e^{i\mathbf{k}\cdot\mathbf{y}} \mathbf{u}_\epsilon\left(\mathbf{x} - \frac{\epsilon\mathbf{y}}{2}\right) \mathbf{u}_\epsilon^*\left(\mathbf{x} + \frac{\epsilon\mathbf{y}}{2}\right). \tag{7}$$

The family $\mathbf{W}_\epsilon$, possibly after extracting a subsequence, has a weak limit as $\epsilon \to 0$ in the sense of Schwartz distributions. The limit matrix $\mathbf{W}(\mathbf{x},\mathbf{k})$, called the Wigner distribution, has a number of important properties, which we summarize in the following proposition.

Proposition 1: The matrix $\mathbf{W}(\mathbf{x},\mathbf{k})$ *is self-adjoint and non-negative. Given a bounded continuous function* $\theta(\mathbf{x})$, *the Wigner distribution of the family* $g_\epsilon(\mathbf{x}) = \theta(\mathbf{x}) f_\epsilon(\mathbf{x})$ *is*

$$\mathbf{W}[g_\epsilon](\mathbf{x},\mathbf{k}) = \theta(\mathbf{x}) \mathbf{W}[f_\epsilon](\mathbf{x},\mathbf{k}) \theta^*(\mathbf{x}). \tag{8}$$

Given a differential operator $L(\mathbf{x},D)$ *with smooth coefficients, the Wigner matrix of the family* $p_\epsilon(\mathbf{x}) = L(\mathbf{x},\epsilon D) f_\epsilon(\mathbf{x})$ *is*

$$\mathbf{W}[p_\epsilon](\mathbf{x},\mathbf{k}) = L(\mathbf{x},i\mathbf{k}) \mathbf{W}[f_\epsilon](\mathbf{x},\mathbf{k}) [L(\mathbf{x},i\mathbf{k})]^*. \tag{9}$$

If the family f_ϵ *is* ϵ*-oscillatory,*[14] *then*

$$\mathrm{Tr} \int \mathbf{W}(\mathbf{x},d\mathbf{k}) = \lim_{\epsilon\to 0} |\mathbf{u}_\epsilon(\mathbf{x})|^2. \tag{10}$$

The property (8) allows one to consider the Wigner distributions of families bounded in $L^2_{\mathrm{loc}}(R^n)$. This is important in dealing with time-harmonic solutions of the acoustic equations. The property (9) allows one to deal with high frequency waves in variable media. The last property shows that the Wigner matrix captures the energy of high frequency waves, which oscillate at a frequency of order ϵ^{-1} at most.

One can also consider the Wigner matrix of two different families $\mathbf{u}_\epsilon, \mathbf{v}_\epsilon$, defined as the limit of

$$\mathbf{W}[\mathbf{u}_\epsilon,\mathbf{v}_\epsilon](\mathbf{x},\mathbf{k}) = \int \frac{d\mathbf{y}}{(2\pi)^n} e^{i\mathbf{k}\cdot\mathbf{y}} \mathbf{u}_\epsilon\left(\mathbf{x} - \frac{\epsilon\mathbf{y}}{2}\right) \mathbf{v}_\epsilon^*\left(\mathbf{x} + \frac{\epsilon\mathbf{y}}{2}\right).$$

It has similar properties.[14]

B. Acoustic wave transport

The time harmonic acoustic equations for the acoustic velocity $\mathbf{v} = (v_1,\ldots,v_n)$, $n = 2,3$, and pressure p, in the high frequency regime, are

$$\begin{aligned} \epsilon\nabla p_\epsilon &= i\omega\rho(\mathbf{x})\mathbf{v}_\epsilon, \\ \epsilon\nabla\cdot\mathbf{v}_\epsilon &= i\omega\kappa(\mathbf{x})p_\epsilon. \end{aligned} \tag{11}$$

Here ρ and κ are the density and compressibility of the medium, and the parameter $\epsilon \ll 1$ is the ratio of the wave length to the typical scale of the variations of $\rho(\mathbf{x})$ and $\kappa(\mathbf{x})$. Equations (11) can be rewritten as a reduced symmetric hyperbolic system

$$\sum_{j=1}^{n} \epsilon D^j \frac{\partial \mathbf{w}_\epsilon}{\partial x_j} - i\omega A(\mathbf{x})\mathbf{w}_\epsilon = 0 \tag{12}$$

for the vector $\mathbf{w}_\epsilon = (\mathbf{v}_\epsilon, p_\epsilon) \in C^{n+1}$. Here the diagonal matrix $A(\mathbf{x}) = \mathrm{diag}(\rho,\rho,\rho,\kappa)$, for $n=3$, and the symmetric matrices D^j are defined appropriately from (11).

Let $\mathbf{w}_\epsilon$ be an ϵ-oscillatory family of solutions of (12) bounded in $L^2_{\mathrm{loc}}(R^n)$. Then the limit Wigner matrix $\mathbf{W}(\mathbf{x},\mathbf{k})$ is described as follows. The dispersion matrix of the system (12) is (in three dimensions)

$$L = A^{-1}\sum_{j=1}^{n} k_j D^j = \begin{pmatrix} 0 & 0 & 0 & k_1/\rho \\ 0 & 0 & 0 & k_2/\rho \\ 0 & 0 & 0 & k_3/\rho \\ k_1/\kappa & k_2/\kappa & k_3/\kappa & 0 \end{pmatrix}. \tag{13}$$

The eigenvector $\mathbf{b}$ of the matrix L that corresponds to forward propagating waves is

$$\mathbf{b} = \left(\frac{\hat{\mathbf{k}}}{\sqrt{2\rho(\mathbf{x})}}, \frac{1}{\sqrt{2\kappa(\mathbf{x})}} \right), \tag{14}$$

where $\hat{\mathbf{k}} = \mathbf{k}/|\mathbf{k}|$. The corresponding eigenfrequency is given by

$$\omega_+ = c(\mathbf{x})|\mathbf{k}|, \quad c(\mathbf{x}) = \frac{1}{\sqrt{\kappa(\mathbf{x})\rho(\mathbf{x})}}. \tag{15}$$

The solutions of (12) have frequency ω, thus we introduce the resonant wave number $K(\mathbf{x}) = \omega\sqrt{\kappa(\mathbf{x})\rho(\mathbf{x})}$. Then, the positive definite matrix $\mathbf{W}(\mathbf{x},\mathbf{k})$ has the form

$$\mathbf{W}(\mathbf{x},\mathbf{k}) = \mu(\mathbf{x},\mathbf{k})\mathbf{b}(\mathbf{x},\mathbf{k})\mathbf{b}^*(\mathbf{x},\mathbf{k}).$$

Here the scalar measure $\mu(\mathbf{x},\mathbf{k})$ is supported on the set

$$\mathcal{S} = \{(\mathbf{x},\mathbf{k}): \ |\mathbf{k}| = K(\mathbf{x})\}. \tag{16}$$

This statement is a generalization of the eikonal equation of geometrical optics. The measure μ satisfies the transport equation (1).[14,2]

The scalar measure $\mu(\mathbf{x},\mathbf{k})$ can be considered as the phase space resolved energy density of acoustic waves in the high frequency limit. Namely, the high frequency limit of the physical space energy density is given by[2]

$$\lim_{\epsilon\to 0} \mathcal{E}_\epsilon(\mathbf{x}) = \int \mu(\mathbf{x}, d\mathbf{k}).$$

Here the acoustic energy density is

$$\mathcal{E}_\epsilon(\mathbf{x}) = \frac{\rho|\mathbf{v}_\epsilon|^2}{2} + \frac{\kappa|p_\epsilon|^2}{2}.$$

The limit energy flux $\mathcal{F} = \frac{1}{2}(p_\epsilon \overline{\mathbf{v}_\epsilon} + \overline{p_\epsilon}\mathbf{v}_\epsilon)$ is expressed via μ by

$$\lim_{\epsilon\to 0}\mathcal{F}_\epsilon(\mathbf{x})=c(\mathbf{x})\int \hat{\mathbf{k}}\mu(\mathbf{x},d\mathbf{k}).$$

C. Acoustic energy transport in a half space

We review here briefly the results of Ref. 8. Let $\mathbf{w}_\epsilon$ be an ϵ-oscillatory family of solutions of the acoustic equations (12) in the half space $x_n>0$ (without imposing any boundary conditions yet). We assume that the family $\mathbf{w}_\epsilon(\mathbf{x})$ and the family

$$\mathbf{r}_\epsilon(\mathbf{x}')=\mathbf{w}_\epsilon(\mathbf{x}',0) \tag{17}$$

of its boundary traces are bounded in $L^2_{\text{loc}}(R^n)$ and $L^2_{\text{loc}}(R^{n-1})$, respectively. Here and below we use the notation $\mathbf{x}',\mathbf{k}'$ for the position on the boundary and the tangential component of the wave vector, respectively. In addition we assume that the measure of the resonant set $B_{cr}=\{(\mathbf{x}',\mathbf{k}')\in R^{n-1}\times R^{n-1}:c(\mathbf{x}')|\mathbf{k}'|=\omega\}$ of grazing rays with respect to the Wigner measure $\nu(\mathbf{x}',\mathbf{k}')$ of the family $\mathbf{r}_\epsilon$ is zero. The grazing rays in the phase space energy context were studied recently in Refs. 15 and 16 and we avoid these technical complications. Physically, our assumption means that grazing rays are not charged. Then we have the following proposition.

Proposition 2: Under the above-mentioned assumptions the following holds.

(i) *The Wigner matrix $\boldsymbol{\nu}(\mathbf{x}',\mathbf{k})$ of the family $\mathbf{r}_\epsilon(\mathbf{x}')$ has the form*

$$\boldsymbol{\nu}=\nu_\alpha\mathbf{b}(\mathbf{k}_-)\mathbf{b}^*(\mathbf{k}_-)+\nu_\beta\mathbf{b}(\mathbf{k}_+)\mathbf{b}^*(\mathbf{k}_+)+\nu_{\alpha\beta}\mathbf{b}(\mathbf{k}_-)\mathbf{b}^*(\mathbf{k}_+)+\overline{\nu_{\alpha\beta}}\mathbf{b}(\mathbf{k}_+)\mathbf{b}^*(\mathbf{k}_-) \tag{18}$$

with $\nu_{\alpha,\beta,\alpha\beta}$ being distributions so that the matrix $\begin{pmatrix}\nu_\alpha & \nu_{\alpha\beta}\\ \overline{\nu_{\alpha\beta}} & \nu_\beta\end{pmatrix}$ is positive definite.

(ii) *The Wigner matrix $\mathbf{W}(\mathbf{x},\mathbf{k})$ of the family $\mathbf{w}_\epsilon(\mathbf{x})$ has the form*

$$\mathbf{W}(\mathbf{x},\mathbf{k})=\mu(\mathbf{x},\mathbf{k})\mathbf{b}(\mathbf{x},\mathbf{k})\mathbf{b}^*(\mathbf{x},\mathbf{k}). \tag{19}$$

The scalar measure μ is supported on the set (16) and satisfies weakly the transport equation

$$\nabla_{\mathbf{k}}\omega_+\cdot\nabla_{\mathbf{x}}\mu-\nabla_{\mathbf{x}}\omega_+\cdot\nabla_{\mathbf{k}}\mu=c\hat{k}_n(\mu_{\text{in}}\delta(k_n-k_n^-)+\mu_{\text{out}}\delta(k_n-k_n^+)). \tag{20}$$

Here the measures μ_{in} and μ_{out} are given by

$$\mu_{\text{in}}(\mathbf{x}',\mathbf{k}')=\nu_\alpha(\mathbf{x}',\mathbf{k}'),\quad \mu_{\text{out}}(\mathbf{x}',\mathbf{k}')=\nu_\beta(\mathbf{x}',\mathbf{k}') \tag{21}$$

with ν_α,ν_β defined by (18), $\omega_+(\mathbf{x},\mathbf{k})=c(\mathbf{x})|\mathbf{k}|$ is given by (2), and the wave vector $\mathbf{k}^\pm$ is defined by (6) with $K=K(\mathbf{x}')=\omega\sqrt{\kappa(\mathbf{x}')\rho(\mathbf{x}')}$.

Note also that if the measure μ is continuous up to the boundary $x_n=0$, then the weak form (20) is equivalent to the boundary value problem:

$$\nabla_{\mathbf{k}}\omega_+\cdot\nabla_{\mathbf{x}}\mu-\nabla_{\mathbf{x}}\omega_+\cdot\nabla_{\mathbf{k}}\mu=0, \tag{22}$$

$$\mu(\mathbf{x}',0,\mathbf{k})=\mu_{\text{in}}\delta(k_n-k_n^-)+\mu_{\text{out}}\delta(k_n-k_n^+). \tag{23}$$

Thus we can interpret ν_α as the phase space resolved energy of the incoming waves at the boundary, and ν_β as the energy of the outgoing waves.

D. Boundary conditions for the transport equation

The boundary conditions for the transport equation (22) can be obtained from those for the acoustic equations (12), by using the relations (21) and (23). Consider the Dirichlet boundary condition in the upper half space

$$w_{\epsilon,n+1}(\mathbf{x}',0)=0. \tag{24}$$

Then the $(\mathbf{n}+1)$-row and column of the matrix $\boldsymbol{\nu}$ vanish. Using (18), the explicit form (14) of the eigenvectors $\mathbf{b}(\mathbf{x},\mathbf{k})$, and (6), we get

$$\nu_\alpha = \nu_\beta = -\nu_{\alpha\beta}. \tag{25}$$

Then (23) implies that

$$\mu(\mathbf{x}',0,\mathbf{k}',k_n) = \mu(\mathbf{x}',0,\mathbf{k}',-k_n), \tag{26}$$

which is the boundary condition for (22).

The Neumann boundary condition

$$w_{\epsilon,n}(\mathbf{x}',0) = 0 \tag{27}$$

implies that the nth row and column of the matrix $\boldsymbol{\nu}$ vanish. Then we obtain, similarly to (25), that

$$\nu_\alpha = \nu_\beta = \nu_{\alpha\beta}, \tag{28}$$

so that (26) still holds. A convenient way to rewrite (26) is

$$\mu_{\text{out}}(\mathbf{x}',\mathbf{k}') = \mu_{\text{in}}(\mathbf{x}',\mathbf{k}'),$$

so that all the energy is reflected specularly as expected.

III. ENERGY REFLECTION AT A ROUGH SURFACE

We consider scattering of acoustic waves described by (12) from a rough surface. The surface ∂H_ϵ is described by the equation $x_n = \epsilon\eta h(\mathbf{x}'/\epsilon)$. The small parameter η is the ratio of the height of the surface to the wavelength. The height of the surface is varying on the scale of the wavelength ϵ. Recall that the random process $h(\mathbf{y}')$ has mean zero, and is stationary, with covariance function $R(\mathbf{y}')$ and with power spectrum $\hat{R}(\mathbf{k}')$, given by (3) and (4), respectively. We assume that $h(\mathbf{y}') \in C^1(R^{n-1})$ a.s. so that solutions of the Dirichlet problem exist for every positive ϵ. The Dirichlet boundary condition for (12), corresponding to a vanishing pressure, is given by

$$w_{\epsilon,n+1}(\mathbf{x}',\epsilon\eta h(\mathbf{x}'/\epsilon)) = 0. \tag{29}$$

We seek the solution $\mathbf{w}_\epsilon$ of (12) as a power series in the parameter η:

$$\mathbf{w}_\epsilon(\mathbf{x}) = \mathbf{w}_\epsilon^0(\mathbf{x}) + \eta\mathbf{w}_\epsilon^1(\mathbf{x}) + \eta^2\mathbf{w}_\epsilon^2(\mathbf{x}) + \cdots \tag{30}$$

with all the terms bounded in $L^2_{\text{loc}}(R^n)$ and their boundary values $\mathbf{r}_\epsilon^j(\mathbf{x}') = \mathbf{w}_\epsilon^j(\mathbf{x}',0)$ bounded in $L^2_{\text{loc}}(R^{n-1})$. Our main assumption is that such an expansion exists, i.e., that the rest in (30) is bounded uniformly in η and ϵ by $o(\eta^2)$. We also assume that the medium is homogeneous above a certain height $x_n = L$ with L arbitrarily large. This assumption is purely technical and allows us to formulate an outgoing condition at infinity. Namely, for $x_n > L$, the plane wave decomposition is valid, and we assume then that the amplitudes of the incoming waves $\alpha_\epsilon(\mathbf{k}')$ are deterministic and given. The correctors $\mathbf{w}_\epsilon^j$ are all outgoing in that region, so that $\alpha_\epsilon^j(\mathbf{k}') = 0$ for $j \geqslant 1$. The characteristics of the transport equation (22) are given by

$$\frac{d\mathbf{x}}{ds} = \nabla_{\mathbf{k}}\omega_+(\mathbf{x},\mathbf{k}),$$

$$\frac{d\mathbf{k}}{ds} = -\nabla_{\mathbf{x}}\omega_+(\mathbf{x},\mathbf{k}).$$

We assume that the characteristics that leave the surface $x_n = 0$ reach the level $x_n = L$ and vice versa. Moreover, we assume that these characteristics do not come back. Then we have the following theorem.

Theorem 1: *Under the above assumptions, the Wigner matrix* $\mathbf{W}(\mathbf{x},\mathbf{k})$ *of the family* $\mathbf{w}_\epsilon$ *of solutions of the Dirichlet problem (29) is supported on the set* $\mathcal{S}=\{(\mathbf{x},\mathbf{k})\colon c(\mathbf{x})|\mathbf{k}|=\omega\}$, *and has the form (19). The scalar measure* μ *has the form* $\mu(\mathbf{x},\mathbf{k})=\mu'(\mathbf{x},\mathbf{k})+o(\eta^2)$. *The measure* $\mu'(\mathbf{x},\mathbf{k})$ *is a solution of the transport equation (20). The average measure* $\langle \mu'_{\mathrm{out}}(\mathbf{x}',\mathbf{k}')\rangle$ *is given in terms of* $\mu'_{\mathrm{in}}(\mathbf{x}',\mathbf{k}')$ *by*

$$\langle \mu'_{\mathrm{out}}(\mathbf{x}',\mathbf{k}')\rangle=\left[1-4\eta^2\int \frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{R}(\mathbf{k}'-\mathbf{p}')p_n^+(\mathbf{x}',\mathbf{p})k_n^+(\mathbf{x}',\mathbf{k}')\right]\mu'_{\mathrm{in}}(\mathbf{x}',\mathbf{k}')$$

$$+4\eta^2\int \frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{R}(\mathbf{k}'-\mathbf{p}')(p_n^+(\mathbf{p}'))^2\mu'_{\mathrm{in}}(\mathbf{x}',\mathbf{p}') \tag{31}$$

with $\mathbf{k}^+$ *and* $\mathbf{p}^+$ *given by (6) with* $K=K(\mathbf{x}')$. *Here* $K(\mathbf{x}')$, *determined by* $c(\mathbf{x}')K(\mathbf{x}')=\omega$, *is the radius of the sphere* $\mathcal{S}$ *of wave vectors, on which the measure* μ *is supported.*

The first term in (31) corresponds to the specular reflection and provides the correction to the reflection coefficient. The second term is the result of the diffuse scattering at the surface. It appears because the boundary is varying on the scale of the wavelength. Note that the total mean flux across the boundary $x_n=0$ vanishes:

$$\langle \mathcal{F}_n(\mathbf{x}',0)\rangle=c(\mathbf{x}')\int d\mathbf{k}\,\hat{k}_n\langle\mu(\mathbf{x}',0,\mathbf{k})\rangle$$

$$=c(\mathbf{x}')\int d\mathbf{k}'[\hat{k}_n^+ +\hat{k}_n^-]\mu_{\mathrm{in}}(\mathbf{x}',\mathbf{k}')+4\eta^2c(\mathbf{x}')\int d\mathbf{k}'\,d\mathbf{p}'\,\hat{R}(\mathbf{k}'-\mathbf{p}')$$

$$\times\{(p_n^+)^2\hat{k}_n^+\mu_{\mathrm{in}}(\mathbf{x}',\mathbf{k}')-p_n^+k_n^+\hat{k}_n^+\mu_{\mathrm{in}}(\mathbf{x}',\mathbf{p}')\}=0.$$

The expression (31) reduces in a homogeneous medium to (5) as one would expect.

The statement regarding the support and form of the matrix $\mathbf{W}(\mathbf{x},\mathbf{k})$ is proved exactly as in Ref. 8, so we will not repeat it here. We derive (31) in the following sections. Note that we can treat more general interface problems in a similar way, at least for incident energy fluxes away from grazing angle. The result is then a direct generalization of the formulas given in Ref. 7 for homogeneous media.

IV. THE PERTURBATION ANALYSIS

Now, we show how the diffusive scattering is obtained when the wave equation satisfies Dirichlet boundary conditions. We derive (31) in three steps. First we analyze the asymptotic expansion (30) in Sec. IV A. Then the diffuse part of the scattered energy is computed in Sec. IV B. At last, we derive the correction to the reflection coefficient in Sec. IV C.

A. The asymptotic expansion

We note that since the series (30) is asymptotic in $L^2_{\mathrm{loc}}(R^n)$, the Wigner matrix $\mathbf{W}$ is approximated by the Wigner matrix of the sum of the first N terms up to order $o(\eta^N)$. Thus we can compute the Wigner measure of the first three terms in the expansion (30) in order to approximate $\mathbf{W}$ up to $o(\eta^2)$. The terms $\mathbf{w}_\epsilon^j$, $j=0,1,2$ solve the acoustic equations (12) in the upper half space with the following boundary conditions:

$$w^0_{\epsilon,n+1}(\mathbf{x}',0)=0, \tag{32}$$

$$w^1_{\epsilon,n+1}(\mathbf{x}',0)=-h\left(\frac{\mathbf{x}'}{\epsilon}\right)\epsilon\frac{\partial w^0_{\epsilon,n+1}}{\partial x_n}(\mathbf{x}',0), \tag{33}$$

$$w^2_{\epsilon,n+1}(\mathbf{x}',0) = -\frac{1}{2}h^2\left(\frac{\mathbf{x}'}{\epsilon}\right)\epsilon^2\frac{\partial^2 w^0_{\epsilon,n+1}}{\partial x_n^2}(\mathbf{x}',0) - h\left(\frac{\mathbf{x}'}{\epsilon}\right)\epsilon\frac{\partial w^1_{\epsilon,n+1}}{\partial x_n}(\mathbf{x}',0). \tag{34}$$

These boundary conditions are constructed so as to satisfy the Dirichlet boundary conditions (29) on the rough surface ∂H_ϵ up to order η^2. The boundary conditions at infinity, where the medium is homogeneous above the level $x_n = L$, are as described above: $\mathbf{w}^0_\epsilon$ has prescribed incoming flux, and $\mathbf{w}^{1,2}_\epsilon$ are outgoing for $x_n > L$.

The boundary conditions (33) and (34) can be rewritten using the acoustic equations for $\mathbf{w}^0_\epsilon$ and $\mathbf{w}^1_\epsilon$ and the Dirichlet boundary condition (32) as

$$r^1_{\epsilon,n+1}(\mathbf{x}') = -i\omega\rho(\mathbf{x}')h\left(\frac{\mathbf{x}'}{\epsilon}\right)r^0_{\epsilon,n}(\mathbf{x}') \tag{35}$$

and

$$r^2_{\epsilon,n+1}(\mathbf{x}') = -\frac{i\epsilon\omega}{2}\frac{\partial\rho}{\partial x_n}(\mathbf{x}',0)h^2\left(\frac{\mathbf{x}'}{\epsilon}\right)r^0_{\epsilon,n}(\mathbf{x}') - i\omega\rho(\mathbf{x}')h\left(\frac{\mathbf{x}'}{\epsilon}\right)r^1_{\epsilon,n}(\mathbf{x}'). \tag{36}$$

Here $\mathbf{r}^j_\epsilon(\mathbf{x}')$ is the value of $\mathbf{w}^j_\epsilon$ at the boundary. The Wigner matrix of the sum of the first three terms in (30) has the form

$$\mathbf{W} = \mathbf{W}_0 + \eta(\mathbf{W}_{01} + \mathbf{W}_{10}) + \eta^2(\mathbf{W}_1 + \mathbf{W}_{02} + \mathbf{W}_{20}) + o(\eta^2), \tag{37}$$

where $\mathbf{W}_0 = \mathbf{W}[\mathbf{w}^0_\epsilon]$, $\mathbf{W}_1 = \mathbf{W}[\mathbf{w}^1_\epsilon]$, $\mathbf{W}_{01} = \mathbf{W}[\mathbf{w}^0_\epsilon, \mathbf{w}^1_\epsilon]$, etc. The leading order term in the expansion (37) has the form

$$\mathbf{W}_0 = \mu_0(\mathbf{x},\mathbf{k})\mathbf{b}(\mathbf{x},\mathbf{k})\mathbf{b}^*(\mathbf{x},\mathbf{k}) \tag{38}$$

since $\mathbf{w}^0_\epsilon$ solves the acoustic equations. The Dirichlet boundary conditions (32) for $\mathbf{w}^0_\epsilon$ imply that the Wigner measure $\boldsymbol{\nu}^0$ of $\mathbf{r}^0_\epsilon$ has the form (18) with the coefficients $\nu_{\alpha,\beta,\alpha\beta}$ related by (25):

$$\boldsymbol{\nu}^0(\mathbf{x}',\mathbf{k}') = \nu^0_\alpha(\mathbf{x}',\mathbf{k}')[\mathbf{b}(\mathbf{k}_+)\mathbf{b}^*(\mathbf{k}_+) + \mathbf{b}(\mathbf{k}_-)\mathbf{b}^*(\mathbf{k}_-) - \mathbf{b}(\mathbf{k}_+)\mathbf{b}^*(\mathbf{k}_-) - \mathbf{b}(\mathbf{k}_-)\mathbf{b}^*(\mathbf{k}_+)]. \tag{39}$$

Then the outgoing Wigner measure μ^0_{out} is

$$\mu^0_{\text{out}} = \mu_{\text{in}}, \tag{40}$$

where the measure μ_{in} is known. Notice that $\langle \mathbf{w}^1_\epsilon(\mathbf{x})\rangle = 0$, and since $\mathbf{w}^0_\epsilon$ is deterministic, we get

$$\langle \mathbf{W}_{01}\rangle = \langle \mathbf{W}_{10}\rangle = 0. \tag{41}$$

Thus we have to compute only $\langle \mathbf{W}_1\rangle$ and $\langle \mathbf{W}_{02} + \mathbf{W}_{20}\rangle$. The first term gives rise to the diffuse scattering and is treated in the following section. The second term produces the correction to the energy reflection coefficient and is considered in Sec. IV C.

B. The diffuse reflected wave

The matrix $\mathbf{W}^1$, being the Wigner matrix of a family of solutions $\mathbf{w}^1_\epsilon$ of the acoustic equations, has the form

$$\mathbf{W}_1 = \mu_1(\mathbf{x},\mathbf{k})\mathbf{b}(\mathbf{x},\mathbf{k})\mathbf{b}^*(\mathbf{x},\mathbf{k}).$$

The solution $\mathbf{w}_\epsilon^1$ is outgoing at infinity. Since we assumed that characteristics do not come back to the boundary, it is equivalent to being outgoing at the boundary. Then we have $\nu_\alpha^1=\mu_{\rm in}^1=0$. Since the matrix $\begin{pmatrix}\nu_\alpha^1 & \nu_{\alpha\beta}^1\\ \bar\nu_{\alpha\beta}^1 & \nu_\beta^1\end{pmatrix}$ is non-negative and definite, we also have $\nu_{\alpha\beta}^1=0$. Then the boundary Wigner measure $\boldsymbol{\nu}^1$ of $\mathbf{r}_\epsilon^1$ has the form

$$\boldsymbol{\nu}^1=\nu_\beta^1\mathbf{b}(\mathbf{k}_+)\mathbf{b}^*(\mathbf{k}_+) \tag{42}$$

and the boundary value of the measure μ_1 is

$$\mu_1(\mathbf{x}',0,\mathbf{k})=\nu_\beta^1(\mathbf{x}',\mathbf{k}')\,\delta(k_n-k_n^+),$$

hence $\mu_{\rm out}^1=\nu_\beta^1(\mathbf{x}',\mathbf{k}')$. The measure ν_β^1 is determined as follows. We have from (42),

$$\nu_{n+1,n+1}^1=\frac{1}{2\kappa(\mathbf{x}')}\,\nu_\beta^1, \tag{43}$$

where the left side is the $(n+1,n+1)$ entry of the matrix $\boldsymbol{\nu}^1(\mathbf{x}',\mathbf{x}')$. This entry is the Wigner measure of $r_{\epsilon,n+1}^1(\mathbf{x}')$, which is explicitly given by the boundary condition (35). Therefore,

$$\nu_{n+1,n+1}^1=\omega^2\rho^2(\mathbf{x}')\,\nu\left[h\left(\frac{\mathbf{x}'}{\epsilon}\right)r_{\epsilon,n}^0\right]$$

and so the average $\langle\nu_{n+1,n+1}^1\rangle$ is given by

$$\langle\nu_{n+1,n+1}^1\rangle=\omega^2\rho^2(\mathbf{x}')\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat R(\mathbf{k}'-\mathbf{p}')\,\nu_{n,n}^0(\mathbf{x}',\mathbf{p}'). \tag{44}$$

We use expression (39) for $\boldsymbol{\nu}^0$ to get

$$\langle\nu_{n+1,n+1}^1\rangle(\mathbf{x}',\mathbf{k}')=\omega^2\rho^2(\mathbf{x}')\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat R(\mathbf{k}'-\mathbf{p}')\,\frac{4\nu_\alpha^0(\mathbf{p}')p_n^{+2}}{2\rho(\mathbf{x}')|\mathbf{p}^+|^2}.$$

Insert this into (43) and use the relations $|\mathbf{p}^+|^2=\omega^2/c^2(\mathbf{x}')=\omega^2\kappa(\mathbf{x}')\rho(\mathbf{x}')$ and $\mu_{\rm in}=\nu_\alpha^0$ yields

$$\langle\mu_{\rm out}^1\rangle=\langle\nu_\beta^1\rangle(\mathbf{x}',\mathbf{k}')=4\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat R(\mathbf{k}'-\mathbf{p}')p_n^{+2}\mu_{\rm in}(\mathbf{x}',\mathbf{p}'). \tag{45}$$

This is the second term in (31).

C. Correction to the reflection coefficient

The correction to the coherent, or specular reflection coefficient arises from the term $\langle\mathbf{W}_{02}+\mathbf{W}_{20}\rangle$ in (37). Our analysis of this term proceeds in several steps. In Sec. IV C 1 we reduce the computation to evaluating the average $\langle\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1]\rangle$ of the cross Wigner distribution of $\mathbf{r}_\epsilon^0$ and the conjugated wave function $\mathbf{u}_\epsilon^1$ defined by (51). This Wigner distribution is described by Lemma 1. We use this lemma in Sec. IV C 2 to derive (31). Finally we prove Lemma 1 in Sec. IV C 3.

1. The reduction to the conjugated wave functions

The functions $\mathbf{w}_\epsilon^0,\mathbf{w}_\epsilon^2$ and $\mathbf{w}_\epsilon^0+\mathbf{w}_\epsilon^2$ are solutions of the acoustic equations, so the Wigner matrices $\mathbf{W}[\mathbf{w}^0],\mathbf{W}[\mathbf{w}^2]$ and $\mathbf{W}[\mathbf{w}^0+\mathbf{w}^2]$ are all of the form (19), and then

$$\mathbf{W}_{02}+\mathbf{W}_{20}=\mathbf{W}[\mathbf{w}^0+\mathbf{w}^2]-\mathbf{W}[\mathbf{w}^0]-\mathbf{W}[\mathbf{w}^2]$$
$$=(\mu[\mathbf{w}^0+\mathbf{w}^2]-\mu[\mathbf{w}^0]-\mu[\mathbf{w}^2])\mathbf{b}(\mathbf{k})\mathbf{b}^*(\mathbf{k})$$
$$=\mu_{02}\mathbf{b}(\mathbf{k})\mathbf{b}^*(\mathbf{k}), \tag{46}$$

so that $\mathbf{W}_{02}+\mathbf{W}_{20}$ has the same form as $\mathbf{W}_0$ and $\mathbf{W}_1$. The value of μ_{02} on the boundary is

$$\mu_{02}(\mathbf{x}',0,\mathbf{k})=(\mu[\mathbf{w}^0+\mathbf{w}^2]-\mu[\mathbf{w}^0]-\mu[\mathbf{w}^2])(\mathbf{x}',0,\mathbf{k})$$
$$=(\nu_\beta[\mathbf{r}^0+\mathbf{r}^2]-\nu_\beta[\mathbf{r}^0]-\nu_\beta[\mathbf{r}^2])(\mathbf{x}',\mathbf{k}')\,\delta(k_n-k_n^+)$$
$$=(\nu_\beta[\mathbf{r}^0,\mathbf{r}^2]+\nu_\beta[\mathbf{r}^2,\mathbf{r}^0])\,\delta(k_n-k_n^+)=\nu_\beta^{02}\,\delta(k_n-k_n^+). \tag{47}$$

The terms involving the incoming wave vector $\mathbf{k}_-$ cancel out since $\mathbf{w}^2$ is outgoing at the boundary and thus $\mu_{\text{in}}[\mathbf{w}^2]=0$, and $\mu_{\text{in}}[\mathbf{w}^0+\mathbf{w}^2]=\mu_{\text{in}}[\mathbf{w}^0]$. Thus we have to find $\nu_\beta^{02}=\nu_\beta[\mathbf{r}^0,\mathbf{r}^2]+\nu_\beta[\mathbf{r}^2,\mathbf{r}^0]$ to finish the computation. We note that the matrix Wigner measure $\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{r}^2]+\boldsymbol{\nu}[\mathbf{r}^2,\mathbf{r}^0]$ has the form

$$\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{r}^2]+\boldsymbol{\nu}[\mathbf{r}^2,\mathbf{r}^0]=(\nu_\beta[\mathbf{r}^0,\mathbf{r}^2]+\nu_\beta[\mathbf{r}^2,\mathbf{r}^0])\mathbf{b}(\mathbf{k}_+)\mathbf{b}^*(\mathbf{k}_+)$$
$$+\nu_{\alpha\beta}^{02}\mathbf{b}(\mathbf{k}_-)\mathbf{b}^*(\mathbf{k}_+)+\overline{\nu}_{\alpha\beta}^{02}\mathbf{b}(\mathbf{k}_+)\mathbf{b}^*(\mathbf{k}_-) \tag{48}$$

since $\mathbf{r}_\epsilon^2$ is outgoing. Evaluating the entry $(n+1,n+1)$ of both sides of (48), we get that ν_β^{02} is real. Moreover, evaluating the entry $(n,\,n+1)$ of the two sides of (48) we obtain

$$\mu_{\text{out}}^{02}=\nu_\beta^{02}=\frac{2\sqrt{\kappa\rho}}{\hat{k}_n^+}\,\text{Re}\,\nu[r_n^0,r_{n+1}^2]. \tag{49}$$

Thus we have reduced our problem to computing $\nu[r_n^0,r_{n+1}^2]$. Recall that $r_{\epsilon,n+1}^2(\mathbf{x}')$ is given by (36) and observe that the first term in (36) vanishes in the limit $\epsilon\to 0$ in L^2_{loc}. Then we have

$$\nu[r_n^0,r_{n+1}^2]=\nu\left[r_n^0(\mathbf{x}'),-h\left(\frac{\mathbf{x}'}{\epsilon}\right)i\omega\rho(\mathbf{x}')r_{\epsilon n}^1(\mathbf{x}')\right]$$
$$=i\omega\rho(\mathbf{x}')\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{h}(\mathbf{p}')\,\nu_{nn}[\mathbf{r}^0,\mathbf{u}^1](\mathbf{x}',\mathbf{k}',\mathbf{p}'). \tag{50}$$

Here we have introduced the conjugated wave functions $\mathbf{u}_\epsilon^1$:

$$\mathbf{u}_\epsilon^1(\mathbf{x}',\mathbf{p}')=\mathbf{r}_\epsilon(\mathbf{x}')e^{-i\mathbf{p}\cdot\mathbf{x}'/\epsilon}, \tag{51}$$

where the vector $\mathbf{p}'$ plays the role of a parameter. The Wigner matrix $\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1]$ is described by the following Lemma.

Lemma 1: The Wigner distribution matrix $\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1]$ *has the form*

$$\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1]=\tilde{\nu}_\alpha\mathbf{b}(\mathbf{k}^+)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+)-\tilde{\nu}_\alpha\mathbf{b}(\mathbf{k}^-)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+), \tag{52}$$

where $\tilde{\nu}_\alpha$ *is some distribution.*

We prove Lemma 1 in Sec. IV C 3.

2. The average reflection coefficient

We use Lemma 1 to evaluate $\nu_{nn}[\mathbf{r}^0,\mathbf{u}^1]$ in (50) and get

$$\left\langle \nu\left[r_n^0,h\left(\frac{\mathbf{x}'}{\epsilon}\right)r_{\epsilon n}^1\right]\right\rangle=\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\langle\hat{h}(\mathbf{p}')\tilde{\nu}_\alpha(\mathbf{x}',\mathbf{k}',\mathbf{p}')\rangle\frac{k_n^+(\mathbf{k}+\mathbf{p})^+}{|\mathbf{k}^+||(\mathbf{k}+\mathbf{p})^+|\rho(\mathbf{x})}. \tag{53}$$

The average $\langle \hat{h}(\mathbf{p})\tilde{\nu}_\alpha\rangle$ is evaluated as follows. We have from the boundary conditions (35),

$$u^1_{\epsilon,n+1}(\mathbf{x}')=e^{-i\mathbf{p}'\cdot\mathbf{x}'/\epsilon}r^1_{\epsilon,n+1}(\mathbf{x}')=-e^{-i\mathbf{p}'\cdot\mathbf{x}'/\epsilon}i\omega\rho(\mathbf{x}')h\left(\frac{\mathbf{x}'}{\epsilon}\right)r^0_{\epsilon,n}(\mathbf{x}')$$

and hence

$$\langle h(\mathbf{p}')\nu_{n,n+1}[\mathbf{r}^0,\mathbf{u}^1]\rangle=i\omega\rho(\mathbf{x})\hat{R}(\mathbf{p}')\nu^0_{nn}(\mathbf{x}',\mathbf{k}'). \tag{54}$$

But we also have from the representation (52),

$$\langle \hat{h}(\mathbf{p}')\nu_{n,n+1}[\mathbf{r}^0,\mathbf{u}^1]\rangle=\langle \hat{h}(\mathbf{p}')\tilde{\nu}_\alpha\rangle\frac{k_n^+}{|\mathbf{k}^+|\sqrt{\rho\kappa}},$$

so that (54) implies

$$\langle \hat{h}(\mathbf{p}')\tilde{\nu}_\alpha\rangle\frac{k_n^+}{|\mathbf{k}^+|\sqrt{\rho}}=\sqrt{\kappa}\hat{R}(\mathbf{p}')i\omega\rho(\mathbf{x}')\nu^0_{nn}.$$

We insert this into (53) and get

$$\begin{aligned}\left\langle \nu\left[r^0_{\epsilon n},h\left(\frac{\mathbf{x}'}{\epsilon}\right)r^1_{\epsilon n}\right]\right\rangle&=\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{R}(\mathbf{p}')\sqrt{\kappa\rho}i\omega\nu^0_{nn}(\mathbf{x}',\mathbf{k}')\frac{(\mathbf{k}+\mathbf{p})_n^+}{|(\mathbf{k}+\mathbf{p})^+|}\\&=\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{R}(\mathbf{k}'-\mathbf{p}')\sqrt{\kappa\rho}i\omega\hat{p}_n^+\nu^0_{nn}(\mathbf{x}',\mathbf{k}').\end{aligned} \tag{55}$$

Then we insert (55) into (50) and obtain

$$\langle \nu[r_n^0,r^2_{n+1}]\rangle=-\omega^2\rho(\mathbf{x}')\sqrt{\kappa(\mathbf{x}')\rho(\mathbf{x}')}\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{R}(\mathbf{k}'-\mathbf{p}')\frac{p_n^+}{|\mathbf{p}_+|}\nu^0_{nn}(\mathbf{x}',\mathbf{k}').$$

Recall that we have from (39)

$$\nu^0_{nn}(\mathbf{x}',\mathbf{k}')=\frac{4\nu^0_\alpha(\hat{k}_n^+)^2}{2\rho}.$$

Then we get from (49) and (55)

$$\begin{aligned}\langle\mu^{02}_{\text{out}}\rangle=\langle\nu^{02}_\beta\rangle&=-\frac{2\sqrt{\kappa\rho}}{\hat{k}_n^+}\frac{\omega^2\rho}{c}\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{R}(\mathbf{k}'-\mathbf{p}')\hat{p}_n^+\frac{4\nu^0_\alpha(\hat{k}_n^+)^2}{2\rho}\\&=-4\nu^0_\alpha(\mathbf{x}',\mathbf{k}')\int\frac{d\mathbf{p}'}{(2\pi)^{n-1}}\hat{R}(\mathbf{k}'-\mathbf{p}')p_n^+k_n^+\end{aligned} \tag{56}$$

because $|\mathbf{k}_+|=|\mathbf{p}_+|=\omega/c$.

Putting together (40), (45), and (56), we get (31).

3. Proof of Lemma 1

The proof of Lemma 1 is similar to that of the first part of Theorem 2, which is given in Ref. 8. The functions $\mathbf{u}^1_\epsilon(\mathbf{x}')$ satisfy the system

$$\sum_{j=1}^{n-1} \epsilon D^j \frac{\partial \mathbf{u}_\epsilon^1}{\partial x_j} + i\sum_{j=1}^{n-1} p_j D^j \mathbf{u}_\epsilon^1 - i\omega A(\mathbf{x}')\mathbf{u}_\epsilon^1 = -\epsilon D^n \frac{\partial \mathbf{w}_\epsilon^1}{\partial x_n}(\mathbf{x}',0), \tag{57}$$

while $\mathbf{r}_\epsilon^0(\mathbf{x}') = \mathbf{w}_\epsilon^0(\mathbf{x}',0)$ satisfy the system

$$\sum_{j=1}^{n-1} \epsilon D^j \frac{\partial \mathbf{r}_\epsilon^0}{\partial x_j} - i\omega A(\mathbf{x}')\mathbf{r}_\epsilon^0 = -\epsilon D^n \frac{\partial \mathbf{w}_\epsilon^0}{\partial x_n}(\mathbf{x}',0). \tag{58}$$

Let us define the reduced dispersion matrix

$$L'(\mathbf{x},\mathbf{k}) = \sum_{j=1}^{n-1} k_j D^j - \omega A(\mathbf{x}).$$

Let P be any matrix such that $PD^n = 0$, then (57) implies that

$$PL'(\mathbf{k}+\mathbf{p})\boldsymbol{\nu}[\mathbf{u}^1,\mathbf{r}^0] = 0$$

and (58) implies that

$$PL'(\mathbf{k})\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1] = 0.$$

Then the Wigner matrix $\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1]$ has the specific form

$$\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1] = \tilde{\nu}_\alpha \mathbf{b}(\mathbf{k}^+)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+) + \tilde{\nu}_\beta \mathbf{b}(\mathbf{k}^{-1})\mathbf{b}^*((\mathbf{k}+\mathbf{p})^-) + \tilde{\nu}_{\alpha\beta}\mathbf{b}(\mathbf{k}^+)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^-)$$
$$+ \tilde{\nu}_{\beta\alpha}\mathbf{b}(\mathbf{k}^-)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+). \tag{59}$$

Recall that $\mathbf{u}_\epsilon^1$ is outgoing, so (59) reduces to

$$\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1] = \tilde{\nu}_\alpha \mathbf{b}(\mathbf{k}^+)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+) + \tilde{\nu}_{\beta\alpha}\mathbf{b}(\mathbf{k}^-)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+). \tag{60}$$

The Dirichlet boundary conditions for $\mathbf{w}_\epsilon^0(\mathbf{x})$ imply that the fourth row of the matrix $\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1]$ vanishes. Then $\tilde{\nu}_{\alpha\beta} = -\tilde{\nu}_\alpha$, and (60) becomes

$$\boldsymbol{\nu}[\mathbf{r}^0,\mathbf{u}^1] = \tilde{\nu}_\alpha \mathbf{b}(\mathbf{k}^+)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+) - \tilde{\nu}_\alpha \mathbf{b}(\mathbf{k}^-)\mathbf{b}^*((\mathbf{k}+\mathbf{p})^+), \tag{61}$$

which is (52).

V. HIGHER ORDER ANISOTROPIC EFFECTS

The scattering cross-sections in the correctors of order η^2 in (31) depend upon the outgoing direction only through the power spectrum $\hat{R}(\mathbf{k}'-\mathbf{p}')$. We show in this section that some anisotropy of the scattered field is captured in the higher order terms. We obtain, in particular, a peak in the backscattering direction. This peak is similar to the coherent backscattering effect in the Neumann and impedance problems in homogeneous media studied in Refs. 17, 12, 13, and 18, although much broader for Dirichlet boundary conditions. The nature of this peak is also the constructive interference between direct and reverse paths of scattered waves. The singular behavior of the reflection operator at grazing angles in the Neumann and impedance problems makes this effect much more pronounced. The study of the coherent backscattering effect for these cases in an inhomogeneous medium requires appropriate modifications and the smoothing method.[19]

A. The backscattering for the Dirichlet problem

The angularly resolved energy density of the second order corrector in the Dirichlet problem is described by the following proposition.

Proposition 3: Assume that $h(\mathbf{y})$ is a mean zero Gaussian process with covariance matrix $R(\mathbf{y})$. Then the Wigner matrix $\mathbf{W}_2$ of the second order term $\mathbf{w}_\epsilon^2$ in the asymptotic expansion (30) of the solution of the Dirichlet problem has the form $\mathbf{W}_2=\mu_2(\mathbf{x},\mathbf{k})\mathbf{b}(\mathbf{x},\mathbf{k})\mathbf{b}^*(\mathbf{x},\mathbf{k})$. The scalar measure $\mu_2(\mathbf{x},\mathbf{k})$ is supported on the sphere $\mathcal{S}(\mathbf{x})=\{\mathbf{k}:|\mathbf{k}|=\omega/c(\mathbf{x})\}$ and satisfies the transport equation (20) with $\mu_2^{\text{in}}=0$ and

$$\langle\mu_2^{\text{out}}(\mathbf{x}',\mathbf{k}')\rangle=4\int\frac{d\mathbf{p}'\,d\mathbf{q}'}{(2\pi)^{2d-2}}\hat{R}(\mathbf{k}'-\mathbf{p}')\hat{R}(\mathbf{p}'-\mathbf{q}')q_n^2p_n[p_n+(\mathbf{q}'+\mathbf{k}'-\mathbf{p}')_n]\mu_{\text{in}}'(\mathbf{q}')$$
$$+4\int\frac{d\mathbf{p}'\,d\mathbf{q}'}{(2\pi)^{2d-2}}\hat{R}(\mathbf{k}'-\mathbf{p}')\hat{R}(\mathbf{k}'-\mathbf{q}')p_nq_nk_n^2\mu_{\text{in}}'(\mathbf{k}'). \tag{62}$$

Here $k_n(\mathbf{x},\mathbf{k}')=\sqrt{\omega^2/c^2(\mathbf{x})}$ is the normal component of the outgoing wave vector in $\mathcal{S}$, which has horizontal component $\mathbf{k}'$ and is pointing upwards, and μ_{in}' is as in Theorem 1.

The first term in (62) corresponds to the diffusive scattering. The second term provides a correction to the reflection coefficient. The differential scattering cross-section in (62) is no longer isotropic, and is centered in the backscattered direction. This can be seen as follows. Let us assume that the incident energy density has the form $\mu_{\text{in}}'(\mathbf{q}')=C(\mathbf{x}')\delta(\mathbf{q}'-\mathbf{q}_0')$, so that waves are coming from a single direction $\mathbf{q}_0$. Then the diffusive scattering is maximal in the direction with tangential component $\mathbf{k}'=-\mathbf{q}_0'$ because both terms in the diffusive scattering cross section are the same. The second term in this cross section is smaller in other directions, because when $\mathbf{k}+\mathbf{q}_0\neq 0$, then the integration in $\mathbf{p}'$ in that term is carried over the region where both $\mathbf{p}'$ and $\mathbf{k}'+\mathbf{q}_0'-\mathbf{p}'$ lie in the disk of radius $K=\omega/c$. This region is shrinking as $\mathbf{k}$ moves away from $-\mathbf{q}_0$, and so the contribution of this term diminishes. In particular, if $\mathbf{q}_0'$ is close to the boundary of the disk, and the incident wave is close to the grazing angle, then the contribution of this term in the forward direction $\mathbf{k}'=\mathbf{q}_0'$ vanishes. This can be interpreted as an enhanced backscattering phenomenon, since the contribution of this term in (62) corresponds to the interference of the direct and reverse paths, as will be seen in the derivation of (62).

B. Derivation of the scattering cross-section

The statements regarding the form of the Wigner matrix $\mathbf{W}_2$ and the support of the measure μ_2 in Proposition 3 follow immediately from Proposition 2, and from the fact that $\mathbf{w}_\epsilon^2$ is outgoing at infinity, together with our assumption that characteristics do not come back to the boundary. Thus, the only part we have to verify is the expression (62) for the measure μ_2 at the boundary. Let $\mathbf{r}_\epsilon^2(\mathbf{x}')=\mathbf{w}_\epsilon^2(\mathbf{x}',0)$ be the boundary value of $\mathbf{w}_2$. Then we have, using (36)

$$\mu_2^{\text{out}}(\mathbf{x}',\mathbf{k}')=2\kappa(\mathbf{x}')\nu[r_{\epsilon,n+1}^2]=2\kappa(\mathbf{x}')\omega^2\rho^2(\mathbf{x}')\nu[h(\mathbf{x}'/\epsilon)r_{\epsilon,n}^1]. \tag{63}$$

This can be rewritten with the help of the conjugated wave functions (51):

$$\nu[h(\mathbf{x}'/\epsilon)r_{\epsilon,n}^1]=\int\frac{d\mathbf{p}'\,d\mathbf{q}'}{(2\pi)^{2d-2}}\hat{h}(\mathbf{p}')\hat{h}(\mathbf{q}')\nu[u_{\epsilon,n}^1(\mathbf{x}',-\mathbf{p}'),u_{\epsilon,n}^1(\mathbf{x}',\mathbf{q}')].$$

It is easy to check that, similarly to Lemma 1 we have for any $\mathbf{p}_1',\mathbf{p}_2'$:

$$\boldsymbol{\nu}[\mathbf{u}_\epsilon^1(\mathbf{p}_1'),\mathbf{u}_\epsilon^1(\mathbf{p}_2')](\mathbf{x}',\mathbf{k}')=\theta(\mathbf{x}',\mathbf{k}',\mathbf{p}_1',\mathbf{p}_2')\mathbf{b}(\mathbf{k}'+\mathbf{p}_1')\mathbf{b}^*(\mathbf{k}'+\mathbf{p}_2') \tag{64}$$

with θ being some unknown distribution. Thus we have

$$\langle\nu[h(\mathbf{x}'/\epsilon)r_{\epsilon,n}^1]\rangle=\int\frac{d\mathbf{p}'\,d\mathbf{q}'}{(2\pi)^{2d-2}}\langle\hat{h}(\mathbf{p}')\hat{h}(\mathbf{q}')\theta(\mathbf{x}',\mathbf{k}',-\mathbf{p}',\mathbf{q}')\rangle b_n(\mathbf{k}'-\mathbf{p}')b_n(\mathbf{k}'+\mathbf{q}') \tag{65}$$

and we need to evaluate the average inside the integral. Expression (64) implies that

$$\langle \hat{h}(\mathbf{p}')\hat{h}(\mathbf{q}')\theta(\mathbf{x}',\mathbf{k}',-\mathbf{p}',\mathbf{q}')\rangle = 2\kappa\langle \hat{h}(\mathbf{p}')\hat{h}(\mathbf{q}')\nu[u^1_{\epsilon,n+1}(-\mathbf{p}'),u^1_{\epsilon,n+1}(\mathbf{q}')](\mathbf{x}',\mathbf{k}')\rangle. \tag{66}$$

The average on the right side can be computed using the expression (35) for $r^1_{\epsilon,n+1}$ in terms of $r^0_{\epsilon,n}$, and the assumption that $h(\mathbf{y})$ is a Gaussian random process so that

$$\langle \hat{h}(\mathbf{p})\hat{h}(\mathbf{q})\hat{h}(\mathbf{p}_2)\hat{h}(\mathbf{q}_2)\rangle = \langle \hat{h}(\mathbf{p})\hat{h}(\mathbf{q})\rangle\langle \hat{h}(\mathbf{p}_2)\hat{h}(\mathbf{q}_2)\rangle + \langle \hat{h}(\mathbf{p})\hat{h}(\mathbf{q}_2)\rangle\langle \hat{h}(\mathbf{q})\hat{h}(\mathbf{p}_2)\rangle + \langle \hat{h}(\mathbf{p})\hat{h}(\mathbf{p}_2)\rangle\langle \hat{h}(\mathbf{q})\hat{h}(\mathbf{q}_2)\rangle. \tag{67}$$

Then we get

$$\langle \hat{h}(\mathbf{p}')\hat{h}(\mathbf{q}')\nu[u^1_{\epsilon,n+1}(-\mathbf{p}'),u^1_{\epsilon,n+1}(\mathbf{q}')](\mathbf{x}',\mathbf{k}')\rangle = \mathrm{I}+\mathrm{II}+\mathrm{III}, \tag{68}$$

where

$$\mathrm{I} = \omega^2\rho^2(\mathbf{x}')\hat{R}(\mathbf{p}')\delta(\mathbf{p}'+\mathbf{q}')\int d\mathbf{p}_1'\hat{R}(\mathbf{p}_1')\nu^0_{nn}(\mathbf{k}'-\mathbf{p}'-\mathbf{p}_1'),$$

$$\mathrm{II} = \omega^2\rho^2(\mathbf{x}')\hat{R}(\mathbf{p}')\hat{R}(\mathbf{q}')\nu^0_{nn}(\mathbf{k}'+\mathbf{q}'-\mathbf{p}'),$$

and

$$\mathrm{III} = \omega^2\rho^2(\mathbf{x}')\hat{R}(\mathbf{p}')\hat{R}(\mathbf{q}')\nu^0_{nn}(\mathbf{k}').$$

The three terms in (68) have a natural interpretation in terms of wave scattering from a collection of discrete random scatterers.[18] The first term in (68) comes from the first term in (67) and corresponds to the interaction of a path with itself. It produces a scattering cross-section that is essentially isotropic. The second term in (68) comes from the second term in (67) and corresponds to the interaction of a path and its reverse one in the discrete picture. It has a peak in the backscattering direction as explained in the previous section. The last term arises from paths scattering twice on the same scatterer. It contributes to the specular reflection coefficient.

Finally we note that $\nu^0_{nn}(\mathbf{k}') = (k_n^2/2\rho|\mathbf{k}|^2)4\mu'_{\mathrm{in}}(\mathbf{x}',\mathbf{k}')$, and putting this together with (63), (65), (66), we obtain (62). This completes the proof of Proposition 3.

VI. CONCLUSIONS

We have derived the boundary conditions for the transport equation for the phase space resolved energy density in an inhomogeneous medium. Our derivation is based on the assumption that the asymptotic expansion (30) holds. These boundary conditions can be used for the radiative transport equation for acoustic waves[1,2] when randomness of the medium is independent of the randomness of the surface. Moreover, one can use similar boundary conditions for more general radiative transport equations[2] for electromagnetic, elastic and other waves in domains with rough boundaries. Our result may also be generalized to reflection and transmission at interfaces between two inhomogeneous media. The results are then a generalization of the diffuse energy reflection and transmission at a rough interface considered in Ref. 7.

The analysis of the Neumann problem in an inhomogeneous medium with a rough boundary requires the smoothing method or any equivalent regularization technique. We plan to address this in a separate note.[19] This allows one to incorporate the coherent backscattering effect into the boundary conditions for the radiative transport equation.

ACKNOWLEDGMENTS

We are grateful to Joseph B. Keller for numerous valuable discussions. L.R. would like to thank Luc Miller for his hospitality and stimulating discussions. The work of Bal and Papanicolaou was supported by AFOSR Grant No. F49620-98-1-0211 and by NSF Grant No. DMS-9622854. The work of Ryzhik was supported in part by the MRSEC Program of the National Science Foundation under Award Number DMR-9400379.

[1] S. Chandrasekhar, *Radiative Transfer* (Oxford University Press, Cambridge, 1960).
[2] L. Ryzhik, G. Papanicolaou, and J. B. Keller, ''Transport equations for elastic and other waves in random media,'' Wave Motion **24**, 327–370 (1996).
[3] J. A. DeSanto and G. S. Brown, ''Analytical techniques for multiple scattering from rough surfaces,'' in *Progress in Optics XXIII*, edited by E. Wolf (Elsevier, Amsterdam, 1986), pp. 1–62.
[4] I. M. Fuks and F. G. Bass, *Wave Scattering from Statistically Rough Surfaces* (Oxford; Pergamon, New York, 1979).
[5] M. Nieto-Vesperinas, *Scattering and Diffraction in Physical Optics* (Wiley, New York, 1991).
[6] A. Voronovich, *Wave Scattering from Rough Surfaces* (Springer–Verlag, Berlin, 1991).
[7] G. Bal, J. B. Keller, G. Papanicolaou, and L. Ryzhik, ''Transport theory for waves with reflection and transmission at interfaces,'' Wave Motion (in press).
[8] L. Ryzhik, G. Papanicolaou, and J. B. Keller, ''Transport equations in a half space,'' Commun. Partial Diff. Eqns. **22**, 1869–1910 (1997).
[9] J. B. Keller and R. Lewis, ''Asymptotic methods for partial differential equations: The reduced wave equation and Maxwell's equations,'' in *Surveys in Applied Mathematics*, edited by J. B. Keller, D. McLaughlin, and G. Papanicolaou (Plenum, New York, 1995).
[10] V. Twersky, ''On scattering and reflection of sound by rough surfaces,'' J. Acoust. Soc. Am. **29**, 209–225 (1957).
[11] J. G. Watson and J. B. Keller, ''Rough surface scattering via the smoothing method,'' J. Acoust. Soc. Am. **75**, 1705–1708 (1984).
[12] A. A. Maradudin and E. R. Mendez, ''Enhanced backscattering of light from weakly rough, random metal surfaces,'' Appl. Opt. **32**, 3335–3343 (1993).
[13] H. Ogura and N. Takahashi, ''Wave scattering from a random rough surface: reciprocal theorem and backscattering enhancement,'' Waves Random Media **5**, 232–242 (1995).
[14] P. Gérard, P. Markowich, N. Mauser, and F. Poupaud, ''Homogenization limits and Wigner transforms,'' Commun. Pure Appl. Math. **50**, 323–380 (1997).
[15] L. Miller, ''Propagation d'ondes semi-classiques a travers une interface et measures 2-microlocales,'' Doctorat de l'Ecole Polytechnique, Palaiseau.
[16] L. Miller, ''Refraction d'ondes semi-classiques par des interfaces franches,'' C. R. Acad. Sci., Ser. I: Math. **325**, 371–376 (1997).
[17] R. M. Fitzgerald, A. A. Maradudin, and F. Pincemin, ''Scattering of a scalar wave from a two-dimensional randomly rough Neumann surface,'' Waves Random Media **5**, 381–411 (1995).
[18] P. Sheng, *Introduction to Wave Scattering, Localization, and Mesoscopic Phenomena* (Academic, New York, 1995).
[19] G. Bal, J. B. Keller, G. Papanicolaou, and L. Ryzhik, ''Coherent Backscattering from a Random Surface via the Smoothing Method,'' preprint, 1999.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Transport equations for a general class of evolution equations with random perturbations

Maozheng Guo
Mathematics Department, Peking University, Beijing, 100871, People's Republic of China

Xiao-Ping Wang
Mathematics Department, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, People's Republic of China

(Received 28 January 1998; accepted for publication 22 March 1999)

We derive transport equations from a general class of equations of form $iu_t = H(X,D)u + V(X,D)u$ where $H(X,D)$ and $V(X,D)$ are pseudodifferential operators (Weyl operator) with symbols $H(x,k)$ and $V(x,k)$, where $H(x,k)$ being polynomial in k and smooth in $x, V(x,k)$ is a mean zero random function and is stationary in space variable. We also consider system of equations in the above form. Such equations cover many of the equations that arise in wave propagations, such as those considered in a paper by Ryzhik, Papanicolaou, and Keller [Wave Motion **24**, 327–370 (1996)]. Our results generalize those by Ryzhik, Papanicolau, and Keller. [S0022-2488(99)03209-0]

I. INTRODUCTION

There have been growing interests in the studies of transport equations for wave propagations. Radiative transport theory was introduced phenomenologically in order to describe the propagation of light energy through the atmosphere.[2] It is well known[1] that radiative transport equations will provide a good description of wave energy transport when (i) typical wavelengths are short compared to macroscopic features of the medium (high frequency approximation), (ii) correlation lengths of the inhomogeneities are comparable to wavelengths, and (iii) the fluctuations of the inhomogeneities are weak.

In a simple form, transport theory is as follows: a wave with wave vector $\mathbf{k}'$ at a point $\mathbf{x}$ in a randomly inhomogeneous medium may be scattered into any direction $\hat{\mathbf{k}}$ with wave vector $\mathbf{k}$, where $\hat{\mathbf{k}} = \mathbf{k}/|\mathbf{k}|$. Therefore, one must consider the angularly resolved, wave vector dependent, scalar energy density $a(t,\mathbf{x},\mathbf{k})$ defined for all $\mathbf{k}$ at each point $\mathbf{x}$ and time t. Energy conservation is expressed by the transport equation

$$\frac{\partial a(t,\mathbf{x},\mathbf{k})}{\partial t} + \nabla_{\mathbf{k}}\omega(\mathbf{x},\mathbf{k}) \cdot \nabla_{\mathbf{x}} a(t,\mathbf{x},\mathbf{k}) - \nabla_{\mathbf{x}}\omega(\mathbf{x},\mathbf{k}) \cdot \nabla_{\mathbf{k}} a(t,\mathbf{x},\mathbf{k})$$
$$= \int \sigma(\mathbf{x},\mathbf{k},\mathbf{k}') a(t,\mathbf{x},\mathbf{k}') d\mathbf{k}' - \Sigma(\mathbf{x},\mathbf{k}) a(t,\mathbf{x},\mathbf{k}). \tag{1}$$

Here $\omega(\mathbf{x},\mathbf{k})$ is the frequency at $\mathbf{x}$ of the wave with wave vector $\mathbf{k}$, $\sigma(\mathbf{x},\mathbf{k},\mathbf{k}')$ is the differential scattering cross-section, the rate at which energy with wave vector $\mathbf{k}'$ is converted to wave energy with wave vector $\mathbf{k}$ at position $\mathbf{x}$, and

$$\int \sigma(\mathbf{x},\mathbf{k}',\mathbf{k}) d\mathbf{k}' = \Sigma(\mathbf{x},\mathbf{k})$$

is the total scattering cross-section. The theory was applied to underwater sound propagation[3] and seismology.[4] It was realized in the 1960s that such theory arises rather naturally in wave propagation through random media. The equation (1) has been derived from equations governing the particular wave motion under consideration by various authors (see Ref. 1). These derivations also determine the functions $\omega(\mathbf{x},\mathbf{k})$ and $\sigma(\mathbf{x},\mathbf{k},\mathbf{k}')$ and show how a is related to the wave field.

In Ref. 1, transport equations for various types of waves and waves in random media were derived and studied more systematically. It is shown that the theory can be derived from the governing equations for light and for other waves of any type, in a randomly inhomogeneous medium. Their derivation motivated us to study this problem in a more general framework.

In this paper, we consider a general class of equations of form

$$i\epsilon u_t = H(X,\epsilon D)u + \sqrt{\epsilon}V\left(\frac{1}{\epsilon}X,\epsilon D\right)u \tag{2}$$

in the high frequency regime. Here $H(X,D)$ is a Weyl quantization of symbol $H(x,k)$ which is polynomial in k and smooth in x. In the case of system, we assume $\mathbf{H}(x,k)$ be a $n\times n$ matrix valued function of $C^\infty(\mathbf{R}_x^n\times\mathbf{R}_k^n)$. We show that transport equation can be derived from (2) in the high frequency limit. Equation (2) is general enough to cover various physical examples given in Ref. 1 by choosing different symbols. Our results generalized those in Ref. 1. By going into a more abstract notion of pseudo-differential operator or Weyl quantization, we can use many of its nice properties and thus simplify the calculations.

As in Ref. 1, an essential step in our approach to deriving radiative transport equations from wave equations is the use of the Wigner distribution. Wigner distribution was introduced first by Wigner[5] in the context of semiclassical quantum mechanics. The paper of Lions and Paul[6] contains many basic facts about the Wigner distribution as well as some interesting applications. The Wigner distribution are closely related to the representation theory of Heisenberg group and pseudodifferential operator theory.[7] The advantage with Wigner distribution is that it can be regarded as wave number resolved wave intensity. In particular, in the case of high frequency limit, it is a good candidate to analyze the evolution of wave energy.

The paper is organized as follows. In Sec. II, we review some of the basic properties of Weyl quantization and Wigner distribution. In Sec. III, we consider a scaled generalized Schrödinger evolution equation with Hamiltonian related to a symbol $H(x,k)$. In the high frequency limit, the limiting Wigner distribution W satisfies a Liouville-type transport equation

$$\frac{\partial W(t)}{\partial t} + \{W(t),H\} = 0. \tag{3}$$

We then consider the equation with small random perturbation. Using the method of multiple scale expansion, we obtain the transport equations for the leading order term $W^{(0)}$ of W^ϵ

$$\frac{\partial W^{(0)}}{\partial t} + \{W^{(0)},H\} = 2\pi\int \mathcal{F}_1 R\left(k-q,k+\frac{q}{2},k+\frac{q}{2}\right)(W^{(0)}(t,x,q) - W^{(0)}(t,x,k))\cdot\delta(H(x,q)-H(x,k))\,dq. \tag{4}$$

Equation (4) is exactly in the form of Eq. (1), where $W^{(0)}$ is the energy density and the symbol Hamiltonian $H(x,k)$ is the frequency $\omega(\mathbf{x},\mathbf{k})$ and the integral kernel of the right-hand side of (4) is the differential scattering cross section σ and total scattering cross section Σ. In Sec. IV, we treat the system of scaled generalized Schrödinger evolution equations. The Wigner distributions are matrix valued. In the high frequency limit, we show that the projections $a^{(s)}$ of limit Wigner distribution on certain eigenspaces satisfy the following Liouville equations:

$$\frac{\partial a^{(s)}}{\partial t} + \{a^{(s)},\lambda_s\} = 0, \tag{5}$$

where λ_s denote dispersion laws of the given Hamiltonian symbol $\mathbf{H}(x,k) = (H_{i,j}(x,k))$. We further show this process also incorporates the system of evolution equations with small random perturbations. We derive in Proposition 9 the system of radiative transport equations which leads to the appearance of the scattering terms on the right-hand side of Eq. (71) similar to (1). We also treat the polarization case which is more physically relevant and can find applications in acoustic wave, electromagnetic waves and elastic waves.

We note that rigorous results have also appeared recently for some simpler cases. In Ref. 8, derivations of transport equations in some cases are proved rigorously for equations without random perturbations. Efforts are also made to include some special random perturbations.[9]

II. WEYL QUANTIZATION AND WIGNER DISTRIBUTION

A. Weyl quantization

The idea of Weyl quantization is to assign to a function (symbol) on the phase space an operator according to the Weyl quantization rule. The procedure was proposed by Weyl not long after the invention of quantum mechanics. Weyl's prescription for assigning an operator $a(X,D)$ to a function $a(x,k)$ amounts to postulating that the exponential function $\exp i(q\cdot x+p\cdot k)((q,p)\in \mathbf{R}^n)$ should correspond to the operator $\rho(q,p)=\exp i(q\cdot X+p\cdot D)$. Once this is granted, one can expand an ''arbitrary'' $a(x,k)$ in terms of exponentials via the Fourier transform,

$$a(X,D)=\frac{1}{(2\pi)^n}\int \hat{a}(q,p)\exp i(q\cdot X+p\cdot D)dq\,dp, \tag{6}$$

where $\hat{a}(q,p)$ denotes the Fourier transform of $a(q,p)$.

For $(q,p)\in \mathbf{R}^n\times\mathbf{R}^n$, let $q\cdot X=\Sigma q_jX_j$, $p\cdot D=\Sigma p_jD_j$ where X_j's are multiplication operators, D_j's are first-order differential operators, i.e., for $\forall f\in L^2(\mathbf{R}^n,dx)$ we have

$$X_jf(x)=x_jf(x), \qquad D_jf(x)=\frac{1}{i}\frac{\partial f(x)}{\partial x_j}. \tag{7}$$

We define a unitary operator $\rho(q,p)=\exp i(q\cdot X+p\cdot D)$ on $L^2(\mathbf{R}^n,dx)$ as follows: For $f\in L^2(\mathbf{R}^n,dx)$,

$$\rho(q,p)f(x)=e^{i(1/2)q\cdot p}e^{iq.x}f(x+p). \tag{8}$$

The function a in (6) is usually called the symbol of the operator $a(X,D)$. The integral is an ordinary Bochner integral if $\hat{a}\in L^1(\mathbf{R}^n\times\mathbf{R}^n)$. If $a\in S$ (the Schwartz class), the Weyl operator is an integral operator

$$a(X,D)f(x)=\frac{1}{(2\pi)^n}\int a\left(\frac{x+y}{2},z\right)e^{i(x-y)z}f(z)dy\,dz \tag{9}$$

and $a(X,D)$ maps S into S. $a(X,D)$ can be defined even for any tempered distribution function a as a continuous linear operator from $S(\mathbf{R}^n)$ to $S'(\mathbf{R}^n)$. In particular, if $a(x,k)$ is a polynomial in (x,k), then $a(X,D)$ is a differential operator.

Example 1. Let $a(x,k)=\frac{1}{2}|k|^2+V(x)$, then

$$a(X,D)=-\tfrac{1}{2}\Delta+V(x).$$

Example 2. Let

$$a(x,k)=\sum_{j=1}^n A_j(x)k_j-\frac{1}{2}i\sum_{j=1}^n\frac{\partial A_j(x)}{\partial x_j},$$

then

$$a(X,D)=\sum_{j=1}^{n} A_j(x)D_j .$$

B. Wigner transformation

Wigner transform plays an important role in the derivation of transport equations from wave equations. We first introduce the so-called Fourier–Wigner transform. For $u,v\in L^2(\mathbf{R}^n)$, the Fourier–Wigner transform is defined as

$$V(u,v)(q,p)=(\rho(q,p)u,v), \tag{10}$$

where the bracket denotes the inner product of $L^2(\mathbf{R}^n)$ and $\rho(q,p)$ is given by (8). $V(u,v)$ is a bounded continuous function on $\mathbf{R}^n\times\mathbf{R}^n$. It maps $S(\mathbf{R}^n)\times S(\mathbf{R}^n)$ into $S(\mathbf{R}^{2n})$ and may be extended to an operator from $S'(\mathbf{R}^n)\times S'(\mathbf{R}^n)$ to $S'(\mathbf{R}^n\times\mathbf{R}^n)$.

Now the Wigner transformation for $u,v\in L^2(\mathbf{R}^n)$ is defined by

$$W(u,v)(x,k)=\frac{1}{(2\pi)^n}\int e^{-i(x\cdot q+k\cdot p)}V(u,v)(q,p)dq\,dp \tag{11}$$

$$=\int e^{-ik\cdot p}u\left(x+\frac{p}{2}\right)\overline{v}\left(x-\frac{p}{2}\right)dp. \tag{12}$$

$W(u,v)$ is a bounded continuous function on the phase space. It maps $S(\mathbf{R}^n)\times S(\mathbf{R}^n)\to S(\mathbf{R}^n\times\mathbf{R}^n)$, and may extend to an operator from $S'(\mathbf{R}^n)\times S'(\mathbf{R}^n)$ to $S'(\mathbf{R}^n\times\mathbf{R}^n)$. When $u=v$, we shall call $W(u,u)$ the Wigner distribution or Wigner transform of u. It is easy to see that

$$\int W(u,u)(x,k)dk=|u(x)|^2,$$

so that we may think of $W(x,k)$ as wave number resolved energy density, if u is a wave function. We list below some of the properties of the Fourier–Wigner and Wigner transform without proving it (see, e.g., Ref. 7). Let u, v and $u_1,v_1,u_2,v_2\in L^2(\mathbf{R}^n)$, we have

(1)

$$\|V(u,v)\|_\infty=\|u\|_2\|v\|_2;$$

(2)

$$(V(u_1,v_1),V(u_2,v_2))=(u_1,u_2)\overline{(v_1,v_2)};$$

(3)

$$V(\rho(a,b)u,v)(q,p)=e^{i/2(pa-bq)}V(u,v)(q+a,p+b),$$

$$V(u,\rho(c,d)v)(q,p)=e^{i/2(pc-qd)}V(u,v)(q-c,p-d);$$

(4)

$$\|W(u,v)\|_\infty=\|u\|_2\|v\|_2;$$

(5)

$$(W(u_1,v_1),W(u_2,v_2))=(u_1,u_2)\overline{(v_1,v_2)};$$

(6)

$$W(v,u)=\overline{W(u,v)},$$

$$W(\hat{u},\hat{v})(x,k)=W(u,v)(-k,x);$$

(7)

$$W(\rho(a,b)u,v)(x,k)=e^{i(x\cdot a+b\cdot k)}W(u,v)\left(x+\frac{b}{2},k-\frac{a}{2}\right),$$

$$W(u,\rho(c,d)v)(x,k)=e^{-i(x\cdot c+k\cdot d)}W(u,v)\left(x+\frac{d}{2},k-\frac{c}{2}\right);$$

(8)

$$\int W(u,v)(x,k)dk=|u(x)|^2,$$

$$\int W(u,v)(x,k)dx=|\hat{u}(k)|^2,$$

$$\int x_j W(u,u)(x,k)dx\,dk=\int x_j|u(x)|^2dx=(X_ju,u),$$

$$\int k_j W(u,u)(x,k)dx\,dk=\int k_j|\hat{u}(x)|^2dk=(D_ju,u).$$

III. THE SCALAR EVOLUTION EQUATIONS

A. Generalized Schrödinger equation

We consider the symbol $H(x,k)\in C^\infty(\mathbf{R}_x^n\times\mathbf{R}_k^n)$ being real valued polynomial in k-variable and Schwartz class in x-variable, then its Weyl operator is a differential operator with smooth coefficient. Let $u(t,x)$ be the solution of the initial value problem of the generalized Schrödinger equation

$$i\frac{\partial u}{\partial t}=H(X,D)u,$$
$$u(0,x)=\varphi(x)\in L^2(\mathbf{R}^n) \tag{13}$$

and its Wigner distribution be

$$W(t,x,k)=W(u(t),u(t))(x,k). \tag{14}$$

Then we have,

Proposition 1: The Wigner distribution $W(t,x,k)$ satisfies

$$i\frac{\partial W(t,x,k)}{\partial t}=\frac{1}{(2\pi)^n}\int \hat{H}(q,p)e^{i(x\cdot q+k\cdot p)}\left[W\left(t,x+\frac{p}{2},k-\frac{q}{2}\right)-W\left(t,x-\frac{p}{2},k+\frac{q}{2}\right)\right]dq\,dp$$
$$W(0,x,k)=W(\varphi,\varphi)(x,k). \tag{15}$$

Proof: Differentiate $W(t,x,k)$ with respect to t, we have

$$\frac{\partial W(t,x,k)}{\partial t}=W\left(\frac{\partial u(t)}{\partial t},u(t)\right)(x,k)+W\left(u(t),\frac{\partial u(t)}{\partial t}\right)(x,k)$$
$$=\frac{1}{i}W(Hu(t),u(t))(x,k)-\frac{1}{i}W(u(t),Hu(t))(x,k).$$

Using (6), (8), (11), and Property (7) in Sec. II B, we have

$$W(H(X,D)u,u)(x,k)=\frac{1}{(2\pi)^n}\int \hat{H}(q,p)W(\rho(q,p)u,u)(x,k)dq\,dp$$
$$=\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{H}(q,p)W(u,u)\left(x+\frac{p}{2},k-\frac{q}{2}\right)dq\,dp,$$

$$W(u,H(X,D)u)(x,k)=\frac{1}{(2\pi)^n}\int \overline{\hat{H}}(q,p)W(u,\rho(q,p)u)(x,k)dq\,dp$$
$$=\frac{1}{(2\pi)^n}\int e^{-i(x\cdot q+k\cdot p)}\hat{H}(-q,-p)W(u,u)\left(x+\frac{p}{2},k-\frac{q}{2}\right)dq\,dp$$
$$=\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{H}(q,p)W(u,u)\left(x-\frac{p}{2},k+\frac{q}{2}\right)dq\,dp.$$

Therefore,

$$i\frac{\partial W(t,x,k)}{\partial t}=\frac{1}{(2\pi)^n}\int \hat{H}(q,p)e^{i(x\cdot q+k\cdot p)}\left[W\left(t,x+\frac{p}{2},k-\frac{q}{2}\right)-W\left(t,x-\frac{p}{2},k+\frac{q}{2}\right)\right]dq\,dp,$$

and we have the Lemma.

We now consider the problem in the high frequency regime. High frequency asymptotics requires the symbol $H(x,k)$ vary slowly in space variable x, i.e., on the long scale. We introduce a dimensionless small parameter $\epsilon>0$ and redefine time and space variables $t\to t/\epsilon$, $x\to x/\epsilon$. The scaled wave function $u^\epsilon(t,x)=u(t/\epsilon,x/\epsilon)$ satisfies the scaled generalized Schrödinger equation

$$i\varepsilon\frac{\partial u^\varepsilon}{\partial t}=H(X,\varepsilon D)u^\varepsilon \qquad u^\varepsilon(0,x)=\varphi(x). \tag{16}$$

Note that there is no small parameter ϵ before the multiplications operator X in the scaled Weyl operator H, which means the symbol is slowly varying in the original space variable. Also the initial data does not depend on ϵ which means the initial data in (13) is slowly varying in the original space variables. We define the rescaled Wigner distribution by

$$W^\varepsilon(t,x,k)=\frac{1}{\varepsilon^n}W(u^\varepsilon(t),u^\varepsilon(t))\left(x,\frac{k}{\varepsilon}\right). \tag{17}$$

With this scaling, Proposition 1 becomes

Proposition 2: The Wigner distribution W^ε satisfies

$$i\varepsilon \frac{\partial w^{\varepsilon}(t,x,k)}{\partial t} = \frac{1}{(2\pi)^n} \int \hat{H}(q,p) e^{i(x\cdot q+k\cdot p)} \left[W^{\varepsilon}\left(t,x+\frac{\varepsilon}{2}p,k-\frac{\varepsilon}{2}q\right) - W^{\varepsilon}\left(t,x-\frac{\varepsilon}{2}p,k+\frac{\varepsilon}{2}q\right)\right] dq\,dp$$

$$W^{\varepsilon}(0,x,k) = \frac{1}{\varepsilon^n} W(\varphi,\varphi)\left(x,\frac{k}{\varepsilon}\right). \tag{18}$$

To calculate the high frequency limit of W^{ϵ} as $\epsilon \to 0$, consider the formal expansion

$$W^{\varepsilon}(t,x,k) = W(t,x,k) + \varepsilon W_1(t,x,k) + \cdots. \tag{19}$$

The equation for the leading order term can be calculated formally and we have

Proposition 3: $W(t,x,k)$ satisfies the following transport equation:

$$\frac{\partial W(t)}{\partial t} + \{W(t),H\} = 0, \tag{20}$$

where the bracket is the Lie bracket on the space of the functions on phase space

$$\{a,b\} = \sum_j \left(\frac{\partial a}{\partial x_j}\frac{\partial b}{\partial k_j} - \frac{\partial a}{\partial k_j}\frac{\partial b}{\partial x_j}\right). \tag{21}$$

Note that $H(x,k)$ plays the role of frequency function $\omega(x,k)$ in (1) and there is no scattering in the equation.

Proof: From (18), we take limit $\varepsilon \to 0$, we have that

$$\begin{aligned} i\frac{\partial W}{\partial t} &= \frac{1}{(2\pi)^n} \int e^{i(x\cdot q+k\cdot q)} \hat{H}(q,p)\left(p\cdot \frac{\partial W}{\partial x} - q\cdot \frac{\partial W}{\partial k}\right) dq\,dp \\ &= \frac{1}{(2\pi)^n} \frac{\partial W}{\partial x} \frac{1}{i} \frac{\partial}{\partial k}\left(\int e^{i(x\cdot q+k\cdot p)} \hat{H}(q,p)\,dq\,dp\right) \\ &\quad - \frac{1}{(2\pi)^n} \frac{\partial W}{\partial k} \frac{1}{i} \frac{\partial}{\partial x}\left(\int e^{i(x\cdot q+k\cdot q)} \hat{H}(q,p)\,dq\,dp\right) \\ &= \frac{1}{i}\left(\frac{\partial W}{\partial x}\frac{\partial H}{\partial k} - \frac{\partial W}{\partial k}\frac{\partial H}{\partial x}\right) = \frac{1}{i}\{W,H\}. \end{aligned}$$

B. Generalized Schrödinger equation with small random perturbation

We now consider the case with random potential. Let $V(x,k)$ be a mean zero random function, stationary in space variable, its correlations are homogeneous in space, so that

$$\langle V(x,k)V(x',k')\rangle = R(x-x',k,k') \tag{22}$$

and

$$\langle \hat{V}(q,p)\hat{V}(q',p')\rangle = \hat{R}(q,p,p')\,\delta(q+q'), \tag{23}$$

where

$$\hat{R}(q,p,p') = \frac{1}{(2\pi)^{3n/2}} \int e^{i(xq+kp+k'p')} R(x,k,k')\,dk\,dk'.$$

We also assume that

$$\operatorname{Im}\mathcal{F}_1R(q,k,k')=0, \tag{24}$$

which is equivalent to

$$\operatorname{Im}\mathcal{F}_1R(q,k,k')=\operatorname{Im}\mathcal{F}_1R(-q,k,k'), \tag{25}$$

where $\mathcal{F}_1R$ stands for the Fourier transform of the correlation function in first variable.

Consider the solution $u^\epsilon(t,x)$ of equation

$$i\epsilon u_t^\epsilon=\left(H(X,\epsilon D)+\sqrt{\epsilon}V\left(\frac{1}{\epsilon}X,\epsilon D\right)\right)u^\epsilon$$

and its scaled Wigner distribution W^ϵ defined by (17). As pointed out in Ref. 1, in order to find the correct result as $\epsilon\to 0$, we need to use multiple scale expansion. Let $\xi=x/\varepsilon$ be the fast space variable and rewrite $W^\varepsilon(t,x,k)=W^\varepsilon(t,x,x/\varepsilon,k)$ as $W^\varepsilon(t,x,\xi,k)$. It is easy to show that $W^\varepsilon(t,x,\xi,k)$ satisfies the following proposition:

Proposition 4:

$$\begin{aligned}
i\frac{\partial W^\varepsilon(t,x,\xi,k)}{\partial t}=&\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{H}(q,p)\frac{1}{\varepsilon}\left[W^\varepsilon\left(t,x+\frac{\varepsilon}{2}p,\xi+\frac{p}{2},k-\frac{\varepsilon}{2}q\right)\right.\\
&\left.-W^\varepsilon\left(t,x-\frac{\varepsilon}{2}p,\xi-\frac{p}{2},k+\frac{\varepsilon}{2}q\right)\right]dq\,dp\\
&+\frac{1}{(2\pi)^n}\int e^{i(\xi\cdot q+k\cdot p)}\hat{V}(q,p)\frac{1}{\sqrt{\varepsilon}}\left[W^\varepsilon\left(t,x+\frac{\varepsilon}{2}p,\xi+\frac{p}{2},k-\frac{q}{2}\right)\right.\\
&\left.-W^\varepsilon\left(t,x,-\frac{\varepsilon}{2}p,\xi-\frac{p}{2},k+\frac{q}{2}\right)\right]dq\,dp.
\end{aligned} \tag{26}$$

To study the limit as $\epsilon\to 0$, we introduce an expansion of form

$$W^\epsilon(t,x,\xi,k)=W^{(0)}(t,x,k)+\sqrt{\varepsilon}W^{(1)}(t,x,\xi,k)+\varepsilon W^{(2)}(t,x,\xi,k)+\cdots, \tag{27}$$

where we assume that the leading term $W^{(0)}$ does not depend upon the fast variable and also it is deterministic. Substitute (27) into (26) and collect the terms at different powers of ϵ, we have that at order $\epsilon^{-1/2}$,

$$\begin{aligned}
&\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{H}(q,p)\left[W^{(1)}\left(t,x,\xi+\frac{p}{2},k\right)-W^{(1)}\left(t,x,\xi-\frac{p}{2},k\right)\right]dq\,dp\\
&\quad+\frac{1}{(2\pi)^n}\int e^{i(\xi\cdot p+k\cdot p)}\hat{V}(q,p)\left[W^{(0)}\left(t,x,k-\frac{q}{2}\right)-W^{(0)}\left(t,x,k+\frac{q}{2}\right)\right]dq\,dp=0,
\end{aligned} \tag{28}$$

at order ϵ^0:

$$i\frac{\partial W^{(0)}(t,x,k)}{\partial t}=\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{H}(q,p)\left[W^{(2)}\left(t,x,\xi+\frac{p}{2},k\right)-W^{(2)}\left(t,x,\xi-\frac{p}{2},k\right)\right]dq\,dp$$
$$+\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{H}(q,p)\left(p\frac{\partial}{\partial x}-q\frac{\partial}{\partial k}\right)W^{(0)}(t,x,k)dq\,dp$$
$$+\frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\hat{V}(q,p)\left[W^{(1)}\left(t,x,\xi+\frac{p}{2},k-\frac{q}{2}\right)\right.$$
$$\left.-W^{(1)}\left(t,x,\xi-\frac{p}{2},k+\frac{q}{2}\right)\right]dq\,dp. \tag{29}$$

Since we have, from ergodicity,

$$\left\langle\frac{\partial W^{(2)}}{\partial \xi}\right\rangle=0,$$

after averaging (29), we get

$$i\frac{\partial W^{(0)}(t,x,k)}{\partial t}+\{W^{(0)},H\}(t,x,k)=\frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\left\langle\hat{V}(q,p)\left[W^{(1)}\left(t,x,\xi+\frac{p}{2},k-\frac{q}{2}\right)\right.\right.$$
$$\left.\left.-W^{(1)}\left(t,x,\xi-\frac{p}{2},k+\frac{q}{2}\right)\right]dq\,dp\right\rangle. \tag{30}$$

In order to get a closed equation for $W^{(0)}$, we need to compute $W^{(1)}$ from (28). Let F be the Fourier transform in ξ of $W^{(1)}$, i.e.,

$$W^{(1)}(t,x,\xi,k)=\frac{1}{(2\pi)^{n/2}}\int e^{i\xi\cdot q'}F(t,x,q',k)dq'.$$

Plug in (28), we have

$$\frac{1}{(2\pi)^{3n/2}}\int\int e^{i\xi\cdot q'}F(t,x,q',k)[e^{i(x\cdot q+(k+q'/2)p)}\hat{H}(q,p)-e^{i(x\cdot q+(k-q'/2)p)}\hat{H}(q,p)]$$
$$\times dq\,dp\,dq'+\frac{1}{(2\pi)^n}\int e^{i(\xi\cdot q+k\cdot p)}\hat{V}(q,p)\left[W^{(0)}\left(t,x,k-\frac{q}{2}\right)-W^{(0)}\left(t,x,k+\frac{q}{2}\right)\right]dq\,dp=0,$$

which implies

$$\frac{1}{(2\pi)^{n/2}}\int e^{i\xi\cdot q'}F(t,x,q',k)[H(x,k+q'/2)-H(x,k-q'/2)]dq'$$
$$+\frac{1}{(2\pi)^n}\int e^{i\xi\cdot q}\int e^{ik\cdot p}\hat{V}(q,p)dp\left[W^{(0)}\left(t,x,k-\frac{q}{2}\right)-W^{(0)}\left(t,x,k+\frac{q}{2}\right)\right]dq=0.$$

Taking inverse Fourier transform in ξ, we get

$$F(t,x,q',k)=\frac{1}{(2\pi)^{n/2}}\frac{\int e^{ik\cdot p}\hat{V}(q',p)dp[W^{(0)}(t,x,k-q'/2)-W^{(0)}(t,x,k+q'/2)]}{H(x,k+q'/2)-H(x,k-q'/2)+i\theta}. \quad (31)$$

The term $i\theta$ is a regularization term. Using (23), we compute the integrand of the right-hand side of (30),

$$\left\langle \hat{V}(q,p)\left[W^{(1)}\left(t,x,\xi+\frac{p}{2},k-\frac{q}{2}\right)-W^{(1)}\left(t,x,\xi-\frac{p}{2},k+\frac{q}{2}\right)\right]\right\rangle$$
$$=\frac{1}{(2\pi)^{n/2}}\left\langle \hat{V}(q,p)\left[\int e^{i(\xi+p/2)q'}F\left(t,x,q',k-\frac{1}{2}\right)dq'\right.\right.$$
$$\left.\left.-\int e^{i(\xi-p/2)q'}F\left(t,x,q',k+\frac{q}{2}\right)dq'\right]\right\rangle$$
$$=\frac{1}{(2\pi)^n}\left\langle \hat{V}(q,p)\int e^{i(\xi+p/2)q'+i(k-q/2)p'}\hat{V}(q',p')\right.$$
$$\left.\times\frac{W^{(0)}\left(t,x,k-\frac{q}{2}+\frac{q'}{2}\right)-W^{(0)}\left(t,x,k-\frac{q}{2}-\frac{q'}{2}\right)}{H\left(x,k-\frac{q}{2}+\frac{q'}{2}\right)-H\left(x,k-\frac{q}{2}-\frac{q'}{2}\right)+i\theta}dq'\,dp'\right\rangle$$
$$-\frac{1}{(2\pi)^n}\left\langle \hat{V}(q,p)\int e^{i(\xi-p/2)q'+i(k+q/2)p'}\hat{V}(q',p')\right.$$
$$\left.\times\frac{W^{(0)}\left(t,x,k+\frac{q}{2}+\frac{q'}{2}\right)-W^{(0)}\left(t,x,k+\frac{q}{2}-\frac{q'}{2}\right)}{H\left(x,k+\frac{q}{2}+\frac{q'}{2}\right)-H\left(x,k+\frac{q}{2}-\frac{q'}{2}\right)+i\theta}dq'\,dp'\right\rangle$$
$$=\frac{1}{(2\pi)^n}\int e^{i(\xi+p/2)q'+i(k-q/2)p'}\hat{R}(q,p,p')\delta(q+q')$$
$$\times\frac{W^{(0)}\left(t,x,k-\frac{q}{2}+\frac{q'}{2}\right)-W^{(0)}\left(t,x,k-\frac{q}{2}-\frac{q'}{q}\right)}{H\left(x,k-\frac{q}{2}+\frac{q'}{2}\right)-H\left(x,k-\frac{q}{2}-\frac{q'}{2}\right)+i\theta}dq'\,dp'$$
$$-\frac{1}{(2\pi)^n}\int e^{i(\xi-p/2)q'+i(k+q/2)p'}\hat{R}(q,p,p')\delta(q+q')$$

$$\times \frac{W^{(0)}\left(t,x,k+\frac{q}{2}+\frac{q'}{2}\right)-W^{(0)}\left(t,x,k+\frac{q}{2}-\frac{q'}{2}\right)}{H\left(x,k+\frac{q}{2}+\frac{q'}{2}\right)-H\left(x,k+\frac{q}{2}-\frac{q'}{2}\right)+i\theta}\,dq'\,dp'$$

$$=e^{-i(\xi+p/2)q}\left(\int e^{i(k-q/2)p'}\hat{R}(q,p,p')dp'\right)\frac{W^{(0)}(t,x,k-q)-W^{(0)}(t,x,k)}{H(x,k-q)-H(x,k)+i\theta}-e^{-i(\xi-p/2)q)}$$

$$\times\left(\int e^{i(k+q/2)p'}\hat{R}(q,p,p')dp'\right)\frac{W^{(0)}(t,x,k)-W^{(0)}(t,x,k+q)}{H(x,k)-H(x,k+q)+i\theta}.$$

Therefore,

$$\left\langle \frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\hat{V}(q,p)\left[W^{(1)}\left(t,x,\xi+\frac{p}{2},k-\frac{q}{2}\right)-W^{(1)}\left(t,x,\xi-\frac{p}{2},k+\frac{q}{2}\right)\right]dq\,dp\right\rangle$$

$$=\int \mathcal{F}_1R\left(q,k-\frac{q}{2},k-\frac{q}{2}\right)\frac{W^{(0)}(t,x,k-q)-W^{(0)}(t,x,k)}{H(x,k-q)-H(x,k)+i\theta}\,dq$$

$$-\int \mathcal{F}_1R\left(q,k+\frac{q}{2},k+\frac{q}{2}\right)\frac{W^{(0)}(t,x,k)-W^{(0)}(t,x,k+q)}{H(x,k)-H(x,k+q)+i\theta}\,dq$$

$$=\int \mathcal{F}_1R\left(k-q,k+\frac{q}{2},k+\frac{q}{2}\right)\frac{W^{(0)}(t,x,q)-W^{(0)}(t,x,k)}{H(x,q)-H(x,k)+i\theta}\,dq$$

$$-\int \mathcal{F}_1R\left(k-q,k+\frac{q}{2},k+\frac{q}{2}\right)\frac{W^{(0)}(t,x,k)-W^{(0)}(t,x,q)}{H(x,k)-H(x,q)+i\theta}\,dq$$

$$=\int \mathcal{F}_1R\left(k-q,k+\frac{q}{2},k+\frac{q}{2}\right)\frac{2i\theta(W^{(0)}(t,x,q)-W^{(0)}(t,x,k))}{(H(x,q)-H(x,k))^2+\theta^2}\,dq.$$

Since

$$\frac{\theta}{x^2+\theta^2}\to\pi\delta(x),\quad \text{as}\quad \theta\to 0$$

we have, from (30), the following Proposition.

Proposition 5: $W^{(0)}$ satisfies the following equation:

$$\frac{\partial W^{(0)}}{\partial t}+\{W^{(0)},H\}=2\pi\int \mathcal{F}_1R\left(k-q,k+\frac{q}{2},k+\frac{q}{2}\right)(W^{(0)}(t,x,q)-W^{(0)}(t,x,k))\cdot\delta(H(x,q)-H(x,k))dq. \tag{32}$$

IV. THE SYSTEM OF EVOLUTION EQUATIONS

A. The generalized vector Schrödinger equations

In this section, we derive transport equations for a system of generalized Schrödinger equations.

Let $\mathbf{H}(x,k)=(H_{ij}(x,k))$ be a $n\times n$ matrix valued function. Assume that $H_{ij}\in C^\infty(\mathbf{R}_x^n\times\mathbf{R}_k^n)$. We always assume that for every $x,k,\mathbf{H}(x,k)$ is diagonalizable. Let $\mathbf{U}^\epsilon(t)=(u_1^\epsilon(t),\ldots,u_n^\epsilon(t))^t$ be the solution of the following system of equations:

$$i\epsilon\frac{\partial \mathbf{U}^\epsilon}{\partial t}=\mathbf{H}(X,\epsilon D)U^\epsilon$$

$$\mathbf{U}^\epsilon(0,x)=\varphi(x), \tag{33}$$

where $\mathbf{H}(X,D)=(H_{ij}(X,D))$ denotes the Weyl quantization of the energy symbol $\mathbf{H}(x,k)$.

Define the scaled Wigner distribution matrix as

$$\mathbf{W}^\epsilon(t,x,k)=(W_{ij}^\epsilon(t,x,k)), \tag{34}$$

where

$$W_{ij}^\varepsilon(t,x,k)=\frac{1}{\varepsilon^n}W(u_i^\varepsilon(t),u_j^\epsilon(t))\left(x,\frac{k}{\varepsilon}\right). \tag{35}$$

The matrix $\mathbf{W}^\epsilon$ is Hermitian as is easily seen by definition (11).

Proposition 6: $\mathbf{W}^\epsilon$ satisfies the following equation:

$$\begin{aligned} i\epsilon\frac{\partial \mathbf{W}^\varepsilon(t,x,k)}{\partial t}&=\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{\mathbf{H}}(q,p)\mathbf{W}^\epsilon\left(t,x+\frac{\varepsilon}{2}p,k-\frac{\varepsilon}{2}q\right)dq\,dp\\ &\quad-\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\mathbf{W}^\epsilon\left(t,x-\frac{\varepsilon}{2}p,k+\frac{\varepsilon}{2}q\right)\widehat{\mathbf{H}^*}(q,p)dq\,dp\\ \mathbf{W}^\varepsilon(0,x,k)&=\frac{1}{\varepsilon^n}\mathbf{W}(\varphi_i,\varphi_j)\left(x,\frac{k}{\varepsilon}\right), \end{aligned} \tag{36}$$

where $\widehat{\mathbf{H}^*}$ is the Fourier transform of Hermitian adjoint matrix of $\mathbf{H}^*$.

The proof of the proposition is similar to that of the scalar case.

Again, we consider the expansion

$$\mathbf{W}^\varepsilon(t,x,k)=\mathbf{W}^{(0)}(t,x,k)+\varepsilon\mathbf{W}^1(t,x,k)+\varepsilon^2\cdots. \tag{37}$$

Substitute (37) into (36) and collect the terms at different powers of ϵ, we have, for the coefficient of ε^{-1} term

$$\mathbf{H}(x,k)\cdot\mathbf{W}^{(0)}(t,x,k)-\mathbf{W}^{(0)}(t,x,k)\cdot\mathbf{H}^*(x,k)=0; \tag{38}$$

and for the coefficient of ε^0 term

$$\begin{aligned} i\frac{\partial \mathbf{W}^{(0)}(t,x,k)}{\partial t}&=\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}[\hat{H}(q,p)\cdot\mathbf{W}^{(1)}(t,x,k)-\mathbf{W}^{(1)}(t,x,k)\cdot\hat{H}^t(q,p)]dq\,dp\\ &\quad+\frac{1}{2(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\hat{H}(q,p)\cdot\left(p\frac{\partial}{\partial x}-q\frac{\partial}{\partial k}\right)\mathbf{W}^{(0)}(t,x,k)dq\,dp\\ &\quad-\frac{1}{2(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\left(-p\frac{\partial}{\partial x}+q\frac{\partial}{\partial k}\right)\mathbf{W}^{(0)}(t,x,k)\cdot\hat{H}^*(q,p)dq\,dp. \end{aligned} \tag{39}$$

Let $\mathbf{A}=(a_{ij}(x,k))$, $\mathbf{B}=(b_{ij}(x,k))$ be matrix valued functions and $f(x,k)$ be scalar function on $\mathbf{R}_x^n\times\mathbf{R}_k^n$. We define a matrix bracket $\{\mathbf{A}\otimes\mathbf{B}\}$ as

$$\{\mathbf{A}\otimes\mathbf{B}\}(x,k)=(c_{ij}(x,k)) \tag{40}$$

with

$$c_{ij}(x,k)=\sum_{l=1}^{n}\{a_{il},b_{lj}\}(x,k), \tag{41}$$

where $\{\cdot,\cdot\}$ is defined by (21). We also define a bracket of a matrix with a scalar function, $\{\mathbf{A},f\}$ as

$$\{\mathbf{A},f\}=(\{a_{ij},f\}).$$

We introduce several operators on $\mathrm{Mat}_n(C)\rightarrow\mathrm{Mat}_n(C)$ as follows:

$$\begin{aligned}
\mathcal{L}(x,k)Z&=\mathbf{H}(x,k)Z-Z\mathbf{H}^*(x,k)\\
\tilde{\mathcal{L}}(x,k)Z&=\mathbf{H}(x,k)Z\\
\mathcal{L}^*(x,k)Z&=\mathbf{H}^*(x,k)Z-Z\mathbf{H}(x,k)\\
\tilde{\mathcal{L}}^*(x,k)Z&=\mathbf{H}^*(x,k)Z.
\end{aligned} \tag{42}$$

For fixed x,k, the kernel of the operator $\mathcal{L}$ is invariant under $\tilde{\mathcal{L}}$. The operator $\tilde{\mathcal{L}}$ is diagonalizable as an operator acting on the vector space $\mathrm{Mat}_n(C)$ if $\mathbf{H}$ is a diagonalizable. That means that any invariant subspace of $\tilde{\mathcal{L}}$ is spanned by its eigenvectors (which are matrices), therefore $\ker\mathcal{L}$ is spanned by some of the eigenvectors of $\tilde{\mathcal{L}}$, namely, by those which belong to $\ker\mathcal{L}$. And each eigenvalue of $\tilde{\mathcal{L}}$ is also an eigenvalue of $\mathbf{H}$.

We will consider two separated cases, i.e., the case without polarization and the case with polarization.

1. Distinct eigenvalue case (without polarizations)

Suppose that the matrix $\mathbf{H}(x,k)$ has distinct real eigenvalues $\lambda_j(x,k)$, $j=1,\ldots,n$, which are C^1 function of x,k. The last assumption is not important but it is convenient. Let $b^{(j)}$ be the corresponding real eigenvectors which are column vectors

$$\mathbf{H}(x,k)b^{(j)}(x,k)=\lambda_j(x,k)b^{(j)}(x,y),\quad j=1,2,\ldots,n.$$

Denote

$$\mathbf{B}^{(j)}(x,k)=b^{(j)}(x,k)b^{(j)*}(x,k), \tag{43}$$

which are Hermitian matrices. Then, we have

$$\tilde{\mathcal{L}}(x,k)\mathbf{B}^{(j)}(x,k)=\lambda_j(x,k)\mathbf{B}^{(j)}(x,k)$$

and

$$\mathcal{L}(x,k)\mathbf{B}^{(j)}(x,k)=0.$$

Moreover $\mathbf{B}^{(j)}(x,k)$, $j=1,\ldots,n$, span the space $\ker\mathcal{L}(x,k)$. Suppose that $c^{(j)}(x,k)$, $j=1,\ldots,n$ are eigenvectors of matrix $\mathbf{H}^*(x,k)$ corresponding to eigenvalues $\lambda_j(x,k)$, $j=1,\ldots n$. $c^{(j)}$ and $b^{(j)}$ satisfy the orthogonal relation

$$c^{(i)*}(x,k)b^{(j)}(x,k)=f_i(x,k)\,\delta_{ij}.$$

Denote

$$\mathbf{C}^{(j)}(x,k)=c^{(j)}(x,k)c^{(j)*}(x,k),$$

we have

$$\tilde{\mathcal{L}}^*(x,k)\mathbf{C}^{(j)}(x,k)=\lambda_j(x,k)\mathbf{C}^{(j)}(x,k),$$

$$\mathcal{L}^*(x,k)\mathbf{C}^{(j)}(x,k)=0,$$

and $\mathbf{C}^{(j)}(x,k)$, $j=1,...,n$, span the space $\ker \mathcal{L}^*(x,k)$.

We introduce an inner product on the space $\mathrm{Mat}_n(C)$ as

$$\langle X,Y\rangle=\mathrm{tr}\, X^*Y.$$

Then $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$.

Using notations introduced above, (38) and (39) can be rewritten as

$$\mathcal{L}\mathbf{W}^{(0)}(t)=0, \tag{44}$$

$$i\frac{\partial \mathbf{W}^{(0)}(t)}{\partial t}-\frac{i}{2}\{\mathbf{H} \mathbin{\S} \mathbf{W}^{(0)}(t)\}+\frac{i}{2}\{\mathbf{W}^{(0)}(t) \mathbin{\S} \mathbf{H}^*\}=\mathcal{L}\mathbf{W}^{(1)}(t), \tag{45}$$

or

$$\mathbf{W}^{(0)}(t)\in \ker \mathcal{L}, \tag{46}$$

$$\frac{\partial \mathbf{W}^{(0)}(t)}{\partial t}-\frac{1}{2}\{\mathbf{H} \mathbin{\S} \mathbf{W}^{(0)}(t)\}+\frac{1}{2}\{\mathbf{W}^{(0)}(t) \mathbin{\S} \mathbf{H}^*\}\in \mathrm{Im}\, \mathcal{L}. \tag{47}$$

Since space $\ker \mathcal{L}$ is spanned by $\mathbf{B}^{(j)}=b^{(j)}b^{(j)*}$, $j=1,...,n$. We can write

$$\mathbf{W}^{(0)}(t,x,k)=\sum_{j=1}^{n} a^{(j)}(t,x,k)\mathbf{B}^{(j)}(x,k), \tag{48}$$

where the coefficients $a^{(j)}(t,x,k)$ are real valued functions. Then we have the following proposition from (44) and (45).

Proposition 7: There exists a normalized bases $\mathbf{B}^{(j)}$ of $\ker \mathcal{L}$ such that the coefficients $a^{(j)}\times(t,x,k)$ satisfy the following Liouville equations

$$\frac{\partial a^{(s)}(t)}{\partial t}+\{a^{(s)}(t),\lambda_s\}=0, \quad s=1,...,n. \tag{49}$$

Proof: Substitute (48) into the left-hand side of (45), we have

$$\begin{aligned}
&\frac{\partial \mathbf{W}^{(0)}(t)}{\partial t}-\frac{1}{2}\{\mathbf{H} \mathbin{\S} \mathbf{W}^{(0)}(t)\}+\frac{1}{2}\{\mathbf{W}^{(0)}(t) \mathbin{\S} \mathbf{H}^*\}\\
&\quad=\frac{\partial a^{(s)}}{\partial t}\mathbf{B}^{(s)}-\frac{1}{2}\{\mathbf{H} \mathbin{\S} a^{(s)}\mathbf{B}^{(s)}\}+\frac{1}{2}\{a^{(s)}\mathbf{B}^{(s)} \mathbin{\S} \mathbf{H}^*\}\\
&\quad=\frac{\partial a^{(s)}}{\partial t}\mathbf{B}^{(s)}-\frac{1}{2}a^{(s)}(\{\mathbf{H} \mathbin{\S} \mathbf{B}^{(s)}\}-\{\mathbf{B}^{(s)} \mathbin{\S} \mathbf{H}^*\})-\frac{1}{2}\{\mathbf{H},a^{(s)}\}\mathbf{B}^{(s)}+\frac{1}{2}\mathbf{B}^{(s)}\{a^{(s)},\mathbf{H}^*\}.
\end{aligned}$$

Since

$$\{\mathbf{H},a^{(s)}\}\mathbf{B}^{(s)}=\{\mathbf{H}\mathbf{B}^{(s)},a^{(s)}\}-\mathbf{H}\{\mathbf{B}^{(s)},a^{(s)}\}$$
$$=\{\lambda_s\mathbf{B}^{(s)},a^{(s)}\}-\mathbf{H}\{\mathbf{B}^{(s)},a^{(s)}\}$$
$$=\{\lambda_s,a^{(s)}\}\mathbf{B}^{(s)}+\lambda_s\{\mathbf{B}^{(s)},a^{(s)}\}-\mathbf{H}\{\mathbf{B}^{(s)},a^{(s)}\}$$

and

$$\mathbf{B}^{(s)}\{a^{(s)},\mathbf{H}^*\}=\{a^{(s)},\mathbf{B}^{(s)}\mathbf{H}^*\}-\{a^{(s)},\mathbf{B}^{(s)}\}\mathbf{H}^*=-\{\lambda_s,a^{(s)}\}\mathbf{B}^{(s)}+\{a^{(s)},\mathbf{B}^{(s)}\}(\lambda_s-\mathbf{H}^*).$$

Therefore

$$\frac{\partial\mathbf{W}^{(0)}(t)}{\partial t}-\frac{1}{2}\{\mathbf{H}\otimes_s\mathbf{W}^{(0)}(t)\}+\frac{1}{2}\{\mathbf{W}^{(0)}(t)\otimes_s\mathbf{H}^*\}$$
$$=\left(\frac{\partial a^{(s)}}{\partial t}-\{\lambda_s,a^{(s)}\}\right)\mathbf{B}^{(s)}-\frac{1}{2}a^{(s)}\{\mathbf{H}\otimes_s\mathbf{B}^{(s)}\}+\frac{1}{2}a^{(s)}\{\mathbf{B}^{(s)}\otimes_s\mathbf{H}^*\}$$
$$-\frac{1}{2}(\lambda_s-\mathbf{H})\{\mathbf{B}^{(s)},a^{(s)}\}+\frac{1}{2}\{a^{(s)},\mathbf{B}^{(s)}\}(\lambda_s-\mathbf{H}^*)\in\operatorname{Im}\mathcal{L},\tag{50}$$

where the right-hand side is the summation over s. Using the fact that $\mathbf{C}^{(p)}$ are orthogonal to $\operatorname{Im}\mathcal{L}$ and

$$\langle\mathbf{C}^{(p)},\mathbf{B}^{(q)}\rangle=f_p\overline{f_p}\delta_{pq},$$

where $f_p=(c^{(p)},b^{(p)})=c^{(p)*}b^{(p)}$. Taking inner product of (50) with $\mathbf{C}^{(p)}$, we have

$$\left(\frac{\partial a^{(s)}}{\partial t}-\{\lambda_s,a^{(s)}\}\right)\delta_{ps}|f_p|^2-\frac{1}{2}a^{(s)}\langle\mathbf{C}^{(p)},\{\mathbf{H}\otimes_s\mathbf{B}^{(s)}\}-\{\mathbf{B}^{(s)}\otimes_s\mathbf{H}^*\}\rangle$$
$$-\frac{1}{2}(\lambda_s-\lambda_p)\langle\mathbf{C}^{(p)},\{\mathbf{B}^{(s)},a^{(s)}\}\rangle+\frac{1}{2}(\lambda_s-\lambda_p)\langle\mathbf{C}^{(p)},\{a^{(s)},\mathbf{B}^{(s)}\}\rangle=0.\tag{51}$$

When $p\neq s$, we have

$$\langle\mathbf{C}^{(p)},\{\mathbf{B}^{(s)},a^{(s)}\}\rangle=\left\langle\mathbf{C}^{(p)},\frac{\partial\mathbf{B}^{(s)}}{\partial x}\right\rangle\frac{\partial a^{(s)}}{\partial k}-\left\langle\mathbf{C}^{(p)},\frac{\partial\mathbf{B}^{(s)}}{\partial k}\right\rangle\frac{\partial a^{(s)}}{\partial x},$$

$$\left\langle\mathbf{C}^{(p)},\frac{\partial\mathbf{B}^{(s)}}{\partial x}\right\rangle=(b^{(s)},c^{(p)})\left(c^{(p)},\frac{\partial b^{(s)}}{\partial x}\right)+\left(\frac{\partial b^{(s)}}{\partial x},c^{p}\right)(c^{(p)},b^{(s)})=0,$$

where a_s are real functions. Similarly,

$$\left\langle\mathbf{C}^{(p)},\frac{\partial\mathbf{B}^{(s)}}{\partial k}\right\rangle=0.$$

It follows that, when $p\neq s$

$$(\lambda_p-\lambda_s)\langle\mathbf{C}^{(p)},\{\mathbf{B}^{(s)},a^{(s)}\}\rangle=0.$$

Similar calculation shows that, when $p\neq s$

$$(\lambda_s-\lambda_p)\langle\mathbf{C}^{(p)},\{a^{(s)},\mathbf{B}^{(s)}\}\rangle=0.$$

Next, we show that for $p\neq s$

$$\langle \mathbf{C}^{(p)}, \{\mathbf{H} \otimes \mathbf{B}^{(s)}\} - \{\mathbf{B}^{(s)} \otimes \mathbf{H}^*\} \rangle = 0. \tag{52}$$

Since $\mathbf{B}^{(s)}\mathbf{C}^{(p)} = 0$, we have

$$\begin{aligned}
0 &= \mathrm{tr}\{\mathbf{H} \otimes \mathbf{B}^{(s)}\mathbf{C}^{(p)}\} \\
&= \mathrm{tr}\{\mathbf{H} \otimes \mathbf{B}^{(s)}\}\mathbf{C}^{(p)} - \mathrm{tr}\{\mathbf{C}^{(p)} \otimes \mathbf{H}\}\mathbf{B}^{(s)} \\
&= \mathrm{tr}\{\mathbf{C}^{(p)}\mathbf{H} \otimes \mathbf{B}^{(s)}\} + \mathrm{tr}\{\mathbf{B}^{(s)} \otimes \mathbf{C}^{(p)}\}\mathbf{H} - (\mathrm{tr}\{\mathbf{C}^{(p)} \otimes \mathbf{H}\mathbf{B}^{(s)}\} + \mathrm{tr}\, \mathbf{H}\{\mathbf{B}^{(s)} \otimes \mathbf{C}^{(p)}\}) \\
&= \mathrm{tr}\{\lambda_p \mathbf{C}^{(p)} \otimes \mathbf{B}^{(s)}\} - \mathrm{tr}\{\mathbf{C}^{(p)} \otimes \lambda_s \mathbf{B}^{(s)}\} \\
&= (\lambda_p - \lambda_s)\mathrm{tr}\{\mathbf{C}^{(p)} \otimes \mathbf{B}^{(s)}\},
\end{aligned}$$

therefore, when $p \neq s$,

$$\mathrm{tr}\{\mathbf{C}^p \otimes \mathbf{B}^{(s)}\} = 0.$$

It follows that

$$\begin{aligned}
\langle \mathbf{C}^{(p)}, \{\mathbf{H} \otimes \mathbf{B}^{(s)}\} \rangle &= \mathrm{tr}\, \mathbf{C}^{(p)}\{\mathbf{H} \otimes \mathbf{B}^{(s)}\} = \mathrm{tr}\{\mathbf{C}^{(p)}\mathbf{H} \otimes \mathbf{B}^{(s)}\} + \mathrm{tr}\{\mathbf{B}^{(s)} \otimes \mathbf{C}^{(p)}\}\mathbf{H} = \mathrm{tr}\{\lambda_p \mathbf{C}^{(p)} \otimes \mathbf{B}^{(s)}\} \\
&\quad + \mathrm{tr}\{\mathbf{B}^{(s)} \otimes \mathbf{C}^{(p)}\}\mathbf{H} = \mathrm{tr}\{\mathbf{B}^{(s)} \otimes \mathbf{C}^{(p)}\}\mathbf{H} = \sum_{m,n,l} \{b_m^{(s)}\overline{b_n^{(s)}}, c_n^{(p)}\overline{c_l^{(p)}}\}\mathbf{H}_{lm} \\
&= \sum_{m,n,l} \{b_m^{(s)}c_n^{(p)}\}\overline{b_n^{(s)}}\,\overline{c_l^{(p)}}\mathbf{H}_{lm} + \sum_{m,n,l} \{\overline{b_n^{(s)}}, \overline{c_l^{(p)}}\}b_m^{(s)}c_n^{(p)}\mathbf{H}_{lm} \\
&\quad + \sum_{m,n,l} \{\overline{b_n^{(s)}}, \overline{c_n^{(p)}}\}\overline{c_l^{(p)}}\mathbf{H}_{lm}b_m^{(s)} = \lambda_p \sum_{m,n,l} \{b_m^{(s)}, c_n^{(p)}\}\overline{b_n^{(s)}}\,\overline{c_m^{(p)}} \\
&\quad + \lambda_s \sum_{l,n} \{\overline{b_n^{(s)}}, \overline{c_l^{(p)}}\}b_l^{(s)}c_n^{(p)}.
\end{aligned}$$

Similarly

$$\langle \mathbf{C}^{(p)}, \{\mathbf{B}^{(s)} \otimes \mathbf{H}^*\} \rangle = \mathrm{tr}\, \mathbf{H}^*\{\mathbf{C}^{(p)} \otimes \mathbf{B}^{(s)}\} = \lambda_p \sum_{nl} \{\overline{c_l^{(p)}}, \overline{b_n^{(s)}}\}b_l^{(s)}c_n^{(p)} + \lambda_s \sum_{mn} \{c_n^{(p)}, b_m^{(s)}\}\overline{b_n^{(s)}}\,\overline{c_m^{(p)}}.$$

Therefore

$$\begin{aligned}
\langle \mathbf{C}^{(p)}, \{\mathbf{H} \otimes \mathbf{B}^{(s)}\} - \{\mathbf{B}^{(s)} \otimes \mathbf{H}^*\} \rangle &= (\lambda_p + \lambda_s)\left(\sum_{mn} \{b_m^{(s)}, c_n^{(p)}\}\overline{b_n^{(s)}}\,\overline{c_m^{(p)}} + \sum_{mn} \{\overline{b_n^{(s)}}\,\overline{c_n^{(p)}}\}b_m^{(s)}c_n^{(p)} \right) \\
&= (\lambda_p + \lambda_s)\mathrm{tr}\{\mathbf{B}^{(s)} \otimes \mathbf{C}^{(p)}\} = 0.
\end{aligned}$$

It follows that (51) becomes

$$\left(\frac{\partial a^{(s)}}{\partial t} - \{\lambda_s, a^{(s)}\} \right)|f_s|^2 - \frac{1}{2}a^{(s)}\langle \mathbf{C}^{(s)}, \{\mathbf{H} \otimes \mathbf{B}^{(s)}\} - \{\mathbf{B}^{(s)} \otimes \mathbf{H}^*\} \rangle = 0.$$

To show that (49) holds, we prove that $b^{(s)}$ can be normalized so that

$$\langle \mathbf{C}^{(s)}, \{\mathbf{H} \otimes \mathbf{B}^{(s)}\} - \{\mathbf{B}^{(s)} \otimes \mathbf{H}^*\} \rangle = 0. \tag{53}$$

To achieve that, let $\tilde{\mathbf{B}}^{(s)} = g(x,k)\mathbf{B}^{(s)}$ where $\mathbf{B}^{(s)}$ is fixed. We will show that there exists a function g such that (53) holds for $\tilde{\mathbf{B}}^{(s)}$:

$$\langle \mathbf{C}^{(s)},\{\mathbf{H}\otimes\tilde{\mathbf{B}}^{(s)}\}-\{\tilde{\mathbf{B}}^{(s)}\otimes\mathbf{H}^*\}\rangle = g\langle \mathbf{C}^{(s)},\{\mathbf{H}\otimes\mathbf{B}^{(s)}\}-\{\mathbf{B}^{(s)}\otimes\mathbf{H}^*\}\rangle + 2\langle \mathbf{C}^{(s)},\mathbf{B}^{(s)}\rangle\{\lambda_s,g\}.$$

Then Eq. (53) becomes

$$2\langle \mathbf{C}^{(s)},\mathbf{B}^{(s)}\rangle\{\lambda_s,g\} + g\langle \mathbf{C}^{(s)},\{\mathbf{H}\otimes\mathbf{B}^{(s)}\}-\{\mathbf{B}^{(s)}\otimes\mathbf{H}^*\}\rangle = 0. \tag{54}$$

This is a first order PDE for g and has always a solution which is nonzero everywhere provided λ_s is not a constant. Therefore (51) reduces to (49).

2. Multiple eigenvalue case (with polarization)

We turn our attention to the case with polarization. We assume that Hamiltonian $\mathbf{H}(x,k)$ is diagonalizable but the eigenvalues are not distinct. We assume throughout that the eigenvalues of $\mathbf{H}(x,k)$ have constant multiplicity independent of x and k. This hypothesis is satisfied in the hyperbolic system of acoustic, electromagnetic and elastic waves. Let $\lambda_s(x,k)$ be a real eigenvalue of $H(x,k)$ of multiplicity r. Denote

$$V_{\lambda_s}(x,k) = \{v \in C^n: \quad H(x,k)v = \lambda_s(x,k)v\}$$

and choose $b^{(s1)},\ldots,b^{(sr)}$ to be an orthogonal basis of V_{λ_s}. Let

$$\mathbf{V}_{\lambda_s}(x,k) = \{Z \in \mathrm{Mat}_n(C): \tilde{\mathcal{L}}(x,k)Z = \lambda_s(x,k)Z, \quad \mathcal{L}(x,k)Z = 0\},$$

then $\dim \mathbf{V}_{\lambda_s}(x,k) = r^2$ and it has a basis $\mathbf{B}_{ij}^{(s)} = b^{(si)}b^{(sj)*}$, $i,j = 1,\ldots,r$. Again, from (44), we assume that

$$\mathbf{W}^{(0)}(t,x,k) = \sum_{s,i,j} a_{i,j}^{(s)}(t,x,k)\mathbf{B}_{i,j}^{(s)}(x,k), \tag{55}$$

where $a_{i,j}^{(s)}(t,x,k)$ are scalar functions. Define the $r\times r$ coherence matrices $\mathbf{A}^{(s)}(t,x,k)$ by

$$\mathbf{A}_{i,j}^{(s)}(t,x,k) = a_{i,j}^{(s)}(t,x,k), \quad i,j = 1,\ldots,r.$$

The multiplicity r of the eigenvalue λ_s depends on s but we do not indicate this explicitly.

Proposition 8: The coherence matrices $\mathbf{A}^{(s)}(t,x,k)$ satisfy the following Liouville type transport equation:

$$\frac{\partial \mathbf{A}^{(s)}(t)}{\partial t} + \{\mathbf{A}^{(s)}(t),\lambda_s\} + \mathbf{A}^{(s)}(t)M^{(s)} - M^{(s)*}\mathbf{A}^{(s)}(t) = 0,$$

where $M^{(s)} = (M_{lm}^{(s)}(x,k))$ and

$$M_{lm}^{(s)}(x,k) = \frac{1}{2}\sum_{\alpha} \frac{C_\alpha^{(sm)}}{\overline{f_m^{(s)}}}\left(\sum_r \{\overline{b_r^{(sl)}},\overline{H}_{\alpha r}\} - \{\lambda_p,\overline{b_\alpha^{(sl)}}\}\right).$$

Proof: The eigenspace of $H^*(x,k)$ corresponding to eigenvalue $\lambda_s(x,k)$ is also r dimensional. Let $c^{(s1)},\ldots,c^{(sr)}$ be the basis of this space which is dual to the basis $b^{(s1)},\ldots,b^{(sr)}$ of $V_{\lambda_s}(x,k)$. Denote $\mathbf{C}_{ij}^{(s)} = c^{(si)}\cdot c^{(sj)*}$ and let $(c^{(sl)},b^{(sm)}) = \delta_{lm}f_m^{(s)}$ then

$$\langle C_{lm}^{(p)},B_{ij}^{(s)}\rangle = \delta_{sp}\delta_{li}\delta_{mj}f_l^{(s)}\overline{f_m^{(s)}}.$$

Substitute (55) into (45), we have

$$\frac{\partial \mathbf{W}^{(0)}(t)}{\partial t}-\frac{1}{2}\{\mathbf{H} \otimes \mathbf{W}^{(0)}(t)\}+\frac{1}{2}\{\mathbf{W}^{(0)}(t) \otimes \mathbf{H}^*\}$$
$$=\sum_{s,i,j}\left(\frac{\partial a_{ij}^{(s)}}{\partial t}-\{\lambda_s, a_{ij}^{(s)}\}\right)\mathbf{B}_{ij}^{(s)}-\frac{1}{2}a_{ij}^{(s)}(\{\mathbf{H}\otimes \mathbf{B}_{ij}^{(s)}\}-\{\mathbf{B}_{ij}^{(s)}\otimes \mathbf{H}^*\})$$
$$-\frac{1}{2}(\lambda_s-\mathbf{H})\{\mathbf{B}_{ij}^{(s)}, a_{ij}^{(s)}\}+\frac{1}{2}\{a_{ij}^{(s)}, \mathbf{B}_{ij}^{(s)}\}(\lambda_s-\mathbf{H}^*)\in \operatorname{Im}\mathcal{L}. \tag{56}$$

Taking inner product of (56) with $\mathbf{C}_{lm}^{(p)}$, we have

$$\left(\frac{\partial a_{lm}^{(p)}}{\partial t}-\{\lambda_p, a_{lm}^{(p)}\}\right) f_l^{(p)}\overline{f_m^{(p)}}-\frac{1}{2}\sum_{s,i,j} a_{ij}^{(s)}\langle \mathbf{C}_{lm}^{(p)}, \{\mathbf{H}\otimes \mathbf{B}_{ij}^{(s)}\}-\{\mathbf{B}_{ij}^{(s)}\otimes \mathbf{H}^*\}\rangle$$
$$-\frac{1}{2}(\lambda_s-\lambda_p)\sum_{s,i,j}\langle \mathbf{C}_{lm}^{(p)}, \{\mathbf{B}_{ij}^{(s)}, a_{ij}^{(s)}\}\rangle+\frac{1}{2}(\lambda_s-\lambda_p)\sum_{s,i,j}\langle \mathbf{C}_{lm}^{(p)}, \{a_{ij}^{(s)}, \mathbf{B}_{ij}^{(s)}\}\rangle=0. \tag{57}$$

Similar calculations as in the case of distinct eigenvalue show that the third and the fourth term of the left-hand side of (57) disappear. In the second term, only terms with $s=p$ remains. Therefore we have

$$\left(\frac{\partial a_{lm}^{(p)}}{\partial t}-\{\lambda_p, a_{lm}^{(p)}\}\right) f_l^{(p)}\overline{f_m^{(p)}}-\frac{1}{2}\sum_{p,i,j} a_{ij}^{(p)}\langle \mathbf{C}_{lm}^{(p)}, \{\mathbf{H}\otimes \mathbf{B}_{ij}^{(s)}\}-\{\mathbf{B}_{ij}^{(p)}\otimes \mathbf{H}^*\}\rangle=0, \tag{58}$$

where

$$\langle \mathbf{C}_{lm}^{(p)}, \{\mathbf{H}\otimes \mathbf{B}_{ij}^{(s)}\}-\{\mathbf{B}_{ij}^{(p)}\otimes \mathbf{H}^*\}\rangle=\sum_{\alpha\beta\gamma} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}(\{\mathbf{H}_{\beta\gamma}, b_\gamma^{(pi)}\overline{b_\alpha^{(pj)}}\}-\{b_\beta^{(pi)}\overline{b_\gamma^{(pj)}}, \overline{H}_{\alpha\gamma}\})=\Delta_1+\Delta_2,$$

with

$$\Delta_1=\sum_{\beta,\gamma}\overline{c_\beta^{(pl)}}\{\mathbf{H}_{\beta\gamma}, b_\gamma^{(pi)}\}\delta_{mj}\overline{f_m^{(p)}}-\sum_{\alpha,\gamma} c_\alpha^{(pm)}\{\overline{b_\gamma^{(pi)}}, \overline{H}_{\alpha\gamma}\}\delta_{li} f_l^{(p)}, \tag{59}$$

$$\Delta_2=\sum_{\alpha,\beta,\gamma} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}} b_\gamma^{(pi)}\{\mathbf{H}_{\beta\gamma}, \overline{b_\alpha^{(pj)}}\}$$
$$-\sum_{\alpha,\beta,\gamma} c_\alpha^{(pm)}\overline{c_\beta^{(pl)} b_\gamma^{(pj)}}\{b_\beta^{(pi)}, \overline{\mathbf{H}}_{\alpha\gamma}\}$$
$$=\sum_{\alpha,\beta,\gamma} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}\{\mathbf{H}_{\beta\gamma} b_\gamma^{(pi)}, \overline{b_\alpha^{(pj)}}\}$$
$$-\sum_{\alpha,\beta,\gamma} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}\{b_\beta^{(pi)}, \overline{\mathbf{H}}_{\alpha\gamma}\overline{b_\gamma^{(pj)}}\}-\sum_{\alpha,\beta,\gamma} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}\mathbf{H}_{\beta\gamma}\{b_\gamma^{(pi)}, \overline{b_\alpha^{(pj)}}\}$$
$$+\sum_{\alpha,\beta,\gamma} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}\overline{\mathbf{H}}_{\alpha\gamma}\{b_\beta^{(pi)}, \overline{b_\gamma^{(pj)}}\}$$
$$=\sum_{\alpha,\beta} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}\{\lambda_p b_\beta^{(pi)}, \overline{b_\alpha^{(pj)}}\}-\sum_{\alpha,\beta} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}\{b_\beta^{(pi)}, \lambda_p\overline{b_\alpha^{(pj)}}\}$$
$$-\sum_{\alpha,\gamma}\lambda_p c_\alpha^{(pm)}\overline{c_\gamma^{(pl)}}\{b_\gamma^{(pi)}, \overline{b_\alpha^{(pj)}}\}+\sum_{\beta,\gamma}\lambda_p c_\gamma^{(pm)}\overline{c_\beta^{(pl)}}\{b_\beta^{(pi)}, \overline{b_\gamma^{(pj)}}\}$$

$$=\sum_{\alpha,\beta} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}}b_\beta^{(pi)}\{\lambda_p,\overline{b_\beta^{(pj)}}\}-\sum_{\alpha,\beta} c_\alpha^{(pm)}\overline{c_\beta^{(pl)}b_\alpha^{(pj)}}\{b_\beta^{(pi)},\lambda_p\}$$

$$=\sum_\alpha c_p^{(pm)}\{\lambda_p,\overline{b_\alpha^{(pj)}}\}\delta_{il}f_l^{(p)}-\sum_\beta \overline{c_\beta^{(pl)}}\{b_\beta^{(pi)},\lambda_p\}\delta_{mj}\overline{f_m^{(p)}}.$$

Therefore

$$-\frac{1}{2}\sum_{i,j} a_{i,j}^{(p)}\langle \mathbf{C}_{lm}^{(p)},\{\mathbf{H}\otimes\mathbf{B}_{ij}^{(p)}\}-\{\mathbf{B}_{ij}^{(p)}\otimes\mathbf{H}^*\}\rangle$$

$$=\frac{1}{2}\sum_j a_{lj}^{(p)}\left(\sum_{\alpha,\gamma} c_\alpha^{(pm)}\{\overline{b_\gamma^{(pj)}},\overline{\mathbf{H}}_{\alpha\gamma}\}-\sum_\alpha c_\alpha^{(pm)}\{\lambda_p,\overline{b_\alpha^{(pj)}}\}\right)f_l^{(p)}$$

$$-\frac{1}{2}\sum_i\left(\sum_{\beta,\gamma}\overline{c_\beta^{(pl)}}\{\mathbf{H}_{\beta\gamma},b_\gamma^{(pi)}\}-\sum_\beta\overline{c_\beta^{(pl)}}\{b_\beta^{(pi)},\lambda_p\}\right)a_{im}^{(p)}\overline{f_m^{(p)}}$$

$$=((A^{(p)}M)_{lm}-(M^*\cdot A^{(p)})_{lm})f_l^{(p)}\overline{f_m^{(p)}},$$

where

$$M^{(p)}=(M_{lm}^{(p)}),$$

$$M_{lm}^{(p)}=\frac{1}{2}\sum_\alpha \frac{c_\alpha^{(pm)}}{\overline{f_m^{(p)}}}\left(\sum_\gamma\{\overline{b_\gamma^{(pl)}},\overline{H}_{\alpha\gamma}\}-\{\lambda_p,\overline{b_\alpha^{(pl)}}\}\right).$$

Thus we have

$$\frac{\partial A^{(p)}}{\partial t}-\{\lambda_p,A^{(p)}\}+A^{(p)}M^{(p)}-M^{(p)*}A^{(p)}=0 \tag{60}$$

and the proof is complete.

B. The vector Schrödinger equations with small random perturbations

Let $\mathbf{V}(x,k)$ be a real matrix valued random function in phase space with mean zero. Suppose that it is stationary in x. We denote the correlation functions by

$$R_{\alpha\beta\lambda\delta}(y-x,k,k')=\langle V_{\alpha\beta}(x,k)V_{\lambda\delta}(y,k')\rangle, \tag{61}$$

which are assumed to be homogeneous in space. It is easy to see that

$$R_{\alpha\beta\lambda\delta}(x,k,k')=R_{\lambda\delta\alpha\beta}(-x,k,k').$$

Denote $\mathcal{F}_1R_{\alpha\beta\lambda\delta}(q,k,k')$ as the Fourier transform of $R_{\alpha\beta\lambda\delta}(x,k,k')$ in x, we have

$$\langle\mathcal{F}_1V_{\alpha\beta}(q,k)\mathcal{F}_1V_{\lambda\delta}(q',k')\rangle=\mathcal{F}_1R_{\alpha\beta\lambda\delta}(q,k,k')\delta(q+q'). \tag{62}$$

We also assume that for any α, β, λ, δ, and all k,k'

$$\operatorname{Im}\mathcal{F}_1R_{\alpha\beta\lambda\delta}(q,k,k')=0,$$

which is equivalent to say that the correlation functions is even in x. Then we have the following symmetry

$$R_{\alpha\beta\lambda\delta}(x,k,k')=R_{\lambda\delta\alpha\beta}(x,k,k')$$

and

$$\mathcal{F}_1 R_{\alpha\beta\lambda\delta}(q,k,k')=\mathcal{F}_1 R_{\lambda\delta\alpha\beta}(q,k,k'). \tag{63}$$

Consider the scaled Wigner distribution matrix

$$\mathbf{W}^{\varepsilon}(t,x,k)=\frac{1}{\varepsilon^n}\mathbf{W}(\mathbf{U}^{\varepsilon}(t),\mathbf{U}^{\varepsilon}(t))\left(x,\frac{k}{\varepsilon}\right), \tag{64}$$

where $\mathbf{U}^{\epsilon}(t)$ be the solution of the following system of equations

$$i\frac{\partial \mathbf{U}^{\epsilon}(t,x)}{\partial t}=\left(\mathbf{H}(x,\epsilon D)+\sqrt{\epsilon}\mathbf{V}\left(\frac{1}{\epsilon}x,\epsilon D\right)\right)\mathbf{U}^{\epsilon}(t,x)$$

$$\mathbf{U}^{\epsilon}(0,x)=\varphi(x). \tag{65}$$

Let $\xi=x/\epsilon$ be the fast space variable and rewrite $\mathbf{W}^{\epsilon}(t,x,k)=\mathbf{W}^{\epsilon}(t,x,x/\epsilon,k)$ by $\mathbf{W}^{\epsilon}(t,x,\xi,k)$, we obtain the equation

$$\begin{aligned}
i\frac{\partial \mathbf{W}^{\varepsilon}(t,x,\xi,k)}{\partial t}=&\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\frac{1}{\epsilon}\left[\hat{\mathbf{H}}(q,p)\mathbf{W}^{\epsilon}\left(t,x+\frac{\varepsilon}{2}p,\xi+\frac{p}{2},k-\frac{\varepsilon}{2}q\right)\right.\\
&\left.-\mathbf{W}^{\epsilon}\left(t,x-\frac{\varepsilon}{2}p,\xi-\frac{p}{2},k+\frac{\varepsilon}{2}q\right)\widehat{\mathbf{H}^*}(q,p)\right]dq\,dp\\
&+\frac{1}{(2\pi)^n}\int e^{i(\xi\cdot q+k\cdot p)}\frac{1}{\sqrt{\epsilon}}\left[\hat{\mathbf{V}}(q,p)\mathbf{W}^{\epsilon}\left(t,x+\frac{\varepsilon}{2}p,\xi+\frac{p}{2},k-\frac{q}{2}\right)\right.\\
&\left.-\mathbf{W}^{\epsilon}\left(t,x-\frac{\varepsilon}{2}p,\xi-\frac{p}{2},k+\frac{q}{2}\right)\hat{\mathbf{V}}^t(q,p)\right]dq\,dp,
\end{aligned} \tag{66}$$

where $\widehat{\mathbf{H}^*}$ is the Fourier transform of $\mathbf{H}^*=\overline{\mathbf{H}}^t$. Consider the formal expansion

$$\mathbf{W}^{\varepsilon}(t,x,\xi,k)=\mathbf{W}^{(0)}(t,x,k)+\sqrt{\epsilon}\mathbf{W}^{(1)}(t,x,\xi,k)+\epsilon\mathbf{W}^{(2)}(t,x,\xi,k)+\cdots.$$

We assume that the leading term does not depend on the fast variable and it is deterministic. We have, as before, at order ϵ^{-1},

$$\mathbf{W}^{(0)}(t,x,k)\in\ker\mathcal{L},$$

where $\mathcal{L}$ is defined by (42). At order $\epsilon^{-1/2}$,

$$\begin{aligned}
&\frac{1}{(2\pi)^n}\int e^{i(x\cdot q+k\cdot p)}\left[\hat{\mathbf{H}}(q,p)\mathbf{W}^{(1)}\left(t,x,\xi+\frac{p}{2},k\right)-\mathbf{W}^{(1)}\left(t,x,\xi-\frac{p}{2},k\right)\widehat{\mathbf{H}^*}(q,p)\right]dq\,dp\\
&\quad+\frac{1}{(2\pi)^n}\int e^{i(\xi\cdot q+k\cdot p)}\left[\hat{\mathbf{V}}(q,p)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right)-\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\hat{\mathbf{V}}^t(q,p)\right]dq\,dp=0.
\end{aligned} \tag{67}$$

At order ϵ^0, we have

$$\begin{aligned}
&\frac{\partial \mathbf{W}^{(0)}(t)}{\partial t}-\frac{i}{2}\{\mathbf{H}\otimes \mathbf{W}^{(0)}(t)\}+\frac{i}{2}\{\mathbf{W}^{(0)}(t)\otimes \mathbf{H}^*\}\\
&\quad=\frac{1}{(2\pi)^n}\int e^{i(xq+kp)}\left[\hat{\mathbf{H}}(q,p)\mathbf{W}^{(2)}\left(t,x,\xi+\frac{p}{2},k\right)-\mathbf{W}^{(2)}\left(t,x,\xi-\frac{p}{2},k\right)\hat{\mathbf{H}}^*(q,p)\right]dq\,dp\\
&\qquad+\frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\left[\hat{\mathbf{V}}(q,p)\mathbf{W}^{(1)}\left(t,x,\xi+\frac{p}{2},k-\frac{q}{2}\right)\right.\\
&\qquad\left.-\mathbf{W}^{(1)}\left(t,x,\xi-\frac{p}{2},k+\frac{q}{2}\right)\hat{\mathbf{V}}^t(q,p)\right]dq\,dp. \qquad (68)
\end{aligned}$$

Similarly, from ergodicity, we have

$$\left\langle \frac{\partial \mathbf{W}^{(2)}(t,x,\xi,k)}{\partial \xi}\right\rangle=0.$$

Averaging (68), we get

$$\begin{aligned}
&i\left(\frac{\partial \mathbf{W}^{(0)}(t)}{\partial t}-\frac{1}{2}\{\mathbf{H}\otimes \mathbf{W}^{(0)}(t)\}+\frac{1}{2}\{\mathbf{W}^{(0)}(t)\otimes \mathbf{H}^*\}\right)\\
&\quad-\frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\left\langle \hat{\mathbf{V}}(q,p)\mathbf{W}^{(1)}\left(t,x,\xi+\frac{p}{2},k-\frac{q}{2}\right)\right\rangle dq\,dp\\
&\quad+\frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\left\langle \mathbf{W}^{(1)}\left(t,x,\xi-\frac{p}{2},k+\frac{q}{2}\right)\mathbf{V}^t(q,p)\right\rangle dq\,dp\\
&\quad=\mathcal{L}(x,k)\langle \mathbf{W}^{(2)}(t,x,\xi,k)\rangle\in \mathrm{Im}\,\mathcal{L}(x,k). \qquad (69)
\end{aligned}$$

Again, we want to express $\mathbf{W}^{(1)}(t)$ in terms of $\mathbf{W}^{(0)}(t)$, then insert it into the above equation and take the inner product with the generators of $\ker \mathcal{L}^*$. Denote $\mathbf{F}(t,x,q,k)$ as the Fourier transform of $\mathbf{W}^{(1)}(t,x,\xi,k)$ in fast space variable ξ, i.e.,

$$\mathbf{W}^{(1)}(t,x,\xi,k)=\frac{1}{(2\pi)^{n/2}}\int e^{i\xi q'}\mathcal{F}(t,x,q',k)dq',$$

we have from (67)

$$\begin{aligned}
&\frac{1}{(2\pi)^{n/2}}\int e^{i(x\cdot q+k\cdot p)+i(\xi+p/2)q'}\hat{\mathbf{H}}(q,p)\mathbf{F}(t,x,q',k)dq'\,dq\,dp\\
&\quad-\frac{1}{(2\pi)^{n/2}}\int e^{i(x\cdot q+k\cdot p)+i(\xi-p/2)q'}\mathbf{F}(t,x,q',k)\hat{\mathbf{H}}^*(q,p)dq'\,dq\,dp\\
&\quad+\int e^{i(\xi\cdot q+k\cdot p)}\left[\hat{\mathbf{V}}(q,p)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right)-\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\hat{\mathbf{V}}^t(q,p)\right]dq\,dp=0,
\end{aligned}$$

$$\int e^{i\xi\cdot q'}\mathbf{H}\left(x,k+\frac{q'}{2}\right)\mathbf{F}(t,x,q',k)dq'-\int e^{i\xi\cdot q'}\mathbf{F}(t,x,q',k)\mathbf{H}^*\left(x,k-\frac{q'}{2}\right)dq'$$
$$+\frac{1}{(2\pi)^{n/2}}\int e^{i(\xi\cdot q+k\cdot p)}\hat{\mathbf{V}}(q,p)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right)dq\,dp$$
$$-\frac{1}{(2\pi)^{n/2}}\int e^{i(\xi\cdot q+k\cdot p)}\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\hat{\mathbf{V}}^t(q,p)dq\,dp=0.$$

Take Fourier inverse transform in ξ, we have

$$\mathbf{H}\left(x,k+\frac{q}{2}\right)\mathbf{F}(t,x,q,k)-\mathbf{F}(t,x,q,k)\mathbf{H}^*\left(x,k-\frac{q}{2}\right)$$
$$+\mathcal{F}_1\mathbf{V}(q,k)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right)-\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}^t(q,k)=0.$$

1. Distinct eigenvalue case

Since $\{\mathbf{B}^{(j)}(x,k),j=1,...,n\}$ [defined as in (43)] spans the space $\ker\mathcal{L}(x,k)$, we write

$$\mathbf{W}^{(0)}(t,x,k)=a_s(t,x,k)\mathbf{B}^{(s)}(x,k).$$

In this case, we define

$$\mathbf{E}^{(s,t)}(x,q,k)=b^{(s)}\left(x,k+\frac{q}{2}\right)b^{(t)*}\left(x,k-\frac{q}{2}\right),$$

and

$$\mathbf{D}^{(s,t)}(x,q,k)=c^{(s)}\left(x,k+\frac{q}{2}\right)c^{(t)*}\left(x,k-\frac{q}{2}\right).$$

The matrices $\mathbf{E}^{(s,t)}$ and $\mathbf{D}^{(s,t)}$ satisfy the bi-orthogonal relations

$$\langle\mathbf{D}^{(r,s)},\mathbf{E}^{(m,n)}\rangle=\delta_{m,r}\delta_{n,s}f_m\left(x,k+\frac{q}{2}\right)\overline{f_n}\left(x,k-\frac{q}{2}\right),$$

and for every fixed x,q,k, the set $\{\mathbf{E}^{(s,t)}(x,q,k),1\leqslant s,t\leqslant n\}$ spans the matrix space $\mathrm{Mat}_n(C)$. Thus let

$$\mathbf{F}(t,x,q,k)=\sum_{i,j}F^{(i,j)}(t,x,q,k)\mathbf{E}^{(i,j)}(x,q,k).$$

We have

$$F^{(i,j)}(t,x,q,k)\left(\mathbf{H}\left(x,k+\frac{q}{2}\right)\mathbf{E}^{(i,j)}(x,q,k)-\mathbf{E}^{(i,j)}(x,q,k)\mathbf{H}^*\left(x,k-\frac{q}{2}\right)\right)$$
$$=\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}^t(q,k)-\mathcal{F}_1\mathbf{V}(q,k)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right),$$

$$F^{(i,j)}(t,x,q,k)\left(\lambda_i\left(x,k+\frac{q}{2}\right)-\lambda_j\left(x,k-\frac{q}{2}\right)\right)\mathbf{E}^{(i,j)}(x,q,k)$$
$$=\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}^t(q,k)-\mathcal{F}_1\mathbf{V}(q,k)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right).$$

Multiplying by $\mathbf{D}^{(s,r)}(x,q,k)$, we obtain

$$F^{(s,r)}(t,x,q,k)\left(\lambda_s\left(x,k+\frac{q}{2}\right)-\lambda_r\left(x,k-\frac{q}{2}\right)\right)f_s\left(x,k+\frac{q}{2}\right)\overline{f_r}\left(x,k-\frac{q}{2}\right)$$
$$=\left\langle\mathbf{D}^{(s,r)}(x,q,k),\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}^t(q,k)-\mathcal{F}_1\mathbf{V}(q,k)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right)\right\rangle$$

and

$$\mathbf{F}(t,x,q,k))=\frac{\left\langle\mathbf{D}^{(s,r)}(x,q,k),\mathbf{W}^{(0)}\left(t,x,k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}^t(q,k)-\mathcal{F}_1\mathbf{V}(q,k)\mathbf{W}^{(0)}\left(t,x,k-\frac{q}{2}\right)\right\rangle}{\left(\lambda_s\left(x,k+\frac{q}{2}\right)-\lambda_r\left(x,k-\frac{q}{2}\right)+i\theta\right)f_s\left(x,k+\frac{q}{2}\right)\overline{f_r}\left(x,k-\frac{q}{2}\right)}$$
$$\times\mathbf{E}^{(s,r)}(x,q,k),$$

where the term $i\theta$ is a regularization term, eventually we will let $\theta\to 0$. Insert into the expansion

$$\mathbf{W}^{(0)}(t,x,k)=a_s(t,x,k)\mathbf{B}^{(s)}(x,k),$$

we get

$$F^{(s,r)}(t,x,q,k)=\frac{1}{\lambda_r\left(x,k-\frac{q}{2}\right)-\lambda_s\left(x,k+\frac{q}{2}\right)+i\theta}$$
$$\cdot\left\{\frac{a_r\left(t,x,k-\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}_{\alpha\beta}(q,k)b_\beta^{(r)}\left(x,k-\frac{q}{2}\right)\overline{c_\alpha^{(s)}}\left(x,k+\frac{q}{2}\right)}{f_s\left(x,k+\frac{q}{2}\right)}\right.$$
$$\left.-\frac{a_s\left(t,x,k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}_{\beta\alpha}^t(q,k)c_\alpha^{(r)}\left(x,k-\frac{q}{2}\right)\overline{b_\beta^{(s)}}\left(x,k+\frac{q}{2}\right)}{\overline{f_r}\left(x,k-\frac{q}{2}\right)}\right\}.$$

Let

$$I_1=\frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\left\langle\hat{\mathbf{V}}(q,p)\mathbf{W}^{(1)}\left(t,x,\xi+\frac{p}{2},k-\frac{q}{2}\right)\right\rangle dq\,dp,$$

$$I_2=\frac{1}{(2\pi)^n}\int e^{i(\xi q+kp)}\left\langle\mathbf{W}^{(1)}\left(t,x,\xi-\frac{p}{2},k+\frac{q}{2}\right)\hat{\mathbf{V}}^t(q,p)\right\rangle dq\,dp.$$

From Eq. (69), we have

$$\left\langle \mathbf{C}^{(r)}(x,k), i\left(\frac{\partial \mathbf{W}^{(0)}(t,x,k)}{\partial t} - \frac{1}{2}\{\mathbf{H}^* \circledS \mathbf{W}^{(0)}(t)\}(x,k) + \frac{1}{2}\{\mathbf{W}^{(0)}(t) \circledS \mathbf{H}^*\} \right) \right\rangle$$
$$= \langle \mathbf{C}^{(r)}(x,k), I_1\rangle - \langle \mathbf{C}^{(r)}(x,k), I_2\rangle. \tag{70}$$

From the last subsection, we know that for appropriate normalized eigenvectors $b^{(m)}(x,k)$, the left-hand side of (70) is

$$i\left(\frac{\partial a_s(t,x,k)}{\partial t} - \{\lambda_s, a_s\}(x,k) \right) \delta_{rs} |f_r(x,k)|^2.$$

Thus, we only need to calculate $\langle \mathbf{C}^{(r)}, I_1\rangle$ and $\langle \mathbf{C}^{(r)}, I_2\rangle$

$$I_1 = \frac{1}{(2\pi)^{3n/2}} \int e^{i(\xi q + kp) + i(\xi + p/2)q'} \left\langle \hat{\mathbf{V}}(q,p) \mathbf{F}\left(t,x,q',k - \frac{q}{2} \right) \right\rangle dq'\, dq\, dp$$
$$= \frac{1}{(2\pi)^n} \int e^{i\xi(q+q')} \sum_{l,m} \left\langle F^{(l,m)}\left(t,x,q',k-\frac{q}{2}\right) \mathcal{F}_1 \mathbf{V}\left(q, k + \frac{q'}{2}\right) \right\rangle \mathbf{E}^{(l,m)}\left(x,q',k-\frac{q}{2}\right) dq'\, dq.$$

$\langle \mathbf{C}^{(r)}, I_1\rangle$

$$= \frac{1}{(2\pi)^n} \int e^{i\xi(q+q')} \sum_{l,m} \mathbf{C}^{(r)}_{\alpha\beta}(x,k)$$
$$\cdot \left\langle F^{(l,m)}\left(t,x,q',k-\frac{q}{2}\right) \mathcal{F}_1 \mathbf{V}_{\beta\delta}\left(q,k+\frac{q'}{2}\right) \right\rangle \mathbf{E}^{(l,m)}_{\delta\alpha}\left(x,q',k-\frac{q}{2}\right) dq'\, dq$$
$$= \frac{1}{(2\pi)^n} \sum_{l,m} \int e^{i\xi(q+q')} \frac{\mathbf{C}^{(r)}_{\alpha\beta}(x,k) \mathbf{E}^{(l,m)}_{\delta\alpha}\left(x,q',k-\frac{q}{2}\right) \delta(q+q')}{\lambda_m\left(x,k-\frac{q}{2}-\frac{q'}{2}\right) - \lambda_l\left(x,k-\frac{q}{2}+\frac{q'}{2}\right) + i\theta}$$
$$\cdot \left\{ \frac{a_m\left(t,x,k-\frac{q}{2}-\frac{q'}{2}\right) \mathcal{F}_1 \mathbf{R}_{\alpha'\beta'\beta\delta}\left(q,k-\frac{q}{2},k+\frac{q'}{2}\right) b^{(m)}_{\beta'}\left(x,k-\frac{q}{2}-\frac{q'}{2}\right) \overline{c^{(l)}_{\alpha'}}\left(x,k-\frac{q}{2}+\frac{q'}{2}\right)}{f_l\left(x,k-\frac{q}{2}+\frac{q'}{2}\right)} \right.$$
$$\left. - \frac{a_l\left(t,x,k-\frac{q}{2}+\frac{q'}{2}\right) \mathcal{F}_1 \mathbf{R}_{\beta'\alpha'\beta\delta}\left(q,k-\frac{q}{2},k+\frac{q}{2}\right) c^{(m)}_{\alpha'}\left(x,k-\frac{q}{2}-\frac{q'}{2}\right) \overline{b^{(l)}_{\beta'}}\left(t,k-\frac{q}{2}+\frac{q'}{2}\right)}{\overline{f_m}\left(x,k-\frac{q}{2}-\frac{q'}{2}\right)} \right\}$$
$$\times dq\, dq'$$
$$= \sum_{lm} \int \frac{c^{(r)}_{\alpha}(x,k) \overline{c^{(r)}_{\beta}}(x,k) b^{(l)}_{\delta}(x,k-q) \overline{b^{(m)}_{\alpha}}(x,k)}{\lambda_m(x,k) - \lambda_l(x,k-q) + i\theta}$$

$$\cdot\left\{\frac{a_m(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q,k-\frac{q}{2},k-\frac{q}{2}\right)b_{\beta}^{(m)}(x,k)\overline{c_{\alpha'}^{(l)}}(x,k-q)}{f_l(x,k-q)}\right.$$

$$\left.-\frac{a_l(t,x,k-q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q,k-\frac{q}{2},k-\frac{q}{2}\right)c_{\alpha'}^{(m)}(x,k)\overline{b_{\beta'}^{(l)}}(t,k-q)}{\overline{f_m}(x,k)}\right\}dq$$

$$=\sum_l\int\frac{\overline{f_r}(x,k)\overline{c_{\beta}^{(r)}}(x,k)b_{\delta}^{(l)}(x,q)}{\lambda_r(x,k)-\lambda_l(x,q)+i\theta}$$

$$\cdot\left\{\frac{a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)b_{\beta'}^{(r)}(x,k)\overline{c_{\alpha'}^{(l)}}(x,q)}{f_l(x,q)}\right.$$

$$\left.-\frac{a_l(t,x,q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{q+k}{2},\frac{q+k}{2}\right)c_{\alpha'}^{(r)}(x,k)\overline{b_{\beta'}^{(l)}}(x,q)}{\overline{f_r}(x,k)}\right\}dq$$

$$=\sum_l\int\frac{a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_r(x,k)-\lambda_l(x,q)+i\theta}\frac{\overline{f_r}(x,k)}{f_l(x,q)}\overline{c_{\alpha'}^{(l)}}(x,q)b_{\beta'}^{(r)}(x,k)\overline{c_{\beta}^{(r)}}(x,k)b_{\delta}^{(l)}$$

$$\times(x,q)dq-\sum_l\int\frac{a_l(t,x,q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{q+k}{2},\frac{q+k}{2}\right)}{\lambda_r(x,k)-\lambda_l(x,q)+i\theta}$$

$$\times c_{\alpha'}^{(r)}(x,k)\overline{b_{\beta'}^{(l)}}(x,q)\overline{c_{\beta}^{(r)}}(x,k)b_{\delta}^{(l)}(x,q)dq.$$

Similarly

$$I_2=\frac{1}{(2\pi)^{3n/2}}\int e^{i(\xi q+kp)+i(\xi-p/2)q'}\left\langle\mathbf{F}\left(t,x,q',k+\frac{q}{2}\right)\hat{\mathbf{V}}^t(q,p)\right\rangle dq'\,dq\,dp$$

$$=\frac{1}{(2\pi)^n}\int e^{i\xi(q+q')}\sum_{l,m}\left\langle F^{(l,m)}\left(t,x,q',k+\frac{q}{2}\right)\mathbf{E}^{(l,m)}\left(x,q',k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}^t\left(q,k-\frac{q'}{2}\right)\right\rangle dq'\,dq,$$

$$\langle\mathbf{C}^{(r)},I_2\rangle$$

$$=\frac{1}{(2\pi)^n}\int e^{i\xi(q+q')}\sum_{l,m}\mathbf{C}_{\alpha\beta}^{(r)}(x,k)\mathbf{E}_{\beta\delta}^{(l,m)}\left(x,q',k+\frac{q}{2}\right)$$

$$\cdot\left\langle F^{(l,m)}\left(t,x,q',k+\frac{q}{2}\right)\mathcal{F}_1\mathbf{V}_{\delta\alpha}^t\left(q,k-\frac{q'}{2}\right)\right\rangle dq'\,dq$$

$$=\frac{1}{(2\pi)^n}\sum_{l,m}\int e^{i\xi(q+q')}\frac{\mathbf{C}_{\alpha\beta}^{(r)}(x,k)\mathbf{E}_{\beta\delta}^{(l,m)}\left(x,q',k+\frac{q}{2}\right)\delta(q+q')}{\lambda_m\left(x,k+\frac{q}{2}-\frac{q'}{2}\right)-\lambda_l\left(x,k+\frac{q}{2}+\frac{q'}{2}\right)+i\theta}$$

$$\cdot\left\{\frac{a_m\left(t,x,k+\frac{q}{2}-\frac{q'}{2}\right)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\delta}\left(q,k+\frac{q}{2},k-\frac{q'}{2}\right)b_{\beta'}^{(m)}\left(x,k+\frac{q}{2}-\frac{q'}{2}\right)\overline{c_{\alpha'}^{(l)}}\left(x,k+\frac{q}{2}+\frac{q'}{2}\right)}{f_l\left(x,k+\frac{q}{2}+\frac{q'}{2}\right)}\right.$$

$$\left.-\frac{a_l\left(t,x,k+\frac{q}{2}+\frac{q'}{2}\right)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\delta}\left(q,k+\frac{q}{2},k-\frac{q'}{2}\right)c_{\alpha}^{(m)}\left(x,k+\frac{q}{2}-\frac{q'}{2}\right)\overline{b_{\beta'}^{(l)}}\left(x,k+\frac{q}{2}+\frac{q'}{2}\right)}{\overline{f_m}\left(x,k+\frac{q}{2}-\frac{q'}{2}\right)}\right\}$$

$$\times dq'\,dq$$

$$=\sum_{l,m}\int\frac{c_{\alpha}^{(r)}(x,k)\overline{c_{\beta}^{(r)}}(x,k)b_{\beta}^{(l)}(x,k)\overline{b_{\delta}^{(m)}}(t,k+q)}{\lambda_m(x,k+q)-\lambda_l(x,k)+i\theta}$$

$$\cdot\left\{\frac{a_m(t,x,k+q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\delta}\left(q,k+\frac{q}{2},k+\frac{q}{2}\right)b_{\beta'}^{(m)}(x,k+q)\overline{c_{\alpha'}^{(l)}}(x,k)}{f_l(x,k)}\right.$$

$$\left.-\frac{a_l(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\delta}\left(q,k+\frac{q}{2},k+\frac{q}{2}\right)c_{\alpha'}^{(m)}(x,k+q)\overline{b_{\beta'}^{(l)}}(x,k)}{\overline{f_m}(x,k+q)}\right\}dq$$

$$=\sum_{m}\int\frac{f_r(x,k)c_{\alpha}^{(r)}(x,k)\overline{b^{(m)}}_{\delta}(x,q)}{\lambda_m(x,q)-\lambda_r(x,k)+i\theta}$$

$$\cdot\left\{\frac{a_m(t,x,q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)b_{\beta'}^{(m)}(x,q)\overline{c_{\alpha'}^{(r)}}(x,k)}{f_r(x,k)}\right.$$

$$\left.-\frac{a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)c_{\alpha'}^{(r)}(x,q)\overline{b_{\beta'}^{(r)}}(x,k)}{\overline{f_m}(x,q)}\right\}dq$$

$$=\sum_{l}\int\frac{f_r(x,k)c_{\beta}^{(r)}(x,k)\overline{b_{\delta}^{(l)}}(x,q)}{\lambda_r(x,k)-\lambda_l(x,q)-i\theta}$$

$$\cdot\left\{\frac{a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)c_{\alpha'}^{(l)}(x,q)\overline{b_{\beta'}^{(r)}}(x,k)}{\overline{f_l}(x,q)}\right.$$

$$\left.-\frac{a_l(t,x,q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)\overline{c_{\alpha'}^{(r)}}(x,k)b_{\beta'}^{(l)}(x,q)}{f_r(x,k)}\right\}dq$$

$$=\sum_l \int \frac{a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_r(x,k)-\lambda_l(x,q)-i\theta}\frac{f_r(x,k)}{\overline{f_l}(x,q)}c_{\alpha'}^{(l)}(x,q)\overline{b_{\beta'}^{(r)}}(x,k)c_\beta^{(r)}(x,k)\overline{b_\delta^{(l)}}$$

$$\times(x,q)dq-\sum_l \int \frac{a_l(t,x,q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_r(x,k)-\lambda_l(x,q)-i\theta}$$

$$\times c_{\alpha'}^{(r)}(x,k)\overline{b_{\beta'}^{(l)}}(x,q)\overline{c_\beta^{(r)}}(x,k)b_\delta^{(l)}(x,q)dq,$$

where we have used the symmetric property (63) for $\mathcal{F}_1R$ in the last equality. Therefore,

$$\langle\mathbf{C}^{(r)},I_1\rangle-\langle\mathbf{C}^{(r)},I_2\rangle=\sum_l \int a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)\mathrm{Re}\left[\frac{f_r(x,k)}{\overline{f_l}(x,q)}c_{\alpha'}^{(l)}(x,q)\overline{b_{\beta'}^{(r)}}\right.$$

$$\left.\times(x,k)c_\beta^{(r)}(x,k)\overline{b_\delta^{(l)}}(x,k)\right]\frac{2i\theta}{(\lambda_r(x,k)-\lambda_l(x,q))^2+\theta^2}dq$$

$$-i\sum_l \int a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)\mathrm{Im}\left[\frac{f_r(x,k)}{\overline{f_l}(x,q)}c_{\alpha'}^{(l)}\right.$$

$$\left.\times(x,q)\overline{b_{\beta'}^{(r)}}(x,k)c_\beta^{(r)}(x,k)\overline{b_\delta^{(l)}}(x,k)\right]\frac{2(\lambda_r(x,k)-\lambda_l(x,q))}{(\lambda_r(x,k)-\lambda_l(x,q))^2+\theta^2}dq$$

$$-\sum_l \int a_l(t,x,q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)c_{\alpha'}^{(r)}(x,k)\overline{b_{\beta'}^{(l)}}(x,q)\overline{c_\beta^{(r)}}$$

$$\times(x,k)b_\delta^{(l)}(x,k)\frac{2i\theta}{(\lambda_r(x,k)-\lambda_l(x,q))^2+\theta^2}dq,$$

as $\theta\to 0,\ \theta/y^2+\pi^2\to\pi\delta(y),$ we have

$$\langle\mathbf{C}^{(r)},I_1\rangle-\langle\mathbf{C}^{(r)},I_2\rangle\to 2\pi i\sum_l \int a_r(t,x,k)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)$$

$$\cdot\mathrm{Re}\left[\frac{f_r(x,k)}{\overline{f_l}(x,q)}c_{\alpha'}^{(l)}(x,q)\overline{b_{\beta'}^{(r)}}(x,k)c_\beta^{(r)}(x,k)\overline{b_\delta^{(l)}}(x,k)\right]\delta(\lambda_r(x,k)$$

$$-\lambda_l(x,q))dq-2i\sum_l \int a_r(t,x,k)\frac{\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_r(x,k)-\lambda_l(x,q)}$$

$$\cdot\mathrm{Im}\left[\frac{f_r(x,k)}{\overline{f_l}(x,q)}c_{\alpha'}^{(l)}(x,q)\overline{b_{\beta'}^{(r)}}(x,k)c_\beta^{(r)}(x,k)\overline{b_\delta^{(l)}}(x,k)\right]dq$$

$$-2\pi i\sum_l \int a_l(t,x,q)\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)\cdot c_{\alpha'}^{(r)}(x,k)\overline{b_{\beta'}^{(l)}}$$

$$\times(x,q)\overline{c_{\beta}^{(r)}}(x,k)b_{\delta}^{(l)}(x,q)\,\delta(\lambda_r(x,k)-\lambda_l(x,k))dq,$$

where the integration over terms involving $1/(\lambda_r(x,k)-\lambda_l(x,q))$ is in certain principal value sense when singularity occurs. Let

$$\sigma_{rl}(x,k,q)=2\pi\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)\frac{c_{\alpha'}^{(r)}(x,k)\overline{b_{\beta'}^{(l)}}(x,q)\overline{c_{\beta}^{(r)}}(x,k)b_{\delta}^{(l)}(x,q)}{|f_r(x,k)|^2}$$
$$\cdot\,\delta(\lambda_l(x,q)-\lambda_r(x,k))$$

$$\Sigma_r=2\pi\int\sum_l\ \mathrm{Re}\left[\frac{c_{\alpha'}^{(l)}(x,q)\overline{b_{\beta'}^{(r)}}(x,k)c_{\beta}^{(r)}(x,k)\overline{b_{\delta}^{(l)}}(x,q)}{\overline{f_l}(x,q)\overline{f_r}(x,k)}\right]$$
$$\cdot\,\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)\delta(\lambda_l(x,q)-\lambda_r(x,k))dq$$

$$-2\int\sum_l\ \mathrm{Im}\left[\frac{c_{\alpha'}^{(l)}(x,q)\overline{b_{\beta'}^{(r)}}(x,k)c_{\beta}^{(r)}(x,k)\overline{b_{\delta}^{(l)}}(x,q)}{\overline{f_l}(x,q)\overline{f_r}(x,k)}\right]\frac{\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\beta\delta}\left(q-k,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_l(x,q)-\lambda_r(x,k)}dq,$$

then we have the following proposition:

Proposition 9: The transport equations for the coefficients a_r are

$$\frac{\partial a_r(t,x,k)}{\partial t}+\{a_r(t),\lambda_r\}(x,k)=\int\sigma_{rj}(x,k,q)a_j(t,x,q)dq-\Sigma_r a_r(t,x,k). \tag{71}$$

2. Multiple eigenvalue case

Now we turn to the multiple eigenvalue case. In this case, we define the matrices

$$E_{l,m}^{(s,r)}(x,q,k)=b^{(sl)}\left(x,k+\frac{q}{2}\right)b^{(rm)*}\left(x,k-\frac{q}{2}\right)$$

$$D_{l,m}^{(s,r)}(x,q,k)=c^{(sl)}\left(x,k+\frac{q}{2}\right)c^{(rm)*}\left(x,k-\frac{q}{2}\right).$$

The matrices $E_{l,m}^{(s,r)}$ and $D_{l,m}^{(s,r)}$ satisfy the following double bi-orthogonal relations

$$\langle D_{l,m}^{(s,r)},E_{p,t}^{(s,r)}\rangle=\delta_{sp}\delta_{rt}\delta_{li}\delta_{mj}f_l^{(s)}\left(x,k+\frac{q}{2}\right)\overline{f_m^{(r)}}\left(x,k-\frac{q}{2}\right),$$

where

$$(c^{(sl)}(x,k),b^{(sl)}(x,k))=f_l^{(s)}(x,k).$$

For every fixed x,q,k, the set of all $E_{l,m}^{(s,r)}(x,q,k)$ spans the space $\mathrm{Mat}_n(\mathbf{C})$. This allows us to express $\mathbf{F}$ as $\mathbf{F}=\Sigma F_{l,m}^{(s,r)}E_{l,m}^{(s,r)}$ where $\mathbf{F}$ is the Fourier transform of $W^{(1)}$ in fast variable ξ. Similar to the calculations in the case of distinct eigenvalue, we have from (67) and the expansion

$$W^{(0)}=\sum a_{ij}^{s}\mathbf{B}_{ij}^{s}. \tag{72}$$

The coefficients of $\mathbf{F}$ in the basis of $\{E_{l,m}^{(s,r)}\}$ are

$$F_{l,m}^{(s,r)}(t,q,k)=\frac{1}{\lambda_s\left(k+\frac{q}{2}\right)-\lambda_r\left(k-\frac{q}{2}\right)+i\theta}$$

$$\times\left\{\sum_j \frac{\alpha_{lj}^{(s)}\left(t,k+\frac{q}{2}\right)}{\overline{f_m^{(r)}}\left(k-\frac{q}{2}\right)}\mathcal{F}_1 V_{\alpha\beta}(q,k)c_\alpha^{(rm)}\left(k-\frac{q}{2}\right)\overline{b_\beta^{(sj)}}\left(k+\frac{q}{2}\right)\right.$$

$$\left.-\sum_j \frac{\alpha_{jm}^{(r)}\left(t,k-\frac{q}{2}\right)}{f_l^{(s)}\left(k+\frac{q}{2}\right)}\mathcal{F}_1 V_{\alpha\beta}(q,k)\overline{c_r^{(sl)}}\left(k+\frac{q}{2}\right)b_\beta^{(rj)}\left(k-\frac{q}{2}\right)\right\}. \tag{73}$$

From (69), we have

$$\left\langle \mathbf{C}_{lm}^{(p)},i\left(\frac{\partial \mathbf{W}^{(0)}(t)}{\partial t}-\frac{1}{2}\{\mathbf{H}\otimes \mathbf{W}^{(0)}(t)\}+\frac{1}{2}\{\mathbf{W}^{(0)}(t)\otimes \mathbf{H}^*\}\right)\right\rangle=\langle \mathbf{C}_{lm}^{(p)},I_1\rangle-\langle \mathbf{C}_{lm}^{(p)},I_2\rangle. \tag{74}$$

From the last section, we know that the left-hand side of the above equation is

$$i\left(\frac{\partial A^{(p)}}{\partial t}-\{\lambda_p,A^{(p)}\}+A^{(p)}M^{(p)}-M^{(p)*}A^{(p)}\right)f_l^{(p)}\overline{f_m^{(p)}}. \tag{75}$$

Thus, we only need to calculate $\langle \mathbf{C}_{lm}^{(p)},I_1\rangle$ and $\langle \mathbf{C}_{lm}^{(p)},I_2\rangle$. After some tedious calculations, we have

$$\langle \mathbf{C}_{lm}^{(p)},I_1\rangle=\int\sum_{s,j}\frac{\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\beta}\left(q,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_s(q)-\lambda_p(k)+i\theta}\left(\sum_h a_{jh}^{(s)}(q)c_{\alpha'}^{(pm)}(k)\overline{b_{\beta'}^{(sh)}}(q)\overline{c_\alpha^{(pl)}}(k)b_\beta^{(sj)}(q)\right.$$

$$\left.-\sum_n a_{n,m}^{(p)}(k)\frac{\overline{f_m^{(p)}}(k)}{f_j^{(s)}(q)}\overline{c_{\alpha'}^{(sj)}}(q)b_{\beta'}^{(pn)}(k)\overline{c_\alpha^{(pl)}}(k)b_\beta^{(sj)}(q)\right)dq \tag{76}$$

and

$$\langle \mathbf{C}_{lm}^{(p)},I_2\rangle=\int\sum_{s,j}\frac{\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\beta}\left(q,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_s(q)-\lambda_p(k)-i\theta}\left(\sum_h a_{jh}^{(s)}(q)c_{\alpha'}^{(pm)}(k)\overline{b_{\beta'}^{(sh)}}(q)\overline{c_\alpha^{(pl)}}(k)b_\beta^{(sj)}(q)\right.$$

$$\left.-\sum_n a_{l,n}^{(p)}(k)\frac{f_l^{(p)}(k)}{\overline{f_j^{(s)}}(q)}c_{\alpha'}^{(sj)}(q)\overline{b_{\beta'}^{(pn)}}(k)c_\alpha^{(pm)}(k)\overline{b_\beta^{(sj)}}(q)\right)dq. \tag{77}$$

On the right-hand side of the above equations, the sum over s runs all eigenvalues λ_s,j runs from 1 to multiplicity of λ_s,n runs from 1 to multiplicity of λ_p. From (74), (75), (76), and (77), we have

Proposition 10: The coherence matrices $A^{(p)}(t,x,k)$ satisfy the following Liouville type transport equation:

$$\frac{\partial A^{(p)}(t)}{\partial t}+\{A^{(p)}(t),\lambda_p\}+A^{(p)}(t)M^{(p)}-M^{(p)*}A^{(p)}(t)$$
$$=2\pi\int\sum_s \sigma^{(p,s)}(A^{(s)}(t))dq+i(\Sigma^{(p)}A^{(p)}-A^{(p)}\Sigma^{(p)})-\Lambda^{(p)}\,\mathrm{diag}\,A^{(p)}, \tag{78}$$

where

$$\sigma^{(p,s)}(A^{(s)}(t))=\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\beta}\left(q,\frac{k+q}{2},\frac{k+q}{2}\right)\sigma^{(p)}_{\alpha\alpha'}(k)\langle\sigma^{(s)}_{\beta\beta'}(q),A^{(s)}(t,q)\rangle,$$

in which the matrices are

$$\sigma^{(s)}_{\beta\beta'}(q)=(\sigma^{(s)}_{\beta\beta',jh}(q))\quad\text{with}\quad\sigma^{(s)}_{\beta\beta',jh}(q)=\overline{b^{(sj)}_{\beta}}(q)b^{(sh)}_{\beta'}(q),$$

$$\sigma^{(p)}_{\alpha\alpha'}(k)=(\sigma^{(p)}_{\alpha\alpha',lm}(k))\quad\text{with}\quad\sigma^{(p)}_{\alpha\alpha',lm}(k)=\frac{c^{(pm)}_{\alpha'}(k)\overline{c^{(pl)}_{\alpha}}(k)}{f^{(p)}_l(k)\overline{f^{(p)}_m}(k)},$$

and where

$$\Sigma^{(p)}(k)=(\Sigma^{(p)}_{lm}(k))\quad\text{with}\quad\Sigma^{(p)}_{lm}(k)$$
$$=\int\sum_{s,j}\frac{\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\beta}\left(q,\frac{k+q}{2},\frac{k+q}{2}\right)}{\lambda_s(q)-\lambda_p(k)}\,\frac{\overline{c^{(sj)}_{\alpha'}}(q)b^{(pm)}_{\beta'}(k)\overline{c^{(pl)}_{\alpha}}(k)b^{(sj)}_{\beta}(q)}{f^{(p)}_l(k)f^{(s)}_j(q)}\,dq,$$

$$\Lambda^{(p)}(k)=\mathrm{diag}(\tau^{(p)}_1,\ldots,\tau^{(p)}_r)\quad\text{with}\quad\tau^{(p)}_l(k)=2\pi\int\Sigma_{s,j}\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\beta}\left(q,\frac{k+q}{2},\frac{k+q}{2}\right)$$
$$\cdot\,\mathrm{Re}\left(\frac{\overline{c^{(sj)}_{\alpha'}}(q)b^{(pm)}_{\beta'}(k)\overline{c^{(pl)}_{\alpha}}(k)b^{(sj)}_{\beta}(q)}{f^{(p)}_l(k)f^{(s)}_j(q)}\right)\delta(\lambda_s(q)-\lambda_p(k))dq,$$

$$\mathrm{diag}\,A^{(p)}=\mathrm{diag}(a^{(p)}_{11},\ldots,a^{(p)}_{rr}).$$

In particular,

$$\sigma^{(p,s)}=\mathcal{F}_1\mathbf{R}_{\alpha'\beta'\alpha\beta}\left(q,\frac{k+q}{2},\frac{k+q}{2}\right)\sigma^{(p)}_{\alpha\alpha'}(k)\otimes\sigma^{(s)}_{\beta\beta'}(q),$$

where r is the multiplicity of λ_p.

ACKNOWLEDGMENTS

We thank G. Papanicolaou and L. Ryzhik for some useful discussions. This work is supported in part through the Research Grant Council of Hong Kong by Grant No. HKUST 586/94P. The work of M. Z. Guo is also supported in part by the Ph.D. station foundation of the State Education Committee.

[1] L. Ryzhik, G. Papanicolaou, and J. Keller, ''Transport equation for elastic and other waves in random media,'' Wave Motion **24**, 327–370 (1996).
[2] S. Chandrasekhar, Radiative Transfer (Dover, New York, 1960).
[3] I. M. Besieris, W. Kohler, and H. Freese, ''A transport-theoretic analysis of pulse propagation through ocean sediments,'' J. Acoust. Soc. Am. **72**, 937–946 (1982).

[4] R. S. Wu, ''Multiple scattering and energy transfer of seismic wave-separation of scattering effect from intrinsic attenuation—I. Theoretical modeling,'' Geophys. J. R. Astron. Soc. **82**, 57–80 (1985).
[5] E. Wigner, ''On the quantum correction for thermodynamic equilibrium,'' Phys. Rev. **40**, 749–759 (1932).
[6] P. L. Lions and T. Paul, ''Sur les Measures de Wigner,'' Revista Mat. Iberoamericana **9**, 553–618 (1993).
[7] G. B. Folland, *Harmonic Analysis in Phase Space* (Princeton University Press, Princeton, NJ, 1989).
[8] P. Gerard, P. Markowich, N. Mauser, and F. Poupaud, ''Homogenization limits and Wigner transforms,'' Commun. Pure Appl. Math. **50**, 323–380 (1996).
[9] L. Erdos and H. T. Yau, Linear Boltzmann equation as scaling limit of quantum Lorentz gas, preprint.

Time harmonic electromagnetic scattering from a bounded obstacle: An existence theorem and a computational method

Lorella Fatone
Dipartimento di Matematica "F Enriques," Università di Milano, Via Saldini 50, 20133 Milano, Italy

Cristina Pignotti
Dipartimento di Matematica, Università di Roma "Tor Vergata," Viale della Ricerca Scientifica, 00133 Roma, Italy

Maria Cristina Recchioni[a]
Istituto di Matematica e Statistica, Università di Ancona, Piazzale Martelli 8, 60100 Ancona, Italy

Francesco Zirilli[b]
Dipartimento di Matematica "G. Castelnuovo," Università di Roma "La Sapienza," Piazzale Aldo Moro 5, 00185 Roma, Italy

(Received 22 January 1998; accepted for publication 2 April 1999)

Let $\Omega\subset\mathbb{R}^3$ be a bounded simply connected obstacle with boundary $\partial\Omega$ locally Lipschitz, we consider the scattering of a time harmonic electromagnetic wave that hits Ω when $\partial\Omega$ is assumed to be perfectly conducting. The scattered electromagnetic field is the solution of an exterior boundary value problem for the vector Helmholtz equation. Under suitable hypotheses we prove the existence and uniqueness of the solution of this boundary value problem and we give a new numerical method to compute this solution. The numerical method proposed is based on a perturbative series and is highly parallelizable. Some numerical results obtained with the numerical method proposed on test problems are presented and discussed from the numerical and the physical point of view. © *1999 American Institute of Physics.* [S0022-2488(99)03409-X]

I. INTRODUCTION

Let $\mathbb{R}$ be the set of real numbers, $\mathbb{R}^3$ be the three dimensional real Euclidean space and $\underline{x}=(x_1,x_2,x_3)^T\in\mathbb{R}^3$ be a generic vector, where the superscript T denotes the transposition operation. For $\underline{x},\underline{y},\underline{z}\in\mathbb{R}^3$ we denote with $(\underline{x},\underline{y})$ the Euclidean scalar product of $\underline{x}$ and $\underline{y}$, with $\|\underline{x}\|$ the Euclidean vector norm of $\underline{x}$, with $[\underline{x},\underline{y}]$ the Euclidean vector product of $\underline{x}$ and $\underline{y}$, and with $(\underline{x},\underline{y},\underline{z})$ the usual triple product, that is $(\underline{x},\underline{y},\underline{z})=([\underline{x},\underline{y}],\underline{z})$. Let $\mathbb{C}$ be the set of complex numbers for $z\in\mathbb{C}$ we denote with $\overline{z}$ the complex conjugate of z, with $|z|$ the modulus of z and with $\mathrm{Re}\, z$, $\mathrm{Im}\, z$ the real and imaginary part of z, respectively. Let $\mathbb{C}^3$ be the three-dimensional complex vector space and $\underline{w}=(w_1,w_2,w_3)^T\in\mathbb{C}^3$ be a generic vector. In the following the symbols $(.,.)$, $[.,.]$ will be used also with complex vectors as arguments, in this case we denote, respectively, the real Euclidean product and the real Euclidean vector product of complex vectors. Let $A\subset\mathbb{R}^3$ be an open set, $\overline{A}$ be the closure of A, we denote with A^c the set $A^c=\mathbb{R}^3\backslash\overline{A}$. Let $\Omega\subset\mathbb{R}^3$ be an open bounded simply connected set, we say that Ω is of class $C^{0,1}$ if the boundary of Ω, $\partial\Omega$, is a locally Lipschitz boundary, that is each point $\underline{x}$ on the boundary of Ω has a neighborhood $U_{\underline{x}}$ such that $\partial\Omega\cup U_{\underline{x}}$ is the graph of a Lipschitz continuous function. Let $\underline{n}(\underline{x})$ be the outward unit normal vector to $\partial\Omega$.

[a]Electronic mail: recchioni@posta.econ.unian.it; Tel: +39+71+2207056; Fax: +39+71+2207058.
[b]Electronic mail: f.zirilli@caspur.it; Tel: +39+6+49913282; Fax: +39+6+44701007.

We note that when Ω is of class $C^{0,1}$ the outward unit normal vector exists almost everywhere on $\partial\Omega$ (see Ref. 1 Lemma 2.4.2 page 88). Let $\underline{g}(\underline{x})=(g_1(\underline{x}),g_2(\underline{x}),g_3(\underline{x}))^T$ be a complex-valued vector field defined on $\partial\Omega$ tangential to $\partial\Omega$, that is such that $(\underline{g}(\underline{x}),\underline{n}(\underline{x}))=0$ almost everywhere for $\underline{x}\in\partial\Omega$. Let $\Omega\in C^{0,1}$ and $\underline{g}(\underline{x})$ be a tangential vector field defined on $\partial\Omega$, we consider the following boundary value problem for the vector Helmholtz equation:

$$(\Delta+k^2)\underline{E}(\underline{x})=\underline{0},\quad \underline{x}\in\Omega^c \tag{1.1}$$

$$\operatorname{div}\underline{E}(\underline{x})=0,\quad \underline{x}\in\Omega^c \tag{1.2}$$

$$[\underline{n},\underline{E}](\underline{x})=\underline{g}(\underline{x}),\quad \underline{x}\in\partial\Omega \tag{1.3}$$

and

$$[\operatorname{curl}\underline{E}(\underline{x}),\hat{\underline{x}}]-ik\underline{E}(\underline{x})=o\left(\frac{1}{\|\underline{x}\|}\right),\quad \|\underline{x}\|\to\infty \tag{1.4}$$

where $\underline{0}=(0,0,0)^T$, $\hat{\underline{x}}=x/\|\underline{x}\|$ for $\underline{x}\neq\underline{0}$, and $o(\cdot)$ is the Landau symbol. The complex constant k is the wave number, later we assume either $\operatorname{Im}k>0$ or $\operatorname{Im}k=0$ and $\operatorname{Re}k>0$. Moreover $\underline{E}(\underline{x})=(E_1(\underline{x}),E_2(\underline{x}),E_3(\underline{x}))^T$, $\operatorname{div}\underline{E}(\underline{x})=\Sigma_{j=1}^3(\partial E_j/\partial x_j)(\underline{x})$, $\Delta(\cdot)=\Sigma_{j=1}^3\partial^2(\cdot)/\partial x_j^2$, $\Delta\underline{E}(\underline{x})=(\Delta E_1(\underline{x}),\Delta E_2(\underline{x}),\Delta E_3(\underline{x}))^T$, and $\operatorname{curl}\underline{E}(\underline{x})=((\partial E_3/\partial x_2)(\underline{x})-(\partial E_2/\partial x_3)(\underline{x}),(\partial E_1/\partial x_3)(\underline{x})-(\partial E_3/\partial x_1)(\underline{x}),(\partial E_2/\partial x_1)(\underline{x})-(\partial E_1/\partial x_2)(\underline{x}))^T$.

We consider a time harmonic electromagnetic field that hits the surface $\partial\Omega$ of the obstacle Ω. Let $\underline{E}^i(\underline{x})$, $\underline{x}\in\Omega^c$, be the part dependent from the spatial coordinates $\underline{x}$ of the incident electric field. In the following we call $\underline{E}^i(\underline{x})$ incident electric field. We assume that the medium that surrounds the obstacle Ω is a homogeneous isotropic medium that does not contain free electric charges, and that the incident electric field $\underline{E}^i$ satisfies (1.1) and (1.2) in $\mathbb{R}^3$. Let $\underline{E}^s(\underline{x})$, $\underline{x}\in\Omega^c$, be the electric field scattered by the perfectly conducting surface $\partial\Omega$ when hit by $\underline{E}^i(\underline{x})$. To be precise $\underline{E}^s(\underline{x})$ is only the part depending from the spatial coordinates of the scattered electric field. Then the scattered field $\underline{E}^s$ is solution of the boundary value problem (1.1), (1.2), (1.3), and (1.4) where the function $\underline{g}(\underline{x})$ appearing in (1.3) is given by (see Ref. 2 page 121):

$$\underline{g}(\underline{x})=-[\underline{n},\underline{E}^i](\underline{x}),\quad \underline{x}\in\partial\Omega. \tag{1.5}$$

In this paper we study the problem of the existence and uniqueness of the solution of problem (1.1), (1.2), (1.3), and (1.4) and we propose a new numerical method to compute it. First of all we prove that under some hypotheses the following formula for the vector field $\underline{E}$ solution of (1.1), (1.2), (1.3), and (1.4) holds:

$$\underline{E}(\underline{x})=\frac{e^{ik\|\underline{x}\|}}{4\pi\|\underline{x}\|}\sum_{n=0}^{\infty}\frac{\underline{E}^{(n)}(\hat{\underline{x}})}{\|\underline{x}\|^n},\quad \|\underline{x}\|\geqslant R \tag{1.6}$$

where $\underline{E}^{(n)}(\hat{\underline{x}})$, $n=0,1,2,\ldots$ are appropriate coefficients of the series expansion (1.6) and R is the radius of a sphere that contains $\overline{\Omega}$. The leading term $\underline{E}^{(0)}(\hat{\underline{x}})$ of the series (1.6) is called far field pattern associated to $\underline{E}(\underline{x})$. Then through the solution of some auxiliary problems we reformulate the boundary value problem (1.1), (1.2), (1.3), and (1.4) as a system of integral equations. That is we reduce problem (1.1), (1.2), (1.3), and (1.4) to the following system of integral equations:

$$(I+\tau)\underline{f}=\underline{F}, \tag{1.7}$$

where the vector functions $\underline{f},\underline{F}$ belong to a Hilbert space, I is identity operator and τ is a compact operator acting on the Hilbert space. Given $\underline{F}$ we prove existence and uniqueness for the solution of (1.7) using the Riesz theory for compact operators. This implies the claimed existence result for the solution of (1.1), (1.2), (1.3), and (1.4). The uniqueness result follows from (1.6) and the assumption on the wave number k. We note that our existence and uniqueness theorem holds for $\Omega \in C^{0,1}$, that is when $\partial\Omega$ is only locally Lipschitz, while the usual existence and uniqueness theorems assume more regularity for $\partial\Omega$. For example, when $\partial\Omega$ is assumed to be of class C^2 the existence and the uniqueness of the solution of the boundary value problems for the Helmholtz equation or for the vector Helmholtz equation can be proved reducing the boundary value problems to boundary integral equations and applying the Riesz–Fredholm theory for compact operators.[2] In fact the operators involved in these boundary integral equations are compact operators acting on the normed space of complex valued continuous functions defined on $\partial\Omega$. When $\partial\Omega$ is only Lipschitz continuous reducing the boundary value problems to boundary integral equations leads to integral equations with singular kernels that are not easy to solve. The study of layer potentials and of boundary value problems in the Lipschitz case goes back to Refs. 3 and 4. Since then it has become an active research field, we mention only Ref. 5 for the study of the Laplace equation, Ref. 6 for the study of the time harmonic Maxwell equations, and Ref. 7 for a survey of the subject.

We deal with the same difficult, i.e., $\partial\Omega$ only locally Lipschitz, when we prove the existence of the solution of the boundary value problem (1.1), (1.2), (1.3), and (1.4) for the vector Helmholtz equation. We overcome the difficulty avoiding layer potentials and singular integral operators. In fact the integral equations (1.7) involves compact integral operators and can be treated using Riesz–Fredholm theory. However, the functions f and F appearing in (1.7) are not defined on $\partial\Omega$ as it will be the case if we had used a boundary integral method but are defined on an annulus surrounding Ω, that is we consider a kind of volume potentials. In the Lipschitz case the use of layer potentials and boundary integral methods to establish the existence of the solution of the elliptic boundary value problems is based on several deep results in analysis. In the case of the boundary value problem (1.1), (1.2), (1.3), and (1.4) our approach establishes the existence of the solution using only elementary results in functional analysis.

Without loss of generality we can assume that Ω contains the origin. Let $a>0$ be a constant, $B_a=\{\underline{x}\in\mathbb{R}^3/\|\underline{x}\|<a\}$ be the sphere of radius a and center the origin and let ∂B_a be its boundary. In the following B_1 and ∂B_1 will be denoted also with B and ∂B. The set Ω contains the sphere $\overline{B_a}$ for some $a>0$. Let (r,θ,ϕ) be the canonical spherical coordinates of $\underline{x}\in\mathbb{R}^3$, for $\underline{x}\neq\underline{0}$, we have

$$\hat{\underline{x}}(\theta,\phi)=\frac{\underline{x}}{\|\underline{x}\|}=(\sin\theta\cos\phi,\sin\theta\sin\phi,\cos\theta)^T, \tag{1.8}$$

where $0\leqslant\theta\leqslant\pi,\ 0\leqslant\phi<2\pi$.

In order to introduce a new numerical method to compute the vector field $\underline{E}(\underline{x})$ solution of the boundary value problem (1.1), (1.2), (1.3), and (1.4), we assume that

$$\Omega=\{\underline{x}=r\hat{\underline{x}}\in\mathbb{R}^3/0\leqslant r<\xi(\hat{\underline{x}}),\hat{\underline{x}}\in\partial B\}, \tag{1.9}$$

where ξ is a single valued sufficiently regular function defined on ∂B. So that

$$\partial\Omega=\{\underline{x}=r\hat{\underline{x}}\in\mathbb{R}^3/r=\xi(\hat{\underline{x}}),\hat{\underline{x}}\in\partial B\} \tag{1.10}$$

is a starlike surface with respect to the origin, that is the boundary of the obstacle $\partial\Omega$ can be represented by a single valued function in spherical coordinates. We note that the choice of spherical coordinates is not essential and that our numerical method can be developed for other coordinate systems. The numerical method introduced here is based on a formalism that generalizes the formalism introduced by Milder in Refs. 8 and 9 to study the scattering of acoustic waves from a rough unbounded surface that divides the three-dimensional space in two ''disturbed'' half spaces and further developed by Milder[10] and by Piccolo, Recchioni, and Zirilli[11] to electromagnetic scattering from unbounded perfectly conducting surfaces of the previous type and by Smith[12] to the scattering of electromagnetic waves from dielectric surfaces. Later Misici, Pacelli, and

Zirilli[13] have generalized Milder's formalism to the scattering of acoustic waves from a bounded obstacle. The numerical method introduced here simplifies the original Milder's formalism since avoids the need to obtain $[\underline{n},\mathrm{curl}\,\underline{E}](\underline{x})$, $\underline{x}\in\partial\Omega$ from the knowledge of $[\underline{n},\underline{E}](\underline{x})$, $\underline{x}\in\partial\Omega$. This simplification gives a substantial reduction in the computational cost of the method proposed. Similar simplified Milder's methods can be developed for the acoustic and electromagnetic problems considered in the previously mentioned papers. The numerical method proposed here constructs the vector field $\underline{E}$ solution of (1.1), (1.2), (1.3), and (1.4) on Ω^c from the knowledge of $[\underline{n},\underline{E}]$ on $\partial\Omega$. Let

$$\Phi(\underline{x},\underline{y})=\frac{e^{ik\|\underline{x}-\underline{y}\|}}{4\pi\|\underline{x}-\underline{y}\|} \tag{1.11}$$

be the fundamental solution of the Helmholtz equation in $\mathbb{R}^3$ with the Sommerfeld radiation condition at infinity. We assume for $\underline{E}(\underline{x})$ solution of (1.1), (1.2), (1.3), and (1.4) the following representation formula:

$$\underline{E}(\underline{x})=\mathrm{curl}_{\underline{x}}\int_{\mathbb{R}^3}\Phi(\underline{x},\underline{y})\,\underline{v}(\hat{\underline{y}})\,\delta(\|\underline{y}\|-a)\,d\underline{y}, \tag{1.12}$$

where $\mathrm{curl}_{\underline{x}}$ is the curl operator with respect to the $\underline{x}$ variable, $\underline{v}(\hat{\underline{y}})$ is a suitable vector density function and $\delta(\|\underline{y}\|-a)$ is a ''Dirac's delta'' concentrated on ∂B_a, finally remember that $a>0$ is chosen such that $\overline{B_a}\subset\Omega$. We obtain for $\underline{v}$ and consequently for $\underline{E}(\underline{x})$ and its associated far field pattern $\underline{E}^{(0)}(\hat{\underline{x}})$, a formal series expansion in ''powers'' of $\delta\xi=\xi-1$ where 1 represents the boundary of the unit sphere whose surface is assumed as reference surface, that is as ''base point'' of the ''power series expansion.'' We note that the surface of the unit sphere is chosen as reference surface only for convenience. A similar formal ''power'' series expansion for $\underline{E}(\hat{\underline{x}})$ or $\underline{E}^{(0)}(\hat{\underline{x}})$ can be obtained when more general surfaces are used as reference surfaces. The terms of the series expansion in ''powers'' of $\delta\xi$ are integrals independent one from the other that can be computed in parallel. So that the efficiency of the numerical method proposed here is due to its highly parallelizable structure. Finally we present some numerical experience on test problems where the incident field is a plane linearly polarized time harmonic wave. When the domain $\Omega\in C^{0,1}$ is bounded with smooth boundary we compare the far field patterns $\underline{E}^{(0)}(\hat{\underline{x}})$ obtained with the method proposed here with those obtained with the *T*-matrix method (see Ref. 14).

This comparison shows that the two methods give similar results. Moreover we compute the far field pattern $\underline{E}^{(0)}(\hat{\underline{x}})$ of the scattered field solution of the boundary value problem (1.1), (1.2), (1.3), and (1.4) when Ω is a domain with locally Lipschitz boundary or with smooth boundary with multiscale corrugations. The obstacles with locally Lipschitz boundary that we consider are polyhedra. The relation between the geometry (i.e., facets, edges, vertices) of the surface of a polyhedron and the corresponding far field pattern is investigated (see Figs. 6 and 7). The obstacles with multiscale corrugations are represented by ''corrugated'' spheres (see Figs. 2 and 3). When the obstacle is a corrugated sphere we show that for some special values of the wave number k there is a ''resonance phenomenon'' due to the corrugation (see Figs. 4 and 5).

In Sec. II we prove formula (1.6) and we prove the existence and uniqueness theorem for the solution of the boundary value problem (1.1), (1.2), (1.3), and (1.4). In Sec. III we describe the numerical method proposed to compute the solution $\underline{E}(\underline{x})$ of (1.1), (1.2), (1.3), and (1.4) and we give the ''power'' series expansion in powers of $\delta\xi=\xi-1$ of $\underline{E}(\underline{x})$ and of the associated far field pattern $\underline{E}^{(0)}(\hat{\underline{x}})$. Finally in Sec. IV the method developed in Sec. III is applied to some test electromagnetic scattering problems, some numerical results are shown and their physical meaning illustrated.

II. AN EXISTENCE AND UNIQUENESS THEOREM FOR THE SOLUTION OF A BOUNDARY VALUE PROBLEM FOR THE VECTOR HELMHOLTZ EQUATION

Let $A \subseteq \mathbb{R}^3$ be an open set, let $C^k(A)$, $k=0,1,2,...$, be the space of real- or complex-valued k-times continuously differentiable functions in A and let $C_0^k(A)$ be the space of functions belonging to $C^k(A)$ with compact support in A. We denote with $C^\infty(A)$ the space of real- or complex-valued infinitely continuously differentiable functions defined in A, and with $C_0^\infty(A)$ the space of functions in $C^\infty(A)$ with compact support in A.

Let $L^p(A)$, $1 \leqslant p \leqslant \infty$, be the usual Lebesgue space of index p of complex valued functions. We denote with $\|\cdot\|_p$ the $L^p(A)$ norm and with $\|\cdot\|_\infty$ the $L^\infty(A)$ norm. Let α_i, $i=1,2,3$ be non-negative integers and $\underline{\alpha}=(\alpha_1,\alpha_2,\alpha_3)^T$ be a multi-index, we denote with $|\underline{\alpha}|=\Sigma_{i=1}^3 \alpha_i$ the length of $\underline{\alpha}$. Let m be a non-negative integer and $W^{m,p}(A)$ be the space of the functions $u \in L^p(A)$ such that $D^{\underline{\alpha}} u \in L^p(A)$ for any multi-index $\underline{\alpha}$ such that $0 \leqslant |\underline{\alpha}| \leqslant m$, where $D^{\underline{\alpha}} u = (\partial^{\alpha_1}/\partial x_1^{\alpha_1})(\partial^{\alpha_2}/\partial x_2^{\alpha_2})(\partial^{\alpha_3}/\partial x_3^{\alpha_3})u$ denotes the weak derivative of u of order given by the multi-index $\underline{\alpha}$. We denote with $\|\cdot\|_{m,p}$ the norm of $W^{m,p}(A)$ that is

$$\|u\|_{m,p}=\left\{\sum_{0\leqslant|\underline{\alpha}|\leqslant m}\|D^{\underline{\alpha}}u\|_p^p\right\}^{1/p}, \quad 1\leqslant p<\infty, \quad u\in W^{m,p}(A), \tag{2.1}$$

$$\|u\|_{m,\infty}=\max_{0\leqslant|\underline{\alpha}|\leqslant m}\|D^{\underline{\alpha}}u\|_\infty, \quad u\in W^{m,\infty}(A). \tag{2.2}$$

Let $\mathcal{W}^{m,p}(A)$ be the space of functions defined on A such that $u \in W^{m,p}(A\cap U)$ for any open set U such that $\overline{U}$ is compact and $A\cap U$ is not empty.

Let

$$(L^p(A))^3=\{\underline{E}=(E_1,E_2,E_3)^T/E_i\in L^p(A), i=1,2,3\}, \tag{2.3}$$

$$(W^{m,p}(A))^3=\{\underline{E}=(E_1,E_2,E_3)^T/E_i\in W^{m,p}(A), i=1,2,3\}. \tag{2.4}$$

In a similar way we define $(\mathcal{W}^{m,p}(A))^3$ and the other product spaces that appear in the following. The scalar products on $(L^2(A))^3$ and $(W^{m,2}(A))^3$ are given by

$$(\underline{E},\underline{F})_{(L^2)^3}=\sum_{i=1}^3\int_A E_i\overline{F_i}\,d\underline{x}=\int_A(\underline{E},\overline{\underline{F}})d\underline{x}, \forall \underline{E},\underline{F}\in(L^2(A))^3, \tag{2.5}$$

$$\begin{aligned}(\underline{E},\underline{F})_{(W^{m,2})^3}&=\sum_{i=1}^3\sum_{0\leqslant|\underline{\alpha}|\leqslant m}\int_A D^{\underline{\alpha}}E_i\overline{D^{\underline{\alpha}}F_i}\,d\underline{x}\\ &=\sum_{0\leqslant|\underline{\alpha}|\leqslant m}\int_A(D^{\underline{\alpha}}\underline{E},\overline{D^{\underline{\alpha}}\underline{F}})d\underline{x}, \forall \underline{E},\underline{F}\in(W^{m,2}(A))^3.\end{aligned} \tag{2.6}$$

Let $U\subset\mathbb{R}^3$ be a bounded open set with locally Lipschitz boundary ∂U and let $u\in W^{1,2}(U)$, from Theorem 4.2, page 84 of Ref. 1 we have that the trace of u on ∂U, $u|_{\partial U}$, belongs to $L^2(\partial U)$. In the following in order to simplify the notation instead of $u|_{\partial U}$ we continue to use u to denote the trace of u on ∂U.

Lemma 2.1: Let U be as above and let $u,v\in W^{1,2}(U)$, we have

$$\int_U v\frac{\partial u}{\partial x_i}d\underline{x}=\int_{\partial U}vun_i\,d\sigma-\int_U u\frac{\partial v}{\partial x_i}d\underline{x}, \quad i=1,2,3, \tag{2.7}$$

where $d\sigma$ denotes the surface measure on ∂U and $\underline{n}(\underline{x})=(n_1(\underline{x}),n_2(\underline{x}),n_3(\underline{x}))^T$ is the outward unit normal vector to ∂U in $\underline{x}\in\partial U$.

Proof: See Ref. 1, page 121, Theorem 1.1. ∎

Lemma 2.2: Let U be as above and let $u \in W^{2,2}(U), v \in W^{1,2}(U)$, then

$$\sum_{i=1}^{3} \int_U v \frac{\partial^2 u}{\partial x_i^2} d\underline{x} = \sum_{i=1}^{3} \int_{\partial U} v \frac{\partial u}{\partial x_i} n_i \, d\sigma - \sum_{i=1}^{3} \int_U \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} d\underline{x}. \tag{2.8}$$

Proof: The proof follows from Lemma 2.1 applied to the functions $\partial u/\partial x_i \in W^{1,2}(U)$, $i = 1,2,3$, and $v \in W^{1,2}(U)$. ■

Moreover when vector functions are considered we have

Theorem 2.3: Let U be as above and let $\underline{E} \in (W^{1,2}(U))^3$, $\underline{F} \in (W^{2,2}(U))^3$ be two vector functions; we have

$$\int_U \{(\underline{E}, \Delta \underline{F}) + (\mathrm{curl}\, \underline{E}, \mathrm{curl}\, \underline{F}) + \mathrm{div}\, \underline{E}\ \mathrm{div}\, \underline{F}\} d\underline{x}$$

$$= \int_{\partial U} \{(\underline{n}, \underline{E}, \mathrm{curl}\, \underline{F}) + (\underline{n}, \underline{E}) \mathrm{div}\, \underline{F}\} d\sigma. \tag{2.9}$$

Proof: Formula (2.9) is the so-called vector Green's formula and follows immediately from the scalar Green's formulas (2.7) and (2.8). ■

Lemma 2.4: Let $k \in \{z \in \mathbb{C} / \mathrm{Im}\, z > 0\} \cup \{z \in \mathbb{C} / \mathrm{Re}\, z > 0, \mathrm{Im}\, z = 0\}$, $\Omega \in C^{0,1}$, and let $u \in \mathcal{W}^{2,2}(\Omega^c)$ be a solution of the scalar Helmholtz equation

$$\Delta u + k^2 u = 0, \quad \underline{x} \in \Omega^c \tag{2.10}$$

satisfying the Sommerfeld radiation condition at infinity:

$$\left(\frac{\underline{x}}{\|\underline{x}\|}, \nabla u \right) - iku = o\left(\frac{1}{\|\underline{x}\|} \right), \quad \|\underline{x}\| \to \infty, \tag{2.11}$$

where $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)^T$ is the gradient operator, and let $R > 0$ be such that $B_R \supset \overline{\Omega}$. Then u has an expansion on the form

$$u(\underline{x}) = \frac{e^{ik\|\underline{x}\|}}{4\pi\|\underline{x}\|} \sum_{n=0}^{\infty} \frac{u^{(n)}(\hat{\underline{x}})}{\|\underline{x}\|^n}, \quad \|\underline{x}\| \geqslant R, \tag{2.12}$$

where $u^{(n)}(\hat{\underline{x}})$, $n = 0,1,2,\ldots$ are suitable functions.

Proof: We note that if $u \in \mathcal{W}^{2,2}(\Omega^c)$ for the couple $u(\underline{x})$, $v(\underline{x}) = 1$ Green's formula (2.8) holds in $\Omega^c \cap U$, for any open set U with locally Lipschitz boundary and such that $\overline{U}$ is compact and $U \supset \overline{\Omega}$ so that we can proceed as in Ref. 2, Theorem 3.6, page 72. ■

Theorem 2.5: Let k be as in Lemma 2.4, let $\underline{E} \in (\mathcal{W}^{2,2}(\Omega^c))^3$ be a solution of the vector Helmholtz equation satisfying the radiation condition (1.4) and let $R > 0$ be such that $B_R \supset \overline{\Omega}$. Then $\underline{E}$ has an expansion on the form

$$\underline{E}(\underline{x}) = \frac{e^{ik\|\underline{x}\|}}{4\pi\|\underline{x}\|} \sum_{n=0}^{\infty} \frac{\underline{E}^{(n)}(\hat{\underline{x}})}{\|\underline{x}\|^n}, \quad \|\underline{x}\| \geqslant R, \tag{2.13}$$

where $\underline{E}^{(n)}(\hat{\underline{x}})$, $n = 0,1,2,\ldots$ are suitable functions.

Proof: We observe that the Cartesian components of $\underline{E} = (E_1, E_2, E_3)^T$ satisfy the scalar Helmholtz equation and the Sommerfeld radiation condition at infinity, therefore we can apply Lemma 2.4 with $u = E_i$, $i = 1,2,3$ and we obtain (2.13). ■

Lemma 2.6: Let k be real and positive and let $u \in \mathcal{W}^{2,2}(\Omega^c)$ be a solution of the scalar Helmholtz equation satisfying the Sommerfeld radiation condition at infinity such that

$$\lim_{R\to+\infty}\int_{\partial B_R}|u|^2 ds=0, \tag{2.14}$$

where ds denotes the surface measure on ∂B_R then $u=0$ in Ω^c.

Proof: The proof follows using (2.12) and arguing as in Ref. 2, Lemma 3.11, page 77. ■

Theorem 2.7: *(Uniqueness):* Let k be as in Lemma 2.4, $\Omega\in C^{0,1}$, $\underline{g}\in(L^2(\partial\Omega))^3$ and let $\underline{E},\underline{F}\in(\mathcal{W}^{2,2}(\Omega^c))^3$ be two solutions of the boundary value problem (1.1), (1.2), (1.3), and (1.4). Then $\underline{E}=\underline{F}$ in $(\mathcal{W}^{2,2}(\Omega^c))^3$.

Proof: We consider the vector function $\underline{W}=\underline{E}-\underline{F}$. The function $\underline{W}$ satisfies the homogeneous problem:

$$(\Delta+k^2)\underline{W}(\underline{x})=\underline{0},\quad \underline{x}\in\Omega^c \tag{2.15}$$

$$\operatorname{div}\underline{W}(\underline{x})=0,\quad \underline{x}\in\Omega^c \tag{2.16}$$

$$[\underline{n},\underline{W}](\underline{x})=\underline{0},\quad \underline{x}\in\partial\Omega \tag{2.17}$$

$$[\operatorname{curl}\underline{W}(\underline{x}),\hat{\underline{x}}]-ik\underline{W}(\underline{x})=o\left(\frac{1}{\|\underline{x}\|}\right),\quad \|\underline{x}\|\to\infty. \tag{2.18}$$

From (2.18) follows that

$$\lim_{R\to\infty}\int_{\partial B_R}\sum_{j=1}^{3}\left|\left[\operatorname{curl}\underline{W},\frac{\underline{x}}{\|\underline{x}\|}\right]_j-ikW_j\right|^2 ds=0, \tag{2.19}$$

where $[\operatorname{curl}\underline{W},\underline{x}/\|\underline{x}\|]_j$, denotes the $\underline{j}$ component of $[\operatorname{curl}\underline{W},\underline{x}/\|\underline{x}\|]$ $j=1,2,3$. From standard arguments using Lemma 2.6 we obtain $\underline{W}=\underline{0}$ in Ω^c, therefore $\underline{E}=\underline{F}$ in $(\mathcal{W}^{2,2}(\Omega^c))^3$. ■

Let U be an open bounded set with locally Lipschitz boundary ∂U, $K(U)$ be the vector subspace of $(L^2(U))^3$ defined by

$$K(U)=\{\underline{E}\in(W^{2,2}(U))^3:\ \operatorname{div}\underline{E}=0\ \text{ on }\ \partial U\ \text{ and }\ [\underline{E},\underline{n}]=\underline{0}\ \text{ on }\ \partial U\}. \tag{2.20}$$

We note that for $\underline{E}\in(W^{2,2}(U))^3$ the trace operators $\underline{E}\to(\operatorname{div}\underline{E})|_{\partial U}$ and $\underline{E}\to[\underline{E},\underline{n}]|_{\partial U}$ are linear bounded operators from $(W^{2,2}(U))^3$ to $L^2(\partial U)$ and from $(W^{2,2}(U))^2$ to $(L^2(\partial U))^3$, respectively (see Ref. 1, Theorem 4.2, page 84).

Definition 2.8: Let

$$T:\ K(U)\subset(L^2(U))^3\to(L^2(U))^3 \tag{2.21}$$

be the linear differential operator defined by

$$T\underline{E}=-\Delta\underline{E}=(-\Delta E_1,-\Delta E_2,-\Delta E_3)^T,\quad \underline{E}\in K(U). \tag{2.22}$$

Theorem 2.9: The linear operator T: $K(U)\subset(L^2(U))^3\to(L^2(U))^3$ defined by (2.22) is closed, symmetric and non-negative.

Proof: The thesis follows using Theorem 2.3, formula (2.8) and from standard arguments. ■

Theorem 2.10: The linear operator T given in Definition 2.8 is self-adjoint.

Proof: The proof follows with an easy computation applying the vector Green's formula (2.9). ■

We have

Theorem 2.11: Let $C\subset\mathbb{R}^3$ be an open cube, let $K(C)$ be defined as in (2.20), the spectrum of the operator

$$T: \ K(C) \subset (L^2(C))^3 \rightarrow (L^2(C))^3 \tag{2.23}$$

defined in Definition 2.8 is discrete.

Proof: Without loss of generality we can restrict our attention to the unit cube $C_1 = \{\underline{x} \in \mathbb{R}^3 : 0 < x_i < 1, i = 1,2,3\}$.

The eigenvalue problem for T becomes

$$(\Delta + \lambda^2)\underline{E}(\underline{x}) = \underline{0}, \quad \underline{x} \in C_1 \tag{2.24}$$

$$\operatorname{div} \underline{E}(\underline{x}) = 0, \quad \underline{x} \in \partial C_1 \tag{2.25}$$

$$[\underline{n}, \underline{E}](\underline{x}) = \underline{0}, \quad \underline{x} \in \partial C_1 . \tag{2.26}$$

We seek the values of the parameter λ^2 such that problem (2.24), (2.25), and (2.26) have a nonzero solution $\underline{E}(\underline{x}) \in K(C_1) \subset (L^2(C_1))^3$.

The eigenvalue problem (2.24), (2.25), and (2.26) is easily solved by separation of variables. The eigenvalues of T are

$$\lambda^2_{m,n,p} = (m^2 + n^2 + p^2)\pi^2, \quad m,n,p = 1,2,..., \tag{2.27}$$

and it is easy to see that the corresponding eigenfunctions are a complete system in $(L^2(C_1))^3$ and the eigenvalues are isolated points of the spectrum and have finite multiplicity. Hence the spectrum of the operator T is discrete. Let $\mu^{(k)}$ be the kth eigenvalue when the eigenvalues in (2.27) are reordered in increasing order, we have $\lim_{k\rightarrow\infty} \mu^{(k)} = \infty$. ■

Theorem 2.12: Let $\Omega \in C^{0,1}$, $C \subset \mathbb{R}^3$ be an open cube such that $\overline{\Omega} \subset C$ and let $V = \Omega^c \cap C$. Let $K(V)$ be defined as in (2.20) then the operator T:

$$T: \ K(V) \subset (L^2(V))^3 \rightarrow (L^2(V))^3 \tag{2.28}$$

defined in (2.22) has discrete spectrum.

Proof: Given $\underline{E} \in (L^2(V))^3$ we define the vector function $\tilde{\underline{E}}$

$$\tilde{\underline{E}}(\underline{x}) = \begin{cases} \underline{E}(\underline{x}), & \underline{x} \in V \\ \underline{0}, & \underline{x} \in \overline{\Omega}. \end{cases} \tag{2.29}$$

Let $\mathcal{Q}_C$ be the quadratic form associated to the operator T on $K(C)$ and $\mathcal{D}(\mathcal{Q}_C)$ be its domain, in a similar way let $\mathcal{Q}_V$ be the quadratic form associated to the operator T on $K(V)$ and $\mathcal{D}(\mathcal{Q}_V)$ be its domain. When $\underline{E} \in \mathcal{D}(\mathcal{Q}_V)$ we have that $\tilde{\underline{E}}$ given by (2.29) belongs to $\mathcal{D}(\mathcal{Q}_C)$ moreover we have $\mathcal{Q}_V(\underline{E},\underline{E}) = \mathcal{Q}_C(\tilde{\underline{E}},\tilde{\underline{E}})$. Hence from Theorem 2.11 and from Theorem XIII.2, page 78,[15] the operator T: $K(V) \subset (L^2(V))^3 \rightarrow (L^2(V))^3$ has discrete spectrum. ■

Let us consider now the boundary value problem (1.1), (1.2), (1.3), and (1.4) for the vector Helmholtz equation. We begin with some preliminary results.

Theorem 2.13: Let k be as in Lemma 2.4, $\Omega \in C^{0,1}$ and let $\underline{E} \in (\mathcal{W}^{2,2}(\Omega^c))^3$ be a solution of the vector Helmholtz equation (1.1) that verifies the radiation condition (1.4) and the boundary condition

$$\operatorname{div} \underline{E}(\underline{x}) = 0, \quad \underline{x} \in \partial\Omega, \tag{2.30}$$

then we have

$$\operatorname{div} \underline{E}(\underline{x}) = 0, \quad \underline{x} \in \Omega^c. \tag{2.31}$$

Proof: Since $\underline{E}$ is solution of the vector Helmholtz equation and $\underline{E}$ satisfies condition (1.4), we have that $u=\operatorname{div}\underline{E}$ verifies the scalar Helmholtz equation (2.10) and the Sommerfeld radiation condition at infinity (2.11). Moreover, we have assumed that $u=\operatorname{div}\underline{E}=0$ on $\partial\Omega$.

Let $R>0$ be such that $\overline{\Omega}\subset B_R$, from (2.8) applied to the domain $\Omega^c\cap B_R$ and the radiation condition (2.11) since $u=0$ on $\partial\Omega$ we have

$$\lim_{R\to\infty}\left\{\int_{\partial B_R}\left[\left|\frac{\partial u}{\partial \underline{n}}\right|^2+|k|^2|u|^2\right]ds+2\,\operatorname{Im}k\int_{\Omega^c\cap B_R}\left[|k|^2|u|^2+\sum_{i=1}^{3}\left|\frac{\partial u}{\partial x_i}\right|^2\right]d\underline{x}\right\}=0,$$

hence if $\operatorname{Im}k>0$ we have

$$\lim_{R\to\infty}\int_{\Omega^c\cap B_R}|u|^2\,d\underline{x}=0,$$

that is $u=0$ in Ω^c, if $\operatorname{Im}k=0$ and $\operatorname{Re}k>0$ we have

$$\lim_{R\to\infty}\int_{\partial B_R}|u|^2\,ds=0$$

so that from Lemma 2.6, we can conclude $u=0$ in Ω^c. ■

Now we reformulate problem (1.1), (1.2), (1.3), and (1.4) in precise mathematical form, that is we state the hypotheses on the vector field $\underline{g}(\underline{x})$ appearing in the boundary condition (1.3) and declare the functional class where we seek the solution of the problem considered.

Problem 2.1: Let k be as in Lemma 2.4, $\Omega\in C^{0,1}$, we seek a vector field $\underline{E}\in(\mathcal{W}^{2,2}(\Omega^c))^3$ such that

$$(\Delta+k^2)\underline{E}(\underline{x})=\underline{0},\quad \underline{x}\in\Omega^c \tag{2.32}$$

$$\operatorname{div}\underline{E}(\underline{x})=0,\quad \underline{x}\in\partial\Omega \tag{2.33}$$

$$[\underline{n},\underline{E}](\underline{x})=\underline{D}(\underline{x}),\quad \underline{x}\in\partial\Omega \tag{2.34}$$

and

$$[\operatorname{curl}\underline{E}(\underline{x}),\hat{\underline{x}}]-ik\underline{E}(\underline{x})=o\left(\frac{1}{\|\underline{x}\|}\right),\quad \|\underline{x}\|\to\infty \tag{2.35}$$

where $\underline{E}\in(L^2(\partial\Omega))^3$ is a given vector field and there exists $\underline{G}=(G_1,G_2,G_3)^T\in(\mathcal{W}^{2,2}(\mathbb{R}^3))^3$ such that

$$\text{(i)}\quad \operatorname{div}\underline{G}(\underline{x})=0,\quad \underline{x}\in\partial\Omega$$

$$\text{(ii)}\quad \underline{D}(\underline{x})=[\underline{n},\underline{G}](\underline{x}),\quad \underline{x}\in\partial\Omega.$$

We note that the datum of the scattering problem mentioned in Sec. I, $\underline{D}(\underline{x})=-[\underline{n},\underline{E}^i](\underline{x})$, satisfies the hypotheses given above when $\underline{E}^i\in(\mathcal{W}^{2,2}(\mathbb{R}^3))^3$. More general classes of data $\underline{D}(\underline{x})$ can be considered. We note that from Theorem 2.13 follows that a solution of Problem 2.1 solves (1.1), (1.2), (1.3), and (1.4) with $\underline{g}(\underline{x})=\underline{D}(\underline{x})$, $\underline{x}\in\partial\Omega$. We restate Problem 2.1 in several equivalent forms.

Let $C\subset\mathbb{R}^3$ an open cube such that $\overline{\Omega}\subset C$ then there exists a function $\phi\in C_0^\infty(\mathbb{R}^3)$ with support in C and such that $\phi(\underline{x})=1$ if $\underline{x}\in\overline{\Omega}$.

We can consider the vector function

$$\hat{\underline{E}}(\underline{x}) = \underline{E}(\underline{x}) - \phi(\underline{x})\underline{G}(\underline{x}), \quad \underline{x} \in \Omega^c. \tag{2.36}$$

It is easy to see that $\underline{E}$ satisfies Problem 2.1 if and only if $\hat{\underline{E}}$ is solution of the following problem.

Problem 2.2: Let k be as in Lemma 2.4 $\Omega \in C^{0,1}$, V as in Theorem 2.12, we seek a vector function $\hat{\underline{E}} \in (\mathcal{W}^{2,2}(\Omega^c))^3$ such that

$$(\Delta + k^2)\hat{\underline{E}}(\underline{x}) = \underline{A}(\underline{x}), \quad \underline{x} \in \Omega^c \tag{2.37}$$

$$\operatorname{div} \hat{\underline{E}}(\underline{x}) = 0, \quad \hat{x} \in \partial\Omega \tag{2.38}$$

$$[\underline{n}, \hat{\underline{E}}](\underline{x}) = \underline{0}, \quad \underline{x} \in \partial\Omega \tag{2.39}$$

and

$$[\operatorname{curl} \hat{\underline{E}}(\underline{x}), \hat{\underline{x}}] - ik\hat{\underline{E}}(\underline{x}) = o\left(\frac{1}{\|\underline{x}\|}\right), \quad \|\underline{x}\| \to \infty \tag{2.40}$$

where

$$\underline{A} = -(\Delta + k^2)(\phi \underline{G}). \tag{2.41}$$

The function $\underline{A}$ is a vector function belonging to $(L^2(\Omega^c))^3$ with support in $V = \Omega^c \cap C$. Note that $\operatorname{div}(\phi \underline{G}) = \Sigma_{i=1}^3 (\partial\phi/\partial x_i) G_i + \phi \operatorname{div} \underline{G} = 0$ on $\partial\Omega$, in fact $\operatorname{div} \underline{G} = 0$ on $\partial\Omega$ and $\partial\phi/\partial x_i = 0$, $i = 1,2,3$ on $\partial\Omega$ in virtue of the properties of ϕ and $\partial\Omega$.

Let $\Phi(\underline{x}, \underline{y})$ be given by (1.11) we have the following result.

Theorem 2.14: Let V be as above, $\underline{a}(\underline{x}) \in (L^2(V))^3$ be a vector function and let

$$\underline{v}(\underline{x}) = \int_V \Phi(\underline{x}, \underline{y}) \underline{a}(\underline{y}) d\underline{y}, \quad \underline{x} \in \mathbb{R}^3 \tag{2.42}$$

then $\underline{v} \in (\mathcal{W}^{2,2}(\mathbb{R}^3))^3$ so that in particular $\underline{v} \in (\mathcal{W}^{2,2}(\Omega^c))^3$ and for any open set U such that $\overline{U}$ is compact and $\Omega^c \cap U$ is not empty we have

$$\|\underline{v}\|_{(W^{2,2}(\Omega^c \cap U))^3} \leq \alpha_U \|\underline{a}\|_{(L^2(V))^3}, \tag{2.43}$$

where α_U is a positive constant that depends on U.

Proof: From Schwarz inequality for $i = 1,2,3$ we have that

$$|v_i(\underline{x})| = \left| \int_V \Phi(\underline{x}, \underline{y}) a_i(\underline{y}) d\underline{y} \right| \leq \left(\int_V |\Phi(\underline{x}, \underline{y})|^2 d\underline{y} \right)^{1/2} \left(\int_V |a_i(\underline{y})|^2 d\underline{y} \right)^{1/2} < \infty, \tag{2.44}$$

therefore, since

$$(\Delta + k^2)\underline{v} = \hat{\underline{a}}, \quad \underline{x} \in \mathbb{R}^3, \tag{2.45}$$

where $\hat{\underline{a}} = -\underline{a}$ in V and $\hat{\underline{a}} = \underline{0}$ in $\mathbb{R}^3 \backslash V$, we have that $\underline{v} \in (W^{2,2}(\Omega^c \cap U))^3$ and $\|\underline{v}\|_{(W^{2,2}(\Omega^c \cap U))^3} \leq \alpha_U \|\underline{a}\|_{(L^2(V))^3}$ for any open set U such that $\overline{U}$ is compact and $\Omega^c \cap U$ is not empty. ■

We introduce the following auxiliary problems.

Problem 2.3: Let V as above, $\underline{a} \in (L^2(V))^3$ and let μ^2 be a constant that is not an eigenvalue of T on $K(V)$, we seek a vector function $\underline{W} \in (W^{2,2}(V))^3$ such that

$$(\Delta + \mu^2)\underline{W}(\underline{x}) = \underline{0}, \quad \underline{x} \in V \tag{2.46}$$

$$\operatorname{div} \underline{W}(\underline{x}) = 0, \quad \underline{x} \in \partial C \tag{2.47}$$

$$\operatorname{div} \underline{W}(\underline{x}) = -\operatorname{div}\left(\int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y} \right), \quad \underline{x} \in \partial\Omega \tag{2.48}$$

$$[\underline{n},\underline{W}](\underline{x}) = \underline{0}, \quad \underline{x} \in \partial C \tag{2.49}$$

$$[\underline{n},\underline{W}](\underline{x}) = -\left[\underline{n}(\underline{x}), \int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y} \right], \quad \underline{x} \in \partial\Omega. \tag{2.50}$$

Problem 2.4: Let V,μ^2 be chosen as in Problem 2.3, and $\underline{a} \in (L^2(V))^3$, we seek a vector function $\underline{\tilde{W}} \in (W^{2,2}(V))^3$ such that

$$(\Delta + \mu^2)\underline{\tilde{W}}(\underline{x}) = \underline{B}(\underline{x}), \quad \underline{x} \in V \tag{2.51}$$

$$\operatorname{div} \underline{\tilde{W}}(\underline{x}) = 0, \quad \underline{x} \in \partial V \tag{2.52}$$

$$[\underline{n},\underline{\tilde{W}}](\underline{x}) = \underline{0}, \quad \underline{x} \in \partial V \tag{2.53}$$

where $\underline{B}(\underline{x}) = (\Delta + \mu^2)(\phi(\underline{x})\int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y})$.

We remember that $\phi \in C_0^\infty(\mathbb{R}^3)$ with support in C and $\phi(\underline{x}) = 1$ if $\underline{x} \in \overline{\Omega}$ so that $\underline{B} \in (L^2(V))^3$.

Theorem 2.15: Let $\underline{a} \in (L^2(V))^3$ be a vector function and let μ^2 be not an eigenvalue of the operator T on $K(V)$ then there exists an unique vector function $\underline{W} \in (W^{2,2}(V))^3$ solution of Problem 2.3.

Proof: We note that Theorem 2.12 implies that there exists μ^2 that satisfies the previous conditions. Later we choose μ^2 such that $\operatorname{Im}\mu^2 \neq 0$. From Theorem 2.14 the vector function $\underline{v}(\underline{x}) = \int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y}$ belongs to $(\mathcal{W}^{2,2}(\Omega^c))^3$ hence there exists the trace of $\underline{v}$ on $\partial\Omega$ and $\underline{v}|_{\partial\Omega}$ belongs to $(L^2(\partial\Omega))^3$.

Let us consider the vector function $\underline{\tilde{W}}$ defined by

$$\underline{\tilde{W}}(\underline{x}) = \underline{W}(\underline{x}) + \phi(\underline{x}) \int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y}, \quad \underline{x} \in V. \tag{2.54}$$

It is easy to see that

$$\underline{W}(\underline{x}) = \underline{\tilde{W}}(\underline{x}) - \phi(\underline{x}) \int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y}, \quad \underline{x} \in V \tag{2.55}$$

is solution of Problem 2.3 if and only if $\underline{\tilde{W}}$ is solution of Problem 2.4.

Since μ^2 is not eigenvalue of the operator T defined on $K(V)$ there exists an unique vector function $\underline{\tilde{W}} \in (W^{2,2}(V))^3$ solution of the Problem 2.4. Therefore $\underline{W}(\underline{x}) = \underline{\tilde{W}}(\underline{x}) - \phi(\underline{x})\int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y}$ is solution of Problem 2.3. ■

We observe that $\underline{W}(\underline{x})$ depends with continuity from the function $\underline{v}(\underline{x}) = \int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y}$, that is the operator which to any $\underline{a} \in (L^2(V))^3$ associates $\underline{W} \in (W^{2,2}(V))^3$ given by (2.55) is a continuous operator from $(L^2(V))^3$ to $(W^{2,2}(V))^3$. We seek a solution of Problem 2.2 in the form

$$\underline{\hat{E}}(\underline{x}) = -\int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y} + \phi(\underline{x})\underline{W}(\underline{x}), \quad \underline{x} \in \Omega^c, \tag{2.56}$$

where $\underline{W}$ is the function given in (2.55) extended with $\underline{0}$ in $\Omega^c \setminus C$. That is we want to determine $\underline{a} \in (L^2(V))^3$ in such a way that $\underline{\hat{E}}(\underline{x})$ is a solution of Problem 2.2.

It is easy verify that $\hat{\underline{E}}$ given by (2.56) satisfies the conditions (2.38) and (2.39), moreover since Φ satisfies the Sommerfeld radiation condition (2.11) and $\underline{a}\in(L^2(V))^3$, $\hat{\underline{E}}$ satisfies the radiation condition (2.40). Therefore $\hat{\underline{E}}$ is solution of Problem 2.2 if $\hat{\underline{E}}$ verifies in Ω^c equation (2.37).

Substituting the expression of $\hat{\underline{E}}$ (2.56) in (2.37) we have that Eq. (2.37) is always verified for $\underline{x}\in\mathbb{R}^3\backslash C$, while for $\underline{x}\in V$ we obtain the following equation that $\underline{a}$ must satisfy

$$\underline{a}+(\Delta+k^2)(\phi\underline{W})=\underline{A},\quad \underline{x}\in V. \tag{2.57}$$

We note that, since $\underline{W}$ is solution of Eq. (2.46) we have

$$(\Delta+k^2)(\phi\underline{W})=\underline{W}\Delta\phi+2(\nabla\phi,\nabla)\underline{W}+(k^2-\mu^2)\phi\underline{W},\quad \underline{x}\in V, \tag{2.58}$$

hence Eq. (2.57) becomes

$$(I+\tau)\underline{a}=\underline{A},\quad \underline{x}\in V \tag{2.59}$$

where I is the identity operator and

$$\tau\underline{a}=\underline{W}\Delta\phi+2(\nabla\phi,\nabla)\underline{W}+(k^2-\mu^2)\phi\underline{W},\quad \underline{x}\in V. \tag{2.60}$$

Theorem 2.16: The operator τ: $(L^2(V))^3\rightarrow(L^2(V))^3$ defined in (2.60) is a compact operator.

Proof: From Theorem 2.14 the operator which associates to any vector function $\underline{a}\in(L^2(V))^3$ the vector function $\underline{W}\in(W^{2,2}(V))^3$ given by (2.55) is continuous, hence τ is a continuous operator from $(L^2(V))^3$ to $(W^{1,2}(V))^3$ since it is composition of continuous operators. Rellich's theorem [see Ref. 16, Theorem 6.2, page 144] implies that the immersion

$$i:\ (W^{1,2}(V))^3\rightarrow(L^2(V))^3$$

is compact, therefore

$$\tau:\ (L^2(V))^3\rightarrow(L^2(V))^3$$

is a compact operator. ■

Theorem 2.17: The operator $I+\tau$: $(L^2(V))^3\rightarrow(L^2(V))^3$ where τ is defined by (2.60) is injective.

Proof: Let $\underline{a}\in(L^2(V))^3$ be such that

$$\underline{a}+\underline{W}\Delta\phi+2(\nabla\phi,\nabla)\underline{W}+(k^2-\mu^2)\phi\underline{W}=\underline{0},\quad \underline{x}\in V, \tag{2.61}$$

then

$$\hat{\underline{E}}(\underline{x})=-\int_V\Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y}+\phi(\underline{x})\underline{W}(\underline{x}),\quad \underline{x}\in\Omega^c, \tag{2.62}$$

is solution of

$$(\Delta+k^2)\hat{\underline{E}}(\underline{x})=\underline{0},\quad \underline{x}\in\Omega^c \tag{2.63}$$

and satisfies (2.38), (2.39), and (2.40), hence from Theorem 2.7 $\hat{\underline{E}}=\underline{0}$ in Ω^c. So that we have

$$\int_V\Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y}=\phi(\underline{x})\underline{W}(\underline{x}),\quad \underline{x}\in\Omega^c, \tag{2.64}$$

therefore

$$\underline{v}(\underline{x}) = \int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y} = \underline{0}, \quad \underline{x} \in \mathbb{R}^3 \backslash C. \tag{2.65}$$

Let us define the vector function $\underline{a}(\underline{x})$ as follows:

$$\underline{\alpha}(\underline{x}) = \begin{cases} \underline{v}(\underline{x}), & \underline{x} \in \Omega, \text{ or } \underline{x} \in \mathbb{R}^3 \backslash C \\ \underline{W}(\underline{x}), & \underline{x} \in V \end{cases} \tag{2.66}$$

where $\underline{v}(\underline{x})$ is given by (2.42) and $\underline{W}(\underline{x})$, $\underline{x} \in V$ is given by (2.55). We observe that

$$(\Delta + k^2)\underline{\alpha}(\underline{x}) = (\Delta + k^2)\underline{v}(\underline{x}) = \underline{0}, \quad \underline{x} \in \Omega \tag{2.67}$$

$$(\Delta + \mu^2)\underline{\alpha}(\underline{x}) = (\Delta + \mu^2)\underline{W}(\underline{x}) = \underline{0}, \quad \underline{x} \in V \tag{2.68}$$

$$(\Delta + k^2)\underline{\alpha}(\underline{x}) = (\Delta + k^2)\underline{v}(\underline{x}) = \underline{0}, \quad \underline{x} \in \mathbb{R}^3 \backslash C. \tag{2.69}$$

Let $B_R \supset \overline{C}$, we have

$$0 = \int_{B_R} ((\Delta + \nu^2)\underline{\alpha}(\underline{x}), \overline{\underline{\alpha}(\underline{x})})d\underline{x} = \int_{B_R} (\Delta \underline{\alpha}(\underline{x}), \overline{\underline{\alpha}(\underline{x})})d\underline{x} + \int_{B_R} \nu^2 \sum_{j=1}^{3} |\alpha_j(\underline{x})|^2 \, d\underline{x},$$

where $\nu^2 = k^2$ if $\underline{x} \in \mathbb{R}^3 \backslash V$, $\nu^2 = \mu^2$ if $\underline{x} \in V$, hence

$$-\int_{B_R} \left(\sum_{j=1}^{3} |(\text{curl}\, \underline{\alpha}(\underline{x}))_j|^2 + |\text{div}\, \underline{\alpha}(\underline{x})|^2 \right) d\underline{x} + \int_{B_R} \nu^2 \sum_{j=1}^{3} |\alpha_j(\underline{x})|^2 \, d\underline{x} = 0.$$

Therefore if we have chosen μ^2 such that $\text{Im}\, \mu^2 \neq 0$ we have

$$\int_V \nu^2 \sum_{j=1}^{3} |\alpha_j(\underline{x})|^2 \, d\underline{x} = 0 \tag{2.70}$$

that is

$$\underline{\alpha}(\underline{x}) = \underline{0}, \quad \underline{x} \in V. \tag{2.71}$$

In particular from (2.71) follows that $\underline{W} = \underline{0}$ in V and from (2.61) $\underline{a} = \underline{0}$ in V, therefore $I + \tau$ is an injective operator. ■

We know that since τ is compact if the operator $I + \tau$ is injective there exists the inverse operator $(I + \tau)^{-1}$ and it is continuous. So that Eq. (2.57) has an unique solution belonging to $(L^2(V))^3$. For any $\underline{A} \in (L^2(V))^3$ exists and is unique a function $\underline{a} \in (L^2(V))^3$ such that $(I + \tau)\underline{a} = \underline{A}$ and therefore Problem 2.2 has a solution $\hat{\underline{E}}$ so that using (2.36) and (2.56) it follows that Problem 2.1 has a solution $\underline{E}$ given by

$$\underline{E}(\underline{x}) = -\int_V \Phi(\underline{x},\underline{y})\underline{a}(\underline{y})d\underline{y} + \phi(\underline{x})\underline{W}(\underline{x}) + \phi(\underline{x})\underline{G}(\underline{x}), \quad \underline{x} \in \Omega^c. \tag{2.72}$$

This concludes the existence proof. ■

III. THE COMPUTATIONAL METHOD

We remind that the content of this section is mainly formal. That is no convergence proof is given for the numerical method proposed. We develop our method assuming that the boundary of the obstacle $\partial\Omega$ is a starlike surface with respect to the origin, i.e. [see (1.10)],

$$\partial\Omega=\{\underline{x}=r\underline{\hat{x}}\in\mathbb{R}^3/r=\xi(\underline{\hat{x}}),\underline{\hat{x}}\in\partial B\},\tag{3.1}$$

where $\underline{\hat{x}}=\underline{\hat{x}}(\theta,\phi)$ is given by (1.8) and $\xi(\underline{\hat{x}})$ is a single valued function. Moreover, we assume that the wave number k is a positive real number. A method analogous to the one described here can be obtained substituting the previous choice of the spherical coordinate system with some other coordinate system. Let $0<a<1$, B_a, $\underline{n}(\underline{x})$, $\underline{x}\in\partial\Omega$ be as specified previously, in particular let $\overline{B_a}\subset\Omega$. The computational method proposed is based on the assumption that the solution $\underline{E}$ of the boundary value problem (1.1), (1.2), (1.3), and (1.4) can be extended to $\underline{x}\in\mathbb{R}^3\backslash\overline{B_a}$ and that this extension $\underline{F}$ can be represented as follows:

$$\underline{F}(\underline{x})=a^2\int_{\partial B}\mathrm{curl}_{\underline{x}}\{\Phi(\underline{x},a\underline{\hat{y}})\underline{v}(\underline{\hat{y}})\}ds(\underline{\hat{y}}),\quad \underline{x}\in\mathbb{R}^3\backslash\overline{B_a},\tag{3.2}$$

where ds is the surface measure on ∂B and $\underline{v}(\underline{\hat{y}})$ is a suitable (complex valued) vector density function defined on ∂B, tangential to ∂B, that is such that

$$(\underline{v}(\underline{\hat{x}}),\underline{\hat{x}})=0,\quad \underline{\hat{x}}\in\partial B.\tag{3.3}$$

We note that when $\underline{v}(\underline{x})$ is a complex vector function with (3.3) we mean real euclidean product of complex vectors. It is easy to see that a vector field $\underline{F}(\underline{x})$ given by (3.2) satisfies

$$(\Delta+k^2)\underline{F}(\underline{x})=\underline{0},\quad \underline{x}\in\mathbb{R}^3\backslash\overline{B_a},\tag{3.4}$$

$$\mathrm{div}\,\underline{F}(\underline{x})=0,\quad \underline{x}\in\mathbb{R}^3\backslash\overline{B_a},\tag{3.5}$$

and the radiation condition at infinity (1.4) for any choice of the density $\underline{v}$ that makes possible differentiation under the integral sign in (3.2). Since $\overline{B_a}\subset\Omega$ from (3.4) and (3.5) follows that (1.1) and (1.2) hold. We impose the ''boundary'' condition (1.3) to the vector field $\underline{F}(\underline{x})$ given by (3.2) that is we impose that $\underline{F}(\underline{x})$ extends $\underline{E}(\underline{x})$, so that we have an equation for the density $\underline{v}$, that is

$$\left[\underline{n}(\underline{x}),a^2\int_{\partial B}\mathrm{curl}_{\underline{x}}\{\Phi(\underline{x},a\underline{\hat{y}})\underline{v}(\underline{\hat{y}})\}ds(\underline{\hat{y}})\right]=\underline{g}(\underline{x}),\quad \underline{x}\in\partial\Omega.\tag{3.6}$$

We note that when $\partial\Omega$ is only Lipschitz continuous (3.6) holds only almost everywhere in $\underline{x}\in\partial\Omega$. The numerical solution of (3.6) as an integral equation in the unknown $\underline{v}(\underline{\hat{y}})$ can be carried out using a boundary integral method. The use of a boundary integral method involves the solution of a computationally very expensive linear system. Instead of using a boundary integral method we solve equation (3.6) with a formal power series using $\partial B=\{\underline{x}\in\mathbb{R}^3|\|\underline{x}\|=1\}$ as base point, that is we look for a solution of (3.6) given by a formal series expansion of the form

$$\underline{v}(\underline{\hat{y}})=\sum_{s=0}^{+\infty}\frac{(\xi(\underline{\hat{y}})-1)^s}{s!}\underline{v}_s(\underline{\hat{y}}),\quad \underline{\hat{y}}\in\partial B,\tag{3.7}$$

where we assume $0!=1$ and the ''coefficients'' $\underline{v}_s(\underline{\hat{y}})$, $\underline{\hat{y}}\in\partial B$, $s=0,1,2,\ldots$, are vector fields tangential to ∂B to be determined. This perturbative procedure is preferable to the solution of the linear system coming from the use of a boundary integral method since the coefficients $\underline{v}_s$, $s=0,1,\ldots$, are defined by recursive formulas involving double integrals. These integrals are independent one from the other so that they can be computed in parallel. We note that $\underline{v}(\underline{\hat{y}})$ given by (3.7) is only an auxiliary unknown and that $\underline{F}(\underline{x})$ can be determined knowing only the action of some appropriate operators on $\underline{v}_s(\underline{\hat{y}})$ without knowing $\underline{v}_s(\underline{\hat{y}})$, $s=0,1,2,\ldots$. We note that the computational method proposed here relies on the assumption $a<1$ that is $B_a\subset B$ since ∂B is chosen as base point of the ''power series'' expansion (3.7). The computational method proposed here can be developed with only minor changes using for example the surface ∂B_R, $R>0$ with $R\neq 1$ as base point of the expansion (in this case $a<R$) or more in general any starlike surface with respect to the origin that defines a bounded region that contains B_a. The choice ∂B_R, $R>a>0$, is more convenient than the choice of a generic starlike surface since for ∂B_R some explicit class of functions such as the vector spherical harmonics and some useful formulas about them are available. In the numerical experience described in Sec. IV we always use ∂B_R as base point, the radius R is chosen in such a way that $\max_{\hat{\underline{x}}\in\partial B}|\xi(\hat{\underline{x}})-R|$ is approximately minimized to improve the ''convergence'' of the series (3.7).

Let $\underline{f}(\hat{\underline{x}})$ be a vector field defined on ∂B and tangential to ∂B; we define the operators p_s, $\hat{p}_s$, $s=0,1,2,\ldots$ as follows:

$$(p_s\underline{f})(\underline{x})=a^2\int_{\partial B}\frac{\partial^s}{\partial\|\underline{x}\|^s}\mathrm{curl}_{\underline{x}}\{\Phi(\underline{x},a\hat{\underline{y}})\underline{f}(\hat{\underline{y}})\}ds(\hat{\underline{y}}),\quad \|\underline{x}\|>a,\quad s=0,1,\ldots, \tag{3.8}$$

$$(\hat{p}_s\underline{f})(\hat{\underline{x}})=(p_s\underline{f})(\hat{\underline{x}}),\quad \hat{\underline{x}}\in\partial B,\quad s=0,1,\ldots. \tag{3.9}$$

For later use we define the operator $\hat{l}_0$ which acts on a vector field $\underline{f}$ defined on ∂B and tangential to ∂B through the following equation:

$$[\hat{\underline{x}},(\hat{p}_0(\hat{l}_0\underline{f}))(\hat{\underline{x}})]=\underline{f}(\hat{\underline{x}}),\quad \hat{\underline{x}}\in\partial B, \tag{3.10}$$

where $\hat{p}_0$ is the operator given in (3.9) when $s=0$. We note that Eq. (3.10) is a special case of Eq. (3.6), that is the case when $\partial\Omega=\partial B$. In Lemma 3.4 under some extra hypotheses a formula to solve Eq. (3.10) is given. Substituting (3.7) in (3.2) using the power series expansion of $\mathrm{curl}_{\underline{x}}\{\Phi(\underline{x},a\hat{\underline{y}})\underline{v}(\hat{\underline{y}})\}$ as a function of $\|\underline{x}\|$ with base point $\|\underline{x}\|=1$ and using (3.8) and (3.9) we obtain the formal expansion of $\underline{F}(\underline{x})$, $\underline{x}\in\mathbb{R}^3\setminus\overline{B_a}$:

$$\underline{F}(\underline{x})=\sum_{s=0}^{+\infty}\sum_{l=0}^{s}\frac{(\|\underline{x}\|-1)^{s-l}}{(s-l)!}\left(\hat{p}_{s-l}\left(\underline{v}_l\frac{(\xi-1)^l}{l!}\right)\right)(\hat{\underline{x}}),\quad \underline{x}=\|\underline{x}\|\hat{\underline{x}},\quad \underline{x}\in\mathbb{R}^3\setminus\overline{B_a}. \tag{3.11}$$

From (3.11) and Theorem 2.5 we obtain the formal series expansion of the far field $\underline{F}^{(0)}(\hat{\underline{x}})$ associated to $\underline{F}(\underline{x})$, in ''powers'' of $(\xi-1)$, that is

$$\underline{F}^{(0)}(\hat{\underline{x}})=\frac{a^2\imath k}{4\pi}\sum_{l=0}^{+\infty}\int_{\partial B}\exp(-\imath ka(\hat{\underline{x}},\hat{\underline{y}}))\left[\hat{\underline{x}},\underline{v}_l(\hat{\underline{y}})\frac{(\xi(\hat{\underline{y}})-1)^l}{l!}\right]ds(\hat{\underline{y}}),\quad \hat{\underline{x}}\in\partial B. \tag{3.12}$$

The problem of determining a formal power series to compute $\underline{F}(\underline{x})$ solution of (1.1), (1.2), (1.3), and (1.4) or the corresponding far field $\underline{F}^{(0)}(\hat{\underline{x}})$ is reduced to the problem of determining the coefficients $\underline{v}_s(\hat{\underline{y}})$, $s=0,1,2,\ldots$, of (3.7) using Eq. (3.6). Let

$$\hat{\underline{x}}_\theta=\frac{\partial\hat{\underline{x}}}{\partial\theta}(\theta,\phi)=(\cos\theta\cos\phi,\cos\theta\sin\phi,-\sin\theta)^T,\quad 0\leqslant\theta\leqslant\pi,\quad 0\leqslant\phi<2\pi, \tag{3.13}$$

$$\hat{\underline{x}}_\phi=\frac{\partial\hat{\underline{x}}}{\partial\phi}(\theta,\phi)=(-\sin\theta\sin\phi,\sin\theta\cos\phi,0)^T,\quad 0\leqslant\theta\leqslant\pi,\quad 0\leqslant\phi<2\pi, \tag{3.14}$$

and when $\xi(\hat{\underline{x}})$ is sufficiently regular we define

$$\begin{gathered}\tilde{\underline{n}}(\underline{x})=\hat{\underline{x}}(\theta,\phi)-\frac{\hat{\underline{x}}_\theta(\theta,\phi)}{\xi(\hat{\underline{x}}(\theta,\phi))}\frac{\partial\xi}{\partial\theta}(\hat{\underline{x}}(\theta,\phi))-\frac{1}{\sin^2\theta}\frac{\hat{\underline{x}}_\phi(\theta,\phi)}{\xi(\hat{\underline{x}}(\theta,\phi))}\frac{\partial\xi}{\partial\phi}(\hat{\underline{x}}(\theta,\phi)),\\ \underline{x}=\xi(\hat{\underline{x}}(\theta,\phi))\hat{\underline{x}}(\theta,\phi)\in\partial\Omega,\quad 0\leqslant\theta\leqslant\pi,\quad 0\leqslant\phi<2\pi\end{gathered} \tag{3.15}$$

so that we have

$$\underline{n}(\underline{x}) = \frac{\underline{\tilde{n}}(\underline{x})}{\|\underline{\tilde{n}}(\underline{x})\|}, \quad \underline{x} \in \partial\Omega, \tag{3.16}$$

where $\underline{n}(\underline{x})$ is the outward unit normal vector to $\partial\Omega$ in $\underline{x} \in \partial\Omega$. When we consider obstacles Ω with $\partial\Omega$ only locally Lipschitz we assume that formulas (3.15) and (3.16) are valid almost everywhere $\underline{x}$ for $\underline{x} \in \partial\Omega$. For example, this assumption is satisfied when Ω is a polyhedron, that is one of the choices considered in Sec. IV. For simplicity in the following abusing the notation we use ξ as an independent variable, so that $O((\xi-1)^l)$, when $\xi \to 1$ becomes a meaningful notation. It is easy to see that

$$\underline{\tilde{n}}(\underline{x}) = \sum_{l=0}^{+\infty} \underline{n}_l(\hat{\underline{x}}), \quad \underline{x} = \|\underline{x}\|\hat{\underline{x}} \in \partial\Omega,$$

where

$$\underline{n}_0(\hat{\underline{x}}) = \hat{\underline{x}}(\theta,\phi), \quad \hat{\underline{x}}\partial B, \tag{3.17}$$

$$\underline{n}_1(\hat{\underline{x}}) = -\left(\frac{\partial \xi}{\partial \theta}(\hat{\underline{x}}(\theta,\phi))\hat{\underline{x}}_\theta(\theta,\phi) + \frac{1}{\sin^2\theta}\frac{\partial \xi}{\partial \phi}(\hat{\underline{x}}(\theta,\phi))\hat{\underline{x}}_\phi(\theta,\phi)\right), \quad \hat{\underline{x}} \in \partial B \tag{3.18}$$

$$\underline{n}_l(\hat{\underline{x}}) = (-1)^{l-1}(\xi(\hat{\underline{x}}(\theta,\phi))-1)^{l-1}\underline{n}_1(\hat{\underline{x}}), \quad l=2,3,\ldots, \quad \hat{\underline{x}} \in \partial B, \tag{3.19}$$

and $\underline{n}_l(\hat{\underline{x}}) = O((\xi(\hat{\underline{x}})-1)^l)$, $l=0,1,2,\ldots$, when $\xi \to 1$, $\hat{\underline{x}} \in \partial B$. We remark that the expansion in ''powers'' of $(\xi-1)$ that we present is only formal and we consider the terms $\partial\xi/\partial\theta, \partial\xi/\partial\phi$ to be $O((\xi-1))$ when $\xi \to 1$ in fact formally $\partial\xi/\partial\theta = \partial/\partial\theta(\xi-1)$ and $\partial\xi/\partial\phi = (\partial/\partial\phi)(\xi-1)$. Since the vector field $\underline{g}$ in (1.3) is tangential to $\partial\Omega$ we can write $\underline{g}$ as follows:

$$\underline{g}(\underline{x}) = -[\underline{n},\underline{b}](\underline{x}), \quad \underline{x} \in \partial\Omega, \tag{3.20}$$

for a suitable choice of the vector field $\underline{b}(\underline{x})$. For example, $\underline{b}(\underline{x}) = [\underline{n},\underline{g}](\underline{x})$, $\underline{x} \in \partial\Omega$ is a possible choice. Using (3.16) and (3.20), Eq. (3.6) can be rewritten as follows:

$$\left[\underline{\tilde{n}}(\underline{x}), a^2 \int_{\partial B} \mathrm{curl}_{\underline{x}}\{\Phi(\underline{x}, a\hat{\underline{y}})\underline{v}(\hat{\underline{y}})\} ds(\hat{\underline{y}})\right] = -[\underline{\tilde{n}},\underline{b}](\underline{x}), \quad \underline{x} \in \partial\Omega. \tag{3.21}$$

Lemma 3.1: Let n_l, $l=0,1,2,\ldots$, be given by (3.17), (3.18), and (3.19), let $\underline{b}$ be as in (3.20) and let $\underline{\tilde{b}}(\hat{\underline{x}}) = \underline{b}(\xi(\hat{\underline{x}})\hat{\underline{x}})$, $\hat{\underline{x}} \in \partial B$, let $\underline{F}$, given by (3.2), be an extension of the solution of (1.1), (1.2), (1.3), and (1.4) and let $\underline{F}(\xi(\hat{\underline{x}})\hat{\underline{x}}) = \Sigma_{l=0}^{+\infty}\underline{F}_l(\hat{\underline{x}})$, with $\underline{F}_l(\hat{\underline{x}}) = O((\xi(\hat{\underline{x}})-1)^l)$, when $\xi \to 1$, $l=0,1,2,\ldots$, then the following recursive formulas for $\underline{F}_l$, $l=0,1,2,\ldots$ hold:

$$[\underline{n}_0,\underline{F}_0](\hat{\underline{x}}) = -[\underline{n}_0,\underline{\tilde{b}}](\hat{\underline{x}}), \quad \hat{\underline{x}} \in \partial B, \tag{3.22}$$

$$[\underline{n}_0,\underline{F}_\nu](\hat{\underline{x}}) = -(\xi-1)^{\nu-1}(-1)^{\nu-1}[\underline{n}_1,\underline{\tilde{b}}](\hat{\underline{x}}) - \sum_{s=0}^{\nu-1}(\xi-1)^{\nu-s-1}(-1)^{\nu-s-1}[\underline{n}_1,\underline{F}_s](\hat{\underline{x}}),$$
$$\nu=1,2,\ldots,\hat{\underline{x}} \in \partial B. \tag{3.23}$$

We observe that formulas (3.22) and (3.23) are compatible if in fact we have

$$\left(\underline{n}_0(\hat{\underline{x}}), \sum_{s=0}^{\nu-1}(\xi-1)^{\nu-s-1}(-1)^{\nu-s-1}[\underline{n}_1,\underline{F}_s](\hat{\underline{x}}) + (\xi-1)^{\nu-1}(-1)^{\nu-1}[\underline{n}_1\underline{\tilde{b}}](\hat{\underline{x}})\right) = 0,$$
$$\nu=1,2,\ldots, \quad \hat{\underline{x}} \in \partial B. \tag{3.24}$$

Proof: Imposing (3.21) order by order in ''powers'' of $(\xi-1)$ we have equations (3.22) and (3.23). We prove (3.24) by induction on ν. From (3.22) we have

$$(\underline{n}_1(\underline{\hat{x}}),[\underline{n}_0,\underline{F}_0+\underline{\tilde{b}}](\underline{\hat{x}}))=0, \quad \underline{\hat{x}}\in\partial B$$

so that when $\nu=1$ (3.24) holds in fact:

$$(\underline{n}_0(\underline{\hat{x}}),[\underline{n}_1,\underline{F}_0+\underline{\tilde{b}}](\underline{\hat{x}}))=(\underline{n}_1(\underline{\hat{x}}),[\underline{n}_0,\underline{F}_0+\underline{\tilde{b}}](\underline{\hat{x}})).$$

Now we assume by induction that formula (3.24) holds when $\nu=k$ and we prove that (3.24) holds when $\nu=k+1$. We have

$$\begin{aligned}&\left(\underline{n}_0(\underline{\hat{x}}),\sum_{s=0}^{k}(\xi-1)^{k-s}(-1)^{k-s}[\underline{n}_1,\underline{F}_s](\underline{\hat{x}})+(\xi-1)^k(-1)^k[\underline{n}_1,\underline{\tilde{b}}](\underline{\hat{x}})\right)-(\xi-1)\\&\quad\times\left(\underline{n}_0(\underline{\hat{x}}),\sum_{s=0}^{k-1}(\xi-1)^{k-s-1}(-1)^{k-s-1}[\underline{n}_1,\underline{F}_s](\underline{\hat{x}})+(\xi-1)^{k-1}(-1)^{k-1}[\underline{n}_1,\underline{\tilde{b}}](\underline{\hat{x}})\right)\\&\quad+(\underline{n}_0(\underline{\hat{x}}),[\underline{n}_1,\underline{F}_k](\underline{\hat{x}}))=(\underline{n}_1(\underline{\hat{x}}),[\underline{n}_0,\underline{F}_k](\underline{\hat{x}})), \quad \underline{\hat{x}}\in\partial B, \end{aligned} \tag{3.25}$$

so that the thesis follows using formula (3.23) for $\underline{F}_k$. ■

Now we can write the formulas which give the coefficients $\underline{v}_s$, $s=0,1,2,...$ of (3.7) using the operator $\hat{l}_0$ defined in (3.10). In Lemma 3.4 we give an explicit formula for $\hat{l}_0$.

Theorem 3.2: Let $\underline{\tilde{b}}$ be as in Lemma 3.1 and $\underline{v}$ be a vector field defined on ∂B satisfying (3.3) such that the vector field $\underline{F}$, given by (3.2), coincides with the solution of the boundary value problem (1.1), (1.2), (1.3), and (1.4) in $\mathbb{R}^3\backslash\Omega$. Then formulas (3.7) and (3.11) hold with

$$\frac{(\xi-1)^s}{s!}\underline{v}_s(\underline{\hat{x}})=(\hat{l}_0\underline{h}_s)(\underline{\hat{x}}), \quad s=0,1,..., \quad \underline{\hat{x}}\in\partial B, \tag{3.26}$$

where

$$\underline{h}_0(\underline{\hat{x}})=-[\underline{n}_0,\underline{\tilde{b}}](\underline{\hat{x}}), \quad \underline{\hat{x}}\in\partial B, \tag{3.27}$$

$$\begin{aligned}\underline{h}_s(\underline{\hat{x}})=&-\left[\underline{n}_0,\sum_{m=0}^{s-1}\frac{(\xi-1)^{s-m}}{(s-m)!}\left(\hat{p}_{s-m}\left(\underline{v}_m\frac{(\xi-1)^m}{m!}\right)\right)\right](\underline{\hat{x}})-\sum_{\nu=0}^{s-1}(\xi(\underline{\hat{x}})-1)^{s-1-\nu}(-1)^{s-1-\nu}\\&\times\left[\underline{n}_1,\sum_{l=0}^{\nu}\frac{(\xi-1)^{\nu-l}}{(\nu-l)!}\left(\hat{p}_{\nu-l}\left(\underline{v}_l\frac{(\xi-1)^l}{l!}\right)\right)\right](\underline{\hat{x}})-(-1)^{s-1}(\xi(\underline{\hat{x}})-1)^{s-1}[\underline{n}_1,\underline{\tilde{b}}](\underline{\hat{x}}),\\&\underline{\hat{x}}\in\partial B, \quad s=1,2,..., \end{aligned} \tag{3.28}$$

where $\hat{p}_s$, $s=0,1,...$, and $\hat{l}_0$ are the operators defined in (3.9) and (3.10).

Proof (formal): It is easy to see that the vector fields $\underline{h}_s$, $s=0,1,2,...$ are tangential to ∂B. The proof follows from the integral representation formula (3.2), Lemma 3.1, and Eq. (3.21) using the Cauchy rule in the product of the series expansion in powers of $(\xi-1)$ of $\underline{\tilde{n}}$ and of $\underline{F}$ given by (3.11). ■

Now we give some Lemmas that will be used in Sec. IV.

Let $\{\underline{B}_{\sigma,m,l}(\underline{\hat{x}}),\underline{C}_{\sigma,m,l}(\underline{\hat{x}}),\underline{P}_{\sigma,m,l}(\underline{\hat{x}})\}_{\sigma=0,1,l=\sigma,\sigma+1,...,m=\sigma,...,l}$ be the complete orthonormal set of $(L^2(\partial B))^3$ made of vector spherical harmonics functions (see Ref. 17, Chap. 13, page 1898). Since $\partial\Omega$ is a starlike surface with respect to the origin and we use an expansion with ∂B as base point, we can expand data and unknown of Eq. (3.21) with respect to this set of vector functions.

This choice makes easy the computation of the coefficients $\underline{v}_s$, $s=0,1,2,\ldots$, of (3.7). In fact in Lemmas 3.3 and 3.4 we show that the operators $\hat{l}_0$, $\hat{p}_s$, $s=0,1,\ldots$, are represented by simple matrices in this basis.

Lemma 3.3: Let $\underline{f}$ be a vector field defined on ∂B and tangential to ∂B, i.e.,

$$(\underline{f}(\underline{\hat{x}}),\underline{\hat{x}})=0, \quad \underline{\hat{x}}\in\partial B. \tag{3.29}$$

Moreover let $\underline{f}$ be given by the following by the following expansion:

$$\underline{f}(\underline{\hat{x}})=\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{+\infty}\sum_{m=\sigma}^{l}\{f_{C_{\sigma,m,l}}\underline{C}_{\sigma,m,l}(\underline{\hat{x}})+f_{B_{\sigma,m,l}}\underline{B}_{\sigma,m,l}(\underline{\hat{x}})\}, \quad \underline{\hat{x}}\in\partial B, \tag{3.30}$$

where $\{f_{C_{\sigma,m,l}},f_{B_{\sigma,m,l}}\}$, $\sigma=0,1$, $l=\sigma,\sigma+1,\ldots$, $m=\sigma,\sigma+1,\ldots,l$ are the generalized Fourier coefficients of $\underline{f}$ and let $\hat{p}_s$, $s=0,1,2,\ldots$, be the operators given by (3.9), then we have

$$\begin{aligned}(\hat{p}_s\underline{f})(\underline{\hat{x}})=\iota k^2a^2\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{+\infty}\sum_{m=\sigma}^{l}\Bigg\{&\underline{P}_{\sigma,m,l}(\underline{\hat{x}})f_{C_{\sigma,m,l}}j_l(ka)\sqrt{l(l+1)}\left(k^s\sum_{q=0}^{s}\binom{s}{q}(-1)^q q!\,\frac{h_l^{(s-q)}(k)}{k^{q+1}}\right)\\&+\underline{B}_{\sigma,m,l}(\underline{\hat{x}})j_l(ka)f_{C_{\sigma,m,l}}\left((l+1)k^s\sum_{q=0}^{s}\binom{s}{q}(-1)^q q!\,\frac{h_l^{(s-q)}(k)}{k^{q+1}}-k^sh_{l+1}^{(s)}(k)\right)\\&+k^s\underline{C}_{\sigma,m,l}(\underline{\hat{x}})h_l^{(s)}(k)\left(\frac{(l+1)j_l(ka)}{ka}-j_{l+1}(ka)\right)f_{B_{\sigma,m,l}}\Bigg\}, \quad \underline{\hat{x}}\in\partial B,\end{aligned} \tag{3.31}$$

where $h_l(z)$, $j_l(z)$, $l=0,1,2,\ldots$ are the spherical Hankel and the spherical Bessel functions, respectively.

Proof: The proof follows by an easy computation using the so called Green's dyadic for free space (see Ref. 17, Chap. 13, page 1875). ■

Lemma 3.4: Let $\underline{f}$ be a vector field defined on ∂B such that (3.29) and (3.30) hold, let $\hat{l}_0$ be the operator given by (3.10) and let the wave number k be such that

$$j_l(ka)\neq 0, \quad l=0,1,2,\ldots, \tag{3.32}$$

$$((l+1)j_l(ka)-kaj_{l+1}(ka))\neq 0, \quad l=0,1,2,\ldots, \tag{3.33}$$

then we have

$$\begin{aligned}(\hat{l}_0\underline{f})(\underline{\hat{x}})=\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{+\infty}\sum_{m=\sigma}^{l}\Bigg\{&\frac{\underline{B}_{\sigma,m,l}(\underline{\hat{x}})f_{B_{\sigma,m,l}}}{\iota kah_l(k)((l+1)j_l(ka)-kaj_{l+1}(ka))}\\&-\frac{\underline{C}_{\sigma,m,l}(\underline{\hat{x}})f_{C_{\sigma,m,l}}}{\iota ka^2j_l(ka)((l+1)h_l(k)-kh_{l+1}(k))}\Bigg\}, \quad \underline{\hat{x}}\in\partial B.\end{aligned} \tag{3.34}$$

Proof: It is easy to see that $h_l(k)\neq 0$ and $(l+1)h_l(k)-kh_{l+1}(k))\neq 0$, $l=0,1,\ldots$, for k real and positive by virtue of the properties of the spherical Hankel functions (see Ref. 18, page 439). The proof follows using (3.31) with $s=0$, Eq. (3.10) and the fact that

$$(\underline{n}_0(\underline{\hat{x}}),\underline{P}_{\sigma,m,l}(\underline{\hat{x}}))=0, \ \ \underline{\hat{x}}\in\partial B, \ \ \sigma=0,1, \ \ l=\sigma,\sigma+1,\ldots, \ \ m=\sigma,\sigma+1,\ldots,l.$$

■

Remark: We observe that the conditions (3.32) and (3.33) are necessary due to the choices made in the factorization of the operators involved in Eq. (3.10). A different factorization will make these conditions not necessary. This will appear more clearly in Lemma 3.5. Moreover it is

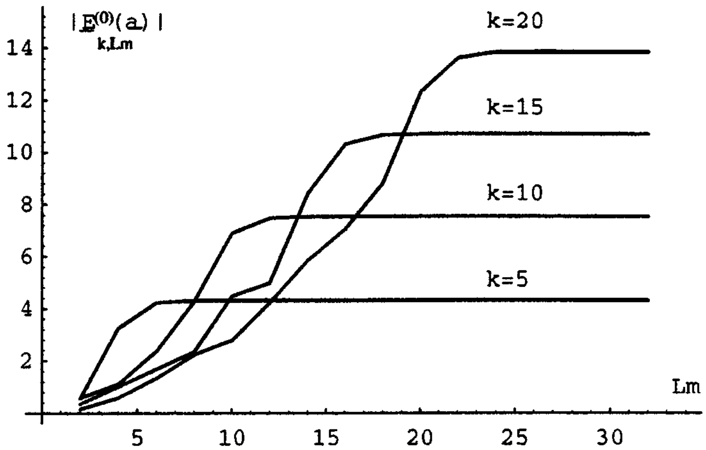

FIG. 1. $|\underline{E}_{k,L_m}(\underline{a})|$ as function of L_m and k relative to the sphere of radius $R=1$.

always possible to choose the radius a such that (3.32) and (3.33) hold, since the zeros of (3.32) and (3.33) are a measure zero set in the interval $\{a\in\mathbb{R}|0<a<1\}$. Finally in practice when using the computational method it is not necessary to choose the radius a to satisfy (3.32) and (3.33) in fact (3.32) and (3.33) are always satisfied numerically.

Lemma 3.5: Let $\hat{p}_s$, $s=0,1,2,\ldots,\hat{l}_0$ be the operators given by (3.9) and (3.10), respectively, and let $\underline{f}$ be a vector field defined on ∂B such that (3.29) and (3.30) hold, then we have

$$(\hat{p}_s(\hat{l}_0\underline{f}))(\underline{\hat{x}})=\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{+\infty}\sum_{m=\sigma}^{l}\Bigg\{-\underline{P}_{\sigma,m,l}(\underline{\hat{x}})\frac{f_{C_{\sigma,m,l}}k\sqrt{l(l+1)}}{((l+1)h_l(k)-kh_{l+1}(k))}$$
$$\times\left(k^s\sum_{q=0}^{s}\binom{s}{q}(-1)^q q!\,\frac{h_l^{(s-q)}(k)}{k^{q+1}}\right)-\underline{B}_{\sigma,m,l}(\underline{\hat{x}})\frac{f_{C_{\sigma,m,l}}k}{((l+1)h_l(k)-kh_{l+1}(k))}$$
$$\times\left((l+1)k^s\sum_{q=0}^{s}\binom{s}{q}(-1)^q q!\,\frac{h_l^{(s-q)}(k)}{k^{q+1}}-k^sh_{l+1}^{(s)}(k)\right)$$
$$+k^s\underline{C}_{\sigma,m,l}(\underline{\hat{x}})h_l^{(s)}(k)\frac{f_{B_{\sigma,m,l}}}{h_l(k)}\Bigg\},\quad \underline{\hat{x}}\in\partial B,\quad s=0,1,2,\ldots,\tag{3.35}$$

where $h_l(z)$, $j_l(z)$, $l=0,1,2,\ldots$ are the spherical Hankel and the spherical Bessel functions, respectively.

Proof: It follows using (3.31) and (3.34) by an easy computation. ■

We note that substituting formula (3.35) into (3.28) we can compute $\underline{h}_s$, given by (3.28), by a recursive formula involving only the vector fields $\underline{h}_\nu$, $\nu=0,1,\ldots,s-1$ and not $\underline{v}_\nu(\xi-1)^\nu/\nu!$, $\nu=0,1,\ldots,s-1$, by doing so we avoid the conditions (3.32), and (3.33) since $\hat{l}_0$ is never considered standing alone but always appears in products such as $\hat{p}_s\hat{l}_0$, $s=0,1,2,\ldots$.

TABLE I. Computational cost.

	T-matrix	Method of Sec. III up to order s
Number of integrals	$(L_m+1)^4$	$\frac{3}{2}s(s+1)(L_m+1)^2$

TABLE II. Accuracy of the far field computed with the perturbation series.

	$\epsilon_{L_2}^{\|E^{(0)}\|}$		
k	Sphere	Ellipsoid	Platelet
2	9.22e-05	2.68e-05	2.11e-05
4	4.00e-04	7.12e-04	6.80e-04
6	1.13e-03	4.13e-03	1.11e-02
8	2.72e-03	2.47e-02	2.44e-02
10	6.53e-03	9.91e-02	9.67e-02

The possibility of computing $\underline{h}_s$ using a recursive formula involving only $\underline{h}_\nu$, $\nu=0,1,...,(s-1)$ [i.e., Eq. (3.28)] is the basis of the computational cost estimates given in Sec. IV (see Table I).

Lemma 3.6: Let $\underline{h}_s(\hat{\underline{x}})$, $\hat{\underline{x}}\in\partial B$, $s=0,1,2,...$ be the vector field given by (3.28) and let $\underline{h}_s$ be given by the following expansion:

$$\underline{h}_s(\hat{\underline{x}})=\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{+\infty}\sum_{m=\sigma}^{l}\{h_{s,B_{\sigma,m,l}}\underline{B}_{\sigma,m,l}(\hat{\underline{x}})+h_{s,C_{\sigma,m,l}}\underline{C}_{\sigma,m,l}(\hat{\underline{x}})\},\quad s=0,1,...,\quad \hat{\underline{x}}\in\partial B, \tag{3.36}$$

where $h_{s,B_{\sigma,m,l}}$, $h_{s,C_{\sigma,m,l}}$, $s=0,1,...$, $\sigma=0,1$, $l=\sigma$, $\sigma+1,...$, $m=\sigma$, $\sigma+1,...,l$, are the generalized Fourier coefficients of $\underline{h}_s(\hat{x})$, then the far field $\underline{F}^{(0)}(\hat{x})$ given by (3.11) has the following expansion:

$$\underline{F}^{(0)}(\hat{\underline{x}})=\sum_{s=0}^{+\infty}\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{+\infty}\sum_{m=\sigma}^{l}\left\{\frac{\underline{B}_{\sigma,m,l}(\hat{\underline{x}})h_{s,C_{\sigma,m,l}}}{\iota^{l+2}((l+1)h_l(k)-kh_{l+1}(k))}+\frac{\underline{C}_{\sigma,m,l}(\hat{\underline{x}})h_{s,B_{\sigma,m,l}}}{\iota^{l+1}kh_l(k)}\right\},\quad \hat{\underline{x}}\in\partial B. \tag{3.37}$$

Proof (formal): The proof follows using the integral representation formula (3.2), the Green's dyadic (Ref. 17, Chap. 13, page 1874) and the following asymptotic expansion of the spherical Hankel functions:

$$h_l(k\|\underline{x}\|)=\iota^{-(l+1)}\frac{\exp(\iota k\|\underline{x}\|)}{k\|\underline{x}\|}+O\left(\frac{1}{(k\|\underline{x}\|)^2}\right),\quad \|\underline{x}\|\to+\infty,\quad l=0,1,\dots. \tag{3.38}$$

■

IV. SOME COMPUTATIONAL RESULTS

We apply the computational method proposed in Sec. III to the electromagnetic scattering problem of Sec. I. That is we consider the scattering problem associated to a linearly polarized electromagnetic plane wave that hits the obstacle Ω. We restrict our attention to the study of the far field pattern associated to the scattered field that is to the computation of (3.37). Let $\underline{E}^i$ be the space dependent part of the electric field associated to the incoming plane wave, we have

$$\underline{E}^i(\underline{x})=\underline{P}\exp(\iota k(\underline{x},\underline{a})), \tag{4.1}$$

where $\underline{P}$, $\underline{a}\in\mathbb{R}^3$, $\|\underline{a}\|=1$ are given and k is the wave number. We assume k to be a positive real number. The vector $\underline{a}$ is the propagation direction of the plane wave and $\underline{P}$ is the polarization vector of the plane wave. We assume

$$\operatorname{div}\underline{E}^i(\underline{x})=\iota k(\underline{P},\underline{a})\exp(\iota k(\underline{x},\underline{a}))=0, \tag{4.2}$$

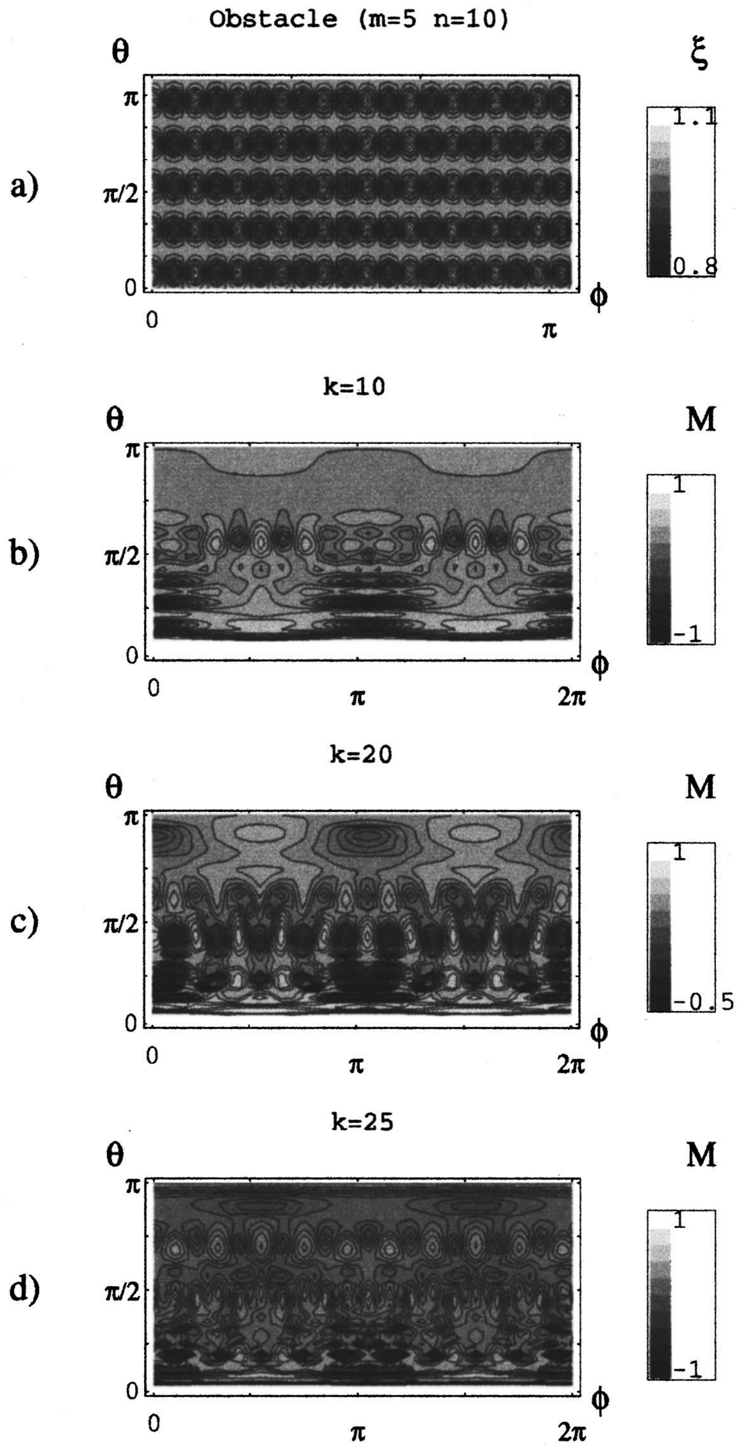

FIG. 2. (a) Contour plot of the corrugated sphere for $m=5$, $n=10$; (b), (c), (d) contour plots of $M(\theta,\phi)$ relative to the geometry for different values of k.

that is $\underline{P}$ and $\underline{q}$ are orthogonal. Let α, β, γ be such that $0\leqslant\alpha\leqslant\pi, 0\leqslant\beta<2\pi, -\pi/2\leqslant\gamma\leqslant\pi/2$ we can parametrize the vectors $\underline{q}$, $\underline{P}$ as follows:

$$\underline{q}=(\sin\alpha\cos\beta,\sin\alpha\sin\beta,\cos\alpha)^T, \tag{4.3}$$

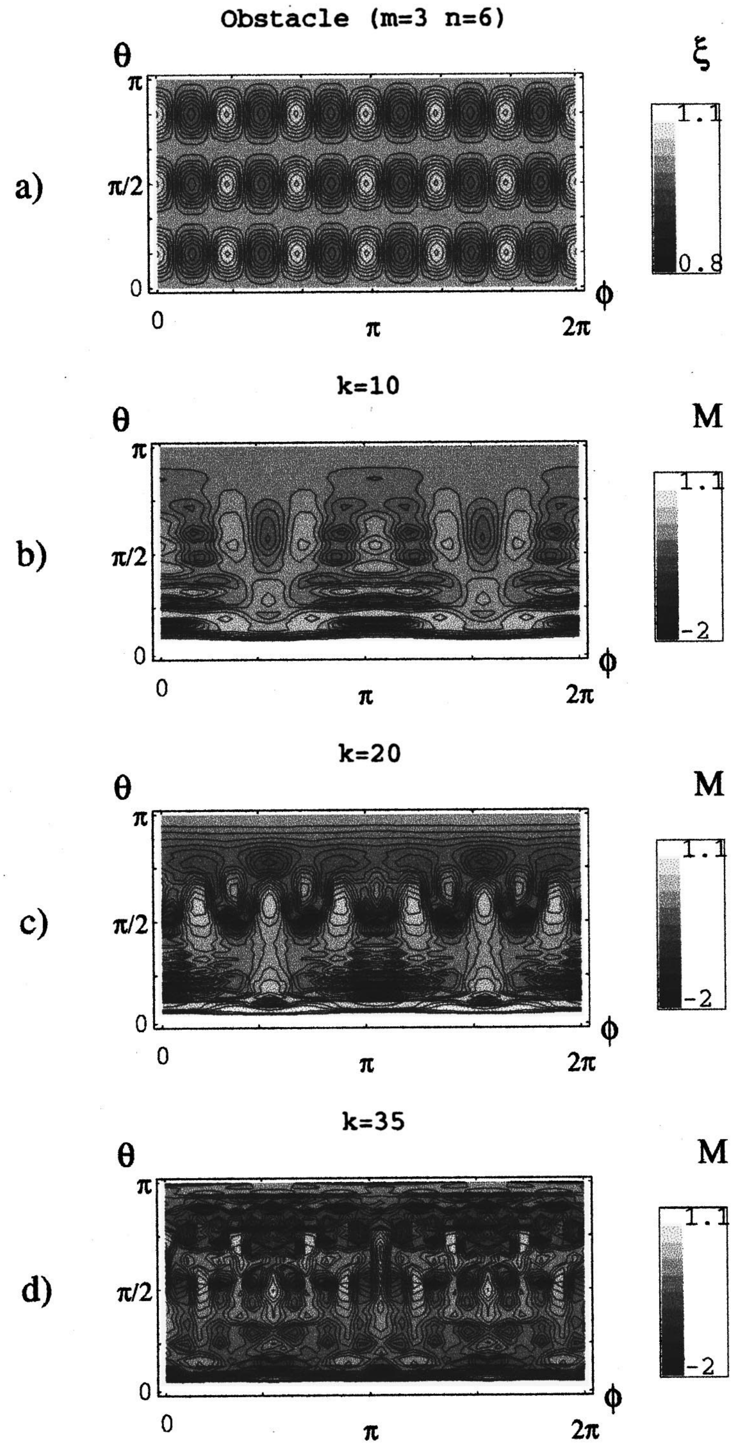

FIG. 3. (a) Contour plot of the corrugated sphere for $m=3$, $n=6$; (b), (c), (d) contour plots of $M(\theta,\phi)$ relative to the geometry for different values of k.

$$\underline{P}=(-\sin\gamma\sin\beta+\cos\gamma\cos\alpha\cos\beta,\sin\gamma\cos\beta+\cos\gamma\cos\alpha\sin\beta,-\cos\gamma\sin\alpha)^{T}. \quad (4.4)$$

We note that when the incoming wave is given by (4.1), the zero-order term of the expansion in powers of $\xi-1$ of the far field $\underline{E}^{(0)}(\hat{\underline{x}})$ associated to the scattering problem (1.1), (1.2), (1.3), and (1.4) with $\underline{g}$ given by (1.3) has an explicit formula given by a series of vector spherical harmonics

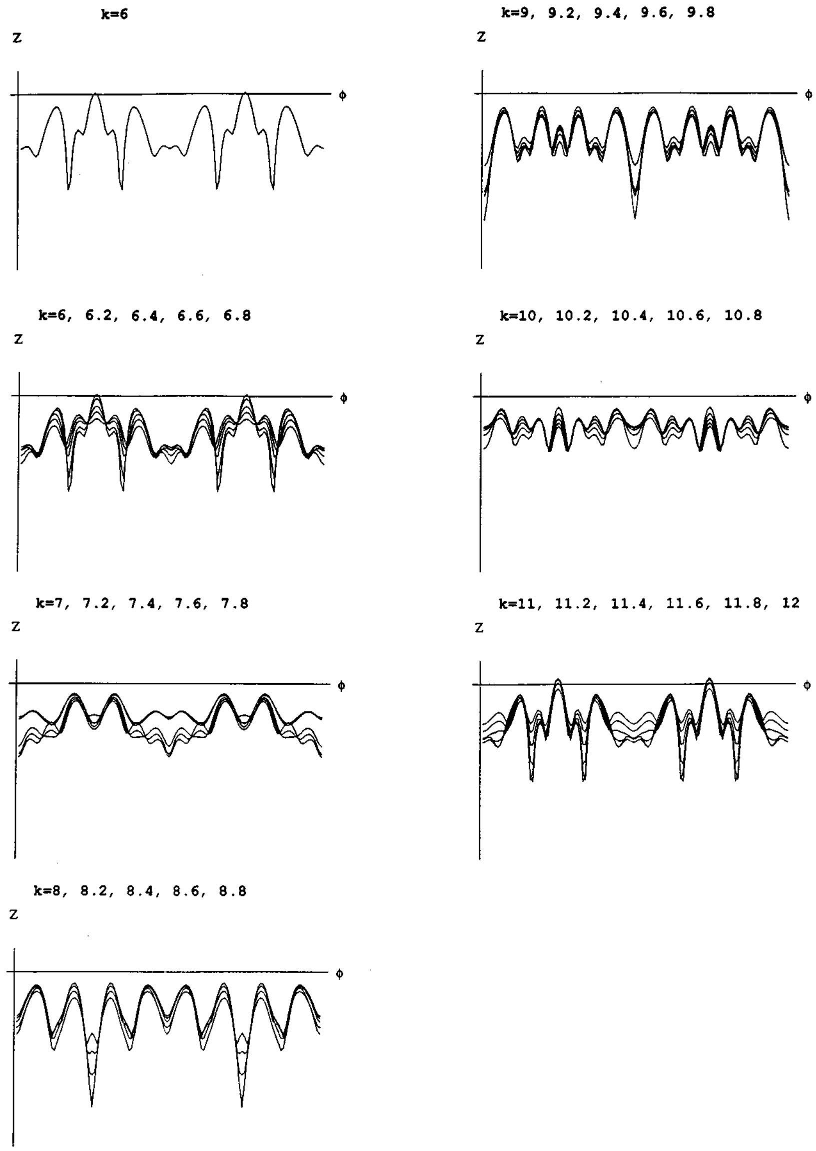

FIG. 4. $Z(\phi)$ versus ϕ relative to a corrugated sphere $m=4$, $n=8$ for different values of k.

(see Ref. 17, Chap. 13, page 1866) and coincides with the formula of the far field pattern generated by the sphere of radius one and center the origin when hit by the incoming wave (4.1).

Using the assumptions and the results of Sec. III we can say that $\underline{E}^{(0)}(\hat{\underline{x}})$ coincides with $\underline{F}^{(0)}(\hat{\underline{x}})$ given by (3.37).

First of all we consider the problem of where to truncate the series expansion (3.37) for $\underline{F}^{(0)}(\hat{\underline{x}})$ to have satisfactory approximation of the far field $\underline{E}^{(0)}(\hat{\underline{x}})$. To do this we consider the series expansion associated with the far field pattern generated by the sphere of radius one (i.e., $\Omega=B$) when hit by the incoming wave (4.1) truncated at $L_m>0$ that is

$$\underline{E}^{(0)}_{k,L_m}(\hat{\underline{x}})=\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{L_m}\sum_{m=\sigma}^{l}\left\{\frac{\underline{C}_{\sigma,m,l}(\hat{\underline{x}})(-1)^{l}\iota^{l+1}h_{0,B_{\sigma,m,l}}}{kh_l(k)}+\frac{\underline{B}_{\sigma,m,l}(\hat{\underline{x}})(-1)^{l+1}\iota^{l+2}h_{0,C_{\sigma,m,l}}}{(l+1)h_l(k)-kh_{l+1}(k)}\right\},$$

$$\hat{\underline{x}}\in\partial B, \tag{4.5}$$

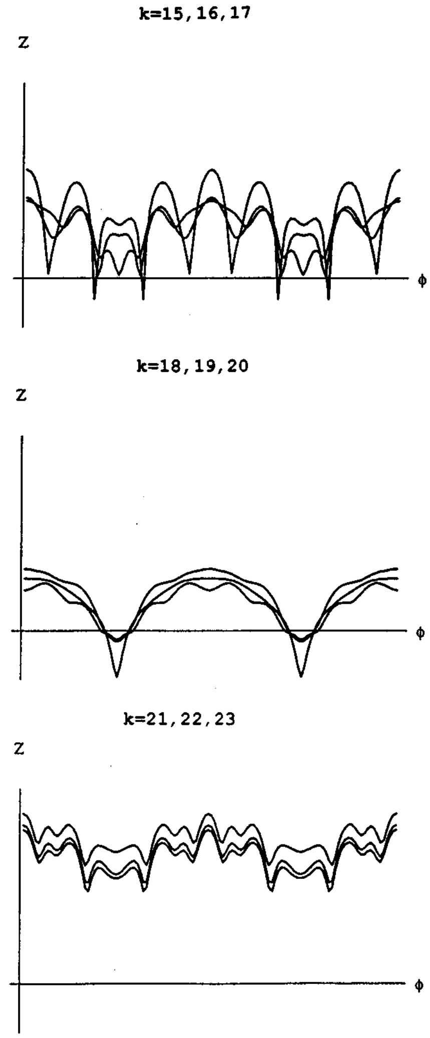

FIG. 5. $Z(\phi)$ versus ϕ relative to a corrugated sphere $m=4$, $n=8$ for different values of k.

we note that $h_l(k)$ and $(l+1)h_l(k)-kh_{l+1}(k)$, $l=0,1,...$, are not zero (see Lemma 3.4). The Fourier coefficients $\{h_{0,C_{\sigma,m,l}},h_{0,B_{\sigma,m,l}}\}$, $\sigma=0,1$, $l=\sigma,\ \sigma+1,...,L_m$, $m=\sigma,\ \sigma+1,...,l$, of the vector field $\underline{h}_0$ defined in (3.27) are given by

$$h_{0,B_{\sigma,m,l}}=-\epsilon_m \iota^l j_l(k)\frac{(2l+1)(l-m)!}{(l+m)!}(\underline{A},\underline{C}_{\sigma,m,l}(\underline{a})), \tag{4.6}$$

$$h_{0,C_{\sigma,m,l}}=-\epsilon_m \iota^{l+1}\frac{(l+1)j_l(k)-kj_{l+1}(k)}{k}\frac{(2l+1)(l-m)!}{(l+m)!}(\underline{A},\underline{B}_{\sigma,m,l}(\underline{a})), \tag{4.7}$$

where

$$\epsilon_m=\begin{cases}1 & m=0\\ 2 & m\neq 0\end{cases}.$$

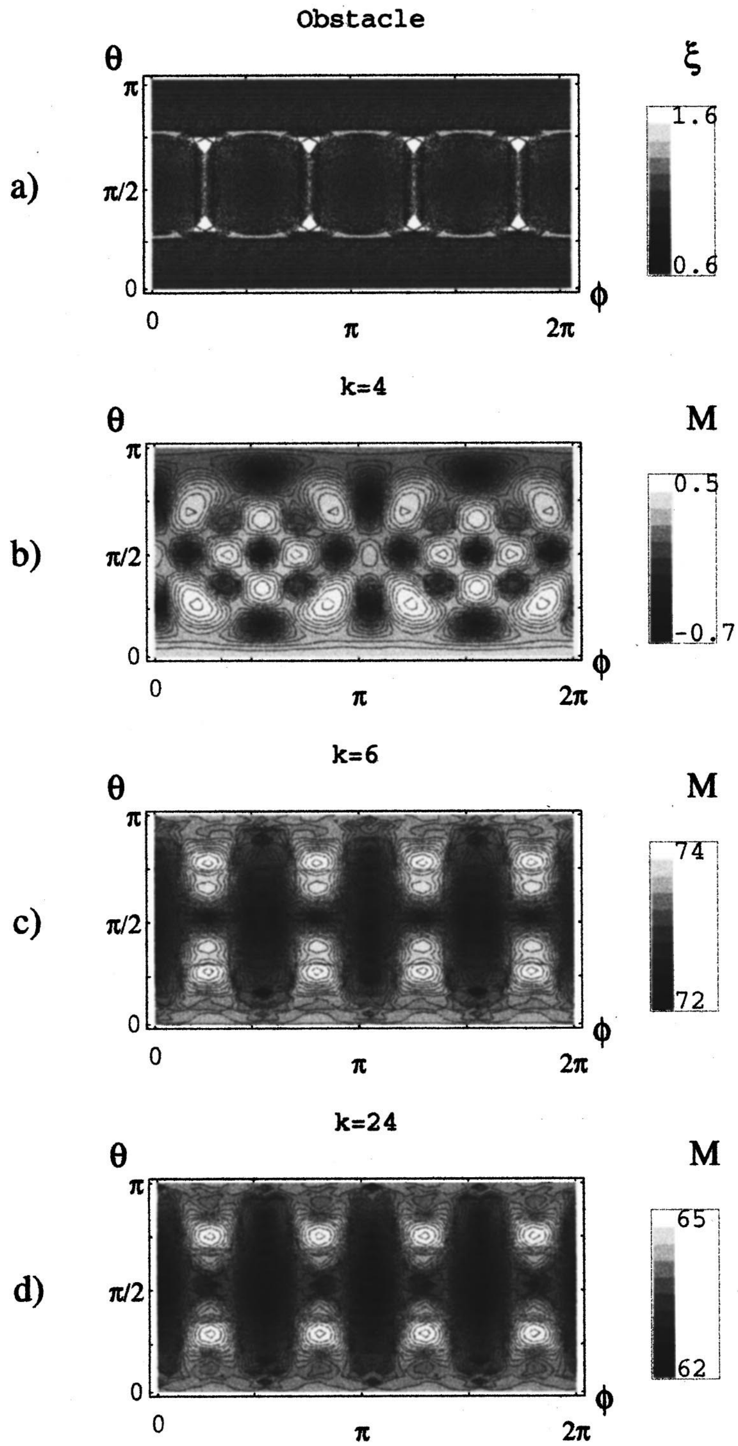

FIG. 6. (a) Contour plot of the cube; (b), (c), (d) contour plots of $M(\theta,\phi)$ relative to the geometry for different values of k.

We look for the smallest value of L_m such that ''numerical convergence'' is reached. We declare that ''numerical convergence'' has been reached when the following test is satisfied:

$$||\underline{E}^{(0)}_{k,L_m+1}(\hat{\underline{x}})|-|\underline{E}^{(0)}_{k,L_m}(\hat{\underline{x}})||\leq 10^{-6}|\underline{E}^{(0)}_{k,L_m}(\hat{\underline{x}})|, \quad \hat{\underline{x}}\in\partial B. \tag{4.8}$$

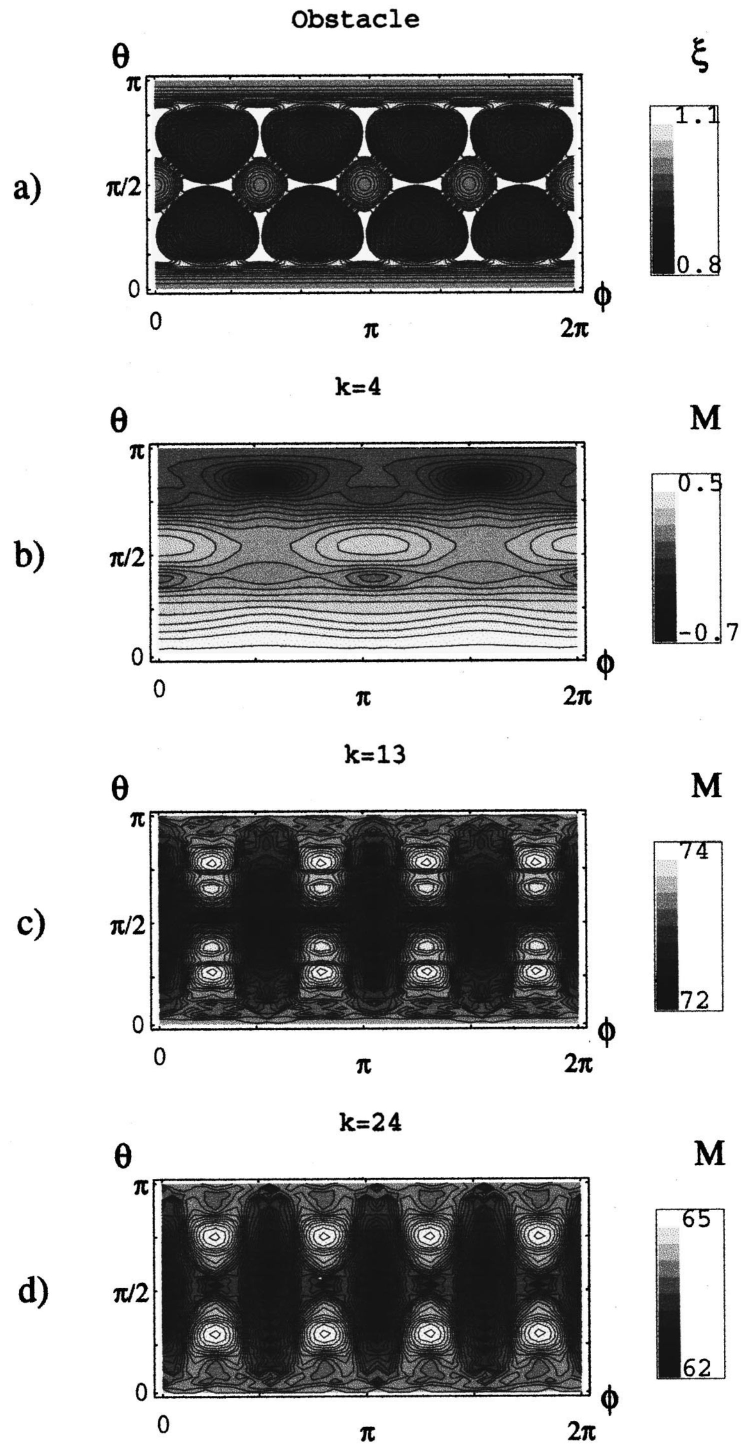

FIG. 7. (a) Contour plot of the octahedron; (b), (c), (d) contour plots of $M(\theta,\phi)$ relative to the geometry for different values of k.

We remind that if $\underline{E}=(E_1,E_2,E_3)^T$ then $|\underline{E}|=[\Sigma_{j=1}^{3}(E_j,\overline{E_j})]^{1/2}$, see Sec. II. From Fig. 1 we can conclude that when k increases the smallest value of L_m such that (4.8) is satisfied increases. The results shown in Fig. 1 are limited to the case of the sphere (i.e., $\Omega=B$) when $\hat{\underline{x}}=\underline{a}$ but these are relevant results since we expect that for a more general Ω and for a general $\hat{\underline{x}}$ the value of L_m that ensures numerical convergence will be greater or equal than the corresponding L_m relative to the case of the sphere when $\hat{\underline{x}}=\hat{a}$. Next we consider the computational cost of the method proposed in Sec. III in terms of the number of double integrals that given L_m must be computed in order to evaluate the expansion in ''powers'' of $(\xi-1)$ of the far field (3.37) up to order s that is

$$\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}})=\sum_{\nu=0}^{s}\sum_{\sigma=0}^{1}\sum_{l=\sigma}^{L_m}\sum_{m=\sigma}^{l}\left\{\underline{C}_{\sigma,m,l}(\hat{\underline{x}})\frac{h_{\nu,B_{\sigma,m,l}}}{\iota^{l+1}kh_l(k)}+\underline{B}_{\sigma,m,l}(\hat{\underline{x}})\frac{1}{\iota^{l+2}}\frac{h_{\nu,C_{\sigma,m,l}}}{(l+1)h_l(k)-kh_{l+1}(k)}\right\},$$

$$\hat{\underline{x}}\in\partial B, \tag{4.9}$$

where $\{h_{\nu,C_{\sigma,m,l}},h_{\nu,B_{\sigma,m,l}}\}$, $\nu=0,1,...,s$ $\sigma=0,1$, $l=\sigma$, $\sigma+1,...,L_m$, $m=\sigma$, $\sigma+1,...,l$, are the coefficients of the expansion of the vector field $\underline{h}_\nu$, $\nu=0,1,...,s$ given by (3.28). This is compared with the number of double integrals necessary to evaluate the T-matrix[14] when the T-matrix is evaluated using an expansion in vector spherical harmonics truncated at L_m. The T-matrix method is an alternative way to compute the far field $\underline{E}^{(0)}(\hat{\underline{x}})$.

Table I shows the result of this comparison. We note that the computational cost of the method proposed in Sec. III is given by a polynomial in s times a polynomial in L_m. That is relatively large values of s and L_m can be used. Moreover at a given order in ''perturbation theory'' (i.e., the contribution of the terms of order i, $i=1,2,...,s$) the computation is fully parallelizable with respect to the number of vector spherical harmonics involved (i.e., the value of L_m chosen). Table II shows a comparison between the far fields obtained with the perturbation series (4.9) and the far fields obtained with the T-matrix method for three axial symmetric obstacles.

We consider the following obstacles:

(1) Sphere: $\xi(\hat{\underline{x}}(\theta,\phi))=1.05$, $0\leqslant\theta\leqslant\pi$, $0\leqslant\phi<2\pi$.
(2) Ellipsoid: $\xi(\hat{\underline{x}}(\theta,\phi))=[(\sin\theta/1.05)^2+(\cos\theta/0.95)^2]^{-1/2}$, $0\leqslant\theta\leqslant\pi$, $0\leqslant\phi<2\pi$.
(3) Platelet: $\xi(\hat{\underline{x}}(\theta,\phi))=1.0+0.05\cos 2\theta$, $0\leqslant\theta\leqslant\pi$, $0\leqslant\phi<2\pi$.

For the incoming wave (4.1) we choose $\underline{a}=(0,0,1)^T$, $\underline{P}=(1,0,0)^T$. In Table II we denote with $\epsilon_{L_2}^{|\underline{E}^{(0)}|}$ the following quantity:

$$\epsilon_{L_2}^{|\underline{E}^{(0)}|}=\left[\frac{\Sigma_{j=0}^{20}||\underline{E}_T^{(0)}(\hat{\underline{x}}_j)|-|\underline{E}_{k,7,3}^{(0)}(\hat{\underline{x}}_j)||^2}{\Sigma_{j=0}^{20}|\underline{E}_T^{(0)}(\hat{\underline{x}}_j)|^2}\right]^{1/2}, \tag{4.10}$$

where $\hat{\underline{x}}_j=(\sin\theta_j,0,\cos\theta_j)^T$, with $\theta_j=j\pi/20$, $j=0,1,...,20$, $\underline{E}_T^{(0)}$ is the far field obtained with the T-matrix method when the T-matrix is approximated with an expansion in vector spherical harmonics truncated at $L_m=7$. Finally $\underline{E}_{k,7,3}^{(0)}$ given by (4.9) is the far field obtained with the expansion in (3.37) up to order $s=3$ in powers of $(\xi-1)$ with $L_m=7$.
We note that $\epsilon_{L_2}^{|\underline{E}^{(0)}|}$ is always small and depends on $k(\xi-1)$, in particular when $k\leqslant 10$ and $\max_{\theta,\phi}|\xi(\hat{\underline{x}}(\theta,\phi))-1|\leqslant 0.05$ we have $\epsilon_{L_2}^{|\underline{E}^{(0)}|}$ less than 3%. The previous choice of obstacles and incoming waves is a generic choice that gives a good sample of the behavior of the method in comparison with other computational methods such as the T-matrix method. We consider now obstacles with more complicated geometries. That is we consider obstacles with multiscale corrugations and obstacles with Lipschitz continuous boundaries such as polyhedra. We consider the following obstacles.

(4) Corrugated sphere: $\xi(\hat{\underline{x}}(\theta,\phi))=r_0+d\sin^2 m\theta\cos n\phi$ where $r_0>0$, $d\in\mathbb{R}$, $|d|<r_0$, m,n are integers [see Figs. 2(a) and 3(a)].
(5) Cube [see Fig. 6(a)].
(6) Cutted octahedron [see Fig. 7(a)].

The analytical expression of $\xi(\hat{\underline{x}}(\theta,\phi))$ for the cube and the cutted octahedron are involved and will be omitted. We only remark that the origin is the center of mass of these obstacles and that the largest sphere contained in the cube has radius 1 and the largest sphere contained in the cutted octahedron has radius 0.965. Figures 2 and 3 are relative to the corrugated sphere for different corrugations that is different choices of *n,m* in the previously given formula for $\xi(\hat{\underline{x}}(\theta,\phi))$.

Figures 2, 3, 6, 7 show the contour plots of the geometry of the obstacles [i.e., Figs. 2(a), 3(a), 6(a), 7(a)] and the contour plots of $\log_{10}|\underline{E}^{(0)}_{k,L_m,s}|$ for different values of the wave number k and of the parameters L_m,s [i.e., Figs. 2(b), 2(c), 2(d), 3(b), 3(c), 3(d), 6(b), 6(c), 6(d), 7(b), 7(c), 7(d)]. In all these figures the parameter s is fixed to be 10 and the parameter L_m is chosen depending on k in such a way that the numerical convergence (4.8) is reached. Moreover, Figs. 2(a), 3(a), 6(a), and 7(a) show $\xi(\hat{\underline{x}}(\theta,\phi))$ and Figs. 2(b), 2(c), 2(d), 3(b), 3(c), 3(d), 6(b), 6(c), 6(d), 7(b), 7(c), 7(d) show the quantity $\log_{10}|\underline{E}^{(0)}_{k,L_m,s}|$. In Figs. 2, 3, 6, 7 we denote with M the function $M(\theta,\phi)=\log_{10}|\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}}(\theta,\phi))|$, and in Figs. 4 and 5 we denote with Z the function $Z(\phi)=\log_{10}|\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}}(\pi/2,\phi))|$. Figures 2 and 3 show that the corrugations of the sphere become visible for values of the wave number k large enough. While for small wave number the corrugated sphere appears as a sphere without corrugation. Moreover we can see that since the size of the corrugated sphere is of order one at values of k of approximately 20 we are already in the geometrical optics limit [see Figs. 2(d) and 3(d)]. Figures 4 and 5 are relative to a corrugated sphere with $r_0=1$, $d=0.15$, $n=8$, $m=4$ and show the occurrence of a resonance phenomenon. This resonance phenomenon is due to the presence of the corrugation. That is

(i) For $k=6$ [Fig. 4(a)] $\log_{10}|\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}}(\pi/2,\phi))|$, $0\leqslant\phi<2\pi$ shows eight peaks.

(ii) Figures 4(b) and 4(c) show with stepsize $\delta k=0.2$ the quantity $\log_{10}|\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}}(\pi/2,\phi))|$, $0\leqslant\phi<2\pi$ as a function of k for $6\leqslant k\leqslant 12$. We note that the peaks at $k=6$ becomes valleys at $k=8$ and come back to peaks at $k=12$.

(iii) Figure 5 shows with stepsize $\delta k=1$ the quantity $\log_{10}|\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}}(\pi/2,\phi))|$, $0\leqslant\phi<2\pi$ as a function of k for $15\leqslant k\leqslant 23$. The same phenomenon discussed in (ii) appears here. Moreover at $k=20$ the far field $|\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}}(\pi/2,\phi))|$, $0\leqslant\phi<2\pi$ looks like the far field of the sphere.

The number of peaks and valleys (i.e., 8 peaks and valleys) shown in Figs. 4 and 5 is related to the corrugation of the obstacle (i.e., $n=8$). The qualitative change in $\log_{10}|\underline{E}^{(0)}_{k,L_m,s}(\hat{\underline{x}}(\pi/2,\phi))|$, $0\leqslant\phi<2\pi$ as a function of k, that is, the sequence: peak, far field of the sphere, valley, peak, that is shown for increasing values of k in Figs. 4(b), 4(c) and 5 is what we call resonance phenomenon. We note that the peaks and the valleys of the far field of the sphere (see Fig. 5) depend on the polarization vector of the incoming wave.

Figures 6 and 7 are relative to the cube and to the cutted octahedron, respectively. In these figures we can see that for increasing values of k appear first the facets then the edges and finally the vertices of the polyhedron considered. In particular, Fig. 6(d) shows that for k large enough the energy irradiated from the cube when hit by an incoming plane wave comes essentially from the vertices. We note that the cut octahedron is more ''similar'' to the sphere than the cube. In fact for $k=4$ the far field of the cutted octahedron is similar to the farfield of the sphere. On the contrary for the same value of k the far field of the cube shows some elements of the geometry of the cube. For $k=6$ the edges and the facets of the cube are already visible and for $k=24$ only the vertices remain visible. The far field of the cutted octahedron shows the same behavior of the far field of the cube but for larger value of k, that is $k=13$ and $k=24$. In particular for $k=24$ the edges and the eight facets of the cutted octahedron are visible.

[1] J. Nečas, *Les Méthodes Directes en théorie des Équations Elliptiques* (Masson and Cie, Paris, 1967).
[2] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory* (Krieger, Malabar, Florida, 1992).
[3] A. P. Calderón, ''Cauchy integrals on Lipschitz curves and related operators,'' Proc. Natl. Acad. Sci. USA **7**, 1324–1327 (1997).
[4] R. R. Coifman, A. McIntosh, and Y. Meyer, ''L' intégrale de Cauchy définit un opérateur borné sur L^2 pour les courbes lipschitziennes,'' Ann. Math. **116**, 361–387 (1982).

[5] G. Verchota, ''Layer potentials and boundary value problems for Laplace's equation in Lipschitz domains,'' J. Funct. Anal. **59**, 572–611 (1984).
[6] M. Mitrea, ''The method of layer potentials in electromagnetic scattering theory on nonsmooth domains,'' Duke Math. J. **77**, 111–133 (1995).
[7] C. Kenig, Elliptic boundary value problems on Lipschitz domains, in *Beijing Lectures in Harmonic Analysis*, Annals of Mathematics Studies (Princeton Univ. Press, Princeton, NJ, 1986), Vol. 112, pp. 131–183.
[8] D. M. Milder, ''An improved formalism for wave scattering from rough surface,'' J. Acoust. Soc. Am. **89**, 529–541 (1991).
[9] D. M. Milder, ''Role of the admittance operator in rough-surface scattering,'' J. Acoust. Soc. Am. **100**, 759–768 (1996).
[10] D. M. Milder, ''An improved formalism for electromagnetic scattering from a perfectly conducting rough surface,'' Radio Sci. **31**, 1369–1376 (1996).
[11] S. Piccolo, M. C. Recchioni, and F. Zirilli, ''The time harmonic electromagnetic field in a disturbed half-space: An existence theorem and a computational method,'' J. Math. Phys. **37**, 2762–2786 (1996).
[12] R. A. Smith, ''The operator expansion formalism for electromagnetic scattering from rough dielectric surfaces,'' Radio Sci. **31**, 1377–1385 (1996).
[13] L. Misici, G. Pacelli, F. Zirilli, ''A new formalism for wave scattering from a bounded obstacle,'' J. Acoust. Soc. Am. **103**, 106–113 (1998).
[14] G. Kristensson and P. C. Waterman, ''The T matrix for acoustic and electromagnetic scattering by circular disk,'' J. Acoust. Soc. Am. **72**, 1612–1625 (1982).
[15] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. IV (Academic, New York, 1978).
[16] R. A. Adams, *Sobolev Spaces* (Academic, New York, 1975).
[17] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Vol. II (McGraw-Hill, New York, 1953).
[18] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

On the electromagnetic scattering problem for an infinite dielectric cylinder of an arbitrary cross section located in the wedge

Yu. K. Podlipenko
Faculty of Cybernetics, Kiev University, 64, Vladimirskaya str., Kiev, Ukraine

Yu. V. Shestopalov[a)]
Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow 119899, Russia

(Received 3 December 1998; accepted for publication 24 May 1999)

Diffraction of time-harmonic E-polarized electromagnetic waves by an infinite dielectric cylinder of arbitrary cross section located inside a wedge parallel to its axis is considered. By the methods of potential theory, the transmission problem is reduced to a system of two one-dimensional Fredholm integral equations with the kernels having logarithmic singularities; integration is performed over the boundary of the cylinder cross section. Existence and uniqueness of solutions are proved both for the system of integral equations and the transmission problem. The kernels of integral equations are represented as rapidly convergent series. © *1999 American Institute of Physics.* [S0022-2488(99)00109-7]

I. INTRODUCTION

Free-space transmission problems were investigated in classical formulation by Kupradze,[1] Werner,[2] Kleinman and Kittappa,[3] Kress and Roach,[4] and other authors. The main attention was paid to the acoustic case, when consideration of boundary value problems for both two- and three-dimensional Helmholtz equations may be physically justified. The amount and quality of the results obtained for electromagnetic problems are much less significant—maybe because of more complicated formulations of the problems (see Macdonald[5] and Colton and Kress[6]). On the other hand, until recently, the uniqueness and solvability of transmission problems for (cylindrical) dielectric obstacles in a wedge have not been proved. This was caused, in particular, by the lack of results concerning the properties of corresponding potentials for the two-dimensional Helmholtz equation. Podlipenko[7,8] developed essential elements of the potential theory for a wedge. In this paper, we apply these results to derive weakly singular boundary integral equations and prove the uniqueness and solvability for the problem of scattering of electromagnetic waves by an infinite dielectric cylinder of arbitrary cross section located in a wedge. In general, this study may be considered as an extension of the approach developed by Kress and Roach[4] as applied to the wedge problem.

II. FORMULATION OF THE PROBLEM

Introduce in $\mathbf{R}^3$ the cylindrical coordinate system r,ϕ,z and denote by $W=\{(r,\phi,z)\in\mathbf{R}^3|r>0,0<\phi<\Phi,-\infty<z<+\infty\}$ a wedge with perfectly conducting walls, the vertex angle $\Phi(0<\Phi\leq 2\pi)$, and the edge coinciding with the z axis. Denote by C an infinite dielectric cylinder with the axis parallel to the edge of the wedge. Domains $W\backslash\overline{C}$ and C are supposed to be filled with homogeneous isotropic media having the permittivities ϵ_1 and ϵ_2, permeabilities μ_1 and μ_2, and conductivities σ_1 and σ_2, respectively.

[a)]Electronic mail:shestop@cs.musu.su

Let Ω and D be the domains obtained as a result of intersection of wedge W and cylinder C by the plane $z=0$, respectively. We will assume that $\overline{D}\subset\Omega$ and domain D is bounded, simply connected, and has a C^2-boundary ∂D. Denote by $\partial\Omega$ the boundary of angular domain Ω.

Let a source of a time-harmonic cylindrical wave be an infinitely long thread located in domain $W\backslash\overline{C}$ parallel to the edge of the wedge; the electric current of a constant amplitude and phase flows along the thread.

We will consider diffraction by cylinder C of the E-field excited by this source. We will mark the components of the initial, scattered, and reflected electromagnetic fields by the upper indices i, s, and r, respectively. The problem is to determine the field components that do not depend on z; these components can be represented in terms of two potential functions $u_1(P)=E_z^{(s)}(P)$, $P\in\Omega\backslash\overline{D}$, and $u_2(P)=E_z^{(r)}(P)$, $P\in D$, $P=(r,\phi)$, solving the following transmission problem:

$$\Delta u_1(r,\phi)+k_1^2u_1(r,\phi)=0, \quad ,(r,\phi)\in\Omega\backslash\overline{D}, \tag{1}$$

$$\Delta u_2(r,\phi)+k_2^2u_2(r,\phi)=0, \ (r,\phi)\in D; \tag{2}$$

$$u_1-u_2=f, \quad \lambda_1\frac{\partial u_1}{\partial n}-\lambda_2\frac{\partial u_2}{\partial n}=g, \ \text{on} \ \partial D; \tag{3}$$

$$u_1=0, \ \text{on} \ \partial\Omega; \tag{4}$$

$$\frac{\partial u_1(r,\phi)}{\partial r}-ik_1u_1(r,\phi)=o\left(\frac{1}{\sqrt{r}}\right), \quad r\to\infty, \tag{5}$$

uniformly with respect to ϕ; and

$$\int_{\Omega\cap\delta}(|u_1|^2+|\text{grad}u_1|^2)dS<\infty. \tag{6}$$

Here, $\lambda_j=1/\mu_j$, $k_j=\omega\sqrt{\epsilon_j'\mu_j}$, $\epsilon_j'=\epsilon_j+i\sigma_j/\omega$, $\Im k_j\geqslant 0$, $j=1,2$, ω is the field frequency (the time dependence $e^{-i\omega t}$ is assumed), Δ is the Laplace operator in polar coordinates, δ is a neighborhood of the origin of the polar coordinate system, n is the unit normal to ∂D drawn in the direction from D to $\Omega\backslash\overline{D}$,

$$f=-E_z^{(i)}|_{\partial D}, \quad g=-\lambda_1\frac{\partial E_z^{(i)}}{\partial n}\bigg|_{\partial D},$$

and

$$E_z^{(i)}(r,\phi)=\frac{-\omega\mu_1 I\pi}{\Phi}\sum_{m=1}^{\infty}\sin(\nu_m\phi^*)\sin(\nu_m\phi)J_{\nu_m}(k_1\min(r^*,r))H_{\nu_m}^{(1)}(k_1\max(r^*,r)),$$

where $(r^*,\phi^*)\in\Omega\backslash\overline{D}$ is the point of intersection of the line source with the plane $z=0$, $J_\nu(x)$ and $H_\nu^{(1)}(x)$ are the Bessel and Hankel functions of the order ν, $\nu_m=m\pi/\Phi$, $m=1,2,\ldots$, and I is the complex amplitude of the current flowing along the thread.

We make the following comments concerning the problem statement: the functions

$$E_z^{(i)}(r,\phi), \quad H_r^{(i)}(r,\phi)=-\frac{i}{r\omega\mu_1}\frac{\partial E_z^{(i)}(r,\phi)}{\partial\phi}, \quad H_\phi^{(i)}(r,\phi)=\frac{i}{\omega\mu_1}\frac{\partial E_z^{(i)}(r,\phi)}{\partial r},$$

are the nonzero components of the field $\mathbf{E}^{(i)}$, $\mathbf{H}^{(i)}$ excited by the source[5] in the wedge W; radiation condition (5) excludes[5] waves coming from infinity; the Meixner condition (6) ensures the absence of energy flux radiated by the edge of the wedge; and boundary condition (4) corresponds to the case of perfectly conducting wedge faces. Transmission condition (3) guarantees continuity of normal components of the total electromagnetic field $\mathbf{E}^{(1)}=(0,0,E_z^1)$, $\mathbf{H}^{(1)}=(H_r^1,H_\phi^1,0)$ and $\mathbf{E}^{(2)}=(0,0,E_z^2)$, $\mathbf{H}^{(2)}=(H_r^2,H_\phi^2,0)$ in the domains $W\setminus\overline{C}$ and C, respectively, when the lateral area (interface) of the cylinder ∂C is crossed; these components are determined by the formulas

$$E_z^{(1)}(r,\phi)=E_z^{(i)}(r,\phi)+u_1(r,\phi),\quad H_r^{(1)}(r,\phi)=H_r^{(i)}(r,\phi)-\frac{i}{r\omega\mu_1}\frac{\partial u_1(r,\phi)}{\partial\phi},$$

$$H_\phi^{(1)}(r,\phi)=H_\phi^{(i)}(r,\phi)+\frac{i}{\omega\mu_1}\frac{\partial u_1(r,\phi)}{\partial r},\quad (r,\phi)\in\Omega\setminus\overline{D},$$

$$E_z^{(2)}(r,\phi)=u_2(r,\phi),\quad H_r^{(2)}(r,\phi)=-\frac{i}{r\omega\mu_2}\frac{\partial u_2(r,\phi)}{\partial\phi},$$

$$H_\phi^{(2)}(r,\phi)=\frac{i}{\omega\mu_2}\frac{\partial u_2(r,\phi)}{\partial r},\quad (r,\phi)\in D.$$

Below, we will consider a more general transmission problem: find functions $u_1\in C^2(\Omega\setminus\overline{D})\cap C^1(\Omega\setminus D)\cap C(\overline{\Omega}\setminus\overline{D})$ and $u_2\in C^2(D)\cap C^1(\overline{D})$ that solve (1)–(6), where $f\in C^{1,\alpha}(\partial D)$, $g\in C^{0,\alpha}(\partial D)$ $(0<\alpha\leq 1)$ are given functions on ∂D, and k_1, k_2, λ_1, and λ_2 are nonzero complex numbers with $0\leq\arg k_j<\pi/2$, $j=1,2$.

Our aim is to establish the uniqueness and existence of solution to (1)–(6) by reducing this problem to an operator equation suitable for further computations.

III. UNIQUENESS

First, we prove the uniqueness theorem for the considered transmission problem.

Theorem 1: *Let* $k_1,k_2\in\mathbf{C}\setminus\{\mathbf{0}\}$, $0\leq\arg k_1<\pi/2$, $0\leq\arg k_2<\pi/2$, *and* $\lambda_1,\lambda_2\in\mathbf{C}\setminus\{\mathbf{0}\}$ *be such that* $\rho:=\lambda_2\overline{k_2^2}(\lambda_1\overline{k_1^2})^{-1}\in\mathbf{R}$, $\rho>0$. *Then, the solution to transmission problem* (1)–(6), *if it exists, is unique.*

Proof: It is sufficient to prove that homogeneous transmission problem (1)–(6) with $f=g=0$ has only the trivial solution.

Assume that u_1 and u_2 satisfy (1), (2), (4)–(6) and the transmission conditions

$$u_1-u_2=0,\quad \lambda_1\frac{\partial u_1}{\partial n}-\lambda_2\frac{\partial u_2}{\partial n}=0,\ \text{on}\ \partial D. \tag{7}$$

Choose a number R_0 so that $\overline{D}\subset\Omega_R:=\{(r,\phi)\in\Omega|r<R\}$ when $R\geq R_0$ and apply the first Green's formula to functions $u_1,\overline{u_1}$ and $u_2,\overline{u_2}$ in domains $\Omega_R\setminus\overline{D}$ and D, respectively, taking into account Eqs. (1) and (2). As a result, we obtain

$$-\overline{k_1^2}\lambda_1\int_{\Omega_R\setminus\overline{D}}|u_1|^2\,dS=-\lambda_1\int_{\partial D}u_1\frac{\partial\overline{u_1}}{\partial n}dl+\lambda_1\int_{C_R}u_1\frac{\partial\overline{u_1}}{\partial n}dl-\lambda_1\int_{\Omega_R\setminus\overline{D}}|\text{grad }u_1|^2\,dS$$

$$-\overline{k_2^2}\lambda_2\int_D|u_2|^2\,dS=\lambda_2\int_{\partial D}u_2\frac{\partial\overline{u_2}}{\partial n}dl-\lambda_2\int_D|\text{grad }u_2|^2\,dS,$$

where $C_R:=\{(r,\phi)\in\mathbf{R}^2|0\leq\phi\leq\Phi,r=R\}$. Note that here we can use the Green's formula in the domain $\Omega_R\setminus\overline{D}$ because $u_1,\text{grad }u_1,\Delta u_1\in L_2(\Omega_R\setminus\overline{D})$ by virtue of the edge condition (6) and Theorem 3.27, Colton and Kress.[6] Using the latter relationships and transmission conditions (7), we obtain

$$\lambda_1 \int_{C_R} u_1 \frac{\partial \overline{u_1}}{\partial n} dl = \lambda_1 \int_{\Omega_R \setminus \overline{D}} |\text{grad } u_1|^2 dS + \lambda_2 \int_D |\text{grad } u_2|^2 \, dS$$

$$- \overline{k_1^2} \lambda_1 \int_{\Omega_R \setminus \overline{D}} |u_1|^2 \, dS - \overline{k_2^2} \lambda_2 \int_D |u_2|^2 \, dS. \tag{8}$$

Dividing both sides of (8) by $\overline{k_1^2}\lambda_1$ and taking the imaginary part, we have

$$\Im\left(\frac{1}{\overline{k_1^2}} \int_{C_R} u_1 \frac{\partial \overline{u_1}}{\partial n} dl \right) = \Im\left(\frac{1}{\overline{k_1^2}} \int_{\Omega_R \setminus \overline{D}} |\text{grad } u_1|^2 \, dS \right) + \rho \Im \int_D \left(\frac{1}{\overline{k_2^2}} |\text{grad } u_2|^2 \right) dS. \tag{9}$$

Using relationship (9), we first prove the assertion of the theorem under the condition $\Im k_1 > 0$. In this case, we will show that the left-hand side of (9) tends to zero for a certain subsequence C_{R_m} as $R_m \to \infty$. Taking the imaginary part of the equality

$$k_1 \int_{C_R} u_1 \frac{\partial \overline{u_1}}{\partial n} dl = k_1 \int_{\partial D} u_1 \frac{\partial \overline{u_1}}{\partial n} dl - \overline{k_1} |k_1|^2 \int_{\Omega_R \setminus \overline{D}} |u_1|^2 \, dS + k_1 \int_{\Omega_R \setminus \overline{D}} |\text{grad } u_1|^2 \, dS,$$

obtained as a result of applying the first Green's formula to functions $u_1, \overline{u_1}$ in the domain $\Omega_R \setminus \overline{D}$ and substituting the resulting expression into the formula

$$0 = \lim_{R \to \infty} \int_{C_R} \left| \frac{\partial u_1}{\partial n} - i k_1 u_1 \right|^2 dl = \lim_{R \to \infty} \int_{C_R} \left\{ \left| \frac{\partial u_1}{\partial n} \right|^2 + |k_1|^2 |u_1|^2 + 2 \Im \left(k_1 u_1 \frac{\partial \overline{u_1}}{\partial n} \right) \right\} dl, \tag{10}$$

that follows from radiation condition (5), we find that

$$\lim_{R \to \infty} \left\{ \int_{C_R} \left(\left| \frac{\partial u_1}{\partial n} \right|^2 + |k_1|^2 |u_1|^2 \right) dl + 2 \Im k_1 \int_{\Omega_R \setminus \overline{D}} (|k_1|^2 |u_1|^2 + |\text{grad } u_1|^2) dS \right\}$$

$$= -2 \Im \left(k_1 \int_{\partial D} u_1 \frac{\partial \overline{u_1}}{\partial n} dl \right). \tag{11}$$

All terms on the left-hand side of (11) are non-negative because $\Im k_1 > 0$. Therefore, each of them is bounded, since their sum tends to a finite limit. In particular,

$$\lim_{R \to \infty} \int_{\Omega_R \setminus \overline{D}} (|k_1|^2 |u_1|^2 + |\text{grad } u_1|^2) dS < \infty,$$

which means that $u_1 \in W_2^{(1)}(\Omega \setminus \overline{D})$. For sufficiently large $R' > R_0$, we have

$$\left| \int_{R'}^{\infty} \int_{C_R} u_1 \frac{\partial \overline{u_1}}{\partial n} dl \, dR \right|^2 \leq \int_{\Omega \setminus \overline{\Omega}_{R'}} |u_1|^2 \, dS \int_{\Omega \setminus \overline{\Omega}_{R'}} \left| \frac{\partial u_1}{\partial n} \right|^2 \, dS$$

$$\leq \int_{\Omega \setminus \overline{D}} |u_1|^2 \, dS \int_{\Omega \setminus \overline{D}} |\text{grad } u_1|^2 \, dS < \infty.$$

Therefore, there exists a sequence C_{R_m} such that

$$\int_{C_{R_m}} u_1 \frac{\partial \overline{u_1}}{\partial n} dl,$$

and the sequence

$$\Im\left(\frac{1}{\overline{k_1^2}}\int_{C_{R_m}} u_1 \frac{\partial \overline{u_1}}{\partial n} dl\right)$$

tends to zero as $R_m \to \infty$. Passing to the limit in (9) over the sequence C_{R_m}, we obtain

$$\frac{\Re k_1 \Im k_1}{|k_1|^4}\int_{\Omega \setminus \overline{D}} |\text{grad } u_1|^2 \, dS + \rho \frac{\Re k_2 \Im k_2}{|k_2|^4}\int_D |\text{grad } u_2|^2 \, dS = 0. \tag{12}$$

Then, from (12), it follows that $\int_{\Omega\setminus\overline{D}} |\text{grad } u_1|^2 \, dS = 0$, and $|\text{grad } u_1| = 0$ in $\Omega\setminus\overline{D}$. Therefore, $u_1 = \text{const}$ in $\Omega\setminus\overline{D}$. The radiation condition yields $u_1 = 0$ in $\Omega\setminus\overline{D}$, and from the transmission conditions we obtain $u_2 = \partial u_2/\partial n = 0$ on ∂D. Consequently, from the integral representation for a solution of the Helmholtz equation, it follows that $u_2 = 0$ in D.

Now we consider the case $\Im k_1 = 0$. Since $\rho > 0$, condition (9) yields

$$\Im \int_{C_R} u_1 \frac{\partial \overline{u_1}}{\partial n} dl \geqslant 0.$$

As a consequence of the radiation condition, we have

$$\frac{\partial \overline{u_1}(R,\phi)}{\partial n} + ik_1 \overline{u_1}(R,\phi) = o\left(\frac{1}{\sqrt{R}}\right), \quad R \to \infty. \tag{13}$$

Multiplying (13) by u_1 and integrating over C_R, we obtain

$$\int_{C_R} u_1 \frac{\partial \overline{u_1}(R,\phi)}{\partial n} \, dl + ik_1 \int_{C_R} |u_1(R,\phi)|^2 dl = \int_{C_R} u_1(R,\phi) o\left(\frac{1}{\sqrt{R}}\right) dl, \quad R \to \infty. \tag{14}$$

Any solution $u_1(r,\phi)$ of the Helmholtz equation (1) that satisfies boundary condition (4), radiation condition (5), and edge condition (6), also satisfies[8] the condition

$$u_1(r,\phi) = O\left(\frac{1}{\sqrt{r}}\right), \quad r \to \infty,$$

uniformly with respect to $\phi \in [0,\Phi]$. Using this fact and taking the imaginary part of Eq. (14), we find that

$$k_1 \int_{C_R} |u_1|^2 \, dl + \Im \int_{C_R} u_1 \frac{\partial \overline{u_1}}{\partial n} dl = o(1),$$

as $R \to \infty$. Taking into account that both summands on the left-hand side are non-negative, we conclude that

$$\int_{C_R} |u_1|^2 \, dl = \int_0^\Phi R |u_1(R,\phi)|^2 \, d\phi = o(1), \tag{15}$$

as $R \to \infty$. At the next stage of the proof, we set $u_m(r,\phi) := H^{(1)}_{\nu_m}(k_1 r)\sin(\nu_m \phi)$, where $\nu_m = m\pi/\Phi$. Applying the second Green's formula to functions $u_1(r,\phi)$ and $u_m(r,\phi)$ in the domains $\Omega_R \setminus \overline{\Omega}_{R_0}$ and taking into account that function $u_m(r,\phi)$ satisfies Eq. (1) in this domain and boundary condition (4), we obtain

$$0=\int_{C_{R_0}}\left\{u_1(r,\phi)\frac{\partial u_m(r,\phi)}{\partial n}-u_m(r,\phi)\frac{\partial u_1(r,\phi)}{\partial n}\right\}dl$$
$$+\int_0^{\Phi}\left\{u_1(R,\phi)\frac{dH^{(1)}_{\nu_m}(k_1R)}{dR}-H^{(1)}_{\nu_m}(k_1R)\frac{\partial u(R,\phi)}{\partial R}\right\}\sin(\nu_m\phi)R\,d\phi=:I_{R_0}+I_R. \qquad (16)$$

One can easily show that

$$I_R=\int_0^{\Phi}u_1(R,\phi)\left\{\frac{dH^{(1)}_{\nu_m}(k_1R)}{dR}-ik_1H^{(1)}_{\nu_m}(k_1R)\right\}\sin(\nu_m\phi)R\,d\phi$$
$$-\int_0^{\Phi}H^{(1)}_{\nu_m}(k_1R)\left\{\frac{\partial u_1(R,\phi)}{\partial R}-ik_1u_1(R,\phi)\right\}\sin(\nu_m\phi)R\,d\phi$$
$$=:J_1+J_2. \qquad (17)$$

Let us estimate integral J_1. Using the relationship

$$\frac{dH^{(1)}_{\nu_m}(k_1r)}{dr}-ik_1H^{(1)}_{\nu_m}(k_1r)=O(r^{-3/2}),\quad r\to\infty, \qquad (18)$$

and applying the Schwarz inequality in combination with (15), we obtain

$$|J_1|=\left|\int_0^{\Phi}u_1(R,\phi)\left\{\frac{dH^{(1)}_{\nu_m}(k_1R)}{dR}-ik_1H^{(1)}_{\nu_m}(k_1R)\right\}\sin(\nu_m\phi)R\;d\phi\right|$$
$$\leqslant\left|\frac{dH^{(1)}_{\nu_m}(k_1R)}{dR}-ik_1H^{(1)}_{\nu_m}(k_1R)\right|R\left(\int_0^{\Phi}|u_1(R,\phi)|^2d\phi\right)^{1/2}\Phi^{1/2}$$
$$=O(R^{-3/2})R\,o(R^{-1/2})=o\left(\frac{1}{R}\right),\quad R\to\infty. \qquad (19)$$

By virtue of radiation condition (5) and the estimate $H^{(1)}_{\nu_m}(k_1R)=O(R^{-1/2}),\quad R\to\infty$, we have

$$|J_2|\leqslant R|H^{(1)}_{\nu_m}(k_1R)|\int_0^{\Phi}\left|\frac{\partial u_1(R,\phi)}{\partial R}-ik_1u_1(R,\phi)\right|d\phi=RO(R^{-1/2})o(R^{-1/2})=o(1),\quad R\to\infty. \qquad (20)$$

Estimates (19) and (20) imply that the second term on the right-hand side of (16) tends to zero as $R\to\infty$. Since the first integral I_{R_0} in (16) does not depend on R, this yields $I_{R_0}=0$. Hence, $I_R=0$ for any R; that is,

$$\frac{dH^{(1)}_{\nu_m}(k_1R)}{dR}R\int_0^{\Phi}u_1(R,\phi)\sin(\nu_m\phi)d\phi-RH^{(1)}_{\nu_m}(k_1R)\int_0^{\Phi}\frac{\partial u_1(R,\phi)}{\partial R}\sin(\nu_m\phi)d\phi=0.$$

Introducing the notation

$$\alpha_m(k_1R):=\int_0^{\Phi}u_1(R,\phi)\sin(\nu_m\phi)d\phi, \qquad (21)$$

we rewrite the last equality as

$$\alpha_m(k_1R)\frac{dH^{(1)}_{\nu_m}(k_1R)}{dR}-H^{(1)}_{\nu_m}(k_1R)\frac{d\alpha_m(k_1R)}{dR}=0;$$

consequently, $\alpha_m(k_1R)=a_mH^{(1)}_{\nu_m}(k_1R)$, where a_m is a constant. Relation (21) means that $\alpha_m(k_1R)$, $m=1,2,\ldots$, are the Fourier coefficients of the function $u_1(R,\phi)$ over the complete system of functions $\{\sin(\nu_m\phi)\}_{m=1}^{\infty}$ on the interval $[0,\Phi]$. Therefore, Parseval's equality yields

$$\int_0^{\Phi}R|u_1(R,\phi)|^2\,d\phi=\frac{\Phi}{2}\sum_{m=1}^{\infty}R|\alpha_m(k_1R)|^2=\frac{\Phi}{2}\sum_{m=1}^{\infty}R|a_m|^2|H^{(1)}_{\nu_m}(k_1R)|^2. \tag{22}$$

The latter equalities in combination with (15) imply that $R|a_m|^2||H^{(1)}_{\nu_m}(k_1R)|^2\to 0$ as $R\to\infty$. However, according to the asymptotic formula

$$H^{(1)}_{\nu_m}(k_1R)\sim\left(\frac{2}{\pi k_1R}\right)^{1/2}e^{i(k_1R-\pi\nu_m/2-\pi/4)},\quad R\to\infty,$$

and, for large values of R, the product remains greater than a certain positive number; hence, $a_m=0$, i.e., $a_m(k_1R)\equiv 0$. Consequently, by virtue of (22), we have $u_1\equiv 0$ on the arc C_R of a sufficiently large radius, and, therefore, $u_1(P)=0$ at all points $P(r_P,\phi_P)\in\Omega$, for which $r_P\geqslant R_0$. We may conclude that $u_1\equiv 0$ everywhere in the domain $\Omega\backslash D$, because a solution of the Helmholtz equation is an analytical function. The theorem is proved.

IV. POTENTIALS

Let us define the potentials

$$u_1(M)=\lambda_1^{-1}\int_{\partial D}\left(\frac{\partial G_{k_1}(M,P)}{\partial n_P}\phi(P)+c_1G_{k_1}(M,P)\psi(P)\right)dl_P,\quad M\in\Omega\backslash\partial D, \tag{23}$$

$$u_2(M)=\lambda_2^{-1}\int_{\partial D}\left(\frac{\partial G_{k_2}(M,P)}{\partial n_P}\phi(P)+c_2G_{k_2}(M,P)\psi(P)\right)dl_P,\quad M\in\Omega\backslash\partial D, \tag{24}$$

with the densities $\phi\in C^{1,\alpha}(\partial D)$ and $\psi\in C^{0,\alpha}(\partial D)$, where $c_1,c_2\in\mathbf{C}\backslash\{\mathbf{0}\}$ are fixed constants and the function $G_{k_j}(M,P)$, $j=1,2$ is defined by the formula

$$G_{k_j}(M,P)=\frac{i\pi}{\Phi}\sum_{m=1}^{\infty}\left\{\begin{matrix}J_{\nu_m}(k_jr_P)H^{(1)}_{\nu_m}(k_jr_M)\\ J_{\nu_m}(k_jr_M)H^{(1)}_{\nu_m}(k_jr_P)\end{matrix}\right\}\sin(\nu_m\phi_M)\sin(\nu_m\phi_P). \tag{25}$$

Here and in all subsequent similar formulas, the upper and lower terms correspond, respectively, to the cases $r_P\leqslant r_M$ and $r_P\geqslant r_M$.

Functions $G_{k_j}(M,P)$ and $\partial G_{k_j}(M,P)/\partial n_P$ can be represented in the form[7,8]

$$G_{k_j}(M,P)=G_0(k_j;M,P)+\frac{1}{4\pi}\ln\Psi(M,P), \tag{26}$$

where $G_0(k_j;M,P)$ is a regular function in Ω, which can be represented as a series

$$G_0(k_j;M,P)=\frac{i\pi}{\Phi}\sum_{m=1}^{\infty}\left\{\begin{bmatrix}J_{\nu_m}(k_jr_P)H^{(1)}_{\nu_m}(k_jr_M)\\ J_{\nu_m}(k_jr_M)H^{(1)}_{\nu_m}(k_jr_P)\end{bmatrix}+\frac{i}{\pi\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right\}\sin(\nu_m\phi_M)\sin(\nu_m\phi_P), \tag{27}$$

whose general term decreases as $O(1/\nu_m^3)$ when $m\to\infty$. Here and below, we use the notation $R=\mathrm{sign}(r_M-r_P)$.

Function $\Psi(M,P)$ can be expressed explicitly,

$$\Psi(M,P)=\frac{\sin^2\frac{\pi}{2\Phi}(\phi_P+\phi_M)+\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_M}{r_P}\right)}{\sin^2\frac{\pi}{2\Phi}(\phi_P-\phi_M)+\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_M}{r_P}\right)},\tag{28}$$

and has a singularity of the type $1/r_{M,P}^2$, where $r_{M,P}$ denotes the distance between points M and P; therefore, the following relationship holds:

$$\frac{1}{4\pi}\ln\Psi(M,P)=\frac{1}{2\pi}\ln\frac{1}{r_{M,P}}+\phi(M,P),$$

where $\phi(M,P)$ is an analytical function with respect to all variables. Thus, we can write the following expressions:

$$\frac{\partial G_{k_j}(M,P)}{\partial n_P}=\frac{1}{4\pi}\frac{\partial}{\partial n_P}\ln\Psi(M,P)+\sum_{q=1}^{3}G_q^{(1)}(k_j;M,P),\quad j=1,2,\tag{29}$$

where

$$G_1^{(1)}(k_j;M,P)=\frac{i\pi}{\Phi}\cos\alpha(P)\sum_{m=1}^{\infty}\left\{k_j\left[\frac{J'_{\nu_m}(k_jr_P)H^{(1)}_{\nu_m}(k_jr_M)}{J_{\nu_m}(k_jr_M)H^{(1)\prime}_{\nu_m}(k_jr_P)}\right]+R\frac{i}{\pi r_P}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right.$$
$$\left.+\frac{i}{\pi r_P}\frac{k_j^2(r_M^2-r_P^2)}{4\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right\}\sin(\nu_m\phi_M)\sin(\nu_m\phi_P),\tag{30}$$

$$G_2^{(1)}(k_j;M,P)=\frac{i\pi}{\Phi}\sin\alpha(P)\sum_{m=1}^{\infty}\left\{\frac{\nu_m}{r_P}\left[\frac{-J_{\nu_m}(k_jr_P)H^{(1)}_{\nu_m}(k_jr_M)}{-J_{\nu_m}(k_jr_M)H^{(1)}_{\nu_m}(k_jr_P)}\right]-\frac{i}{\pi r_P}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right.$$
$$\left.-R\frac{i}{\pi r_P}\frac{k_j^2(r_M^2-r_P^2)}{4\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right\}\sin(\nu_m\phi_M)\sin(\nu_m\phi_P),\tag{31}$$

$$G_3^{(1)}(k_j;M,P)=\cos\alpha(P)\frac{k_j^2(r_M^2-r_P^2)}{16\pi r_P}\ln\Psi(M,P)$$
$$-R\sin\alpha(P)\frac{k_j^2(r_M^2-r_P^2)}{8\pi r_P}(\Phi_1(M,P)+\Phi_2(M,P));\tag{32}$$

functions $\Phi_i(M,P),\ i=1,2$, are defined by the formulas

$$\Phi_{1,2}(M,P)=\arctan\frac{\sin\frac{\pi}{\Phi}(\phi_M\pm\phi_P)}{\left(\frac{r_P}{r_M}\right)^{-R(\pi/\Phi)}-\cos\frac{\pi}{\Phi}(\phi_M\pm\phi_P)},$$

$\alpha(P)$ is the angle between the vector of point P and the unit normal n_P to the curve ∂D. The terms in series (30) and (31) decrease as $O[(1/\nu_m^2)(r_P/r_M)^{\nu_mR}]$.

From relationships (26) and (29), it follows that potentials (23) and (24) are well defined on the curve ∂D. The following estimates are valid:[8]

$$|G_{k_1}(M,P)|\leqslant Ar_M^{\pi/\Phi}, \quad \left|\frac{\partial G_{k_1}(M,P)}{\partial n_P}\right|\leqslant Br_M^{\pi/\Phi},$$

where A and B are constants that do not depend on $M\in\delta\cap\Omega$, $P\in\partial D$, and δ denotes a neighborhood of the origin O of the polar coordinate system. It is easy to see that functions u_1 and u_2 defined by formulas (23) and (24) belong to the classes $C^2(\Omega\setminus\overline{D})\cap C^1(\Omega\setminus D)\cap C(\overline{\Omega}\setminus\overline{D})$ and $C^2(D)\cap C^1(\overline{D})$, respectively, if we set by continuity $u_1(O)=0$.

Functions $G_{k_1}(M,P)$ and $\partial G_{k_1}(M,P)/\partial n_P$ satisfy radiation condition (6) uniformly[8] in $0<\phi_M<\Phi$, $P\in\partial D$. This implies that potentials (23) and (24) satisfy radiation condition (5) and edge condition (6). In addition to this, potentials (23) and (24) satisfy[8] Eqs. (1) and (2), respectively, and vanish on $\partial\Omega$.

We will need also a representation similar to (26) for the functions

$$\frac{\partial^2 G_{k_j}(M,P)}{\partial n_M\ \partial n_P}, \quad j=1,2,$$

where n_M is the outward (with respect to D) unit normal on ∂D at the point $M\in\Omega$. To this end, we use the following result.

Lemma 1: For the Bessel, $J_\nu(z)$, and Hankel, $H_\nu(z)$, functions at fixed $z\in\mathbf{C}\setminus\mathbf{0}$, $0\leqslant\arg z<\pi/2$, the following asymtotic estimates hold:

$$J_\nu(z)=\frac{1}{(2\pi\nu)^{1/2}}\left(\frac{ez}{2\nu}\right)^\nu\left(1-\frac{1+3z^2}{12\nu}+\frac{9z^4+78z^2+1}{288\nu^2}+O\left(\frac{1}{\nu^3}\right)\right), \quad \nu\to+\infty, \tag{33}$$

$$H_\nu^{(1)}(z)=\frac{1}{i}\left(\frac{2}{\pi\nu}\right)^{1/2}\left(\frac{2\nu}{ez}\right)^\nu\left(1+\frac{1+3z^2}{12\nu}+\frac{9z^4+78z^2+1}{288\nu^2}+O\left(\frac{1}{\nu^3}\right)\right), \quad \nu\to+\infty. \tag{34}$$

Differentiating formula (29) with respect to the normal n_M and separating, with the help of Lemma 2, in the resulting relationship the principal asymptotical term of the series on the right-hand side, we obtain the representations 7 for $\partial^2 G_{k_j}(M,P)/\partial_M\partial n_P$, $j=1,2$ similar to those for functions $G_{k_j}(M,P)$ [see (26)] but much more bulky. From these expressions, it follows that the considered functions have a logarithmic singularity when their arguments coincide.

V. REDUCTION TO INTEGRAL EQUATIONS

Now we reduce the transmission problem (1)–(6) to a system of Fredholm integral equations. Then, we will use Theorem 1 to prove the unique solvability both of this system and initial problem (1)–(6).

Looking for a solution of transmission problem (1)–(6) in the form of potentials (23) and (24) in the domains $\Omega\setminus\overline{D}$ and D, respectively, taking into account the representations for functions $G_{k_j}(M,P)$ and their normal derivatives obtained above, and using the properties of harmonic potentials,[9] we derive the relationships that are valid on ∂D:

$$u_1(M)-u_2(M)=\frac{1}{2}(\lambda_1^{-1}+\lambda_2^{-1})\phi(M)+\int_{\partial D}\left(\lambda_1^{-1}\frac{\partial G_{k_1}(M,P)}{\partial n_P}-\lambda_2^{-1}\frac{\partial G_{k_2}(M,P)}{\partial n_P}\right)\phi(P)dl_P$$
$$+\int_{\partial D}(\lambda_1^{-1}c_1G_{k_1}(M,P)-\lambda_2^{-1}c_2G_{k_2}(M,P))\psi(P)dl_P, \tag{35}$$

$$\lambda_1 \frac{\partial u_1(M)}{\partial n_M} - \lambda_2 \frac{\partial u_2(M)}{\partial n_M} = -\frac{1}{2}(c_1+c_2)\psi(M) + \int_{\partial D}\left(\frac{\partial^2 G_{k_1}(M,P)}{\partial n_M \partial n_P} - \frac{\partial^2 G_{k_2}(M,P)}{\partial n_M \partial n_P}\right)\phi(P)\,dl_P$$
$$+ \int_{\partial D}\left(c_1 \frac{\partial G_{k_1}(M,P)}{\partial n_M} - c_2 \frac{\partial G_{k_2}(M,P)}{\partial n_M}\right)\psi(P)\,dl_P, \quad M \in \partial D, \quad (36)$$

where the following representations are valid:

$$\frac{\partial^2 G_{k_1}(M,P)}{\partial n_M\, \partial n_P} - \frac{\partial^2 G_{k_2}(M,P)}{\partial n_M\, \partial n_P} = \sum_{q=1}^{8} G_q^{(2)}(k_1,k_2;M,P), \quad (37)$$

$$G_1^{(2)}(k_1,k_2;M,P) = \sum_{m=1}^{\infty} \left\{ \frac{i\pi}{\Phi} \begin{bmatrix} k_1^2 J'_{\nu_m}(k_1 r_P) H_{\nu_m}^{(1)\prime}(k_1 r_M) - k_2^2 J'_{\nu_m}(k_2 r_P) H_{\nu_m}^{(1)\prime}(k_2 r_M) \\ k_1^2 J'_{\nu_m}(k_1 r_M) H_{\nu_m}^{(1)\prime}(k_1 r_P) - k_2^2 J'_{\nu_m}(k_2 r_M) H_{\nu_m}^{(1)\prime}(k_2 r_P) \end{bmatrix} \right.$$
$$\left. + RA(r_M,r_P;k_1,k_2)\left(\frac{r_P}{r_M}\right)^{\nu_m R} + B_1^{(-)}(r_M,r_P;k_1,k_2)\frac{1}{\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_m R} \right\}$$
$$\times \sin(\nu_m \phi_M)\sin(\nu_m \phi_P)\cos\alpha(M)\cos\alpha(P), \quad (38)$$

$$G_2^{(2)}(k_1,k_2;M,P) = -\sum_{m=1}^{\infty} \left\{ \frac{i\pi\nu_m}{r_M \Phi} \begin{bmatrix} k_1 J'_{\nu_m}(k_1 r_P) H_{\nu_m}^{(1)}(k_1 r_M) - k_2 J'_{\nu_m}(k_2 r_P) H_{\nu_m}^{(1)}(k_2 r_M) \\ k_1 J_{\nu_m}(k_1 r_M) H_{\nu_m}^{(1)\prime}(k_1 r_P) - k_2 J_{\nu_m}(k_2 r_M) H_{\nu_m}^{(1)\prime}(k_2 r_P) \end{bmatrix} \right.$$
$$\left. - A(r_M,r_P;k_1,k_2)\left(\frac{r_P}{r_M}\right)^{\nu_m R} - RB_2^{(+)}(r_M,r_P;k_1,k_2)\frac{1}{\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_m R} \right\}$$
$$\times \cos(\nu_m \phi_M)\sin(\nu_m \phi_P)\sin\alpha(M)\cos\alpha(P), \quad (39)$$

$$G_3^{(2)}(k_1,k_2;M,P) = -\sum_{m=1}^{\infty} \left\{ \frac{i\pi\nu_m}{r_P \Phi} \begin{bmatrix} k_1 J_{\nu_m}(k_1 r_P) H_{\nu_m}^{(1)\prime}(k_1 r_M) - k_2 J_{\nu_m}(k_2 r_P) H_{\nu_m}^{(1)\prime}(k_2 r_M) \\ k_1 J'_{\nu_m}(k_1 r_M) H_{\nu_m}^{(1)}(k_1 r_P) - k_2 J'_{\nu_m}(k_2 r_M) H_{\nu_m}^{(1)}(k_2 r_P) \end{bmatrix} \right.$$
$$\left. + A(r_M,r_P;k_1,k_2)\left(\frac{r_P}{r_M}\right)^{\nu_m R} + RB_2^{(-)}(r_M,r_P;k_1,k_2)\frac{1}{\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_m R} \right\}$$
$$\times \sin(\nu_m \phi_M)\cos(\nu_m \phi_P)\cos\alpha(M)\sin\alpha(P), \quad (40)$$

$$G_4^{(2)}(k_1,k_2;M,P) = \sum_{m=1}^{\infty} \left\{ \frac{i\pi\nu_m^2}{r_M r_P \Phi} \begin{bmatrix} J_{\nu_m}(k_1 r_P) H_{\nu_m}^{(1)}(k_1 r_M) - J_{\nu_m}(k_2 r_P) H_{\nu_m}^{(1)}(k_2 r_M) \\ J_{\nu_m}(k_1 r_M) H_{\nu_m}^{(1)}(k_1 r_P) - J_{\nu_m}(k_2 r_M) H_{\nu_m}^{(1)}(k_2 r_P) \end{bmatrix} \right.$$
$$\left. - RA(r_M,r_P;k_1,k_2)\left(\frac{r_P}{r_M}\right)^{\nu_m R} - B_1^{(+)}(r_M,r_P;k_1,k_2)\frac{1}{\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_m R} \right\}$$
$$\times \cos(\nu_m \phi_M)\cos(\nu_m \phi_P)\sin\alpha(M)\sin\alpha(P), \quad (41)$$

$$G_5^{(2)}(k_1,k_2;M,P)=[-\operatorname{sign}(r_M-r_P)\tilde{A}(r_M,r_P;k_1,k_2)\Psi_1(M,P) \\ -\tilde{B}_1^{(-)}(r_M,r_P;k_1,k_2)\Psi_2(M,P)]\cos\alpha(M)\cos\alpha(P), \tag{42}$$

$$G_6^{(2)}(k_1,k_2;M,P)=[-\tilde{A}(r_M,r_P;k_1,k_2)\Psi_3(M,P) \\ -R\tilde{B}_2^{(+)}(r_M,r_P;k_1,k_2)\Psi_4(M,P)]\sin\alpha(M)\cos\alpha(P), \tag{43}$$

$$G_7^{(2)}(k_1,k_2;M,P)=[\tilde{A}(r_M,r_P;k_1,k_2)\Psi_5(M,P) \\ +R\tilde{B}_2^{(-)}(r_M,r_P;k_1,k_2)\Psi_6(M,P)]\cos\alpha(M)\sin\alpha(P), \tag{44}$$

$$G_8^{(2)}(k_1,k_2;M,P)=[R\tilde{A}(r_M,r_P;k_1,k_2)\Psi_7(M,P) \\ +\tilde{B}_1^{(+)}(r_M,r_P;k_1,k_2)\Psi_8(M,P)]\sin\alpha(M)\sin\alpha(P), \tag{45}$$

$$A(r_M,r_P;k_1,k_2)=(k_1^2-k_2^2)\frac{r_M^2-r_P^2}{4\Phi r_M r_P},$$

$$B_1^{(\pm)}(r_M,r_P;k_1,k_2)=\frac{k_1^2-k_2^2}{32\Phi r_M r_P}[(k_1^2+k_2^2)(r_M^2-r_P^2)^2\pm 8(r_M^2+r_P^2)],$$

$$B_2^{(\pm)}(r_M,r_P;k_1,k_2)=\frac{k_1^2-k_2^2}{32\Phi r_M r_P}[(k_1^2+k_2^2)(r_M^2-r_P^2)\pm 8](r_M^2-r_P^2),$$

$$\tilde{A}(r_M,r_P;k_1,k_2)=\tfrac{1}{8}A(r_M,r_P;k_1,k_2),$$

$$\tilde{B}_1^{(\pm)}=\frac{\Phi}{4\pi}B_1^{(\pm)}(r_M,r_P;k_1,k_2),\qquad \tilde{B}_2^{(\pm)}=\frac{\Phi}{2\pi}B_2^{(\pm)}(r_M,r_P;k_1,k_2),$$

$$\Psi_{1,7}(M,P)=\mp\frac{\cos\frac{\pi}{\Phi}(\phi_M+\phi_P)-\left(\frac{r_P}{r_M}\right)^{R(\pi/\Phi)}}{\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_P}{r_M}\right)+\sin^2\frac{\pi}{2\Phi}(\phi_M+\phi_P)}+\frac{\cos\frac{\pi}{\Phi}(\phi_M-\phi_P)-\left(\frac{r_P}{r_M}\right)^{R(\pi/\Phi)}}{\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_P}{r_M}\right)+\sin^2\frac{\pi}{2\Phi}(\phi_M-\phi_P)},$$

$$\Psi_{3,5}(M,P)=\frac{\sin\frac{\pi}{\Phi}(\phi_M+\phi_P)}{\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_P}{r_M}\right)+\sin^2\frac{\pi}{2\Phi}(\phi_M+\phi_P)}\mp\frac{\sin\frac{\pi}{\Phi}(\phi_M-\phi_P)}{\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_P}{r_M}\right)+\sin^2\frac{\pi}{2\Phi}(\phi_M-\phi_P)},$$

$$\Psi_{4,6}(M,P)=\Phi_1(M,P)\mp\Phi_2(M,P),\qquad \Psi_2(M,P)=\ln\Psi(M,P),$$

$$\Psi_8(M,P)=-\ln\left\{16\left[\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_P}{r_M}\right)+\sin^2\frac{\pi}{2\Phi}(\phi_M-\phi_P)\right]\right. \\ \left.\times\left[\sinh^2\left(\frac{\pi}{2\Phi}\ln\frac{r_P}{r_M}\right)+\sin^2\frac{\pi}{2\Phi}(\phi_M+\phi_P)\right]\right\}-R\frac{2\pi}{\Phi}\ln\frac{r_P}{r_M},$$

$$\frac{\partial G_{k_j}(M,P)}{\partial n_M}=\frac{1}{4\pi}\frac{\partial}{\partial n_M}\ln\Psi(M,P)+\sum_{q=1}^{3}G_q^{(3)}(k_j;M,P),\quad j=1,2; \tag{46}$$

in the latter expression,

$$G_1^{(3)}(k_j;M,P)=\frac{i\pi}{\Phi}\cos\alpha(M)\sum_{m=1}^{\infty}\left\{k_j\begin{bmatrix}J_{\nu_m}(k_jr_P)H_{\nu_m}^{(1)\prime}(k_jr_M)\\J'_{\nu_m}(k_jr_M)H_{\nu_m}^{(1)}(k_jr_P)\end{bmatrix}-R\frac{i}{\pi r_M}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right.$$
$$\left.-\frac{i}{\pi r_m}\frac{k_j^2(r_M^2-r_P^2)}{4\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right\}\sin(\nu_m\phi_M)\sin(\nu_m\phi_P), \tag{47}$$

$$G_2^{(3)}(k_j;M,P)=\frac{i\pi}{\Phi}\sin\alpha(M)\sum_{m=1}^{\infty}\left\{\frac{\nu_m}{r_M}\begin{bmatrix}-J_{\nu_m}(k_jr_P)H_{\nu_m}^{(1)}(k_jr_M)\\-J_{\nu_m}(k_jr_M)H_{\nu_m}^{(1)}(k_jr_P)\end{bmatrix}-\frac{i}{\pi r_M}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right.$$
$$\left.-R\frac{i}{\pi r_M}\frac{k_j^2(r_M^2-r_P^2)}{4\nu_m}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right\}\cos(\nu_m\phi_M)\sin(\nu_m\phi_P), \tag{48}$$

$$G_3^{(3)}(k_j;M,P)=-\cos\alpha(M)\frac{k_j^2(r_M^2-r_P^2)}{16\pi r_M}\ln\Psi(M,P)$$
$$-\operatorname{sign}(r_M-r_P)\sin\alpha(M)\frac{k_j^2(r_M^2-r_P^2)}{8\pi r_M}(\Phi_1(M,P)-\Phi_2(M,P)).$$

As follows from Lemma 1, general terms of series (38)–(41), (47), and (48) decrease as

$$O\left[\frac{1}{\nu_m^2}\left(\frac{r_P}{r_M}\right)^{\nu_mR}\right].$$

VI. SOLVABILITY OF INTEGRAL EQUATIONS AND BOUNDARY VALUE PROBLEM

Introducing the integral operators

$$(A_{11}\phi)(M):=2\int_{\partial D}\left(\lambda_1^{-1}\frac{\partial G_{k_1}(M,P)}{\partial n_P}-\lambda_2^{-1}\frac{\partial G_{k_2}(M,P)}{\partial n_P}\right)\phi(P)dl_P, \tag{49}$$

$$(A_{12}\psi)(M):=2\int_{\partial D}(\lambda_1^{-1}c_1G_{k_1}(M,P)-\lambda_2^{-1}c_2G_{k_2}(M,P))\psi(P)dl_P, \tag{50}$$

$$(A_{21}\phi)(M):=2\int_{\partial D}\left(\frac{\partial^2G_{k_1}(M,P)}{\partial n_M\,\partial n_P}-\frac{\partial^2G_{k_2}(M,P)}{\partial n_M\,\partial n_P}\right)\phi(P)dl_P, \tag{51}$$

$$(A_{22}\psi)(M):=2\int_{\partial D}\left(c_1\frac{\partial G_{k_1}(M,P)}{\partial n_M}-c_2\frac{\partial G_{k_2}(M,P)}{\partial n_M}\right)\psi(P)dl_P, \tag{52}$$

defined by the right-hand sides of relationships (35) and (36), we see that these operators have weakly singular kernels. In fact, for operators A_{11}, A_{12}, A_{22}, this statement follows from representations (26), (29), and (46) (the kernel of operator A_{12} has a logarithmic singularity and the kernels of A_{11}, A_{22} are continuous), and for operator A_{21}, it follows from the fact that the sum of two last summands on the right-hand sides of (42) and (45) have singularities of the form

$(1/4\pi)(k_1^2-k_2^2)\ln(1/r_{MP})$while the rest of summands in (37) are bounded as $P\rightarrow M$. Therefore, these operators are compact in the space $C(\partial D)$ and they map $C(\partial D)$ into $C^{0,\alpha}(\partial D)$ and $C^{0,\alpha}(\partial D)$, into $C^{1,\alpha}(\partial D)$.

Hence, using operator notation (49)–(52), relationships (35)–(36), and the transmission conditions, we can prove the following assertion.

Theorem 2: *The potentials u_1 and u_2 defined by formulas (23) and (24) solve transmission problem (1)–(6) provided that densities ϕ and ψ solve the system of integral equations with weakly singular kernels,*

$$(\lambda_1^{-1}+\lambda_2^{-1})\phi+A_{11}\phi+A_{12}\psi=2f,$$

$$(c_1+c_2)\psi-A_{21}\phi-A_{22}\psi=-2g. \tag{53}$$

Introduce in the space $C(\partial D)\times C(\partial D)$ the operators E and A defined as

$$E:=\begin{bmatrix}(\lambda_1^{-1}+\lambda_2^{-1})I & 0\\ 0 & (c_1+c_2)I\end{bmatrix},\quad A:=\begin{bmatrix}-A_{11} & -A_{12}\\ A_{21} & A_{22}\end{bmatrix},$$

where I is the identity operator on $C(\partial D)$. If we set $\Xi:=[{\psi \atop \phi}]$, then the system of integral equations (53) can be rewritten in the form

$$(E-A)\Xi=F, \tag{54}$$

where

$$F:=\begin{bmatrix}f\\ -g\end{bmatrix}.$$

Theorem 3: *Under the conditions of Theorem 1, system of integral equations (53) [or what is the same, integral equation (54)] has a unique solution $\phi\in C^{1,\alpha}(\partial D)$, $\psi\in C^{0,\alpha}(\partial D)$ if $\lambda_1+\lambda_2\neq 0$ and the numbers c_1 and c_2 are chosen so that the following conditions are valid:*

$$(i)\quad c_1+c_2\neq 0,$$

$$(ii)\quad \eta:=\frac{c_2\overline{k_1^2}}{c_1\overline{k_2^2}}\in\mathbf{R},\quad \text{where}\quad \eta>0.$$

Proof. A is a compact operator; therefore, using the assumptions introduced above, the fact that E is a bounded and invertible operator in $C(\partial D)\times C(\partial D)$, and Corollaries 1.17 and 1.20, Colton and Kress,[6] we can show that the inhomogeneous integral equation (54) has the unique solution $\Xi=({\psi \atop \phi})\in C(\partial D)\times C(\partial D)$ if the homogeneous equation,

$$(E-A)\Xi=0, \tag{55}$$

has only the trivial solution. Moreover, in this case, due to properties of the operator of system (53) and the smoothness condition $f\in C^{1,\alpha}(\partial D)$, $g\in C^{0,\alpha}(\partial D)$, we obtain that $\phi\in C^{1,\alpha}(\partial D)$, $\psi\in C^{0,\alpha}(\partial D)$.

So, it remains to prove that homogeneous integral Eq. (55) has only the trivial solution. Let $\Xi=[{\psi \atop \phi}]$ be a solution to Eq. (55). Then, potentials u_1 and u_2 defined by (23) and (24) solve homogeneous transmission problem (1)–(6). Therefore, from the uniqueness Theorem 1, we have

$$u_1=0,\ \text{in}\ \Omega\setminus\overline{D},\quad u_2=0,\ \text{in}\ \overline{D}. \tag{56}$$

Using formulas (23) and (24), representations for function $G_{k_j}(M,P)$ and its normal derivatives obtained above, and the discontinuity properties of potentials, we obtain the equalities

$$u_i^+ - u_i^- = \lambda_i^{-1}\phi, \quad \text{on} \quad \partial D, \quad \frac{\partial u_i^+}{\partial n} - \frac{\partial u_i^-}{\partial n} = -\lambda_i^{-1} c_i \psi, \quad \text{on} \quad \partial D, \quad i=1,2, \tag{57}$$

where $u_i^+(M) := \lim_{P\to M, P\in\Omega\setminus\overline{D}} u(P)$, $u_i^-(M) := \lim_{P\to M, P\in D} u(P), M\in\partial D$. Relationships (57) and (56) imply that

$$\lambda_2 u_2^+ + \lambda_1 u_1^- = 0, \quad \frac{\lambda_2}{c_2}\frac{\partial u_2^+}{\partial n} + \frac{\lambda_1}{c_1}\frac{\partial u_1^+}{\partial n} = 0. \tag{58}$$

Now, let us introduce the functions

$$v_2(M) = \lambda_2 u_2(M) = \int_{\partial D}\left(\frac{\partial G_{k_2}(M,P)}{\partial n_P}\phi(P) + c_2 G_{k_2}(M,P)\psi(P)\right) dl_P, \quad M\in\Omega\setminus\overline{D}, \tag{59}$$

$$v_1(M) = -\lambda_1 u_1(M) = -\int_{\partial D}\left(\frac{\partial G_{k_1}(M,P)}{\partial n_P}\phi(P) + c_1 G_{k_1}(M,P)\psi(P)\right) dl_P, \quad M\in D. \tag{60}$$

Then, from the above considerations and formulas (58)–(60), it follows that functions $v_1 \in C^2(D)\cap C^1(\overline{D})$ and $v_2\in C^2(\Omega\setminus\overline{D})\cap C^1(\Omega\setminus D)\cap C(\overline{\Omega}\setminus\overline{D})$ satisfy the homogeneous transmission problem

$$\Delta v_1 + k_1^2 v_1 = 0, \quad \text{in} \quad D, \quad \Delta v_2 + k_2^2 v_2 = 0, \quad \text{in} \quad \Omega\overline{D},$$

$$v_2 - v_1 = 0, \ \text{on } \partial D, \quad \frac{1}{c_2}\frac{\partial v_2}{\partial n} - \frac{1}{c_1}\frac{\partial v_1}{\partial n} = 0, \ \text{on } \partial D, \quad v_1 = 0, \quad \text{on } \partial\Omega,$$

$$\frac{\partial v_2}{\partial r} - ik_2 v_2 = o\left(\frac{1}{\sqrt{r}}\right), \quad r\to\infty,$$

$$\int_{\Omega\cap\delta}(|v_2|^2 + |\text{grad} v_2|^2)\, dS < \infty.$$

Condition (ii) enables us to apply in this case the uniqueness theorem, so that $v_1=0$ in $\overline{D}$ and $v_2=0$ in $\Omega\setminus\overline{D}$; therefore, $u_1=0$ in $\overline{D}$ and $u_2=0$ in $\Omega\setminus\overline{D}$. Using relations (57) and (56), we see that $\phi=\psi=0$ on ∂D. The theorem is proved.

Theorems 3 and 2 imply the following result.

Corollary 1:

Assume that conditions of Theorem 3 are valid and $\phi\in C^{1,\alpha}(\partial D)$, $\psi\in C^{0,\alpha}(\partial D)$ *is a solution of the system of integral equations* (53). *Then, potentials* (23) *and* (24) *with densities* ϕ *and* ψ *solve transmission problem* (1)–(6).

In conclusion, we make the following remarks.

1. The above analysis shows that representations (26), (29), (46), and (37) for functions $G_{k_j}(M,P)$ and their normal derivatives are not only of theoretical but also of practical interest, since they enable one to compute the kernels of the system of integral equations in the case when their arguments are close. In this connection, it should be noted that formula (25) and the representations obtained as a result of its differentiation are not suitable for this purpose. In fact, series (25) are not absolutely convergent, its general term decreases as $O(1/\nu_M)$ for $r_P=r_M$, ϕ_P

$\neq \phi_M$, whereas the series obtained by termwise differentiation with respect to r and ϕ are divergent for $r_P = r_M$, $\phi_P \neq \phi_M$. In addition to this, unlike formulas (26) and (37), the logarithmic singularity is not extracted explicitly from series (25).

2. Transmission problem (1)–(6) may be reduced to a uniquely solvable system of integral equations using another approach, when a solution to (1)–(6) is sought in the form (23) and (24), where $G_{k_2}(M,P) = (i/4) H_0^{(1)}(k_2 r_{M,P})$. In order to verify the validity of Theorems 2 and 3 in this case, one should assume, at the end of the proof of Theorem 3 when referring to Theorem 1 that $\Omega = \mathbf{R}^2$ (the corresponding result can be obtained by literally repeating the arguments).[4]

3. All the results obtained in this paper can be generalized without any serious changes to the transmission problem in the case when the boundary ∂D consists of a finite number of closed nonintersecting Lyapynov contours.

ACKNOWLEDGMENTS

This paper is based on the research carried out under the program Research in Pairs supported by Mathematisches Forschungsinstitut Oberwolfach (Germany, January, 1996).

[1] W. D. Kupradze, *Randwertaufgaben der Scwingungstheorie und, Integraleichungen* (Deutshe-Verlag der Wissenschaften, Berlin, 1956).
[2] P. Werner, Arch. Ration. Mech. Anal. **6**, 231 (1960).
[3] R. Kittappa and R. E. Kleinman, J. Math. Phys. **16**, 421 (1975).
[4] R. Kress and G. F. Roach, J. Math. Phys. **19**, 1433 (1978).
[5] H. M. Macdonald, *Electric Waves* (Cambridge U. P., Cambridge, 1902).
[6] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory* (Wiley, New York, 1983).
[7] Yu. K. Podlipenko, Ukr. Mat. Zh. **45**, 430 (1993).
[8] Yu. K. Podlipenko, Ukr. Mat. Zh. **45**, 535 (1993).
[9] N. M. Günter, *Die Potentialtheorie und ihre Anwendungen Auf der Mathematischen Physik* (Teubner, Leipzig, 1957).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Jacobi's principle for magnetic interactions

G. F. Torres del Castillo[a)]
Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, 72570 Puebla, Pue., Mexico

(Received 24 September 1998; accepted for publication 3 June 1999)

It is shown that the trajectories of a charged particle in a static magnetic field and a velocity-independent potential in a three-dimensional space are (the projection of) the geodesics of a suitably defined metric in a four-dimensional space. It is shown that each one-parameter group of isometries of the original configuration space that leaves the magnetic field invariant gives rise to a one-parameter group of isometries of the metric defined on the four-dimensional space and, hence, to a constant of the motion. It is also shown, similarly, that the Schrödinger equation for a charged particle in a static magnetic field is equivalent to the Schrödinger equation for a free particle in the four-dimensional space mentioned above. © *1999 American Institute of Physics.* [S0022-2488(99)02809-1]

I. INTRODUCTION

As is well known in classical mechanics, the orbits in the configuration space of a particle, or a system of particles, whose Hamiltonian is of the form

$$H=\tfrac{1}{2}g^{ij}p_ip_j+V(q^i),$$

$(i,j,\ldots=1,\ldots,n)$, are the geodesics of the metric

$$(E-V)g_{ij}\,dq^i\,dq^j$$

(Jacobi's principle). This result does not apply to the case where there are magnetic forces present; however, as shown in Refs. 1 and 2, in the latter case, the following result holds: the orbits corresponding to the Hamiltonian,

$$H=\frac{1}{2}\,\gamma^{ij}\left(p_i-\frac{e}{c}A_i\right)\left(p_j-\frac{e}{c}A_j\right)+V,$$

where γ^{ij}, A_i, and V depend only on the coordinates q^i, are projections of the geodesics of an $(n+1)$-dimensional space with a metric tensor made out of γ^{ij}, A_i, and V. [Actually, in Refs. 1 and 2, only the case $n=3$ was considered, but it is easy to see that the expressions obtained there apply to all values of n; furthermore, in Refs. 1 and 2 it was assumed that the $(n+1)$th coordinate is a time coordinate, but one can leave the nature of the additional variable unspecified.]

Thus, by increasing the dimension of the space, one can absorb the electromagnetic forces, and other conservative forces present, into the geometry. This geometrization of the electromagnetic interaction also allows one to deal with the gauge transformations in the Hamiltonian formalism in an easy way (see below) (cf. Ref. 3). The manner in which the magnetic interaction is incorporated into the geometry of a space of higher dimension here and in Refs. 1 and 2 is similar to that followed in the Kaluza–Klein theory, where the electromagnetic and gravitational fields in the four-dimensional spacetime of general relativity are incorporated into the metric of a five-dimensional space in such a way that the Einstein–Maxwell equations correspond to the vanishing of the Ricci tensor of the five-dimensional metric. In fact, some of the equations given in Sec. II

a)Electronic mail: gtorres@fcfm.buap.mx

[e.g., Eqs. (3)–(5), (9), and (10)] are analogous to the corresponding formulas in the Kaluza–Klein theory; however, since we are not interested in the curvature of the auxiliary higher dimension space, we have some freedom in the definition of the metric of the auxiliary space, which is employed to make the parameter of the geodesics of the auxiliary space to coincide with the time, even when there is a velocity-independent potential present [Eq. (6) below].

Our aim in this paper is to show that a similar connection exists in the case of the Schrödinger equation; specifically, we shall show that the solution of the (time-independent) Schrödinger equation for a charged particle in a static magnetic field can be obtained from the solution of the Schrödinger equation for a free particle in an $(n+1)$-dimensional space. It should be noticed that in the context of the Kaluza–Klein theory, one would consider an equation like the Klein–Gordon equation, rather than the Schrödinger equation. In Sec. II we consider the motion of a charged particle in a magnetic field in the framework of classical mechanics, showing the relationship of this problem with the motion of a free particle in a space with an extra dimension. The metric of this higher dimension space is different from that introduced in Refs. 1 and 2 and, as pointed out above, it has the advantage that the time parameters conjugate to the original and the auxiliary Hamiltonians coincide. We show that even though the vector potential may not be invariant under a group of isometries that leave the magnetic field invariant, one can find a local group of isometries of the metric defined in the extended space, which, however, may not be isomorphic to the original group. Two examples are given, which turn out to be related to metrics in four-dimensional spaces of interest in general relativity. A similar result can be obtained in the Kaluza–Klein theory; for any group of isometries of the spacetime metric that leave the electromagnetic field tensor invariant, it is possible to find a local group of isometries of the metric of the five-dimensional space. In Sec. III we consider the Schrödinger equation for a charged particle in a static magnetic field, showing that the solutions of this equation can be obtained from those of the Schrödinger equation for a free particle in the extended space employed in Sec. II. The generators of the isometries of this latter space correspond to operators that commute with the Hamiltonian.

II. MAGNETIC FIELDS AND GEOMETRY

The Hamiltonian of a particle of mass m and electric charge e in a static magnetic field and a velocity-independent potential ϕ is given by

$$H=\frac{1}{2m}\,\gamma^{ij}\left(p_i-\frac{e}{c}A_i\right)\left(p_j-\frac{e}{c}A_j\right)+\phi, \tag{1}$$

where (γ^{ij}) is the inverse of the matrix (γ_{ij}) formed by the components of the metric tensor with respect to a coordinate system x^i $(i,j,\ldots=1,2,3)$, A_i are the components of a (time-independent) vector potential corresponding to the given magnetic field and $\phi=\phi(x^i)$. We now introduce the auxiliary Hamiltonian,

$$h\equiv\frac{1}{2m}g^{\alpha\beta}p_\alpha p_\beta, \tag{2}$$

of a free particle of mass m in a four-dimensional space $(\alpha,\beta,...=0,1,2,3)$, with

$$g^{00}=\frac{1}{g_{00}}+\frac{A_iA^i}{c^2},\quad g^{0i}=-\frac{A^i}{c},\quad g^{ij}=\gamma^{ij}, \tag{3}$$

where g_{00} is a function of x^i only, to be specified later, and $A^i=\gamma^{ij}A_j$. It can be readily verified that the inverse of $(g^{\alpha\beta})$, denoted by $(g_{\alpha\beta})$, is given by

$$g_{0i}=g_{00}\frac{A_i}{c},\quad g_{ij}=\gamma_{ij}+g_{00}\frac{A_iA_j}{c^2}; \tag{4}$$

therefore, the orbits corresponding to the Hamiltonian (2) are the geodesics of the metric,

$$ds^2 = g_{\alpha\beta}\, dx^\alpha\, dx^\beta = g_{00}\left(dx^0 + \frac{A_i}{c} dx^i \right)^2 + \gamma_{ij}\, dx^i\, dx^j. \tag{5}$$

From Eqs. (1)–(3) it is easy to see that

$$h = H + \frac{(p_0 - e)}{2m} \frac{A^i}{c}\left(\frac{A_i}{c}(p_0 + e) - 2p_i \right) + \frac{p_0^2}{2mg_{00}} - \phi. \tag{6}$$

Thus, choosing g_{00} in such a way that

$$\frac{e^2}{2mg_{00}} = \phi + \text{const}, \tag{7}$$

it follows that on the submanifold

$$p_0 = e, \tag{8}$$

$\partial h/\partial x^i = \partial H/\partial x^i$ and $\partial h/\partial p_i = \partial H/\partial p_i$. Hence, the projection on the phase space of the original Hamiltonian (1) of the orbits determined by h [Eq. (2)], with $p_0 = e$, coincides with the orbits of H. Since the Hamiltonian h does not depend on x^0, p_0 is a constant of the motion.

Thus, we have shown that the solutions of the equations of motion determined by the Hamiltonian (1) correspond to the geodesics of the metric (5). For a given magnetic field, the vector potential A_i is defined up to the gauge transformations,

$$A_i' = A_i + \partial_i \xi, \tag{9}$$

where ξ is an arbitrary function of the x^i. From Eq. (5) we see that the gauge transformation (9) is equivalent to the substitution of the coordinate x^0 by

$$x'^0 = x^0 - \xi/c. \tag{10}$$

Applying Hamilton's equations to the Hamiltonian (2), it follows that $p_\alpha = mg_{\alpha\beta}\dot{x}^\beta$. Making use of Eqs. (4), this leads to

$$p_0 = mg_{00}\left(\dot{x}^0 + \frac{A_i \dot{x}^i}{c} \right), \quad p_i = m\gamma_{ij}\dot{x}^j + mg_{00}\frac{A_i}{c}\left(\dot{x}^0 + \frac{A_j \dot{x}^j}{c} \right),$$

hence, $p_i = m\gamma_{ij}\dot{x}^j + A_i p_0/c$, which, on the submanifold $p_0 = e$ [Eq. (8)] reduces to the well-known relation

$$p_i = m\gamma_{ij}\dot{x}^j + \frac{e}{c} A_i, \tag{11}$$

which also follows from the Hamiltonian (1).

In the rest of this paper we shall assume that only the magnetic field is present in (1), taking $\phi = 0$; therefore, in what follows, g_{00} will be taken as a constant [see Eq. (7)].

The symmetry of the magnetic field need not be shared by the vector potential; however, if the magnetic field is invariant under a one-parameter group of motions on the original configuration space, one can find a one-parameter (local) group of isometries of (5), which leads to a constant of the motion for both Hamiltonians. More precisely, let $K^i\partial_i$ be a Killing vector field (i.e., the generator of a one-parameter group of isometries) of the metric $\gamma_{ij}\, dx^i\, dx^j$ of the original configuration space,

$$K^i \partial_i \gamma_{jk} + 2\gamma_{i(j}\partial_{k)}K^i = 0, \tag{12}$$

where the parentheses denote symmetrization on the indices enclosed, which leaves the magnetic field 2-form $B = d(A_i\,dx^i)$ invariant; this means that the Lie derivative of $d(A_i\,dx^i)$ along $K^i\partial_i$ vanishes. Then, since the Lie derivative commutes with the differential, the Lie derivative of $A_i\,dx^i$ along $K^i\partial_i$ must be locally exact, i.e., there exists (locally) a function K^0 of the x^i only, such that $\pounds_{K^i\partial_i}(A_j\,dx^j) = -c\,dK^0$, or

$$K^i \partial_i A_j + A_i \partial_j K^i = -c\,\partial_j K^0. \tag{13}$$

Then, we can show that the vector field $K^\alpha\partial_\alpha = K^0\partial_0 + K^i\partial_i$ is a Killing vector field of the metric $g_{\alpha\beta}\,dx^\alpha\,dx^\beta$. Indeed, the components of the Lie derivative of $g_{\alpha\beta}\,dx^\alpha\,dx^\beta$ along $K^\alpha\partial_\alpha$,

$$K^\rho \partial_\rho g_{\alpha\beta} + 2g_{\rho(\alpha}\partial_{\beta)}K^\rho = K^i \partial_i g_{\alpha\beta} + 2g_{0(\alpha}\partial_{\beta)}K^0 + 2g_{i(\alpha}\partial_{\beta)}K^i,$$

are

$$K^i \partial_i g_{00} + 2g_{00}\partial_0 K^0 + 2g_{i0}\partial_0 K^i,$$

which vanishes since the components K^α do not depend on x^0 and in the present case ($\phi=0$) g_{00} is constant,

$$K^i \partial_i g_{0j} + g_{00}\partial_j K^0 + g_{i0}\partial_j K^i = K^i \partial_i\left(g_{00}\frac{A_j}{c}\right) + g_{00}\partial_j K^0 + g_{00}\frac{A_i}{c}\partial_j K^i = 0,$$

by virtue of Eqs. (4) and (13), and

$$\begin{aligned} &K^i \partial_i g_{jk} + 2g_{0(j}\partial_{k)}K^0 + 2g_{i(j}\partial_{k)}K^i \\ &\quad = K^i \partial_i \gamma_{jk} + 2\gamma_{i(j}\partial_{k)}K^i + 2\frac{g_{00}}{c^2}A_{(j}(K^i \partial_i A_{k)} + c\,\partial_{k)}K^0 + A_i\,\partial_{k)}K^i) = 0, \end{aligned}$$

where we have made use of Eqs. (4), (12), and (13). [The conclusion can also be obtained directly from Eq. (5), by noting that Eq. (13) means that the Lie derivative of $dx^0 + A_i\,dx^i/c$ along $K^\alpha\partial_\alpha$ vanishes.] It may be noticed that the same conclusion applies if ϕ is invariant under the group generated by $K^i\partial_i$, not necessarily a constant.

The invariance of the Hamiltonian (2) under the flow generated by $K^\alpha\partial_\alpha$ implies the conservation of $K^\alpha p_\alpha$, therefore, taking into account the constraint (8) and the fact that the Poisson bracket between $K^\alpha p_\alpha$ and p_0 vanishes, the function

$$K^i p_i + eK^0 \tag{14}$$

is also a constant of the motion for the original Hamiltonian (1) (see the examples below). Note that K^0 is defined by Eq. (13) up to an additive constant.

If there exist r Killing vector fields, $K^i_{(a)}\partial_i$ $(a,b,\ldots=1,2,\ldots,r)$, of $\gamma_{ij}\,dx^i\,dx^j$ that leave the magnetic field invariant such that

$$[K^i_{(a)}\partial_i, K^j_{(b)}\partial_j] = c^d_{ab}K^i_{(d)}\partial_i, \tag{15}$$

for some constants c^d_{ab}, then each Killing vector field can be locally ''lifted'' to a Killing vector field, $K^\alpha_{(a)}\partial_\alpha$ of the metric (5), with $K^0_{(a)}$ given by Eq. (13) up to an additive constant. Making use of Eq. (13) one finds that the commutation relations of the vector fields $K^\alpha_{(a)}\partial_\alpha$ are given by

$$[K^\alpha_{(a)}\partial_\alpha, K^\beta_{(b)}\partial_\beta] = (K^i_{(a)}\partial_i K^0_{(b)} - K^i_{(b)}\partial_i K^0_{(a)})\partial_0 + c^d_{ab}K^i_{(d)}\partial_i = c^d_{ab}K^\alpha_{(d)}\partial_\alpha + f_{ab}\partial_0, \tag{16}$$

where

$$f_{ab} \equiv -c^d_{ab} K^0_{(d)} + K^i_{(a)}\, \partial_i K^0_{(b)} - K^i_{(b)}\, \partial_i K^0_{(a)} \tag{17}$$

are constants that can be different from zero [see, e.g., Eq. (28) below]. (The fact that the f_{ab} are constant can be proven making use of Eq. (13) and of the property of the Lie derivative $\pounds_{[X,Y]} = [\pounds_X, \pounds_Y]$.) Since the vector fields $K^\alpha_{(a)}\partial_\alpha$ (and, hence, the constants of the motion $K^\alpha_{(a)} p_\alpha$ with the Poisson bracket) satisfy the commutation relations (16), if $f_{ab} \neq 0$, they generate a central extension of the Lie algebra generated by the vector fields $K^i_{(a)}\partial_i$. (Note that ∂_0 commutes with the Killing vector fields $K^\alpha_{(a)}\partial_\alpha$.)

Examples

The vector potential given by

$$A_i\, dx^i = -g \cos\theta\, d\varphi, \tag{18}$$

in terms of the spherical coordinates (r, θ, φ), corresponds to the field of a magnetic monopole of charge g. Substituting Eq. (18) and the metric of the Euclidean three-dimensional space into Eq. (5) one obtains the metric

$$ds^2 = g_{00}(g/c)^2 (d\chi + \cos\theta\, d\varphi)^2 + dr^2 + r^2 (d\theta^2 + \sin^2\theta\, d\varphi^2), \tag{19}$$

where we have introduced the dimensionless variable $\chi \equiv -(c/g)x^0$. The magnetic field of the monopole is invariant under the rotations about the origin [though the potential (18) is not]; therefore the vector fields,

$$\begin{aligned} K_{(1)} &\equiv K^i_{(1)}\partial_i = -\sin\varphi\,\partial_\theta - \cot\theta\cos\varphi\,\partial_\varphi, \\ K_{(2)} &\equiv K^i_{(2)}\partial_i = \cos\varphi\,\partial_\theta - \cot\theta\sin\varphi\,\partial_\varphi, \\ K_{(3)} &\equiv K^i_{(3)}\partial_i = \partial_\varphi, \end{aligned} \tag{20}$$

which generate rotations about the coordinate axes in the three-dimensional Euclidean space, give rise to Killing vector fields of (19). Substituting Eqs. (18) and (20) into Eq. (13), one finds that $K^0_{(1)} = -(g/c)\operatorname{cosec}\theta\cos\varphi$, $K^0_{(2)} = -(g/c)\operatorname{cosec}\theta\sin\varphi$, $K^0_{(3)} = 0$ (setting to zero the integration constants); thus, the vector fields,

$$\begin{aligned} \tilde{K}_{(1)} &\equiv K^\alpha_{(1)}\partial_\alpha = -\sin\varphi\,\partial_\theta - \cot\theta\cos\varphi\,\partial_\varphi + \operatorname{cosec}\theta\cos\varphi\,\partial_\chi, \\ \tilde{K}_{(2)} &\equiv K^\alpha_{(2)}\partial_\alpha = \cos\varphi\,\partial_\theta - \cot\theta\sin\varphi\,\partial_\varphi + \operatorname{cosec}\theta\sin\varphi\,\partial_\chi, \\ \tilde{K}_{(3)} &\equiv K^\alpha_{(3)}\partial_\alpha = \partial_\varphi, \end{aligned} \tag{21}$$

are Killing vector fields of (19). One can easily verify that the vector fields $K_{(a)}$ and $\tilde{K}_{(a)}$ obey the commutation relations $[K_{(a)}, K_{(b)}] = -\epsilon_{abc}K_{(c)}$ and $[\tilde{K}_{(a)}, \tilde{K}_{(b)}] = -\epsilon_{abc}\tilde{K}_{(c)}$, respectively (i.e., $f_{ab} = 0$). The three constants of the motion associated with the isometries (21) are [see Eq. (14)]

$$\begin{aligned} L_1 &= -\sin\varphi\, p_\theta - \cot\theta\cos\varphi\, p_\varphi - (eg/c)\operatorname{cosec}\theta\cos\varphi, \\ L_2 &= \cos\varphi\, p_\theta - \cot\theta\sin\varphi\, p_\varphi - (eg/c)\operatorname{cosec}\theta\sin\varphi, \\ L_3 &= p_\varphi, \end{aligned} \tag{22}$$

or, making use of Eq. (11), one finds that $(L_1, L_2, L_3) = \mathbf{r} \times m\dot{\mathbf{r}} - (eg/c)\mathbf{r}/r$. Since the mapping $K^\alpha\partial_\alpha \mapsto K^\alpha p_\alpha$ is a Lie algebra homomorphism, it follows that the functions L_a satisfy the Poisson bracket relations $\{L_a, L_b\} = -\epsilon_{abc}L_c$. It may be pointed out, finally, that the metric (19), with

$g_{00}>0$, is a Euclidean Taub–NUT metric (see, e.g., Ref. 4) and that the vector fields (21) form a basis for the right-invariant vector fields on SO(3), with φ, θ, and χ being Euler angles.

As a second example, we shall consider the potential,

$$A_i\,dx^i=\tfrac{1}{2}B_0(x\,dy-y\,dx), \tag{23}$$

in Cartesian coordinates, where B_0 is a constant, which corresponds to a uniform magnetic field. In this case the four-dimensional metric (5) is given by

$$ds^2=g_{00}(dx^0+(B_0/2c)(x\,dy-y\,dx))^2+dx^2+dy^2+dz^2. \tag{24}$$

The vector fields,

$$K_{(1)}\equiv\partial_x,\quad K_{(2)}\equiv\partial_y,\quad K_{(3)}\equiv\partial_z,\quad K_{(4)}\equiv x\partial_y-y\partial_x, \tag{25}$$

are Killing vector fields of the three-dimensional Euclidean space, that leave invariant the magnetic field generated by the vector potential (23) and whose nonvanishing commutators are

$$[K_{(1)},K_{(4)}]=K_{(2)},\quad [K_{(2)},K_{(4)}]=-K_{(1)}. \tag{26}$$

Making use of Eqs. (13), (23), and (25) one obtains the following Killing vector fields of the metric (24):

$$\tilde{K}_{(1)}\equiv\partial_x-\frac{B_0y}{2c}\partial_0,\quad \tilde{K}_{(2)}\equiv\partial_y+\frac{B_0x}{2c}\partial_0,\quad \tilde{K}_{(3)}\equiv\partial_z,\quad \tilde{K}_{(4)}\equiv x\partial_y-y\partial_x. \tag{27}$$

The only nonvanishing commutators of these vector fields are

$$[\tilde{K}_{(1)},\tilde{K}_{(2)}]=\frac{B_0}{c}\partial_0,\quad [\tilde{K}_{(1)},\tilde{K}_{(4)}]=\tilde{K}_{(2)},\quad [\tilde{K}_{(2)},\tilde{K}_{(4)}]=-\tilde{K}_{(1)}. \tag{28}$$

Thus, in this case $f_{12}=B_0/c$ [Eq. (16)] and the vector fields $\tilde{K}_{(a)}$ are part of a basis of a Lie algebra of dimension 5, which is not isomorphic to the Lie algebra generated by the vector fields (25). The constants of the motion corresponding to the Killing vector fields (27) on the submanifold $p_0=e$, are

$$\begin{gathered}p_x-\frac{eB_0}{2c}y=m\dot{x}-\frac{eB_0}{c}y,\quad p_y+\frac{eB_0}{2c}x=m\dot{y}+\frac{eB_0}{c}x,\\ p_z=m\dot{z},\\ xp_y-yp_x=m(x\dot{y}-y\dot{x})+\frac{eB_0}{2c}(x^2+y^2).\end{gathered} \tag{29}$$

As pointed out in Ref. 1, the metric (24) with $g_{00}=-1$ corresponds to the Som–Raychaudhuri spacetime;[5] thus, as a byproduct, we have shown that this spacetime possesses at least five Killing vector fields.

III. MAGNETIC INTERACTIONS IN THE SCHRÖDINGER EQUATION

Now we shall consider the time-independent Schrödinger equation for a particle of mass m and electric charge e in a static magnetic field with a (time-independent) vector potential **A**,

$$\frac{1}{2m}\left(\frac{\hbar}{i}\boldsymbol{\nabla}-\frac{e}{c}\mathbf{A}\right)^2\psi=E\psi, \tag{30}$$

or, equivalently,

$$-\hbar^2\nabla^2\psi - \frac{\hbar e}{ic}\boldsymbol{\nabla}\cdot(\mathbf{A}\psi) - \frac{\hbar e}{ic}\mathbf{A}\cdot\boldsymbol{\nabla}\psi + \left(\frac{e}{c}\right)^2\mathbf{A}^2\psi = 2mE\psi \tag{31}$$

(without imposing any gauge condition on **A**). If, as in the preceding section, γ_{ij} are the components of the metric tensor with respect to a coordinate system x^i, $(i,j,\ldots=1,2,3)$ and (γ^{ij}) denotes the inverse of (γ_{ij}), then Eq. (31) reads as

$$-\hbar^2\frac{1}{\sqrt{\gamma}}\partial_i(\sqrt{\gamma}\gamma^{ij}\partial_j\psi) - \frac{\hbar e}{ic}\frac{1}{\sqrt{\gamma}}\partial_i(\sqrt{\gamma}A^i\psi) - \frac{\hbar e}{ic}A^i\,\partial_i\psi + \left(\frac{e}{c}\right)^2 A_iA^i\psi = 2mE\psi, \tag{32}$$

where $\gamma\equiv\det(\gamma_{ij})$.

From Eqs. (4) it follows that $g\equiv\det(g_{\alpha\beta})$ is related to γ by means of

$$g = g_{00}\gamma. \tag{33}$$

(Note that g is used with two different meanings.) Therefore, making use of Eqs. (3) and (33), with g_{00} constant, one finds that Eq. (32) amounts to

$$-\hbar^2\frac{1}{\sqrt{g}}\partial_\alpha(\sqrt{g}g^{\alpha\beta}\partial_\beta\Psi) = (2mE + e^2/g_{00})\Psi, \tag{34}$$

provided that

$$\Psi(x^0,x^1,x^2,x^3) = \psi(x^1,x^2,x^3)\exp(iex^0/\hbar). \tag{35}$$

Equation (34) is an explicit expression of

$$-\frac{\hbar^2}{2m}{}^{(4)}\nabla^2\Psi = E'\Psi, \tag{36}$$

where ${}^{(4)}\nabla^2$ is the Laplace operator for the metric (5) and $E' = E + e^2/(2mg_{00})$. Equation (36) is the Schrödinger equation for a free particle of mass m (or the Helmholtz equation) in the four-dimensional space with metric (5).

Note that, by contrast with the operator appearing on the left-hand side of Eq. (30), the left-hand side of Eq. (36) does not contain the charge of the particle. Since the metric (5) does not depend on x^0, the operator $C\equiv -i\hbar\partial_0$ commutes with the free particle Hamiltonian $(-\hbar^2/2m){}^{(4)}\nabla^2$; the functions of the form (35) are the eigenfunctions of C,

$$C\Psi = e\Psi, \tag{37}$$

and the eigenvalue of C is the electric charge of the particle [cf. Eq. (8)].

As pointed out in Sec. II, the effect of a gauge transformation (9) is equivalent to the coordinate change (10); therefore, from Eq. (35) we see that a gauge transformation produces the change,

$$\psi'(x^i) = \psi(x^i)\exp\left(\frac{ie}{\hbar c}\xi\right), \tag{38}$$

as is well known. Furthermore, a Killing vector field $K^i\partial_i$ of the metric $\gamma_{ij}\,dx^i\,dx^j$ that leaves the magnetic field invariant, locally yields a Killing vector field $K^\alpha\partial_\alpha$ of the metric (5), which means that the operator,

$$\mathcal{K}\equiv -i\hbar K^\alpha\partial_\alpha, \tag{39}$$

commutes with the operator $(-\hbar^2/2m)^{(4)}\nabla^2$. In the case where we have a set of Killing vector fields $K_{(a)}=K^i_{(a)}\partial_i$ of $\gamma_{ij}\,dx^i\,dx^j$ that leave the magnetic field invariant satisfying Eq. (15), from Eqs. (16) and (39) it follows that the corresponding operators $\mathcal{K}_{(a)}$ obey the commutation relations,

$$[\mathcal{K}_{(a)},\mathcal{K}_{(b)}]=-i\hbar(c^d_{ab}\mathcal{K}_{(d)}+f_{ab}C). \tag{40}$$

For example, making use of Eqs. (3), (18), (33), and (34), one finds that the Schrödinger equation for a free particle in the four-dimensional geometry (19) is given by

$$\frac{1}{r^2}\partial_r(r^2\,\partial_r\Psi)+\frac{1}{r^2}\left[\frac{1}{\sin\theta}\partial_\theta(\sin\theta\,\partial_\theta\Psi)+\frac{1}{\sin^2\theta}(\partial^2_\chi\Psi-2\cos\theta\partial_\chi\,\partial_\varphi\Psi+\partial^2_\varphi\Psi)\right]$$
$$+\left(\frac{(c/g)^2}{g_{00}}-\frac{1}{r^2}\right)\partial^2_\chi\Psi=-\frac{2mE'}{\hbar^2}\Psi. \tag{41}$$

This equation admits separable solutions of the form

$$\Psi(\chi,r,\theta,\varphi)=R(r)e^{is\chi}{}_sY_{jm}(\theta,\varphi), \tag{42}$$

where the ${}_sY_{jm}$ are spin-weighted spherical harmonics[6] [which can be related to the Wigner D functions[7,8] and to the Jacobi polynomials (see, e.g., Refs. 9 and 10)] and $R(r)$ is a spherical Bessel function (cf. Refs. 11 and 12).

A comparison of Eqs. (35) and (42), taking into account that $\chi=-(c/g)x^0$, shows that $\psi(r,\theta,\varphi)=R(r){}_sY_{jm}(\theta,\varphi)$ is a separable solution of the Schrödinger equation (30) in the field of a magnetic monopole, with $s=-(eg)/(\hbar c)$. The operators $\mathcal{K}_{(a)}=-i\hbar K^\alpha_{(a)}\partial_\alpha$ $(a=1,2,3)$ [Eq. (39)], where the $K^\alpha_{(a)}\partial_\alpha$ are the Killing vector fields of (19) given by Eqs. (21), restricted to the subspace (37), reduce to the ‘‘generalized angular momentum’’ operators given, e.g., in Refs. 10 and 12.

ACKNOWLEDGMENT

The author would like to thank the referee for helpful comments.

1. V. Perlick, Class. Quantum Grav. **7**, 1849 (1990).
2. G. F. Torres del Castillo, Rev. Mex. Fis. **44**, 540 (1998).
3. H. V. Mclntosh and A. Cisneros, J. Math. Phys. **11**, 896 (1970).
4. T. Iwai and N. Katayama, J. Geom. Phys. **12**, 55 (1993); J. Phys. A **27**, 3179 (1994).
5. M. Som and A. Raychaudhuri, Proc. R. Soc. London, Ser. A **304**, 81 (1968).
6. E. T. Newman and R. Penrose, J. Math. Phys. **7**, 863 (1966).
7. J. N. Goldberg, A. J. Macfarlane, E. T. Newman, F. Rohrlich, and E. C. G. Sudarshan, J. Math. Phys. **8**, 2155 (1967).
8. G. F. Torres del Castillo and A. Hernández-Guevara, Rev. Mex. Fis. **41**, 139 (1995).
9. G. F. Torres del Castillo and L. C. Cortés-Cuautli, J. Math. Phys. **38**, 2996 (1997).
10. L. C. Cortés-Cuautli, Rev. Mex. Fis. **43**, 527 (1997).
11. Ig. Tamm, Z. Phys. **71**, 141 (1931).
12. T. T. Wu and C. N. Yang, Nucl. Phys. B **107**, 365 (1976).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Uniqueness theorems for classical four-vector fields in Euclidean and Minkowski spaces

Dale A. Woodside[a)]
Department of Physics, Macquarie University–Sydney, New South Wales 2109, Australia

(Received 8 December 1998; accepted for publication 14 July 1999)

Euclidean and Minkowski four-space uniqueness theorems are derived which yield a new perspective of classical four-vector fields. The Euclidean four-space uniqueness theorem is based on a Euclidean four-vector identity which is analogous to an identity used in Helmholtz's theorem on the uniqueness of three-vector fields. A Minkowski space identity and uniqueness theorem can be formulated from first principles and the space components of this identity turn out to reduce to the three-vector Helmholtz's identity in a static Newtonian limit. A further result is a uniqueness theorem for scalar fields based on an identity which is proved to be a static Newtonian limit of the zeroth or scalar component of the Minkowski space extension of the Helmholtz identity. Last, the three-vector Helmholtz identity and uniqueness theorem and their four-space extensions to Minkowski space are generalized to mass damped fields.
[S0022-2488(99)00810-5]

I. INTRODUCTION

In Sec. II a review of Helmholtz's theorem on the uniqueness of three-vector fields is first presented. Helmholtz's theorem is concerned with irrotational fields, which have zero curl everywhere in space, and solenoidal fields, which have zero divergence everywhere in space. Now, the divergence and curl of a vector field over all of a Euclidean three-space uniquely determine this vector field. Based on these properties a theorem, that will be called the Helmholtz theorem, states that the most general continuous three-vector field defined everywhere in a Euclidean three-space, that along with its first derivatives vanishes sufficiently rapidly at infinity, may be uniquely represented as a sum of an irrotational and a solenoidal part, up to a possible additive vector constant. The theorem can be extended to finite volumes as well. In proving Helmholtz's theorem, a vector identity is used that is commonly referred to as the Helmholtz identity, which in turn is derived from a delta function property of the Laplacian operator.

A comparison of the three-space Laplacian operator with the Minkowski space d'Alembertian operator then suggests that a three-divergence and a three-curl naturally generalize into a four-divergence and a four-curl when a fourth dimension is added. A review of the Sommerfeld four-space integral solution of the four-vector potential d'Alembertian wave equation, translated into modern notation, is then presented.

In Secs. III and IV a number of theorems are put forward. Of these theorems, Theorems I, II, IV, V, and VII–XII appear to be new results for three-vector and four-vector fields in a flat space. The new four-vector results of this paper follow from a different starting point than another series of attempts to extend Helmholtz's theorems to Minkowski four-space[1] in that the present paper adopts a definition of the four-curl as given by the Maxwell field tensor $F^{\mu\nu}$, Eq. (14b), following the definition in Møller,[2] while these other attempts choose the dual of the Maxwell field tensor as their four-curl. This later definition appears to be oriented toward the analysis of the uniqueness of electromagnetic fields in the presence of hypothetical magnetic monopoles. The advantage of the approach of the present paper, on the other hand, is that it leads to a more natural generalization

[a)]Electronic mail: dalew@physics.mq.edu.au

of Helmholtz's three-space theorems to four-space. In particular, it yields a four-space Helmholtz identity whose space components reduce to the three-space Helmholtz identity in a static Newtonian limit.

In Sec. III a comparison of Sommerfeld's four-space integral solution with a Green's function approach yields a four-space Euclidean Laplacian delta function identity. On observing that the original three-space Helmholtz identity is commonly derived from an analogous three-space Laplacian operator delta function identity, a parallel approach is taken and the four-vector Euclidean Laplacian delta function identity is used to derive a Euclidean four-space generalization of the Helmholtz identity. As with Sommerfeld's Euclidean four-space integral, the Euclidean four-space delta function identity can be analytically continued to Minkowski space by choosing an appropriate integration contour which takes timelike causality into account. Rather than working with this analytic continuation technique to derive a Minkowski space result, it turns out to be more convenient to derive the Minkowski four-space extension of the Helmholtz identity from first principles using a Minkowski space retarded Green's function. These Euclidean and Minkowski four-space extensions of the three-space Helmholtz identity are explicitly stated in Theorems I and II, respectively, and form the basis for several uniqueness theorems for four-vector fields.

Next, it is shown that this Minkowski four-space identity reduces to the three-space Helmholtz identity in a static Newtonian limit. The zeroth or scalar component of this Minkowski four-space extension of the Helmholtz identity is then shown to reduce to a three-vector integral identity for scalar fields in a static Newtonian limit. This three-vector integral identity for scalar fields, explicitly stated in Theorem III, is also proved by direct application of the three-space Laplacian operator delta function identity.

Uniqueness theorems in Euclidean and Minkowski four-spaces, Theorems IV and V, are then proved using their respective four-vector Helmholtz identities. It is found that the specification of the *four-curl* and *four-divergence* of the four-vector field throughout the four-volume V_4, as well as the *four-tangential* and *four-normal projections* of the four-vector field everywhere on the bounding three-surface Σ, are sufficient to obtain a unique four-vector field. A further result based on Theorems IV and V, and stated later in Theorem X in Sec. IV, is that a four-vector field is uniquely specified by the sum of a *four-irrotational* and a *four-solenoidal* part. This latter theorem corresponds to a four-space generalization of Helmholtz's uniqueness theorem.

Also in Sec. III, a uniqueness theorem for scalar fields, Theorem VI, is proved using the scalar field integral identity of Theorem III. In this three-vector field case it is found that the gradient of the scalar field throughout the volume V, as well as the magnitude of the scalar field on the bounding surface S, are sufficient to obtain a unique scalar field in a Euclidean three-space.

Finally, in Sec. IV there is a discussion on whether or not the Helmholtz identity and its relativistic extensions to Euclidean and Minkowski spaces can be generalized to fields with mass. First, by adding an exponential damping factor for the mass to the Euclidean three-space Laplacian delta function identity, an exponentially damped three-vector Helmholtz identity is stated in Theorem VII. A uniqueness theorem, Theorem VIII, is then proved using this three-vector identity. Next, in the Euclidean four-vector case it is shown that adding an exponential damping factor for the mass leads to a Euclidean four-space Laplacian delta function identity with an additional cross term which makes the development of an analogous four-vector identity problematic. On the other hand, in the Minkowski four-space case, the existence of a massive scalar Green's function over timelike separations allows one to obtain an exponentially damped Minkowski space four-vector identity which is explicitly stated in Theorem IX. But the four-space extension of Helmholtz's uniqueness theorem does not appear to carry over to the exponentially damped case. However, an exponentially damped version of uniqueness Theorem V is stated in Theorem XII, which relies on a theorem on the vanishing of a four-vector field, i.e., Theorem XI.

Although it is tempting to interpret the exponentially damped results as generalizations of the earlier results to massive four-vector fields, no limitation was imposed on their derivations other than that the four-vector fields are assumed to be sufficiently smooth. The exponentially damped results do, however, appear to be oriented toward application to massive vector fields or alternately to fields undergoing spatial diffusion in three or four dimensions, respectively.

II. HISTORICAL AND MATHEMATICAL BACKGROUND

A. Historical survey of uniqueness theorems in three dimensions

A natural starting point in the development of a uniqueness theorem for classical four-vector fields is an examination of the various statements of uniqueness theorems for ordinary three-vector fields in a flat Euclidean three-space. Historically, the first statement of a uniqueness theorem for three-vector fields is found in Stokes' article on diffraction in 1849.[3] The next statement is found in Helmholtz's article on the hydrodynamics of vortex motion in 1858.[4] Modern texts often attach Helmholtz's name to a uniqueness theorem for three-vector fields, but Sommerfeld[5] makes a point of acknowledging Stokes' contribution as well. Also, modern statements of uniqueness theorems for three-vector fields adopt vector notation, which is absent from their 1850's counterparts, and in addition differ in their emphasis and presentation.[6–9] A short review of these modern uniqueness statements is therefore warranted.

First, it is important to note that the defining properties of irrotational and solenoidal fields are the essential underpinnings of the uniqueness theorems for three-vector fields. Namely, an irrotational field has zero curl everywhere in space, i.e.,

$$0=\nabla\times\mathbf{A}, \tag{1a}$$

while a solenoidal field has zero divergence everywhere in space, i.e.,

$$0=\nabla\cdot\mathbf{A}. \tag{1b}$$

Next, a preliminary uniqueness theorem can be stated as follows:

Theorem U: *The divergence and curl of a three-vector field over a volume V in a Euclidean three-space, along with its normal components on a closed surface S bounding the volume V, uniquely determines the three-vector field over the volume V and on the surface S.*

In other words, one must specify

$$\nabla\times\mathbf{F}=\mathbf{j}(x,y,z), \tag{2a}$$

$$\nabla\cdot\mathbf{F}=\rho(x,y,z), \tag{2b}$$

over the volume V, and the normal component $\mathbf{F}_n(x,y,z)$ on the surface S, where for example in an electromagnetic context, $\mathbf{j}$ is a ''circulation current density'' and ρ is a ''source charge density.'' A proof of this theorem is given in Arfken.[8]

A uniqueness theorem that can be attributed to Helmholtz can now be stated as follows:

Theorem H1 (Helmholtz's theorem on the uniqueness of three-vector fields over all of a Euclidean three-space): *A general continuous three-vector field defined everywhere in a Euclidean three-space, that along with its first derivatives vanishes sufficiently rapidly at infinity, may be uniquely represented as a sum of an irrotational and a solenoidal part, up to a possible additive constant vector*.

A proof of this theorem is given in Sommerfeld.[5]

A modern alternate form of this theorem can be stated by restricting the domain of definition of the three-vector field to a finite volume as follows:

Theorem H2 (Alternate form of Theorem H1 for a finite volume in a Euclidean three-space): *A general continuous three-vector field that is defined everywhere in a finite volume V of a Euclidean three-space and whose tangential and normal components on the bounding closed surface S are given may be uniquely represented as a sum of an irrotational and a solenoidal part*.

In order to prove either Theorem H1 or Theorem H2, it turns out that it is sufficient to prove only that the three-vector $\mathbf{F}$ can be written as

$$\mathbf{F}(x,y,z)=-\nabla\Phi(x,y,z)+\nabla\times\mathbf{A}(x,y,z), \tag{3}$$

since by the three-vector identity

$$\nabla\times\nabla\Phi\equiv 0, \tag{4}$$

one has by (1a) that $-\nabla\Phi$ is irrotational, and since by the three-vector identity

$$\nabla\cdot\nabla\times\mathbf{A}\equiv 0, \tag{5}$$

one has by (1b) that $\nabla\times\mathbf{A}$ is solenoidal. In fact, Eq. (3) is used by King as a brief statement of his version of Helmholtz's theorem.[9] In another notable case, Eq. (3) is referred to in the very extensive treatment by Plonsey and Collin as ''the mathematical statement of the second part of Helmholtz's theorem.''[7]

The proof of (3) is based on the assumption that there exists a solution for the three-vector potential **A** of the vector Poisson equation in Cartesian coordinates, which in the electromagnetic case reads

$$\nabla^2\mathbf{A}(x,y,z)=-\mu_0\mathbf{J}(x,y,z), \tag{6}$$

where $\mu_0=1/\epsilon_0 c^2$ is the free space permeability. Now, in Cartesian coordinates, the three components of **A** are each separately a solution of a scalar Poisson equation. The scalar Poisson equation can then be solved in terms of a two-point scalar Green's function $G(\mathbf{r},\mathbf{r}')$ which connects its unit delta function source located at the source point $\mathbf{r}'=(x',y',z')$ to a measurement at the field point $\mathbf{r}=(x,y,z)$, i.e.,

$$\nabla^2 G(\mathbf{r},\mathbf{r}')=-\delta^3(\mathbf{r}-\mathbf{r}')=-\delta(x-x')(y-y')(z-z'). \tag{7}$$

The well-known identity over the Euclidean three-space $\mathbb{R}^3$, namely

$$\nabla^2\frac{1}{4\pi r}\equiv-\delta^3(\mathbf{r}-\mathbf{r}')\quad\forall\mathbf{r},\mathbf{r}'\in\mathbb{R}^3, \tag{8}$$

where $r\equiv|\mathbf{r}-\mathbf{r}'|$, yields by comparison with (7) for the case of an infinite spatial domain, the Green's function of the Laplacian operator as $G(\mathbf{r},\mathbf{r}')=1/4\pi r$. The inhomogeneous solution of (6) for the vector potential then follows from the integral

$$\mathbf{A}(\mathbf{r})=\int_{V'}\mu_0\mathbf{J}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')d^3\mathbf{r}'=\int_{V'}\frac{\mu_0\mathbf{J}(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|}d^3\mathbf{r}'. \tag{9}$$

Although (9) is an inhomogeneous solution to the vector Poisson equation (6) in an infinite spatial domain, it is the delta function property of the vector identity (8) which is of importance in the proof of Theorems H1 or H2, namely

$$\mathbf{F}(x,y,z)=\int_{V'}\mathbf{F}(x',y',z')\delta^3(\mathbf{r}-\mathbf{r}')d^3\mathbf{r}'=\int_{V'}\mathbf{F}(x',y',z')\nabla^2\left(\frac{-1}{4\pi r}\right)d^3\mathbf{r}'. \tag{10}$$

Then, since the Laplacian operator acts only on the field coordinates, it can be brought outside of the integration. At this point, one makes a decomposition of the Laplacian operator using identity (20), the significance of which is discussed in Sec. II B. The identity $\nabla(1/r)=-\nabla'(1/r)$ and a fair amount of vector analysis then yields the identity[6,7]

$$\mathbf{F}(x,y,z)=-\nabla\left[\int_{V'}\frac{\nabla'\cdot\mathbf{F}(x',y',z')}{4\pi r}dV'-\oint_{S'}\frac{\mathbf{F}(x',y',z')\cdot\mathbf{n}'}{4\pi r}dS'\right]$$
$$+\nabla\times\left[\int_{V'}\frac{\nabla'\times\mathbf{F}(x',y',z')}{4\pi r}dV'+\oint_{V'}\frac{\mathbf{F}(x',y',z')\times\mathbf{n}'}{4\pi r}dS'\right], \tag{11a}$$

where $\mathbf{n}'$ is the unit surface normal pointing out of the volume V' bounded by the closed surface S'. The proof of the volume integral terms of (11a) can be traced back to Stokes.[3] Proof of the surface integral terms of (11a) can be found in *Field Theory of Guided Waves* by Collin.[6] Equation (11a) is of the desired form (3) and can therefore be considered as completing the proof of Theorem H2, provided of course that the integrals are well defined. In order for the integrals to be well defined, one must make the additional assumption that the field $\mathbf{F}$ must vanish sufficiently rapidly at infinity, i.e., at least as fast as $1/r^2$ in order to avoid logarithmic divergences. To prove H1, one takes the surface S as going to infinity to include all of Euclidean three-space. The surface integral terms vanish as $r\rightarrow\infty$ under the same assumption that the field $\mathbf{F}$ falls off at least as fast as $1/r^2$. Equation (11a), but without the surface integral terms, is sometimes referred to as the ''Helmholtz identity'' (cf. Ref. 10). However, the more general result, the full equation (11a), will be referred to as the Helmholtz identity in this article.

It should be noted in passing that application of identity (11a) to a field $\mathbf{F}$ falling off only as fast as $1/r$ (e.g., a potential), would presumably require a cut-off procedure in the integrals. However, the vector derivatives which stand in front of the integrals act only on the field point coordinates so that, for a twice continuously differentiable vector field $\mathbf{F}$, one can move these derivatives inside of the integrals over the source point coordinates to give

$$\mathbf{F}(x,y,z)=-\left[\int_{V'}\nabla\left(\frac{\nabla'\cdot\mathbf{F}(x',y',z')}{4\pi r(\mathbf{r},\mathbf{r}')}\right)dV'-\oint_{S'}\nabla\left(\frac{\mathbf{F}(x',y',z')\cdot\mathbf{n}'}{4\pi r(\mathbf{r},\mathbf{r}')}\right)dS'\right]$$
$$+\left[\int_{V'}\nabla\times\left(\frac{\nabla'\times\mathbf{F}(x',y',z')}{4\pi r(\mathbf{r},\mathbf{r}')}\right)dV'+\oint_{S'}\nabla\times\left(\frac{\mathbf{F}(x',y',z')\times\mathbf{n}'}{4\pi r(\mathbf{r},\mathbf{r}')}\right)dS'\right], \quad (11b)$$

as an alternate form of the Helmholtz identity. Now, since the integrands of identity (11b) involve an extra vector derivative over the field point coordinates $\mathbf{r}$, the convergence properties of the integrands are improved and the integrals are now well defined for vector potentials $\mathbf{F}$ falling off only as fast as $1/r$.

Note, Eqs. (11a) and (11b) should be thought of as identities for representing a general (static) three-vector field rather than as a general solution to a partial differential equation. Indeed, Eqs. (11a) and (11b) follow from the vector identity (8) over an infinite spatial domain and its subsequent use in the delta function property (10). Consequently, Eqs. (11a) and (11b) are vector identities that apply to all of the Euclidean three-space, and so must hold for a finite subvolume of it as well. The Green's function for an inhomogeneous vector Poisson equation in an infinite spatial domain is only mentioned in passing, and certainly no use is made of the solutions of the associated source free homogeneous vector Laplace equation. This does not necessarily reduce the utility or applicability of (11a), for example. In fact, if one makes the replacements (2), for suitable volume source densities, and then makes the replacements

$$\sigma\equiv-\mathbf{F}\cdot\mathbf{n}, \quad (12a)$$

$$\mathbf{K}\equiv\mathbf{F}\times\mathbf{n}, \quad (12b)$$

with for example σ taken as a surface ''charge'' density and $\mathbf{K}$ taken as a surface ''current'' density, one can deduce from a general point of view basically all of the (static) macroscopic integral equations for an electromagnetic field in material media expressed in terms of its sources. This program, i.e., using (11a), is carried to its logical completion for the electromagnetic case in the thorough treatment of Plonsey and Collin.[7]

From the point of view of the present article, it turns out that identity (11a), as well as Theorems H1 or H2, are actually static, (i.e., nontime varying), cases of a more general identity and theorem in a pseudo-Euclidean $3+1$ space, (hereafter taken to be Minkowski space). Indeed, in the nontime varying case, identity (11a) follows as the space components of a four-vector identity in a static Newtonian limit, and in addition, a three-space identity arises from the fourth or scalar component of the four-vector identity. In the general time varying case in Minkowski space, the new identity and uniqueness theorem imply that the three-space notions of uniquely specifying a general vector field by its divergence and curl in a Euclidean volume V, along with its normal and tangential components on the closed bounding surface S, must be generalized to specifying its *four-divergence* and its *four-curl* in a four-volume V_4 of Minkowski space, along with its *four-normal* and *four-tangential* components on the closed bounding three-surface Σ.

B. Lagrangian formulation of four-vector fields in the flat space–time of special relativity

Before proceeding, a few preliminary definitions and assumptions are made. First, the nonzero components of the flat space Minkowski metric tensor $\eta_{\mu\nu}$ are taken as: $-\eta_{00}=\eta_{11}=\eta_{22}=\eta_{33}=1$. So, the ordinary four-vector derivatives are taken as: $\partial_\mu=((1/c)\partial/\partial t,\nabla)$ and $\partial^\mu=(-(1/c)\partial/\partial t,\nabla)$. Similarly, the position four-vector $x^\nu=(ct,x,y,z)$, and so $x_\nu=(-ct,x,y,z)$. [The adoption of this $(-+++)$ signature metric in this article will aid in the comparison of the results of this article with historical results expressed using complex Minkowski space notation, here taken as $x_\nu=(ict,x,y,z)$.]

Next, it is assumed that the most general form of Lagrangian density for a four-vector field, which is no more than quadratic in its variables and their derivatives, is given by the so-called Stueckelberg Lagrangian density,[11,12] (in SI units where ϵ_0 is the free space permittivity and c is the speed of light),

$$\mathcal{L}=-\frac{\epsilon_0 c^2}{4}F_{\mu\nu}F^{\mu\nu}+j_\mu A^\mu-\frac{\lambda\epsilon_0 c^2}{2}(\partial_\mu A^\mu)^2-\frac{\epsilon_0 c^2\mu^2}{2}(A_\mu A^\mu), \tag{13}$$

where $j^\mu=(\rho c,\mathbf{j})$ is the usual four-vector current, where the positive real constant λ is a Lagrange multiplier for the Lorentz constraint term, and where $\mu=2\pi/\lambda_C=2\pi mc/h$ is the Compton wave number for photons of mass m. A choice of $\lambda=0$ and $\mu=0$ yields what many physicists believe to be the electromagnetic theory, with its massless photons, i.e., when an appropriate constraint is externally imposed. However, the choice of $\lambda=0$ has the distinct disadvantage of implying a vanishing momentum canonically conjugate to the zeroth component of the four-vector potential $A^\nu=(\phi/c,\mathbf{A})$. The incorporation of the Lorentz constraint term, with its $\partial\phi/\partial t$ functionality, eliminates this deficiency, and yields an added bonus in terms of the ease of renormalization of the theory. A particularly simple choice of $\lambda=1$, (and $\mu=0$), then yields a Lagrangian density which is equivalent (i.e., differs by no more than a four-divergence), to the so-called Fermi Lagrangian density.[13] The Fermi Lagrangian density is the most straightforward take off point for field quantization in terms of harmonic oscillators which correspond to massless photons (cf. Ref. 13). The Stueckelberg Lagrangian density (13) also has the advantage of explicitly including the four-divergence and four-curl of A^μ, which are in turn sufficient for the unique specification of a four-vector field as is shown in Secs. II F and III B. Therefore, the point of view is taken that the choice $\lambda=1$ in (13) is the most natural choice for the development of a uniqueness theorem for four-vector fields.

The Maxwell field tensor $F_{\mu\nu}$ in terms of A_ν is taken as following from a Bianchi identity, namely

$$\partial_\lambda F_{\mu\nu}+\partial_\mu F_{\nu\lambda}+\partial_\nu F_{\lambda\mu}=0 \tag{14a}$$

$$\Rightarrow F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu. \tag{14b}$$

The covariant form of the Euler–Lagrange equations of motion

$$\partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu A_\nu)}\right)-\frac{\partial\mathcal{L}}{\partial A_\nu}=0, \tag{15}$$

then yields for the Stueckelberg Lagrangian density (13), the covariant equation of motion

$$(\partial_\mu \partial^\mu - \mu^2) A^\nu - (1-\lambda) \partial^\nu (\partial_\sigma A^\sigma) = - j^\nu / \epsilon_0 c^2. \tag{16}$$

In the so-called Feynman gauge, one takes $\lambda = 1$ so that (16) reduces to the following:

$$\Box A^\nu - \mu^2 A^\nu = - j^\nu / \epsilon_0 c^2, \tag{17}$$

where $\Box \equiv \partial_\mu \partial^\mu$ is the d'Alembertian operator.

One can now rewrite the d'Alembertian operator acting on A^ν by adding and subtracting a term, i.e.,

$$\Box A^\nu = \partial_\mu \partial^\mu A^\nu - \partial_\mu \partial^\nu A^\mu + \partial_\mu \partial^\nu A^\mu. \tag{18}$$

The first two terms combine naturally using the field tensor $F^{\mu\nu}$, while the ordinary four-vector derivatives in the last term commute in flat space–time to yield the identity

$$\Box A^\nu \equiv \partial_\mu \partial^\mu A^\nu = \partial_\mu F^{\mu\nu} + \partial^\nu (\partial_\mu A^\mu). \tag{19}$$

Equation (19) is a special relativistic generalization to $3+1$ space–time of the well-known three-space identity

$$\nabla^2 \mathbf{A} \equiv - \nabla \times (\nabla \times \mathbf{A}) + \nabla (\nabla \cdot \mathbf{A}). \tag{20}$$

Specifically, the curl of **A** is generalized into the four-curl of A^ν (i.e., $F^{\mu\nu}$), the divergence of **A** is generalized into the four-divergence of A^μ (i.e., $\partial_\mu A^\mu$), and the Laplacian is generalized into the d'Alembertian.

This special relativistically invariant decomposition, using the identity (19), is the defining property which leads to a uniqueness theorem for four-vector fields in a flat space–time. Indeed, in the same way that the delta function property (10), with its Laplacian operator, leads to the Helmholtz identity (11a), so too does a more general Minkowski space delta function property, with a d'Alembertian operator, lead to a new special relativistically invariant identity.

C. Integration of the four-vector potential wave equation

Assuming that the four-vector potential wave equation (17), as following from the Stueckelberg Lagrangian (13) with $\lambda = 1$, is the most natural starting point for a four-dimensional analysis, one now proceeds in a manner analogous to the three-dimensional case outlined in Sec. II A. Initially, however, μ is taken as zero yielding the massless four-vector potential wave equation as

$$\Box A^\mu(x^\nu) \equiv \left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) A^\mu(x^\nu) = - \frac{j^\mu(x^\nu)}{\epsilon_0 c^2}. \tag{21}$$

In Sec. IV B the mass term is restored and an attempt at a more general result is made. Now, the first step in the analysis leading to a uniqueness theorem for classical four-vector fields is to integrate the four-vector potential wave equation (21).

The integration of the massless four-vector potential wave equation (21) was first done in a special relativistically invariant way by Arnold Sommerfeld in an important article in 1910.[14] The method was later translated into English and appeared in Volume III: Electrodynamics of his well-known *Lectures on Theoretical Physics*.[15] Additional revealing descriptions of the method were given by Møller[2] and Stratton[16] in their classic texts.

To formulate the integral, Sommerfeld replaces Newton's $1/r$ potential in three dimensions by an analogous four-dimensional scalar potential, i.e.,

$$U = \frac{1}{R^2}, \tag{22}$$

where R is the distance in a Euclidean four-space between the source point x'^{ν} and field point x^{ν},

$$R \equiv |x^{\nu}-x'^{\nu}| = \sqrt{(x_{\nu}-x'_{\nu})(x^{\nu}-x'^{\nu})}, \tag{23}$$

and where the Euclidean metric tensor is just $\eta_{\mu\nu}=\mathbf{I}$, the identity matrix. As pointed out by Møller,[2] Sommerfeld derives the integral in Euclidean space, and then in order to get physically reasonable results, he analytically continues the four-current density by deforming the time integration from the real axis onto the imaginary axis in such a way as to take timelike causality into account as is appropriate for the *complex* Minkowski space used in his treatment. On the other hand, modern notation using covariant and contravariant indices is used in this partial review of Sommerfeld's method. Therefore, after the analytic continuation from Euclidean four-space to complex Minkowski space is done, the results are adjusted to take into account the replacement of the Euclidean metric tensor by the Minkowski metric tensor. The contour integration itself is described in Refs. 15 and 2.

Just as the three-space Laplacian of Newton's $1/r$ potential is zero everywhere except at the source point $\mathbf{r}'$, [see (8)], the *Euclidean* version of the d'Alembertian, (i.e., the four-space Laplacian Δ^{ν}_{ν}), of the scalar potential U is likewise zero everywhere except at the four-space source point x'^{ν}. To demonstrate this one first calculates

$$\frac{\partial}{\partial x^{\nu}}\left(\frac{1}{R^2}\right) = -\frac{2}{R^3}\frac{\partial R}{\partial x^{\nu}} = -\frac{2(x_{\nu}-x'_{\nu})}{R^4}, \tag{24}$$

where use is made of the four-vector derivative of R as follows:

$$\frac{\partial R}{\partial x^{\nu}} = \frac{1}{2R}\frac{\partial((x_{\nu}-x'_{\nu})(x^{\nu}-x'^{\nu}))}{\partial x^{\nu}} = \frac{(x_{\nu}-x'_{\nu})}{R}. \tag{25}$$

It is obvious from the calculation (24) that one also has the useful identity:

$$\frac{\partial}{\partial x'^{\nu}}\left(\frac{1}{R^2}\right) = -\frac{\partial}{\partial x^{\nu}}\left(\frac{1}{R^2}\right). \tag{26}$$

The Euclidean four-space Laplacian of U now follows from a four-vector derivative contraction of (24) as

$$\Delta^{\nu}_{\nu}\left(\frac{1}{R^2}\right) \equiv \frac{\partial}{\partial x_{\nu}}\frac{\partial}{\partial x^{\nu}}\left(\frac{1}{R^2}\right) = -\frac{\partial}{\partial x_{\nu}}\left(\frac{2(x_{\nu}-x'_{\nu})}{R^4}\right) = \frac{-2\cdot 4}{R^4} + \frac{2\cdot 4(x_{\nu}-x'_{\nu})}{R^5}\frac{\partial R}{\partial x_{\nu}}. \tag{27}$$

Substitution of a contravariant version of (25) into (27) yields finally

$$\Delta^{\nu}_{\nu}\left(\frac{1}{R^2}\right) = 0 \quad \forall x^{\nu} \neq x'^{\nu}, \tag{28}$$

that is except at the singular point at $R=0$.

Next, using a Green's theorem integral approach, Sommerfeld obtains the Euclidean four-space integral: (cf. Refs. 15, 2, and 16 for details)

$$\phi(x^{\nu}) = -\int_{V'_4}\frac{\partial'^{\mu}\partial'_{\mu}\phi(x'^{\nu})}{4\pi^2 R^2(x^{\nu},x'^{\nu})}\,d^4x', \tag{29}$$

where the surface area $2\pi^2R^2$ of a three-sphere of radius R is incorporated in the denominator of (29). As pointed out by Møller[2] Eq. (29) holds for any regular (i.e., analytic), function ϕ, so if $\phi = A^0 c$ is taken as satisfying the zeroth component of (21), i.e., of $\partial'^{\mu}\partial'_{\mu}\phi(x') = -\rho(x')/\epsilon_0$, one obtains

$$\phi(x^{\nu}) = \int_{V'_4} \frac{\rho(x'^{\nu})}{4\pi^2\epsilon_0 R^2(x^{\nu},x'^{\nu})} d^4x'. \tag{30a}$$

Equation (30a) allows one to calculate ϕ at every point in V'_4 when the source charge density ρ is known over all of V'_4. However, in real life physical problems ρ is given only for purely imaginary x'^0 values corresponding to $\mathrm{Im}(x^0) > \mathrm{Im}(x'^0)$ i.e., the zeroth component of (21) is a Minkowski space relation and the integral (30a) can no longer be limited to Euclidean space. This timelike causality assumption requires that a signal from the source point reaches the field point only after traveling the distance R at a the finite speed c. Therefore, to take into account the timelike causal data required for the charge density ρ, one can then analytically continue in $x'^0 = ct'$ in the integrand of (30a). The analytically continued version of (30a) is therefore

$$\phi(x^{\nu}) = \int_{V'} \left[\oint_{t'} \frac{\rho(x'^{\nu})}{4\pi^2\epsilon_0 R^2(x^{\nu},x'^{\nu})} c\,dt' \right] d^3x', \tag{30b}$$

where a suitable integration contour in the complex t' plane has to be chosen to satisfy timelike causality of the data. That is, such a contour integration is taken as being compatible with timelike causality. Performance of the integral (30b) under the specified timelike causal data yields, in terms of a ''retarded'' charge density, a Minkowski space result. In order to obtain results in modern metric notation, the complex Minkowski space result is then mapped to a Minkowski metric space by replacement of the Euclidean metric tensor by the Minkowski metric tensor (thus absorbing any factors of the imaginary unit i), which here amounts to retention of raised and lowered index notation. An analytic continuation in the complex t' plane of a time integral of the type in (30b) is detailed in Refs. 15 and 2.

Again, as pointed out by Møller,[2] since Eq. (29) holds for any regular function of ϕ it will hold for the A^1, A^2, and A^3 spatial Cartesian components of A^{μ} in the massless four-vector potential wave equation (21). The combined result for the four-vector potential A^{μ} at the space–time field point x^{ν}, is therefore given by a four-dimensional integration over the space–time source coordinates x'^{ν} of the timelike causal four-vector current density j^{μ}, i.e.,

$$A^{\mu}(x^{\nu}) = \int_{V'_4} \frac{j^{\mu}(x'^{\nu})}{4\pi^2\epsilon_0 c^2 R^2(x^{\nu},x'^{\nu})} d^4x' = \int_{V'_4} \frac{\mu_0 j^{\mu}(x'^{\nu})}{4\pi^2|x^{\nu}-x'^{\nu}|^2} d^4x', \tag{31}$$

where the integral (31) is to be interpreted as a Minkowski space integral in the same sense as in (30b). Equation (31) is the integral result first obtained by Sommerfeld,[14] retraced here in modern notation, as desired.

III. DERIVATION OF UNIQUENESS THEOREMS

A. Green's function approach and a delta function identity

Rather than pursuing further the inhomogeneous four-vector wave equation (21), it turns out to be more convenient for the purpose of this article to focus instead on the inhomogeneous *scalar* wave equation for the scalar potential ϕ,

$$\Box\phi(x^{\nu}) = \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} \right) \phi(x^{\nu}) = -\frac{\rho(x^{\nu})}{\epsilon_0}. \tag{32}$$

The wave equation (32) can then be solved in terms of a two-point scalar Green's function $G(x^\nu, x'^\nu)$ which connects its unit delta function source located at the space–time source point x'^ν to a measurement at the space–time field point x^ν in Minkowski space, i.e.,

$$\Box G(x^\nu, x'^\nu) = \left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) G(x^\nu, x'^\nu) = -\delta^{(4)}(x^\nu - x'^\nu), \tag{33}$$

where $\delta^{(4)}(x^\nu - x'^\nu) = \delta(x^0 - x'^0)\delta^3(\mathbf{r} - \mathbf{r}')$. If the Green's function $G(x^\nu, x'^\nu)$ is known *over all space–time*, the inhomogeneous solution of (32) then follows from the Minkowski space integral

$$\phi(x^\nu) = \int_{V_4'} \frac{\rho(x'^\nu)}{\epsilon_0} G(x^\nu, x'^\nu) d^4x'. \tag{34}$$

Interestingly, Eq. (34) is of the same general form as the Euclidean space integral (30a). Therefore, if one starts with (30b), i.e., the analytic continuation of (30a) to Minkowski space, it should be possible to extract a Green's function $G(x^\nu, x'^\nu)$ through simple comparison of (34) and (30b), which would integrate to the same result under mutually compatible causality conditions. As additional evidence notice that for $x^\nu \neq x'^\nu$, Eq. (33) reduces to the following Minkowski space result:

$$\Box G(x^\nu, x'^\nu) = 0 \quad \forall x^\nu \neq x'^\nu. \tag{35}$$

The Euclidean four-space result (28) has the same general form as Eq. (35). And, the result (28) holds in Minkowski space as well, (since, e.g., $\partial^\nu x_\nu = 4$ in either space). This implies (by the uniqueness of the solutions of the Cauchy problem for constant coefficient wave equations) that the Green's function $G(x^\nu, x'^\nu)$ can be taken as the scalar field $U = 1/R^2$ times a numerical constant. It is reasonable to assume therefore that the numerical constant can be obtained from a comparison of (30b) and (34) and is just $1/4\pi^2$. The desired Green's function $G(x^\nu, x'^\nu)$ is therefore given by the ansatz

$$G(x^\nu, x'^\nu) = \frac{1}{4\pi^2 R^2}, \tag{36}$$

where $R^2 = |x^\nu - x'^\nu|^2$. It should be emphasized that (36) is an appropriate Green's function for Minkowski space only in the context of an analytically continued time integral as in (30b), and therefore appropriate causality conditions must subsequently be applied to obtain physically reasonable results. Equation (36) would of course be suitable in a Euclidean four-space without any assumptions involving analytic continuation.

The result (36) differs from the familiar retarded Green's function as derived for example in Cushing[17] or Jackson,[18] which for the metric signature $(-+++)$ and sign of the source term in (32) would be

$$G_{\text{ret}}(\mathbf{r}, \mathbf{r}'; t, t') = \frac{1}{4\pi} \frac{\delta(|\mathbf{r} - \mathbf{r}'| - c(t - t'))}{|\mathbf{r} - \mathbf{r}'|} \quad \forall t > t'. \tag{37}$$

This is because the retarded Green's function (37) is derived via a spectral decomposition of the delta function, while also taking into account homogeneous boundary conditions on a closed spatial surface, as well as timelike causality with the initial conditions

$$G(\mathbf{r}, \mathbf{r}'; t, t') = 0, \quad \partial G(\mathbf{r}, \mathbf{r}'; t, t')/\partial t = 0 \quad \forall t > t', \tag{38}$$

and with the Green's function symmetry relation

$$G(\mathbf{r},\mathbf{r}';t,t')=G(\mathbf{r}',\mathbf{r};-t',-t). \tag{39}$$

On the other hand, in obtaining the result (36), no such spectral decomposition is made and the calculation of the time integral, with an appropriate contour taking causality into account, is performed only later in this article. In contradistinction to the Green's function (37), (36) might therefore be referred to as an ''acausal'' Minkowski space Green's function, or more simply as a Euclidean four-space Green's function.

Substitution of (36) into a Euclidean space version of (33) then yields a delta function identity over all of the Euclidean four-space $\mathbb{R}^4$ as follows:

$$\Delta_\alpha{}^\alpha\left(\frac{1}{4\pi^2R^2}\right)=-\delta^{(4)}(x^\nu-x'^\nu),\quad \forall x^\nu,x'^\nu\in\mathbb{R}^4. \tag{40}$$

Although (40) follows from ostensibly Euclidean space calculations it can be applied to a space–time integral when an analytic continuation of the time integral is performed under appropriate causality conditions in order to obtain a Minkowski space result.

Equation (34) with the Green's function (36) is, under suitable analytic continuation of the time integral, the same result as (30b). [Indeed, the numerical constant $1/4\pi^2$ in (36) was obtained by a short cut comparison between the two.] And since the A^1, A^2, A^3 spatial Cartesian components of (21) separately satisfy a scalar wave equation like (32), one can then write the inhomogeneous solution of (21) for the four-vector potential A^μ as

$$A^\mu(x^\nu)=\int_{V_4'}\frac{j^\mu(x'^\nu)}{\epsilon_0c^2}G(x^\nu,x'^\nu)d^4x'=\int_{V'}\left[\oint_{t'}\frac{j^\mu(x'^\nu)}{4\pi^2\epsilon_0c^2R^2(x^\nu,x'^\nu)}c\,dt'\right]d^3x', \tag{41}$$

which is the same result as (31). Compare (36), (40), and (41) with the three-dimensional case (8) and (9). It is shown in Sommerfeld[15,14] in terms of retarded potentials that an expression like (9), as well as a scalar potential integral, follows from (31) via a contour integration over the time coordinate.

Although the result (31) is a historically important result, it is the delta function property of identity (40) which is of importance in the proof of a vector identity in Minkowski space analogous to the Helmholtz identity, namely

$$A^\mu(x^\nu)=\int_{V_4'}A^\mu(x'^\nu)\delta^{(4)}(x^\nu-x'^\nu)d^4x'=\int_{V'}\left[\oint_{t'}A^\mu(x'^\nu)\Delta_\alpha{}^\alpha\left(\frac{-1}{4\pi^2R^2}\right)c\,dt'\right]d^3x'. \tag{42}$$

Equation (42) and a Euclidean four-space version of (42) are used by the author in the next section to derive four-vector identities analogous to the three-vector Helmholtz identity (11a).

B. Euclidean and Minkowski four-space analogs of the Helmholtz identity

Equation (42) is now used by the author to derive four-vector identities analogous to, but more general than, the three-vector Helmholtz identity (11a). To a certain extent, the derivation parallels the three-vector derivation of (11a) as detailed in Refs. 6 and 7.

Initially, a theorem containing a four-vector identity is stated and proved for Euclidean space. Next, a corollary to the Euclidean theorem is stated for four-vectors in Minkowski space based on the contour integral used in (42). Subsequently, a second theorem containing a four-vector identity is stated which is based on a direct relativistically invariant integration over Minkowski space involving a Minkowski space Green's function of the form (37).

It is assumed in what follows, unless otherwise stated, that the term ''sufficiently smooth'' refers to functions which are scalar fields or components of vector fields that are $C^2(\overline{V}_4)$, i.e., twice continuously differentiable functions on the closure of V_4 which in turn is comprised of the union of V_4 and its bounding three-surface Σ (i.e., $\overline{V}_4 = V_4 \cup \Sigma$). The boundary three-surface Σ is itself assumed to be sufficiently smooth.

Theorem I: *The following identity holds for sufficiently smooth four-vector fields $A^\mu(x^\sigma)$ in the Euclidean four-space $\mathbb{R}^4$:*

$$A^\mu(x^\sigma) = -\partial^\mu \left[\int_{V_4'} \frac{\partial_\nu' A^\nu(x'^\sigma)}{4\pi^2 R^2(x^\sigma, x'^\sigma)} d^4x' - \oint_{\Sigma'} \frac{A^\nu(x'^\sigma) n_\nu'}{4\pi^2 R^2(x^\sigma, x'^\sigma)} d\Sigma' \right]$$
$$- \partial_\alpha \left[\int_{V_4'} \frac{\partial'^\alpha A^\mu(x'^\sigma) - \partial'^\mu A^\alpha(x'^\sigma)}{4\pi^2 R^2(x^\sigma, x'^\sigma)} d^4x' + \oint_{\Sigma'} \frac{A^\alpha(x'^\sigma) n'^\mu - A^\mu(x'^\sigma) n'^\alpha}{4\pi^2 R^2(x^\sigma, x'^\sigma)} d\Sigma' \right], \tag{43}$$

where $R^2(x^\sigma, x'^\sigma) = |x^\sigma - x'^\sigma|^2$, and where n'^μ is the four-vector outward unit normal of the three-surface Σ' which encloses the four-volume V_4'.

Proof: The proof is based on a Euclidean space version of the four-vector delta function property (42), namely

$$A^\mu(x^\nu) = \int_{V_4'} A^\mu(x'^\nu)\, \delta^{(4)}(x^\nu - x'^\nu) d^4x' = \int_{V_4'} A^\mu(x'^\nu) \Delta_\alpha{}^\alpha \left(\frac{-1}{4\pi^2 R^2} \right) d^4x'. \tag{44}$$

The important thing to realize about (44) is that the Euclidean four-space Laplacian operator acts only on the four-space field coordinates x^ν and so for sufficiently smooth four-vector fields $A^\mu(x'^\nu)$ it can be brought outside of the integration over the four-space source coordinates x'^ν as follows:

$$A^\mu(x^\nu) = \int_{V_4'} A^\mu(x'^\nu) \Delta_\alpha{}^\alpha \left(\frac{-1}{4\pi^2 R^2} \right) d^4x' = -\Delta_\alpha{}^\alpha \int_{V_4'} \frac{A^\mu(x'^\nu)}{4\pi^2 R^2} d^4x'. \tag{45}$$

At this point, one makes a decomposition of the Euclidean Laplacian operator Δ_α^α, analogous to that discussed in (18), by adding and subtracting a term in (45) yielding

$$A^\mu(x) = -\partial_\alpha \partial^\alpha \int_{V_4'} \frac{A^\mu(x')}{4\pi^2 R^2} d^4x' + \partial_\alpha \partial^\mu \int_{V_4'} \frac{A^\alpha(x')}{4\pi^2 R^2} d^4x' - \partial_\nu \partial^\mu \int_{V_4'} \frac{A^\nu(x')}{4\pi^2 R^2} d^4x', \tag{46}$$

where the four-vector superscripts in the functional dependencies are again suppressed. Now, in Euclidean space, (and also in flat Minkowski space), one can commute the derivatives in the third term of (46). And since the four-vector derivatives act only on the field coordinates they can be passed inside the integrations over the source coordinates. Also identity (26) and its contravariant derivative counterpart

$$\partial^\mu \left(\frac{1}{R^2} \right) = -\partial'^\mu \left(\frac{1}{R^2} \right), \tag{47}$$

allow one to change the unprimed field point derivatives of $1/R^2$ in (46) into primed source point derivatives. The net result is that one can rewrite (46) as

$$A^{\mu}(x)=\partial_{\alpha}\left[\int_{V_4'}\frac{A^{\mu}(x')}{4\pi^2}\partial'^{\alpha}\left(\frac{1}{R^2}\right)d^4x'-\int_{V_4'}\frac{A^{\alpha}(x')}{4\pi^2}\partial'^{\mu}\left(\frac{1}{R^2}\right)d^4x'\right]$$
$$+\partial^{\mu}\left[\int_{V_4'}\frac{A^{\nu}(x')}{4\pi^2}\partial'_{\nu}\left(\frac{1}{R^2}\right)d^4x'\right]. \tag{48}$$

Setting the bracketed part of the first term of (48) equal to $A^{\alpha\mu}$, i.e.,

$$A^{\alpha\mu}\equiv\left[\int_{V_4'}\frac{A^{\mu}(x')}{4\pi^2}\partial'^{\alpha}\left(\frac{1}{R^2}\right)d^4x'-\int_{V_4'}\frac{A^{\alpha}(x')}{4\pi^2}\partial'^{\mu}\left(\frac{1}{R^2}\right)d^4x'\right], \tag{49}$$

and the bracketed part of the second term of (48) equal to A, i.e.,

$$A\equiv\left[\int_{V_4'}\frac{A^{\nu}(x')}{4\pi^2}\partial'_{\nu}\left(\frac{1}{R^2}\right)d^4x'\right], \tag{50}$$

allows one to write (48) as

$$A^{\mu}=\partial_{\alpha}A^{\alpha\mu}+\partial^{\mu}A, \tag{51}$$

which is reminiscent of the decomposition (19).

Working on (50) first using the four-vector derivative product rule while using a Euclidean four-space version of Gauss' divergence theorem on the appropriate term of the resulting equation then yields

$$A=\oint_{\Sigma'}\frac{A^{\nu}(x')n'_{\nu}}{4\pi^2R^2}d\Sigma'-\int_{V_4'}\frac{\partial'_{\nu}A^{\nu}(x')}{4\pi^2R^2}d^4x', \tag{52}$$

where n'_{ν} is the four-vector outward unit normal of the three-surface Σ' which encloses the four-volume V_4'. Equation (52) is in its final form.

Working on (49) next using the four-vector derivative product rule on both terms yields after some rearrangement

$$A^{\alpha\mu}=\int_{V_4'}\partial'^{\alpha}\left(\frac{A^{\mu}(x')}{4\pi^2R^2}\right)d^4x'-\int_{V_4'}\partial'^{\mu}\left(\frac{A^{\alpha}(x')}{4\pi^2R^2}\right)d^4x'-\int_{V_4'}\frac{\partial'^{\alpha}A^{\mu}(x')-\partial'^{\mu}A^{\alpha}(x')}{4\pi^2R^2}d^4x'. \tag{53}$$

The third term of (53) is in its final form. It remains then to show that the four-vector derivative ∂_{α} [see (51)] of the first two terms of (53) combine to yield the Euclidean four-space identity

$$\partial_{\alpha}\int_{V_4'}\left(\partial'^{\alpha}\left(\frac{A^{\mu}(x')}{4\pi^2R^2}\right)-\partial'^{\mu}\left(\frac{A^{\alpha}(x')}{4\pi^2R^2}\right)\right)d^4x'=-\partial_{\alpha}\oint_{\Sigma'}\frac{A^{\alpha}(x')n'^{\mu}-A^{\mu}(x')n'^{\alpha}}{4\pi^2R^2}d\Sigma'. \tag{54}$$

To prove (54), one can define a four-vector a_{μ} with constant magnitude and constant but arbitrary direction in Euclidean four-space which is chosen once and then held fixed. Next consider the two-point function

$$K_{\mu}(x^{\sigma},x'^{\sigma})\equiv a_{\mu}\left(\partial_{\alpha}\left(\frac{A^{\alpha}(x'^{\sigma})}{4\pi^2R^2(x^{\sigma},x'^{\sigma})}\right)\right). \tag{55}$$

Using a Euclidean four-space version of Gauss' divergence theorem on (55), namely

$$\int_{V_4'} (\partial'^{\mu} K_{\mu}) d^4x' = \oint_{\Sigma'} K_{\mu} n'^{\mu} d\Sigma', \tag{56}$$

where in (56) and in what follows n'^{μ} is the four-vector outward unit normal of the three-surface Σ' which encloses the four-volume V_4', one obtains

$$a_{\mu} \int_{V_4'} \left(\partial'^{\mu} \left(\partial_{\alpha} \left(\frac{A^{\alpha}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) \right) \right) d^4x' = a_{\mu} \oint_{\Sigma'} \left(\partial_{\alpha} \left(\frac{A^{\alpha}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) \right) n'^{\mu} d\Sigma', \tag{57}$$

where the four-vector a_{μ} factors out of the integral since it is a constant. Now, since $|a_{\mu}| \neq 0$ and a_{μ} has an arbitrary fixed direction, then its four-contractions in (57) cannot everywhere vanish and so (57) reduces to the identity

$$\partial_{\alpha} \int_{V_4'} \left(\partial'^{\mu} \left(\frac{A^{\alpha}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) \right) d^4x' = \partial_{\alpha} \oint_{\Sigma'} \left(\frac{A^{\alpha}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) n'^{\mu} d\Sigma', \tag{58}$$

where the field point derivatives ∂_{α} have been moved out of the source point integrations and where the field and source point derivatives on the left-hand side (lhs) of (58) have been commuted. Equation (58) shows that the second term on the lhs of (54) is equal to the first term on the right-hand side (rhs) of (54). The remaining two terms in (54) follow in a similar fashion. One can again define a four-vector a_{μ} with constant magnitude and constant but arbitrary direction in the Euclidean four-space which is chosen once and then held fixed. Next consider the two-point function

$$I_{\alpha}(x^{\sigma},x'^{\sigma}) \equiv a_{\mu} \left(\partial_{\alpha} \left(\frac{A^{\mu}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) \right). \tag{59}$$

Using (59) in the Euclidean four-space Gauss' divergence theorem (56) one obtains

$$a_{\mu} \int_{V_4'} \left(\partial'^{\alpha} \left(\partial_{\alpha} \left(\frac{A^{\mu}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) \right) \right) d^4x' = a_{\mu} \oint_{\Sigma'} \left(\partial_{\alpha} \left(\frac{A^{\mu}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) \right) n'^{\alpha} d\Sigma', \tag{60}$$

where the four-vector a_{μ} factors out of the integral since it is a constant. Now, since $|a_{\mu}| \neq 0$ and a_{μ} has an arbitrary fixed direction, then its four-contractions in (60) cannot everywhere vanish and so (60) reduces to the identity

$$\partial_{\alpha} \int_{V_4'} \left(\partial'^{\alpha} \left(\frac{A^{\mu}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) \right) d^4x' = \partial_{\alpha} \oint_{\Sigma'} \left(\frac{A^{\mu}(x'^{\sigma})}{4\pi^2 R^2(x^{\sigma},x'^{\sigma})} \right) n'^{\alpha} d\Sigma', \tag{61}$$

where the field point derivatives ∂_{α} have been moved out of the source point integrations and where the field and source point derivatives on the lhs of (60) have been commuted. Equation (61) shows that the first term on the lhs of (54) is equal to the second term on the rhs of (54), which when combined with (58) completes the proof of the Euclidean four-space identity (54).

Combining the results (51), (52), (53), and (54) completes the proof of the Euclidean four-space identity (43) and Theorem I. ♣

Note the similarity in structure of identity (43) and the Helmholtz identity (11a). In particular, the factors of $1/4\pi r$ in (11a) appropriate for spherically symmetric functions in $\mathbb{R}^3$ are replaced in (43) by factors of $1/4\pi^2 R^2$ which are appropriate for hyperspherically symmetric functions in $\mathbb{R}^4$. It appears then that identity (43) is a Euclidean four-space generalization of the (Euclidean three-space) Helmholtz identity (11a). Therefore, just as (11a) can be used to prove the Helmholtz uniqueness theorems H2 and H1 of (static) three-vector fields, identity (43) is used later in this article to prove a uniqueness theorem for four-vector fields in the Euclidean four-space $\mathbb{R}^4$.

A corollary is now stated which extends the Euclidean four-space Theorem I to Minkowski space.

Corollary I to Theorem I: When, in the complex t' plane, a time integration contour is taken which is compatible with timelike causality, and for sufficiently smooth four-vector fields $A^{\mu}(x^{\sigma})$, the following identity holds in the Minkowski space $\mathbb{R}^{3+1}$:

$$A^{\mu}(x^{\sigma}) = -\partial^{\mu}\left[\int_{V'}\left[\oint_{t'}\frac{\partial'_{\nu}A^{\nu}(x'^{\sigma})}{4\pi^2R^2}c\,dt'\right]d^3x' - \int_{V'}\left[\oint_{t'}\partial'_{\nu}\left(\frac{A^{\nu}(x'^{\sigma})}{4\pi^2R^2}\right)c\,dt'\right]d^3x'\right]$$
$$-\partial_{\alpha}\left[\int_{V'}\left[\oint_{t'}\frac{\partial'^{\alpha}A^{\mu}(x'^{\sigma})-\partial'^{\mu}A^{\alpha}(x'^{\sigma})}{4\pi^2R^2(x^{\sigma},x'^{\sigma})}c\,dt'\right]d^3x'\right.$$
$$\left.-\int_{V'}\left[\oint_{t'}\partial'^{\alpha}\left(\frac{A^{\mu}(x'^{\sigma})}{4\pi^2R^2}\right)c\,dt'\right]d^3x' + \int_{V'}\left[\oint_{t'}\partial'^{\mu}\left(\frac{A^{\alpha}(x'^{\sigma})}{4\pi^2R^2}\right)c\,dt'\right]d^3x'\right], \qquad (62)$$

where $R^2(x^{\sigma},x'^{\sigma}) = |x^{\sigma}-x'^{\sigma}|^2$ in $\mathbb{R}^4$, and where V' is a spatial volume in $\mathbb{R}^3$.

Proof: The proof is based on the four-vector delta function property (42) and in all important respects parallels the proof of the Euclidean four-space identity (43) of Theorem I. The details of the basic approach are not repeated here. However, in order to express the time integration in terms of a contour integral over the complex t' plane, it is convenient to forgo the transformations from four-volume to three-surface integrals as in (52) and (54). The result is identity (62). An integration contour compatible with timelike causality is specified in order to interpret the identity as a Minkowski space integral. These contours are discussed in general in Refs. 15 and 2. The contour integration is performed using the residue theorem and assumes that the integrand, aside from the poles in the $1/R^2$ factor, is analytic and without singularities over the domain enclosed by or on the chosen contour. The actual details of the choice of contour, as well as performing the integration itself are not necessary for the proof of Corollary I to Theorem I. It is only necessary for the sake of a physical interpretation that a contour compatible with timelike causality is chosen. ♣

The similarity in structure of identity (62) and the Helmholtz identity (11a) is now only apparent through the four-volume integrals in (62) containing the four-divergence and four-curl of A^{ν}. The normal and tangential three-surface integrals in (43) no longer appear in (62). That (62) can be interpreted as a Minkowski space integral instead of an integral in $\mathbb{R}^4$ is a result of the analytic continuation of the time integral. Identity (62) could be interpreted as a Minkowski space generalization of the (Euclidean three-space) Helmholtz identity (11a).

However, identity (62) is not of the same general form as identity (11a) because it lacks surface integral terms. Consequently, identity (62) is not a convenient starting point for proving a four-vector uniqueness theorem analogous to the three-space Helmholtz uniqueness theorem H2. Therefore, a second theorem is now developed in $\mathbb{R}^{3+1}$, through a direct integration in Minkowski space, which states an identity that is formally analogous to identity (43) in $\mathbb{R}^4$.

It is central to the development of this second theorem that identities for the four-vector derivatives of the Green's functions satisfying (33) in $\mathbb{R}^{3+1}$ analogous to (26) and (47) in $\mathbb{R}^4$ be obtained. A reciprocal four-space integral representation of the Green's functions satisfying (33) appears to be the easiest way to obtain the desired derivative properties. Therefore, a spectral decomposition of the delta function $\delta^{(4)}(x^{\nu}-x'^{\nu})$ appearing in (33) is made via the Fourier integral

$$\delta^{(4)}(x^{\nu}-x'^{\nu}) = \frac{1}{\sqrt{(2\pi)^4}}\int\left[\frac{e^{-ik_{\mu}x'^{\mu}}}{\sqrt{(2\pi)^4}}\right]e^{ik_{\mu}x^{\mu}}d^4k = \frac{1}{(2\pi)^4}\int e^{ik_{\mu}(x^{\mu}-x'^{\mu})}d^4k. \qquad (63)$$

In the usual way, Green's functions satisfying (33) then follow from

$$G(x^\nu,x'^\nu)=-\Box^{-1}\left(\frac{1}{(2\pi)^4}\int e^{ik_\mu(x^\mu-x'^\mu)}d^4k\right)=\frac{1}{(2\pi)^4}\int\frac{e^{ik_\mu(x^\mu-x'^\mu)}}{k^2}d^4k. \tag{64}$$

The desired covariant and contravariant derivative properties in $\mathbb{R}^{3+1}$ follow immediately from the result (64), for any appropriate contour, as

$$\partial_\mu G(x^\nu,x'^\nu)=-\partial'_\mu G(x^\nu,x'^\nu), \tag{65a}$$

$$\partial^\mu G(x^\nu,x'^\nu)=-\partial'^\mu G(x^\nu,x'^\nu), \tag{65b}$$

which are of the same general form as the properties (26) and (47) in $\mathbb{R}^4$, as might be expected.

An example of a Green's function which satisfies (64) is the retarded Green's function (37). However, as a manifestly covariant identity is desired for this theorem, the retarded Green's function, and optionally an advanced Green's function, must be able to be restated in relativistically covariant form. This is readily done using the delta function identity[18]

$$\begin{aligned}\delta((x^\nu-x'^\nu)^2)&=\delta(|\mathbf{r}-\mathbf{r}'|^2-c^2(t-t')^2)\\&=\delta((|\mathbf{r}-\mathbf{r}'|-c(t-t'))(|\mathbf{r}-\mathbf{r}'|+c(t-t')))\\&=\left[\frac{\delta((|\mathbf{r}-\mathbf{r}'|-c(t-t'))+\delta(|\mathbf{r}-\mathbf{r}'|+c(t-t')))}{2|\mathbf{r}-\mathbf{r}'|}\right].\end{aligned} \tag{66}$$

Using (66), the retarded Green's function (37), along with an advanced Green's function $G_{\rm adv}$, can be stated in relativistically covariant form as follows:[18]

$$G_{\rm ret}(x,x')=\frac{1}{2\pi}\theta(x^0-x'^0)\,\delta((x-x')^2), \tag{67a}$$

$$G_{\rm adv}(x,x')=\frac{1}{2\pi}\theta(x'^0-x^0)\,\delta((x-x')^2), \tag{67b}$$

where the theta function is defined as follows:

$$\theta(x^0-x'^0)=\begin{cases}1 & \text{for } x^0>x'^0\\ 0 & \text{for } x'^0>x^0.\end{cases} \tag{68}$$

A theorem can now be stated for four-vector fields in Minkowski space.

Theorem II: *Given a covariant scalar two-point Green's function $G(x^\nu,x'^\nu)$ which is a solution of (33) and which satisfies the derivative properties (65a), (65b), the following identity holds for sufficiently smooth four-vector fields $A^\mu(x^\sigma)$ in the Minkowski space $\mathbb{R}^{3+1}$:*

$$\begin{aligned}A^\mu(x)=&-\left[\int_{V'_4}\partial^\mu((\partial'_\nu A^\nu(x'))G(x,x'))d^4x'-\oint_{\Sigma'}\partial^\mu((A^\nu(x')n'_\nu)G(x,x'))d\Sigma'\right]\\&-\left[\int_{V'_4}\partial_\alpha((\partial'^\alpha A^\mu(x')-\partial'^\mu A^\alpha(x'))G(x,x'))d^4x'+\oint_{\Sigma'}\partial_\alpha((A^\alpha(x')n'^\mu\right.\\&\left.-A^\mu(x')n'^\alpha)G(x,x'))d\Sigma'\right],\end{aligned} \tag{69}$$

where n'^μ is the four-vector outward unit normal of the three-surface Σ' which encloses the four-volume V'_4, and where the three-surface Σ' is defined covariantly with respect to a general Lorentz transformation.

Proof: In contradistinction to identity (43) of Theorem I, the unprimed four-vector derivatives are now included in the integrands of identity (69) of Theorem II. Consequently, the convergence properties of these integrands are improved in comparison to those in (43) and so the integrals in (69) are well defined for four-vector fields A^μ falling off only as fast as $1/r$. For sufficiently smooth four-vector fields the unprimed derivatives can still be factored out of the integrals over the primed coordinates if required.

The proof is based on a Minkowski space version of the four-vector delta function property (42), but now written in terms of a covariant scalar two-point Green's function $G(x^\nu,x'^\nu)$ which is assumed to be a solution of (33), as follows:

$$A^\mu(x^\nu)=\int_{V_4'}A^\mu(x'^\nu)\delta^{(4)}(x^\nu-x'^\nu)d^4x'=\int_{V_4'}A^\mu(x'^\nu)\Box(-G(x^\nu,x'^\nu))d^4x'. \tag{70}$$

At this point, one makes a decomposition of the d'Alembertian operator as in (18), by adding and subtracting a term in (70), yielding

$$A^\mu(x)=-\int_{V_4'}\partial_\alpha\partial^\alpha(A^\mu(x')G(x,x'))d^4x'+\int_{V_4'}\partial_\alpha\partial^\mu(A^\alpha(x')G(x,x'))d^4x'$$
$$-\int_{V_4'}\partial_\nu\partial^\mu(A^\nu(x')G(x,x'))d^4x', \tag{71}$$

where the four-vector superscripts in the functional dependencies are again suppressed. The rest of the proof proceeds in a manner parallel to the proof of the Euclidean four-space identity (43) of Theorem I and most of the details will be condensed. In flat Minkowski space one can commute the derivatives in the third term of (71). Then certain unprimed derivatives can be commuted with primed coordinate dependent fields. Then using identities (65a) and (65b) one can change the unprimed field point derivatives of $G(x,x')$ in (71) into primed source point derivatives. The net effect of these changes allows one to rewrite (71) as

$$A^\mu(x)=\partial_\alpha\left[\int_{V_4'}A^\mu(x')\partial'^\alpha G(x,x')d^4x'-\int_{V_4'}A^\alpha(x')\partial'^\mu G(x,x')d^4x'\right]$$
$$+\partial^\mu\left[\int_{V_4'}A^\nu(x')\partial_\nu' G(x,x')d^4x'\right], \tag{72}$$

where for sufficiently smooth four-vector fields, certain unprimed derivatives have been factored out of the primed coordinate integrals of (72) for later reference. Setting the first bracketed term of (72) equal to $A^{\alpha\mu}$, i.e.,

$$A^{\alpha\mu}\equiv\left[\int_{V_4'}A^\mu(x')\partial'^\alpha G(x,x')d^4x'-\int_{V_4'}A^\alpha(x')\partial'^\mu G(x,x')d^4x'\right], \tag{73}$$

and the bracketed part of the second term of (72) equal to A, i.e.,

$$A\equiv\left[\int_{V_4'}A^\nu(x')\partial_\nu' G(x,x')d^4x'\right], \tag{74}$$

allows one to write (72) as

$$A^\mu=\partial_\alpha A^{\alpha\mu}+\partial^\mu A, \tag{75}$$

which is reminiscent of the decomposition (19). This decomposition of a four-vector field into the sum of a four-irrotational and a four-solenoidal part will be used in Sec. IV B in connection with Theorem X.

Next, working on the second bracketed term of (72) first (but without the unprimed derivatives factored out), using the four-vector derivative product rule, while at the same time using a Minkowski four-space version of Gauss' divergence theorem as appropriate for a second rank tensor integrand, (cf. p. 130 of Ref. 2), namely

$$\int_{V_4'} \partial_\nu' Z^{\mu\nu}\, d^4x' = \oint_{\Sigma'} Z^{\mu\nu}\, d\sigma_\nu' = \oint_{\Sigma'} Z^{\mu\nu} n_\nu'\, d\Sigma', \tag{76}$$

on the appropriate term of the resulting equation (i.e., after commuting ∂^μ and ∂_ν'), then yields

$$\int_{V_4'} \partial^\mu (A^\nu(x')\partial_\nu' G(x,x'))d^4x' = \oint_{\Sigma'} \partial^\mu (A^\nu(x') n_\nu' G(x,x'))d^4x' - \int_{V_4'} \partial^\mu ((\partial_\nu' A^\nu(x'))G(x,x'))d^4x', \tag{77}$$

where n_ν' is the four-vector outward unit normal of the three-surface Σ' which encloses the four-volume V_4'. Equation (77) is in its final form.

Working on the first bracketed term of (72) next (but again without the unprimed derivatives factored out), using the four-vector derivative product rule on both terms yields after some rearrangement

$$\begin{aligned}\int_{V_4'} \partial_\alpha (A^\mu(x')\partial'^\alpha G(x,x'))d^4x' - \int_{V_4'} \partial_\alpha (A^\alpha(x')\partial'^\mu G(x,x'))d^4x' \\ = \int_{V_4'} \partial_\alpha \partial'^\alpha (A^\mu(x') G(x,x'))d^4x' - \int_{V_4'} \partial_\alpha \partial'^\mu (A^\alpha(x') G(x,x'))d^4x' \\ - \int_{V_4'} \partial_\alpha ((\partial'^\alpha A^\mu(x') - \partial'^\mu A^\alpha(x'))G(x,x'))d^4x'.\end{aligned} \tag{78}$$

The third term on the rhs of (78) is in its final form. It can be shown in an entirely analogous manner as for the Euclidean four-space identity (54), (using a four-vector a_μ with constant magnitude and constant but arbitrary direction in Minkowski space), that the first two terms on the rhs of (78) combine to yield the Minkowski space identity

$$\begin{aligned}\int_{V_4'} \partial_\alpha \partial'^\alpha (A^\mu(x') G(x,x'))d^4x' - \int_{V_4'} \partial_\alpha \partial'^\mu (A^\alpha(x') G(x,x'))d^4x' \\ = - \oint_{\Sigma'} \partial_\alpha ((A^\alpha(x') n'^\mu - A^\mu(x') n'^\alpha) G(x,x'))d\Sigma'.\end{aligned} \tag{79}$$

Combining the results (77), (78), and (79) completes the proof of the Minkowski space identity (69) and Theorem II. ♣

In passing, consider a representative case where the Green's function is chosen to be the covariant retarded Green's function (67a). A convenient covariant three-surface Σ', (bounding a Minkowski four-volume V_4'), could then be taken as the union of a finite section of the forward light cone Σ_C', (a lightlike hypersurface), and its end cap comprised for example by an intersecting three-spherical conic section Σ_{cap}' (a spacelike hypersurface). The outward surface normal n'^μ to the forward light cone is a spacelike unit four-vector, while the outward surface normal to the three-spherical end cap of a finite section of the forward light cone is a timelike unit four-vector.

But since the forward light cone is the only surface where the retarded Green's function is different than zero, only the surface integrals over the finite section of the forward light cone Σ_C' can be nonzero!

Finally, note the similarity in structure of the Minkowski space identity (69) and the Euclidean four-space identity (43). Aside from the incorporation of the unprimed derivatives in the integrals of (69) (which was done to improve the convergence of the integrals), the only difference is that the Euclidean four-space Green's function $1/4\pi^2R^2$ is replaced by an appropriate Minkowski space Green's function $G(x,x')$.

It will be shown in Sec. III D that identity (69) is a Minkowski space generalization of the (Euclidean three-space) Helmholtz identity (11a). Therefore, just as (11a) can be used to prove the Helmholtz uniqueness theorems H2 and H1 of (static) three-vector fields, identity (69) will be used in Secs. III E and IV B to prove uniqueness theorems for four-vector fields in $\mathbb{R}^{3+1}$.

C. Scalar field identity in Euclidean three-space

The Euclidean three-space identity (8) is next substituted into a delta function property as follows:

$$\phi(\mathbf{r})=\int_{V'}\phi(\mathbf{r}')\,\delta^3(\mathbf{r}-\mathbf{r}')\,d^3\mathbf{r}'=\int_{V'}\phi(\mathbf{r}')\nabla^2\left(\frac{-1}{4\pi r}\right)d^3\mathbf{r}'. \tag{80}$$

Identity (80) is now used to state a theorem.

Theorem III: *The following identity holds for twice continuously differentiable (static) scalar fields ϕ in the Euclidean three-space $\mathbb{R}^3$:*

$$\phi(\mathbf{r})=\nabla\cdot\left[\int_{V'}\frac{-\nabla'\phi(\mathbf{r}')}{4\pi r}\,dV'+\oint_{S'}\frac{\phi(\mathbf{r}')\mathbf{n}'}{4\pi r}\,dS'\right], \tag{81}$$

where the three-dimensional distance $r\equiv|\mathbf{r}-\mathbf{r}'|$, and where $\mathbf{n}'$ is the three-vector outward unit normal of the two-surface S' which encloses the three-volume V'.

Proof: Starting with (80), for a twice continuously differentiable field ϕ, one can factor part of the Laplacian operator acting on the unprimed field point coordinates out of the integration over the primed source point coordinates to yield

$$\phi(\mathbf{r})=\nabla\cdot\left[\int_{V'}\frac{-\phi(\mathbf{r}')}{4\pi}\nabla\left(\frac{1}{r}\right)d^3x'\right]. \tag{82}$$

Using the vector identity

$$\nabla\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right)=-\nabla'\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right), \tag{83}$$

one can rewrite (82) as

$$\phi(\mathbf{r})=\nabla\cdot\left[\int_{V'}\frac{\phi(\mathbf{r}')}{4\pi}\nabla'\left(\frac{1}{r}\right)d^3x'\right]. \tag{84}$$

The vector identity

$$\nabla(\phi\psi)=\phi\nabla\psi+\psi\nabla\phi, \tag{85}$$

with $\psi\equiv1/|\mathbf{r}-\mathbf{r}'|$ becomes

$$\nabla\left(\frac{\phi}{|\mathbf{r}-\mathbf{r}'|}\right)=\phi\nabla\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right)+\frac{\nabla\phi}{|\mathbf{r}-\mathbf{r}'|},\tag{86}$$

allowing one to rewrite (84) as

$$\phi(\mathbf{r})=\nabla\cdot\left[\int_{V'}\frac{1}{4\pi}\left(\nabla'\left(\frac{\phi(\mathbf{r}')}{r}\right)-\frac{\nabla'\phi(\mathbf{r}')}{r}\right)d^3x'\right].\tag{87}$$

Now, using the vector identity[7]

$$\int_V \nabla\Phi\, dV=\oint_S \Phi\mathbf{n}\, dS,\tag{88}$$

which follows from Gauss' divergence theorem in a *flat space–time*, on the first term of (87) yields immediately identity (81), proving Theorem III. Identity (81) is of a form which is reminiscent of the Helmholtz identity (11a), but is applicable to static scalar fields. ♣

D. Derivation of the Helmholtz identity and a scalar field identity from a Minkowski space Helmholtz identity

The author will now show that identity (69) of Theorem II is a Minkowski space generalization of the (Euclidean three-space) Helmholtz identity (11a) in a static Newtonian limit. In addition, the fourth or scalar field component of (69) will be shown to yield identity (81) in a similar static Newtonian limit.

To prove that the Helmholtz identity (11a) follows from identity (69) in a static Newtonian limit, it is convenient to start with an identity which follows from (69), i.e.,

$$\begin{aligned}A^\mu(x)=&-\partial^\mu\left[\int_{V_4'}\frac{\partial_\nu' A^\nu(x')}{4\pi r}\,\delta(r-c(t-t'))d^4x'-\int_{V_4'}\partial_\nu'\left(\frac{A^\nu(x')}{4\pi r}\,\delta(r-c(t-t'))\right)d^4x'\right]\\&-\partial_\alpha\left[\int_{V_4'}\frac{\partial'^\alpha A^\mu(x')}{4\pi r}\,\delta(r-c(t-t'))d^4x'-\int_{V_4'}\frac{\partial'^\mu A^\alpha(x')}{4\pi r}\,\delta(r-c(t-t'))d^4x'\right.\\&\left.-\int_{V_4'}\partial'^\alpha\left(\frac{A^\mu(x')}{4\pi r}\,\delta(r-c(t-t'))\right)d^4x'+\int_{V_4'}\partial'^\mu\left(\frac{A^\alpha(x')}{4\pi r}\,\delta(r-c(t-t'))\right)d^4x'\right],\end{aligned}\tag{89}$$

where $r\equiv|\mathbf{r}-\mathbf{r}'|$. Identity (89) follows from (69) as a result of using the retarded Green's function (37), Gauss' divergence theorem on the surface integral in the first bracketed term of (69), and identity (79) on the surface integral in the second bracketed term of (69), while at the same time factoring the unprimed derivatives out of all the integrals.

One next takes the following as an intermediate approximation of the space components of identity (89):

$$\begin{aligned}A^j(x)=&-\partial^j\left[\int_{V_4'}\frac{\partial_k' A^k(x')}{4\pi r}\,\delta(r-c(t-t'))d^4x'-\int_{V_4'}\partial_k'\left(\frac{A^k(x')}{4\pi r}\,\delta(r-c(t-t'))\right)d^4x'\right]\\&-\partial_i\left[\int_{V_4'}\frac{\partial'^i A^j(x')}{4\pi r}\,\delta(r-c(t-t'))d^4x'-\int_{V_4'}\frac{\partial'^j A^i(x')}{4\pi r}\,\delta(r-c(t-t'))d^4x'\right.\\&\left.-\int_{V_4'}\partial'^i\left(\frac{A^j(x')}{4\pi r}\,\delta(r-c(t-t'))\right)d^4x'+\int_{V_4'}\partial'^j\left(\frac{A^i(x')}{4\pi r}\,\delta(r-c(t-t'))\right)d^4x'\right].\end{aligned}\tag{90}$$

Note, all terms involving partial time derivatives, e.g., $\partial_0 = (1/c)\partial/\partial t$, are omitted in (90) since a detailed dimensional analysis in the speed of propagation c shows that these terms vanish in a Newtonian limit where $c \to \infty$. The author will delay the application of the Newtonian limit for the terms retained in (90), however, until later in this derivation. As usual, Roman indices i,j,k, etc., are used here and in what follows to denote three-vectors. One next performs the t' time integrations in (90) yielding

$$\mathbf{A}(\mathbf{r},t) = -\nabla\left[\int_{V'} \frac{\nabla'\cdot\mathbf{A}(\mathbf{r}',t-r/c)}{4\pi r}dV' - \int_{V'} \nabla'\cdot\left(\frac{\mathbf{A}(\mathbf{r}',t-r/c)}{4\pi r}\right)dV'\right] + \nabla\times\left[\int_{V'} \frac{\nabla'\times\mathbf{A}(\mathbf{r}',t-r/c)}{4\pi r}dV' - \int_{V'} \nabla'\times\left(\frac{\mathbf{A}(\mathbf{r}',t-r/c)}{4\pi r}\right)dV'\right]. \tag{91}$$

It should be noted in passing that (91) is only an intermediate result where time derivative terms have already been neglected. Taking a Newtonian limit where $c \to \infty$ reduces (91) further to

$$\mathbf{A}(\mathbf{r},t) = -\nabla\left[\int_{V'} \frac{\nabla'\cdot\mathbf{A}(\mathbf{r}',t)}{4\pi r}dV' - \int_{V'} \nabla'\cdot\left(\frac{\mathbf{A}(\mathbf{r}',t)}{4\pi r}\right)dV'\right] + \nabla\times\left[\int_{V'} \frac{\nabla'\times\mathbf{A}(\mathbf{r}',t)}{4\pi r}dV' - \int_{V'} \nabla'\times\left(\frac{\mathbf{A}(\mathbf{r}',t)}{4\pi r}\right)dV'\right]. \tag{92}$$

Now, applying Gauss' divergence theorem to the second term of the first bracketed term of (92), while using the three-vector identity[6,7]

$$-\int_{V'} \nabla'\times\frac{\mathbf{A}(\mathbf{r}')}{4\pi r}dV' = \oint_{S'} \frac{\mathbf{A}(\mathbf{r}')\times\mathbf{n}'}{4\pi r}dS' \tag{93}$$

on the second term of the second bracketed term of (92), and taking $t=0$ for convenience, yields identity (11) as desired (since the t dependence is the same on both sides of the resulting equation and can be ignored in what amounts to a static field assumption). The Helmholtz identity (11a) therefore follows from the space components of (69) in a static Newtonian limit.

Next, to prove that identity (81) of Theorem III follows directly from the Minkowski space Helmholtz identity (69) in a static Newtonian limit, it is convenient to again start with identity (89) which follows from (69). One next takes the following as an intermediate approximation of the zeroth component of identity (89):

$$A^0(x) = -\partial_i\left[\int_{V_4'} \frac{\partial'^i A^0(x')}{4\pi r}\delta(r-c(t-t'))d^4x' - \int_{V_4'} \partial'^i\left(\frac{A^0(x')}{4\pi r}\delta(r-c(t-t'))\right)d^4x'\right]. \tag{94}$$

Note, all terms involving partial time derivatives, e.g., $\partial_0 = (1/c)\partial/\partial t$, are omitted in (94) since a detailed dimensional analysis in the speed of propagation c shows that these terms vanish in a Newtonian limit where $c \to \infty$. The author will delay the application of the Newtonian limit for the terms retained in (94), however, until later in this derivation. One next performs the t' time integrations in (94), while using three-vector notation and setting $\phi = cA^0$, yielding

$$\phi(\mathbf{r},t) = \nabla\cdot\left[\int_{V'} \frac{-\nabla'\phi(\mathbf{r}',t-r/c)}{4\pi r}dV' + \int_{V'} \nabla'\cdot\left(\frac{\phi(\mathbf{r}',t-r/c)}{4\pi r}\right)dV'\right]. \tag{95}$$

It should be noted in passing that (95) is only an intermediate result where time derivative terms have already been neglected. Taking a Newtonian limit where $c \to \infty$ reduces (95) further to

$$\phi(\mathbf{r},t)=\nabla\cdot\left[\int_{V'}\frac{-\nabla'\phi(\mathbf{r}',t)}{4\pi r}dV'+\int_{V'}\nabla'\cdot\left(\frac{\phi(\mathbf{r}',t)}{4\pi r}\right)dV'\right]. \tag{96}$$

Now, applying the three-vector Gauss' divergence theorem to the second term on the rhs of (96) and taking $t=0$ for convenience, yields identity (81) as desired, (since the t dependence is the same on both sides of the resulting equation and can be ignored in what amounts to a static field assumption). The scalar field identity (81) of Theorem III therefore follows from the zeroth or time component of (69) in a static Newtonian limit.

So, in a static Newtonian limit the space components of the Minkowski space identity (69) reduce to the Helmholtz identity (11a) and the zeroth component of (69) reduces to the scalar field identity (81). It seems reasonable to conclude therefore that identity (69) is a Minkowski space generalization of the (three-space) Helmholtz identity (11a).

E. Uniqueness theorems for four-vector fields in Euclidean and Minkowski four-spaces

In this section, two uniqueness theorems will be proved. First, a uniqueness theorem for four-vector fields in Euclidean space will be proved using identity (43) of Theorem I, i.e., using the Euclidean four-space generalization of the Helmholtz identity. Then, a uniqueness theorem for four-vector fields in Minkowski space will be proved using identity (69) of Theorem II, i.e., using the Minkowski four-space generalization of the Helmholtz identity.

A uniqueness theorem for four-vector fields in Euclidean four-space is now stated.

Theorem IV: *A sufficiently smooth four-vector field A^ν in the Euclidean four-space $\mathbb{R}^4$ is uniquely specified by giving its four-divergence and its four-curl within a four-space region V_4, as well as its normal and tangential components on the bounding three-surface Σ. That is, one must specify the following:*

$$\partial_\nu A^\nu(x^\sigma)\equiv s, \tag{97a}$$

$$A^\nu(x^\sigma)n_\nu\equiv A_{\text{norm}}(x^\sigma), \tag{97b}$$

$$\partial^\alpha A^\mu(x^\sigma)-\partial^\mu A^\alpha(x^\sigma)\equiv c^{\alpha\mu}, \tag{97c}$$

$$A^\alpha(x^\sigma)n^\mu-A^\mu(x^\sigma)n^\alpha\equiv A^{\alpha\mu}_{\text{tang}}(x^\sigma), \tag{97d}$$

where n^ν is the four-vector outward unit normal of the three-surface Σ which encloses the four-volume V_4.

Proof: In order to demonstrate the uniqueness of the four-vector field A^ν, one first postulates the existence of a second four-vector B^ν, which also satisfies Eq. (97). That is, one only replaces A^ν on the lhs of Eq. (97) by B^ν while the rhs of Eq. (97) remains unchanged. The four-vector field A^ν is unique if one can show that

$$W^\nu\equiv A^\nu-B^\nu=0. \tag{98}$$

Now, taking the four-divergence of W^ν and using (97a) yields

$$\partial_\nu W^\nu=\partial_\nu A^\nu-\partial_\nu B^\nu=s-s=0, \tag{99}$$

everywhere in the four-volume V_4. Also, calculating the magnitude of the normal component of W^ν along the surface normal n^ν and using (97b) yields

$$W_{\text{norm}}\equiv W^\nu n_\nu=A^\nu n_\nu-B^\nu n_\nu=A_{\text{norm}}-A_{\text{norm}}=0, \tag{100}$$

everywhere on the bounding three-surface Σ. Next, taking the four-curl of W^ν and using (97c) yields

$$\partial^{\alpha}W^{\mu}-\partial^{\mu}W^{\alpha}=(\partial^{\alpha}A^{\mu}-\partial^{\mu}A^{\alpha})-(\partial^{\alpha}B^{\mu}-\partial^{\mu}B^{\alpha})=c^{\alpha\mu}-c^{\alpha\mu}=0, \tag{101}$$

everywhere in the four-volume V_4. Also, calculating the rank two tangential components with respect to n^{ν} of W^{ν} and using (97d) yields

$$W^{\alpha}n^{\mu}-W^{\mu}n^{\alpha}=(A^{\alpha}n^{\mu}-A^{\mu}n^{\alpha})-(B^{\alpha}n^{\mu}-B^{\mu}n^{\alpha})=A^{\alpha\mu}_{\text{tang}}-A^{\alpha\mu}_{\text{tang}}=0, \tag{102}$$

everywhere on the bounding three-surface Σ. Finally, substituting the results (99)–(102) for the four-vector field W^{ν} into identity (43) of Theorem I, one obtains the result $W^{\nu}=0$ which implies [via (98)] that $A^{\nu}=B^{\nu}$ everywhere in the four-space region V_4' and on its bounding three-surface Σ'. This proves that the four-vector field A^{ν} is uniquely determined by Eq. (97) thus proving Theorem IV on the uniqueness of four-vector fields in Euclidean four-space. ♣

A uniqueness theorem for four-vector fields in Minkowski space is now stated.

Theorem V: *A sufficiently smooth four-vector field $A^{\mu}(x^{\sigma})$ in the Minkowski space $\mathbb{R}^{3+1}$ which satisfies identity* (69) (*i.e., Theorem II*) *is uniquely specified by giving its four divergence and its four-curl within a space–time region V_4, as well as its normal and tangential components on the bounding three-surface Σ. That is, one must specify the following:*

$$\partial_{\nu}A^{\nu}(x^{\sigma})\equiv s, \tag{103a}$$

$$\partial^{\alpha}A^{\mu}(x^{\sigma})-\partial^{\mu}A^{\alpha}(x^{\sigma})\equiv c^{\alpha\mu}, \tag{103b}$$

throughout the space–time region V_4, as well as

$$A^{\nu}(x^{\sigma})n_{\nu}\equiv A_{\text{norm}}(x^{\sigma}), \tag{103c}$$

$$A^{\alpha}(x^{\sigma})n^{\mu}-A^{\mu}(x^{\sigma})n^{\alpha}\equiv A^{\alpha\mu}_{\text{tang}}(x^{\sigma}), \tag{103d}$$

everywhere on the bounding three-surface Σ, where n^{ν} is the four-vector outward unit normal of the three-surface Σ which encloses the space–time four-volume V_4.

Proof: The proof is based on the Minkowski space Theorem II. The proof proceeds in a parallel manner to the proof of the Euclidean four-space Theorem IV. One postulates the existence of a second four-vector B^{ν} which also satisfies Eq. (103). The four-vector field A^{ν} is unique if one can show, as in Theorem IV, that

$$W^{\nu}\equiv A^{\nu}-B^{\nu}=0. \tag{104}$$

Equation (103) then leads, as before with (97), to results analogous to (99)–(102), which when substituted into identity (69) shows that $W^{\nu}=0$ which implies [via the definition of W^{ν} in (104)] that $A^{\nu}=B^{\nu}$ everywhere in the Minkowski space region V_4' and on its bounding three-surface Σ'. This proves that the four-vector field A^{ν} is uniquely determined by Eq. (103) specified over Minkowski space, thus proving Theorem V on the uniqueness of four-vector fields in Minkowski space. ♣

It has already been demonstrated that identity (69) of Theorem II reduces to the Helmholtz identity (11a) when an appropriate (static) Newtonian limit is taken. And it is known that the Helmholtz identity (11a) can be used to prove the Helmholtz uniqueness theorem for a finite volume of $\mathbb{R}^3$, i.e., Theorem H2. Theorems IV and V are later used in Sec. IV B to prove Theorem X, which extends Theorem H2 to four-vector fields in Euclidean and Minkowski spaces. [If one makes the usual assumption that the four-vector fields A^{μ} vanish sufficiently rapidly at infinity, then the surface integral terms in (43) or (69) vanish and so a four-space generalization of the Helmholtz uniqueness Theorem H1 over the entire four-volume of Euclidean or Minkowski space follows from Theorem X as well.] Also, Theorem's IV and V can be interpreted as extensions of Theorem U in Sec. II A to four-vector fields in Euclidean and Minkowski spaces, respectively. The Helmholtz identity (11a) and the three-vector uniqueness theorems U, H1, and H2 of Sec. II A can therefore be generalized readily to four-vector fields in Euclidean and Minkowski spaces.

F. Uniqueness theorem for scalar fields in Euclidean three-space

In this section, a uniqueness theorem for scalar fields in Euclidean three-space will be proved using identity (81) of Theorem III. The theorem will now be stated.

Theorem VI: *A twice continuously differentiable (static) scalar field $\phi(\mathbf{r})$ in the Euclidean three-space $\mathbb{R}^3$ is uniquely specified by giving its gradient everywhere within a spatial volume V, as well as its value on the bounding surface S.* That is, one must specify the following:

$$-\nabla\phi(\mathbf{r})\equiv\mathbf{E}(\mathbf{r}) \tag{105a}$$

throughout the volume V, as well as

$$\phi(\mathbf{r})|_S\equiv\phi_S(\mathbf{r}) \tag{105b}$$

on the bounding surface S.

Proof: In order to demonstrate the uniqueness of the scalar field $\phi(\mathbf{r})$, one first postulates the existence of a second scalar field $\xi(\mathbf{r})$, which also satisfies Eq. (105). That is, one replaces $\phi(\mathbf{r})$ on the lhs of Eq. (105) by $\xi(\mathbf{r})$, while the rhs of Eq. (105) remains unchanged. The scalar field $\phi(\mathbf{r})$ is unique if one can show that

$$\Phi(\mathbf{r})\equiv\phi(\mathbf{r})-\xi(\mathbf{r})=0. \tag{106}$$

Now, taking minus one times the gradient of $\Phi(\mathbf{r})$ and using (105a) yields

$$-\nabla\Phi(\mathbf{r})=-\nabla\phi(\mathbf{r})+\nabla\xi(\mathbf{r})=\mathbf{E}(\mathbf{r})-\mathbf{E}(\mathbf{r})=0 \tag{107}$$

for all $\mathbf{r}$ in the spatial volume V. Next, evaluating Φ on the boundary surface S and using (105b) yields

$$\Phi(\mathbf{r})|_S=\phi(\mathbf{r})|_S-\xi(\mathbf{r})|_S=\phi_S(\mathbf{r})-\phi_S(\mathbf{r})=0, \tag{108}$$

for all $\mathbf{r}$ on the surface S. Finally, using the results (107) and (108) for the scalar field $\Phi(\mathbf{r})$ in identity (81) of Theorem III, one obtains the result $\Phi(\mathbf{r})=0$ which implies [via the definition of $\Phi(\mathbf{r})$ in (106)] that $\phi(\mathbf{r})=\xi(\mathbf{r})$ everywhere in the three-space region V' and on its bounding surface S'. This proves that the scalar field $\phi(\mathbf{r})$ is uniquely determined by Eq. (105) thus proving Theorem VI on the uniqueness of scalar fields in Euclidean three-space. ♣

Parenthetically, one should take note of the definition of the vector field $\mathbf{E}$ in (105a), which for example in electromagnetism could be interpreted as the static electric field. From this point of view, the uniqueness of a static electric scalar potential $\phi(\mathbf{r})$ requires the specification of the static electric field in a volume V, as well as the value of the scalar potential evaluated on the bounding surface S.

It should also be noted that Theorem VI bears little resemblance to a typical statement of a uniqueness theorem for a scalar field in Euclidean three-space (cf. pp. 38–45 in Jackson—Ref. 18). This is because Theorem VI is based on identity (81) of Theorem III, which is based on identity (80) alone. Equation (80) in turn follows from a solution of the *inhomogeneous* scalar wave equation, i.e., from an *inhomogeneous* Green's function approach. The *homogeneous* solutions which would have lead to Dirichlet or Neumann boundary conditions have not been included. The importance of Theorem VI appears to be that it emphasizes the role of the electric field, i.e., the negative gradient of ϕ, in obtaining a unique scalar potential.

Essentially the same remarks as above would apply to Theorems IV and V, but in a four-vector potential context. The importance of the four-vector result appears to be that it emphasizes the role of the four-divergence and the four-curl of A^μ in obtaining a unique four-vector potential.

IV. EXTENSION TO FIELDS WITH MASS

A. Helmholtz identity and uniqueness theorem for fields with mass

In order to extend the Helmholtz identity (11a) to vector fields with mass it is convenient to start by adding a μ^2 mass term to the *scalar* Poisson equation for the static scalar potential ϕ, as follows:

$$(\nabla^2-\mu^2)\phi(x,y,z)=-\frac{\rho(x,y,z)}{\epsilon_0}. \tag{109}$$

The massive scalar Poisson equation (109) can be solved in terms of a two-point scalar Green's function $G(\mathbf{r},\mathbf{r}')$ which connects its unit delta function source located at the source point $\mathbf{r}'=(x',y',z')$ to a measurement at the field point $\mathbf{r}=(x,y,z)$, i.e.,

$$(\nabla^2-\mu^2)G(\mathbf{r},\mathbf{r}')=-\delta^3(\mathbf{r}-\mathbf{r}'). \tag{110}$$

The well-known identity (cf. Ref. 19) over the Euclidean three-space $\mathbb{R}^3$, namely

$$(\nabla^2-\mu^2)\frac{e^{-\mu r}}{4\pi r}\equiv-\delta^3(\mathbf{r}-\mathbf{r}') \quad \forall \mathbf{r},\mathbf{r}'\in\mathbb{R}^3, \tag{111}$$

where $r\equiv|\mathbf{r}-\mathbf{r}'|$, yields by comparison with (110) for the case of an infinite spatial domain, the Green's function of the $(\nabla^2-\mu^2)$ operator as

$$G(\mathbf{r},\mathbf{r}')=\frac{e^{-\mu r}}{4\pi r}=\frac{e^{-\mu|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}. \tag{112}$$

One recognizes in (111) and (112) an exponential damping factor depending in this example on the Compton wave number $\mu=2\pi/\lambda_C=mc/\hbar$ of interaction bosons of mass m. The inhomogeneous solution of (109) for the scalar potential then follows from the integral

$$\phi(\mathbf{r})=\int_{V'}\frac{\rho(\mathbf{r}')}{\epsilon_0}G(\mathbf{r},\mathbf{r}')d^3\mathbf{r}'=\int_{V'}\frac{\rho(\mathbf{r}')e^{-\mu|\mathbf{r}-\mathbf{r}'|}}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|}d^3\mathbf{r}'. \tag{113}$$

Although (113) is an inhomogeneous solution to the massive scalar Poisson equation (109) in an infinite spatial domain, it is the delta function property of identity (111) which is of importance in the proof of the new identity, namely

$$\mathbf{F}(\mathbf{r})=\int_{V'}\mathbf{F}(\mathbf{r}')\delta^3(\mathbf{r}-\mathbf{r}')d^3\mathbf{r}'=\int_{V'}\mathbf{F}(\mathbf{r}')(\nabla^2-\mu^2)\left(\frac{-e^{-\mu|\mathbf{r}-\mathbf{r}'|}}{4\pi r}\right)d^3\mathbf{r}'. \tag{114}$$

In addition, the identity $\nabla(1/r)=-\nabla'(1/r)$ used in the derivation of (11a) must be replaced by the identity

$$\nabla G(\mathbf{r},\mathbf{r}')=-\nabla' G(\mathbf{r},\mathbf{r}'), \tag{115}$$

which follows from the Green's function (112). A new identity now follows from (114) and (115) which allows one to state the following theorem.

Theorem VII: *The following identity holds for a continuous (static) three-vector field* $\mathbf{F}(\mathbf{r})$ *in a Euclidean three-space* $\mathbb{R}^3$:

$$\mathbf{F}(\mathbf{r}) = -\nabla\left[\int_{V'} \frac{\nabla'\cdot\mathbf{F}(\mathbf{r}')}{4\pi r} e^{-\mu r}\, dV' - \oint_{S'} \frac{\mathbf{F}(\mathbf{r}')\cdot\mathbf{n}'}{4\pi r} e^{-\mu r}\, dS'\right] + \nabla\times\left[\int_{V'} \frac{\nabla'\times\mathbf{F}(\mathbf{r}')}{4\pi r} e^{-\mu r} dV' + \oint_{S'} \frac{\mathbf{F}(\mathbf{r}')\times\mathbf{n}'}{4\pi r} e^{-\mu r}\, dS'\right] + \mu^2 \int_{V'} \frac{\mathbf{F}(\mathbf{r}')}{4\pi r} e^{-\mu r}\, dV', \tag{116}$$

where $r \equiv |\mathbf{r} - \mathbf{r}'|$ *and where* $\mathbf{n}'$ *is the unit surface normal pointing out of the volume* V' *bounded by the closed surface* S'.

Proof: Starting with (114), since the Laplacian operator acts only on the field coordinates, it can be brought outside of the integration. At this point, one makes a decomposition of the Laplacian operator using identity (20) as before in the derivation of (11a).[6,7] Then, since identity (115) retains the same functional form as the identity $\nabla(1/r) = -\nabla'(1/r)$ with respect to the overall derivation, the derivation retains the same form as the derivation of (11), with the minor exception that an extra μ^2 term is carried along unchanged. Equation (116) therefore follows readily as a new identity for static three-vector fields thus proving Theorem VII. ♣

In contrast to the Helmholtz identity (11a), the integrands of identity (116) contain an additional exponential mass damping factor which improves their convergence. The integrals are therefore well defined even for fields **F** falling off only as fast as $1/r$ (e.g., potentials).

Note, Eq. (116) should be thought of as an identity for representing a general (static) three-vector field rather than as a general solution to a partial differential equation. Indeed, (116) follows from the identity (111) over an infinite spatial domain and its subsequent use in the delta function property (114). Consequently, (116) is a vector identity that applies to all of the Euclidean three-space, and so must hold for a finite volume of it as well. The Green's function for an *inhomogeneous* massive scalar Poisson equation in an infinite spatial domain is only mentioned in passing, and certainly no use is made of the solutions of the associated source free *homogeneous* scalar equation.

It is interesting at this point to inquire whether or not identity (116) of Theorem VII can be used to obtain an alternate version of Theorem H2 of Sec. II A. However, the first thing to note about identity (116) is that **F** is no longer simply of the form (3) i.e., involving only the gradient of a scalar field (an irrotational part), and the curl of a vector field (a solenoidal part), but is now of the form

$$\mathbf{F}(\mathbf{r}) = -\nabla\Phi(\mathbf{r}) + \nabla\times\mathbf{A}(\mathbf{r}) + \mu^2 \int_{V'} \frac{\mathbf{F}(\mathbf{r}')}{4\pi r} e^{-\mu r}\, dV', \tag{117}$$

where the first and second bracketed terms on the rhs of (116) are set equal to $\Phi(\mathbf{r})$ and $\mathbf{A}(\mathbf{r})$, respectively, and where the last term on the rhs of (117) is an extra nonzero term which is neither irrotational nor solenoidal. Consequently, an alternate version of Theorem H2 does not appear to follow from identity (116).

On the other hand, one can use identity (116) to prove the following uniqueness theorem for three-vector fields.

Theorem VIII: *A twice continuously differentiable (static) three-vector field* $\mathbf{F}(\mathbf{r})$ *in the Euclidean three-space* $\mathbb{R}^3$, *which satisfies identity* (116), *is uniquely specified by giving its divergence and curl within the volume* V, *its normal and tangential components on the bounding surface S, and the value of the real constant* μ. *That is, one must specify the constant* μ *and the following:*

$$\nabla\cdot\mathbf{F} = \rho, \tag{118a}$$

$$\nabla\times\mathbf{F} = \mathbf{j}, \tag{118b}$$

throughout the volume V, *along with the normal and tangential components*

$$\mathbf{F}\cdot\mathbf{n}=-\sigma, \tag{118c}$$

$$\mathbf{F}\times\mathbf{n}=\mathbf{K}, \tag{118d}$$

respectively, on the surface S bounding the volume V, where **n** *is the outward unit normal vector of the surface S which encloses the volume V.*

Proof: In order to demonstrate the uniqueness of massive vector fields which satisfy identity (116) and Eqs. (118a)–(118d), one first postulates the existence of a second vector **G** which also satisfies identity (116) and Eqs. (118a)–(118d). That is, one replaces **F** on the lhs of Eqs. (118a)–(118d) by **G** while the rhs of these equations remain unchanged. The vector field **F** is unique if one can show that

$$\mathbf{W}\equiv\mathbf{F}-\mathbf{G}=0. \tag{119}$$

Now, taking the divergence of **W** and using (118a) yields

$$\nabla\cdot\mathbf{W}=\nabla\cdot\mathbf{F}-\nabla\cdot\mathbf{G}=\rho-\rho=0, \tag{120}$$

everywhere in the volume V. Also, calculating the magnitude of the normal component of **W** along the surface normal **n** and using (118c) yields

$$\mathbf{W}_{\text{norm}}\equiv\mathbf{W}\cdot\mathbf{n}=\mathbf{F}\cdot\mathbf{n}-\mathbf{G}\cdot\mathbf{n}=-\sigma-(-\sigma)=0, \tag{121}$$

everywhere on the bounding surface S. Next, taking the curl of **W** and using (118b) yields

$$\nabla\times\mathbf{W}=\nabla\times\mathbf{F}-\nabla\times\mathbf{G}=\mathbf{j}-\mathbf{j}=0, \tag{122}$$

everywhere in the volume V. Also, calculating the tangential components with respect to **n** of **W** and using (118d) yields

$$\mathbf{W}_{\text{tang}}\equiv\mathbf{W}\times\mathbf{n}=\mathbf{F}\times\mathbf{n}-\mathbf{G}\times\mathbf{n}=\mathbf{K}-\mathbf{K}=0, \tag{123}$$

everywhere on the bounding surface S. Finally, substituting the results (120)–(123) for the vector field **W** into identity (116) yields

$$\mathbf{W}(\mathbf{r})=\mu^2\int_{V'}\frac{\mathbf{W}(\mathbf{r}')}{4\pi r}e^{-\mu r}\,dV'. \tag{124}$$

Then, if one applies the operator $(\nabla^2-\mu^2)$, which acts only on the unprimed coordinates, to (124), while using (111) and (114), one obtains

$$(\nabla^2-\mu^2)\mathbf{W}(\mathbf{r})=\mu^2\int_{V'}\mathbf{W}(\mathbf{r}')(\nabla^2-\mu^2)\frac{e^{-\mu r}}{4\pi r}\,dV'=-\mu^2\int_{V'}\mathbf{W}(\mathbf{r}')\,\delta(\mathbf{r}-\mathbf{r}')\,dV'=-\mu^2\mathbf{W}(\mathbf{r}). \tag{125}$$

The μ^2 terms in (125) cancel and so **W** satisfies the vector Laplace equation

$$\nabla^2\mathbf{W}(\mathbf{r})=0. \tag{126}$$

Equation (126) also follows from a direct expansion of the Laplacian operator using (20) as follows:

$$\nabla^2\mathbf{W}=\nabla(\nabla\cdot\mathbf{W})-\nabla\times(\nabla\times\mathbf{W})=0, \tag{127}$$

where the first term in (127) vanishes by (120) and the second term in (127) vanishes by (122). A further restriction on **W** is that its normal component vanishes on the surface S by (121), which combined with the fact that **W** is irrotational by (122) and is therefore expressible as **W**

$=-\nabla\phi$, implies via Green's theorem that $\mathbf{W}=0$ throughout the volume V as well.[8] So, only the trivial solution $\mathbf{W}=0$ of (124) satisfies the boundary conditions and therefore [via (119)], the vector field $\mathbf{F}=\mathbf{G}$ everywhere in the volume V and on its bounding surface S. This proves that a static vector field $\mathbf{F}$ in a Euclidean three-space satisfying identity (116) is uniquely determined by specifying the real scalar constant μ and relations (118a)–(118d), thus proving Theorem VIII.♣

It is tempting to interpret Theorem VIII as a uniqueness theorem for static massive vector fields, i.e., those which satisfy the inhomogeneous massive *vector* Poisson equation

$$(\nabla^2-\mu^2)\mathbf{F}(\mathbf{r})=-\mathbf{j}(\mathbf{r}). \tag{128}$$

However, the Green's function which was used to derive identity (116) followed from an inhomogeneous massive scalar Poisson equation which is a much simpler problem. That is, equations of the form (128) would in general use, for example, a solution technique which involves a two-point dyadic Green's function, i.e., a Green's function which is not a three-vector.[20] Nevertheless, a close inspection of the proofs of identity (116) and Theorem VIII reveals that no limitations are imposed on the vector field $\mathbf{F}$ other than that its components must be twice continuously differentiable. The vector field $\mathbf{F}$ could therefore be either a massless or a massive vector field. Naturally, identity (116) appears to be oriented toward application to massive vector fields due to its incorporation of a mass damping factor. Alternately, identity (116) could be used in situations involving spatial diffusion, where the parameter μ would be interpreted as a diffusion parameter.

B. Extending the four-space Helmholtz identities and uniqueness theorems to four-vector fields with mass

In attempting to extend the four-space Helmholtz identities to four-vector fields with mass, one can start by adding a μ^2 mass term to (32), the inhomogeneous scalar wave equation for the scalar potential ϕ, as follows:

$$\Box\phi(x^\nu)-\mu^2\phi(x^\nu)=-\rho(x^\nu)/\epsilon_0, \tag{129}$$

where $\Box\equiv\partial_\mu\partial^\mu$ is the d'Alembertian operator, and where the $(-\ +\ +\ +)$ metric signature is again convenient. One can then attempt a solution of the massive inhomogeneous scalar wave equation (129) in terms of a two-point massive scalar Green's function $G(x^\nu,x'^\nu)$ as in (33), but now with a μ^2 mass term as follows:

$$(\Box-\mu^2)G(x^\nu,x'^\nu)=\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}-\mu^2\right)G(x^\nu,x'^\nu)=-\delta^{(4)}(x^\nu-x'^\nu), \tag{130}$$

where $\delta^{(4)}(x^\nu-x'^\nu)=\delta(x^0-x'^0)\,\delta^3(\mathbf{r}-\mathbf{r}')$. For the previous scalar field case with the mass factor $\mu=0$, comparison of the Euclidean space analytically continued integral (30b) with the Minkowski space integral (34) yielded the Euclidean four space Green's function (36). One could therefore try to follow a course paralleling the three-space analysis in Sec. IV A by simply adding a mass damping factor to the Euclidean four-space Green's function (36) thereby obtaining a trial Green's function

$$G_{\text{trial}}(x^\nu,x'^\nu)\equiv\frac{e^{-\mu R}}{4\pi^2R^2}=\frac{e^{-\mu|x^\nu-x'^\nu|}}{4\pi^2|x^\nu-x'^\nu|^2}, \tag{131}$$

where R is the distance in a *Euclidean* four-space between the source point x'^ν and field point x^ν as defined previously in (23). However, it turns out that G_{trial} satisfies a different identity, namely

$$\left(\Box-\mu^2-\frac{\mu}{R}\right)G_{\text{trial}}(x^\nu,x'^\nu)=-\delta^{(4)}(x^\nu-x'^\nu), \tag{132}$$

where the third term on the lhs of (132) is an additional cross term. To demonstrate this one uses the result $\partial^\nu x_\nu = 4$ to calculate

$$\Box\left(\frac{e^{-\mu R}}{R^2}\right) = -4\pi^2\delta^{(4)}(x^\nu - x'^\nu)e^{-\mu R} + (2-4+3)\frac{\mu}{R^3}e^{-\mu R} + \frac{\mu^2}{R^2}e^{-\mu R}, \tag{133}$$

which proves (132) since the exponential in the first term on the rhs of (133) drops out in a delta function distribution integral context. Contrast (133) with the Euclidean three-space identity (110) rewritten in index notation, and using the result $\partial^j x_j = 3$, as

$$\frac{\partial}{\partial x_j}\frac{\partial}{\partial x^j}\left(\frac{e^{-\mu r}}{r}\right) = -4\pi\delta^3(x^j - x'^j)e^{-\mu r} + (1-3+2)\frac{\mu}{r^2}e^{-\mu r} + \frac{\mu^2}{r}e^{-\mu r}, \tag{134}$$

where the additional cross term, the middle term on the rhs of (134), clearly drops out.

However, a solution for the massive scalar two-point Green's function has been obtained by DeWitt[21] using a Fourier ''momentum'' space method for the case of timelike separations $(x^\nu - x'^\nu)^2 < 0$ as follows:

$$G_M(x^\nu, x'^\nu) = -\frac{\mu^2}{8\pi}\frac{H_1^{(2)}(i\mu R)}{i\mu R}, \tag{135}$$

where $H_1^{(2)}$ is the Hankel function of the second kind of order 1 with an integral representation[21]

$$H_1^{(2)}(z) = \frac{1}{i\pi}\int_C \frac{1}{u^2}\exp\left[\frac{1}{2}z\left(1-\frac{1}{u}\right)\right]du, \tag{136}$$

over the contour C defined in Ref. 21. Since $z = i\mu R = i\mu|x^\nu - x'^\nu|$ in (136), the four-vector derivative of the massive scalar Green's function (135) has the property

$$\frac{\partial}{\partial x'^\nu}G_M(x^\nu, x'^\nu) = -\frac{\partial}{\partial x^\nu}G_M(x^\nu, x'^\nu), \tag{137}$$

which is of the same functional form as properties (65a) and (65b). Consequently, (137) can be used in the same way as in the derivation of identity (69). An alternate form of Theorem II can therefore be stated as follows:

Theorem IX: *Given that a massive scalar two-point Green's function $G_M(x^\nu, x'^\nu)$ exists for timelike separations $(x^\nu - x'^\nu)^2 < 0$, whose dependencies on the coordinates x^ν and x'^ν occur only through the variable $R = |x^\nu - x'^\nu|$, the following identity holds for sufficiently smooth four-vector fields $A^\mu(x^\sigma)$ in the Minkowski space $\mathbb{R}^{3+1}$ for timelike separations $(x^\nu - x'^\nu)^2 < 0$:*

$$\begin{aligned} A^\mu(x) = &-\left[\int_{V_4'}\partial^\mu((\partial'_\nu A^\nu(x'))G_M(x,x'))d^4x' - \oint_{\Sigma'}\partial^\mu((A^\nu(x')n'_\nu)G_M(x,x'))d\Sigma'\right] \\ &-\left[\int_{V_4'}\partial_\alpha((\partial'^\alpha A^\mu(x') - \partial'^\mu A^\alpha(x'))G_M(x,x'))d^4x' + \oint_{\Sigma'}\partial_\alpha((A^\alpha(x')n'^\mu \right. \\ &\left. - A^\mu(x')n'^\alpha)G_M(x,x'))d\Sigma'\right] + \mu^2\int_{V_4'}A^\mu(x')G_M(x,x')d^4x', \end{aligned} \tag{138}$$

where n'^μ is the four-vector outward unit normal of the three-surface Σ' which encloses the four-volume V_4', and where the three-surface Σ' is defined covariantly with respect to a general Lorentz transformation.

Proof: The proof is based on the four-vector delta function property (42) which in the present case uses (130) for representing the four-space delta function as follows:

$$A^{\mu}(x)=\int_{V_4'}A^{\mu}(x')\,\delta^{(4)}(x-x')d^4x'=-\int_{V_4'}A^{\mu}(x')(\Box-\mu^2)G_M(x,x')d^4x'. \tag{139}$$

Then since the property (137) is of the same functional form as in (65), the proof of (138) parallels the proof of (69) in all important respects, except that an extra μ^2 mass term is carried along unchanged from (139), and so the details will not be repeated. ♣

It is interesting at this point to inquire whether or not one can use identity (138) of Theorem IX to state a four-space analog of the three-vector Theorem H2 of Sec. II A. For a sufficiently smooth four-vector field one can factor the unprimed field point derivatives out of the integrals over the primed source point coordinates in (138). However, the first thing to note about such a factored version of identity (138) is that A^{μ} is no longer simply of the form (75), i.e., involving only the four-gradient of a scalar field (a four-irrotational part), and the four-curl of a four-vector field (a four-solenoidal part), but is now of the form

$$A^{\mu}(x)=\partial^{\mu}A(x)+\partial_{\alpha}A^{\alpha\mu}(x)+\mu^2\int_{V_4'}A^{\mu}(x')G_M(x,x')d^4x', \tag{140}$$

where the first and second bracketed terms on the rhs of (138) are set equal to $-A(x)$ and $-A^{\alpha\mu}(x)$, respectively, and where the last term on the rhs of (140) is an extra nonzero term which is neither four-irrotational nor four-solenoidal. Therefore a four-space analog of the three-vector Theorem H2 of Sec. II A does not appear to follow from (138). However, it is possible to state a theorem for four-vector fields based on identity (43) of Theorem I or on identity (69) of Theorem II as follows:

Theorem X: *A sufficiently smooth four-vector field $A^{\mu}(x^{\sigma})$ that is defined everywhere in a finite volume V_4 in a Euclidean four-space $\mathbb{R}^4$ or in a Minkowski space $\mathbb{R}^{3+1}$ and whose tangential and normal components on the bounding three-surface Σ are given may be uniquely represented as a sum of a four-irrotational and a four-solenoidal part.*

Proof: It has already been shown that (44) leads to (51) in the Euclidean case, while in a similar fashion (70) leads to (75) in the Minkowski case. Now, the second term of (51) or (75), is four-irrotational

$$\partial^{\mu}(\partial^{\nu}A)-\partial^{\nu}(\partial^{\mu}A)=0, \tag{141}$$

i.e., its four-curl is zero. Also, the first term of (51) or (75) is four-solenoidal

$$\partial_{\mu}(\partial_{\alpha}A^{\alpha\mu})=0, \tag{142}$$

i.e., its four-divergence is zero, since it is a contraction of a symmetric factor $\partial_{\mu}\partial_{\alpha}$ and an antisymmetric factor (53) or (73). The decomposition defined by Eq. (51) or (75) is therefore a sum of a four-irrotational and a four-solenoidal part and by identity (43) and (69) and by the arguments of the Euclidean uniqueness Theorem IV and the Minkowski space uniqueness Theorem V the field A^{μ} is unique under this decomposition, thereby proving Theorem X. ♣

A theorem will now be stated that will be used later in this section.

Theorem XI: *A sufficiently smooth four-vector field $A^{\mu}(x^{\sigma})$ vanishes in a compact nonempty region $\overline{V_4}=V_4\cup\Sigma$ of the Minkowski space $\mathbb{R}^{3+1}$ when the four-divergence and four-curl of $A^{\mu}(x^{\sigma})$ vanish over the space–time region V_4 and the four-normal and four-tangential components of $A^{\mu}(x^{\sigma})$ vanish over the bounding three-surface Σ. That is, one must specify the following:*

$$\partial_{\nu}A^{\nu}(x^{\sigma})=0, \tag{143a}$$

$$\partial^{\alpha}A^{\mu}(x^{\sigma})-\partial^{\mu}A^{\alpha}(x^{\sigma})=0, \tag{143b}$$

throughout the space–time region V_4*, as well as*

$$A_{\text{norm}}(x^\sigma)\equiv A^\nu(x^\sigma)n_\nu=0, \tag{143c}$$

$$A_{\text{tang}}^{\alpha\mu}(x^\sigma)\equiv A^\alpha(x^\sigma)n^\mu - A^\mu(x^\sigma)n^\alpha=0, \tag{143d}$$

everywhere on the bounding three-surface Σ*, where* n^ν *is the four-vector outward unit normal of the three-surface* Σ *which encloses the space–time four-volume* V_4.

Proof: Since it is assumed that A^μ and n^ν are four-vectors, then $A_{\text{norm}}(x^\sigma)$ and $A_{\text{tang}}^{\alpha\mu}(x^\sigma)$ are covariant with respect to general Lorentz transformations and identity (69) of Theorem II, which is defined under similar constraints, can be used. Substitution of (143a)–(143d) into identity (69) then yields the result $A^\mu(x^\sigma)=0$ throughout $\overline{V_4}$, thus proving Theorem XI. ♣

The next step is to state a four-vector uniqueness theorem analogous to the three-vector Theorem VIII as follows:

Theorem XII: *A sufficiently smooth four-vector field* $A^\mu(x^\sigma)$ *in the Minkowski space* $\mathbb{R}^{3+1}$ *which satisfies identity (138) (i.e., Theorem IX), is uniquely specified by giving its four-divergence and four-curl within the space-time region* V_4*, its normal and tangential components on the bounding three-surface* Σ*, and the value of the real constant* μ*. That is, one must specify the constant* μ *and the following:*

$$\partial_\nu A^\nu(x^\sigma)\equiv s, \tag{144a}$$

$$\partial^\alpha A^\mu(x^\sigma) - \partial^\mu A^\alpha(x^\sigma)\equiv c^{\alpha\mu}, \tag{144b}$$

throughout the space–time region V_4*, as well as*

$$A_{\text{norm}}(x^\sigma)\equiv A^\nu(x^\sigma)n_\nu, \tag{144c}$$

$$A_{\text{tang}}^{\alpha\mu}(x^\sigma)\equiv A^\alpha(x^\sigma)n^\mu - A^\mu(x^\sigma)n^\alpha, \tag{144d}$$

everywhere on the bounding three-surface Σ*, where* n^ν *is the four-vector outward unit normal of the three-surface* Σ *which encloses the space–time four-volume* V_4.

Proof: The proof proceeds in a parallel manner to the proof of Theorem V. One postulates the existence of a second four-vector B^ν which also satisfies Eqs. (144a)–(144d). The four-vector field A^ν is unique if one can show, as in Theorem V, that

$$W^\nu\equiv A^\nu - B^\nu=0. \tag{145}$$

Equations (144a)–(144d) then lead, as before with (97a)–(97d), to results analogous to (99)–(102), which when substituted into identity (138) yields the result

$$W^\mu(x)=\mu^2\int_{V_4'} W^\mu(x')G_M(x,x')d^4x'. \tag{146}$$

Then, if one applies the operator $(\Box-\mu^2)$, which acts only on the unprimed coordinates x^ν, to (146) while using (139), one obtains

$$\begin{aligned}(\Box-\mu^2)W^\mu(x)&=\mu^2\int_{V_4'} W^\mu(x')(\Box-\mu^2)G_M(x,x')d^4x'\\ &=-\mu^2\int_{V_4'} W^\mu(x')\,\delta^{(4)}(x-x')d^4x' = -\mu^2 W^\mu(x).\end{aligned} \tag{147}$$

The μ^2 terms in (147) cancel and so $W^\mu(x)$ satisfies the homogeneous wave equation

$$\Box W^{\mu}(x)=0. \tag{148}$$

Equation (148) also follows from a direct expansion of the d'Alembertian operator using (19) as follows:

$$\Box W^{\mu}=\partial_{\nu}(\partial^{\nu}W^{\mu}-\partial^{\mu}W^{\nu})+\partial^{\mu}(\partial_{\nu}W^{\nu})=0, \tag{149}$$

where the first bracketed term of (149) vanishes because the four-curl vanishes by a result analogous to (101) and the second bracketed term of (146) vanishes because the four-divergence vanishes by a result analogous to (99). It should be clear from the result (149) that it is not necessary to solve the Cauchy problem for the wave equation (148) to find $W^{\mu}(x)$ since each bracketed term in (149) separately vanishes. In fact, each of the conditions (143a)–(143d) are satisfied for $W^{\mu}(x)$ throughout $\overline{V}_4$, and so by Theorem XI $W^{\mu}(x)=0$ throughout $\overline{V}_4$. Thus, only the trivial solution $W^{\mu}(x)=0$ of (148) satisfies the boundary conditions and therefore, [via (145)], the four-vector field $A^{\mu}(x)=B^{\mu}(x)$ everywhere in the four-volume V_4 and on its bounding three-surface Σ. This proves that a four-vector field $A^{\mu}(x)$ in a Minkowski $3+1$ space–time satisfying identity (138) is uniquely specified by specifying the real scalar constant μ and the relations (144a)–(144d), thus proving Theorem XII. ♣

It is tempting to interpret Theorem XII as a uniqueness theorem for massive four-vector fields, i.e., those which satisfy the inhomogeneous massive four-*vector* wave equation

$$(\Box-\mu^2)F^{\mu}(x)=-j^{\mu}(x). \tag{150}$$

However, the Green's function $G_M(x,x')$, e.g., (135), which was used in identity (138) followed from an inhomogeneous massive *scalar* wave equation which is a much simpler problem. That is, equations of the form (150) would in general use a solution technique which involves a two-point second rank tensor Green's function.[22,23] Nevertheless, a close inspection of the proofs of identity (138) and Theorem XII reveals that no limitations are imposed on the four-vector field $A^{\mu}(x)$ other than that its components must be sufficiently smooth. The four-vector field $A^{\mu}(x)$ could therefore be either a massless or a massive four-vector field. Naturally, identity (138) appears to be oriented toward application to massive four-vector fields $A^{\mu}(x)$ due to its incorporation of a mass damping factor.

V. CONCLUSION

In conclusion, the three-space Helmholtz identity and its associated uniqueness theorems, which focus on the curl and divergence of a vector field, provide insight into irrotational and solenoidal fields. The extension of the Helmholtz identity and associated uniqueness theorems to Euclidean and Minkowski four-spaces presented in this article demonstrates that the curl and divergence of a three-vector field generalize into the four-curl and four-divergence of a four-vector field, and that irrotational and solenoidal three-vector fields naturally generalize into four-irrotational and four-solenoidal four-vector fields, respectively.

Now, a four-solenoidal field is essentially a four-vector field in the Lorentz gauge (sometimes referred to as a *relativistic transverse gauge*), with zero four-divergence, as for example in the case of the electromagnetic field. The author is currently investigating the associated concept of a four-irrotational four-vector field. This leads to the development of what the author shall call a ''*relativistic longitudinal gauge*'' where the Maxwell field tensor itself is set to zero, while the four-divergence of the four-vector field can in general be nonzero. Consider the case of a so-called ''pure gauge'' field which arises in connection with the Meissner effect deep in a superconductor, where the magnetic field is required to vanish (cf. Ref. 24). Interestingly, pure gauge fields which are defined by the relation

$$A^{\mu}\equiv\partial^{\mu}\Lambda, \tag{151}$$

satisfy a ''*relativistic longitudinal gauge condition*'' as defined by the vanishing of its Maxwell field tensor:

$$F^{\mu\nu}=\partial^{\mu}A^{\nu}-\partial^{\nu}A^{\mu}=\partial^{\mu}\partial^{\nu}\Lambda-\partial^{\nu}\partial^{\mu}\Lambda=0, \tag{152}$$

and further are four-irrotational since the four-curl of A^{μ}, i.e., $F^{\mu\nu}$, is zero by (152). If they also have nonzero four-divergence throughout a region V_4 where (152) holds, i.e.,

$$\partial_{\mu}A^{\mu}(x)=\partial_{\mu}\partial^{\mu}\Lambda(x)\neq 0 \quad \forall x\in V_4, \tag{153}$$

they would provide an example of a four-vector field in the relativistic longitudinal gauge. Indeed, if a pure gauge field, defined for example over an unbounded space–time region, satisfied $\partial_{\mu}A^{\mu}=0$, then by Theorem XI one would have the field A^{μ} vanishing everywhere! In a finite space–time region, on the other hand, one could still possibly have $\partial_{\mu}A^{\mu}=0$, $F^{\mu\nu}=0$, and $A^{\mu}\neq 0$ all holding true if either (143c) and/or (143d) were nonzero on the three-surface Σ bounding the four-volume V_4.

ACKNOWLEDGMENTS

The style and presentation of this article has benefited from detailed discussions with the author's thesis adviser Associate Professor John Corbett. The author is deeply indebted to Professor Corbett for allowing him to freely develop his own ideas while at the same time making him aware of their mathematical significance and limitations. This work is a partial reporting of the results of the author's Ph.D. thesis. The thesis was submitted on 28 August, 1998 to the Department of Physics at Macquarie University–Sydney, Australia, in partial fulfillment of the requirements for the degree Doctor of Philosophy. Tuition exemption from the Higher Education Contribution Scheme (HECS) was obtained via the Ph.D. by Research program.

[1] W. Hauser, Am. J. Phys. **38**, 80 (1970); D. H. Kobe, *ibid.* **52**, 354 (1984); **54**, 552 (1986); J. A. Heras, *ibid.* **58**, 154 (1990).
[2] C. Møller, *The Theory of Relativity*, corrected 1st ed. (Clarendon, Oxford, 1969) pp. 126, 144–147.
[3] G. Stokes, Trans. Cambridge Philos. Soc. **9**, 1 (1849).
[4] H. L. F. Helmholtz, Crelles J. **55**, 25 (1858).
[5] A. Sommerfeld, *Mechanics of Deformable Bodies*, Lectures on Theoretical Physics Vol. II, English translation of the 2nd German ed. (Academic, New York, 1964), pp. 147–151.
[6] R. E. Collin, *Field Theory of Guided Waves* (McGraw–Hill, New York, 1960), pp. 564–568.
[7] R. Plonsey and R. E. Collin, *Principles and Applications of Electromagnetic Fields* (McGraw–Hill, New York, 1961), pp. 29–36.
[8] G. Arfken, *Mathematical Methods for Physicists*, 3rd ed. (Academic, Orlando, FL, 1985), pp. 78–84.
[9] R. W. P. King, *Fundamental Electromagnetic Theory*, retitled 2nd ed. (Dover, New York, 1963), p. 163.
[10] R. F. Harrington, *Time-Harmonic Electromagnetic Fields* (McGraw–Hill, New York, 1961), p. 450.
[11] C. Itzykson and J. Zuber, *Quantum Field Theory* (McGraw–Hill, New York, 1980), pp. 11–12, 136, 576, 10.
[12] E. C. G. Stueckelberg, Helv. Phys. Acta **11**, 225 (1938); **11**, 299 (1938).
[13] F. Mandl and G. Shaw, *Quantum Field Theory*, revised ed. (Wiley, New York, 1994).
[14] A. Sommerfeld, Ann. Phys. (Leipzig) **33**, 649 (1910).
[15] A. Sommerfeld, *Electrodynamics*, Lectures on Theoretical Physics Vol. III, English translation of the German ed. (Academic, New York, 1964), pp. 245–249.
[16] J. A. Stratton, *Electromagnetic Theory* (McGraw–Hill, New York, 1941), pp. 470–473.
[17] J. T. Cushing, *Applied Analytical Mathematics for Physical Scientists* (Wiley, New York, 1975), pp. 511–516.
[18] J. D. Jackson, *Classical Electrodynamics*, 2nd ed. (Wiley, New York, 1975), pp. 608–612, 38–45, 219–223.
[19] W. Greiner and J. Reinhardt, *Quantum Electrodynamics*, 2nd corrected ed. (Springer, New York, 1996), p. 252.
[20] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw–Hill, New York, 1953).
[21] B. S. DeWitt, in *Relativity, Groups and Topology*, edited by C. DeWitt and B. DeWitt (Blackie, London, 1964), pp. 620, 615–624.
[22] B. S. DeWitt and R. W. Brehme, Ann. Phys. (N.Y.) **9**, 220 (1960).
[23] F. G. Friedlander, *The Wave Equation on a Curved Space-Time* (Cambridge University Press, Cambridge, 1975).
[24] S. Weinberg, *The Quantum Theory of Fields*, Modern Applications Vol. II (Cambridge University Press, Cambridge, 1996), pp. 333–334, 24–27.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Grassmann algebra and fermions at finite temperature

I. C. Charret
Departamento de Ciências Exatas, Universidade Federal de Lavras, Caixa Postal 37, CEP: 37200-000, Lavras, MG, Brazil

E. V. Corrêa Silva
Centro Brasileiro de Pesquisas Físicas, R. Dr. Xavier Sigaud no. 150, CEP: 22290-180, Rio de Janeiro, RJ, Brazil

S. M. de Souza
Departamento de Ciências Exatas, Universidade Federal de Lavras, Caixa Postal 37, CEP: 37200-000, Lavras, MG, Brazil

O. Rojas Santos and M. T. Thomaz
Instituto de Física, Universidade Federal Fluminense, Av. Gal. Milton Tavares de Souza s/no., Niterói, CEP: 24210-340, Niterói, RJ, Brazil

(Received 25 September 1998; accepted for publication 27 April 1999)

For any d-dimensional self-interacting fermionic model, all coefficients in the high-temperature expansion of its grand canonical partition function can be put in terms of multivariable Grassmann integrals. A new approach to calculate such coefficients, based on direct exploitation of the Grassmannian nature of fermionic operators, is presented. We apply the method to the soluble Hatsugai–Kohmoto model, reobtaining well-known results.
[S0022-2488(99)00310-2]

I. INTRODUCTION

A quantum system at thermal equilibrium can be completely described provided that one knows its grand canonical partition function, which can be expressed as a path integral. For bosonic systems, an advantageous feature of the path integral approach is that of employing commuting functions instead of noncommuting operators. For fermionic systems, however, such an advantage is not obvious to hold, as the integration variables are also noncommuting.

In 1980, Kubo[1] used the path integral approach to calculate the grand canonical partition function of the Hubbard model, using the strong coupling limit and performing a perturbative expansion in the hopping constant (t). Even though his result is valid for any temperature, one does not have the exact coefficient of $\beta(\beta=1/kT)$ of the high-temperature expansion of the partition function. Since then, improvements on the calculation of the high-temperature expansion up to order $(\beta t)^9$ for the Hubbard model in two and three dimensions have been reported in the literature.[2]

Recently, Grandati *et al.*[3] presented a method to calculate the grand canonical partition function of self-interacting fermions by writing that function on a lattice and using the properties of the Grassmann algebra to calculate its expansion in powers of the coupling constant. They calculated the first two terms for the bi-dimensional chiral Gross–Neveu model, obtaining an analytical result; however, their approach is model dependent. More recently, Creutz[4] used a numerical algorithm to calculate the generating functional of a fermionic model, rewritten on a lattice. He applied his algorithm to a unidimensional fermionic system involving a thousand Grassmannian variables. He pointed out that this approach does not have the sign problems that generally hamper the application of the Monte Carlo method to fermionic models.

We do not write the grand canonical partition function of a self-interacting fermionic model on a lattice; instead, we present a new method to obtain the coefficients of its high-temperature expansion in d dimensions, where $d\geqslant 1$. However, for the expansion in β, this method does *not* involve any other perturbative expansion (such as, say, in the coupling constant of the model). In Sec. II we present the method, an extension to the one used to calculate the grand canonical partition function of the anharmonic fermionic oscillator,[5] a quantum model in $d=0$ space dimension. In Ref. 6, the properties of the Grassmann algebra were used to calculate the moments of Grassmannian Gaussian integrals. This general result, together with the diagonalization of matrices $\mathbf{A}^{\sigma\sigma}(\sigma=\uparrow,\downarrow)$—matrices that appear when the trace of any fermionic operator is expressed in terms of a multivariable Grassmann integral [see Eq. (9)]—allows us to develop a general approach to obtain analytical expressions for the coefficients of the high-temperature expansion of the grand canonical partition function of any self-interacting fermionic model in d dimensions, even in the thermodynamical limit. In Sec. III we apply the method to the Hatsugai–Kohmoto model, a simple toy model that was used by Hatsugai and Kohmoto to explain the metal-insulator transition. The solution of this model is very simple, and does not require all of the features developed in Sec. II. In Sec. IV we present our conclusions and future applications of the present approach. In the Appendix, the diagonalization of the matrices $\mathbf{A}^{\sigma\sigma}$, for arbitrary lattice dimension and arbitrary number of points in the lattice, is described.

II. EXPANSION IN THE HIGH-TEMPERATURE LIMIT AND THE GRASSMANN MULTIVARIABLE INTEGRALS

The grand canonical partition function of any quantum system in the high-temperature limit can be expanded in terms of β as

$$\mathcal{Z}(\beta;\mu)=\mathrm{Tr}\,(e^{-\beta\mathbf{K}})=\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\mathrm{Tr}[\mathbf{K}^n]\beta^n, \tag{1}$$

where $\mathbf{K}$ is given by

$$\mathbf{K}=\mathbf{H}-\mu\mathbf{N}, \tag{2}$$

$\mathbf{H}$ is the Hamiltonian of the system, μ is the chemical potential, and $\mathbf{N}$ is the total number of particles operator.

The fermionic creation ($\mathbf{a}_i^\dagger$) and destruction ($\mathbf{a}_j$) operators can be mapped into generators of the Grassmann algebra $\{\overline{\eta_i},\eta_j\}$ as follows:[7–9]

$$\mathbf{a}_i^\dagger\rightarrow\overline{\eta_i} \quad \text{and} \quad \mathbf{a}_j\rightarrow\frac{\partial}{\partial\overline{\eta_j}}, \tag{3}$$

where $i,j=1,2,...,\mathcal{N}$. The generators of this Grassmann algebra of dimension $2^{2\mathcal{N}}$, written explicitly as $\{\overline{\eta_1},\ldots,\overline{\eta_\mathcal{N}};\eta_1,\ldots,\eta_\mathcal{N}\}$, satisfy the following anticommutation relations:

$$\{\eta_i,\eta_j\}=0, \quad \{\overline{\eta_i},\overline{\eta_j}\}=0, \quad \text{and}\ \{\overline{\eta_i},\eta_j\}=0. \tag{4}$$

The trace of any normal-ordered fermionic operator $\mathbf{O}$ is[7]

$$\mathrm{Tr}[\mathbf{O}]=\int\prod_{i=1}^{\mathcal{N}}d\eta_i d\overline{\eta_i}\,\mathcal{O}^{ⓝ}(\overline{\eta},\eta)e^{\Sigma_{j=1}^{\mathcal{N}}2\overline{\eta_j}\eta_j}, \tag{5}$$

where we use the shorthand notation, $\overline{\eta}\equiv\{\overline{\eta_1},\ldots,\overline{\eta_\mathcal{N}}\}$ and $\eta\equiv\{\eta_1,\ldots,\eta_\mathcal{N}\}$, and $\mathcal{O}^{ⓝ}(\overline{\eta},\eta)$ is the kernel of the fermionic operator $\mathbf{O}$ in the normal order. (By ''normal-ordered operator'' we mean an operator in which all destruction operators are to placed to the right of all creation operators.) Naively, it can be said that the Grassmannian function $\mathcal{O}^{ⓝ}(\overline{\eta},\eta)$ is obtained by replacing $\mathbf{a}_i^\dagger\rightarrow\overline{\eta_i}$ and $\mathbf{a}_i\rightarrow\eta_i$ in operator $\mathbf{O}$.[5,7]

Let us consider from now on the case where the creation and destruction operators are characterized by the indices $(\vec{\ell};\sigma)$, where $\vec{\ell}$ is a d-dimensional lattice vector ($d=1,2,3,...$) and σ is the spin component. The components of vector $\vec{\ell}$ need not be orthogonal. This lattice vector could equally represent either the space vector $\vec{\mathbf{x}}$ or the momentum vector $\vec{\mathbf{k}}$. If the fermionic operator $\mathbf{O}$ is a product of n normal-ordered fermionic operators $\mathbf{Q}$, we have[5]

$$\mathrm{Tr}[\mathbf{Q}^n]=\int \prod_{\vec{\ell}} \prod_{\sigma=\pm 1} \prod_{\alpha=0}^{n-1} d\eta_\sigma(\vec{\ell};\alpha)d\overline{\eta}_\sigma(\vec{\ell};\alpha)e^{\Sigma_{\vec{\ell}}\Sigma_{\nu=0}^{n-1}\overline{\eta}_\sigma(\vec{\ell};\nu)[\eta_\sigma(\vec{\ell};\nu)-\eta_\sigma(\vec{\ell};\nu+1)]}$$
$$\times \mathcal{Q}^{ⓝ}(\overline{\eta}_\sigma(\vec{\ell};0),\eta_\sigma(\vec{\ell};0))\mathcal{Q}^{ⓝ}(\overline{\eta}_\sigma(\vec{\ell};1),\eta_\sigma(\vec{\ell};1))\times\cdots$$
$$\times \mathcal{Q}^{ⓝ}(\overline{\eta}_\sigma(\vec{\ell};n-1),\quad \eta_\sigma(\vec{\ell};n-1)), \tag{6}$$

where we define $\sigma=\uparrow\equiv+1$ and $\sigma=\downarrow\equiv-1$. The Grassmann variables in Eq. (6) satisfy the boundary conditions

$$\eta_\sigma(\vec{\ell};n)=-\eta_\sigma(\vec{\ell};0) \quad \text{and} \quad \eta_\sigma(\vec{\ell};\nu)=0, \quad \text{for} \quad \nu>n, \tag{7}$$

with $\sigma=\pm 1$ and $\vec{\ell}$ stands for any vector on the lattice. Equation (6) is still valid for a product of n-ordered operators, not necessarily equal.

Relation (6) is used to write the terms of the expansion of the grand canonical partition function in the high-temperature limit as multivariable Grassmann integrals. For a d-dimensional fermionic model, the coefficients of the expansion $\mathcal{Z}(\beta,\mu)$ in Eq. (1) become

$$\mathrm{Tr}[\mathbf{K}^n]=\int \prod_{\vec{\ell}} \prod_{\sigma=\pm 1} \prod_{\alpha=0}^{n-1} d\eta_\sigma(\vec{\ell};\alpha)d\overline{\eta}_\sigma(\vec{\ell};\alpha)e^{\Sigma_{\vec{\ell}}\Sigma_{\sigma=\pm 1}\Sigma_{\nu=0}^{n-1}\overline{\eta}_\sigma(\vec{\ell};\nu)[\eta_\sigma(\vec{\ell};\nu)-\eta_\sigma(\vec{\ell};\nu+1)]}$$
$$\times \mathcal{K}^{ⓝ}(\overline{\eta}_\sigma(\vec{\ell};0),\eta_\sigma(\vec{\ell};0))\mathcal{K}^{ⓝ}(\overline{\eta}_\sigma(\vec{\ell};1),\eta_\sigma(\vec{\ell};1))\times\cdots$$
$$\times \mathcal{K}^{ⓝ}(\overline{\eta}_\sigma(\vec{\ell};n-1),\quad \eta_\sigma(\vec{\ell};n-1)). \tag{8}$$

The boundary conditions (7) still hold for the generators $\eta_\sigma(\vec{\ell};\nu)$.

It is much easier to handle generators with one index. Then we map the generators $\eta_\sigma(\vec{\ell},\nu)$ and $\overline{\eta}_\sigma(\vec{\ell},\nu)$ into single-indexed anti-commuting variables. The sum in the exponential on the rhs of Eq. (8) can be written as

$$\sum_{\vec{\ell}} \sum_{\sigma=\pm 1} \sum_{\nu=0}^{n-1} \overline{\eta}_\sigma(\vec{\ell};\nu)[\eta_\sigma(\vec{\ell};\nu)-\eta_\sigma(\vec{\ell};\nu+1)]\equiv\sum_{I,J=1}^{2nN^d} \overline{\eta}_I A_{IJ}\eta_J. \tag{9}$$

Note that the argument of the exponential on the rhs of Eq. (8) is diagonal in the indices $\vec{\ell}$ and σ. Having the components of the column vector η_J (or the line vector $\overline{\eta}$) grouped according to the values of σ, and then each subset ordered according to ν, and finally each subsubset ordered according to $\vec{\ell}$, the matrix $\mathbf{A}$ will have the block-structure

$$\mathbf{A}=\begin{pmatrix} \mathbf{A}^{\uparrow\uparrow} & \mathbb{O} \\ \mathbb{O} & \mathbf{A}^{\downarrow\downarrow} \end{pmatrix}, \tag{10}$$

whose entries are matrices of dimension $nN^d\times nN^d$ and d is the dimension of the vector $\vec{\ell}$. The indices I,J are such that $I,J=1,2,...,2nN^d$, where N^d is the number of points in the lattice. The matrices $\mathbf{A}^{\uparrow\uparrow}$ and $\mathbf{A}^{\downarrow\downarrow}$ are identical. Taking into account the anti-periodic condition in temperature (7) in Eq. (9), the matrices $\mathbf{A}^{\sigma\sigma}$, where $\sigma=\uparrow,\downarrow$, are found to have the following block-structure:

$$\mathbf{A}^{\uparrow\uparrow}=\mathbf{A}^{\downarrow\downarrow}=\begin{pmatrix} \mathbb{1}_{N^d\times N^d} & -\mathbb{1}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \cdots & \mathbb{0}_{N^d\times N^d} \\ \mathbb{0}_{N^d\times N^d} & \mathbb{1}_{N^d\times N^d} & -\mathbb{1}_{N^d\times N^d} & \cdots & \mathbb{0}_{N^d\times N^d} \\ \vdots & & & & \vdots \\ \mathbb{1}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \cdots & \mathbb{1}_{N^d\times N^d} \end{pmatrix}. \tag{11}$$

The symbols $\mathbb{1}_{N^d\times N^d}$ and $\mathbb{0}_{N^d\times N^d}$ stand for the identity and null matrices of dimension $N^d\times N^d$, respectively. Any lattice vector $\vec{\ell}$ can be written as

$$\vec{\ell}=\ell_1\,\vec{\mathbf{u}}_1+\ell_2\,\vec{\mathbf{u}}_2+\cdots+\ell_d\,\vec{\mathbf{u}}_d, \tag{12}$$

where $\ell_i=1,2,...,N(i=1,2,...,d)$, and N is the number of points in the lattice in the direction of the d-dimensional basis vector $\vec{\mathbf{u}}_i$. A particular basis, for which $\mathbf{A}^{\sigma\sigma}$ has the block-form shown in (11), yields the following mapping:

$$\eta_\sigma(\vec{\ell};\nu)\rightarrow\eta_{[[(1-\sigma)/2]n+\nu]N^d+\ell_1+(\ell_2-1)N+\cdots+(\ell_d-1)N^{(d-1)}}. \tag{13}$$

The generators $\overline{\eta}_\sigma(\vec{\ell};\nu)$ have an analogous mapping. With the newly indexed generators, the expression of $\mathrm{Tr}[\mathbf{K}^n]$ [Eq. (8)] becomes

$$\mathrm{Tr}[\mathbf{K}^n]=\int\prod_{I=1}^{2nN^d}d\eta_I\,d\overline{\eta}_I\;e^{\sum_{I,J=1}^{2nN^d}\overline{\eta}_I A_{IJ}\eta_J}\times\mathcal{K}^{\circledR}(\overline{\eta},\eta;\nu=0)\mathcal{K}^{\circledR}(\overline{\eta},\eta;\nu=1)\cdots\mathcal{K}^{\circledR}(\overline{\eta},\eta;\nu=n-1). \tag{14}$$

Note that expression (14), up to the constant $-1/n!$, is the coefficient at order β^n of the expansion in the high-temperature limit of the grand canonical partition function for any self-interacting fermionic model. The specific model to be studied is represented by the Grassmannian function $\mathcal{K}^{\circledR}$, but the matrix $\mathbf{A}$ is the same for all fermionic models. Once the submatrices $\mathbf{A}^{\uparrow\downarrow}$ and $\mathbf{A}^{\downarrow\uparrow}$ are null, the multivariable integral (14) is equal to the product of the contributions coming from the sectors: $\sigma\sigma=\uparrow\uparrow$ and $\sigma\sigma=\downarrow\downarrow$ separately. The Grassmann functions $\mathcal{K}^{\circledR}$ are polynomials in the generators of the algebra. Therefore, the rhs of Eq. (14) are moments of the multivariable Grassmann Gaussian integrals. In Ref. 6 it is shown that these integrals can be written as cofactors of the matrix $\mathbf{A}$.

The integrals in Eq. (14), for sector $\sigma\sigma=\uparrow\uparrow$, have the form

$$M(L,K)=\int\prod_{i=1}^{nN^d}d\eta_i d\overline{\eta}_i\;\overline{\eta}_{l_1}\eta_{k_1}\cdots\overline{\eta}_{l_m}\eta_{k_m}e^{\sum_{i,j=1}^{nN^d}\overline{\eta}_i A_{ij}^{\uparrow\uparrow}\eta_J}, \tag{15}$$

with $L=\{l_1,\ldots,l_m\}$ and $K=\{k_1,\ldots,k_m\}$. The products $\overline{\eta}\eta$ are ordered in such a way that $l_1<l_2<\cdots<l_m$ and $k_1<k_2<\cdots<k_m$. From Ref. 6, the result of this type of integrals is equal to

$$M(L,K)=(-1)^{(l_1+l_2+\cdots+l_m)+(k_1+k_2+\cdots+k_m)}A(L,K), \tag{16}$$

where $A(L,K)$ is the determinant of the matrix obtained from matrix $\mathbf{A}^{\uparrow\uparrow}$ by deleting the lines $\{l_1,\ldots,l_m\}$ and the columns $\{k_1,\ldots,k_m\}$. $M(L,K)$ is a cofactor of matrix $\mathbf{A}^{\uparrow\uparrow}$. The Grassmann integrals to be calculated in sector $\downarrow\downarrow$ are the same type as Eq. (15). Evaluating determinants of nondiagonal matrices of dimension $nN^d\times nN^d$ is still a hard task, even if we have restricted ourselves to multivariable integrals of a fixed sector $\sigma\sigma$. Calculating such determinants is a suitable task for computers, and it obviously depends on hardware and software resources. Fixing n, for instance, there is an upper practical limit for N, so that the calculation of determinants is feasible. One possibility for evaluating Eq. (14) is that of assigning different values for N and, from the results obtained, trying to extrapolate for an arbitrary value of N. If we are lucky, some recursion expression for Eq. (14) for all N could be recognized.

Our approach to calculate the integral (15) is, for fixed n and arbitrary N, to explore the block-structure of matrices $\mathbf{A}^{\sigma\sigma}$, $\sigma=\uparrow$ and $\sigma=\downarrow$, diagonalizing it through a similarity transformation

$$\mathbf{P}^{-1}\mathbf{A}^{\sigma\sigma}\mathbf{P}=\mathbf{D}, \tag{17}$$

where the matrix $\mathbf{D}$ is

$$\mathbf{D}=\begin{pmatrix} \lambda_1 \mathbb{1}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \cdots & \mathbb{0}_{N^d\times N^d} \\ \mathbb{0}_{N^d\times N^d} & \lambda_2 \mathbb{1}_{N^d\times N^d} & \cdots & \mathbb{0}_{N^d\times N^d} \\ \vdots & & & \vdots \\ \mathbb{0}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \cdots & \lambda_n \mathbb{1}_{N^d\times N^d} \end{pmatrix}, \tag{18}$$

and λ_i, $i=1,2,...,n$, are the eigenvalues of matrices $\mathbf{A}^{\sigma\sigma}$, $\sigma=\uparrow,\downarrow$, and calculate the cofactors of the matrix $\mathbf{D}$. The jth column of matrix $\mathbf{P}$ is the eigenvector of $\mathbf{A}^{\sigma\sigma}$ associated to the eigenvalue λ_j. Each eigenvalue of matrix $\mathbf{A}^{\sigma\sigma}$ has degeneracy N^d. The matrices are not Hermitian, thus some eigenvalues are complex. In the following, we will be working on the $\sigma\sigma=\uparrow\uparrow$ sector; however, the results for the $\sigma\sigma=\downarrow\downarrow$ sector are analogous since $\mathbf{A}^{\uparrow\uparrow}=\mathbf{A}^{\downarrow\downarrow}$.

We will apply the following transformation of variables,

$$\eta'=\mathbf{P}^{-1}\eta \quad \text{and} \quad \overline{\eta'}=\overline{\eta}\mathbf{P}, \tag{19}$$

where $\eta'\equiv\{\eta'_1,\dots,\eta'_{nN^d}\}$ and $\overline{\eta'}\equiv\{\overline{\eta'_1},\dots,\overline{\eta'_{nN^d}}\}$. The Jacobian of the transformation (19) is equal to one.

Due to the fact that $\mathbf{A}^{\uparrow\uparrow}$ is a block matrix, the matrix $\mathbf{P}$ also has a block structure. This fact implies that transformations (19) do not mix up lattice indices.

In a schematic way, the integrals $M(L,K)$ [Eq. (15)] become

$$M(L,K)=\int \prod_{i=1}^{nN^d} d\eta_i \, d\overline{\eta_i}\, (\overline{\eta}\mathbf{P}^{-1})_{l_1}(\mathbf{P}\eta)_{k_1}\cdots(\overline{\eta}\mathbf{P}^{-1})_{l_m}(\mathbf{P}\eta)_{k_m} e^{\Sigma_{i,j=1}^{nN^d}\overline{\eta}_i D_{ij}\eta_j}, \tag{20}$$

where D_{ij} are the entries of the diagonal matrix $\mathbf{D}$. The expression $M(L,K)$ fits into the form of Eq. (15), and hence corresponds to some cofactor of the diagonalized matrix $\mathbf{D}$ [Eq. (16)]. It is very simple to calculate these cofactors, and the matrix $\mathbf{P}$ is the same for any self-interacting fermionic model.

In the Appendix we present the derivation of the eigenvalues and eigenvectors of matrix $\mathbf{D}$ for arbitrary values of n and N. From Eqs. (A25) and (A26), for arbitrary value of n, we have that

$$p_{\nu\nu'}^{(n)}=\frac{1}{\sqrt{n}} e^{(i\pi/n)(2\nu'+1)(\nu+1)}, \tag{21}$$

and

$$q_{\nu'\nu}^{(n)}=\frac{1}{\sqrt{n}} e^{-(i\pi/n)(2\nu'+1)(\nu+1)}, \tag{22}$$

with ν, $\nu'=0,1,...,n-1$, and

$$\mathbf{P}=\begin{pmatrix} p_{00}^{(n)}\mathbb{1}_{N^d\times N^d} & \cdots & p_{0,n-1}^{(n)}\mathbb{1}_{N^d\times N^d} \\ \vdots & & \vdots \\ p_{n-1,0}^{(n)}\mathbb{1}_{N^d\times N^d} & \cdots & p_{n-1,n-1}^{(n)}\mathbb{1}_{N^d\times N^d} \end{pmatrix} \tag{23}$$

and

$$\mathbf{P}^{-1}=\begin{pmatrix} q_{00}^{(n)}\mathbb{1}_{N^d\times N^d} & \cdots & q_{0,n-1}^{(n)}\mathbb{1}_{N^d\times N^d} \\ \vdots & & \vdots \\ q_{n-1,0}^{(n)}\mathbb{1}_{N^d\times N^d} & \cdots & q_{n-1,n-1}^{(n)}\mathbb{1}_{N^d\times N^d} \end{pmatrix}. \tag{24}$$

The diagonal elements of matrix $\mathbf{D}$ are

$$\lambda_\nu^{(n)}=1-e^{(i\pi/n)(2\nu+1)}, \quad \nu=0,1,..., n-1, \tag{25}$$

where the eigenvalues are N^d-fold degenerated, N^d being the number of lattice sites. Due to lattice translation symmetry, we should note that the elements $p_{\nu\nu'}^{(n)}$ and $q_{\nu\nu'}^{(n)}$ do not carry any lattice site index.

This is a general approach, and it can be applied to any self-interacting fermionic model. The important point here is that the relations (21)–(25) are valid for any self-interacting fermionic model with space translation symmetry.

III. APPLICATION TO HATSUGAI–KOHMOTO MODEL

The calculation of an exactly soluble model is a nice way to test a new approach. Hatsugai and Kohmoto[10] proposed a toy model (HK model) that shares the atomic and band limits of the Hubbard model.[11] Using the Green's function and path integral approaches, Nogueira and Anda[12] established the equivalence of this model (with unrestricted hopping) and the Hubbard model (with infinite-range hopping).

In this section we derive the grand canonical partition function of the HK model using the results presented in Sec. II. The Hamiltonian of the HK model in momentum space is[12]

$$\mathbf{H}=\sum_{\vec{k}}\sum_{\sigma=\uparrow,\downarrow}\varepsilon(\vec{k})\mathbf{n}_\sigma(\vec{k})+U\sum_{\vec{k}}\mathbf{n}_\uparrow(\vec{k})\mathbf{n}_\downarrow(\vec{k})\equiv\sum_{\vec{k}}\mathbf{H}(\vec{k}), \tag{26}$$

where $\mathbf{n}_\sigma(\vec{k})\equiv\mathbf{a}_\sigma^\dagger(\vec{k})\mathbf{a}_\sigma(\vec{k})$ and $\mathbf{a}_\sigma^\dagger(\vec{k})[\mathbf{a}_\sigma(\vec{k})]$ is the creation (destruction) operator of an electron with momentum $\vec{k}$ and spin σ. The function $\varepsilon(\vec{k})=-2t\Sigma_{i=1}^3\cos k_i$, $\vec{k}=(k_1,k_2,k_3)$, corresponds to the nearest hopping of the electrons in the dual-space lattice. U is the strength of the repulsion between electrons with the same momentum $\vec{k}$ but opposite spin conponents.

From Eqs. (1) and (26), the grand canonical partition function of the HK model is

$$\mathcal{Z}(\beta;\mu)=\mathrm{Tr}\left[\prod_{\vec{k}} e^{-\beta\mathbf{K}(\vec{k})}\right]=\prod_{\vec{k}}[\mathrm{Tr}_{\vec{k}}\, e^{-\beta\mathbf{K}(\vec{k})}]. \tag{27}$$

We have

$$\mathbf{K}(\vec{k})=\sum_{\sigma=\uparrow,\downarrow}\Delta(\vec{k})\mathbf{n}_\sigma(\vec{k})+U\mathbf{n}_\uparrow(\vec{k})\mathbf{n}_\downarrow(\vec{k}), \tag{28}$$

where we define $\Delta(\vec{k})\equiv\varepsilon(\vec{k})-\mu$, and μ is the chemical potential. In Eq. (27), the symbol $\mathrm{Tr}_{\vec{k}}$ stands for the trace for a fixed vector $\vec{k}$, whereas Tr represents the trace for all $\vec{k}$'s.

The high-temperature expansion for the grand canonical partition function $\mathcal{Z}(\beta;\mu)$ is

$$\mathrm{Tr}_{\vec{k}}[e^{-\beta\mathbf{K}(\vec{k})}]=\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\beta^n\,\mathrm{Tr}_{\vec{k}}[\mathbf{K}^n(\vec{k})]. \tag{29}$$

Since all the operators on the rhs of Eq. (28) commute, we can apply the Newton's multinomial expression to write $\mathrm{Tr}_{\vec{k}}[\mathbf{K}^n(\vec{k})]$ as

$$\mathrm{Tr}_{\vec{k}}[\mathbf{K}^n(\vec{k})]=\sum_{n_1,n_2,n_3=0}^{n}{}'\ \frac{n!}{n_1!n_2!n_3!}\Delta^{n_1+n_2}(\vec{k})U^{n_3}\,\mathrm{Tr}_{\vec{k}}[\mathbf{n}_\uparrow^{n_1+n_3}(\vec{k})\mathbf{n}_\downarrow^{n_2+n_3}(\vec{k})]. \tag{30}$$

The symbol Σ' means that the summation indices satisfy the condition $n_1+n_2+n_3=n$.

Let l_1, l_2, and l_3 be the integers that determine the lattice vector $\vec{k}$. The mapping (13) takes the index $\vec{k}$ into the index $L\equiv l_1+(l_2-1)N+(l_3-1)N^2$, where N is the number of points in the momentum lattice in each direction. In the sum on the rhs of Eq. (30), we calculate the trace for a fixed $\vec{k}$, which means that in Eq. (6) we take a single point in the momentum lattice $(N=1)$. Then,

$$\mathrm{Tr}_{\vec{k}}[\mathbf{n}_\uparrow^{n_1+n_3}(\vec{k})\mathbf{n}_\downarrow^{n_2+n_3}(\vec{k})]=\mathcal{I}^{\uparrow\uparrow}_{n_1,n_3}\times\mathcal{I}^{\downarrow\downarrow}_{n_2,n_3}, \tag{31}$$

where

$$\begin{aligned}\mathcal{I}^{\uparrow\uparrow}_{n_1,n_3}\equiv\int\prod_{I=1}^{n}\ d\eta_I(L)d\overline{\eta_I}(L)\ e^{\sum_{I,J=1}^{n}\overline{\eta}_I(L)A^{\uparrow\uparrow}_{IJ}\eta_J(L)}\\ \times\overline{\eta}_0(L)\,\eta_0(L)\cdots\overline{\eta}_{n_1-1}(L)\,\eta_{n_1-1}(L)\,\overline{\eta}_{n_1+n_2}(L)\,\eta_{n_1+n_2}(L)\cdots\overline{\eta}_{n-1}(L)\,\eta_{n-1}(L),\end{aligned} \tag{32}$$

and

$$\begin{aligned}\mathcal{I}^{\downarrow\downarrow}_{n_2,n_3}\equiv\int\prod_{J=n+1}^{2n}\ d\eta_J(L)d\overline{\eta_J}(L)\ e^{\sum_{I,J=n+1}^{2n}\overline{\eta}_I(L)A^{\downarrow\downarrow}_{IJ}\eta_J(L)}\\ \times\overline{\eta}_{n+n_1}(L)\,\eta_{n+n_1}(L)\cdots\overline{\eta}_{n+n_1+n_2-1}(L)\,\eta_{n+n_1+n_2-1}(L)\\ \times\overline{\eta}_{n+n_1+n_2}(L)\,\eta_{n+n_1+n_2}(L)\cdots\overline{\eta}_{2n-1}(L)\,\eta_{2n-1}(L).\end{aligned} \tag{33}$$

The matrices $\mathbf{A}^{\sigma\sigma}$, $\sigma=\uparrow,\downarrow$, are given by Eq. (11) with $N=1$. According to Eqs. (15) and (16), the presence of $\overline{\eta}$'s (and η's) in the integrand on the rhs of Eqs. (32) and (33) allows one to evaluate the integrals as the determinants of matrices obtained after deletion of lines (and columns) of the matrices $\mathbf{A}^{\sigma\sigma}$, $\sigma=\uparrow,\downarrow$. For this particular model, it turns out easier to apply Eq. (16) directly, rather than using the similarity transformation (17), since the lattice is unidimensional. We should mention that for $N=1$, we recover the case of the anharmonic fermionic oscillator, which has been considered in a previous work.[5] Now we discuss the values of $\mathcal{I}^{\uparrow\uparrow}_{n_1,n_3}$ [Eq. (32)], in view of the possible values of the indices (n_1,n_2,n_3).

(i) $n_1=n$, $n_2=0$, and $n_3=0$.

In this case the first n lines and the first n columns of matrix $\mathbf{A}^{\uparrow\uparrow}$ are deleted; hence,

$$\mathcal{I}^{\uparrow\uparrow}_{n,0}=1. \tag{34}$$

(ii) $n_1=0$, $n_2=n$, and $n_3=0$.

In this case no lines or columns are deleted in $\mathbf{A}^{\uparrow\uparrow}$; so,

$$\mathcal{I}^{\uparrow\uparrow}_{0,0}=\det(\mathbf{A}^{\uparrow\uparrow})=2. \tag{35}$$

(iii) $n_1=0$, $n_2=0$, and $n_3=n$.

This case is equal to case i and, therefore,

$$\mathcal{I}_{0,n}^{\uparrow\uparrow}=1. \tag{36}$$

(iv) $n_1\neq 0$, $n_2\neq 0$, and $n_3\neq 0$.

In this case, the first n_1 lines and columns are deleted, as well as the last n_3 lines and columns of matrix $\mathbf{A}^{\uparrow\uparrow}$. The triangular matrix thus obtained has its determinant equal to 1, for any value of n. Then,

$$\mathcal{I}_{n_1,n_3}^{\uparrow\uparrow}=1. \tag{37}$$

Equivalent results are valid for $\mathcal{I}_{n_2,n_3}^{\downarrow\downarrow}$.

From the results (34)–(37) and the equivalent results for $\mathcal{I}_{n_2,n_3}^{\downarrow\downarrow}$, we have

$$\mathrm{Tr}_{\vec{k}}[\mathbf{n}_{\uparrow}^{n_1+n_3}(\vec{k})\mathbf{n}_{\downarrow}^{n_2+n_3}(\vec{k})]=(1+\delta_{n_1+n_3,0})(1+\delta_{n_2+n_3,0}), \tag{38}$$

which, substituted in Eq. (30), gives

$$\mathrm{Tr}_{\vec{k}}[\mathbf{K}^n(\vec{k})]=[2\Delta(\vec{k})+U]^n+2\Delta^n(\vec{k}). \tag{39}$$

Returning to Eqs. (27) and (29), we finally get

$$\mathcal{Z}(\beta;\mu)=\prod_{\vec{k}}\,[1+e^{-\beta(2\Delta(\vec{k})+U)}+2e^{-\beta\Delta(\vec{k})}], \tag{40}$$

which gives the same free energy density found in Ref. 12.

IV. CONCLUSIONS

Calculations involving fermionic fields do demand some extra care, in comparison to the manipulation of bosonic fields. For this reason, fermionic models are usually bosonized, in a strategy designed to avoid the ''annoying'' fermionic features. However, moments of Grassmannian multivariable integrals can be easily calculated, as shown in Eqs. (15) and (16). In this paper we have presented a new approach, based on the explicit use of Grassmann algebra properties, to the problem of calculating the coefficients of the high temperature expansion of the grand canonical partition function for any d-dimensional self-interacting fermionic model ($d=1,2,3,...$). We have explored the results (15) and (16) and the possibility of performing the similarity transformation (17) for a system with arbitrary dimension d and arbitrary number of lattice points N^d. It is important to point out that the matrices $\mathbf{A}^{\sigma\sigma}(\sigma=\uparrow,\downarrow)$ are model independent; they are solely related to kinetical aspects of the approach. To simplify the notation, we considered that the number of points in each direction of the lattice is the same, but the results derived are still valid if this is not true. The fact that our results are analytical allows us to obtain the thermodynamical limit for any self-interacting fermionic model.

As a simple example of application of the method (that does not explore all of its features, though), we have considered the Hatsugai–Kohmoto model, which is diagonal in momentum space and had been solved by other approaches. We have derived its grand canonical partition function, and obtained the same free energy density found in the literature.[12]

The most important features of this method appear when the Hamiltonian has noncommuting terms and, consequently, Newton's multinomial expansion does not apply. Equations (15) and (16) and the similarity transformation (17) then become the keystone of our analytical results. That is the case of the Hubbard model;[11] the grand canonical partition function for the unidimensional version of this model is known in integral form.[13] A closed expression for this function was obtained by Takahashi,[13] within certain limits only. By the application of the approach we have presented here, we are currently calculating the coefficients of that partition function, up to order β^5 and for any value of the parameters of the model, as well for any value of the chemical potential. These calculations will soon be submitted for publication.

ACKNOWLEDGMENTS

EVCS thanks CNPq, and ORS thanks CAPES for total financial support. SMdeS and MTT thank CNPq for partial financial support. MTT also thanks FINEP and FAPERJ, and ICC and SMdeS thank FAPEMIG for partial financial support.

APPENDIX: CALCULATION OF EIGENVALUES AND EIGENVECTORS OF MATRIX $\mathbf{A}^{\sigma\sigma}$

This appendix is devoted to calculating the eigenvalues and eigenvectors of the matrix $\mathbf{A}^{\sigma\sigma}$, defined in Eq. (11), as well as determining the matrices $\mathbf{P}$ and $\mathbf{P}^{-1}$ that diagonalize it [see Eq. (17)].

The characteristic equation for $\mathbf{A}^{\sigma\sigma}$ is

$$\det\begin{pmatrix} (1-\lambda)\mathbb{1}_{N^d\times N^d} & -\mathbb{1}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \cdots & \mathbb{0}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} \\ \mathbb{0}_{N^d\times N^d} & (1-\lambda)\mathbb{1}_{N^d\times N^d} & -\mathbb{1}_{N^d\times N^d} & \cdots & \mathbb{0}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbb{0}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \cdots & (1-\lambda)\mathbb{1}_{N^d\times N^d} & -\mathbb{1}_{N^d\times N^d} \\ \mathbb{1}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \mathbb{0}_{N^d\times N^d} & \cdots & \mathbb{0}_{N^d\times N^d} & (1-\lambda)\mathbb{1}_{N^d\times N^d} \end{pmatrix}=0. \tag{A1}$$

Observe that this matrix (of total dimension $nN^d\times nN^d$) consists of a $n\times n$ block matrix, each block having dimension $N^d\times N^d$. Moreover, these blocks are either null matrices $\mathbb{0}_{N^d\times N^d}$ or proportional to the identity matrix $\mathbb{1}_{N^d\times N^d}$.

We will demonstrate a useful property of the determinant of a block matrix in which all blocks are diagonal, such as the previous matrix. Take a block-matrix $\mathbf{M}$ composed of blocks $\mathbf{B}^{[i,j]}$, namely,

$$\mathbf{M}=\begin{pmatrix} \mathbf{B}^{[1,1]} & \mathbf{B}^{[1,2]} & \cdots & \mathbf{B}^{[1,n]} \\ \mathbf{B}^{[2,1]} & \mathbf{B}^{[2,2]} & \cdots & \mathbf{B}^{[2,n]} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}^{[n,1]} & \mathbf{B}^{[n,2]} & \cdots & \mathbf{B}^{[n,n]} \end{pmatrix}, \tag{A2}$$

where each block $\mathbf{B}^{[i,j]}$ is diagonal:

$$(\mathbf{B}^{[i,j]})_{\alpha\beta}=\delta_{\alpha\beta}\mathbf{b}_\alpha^{[i,j]}, \tag{A3}$$

where $\alpha,\beta=1,2,...,N^d$. (No summation over repeated indices is implied.) We define a ''determinant like'' matrix function $\mathbf{F}$ upon the blocks $\mathbf{B}^{[i,j]}$ as

$$\mathbf{F}\equiv\sum_{\theta_1,\theta_2,\dots,\theta_n=1}^{n}\varepsilon_{\theta_1,\theta_2,\dots,\theta_n}\mathbf{B}^{[1,\theta_1]}\mathbf{B}^{[2,\theta_2]}\cdots\mathbf{B}^{[n,\theta_n]}, \tag{A4}$$

where $\varepsilon_{\theta_1,\theta_2,\dots,\theta_n}$ is the Levi-Civita symbol in n-dimension. Obviously, $\mathbf{F}$ and the blocks $\mathbf{B}^{[i,j]}$ have all the same dimensions, $N^d\times N^d$. Using Eq. (A3), we have

$$F_{ab}=\delta_{a,b}\sum_{\theta_1,\theta_2,\dots,\theta_n=1}^{n}\varepsilon_{\theta_1,\theta_2,\dots,\theta_n}(\mathbf{b}_a^{[1,\theta_1]}\mathbf{b}_a^{[2,\theta_2]}\cdots\mathbf{b}_a^{[n-1,\theta_{n-1}]}\mathbf{b}_a^{[n,\theta_n]}). \tag{A5}$$

From Eq. (A5) and the definition of the determinant, we obtain

$$\det \mathbf{F} = \det \boldsymbol{\Gamma}_1 \det \boldsymbol{\Gamma}_2 \cdots \det \boldsymbol{\Gamma}_n, \tag{A6}$$

where we have defined n matrices $\boldsymbol{\Gamma}_p$ of dimension $N^d \times N^d$ as

$$(\boldsymbol{\Gamma}_p)_{uv} \equiv \mathbf{b}_p^{[u,v]}, \tag{A7}$$

so that

$$\det \boldsymbol{\Gamma}_p = \sum_{\omega_1,\omega_2,\ldots,\omega_N=1}^{N} \varepsilon_{\omega_1,\omega_2,\ldots,\omega_n} (\mathbf{b}_p^{[1,\omega_1]} \mathbf{b}_p^{[2,\omega_2]} \cdots \mathbf{b}_p^{[n-1,\omega_{n-1}]} \mathbf{b}_p^{[n,\omega_n]}). \tag{A8}$$

Thus, the evaluation of $\det \mathbf{F}$ is equivalent to the evaluation of the determinant of a block matrix $\boldsymbol{\Gamma}$, defined as

$$\boldsymbol{\Gamma} \equiv \begin{pmatrix} \boldsymbol{\Gamma}_1 & \mathbb{O}_{N^d \times N^d} & \cdots & \mathbb{O}_{N^d \times N^d} \\ \mathbb{O}_{N^d \times N^d} & \boldsymbol{\Gamma}_2 & \cdots & \mathbb{O}_{N^d \times N^d} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{O}_{N^d \times N^d} & \mathbb{O}_{N^d \times N^d} & \cdots & \boldsymbol{\Gamma}_n \end{pmatrix}. \tag{A9}$$

However, $\boldsymbol{\Gamma}$ and $\mathbf{M}$ only differ by an even number of permutations of lines and columns! More specifically, $\mathbf{M}$ can be recovered from $\boldsymbol{\Gamma}$ if we reorder the lines of the latter according to the pattern

$$\begin{aligned}(1,2,\ldots,nN^d) \rightarrow (&1,N^d+1,\, 2N^d+1,\ldots,(n-1)N^d+1,\, 2,N^d+2,2N^d+2,\ldots,(n-1)N^d+2,\ldots,\\ &N^d-1,2N^d-1,3N^d-2,\ldots,nN^d-1,N^d,2N^d,3N^d,\ldots,nN^d),\end{aligned} \tag{A10}$$

i.e., the 1st line is left untouched, the 2nd line is replaced by the $(N^d+1)^{th}$ line, etc., and then have the *columns* of the resulting matrix reordered in the same fashion. (The same result is obtained if we reorder columns before lines.) As the total number of permutations is even, we have

$$\det \mathbf{M} = \det \boldsymbol{\Gamma}. \tag{A11}$$

Combining (A11), (A6), and (A4), we finally obtain

$$\det \mathbf{M} = \det \mathbf{F} = \det \left(\sum_{\theta_1,\theta_2,\ldots,\theta_n=1}^{n} \varepsilon_{\theta_1\theta_2,\ldots,\theta_n} \mathbf{B}^{[1,\theta_1]} \mathbf{B}^{[2,\theta_2]} \cdots \mathbf{B}^{[n,\theta_n]} \right). \tag{A12}$$

In conclusion, if the matrix $\mathbf{M}$ is composed of diagonal blocks $\mathbf{B}^{[i,j]}$, the determinant of $\mathbf{M}$ is equal to the determinant of the matrix $\mathbf{F}$, defined as a ''determinantlike'' function upon the blocks $\mathbf{B}^{[i,j]}$.

Turning our attention back to Eq. (A1), we expand the determinant in terms of ''cofactors'', based on the last ''line'' of blocks:

$$\det \left((-1)^{1+n} \mathbb{1}_{N^d \times N^d} (-\mathbb{1}_{N^d \times N^d})^{n-1} + (-1)^{n+n} (1-\lambda) \mathbb{1}_{N^d \times N^d} (1-\lambda)^{n-1} \mathbb{1}_{N^d \times N^d}^{n-1}\right) = 0, \tag{A13}$$

which yields the characteristic equation

$$(1 + (1-\lambda)^n)^{N^d} = 0. \tag{A14}$$

There are n distinct eigenvalues λ_k, each one with multiplicity N^d, given by

$$\lambda_k = 1 - e^{i(\pi/n)(2k+1)}, \tag{A15}$$

where $k=0,1,2,...,n-1$. Observe that if λ_k is an eigenvalue, so is its complex conjugate: $\lambda_k^* = \lambda_{n-k-1}$. Let us denote by $\mathbf{V}_k$ an eigenvector of $\mathbf{A}^{\sigma\sigma}$ associated to λ_k. It has the structure

$$\mathbf{V}_k = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \end{pmatrix}, \tag{A16}$$

where each $\mathbf{v}_i$, $i=1,2,...,n$, is a $1\times N^d$ matrix. We obtain

$$\mathbf{v}_i = -(1-\lambda_k)^i \xi, \quad \text{where } i=1,...,n-1 \tag{A17}$$

and $\xi=\mathbf{v}_n$ is an *arbitrary* column vector of dimension N^d. There are N^d possible linearly independent choices for ξ, corresponding to N^d distinct eigenvectors associated to the same eigenvalue λ_k. We choose them to be

$$\xi_k^{(1)} = \begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \xi_k^{(2)} = \begin{pmatrix} 0 \\ -1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \ldots, \quad \xi_k^{(N^d-1)} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ -1 \\ 0 \end{pmatrix}, \quad \xi_k^{(N^d)} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix}, \tag{A18}$$

so that each $\xi_k^{(l)}$ corresponds to an eigenvector $\mathbf{V}_k^{(l)}$, where $l=1,2,...,N^d$, associated to the eigenvalue λ_k.

The matrix $\mathbf{P}$ that diagonalizes $\mathbf{A}^{\sigma\sigma}$ can be obtained by concatenating all eigenvectors $\mathbf{V}_k^{(l)}$ for all eigenvalues λ_k, $k=0,1,...,n-1$, up to a normalizing factor R:

$$\mathbf{P} = \begin{pmatrix} p_{00}^{(n)} \mathbb{1}_{N^d\times N^d} & \cdots & p_{0,n-1}^{(n)} \mathbb{1}_{N^d\times N^d} \\ \vdots & & \vdots \\ p_{n-1,0}^{(n)} \mathbb{1}_{N^d\times N^d} & \cdots & p_{n-1,n-1}^{(n)} \mathbb{1}_{N^d\times N^d} \end{pmatrix}, \tag{A19}$$

where

$$p_{\nu\nu'}^{(n)} = R\ e^{(i\pi/n)(2\nu'+1)(\nu+1)}, \tag{A20}$$

with $\nu,\nu'=0,1,...,n-1$. The matrix

$$\mathbf{P}^{-1} = \begin{pmatrix} q_{00}^{(n)} \mathbb{1}_{N^d\times N^d} & \cdots & q_{0,n-1}^{(n)} \mathbb{1}_{N^d\times N^d} \\ \vdots & & \vdots \\ q_{n-1,0}^{(n)} \mathbb{1}_{N^d\times N^d} & \cdots & q_{n-1,n-1}^{(n)} \mathbb{1}_{N^d\times N^d} \end{pmatrix}, \tag{A21}$$

where

$$q_{\nu'\nu}^{(n)} = R' e^{-(i\pi/n)(2\nu'+1)(\nu+1)} \tag{A22}$$

is the inverse of $\mathbf{P}$, upon a suitable choice of R and R'; i.e.,

$$R = R' = 1/\sqrt{n}, \tag{A23}$$

so that they satisfy the relation

$$\sum_{\nu=0}^{n-1} p_{\nu_1\nu}^{(n)} q_{\nu\nu_2}^{(n)} = \delta_{\nu_1\nu_2}. \tag{A24}$$

Hence,

$$p^{(n)}_{\nu\nu'} = \frac{1}{\sqrt{n}} e^{(i\pi/n)(2\nu'+1)(\nu+1)}, \tag{A25}$$

$$q^{(n)}_{\nu'\nu} = \frac{1}{\sqrt{n}} e^{-(i\pi/n)(2\nu'+1)(\nu+1)}. \tag{A26}$$

[1] K. Kubo, Prog. Theor. Phys. Suppl. **69**, 290 (1980).
[2] J. A. Henderson, J. Oitmaa, and M. C. B. Ashley, Phys. Rev. B **46**, 6328 (1992), and references therein.
[3] Y. Grandati, A. Bérard, and P. Grangé, J. Math. Phys. **33**, 1082 (1992).
[4] M. Creutz, *Evaluating Grassmann Integrals*, pre-print hep-lat/9806037.
[5] I. C. Charret, E. V. Corrêa Silva, S. M. de Souza, O. R. Santos, M. T. Thomaz, and C. E. I. Carneiro, Physica A **264**, 204 (1999).
[6] I. C. Charret, S. M. de Souza, and M. T. Thomaz, Braz. J. Phys. **26**, 720 (1996).
[7] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980), Sec. 9.1.3.
[8] M. S. Swanson, *Path Integrals and Quantum Processes* (Academic, New York, 1992), Chap. 5.
[9] A. Das, *Field Theory—A Path Integral Approach* (World Scientific, Singapore, 1993), Sec. 5.2.
[10] Y. Hatsugai and M. Kohmoto, J. Phys. Soc. Jpn. **61**, 2056 (1992).
[11] J. Hubbard, Proc. R. Soc. London, Ser. A **277**, 237 (1963); **281**, 401 (1964); M. Gutzwiller, Phys. Rev. Lett. **10**, 159 (1963); Phys. Rev. A **137**, 1726 (1965).
[12] F. S. Nogueira and E. V. Anda, Int. J. Mod. Phys. B **10**, 3705 (1996).
[13] M. Takahashi, Prog. Theor. Phys. **47**, 69 (1972).

Uniqueness of Gibbs states in one-dimensional antiferromagnetic model with long-range interaction

Azer Kerimov
Department of Mathematics, Bilkent University, 06533 Bilkent, Ankara, Turkey

(Received 28 May 1998; accepted for publication 2 April 1999)

Uniqueness of Gibbs states in the one-dimensional antiferromagnetic model with very long-range interaction is established. © *1999 American Institute of Physics.* [S0022-2488(99)03309-5]

I. INTRODUCTION

We study a model on the lattice $\mathbf{Z}^1$ with the Hamiltonian

$$\mathbf{H}(\varphi(x)) = \sum_{x,y\in\mathbf{Z}^1;x>y} U(x-y)\varphi(x)\varphi(y) - \mu\sum_{x\in\mathbf{Z}^1}\varphi(x), \tag{1}$$

where the spin variable $\varphi(x)$ takes the values 0 and 1, μ is a chemical potential. The antiferromagnetic potential $U(x)>0$ satisfies the following conditions:

(1) $U(x+y)+U(x-y)>2U(x);x,y\in\mathbf{Z}^1,x>y$.

(2) The function $U(x)$ can be extended to a twice continuously differentiable function such that $U(x)\sim A(x)^{-\gamma}$, $U'\sim -A\gamma x^{-\gamma-1}$ and $U''(x)\sim A\gamma(\gamma+1)x^{-\gamma-2}$ at $x\to\infty$; where $\gamma>1$, and A is a strong positive constant.

The first convexity condition plays a significant role for the structure of the set of all ground states of the model (1). The second condition determines the character of the potential's decrease at infinity and is important in further calculations.

The hypothesis on the uniqueness of the Gibbs states in the model (1) was stated by Sinai in 1983 (see Ref. 1, Problem 1).

It is well known that the condition $\Sigma_{x\in\mathbf{Z}^1,x>0}xU(x)<\infty$ automatically implies the uniqueness of the Gibbs states.[2–4] We investigate the phase transition problem in the model (1) in the alternative case, when $U(x)\sim Ax^{-\gamma}$, where $\gamma=1+\alpha$, $0<\alpha<1$.

The ferromagnetic version of this model [when the potential $U(x)$ is negative] was considered by Dyson in his well-known papers.[5,6] He proved the existence of two extreme limit Gibbs states P^+ and P^- corresponding to the ground states $\varphi(x)=+1$ and $\varphi(x)=-1$ at low temperatures.

A series of papers has been devoted to the investigation of the antiferromagnetic model (1).[1,7–13]

The validity of Sinai's hypothesis for rational values of the density (for almost each value of the external field) at low temperatures was proved in Ref. 13.

The main purpose of this paper is to extend the result of Ref. 13 to all values of the external field and to all values of the temperature.

Theorem 1: *The model (1) has a unique limit Gibbs state at all values of the temperature* β^{-1}.

Let us introduce necessary definitions. The set of all periodic configurations we denote by Φ^{per}. For every $\varphi\in\Phi^{\mathrm{per}}$, we define $q=\Sigma_{y=x+1}^{x+p}\varphi(x)$, where p is the period of φ. It is obvious that q does not depend on x. Therefore, the density of each periodic configuration is $\kappa=q/p$. It is more convenient to work with the reciprocal of the density, $\eta(\varphi(x))=p/q$, which represents the average distance between neighboring points at which $\varphi(x)=1$. For every configuration $\varphi\in\Phi^{\mathrm{per}}$ the mean energy $h(\varphi)$ is defined as follows:

$$h(\varphi(x))=\frac{1}{p}\sum_{y=x+1}^{x+p}\varphi(x)\sum_{z>0}U(z)\varphi(y+z).$$

The last expression does not depend on x.

The following definition is useful for describing the zero temperature phase diagram of the model (1).

We fix a positive rational number p/q.

A configuration $\varphi_0(x)\in\Phi^{\mathrm{per}}$ with $\eta(\varphi_0(x))=p/q$ is called a special ground state[1] if

$$h(\varphi(x))=\inf_{\varphi\in\Phi^{\mathrm{per}},\,\eta(\varphi)=p/q}h(\varphi).$$

Hubbard's criterion (Refs. 1 and 7): Let $\varphi\in\Phi^{\mathrm{per}}$ and $r_i(x;\varphi)$ denotes the distance between a particle placed at $x\in\mathbf{Z}^1$ and ith particle on the right. If for each x and i

$$[i\eta]\leqslant r_i(x;\varphi)\leqslant[i\eta]+1,$$

(the square brackets denote the integer part of the enclosed number) then φ is a special ground state.

The existence of configurations satisfying Hubbard's criterion (the special ground states) is shown in Ref. 1. The remarkable elegant formula for the special ground states was offered by Aubry. Here we give the construction of the special ground states for each fixed rational value of the density κ.[1]

Every rational number p/q has a unique decomposition into a finite continued fraction:
$p/q=[n_0,n_1,\dots,n_s]$, this means that

$$n_0+\cfrac{1}{n_1+\cfrac{1}{n_2+\dots+\cfrac{1}{n_s}}}.$$

The ground state for a configuration with $\eta=[n_0,n_1,\dots,n_s]$ will be constructed by recursion.

(1) $\eta=n_0\geqslant 1$, n_0 is an integer. The periodic configuration with equally distant x at which $\varphi(x)=1$ satisfies Hubbard's criterion i.e., is a special ground state. In this case $r_i(x;\varphi)=in_0$, $i>0$.

(2) $\eta=n_0+1/n_1$, where n_0 and n_1 are integers, $n_0\geqslant 1$, $n_1>1$. Then the (n_0n_1+1) periodic configuration

$$0\,\underbrace{\underbrace{0\dots01}_{n_0}\dots\underbrace{0\dots01}_{n_0}}_{n_1\ \text{times}}$$

also satisfies Hubbard's criterion and is a special ground state.

(3) $\eta=[n_0,n_1,\dots,n_s]$, where $n_0,n_1,\dots,n_s$ are integers, $n_0,n_1,\dots,n_s\geqslant 1$. For $s=0$ and $s=1$ the required configurations are already constructed. Suppose we have already constructed a ground state with $s=m$ and $\kappa=[n_0,n_1,\dots,n_m]$. Then the following configuration with $s=m+1$ and $\kappa=[n_0,n_1,\dots,n_{m+1}]$ is constructed as

$$\varphi(n_0,\dots,n_{m+1})=\varphi(n_0,\dots,n_{m-1})\,\underbrace{\varphi(n_0,\dots n_m)\dots\varphi(n_0,\dots,n_m)}_{n_{m+1}\ \text{times}}.$$

Here, $\varphi(n_0,\dots,nj)$, $j=m-1$, $m,m+1$, are the blocks from which the ground states for $\eta=[n_0,\dots,n_j]$ are obtained by periodic continuations.

The constructed configuration satisfies Hubbard's criterion and therefore is a special ground state for $\eta=[n_0,n_1,\dots,n_m,n_{m+1}]$.[1]

The following explicit expression for the mean energy of the special ground state follows from Hubbard's criterion:[1]

$$h_\kappa = \kappa \sum_{i=1}^{\infty} U(m_i)\pi_i + U(m_i+1)(1-\pi_i), \tag{2}$$

where $m_i = [i\eta]$, $\pi_i = 1 + m_i - i\eta$.

This formula shows that the function of mean energy as a function of the density κ is continuous on the set of all rationals and can be extended to a continuous function defined on whole segment [0, 1].

Theorem 2: *(Refs. 9 and 1.)* (1) *The function h_κ is convex.*

(2) *In each rational point the function h_κ has a left-hand derivative μ_κ^- and a right-hand derivative μ_κ^+, with $\mu_\kappa^+ > \mu_\kappa^-$.*

(3) *The Lebesgue measure of the complement of the set $\cup_\kappa(\mu_\kappa^-, \mu_\kappa^+)$ in the real line $\mathbf{R}$ is zero.*

The following theorem gives the full description of the set of all special ground states of the model (1) at rational densities.

Theorem 3: *(Ref. 12.) Suppose that the value of the external field μ of the model (1) belongs to the interval $(\mu_\kappa^-, \mu_\kappa^+)$ for some number $\kappa = q/p$. Then the special ground state of the model* (1) *is unique up to translations.*

Following Theorem 4 generalizes the main result of Ref. 13 for all values of the temperature and is a special case (rational densities) of Theorem 1.

Theorem 4: Suppose that the value of the external field μ of the model (1) belongs to the interval $(\mu_\kappa^-, \mu_\kappa^+)$ for some number $\kappa = q/p$.

Then the model (1) has a unique limit Gibbs state at all values of the temperature β^{-1}.

Suppose that the value of the external field μ of the model (1) belongs to the interval $(\mu_\kappa^-, \mu_\kappa^+)$ for some number $\kappa = q/p$.

Let us consider an arbitrary configuration $\varphi(x)$. We say that $\varphi([a,b])$; $a,b \in \mathbf{Z}^1$ is a preregular phase, if there exists a special ground state φ_κ, such that the restriction of this configuration to $[a,b]$ coincides with $\varphi([a,b])$. We say that $\varphi([c,d])$; $c,d \in \mathbf{Z}^1$ is a regular phase, if there exists a preregular phase $\varphi([a,b])$; $a,b \in \mathbf{Z}^1$, such that $c - a > d_0 p$ and $b - d > d_0 p$. Thus, right and left $d_0 p$ extensions of a regular phase are ground states.

Let us consider a set $A = \cup_i [a_i, b_i]$, where $\varphi([a_i, b_i])$ is a regular phase and supp PB is the complement of A in $\mathbf{Z}^1$. The connected components of supp PB defined in such a way are called supports of precontours and are denoted by supp PK: supp $PK = \cup_{i \in \mathrm{Ind}}$ supp PK_i.

For each fixed rational density κ the constant d_0 satisfies some technical conditions.[13] In this work we do not need the explicit value of d_0.

Definition 1 (Ref. 13): The pair $PK = (\mathrm{supp}\, PK, \varphi'(\mathrm{supp}\, PK))$ is called a precontour. The set of all precontours is called a preboundary PB of the configuration $\varphi'(x)$. Two precontours PK_1 and PK_2 are said to be connected if dist(supp PK_1, supp PK_2) $< N_b$. The set of precontours (PK_i; $i \in$ Ind) is called connected if for any two precontours PK_c and PK_d; $c,d \in$ Ind there exists a collection $(PK_{j_1} = PK_c, \ldots, PK_{j_i}, \ldots, PK_{j_{n-1}}, PK_{j_n} = PK_d)$; $j_i \in$ Ind, $i = 1,...,n$; such that any two precontours PK_{j_i} and $PK_{j_{i+1}}$, $i = 1,...,n-1$ are connected. Let $\cup_{i=1}^n PK_i$ be some maximal connected component of the preboundary PB. Suppose that supp $PK_i = [a_i, b_i]$ and $b_i < a_{i+1}$; $i = ,\ldots, n-1$.

The pair $K = (\mathrm{supp}\, K, \varphi'(\mathrm{supp}\, PK))$, where supp $K = [a_1, b_n]$ is called a contour. The set of all contours is called a boundary B of the configuration $\varphi'(x)$.

In this work we do not need the exact value of the constant N_b.[12] From Ref. 12 it becomes clear that $\lim_{p\to\infty} N_b = \infty$. Thus, for irrational values of the density κ N_b is not defined, but as will be seen below, we do not need to define N_b for irrational densities.

Note that supp $K = (\cup_{i=1}^n \mathrm{supp}\, PK_i) \cup ([a_1, b_n] - (\cup_{i=1}^n \mathrm{supp}\, PK_i)) = \mathrm{supp}^1 K \cup \mathrm{supp}^2 K$.

The sets $\mathrm{supp}^1 K$ and $\mathrm{supp}^2 K$ will be, respectively, called the essential and regular parts of the support $\mathrm{supp}\, K$.

Let the boundary conditions $\overline{\varphi}(x)=[\varphi(x),x\in(-\infty,-V-1]\cup[V+1,\infty)]$ be fixed. The set of all configurations $\varphi(x)$; $x\in[-V,V]$ we denote via $\Phi(V)$.

It is obvious that for each contour K, such that $\mathrm{supp}\, K\in[-V+(d_0+1)p,V-(d_0+1)p]$, there exists a configuration $\psi_K([-V,V])$ such that the boundary of the configuration $\psi_K([-V,V])$ includes the contour K only:

$$B(\psi_K([-V,V]))=K.$$

Let $\mathrm{supp}\, K=[a,b]$. It is obvious that the restrictions of the configuration $\psi_K([-V,V])$ to the segments $[-V,a-1]$ and $[b+1,V]$ coincide with two ground states $\varphi_\kappa^1(x)$ and $\varphi_\kappa^2(x)$.

A contour K is called an interface contour, if $\varphi_\kappa^1(x)\neq\varphi_\kappa^2(x)$.

Note that, $\varphi_\kappa^1(x)$ can be obtained by some shifting of the configuration $\varphi_\kappa^2(x)$.

An interface contour will be denoted as IK.

Let K be a usual contour (not an interface contour) $K,\mathrm{supp}\, K\subset[-V,V]$ and $\psi K(x)=\psi([-V,V])$ *if* $x\in[-V,V]$, and $\overline{\varphi}(x)$ if $x\in(-\infty,-V-1]\cup[V+1,\infty)$; $IK,\mathrm{supp}\, IK\subset[-V,V]$ be an interface contour and $\psi_{IK}(x)=\psi([-V,V])$ if $x\in[-V,V]$, and $\overline{\varphi}(x)$ if $x\in(-\infty,V-1]\cup[V+1,\infty)$; and $\overline{\varphi}_\kappa^1(x)=\varphi_\kappa^1(x)$, if $x\in[-V,V]$, and $\overline{\varphi}(x)$ if $x\in(-\infty,-V-1]\cup[V+1,\infty)$.

Below the configuration $\overline{\varphi}_\kappa^1(x)$ defined for usual contours will be denoted by $\overline{\varphi}_\kappa(x)$.

The weights of the usual contour K and interface contour IK will be calculated by the following formulas:

$$\gamma(K)=H(\psi_K(x))-H(\overline{\varphi}_\kappa(x)), \tag{3}$$

$$\gamma(IK)=H(\psi_{IK}(x))-H(\overline{\varphi}_\kappa^1(x)). \tag{4}$$

The proof of Theorem 4 is based on the following idea. Let the boundary conditions $\overline{\varphi}(x)=[\varphi(x),x\in(-\infty,-V-1]\cup[V+1,\infty)]$ be fixed. The set of all configurations $\varphi(x)$; $x\in[-V,V]$ we denote via $\Phi(V)$. Suppose a configuration $\varphi_{\min}(x)\in\Phi(V)$ be a configuration with the minimal energy:

$$H(\varphi_{\min}(x))=\min_{\varphi(x)\in\Phi(V)}H(\varphi(x))\,.$$

Then the configuration $\varphi_{\min}(x)$ almost coincides with a special ground state of the model (1) (Lemma 1 in Sec. II). This fact allows us, based on special ground states, to define a common (for all boundary conditions) contour model and after that by using well-known trick[14] (this trick, which was introduced in Ref. 14 for some special extensions of Pirogov–Sinai theory, is directly applicable to one-dimensional models with long-range interaction) to come to noninteracting clusters from interacting contours. Consider an arbitrary segment I, a sufficiently large volume V, two arbitrary boundary conditions $\varphi^1(x)$ and $\varphi^2(x)$. It turns out that the dependence of the expression $\mathbf{P}^1(\varphi'(I))/\mathbf{P}^2(\varphi'(I))$ on the boundary conditions $\varphi^1(x)$ and $\varphi^2(x)$ can be estimated through the sum of statistical weights of super clusters connecting the segment I with the boundary and this sum is negligible. Thus, two arbitrary extreme Gibbs states are relatively continuous and hence coincide. In Ref. 13 we developed this method [the estimation of dependence of the expression $\mathbf{P}^1(\varphi'(I))/\mathbf{P}^2(\varphi'(I))$ on the boundary conditions through the sum of statistical weights of super clusters connecting the segment I with the boundary] at low temperatures. It turns out that after some modification the method works at all temperatures.

The contents of this paper are as follows. In Sec. II we prove Theorem 4, in Sec. III we complete the proof of Theorem 1.

II. UNIQUENESS OF GIBBS STATES: THE DENSITY κ IS p/q

Let us now introduce some necessary facts.

Suppose that the value of the external field μ of the model (1) belongs to the interval $(\mu_\kappa^-, \mu_\kappa^+)$ for some number $\kappa = q/p$.

Let the boundary conditions $\varphi^1(x) = [\varphi^1(x), x \in (-\infty, -V-1] \cup [V+1, \infty)]$ be fixed and

$$H(\varphi(x)|\varphi^1(x)) = -\mu \sum_{x \in \mathbf{Z}^1, x \in [-V,V]} \varphi(x) + \sum_{x,y \in \mathbf{Z}^1, x>y; x,y \in [-V,V]} U(x-y)\varphi(x)\varphi(y)$$
$$+ \sum_{x,y \in \mathbf{Z}^1, x>y; x \in [-V,V]; y \notin [-V,V]} U(x-y)\varphi(x)\varphi^1(y)$$
$$+ \sum_{x,y \in \mathbf{Z}^1, x>y; x \notin [-V,V], y \in [-V,V]} U(x-y)\varphi^1(x)\varphi(y). \tag{5}$$

Lemma 1: Let $\varphi_{\min}(x) \in \Phi(V)$ *be a configuration with the minimal energy:*

$$H(\varphi_{\min}(x)|\varphi^1(x)) = \min_{\varphi(x) \in \Phi(V)} H(\varphi(x)|\varphi^1(x)) .$$

Then the configuration $\varphi_{\min}(x)$ *has the following structure.*

The restriction of the configuration $\varphi_{\min}(x)$ on the set $[-V+N_b, V-N_b]$ *contains at most* $p-1$ *contours, moreover, all of them are interface contours* $IK_{i,}$, $i=1,...,m$, *where* $m<p-1$ *and* $|\mathrm{supp}\, IK_i| < 3d_0 p + N_b$.

Lemma 1 was proved in Ref. 13 [see Lemma 12 (Ref. 13) and Sec. 5 of Ref. 13].

Let $H(\varphi(x)|\varphi^1(x), \varphi_{\min}(x))$ denote the relative energy of a configuration $\varphi(x)$ [with respect to $\varphi_{\min}(x)$]:

$$H(\varphi(x)|\varphi^1(x), \varphi_{\min}(x)) = H(\varphi(x)|\varphi^1(x)) - H(\varphi_{\min}(x)|\varphi^1(x)).$$

Consider the Gibbs distribution $\mathbf{P}^1$ on $\Phi(V)$ corresponding to the boundary conditions $\varphi^1(x) = [\varphi^1(x), x \in (-\infty, -V-1] \cup [V+1, \infty)]$:

$$\mathbf{P}^1(\varphi'(x)) = \frac{\exp(-\beta(H(\varphi'(x)|\varphi^1(x), \varphi_{\min}(x))))}{\sum_{\varphi(x) \in \Phi(V)} \exp(-\beta(H(\varphi(x)|\varphi^1(x), \varphi_{\min}(x))))}. \tag{6}$$

Let $\varphi(x) \in \Phi(V)$ be an arbitrary configuration, the boundary of the $\varphi(x)$ includes a finite number of usual contours K_i; $i=1,...,n$, and a finite number of interface contours IK_i; $i=n+1,...,n+m$. Let $K_i = K_i$; $i=1,...,n$; $K_i = IK_i$; $i=n+1,...,n+m$. The set of all contours of the boundary conditions $\varphi^1(x)$ will be denoted by K_0.

The statistical weights of contours and interface contours are

$$w(K_i) = \exp(-\beta\gamma(K_i)). \tag{7}$$

The following equation is a direct consequence of the formulas (3), (4), and (7)

$$\exp(-\beta H(\varphi(x)|\varphi^1(x), \varphi_{\min}(x))) = Q_1 \prod_{i=1}^{n+m} w(K_i) \exp(-\beta G(K_0, K_1, \ldots, K_{n+m}), \tag{8}$$

where the multiplier $G(K_0, K_1, \ldots, K_{n+m})$ corresponds to the interaction between contours (usual and interface), and with the boundary conditions $\varphi^1(x)$

$$G(K_0, K_1, \ldots, K_{n+m}) = \sum_{i,j=0; i<j}^{n+m} G(K_i, K_j) = \sum_{i,j; i<j} \sum_{(x,y) \in \mathrm{Int}(K_i, K_j)} f(x,y,\varphi) \tag{9}$$

and the multiplier $Q_1 = Q_1(V, \varphi(x), \varphi^1(x))$ is uniformly bounded from below and above: $0 < \mathrm{const}_1 < Q_1 < \mathrm{const}_2$. The factor Q_1 appears due to the facts that the configuration $\varphi_{\min}(x)$ not necessarily coincides with a special ground state and is bounded due to Lemma 1.

Now we write down the value of the interaction between the contours K_i and K_j, the value of the interaction between the interface contours IK_i and IK_j and the value of the interaction between contour K_i and interface contour IK_j.

Suppose $\operatorname{supp} K_l=[a_l,b_l]$; $\operatorname{supp} IK_l=[a_l,b_l]$.

Let

$$\operatorname{supp} IK_i^+=[b_i,a_{i+1}] \quad \text{and} \quad \operatorname{supp} IK_i^-=[b_{i-1},a_i],$$

where $b_0=c$, if there exists $K\in B(\varphi'(x))$, such that $\operatorname{supp} K=[-\infty,c]$ and $b_0=-\infty$ otherwise; $a_{m+1}=d$, if there exists $K\in B(\varphi'(x))$, such that $\operatorname{supp} K=[d,\infty]$ and $a_{m+1}=\infty$ otherwise.

(1) The contour $K_i\in B(\varphi'(x))$ interacts with the contour $K_j\in B(\varphi'(x))$ through all pairs (x,y), such that $(x,y)\in \operatorname{Int}(K_i,K_j)$ and $f'(x,y,\varphi)\neq 0$ where

$$\operatorname{Int}(K_i,K_j)=[(x,y):x,y\in \mathbf{Z}^1; x\in \operatorname{supp} K_i, y\in \operatorname{supp} K_j].$$

The value of the interaction

$$f'(x,y,\varphi)=U(x-y)(\varphi'(x)\varphi'(y)-\psi_{K_i}(x)\psi_{K_i}(y)+\overline{\varphi^i_\kappa}(x)\overline{\varphi^i_\kappa}(y)$$
$$-\psi_{K_j}(x)\psi_{K_j}(y)+\overline{\varphi^j_\kappa}(x)\overline{\varphi^j_\kappa}(y)).$$

(2) The interface contour $IK_i\in B(\varphi'(x))$ interacts with the interface contour $IK_j\in B(\varphi'(x))$ (let $a_j>b_i$) through all pairs (x,y), such that $(x,y)\in \operatorname{Int}(IK_i,IK_j)$ and $f''(x,y)\neq 0$, where

$$\operatorname{Int}(IK_i,IK_j)=\operatorname{Int}^1(IK_i,IK_j)+\operatorname{Int}^2(IK_i,IK_j)+\operatorname{Int}^3(IK_i,IK_j)+\operatorname{Int}^4(IK_i,IK_j),$$

$$\operatorname{Int}^1(IK_i,IK_j)=[(x,y):x,y\in \mathbf{Z}^1; x\in \operatorname{supp} IK_i \text{ and } y\in \operatorname{supp} IK_j],$$

$$\operatorname{Int}^2(IK_i,IK_j)=[(x,y):x,y\in \mathbf{Z}^1; x\in \operatorname{supp} IK_i \text{ and } y\in \operatorname{supp} IK_j^+],$$

$$\operatorname{Int}^3(IK_i,IK_j)=[(x,y):x,y\in \mathbf{Z}^1; x\in \operatorname{supp} IK_i^- \text{ and } y\in \operatorname{supp} IK_j],$$

$$\operatorname{Int}^4(IK_i,IK_j)=[(x,y):x,y\in \mathbf{Z}^1; x\in \operatorname{supp} IK_i^- \text{ and } y\in \operatorname{supp} IK_j^+].$$

The value of the interaction

$$f''(x,y,\varphi)=f''_1(x,y)=U(x-y)(\varphi'(x)\varphi'(y)-\psi_{IK_i}(x)\psi_{IK_i}(y)$$
$$+\overline{\varphi^i_\kappa}(x)\overline{\varphi^i_\kappa}(y)-\psi_{IK_j}(x)\psi_{IK_j}(y)+\overline{\varphi^j_\kappa}(x)\overline{\varphi^j_\kappa}(y))$$

if $(x,y)\in \operatorname{Int}^2(IK_i,IK_j)$,

$$f''(x,y)=f''_2(x,y)=U(x-y)(\varphi'(x)\varphi'(y)-\psi_{IK_i}(x)\psi_{IK_i}(y)+\overline{\varphi^i_\kappa}(x)\overline{\varphi^i_\kappa}(y))$$

if $(x,y)\in \operatorname{Int}^2(IK_i,IK_j)$,

$$f''(x,y)=f''_3(x,y)=U(x-y)(\varphi'(x)\varphi'(y)-\psi_{IK_j}(x)\psi_{IK_j}(y))+\overline{\varphi^j_\kappa}(x)\overline{\varphi^j_\kappa}(y))$$

if $(x,y)\in \operatorname{Int}^3(IK_i,IK_j)$,

$$f''(x,y)=f''_4(x,y)=U(x-y)(\varphi'(x)\varphi'(y)-\overline{\varphi^{1,i}_\kappa}(x)\overline{\varphi^{1,i}_\kappa}(y)-\overline{\varphi^{2,j}_\kappa}(x)\overline{\varphi^{2,j}_\kappa}(y))$$

if $(x,y)\in \operatorname{Int}^4(IK_i,IK_j)$.

(3) The contour $K_i \in B(\varphi'(x))$ interacts with the interface contour $IK_j \in B(\varphi'(x))$ through all pairs (x,y), such that $(x,y) \in \mathrm{Int}(K_i, IK_j)$ and $f'''(x,y) \neq 0$, where

$$\mathrm{Int}(K_i, IK_j) = \mathrm{Int}^1(K_i, IK_j) + \mathrm{Int}^2(K_i, IK_j),$$

$$\mathrm{Int}^1(K_i, IK_j) = [(x,y): x,y \in \mathbf{Z}^1; x \in \mathrm{supp}\, K_i \quad \text{and} \quad y \in \mathrm{supp}\, IK_j],$$

$$\mathrm{Int}^2(K_i, IK_j) = [(x,y): x,y \in \mathbf{Z}^1; x \in \mathrm{supp}\, K_i \quad \text{and} \quad y \in \mathrm{supp}\, IK_j^+]$$

if $a_j > b_i$, and

$$\mathrm{Int}^2(K_i, IK_j) = [(x,y): x,y \in \mathbf{Z}^1; x \in \mathrm{supp}\, K_i \quad \text{and} \quad y \in \mathrm{supp}\, IK_j^-]$$

if $a_i > b_j$.
The value of the interaction

$$f'''(x,y) = f_1'''(x,y) = U(x-y)(\varphi'(x)\varphi'(y) - \psi_{K_i}(x)\psi_{K_i}(y) + \overline{\varphi^i_\kappa}(x)\overline{\varphi^i_\kappa}(y) - \psi_{IK_j}(x)\psi_{IK_j}(y)) + \overline{\varphi^j_\kappa}(x)\overline{\varphi^j_\kappa}(y))$$

if $(x,y) \in \mathrm{Int}^1(K_i, IK_j)$,

$$f'''(x,y) = f_2'''(x,y) = U(x-y)(\varphi'(x)\varphi'(y) - \psi_{K_i}(x)\psi_{K_i}(y) + \overline{\varphi^i_\kappa}(x)\overline{\varphi^i_\kappa}(y))$$

if $(x,y) \in \mathrm{Int}^2(K_i, IK_j)$.

For simplicity K_i, $i = 1,...,n+m$ will be denoted by K_i, $i \in \mathrm{Ind}$, where the statistical weights $w(K_i)$ are defined by the formulas (7), (3), and (4). Thus, the formula (8) has the form

$$\exp(-\beta H(\varphi(x)|\varphi^1(x), \varphi_{\min}(x))) = Q_1 \prod_{i \in \mathrm{Ind}} w(K_i) \exp(-\beta G(K_0, K_1, \ldots, K_{n+m})). \tag{10}$$

The set of all pairs (x,y) in the double sum (9) will be denoted by $Y = Y(K_0, K_1, \ldots, K_{n+m})$. Write (10) as follows:

$$\exp(-\beta H(\varphi(x)|\varphi^1(x), \varphi_{\min}(x))) = Q_1 \prod_{i \in \mathrm{Ind}} w(K_i) \prod_{(x,y) \in Y} (1 + \exp(-\beta f(x,y,\varphi) - 1). \tag{11}$$

From (11) we get

$$\exp(-\beta H(\varphi(x)|\varphi^1(x), \varphi_{\min}(x))) = Q_1 \prod_{G' \subset G} \prod_{i \in \mathrm{Ind}} w(K_i) \prod_{(x,y) \in Y'; f(x,y,\varphi) \neq 0} g(x,y), \tag{12}$$

where the summation is taken over all subsets Y' (including the empty set) of the set Y, and $g(x,y,\varphi) = \exp(-\beta f(x,y,\varphi)) - 1$.

Consider an arbitrary term of the sum (12), which corresponds to the subset $Y' \subset Y$. Let the bond $(x,y) \in Y'$. Below, contours and interface contours will be called contours. Consider the set $\mathbf{K}$ of all contours such that for each contour $K \subset \mathbf{K}$, the set $\mathrm{supp}\, K \cap (x \cup y)$ contains one point. We call any two contours from $\mathbf{K}$ connected. The set of contours $\mathbf{K}$ is called Y' connected if for any two contours K_a and K_b there exists a collection $(K_1 = K_a, K_2, \ldots, K_n = K_b)$ such that any two contours K_i and K_{i+1}, $i = 1,...,n-1$, are connected by some bond $(x,y) \in Y'$.

The pair $D = [(K_i, i = 1,...,s); Y']$, where Y' is some set of bonds, is called a cluster provided there exists a configuration $\varphi(x)$ such that $K_i \in B(\varphi(x))$; $i = 1,...,s$; $Y' \subset Y$; and the set $(K_i$, $i = 1,...,s)$ is Y' connected. The statistical weight of a cluster D is defined by the formula.

$$w(D)=\prod_{i=1}^{s} w(K_i) \prod_{(x,y)\in Y'} g(x,y,\varphi). \tag{13}$$

Two clusters D_1 and D_2 are called compatible provided any two contours K_1 and K_2 belonging to D_1 and D_2, respectively, are compatible and not connected. A set of clusters is called compatible provided any two clusters of it are compatible.

If $D=[(K_i,i=1,...,s);Y']$, then we say that $K_i\in D$; $i=1,...,s$.

The following lemma is a direct consequence of the definitions.

Lemma 2: Let the boundary conditions $\varphi^1(x)=[\varphi^1(x),x\in(-\infty,-V-1]\cup[V+1,\infty)]$ *be fixed.*

If $[D_1,\ldots,D_m]$ *is a compatible set of clusters and* $\cup_{i=1}^m \operatorname{supp} D_i\subset[-V,V]$, *then there exists a configuration* $\varphi(x)$ *which contains this set of clusters. For each configuration* $\varphi(x)$ *we have*

$$\exp(-\beta H(\varphi(x)|\varphi^1(x),\varphi_{\min}(x)))=Q_1\sum_{Y'\subset Y}\prod w(D_i),$$

where the clusters D_i *are completely determined by the set* Y'. *The partition function is*

$$\Xi(\varphi^1(x))=Q\sum w(D_1)\cdots w(D_m),$$

where the summation is taken over all nonordered compatible collections of clusters and the factor $Q=Q(V,\varphi^1(x))$ *is uniformly bounded:* $0<\text{const}<Q<\text{const}_2$.

Lemma 2 shows that we come to noninteracting clusters from interacting contours.

Let $\mathbf{P}^1$ and $\mathbf{P}^2$ be two Gibbs states of the model (1) corresponding to the boundary conditions $\varphi^1(x)$ and $\varphi^2(x)$, respectively.

The following lemma has a key role in the proof of Theorem 4.

Lemma 3: Suppose that the value of the external field μ *of the model (1) belongs to the interval* $(\mu_\kappa^-,\mu_\kappa^+)$ *for some number* $\kappa=q/p$.

Then the measures $\mathbf{P}^1$ *and* $\mathbf{P}^2$ *are absolutely continuous with respect to each other.*

Proof: Let $I=[a,b]$ be an arbitrary segment and $\varphi'(I)$ be an arbitrary configuration.

In order to prove the lemma we show that there exist two positive constants s and S not depending on I, $\varphi^1(x)$, $\varphi^2(x)$ and $\varphi'(I)$, such that

$$s\leq \mathbf{P}^1(\varphi'(I))/\mathbf{P}^2(\varphi'(I))\leq S. \tag{14}$$

Let $\mathbf{P}_\mathbf{V}^\mathbf{1}$ and $\mathbf{P}_\mathbf{V}^\mathbf{2}$ be Gibbs measures corresponding to the boundary conditions $\varphi^1(x)$, and $\varphi^2(x)$, $x\in\mathbf{Z}^1-I_V$, respectively, where $I_V=[-V,V]$.

Therefore,

$$\lim_{V\to\infty}\mathbf{P}_\mathbf{V}^\mathbf{1}=\mathbf{P}^\mathbf{1} \quad \text{and} \quad \lim_{V\to\infty}\mathbf{P}_\mathbf{V}^\mathbf{2}=\mathbf{P}^\mathbf{2},$$

where by convergence we mean weak convergence of probability measures.

In order to establish the inequality (14) it will be proved that for each fixed interval I, $I\subset[-M,M]$ there exists a number $V_0(M)$, which depends on M only, such that

$$s\leq \mathbf{P}_\mathbf{V}^\mathbf{1}(\varphi'(I))/\mathbf{P}_\mathbf{V}^\mathbf{2}(\varphi'(I))\leq S \tag{15}$$

if $V>V_0$.

Consider

$$\mathbf{P}_{\mathbf{V}}^{\mathbf{1}}(\varphi'(I))=\frac{\sum_{\varphi(I_V):\varphi(I)=\varphi'(I)}\exp(-\beta H(\varphi(I_V)|\varphi^1(x),\varphi_{\min}(x)))O(\varphi(I),V,\varphi^1)}{\sum_{\varphi(I_V)}\exp(-\beta H(\varphi(I_V)|\varphi^1(x),\varphi_{\min}(x)))O(\varphi(I),V,\varphi^1)}$$
$$=\frac{\Xi(I_V-I|\varphi^1(x),\varphi'(I),\varphi_{\min}(x))O(\varphi(I),V,\varphi^1)}{\sum_{\varphi''(I)}\Xi(I_V-I|\varphi^1(x),\varphi''(I),\varphi_{\min}(x))O(\varphi(I),V,\varphi^1)}$$

where $\Xi(I_V-I|\varphi^1(x),\varphi'(I),\varphi_{\min}(x))$ denotes the partition function corresponding to the boundary conditions $\varphi^1(x),\ x\in\mathbf{Z}^1-I_V,\ \varphi'(I),\ x\in I$ and

$$O(\varphi(I),V,\varphi^1)=\exp(-\beta\sum_{x,y\in\mathbf{Z}^1;x\in\mathbf{Z}^1-I_V,y\in I}U(x-y)(\varphi^1(x)\varphi(y)-\varphi^1(x)\varphi_{\min}(x))).$$

We can express $\mathbf{P}_{\mathbf{V}}^{\mathbf{2}}(\varphi'(I))$ in just the same way.

In order to prove the inequality (15) it is enough to establish inequality (16) and inequality (17):

$$1/2<O(\varphi(I),V,\varphi^i(x))<2,\quad i=1,2 \tag{16}$$

[where the inequalities in (16) are held uniformly with respect to $\varphi(I)$ and φ^i: for each I there exists V, not depending on $\varphi(I)$ and φ^i] and

$$1/S\leqslant\frac{\Xi(I_V-I|\varphi^1(x),\varphi''(I),\varphi_{\min}(x))}{\Xi(I_V-I|\varphi^1(x),\varphi'(I),\varphi_{\min}(x))}\Bigg/\frac{\Xi(I_V-I|\varphi^2(x),\varphi''(I),\varphi_{\min}(x))}{\Xi(I_V-I|\varphi^2(x),\varphi'(I),\varphi_{\min}(x))}\leqslant1/s \tag{17}$$

for arbitrary $\varphi''(I)$.

Indeed, if the inequality (17) holds, then

$$\frac{\Xi(I_V-I|\varphi^1(x),\varphi'(I),\varphi_{\min}(x))}{\sum_{\varphi''(I)}\Xi(I_V-I|\varphi^1(x),\varphi''(I),\varphi_{\min}(x))}\Bigg/\frac{\Xi(I_V-I|\varphi^2(x),\varphi'(I),\varphi_{\min}(x))}{\sum_{\varphi''(I)}\Xi(I_V-I|\varphi^2(x),\varphi''(I),\varphi_{\min}(x))}$$
$$=\mathbf{A}_{\mathbf{V}}^{\mathbf{1}}(\varphi'(I))/\mathbf{A}_{\mathbf{V}}^{\mathbf{2}}(\varphi'(I))$$
$$=1\Bigg/\left(\frac{\sum_{\varphi''(I)}\Xi(I_V-I|\varphi^1(x),\varphi''(I),\varphi_{\min}(x))}{\Xi(I_V-I|\varphi^1(x),\varphi'(I),\varphi_{\min}(x))}\Bigg/\frac{\sum_{\varphi''(I)}\Xi(I_V-I|\varphi^2(x),\varphi''(I),\varphi_{\min}(x))}{\Xi(I_V-I|\varphi^2(x),\varphi'(I),\varphi_{\min}(x))}\right)$$
$$=1\Bigg/\frac{(\sum_{\varphi''(I)}\Xi(I_V-I|\varphi^1(x),\varphi''(I,\varphi_{\min}(x)))\Xi(I_V-I|\varphi^2(x),\varphi'(I),\varphi_{\min}(x))}{(\sum_{\varphi''(I)}\Xi(I_V-I|\varphi^2(x),\varphi''(I),\varphi_{\min}(x)))\Xi(I_V-I|\varphi^1(x),\varphi'(I),\varphi_{\min}(x))}.$$

Therefore,

$$1/(1/s)\leqslant\mathbf{A}_{\mathbf{V}}^{\mathbf{1}}(\varphi'(I))/\mathbf{A}_{\mathbf{V}}^{\mathbf{2}}(\varphi'(I))\leqslant1/(1/S)$$

since the quotient of $\sum_{i=1}^n a_i/\sum_{i=1}^n b_i$ lies between $\min(a_i/b_i)$ and $\max(a_i/b_i)$.

Thus, if in addition, the inequality (16) holds, then

$$2^{-4}s<\mathbf{P}_{\mathbf{V}}^{\mathbf{1}}(\varphi'(I)):\mathbf{P}_{\mathbf{V}}^{\mathbf{2}}(\varphi'(I))<2^4S.$$

Now we start to prove the inequalities (16) and (17).

It can be easily shown that (16) is a direct consequence of the condition $U(x)\sim Ax^{-\gamma}$, at $x\to\infty$; where $\gamma>1$, and A is a strong positive constant.

So, in order to complete the proof of Lemma 3 we must establish the following inequality [which is just transformed inequality (17)]:

$$1/S \leqslant \frac{\Xi(I_V-I|\varphi^1(x),\varphi''(I),\varphi_{\min}(x)))\Xi(I_V-I|\varphi^2(x),\varphi'(I),\varphi_{\min}(x))}{\Xi(I_V-I|\varphi^2(x),\varphi''(I),\varphi_{\min}(x)))\Xi(I_V-I|\varphi^1(x),\varphi'(I),\varphi_{\min}(x))} = \frac{\Xi^{1,''}\Xi^{2,'}}{\Xi^{2,''}\Xi^{1,'}} \leqslant 1/s. \tag{18}$$

Consider

$$\Xi^{1,''}\Xi^{2,'} = \Xi(I_V-I|\varphi^1(x),\varphi''(I),\varphi_{\min}(x))\Xi(I_V-I|\varphi^2(x),\varphi'(I),\varphi_{\min}(x)).$$

The following generalization of the definition of the compatibility allows us to represent $\Xi^{1,''}\Xi^{2,'}$ as a single partition function.

A set of clusters is called super compatible provided any of its two parts coming from two partitions sums is compatible. In other words, in super compatibility an intersection of supports of two clusters is allowed.

The following lemma is an analogue of Lemma 2.

Lemma 4: Let boundary conditions $\varphi^1(x)=[\varphi^1(x), x\in(-\infty,-V-1]\cup[V+1,\infty)]$ *and* $\varphi^2(x)=[\varphi^2(x), x\in(-\infty,-V-1]\cup[V+1,\infty)]$ *be fixed.*

If $[D_1,\ldots,D_m]$ *is a super compatible set of clusters and* $\cup_{i=1}^m \operatorname{supp} D_i \subset [-V,V]$, *then there exist two configurations* $\varphi^3(x)$ *and* $\varphi^4(x)$ *which contain this set of clusters. For each two configurations* $\varphi^3(x)$ *and* $\varphi^4(x)$ *we have*

$$\exp(-\beta H(\varphi^3(x)|\varphi^1(x),\varphi_{\min}(x))\exp(-\beta H(\varphi^4(x)|\varphi^1(x),\varphi_{\min}(x)) = Q_1 \sum_{G'\subset G, G''\subset G} \prod w(D_i),$$

where the clusters D_i *are completely determined by the sets* G' *and* G''. *The super partition function is*

$$\Xi^{1,'',2,'} = \Xi^{1,''}\Xi^{2,'} = Q\sum w(D_1)\cdots w(D_m),$$

where the summation is taken over all nonordered super compatible collections of clusters and the factor $Q=Q(V,\varphi^1(x),\varphi^2(x))$ *is uniformly bounded:* $0<\text{const}_1<Q<\text{const}_2$.

Lemma 4 is a direct consequence of the definitions.

An arbitrary connected component of an arbitrary super compatible set of clusters will be called a super clusters. A super cluster $SD=[(K_i, i=1,...,r);G']$ is said to be long if the intersection of the set $(\cup_{i=1}^m \operatorname{supp} K_i))\cup G'$ with both I and $\mathbf{Z}^1-I_V=(-\infty,-V-1]\cup[V+1,\infty)$ is nonempty. In other words, a long super cluster connects the boundary with the segment I.

A set of super clusters is called compatible provided the set of all clusters belonging to these super clusters are super compatible.

It turns out that in our estimates long super clusters are negligible.

Lemma 5: For each fixed interval I, there exists a number $V_0(I)$, *which depends on I only, such that if* $V>V_0(I)$

$$1/2\Xi^{1,',2,''} < \Xi^{1,',2,'',(n.l.)} = \sum w(SD_1)\cdots w(SD_m) < 3/2\Xi^{1,',2,''},$$

where the summation is taken over all nonlong, nonordered compatible collections of super clusters $[SD_1,\ldots,SD_m]$, $\cup_{i=1}^m \operatorname{supp}(SD_i)\subset I_N-I$ *corresponding to the boundary conditions* $\varphi^1(x),\varphi^2(x)$, $x\in\mathbf{Z}^1-I_V$; $\varphi'(x)$ *and* $\varphi''(x)$, $x\in I$.

Consider a collection of contours $K_0,K_1,\ldots,K_n$. The value of the interaction of the contour K_0 with the contours $K_1,\ldots,K_n$ we denote by $G(K_0|K_1,\ldots,K_n)$:

$$G(K_0|K_1,\ldots,K_n) = \prod_{B\in IG(0|1,\ldots,n)} (1+\exp(-\beta f(B)-1)), \tag{19}$$

where $IG(0|1,...,n)$ is the set of all interaction elements intersecting the support of the contour K_0.

Lemma 6:

$$G(K_0|K_1,\ldots,K_n)=\prod_{B\in IG(0|1,...,n)}|(1+\exp(-\beta f(B)-1))$$
$$\leqslant \mathrm{const}(\mathrm{dist}(0|1,...,n))^{-\alpha}(|\mathrm{supp}(K_0)|)^{1-\alpha}, \tag{20}$$

where $\mathrm{dist}(0|1,...,n)$ *is the distance between the support of* K_0 *and the union of the supports of contours* $K_1,\ldots,K_n$.

In other words, the interaction of $K_1,\ldots,K_n$ on K_0 tends to zero when the distance between them increases, and value of the interaction increases with a rate less than the length of the support of K_0.

The technical Lemma 6 follows from the decreasing conditions of the potential $U(x)$. For the rigorous proof see Ref. 13, Lemma 4.

The following lemma is an analogue of Lemma 5 for clusters (not super clusters).

Lemma 7: For each fixed interval I, there exists a number $V_0(I)$, *which depends on I only, such that if* $V>V_0(I)$

$$1/2\Xi^{1,'}<\Xi^{1,',(n.l.)}=\sum w(D_1)\ldots w(D_m)<3/2\Xi^{1,'},$$

where the summation is taken over all nonlong, nonordered compatible collections of clusters $[D_1,\ldots,D_m]$, $\cup_{i=1}^m \mathrm{supp}\, D_i\subset I_N-I$ *corresponding to the boundary conditions* $\varphi^1(x)$, $x\in \mathbf{Z}^1-I_V$; $\varphi'(x)$, $x\in I$.

Proof:

$$\Xi^{1,'}=\Xi^{1,',(n.l.)}+(\Xi^{1,'}-\Xi^{1,',(n.l.)})=\Xi^{1,',(n.l.)}+\Xi^{1,',(l.)},$$

where the summation in $\Xi^{1,',(l.)}$ is taken over all nonordered compatible collections of clusters $[D_1,\ldots,D_m]$ containing at least one long cluster, $\cup_{i=1}^m \mathrm{supp}\, D_i\subset I_N-I$ corresponding to the boundary conditions $\varphi^1(x)$, $x\in\mathbf{Z}^1-I_V$; $\varphi'(x)$, $x\in I$.

By dividing both sides of the last equality by $\Xi^{1,'}$, we get

$$1=\Xi^{1,',(n.l.)}/\Xi^{1,'}+\Xi^{1,',(l.)}/\Xi^{1,'}. \tag{21}$$

Now we are going to show that the second term (which is not necessarily positive) is negligible, that is the absolute value of it is less than 1/2 (actually we can show that the absolute value of the second term is less than any fixed positive number at sufficiently large values of V).

The term $\Xi^{1,',(l.)}/\Xi^{1,'}$ can be interpreted as a ''probability'' P (Long) of the event that there exists at least one long cluster.

We show that the absolute value of this ''probability'' is less than 1/2 by the following method. We estimate the density of long clusters: the probability that a given segment belongs to the support of some long cluster. Since some statistical weights of clusters are positive and some negative, we estimate the absolute values of these ''probabilities.'' We show that for a fixed segment the ''probability'' that this segment belongs to the support of some long cluster with positive ''probability'' minus the ''probability'' that this segment belongs to the support of some long cluster with negative ''probability'' is less than one. Since the density is less than one, by the law of large numbers a ''typical'' long cluster has not very long support, and therefore has long bonds. When V tends to infinity, the total length of bonds tends to infinity, and the impact of these bonds tends to zero.

Now we replace a statistical weight $w(D_i)$ of each cluster D_i belonging to the configuration containing at least one long cluster with its absolute value (and ''probability'' of long cluster becomes positive) and the expression $\Xi^{1,',(l.)}/\Xi^{1,'}$ transfers into $\Xi^{1,',(l.\mathrm{abs})}/\Xi^{1,',(\mathrm{abs})}$. It can be easily shown that, without loss of generality we can suppose that $\Xi^{1,',(l.)}\geqslant 0$. Obviously,

$$|\Xi^{1,',(l.)}/\Xi^{1,'}|\leqslant \Xi^{1,',(l.\mathrm{abs})}/\Xi^{1,',(\mathrm{abs})}.$$

Now the expression $\Xi^{1,',(l.\mathrm{abs})}/\Xi^{1,',(\mathrm{abs})}$ can be interpreted as a ''absolute probability'' P^{abs} (Long) of the event that there is at least one long cluster.

Now our aim is to estimate the ''absolute probability'' P^{abs} of the event that a given segment belongs to the support of long cluster. In other words, we are going to estimate the statistical weights of long clusters after replacing of the values of all negative bonds in configurations containing at least one long cluster with their absolute values.

Let $\varphi(I_V-I)$ be an arbitrary subconfiguration which contains contours $K_1,\ldots,K_l$, belonging to long clusters, $\mathbf{K}=\cup_1^l \mathrm{supp}^1 K_i$, $\mathbf{K^1}=\mathbf{K}\cap[-V,-(|I|/2)]$ and $\mathbf{K^2}=\mathbf{K}\cap[|I|/2,V]$.

Put $C^1(\varphi(I_V-I))=|\mathbf{K^1}|$ and $C^2(\varphi(I_V-I))=|\mathbf{K^2}|$. We have

$$\begin{aligned}|P(\mathrm{Long})|&=|\Xi^{1,',(l.)}/\Xi^{1,'}|\\ &\leqslant P^{\mathrm{abs}}(\mathrm{Long})\\ &=\sum w^{\mathrm{abs}}(D_1)\cdots w(D_m)/\Xi^{1,',(\mathrm{abs})}\\ &=\sum^{p,1} w^{\mathrm{abs}}(D_1)\ldots w^{\mathrm{abs}}(D_m)/\Xi^{1,',(\mathrm{abs})}+\sum^{p,2} w^{\mathrm{abs}}(D_1)\ldots w^{\mathrm{abs}}(D_m)/\Xi^{1,',(\mathrm{abs})}\\ &=P^{\mathrm{abs}}(\mathrm{Long},>p)+P^{\mathrm{abs}}(\mathrm{Long},\leqslant p),\end{aligned}$$

where $w^{\mathrm{abs}}(D_i)=|w(D_i)|$ for all clusters belonging to the configuration containing at least one long cluster and $w^{\mathrm{abs}}(D_i)=w(D_i)$ for other clusters [note that the statistical weight $w^{\mathrm{abs}}(D_i)$ of fixed cluster in one configuration can be positive, in other negative], last two summations are taken over all nonordered compatible collections of clusters $[D_1,\ldots,D_m]$ containing at least one long cluster, $\cup_{i=1}^m \mathrm{supp}\, D_i\subset I_V-I$ corresponding to the boundary conditions $\{\varphi^1(x),x\in\mathbf{Z}^1-I_V;\varphi'(x),x\in I\}$, the summation in $\Sigma^{p,1}$ is taken over all configurations $\varphi(I_V):\varphi(I)=\varphi'(I)$; $2C^1(\varphi(I_V-V))/(|I_V|-|I|)>p$; $2C^2(\varphi(I_V-V))/(|I_V|-|I|)>p$, the summation in $\Sigma^{p,2}$ is taken over all configurations $\varphi(I_V):\varphi(I)=\varphi'(I)$; $2C^1(\varphi(I_V-V))/(|I_V|-|I|)\leqslant p$; $2C^2(\varphi(I_V-V))/(|I_V|-|I|)\leqslant p$. It means that the density of contours belonging to long clusters in each configuration from $\Sigma^{p,1}$ ($\Sigma^{p,2}$) in both segments $[-V,-(|I|/2)]$ and $[|I|/2,V]$ is greater than p (is not greater than p).

We fixed the value of p as $1-q/2l$, where the values of q and l will be defined in the proof of Lemma 9.

It turns out that the long clusters are negligible.

Lemma 8: For each fixed interval I there exists a value of V_0, such that if $V>V_0$

$$P^{\mathrm{abs}}(\mathrm{Long})=P^{\mathrm{abs}}(\mathrm{Long},>p)+P^{\mathrm{abs}}(\mathrm{Long},\leqslant p)<1/2. \tag{22}$$

Lemma 8 is a consequence of the following two lemmas.

Lemma 9: For each fixed interval I there exists a value of V_0, such that if $V>V_0$

$$P^{\mathrm{abs}}(\mathrm{Long},>p)<1/4.$$

Lemma 10: For each fixed interval I there exists a value of V_0, such that if $V>V_0$

$$P^{\mathrm{abs}}(\mathrm{Long},\leqslant p)<1/4.$$

Proof of Lemma 9: Consider the partition of $\mathbf{Z}^1$ into segments $T_k = T_k(lp)$, where $T_k(lp)$ is the segment with the center at $x=(lp/2)+klp$ and with the length lp (T_k consists of l segments I_k with the length p, where p is the period of the special ground state). The value of l will be defined later. Let us consider an arbitrary configuration $\varphi(x)$. We say that a segment I_k is regular, if I_k does not belong to the support of some long cluster. We say that a segment T_k is super-regular, if T_k contains at least one regular segment.

Let $\mathbf{P_V}$ be a Gibbs measure corresponding to the boundary conditions $\varphi^1(x)$, $x\in\mathbf{Z^1}$, $\varphi'(I)$, $x\in I$.

Let the segment $I_V - I$ consist of n segments T_k; $k=1,...,n$.

We define a sample space Ω consisting of 2^n elementary events $A^j=[\sigma(1),\ldots,\sigma(n)]$, where $\sigma(k)$, $k=1,...,n$ takes two values: $\sigma(k)=0$ corresponds to the case when the segment T_k is super-regular and $\sigma(k)=1$ corresponds to the case when the segment T_k is not super-regular. On the sample space Ω we define two different probability spaces $(\Omega,\mathbf{P_1})$ and $(\Omega,\mathbf{P_2})$ by the following formulas:

$$\mathbf{P_1}(A^j)=\mathbf{P_1}[\sigma(1),\ldots,\sigma(n)]=\mathbf{P_V}[\sigma(1),\ldots,\sigma(n)],$$

where $\mathbf{P_V}$ is the Gibbs distribution $\mathbf{P_V}$, corresponding to the boundary conditions $\varphi^1(x)$, $x\in\mathbf{Z^1}$, $\varphi'(I)$, $x\in I$ and

$$\mathbf{P_2}(A^j)=\mathbf{P_2}[\sigma(1),\ldots,\sigma(n)]=q^{n-s}(1-q)^s,$$

where s denotes the total number of 1 entries of the vector $A^j=[\sigma(1),\ldots,\sigma(n)]$.

We define a random vector $(\eta(1),\eta(2),\ldots,\eta(n))$ on the probability space $(\Omega,\mathbf{P_1})$ and, respectively, a random vector $(\xi(1),\xi(2),\ldots,\xi(n))$ on the probability space $(\Omega,\mathbf{P_2})$ by the formulas:

$$\eta(k)(A^j)=\sigma(k) \quad \text{and} \quad \xi(k)(A^j)=\sigma(k)\,.$$

The random variables $\eta(k)$ and $\xi(k)$ are defined on the same sample space but on different probability spaces.

Due to the definitions, the random variables $\eta(k)$ are dependent, and the random variables $\xi(k)$ are independent and identically distributed.

Consider the two sums $\Sigma_{k=1}^n \eta(k)$ and $\Sigma_{k=1}^n \xi(k)$.

Suppose that

$$\mathbf{P}(\eta(m)=1|\text{any conditions outside } T_m)\leqslant 1-q. \tag{23}$$

Note that $\mathbf{P}(\eta(m)=1|\text{any conditions outside} T_m)\leqslant 1-q=\mathbf{P}(\xi(m)=1)$ and therefore the following natural lemma holds.

Lemma 11:

$$\mathbf{P}\left(\sum_{k\in K}\eta(k)\geqslant l\right)\leqslant\mathbf{P}\left(\sum_{k\in K}\xi(k)\geqslant l\right)$$

for all natural values of l.

The proof of the probabilistically clear Lemma is omitted. For the detailed proof see the Proposition in Ref. 15.

The random variables $\xi(k)$ are independent and identically distributed. The mathematical expectation of $\xi(k)$ equals $1-q$.

Now we show that

$$P^{\text{abs}}(\eta(m)=1|\text{any conditions outside } T_m)\leqslant 1-q. \tag{24}$$

Let $\mathbf{P_V}$ be a Gibbs measure corresponding to arbitrary boundary conditions and T_k be an arbitrary segment. Consider the set of all configurations on the interval T_k and the restriction of the measure $\mathbf{P_V}$ on this set. We show that at some value of l the ''absolute probability'' P^{abs} that in T_k there is at least one regular segment I_k is greater than $q>0$ for some constant q not depending on k. The event $\eta(k)=1$ means that all segments belonging to T_k are nonregular.

Suppose that a fixed configuration $\varphi'(T_m)$ does not coincide with the ground state at all $I_i \in T_m$.

The Peierls argument method directly imply that for some positive constant t_0

$$P^{\text{abs}}(\varphi'(T_m)|\text{conditions outside } T_m \text{ are } \varphi^{gr}(x)) \leqslant \exp(-\beta t_0 l).$$

Note that when we increase the value of l the influence of the conditions outside T_m on the configuration in T_m increases with the rate less than l and therefore at some value of l and for some positive constant t we have

$$P^{\text{abs}}(\varphi'(T_k)|\text{any conditions outside } T_m) \leqslant \exp(-\beta t l) \leqslant 1-q_0.$$

Thus, the probability $P^{\text{abs}}(\eta(m)=1|\text{any conditions outside } T_m)$ as a union of at most 2^{lp} events with probabilities less than $1-q_0$, is bounded by some number $1-q$. The inequality (24) is proved.

Now Lemma 9 is a direct consequence of the strong law of large numbers for $\xi(k)$ and the Lemma 11. Indeed, consider independent Bernoulli trials when the probability of success at each trial is $1-q$. According to the law of large numbers, the probability of the event that the density of successes exceeds $1-q'$; $0<q'<q$, is less than 1/4, when V tends to infinity. It means that the ''absolute probability'' of the event that the density of non-super-regular segments T_k is greater than $1-q'$ is less than 1/4. Due to Lemma 11, this probability is greater than the P^{abs} probability of the event that the density of non-super-regular segments T_m is greater than $1-q'$. In other words, the P^{abs} probability of the event that the density of super-regular segments T_m is less than $1-q'$ is less than 1/4. Thus, the P^{abs} probability of the event that the density of super-regular segments T_m is greater than $1-q'$ is greater than 1/4. Taking into account that each super-regular segment T_m contains at least one regular segment, one can see that the last statement implies the Lemma 9 if the parameter p is chosen from the open interval $(1-q'/l,1)$. We choose the value of p as $1-q/2l$.

Lemma 9 is proved.

Proof of Lemma 10: Let us consider the set of all long clusters D_i with the density of supports less than p. Let $\text{supp}(D)=\cup_{i=j}^{r}\text{supp}(K_j)$. These supports K_i are connected between themselves and with the boundary. Since the density of supports is not greater than $p<1$, the sum of the lengths of bonds in both halves $[-V,-|I|/2$ and $[|I|/2,V]$ is not less than $(V-|I|/2)(1-p)$. When V goes to infinity the sum of lengths of bonds of any long cluster with the density less than p tends to infinity. As it becomes apparent from the proof of Lemma 8 $P^{\text{abs}}(\text{Long},>p)$ does not exceed one. And it does not exceed one, if we omit the factor $g(x,y)$ corresponding to the long bond and since $g(x,y,\varphi)=\exp(-\beta f(x,y,\varphi))-1$ [see (12)] the impact of these bonds tends to zero. By choosing the appropriate value of V we complete the proof of Lemma 10.

Lemma 10 is proved.

We omit the huge proof of Lemma 5 since it is absolutely analogous to the proof of Lemma 6. The only difference is the fact that in $\Xi^{1',2''}$ overlapped clusters are allowed, so the density of nonregular segments of typical configurations in Lemmas 8,9 instead of p will be a number less than $1-(1-p)(1-p)$.

Partition functions including only non-long-super clusters satisfy the following key lemma which has a geometrically-combinatorial explanation.

Lemma 12:

$$\Xi^{1',2'',(n.l.)}=Q\,\Xi^{1'',2',(n.l.)}$$

where the factor $Q = Q(\varphi^1(x), \varphi^2(x), \varphi'(x), \varphi''(x))$ *is uniformly bounded:* $0 < \mathrm{const}_1 < Q < \mathrm{const}_2$.

The factor appears due to the fact that configurations with minimal energy corresponding to the different boundary conditions do not coincide everywhere (they coincide to within shifts, everywhere but finite area).

Proof of Lemma 12: Due to the constant Q without loss of generality we assume that the configurations with minimal energy $\varphi_{\min}$ for both boundary conditions coincide.

According to the definitions and Lemma 4

$$\Xi^{1,'',2',(n.l.)} = Q' \sum^{*} w(SD_1) \cdots w(SD_m),$$

where the summation is taken over all nonlong, nonordered compatible collections of super clusters.

According to the definition of the super cluster

$$Q' \sum^{*} w(SD_1) \cdots w(SD_m) = Q' \sum^{1,',*} w(D_1) \cdots w(D_k) \sum^{2,'',*} w(D_1) \cdots w(D_l)$$

in $\Sigma^{1,',*}$ and $\Sigma^{2,'',*}$ the summation is taken over all nonordered collections of clusters $w(D_1^{1,'}) \cdots w(D_k^{1,'})$ and $w(D_1^{2,''}) \cdots w(D_l^{2,''})$ such that their product belongs to Σ^*.

Similarly,

$$\Xi^{1,'',2,',(n.l.)} = Q'' \sum^{**} w(SD_1) \cdots w(SD_m)$$
$$= Q'' \sum^{1,'',**} w(D_1) \cdots w(D_k) \sum^{2,',**} w(D_1) \cdots w(D_l).$$

In order to prove Lemma 12 we put one-to-one correspondence between $\Sigma^* w(SD_1) \ldots w(SD_m)$ and $\Sigma^{**} w(SD_1) \ldots w(SD_m)$.

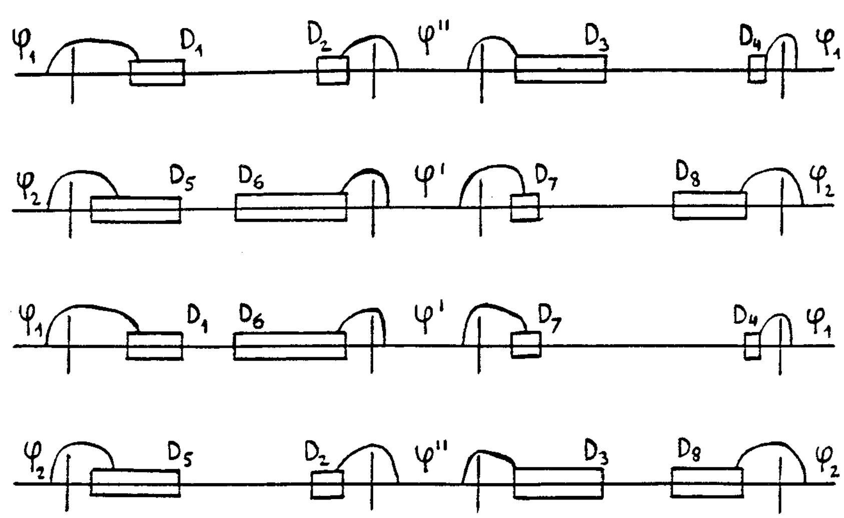

FIG. 1.

Let us consider an arbitrary term $U=w(SD_1)\ldots w(SD_a)$ of Σ^*. By definitions

$$U=w(D_1^{1,\prime})\cdots w(D_m^{1,\prime})w(D_1^{2,\prime\prime})\cdots w(D_k^{2,\prime\prime}),$$

where the factor $w(D_1^{1,\prime})\cdots w(D_m^{1,\prime})$ belongs to the $\Sigma^{1,\prime,*}$ and the factor $w(D_1^{2,\prime\prime})\cdots w(D_k^{2,\prime\prime})$ belongs to the $\Sigma^{2,\prime\prime,*}$.

A cluster $D=[(K_i,i=1,...,r);G'']$ is said to be basic, if the set $((\cup_{i=1}^m \mathrm{supp}\, D_i)\cup G'')\cap((\mathbf{Z}^1-I_N)\cup I))$ is not empty. In Fig. 1 all clusters are basic.

Consider the set of all clusters $\mathbf{W}(U)$ of the term U: $\mathbf{W}(U)=\cup_{i=1}^m D_i^{1,\prime\prime}\cup_{i=1}^k D_i^{2,\prime}$ and four subsets of $\mathbf{W}(U)$:

$$W'=\left[D'=[(K_i,i=1,...,r);G']\in\Xi^{2,\prime}:\left(\left(\bigcup_{i=1}^{k}\mathrm{supp}\, D_i\right)\cup G'\right)\cap I \text{ is not empty}\right],$$

$$W''=\left[D''=[(K_i,i=1,...,r);G']\in\Xi^{1,\prime\prime}:\left(\left(\bigcup_{i=1}^{m}\mathrm{supp}\, D_i\right)\cup G'\right)\cap I \text{ is not empty}\right],$$

$$W^1=\left[D^1=[(K_i,i=1,...,r);G']\in\Xi^{1,\prime\prime}:\left(\left(\bigcup_{i=1}^{m}\mathrm{supp}\, D_i\right)\cup G'\right)\cap(\mathbf{Z}^1-I_N) \text{ is not empty}\right],$$

$$W^2=\left[D^2=[(K_i,i=1,...,r);G']\in\Xi^{2,\prime}:\left(\left(\bigcup_{i=1}^{k}\mathrm{supp}\, D_i\right)\cup G'\right)\cap(\mathbf{Z}^1-I_N) \text{ is not empty}\right].$$

Note that the subsets W',W'',W^1,W^2 contain only basic clusters and the union of them contain all basic clusters of the term U.

Let us consider an arbitrary term $U=w(SD_1)\cdots w(SD_b)$ of Σ^{**}. By the definitions

$$U'=w(D_1^{1,\prime\prime})\cdots w(D_l^{1,\prime\prime})w(D_1^{2,\prime})\cdots w(D_n^{2,\prime}),$$

where the factor $w(D_1^{1,\prime\prime})\cdots w(D_m^{1,\prime\prime})$ belongs to the $\Sigma^{1,\prime\prime,**}$ and the factor $w(D_1^{2,\prime})\cdots w(D_k^{2,\prime})$ belongs to the $\Sigma^{2,\prime,**}$.

Consider the set of all clusters $W(U')$ of the term $U':W(U')=\cup_{i=1}^m D_i^{1,\prime\prime}\cup_{i=1}^k D_i^{2,\prime}$. In just the same way we can define four subsets of $W(U')$.

Consider a term $U=w(D_1)\cdots w(D_k)\in\Sigma^*$, containing only basic clusters. By definition $\cup_{i=1}^k D_i$ can be represented as $\cup_{i=1}^k D_i=(\cup_{i=1}^m D_i)\cup(\cup_{i=m+1}^k D_j)$, where the clusters $\cup_{i=1}^m D_i =W^1\cup W'$; and $\cup_{i=m+1}^k D_j=W^2\cup W''$.

From the definition of nonlong clusters and W',W'',W^1,W^2 it easily follows that there exists the same term $U'=w(D_1)\ldots w(D_k)\in\Sigma^{**}$, such that $\cup_{i=1}^k D_i=(\cup_{i=1}^m D_i)\cup(\cup_{i=m+1}^k D_j)$, where the clusters $\cup_{i=1}^m D_i=W^1\cup W'$; and $\cup_{i=m+1}^k D_j=W^2\cup W''$.

Figure 1 shows four collections of clusters $\mathrm{COL}_1=[D_1^{1,\prime\prime},D_2^{1,\prime\prime},D_3^{1,\prime\prime},D_4^{1,\prime\prime}]$, $\mathrm{COL}_2 =[D_5^{2,\prime},D_6^{2,\prime},D_7^{2,\prime},D_8^{2,\prime}]$, $\mathrm{COL}_3=[D_1^{1,\prime},D_6^{1,\prime},D_7^{1,\prime},D_4^{1,\prime}]$, $\mathrm{COL}_4=[D_5^{2,\prime\prime},D_2^{2,\prime\prime},D_3^{2,\prime\prime},D_8^{2,\prime\prime}]$.

Two coincident terms $U=U'=\Pi_{i=1}^8 w(D_i)$ belonging to the sums Σ^* and Σ^{**} are constructed by the Cartesian product of the collections COL_1, COL_2, and COL_3, COL_4, respectively.

We see that between terms $U\in\Sigma^*$ and $U'\in\Sigma^{**}$ containing only basic clusters we easily can put a one-to-one correspondence.

Consider a term $U=w(D_1)\cdots w(D_k)w(D_{k+1})\cdots w(D_n)\in\Sigma^*$, containing basic clusters $D_1\cdots D_k$ and not basic clusters $D_{k+1}\cdots D_n$.

It can be easily shown that there exists a term $U'=w(D_1)\cdots w(D_k)w(D_{k+1})\cdots w(D_n)\in\Sigma^{**}$ coinciding with the term $U\in\Sigma^*$. Indeed, suppose that there is no term U'

$=w(D_1)\cdots w(D_k)w(D_{k+1})\cdots w(D_n)\in\Sigma^{**}$ coinciding with the term $U\in\Sigma^*$. Then, according to the definition of the long clusters, we directly get that, the term U contains long super cluster, which contradicts the definition of Σ^*.

Lemma 12 is proved.

Remark: The essential point of the proof of the important Lemma 12 (therefore, of this paper) is the amusing fact that $\Sigma^* w(SD_1)\cdots w(SD_m)$ and $\Sigma^{**} w(SD_1)\cdots w(SD_m)$ coincide.

Now the demanded inequality (18) is a direct consequence of Lemmas 5 and 12. The inequality (18), therefore Lemma 3 is proved.

Let $\mathbf{P^1}$ and $\mathbf{P^2}$ be two different extreme Gibbs states of the model (1) corresponding to the boundary conditions $\varphi^1(x)$ and $\varphi^2(x)$, respectively.

Theorem 5: (Ref. 16.) $\mathbf{P^1}$ *and* $\mathbf{P^2}$ *are singular or coincide*.

Proof of Theorem 4: Let $\mathbf{P^1}$ and $\mathbf{P^2}$ be two different extreme Gibbs states of the model (1) corresponding to the boundary conditions $\varphi^1(x)$ and $\varphi^2(x)$ respectively. According to Lemma 3 $\mathbf{P^1}$ and $\mathbf{P^2}$ are not singular. Therefore, according to Theorem 5 $\mathbf{P^1}$ and $\mathbf{P^2}$ coincide, which contradicts the assumption. Theorem 4 is proved.

III. UNIQUENESS OF GIBBS STATES

In this section we prove the main Theorem 1.

The statement of Theorem 1 for rational densities coincides with Theorem 4. Thus, in order to complete the proof of Theorem 1, we have to prove the following theorem, which covers the case when the density of the special ground state is irrational.

Theorem 6: *Suppose that the value of the external field* μ *of the model (1) belongs to the set* $\mathbf{C^{ir}}=\mathbf{R^1}-\cup_\kappa(\mu_\kappa^-,\mu_\kappa^+)$. *Then the model (1) has a unique Gibbs state at all values of the temperature* β^{-1}.

It can be easily shown that the special ground states of the model (1) are not stable when the density is irrational. In other words, the Peierls constant t for the special ground state tends to zero, when $p\to\infty$. The essence of this fact is the following.

For the fixed irrational number $\eta=[n_0,n_1,\ldots,n_s,\ldots]$ consider the corresponding special ground state $\varphi_\kappa(x)$ and its arbitrary perturbation $\varphi_\kappa'(x)$. The configuration $\varphi_\kappa'(x)$ is not a special ground state, therefore for some pair of points, say x and $y\in\mathbf{Z^1}$; $\varphi_\kappa'(x)=\varphi_\kappa'(y)=1$, we have a violation of Hubbard's criterion. Let x and y be closest points with this property. When the distance between x and y tends to infinity, the Peierls constant tends to zero.

In the irrational case the special ground states are not stable, but this fact is not crucial for our method. Since the essence of our method is the estimation of long super clusters connecting the boundary with the segment I, small clusters not satisfying Peierls condition cannot ''help'' to connect the boundary with I, and it turns out that big clusters satisfy the Peierls stability condition and the method works. One can say that the special ground states in the irrational case are ''stable in general.''

Below we give the mathematical details of the last observation.

Consider $\eta(s)=[n_0,n_1,\ldots,n_s]$.

Lemma 13: Suppose that the value of the external field μ *of the model (1) belongs to the interval* $(\mu_{\kappa(s)}^-,\mu_{\kappa(s)}^+)$ *for some number* $\kappa(s)=\eta(s)^{-1}$. *Let* $\varphi'(x)$ *be an arbitrary finite perturbation of the special ground state* $\varphi_{\kappa(s)}(x)$ *such that the boundary B of the configuration* $\varphi'(x)$ *includes a unique contour K. Then there exists a positive constant* t_s *depending only on the Hamiltonian (1), such that*

$$H(\varphi'(x))-H(\varphi_{\kappa(s)}(x))\geq t_s|\mathrm{supp}\,B|$$

where $|\mathrm{supp}\,B|$ *is the total area of the support of the boundary*.

Lemma 13 was proved in Ref. 13 [see Lemma 1 and Sec. 5 (Ref. 13)].

Thus, for each nonnegative integer s the number t_s is defined. Suppose that a positive number t less than t_1 is fixed. Let s be the maximal number meeting the condition $t_s>t$.

Now we are ready to define the notion of a contour in the irrational case.

Let us consider an arbitrary configuration $\varphi(x)$. Let $C=\cup_{i\in \mathrm{Ind}}[x_i,y_i]$, where $x_i,y_i\in \mathbf{Z}^1$ and $x_i\neq y_i$ has the following properties:

(1) For each segment $[a_i,b_i]$ from the set $\mathbf{Z}^1-C$ there exists a special ground state φ_κ, such that the restriction of this configuration on $[a_i,b_i]$ coincides with $\varphi([a_i,b_i])$.

(2) For any $C'\subset C$; $C'\neq C$ the property 1 is not held.

It can be easily shown that the set $C=C(\varphi(x))$ is not uniquely defined. Suppose that, some rule uniquely determines the set C for each configuration $\varphi(x)$. Let $\mathbf{Z}^1-C=\cup_i[a_i,b_i]$. We say that $\varphi([a_i,b_i])$; is a preregular phase. Consider any segment $[x_i,y_i]$ belonging to C. The segment $[x_i,y_i]$ is said to be t-negligible, if for each segment $[v_i,w_i]$ covering $[x_i,y_i], w_i-v_i=p$ [p is the numerator of $\eta(s)$] there exists a special ground state $\varphi_{\kappa(s)}$, such that the restriction of this configuration on $[v_i,w_i]$ coincides with $\varphi([v_i,w_i])$. Let $C=\cup_{i\in \mathrm{Ind}}[x_i,y_i]=(\cup_{i\in \mathrm{Ind}(t)}\times[x_i,y_i])\cup(\cup_{i\in \text{Ind-Ind}(t)}[x_i,y_i])$, where $\mathrm{Ind}(t)$ means that the union is taken over all t-negligible segments. The support of the preboundary $\mathrm{supp}\,PB$ of the configuration $\varphi(x)$ will be defined as $\mathrm{supp}\,PB=(\cup_{i\in \mathrm{Ind}(t)}[x_i,y_i])\cup(\cup_{i\in \text{Ind-Ind}(t)}[x_i-d_0p,y_i+d_0p])=\mathrm{supp}\,PB(\mathrm{main})\cup\mathrm{supp}\,PB(t)$. Each segment belonging to the union $\mathrm{supp}\,PB$ will be called a support of a precontour and is denoted by $\mathrm{supp}\,PK$. The support $[x_i,y_i]$ of a precontour is said to be t-negligible, if $[x_i,y_i]$ belongs to $\mathrm{supp}\,PB(t)$.

We define contours as in the Definition 1. The constants p,d_0 and N_b for irrational density η^{-1} will be constants defined for rational density $\eta(s)^{-1}$.

The pair $PK=(\mathrm{supp}\,PK,\varphi'(\mathrm{supp}\,PK))$ is called a precontour. The set of all precontours is called a preboundary PB of the configuration $\varphi'(x)$. Two precontours PK_1 and PK_2 are said to be connected if $\mathrm{dist}(\mathrm{supp}\,PK_1,\mathrm{supp}\,PK_2)<N_b$ and at least one of them is not t-negligible. The set of precontours $(PK_i;i\in \mathrm{Ind})$ is called connected if for any two precontours PK_c and $PK_d;c,d\in \mathrm{Ind}$ there exists a collection $(PK_{j_1}=PK_c,\ldots,PK_{j_i},\ldots,PK_{j_{n-1}},PK_{j_n}=PK_d)$; $j_i\in \mathrm{Ind}$, $i=1,...,n$; such that any two precontours PK_{j_i} and $PK_{j_{i+1}}$, $i=1,...,$ $n-1$ are connected. Let $\cup_{i=1}^n PK_i$ be some maximal connected component of the preboundary PB. Suppose that $\mathrm{supp}\,PK_i=[a_i,b_i]$ and $b_i<a_{i+1}$; $i=,\ldots,n-1$.

The pair $K=(\mathrm{supp}\,K,\varphi'(\mathrm{supp}\,PK))$, where $\mathrm{supp}\,K=[a_1,b_n]$ is called a contour. The set of all contours is called a boundary B of the configuration $\varphi'(x)$.

A contour is said to be t-negligible, if its support is t-negligible.

By the definitions, the distance between the supports of two t-negligible contours exceeds p, where p is the numerator of $\eta(s)$ and the length of the support of any t-negligible contour is one.

The following lemma is reformulation of Lemma 13 for irrational densities.

Lemma 14: Suppose that the value of the external field μ *of the model (1) belongs to the set* $\mathbf{C}^{\mathbf{ir}}=\mathbf{R}^1-\cup_\kappa(\mu_\kappa^-,\mu_\kappa^+)$. *Let* $\varphi'(x)$ *be an arbitrary finite perturbation of the special ground state* $\varphi_\kappa(x)$ *such that the boundary B of the configuration* $\varphi'(x)$ *includes a unique contour (not t-negligible contour) K. Then there exists a positive constant* t_s *depending only on the Hamiltonian (1), such that*

$$H(\varphi'(x))-H(\varphi_\kappa(x))\geqslant t_s|\mathrm{supp}\,B|$$

where $|\mathrm{supp}\,B|$ *is the total area of the support of the boundary.*

Suppose that the value of the external field μ of the model (1) belongs to the set $\mathbf{C}^{\mathbf{ir}}=\mathbf{R}^1-\cup_\kappa(\mu_\kappa^-,\mu_\kappa^+)$. Let $t,0<t<t_1$ is fixed and t_s is chosen as above.

Lemma 15: Let $\varphi_{\min}(x)\in\Phi(V)$ *be a configuration with the minimal energy:*

$$H(\varphi_{\min}(x)|\varphi^1(x))=\min_{\varphi(x)\in\Phi(V)}H(\varphi(x)|\varphi^1(x)).$$

Then the configuration $\varphi_{\min}(x)$ *has the following structure:*

The restriction of the configuration $\varphi_{\min}(x)$ *on the set* $[-V+N_b,V-N_b]$ *contains t-negligible contours and* $p-1$ *non t-negligible contours, moreover the sum of weights of all t-negligible* *contours is bounded by constant, not depending on the boundary conditions, all of* $p-1$ *non t-negligible contours are interface contours* IK_i, $i=1,...,m$, *where* $m<p-1$ *and* $|\mathrm{supp}\, IK_i| < 3d_0p+N_b$.

The proof of Lemma 15 is very similar to the proof of Lemma 1[13] and will be omitted.

From Lemma 15 follows that the density of possible *t*-negligible contours of $\varphi_{\min}(x)$ tends to zero, when V goes to infinity.

Now the proof of Theorem 6 principally coincides with the proof of Theorem 3 and will be omitted. Theorem 6, and hence main Theorem 1 is proved.

IV. FINAL REMARKS

The unique limit Gibbs state of the model (1) is translationally invariant. This result was proved independently in Ref. 1 by using of the method of the equivalence of boundary conditions,[17] and in Ref. 11 by using of energy–entropy inequalities.

At low temperatures, the sum of the statistical weights of all clusters having fixed support has an exponential estimation (see Lemma 16, Ref. 13) and each limit Gibbs state of the model (1) is a ''small perturbation of special ground states'' (see Lemma 17, Ref. 13).

The essential points in the proof of the uniqueness of Gibbs states are the geometrically combinatorial Lemma 12 and the estimation of long super clusters, connecting the boundary with the segment *I*. This estimation mainly works due to the fact that ground states of the model (1) degenerate. In Ref. 13 we proved Theorem 4 at low temperatures. The temperature restriction was related with the fact that at low temperatures the weight of the support of a cluster has an exponential estimation [Lemmas 16 and 17 (Ref. 13)] and hence long clusters are negligible (Ref. 13). But at any temperature an exponential estimation is absent. In the general case, when we estimate the statistical weight of long super clusters, a key role plays the Lemma 6 on the estimation of the value of the interaction between contours.

In Ref. 15 at low temperatures the result of Ref. 13 is extended to more abstract models. The method of the proof of Theorem 1 shows that the result of Ref. 15 can be extended to all values of the temperatures.

[1] S. E. Burkov and G. Ya. Sinai, Usp. Mat. Nauk **38**, 205 (1983).
[2] R. L. Dobrushin, Teor. Veroyatn. Primen. **18**, 201 (1968).
[3] R. L. Dobrushin, Funk. Anal. Pril. **2**, 44 (1968).
[4] D. Ruelle, Commun. Math. Phys. **9**, 267 (1968).
[5] F. Dyson, Commun. Math. Phys. **12**, 91 (1969).
[6] F. Dyson, Commun. Math. Phys. **21**, 269 (1971).
[7] J. Hubbard, Phys. Rev. B **17**, 494 (1978).
[8] P. L. Pokrovsky and G. V. Uimin, J. Phys. C **11**, 3535 (1978).
[9] P. Bak, ''Incommensurate, commensurate and chaotic phases,'' preprint, Nordita, 1982.
[10] P. Bak and R. Bruinsma, Phys. Rev. Lett. **49**, 249 (1982).
[11] M. Fannes, P. Vanheuverzwijn, and A. Verbeure, J. Stat. Phys. **29**, 547 (1982).
[12] A. A. Kerimov, Teor. Mat. Fiz. **58**, 473 (1984).
[13] A. A. Kerimov, J. Stat. Phys. **72**, 571 (1993).
[14] J. Bricmont, K. Kuroda, and J. L. Lebowitz, Commun. Math. Phys. **101**, 501 (1985).
[15] A. A. Kerimov, Physica A **258**, 183 (1998).
[16] H.-O. Georgee, *Gibbs Measures and Phase Transitions* (de Gruyter, Berlin, New York, 1988).
[17] J. Bricmont, J. L. Lebowitz, and C. E. Pfister, J. Stat. Phys. **21**, 573 (1979).

The Lanczos algorithm for extensive many-body systems in the thermodynamic limit

N. S. Witte[a)]
Research Centre for High Energy Physics, School of Physics, University of Melbourne, Parkville, Victoria 3052, Australia

D. Bessis
Service de Physique Théorique, Centre d'Etudes Nucléaires de Saclay, F-91191 Gif-Sur-Yvette Cedex, France

(Received 24 March 1999; accepted for publication 5 May 1999)

We establish rigorously the scaling properties of the Lanczos process applied to an arbitrary extensive many-body system which is carried to convergence $n\rightarrow\infty$ and the thermodynamic limit $N\rightarrow\infty$ taken. In this limit the solution for the limiting Lanczos coefficients are found exactly and generally through two equivalent sets of equations, given initial knowledge of the exact cumulant generating function. The measure and the orthogonal polynomial system associated with the Lanczos process in this regime are also given explicitly. Some important representations of these Lanczos functions are given, including Taylor series expansions, and theorems controling their general properties are proven. © *1999 American Institute of Physics.* [S0022-2488(99)01510-8]

I. INTRODUCTION

The Lanczos algorithm is one of the few reliable and general methods for computing the ground state and excited state properties of strongly interacting quantum many-body systems. It has been traditionally employed as a numerical technique on small finite systems, with attendant round-off error problems, although the main obstacle to its further development has been the rapid growth of the number of basis states with system size. The reader is referred to a review of the applications of this method[1] in strongly correlated electron problems. In this work we examine the Lanczos process in the context of the extensive quantum many-body systems, where it is employed entirely in an exact manner and where the thermodynamic limit is taken. So, in complete contrast to the traditional use of the Lanczos algorithm, we completely circumvent the issues of loss of orthogonality due to round-off errors and the inability to approach the thermodynamic limit because of the requirement to construct a full basis on the cluster. The systems we have in mind are those with an infinite number of degrees of freedom, yet are extensive, in that all total averages of any physical quantity scale linearly with the numbers of degrees of freedom however quantified. These would include all condensed matter systems with sufficiently local interactions (the precise conditions need to be clarified, but it is clear which specific systems obey extensivity) and quantum field theories, with the proviso that the spectrum is bounded below (in some cases there is also an upper bound, too).

After noting some of the advantageous features of the algorithm in general we discuss the scaling behavior of the Lanczos process as it approaches convergence and as the thermodynamic limit is taken. Central to this approach is the manifestation of extensivity through a description based on the cumulant generating function (CGF) which we take to be given. We then derive a set of general integral equations which define the scaled Lanczos functions in the thermodynamic limit, which can be explicitly and exactly solved for certain integrable models, or employed in a truncated manner for nonintegrable models. An alternative formulation is also given which ex-

[a)]Electronic mail: nsw@physics.unimelb.edu.au

presses the equivalence of the Lanczos process with the continuum Toda lattice model treated as a boundary value problem. Finally, we state some general results concerning the behavior of the Lanczos functions.

II. THE LANCZOS PROCESS, ORTHOGONAL POLYNOMIALS, AND MOMENTS

The Lanczos algorithm or process[2–4] begins with a trial state $|\psi_0\rangle$ appropriate to the model and the symmetries of the phase being investigated. From this the Lanczos recurrence generates a sequence of orthonormal states $\{|\psi_n\rangle\}_{n=1}^{\infty}$ and Lanczos coefficients $\{\alpha_n(N)\}_{n=0}^{\infty}$ and $\{\beta_n(N)\}_{n=1}^{\infty}$, thus

$$\hat{H}|\psi_n\rangle=\beta_n|\psi_{n-1}\rangle+\alpha_n|\psi_n\rangle+\beta_{n+1}|\psi_{n+1}\rangle, \tag{1}$$

with the Lanczos coefficients being defined

$$\begin{aligned}\alpha_n&=\langle\psi_n|\hat{H}|\psi_n\rangle,\\ \beta_n&=\langle\psi_{n-1}|\hat{H}|\psi_n\rangle.\end{aligned} \tag{2}$$

We distinguish a total or extensive operator or variable such as H from its density or intensive counterpart by $\hat{H}$. In this basis the transformed Hamiltonian takes the following tridiagonal form

$$T_n=\begin{pmatrix}\alpha_0 & \beta_1 & & & & \\ \beta_1 & \alpha_1 & \beta_2 & & & \\ & & \ddots & & & \\ & & & \beta_{n-1} & \alpha_{n-1} & \beta_n \\ & & & & \beta_n & \alpha_n\end{pmatrix}. \tag{3}$$

As such the Lanczos process is one of the Krylov subspace methods[5] in that at a finite step n, the eigenvectors belong to the Krylov subspace Span$\{|\psi_0\rangle,\hat{H}|\psi_0\rangle,\hat{H}^2|\psi_0\rangle,\ldots,\hat{H}^n|\psi_0\rangle\}$.

In the many-body context one would iterate the Lanczos process until termination whereupon the Hilbert space is exhausted (at this point one of the $\beta_{n_T}=0$, where n_T is the dimension of the Hilbert space in the sector defined by the ground state), or until the process has converged according to some arbitrary criteria $n\to n_C$. Then one would perform the thermodynamic limit $N\to\infty$ where it should be understood that the above conclusion of the Lanczos process is also dependent on the system size, that is to say, $n_T(N),n_C(N)$. These cutoffs are monotonically increasing functions of the system size so they will all tend to ∞ in the thermodynamic limit as well. Taking the limits in the reverse order clearly leads to nonsensical results, as taking $N\to\infty$ with n fixed produces $\alpha_n\to c_1$ and $\beta_n\to 0$. The great virtue of the Lanczos process is that it can be shown to converge essentially exponentially fast with respect to iteration number, using the Kaniel–Paige–Saad exact bounds[6–8] for the rate of convergence. This means that convergence occurs within a very small subspace of the total Hilbert space, so that $n_C\ll n_T$.

The Lanczos process is entirely equivalent to the three-term recurrence for an orthogonal polynomial system (OPS),[9–11] however, we consider a slight generalization of the preceding process to one with a single parameter evolution (a ''time'' t). In this construction we are continuing a development begun by Lindsay[12] and Chen and Ismail,[13] which will lead to some powerful tools in treating the Lanczos process. The measure, or that component which is absolutely continuous, is defined by the weight function

$$w(\epsilon,t)=e^{-u(\epsilon)+tN\epsilon}, \tag{4}$$

on the real line $\epsilon \in \mathbb{R}$. Our system under study is described by the initial value of the system at $t=0$ and often we will suppress this argument for the sake of simplicity. This measure defines a system of monic orthogonal polynomials $\{P_n(\epsilon,t)\}_{n=0}^{\infty}$ with an orthogonality relation

$$\int_{-\infty}^{+\infty} d\epsilon\, w(\epsilon,t) P_m(\epsilon,t) P_n(\epsilon,t) = h_n(t)\,\delta_{mn}, \tag{5}$$

and normalization $h_n(t)$. This is equivalent to the following three-term recurrence relation

$$P_{n+1}(\epsilon,t) = (\epsilon - \alpha_n(t)) P_n(\epsilon,t) - \beta_n^2(t) P_{n-1}(\epsilon,t), \tag{6}$$

with the recursion coefficients $\alpha_n(t)$ real for $n \geqslant 0$ and $\beta_n^2(t)$ real and positive for $n > 0$. By convention we take $\beta_0^2 = 1$. It can be readily shown that the Lanczos coefficients are given in terms of the normalization, thus

$$\begin{aligned} \alpha_n(t) &= \frac{1}{N h_n(t)} \frac{d}{dt} h_n(t), \\ \beta_n^2(t) &= \frac{h_n(t)}{h_{n-1}(t)}. \end{aligned} \tag{7}$$

The direct connection between the Lanczos process and the OPS is given by the determinant relation of the characteristic polynomial

$$P_{n+1}(\epsilon) = (-)^{n+1} |T_n - \epsilon I_{n+1}|, \tag{8}$$

so that the zeros of the orthogonal polynomial are eigenvalues of Hamiltonian.

Some comments are in order regarding the differences, or more accurately the special character, of these orthogonal polynomials with respect to the generic OPS or with some of the scaling versions of the OPS.[14] These OPSs have been termed many-body OPSs, but could be equally described as extensive OPSs. They all have an additional, essential parameter to the generic OPS, the system size N, which appears in both the gross scaling factors (the ''external'' scaling such as in the energy densities ϵ defined by $E = N\epsilon$), but also internally in the three-term recurrence coefficients, in the polynomials themselves and in other derived quantities. The internal dependence in the Lanczos coefficients on the system size is not at all apparent, and the most transparent way that extensive scaling properties can be exhibited is through the cumulant generating Function (CGF), which hitherto has played no role in orthogonal polynomial theory. In fact, the CGF is central to this class of OPS, rather than the moments, and is in a practical sense the starting point in any application of the formalism to physical models. For all models it is clear that the ground state energy E_0 is proportional to N and unbounded in the thermodynamic limit, and similarly the total Lanczos coefficients (as opposed to the densities) are unbounded as $n \to \infty$ for fixed N. When everything is recast in terms of densities, the spectrum is bounded below by ϵ_0 and in many models will also be bounded above, and similarly the density Lanczos coefficients are bounded. Another difference that many-body OPSs exhibit in comparison to general OPS is, as we have noted above, the three-term recurrence will terminate exactly at $n = n_T$, although this will never present any problems as this is exponentially large.

The Lanczos process is intimately connected with the Hamburger moment problem[15,16] via the Resolvent operator

$$R(\epsilon) = \left\langle \frac{1}{\epsilon - \hat{H}} \right\rangle, \quad \epsilon \notin \text{Supp}\,[d\rho]. \tag{9}$$

Its formal Laurent series establishes a direct link with Hamiltonian moments

$$R(\epsilon)=\sum_{i=0}^{\infty}\frac{\mu_i}{\epsilon^{i+1}}, \tag{10}$$

where these moments are defined as expectation values with respect to the trial state referred to above

$$\mu_n \equiv \langle \hat{H}^n \rangle, \quad \mu_0 = 1. \tag{11}$$

The resolvent has a real Jacobi-fraction continued fraction representation[17,18]

$$R(\epsilon)=-\mathbf{K}_{n=0}^{\infty}-\left(\frac{\beta_n^2}{\epsilon-\alpha_n}\right), \tag{12}$$

with elements coming from the Lanczos coefficients.

An equivalent description to that of the Hamiltonian moments is to formulate everything in terms of cumulants or connected moments[19,20] $\{\nu_n\}_{n=1}^{\infty}$, and to ignore all corrections which vanish in the thermodynamic limit $N\to\infty$. Cumulants scale directly with the size of the system so that for the extensive many-body problem we have

$$\nu_n = c_n N + o(1) \tag{13}$$

in the ground state sector, or

$$\nu_n = c_n N + m_n + o(1) \tag{14}$$

in any other sector.[21] This also means that no finite-size scaling can be performed given that only the limiting quantities are retained here and boundary condition effects do not appear. The foundation ingredient is the moment generating function which is related to the cumulant generating function in the following way.

Definition 1: The moment generating function (MGF) $M(t)$ and the cumulant generating functions (CGF) $F(t)$ are defined by

$$M(t)\equiv\langle e^{tH}\rangle=\sum_{n=0}^{\infty}\mu_n\frac{t^n}{n!}=\exp\left(\sum_{n=1}^{\infty}\nu_n\frac{t^n}{n!}\right)\equiv\exp(NF(t)). \tag{15}$$

Some examples of cumulant generating functions include the isotropic *XY* model using the z-polarized Néel state as the trial state,[22]

$$F(t)=\frac{1}{\pi}\int_0^{\pi/2} dq\,\log\cosh(t\cos q), \tag{16}$$

and the Ising model in a transverse field using the disordered state as the trial state, and coupling constant x (Ref. 23),

$$F(t)=\frac{1}{2\pi}\int_0^{\pi} dq\,\ln\left[\cosh(2t\epsilon_q)-\frac{(\cos q+x)}{\epsilon_q}\sinh(2t\epsilon_q)\right], \tag{17}$$

where the quasiparticle energies ϵ_k are defined by $\epsilon_q^2=1+x^2+2x\cos q$.

Definition 2: The determinants of the moment matrices $\Delta_n(t)$ for $n\geqslant 0$ are defined by the Hankel form:

$$\Delta_n(t)=\left|M^{(i+j-2)}(t)\right|_{i,j=1}^{n+1}. \tag{18}$$

The direct relationship from moments to the Lanczos coefficients which is established in this way is via the construction of a sequence of Hankel determinants of the moment matrices and their Selberg-type integral representation[9]

$$\Delta_n(t)=\frac{1}{(n+1)!}\int_{-\infty}^{+\infty}\prod_{k=1}^{n+1} d\epsilon_k\, w(\epsilon_k,t)\prod_{1\leq i<j\leq n+1}|\epsilon_i-\epsilon_j|^2. \tag{19}$$

These determinants are related to the normalizations via

$$\Delta_n(t)=\prod_{j\leq n} h_j(t). \tag{20}$$

Definition 3: Our final definition, that of the Lanczos L-function, is

$$N^2L_n(t)=\frac{\Delta_n(t)\Delta_{n-2}(t)}{\Delta_{n-1}^2(t)}, \tag{21}$$

for $n\geq 1$ *and* $L_0(t)=M(t)$.

The converse result is then

$$\Delta_n(t)=N^{n(n+1)}\prod_{k=0}^{n} L_k^{n+1-k}(t), \tag{22}$$

for $n\geq 1$. From these the Lanczos coefficients are given simply by

$$\alpha_n(t)=\frac{1}{N}\sum_{j=0}^{n}\frac{L_j'(t)}{L_j(t)},\qquad \beta_n^2(t)=L_n(t). \tag{23}$$

Theorem 1: *The equation of motion for the Lanczos L-function is*

$$L_n(t)=\frac{1}{N}\sum_{j=1}^{n}\frac{j}{N}D_t^2\log L_{n-j}(t), \tag{24}$$

with the initial condition on the recurrence given by $\log L_0(t)=NF(t)$ *for all t.*

The advantage of introducing evolution into the Lanczos process is that Sylvester's theorem applied to the Hankel determinants,[24]

$$\Delta_{n+1}(t)\Delta_{n-1}(t)=\Delta_n(t)\Delta_n''(t)-(\Delta_n'(t))^2, \tag{25}$$

so that the theorem follows directly from this. □

The first few members of the Lanczos L-sequence are

$$L_1(t)=\frac{1}{N}F''(t),\qquad L_2(t)=\frac{2}{N}F''(t)+\frac{1}{N^2}\frac{F^{(2)}F^{(4)}-(F^{(3)})^2}{(F^{(2)})^2}. \tag{26}$$

The consequence of Sylvester's theorem for the evolution of the Δ_n is the following theorem.

Theorem 2: *The* $\Delta_n(t)$ *obey the following differential-difference equation,*

$$\exp\{\log \Delta_{n+1}+\log \Delta_{n-1}-2 \log \Delta_n\}=D_t^2 \log \Delta_n, \tag{27}$$

with the boundary value $\log \Delta_0 = NF(t)$ *and conventionally* $\Delta_{-1}=1$.

This follows directly from Sylvester's identity. □

This evolution equation is just the finite Toda lattice equation of motion,[25] and this point has been previously noted in Ref. 13.

III. SCALING IN THE THERMODYNAMIC LIMIT

As was discussed earlier, there are two limiting processes that one must consider when the thermodynamic limit is taken in the Lanczos algorithm, both $n, N\to\infty$, and the issue then is what mutual relationship exists between them in the limit. One can view this limiting process in the $1/n$ vs $1/N$ plane and then consider along what types of paths must one approach the origin. We shall find that the general relationship is $n, N\to\infty$ with $s\equiv n/N$ fixed, although for systems at criticality it seems inevitable that s will become unbounded in the analysis. A consequence of these ideas is the confluence property of the Lanczos coefficients as $n, N\to\infty$ at fixed $s=n/N$.

$$\begin{aligned}\alpha_n(N)&=\alpha(s)+O(1/N),\\ \beta_n^2(N)&=\beta^2(s)+O(1/N).\end{aligned} \tag{28}$$

There are a number of ways to see this approach to the thermodynamic limit.

Using the explicit forms connecting cumulants and moments, and a direct evaluation of the Hankel determinants, one can prove[26] for general n and N that the Lanczos coefficients have a leading order scaling in $s=n/N$ for the first two orders of an expansion in large N. Actually, this expansion is valid for all n not just for large values and thus includes all the subdominant contributions. Thus

$$\alpha_n=c_1N+n\left[\frac{c_3}{c_2}\right]+\frac{1}{2}\,n(n-1)\left[\frac{3c_3^3-4c_2c_3c_4+c_2^2c_5}{2c_2^4}\right]\frac{1}{N}+\cdots \tag{29}$$

for $n\geqslant 0$, and

$$\begin{aligned}\beta_n^2={}&nc_2N+\frac{1}{2}\,n(n-1)\left[\frac{c_2c_4-c_3^2}{c_2^2}\right]+\frac{1}{6}\,n(n-1)(n-2)\\ &\times\left[\frac{-12c_3^4+21c_2c_3^2c_4-4c_2^2c_4^2-6c_2^2c_3c_5+c_2^3c_6}{2c_2^5}\right]\frac{1}{N}+\cdots\end{aligned} \tag{30}$$

for $n\geqslant 1$. However, this approach cannot be generalized to higher orders and therefore for the full exact Lanczos coefficients. The first two terms in the above expansions were also proven by Lindsay using the Sylvester Identity in the statistical context[12] but no further, while this form for the higher terms (but finite numbers) was conjectured in Ref. 27. We shall find that use of the Sylvester identity allows one to very easily recover this result, to in fact go to much higher orders in constructing explicit forms, and to prove this type of scaling in a completely general way.

Lemma 1: The Lanczos L-function $L_n(t,N)$ *is a rational function of* $1/N$ *for fixed n, and all t.*

The difference-differential equation (24) is of finite order in j/N and t, so the result follows. □

Also for fixed n we have

$$\lim_{N\to\infty} L_n(t,N)=0, \tag{31}$$

and specifically the leading order term is $O(N^{-1})$ which arises from the $j=n$ term in the sum. Therefore we can expand this function in a descending series in N^{-1}, thus

$$L_n(t,N)=\sum_{p\geqslant 1}\frac{l_{np}(t)}{N^p}, \tag{32}$$

and defining the connected series related by

$$\sum_{p\geqslant 1}\frac{m_{np}(t)}{N^p}\equiv\log\left(1+\sum_{p\geqslant 1}\frac{l_{np+1}/l_{n1}}{N^p}\right). \tag{33}$$

This last relation can be rendered into an explicit form

$$m_{np}=-\sum_{\Sigma_i q_i r_i=p}\left(\sum_i q_i-1\right)!\prod_i \frac{1}{q_i!}\left(\frac{-l_{nr_i+1}}{l_{n1}}\right)^{q_i}. \tag{34}$$

It is actually necessary to perform an expansion of this type because it combines the iteration number (n) dependence of the numerator and denominator which are both essential in the following results.

Then one can establish a hierarchy of equations for these coefficients:

$$l_{n1}(t)=nF''(t),$$

$$l_{n2}(t)=\sum_{j=1}^{n-1} jD_t^2 \log l_{n-j1}(t), \tag{35}$$

$$l_{np}(t)=\sum_{j=1}^{n-1} jm''_{n-jp-2}(t) \quad \text{for } p\geqslant 3,$$

for $n\geqslant 1$ while for $n=0$ we have $l_{np}(t)=0$ as $L_0(t,N)=\exp(NF(t))$. The first members of this hierarchy can be easily solved for yielding

$$l_{n1}(t)=nF''(t),$$

$$l_{n2}(t)=\frac{1}{2}n(n-1)\frac{F^{(2)}F^{(4)}-(F^{(3)})^2}{(F^{(2)})^2}, \tag{36}$$

$$l_{n3}(t)=\frac{1}{12}n(n-1)(n-2)\left(\frac{F^{(2)}F^{(4)}-(F^{(3)})^2}{(F^{(2)})^3}\right)^{(2)},$$

and from these it is easy to establish the leading-order terms already found in Eqs. (29) and (30).

Lemma 2: The hierarchy coefficients $l_{np}(t)$, $m_{np}(t)$ are polynomials in n.

These coefficients are constructed from a finite difference equation in *n*. □

Theorem 3: *The hierarchy coefficients $l_{np}(t)$, $m_{np}(t)$ are polynomials of degree p in n.*

This is proved by induction on *p* using the hierarchy equations. If we take $l_{jq}(t)$ to be of degree $q\leqslant p-2$ in *n*, then similarly for $m_{jq}(t)$ and $m''_{jq}(t)$. Now for any polynomial $P(n)$ of degree $p-2$ in its argument, then

$$\sum_{j=1}^{n-1} jP(n-j). \tag{37}$$

is a pth degree polynomial. Thus the recurrence, Eq. (35), establishes that l_{n+1p} is also a pth degree polynomial. □

From this result it is clear that the limiting forms of the Lanczos coefficients $\alpha_n(N)$, $\beta_n^2(N)$ exist when $n, N \to \infty$ with n/N fixed. If the ratio is not kept constant in this limiting operation, say with $n = o(N)$, then the Lanczos coefficients will vanish in the limit, while if the reverse is true, $N = o(n)$, then there will be divergent terms in the limit.

Given that the scaling Lanczos coefficients have been established, then all the exact theorems for the ground state properties[28,29] that were predicated on this result now are established. The first example of these theorems was the one for the ground state energy density,

$$\epsilon_0 = \inf_{s \in \mathbb{R}^+} [\alpha(s) - 2\beta(s)], \tag{38}$$

which also has an analog for the top of the spectrum, if this exists,

$$\epsilon_\infty = \sup_{s \in \mathbb{R}^+} [\alpha(s) + 2\beta(s)]. \tag{39}$$

For many models these Lanczos functions will be bounded on the positive real axis, and have limits as $s \to \infty$ on the real line. So, there is a superficial similarity to classes of orthogonal polynomials whose three-term recurrence coefficients have limiting values, such as the S class, the M class, or the $M(a,b)$ classes.[14]

IV. THE EXTENSIVE MEASURE

It is necessary to determine the OPS measure and its weight function $w(\epsilon)$, and this is not generally known at the outset, but rather the cumulant generating function is. In fact, it seems to be the case that the measures are not exactly expressible in simple terms, but the CGF or characteristic functions are. There is, of course, a direct route from a model system and a trial state to the Lanczos coefficients, but from many points of view, including practical considerations, the route beginning with a cumulant description is more useful.

Theorem 4: *Given that the cumulant generating function $F(-t)$ is analytic for $\mathcal{R}(t) > 0$ and in the neighborhood of the origin $t = 0$, the OPS weight function $w(\epsilon)$ has the following asymptotic development in the thermodynamic limit $N \to \infty$,*

$$w(\epsilon) = \sqrt{\frac{N}{2\pi F^{(2)}(\xi)}}\, e^{N[-\epsilon\xi + F(\xi)]} + O(N^{-1/2}), \tag{40}$$

where the function $\xi(\epsilon)$ is defined implicitly by

$$\epsilon = F'(\xi). \tag{41}$$

Starting with the definition of the cumulant generating function $F(t)$,

$$\langle e^{tH} \rangle \equiv \exp\{NF(t)\} = \exp\left\{ N \sum_{n=1}^{\infty} \frac{c_n}{n!} t^n \right\} \tag{42}$$

We assume here that this infinite series is not just formal but actually exists, that is, it has a finite radius of convergence in addition to its analytic character for $\mathcal{R}(t) < 0$. However, the moment generating function is simply the analytic continuation of the characteristic function and this continuation is possible given its analyticity, so that a Fourier inversion of this will yield the weight function,

$$w(\epsilon) = \frac{N}{2\pi} \int_{i\gamma-\infty}^{i\gamma+\infty} dt\ e^{N[-it\epsilon + F(it)]} = \frac{N}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} dt\ e^{N[t\epsilon + F(-t)]} \quad \mathcal{R}(\gamma) > 0. \tag{43}$$

One does not require the exact inversion, but only the leading order in N in a steepest descent approximation. In an asymptotic analysis the relevant function is

$$g(t)=t\epsilon+F(-t), \tag{44}$$

which is analytic for all $\mathcal{R}(t)>0$. We will assume the existence of a stationary point which occurs at t_0,

$$\epsilon=F'(-t_0), \tag{45}$$

and is assumed to be unique. This point is evidently real because the energy density is real and the CGF is a real function of a real argument (here we define $\xi=-t_0$ for convenience). One requires the inversion of this relation for $\xi(\epsilon)$ and this is guaranteed by the implicit function theorem because $F^{(2)}(\xi)>0$. This latter condition also implies that the saddle point is of order unity. Indeed, one clearly has the case of $F^{(2)}(t)>0$ for real values of t in the neighborhood of the saddle point and $F^{(2)}(t)<0$ for imaginary values of t in the same neighborhood. Thus the path of steepest descent through the saddle point is parallel to the imaginary axis. One can then apply the standard saddle point analysis (see Ref. 30, Sect. II.4) to arrive at the stated result. □

The corresponding example of the saddle point equation for the isotropic XY model is

$$\epsilon=\frac{1}{\pi}\int_0^{\pi/2}dq\ \cos q \tanh(\xi\cos q), \tag{46}$$

and that for the Ising model in a transverse field is

$$\epsilon=\frac{1}{\pi}\int_0^{\pi}dq\ \epsilon_q\ \frac{[(x+\cos q)/\epsilon_q]+\tanh(2\xi\epsilon_q)}{1+[(x+\cos q)/\epsilon_q]\ \tanh(2\xi\epsilon_q)}. \tag{47}$$

The first of the more obvious properties concerns the convexity of the measure arising in the thermodynamic limit,

Theorem 5: *The leading order of the negative logarithm of the weight function $u(\epsilon)$ is convex for all real energies ϵ.*

This follows from the relationship of $u(\epsilon)$ to the stationary point

$$\frac{d}{d\epsilon}u(\epsilon)=N\xi(\epsilon), \tag{48}$$

and the definition

$$\epsilon=F'(\xi). \tag{49}$$

Now it can be easily seen that $F''(t)>0$ for t real and the Hermitian Hamiltonian using the definition of $F(t)$ in terms of the expectation value $NF(t)=\ln\langle\exp(tH)\rangle$. □

Some detailed, yet general information, concerning the extensive measure in the neighborhood of the ground state is available. This arises from consideration of the overlap of the trial state with the true ground state,[31] and its relation to the Horn–Weinstein function $E(t)\equiv F'(-t)$ via

$$|\langle\Psi_{\mathrm{GS}}|\psi_0\rangle|^2=\exp\left\{-N\int_0^{\infty}dt\ [E(t)-E(\infty)]\right\}. \tag{50}$$

In general, the limit $E(t)$ as $t\to\infty$ will exist, and is the ground state energy, and so the asymptotic properties of $E(t)$ for $\mathcal{R}(t)>0$ as this tends to infinity is a means of classifying systems. This equivalent to the asymptotic properties of $\epsilon(\xi)-\epsilon(-\infty)$ as $\xi\to-\infty$ [we denote the ground state energy by ϵ_0, which is also the same as $\epsilon(-\infty)$]. In general the overlap is nonzero, so that $E(t)-E(\infty)\in L^1[0,\infty)$, but it is possible at isolated points that this is not true (critical points in the model, for example) and the overlap may vanish. For example, the overlap squared in the case of the isotropic *XY* model is $2^{-N/2}$ and that for the Ising model in a transverse field is

$$\exp\left\{\frac{N}{2\pi}\int_0^{\pi}dq\,\ln\left(\frac{\epsilon_q+x+\cos q}{2\epsilon_q}\right)\right\}. \tag{51}$$

Where the overlap is nonzero, then several possibilities for the asymptotic behavior exist, which do actually arise in the exact solutions of the example models.

(i) *gapless case*, isotropic *XY* and critical Ising model in a transverse field (Refs. 32 and 23): At a critical point, the first excited state gap vanishes and

$$\epsilon-\epsilon_0\sim A|\xi|^{-\gamma}, \tag{52}$$

as $\xi\to-\infty$ and, if the overlap is finite, then $\mathcal{R}(\gamma)>1$. Therefore, the weight function at the bottom of the spectrum takes the following form:

$$w(\epsilon)\sim(\epsilon-\epsilon_0)^{-(1+\gamma)/2\gamma}\exp\left\{N\,\frac{b}{1-1/\gamma}\,(\epsilon-\epsilon_0)^{1-1/\gamma}\right\}. \tag{53}$$

This measure is integrable on $(\epsilon_0,\epsilon_\infty)$ because of the above condition $\mathcal{R}(\gamma)>1$ and has a branch point at the ground state energy ϵ_0.

(ii) *gapped case 1*, Ising model in a transverse field, in the ordered phase with the disordered trial state (Ref. 23):

If the gap is finite, then one possibility is that

$$\epsilon-\epsilon_0\sim Ae^{-\Delta|\xi|}, \tag{54}$$

as $\xi\to-\infty$ and where the excited state gap $\Delta>0$. One can show that the weight function near the bottom edge of the spectrum is analytic, having the form

$$w(\epsilon)\sim\frac{1}{\Gamma(N[\epsilon-\epsilon_0]/\Delta+1)}. \tag{55}$$

(iii) *gapped case 2*, Ising model in a transverse field, in the disordered phase with the disordered trial state (Ref. 23):

Yet another type of gap behavior exists,

$$\epsilon-\epsilon_0\sim A|\xi|^{-\gamma}e^{-\Delta|\xi|}. \tag{56}$$

The leading-order behavior of the weight function in this case is

$$w(\epsilon)\sim(\epsilon-\epsilon_0)^{-1/2-N(\epsilon-\epsilon_0)}[-\log(\epsilon-\epsilon_0)]^{-N\gamma(\epsilon-\epsilon_0)}, \tag{57}$$

which again has a branch point at the bottom edge of the spectrum.

So generally we find the support of the measure is bounded, which excludes a number of weight function types such as the Freud or Erdös weights, but that the weight functions belong to the Szegö class on $[\epsilon_0,\ \epsilon_\infty]$,

$$\int_{\epsilon_0}^{\epsilon_\infty}d\epsilon\,\frac{\log w(\epsilon)}{\sqrt{[\epsilon_\infty-\epsilon][\epsilon-\epsilon_0]}}>-\infty. \tag{58}$$

V. EXACTLY SOLVABLE LANCZOS PROCESS

In this section we derive how the exact Lanczos functions $\alpha(s)$ and $\beta^2(s)$ can be constructed directly from the knowledge of the connected moments or cumulants, or more specifically from the cumulant generating function. This is the initial data that one uses in any analysis of quantum many-body systems with this approach, and for soluble models the full generating function may be available. However, if this is not the case, then one would use a set of low-order cumulants, up to a given order.

As a first step we recast the Hankel determinants into Selberg integral form, from the classical result[9]

$$\Delta_n(t)=\frac{1}{(n+1)!}\int_{-\infty}^{+\infty}\prod_{k=1}^{n+1} d\rho(\epsilon_k)e^{Nt\Sigma_{k=1}^{n+1}\epsilon k}\prod_{1\leq i<j\leq n+1}|\epsilon_i-\epsilon_j|^2. \tag{59}$$

For the steps leading to the two conditions which will define the Lanczos functions we follow Chen and Ismail.[13] A similar approach, but just confined to the evaluation of the Hankel determinants, was taken in Refs. 33 and 34. The Hankel determinant can be recast into the form of a partition function, which is

$$\Delta_n(t)=\frac{1}{(n+1)!}\int_{-\infty}^{+\infty}\prod_{i}^{n+1} d\epsilon_i \exp\left\{-\sum_i^{n+1} u(\epsilon_i)+Nt\sum_i^{n+1}\epsilon_i+2\sum_{i<j}^{n+1}\ln|\epsilon_i-\epsilon_j|\right\}. \tag{60}$$

One should observe that both $\Sigma_i^{n+1}u(\epsilon_i)$ and $Nt\Sigma_i^{n+1}\epsilon_i$ are of order $(n+1)N$, while the remaining term in the argument $\Sigma_{i<j}^{n+1}\ln|\epsilon_i-\epsilon_j|$ is of order $(n+1)^2$, so that the only relative scaling that remains nontrivial is one in which n/N is fixed. The alternatives would lead to completely trivial consequences. The leading-order term for this Hankel determinant as n, $N\rightarrow\infty$ is given by a steepest descent approximation (see Ref. 30, Sec. IX.5)

$$\Delta_n(t)=\frac{(2\pi)^{n+1}}{(n+1)!}\left|\frac{\partial^2 f}{\partial\epsilon_i^0\partial\epsilon_j^0}\right|^{-1/2}e^{-f(\epsilon^0)}[1+O(1/n,1/N)], \tag{61}$$

where the function $f(\epsilon)$ is defined as

$$f(\epsilon)=\sum_i^{n+1}u(\epsilon_i)-Nt\sum_i^{n+1}\epsilon_i-2\sum_{i<j}^{n+1}\ln|\epsilon_i-\epsilon_j|, \tag{62}$$

and the saddle points $\{\epsilon_i^0\}_{i=1}^{n+1}$ are given by

$$u'(\epsilon_i^0)=Nt+2\sum_{i\neq j}^{n+1}\frac{1}{\epsilon_i^0-\epsilon_j^0}. \tag{63}$$

One can easily show that the Hessian in Eq. (61) is positive definite given that $u(\epsilon)$ is convex. One can carry the continuum limit further by describing the saddle points as a charged fluid whose dynamics are governed by an energy functional $F[\sigma]$,

$$\exp(-f(\epsilon^0))\xrightarrow[n,N\rightarrow\infty]{}\exp(-F[\sigma_0]), \tag{64}$$

with a charge density $\sigma(\epsilon)$ defined on an interval of integration which is to be determined, $I=(\epsilon_-,\epsilon_+)$. The energy functional takes the following form

$$F[\sigma]=\int_I d\epsilon\,\sigma(\epsilon)[u(\epsilon)-Nt\epsilon]-\int_I d\epsilon\int_I d\epsilon'\,\sigma(\epsilon)\ln|\epsilon-\epsilon'|\sigma(\epsilon'), \tag{65}$$

where the single particle confining potential is controlled by the OPS measure and the two-body interaction is a logarithmic type. The result of minimizing this functional yields the following singular integral equation for the charge density,

$$u'(\epsilon)-Nt=2\mathrm{PV}\int_I d\epsilon' \, \frac{\sigma_0(\epsilon')}{\epsilon-\epsilon'}. \tag{66}$$

The solution of this integral equation for the minimal charge density $\sigma_0(\epsilon)$ can be found exactly and is

$$\sigma_0(\epsilon)=\frac{\sqrt{(\epsilon_+-\epsilon)(\epsilon-\epsilon_-)}}{2\pi^2}\,\mathrm{PV}\int_I d\epsilon' \, \frac{u'(\epsilon')-Nt}{(\epsilon'-\epsilon)\sqrt{(\epsilon_+-\epsilon')(\epsilon'-\epsilon_-)}}. \tag{67}$$

There are two conditions arising from this solution:

(i) the first is a supplementary condition which is necessary for the charge density solution to be well defined throughout the interval I,

$$0=\int_I d\epsilon \, \frac{u'(\epsilon)-Nt}{\sqrt{(\epsilon_+-\epsilon)(\epsilon-\epsilon_-)}}; \tag{68}$$

(ii) and the normalization condition which simply counts the number of Lanczos steps,

$$n=\frac{1}{2\pi}\int_I d\epsilon \; \epsilon \, \frac{u'(\epsilon)-Nt}{\sqrt{(\epsilon_+-\epsilon)(\epsilon-\epsilon_-)}}. \tag{69}$$

Using this solution for the charge density one can substitute this into the original defining equations for the Hankel determinants (the leading-order approximations) and establish that the Lanczos functions are simply defined by the interval I in this way, $\epsilon_\pm=\alpha\pm2\beta$.

Theorem 6: *The Lanczos functions are given implicitly by the two integral equations*

$$0=\int_{\alpha-2\beta}^{\alpha+2\beta} d\epsilon \, \frac{\xi(\epsilon)}{\sqrt{4\beta^2-(\epsilon-\alpha)^2}}, \tag{70}$$

$$s=\frac{1}{2\pi}\int_{\alpha-2\beta}^{\alpha+2\beta} d\epsilon \, \frac{\epsilon\xi(\epsilon)}{\sqrt{4\beta^2-(\epsilon-\alpha)^2}}, \tag{71}$$

where the model-dependent equation for the stationary point $\xi(\epsilon)$ is given by Eq. (49).

This theorem follows from the previous conditions, namely Eqs. (68) and (69), and the result for the logarithmic derivative of the weight function,

$$u'(\epsilon)=N\xi(\epsilon)+O(\log N). \tag{72}$$

□

Usually this later equation for the saddle point is also an implicit equation and invariably a nonlinear one. In our derivation the scaling $s=n/N$ remains finite while $n,N\to\infty$ emerges naturally and, in fact, it is difficult to see how one could avoid this confluence.

We now give an alternative result for the Lanczos functions which is based on the time evolution of the Lanczos L-function.

Theorem 7: *The Lanczos L-function, in the thermodynamic limit, is the solution of the following integro-differential equation*

$$L(s,t)=\int_0^s dr \; rD_t^2 \log L(s-r,t)+sF^{(2)}(t), \tag{73}$$

and the two Lanczos functions are derivable from this via

$$\alpha(s)=\int_0^s dr\ D_t \log L(r,0)+F'(0),$$
$$\beta^2(s)=L(s,0). \tag{74}$$

The integro-differential equation is simply derived from the discrete recurrence, namely Eq. (24), after making the observation that the $j=n$ term involving $L_0(t)$ has to be separated from the sum because it encompasses the initial conditions and is itself not generated by the recurrence. □

Finally, we give a result equivalent to the theorem above, but which involves only scaled forms of the Hankel determinants $\Delta_n(N,t)$and is the differential analog of the above theorem.

Definition 4: We make the following definition for $\delta(n,N,t)$ in terms of the Hankel determinant,

$$\Delta_n(N,t)=N^{n(n+1)}[\delta(n,N,t)]^{N^2}, \tag{75}$$

for $n\geqslant 1$ and $\Delta_0(t)=[\delta(0,t)]^N$.

Lemma 3: The function $\delta(n,N,t)$ is well defined in the scaling limit $n,N\to\infty$.

This follows naturally from the relation of the $\Delta_n(t)$and the Lanczos L-function as given in Eq. (22), and the well-defined scaling of this latter function as demonstrated in Theorem 3 above. □

Then we have the following result.

Theorem 8: *The Lanczos $\delta(s,t)$-function satisfies the following partial differential equation in the thermodynamic limit,*

$$\exp\{D_s^2 \log \delta(s,t)\}=D_t^2 \log \delta(s,t), \tag{76}$$

with the boundary condition

$$\lim_{s\to 0+}\frac{\log \delta(s,t)}{s}=F(t) \quad \forall t\in\mathbb{R}^+. \tag{77}$$

The Lanczos functions are given by

$$\alpha(s)=D_tD_s \log \delta(s,t)\big|_{t=0},$$
$$\beta^2(s)=\exp\{D_s^2 \log \delta(s,0)\}. \tag{78}$$

Using the scaling relation above, Eq. (75), and the equation of motion for $\Delta_n(t)$, Eq. (27), the result follows. □

These last two theorems relate to the dynamics of a nonlinear continuum Toda lattice in one space domain $s\in\mathbb{R}^+$ and one time domain t, with boundary conditions defined at the origin $s=0$ for all times t by the cumulant generating function $F(t)$. The object is then to find the Lanczos functions $\alpha(s),\beta^2(s)$ from a solution of this system, wherein these functions are directly related to the solution at a given time $t=0$ over all spatial points $s>0$.

VI. THE TAYLOR SERIES EXPANSION

The investigation of the Taylor series expansion of the Lanczos coefficients about $s=0$ is an essential element in the application of this Lanczos method, as was indicated earlier, where one has only a finite set of low-order cumulants available, say for nonintegrable models. Therefore, in this case, one can only construct a truncated Taylor series expansion and so issues concerning convergence, the radius of convergence of the series, and whether one can extrapolate immediately arise. In addition, one would like a direct algorithm relating the cumulants to the Lanczos functions from a purely practical point of view.

We define the Taylor series expansion of the two Lanczos functions by two new sequences of coefficients,

$$\alpha(s)=c_1+\sum_{n=0}^{\infty} a_n s^{n+1},$$
$$\beta^2(s)=\sum_{n=0}^{\infty} b_n s^{n+1}, \tag{79}$$

In order to find these coefficients one could use either of the two general solutions for the Lanczos process, Eqs. (70) and (71) or Eq. (73), and the two methods are presented below.

The first step involves the inversion of the following Taylor series expansion,

$$\epsilon=c_1+\sum_{n=1} \frac{c_{n+1}}{n!}\xi^n, \tag{80}$$

for $\xi(\epsilon)$, namely the coefficients e_k appearing in

$$\xi=\sum_{k=1} e_k(\epsilon-c_1)^k. \tag{81}$$

The coefficients c_n appearing in Eq. (80) are the cumulant coefficients. The existence of this inverse function is guaranteed because the second cumulant $c_2>0$ in all systems and we assume that the saddle point function, Eq. (49), is analytic in the neighborhood of $\xi=0$. The next step involves the solution of the two recurrences

$$0=\sum_{k=1} e_k \sum_{m=0}^{\lfloor k/2\rfloor} \binom{k}{2m} \frac{(1/2)_m}{m!}(\alpha-c_1)^{k-2m}(4\beta^2)^m,$$
$$2s=\sum_{k=1} e_k \sum_{m=0}^{\lfloor (k-1)/2\rfloor} \binom{k}{2m+1} \frac{(1/2)_{m+1}}{(m+1)!}(\alpha-c_1)^{k-2m-1}(4\beta^2)^{m+1}, \tag{82}$$

which are used to solve for the coefficients a_n, b_n appearing in Eq. (79).

In the second method we define a continuum version of the coefficients that are defined in Eq. (33) in the following way,

$$\log\frac{L(s,t)}{sl_1(t)}=\log\left(1+\sum_{p\geqslant 1}\frac{l_{p+1}}{l_1}s^p\right)\equiv\sum_{p\geqslant 1} m_p(t)s^p, \tag{83}$$

and the inverse of Eq. (34) in an explicit form,

$$\frac{l_{p+1}}{l_1}=\sum_{\Sigma_i q_i r_i=p}\prod_i \frac{1}{q_i!}m_{r_i}^{q_i}. \tag{84}$$

From these relations one can find a hierarchy of equations for these coefficients:

$$l_1(t)=F''(t),$$

$$l_2(t)=\frac{F^{(2)}F^{(4)}-(F^{(3)})^2}{2(F^{(2)})^2}, \tag{85}$$

$$l_{p+2}(t)=\frac{m_p''(t)}{(p+2)(p+1)}=l_1(t)\sum_{\Sigma_i q_i r_i=p+1}\prod_i \frac{m_{r_i}^{q_i}}{q_i!} \quad \text{for } p\geqslant 1.$$

Thus one can verify from the solution for the initial value problem above that the general Taylor series coefficients are given by

$$[(n+1)!]^2 c_2^{3n+1} a_n=\sum_{\lambda\vdash 2n+1} A(n;\lambda)\prod_{i=0}^{2n+1} c_{2+i}^{a_i},$$
$$(n+1)!n!c_2^{3n-1} b_n=\sum_{\lambda\vdash 2n} B(n;\lambda)\prod_{i=0}^{2n} c_{2+i}^{a_i}, \tag{86}$$

where the coefficients labeled by the partition $\lambda=(1^{a_1},2^{a_2},\ldots,i^{a_i})$, denoted by $A(n;\lambda)$, $B(n;\lambda)$, are listed in the table in the Appendix. There are constraints operating in the above equations, namely $\Sigma_{i=1}^{2n+1} i a_i=\Sigma_{i=0}^{2n+1} a_i=2n+1$ for the first relation and $\Sigma_{i=1}^{2n} i a_i=\Sigma_{i=0}^{2n} a_i=2n$ for the second.

Clearly the Taylor series expansion of the Lanczos functions has low-order coefficients which are constructed from the low-order cumulants, and is a form of a linked cluster expansion. However, it is not just a simple linked cluster expansion as in the Taylor series expansion of the cumulant generating function, but involves a subtle interplay and cancellation of all cumulants below a given order.

VII. GENERAL PROPERTIES

There are some very general properties that the Lanczos process in the thermodynamic limit and the associated Lanczos functions satisfy and we examine these now. Some are quite obvious and not particularly surprising; however, we state these for completeness sake, while there are some other properties which are not so immediate, but very important nevertheless.

The next, and natural, property concerns the monotonicity of the two envelope function $\epsilon_{\pm}(s)=\alpha(s)\pm 2\beta(s)$.

Theorem 9: *The envelope functions* $\epsilon_+(s),\epsilon_-(s)$ *are monotonically increasing and decreasing functions of real, positive s, respectively*.

This follows from a recasting of the normalization condition in the following way,

$$2\pi s=\int_{\xi-}^{\xi+} d\xi\ \sqrt{[\epsilon(\xi_+)-\epsilon(\xi)][\epsilon(\xi)-\epsilon(\xi_-)]}, \tag{87}$$

where the $\xi_{\pm}$ are defined by $\epsilon(\xi_{\pm})=\epsilon_{\pm}$. Now it is straightforward to write the explicit forms for the derivatives of the envelope functions with respect to s as

$$\frac{d\epsilon_+}{ds}=4\pi \Bigg/ \int_{\xi_-}^{\xi_+} d\xi\ \sqrt{\frac{\epsilon(\xi)-\epsilon(\xi_-)}{\epsilon(\xi_+)-\epsilon(\xi)}},$$
$$\frac{d\epsilon_-}{ds}=-4\pi \Bigg/ \int_{\xi_-}^{\xi_+} d\xi\ \sqrt{\frac{\epsilon(\xi_+)-\epsilon(\xi)}{\epsilon(\xi)-\epsilon(\xi_-)}}, \tag{88}$$

so that the stated properties are evident. □

It is clear that the envelope functions $\epsilon_{\pm}(s)$ are bounded in the following ways, $\epsilon_{-}(s) \geqslant \epsilon_0$ and $\epsilon_{+}(s) \leqslant \epsilon_{\infty}$.

The three-term recurrence which serves as one of the definitions of the orthogonal polynomials themselves is now going to take a definite limiting form when $n,N \to \infty$ such that s is finite. This is going to lead to a scaling form for one set of the polynomials themselves, which would be more correctly termed orthogonal functions $p(s,\epsilon)$. Heuristically one can see how this arises by the following argument. If one ensures that Lanczos densities are employed and the following scaling of the polynomials thus $P_n(E) = N^n p_n(\epsilon)$, then the three-term recurrence becomes

$$p_{n+1}(\epsilon)/p_n(\epsilon) + \beta_n^2 \frac{1}{p_n(\epsilon)/p_{n-1}(\epsilon)} = \epsilon - \alpha_n . \tag{89}$$

Now these ratios are approximated by

$$\frac{p_{n+1}(\epsilon)}{p_n(\epsilon)} \sim \exp\left(\frac{1}{N}\frac{\partial}{\partial s} \ln p(s,\epsilon)\right), \tag{90}$$

for arguments $\epsilon \in \mathbb{C} \backslash \mathrm{Supp}\,[d\rho]$, so that in the asymptotic regime the recurrence becomes

$$\exp\left(\frac{1}{N}\frac{\partial}{\partial s} \ln p(s,\epsilon)\right) + \beta^2(s) \exp\left(-\frac{1}{N}\frac{\partial}{\partial s} \ln p(s,\epsilon)\right) \sim \epsilon - \alpha(s), \tag{91}$$

whose solutions are

$$p^{\pm}(s,\epsilon) \sim p(0) \exp\left\{ N \int^{s} dt \ln \frac{1}{2} \left[\epsilon - \alpha(t) \pm \sqrt{(\epsilon - \alpha(t))^2 - 4\beta^2(t)}\right]\right\}. \tag{92}$$

These are the corresponding results for the ratio $P_n(x)/P_{n+1}(x)$ or nth root $\sqrt[n]{P_n(x)}$ asymptotics of generic orthogonal polynomials as $n \to \infty$ (Refs. 9, 11, and 35–37), or the scaled orthogonal polynomials,[14] but are rather different due to the particular nature of many-body orthogonal polynomials.

Theorem 10: *Given the scaling behavior of the Lanczos coefficients, and that they are bounded for $n,N \to \infty$, then the nth roots of the denominator orthogonal polynomials $p_n(\epsilon)$ have the limiting form uniformly for ϵ in compact subsets of $\mathbb{C} \backslash \mathrm{Supp}\,[d\rho]$:*

$$p(s,\epsilon) \equiv \lim_{n,N \to \infty} |p_n(N,\epsilon)|^{1/N} = \exp\left\{ \int_0^{s} dt \ln \frac{1}{2} \left[\epsilon - \alpha(t) + \sqrt{(\epsilon - \alpha(t))^2 - 4\beta^2(t)}\right]\right\}. \tag{93}$$

The proof of this parallels the one constructed by van Assche in Ref. 14 through the use of Turán determinants,

$$D_n \equiv p_n^2 - p_{n+1} p_{n-1} . \tag{94}$$

One can show that these obey the following recurrence relation

$$D_n = \beta_n^2 D_{n-1} + (\alpha_n - \alpha_{n-1}) p_n p_{n-1} + (\beta_n^2 - \beta_{n-1}^2) p_n p_{n-2} . \tag{95}$$

Using the partial fraction decomposition of the ratio of two successive orthogonal polynomials one can also find a bound on this ratio

$$\left|\frac{p_{n-1}(\epsilon)}{p_n(\epsilon)}\right| \leqslant \frac{C}{d} \quad \forall n \tag{96}$$

for all $\epsilon \in K$ where the compact set $K \subset \mathbb{C} \backslash \mathrm{supp}\,[d\rho]$ and d is the distance between this set and the interval $[\epsilon_0, \epsilon_\infty]$, and C is a positive constant. Using Eq. (95) we have

$$\left|\frac{D_n}{p_n^2}\right| \leqslant \sup_n(\beta_n^2)\,\frac{C^2}{d^2}\left|\frac{D_{n-1}}{p_{n-1}^2}\right| + |\alpha_n - \alpha_{n-1}|\,\frac{C}{d} + |\beta_n^2 - \beta_{n-1}^2|\,\frac{C^2}{d^2}. \tag{97}$$

Given the scaling form of the Lanczos coefficients the ratio $|D_n/p_n^2| \to 0$ as $n, N \to \infty$ uniformly in ϵ whenever d is large enough. This means that $|p_{n-1}/p_n|$ and $|p_n/p_{n+1}|$ tend to the same accumulation point which we denote by $p(s,\epsilon)$. This point is given by the solution of the quadratic equation $p + \beta^2(s)/p = \epsilon - \alpha(s)$, and the positive branch of the solution must be taken as $p \to \infty$ when $\epsilon \to \infty$. The functions $p(s,\epsilon)$ are analytic functions of $\epsilon \in K$ which are uniformly bounded, so the restriction on d can be lifted to being only nonzero. The behavior of the nth ratio then gives the nth root behavior directly as

$$|p_n|^{1/N} = \exp\left\{\frac{1}{N}\sum_{k=1}^{n} \log\left|\frac{p_k(\epsilon)}{p_{k-1}(\epsilon)}\right|\right\}, \tag{98}$$

The asymptotic behavior that we have found applies to the denominator OP only as can be seen from the observation that $p_1 = \epsilon - c_1$ and $p_2 = (\epsilon - c_1)^2 - c_3/c_2 N(\epsilon - c_1) - c_2/N$, while

$$[\epsilon - \alpha(s) + \sqrt{(\epsilon - \alpha(s))^2 - 4\beta^2(s)}] \xrightarrow[s\to 0]{} \frac{1}{\epsilon - c_1}\,((\epsilon - c_1)^2 - c_3/c_2 N(\epsilon - c_1) - c_2/N). \tag{99}$$

This establishes the result. □

VIII. SUMMARY

In this work we have demonstrated the general scaling behavior of the Lanczos process as applied to many-body systems when the process is taken to convergence and the thermodynamic limit taken. We also find explicit constructions of the limiting Lanczos coefficients in two equivalent formulations, from an initial exact solution of the moment problem, that is to say, the cumulant generating function for the system. There are explicit examples where the CGF can be found and the whole Lanczos process explicitly realized. Furthermore, we have given the corresponding results for the associated orthogonal polynomial system and the measure in this regime quite generally. However, we must emphasize that these results apply only to the bulk properties, that is to say, the ground state properties that scale extensively and the spectral properties in the interior (the ‘‘bulk’’) of the spectrum. So this does not include the delicate scaling behavior at the edges of the spectrum, nor in the neighborhood of singularities—this theory would have to be extended to treat the excited state gaps near the bottom of the spectrum. A number of general theorems are given which constrain the behavior of the Lanczos functions, and the process in general. We also indicate how a number of such constraints operating can lead to some concrete realizations or scenarios that the Lanczos process can present, namely its behavior at a critical point in the model under study. This is a significant step on the way to the goal of a rigorous classification of many-body systems in terms of their character via the Lanczos process. Other important questions that arise in the treatment of nonintegrable models, for which the general results presented here have suggested some answers, are the questions of the choice of trial state, the rate of convergence of the truncated Lanczos process, and how one might accelerate its convergence given some independent qualitative knowledge.

TABLE I. The coefficients in the Taylor series expansion for the Lanczos functions $\alpha(s)$ and $\beta^2(s)$, as defined in Eq. (86), and the labels denoting the partitions λ of the positive integers.

	$\lambda =$	$A(0;\lambda)=$
1 a_0		
	1	1
		$B(1;\lambda)=$
2 b_1	2	1
	1^2	−1
		$A(1;\lambda)=$
3 a_1	1^3	3
	2.1	−4
	3	1
		$B(2;\lambda)=$
4 b_2	1^4	−12
	2.1^2	21
	2^2	−4
	3.1	−6
	4	1
		$A(2;\lambda)=$
5 a_2	1^5	81
	2.1^3	−174
	3.1^2	48
	$2^2.1$	70
	4.1	−9
	3.2	−17
	5	1
		$B(3;\lambda)=$
6 b_3	1^6	−567
	2.1^4	1449
	3.1^3	−414
	$2^2.1^2$	−872
	4.1^2	84
	3.2.1	304
	5.1	−12
	2^3	70
	4.2	−26
	3^2	−17
	6	1
		$A(3;\lambda)=$
7 a_3	1^7	5805
	2.1^5	−17 190
	3.1^4	4815
	$2^2.1^3$	13 940
	4.1^3	−990
	5.1^2	150
	$3.2.1^2$	−5470
	$2^3.1$	−2680
	$3^2.1$	425
	4.2.1	680
	6.1	−16
	3.2^2	640
	5.2	−44
	4.3	−66
	7	1

TABLE I. (*Continued.*)

		$B(4;\lambda)=$
8 b_4	1^8	−58 050
	2.1^6	195 345
	3.1^5	−55 710
	$2^2.1^4$	−197 470
	$3.2.1^3$	85 430
	4.1^4	11 745
	5.1^3	−1890
	$2^3.1^2$	60 580
	$3^2.1^2$	−8020
	$4.2.1^2$	−12 520
	6.1^2	230
	$3.2^2.1$	−22 820
	4.3.1	1860
	5.2.1	1200
	7.1	−20
	2^4	−2680
	4.2^2	1320
	$3^2.2$	1705
	6.2	−60
	5.3	−110
	4^2	−66
	8	1

ACKNOWLEDGMENTS

One of the authors (NSW) would like to acknowledge the support of an Australian Research Council Large Grant while this work was performed, and the hospitality of Service de Physique Théorique, Center d'Etudes Nucléaires de Saclay.

APPENDIX: COEFFICIENTS OF THE TAYLOR SERIES

We list here in the Table I the coefficients of the Taylor series expansion for the Lanczos coefficients, labeled by the partitions of integers, according to the definition of Eq. (86).

[1] E. Dagotto, ''Correlated Electrons in High-Temperature Superconductors,'' Rev. Mod. Phys. **66**, 763–840 (1994).
[2] C. Lanczos, J. Res. Natl. Bur. Stand. **45**, 255–282 (1950).
[3] B. N. Parlett, *The Symmetric Eigenvalue Problem* (Prentice–Hall, Englewood Cliffs, NJ, 1980).
[4] F. Chatelin, *Eigenvalues of Matrices* (Wiley, Chichester, 1993).
[5] Y. Saad, *Numerical Methods for Large Eigenvalue Problems.* Manchester University Press Series in Algorithms and Architectures for Advanced Scientific Computing (1991).
[6] S. Kaniel, ''Estimates for Some Computational Techniques in Linear Algebra,'' Math. Comput. **20**, 369–378 (1966).
[7] C. C. Paige, ''Error Analysis of the Lanczos Algorithm for Tridiagonalizing a Symmetric Matrix,'' J. Inst. Math. Appl. **18**, 341–349 (1976).
[8] Y. Saad, ''On the Rates of Convergence of the Lanczos and the Block-Lanczos Methods,'' SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **17**, 687–706 (1980).
[9] G. Szegö, *Orthogonal Polynomials*, Colloquium Publications Vol. 23 (American Mathematical Society, Providence, RI, 1975), 4th ed.
[10] T. S. Chihara, *An Introduction to Orthogonal Polynomials* (Gordon and Breach, New York, 1978).
[11] G. Freud, *Orthogonal Polynomials* (Pergamon, Oxford, 1971).
[12] B. G. Lindsay, ''On the Determinants of Moment Matrices,'' Ann. Stat. **17**, 711–721 (1989).
[13] Y. Chen and M. E. H. Ismail, ''Thermodynamic Relations of the Hermitian Matrix Ensembles,'' J. Phys. A **30**, 6633–6654 (1997).
[14] W. van Assche, ''Asymptotics for Orthogonal Polynomials and Three-Term Recurrences,'' in *Orthogonal Polynomials*, edited by P. Nevai (Kluwer Academic, Dordrecht, 1990), pp. 435–462.
[15] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments* (American Mathematical Society, Providence, RI, 1943).
[16] N. I. Akhiezer, *The Classical Moment Problem* (Oliver and Boyd, London, 1965).
[17] W. B. Jones and W. J. Thron, *Continued Fractions - Analytic Theory and Applications* (Addison–Wesley, Reading, MA, 1980).
[18] L. Lorentzen and H. Waadeland, *Continued Fractions with Applications* (North-Holland, Amsterdam, 1992).
[19] M. G. Kendall, *The Advanced Theory of Statistics, Vol. 1* (Edward Arnold and Halstead, London, 1994), 6th ed.

[20] R. Kubo, ''Generalized Cumulant Expansion Method,'' J. Phys. Soc. Jpn. **17**, 1100–1120 (1962).
[21] L. C. L. Hollenberg, M. P. Wilson, and N. S. Witte, ''General Non-Perturbative Massgap to First Order in $1/V$,'' Phys. Lett. B **361**, 81–86 (1995).
[22] N. S. Witte, ''Analytic Solution to the Moment Problem for the *XY* Chain,'' Int. J. Mod. Phys. B **11**, 1503–1517 (1997).
[23] N. S. Witte, ''Moment Formalisms Applied to a Solvable Model with a Quantum Phase Transition. II. Geometrical Moment Methods,'' submitted to Nucl. Phys. B (1999).
[24] S. Karlin, *Total Positivity, Vol. 1* (Stanford U. P., Stanford, 1968).
[25] M. Toda, *Theory of Nonlinear Lattices*, Springer Series in Solid-State Sciences 20 (Springer-Verlag, Berlin, 1989), 2nd ed.
[26] N. S. Witte and L. C. L. Hollenberg, ''Plaquette Expansion Proof and Interpretation,'' Z. Phys. B **95**, 531–539 (1994).
[27] L. C. L. Hollenberg, ''Plaquette Expansion in Lattice Hamiltonian Models,'' Phys. Rev. D **47**, 1640–1644 (1993).
[28] L. C. L. Hollenberg and N. S. Witte, ''Analytic Solution for the Ground State Energy of the Extensive Many-Body Problem,'' Phys. Rev. B **54**, 16309–16312 (1996).
[29] N. S. Witte, L. C. L. Hollenberg, and Z. Weihong, ''Two-dimensional XXZ Model Ground State Properties Using an Analytic Lanczos Expansion,'' Phys. Rev. B **55**, 10412–10419 (1997).
[30] R. Wong, *Asymptotic Approximations of Integrals* (Academic P, Boston, 1989).
[31] J. Cioslowski, ''Estimation of the Overlap Between the Approximate and Exact Wave Function of the Ground State from the Connected-Moments Expansion,'' Phys. Rev. A **36**, 3441–3442 (1987).
[32] N. S. Witte, ''The Exact Realisation of the Lanczos Method for a Quantum Many-Body System,'' Phys. Lett. A **254**, 18–23 (1999).
[33] E. Brézin, C. Itzykson, G. Parisi, and J. B. Zuber, ''Planar Diagrams,'' Commun. Math. Phys. **59**, 35–51 (1978).
[34] G. Baker, Jr., D. Bessis, and P. Moussa, ''Asymptotic Behaviour of Some Hankel-Toeplitz Determinants,'' Rev. Math. Phys. **4**, 65–94 (1992).
[35] P. G. Nevai, *Orthogonal Polynomials* (American Mathematical Society, Providence, Rhode Island, 1979).
[36] P. G. Nevai, ''Distribution of Zeros of Orthogonal Polynomials,'' Trans. Am. Math. Soc. **249**, 341–361 (1979).
[37] W. van Assche, *Asymptotics for Orthogonal Polynomials*, Lecture Notes in Mathematics Vol. 1265 (Springer Verlag, Berlin, 1987).

The stationary KdV hierarchy and $so(2,1)$ as a spectrum generating algebra

H.-D. Doebner[a)]
Arnold-Sommerfeld Institute for Mathematical Physics, TU Clausthal, Leibnizstraße 10, 38678 Clausthal-Zellerfeld, Germany

R. Z. Zhdanov[b)]
Institute of Mathematics of the National Academy of Sciences of Ukraine, Tereshchenkivska Street 3, 252004 Kyiv, Ukraine

(Received 30 November 1998; accepted for publication 3 June 1999)

The family $\mathcal{F}_L^\kappa$ of all potentials $V(x)$ for which the Hamiltonian $H=-d^2/dx^2+V(x)$ in one space dimension possesses a high-order Lie symmetry is determined. A subfamily $\mathcal{F}_{\mathrm{SGA}}^{(2)}$ of $\mathcal{F}_L^\kappa$, which contains a class of potentials allowing a realization of $so(2,1)$ as spectrum generating algebra of H through differential operators of finite order, is identified. Furthermore and surprisingly, the families $\mathcal{F}_{\mathrm{SGA}}^{(2)}$ and $\mathcal{F}_L^\kappa$ are shown to be related to the stationary KdV hierarchy. Hence, the ''harmless'' Hamiltonian H connects different mathematical objects: high-order Lie symmetry, the realization of $so(2,1)$-spectrum generating algebra and families of nonlinear differential equations. We describe in a physical context the interplay between these objects. [S0022-2488(99)02710-3]

I. MOTIVATION AND BACKGROUND

In the two internal reports for the International Center of Theoretical Physics (ICTP),[1,2] written in the clearly 1970s, a complete classification of symmetric Hamiltonians in one space dimension on $L^2(\mathbf{R}_x^1,dx)$

$$H=\gamma\frac{d^2}{dx^2}+V(x),\qquad \gamma<0, \tag{1}$$

having the Lie algebra $so(2,1)$ as a ''spectrum generating algebra'' (SGA), has been obtained. This result has been published only recently in connection with a Barut-Memorial Lecture.[3] In Refs. 1, 2, the following definition of SGA is used: a differential operator A of the order n' has a spectrum generating (Lie) algebra L with generators $g_i\,(i=1,...,m,\ m=\dim L)$ if there exists a realization R of L through differential operators of an order $n\geqslant n'$, such that

$$A=\sum_{i=1}^{m}\alpha_i R(g_i),\qquad \alpha_i\in\mathbf{R}. \tag{2}$$

The Hamiltonian (1) is a differential operator on a suitable complex function space $\mathcal{G}$ over x with $n'=2$. A realization R of $so(2,1)$ with standard basis $\mathcal{M}$ spanned by g_1,g_2,g_3 through nth-order differential operators on $\mathcal{G}$ reads as

$$R(g_i)=\sum_{j=0}^{n} a_{ij}(x)\frac{d^j}{dx^j},\qquad i=1,2,3, \tag{3}$$

[a)]Electronic mail: asi@pt.tu-clausthal.de
[b)]Electronic mail: renat@imath.kiev.ua

where $a_{ij}(x)$ are complex functions, such that the commutation relations,

$$[R(g_1),R(g_2)]=-R(g_3),\quad [R(g_2),R(g_3)]=R(g_1),\quad [R(g_3),R(g_1)]=-R(g_2), \tag{4}$$

are fulfilled. So $R(so(2,1))$ is a SGA of (1) on $\mathcal{G}$ if there exist constants $\alpha_i\in\mathbf{R}\,(i=1,2,3)$, such that the equality,

$$H\equiv\gamma\frac{d^2}{dx^2}+V(x)=\sum_{j=1}^{3}\alpha_j R(g_j), \tag{5}$$

holds.

The relations (3)–(5) impose restrictions both on the coefficients $a_{ij}(x)$ and on the potential $V(x)$. Solving these restrictions we find two different families $\mathcal{F}_{\mathrm{SGA}}=\{\mathcal{F}^{(1)}_{\mathrm{SGA}},\mathcal{F}^{(2)}_{\mathrm{SGA}}\}$ of those potentials, which allow $so(2,1)$ as SGA for (1), and we can use properties of $R(so(2,1))$ to calculate the spectrum of *H*, if *H* and $R(so(2,1))$ act on $L^2(\mathbf{R}^1_x,dx)$. If, furthermore, $R(g_i)$ are essentially self-adjoint in a common dense domain and if the representation is integrable, then one can use the known theory for unitary representations of $so(2,1)$. This is the background of the term ''spectrum generating algebra'' as suggested in Refs. 4, 5. There were many results in this field for different Hamiltonians, Lie algebras, and physical systems (see, e.g., the recent review[6]), but no general study in the sense of Refs. 1 and 2.

The motivation of the present paper is to show (in Sec. III) that the family $\mathcal{F}^{(2)}_{\mathrm{SGA}}$ can be read from the stationary KdV hierarchy. This surprising connection between the KdV hierarchy and $so(2,1)$ has its origin in a certain higher-order Lie symmetry of the Hamiltonian (5), which has $so(2,1)$ as SGA. In Sec. II we sketch the results of Refs. 1 and 2 on which our discussion is based.

II. ON SOME KNOWN RESULTS

To classify those $V(x)$ that are solutions of (3)–(5), we reduce the calculation to special choices of parameters $\alpha_1,\alpha_2,\alpha_3$ through basis transformations, which leave $\mathcal{M}$ invariant. As a result, we only have to treat the following two cases ($\lambda\in\mathbf{R},\ \lambda\neq 0$):

$$\text{Case 1.}\quad \alpha_1^2+\alpha_2^2\neq\alpha_3^2,\quad \text{with } H=\lambda R(g_i),\quad i=1,2, \tag{6}$$

$$\text{Case 2.}\quad \alpha_1^2+\alpha_2^2=\alpha_3^2,\quad \text{with } H=\lambda(R(g_1)+R(g_3)). \tag{7}$$

Case 2 is denoted as the ''light cone case.'' Both cases (3)–(5) lead for a fixed *n* to set of coupled differential equations of the order *n* for $a_{ij}(x)$ $(i=1,2,3,\ j=0,...,n)$ and $V(x)$. We assume that $R(g_i)$ are symmetric operators and that $V(x)$ is a real function. A clumsy but straightforward calculation shows that in Case 1 a solution exists only for $n=2$ with a family $\mathcal{F}^{(1)}_{\mathrm{SGA}}$ of corresponding potentials,

$$\mathcal{F}^{(1)}_{\mathrm{SGA}}(\lambda_1,\lambda_2,c)=\{V(x)|V(x)=\lambda_1(x-c)^2+\lambda_2(x-c)^{-2},\lambda_1,\lambda_2,c\in\mathbf{R},\lambda_1\neq 0\}. \tag{8}$$

In Case 2 a solution exists for all $n\geqslant 2$. The corresponding family $\mathcal{F}^{(2)}_{\mathrm{SGA}}$ consists of potentials that are solutions of nonlinear differential equations,

$$\mathcal{F}^{(2)}_{\mathrm{SGA}}=\left\{V(x)\left|-\left(\frac{x}{2}V'+V\right)+\sum_{j=0}^{N-1}c_jF_j+F_N=0\right.\right\}, \tag{9}$$

where $N=[(n-1)/2]$, F_j are some polynomials of $V(x)$ and its derivatives up to the order $2j+1$ $(j=0,1,...,N)$, which will be given later in another context [see (12) and (13)]; $c_0,c_1,\ldots,c_{N-1}$ are arbitrary real constants.

The family $\mathcal{F}^{(2)}_{\mathrm{SGA}}$ has a peculiar structure. The facts that the equations are equal for $n=2N+1$ and $n=2N+2$, $N=1,2,...$, and that the equation for $n=n_1>n_2$ contains all terms of the equation for $n=n_2$ are two of these peculiarities. This structure was not elucidated in Refs. 1, 2, also relations to other mathematical notions and objects were not found. This was the reason why the authors of Refs. 1, 2 decided to present the results as internal reports. In the present paper we fill this gap.

In order to simplify the following calculations we scale the variable x and thus get $\gamma=-1$. So the Schrödinger operator (1) takes the form

$$H=-\frac{d^2}{dx^2}+V(x). \tag{10}$$

III. SGA, LIE SYMMETRIES, AND KdV HIERARCHY

A. Aim and strategy

We present a view on the family $\mathcal{F}_{\rm SGA}^{(2)}$ through a high-order Lie symmetry Q of type C_κ, i.e., with $[Q,H]=\kappa H$ (see Sec. III B) of the Schrödinger equation,

$$\left(-\frac{d^2}{dx^2}+V(x)\right)f(x)=0, \quad f\in\mathcal{G}, \tag{11}$$

and we show that because of this view the F_j, $(j=0,1,...,N)$ in (9) appear, surprisingly, in the stationary KdV hierarchy,

$$\sum_{j=0}^{N-1} c_j F_j+F_N=0.$$

We remind the reader that the stationary KdV hierarchy is obtained successively by the repeated action of the integrodifferential operator (the second recursive operator for the KdV equation[7–9]),

$$\mathcal{R}=-\frac{1}{4}\frac{d^2}{dx^2}-V(x)-\frac{1}{2}V'(x)\left(\frac{d}{dx}\right)^{-1}, \tag{12}$$

generating F_j, $(j=0,1,...)$ through

$$F_{j+1}=\mathcal{R}F_j, \quad j=0,1,..., \quad \text{with } F_0=-\tfrac{1}{2}V'(x). \tag{13}$$

Our strategy is the following. We construct as a first step the family $\mathcal{F}_L^0$ of all potentials $V(x)$ for which $Hf=0$ allows an nth-order Lie symmetry $Q\in\mathcal{L}$ of type C_0 (Theorem 1). In a second step, we generalize this result for a Lie symmetry of type C_κ, κ is arbitrary, (Theorem 2) and get the family $\mathcal{F}_L^\kappa$. In the special case $\kappa=-1$ we find the families $\mathcal{F}_{\rm SGA}^{(2)}$.

B. Lie symmetries

An nth-order differential operator on a suitable function space $\mathcal{G}$ over $\mathbf{R}_x^1$,

$$\tilde{Q}^{(n)}=\sum_{j=0}^{n} q_j(x)^{(n)}\frac{d^j}{dx^j}, \quad \frac{d^0}{dx^0}\overset{\rm def}{=}1, \tag{14}$$

is called an nth-order symmetry of the Schrödinger equation (11) if it transforms the set of its solutions Γ_H into itself, i.e., $\tilde{Q}^{(n)}\Gamma_H=\Gamma_H$ $(\mathcal{G}\supset\Gamma_H)$. It is reasonable to impose further properties on $\tilde{Q}^{(n)}$. A useful choice is a Lie symmetry $Q^{(n)}$ of H. Take an mth-order differential operator P and assume

$$[Q^{(n)},H]=PH, \quad \text{on } \mathcal{G}. \tag{15}$$

Obviously $Q^{(n)}$ on Γ_H is a symmetry $\tilde{Q}^{(n)}$ of (11).

In the following we use $P=\kappa\mathbf{1}$, κ is a real constant, i.e.,

$$[Q^{(n)},H]=\kappa H, \quad \text{on } \mathcal{G}, \tag{16}$$

and denote this as a Lie symmetry of type C_κ.

Together with $Q^{(n)}$, also

$$Q^{(n')}=Q^{(n)}+\sum_{j=0}^{m} \gamma_j H^j,$$

$n'<2m+n$ with arbitrary constants γ_j is a Lie symmetry of H. Hence, $Q^{(n')}$ gives no new information and is excluded from further considerations. To formalize this we introduce the equivalence relation

$$Q^{(n')}\sim Q^{(n)}, \quad \text{if } Q^{(n')}-Q^{(n)}=\sum_{j=0}^{m} \gamma_j H^j,$$

and consider representatives of these equivalence classes in the linear space of differential operators $Q^{(n)}$ satisfying (16). We denote this quotient space as $\mathcal{L}$.

The $Q^{(n)}$ of type C_κ depends on $V(x)$ and on κ. As mentioned, we denote the family of such potentials (for finite κ) as $\mathcal{F}_L^\kappa$. There is a connection between $\mathcal{F}_{\mathrm{SGA}}^{(2)}$ and $\mathcal{F}_L^\kappa$ (see Sec. III E):

$$\mathcal{F}_{\mathrm{SGA}}^{(2)}\subset\mathcal{F}_L^\kappa,$$

because in the singular case (16), the realization $R(g_2)$ of $g_2\in so(2,1)$ is a Lie symmetry (16) of H for $\kappa=-1$

$$[R(g_2),H]=-H. \tag{17}$$

It is just this higher-order Lie symmetry of H that is ''responsible'' for $\mathcal{F}_{\mathrm{SGA}}^{(2)}$. We add that $R(g_j)$, $j=1,3$ can be constructed if $R(g_2)$ is known.

C. The reduction lemma

Because a first-order Lie symmetry $Q^{(1)}$ of H is easier to analyze than an nth-order one, it would be convenient to reduce the order of $Q^{(n)}$ under the condition that no information is lost. This is possible if we use, in addition to H, the operator $H+\epsilon\mathbf{1}$. The following result holds.

Lemma 1: The Schrödinger equation $Hf=0, H=-d^2/dx^2+V(x)$ *in* $\mathcal{G}\supset\Gamma_H\cup\Gamma_{H+\varepsilon}$ *allows an* nth-*order Lie symmetry* $Q^{(n)}$ *of type* C_0 *if and only if*

$$\left(-\frac{d^2}{dx^2}+V(x)+\varepsilon\right)f=0, \quad \varepsilon\neq 0, \quad \text{real}, \tag{18}$$

admits a first-order Lie symmetry $\hat{Q}^{(1)}$ *of the form*

$$\hat{Q}^{(1)}=a(x,\varepsilon)\frac{d}{dx}+b(x,\varepsilon)\equiv\left(\sum_{j=0}^{N} a_j(x)\varepsilon^j\right)\frac{d}{dx}+\sum_{j=0}^{N} b_j(x)\varepsilon^j, \tag{19}$$

where $N=[(n-1)/2]$, $a_N=1$, *and the relations*

$$a''(x,\varepsilon)+2b'(x,\varepsilon)=0, \quad b''(x,\varepsilon)+a(x,\varepsilon)V'(x)+2a'(x,\varepsilon)(V(x)+\varepsilon)=0, \tag{20}$$

hold. Primes denote differentiations with respect to x.

Proof: (1) Consider $Q^{(n)}$. On $\mathcal{G}$ the commutator $[Q^{(n)},H]=0$ yields

$$\sum_{j=0}^{n} q_j^{(n)}\left[\frac{d^j}{dx^j},V(x)\right]+\sum_{j=0}^{n}\left(q_j^{(n)\prime\prime}+2q_j^{(n)\prime}\frac{d}{dx}\right)\frac{d^j}{dx^j}=0.$$

Equating the coefficients of d^{n+1}/dx^{n+1}, we get $q_n^{(n)}=\text{const}$. Without loss of generality we may choose $q_n^{(n)}=1$,

$$Q^{(n)}=\frac{d^n}{dx^n}+\sum_{j=0}^{n-1} q_j^{(n)}\frac{d^j}{dx^j}. \tag{21}$$

We use this and $Q^{(n)}\subset\mathcal{L}$ to construct for even n an equivalent $Q^{(n-1)}$,

$$Q^{(n-1)}=\sum_{j=0}^{n-1} q_j^{(n-1)}\left(\frac{d}{dx}\right)^j.$$

Hence, it is necessary to consider only odd n, i.e., $n=2N+1, N\in\mathbf{N}$. As $Q^{(n)}$ commutes with H it commutes with $H+\varepsilon\mathbf{1}$.

Our aim is to reduce the power n of $Q^{(n)}$ (which is a Lie symmetry with $P\equiv 0$) in an iterative procedure for the price that $P\neq 0$. Assume that $Q^{(n)}$ is a Lie symmetry of $H+\varepsilon$. Introduce a differential operator R (see Ref. 10) such that $Q^{n-1}=Q^{(n)}+R(H+\varepsilon)$ is again a Lie symmetry but of the order $n-2$. We find for R,

$$R=-q_n\left(\frac{d}{dx}\right)^{n-2},$$

and coefficients of Q^{n-1} are first-order polynomials in ε. Repeating this procedure $N-1$ times yields a first-order Lie symmetry $\hat{Q}^{(1)}$ whose coefficients are polynomials in ε of the order N, i.e., we get for $\hat{Q}^{(1)}$ an expression of the form (19). One shows by induction that $a_N=1$. The relation (20) for a_j, b_j are implied by the fact that $\hat{Q}^{(1)}$ is a Lie symmetry for $H+\varepsilon$,

$$[\hat{Q}^{(1)},H+\varepsilon]=R(H+\varepsilon), \tag{22}$$

with R as a multiplication operator.

(2) Start with $\hat{Q}^{(1)}=\Sigma_{j=0}^{N}A_j\varepsilon^j$, $A_j=a_j(x)(d/dx)+b_j(x)$. We want to cancel the ε^j and to introduce high-order differential operators through a stepwise procedure. Construct from $\hat{Q}^{(1)}$,

$$Q_1^{(2N+1)}=\hat{Q}^{(1)}-A_N(H+\varepsilon)^N=\sum_{j=0}^{N-1} A_{j,1}\varepsilon^j.$$

Because of (22) this is a Lie symmetry of order $2N+1$,

$$[Q_1^{(2N+1)},H+\varepsilon]=R_1(H+\varepsilon);$$

R_1 is an ε-dependent differential operator. A corresponding property holds for

$$Q_2^{(2N+1)}=Q_1^{(2N+1)}-A_{N-1,1}(H+\varepsilon)^{N-1},$$

with some R_2. Continuing this process we get after $N-2$ steps an ε-independent $Q_N^{(2N+1)}$ with

$$[Q_N^{(2N+1)},H+\varepsilon]=[Q_N^{(2N+1)},H]=R_N(H+\varepsilon).$$

The left-hand side is independent of ε, which implies that $R_N=0$. ▷

D. The family $\mathcal{F}_L^0$ of potentials that allow Lie symmetries Q of type C_0

The reduction lemma restricts the potentials $V(x)$ that allow an nth-order Lie symmetry of type C_0. We show that this set is the solution set of a family of nonlinear differential equations. Integrodifferential operators are applied in order to characterize the structure of this family.

Our main result is the following.

Theorem 1: *The Schrödinger equation $Hf=0, H=-d^2/dx^2+V(x)$ admits an nth-order Lie symmetry of type C_0 if and only if the potential $V(x)$ satisfies a nonlinear differential equation contained in the family*

$$\mathcal{F}_L^0=\left\{ V(x) \middle| G(V)\equiv\sum_{j=0}^{N-1} c_j F_j+F_N=0\right\}, \tag{23}$$

where $N=[(n-1)/2]$; $F_j=\mathcal{R}^j F_0$, $F_0=-\frac{1}{2}V'$, i.e., the F_j represent the stationary KdV hierarchy (13), and c_j are real constants.

Proof: The proof is simplified substantially if we use Lemma 1 and well-known techniques of soliton theory.

In the reduction lemma we insert the polynomial ε dependence for a_j, b_j and find, after integration for b_j (B_j are integration constants),

$$b_j(x)=-\tfrac{1}{2}a_j'(x)+B_j, \quad j=0,1,...,N, \tag{24}$$

and $N+2$ recurrence relations for a_j depending on $V(x)$ and on its derivatives,

$$a_N(x)=1, \quad a_{j-1}'(x)=-\tfrac{1}{4}a_j'''(x)-V(x)a_j'(x)-\tfrac{1}{2}V'(x)a_j(x), \quad a_{-1}(x)\overset{\text{def}}{=}0, \tag{25}$$

where $j=0,1,...,N$.

The first $N+1$ relations of (25) are solved by subsequent integrations yielding expressions for the functions $a_0(x),\dots,a_{N-1}(x)$ through $V(x)$ and its derivatives. Substituting these results into the last equation for $j=0$, i.e., $a_{-1}=0$, we arrive at a nonlinear differential equation for $V(x)$ of order $2N+1$. Its solutions generate all solutions of (24) and (25).

To reveal the structure of (25), especially of $a_{-1}=0$, we introduce new functions $\mathcal{U}_0(x)$, $\mathcal{U}_1(x),\dots,$ by the following recurrence relation (D_x denotes d/dx):

$$\mathcal{U}_j(x)=\mathcal{P}\mathcal{U}_{j-1}, \quad \mathcal{U}_{-1}\equiv 1, \quad \mathcal{P}=-\tfrac{1}{4}D_x^2-V(x)+\tfrac{1}{2}D_x^{-1}V'(x), \tag{26}$$

where $j=0,1,...,$ and $\mathcal{U}_{-1}(x)\overset{\text{def}}{=}1$. Note that $\mathcal{P}$ is the first recursive (integrodifferential) operator for the KdV equation (see, e.g., Refs. 7–9). The action of $\mathcal{P}$ on some (initial conserved density) $\mathcal{U}_0=-\frac{1}{2}V(x)$ yields the whole hierarchy of the conserved densities $\mathcal{U}_1,\mathcal{U}_2,\dots$. The second recursive operator $\mathcal{R}$ in (12) is related to $\mathcal{P}$ through

$$\mathcal{R}=D_x\circ\mathcal{P}\circ D_x^{-1}.$$

With this $\mathcal{U}_j(x)$ the $a_j(x)$ in (25) can be written as

$$a_{j-1}(x)=\mathcal{U}_{N-j}+\sum_{k=1}^{N-j+1} c_{N-k}\mathcal{U}_{N-j-k}, \quad j=1,...,N, \tag{27}$$

where $c_0,\dots,c_{N-1}$ are integration constants (independent of the B_j).

The equation $a_{-1}=0$ can be rewritten in a more transparent form as

$$D_x\circ\left(\sum_{j=0}^{N-1} c_j\mathcal{P}^j+\mathcal{P}^N\right)\mathcal{U}_0=0, \tag{28}$$

which shows a direct relation to the stationary higher KdV equation. Using the operator identity,

$$D_x \circ \mathcal{P}^j \equiv (D_x \circ \mathcal{P} \circ D_x^{-1})^j \circ D_x = \mathcal{R} \circ D_x,$$

(28) takes the following form:

$$\left(\sum_{j=0}^{N-1} c_j \mathcal{R}^j + \mathcal{R}^N \right) \circ D_x \mathcal{U}_0 = 0, \quad \mathcal{U}_0 = -\frac{1}{2} V'(x).$$

Because of $D_x \mathcal{U}_0 = -\frac{1}{2} V' = F_0$ and $F_j = \mathcal{R}^j F_0$ we get finally (23). ▷

E. The family $\mathcal{F}_L^\kappa$ of potentials that allow Lie symmetries of type C_κ

The results in Sec. III D for nth-order Lie symmetries of type C_0 can be utilized to construct also nth-order Lie symmetries of type C_κ for $\kappa \neq 0$. We have the following.

Theorem 2: *The Schrödinger equation* $Hf = 0, H = -d^2/dx^2 + V(x)$ *admits an* nth-*order Lie symmetry* $Q \in \mathcal{L}$ *of type* C_κ, *i.e.*, $[Q,H] = \kappa H$, *if and only if the potential* $V(x)$ *satisfies a nonlinear differential equation contained in the family*

$$\mathcal{F}_L^\kappa = \left\{ V(x) \middle| \kappa \left(\frac{x}{2} V' + V \right) + \sum_{j=0}^{N-1} c_j F_j + F_N = 0 \right\}, \tag{29}$$

where $N = [(n-1)/2]$, $F_j = \mathcal{R}^j F_0$, $F_0 = -\frac{1}{2} V'$; c_j are real constants.

Proof: With the same arguments as in the proof of the reduction lemma we get for Q [see (14)] $q_n = 1$. It is only necessary to consider odd n. We insert this Q and H in (16):

$$\sum_{j=0}^{n-1} \binom{j}{n} V^{(n-j)} \frac{d^j}{dx^j} + \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} q_i \binom{j}{i} V^{(i-j)} \frac{d^j}{dx^j} + \sum_{j=0}^{n-1} \left(2 q_j' \frac{d}{dx} + q_j'' \right) \frac{d^j}{dx^j} = \kappa \left(-\frac{d^2}{dx^2} + V(x) \right). \tag{30}$$

Comparing coefficients in front of the linearly independent operators d^j/dx^j, $(j = 1,...,n)$ yields n recurrence integrodifferential relations for the coefficients $q_i(x) = q_i(x, V(x), \kappa)$, $(i = 0,1,...,n-1)$ in (14),

$$q_{n-1}(x) = \hat{c}_{n-1},$$

$$q_{j-1}(x) = -\frac{1}{2} \left(q_j'(x) + \binom{j}{n} V^{(n-j-1)}(x) + \sum_{i=j+1}^{n-1} \binom{j}{i} \int^x q_i(y) V^{(i-j)}(y) dy + \kappa \delta_{j2} x \right) + \hat{c}_{j-1}, \tag{31}$$

where $\hat{c}_j$ are integration constants and

$$\hat{c}_j = 0, \quad \text{for } j \text{ even}.$$

Note that only q_1 depends on κ through an additive part $\frac{1}{2}\kappa x$, i.e.,

$$q_1(x, V(x), \kappa) = \hat{q}_1(x, V(x)) + \tfrac{1}{2} \kappa x.$$

Collecting the terms with $(d/dx)^0$ in (30), we get a differential equation $G(V,\kappa)$ for $V(x)$ of the type

$$G(V,\kappa) \equiv q_0'' + V^{(n)} + \hat{q}_1 V' + \kappa \frac{x}{2} V' + \sum_{j=2}^{n-1} q_j V^{(j)} + \kappa V = 0. \tag{32}$$

From Theorem 1 we know the differential equation for $V(x)$ if $\kappa = 0$. Hence

$$G(V,0)=G(V). \tag{33}$$

Because the κ dependence in (32) is explicitly known, we have

$$G(V,\kappa)\equiv G(V)+\kappa\left(\frac{x}{2}V'+V\right)=0; \tag{34}$$

so we arrive at (29). ▷

This concludes the discussion of high-order Lie symmetry of the Hamiltonian $H=-d^2/dx^2+V(x)$.

IV. RELATION TO SPECTRUM GENERATING ALGEBRAS

The Lie symmetry $[Q,H]=\kappa H$ (Theorem 2) is related to the spectrum generating algebra $so(2,1)$ of H through (17), where $\kappa=-1$. The realization $R(g_2)$ of $g_2\in so(2,1)$ is given by Q of the form (14), (31) through solutions of (29) with $\kappa=-1$ and

$$\mathcal{F}^{(2)}_{\mathrm{SGA}}=\mathcal{F}^{\kappa}_{L}, \quad \kappa=-1 \tag{35}$$

holds. As $R(g_2)$ is explicitly known, we can insert $R(g_2)$ and the differential operators $R(g_1)$, $R(g_3)$ into the commutation relations of the algebra $so(2,1)$ and thus find the latter (for further details, see Refs. 1 and 2).

With the results obtained in Sec. III we can elucidate the peculiar features of the nonlinear differential equation (9) mentioned in Sec. II. The fact that the potentials $V(x)$ are identical for $n=2N+1$ and $n=2N+2$ is explained as follows. In a symmetry operator Q of order $2N+2$, the term with $(d/dx)^{2N+2}$ can be canceled because of $Q\in\mathcal{L}$, thus getting a symmetry operator of order $2N+1$ that is again a Lie symmetry of the Schrödinger equation. The stronger statement that the equations for $V(x)$ in $\mathcal{F}^{(2)}_{\mathrm{SGA}}$ are identical for $n=4k$, $4k+1$, $4k+2$, $4k+3$, $k=1,2,...$, is valid because a SGA symmetry is stronger than a Lie symmetry. The way of constructing the equations for $V(x)$ used while proving Lemma 1, makes it also evident, why this equation with some fixed $n=n_1$ contains all the terms of an equation for $V(x)$ under $n=n_2<n_1$. Indeed, the equation for $n=n_1$ is obtained from one for n_1-1 by the action of the recursive operator $\mathcal{R}$, which transforms a term F_j into F_{j+1}.

V. INTEGRABILITY

Hamiltonians (10) admitting $so(2,1)$-spectrum generating algebra have a further useful property: they are *integrable* in the sense that the corresponding Schrödinger equation (11) can be integrated by quadratures. This is so because (11) admits a first-order Lie symmetry of the form $\hat{Q}=\xi(x)(d/dx)+\eta(x)$, with

$$-\eta''+2V\xi'+V'\xi=0, \quad 2\eta'+\xi''=0, \tag{36}$$

and because one can apply for an integration of (11) the classical method (see, e.g., Refs. 10 and 11) based on its Lie symmetry. The first integral for system (36) is

$$-\eta'(\xi)+V\xi^2-\tfrac{1}{4}=\alpha\equiv\text{const.}$$

Depending on the sign of α, the general solution of the equation (11) reads as

$$f(x)=\sqrt{\xi(x)}\begin{cases} C_1\rho(x)+C_2, & \alpha=0,\\ C_1\cos a\,\rho(x)+C_2\sin a\,\rho(x), & \alpha=a^2>0,\\ C_1\cosh a\,\rho(x)+C_2\sinh a\,\rho(x), & \alpha=-a^2<0,\end{cases}$$

where

$$\rho(x)=\int \frac{dx}{\xi(x)}.$$

Now, inserting the explicit expressions for $\xi(x)$, $\eta(x)$ from (19) into the above formulas yields the general solution of (11), provided the function $V(x)$ fulfills an equation of the form (9).

VI. CONCLUDING REMARKS

Given a physical observable quantized through a linear differential operator A in $L^2(\mathbf{R}^d,dx^d)$, e.g., a Hamiltonian H (1) in one space dimension, a spectrum generating algebra for A is specified through a Lie algebra L with $\dim L=m$ with generators g_i and a realization through differential operators of the order n such that $A=\Sigma_{k=1}^{m}\alpha_k R(g_k)$, e.g., $L=so(2)$, $m=3$. A high-order Lie symmetry for the linear operator A is defined through finite-order differential operators Q,P with $[Q,A]=PA$, e.g., $A=H$ and $P=\kappa$. Both SGA and high-order Lie symmetry are different methods to model a physical symmetry with different mathematical structures.

We have shown that for the Hamiltonian (1) an $so(2,1)$ SGA and an nth-order Lie symmetry with $P=-1$ are directly related via (17). However, a Lie symmetry is more general than an $so(2,1)$ SGA symmetry. The interesting result is the connection between an nth-order C_κ Lie-symmetry of the Hamiltonian (1) and the stationary KdV hierarchy. It is understandable that for the singular case of $so(2,1)$, which reflects the light cone structure in $so(2,1)$, a family of nonlinear differential equations for $V(x)$ appears. But it is, as we already mentioned, surprising that this family is related to the stationary KdV hierarchy, a mathematical object not connected directly to a symmetry concept of observables. We suspect that this connection to the KdV hierarchy is somehow encoded in the geometry of $so(2,1)$ and its realizations. An investigation of Hamiltonians of the type (1) in higher space dimensions, their Lie symmetry in the above sense, and L-SGA with noncompact L, $m>3$, seems to be appropriate.

ACKNOWLEDGMENTS

One of the authors (R.Zh.) gratefully acknowledges support from the Alexander von Humboldt Foundation and the Technical University of Clausthal.

[1] H.-D. Doebner and B. Pirrung, ''Spectrum generating algebras and canonical realizations,'' International Center for Theoretical Physics, Trieste Report IC/72/77, 1972.
[2] H.-D. Doebner and B. Pirrung, ''A new class of Hamiltonians with $so(2,1)$ as spectrum generating algebra,'' International Center for Theoretical Physics, Trieste Report IC/75/112, 1972.
[3] H.-D. Doebner and H.-J. Mann, in *Proceedings of the VIth Wigner Symposium*, edited by N. M. Atakashiev, T. H. Seligman, and K. B. Wolf (World Scientific, Singapore, 1996), p. 11–23.
[4] A. Barut and A. Böhm, Phys. Rev. B **139**, 1107 (1965).
[5] Y. Dotham, M. Gell-Mann, and Y. Ne'eman, Phys. Lett. **17**, 48 (1965).
[6] A. O. Barut, A. Bohm, and Y. Ne'eman, *Dynamical Groups and Spectrum Generating Algebras* (World Scientific, Singapore, 1988).
[7] N. Kh. Ibragimov, *Transformation Groups Applied to Mathematical Physics* (Reidel, Dordrecht, 1985).
[8] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[9] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevski, *Theory of Solitons: the Inverse Scattering Method* (Consultants Bureau, New York, 1980).
[10] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer-Verlag, New York, 1986).
[11] E. Cartan, *Les Systémes Differentiels Extèrieurs et Leur Applications Géometriques* (Hermann, Paris, 1945).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Algebraic exact solvability of trigonometric-type Hamiltonians associated to root systems

Niky Kamran[a)]
Department of Mathematics, McGill University, Montreal, Quèbec H3A 2K6, Canada

Robert Milson[b)]
Department of Mathematics, Dalhousie University, Halifax, Nova Scotia B3J 3J5 Canada

(Received 3 November 1998; accepted for publication 30 June 1999)

In this article, we study and settle several structural questions concerning the exact solvability of the Olshanetsky–Perelomov quantum Hamiltonians corresponding to an arbitrary root system. We show that these operators can be written as linear combinations of certain basic operators admitting infinite flags of invariant subspaces, namely the Laplacian and the logarithmic gradient of invariant factors of the Weyl denominator. The coefficients of the constituent linear combination become the coupling constants of the final model. We also demonstrate the L^2 completeness of the eigenfunctions obtained by this procedure, and describe a straightforward recursive procedure based on the Freudenthal multiplicity formula for constructing the eigenfunctions explicitly.
[S0022-2488(99)01110-X]

I. INTRODUCTION

The potentials first discovered by Calogero and Sutherland[1,2] and subsequently generalized to arbitrary root systems by Olshanetsky and Perelomov[3] play a central role in the theory of classical and quantum completely integrable systems. One of the main themes of the original work by Olshanetsky and Perelomov was to establish quantum complete integrability, that is, the existence of complete sets of commuting operators. The actual eigenfunctions of the corresponding Hamiltonians were discussed in numerous subsequent publications.[4–7]

Our purpose in this paper is study and settle a certain number of basic structural questions concerning the exact solvability of the Olshanetsky-Perelomov Hamiltonians. In order to outline the main results of our paper, we first need to give a precise definition of what we mean by exact solvability. We will adopt a promising approach, which has recently arisen in the framework of the theory of quasiexactly solvable potentials,[8–11] by defining a quantum Hamiltonian $\mathcal{H}$ to be *algebraically exactly solvable* if one can explicitly construct an ordered basis for the underlying Hilbert space such that the corresponding flag of subspaces is $\mathcal{H}$ invariant. In terms of this approach, the first step in the treatment of an exactly solvable operator must be the construction of an infinite flag of finite-dimensional vector spaces ordered by inclusion, the determination of a collection of basic operators that preserve this flag, and the demonstration that the operator in question is generated by the basic ones. The second step is to prove the L^2 completeness in the underlying Hilbert space of this family of subspaces.

In order to fit the Olshanetsky–Perelomov Hamiltonians of trigonometric type into this framework, we first recall that these Hamiltonians are indexed by irreducible root systems, with the Calogero–Sutherland potentials corresponding to type A_n root systems. We thus consider the vector space of trigonometric functions that are invariant under the Weyl group W of the given root system R. The partial order relation on dominant weights gives rise to a natural flag of finite-dimensional subspaces of this infinite-dimensional vector space. It is quite evident that the

[a)]Electronic mail: nkamran@math.mcgill.ca
[b)]Current address: McGill University, Montreal. Electronic mail: milson@math.mcgill.ca

flag in question is preserved by the ordinary, multidimensional Laplacian. Less evident is the fact that one can obtain other flag-preserving operators by factoring the Weyl denominator,

$$A=\prod_{\alpha\in R^+} e^{\alpha/2}-e^{-\alpha/2},$$

into factors corresponding to the various orbits of the Weyl group on R. It turns out (see Proposition 12) that the gradient of the logarithm of each of the resulting factors also preserves the flag in question. More generally, one obtains other flag-preserving second-order operators by taking linear combinations of the Laplacian and of these gradients. The Olshanetsky–Perelomov Hamiltonians are then obtained by a ground-state conjugation. This approach also sheds light on the presence of multiple coupling constants in some of the models; the number of coupling constants is precisely the number of invariant factors of A, i.e., the number of Weyl group orbits in R, or, equivalently, the number of distinct root lengths. We then show that if all the coupling constants are positive, then the action of the Hamiltonian on each subspace of the flag is diagonalizable. This is the first main result of our paper; it is given in Theorem 1. The second main result concerns the L^2 completeness of the resulting eigenfunctions in the underlying Hilbert space of L^2 functions on the alcove of the root system R.

It is also interesting to note that if all the coupling constants are equal to 1, then one recovers a second-order differential operator whose eigenfunctions are precisely the characters of the corresponding simple Lie algebras. For certain other values of the coupling constants, one recovers the spherical functions associated to any symmetric space G/K, where G is a semisimple real Lie group and K is a suitable compact subgroup. If the restricted root system of the symmetric space is of type A_{n-1} and m is the multiplicity of each restricted root, then the eigenfunctions corresponding to the value $k_c=m/2$ of the deformation parameter are the zonal spherical functions on G/K, as pointed out by Macdonald.[12,13] Thus the coupling constants can be regarded as parameters in a deformation of the classical characters.

In the classical case, if one reexpresses the gradient of $\log A$ in terms of a formal power series, one obtains Freudenthal's recursion formula for the character coefficients. This trick also works for the deformed characters, and leads to a recursion formula that allows one to straightforwardly compute the eigenfunctions of the Olshanetsky–Perelomov Hamiltonians. This result is presented in Sec. IV.

We should point out that the Weyl-invariant deformed characters that appear in the expressions of the eigenfunctions of the Olshanetsky–Perelomov trigonometric Hamiltonians are related by a change of variables to the multivariate Jacobi polynomials that have been investigated by Heckman and Opdam.[14] In particular, the analog of the Freudenthal multiplicity formula that is at the basis of the recursion formula we give in Proposition 19 for the eigenfunctions of the Hamiltonians also appears in the context of their study. We should also mention the interesting recent contributions of Brink, Turbiner, and Wyllard[15] in the general effort aimed at understanding the exact solvability for multidimensional systems in an algebraic context.

II. TRIGONOMETRIC-TYPE POTENTIALS ASSOCIATED TO ROOT SYSTEMS

We first recall the abstract definition of the trigonometric Olshanetsky–Perelomov Hamiltonians in terms of root systems. Let $\mathbf{V}$ be a finite-dimensional real vector space endowed with a positive-definite inner product $(u,v)\in\mathbb{R}$, u, $v\in\mathbf{V}$. We use this inner product to identify $\mathbf{V}$ with $\mathbf{V}^*$. The induced positive-definite inner product on $\mathbf{V}^*$ will also be denoted by $(\cdot,\cdot)$. Let $\Delta:C^\infty(\mathbf{V};\mathbb{R})\to C^\infty(\mathbf{V};\mathbb{R})$ and $\boldsymbol{\nabla}:C^\infty(V;\mathbb{R})\to\Gamma(TV)$ denote the corresponding Laplace–Beltrami and gradient operators.

For a nonzero $\alpha\in\mathbf{V}^*$, we set $\check{\alpha}=2\alpha/(\alpha,\alpha)$ and let s_α denote the reflection across the hyperplane orthogonal to α:

$$s_\alpha(\beta)=\beta-(\check{\alpha},\beta)\alpha,\quad \beta\in\mathbf{V}^*.$$

By a root system, we mean a finite, spanning subset R of $\mathbf{V}^*$ such that $0 \in R$, $s_\alpha(R) \subset R$ for all $\alpha \in R$ and $(\breve{\alpha},\beta) \in \mathbb{Z}$ for all $\alpha,\beta \in R$. A root system R is said to be irreducible if it cannot be partitioned into a union of root systems spanning orthogonal subspaces of $\mathbf{V}$.

To any root system R corresponds a root lattice $Q = \{\Sigma_R m_\alpha \alpha : m_\alpha \in \mathbb{Z}\}$ and a weight lattice $P = \{\lambda \in \mathbf{V}^* : (\breve{\alpha},\lambda) \in \mathbb{Z}, \forall \alpha \in R\}$. The Weyl group of R, generated by $s_\alpha, \alpha \in R$, will be denoted by W. The subgroup of W fixing a particular $\lambda \in \mathbf{V}^*$ will be denoted by W_λ.

The hyperplanes $\{\lambda \in \mathbf{V}^* : (\alpha,\lambda) = 0\}, \alpha \in R$ define a set of open Weyl chambers in $\mathbf{V}^*$. We choose a Weyl chamber C and let $R^+ = R \cap \overline{C}$ denote the corresponding subset of positive roots. Let $B \subset R^+$ denote the set of simple roots, i.e., the positive roots that cannot be written as the sum of two positive roots. Let $P^+ = R \cap \overline{C}$ denote the set of dominant weights.

We will say that a real number $c > 0$ is a *root length* if there exists a $\alpha \in R$ such that $c = \|\alpha\|$. Let c be a root length, and set

$$R_c = \{\alpha \in R : \|\alpha\| = c\},$$

$$R_c^+ = R_c \cap R^+,$$

$$U_c = \frac{c^2}{4} \sum_{\alpha \in R_c^+} \cos^2 \frac{\alpha}{2}.$$

Note[16] that if c is a root length, then, R_c is nothing but the W orbit of α.

The Olshanetsky–Perelomov Hamiltonians with trigonometric potentials associated to a root system R are defined in terms of the above data by

$$\mathcal{H} = -\Delta + \sum_c a_c U_c,$$

where the sum is taken over all root lengths, c, and where the a_c's are real coupling constants.

III. THE ALGEBRAIC EXACT SOLVABILITY OF $\mathcal{H}$

The affine hyperplanes $\{\lambda \in \mathbf{V}^* : (\alpha,\lambda) \in 2\pi\mathbb{Z}\}$ determine in $\mathbf{V}^*$ a set of isometric open bounded subsets called alcoves. Let A denote the unique alcove (usually referred to as the fundamental alcove) that is contained in C and that has the origin as a boundary point. Let m denote the Lebesgue measure on A. From now on we use the inner product to identify A with the corresponding subset of $\mathbf{V}$ and restrict the domain of functions introduced subsequently to A. Our goal is to construct a basis for the underlying Hilbert space $L^2(A,m)$ in which the algebraic exact solvability of $\mathcal{H}$ is manifest. The elements of this basis will be products of W-invariant trigonometric functions of certain linear forms on $\mathbf{V}$ with a common gauge factor vanishing along the walls $\{u \in \mathbf{V} : \alpha(u) \in 2\pi\mathbb{Z}\}$, $\alpha \in R$ of the potential terms U_c.

We now proceed to define this basis. Recall that a choice of positive roots naturally induces a partial order relation, $\leqslant$, on the weight lattice. For $\lambda \in P^+$ set

$$P_\lambda = \cup_{w \in W} \{w(\mu) : \mu \in P^+ \text{ and } \mu \leqslant \lambda\},$$

$$P_{\lambda -} = \cup_{w \in W} \{w(\mu) : \mu \in P^+ \text{ and } \mu \lneq \lambda\}.$$

For $S \subset \mathbf{V}^*$ let trig(S) denote the complex vector space spanned by functions of the form $e^{i\lambda}, \lambda \in S$. If S is a W-invariant subset of $\mathbf{V}^*$, then there is a well-defined action of W on trig(S), namely

$$w \cdot e^{i\lambda} = e^{iw(\lambda)}, \quad w \in W, \quad \lambda \in S.$$

In this case, let $\mathrm{trig}(S)^W$ denote the subspace of W-invariant functions.

Recall that a root system R is said to be reduced if for every $\alpha \in R$, the only roots homothetic to α are $-\alpha$ and α itself. A root α will be called nondivisible if $\alpha/2$ is not a root. Similarly, α will be called nonmultiplicable if 2α is not a root. Of course, if R is reduced, then all roots are both nondivisible and nonmultiplicable. An irreducible nonreduced system must be isomorphic to a root system of type BC_n for some n. To describe the latter, take $\mathbf{V} = \mathbb{R}^n$ and let $\epsilon_1, \dots, \epsilon_n$ denote the dual basis of the standard basis of $\mathbb{R}^n$. The root system in question consists of three types of roots: short roots $\pm \epsilon_i$, medium roots $\pm \epsilon_i \pm \epsilon_j$, $i \neq j$, and long roots $\pm 2\epsilon_i$.

For reasons that will become clear later, it is convenient to reexpress the coupling constants a_c appearing in $\mathcal{H}$ as follows. We let $a_c = k_c(k_c - 1)$ if c is the length of a nonmultiplicable root, and $a_c = k_c(k_c + k_{2c} - 1)$ if R is nonreduced and c is the length of the short roots. Let

$$A_c = \prod_{\alpha \in R_c^+} \sin \frac{\alpha}{2}, \quad F = \prod_c |A_c|^{k_c}, \quad \rho_c = \frac{1}{2} \sum_{\alpha \in R_c^+} \alpha, \quad \rho = \sum_c k_c \rho_c .$$

The following theorems, which are the main results of our paper, shows that the Olshanetsky–Perelomov trigonometric Hamiltonians $\mathcal{H}$ are exactly solvable in the algebraic sense, and that the corresponding eigenfunctions are physically meaningful.

Theorem 1: *Let λ be a dominant weight. If $k_c \geqslant 0$ for each root length c, then there exists a unique $\phi_\lambda \in \mathrm{trig}(P_\lambda)^W$ such that $F\phi_\lambda$ is an eigenfunction of $\mathcal{H}$ with eigenvalue $\|\lambda + \rho\|^2$. Furthermore, if $F\phi, \phi \in \mathrm{trig}(P)^W$ is an eigenfunction of $\mathcal{H}$, then $\phi = \phi_\lambda$ for some $\lambda \in P^+$.*

Theorem 2: *The subspace $F\,\mathrm{trig}(P)^W$ is dense in $L^2(A,m)$. Moreover, if $k_c \geqslant 0$ for all root lengths c, then the operator $\mathcal{H}$ is essentially self-adjoint on the domain $F\,\mathrm{trig}(P)^W \subset L^2(A,m)$.*

We begin with the proof of Theorem 2, assuming Theorem 1 to be true. We first have the following.

Lemma 3: Let D be an open, bounded subset of Euclidean space, and $f: D \to \mathbb{R}$ a bounded continuous function that does not vanish on D (but may vanish on the boundary). With these assumptions, $fL^2(D,m)$ is a dense subset of $L^2(D,m)$.

Proof: Let D_0, an open subset of D, be given, and choose D_1 such that $\overline{D_1} \subset D_0$ and such that $m(D_0) - m(D_1)$ is smaller than a given $\epsilon > 0$. Note that $h = f^{-1}\chi_{D_1}$ is a well-defined element of $L^2(D)$ and that $fh = \chi_{D_1}$. Consequently, χ_{D_0} lies in the closure of $fL^2(D)$. The conclusion follows from the fact that the characteristic functions form a dense subset of $L^2(D)$. □

Proof of Theorem 2: Let $\mathbf{T}$ denote the torus $\mathbf{V}^*/(2\pi Q)$. We use the inner product on $\mathbf{V}$ to identify $\mathbf{T}$ with the identical quotient of $\mathbf{V}$. Recall that $\mathrm{trig}(P)$ is dense in $L^2(\mathbf{T})$ by the Fourier representation theorem. Now W acts on $\mathbf{T}$ and A serves as a fundamental region for this action (Ref. 17, Chap. VI, No. 2.1). Consequently, $\mathrm{trig}(P)^W$ is dense in $L^2(T)^W$ and the latter is naturally isomorphic to $L^2(A,m)$. We therefore conclude that $F\,\mathrm{trig}(P)^W$ is dense in $L^2(A,m)$ by applying the preceding Lemma with $f = F$.

We now prove the essential self-adjointness of $\mathcal{H}$ on the domain $F\,\mathrm{trig}(P)^W$. Let $A_0 \subset A$ be an open subset with a piecewise smooth boundary. Let $\phi_1, \phi_2 \in \mathrm{trig}(P)^W$ be given. Setting $\psi_i = F\phi_i$, $i = 1,2$, we have

$$\int_{A_0} \mathcal{H}(\psi_1)\psi_2 - \psi_1 \mathcal{H}(\psi_2) = \int_{A_0} \mathrm{div}(\psi_2 \nabla \psi_1 - \psi_1 \nabla \psi_2) = \int_{\partial A_0} F^2(\phi_2 \nabla \phi_1 - \phi_1 \nabla \phi_2).$$

Hence, as the boundary of A_0 approaches the boundary of A, the above integrals tend to zero, so that the operator $\mathcal{H}$ is a symmetric. By Theorem 1 and the density of $F\,\mathrm{trig}(P)^W$ in $L^2(A,m)$, the span of eigenfunctions of $\mathcal{H}$ is dense in $L^2(A)$, and therefore $\mathcal{H}$ must be essentially self-adjoint.□

We now proceed with the proof of Theorem 1. The strategy behind the proof of this theorem is to conjugate the Olshanetsky–Perelomov Hamiltonians $\mathcal{H}$ by a suitable multiplication operator chosen in such a way that the resulting operator has a simple action on the space $\mathrm{trig}(P)^W$. This will give rise to an essential intertwining relation that will, in turn, imply the algebraic exact solvability. In order to determine this multiplicative factor, we need a series of facts about root lengths.

Let $M_c : W \to \{\pm 1\}$ be the class function defined by

$$M_c(s_\alpha) = \begin{cases} -1, & \text{if} \quad \alpha \in B \cap R_c, \\ 1 & \text{if} \quad \alpha \in B \backslash R_c. \end{cases}$$

The following result is a straightforward consequence of the definition of A_c.

Proposition 4: For $w \in W$ one has $w(A_c) = M_c(w) A_c$. In other words, A_c is a relative invariant of W with multiplier M_c. Moreover, we have the following.

Proposition 5: Let c be a root length. If $\alpha \in B$, then $(\check{\alpha}, \rho_c)$ takes one of four possible values: 1 if $\|\alpha\| = c$, 2 if $\|\alpha\| = c/2$, 1/2 if $\|\alpha\| = 2c$, 0 in all other cases.

Proof: Let $\alpha \in B$ be given. The action of s_α maps α to $-\alpha$ and permutes the elements of R^+ not homothetic to α (Ref. 17, Chap. VI, No. 1.6). Let $\beta \in R_c^+$ be given and set $\beta' = s_\alpha(\beta)$. Note that if $\beta = \beta'$, then $(\check{\alpha}, \beta) = 0$; and that if $\beta' \neq \beta$, then $(\check{\alpha}, \beta + \beta') = 0$. If $\|\alpha\| \in \{c, 2c, c/2\}$, then α is not homothetic to any element of R_c, and hence one can break up ρ_c into subterms of length one and two such that each subterm is annihilated by $\check{\alpha}$. This proves the fourth assertion of the proposition. If $\|\alpha\| = c$, then ρ_c is the sum of $\alpha/2$ and a remainder perpendicular to $\check{\alpha}$. Consequently, $(\check{\alpha}, \rho_c) = 1$, thereby proving the first assertion. If $\|\alpha\| = c/2$, then 2α is also a root, and, consequently, ρ_c is the sum of α and a remainder perpendicular to $\check{\alpha}$. This implies the second assertion. The case three assertion is proven similarly. □

Corollary 6: If c is the length of a nonmultiplicable root, then ρ_c is a weight. If R is nonreduced, and c is the length of the short roots, then ρ_c is merely a half-weight.

Corollary 7: Let c be a root length. Then for all $\alpha \in R_c$, one has $(\check{\alpha}, \rho_c) \in \mathbb{Z}$.

Proof: If c is the length of a nonmultiplicable root, then the claim follows from the preceding corollary. Suppose then that $2c$ is also a root length. For $\alpha \in R_c$ note that $2(2\alpha)\check{} = \check{\alpha}$ and that $2\rho_c = \rho_{2c}$. Hence

$$(\check{\alpha}, \rho_c) = ((2\alpha)\check{}, \rho_{2c}).$$

Since 2α is nonmultiplicable, the right-hand side is an integer by the preceding corollary. □

Corollary 8: Let c be a root length and $w \in W$. Then, $w(\rho_c) \in Q - \rho_c$.

Proof: Note that

$$w(\rho_c) = \frac{1}{2} \sum_{\alpha \in R_c^+} \sigma_\alpha(w) \alpha,$$

where $\sigma_\alpha(w)$ is either 1 or -1. Hence, $\rho_c + w(\rho_c)$ is the sum of all $\alpha \in R_c^+$ such that $\sigma_\alpha(w) = 1$. □

We are now ready for the next step leading to the required intertwining relation, which is to show that $\mathrm{trig}(P_\lambda)^W$ is an invariant subspace of $\nabla \log|A_c|$. First, we have the following.

Proposition 9: Let c be a root length. If $\phi \in \mathrm{trig}(P - \rho_c)$ is a relative invariant of W with multiplier M_c, then $\phi = A_c \phi_0$ for some $\phi_0 \in \mathrm{trig}(P)^W$.

Proof: By assumption, $\phi_1 = e^{i\rho_e}\phi$ is an element of $\mathrm{trig}(P)$. Let $\alpha \in R_c^+$ be given. The first claim is that ϕ_1 is divisible by $e^{i\alpha} - 1$ in $\mathrm{trig}(P)$. By assumption, ϕ is a linear combination of expressions of the form $e^{i\lambda} - e^{i\lambda'}$, where $\lambda + \rho_c \in P$, and $\lambda' = s_\alpha(\lambda)$. Since λ is the difference of a weight and ρ_c, Corollary 7 shows that $(\check{\alpha}, \lambda) \in \mathbb{Z}$. By switching λ and λ', if necessary, one may assume without loss of generality that $-(\check{\alpha}, \lambda) \in \mathbb{N}$. The claim follows by noting that

$$e^{i\lambda} - e^{i\lambda'} = e^{i\lambda}(1 - e^{-i(\check{\alpha}, \lambda)\alpha}),$$

and by factoring the right-hand side in the usual fashion.

Note that trig(P) with the natural function multiplication is a unique factorization domain (Ref. 17, Chap. VI, No. 3.1). Hence, the preceding claim implies that there exists a $\phi_0 \in \mathrm{trig}(P)$, such that

$$\phi_1 = \phi_0 \prod_{\alpha \in R_c^+} (e^{i\alpha} - 1).$$

The proof is concluded by noting that up to a constant factor, A_c is equal to

$$e^{-i\rho_c} \prod_{\alpha \in R_c^+} (e^{i\alpha} - 1).$$

The W invariance of ϕ_0 follows from the fact that A_c and ϕ are relative invariants with the same multiplier. □

We have:

Corollary 10: Let c be a root length. One has

$$(2i)^{\sharp R_c} A_c = \frac{1}{\sharp W_{\rho_c}} \sum_{w \in W} M_c(w) e^{iw(\rho_c)}. \tag{1}$$

Proposition 11: The differential operator $\nabla \log|A_c|$ *has a well-defined action on* $\mathrm{trig}(P)^W$.

Proof: Let $\phi \in \mathrm{trig}(P)^W$. The claim is that $(\nabla \log|A_c|)(\phi) \in \mathrm{trig}(P)^W$. By Corollaries 8 and 10, $A_c \in \mathrm{trig}(Q - \rho_c)$, and hence $\nabla A_c(\phi) \in \mathrm{trig}(P - \rho_c)$. Since ∇ is a W-invariant operator, $\nabla A_c(\phi)$ is a relative invariant of W with multiplier M_c. Hence, by Proposition 9, there exists a $\phi_0 \in \mathrm{trig}(P)^W$ such that $\nabla A_c(\phi) = A_c \phi_0$. □

We now have the following.

Proposition 12: If $\lambda \in P^+$, *then* $\mathrm{trig}(P_\lambda)^W$ *is an invariant subspace of* $\nabla \log|A_c|$.

Proof: Let $\phi \in \mathrm{trig}(P_\lambda)^W$ be given. Set $\phi_0 = (\nabla \log|A_c|)(\phi)$. By Proposition 11, $\phi_0 \in \mathrm{trig}(P)^W$. Let μ be a maximal element of $\mathrm{supp}(\phi_0)$. Consequently, $\mu + \rho_c$ is a maximal element of $\mathrm{suppl}(A_c \phi_0)$. Now

$$A_c = b_1 e^{i\rho_c} + \text{lower-order terms},$$

$$\phi = b_2 e^{i\lambda} + \text{lower-order terms},$$

where b_1, b_2 are nonzero constants, and hence,

$$(\nabla A_c)(\phi) = -b_1 b_2 (\rho_c, \lambda) e^{i(\rho_c + \lambda)} + \text{lower-order terms}.$$

Since $(\rho_c, \lambda) > 0$, one must have $\rho_c + \lambda = \rho_c + \mu$. Therefore $\mu = \lambda$, and $\phi_0 \in \mathrm{trig}(P_\lambda)^W$.

The basic identity that will give rise to the intertwining relation that we are looking for is given in the following proposition.

Proposition 13: Let $f_1, \ldots, f_n$ *be smooth real-valued functions on* $\mathbf{V}$; *let* $k_1, \ldots, k_n$ *be real constants; and let*

$$X = \sum_{i=1}^{n} 2k_i \nabla \log|f_i|, \quad F = \prod_{i=1}^{n} |f_i|^{k_i}.$$

We have the identity

$$F(-\Delta - X) = (-\Delta + U)F,$$

where

$$U=\sum_i k_i(k_i-1)\frac{\|\nabla f_i\|^2}{f_i^2}+\sum_{i\neq j} k_ik_j\frac{(\nabla f_i,\nabla f_j)}{f_if_j}+\sum_i k_i\frac{\Delta f_i}{f_i}.$$

The application of this proposition to the Olshanetsky–Perelomov Hamiltonians $\mathcal{H}$ requires a number of intermediate formulas.

Proposition 14: Let c be a root length. One has

$$\Delta A_c=-\|\rho_c\|^2A_c, \tag{2}$$

$$\|\nabla A_c\|^2=(U_c-\|\rho_c\|^2)A_c^2. \tag{3}$$

Proof: Note that for $\lambda\in\mathbf{V}^*$ one has $\Delta e^{i\lambda}=-\|\lambda\|^2e^{i\lambda}$. Formula (2) follows immediately from (1). Note that

$$\nabla A_c=\frac{A_c}{2}\sum_{\alpha\in R_c^+}\cot\frac{\alpha}{2}\nabla\alpha. \tag{4}$$

Consequently,

$$\|\nabla A_c\|^2=\left(\frac{c^2}{4}\sum_\alpha\cot^2\frac{\alpha}{2}+\frac{1}{4}\sum_{\alpha\neq\beta}(\alpha,\beta)\cot\frac{\alpha}{2}\cot\frac{\beta}{2}\right)A_c^2. \tag{5}$$

Taking the divergence of (4), one obtains

$$\frac{\Delta A_c}{A_c}=-\frac{(\sharp R_c)c^2}{4}+\frac{1}{4}\sum_{\alpha\neq\beta}(\alpha,\beta)\cot\frac{\alpha}{2}\cot\frac{\beta}{2}.$$

Solving for the second term of the right-hand side of the latter equation, substituting into (5) and applying (2), we obtain (3). □

Proposition 15:If c_1, c_2 are distinct root lengths such that the corresponding roots are not homothetic, then

$$(\nabla A_{c_1},\nabla A_{c_2})=-(\rho_{c_1},\rho_{c_2})A_{c_1}A_{c_2}. \tag{6}$$

If R is nonreduced and c is the length of the short roots, then

$$(\nabla A_c,\nabla A_{2c})=[U_c-(\rho_c,\rho_{2c})]A_cA_{2c}. \tag{7}$$

Proof: Let c_1, c_2 be given. A straightforward generalization of the argument in Proposition 9 yields

$$A_{c_1}A_{c_2}=\frac{1}{\sharp W_{\rho_{c_1}+\rho_{c_2}}}\sum_{w\in W}M_{c_1}(w)M_{c_2}(w)e^{iw(\rho_{c_1}+\rho_{c_2})}.$$

Hence,

$$\Delta(A_{c_1}A_{c_2})=-\|\rho_{c_1}+\rho_{c_2}\|^2A_{c_1}A_{c_2},$$

and the desired conclusion follows immediately from the usual product rule for the Laplacian.

Next, assume that the second of the proposition's hypotheses holds. Set $S_c=\Pi_{\alpha\in R_c}\cos(\alpha/2)$, and note that $A_{2c}=2A_cS_c$. Since R is of type BC_n, a direct calculation will show that $\Delta S_c=-\|\rho_c\|^2S_c$. Consequently,

$$2(\nabla A_c,\nabla S_c)=\tfrac{1}{2}\Delta A_{2c}-A_c\Delta S_c-S_c\Delta A_c=-\|\rho_c\|^2A_{2c}.$$

$$(\nabla A_c, \nabla A_{2c}) = -\|\rho_c\|^2 A_{2c} + 2S_c \|\nabla A_c\|^2.$$

The formula to be proved now follows from (3). □

We can now state and prove the intertwining relation, which is fundamental to the proof of our main result.

Proposition 16: Let

$$\tilde{\mathcal{H}} = -\Delta - \sum_c 2k_c \nabla \log|A_c|.$$

We have

$$F\tilde{\mathcal{H}} = \mathcal{H}F - \|\rho\|^2.$$

Proof: Apply Propositions 13, 14, and 15. □

Finally, we are ready to give the proof of Theorem 1, that is of the algebraic exact solvability of the Olshanetsky–Perelomov Hamiltonian $\mathcal{H}$. We begin with the following simple result from linear algebra.

Proposition 17: Let $\mathbf{V}$ *a finite-dimensional vector space over* $\mathbb{C}$, *and* $\mathbf{V}_1 \subset \mathbf{V}$ *a codimension 1 subspace. Let T be an endomorphism of* $\mathbf{V}$ *such that* $\mathbf{V}_1$ *is an invariant subspace, and let* $\kappa \in \mathbb{C}$ *denote the unique eigenvalue of the corresponding endomorphism of* $\mathbf{V}/\mathbf{V}_1$. *If* κ *is not an eigenvalue of* $T|_{\mathbf{V}_1}$, *then* κ *is a multiplicity 1 eigenvalue of T.*

It should be noted that the assumption $k_c \geqslant 0$ in Theorem 1 is crucial. The necessity of this assumption is explained by the following proposition. Indeed, one should remark that there exist certain negative values of k_c for which the action of $\mathcal{H}$ fails to be diagonalizable.

Proposition 18: Let $\mu < \lambda$ *be dominant weights. If* $k_c \geqslant 0$ *for each root length c, then* $\|\lambda + \rho\| > \|\mu + \rho\|$.

Proof: Note that

$$\|\lambda + \rho\|^2 - \|\mu + \rho\|^2 = \|\lambda\|^2 - \|\mu\|^2 + 2(\lambda - \mu, \rho).$$

Using the fact that $\lambda - \mu \in P^+$, one can easily show that $\|\lambda\| > \|\mu\|$. Furthermore, since $\lambda - \mu$ is a linear combination of basic roots with positive coefficients, Proposition 5 implies that $(\lambda - \mu, \rho) > 0$. □

Finally, we have the following.

Proof of Theorem 1: Let λ be a dominant weight. By Proposition 12, $\operatorname{trig}(P_\lambda)^W$ is an invariant subspace of $\tilde{\mathcal{H}}$. Using an argument similar to the one given in the proof of Proposition 12, it is not hard to verify that if $\phi \in \operatorname{trig}(P_\lambda)^W$, then

$$(\tilde{\mathcal{H}} - \|\lambda\|^2 - 2(\rho, \lambda))(\phi) \in \operatorname{trig}(P_{\lambda -})^W. \tag{8}$$

Note that $\operatorname{trig}(P_{\lambda -})^W$ is a codimension 1 subspace of $\operatorname{trig}(P_\lambda)^W$. Furthermore, by Proposition 18,

$$\|\lambda\|^2 + 2(\lambda, \rho) > \|\mu\|^2 + 2(\mu, \rho),$$

for all dominant weights $\mu < \lambda$. Hence, by Proposition 17, there exists a unique $\phi_\lambda \in \operatorname{trig}(P_\lambda)^W$ such that $\tilde{\mathcal{H}}\phi_\lambda = (\|\lambda\|^2 + 2(\rho, \lambda))\phi$. The first of the desired conclusions now follows by Proposition 16.

To prove the converse let $F\phi$ with $\phi \in \operatorname{trig}(P)^W$ be an eigenfunction of $\mathcal{H}$ with eigenvalue κ. Let $\lambda \in P^+$ be a maximal element of $\operatorname{supp}(\phi)$. Since $\operatorname{trig}(P_{\lambda -})^W$ is a codimension 1 subspace of $\operatorname{trig}(P_\lambda)^W$, (8) implies that $\kappa = \|\lambda\|^2 + 2(\lambda, \rho)$. Consequently, λ is the unique maximal element of $\operatorname{supp}(\phi)$. By Proposition 17, κ has multiplicity 1, and this gives the desired conclusion. □

IV. A RECURSION FORMULA FOR THE EIGENFUNCTIONS OF $\tilde{\mathcal{H}}$

In the present section we show how to explicitly compute the eigenfunctions of the Olshanetsky–Perelomov Hamiltonian by using a k_c-parametrized analog of the Freudenthal multiplicity formula. The generalized formula actually yields the eigenfunctions ϕ_λ of the related operator $\tilde{\mathcal{H}}$. One should mention that the eigenfunctions ϕ_λ first appeared in the investigations of Heckman and Opdam,[14] who regard these functions as multivariable generalizations of the Jacobi polynomials. The eigenfunctions of $\mathcal{H}$ are, of course, obtained by multiplication with the gauge factor F.

By way of motivation it will be useful to recall the context of the original Freudenthal formula. Suppose that R is reduced and let χ_λ, $\lambda \in P^+$ denote a character of the corresponding compact, simply connected Lie group. The Weyl character formula states that

$$\chi_\lambda = \frac{\Sigma_{w\in W}\,\mathrm{sgn}(w)e^{iw(\lambda+\tilde{\rho})}}{\Sigma_{w\in W}\,\mathrm{sgn}(w)e^{iw(\lambda)}}, \tag{9}$$

where $\tilde{\rho}$ is the half-sum of the positive roots. Now if $k_c=1$ for all c, then the potential term of $\mathcal{H}$ is zero, and the gauge factor F is nothing but the W-antisymmetric denominator of (9). Furthermore, the numerator in (9) is the unique W-antisymmetric eigenfunction of Δ with highest-order term $e^{i(\lambda+\tilde{\rho})}$. Hence, by the intertwining relation described in Proposition 16, the Weyl character formula is equivalent to the statement that χ_λ is an eigenfunction of $\tilde{\mathcal{H}}$ with eigenvalue $(\lambda,\lambda+2\tilde{\rho})$. This observation leads directly to the classical Freudenthal formula for the multiplicities of χ_λ, and to the following generalization involving the parameters k_c. (See Ref. 18 for more details regarding the Weyl and Freudenthal formulas.)

Proposition 19: Let $\phi_\lambda = e^{i\lambda} + \Sigma_{\mu<\lambda} n_\mu e^{i\mu}$ *be the eigenfunction of* $\tilde{\mathcal{H}}$ *described in the statement and proof of Theorem 1. Setting* $n_\lambda = 1$ *and* $n_\nu = 0$ *for* $\nu \not\leq \lambda$, *the remaining coefficients* n_μ, $\mu<\lambda$, *are given by the following recursion formula*:

$$(\|\lambda+\rho\|^2 - \|\mu+\rho\|^2)n_\mu = 2\sum_{\alpha\in R^+}\sum_{j\geq 1} k_{|\alpha|}(\alpha,\mu+j\alpha)n_{\mu+j\alpha}. \tag{10}$$

Proof: Rewriting

$$A_c = e^{i\rho_c}\prod_{\alpha\in R_c^+}(1-e^{-i\alpha}),$$

one obtains

$$\tilde{\mathcal{H}} = -\Delta - i\,\nabla\rho - 2i\sum_{\alpha\in R^+} k_{|\alpha|}\frac{e^{-i\alpha}}{1-e^{-i\alpha}}\nabla\alpha.$$

Let $\mathrm{trig}((P))$ denote the vector space of formal power series $\Sigma_{\mu\in P} c_\mu e^{i\mu}$. Since elements of $\mathrm{trig}(P)$ are finitely supported sums, one has a well-defined multiplication operation $\mathrm{trig}((P))\times\mathrm{trig}(P)\rightarrow\mathrm{trig}((P))$. Thus, setting the domain of $\tilde{\mathcal{H}}$ to be $\mathrm{trig}(P)$, one can extend the operator's coefficient ring and write

$$\tilde{\mathcal{H}} = -\Delta - i\,\nabla\rho - 2i\sum_{\alpha\in R^+}\sum_{j\geq 1} k_{|\alpha|}e^{-ji\alpha}\,\nabla\alpha.$$

However, because of Proposition 11 one can take the codomain of $\tilde{\mathcal{H}}$ to be $\mathrm{trig}(P)$ rather than all of $\mathrm{trig}((P))$. Acting with the right-hand side of the latter equation on ϕ_λ, collecting like terms, and using the fact that ϕ_λ is an eigenfunction with eigenvalue $(\lambda,\lambda+2\rho)$ immediately yields (10). □

It is important to remark that by Proposition 18 the coefficient of n_μ appearing in (10) is never zero. Consequently, (10) can indeed be used as a recursive formula for the coefficients n_μ. One should also remark that the W symmetry of ϕ_λ means that it suffices to use formula (10) to calculate n_μ with $\mu\in P^+$.

[1] F. Calogero, ''Solution of the one-dimensional *N*-body problems with quadratic and/or inversely quadratic pair potentials,'' J. Math. Phys. **12**, 419 (1971).
[2] B. Sutherland, ''Exact results for a quantum many-body problem in one dimension II,'' Phys. Rev. A **5**, 1372 (1972).
[3] M. Olshanetsky and A. Perelomov, ''Quantum integrable systems related to Lie algebras,'' Phys. Rep. **94**, 313 (1983).
[4] A. Perelomov, E. Ragoucy, and P. Zaugg, ''Explicit solution of the quantum three-body Calogero–Sutherland model,'' J. Phys. A **31**, L559 (1998).
[5] W. Rühl and A. Turbiner, ''Exact solvability of the Calogero and Sutherland models,'' Mod. Phys. Lett. A **10**, 2213 (1995).
[6] D. Serban, ''Some properties of the Calogero-Sutherland model with reflections,'' J. Phys. A **30**, 4215 (1997).
[7] K. Sogo, ''Eigenstates of Calogero–Sutherland–Moser model and generalized Schur functions,'' J. Math. Phys. **35**, 2282 (1994).
[8] M. Shifman, ''New findings in quantum mechanics (partial algebraization of the spectrum),'' Int. J. Mod. Phys. A **126**, 2897 (1989).
[9] A. Turbiner, ''Quasi-exactly solvable problems and sl_2 algebra,'' Commun. Math. Phys. **118**, 467 (1989).
[10] A. Gonzalez-Lopez, N. Kamran, and P. J. Olver, ''New quasi-exactly solvable Hamiltonians in two dimensions,'' Commun. Math. Phys. **159**, 503 (1994).
[11] A. Gonzalez-Lopez, N. Kamran, and P. J. Olver, ''Real Lie algebras of differential operators and quasi-exactly solvable potentials,'' Philos. Trans. R. Soc. London, Ser. A **354**, 1165 (1996).
[12] I. G. Macdonald, *Symmetric Functions and Hall Polynomials* (Oxford University Press, Oxford, 1995).
[13] I. G. Macdonald, ''Commuting differential operators and zonal spherical functions,'' in *Springer Lecture Notes in Mathematics*, 1987, Vol. 1271, pp. 189.
[14] G. J. Heckman and E. M. Opdam, ''Root systems and hypergeometric functions I & II,'' Compositio Mathematica **64**, 329 (1987).
[15] L. Brink, A. Turbiner, and N. Wyllard, ''Hidden algebras of the (super) Calogero and Sutherland models,'' J. Math. Phys. **39**, 1285 (1998).
[16] J. Humphreys, *Introduction to Lie Algebras and Representation theory* (Springer-Verlag, Berlin, 1972).
[17] N. Bourbaki, *Groupes et Algèbres de Lie* (Hermann, Paris, 1968), Chaps. 4–9.
[18] W. Fulton and J. Harris, *Representation Theory* (Springer-Verlag, Berlin, 1996).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

A loop group approach to the C. Neumann problem and Moser–Veselov factorization

Saša Krešić-Jurić
Symbol Technologies, Inc., Holtsville, New York 11742-1300

(Received 9 December 1998; accepted for publication 28 April 1999)

A geometrical description of continuous and discrete versions of the Neumann oscillator in terms of a loop group framework is investigated. It is shown that the continuous Neumann oscillator can be integrated by the Riemann–Hilbert factorization on the ''twisted'' loop group of $O(3)$, $LO(3)$. The solution of the problem is given in terms of a special class of flows on the quotient space $LO(3)/LO(3)_+$ where $LO(3)_+$ is a subgroup of positive loops in $LO(3)$. It is also shown that the Moser–Veselov algorithm for integrating a discrete version of the Neumann oscillator (Heisenberg *XYZ* chain) is induced by a discrete flow in this space. The flow can be explicitly integrated by solving the matrix Riccati equation. In both cases, discrete and continuous, conservation laws are derived from time invariance of a relation that holds for coefficients of the Fourier expansion of the flows. © *1999 American Institute of Physics.* [S0022-2488(99)03110-2]

I. INTRODUCTION

Recently there has been a growth of interest in discretization algorithms of dynamical systems which preserve certain properties of the system. For example, one may be interested in discretizations which preserve the energy or momentum, or perhaps the symplectic structure of the system. A class of such algorithms called variational integrators can be obtained by means of discretizing the Hamilton's principle which leads to discrete Euler–Lagrange equations.[1]

In this article we consider discretizations of integrable systems which have a zero-curvature representation on the Lie algebra of a loop group G. Roughly speaking, if G has a unique factorization into subgroups $G=G_-G_+$, then one can usually associate to the system a flow $g(t)G_+$ on the homogeneous space G/G_+. If the flow can be mapped into the phase space flow φ_t, then the system can be integrated by calculating $g(t)$ which is equivalent to solving a Riemann–Hilbert factorization problem on G. Now, suppose that ϕ_n is a given discretization of the continuous flow φ_t. Then one can ask if it is possible to find a discrete flow in G/G_+ which can be mapped into ϕ_n. In other words, one would like to know if the discretized system can be integrated by stepping along a discrete flow in G/G_+. This is in general a very difficult question and has to be dealt with separately for each system under consideration.

In this paper we show that it is possible to carry out the above construction for the Neumann oscillator

$$\frac{d^2S}{dt^2}=-JS+\lambda S, \quad S\in\mathbb{R}^n, \quad |S|=1, \tag{1}$$

where $J=\mathrm{diag}(J_1,...,J_n)$, $\lambda=\langle JS,S\rangle-|S_t|^2$, and its discrete version

$$J(S_{k-1}+S_{k+1})=\lambda_k S_k, \quad S_k\in\mathbb{R}^n, \quad |S_k|=1. \tag{2}$$

Here $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$ and the subscript t denotes the derivative with respect to t. In classical mechanics Eq. (1) describes the motion of a particle with the Hamiltonian $H=\frac{1}{2}|P|^2+\frac{1}{2}\langle JS,S\rangle$ constrained to the cotangent bundle of the $(n-1)$-sphere. In solid state physics

(1) is equivalent to the stationary Landau–Lifshitz equation[2] (t being a spatial coordinate), and the difference equation (2) represents the Heisenberg chain with classical spins.[3]

For $n=3$, Eq. (1) was originally studied by C. Neumann[4] in 1859 who integrated it by separation of variables in the Hamilton–Jacobi equation. The interest in the problem was revived when Moser[5] established the connection between (1) and certain rank-2 perturbations of a symmetric matrix. He showed that the integrals of the Neumann oscillator can be obtained in terms of the eigenvalues of the perturbed matrix. Moser's results were generalized by Adams, Harnad, and Previato[6] to general rank-r perturbations. In a recent paper[3] the discrete system (2) was investigated by Moser and Veselov through factorization of matrix polynomials $M(z)=J^2z^2+(x\otimes Jy-Jy\otimes x)z-x\otimes x,\ x,y\in\mathbb{R}^n$. They showed that Eq. (2) can be integrated by isospectral deformations of M based on a factorization of M into two first-order polynomials. We also mention the work of Deift, Li, and Tomei[7] in which the Moser–Veselov ideas were interpreted in terms of a loop group framework. In their paper a continuous flow is constructed which at integer times interpolates the Moser–Veselov algorithm.

The outline of the present paper is as follows: Sec. II begins with a brief review of the zero-curvature formulation of partial differential equations (PDE's) which is needed in our approach to the Neumann oscillator. We then introduce the twisted loop group of $O(3)$, $LO(3)$, and show that an action of $\mathbb{R}^2$ on a subset of $LO(3)$ induces Eq. (1) for $n=3$. Solutions of this equation are given in terms of a special class of flows on the quotient space $LO(3)/LO(3)_+$, where $LO(3)_+$ is the subgroup of positive loops in $LO(3)$. We remark that in this picture the Neumann oscillator appears as a special case of an infinite dimensional systems which is generated by a generic flow in $LO(3)/LO(3)_+$. The conservation laws for Eq. (1) can be deduced from algebraic relations which hold for coefficients of the Fourier expansion of the flows.

In Sec. III we present a loop group approach to the discrete system (2). We show that after a simple transformation the Moser–Veselov polynomials M become elements of the Lie algebra of $LO(3)$. This enables us to find a discrete flow $g_-^{(k)}\mapsto g_-^{(k+1)}$ in $LO(3)$ which induces the aforementioned isospectral deformations of M. The flow can be explicitly integrated by solving the matrix Riccati equation for the first coefficient of the Fourier expansion of $g_-^{(k+1)}$. The integrals of the discrete system which were discovered by Granovskii and Zhedanov[8] can be found by the same method as in Sec. II. The results of Secs. II and III generalize in a straightforward manner to arbitrary dimensions by considering the action of $\mathbb{R}^2$ on $LO(n)$.

We mentioned in the Introduction that the Moser–Veselov algorithm was interpreted in terms of a loop group framework by Deift, Li, and Tomei. We should emphasize, however, that our approach has the advantage of describing both systems (1) and (2) from a single point of view. In each case the dynamics is given by a special class of flows (Neumann flows) on the quotient space $LO(3)/LO(3)_+$, and the same method can be used to obtain the integrals of motion.

II. LOOP GROUPS AND THE NEUMANN OSCILLATOR

A. Zero-curvature representation

In this section we review the zero curvature representation on loop algebras as a method for constructing integrable systems (see Refs. 2, 9).

Let G be a Banach Lie group which contains closed Lie subgroups G_- and G_+ such that $G_-\cap G_+=\{I\}$. The set $G_-G_+=\{gh|g\in G_-,h\in G_+\}$ is open in G if and only if the Lie algebra of G splits into a direct sum of vector spaces $\mathfrak{g}=\mathfrak{g}_-\oplus\mathfrak{g}_+$. Here $\mathfrak{g}_\pm$ is the Lie algebra of $G_\pm$. Suppose that G_-G_+ is open in G and consider a differentiable action of $\mathbb{R}^n$ on G defined by

$$\mathbf{t}g=\exp\left(\sum_{i=1}^{n} t_iX_i\right)g. \tag{3}$$

Here $\mathbf{t}=(t_1,\ldots,t_n)\in\mathbb{R}^n$ and $X_1,\ldots,X_n$ are pairwise commuting elements of $\mathfrak{g}_+$. Note that (3) descends to an action of $\mathbb{R}^n$ on G/G_+ by

$$\mathbf{t}(gG_+)=(\mathbf{t}g)G_+. \tag{4}$$

Now let $g\in G_-G_+$. Then there exists a neighborhood $U\subset\mathbb{R}^n$ of zero such that $\mathbf{t}g\in G_-G_+$ for all $\mathbf{t}\in U$. Hence $\mathbf{t}g$ can be written in a unique way as

$$\mathbf{t}g=g_-(\mathbf{t})g_+(\mathbf{t}), \tag{5}$$

where $g_-\in G_-$ and $g_+\in G_+$. We shall refer to (5) as the ''Riemann–Hilbert'' factorization of $\mathbf{t}g$. In view of (4) we see that the action of $\mathbb{R}^n$ induces the flow $g_-(\mathbf{t})G_+$ on the homogeneous space G/G_+.

Next we show that $g_-(\mathbf{t})$ is a ''solution'' to a hierarchy of PDE's associated with the Riemann–Hilbert factorization problem. By differentiating the Riemann–Hilbert splitting (5) and using (3) we obtain

$$g_-^{-1}X_ig_-=g_-^{-1}\frac{\partial g_-}{\partial t_i}+\frac{\partial g_+}{\partial t_i}g_+^{-1}.$$

Since the first term on the right-hand side is in $\mathfrak{g}_-$ and the second is in $\mathfrak{g}_+$, we have

$$\Pi_+(g_-^{-1}X_ig_-)=\frac{\partial g_+}{\partial t_i}g_+^{-1}, \tag{6}$$

where Π_+ is the projection of $\mathfrak{g}$ onto $\mathfrak{g}_+$. Denote

$$M_i(\mathbf{t})=\Pi_+(g_-^{-1}(\mathbf{t})X_ig_-(\mathbf{t})) \tag{7}$$

for $i=1,2,...,n$. Since the flows induced by any two vectors X_k, X_l commute, the system of equations

$$\frac{\partial g_+}{\partial t_k}(\mathbf{t})=M_k(\mathbf{t})g_+(\mathbf{t}),\quad \frac{\partial g_+}{\partial t_l}(\mathbf{t})=M_l(\mathbf{t})g_+(\mathbf{t}) \tag{8}$$

satisfies the compatibility condition $(g_+)_{kl}=(g_+)_{lk}$ which yields the zero-curvature equation

$$\frac{\partial M_k}{\partial t_l}-\frac{\partial M_l}{\partial t_k}+[M_k,M_l]=0. \tag{9}$$

Usually, G is the loop group of a linear group and the zero-curvature condition turns out to be equivalent with a nonlinear PDE satisfied by a matrix element $u(\mathbf{t})$ of M_k and M_l. Since u can be explicitly calculated in terms of $g_-(\mathbf{t})$ via (7), we have the mappings $g\mapsto g_-(\mathbf{t})\mapsto u(\mathbf{t})$. Here $u(\mathbf{t})$ is regarded as a function of t_k and t_l while the other variables are kept fixed. The Riemann–Hilbert factorization establishes local existence of $u(\mathbf{t})$, thus producing a solution of (9). Moreover, it linearizes the equation for u since multiplying g by $\exp(\Delta t_iX_i)$ corresponds to $u(\mathbf{t})$ flowing in the ''t_i-direction'' by the amount Δt_i. By choosing different values of k and l the flow $g_-(\mathbf{t})$ yields solutions to an entire hierarchy of equations represented by (9). Since the map $g\mapsto g_-(\mathbf{t})$ is invariant under the right multiplication of g by an element of G_+, we may assume that $g\in G_-$. Then by uniqueness of the splitting $g=g_-(0)$, so g contains the initial values of u.

For applications of the above method it is most convenient to work with subgroups of the group of continuous loops in $GL(n,\mathbb{C})$. More precisely, G is constructed as follows. Let $\mathcal{A}$ denote the Banach algebra,

$$\mathcal{A}=\left\{f:S^1\to\mathbb{C}\,\middle|\,f(z)=\sum_{n\in\mathbb{Z}}a_nz^n,\quad \sum_{n\in\mathbb{Z}}|a_n|<\infty,a_n\in\mathbb{C}\right\}, \tag{10}$$

with respect to the norm $\|f\|_1=\Sigma_{n\in Z}|a_n|$ and pointwise addition and multiplication. Alternatively, $\mathcal{A}$ consists of continuous functions on S^1 which have an absolutely convergent Fourier series. Let $M(n,\mathcal{A})$ be the Banach algebra of $n\times n$ matrices $g=(g_{ij})$ relative to the norm $\|g\|=\Sigma_{i,j=1}^n\|g_{ij}\|_1$. Various integrable systems can be constructed by introducing an involution on $M(n,\mathcal{A})$ such that $\|g^*\|=\|g\|$, and considering subgroups

$$G=\{g\in GL(n,\mathcal{A})|gg^*=I\} \tag{11}$$

of the Banach Lie group

$$GL(n,\mathcal{A})=\{g(z)\in M(n,\mathcal{A})|\det(g(z))\neq 0 \quad \text{for all} \quad z\in S^1\}.$$

A standard result from the theory of Banach manifolds[10] asserts that G is a regular Lie subgroup of $GL(n,\mathcal{A})$ with the Lie algebra

$$\mathfrak{g}=\{X\in M(n,\mathcal{A})|X+X^*=0\}. \tag{12}$$

Various examples of loop groups and integrable systems obtained by this method can be found in Refs. 11, 12, and 13.

B. Integration of the Neumann problem by the Riemann–Hilbert factorization

As we mentioned in the Introduction, the Neumann problem

$$\frac{d^2S}{dt^2}=-JS+\lambda S, \quad S\in\mathbb{R}^3, \quad |S|=1, \tag{13}$$

where $J=\operatorname{diag}(J_1,J_2,J_3)$ and $\lambda=\langle JS,S\rangle-|S_t|^2$, is a completely integrable Hamiltonian system which is obtained by constraining the Hamiltonian $H=\frac{1}{2}|P|^2+\frac{1}{2}\langle JS,S\rangle$ to the cotangent bundle of the 2-sphere. We assume that the spring constants J_i are nondegenerate, i.e., they satisfy $J_1<J_2<J_3$.

In order to motivate the connection between the Neumann oscillator and loop groups we state the following result attributed to Uhlenbeck.[14]

Lemma 1: Define matrices $Q=(S_iS_j)$, $L=(P_iS_j-S_iP_j)$, *and set* $U(z)=Jz^2+Lz-Q$, $V(z)=Jz+L$ *with a parameter* $z\in\mathbb{C}$. *Then the Neumann system*

$$\frac{dS_i}{dt}=P_i, \quad \frac{dP_i}{dt}=-J_iS_i+[\langle JS,S\rangle-|P|^2]S_i, \quad |S|=1, \tag{14}$$

is equivalent to the Lax equation

$$\frac{dU}{dt}=[V,U] \quad \textit{subject to constraints} \quad |S|=1,\ \langle P,S\rangle=0. \tag{15}$$

The proof is a direct computation. According to the discussion in the previous section we would like $U(z)$ and $V(z)$ to be elements of the Lie algebra of a loop group, and the Lax Eq. (15) to be a special case of the zero-curvature condition on this algebra. It turns out that this is in principle true. Since the details are somewhat nontrivial and do not seem to appear in the literature we present them here. This is instrumental in understanding the second part of the paper which deals with the discretized Neumann oscillator.

Let $\mathcal{A}_{\mathbb{R}}$ be the function algebra introduced in (10) with real-valued Fourier coefficients. Define the group $G=LO(3)$ as in (11) with $\mathcal{A}=\mathcal{A}_{\mathbb{R}}$ and the involution on $M(3,\mathcal{A}_{\mathbb{R}})$ given by

$$g^*(z)=g^T(-z).$$

We call G the ''twisted'' loop group of $O(3)$. Note that if $X(z)=\Sigma_{n\in Z}C_n z^n$ is an element of the Lie algebra of G, then $C_n^T=C_n$ for odd n, and $C_n^T=-C_n$ for even n. In order to define the Riemann–Hilbert factorization consider the closed subgroups of G,

$$G_+=\left\{g\in LO(3)\,\middle|\,g(z)=\sum_{n\geqslant 0}A_n z^n\right\},$$

$$G_-=\left\{g\in LO(3)\,\middle|\,g(z)=I+\sum_{n<0}A_n z^n\right\}.$$

Clearly, $G_-\cap G_+=\{I\}$. If $\mathfrak{g}_\pm$ denotes the Lie algebra of $G_\pm$, then $\mathfrak{g}=\mathfrak{g}_-\oplus\mathfrak{g}_+$ so the set G_-G_+ is open in G. Now let $X_1(z)=Jz$, $X_2(z)=Jz^3$ and consider an action of $\mathbb{R}^2$ on G defined by

$$(t,x)g=\exp(tX_1(z)+xX_2(z))g. \tag{16}$$

The action (16) gives rise to a unique factorization of $(t,x)g$,

$$\exp(tX_1(z)+xX_2(z))g=g_-(t,x)g_+(t,x)\quad\text{for}\quad g\in G_-G_+. \tag{17}$$

The matrices $M_i=\Pi_+(g_-^{-1}X_i g_-)$, $i=1,2$, satisfy the zero-curvature equation

$$\frac{\partial M_2}{\partial t}-\frac{\partial M_1}{\partial x}+[M_2,M_1]=0, \tag{18}$$

where $\Pi_+:\mathfrak{g}\to\mathfrak{g}_+$ is the canonical projection with kernel $\mathfrak{g}_-$. The system of differential equations represented by Eq. (18) is in general very complicated and it is not known whether it describes a physical system. However, we will show that for special initial values of the flow $(t,x)g$ this equation is equivalent with Neumann problem (13).

Since we may assume $g\in G_-$, throughout this section the triple (g,g_-,g_+) will be *always* given by (17) with $g=g_-(0,0)$. Thus $M_1(0,0)=\Pi_+(g^{-1}Jzg)$ and $M_2(0,0)=\Pi_+(g^{-1}Jz^3g)$.

Let $I+\Sigma_{n=1}^{\infty}A_n z^{-n}$ be the Fourier expansion of g. The group law $gg^*=I$ implies $g^{-1}=\Sigma_{n=1}^{\infty}(-1)^n A_n^T z^{-n}$, where the matrices A_n satisfy $A_1^T=A_1$ and

$$A_n+(-1)^n A_n^T=-\sum_{k=1}^{n-1}(-1)^k A_{n-k}A_k^T\quad n\geqslant 2. \tag{19}$$

From (19) we find by a straightforward computation that

$$M_1(0,0)=Jz+L_0,\quad M_2(0,0)=Jz^3+L_0z^2-K_0z+N_0, \tag{20}$$

where $L_0=[J,A_1]$ and

$$K_0=A_1[J,A_1]-[J,A_2], \tag{21}$$

$$N_0=[J,A_3]-A_1[J,A_2]+A_2^T[J,A_1]. \tag{22}$$

We can prove now the following preliminary result.

Lemma 2: There exist elements $g=I+\Sigma_{n=1}^{\infty}A_n z^{-n}\in G_-$ *such that*

$$M_2(0,0)=\Pi_+(g^{-1}Jz^3g)=Jz^3+L_0z^2-K_0z,$$

where $K_0+\frac{1}{3}I$ *is the tensor product* $S_0\otimes S_0$ *for some unit vector* S_0 in $\mathbb{R}^3$.

Proof: First we show that one can find A_1 and A_2 so that $K_0+\frac{1}{3}I=S_0\otimes S_0$ for some $|S_0|=1$. Note that for $n=2$ Eq. (19) implies $A_2=\frac{1}{2}A_1^2+\frac{1}{2}T$, where $T=\frac{1}{2}(A_2-A_2^T)$. By substituting this into (21) we obtain

$$K_0 = \tfrac{1}{2}[A_1,[J,A_1]] - \tfrac{1}{2}[J,T]. \tag{23}$$

Denote

$$A_1 = \begin{pmatrix} a_{11} & a_3 & -a_2 \\ a_3 & a_{22} & a_1 \\ -a_2 & a_1 & a_{33} \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 0 & t_3 & t_2 \\ -t_3 & 0 & t_1 \\ -t_2 & -t_1 & 0 \end{pmatrix}. \tag{24}$$

Then the diagonal of K_0 is given by

$$(K_0)_{11} = (J_3 - J_1)a_2^2 + (J_2 - J_1)a_3^2,$$

$$(K_0)_{22} = (J_3 - J_2)a_1^2 + (J_1 - J_2)a_3^2,$$

$$(K_0)_{33} = (J_1 - J_3)a_2^2 + (J_2 - J_3)a_1^2, \tag{25}$$

and for the off-diagonal elements we have

$$(K_0)_{ij} = \tfrac{1}{2}(-1)^{i+j+1}[(J_i - J_j)(a_{ii} - a_{jj})a_k + (J_i + J_j - 2J_k)a_i a_j - (J_i - J_j)t_k]. \tag{26}$$

Here (i,j,k) is a cyclic permutation of (1,2,3). Now pick $S_0 = (S_1, S_2, S_3)^T \in S^2$ such that $S_1^2 - \frac{1}{3} \geqslant 0$ and $S_3^2 - \frac{1}{3} \leqslant 0$. Recalling that $J_1 < J_2 < J_3$ it is clear from (25) and (26) that one can find A_1 and T such that $K_0 = S_0 \otimes S_0 - \frac{1}{3}I$.

By similar reasoning one can show that there exists A_3 satisfying (19) which makes the free term in $M_2(0,0)$ vanish,

$$[J,A_3] = A_1[J,A_2] - A_2^T[J,A_1]. \tag{27}$$

Substituting $A_2 = \frac{1}{2}A_1^2 + \frac{1}{2}T$ into (19) with $n = 3$ yields $A_3 = \frac{1}{2}U + \frac{1}{4}(TA_1 + A_1T)$, where $U = \frac{1}{2}(A_3 + A_3^T)$. Then (27) is equivalent to

$$[J,U] = \tfrac{1}{2}[T,[J,A_1]] + \tfrac{1}{2}[A_1,[J,T]] + A_1[J,A_1]A_1. \tag{28}$$

Clearly, this equation determines the off-diagonal elements of U, while the diagonal can be chosen arbitrary. ■

For given $J = \mathrm{diag}(J_1,J_2,J_3)$ let G_J^i denote the set of all $g \in G_-$ from Lemma 2 (i stands for the special initial conditions imposed on M_2). Since the Neumann problem is given by an ordinary differential equation we will be interested in loops $g_-(t) \in G_-$ which depend only on t. The following lemma is easily proved and provides a simple characterization of such loops.

Lemma 3: Let $g \in G_-$. Then $\exp(tX_1 + xX_2)g = g_-(t)g_+(t,x)$ if and only if $g^{-1}X_2 g \in \mathfrak{g}_+$.

Let G_J^t be the set of all $g \in G_-$ such that $g^{-1}X_2 g \in \mathfrak{g}_+$. This notation is to remind us that if $g \in G_J^t$, then g_- flows only in the "t-direction." Now define the set of "Neumann loops" by $G_J^N = G_J^i \cap G_J^t \subset G_-$. One can show that the action of $\mathbb{R}^2$ on G/G_+ can be restricted to G_J^N/G_+ in the sense that if $gG_+ \in G_J^N/G_+$, then $g_-(t)G_+ \in G_J^N/G_+$ for all t sufficiently close to zero. Roughly speaking, as a consequence of this result the matrices $M_1(t)$ and $M_2(t)$ evolve along the orbits of the Neumann oscillator. This is the content of the following theorem.

Theorem 4: *Let $g \in G_J^N$, and let $M_i(t) = \Pi_+(g_-^{-1}(t)X_i g_-(t))$, $i = 1,2$. Then the matrices M_1 and M_2 satisfy the Lax equation $dM_2/dt = [M_1, M_2]$ which is equivalent with the Neumann problem (14).*

Proof: Let M_i^0 denote the initial value of $M_i(t)$, i.e., $M_i^0 = \Pi_+(g^{-1}X_i g)$. Then $M_1^0 = Jz + L_0$ with $L_0^T = -L_0$, and by Lemma 2,

$$M_2^0 = Jz^3 + L_0 z^2 - K_0 z, \quad \text{where} \quad K_0 = S_0 \otimes S_0 - \tfrac{1}{3}I, \quad |S_0| = 1. \tag{29}$$

We show that $M_2(t)$ retains the form (29) for all t for which the Riemann–Hilbert splitting is defined. Recall from (8) that $M_2(t)=(\partial g_+/\partial x)g_+^{-1}$. On the other hand, by differentiating (17) and rearranging we have $g_+(g^{-1}X_2g)g_+^{-1}=(\partial g_+/\partial x)g_+^{-1}$ because $\partial g_-/\partial x=0$. Hence $M_2(t)=g_+M_2^0g_+^{-1}$, so in view of (29) the matrix $M_2(t)$ has vanishing zero-order term. Therefore we can write

$$M_1(t)=Jz+L(t), \quad M_2(t)=Jz^3+L(t)z^2-K(t)z,$$

where L and K satisfy the initial conditions $L(0)=L_0$ and $K(0)=K_0$. Recall that M_1 and M_2 satisfy the zero-curvature Eq. (18) which reduces to

$$\frac{dM_2}{dt}=[M_1,M_2]. \tag{30}$$

We claim that $K(t)=S(t)\otimes S(t)-\frac{1}{3}I$ for some unit vector $S(t)$. It follows from 30 that $dL/dt=[K,J]$ and $dK/dt=[L,K]$. Define $Q(t)=K(t)+\frac{1}{3}I$. Then the time evolution of Q is also given by $dQ/dt=[L,Q]$ with the initial condition $Q(0)=S_0\otimes S_0$. Hence the spectrum of Q is independent of t and is given by the roots of

$$\det(Q(0)-\lambda I)=|S_0|^2\lambda^2-\lambda^3=\lambda^2-\lambda^3. \tag{31}$$

Since Q is real and symmetric this implies that $Q(t)=\omega(t)\mathrm{diag}(1,0,0)\omega^{-1}(t)$ for some orthogonal matrix $\omega(t)$. Thus we conclude $Q(t)=S(t)\otimes S(t)$, where $S(t)\in\mathbb{R}^3$. By invariance of the spectrum we have $\det(Q(t)-\lambda I)=\det(Q(0)-\lambda I)$ which implies $|S(t)|=1$ for all t.

To finish the proof let L' be the image of $L\in\mathfrak{so}(3)$ under the Lie algebra isomorphism $\mathfrak{so}(3)\approx\mathbb{R}^3$, and define $P=L'\times S\in\mathbb{R}^3$. Since L can be chosen so that L' is perpendicular to S we have $L'=P\times S$, thus $L=P\otimes S-S\otimes P$. Now set

$$V(t)=Jz+L(t) \ \text{ and } \ U(t)=Jz^2+L(t)z-Q(t).$$

Since $dM_2/dt=[M_1,M_2]$ if and only if $dU/dt=[V,U]$ the theorem follows from Lemma 1. ■

Theorem 4 shows that the Neumann problem is explicitly integrable in the sense that for $g\in G_J^N$ the differential Eq. (13) can be solved by the factorization of $\exp(tX_1+xX_2)g$. It is well known that this problem is also integrable in the sense of Liouville. Moser showed that the integrals of motion can be obtained from eigenvalues of certain rank-2 perturbations of a symmetric matrix.[5] Here we give a new proof based on the fact that the integrals are simply the diagonal elements of the matrix $S(t)\otimes S(t)-K(t)$ which, because of the special initial values of the flow $g_-(t)G_+$, are preserved along this flow.

Corollary 5: If the initial values of the Neumann problem (13) are determined by $g\in G_J^N$, then the system has the following integrals of motion:

$$F_i=S_i^2+\sum_{\substack{k=1\\k\neq i}}^{3}\frac{(P_iS_k-S_iP_k)^2}{J_i-J_k}, \quad i=1,2,3. \tag{32}$$

Proof: We have shown in Theorem 4 that $K(t)=S(t)\otimes S(t)-\frac{1}{3}I$, when $g\in G_J^N$. Hence $S\otimes S-K$ is constant along the flow. We claim that the diagonal elements $F_i=(S\otimes S-K)_{ii}$ have the form (32). To see this recall that $L(t)=[J,A_1(t)]$, where we label the elements of $A_1(t)$ as in (24). On the other hand, $L=P\otimes S-S\otimes P$, hence $[J,A_1]=P\otimes S-S\otimes P$. By solving for the off-diagonal elements of A_1 we obtain

$$a_k=(-1)^{i+j+1}\frac{P_iS_j-S_iP_j}{J_i-J_j},$$

where (i,j,k) is a cyclic permutation of (1,2,3). Substituting this into Eqs. (25) we find that $(S\otimes S)_{ii}-K_{ii}$ is given by (32). ■

We conclude our discussion with the remark that Eqs. (25) imply $S_1^2(t)\geqslant\frac{1}{3}$ and $S_3^2(t)\leqslant\frac{1}{3}$ because $J_1<J_2<J_3$. These conditions restrict the motion of the Neumann oscillator to two disjoint regions of the sphere. This corresponds to one of the four types of motion described in the original paper of Neumann (see p. 52 in Ref. 4).

III. A LOOP GROUP DESCRIPTION OF THE DISCRETIZED NEUMANN OSCILLATOR

This section is devoted to the study of a discrete version of the Neumann problem

$$J(S_{n-1}+S_{n+1})=\lambda_n S_n, \quad |S_n|=1. \tag{33}$$

S_n is the position of a point on the 2-sphere at integer time n, and the multiplier λ_n is determined from the condition $|S_{n+1}|=1$. In contrast to the continuous case λ_n is not unique,

$$\lambda_n=0 \quad \text{or} \quad \lambda_n=2\langle S_{n-1},J^{-1}S_n\rangle/|J^{-1}S_n|^2.$$

We shall assume that for all integer times we have $\lambda_n\neq 0$ since only in this case Eq. (33) corresponds to the correct limit describing continuous motion.[3]

We have shown in Sec. II that the Neumann oscillator can be integrated by constructing a continuous flow in G/G_+ where we recall $G=LO(3)$ is the twisted loop group of $O(3)$. It is natural to ask whether this property is conserved under the above discretization. More precisely, we would like to know if there is a discrete flow $g_-^{(n)}G_+\mapsto g_-^{(n+1)}G_+$ in G/G_+ which integrates Eq. (33). In this paper by integrability of (33) we mean that S_{n+1} can be calculated from S_n, i.e., that the order of the difference equation can be reduced by one. For another interpretation of integrability based on a discrete version of Liouville's theorem, see Ref. 15. Before we answer the above question we will need the following background information.

A. Moser–Veselov factorization of matrix polynomials

Moser and Veselov[3] investigated integrability of Eq. (33) with the help of an isospectral technique based on factorization of matrix polynomials $U(z)=C_0+C_1z+C_2z^2$, $z\in\mathbb{C}$. The idea is similar to the QR-algorithm of Francis designed to calculate the eigenvalues of a complex matrix. The matrix $U(z)$ is factored into a product of two first order polynomials

$$U(z)=(A_0+A_1z)(B_0+B_1z), \tag{34}$$

and the isospectral mapping Ψ is defined by exchanging the factors,

$$\Psi:U(z)\mapsto\hat{U}(z)=(B_0+B_1z)(A_0+A_1z).$$

They have shown that this technique integrates several well known classical problems, e.g., the rigid body and the billiard inside an ellipsoid. In the case of the Neumann oscillator it induces a map $(S_{n-1},S_n)\mapsto(S_n,S_{n+1})$ such that the triple (S_{n-1},S_n,S_{n+1}) satisfies the difference Eq. (33).

Our exposition of the Moser–Veselov algorithm for the Neumann oscillator is similar to the one for the rigid body problem in Ref. 3. This makes it more suitable for interpretation in terms of a discrete flow in G/G_+.

Let $\mathcal{P}$ denote the class of polynomials $U_n(z)=J^2z^2+L_nz-S_nS_n^T$, where $L_n^T=-L_n$ and $S_n=(S_n^1,S_n^2,S_n^3)^T$ is a unit vector in $\mathbb{R}^3$. Define $\Psi:\mathcal{P}\to\mathcal{P}$ as follows. Suppose that $U_n(z)$ can be factored as

$$U_n(z)=(zJ+\omega_n^T)(zJ-\omega_n). \tag{35}$$

This is possible if and only if ω_n satisfies the equations

$$\omega_n^T J - J\omega_n = L_n \quad \text{and} \quad \omega_n^T \omega_n = S_n S_n^T. \tag{36}$$

Now define $U_{n+1} = \Psi(U_n)$ by exchanging the factors in (35), i.e.,

$$U_{n+1}(z) = (zJ - \omega_n)(zJ + \omega_n^T). \tag{37}$$

Assuming for the moment that such ω_n exists we show that Ψ is well-defined, i.e., $\mathrm{Im}(\Psi) \subset \mathcal{P}$. Since

$$U_{n+1}(z) = J^2 z^2 + (J\omega_n^T - \omega_n J)z - \omega_n \omega_n^T,$$

we have $L_{n+1} = J\omega_n^T - \omega_n J$, thus L_{n+1} is antisymmetric. It remains to show that $\omega_n \omega_n^T = S_{n+1} S_{n+1}^T$ for some unit vector S_{n+1} in $\mathbb{R}^3$. This follows from the isospectrality of Ψ. Namely,

$$U_{n+1}(z) = (zJ + \omega_n^T)^{-1} U_n(z)(zJ + \omega_n^T)$$

so when $z = 0$ we have $\omega_n \omega_n^T = (\omega_n^T)^{-1}(S_n S_n^T)\omega_n^T$. Denote $T_n = \omega_n \omega_n^T$. Then T_n is a real symmetric matrix with the characteristic polynomial $\det(T_n - \lambda I) = |S_n|^2 \lambda^2 - \lambda^3 = \lambda^2 - \lambda^3$. By the argument following Eq. (31) in Sec. II, there is a unit vector S_{n+1} in $\mathbb{R}^3$ such that $T_n = S_{n+1} S_{n+1}^T$. Therefore $\mathrm{Im}(\Psi) \subset \mathcal{P}$.

Moser and Veselov now show the following. Let S_n and S_{n+1} be unit vectors in $\mathbb{R}^3$, and let $U_n(z)$ be defined by (35) with $\omega_n = S_{n+1} S_n^T$. Note that the set Σ of zeros of the polynomial $p(z) = \det(U_n(z))$ splits into $\Sigma = \Sigma_- \cup \Sigma_+$, where Σ_+, Σ_- satisfy the conditions $\overline{\Sigma_\pm} = \Sigma_\pm$ and $\Sigma_+ = -\Sigma_-$. If such a splitting is fixed and all zeros of $p(z)$ are distinct, then one can find $\omega_{n+1} = S_{n+2} S_{n+1}^T$ with $|S_{n+2}| = 1$ such that the image of U_n under Ψ is given by

$$U_{n+1}(z) = (zJ + \omega_{n+1}^T)(zJ - \omega_{n+1}). \tag{38}$$

A comparison of (37) and (38) shows that

$$J\omega_n^T - \omega_n J = \omega_{n+1}^T J - J\omega_{n+1}.$$

From here it follows at once that the vectors S_n, S_{n+1}, and S_{n+2} satisfy the difference Eq. (33). Now this procedure can be iterated starting with U_{n+1}, hence the isospectral mapping Ψ induces a discrete map $\omega_n \mapsto \omega_{n+1}$ which integrates the Neumann oscillator. This map is multivalued because it depends on a particular splitting of $\Sigma = \Sigma_- \cup \Sigma_+$.

In the next section we shall prove the existence of a discrete flow in G/G_+ which via conjugation induces the map Ψ. The multivaluedness of Ψ will correspond to the fact that the flow in G/G_+ is also multivalued because it is determined by symmetric solutions of a matrix Riccati equation. It is well known that these solutions depend on a particular factorization of the characteristic polynomial of the coefficient matrix of the equation.[16]

B. A loop group approach to the Moser–Veselov Algorithm

Here we give a group theoretic interpretation of the results described in the previous section. The motivation for this comes from the observation that the isospectral mapping $\Psi: U_n \mapsto U_{n+1}$ can be viewed as a discrete version of the Lax pair equation. Recall from Lemma 1 that the Neumann oscillator has the Lax representation

$$\frac{d}{dt}(Jz^2 + L(t)z - Q(t)) = [Jz + L(t), Jz^2 + L(t)z - Q(t)], \tag{39}$$

where $Q = (S_i S_j)$ and $L = (P_i S_j - S_i P_j)$, which is an isospectral deformation of the matrix $U(z) = Jz^2 + Lz - Q$. If we think of

$$U_n(z) = J^2 z^2 + L_n z - S_n S_n^T \tag{40}$$

as being $U(z)$ at time n (modulo the difference in the powers of J), then the mapping $\Psi: U_n \mapsto U_{n+1}$ can be viewed as a discrete analog of Eq. (39).

It was shown in Sec. II that this equation is induced by a special class of flows $g_-(t)G_+$ in G/G_+ whose initial values belong to the set of Neumann loops G_J^N. In analogy with this it is reasonable to expect that the mapping Ψ is induced by a class of discrete flows

$$\phi: g_-^{(n)} \mapsto g_-^{(n+1)} \tag{41}$$

with initial values $g_-^{(0)} \in G_{J^2}^N$. Here $G_{J^2}^N$ is the set of Neumann loops with J replaced by J^2. In other words, we would like to say that there is a mapping $g_-^{(n)} \mapsto U_n$ such that ϕ gives rise to the isospectral transformation $U_n \mapsto U_{n+1}$.

In what follows we show that such a correspondence exists and we develop an explicit algorithm for computing ϕ. Let $g_-^{(n)} \in G_{J^2}^N$, and define

$$M_2^{(n)}(z) = (g_-^{(n)}(z))^{-1} J^2 z^3 g_-^{(n)}(z). \tag{42}$$

Clearly, $M_2^{(n)}$ is an element of $\mathfrak{g}_+$, and by Lemma 2 we have

$$M_2^{(n)}(z) = J^2 z^3 + L_n z^2 - \left(S_n S_n^T - \tfrac{1}{3} I \right) z, \tag{43}$$

where $L_n^T = -L_n$ and $|S_n| = 1$. If $I + \Sigma_{i=1}^{\infty} A_i z^{-i}$ is the Fourier expansion of $g_-^{(n)}$, then following Eqs. (20) and (21) the coefficients of $M_2^{(n)}$ are given by

$$L_n = [J^2, A_1], \quad S_n S_n^T - \tfrac{1}{3} I = A_1 [J^2, A_1] - [J^2, A_2]. \tag{44}$$

Note that Eqs. (42) and (43) imply that

$$U_n(z) = (g_-^{(n)}(z))^{-1} \left(J^2 z^2 - \tfrac{1}{3} I \right) g_-^{(n)}(z). \tag{45}$$

This suggests that the flow $\phi: G_{J^2}^N \to G_{J^2}^N$ should be constructed so that it induces Ψ via the above conjugation. By writing $J^2 z^2 - \frac{1}{3} I = J^2 z^2 (I - \frac{1}{3} J^{-2} z^{-2})$ and using the Taylor's expansion for $(1-x)^{1/2}$ we obtain

$$J^2 z^2 - \frac{1}{3} I = \left[Jz \left(I - \sum_{k=1}^{\infty} \frac{(2k-3)!!}{3^k (2k)!!} J^{-2k} z^{-2k} \right) \right]^2. \tag{46}$$

The series can always be made convergent in $M(3, \mathcal{A}_{\mathbb{R}})$ because we can multiply J by a sufficiently large constant without effecting Eq. (33). Let C_k denote the coefficients in the above series. For $g_-^{(n)} \in G_{J^2}^N$ consider the following factorization:

$$Jz \left(I - \sum_{k=1}^{\infty} C_k J^{-2k} z^{-2k} \right) g_-^{(n)}(z) = h(z)(Jz - \Omega_n) \tag{47}$$

for some $h(z) = I + \Sigma_{k=1}^{\infty} B_k z^{-k}$ and Ω_n subject to conditions

$$B_1^T = B_1 \quad \text{and} \quad \Omega_n^T \Omega_n = S_n S_n^T.$$

Some algebraic manipulation shows that given B_1 the Fourier coefficients of h can be found recursively from

$$B_2 J = J A_2 + B_1 \Omega_n - C_1 J^{-1},$$

$$B_kJ=JA_k+B_{k-1}\Omega_n-\sum_{2i+j=k}C_iJ^{-2i+1}A_j-\frac{1}{2}(1+(-1)^k)C_{k/2}J^{-k+1},\quad k>2.$$

Furthermore, it follows from (47) that $\Omega_n=B_1J-JA_1$. Hence the factorization (47) can be carried out provided we can find B_1. In order to do this note that $\Omega_n^T\Omega_n=S_nS_n^T$ if and only if B_1 is a symmetric solution of the matrix Riccati equation

$$B_1^2+P^TB_1+B_1P+Q=0, \tag{48}$$

where $P=-JA_1J^{-1}$ and $Q=J^{-1}(A_1J^2A_1-S_nS_n^T)J^{-1}$. Equation (48) is indeed a Riccati equation for B_1 because Q is symmetric and (P,I) is controllable. It is well known that this equation has symmetric solutions provided the coefficient matrix

$$M=\begin{pmatrix} P & I \\ -Q & -P^T \end{pmatrix}$$

has no purely imaginary eigenvalues. In this case it can be shown[16] that $\det(tI-M)=(-1)^3p(t)p(-t)$, where the real polynomial $p(t)$ is monic and relatively prime to $p(-t)$. Then the Riccati equation has a unique symmetric solution B_1 such that $\det(tI-(P+B_1))=p(t)$, and a unique symmetric solution B_1' such that $\det(tI-(P+B_1'))=(-1)^3p(-t)$.

Suppose that B_1 has been found by solving Eq. (48) and calculate $\Omega_n=B_1J-JA_1$. Then we have $\Omega_n^T\Omega_n=S_nS_n^T$ and, since A_1 and B_1 are symmetric, it follows that $\Omega_n^TJ-J\Omega_n=[J^2,A_1]=L_n$. Thus the matrix polynomial (40) can be factored as

$$U_n(z)=(zJ+\Omega_n^T)(zJ-\Omega_n), \tag{49}$$

which is precisely the Moser–Veselov factorization. Moreover, a straightforward computation using (46), (47), and (49) shows that h is an element of G,

$$\begin{aligned}(h(z))^*h(z)&=(Jz+\Omega_n^T)^{-1}(g_-^{(n)}(z))^{-1}(J^2z^2-\tfrac{1}{3}I)g_-^{(n)}(z)(Jz-\Omega_n)^{-1}\\&=(Jz+\Omega_n^T)^{-1}U_n(z)(Jz-\Omega_n)^{-1}=I.\end{aligned}$$

Let $g_-^{(n+1)}=h$ and define

$$U_{n+1}(z)=(g_-^{(n+1)}(z))^{-1}(J^2z^2-\tfrac{1}{3})g_-^{(n+1)}(z).$$

As desired, one can show that U_{n+1} belongs to $\mathcal{P}$ and is given by switching the terms in Eq. (49). Indeed, the group law in G and (47) imply that

$$\begin{aligned}U_{n+1}(z)&=(Jz+\Omega_n^T)^{-1}(g_-^{(n)}(z))^{-1}(J^2z^2-\tfrac{1}{3})^2g_-^{(n)}(z)(Jz-\Omega_n)^{-1}\\&=(Jz+\Omega_n^T)^{-1}U_n^2(z)(Jz-\Omega_n)^{-1}=(Jz-\Omega_n)(Jz+\Omega_n^T).\end{aligned}$$

Recall from the previous section that $\Omega_n\Omega_n^T=S_{n+1}S_{n+1}^T$ for some unit vector S_{n+1}. Thus $U_{n+1}\in\mathcal{P}$, so in fact $g_-^{(n+1)}$ belongs to $G_{J^2}^N$. Now define $\phi:G_{J^2}^N\rightarrow G_{J^2}^N$ by $\phi(g_-^{(n)})=g_-^{(n+1)}$. Note that ϕ depends on the factorization of $\det(tI-M)$ which determines a particular solution of the Riccati equation for B_1. For a fixed branch of ϕ we can summarize the above results in the following theorem.

Theorem 6: *The flow $\phi:g_-^{(n)}\mapsto g_-^{(n+1)}$ induces the Moser–Veselov map $\Psi:U_n\mapsto U_{n+1}$ and gives the commutative diagram*

$$\begin{array}{ccc} \mathcal{P} & \stackrel{\Psi}{\rightarrow} & \mathcal{P} \\ C\uparrow & & \uparrow C \\ G^N_{J^2} & \stackrel{\phi}{\rightarrow} & G^N_{J^2} \end{array}$$

where $C(g)=g^{-1}(J^2z^2-\frac{1}{3})g$.

In complete analogy with the continuous model, the discretized Neumann oscillator is integrable in the sense that there are constants of the motion independent of n. As in the continuous case this is a consequence of the fact that the diagonal elements of $S_nS_n^T-K_n$ are preserved under the flow. A proof identical to that of Corollary 5 with the exception that J is replaced by J^2 shows that the diagonal elements of $S_nS_n^T-K_n$ are given by

$$\tilde{F}_i=(S_n^i)^2+\sum_{\substack{k=1\\k\neq i}}^{3}\frac{(J_iS_{n+1}^iS_n^k-J_kS_{n+1}^kS_n^i)^2}{J_i^2-J_k^2},\quad i=1,2,3.$$

It turns out that these are the integrals found by Granovskii and Zhedanov,[8] and later generalized by Veselov[17] to the n-dimensional case. The algebraic structure of $\tilde{F}_i$ clearly resembles that of F_i, the integrals of the continuous Neumann problem. Hence $\tilde{F}_i$ is a natural discrete analog of F_i.

ACKNOWLEDGMENT

The author wishes to thank his adviser Professor M. Adams for invaluable help during the preparation of this work.

[1] J. M. Wendlandt and J. E. Marsden, ''Mechanical integrators derived from a discrete variational principle,'' Physica D **106**, 223–246 (1997).
[2] L. D. Fadeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer-Verlag, New York, 1987).
[3] J. Moser and A. P. Veselov, ''Discrete versions of some classical integrable systems and factorization of matrix polynomials,'' Commun. Math. Phys. **139**, 217–243 (1991).
[4] C. Neumann, ''De problemate quodam mechanico, quod ad primam integralium ultraellipticorum classem revocatur,'' J. Reine Angew. Math. **56**, 46–63 (1859).
[5] J. Moser, ''Geometry of quadratics and spectral theory,'' in *The Chern Symposium*, Berkeley, June 1979 (Springer-Verlag, New York, 1980).
[6] M. R. Adams, J. Harnad, and E. Previato, ''Isospectral Hamiltonian flows in finite and infinite dimensions,'' Commun. Math. Phys. **117**, 451–500 (1988).
[7] P. Deift, L. C. Li, and C. Tomei, ''Matrix factorizations and integrable systems,'' Commun. Pure Appl. Math. **42**, 443–521 (1989).
[8] Y. I. Granovskii and A. S. Zhedanov, ''Solutions of domain type in anisotropic magnetic chains,'' Teor. Mat. Fiz. **71**, 438–445 (1987).
[9] J. Szmigielski, Ph.D. thesis, University of Georgia, 1987.
[10] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications* (Springer-Verlag, New York, 1988).
[11] J. Dorfmeister, ''Banach manifolds of solutions to nonlinear partial differential equations, and relations with finite-dimensional manifolds,'' Proc. Symp. Pure Math. **54**, 121–139 (1993).
[12] H. Flaschka, A. C. Newell, and T. Ratiu, ''Kac–Moody Lie algebras and soliton equations,'' Physica D **9**, 300–323 (1983).
[13] S. Krešić-Jurić, Ph.D. thesis, University of Georgia, 1995.
[14] T. Ratiu, ''The C. Neumann problem as a completely integrable system on an adjoint orbit,'' Trans. Am. Math. Soc. **246**, 321–329 (1981).
[15] A. P. Veselov, ''Integrable discrete-time systems and difference operators,'' Funct. Anal. Appl. **22**, 83–93 (1988).
[16] W. A. Coppel, ''Matrix quadratic equations,'' Bull. Aust. Math. Soc. **10**, 377–401 (1974).
[17] A. P. Veselov, ''Integration of the stationary problem for a classical spin chain,'' Teor. Mat. Fiz. **71**, 446–450 (1987).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Superintegrable systems on the two-dimensional sphere S^2 and the hyperbolic plane H^2

Manuel F. Rañada
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, 50009 Zaragoza, Spain

Mariano Santander
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Valladolid, 47011 Valladolid, Spain

(Received 1 March 1999; accepted for publication 5 April 1999)

The existence of superintegrable systems with $n=2$ degrees of freedom possessing three independent globally defined constants of motion which are quadratic in the velocities is studied on the two-dimensional sphere S^2 and on the hyperbolic plane H^2. The approach used is based on enforcing the conditions for the existence of two independent integrals (further than the energy). This is done in a way which allows us to discuss at once the cases of the sphere S^2 and the hyperbolical plane H^2, by considering the curvature κ as a parameter. Different superintegrable potentials are obtained as the solutions of certain systems of two κ-dependent second order partial differential equations. The Euclidean results are directly recovered for $\kappa=0$, and the superintegrable potentials on either the standard unit sphere (radius $R=1$) or the unit Lobachewski plane (''radius'' $R=1$) appear as the particular values of the κ-dependent superintegrable potentials for the values $\kappa=1$ and $\kappa=-1$. Some new superintegrable potentials are found, both on S^2 and H^2. The correspondence between superintegrable systems in spaces of zero and nonzero curvature is discussed.
[S0022-2488(99)02110-6]

I. INTRODUCTION

A superintegrable system is a system that is integrable (in the sense of Liouville–Arnold) and that, in addition to this, possesses more constants of motion than degrees of freedom. If the number N of independent constants takes the value $N=2n-1$ (n the number of degrees of freedom) then the system is called maximally superintegrable (see Refs. 1–23). There are three well known examples of this very particular class of systems, namely, the Kepler problem, the isotropic harmonic oscillator, and the nonisotropic oscillator with commensurable frequencies. The $n=3$ Kepler possesses not only the energy and the angular momentum as constants of motion, but also the Runge–Lenz vector; five of these integrals are functionally independent. This property is also true for n arbitrary. Concerning the harmonic oscillator, it is a system trivially integrable since it can be considered as a kind of ''direct sum'' of systems with one degree of freedom. If the oscillator is isotropic then it has the angular momentum (and/or the Fradkin tensor[24]) as an additional integral of motion. If the oscillator is nonisotropic the angular momentum is not preserved as the potential is not central; nevertheless when the quotients of the frequencies are rational the system has other additional integrals. In these three cases it is well known that all the orbits became closed for the case of bounded motions. This high degree of regularity (existence of periodic motions) is a consequence of the superintegrable character.

Most of known integrable systems are Hamilton–Jacobi separable, that is, systems with an associate Hamilton–Jacobi equation that can be solved by separation of variables after an appropriate coordinate system has been found. An important point is that separable systems have constants of motion which are linear or quadratic in the velocities. Fris *et al.*[2] studied in 1965 the two-dimensional Hamiltonians with standard Euclidean kinetic term for which the Hamilton–Jacobi equations separates in more than one coordinate system in E^2 and obtained four different families, each with a potential which is a ''linear superposition'' of three simpler potentials (they were mainly interested in the quantum viewpoint but the results obtained are also valid at the classical level). This Hamilton–Jacobi approach was also used in Ref. 3 and, later on, by other authors as Evans,[9] Grosche *et al.*[14–16] and Kalnins *et al.*[17,18] Other approaches recently studied have been the dimensional reduction of simpler systems,[12,13] or the direct obtaining of the conditions for the existence of two quadratic independent integrals[20] (further than the energy).

The dynamics for the cases where the configuration space Q is either the sphere S^2 or the hyperbolic (Lobachewski) plane H^2 is not so well known as it is in the Euclidean case E^2, and most of the studies done on the sphere have focused the attention on spherical central potentials.[25–27] There are some noncentral but rather simple problems as, e.g., the nonisotropic oscillator, that still remain as very partially understood in manifolds of nonzero curvature. In this article we will study superintegrable Lagrangian systems with quadratic constants of motion for the case in which the configuration space Q is either the two-dimensional sphere S^2 or the hyperbolic plane H^2. We will use as an approach the same strategy that proved to be successful in Ref. 20 for two different planar manifolds, $Q=\mathbb{R}^2$ with a Euclidean metric, this is E^2, and $Q=\mathbb{R}^2$ with a Minkowskian metric.

The article is organized as follows: In Sec. II we present some geometric properties of Riemannian 2D manifolds of constant curvature, taking as leading idea the introduction of a formalism with the curvature κ as a parameter. This is made so that the Euclidean results are directly recovered for $\kappa=0$ (in a natural way without the problems of a limit process) and the standard sphere of radius $R=1$ or the Lobachewski plane of ''radius'' $R=1$ will correspond to the particular values $\kappa=1$ or $\kappa=-1$. Indeed, if κ is looked as a parameter, the approach can be done simultaneously for the three cases; most computations might even be done simultaneously, at the small price of using κ dependent trigonometric functions. In Sec. III, we recall the properties of the case $Q=E^2$ from the present viewpoint.

In Sec. IV we look for superintegrable systems on both $Q=S^2$ and H^2 (with curvature κ) and we find several families which have the superintegrable Euclidean potentials (analyzed in Sec. III) as flat limits. In the first subsection the study is done in polar coordinates, and in the second subsection in parallel coordinates. Some of the superintegrable potentials we obtain are already known,[14–16] but other are, as far as we know, new. We find several noncentral superintegrable spherical potentials, and in particular the spherical and hyperbolical superintegrable versions of the anisotropic 2:1 oscillator and several new superintegrable potentials in the hyperbolic plane. Finally, Sec. V provides a discussion and an outlook to the results obtained; it can be read independently and gives information enough for a reader who is not interested in the details of the derivation.

II. GEOMETRY AND DYNAMICS ON THE SPHERE S^2 AND THE HYPERBOLIC PLANE H^2

On any general two-dimensional Riemannian space (not necessarily of constant curvature) there are two distinguished types of local coordinate systems, ''geodesic parallel'' and ''geodesic polar'' coordinates. They reduce to the familiar cartesian and polar coordinates on the Euclidean plane (see, e.g., Klinberger[28]).

Both these systems are based on an origin point O and an oriented geodesic l_1 through O (Fig. 1). For any point P in some suitable neighborhood of O, there is a unique geodesic l joining O and P. The (geodesic) polar coordinates (r,ϕ) of P, relative to the origin O and the positive geodesic ray of l_1, are the (positive) distance r between O and P measured along l, and the angle ϕ between l and the positive ray l_1, measured around O. These coordinates are defined in a neighborhood of O not extending beyond the cut locus of O; polar coordinates are singular at O, and ϕ is discontinuous on the positive ray of l_1.

Now, for any point P (in some suitable strip neighborhood of l_1) consider the geodesic l_2' through P and orthogonal to l_1 and let P_1 be the intersection point of l_2' and l_1 nearest to P. The

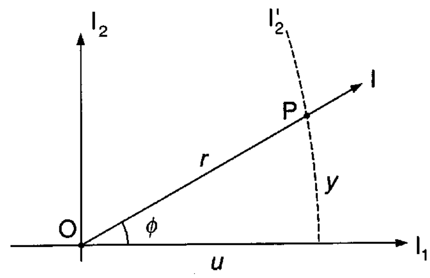

FIG. 1. Polar (r,ϕ) and parallel (u,y) coordinates based on the oriented geodesic l_1 and reference point O. All these coordinates are lengths or angles measured in the intrinsic metric of the space of constant curvature. The figure follows the pattern of a stereographic projection of the sphere from the South pole, with O at the North pole, but the geometrical meaning of these coordinates holds for any value of the curvature.

(geodesic) parallel coordinates (u,y) of P, relative to the origin O and base geodesic l_1, are defined as the distance u between O and P_1, measured along l_1, and the distance y, between P_1 and P, measured along l_2'. Again these coordinates will be regular and without singularities in some strip centered in l_1. If instead of l_1 another line is taken as base, we obtain another system of geodesic parallel coordinates. Figure 2 also displays the particular case with base l_2, orthogonal to l_1 through O; these second set of parallel coordinates will be denoted (v,x) and will also play some role in our discussion.

These systems are suitable for most general purposes, because the coordinates (r,ϕ), (u,y), and (v,x) have a *direct* geometric significance, as distances and angles measured in the intrinsic metric of the surface. Closed expressions are usually only possible for spaces of *constant curvature*. In the constant *positive* curvature case, i.e., the sphere, the geodesics are great circles, and relations among distances and angles are the subject of spherical geometry. Polar coordinates on the sphere are singular at the origin (pole) O and also at its antipodal point (the cut locus of O). Parallel coordinates are singular in the two poles of the base geodesic. While in the Euclidean plane a line orthogonal to both l_1 and l_2 do not exist (nor does it exist in the hyperbolic plane), there is such a line l_3 for the sphere (the polar of the point O), so we have here a third set of parallel coordinates. These three sets are based on three geodesics mutually orthogonal by pairs and the third system with base l_3 is essentially equivalent to the polar coordinates whose center is the pole of l_3.

The notation has been chosen to emphasize the similarities with the Euclidean case. For a point P, r is the distance measured in either S^2 or H^2 (with curvature κ) from P to the origin point O, and ϕ determines the orientation of the line OP through O. On the other side, x, y are the geodesic distances from P to the two ''coordinate axes'' l_1,l_2; there are other two quantities, u, v which are distances, measured along l_1,l_2, between O and the orthogonal projections of P on l_1,l_2. In the Euclidean case, we have the identities $x=u$, $y=v$, but once we deal with nonzero

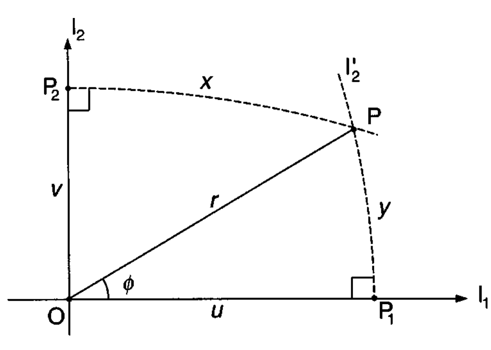

FIG. 2. The three coordinate systems (r,ϕ), (u,y), and (v,x) of a point P. Relationships among these coordinates are discussed in the text for any curvature value κ.

curvature these equalities are no longer true; recall that y is the distance from P to the ''x'' coordinate axe, but u *is not* the distance from P to the ''y'' coordinate axe. Both polar (r,ϕ) and the two systems of parallel coordinates (u,y) and (v,x) are always *orthogonal*; however the coordinate system (x,y) made up of the distances to the two coordinate axes is only orthogonal in the Euclidean plane, but *not* in S^2 nor in H^2.

For a sphere of radius R (curvature $\kappa = 1/R^2$), the ''geographic'' coordinates (θ,ϕ) (where θ is the latitude and ϕ the longitude) are closely related to both polar and parallel type coordinate systems; $(R(\pi/2-\theta),\phi)$ are *polar* coordinates with its origin in the North pole, while $(R\phi,R\theta)$ are *parallel* coordinates with the equator as the baseline. This equivalence does not exist in the Euclidean and hyperbolic case, where polar and parallel coordinates are different, so there are reasons to keep their consideration separate, even for the sphere, in the context we are working. The structure of the superintegrable systems on planes of constant curvature and in their limiting Euclidean case will be more clearly seen this way.

The metric of the sphere of curvature $\kappa = 1/R^2$ is given in parallel and polar coordinates by

$$ds^2 = \cos^2(y/R)\,du^2 + dy^2, \quad ds^2 = dr^2 + R^2 \sin^2(r/R)\,d\phi^2$$

reducing to $du^2 + dy^2$ and $dr^2 + r^2 d\phi^2$ when $\kappa \to 0$ (or, equivalently $R \to \infty$). It is possible to write these expressions in a form which holds simultaneously for the sphere, the Euclidean plane and the hyperbolic plane, by introducing the ''tagged'' trigonometric functions $\mathrm{C}_\kappa(x)$, $\mathrm{S}_\kappa(x)$, and $\mathrm{T}_\kappa(x)$ defined by

$$\mathrm{C}_\kappa(x) = \begin{cases} \cos\sqrt{\kappa}x & \text{if } \kappa > 0, \\ 1 & \text{if } \kappa = 0, \\ \cosh\sqrt{-\kappa}x & if \ \kappa < 0, \end{cases} \qquad \mathrm{S}_\kappa(x) = \begin{cases} \dfrac{1}{\sqrt{\kappa}} \sin\sqrt{\kappa}x & \text{if } \kappa > 0, \\ x & \text{if } \kappa = 0, \\ \dfrac{1}{\sqrt{-\kappa}} \sinh\sqrt{-\kappa}x & \text{if } \kappa < 0, \end{cases}$$

and

$$\mathrm{T}_\kappa(x) = \frac{\mathrm{S}_\kappa(x)}{\mathrm{C}_\kappa(x)}.$$

When the constant κ is nonzero, it can be reduced to either 1 or -1 by length rescaling (this is tantamount to choose units of length so that $R=1$). When $\kappa=1$ the three ''tagged'' functions are the ordinary trigonometrical functions sine, cosine and tangent, i.e., $\mathrm{S}_1(x) = \sin x$, $\mathrm{C}_1(x) = \cos x$, $\mathrm{T}_1(x) = \tan x$. For $\kappa = 0$ one gets the ''parabolic'' sine $\mathrm{S}_0(x) = x$, cosine $\mathrm{C}_0(x) = 1$, and tangent $\mathrm{T}_0(x) = x$. For $\kappa = -1$, these functions are the hyperbolic cosine, sine, and tangent. Therefore, in the flat case $\kappa = 0$ all $\mathrm{C}_\kappa(x)$ are replaced by 1, while all $\mathrm{S}_\kappa(x)$, $\mathrm{T}_\kappa(x)$ are replaced by its variable x; this suggests that in the curved case, $\mathrm{C}_\kappa(x)$ should be looked at as a kind of ''curved'' deformation of the function 1, while both $\mathrm{S}_\kappa(x)$ and $\mathrm{T}_\kappa(x)$ are two kinds of deformations of the linear function x.

The consistent use of the ''tagged'' functions allow to describe simultaneously the geometry of spaces with a positive-definite metric of constant curvature κ. For instance, the metric of either the sphere, the Euclidean plane or the hyperbolic plane (according as $\kappa > 0, = 0, < 0$) is written in parallel and polar coordinates as

$$ds^2 = \mathrm{C}_\kappa^2(y)\,du^2 + dy^2, \quad ds^2 = dr^2 + \mathrm{S}_\kappa^2(r)\,d\phi^2,$$

and the nonzero Christoffel symbols are, in either coordinate systems,

$$\Gamma^y_{uu} = \kappa \mathrm{S}_\kappa(y)\mathrm{C}_\kappa(y), \quad \Gamma^u_{uy} = -\kappa \mathrm{T}_\kappa(y),$$

$$\Gamma^r_{\phi\phi} = -\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r), \quad \Gamma^\phi_{r\phi} = 1/\mathrm{T}_\kappa(r).$$

The relations between the different coordinates (u,y), (v,x), (r,ϕ), of a point P are obtained from trigonometry. In fact, the points O, P, and P_1 or P_2 determine (see Fig. 2) two rectangular triangles P_1PO (resp. P_2PO). We state these relations in a way that holds regardless of the value of κ; in the cases $\kappa>0$ or $\kappa<0$ they reduce to standard formulas of spherical or hyperbolic trigonometry, while for $\kappa=0$ they are well-known Euclidean relations.

In any rectangular triangle, as P_1PO in the figure, the three sides r,u,y and the angle ϕ at O are related by the following equations:

$$\mathrm{S}_\kappa(y) = \mathrm{S}_\kappa(r)\sin\phi, \quad \mathrm{C}_\kappa(r) = \mathrm{C}_\kappa(u)\mathrm{C}_\kappa(y),$$

$$\mathrm{T}_\kappa(u) = \mathrm{T}_\kappa(r)\cos\phi, \quad \mathrm{T}_\kappa(y) = \mathrm{S}_\kappa(u)\tan\phi.$$

When these equations are applied to the two rectangular triangles P_1PO (sides r,u,y, angle ϕ at O) and P_2PO (sides r,v,x, angle $\pi/2-\phi$ at O), we get many relations with a rather symmetrical appearance in the pairs x,y, and u,v. In particular,

$$\mathrm{S}^2_\kappa(r) = \mathrm{S}^2_\kappa(x) + \mathrm{S}^2_\kappa(y), \quad \mathrm{C}^2_\kappa(r) = \mathrm{C}_\kappa(x)\mathrm{C}_\kappa(y)\mathrm{C}_\kappa(u)\mathrm{C}_\kappa(v), \quad \mathrm{T}^2_\kappa(r) = \mathrm{T}^2_\kappa(u) + \mathrm{T}^2_\kappa(v),$$

$$\mathrm{S}_\kappa(x) = \mathrm{C}_\kappa(y)\mathrm{S}_\kappa(u) = \mathrm{C}_\kappa(r)\mathrm{T}_\kappa(u), \quad \mathrm{C}_\kappa(x) = \frac{\mathrm{C}_\kappa(r)}{\mathrm{C}_\kappa(v)}, \quad \mathrm{T}_\kappa(x) = \mathrm{C}_\kappa(v)\mathrm{T}_\kappa(u),$$

$$\mathrm{S}_\kappa(y) = \mathrm{C}_\kappa(x)\mathrm{S}_\kappa(v) = \mathrm{C}_\kappa(r)\mathrm{T}_\kappa(v), \quad \mathrm{C}_\kappa(y) = \frac{\mathrm{C}_\kappa(r)}{\mathrm{C}_\kappa(u)}, \quad \mathrm{T}_\kappa(y) = \mathrm{C}_\kappa(u)\mathrm{T}_\kappa(v),$$

$$\sin\phi = \frac{\mathrm{T}_\kappa(v)}{\mathrm{T}_\kappa(r)} = \frac{\mathrm{S}_\kappa(y)}{\mathrm{S}_\kappa(r)}, \quad \cos\phi = \frac{\mathrm{T}_\kappa(u)}{\mathrm{T}_\kappa(r)} = \frac{\mathrm{S}_\kappa(x)}{\mathrm{S}_\kappa(r)}, \quad \tan\phi = \frac{\mathrm{T}_\kappa(v)}{\mathrm{T}_\kappa(u)} = \frac{\mathrm{T}_\kappa(y)}{\mathrm{S}_\kappa(u)} = \frac{\mathrm{S}_\kappa(v)}{\mathrm{T}_\kappa(x)},$$

as well as

$$\mathrm{T}_\kappa(x)\mathrm{T}_\kappa(y) = \mathrm{S}_\kappa(u)\mathrm{S}_\kappa(v), \quad \mathrm{S}_\kappa(x)\mathrm{S}_\kappa(y) = \mathrm{C}^2_\kappa(r)\mathrm{T}_\kappa(u)\mathrm{T}_\kappa(v).$$

The change from *polar* to *parallel* coordinates in any constant curvature plane can be read from these equations, and allow several equivalent expressions. Perhaps the way closer to the Euclidean one is

$$\mathrm{T}_\kappa(u) = \mathrm{T}_\kappa(r)\cos\phi, \quad \mathrm{S}^2_\kappa(r) = \mathrm{C}^2_\kappa(y)\mathrm{S}^2_\kappa(u) + \mathrm{S}^2_\kappa(y),$$

$$\mathrm{S}_\kappa(y) = \mathrm{S}_\kappa(r)\sin\phi, \quad \tan\phi = \frac{\mathrm{T}_\kappa(y)}{\mathrm{S}_\kappa(u)}.$$

Another form for the formulas in the first line is

$$\mathrm{S}_\kappa(u) = \frac{\mathrm{S}_\kappa(r)\cos\phi}{\mathrm{C}_\kappa(y)} = \frac{\mathrm{S}_\kappa(r)\cos\phi}{\sqrt{1-\kappa\mathrm{S}^2_\kappa(r)\sin^2\phi}},$$

$$\mathrm{T}^2_\kappa(r) = \mathrm{T}^2_\kappa(u) + \mathrm{T}^2_\kappa(v) = \mathrm{T}^2_\kappa(y) + \frac{\mathrm{T}^2_\kappa(u)}{\mathrm{C}^2_\kappa(y)}.$$

Indeed $\mathrm{T}_\kappa^2(r)$ allows several alternative expressions in terms of y,u,x,v which can be considered as constant curvature versions of Pythagorean theorem,

$$\mathrm{T}_\kappa^2(r)=\mathrm{T}_\kappa^2(u)+\frac{\mathrm{T}_\kappa^2(y)}{\mathrm{C}_\kappa^2(u)}=\frac{\mathrm{T}_\kappa^2(u)}{\mathrm{C}_\kappa^2(y)}+\mathrm{T}_\kappa^2(y)=\mathrm{T}_\kappa^2(v)+\frac{\mathrm{T}_\kappa^2(x)}{\mathrm{C}_\kappa^2(v)}=\frac{\mathrm{T}_\kappa^2(v)}{\mathrm{C}_\kappa^2(x)}+\mathrm{T}_\kappa^2(x).$$

The Euclidean case corresponds to $\kappa=0$; there some of the equations imply $u=x$ and $v=y$, and the nontrivial remaining equations reduce to the Euclidean ones.

III. SUPERINTEGRABLE SYSTEMS ON THE EUCLIDEAN PLANE WITH QUADRATIC CONSTANTS OF MOTION

In this section we give a short résumé of the approach developed in Ref. 20, where superintegrable systems with standard kinetic term on the Euclidean plane and integrals of motion (further than the energy) quadratic in the velocities are determined. This can be considered as the $\kappa=0$ particular case, for the later study in S^2 or H^2 where $\kappa\neq 0$; we shall emphasize those aspects which will turn out relevant for the case of nonzero curvature.

The systems we are studying are described by a Lagrangian of mechanical class living in the Euclidean plane. Let us first use Cartesian coordinates (x,y). The parallel coordinate system (u,y) reduces to the Cartesian one (x,y) in the Euclidean plane because $u=x$, so there is no reason here to depart from the conventional notation. The Lagrangian is

$$L=\tfrac{1}{2}(v_x^2+v_y^2)-V(x,y).$$

Assume there exists a constant of motion $I=I_{22}+I_{20}$, where I_{22} is the quadratic term in the velocities v_x, v_y, and I_{20} is a velocity-independent term. Write $I_{22}=av_x^2+2bv_xv_y+cv_y^2$, with a, b, c, depending on x,y. Then the functions a, b, c, and the term I_{20} cannot be arbitrary, but must satisfy the equations

$$a_x=0,\quad a_y+2b_x=0,\quad c_x+2b_y=0,\quad c_y=0,$$

$$I_{20x}=2aV_x+2bV_y,\quad I_{20y}=2bV_x+2cV_y.$$

The compatibility condition for the last two equations leads to

$$b(V_{yy}-V_{xx})+(a-c)V_{xy}+(a_y-b_x)V_x+(b_y-c_x)V_y=0,$$

so a potential $V(x,y)$ having *any* constant of motion quadratic in the velocities must satisfy this very particular differential equation. The equation is V-independent, so the solution for V is determined up to an additive constant k_0; this is expected for a potential.

The first set of four equations, which are independent of the potential $V(x,y)$, determine the possible forms for a, b, c. By integration we obtain

$$a(x,y)=a_0+a_1y+a_2y^2,$$

$$b(x,y)=(\tfrac{1}{2})(b_0-a_1x-c_1y-2a_2xy),$$

$$c(x,y)=c_0+c_1x+a_2x^2,$$

where $\{a_0,b_0,c_0;a_1,c_1;a_2\}$ are real parameters; the subindex making reference to the order in x or y of each term. Every choice of these parameters determine a partial differential equation and, hence, a family of potentials $V=V[f(x,y),g(x,y)]$ depending on two functions, and such that all these potentials have a constant of motion I_2 quadratic in the velocities. For the values $\{a_0=e_0,0,\ c_0=e_0;0,0;0\}$ the equation reduces to an identity, so it is satisfied for an arbitrary potential V, and the integral I_2 reduces to the energy $I=(1/2)(v_x^2+v_y^2)+V(x,y)$. Any other ''nontrivial'' choice of $\{a_0,b_0,c_0;a_1,c_1;a_2\}$ will correspond to constants of motion further than the energy. The fundamental equation for V can be written in the alternative way,

$$b(V_{yy}-V_{xx})+(a-c)V_{xy}-3b_xV_x+3b_yV_y=0.$$

For a fixed set of parameters $\{a_0,b_0,c_0;a_1,c_1;a_2\}$ this is a linear equation for V; conversely, for fixed V, it is a *linear* equation for these six parameters, whose solutions will determine a linear subspace in the ''parameter space'' $\mathbb{R}^6$ with coordinates $\{a_0,b_0,c_0;a_1,c_1;a_2\}$. The dimension m of this submanifold will depend on the potential, and it is at least equal to 1 (generated by the vector {101;00;0}) for *any* potential. A dimension $m=2$ will mean the existence, further than the energy, of one (and only one) additional integral of motion quadratic in the velocities. For the characterization of such potentials see Perelomov;[29] they belong to one of the four types,

$$V=\frac{1}{r_1r_2}(A(r_1+r_2)+B(r_1-r_2)),$$

$$V=A(r)+\frac{1}{r^2}B(\phi),$$

$$V=\frac{1}{r}(A(r+x)+B(r-x)),$$

$$V=A(x)+B(y),$$

where r_1, r_2, are the two distances to two fixed points in the plane, and r,ϕ; x,y, are a suitable choice of polar and Cartesian coordinates. The first case is the generic one, the others being limiting cases where the two points either coincide, one go to infinity or two go to infinity. Which of the four cases applies is determined by the values of the parameters $\{a_0,b_0,c_0;a_1,c_1;a_2\}$; except for the trivial case {101;00;0}, any other parameter vector can be considered as determining the position of *two* points in the (completed) Euclidean plane and in either case the potential depends on *two* arbitrary functions of a *single* variable.

The next case will be $m=3$; here the potential $V(x,y)$ should be solution of a system of *two* partial differential equations, determined by *two* different sets of parameters (further than {101;00;0}), which will be denoted by lower case letters $(a_0,b_0,c_0;a_1,c_1;a_2)$ and the corresponding upper case ones $(A_0,B_0,C_0;A_1,C_1;A_2)$. We shall consider here *seven* cases (labeled V^a, $\tilde{V}^a$, V^b, V^c, $\tilde{V}^c$, V^d, and V^e). Within each case, the potential which solves the system appear as a ''linear combination'' of three particular solutions and a constant term

$$V^r=k_0+k_1V_1^r+k_2V_2^r+k_3V_3^r,\quad V^r=V^a,\tilde{V}^a,V^b,V^c,\tilde{V}^c,V^d,V^e.$$

Table I contains the seven sets of ''basic'' potentials V_1^r, V_2^r, V_3^r, and for each set, the three vectors generating the associated three-dimensional parameter subspace in the space $(a_0,b_0,c_0;a_1,c_1;a_2)$.

Several comments are in order.

(1) The pairs of families denoted *with* and *without* a tilde are clearly equivalent, the transformation relating them being simply the interchange $(x,y)\leftrightarrow(y,x)$ which geometrically is the reflection in the line $x=y$. The three families V^b, V^d, V^e, are invariant under this transformation.

(2) Each family V^r includes one of the three ''fundamental'' superintegrable potentials with quadratic constants of motion (nonisotropic 2:1 oscillator, isotropic oscillator, and Kepler problem). The general potential in each family may be considered as a kind of ''superintegrable deformation'' of these ''basic'' potentials (the values k_2, k_3 representing the ''intensity'' of the deformation). This has been known for some time, and the existence of ''super-

TABLE I. The Euclidean maximally superintegrable potentials.

$\begin{pmatrix}1,0,0;0,0;0\\0,0,1;0,0;0\\0,0,0;0,1;0\end{pmatrix}$	$(a_0,c_0,c_1)\equiv$ (a_0,c_1) or (c_0,c_1)	V^a	$4x^2+y^2$,	x,	$\frac{1}{y^2}$
$\begin{pmatrix}1,0,0;0,0;0\\0,0,1;0,0;0\\0,0,0;1,0;0\end{pmatrix}$	$(a_0,c_0,a_1)\equiv$ (a_0,a_1) or (c_0,a_1)	$\tilde{V}^a$	x^2+4y^2,	y,	$\frac{1}{x^2}$
$\begin{pmatrix}1,0,0;0,0;0\\0,0,1;0,0;0\\0,0,0;0,0;1\end{pmatrix}$	$(a_0,c_0,a_2)\equiv$ (a_0,a_2) or (c_0,a_2)	V^b	x^2+y^2,	$\frac{1}{x^2}$,	$\frac{1}{y^2}$
$\begin{pmatrix}1,0,1;0,0;0\\0,0,0;0,1;0\\0,0,0;0,0;1\end{pmatrix}$	(c_1,a_2)	V^c	$\frac{1}{\sqrt{x^2+y^2}}$,	$\frac{x}{y^2\sqrt{x^2+y^2}}$,	$\frac{1}{y^2}$
$\begin{pmatrix}1,0,1;0,0;0\\0,0,0;1,0;0\\0,0,0;0,0;1\end{pmatrix}$	(a_1,a_2)	$\tilde{V}^c$	$\frac{1}{\sqrt{x^2+y^2}}$,	$\frac{y}{x^2\sqrt{x^2+y^2}}$,	$\frac{1}{x^2}$
$\begin{pmatrix}1,0,1;0,0;0\\0,0,0;1,0;0\\0,0,0;0,1;0\end{pmatrix}$	(a_1,c_1)	V^d	$\frac{1}{\sqrt{x^2+y^2}}$,	$\frac{[\sqrt{(x^2+y^2)}+x]^{1/2}}{\sqrt{x^2+y^2}}$,	$\frac{[\sqrt{(x^2+y^2)}-x]^{1/2}}{\sqrt{x^2+y^2}}$
$\begin{pmatrix}1,0,0;0,0;0\\0,1,0;0,0;0\\0,0,1;0,0;0\end{pmatrix}$	$(a_0,b_0,c_0)\equiv$ (a_0,b_0) or (b_0,c_0)	V^e	x^2+y^2,	x,	y

integrable deformations'' which do not destroy a fragile property like the superintegrability of the Kepler and oscillator potentials, is remarkable in itself. From a purely mathematical point of view, every V_i^r is an element of a basis in a vector space; hence the choice of certain particular potentials as fundamental ones is due to properties related with dynamics.

(3) For any fixed nonconstant potential V, the maximal dimension for the associated parameter subspace appears to be $m=4$ (the trivial constant potential $V_0=1$ has $m=6$). Only four potentials reach the maximal dimension. Next we give these potentials together with the general form of the associated vectors

$$V=x^2+y^2 \quad (a_0,b_0,c_0;0,0;a_2),$$
$$V=y \quad (a_0,b_0,c_0;a_1,0;0),$$
$$V=1/y^2 \quad (a_0,0,c_0;0,c_1;a_2),$$
$$V=1/\sqrt{x^2+y^2} \quad (e_0,0,e_0;a_1,c_1;a_2).$$

(4) The Kepler potential admits two possibilities of deformation preserving superintegrability since it belongs to two different families, V^c and V^d. The same is also true for the isotropic oscillator that belongs to both V^b and V^e (however in the case of V^e the ''deformed'' oscillator is again another oscillator, unlike in V^b). The anisotropic 2:1 oscillator admits only one family of superintegrable deformations. The reason for this lies in the different value for the dimension of the associated subspaces, which is $m=4$ for Kepler and oscillator, but $m=3$ for the 2:1 anisotropic oscillator.

(5) The potential $1/y^2$ (or $1/x^2$) appear also in three different families. This potential is important because it is superintegrable by itself and also because it turns out to be ''linearly compatible'' with all the three well known superintegrable potentials.

(6) The family V^e was not considered in Refs. 2, 14, or 20 (neither in Ref. 9 that studies the $n=3$ systems). It can be considered as rather trivial since the general potential in V^e is just an isotropic oscillator with center at an arbitrary point (with another choice of origin this has

been already considered in V^b). Nevertheless we have decided to include it because its extension to the nonzero curvature case will prove to be interesting (''curvature versions'' of simple potentials in E^2 can become rather complicate functions in S^2 or H^2).

Under a general change of coordinates, the potential V will behave as a scalar field, and the functions a, b, c, are the three components of a covariant symmetric tensor. As a previous step for the study in S^2 or H^2 we will obtain the explicit expressions in planar polar coordinates. The Lagrangian is given by

$$L=\tfrac{1}{2}(v_r^2+r^2v_\phi^2)-V(r,\phi)$$

and the constant of motion I can be written as, $I=I_{22}+I_{20}(r,\phi)$, $I_{22}=av_r^2+2bv_rv_\phi+cv_\phi^2$, where now a, b, c, are functions of r, ϕ (notice that these new three functions a, b, c, are not simply equal to the former ones expressed in terms of (r,ϕ), because of the tensorial character). These four functions a, b, c, and I_{20}, must satisfy the following equations:

$$a_r=0, \quad 2b_r-(4/r)b+a_\phi=0,$$

$$c_r-(4/r)c+2ra+2b_\phi=0, \quad c_\phi+2rb=0,$$

$$I_{20r}=2aV_r+2(b/r^2)V_\phi, \quad I_{20\phi}=2bV_r+2(c/r^2)V_\phi.$$

The first set of equations can be solved for a, b, c, and leads to

$$a(r,\phi)=(\tfrac{1}{2})[(a_0+c_0)+(a_0-c_0)\cos 2\phi+b_0\sin 2\phi],$$

$$b(r,\phi)=(\tfrac{1}{2})r[(c_0-a_0)\sin 2\phi+b_0\cos 2\phi]+(\tfrac{1}{2})r^2(-a_1\cos\phi+c_1\sin\phi),$$

$$c(r,\phi)=(\tfrac{1}{2})r^2[(a_0+c_0)+(c_0-a_0)\cos 2\phi-b_0\sin 2\phi]+r^3(c_1\cos\phi+a_1\sin\phi)+a_2r^4.$$

In Cartesian coordinates, $a(x,y)$, $b(x,y)$, and $c(x,y)$ were nonhomogeneous quadratic polynomials in x, y (with some special relations for the coefficients); in polar coordinates $a(r,\phi)$, $b(r,\phi)$, $c(r,\phi)$, have Fourier series for the ϕ dependence up to the terms in $\cos(2\phi)$, $\sin(2\phi)$ (also with special relations for the coefficients and the radial dependence).

Making use of these expressions we get the most general form of the function I_{22} as

$$I_{22}=a_0I_{22}(a_0)+b_0I_{22}(b_0)+c_0I_{22}(c_0)+a_1I_{22}(a_1)+c_1I_{22}(c_1)+a_2I_{22}(a_2),$$

where

$$I_{22}(a_0)=(v_r\cos\phi-rv_\phi\sin\phi)^2,$$

$$I_{22}(b_0)=(v_r\cos\phi-rv_\phi\sin\phi)(v_r\sin\phi+rv_\phi\cos\phi),$$

$$I_{22}(c_0)=(v_r\sin\phi+rv_\phi\cos\phi)^2,$$

$$I_{22}(a_1)=(v_r\cos\phi-rv_\phi\sin\phi)(r^2v_\phi),$$

$$I_{22}(c_1)=(v_r\sin\phi+rv_\phi\cos\phi)(r^2v_\phi),$$

$$I_{22}(b_2)=(r^2v_\phi)^2$$

(notice that in the above expressions the coefficients in brackets are just labels and not variables). The function I_{20} exists if and only if V satisfies the following equation:

TABLE II. Polar expressions of the basic Euclidean superintegrable potentials.

$(a_0,c_0,c_1)\equiv$ (a_0,c_1) or (c_0,c_1)	V^a	$r^2(1+3\cos^2\phi)$	$r\cos\phi$	$\frac{1}{r^2\sin^2\phi}$
$(a_0,c_0,a_1)\equiv$ (a_0,a_1) or (c_0,a_1)	$\tilde{V}^a$	$r^2(1+3\sin^2\phi)$	$r\sin\phi$	$\frac{1}{r^2\cos^2\phi}$
$(a_0,c_0,a_2)\equiv$ (a_0,a_2) or (c_0,a_2)	V^b	r^2	$\frac{1}{r^2\cos^2\phi}$	$\frac{1}{r^2\sin^2\phi}$
(c_1,a_2)	V^c	$\frac{1}{r}$	$\frac{\cos\phi}{r^2\sin^2\phi}$	$\frac{1}{r^2\sin^2\phi}$
(a_1,a_2)	$\tilde{V}^c$	$\frac{1}{r}$	$\frac{\sin\phi}{r^2\cos^2\phi}$	$\frac{1}{r^2\cos^2\phi}$
(a_1,c_1)	V^d	$\frac{1}{r}$	$\frac{\sqrt{1+\cos\phi}}{r^{1/2}}$	$\frac{\sqrt{1-\cos\phi}}{r^{1/2}}$
$(a_0,b_0,c_0)\equiv$ (a_0,b_0) or (b_0,c_0)	V^e	r^2	$r\cos\phi$	$r\sin\phi$

$$b(V_{\phi\phi}-r^2V_{rr})+(c-r^2a)V_{r\phi}+\left(\frac{r^2}{2}\right)\left[\left(\frac{4}{r}\right)b-3a_\phi\right]V_r+\left[\left(\frac{2}{r}\right)(c-r^2a)-3b_\phi\right]V_\phi=0.$$

Consequently, superintegrability arises for potentials satisfying *two* (nontrivial) equations of such a particular form. The following proposition states this property.

Proposition 1: Let the function V be the solution of the following system of two partial differential equations:

$$b(V_{\phi\phi}-r^2V_{rr})+(c-r^2a)V_{r\phi}+\left(\frac{r^2}{2}\right)\left[\left(\frac{4}{r}\right)b-3a_\phi\right]V_r+\left[\left(\frac{2}{r}\right)(c-r^2a)-3b_\phi\right]V_\phi=0,$$

$$B(V_{\phi\phi}-r^2V_{rr})+(C-r^2A)V_{r\phi}+\left(\frac{r^2}{2}\right)\left[\left(\frac{4}{r}\right)B-3A_\phi\right]V_r+\left[\left(\frac{2}{r}\right)(C-r^2A)-3B_\phi\right]V_\phi=0,$$

where $(a(r,\phi),b(r,\phi),c(r,\phi))$, *and* $(A(r,\phi),B(r,\phi),C(r,\phi))$, *are two sets of three functions determined by two sets of nontrivial real constants* (a_i,b_i,c_i), *and* (A_i,B_i,C_i), *respectively. Suppose that these two sets of constants are such that the two above second-order equations are independent. Then, if T denotes the Euclidean kinetic function* $T=(1/2)(v_r^2+r^2v_\phi^2)$, *the Lagrangian* $L=T-V(r,\phi)$ *is superintegrable with quadratic constants of motion.*

For further convenience, we sum up in the Table II the polar expressions for the sets of ''basic'' potentials V_1, V_2, V_3, in each of the families, as well as the associated sets of parameters.

We close this section with some observations. Any superintegrable potential should be associated to *two* different sets of constants (a_i,b_i,c_i) and (A_i,B_i,C_i), in addition to the trivial one $\{a_0=e_0,0,\ c_0=e_0;0,0;0\}$, which corresponds to the energy for all potentials. These two sets of values are (c_1,a_2) (i.e., $\{0,0,0;0,c_1;0\}$ and $\{0,0,0;0,0;a_2\}$) for V^c, (a_1,a_2) for $\tilde{V}^c$ and (a_1,c_1) for V^d. The situation for other families is a bit more subtle. V^b appears for two different choices for the two sets of constants, say (a_0,a_2) *or* (c_0,a_2), and V^e appears for (a_0,b_0) *or* (b_0,c_0). In both cases this twofold description is possible because the integral of motion associated to a_0, together with the energy—which is always present—implies the presence of another constant of motion associated to c_0, and conversely. Indeed, one can check that the systems of two fundamental equations for (a_0,a_2) and for (c_0,a_2) are equivalent. As far as V^a is concerned, the two possibilities (a_0,c_1) and (c_0,c_1) coincide again for the same reason, and the same happens for $\tilde{V}^a$. This will be relevant when discussing the situation for the case of nonzero curvature.

Overall, the simplest expression for these potentials is not in terms of any system of coordinates, but of *distances* to a given point O and/or to a pair of mutually orthogonal lines through O. In terms of these geometrical quantities r; $x=v$, $y=u$, we have

$$V^a=k_0+k_1(4x^2+y^2)+k_2x+k_3\frac{1}{y^2},$$

$$V^b=k_0+k_1r^2+k_2\frac{1}{x^2}+k_3\frac{1}{y^2},$$

$$V^c=k_0+k_1\frac{1}{r}+k_2\frac{1}{r}\frac{x}{y^2}+k_3\frac{1}{y^2},$$

$$V^d=k_0+k_1\frac{1}{r}+k_2\frac{1}{r}(r+x)^{1/2}+k_3\frac{1}{r}(r-x)^{1/2},$$

$$V^e=k_0+k_1r^2+k_2x+k_3y.$$

As we will see, superintegrable systems with two quadratic constants of motion in 2D spaces of constant curvature will have a comparable description.

IV. SUPERINTEGRABLE SYSTEMS ON THE TWO-DIMENSIONAL SPHERE S^2 AND HYPERBOLIC PLANE H^2 WITH QUADRATIC CONSTANTS OF MOTION

As observed in the Introduction, existing studies of potentials on the sphere have been mainly concerned with central potentials. There are some rather simple noncentral problems as, e.g., the nonisotropic oscillator or the Hénon–Heiles system, that have been highly studied in the plane but still remain very little analyzed in the sphere or the hyperbolic plane.

We now perform for the case of constant curvature κ a study similar to the one outlined for the Euclidean plane. We first use polar coordinates, and afterwards we shall give the expressions in parallel coordinates.

A. Polar coordinates

Let us consider the following Lagrangian:

$$L=\tfrac{1}{2}(v_r^2+\mathrm{S}_\kappa^2(r)v_\phi^2)-U(r,\phi)$$

corresponding to a system with configuration space Q (S^2 if $\kappa>0$ or H^2 if $\kappa<0$). In both cases κ is the curvature of Q. We will follow the same two step approach developed in Sec. III. The first step is concerned with the existence of *one* quadratic integral of motion (integrability). The second step corresponds with the existence of two quadratic integrals of motion (superintegrability). The following proposition summarizes the first step:

Proposition 2: Let L be the following two-degrees of freedom Lagrangian:

$$L=(\tfrac{1}{2})(v_r^2+\mathrm{S}_\kappa^2(r)v_\phi^2)-U(r,\phi)$$

corresponding to a system with a configuration space of constant curvature κ, and suppose that L has a constant of the motion $I=I(r,\phi,v_r,v_\phi)$ that is quadratic in the velocities

$$I=I_{22}+I_{20}(r,\phi),\quad I_{22}=av_r^2+2bv_rv_\phi+cv_\phi^2,$$

where a, b, and c, are functions of r and ϕ. Then,

(i) The three functions a, b, and c must take the form

$$a = a_0 \cos^2\phi + c_0 \sin^2\phi + b_0 \sin\phi\cos\phi,$$

$$b = (\tfrac{1}{2})\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r)[(c_0-a_0)\sin 2\phi + b_0\cos 2\phi] + (\tfrac{1}{2})\mathrm{S}_\kappa^2(r)(-a_1\cos\phi + c_1\sin\phi),$$

$$c = (\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r))^2(a_0\sin^2\phi + c_0\cos^2\phi - b_0\sin\phi\cos\phi) + \mathrm{S}_\kappa^3(r)\mathrm{C}_\kappa(r)(c_1\cos\phi + a_1\sin\phi) + a_2\mathrm{S}_\kappa^4(r),$$

where a_0, b_0, c_0; a_1, c_1; a_2 are real parameters.

(ii) The potential $U(r,\phi)$ must be solution of the following differential equation:

$$b(\mathrm{S}_\kappa^2(r)U_{rr} - U_{\phi\phi}) + (c - a\mathrm{S}_\kappa^2(r))U_{r\phi} + \mathrm{S}_\kappa^2(r)\left[2\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)b - \left(\frac{3}{2}\right)a_\phi\right]U_r + \left[2\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)(c - a\mathrm{S}_\kappa^2(r)) - 3b_\phi\right]U_\phi = 0.$$

Proof: Assume there exists a constant of motion $I = I_{22} + I_{20}$, where I_{22} takes the form $I_{22} = a v_x^2 + 2b v_x v_y + c v_y^2$, with the functions a, b, c depending on r,ϕ. These functions must satisfy the following four equations:

$$a_r = 0,$$

$$2b_r - 4\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)b + a_\phi = 0,$$

$$c_r - 4\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)c + (2\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r))a + 2b_\phi = 0,$$

$$c_\phi + (2\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r))b = 0.$$

Solving this system we obtain the expressions in statement (i) in the Proposition.

The function I_{20} is related to the potential U by the following system:

$$I_{20r} = 2aU_r + \frac{2bU_\phi}{\mathrm{S}_\kappa^2(r)}, \quad I_{20\phi} = 2bU_r + \frac{2cU_\phi}{\mathrm{S}_\kappa^2(r)}.$$

Taking derivatives, and using (i), we obtain the equation in point (ii) for the potential U.

Concerning (i), the most remarkable property of the expressions obtained for the three functions a, b, c, is that they depend on the *same* number of parameters $\{a_0, b_0, c_0;\ a_1, c_1;\ a_2\}$, as in the flat $\kappa=0$ case (then $\mathrm{S}_\kappa(r)\to r$ and $\mathrm{C}_\kappa(r)\to 1$ so all these equations coincide with the ones obtained in Sec. III for $Q=E^2$). In this κ-dependent form, the case where the configuration space is Euclidean appears not as a limit but simply as the particular case $\kappa=0$; generically r appears through $\mathrm{S}_\kappa(r)$, and there are also some factors which in the curved case appear through a tagged cosine of r, $\mathrm{C}_\kappa(r)$ which in the Euclidean case degenerates to $\mathrm{C}_0(r)\equiv 1$ and which therefore becomes invisible; of course these terms turn visible once we deal with the case of nonzero curvature κ.

As a, b, c are linear in the parameters a_i, b_i, c_i, the most general form for I_{22} is

$$I_{22} = a_0 I_{22}(a_0,\kappa) + b_0 I_{22}(b_0,\kappa) + c_0 I_{22}(c_0,\kappa) + a_1 I_{22}(a_1,\kappa) + c_1 I_{22}(c_1,\kappa) + a_2 I_{22}(a_2,\kappa),$$

where

$$I_{22}(a_0,\kappa)=(\cos\phi v_r-\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r)\sin\phi v_\phi)^2,$$

$$I_{22}(b_0,\kappa)=(\cos\phi v_r-\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r)\sin\phi v_\phi)(\sin\phi v_r+\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r)\cos\phi v_\phi),$$

$$I_{22}(c_0,\kappa)=(\sin\phi v_r+\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r)\cos\phi v_\phi)^2,$$

$$I_{22}(a_1,\kappa)=\mathrm{S}_\kappa^2(r)(\cos\phi v_r-\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r)\sin\phi v_\phi)v_\phi,$$

$$I_{22}(c_1,\kappa)=\mathrm{S}_\kappa^2(r)(\sin\phi v_r+\mathrm{S}_\kappa(r)\mathrm{C}_\kappa(r)\cos\phi v_\phi)v_\phi.$$

$$I_{22}(a_2,\kappa)=\mathrm{S}_\kappa^4(r)v_\phi^2.$$

A direct calculation shows that

$$I_{22}(a_0,\kappa)+I_{22}(c_0,\kappa)+\kappa[I_{22}(a_2,\kappa)]=v_r^2+\mathrm{S}_\kappa^2(r)v_\phi^2$$

so in any configuration space of nonzero constant curvature κ the kinetic term can always be written as a sum of *three* summands, one of which carries the curvature κ and vanishes into the limit $\kappa\to 0$. In the particular cases where the total energy $E=T+U$ becomes totally separable, it will appear as a sum, not of two (as in E^2), but of three components. This is an interesting characteristic of this trivial first integral of the motion making the analysis of the different families of superintegrable potentials in the sphere and the hyperbolic plane richer than in the Euclidean case.

The above Proposition 2 relates integrability with the property of satisfying a *single* differential equation for the potential. In $Q=S^2$ or $Q=H^2$, as in $Q=E^2$, two independent integrals of motion (further than the energy) are required for the superintegrability of the system. The following proposition 3 states the relation of superintegrability in configuration spaces of constant curvature with the fact that the potential should satisfy a system of two such equations.

Proposition 3: Let the potential U be a solution of the following system of two partial differential equations,

$$b(\mathrm{S}_\kappa^2(r)U_{rr}-U_{\phi\phi})+(c-a\mathrm{S}_\kappa^2(r))U_{r\phi}+\mathrm{S}_\kappa^2(r)\left[2\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)b-\left(\frac{3}{2}\right)a_\phi\right]U_r$$
$$+\left[2\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)(c-a\mathrm{S}_\kappa^2(r))-3b_\phi\right]U_\phi=0,$$

$$B(\mathrm{S}_\kappa^2(r)U_{rr}-U_{\phi\phi})+(C-A\mathrm{S}_\kappa^2(r))U_{r\phi}+\mathrm{S}_\kappa^2(r)\left[2\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)B-\left(\frac{3}{2}\right)A_\phi\right]U_r$$
$$+\left[2\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}\right)(C-A\mathrm{S}_\kappa^2(r))-3B_\phi\right]U_\phi=0,$$

where $(a(r,\phi),b(r,\phi),c(r,\phi))$, *and* $(A(r,\phi),B(r,\phi),C(r,\phi))$, *are two sets of three functions determined as in Prop. 2(i) by two sets of nontrivial real constants* (a_i,b_i,c_i), *and* (A_i,B_i,C_i), *respectively. Suppose that these two sets of constants are not only different but such that the two above second-order equations are independent. Then, if* $T=(1/2)(v_r^2+\mathrm{S}_\kappa^2(r)v_\phi^2)$ *denotes the kinetic function in a configuration space of constant curvature* κ, *the Lagrangian* $L=T-U(r,\phi)$ *describes a superintegrable system on either the sphere* $(\kappa>0)$, *the Euclidean plane* $(\kappa=0)$ *or the hyperbolic plane* $(\kappa<0)$, *with quadratic constants of motion.*

We will study now the spherical or hyperbolic versions of the superintegrable potentials already presented for the Euclidean plane. For consistency with the flat case we will denote them as U^a, $\widetilde{U}^a$, U^{aa}, $\widetilde{U}^{aa}$, U^b, U^c, U^d, U^e, and $\widetilde{U}^e$. They correspond to the choices of parameters given in Table III.

TABLE III. Parameters and non-Euclidean families of potentials.

$(a_0,c_1);(c_0,a_1)$	$U^a;\tilde{U}^a$
$(c_0,c_1);(a_0,a_1)$	$U^{aa};\tilde{U}^{aa}$
$(a_0,c_0,a_2)\equiv(a_0,a_2)$ or (c_0,a_2) or (a_0,c_0)	U^b
$(c_1,a_2);(a_1,a_2)$	$U^c;\tilde{U}^c$
(a_1,c_1)	U^d
$(a_0,b_0);(b_0,c_0)$	$U^e;\tilde{U}^e$

The family U^b is again characterized by three independent parameter vectors {100;00;0}, {001;00;0} and {000;00;1}; the vector corresponding to the energy is already contained in the subspace generated by these three vectors. In other cases, the parameter vector corresponding to the energy will be assumed without explicit mention. The family V^a, which in the flat case corresponded *simultaneously* to two pairs pairs (a_0,c_1) and (c_0,c_1) will now split into two *different* families U^a and U^{aa}, associated, respectively, with (a_0,c_1) and (c_0,c_1). A similar splitting occurs for the flat families $\tilde{V}^a$ and V^e. As we shall see shortly, and likewise as in $Q=E^2$, it is again remarkable that the general solution for U appears also as a linear combination of simpler particular nontrivial solutions (three for almost all families), in addition to the constant potential $U_0=1$, which is of course superintegrable and has associated the full the six-dimensional parameter space.

1. Superintegrable systems on the unit sphere

We first present a detailed study of the resolution of the equations for the particular case $\kappa=1$. The complete rewriting of the solutions for any arbitrary (positive or negative) value of κ is done afterwards.

a. Family U^a. Notice that the case (a_0,c_0,c_1) is not possible for $\kappa\neq 0$. In fact we have obtained that it splits in two different subcases (a_0,c_1) and (c_0,c_1).

We first consider the two linear second-order equations associated to (a_0,C_1), that is, to the vectors $\{a_0 00;00;0\}$ and $\{000;0C_1;0\}$. When $\kappa=1$ they are

$$\left(\frac{\cos r}{\sin r}\right)(\sin\phi\cos\phi)(\sin^2 r\,U_{rr}-U_{\phi\phi})+(\cos^2\phi-\cos^2 r\sin^2\phi)U_{r\phi}$$
$$+(2\cos^2 r-3)(\sin\phi\cos\phi)U_r+\left(\frac{\cos r}{\sin r}\right)[(3-2\cos^2 r)\sin^2\phi-\cos^2\phi]U_\phi=0,$$

$$\sin\phi(\sin^2 r\,U_{rr}-U_{\phi\phi})+2(\sin r\cos r\cos\phi)U_{r\phi}+2(\sin r\cos r\sin\phi)U_r$$
$$-\cos\phi(3-4\cos^2 r)U_\phi=0.$$

The first equation, that corresponds to $a_0\neq 0$, can be reduced to canonical form by means of the following change of variables

$$(r,\phi)\to(w_a,z_a),\quad w_a=\sin r\sin\phi,\quad z_a=\tan r\cos\phi.$$

The general solution takes the form

$$U(r,\phi;a_0)=F(w_a)+\frac{G(z_a)}{1-w_a^2}=F(\sin r\sin\phi)+\frac{G(\tan r\cos\phi)}{1-(\sin r\sin\phi)^2}.$$

Substituting in the equation for $C_1\neq 0$ we have found the following potentials:

$$U^a=U^a(r,\phi;a_0,C_1)=k_0+k_1U_1^a+k_2U_2^a+k_3U_3^a,$$

$$U_1^a=\frac{1}{1-w_a^2}\left[4\left(\frac{z_a}{1-z_a^2}\right)^2+w_a^2\right]=\frac{1}{1-(\sin r\sin\phi)^2}\left[4\left(\frac{\tan r\cos\phi}{1-(\tan r\cos\phi)^2}\right)^2+(\sin r\sin\phi)^2\right],$$

$$U_2^a=\frac{1}{1-w_a^2}\left[\frac{(1+z_a^2)z_a}{(1-z_a^2)^2}\right]=\frac{\tan r\cos\phi}{(\cos^2 r)[1-(\tan r\cos\phi)^2]^2},$$

$$U_3^a=\frac{1}{w_a^2}=\frac{1}{(\sin r\sin\phi)^2},$$

where k_0 and k_i, $i=1,2,3$, are arbitrary constants (we display explicitly the constant term k_0, linked to the constant potential $U_0=1$ which will appear in all families). The two constants of motion, I_2^a and I_3^a, take the form

$$I_2^a=(\cos\phi v_r-\sin r\cos r\sin\phi v_\phi)^2+\frac{8k_1(\tan r\cos\phi)^2}{[1-(\tan r\cos\phi)^2]^2}+\frac{2k_2[1+(\tan r\cos\phi)^2](\tan r\cos\phi)}{[1-(\tan r\cos\phi)^2]^2},$$

$$I_3^a=(\sin r)^2(\sin\phi v_r+\sin r\cos r\cos\phi v_\phi)v_\phi-\frac{2k_1(\tan^3 r\cos\phi\sin^2\phi)}{[1-(\tan r\cos\phi)^2]^2}-\frac{k_2[1+(\tan r\cos\phi)^2](\tan r\sin\phi)^2}{2[1-(\tan r\cos\phi)^2]^2}+\frac{2k_3\cos r\cos\phi}{\sin r\sin^2\phi}.$$

b. Family $\tilde{U}^a$. We can also consider the family of potentials associated with (c_0,A_1). The equations are

$$\left(\frac{\cos r}{\sin r}\right)(\sin\phi\cos\phi)(\sin^2 r U_{rr}-U_{\phi\phi})+(\cos^2 r\cos^2\phi-\sin^2\phi)U_{r\phi}+(2\cos^2 r-3)(\sin\phi\cos\phi)U_r+\left(\frac{\cos r}{\sin r}\right)[\sin^2\phi-(3-2\cos^2 r)\cos^2\phi]U_\phi=0,$$

$$\cos\phi(\sin^2 r U_{rr}-U_{\phi\phi})-2(\sin r\cos r\sin\phi)U_{r\phi}+2(\sin r\cos r\cos\phi)U_r-\sin\phi(3-4\cos^2 r)U_\phi=0.$$

The first equation, that corresponds to $c_0\neq 0$, can be reduced to canonical form making use of the following change:

$$(r,\phi)\to(w_c,z_c),\quad w_c=\sin r\cos\phi,\quad z_c=\tan r\sin\phi.$$

The general solution takes the form

$$\tilde{U}(r,\phi;c_0)=\tilde{F}(w_c)+\frac{\tilde{G}(z_c)}{1-w_c^2}=\tilde{F}(\sin r\cos\phi)+\frac{\tilde{G}(\tan r\sin\phi)}{1-(\sin r\cos\phi)^2}.$$

Substituting this general expression into the equation for $A_1\neq 0$ we have found the following potentials:

$$\tilde{U}^a=\tilde{U}^a(r,\phi;c_0,A_1)=k_0+k_1\tilde{U}_1^a+k_2\tilde{U}_2^a+k_3\tilde{U}_3^a,$$

$$\tilde{U}_1^a = \frac{1}{1-w_c^2}\left[w_c^2 + 4\left(\frac{z_c}{1-z_c^2} \right)^2 \right] = \frac{1}{1-(\sin r \cos \phi)^2}\left[(\sin \cos \phi)^2 + 4\left(\frac{\tan r \sin \phi}{1-(\tan r \sin \phi)^2} \right)^2 \right],$$

$$\tilde{U}_2^a = \frac{1}{1-w_c^2}\left[\frac{(1+z_c^2)z_c}{(1-z_c^2)^2} \right] = \frac{\tan r \sin \phi}{(\cos^2 r)[1-(\tan r \sin \phi)^2]^2},$$

$$\tilde{U}_3^a = \frac{1}{w_c^2} = \frac{1}{(\sin r \cos \phi)^2}.$$

c. Family U^{aa}. When $\kappa \neq 0$ we can also consider the system of second-order equations associated with (c_0, C_1), that is, to the vectors $\{00c_0;00;0\}$ and $\{000;0C_1;0\}$. They are

$$\left(\frac{\cos r}{\sin r} \right)(\sin \phi \cos \phi)(\sin^2 r U_{rr} - U_{\phi\phi}) + (\cos^2 r \cos^2 \phi - \sin^2 \phi)U_{r\phi} + (2\cos^2 r - 3)(\sin \phi \cos \phi)U_r + \left(\frac{\cos r}{\sin r} \right)[\sin^2 \phi - (3 - 2\cos^2 r)\cos^2 \phi]U_\phi = 0,$$

$$\sin \phi(\sin^2 r U_{rr} - U_{\phi\phi}) + 2(\sin r \cos r \cos \phi)U_{r\phi} + 2(\sin r \cos r \sin \phi)U_r - \cos \phi(3 - 4\cos^2 r)U_\phi = 0.$$

We have found the following solutions:

$$U^{aa} = U^{aa}(r,\phi;c_0,C_1) = k_0 + k_1 U_1^{aa} + k_2 U_2^{aa} + k_3 U_3^{aa},$$

$$U_1^{aa} = \frac{w_c}{\sqrt{1-w_c^2}} = \frac{\sin r \cos \phi}{\sqrt{1-(\sin r \cos \phi)^2}},$$

$$U_2^{aa} = \frac{\sqrt{1+z_c^2}}{(1-w_c^2)z_c^2} = \frac{\cos r}{(\sin r \sin \phi)^2 \sqrt{1-(\sin r \cos \phi)^2}},$$

$$U_3^{aa} = \frac{1+z_c^2}{(1-w_c^2)z_c^2} = \frac{1}{(\sin r \sin \phi)^2},$$

and the following two constants of motion:

$$I_2^{aa} = (\sin \phi v_r + \sin r \cos r \cos \phi v_\phi)^2 + \frac{2k_2 \cos r \sqrt{1-(\sin r \cos \phi)^2}}{(\sin r \sin \phi)^2} + \frac{2k_3 \cos^2 r}{(\sin r \sin \phi)^2},$$

$$I_3^{aa} = (\sin r)^2(\sin \phi v_r + \sin r \cos r \cos \phi v_\phi)v_\phi + \frac{k_1 \cos r}{\sqrt{1-(\sin r \cos \phi)^2}} + \frac{k_2 \cos \phi[2\cos^2 r + (\sin r \sin \phi)^2]}{(\sin r \sin \phi)^2 \sqrt{1-(\sin r \cos \phi)^2}} + \frac{2k_3 \cos r \cos \phi}{\sin r \sin^2 \phi}.$$

d. Family $\tilde{U}^{aa}$. Similarly, the general solution of the system $(a_0 \neq 0,\ A_1 \neq 0)$, that we will denote by $\tilde{U}^{aa} = \tilde{U}^{aa}(r,\phi;a_0,A_1)$, is given by

$$\widetilde{U}^{aa}=\widetilde{U}^{aa}(r,\phi;a_0,A_1)=k_0+k_1\widetilde{U}_1^{aa}+k_2\widetilde{U}_2^{aa}+k_3\widetilde{U}_3^{aa},$$

$$\widetilde{U}_1^{aa}=\frac{w_a}{\sqrt{1-w_a^2}}=\frac{\sin r\sin\phi}{\sqrt{1-(\sin r\sin\phi)^2}},$$

$$\widetilde{U}_2^{aa}=\frac{\sqrt{1+z_a^2}}{(1-w_a^2)z_a^2}=\frac{\cos r}{(\sin r\cos\phi)^2\sqrt{1-(\sin r\sin\phi)^2}},$$

$$\widetilde{U}_3^{aa}=\frac{1+z_a^2}{(1-w_a^2)z_a^2}=\frac{1}{(\sin r\cos\phi)^2}.$$

e. Family U^b. The two linear second-order equations are those associated with (a_0,C_0), that is, to the vectors $\{a_000;00;0\}$ and $\{00C_0;00;0\}$. Before proceeding, we note that any of the alternative choices (a_0,A_2) or (c_0,A_2) would lead exactly to the same family. The two equations for a_0 and C_0 are

$$\left(\frac{\cos r}{\sin r}\right)(\sin\phi\cos\phi)(\sin^2 rU_{rr}-U_{\phi\phi})+(\cos^2\phi-\cos^2 r\sin^2\phi)U_{r\phi}+(2\cos^2 r-3)$$
$$\times(\sin\phi\cos\phi)U_r+\left(\frac{\cos r}{\sin r}\right)[(3-2\cos^2 r)\sin^2\phi-\cos^2\phi]U_\phi=0,$$

$$\left(\frac{\cos r}{\sin r}\right)(\sin\phi\cos\phi)(\sin^2 rU_{rr}-U_{\phi\phi})+(\cos^2 r\cos^2\phi-\sin^2\phi)U_{r\phi}+(2\cos^2 r-3)$$
$$\times(\sin\phi\cos\phi)U_r+\left(\frac{\cos r}{\sin r}\right)[\sin^2\phi-(3-2\cos^2 r)\cos^2\phi]U_\phi=0.$$

The general solution of the first equation, that corresponds to $a_0\neq 0$, takes the form

$$U(r,\phi;a_0)=F(w_a)+\frac{G(z_a)}{1-w_a^2}=F(\sin r\sin\phi)+\frac{G(\tan r\cos\phi)}{1-(\sin r\sin\phi)^2}.$$

Notice that, in the parallel coordinates (u,y), the canonical variables can be written as $w_a=\sin y$, $z_a=\tan u$, so $U(r,\phi;a_0)$ can also be written as

$$U(r,\phi;a_0)=f(\sin y)+\frac{g(\tan u)}{\cos^2 y}.$$

Concerning the second equation, corresponding to $C_0\neq 0$, the canonical coordinates are $w_c=\sin r\cos\phi$, $z_c=\tan r\sin\phi$, which can be alternatively rewritten as $w_c=\sin x$, $z_c=\tan v$. The corresponding general solution takes the form

$$\widetilde{U}(r,\phi;c_0)=\widetilde{F}(w_c)+\frac{\widetilde{G}(z_c)}{1-w_c^2}=\widetilde{F}(\sin r\cos\phi)+\frac{\widetilde{G}(\tan r\sin\phi)}{1-(\sin r\cos\phi)^2}=\widetilde{f}(\sin x)+\frac{\widetilde{g}(\tan v)}{\cos^2 x}.$$

The corresponding general solution for the potential U, that we will denote by $U^b=U^b(r,\phi;a_0,C_0)$, must be expressible simultaneously as $U(r,\phi;a_0)$ and as $U(r,\phi;C_0)$; this is enforced by requiring that the explicit form for the general solution of the equation for $a_0\neq 0$ solves the equation for $C_0\neq 0$. This leads to the following solutions for U:

$$U^b=U^b(r,\phi;a_0,C_0)=k_0+k_1U_1^b+k_2U_2^b+k_3U_3^b,$$

$$U_1^b=\frac{w_a^2+z_a^2}{1-w_a^2}=\frac{w_c^2+z_c^2}{1-w_c^2}=\left(\frac{\sin r}{\cos r}\right)^2,$$

$$U_2^b=\frac{1+z_a^2}{(1-w_a^2)z_a^2}=\frac{1}{w_c^2}=\frac{1}{(\sin r\cos\phi)^2},$$

$$U_3^b=\frac{1}{w_a^2}=\frac{1+z_c^2}{(1-w_c^2)z_c^2}=\frac{1}{(\sin r\sin\phi)^2},$$

where k_0 and k_i, $i=1,2,3$, are arbitrary constants. As indicated, these potentials are automatically solution of the equation for $a_2\neq 0$.

The two constants of motion, I_2^b and I_3^b, take the form

$$I_2^b=(\cos\phi v_r-\sin r\cos r\sin\phi v_\phi)^2+2k_1\left(\frac{\sin r}{\cos r}\right)^2\cos^2\phi+2k_2\left(\frac{\cos r}{\sin r\cos\phi}\right)^2,$$

$$I_3^b=(\sin\phi v_r+\sin r\cos r\cos\phi v_\phi)^2+2k_1\left(\frac{\sin r}{\cos r}\right)^2\sin^2\phi+2k_3\left(\frac{\cos r}{\sin r\sin\phi}\right)^2.$$

Notice that, in this particular case, we can consider as the first integral of motion the following function:

$$I_1^b=(\sin r)^4v_\phi^2+2\left(\frac{k_2}{\cos^2\phi}\right)+2\left(\frac{k_3}{\sin^2\phi}\right).$$

f. Family U^c. The second-order equations correspond to $\{000;00;a_2\}$ and $\{000;0C_1;0\}$. They are

$$\sin r U_{r\phi}+2\cos r U_\phi=0,$$

$$\sin\phi(\sin^2 r U_{rr}-U_{\phi\phi})+2(\sin r\cos r\cos\phi)U_{r\phi}+2(\sin r\cos r\sin\phi)U_r$$
$$-\cos\phi(3-4\cos^2 r)U_\phi=0.$$

The first equation, that corresponds to $a_2\neq 0$, has as general solution the following family of functions:

$$U(r,\phi;a_2)=F(r)+\frac{G(\phi)}{(\sin r)^2}.$$

Substituting $U(r,\phi;a_2)$ in the second equation, that corresponds to $C_1\neq 0$, we obtain the following potentials:

$$U^c=U^c(r,\phi;a_2,C_1)=k_0+k_1U_1^c+k_2U_2^c+k_3U_3^c,$$

$$U_1^c=\frac{\cos r}{\sin r},\quad U_2^c=\frac{\cos\phi}{(\sin r\sin\phi)^2},\quad U_3^c=\frac{1}{(\sin r\sin\phi)^2}.$$

The two constants of motion I_2^c, and I_3^c, take the form

$$I_2^c = (\sin r)^4 v_\phi^2 + 2k_2\left(\frac{\cos\phi}{\sin^2\phi}\right) + 2k_3\left(\frac{\cos\phi}{\sin\phi}\right)^2,$$

$$I_3^c = (\sin r)^2(\sin\phi v_r + \sin r\cos r\cos\phi v_\phi)v_\phi + k_1\cos\phi$$
$$+k_2\left(\frac{\cos r(1+\cos^2\phi)}{\sin r\sin^2\phi}\right) + 2k_3\left(\frac{\cos r\cos\phi}{\sin r\sin^2\phi}\right).$$

g. Family $\tilde{U}^c$. The second-order equations correspond to $\{000;00;a_2\}$ and $\{000;0A_1;0\}$. They are

$$\sin r U_{r\phi} + 2\cos r U_\phi = 0,$$

$$\cos\phi(\sin^2 r U_{rr} - U_{\phi\phi}) - 2(\sin r\cos r\sin\phi)U_{r\phi}$$
$$+2(\sin r\cos r\cos\phi)U_r + \sin\phi(3-4\cos^2 r)U_\phi = 0.$$

In a rather similar way to the previous case, we have obtained the following potentials:

$$\tilde{U}^c = \tilde{U}^c(r,\phi;a_2,A_1) = k_0 + k_1\tilde{U}_1^c + k_2\tilde{U}_2^c + k_3\tilde{U}_3^c,$$

$$\tilde{U}_1^c = \frac{\cos r}{\sin r}, \quad \tilde{U}_2^c = \frac{1}{(\sin r\cos\phi)^2}, \quad \tilde{U}_3^c = \frac{\sin\phi}{(\sin r\cos\phi)^2}.$$

h. Family U^d. The two linear second-order equations corresponding to $\{000;a_10;0\}$ and to $\{000;0C_1;0\}$ are the following:

$$\sin\phi(\sin^2 r U_{rr} - U_{\phi\phi}) + 2(\sin r\cos r\cos\phi)U_{r\phi}$$
$$+2(\sin r\cos r\sin\phi)U_r - \cos\phi(3-4\cos^2 r)U_\phi = 0,$$

$$\cos\phi(\sin^2 r U_{rr} - U_{\phi\phi}) - 2(\sin r\cos r\sin\phi)U_{r\phi}$$
$$+2(\sin r\cos r\cos\phi)U_r + \sin\phi(3-4\cos^2 r)U_\phi = 0,$$

that can be equivalently rewritten as

$$2\sin r\cos r U_{r\phi} + (4\cos^2 r - 3)U_\phi = 0,$$

$$2\sin r\cos r U_r + (\sin^2 r U_{rr} - U_{\phi\phi}) = 0.$$

The general solution for the first equation is $F(r) + G(\phi)/\sqrt{\sin r\cos^3 r}$. Substituting this expression in the second one we arrive at $F(r) = k_0 + k_1(\cos r/\sin r)$, and $G=0$. Thus the most general form of a potential in the family U^d is given by

$$U^d = U^d(r,\phi;a_1,C_1) = k_0 + k_1 U_1^d,$$

$$U_1^d = \frac{1}{\tan r},$$

where k_0 and k_1 are again arbitrary constants.

The two constants of motion, I_2^d and I_3^d, take the form

$$I_2^d = (\sin r)^2(\sin\phi v_r - \sin r\cos r\cos\phi v_\phi)v_\phi - k_1\sin\phi,$$

$$I_3^d = (\sin r)^2(\cos\phi v_r + \sin r\cos r\sin\phi v_\phi)v_\phi + k_1\cos\phi.$$

i. Family U^e. The two linear second-order equations are those associated with $(a_0\neq 0,\ B_0\neq 0)$, that is, to the vectors $\{a_0 00;00;0\}$ and $\{0B_0 0;00;0\}$. The two equations for a_0 and B_0 are

$$\left(\frac{\cos r}{\sin r}\right)(\sin\phi\cos\phi)(\sin^2 r\,U_{rr}-U_{\phi\phi})+(\cos^2\phi-\cos^2 r\,\sin^2\phi)U_{r\phi}$$
$$+(2\cos^2 r-3)(\sin\phi\cos\phi)U_r+\left(\frac{\cos r}{\sin r}\right)[(3-2\cos^2 r)\sin^2\phi-\cos^2\phi]U_\phi=0,$$

and

$$\left(\frac{\cos r}{\sin r}\right)(\cos^2\phi-\sin^2\phi)(\sin^2 r\,U_{rr}-U_{\phi\phi})-2(1+\cos^2 r)(\cos\phi\sin\phi)U_{r\phi}$$
$$+(\cos^2 r-\sin^2 r-2)(\cos^2\phi-\sin^2\phi)U_r+4\left(\frac{\cos r}{\sin r}+\cos r\sin r\right)(\cos\phi\sin\phi)U_\phi=0.$$

The general solution of this system, that we will denote by $U^e=U^e(r,\phi;a_0,B_0)$, turns out to be

$$U^e=U^e(r,\phi;a_0,B_0)=k_0+k_1U_1^e+k_2U_2^e+k_3U_3^e,$$

$$U_1^e=\frac{w_a^2+z_a^2}{1-w_a^2}=(\tan r)^2,$$

$$U_2^e=\frac{z_a\sqrt{1+z_a^2}}{1-w_a^2}=\frac{\tan r\cos\phi}{(\cos r)\sqrt{1-(\sin r\sin\phi)^2}},$$

$$U_3^e=\frac{w_a}{\sqrt{1-w_a^2}}=\frac{\sin r\sin\phi}{\sqrt{1-(\sin r\sin\phi)^2}},$$

and the associated two constants of motion, I_2^e and I_3^e, are given by

$$I_2^e=(\cos\phi v_r-\sin r\cos r\sin\phi v_\phi)^2+2k_1(\tan r\cos\phi)^2+2k_2\left(\frac{\sin r\cos\phi}{\cos^2 r}\right)\sqrt{1-(\sin r\sin\phi)^2},$$

$$I_3^e=(\cos\phi v_r-\sin r\cos r\sin\phi v_\phi)(\sin\phi v_r+\sin r\cos r\cos\phi v_\phi)+2k_1\tan^2 r\sin\phi\cos\phi$$
$$+k_2\sin r\sin\phi\left(\frac{1+2\tan^2 r\cos^2\phi}{\sqrt{1-(\sin r\sin\phi)^2}}\right)+\frac{k_3\sin r\cos\phi}{\sqrt{1-(\sin r\sin\phi)^2}}.$$

j. Family $\tilde U^e$. Similarly, the general solution of the system $(b_0\neq 0,\ C_0\neq 0)$, that we will denote by $\tilde U^e=\tilde U^e(r,\phi;b_0,C_0)$, is given by

$$\tilde U^e=U^e(r,\phi;b_0,C_0)=k_0+k_1\tilde U_1^e+k_2\tilde U_2^e+k_3\tilde U_3^e,$$

$$\tilde U_1^e=\frac{w_c^2+z_c^2}{1-w_c^2}=(\tan r)^2,$$

$$\tilde U_2^e=\frac{w_c}{\sqrt{1-w_c^2}}=\frac{\sin r\cos\phi}{\sqrt{1-(\sin r\cos\phi)^2}},$$

$$\widetilde{U}_3^e = \frac{z_c\sqrt{1+z_c^2}}{1-w_c^2} = \frac{\tan r \sin\phi}{(\cos r)\sqrt{1-(\sin r\cos\phi)^2}}.$$

2. Superintegrable systems on a 2D space of constant curvature

Next we present the results obtained for the potentials U^r and the integrals of motion I_2, I_3, for more general values (positive, zero or negative) of the curvature κ. Thus, the following expressions simultaneously includes spherical, Euclidean, and hyperbolic potentials. We have not included the tilded families $\widetilde{U}^a$, $\widetilde{U}^{aa}$, $\widetilde{U}^c$, and $\widetilde{U}^e$, which for all values of κ can be obtained from U^a, U^{aa}, U^c, and U^e by reflection in the geodesic $x=y$, which in polar coordinates amounts to the interchange $\cos\phi \leftrightarrow \sin\phi$.

a. Family U^a. The general κ-dependent form of the potential U^a is given by

$$U^a = U^a(r,\phi,\kappa;a_0,C_1) = k_0 + k_1 U_1^a + k_2 U_2^a + k_3 U_3^a,$$

$$U_1^a = \frac{1}{1-\kappa(\mathrm{S}_\kappa(r)\sin\phi)^2}\left[4\left(\frac{\mathrm{T}_\kappa(r)\cos\phi}{1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2}\right)^2 + (\mathrm{S}_\kappa(r)\sin\phi)^2\right],$$

$$U_2^a = \frac{\mathrm{T}_\kappa(r)\cos\phi}{\mathrm{C}_k^2(r)[1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2]^2},$$

$$U_3^a = \frac{1}{\mathrm{S}_\kappa^2(r)\sin^2\phi},$$

and the two κ-dependent constants of motion, I_2^a and I_3^a, take the form

$$I_2^a = I_{22}(a_0,\kappa) + \frac{8k_1(\mathrm{T}_\kappa(r)\cos\phi)^2}{[1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2]^2} + \frac{2k_2[1+\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2](\mathrm{T}_\kappa(r)\cos\phi)}{[1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2]^2},$$

$$I_3^a = I_{22}(c_1,\kappa) - \frac{2k_1\mathrm{T}_\kappa^3(r)\cos\phi\sin^2\phi}{[1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2]^2} - \frac{k_2[1+\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2](\mathrm{T}_\kappa(r)\sin\phi)^2}{2[1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2]^2} + \frac{2k_3\mathrm{C}_\kappa(r)\cos\phi}{\mathrm{S}_\kappa(r)\sin^2\phi}.$$

Notice that for the particular case $\kappa=0$ we obtain

$$U^a = k_0 + k_1 r^2(4\cos^2\phi + \sin^2\phi) + k_2 r\cos\phi + \frac{k_3}{(r\sin\phi)^2},$$

$$I_2^a = (v_r\cos\phi - r v_\phi\sin\phi)^2 + 8k_1(r\cos\phi)^2 + 2k_2 r\cos\phi,$$

$$I_3^a = (v_r\sin\phi + r v_\phi\cos\phi)(r^2 v_\phi) - 2k_1 r^3\cos\phi\sin^2\phi - \left(\frac{1}{2}\right)k_2(r\sin\phi)^2 + \frac{2k_3\cos\phi}{r\sin^2\phi}$$

that coincide with the corresponding expressions obtained in Sec. III for the Euclidean plane.

b. Family U^{aa}. The general κ-dependent form of the potential U^{aa} is given by

$$U^{aa} = U^{aa}(r,\phi,\kappa;c_0,C_1) = k_0 + k_1 U_1^{aa} + k_2 U_2^{aa} + k_3 U_3^{aa},$$

$$U_1^{aa} = \frac{\mathrm{S}_\kappa(r)\cos\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2}},$$

$$U_2^{aa} = \frac{1}{[\mathrm{S}_\kappa(r)\sin\phi]^2}\,\frac{\mathrm{C}_\kappa(r)}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2}},$$

$$U_3^{aa} = \frac{1}{[\mathrm{S}_\kappa(r)\sin\phi]^2},$$

and the two κ-dependent constants of motion, I_2^{aa} and I_3^{aa}, take the form

$$I_2^{aa} = I_{22}(c_0,\kappa) + \frac{2k_2\mathrm{C}_\kappa(r)}{[\mathrm{S}_\kappa(r)\sin\phi]^2}\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2} + \frac{2k_3\mathrm{C}_\kappa^2(r)}{[\mathrm{S}_\kappa(r)\sin\phi]^2},$$

$$I_3^{aa} = I_{22}(c_1,\kappa) + \frac{k_1\mathrm{C}_\kappa(r)}{\kappa\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2}} + \frac{k_2\cos\phi[2\mathrm{C}_\kappa^2(r)+\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2]}{\mathrm{S}_\kappa(r)\sin^2\phi\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2}} + \frac{2k_3\mathrm{C}_\kappa(r)\cos\phi}{\mathrm{S}_\kappa(r)\sin^2\phi}.$$

c. Family U^b. The general κ-dependent form of the potential U^b is given by

$$U^b = U^b(r,\phi,\kappa;a_0,C_0) = k_0 + k_1U_1^b + k_2U_2^b + k_3U_3^b,$$

$$U_1^b = \mathrm{T}_\kappa^2(r), \quad U_2^b = \frac{1}{(\mathrm{S}_\kappa(r)\cos\phi)^2}, \quad U_3^b = \frac{1}{(\mathrm{S}_\kappa(r)\sin\phi)^2},$$

and the two κ-dependent constants of motion, I_2^b and I_3^b, take the form

$$I_2^b = I_{22}(a_0,\kappa) + 2k_1\mathrm{T}_\kappa^2(r)\cos^2\phi + 2k_2\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)\cos\phi}\right)^2,$$

$$I_3^b = I_{22}(c_0,\kappa) + 2k_1\mathrm{T}_\kappa^2(r)\sin^2\phi + 2k_3\left(\frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)\sin\phi}\right)^2.$$

d. Family U^c. The general κ-dependent form of the potential U^c is given by

$$U^c = U^c(r,\phi,\kappa;a_2,C_1) = k_0 + k_1U_1^c + k_2U_2^c + k_3U_3^c,$$

$$U_1^c = \frac{\mathrm{C}_\kappa(r)}{\mathrm{S}_\kappa(r)}, \quad U_2^c = \frac{\cos\phi}{(\mathrm{S}_\kappa(r)\sin\phi)^2}, \quad U_3^c = \frac{1}{(\mathrm{S}_\kappa(r)\sin\phi)^2},$$

and the two κ-dependent constants of motion, I_2^c and I_3^c, take the form

$$I_2^c = I_{22}(a_2,\kappa) + 2k_2\left(\frac{\cos\phi}{\sin^2\phi}\right) + 2k_3\left(\frac{\cos\phi}{\sin\phi}\right)^2$$

$$I_3^c = I_{22}(c_1,\kappa) + k_1\cos\phi + k_2\left(\frac{\mathrm{C}_\kappa(r)(1+\cos^2\phi)}{\mathrm{S}_\kappa(r)\sin^2\phi}\right) + 2k_3\left(\frac{\mathrm{C}_\kappa(r)\cos\phi}{\mathrm{S}_\kappa(r)\sin^2\phi}\right).$$

e. Family U^d. The general κ-dependent form of the potential U^d is given by

TABLE IV. Superintegrable potentials with curvature κ: Polar coordinates.

U^a	$\dfrac{4\left(\dfrac{\mathrm{T}_\kappa(r)\cos\phi}{1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2}\right)^2+(\mathrm{S}_\kappa(r)\sin\phi)^2}{1-\kappa(\mathrm{S}_\kappa(r)\sin\phi)^2}$	$\dfrac{[\mathrm{T}_\kappa(r)/\mathrm{C}_k^2(r)]\cos\phi}{[1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2]^2}$	$\dfrac{1}{\mathrm{S}_\kappa^2(r)\sin^2\phi}$
U^{aa}	$\dfrac{\mathrm{S}_\kappa(r)\cos\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2}}$	$\dfrac{\mathrm{C}_\kappa(r)/[\mathrm{S}_\kappa(r)\sin\phi]^2}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2}}$	$\dfrac{1}{\mathrm{S}_\kappa^2(r)\sin^2\phi}$
U^b	$\mathrm{T}_\kappa^2(r)$	$\dfrac{1}{\mathrm{S}_\kappa^2(r)\cos^2\phi}$	$\dfrac{1}{\mathrm{S}_\kappa^2(r)\sin^2\phi}$
U^c	$\dfrac{1}{\mathrm{T}_\kappa(r)}$	$\dfrac{\cos\phi}{\mathrm{S}_\kappa^2(r)\sin^2\phi}$	$\dfrac{1}{\mathrm{S}_\kappa^2(r)\sin^2\phi}$
U^d	$\dfrac{1}{\mathrm{T}_\kappa(r)}$		
U^e	$\mathrm{T}_\kappa^2(r)$	$\dfrac{[\mathrm{T}_\kappa(r)/\mathrm{C}_k(r)]\cos\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\sin\phi]^2}}$	$\dfrac{\mathrm{S}_\kappa(r)\sin\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\sin\phi]^2}}$

$$U^d = U^d(r,\phi,\kappa;a_1,C_1) = k_0 + k_1 U_1^d,$$

$$U_1^d = \frac{1}{\mathrm{T}_\kappa(r)},$$

and the two κ-dependent constants of motion, I_2^d and I_3^d, take the form

$$I_2^d = I_{22}(a_1,\kappa) - k_1 \sin\phi,$$

$$I_3^d = I_{22}(c_1,\kappa) + k_1 \cos\phi.$$

f. Family U^e. The general κ-dependent form of the potential U^e is given by

$$U^e = U^e(r,\phi,\kappa;a_0,B_0) = k_0 + k_1 U_1^e + k_2 U_2^e + k_3 U_3^e,$$

$$U_1^e = \mathrm{T}_\kappa^2(r), \quad U_2^e = \frac{\mathrm{T}_\kappa(r)\cos\phi}{\mathrm{C}_k(r)\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\sin\phi]^2}}, \quad U_3^e = \frac{\mathrm{S}_\kappa(r)\sin\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\sin\phi]^2}},$$

and the two κ-dependent constants of motion, I_2^e and I_3^e, take the form

$$I_2^e = I_{22}(a_0,\kappa) + 2k_1(\mathrm{T}_\kappa(r)\cos\phi)^2 + 2k_2\left(\frac{\mathrm{S}_\kappa(r)\cos\phi}{\mathrm{C}_k^2(r)}\right)\sqrt{1-\kappa(\mathrm{S}_\kappa(r)\sin\phi)^2},$$

$$I_3^e = I_{22}(b_0,\kappa) + 2k_1\mathrm{T}_\kappa^2(r)\sin\phi\cos\phi + k_2\mathrm{S}_\kappa(r)\sin\phi\left(\frac{1+2\kappa\mathrm{T}_\kappa^2(r)\cos^2\phi}{\sqrt{1-\kappa(\mathrm{S}_\kappa(r)\sin\phi)^2}}\right)$$
$$+\frac{k_3\mathrm{S}_\kappa(r)\cos\phi}{\sqrt{1-\kappa(\mathrm{S}_\kappa(r)\sin\phi)^2}}.$$

We close this subsection with Table IV that summarizes the expressions obtained for the κ-dependent ''basic'' potentials, U^r, $r = a, aa, b, c, d, e$.

B. Parallel coordinates

When looking for superintegrable potentials in the Euclidean plane, Cartesian coordinates allow a simpler approach than polar ones. On the sphere or hyperbolic plane the situation is reversed. However, we can make an independent check to the results obtained so far by discussing the problem on the sphere in terms of the parallel coordinates (u,y) introduced in Sec. II. The Lagrangian is

$$L=(\tfrac{1}{2})(\mathrm{C}_\kappa^2(y)v_u^2+v_y^2)-U(u,y).$$

Denoting now $I_{22}=av_u^2+2bv_uv_y+cv_y^2$, assume as in previous sections that $I=I_{22}+I_{20}(u,y)$ is a constant of motion. Then a, b, c, and I_{20}, must satisfy

$$a_u-\kappa(2b\mathrm{S}_\kappa(y)\mathrm{C}_\kappa(y))=0,$$

$$a_y+2b_u+\kappa(4a\mathrm{T}_\kappa(y)-2c\mathrm{S}_\kappa(y)\mathrm{C}_\kappa(y))=0,$$

$$c_u+2b_y+\kappa(4b\mathrm{T}_\kappa(y))=0,$$

$$c_y=0,$$

and

$$I_{20u}=\frac{2aU_u}{\mathrm{C}_\kappa^2(y)}+2bU_y\,,\quad I_{20y}=\frac{2bU_u}{\mathrm{C}_\kappa^2(y)}+2cU_y\,.$$

The compatibility condition of the two last equations leads to the following differential equation for the potential:

$$(a-\mathrm{C}_\kappa^2(y)c)U_{uy}+b(\mathrm{C}_\kappa^2(y)U_{yy}-U_{uu})+(a_y-b_u+2\kappa a\mathrm{T}_\kappa(y))U_u+\mathrm{C}_\kappa^2(y)(b_y-c_u)U_y=0.$$

Again the terms involving the first derivatives of the potential U can be written in different alternative ways. When $\kappa=0$, all these equations reduce to the expressions obtained in Sec. III for $Q=E^2$ with Cartesian coordinates (x,y). When comparing with the approach of Sec. IV using polar coordinates on the sphere or the hyperbolic plane, we must remark the *explicit* appearance of the curvature κ; in polar coordinates κ only appeared in the ‘‘tagged’’ trigonometric functions.

The first set of four equations can be solved for $a=a(u,y)$, $b=b(u,y)$, and $c=c(u,y)$. We have obtained

$$a=\alpha_0\mathrm{C}_\kappa^2(y)+\kappa\alpha_1\mathrm{C}_\kappa^2(y)\mathrm{S}_\kappa(y)+\kappa^2c_0(\mathrm{S}_\kappa(u)\mathrm{C}_\kappa(y)\mathrm{S}_\kappa(y))^2,$$

$$b=(\tfrac{1}{2})[\beta_0\mathrm{C}_\kappa(y)+\kappa\beta_1\mathrm{C}_\kappa(u)\mathrm{S}_\kappa(u)\mathrm{S}_\kappa(y)],$$

$$c=c_0\mathrm{C}_\kappa^2(u)+c_1\mathrm{C}_\kappa(u)\mathrm{S}_\kappa(u)+a_2\mathrm{S}_\kappa^2(u),$$

where α_0, α_1, β_0, and β_1, are given by

$$\alpha_0=a_0\mathrm{C}_\kappa^2(y)+a_1\mathrm{C}_\kappa(u)\mathrm{C}_\kappa(y)\mathrm{S}_\kappa(y)+a_2\mathrm{C}_\kappa^2(u)\mathrm{S}_\kappa^2(y),$$

$$\alpha_1=b_0\mathrm{S}_\kappa(u)\mathrm{C}_\kappa(y)-c_1\mathrm{C}_\kappa(u)\mathrm{S}_\kappa(u)\mathrm{S}_\kappa(y),$$

$$\beta_0=b_0\mathrm{C}_\kappa(u)\mathrm{C}_\kappa(y)-a_1\mathrm{S}_\kappa(u)\mathrm{C}_\kappa(y)-c_1\mathrm{C}_\kappa^2(u)\mathrm{S}_\kappa(y)-2a_2\mathrm{C}_\kappa(u)\mathrm{S}_\kappa(u)\mathrm{S}_\kappa(y),$$

$$\beta_1=2c_0\mathrm{C}_\kappa(u)+c_1\mathrm{S}_\kappa(u).$$

TABLE V. Superintegrable potentials with curvature κ: Parallel coordinates.

U^a	$\mathrm{T}_\kappa^2(y)+\dfrac{\mathrm{T}_\kappa^2(2u)}{\mathrm{C}_\kappa^2(y)}$	$\mathrm{T}_\kappa(u)\mathrm{C}_\kappa^2(2u)\mathrm{C}_\kappa^2(y)$	$\dfrac{1}{\mathrm{S}_\kappa^2(y)}$
U^{aa}	$\dfrac{\mathrm{C}_\kappa(y)\mathrm{S}_\kappa(u)}{\sqrt{1-\kappa[\mathrm{C}_\kappa(y)\mathrm{S}_\kappa(u)]^2}}$	$\dfrac{\mathrm{C}_\kappa(u)\mathrm{C}_\kappa(y)}{\mathrm{S}_\kappa^2(y)\sqrt{1-\kappa[\mathrm{C}_\kappa(y)\mathrm{S}_\kappa(u)]^2}}$	$\dfrac{1}{\mathrm{S}_\kappa^2(y)}$
U^b	$\mathrm{T}_\kappa^2(y)+\dfrac{\mathrm{T}_\kappa^2(u)}{\mathrm{C}_\kappa^2(y)}$	$\dfrac{1}{\mathrm{C}_\kappa^2(y)\mathrm{S}_\kappa^2(u)}$	$\dfrac{1}{\mathrm{S}_\kappa^2(y)}$
U^c	$\left(\mathrm{T}_\kappa^2(y)+\dfrac{\mathrm{T}_\kappa^2(u)}{\mathrm{C}_\kappa^2(y)}\right)^{-1/2}$	$\dfrac{\mathrm{C}_\kappa(y)\mathrm{S}_\kappa(u)}{\mathrm{S}_\kappa^2(y)\sqrt{\mathrm{C}_\kappa^2(y)\mathrm{S}_\kappa^2(u)+\mathrm{S}_\kappa^2(y)}}$	$\dfrac{1}{\mathrm{S}_\kappa^2(y)}$
U^d	$\left(\mathrm{T}_\kappa^2(y)+\dfrac{\mathrm{T}_\kappa^2(u)}{\mathrm{C}_\kappa^2(y)}\right)^{-1/2}$		
U^e	$\mathrm{T}_\kappa^2(y)+\dfrac{\mathrm{T}_\kappa^2(u)}{\mathrm{C}_\kappa^2(y)}$	$\dfrac{\mathrm{S}_\kappa(u)}{\mathrm{C}_\kappa^2(u)\mathrm{C}_\kappa^2(y)}$	$\mathrm{T}_\kappa(y)$

Alternatively, these can also be found by transforming the polar expressions by means of the tensor transformation laws. When $\kappa=0$ these reduce to the Euclidean expressions for a,b,c in Cartesian coordinates.

The explicit form of the integrable and superintegrable potentials are obtained by solving the different particular cases of the equation for $U(u,y)$. We simply remark the relation of the auxiliar variables w_a,z_a and w_c,z_c with the parallel coordinates (u,y) and (v,x) in the general case of curvature κ,

$$w_a=\mathrm{S}_\kappa(y),\quad z_a=\mathrm{T}_\kappa(u),\quad w_c=\mathrm{S}_k(x),\quad z_c=\mathrm{T}_\kappa(v).$$

We omit the details and give directly the results obtained in Table V. These potentials coincide with the expressions obtained in Sec. IV A in polar coordinates. This can be checked directly by making use of the geometrical relations given in Sec. II. For most potentials, the expressions in parallel coordinates are much less transparent than in polar ones, though the 2:1 anisotropic oscillator U_1^a and the potential U_3^e are simpler here. We recall that the potentials in the families $\tilde{U}^a$, $\tilde{U}^{aa}$, $\tilde{U}^c$, and $\tilde{U}^e$ are obtained from U^a, U^{aa}, U^c, and U^e simply by reflection in the geodesic $x=y$. This reflection maps the parallel coordinate system (u,y) into the parallel system (v,x); therefore the expression for these ''tilded'' potentials is obtained from their ''untilded'' family in Table V by the replacements $u\leftrightarrow v$, $y\leftrightarrow x$.

V. DISCUSSION OF RESULTS AND COMMENTS

When using the three distances r, x, y, from P to O, l_2, and l_1, and the two distances u, v, between O and the orthogonal projection of P on l_1, and l_2, the families of potentials can be rewritten in the ''simplest'' way as follows:

$$U^a=k_0+k_1\left(\mathrm{T}_\kappa^2(y)+\frac{\mathrm{T}_\kappa^2(2u)}{\mathrm{C}_\kappa^2(y)}\right)+k_2\mathrm{T}_\kappa(u)\mathrm{C}_\kappa^2(2u)\mathrm{C}_\kappa^2(y)+k_3\frac{1}{\mathrm{S}_\kappa^2(y)},$$

$$U^{aa}=k_0+k_1\mathrm{T}_\kappa(x)+k_2\frac{\mathrm{C}_\kappa(v)}{\mathrm{S}_\kappa^2(y)}+k_3\frac{1}{\mathrm{S}_\kappa^2(y)},$$

$$U^b=k_0+k_1\mathrm{T}_\kappa^2(r)+k_2\frac{1}{\mathrm{S}_\kappa^2(x)}+k_3\frac{1}{\mathrm{S}_\kappa^2(y)},$$

$$U^c = k_0 + k_1 \frac{1}{\mathrm{T}_\kappa(r)} + k_2 \frac{\mathrm{S}_\kappa(x)}{\mathrm{S}_\kappa(r)\mathrm{S}_\kappa^2(y)} + k_3 \frac{1}{\mathrm{S}_\kappa^2(y)},$$

$$U^d = k_0 + k_1 \frac{1}{\mathrm{T}_\kappa(r)},$$

$$U^e = k_0 + k_1 \mathrm{T}_\kappa^2(r) + k_2 \frac{\mathrm{S}_\kappa(u)}{\mathrm{C}_\kappa^2(r)} + k_3 \mathrm{T}_\kappa(y).$$

These potentials extend, and include as a particular case, the Euclidean expressions obtained in Sec. III for $\kappa=0$.

The use of the curvature κ as a parameter has lead to an unified approach that highlights some facts which would be more difficult to appreciate with a separate analysis of the three spherical, Euclidean and hyperbolic cases. Contractions are built-in in the procedure, so there is not any additional need to consider separately the study of contraction of spherical or hyperbolic superintegrable potentials to the Euclidean ones. We now make several comments which are grouped in two sets. First, in Sec. A, we give some comments on the potentials we have obtained, and then, in the following subsections, a group of observations of somewhat broader scope.

A. Comments on the potentials obtained

(a) The results we have obtained give several families of potentials which remain superintegrable when the configuration space curvature κ varies and which are not trivial in the flat limit (here simply described as the particular case $\kappa=0$). We are not implying that these families must be considered as a *complete and not redundant* set of families of superintegrable potentials with quadratic constants of motion on 2D spaces of constant curvature (see below).

(b) There are two *central* potentials for any value of the curvature, either on the sphere and on the hyperbolic plane. The potential

$$U_K = -kU_1^c = -kU_1^d = -\frac{k}{\mathrm{T}_\kappa(r)}, \quad k>0$$

is the ''Kepler potential,'' which on the sphere was first studied by Schrödinger in Ref. 25. Likewise, the potential

$$U_{HO} = (\tfrac{1}{2})k^2 U_1^b = (\tfrac{1}{2})k^2 U_1^e = (\tfrac{1}{2})k^2 \mathrm{T}_\kappa^2(r)$$

plays the role of the ''harmonic oscillator'' on either the sphere or the hyperbolic plane. The spherical version was studied by Higgs in Ref. 26 and is known as the Higgs oscillator.

(c) All other potentials found appear to be ''noncentral'' in the sense they depend not only on the distance to the origin point, but also on the angular coordinate. However, a function involving ϕ may turn out depending only on the distance to some given point (not the origin). This is what happens in the case $\kappa>0$ (sphere) with the two potentials, $U_3^a = 1/(\mathrm{S}_\kappa(r)\sin\phi)^2 = 1/\mathrm{S}_\kappa^2(y)$ and $U_3^e = \mathrm{T}_k(y)$. These two particular potentials, that depend only on x or on y, can be considered as ''central'' if this is understood as ''depending on the distance to some fixed point in the sphere,'' not necessarily the coordinate origin. Notice that this property is a consequence of the geometric properties of S^2. Thus it is not true for the Euclidean or the hyperbolic cases.

(d) The spherical and hyperbolic families U^b and U^c, whose potentials are three parametrical superintegrable deformations of the harmonic oscillator and of the ''Kepler problem'' appear, e.g., in Grosche–Pogosyan Refs. 14–16. In the spherical case these are the only previously known families.

(e) An interesting—and as far as we know, new—outcome is the existence of superintegrable spherical and hyperbolic versions of the Euclidean anisotropic 2:1 oscillator $4x^2+y^2$ (or x^2+4y^2). They are the two potentials U_1^a, (or $\tilde{U}_1^a$),

$$U_1^a=\frac{1}{1-\kappa(\mathrm{S}_\kappa(r)\sin\phi)^2}\left[4\left(\frac{\mathrm{T}_\kappa(r)\cos\phi}{1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2}\right)^2+(\mathrm{S}_\kappa(r)\sin\phi)^2\right]=\mathrm{T}_\kappa^2(y)+\frac{\mathrm{T}_\kappa^2(2u)}{\mathrm{C}_\kappa^2(y)},$$

$$\tilde{U}_1^a=\frac{1}{1-\kappa(\mathrm{S}_\kappa(r)\cos\phi)^2}\left[(\mathrm{S}_\kappa(r)\cos\phi)^2+4\left(\frac{\mathrm{T}_\kappa(r)\sin\phi}{1-\kappa(\mathrm{T}_\kappa(r)\sin\phi)^2}\right)^2\right]=\mathrm{T}_\kappa^2(x)+\frac{\mathrm{T}_\kappa^2(2v)}{\mathrm{C}_\kappa^2(x)}.$$

These two functions both have the correct Euclidean limit, i.e., $\lim_{\kappa\to 0} U_1^a=r^2(4\cos^2\phi+\sin^2\phi)=4x^2+y^2$ and $\lim_{\kappa\to 0}\tilde{U}_1^a=r^2(\cos^2\phi+4\sin^2\phi)=x^2+4y^2$. However, their structure in polar coordinates is complicated, and do not reduce to a radial harmonic oscillator with a purely angular modulation as it is in the flat case. If the Higgs isotropic oscillator $U_1^b=\mathrm{T}_\kappa^2(r)$ is rewritten as $U_1^b=[\mathrm{T}_\kappa^2(r)/Z_1]Z_2$ with $Z_1=1-\kappa(\mathrm{S}_\kappa(r)\sin\phi)^2$, $Z_2=\cos^2\phi+\mathrm{C}_\kappa^2(r)\sin^2\phi$, then the changes required to obtain the superintegrable 2:1 anisotropic oscillator are, not only the numerical coefficient 4 (as is the case in E^2), but also the κ-dependent function $1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2$ in the denominator of one of the two summands of Z_2. It is important to notice that the same process is easier in parallel coordinates; starting from the Higgs oscillator $U_1^b=\mathrm{T}_\kappa^2(r)=\mathrm{T}_\kappa^2(u)/\mathrm{C}_\kappa^2(y)+\mathrm{T}_\kappa^2(y)$, the replacement of u by $2u$ as the argument of the function $\mathrm{T}_\kappa(u)$ gives directly the 2:1 anisotropic oscillator.

(f) The family U^a in the curved case is new. The two families U^b,U^c were already known in either S^2, E^2, H^2, and to the family U^d is rather trivial in the curved case, as it contains only the Kepler potential. Further to these families, we also obtain two additional families of superintegrable potentials, U^{aa} and U^e, which both in the sphere or in the hyperbolic plane are also new. The family U^e includes the Higgs harmonic oscillator and two new superintegrable spherical and hyperbolic potentials whose flat limits are the two linear potentials $r\cos\phi=x$ and $r\sin\phi=y$. The planar family V^a admits two different curved deformations, $U^a(a_0,C_1)$ and $U^{aa}(c_0,C_1)$. Notice that the Euclidean limits of U^a and U^{aa} are given by

$$\lim_{\kappa\to 0} U^a=k_1V_1^a+k_2V_2^a+k_3V_3^a,$$

$$\lim_{\kappa\to 0} U^{aa}=k_1V_2^a+k_2V_3^a+k_3V_3^a,$$

so the potential U^a must be considered as the appropriate ''curved version'' of V^a. Concerning U^{aa} it must be considered, not as a ''curved version'' of V^a taken as a whole, but only of the subspace generated by the the two potentials $V_2^a=x$ and $V_3^a=1/y^2$ (the 2:1 oscillator V_1^a is not present).

(g) In Sec. III we have seen that the ''Stark'' potential $V(x,y)=x$ is superintegrable in E^2 and is linearly compatible with both the nonisotropic oscillator $V_1^a=4x^2+y^2$ and the isotropic one $V_1^e=x^2+y^2$. It is interesting that this potential admits *three* different superintegrable spherical or hyperbolic versions given by

$$U_2^a=\frac{[\mathrm{T}_\kappa(r)/\mathrm{C}_k^2(r)]\cos\phi}{[1-\kappa(\mathrm{T}_\kappa(r)\cos\phi)^2]^2}=\mathrm{T}_\kappa(u)\mathrm{C}_\kappa(2u)\mathrm{C}_\kappa^2(y)\qquad \lim_{\kappa\to 0} U_2^a=r\cos\phi=x,$$

$$U_2^e=\frac{[\mathrm{T}_\kappa(r)/\mathrm{C}_k(r)]\cos\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\sin\phi]^2}}=\frac{\mathrm{S}_\kappa(u)}{\mathrm{C}_\kappa^2(u)\mathrm{C}_\kappa^2(y)}\qquad \lim_{\kappa\to 0} U_2^e=r\cos\phi=x,$$

$$\tilde{U}_3^e=\frac{\mathrm{S}_\kappa(r)\cos\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\cos\phi]^2}}=\mathrm{T}_\kappa(x)\qquad \lim_{\kappa\to 0}\tilde{U}_2^e=r\cos\phi=x.$$

Each of these three constant curvature versions of the linear potential is linearly compatible with one of the two oscillators; U_2^a with the 2:1 nonisotropic oscillator, and U_2^e and $\tilde{U}_3^e$ with the isotropic oscillator.

(h) For any fixed nonconstant potential V in a configuration space of a constant curvature κ, the maximal dimension for the associated parameter subspace appears to be $m=4$, independently of κ. The following four potentials reach the maximal dimension:

$$U_1^b = U_1^e = \mathrm{T}_k^2(r) = \mathrm{T}_\kappa^2(y) + \frac{\mathrm{T}_\kappa^2(2u)}{\mathrm{C}_\kappa^2(y)} \quad (a_0, b_0, c_0; 0,0; a_2),$$

$$U_3^e = \widetilde{U}_1^{aa} = \frac{\mathrm{S}_\kappa(r)\sin\phi}{\sqrt{1-\kappa[\mathrm{S}_\kappa(r)\sin\phi]^2}} = \mathrm{T}_k(y) \quad (a_0, b_0, e_0; a_1, 0; \kappa e_0),$$

$$U_3^a = U_3^b = U_3^c = \frac{1}{\mathrm{S}_\kappa^2(r)\sin^2\phi} = \frac{1}{\mathrm{S}_\kappa^2(y)} \quad (a_0, 0, c_0; 0, c_1; a_2),$$

$$U_1^c = U_1^d = \frac{1}{\mathrm{T}_k(r)} = \left(\mathrm{T}_\kappa^2(y) + \frac{\mathrm{T}_\kappa^2(u)}{\mathrm{C}_\kappa^2(y)} \right)^{-1/2} \quad (e_0, 0, e_0; a_1, c_1; a_2).$$

This means that *all* flat superintegrable potentials with the maximal dimension of the associated parameter subspace allow a ''curvature'' version. This remark also accounts for the special role played by these four curvature versions of the flat harmonic oscillator, Kepler potential, linear potential, and centrifugal barrier. The curvature version of the potential $1/y^2$, i.e., $U_3^a = U_3^b = U_3^c = 1/(\mathrm{S}_\kappa(r)\sin\phi)^2 = 1/\mathrm{S}_\kappa^2(y)$, is a very remarkable system endowed with a level of superintegrability compatible with the constant curvature versions of either the oscillator, Kepler problem, and also the anisotropic 2:1 oscillator.

After these particular comments, we now discuss some more general questions.

B. Equivalence

A complete *classification* of superintegrable potentials with constants of motion which are quadratic in the velocities, in either the sphere or the hyperbolic plane would require to introduce some equivalence criteria. We do not intend to do this here, but some remarks may be relevant. From the present point of view a most natural idea is to look a superintegrable family in the space of constant curvature κ, associated with a pair of sets of parameter values (a_i, b_i, c_i) and (A_i, B_i, C_i) as *equivalent* to another one, with parameters (a_i', b_i', c_i') and (A_i', B_i', C_i') if the equation for (a_i, b_i, c_i) (resp. (A_i, B_i, C_i)) is transformed into the equation for (a_i', b_i', c_i') (resp. (A_i', B_i', C_i')) under the regular action of the isometry group of the configuration space. As a consequence, the general potential in the first family is transformed into the general potential in the second family. These transformations preserve the kinetic term, and clearly two families of potentials which can be transformed among themselves by a rigid motion of the configuration space should be considered as essentially equivalent. This has been implicitly done when denoting with a tilde some families which are equivalent in this sense to those without a tilde; the transformation of the configuration space realizing the equivalence is the reflection in the line $x = y$.

Forgetting about the tilded families which are equivalent to the untilded ones, we are left with the families U^a, U^{aa}, U^b, U^c, U^d, U^e, which are all different, and the question of their possible equivalence should be addressed. The classification will produce different results in the three cases $\kappa > 0$, $\kappa = 0$, $\kappa < 0$, and the number of equivalence classes of families should be expected to be largest in the hyperbolic plane $\kappa < 0$. Grosche and Pogosyan give our hyperbolic families U^b, U^c, and three hyperbolic families we have not discussed here, even though it is easy to check that their potentials satisfy the basic equation for suitable choices of pairs of parameter vectors (these are called V_3, V_4, V_5, in Refs. 15, 16). The hyperbolic families U^a, U^{aa}, U^e, do not appear in Refs. 15 and 16. All these hyperbolic families seem to be inequivalent, and there might be still more inequivalent superintegrable families of potentials on the hyperbolic plane.

However, in the spherical $\kappa > 0$ case, the three families U^{aa}, U^c, and U^e are equivalent. The transformation of the configuration space which permutes cyclically the three families, i.e., $U^{aa} \to U^e \to \widetilde{U}^c \to U^{aa}$ ($\widetilde{U}^{aa} \to \widetilde{U}^e \to U^c \to \widetilde{U}^{aa}$), is the spherical rotation which permutes cyclically the three mutually orthogonal coordinate axes in the ambient space (around the center of the positive octant, with an angle of $2\pi/3$). Therefore, as far as equivalence classes, only the four families U^a, U^b, U^c, U^d, must be considered. We strongly believe that these four families are inequivalent. If we restrict ourselves to the particular case where each of the two extra constants of motion is associated to a parameter vector which has a *single* nonzero component in the natural basis $(a_0,b_0,c_0,\ldots)$, like the ones in Table III, then we can state that the four families of potentials we have given *exhaust* this type of superintegrable potentials in the sphere. As the reader can check, the choices (b_0,a_2), (b_0,a_1), and (c_0,a_1) which have been not discussed in the paper, are equivalent on the sphere to some of those discussed (respectively, to U^a, U^d, and U^d).

C. Superintegrable potentials with constant flat limit

In addition to the constant potential $U_0 \equiv 1$, which trivially belongs to all the families of superintegrable potentials, the families U^b, U^e, and $\tilde{U}^e$, contain one nonconstant superintegrable deformation of the trivial planar potential $V_0=1$, and the families U^a, and $\tilde{U}^a$, contain two different (hence a one parameter infinite family) of such nonconstant superintegrable deformations of $V_0=1$. They have the following expressions:

$$U_0^b = U_0^e = \tilde{U}_0^e = \frac{1}{\mathrm{C}_\kappa^2(r)},$$

$$U_{0(\pm)}^a = \frac{1}{[\mathrm{C}_\kappa(r) \pm \sqrt{\kappa}\mathrm{S}_\kappa(r)\cos\phi]^2} = \frac{1}{[\mathrm{C}_\kappa(r) \pm \sqrt{\kappa}\mathrm{S}_\kappa(x)]^2},$$

$$\tilde{U}_{0(\pm)}^a = \frac{1}{[\mathrm{C}_\kappa(r) \pm \sqrt{\kappa}\mathrm{S}_\kappa(r)\sin\phi]^2} = \frac{1}{[\mathrm{C}_\kappa(r) \pm \sqrt{\kappa}\mathrm{S}_\kappa(y)]^2}.$$

These functions represent superintegrable potentials with quadratic integrals (one of them of the Noether class). In the limit $\kappa \to 0$, all of them go into the trivial constant potential,

$$\lim_{\kappa\to 0} U_{0(\pm)}^a = \lim_{\kappa\to 0} \tilde{U}_{0(\pm)}^a = \lim_{\kappa\to 0} U_0^b = U_0 \equiv 1,$$

so each of these potentials can be considered as a nontrivial superintegrable ''curvature version'' of the constant planar potential (notice that in the hyperbolic case $U_{0(\pm)}^a$ and $\tilde{U}_{0(\pm)}^a$ are complex). This property (existence of nontrivial superintegrable dynamics in $Q=S^2$ or $Q=H^2$ reducing to the free motion when $\kappa \to 0$) points out the great level of complexity of the dynamics in configuration spaces of constant curvature.

Each of the potentials $U_{0(\pm)}^a$, $\tilde{U}_{0(\pm)}^a$, and U_0^b can be expressed in terms of the constant potential U_0 and the three ''basic'' solutions U_1, U_2, U_3, within each family. Alternatively, by inverting the former expressions, some of the ''basic solutions'' U_1, U_2, U_3, within each family can be expressed as a ''linear superposition'' of these potentials. In particular, the isotropic and 2:1 anisotropic oscillators can be rewritten in the following form:

$$U_1^b = \mathrm{T}_k^2(r) = \left(\frac{1}{\kappa}\right)(U_0^b - U_0),$$

$$U_1^a = \mathrm{T}_\kappa^2(y) + \frac{\mathrm{T}_\kappa^2(2u)}{\mathrm{C}_\kappa^2(y)} = \left(\frac{1}{\kappa}\right)(U_0^a - U_0),$$

where we have used the notation

$$U_0^a = (\tfrac{1}{2})[U_{0(+)}^a + U_{0(-)}^a]$$

(a similar relation is obtained for $\tilde{U}_1^a$ and $\tilde{U}_{0(\pm)}^a$). Thus, the isotropic and the anisotropic 2:1 oscillator potentials in a curved configuration space are proportional to the difference between two superintegrable potentials in the same family and with a common Euclidean limit.

The right-hand side of the above two expressions is well defined even for $\kappa \to 0$. This can be checked directly in these particular examples, but it is indeed a *general* property. If two superintegrable potentials U_1, U_2 in a configuration space of constant curvature κ are in the same family of superintegrable potentials (characterized by a given pair of sets of parameter vectors), and have the same Euclidean limit, $\lim_{\kappa\to 0} U_1 = \lim_{\kappa\to 0} U_2$, then we can write an expansion for the difference $U_1 - U_2$ in powers of κ, which starts in the linear term

$$U_1 - U_2 = \kappa\left(\left[\frac{dU_1}{d\kappa}\right]_{\kappa=0} - \left[\frac{dU_2}{d\kappa}\right]_{\kappa=0}\right) + O(\kappa^2).$$

Hence the function V_{12} given by

$$V_{12} = \lim_{k\to 0}\left[\left(\frac{1}{\kappa}\right)(U_1 - U_2)\right] = \left[\frac{dU_1}{d\kappa}\right]_{\kappa=0} - \left[\frac{dU_2}{d\kappa}\right]_{\kappa=0}$$

is well defined and superintegrable in E^2, with the same pair of parameter vectors as U_1, U_2. For the two families, U^b and U^a, the Euclidean potential V_{12} represent the isotropic oscillator in the first case ($U_1 = U_0^b$, $U_2 = U_0$) and the anisotropic 2:1 in the second case ($U_1 = U_0^a$, $U_2 = U_0$). So, in a sense, these two well-known superintegrable Euclidean oscillators can be considered as a kind of ''residue'' in E^2 of the existence of different superintegrable potentials in configuration spaces of constant curvature within the same family and with the same Euclidean limit.

D. ''Correspondence'' between potentials on constant nonzero and zero curvature spaces

Our procedure gives directly a correspondence between *superintegrable* potentials in the nonzero curvature spaces and in the flat Euclidean space. All superintegrable flat potentials in the families V^a, V^b, V^c, V^e, admit a superintegrable ''curvature'' version on either the sphere and the hyperbolic plane. This property fails for the flat family V^d, where we have not found any curvature version of the superintegrable potentials V_2^d, V_3^d. If such curvature version exists, it should necessarily be associated with *mixed* parameter vectors, which should exhibit an explicit dependence of the curvature κ similar to the one found in the energy.

A somewhat surprising result is that a flat superintegrable potential may have several *different* ''curvature versions'' which still are *superintegrable*. For the constant potential $V_0 = 1$ this has been already discussed in the previous paragraph. But once at least two (and then a one-parameter infinity of) different superintegrable versions of the constant potential exists in a given superintegrable family, each of their members admits also a one-parameter infinity of different ''curvature versions'' of the same flat potential, all of which belong to the same family. If we drop the requirement of being in the same family, there are even more examples of *different* curvature versions of the same flat potential. For instance, U_2^a, U_2^e, and $\tilde{U}_3^e$, are all superintegrable curved versions of the linear flat potential $V = x$.

Given a potential $V = V(r,\phi)$ in E^2 we can construct many different potentials in the space of constant curvature κ ($Q = S^2$ or $Q = H^2$), $U = U(r,\phi;\kappa)$ with V as a Euclidean limit, i.e., $\lim_{\kappa\to 0} U = V$. Additional requirements may successively narrow the choice; for instance, if V is integrable (superintegrable) in E^2, we may require U to be integrable (superintegrable) in S^2 or H^2 and so on. As an example, let us consider the general anisotropic planar oscillator, $V = Ax^2 + By^2$. Consider the two following spherical or hyperbolic potentials:

$$\frac{A\mathrm{T}_\kappa^2(r)\cos^2\phi + B\mathrm{S}_\kappa^2(r)\sin^2\phi}{1 - \kappa(\mathrm{S}_\kappa(r)\sin\phi)^2} = A\frac{\mathrm{T}_\kappa^2(u)}{\mathrm{C}_\kappa^2(y)} + B\mathrm{T}_\kappa^2(y),$$

$$\frac{A\mathrm{S}_\kappa^2(r)\cos^2\phi+B\mathrm{T}_\kappa^2(r)\sin^2\phi}{1-\kappa(\mathrm{S}_\kappa(r)\cos\phi)^2}=A\mathrm{T}_\kappa^2(x)+B\frac{\mathrm{T}_\kappa^2(v)}{\mathrm{C}_\kappa^2(x)}.$$

Both are integrable, with constant of motion associated with a_0 and c_0, respectively, and they are apparently reasonable spherical/hyperbolic versions of the general anisotropic planar oscillator V in the sense that both satisfy the correct Euclidean limit, i.e., Ax^2+By^2, and they reduce to the spherical/hyperbolic harmonic oscillator $\mathrm{T}_\kappa^2(r)$ when $B=A$. Nevertheless if we look for *superintegrable* spherical/hyperbolic versions of the general anisotropic planar oscillator, we must reject these candidates in S^2 and H^2 since they are integrable but not superintegrable, and the orbits will be open curves. The appropriate superintegrable spherical/hyperbolic versions of the oscillators $(A=4,\ B=1)$ and $(A=1,\ B=4)$ are the potentials U_1^a and $\tilde{U}_1^a$. As far as we know, whether or not there exists superintegrable $\kappa\neq0$ versions of the special known cases of superintegrable anisotropic planar potentials (with nonquadratic constants of motion) is an open problem.

E. Coordinate systems and superintegrability

The approach we have presented here can be compared with the search for superintegrable potentials by requiring separation in at least two of those coordinate systems which allow complete separation of variables for the Laplace–Beltrami equation. This requirement for the potential turns out to be equivalent to satisfying the differential equation for the potential U for some choice of the parameter vector. Therefore, the nontrivial parameter vectors are in correspondence with the possible coordinate systems of the required type. In the simplest cases, this is clear from the expressions we have given; a potential which satisfies the a_2 (resp. a_0, c_0) equation separates in the polar (r,ϕ) coordinates (resp. in the parallel system (u,y), (x,v)). If one starts from the flat case, several coordinate systems which are *different* whenever $\kappa\neq0$ (actually, all those associated to the parameter vectors $\{a_0,0,c_0;00;0\}$) coalesce into a single very ''degenerate'' Cartesian system. Therefore the analysis of superintegrability in terms of separability in several coordinate systems should be better done in the generic $\kappa\neq0$ case, where the coordinate systems which coalesce in the $\kappa=0$ case are still different.

ACKNOWLEDGMENTS

We thank J. F. Cariñena for comments on the geometric approach to the theory of superintegrable systems. This article is part of a project supported by Grants Nos. PB-94-115, PB-96-0717, and AEN-97-1680 (DGICYT, Madrid).

[1] M. Henon, ''Numerical exploration of Hamiltonian systems'' in *Chaotic Behaviour of Deterministic Systems*, edited by G. Iooss, R. H. Helleman, and R. Stora (North-Holland, Amsterdam, 1983).
[2] T. I. Fris, V. Mandrosov, Y. A. Smorodinsky, M. Uhlir, and P. Winternitz, Phys. Lett. **16**, 354 (1965).
[3] A. A. Makarov, Y. A. Smorodinsky, K. Valiev, and P. Winternitz, Nuovo Cimento **52**, 1061 (1967).
[4] S. Wojciechowski, Phys. Lett. A **95**, 279 (1983).
[5] L. Chetouani, L. Guechi, and T. F. Hammann, Phys. Lett. A **125**, 277 (1987).
[6] M. V. Carpio-Bernido and C. C. Bernido, Phys. Lett. A **134**, 395 (1987).
[7] J. Hietarinta, Phys. Rep. **147**, 87 (1987).
[8] M. Kibler, and P. Winternitz, Phys. Lett. A **147**, 338 (1990).
[9] N. W. Evans, Phys. Rev. A **41**, 5666 (1990).
[10] N. W. Evans, Phys. Lett. A **147**, 483 (1990).
[11] N. W. Evans, J. Math. Phys. **32**, 3369 (1991).
[12] M. A. del Olmo, M. A. Rodríguez, and P. Winternitz, J. Math. Phys. **34**, 5118 (1993).
[13] J. A. Calzada, M. A. del Olmo, and M. A. Rodríguez, J. Geom. Phys. **23**, 14 (1997).
[14] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, Fortschr. Phys. **43**, 453 (1995).
[15] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, Fortschr. Phys. **43**, 523 (1995).
[16] C. Grosche, G. S. Pogosyan, and A. N. Sissakian, Phys. Part. Nuclei **27**, 244 (1996).
[17] E. G. Kalnins, W. Miller, Jr., and G. S. Pogosyan, J. Math. Phys. **37**, 6439 (1996).
[18] E. G. Kalnins, W. Miller, Jr., and G. S. Pogosyan, J. Math. Phys. **37**, 5416 (1997).
[19] A. L. Salas-Brito, H. N. Núñez-Yépez, and R. P. Martínez-Romero, Int. J. Mod. Phys. **12**, 271 (1997).
[20] M. F. Rañada, J. Math. Phys. **38**, 4165 (1997).
[21] C. Gonera, J. Phys. A **31**, 4465 (1998).

[22] M. F. Rañada, J. Math. Phys. **40**, 236 (1999).
[23] C. López, E. Martínez, and M. F. Rañada, ''Dynamical symmetries, non-Cartan symmetries, and superintegrability of the *n*-dimensional harmonic oscillator,'' J. Phys. A **32**, 1241 (1999).
[24] D. M. Fradkin, Am. J. Phys. **33**, 207 (1965).
[25] E. Schroedinger, Proc. R. Ir. Acad. Sect. A, Math. Astron. Phys. Sci. **A XLVI**, 9 (1940).
[26] P. W. Higgs, J. Phys. A **12**, 309 (1979).
[27] H. I. Leemon, J. Phys. A **12**, 489 (1979).
[28] W. Klinberger, *A Course in Differential Geometry* (Springer, Graduate texts in Mathematics, New York, 1978).
[29] A. M. Perelomov, *Integrable Systems of Classical Mechanics and Lie Algebras* (Birkhauser, Basel, 1990).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Degenerate Frobenius manifolds and the bi-Hamiltonian structure of rational Lax equations

I. A. B. Strachan[a)]
Department of Mathematics, University of Hull, Hull, HU6 7RX, England

(Received 25 March 1999; accepted for publication 20 May 1999)

The bi-Hamiltonian structure of certain multicomponent integrable systems, generalizations of the dispersionless Toda hierarchy, is studied for systems derived from a rational Lax function. One consequence of having a rational rather than a polynomial Lax function is that the corresponding bi-Hamiltonian structures are degenerate, i.e., the metric that defines the Hamiltonian structure has a vanishing determinant. Frobenius manifolds provide a natural setting in which to study the bi-Hamiltonian structure of certain classes of hydrodynamic systems. Some ideas on how this structure may be extended to include degenerate bi-Hamiltonian structures, such as those given in the first part of the paper, is given. © *1999 American Institute of Physics.* [S0022-2488(99)02210-0]

I. INTRODUCTION

Poisson brackets of a hydrodynamic type were introduced by Dubrovin and Novikov in Ref. 1, where they gave a complete description of Poisson brackets of the form

$$\{u^i(x),u^j(y)\}=g^{ij}[u(x)]\,\delta'(x-y)+\Gamma^{ij}_k[u(x)]u^k(x)\,\delta(x-y), \tag{1}$$

under the nondegenerate condition $\det(g^{ij})\neq 0$. This defines a skew-symmetric Poisson bracket on functionals,

$$\{I,J\}=\int dx\,\frac{\delta I}{\delta u^i(x)}\widehat{A^{ij}}\frac{\delta J}{\delta u^j(x)},$$

where

$$\widehat{A^{ij}}=g^{ij}[u(x)]\,\frac{d}{dx}+\Gamma^{ij}_k[u(x)]u^k_x(x).$$

The conditions on g^{ij} and Γ^{ij}_k necessary in order for (1) to define a Hamiltonian structure, under the nondegenerate condition $\det(g^{ij})\neq 0$, have a natural geometric interpretation.[1]

Theorem 1: *Under the nondegenerate condition* $\det(g^{ij})\neq 0$, *the bracket* (1) *defines a Hamiltonian structure if and only if (a)* $\mathbf{g}=(g^{ij})^{-1}$ *defines a (pseudo-) Riemannian metric; (b)* $\Gamma^{ij}_k=-g^{is}\Gamma^j_{sk}$, *where* Γ^j_{sk} *are the Christoffel symbols of the Riemannian connection defined by* $\mathbf{g}$; *(c) the Riemann curvature tensor of* $\mathbf{g}$ *vanishes*.

This result, and its interpretation in terms of differential geometry, rests on the nondegeneracy condition on the metric. However, this is not a necessary condition for (1) to define a Hamiltonian structure and the full result, with no *a priori* restriction on g^{ij} was derived by Grinberg[2] and Dorfmann.[3]

Theorem 2: *The bracket* (1) *defines a Hamiltonian structure if and only if the pair* (g,Γ) *satisfy the conditions*

[a)]Electronic mail: i.a.b.strachan@hull.ac.uk

$$g^{ij} = g^{ji}; \tag{2}$$

$$\frac{\partial g^{ij}}{\partial u^k} = \Gamma_k^{ij} + \Gamma_k^{ji}; \tag{3}$$

$$g^{ij}\Gamma_i^{rs} = g^{ri}\Gamma_i^{js}; \tag{4}$$

$$\Gamma_t^{ij}\Gamma_r^{tk} - \Gamma_t^{ik}\Gamma_r^{tj} = g^{ti}\left(\frac{\partial\Gamma_r^{jk}}{\partial u^t} - \frac{\partial\Gamma_t^{jk}}{\partial u^r}\right), \tag{5}$$

and

$$\sum_{\text{cyclic sum on } i,j,k}\left[\left(\frac{\partial\Gamma_t^{ij}}{\partial u^q} - \frac{\Gamma_q^{ij}}{\partial u^t}\right)\Gamma_r^{tk} + \left(\frac{\partial\Gamma_t^{ij}}{\partial u^r} - \frac{\Gamma_r^{ij}}{\partial u^t}\right)\Gamma_q^{tk}\right] = 0. \tag{6}$$

If $\det g^{ij} \neq 0$ *then the last equation is a consequence of the earlier equations.*
[*N.B.:* there is a minor error in Ref. 2 in the order of the indices in Eq. (4)]. In this more general situation it is not possible to give a clear geometric interpretation of these equations. They define an integrable distribution, but their differential geometric content is less clear. One can define a covariant derivative-like object,

$$\nabla^i \xi^j = \partial^i \xi^j - \Gamma_k^{ij}\xi^k,$$

where $\partial^i = g^{ij}\partial_j$, with the property (when suitably extended to tensors) that $\nabla^i g^{jk} = 0$, though the ''connection'' cannot be defined in terms of the ''metric.'' With such a covariant derivative one can introduce a ''curvature'' by the equation

$$(\nabla^r\nabla^s - \nabla^s\nabla^r)\xi^t = -R_k^{rst}\xi^k,$$

and the third equation above is now just the vanishing of this curvature. Such a description is not very natural; one cannot lower indices and the interpretation of the last equation remains unclear. However, the terms ''metric'' and ''connection'' will be used to denote these objects, and a pair satisfying these equations will be called a (g,Γ) pair.

Our purpose in this paper is to study the bi-Hamiltonian structure of dispersionless integrable systems defined by the Lax equation (the variables τ_n will be used to denote the times, t being reserved for flat coordinates in which the components η^{ij} are constants),

$$\frac{\partial\mathcal{L}}{\partial\tau_n} = \{(\mathcal{L}^{n/(N-M)})_+, \mathcal{L}\}_{PB} \tag{7}$$

where $\{f,g\}_{PB} = p(\partial_p f\,\partial_x g - \partial_x f\,\partial_p g)$, $\mathcal{L}$ is given by a rational function,

$$\mathcal{L} = \frac{\text{polynomial of degree } N}{\text{polynomial of degree } M}$$

with the single constraint $N > M$, and $(\quad)_+$ denotes the projection onto non-negative powers of p under a formal expansion in powers of p. In an earlier paper[4] this system was studied but a complete description of the Hamiltonian structure was not given. The simplest example of such a system is the continuum Toda equations

$$S_\tau = P_x,$$

$$P_\tau = PS_x, \tag{8}$$

which is generated for the above Lax equation (7) with a Lax function

$$\mathcal{L}=p^2+S(x,t)+\frac{P(x,t)}{p}.$$

This paper aims to extend these earlier results from polynomial Lax functions to rational Lax functions and to relate these results to the theory of Frobenius manifolds.[5] It will turn out that in the rational case the Hamiltonian structure is degenerate, so the more general description of Grinberg and Dorfmann will have to be utilized to give a complete description of the bi-Hamiltonian structure of the hierarchy. This in turn implies that a new concept of a degenerate Frobenius manifold is required.

In the next section a summary of the pertinent result of Ref. 4 will be given, and this will also serve to fix the notation used. Full details will not be given and the reader should consult the earlier paper for the proofs. In Sec. III the polynomial case will be studied in more detail (and this will relate the results of Ref. 4 to more recent work of Dubrovin and Zhang[6]) before the full rational case is studied in Sec. IV. The properties of a degenerate Frobenius manifold are introduced by way of an extended example in Sec. V.

Throughout this paper various different coordinate systems will be used, and the resulting transformations from one system to another will be important. The notation $g^{ij}(s)$ will be used to denote the components of the metric in the s^i-coordinate system, so the transformation from s^i to t^i coordinates will be written

$$g^{ij}(t)=\frac{\partial t^i}{\partial s^p}\frac{\partial t^j}{\partial s^q}g^{pq}(s),$$

rather than using different fonts and alphabets for the different coordinate systems.

II. CONSERVATION LAWS AND EVOLUTION EQUATIONS

In order to study rational functions it is convenient, and indeed necessary in order to obtain some results, to factorize the numerator and denominator of the rational function, so

$$\mathcal{L}=\frac{\Pi_{i=1}^{N}(p+u^i)}{\Pi_{i=N+1}^{N+M}(p+u^i)},$$

$$=\prod_{i=1}^{N+M}(p+u^i)^{\varepsilon_i}.$$

Here it will be assumed that $\varepsilon_i=\pm 1$ and that the numerator and denominator have no common root. With these conditions and $N>M$ the Lax function is of the general form

$$\mathcal{L}=\text{polynomial of degree }(N-M)+\sum_{i=N+1}^{N+M}\text{ simple poles.}$$

Such a factorization of the Lax function was introduced by Kupershmidt[7] (though this could also be viewed as a Vieté transformation) and the variables u^i will be called modified variables. One advantage of such a factorization is that it puts all the fields on an egalitarian footing, i.e., the permutation group S^N acts on the zeros of $\mathcal{L}$ and the permutation group S^M acts on the roots of $\mathcal{L}$, and this drastically reduces the complexity of the calculations.

The flows are given by the Lax equation (7) which may be calculated explicitly

$$u^i_{\tau_n}=A_i^{(n)}u^i_x+\sum_{j\neq i}u^iB_{ij}^{(n)}u^j_x, \tag{9}$$

where

$$A_i^{(n)}=\left(\frac{\varepsilon_i n}{N-M}-1\right)\sum_{\{r_j:\Sigma_{j=1}^{N+M}r_j=n\}}\left[\prod_{\substack{k=1\\k\neq i}}^{N+M}\binom{\frac{\varepsilon_k n}{N-M}}{r_k}(u^k)^{r_k}\right]\binom{\frac{\varepsilon_i n}{N-M}-2}{r_i-1}(u^i)^{r_i}$$

and

$$B_{ij}^{(n)}=\frac{\varepsilon_j n}{N-M}\sum_{\{r_j:\Sigma_{j=1}^{N+M}r_j=n-1\}}\left[\prod_{\substack{k=1\\k\neq i,j}}^{N+M}\binom{\frac{\varepsilon_k n}{N-M}}{r_k}(u^k)^{r_k}\right]\binom{\frac{\varepsilon_i n}{N-M}-1}{r_i}(u^i)^{r_i}\binom{\frac{\varepsilon_j n}{N-M}-1}{r_j}(u^j)^{r_j}.$$

Care has to be taken in evaluating the binomial coefficients for negative and fractional numbers. These must be interpreted in terms of Γ-function, so

$$\binom{a}{b}=\frac{\Gamma(a+1)}{\Gamma(a-b+1)\Gamma(b+1)}.$$

It also follows from the proof of these results (though not explicitly mentioned in Ref. 4) that

$$\mathcal{C}=\mathcal{L}|_{p=0},$$

$$=\prod_{i=1}^{M+N}(u^i)^{\varepsilon_i}$$

is independent of all the times, i.e.,

$$\frac{\partial\mathcal{C}}{\partial\tau_n}=0\qquad n=1,...,\infty.$$

The functions $\mathcal{C}$ will turn out to be a Casimir for the bi-Hamiltonian structure of this hierarchy.

Conservation laws are similarly defined, the conserved charges being given by

$$Q^{(n)}=\frac{1}{2\pi i}\oint\mathcal{L}^{n/(N-M)}\frac{dp}{p}. \tag{10}$$

These may be derived explicitly

$$Q^{(n)}=\sum_{\{r_i:\Sigma_{i=1}^{N+M}r_i=n\}}\left\{\prod_{i=1}^{N+M}\binom{\frac{\varepsilon_i n}{N-M}}{r_i}(u^i)^{r_i}\right\}.$$

Under a suitable change of variable, these polynomials take the form of generalized hypergeometric functions, a result which remains to be exploited. The corresponding functionals

$$H^{(n)}=\int Q^{(n)}dx \tag{11}$$

will turn out to be the Hamiltonians of the system (9).

III. POLYNOMIAL LAX EQUATIONS

In this section the bi-Hamiltonian structure of the hierarchy defined by a Lax function,

$$\mathcal{L}=p^{-M}\prod_{i=1}^{N}(p+u_i),\quad 0<M<N,$$

will be derived, this generalizing the results of Ref. 4, where the special case $M=1$ was studied. Having derived one Hamiltonian structure, the intersection form in the language of a Frobenius manifold, one may use a result of Dubrovin to find a second compatible Hamiltonian structure.

Proposition 3: The Hamiltonian structure of the hierarchy defined by Eq. (7) is given by the nondegenerate metric,

$$g^{ij}(u)=\begin{cases}[1-(N-M)]u^iu^i, & \text{if } i=j,\\ u^iu^j, & \text{if } i\neq j.\end{cases}\tag{12}$$

Comment: This is clearly a flat, nondegenerate metric, and so defines a Hamiltonian structure. What is less clear is whether this structure, coupled to the Hamiltonians given by (11), gives rise to the flows defined by (7). This may be shown to be the case by direct calculation. An alternative proof, viewing the polynomial as a reduction of the rational case, will follow from the Theorem 7 in Sec. IV.

A bi-Hamiltonian structure is more than just two Hamiltonian structures; the two structures $\{,\}_1$ and $\{,\}_2$ have to be compatible, i.e., $\{,\}=\{,\}_1+\lambda\{,\}_2$ must be a Hamiltonian structure for all values of λ. For nondegenerate Poisson brackets of a hydrodynamic type, this compatibility condition implies that, for arbitrary λ, (a) the metric $g^{ij}=g_1^{ij}+\lambda g_2^{ij}$ is flat (such a metric is sometimes referred to as a flat pencil); (b) the metric connection for this metric has the form $\Gamma_k^{ij}=\Gamma_{1k}^{ij}+\lambda\Gamma_{2k}^{ij}$.

A result of Dubrovin[5] (actually a special case of a more general result of Magri[8]) will enable the bi-Hamiltonian structure to be found.

Lemma 4: If for a flat metric in some coordinate system $x^1,\ldots,x^n$, both the components $g^{ij}(x)$ of the metric and $\Gamma_k^{ij}(x)$ of the corresponding metric connection depend linearly on the coordinate $x^{\bullet}$, then the metrics,

$$g_1^{ij}=g^{ij},$$

$$g_2^{ij}=\partial_{\bullet}g^{ij},$$

form a flat pencil, under the assumption that $\det[g_2^{ij}]\neq 0$. *The corresponding metric connection has the form*

$$\Gamma_{1k}^{ij}=\Gamma_k^{ij},$$

$$\Gamma_{2k}^{ij}=\partial_{\bullet}\Gamma_k^{ij}.$$

The proof of this result is straightforward, and an alternative proof to that given in Ref. 5 will follow from a result given in the next section where this lemma is extended to degenerate Hamiltonian structures.

In order to find such a coordinate system, it is necessary to perform a number of coordinate transformations on the metric (12). This will be achieved in two stages. First, define variables,[6]

$$z^1=+x^1,$$

$$z^i=+x^i-x^{i-1},\quad i=1,\ldots,N-1,$$

$$z^N=-x^{N-1}$$

(so $\Sigma_{i=1}^{N} z^i = 0$), and then

$$u^i = e^{(1/N)x^N - z^i}, \quad i = 1,...,N.$$

Such a coordinate transformation has a nature interpretation in terms of the Weyl group $W(A_{N-1})$, which act by permutation of the coordinates z^i on the hyperplane $\Sigma_{i=1}^{N} z^i = 0$. In this x-coordinate system the components of the metric become, up to an overall factor of $(M-N)$,

$$g^{ij}(x) = \left(\begin{array}{ccccc|c} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & 0 \\ \hline 0 & 0 & 0 & \cdots & 0 & -\dfrac{M}{N(N-M)} \end{array} \right) = \left(\begin{array}{c|c} \text{Cartan matrix of } A_{N-1} & 0 \\ \hline 0 & -d_M^{-1} \end{array} \right).$$

The final entry is defined naturally using the Weyl group structure on A_{N-1},

$$\begin{aligned} d_M &= \frac{M(N-M)}{N}, \\ &= (\omega_M, \omega_M), \end{aligned}$$

where (,) is the Euclidean inner product and ω_i are the fundamental weights.[6]

What these coordinate transformation show is that the Hamiltonian structure coincides with those found by Dubrovin and Zhang, so their results may be used to complete the second part of this argument. In particular, they show that in terms of the symmetric functions,

$$\begin{aligned} s^1 &= \sum_i u^i, \\ s^2 &= \sum_{i<j} u^i u^j, \\ &\vdots \qquad \vdots \\ s^N &= \prod_i u^i, \end{aligned}$$

the metric (12) will be linear in the variable s^M, and hence Lemma 4 may be used to find the bi-Hamiltonian structure. The Jacobian of this transformation from modified to the original variables is just the Vandemonde determinant,

$$\frac{\partial(s^1,\ldots,s^N)}{\partial(u^1,\ldots,u^N)} = \left| \begin{array}{cccc} 1 & 1 & \cdots & 1 \\ \sum_{i\neq 1} u^i & \sum_{i\neq 2} u^i & \cdots & \sum_{i\neq N} u^i \\ \vdots & \vdots & \ddots & \vdots \\ \prod_{i\neq 1} u^i & \prod_{i\neq 2} u^i & \cdots & \prod_{i\neq N} u^i \end{array} \right| = \prod_{i<j} (u^i - u^j).$$

This defines the discriminant hypersurface, a caustic, on which $\mathcal{L}$ has multiple roots. By assumption $\varepsilon_i = \pm 1$, so all the roots are simple and hence the fields are well defined away from this surface. Hence Ref. 6.

Lemma 5 The metric (12), *when written in terms of the symmetric variables s^i, is linear in the variable s^M.*

Before performing these calculations one should note that in terms of these symmetric variables, the Lax function takes the more familiar form,

$$\mathcal{L} = p^{-M}[p^N + p^{N-1}s^1 + \cdots + s^N],$$

these symmetric variables coinciding with the original, unmodified variables. It also follows from the Lax equation (7) that the variable s^M is special for another reason, namely, it is the single variable for which the conserved charges $Q^{(n)}$ obey the relation

$$Q^{(n-1)} = \text{const}\,\frac{\partial Q^{(n)}}{\partial_s M}.$$

Proposition 6: The first Hamiltonian structure, in terms of the modified variables, is given by

$$\eta^{ij}(u) = \mathcal{L}_{\partial/\partial s^\bullet} g^{ij}(u),$$

$$= \frac{\partial}{\partial s^\bullet} g^{ij} - \frac{\partial \alpha^i_\bullet}{\partial u^k} g^{kj} - \frac{\partial \alpha^j_\bullet}{\partial u^k} g^{ik},$$

where the functions $\alpha^i_\bullet(u)$ are defined by

$$\frac{\partial}{\partial s^\bullet} = \alpha^i_\bullet(u)\,\frac{\partial}{\partial u^i},$$

and $\mathcal{L}_{\partial/\partial s^\bullet}$ is the Lie derivative along the vector field $\partial/\partial s^\bullet$.

Proof: The transformation between the modified variables and the symmetric variables induces the transformation

$$\begin{pmatrix} \dfrac{\partial}{\partial u^1} \\ \vdots \\ \dfrac{\partial}{\partial u^N} \end{pmatrix} = \begin{pmatrix} 1 & \sum_{j\neq 1} u^j & \cdots & \prod_{j\neq 1} u^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \sum_{j\neq N} u^j & \cdots & \prod_{j\neq N} u^j \end{pmatrix} \begin{pmatrix} \dfrac{\partial}{\partial s^1} \\ \vdots \\ \dfrac{\partial}{\partial s^N} \end{pmatrix},$$

and hence by inverting the Vandemonde determinant,

$$\frac{\partial}{\partial s^\bullet} = \alpha^i_\bullet(u)\,\frac{\partial}{\partial u^i},$$

this defining the functions $\alpha^i_\bullet(u)$. By Lemmas 4 and 5, it follows, by starting with the metric (12) in the u^i variables, transforming to the s^i variables, differentiating with respect to $s^\bullet$, and then transforming back to the u^i variables, that

$$\eta^{ij}(u) = \frac{\partial u^i}{\partial s^m}\frac{\partial u^j}{\partial s^n}\frac{\partial}{\partial s^\bullet}\left[\frac{\partial s^m}{\partial u^r}\frac{\partial s^n}{\partial u^s} g^{rs}(u)\right]$$

is the required flat metric, which defines the second Hamiltonian structure. Expanding yields

$$\eta^{ij}(u)=\frac{\partial}{\partial s^{\bullet}}g^{ij}+\frac{\partial u^i}{\partial s^m}\left(\frac{\partial}{\partial s^{\bullet}}\frac{\partial s^m}{\partial u^r}\right)g^{rj}+\frac{\partial u^j}{\partial s^n}\left(\frac{\partial}{\partial s^{\bullet}}\frac{\partial s^n}{\partial u^s}\right)g^{is}.$$

But

$$\left[\frac{\partial}{\partial s^{\bullet}},\frac{\partial}{\partial u^k}\right]=\left[\alpha^i_{\bullet}\frac{\partial}{\partial u^i},\frac{\partial}{\partial u^k}\right],$$
$$=-\frac{\partial\alpha^j_{\bullet}}{\partial u^k}\frac{\partial}{\partial u^j}.$$

This, together with the definition of the Lie derivative and

$$\frac{\partial s^{\beta}}{\partial s^{\bullet}}=\delta^{\beta}_{\bullet},$$

yields the result.

□

This proposition is just an application of Magri's more general result.[8]

Example 1: For arbitrary N and $M=1$ the distinguished coordinate is s^N (so $\bullet=N$ in the above formulas) and the functions α^i_N are

$$\alpha^i_N=\frac{u^i}{\Pi_{r\neq i}(u^N-u^r)},$$

and hence one may calculate η^{rs} explicitly:

$$\eta^{ij}(u)=(N-1)\frac{u^iu^j}{u^i-u^i}[\alpha^i_N-\alpha^j_N],\quad r\neq s,$$

$$\eta^{ii}(u)=2(N-1)\frac{(u^i)^2}{\Pi_{k\neq i}(u^i-u^k)}\left[1-u^i\sum_{n\neq i}\frac{1}{u^i-u^n}\right].$$

This, together with (12), *constitutes the bi-Hamiltonian structure for the hierarchy* (7), *also known as the continuum Toda hierarchy.*

If $N=2$, then

$$\eta^{ij}(u)=\frac{uv}{(u-v)^2}\begin{pmatrix}-2u & u+v\\ u+v & -2v\end{pmatrix}.$$

This example also shows an interesting result of the transformation from the original to the modified variables; in the original variables the form of g^{ij} is more complicated than the form of η^{ij} while in the modified variables the complexities are interchanged.

These results depend crucially on the properties of d_M. To see this consider the flat metric,

$$h^{ij}(u,v,w)=\begin{pmatrix}au^2 & uv & uw\\ uv & av^2 & vw\\ uw & vw & aw^2\end{pmatrix},$$

this being (12) with $N=3$ and $1-(N-M)$ replaced with an arbitrary constant a. We assume that this metric is invertible (so $a\neq 1,-2$). In terms of symmetric variables $S=u+v+w$, $P=uv+vw+wu$, and $Q=uvw$ this takes the form

$$h^{ij}(S,P,Q)=\begin{pmatrix} as^2+2(1-a)P & (1+a)SP+3(1-a)Q & (2+a)SQ \\ (1+a)SP+3(1-a)Q & 2(1+a)P^2+2(1-a)SQ & 2(2+a)PQ \\ (2+a)SQ & 2(2+a)PQ & 3(2+a)Q^2 \end{pmatrix}.$$

For general values of a, the entries are not linear in any of the variables. The metric cannot depend on Q linearly, as this would imply $a=-2$. For the entries to depend linearly on Q would imply $a=-1$, and this corresponds to (12) with $M=1$. For the entries to depend linearly on S would imply $a=0$, and this corresponds to (12) with $M=2$. Thus, any requirement that the metric depends linearly on one of the symmetric variables forces the metric to take one of the above known forms. Of course, this does not rule out the possibility that in some other coordinate systems the components of the metric do become linear in some variable.

IV. RATIONAL LAX EQUATIONS

In this section the evolution equations (9) will be written in Hamiltonian form. The resulting Hamiltonian structure turns out to be degenerate, so the results of Dubrovin used in the last section to derive the bi-Hamiltonian structure cannot be used without modification. These modifications turn out to be minor and a version of Lemma 4 will hold for degenerate Hamiltonian systems.

Theorem 7: (a) *In terms of the variables* $\tilde{u}^i=\log u^i$ *the evolution equations (9) may be written in Hamiltonian form,*

$$\tilde{u}^i_{\tau_n}=\sum_j m^{ij}\mathcal{D}\left(\frac{\delta H^{(n)}}{\delta\tilde{u}^j}\right),$$

where m^{ij} *is the constant matrix,*

$$m^{ij}=\begin{pmatrix} \alpha_1 & 1 & 1 & \cdots & 1 \\ 1 & \alpha_2 & 1 & \cdots & 1 \\ 1 & 1 & \alpha_3 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & \alpha_{N-M} \end{pmatrix}, \tag{13}$$

$\alpha_i=1-\varepsilon_i(N-M)$, *and*

$$H^{(n)}=\int dx \sum_{\{r_i:\Sigma_{i=1}^{N+M}r_i=n\}}\left\{\prod_{i=1}^{N+M}\begin{pmatrix}\frac{\varepsilon_i n}{N-M}\\ r_i\end{pmatrix}e^{r_i\tilde{u}^i}\right\}.$$

(B) In terms of the original variables the (g,Γ) *pair,*

$$g^{ij}(u)=m^{ij}u^iu^j,$$

$$\Gamma^{ij}_k(u)=\delta^j_k m^{ij}u^i,$$

define a degenerate Hamiltonian structure, satisfying the conditions of Theorem 2.

Proof: (a) In terms of the u^i variables, the system,

$$\tilde{u}^i_{\tau_n}=\sum_j m^{ij}\mathcal{D}\left(\frac{\delta H^{(n)}}{\delta\tilde{u}^j}\right), \tag{14}$$

becomes

$$u^i_{\tau_n} = \sum_i \alpha_i u^i \mathcal{D}\left(u^i \frac{\delta H^{(n)}}{\delta u^i} \right) + \sum_{j \neq i} u^i \mathcal{D}\left(u^j \frac{\delta H^{(n)}}{\delta u^j} \right).$$

Expanding this yields

$$u^i_{\tau_n} = \left[\frac{1}{n} \sum_{\{r_i : \Sigma_{i=1}^{N+M} r_i = n\}} [\alpha_i r_i^2 + r_i(n - r_i)] \prod_{k=1}^{N+M} \binom{\frac{\varepsilon_k n}{N-M}}{r_k} (u^k)^{r_k} \right] u^i_x$$
$$+ \sum_{j \neq i} \left[\frac{1}{n} \sum_{\{r_i : \Sigma_{i=1}^{N+M} r_i = n\}} [\alpha_i r_i r_j + r_j(n - r_i)](u^j)^{-1} \prod_{k=1}^{N+M} \binom{\frac{\varepsilon_k n}{N-M}}{r_k} (u^k)^{r_k} \right] u^j_x .$$

Using $\alpha_i = 1 - \varepsilon_i (N-M)$ and various binomial identities reduces this to

$$u^i_{\tau_n} = A_i^{(n)} u^i_x + \sum_{j \neq i} u^i B_{ij}^{(n)} u^j_x ,$$

where

$$A_i^{(n)} = \left(\frac{\varepsilon_i n}{N-M} - 1 \right) \sum_{\{r_j : \Sigma_{j=1}^{N+M} r_j = n\}} \left[\prod_{\substack{k=1 \\ k \neq i}}^{N+M} \binom{\frac{\varepsilon_k n}{N-M}}{r_k} (u^k)^{r_k} \right] \binom{\frac{\varepsilon_i n}{N-M} - 2}{r_i - 1} (u^i)^{r_i},$$

and

$$B_{ij}^{(n)} = \frac{\epsilon_j n}{N-M} \sum_{\{r_j : \Sigma_{j=1}^{N+M} r_j = n-1\}} \left[\prod_{\substack{k=1 \\ k \neq i,j}}^{N+M} \binom{\frac{\epsilon_k n}{N-M}}{r_k} (u^k)^{r_k} \right]$$
$$\times \binom{\frac{\epsilon_j n}{N-M} - 1}{r_i} (u^i)^{r_i} \binom{\frac{\epsilon_j n}{N-M} - 1}{r_j} (u^j)^{r_j},$$

that is, to the equations obtained from the Lax equation (7). Hence the result.

(b) Rewriting (14) in terms of a (g,Γ) pair yields

$$g^{ij}(u) = m^{ij} u^i u^j,$$

$$\Gamma^{ij}_k(u) = \delta^j_k m^{ij} u^i.$$

The above argument does not show that the pair (g,Γ) defines a Hamiltonian structure, as the corresponding bracket (1) must define a Hamiltonian structure for all functionals, not just the specific functionals used above. In order to show that this pair does define such a structure, one must verify that the equations (2)–(6) hold. This is entirely straightforward, so the details will be omitted. The degeneracy of the metric follows from the result, easily proved using elementary row and column operations, that

$$\det\left(\begin{array}{cccc|ccc} 1-a & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & 1-a & \cdots & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1-a & 1 & \cdots & 1 \\ \hline 1 & 1 & \cdots & 1 & 1+a & \cdots & 1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1+a \end{array}\right) = (-1)^N a^{N+M-1}[a-(N-M)],$$

where the diagonal blocks are $N\times N$ and $M\times M$ matrices. For the matrix m_{ij} $a=N-M$ [since $\alpha_i = 1-\varepsilon_i(N-M)$], and hence $\det(g^{ij})=0$. It also follows from these operations that $\mathrm{rank}(g^{ij}) = (N+M)-1$. □

This also shows that this system is only mildly degenerate; the coordinate transformation that reduces g^{ij} to a metric with constant entries simultaneously reduce the Γ^{ij}_k to zero. For a degenerate metric this need not be the case, and some nonzero Γ^{ij}_k can remain.[2]

Lemma 8: Let the pair (g,Γ) *define a degenerate Hamiltonian structure. If the components of the pair* (g,Γ) *in some coordinate system* $x^1,\ldots,x^n$ *depend linearly on the coordinate* $x\bullet$, *then the pair,*

$$(g+\lambda\,\partial_\bullet g, \Gamma+\lambda\,\partial_\bullet\Gamma), \tag{15}$$

defines a degenerate Hamiltonian structure for all values of λ. *Hence, one obtains a degenerate bi-Hamiltonian structure.*

Proof: All that is required is to show that the pair (15) satisfies the conditions of Theorem 2, given the original (g,Γ) pair. This is straightforward, the first two conditions being trivial. Consider, for example, condition (4) in Theorem 2:

$$[(\Gamma^{ij}_k + \lambda_\bullet \Gamma^{ij}_t)(g^{tk}+\lambda\,\partial_\bullet g^{tk}) - (k\leftrightarrow i)] = \left(1+\lambda\partial_\bullet + \frac{\lambda^2}{2}\partial_\bullet^2\right)[(\Gamma^{ij}_k - (k\leftrightarrow i)],$$

this following from that fact that if g and Γ depend linearly on $x^\bullet$, then

$$\partial_\bullet^2(\Gamma g) = 2\,\partial_\bullet\Gamma\,\partial_\bullet g.$$

Hence, if (g,Γ) satisfies condition (4), so does (15). The remaining conditions are all quadratic in g and Γ, and so the proof is identical. □

One may perform a similar sequence of coordinate transformation to those in Sec. III. Explicitly, let

$$z^1 = +x^1,$$

$$z^i = +x^i - x^{i-1}, \quad i=1,\ldots,N,$$

$$z^N = -x^{N-1},$$

$$z^{N+1} = +x^{N+1},$$

$$z^i = +x^i - x^{i-1}, \quad i=N+1,\ldots,N+M,$$

$$z^{N+M} = -x^{N+M-1},$$

and

$$u^i = e^{(1/N)x^N - z^i}, \quad i=1,\ldots,N+M.$$

After a permutation is the labels the metric becomes (up to an overall factor)

$$g^{ij}(x)=\left(\begin{array}{c|c|c} +\begin{pmatrix}\text{Cartan matrix}\\ \text{of } A_{N-1}\end{pmatrix} & 0 & 0 \\ \hline 0 & -\begin{pmatrix}\text{Cartan matrix}\\ \text{of } A_{N+M-1}\end{pmatrix} & 0 \\ \hline 0 & 0 & -d_M\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}\end{array}\right).$$

This is a considerable scope for the investigation of bi-Hamiltonian structures based on such block decompositions.

Having derived one Hamiltonian structure it is necessary, before the above lemma can be applied, to find a suitable coordinate system in which the metric given in Theorem 7 becomes linear in one of the coordinates. This will be done only for the $M=1$ case, i.e., a rational Lax function with a single pole. Extending these results to an arbitrary number of poles presents certain problems, which will be discussed later.

The new variable s^i are defined by the following expansion of the rational function $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}&=\text{polynomial of degree}(N-1)+\frac{\text{function}}{p+\text{pole}},\\ &=\sum_{n=0}^{N-1} p^n s^{N-1-n}+\frac{s^N}{p+s^{N+1}}.\end{aligned}$$

To express the s^i as functions of the variables u^i, it is convenient to introduce the basic symmetric functions of the variables $u^1,\ldots,u^N$:

$$\sigma^0=1,\quad \sigma^1=\sum_i u^i,\quad \sigma^2=\sum_{i<j} u^i u^j,\ldots,\sigma^N=\prod_i u^i,$$

so

$$\prod_{i=1}^{N}(p+u^i)=\sum_{i=0}^{N} p^i\sigma^{N-i}.$$

By expanding the various expressions for $\mathcal{L}$, one obtains

$$\begin{aligned}s^0&=1,\\ s^r&=\sum_{n=0}^{r}(-1)^n\sigma^{r-n}(u^{N+1})^n,\quad r=1,\ldots,N,\\ s^{N+1}&=u^{N+1}.\end{aligned}\tag{16}$$

It is in these variables that the pair (g,Γ) will become linear in one of the variables.

Example 2: For $N=3$, $M=1$,

$$\begin{aligned}\mathcal{L}&=\frac{(p+u)(p+v)}{p+w},\\ &=p+(u+v-w)+\frac{(u-w)(v-w)}{p+w},\end{aligned}$$

and hence

$$s^1 = u + v - w,$$

$$s^2 = uv - w(u+v) + w^2,$$

$$s^3 = w,$$

in accordance with (16).

The following result will be required in the next theorem.

Lemma 9: With the variables s^m defined above,

$$\frac{\partial s^m}{\partial u^i} + \frac{\partial s^m}{\partial u^{N+1}} = (u^{N+1} - u^i)\{\text{polynomial of degree } \alpha - 2\}, \quad m,i = 1,\dots,N;$$

Proof: One may write σ^n as $\sigma^n = u^i \tilde{\sigma}^{n-1} + \tilde{\sigma}^n$ so

$$\frac{\partial \sigma^n}{\partial u^i} = \tilde{\sigma}^{n-1}.$$

Hence

$$\left.\frac{\partial s^m}{\partial u^{N+1}}\right|_{u^{N+1}=u^i} = \sum (-1)^n (u^i \tilde{\sigma}^{n-1} + \tilde{\sigma}^n)(\alpha - n)(u^i)^{\alpha - n - 1},$$

$$= \sum (-1)^{n+1} \tilde{\sigma}^{n-1} (u^i)^{\alpha - n}.$$

So

$$\left.\left(\frac{\partial s^m}{\partial u^i} + \frac{\partial s^m}{\partial u^{N+1}}\right)\right|_{u^{N+1}=u^i} = 0.$$

The result now follows from the homogeneities of the functions involved. □

Theorem 10: The terms of the coordinates s^i defined above (16) the (g,Γ) pair depend linearly on the variable s^{N-1}.

Proof: In terms of the u^i variables the (g,Γ) pair is given by

$$g^{ij}(u) = m^{ij} u^i u^j,$$

$$\Gamma^{ij}_k(u) = \delta^j_k m^{ij} u^i,$$

where

$$m^{ij} = \begin{pmatrix} 2-N & 1 & \cdots & 1 & 1 \\ 1 & 2-N & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 2-N & 1 \\ 1 & 1 & \cdots & 1 & N \end{pmatrix}.$$

The components of the metric in terms of the s^i coordinates are given by

$$g^{ij}(s)=\frac{\partial s^i}{\partial u^p}\,\frac{\partial s^j}{\partial u^q}\,g^{pq}(u)$$

and it follows from the symmetry of this equation that the entries will be polynomial in the new variables. It also follows from this that the degrees of the entries are

$$\deg g^{ij}(s)=\begin{cases} i+j & \text{if } 1\leqslant i,j\leqslant N,\\ 1+j & \text{if } i=N+1 \text{ and } j\neq N+1,\\ 1+i & \text{if } j=N+1 \text{ and } i\neq N+1,\\ 2 & \text{if } i=j=N+1. \end{cases}$$

The degrees of the terms in the lower right corner of $g^{ij}(u)$ are given schematically below:

$$\left(\begin{array}{cccc|c} & & & & \vdots\\ & & & 2N-2 & N-1\\ & & 2N-2 & 2N-1 & N\\ & 2N-2 & 2N-1 & 2N & N+1\\ \hline \dots & N-1 & N & N+1 & 2 \end{array}\right). \tag{17}$$

Thus there are only four terms where $g^{ij}(s)$ could possibly contain a term quadratic in s^{N-1} (or six terms if $N=2$ or five terms if $N=3$, but these special cases may be disposed of by direct computation). The result will follow if it can be shown that these terms contain a factor s^N, that is if

$$g^{N-1,N-1}(s)=s^N\{\text{polynomial of degree } N-2\},$$

$$g^{N,k}(s)=s^N\{\text{polynomial of degree } k\},\quad k=N,N-1,N-2,$$

since the polynomials cannot be quadratic in s^{N-1} without violating the overall degree of the term.

From these formulas,

$$g^{N,\alpha}=\sum_{i,j=1}^{N}\frac{\partial s^N}{\partial u^i}\,\frac{\partial s^\alpha}{\partial u^j}\,g^{ij}+\sum_{i=1}^{N}\frac{\partial s^N}{\partial u^{N+1}}\,\frac{\partial s^\alpha}{\partial u^i}\,g^{N-1,i}\sum_{i=1}^{N}\frac{\partial s^N}{\partial u^i}\,\frac{\partial s^\alpha}{\partial u^{N+1}}\,g^{N-1,i}+\frac{\partial s^N}{\partial u^{N-1}}\,\frac{\partial s^\alpha}{\partial u^{N-1}}\,g^{N-1,N-1}.$$

Since

$$s^N=\prod_{i=1}^{N}(u^i-u^{N+1}),$$

it follows from Euler's theorem that

$$\sum_{i=1}^{N+1}u^i\frac{\partial s^N}{\partial u^i}=Ns^N,$$

and these may be used to simplify the above. After somewhat tedious calculations one obtains

$$\frac{\partial s^N}{\partial u^{N-1}}=\alpha s^N s^\alpha+(1-N)s^N\left\{(u^{N+1})^2\left[\frac{\dfrac{\partial s^\alpha}{\partial u^i}+\dfrac{\partial s^\alpha}{\partial u^{N+1}}}{u^i-u^{N+1}}\right]-u^{N+1}\left[\frac{\partial s^\alpha}{\partial u^{N+1}}-\frac{\partial s^\alpha}{\partial u^i}\right]\right\}.$$

The result now follows from the above lemma.

The corresponding result for $g^{N-1,N-1}$ is similar and rest on proving, in a similar manner as above, that

$$g^{N-1,N-1}(s)=(N-1)s^N \underbrace{\sum_{i\neq j}\prod_{\substack{r=1\\ r\neq i,j}}^{N}(u^r-u^{N+1})}_{\text{symmetric polynomial in the } s \text{ variables of degree } (N-2)}.$$

These results show that in the s^i-coordinates the metric is linear in the coordinate s^{N-1}. The second half of the proof, showing that $\Gamma^{ij}_k(s)$ is also linear in s^{N-1} is similar, and follows from the transformation properties of Γ^{ij}_k. □

In what follows this second degenerately flat metric $\partial g^{ij}/\partial s^{\bullet}$ will be denoted by η^{ij}.

Example 3: N=2, *M*=1. With these values,

$$m^{ij}=\begin{pmatrix} 0 & 1 & 1\\ 1 & 0 & 1\\ 1 & 1 & 2\end{pmatrix},$$

and a short computation yields

$$g^{ij}(s)=\begin{pmatrix} 2s^2 & s^2(s^1-3s^3) & s^3(s^1-s^3)\\ s^2(s^1-3s^3) & 2s^2(s^2-s^1s^3+2(s^3)^2) & s^3(2s^2-s^1s^3)\\ s^3(s^1-s^3) & s^3(2s^2-s^1s^3) & 2(s^3)^2\end{pmatrix}.$$

This is linear in s^1 and hence

$$\eta^{ij}(s)=\frac{\partial g^{ij}(s)}{\partial s^1},$$

$$=\begin{pmatrix} 0 & s^2 & s^3\\ s^2 & -2s^2s^3 & -(s^3)^2\\ s^3 & -(s^3)^2 & 0\end{pmatrix}.$$

One may easily introduce degenerate flat coordinates in which the entries of η^{ij} are constant. These flat coordinates are

$$t^1=s^1,$$

$$t^2=\frac{s^2}{s^3},$$

$$t^3=\log(s^3)$$

and in these coordinates

$$g^{ij}(t)=\begin{pmatrix} 2t^2e^{t^3} & -2t^2e^{t^3} & +t^1-e^{t^3}\\ -2t^2e^{t^3} & +2t^2e^{t^3} & -t^1+e^{t^3}\\ +t^1-e^{t^3} & -t^1+e^{t^3} & 2\end{pmatrix}.$$

This is linear in t^1 and hence

$$\eta^{ij}(t)=\begin{pmatrix} 0 & 0 & +1 \\ 0 & 0 & -1 \\ +1 & -1 & 0 \end{pmatrix}.$$

Example 4: $N=3$, $M=1$. With these values,

$$m^{ij}=\begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & +3 \end{pmatrix},$$

and a short computation yields the degenerate metric $g^{ij}(s)$. These components are linear in s^2 and so define a new metric

$$\eta^{ij}(s)=\frac{\partial g^{ij}(s)}{\partial s^2},$$

$$=\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 0 & 2s^3 & 2s^4 \\ 0 & 2s^3 & -4s^3s^4 & -2(s^4)^2 \\ 0 & 2s^4 & -2(s^4)^2 & 0 \end{pmatrix}.$$

The degenerate flat coordinates are defined by

$$t^1=s^1,$$

$$t^2=s^2,$$

$$t^3=\frac{s^3}{s^4},$$

$$t^4=\log(s^4),$$

and in these flat coordinates the original metric metric takes the form

$$g^{ij}(t)=\begin{pmatrix} -(t^1)^2+4t^2 & +6t^3e^{t^4} & -6t^3e^{t^4} & t^1-2e^{t^4} \\ +6t^3e^{t^4} & +4t^1t^3e^{t^4}-8t^3e^{2t^4} & -4t^1t^3e^{t^4}+8t^3e^{2t^4} & +2t^2-2t^1e^{t^4}+2e^{2t^4} \\ -6t^3e^{t^4} & -4t^1t^3e^{t^4}+8t^3e^{2t^4} & +4t^1t^3e^{t^4}-8t^3e^{2t^4} & -2t^2+2t^1e^{t^4}-2e^{2t^4} \\ t^1-2e^{t^4} & +2t^2-2t^1e^{t^4}+2e^{2t^4} & -2t^2+2t^1e^{t^4}-2e^{2t^4} & 3 \end{pmatrix}.$$

The entries are linear in t^2 and, hence, one obtains the second metric,

$$\eta^{ij}(t)=\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & -2 \\ 0 & 2 & -2 & 0 \end{pmatrix}. \tag{18}$$

V. DEGENERATE FROBENIUS MANIFOLDS

The natural geometric setting in which to understand the bi-Hamiltonian structure of hydrodynamic systems is the Frobenius manifold.[5] One way to define such manifolds is to construct a function $F(t^1,\ldots,t^n)$ such that the associated functions,

$$c_{ijk}=\frac{\partial^3 F}{\partial t^i\,\partial t^j\,\partial t^k},$$

satisfy the following conditions.

(i) The matrix $\eta_{ij}=c_{1ij}$ is constant and nondegenerate. This together with the inverse matrix η^{ij} are used to raise and lower indices. On such a manifold one may interpret η_{ij} as a flat metric.

(ii) The functions $c^i_{jk}=\eta^{ir}c_{rjk}$ defined an associative commutative algebra with a unity element. This defines a Frobenius algebra on each tangent space $T^t\mathcal{M}$. This multiplication will be denoted by $u\cdot v$.

(iii) The functions F satisfies a quasihomogeneity condition, which may be expressed as

$$\mathcal{L}_{\mathrm{E}}F=d_F F+\{\text{quadratic terms}\}, \tag{19}$$

where E is a vector field known as the Euler vector field.

These conditions constitute the Witten–Dijkgraaf–Verlinde–Verlinde (or WDVV) equations. On such a manifold one may introduce a second flat metric defined by

$$g^{ij}=E(dt^i\cdot dt^j). \tag{20}$$

This metric, together with the original metric η^{ij}, define a flat pencil (i.e., $\eta^{ij}+\lambda g^{ij}$ is flat for all values of λ). Thus, one automatically obtains a bi-Hamiltonian structure from a Frobenius manifold. The corresponding Hamiltonians are defined recursively by the formula

$$\frac{\partial^2 h^{(n)}}{\partial t^i\,\partial t^j}=c^k_{ij}\frac{\partial h^{(n-1)}}{\partial t^k}. \tag{21}$$

The integrability conditions for this systems are automatically satisfied when the c^k_{ij} are defined as above.

One basic assumption in this definition is that the metric η_{ij} is nondegenerate, and it follows from this that the bi-Hamiltonian structures are also nondegenerate. Thus the degenerate bi-Hamiltonian structures obtained in the preceding section cannot be obtained from this construction. However, one may formulate the new notion of a degenerate Frobenius manifold in which the corresponding bi-Hamiltonian structures are degenerate.

Rather than develop the theory of degenerate Frobenius manifolds in full generality, an extended example will be given here based on the study of the hydrodynamic system,

$$u_\tau=u(v_x-w_x),$$

$$v_\tau=v(u_x-w_x),$$

$$w_\tau=w(u_x+v_x-2w_x),$$

obtained from the rational Lax function

$$\mathcal{L}=\frac{(p+u)(p+v)}{p+w}.$$

The bi-Hamiltonian structure of this system has already been derived in Example 3 in Sec. IV. To recapitulate, in the flat coordinates given by

$$t^1 = u + v - w,$$

$$t^2 = \frac{(u-w)(v-w)}{w},$$

$$t^3 = \log w,$$

the degenerate metrics that give rise to the degenerate bi-Hamiltonian structures are

$$\eta^{ij}(t) = \begin{pmatrix} 0 & 0 & +1 \\ 0 & 0 & -1 \\ +1 & -1 & 0 \end{pmatrix}, \tag{22}$$

$$g^{ij}(t) = \begin{pmatrix} 2t^2 e^{t^3} & -2t^2 e^{t^3} & +t^1 - e^{t^3} \\ -2t^2 e^{t^3} & +2t^2 e^{t^3} & -t^1 + e^{t^3} \\ +t^1 - e^{t^3} & -t^1 + e^{t^3} & 2 \end{pmatrix}. \tag{23}$$

The first few Hamiltonian densities (suitably normalized) are given by the formula (10), and in the flat coordinates these become

$$h^{(1)} = t^1,$$

$$h^{(2)} = \tfrac{1}{2}[(t^1)^2 + 2t^2 e^{t^3}],$$

$$h^{(3)} = \tfrac{1}{6}[(t^1)^3 + 6t^1 t^2 e^{t^3} - 3t^2 e^{2t^3}],$$

$$h^{(4)} = \tfrac{1}{24}[(t^1)^4 + 12(t^1)^2 t^2 e^{t^3} + 6(t^2)^2 e^{2t^3} - 12t^1 t^2 e^{2t^3} + 4t^2 e^{3t^3}].$$

From these and the recursion equation (21), one may reconstruct the structure functions c^i_{jk} and verify that they form a commutative and associative algebra with a unity element. Explicitly the structure constants are given by $c^i_{1j} = \delta^i_j$ and

$$\begin{pmatrix} c^1_{22} & c^2_{22} & c^3_{22} \\ c^1_{23} & c^2_{23} & c^3_{23} \\ c^1_{33} & c^2_{33} & c^3_{33} \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1/t^2 \\ +e^{t^3} & -e^{t^3} & 0 \\ +t^2 e^{t^3} & -t^2 e^{t^3} & -e^{t^3} \end{pmatrix}.$$

From these structure functions one may raise an index using η^{ij} and determine the Euler vector field from Eq. (20). For this example this vector field is

$$E = t^1 \frac{\partial}{\partial t^1} + t^2 \frac{\partial}{\partial t^2} + \frac{\partial}{\partial t^3}.$$

In addition, the structure functions satisfy the relations

$$\frac{\partial c^r_{jk}}{\partial t^i} - \frac{\partial c^r_{ik}}{\partial t^k} = 0,$$

and this, together with the symmetry $c^k_{ij} = c^k_{ji}$, enables one to write them as

$$c^i_{jk} = \frac{\partial^2 f^i}{\partial t^j\,\partial t^k},$$

for some set of functions f^i. For the above structure constants these turn out to be (up to linear terms)

$$f^1 = \tfrac{1}{2}(t^1)^2 + t^2 e^{t^3},$$

$$f^2 = \tfrac{1}{2}(t^2)^2 + t^1 t^2 - t^2 e^{t^3},$$

$$f^3 = t^1 t^3 - t^2 \log t^2 - e^{t^3}.$$

At this stage one normally lowers the i index and uses another symmetry to write c_{ijk} as the third derivative of some function F. This, however, assumes that the metric η^{ij} is invertible, which, for the metric given by (22), is not the case. However, one may write the f^i as

$$f^i(t) = \eta^{ij}\frac{\partial F}{\partial t^j} + h^i(t^1 + t^2).$$

Since the matrix η^{ij} is of rank 2, it follows it has a nontrivial kernel, so there exists a nonzero vector ζ_i such that $\eta^{ij}\zeta_j = 0$, and the functions h^i are functions of the combination $\zeta_i t^i$, which in this example is just $t^1 + t^2$. These functions h^i satisfy the single constraint $h^1 + h^2 = 1/2(t^1 + t^2)^2$. To obtain the above structure functions one possible such F is

$$F = \tfrac{1}{2}(t^1)^2 t^3 + t^2 e^{t^3} + \tfrac{1}{2}(t^2)^2 \log t^2,$$

and

$$h^1 = 0,$$

$$h^2 = \tfrac{1}{2}(t^1 + t^2)^2,$$

$$h^3 = 0,$$

and this satisfies the homogeneity condition (19) with $d_F = 2$. There is much freedom in these functions. One may transform, for arbitrary constant k,F,

$$F \to F + k t^3 (t^1 + t^2)^2,$$

and the homogeneity property is unchanged. This induced a change in the functions h^i but leaves unchanged the structure functions defining the Frobenius algebra.

From this extended example one may distill the basic properties of a degenerate Frobenius manifold. One starts with a basic function $F(t^i)$ satisfying some homogeneity condition and degenerate metric η^{ij}, the entries of which are constant in the t^i coordinates. The metric is not related to the third derivatives of F, as for nondegenerate Frobenius manifolds. The structure functions, which form a Frobenius algebra with a degenerate inner product, are defined by

$$c^i_{jk} = \eta^{ir}\partial_r \partial_j\,\partial_k F + \partial_j \partial_k h^i, \tag{24}$$

where the functions h^i are functions that depend on the kernel of the degenerate matrix η^{ij}. Thus, for degenerate Frobenius manifolds one has a set of extra functions related to the fact that the matrix η^{ij} is not of maximal rank. The associativity conditions result in a complicated set over-determined partial differential equations for F, the degenerate analog of the WDVV equations. One avenue for future research is to develop the concept of a degenerate Frobenius manifold more axiomatically.

Example 5: For $N=3$, $M=1$ the metrics $g^{ij}(t)$ and $\eta^{ij}(t)$ have been calculated in Example 4. One may repeat the calculations above and obtain the following: Euler vector field:

$$E=\frac{t^1}{2}\frac{\partial}{\partial t^1}+t^2\frac{\partial}{\partial t^2}+t^3\frac{\partial}{\partial t^3}+\frac{1}{2}\frac{\partial}{\partial t^4};$$

prepotential F:

$$F=\tfrac{1}{8}(t^1)^2t^2+\tfrac{1}{4}(t^2)^3-\tfrac{1}{192}(t^1)^4+\tfrac{1}{2}t^1t^3e^{t^4}-\tfrac{1}{4}t^3e^{2t^4}-\tfrac{1}{4}(t^3)^2\log t^3;$$

and associated nonzero potentials h^i:

$$h^3=\tfrac{1}{2}(t^2+t^3)^2.$$

From these, and the constant matrix η^{ij} given by (18), one may construct a degenerate Frobenius algebra with structure functions given by Eq. (24) and second degenerate flat metric given by (20).

In Sec. IV the bi-Hamiltonian structures were shown to exist for arbitrary N, but $M=1$. It is clear that the ideas will generalize to arbitrary M, and hence to degenerate Frobenius manifolds for arbitrary N and M. The following example is for $N=3$, $M=2$.

Example 6: For $N=3$, $M=2$ the flat coordinates are defined by the expansion

$$\mathcal{L}=p+t^1+\frac{t^2e^{t^4}}{p+e^{t^4}}+\frac{t^3e^{t^5}}{p+e^{t^5}}.$$

In these coordinates,

$$\eta^{ij}(t)=\begin{pmatrix}0&0&0&1&1\\0&0&0&-1&0\\0&0&0&0&-1\\1&-1&0&0&0\\1&0&-1&0&0\end{pmatrix},$$

and the Frobenius data is the following: Euler vector field:

$$E=t^1\frac{\partial}{\partial t^1}+t^2\frac{\partial}{\partial t^2}+t^3\frac{\partial}{\partial t^3}+\frac{\partial}{\partial t^4}+\frac{\partial}{\partial t^5};$$

prepotential F:

$$\begin{aligned}F&=+t^2e^{t^4}+t^3e^{t^5}+\tfrac{1}{2}t^2t^3\log(e^{t^4}-e^{t^5})^2+\tfrac{1}{2}((t^2)^2\log t^2+(t^3)^2\log t^3)\\&=-\tfrac{1}{2}t^2t^4(2t^1+t^2+2t^3)-\tfrac{1}{2}t^3t^5(2t^1+2t^2+t^3);\end{aligned}$$

and associated nonzero potentials h^i:

$$h^1=\tfrac{1}{2}(t^1+t^2+t^3)^2.$$

From this data the Frobenius algebra structure functions given by Eq. (24) and the second degenerate flat metric given by (20).

The form of these results suggest the following.

Conjecture 1: The metric given in Theorem 7 is linear in the coordinate s^{N-M}, where the coordinates s^i are defined in terms of the expansion of the rational Lax function,

$$\mathcal{L}=p^{N-M}+s^1p^{N-M-1}+\cdots+s^{N-M}+\frac{s^{N-M+1}}{p+s^{N+1}}+\cdots+\frac{s^N}{p+s^{N+M}}.$$

Moreover, there exist flat coordinates t^i *such that the variables* s^i *are polynomial functions of the variables* $t^1,\ldots,t^N,e^{t^{N+1}},\ldots e^{t^M}$, *and in which the entries* $\eta^{ij}(t)$ *are all constants.*

One would hope to be able to modify the results of Ref. 6 to prove this conjecture; the vanishing of the determinants of the metrics means that the results cannot be used directly. One should be able to modify the Gauss–Manin equations for the flat coordinates to include these degenerate examples.

VI. COMMENTS

One notable difference between the bi-Hamiltonian structure of the hierarchies considered here, these being multicomponent generalizations of Toda and Benney hierarchies,[4] and the bi-Hamiltonian structures of dispersionless KP-type hierarchies, is the degeneracy of the structures. The dispersionless KP-type hydrodynamic systems involve rationale such as (see, for example, those in Ref. 9)

$$\mathcal{L}=\frac{1}{2}p^2+S(x,t)+\frac{P(x,t)}{p-Q(x,t)},$$

and the Lax equation similar to Eq. (7), but with a Poisson bracket,

$$\{f,g\}_{PB}=(\partial_p f\,\partial_x g-\partial_x f\,\partial_p g).$$

The bi-Hamiltonian structure of these equations is not degenerate.[5,9] These rational Lax functions may be considered as a reduction of an infinite component Lax function $\mathcal{L}=\Sigma_{i=-\infty}^{N}s^ip^i$ and it may be of interest to see how constraining the resulting Hamiltonian structures results in the degenerate structures studied here.

The existence of a nontrivial Casimir for these systems is of interest. One possible reduction of these systems is to restrict the dynamics to the surface given by

$$\mathcal{C}=\text{const},$$

for example, the $(N=2,\ M=1)$ system,

$$u_\tau=u(v_x-w_x),$$

$$v_\tau=v(u_x-w_x),$$

$$w_\tau=w(u_x+v_x-2w_x),$$

when restricted to the surface $w=uv$ results in the system

$$u_\tau=u[(1-u)v_x-vu_x],$$

$$v_\tau=v[(1-v)u_x-uv_x].$$

How the Hamiltonian structure behaves under such a constraint is unknown. For nondegenerate Hamiltonian structures one may use the result of Ferapontov,[10] though this work would need to be generalized to include a degenerate Hamiltonian structure such as those considered here. More generally, one may restrict the above system to the surface $w=uvf(x)$ for some arbitrary function $f(x)$ (i.e., $\mathcal{C}=f^{-1}$.). This results in the system

$$u_\tau=u[(1-u)v_x-vu_x]-u^2vf'(x),$$

$$v_\tau = v[(1-v)u_x - uv_x] - v^2 u f'(x),$$

an example of inhomogeneous hydrodynamic system with specific x dependence. It may also be possible to obtain the Hamiltonian structure of these systems.[11]

The idea of a degenerate Frobenius manifold requires further elucidation. One complicating factor is that for a degenerate structure the transformation that reduces the components of the metric to constants will not, in general, reduce all of the components Γ_k^{ij} to zero.[2] The systems in this paper are special in this respect since in flat coordinates the components of Γ_k^{ij} are automatically zero, which is not the generic situation; the systems here are doubly degenerate. It has recently been shown by Kodama[12] that the degenerate Frobenius manifold constructed here may be embedded in a nondegenerate, higher-dimensional, Frobenius manifold under a somewhat singular limit. Clearly these ideas require further work.

Finally, this paper has only dealt with dispersionless systems. For polynomial Lax equations one has discrete counterparts, the simplest example being the Toda Lattice,

$$S_{n,\tau} = P_n - P_{n+1},$$

$$P_{n,\tau} = P_n(S_{n-1} - S_n),$$

which reduces to (8) in the continuum limit; the lattice variable becoming the continuous variable x. The bi-Hamiltonian structure of such systems have been studied in Ref. 7. Indeed, the structures obtained here could also be derived by taking certain limits of those structures, if they where known explicitly for arbitrary M and N. How to extend these results to rational discrete systems is unclear. One approach would be to use the ideas in Ref. 13, which deals with the interpretation of the inverse operator $(e^\partial + u)^{-1}$, or the ideas of Ref. 14, where one would consider term-by-term deformation of the underlying dispersionless system.

ACKNOWLEDGMENTS

I would like to thank Professor A. P. Fordy for informing me that the result of Dubrovin[5] was a special case of a result of Magri.[8]

[1] B. Dubrovin and S. P. Novikov, ''Hydrodynamics of weakly deformed soliton lattices. Differential geometry and Hamiltonian theory,'' Russ. Math. Surv. **44**, 35–124 (1989).

[2] N. I. Grinberg, ''On Poisson brackets of hydrodynamic type with a degenerate metric,'' Russ. Math. Surv. **40**, 231–232 (1985).

[3] I. Dorfmann, *Dirac Structures and Integrability of Nonlinear Evolution Equations* (Wiley, New York, 1993).

[4] D. B. Fairlie and I. A. B. Strachan, ''The algebraic and Hamiltonian structure of the dispersionless Benney and Toda hierarchies,'' Inverse Probl. **12**, 885–908 (1996).

[5] B. Dubrovin, ''*Geometry of 2D topological field theories, in Integrable Systems and Quantum Groups*, Springer Lecture Notes in Mathematics, edited by M. Francaviglia and S. Greco, 1996, Vol. 1620, pp. 120–3348.

[6] B. Dubrovin and Y. Zhang, ''Extended affine Weyl groups and Frobenius manifolds,'' Compositio Mathematica, **111**, 167–219 (1998).

[7] B. A. Kupershmidt, ''Discrete Lax equations and differential difference calculus,'' Astérisque **123**, 1–212 (1985).

[8] F. Magri, ''A simple model of the integrable Hamiltonian equation,'' J. Math. Phys. **19**, 1156–1162 (1978).

[9] S. Aoyama and Y. Kodama, ''Topological Landau–Ginzburg theory with a rational potential and the dispersionless KP hierarchy,'' Commun. Math. Phys. **182**, 185–219 (1996).

[10] E. V. Ferapontov, ''Dirac reduction of the operator $\delta^{ij} d/dx$ to a submanifold with flat normal connection,'' Funct. Anal. Appl. **26**, 83–86 (1992).

[11] M. V. Polyak, ''One-dimensional Hamiltonian systems of hydrodynamic type with an explicit dependence on the spacial variable,'' Russ. Math. Surv. **42**, 229–230 (1987).

[12] Y. Kodama (private communication).

[13] B. Enriquez, A. Yu Orlov, and V. N. Rubtsov, ''Dispersionful analogues of Benney's equations and *N*-wave systems,'' Inverse Probl. **12**, 241–50 (1996).

[14] T. Eguchi, Y. Yamada, and S. K. Yang, ''On the genus expansion in the topological string theory,'' Rev. Math. Phys. **7**, 279 (1995).

Simultaneous Beltrami conditions in coupled vortex dynamics

Z. Yoshida[a)]
Graduate School of Frontier Sciences, The University of Tokyo, Tokyo 113-8656, Japan

S. M. Mahajan
Institute for Fusion Studies, The University of Texas at Austin, Austin, Texas 78712

(Received 13 April 1999; accepted for publication 1 July 1999)

The two-fluid model of a plasma describes the strong coupling between the magnetic and the fluid aspects of the plasma. The Beltrami condition that demands alignment of vortices and flows becomes a system of simultaneous equations in the magnetic field and the flow velocity. Combining these equations yields the double curl Beltrami equation. General solvability of the equation has been proved using the spectral theory of the curl operator. The set of solutions contains field configurations that can be qualitatively different from the conventional constant-α-Beltrami fields (which are naturally included in the set). The larger new set may help us understand a variety of structures generated in plasmas. © *1999 American Institute of Physics.* [S0022-2488(99)02810-8]

I. INTRODUCTION

The Beltrami condition, an expression of the alignment of a vorticity with its flow, describes the simplest and perhaps the most fundamental equilibrium state in a vortex dynamics system (Sec. II). The resulting Beltrami fields constitute a null set for the generator of the evolution equation describing the vortex dynamics. It is also believed that the Beltrami fields are accessible and robust in the sense that they emerge as the nonlinear dynamics of vortices tends to self-organize the system through a weakly dissipative process (Appendix A).

The simplest example of a Beltrami condition is provided by a three-dimensional solenoidal field (flow) $\mathbf{u}$, obeying

$$\begin{cases} \nabla\times\mathbf{u}=\lambda\mathbf{u} \quad (\text{in } \Omega), \\ \mathbf{n}\cdot\mathbf{u}=0 \quad (\text{on } \partial\Omega), \end{cases} \tag{1}$$

where λ is a real (or complex) constant number, Ω $(\subset\mathbf{R}^3)$ is a bounded domain with a smooth boundary $\partial\Omega$ and $\mathbf{n}$ is the unit normal vector onto $\partial\Omega$. This system of linear equations is regarded as an eigenvalue problem with respect to the curl operator. The spectral theory of the curl operator reveals an interesting relation of this problem with the cohomology theory.[1] We have the following theorem.

(i) If Ω is simply connected, then (1) has a nonzero solution for special λ included in a set of discrete real numbers; these numbers represent the point spectrum of the self-adjoint part of the curl operator.

(ii) If Ω is multiply connected, then (1) has a nonzero solution for every $\lambda\in\mathbf{C}$.[2]

Our aim in this paper is to generalize this theory for ''coupled'' (or higher-order) Beltrami conditions[3] that describe structures far richer than the ones contained in the single curl Beltrami equation (1). In an ideal plasma, the coupling between the magnetic field and the plasma flow yields the ''double curl Beltrami equation,''

[a)]Electronic mail: yoshida@plasma.q.t.u-tokyo.ac.jp

$$\begin{aligned} &\nabla\times(\nabla\times\mathbf{u})+\alpha\nabla\times\mathbf{u}+\beta\mathbf{u}=0 \quad (\text{in } \Omega),\\ &\mathbf{n}\cdot\mathbf{u}=0, \quad \mathbf{n}\cdot(\nabla\times\mathbf{u})=0 \quad (\text{on } \partial\Omega), \end{aligned} \tag{2}$$

where $\mathbf{u}$ is either the magnetic field or the flow velocity of the plasma (Sec. III). Applying the spectral theory of the curl operator, we will show that (2) has a nonzero solution for arbitrary complex numbers α and β, if the domain Ω is multiply connected (Sec. IV). The method of present theory applies for general multicurl Beltrami equations obtained from simultaneous Beltrami conditions in coupled systems.

II. VORTEX DYNAMICS AND BELTRAMI CONDITION

We start with reviewing the prototype equation for vortex dynamics. Let $\boldsymbol{\omega}$ be a three-dimensional vector field representing a certain vorticity (contravariant vector field) in $\mathbf{R}^3$. We consider an incompressible flow $\mathbf{U}$ that transports $\boldsymbol{\omega}$. When the circulation associated with the vorticity is conserved everywhere, this $\boldsymbol{\omega}$ obeys the equation

$$\frac{\partial}{\partial t}\boldsymbol{\omega}-\nabla\times(\mathbf{U}\times\boldsymbol{\omega})=0. \tag{3}$$

In $\mathbf{R}^2$, the vorticity becomes a pseudoscalar field ω, and the vortex dynamics equation can be cast in the form of a Liouville equation,

$$\frac{\partial}{\partial t}\omega+\{\phi,\omega\}=0, \tag{4}$$

where ϕ is the Hamiltonian of an incompressible flow and $\{\ ,\ \}$ is the Poisson bracket, i.e.,

$$\mathbf{U}=\begin{pmatrix} \partial\phi/\partial y\\ -\partial\phi/\partial x\end{pmatrix}, \quad \{\phi,\omega\}=\frac{\partial\phi}{\partial y}\frac{\partial\omega}{\partial x}-\frac{\partial\phi}{\partial x}\frac{\partial\omega}{\partial y}. \tag{5}$$

The Beltrami condition with respect to (3) is

$$\mathbf{U}=\mu\boldsymbol{\omega}, \tag{6}$$

where μ is a certain scalar function. This condition assures the vanishing of the generator of the vortex dynamics equation (3). For (4), the Beltrami condition is simply

$$\phi=f(\omega), \tag{7}$$

which implies the commutation of the vorticity and the Hamiltonian of the flow.

The simplest example of the vortex dynamics equation is that of the Euler equation of incompressible ideal flows. Let $\mathbf{U}$ be an incompressible flow that obeys

$$\frac{\partial}{\partial t}\mathbf{U}+(\mathbf{U}\cdot\nabla)\mathbf{U}=-\nabla p, \tag{8}$$

where p is the pressure. Taking the curl of (8), we obtain the evolution equation for the vorticity $\boldsymbol{\omega}=\nabla\times\mathbf{U}$, which reads, in $\mathbf{R}^3$, as (3), and in $\mathbf{R}^2$, as (4). In the Beltrami flow, $\boldsymbol{\omega}$ parallels $\mathbf{U}$, i.e.,

$$\nabla\times\mathbf{U}=\mu\mathbf{U}. \tag{9}$$

We note that (9) is not Galilean invariant. We thus consider a bounded domain and impose a boundary condition [see (1)] to remove the freedom of the Galilei transform. Taking the divergence of (9), we find that the scalar function μ must satisfy

$$\mathbf{U}\cdot\nabla\mu=0, \tag{10}$$

demanding that μ must remain constant along each streamline of the flow $\mathbf{U}$. An analysis of the nonlinear system of elliptic–hyperbolic partial differential equations (9)–(10) involves extremely difficult mathematical issues. The characteristic curve of (10) is the streamline of the unknown flow $\mathbf{U}$, which can be chaotic (nonintegrable) in general three-dimensional problems. If we assume, however, that μ is a constant number, the analysis reduces into a simple but nontrivial problem, i.e., the eigenvalue problem of the curl operator. In this paper, our analysis is restricted to this mathematically well-defined subclass of Beltrami fields.

We end this section by reviewing another example of vortex dynamics; the magnetohydrodynamic (MHD) description of a plasma. The two principal equations of the ideal (dissipationless) conducting-fluid model are

$$\mathbf{E}+\mathbf{U}\times\mathbf{B}=0, \tag{11}$$

$$\frac{\partial}{\partial t}\mathbf{U}+(\mathbf{U}\cdot\nabla)\mathbf{U}=\frac{1}{\rho}(\mathbf{J}\times\mathbf{B}-\nabla p), \tag{12}$$

where $\mathbf{U}$, $\mathbf{J}$, $\mathbf{E}$, and $\mathbf{B}$ are, respectively, the flow velocity, the current density, the electric field, and the magnetic field measured in certain fixed coordinates, and ρ is the fluid mass density that is assumed to be constant. We may write

$$\mathbf{E}=-\frac{\partial}{\partial t}\mathbf{A}-\nabla\phi, \tag{13}$$

$$\mathbf{J}=\frac{1}{\mu_0}\nabla\times\mathbf{B}, \tag{14}$$

in terms of a vector potential $\mathbf{A}$ (such that $\nabla\times\mathbf{A}=\mathbf{B}$) and a scalar potential ϕ. Using Faraday's law,

$$\partial\mathbf{B}/\partial t=-\nabla\times\mathbf{E},$$

and taking the curl of (11) and (12), we obtain

$$\frac{\partial}{\partial t}\mathbf{B}-\nabla\times(\mathbf{U}\times\mathbf{B})=0, \tag{15}$$

$$\frac{\partial}{\partial t}\boldsymbol{\omega}-\nabla\times\left(\frac{\mathbf{J}\times\mathbf{B}}{\rho}+\mathbf{U}\times\boldsymbol{\omega}\right)=0, \tag{16}$$

where $\boldsymbol{\omega}=\nabla\times\mathbf{U}$. The Beltrami conditions for this system of vortex dynamics equations are

$$\mathbf{J}=\mu_1\mathbf{B}=\mu_2\mathbf{U}=\mu_3\boldsymbol{\omega}. \tag{17}$$

Using (14) in the first equality of (17), we get

$$\nabla\times\mathbf{B}=\mu\mathbf{B}, \tag{18}$$

which implies that $\mathbf{B}$ parallels its own vorticity [cf. (9)]. This configuration, for which the magnetic stress $\mathbf{J}\times\mathbf{B}$ vanishes, is aptly called ''force-free.''

In order to characterize the stellar magnetic fields, solutions to (18) were intensively studied in the 1950s.[4–6] For $\mu\neq0$, the magnetic field $\mathbf{B}$ has a finite curl, and, hence, the field lines are twisted. The current (proportional to $\nabla\times\mathbf{B}$), flowing parallel to the twisted field lines, creates what may be termed as ''paramagnetic'' structures. Such twisted magnetic field lines appear commonly in many different plasma systems such as the magnetic ropes created in solar and geomagnetic systems,[7] and galactic jets.[8] Some laboratory experiments have also shown that the ''relaxed state'' generated through turbulence is well described as solutions of the force-free equation.[9,10]

In the next section, we will show that a more adequate formulation of the plasma dynamics allows a much wider class of special equilibrium solutions. The set of new solutions contains field configurations that can be qualitatively different from the force-free magnetic fields.

III. DOUBLE CURL BELTRAMI FIELD

The two-fluid model for the macroscopic dynamics of a plasma differentiates between the electron and ion velocities. Denoting the electron (ion) flow velocity by $\mathbf{V}_e(\mathbf{V}_i)$, the macroscopic evolution equations become

$$\frac{\partial}{\partial t}\mathbf{V}_e+(\mathbf{V}_e\cdot\boldsymbol{\nabla})\mathbf{V}_e=\frac{-e}{m}(\mathbf{E}+\mathbf{V}_e\times\mathbf{B})-\frac{1}{mn}\boldsymbol{\nabla}p_e, \tag{19}$$

$$\frac{\partial}{\partial t}\mathbf{V}_i+(\mathbf{V}_i\cdot\boldsymbol{\nabla})\mathbf{V}_i=\frac{e}{M}(\mathbf{E}+\mathbf{V}_i\times\mathbf{B})-\frac{1}{Mn}\boldsymbol{\nabla}p_i, \tag{20}$$

where $\mathbf{E}$ is the electric field, p_e and p_i are, respectively, the electron and the ion pressures, e is the elementary charge, n is the number density of both electrons and ions (we consider a quasineutral plasma with singly charged ions), and m and M are, respectively, the electron and the ion masses. In the electron equation, the inertial terms [the left-hand side of (19)] can be safely neglected, because of their small mass $(m\ll M)$.[11] Therefore, (19) reduces to

$$\mathbf{E}+\mathbf{V}_e\times\mathbf{B}+\frac{1}{en}\boldsymbol{\nabla}p_e=0. \tag{21}$$

When electron mass is neglected, $\mathbf{V}_i=\mathbf{V}$, the fluid velocity. We introduce the following set of dimensionless variables:

$$\begin{gathered}\mathbf{x}=\lambda_i\hat{\mathbf{x}},\quad \mathbf{B}=B_0\hat{\mathbf{B}},\\ t=(\lambda_i/V_{\mathrm{A}})\hat{t},\quad p=(B_0^2/\mu_0)\hat{p},\quad \mathbf{V}=V_{\mathrm{A}}\hat{\mathbf{V}},\\ \mathbf{A}=(\lambda_iB_0)\hat{\mathbf{A}},\quad \phi=(V_{\mathrm{A}}\lambda_iB_0)\hat{\phi},\end{gathered} \tag{22}$$

where the ion skin depth,

$$\lambda_i=\frac{c}{\omega_{pi}}=\frac{V_{\mathrm{A}}}{\omega_{ci}}=\sqrt{\frac{M}{\mu_0ne^2}},$$

is a characteristic length scale of the system, and the Alfvén speed is given by $V_{\mathrm{A}}=B_0/\sqrt{\mu_0Mn}$ (we assume $n=$const, for simplicity), with B_0 as an appropriate measure of the magnetic field.

Writing $\hat{\mathbf{E}}=-\partial\hat{\mathbf{A}}/\partial\hat{t}-\hat{\boldsymbol{\nabla}}\hat{\phi}$, the dimensionless version of (21) and (20) now reads as

$$\frac{\partial}{\partial\hat{t}}\hat{\mathbf{A}}=(\hat{\mathbf{V}}-\hat{\boldsymbol{\nabla}}\times\hat{\mathbf{B}})\times\hat{\mathbf{B}}-\hat{\boldsymbol{\nabla}}(\hat{\phi}+\hat{p}_e), \tag{23}$$

$$\frac{\partial}{\partial \hat{t}}(\hat{\mathbf{V}}+\hat{\mathbf{A}})=\hat{\mathbf{V}}\times(\hat{\mathbf{B}}+\hat{\boldsymbol{\nabla}}\times\hat{\mathbf{V}})-\hat{\boldsymbol{\nabla}}(\hat{V}^2/2+\hat{p}_i+\hat{\phi}). \tag{24}$$

In what follows, we shall drop the overcaret for a simpler notation. Taking the curl of (23) and (24), we can cast them in a revealing symmetric form,

$$\frac{\partial}{\partial t}\boldsymbol{\omega}_j-\boldsymbol{\nabla}\times(\mathbf{U}_j\times\boldsymbol{\omega}_j)=0 \quad (j=1,2), \tag{25}$$

in terms of a pair of generalized vorticities,

$$\boldsymbol{\omega}_1=\mathbf{B}, \quad \boldsymbol{\omega}_2=\mathbf{B}+\boldsymbol{\nabla}\times\mathbf{V},$$

and the effective flows,

$$\mathbf{U}_1=\mathbf{V}-\boldsymbol{\nabla}\times\mathbf{B}, \quad \mathbf{U}_2=\mathbf{V}.$$

The simplest equilibrium solution to (25) is given by the ''Beltrami conditions,''

$$\mathbf{U}_j=\mu_j\boldsymbol{\omega}_j \quad (j=1,2), \tag{26}$$

which implies the alignment of the vorticities and the corresponding flows. Writing $a=1/\mu_1$ and $b=1/\mu_2$, and assuming that a and b are constants, the Beltrami conditions (26) read as a system of simultaneous linear equations in **B** and **V**,

$$\mathbf{B}=a(\mathbf{V}-\boldsymbol{\nabla}\times\mathbf{B}), \tag{27}$$

$$\mathbf{B}+\boldsymbol{\nabla}\times\mathbf{V}=b\mathbf{V}. \tag{28}$$

These equations have a simple and significant connotation; the electron flow $(\mathbf{V}-\boldsymbol{\nabla}\times\mathbf{B})$ parallels the magnetic field **B**, while the ion flow **V** follows the ''generalized magnetic field'' $(\mathbf{B}+\boldsymbol{\nabla}\times\mathbf{V})$. This generalized magnetic field contains the Coriolis' force induced by the ion inertia effect on a circulating flow.

Combining (27) and (28) yields a second-order partial differential equation,

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{B})+\alpha\boldsymbol{\nabla}\times\mathbf{B}+\beta\mathbf{B}=0, \tag{29}$$

where

$$\alpha=\frac{1}{a}-b, \quad \beta=1-\frac{b}{a}.$$

The double curl Beltrami equation (29) encompasses a wide class of steady-state equations of mathematical physics. The conventional force-free-field equation (18), which describes paramagnetic fields, is included in this system as a special case: $\alpha=0$ and $\beta<0$. On the other hand, when $\alpha=0$ and $\beta>0$, (29) resembles London's equation of superconductivity with its well-known fully diamagnetic solutions. We note that, in this version of the London equation, the characteristic shielding length for the magnetic field is the ion skin depth c/ω_{pi}, instead of the usual electron skin depth c/ω_{pe}, because it is the ion dynamics that brings about the coupling of the magnetic field with the collective motion of the medium.

In the next section, we will study the mathematical structure of the double curl Beltrami equation with arbitrary complex α and β.[12]

IV. BELTRAMI FIELDS AND HARMONIC FIELDS

The single Beltrami condition (1) is known to have a nonzero solution for arbitrary complex number λ, if the domain Ω is multiply connected.[1] The harmonic field that represents the cohomology class of the differential forms in Ω plays an essential role to generate the Beltrami field. A similar relation holds in the double curl Beltrami equations (2). Here, we study the relation between the topology of the domain Ω and the degree of freedom in the solution of the double curl Beltrami fields.

It is convenient to denote the curl derivative $\nabla\times$ by ''curl'' to use it as an operator. Let us rewrite the differential equation of (2) in the form

$$(\mathrm{curl}-\lambda_+)(\mathrm{curl}-\lambda_-)\mathbf{u}=0, \tag{30}$$

where

$$\lambda_\pm=\tfrac{1}{2}[-\alpha\pm(\alpha^2-4\beta)^{1/2}]. \tag{31}$$

Because the two operators $(\mathrm{curl}-\lambda_\pm)$ commute, the general solution to (30) is given by a linear combination of two Beltrami fields. Let $\mathbf{G}_\pm$ be the Beltrami field, such that

$$\begin{cases}(\mathrm{curl}-\lambda_\pm)\mathbf{G}_\pm=0 \quad (\text{in } \Omega),\\ \mathbf{n}\cdot\mathbf{G}_\pm=0 \quad (\text{on } \partial\Omega).\end{cases}$$

Then, for arbitrary constants $c_\pm$, the sum

$$\mathbf{u}=c_+\mathbf{G}_+ + c_-\mathbf{G}_- \tag{32}$$

solves (30). Since $\mathbf{n}\cdot(\nabla\times\mathbf{G}_\pm)=\lambda_\pm\mathbf{n}\cdot\mathbf{G}_\pm=0$ on $\partial\Omega$, $\mathbf{u}$ satisfies the boundary conditions given in (2). Therefore, the existence of a nontrivial solution to the double curl Beltrami equations (2) will be predicated on the existence of the appropriate pair of single Beltrami fields (cf. Appendix B). Let us briefly review the mathematical theory of single Beltrami fields.[1]

Suppose that $\Omega(\subset\mathbf{R}^3)$ is a bounded domain with a smooth boundary $\partial\Omega=\cup_{i=1}^n\Gamma_i$. We consider cuts of the domain Ω. Let $\Sigma_1,\ldots,\Sigma_\nu$ $(\nu\geqslant 0)$ be the cuts such that $\Sigma_i\cap\Sigma_j=\emptyset$ $(i\neq j)$, and such that $\Omega\setminus(\cup_{j=1}^\nu\Sigma_j)$ becomes a simply connected domain. The number ν of such cuts is the first Betti number of Ω. When $\nu>0$, we define the flux through each cut by

$$\Phi_j(\mathbf{u})=\int_{\Sigma_j}\mathbf{n}\cdot\mathbf{u}\,ds \quad (j=1,\ldots,\nu),$$

where $\mathbf{n}$ is the unit normal vector on Σ_j with an appropriate orientation. By Gauss' formula, $\Phi_j(\mathbf{u})$ is independent of the place of the cut Σ_j, if $\nabla\cdot\mathbf{u}=0$ in Ω and $\mathbf{n}\cdot\mathbf{u}=0$ on $\partial\Omega$.

Let $L^2(\Omega)$ the Lebesgue space of square-integrable (complex) vector fields in Ω, which is endowed with the standard inner product,

$$(\mathbf{a},\mathbf{b})=\int_\Omega \mathbf{a}\cdot\overline{\mathbf{b}}\,dx.$$

We define the following subspaces of $L^2(\Omega)$:

$$L^2_\Sigma(\Omega)=\{\mathbf{w};\nabla\cdot\mathbf{w}=0 \text{ in } \Omega,\ \mathbf{n}\cdot\mathbf{w}=0 \text{ on } \partial\Omega,\ \Phi_j(\mathbf{w})=0\ (j=1,\ldots,\nu)\},$$

$$L^2_H(\Omega)=\{\mathbf{h};\nabla\cdot\mathbf{h}=0,\ \nabla\times\mathbf{h}=0 \text{ in } \Omega,\ \mathbf{n}\cdot\mathbf{h}=0 \text{ on } \partial\Omega\},$$

$$L^2_G(\Omega)=\{\nabla\phi;\ \Delta\phi=0 \text{ in } \Omega\},$$

$$L_F^2(\Omega)=\{\nabla\phi;\ \ \phi=c_i(\in\mathbf{C})\ \text{ on }\ \Gamma_i\ \ (i=1,...,n)\},$$

in terms of which we have an orthogonal decomposition,[13]

$$L^2(\Omega)=L_\Sigma^2(\Omega)\oplus L_H^2(\Omega)\oplus L_G^2(\Omega)\oplus L_F^2(\Omega).$$

The space of the solenoidal vector fields with vanishing normal components on $\partial\Omega$ is

$$L_\sigma^2(\Omega)=L_\Sigma^2(\Omega)\oplus L_H^2(\Omega).$$

The subspace $L_H^2(\Omega)$ corresponds to the cohomology class, whose member is a harmonic vector field and $\dim L_H^2(\Omega)=\nu$ (the first Betti number of Ω). When Ω is simply connected, then $\nu=0$ and $L_H^2(\Omega)=\emptyset$. We have the following expression:

$$L_\Sigma^2(\Omega)=\{\nabla\times\mathbf{w};\mathbf{w}\in H^1(\Omega),\ \ \nabla\cdot\mathbf{w}=0\ \text{ in }\ \Omega,\ \ \mathbf{n}\times\mathbf{w}=0\ \text{ on }\ \partial\Omega\},$$

where $H^1(\Omega)$ is the Sobolev space of first order. This says that a member of $L_\Sigma^2(\Omega)$ can be expressed as the curl of a vector potential with the boundary condition $\mathbf{n}\times\mathbf{w}=0$.

The spectral theory of the curl operator provides the basic understanding of the mathematical structure of the Beltrami equations. We repeat Theorem 1 of Yoshida–Giga.[1]

Theorem 1: *Suppose that Ω is a smoothly bounded domain in $\mathbf{R}^3$. We define a curl operator $\mathcal{S}$ in the Hilbert space $L_\Sigma^2(\Omega)$ by*

$$\mathcal{S}\mathbf{u}=\nabla\times\mathbf{u},$$

$$D(\mathcal{S})=\{\mathbf{u}\in L_\Sigma^2(\Omega);\ \ \nabla\times\mathbf{u}\in L_\Sigma^2(\Omega)\}.$$

The $\mathcal{S}$ is a self-adjoint operator. The spectrum of $\mathcal{S}$ consists of only point spectrum $\sigma_p(\mathcal{S})$, which is a discrete set of real numbers.

This theorem says that the Beltrami equation (1) together with the zero-flux condition [see the definition of the space $L_\Sigma^2(\Omega)$] has a nonzero solution only for special discrete real numbers $\lambda\in\sigma_p(\mathcal{S})$. If Ω is simply connected $(\nu=0)$, the topological flux $\Phi_j(\ )$ does not exist, so that $L_\Sigma^2(\Omega)=L_\sigma^2(\Omega)$. If Ω is multiply connected $(\nu\geqslant 1)$, however, we can remove the zero-flux condition assumed in Theorem 1, and consider a wider set of functions to find solutions of (1). This is done by considering the curl operator defined in the space $L_\sigma^2(\Omega)$. Let us trace the method of Yoshida–Giga.[1]

Lemma 1: For every $\mathbf{f}\in L_\sigma^2(\Omega)$, the equation,

$$\nabla\times\mathbf{u}=\mathbf{f}\quad(\text{in }\ \Omega),\tag{33}$$

has a unique solution in $L_\Sigma^2(\Omega)$.

Proof: Let $\mathbf{f}$ be the 0-extension of $\mathbf{f}$ over $\mathbf{R}^3$, i.e.,

$$\tilde{\mathbf{f}}(x)=\begin{cases}\mathbf{f}(x), & x\in\Omega,\\ 0, & x\notin\Omega.\end{cases}$$

Since $\mathbf{f}\in L_\sigma^2(\Omega)$, we have $\nabla\cdot\tilde{\mathbf{f}}=0$ in $\mathbf{R}^3$. We denote by $(-\Delta)^{-1}$ the vector Newtonian potential. We define

$$\mathbf{w}_0=\nabla\times[(-\Delta)^{-1}\tilde{\mathbf{f}}],\quad\text{in }\ \Omega.$$

We denote by $\mathcal{P}_\Sigma$ the orthogonal projection in $L^2(\Omega)$ onto $L_\Sigma^2(\Omega)$, and define $\mathbf{u}_0=\mathcal{P}_\Sigma\mathbf{w}_0$. Since $L_\Sigma^2(\Omega)$ is orthogonal to Ker(curl), we observe

$$\nabla\times\mathbf{u}_0=\nabla\times\mathbf{w}_0=\nabla\times\{\nabla\times[(-\Delta)^{-1}\tilde{\mathbf{f}}]\}.$$

Since $\nabla\cdot[(-\Delta)^{-1}\tilde{\mathbf{f}}]=0$, we obtain

$$\nabla\times\{\nabla\times[(-\Delta)^{-1}\tilde{\mathbf{f}}]\}=-\Delta[(-\Delta)^{-1}\tilde{\mathbf{f}}]=\tilde{\mathbf{f}}.$$

We thus find that $\mathbf{u}_0$ $[\in L^2_\Sigma(\Omega)]$ is the solution of (33). Since $L^2_\Sigma(\Omega)$ is orthogonal to Ker(curl), this $\mathbf{u}_0$ is the unique solution. □

This lemma shows that every solenoidal vector field [member of $L^2_\sigma(\Omega)$] has a unique vector potential in the space $L^2_\Sigma(\Omega)$. We apply this result to determine the vector potential of the harmonic field [member of $L^2_H(\Omega)$]. Let $\nu(\geqslant 1)$ be the dimension of $L^2_H(\Omega)$ (first Betti number of Ω), and $\mathbf{h}_j$ $(j=1,...,\nu)$ be the orthogonal basis of $L^2_H(\Omega)$, such that

$$\Phi_i(\mathbf{h}_j)=\int_{\Sigma_i}\mathbf{n}\cdot\mathbf{h}_j\,ds=\delta_{i,j}\,. \tag{34}$$

By solving (33) for $\mathbf{f}=\mathbf{h}_j$, we obtain the corresponding vector potential, which we denote by $\mathbf{g}_j$, i.e.,

$$\nabla\times\mathbf{g}_j=\mathbf{h}_j \quad (\text{in }\ \Omega), \quad \mathbf{g}_j\in L^2_\Sigma(\Omega) \quad (j=1,...,\nu).$$

Let us consider an arbitrary harmonic field and its vector potential, and write them as

$$\mathbf{h}=\sum_{j=1}^{\nu}\xi_j\mathbf{h}_j\,, \quad \mathbf{g}=\sum_{j=1}^{\nu}\xi_j\mathbf{g}_j\,. \tag{35}$$

For every $\lambda\notin\sigma_p(\mathcal{S})$, the resolvent operator $(\mathcal{S}-\lambda)^{-1}$ defines a unique continuous map on $L^2_\Sigma(\Omega)$. We consider

$$\mathbf{v}=\lambda\mathbf{g}+\lambda^2(\mathcal{S}-\lambda)^{-1}\mathbf{g}.$$

This $\mathbf{v}$ is the unique solution [in $L^2_\Sigma(\Omega)$] of

$$(\text{curl}-\lambda)\mathbf{v}=\lambda\mathbf{h} \quad (\text{in }\ \Omega). \tag{36}$$

Now we find that $\mathbf{u}=\mathbf{v}+\mathbf{h}$ solves

$$\begin{cases}(\text{curl}-\lambda)\mathbf{u}=0 & (\text{in }\ \Omega),\\ \mathbf{n}\cdot\mathbf{u}=0 & (\text{on }\ \partial\Omega).\end{cases}$$

Since $\mathbf{h}[\in L^2_H(\Omega)]$ and $\mathbf{v}[\in L^2_\Sigma(\Omega)]$ are orthogonal, $\mathbf{u}\not\equiv 0$.

We have shown that the single curl Beltrami equation (1) has a nonzero solution for every complex number λ, if the domain Ω is multiply connected. For $\lambda\notin\sigma_p(\mathcal{S})$, the solution is uniquely determined by the harmonic field $\mathbf{h}$. Although (1) appears as a homogeneous equation, the harmonic field (member of the kernel of curl) plays a role of a hidden inhomogeneous term; see (36). On the other hand, for $\lambda\in\sigma_p(\mathcal{S})$, the solution is given by the eigenfunction of the self-adjoint curl operator $\mathcal{S}$. Therefore, the solution is a zero-flux field, and $\mathbf{h}$ must be set to zero. The solution is not unique in the sense that any constant multiple of the eigenfunction is a solution.

Because of (32), it is now straightforward to generalize the theory for the double curl (and multicurl) Beltrami equations.

Theorem 2: *For a multiply connected smoothly bounded domain Ω, and for all complex numbers λ_1 and λ_2, the equation,*

$$(\text{curl}-\lambda_1)(\text{curl}-\lambda_2)\mathbf{u}=0, \tag{37}$$

has a nonzero solution.

Let us examine the relations among the solutions, the harmonic fields and the fluxes. If $\lambda_1,\lambda_2 \notin \sigma_p(\mathcal{S})$, then the solution is given by

$$\mathbf{u}=c_1\mathbf{u}_1+c_2\mathbf{u}_2,$$

$$\mathbf{u}_j=\mathbf{h}+\lambda_j\mathbf{g}+\lambda_j^2(\mathcal{S}-\lambda_j)^{-1}\mathbf{g} \quad (j=1,2),$$

where $\mathbf{h}\in L^2_H(\Omega)$, $\nabla\times\mathbf{g}=\mathbf{h}$, and $\mathbf{g}\in L^2_\Sigma(\Omega)$. Let us decompose $\mathbf{h}$ in terms of the normalized bases as (35). The coefficients $c_1,c_2,\xi_1,\ldots,\xi_\nu$ are related to the fluxes of $\mathbf{u}$ and $\nabla\times\mathbf{u}$ by

$$\begin{aligned}(c_1+c_2)\xi_j&=\Phi_j(\mathbf{u}),\\ (c_1\lambda_1+c_2\lambda_2)\xi_j&=\Phi_j(\nabla\times\mathbf{u}),\end{aligned} \quad (j=1,\ldots,\nu),$$

where $\Phi_j(\,)$ is the flux through the cut Σ_j. When $\nu=1$ (as in the case of a simple toroid), we can give the fluxes of both $\mathbf{u}$ and $\nabla\times\mathbf{u}$ independently to determine ξ_1 and c_1 with setting $c_2=1-c_1$ (cf. Appendix B). For $\nu>1$, the fluxes of $\nabla\times\mathbf{u}$ are not totally independent.

If $\lambda_1\in\sigma_p(\mathcal{S})$ and $\lambda_2\notin\sigma_p(\mathcal{S})$, we take $\mathbf{u}_1$ to be the eigenfunction corresponding to λ_1. Then, $\mathbf{u}_1$ is a zero-flux function, and, hence, c_1 is an arbitrary constant. The other component $\mathbf{u}_1$ carries fluxes. Taking $c_2=1$, we can determine

$$\xi_j=\Phi_j(\mathbf{u}) \quad (j=1,\ldots,\nu).$$

If $\lambda_1,\lambda_2\in\sigma_p(\mathcal{S})$, then both $\mathbf{u}_1$ and $\mathbf{u}_2$ are the corresponding eigenfunctions. A solution exists only for $\xi_j=0$ $(j=1,\ldots,\nu)$.

V. SUMMARY

The study of the solvability of the double curl equation is warranted both by physical as well as mathematical considerations. A more adequate modeling of plasma dynamics, containing a coupling of the magnetic and fluid aspects of a plasma, necessarily leads to a departure from the conventional single Beltrami equilibria (1), which are restricted to only force-free equilibria. This departure, then, leads to an immensely larger class of physically interesting equilibria, which can be constructed by a superposition of several different Beltrami fields. In the example dealt with in this paper (where the coupling is introduced by the Hall term), a superposition of two Beltrami fields suffices. Notice that in the nonlinear vortex dynamics models such as (3) with coupled $\boldsymbol{\omega}$ and $\mathbf{U}$, a linear combination of Beltrami fields is no longer a Beltrami field. Hence, a finite pressure and coupled flows can exist in conjunction with the magnetic field, and the structures that are far richer than those of single Beltrami fields come within the scope of the theory.[3]

The mathematical content of the paper may be summarized as follows: We have elucidated the general relation between the (double curl) Beltrami fields and the harmonic fields, which, being members of Ker(curl), play the role of a hidden inhomogeneous term in the Beltrami equations. The existence of harmonic fields invokes the multiply connectedness of the domain. For every $\lambda\in\mathbf{C}\backslash\sigma_p(\mathcal{S})$ (point spectrum of the self-adjoint curl operator), a harmonic field generates a nonzero unique Beltrami field corresponding to λ. When $\lambda\in\sigma_p(\mathcal{S})$, the corresponding eigenfunction gives the Beltrami field. The linear combination of two Beltrami fields yields the double curl Beltrami field.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Science and Culture No. 09308011. A part of this study was done at the workshop ''Various Approaches to Complex Systems'' held at the International Institute for Advanced Studies in Kyoto.

APPENDIX A: VARIATIONAL PRINCIPLE AND RELAXATION THEORY

The Beltrami condition can be derived by a variational principle invoking the ‘‘helicity.’’ Woltjer[14] derived the force-free equation (18) by minimizing the magnetic energy with the constraint that the magnetic helicity is conserved. Here, the magnetic helicity is, for a magnetic field **B** and its vector potential **A**,

$$H=\frac{1}{2}\int_\Omega \mathbf{A}\cdot\mathbf{B}\,dx.$$

Minimization of the magnetic field energy,

$$E=\frac{1}{2}\int_\Omega |\mathbf{B}|^2\,dx=\frac{1}{2}\int_\Omega |\nabla\times\mathbf{A}|^2\,dx,$$

with keeping H constant is represented by the variational principle

$$\delta(E-\lambda H)=0, \tag{A1}$$

where λ is the Lagrange multiplier. Assuming a boundary condition,

$$\mathbf{n}\times\mathbf{A}=0 \quad (\text{on } \partial\Omega) \tag{A2}$$

[note that $\mathbf{n}\cdot\mathbf{B}=0$ (on $\partial\Omega$) follows from (A2)], the formal Euler–Lagrange equation with respect to (A1) yields (18); see Ref. 15 for a more rigorous treatment of the variational principle.

Using Maxwell's equations, we obtain

$$\frac{\partial}{\partial t}(\mathbf{A}\cdot\mathbf{B})=-2\mathbf{E}\cdot\mathbf{B}-\nabla\cdot(\phi\mathbf{B}+\mathbf{E}\times\mathbf{A}).$$

In an ideal plasma, $\mathbf{E}\cdot\mathbf{B}=0$ [see (11)]. When Ω is surrounded by a perfectly conducting wall, we find that H is a constant of motion. Taylor[16] introduced the far-reaching concept of relaxation; he conjectured that a small amount of resistivity would tend to relax all constraints restricting an ideal plasma leaving only the ‘‘rugged’’ constraint on the global helicity H. When the magnetic energy achieves its minimum under the constraint on H, the ‘‘relaxed state’’ is characterized by the variational principle (A1), and, hence, the magnetic field satisfies the force-free equation (18). Many authors have examined the selective dissipation of the magnetic field energy E with respect to the helicity H (see Hasegawa[17] and papers cited there). Montgomery *et al.*[18] studied the statistical mechanical properties of the relaxed state using the Beltrami functions to expand fields (see also Ref. 19).

The conservation of helicity applies for general vortex dynamics. Let $\boldsymbol{\omega}$ be a vorticity that satisfies (3) and boundary condition,

$$\mathbf{n}\times(\mathbf{U}\times\boldsymbol{\omega})=0 \quad (\text{on } \partial\Omega).$$

The general ‘‘helicity’’ is defined as

$$\mathcal{H}=\frac{1}{2}\int_\Omega (\text{curl}^{-1}\,\boldsymbol{\omega})\cdot\boldsymbol{\omega}\,dx,$$

where curl^{-1} is the inverse operator of the curl that is represented by the Biot–Savart integral (see Lemma 1 for a more suitable treatment). By this definition, we easily verify the conservation of $\mathcal{H}$. For our two-fluid MHD model, we have two helicities:

$$H_1=\frac{1}{2}\int_\Omega \mathbf{A}\cdot\mathbf{B}\,dx, \qquad H_2=\frac{1}{2}\int_\Omega (\mathbf{A}+\mathbf{V})\cdot(\mathbf{B}+\nabla\times\mathbf{V})dx.$$

The total energy,

$$E=\frac{1}{2}\int_\Omega (|\mathbf{B}|^2+|\mathbf{V}|^2)dx,$$

is also conserved. The variation $\delta(E-\mu_1 H_1-\mu_2 H_2)=0$ will directly lead us to (27) and (28) (cf. Ref. 20).

APPENDIX B: EXAMPLES OF SOLUTIONS

Some explicit forms of the Beltrami fields may help understanding of the structures of the solutions.

When we consider a cubic volume that has sides of length a and assume the periodic boundary condition, we have the so-called *ABC* flow. Let *A, B,* and *C* be real (complex) constants and $\lambda=2\pi n/a$ $(n\in\mathbf{N})$. In the Cartesian coordinates, we define

$$\mathbf{u}=\begin{pmatrix} A\sin\lambda z+C\cos\lambda y\\ B\sin\lambda x+A\cos\lambda z\\ C\sin\lambda y+B\cos\lambda x\end{pmatrix}. \tag{B1}$$

We easily verify that (B1) gives an eigenfunction of the curl belonging to an eigenvalue λ. The linear combination of two *ABC* flows give the double curl Beltrami flow.

Solutions with the zero-normal boundary conditions are known for a cylindrical domain. In the (r,θ,z) cylindrical coordinates, the Chandrasekhar–Kendall function[6] is defined as

$$\mathbf{u}=\lambda(\nabla\psi\times\nabla z)+\nabla\times(\nabla\psi\times\nabla z), \tag{B2}$$

with

$$\lambda=\pm(\mu^2+k^2)^{1/2} \tag{B3}$$

$$\psi=J_m(\mu r)e^{i(m\varphi-kz)}, \quad k=2\pi n/L, \quad m,n\in\mathbf{N}, \tag{B4}$$

where J_m is the ordinary Bessel function and L is the length of the periodic cylinder. We find that $\mathbf{u}$ is an eigenfunction of the curl corresponding to the eigenvalue $\lambda(\in\mathbf{R})$. The eigenvalue is determined by the boundary condition that the normal component of $\mathbf{u}$ vanishes on the surface of the cylindrical domain. This condition becomes trivial when $k=m=0$. For these axisymmetric modes, we impose the ''zero-flux condition,''

$$\Phi(\mathbf{u})=\int_\Sigma \mathbf{n}\cdot\mathbf{u}\,ds=0, \tag{B5}$$

where Σ is a cut of the cylinder (cf. Theorem 1).

When we do not impose the zero-flux condition, however, the eigenvalue μ can be an arbitrary real (and even complex) number for the $k=m=0$ mode.[2] Therefore, we have nonzero Beltrami fields for arbitrary λ. For such a solution that has a finite flux $\Phi(\mathbf{u})$, the flux can be regarded as the variable of state. The double curl Beltrami field is a combination of two Beltrami fields, and, hence, the degree of freedom is two and two fluxes $\Phi(\mathbf{u})$ and $\Phi(\nabla\times\mathbf{u})$ can be assigned.

In a two-dimensional system, we can apply the Clebsch representation of solenoidal vector fields [cf. (4) and (5)]. For example, let us assume that the fields are homogeneous in the direction of z in the Cartesian coordinates x-y-z. We write $\mathbf{B}$ in a contravariant–covariant combination form,

$$\mathbf{B}=\nabla\psi\times\mathbf{e}+\phi\mathbf{e}, \tag{B6}$$

where $\mathbf{e}=\nabla z$. The ψ and ϕ are scalar functions of x and y. We have

$$\nabla\times\mathbf{B}=\nabla\phi\times\mathbf{e}+(-\Delta\psi)\mathbf{e},$$

$$\nabla\times(\nabla\times\mathbf{B})=\nabla(-\Delta\psi)\times\mathbf{e}+(-\Delta\phi)\mathbf{e}.$$

Using these expressions in the double curl Beltrami equation (29), we obtain a system of coupled Helmholtz equations,

$$-\Delta\psi+\alpha\phi+\beta\psi=C, \qquad -\Delta\phi-\alpha\Delta\psi+\beta\phi=0, \tag{B7}$$

where C is a constant. Biasing the potential ψ with $-C/\beta$, we can eliminate this constant. The system (B7) can be casted into a symmetric form,

$$\Delta\begin{pmatrix}\psi\\ \phi\end{pmatrix}=\begin{pmatrix}\beta & \alpha\\ -\alpha\beta & \beta-\alpha^2\end{pmatrix}\begin{pmatrix}\psi\\ \phi\end{pmatrix}. \tag{B8}$$

Similar algebra applies for the case of axisymmetric (toroidal) systems, where we must take $\mathbf{e}=\nabla\theta$ in (B6) and assume that ψ and ϕ are functions of r and z in the r-θ-z cylindrical coordinates. Then, the Laplacian Δ is replaced by the Grad–Shafranov operator,

$$L=r\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}\right)+\frac{\partial^2}{\partial z^2}.$$

The coupled Grad–Shafranov equation of the type (B7) was derived previously for the analysis of toroidal equilibrium in a plasma–beam system, where the inertia force of the beam particles brings about coupling of the magnetic field and the beam flow.[21]

[1] Z. Yoshida and Y. Giga, Math. Z. **204**, 235 (1990).
[2] Without the zero-flux condition, the curl operator is not self-adjoint,[1] and, hence, the potency of the set of the eigenvalues (point spectrum) can be uncountable without violating the separability of the Hilbert space $L^2(\Omega)$.
[3] S. M. Mahajan and Z. Yoshida, Phys. Rev. Lett. **81**, 4863 (1998).
[4] R. Lüst and A. Schlüter, Z. Astrophys. **34**, 263 (1954).
[5] S. Chandrasekhar, Proc. Natl. Acad. Sci. USA **42**, 1 (1956).
[6] S. Chandrasekhar and P. C. Kandall, Astrophys. J. **126**, 457 (1957).
[7] L. Acton *et al.*, Science **258**, 618 (1992).
[8] A. Königl and A. R. Choudhuri, Astrophys. J. **289**, 173 (1985).
[9] J. B. Taylor, Rev. Mod. Phys. **58**, 741 (1986).
[10] Z. Yoshida *et al.*, J. Plasma Phys. **59**, 103 (1998).
[11] When the electron mass is neglected, the two-fluid model may be called Hall MHD.
[12] We remark that the solution $\mathbf{u}$ is a real function, if α and β are real. Indeed, if $\mathbf{u}$ is a solution for real α and β, then $\overline{\mathbf{u}}$ is also a solution (take the complex conjugate of the equations). Since the solution is unique (see Sec. IV), $\mathbf{u}=\overline{\mathbf{u}}$, and hence $\mathbf{u}$ is a real function.
[13] R. Temam, *Navier–Stokes Equations* (North-Holland, Amsterdam, 1984).
[14] L. Woltjer, Proc. Natl. Acad. Sci. USA **44**, 489 (1958).
[15] P. Laurence and M. Avellaneda, J. Math. Phys. **32**, 1240 (1991).
[16] J. B. Taylor, Phys. Rev. Lett. **33**, 1139 (1974).
[17] A. Hasegawa, Adv. Phys. **34**, 1 (1985).
[18] D. Montgomery, L. Turner, and G. Varala, Phys. Fluids **21**, 757 (1978).
[19] N. Ito and Z. Yoshida, Phys. Rev. E **53**, 5200 (1996).
[20] L. C. Steinhauer and A. Ishida, Phys. Rev. Lett. **79**, 3423 (1997).
[21] Z. Yoshida, Phys. Fluids B **1**, 1702 (1989).

Qualitative analysis of early universe cosmologies

Andrew P. Billyard[a)]
Department of Physics, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada

Alan A. Coley[b)]
Department of Physics, Dalhousie University, Halifax, Nova Scotiae B3H 3J5, Canada and Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada

James E. Lidsey[c)]
Astronomy Centre and Centre for Theoretical Physics, University of Sussex, Brighton, BN1 9QJ, United Kingdom

(Received 25 March 1999; accepted for publication 30 June 1999)

A qualitative analysis is presented for a class of homogeneous cosmologies derived from the string effective action when a cosmological constant is present in the matter sector of the theory. Such a term has significant effects on the qualitative dynamics. For example, models exist which undergo a series of oscillations between expanding and contracting phases due to the existence of a heteroclinic cycle in the phase space. Particular analytical solutions corresponding to the equilibrium points are also found.
[S0022-2488(99)00910-X]

I. INTRODUCTION

Very early universe cosmology provides one of the few environments where the predictions of fundamental theories of physics, and in particular string theories, can be investigated. String theory is the most promising candidate for a unified theory of the fundamental interactions. It introduces significant modifications to the standard, hot big bang model based on conventional Einstein gravity and a study of string-inspired cosmologies is therefore important.

String theories predict the existence of a graviton, $g_{\mu\nu}$, a scalar ''dilaton'' field, Φ, and an antisymmetric two-form potential, $B_{\mu\nu}$, with a field strength $H_{\mu\nu\lambda}\equiv\partial_{[\mu}B_{\nu\lambda]}$.[1,2] In four dimensions, the three-form field strength is dual to a one-form, $\nabla_\mu\sigma$, such that $H^{\mu\nu\lambda}\equiv e^{\Phi}\epsilon^{\mu\nu\lambda\kappa}\nabla_\kappa\sigma$, where $\epsilon^{\mu\nu\lambda\kappa}$ is the covariantly constant four-form.[3] The one-form may be interpreted as the gradient of a scalar ''axion'' field. The string field equations can then be derived from the effective action,[3]

$$S=\int d^4x\sqrt{-g}e^{-\Phi}\left[R+(\nabla\Phi)^2-\frac{1}{2}e^{2\Phi}(\nabla\sigma)^2\right]+S_M, \tag{1}$$

where S_M represents the action for perfect fluid matter sources, R is the Ricci curvature of the space–time and $g\equiv\det g_{\mu\nu}$. The dilaton-graviton sector of action (1) may be interpreted as a Brans–Dicke theory, where the coupling parameter between the two fields takes the specific value $\omega=-1$.[4] The value of the dilaton field determines the effective value of Newton's ''constant,'' $G_{\rm eff}\propto e^{\Phi}$.

The general solutions to the field equations of action (1) are known analytically when $S_M=0$ for both the spatially flat and isotropic Friedmann–Robertson–Walker (FRW) universes and

[a)]Electronic mail: jaf@mscs.dal.ca
[b)]Electronic mail: aac@mscs.dal.ca
[c)]Electronic mail: jlidsey@astr.cpes.susx.ac.uk

the anisotropic Bianchi type I models.[5,6] The purpose of the present paper is to qualitatively investigate the consequences of introducing a cosmological constant, Λ_M, into the matter sector of Eq. (1),

$$S=\int d^4x\sqrt{-g}\left\{e^{-\Phi}\left[R+(\nabla\Phi)^2-\frac{1}{2}e^{2\Phi}(\nabla\sigma)^2\right]-\Lambda_M\right\}. \tag{2}$$

This term may be interpreted as a perfect fluid matter stress with an equation of state $p=-\rho$. It could be generated by a slowly moving scalar field, with a kinetic energy contribution dominated by a self-interaction potential, $p\approx -V\approx -\rho$. Analytical FRW solutions have not been found for this model when the axion field is trivial and $\Lambda_M>0$.[7,8] Moreover, the combined effects of the cosmological constant and axion field have not been considered previously.

We determine the general structure of the phase space of solutions for spatially flat FRW and axisymmetric Bianchi type I cosmologies derived from action (2) for arbitrary Λ_M. This complements the work of Refs. 9–13, where the qualitative effects of introducing a cosmological constant, $\Lambda_M\propto e^{-\Phi}$, into the gravitational sector of Eq. (1) were determined.

The paper is organized as follows. In Sec. II, the cosmological field equations and solutions for a zero cosmological constant are presented. The qualitative behavior of the models with positive and negative Λ_M is determined in Secs. III and IV, respectively. The phase portraits are interpreted in Sec. V and we conclude with a discussion in Sec. VI.

II. COSMOLOGICAL FIELD EQUATIONS

The metric for the Bianchi type I model may be written in the form

$$ds^2=-dt^2+h_{ab}dx^a dx^b, \quad a,b=1,2,3, \tag{3}$$

where $h_{ab}(t)$ is a function of cosmic time t only and represents the metric on the surfaces of homogeneity. The axisymmetric model may be parametrized by $h_{ab}=e^{2\alpha(t)}(e^{2\beta(t)})_{ab}$, where $e^{3\alpha}$ denotes the effective spatial volume of the universe. The traceless, diagonal matrix $\beta_{ab}\equiv\mathrm{diag}[\beta,\beta,-2\beta]$ determines the shear of the models and we refer to β as the shear parameter.[14] The spatially flat, isotropic FRW model is recovered in the limit where $\beta=0$ and, in this case, e^{α} represents the scale factor of the universe.

Substituting the metric (3) into the action (2) and integrating over the spatial variables implies that

$$S=\int dt\, e^{3\alpha}\left\{e^{-\Phi}\left[6\dot{\alpha}\dot{\Phi}-6\dot{\alpha}^2+6\dot{\beta}^2-\dot{\Phi}^2+\frac{1}{2}e^{2\Phi}\dot{\sigma}^2\right]-\Lambda_M\right\}, \tag{4}$$

where the comoving volume has been normalized to unity without loss of generality and a dot denotes differentiation with respect to t. The field equations derived from Eq. (4) are given by

$$\ddot{\alpha}=\dot{\alpha}\dot{\varphi}+\dot{\varphi}^2-3\dot{\alpha}^2-6\dot{\beta}^2-\tfrac{3}{2}\Lambda_M e^{\varphi+3\alpha}, \tag{5}$$

$$\ddot{\varphi}=3\dot{\alpha}^2+6\dot{\beta}^2+\tfrac{1}{2}\Lambda_M e^{\varphi+3\alpha}, \tag{6}$$

$$\ddot{\sigma}=-(\dot{\varphi}+6\dot{\alpha})\dot{\sigma}, \tag{7}$$

$$\ddot{\beta}=\dot{\beta}\dot{\varphi}, \tag{8}$$

where

$$\varphi\equiv\Phi-3\alpha \tag{9}$$

defines the ''shifted'' dilaton field and the generalized Friedmann constraint takes the form

$$3\dot{\alpha}^2-\dot{\varphi}^2+6\dot{\beta}^2+\tfrac{1}{2}\dot{\sigma}^2e^{2\varphi+6\alpha}+\Lambda_M e^{\varphi+3\alpha}=0. \tag{10}$$

Equations (5)–(10) may be simplified by introducing the new time coordinate

$$\frac{d}{d\theta}\equiv e^{-(\varphi+3\alpha)/2}\frac{d}{dt} \tag{11}$$

and employing the generalized Friedmann constraint Eq. (10) to eliminate the axion field. The remaining field equations are then given by

$$\alpha''=\varphi'^2-\tfrac{9}{2}\alpha'^2+\tfrac{1}{2}\alpha'\varphi'-6\beta'^2-\tfrac{3}{2}\Lambda_M, \tag{12}$$

$$\varphi''=3\alpha'^2+6\beta'^2-\tfrac{1}{2}\varphi'^2-\tfrac{3}{2}\alpha'\varphi'+\tfrac{1}{2}\Lambda_M, \tag{13}$$

$$\beta''=\tfrac{1}{2}\beta'(\varphi'-3\alpha'), \tag{14}$$

where a prime denotes differentiation with respect to θ.

The general solution to Eqs. (5)–(10) is known when the cosmological constant vanishes.[5] It is given by

$$e^{\alpha}=e^{\alpha_*}\left|\frac{s}{s_*}\right|^{1/2}\left[\left|\frac{s}{s_*}\right|^{r}+\left|\frac{s}{s_*}\right|^{-r}\right]^{1/2},$$

$$e^{\Phi}=\frac{e^{\Phi_*}}{2}\left[\left|\frac{s}{s_*}\right|^{r}+\left|\frac{s}{s_*}\right|^{-r}\right],$$

$$\sigma=\sigma_*\pm e^{-\Phi_*}\left[\frac{|s/s_*|^{-r}-|s/s_*|^{r}}{|s/s_*|^{-r}+|s/s_*|^{r}}\right],$$

$$e^{\beta}=e^{\beta_*}\left|\frac{s}{s_*}\right|^{q}, \tag{15}$$

where $s\equiv\int^t dt'\,e^{-\alpha(t')}$ is conformal time, $\{\alpha_*,s_*,\Phi_*,\sigma_*,\beta_*\}$ are arbitrary constants, and $\{r,q\}$ satisfy the constraint equation $r=(3-12q^2)^{1/2}$.

The solutions to Eqs. (5)–(10) for a trivial axion field and zero cosmological constant have a power-law form,

$$e^{\alpha}=e^{\alpha_*}|t|^{\pm h_*},$$

$$e^{\Phi}=e^{\Phi_*}|t|^{\pm 3h_*-1},$$

$$e^{\beta}=e^{\beta_*}|t|^{\pm\sqrt{(1-3h_*^2)/6}}, \tag{16}$$

where h_* is a constant such that $|h_*|\leq 1/\sqrt{3}$. Solution (15) asymptotes to these power-law models at early and late times and the axion field is therefore dynamically negligible in these limits. When an axion field it present, as in Eq. (15), the universe undergoes a smooth transition between the two power-law solutions (16) and exhibits a bounce when $s\approx s_*$. In the isotropic limit, $h_*^2=1/3$, and the time-reversal of the $e^{\alpha}\propto|t|^{-1/\sqrt{3}}$ solution is inflationary. It corresponds to the prebig bang cosmology, where the inflationary expansion is driven by the kinetic energy of the dilaton field.[15]

In the next section we determine the phase portraits for the generalized model with a nontrivial axion field and $\Lambda_M>0$. The effect of the cosmological constant on the solutions (15) can then be established.

III. POSITIVE COSMOLOGICAL CONSTANT

When $\Lambda_M > 0$, we can rewrite Eqs. (12)–(14) using new variables defined by

$$h \equiv \alpha', \quad \psi \equiv \varphi', \quad N \equiv \beta'. \tag{17}$$

Equation (10) then implies that

$$\psi^2 \geqslant 3h^2 + 6N^2 + \Lambda_M \geqslant 0, \tag{18}$$

and consequently we may normalize with ψ. We therefore define

$$x \equiv \frac{\sqrt{3}h}{\psi}, \tag{19}$$

$$y \equiv \frac{6N^2}{\psi^2}, \tag{20}$$

$$z \equiv \frac{\Lambda_M}{\psi^2}, \tag{21}$$

$$\frac{d}{d\Theta} \equiv \frac{1}{\psi}\frac{d}{d\theta}, \tag{22}$$

and assume that $\psi > 0$. (The case $\psi < 0$ is related to a time-reversal of the system and the qualitative behavior is similar.) The three-dimensional system (12)–(14) is therefore given by

$$\frac{dx}{d\Theta} = (x + \sqrt{3})[1 - x^2 - y - z] + \frac{1}{2}z[x - \sqrt{3}], \tag{23}$$

$$\frac{dy}{d\Theta} = 2y\left\{[1 - x^2 - y - z] + \frac{1}{2}z\right\}, \tag{24}$$

$$\frac{dz}{d\Theta} = 2z\left\{[1 - x^2 - y - z] - \frac{1}{2}(1 - z - \sqrt{3}x)\right\}. \tag{25}$$

It follows from definitions (19)–(21) that the phase space is bounded with $0 \leqslant \{x^2, y, z\} \leqslant 1$ subject to the constraint $1 - x^2 - y - z \geqslant 0$. The invariant set $1 - x^2 - y - z = 0$ corresponds to a zero axion field. The dynamics of the system (23)–(25) is determined primarily by the dynamics in the invariant sets $y = 0$ and $z = 0$. These correspond to a zero shear parameter and a zero cosmological constant, respectively. The dynamics is also determined by the fact that the right-hand side of Eq. (24) is positive-definite so that y is a monotonically increasing function. This guarantees that there are no closed or recurrent orbits in the three-dimensional phase space.

A. Isotropic model for $\Lambda_M > 0$

The isotropic FRW cosmology corresponds to the invariant set $y = 0$, where the shear parameter is trivial. The system (23)–(25) reduces to the following plane system in this case:

$$\frac{dx}{d\Theta} = (x + \sqrt{3})[1 - x^2 - z] + \frac{1}{2}z[x - \sqrt{3}], \tag{26}$$

$$\frac{dz}{d\Theta} = 2z\left\{[1 - x^2 - z] - \frac{1}{2}(1 - z - \sqrt{3}x)\right\}. \tag{27}$$

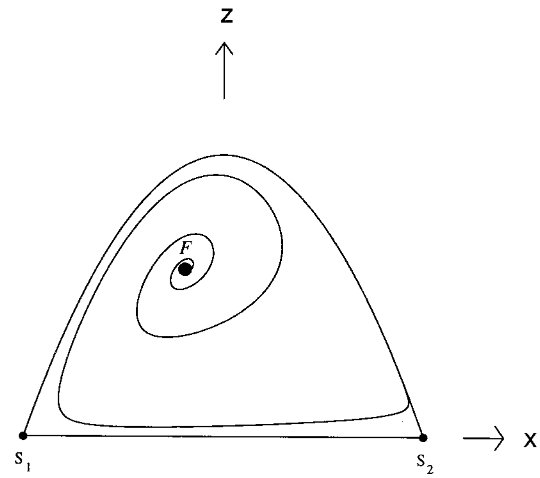

FIG. 1. Phase portrait of the system (26)–(27), corresponding to the isotropic FRW model with $\Lambda_M>0$. Equilibrium points are denoted by dots and the labels in all figures correspond to those equilibrium points (and hence the exact solutions they represent) discussed in the text. We shall adopt the convention throughout that large black dots represent sources (i.e., repellors), large gray-filled dots represent sinks (i.e., attractors), and small black dots represent saddles. Arrows on the trajectories have been suppressed since the direction of increasing time is clear using this notation. Note that in this phase space orbits are future asymptotic to a heteroclinic cycle.

The equilibrium points and their associated eigenvalues are given by

$$S_1: x=-1,\ \ z=0;\ \ \lambda_1=2(\sqrt{3}-1),\ \ \lambda_2=-(1+\sqrt{3}), \tag{28}$$

$$S_2: x=1,\ \ z=0;\ \ \lambda_1=-2(\sqrt{3}+1),\ \ \lambda_2=(\sqrt{3}-1), \tag{29}$$

$$F: x=-\frac{1}{3\sqrt{3}},\ \ z=\frac{16}{27};\ \ \lambda_{1,2}=\frac{1}{3}\pm\frac{i}{9}\sqrt{231}. \tag{30}$$

The points S_1 and S_2 are saddles and F is a repelling focus. The phase portrait is given in Fig. 1.

In the invariant set $1-x^2-z=0$, corresponding to the case of a zero axion field, Eqs. (26) and (27) reduce to the single ordinary differential equation,

$$\frac{dx}{d\Theta}=\frac{1}{2}(1-x^2)(x-\sqrt{3}), \tag{31}$$

which can be integrated to yield an exact solution in terms of Θ-time.

B. Anisotropic model for $\Lambda_M>0$

In the full system (23)–(25), corresponding to the anisotropic model with a nontrivial shear parameter, there exists the isolated equilibrium point (and their associated eigenvalues)

$$F:\quad x=-\frac{1}{3\sqrt{3}},\quad y=0,\quad z=\frac{16}{27}$$

$$(\lambda_1,\lambda_2,\lambda_3)=\left(\frac{1}{3}+\frac{i}{9}\sqrt{231},\frac{1}{3}-\frac{i}{9}\sqrt{231},\frac{4}{3}\right), \tag{32}$$

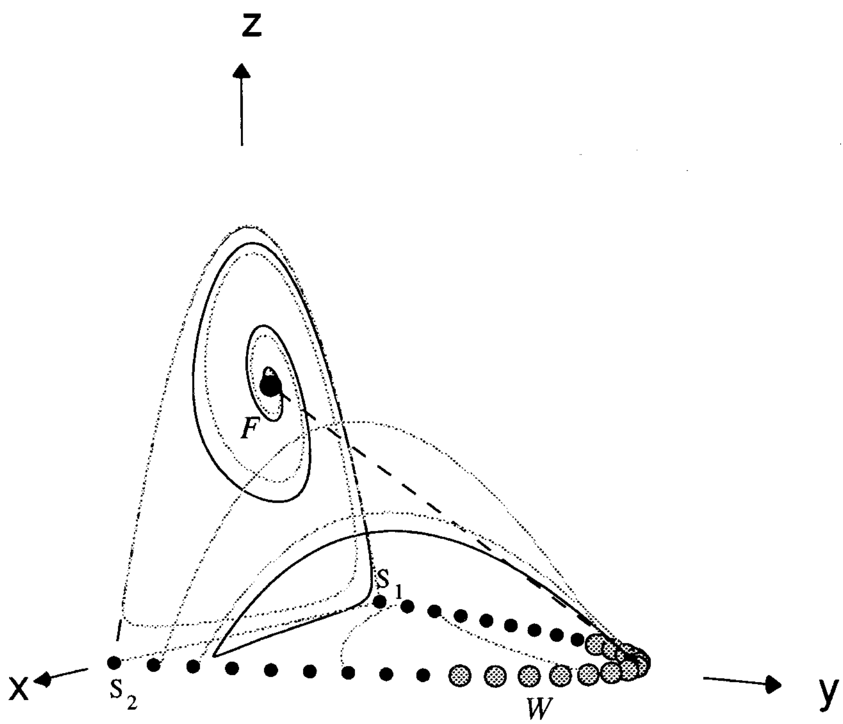

FIG. 2. Phase portrait of the system (23)–(25) corresponding to the axisymmetric Bianchi type I cosmology with $\Lambda_M > 0$. Note that W denotes a line of nonisolated equilibrium points. The dashed line represents the exact $x=$constant solution (34)–(36). See caption to Fig. 1 for notation. Gray lines represent typical trajectories found within the two-dimensional invariant sets, and solid black lines are typical trajectories within the full three-dimensional phase space.

$$W: \quad y=1-x^2, \quad z=0 \quad (x \text{ arbitrary})$$

$$(\lambda_1,\lambda_2,\lambda_3)=\left(-2\sqrt{3}\left[x+\frac{1}{\sqrt{3}}\right],\sqrt{3}\left[x-\frac{1}{\sqrt{3}}\right],0\right). \tag{33}$$

Hence, F is a global source. The set W lies in the invariant set $z=0$ on the boundary $y=1-x^2$. Points on W with $x\in(-1/\sqrt{3},1/\sqrt{3})$ are local sinks, while the remaining points are saddles in the full three-dimensional phase space {in the invariant set $z=0$ equilibrium points with $x\in[-1,-1/\sqrt{3})$ are repelling and those with $x\in(-1/\sqrt{3},1]$ are attracting}. The phase portrait for this system is given in Fig. 2 and Table I lists each equilibrium set and its stability.

We note that there exists an exact, anisotropic solution of Eqs. (23)–(25), where

$$x=-\frac{1}{3\sqrt{3}}=\text{constant} \tag{34}$$

TABLE I. Equilibrium sets for anisotropic model with $\Lambda_M > 0$, and their stability (the equations where each sets is defined is also listed).

Equilibrium point	Stability
F: Eq. (32)	Repellor (source)
W: Eq. (33)	Attractor (sink) for $(-1/\sqrt{3})<x<(1/\sqrt{3})$
	Saddle otherwise

and

$$y = -\frac{13}{8}\left(z - \frac{16}{27}\right). \tag{35}$$

This implies that

$$\frac{dz}{d\Theta} = \frac{9}{4} z\left(z - \frac{16}{27}\right) \tag{36}$$

and Eq. (36) can be integrated explicitly in terms of Θ-time.

In the following section we determine the effects of a negative cosmological constant. This allows a direct comparison to be made with the $\Lambda_M > 0$ models considered above.

IV. NEGATIVE COSMOLOGICAL CONSTANT

In the case where $\Lambda_M < 0$, the generalized Friedmann constraint Eq. (10) implies that

$$\psi^2 - \Lambda_M \geqslant 3h^2 + 6N^2 \geqslant 0. \tag{37}$$

We may therefore normalize by employing the quantity $\sqrt{\psi^2 - \Lambda_M}$. Defining the new variables

$$u \equiv \frac{\sqrt{3}h}{\sqrt{\psi^2 - \Lambda_M}}, \tag{38}$$

$$v \equiv \frac{\psi}{\sqrt{\psi^2 - \Lambda_M}}, \tag{39}$$

$$w \equiv \frac{6N^2}{\psi^2 - \Lambda_M}, \tag{40}$$

where $0 \leqslant \{u^2, v^2, w\} \leqslant 1$, and the new time variable

$$\frac{d}{d\Xi} = \frac{1}{\sqrt{\psi^2 - \Lambda_M}} \frac{d}{d\Theta} \tag{41}$$

implies that Eqs. (12)–(14) become

$$\frac{du}{d\Xi} = \frac{\sqrt{3}}{2}(1-u^2)(1-v^2) + (1-u^2-w)(\sqrt{3}+uv), \tag{42}$$

$$\frac{dv}{d\Xi} = -\frac{1}{2}(1-v^2)(1-2u^2-2w+\sqrt{3}uv), \tag{43}$$

$$\frac{dw}{d\Xi} = w[2v(1-u^2-w) - \sqrt{3}u(1-v^2)]. \tag{44}$$

The phase space is bounded by the sets $v = \pm 1$ and $w = 1 - u^2$, where the latter corresponds to a zero axion field. The dynamics is determined by the fact that the right-hand side of Eq. (42) is positive definite so that u is a monotonically increasing function.

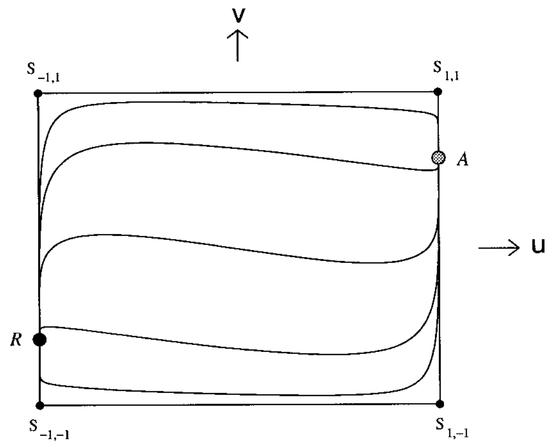

FIG. 3. Phase portrait of the system (45)–(46) corresponding to the isotropic FRW model with $\Lambda_M<0$. See caption to Fig. 1.

A. Isotropic model for $\Lambda_M<0$

In the invariant set $w=0$, corresponding to the isotropic FRW model $(\beta=0)$, the system (42)–(44) reduces to the following two-dimensional system:

$$\frac{du}{d\Xi}=\frac{1}{2}(1-u^2)(2uv+\sqrt{3}[3-v^2]), \tag{45}$$

$$\frac{dv}{d\Xi}=\frac{1}{2}(1-v^2)(2u^2-1-\sqrt{3}uv). \tag{46}$$

The lines $u^2=1$ and $v^2=1$ are invariant sets, containing four equilibrium points $S_{u,v}$. These points are all saddles and are located at the intersections of the lines. Their eigenvalues are given by

$$S_{1,1}: \lambda_1=\sqrt{3}-1,\ \ \lambda_2=-2(\sqrt{3}+1), \tag{47}$$

$$S_{-1,1}: \lambda_1=2(\sqrt{3}-1),\ \ \lambda_2=-(\sqrt{3}+1), \tag{48}$$

$$S_{1,-1}: \lambda_1=-2(\sqrt{3}-1),\ \ \lambda_2=(\sqrt{3}+1), \tag{49}$$

$$S_{-1,-1}: \lambda_1=-(\sqrt{3}-1),\ \ \lambda_2=2(\sqrt{3}+1). \tag{50}$$

The remaining two equilibrium points and their eigenvalues are

$$R: (u_-,v_-)=\left(-1,-\frac{1}{\sqrt{3}}\right);\ \ \lambda_1=\frac{1}{\sqrt{3}},\ \ \lambda_2=\frac{10}{\sqrt{3}}, \tag{51}$$

$$A: (u_+,v_+)=\left(1,\frac{1}{\sqrt{3}}\right);\ \ \lambda_1=-\frac{1}{\sqrt{3}},\ \ \lambda_2=-\frac{10}{\sqrt{3}}. \tag{52}$$

Consequently, R is a source and A is a sink. Figure 3 depicts the phase plane of the system (45)–(46).

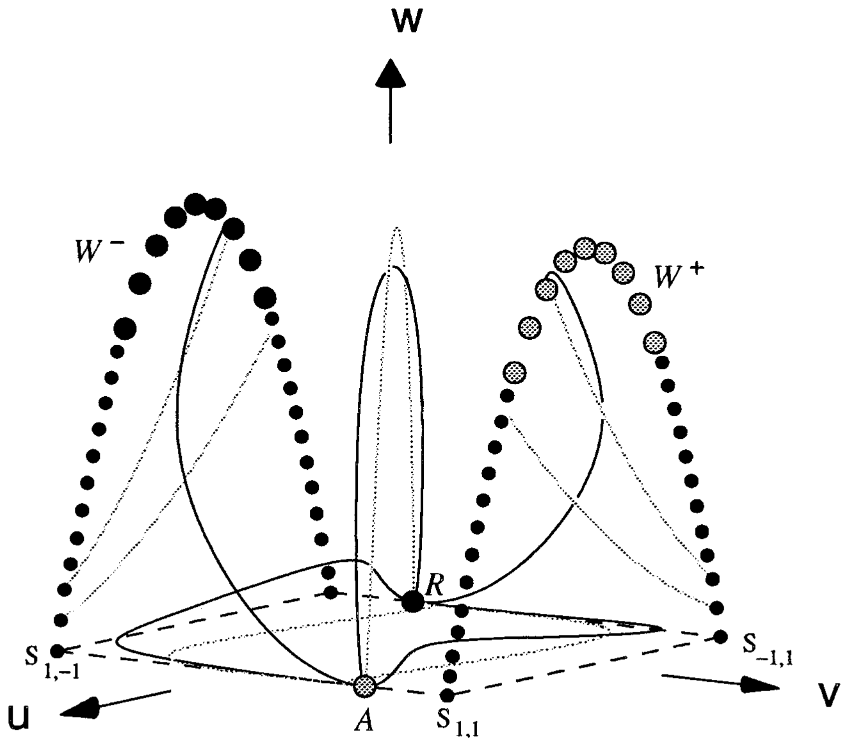

FIG. 4. Phase portrait of the system (42)–(44) corresponding to the axisymmetric Bianchi type I model with $\Lambda_M<0$. Note that $W^{\pm}$ denote lines of nonisolated equilibrium points. See captions to Figs. 1 and 2.

B. Anisotropic model for $\Lambda_M<0$

In the full system (42)–(44) with a nontrivial shear parameter, the equilibrium points and their respective eigenvalues are

$$W^{\pm}: v=\pm 1,\ u^2+w=1;\ (\lambda_1,\lambda_2,\lambda_3)=\left(0,\sqrt{3}\left[u\mp\frac{1}{\sqrt{3}}\right],-2\sqrt{3}\left[u\pm\frac{1}{\sqrt{3}}\right]\right), \tag{53}$$

$$R: v=-\frac{1}{\sqrt{3}},\ u=-1,\ w=0;\ (\lambda_1,\lambda_2,\lambda_3)=\frac{1}{\sqrt{3}}(1,2,10), \tag{54}$$

$$A: v=\frac{1}{\sqrt{3}},\ u=1,\ w=0;\ (\lambda_1,\lambda_2,\lambda_3)=-\frac{1}{\sqrt{3}}(1,2,10). \tag{55}$$

The saddle points $S_{\pm 1,-1}$ in Sec. IV A are the endpoints to the line W^-. This line represents early-time attracting solutions for $-1/\sqrt{3}<u<1/\sqrt{3}$ and saddles otherwise. The saddle points $S_{\pm 1,1}$ are the endpoints to the line W^+. This corresponds to late-time attracting solutions for $-1/\sqrt{3}<u<1/\sqrt{3}$ and saddles otherwise. Hence, there are two early-time attractors given by the point R and the line W^- for $-1/\sqrt{3}<u<1/\sqrt{3}$. There are also two late-time attractors corresponding to the point A and the line W^+ for $-1/\sqrt{3}<u<1/\sqrt{3}$. Figure 4 depicts the three-dimensional phase space and Table II lists each equilibrium set and its stability.

This concludes the derivation of the phase portraits for the spatially flat and homogeneous cosmologies derived from Eq. (2). We proceed in the following section to discuss their properties.

TABLE II. Equilibrium sets for anisotropic model with $\Lambda_M<0$, and their stability (the equations where each set is defined is also listed).

Equilibrium point	Stability
W^-: Eq. (53)	Repellor (source) for $(-1/\sqrt{3})<u<(1/\sqrt{3})$ Saddle otherwise
W^+: Eq. (53)	Attractor (sink) for $(-1/\sqrt{3})<u<(1/\sqrt{3})$ Saddle otherwise
R: Eq. (54)	Repellor (source)
A: Eq. (55)	Attractor (sink)

V. INTERPRETATION OF THE PHASE PORTRAITS

The dynamics of the isotropic cosmology described by the system (26)–(27) is of interest from a mathematical point of view due to the existence of the quasiperiodic behavior. The orbits are future asymptotic to a *heteroclinic cycle*, consisting of the two saddle equilibrium points S_1 and S_2 and the single (boundary) orbits in the invariant sets $z=0$ and $1-x^2-z=0$ joining S_1 and S_2 (see Fig. 1). The former set corresponds to the zero Λ_M solution [given by Eqs. (15) with $q=0$] and the latter to the solution with constant axion field [see Eq. (31)]; to our knowledge this exact solution was not previously known. In a given ''cycle,'' an orbit spends a long time close to S_1 and then moves quickly to S_2 shadowing the orbit in the invariant set $z=0$. It is then again quasistationary and remains close to the equilibrium point S_2 before quickly moving back to S_1 shadowing the orbit in the invariant set $1-x^2-z=0$. We stress that the motion is *not* periodic, and on each successive cycle a given orbit spends more and more time in the neighborhood of the equilibrium points S_1 and S_2.

In Fig. 1, the exact solution corresponding to the equilibrium point F is a power-law solution,

$$a=a_*(-t)^{1/3},$$

$$\Phi=\ln\left(\frac{16}{3\Lambda_M}\right)-2\ln(-t),$$

$$\sigma=\sigma_*\pm\frac{\sqrt{15}\Lambda_M}{16}(-t)^2,$$

$$\dot{\beta}=0, \tag{56}$$

where t is defined over the range $-\infty<t<0$ by a suitable choice of an integration constant. This new solution represents a cosmology that collapses monotonically to zero volume at $t=0$. The curvature and coupling are both singular at this point. The universe is initially in a weak coupling regime, since $G_{\rm eff}\to 0$ as $t\to-\infty$, and the effective energy density of the axion field also vanishes in this limit.

All orbits in Fig. 1 begin at F. The cyclical nature of these orbits can be physically understood by reinterpreting the axion field in terms of a membrane. The homogeneity of the axion field, $\sigma=\sigma(t)$, implies that the two-form potential, $B_{\mu\nu}$, must be independent of cosmic time, and this in turn implies that its field strength must be proportional to the volume form of the three-space. If the topology of the spatial sections is given by a three-torus, $S^1\times S^1\times S^1$, the behavior of the axion field is dynamically equivalent to that of a membrane that has been wrapped around this torus.[16] The collapse is resisted by this membrane and the universe undergoes a bounce. As the volume increases, however, the influence of the membrane is diminished, because the energy density of the axion is rapidly red-shifted away. Consequently, the cosmological constant becomes important.

The subsequent effect of the cosmological constant can be determined by viewing Eq. (2) in terms of a Brans–Dicke action, where the dilaton-graviton coupling parameter is given by $\omega=-1$.[4] The Brans–Dicke FRW models containing only a cosmological constant in the matter sector have been discussed previously by Barrow and Maeda, but their solutions only apply for $\omega>-5/6$.[7,8] The behavior of the general solution for $\omega<-5/6$ is different and can be established by performing a conformal transformation to a frame where the dilaton field is minimally coupled to gravity. In such a frame the term containing Λ_M may be viewed as an exponential, self-interaction potential for the dilaton, where the exponent is uniquely determined by the value of ω.[17] When $\omega>-5/6$, the late-time attractor is a scaling solution, where the kinetic and potential energies of the dilaton field redshift in direct proportion.[18] For $\omega<-5/6$, however, the potential is so steep that the dilaton effectively becomes massless.[19] Thus, the late-time attractor when $\omega=-1$ corresponds to the solution (16) where $h_*^2=1/3$.

Further insight may be gained by defining new variables in the reduced action (4),

$$\chi\equiv 4\alpha-\Phi,$$

$$\gamma\equiv\Phi-6\alpha,$$

$$\tilde{t}=\int dt e^{\varphi}. \tag{57}$$

In the case where $\dot{\beta}=\dot{\sigma}=0$, Eq. (4) reduces to

$$S=\int d\tilde{t}\left[-\frac{3}{2}\left(\frac{d\chi}{d\tilde{t}}\right)^2+\frac{1}{2}\left(\frac{d\gamma}{d\tilde{t}}\right)^2-\Lambda_M e^{-\gamma}\right]. \tag{58}$$

The momentum conjugate to the variable χ is constant, i.e., $d\chi/d\tilde{t}=C$, and the field equation for γ is a Liouville equation,

$$\frac{d^2\gamma}{d\tilde{t}^2}=\Lambda_M e^{-\gamma}. \tag{59}$$

The general solution to Eq. (59) satisfying the Hamiltonian constraint can be found. When $C>0$, it can be shown that $\gamma\propto\sqrt{3C^2}\tilde{t}$ in the late-time limit. Since the Hubble parameter is given by

$$\dot{\alpha}=-\frac{1}{2}e^{\varphi}\left(C+\frac{d\gamma}{d\tilde{t}}\right), \tag{60}$$

the late-time attractor corresponds to the *collapsing* solution in Eq. (16).

In effect, therefore, the cosmological constant resists the expansion and ultimately causes the universe to recollapse and asymptotically approach the saddle point S_1. On the other hand, the collapse causes the axion field to become relevant once more and a further bounce ensues. The process is then repeated with the universe undergoing a series of bounces. The orbits move progressively closer towards the two saddles, $S_{1,2}$, and spend increasingly more time near to these points. This behavior is related to the fact that the kinetic energy of the shifted dilaton field increases monotonically with time, since Eq. (6) implies that $\ddot{\varphi}>0$.

When shear is included $(y\neq 0)$, F still represents the *only* source in the system. The orbits follow cyclical trajectories in the neighborhood of the invariant set $y=0$ and they spiral outwards monotonically, since Eq. (24) implies that $dy/d\Theta>0$. After a finite (but arbitrarily large) number of cycles the kinetic energy associated with the shear parameter, β, begins to dominate the axion and cosmological constant. The orbits then asymptote to the power-law solutions (16). All orbits in the full three-dimensional phase space actually spiral outwards around the orbit represented by the dashed line in Fig. 2 which corresponds to the exact solutions (34)–(36) with $x=\text{constant}$. In terms of cosmic time, t, this exact solution satisfies

$$\Lambda_M - \frac{16}{27}\psi^2 + \frac{48}{13}N^2 = 0 \tag{61}$$

and

$$\dot{\alpha} = -\frac{1}{9}\dot{\varphi}, \tag{62}$$

whence from Eqs. (5)–(10) we obtain

$$\ddot{\varphi} = \dot{\varphi}^2 + k_\varphi^2 e^{2\varphi}, \tag{63}$$

where k_φ is an integration constant. Defining

$$\varrho \equiv \frac{e^{-\varphi}}{k_\varphi} \tag{64}$$

simplifies Eq. (63) to

$$\varrho\ddot{\varrho} = -1 \tag{65}$$

and Eq. (65) can be integrated exactly to obtain $\dot{\varphi}$.[20] A second integration then yields φ in terms of the inverse error function, so that in principle we can obtain the scale factor as a function of time, t, from Eq. (62).

This cyclical behavior does not arise if $\Lambda_M < 0$ (see Fig. 3). The equilibrium points A and R represent the power-law solutions,

$$\begin{aligned} a &= \frac{a_*}{\sqrt{\pm 2t}}, \\ \Phi &= \Phi_* - \ln[\pm\sqrt{-2\Lambda_M t}]^2, \\ \beta &= \beta_*, \\ \sigma &= \sigma_*, \end{aligned} \tag{66}$$

where the $+$ sign corresponds to the point R and the $-$ sign to the point A. Initially the universe is collapsing and the axion field induces a bounce, but this field can not dominate the dynamics again once the volume of the universe has increased sufficiently.

Figure 4 depicts the axisymmetric Bianchi type I model when $\Lambda_M < 0$. In this phase space, for $-1/\sqrt{3} < u < 1/\sqrt{3}$ the line W^- represents the positive branch of the solution (16) for $h_* \in (-1/3, 1/3)$. Likewise, for $-1/\sqrt{3} < u < 1/\sqrt{3}$ the line W^+ represents the ''$-$'' solution in Eq. (16) for $h_* \in (-1/3, 1/3)$. The four saddle points $S_{u,v}$ correspond to the power-law solutions (16) with $h_* = \pm 1/\sqrt{3}$. From Fig. 4 we see that generically trajectories asymptote away from either the line W^- or the point R and move towards the expanding power-law solutions W^+ or A. Hence, the cosmological constant is important in determining both the early- and late-time dynamics. Since u is monotonically increasing [see (42)] we note that the occurrence of a bounce in these cosmological models is a typical feature.

VI. DISCUSSION

In this paper we have presented a qualitative analysis of spatially flat FRW and Bianchi type I cosmologies containing nontrivial dilaton and axion fields with a cosmological constant in the matter sector of the theory. The action we considered reduces to the string effective action when the cosmological constant vanishes. A complete stability analysis was performed in all cases by finding variables that led to a compactification of the phase space. We found that a cosmological constant has a significant effect on the dynamics of the string cosmologies (15).

One of the more interesting mathematical features of the models we have considered is the existence of quasi-periodic behavior. This occurs in the isotropic cosmologies, where the orbits are future asymptotic to a heteroclinic cycle (see Fig. 1). The solutions interpolate between the saddles S_1 and S_2 corresponding to the power-law models (16) with $|h_*|=1/\sqrt{3}$. It would be interesting to consider the implications of this behavior for the prebig bang inflationary scenario.[15] We note that the phase portrait depicted in Fig. 1 is similar to that of Fig. 1(e) in Ref. 21 that describes the locally rotationally symmetric submanifold of the stationary Bianchi type I perfect fluid models in general relativity, although in this latter case the independent variable is spacelike.

The general Bianchi cosmology, where the shear matrix is given by

$$\beta_{ab}=\text{diag}[\beta_+ + \sqrt{3}\beta_- , \beta_+ - \sqrt{3}\beta_- , -2\beta_+] \tag{67}$$

can be analyzed directly by defining the variable N in Eqs. (17) and (20) via $N^2=\Sigma_{i=\pm}N_i^2$. Orbits in the full phase space of Fig. 2 with nontrivial shear term (represented by the variable y) are repelled from the source F. The variable y increases monotonically and the orbits spiral around the exact solution given by Eqs. (61)–(63), as represented by the dashed line in Fig. 2. [See also Fig. 1(f) and the Appendix in Ref. 21.] This implies that solutions are asymptotic in the past to the solution given by Eq. (56). At early times the orbits ''shadow'' the orbits in the invariant set $y=0$ and undertake cycles between the saddles (in three-dimensional phase space) on the equilibrium set W close to S_1 and S_2. These saddles on W may be interpreted as Kasner-type solutions.[21,22] Note that $y=0$ at S_1 and S_2, however, and there is no shear term in these cases. The orbits thus experience a finite number of cycles in which the solutions interpolate between different Kasner-type states. The orbits eventually asymptote towards a source on the line W.

This is perhaps reminiscent of the mixmaster behavior that occurs in the Bianchi type VIII and IX cosmologies.[22,23] These are the most general models in the Bianchi class A of spatially homogeneous universes.[24] In these models, Taub orbits joining equilibrium points of the Kasner set K lead to the existence of infinite heteroclinic sequences which approximate the past asymptotic behavior of generic orbits. (These heteroclinic sequences are defined by a map of K onto itself.) Mixmaster oscillations also occur in less general (i.e., lower-dimensional) Bianchi models with a magnetic field[25] or Yang–Mills fields.[26] It is interesting to note in the string context that mixmaster behavior also occurs in scalar-tensor theories of gravity in general and in the Brans–Dicke theory in particular.[27]

This analogy is only suggestive. We note that if a nonzero central charge deficit is included, the quasiperiodic behavior in the full (higher-dimensional) phase space does indeed persist.[28] Unlike the mixmaster oscillations, however, the orbits in Fig. 2 eventually spiral away from $y=0$, although there are orbits that experience a finite but arbitrarily large number of oscillations. However, it would be interesting to further explore any correspondence with possible mixmaster behavior, particularly by including additional anisotropic or matter degrees of freedom.

Some of the dynamics discussed in this paper is also relevant to higher-dimensional cosmological models. Kaluza–Klein compactification of ten-dimensional supergravity theories[2] onto an isotropic six-torus of radius e^{β} introduces an additional modulus field into the effective four-dimensional action (1). Integration over the spatial variables for a spatially flat FRW model then leads to an action that is formally identical to that of Eq. (4) when we specify $\Lambda_M=0$. In this sense, therefore, the action (4) can be recast into a higher-dimensional context, where the shear term β plays the role of the modulus field and Λ_M is interpreted as a cosmological constant that is introduced after compactification.

More generally, type II supergravity theories contain Ramond–Ramond form-fields that do not couple directly to the dilaton field in the string frame.[2] Under dimensional reduction, these fields give rise to terms in the effective action of the form $Q^2 \exp(c\beta)$, where Q and c are constants;[29] i.e., Ramond–Ramond charges give rise to exponential potentials for the modulus field rather than a simple constant term such as that considered in this work. However, from the analysis in Sec. III A, there will be string solutions containing Ramond–Ramond fields that asymptote towards solutions with $\beta=$constant ($y=0$ in Fig. 2), in which case $Q^2 \exp(c\beta)$ is effectively constant. It might then be expected that the heteroclinic cycle that occurs in the invariant set $y=0$ (see Fig. 1) will play an important rôle in describing the dynamics of these string cosmologies.

ACKNOWLEDGMENTS

We would like to thank Ulf Nilsson for helpful comments. A.P.B. is supported by Dalhousie University, A.A.C. is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and J.E.L. is supported by the Royal Society.

[1] E. S. Fradkin and A. Tseytlin, Phys. Lett. B **158**, 316 (1985); C. G. Callan, D. Friedan, E. J. Martinec, and M. J. Perry, Nucl. Phys. B **262**, 593 (1985); C. Lovelace, *ibid.* **273**, 413 (1986).
[2] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, 1987).
[3] A. Shapere, S. Trivedi, and F. Wilczek, Mod. Phys. Lett. A **6**, 2677 (1991); A. Sen, *ibid.* **8**, 2023 (1993).
[4] C. Brans and R. H. Dicke, Phys. Rev. **124**, 925 (1961).
[5] E. J. Copeland, A. Lahiri, and D. Wands, Phys. Rev. D **50**, 4868 (1994); **51**, 1569 (1995).
[6] K. A. Meissner and G. Veneziano, Mod. Phys. Lett. A **6**, 3397 (1991).
[7] J. D. Barrow and K. Maeda, Nucl. Phys. B **341**, 294 (1990).
[8] J. E. Lidsey, Phys. Rev. D **55**, 3303 (1997).
[9] D. S. Goldwirth and M. J. Perry, Phys. Rev. D **49**, 5019 (1993); K. Behrndt and S. Forste, Phys. Lett. B **320**, 253 (1994); Nucl. Phys. B **430**, 441 (1994).
[10] N. Kaloper, R. Madden, and K. A. Olive, Nucl. Phys. B **452**, 677 (1995).
[11] R. Easther, K. Maeda, and D. Wands, Phys. Rev. D **53**, 4247 (1996).
[12] N. Kaloper, R. Madden, and K. A. Olive, Phys. Lett. B **371**, 34 (1996).
[13] A. P. Billyard, A. A. Coley, and J. E. Lidsey, Phys. Rev. D **59**, 123505 (1999).
[14] M. P. Ryan and L. S. Shepley, *Homogeneous Relativistic Cosmologies* (Princeton University Press, Princeton, 1975).
[15] M. Gasperini and G. Veneziano, Astropart. Phys. **1**, 317 (1992).
[16] N. Kaloper, Phys. Rev. D **55**, 3394 (1997).
[17] A. A. Coley, J. Ibáñez, and R. J. van den Hoogen, J. Math. Phys. **38**, 5256 (1997); A. P. Billyard, A. A. Coley, and J. Ibáñez, Phys. Rev. D **59**, 023507 (1999).
[18] A. R. Liddle, Phys. Lett. B **220**, 502 (1989).
[19] J. E. Lidsey, Gen. Relativ. Gravit. **25**, 399 (1993).
[20] J. E. Lidsey and I. Waga, Phys. Rev. D **51**, 444 (1995).
[21] U. S. Nilsson and C. Uggla, J. Math. Phys. **38**, 2611 (1997).
[22] J. Wainwright and G. F. R. Ellis, *Dynamical Systems in Cosmology* (Cambridge University Press, Cambridge, 1997).
[23] D. Hobill, A. Burd, and A. Coley, *Deterministic Chaos in General Relativity* (Plenum, New York, 1994).
[24] G. F. R. Ellis and M. A. H. MacCallum, Commun. Math. Phys. **12**, 108 (1969).
[25] V. G. LeBlanc, D. Kerr, and J. Wainwright, Class. Quantum Grav. **12**, 513 (1995).
[26] J. D. Barrow and J. Levin, Phys. Rev. Lett. **80**, 656 (1998).
[27] R. Carretero-Gonzalez, H. N. Nunez-Yepez, and A. L. Salas Brito, Phys. Lett. A **188**, 48 (1994).
[28] A. P. Billyard, Ph.D. thesis, Department of Physics, Dalhousie University, 1999.
[29] J. Scherk and J. H. Schwarz, Nucl. Phys. **B153**, 61 (1979).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Quantum scalar field in *D*-dimensional static black hole space–times

Daniele Binosi[a)]
Dipartimento di Fisica, Università di Trento, Italy

Sergio Zerbini[b)]
Dipartimento di Fisica, Università di Trento, Italy
and Istituto Nazionale di Fisica Nucleare, Gruppo Collegato di Trento, Italy

(Received 1 March 1999; accepted for publication 14 July 1999)

An Euclidean approach for investigating quantum aspects of a scalar field living on a class of *D*-dimensional static black hole space–times, including the extremal ones, is reviewed. The approach makes use of a near-horizon approximation of the metric and ζ-function formalism for evaluating the partition function and the expectation value of the field $\langle \phi^2(x) \rangle$. After a review of the nonextreme black hole case, the extreme one is considered in some detail. In this case, there is no conical singularity, but the finite imaginary time compactification introduces a cusp singularity. It is found that the ζ-function regularized partition function can be defined, and the vacuum expectation value of the field, is finite on the horizon, as soon as the cusp singularity is absent, namely, the manifold is smooth and the corresponding temperature is $T=0$. It is suggested that the requirement of having a smooth near-horizon geometry always selects the correct black hole equilibrium temperature. © *1999 American Institute of Physics.* [S0022-2488(99)03910-9]

I. INTRODUCTION

The issue of determining the equilibrium (Unruh–Hawking) temperature of a black hole, is important. In fact, one can extract thermodynamical information from its knowledge, for example, the Bekenstein–Hawking entropy (i.e., the tree-level contribution to the entropy) can be defined as the response of the free energy of the black hole to the change of this equilibrium temperature. Furthermore, it defines the admissible temperatures of thermal states of free scalar fields in a static and globally hyperbolic space–time region with horizons.

As is well known, there exist several methods for evaluating the possible equilibrium temperature of a stationary black hole. Within the simplest of these methods, one has to make a Wick rotation of the time coordinate (passing in this way to the Euclidean time $\tau=it$), and eliminate all the metric (conical) singularities connected to the horizon by an opportune choice of the time periodicity β_M.[1] Then, one has to impose the KMS condition for thermal states,[2–5] i.e., to impose the periodicity condition on the imaginary time dependence of the thermal Wightman functions, and interpret the common period β_T as the inverse of the temperature T of the state. Although this procedure determines the correct Unruh–Hawking temperature in the case of a nonextreme black hole, it does not apply to the extreme case (for example, to the case of an extreme Reissner–Nordström black hole), since one is unable to determine the time periodicity of the manifold β_M.

Later, a more sophisticated Lorentzian method was introduced in Ref. 6 and successively developed in Refs. 7–9. Without entering in the details of this approach, we only recall that the method is connected to the well known Hadamard expansion of the two-point Green functions in a curved background and in the limit of coincidence of the arguments. Basically in Ref. 6 was proved that assuming fairly standard axioms of quantum (quasifree) field theory (such as local

[a)]Electronic mail: binosi@alpha.science.unitn.it
[b)]Electronic mail: zerbini@science.unitn.it

definiteness and local stability in a stationary space–time region), in the scaling limit, the thermal Wightman functions in the interior of this region will transform into nonthermal and massless Wightman functions in Minkowski space–time. This point coincidence behavior of the Wightman functions, must hold for any physically sensible state (thermal or not), and, in the case that the space–time region one is dealing with is just a part of the whole manifold separated by event horizons, it must hold on the horizons. This constraint actually selects the correct temperatures $T=\beta_T^{-1}=\beta_M^{-1}$, in the case of Rindler and Schwarzschild space–times. Then these results have been generalized in Ref. 7 to a large class of space–times admitting an appropriate reflection isometry.

Both Haag's method[6] and Kay–Wald approach,[7] which, as they stand, work only for space–times with an intersection between past and future horizons, were extended in Refs. 8, 9 for working in more physical situations than the ethernal black holes one, including the extreme case.

However, all of these ''Lorentzian'' methods involve a certain amount of calculations (for example the procedure developed in Ref. 9 requires the evaluation of all the possible geodesics which start from the horizon); for this fact, even if it is not difficult to foresee a possible generalization to, say, *D*-dimensional extreme black holes, the concrete computation does not appear an easy task. In this paper, making use of Euclidean approach, we would like to obtain some information about the equilibrium temperature of a class of static, *D*-dimensional black holes, evaluating quantities such as the field fluctuations and the one-loop partition function. The inverse of the temperature is formally introduced as the period of the compactified imaginary time $0\leq\tau\leq\beta$.

In order to deal with explicit calculations, we will make use of a near-horizon approximation of the metric. This approximation may also be justified observing that only near the horizon interesting physical effects are supposed to be relevant.

First, the case of nonextreme black holes is reconsidered. Here, as is well known, a conical singularity is present. We will show that its presence leads to divergences of the vacuum expectation value of the field on the horizon. However in this case it is known that as soon as the smoothness of the manifold is required, the Hawking temperature, as well as the absence of the divergences, is recovered.

The analysis is extended to the extreme black holes case. In this case, our mathematical results are, no conical singularity is present, but the compactification of the imaginary time and the related periodic identification, induces an isometry containing parabolic elements (translation in τ), so that a cusp singularity appears. Its presence leads to the following features:

(1) The vacuum expectation value of the scalar field has divergences on the horizon;
(2) The global ζ-function, besides the horizon divergences, does not exist, and requires a further regularization.

These undesired features disappear as soon as the imaginary time period is taken to be ∞, namely the associated temperature is to be $T=0$, in agreement with the four-dimensional results obtained in Refs. 10, 9.

The main objection to the approach proposed here could be the use of a near-horizon approximation of the metric. However, we stress that in Ref. 6 only the limit form of the metric near the horizon was used in order to obtain the Unruh–Hawking temperature; moreover also the results in Refs. 10, 9 were derived in a near horizon approximation contest.

The paper is organized as follows: In Sec. II we review the evaluation of the vacuum expectation value of the field within the ζ-function regularization procedure, while in Sec. III we will derive a near-horizon approximation of the generic line element describing a nonextreme and an extreme black hole. Then in Secs. IV and V we will discuss in detail these two cases, taking advantage of the approximation done. The paper ends with some concluding remarks in Sec. VI.

II. EVALUATING THE VACUUM EXPECTATION VALUE OF THE FIELD WITHIN THE ζ-FUNCTION PROCEDURE

In this section, we evaluate the vacuum expectation value of the field $\langle\phi^2(x)\rangle$, in the framework of the ζ-function regularization procedure. Within this approach (see Ref. 11 for an exhaustive discussion) one has[12,13]

$$\langle\phi^2(x)\rangle=-\frac{2}{\sqrt{g(x)}}\frac{\delta S_{\mathrm{eff}}}{\delta J(x)}\bigg|_{J(x)\equiv 0}=\frac{1}{\sqrt{g(x)}}\frac{d}{ds}\left[\frac{\delta\zeta(s|Al^2)}{\delta J(x)}\right]\bigg|_{s=J(x)=0}, \tag{2.1}$$

where $J(x)$ is a classical source, and l is the usual arbitrary parameter (with the dimension of mass^{-1}) necessary from dimensional considerations.

By a direct calculation, it follows that the ζ-regularized field fluctuations turns out to be

$$\langle\phi^2(x)\rangle_{\mathrm{ren}}=l^2\frac{d}{ds}[s\zeta(s+1;x|Al^2)]\bigg|_{s=0}, \tag{2.2}$$

with the ζ-function evaluated when the source $J(x)$ vanishes.

By making use of the Laurent expansion of $\zeta(s+1;x|Al^2)$, and extracting from the ζ-function the l^2 dependence, we can rewrite (2.2) as

$$\langle\phi^2(x)\rangle_{\mathrm{ren}}=\lim_{s\to 0}\left[\zeta(s+1;x|A)-\frac{1}{s}\mathrm{Res}\ \ \zeta(s+1;x|A)-\mathrm{Res}\,\zeta(s+1;x|A)\ln l^2\right]. \tag{2.3}$$

Notice that when the manifold is smooth, the meromorphic structure of the ζ-function is known (Seeley's Theorem). In particular for a differential elliptic operator of the second order (Laplacian) one has

$$\Gamma(z)\zeta(z;\mathbf{x}|L_N)=\sum_{r=0}^{\infty}\frac{A_r(\mathbf{x}|L_N)}{z+r-\frac{N}{2}}+\text{analytic part},\quad A_r(\mathbf{x}|L_N)=\frac{a_r(\mathbf{x}|L_N)}{(4\pi)^{N/2}}. \tag{2.4}$$

The spectral coefficients $a_r(\mathbf{x}|L_N)$ are computable functions known as the Seeley–de Witt coefficients.

As a consequence, if the dimension of the smooth manifold is odd, the ζ-function is regular at $z=1$ $(s=0)$ and the dependence on the scale parameter l disappears and one gets

$$\langle\phi^2(x)\rangle_{\mathrm{ren}}=\zeta(1;x|A). \tag{2.5}$$

On the other hand, if the dimension is even, there is a simple pole at $z=1$, and the l ambiguity will be present.

III. NEAR-HORIZON APPROXIMATION OF THE METRIC

The metric for a general static spherically symmetric D-dimensional space–time, analytically continued into the Euclidean space, reads

$$ds^2=f(r)d\tau^2+\frac{1}{h(r)}dr^2+r^2d\Sigma_N^2,\quad x=(\tau,r,\mathbf{x}), \tag{3.1}$$

where $\tau=it$ is the Euclidean time, f and h are arbitrary functions of r (which are constant in the $r\to\infty$ limit, if the space–time has to be asymptotically flat), and $d\Sigma_N^2$ represent the line element of a smooth N-dimensional (transverse) manifold without boundary ($\mathbf{x}$ are the transverse coordinates).

For this metric representing a black hole, one demands the presence, at $r=r_+$, of a zero in both f and h, so that, according to the nature of this zero, one finds the following two interesting cases.

A. Nonextremal case

In the case of a nonextremal black hole, one has a simple zero at $r=r_+$, so that the functions f and h can be expanded as[14]

$$f(r)\simeq f'(r_+)(r-r_+), \quad h(r)\simeq h'(r_+)(r-r_+). \tag{3.2}$$

Thus, after changing to the coordinates (ρ, θ, $\mathbf{x}$) by means of

$$\rho^2=\frac{4}{h'(r_+)}(r-r_+), \quad \theta=\tfrac{1}{2}\sqrt{f'(r_+)h'(r_+)}\tau, \tag{3.3}$$

the geometry near the event horizon is described by the approximated line element

$$ds^2\simeq d\rho^2+\rho^2 d\theta^2+r_+^2 d\Sigma_N^2. \tag{3.4}$$

We may generalize the argument to black hole solutions in semiclassical gravity. In this case, near the horizon, one has

$$f(r)\simeq C_f(r-r_+)^{c_1}, \quad h(r)\simeq C_h(r-r_+)^{c_2}, \tag{3.5}$$

with the constants $C_f>0, C_h>0, c_1>0, 0<c_2<2$. Here c_1 may be less than one, and the first derivative may not exist at the horizon. However, if $c_1=2-c_2$, it is easy to show that by means of the following coordinates transformation

$$(r-r_+)^{c_1/2}=\frac{c_1}{2}\sqrt{C_h}\rho, \quad \theta=\frac{c_1}{2}\sqrt{C_hC_f}\tau, \tag{3.6}$$

the line black hole element reduces again to the line element (3.4). For example, the previous case corresponds to $c_1=c_2=1$, and very recently, in Ref. 15 the case $c_1=1/2$ and $c_2=3/2$ have been considered; in any case notice that since $c_2<2$, the proper radial distance to the horizon is finite.

Now, finite temperature effects are assumed to arise when the Euclidean time τ (correspondingly θ) is compactified requiring $0\leqslant\tau\leqslant\beta$ $(0\leqslant\theta\leqslant\gamma)$, with β the inverse of the temperature. So, for arbitrary β (γ), the manifold $\mathcal{M}^D$ shows, near the horizon, the topology of $\mathcal{C}_\gamma\times\Sigma^N$, $\mathcal{C}_\gamma$ being the simple two-dimensional flat cone with deficit angle $2\pi-\gamma$.

In such a space–time, one usually determines the temperature of the black hole, by requiring the absence of the conical singularity;[1] the manifold, in fact, is not smooth, showing a conical singularity at $\rho=0$ unless $\gamma=2\pi$. In this way the temperature is found to be

$$T=\frac{\sqrt{f'(r_+)h'(r_+)}}{4\pi}, \quad T=\frac{c_1}{4\pi}\sqrt{C_hC_f}, \tag{3.7}$$

respectively, which are the Unruh–Hawking temperatures of the black holes.

We will show that $\gamma=2\pi$ is the only possible requirements for having a well-behaved $\langle\phi^2(x)\rangle$ on the horizon.

B. Extreme case

In the case of an extreme black hole, one has a double zero at $r=r_+$, so that the behavior of f and h near the horizon, is[14]

$$f(r)\simeq \tfrac{1}{2}f''(r_+)(r-r_+)^2, \quad h(r)\simeq \tfrac{1}{2}h''(r_+)(r-r_+)^2. \tag{3.8}$$

Thus, if we define the new coordinates (ρ, θ, $\mathbf{x}$) by means of

$$\rho=\sqrt{\frac{2}{f''(r_+)}}(r-r_+)^{-1}, \quad \theta=\sqrt{\frac{h''(r_+)}{2}}\tau=\frac{\tau}{b}, \tag{3.9}$$

we get the approximated line element

$$ds^2\simeq\frac{b^2}{\rho^2}(d\rho^2+d\theta^2)+r_+^2 d\Sigma_N^2. \tag{3.10}$$

So, once the compactification in the Euclidean time is carried over, the manifold shows the topology $H^2/\Gamma\times\Sigma^N$, H^2 being the two-dimensional hyperbolic space, and Γ being the (discontinuous and fixed-point-free) group of isometry induced by the identification $\theta\sim\theta+n\gamma$.

Notice that in this case it is not possible to determine the temperature by using the method of the conical singularity, since no conical singularity is present. However, it does not seem correct to deduce from this fact that such a manifold admits any temperature.[16] In fact, the evaluation of the stress-energy tensor of a free quantized field,[10] which is found to be finite only for $T=0$, and the scaling argument of Ref. 9 suggest that $T=0$ is the only physical admissible temperature in the case of a four-dimensional extreme Reissner–Nordström black hole (which can be recovered by setting $d\Sigma_N^2=d\Omega_2$ in (3.10)). Further evidence in favor of this fact comes from the absence of the Hawking radiation in the extreme Reissner–Nordström black hole.[17]

Again we will see that when $\beta=\infty$, corresponding to $T=0$, one gets a well behaved $\langle\phi^2(x)\rangle$ on the horizon, and the related extreme black hole Euclidean manifold becomes smooth.

Without loss of generality, we set b^2 as well as r_+ equal to 1.

IV. FIRST CASE: $\mathcal{M}^D=\mathcal{C}_\gamma\times\Sigma^N$

As we mentioned in the Introduction, we now evaluate the expectation value of the squared of the scalar field, using the ζ-function regularization technique. We start by reviewing the evaluation of the heat kernel and the local ζ-function on $\mathcal{M}^D=\mathcal{C}_\gamma\times\Sigma^N$ (for a complete discussion, see Ref. 18).

We consider a massless and minimally coupled scalar field on $\mathcal{C}_\gamma$, so that the associated operator is the pure Laplacian $L_\gamma=-\nabla_\gamma=\partial_\rho^2+(1/\rho)\partial_\rho+(1/\rho^2)\partial_\theta^2$; the spectral properties of this operator are well known, and, in fact, a complete set of normalized eigenfunctions is easily found to be

$$\psi_{n\lambda}=\frac{1}{\sqrt{\gamma}}e^{(2\pi ni/\gamma)}J_{\nu_n}(\lambda\rho), \quad \nu_n=\frac{2\pi|n|}{\gamma}, \quad n\in Z, \tag{4.1}$$

together with its complex conjugate.

Here λ^2 ($\lambda\geqslant 0$) is the eigenvalue corresponding to ψ and ψ^*, while J_ν is the regular Bessel function. So, using the standard separation of variables, it is easy to get the spectrum and eigenfunctions of the operator $L_D=-\nabla_\gamma+L_N$ on $\mathcal{M}^D=\mathcal{C}_\gamma\times\Sigma^N$, L_N being a Laplace-type operator on Σ^N including, eventually, a mass and a scalar curvature coupling term. Moreover, since we suppose Σ^N an arbitrary smooth manifold without boundary, all known results concerning the heat kernel and the ζ-function for L_N on Σ^N (which we assume to be known) are applicable.

In particular the heat kernel has the usual asymptotic expansion (see also Sec. II),

$$K(t;\mathbf{x}|L_N)\simeq\sum_{r=0}^{\infty}A_r(\mathbf{x}|L_N)t^{r-(N/2)}, \tag{4.2}$$

and the meromorphic structure of the local ζ-function reads

$$\Gamma(s)\zeta(s;\mathbf{x}|L_N)=\sum_{r=0}^{\infty}\frac{A_r(\mathbf{x}|L_N)}{s+r-\frac{N}{2}}+J(s;\mathbf{x}|L_N), \tag{4.3}$$

where $J(s;\mathbf{x}|L_N)$ is the (generally unknown) analytic part. Here we have supposed the absence of zero modes, but one can easily take them into account with a simple modification of the formulas.

We can now derive the meromorphic structure of $\zeta_\gamma(s;x|L_D)$ on $\mathcal{M}^D=\mathcal{C}_\gamma\times\Sigma^N$. To this aim, one can use the factorization property of the heat kernel

$$K_\gamma(t;x|L_D)=K(t;\theta,\rho|L_\gamma)K(t;\mathbf{x}|L_D), \tag{4.4}$$

in which the heat kernels of the Laplace-type operators on $\mathcal{M}^D$, $\mathcal{C}_\gamma$, and Σ^N, respectively, appear. By taking the Mellin transform of (4.4), one usually gets the Dikii–Gelfand representation of the ζ-function, from which the meromorphic structure can be deduced.

Anyway in dealing with the conical manifold one has a convergence obstruction, in the meaning that there are no values of s for which the Mellin transform of (4.4) is a finite quantity. The solution to this problem has been suggested by Cheeger,[19] and simply consist in a separation between higher and lower eigenvalues. In practice we split the sum which appears in the heat kernel (and in the related ζ-function) in two sums, the first over the lower eigenvalues, and the second over the higher ones; then, after the analytic continuation is performed, one may define the full ζ-function by summing up the two contributions obtained in this way (of course such a definition has all the requested properties and coincides with the usual one if the manifold is smooth).

So we set

$$\zeta_<(s;x|L_D)=\int_0^\infty dt t^{s-1}K_<(t;\theta,\rho|L_\gamma)K(t;\mathbf{x}|L_N), \tag{4.5}$$

$$\zeta_>(s;x|L_D)=\int_0^\infty dt t^{s-1}K_>(t;\theta,\rho|L_\gamma)K(t;\mathbf{x}|L_N), \tag{4.6}$$

where $K_<(t;\theta,\rho|L_D)$ and $K_>(t;\theta,\rho|L_D)$ are, respectively, the ''lower'' and the ''higher'' heat kernels, which are related to the corresponding ζ-function by the relations

$$K_<(t;\theta,\rho|L_\gamma)=\frac{1}{2\pi i}\int_{1/2<\mathrm{Re}(s)<1} ds t^{-s}\Gamma(s)\zeta_<(s;\theta,\rho|L_\gamma), \tag{4.7}$$

$$K_>(t;\theta,\rho|L_\gamma)=\frac{1}{2\pi i}\int_{1/2<\mathrm{Re}(s)<1+\nu_1} ds t^{-s}\Gamma(s)\zeta_>(s;\theta,\rho|L_\gamma), \tag{4.8}$$

$$\zeta_<(s;\theta,\rho|L_\gamma)=\frac{1}{\Gamma(s)}\int_0^\infty dt t^{s-1}K_<(t;\theta,\rho|L_\gamma),\quad \tfrac{1}{2}<\mathrm{Re}(s)<1, \tag{4.9}$$

$$\zeta_>(s;\theta,\rho|L_\gamma)=\frac{1}{\Gamma(s)}\int_0^\infty dt t^{s-1}K_>(t;\theta,\rho|L_\gamma),\quad \tfrac{1}{2}<\mathrm{Re}(s)<1+\nu_1, \tag{4.10}$$

and, by definition,

$$K_\gamma(t;\theta,\rho|L_\gamma)=K_<(t;\theta,\rho|L_\gamma)+K_>(t;\theta,\rho|L_\gamma), \tag{4.11}$$

$$\zeta_\gamma(s;\theta,\rho|L_\gamma)=\zeta_<(s;\theta,\rho|L_\gamma)+\zeta_>(s;\theta,\rho|L_\gamma). \tag{4.12}$$

Now, making use of the Mellin–Parseval identity and paying attention to the range of convergence, one gets, for $\mathrm{Re}(s)>1+(N/2)$, the following representation:[18]

$$\zeta_\gamma(s;x|L_D)\simeq\frac{\zeta(s-1;\mathbf{x}|L_N)}{2\gamma(s-1)}+\frac{1}{\gamma\Gamma(s)}\sum_{r=0}^{P}A_r(\mathbf{x}|L_N)I_\gamma\left(s+r-\frac{N}{2}\right)\rho^{2s+2r-D}+\mathcal{O}(\rho^{2s+2P-D}),\tag{4.13}$$

where P is an arbitrary large integer, while

$$I_\gamma(s)=\frac{\Gamma(s-\frac{1}{2})}{\sqrt{\pi}}[G_\gamma(s)+G_{2\pi}(s)].\tag{4.14}$$

For $\mathrm{Re}(s)>1$,

$$G_\gamma(s)=\sum_{n=1}^{\infty}\frac{\Gamma(\nu_n-s+1)}{\Gamma(\nu_n+s)},\quad G_{2\pi}=-\frac{\Gamma(1-s)}{2\Gamma(s)}.\tag{4.15}$$

It is possible to show that $G_\gamma(s)$ admits an analytical continuation, and the properties of I_γ and G_γ on the whole complex plane are studied in detail in the Appendix of Ref. 18; an important property is that the analytical continued I_γ as well as G_γ, has only a simple pole at $s=1$, with residue

$$\mathrm{Res}\, I_\gamma(s)|_{s=1}=\frac{1}{2}\left(\frac{\gamma}{2\pi}-1\right).\tag{4.16}$$

Having found the meromorphic structure of the ζ-function on our manifold, we can determine the vacuum expectation value of the fluctuation of a scalar field. With regard to this, it is convenient to distinguish between odd- and even-dimensional space-times.

We first consider the case in which N (or, equivalently, D) is odd, so that I_γ is finite at $s=1$. To begin with, notice that the first term in (4.13) depends only on the transverse coordinates and is finite on the horizon. As a result, making use of the meromorphic structure of $\zeta(s;\mathbf{x}|L_N)$, we get

$$\langle\phi^2(x)\rangle\simeq\frac{1}{2\gamma}\left[\sum_{r=0}^{\infty}\frac{A_r(\mathbf{x}|L_N)}{r-\frac{N}{2}}+J(0;\mathbf{x}|L_n)\right]+\frac{1}{\gamma}\sum_{r=0}^{P}A_r(\mathbf{x}|L_N)I_\gamma\left(1+r-\frac{N}{2}\right)\rho^{2r-N}+\mathcal{O}(\rho^{2P-N}).\tag{4.17}$$

It is now easy to see that the above expression contains $[N/2]$ (where $[\,]$ means ‘‘integer part’’) terms which are divergent as ρ^{2r-N} $(r<[N/2])$ in the limit $\rho\to0$, and so on the horizon (see (3.3)). Thus, if we want a good behavior on it, we must demand that all the I_γ's vanish for $r<[N/2]$, i.e., $\gamma=2\pi$. In particular notice that within this value of γ all of the I_γ's actually vanish.

We now come to the case in which $N(D)$ is even. In this case, the ζ-function (4.13) has a pole at $s=1$, coming from the first and the I_γ term in (4.13). From (2.3) one gets

$$\langle\phi^2(x)\rangle_{\mathrm{ren}}\simeq\frac{1}{2\gamma}\left[\sum_{r=0}^{\infty}{}'\,\frac{A_r(\mathbf{x}|L_N)}{r-\frac{N}{2}}+J(0;\mathbf{x}|L_n)\right]-\frac{1}{4\pi}A_{N/2}(\mathbf{x}|L_N)\ln\mu^2$$
$$+\frac{1}{\gamma}\sum_{r=0}^{P}{}'\,A_r(\mathbf{x}|L_N)I_\gamma\left(1+r-\frac{N}{2}\right)\rho^{2r-N}+\mathcal{O}(\rho^{2P-N}),\tag{4.18}$$

where the $'$ in the sums means omission of the $r=N/2$ term.

Again, as long as $r<N/2$, we get divergent terms on the horizon, unless we require the I_γ's to vanish, i.e., $\gamma=2\pi$ as in the odd-dimensional case.

As a result, for a manifold $\mathcal{M}^D$ whose near-horizon geometry is described by $\mathcal{C}_\gamma\times\Sigma^N$, the requirement of having a well behaved $\langle\phi^2(x)\rangle$ on the horizon, selects $\gamma=2\pi$, and so the Unruh–Hawking temperature, according to the conical singularity method. We are also reminded that the choice $\gamma=2\pi$ makes the manifold smooth, getting rid of the conical singularity otherwise present in $\rho=0$.

The computation of the partition function for arbitrary γ has been done in Ref. 18, and it has been used in order to discuss thermodynamical properties. Only the horizon divergences are present, and these are still present in the on-shell ($\gamma=2\pi$) entropy.

V. SECOND CASE: $\mathcal{M}^D=H^2/\Gamma\times\Sigma^N$

After having checked that our procedure works at least in the case of nonextreme black holes, we can tackle the case of the extreme ones, i.e., manifold whose topology near the horizon is described by $H^2/\Gamma\times\Sigma^N$.

Again, one can start by making use of the factorization property of the heat kernel, writing that

$$K_\gamma(t;x|L_D)=K(t;\theta,\rho|L_\gamma)K(t;\mathbf{x}|L_N), \tag{5.1}$$

where the heat kernels of the Laplace-type operators on $\mathcal{M}^D$, H^2/Γ, and Σ^N, respectively appear. As in the previous case, we suppose that L_γ is the operator associated with a massless and minimally coupled scalar field on H^2, while L_N is a Laplace-type operator on Σ^N including eventually, mass and scalar curvature coupling term (so that the expansions (4.2) and (4.3) are still valid). In this way, $L_\gamma=-\Delta_\gamma=-\rho^2(\partial_\theta^2+\partial_\rho^2)$, and a complete set of normalized eigenfunctions is easily found to be

$$\psi_{\lambda k}=\sqrt{\frac{y}{2\pi}}e^{ik\theta}K_{i\lambda}(|k|y), \tag{5.2}$$

together with its complex conjugate. Here λ^2 ($\lambda\geqslant0$) is the eigenvalue corresponding to ψ and ψ^*, while K_ν is the MacDonald function. Thus the spectral representation of the (off-diagonal) heat kernel associated with L_γ on H^2 reads (see, for example, Ref. 20)

$$K_{H^2}(t;\theta,\rho;\theta',\rho'|L_\gamma)=\frac{1}{2\pi}\int_0^\infty d\lambda\lambda\tanh(\pi\lambda)e^{-(\lambda^2+(1/4))t}P_{i\lambda-(1/2)}(\cosh\sigma), \tag{5.3}$$

where P is the associated Legendre function, while σ is the H^2 geodesic distance between (θ,ρ) and (θ',ρ').

As previously remarked, for studying thermal effects, one has to deal with the quotient space H^2/Γ, with Γ the (discontinuous and fixed-point-free) group of isometry induced by the time compactification $0\leqslant\theta\leqslant\gamma$. In our case we have translations, corresponding to parabolic elements. By applying the method of images, the diagonal heat kernel turns out to be

$$K(t;\theta,\rho|L_\gamma)=\frac{1}{2\pi}\int_0^\infty d\lambda\lambda\tanh(\pi\lambda)e^{-(\lambda^2+(1/4))t}$$
$$+\frac{1}{\pi}\sum_{n=1}^\infty\int_0^\infty d\lambda\lambda\tanh(\pi\lambda)e^{-(\lambda^2+(1/4))t}P_{i\lambda-(1/2)}(\cosh\sigma_n), \tag{5.4}$$

where now

$$\cosh\sigma_n = 1 + \frac{n^2\gamma^2}{2\rho^2}. \tag{5.5}$$

Let us show that the partition function does not exist, and requires, besides the horizon divergence regularization, a further regularization. The partition function is proportional to the first derivative at zero of the ζ-function, which may be defined by the Mellin transform of the heat kernel trace. The latter may be obtained integrating over the manifold coordinates. As a result,

$$\zeta(s|L_\gamma) = \frac{\gamma}{4\pi\epsilon}\frac{1}{s-1}\zeta\left(1-s\middle|L_N+\frac{1}{4}\right) + \frac{\gamma}{2\pi\epsilon}\int_0^\infty d\lambda\,\frac{\lambda}{1+e^{2\pi\lambda}}\zeta\left(s\middle|L_N+\lambda^2+\frac{1}{4}\right)$$
$$+\frac{1}{\sqrt{\pi}}\zeta_R(1+\delta)\zeta\left(s-\frac{1}{2}\middle|L_N+\frac{1}{4}\right)+\mathcal{O}(\delta), \tag{5.6}$$

where ζ_R is the Riemann ζ-function, and we have introduced the horizon cutoff ϵ in the identity contribution, and the cusp regularization parameter $\delta>0$ in the topological contribution. With regard to this, it should be noticed the divergence for $\delta=0$, which is usually present when one is dealing with parabolic elements.[21]

As far as the vacuum expectation value of the field is concerned, we only need the expression of the local ζ-function near the horizon. Thus with regard to the sum over n, we may apply the simplest version of the Euler–MacLaurin resummation formula, namely,

$$\sum_{n=1}^{\infty} f(n) = \int_1^\infty dx f(x) - \frac{1}{2}f(1) + \int_1^\infty dx\left(x-[x]-\frac{1}{2}\right)f'(x). \tag{5.7}$$

As a result, for large ρ,

$$\sum_{n=1}^{\infty} P_{i\lambda-(1/2)}(\cosh\sigma_n) = \frac{\rho}{\sqrt{2}\gamma}\frac{1}{\lambda\tanh(\pi\lambda)} + C(\lambda) + \mathcal{O}\left(\frac{\gamma}{\rho}\right), \tag{5.8}$$

where $C(\lambda)$ does not depend on ρ. Thus, the diagonal part of the heat-kernel may be rewritten as

$$K(t;\theta,\rho|L_\gamma) = \frac{1}{2\pi}\int_0^\infty d\lambda\,\lambda\tanh(\pi\lambda)e^{-(\lambda^2+(1/4))t} + \frac{\rho}{2\gamma\sqrt{2\pi t}}e^{-(t/4)} + \mathcal{O}(1) + \mathcal{O}\left(\frac{\gamma}{\rho}\right). \tag{5.9}$$

As a result, the related local ζ-function reads

$$\zeta_\gamma(s;x|L_D) = \zeta(s;\mathbf{x}|L_D) + \frac{\rho}{2\gamma\sqrt{2\pi}}\frac{\Gamma(s-\frac{1}{2})}{\Gamma(s)}\zeta\left(s-\frac{1}{2};\mathbf{x}\middle|L_N+\frac{1}{4}\right) + \mathcal{O}(1) + \mathcal{O}\left(\frac{\gamma}{\rho}\right). \tag{5.10}$$

The first term on the right-hand side of the above relation, is the local ζ-function associated with the smooth manifold $\mathcal{M}^D = H^2\times\Sigma^N$ and it depends only on the transverse coordinates $\mathbf{x}$. The other terms are the asymptotic contribution for large ρ. As a result, we have obtained the meromorphic structure of $\zeta(s;x|L_D)$ via the meromorphic structure of $\zeta_\gamma(s;\mathbf{x}|L_D)$ and $\zeta(s-\frac{1}{2};\mathbf{x}|L_N+\frac{1}{4})$. It should be noticed that the second term, which is the relic of the sum over images, contains the shift $s-1/2$. This means that, with regard to the evaluation of the vacuum expectation value of the field, one has a simple pole at $s=1$ for any D, not only for D even. This violation of Seeley's Theorem is related to the presence of parabolic elements, which make the manifold singular.

On the other side, no matter the dimension of the space-time, the vacuum expectation value of the field contains terms proportional to ρ/γ which are divergent on the horizon ($\rho\to\infty$, see (3.9)), unless we demand $\gamma=\infty$, and so $T=0$, according to the result obtained in Refs. 9, 10 in the four-dimensional case. We finally notice that if $\gamma=\infty$, the partition function contains only the usual volume divergence associated with the noncompact nature of the Euclidean section.

VI. CONCLUSIONS

In this paper, making use of an Euclidean approach, we have evaluated the expectation value and the global zeta function, whose derivative with respect to s evaluated at $s=0$ gives the partition function, of a scalar field in a D-dimensional static black hole. A near-horizon approximation has lead to quite explicit expressions for the local ζ-function related to such a field propagating in the Euclidean section of the black hole space–time. The period of the compactified imaginary time has been interpreted as the inverse of the temperature. The fact is then stressed that this quantity is finite on the horizon as soon as one selects a distinguished period of the imaginary time (temperature), which coincides with the Hawking–Unruh temperature in the nonextreme case, and with the zero temperature in the extreme one. With regard to the partition function, no problem exists in the conical singularity case, while in the presence of the cusp singularity, a new divergence shows up in the global ζ-function, and, strictly speaking, the partition function does not exist. This drawback disappears if the cusp singularity is absent, the near-horizon geometry is smooth and namely the temperature is again zero.

With regard to the local quantities, we also notice that the regular behavior of the vacuum expectation value of the fields on the horizon, leads also to the regular behavior of the expectation value of the stress tensor. This can be verified by means of a direct calculation, starting from the local off-diagonal ζ-function which can be obtained with our approach.

As far as the extreme black holes are concerned, all the properties we have been deriving, and the lack of the Hawking radiation,[17] strongly suggest that the only admissible temperature is the zero one, in agreement with the four-dimensional case studied in Ref. 10. As a consequence, within the Euclidean approach, requiring the smoothness of the near-horizon geometry seems to select, in general, the correct Hawking temperature. The only class of space–times for which our analysis seems to have no direct application is the one in which the double zero occurs at $r_+=0$. As an example, we may recall the so called massless ground state of the asymptotically AdS toroidal black holes.[22–25] In fact these black holes have

$$f(r)=\frac{1}{h(r)}=\left(\frac{l^2}{r^2}-\frac{C_D M}{r^{D-1}}\right), \tag{6.1}$$

where C_D is a constant, M is the mass of the black hole, and the parameter l is related to the cosmological constant, namely, $\Lambda=-l^{-2}$. For $D=3$, one recovers the celebrated BTZ black hole.[26] The ground state of this class of black holes is the zero mass solution, and the Euclidean metric becomes

$$ds^2=\frac{r^2}{l^2}d\tau^2+\frac{l^2}{r^2}dr^2+r^2 dT_N^2, \tag{6.2}$$

where dT_N^2 represents the metric of an N-dimensional torus. In the above metric, $r=0$ is a naked coordinate singularity. If one compactifies the Euclidean time and makes the coordinate transformation $r=l^2/\rho$, one gets

$$ds^2=\frac{l^2}{\rho^2}[d\rho^2+d\tau^2+l^2 dT_N^2]. \tag{6.3}$$

This metric describes locally the D-dimensional hyperbolic space H^D.

For the zero temperature case and in $D=3$ and $D=4$, the vacuum expectation value of the field, as well as the expectation value of the stress tensor has been computed in Refs. 27, 21, and 28, respectively, and divergences have been found as ρ goes to infinity. In this case, our analysis does not select any distinguished temperature. However, it should be noticed that in this case, it is not reasonable to neglect the back-reaction effects. In fact, in Refs. 27, 21, 28 it has been shown that there is a quantum implementation of the Cosmic Censorship Principle due to the back-reaction on the metric.

ACKNOWLEDGMENTS

Useful discussions with V. Moretti and L. Vanzo are gratefully acknowledged.

[1] C. W. Gibbons and S. W. Hawking, Commun. Math. Phys. **15**, 2752 (1977).
[2] R. Kubo, J. Math. Soc. Jap. **12**, 570 (1957).
[3] P. C. Martin and J. Schwinger, Phys. Rev. **115**, 1342 (1959).
[4] R. Haag, N. M. Hugenholtz, and M. Winnink, Commun. Math. Phys. **5**, 215 (1967).
[5] R. Haag, *Local Quantum Physics* (Springer, Berlin, 1992).
[6] R. Haag, H. Narnhofer, and U. Stein, Commun. Math. Phys. **94**, 219 (1984).
[7] B. S. Kay and R. M. Wald, Phys. Rep. **207**, 49 (1991).
[8] H. Hessling, Nucl. Phys. B **415**, 243 (1994).
[9] V. Moretti, Class. Quantum Grav. **13**, 985 (1996).
[10] P. R. Anderson, W. A. Hiscock, and D. J. Loranz, Phys. Rev. Lett. **74**, 4365 (1995).
[11] V. Moretti, Commun. Math. Phys. **201**, 327 (1999).
[12] G. Cognola, L. Vanzo, and S. Zerbini, Phys. Lett. B **241**, 381 (1990).
[13] D. Iellici and V. Moretti, Phys. Lett. B **425**, 33 (1998).
[14] A. E. Mayo and J. D. Bekenstein, Phys. Rev. D **54**, 5059 (1996).
[15] P. R. Anderson and C. D. Mull, Phys. Rev. D **59**, 044007 (1999).
[16] S. W. Hawking, G. T. Horowitz, and S. F. Ross, Phys. Rev. D **51**, 4302 (1995).
[17] L. Vanzo, Phys. Rev. D **55**, 2192 (1997).
[18] S. Zerbini, G. Cognola, and L. Vanzo, Phys. Rev. D **54**, 2699 (1996).
[19] J. Cheeger, J. Diff. Geom. **18**, 575 (1983).
[20] A. A. Bytsenko, G. Cognola, L. Vanzo, and S. Zerbini, Phys. Rep. **266**, 1 (1996).
[21] D. Binosi, V. Moretti, L. Vanzo, and S. Zerbini, Phys. Rev. D **59**, 104017 (1996).
[22] L. Vanzo, Phys. Rev. D **56**, 6475 (1997).
[23] R. B. Mann, Class. Quantum Grav. **14**, 109 (1997).
[24] D. R. Brill, J. Louko, and P. Peldan, Phys. Rev. D **56**, 3600 (1997).
[25] D. Birmingham, Class. Quantum Grav. **16**, 1197 (1999).
[26] M. Banados, C. Teitelboin, and J. Zanelli, Phys. Rev. Lett. **69**, 1849 (1992).
[27] G. Lyfshitz and M. Ortiz, Phys. Rev. D **49**, 1929 (1994).
[28] M. Caldarelli, Phys. Rev. D **58**, 4008 (1998).

Linearization stability of the Einstein equation for Robertson–Walker models. I

Lluís Bruna
Departament de Física Aplicada, E.T.S. d'Enginyers de Telecomunicacions, Universitat Politècnica de Catalunya, C/ Jordi Girona s/n, 08034-Barcelona, Spain

Joan Girbau
Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

(Received 24 November 1998; accepted for publication 15 January 1999)

This is the first part of a series of two papers. In this article we study the linearization stability of the Einstein equation in the presence of matter. We have slightly changed the classic definition of this concept for the vacuum spacetime and a more general one adapted to our case is given. We consider a Robertson–Walker model (V,g,T) where V stands for the spacetime, g for a Robertson–Walker metric, and T for a stress-energy tensor of a perfect fluid. We write $V=S\times I$ where S is a spacelike hypersurface of V and I an $\mathbb{R}$-interval. We show that in the case S has a constant curvature K equal to 0, the Einstein equation $G(g)=\chi T$ is linearization stable at g. In a subsequent paper we shall prove that in the case $K=1$ the opposite occurs. The case $K=-1$ remains as an open question. © *1999 American Institute of Physics.* [S0022-2488(99)02109-X]

I. INTRODUCTION

This is the first part of a series of two papers. The aim of this series is to study the linearization stability of the Einstein equation when a Robertson–Walker model is considered. On this subject we have to cite D'Eath[1] who showed the existence of certain families of solutions to the Einstein equation in a small neighborhood of a Robertson–Walker metric and also discussed its linearization stability. Nevertheless, our initial point of view is different from that of D'Eath, and hence we can say that we are dealing with a very different problem. The question we consider is the following: interpret the universe as a Robertson–Walker model of metric g and stress-energy tensor T, related by the Einstein equation $G(g)=\chi T$, where $G(g)=\mathrm{Ric}(g)-\frac{1}{2}\mathrm{R}g$ is the Einstein tensor of g, and χ is a universal constant. Let δT be a small perturbation of the tensor T, originated, for example, by a distant supernova explosion (though T is the stress-energy tensor of a perfect fluid, $T+\delta T$ is not necessarily so). This perturbation of T gives raise to another perturbation δg of g. In order to study the propagation in the universe of the effects of that explosion, we wonder if it is admissible to deal with the linearized Einstein equation $(D_g G)(\delta g)=\chi\delta T$, instead of dealing with the true equation $G(g+\delta g)=\chi(T+\delta T)$. In other words, we want to know whether the Einstein equation is linearization stable at the initial Robertson–Walker metric g.

The linearization stability of the *empty space* Einstein equation has originated many research articles (in Section 3 of Refs. 2 and 3 one can find a brief description of the history of the subject, as well as the basic bibliography). Here we mention the works of Moncrief,[4,5] who established a necessary and sufficient condition for the linearization stability in terms of the geometry of the spacetime. However, not much literature has been published about the linearization stability of the Einstein equation *with matter*. We have to cite the article of Bao, Marsden, and Walton[6] who pointed out that linearization stability methods do not work in the presence of matter. The first difficulty we have encountered is the stability definition itself. The definition used by the authors that deal with vacuum spacetime is not entirely valid to study the linearization stability in the presence of matter. Although in Sec. IV of this paper we discuss in detail this subject, we now summarize the principal ideas we develop in that section.

Basically, the main problem of the linearization stability of the *empty space* Einstein equation is the following: Let $V=M\times I$ be a product manifold, where M is a 3-manifold and I an $\mathbb{R}$-interval. One starts from an initial Lorentz metric g on V that satisfies $G(g)=0$. Then one is interested in the metrics $g'=g+h$ close to g that satisfy $G(g')=0$. Now, in order to solve $G(g')=0$, it is common to linearize this equation, $(D_gG)(h)=0$, where $(D_gG)(h)$ means the linear term in h of $G(g+h)$, in such a way that $G(g+h)=G(g)+(D_gG)(h)+\cdots$. This simplification, however, only makes sense if each solution h of the linearized equation $(D_gG)(h)=0$ is tangent to a curve $\lambda\to g'(\lambda)$ of exact solutions of the true equation $G(g')=0$. That is, for any solution h of $(D_gG)(h)=0$ there exists a curve $g'(\lambda)$ such that $G(g'(\lambda))=0$, with $g'(\lambda)=g+h\lambda+\cdots$. When this is the case, one says that the Einstein equation is linearization stable at the initial metric g (see, for example, Ref. 7).

Let us see what happens in the presence of matter. For that purpose, we remember briefly the initial Einstein's point of view about the linearization of his equation in a small neighborhood of the Minkowski metric [see Refs. 8 and 9]. The Minkowski metric η of $\mathbb{R}^4$ satisfies the empty space Einstein equation $G(\eta)=0$. Now, let T' be a stress-energy tensor close to zero and look for the metrics g' close to η satisfying the Einstein equation with matter $G(g')=\chi T'$. Einstein solved the linearized equation $(D_\eta G)(h)=\chi T'$ of the previous equation and he wrote the solutions h (with vanishing asymptotic conditions) in terms of T'. But this analysis would be valid if for each solution h of $(D_\eta G)(h)=\chi T'$, $\eta+h$ were *close* to a solution g' of the true equation $G(g')=\chi T'$. Choquet-Bruhat and Deser,[10] and later Choquet-Bruhat, Fischer, and Marsden[11] showed that the Einstein equation is linearization stable at the Minkowski metric η (stable in the sense of the previous paragraph). Because of this, one can substitute the equation $G(g')=0$ by its linear one $(D_\eta G)(h)=0$. But this has nothing to do with Einstein's procedure because he does not solve the equation $(D_\eta G)(h)=0$, but he deals with $(D_\eta G)(h)=\chi T'$ for a small T'. However, we must say that this difficulty is somewhat formal because the results of Ref. 11 really do answer the question of whether the initial Einstein's point of view is valid or not, as Theorem 1 of Sec. IV assures.

At the end of Sec. IV we give a definition of linearization stability adapted to the matter case. Shortly speaking it says that the Einstein equation $G(g)=\chi T$ is linearization stable at the initial metric g in the direction of $\mathcal{F}$, $\mathcal{F}$ being a vector subspace of the space of stress-energy tensors, if for any $H\in\mathcal{F}$ *close* to zero and for any small solution h of the linear equation $(D_gG)(h)=\chi H$, $g+h$ is *close* to a solution g' of the true equation $G(g')=\chi(T+H)$ (later we will explain in detail what we mean by the word *close*). Note that if $\mathcal{F}=\{0\}$ and $T=0$ one obtains the old definition. If the Einstein equation is linearization stable at a given metric g in the direction of any subspace $\mathcal{F}$ of stress-energy tensors, we simply say that it is linearization stable.

Now we summarize the rest of the sections. Section I is devoted to the expression of the Einstein equation using Gauss coordinates and in Sec. II we deal with the Cauchy problem for the Einstein equation with matter. Finally, in Secs. V and VI we express the main result of this paper: Let $V=M\times I$ be a Robertson-Walker model (I an $\mathbb{R}$-interval and M a connected 3-Riemannian manifold of constant curvature K). We assume M to be simply connected. We show that if $K=0$ then the Einstein equation is linearization stable at the Robertson–Walker metric g. In a subsequent paper we will show that if $K>0$ there exists no vector subspace $\mathcal{F}$ of the space of stress-energy tensors for which the Einstein equation is linearization stable at g in the direction of $\mathcal{F}$. In the proof of our result for curvature $K=0$ we use a theorem of Cantor (Refs. 12 and 13) concerning the usual Laplace operator in $\mathbb{R}^3$. If there existed an analogous result for the De Rham Laplace operator (acting on 1-forms) associated to the metric of the 3-hyperbolic space of curvature -1 we would be able to adapt our proof of stability in the case $K=0$ to curvature $K<0$, but at present we do not know such a result.

II. THE EXPRESSION OF EINSTEIN TENSOR IN GAUSS COORDINATES

Let $(V,\tilde{g})$ be a time-orientable Lorentz manifold (we assume the signature of $\tilde{g}$ to be $+++-$). Let M be a spacelike hypersurface of V. To each $x\in M$ let N_x be the normal unit future-pointing timelike vector of M [this means that $\tilde{g}_x(N_x, N_x) = -1$]. Let $\varphi_t(x)$ be the geodesic at x with N_x as a tangent vector. We will assume that there exists a $\epsilon > 0$ so that the mapping

$$\Psi:\ M \times (-\epsilon, \epsilon) \to V$$

$$(x,t) \to \varphi_t(x)$$

gives a diffeomorphism between $M \times (-\epsilon, \epsilon)$ and a certain neighborhood V_M of M in V. This diffeomorphism is called the Gauss representation of V_M associated to the hypersurface M. We denote by t the coordinate of the interval $(-\epsilon, \epsilon)$. The notation $\partial/\partial t$ will stand for the field in V_M that corresponds to the field $\partial/\partial t$ of $M \times (-\epsilon, \epsilon)$ in the diffeomorphism Ψ. The field $\partial/\partial t$ when $t=0$ is orthogonal to M by construction. The well-known Gauss lemma assures that $\partial/\partial t$ is orthogonal to $M_t = \varphi_t(M)$ for all t.

We denote by g_t the restriction of $\tilde{g}$ to the hypersurface M_t, by ∇_t the connection on M_t associated to g_t, by Ric_t the Ricci tensor of g_t and by R_t the scalar curvature of g_t. Sometimes we will drop the subindex t of g_t, ∇_t, Ric_t and R_t assuming any M_t, without specifying t. We will write $\tilde{\nabla}$, $\widetilde{\mathrm{Ric}}$ and $\tilde{R}$ for the connection on V_M, the Ricci tensor and the scalar curvature of $\tilde{g}$, respectively. Let S_t be the second fundamental form of M_t; recall that this is defined by $S_t(X,Y) = -\tilde{g}(\tilde{\nabla}_X Y, \partial/\partial t)$ for X and Y vector fields on M_t. We write $k_t = 2S_t$.

Every local coordinate system (x^1, x^2, x^3) of M induces by means of the Gauss representation Ψ a coordinate system (x^1, x^2, x^3, t) in a certain open set of V_M; such systems will be called Gauss normal coordinates. Sometimes we write $x^4 = t$. Throughout this paper we assume that latin indices $i, j, k, \ldots$ run from 1 to 3, and greek indices $\alpha, \beta, \gamma, \ldots$ run from 1 to 4. With all these conventions in mind, the components $\tilde{R}_{\alpha\beta}$ of $\widetilde{\mathrm{Ric}}$ are expressed in Gauss normal coordinates in the following way:

$$\tilde{R}_{ij} = \frac{1}{2}\frac{\partial k_{ij}}{\partial t} - \frac{1}{2}(k \times k)_{ij} + R_{ij} + \frac{1}{4}(\mathrm{tr}_g\, k) k_{ij}, \tag{1}$$

$$\tilde{R}_{4i} = \tfrac{1}{2}\nabla^r (k_{ir} - (\mathrm{tr}_g\, k) g_{ir}), \tag{2}$$

$$\tilde{R}_{44} = -\frac{1}{2} g^{ij}\left(\frac{\partial k_{ij}}{\partial t} - \frac{1}{2}(k \times k)_{ij}\right) \tag{3}$$

(see, for instance Ref. 14). The expression $k \times k$ means $(k \times k)_{ij} = g^{rs} k_{ir} k_{js}$, or in other words, $(k \times k)(X,Y) = k(X, P(Y))$, where P is the associated endomorphism to k relative to the metric g. The expression $\mathrm{tr}_g\, k$ means the trace of k relative to the metric g, that is $g^{ij} k_{ij}$. Formulas (1), (2), and (3) are a consequence of the Gauss and Codazzi equations of the hypersurfaces M_t.

Let $\tilde{G}(\tilde{g})$ be the Einstein tensor of $\tilde{g}$ defined by $\tilde{G}(\tilde{g}) = \widetilde{\mathrm{Ric}} - (1/2)\tilde{R}\tilde{g}$. Later we will be interested in similar expressions to (1), (2), and (3) for the components $\tilde{G}_{4i}$ and $\tilde{G}_{44}$ of Einstein tensor. Bearing in mind that $\tilde{g}_{4i} = 0$ is fulfilled in Gauss coordinates, we get

$$\tilde{G}_{4i} = \tilde{R}_{4i} - \tfrac{1}{2}\tilde{R}\tilde{g}_{4i} = \tilde{R}_{4i}. \tag{4}$$

On the other hand, $\tilde{R} = \tilde{g}^{\alpha\beta}\tilde{R}_{\alpha\beta} = -\tilde{R}_{44} + g^{ij}\tilde{R}_{ij}$, and so by (1) and (3) one obtains

$$\tilde{G}_{44} = -\tfrac{1}{8} k \cdot k + \tfrac{1}{8}(\mathrm{tr}_g\, k)^2 + \tfrac{1}{2} R, \tag{5}$$

where $k \cdot k$ indicates the dot product of k by k given by the metric g, that is, $g^{ir} g^{js} k_{rs} k_{ij} = k^{ij} k_{ij}$.

III. THE CAUCHY PROBLEM FOR THE EINSTEIN EQUATION WITH MATTER

We want to study the Cauchy problem for the Einstein equation $\tilde{G}(\tilde{g})=\chi\tilde{T}$, where $\tilde{T}$ is the stress-energy tensor of matter and χ a universal constant.

Let V be a 4-manifold. Let M be a spacelike hypersurface of V. We assume the diffeomorphism

$$\Psi:\ M\times(-\epsilon,\epsilon)\rightarrow V_M,$$

is given, where V_M is a certain neighborhood of M in V. Besides M and the diffeomorphism Ψ, we will assume also that the following is given:

(a) a 2-covariant symmetric tensor field T on V_M, satisfying $i(\partial/\partial t)T=0$, where $i(\vec{v})$ indicates the inner contraction of $\vec{v}$ (later, we will choose the components of T in Gauss coordinates as the components $\tilde{T}_{ij}$ of the stress-energy tensor).
(b) A function F on M (later it will be chosen as $\tilde{T}_4^4$).
(c) A vector field X on M (chosen as $X^i=\tilde{T}_4^i$).
(d) A Riemannian metric g on M, and a 2-covariant symmetric tensor field k on M.

With this data we want to find a Lorentz metric $\tilde{g}$ and a 2-covariant symmetric tensor field $\tilde{T}$ on V_M (or possibly in a certain neighborhood of M on V_M), in such a way that the following conditions are fulfilled:

(1) $\tilde{G}(\tilde{g})=\chi\tilde{T}$.
(2) $\tilde{g}(\partial/\partial t,\partial/\partial t)=-1$.
(3) For all t, the hypersurface $M_t=\Psi(M\times\{t\})$ is orthogonal to $\partial/\partial t$ in the metric $\tilde{g}$.
(4) The restriction of $\tilde{g}$ to M is g.
(5) Twice the second quadratic form of M with respect to the metric $\tilde{g}$ is k.
(6) In Gauss normal coordinates the relations $\tilde{T}_{ij}=T_{ij}$ hold on V_M.
(7) In Gauss normal coordinates the relations $\tilde{T}_4^4=F$ and $\tilde{T}_4^i=X^i$ hold on M, where $\tilde{T}_4^\alpha$ means $\tilde{g}^{\alpha\beta}\tilde{T}_{4\beta}$.

In order to deal with this problem we begin by noting that conditions (2) and (3) are written $\tilde{g}_{44}=-1$ and $\tilde{g}_{4i}=0$ in Gauss normal coordinates. It is easy to see that both conditions force the field $\partial/\partial t$ to be geodesic and hence we can use the formulas of the last section. Condition (7) gives $\tilde{T}_{44}=-F$ and $\tilde{T}_{4i}=X_i$ because $\tilde{g}_{44}=-1$ and $\tilde{g}_{4i}=0$.

The metric $\tilde{g}$ we look for will be determined by the knowledge of the $g_t=\tilde{g}|_{M_t}$ for all t in an interval of the origin of $\mathbb{R}$ [because $\tilde{g}_{44}=-1$, $\tilde{g}_{4i}=0$, and $\tilde{g}_{ij}(x,t)=(g_t)_{ij}(x)$]. In a similar way the knowledge of the tensor $\tilde{T}$ we look for will be determined by the knowledge, for all t, of the function F_t defined on M_t by $F_t=\tilde{T}_4^4|_{M_t}$, and by the family of vector fields X_t on each M_t defined by $(X_t)^i(x)=\tilde{T}_4^i(x,t)$ (since the components $\tilde{T}_{ij}=T_{ij}$ are given).

The Einstein equation $\tilde{G}(\tilde{g})=\chi\tilde{T}$ can be written in the equivalent form:

$$\tilde{R}_{\alpha\beta}=\chi(\tilde{T}_{\alpha\beta}-\tfrac{1}{2}(\mathrm{tr}_{\tilde{g}}\tilde{T})\tilde{g}_{\alpha\beta}). \tag{6}$$

We will write Einstein equation in the form $\tilde{G}(\tilde{g})=\chi\tilde{T}$ for some components, whereas we will use the form (6) for others. Concretely, we will use

$$\begin{aligned}\tilde{G}_{4\alpha}&=\chi\tilde{T}_{4\alpha},\\ \tilde{R}_{ij}&=\chi(\tilde{T}_{ij}-\tfrac{1}{2}(\mathrm{tr}_{\tilde{g}}\tilde{T})\tilde{g}_{ij}).\end{aligned} \tag{7}$$

By means of (4), (5), and (2), the four equations in (7) for the $\tilde{G}_{4\alpha}$ can be written in the following way:

$$\begin{gathered} \tfrac{1}{8}(k\cdot k-(\mathrm{tr}_g\, k)^2-4R)=\chi F, \\ \tfrac{1}{2}\nabla^r(k_{ir}-(\mathrm{tr}_g\, k)g_{ir})=\chi X_i\,. \end{gathered} \tag{8}$$

This means that the Cauchy data (g,k,F,X) on M are not independent but must satisfy the constraints (8). Since $\tilde{T}$ equals a constant times the Einstein tensor, it must have zero divergence. Taking into account that in Gauss normal coordinates one has

$$\tilde{\Gamma}^k_{ij}=\Gamma^k_{ij}, \quad \tilde{\Gamma}^4_{4j}=0, \quad \tilde{\Gamma}^\alpha_{44}=0,$$

this fact is written

$$0=\partial_4\tilde{T}^{4\beta}+\partial_i\tilde{T}^{i\beta}+\tilde{\Gamma}^i_{ij}\tilde{T}^{j\beta}+\tilde{\Gamma}^i_{i4}\tilde{T}^{4\beta}+2\tilde{\Gamma}^\beta_{4i}\tilde{T}^{4i}+\tilde{\Gamma}^\beta_{ij}\tilde{T}^{ij}, \tag{9}$$

where ∂_α indicates $\partial/\partial x^\alpha$. Now, $\tilde{T}^{44}=\tilde{T}_{44}=-F$ and $\tilde{T}^{i4}=-\tilde{T}^i_4=-X^i$. Moreover, $\tilde{\Gamma}^4_{ij}=S_{ij}$ where $S_{ij}=k_{ij}/2$ is the second fundamental form of the hypersurfaces M_t. One has $\tilde{\Gamma}^r_{4j}=S^r_j$, too. Writing now condition (9) for $\beta=4$ one obtains

$$\partial_t F=-\mathrm{div}_g\,\chi-\tfrac{1}{2}(\mathrm{tr}_g\, k)F+\tfrac{1}{2}k\cdot T, \tag{10}$$

For $\beta=r$, condition (9) is written in a similar way

$$\partial_t X^r=\nabla_i T^{ir}-\tfrac{1}{2}(\mathrm{tr}_g\, k)X^r-k^r_i X^i. \tag{11}$$

The second equation of (7) is written, by (1),

$$\partial_t k_{ij}=\chi T_{ij}-\tfrac{1}{2}\chi(F+\mathrm{tr}_g\, T)g_{ij}+(k\times k)_{ij}-2R_{ij}-\tfrac{1}{2}(\mathrm{tr}_g\, k)k_{ij}\,. \tag{12}$$

It is well known that in Gauss coordinates one has

$$\partial_t g_{ij}=k_{ij}\,. \tag{13}$$

Equations (10), (11), (12), and (13) give the evolution of F_t, X_t, k_t, and g_t with respect to t, respectively. The system of equations (10), (11), (12), and (13) is similar to the one studied in Refs. 15 and 16. It is well-known the existence, for small t, of a unique solution (F_t,X_t,k_t,g_t) with initial conditions (F,X,k,g) on M (corresponding to $t=0$) that satisfy constraints (8). Also, it is well-known that the solution (F_t,X_t,k_t,g_t) satisfies then (8) for all t, and hence the solution (F_t,X_t,k_t,g_t) defines a metric $\tilde{g}$ and a tensor $\tilde{T}$ satisfying the Einstein equation.

Remark 1: The Cauchy problem splits into a problem of initial conditions (the search for Cauchy data satisfying $\tilde{G}_{4\alpha}=\chi\tilde{T}_{4\alpha}$ on M), and a problem of evolution (the integration of the equations $\tilde{G}_{ij}=\chi\tilde{T}_{ij}$ in V_M for the previous initial conditions). This is the reason why the components $\tilde{T}_{ij}$ have to be given on V_M, not only on M.

Remark 2: If we require the stress-energy tensor $\tilde{T}$ to be that of a perfect fluid with null pressure, that is, $\tilde{T}=\rho u\otimes u$, with u a timelike unit vector $(\tilde{g}(u,u)=-1)$, then $\tilde{T}_{ij}$ remains determined by the initial conditions (g,k). Indeed, write $u=r(\partial/\partial t)+v$ with v tangent to M_t. One has $u_i=\tilde{g}_{i\alpha}u^\alpha=g_{ij}u^j=g_{ij}v^j=v_i$, $u_4=\tilde{g}_{44}u^4=-r$. Then $X_i=\tilde{T}_{4i}=\rho u_4 u_i=-\rho r v_i$, $F=-\tilde{T}_{44}=-\rho r^2$. On the other hand, $\tilde{g}(u,u)=-1$, hence $g(v,v)=-1+r^2$. Then the 5 scalar equations

$$X_i=-\rho r v_i, \quad F=-\rho r^2, \quad g(v,v)=-1+r^2$$

determine the 5 variables ρ, r, and v_i in terms of F, X_i, and g. By (8) F and X are related to (g,k), and hence T_{ij} depends only on (g,k). The same occurs in a perfect fluid with pressure when a state equation relates both the pressure and density (see Ref. 15).

IV. REVISION OF THE CONCEPT OF LINEARIZATION STABILITY OF AN EQUATION

Before the definition of this concept we remember the initial Einstein's point of view with respect to the linearization of his equation in a small neighborhood of the Minkowski metric (see Refs. 8 and 9). Let

$$\tilde{G}(\tilde{g}')=\chi\tilde{T}' \tag{14}$$

be the Einstein equation. The metric $\tilde{g}'=\eta=\mathrm{Diag}[1,1,1,-1]$ in $\mathbb{R}^4$ satisfies this equation without matter $(\tilde{T}'=0)$. Let $\tilde{T}'$ be a small stress-energy tensor (components close to zero) and look for the metrics $\tilde{g}'$ close to η that fulfills (14). Since Einstein tensor has zero divergence, $\mathrm{div}_{\tilde{g}'}\tilde{T}'=0$ has to be satisfied for any $\tilde{g}'$ solution of (14). This is a condition that links $\tilde{T}'$ and the solutions $\tilde{g}'$ of (14) corresponding to this $\tilde{T}'$. Let $\tilde{g}'=\eta+\tilde{h}$ be a metric close to η. We design by ψ the 2-tensor defined in terms of $\tilde{h}$ by

$$\psi_{\alpha\beta}=\tilde{h}_{\alpha\beta}-\tfrac{1}{2}(\mathrm{tr}_\eta\tilde{h})\,\eta_{\alpha\beta}.$$

Einstein realized that if $\tilde{h}$ satisfies $\mathrm{div}_\eta\psi=0$, the linear equation in $\tilde{h}$ obtained from $\tilde{G}(\eta+\tilde{h})=\chi\tilde{T}'$ can be written with respect to ψ in the following form:

$$-\frac{1}{2}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial x_3^2}-\frac{\partial^2}{\partial x_4^2}\right)\psi_{\alpha\beta}=\chi\tilde{T}'_{\alpha\beta}. \tag{15}$$

Einstein also realized that for any $\tilde{h}$ there exists an appropriate *gauge* in which the condition $\mathrm{div}_\eta\psi=0$ is fulfilled, and therefore this supplementary condition is physically irrelevant. Then Einstein gave the solutions ψ of (15) vanishing at infinity in terms of $\tilde{T}'$ using the retarded-potential method. Briefly, given a tensor $\tilde{T}'$ close to zero for which (14) has a solution, he solves the linear equation $(d\tilde{G})_\eta(\tilde{h})=\chi\tilde{T}'$ instead of $\tilde{G}(\eta+\tilde{h})=\chi\tilde{T}'$ for the same $\tilde{T}'$.

In general, we are dealing with the following problem: we begin with spacetime $(V,\tilde{g})$ where the metric $\tilde{g}$ satisfies the Einstein equation $\tilde{G}(\tilde{g})=\chi\tilde{T}$ for a given stress-energy tensor $\tilde{T}$. Then we consider stress-energy tensors $\tilde{T}'=\tilde{T}+\delta\tilde{T}$ close to $\tilde{T}$ (small variations of $\tilde{T}$) and we look for metrics $\tilde{g}'=\tilde{g}+\tilde{h}$ close to $\tilde{g}$ satisfying (14). But instead of working with $\tilde{G}(\tilde{g}+\tilde{h})=\chi(\tilde{T}+\delta\tilde{T})$, we deal with $(d\tilde{G})_{\tilde{g}}(\tilde{h})=\chi\delta\tilde{T}$. Is it admissible ?

So let us express this idea in a more exact form using the results of the preceding section. Consider a spacelike hypersurface M of $(V,\tilde{g})$. The tensors $\tilde{T}'=\tilde{T}+\delta\tilde{T}$ close to $\tilde{T}$ for which (14) has a solution in a certain neighborhood of M in V can be given in the following form: (a) Give a diffeomorphism Ψ of $M\times(-\epsilon,\epsilon)$ on an open neighborhood V_M of M in V. (b) Give a vector field X' and a function F' on M close to the field $(\tilde{T})^i_4|_M$ and to the function $(\tilde{T})^4_4|_M$ on M, respectively. (c) Give a 2-symmetric tensor T' in V_M satisfying $i(\partial/\partial t)T'=0$, in such a way that T'_{ij} be close to $(\tilde{T})_{ij}$. With this data, any (g',k') solution of the equations obtained from (8) replacing g, k, ∇, R, F, and X by g', k', ∇', R', F', and X', defines a metric $\tilde{g}'$ in a neighborhood of M in V and a tensor $\tilde{T}'$ such that $\tilde{G}(\tilde{g}')=\chi\tilde{T}'$. Thus the only equations we have to study for the linearization stability of a given F' and X' are the ones in (8).

The notations $S_2(M)$, $\mathcal{F}(M)$, and $\mathcal{X}(M)$ will stand for the 2-covariant symmetric tensors spaces, function spaces and vector field spaces on M, respectively. Later we will specify the topology of these spaces as well as the derivability and asymptotic conditions on the coefficients. For the moment we assume that all these spaces are Banach spaces. Let $\mathcal{H}(g',k')$ be the first member of the first equation of (8), that is

$$\mathcal{H}(g',k')=(1/8)(k'\cdot_{g'}k'-(\mathrm{tr}_{g'}k')^2-4R').$$

Let $\gamma(g',k')$ be the 1-form on M given by the first member of the second equation of (8), that is,

$$\gamma(g',k')_i=(1/2)\nabla'^r(k'_{ir}-(\mathrm{tr}_{g'}k')g'_{ir}).$$

1-forms on M are related to vector fields by the metric g'. Consider the mapping

$$\Phi:\ S_2(M)\times S_2(M)\to\mathcal{F}(M)\times\mathcal{X}(M)$$

$$(g',k')\to(\mathcal{H}(g',k'),\gamma(g',k')).$$

[The mapping Φ is only defined on the open set $S_2(M)\times S_2(M)$ of the pairs (g',k') such that g' is positive defined.] The initial metric $\tilde{g}$ of V corresponds to an initial pair (g,k). Given an element $\mathcal{T}'=(F',X')$ of $\mathcal{F}(M)\times\mathcal{X}(M)$ close to $((\tilde{T})_4^4|_M,(\tilde{T})_4^i|_M)$, we want to solve the linear equation $(d\Phi)_{(g,k)}(h,K)=\chi(\mathcal{T}'-\mathcal{T})$ instead of the true equation $\Phi(g',k')=\chi\mathcal{T}'$ in a neighborhood of (g,k). Obviously, in order to be meaningful, the solutions of the linear equations would have to be close to those of the true equation. We now write this in terms of a mapping between two Banach spaces.

Hence, let E_1 and E_2 be two Banach spaces. Let $p_0\in E_1$. Let U be a neighborhood of p_0 in E_1 and f: $U\to E_2$ a continuously differentiable mapping. For all $q\in E_2$ close to the origin, we want to compare the set H_q of $p\in U$ that are solution of $f(p)-f(p_0)=q$ with the set L_q of $p\in U$ that are solution of the linear equation $(df)_{p_0}(p-p_0)=q$ and make this comparison only for those p close enough to p_0. With respect to this problem, we enunciate the following theorem, a consequence of the inverse function theorem:

Theorem 1: *Let E_1 and E_2 be two Banach spaces. Let $p_0\in E_1$. Let U be a neighborhood of p_0 in E_1 and let f: $U\to E_2$ be a continuously differentiable mapping. For any $q\in E_2$ let H_q be the set of $p\in U$ such that $f(p)-f(p_0)=q$ and L_q the set of $p\in E_1$ such that $(df)_{p_0}(p-p_0)=q$. Let L be the kernel of $(df)_{p_0}$. We assume that L has a splitting kernel, that is, it has a topological complement S in E_1 in such a way that $E_1=L\oplus S$ [this enables us to write every element of E_1 as a pair (x,y) with $x\in L$ and $y\in S$]. If the tangent linear mapping $(df)_{p_0}$: $E_1\to E_2$ is surjective, then there exists an open neighborhood U' of p_0 in U, an open neighborhood V of the origin in E_2, an open neighborhood W of the origin in L, a linear mapping α: $E_2\to S$ and a differentiable mapping β: $V\times W\to S$ such that for any $q\in V,H_q\cap U'$ is a differentiable submanifold of U' parametrized by*

$$\varphi_q:\ W\to H_q\cap U'$$

$$x\to p_0+(x,\beta(q,x))$$

and L_q is a linear submanifold of E_1 parametrized by

$$\psi_q:\ L\to L_q$$

$$x\to p_0+(x,\alpha(q)),$$

and if for any $q\in V$ we denote by $E_q(x)$ the error done when considering $\psi_q(x)\in L_q$ instead of $\varphi_q(x)\in H_q\cap U'$ [that is, $E_q(x)=\varphi_q(x)-\varphi_q(x)=(0,\beta(q,x)-\alpha(q))$], one can assure that

$$\lim_{(x,q)\to(0,0)} \frac{E_q(x)}{\sqrt{\|x\|^2+\|q\|^2}}=0. \tag{16}$$

In other words, the error $E_q(x)$ is a high-order infinitesimal with respect to the distance of $(x,q)\in L\times E_2$ to the origin.

This theorem (whose proof is given later) leads us to introduce the following two definitions.

Definition 1: Let f: $U\to E_2$ be a continuously differentiable mapping between an open set U of a Banach space E_1 and another Banach space E_2. Let $p_0\in U$. Let F be a closed vector subspace E_2. For any $q\in F$ we denote by $H_q=\{p\in U$ such that $f(p)-f(p_0)=q\}$ and by $L_q=\{p\in E_1$ such that $(df)_{p_0}(p-p_0)=q\}$. We say f is linearization stable at the initial point p_0 in the direction of F if $(df)_{p_0}$ has a splitting kernel L and there exists an open neighborhood U' of *p_0 in U, an open neighborhood V of the origin in F, an open neighborhood W of the origin in L, a linear mapping α: $F\to S$ and a differentiable mapping $\beta: V\times W\to S$ such that for any $q\in V, H_q\cap U'$ is a differentiable submanifold of U' parametrized by*

$$\varphi_q:\; W\to H_q\cap U'$$

$$x\to p_0+(x,\beta(q,x)),$$

L_q is a linear submanifold of E_1 parametrized by

$$\psi_q:\; L\to L_q$$

$$x\to p_0+(x,\alpha(q)),$$

and for any $q\in V$ the error $E_q(x)$ done in considering $\psi_q(x)\in L_q$ instead of $\varphi_q(x)\in H_q\cap U'$ is a high-order infinitesimal with respect to the distance of $(x,q)\in L\times F$ to the origin [that is, (16) is fulfilled].

Definition 2: Let f: $U\to E_2$ be a continuously differentiable mapping between an open set U of a Banach space E_1 and another Banach space E_2. Let $p_0\in U$. We say f is linearization stable at the initial point p_0 if f is linearization stable at p_0 in the direction of E_2 (that is, when the previous definition is fulfilled for $F=E_2$).

Remarks on these definitions: (1) The fact that a mapping *f* between an open set *U* of a Banach space E_1 and another Banach space E_2 is linearization stable at $p_0\in U$ in the direction of a subspace F of E_2 means that, instead of solving the equation $f(p_0+h)=f(p_0)+q$, for small $q\in F$ and $h\in E_1$, one can deal with the linearized equation $(df)_{p_0}(h)=q$ without making a serious error because by (16) the solutions of each one are close enough.

(2) With the previous definitions in mind, Theorem 1 can be rewritten in the following form: if $(df)_{p_0}$ is surjective and has a splitting kernel, then *f* is linearization stable at p_0.

(3) In the case of Einstein equation, the mapping whose linearization stability we want to study is

$$\Phi:\; S_2(M)\times S_2(M)\to\mathcal{F}(M)\times\mathcal{X}(M)$$

$$(g',k')\to(\mathcal{H}(g',k'),\gamma(g',k')),$$

which is the same mapping one encounters when dealing with *empty space* Einstein equation. Nevertheless, in the empty space case this mapping is studied in a neighborhood of a pair (g,k) such that $\Phi(g,k)=0$, whereas in the presence of matter Φ must be studied in a neighborhood of a pair (g,k) such that $\Phi(g,k)=\chi\mathcal{T}$.

(4) If the subspace F of E_2 is $F=\{0\}$, definition 1 agrees with that of Ref. 7. These authors required every solution of the linear equation $(df)_{p_0}(p-p_0)=0$ to be tangent to a curve $p(t)$ of exact solutions of $f(p)-f(p_0)=0$. Let us see that the definition of these authors can be deduced

from condition (16). For that purpose let $v=p-p_0$ be a solution of the linear equation $(df)_{p_0}(v)=0$ (v will be in L). Let $p'(t)$ be the the straight line in L_0 defined by $p'(t)=p_0+tv$. Using $E_1=L\oplus S$ the straight line $p'(t)$ is written as $p'(t)=p_0+(tv,0)$. Consider the curve $p(t)=p_0+(tv,\beta(0,tv))$ of H_0. We wonder if v is the tangent vector to this curve at the point of parameter $t=0$. So one has to show that $(d\beta(0,tv)/dt)_{t=0}=0$. First, notice that (16) implies that $E_0(0)=0$, and since $E_0(0)=(0,\beta(0,0)-\alpha(0))$ and $\alpha(0)=0$, one has $\beta(0,0)=0$. By Taylor expansion,

$$\beta(0,tv)=\left(\frac{d\beta(0,tv)}{dt}\right)_{t=0}t+t^2h(t).$$

Hence,

$$\lim_{t\to 0}\frac{E_0(tv)}{\|tv\|}=\lim_{t\to 0}\frac{\beta(0,tv)}{\|tv\|}=\pm\frac{1}{\|v\|}\left(\frac{d\beta(0,tv)}{dt}\right)_{t=0}.$$

Therefore, if $(d\beta(0,tv)/dt)_{t=0}$ were $\neq 0$ the previous limit would be $\neq 0$. On the other hand, condition (16) requires this limit to be zero. This completes the proof.

Finally, notice that if f is linearization stable at p_0 according to Definition 2, then, even though H_0 is tangent to L_0, H_q is not tangent to L_q for $q\neq 0$. However, condition (16) assures us that the error done in considering L_q instead of H_q is small.

Proof of the theorem: If $(df)_{p_0}$: $E_1=L\oplus S\to E_2$ is surjective, its restriction to S gives raise to an isomorphism from S to E_2. We denote by α: $E_2\to S$ the inverse isomorphism. For any $p\in E_1$, $p-p_0$ factorizes as $x+y$ with $x\in L$ and $y\in S$. For all $q\in E_2$ the submanifold $L_q=\{p\in E_1$ such that $(df)_{p_0}(p-p_0)=q\}$ is parametrized by ψ_q: $x\to p=p_0+x+\alpha(q)$, since

$$(df)_{p_0}(p-p_0)=(df)_{p_0}(x+\alpha(q))=((df)_{p_0}\alpha)(q)=q.$$

f being differentiable at p_0, we have $f(p_0+x+y)=f(p_0)+(df)_{p_0}(y)+\epsilon(x,y)$, with

$$\lim_{(x,y)\to(0,0)}\frac{\epsilon(x,y)}{\sqrt{\|x\|^2+\|y\|^2}}=0.$$

Consider the mapping

$$\gamma:\ E_1\to L\times E_2$$

$$p_0+x+y\to(x,f(p_0+x+y)-f(p_0)).$$

Since $(d\gamma)_{p_0}$ is an isomorphism, the inverse function theorem states the existence of a neighborhood U' of p_0 in E_1 and of a neighborhood $W\times V$ of (0, 0) in $L\times E_2$ in such a way that γ gives an isomorphism between U' and $W\times V$. Let

$$\gamma^{-1}:\ W\times V\to U'\subset E_1$$

$$(x,q)\to\gamma^{-1}(x,q)$$

be the inverse isomorphism. Obviously, for any $q\in V,\varphi_q$: $x\to\gamma^{-1}(x,q)$ is a parametrization of $H_q\cap U'$. Now, we claim that $\gamma^{-1}(x,q)$ is written as p_0+x+y with $y\in S$. Indeed, we have

$$\gamma^{-1}(x,q)=p_0+x'+y$$

with $x'\in L$ and $y\in S$. We have to show that $x'=x$. To do this we write

$$(x,q)=\gamma\gamma^{-1}(x,q)=\gamma(p_0+x'+y)=(x',f(p_0+x'+y)-f(p_0)).$$

This implies $x=x'$. In $\gamma^{-1}(x,q)=p_0+x+y$ the element $y\in S$ depends differentiably on x and q. We write $\gamma^{-1}(x,q)=p_0+x+\beta(q,x)$. This is the function β whose existence the theorem claims. Only (16) remains to be shown. Notice that

$$\gamma\varphi_q(x)=\gamma\gamma^{-1}(x,q)=(x,q)$$

and that

$$\gamma\psi_q(x)=\gamma(p_0+x+\alpha(q))=(x,f(p_0+x+\alpha(q))-f(p_0)).$$

Hence

$$\gamma\psi_q(x)-\gamma\varphi_q(x)=(0,f(p_0+x+\alpha(q))-f(p_0)-q).$$

Now,

$$f(p_0+x+\alpha(q))-f(p_0)-q=(df)_{p_0}(\alpha(q))+\epsilon(x,\alpha(q))-q.$$

But $(df)_{p_0}(\alpha(q))=q$ and thus $f(p_0+x+\alpha(q))-f(p_0)-q=\epsilon(x,\alpha(q))$. Therefore, we have

$$\|\gamma\varphi_q(x)-\gamma\psi_q(x)\|_{L\times E_2}=\|\epsilon(x,\alpha(q))\|_{E_2}.$$

Since the inverse function theorem states that the mapping

$$(d\gamma)^{-1}:\ W\times V\to\mathcal{L}(L\times E_2,E_1)$$

is continuous, given a $K>0$ there exists a neighborhood A of (0, 0) in $W\times V$ such that for any $z\in A$ one has $\|(d\gamma)^{-1}(z)\|<K$. If z_1 and z_2 belong to A, by the mean value theorem $\|\gamma^{-1}(z_1)-\gamma^{-1}(z_2)\|<K\|z_1-z_2\|$. Apply this when $z_1=\gamma\varphi_q(x)$ and $z_2=\gamma\psi_q(x)$. Suppose that x and q are close enough to the origins of L and E_2 in order that z_1 and z_2 belong to A. Then

$$\|\varphi_q(x)-\psi_q(x)\|<K\|\gamma\varphi_q(x)-\gamma\psi_q(x)\|=K\|\epsilon(x,\alpha(q))\|.$$

Now, let $C=1/\|\alpha\|$. Since $\|\alpha(q)\|\leqslant\|\alpha\|\|q\|$, we have

$$\frac{\|\varphi_q(x)-\psi_q(x)\|}{\sqrt{\|x\|^2+\|q\|^2}}<\frac{K\|\epsilon(x,\alpha(q))\|}{\sqrt{\|x\|^2+C^2\|\alpha(q)\|^2}}.$$

If $C^2\geqslant 1$ one has $\sqrt{\|x\|^2+\|\alpha(q)\|^2}\leqslant\sqrt{\|x\|^2+C^2\|\alpha(q)\|^2}$, and if $C^2<1$ then $C\sqrt{\|x\|^2+\|\alpha(q)\|^2}\leqslant\sqrt{\|x\|^2+C^2\|\alpha(q)\|^2}$. In both cases we get, finally

$$\frac{\|\varphi_q(x)-\psi_q(x)\|}{\sqrt{\|x\|^2+\|q\|^2}}<K'\frac{\|\epsilon(x,\alpha(q))\|}{\sqrt{\|x\|^2+\|\alpha(q)\|^2}}\to 0.$$

This ends the proof.

V. LINEARIZATION STABILITY IN ROBERTSON–WALKER MODELS: EXPOSITION OF RESULTS

(S,g) will stand for a 3-Riemannian manifold with constant curvature K. Let $V=S\times I$, where I is an $\mathbb{R}$-interval. A Lorentzian metric of the form $\tilde{g}=-dt^2+\zeta(t)^2g$ is given on V, where t is the coordinate of I. In V we consider a perfect fluid with $u=\partial/\partial t$ as a velocity field, so that the stress-energy tensor of the fluid has the form $\tilde{T}=(\rho+p)u^*\otimes u^*+p\tilde{g}$, where u^* is the 1-form associated to u by $\tilde{g}$. Obviously, the metric $\tilde{g}$ and the tensor $\tilde{T}$ are related by the Einstein equation $\tilde{G}(\tilde{g})=\chi\tilde{T}$, which gives a well-known relations between ρ, p, ζ and the curvature K of g. From these relations one obtains, in particular, that ρ and p depend only on t. The Lorentz manifold $(V,\tilde{g})$ with the stress-energy tensor $\tilde{T}$ is called a Robertson–Walker model. We suppose that the Riemannian manifold (S,g) is connected and simply connected, and hence S will be a sphere in the case $K>0$, the Euclidean space $\mathbb{R}^3$ in the case $K=0$ or the hyperbolic space of curvature K in the case $K<0$.

We wonder if the equation $\tilde{G}(\tilde{g}')=\chi\tilde{T}'$ can be linearizated at the initial metric $\tilde{g}$ corresponding to a Robertson–Walker model $(V,\tilde{g},\tilde{T})$. That is, given a perturbation $\delta\tilde{T}$ of a stress-energy tensor $\tilde{T}$, is it legitimate to deal with the linear equation $(d\tilde{G})_{\tilde{g}}(\tilde{\delta}g)=\chi\delta\tilde{T}$ instead of working with the true equation $\tilde{G}(\tilde{g}+\tilde{h})=\chi(\tilde{T}+\delta\tilde{T})$? In the next section we show that if the curvature K of (S,g) vanishes then the Einstein equation is linearization stable at the initial metric $\tilde{g}$ of Robertson–Walker. In a subsequent paper we will show that if $K>0$ there exists no vector subspace $\mathcal{F}$ of the space of stress-energy tensors for which the Einstein equation is linearization stable at g in the direction of $\mathcal{F}$.

VI. RESULTS FOR $K=0$

Let $\mathcal{C}_c^\infty$ be the space of functions C^∞ with compact support on $\mathbb{R}^3$. For any positive integer p consider the norm $|\ |_p$ on $\mathcal{C}_c^\infty$ defined by

$$|f|_p=\left(\left|\int_{\mathbb{R}^3} f(x)^p\, dx\right|\right)^{1/p}.$$

Let $\mathbb{R}^+=\{x\in\mathbb{R} \text{ with } x\geq 0\}$ and $\mathbb{Z}^+=\mathbb{Z}\cap\mathbb{R}^+$. For all $s\in\mathbb{Z}^+$, p positive integer and $\delta\in\mathbb{R}^+$ consider on $\mathcal{C}_c^\infty$ the norm $|\ |_{p,s,\delta}$ defined by

$$|f|_{p,s,\delta}=\sum_{|\alpha|\leq s}|(D^\alpha f)(x)(\sqrt{1+|x|^2})^{|\alpha|+\delta}|_p .$$

Let $\mathcal{F}^p_{s,\delta}$ be the completion of $\mathcal{C}_c^\infty$ with respect to $|\ |_{p,s,\delta}$ (weighted Sobolev spaces). See Refs. 12 and 13 for the properties of these spaces. We mention here that for $s>(3/p)+k$ one has $\mathcal{F}^p_{s,\delta}\subset\mathcal{C}^k$, where $\mathcal{C}^k$ is the space of functions of class $\mathcal{C}^k$ on $\mathbb{R}^3$. Also, when $s>3/p$ the pointwise multiplication of functions $(f,g)\to f\cdot g$ induces a continuous mapping

$$\mathcal{F}^p_{s,\delta}\times\mathcal{F}^p_{s,\delta}\to\mathcal{F}^p_{s,\delta}.$$

From now on we will always suppose that $s>3/p$ so that the previous property will be satisfied and one has $\mathcal{F}^p_{s,\delta}\subset\mathcal{C}^0$.

We denote by $S^p_{s,\delta}$ and by $\mathcal{X}^p_{s,\delta}$ the spaces of 2-covariant symmetric tensors over $\mathbb{R}^3$ and vector fields over $\mathbb{R}^3$, respectively, whose components are functions of $\mathcal{F}^p_{s,\delta}$.

We will suppose given in $\mathbb{R}^3$ a Riemannian metric g, a 2-covariant symmetric tensor k, a vector field X and a function F such that the data (g,k,F,X) satisfy (8). $\mathcal{F}^p_{s,\delta}(F)$ will stand for the set of continuous functions F' over $\mathbb{R}^3$ such that $F'-F\in\mathcal{F}^p_{s,\delta}$. Hence we have $\mathcal{F}^p_{s,\delta}(F)=F+\mathcal{F}^p_{s,\delta}$. The space $\mathcal{F}^p_{s,\delta}(F)$ may be topologized by declaring the mapping $f\to F+f$ from $\mathcal{F}^p_{s,\delta}$ to $\mathcal{F}^p_{s,\delta}(F)$ to be an homeomorphism.

Similarly, we define $\mathcal{R}^p_{s,\delta}(g)$, $S^p_{s,\delta}(k)$, and $\mathcal{X}^p_{s,\delta}(X)$ by the following manner: $\mathcal{R}^p_{s,\delta}(g)$ is the set of continuous Riemannian metrics g' of $\mathbb{R}^3$ such that $g'-g\in S^p_{s,\delta}$; $S^p_{s,\delta}(k)$ is the set of continuous 2-covariant symmetric tensors k' over $\mathbb{R}^3$ such that $k'-k\in S^p_{s,\delta}$; $\mathcal{X}^p_{s,\delta}(X)$ is the set of continuous vector fields X' over $\mathbb{R}^3$ such that $X'-X\in\mathcal{X}^p_{s,\delta}$.

Consider the mapping

$$\Phi:\ \mathcal{R}^p_{s,\delta}(g)\times S^p_{s-1,\delta+1}(k)\to\mathcal{F}^p_{s-2,\delta+2}(F)\times\mathcal{X}^p_{s-2,\delta+2}(X)$$

$$(g',k')\to(\mathcal{H}(g',k'),\gamma(g',k'))$$

where, as usual,

$$\mathcal{H}(g',k')=(1/8)(k'\cdot_{g'}k'-(\mathrm{tr}_{g'}\,k')^2-4R(g'))$$

$$\gamma(g',k')_i=(1/2)\nabla'^r(k'_{ir}-(\mathrm{tr}_{g'}\,k')g'_{ir}).$$

With these conventions in mind we can state the following theorem.

Theorem 2: *If* $p>3$, $s>(3/p)+2$ *and* $0\leqslant\delta<(3(p-1)/p)-2$ *and if the initial data* (g,k,F,X) *satisfy* $\mathrm{tr}_g\,k=\text{constant}$, $X=0$ *and g is the Euclidean metric then the differential of* Φ *at the point* (g,k)

$$D_{(g,k)}\Phi:\ S^p_{s,\delta}\times S^p_{s-1,\delta+1}\to\mathcal{F}^p_{s-2,\delta+2}\times\mathcal{X}^p_{s-2,\delta+2}$$

is surjective and has a splitting kernel.

Remark: This theorem is a generalization of that of Choquet-Bruhat, Fischer, and Marsden[11] in the presence of matter and its proof reduces to show that the statements and calculus of Ref. 11 still remain valid in this new situation. Since Ref. 11 always uses (explicit and implicit) the condition $\mathcal{H}(g,k)=0$, which now is not satisfied, we think is worth briefly summarizing again the calculus of Ref. 11 in our situation.

Proof of the theorem: We begin by evaluating the differentials $D_{(g,k)}\mathcal{H}$ and $D_{(g,k)}\gamma$. To do this we replace g' by $g+h$ and k' by $k+K$ in the expressions of $\mathcal{H}(g',k')$ and of $\gamma(g',k')$ and only keep the linear terms in h and K bearing in mind that the difference between the Christoffel symbols Γ'^r_{ij} of g' and the symbols Γ^r_{ij} of g is written

$$\Gamma'^r_{ij}-\Gamma^r_{ij}=\tfrac{1}{2}g'^{rs}(\nabla_i h_{js}+\nabla_j h_{si}-\nabla_s h_{ij}),$$

and hence the linear term in h of this expression is $(1/2)g^{rs}(\nabla_i h_{js}+\nabla_j h_{si}-\nabla_s h_{ij})$. Also, the linear term in h of the difference $R(g')-R(g)$ between the scalar curvatures of g' and g is written as

$$R(g')-R(g)\cong -h^{ij}R_{ij}+\nabla^i\nabla^s h_{is}-\nabla^s\nabla_s\,\mathrm{tr}\,h$$

(the sign $\cong$ means equality up to terms of order >1 in h and K). Then a straight-forward computation gives

$$(D_{(g,k)}\mathcal{H})(h,K)=\tfrac{1}{8}[-2h\cdot(k\times k)+2k\cdot K-2(\mathrm{tr}\,k)(\mathrm{tr}\,K)+2h\cdot k\,\mathrm{tr}\,k$$
$$+4h\cdot\mathrm{Ric}(g)-4\nabla^i\nabla^j h_{ij}+4\nabla^i\nabla_i\,\mathrm{tr}\,h],$$

$$(D_{(g,k)}\gamma)(h,K)_i=\tfrac{1}{2}[\nabla^s K_{is}-\partial_i(\mathrm{tr}\,K)-h^{rs}\nabla_r k_{is}+h^{rs}\nabla_i k_{rs}$$
$$+\tfrac{1}{2}(\nabla_i h_{rs})k^{rs}-(\nabla^s h_{sm})k^m_i+\tfrac{1}{2}\nabla^l(\mathrm{tr}\,h)k_{il}].$$

For each function τ on $\mathbb{R}^3$ and each vector field Y on $\mathbb{R}^3$ we consider $h=2g\tau$ and $K=L_Y g-(\mathrm{div}\,Y)g-\tau k+\tau(\mathrm{tr}\,k)g$ where $\mathrm{div}\,Y$ means $\nabla^i Y_i$. For the moment, we only deal with these pairs (h,K) that depends on (τ,Y). Replacing these h and K in $D_{(g,k)}\mathcal{H}$ and $D_{(g,k)}\gamma$ we get

$$(D_{(g,k)}\mathcal{H})(Y,\tau)=\tfrac{1}{8}(-6\tau k\cdot k+2k\cdot L_Y g+2\tau(\mathrm{tr}\,k)^2+8\tau R(g)-16\Delta\tau),$$

where $\Delta\tau$ means $-g^{ij}\nabla_i\nabla_j\tau$. Analogously

$$(D_{(g,k)}\gamma)(Y,\tau)_i=\tfrac{1}{2}(-(\delta L_Y g)_i-3\tau\nabla^s k_{is}-\tau\nabla_i\,\mathrm{tr}\,k),$$

where δ means the Hodge operator that assigns to each 2-covariant symmetric tensor ω the 1-form $(\delta\omega)_i = -\nabla^j \omega_{ji}$. Since the initial data satisfy (8), we can replace $(\operatorname{tr} k)^2$ by $k\cdot k - 4R - 8\chi F$ and $\nabla^s k_{is}$ by $2\chi X_i + \nabla_i \operatorname{tr} k$. Using now the hypothesis $\operatorname{tr} k = \text{constant}$ and $X=0$, we finally have

$$(D_{(g,k)}\mathcal{H})(Y,\tau) = \tfrac{1}{8}(-4\tau k\cdot k + 2k\cdot L_Y g - 16\chi\tau F - 16\Delta\tau)$$

$$(D_{(g,k)}\gamma)(Y,\tau)_i = -\tfrac{1}{2}(\delta L_Y g)_i .$$

As a consequence of a theorem of Cantor[13,12] for those p, s, δ satisfying the hypothesis of the theorem, the operators $Y \to \delta L_Y g$ of $\mathcal{X}^p_{s,\delta} \to \mathcal{X}^p_{s-2,\delta+2}$ and $\tau \to \Delta\tau$ of $\mathcal{F}^p_{s,\delta} \to \mathcal{F}^p_{s-2,\delta+2}$ are isomorphisms.

Now we are ready to show that $D_{(g,k)}\Phi$ is surjective. Given a $(f,Z) \in \mathcal{F}^p_{s-2,\delta+2} \times \mathcal{X}^p_{s-2,\delta+2}$ there exists a unique Y such that $(D_{(g,k)}\gamma)(Y,\tau) = -(1/2)\delta L_Y g = Z$. Now, we look for a function τ such that

$$\tfrac{1}{8}(-4\tau k\cdot k + 2k\cdot L_Y g - 16\chi\tau F - 16\Delta\tau) = f.$$

or what is the same

$$-4\tau k\cdot k - 16\chi\tau F - 16\Delta\tau = 8f - 2k\cdot L_Y g.$$

Since Δ is an isomorphism, so is $-16\Delta - 4k\cdot k - 16\chi F$. Therefore there exists a unique τ such that $(D_{(g,k)}\mathcal{H})(\tau,Y) = f$. From this function τ and this field Y we form h and K given by $h = 2g\tau$ and $K = L_Y g - (\operatorname{div} Y)g - \tau k + \tau(\operatorname{tr} k)g$. Clearly, $(D_{(g,k)}\Phi)(h,K) = (f,Z)$. Therefore $D_{(g,k)}\Phi$ is surjective. To see that $D_{(g,k)}\Phi$ has a splitting kernel for any pair (h,K) we consider the unique pair (h',K') of the form $h' = 2g\tau$ and $K' = L_Y g - (\operatorname{div} Y)g - \tau k + \tau(\operatorname{tr} k)g$ such that $(D_{(g,k)}\Phi) \times (h,K) = (D_{(g,k)}\Phi)(h',K')$ [unique because $D_{(g,k)}\Phi$ is an isomorphism restricted to the pairs (h',K')]. Then

$$(h,K) = (h-h', K-K') + (h',K')$$

gives us the desired splitting. This completes the proof.

Let us apply theorem 2 to a $(V = S\times I, \tilde{g}, \tilde{T})$ Robertson–Walker model with curvature $K=0$. We assume that S is connected and simply connected. If the curvature K of (S,g) is zero then S is $\mathbb{R}^3$ and g the Euclidian metric. From the form of the metric $\tilde{g} = -dt^2 + \zeta(t)^2 g$ it follows that the vector field $\partial/\partial t$ is geodesic and therefore the Gauss representation of the hypersurface $M_0 = S \times \{0\}$ is the identity $S\times I \to S\times I$. The restriction of $\tilde{g}$ to the hypersurfaces $M_t = S\times\{t\}$ is $g_t = \zeta(t)^2 g$. Therefore, from (13), we have $k_t = \partial_t g_t = 2\zeta\dot{\zeta} g = 2(\dot{\zeta}/\zeta) g_t$. That is, when $t=0$ we get $g_0 = \zeta(0)^2 g$ and $k_0 = 2(\dot{\zeta}(0)/\zeta(0)) g_0$. Note that $\operatorname{tr} k_0 = 6(\dot{\zeta}(0)/\zeta(0))$ is constant. Since the stress-energy tensor $\tilde{T}$ of the Robertson–Walker model satisfies $\tilde{T}_{44} = \rho$ and $\tilde{T}_{4i} = 0$, we have $F_t = \tilde{T}^4_4|_{M_t} = -\rho_t$ and $X^i_t(x) = \tilde{T}^i_4(x) = 0$. Thus, the hypotheses of Theorem 2 are all fulfilled. Then, Theorem 1 assures that the Einstein equations are linearization stable at the initial Robertson–Walker metric $\tilde{g}$.

[1] P. D. D'Eath, ''On the existence of perturbed Robertson–Walker universes,'' Ann. Phys. (N.Y.) **98**, 237–263 (1976).

[2] J. E. Marsden, ''The initial value problem and dynamics of gravitational fields,'' in *Proceedings of the Ninth International Conference on General Relativity and Gravitation*, edited by E. Schmutzer (Cambridge University Press, Cambridge, 1983), pp. 115–126.

[3] J. E. Marsden, ''Lectures on Geometric Methods in Mathematical Physics,'' Regional Conference Series in Applied Mathematics, Commun. Math. Phys. .

[4] V. Moncrief, ''Space-time symmetries and linearization stability of the Einstein equations (I),'' J. Math. Phys. **16**, 493–498 (1975).

[5] V. Moncrief, ''Space-time symmetries and linearization stability of the Einstein equations (II),'' J. Math. Phys. **17**, 1893–1902 (1976).

[6] D. Bao, J. E. Marsden, and R. Walton, ''The Hamiltonian structure of general relativistic perfect fluids,'' Commun. Math. Phys. **99**, 319–345 (1985).

[7] A. Fischer and J. E. Marsden, ''Linearization stability of nonlinear partial differential equations,'' Proc. Symp. Pure Math. **27**, 219–263 (1975).
[8] A. Einstein, *Näherungsweise Integration der Feldgleichungen der Gravitation* (Preussiche Akademie der Wissenschaften, Sitzungsberiche, 1916), pp. 688–696.
[9] A. Einstein, *Über Gravitationswellen* (Preussiche Akademie der Wissenschaften, Sitzungsberiche, 1918), pp. 154–167.
[10] Y. Choquet-Bruhat and S. Deser, ''On the stability of flat space,'' Ann. Phys. (N.Y.) **81**, 165–168 (1973).
[11] Y. Choquet-Bruhat, A. Fischer, and J. E. Marsden, ''Équations des constraintes sur une variété non compacte,'' C. R. Acad. Sci., Ser. I: Math. **284**, A975–A978 (1977).
[12] M. Cantor, ''Spaces of functions with asymptotic conditions on $\mathbb{R}^n$,'' Indiana Univ. Mat. J. **24**, 897–902 (1975).
[13] M. Cantor, ''Some problems of global analysis on asymptotically simple manifolds,'' Compositio Mathematica **38**, 3–35 (1979).
[14] A. Fischer and J. E. Marsden, ''The Einstein equations of evolution—a geometric approach,'' J. Math. Phys. **13**, 546–568 (1972).
[15] A. Lichnerowicz, *Théories Relativistes de la Gravitation et de l'Électromagnétisme* (Masson and Cie Paris, 1955).
[16] A. Fischer and J. E. Marsden, ''The Einstein evolution equations as a first-order quasi-linear symmetric hyperbolic system,'' Commun. Math. Phys. **28**, 1–38 (1972).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Linearization stability of the Einstein equation for Robertson–Walker models. II

Lluís Bruna
Departament de Física Aplicada, E.T.S. d'Enginyers de Telecomunicacions, Univeristat Politècnica de Catalunya, C/Jordi Girona s/n, 08034-Barcelona, Spain

Joan Girbau
Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

(Received 24 November 1998; accepted for publication 15 January 1999)

In a previous paper (referred to here as paper I) we showed that the Einstein equation is linearization stable when a Robertson–Walker model of curvature $K=0$ is considered. For that purpose, a slightly different definition of linearization stability was introduced. In this paper we show that in case the curvature K is equal to 1 the Einstein equation $G(g)=\chi T$ is not linearization stable at the Robertson–Walker metric. © *1999 American Institute of Physics.* [S0022-2488(99)02209-4]

I. INTRODUCTION

This is the second part of a series of two papers. For motivations, notations, and references see Paper I.[1] Recall that in Paper I we dealt with a Robertson–Walker model $V=S\times I$ where (S,g) was a Riemannian 3-manifold of constant curvature K and V was endowed with a Lorentz metric of the form $\tilde{g}=-dt^2+\zeta(t)g$. In V we considered also a perfect fluid with stress-energy tensor $\tilde{T}$. We wondered whether the Einstein equation $\tilde{G}(\tilde{g}')=\chi\tilde{T}'$ could be linearized at the initial metric $\tilde{g}$ corresponding to the Robertson–Walker model $(V,\tilde{g},\tilde{T})$. The main result of this article (contained in Sec. III) is the linearization instability of the Einstein equation at the initial metric $\tilde{g}$ when the curvature K of (S,g) is positive. In Sec. I we write the notation used throughout this paper and we express the mapping $\Phi(g,k)=(\mathcal{H}(g,k),\gamma(g,k))$ of paper I in terms of a more suitable pair of variables (g,p) for subsequent calculus. In Sec. II we compute the expressions for $D_{(g,p)}\Phi$, $(D_{(g,p)}\Phi)^*$ and $D^2_{(g,p)}\gamma$ needed in the proof of the main theorem.

II. THE Φ MAPPING

Let M be a compact 3-manifold. Unlike paper 1, we use common Sobolev spaces with $p=2$ without weights. For all $s\in\mathbb{R}^+$, $\mathcal{F}^s(M)$ will stand for the space of functions on M of Sobolev class s, $S_2^s(M)$ the space of 2-covariant symmetric tensor fields over M of Sobolev class s, $\mathcal{R}^s(M)$ the open set in $S_2^s(M)$ of the Riemannian metrics, and $\mathcal{X}^s(M)$ the space of vector fields on M of Sobolev class s. Suppose that a Riemannian metric g, a 2-symmetric covariant tensor k, a function F and a vector field X are given on M in such a way that the data (g,k,F,X) satisfy

$$\begin{cases}\frac{1}{8}(k\cdot k-(\mathrm{tr}_g\,k)^2-4R)=\chi F\\ \frac{1}{2}\nabla^r(k_{ir}-(\mathrm{tr}_g\,k)g_{ir})=\chi X_i\end{cases}. \tag{1}$$

Consider the mapping

$$\Phi:\mathcal{R}^s(M)\times S_2^s(M)\to\mathcal{F}^{s-2}(M)\times\mathcal{X}^{s-1}(M)$$

$$(g',k')\to(\mathcal{H}(g',k'),\gamma(g',k'))$$

where $\mathcal{H}$ and γ denote the left hand side of (1)

In order to simplify the calculus we define $p' = k' - (\mathrm{tr}_{g'} k')g'$ and we will work with the new variables (g',p') instead of the old ones (g',k'). A straightforward computation shows that $\mathcal{H}$ and γ are then given by

$$\mathcal{H}(g',p') = \tfrac{1}{8}(p' \cdot_{g'} p' - 4R(g') - \tfrac{1}{2}(\mathrm{tr}_{g'} p')^2)$$

$$\gamma(g',p')_i = \tfrac{1}{2}\nabla'^j p'_{ij}.$$

The initial metric g induces an inner product at each point $x \in M$ in the following way:

$$(h,P)_x \cdot (h',P')_x = g^{ir}(x) g^{js}(x)(h_{ir} h'_{js}(x) + P_{ir} P'_{js}(x)).$$

We then extend it to $S_2^s(M) \times S_2^s(M)$ by

$$\langle (h,P),(h',P') \rangle = \int_{x \in M} (h,P)_x \cdot (h',P')_x dx,$$

where dx is the volume element of M by g. Also we define in $\mathcal{F}^{s-2}(M) \times \mathcal{X}^{s-1}(M)$ an inner product by

$$\langle (f,Y),(f',Y') \rangle = 4\int_M f(x) f'(x) dx + 4\int_M g(Y,Y')_x dx.$$

(The factor 4 will simplify subsequent formulas.)

The linear tangent mapping of Φ at (g,p) induces a mapping

$$D_{(g,p)}\Phi: \ S_2^\infty(M) \times S_2^\infty(M) \to \mathcal{F}^\infty(M) \times \mathcal{X}^\infty(M),$$

with an adjoint operator with respect to both inner products

$$(D_{(g,p)}\Phi)^*: \ \mathcal{F}^\infty(M) \times \mathcal{X}^\infty(M) \to S_2^\infty(M) \times S_2^\infty(M)$$

defined by

$$\langle (D_{(g,p)}\Phi)(h,P),(f,Y) \rangle = \langle (h,P),(D_{(g,k)}\Phi)^*(f,Y) \rangle. \tag{2}$$

This operator induces an operator for any Sobolev class s

$$(D_{(g,p)}\Phi)^*: \ \mathcal{F}^{s-2}(M) \times \mathcal{X}^{s-1}(M) \to S_2^{s-4}(M) \times S_2^{s-4}(M).$$

III. EXPRESSIONS FOR $D_{(g,p)}\Phi$, $(D_{(g,p)}\Phi)^*$ AND $D^2_{(g,p)}\gamma$

Later on we will need the expressions of $D_{(g,p)}\Phi$, $(D_{(g,p)}\Phi)^*$ and $D^2_{(g,p)}\gamma$ (γ being the second component of $\Phi = (\mathcal{H},\gamma)$). The expressions of these operators can be found in several papers (Refs. 2–5 for instance) when the initial conditions (g,p) satisfy $\mathcal{H}(g,p) = 0$, $\gamma(g,p) = 0$. However, now $\mathcal{H}(g,p) = \chi F$ and $\gamma(g,p) = \chi X$.

In Sec. VI of paper I we have already calculated the expressions of $D_{(g,k)}\mathcal{H}$ and $D_{(g,k)}\gamma$ using the (g,k) variables. In the same way, working with the (g,p) variables, we must calculate $\mathcal{H}(g + h, p + P)$ and $\gamma(g + h, p + P)$ and keep the linear terms in h and P. Then

$$(D_{(g,p)}\mathcal{H})(h,P)=\tfrac{1}{8}[2p\cdot P-2h\cdot(p\times p)+4h\cdot Ric(g)-4\nabla^i\nabla^j h_{ij}$$
$$+4\nabla^i\nabla_i \operatorname{tr} h-(\operatorname{tr} p)(\operatorname{tr} P)+(\operatorname{tr} p)h\cdot p]$$
$$(D_{(g,p)}\gamma)(h,P)_i=\tfrac{1}{2}[\nabla^s P_{is}-h^{rs}\nabla_r p_{is}+h^{rs}\nabla_i k_{rs}-\tfrac{1}{2}(\nabla_i h_{rm})p^{rm}-(\nabla^s h_{sm})p_i^m+\tfrac{1}{2}\nabla_m(\operatorname{tr} h)p_i^m]. \tag{3}$$

Keeping the quadratic terms in h and P of $\gamma(g+h,p+P)$ we obtain

$$(D^2_{(g,p)}\gamma)(h,P)_i=\tfrac{1}{2}[\nabla_r((h\times h)^{rs}p_{is})+h^{rs}(\nabla_i h_{rm})p_s^m-\tfrac{1}{2}h^{rs}(\nabla_m h_{rs})p_i^m-\tfrac{1}{2}h^{lm}(\nabla_m \operatorname{tr} h)p_{il}$$
$$-\tfrac{1}{2}(\nabla_i h_{rm})P^{mr}+\tfrac{1}{2}(\nabla_m \operatorname{tr} h)P_i^m-\nabla_r(h^{rs}P_{is})]. \tag{4}$$

Let us now evaluate $(D_{(g,p)}\Phi)^*$, defined by the identity (2). If we denote by A and B the two components of $(D_{(g,p)}\Phi)^*$, the identity (2) is written

$$4\int_M (D_{(g,p)}\mathcal{H})(h,P)_x f(x)dx+4\int_M g((D_{(g,p)}\gamma)(h,P)_x,Y_x)dx$$
$$=\int_M hA(f,Y)_x\,dx+\int_M P\cdot B(f,Y)_x\,dx. \tag{5}$$

In order to calculate the two integrals of the first member of (5), once (3) is used, one must handle with care the terms containing derivatives in h or in P. For example, consider the expression of $(D_{(g,p)}\mathcal{H})(h,P)f$ and its term $-(1/2)\nabla^i\nabla^j h_{ij}f$ containing two derivatives on h. This term can be written

$$-\tfrac{1}{2}(\nabla^i\nabla^j h_{ij})f=\tfrac{1}{2}(\nabla^j h_{ij})\nabla^i f+\text{divergence}=-\tfrac{1}{2}h_{ij}\nabla^j\nabla^i f+\text{divergence}.$$

Doing the same in all the terms containing derivatives in h or in P and bearing in mind that the integral of a divergence is zero (because M is compact), we obtain the following expressions for A and B:

$$A(f,Y)=-fp\times p+2f\,\text{Ric}(g)-2\,\text{Hess}(f)-2(\Delta f)g+\tfrac{1}{2}f(\operatorname{tr} p)p$$
$$+L_Y p+(\operatorname{div} Y)p-\tfrac{1}{2}(p\cdot L_Y g)g-2g(\gamma(g,p),Y)$$
$$B(f,Y)=-L_Y g+fp-\tfrac{1}{2}f(\operatorname{tr} p)g, \tag{6}$$

where $(\text{Hess}\,f)_{ij}$ means the hessian $\nabla_i\nabla_j f$ and Δf the Laplacian $-\nabla^i\nabla_i f$. In the computations of (6) the following expression for the Lie derivative $L_Y\alpha$ of a 2-covariant tensor α has been used:

$$(L_Y\alpha)_{ij}=Y^k\nabla_k\alpha_{ij}+(\nabla_i Y^k)\alpha_{kj}+(\nabla_j Y^k)\alpha_{ik}\,.$$

Since the initial data (g,p) satisfy $\gamma(g,p)=\chi X$, the last term of A in (6) can be written in the form $-2\chi g(X,Y)$.

IV. THE MAIN RESULT

Consider a Robertson-Walker model $V=S^3\times I$, where S^3 is the 3-sphere endowed with a Riemannian metric g of constant curvature $K=1$ on S^3 and I an $\mathbb{R}$-interval. Let $\tilde{g}$ be a Lorentzian metric on V of the form $\tilde{g}=-dt^2+\zeta(t)^2 g$ and $\tilde{T}=(\rho+p)dt\otimes dt+p\tilde{g}$ the stress-energy tensor of a perfect fluid. Following the analysis of paper I, the hypersurface M of V is $M=S^3\times\{0\}$ and the Gauss representation of M is the identity $S^3\times I\to S^3\times I$ because $\partial/\partial t$ is a geodesic vector field of $\tilde{g}$. Since the hypersurface M has constant curvature 1 its curvature tensor satisfies

$$R^r_{kij}=\delta^r_i g_{kj}-\delta^r_j g_{ki}\,.$$

By contracting, $R_{kj}=R^r_{krj}=2g_{kj}$ and contracting again, $R=g^{ij}R_{ij}=6$. The restriction of $\tilde{g}$ to the hypersurface $M_t=S^3\times\{t\}$ is $g_t=\zeta(t)^2 g$. We have $k_t=\partial_t g_t=2(\dot{\zeta}/\zeta)g_t$. Then,

$$\begin{aligned} &\mathrm{Ric}(g_t)=\frac{2g_t}{\zeta(t)^2}; \quad R(g_t)=\frac{6}{\zeta(t)^2}; \quad k_t=\frac{\partial g_t}{\partial t}=\frac{2\dot{\zeta}(t)g_t}{\zeta(t)};\\ &k_t\times_{g_t}k_t=4\left(\frac{\dot{\zeta}(t)}{\zeta(t)}\right)^2 g_t; \quad k_t\cdot_{g_t}k_t=12\left(\frac{\dot{\zeta}(t)}{\zeta(t)}\right)^2; \quad \mathrm{tr}_{g_t}k_t=\frac{6\dot{\zeta}(t)}{\zeta(t)}. \end{aligned} \tag{7}$$

In our case $F_t=\tilde{T}^4_4|_{M_t}=-\rho_t$ and for any $x\in M_t$ we have $X^i_i(x)=\tilde{T}^i_4(x,t)=0$. The linearization stability of the Einstein equation at the initial Robertson–Walker metric $\tilde{g}$ leads us to the study of the stability of the mapping

$$\Phi:\ \mathcal{R}^s(S^3)\times S^s_2(S^3)\to\mathcal{F}^{s-2}(S^3)\times\mathcal{X}^{s-1}(S^3)$$

$$(g',p')\to(\mathcal{H}(g',p'),\gamma(g',p'))$$

at (g_0,p_0). Now, $p_0=k_0-(\mathrm{tr}_{g_0}k_0)g_0=-4(\dot{\zeta}(0)/\zeta(0))g_0$ and $\Phi(g_0,p_0)=(-\chi\rho_0,0)$ since $F=-\rho$ and $X=0$.

We state the following theorem.

Theorem 1: *There is no neighborhood U of (g_0,p_0) in $\mathcal{R}^s(S^3)\times S^s_2(S^3)$ such that the set of the pairs $(g',p')\in U$ satisfying $\Phi(g',p')=\chi(-\rho_0,0)$ is a differentiable submanifold of U.*

As a consequence of this theorem, we have:

Corollary. There exists no subspace $\mathcal{F}$ of $\mathcal{F}^{s-2}(S^3)\times\mathcal{X}^{s-1}(S^3)$ such that Φ is linearization stable at the initial point (g_0,p_0) in the direction of $\mathcal{F}$.

Proof of the corollary. The definition of linearization stability relative to any subspace $\mathcal{F}$ given in paper I requires $\Phi^{-1}(\Phi(g_0,p_0))$ to be a differential manifold in a neighborhood of (g_0,p_0), which is inconsistent with the theorem.

Proof of theorem 1. Suppose there exists a neighborhood U of (g_0,p_0) in $S^s_2(S^3)\times S^s_2(S^3)$ such that the set of $(g',p')\in U$ satisfying $\Phi(g',p')=-\chi(\rho_0,0)$ has the structure of a differential manifold of U. Let $\lambda\to(g'(\lambda),p'(\lambda))$ be a curve of this manifold passing through (g_0,p_0) for $\lambda=0$. Let $(h,P)\in S^s_2(S^3)\times S^s_2(S^3)$ be the tangent vector to this curve at (g_0,p_0). Denote by (h',P') the second derivative $d^2(g'(\lambda),p'(\lambda))/d\lambda^2$ at $\lambda=0$. Since $\Phi(g'(\lambda),p'(\lambda))=-\chi(\rho_0,0)$ for any λ, evaluating its first derivative at $\lambda=0$ one obtains $(D_{(g_0,p_0)}\Phi)(h,P)=0$. Differentiating again at $\lambda=0$ one gets

$$(D^2_{(g_0,p_0)}\Phi)(h,P)+(D_{(g_0,p_0)}\Phi)(h',P')=0. \tag{8}$$

Now, let (f,Y) be any element of $\mathcal{F}^\infty(S^3)\times\mathcal{X}^\infty(S^3)$. Computing the inner product of (f,Y) with both members of (8), one gets

$$\langle(D^2_{(g_0,p_0)}\Phi)(h,P),(f,Y)\rangle+\langle(D_{(g_0,p_0)}\Phi)(h',P'),(f,Y)\rangle=0.$$

Taking into account that

$$\langle(D_{(g_0,p_0)}\Phi)(h',P'),(f,Y)\rangle=\langle(h',P'),(D_{(g_0,p_0)}\Phi)^*(f,Y)\rangle$$

it follows that for any $(h,P)\in S^s_2(S^3)\times S^s_2(S^3)$ in $\ker(D_{(g_0,p_0)}\Phi)$ and for any $(f,Y)\in\mathcal{F}^\infty(S^3)\times\mathcal{X}^\infty(S^3)$ in $\mathrm{Ker}(D_{(g_0,p_0)}\Phi)^*$

$$\langle(D^2_{(g_0,p_0)}\Phi)(h,P),(f,Y)\rangle=0. \tag{9}$$

is fulfilled.

To prove Theorem 1 (making use of an idea of Ref. 6) we look for a pair (h,P) of the kernel of $D_{(g_0,p_0)}\Phi$ and a pair (f,Y) of the kernel of $(D_{(g_0,p_0)}\Phi)^*$ for which (9) is not satisfied. From previous calculus we know that $(D_{(g_0,p_0)}\Phi)^*(f,Y)=(A,B)$, where A and B are given by (6). In our case, using (7) and the fact that $X=0$ we have

$$A(f,Y)=8\left(\frac{\dot{\zeta}(0)^2+4}{\zeta(0)^2}\right)fg_0-4\frac{\dot{\zeta}(0)}{\zeta(0)}L_Yg_0-2\text{ Hess }\ f-2(\Delta f)g_0$$

$$B(f,Y)=2\frac{\dot{\zeta}(0)}{\zeta(0)}fg_0-L_Yg_0.$$

The metric $g_0=\zeta(0)^2g$ is the restriction of the euclidean metric of $\mathbb{R}^4$ to the sphere of radius $\zeta(0)$ and center at the origin. Let us denote by $S^3(\zeta(0))$ this sphere. We choose $f=0$ and Y a Killing field on $S^3(\zeta(0))$ relative to the metric g_0 (this means $L_Yg_0=0$) as an element of $\mathrm{Ker}(D_{(g_0,p_0)}\Phi)^*$. For any $(h,P)\in\mathrm{Ker}(D_{(g_0,p_0)}\Phi)$ and a pair of the form $(0,Y)\in\mathrm{Ker}(D_{(g_0,p_0)}\Phi)^*$ we have

$$\langle(D^2_{(g_0,p_0)}\Phi)(h,P),(0,Y)\rangle=4\int_M g((D^2_{(g_0,p_0)}\gamma)(h,P),Y)dv,$$

where dv stands for the volume element of $M=S^3(\zeta(0))$ by g_0. To evaluate $(D^2_{(g_0,p_0)}\Phi)\times(h,P)$ given by (4) when a Robertson–Walker model is considered, we make use of the formulas $p_{ij}=-4C(g_0)_{ij}$, where $C=\dot{\zeta}(0)/\zeta(0)$. Then

$$(D^2_{(g_0,p_0)}\gamma)(h,P)_i=\tfrac{1}{2}[-4C\nabla^r(h\times h)_{ri}-2Ch^{rm}\nabla_ih_{rm}+2Ch_{im}\nabla^m\operatorname{tr}h-\tfrac{1}{2}(\nabla_ih^{rm})P_{mr}+\tfrac{1}{2}(\nabla^m\operatorname{tr}h)P_{im}-\nabla_r(h^{rs}P_{is})].\tag{10}$$

Hence, $g((D^2_{(g_0,p_0)}\gamma)(h,P),Y)$ can be written in the following form:

$$g((D^2_{(g_0,p_0)}\gamma)(h,P),Y)=\tfrac{1}{2}\{4C(\nabla^rY^i)(h\times h)_{ri}-2Ch^{rm}(\nabla_Yh)_{rm}-2C(\nabla^mY^i)h_{im}\operatorname{tr}h-2CY^i(\nabla^mh_{im})\operatorname{tr}h-\tfrac{1}{2}(\nabla_Yh)^{rm}P_{mr}-\tfrac{1}{2}Y^i(\nabla^mP_{im})\operatorname{tr}h-\tfrac{1}{2}(\nabla^mY^i)P_{im}\operatorname{tr}h-Y^i\nabla_r(h^{rs}P_{is})\}+\mathrm{div}.\tag{11}$$

Now we require (h,P) to belong to the kernel of $D_{(g_0,p_0)}\Phi$; therefore both components of (3) must be zero. From the vanishing of the second expression of (3) [and since $p_{ij}=-4C(g_0)_{ij}$], one obtains $\nabla^mP_{im}=-4C\nabla^mh_{im}$. Substituting this into (11),

$$g((D^2_{(g_0,p_0)}\gamma)(h,P),Y)=\tfrac{1}{2}\{-2Ch\cdot\nabla_Yh-\tfrac{1}{2}(\nabla_Yh)\cdot P-Y^i\nabla_r(h^{rs}P_{is})\}+\mathrm{div}.$$

When integrating the previous identity on $M=S^3(\zeta(0))$, the term $\int_M(h\cdot\nabla_Yh)dv$ will be zero. Indeed

$$h\cdot\nabla_Yh=h_{rm}Y^i\nabla_ih^{rm}=-h_{rm}(\nabla_iY^i)h^{rm}-(\nabla_ih_{rm})Y^ih^{rm}+\mathrm{div}.$$

Since Y is a Killing field, it has zero divergence and therefore the first term of the second member vanishes. Integrating on M one gets $\int_M(h\cdot\nabla_Yh)dv=-\int_M(h\cdot\nabla_Yh)dv$ and so this integral is zero. Finally:

$$4\int_M g((D^2_{(g_0,p_0)}\gamma)(h,P),Y)dv=-\int_M(\nabla_Yh\cdot P)dv-2\int_M Y^i\nabla_r(h^{rs}P_{is})dv.\tag{12}$$

So our problem consists in finding h and P in $\operatorname{Ker} D_{(g_0,p_0)}\Phi$ and a Killing field Y on the sphere $S^3(\zeta(0))$ for which (12) is not zero.

Consider $S^3(\zeta(0))\subset\mathbb{R}^4$. Let (x,y,z,t) be the canonical coordinates of $\mathbb{R}^4$. Let e_1, e_2, e_3 be three vector fields of $\mathbb{R}^4$ with the following components in the canonical basis of $\mathbb{R}^4$: $e_1=(y,-x,t,-z)$, $e_2=(z,-t,-x,y)$, $e_3=(t,z,-y,-x)$. The Lie brackets of these fields are $[e_1,e_2]=2e_3$, $[e_2,e_3]=2e_1$, $[e_3,e_1]=2e_2$. Consider the radial vector field $N=(x,y,z,t)$. The fields $e_i (i=1,2,3)$ are orthogonal to N by the Euclidean metric of $\mathbb{R}^4$. Hence, their restriction to the sphere $S^3(\zeta(0))$ are tangent fields that at each point are mutually orthogonal and have norm $\zeta(0)$. The formula

$$2g_0(\nabla_U V,W)=U(g_0(V,W))+V(g_0(W,U))-W(g_0(U,V))$$
$$-g_0(U,[V,W])-g_0(V,[U,W])-g_0(W,[V,U])$$

that links the covariant derivative with the metric, gives $\nabla_{e_i}e_j=-\nabla_{e_j}e_i=e_k$ provided that i,j,k be a cyclic permutation of 1,2,3. Also $\nabla_{e_i}e_i=0$. If Γ_{ij}^k stands for the Christoffel symbols in the basis $\{e_1,e_2,e_3\}$ defined by $\nabla_{e_i}e_j=\Gamma_{ij}^k e_k$, one has $\Gamma_{ij}^k=\epsilon(i,j,k)$, where $\epsilon(i,j,k)$ indicates the sign of the permutation (i,j,k) when the three indices are different, and zero otherwise. A Killing vector field Y of $S^3(\zeta(0))$ satisfies $g_0(\nabla_{e_i}Y,e_j)=-g_0(\nabla_{e_j}Y,e_i)$. In particular, the fields e_1, e_2, e_3 are Killing fields.

Now, choose $Y=e_1$, $h=e_1\otimes e_2+e_2\otimes e_1$, $P=e_1\otimes e_3+e_3\otimes e_1$. We want to see that (Y,h,P) satisfy the desired conditions. We begin by showing that these h and P make the right hand side of (3) to vanish. In a Robertson-Walker model, the expressions (3) take the form:

$$(D_{(g_0,p_0)}\mathcal{H})(h,P)=\frac{1}{8}\left[4C\operatorname{tr}P+\left(16C^2+\frac{8}{\zeta(0)^2}\right)\operatorname{tr}h-4\nabla^i\nabla^j h_{ij}-4\nabla\operatorname{tr}h\right]$$
$$(D_{(g_0,p_0)}\gamma)(h,P)_i=\tfrac{1}{2}\nabla^s(4Ch+P)_{is}. \tag{13}$$

Here, both the trace of h and P vanish since, for instance, $\operatorname{tr}h=(g_0)^{ij}h_{ij}=(1/\zeta(0)^2)(\Sigma_i h_{ii})=0$. On the other hand, if both h and P have zero divergence then the expressions (13) would be consequently satisfied. Hence, let $A=A^{ij}e_i\otimes e_j$ be a symmetric tensor with constant components A^{ij}. Then $\nabla_i A^{ij}=e_i(A^{ij})+\Gamma_{ir}^i A^{rj}+\Gamma_{ir}^j A^{ir}$. The first term is zero because A^{ij} are constant. The second term is zero because $\Gamma_{ij}^k=\epsilon(i,j,k)$, and the third is also zero because Γ_{ij}^k is skew symmetric in the indices i,j but A^{ij} is symmetric in the same indices.

Finally, let us to see that (12) is not zero. We have $\nabla_Y h=\nabla_{e_1}(e_1\otimes e_2+e_2\otimes e_1)=e_1\otimes e_3+e_3\otimes e_1=P$. Hence, $(\nabla_Y h)\cdot P=P\cdot P=2/\zeta(0)^2$. On the other hand, since $Y=e_1$ and h has zero divergence, we get

$$Y^i\nabla_r(h^{rs}P_{is})=h^{rs}\nabla_r P_{1s}=h^{12}\nabla_1 P_{12}+h^{21}\nabla_2 P_{11}=(1/\zeta(0))^2(\nabla_1 P_{12}+\nabla_2 P_{11}).$$

A short computation shows that $\nabla_1 P_{12}=-1$ and $\nabla_2 P_{11}=2$. Therefore,

$$Y^i\nabla_r(h^{rs}P_{is})=1/\zeta(0)^2.$$

Substituting these results in (12) we get

$$4\int_M g((D^2_{(g_0,p_0)}\gamma)(h,P),Y)dv=-\int_M(\nabla_Y h\cdot P)dv-2\int_M Y^i\nabla_r(h^{rs}P_{is})dv$$
$$=(1/\zeta(0))^2(-4\operatorname{vol}(S^3(\zeta(0))))\neq 0.$$

This completes the proof of the theorem.

[1] Ll. Bruna and J. Girbau, ''Linearization stability of the Einstein equation for Robertson–Walker models. I,'' J. Math. Phys. **40**, 5117 (1999), preceding paper.
[2] A. Fischer and J. E. Marsden, ''The Einstein evolution equations as a first-order quasi-linear symmetric hyperbolic system,'' Commun. Math. Phys. **28**, 1–38 (1972).
[3] A. Fischer and J. E. Marsden, ''Linearization stability of nonlinear partial differential equations,'' Proc. Symp. Pure Math. **27**, 219–263 (1975).
[4] V. Moncrief, ''Space-time symmetries and linearization stability of the Einstein equations (I),'' J. Math. Phys. **16**, 493–498 (1975).
[5] V. Moncrief, ''Space-time symmetries and linearization stability of the Einstein equations (II),'' J. Math. Phys. **17**, 1893–1902 (1976).
[6] J. M. Arms and J. E. Marsden, ''The absence of Killing fields is necessary for linearization stability of Einstein's equations,'' Indiana Univ. Math. J. **28**, 119–125 (1979).

On the differentiability of Cauchy horizons

Robert J. Budzyński
Department of Physics, Warsaw University, Hoża 69, 00-681 Warsaw, Poland

Witold Kondracki and Andrzej Królak
Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, 00-950 Warsaw, Poland

(Received 9 March 1999; accepted for publication 14 June 1999)

Chruściel and Galloway constructed a Cauchy horizon that is nondifferentiable on a dense set. We prove that in a certain class of Cauchy horizons densely nondifferentiable Cauchy horizons form a dense set. We show that our class of densely nondifferentiable Cauchy horizons implies the existence of densely nondifferentiable Cauchy horizons arising from partial Cauchy surfaces and also the existence of densely nondifferentiable black hole event horizons. © *1999 American Institute of Physics.* [S0022-2488(99)01909-X]

I. INTRODUCTION

Recently Chruściel and Galloway[1] have constructed an example of a Cauchy horizon which fails to be differentiable on a dense subset. In this paper we show that densely nondifferentiable Cauchy horizons appear to be generic in a certain class of Cauchy horizons. Chruściel and Galloway have also shown that their example implies the existence of a densely nondifferentiable black hole event horizon. They point out that these examples raise definite questions concerning some major arguments that have been given in the past where smoothness assumptions were implicitly made. In the light of these new examples, it is clear that there is a real need for a deeper understanding of the differentiability properties of horizons.

In a spacetime with a partial Cauchy surface S the Cauchy horizon $H(S)$ is the boundary of the set of points where, in theory, one may calculate everything in terms of the initial data on S. Cauchy horizons are *achronal* (i.e., no two points on the horizon may be joined by a timelike curve) and this implies that Cauchy horizons (locally) satisfy a Lipschitz condition. This, in turn, implies that Cauchy horizons are differentiable almost everywhere. Because they are differentiable except for a set of (three-dimensional) measure zero, it seems that they have often been assumed to be smooth except for a set which may be more or less neglected. However, one must remember in the above that (1) differentiable only refers to being differentiable at a single point and (2) sets of measure zero may be quite widely distributed.

For a closed achronal set S each point p of a Cauchy horizon $H^+(S)$ lies on at least one null generator.[2] However, null generators may or may not remain on the horizon when they are extended in the future direction. If a null generator leaves the horizon, then there is a last point where it remains on the horizon. This last point is said to be an *endpoint* of the horizon. Endpoints where two or more null generators leave the horizon are points where the horizon must fail to be differentiable.[3,1] In addition, Chruściel and Galloway[1] have shown that Cauchy horizons are differentiable at points which are not endpoints. Beem and Królak have shown[4] that Cauchy horizons are differentiable at endpoints where only one generator leaves the horizon. These results give a complete classification of (pointwise) differentiability for Cauchy horizons in terms of null generators and their endpoints. Beem and Królak have also shown[4] that if we consider an open subset W of the Cauchy horizon $H^+(S)$ and assume that the horizon has no endpoints on W, then the horizon must be differentiable at each point of W and, in fact, that the horizon must be at least of class C^1 on W. Conversely, the differentiability on an open set W implies there are no endpoints on W.

For general spacetimes, horizons may fail to be stable under small metric perturbations; however, some sufficiency conditions for various stability questions have been obtained.[5,6]

II. PRELIMINARIES

Definition 1: A space-time (M,g) is a smooth n-dimensional, Hausdorff manifold M with a semi-Riemannian metric g of signature (−,+,...,+), a countable basis, and a time orientation.

A set S is said to be *achronal* if there are no two points of S with timelike separation.

We give definitions and state our results in terms of the future horizon $H^+(S)$, but similar results hold for any past Cauchy horizon $H^-(S)$.

Definition 2: The future Cauchy development $D^+(S)$ *consists of all points* $p\in M$ *such that each past endless and past directed causal curve from p intersects the set S. The future Cauchy horizon is* $H^+(S)=\overline{(D^+(S))}-I^-(D^+(S))$.

Let p be a point of the Cauchy horizon; then there is at least one null generator of $H^+(S)$ containing p. Each null generator is at least part of a null geodesic of M. When a null generator of $H^+(S)$ is extended into the past it either has no past endpoint or has a past endpoint on edge (S) (see Ref. 2, p. 203). However, if a null generator is extended into the future it may have a last point on the horizon which is then said to be an *endpoint* of the horizon. We define the *multiplicity* (see Ref. 4) of a point p in $H^+(S)$ to be the number of null generators containing p. Points of the horizon which are not endpoints must have multiplicity one. The multiplicity of an endpoint may be any positive integer or infinite. We call the set of endpoints of multiplicity two or higher the *crease set*, compare Ref. 1. By a basic Proposition due to Penrose (Ref. 2, Proposition 6.3.1) $H^+(S)$ is an $n-1$ dimensional Lipschitz topological submanifold of M and is achronal. Since a Cauchy horizon is Lipschitz it follows from a theorem of Rademacher that it is differentiable almost everywhere (i.e., differentiable except for a set of $n-1$ dimensional measure zero). This does not exclude the possibility that the set of nondifferentiable points is a dense subset of the horizon. An example of such a behavior was given by Chruściel and Galloway.[1]

Following Ref. 4 let us introduce the notion of differentiability of a Cauchy horizon. Consider any fixed point p of the Cauchy horizon $H^+(S)$ and let x^0,x^1,x^2,x^3 be local coordinates defined on an open set about $p=(p^0,p^1,p^2,p^3)$. Let $H^+(S)$ be given near p by an equation of the form

$$x^0=f_H(x^1,x^2,x^3).$$

The horizon $H^+(S)$ is *differentiable* at the point p iff the function f_H is differentiable at the point (p^1,p^2,p^3). In particular, if $p=(0,0,0,0)$ corresponds to the origin in the given local coordinates and if

$$\Delta x=(x^1,x^2,x^3)$$

represents a small displacement from p in the $x^0=0$ plane, then $H^+(S)$ is differentiable at p iff one has

$$f_H(\Delta x)=f_H(0)+\sum a_i x^i+R_H(\Delta x)=0+\sum a_i x^i+R_H(\Delta x),$$

where the ratio $R_H(\Delta x)/|\Delta x|$ converges to zero as $|\Delta x|$ goes to zero. Here we use

$$|\Delta x|=\sqrt{(x^1)^2+(x^2)^2+(x^3)^2}.$$

If $H^+(S)$ is differentiable at the point p, then there is a well-defined three-dimensional linear subspace N_0 in the tangent space $T_p(M)$ such that N_0 is tangent to the three-dimensional surface $H^+(S)$ at p. In the above notation a basis for N_0 is given by $\{a_i\partial/\partial x^0+\partial/\partial x^i|i=1,2,3\}$.

Theorem 1: *(Chruściel and Galloway[1]). There exists a connected set* $K\subset R^2=\{t=0\}\subset R^{2,1}$, *where* $R^{2,1}$ *is a* $2+1$ *dimensional Minkowski space-time, with the following properties:*

(1) *The boundary* $\partial K = \overline{K} - \mathrm{int}\, K$ *of K is a connected, compact, Lipschitz topological submanifold of* R^2. *K is the complement of a compact set in* R^2.
(2) *There exists no open set* $\Omega \subset R^{2,1}$ *such that* $\Omega \cap H^+(K) \cap \{0 < t < 1\}$ *is a differentiable submanifold of* $R^{2,1}$.

Proposition 1 (Beem and Królak[4]): Let W be an open subset of the Cauchy horizon $H^+(S)$. *Then the following are equivalent:*

(1) $H^+(S)$ *is differentiable on W.*
(2) $H^+(S)$ *is of class* C^r *on W for some* $r \geq 1$.
(3) $H^+(S)$ *has no endpoints on W.*
(4) *All points of W have multiplicity one.*

Note that the four parts of Proposition 1 are logically equivalent for an open set *W*, but that, in general, they are not necessarily equivalent for sets which fail to be open. Using the equivalence of parts (1) and (3) of Proposition 1, it now follows that near each endpoint of multiplicity one there must be points where the horizon fails to be differentiable. Hence, each neighborhood of an endpoint of multiplicity one must contain endpoints of higher multiplicity. This yields the following corollary.

Corollary 1 (Ref. 4): If p is an endpoint of multiplicity one on a Cauchy horizon $H^+(S)$, *then each neighborhood* $W(p)$ *of p on* $H^+(S)$ *contains points where the horizon fails to be differentiable. Hence, the set of endpoints of multiplicity one is in the closure of the crease set.*

III. A GENERIC DENSELY NONDIFFERENTIABLE CAUCHY HORIZON

We shall construct a densely nondifferentiable Cauchy horizon in the three-dimensional Minkowski space-time $R,^{2,1}$ but our construction can be generalized in a natural way to higher dimensions. Let Σ be the surface $t=0$, and let K be a compact, convex subset of Σ. Let ∂K denote the boundary of K. Let $\rho(x,R)$ and $D(x,R)$ be, respectively, a circle and a disc with center at x and radius R.

Definition 3: A circle $\rho(x,R)$ *is internally tangent to the boundary* ∂K *of K if the disc enclosed by* ρ *is contained in K and for all* $\epsilon > 0$ *the disc of radius* $R+\epsilon$ *and center x is not contained in K.*

Let $\rho(x,R)$ be internally tangent to ∂K; then the point $(x,R) \in R^{2,1}$ belongs to the future Cauchy horizon $H^+(K)$ and conversely, if a point $(x,R) \in R^{2,1}$ belongs to $H^+(K)$ then the circle $\rho(x,R)$ is internally tangent to ∂K. If $\rho(x,R)$ is internally tangent in at least two points of ∂K then it follows from Proposition 1 that $H^+(K)$ is not differentiable at the point (x,R) and the point (x,R) has multiplicity at least two.

We shall first construct a continuous curve that is not differentiable on any open subset. Let us take a line segment l_0 and let us consider an isosceles triangle with base l_0 and let α_0 be the angle at the base and let l_1 denote the broken line consisting of two equal arms of the triangle. In the next step we construct two isosceles triangles with bases that are segments of the broken line l_1 and we choose the angles α_1 at the base equal $q \times \alpha_0$ where $q < 1/2$. We iterate the above construction. At the Nth step of the construction the number of nondifferentiable points of the curve increases by $2N-1$. After the Nth step of the iterative procedure the vertex angle of the isosceles triangle obtained in the ith step is given by

$$\angle_N(x_i) = \pi - 2\alpha_1 \left[q^{i-1} - \frac{q^i - q^N}{1-q} \right]. \tag{1}$$

In the limit $N \to \infty$ the ith vertex angle is given by $\pi - 2\alpha_1 q^i[(q^{-1}-2)/(1-q)]$ and is strictly less than π as $q < 1/2$.

Let us call the nowhere differentiable continuous curve constructed above a *rough curve*. Let us call a region of Σ that is bounded by a rough curve and two straight lines perpendicular to the rough curve at its two endpoints (this notion is unambiguous, as the slope of the rough curve at an endpoint is given by a well-defined limit) a *fan*. The above construction can be generalized to higher dimensions, for example in the four-dimensional Minkowski space-time we construct a *rough surface* in the following way. We consider a triangle and the first step is to construct a pyramid with the triangle as a base and all angles between the base and the sides of the pyramid equal to the same angle α_1; we then iterate the construction decreasing at each step the angle α between the base and the sides of the pyramid by a factor $q<1/2$ as in the three-dimensional case. As a result we obtain a nowhere differentiable surface and we define a three-dimensional fan as the region of Σ bounded by the rough surface and planes perpendicular to the rough surface passing through the sides of the initial triangle.

Theorem 2: *Let b be a rough curve and F the corresponding fan. Then the set of points of F that are centers of circles tangent to b in at least two points of b is dense in the interior of the fan F.*

Proof: Each point of F is the center of a circle tangent to b at at least one point. If the claim of the theorem were false, then there would exist a disc $D(x,R)$ with nonempty interior with the property that every point $a\in \operatorname{int} D$ is the center of a circle tangent to the rough curve at exactly one point.

(1) A vertex point cannot be a point of tangency of any circle with center in $\operatorname{int} F$.
(2) By construction the set of vertices of b is dense in b. Thus the complement of the set of vertices in b is totally disconnected (i.e., only one-element subsets are connected).

Let us consider a map P from the disc to b that assigns to every point y of D a point on b that is tangent to the circle centered at y. By assumption this point is unique and thus the map is well-defined.

Let us show that the map P is continuous. It is enough to prove that if $a_n \rightarrow a$ then $P(a_n) \rightarrow P(a)$. As b is compact, $P(a_n)$ has a subsequence that converges to a point c on b. Since the distance $d(a_n, P(a_n))$ is continuous on D we have $d(c,a)=d(a,P(a))$. Hence c is a tangency point of a circle centered at a and consequently $c=P(a)$.

By the Darboux theorem the image $P(D(x,R))$ is connected and by 1. and 2. above, it is a one-point set. It then follows that $R=0$ which is a contradiction. **QED**

The above theorem generalizes to the three-dimensional case. In the case of a three-dimensional fan F there exists a dense subset of F such that every ball with the center in this subset has at least two tangency points to the rough surface. All steps of the proof of Theorem 2 carry over to this case in the natural way.

Let $\mathcal{H}$ be the set of Cauchy horizons arising from compact convex sets $K\subset\Sigma$. The topology on $\mathcal{H}$ is induced by the Hausdorff distance on the set of compact and convex regions K.

Theorem 3: *Let $\mathcal{H}$ be the set of future Cauchy horizons $H^+(K)$ where K are compact and convex regions of Σ. The subset of densely nondifferentiable horizons is dense in $\mathcal{H}$.*

Proof: Any compact and convex region K can be approximated in the sense of Hausdorff distance by a (sequence of) convex polygons contained in K. Each of the vertex angles of such a polygon is strictly less than π. Over each side of the polygon we construct a rough curve in such a way that the fans corresponding to the rough curves cover the polygon. This is always possible, since we may choose the starting angle α_1 in the rough curve's construction to obey the condition

$$\phi+\frac{2\alpha_1}{1-q}<\pi, \tag{2}$$

where ϕ is the largest vertex angle of the original polygon. When α_1 decreases to 0 the rough-edged polygon converges to the original polygon in the sense of Hausdorff topology. **QED**

It is clear that the above theorem generalizes to higher dimensions.

IV. SOME EXAMPLES OF DENSELY NONDIFFERENTIABLE HORIZONS

In this section we show that the construction of the preceding section implies the existence of densely nondifferentiable Cauchy horizons of partial Cauchy surfaces and also the existence of black hole event horizons.

Definition 4: A partial Cauchy surface S is a connected, acausal, edgeless $n-1$ *dimensional submanifold of (M,g).*

Example 1: A rough wormhole.

Let $R^{3,1}$ be the four-dimensional Minkowski space-time and let K be a compact subset of the surface $\{t=0\}$ such that its Cauchy horizon is nowhere differentiable in the sense of the construction given in Sec. III. We consider a space-time obtained by removing the complement of the interior of the set K in the surface $t=0$ from the Minkowski space-time. Let us consider the partial Cauchy surface $S=\{t=-1\}$. The future Cauchy horizon of S is the future Cauchy horizon of set $K-\text{edge}(K)$, since $\text{edge}(K)$ has been removed from the space-time. Thus the future Cauchy horizon is nowhere differentiable and it is generated by past-endless null geodesics. The interior of the set K can be thought of as a ''wormhole'' that separates two ''worlds,'' one in the past of surface $\{t=0\}$ and one in its future.

Example 2: A transient black hole.

Let $R^{3,1}$ be the four-dimensional Minkowski space-time and let K be a compact subset of the surface $\{t=0\}$ such that its *past* Cauchy horizon is nowhere differentiable in the sense of the construction given in Sec. III. We consider a space-time obtained by removing from Minkowski space-time the closure of the set K in the surface $t=0$. Let us consider the event horizon $E \coloneqq \dot{J}^-(\mathcal{J}^+)$. The event horizon E coincides with $H^-(K)-\text{edge}(K)$ and thus it is not empty and nowhere differentiable. The event horizon disappears in the future of surface $\{t=0\}$ and thus we can think of the black hole (i.e., the set $B\coloneqq R^{3,1}-J^-(\mathcal{J}^+)$) in the space-time as ''transient.''

ACKNOWLEDGMENTS

The authors would like to thank P. T. Chruściel for many helpful discussions and a careful reading of the manuscript. This work was supported by the Polish Committee for Scientific Research through Grants Nos. 2 P03B 130 16 and 2 P03B 073 15.

[1] P. T. Chruściel and G. Galloway, ''Horizons non-differentiable on a dense set,'' Commun. Math. Phys. **193**, 449–470 (1998).
[2] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge University Press, Cambridge, 1973).
[3] S. W. Hawking, ''Chronology protection conjecture,'' Phys. Rev. D **46**, 603–611 (1992).
[4] J. K. Beem and A. Królak, ''Cauchy horizon endpoints and differentiability,'' J. Math. Phys. **39**, 6001–6010 (1998).
[5] J. K. Beem, ''Causality and Cauchy horizons,'' Gen. Relativ. Gravit. **27**, 93–108 (1995).
[6] P. T. Chruściel and J. Isenberg, ''On the stability of differentiability of Cauchy horizons,'' Commun. Anal. Geom. **5**, 249–277 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Well-posed forms of the 3+1 conformally-decomposed Einstein equations

Simonetta Frittelli[a)]
Physics Department, Duquesne University, Pittsburgh, Pennsylvania 15282

Oscar A. Reula[b)]
FaMAF, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

(Received 23 April 1999; accepted for publication 25 May 1999)

We show that well-posed, conformally-decomposed formulations of the 3+1 Einstein equations can be obtained by densitizing the lapse and by combining the constraints with the evolution equations. We compute the characteristics structure and verify the constraint propagation of these new well-posed formulations. In these formulations, the trace of the extrinsic curvature and the determinant of the 3-metric are singled out from the rest of the dynamical variables, but are evolved as part of the well-posed evolution system. The only free functions are the lapse density and the shift vector. We find that there is a 3-parameter freedom in formulating these equations in a well-posed manner, and that part of the parameter space found consists of formulations with causal characteristics, namely, characteristics that lie only within the lightcone. In particular there is a 1-parameter family of systems whose characteristics are either normal to the slicing or lie along the lightcone of the evolving metric.
[S0022-2488(99)01910-6]

I. INTRODUCTION

Analytical work in recent years has produced a number of systems of evolution equations which are equivalent to the Einstein equations at the constraint manifold, and which have a well posed initial value formulation.[1–7]

What motivates interest in this type of result is a general understanding (see, for instance, Ref. 8) that explicit well-posedness would be relevant in implementing consistent and stable numerical algorithms to integrate blackhole space–times.

The well-posed schemes for which a numerical code has been implemented appear not to exibit significant improvements over other methods, there being several factors relevant to numerical implementation which play a significant role. What is puzzling, however, is that, on the other hand, there have been numerical simulations with apparently better behavior, but which are based on systems which do not seem to have the well-posed character. One preponderant feature of these numerically more robust schemes is that they are built on a decomposition of the intrinsic metric into a metric of unit determinant and the determinant itself, and of the extrinsic curvature into trace and trace-free part. With slight variations, this way of evolving the 3+1 Einstein equations has been considered by Refs. 9 and 10. Quite recently, this form has been shown to possess striking computational advantages over the standard form.[11] We refer to this general scheme as a conformally-decomposed formulation of the 3+1 Einstein equations.

It is difficult to explain the success of these systems as opposed to the well-posed evolution schemes, or even to the standard (ADM) evolution schemes. The relative sizes of the fields of well-posed systems are roughly the same for different spectral frequencies in a Fourier represen-

a)Electronic mail: simo@mayu.physics.duq.edu
b)Electronic mail: reula@fis.uncor.edu

tation, which helps explain the stability of the system via numerical analysis. However, in the conformally-decomposed systems this does not happen in general (for standard norms), as it does not happen for the standard ADM system, thus making it more difficult to justify their relative better behavior.

We can speculate on two features that can possibly bear relevance to well-behaved numerical evolution. One feature is that good behavior in evolution is related to constraint violations. If the system preserves more accurately the constraints, then the evolution remains closer to the constraint submanifold, which contains the physical solutions. Outside this submanifold the solutions are unphysical; thus, there is no compelling reason to rule out fast growths for seemingly tame initial data for unphysical solutions. Thus, we suggest that controlling the constraint violations may lead to well-behaved numerical evolution. In this respect, it has been shown,[12] that (at least in the linearized case) there are well-posed modifications of the Einstein equations outside the constraint submanifold which make that submanifold an attractor, thus improving the chances of building numerical codes with better behaved constraint propagation. Another possible cause of concern for generating numerical instabilities is the nature of the boundary conditions which are usually imposed. There are only two existing treatments of the initial-boundary value problem in general relativity (see Ref. 13 for a complete theory of boundary values for the conformal Einstein equations in frame variables, and Ref. 14 for a linearized study of the ADM equations in well-posed first-order form), and no numerical simulations applying these treatments have yet appeared in the literature. The initial-boundary value problem for systems used in numerical simulations of the Einstein equations were instabilities have been found has generally not been studied, thus the set of boundary conditions for which the constraint equations are satisfied is not known for these cases. In dealing with this problem, establishing well-posedness for the Cauchy problem is a necessary first step.

The other feature which could give rise to numerical instabilities is the relative sizes of the ''longitudinal'' and the ''radiative'' modes in general relativity. In all nontrivial asymptotically flat solutions (either vacuum or with matter satisfying the appropriate energy conditions) the positivity of the mass implies the existence of longitudinal modes, and there are many astrophysically relevant cases where there is an approximate local notion of longitudinal vs transverse modes, and where the former are several orders of magnitude bigger than the latter. If they are not properly separated in the numerical algorithms, the errors caused by finite differencing might be of the order of the ''radiative'' modes, and bad behavior can be expected. The separation of the conformal freedom in the more successful codes can perhaps be thought of as a way of dealing with this issue, or at least isolating it.

In this work, we focus on this latter aspect. A technique for taking advantage of the conformal factor to partially decouple the ''longitudinal'' and ''transversal'' modes was used to obtain results on the Newtonian limit of general relativity.[2] In that case the conformal field was fixed via an elliptic equation, decoupling in this way the more prominent Newtonian potential to first order from the radiative degrees of freedom. Further studies on this problem would be critical to obtain realistic simulations of most astrophysically relevant problems.

Here we construct 3-parameter families of first-order well-posed systems which share some of the properties of the more successful systems, such as the conformal decomposition of the fundamental fields, in the hope that their study would help understand what is causing them to behave better than others. In Sec. II we apply techniques similar to those we used in Refs. 5 and 7 in order to obtain versions of the $3+1$ equations that are conformally-decomposed but which are well posed. Additionally, we calculate the structure of characteristics and show that for a open region in parameter space the resulting equations are metric-causal, namely they have all propagation cones inside or coincident with the light cone. There is even a one parameter subfamily for which propagation is either along the light cone or normal to the slices.

Furthermore we show that the constraints are propagated by these well-posed evolution equations. As opposed to Ref. 15, where also analytical studies of systems with this decomposition have been done, in this work the trace of the extrinsic curvature and the determinant of the intrinsic metric are considered dynamical variables and are evolved jointly with the rest of the system.

II. SYSTEM II

The conformally-decomposed system that we take as starting point has appeared in Ref. 11, and is a variation of the system used by Shibata and Nakamura.[10] It is a system of 15 equations for 15 variables ($\phi, K, \tilde{\gamma}_{ij}, \tilde{A}_{ij}, \tilde{\Gamma}^i$), and is referred to as System II in Ref. 11, to distinguish it from the standard 3+1 Einstein equations.[16] These variables are related to the intrinsic metric γ_{ij} and extrinsic curvature K_{ij} as follows:

$$e^{4\phi}=\det(\gamma_{ij})^{1/3}, \tag{1a}$$

$$\tilde{\gamma}_{ij}=e^{-4\phi}\gamma_{ij}, \tag{1b}$$

$$K=\gamma^{ij}K_{ij}, \tag{1c}$$

$$\tilde{A}_{ij}=e^{-4\phi}(K_{ij}-\tfrac{1}{3}\gamma_{ij}K), \tag{1d}$$

$$\tilde{\Gamma}^i=-\tilde{\gamma}^{ij}{}_{,j}, \tag{1e}$$

where $\tilde{\gamma}^{ij}$ is the inverse of $\tilde{\gamma}_{ij}$. The Einstein equations in terms of these variables are equivalent to the following:

$$\frac{d}{dt}\phi=-\frac{1}{6}\alpha K, \tag{2a}$$

$$\frac{d}{dt}\tilde{\gamma}_{ij}=-2\alpha\tilde{A}_{ij}, \tag{2b}$$

$$\frac{d}{dt}K=-\gamma^{ij}D_iD_j\alpha+\alpha\left(\tilde{A}_{ij}\tilde{A}^{ij}+\frac{1}{3}K^2\right)+\frac{1}{2}\alpha(\rho+S), \tag{2c}$$

$$\frac{d}{dt}\tilde{A}_{ij}=e^{-4\phi}(-(D_iD_j\alpha)^{\mathrm{TF}}+\alpha(R_{ij}^{\mathrm{TF}}-S_{ij}^{\mathrm{TF}}))+\alpha(K\tilde{A}_{ij}-2\tilde{A}_{il}\tilde{A}^l_j), \tag{2d}$$

$$\frac{\partial}{\partial t}\tilde{\Gamma}^i=-2\tilde{A}^{ij}\alpha_j+2\alpha\left(\tilde{\Gamma}^i_{jk}\tilde{A}^{kj}-\frac{2}{3}\tilde{\gamma}_{ij}K_{,j}-\tilde{\gamma}^{ij}S_j+6\tilde{A}^{ij}\phi_{,j}\right)$$
$$-\frac{\partial}{\partial x^j}\left(\beta^l\tilde{\gamma}^{ij}{}_{,l}-2\tilde{\gamma}^{m(j}\beta^{i)}{}_{,m}+\frac{2}{3}\tilde{\gamma}^{ij}\beta^l{}_{,l}\right). \tag{2e}$$

Here α is the lapse function, β^i is the shift vector, and $\tilde{\Gamma}^i_{jk}$ are the connection coefficients of $\tilde{\gamma}_{ij}$. The superscript TF denotes trace-free part, e.g., $R_{ij}^{\mathrm{TF}}=R_{ij}-\gamma_{ij}R/3$. Indices are raised and lowered with $\tilde{\gamma}^{ij}$ and its inverse. We use the shorthand notation

$$\frac{d}{dt}\equiv\frac{\partial}{\partial t}-\mathcal{L}_\beta, \tag{2f}$$

where $\mathcal{L}_\beta$ is the Lie derivative along β^i. We have, as well,

$$R_{ij}^{\mathrm{TF}}=R_{ij}-\tfrac{1}{3}\gamma^{ij}\gamma^{kl}R_{kl}, \tag{2g}$$

$$R_{ij}=\tilde{R}_{ij}+R^{\phi}_{ij}, \tag{2h}$$

$$R^{\phi}_{ij}=-2\overline{D}_i\overline{D}_j\phi-2\overline{\gamma}_{ij}\overline{\gamma}^{kl}\overline{D}_k\overline{D}_l\phi+4\overline{D}_i\phi\overline{D}_j\phi-4\overline{\gamma}_{ij}\overline{\gamma}^{kl}\overline{D}_k\phi\overline{D}_l\phi, \tag{2i}$$

$$\tilde{R}_{ij}=-\tfrac{1}{2}\tilde{\gamma}^{kl}\tilde{\gamma}_{ij,kl}+\tilde{\gamma}_{k(i}\tilde{\Gamma}^{k}{}_{,j)}+\tilde{\Gamma}^{k}\tilde{\Gamma}_{(ij)k}+\tilde{\gamma}^{lm}(2\tilde{\Gamma}_{l(i}\tilde{\Gamma}_{j)km}+\tilde{\Gamma}^{k}{}_{il}\tilde{\Gamma}_{kmj}), \tag{2j}$$

$$\tilde{\Gamma}^{k}_{ij}=\tfrac{1}{2}\tilde{\gamma}^{kl}(\tilde{\gamma}_{il,j}+\tilde{\gamma}_{jl,i}-\tilde{\gamma}_{ij,l}). \tag{2k}$$

This system is first-order in time and second-order in space, thus it is of second order overall. We show how System II can be handled in order to be turned into a well-posed form. First, we reduce the system to a straightforward first order form, and subsequently we densitize the lapse and combine the constraints into the evolution equations.

A. System II reduced to first-order form

We define a set of 12 additional variables

$$V_{ijk}\equiv\tilde{\gamma}_{ij,k}-\tfrac{3}{5}\tilde{\gamma}_{k(i}\tilde{\gamma}_{j)n,s}\tilde{\gamma}^{ns}+\tfrac{1}{5}\tilde{\gamma}_{ij}\tilde{\gamma}_{kn,s}\tilde{\gamma}^{ns}, \tag{3a}$$

which is trace free in all its indices, namely, $V_{ijk}\tilde{\gamma}^{ij}=0$ and $V_{ijk}\tilde{\gamma}^{jk}=V_{ijk}\tilde{\gamma}^{ik}=0$, and another set of 3 additional variables,

$$Q_i\equiv\phi_{,i}\,. \tag{3b}$$

Evolution equations for these new variables are obtained by taking a time derivative of (3) and commuting time and spatial derivatives in the resulting right-hand sides. The complete system of equations is now

$$\frac{d}{dt}\phi=-\frac{1}{6}\alpha K, \tag{4a}$$

$$\frac{d}{dt}\tilde{\gamma}_{ij}=-2\alpha\tilde{A}_{ij}, \tag{4b}$$

$$\frac{d}{dt}K=-\tilde{\gamma}^{ij}D_iD_j\alpha+\alpha\left(\tilde{A}_{ij}\tilde{A}^{ij}+\frac{1}{3}K^2\right)+\frac{1}{2}\alpha(\rho+S), \tag{4c}$$

$$\frac{d}{dt}\tilde{A}_{ij}=e^{-4\phi}(-(D_iD_j\alpha)^{\mathrm{TF}}+\alpha(R^{\mathrm{TF}}_{ij}-S^{\mathrm{TF}}_{ij}))+\alpha(K\tilde{A}_{ij}-2\tilde{A}_{il}\tilde{A}^{l}{}_{j}), \tag{4d}$$

$$\begin{aligned}\dot{\tilde{\Gamma}}^{i}-\beta^{l}\tilde{\Gamma}^{i}{}_{,l}={}&-2\tilde{A}^{ij}\alpha_{,j}+2\alpha(\tilde{\Gamma}^{i}{}_{jk}\tilde{A}^{kj}-\tfrac{2}{3}\tilde{\gamma}^{ij}K_{,j}-\tilde{\gamma}^{ij}S_j+6\tilde{A}^{ij}Q_j)-\beta^{l}{}_{,j}\tilde{\gamma}^{ij}{}_{,l}+\tilde{\Gamma}^{m}\beta^{i}{}_{,m}\\&+\tilde{\gamma}^{mi}{}_{,j}\beta^{j}{}_{,m}+2\tilde{\gamma}^{m(i}\beta^{j)}{}_{,mj}+\tfrac{2}{3}\tilde{\Gamma}^{i}\beta^{l}{}_{,l}+\tfrac{2}{3}\tilde{\gamma}^{ij}\tilde{\Gamma}^{l}{}_{,lj},\end{aligned} \tag{4e}$$

$$\begin{aligned}\dot{V}_{ijk}-\beta^{l}V_{ijk,l}={}&-2\alpha\tilde{A}_{ij,k}+\tfrac{6}{5}\alpha\tilde{\gamma}_{k(i}\tilde{A}_{j)m,n}\tilde{\gamma}^{mn}-\tfrac{2}{5}\alpha\tilde{\gamma}_{ij}\tilde{A}_{km,n}\tilde{\gamma}^{mn}+\beta^{l}{}_{,k}V_{ijl}+\beta^{l}{}_{,i}V_{ljk}\\&+\beta^{l}{}_{,j}V_{ilk}-2\alpha_{,k}\tilde{A}_{ij}\end{aligned} \tag{4f}$$

$$+\tfrac{6}{5}\tilde{\gamma}_{k(i}\tilde{A}^{n}{}_{j)}\alpha_{,n}+\tfrac{6}{5}\alpha\tilde{A}_{k(i}\tilde{\Gamma}_{j)}-\tfrac{2}{5}\alpha\tilde{A}_{ij}\tilde{\Gamma}_{k} \tag{4g}$$

$$+\tilde{\gamma}_{l(i}\beta^{l}{}_{,j)k}-\tfrac{3}{5}\beta^{l}{}_{,l(j}\tilde{\gamma}_{i)k}+\tfrac{1}{5}\tilde{\gamma}_{ij}\beta^{l}{}_{,lk} \tag{4h}$$

$$-\beta^{l}{}_{,ns}\tilde{\gamma}^{ns}(\tfrac{3}{5}\tilde{\gamma}_{k(i}\tilde{\gamma}_{j)l}-\tfrac{1}{5}\tilde{\gamma}_{ij}\tilde{\gamma}_{kl}) \tag{4i}$$

$$+(2\alpha\tilde{A}^{ns}+\tilde{\gamma}^{ns}\beta^{s}{}_{,m})(\tfrac{3}{5}\tilde{\gamma}_{k(i}\tilde{\gamma}_{j)n,s}-\tfrac{1}{5}\tilde{\gamma}_{ij}\tilde{\gamma}_{ks,n}), \tag{4j}$$

$$\dot{Q}_i-\beta^l Q_{i,l}=-\tfrac{1}{6}\alpha K_{,i}+\beta^l{}_{,i}Q_l-\tfrac{1}{6}\alpha_{,i}K, \tag{4k}$$

where, as before, $R_{ij}^{\mathrm{TF}}=R_{ij}-\frac{1}{3}\gamma^{ij}\gamma^{kl}R_{kl}$, with $R_{ij}=\tilde{R}_{ij}+R_{ij}^{\phi}$, and

$$R_{ij}^{\phi}=-2\tilde{D}_iQ_j-2\tilde{\gamma}_{ij}\tilde{\gamma}^{kl}\tilde{D}_kQ_l+4Q_iQ_j-4\tilde{\gamma}_{ij}\tilde{\gamma}^{kl}Q_kQ_l, \tag{5a}$$

$$\begin{aligned}\tilde{R}_{ij}=&-\tfrac{1}{2}V_{ijk,}{}^{k}+\tfrac{7}{10}\tilde{\gamma}_{k(i}\tilde{\Gamma}^{k}{}_{,j)}+\tfrac{1}{10}\tilde{\gamma}_{ij}\tilde{\Gamma}^{k}{}_{,k}-\tfrac{3}{10}(\tilde{\Gamma}^{k}V_{k(ij)}+\tilde{\Gamma}_{(i}V_{j)k}{}^{k}+\tfrac{9}{10}\tilde{\Gamma}_{i}\tilde{\Gamma}_{j})\\&-\tfrac{1}{5}(V_{ijk}\tilde{\Gamma}^{k}+\tilde{\gamma}_{ij}\tilde{\Gamma}^{k}V_{km}{}^{m}-\tfrac{1}{10}\tilde{\gamma}_{ij}\tilde{\Gamma}^{k}\tilde{\Gamma}_{k)}+\tilde{\Gamma}^{k}\tilde{\Gamma}_{(ij)k}+2\tilde{\Gamma}^{kl}{}_{(i}\tilde{\Gamma}_{j)kl}+\tilde{\Gamma}^{k}{}_{il}\tilde{\Gamma}_{kj}{}^{l},\end{aligned} \tag{5b}$$

$$\tilde{\Gamma}^{k}{}_{ij}=V^{k}{}_{(ij)}-\tfrac{1}{2}V_{ij}{}^{k}-\tfrac{1}{5}\delta^{k}_{(i}\tilde{\Gamma}_{j)}+\tfrac{2}{5}\tilde{\gamma}_{ij}\tilde{\Gamma}^{k}, \tag{5c}$$

and indices are raised and lowered with $\tilde{\gamma}^{ij}$ and $\tilde{\gamma}_{ij}$, respectively. The derivatives of the form $\tilde{\gamma}_{ij,k}$ that appear in the right-hand sides of (4) must be interpreted simply as shorthands for combinations of the fields $\tilde{\Gamma}^i$ and V_{ijk}, via

$$\tilde{\gamma}_{ij,k}=V_{ijk}+\tfrac{3}{5}\tilde{\gamma}_{k(i}\tilde{\Gamma}_{j)}-\tfrac{1}{5}\tilde{\gamma}_{ij}\tilde{\Gamma}_{k}. \tag{6}$$

For this first-order system to be equivalent to the Einstein equations (in the sense that its set of solutions in the same as that of the Einstein equations), the following sets of constraints must be imposed on the initial data (and are subsequently preserved by the evolution, as will be shown in the next section):

$$\mathcal{H}=\gamma^{ij}R_{ij}-\tilde{A}_{ij}\tilde{A}^{ij}+\tfrac{2}{3}K^2-2\rho, \tag{7a}$$

$$\mathcal{P}_i=\tilde{\gamma}^{jl}D_l\tilde{A}_{ij}-\tfrac{2}{3}D_iK+4Q_l\tilde{A}^{l}{}_{i}-S_i, \tag{7b}$$

$$\mathcal{G}^i=\tilde{\Gamma}^i+\tilde{\gamma}^{ij}{}_{,j}, \tag{7c}$$

$$\mathcal{Q}_i=Q_i-\phi_{,i}, \tag{7d}$$

$$\mathcal{V}_{ijk}=V_{ijk}-\tilde{\gamma}_{ij,k}+\tfrac{3}{5}\tilde{\gamma}_{k(i}\tilde{\gamma}_{j)n,s}\tilde{\gamma}^{ns}-\tfrac{1}{5}\tilde{\gamma}_{ij}\tilde{\gamma}_{kn,s}\tilde{\gamma}^{ns}, \tag{7e}$$

where

$$\gamma^{ij}R_{ij}=e^{-4\phi}(\tilde{\Gamma}^{l}{}_{,l}-8\tilde{D}^{l}Q_l-Q^lQ_l-\tfrac{1}{2}V_{ijl}V^{ijl}-\tfrac{15}{10}\tilde{\Gamma}^{k}V_{km}{}^{m}-\tfrac{1}{5}V^{m}{}_{mk}\tilde{\Gamma}^{k} \tag{7f}$$

$$-\frac{21}{100}\tilde{\Gamma}^{k}\tilde{\Gamma}_{k}+\tilde{\Gamma}^{k}\tilde{\Gamma}^{m}{}_{mk}+2\tilde{\Gamma}^{klm}\tilde{\Gamma}_{mkl}+\tilde{\Gamma}^{mkl}\tilde{\Gamma}_{mkl}). \tag{7g}$$

Constraints (7a) and (7b) are the Hamiltonian and momentum constants of the 3+1 decomposition of the Einstein equations, written in our choice of variables. Constraints (7c), (7d), and (7e) arise in turning the original second-order system into first order.

In (4) and (7), the derivative D_l is the covariant derivative with respect to γ_{ij}, and is related to $\tilde{D}_l$ by undifferentiated terms,

$$\Gamma^{k}{}_{ij}=\tilde{\Gamma}^{k}{}_{ij}+2(Q_i\delta^{k}_{j}+Q_j\delta^{k}_{i}-Q_l\tilde{\gamma}^{kl}\tilde{\gamma}_{ij}). \tag{8}$$

B. Taking advantage of the available freedom

In this section we take advantage of two facts that have been used successfully in similar problems.[5,7,17] First, we densitize the lapse α (and in doing so we introduce a free function, referred to as ''slicing density'' in Ref. 18),

$$\alpha = e^{4a\phi}\sigma. \tag{9}$$

Like the shift vector β^i, the lapse density σ will be considered arbitrary but fixed, a source function independent of the dynamical fields. Here a is a numerical parameter, not a function of the point.

Second, the evolution equations can be combined with the constraints without altering the set of solutions. We add the scalar constraint with a factor $b\alpha$ to the evolution equation for K and we add the vector constraint to the evolution equations for $\tilde{\Gamma}^i$ and Q_i, with factors of $c\alpha$ and $d\alpha$, respectively. Here b,c,d are numerical parameters, not functions of the point. In this manner we obtain a system of the form

$$\dot{u} = \mathbf{A}^i(u)\nabla_i u + B(u). \tag{10}$$

A system of this form is known to be well posed if the matrix-valued vector $\mathbf{A}^i(u)$ admits a symmetrizer, namely, a positive definite, symmetric, bilinear form $\mathbf{H}$, in the space of the fields u, whose product with $\mathbf{A}^i(u)$ yields a symmetric-bilinear-form-valued vector. Thus, in order to determine well-posedness, it suffices to consider the principal part of the system.

In this case, the principal terms are

$$\dot{\phi} = \beta^l \phi_{,l}, \tag{11a}$$

$$\dot{\tilde{\gamma}}_{ij} = \beta^l \tilde{\gamma}_{ij,l}, \tag{11b}$$

$$\dot{K} = \beta^l K_{,l} - \alpha(4a+8b)e^{-4\phi}\tilde{\gamma}^{kl}Q_{k,l} + \alpha b e^{-4\phi}\tilde{\Gamma}^l{}_{,l}, \tag{11c}$$

$$\dot{\tilde{A}}_{ij} = \beta^l \tilde{A}_{ij,l} + e^{-4\phi}\alpha\left(-\tfrac{1}{2}\tilde{\gamma}^{kl}V_{ijk,l} + \tfrac{7}{10}\left(\tilde{\gamma}_{k(i}\tilde{\Gamma}^k{}_{,j)} - \tfrac{1}{3}\tilde{\gamma}_{ij}\tilde{\Gamma}^k{}_{,k}\right)\right) - 2(2a+1)e^{-4\phi}\alpha\left(Q_{(i,j)} - \tfrac{1}{3}\tilde{\gamma}_{ij}\tilde{\gamma}^{kl}Q_{k,l}\right), \tag{11d}$$

$$\dot{\tilde{\Gamma}}^i = \beta^l \tilde{\Gamma}^i{}_{,l} + \alpha c \tilde{A}^{il}{}_{,l} - \tfrac{2}{3}(c+2)\alpha\tilde{\gamma}^{il}K_{,l}, \tag{11e}$$

$$\dot{V}_{ijk} = \beta^l V_{ijk,l} - 2\alpha\tilde{A}_{ij,k} + \tfrac{6}{5}\alpha\tilde{\gamma}_{k(i}\tilde{A}_{j)m,n}\tilde{\gamma}^{mn} - \tfrac{2}{5}\alpha\tilde{\gamma}_{ij}\tilde{A}_{km,n}\tilde{\gamma}^{mn}, \tag{11f}$$

$$\dot{Q}_i = \beta^l Q_{i,l} - \frac{\alpha}{6}(1+4d)K_{,i} + d\alpha\tilde{\gamma}^{jl}\tilde{A}_{ij,l}. \tag{11g}$$

Our aim is to show that there exist choices of the numerical factors a,b,c,d such that the system (11) is symmetrizable, therefore, well posed. We established this result by defining a candidate symmetrizer $\mathbf{H}$ given as

$$\tilde{u}\mathbf{H}u = \phi^2 + \delta^{ik}\delta^{jl}\tilde{\gamma}_{ij}\tilde{\gamma}_{kl} + n_1^2 e^{-4\phi}K^2 + \tilde{A}^{ij}\tilde{A}_{ij} + n_2^2 e^{-4\phi}\tilde{\Gamma}^i\tilde{\Gamma}_i + \frac{e^{-4\phi}}{4}V_{ijk}V^{ijk} + n_3^2 e^{-4\phi}Q^iQ_i, \tag{12}$$

where n_1, n_2, n_3 are any fixed real numbers different from zero and bounded, so that $C^{-1}\mathbf{I} \leqslant \mathbf{H} \leqslant C\mathbf{I}$, where C is a positive constant and $\mathbf{I}$ is the identity operator on the space of u.

We can easily arrange the values of a,b,c,d so that symmetry of $\mathbf{HA^i}(u)$ is attained. To this effect, they must satisfy

$$\tfrac{7}{10}=n_2^2 c, \tag{13a}$$

$$-2(2a+1)=n_3^2 d, \tag{13b}$$

$$-(4a+8b)n_1^2=-\tfrac{1}{6}(1+4d)n_3^2, \tag{13c}$$

$$n_1^2 b=-\tfrac{2}{3}n_2^2(c+2). \tag{13d}$$

There is clearly plenty of freedom in the choice of *a,b,c,d,* since any choice that results in nonvanishing n_1, n_2, n_3 is allowed. The freedom is thus parametrized by the values of n_1^2, n_2^2, n_3^2, since these can take independent positive values. Thus our four parameters *a,b,c,d* are not all independent, but there is a relationship between them that reduces the freedom to 3 independent parameters. We can solve (13) for *a,b,c,d* in terms of n_1, n_2, n_3, which yields

$$a=\frac{9/5+8n_2^2+n_3^2/8}{(3n_1^2+2)}, \tag{14a}$$

$$b=-\frac{2}{3n_1^2}\left(\frac{7}{10}+2n_2^2\right), \tag{14b}$$

$$c=\frac{7}{10n_2^2}, \tag{14c}$$

$$d=-\frac{2}{n_3^2}\left(2\,\frac{9/5+8n_2^2+n_3^2/8}{(3n_1^2+2)}+1\right). \tag{14d}$$

It is clear from (14) that a and c will take only strictly positive values, and b and d will take only strictly negative values, for all real values of n_1, n_2, n_3 different from zero.

C. Structure of characteristics

The system (11) is of the form

$$\mathbf{A}^a\frac{\partial u}{\partial x^a}=0. \tag{15}$$

The characteristic covectors are covectors $\xi_a=(\xi_{i,}-v)$ such that $\xi_i\xi_j\gamma^{ij}=1$ and such that

$$\det(\mathbf{A}^a\xi_a)=0. \tag{16}$$

The values of v that satisfy (16) for every direction ξ_i are the characteristic speeds in that direction. In order to find these values we set up an eigenvalue problem for the principal symbol $\mathbf{A}^a\xi_a$ and find the null eigenvectors. The eigenvalue problem is

$$n^a\xi_a\phi=0, \tag{17a}$$

$$n^a\xi_a\tilde{\gamma}_{ij}=0, \tag{17b}$$

$$n^a\xi_a K=-(4a+8b)e^{-4\phi}\xi^k Q_k+be^{-4\phi}\xi_l\tilde{\Gamma}^l, \tag{17c}$$

$$n^a\xi_a\tilde{A}_{ij}=e^{-4\phi}(-\tfrac{1}{2}\xi^k V_{ijk}+\tfrac{7}{10}(\tilde{\gamma}_{k(i}\xi_{j)}\tilde{\Gamma}^k-\tfrac{1}{3}\tilde{\gamma}_{ij}\xi_k\tilde{\Gamma}^k))-2(2a+1)e^{-4\phi}(\xi_{(j}Q_{i)}-\tfrac{1}{3}\tilde{\gamma}_{ij}\xi^k Q_k), \tag{17d}$$

$$n^a\xi_a\tilde{\Gamma}^i = c\,\xi_l\tilde{A}^{il} - \tfrac{2}{3}(c+2)\xi^i K, \tag{17e}$$

$$n^a\xi_a V_{ijk} = -2\xi_k\tilde{A}_{ij} + \tfrac{6}{5}\tilde{\gamma}_{k(i}\tilde{A}_{j)m}\xi^m - \tfrac{2}{5}\tilde{\gamma}_{ij}\xi^m\tilde{A}_{km}, \tag{17f}$$

$$n^a\xi_a Q_i = -\tfrac{1}{6}(1+4d)\xi_i K + d\,\xi^j\tilde{A}_{ij}, \tag{17g}$$

where $n^a=(1,-\beta^i)\alpha^{-1}$ is the normal to the slice. Clearly, $n^a\xi_a=0$ allows for 18 eigenvectors. This is because (17e), (17f), and (17g) in this case constitute an overdetermined system of 18 homogeneous equations for 6 unknowns $(\tilde{A}_{ij},K)$, with zero as the only solution, whereas (17c) and (17d) constitute a system of 6 equations for 18 variables, which leaves out 12 of the 18 fields $(V_{ijk},\tilde{\Gamma}^i,Q)$ free. Lastly, (17a) and (17b) leave the 6 variables $(\tilde{\gamma}_{ij},\phi)$ free. If we represent the eigenvectors in the form,

$$(\tilde{\gamma}_{ij},\phi,Q_i,\tilde{\Gamma}^{(L)},\tilde{\Gamma}_i^{(T)},V_{ij}^{(L)},V_{ijk}^{(T)},K,\tilde{A}^{(LL)},\tilde{A}_i^{(LT)},\tilde{A}_{ij}^{(TT)}), \tag{18}$$

where $\tilde{\Gamma}^{(L)}:=\tilde{\Gamma}^i\xi_i$, $\tilde{\Gamma}_i^{(T)}:=\tilde{\Gamma}_i - e^{-4\phi}\xi_i\tilde{\Gamma}^k\xi_k$, $V_{ij}^{(L)}:=V_{ijk}\xi^k$, $V_{ijk}^{(T)}:=V_{ijk}-e^{-4\phi}\xi_k V_{ijl}\xi^l$, $\tilde{A}^{(LL)}:=\tilde{A}^{ij}\xi_i\xi_j$, $\tilde{A}_i^{(LT)}:=\tilde{A}_{ij}\xi^j - e^{-4\phi}\xi_i\tilde{A}^{kl}\xi_k\xi_l$, $\tilde{A}_{ij}^{(TT)}:=\tilde{A}_{ij}-2e^{-4\phi}\xi_{(i}\tilde{A}_{j)l}\xi^l + e^{-8\phi}\xi_i\xi_j\tilde{A}^{kl}\xi_k\xi_l$, then we have 5 eigenvectors corresponding to the five components of the conformal metric,

$$(\tilde{\gamma}_{ij},0,0,0,0,0,0,0,0,0,0); \tag{19a}$$

we have the determinant as an eigenfield,

$$(0,\phi,0,0,0,0,0,0,0,0,0); \tag{19b}$$

we have 7 eigenvectors corresponding to the seven transverse components of V_{ijk},

$$(0,0,0,0,0,0,V_{ijk}^{(T)},0,0,0); \tag{19c}$$

we have 3 eigenvectors corresponding essentially to the three components of Q_i,

$$\left(0,0,Q_i,\frac{4a+8b}{b}\xi^i Q_{i,}0_{,}-2(2a+1)\xi_{(i}Q_{j)}-\frac{1}{3}\tilde{\gamma}_{ij}\left(\frac{7(4a+8b)}{10b}-2(2a+1)\right)Q_l\xi^l,0,0,0,0\right); \tag{19d}$$

and we have 2 eigenvectors corresponding essentially to the two components of the transverse part of $\tilde{\Gamma}_i$,

$$(0,0,0,0,\tilde{\Gamma}_i^{(T)},\tfrac{7}{10}\xi_{(i}\tilde{\Gamma}_{j)}^{(T)},0,0,0,0). \tag{19e}$$

If $n^a\xi_a\neq 0$, then $\tilde{\gamma}_{ij}=\phi=0$, and we can solve (17e), (17f), and (17g) for $(V_{ijk},\tilde{\Gamma}^i,Q)$ in terms of ξ_a and $(\tilde{A}_{ij},K)$. We can substitute $(V_{ijk},\tilde{\Gamma}^i,Q)$ into (17c) and (17d), obtaining thus a system of 6 equations for the 6 variables $(\tilde{A}_{ij},K)$ as follows:

$$0=Ke^{4\phi}((n^a\xi_a)^2-\tfrac{1}{6}(4a+8b)(1+4d)+\tfrac{2}{3}b(c+2))+\xi\cdot\tilde{A}\cdot\xi((4a+8b)d-bc), \tag{20a}$$

$$0=\tilde{A}_{ij}e^{4\phi}(1-(n^a\xi_a)^2)-\frac{K}{3}\left(\frac{7}{5}(c+2)-(2a+1)(1+4d)\right)\left(\xi_i\xi_j-\frac{1}{3}e^{4\phi}\tilde{\gamma}_{ij}\right)$$
$$+\left(\frac{7c}{10}-\frac{3}{5}-2d(2a+1)\right)\left(\xi^l\tilde{A}_{l(i}\xi_{j)}-\frac{1}{3}\tilde{\gamma}_{ij}\xi\cdot\tilde{A}\cdot\xi\right), \tag{20b}$$

where we have used the notation

$$\xi\cdot\tilde{A}\cdot\xi := \xi_i\tilde{A}^{ij}\xi_j. \tag{21}$$

If $1-(n^a\xi_a)^2=0$, then $K=\xi\cdot\tilde{A}\cdot\xi=0$ by (20a), which implies $\xi^l\tilde{A}_{li}=0$ by (20b). However, two of the five components of $\tilde{A}_{ij}$ are thus free, which means that there are 4 eigenvectors, essentially labeled by the transverse components of $\tilde{A}_{ij}$. We have 2 eigenvectors for $n^a\xi_a=1$,

$$(0,0,0,0,0,-2\tilde{A}_{ij}^{(TT)},0,0,0,0,\tilde{A}_{ij}^{(TT)}); \tag{22a}$$

and 2 eigenvectors for $n^a\xi_a=-1$,

$$(0,0,0,0,0,2\tilde{A}_{ij}^{(TT)},0,0,0,0,\tilde{A}_{ij}^{(TT)}). \tag{22b}$$

If $1-(n^a\xi_a)^2\neq 0$, then contracting (20b) with ξ^i yields

$$0=e^{4\phi}\xi^l\tilde{A}_{lj}\left(1+\frac{1}{2}\left(\frac{7c}{10}-\frac{3}{5}-2d(2a+1)\right)-(n^a\xi_a)^2\right)-\frac{2}{9}e^{4\phi}K\xi_j\left(\frac{7}{5}(c+2)-(2a+1)(1+4d)\right)$$
$$+\frac{1}{6}\xi\cdot\tilde{A}\cdot\xi\xi_j\left(\frac{7}{10}c-\frac{3}{5}-2d(2a+1)\right). \tag{23}$$

Thus, if $1+\frac{1}{2}(7c/10-3/5-2d(2a+1))-(n^a\xi_a)^2=0$, then $K=\xi\cdot\tilde{A}\cdot\xi=0$ by (20a), which implies that (23) is identically satisfied, thus three out of the five equations (20b) are identities, the remaining two determining two components of $\tilde{A}_{ij}$. Thus two of the five components of $\tilde{A}_{ij}$ are free, which means that there are 4 eigenvectors, essentially labeled by the two longitudinal-transverse components of $\tilde{A}_{ij}$. We have 2 eigenvectors for $n^a\xi_a=\sqrt{(3/5-7c/10-2d(2a+1))/2}$, namely,

$$\left(0,0,\frac{d}{C_1}\tilde{A}_i^{(LT)},0,\frac{c}{C_1}\tilde{A}_i^{(LT)},-\frac{4}{5C_1}\xi_{(i}\tilde{A}_{j)}^{(LT)},\frac{6}{5C_1}(\tilde{\gamma}_{k(i}\tilde{A}_{j)}^{(LT)}-\xi_k\xi_{(i}\tilde{A}_{j)}^{(LT)}\right.$$
$$\left.-\frac{1}{3}\tilde{\gamma}_{ij}\tilde{A}_k^{(LT)}),0,0,\tilde{A}_i^{(LT)},0\right), \tag{24a}$$

and 2 eigenvectors for $n^a\xi_a=-\sqrt{(3/5-7c/10-2d(2a+1))/2}$, namely,

$$\left(0,0,-\frac{d}{C_1}\tilde{A}_i^{(LT)},0,-\frac{c}{C_1}\tilde{A}_i^{(LT)},\frac{4}{5C_1}\xi_{(i}\tilde{A}_{j)}^{(LT)},-\frac{6}{5C_1}(\tilde{\gamma}_{k(i}\tilde{A}_{j)}^{(LT)}-\xi_k\xi_{(i}\tilde{A}_{j)}^{(LT)}\right.$$
$$\left.-\frac{1}{3}\tilde{\gamma}_{ij}\tilde{A}_k^{(LT)}),0,0,\tilde{A}_i^{(LT)},0\right), \tag{24b}$$

where we have used the shorthand notation

$$C_1 := \sqrt{(3/5-7c/10-2d(2a+1))/2}. \tag{25}$$

But if $1+\frac{1}{2}(7c/10-3/5-2d(2a+1))-(n^a\xi_a)^2\neq 0$, then $\xi^l\tilde{A}_{lj}$ is determined by the values of K and $\xi\cdot\tilde{A}\cdot\xi$ by (23), and if plugged back into (20b) it follows that all the components of $\tilde{A}_{ij}$ are determined by K and $\xi\cdot\tilde{A}\cdot\xi$. Therefore it is necessary that K and $\xi\cdot\tilde{A}\cdot\xi$ be nonvanishing. Contracting (23) with ξ^j we obtain

$$0=-\tfrac{2}{9}e^{4\phi}K(\tfrac{7}{5}(c+2)-(2a+1)(1+4d))+\xi\cdot\tilde{A}\cdot\xi(1+\tfrac{2}{3}(\tfrac{7}{10}c-\tfrac{3}{5}-2(2a+1)d)-(n^a\xi_a)^2). \tag{26}$$

Equations (20a) and (26) form a system of two homogeneous equations for K and $\xi\cdot\tilde{A}\cdot\xi$. Thus, for K and $\xi\cdot\tilde{A}\cdot\xi$ to be nonvanishing, it is necessary that the determinant of the system be zero. The determinant is

$$\tfrac{1}{45}(-3(n^a\xi_a)^2+2a)(15(n^a\xi_a)^2-9+10bc-80bd+20d-7c). \tag{27}$$

It can be seen that, because a and c are strictly positive and b and d are strictly negative, the four roots of the determinant are real. For the roots $n^a\xi_a$ of the determinant, we have

$$K=\frac{9}{2}\,\frac{1-(n^a\xi_a)^2+\dfrac{2}{3}\left(\dfrac{7c}{10}-\dfrac{3}{5}-2d(2a+1)\right)}{\dfrac{7}{5}(c+2)-(2a+1)(1+4d)}\,e^{-4\phi}\xi\cdot\tilde{A}\cdot\xi, \tag{28a}$$

$$Q_i=\left(d-\frac{3(1+4d)}{4}\,\frac{1-(n^a\xi_a)^2+\dfrac{2}{3}\left(\dfrac{7c}{10}-\dfrac{3}{5}-2d(2a+1)\right)}{\dfrac{7}{5}(c+2)-(2a+1)(1+4d)}\right)\frac{e^{-4\phi}}{n^a\xi_a}\,\xi_i\xi\cdot\tilde{A}\cdot\xi, \tag{28b}$$

$$\tilde{\Gamma}^{(L)}=\left(c-3(c+2)\,\frac{1-(n^a\xi_a)^2+\dfrac{2}{3}\left(\dfrac{7c}{10}-\dfrac{3}{5}-2d(2a+1)\right)}{\dfrac{7}{5}(c+2)-(2a+1)(1+4d)}\right)\frac{\xi\cdot\tilde{A}\cdot\xi}{n^a\xi_a}, \tag{28c}$$

$$\tilde{\Gamma}_i^{(T)}=0, \tag{28d}$$

$$V_{ij}^{(L)}=-\frac{9}{5n^a\xi_a}\left(e^{-4\phi}\xi_i\xi_j-\frac{1}{3}\tilde{\gamma}_{ij}\right)\xi\cdot\tilde{A}\cdot\xi, \tag{28e}$$

$$V_{ijk}^{(T)}=-\frac{6}{5n^a\xi_a}(\xi_i\xi_j-\tilde{\gamma}_{k(i}\xi_{j)})e^{-4\phi}\xi\cdot\tilde{A}\cdot\xi, \tag{28f}$$

$$\tilde{A}_i^{(LT)}=0, \tag{28g}$$

$$\tilde{A}_{ij}^{(TT)}=\frac{e^{-4\phi}}{2}(e^{-4\phi}\xi_i\xi_j-\tilde{\gamma}_{ij})\xi\cdot\tilde{A}\cdot\xi. \tag{28h}$$

These are clearly four distinct eigenvectors, since there are four distinct values of $n^a\xi_a$ given by the roots of the determinant (27). These can be thought as being labeled, essentially, by K or $\xi\cdot\tilde{A}\cdot\xi$ indistinctly, or we can associate one characteristic speed to K and the other one to $\xi\cdot\tilde{A}\cdot\xi$, as we prefer to do below.

Summarizing, we have characteristic speeds obtained from the following distinct values of $n^a\xi_a$:

(a) $n^a\xi_a=0$, timelike, with eigenfields (essentially) $\tilde{\gamma}_{ij}$, ϕ, Q_i, $\tilde{\Gamma}^{(T)}$, $V_{ijk}^{(T)}$.
(b) $n^a\xi_a=1$, null, with eigenfields (essentially) $\tilde{A}_{ij}^{(TT)}$.
(c) $n^a\xi_a=(1+\frac{1}{2}(7c/10-3/5-2d(2a+1)))^{1/2}\equiv C_1$, with eigenfields (essentially) $\tilde{A}_i^{(TL)}$.
(d) $n^a\xi_a=(2a/3)^{1/2}\equiv C_2$, with eigenfield (essentially) K.
(e) $n^a\xi_a=(3/5-2bc/3+16bd/3-4d/3+7c/15)^{1/2}\equiv C_3$, with eigenfield (essentially) $\tilde{A}^{(LL)}$.

In the expressions for the characteristic speeds (c), (d), and (e), the parameters *a,b,c,d* are given in terms of n_1, n_2, n_3 via (14). These speeds may be superluminal or causal, depending on the values of n_1, n_2, n_3. We can choose n_1, n_2, n_3 so that C_1, C_2 and C_3 are all equal to 1. This is achieved by setting

$$n_1^2=\frac{4}{15}\,\frac{280n_2^2+49+400n_2^4}{60n_2^2-49}, \tag{29a}$$

$$n_3^2=\frac{6400n_2^2}{60n_2^2-49}, \tag{29b}$$

for any value of n_2^2 greater than 49/60. This means that there is a one-parameter family of well-posed conformally-decomposed systems with ''physical'' characteristics. From the analytical point of view, there does not appear to exist an argument for singling out a preferred value of n_2. It is likely that a preferred value of n_2 will be dictated by optimal numerical behavior. The expressions for *a,b,c,d* in terms of n_2, with n_1 and n_3 as above (29), are as follows:

$$a=\frac{3}{2}, \tag{30a}$$

$$b=-\frac{60n_2^2-49}{4(7+20n_2^2)}, \tag{30b}$$

$$c=\frac{7}{10n_2^2}, \tag{30c}$$

$$d=-\frac{60n_2^2-49}{800n_2^2}. \tag{30d}$$

D. Propagation of the constraints

The propagation of the constraints can be calculated by taking a time derivative of each one of the constraint expressions, and subsequently using the evolution equations (11) to eliminate the time derivative of the fields in the right-hand side in favor of spatial derivatives, which recombine to yield back the constraints. We obtain

$$\dot{\mathcal{H}}=\beta^l\mathcal{H}_{,l}+(c-8d)\alpha e^{-4\phi}\tilde{\gamma}^{kl}\mathcal{P}_{k,l}+\cdots, \tag{31a}$$

$$\dot{\mathcal{P}}_i=\beta^l\mathcal{P}_{i,l}+\frac{\alpha}{6}(1-4b)\mathcal{H}_{,i}-\frac{\alpha}{2}e^{-4\phi}\left(\tilde{\gamma}^{jl}\left(\tilde{\gamma}^{kr}\mathcal{V}_{ijk,rl}-\frac{7}{10}\tilde{\gamma}_{im}\mathcal{G}^m{}_{,jl}\right)+\frac{1}{10}\mathcal{G}^m{}_{,mi}\right)$$
$$-2\alpha(2a+1)e^{-4\phi}\tilde{\gamma}^{jl}\mathcal{Q}_{[i,l]j}+\cdots, \tag{31b}$$

$$\dot{\mathcal{G}}^i=\beta^l\mathcal{G}^i{}_{,l}+\cdots, \tag{31c}$$

$$\dot{\mathcal{Q}}_i=\beta^l\mathcal{Q}_{i,l}+\cdots, \tag{31d}$$

$$\dot{\mathcal{V}}_{ijk}=\beta^l\mathcal{V}_{ijk,l}+\cdots, \tag{31e}$$

where $\cdots$ denotes undifferentiated terms proportional to the constraints themselves. To analyze the constraint propagation we proceed to turn (31) into first order by defining several sets of variables which represent all the spatial derivatives of $\mathcal{V}_{ijk}$, $\mathcal{G}^i$, and $\mathcal{Q}_i$,

$$\mathcal{W}_{ij}=\mathcal{V}_{ijk,}{}^{k}-\tfrac{1}{5}\mathcal{G}_{(i,j)}+\tfrac{1}{15}\tilde{\gamma}_{ij}\mathcal{G}^{k}{}_{,k}\,, \tag{32a}$$

$$\mathcal{X}_{ijkl}=\mathcal{V}_{ijk,l}-\tfrac{1}{3}\tilde{\gamma}_{kl}\mathcal{V}_{ijm,}{}^{m}, \tag{32b}$$

$$\mathcal{U}_{ij}=\mathcal{Q}_{[i,j]}\,, \tag{32c}$$

$$\mathcal{Z}_{ij}=\mathcal{Q}_{(i,j)}\,, \tag{32d}$$

$$\mathcal{T}_{ij}=\mathcal{G}_{[i,j]}\,, \tag{32e}$$

$$\mathcal{J}_{ij}=\mathcal{G}_{(i,j)}+\tfrac{30}{7}A\,\mathcal{Q}_{(i,j)}\,, \tag{32f}$$

where A is a constant which will be fixed shortly. Calculating the time derivative of these we obtain the resulting first-order system of evolution of the constraints,

$$\dot{\mathcal{H}}=\beta^{l}\mathcal{H}_{,l}+\alpha(c-8d)e^{-4\phi}\mathcal{P}_{l,}{}^{l}+\cdots, \tag{33a}$$

$$\dot{\mathcal{P}}_i=\beta^{l}\mathcal{P}_{i,l}+\frac{\alpha}{6}(1-4b)\mathcal{H}_{,i}-\frac{\alpha}{2}e^{-4\phi}\mathcal{W}_{il,}{}^{l}-2\alpha(2a+1)e^{-4\phi}\mathcal{U}_{il,}{}^{l}-A\alpha e^{-4\phi}\mathcal{Z}_{il,}{}^{l}+\frac{7\alpha}{30}e^{-4\phi}\mathcal{J}_{il,}{}^{l}$$
$$+\frac{11\alpha}{30}e^{-4\phi}\mathcal{T}_{il,}{}^{l}+\cdots, \tag{33b}$$

$$\dot{\mathcal{W}}_{ij}=\beta^{l}\mathcal{W}_{ij,l}-\frac{c\alpha}{5}\mathcal{P}_{(i,j)}+\frac{c\alpha}{15}\tilde{\gamma}_{ij}\mathcal{P}_{l,}{}^{l}+\cdots, \tag{33c}$$

$$\dot{\mathcal{X}}_{ijkl}=\beta^{m}\mathcal{X}_{ijkl,m}+\cdots, \tag{33d}$$

$$\dot{\mathcal{U}}_{ij}=\beta^{l}\mathcal{U}_{ij,l}+\alpha d\mathcal{P}_{[i,j]}+\cdots, \tag{33e}$$

$$\dot{\mathcal{Z}}_{ij}=\beta^{l}\mathcal{Z}_{ij,l}+\alpha d\mathcal{P}_{(i,j)}+\cdots, \tag{33f}$$

$$\dot{\mathcal{T}}_{ij}=\beta^{l}\mathcal{T}_{ij,l}+\alpha c\mathcal{P}_{[i,j]}+\cdots, \tag{33g}$$

$$\dot{\mathcal{J}}_{ij}=\beta^{l}\mathcal{J}_{ij,l}+\alpha\left(c+\frac{30}{7}Ad\right)\mathcal{P}_{(i,j)}+\cdots, \tag{33h}$$

$$\dot{\mathcal{G}}^{i}=\beta^{l}\mathcal{G}^{i}{}_{,j}+\cdots, \tag{33i}$$

$$\dot{\mathcal{Q}}_i=\beta^{l}\mathcal{Q}_{i,l}+\cdots, \tag{33j}$$

$$\dot{\mathcal{V}}_{ijk}=\beta^{l}\mathcal{V}_{ijk,l}+\cdots. \tag{33k}$$

For this system there is a symmetrizer given by

$$\overline{u}\mathbf{H}_c u=e^{4\phi}\frac{(1-4b)}{6(c-8d)}\mathcal{H}^2+\mathcal{P}_i\mathcal{P}^i+e^{-4\phi}\frac{5}{2c}\mathcal{W}_{ij}\mathcal{W}^{ij}+\mathcal{X}_{ijkl}\mathcal{X}^{ijkl}-e^{-4\phi}\frac{2(2a+1)}{d}\mathcal{U}_{ij}\mathcal{U}^{ij}$$
$$-e^{-4\phi}\frac{A}{d}\mathcal{Z}_{ij}\mathcal{Z}^{ij}+e^{-4\phi}\frac{11}{30c}\mathcal{T}_{ij}\mathcal{T}^{ij}+e^{-4\phi}\frac{7/30}{c+Ad30/7}\mathcal{J}_{ij}\mathcal{J}^{ij}+\mathcal{G}_i\mathcal{G}^i+\mathcal{Q}_i\mathcal{Q}^i+\mathcal{V}_{ijk}\mathcal{V}^{ijk}. \tag{34}$$

Taking $A=\frac{7}{60}(-c/d)$, $\mathbf{H}_C$ is positive definite because, under the conditions (14), all the factors accompanying the squares of the fields are strictly positive. This shows that no additional restrictions on the ranges of the parameters *a,b,c,d* are necessary in order to have well posed constraint evolution.

III. CONCLUSION

We have derived a 3-parameter family of well posed versions of the conformally-decomposed 3+1 equations, perhaps amenable to successful numerical integration. One might object that there is no need for it in view of the results in Ref. 11, but we can argue rather strongly that these results may prove helpful in pinning-down the main cause of numerical instabilities. This well posed version requires the lapse to be proportional to the determinant of the intrinsic geometry of the surfaces, and requires combinations of the constraints with the evolution equations. The lapse density σ and the shift vector β^i are arbitrary nondynamical variables, which means that they must be specified as free source functions. This well posed version uses the same variables as the original system (except for the addition of the first spatial derivatives of the densitized 3-metric, referred to as ''conformal metric'' by the authors Ref. 11). In addition, this well-posed version of the original equations propagates the constraints in a stable manner, which is relevant to unconstrained evolution. We think that this is the least invasive way to turn the original conformally-decomposed system into a well posed one. In practice, a choice of the numerical parameters n_1, n_2, n_3 must be made. The characteristic speeds depend on this choice.

Optimal choices of the parameters n_1, n_2, n_3 for numerical evolution are those that ensure that the characteristics are all either null or timelike. With such a choice, the formulation would be suited to evolve blackhole space–times outside the event horizon. Among these choices, it has been suggested[19] that the preferred one would be the one for which the characteristics are all ''physical,'' namely, either null or normal to the slices. We have shown that such a choice is possible for an arbitrary $n_2>\sqrt{49/60}$.

The systems obtained in this work are not contained in our previous work.[5,7] The choice of variables in Refs. 5,7 is inadequate for decomposing the trace and trace free part of the extrinsic curvature, as well as for extracting the determinant of the 3-metric. This is clear from the fact that, in that work, the available parameters α and β are not allowed to take the value $-1/3$ without the argument breaking down.

The systems obtained here differ significantly from the system obtained in Ref. 15 by considering the trace of the extrinsic curvature K and the determinant of the intrinsic metric (and its derivatives) as dynamical variables on equal footing with the rest, rather than as free source functions. Furthermore, we have obtained a 3-parameter family of systems, one system for each appropriate choice of n_1, n_2, n_3, whereas in Ref. 15 there is only one system which preserves the trace conditions. Additionally, we have separated the divergence of the intrinsic metric $\tilde{\Gamma}^i$ from the divergence-free part of the metric. This decomposition keeps up with the spirit of Ref. 11.

We have found that in obtaining these well-posed formulations the lapse must be proportional to some power of the determinant of the intrinsic metric, since the parameter *a* cannot take the value 0. This is similar to our findings in Refs. 2,5,7, as well as other notable cases.[6,20,21,18] In our present case this is remarkable, since we have used quite general energy norms.

ACKNOWLEDGMENTS

We thank Roberto Gómez for enlightening discussions. This work has been supported by CONICOR, CONICET, SeCyT (UNC), and by NSF under Grants Nos. PHY-9803301 and PHY-9407194.

[1] C. Bona and J. Masso, Phys. Rev. Lett. **68**, 1097 (1992).
[2] S. Frittelli and O. A. Reula, Commun. Math. Phys. **166**, 221 (1994).
[3] C. Bona, J. Masso, E. Seidel, and J. Stela, Phys. Rev. Lett. **75**, 600 (1995).
[4] A. Abrahams, A. Anderson, Y. Choquet-Bruhat, and J. W. York, Phys. Rev. Lett. **75**, 3377 (1995).
[5] S. Frittelli and O. A. Reula, Phys. Rev. Lett. **76**, 4667 (1996).

[6] H. Friedrich, Class. Quantum Grav. **13**, 1451 (1996).
[7] S. Frittelli and O. A. Reula, ''Well posed first-order forms of the 3+1 Einstein equations'' (to be published).
[8] B. Gustaffson, H.-O. Kreiss, and J. Oliger, *Time-Dependent Problems and Difference Methods* (Wiley, New York, 1995).
[9] H. Kurki-Suonio, P. Laguna, and R. A. Matzner, Phys. Rev. D **48**, 3611 (1993).
[10] M. Shibata and T. Nakamura, Phys. Rev. D **52**, 5428 (1995).
[11] T. W. Baumgarte and S. L. Shapiro, Phys. Rev. D **59**, 024007 (1999).
[12] O. Brodbeck, S. Frittelli, P. Hübner, and O. A. Reula, J. Math. Phys. **40**, 909 (1999).
[13] H. Friedrich and G. B. Nagy, ''The initial boundary value problem for Einstein's field equations,'' AEI preprint 078, 1998.
[14] J. M. Stewart, Class. Quantum Grav. **15**, 2865 (1998).
[15] M. Alcubierre, B. Brugmann, M. Miller, and W.-M. Suen, ''A conformal hyperbolic formulation of the Einstein equations,'' preprint, gr-qc/9903030 (1999).
[16] J. W. York, in *Sources of Gravitational Radiation* (Cambridge University Press, Cambridge, 1979).
[17] Y. Choquet-Bruhat and T. Ruggeri, Commun. Math. Phys. **89**, 269 (1983).
[18] A. Anderson and James W. York, Phys. Rev. Lett. **81**, 1154 (1998).
[19] A. Anderson and James W. York, Phys. Rev. Lett. **82**, 4384 (1999).
[20] M. Iriondo, E. Leguizamón, and O. Reula, Phys. Rev. Lett. **79**, 4732 (1997).
[21] M. Iriondo, E. Leguizamón, and O. Reula, Adv. Theor. Math. Phys. **2**, 1075 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

The effective σ-model of multidimensional gravity

Martin Rainer[a)]
Center for Gravitational Physics and Geometry, 104 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802-6300 and Gravitationsprojekt, Mathematische Physik I, Institut für Mathematik, Universität Potsdam, PF 601553, D-11415 Potsdam, Germany

(Received 14 December 1998; accepted for publication 13 May 1999)

The properties of the effective σ-model for D-dimensional Einstein gravity based on multidimensional geometries is analyzed. Besides pure geometry additional minimally coupled scalars and $(p+2)$-forms are considered which yield an extended target space after reduction to the effective D_0-dimensional geometry. The target space is always a homogeneous space. Exact solutions exist provided an orthobrane condition is satisfied which geometrically makes the target space a locally symmetric one. New solutions with scalar fields are found which may inflate not only in time-like but in also in additional spatial directions of the effective geometry. Static spherically symmetric solutions with a particular configuration of intersecting electric and magnetic branes are investigated both, for the orthobrane case and for degenerated charges. In both cases T_H depends critically on the intersection dimension of the branes. Finally, the role of the Einstein frame for 4-geometries is addressed, and the physical frame transformation for cosmological geometries is given.
[S0022-2488(99)03408-8]

I. INTRODUCTION

Historically σ-models have turned out to be a very powerful tool in many areas of physics. In gravity the importance was soon realized[1] in the context of solution generating techniques.[2] More recently, σ-models have been also discussed in the context of string theory.[3–5]

The purpose of this paper is to clarify the geometric structure of the effective σ-model for multidimensional Einstein geometry and to demonstrate its applicability in such different directions as cosmology, (extended) string theory, and quantization of certain higher-dimensional geometric actions.

In fact it turns out to be a very powerful tool which, on one side, allows to test the geometric content of string and M-theory down to their concrete physical imprints in the physical space-time and, on the other side, prepares a well-defined class of classical higher-dimensional geometries for the canonical quantization program in dimension $D_0 \leqslant 4$ whenever this is applicable to pure Einstein gravity itself. In principle all cases with infinite number of degrees of freedom in dimension $D_0 = 4$ which can be canonically quantized have some analogous cases where additional extra dimensions add only a finite number of degrees of freedom without disturbing the integrability of the problem. These cases include of course also recently investigated midisuperspace 4-geometries. In the case of spherical symmetries, and more particular in the static case, one can find particular solutions to a classical system of the multidimensional Einstein action with scalar and antisymmetric $p+2$-form fields which are multidimensional extensions of black hole solutions. It turns out that the standard surface gravity and the Hawking temperature T_H as calculated from a Komar-like integral depend sensitively on the intersection dimension of the p-branes involved in the solution. This provides, at least in principle, an observational window to very direct geometrical properties of possible extra dimensions. Apart from that, the multidimensional

[a)]Electronic mail: rainer@phys.psu.edu

σ-model contains all kinds of multidimensional spatially homogeneous cosmological models as degenerate minisuperspace cases with a finite number of degrees of freedom only.

Below, the effective D_0-dimensional σ-model is derived from a multidimensional action of Einstein type in a higher dimension D, first for pure geometry, then with additional scalar and antisymmetric $p+2$-form matter fields. The domains of the $p+1$-form potentials of the antisymmetric $p+2$-forms are the world-sheets of p-branes. In extended string and M-theory[6–8] strings are generalized to membranes as higher-dimensional objects. Most of these unified models are modeled initially on a higher-dimensional space-time manifold, say of dimension $D>4$, which then undergoes some scheme of spontaneous compactification.

The geometric structure of the target-space is clarified. In particular it is shown that it is a always a homogeneous space. It is furthermore locally symmetric if and only if the characteristic target-space vectors satisfy a particular orthogonality condition, called the *orthobrane* relation whenever they are not identical. In any case, it turns out possible to express the *general* exact solutions in terms of elementary functions, provided the input parameters of the model satisfy them, whence the target space is locally symmetric.

Solutions of the corresponding field equations are discussed generally and with concrete examples. Particular solutions for the subcases with Ricci flat internal spaces with scalar fields only, and with intersecting p-branes are presented. In the subcase of spherically symmetric solutions the relation to particles and black p-branes is given. Although a priori one might admit all possible types of components of F-fields compatible with spherical symmetry, namely, electric, magnetic and quasiscalar ones, we concentrate on true electric and magnetic type fields, since these are the ones which admit black hole solutions.

Besides the orthobrane solutions which by now became popular in string theory, there are further families of solutions, which have another additional symmetry, e.g., coinciding F-field charges for the electromagnetic solutions. In target space this additional symmetry is expressed by a linear relation between certain column vectors of the coupling matrix. In this case the original orthobrane conditions reduce to some weaker set of orthogonality conditions.

In the case of static, spherical symmetric solutions it is demonstrated that the formal Hawking temperature T_H (as it might appear to an observer at infinity) depends sensitively on the intersection dimension of the p-branes. Hence solutions to the multidimensional σ-model allow to detect possible imprints from extra-dimensional internal factor spaces within the physical dimension $D_0=4$. The black hole solutions depend on 3 integration constants, related to the electric, the magnetic, and the mass charge. It is also shown that the Hawking temperature of such black holes depends on the intersection dimension $d_{\rm int}$ of the corresponding p-branes. In an extremal limit of the charges, the black hole temperature turns out to converge to zero for $d_{\rm int}=0$, to a finite limit for $d_{\rm int}=1$, and to infinity for $d_{\rm int}>1$.

Finally it is shown how the geometries of well-known solutions in a Brans–Dicke frame can be transformed to the physically relevant Einstein frame.

II. PURE MULTIDIMENSIONAL GRAVITY

For the purpose of this paper let a (C^∞-) *multidimensional* (MD) manifold N be topologically just defined by a C^∞-fiber bundle

$$M \hookrightarrow N \to \overline{M}_0 \tag{2.1}$$

with a direct product

$$M := \times_{i=1}^{n} M_i \tag{2.2}$$

of internal C^∞ factor spaces M_i, $i=1,\dots,n$, as a standard fiber, and a distinguished C^∞ base manifold $\overline{M}_0$. (Later, for considerations of dynamics and cosmology we will set in particular $\overline{M}_0 := \mathbb{R}\times M_0$, and for the connection representation of Einstein gravity $D_0 := 4$ will be required.)

The MD manifold N is called *internally homogeneous* if there exists a direct product group $G:=\otimes_{i=0}^{n}G_i$ with a direct product realization $\tau:=\otimes_{i=0}^{n}\tau_i$ on $\mathrm{Diff}(M):=\otimes_{i=0}^{n}\mathrm{Diff}(M_i)$ such that for $i=0,\ldots,n$ the realization

$$\tau_i: G_i \to \mathrm{Diff}(M_i) \tag{2.3}$$

yields a transitive action of $\tau_i(G_i)$ on M_i.

Definition: A (C^∞) Riemannian manifold (M,g) (of arbitrary signature) is a C^∞ manifold M equipped with a symmetric bilinear C^∞ section $g: M \to \mathfrak{T}_2^0 M$ called metric. Unless specified otherwise the metric g will always be assumed to be nondegenerate. □

Definition: Given a Riemannian manifold (M,g), a diffeomorphism $\chi \in \mathrm{Diff}(M)$ is called an *isometry* of (M,g) whenever it leaves g invariant, i.e., whenever

$$g_{\chi(p)}=g_p \;\; \forall p \in M. \tag{2.4}$$

The very fact that a given diffeomorphism $\chi \in \mathrm{Diff}(M)$ may be an isometry on some metric but not on another one is the reason why the action of $\mathrm{Diff}(M)$ is not free on the space $\mathrm{Met}(M)$ of C^∞-metrics on M, whence $\mathrm{Geom}(M):=\mathrm{Met}(M)/\mathrm{Diff}(M)$ is in general not a manifold.

Definition: A Riemannian manifold (M,g) is called homogeneous, whenever M is homogeneous with a corresponding group G having a transitive realization $\tau(G)\subset \mathrm{Diff}(M)$ which leaves g invariant, i.e.,

$$g_{\chi(p)}=g_p \;\; \forall p \in M \;\; \forall \chi \in \tau(G). \tag{2.5}$$

□

Now for $i=0,\ldots,n$, let each factor space M_i be equipped with a smooth homogeneous metric $g^{(i)}$, rendering it into a homogeneous Riemannian manifold. Furthermore, let $\bar{M}_0$ be equipped with an arbitrary C^∞-metric $\bar{g}^{(0)}$, and let $\bar{\gamma}$ and β^i, $i=1,\ldots,n$ be smooth scalar fields on $\bar{M}_0$.

Then, under any projection $\mathrm{pr}: N \to \bar{M}_0$ a pullback of $e^{2\bar{\gamma}}\bar{g}^{(0)}$ from $x \in \bar{M}_0$ to $z \in \mathrm{pr}^{-1}\{x\}\subset M$, consistent with the fiber bundle (2.1) and the homogeneity of internal spaces, is given by

$$g_{(z)}:=e^{2\bar{\gamma}(x)}\bar{g}^{(0)}_{(x)} \oplus_{i=1}^{n} e^{2\beta^i(x)} g^{(i)}. \tag{2.6}$$

The function $\bar{\gamma}$ fixes a *gauge* for the (Weyl) *conformal frame* on $\bar{M}_0$, corresponding just to a particular choice of geometrical variables.

$\bar{\gamma}$ uniquely defines the form of the effective D_0-dimensional theory. For example $\bar{\gamma}:=0$ defines the Brans–Dicke frame.

Let us now consider a multidimensional manifold N (2.1) of dimension $D=D_0+\Sigma_{i=0}^{n}d_i$, equipped with a (pseudo) Riemannian metric (2.6) where

$$g^{(i)} \equiv g_{m_i n_i}(y_i)\, dy_i^{m_i} \otimes dy_i^{n_i}, \tag{2.7}$$

are R-homogeneous Riemannian metrics on M_i (i.e., the Ricci scalar $R[g^{(i)}]\equiv R_i$ is a constant on M_i), in coordinates $y_i^{n_i}$, $n_i=1,\ldots,d_i$, and

$$x \mapsto \bar{g}^{(0)}(x)=\bar{g}^{(0)}_{\mu\nu}(x)\, dx^\mu \otimes dx^\nu \tag{2.8}$$

yielding a general, not necessarily R-homogeneous, (pseudo) Riemannian metric on $\bar{M}_0$.

With (2.6) is a multidimensional generalization of the warped product of Ref. 9, namely $N=\bar{M}_0 \times_a M$, where $a:=e^{\beta}$ is now a *vector-valued* root warping function, given by

$$\beta := \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}. \tag{2.9}$$

Below sometimes, in particular for physical application to the Diff($\overline{M}_0$)-invariant case with $D_0 = 4$, we will assume the $i=0$ geometry to be empty and omit corresponding empty contributions to tensors, summations, etc. For later convenience we also define

$$\varepsilon(I) := \prod_{i\in I} \varepsilon_i, \quad \sigma_0 := \sum_{i=0}^{n} d_i\beta_i, \quad \sigma_1 := \sum_{i=1}^{n} d_i\beta_i, \quad \sigma(I) := \sum_{i\in I} d_i\beta_i, \tag{2.10}$$

where $\varepsilon_i := \mathrm{sign}(|g^{(i)}|)$ and $M_i \subset M$ for $i=0,\ldots,n$ are all homogeneous factor spaces. Here and below, we use the shorthand $|g| := |\det(g_{MN})|$, $|\overline{g^{(0)}}| := |\det(\overline{g^{(0)}_{\mu\nu}})|$, and analogously for all other metrics including $g^{(i)}$, $i=1,\ldots,n$.

Further, a $\overline{g^{(0)}}$-covariant derivative of a given function α w.r.t. x^μ is denoted by $\alpha_{;\mu}$, its partial derivative also by $\alpha_{,\mu}$, and $(\partial\alpha)(\partial\beta) := \overline{g^{(0)\mu\nu}}\alpha_{,\mu}\beta_{,\nu}$.

On $\overline{M}_0$, the Laplace–Beltrami operator

$$\Delta[\overline{g^{(0)}}] = \frac{1}{\sqrt{|\overline{g^{(0)}}|}} \frac{\partial}{\partial x^\mu}\left(\sqrt{|\overline{g^{(0)}}|}\,\overline{g^{(0)\mu\nu}} \frac{\partial}{\partial x^\nu}\right)$$

transforms under the conformal map $\overline{g^{(0)}} \mapsto e^{2\overline{\gamma}}\overline{g^{(0)}}$ according to

$$\begin{aligned}\Delta[e^{2\overline{\gamma}}\overline{g^{(0)}}] &= e^{-2\overline{\gamma}}\Delta[\overline{g^{(0)}}] - e^{-2\overline{\gamma}}\overline{g^{(0)\mu\nu}}(\Gamma[e^{2\overline{\gamma}}\overline{g^{(0)}}] - \Gamma[\overline{g^{(0)}}])^\lambda_{\mu\nu}\frac{\partial}{\partial x^\lambda} \\ &= e^{-2\overline{\gamma}}\left(\Delta[\overline{g^{(0)}}] + (D_0-2)g^{(0)\mu\nu}\frac{\partial\overline{\gamma}}{\partial x^\mu}\frac{\partial}{\partial x^\nu}\right),\end{aligned} \tag{2.11}$$

where Γ denotes the Levi-Civita connection.

The Levi-Civita connection Γ corresponding to (2.6) does *not* decompose multidimensionally, and neither does the Riemann tensor. The latter is a section in $\mathfrak{T}^1_3 M$ which is not given as a pullback to $\overline{M}_0$ of a section in the direct sum $\oplus_{i=1}^n \mathfrak{T}^1_3 M_i$ of corresponding tensor bundles over the factor manifolds.

However, with (2.6) the Ricci tensor decomposes again multidimensionally:

$$\mathrm{Ric}[g] = \mathrm{Ric}^{(0)}[g^{(0)},\overline{\gamma};\phi] \oplus_{i=1}^n \mathrm{Ric}^{(i)}[g^{(0)},\overline{\gamma};g^{(i)},\phi], \tag{2.12}$$

where

$$\begin{aligned}\mathrm{Ric}^{(0)}_{\mu\nu} &:= R_{\mu\nu}[g^{(0)}] + g^{(0)}_{\mu\nu}\left\{ -\Delta[g^{(0)}]\overline{\gamma} + (2-D_0)(\partial\overline{\gamma})^2 - \partial\overline{\gamma}\sum_{j=1}^n d_j\partial\phi^j\right\} \\ &\quad + (2-D_0)(\overline{\gamma}_{;\mu\nu} - \overline{\gamma}_{,\mu}\overline{\gamma}_{,\nu}) - \sum_{i=1}^n d_i(\phi^i_{;\mu\nu} - \phi^i_{,\mu}\overline{\gamma}_{,\nu} - \phi^i_{,\nu}\overline{\gamma}_{,\mu} + \phi^i_{,\mu}\phi^i_{,\nu}),\end{aligned}$$

$$\mathrm{Ric}^{(i)}_{m_i n_i} := R_{m_i n_i}[g^{(i)}] - e^{2\phi^i - 2\overline{\gamma}} g^{(i)}_{m_i n_i}\left\{ \Delta[g^{(0)}]\phi^i + (\partial\phi^i)\left[(D_0-2)\partial\overline{\gamma} + \sum_{j=1}^n d_j\partial\phi^j\right]\right\}, \quad i=1,\ldots,n, \tag{2.13}$$

The corresponding Ricci curvature scalar reads

$$R[g]=e^{-2\bar{\gamma}}R[\bar{g}^{(0)}]+\sum_{i=1}^{n} e^{-2\beta^i}R[g^{(i)}]-e^{-2\bar{\gamma}}\bar{g}^{(0)\mu\nu}\Bigg((D_0-2)(D_0-1)\frac{\partial\bar{\gamma}}{\partial x^\mu}\frac{\partial\bar{\gamma}}{\partial x^\nu}$$
$$+\sum_{i,j=1}^{n}(d_i\delta_{ij}+d_id_j)\frac{\partial\beta^i}{\partial x^\mu}\frac{\partial\beta^j}{\partial x^\nu}+2(D_0-2)\sum_{i=1}^{n}d_i\frac{\partial\bar{\gamma}}{\partial x^\mu}\frac{\partial\beta^i}{\partial x^\nu}\Bigg)$$
$$-2e^{-2\bar{\gamma}}\Delta[\bar{g}^{(0)}]\left((D_0-1)\bar{\gamma}+\sum_{i=1}^{n}d_i\beta^i\right). \tag{2.14}$$

Let us now set

$$f\equiv f[\bar{\gamma},\beta]:=(D_0-2)\bar{\gamma}+\sum_{j=1}^{n}d_j\beta^j, \tag{2.15}$$

where β is the vector field with the dilatonic scalar fields β^i as components. (Note that f can be resolved for $\bar{\gamma}\equiv\bar{\gamma}[f,\beta]$ if and only if $D_0\neq 2$. The singular case $D_0=2$ is discussed in Ref. 10.) Then, (2.14) can also be written as

$$R[g]-e^{-2\bar{\gamma}}R[\bar{g}^{(0)}]-\sum_{i=1}^{n}e^{-2\beta^i}R_i$$
$$=-e^{-2\bar{\gamma}}\left\{\sum_{i=1}^{n}d_i(\partial\beta^i)^2+(\partial f)^2+(D_0-2)(\partial\bar{\gamma})^2+2\Delta[\bar{g}^{(0)}](f+\bar{\gamma})\right\}$$
$$=-e^{-2\bar{\gamma}}\left\{\sum_{i=1}^{n}d_i(\partial\beta^i)^2+(D_0-2)(\partial\bar{\gamma})^2-(\partial f)\partial(f+2\bar{\gamma})+R_B\right\}, \tag{2.16}$$

$$R_B:=\frac{1}{\sqrt{|\bar{g}^{(0)}|}}e^{-f}\partial_\mu[2e^f\sqrt{|\bar{g}^{(0)}|}\bar{g}^{(0)\mu\nu}\partial_\nu(f+\bar{\gamma})], \tag{2.17}$$

where the last term will yield just a boundary contribution (2.22) to the action (2.21) below.

Let us assume all M_i, $i=1,\dots,n$, to be connected and oriented. The Riemann–Lebesgue volume form on M_i is denoted by

$$\tau_i:=\mathrm{vol}(g^{(i)})=\sqrt{|g^{(i)}(y_i)|}dy_i^1\wedge\cdots\wedge dy_i^{d_i}, \tag{2.18}$$

and the total internal space volume by

$$\mu:=\prod_{i=1}^{n}\mu_i,\quad \mu_i:=\int_{M_i}\tau_i=\int_{M_i}\mathrm{vol}(g^{(i)}). \tag{2.19}$$

If all of the spaces M_i, $i=1,...,n$ are compact, then the volumes μ_i and μ are finite, and so are also the numbers $\rho_i=\int_{M_i}\mathrm{vol}(g^{(i)})R[g^{(i)}]$. However, a non-compact M_i might have infinite volume μ_i or infinite ρ_i. Nevertheless, by the R-homogeneity of $g^{(i)}$ (in particular satisfied for Einstein spaces), the ratios $\rho_i/\mu_i=R[g^{(i)}]$, $i=1,\dots,n$, are just finite constants. In any case, the D-dimensional coupling constant κ can be tuned such that, under the dimensional reduction $\mathrm{pr}:M\to\bar{M}_0$,

$$\kappa_0:=\kappa\cdot\mu^{-\frac{1}{2}} \tag{2.20}$$

becomes the D_0-dimensional physical coupling constant. If $D_0=4$, then $\kappa_0^2=8\pi G_N$, where G_N is the Newton constant. The limit $\kappa\to\infty$ for $\mu\to\infty$ is in particular harmless, if D-dimensional gravity is given purely by curvature geometry, without additional matter fields. If, however, this geometry is coupled with finite strength to additional (matter) fields, one should indeed better take care to have all internal spaces M_i, $i=1,...,n$ compact. Often this can be achieved by factorizing with an appropriate finite symmetry group.

With the total dimension D, κ^2 a D-dimensional gravitational constant we consider a purely gravitational action of the form

$$S=\frac{1}{2\kappa^2}\int_N d^D z\sqrt{|g|}\{R[g]\}+S_{\mathrm{GHY}}. \tag{2.21}$$

Here a (generalized) Gibbons–Hawking–York[11,12] type boundary contribution S_{GHY} to the action is taken to cancel boundary terms. Equations (2.16) and (2.16) show that S_{GHY} should be taken in the form

$$S_{\mathrm{GHY}}:=\frac{1}{2\kappa^2}\int_N d^D z\sqrt{|g|}\{e^{-2\bar{\gamma}}R_B\}=\frac{1}{\kappa_0^2}\int_{\bar{M}_0} d^{D_0}x\,\frac{\partial}{\partial x^\lambda}\left(e^f\sqrt{|\overline{g^{(0)}}|}\,\overline{g}^{(0)\lambda\nu}\frac{\partial}{\partial x^\nu}(f+\bar{\gamma})\right), \tag{2.22}$$

which is just a pure boundary term in form of an effective D_0-dimensional flow through $\partial\bar{M}_0$.

After dimensional reduction the action (2.21) reads

$$\begin{aligned}S=\frac{1}{2\kappa_0^2}\int_{\bar{M}_0} d^{D_0}x\sqrt{|\overline{g^{(0)}}|}e^f\Bigg\{&R[\overline{g^{(0)}}]+(\partial f)(\partial[f+2\bar{\gamma}])\\&-\sum_{i=1}^n d_i(\partial\beta^i)^2-(D_0-2)(\partial\bar{\gamma})^2+e^{2\bar{\gamma}}\left[\sum_{i=1}^n e^{-2\beta^i}R_i\right]\Bigg\},\end{aligned} \tag{2.23}$$

where e^f is a dilatonic scalar field coupling to the D_0-dimensional geometry on $\bar{M}_0$.

According to the considerations above, due to the conformal reparametrization invariance of the geometry on $\bar{M}_0$, we should fix a conformal frame on $\bar{M}_0$. But then in (2.23) $\bar{\gamma}$, and with (2.15) also f, is no longer independent from the vector field β, but rather

$$\bar{\gamma}\equiv\bar{\gamma}[\beta],\quad f\equiv f[\beta]. \tag{2.24}$$

Then, modulo the conformal factor e^f, the dilatonic kinetic term of (2.23) takes the form

$$(\partial f)(\partial[f+2\bar{\gamma}])-\sum_{i=1}^n d_i(\partial\beta^i)^2-(D_0-2)(\partial\bar{\gamma})^2=-G_{ij}(\partial\beta^i)(\partial\beta^j), \tag{2.25}$$

with $G_{ij}\equiv{}^{(\bar{\gamma})}G_{ij}$, where

$${}^{(\bar{\gamma})}G_{ij}:={}^{(\mathrm{BD})}G_{ij}-(D_0-2)(D_0-1)\frac{\partial\bar{\gamma}}{\partial\beta^i}\frac{\partial\bar{\gamma}}{\partial\beta^j}-2(D_0-1)d_{(i}\frac{\partial\bar{\gamma}}{\partial\beta^{j)}}, \tag{2.26}$$

$${}^{(\mathrm{BD})}G_{ij}:=\delta_{ij}d_i-d_id_j. \tag{2.27}$$

For $D_0\neq2$, we can write equivalently $G_{ij}\equiv{}^{(f)}G_{ij}$, where

$${}^{(f)}G_{ij}:={}^{(\mathrm{E})}G_{ij}-\frac{D_0-1}{D_0-2}\frac{\partial f}{\partial\beta^i}\frac{\partial f}{\partial\beta^j}, \tag{2.28}$$

$$^{(\mathrm{E})}G_{ij} := \delta_{ij} d_i + \frac{d_i d_j}{D_0 - 2}. \tag{2.29}$$

For $D_0 = 1$, $G_{ij} = {}^{(\mathrm{E})}G_{ij} = {}^{(\mathrm{BD})}G_{ij}$ is independent of $\overline{\gamma}$ and f. Note that the metrics (2.27) and (2.29) (with $D_0 \neq 2$) may be diagonalized to $(\mp(\pm)^{\delta_{1D_0}})^{\delta_{1i}}\delta_{ij}$ respectively, by homogeneous linear minisuperspace coordinate transformations $\beta \overset{T}{\mapsto} z$ and $\beta \overset{Q}{\mapsto} \varphi$, explicitly given by components

$$z^1 := {}^{(\mathrm{BD})}q^{-1} \sum_{j=1}^{n} d_j \beta^j, \quad \varphi^1 := {}^{(\mathrm{E})}q^{-1} \sum_{j=1}^{n} d_j \beta^j,$$

$$z^i \equiv \varphi^i := [d_{i-1}/\Sigma_{i-1}\Sigma_i]^{1/2} \sum_{j=i}^{n} d_j (\beta^j - \beta^{i-1}), \tag{2.30}$$

$i = 2,\ldots,n$, where with $D' := D - D_0$ and $\Sigma_k := \Sigma_{i=k}^{n} d_i$,

$$^{(\mathrm{BD})}q := \sqrt{\frac{D'}{D'-1}}, \quad {}^{(\mathrm{E})}q := \sqrt{\frac{D'(D_0-2)}{D'+D_0-2}}. \tag{2.31}$$

So, after fixing a conformal reparametrization gauge for the geometry on M_0, (2.21) becomes a σ-model, where the vector field β (or z resp. φ) defines the coordinates of its n-dimensional target space. In the following, we will simplify notation by a summation convention for tensors over target space.

In general, for $n > 2$ and nonconstant functional $\gamma[\beta]$, the minisuperspace metric given by (2.25) and the conformally related target space metric may not even be conformally flat. However, for constant $\overline{\gamma}$, (2.26) reduces to (2.27), whence target space is conformally flat, namely it is related to n-dimensional Minkowski space by a conformal scale factor

$$\varphi \equiv \varphi(\beta) := \prod_{l=1}^{n} e^{d_l \beta^l} = e^{{}^{(\mathrm{BD})}q z^1} = e^{{}^{(\mathrm{E})}q \varphi^1}, \tag{2.32}$$

which is proportional to the total internal space volume.

In the case $D_0 \neq 2$, for nonconstant functional $f[\beta]$, the target space may again in general not be conformally flat for $n > 2$. However, for constant f, (2.28) reduces to (2.29), whence, target space is a flat n-dimensional space, namely an Euclidean one for $D_0 > 2$, and a Minkowskian one for $D_0 = 1$.

After gauging $\overline{\gamma}$, setting $m := \kappa_0^{-2}$, (2.23) yields a σ-model in the form

$$^{(\overline{\gamma})}S = \int_{\overline{M}_0} d^{D_0}x \sqrt{|\overline{g}^{(0)}|}\, {}^{(\overline{\gamma})}N^{D_0} \quad \varphi(\beta) \left\{ \frac{m}{2} {}^{(\overline{\gamma})}N^{-2} [R[\overline{g}^{(0)}] - {}^{(\overline{\gamma})}G_{ij}(\partial\beta^i)(\partial\beta^j)] - {}^{(\mathrm{BD})}V(\beta) \right\}, \tag{2.33}$$

$$\text{where} \quad {}^{(\mathrm{BD})}V(\beta) := m \left[-\frac{1}{2} \sum_{i=1}^{n} R[g^{(i)}] e^{-2\beta^i} \right], \tag{2.34}$$

$$^{(\overline{\gamma})}N := e^{\overline{\gamma}}. \tag{2.35}$$

Note that, the potential (2.34) and the conformal factor $\phi(\beta) := \Pi_{i=1}^{n} e^{d_i \beta^i}$ are gauge invariant.

Analogously, the σ-model action from (2.23) gauging f can also be written as

$$^{(f)}S=\int_{\overline{M}_0} d^{D_0}x\sqrt{|\overline{g^{(0)}}|}\,{}^{(f)}N^{D_0}\left\{\frac{m}{2}{}^{(f)}N^{-2}[R[\overline{g^{(0)}}]-{}^{(f)}G_{ij}(\partial\beta^i)(\partial\beta^j)]-{}^{(\mathrm{E})}V(\beta)\right\}, \quad (2.36)$$

$$^{(\mathrm{E})}V(\beta):=m\Omega^2\left[-\frac{1}{2}\sum_{i=1}^{n} R[g^{(i)}]e^{-2\beta^i}\right], \quad (2.37)$$

$$^{(f)}N:=e^{\frac{f}{D_0-2}}, \quad (2.38)$$

where the function Ω on $\overline{M}_0$ is defined as

$$\Omega:=\varphi^{\frac{1}{2-D_0}}. \quad (2.39)$$

Note that, with Ω also the potential (2.37) is gauge invariant, and the dilatonic target-space, though not even conformally flat in general, is flat for constant f.

In fact, Eqs. (2.33)–(2.35) and (2.36)–(2.38) show that there are at least two special frames.

The first one corresponds to the gauge $\overline{\gamma}\overset{!}{=}0$. In this case $^{(\overline{\gamma})}N=1$, the minisuperspace metric (2.26) reduces to the Minkowskian (2.27), the dilatonic scalar field becomes proportional to the internal space volume, $e^{f[\beta]}=\varphi(\beta)=\Pi_{i=1}^{n}e^{d_i\beta^i}$, and (2.33) describes a generalized σ-model with conformally Minkowskian target space. The Minkowskian signature implies a negative sign in the dilatonic kinetic term. This frame is usually called the Brans–Dicke one, because $\varphi=e^f$ here plays the role of a Brans–Dicke scalar field.

The second distinguished frame corresponds to the gauge $f\overset{!}{=}0$, where $\overline{\gamma}=[1/(2-D_0)]\Sigma_{i=1}^{n}d_i\beta^i$ is well defined only for $D_0\neq 2$. In this case $^{(f)}N=1$, the minisuperspace metric (2.28) reduces to the Euclidean (2.29), and (2.36) describes a self-gravitating σ-model with Euclidean target space. Hence all dilatonic kinetic terms have positive signs. This frame is usually called the Einstein one, because it describes an effective D_0-dimensional Einstein theory with additional minimally coupled scalar fields. For multidimensional geometries with $D_0=2$ the Einstein frame fails to exist, which reflects the well-known fact that two-dimensional Einstein equations are trivially satisfied without implying any dynamics.

For $D_0=1$, the action of both (2.33) and (2.36) was shown in Ref. 13 (and previously in Refs. 14 and 15) to take the form of a classical particle motion on minisuperspace, whence different frames correspond are just related by a time reparametrization. More generally, for $D_0\neq 2$ and $(\overline{M}_0,\overline{g}(0))$ a vacuum space-time, the σ-model (2.36) with the gauge $f\overset{!}{=}0$ describes the dynamics of a massive (D_0-1)-brane within a potential (2.37) on its target minisuperspace.

In fact, the target space is in general a conformally homogeneous space, and in the Einstein frame a homogeneous one. Once its isometry group $\mathfrak{G}$ and isotropy group $\mathfrak{H}$ are known, it is clear that the sigma model (2.36) can also be written in matrix form

$$^{(f)}S=\int_{\overline{M}_0} d^{D_0}x\sqrt{|\overline{g^{(0)}}|}N^{D_0}(\mathcal{M})\left\{\frac{m}{2}N^{-2}(\mathcal{M})[R[\overline{g^{(0)}}]+g^{(0)\mu\nu}B\,\mathrm{Tr}_{\rho}(\partial_\mu\mathcal{M}\partial_\nu\mathcal{M}^{-1})]\right.$$
$$\left.-{}^{(\mathrm{E})}U(\mathcal{M})\right\}, \quad (2.40)$$

with $\mathcal{M}\in\rho(G)$ where ρ is an appropriate coset representation of the target space $\mathfrak{M}:=\mathfrak{G}/\mathfrak{H}$, $^{(\mathrm{E})}U$ is now the corresponding potential on $\mathfrak{M}$, N a gauge function on $\mathfrak{M}$, and B a normalization.

For $D_0=4$, Eq. (2.40) can also be written in the Einstein frame as

$$^{(E)}S=\int_{\overline{M}_0}\left\{\frac{m}{2}[\mathrm{Tr}\,\Omega\wedge*\Sigma+B\,\mathrm{Tr}_{\rho}d\mathcal{M}\wedge*d\mathcal{M}^{-1}]-{}^{(\mathrm{E})}U(\mathcal{M})*1\right\}, \quad (2.41)$$

where Ω is the curvature 2-form, $\Sigma := e \wedge e$ and $\overline{g^{(0)}}$ are given by the D_0-dimensional soldering 1-form e, and the Hodge star is taken w.r.t. $(\overline{M}, g^{(0)})$. The form (2.41) is a then a convenient starting point for the canonical quantization procedure.

In the purely gravitational model consider so far $\mathfrak{M}$ is a finite dimensional and homogeneous with a transitive Abelian group. In the following section let us add minimally coupled scalar and $p+2$-form matter fields and investigate the extension of the resulting target space $\mathfrak{M}$.

III. σ-MODEL WITH MINIMALLY COUPLED SCALARS AND $p+2$-FORMS

We now couple the purely gravitational action (2.21) to additional matter fields of scalar and generalized Maxwell type, i.e. we consider now the action

$$2\kappa^2[S[g,\phi,F^a]-S_{\mathrm{GHY}}]=\int_N d^Dz\,\sqrt{|g|}\Bigg\{R[g]-C_{\alpha\beta}g^{MN}\partial_M\Phi^\alpha\partial_N\Phi^\beta - \sum_{a\in\Delta}\frac{\eta_a}{n_a!}\exp[2\lambda_a(\Phi)](F^a)^2\Bigg\} \tag{3.1}$$

of a self-gravitating σ model on M with topological term S_{GHY}. Here the l-dimensional target space, defined by a vector field ϕ with scalar components ϕ^α, $\alpha=1,\ldots,l$, is coupled to several antisymmetric n_a-form fields F^a via 1-forms λ_a, $a\in\Delta$. For consistency, we have to demand of course that all fields are internally homogeneous. We will see below how this gives rise to an effective $l+|\Delta|$-dimensional target-space extension. Note also that for convenience here we work with fields ϕ and F which differ from the actual (physical) matter fields by a rescaling with the square root of the coupling constant.

With $I\subset\{1,\ldots,n\}$, the generalized Maxwell fields F^a are located on (n_a-1)-dimensional world sheets

$$M_I:=\prod_{i\in I}M_i=M_{i_1}\times\cdots\times M_{i_k}, \tag{3.2}$$

$$n_a-1=D(I):=\sum_{i\in I}d_i=d_{i_1}+\cdots+d_{i_k} \tag{3.3}$$

of different $(n_a-2)-$branes, labeled for each a by the sets I in a certain subset Ω_a of the power set of $\{1,\ldots,n\}$. Variation of (3.1) yields the field equations

$$R_{MN}-\tfrac{1}{2}g_{MN}R=T_{MN}, \tag{3.4}$$

$$C_{\alpha\beta}\Delta[g]\phi^\beta-\sum_{a\in\Delta}\frac{\eta_a\lambda_a^\alpha}{n_a!}e^{2\lambda_a(\phi)}(F^a)^2=0, \tag{3.5}$$

$$\nabla_{M_1}[g](e^{2\lambda_a(\phi)}F^{a,M_1M_2\ldots M_{n_a}})=0, \tag{3.6}$$

$a\in\Delta$, $\alpha=1,\ldots,l$.

In (3.4) the D-dimensional energy-momentum resulting from (3.1) is given by a sum

$$T_{MN}:=\sum_{\alpha=1}^{l}T_{MN}[\phi^\alpha,g]+\eta_a\sum_{a\in\Delta}e^{2\lambda_a(\phi)}T_{MN}[F^a,g], \tag{3.7}$$

of contributions from scalar and generalized Maxwell fields,

$$T_{MN}[\phi^\alpha,g]:=C_{\alpha\beta}\partial_M\phi^\alpha\partial_N\phi^\alpha-\tfrac{1}{2}g_{MN}\partial_P\phi^\alpha\partial^P\phi^\alpha, \tag{3.8}$$

$$T_{MN}[F^a,g]:=\frac{1}{n_a!}\left[-\frac{1}{2}g_{MN}(F^a)^2+n_aF^a_{MM_2\ldots M_{n_a}}F^a{}_N{}^{M_2\ldots M_{n_a}}\right]. \tag{3.9}$$

We give now a sufficient criterion for the energy-momentum tensor (3.7) to decompose multidimensionally.

Let $W_1:=\{i|i>0,\ d_i=1\}$ be the label set of 1-dimensional factor spaces of the multidimensional decomposition, and set $n_1:=|W_1|$. Define

$$W(a;i,j):=\{(I,J)|I,J\in\Omega_a,(I\cap J)\cup\{i\}=I\not\ni j,(I\cap J)\cup\{j\}=J\not\ni i\} \tag{3.10}$$

Then the following holds.

Theorem. If for $n_1>1$ the p-branes satisfy the condition for all $a\in\Delta,\ i,j\in W_1$ with $i\neq j$, the condition

$$W(a;i,j)\overset{!}{=}\emptyset\forall a\in\Delta\forall i,j\in W_1, \tag{3.11}$$

then the energy-momentum (3.7) decomposes multidimensionally without further constraints.

Proof: The only possible obstruction to the multidimensional decomposition of (3.7) comes from the second term of (3.9), $F^a{}_{MM_2\ldots M_{n_a}}F^a{}_N{}^{M_2\ldots M_{n_a}}$ when the indices M and N take values in different index sets labeling different 1-dimensional factor spaces. The theorem then follows just from the antisymmetry of the F-fields. □

Corollary: A sufficient condition for the multidimensional decomposition of (3.7) is

$$n_1\overset{!}{\leqslant}1. \tag{3.12}$$

If condition (3.11) does not hold, multidimensional decomposability of (3.7) may impose additional nontrivial constraints on the $p+2$-form fields.

Let us now specify the components of the F-fields of generalized electric and magnetic type.

Antisymmetric fields of generalized electric type, are given by scalar potential fields $\Phi^{a,I}$, $a\in\Delta,\ I\in\Omega_a$, which compose to a $(\Sigma_{a\in\Delta}|\Omega_a|)$-dimensional vector field Φ. Magnetic type fields are just given as the duals of appropriate electric ones.

$$F^{e,I}=d\Phi^{e,I}\wedge\tau(I), \tag{3.13}$$

$$F^{m,I}=e^{-2\lambda_a(\phi)}*(d\Phi^{m,I}\wedge\tau(J)). \tag{3.14}$$

In the Einstein frame, the action then reduces to

$$\begin{aligned}{}^{(\mathrm{E})}S[g^{(0)},\beta,\phi,\Phi]=\int_{M_0}d^{D_0}x\sqrt{|g^{(0)}|}\Bigg\{\frac{m}{2}\Bigg[&R[g^{(0)}]-G_{ij}(\partial\beta^i)(\partial\beta^j)-C_{\alpha\beta}(\partial\phi^\alpha)(\partial\phi^\beta)\\ &-\sum_{a\in\Delta,I\in\Omega_a}\varepsilon_{a,I}e^{2(\lambda_a(\phi)-d_i\beta^i)}(\partial\Phi^{a,I})^2\Bigg]-{}^{(\mathrm{E})}V(\beta)\Bigg\},\end{aligned} \tag{3.15}$$

which corresponds to an purely Einsteinian σ-model on M_0 with extended $(n+l+\Sigma_{a\in\Delta}|\Omega_a|)$-dimensional target space and dilatonic potential (2.37). Here and below we will consider by default the Einstein frame, and set correspondingly $G_{ij}:={}^{(\mathrm{E})}G_{ij}$. In (3.15) and below a summation convention is assumed also on the extended target space.

For convenience, let us introduce the topological numbers

$$l_{jI}:=-\sum_{i\in I} D_i\delta^i_j, \quad j=1,\ldots,n, \tag{3.16}$$

and with $N:=n+l$ define and define a $N\times|S|$-matrix

$$L=(L_{As})=\begin{pmatrix} L_{is} \\ L_{\alpha s}\end{pmatrix}:=\begin{pmatrix} l_{iI} \\ \lambda_{\alpha a}\end{pmatrix}, \tag{3.17}$$

a N-dimensional vector field $(\sigma^A):=(\beta^i,\phi^\alpha)$, $A=1,\ldots,n,n+1,\ldots,N$, composed by dilatonic and matter scalar fields, and a nondegenerate (block-diagonal) $N\times N$-matrix

$$\hat{G}=(\hat{G}_{AB})=\begin{pmatrix} G_{ij} & 0 \\ 0 & C_{\alpha\beta}\end{pmatrix}. \tag{3.18}$$

With these definitions, (3.15) takes the form

$$S_0=\int_{M_0} d^{D_0}x\sqrt{|g^{(0)}|}\left\{\frac{m}{2}\left[R[g^{(0)}]-\hat{G}_{AB}\partial\sigma^A\partial\sigma^B-\sum_{s\in S}\varepsilon_s e^{2L_{As}\sigma^A}(\partial\Phi^s)^2\right]-{}^{(\mathrm{E})}V(\sigma)\right\}. \tag{3.19}$$

IV. SOLUTION WITH ABELIAN TARGET-SPACE

In the this section we consider the σ-model (3.15) without the Φ fields from the $p+2$-forms, whence the target-space is the $n+l$-dimensional Abelian one, and present a particularly interesting vacuum solution.

We derive for $D_0\neq 2$ an new exact Ricci flat multidimensional solution for the effective σ-model (3.15) in the harmonic gauge $(2-D_0)\overline{\gamma}\overset{!}{=}d_i\beta^i$ with zero potential (2.37) and zero Φ. The field equation then read

$$G_{ij}\partial_\mu\beta^i\partial_\nu\beta^j+C_{\alpha\beta}\partial_\mu\phi^\alpha\partial_\nu\phi^\beta=0,\quad \mu,\nu=0,\ldots,D_0-1, \tag{4.1}$$

$${}^{(\mathrm{E})}G_{ij}\Delta[\overline{g^{(0)}}]\beta^j=0,\quad i=1,\ldots,n, \tag{4.2}$$

$$C_{\alpha\beta}\Delta[\overline{g^{(0)}}]\phi^\beta=0,\quad \alpha=1,\ldots,l. \tag{4.3}$$

In particular, we now solve these equations with flat $(\overline{M}_0,\overline{g^{(0)}})$. In this case, there exist g-harmonic $\overline{M}_0$-coordinates τ^μ, $\mu=0,\ldots,D_0-1$. Let $g^{(0)}=e^{-2\gamma}\eta_{\mu\nu}d\tau^\mu d\tau^\nu$. In such harmonic coordinates, Eqs. (4.2) and (4.3) are solved by

$$\beta^i=b^i_\mu\tau^\mu+c^i,\qquad i=1,\ldots,n, \tag{4.4}$$

$$\phi^\alpha=b^{n+\alpha}_\mu\tau^\mu+c^{n+\alpha},\qquad \alpha=1,\ldots,l. \tag{4.5}$$

We set

$$\varphi^i_\mu:=\frac{\partial}{\partial\tau^\mu}\varphi^i,\quad \mu=0,\ldots,D_0-1. \tag{4.6}$$

With (4.4), the harmonic gauge condition reads

$$A_\mu:={}^{(\mathrm{E})}q\varphi^1_\mu=\sum_i d_i b^i_\mu\overset{!}{=}0,\qquad \mu=0,\ldots,D_0-1. \tag{4.7}$$

With the harmonic gauge constraint (4.7), Eq. (4.1) then reads

$$\sum_{i=2}^{n} \varphi^i_\mu \varphi^i_\nu + \sum_{\alpha\beta=1}^{l} C_{\alpha\beta} b^\alpha_\mu b^\beta_\nu = \sum_{i=1}^{n} d_i b^i_\mu b^i_\nu + \sum_{\alpha\beta=1}^{l} C_{\alpha\beta} b^\alpha_\mu b^\beta_\nu \overset{!}{=} 0,$$

$$\mu,\nu = 0,\dots,D_0-1. \tag{4.8}$$

For convenience, one can set $c^A := 0$, $A=1,\dots,n+l$. Then $\gamma=0$, whence the harmonic coordinates are simultaneously proper coordinates, and the solution reads explicitly,

$$g = \eta_{\mu\nu} d\tau^\mu \otimes d\tau^\nu + \sum_{i=1}^{n} e^{2b^i_\lambda \tau^\lambda} g^{(i)}, \tag{4.9}$$

with linear coefficients b^i_μ, $i=1,\dots,n$, $\mu=0,\dots,D_0-1$, satisfying D_0 linear constraints (4.7) (the harmonic gauge) and ${D_0}^2$ quadratic constraints (4.8) (the harmonic Wheeler–de Witt constraints).

This solution shows a generalized inflationary behavior, which extends the familiar notion of inflation w.r.t. time, as in cosmology, to inflation w.r.t. the internal degrees of freedom on the D_0-dimensional world manifold of an extended object. The constraint (4.7) implies that the total $(D-D_0)$-dimensional volume remains constant (like in a steady state universe[16]) on the world manifold M_0, although here (unlike the stationary case[16]) individual factor spaces may undergo inflationary expansion or contraction in particular directions on M_0. In the standard cosmological case $D_0=1$, this solution agrees with the one described in Ref. 17.

V. ORTHOBRANE SOLUTIONS WITH $^{(E)}V=0$

Now we present a class of solutions with $^{(E)}V=0$, where the field equations read

$$R_{\mu\nu}[g^{(0)}] = \hat{G}_{AB} \partial_\mu \sigma^A \partial_\nu \sigma^B + \sum_{s\in S} \varepsilon_s e^{2L_{As}\sigma^A} \partial_\mu \Phi^s \partial_\nu \Phi^s, \quad \mu,\nu = 1,\dots,D_0, \tag{5.1}$$

$$\hat{G}_{AB} \Delta[g^{(0)}] \sigma^B - \sum_{s\in S} \varepsilon_s L_{As} e^{2L_{Cs}\sigma^C} (\partial \Phi^s)^2 = 0, \quad A=1,\dots,N, \tag{5.2}$$

$$\partial_\mu \left(\sqrt{|g^{(0)}|} g^{(0)\mu\nu} e^{2L_{As}\sigma^A} \partial_\nu \Phi^s \right) = 0, \quad s\in S. \tag{5.3}$$

For the Abelian part of the target space metric we set $(\hat{G}^{AB}) := (\hat{G}_{AB})^{-1}$.

$$\langle X, Y \rangle := X_A \hat{G}^{AB} X_B. \tag{5.4}$$

For $s\in S$ let us now consider vectors

$$L_s = (L_{As}) \in \mathbb{R}^N. \tag{5.5}$$

Definition. A non-empty set S is called an *orthobrane* index set, iff there exists a family of real nonzero coefficients $\{\nu_s\}_{s\in S}$, such that

$$\langle L_s, L_r \rangle = (L^T \hat{G}^{-1} L)_{sr} = -\varepsilon_s (\nu_s)^{-2} \delta_{sr}, \quad s,r \in S. \tag{5.6}$$

For $s\in S$ and $A=1,\dots,N$, we set

$$\alpha^A_s := -\varepsilon_s (\nu_s)^2 \hat{G}^{AB} L_{Bs}. \tag{5.7}$$

Here, (5.6) is just an orthogonality condition for the vectors L_s, $s\in S$. Note that $\langle L_s, L_s\rangle$ has just the opposite sign of ε_s, $s\in S$. With the definition above, we obtain an existence criterion for solutions.

Theorem: Let S be an orthobrane index set with coefficients (5.7). If for any $s\in S$ there is a function $H_s>0$ on M_0 such that

$$\Delta[g^{(0)}]H_s=0, \tag{5.8}$$

i.e., H_s is harmonic on M_0, then, the field configuration

$$R_{\mu\nu}[g^{(0)}]=0, \quad \mu,\nu=1,\ldots,D_0, \tag{5.9}$$

$$\sigma^A=\sum_{s\in S}\alpha_s^A \ln H_s, \quad A=1,\ldots,N, \tag{5.10}$$

$$\Phi^s=\frac{\nu_s}{H_s}, \quad s\in S, \tag{5.11}$$

satisfies the field equations (5.1)–(5.3). □

This theorem follows just from substitution of (5.6)–(5.11) into the equations of motion (5.1)–(5.3). From (3.18), (3.17), and (5.4) we get

$$\langle L_s, L_r\rangle=G^{ij}l_{iI}l_{jJ}+C^{\alpha\beta}\lambda_{\alpha a}\lambda_{\beta b}, \tag{5.12}$$

with $s=(a,I)$ and $r=(b,J)$ in S $(a,b\in\Delta,\ I\in\Omega_a,\ J\in\Omega_b)$. Here, the inverse of the dilatonic midisuperspace metric G_{ij} is given by

$$G^{ij}=\frac{\delta_{ij}}{D_i}+\frac{1}{2-D}, \tag{5.13}$$

whence, for $I,J\in\Omega$, with topological numbers l_{iI} from (3.16), we obtain

$$G^{ij}l_{iI}l_{jJ}=D(I\cap J)+\frac{D(I)D(J)}{2-D}, \tag{5.14}$$

which is again a purely topological number.

We set $\nu_{a,I}:=\nu_{(a,I)}$. Then, due to (5.12) and (5.14), the orthobrane condition (5.6) reads

$$D(I\cap J)+\frac{D(I)D(J)}{2-D}+C^{\alpha\beta}\lambda_{\alpha a}\lambda_{\beta b}=-\varepsilon(I)(\nu_{a,I})^{-2}\delta_{ab}\delta_{I,J}, \tag{5.15}$$

for $a,b\in\Delta$, $I\in\Omega_a$, $I\in\Omega_b$. With $(a,I)=s\in S$, the coefficients (5.7) are

$$\alpha_s^i=-\varepsilon(I)G^{ij}l_{jI}\nu_{a,I}^2=\varepsilon(I)\left(\sum_{j\in I}\delta_j^i+\frac{D(I)}{2-D}\right)\nu_{a,I}^2, \quad i=1,\ldots,n, \tag{5.16}$$

$$\alpha_s^\beta=-\varepsilon(I)C^{\beta\gamma}\lambda_{\gamma a}\nu_{a,I}^2, \quad \beta=1,\ldots,l. \tag{5.17}$$

With $(\sigma^A)=(\phi^i,\varphi^\beta)$, according to (5.10),

$$\beta^i=\sum_{s\in S}\alpha_s^i \ln H_s, \quad i=1,\ldots,n, \tag{5.18}$$

$$\phi^{\beta}=\sum_{s\in S}\alpha_s^{\beta}\ln H_s\,,\quad \beta=1,\ldots,l, \tag{5.19}$$

and the harmonic gauge reads

$$\gamma=\sum_{s\in S}\alpha_s^{0}\ln H_s\,, \tag{5.20}$$

where

$$\alpha_s^{0}:=\varepsilon(I)\frac{D(I)}{2-D}\nu_{a,I}^{2}\,. \tag{5.21}$$

With $H_{a,I}:=H_{(a,I)}$, (5.16), (5.17), and (5.21), the solution of (5.9)–(5.11) reads

$$\begin{aligned} g&=\left(\prod_{s\in S}H_s^{2\alpha_s^{0}}\right)g^{(0)}+\sum_{i=1}^{n}\left(\prod_{s\in S}H_s^{2\alpha_s^{i}}\right)g^{(i)}\\ &=\left(\prod_{(a,I)\in S}H_{a,I}^{\varepsilon(I)2D(I)\nu_{a,I}^{2}}\right)^{1/(2-D)}\left\{g^{(0)}+\sum_{i=1}^{n}\left(\prod_{(a,I)\in S,I\ni i}H_{a,I}^{\varepsilon(I)2\nu_{a,I}^{2}}\right)g^{(i)}\right\},\\ &\text{with}\ \mathrm{Ric}[g^{(0)}]=0,\ \ \mathrm{Ric}[g^{(i)}]=0,\ \ i=1,\ldots n, \end{aligned} \tag{5.22}$$

$$\phi^{\beta}=\sum_{s\in S}\alpha_s^{\beta}\ln H_s=-\sum_{(a,I)\in S}\varepsilon(I)C^{\beta\gamma}\lambda_{\gamma a}\nu_{a,I}^{2}\ln H_{a,I}\,,\quad \beta=1,\ldots,l, \tag{5.23}$$

$$A^{a}=\sum_{I\in\Omega_a}\frac{\nu_{a,I}}{H_{a,I}}\tau_I\,,\quad a\in\Delta, \tag{5.24}$$

where forms τ_I are defined in (2.18), parameters $\nu_s\neq0$ and λ_a satisfy the orthobrane condition (5.15), H_s are positive harmonic functions on M_0, and $\mathrm{Ric}[g^{(i)}]$ denotes the Ricci-tensor of $g^{(i)}$. Finally recall that these solutions are subject to the *orthobrane* constraints

$$D(I\cap J)+\frac{D(I)D(J)}{2-D}+C^{\alpha\beta}\lambda_{\alpha a}\lambda_{\beta b}=-\varepsilon(I)(\nu_{a,I})^{-2}\delta_{ab}\delta_{I,J}\,,\quad 0\neq\nu_{a,I}\in\mathbb{R}, \tag{5.25}$$

for $a,b\in\Delta,\ I\in\Omega_a,\ I\in\Omega_b$. These condition lead to specific intersection rules for the p-branes involved. Some concrete examples of *orthobrane* solutions have been elaborated in Ref. 18.

For positive definite $(C_{\alpha\beta})$ [or $(C^{\alpha\beta})$] and $D_0\geqslant2$, (5.25) implies

$$\varepsilon(I)=-1, \tag{5.26}$$

for all $I\in\Omega_a,\ a\in\Delta$. Then, the restriction $g_{|M_I}$ of the metric (5.22) to a membrane manifold M_I has an odd number of negative eigenvalues, i.e., linearly independent timelike directions. However, if the metric $(C_{\alpha\beta})$ in the space of scalar fields is not positive definite, then (5.26) may be violated for sufficiently negative $C^{\alpha\beta}\lambda_{\alpha a}\lambda_{\beta b}<0$.

VI. TARGET SPACE STRUCTURE

Theorem. The target space $(\mathfrak{M},\mathfrak{g})$ is a homogeneous space.

Proof. The Killing vectors of a transitive subgroup of Isom$(\mathfrak{M})$ can be determined explicitly

$$V_s:=\frac{\partial}{\partial\Phi^{s}},\quad s\in S,$$

$$U_A := \frac{\partial}{\partial x^A} - \sum_{s\in S} U_A^s \Phi^s \frac{\partial}{\partial \Phi^s}, \quad A=1,\ldots,N. \tag{6.1}$$

Moreover, the Lie-algebra of the transitive group of isometries generated by (6.1) reads

$$[U,U]=[V,V]=0$$

$$[U_A, V_s] = L_A^s V_s, \; A=1,\ldots,N, \quad s\in S. \tag{6.2}$$

Theorem: The target space $(\mathfrak{M},\mathfrak{g})$ is locally symmetric if and only if $\langle L^s, L^r\rangle_{\hat{G}}(L^s - L^r) = 0$.

Proof: Let $\mathfrak{R}iem$ denote the Riemann tensor of $(\mathfrak{M},\mathfrak{g})$. The latter is locally symmetric, if and only if

$$\nabla \mathfrak{R}iem = 0, \tag{6.3}$$

where ∇ denotes the covariant derivative w.r.t $\mathfrak{g}$. However, the only nontrivial equations (6.3) are

$$\nabla_p \mathfrak{R}_{srqA} = k_{psrq}\langle L^s, L^r\rangle_{\hat{G}}(L_A^r - L_A^s) = 0, \; A=1,\ldots,N, \quad p,q,r,s\in S \tag{6.4}$$

with $k_{psrq} := \varepsilon_s \varepsilon_r e^{2U^s + 2U^r}(\delta_{ps}\delta_{rq} + \delta_{pr}\delta_{sq})$ nonzero for fixed s,r.

VII. SCALAR PLUS p-BRANES WITH SPHERICAL SYMMETRY

Let us now examine static, spherically symmetric, multidimensional space-times with

$$M = M_{-1}\times M_0 \times M_1 \times \cdots \times M_N, \quad \dim M_i = d_i, i=0,\ldots,N, \tag{7.1}$$

where $M_{-1}\subset \mathbb{R}$ corresponds to a radial coordinate u, $M_0 = S^2$ is a 2-sphere, $M_1 \subset \mathbb{R}$ is time, and $M_i, i>1$ are internal factor spaces. The metric is assumed correspondingly to be

$$ds^2 = \mathrm{e}^{2\alpha(u)} du^2 + \sum_{i=0}^{N} \mathrm{e}^{2\beta_i(u)} ds_i^2 \equiv -\mathrm{e}^{2\gamma(u)} dt^2 + \mathrm{e}^{2\alpha(u)} du^2 + \mathrm{e}^{2\beta_0(u)} d\Omega^2 + \sum_{i=2}^{N} \mathrm{e}^{2\beta_i(u)} ds_i^2, \tag{7.2}$$

where $ds_0^2 \equiv d\Omega^2 = d\theta + \sin^2\theta d\phi^2$ is the line element on S^2, $ds_1^2 \equiv -dt^2$ with $\beta_1 =: \gamma$, and ds_i^2, $i>1$, are u-independent line elements of internal Ricci-flat spaces of arbitrary dimensions d_i and signatures ε_i.

For simplicity here let us only consider a single scalar field denoted as φ.

An electric-type $p+2$-form $F_{\mathrm{e}I}$ has a domain given by a product manifold

$$M_I = M_{i_1} \times \cdots \times M_{i_k}, \tag{7.3}$$

where

$$I = \{i_1, \ldots, i_k\} \subset I_0 \overset{\text{def}}{=} \{0,1,\ldots,N\}. \tag{7.4}$$

The corresponding dimensions are

$$d(I) \overset{\text{def}}{=} \sum_{i\in I} d_i, \quad d(I_0) = D-1. \tag{7.5}$$

A magnetic-type F-form of arbitrary rank k may be defined as a form on a domain $M_{\bar{I}}$ with $\bar{I} \overset{\text{def}}{=} I_0 - I$, dual to an electric-type form,

TABLE I. Different types of antisymmetric $p+2$-form fields.

E	Electric ($1\in I$)	$F_{tuA_3\ldots A_n}$	A_k (coordinate) index of M_I
M	Magnetic ($1\in I$)	$F_{\theta\phi B_3\ldots B_n}$	B_l (coordinate) index of M_I^-
EQ	Electric quasiscalar ($1\notin I$)	$F_{uA_2\ldots A_n}$	A_k (coordinate) index of M_I
MQ	Magnetic quasiscalar ($1\notin I$)	$F_{t\theta\phi B_4\ldots B_n}$	B_l (coordinate) index of M_I^-

$$F_{\mathrm{m}I,\,M_1\cdots M_k}=\mathrm{e}^{-2\lambda\varphi}(*F)_{\mathrm{e}I,\,M_1\cdots M_k}\equiv \mathrm{e}^{-2\lambda\varphi}\frac{\sqrt{g}}{k!}\varepsilon_{M_1\cdots M_kN_1\cdots N_{D-k}}F_{\mathrm{e}I}^{N_1\cdots N_{D-k}}, \tag{7.6}$$

where $*$ is the Hodge operator and ε is the totally antisymmetric Levi–Civita symbol.

For simplicity we now considering a just a single n-form, i.e., a single electric type and a single dual magnetic component, whence

$$\mathrm{rank}F_{\mathrm{m}I}=D-\mathrm{rank}F_{\mathrm{e}I}=d(\overline{I}), \tag{7.7}$$

whence $k=n$ in (7.6) and

$$d(I)=n-1 \text{ for } F_{\mathrm{e}I}, \quad d(I)=d(I_0)-n=D-n-1 \text{ for } F_{\mathrm{m}I}. \tag{7.8}$$

All fields must be compatible with spherical symmetry and staticity. Correspondingly, the vector φ of scalars and the $p+2$-forms valued fields depend (besides on their domain as forms) on the radial variable u only.

Furthermore, the domain of the electric form $F_{\mathrm{e}I}$ does not include the sphere $M_0=S^2$, and $F_{\mathrm{e}I}$ is specified by a u-dependent potential form,

$$F_{\mathrm{e}I,\,uL_2\cdots L_n}=\partial_{[u}U_{L_2\cdots L_n]}, \quad U=U_{L_2,\ldots,L_n}dx^{L_2}\wedge\cdots\wedge dx^{L_n}. \tag{7.9}$$

Since the time manifold M_1 is a factor space of M_I, the form (7.9) describes an electric $(n-2)$-brane in the remaining subspace of M_I. Similarly (7.6) describes a magnetic $(D-n-2)$-brane in M_I.

Let us label all nontrivial components of F by a collective index $s=(I_s,\chi_s)$, where $I=I_s\subset I_0$ characterizes the subspace of M as described above and $\chi_s=\pm 1$ according to the rule

$$\mathrm{e}\mapsto\chi_s=+1, \quad \mathrm{m}\mapsto\chi_s=-1. \tag{7.10}$$

If $1\in I$, the corresponding p-brane evolves with t and we have a true electric or magnetic field, otherwise the potential (7.9) does not depend on $\overline{M}_0$, i.e., it is just a scalar in four dimensions. In this case we call the corresponding electric-type F component (7.9) *electric quasiscalar* and its dual, magnetic-type, F component (7.6) *magnetic quasiscalar*. So there are in general four types of F-field components (summarized in Table I): electric (E), magnetic (M), electric quasiscalar (EQ), magnetic quasiscalar (MQ). The choice of subsets I_s is only constrained by the multidimensional decomposition condition 3.11 for the energy-momentum tensor. Since antisymmetric $p+2$-form field components of type E and M (and type EQ and MQ, respectively) just complement each other, they should be considered as independent of each other. In the following we consider all F_s as independent fields (up to index permutations) each with a single nonzero component.

Let us assume Ricci-flat internal spaces. With spherical symmetry and staticity all field become independent of M_0 and M_0, respectively. And the variation reduces further from $\overline{M}_0$ to the radial manifold M_{-1}.

The reparametrization gauge on the lower dimensional manifold here is chosen as the (generalized) harmonic one.[13] Since M_{-1} is one-dimensional u is a harmonic coordinate, $\Box u=0$, such that

$$\alpha(u)=\sigma_0(u). \tag{7.11}$$

The nonzero Ricci tensor components are

$$e^{2\alpha}R^t_t = -\gamma'',$$

$$e^{2\alpha}R^u_u = -\alpha'' + \alpha'^2 - \gamma'^2 - 2\beta'^2 - \sum_{i=2}^{N} d_i \beta_i'^2,$$

$$e^{2\alpha}R^\theta_\theta = e^{2\alpha}R^\phi_\phi = e^{2\alpha-2\beta} - \beta'',$$

$$e^{2\alpha}R^{b_i}_{a_j} = -\delta^{b_i}_{a_j}\beta_i'' \quad (i,j=1,\dots,N), \tag{7.12}$$

where a prime denotes d/du and the indices a_i, b_i belong to the ith internal factor space. The Einstein tensor component G^1_1 does not contain second-order derivatives:

$$e^{2\alpha}G^1_1 = -e^{2\alpha-2\beta} + \frac{1}{2}\alpha'^2 - \frac{1}{2}\left(\gamma'^2 + 2\beta'^2 + \sum_{i=2}^{N} d_i \beta_i'^2\right). \tag{7.13}$$

The corresponding component of the Einstein equations is an integral of other components, similar to the energy integral in cosmology.

The generalized Maxwell equations give

$$F^{uM_2\dots M_n}_{eI} = Q_{eI} e^{-2\alpha-2\lambda\varphi}, \quad Q_{eI} = \text{const}, \tag{7.14}$$

$$F_{mI,\, uM_1\dots M_{d(\bar I)}} = Q_{mI}\sqrt{|g_{\bar I}|}, \quad Q_{mI} = \text{const}, \tag{7.15}$$

where $|g_{\bar I}|$ is the determinant of the u-independent part of the metric of $M_{\bar I}$ and Q_s are charges. These solutions provide then the energy momentum tensors, of the electric and magnetic $p+2$-forms written in matrix form,

$$e^{2\alpha}(T^N_M[F_{eI}]) = -\tfrac{1}{2}\eta_F \varepsilon(I) Q^2_{eI} e^{2y_{eI}} \text{diag}(+1,[1]_I,[-1]_{\bar I}),$$

$$e^{2\alpha}(T^N_M[F_{mI}]) = \tfrac{1}{2}\eta_F \varepsilon(\bar I) Q^2_{mI} e^{2y_{mI}} \text{diag}(1,[1]_I,[-1]_{\bar I}), \tag{7.16}$$

where the first position belongs to u and f operating over M_J is denoted by $[f]_J$. The functions $y_s(u)$ are

$$y_s(u) = \sigma(I_s) - \chi_s \lambda \varphi. \tag{7.17}$$

The scalar field EMT is

$$e^{2\alpha}T^N_M[\varphi] = \tfrac{1}{2}(\varphi^a)'^2 \text{diag}(+1,[-1]_{I_0}). \tag{7.18}$$

The sets $I_s \in I_0$ may be classified by types E, M, EQ, MQ according to the description in the preceding section. Denoting I_s for the respective types by I_E, I_M, I_{EQ}, I_{MQ}, we see from (7.16) that, positive electric and magnetic energy densities require

$$\eta_f = -\varepsilon(I_E) = \varepsilon(\overline{I_M}) = \varepsilon(I_{EQ}) = -\varepsilon(\overline{I_{MQ}}). \tag{7.19}$$

If t is the only time coordinate, (7.19) with $\eta_F = 1$ holds for any choices of I_s. If there exist other times, then the relations (7.19) constrain the subspaces where the different F components may be specified.

Since the total EMT on the r.h.s. of the Einstein equations has the property

$$T_u^u + T_\theta^\theta = 0, \tag{7.20}$$

the corresponding combination on the lhs becomes an integrable Liouville form

$$G_u^u + G_\theta^\theta = \mathrm{e}^{-2\alpha}[-\alpha'' + \beta_0'' + \mathrm{e}^{2\alpha - 2\beta_0}] = 0,$$

$$\mathrm{e}^{\beta_0 - \alpha} = s(k,u), \tag{7.21}$$

where k is an integration constant (IC) and the function $s(k,.)$ is defined as follows:

$$s(k,u) \overset{\text{def}}{=} \begin{cases} k^{-1}\sinh ku, & k>0 \\ u, & k=0 \\ k^{-1}\sin ku, & k<0. \end{cases} \tag{7.22}$$

Another IC is suppressed by adjusting the origin of the u coordinate.

With (7.21) the D-dimensional line element may be written in the form

$$ds^2 = \frac{\mathrm{e}^{-2\sigma_1}}{s^2(k,u)}\left[\frac{du^2}{s^2(k,u)} + d\Omega^2\right] + \sum_{i=1}^{N} \mathrm{e}^{2\beta_i} ds_i^2 \tag{7.23}$$

where σ_1 has been defined in (2.10).

Let us treat the whole set of unknowns $\beta_i(u), \varphi(u)$ as a real-valued vector function $x(u)$ in an $(N+1)$-dimensional vector space V, with components $x^A = \beta_A$ for $A = 1,\ldots,N$ and $x^{N+1} = \varphi$.

Then the field equations for β_i and φ coincide with the equations of motion corresponding to the Lagrangian of a Euclidean Toda-type system

$$L = \bar{G}_{AB} x'^A x'^B - V_Q(y), \quad V_Q(y) = \sum_s \theta_s Q_s^2 \mathrm{e}^{2y_s}, \tag{7.24}$$

where, according to (7.19), $\theta_s = 1$ if F_s is a true electric or magnetic field and $\theta_s = -1$ if F_s is quasiscalar. The nondegenerate, symmetric matrix

$$(\bar{G}_{AB}) = \begin{pmatrix} G_{ij} & 0 \\ 0 & 1 \end{pmatrix}, \quad G_{ij} = d_i d_j + d_i \delta_{ij} \tag{7.25}$$

defines a positive-definite metric in V. The energy constraint corresponding to (7.24) is

$$E = \sigma_1'^2 + \sum_{i=1}^{N} d_i \beta_i'^2 + \varphi'^2 + V_Q(y) = \bar{G}_{AB} x'^A x'^B + V_Q(y) = 2k^2 \mathrm{sign} k, \tag{7.26}$$

with k from (7.21). The integral (7.26) follows here from the (uu)-component of (3.4).

The functions $y_s(u)$ (7.17) can be represented as scalar products in V (recall that $s = (I_s, \chi_s)$):

$$y_s(u) = Y_{s,A} x^A, \quad (Y_{s,A}) = (d_i \delta_{iI_s}, -\chi_s \lambda), \tag{7.27}$$

where $\delta_{iI} := \Sigma_{j\in I} \delta_{ij}$ is an indicator for i belonging to I (1 if $i \in I$ and 0 otherwise).

The contravariant components of Y_s are found using the matrix $\bar{G}^{AB}$ inverse to $\bar{G}_{AB}$:

$$(\bar{G}^{AB}) = \begin{pmatrix} G^{ij} & 0 \\ 0 & 1 \end{pmatrix}, \quad G^{ij} = \frac{\delta^{ij}}{d_i} - \frac{1}{D-2} \tag{7.28}$$

$$(Y_s{}^A) = \left(\delta_{iI_s} - \frac{d(I_s)}{D-2}, -\chi_s \lambda \right), \tag{7.29}$$

and the scalar products of different Y_s, whose values are of primary importance for the integrability of our system, are

$$Y_{s,A} Y_{s'}{}^A = d(I_s \cap I_{s'}) - \frac{d(I_s) d(I_{s'})}{D-2} + \chi_s \chi_{s'} \lambda^2. \tag{7.30}$$

VIII. PURELY EM BLACK HOLE SOLUTIONS

In Ref. 19 it was shown that quasiscalar components of the F-fields are incompatible with orthobrane black holes. Therefore let us now consider only two F-field components, Type E and Type M according to the classification above. They will be electric as $F_{\rm e}$ and $F_{\rm m}$ and the corresponding sets $I_s \subset I_0$ as $I_{\rm e}$ and $I_{\rm m}$. Then a minimal configuration (7.1) of the manifold M compatible with an arbitrary choice of I_s has the following form:

$$N=5, \ I_0 = \{0,1,2,3,4,5\}, \quad I_e = \{1,2,3\}, \quad I_m = \{1,2,4\}, \tag{8.1}$$

so that

$$(I_0) = D-1, \ d(I_{\rm e}) = n-1, \ d(I_{\rm m}) = D-n-1, \ d(I_{\rm e} \cap I_{\rm m}) = 1 + d_2;$$

$$d_1 = 1, \ d_2 + d_3 = d_3 + d_5 = n-2. \tag{8.2}$$

The relations (8.2) show that, given D and d_2, all d_i are known.

This corresponds to an electric $(n-2)$-brane located on the subspace $M_2 \times M_3$ and a magnetic $(D-n-2)$-brane on the subspace $M_2 \times M_4$. Their intersection dimension $d_{\rm int} = d_2$ turns out to determine qualitative properties of the solutions.

The index s now takes the two values e and m and

$$Y_{{\rm e},A} = (1, d_2, d_3, 0, 0, -\lambda),$$

$$Y_{{\rm m},A} = (1, d_2, 0, d_4, 0, \lambda),$$

$$Y_{\rm e}^A = (1,1,1,0,0,-\lambda) - \frac{n-1}{D-2}(1,1,1,1,1,0),$$

$$Y_{\rm m}^A = (1,1,0,1,0,\lambda) - \frac{D-n-1}{D-2}(1,1,1,1,1,0), \tag{8.3}$$

where the last component of each vector refers to $x^{N+1} = x^6 = \varphi$.

In the solutions presented below the set of ICs will be reduced by the condition that the space-time be asymptotically flat at spatial infinity $(u=0)$ and by a choice of scales in the relevant directions. Namely, we put

$$\beta_i(0) = 0 = \varphi(0), \quad i = 1,2,3,4,5. \tag{8.4}$$

The requirement $\varphi(0)=0$ is convenient and may be always satisfied by a redefinition of the charges. The conditions $\beta_i(0)=0$ $(i>1)$ mean that the real scales of the extra dimensions are hidden in the internal metrics ds_i^2 independent of whether or not they are assumed to be compact.

In the following, both cases, orthobrane solutions and solutions with degenerate charges, are considered first generally and then for the minimal configuration (8.1)–(8.4).

A. Orthobrane black hole solutions

Assuming that the vectors Y_s are mutually orthogonal with respect to the metric $\overline{G_{AB}}$, i.e.,

$$Y_{s,A}Y_{s'}{}^{A}=\delta_{ss'}N_s^2, \tag{8.5}$$

the number of functions y_s does not exceed the number of equations, and the system becomes integrable. Due to (7.8), the norms N_s are actually s-independent:

$$N_s^2=d(I_s)\left[1-\frac{d(I_s)}{D-2}\right]+\lambda^2=\frac{(n-1)(D-n-1)}{D-2}+\lambda^2\stackrel{\text{def}}{=}\frac{1}{\nu}, \tag{8.6}$$

$\nu>0$.

Due to (8.5), the functions $y_s(u)$ obey the decoupled equations

$$y_s''=\theta_s\frac{Q_s^2}{\nu}e^{2y_s}, \tag{8.7}$$

whence

$$e^{-y_s(u)}=\begin{cases}(|Q_s|/\sqrt{\nu})s(h_s,u+u_s), & \theta=+1,\\ [|Q_s|/(\sqrt{\nu}h_s)]\cosh[h_s(u+u_s)], & h_s>0,\quad \theta=-1.\end{cases} \tag{8.8}$$

where h_s and u_s are ICs and the function s was defined in (7.22). For the functions $x^A(u)$ we obtain:

$$x^A(u)=\nu\sum_s Y_s{}^{A}y_s(u)+c^Au+\overline{c^A}, \tag{8.9}$$

where the vectors of ICs c^A and $\overline{c^A}$ satisfy the orthogonality relations $c^AY_{s,A}=\overline{c^A}Y_{s,A}=0$, or

$$c^id_i\delta_{iI_s}-\lambda c^{N+1}\chi_s=0,\quad \overline{c^i}d_i\delta_{iI_s}-\lambda\overline{c^{N+1}}\chi_s=0. \tag{8.10}$$

Specifically, the logarithms of the scale factors β_i and the scalar field φ are

$$\beta_i(u)=\nu\sum_s\left[\delta_{iI_s}-\frac{d(I_s)}{D-2}\right]y_s(u)+c^iu+\overline{c^i}, \tag{8.11}$$

$$\varphi(u)=-\lambda\nu\sum_s y_s(u)+c^{N+1}u+\overline{c^{N+1}}, \tag{8.12}$$

and the function σ_1 which appears in the metric (7.23) is

$$\sigma_1=-\frac{\nu}{D-2}\sum_s d(I_s)\,y_s(u)+c^0u+\overline{c^0} \tag{8.13}$$

with

$$c^0=\sum_{i=1}^N d_ic^i,\quad \overline{c^0}=\sum_{i=1}^N d_i\overline{c^i}. \tag{8.14}$$

Finally, (7.26) now reads

TABLE II. Orthobrane solutions with $\lambda=0$.

	n	$d(I_e)$	$d(I_m)$	d_2	B	C
$D=4m+2$ ($m\in\mathbb{N}$)	2m+1	2m	2m	m−1	1/m	1/m
D= 11	4	3	6	1	2/3	1/3
	7	6	3	1	1/3	2/3

$$E=\nu\sum_s h_s^2\mathrm{sign}h_s+\overline{G}_{AB}c^Ac^B=2k^2\mathrm{sign}k. \tag{8.15}$$

The relations (7.11), (7.14), (7.15), (7.21), (7.23), (8.8)–(8.15), along with the definitions (7.22) and (8.6) and the restriction (8.5), entirely determine the general solution.

For the minimal configuration (8.1)–(8.4), the orthogonality condition (8.5) reads

$$\lambda^2=d_2+1-\frac{1}{D-2}(n-1)(D-n-1). \tag{8.16}$$

In particular, in dilaton gravity $n=2,d_2=0$ and the integrability condition (8.16) just reads $\lambda^2=1/(D-2)$, which is a well-known relation of string gravity. The familiar Reissner-Nordström solution, $D=4,n=2$, $\lambda=0$, $d_2=0$ does *not* satisfy Eq. (8.16). (It will be recovered indeed as a degenerate case below.) Some examples of configurations satisfying the orthogonality condition (8.16) in the purely topological case $\lambda=0$ are summarized in Table II [including the values of the constants B and C from (8.28)]. In this case (8.16) is just a Diophantus equation for D, n and d_2.

The solution is entirely determined by inserting (8.3) into (8.9) with $\overline{c^A}=0$ due to (8.4),

$$x^A(u)=\nu\sum_s Y_s{}^A y_s(u)+c^Au;\quad \mathrm{e}^{-y_s(u)}=(|Q_s|/\sqrt{\nu})s(h_s,u+u_s). \tag{8.17}$$

Due to (8.16) the parameter ν is

$$\nu=1/\sqrt{1+d_2}. \tag{8.18}$$

The constants are connected by the relations

$$(|Q_{\mathrm{e,m}}|/\nu)\,s(h_{\mathrm{e,m}},u_{\mathrm{e,m}})=1;$$

$$c^1+d_2c^2+d_3c^3-\lambda c^6=0;\quad c^1+d_2c^2+d_4c^4+\lambda c^6=0;$$

$$\frac{h_{\mathrm{e}}^2\mathrm{sign}h_{\mathrm{e}}+h_{\mathrm{m}}^2\mathrm{sign}h_{\mathrm{m}}}{1+d_2}+G_{ij}c^ic^j+(c^6)^2=2k^2\mathrm{sign}k, \tag{8.19}$$

where the matrix G_{ij} is given in (7.25) and all $\overline{c^A}=0$ due to the boundary conditions (8.4). The fields φ and F are given by Eqs. (7.14), (7.15), (8.12).

This solution contains 8 nontrivial, independent ICs, namely, $Q_{\mathrm{e}},Q_{\mathrm{m}},h_{\mathrm{e}},h_{\mathrm{m}}$ and 4 others from the set c^A constrained by (8.19).

For black holes, we require that all $|\beta_i|<\infty$, $i=2,\dots,N$ (regularity of extra dimensions), $|\varphi|<\infty$ (regularity of the scalar field) and $|\beta_0|<\infty$ (finiteness of the spherical radius) as $u\to\infty$. With $y_s(u)\sim -h_su$, this leads to the following constraints on the ICs:

$$c^A=-k\sum_s(\delta_{1I_s}+\nu Y_s{}^A h_s), \tag{8.20}$$

where $A=1$ corresponds to $i=1$. Via orthonormality relations (8.10) for c^A, we obtain

$$h_s = k\delta_{1I_s}, \tag{8.21}$$

$$c^A = -k\delta_1^A + k\nu \sum_s \delta_{1I_s} Y_s{}^A, \tag{8.22}$$

and (8.15) then holds automatically.

Let us now consider the case where (8.21) and (8.22) with $\delta_{1I_s}=1$ hold. After a transformation $u \mapsto R$, to isotropic coordinates given by the relation

$$\mathrm{e}^{-2ku} = 1 - 2k/R, \tag{8.23}$$

we obtain

$$ds^2 = -\frac{1-2k/R}{P_\mathrm{e}^B P_\mathrm{m}^C} dt^2 + P_\mathrm{e}^C P_m^B \left(\frac{dR^2}{1-2k/R} + R^2 d\Omega^2 \right) + \sum_{i=2}^{5} \mathrm{e}^{2\beta_i(u)} ds_i^2, \tag{8.24}$$

$$\mathrm{e}^{2\beta_2} = P_\mathrm{e}{}^{-B} P_\mathrm{m}{}^{-C}, \quad \mathrm{e}^{2\beta_3} = (P_\mathrm{m}/P_\mathrm{e})^B,$$

$$\mathrm{e}^{2\beta_4} = (P_\mathrm{e}/P_\mathrm{m})^C, \quad \mathrm{e}^{2\beta_5} = P_\mathrm{e}{}^C P_\mathrm{m}{}^B, \tag{8.25}$$

$$e^{2\lambda\varphi} = (P_\mathrm{e}/P_\mathrm{m})^{2\lambda^2/(1+d_2)}, \tag{8.26}$$

$$F_{01M_3\ldots M_n} = -Q_\mathrm{e}/(R^2 P_\mathrm{e}), \quad F_{23M_3\ldots M_n} = Q_\mathrm{m} \sin\theta, \tag{8.27}$$

with the notations

$$P_\mathrm{e,m} = 1 + p_\mathrm{e,m}/R, \quad p_\mathrm{e,m} = \sqrt{k^2 + (1+d_2) Q_\mathrm{e,m}^2} - k;$$

$$B = \frac{2(D-n-1)}{(D-2)(1+d_2)}, \quad C = \frac{2(n-1)}{(D-2)(1+d_2)}. \tag{8.28}$$

The BH gravitational mass as determined from a comparison of (8.24) with the Schwarzschild metric for $R \to \infty$ is

$$G_N M = k + \tfrac{1}{2}(B p_\mathrm{e} + C p_\mathrm{m}), \tag{8.29}$$

where G_N is the Newtonian gravitational constant. This expression, due to $k>0$, provides a restriction upon the charge combination for a given mass, namely,

$$B|Q_\mathrm{e}| + C|Q_\mathrm{m}| < 2G_N M/\sqrt{1+d_2}. \tag{8.30}$$

The inequality is replaced by equality in the extreme limit $k=0$. For $k=0$ our BH turns into a naked singularity (at the center $R=0$) for any $d_2>0$, while for $d_2=0$ the zero value of R is not a center ($g_{22}\neq 0$) but a horizon. In the latter case, if $|Q_\mathrm{e}|$ and $|Q_\mathrm{m}|$ are different, the remaining extra-dimensional scale factors are smooth functions for all $R \geqslant 0$.

For a static, spherical BH one can define a Hawking temperature $T_H := \kappa/2\pi$ as given by the surface gravity κ. With a generalized Komar integral (see e.g., Ref. 20)

$$M(r) := -\frac{1}{8\pi} \int_{S_r} * d\xi \tag{8.31}$$

over the timelike Killing form ξ, the surface gravity can be evaluated as

$$\kappa = M(r_H)/(r_H)^2 = (\sqrt{|g_{00}|})'/\sqrt{g_{11}}|_{r=r_H} = e^{\gamma-\alpha}|\gamma'||_{r=r_H}, \tag{8.32}$$

where a prime, α, and γ are understood in the sense of the general metric (7.2) and k_B is the Boltzmann constant. The expression (8.32) is invariant with respect to radial coordinate reparametrization, as is necessary for any quantity having a direct physical meaning. It is also invariant under conformal mappings with a conformal factor which is smooth at the horizon.

Substituting g_{00} and g_{11} from (8.24), one obtains

$$T_H = \frac{1}{2\pi k_B}\frac{1}{4k}\left[\frac{4k^2}{(2k+p_e)(2k+p_m)}\right]^{1/(d_2+1)}. \tag{8.33}$$

If $d_2=0$ and both charges are nonzero, this temperature tends to zero in the extreme limit $k \to 0$; if $d_2=1$ and both charges are nonzero, it tends to a finite limit, and in all other cases it tends to infinity. Remarkably, it is determined by the p-brane intersection dimension d_2 rather than the whole space-time dimension D.

B. The solution for $Q_e^2 = Q_m^2$

In this degenerate case, solutions can be found which need not satisfy the orthobrane condition (8.5). Let us suppose that two functions (7.17), say, y_1 and y_2, coincide up to an addition of a constant (which may be then absorbed by re-defining a charge Q_1 or Q_2) while corresponding vectors Y_1 and Y_2 are neither coinciding, nor orthogonal (otherwise we would have the previously considered situation). Substituting $y_1 \equiv y_2$ into (7.27), one obtains

$$(Y_{1,A} - Y_{2,A})x^A = 0. \tag{8.34}$$

This is a constraint reducing the number of independent unknowns x^A. Furthermore, substituting (8.34) to the Lagrange equations for x^A,

$$-(Y_{1,A} - Y_{2,A})x''^A = \sum_s \theta_s Q_s^2 e^{2y_s} Y_s^A (Y_{1,A} - Y_{2,A}) = 0. \tag{8.35}$$

In this sum all coefficients of different functions e^{2y_s} must be zero. This yields new orthogonality conditions

$$Y_s^A(Y_{1,A} - Y_{2,A}) = 0, \quad s \neq 1,2, \tag{8.36}$$

now for the difference $Y_1 - Y_2$ and other Y_s, and with Eq. (8.6) the relation

$$(\nu^{-1} - Y_1^A Y_{2,A})(\theta_1 Q_1^2 - \theta_2 Q_2^2) = 0. \tag{8.37}$$

The first multiplier in (8.37) is positive ($\overline{G}_{AB}$ is positive-definite, hence a scalar product of two different vectors with equal norms is smaller than their norm squared). Therefore

$$\theta_1 = \theta_2, \quad Q_1^2 = Q_2^2. \tag{8.38}$$

Imposing the constraints (8.34), (8.36), (8.38), reduces the numbers of unknowns and integration constants, and simultaneously also reduces the number of restrictions on the input parameters [by the orthogonality conditions (8.5)]. Due to (8.38), this is only possible when the two components with coinciding charges are of equal nature: both must be either true electric/magnetic ones ($\theta_s = 1$), or quasiscalar ones ($\theta_s = -1$). Correspondingly, we now set $y(u) := y_e = y_m$ and $Q^2 := Q_e^2 = Q_m^2$.

For the minimal configuration (8.1)–(8.3), Eq. (8.34) yields

$$d_3\beta_3 - d_4\beta_4 - 2\lambda\varphi = 0. \tag{8.39}$$

Equations (8.36) are irrelevant here since we are dealing with two functions y_s only. The equations of motion for x^A now take the form

$$x^{A''}=Q^2 e^{2y}(Y^A_e+Y^A_m). \tag{8.40}$$

Their proper combination gives $y''=(1+d_2)Q^2e^{2y}$, whence

$$e^{-y}=\sqrt{(1+d_2)Q^2}s(h,u+u_1), \tag{8.41}$$

where the function s is defined in (7.22) and h,u_1 are ICs and, due to (8.4),$\sqrt{(1+d_2)Q^2}s(h,u_1)=1$. Other unknowns are easily determined using (8.40) and (8.4):

$$x^A=\nu Y^A y+c^A;\quad Y^A=Y^A_e+Y^A_m=(1,1,0,0,-1,0); \tag{8.42}$$

$$\sigma_1=-\nu y+c_0 u.$$

Here, as in (8.18), $\nu=1/(1+d_2)$, but it is now just a notation. The constants $c_0,h,c^A(A=1,\ldots,6)$ and k [see (7.21)] are related by

$$-c^0+\sum_{i=1}^{5} d_i c^i=0,\quad c^1+d_2c^2+d_3c^3-\lambda c^6=0,\quad c^1+d_2c^2+d_4c^4+\lambda c^6=0,$$

$$2k^2\text{sign}k=\frac{2h^2\text{sign}h}{1+d_2}(c^0)^2+\sum_{i=1}^{5} d_i(c^i)^2+(c^6)^2. \tag{8.43}$$

Extra-dimensional scale factors remain finite as $u\to u_{\max}$ in the case of a BH. It is specified by the following values of the ICs:

$$k=h>0,\quad c^3=c^4=c^6=0,\quad c_2=-c_5=-\frac{k}{1+d_2},\quad c_0=c^1=-\frac{d_2k}{1+d_2}. \tag{8.44}$$

The event horizon occurs at $u=\infty$. After the same transformation (8.23) the metric takes the form

$$ds^2_D=-\frac{1-2k/R}{(1+p/R)^{2\nu}}dt^2+(1+p/R)^{2\nu}\left(\frac{dR^2}{1-2k/R}+R^2d\Omega^2\right)$$
$$+(1+p/R)^{-2\nu}ds_2^2+ds_3^2+ds_4^2+(1+p/R)^{2\nu}ds_5^2 \tag{8.45}$$

with the notation

$$p=\sqrt{k^2+(1+d_2)Q^2}-k. \tag{8.46}$$

The fields φ and F are determined by the relations

$$\varphi\equiv 0,\quad F_{01L_3\cdots L_n}=-\frac{Q}{R^2(1+p/R)},\quad F_{23L_3\cdots L_n}=Q\sin\theta. \tag{8.47}$$

$$G_N M=k+p/(1+d_2), \tag{8.48}$$

The Hawking temperature can be calculated as before,

$$T=\frac{1}{2\pi k_B}\frac{1}{4k}\left(\frac{2k}{2k+p}\right)^{2/(d_2+1)}. \tag{8.49}$$

The well-known results for the Reissner–Nordström metric are recovered when $d_2=0$. In this case $T\to 0$ in the extreme limit $k\to 0$. For $d_2=1$, T tends to a finite limit as $k\to 0$ and for $d_2>1$ it tends to infinity. As is the case with two different charges, T does not depend on the space-time dimension D, but depends on the p-brane intersection dimension d_2.

IX. THE EINSTEIN FRAME FOR DYNAMICS AND COSMOLOGY

In this final section we discuss the issue of the physical frame for the particularly important case $D_0=4$. First of all, in this case a selfdual canonical formulation of dynamics is at hand, due to the particular spinor decomposition $\mathrm{so}(1,3)=\mathrm{su}(2)\oplus\mathrm{su}(2)$ of the tangent Lorentz symmetry. In the Einstein frame, the effective σ-model with $D_0=4$ admits in principle a canonical quantization of the geometry on $\overline{M}_0=\mathbb{R}\times M_0$ to the same extend and under the same assumptions as pure Einstein gravity does.

Since for a multidimensional geometry as defined above the imprint of the internal factor spaces is only by their scale factors, configuration space and phase space of such geometries will only be extended by finite a finite number of dilatonic midisuperspace fields. However, only in the Einstein frame the coupling of the dilatonic fields to the $\overline{D}_0$-geometry will be minimal such that the quantization of the latter can be executed practically independently.

Let us denote the external space-time metric $\overline{g^{(0)}}$ in the Brans–Dicke frame with $\overline{\gamma}\overset{!}{=}0$ as $\overline{g^{(\mathrm{BD})}}$ and in the Einstein frame with $f\overset{!}{=}0$ as $\overline{g^{(\mathrm{E})}}$. It can be easily seen that they are connected with each other by a conformal transformation

$$\overline{g^{(\mathrm{E})}}\mapsto\overline{g^{(\mathrm{BD})}}=\Omega^2\overline{g^{(\mathrm{E})}} \tag{9.1}$$

with Ω from (2.39).

In particular, also for spatially homogeneous cosmological models (and with $t\leftrightarrow u$ for spherically symmetric static models) solutions have to be transformed to the Einstein frame before a physical interpretation can be given.

Under any projection $\mathrm{pr}_0:\overline{M_0}\to\mathbb{R}$ a consistent pullback of the metric $-e^{2\gamma(\tau)}d\tau\otimes d\tau$ from $\tau\in\mathbb{R}$ to $x\in\mathrm{pr}_0^{-1}\{\tau\}\subset\overline{M}_0$ is given by

$$\overline{g^{(\mathrm{BD})}}(x):=-e^{2\gamma(\tau)}d\tau\otimes d\tau+e^{2\beta^0(x)}g^{(0)}. \tag{9.2}$$

In particular, for spatially (metrically-)homogeneous cosmological models all scale factors $a_i:=e^{\beta^i}$, $i=0,\dots,n$, depend only on $\tau\in\mathbb{R}$.

With (9.2) and (9.1), Eq. (2.6) reads

$$\begin{aligned} g&=-e^{2\gamma(\tau)}d\tau\otimes d\tau+a_0^2g^{(0)}+\sum_{i=1}^{n}e^{2\beta^i}g^{(i)}\\ &=-dt_{\mathrm{BD}}\otimes dt_{\mathrm{BD}}+a_{\mathrm{BD}}^2g^{(0)}+\sum_{i=1}^{n}e^{2\beta^i}g^{(i)}\\ &=-\Omega^2dt_{\mathrm{E}}\otimes dt_{\mathrm{E}}+\Omega^2a_{\mathrm{E}}^2g^{(0)}+\sum_{i=1}^{n}e^{2\beta^i}g^{(i)}, \end{aligned} \tag{9.3}$$

where $a_0:=a_{\mathrm{BD}}$ and a_{E} are the external space scale factor functions depending respectively on the cosmic synchronous time t_{BD} and t_{E} in the Brans–Dicke and Einstein frame. With (2.39) the latter is related to the former by

$$a_{\mathrm{E}}=\Omega^{-1}a_{\mathrm{BD}}=\left(\prod_{i=1}^{n} e^{d_i\beta^i}\right)^{\frac{1}{D_0-2}} a_{\mathrm{BD}}, \tag{9.4}$$

and the cosmic time of the Einstein frame is given by

$$\pm dt_{\mathrm{E}}=\Omega^{-1}e^{\gamma}d\tau=\left(\prod_{i=1}^{n} e^{d_i\beta^i}\right)^{\frac{1}{D_0-2}} dt_{\mathrm{BD}}. \tag{9.5}$$

Since $a_{\mathrm{BD}}^2(d\eta_{\mathrm{BD}})^2=\Omega^2 a_{\mathrm{E}}^2(d\eta_{\mathrm{E}})^2$, the conformal times of the Einstein and the Brans–Dicke frame agree (up to time reversal). This has sometimes guided authors to compare the frames in conformal time (see e.g., Ref. 21). However (at least for cosmology) the physical relevant time is the cosmic synchronous time, which is different for different frames, in particular for the Einstein and Brans–Dicke frame.

In Ref. 22 several reasons have been listed why minimal coupling between geometry and matter and hence the Einstein frame is the preferred choice. There also a general prescription for the transformation of well known solutions from the Brans–Dicke frame to the Einstein frame has been given. It was demonstrated explicitly that qualitative cosmological features change significantly under this transformation. This was shown for a couple of examples, including the general multidimensional Kasner solution and a special inflationary solution with constant internal volume. In particular it was shown that inflationary solutions in the Brans–Dicke frame transform into noninflationary ones in the Einstein frame. It is to be expected that this is a rather general feature, whence the multitude of solutions which appear inflationary in the Brans–Dicke frame will be indeed non-inflationary when considered in the Einstein frame.

X. DISCUSSION

The Einstein action with minimally coupled scalars and p-branes in higher dimension D can be reduced to an effective model in lower dimension D_0. This results in a (generalized) σ-model with conformally flat target space. With a purely geometrical dilaton field f, it provides a natural generalization for the well-known Brans–Dicke theory.

The orthobrane condition (5.6) allows us to find exact solutions. Furthermore, the orthobrane solution is a sufficient condition for the target space of the σ-model to be a locally symmetric space. The orthobrane case is the generic one [apart from cases with degenerate coupling matrix (3.17)] where the target space is locally symmetric.

Examples of a certain minimal static, spherically symmetric p-brane configuration are given with just one electric and one magnetic antisymmetric F component (since in four dimensions we only deal with a single electromagnetic field), which in general intersect and interact with a single scalar field. Spherical symmetry here is considered in the physical relevant $D_0=4$ case of S^2 spheres, although the extension to arbitrary spheres is straightforward.

Besides popular families of orthobrane solutions there are further families of solutions, which have another additional symmetry, e.g., coinciding F-field charges for the electromagnetic solutions. In the target space this additional symmetry is expressed by a linear dependency (8.34) between column vectors Y of the coupling matrix L defined in (3.17).

For the mentioned static solutions, Hawking temperature T_H can be formally calculated by surface gravity via a Komar-type integral. For both, the orthobrane case and the case of equal charges $Q_{\mathrm{e}}^2=Q_{\mathrm{m}}^2$, the expressions of T_H depend characteristically on the intersection dimension. This results are also interesting in the context of recent increased interest in extremal p-brane configurations with black holes.[23]

The interpretation of the extremal limit $k\to 0$ is delicate. The solutions above have been described in isotropic coordinates which cover just the asymptotically flat exterior of the black hole. A better understanding of their global causality structure would require an investigation of the maximal extension of the space-time rather than only of its exterior part. The limit $k\to 0$ was here called extremal, since via (8.29) and (8.48) in this limit the effective asymptotical Schwarzschild mass M is just given by the charges, $G_N M = \frac{1}{2}(Bp_{\rm e} + Cp_{\rm m})$ and $G_N M = p/(1+d_2)$, respectively. Further work is required to understand this type of extremality, and the related asymptotics of T_H, which remains finite for intersection dimension $d_2 = 1$ and becomes infinite for $d_2 \geqslant 2$. As it was pointed out recently in Ref. 24 particular care is needed in order to associate the correct physical charges and thermal properties of a black hole correctly with its horizon.

The multidimensional σ-model opens the door for further investigations, in particular also for covariant and canonical quantization. The effective σ-model reduction appears as a possible clue to canonical quantization within a large well defined class of higher dimensional geometries, namely the multidimensional ones. Covariant quantization techniques can be applied in any dimension D_0. In particular, they are well applicable to our new solutions with scalar fields only, when the target space is flat. The effective geometry of the D_0-dimensional model can be reformulated in terms of connections and soldering forms. For $D_0 = 4$, not only a canonical $1+3$ split can be performed, but moreover the canonical quantization of the D_0-geometry can be performed with self-dual variables in the usual manner.

Finally, in analogy to investigations in Ref. 4, it should be possible to apply solution generating techniques like the Ehlers–Harrison transformation also in the context of the multidimensional σ-model.

ACKNOWLEDGMENTS

I would like to thank to A. Ashtekar, O. Brodbeck, K. Bronnikov, S. Das, L. Freidel, S. Gupta, V. Ivashchuk, J. Pullin, R. Puzio, and L. Smolin for discussions on the subject, and to N. Guerras for permanent encouragement. This work was supported by NSF Grant No. PHY-9514240 to The Pennsylvania State University and a gift from the Jesse Phillips Foundation.

[1] P. Breitenlohner, D. Maison, and G. Gibbons, Commun. Math. Phys. **120**, 295 (1988).
[2] R. Geroch, J. Math. Phys. **12**, 918 (1971); J. Math. Phys. **13**, 394 (1972).
[3] D. V. Gal'tsov, Phys. Rev. Lett. **74**, 2863 (1995).
[4] D. V. Gal'tsov and O. A. Rytchkov, Generating branes via sigma-models, hep-th/9801160.
[5] D. Korotkin and H. Samtleben, Yangian symmetry in integrable quantum gravity, hep-th/9710210.
[6] A. Strominger and C. Vafa, Phys. Lett. B **379**, 99 (1996).
[7] M. J. Duff, Int. J. Mod. Phys. A **11**, 5623 (1996).
[8] J. Scherk and J. H. Schwarz, Phys. Lett. B **57**, 463 (1975).
[9] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian Geometry, Pure and Applied Mathematics 202*, 2nd ed. (Marcel Dekker, New York, 1996).
[10] M. Rainer and A. Zhuk, Phys. Rev. D **54**, 6186 (1996).
[11] G. W. Gibbons and S. W. Hawking, Phys. Rev. D **15**, 2752 (1977).
[12] J. W. York, Phys. Rev. Lett. **28**, 1082 (1972); Found. Phys. **16**, 249 (1986).
[13] M. Rainer, Hadronic J. **21**, 351 (1998); Proceedings of the International Workshop on Modern Modified Theories of Gravitation and Cosmology, Beer Sheva, June 29–30, 1997.
[14] M. Rainer, Int. J. Mod. Phys. D **4**, 397 (1995).
[15] M. Rainer, Gravitation and Cosmology **1**, 121 (1995).
[16] H. Bondi and T. Gold, Mon. Not. R. Astron. Soc. **108**, 252 (1948).
[17] U. Bleyer and A. Zhuk, Astron. Nachr. **317**, 161 (1996).
[18] V. D. Ivashchuk, V. N. Melnikov, and M. Rainer gr-gc/9708058, Grav. and Cosmol. **4**, 73–82 (1998).
[19] K. A. Bronnikov, U. Kasper, and M. Rainer, Intersecting electric and magnetic *p*-branes: spherically symmetric solutions, preprint P-Math-97/21, gr-qc/9708058 (to appear in Gen. Rel. Grav.).
[20] M. Heusler, *Black Hole Uniqueness Theorems* (Cambridge Lecture Notes in Physics 6, Cambridge University Press, 1996).
[21] M. Gasperini and G. Veneziano, Mod. Phys. Lett. A **8**, 3701 (1993); Phys. Rev. D **50**, 2519 (1994).
[22] M. Rainer and A. Zhuk, Einstein and Brans–Dicke frames in multidimensional cosmology, gr-qc/9808073.
[23] Y. Satoh, BTZ black holes and their near-horizon geometry of higher-dimensional black holes, hep-th/9810135.
[24] A. Ashtekar, J. Baez, A. Corichi, and K. Krasnov, Phys. Rev. Lett. **80**, 904 (1998).

Frames from imprimitivity systems

Paolo Aniello[a] and Gianni Cassinelli[b]
Dipartmento di Fisica, Universita di Genova, INFN, Sezione di Genova, via Dodecaneso 33, 16146 Genova, Italy

Ernesto De Vito[c]
Dipartimento di Matematica, Universita di Modena, via Campi 213/B, 41100 Modena, Italy and INFN, Sezione di Genova, via Dodecaneso 33, 16146 Genova, Italy

Alberto Levrero[d]
Dipartimento di Fisica, Universita di Genova, INFN, Sezione di Genova, via Dodecaneso 33, 16146 Genova, Italy

(Received 18 March 1999; accepted for publication 26 April 1999)

Let (P,V) be an irreducible imprimitivity system for a group H based on a dual group $\hat{A}$ of an Abelian group A and acting on a Hilbert space $\mathcal{H}$. Given $\psi \in \mathcal{H}$, we find necessary and sufficient conditions in order that the set of vectors $\{\int_{\hat{A}} \overline{\langle \hat{a},a\rangle} V_h^{-1}\, dP(\hat{a})\psi : a\in A,\, h\in H\}$ be a frame in $\mathcal{H}$. Moreover, we apply these results to some examples that are considered in the literature in the context of square-integrability modulo a coset space. © *1999 American Institute of Physics.* [S0022-2488(99)00610-6]

I. INTRODUCTION

It is well known that a large class of continuous frames is defined in terms of square-integrable representations U of locally compact groups G by means of the equation

$$G \ni g \mapsto U_g \psi \in \mathcal{H}, \tag{1}$$

where $\mathcal{H}$ is the Hilbert space carrying the representation U and ψ is any nonzero vector belonging to the domain of $K_U^{-1/2}$, where K_U is the formal degree of U. The space labelling the frame is the group G endowed with a left invariant Haar measure μ_G^l. As a consequence of the properties of square-integrable representations,[1] the above frames are tight and the frame bound can be computed using the orthogonality relation

$$\int_G \langle \phi, U_g \psi\rangle \overline{\langle \phi', U_g \psi'\rangle}\, d\mu_G^l(g) = \langle \phi,\phi'\rangle \langle K_U^{-1/2}\psi', K_U^{-1/2}\psi\rangle. \tag{2}$$

The groups that are semidirect products with an Abelian normal factor, i.e., $G=A\times' H$, are of particular interest in the applications. For these groups, there is a one-to-one correspondence between the representations U of G and the imprimitivity systems for H based on the dual group $\hat{A}$ of A, explicitly given by

$$U_a = \int_{\hat{A}} \langle \hat{a},a\rangle\, dP(\hat{a}),$$

$$V_h = U_h,$$

[a]Electronic mail: aniello@ge.infn.it
[b]Electronic mail: cassinelli@ge.infn.it
[c]Electronic mail: devito@unimo.it
[d]Electronic mail: levrero@ge.infn.it

where $a\in A$ and $h\in H$. Moreover, a right invariant Haar measure μ_G^r for G is simply given by $\mu_G^r=\mu_A^r\otimes\mu_H^r$, where μ_A^r and μ_H^r are right invariant Haar measures on A and H, respectively; hence it is convenient to perform in Eq. (1) the change of variable $g\mapsto g^{-1}$ in order to use the measure μ_G^r instead of μ_G^l. Taking into account these facts Eq. (1) becomes

$$A\times H\ni(a,h)\mapsto\left(\int_{\hat{A}}\overline{\langle\hat{a},a\rangle}V_h^{-1}\,dP(\hat{a})\psi\right)\in\mathcal{H},\tag{3}$$

and the corresponding frame is labeled by the points of the topological direct product $A\times H$ with the product measure $\mu_A^r\otimes\mu_H^r$.

The structure of the above relation suggests the possibility of defining frames by means of Eq. (3) also in the case that H acts on $\hat{A}$ without preserving the composition law of $\hat{A}$.

The need to consider the above classes of frames is due to the fact that there are groups G of interest in the applications, as the Euclidean group, the Galilei and Poincaré groups, that are semidirect products, but do not admit square integrable representations.

In this paper, we consider two locally compact second countable (lcsc) groups A, H such that A is Abelian and H acts on $\hat{A}$, the dual group of A, and an irreducible imprimitivity system (P,V) for H based on $\hat{A}$ and acting on a Hilbert space $\mathcal{H}$. We assume that the orbits of H in $\hat{A}$ are locally closed, so that, due to the theorem of Mackey about imprimitivity systems, see, for example, Ref. 2, there is a one-to-one correspondence between the equivalence classes of irreducible imprimitivity systems (P,V) and the couples $(X,[m])$, where X is an orbit in $\hat{A}$ and $[m]$ is an equivalence class of irreducible representations of the stability subgroup at a fixed $x_0\in X$. Hence, we can (and we do) associate the maps defined by Eq. (3) with the couples $(X,[m])$ instead of (P,V).

In Sec. III, we prove that the set of vectors given by Eq. (3) is a frame for some $\psi\in\mathcal{H}$ if and only if the Haar measure of $\hat{A}$ restricted to X is not singular with respect to the H-quasi-invariant measures on X and m is a square-integrable representation. Moreover we show that these frames are tight, that their frame bound can be computed by a *generalized orthogonality relation* and we characterize the set of admissible vectors. We stress that, with our assumptions, we have to show these last properties since orthogonality relations are not proved to exist for imprimitivity systems, but only for square-integrable representations. In the particular case that H preserves the group law of $\hat{A}$ (so that, by duality, it is defined the semidirect product $A\times' H$) we obtain the results of Ref. 3.

In Sec. IV, we restrict ourselves to the case of imprimitivity systems such that A and H are Lie groups and the stability subgroup H_0 is compact and we consider the sets of vectors of the form

$$A\times(H/H_0)\times J\ni(a,x,i)\mapsto\left(\int_{\hat{A}}\langle\hat{a},a\rangle V_{q(x)}\,dP(\hat{a})\psi_i\right)\in\mathcal{H},$$

where J is the set of the first n numbers with the counting measure μ_c (n is the dimension of the Hilbert space where m acts) and q is any measurable section from the quotient space H/H_0 to H. We prove that, with a suitable choice of the vectors ψ_i, the above set is a tight frame labeled by the points of the space $(A\times H\times J)$ endowed with the measure $\mu_A^r\otimes\mu_H^r\otimes\mu_c$.

In the literature there are attempts to define frames by the equation

$$G/G_0\ni x\mapsto U_{q(x)}\psi\in\mathcal{H},\tag{4}$$

where G_0 is a closed subgroup of G, q is a suitable section from the quotient space G/G_0 to G and the frame is labeled by the points of G/G_0 endowed with a G-quasi invariant measure (in this case U is called *square-integrable modulo a coset space*), see Ref. 4 and references therein. Nevertheless, there are no general results on the square-integrable representations modulo a coset space and one has to prove for each representation both the fact that the set given by Eq. (4) is a frame and that the frame is tight with an equation for computing the frame bound. In the examples we give in the final section we show that the frames defined by Eq. (4) are particular cases of our procedure, so that our results provide a rigorous and general framework to this class of frames. An alternative approach to this problem can be found in Ref. 5.

In the following section we introduce the mathematical notations and we give some preliminary results, as standard reference we use Ref. 2.

II. MATHEMATICAL PRELIMINARIES

Let X be a locally compact second countable Hausdorff topological space. We denote by $\mathcal{K}(X)$ the vector space of continuous functions with compact support on X, by $\mathcal{B}(X)$ the Borel σ-algebra of X and by $M(X)$ the Banach space of complex measures on $\mathcal{B}(X)$.

By lcsc group G we mean a locally compact second countable Hausdorff topological group. We denote by μ_G a right invariant Haar measure and by Δ_G its modular function.

Let A be an Abelian lcsc group. We denote by $\hat{A}$ the dual group of A. Let $\mathcal{F}$ be the Fourier transform on $M(\hat{A})$ defined as

$$(\mathcal{F}\nu)(a)=\int_{\hat{A}}\overline{\langle x,a\rangle}d\nu(x),\quad a\in A,\quad \nu\in M(\hat{A}).$$

The same symbol $\mathcal{F}$ denotes the Fourier–Plancherel operator on $L^2(\hat{A},\mu_{\hat{A}})$. We choose the Haar measure on $\hat{A}$ in such a way that $\mathcal{F}$ is unitary from $L^2(\hat{A},\mu_{\hat{A}})$ onto $L^2(A,\mu_A)$.

Let G be a lcsc group and X be a lcsc continuous G-space. We denote the action of $g\in G$ on $x\in X$ by $g[x]$, the G-orbit of $x\in X$ by $G[x]$ and the stability subgroup of G at x by G_x.

We recall the following standard results about transitive G-spaces.

Lemma 1: Let G be a lcsc topological group and $X=G[x_0]$ a transitive G-space.

(1) There exists a regular section based on x_0, i.e., a measurable map q from X to G such that

$$q(x_0)=e,$$
$$q(x)[x_0]=x\quad \forall x\in X$$

and, for any compact set K in G, the set $\{q(g[x_0]):g\in K\}$ has compact closure in G.

(2) There exists a strongly quasi-invariant measure ν on X, i.e., a Radon measure ν such that

$$\nu(g[E])=\int_E\lambda_\nu(g,y)d\nu(y),\quad g\in G,\ E\in\mathcal{B}(X),$$

where λ_ν is the cocycle of ν, i.e., it is a continuous function from $G\times X$ to $(0,\infty)$ satisfying

$$\lambda_\nu(g_1g_2,y)=\lambda_\nu(g_1,g_2[y])\lambda_\nu(g_2,y),\quad g_1,g_2\in G,\quad y\in X,$$
$$\lambda_\nu(h,x_0)=\frac{\Delta_{G_{x_0}}(h)}{\Delta_G(h)},\quad h\in G_{x_0}.$$

(3) Two strongly quasi-invariant measures on X are mutually absolutely continuous and the corresponding density is a continuous positive function.

The following lemma is a particular case of the Mackey–Bruhat formula, which allows us to compute the integrals on G as integrals on $G/H\times H$, where H is a closed subgroup of G, see Ref. 3 for an elementary proof based on the fact that G is a second countable space.

Lemma 2: Let G be a lcsc group, $X=G[x_0]$ be a transitive G-space and G_{x_0} be the stability subgroup at x_0. Given a strongly quasi-invariant measure ν and a regular section q based on x_0, define $\beta:X\times H\to G_{x_0}$ as the map

$$\beta(x,h)=q(x)h,\quad x\in X,\quad h\in G_{x_0}.$$

Then the map β is an isomorphism of measurable spaces and the image measure of

$$\frac{1}{\Delta_G(q(x))\lambda_\nu(q(x),x_0)}\,\nu(x)\otimes\mu_{G_{x_0}}(h),$$

under the map β, is a Haar measure of G.

By *Hilbert space* we mean a complex separable Hilbert space with scalar product $\langle\cdot,\cdot\rangle$, linear in the first argument. We use the word *representation* to mean a continuous unitary representation of G acting in a Hilbert space $\mathcal{H}$. If a representation U is square-integrable, we denote by K_U the corresponding *formal degree*.

Finally, we recall the notion of *square-integrability modulo, a coset space* (see Ref. 4 and references therein).

Let G be a lcsc group, X a transitive G-space and U an irreducible representation of G acting in a Hilbert space $\mathcal{H}$. The representation U is said to be *square-integrable modulo X* if there exist n nonzero vectors

$$\psi_1,\ldots,\psi_n\in\mathcal{H}$$

and a section q from X to G such that

$$\Gamma:X\times\{1,...,n\}\ni(x,i)\mapsto U_{q(x)}\psi_i\in\mathcal{H} \tag{5}$$

is a frame over the space $(X\times\{1,...,n\},\mu\otimes\mu_c)$, where μ is a strongly quasi-invariant measure on X and μ_c the counting measure on $\{1,...,n\}$.

III. FRAMES ON A DIRECT PRODUCT

Let H be a lcsc group and A an Abelian lcsc group such that H acts continuously on the dual group $\hat{A}$ of A. Let M be the direct topological product $M=A\times H$, endowed with the measure $\mu^M=\mu_A\otimes\mu_H$.

Let $x_0\in\hat{A}$ and $X=H[x_0]$ be the corresponding orbit. We assume that X is locally closed in $\hat{A}$. Fix a strongly quasi-invariant measure ν on X with cocycle λ and a regular section q from X to H based on x_0.

Let m be an irreducible representation of the stability subgroup H_0 at x_0. Define μ_X as the restriction of $\mu_{\hat{A}}$ to X (since X is locally closed, μ_X is a Radon measure) and (P,V) be the imprimitivity system based on $\hat{A}$ for the group H acting in $\mathcal{H}=L^2(X,\nu,\mathcal{K})$ as

$$(P_E\phi)(x)=\mathcal{X}(x)E\cap X(x)\phi(x), \tag{6}$$

$$(U_a\phi)(x)=\langle x,a\rangle\phi(x), \tag{7}$$

$$(V_h\phi)(x)=(\lambda(h^{-1},x))^{1/2}m(q(x)^{-1}hq(h^{-1}[x]))\phi(h^{-1}[x]), \tag{8}$$

where $x\in X$, $a\in A$, $h\in H$, $E\in\mathcal{B}(\hat{A})$, and $\phi\in L^2(X,\nu,\mathcal{K})$.

Let $\psi\in\mathcal{H}$ and define $\mathbb{F}_\psi$ as the map

$$M\ni(a,h)\mapsto V_{h^{-1}}U_{-a}\psi\in\mathcal{H},$$

and, for all $\phi\in\mathcal{H}$, the map $c_{\phi,\psi}$

$$M\ni(a,h)\mapsto\langle\phi,V_{h^{-1}}U_{-a}\psi\rangle_{\mathcal{H}}\in\mathbb{C}. \tag{9}$$

As in the case of square-integrable representations, we say that ψ is an *admissible vector* for (P,V) if the map $\mathbb{F}_\psi$ is a frame over (M,μ^M), i.e., if the following conditions are satisfied:

(1) for all $\phi\in\mathcal{H},c_{\phi,\psi}\in L^2(M,\mu^M)$;
(2) there exist two positive numbers α and β such that

$$\alpha\|\phi\|^2_{\mathcal{H}}\leqslant\|c_{\phi,\psi}\|^2_{L^2(M,\mu^M)}\leqslant\beta\|\phi\|^2_{\mathcal{H}},\quad \forall\phi\in\mathcal{H},$$

We are now in position to state the main result of the paper. We give necessary and sufficient conditions on the imprimitivity system in order that the set of admissible vectors for (P,V) is not void. Moreover, we show that the corresponding frame $\mathbb{F}_\psi$ is tight, i.e., $\alpha=\beta$, and its frame bound α can be computed by means of a sort of *orthogonality relation*, as Eq. (2).

Theorem 1: There is $\psi\in\mathcal{H}$ such that $\mathbb{F}_\psi$ is a frame in $\mathcal{H}$ over (M,μ^M) if and only if ν and μ_X are not disjoint and m is a square-integrable representation of H_0.

This theorem is a consequence of the following proposition. Let

$$\nu=\nu_s+f\mu_X \tag{10}$$

be the decomposition of ν with respect to μ_X given by the Lebesgue–Radon–Nikodym theorem where f is non-negative and measurable. Defined $\gamma:X\to(0,\infty)$ as

$$\gamma(x)=\frac{f(x)}{\Delta_H(q(x))\lambda(q(x),x_0)},$$

the following result holds.

Proposition 1: Let $\psi,\phi\in\mathcal{H}$, $\psi,\phi\neq 0$. The function $c_{\phi,\psi}$ is in $L^2(M,\mu^M)$ if and only if the following conditions hold:

(a) m is a square-integrable representation of H_0 (we denote by K_m its formal degree);
(b) $\psi(x)=0$ for ν_s-almost all $x\in X$;
(c) $\psi(x)\in\mathrm{Dom}\,K_m^{-1/2}$ for ν-almost all $x\in X$;
(d) the function

$$X\ni x\mapsto\gamma(x)\|K_m^{-1/2}\psi(x)\|^2_{\mathcal{K}}$$

is in $L^1(X,\nu)$.

In this case, we have

$$\|c_{\phi,\psi}\|^2_{L^2(G,\mu_G)}=\|\phi\|^2_{\mathcal{H}}\int_X\gamma(x)\|K_m^{-1/2}\psi(x)\|^2_{\mathcal{K}}d\nu(x). \tag{11}$$

Remark 1: We observe that

(1) the measure ν defines uniquely the density f, hence also γ, up to a ν-negligible set;
(2) the function γ is invariant, up to a positive constant factor and a ν-negligible set, if we change ν with another quasi-invariant measure;
(3) the condition that ν and μ_X are not disjoint implies that $\mu_X(X)>0$.

Proof: Let ϕ, $\psi\in\mathcal{H}$, $\psi,\phi\neq 0$. Then, for all $(a,h)\in M$,

$$c_{\phi,\psi}(a,h)=\int_X\langle x,a\rangle\langle(V_h\phi)(x),\psi(x)\rangle_{\mathcal{K}}\,d\nu(x)=(\mathcal{F}\nu_h)(-a),$$

where ν_h is the canonical extension to $\hat{A}$ of the complex measure on X having $\langle(V_h\phi)(x),\psi(x)\rangle_{\mathcal{K}}$ as density with respect to ν. Applying Fubini theorem and taking into account that A is unimodular, the condition

$$c_{\phi,\psi}\in L^2(M,\mu^M)$$

is equivalent to the following two conditions.

(1) For μ_H-almost all $h \in H$ the function

$$A \ni a \mapsto (\mathcal{F}\nu_h)(a)$$

is in $L^2(A,\mu_A)$;

(2) the function

$$H \ni h \mapsto \int_A |(\mathcal{F}\nu_h)(a)|^2 \, d\mu_A(a)$$

is in $L^1(H,\mu_H)$.

Using a standard result on Fourier transform (see, for example, Theorem 31.33 of Ref. 6) and the fact that the Fourier-Plancherel operator is unitary, the two conditions above turn out to be equivalent to the following.

(1) For μ_H almost all $h \in H$, there is $l_h \in L^2(\hat{A},\mu_{\hat{A}})$ such that

$$l_h \mu_{\hat{A}} = \nu_h .$$

(2) The function

$$H \ni h \rightarrow \int_{\hat{A}} |l_h(x)|^2 \, d\mu_{\hat{A}}(x)$$

is in $L^1(H,\mu H)$.

Since $\nu_h(\hat{A} \setminus X) = 0$, $l_h = 0$ for $\mu_{\hat{A}}$-almost any $x \notin X$ and

$$l_h(x)\mu_X = \langle (V_h\phi)(x), \psi(x) \rangle (\nu_s + f(x)\mu_X).$$

This last equation shows that the above conditions are in fact equivalent to

(a) For μ_H-almost all $h \in H$ and ν_s-almost all $x \in X$,

$$\langle (V_h\phi)(x), \psi(x) \rangle_{\mathcal{K}} = 0,$$

(b) For μ_H-almost all $h \in H$ the function

$$X \ni x \mapsto \langle (V_h\phi)(x), \psi(x) \rangle_{\mathcal{K}} f(x)$$

is in $L^2(X,\mu_X)$;

(c) the function

$$H \ni h \mapsto \int_X |\langle (V_h\phi)(x), \psi(x) \rangle_{\mathcal{K}} f(x)|^2 d\mu_X(x)$$

is in $L^1(H,\mu_H)$.

Denoted by $\mathcal{P}$ the orthogonal projection on the singular part ν_s, we claim that the first of the above conditions is equivalent to the fact that $\mathcal{P}\psi = 0$. Indeed, the first condition implies that for μ_H-almost all $h \in H$, for all $E \in \mathcal{B}(\hat{A})$ and for ν-almost all $x \in X$

$$\langle (P_E V_h \phi)(x), (\mathcal{P}\psi)(x) \rangle_{\mathcal{K}} = 0,$$

hence, by integration over X and taking into account the fact that the map $h \mapsto V_h$ is continuous with respect to the strong operator topology, one has that, for all $h \in H$ and for all $E \in \mathcal{B}(\hat{A})$,

$$\langle P_E V_h \phi, \mathcal{P}\psi \rangle_{\mathcal{H}} = 0.$$

Since $\phi \neq 0$ and (P,V) is irreducible, it follows that $\mathcal{P}\psi = 0$. The converse implication is evident. Taking into account Eq. (10), one obtains the following conditions:

(1) $\mathcal{P}\psi=0$,
(2) for μ_H-almost all $h\in H$ the function

$$X\ni x\mapsto\langle (V_h\phi)(x),\psi(x)\rangle_{\mathcal{K}}\sqrt{f(x)}$$

is in $L^2(X,\nu)$;
(3) the function

$$H\ni h\mapsto\int_X|\langle (V_h\phi)(x),\psi(x)\rangle_{\mathcal{K}}|^2 f(x)d\nu(x)$$

is in $L^1(H,\mu_H)$.

Fixed $h\in H$, performing the change of variables $x\rightarrow h[x]$, these conditions are equivalent to

(1) $\mathcal{P}\psi=0$,
(2) for μ_H-almost all $h\in H$ the function

$$X\ni x\mapsto|\langle \phi(x),m(q(x)^{-1}h^{-1}q(h[x]))\psi(h[x])\rangle_{\mathcal{K}}|^2 f(h[x])$$

is in $L^1(X,\nu)$;
(3) the function

$$H\ni h\mapsto\int_X|\langle (\phi(x),m(q(x)^{-1}h^{-1}q(h[x]))\psi(h[x])\rangle_{\mathcal{K}}|^2 f(h[x])d\nu(x)$$

is in $L^1(H,\mu_H)$.

Again by Fubini theorem and, fixed $x\in X$, performing the change of variables $h\mapsto hq(x)^{-1}$, the above conditions turn out to be equivalent to

(1) $\mathcal{P}\psi=0$,
(2) for ν-almost all $x\in X$ the function

$$H\ni h\mapsto|\langle \phi(x),m(h^{-1}q(h[x_0]))\psi(h[x_0])\rangle_{\mathcal{K}}|^2 f(h[x_0])$$

is in $L^1(H,\mu_H)$;
(3) the function

$$X\ni x\mapsto\int_H|\langle \phi(x),m(h^{-1}q(h[x_0]))\psi(h[x_0])\rangle_{\mathcal{K}}|^2 f(h[x_0])d\mu_H(h)$$

is in $L^1(X,\nu)$.

Finally, using Lemma 2 to compute the integral in the variable h and Fubini theorem, the last conditions are equivalent to

(1) $\mathcal{P}\psi=0$,
(2) for ν-almost all $x,y\in X$ the function

$$H_0\ni s\mapsto\gamma(y)|\langle \phi(x),m(s^{-1})\psi(y)\rangle_{\mathcal{K}}|^2$$

is in $L^1(H_0,\mu_{H_0})$;
(3) for ν-almost all $x\in X$ the function

$$X\ni y\mapsto\gamma(y)\int_{H_0}|\langle \phi(x),m(s^{-1})\psi(y)\rangle_{\mathcal{K}}|^2\, d\mu_{H_0}(s)$$

is in $L^1(X,\nu)$;
(4) the function

$$X \ni x \mapsto \int_X \gamma(y) \int_{H_0} |\langle \phi(x), m(s^{-1}) \psi(y) \rangle_{\mathcal{K}}|^2 d\mu_{H_0}(s) d\nu(y)$$

is in $L^1(X,\nu)$.

The equivalence of these last four conditions to the ones contained in the statement of the proposition is now consequence of the properties of the formal degree operator, see Theorem 3 of Ref. 1. Equation (11) follows from Fubini theorem. □

Proof of theorem: ''only if'': By definition of frame, there are $\psi,\phi \in \mathcal{H}$, $\psi,\phi \neq 0$, such that $c_{\phi,\psi} \in L^2(M,\mu^M)$. Applying Proposition 1, one has that $\psi(x)=0$ ν_s-almost all $x \in X$ and m is square-integrable. The first condition and the fact that $\psi \neq 0$ imply that the measures ν and μ_X are not disjoint.

''If'': Since the measures ν and μ_X are not disjoint, there is a set $C \in \mathcal{B}(X)$ such that

$$\nu(C) > 0,$$

$$\nu_s(C) = 0,$$

$$f(x) > 0 \quad \forall x \in C.$$

Since ν is a regular measure, we can always assume that C is compact. Moreover, due to the fact that m is square-integrable, there is $v \in \mathcal{K}$, $v \neq 0$, such that $v \in \mathrm{Dom}\, K_m^{-1/2}$. Define ψ: $X \to \mathcal{K}$ as

$$\phi(x) = \begin{cases} \dfrac{v}{\sqrt{f(x)}} & x \in C \\ 0 & x \notin C \end{cases}.$$

Obviously $\psi \in \mathcal{H}$, $\psi \neq 0$. Taking into account that $(\Delta_G(q(x))\lambda_\nu(q(x),x_0))^{-1}$ is bounded on the compact sets and the definition of γ, one has that the four conditions of Proposition 1 are satisfied for all $\phi \in \mathcal{H}$, $\phi \neq 0$. Hence, $c_{\phi,\psi}$ is in $L^2(M,\mu^M)$ for all $\phi \in \mathcal{H}$. □

Corollary 1: With the notations of Theorem 1, there is $\psi \in \mathcal{H}$ such that $\mathbb{F}_\psi$ is a frame in $\mathcal{H}$ over (M,μ^M) if and only if there are $\psi,\phi \in \mathcal{H},\psi$, $\phi \neq 0$, such that $c_{\phi,\psi}$ is in $L^2(M,\mu^M)$.

The admissible vectors ψ are the ones satisfying

$$\psi \neq 0,$$

$$\psi(x) = 0, \quad \nu_s \text{ almost all } x \in X,$$

$$\int_X \gamma(x) \|K_m^{-1/2} \psi(x)\|_{\mathcal{K}}^2 \, d\nu(x) < \infty,$$

and the corresponding frame $\mathbb{F}_\psi$ is tight with frame bound

$$\alpha = \int_X \gamma(x) \|K_m^{-1/2} \psi(x)\|_{\mathcal{K}}^2 d\nu(x),$$

where K_m is the formal degree of m.

Proof: It follows easily from Proposition 1. □

We are unable to prove that for the frames $\mathbb{F}_\psi$ given by Theorem 1 the singular part ν_s of the strongly quasi-invariant measure ν is zero nor can we find an example of such frames so that $\nu_s \neq 0$. Nevertheless, there are two partial results in this direction.

The first one applies to the case that H and A are Lie groups.

Corollary 2: Let H and A be Lie groups such that the action of H on $\hat{A}$ is smooth and the orbits are locally closed. There is $\psi \in \mathcal{H}$ such that $\mathbb{F}_\psi$ is a frame in $\mathcal{H}$ over (M, μ^M) if and only if the orbit X is open in $\hat{A}$ and m is a square-integrable representation of H_0.

In particular, the restriction μ_X of the Haar measure $\mu_{\hat{A}}$ to X is a strongly quasi-invariant measure, so that in Eq. (10) $\nu_s = 0$ and f can be chosen continuous and strictly positive.

The proof of the corollary is based on the following standard results on the Lebesgue measures. We recall that, if M is a manifold of dimension n, a Radon measure μ on M is said to be a *Lebesgue measure* if for each chart (U, ϕ) the image under ϕ of the induced measure μ_U is equivalent to the restriction of the Lebesgue measure of $\mathbb{R}^n$ to $\phi(U)$ with a C^∞ density. Then

(1) there exist Lebesgue measures and any two of them are equivalent (see 16.22.2 of Ref. 7).
(2) If N is a submanifold with dimension strictly lower than n and it is locally closed in M, then N is negligible with respect to any Lebesgue measure (see 16.22 of Ref. 7).
(3) If G is a Lie group, then the Haar measure of G is a Lebesgue measure (see 19.6 of Ref. 7).
(4) If G is a Lie group acting on M and the action is transitive and smooth, then a Radon measure μ is G-quasi-invariant if and only if it is equivalent to a Lebesgue measure (see, for example, Ref. 8).

Proof: As a consequence of Theorem 1, there is $\psi \in \mathcal{H}$ such that $\mathbb{F}_\psi$ is a frame in $\mathcal{H}$ over (M, μ^M) if and only if m is a square-integrable representation of H_0 and there is a measurable subset C of X such that $\nu(C) > 0$ and $\mu_{\hat{A}}(C) > 0$. So, we can assume, without loss of generality, that $\mu_{\hat{A}}(X) > 0$. Under this assumption, we show that the restriction μ_X of $\mu_{\hat{A}}$ to X is quasi-invariant. Indeed, by means of a standard result of Lie groups, the orbit X is a (locally closed) submanifold and, since $\mu_{\hat{A}}$ is a Lebesgue measure and $\mu(X) > 0$, then $\dim X = \dim \hat{A}$, i.e. X is open in $\hat{A}$, and μ_X is a Lebesgue measure (with respect to the manifold X). By the results referred above, the claim is now evident.

The statement of the corollary is now clear. □

The same conclusion can be proved if we assume that the action of H on $\hat{A}$ preserves the composition law of $\hat{A}$.

Corollary 3: Let H and A be lcsc groups with A abelian such that the action of H on $\hat{A}$ satisfies

$$h[x_1 + x_2] = h[x_1] + h[x_2], \quad h \in H, \;\; x_1, x_2 \in \hat{A}, \tag{12}$$

and the corresponding orbits are locally closed. There is $\psi \in \mathcal{H}$ such that $\mathbb{F}_\psi$ is a frame in $\mathcal{H}$ over (M, μ^M) if and only if $\mu_{\hat{A}}(X) > 0$ and m is a square-integrable representation of H_0.

In particular, the restriction μ_X of the Haar measure $\mu_{\hat{A}}$ to X is a strongly quasi-invariant measure, so that in Eq. (10) $\nu_s = 0$ and f can be chosen continuous and strictly positive.

The proof is based on the following Lemma, see, for example, Ref. 3, for the proof.

Lemma 3: For any $h \in H$ and $E \in \mathcal{B}(\hat{A})$ we have

$$\mu_{\hat{A}}(h[E]) = \rho(h^{-1}) \mu_{\hat{A}}(E),$$

where ρ: $H \to (0, \infty)$ is a continuous group homomorphism.

Proof of the corollary: Arguing as in the proof of the previous corollary, we can assume, without loss of generality, that $\mu_{\hat{A}}(X) > 0$. Hence, due to the above Lemma, the measure μ_X is a strongly quasi-invariant measure on X. The thesis in now evident. □

We observe that, if Eq. (12) holds, the frame (9) is associated in a natural way with a square-integrable representation. Indeed, since the action of H on $\hat{A}$ satisfies (12), by duality, H acts also on A in such a way that

$$h[a_1+a_2]=h[a_1]+h[a_2], \quad h\in H, \ a_1,a_2\in A,$$

and the set M acquires a structure of a lcsc group with respect to the group law

$$(a_1,h_1)(a_2,h_2)=(a_1+h_1[a_2],h_1h_2).$$

To stress the structure of group, we denote M by G in the following.

By definition, the group G is the semidirect product of A and H, and the measure $\mu_A\otimes\mu_H$ is a right invariant Haar measure. Moreover there is a one-to-one correspondence between the irreducible imprimitivity systems (P,V) and the irreducible representations U of G, explicitly given by

$$(P,V)\mapsto U_{(a,h)}:=U_aV_h,$$

where $U_a=\int_{\hat{A}}\langle x,a\rangle dP(x)$, $a\in A$. As a consequence, the fact that there is $\psi\in\mathcal{H}$ such that the map $\mathbb{F}_\psi$ given by Eq. (9) is a frame is precisely the fact that U is a square-integrable representation of G and, hence, Corollary 3 characterizes completely the square-integrable representations of groups that are semidirect products with an Abelian normal factor.[9,3]

IV. FRAMES ON A QUOTIENT SPACE

In this section, we assume that A is an Abelian Lie group and H is a Lie group acting smoothly on $\hat{A}$, the dual group of A.

Fix $x_0\in\hat{A}$ and an irreducible representation m of the stability subgroup H_0 at x_0. Let $X=H[x_0]$ be the orbit of x_0 (as usual we assume that X is locally closed) and $\mathcal{K}$ the Hilbert space where m acts. We suppose that

(1) the stability subgroup H_0 is compact;

(2) the orbit X of x_0 has positive measure with respect to the Haar measure $\mu_{\hat{A}}$ of $\hat{A}$.

From these assumptions one has the following properties.

(1) The stability subgroup H_0 is unimodular and we can normalize the Haar measure μ_{H_0} in such a way that $\mu_{H_0}(H_0)=1$.

(2) The orbit X admits an H-invariant measure ν, which is equivalent to the restriction μ_X of $\mu_{\hat{A}}$ to X. Let f be the continuous and positive function from X to $\mathbb{R}$ such that

$$\nu=f\mu_X,$$

so that the cocycle λ of μ_X is

$$\lambda(h,x)=\frac{f(x)}{f(h[x])}, \quad x\in X, \ h\in H$$

(since ν is invariant the cocycle of ν is the identity).

(3) The Hilbert space $\mathcal{K}$, in which the representation m acts, is finite dimensional, so we can assume that $\mathcal{K}=\mathbb{C}^n$. Moreover, the representation m is square-integrable and the corresponding operator of formal degree is proportional to the identity.

Let (P,V) be the imprimitivity system for H based on $\hat{A}$ and acting on $\mathcal{H}=L^2(X,\nu,\mathbb{C}^n)$ whose equivalence class corresponds to $(X,[m])$. If q is a regular section from X to H based on x_0, then the system (P,V) is explicitly given by Eqs. (6)–(8).

Choose $\Psi\in\mathcal{K}(X),\Psi\neq 0$, and define ψ from X to $\mathbb{C}$ as

$$\psi(x)=\int_{H_0}\sqrt{\frac{f(s[x])}{f(x)}}\,\Psi(s[x])\,d\mu_{H_0}(s), \tag{13}$$

the integral being finite since Ψ has compact support, f is continuous and H_0 is compact. Taking into account that H_0 is unimodular, for all $s'\in H_0$ and $x\in X$

$$\psi(s'[x])=\sqrt{\lambda(s',x)}\,\psi(x). \tag{14}$$

Indeed, we have

$$\begin{aligned}\psi(s'[x])&=\int_{H_0}\sqrt{\frac{f(ss'[x])}{f(s'[x])}}\,\Psi(ss'[x])\,d\mu_{H_0}(s)\\&=(s\mapsto ss'^{-1})\\&=\int_{H_0}\sqrt{\frac{f(s[x])}{f(s'[x])}}\,\Psi(s[x])\,d\mu_{H_0}(s)\\&=\sqrt{\frac{f(x)}{f(s'[x])}}\int_{H_0}\sqrt{\frac{f(s[x])}{f(x)}}\,\Psi(s[x])\,d\mu_{H_0}(s)\\&=\sqrt{\lambda(s',x)}\,\psi(x).\end{aligned}$$

Moreover, since H_0 is compact and $\Psi\in\mathcal{K}(X)$, still $\psi\in\mathcal{K}(X)$ and, recalling that the cocycle of ν is equal to 1, one has that

$$\alpha:=\int_X\frac{f(x)}{\Delta_H(q(x))}|\psi(x)|^2\,d\nu(x)<\infty. \tag{15}$$

Finally, let $e_1,\dots,e_n$ be the standard basis of $\mathbb{C}^n$, J be the set of the first n numbers and μ^c the corresponding counting measure.

Corollary 4: With the above notations, the map

$$A\times X\times J\ni(a,x,i)\mapsto V_{q(x)}U_a(\psi e_i)\in\mathcal{H}$$

is a tight frame in $\mathcal{H}$ over $(A\times X\times J,\mu_{\hat{A}}\otimes\nu\otimes\mu^c)$, with frame bound α.

Proof: Let V^0 and U^0 be the representations of H and A corresponding to the choice of the trivial representation m^0 of H_0 ($\mathcal{K}^0=\mathbb{C}$). Since ψ, defined as above, satisfies Eq. (15), then

$$(a,h)\mapsto V^0_{h^{-1}}U^0_{-a}\psi \tag{16}$$

is a tight frame in $L^2(X,\nu)$ over $(A\times H,\mu_{\hat{A}}\otimes\mu_H)$, with frame bound α. Now, let ϕ be any vector in $\mathcal{H}$. Then, by the fact that (16) is a frame, it follows that:

$$\begin{aligned}\alpha\|\phi\|^2_{\mathcal{H}}&=\int_{A\times H}\langle\|\phi(x)\|_{\mathbb{C}^n},V^0_{h^{-1}}U^0_{-a}\psi\rangle_{L^2(X,\nu)}\,d\mu_A(a)\otimes d\mu_H(h)\\&=\int_H\int_A\left|\int_X\|\phi(x)\|_{\mathbb{C}^n}\overline{(V^0_{h^{-1}}U^0_{-a}\psi)(x)}\,d\nu(x)\right|^2 d\mu_A(a)\,d\mu_H(h)=(x\mapsto h^{-1}[x])\\&=\int_H\int_A\left|\int_X\|\phi(h^{-1}[x])\|_{\mathbb{C}^n}\overline{(U^0_{-a}\psi)(x)}\,d\nu(x)\right|^2 d\mu_A(a)\,d\mu_H(h)\end{aligned}$$

$$=(\text{Fourier transform and } \nu=f\mu_X)=\int_H\int_X \|\phi(h^{-1}[x])\|^2_{\mathbb{C}^n}|\psi(x)|^2 f(x)\,d\nu(x)\,d\mu_H(h)$$

$$=(h\mapsto h^{-1})=\int_H\int_X \|\phi(h[x])\|^2_{\mathbb{C}^n}|\psi(x)|^2 f(x)\,d\nu(x)\,d\mu_H(h).$$

Now, using the Mackey–Bruhat formula in order to compute the integral on H as an integral on $X\times H_0$, we have

$$\alpha\|\phi\|^2_{\mathcal{H}}=\int_X\int_{H_0}\int_X \|\phi((q(y)s)[x])\|^2_{\mathbb{C}^n}|\psi(x)|^2 f(x)\,d\nu(x)\,d\mu_{H_0}(s)\,d\nu(y)$$

$$=(x\mapsto s^{-1}[x] \text{ and unitarity of } m)=\int_X\int_{H_0}\int_X \|(V_{q(y)^{-1}}\phi)(x)\|^2_{\mathbb{C}^n}$$

$$\times|\psi(s^{-1}[x])|^2 f(s^{-1}[x])\,d\nu(x)\,d\mu_{H_0}(s)\,d\nu(y)$$

$$=[\text{Eq.}(14)]=\int_X\int_{H_0}\int_X \|(V_{q(y)^{-1}}\phi)(x)\|^2_{\mathbb{C}^n}|\psi(x)|^2 f(x)\,d\nu(x)\,d\mu_{H_0}(s)\,d\nu(y).$$

At this point, recalling that we have set $\mu_{H_0}(H_0)=1$, we can perform the integral over H_0; thus, we obtain that

$$\alpha\|\phi\|^2_{\mathcal{H}}=\int_X\int_X \|(V_{q(y)^{-1}}\phi)(x)\|^2_{\mathbb{C}^n}|\psi(s^{-1}[x])|^2 f(x)\,d\nu(x)\,d\nu(y).$$

Then, since $\{e_1,\ldots,e_n\}$ is the canonical basis in $\mathbb{C}^n$, we find

$$\alpha\|\phi\|^2_{\mathcal{H}}=\int_X\int_X\sum_i |\langle (V_{q(y)^{-1}}\phi)(x),\psi(x)e_i\rangle_{\mathbb{C}^n}|^2 f(x)\,d\nu(x)\,d\nu(y)$$

$$=(\text{Fourier transform and } \nu=f\mu_X)$$

$$=\int_X\int_A\sum_i |\langle (V_{q(y)^{-1}}\phi),U_{-a}(\psi e_i)\rangle_{\mathcal{H}}|^2 d\mu_A(a)\,d\nu(y)=(a\mapsto -a)$$

$$=\int_X\int_A\sum_i |\langle \phi,V_{q(y)}U_a(\psi e_i)\rangle_{\mathcal{H}}|^2 d\mu_A(a)\,d\nu(y).$$

The thesis is now evident. □

V. EXAMPLES

In this section we apply the previous results to some specific groups that are considered in the literature.

A. The causal group

Let G be the *causal group*, namely the semidirect product

$$G=\mathbb{R}^4\times'(\mathrm{SO}_0(3,1)\times\mathbb{R}^+_*),$$

where $\mathrm{S0}_0(3,1)$ is the connected component with the identity of the Lorentz group and $\mathbb{R}^+_*$ is the multiplicative group of strictly positive real numbers. This group has been considered in the context of wavelet electrodynamics.[10]

In order to take into account the projective representations of G, we will study the universal covering group of G, i.e.,

$$G^* = \mathbb{R}^4 \times' (\mathrm{SL}(2,\mathbb{C}) \times \mathbb{R}_*^+).$$

The group G^* is the semidirect product of the Abelian group $A = \mathbb{R}^4$ and the group $H = \mathrm{SL}(2,\mathbb{C}) \times \mathbb{R}_*^+$, so by duality H acts on the dual group $\mathbb{P}^4$ of $\mathbb{R}^4$, preserving the composition law of $\mathbb{P}^4$. Hence, we can consider unitary representations of G^* instead of imprimitivity systems for H based on $\mathbb{P}^4$.

In order to apply Corollary 3, we identify the dual group $\mathbb{P}^4$ of $\mathbb{R}^4$ with $\mathbb{R}^4$ by means of the pairing

$$\langle p,a \rangle = p_0 a_0 - p_1 a_1 - p_2 a_2 - p_3 a_3 .$$

The Haar measure on $\mathbb{P}^4$ is the Lebesgue measure dp and the action of H on $\mathbb{P}^4$ is

$$(h,d)[p] = \frac{1}{d} L(h) a, \quad (h,d) \in \mathrm{SL}(2,\mathbb{C}) \times \mathbb{R}_*^+, \quad p \in \mathbb{P}^4,$$

where $L: \mathrm{SL}(2,\mathbb{C}) \to \mathrm{SO}_0(3,1)$ is the covering homomorphism. The corresponding orbits of $\mathrm{SL}(2,\mathbb{C}) \times \mathbb{R}_*^+$ can be classified in the following way:

(1) two ''massive'' orbits O^+, O^- defined by

$$O^{\pm} = G^*[p^{\pm}] = \{p \in \mathbb{R}^4 : \pm p_0 > 0, \|p\| > 0\},$$

where $p^{\pm} = (\pm 1,0,0,0)$ and $\|p\| = \sqrt{p_0^2 - p_1^2 - p_2^2 - p_3^2}$;

(2) the orbit O defined by

$$O = G^*[(0,0,0,1)] = \{p \in \mathbb{R}^4 : \|p\| < 0\};$$

(3) the singleton orbit $\{(0, 0, 0, 0)\}$.

Since the orbits are locally closed, we can apply Corollary 3 to select the square-integrable representations. The orbits having positive Lebesgue measure are $O^{\pm}, O$. The stability subgroup at $(0,0,0,1) \in O$ is the Euclidean group $E(2)$, which does not admit square-integrable representations, whereas the one associated with $O^{\pm}$ is isomorphic to SU(2). Since SU(2) is compact, every irreducible representation is square-integrable and let

$$\{D^j : j = 0, \tfrac{1}{2}, 1, \ldots\} \tag{17}$$

be the canonical maximal set of inequivalent irreducible representations of SU(2) where each D^j acts on $\mathbb{C}^{2j+1}$. Then, a maximal set of inequivalent square-integrable representations of G^* is

$$\{U^{\pm,j} := \mathrm{Ind}_{\mathbb{R}^4 \times' \mathrm{SU}(2)}^{G^*} (p^{\pm} D^j) : j = 0, \tfrac{1}{2} 1, \ldots\}.$$

Observing that the restriction ν of the measure $\|p\|^{-4}\, dp$ to the orbit $O^{\pm}$ is invariant with respect to the action of G^*, for each j the representation $U^{\pm,j}$ acts on the Hilbert space $\mathcal{H}^j := L^2(O^{\pm}, \nu, \mathbb{C}^{2j+1})$ as

$$(U^j_{(a,h,d)} \psi)(p) = e^{i\langle p,a \rangle} D^j (q(p)^{-1} (hd) q((hd)^{-1}[p])) \psi((hd)^{-1}[p])$$

where $\psi \in \mathcal{H}^j$, $(a,h,d) \in G^*$ and q is a regular section from $O^{\pm}$ into $\mathrm{SL}(2,\mathbb{C}) \times \mathbb{R}_*^+$. The corresponding set $\mathcal{A}^j$ of admissible vectors is given by Corollary 1. To this aim, notice that ν has density $f(p)=\|p\|^{-4}$ with respect to dp (according to Corollary 3, the singular part ν_s of ν with respect to dp is zero), $SL(2,\mathbb{C})\times\mathbb{R}_*^{\pm}$ is unimodular and the formal degree of the representation D^j is proportional to the identity, then

$$\mathcal{A}^j=\left\{\psi\in\mathcal{H}^j: \quad \int_{O\pm}\|\psi(p)\|^2_{\mathbb{C}^{2j+1}}\|p\|^{-4}\,d\mu(p)<\infty\right\}.$$

B. Lorentz invariant frames

It is well known that the Poincaré group does not admit square integrable representations, so that there are no frames given by Eq. (9), which, obviously, are invariant with respect to the full Poincaré group. Nevertheless we can define frames that are invariant with respect to the Lorentz group by means of the imprimitivity systems for the Lorentz group based on $\mathbb{R}^3$. The action of the Lorentz group on $\mathbb{R}^3$ is defined by identifying, as a manifold, $\mathbb{R}^3$ with with the *massive* orbit in the momentum space

$$X_m=\{(p_0,\vec{p})\in\mathbb{R}^4: p_0^2-\vec{p}^2=m^2,\ \ p_0>0\}.$$

These kinds of frames were introduced by means of suitable representations of the Poincaré group that are square-integrable modulo a coset space, see Ref. 4 and references therein.

As usual we consider the universal covering group $H=SL(2,\mathbb{C})$ of the connected component $SO_0(3,1)$ of the Lorentz group. We denote by $\mu_{SL(2,\mathbb{C})}$ a Haar measure of $SL(2,\mathbb{C})$, which is a unimodular group, and by L the covering homomorphism from $SL(2,\mathbb{C})$ onto $SO_0(3,1)$.

Let A be the Abelian Lie group $\mathbb{R}^3$. As usual, we identify the dual group $\mathbb{P}^3$ of A with $\mathbb{R}^3$ by means of the Euclidean scalar product. Define the diffeomorphism $\overline{\omega_m}$ from $\mathbb{R}^3$ onto X_m as

$$\overline{\omega_m}(\vec{p})=(\sqrt{m^2+\vec{p}^2},\vec{p})$$

with inverse

$$\overline{\omega_m}^{-1}(p_0,\vec{p})=\vec{p}.$$

The space $\mathbb{P}^3$ becomes a transitive H-space with respect to the smooth action

$$h[\vec{p}]=\overline{\omega_m}^{-1}(L(h)\overline{\omega_m}(\vec{p}))$$

with $h\in SL(2,\mathbb{C})$ and $\vec{p}\in\mathbb{P}^3$.

We recall the following facts about $\mathbb{P}^3$ as an H-space.

(1) The H-invariant measure on $\mathbb{P}^3$ is

$$\nu_m=\frac{d\vec{p}}{\sqrt{m^2+\vec{p}^2}},$$

with density $f(\vec{p})=1/\sqrt{m^2+\vec{p}^2}$ with respect to the Lebesgue measure $d\vec{p}$ of $\mathbb{P}^3$.

(2) The stability subgroup at the origin $\vec{0}$ is $H_0=SU(2)$, which is compact.

(3) The set $\{D^j\}$ given by Eq. (17) is a set of inequivalent representations of SU(2).

Fixed $j=0,\frac{1}{2},1,\ldots$, each couple $(\mathbb{P}^3,[D^j])$ defines two classes of frames in the Hilbert space $\mathcal{H}^j=L^2(\mathbb{P}^3,\nu_m,\mathbb{C}^{2j+1})$. Indeed, let

$$(U^j_{\vec{a}}\phi)(\vec{p})=e^{i\vec{a}\cdot\vec{p}}\phi(\vec{p}),$$

$$(V^j_h\phi)(\vec{p})=D^j(q(\vec{p})^{-1}hq(h^{-1}[\vec{p}]))\phi(h^{-1}[\vec{p}]),$$

where $\vec{p}\in\mathbb{P}^3$, $\phi\in\mathcal{H}^j$, $h\in H$, $\vec{a}\in A$ and q is any regular section from $\mathbb{P}^3$ to H.

We observe that the formal degree of the representations D^j is proportional to the identity and that

$$\gamma(\vec{p}) = f(\vec{p}) = \frac{1}{\sqrt{m^2+\vec{p}^2}}$$

since H is unimodular and ν_m is invariant. By means of Corollary 2 and Corollary 1, if $\psi \in \mathcal{A}^j$, where

$$\mathcal{A}^j = \left\{ \psi \in \mathcal{H} : \psi \neq 0, \frac{\psi(\vec{p})}{(m^2+\vec{p}^2)^{1/2}} \in L^2(\mathbb{P}^3, d\vec{p}, \mathbb{C}^{2j+1}) \right\},$$

then the map

$$\mathbb{R}^3 \times \mathrm{SL}(2,\mathbb{C}) \ni (a,h) \mapsto V_h^{-1} U_a^{-1} \psi \in \mathcal{H}^j \tag{18}$$

is a tight frame over $(\mathbb{R}^3 \times \mathrm{SL}(2,\mathbb{C}), d\vec{a} \otimes \mu_{\mathrm{SL}(2,\mathbb{C})})$ with frame bound

$$\alpha = \int_{p^3} \frac{\|\psi(\vec{p})\|^2_{\mathbb{C}^{2j+1}}}{(m^2+\vec{p}^2)} d\vec{p}.$$

Moreover, we notice that the measure ν_m is an H-invariant measure on $\mathbb{P}^3$ with density

$$f(\vec{p}) = \frac{1}{\sqrt{m^2+\vec{p}^2}}$$

with respect to the Lebesgue measure $d\vec{p}$ and we define J as the set of the first n numbers and μ^c the corresponding counting measure.

Let $(e_i)_{i=1}^{2j+1}$ be the standard basis of $\mathbb{C}^{2j+1}$ and $\Psi \in \mathcal{K}(\mathbb{P}^3)$. Define ψ by means of Eq. (13), then, as a consequence of Corollary 4, the map

$$\mathbb{R}^3 \times \mathbb{P} \times J \ni (\vec{a}, \vec{p}, i) \mapsto V^j_{q(\vec{p})} U^j_{\vec{a}} (\psi e_i) \in \mathcal{H}^j \tag{19}$$

is a tight frame in $L^2(\mathbb{P}^3, \nu_m, \mathbb{C}^{2j+1})$ over $(A \times \mathbb{P}^3 \times J, d\vec{a} \otimes \nu_m \otimes \mu^c)$ with frame bound

$$\alpha = \int_{p^3} \frac{\|\psi(\vec{p})\|^2_{\mathbb{C}^{2j+1}}}{(m^2+\vec{p}^2)} d\vec{p}.$$

These classes of frames were introduced in Ref. 4 by the use of an irreducible representation W of the universal covering group G of the connected component of the Poincaré group such that W be *square integrable modulo a section*. We now show the relationship between the two constructions.

Consider first the frame (18) and define $\mathbb{T}$ as the subgroup of the time translations, i.e.,

$$\mathbb{T} = \{ (p_0, \vec{0}, I) \in G : p_0 \in \mathbb{R} \}.$$

Then, the manifold $A \times H$ is diffeomorphic to the space of left cosets $G/\mathbb{T}$ and it becomes a transitive G-space. One has that the map $c: A \times H \to G$ given by

$$(\vec{a}, h) \mapsto (h[(0,\vec{a})], h)$$

is a smooth section and the measure $d\vec{a} \otimes \mu_{\mathrm{SL}(2,\mathbb{C})}$ is G-invariant.

Let $G_0 = \mathbb{R}^4 \times' \mathrm{SU}(2)$ and χ_m be the character of $\mathbb{R}^4$

$$\chi_m(a_0,\vec{a}) = e^{ia_0 m}, \quad (a,\vec{a}) \in \mathbb{R}^4,$$

then $\chi \otimes D^j$ is an irreducible representation of G_0 and we can consider the representation $W^{m,j}$ unitarily induced from G_0 to G by $\chi_m \otimes D^j$, which is explicitly given by

$$(W^{m,j}_{(a,h)}\phi)(\vec{p}) = e^{i(a_0\sqrt{m^2+\vec{p}^2} - \vec{a}\cdot\vec{p})} D^j(q_m(\vec{p})^{-1} h q_m(h^{-1}[\vec{p}]))\phi(h^{-1}[\vec{p}]),$$

where $a \in \mathbb{R}^4$, $h \in H$, $p \in \mathbb{P}^3$, and $\phi \in \mathcal{H}^j$. Comparing the form of $W^{m,j}$ with the one of U^j and V^j, it follows that

$$W^{m,j}_{c(-\vec{a},h)} = V^j_h U^j_{\vec{a}}.$$

Hence, we can conclude that, given any $\psi \in \mathcal{A}^j$, the representation $W^{m,j}$ is square integrable mod. $(\mathbb{R}^3 \times \mathrm{SL}(2,\mathbb{C}), c, \psi)$ and that the frame (18) coincides with the one defined in Ref. 4.

Now consider the frame (19). The space $A \times \mathbb{P}^3$ is a transitive G-space with respect to the action

$$(b,h)[(\vec{a},\vec{p})] := (pr_A(q(h[\vec{p}])^{-1}[b] + (q(\vec{p})^{-1} h q(h^{-1}[\vec{p}]))[\vec{a}]), q(\vec{p})),$$

where $b \in \mathbb{R}^4$, $h \in \mathrm{SL}(2,\mathbb{C})$, $\vec{a} \in A$, $\vec{p} \in \mathbb{P}^3$ and pr_A is the canonical projection from $\mathbb{R}^4$ to A, i.e.,

$$pr_A(b_0,\vec{b}) = \vec{b}.$$

Observe that

(1) the stability subgroup at $(\vec{0},\vec{0})$ is $\mathbb{T} \times' \mathrm{SU}(2)$, where $\mathbb{T}$ is the subgroup of the time translations and, since SU(2) does not act on $\mathbb{T}$ the semidirect product is in fact a direct product;

(2) a continuous section from $A \times \mathbb{P}^3$ to G is
$$\tilde{q}(\vec{a},\vec{p}) = (q(\vec{p})[(0,\vec{a})], q(\vec{p}));$$

(3) the G-invariant measure is $d\vec{a} \otimes \nu_m$.

Hence one has that

$$W^{m,j}_{\tilde{q}(-\vec{a},\vec{p})} = V^j_{q(\vec{p})} U^j_{\vec{a}}.$$

Since the map (19) is a frame, then the representation $W^{m,j}$ is a square-integrable modulo $(A \times \mathbb{P}^3, \tilde{q}_m, \psi e_1, \ldots, \psi e_{2j+1})$.

C. The Galilei group

We now consider a class of frames that are invariant under the homogeneous Galilei group.

Let $\mathcal{V} := \mathbb{R}^3$ and $d\vec{v}$ be the corresponding Lebesgue measure. Let SU(2) be the universal covering of the rotation group in $\mathbb{R}^3$ and $\mu_{\mathrm{SU}(2)}$ be the corresponding normalized Haar measure. The group SU(2) acts on $\mathcal{V}$ by means of the covering homomorphism δ and we can consider the semidirect product,

$$H := \mathcal{V} \times' \mathrm{SU}(2).$$

The group H is unimodular with Haar measure $d\vec{v} \otimes \mu_{\mathrm{SU}(2)}$.

Moreover, let $A = \mathbb{R}^3$ and $\mathbb{P}^3$ be the corresponding dual group identified with $\mathbb{R}^3$ by means of the Euclidean scalar product. The corresponding Lebesgue measures are denoted by $d\vec{a}$ and $d\vec{p}$, respectively.

Fixed $m \in \mathbb{R}$, $m \neq 0$, the space $\mathbb{P}^3$ is a transitive G space with respect to the nonlinear transitive smooth action of H

$$(\vec{v},h)[\vec{p}]=\delta(h)\vec{p}+m\vec{v}. \tag{20}$$

The measure $d\vec{p}$ is H-invariant and the stability subgroup H_0 at $\vec{0}$ is SU(2).

As in the previous case, for each $j=0,\frac{1}{2},1,...$, the couple $(\mathbb{P}^3,[D^j])$ defines two classes of frames in the Hilbert space $\mathcal{H}^j=L^2(\mathbb{P}^3,d\vec{p},\mathbb{C}^{2j+1})$. Indeed, let

$$(U^j_{\vec{a}}\phi)(\vec{p})=e^{i\vec{a}\cdot\vec{p}}\phi(\vec{p}),$$

$$(V^j_h\phi)(\vec{p})=D^j(q(\vec{p})^{-1}hq(h^{-1}[\vec{p}]))\phi(h^{-1}[\vec{p}])$$

where $\vec{a}\in A,\ h\in H,\ \vec{p}\in\mathbb{P}^3,\ \phi\in\mathcal{H}^j$ and $q:\mathbb{P}^3\rightarrow H$ is a section for the action of H on $\mathbb{P}^3$ [for example, $q(\vec{p})=(\vec{p}/m)I)$].

Indeed, by Corollary 2, for *any* nonzero $\psi\in\mathcal{H}^j$, the map

$$A\times H\ni(\vec{a},h)\mapsto V^j_{h^{-1}}U^j_{-\vec{a}}\psi\in\mathcal{H}^j \tag{21}$$

is a tight frame in $\mathcal{H}$ over $(A\times H,d\vec{a}\otimes\mu_{\mathrm{SU}(2)})$ with frame bound $\|\psi\|^2$. It is easy to prove that two frames of the type (21) corresponding to the same value of j and to the same analyzing vector ψ, but differing for the value of the mass $m\neq0$, are unitarily equivalent.

Moreover, let $(e_i)_{i=1}^{2j+1}$ be the standard basis of $\mathbb{C}^{2j+1}$, J the set of the first n numbers, and μ^c the corresponding counting measure. Given $\Psi\in\mathcal{K}(\mathbb{P}^3)$, define ψ by means of Eq. (13), then, as a consequence of Corollary 4, the map

$$(\vec{a},\vec{p},i)\mapsto V^j_{q(\vec{p})}U^j_{\vec{a}}(\psi e_i) \tag{22}$$

is a tight frame in $L^2(\mathbb{P}^3,d\vec{p},\mathbb{C}^{2j+1})$ over $(A\times\mathbb{P}^3\times J,\ d\vec{a}\otimes d\vec{p}\otimes\mu^c)$.

In this particular case, one can check by direct computation that in Eq. (22) any vector $\psi\in L^2(\mathbb{P}^3,d\vec{p})$ can be used to define a frame.

As in the previous example, we can show that the above frames can be obtained using a (projective) representation of covering group G of the full Galilei group which is square-integrable modulo a section. Let $W^{m,j}$ be the (projective) representation of G associated with the Galilei invariant quantum particle of mass m and spin j (for an explicit description of this representations see, for example, Ref. 11), then one can show, as in the case of the Poincaré group, that $W^{m,j}$ is square integrable modulo the left coset space $G/\mathbb{T}$ (where $\mathbb{T}$ is the subgroup of the time translations) since the map (21) is a frame and that $W^{m,j}$ is square integrable modulo the left coset space $G/(\mathbb{T}\times'\mathrm{SU}(2))$ since the map (22) is a frame.

D. The Weyl–Heisenberg group

Let us consider the $(2n+1)$-dimensional Weyl–Heisenberg group H_n, which is the group associated with the canonical quantization of a classical mechanical system.

Fixed $n\in\mathbb{N}$, $n>0$, let $T=\mathbb{R}$ and $A_n=\mathcal{V}_n=\mathbb{R}^n$ be the usual vector groups, then the Lie group H_n is the manifold $T\times A_n\times\mathcal{V}_n$ with the composition law

$$(\tau_1,\vec{a}_1,\vec{u}_1)(\tau_2,\vec{a}_2,\vec{u}_2)=(\tau_1+\tau_2-\tfrac{1}{2}(\vec{a}_1\cdot\vec{u}_2-\vec{a}_2\cdot\vec{u}_1),\vec{a}_1+\vec{a}_2,\vec{u}_1+\vec{u}_2).$$

The group H_n is the semidirect product of the Abelian group $T\times A_n$ and the group $\mathcal{V}_n$, whose action is given by

$$\vec{u}[(\tau,\vec{a})]=(\tau+\vec{a}\cdot\vec{u},\vec{a}),\quad (\tau,\vec{a})\in T\times A_n,\quad \vec{u}\in\mathcal{V}_n,$$

where the dot denotes the Euclidean product. Identifying the dual group $\mathbb{P}\times\mathbb{P}^n$ of $T\times A_n$ with $T\times A_n$ itself by means of the Euclidean product, the dual action of $\mathcal{V}_n$ on $\mathbb{P}\times\mathbb{P}^n$ is given by

$$\vec{u}[(h,\vec{p})]=(h,\vec{p}-h\vec{u}), \quad (h,\vec{p})\in\mathbb{P}\times\mathbb{P}^n, \;\; \vec{u}\in\mathcal{V}_n. \tag{23}$$

Then, the $\mathcal{V}_n$-orbits in $\mathbb{P}\times\mathbb{P}^n$ are the following:

(1) for all $\vec{p}\in\mathbb{R}^n$, the singleton orbits

$$\mathcal{O}_{0,\vec{p}}=\{(0,\vec{p})\};$$

(2) for all $h\in\mathbb{R}$, $h\neq 0$, the orbits

$$\mathcal{O}_h=\{(h,\vec{p})|\vec{p}\in\mathbb{R}^n\}.$$

Since all the orbits are closed, according to Corollary 3, the group H_n does not have square-integrable representations. Nevertheless, fixed $h\neq 0$, the orbit $\mathcal{O}_h$ is canonically identified with $\mathbb{P}^n$, which is the dual group of $\mathbb{R}^n$, so that we can use our procedure to define frames associated with irreducible imprimitivity systems for the group $\mathcal{V}_n$ based on the transitive $\mathcal{V}_n$-space $\mathbb{P}^n$.

Since the stability subgroup at $\vec{0}\in\mathbb{P}_n$ is the identity, we have only the couple $(\mathbb{P}_n,I)$ and it defines a class of tight frames in $\mathcal{H}=L^2(\mathbb{P}^n,d\vec{p})$.

Indeed, let

$$(U_{\vec{a}}\phi)(\vec{p})=e^{i\vec{a}\cdot\vec{p}},$$

$$(V_{\vec{u}}\phi)(\vec{p})=\phi(\vec{p}+h\vec{u}),$$

where $\vec{p}\in\mathbb{P}_n$, $\phi\in\mathcal{H}$, $\vec{u}\in\mathcal{V}_n$, and $\vec{a}\in\mathbb{R}^n$, and observe that the Lebesgue measure $d\vec{p}$ is clearly invariant with respect to the action (23).

Hence, since $\mathcal{V}_n$ is unimodular and its Haar measure is $d\vec{u}$, it follows from Corollary 2 and Corollary 1 that for any nonzero vector $\psi\in\mathcal{H}$, the map

$$\mathbb{F}^h_\psi: \quad A_n\times\mathcal{V}_n\ni(\vec{a},\vec{u})\mapsto V^h_{-\vec{u}}U_{-\vec{a}}\psi\in\mathcal{H}$$

is a tight frame in $\mathcal{H}$ over the space $(A_n\times\mathcal{V}_n,\, d\vec{a}\otimes d\vec{u})$ with frame bound $\|\psi\|$.

We remark that, given the unitary operator $\mathfrak{U}_h$, $h\in\mathbb{R}-\{0\}$, in $\mathcal{H}$ defined by

$$(\mathfrak{U}_h\phi)(\vec{p})=|h|^{-n/2}\phi(h^{-1}\vec{p}), \quad \phi\in\mathcal{H},$$

then, for any $\psi\in\mathcal{H}$ and $h_1,h_2\in\mathbb{R}-\{0\}$, we have

$$\mathbb{F}^{h_1}_\psi=\mathfrak{U}_{h_1/h_2}\circ\mathbb{F}^{h_2}_{\psi'} \qquad \text{if} \quad \psi'=\mathfrak{U}^*_{h_1/h_2}\psi,$$

namely, the frames $\mathbb{F}^{h_1}_\psi$ and $\mathbb{F}^{h_2}_{\psi'}$ are unitarily equivalent. Moreover we notice explicitly that, for $n=3$ and $h=m\in\mathbb{R}^+$, the frame $\mathbb{F}^h_\psi$ is exactly the frame (22) over the coset space $G/(\mathbb{T}\times\mathrm{SU}(2))$, where G is the Galilei group, with the choice $j=0$.

Also in this case, the frames $\mathbb{F}^h_\psi$ can be defined in terms of a representation of the group H_n that is square-integrable modulo a coset space. Indeed, for sake of simplicity, fix $h=1$ and let W be the irreducible representation of H_n acting on $\mathcal{H}=L^2(\mathbb{P}^n,d\vec{p})$ as

$$(W_{(\tau,\vec{a},\vec{u})}\phi)(\vec{p})=e^{i(\tau+(1/2)\vec{a}\cdot\vec{u}-\vec{a}\cdot\vec{p})}\phi(\vec{p}-\vec{u}), \quad \phi\in\mathcal{H}.$$

The left coset space $H_n/\mathbb{T}$ is clearly diffeomorphic to $A_n\times\mathcal{V}_n$, so that $A_n\times\mathcal{V}_n$ turns out to be a H-space where the action is explicitly given by

$$(\tau,\vec{a},\vec{u})[(\vec{a}',\vec{u}')]=(\vec{a}+\vec{a},\vec{u}+\vec{u}) \quad (\tau,\vec{a},\vec{u})\in H_n, \quad (\vec{a}',\vec{u}')\in A_n\times\nu_n.$$

One can easily check that the map

$$q\colon A_n\times\mathcal{V}_n\ni(\vec{a},\vec{u})\mapsto(\vec{u}[(0,\vec{a})],\vec{u})=(\vec{a}\cdot\vec{u},\vec{a},\vec{u})\in H_n$$

is a smooth section and $d\vec{a}\otimes d\vec{u}$ is an invariant measure on $A_n\times\mathcal{V}_n$. Finally, for any $\psi\in\mathcal{H}$, we have

$$\mathbb{F}_\psi(\vec{a},\vec{u})=W_{q(\vec{a},\vec{u})}\psi.$$

Thus, since, for any nonzero vector $\psi\in\mathcal{H},\mathbb{F}_\psi$ is a frame over the measured space $(A_n\times\mathcal{V}_n,d\vec{a}\otimes d\vec{u})$, it follows that W is square-integrable mod$(A_n\times\mathcal{V}_n,q;\psi)$.

Moreover, one can show that, for $n=3$, if the analyzing vector ψ is chosen to be a *Gaussian function*, the family of vectors

$$\{W_{q(\vec{a},\vec{u})}\psi|\vec{a}\in\mathbb{R}^3,\vec{u}\in\mathbb{R}^3\}$$

is nothing but the classical canonical family of coherent states associated to the quantum harmonic oscillator.

[1] M. Duflo and C. C. Moore, ''On the regular representations of a nonunimodular locally compact group,'' J. Funct. Anal. **21**, 209–243 (1976).
[2] G. B. Folland, *A Course in Abstract Harmonic Analysis* (CRC, Boca Raton, FL, 1995).
[3] P. Aniello, C. Cassinelli, E. De Vito, and A. Levrero, ''Square-integrability of induced representations of semidirect products,'' Rev. Math. Phys. **10**, 301–313 (1998).
[4] S. T. Ali, J.-P. Antoine, J.-P. Gazeau, and U. A. Mueller, ''Coherent states and their generalizations: a mathematical overview,'' Rev. Math. Phys. **7**, 1013–1104 (1995).
[5] S. T. Ali, ''A general theorem on square-integrability: vector coherent states,'' J. Math. Phys. **39**, 3954–3964 (1999).
[6] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis*, Vol. II (Springer, New York, 1970).
[7] J. Dieudonné, *Elements d'Analyse*, Vols. 3 and 4 (Gauthiers-Villars, Paris, 1970 and 1971).
[8] G. Warner, *Harmonic Analysis on Semisimpie Lie Groups*, Vols. I and II (Springer, New York, 1972).
[9] A. Kleppner and R. L. Lipsman, ''The Plancherel formula for group extensions,'' Ann. Sci. Ec. Norm. Sup. **5**, 459–516 (1972).
[10] G. Kaiser, ''Wavelet electrodynamics,'' Phys. Lett. A **168**, 28–34 (1992).
[11] V. S. Varadarajan, *Geometry of Quantum Theory*, second ed. (Springer, New York, 1985).

General solution of string inspired nonlinear equations

I. Bandos[a)]
The Abdus Salam ICTP, P.O. Box 586, 34100, Trieste, Italy and Institute for Theoretical Physics, NSC Kharkov Institute of Physics and Technology, 310108, Kharkov, Ukraine

E. Ivanov[b)]
Bogoliubov Laboratory of Theoretical Physics, JINR, 141 980 Dubna, Moscow Region, Russia

A. A. Kapustnikov[c)] and S. A. Ulanov[d)]
Department of Physics, Dnepropetrovsk University, 320625, Dnepropetrovsk, Ukraine

(Received 12 February 1999; accepted for publication 4 May 1999)

We present the general solution of the system of coupled nonlinear equations describing dynamics of D-dimensional bosonic string in the geometric (or embedding) approach. The solution is parametrized in terms of the left- and right-moving Lorentz harmonic variables providing a special coset space realization of the product of two $(D-2)$-dimensional spheres $S^{D-2}=[\mathrm{SO}(1,D-1)/\mathrm{SO}(1,1)\times \mathrm{SO}(D-2)\circledS K_{D-2}]$. © *1999 American Institute of Physics.* [S0022-2488(99)02309-9]

I. INTRODUCTION

The bosonic string (and p-brane) theory allows a geometric description in terms of extrinsic geometry of the worldsheet treated as a surface embedded into a flat D-dimensional Minkowski space.[1–4] (See Ref. 3 for a supersymmetric generalization of the classical surface theory and its application to superstrings and $N=1$ superbranes.)

In this approach the dynamics of a free relativistic string is described by the Maurer–Cartan equation supplemented with the additional conditions which insure the string worldsheet to be a minimal surface embedded into a flat target space-time. All these additional constraints can be solved algebraically, after which one is left with some $\mathrm{SO}(1,D-1)$ valued connection form whose curvature vanishes due to the Maurer–Cartan equations.[5] Thus, though the string in the geometric approach is described by nonlinear equations,[1,2,5] the latter are finally reduced to the zero curvature conditions for the $\mathrm{SO}(1,D-1)$ valued connection form properly specified in terms of the independent field variables.[5]

The system of independent equations describing a free string theory in the D-dimensional Minkowski space was derived, for the first time, by Zheltukhin.[1] In the form when all the gauge symmetries inherent in the string theory are kept unfixed, this system is constituted by the WZNW sigma-model-type equation

$$\partial_{(--)}((\partial_{(++)}G)G^T)^{ij}=e^{2W}G^{[i|k}M^{(++)k}_{(--)}M^{(--)|j]}_{(++)} \tag{1}$$

for the $\mathrm{SO}(D-2)$ valued matrix field G^{ij},

$$GG^T=I,$$

and the Liouville-type equation

[a)]Lise Meitner Fellow of the ''Fonds zur Förderung der wissenschaftlichen Forschung'' at the Institut für Theoretische Physik, Technische Universität Wien, A-1040 Vienna Austria. Electronic mail: bandos@tph32.tuwien.ac.at
[b)]Electronic mail: eivanov@thsun1.jinr.ru
[c)]Electronic mail: alexandr@ff.dsu.dp.ua
[d)]Electronic mail: theorph@ff.dsu.dp.ua

$$\partial_{(++)}\partial_{(--)}W=\tfrac{1}{4}M^{(--)i}_{(++)}G^{ij}M^{(++)j}_{(--)}e^{2W} \tag{2}$$

for a ''scalar density'' e^W.[5] Thus, the considered system involves these two independent fields, as well as two chiral SO$(D-2)$ vector fields $M^{(--)i}_{(++)}$, $M^{(++)i}_{(++)}$,

$$\partial_{(--)}M^{(--)j}_{(++)}=0, \quad \partial_{(++)}M^{(++)j}_{(--)}=0, \tag{3}$$

which appear on the right-hand side.

Already at the early stage of working out the geometric approach it was observed that at least for the low dimensions D this system is exactly solvable. It is reduced to the Liouville equation for $D=3$ and a complex Liouville equation for $D=4$. The general solution for $D=5$ was found in Ref. 2. However, for the generic case the general solution of the Eqs. (1), (2), and (3) was so far unknown.

Here we present the general solution of these equations for any value of D. This proves them to provide a new example of a nontrivial system of exactly solvable nonlinear equations and opens an opportunity to study the classical and quantum string theory in terms of a new left- and right-moving variables which parametrize the general solution.

The meaning and origin of our result require a few comments.

The standard equations of motion in the string theory become linear in the conformal gauge. The general solution of these linear equations is given by the sum of chiral and antichiral functions (subjected to the Virasoro constraints). The string-inspired nonlinear equations (i.e., the equations describing bosonic string theory in the geometric approach) encode the information about just the same dynamical system. Thus it is natural to expect (and, as we demonstrate here, this is indeed the case) that these equations are exactly solvable and that their general solution should be expressible in terms of chiral data.

Thus, the first step was to seek for an adequate set of chiral variables appropriate for constructing the general solution. It turned out that the necessary variables are provided by the harmonic approach[6] adapted to the case of Lorentz groups in Refs. 7 and 8.

Namely, in Ref. 9 it was found that the ''constrained chiral twistorlike variables'' (which can be identified with the spinor SO$(1,D-1)$ Lorentz harmonics[10,11,8]) can be used to obtain a covariant solution of the chiral Virasoro constraints

$$\partial_{(++)}X^{\underline{m}}_L\partial_{(++)}X^{\underline{m}}_L=0, \quad \partial_{(--)}X^{\underline{m}}_R\partial_{(--)}X^{\underline{m}}_R=0,$$

$$\partial_{(--)}X^{\underline{m}}_L=0, \quad \partial_{(++)}X^{\underline{m}}_R=0 \quad (\underline{m}=0,1,...,D-1)$$

which are to be imposed on the solutions $X^{\underline{m}}_L, X^{\underline{m}}_R$ of the string equations of motion in the conformal gauge. The $D=3,4,6$ strings were treated in this way. It was analyzed how the fields composed from such chiral twistors are related to the corresponding string-inspired nonlinear equations.

Here, instead of solving the chiral Virasoro constraints of D-dimensional string theory, we *construct two chiral moving frames* given by the two sets of chiral SO$(1,D-1)$ Lorentz harmonic variables,[7] and identify the left- and right-moving vectors $\partial_{(++)}X^{\underline{m}}_L$ and $\partial_{(--)}X^{\underline{m}}_R$ with the lightlike components of the relevant harmonic matrix. Each set of the moving frame variables parametrizes a special coset space of the group SO$(1,D-1)$. Despite the fact that the group SO$(1,D-1)$ on its own is noncompact, this coset space is compact and is isomorphic to the D-dimensional sphere[11]

$$S^{D-2}=\mathrm{SO}(1,D-1)/[\mathrm{SO}(1,1)\times \mathrm{SO}(D-2)\otimes K_{D-2}].$$

It provides us with the appropriate chiral data for constructing the general solution of the string-inspired nonlinear equations (1)–(3). We construct such a solution explicitly and argue that this is the *general* one.

Basic notation. Our conventions basically coincide with those of Ref. 5. The indices $m,n,\ldots=0,\ 1=(++),\ (--)(\{v^m\}\equiv\{v^{\pm\pm}=v^0\pm v^1\})$ label the worldsheet vectors, while $\underline{m},\underline{n},\ldots=0,1,\ldots,(D-1)$ are the flat target space vector indices.

For the tangent space indices we use the notations $a,b,\ldots=0,1$ and $\underline{a},\underline{b},\ldots=0,1,\ldots,(D-1)$. Thus, underlined indices always correspond to the D-dimensional target manifold and its tangent space, while the nonunderlined ones refer to the two-dimensional ($d=2$) worldsheet and its tangent.

The indices $+,\ -$ denote the weights of the tangent space vectors and spinors with respect to both the worldsheet Lorentz group SO(1,1) and the SO(1,1) subgroup of the target space Lorentz group $SO(1,D-1)$. These two subgroups are identified with each other in that version of the geometric approach to string theory which we follow in the present consideration (see Refs. 3 and 5, and references therein). For example, we write $V^a=(V^0,V^1)=(\frac{1}{2}(V^{++}+V^{--}),\frac{1}{2}(V^{++}-V^{--}))$ for the $d=2$ tangent space vectors [and reserve the notations $\psi^\alpha=(\psi^+,\psi^-)$ for the $d=2$ spinors which appear in supersymmetric generalizations of our approach]. The indices $+,\ -$ within the parentheses denote the weights with respect to the $d=2$ conformal symmetry, e.g., $d\xi^m=(d\xi^0,d\xi^1)=(\frac{1}{2}(d\xi^{(++)}+d\xi^{(--)}),\frac{1}{2}(d\xi^{(++)}-d\xi^{(--)}))$. The weights with respect to the chiral affine $SO(1,1)_L$ and $SO(1,1)_R$ transformations are indicated in the same way.

We use the subscripts L and R to denote the chiral functions of the string worldsheet coordinates $\xi^m=(\xi^{(++)},\xi^{(--)})$

$$f_L=f_L(\xi^{(++)}),\quad f_R=f_R(\xi(--)),$$

$$\partial_{(--)}f_L=0,\quad \partial_{(++)}f_R=0.$$

They should not be confused with the calligraphic subscripts $\mathcal{L}$ and $\mathcal{R}$ carried by some fields. The chiral Lorentz harmonics (chiral moving frame variables) are denoted by the letters l and r, while the generic (nonchiral) Lorentz harmonics are denoted by u.

II. GEOMETRIC APPROACH TO STRING DYNAMICS AND STRING-INSPIRED NONLINEAR EQUATIONS

A. Lorentz harmonics

We begin with the definition of the moving frame variables (Lorentz harmonics[6,7]) which are the basic entities of the geometric approach to D-dimensional bosonic string theory[8,3,5]

$$u_{\underline{m}}^{\underline{a}}(\xi)\equiv(\tfrac{1}{2}(u_{\underline{m}}^{++}+u_{\underline{m}}^{--}),\quad u_{\underline{m}}^{i},\tfrac{1}{2}(u_{\underline{m}}^{++}-u_{\underline{m}}^{--}))$$
$$\underline{m}=0,1,\ldots,D-1,\quad \underline{a}=0,1,\ldots,D-1. \tag{4}$$

These objects are subjected to the following orthonormality conditions:

$$u_{\underline{a}}^{\underline{m}}u_{\underline{m}\underline{b}}=\eta_{\underline{a}\underline{b}}\equiv\mathrm{diag}(1,-1,\ldots,-1)\Leftrightarrow\begin{cases}u_{\underline{m}}^{++}u^{\underline{m}++}=0,\quad u_{\underline{m}}^{--}u^{\underline{m}--}=0,\\ u_{\underline{m}}^{++}u^{\underline{m}--}=2,\\ u_{\underline{m}}^{++}u^{\underline{m}i}=0,\quad u_{\underline{m}}^{--}u^{\underline{m}i}=0,\\ u_{\underline{m}}^{i}u^{\underline{m}j}=-\delta^{ij}\end{cases} \tag{5}$$

which imply that the $D\times D$ matrix $u_{\underline{m}}^{\underline{a}}$ (4) belongs to the group $SO(1,D-1)$

$$u_{\underline{m}}^{\underline{a}}(\xi)\in SO(1,D-1). \tag{6}$$

The completeness condition

$$u_{\underline{m}}^{\underline{a}}u_{\underline{n}\underline{a}}\equiv\tfrac{1}{2}u_{\underline{m}}^{++}u_{\underline{n}}^{--}+\tfrac{1}{2}u_{\underline{m}}^{--}u_{\underline{n}}^{++}-u_{\underline{m}}^{i}u_{\underline{n}}^{i}=\eta_{\underline{m}\underline{n}}\equiv\mathrm{diag}(1,-1,\ldots,-1) \tag{7}$$

follows from Eq. (5). In Eqs. (4) and (7) the lightlike notation is used

$$u_{\underline{m}}^{++} \equiv u_{\underline{m}}^{0} + u_{\underline{m}}^{D-1} = u_{0\underline{m}} - u_{D-1\underline{m}} = u_{--\underline{m}},$$

$$u_{\underline{m}}^{--} \equiv u_{\underline{m}}^{0} - u_{\underline{m}}^{D-1} = u_{0\underline{m}} + u_{D-1\underline{m}} = u_{++\underline{m}}. \tag{8}$$

The constraints (5) and (7) as they stand are invariant under the local SO$(1,D-1)$ transformations acting on the tangent space indices $\underline{a},\underline{b},\ldots$. Below we will see that for constructing the geometric approach description of the string theory this symmetry should be restricted to SO(1,1)$\otimes$SO$(D-2)$ in accordance with the splitting (4). Just this local symmetry is respected by the basic geometric postulate of such a description, namely by the condition that the Lorentz harmonic frame should be adapted to the string worldsheet [see Eq. (23) below]. This means that two of $(D-2)$ vectors $u_{\underline{m}}^{\underline{a}}$ are chosen to be tangent to the worldsheet while the remaining ones $u_{\underline{m}}^{i}$ are orthogonal to it. The local (gauge) symmetry SO(1,1)$\otimes$SO$(D-2)$ reflects, first, the freedom of the $d=2$ Lorentz rotation of the vectors $u_{\underline{m}}^{0,(D-1)}$ tangent to the worldsheet and, second, the freedom of SO$(D-2)$ rotations of the vectors $u_{\underline{m}}^{i}$ orthogonal to the worldsheet [in the lightlike notation (8) the $d=2$ rotations are realized as the opposite weights scaling transformations of the vectors u^{++} and u^{--}].

Thus the vectors $u_{\underline{m}}^{\underline{a}}$ appropriate for the description of the external geometry of the bosonic string worldsheet parametrize the *noncompact* coset space

$$\frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times \mathrm{SO}(D-2)}. \tag{9}$$

In other words, the harmonics (4) regarded as the worldsheet fields define a map of the worldsheet $\mathcal{M}^{(1,1)}=\{\xi^m\}$ onto the noncompact coset (9)

$$u_{\underline{m}}^{\underline{a}}:\ \ \mathcal{M}^{(1,1)}=\{\xi^m\}\rightarrow \frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times \mathrm{SO}(D-2)}. \tag{10}$$

Below we will see that the basic ingredients of the sought general solution of the string-inspired equations will be smaller sets of left- and right-moving Lorentz harmonics which parametrize some compact subspaces in two chiral copies of the coset space (9).

1. Cartan 1-forms

Differentials of the harmonic variables can be calculated with taking into account the conditions (4). Differentiating Eq. (4) produces the equation

$$du_{\underline{m}}^{\underline{a}}\, u^{\underline{b}\,\underline{m}} + u^{\underline{a}\,\underline{m}}\, du_{\underline{m}}^{\underline{b}} = 0$$

which can be solved as follows

$$du_{\underline{m}}^{\underline{a}} = u_{\underline{m}}^{\underline{b}}\Omega_{\underline{b}}^{\ \underline{a}}(d) \Leftrightarrow \begin{cases} du_{\underline{m}}^{++} = u_{\underline{m}}^{++}\omega + u_{\underline{m}}^{i} f^{++i}(d),\\ du_{\underline{m}}^{--} = -u_{\underline{m}}^{--}\omega + u_{\underline{m}}^{i} f^{--i}(d),\\ du_{\underline{m}}^{i} = -u_{\underline{m}}^{j}A^{ji} + \tfrac{1}{2}u_{\underline{m}}^{++} f^{--i}(d) + \tfrac{1}{2}u_{\underline{m}}^{--} f^{++i}(d). \end{cases} \tag{11}$$

Here

$$\Omega_{\underline{b}}^{\underline{a}} \equiv u_{\underline{b}}^{\underline{m}} du_{\underline{m}}^{\underline{a}} = \begin{pmatrix} \omega & 0 & \frac{1}{\sqrt{2}} f^{--i}(d) \\ 0 & -\omega & \frac{1}{\sqrt{2}} f^{++i}(d) \\ \frac{1}{\sqrt{2}} f^{++i}(d) & \frac{1}{\sqrt{2}} f^{--i}(d) & A^{ji}(d) \end{pmatrix}, \quad \Omega^{\underline{a}\,\underline{b}} \equiv \eta^{\underline{a}\,\underline{c}} \Omega_{\underline{c}}^{\underline{b}} = -\Omega^{\underline{b}\,\underline{a}} \tag{12}$$

are the SO(1,$D-1$) Cartan forms [in the vector representation of the SO(1,$D-1$) generators]. Due to (4), they are naturally divided into the three subsets: (i) the SO(1,1)$\times$SO($D-2$) covariant forms

$$f^{++i} \equiv u_{\underline{m}}^{++}\, du^{\underline{m}i}, \tag{13}$$

$$f^{--i} \equiv u_{\underline{m}}^{--}\, du^{\underline{m}i}, \tag{14}$$

which constitute a vielbein of the noncompact coset space SO(1,$D-1$)/SO(1,1)$\otimes$SO($D-2$), (ii) the SO(1,1) (spin) connection

$$\omega \equiv \tfrac{1}{2} u_{\underline{m}}^{--}\, du^{\underline{m}++}, \tag{15}$$

and (iii) the SO($D-2$) connections (gauge fields)

$$A^{ij} \equiv u_{\underline{m}}^{i}\, du^{\underline{m}j} \equiv -u_{\underline{m}}^{j}\, du^{\underline{m}i}. \tag{16}$$

2. Parabolic subgroup

As was mentioned above, the choice of the tangent space local group as SO(1,1)$\otimes$SO(D) [and the coset space (9) for Lorentz harmonics] is motivated by the adaptation postulate which ''solders'' harmonics to the worldsheet and is thus relevant just to the geometric description of strings. Formally, the same harmonic (moving frame) variables (4) can be used to give a geometric description of the massless particle. But in this case the adaptation of the moving frame would consist in requiring that one of the lightlike vectors, e.g., $u_{\underline{m}}^{++}$, is tangent to the worldline, while u^i and u^{--} are orthogonal to it. Such an adaptation is covariant with respect to the following right gauge transformations[11] (see also Ref. 10, where the Hamiltonian form of the corresponding $D=4$ transformations was presented)

$$u_{\underline{m}}^{++\prime} = u_{\underline{m}}^{++} V^{-1},$$

$$u_{\underline{m}}^{i\prime} = (u_{\underline{m}}^{++} V^{--j} + u_{\underline{m}}^{j}) V^{ji},$$

$$u_{\underline{m}}^{--\prime} = (u_{\underline{m}}^{--} + u_{\underline{m}}^{++} V^{--i} V^{--i} + 2 u_{\underline{m}}^{i} V^{--j}) V. \tag{17}$$

The transformations (17) form the maximal proper subgroup (parabolic subgroup) SO(1,1)$\times$SO($D-2$)$\otimes$$K_{D-2}$ of the Lorentz group SO(1,$D-1$).[11] An arbitrary element of this subgroup is characterized by the SO(1,1) transformation $V=e^{\alpha}$, the matrix V^{ij} of SO($D-2$)-orthogonal rotations and the parameters V^{--i} of the boosts K_{D-2}. Thus, in the geometric description of massless particle, *as well as in any case when the adaptation of the moving frame involves only one lightlike moving frame vector*, the harmonics $u_{\underline{m}}^{\underline{a}}$ (4) can be regarded as parameters of the compact coset space

$$S^{D-2} = \frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times \mathrm{SO}(D-2) \otimes K_{D-2}}. \tag{18}$$

It is isomorphic to a $(D-2)$-dimensional sphere S^{D-2}.[11]

The Cartan forms (13), (14), (15), and (16) are transformed under (17) as follows:

$$f^{++i\prime}=f^{++j}V^{ji}V^{-1}, \tag{19}$$

$$f^{--i\prime}=(f^{--j}+2\mathcal{D}V^{--j}-2f^{++k}(V^{--k}V^{--j}-\tfrac{1}{2}\delta^{kj}V^{--l}V^{--l}))V^{ji},$$

$$\mathcal{D}V^{--j}\equiv dV^{--j}+\omega V^{--j}-V^{--k}A^{kj}, \tag{20}$$

$$\omega'=\omega-f^{++i}V^{--i}+V\,dV^{-1}, \tag{21}$$

$$A^{ij\prime}=(V^{-1}AV)^{ij}-(V^{-1}\,dV)^{ij}-2f^{++k}V^{k[i|}V^{--l}V^{l|j]}. \tag{22}$$

Though the K_{D-2} transformations (17) with the parameters V^{--i} *are not* the gauge symmetries of the whole bosonic string theory, they play a crucial role for understanding the group-theoretical structure of the general solution of the nonlinear equations (1) and (2) (see Sec. III) [It was noticed in Ref. 5 that the boost symmetry allows one to introduce a nontrivial dependence on a spectral parameter into the connection 1-forms entering the zero curvature representation for the nonlinear equations (1) and (2).] Moreover, to define the general solution, we introduce two chiral sets of moving frame variables, $l_{\underline{m}}^{(\underline{a})}$ and $r_{\underline{m}}^{(\underline{a})}$ (see Sec. II B). Their ''adaptation'' is realized just in the ''particlelike'' fashion and thus respects covariance under the chiral counterparts of the maximal parabolic symmetry (17) [cf. (82) and (83)].

B. The first-order form of string equations and the geometric approach to string theory

The Lorentz harmonics give us a possibility to rewrite the string equations of motion in the first order form, namely, as the following set of equations[8,3,5]

$$dX^{\underline{m}}=\tfrac{1}{2}e^{++}u^{--\underline{m}}+\tfrac{1}{2}e^{--}u^{++\underline{m}}, \tag{23}$$

$$d(e^{++}u^{--\underline{m}}-e^{--}u^{++\underline{m}})=0. \tag{24}$$

Here $e^{\pm\pm}=d\xi^m e_m^{\pm\pm}$ is a worldsheet vielbein. While dealing with (23) and (24), one should take into account the restrictions (11) on the differentials of the harmonic variables.

Equation (23) is the adaptation relation already mentioned earlier. It plays the basic role in the geometric approach to strings. In particular, it implies that the intrinsic worldsheet metric is identified with the induced one

$$\tfrac{1}{2}(e_m^{++}e_n^{--}+e_m^{--}e_n^{++})=g_{mn}=\partial_m X^{\underline{m}}\partial_n X_{\underline{m}}. \tag{25}$$

The geometric meaning of Eq. (23) consists in that the string worldsheet is identified with a surface embedded into the D-dimensional Minkowski spacetime. On its own right, it has no dynamical content and gives rise to the purely geometric corollaries.

The integrability conditions $(d\,dX\equiv 0)$ for Eq. (23) are as follows:

$$T^{++}\equiv de^{++}-e^{++}\wedge\omega=0, \quad T^{--}\equiv de^{--}+e^{--}\wedge\omega=0, \tag{26}$$

$$e^{++}\wedge f^{--i}+e^{--}\wedge f^{++i}=0 \Leftrightarrow f_{--}^{--i}-f_{++}^{++i}=0. \tag{27}$$

Thus they require the torsion 2-forms $T^{\pm\pm}$ to vanish, thereby imposing proper constraints on the induced spin connection ω. Besides, they imply the SO(1,1) invariant components of the covariant 1-forms f^{++i} and f^{--i} to coincide with each other

$$f_{--}^{--i}=f_{++}^{++i}\equiv\tfrac{1}{2}h^i. \tag{28}$$

The quantity h^i can be easily recognized as the mean curvatures of the embedded surface.[4] Indeed, using Eq. (23) to express the lightlike harmonic vectors in terms of derivatives of the embedding functions X, we arrive at the standard expression for the mean curvatures

$$h^i = g^{mn} K^i_{mn}, \quad K^i_{mn} = -\partial_m \partial_n X^{\underline{m}} u^i_{\underline{m}}. \tag{29}$$

1. Minimal embedding

Using Eq. (11), one finds that Eq. (24) implies the vanishing of the mean curvatures (28)

$$h^i = 0. \tag{30}$$

Thus the surface defined by Eq. (24) is minimal.[4]

On the other hand, expressing all the auxiliary variables u and e through the derivatives of the embedding functions $X(\xi)$ and using the induced metric (25)

$$g_{mn} \equiv \partial_m X^{\underline{m}} \partial_n X_{\underline{m}}, \tag{31}$$

and its inverse g^{mn}, one can rewrite Eq. (30) [or (24)] in the form

$$\partial_m (\sqrt{-g} g^{mn} \partial_n X^{\underline{m}}) = 0, \tag{32}$$

which is the standard string equations of motion pertinent to the Nambu–Goto action (see Ref. 5 for details). Thus the strings dynamics in the geometric approach is contained just in Eq. (24).

Note that it is important for the geometric approach description that the minimal embedding is described by the covariant Cartan forms f^{++i}, f^{--i} containing in their decomposition only one of the two basic forms e^{--}, e^{++}

$$f^{++i} = e^{--} f^{++i}_{--}, \quad f^{--i} = e^{++} f^{--i}_{++} \tag{33}$$

[cf. Eqs. (28) and (30)]. Actually, Eqs. (33) encode three previous equations: (27), (28), and (30) [which amounts to (24)]. Thus the string dynamics proves to be eventually encoded just in Eqs. (33).

C. Maurer–Cartan equation and the string-inspired nonlinear equations

The integrability conditions for Eqs. (11) produce the Maurer–Cartan equations

$$d\Omega_{\underline{a}}{}^{\underline{b}} - \Omega_{\underline{a}}{}^{\underline{c}} \wedge \Omega_{\underline{c}}{}^{\underline{b}} = 0. \tag{34}$$

With making use of Eq. (4), Eqs. (34) naturally split into the following equations for the coset vielbeins $f^{\pm\pm i}$ (13) and (14) and the connection 1-forms ω, A^{ij} (15) and (16):

$$\mathcal{D} f^{++i} \equiv d f^{++i} - f^{++i} \wedge \omega + f^{++j} \wedge A^{ji} = 0, \tag{35}$$

$$\mathcal{D} f^{--i} \equiv d f^{--i} + f^{--i} \wedge \omega + f^{--j} \wedge A^{ji} = 0, \tag{36}$$

$$\mathcal{R} = d\omega = \tfrac{1}{2} f^{--i} \wedge f^{++i}, \tag{37}$$

$$R^{ij} \equiv dA^{ij} + A^{ik} \wedge A^{kj} = -f^{--[i} \wedge f^{++j]}. \tag{38}$$

Equations (35)–(38) amount to the Peterson-Codazzi, Gauss, and Ricci equations of the classical surface theory.[4]

Thus, in the geometric approach framework, the dynamics of string is described by the vielbein $e^{\pm\pm}$ and the set of Cartan forms ω, $f^{\pm\pm i}, A^{ij}$ which satisfy Eqs. (26) and (33) and the Maurer–Cartan Eqs. (35)–(38).[4,1,2,3,5]

The most essential general feature of this approach is that Eqs. (33), (26), (35), and (36) can be solved algebraically.[5]

Indeed, using the fact that any connection is integrable on any one-dimensional subspace, we can specify the expressions for the SO(1,1) and SO$(D-2)$ connection 1 forms in the following way:

$$\omega = e^{++}\nabla_{++}(W-L) - e^{--}\nabla_{--}(W+L), \tag{39}$$

$$A^{ij} = e^{++}\nabla_{++}G^{ik}_{\mathcal{R}}G^{jk}_{\mathcal{R}} + e^{--}\nabla_{--}G^{ik}_{\mathcal{L}}G^{jk}_{\mathcal{L}}, \tag{40}$$

where $G_{\mathcal{L}}$ and $G_{\mathcal{R}}$ are some SO$(D-2)$ group matrices,

$$G_{\mathcal{L}}G^T_{\mathcal{L}} = I, \quad G_{\mathcal{R}}G^T_{\mathcal{R}} = I. \tag{41}$$

Equations (39) and (40) provide a possibility to rewrite Eqs. (26), (35), and (36) as the conditions of closeness of some 1-forms

$$d(e^{++}\exp(W+L)) = 0, \quad d(e^{--}\exp(W-L)) = 0. \tag{42}$$

$$d(f^{++i}G^{ji}_{\mathcal{R}}\exp(-W+L)) = 0, \quad d(f^{--i}G^{ji}_{\mathcal{L}}\exp(-W-L)) = 0 \tag{43}$$

[Eqs. (33) have been used when deriving (43)].

The general solution to Eqs. (42) (up to some possible topological subtleties which are unessential for the present study) is provided by

$$\begin{aligned} e^{++} &= d\xi^{(++)}M^{(++)}_{(++)}(\xi^{(++)})\exp(-W-L), \\ e^{--} &= d\xi^{(--)}M^{(--)}_{(--)}(\xi^{(--)})\exp(-W+L). \end{aligned} \tag{44}$$

Here, $\xi^{(\pm\pm)}$ are some functions of the string worldsheet coordinates with the only defining demand that their differentials are linearly-independent 1-forms. It is convenient, however, to choose $\xi^{(\pm\pm)}$ as a set of local coordinates on the worldsheet

$$\xi^m = \xi^{(\pm\pm)}. \tag{45}$$

This choice fixes a gauge with respect to the worldsheet reparametrizations (general coordinate transformations), so that only two-dimensional conformal reparametrizations survive [see Eq. (61) below].

In the holonomic basis $d\xi^{(\pm\pm)}$ the components of the vielbein form $e^{\pm\pm}$ are

$$\begin{gathered} e^{++}_{(++)} = M^{(++)}_{(++)}\exp(-W-L), \quad e^{++}_{(--)} = 0, \\ e^{--}_{(++)} = 0, \quad e^{--}_{(--)} = M^{(--)}_{(--)}\exp(-W+L), \\ e^{(++)}_{--} = 0, \quad e^{(--)}_{--} = 2(M^{(--)}_{(--)})^{-1}\exp(W-L), \\ e^{(++)}_{++} = 2(M^{(++)}_{(++)})^{-1}\exp(W+L), \quad e^{(--)}_{++} = 0, \end{gathered} \tag{46}$$

the induced metric (25) is conformally flat

$$ds^2 = d\xi^m\, d\xi^n\, g_{mn} = d\xi^{(++)}\, d\xi^{(--)}\, M^{(++)}_{(++)}M^{(--)}_{(--)}e^{2W} \tag{47}$$

and the covariant derivatives are proportional to the corresponding holonomic ones

$$e^{++}\nabla_{++} = d\xi^{(++)}\,\partial_{(++)}, \quad e^{--}\nabla_{--} = d\xi^{(--)}\,\partial_{(--)}. \tag{48}$$

Hence, the expressions (39) and (40) for the gauge connections can be rewritten as follows:

$$\omega = d\xi^{(++)}\,\partial_{(++)}(W-L) - d\xi^{(--)}\,\partial_{(--)}(W+L), \tag{49}$$

$$A^{ij} = d\xi^{(++)}\,\partial_{(++)}G_{\mathcal{R}}^{ik}G_{\mathcal{R}}^{jk} + d\xi^{(--)}\,\partial_{(--)}G_{\mathcal{L}}^{ik}G_{\mathcal{L}}^{jk}. \tag{50}$$

Finally, using the holonomic basis, one can write down the general solution of the Peterson–Codazzi equations (43) for the covariant forms (33) as

$$f^{++i} = d\xi^{(--)}\,e^{W-L}G_{\mathcal{R}}^{ij}M_{(--)}^{(++)j}(\xi^{(--)}), \tag{51}$$

$$f^{--i} = d\xi^{(++)}\,e^{W+L}G_{\mathcal{L}}^{ij}M_{(++)}^{(--)j}(\xi^{(++)}), \tag{52}$$

where vector fields $M_{(--)}^{(++)j}, M_{(++)}^{(--)j}$ are chiral, similarly to the parameters $M_{(++)}^{(++)}, M_{(--)}^{(--)}$ of the solutions (44) of Eqs. (26):

$$\partial_{(++)}M_{(--)}^{(++)j} = 0, \quad \partial_{(--)}M_{(++)}^{(--)j} = 0.$$

Thus, following Ref. 5, we have solved algebraically all the equations except for the Gauss and Ricci ones (37) and (38). Substituting (49), (50), (51), and (52) into Eqs. (37) and (38), we obtain the set of nonlinear equations

$$\partial_{(++)}\partial_{(--)}W = \tfrac{1}{4}M_{(--)}^{(++)i}(G_{\mathcal{L}}^{T}G_{\mathcal{R}})^{ij}M_{(++)}^{(--)j}e^{2W}, \tag{53}$$

$$\begin{aligned}\partial_{(--)}((\partial_{(++)}G_{\mathcal{L}})G_{\mathcal{L}}^{T})^{ij} - \partial_{(++)}((\partial_{(--)}G_{\mathcal{R}})G_{\mathcal{R}}^{T})^{ij} + [(\partial_{(++)}G_{\mathcal{L}})G_{\mathcal{L}}^{T}, (\partial_{(--)}G_{\mathcal{R}})G_{\mathcal{R}}^{T}]^{ij}\\ = e^{2W}(G_{\mathcal{L}}M_{(--)}^{(++)})^{[i}(G_{\mathcal{R}}M_{(++)}^{(--)})^{j]}\end{aligned} \tag{54}$$

describing the extrinsic geometry of the string worldsheet embedded into a D-dimensional Minkowski space.

D. A zero curvature representation and the associated linear system

As we saw, most of the equations of the geometric approach (33), (26), (35), and (36) can be solved algebraically and the final set of the string-inspired nonlinear equations (53) and (54) emerges as the result of substitution of these algebraic solutions into the Gauss and Ricci equations (37) and (38). Since the latter constitute a part of the Maurer–Cartan equation (34), one can conclude that a zero curvature representation for the nonlinear equations (53) and (54) is given by the Maurer–Cartan equation (34) for the SO$(1,D-1)$ valued connection 1-forms specified by Eqs. (12), (49), (50), (51), and (52).

The associated linear system is provided by Eq. (11) with the 1-forms (34) specified by Eqs. (12), (49), (50), (51), and (52).

A nontrivial dependence on a spectral parameter can be introduced into the associated linear system by means of the parabolic subgroup transformations (17).[5]

Thus the nonlinear equations (53) and (54) possess all the features inherent in the equations which can be solved by the inverse scattering method.[12] Below we will prove that they are solvable even in a more strong sense, like the Liouville, Toda or WZNW sigma-model equations. Namely, we will deduce an explicit form of the general solution. It is interesting that the solution can be obtained by exploiting the parabolic group transformations (17). We will demonstrate this in the last section of this paper.

E. Symmetries and bridges

The obtained nonlinear equations and their zero curvature representation (34), (12), (49), (50), (51), and (52) possess a number of powerful symmetries.

As was already mentioned, the local (gauge) symmetries of the string model form the $SO(1,1)\times SO(D-2)$ group

$$\omega' = \omega + V\,dV^{-1} = \omega + \alpha, \tag{55}$$

$$A^{ij\prime} = (V^{-1}AV)^{ij} - (V^{-1}\,dV)^{ij}, \tag{56}$$

$$f^{++i\prime} = V^{-1} f^{++j} V^{ji}, \tag{57}$$

$$f^{--i\prime} = V f^{--j} V^{ji}. \tag{58}$$

The matrix fields $G_{\mathcal{L},\mathcal{R}}$ appearing in Eqs. (51), (52), and (50) are transformed homogeneously under the $SO(D-2)$ gauge symmetry

$$G^{ij\prime}_{\mathcal{L},\mathcal{R}} = G^{kj}_{\mathcal{L},\mathcal{R}} V^{ki} = (V^{-1})^{ik} G^{kj}_{\mathcal{L},\mathcal{R}}, \tag{59}$$

whereas the field L is pure gauge

$$L' = L - \alpha, \quad V = e^{-\alpha}. \tag{60}$$

In other words, it is a compensator (or Nambu–Goldstone field) for the SO(1,1) gauge symmetry.

Examining the expressions (44), (51), and (52), one concludes that our system possesses two types of infinite-dimensional global symmetries whose parameters can be combined into left-moving and right-moving (chiral) functions. One of them is the $d=2$ *conformal symmetry*

$$\begin{gathered} d\xi^{(++)\prime} = d\xi^{(++)}\, s_L(\xi^{(++)}), \quad d\xi^{(--)\prime} = d\xi^{(--)}\, s_R(\xi^{(--)}), \\ \partial_{(--)} s_L = 0, \quad \partial_{(++)} s_R = 0. \end{gathered} \tag{61}$$

The second one is realized as chiral rescalings of the chiral fields $M^{(++)}_{(++)}$ and $M^{(--)}_{(--)}$ present in Eqs. (51), and (52)

$$\begin{gathered} M^{(++)\prime}_{(++)} = M^{(++)}_{(++)} e^{h_L}, \quad M^{(--)\prime}_{(--)} = M^{(--)}_{(--)} e^{h_R}, \\ \partial_{(--)} h_L = 0, \quad \partial_{(++)} h_R = 0. \end{gathered} \tag{62}$$

Clearly, these rescalings can be treated as a sort of affine (or Kac–Moody) $SO(1,1)_L$ and $SO(1,1)_R$ symmetry transformations [i.e., as SO(1,1) transformations with the parameters depending on $\xi^{(++)}$ and $\xi^{(--)}$, respectively]. The full set of nontrivial transformations of these symmetries on the involved fields is given by

$$\begin{gathered} M^{(++)\prime}_{(++)} = M^{(++)}_{(++)} s_L^{-1} e^{h_L}, \quad M^{(--)\prime}_{(--)} = M^{(--)}_{(--)} s_R^{-1} e^{h_R}, \\ (W+L)' = (W+L) + h_L - \alpha, \quad (W-L)' = (W-L) + h_R + \alpha, \\ (M^{(--)i}_{(++)})' = M^{(--)i}_{(++)} s_L^{-1} e^{-h_L}, \quad (M^{(++)i}_{(--)})' = M^{(++)i}_{(--)} s_R^{-1} e^{-h_R}. \end{gathered} \tag{63}$$

We observe that the chiral fields $M^{(++)}_{(++)}, M^{(--)}_{(--)}$ can be regarded as the ''bridges'' relating the affine $SO(1,1)_L, SO(1,1)_R$ groups to the corresponding chiral parts of the two-dimensional conformal group, while

$$e^{W+L} = (e^{W+L})^{(++)}_{++}, \quad e^{W-L} = (e^{W-L})^{(--)}_{--}$$

as the bridges relating affine $SO(1,1)_L$ and $SO(1,1)_R$ to the gauge SO(1,1) symmetry. Since the symmetries (61) and (62) offer the possibility to choose the gauge [they also make it possible to fix the value of the norm of the chiral vector fields $M^{\pm\pm i}_{(\mp\mp)}$ (see Ref. 5)] $M^{(++)}_{(++)} = M^{(--)}_{(--)} = 1$, we, for simplicity, will make no distinction between the $SO(1,1)_{L,R}$ indices and conformal ones in what follows. We will also use the superscript $(--)$ instead of the $SO(1,1)_L$ subscript $(++)$ for the chiral vector field $M^{(--)i}_{(++)}$ (the chirality property of the latter field, $\partial_{(--)}M^{(--)i}_{(++)}=0$, excludes any confusion).

In addition, our equations possess an invariance under right multiplication of the $SO(D-2)$ valued fields $G_{\mathcal{L}}$ and $G_{\mathcal{R}}$ by chiral $SO(D-2)$ matrices H_R and H_L. So, the complete form of the appropriate symmetry transformations is

$$G'_{\mathcal{L}}=V^{-1}G_{\mathcal{L}}H_L, \quad G'_{\mathcal{R}}=V^{-1}G_{\mathcal{R}}H_R, \tag{64}$$

$$M^{(--)i\,\prime}_{(++)}=H^{ij}_L M^{(++)j}_{(--)}, \quad M^{(--)j\,\prime}_{(++)}=H^{ij}_R M^{(--)j}_{(++)}, \tag{65}$$

$$H_L H_L^T=H_R H_R^T=I, \quad \partial_{(--)}H_L=\partial_{(++)}H_R=0. \tag{66}$$

Thus the orthogonal matrix fields $G_{\mathcal{L}}$ and $G_{\mathcal{R}}$ can be regarded as bridges between the gauge $SO(D-2)$ transformations and affine chiral $SO(D-2)_L$ and $SO(D-2)_R$ transformations, respectively.

F. A simplified form of the nonlinear equations

The system of nonlinear equations (53) and (54) can be significantly simplified by using the $SO(D-2)$ gauge symmetry with parameters V^{ij} to fix the gauge

$$G_{\mathcal{L}}=1, \quad G_{\mathcal{R}}=G. \tag{67}$$

Then the sigma-model-type equation (54) acquires the simplest WZNW sigma-model-type form (1):

$$\partial_{(--)}((\partial_{(++)}G)G^T)^{ij}=e^{2W}G^{[i|k}M^{(++)k}_{(--)}M^{(--)|j]}_{(++)}, \tag{68}$$

whereas the Liouville-type equation (53) becomes form (2):

$$\partial_{(++)}\partial_{(--)}W=\tfrac{1}{4}M^{(--)i}_{(++)}G^{ij}M^{(++)j}_{(--)}e^{2W}. \tag{69}$$

The gauge (67) is invariant under the action of two chiral affine $SO(D-2)$ symmetries with the parameters $H_L(\xi^{(++)})$ and $H_R(\xi^{(--)})$. They act on the matrix field G as follows:

$$G'=H_L^{-1}GH_R, \quad \partial_{(--)}H_L=\partial_{(++)}H_R=0. \tag{70}$$

III. GENERAL SOLUTION OF THE STRING-INSPIRED NONLINEAR EQUATIONS

A. Standard string equations of motion, their solution and Virasoro constraints

We start by discussing the familiar string equations of motion and Virasoro conditions in the standard setting. The study of the relation between the solutions of these equation and the Lorentz harmonics (which, as was already mentioned, provide the associated linear system for the considered nonlinear equations) opens a possibility to construct the general solution of the nonlinear equations (69) and (68).

The equations of motion of the D-dimensional bosonic string following from the Nambu–Goto action (see Ref. 13 and references therein) has the form (32)

$$\partial_m(\sqrt{-g}g^{mn}\partial_n X^{\underline{m}})=0, \tag{71}$$

where [cf. Eq. (31)]

$$g_{mn}\equiv\partial_m X^{\underline{m}}\partial_n X_{\underline{m}} \tag{72}$$

is the induced metric, g^{mn} is its inverse and $g=\det(g_{mn})$.

In the conformal gauge (see, e.g., Ref. 13) the string equation (71) becomes linear

$$\tilde{\partial}_{(++)}\tilde{\partial}_{(--)}X^{\underline{m}}\equiv\frac{\partial}{\partial\tilde{\xi}^{(++)}}\frac{\partial}{\partial\tilde{\xi}^{(--)}}X^{\underline{m}}=0 \tag{73}$$

and has the following general solution

$$X^{\underline{m}}=X_L^{\underline{m}}+X_R^{\underline{m}}, \quad \tilde{\partial}_{(--)}X_L^{\underline{m}}=0, \quad \tilde{\partial}_{(++)}X_R^{\underline{m}}=0. \tag{74}$$

The chiral functions $X_L^{\underline{m}}(\xi^{(\tilde{+}+)}),X_R^{\underline{m}}(\xi^{(\tilde{-}-)})$ are subjected to the Virasoro constraints

$$\tilde{\partial}_{(++)}X_L^{\underline{m}}\tilde{\partial}_{(++)}X_{L\underline{m}}=0, \quad \tilde{\partial}_{(--)}X_R^{\underline{m}}\tilde{\partial}_{(--)}X_{R\underline{m}}=0. \tag{75}$$

Let us compare the solution (74) with the expressions (23) and (44) obtained in the geometric approach

$$dX^{\underline{m}}=\tfrac{1}{2}d\xi^{(++)}e^{-W-L}M_{(++)}^{(++)}u^{--\underline{m}}+\tfrac{1}{2}d\xi^{(--)}\,e^{-W+L}M_{(--)}^{(--)}u^{++\underline{m}}. \tag{76}$$

Since, due to Eq. (44), the induced metric is conformally flat in the coordinate frame $\xi^{(\pm\pm)}$, we can identify these worldsheet coordinates with the ''conformal coordinates'' $\tilde{\xi}^{(\pm\pm)}$ used in the standard string description (71)–(75)

$$\xi^{(\pm\pm)}=\tilde{\xi}^{(\pm\pm)}.$$

Thus Eqs. (76) and (74) result in

$$\begin{aligned}\partial_{(++)}X_L^{\underline{m}}(\xi^{(++)})&=\tfrac{1}{2}e^{-W-L}M_{(++)}^{(++)}(\xi^{(++)})u^{--\underline{m}},\\ \partial_{(--)}X_R^{\underline{m}}(\xi^{(--)})&=\tfrac{1}{2}e^{-W+L}M_{(--)}^{(--)}(\xi^{(--)})u^{++\underline{m}}.\end{aligned} \tag{77}$$

It is easy to verify that the Virasoro constraints (75) are satisfied for the functions (77).

It is worth noticing that the lhs of Eqs. (77) includes the chiral functions only [cf. (74)], while the rhs involves the functions $W,L,u^{\pm\pm}$ which from the very beginning were assumed to depend on both coordinates $\xi^{(\pm\pm)}$.

We will demonstrate below that the origin of this fact lies in that any solution of the string equation produces a solution of the string-inspired nonlinear equations, i.e., of the equations (53) and (68) which describe the extrinsic geometry of the string worldsheet.

B. Chiral harmonics

In Ref. 9 it was discussed how to find an appropriate set of chiral functions for obtaining explicit expressions for the fields W and $G^{ij}=(G^{-1})^{ji}$ which enter the nonlinear equations (53) and (68). Constrained twistors have been proposed as such variables for the case of bosonic string theories in dimensions $D=3,4,6$. Such twistors can be regarded as spinor Lorentz harmonics.[11,8] Their only property to be essential for our purposes is that they can be used to define the appropriate vector moving frame. This allows one to avoid complicated calculations associated with the use of the spinor moving frame or spinor harmonic formalism.[8] In this way, the solution of nonlinear equations describing the extrinsic geometry of bosonic string which moves in the Minkowski space of *arbitrary* dimension D can be obtained in terms of the vector Lorentz harmonics only.

Let us introduce two extra sets of Lorentz harmonics (moving frame variables)

$$
\begin{aligned}
r_{\underline{m}}^{(\underline{a})}(\xi^{(--)}) &= (r_{\underline{m}}^{(\pm\pm)}, r_{\underline{m}}^{i}) \in \mathrm{SO}(1,D-1),\\
l_{\underline{m}}^{(\underline{a})}(\xi^{(++)}) &= (l_{\underline{m}}^{(\pm\pm)}, l_{\underline{m}}^{i}) \in \mathrm{SO}(1,D-1),
\end{aligned}
\tag{78}
$$

each depending only on the $\xi^{(--)}$ or $\xi^{(++)}$ coordinates of the string worldsheet.

Recall that the condition (78) means

$$
r^{\underline{m}}{}_{(\underline{a})} r_{\underline{m}\,\underline{b}} = \eta_{(\underline{a})(\underline{b})} \equiv \mathrm{diag}(1,-1,\ldots,-1) \Leftrightarrow
\begin{cases}
r_{\underline{m}}^{(++)} r^{\underline{m}(++)} = 0, \quad r_{\underline{m}}^{(--)} r^{\underline{m}(--)} = 0,\\
r_{\underline{m}}^{(++)} r^{\underline{m}(--)} = 2,\\
r_{\underline{m}}^{(++)} r^{\underline{m}i} = 0, \quad r_{\underline{m}}^{(--)} r^{\underline{m}i} = 0,\\
r_{\underline{m}}^{i} r^{\underline{m}j} = -\delta^{ij}.
\end{cases}
\tag{79}
$$

$$
l^{\underline{m}}{}_{(\underline{a})} l_{\underline{m}\,\underline{b}} = \eta_{(\underline{a})(\underline{b})} \equiv \mathrm{diag}(1,-1,\ldots,-1) \Leftrightarrow
\begin{cases}
l_{\underline{m}}^{(++)} l^{\underline{m}(++)} = 0, \quad l_{\underline{m}}^{(--)} l^{\underline{m}(--)} = 0,\\
l_{\underline{m}}^{(++)} l^{\underline{m}(--)} = 2,\\
l_{\underline{m}}^{(++)} l^{\underline{m}i} = 0, \quad l_{\underline{m}}^{(--)} l^{\underline{m}i} = 0,\\
l_{\underline{m}}^{i} l^{\underline{m}j} = -\delta^{ij}.
\end{cases}
\tag{80}
$$

Further, we identify the chiral vectors $\partial_{(--)}X_R^{\underline{m}}$ and $\partial_{(++)}X_L^{\underline{m}}$ with the components $r^{(++)} = r^{(++)}(\xi^{(--)})$ and $l^{(--)} = l^{(--)}(\xi^{(++)})$ of these sets of chiral harmonics

$$
\partial_{(++)}X_L^{\underline{m}} = \tfrac{1}{2} M_{(++)}^{(++)}(\xi^{(++)}) l^{(--)\underline{m}}(\xi^{(++)}), \quad \partial_{(--)}X_R^{\underline{m}} = \tfrac{1}{2} M_{(--)}^{(--)}(\xi^{(--)}) r^{(++)\underline{m}}(\xi^{(--)}).
\tag{81}
$$

In such a way we adapt the chiral frames to the left and right sectors of the string worldsheet. Since other components of the left- and right-moving frame variables $l_{\underline{m}}^{(\underline{a})}, r_{\underline{m}}^{(\underline{a})}$ remain arbitrary, we face just a "particlelike" situation in the present case. Hence, Eqs. (81) possess the invariance under the affine

$$
(\mathrm{SO}(1,1) \otimes \mathrm{SO}(D-2) \circledS K_{D-2})_L \quad \text{and} \quad (\mathrm{SO}(1,1) \otimes \mathrm{SO}(D-2) \circledS K_{D-2})_R
$$

symmetries with chiral parameters

$$
V_L(\xi^{(++)}) = e^{h_L}, \quad V_R^{(++)i}(\xi^{(++)}), \quad V^{ij}(\xi^{(++)}) = V^{-1ji}(\xi^{(++)})
$$

and

$$
V_R(\xi^{(--)}) = e^{h_R}, \quad V_R^{(--)i}(\xi^{(--)}), \quad V^{ij}(\xi^{(--)}) = V^{-1ji}(\xi^{(--)}),
$$

respectively. [The affine symmetry $\mathrm{SO}(1,1)_L \otimes \mathrm{SO}(1,1)_R$ proves to be "soldered" to the worldsheet conformal symmetry when the gauge $M_{(++)}^{(++)} = 1 = M_{(--)}^{(--)}$ is imposed.] The corresponding transformations read

$$
\begin{aligned}
r_{\underline{m}}^{(++)\prime} &= r_{\underline{m}}^{(++)} V_R,\\
r_{\underline{m}}^{i\prime} &= (r_{\underline{m}}^{j} + r_{\underline{m}}^{(++)} V_R^{(--)j}) V_R^{ji},\\
r_{\underline{m}}^{(--)\prime} &= (r_{\underline{m}}^{(--)} + r_{\underline{m}}^{(++)} V_R^{(--)i} V_R^{(--)i} + 2 r_{\underline{m}}^{i} V_R^{--i}) V_R^{-1},
\end{aligned}
\tag{82}
$$

$$
\begin{aligned}
l_{\underline{m}}^{(--)\prime} &= l_{\underline{m}}^{(--)} V_L,\\
l_{\underline{m}}^{i\prime} &= (l_{\underline{m}}^{j} + l_{\underline{m}}^{(--)} V_L^{(++)j}) V_L^{ji},\\
l_{\underline{m}}^{(++)\prime} &= (l_{\underline{m}}^{(++)} + l_{\underline{m}}^{(--)} V_L^{(++)i} V_L^{(++)i} + l_{\underline{m}}^{i} V_L^{(++)i}) V_L^{-1}.
\end{aligned}
\tag{83}
$$

Hence, each set of chiral harmonics parametrizes the sphere (18). As they depend only on one of the worldsheet coordinates, $\xi^{(--)}$ or $\xi^{(++)}$, they map one of the lightlike sectors, $\mathcal{M}^{(0,1)}=\{\xi^{(--)}\}$ or $\mathcal{M}^{(1,0)}=\{\xi^{(++)}\}$, of the worldsheet $\mathcal{M}^{(1,1)}=\{\xi^m\}=\{\xi^{(++)},\xi^{(--)}\}$ onto two copies of this sphere

$$r_{\underline{m}}^{(\underline{a})}:\quad \mathcal{M}^{(0,1)}=\{\xi^{(--)}\}\rightarrow S^{D-2}=\frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times \mathrm{SO}(D-2)\otimes K_{D-2}}, \tag{84}$$

$$l_{\underline{m}}^{(\underline{a})}:\quad \mathcal{M}^{(1,0)}=\{\xi^{(++)}\}\rightarrow S^{D-2}=\frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times \mathrm{SO}(D-2)\otimes K_{D-2}}. \tag{85}$$

Below we denote the spaces of all possible images of these maps by $S_L^{(D-2)}$ and $S_R^{(D-2)}$, respectively.

The chiral counterparts of the Cartan forms (12) contain only one of the chiral holonomic basic 1-forms $d\xi^{(--)}$ or $d\xi^{(++)}$

$$f_R^{(++)i}=d\xi^{(--)}\, f_{(--)R}^{(++)i}=r^{(++)\underline{m}}\, dr_{\underline{m}}^{i}, \quad f_{(--)R}^{(++)i}=r^{(++)\underline{m}}\partial_{(--)}r_{\underline{m}}^{i}, \tag{86}$$

$$f_R^{(--)i}=d\xi^{(--)}\, f_{(--)R}^{(--)i}=r^{(--)\underline{m}}\, dr_{\underline{m}}^{i}, \quad f_{(--)R}^{(--)i}=r^{(--)\underline{m}}\partial_{(--)}r_{\underline{m}}^{i}, \tag{87}$$

$$\omega_R=d\xi^{(--)}\,\omega_{(--)R}=\tfrac{1}{2}r^{(--)\underline{m}}\, dr_{\underline{m}}^{(++)}, \quad \omega_{(--)R}=\tfrac{1}{2}r^{(--)\underline{m}}\partial_{(--)}r_{\underline{m}}^{(++)}, \tag{88}$$

$$A_R^{ij}=d\xi^{(--)}\, A_{(--)R}^{ij}=\tfrac{1}{2}r^{i\underline{m}}\, dr_{\underline{m}}^{j}, \quad A_{(--)R}^{ij}=\tfrac{1}{2}r^{i\underline{m}}\partial_{(--)}r_{\underline{m}}^{j}, \tag{89}$$

$$f_L^{(++)i}=d\xi^{(++)}\, f_{(++)L}^{(++)i}=l^{(++)\underline{m}}dl_{\underline{m}}^{i}, \quad f_{(++)L}^{(++)i}=l^{(++)\underline{m}}\partial_{(++)}l_{\underline{m}}^{i}, \tag{90}$$

$$f_L^{(--)i}=d\xi^{(++)}\, f_{(++)L}^{(--)i}=l^{(--)\underline{m}}dl_{\underline{m}}^{i}, \quad f_{(++)L}^{(--)i}=l^{(--)\underline{m}}\partial_{(++)}l_{\underline{m}}^{i}, \tag{91}$$

$$\omega_L=d\xi^{(++)}\,\omega_{(++)L}=\tfrac{1}{2}l^{(--)\underline{m}}\, dl_{\underline{m}}^{(++)}, \quad \omega_{(++)L}=\tfrac{1}{2}l^{(--)\underline{m}}\partial_{(++)}l_{\underline{m}}^{(++)}, \tag{92}$$

$$A_L^{ij}=d\xi^{(++)}\, A_{(++)L}^{ij}=\tfrac{1}{2}l^{i\underline{m}}\, dl_{\underline{m}}^{j}, \quad A_{(++)L}^{ij}=\tfrac{1}{2}l^{i\underline{m}}\partial_{(++)}l_{\underline{m}}^{j}. \tag{93}$$

The transformations of these 1-forms under the left and right affine $\mathrm{SO}(1,1)\otimes\mathrm{SO}(D-2)\subset\times K_{D-2}$ symmetries (82) and (83) are determined by the chiral version of Eqs. (19)–(22) and its evident ''left'' counterpart, respectively. It is worth noting that only the forms

$$f_R^{(++)i}=d\xi^{(--)}\, f_{(--)R}^{(++)i} \quad \text{and} \quad f_R^{(--)i}=d\xi^{(++)}\, f_{(++)L}^{(--)i}$$

transform covariantly under (82) and (83). These forms are vielbeins of the ''chiral spheres'' S_R^{D-2} (84) and S_R^{D-2} (85), respectively.

C. Relation of general and chiral harmonics: Solving the Liouville-type equation

Substituting Eq. (81) into Eq. (77), one obtains the expression for the chiral lightlike moving frame vector fields $l^{(--)\underline{m}}(\xi^{(++)})$, $r^{(++)\underline{m}}(\xi^{(--)})$ in terms of generic lightlike harmonics $u^{--\underline{m}},u^{++\underline{m}}$, the Liouville field W and compensator L

$$l^{(--)\underline{m}}(\xi^{(++)})=e^{-W-L}u^{--\underline{m}}, \quad r^{(++)\underline{m}}(\xi^{(--)})=e^{-W+L}u^{++\underline{m}}. \tag{94}$$

Contracting both sides of these two equations in indices $\underline{m}$, we get the expression for the field W in terms of chiral harmonics

$$e^{-2W}=\tfrac{1}{2}l^{(--)\underline{m}}r_{\underline{m}}^{(++)}. \tag{95}$$

For $D=3$, Eq. (95) produces the general solution of the Liouville equation in a special parametrization (see the Appendix).

A similar solution of the Liouville equation in the conformal gauge was presented in Ref. 9. The Cartan–Penrose representation in terms of bosonic spinors was used there for chiral lightlike vectors $l^{(--)\underline{m}}$ and $r^{(++)}_{\underline{m}}$.

In the generic case of higher D it is necessary to have the suitable representation for the $SO(D-2)$ matrices G^{ij} as well.

D. Relation of general and chiral harmonics: Solving the sigma-model-type equation

To obtain the expression for $SO(D-2)$ matrix field G, let us analyze the consequences of Eq. (94) for the moving frame vectors u^i. First of all, one finds

$$l^{(--)\underline{m}}u^i_{\underline{m}}=0, \quad r^{(++)\underline{m}}u^i_{\underline{m}}=0. \tag{96}$$

Equations (96) mean that the decompositions of the u^i harmonic over the chiral left- and right-moving ones involve no terms proportional to $l^{(++)}$ and $r^{(--)}$, respectively

$$u^i_{\underline{m}}=-(l^j_{\underline{m}}-V^j_{(--)}l^{(--)}_{\underline{m}})U^{ji}_{\mathcal{L}}, \tag{97}$$

$$u^i_{\underline{m}}=-(r^j_{\underline{m}}+V^i_{(++)}r^{(++)}_{\underline{m}})U^{ji}_{\mathcal{R}}. \tag{98}$$

The newly introduced matrices $U^{ji}_{\mathcal{L}}$ and $U^{ji}_{\mathcal{L}}$ are expressed through the contractions of chiral harmonics with the generic ones:

$$U^{ji}_{\mathcal{L}}=l^{j\underline{m}}u^i_{\underline{m}}, \quad U^{ji}_{\mathcal{R}}=r^{j\underline{m}}u^i_{\underline{m}}. \tag{99}$$

So they can easily be checked to be orthogonal matrices

$$U^{jk}_{\mathcal{L}}U^{ik}_{\mathcal{L}}=\delta^{ji}, \quad U^{jk}_{\mathcal{L}}U^{ik}_{\mathcal{L}}=\delta^{ji}. \tag{100}$$

The proof uses the unity decomposition (7) and the corollaries of Eqs. (94)

$$u^{--\underline{m}}l^i_{\underline{m}}=0, \quad u^{++\underline{m}}r^i_{\underline{m}}=0. \tag{101}$$

To find the relations between $U_{\mathcal{L},\mathcal{R}}$ and $U_{\mathcal{R},\mathcal{L}}$, one should consider the derivatives of the $U_{\mathcal{L},\mathcal{R}}$ fields

$$\partial_{(--)}U^{ji}_{\mathcal{L}}=l^{j\underline{m}}\partial_{(--)}u^i_{\underline{m}} \quad \text{and} \quad \partial_{(++)}U^{ji}_{\mathcal{R}}=r^{j\underline{m}}\partial_{(++)}u^i_{\underline{m}}$$

and use the decomposition (11) to express the derivatives of the generic harmonics in terms of components of the Cartan forms (51)–(50). In such a way one arrives at the relations

$$(\partial_{(--)}U^{kj}_{\mathcal{L}})U^{ki}_{\mathcal{L}}=(\partial_{(--)}G^{jk}_{\mathcal{R}})G^{ik}_{\mathcal{R}}, \quad (\partial_{(++)}U^{kj}_{\mathcal{R}})U^{ki}_{\mathcal{R}}=(\partial_{(++)}G^{jk}_{\mathcal{L}})G^{ik}_{\mathcal{L}}. \tag{102}$$

Equations (102) mean that the field $G_{\mathcal{L}}(G_{\mathcal{R}})$ differs from the (transposed) matrix field $U_{\mathcal{R}}(U_{\mathcal{L}})$ by an affine $SO(D-2)_L(SO(D-2)_R)$ transformation only [cf. (70)]

$$G^{ij}_{\mathcal{R}}=H^{jk}_L(\xi^{(++)})U^{ki}_{\mathcal{L}}=u^{i\underline{m}}l^k_{\underline{m}}H^{kj}_L(\xi^{(++)}), \tag{103}$$

$$G^{ij}_{\mathcal{L}}=H^{jk}_R(\xi^{(--)})U^{ki}_{\mathcal{R}}=u^{i\underline{m}}r^k_{\underline{m}}H^{jk}_R(\xi^{(--)}),$$

$$H_LH_L^T=I, \quad \partial_{(--)}H_L=0, \quad H_RH_R^T=I, \quad \partial_{(++)}H_R=0. \tag{104}$$

The expression for $V^i_{(++)}$ follows from the first equation in (96) upon substituting (98) and using Eqs. (94), (103), (95). In this way one gets

$$V^i_{(++)}=\tfrac{1}{2}e^{2W}l^{(--)\underline{m}}r^i_{\underline{m}}. \tag{105}$$

In the same manner one can obtain

$$V^i_{(--)}=\tfrac{1}{2}e^{2W}r^{(++)\underline{m}}l^i_{\underline{m}} \tag{106}$$

from Eq. (97) and the second of Eqs. (96).

Now we can rewrite Eqs. (97) and (98) in terms of chiral harmonics and the functions present in the nonlinear equations (53) and (54)

$$u^i_{\underline{m}}=G^{ik}_{\mathcal{R}}H^{kj}_L(-l^j_{\underline{m}}+e^{2W}r^{(++)\underline{n}}l^j_{\underline{n}}l^{(--)}_{\underline{m}}), \tag{107}$$

$$u^i_{\underline{m}}=G^{ik}_{\mathcal{L}}H^{kj}_R(-r^j_{\underline{m}}+e^{2W}l^{(--)\underline{n}}r^j_{\underline{n}}r^{(++)}_{\underline{m}}). \tag{108}$$

From Eqs. (107) and (108) we obtain the expression

$$(G^T_{\mathcal{L}}G_{\mathcal{R}})^{ij}=H^{ik}_L\left(-l^{k\underline{m}}r^l_{\underline{m}}+\frac{2(r^{(++)\underline{n}}l^k_{\underline{n}})(l^{(--)\underline{n}'}r^l_{\underline{n}'})}{(l^{(--)\underline{n}''}r^{(++)}_{\underline{n}''})}\right)H^{jl}_R, \tag{109}$$

which provides the general solution for the sigma-model-like equation (1) in the gauge (67)

$$G^{ij}_{\mathcal{R}}=\delta^{ij}, \quad G^{ij}_{\mathcal{L}}\equiv G^{ij},$$

$$G^{ij}=H^{ik}_L\left(-l^{k\underline{m}}r^l_{\underline{m}}+\frac{2(r^{(++)\underline{n}}l^k_{\underline{n}})(l^{(--)\underline{n}'}r^l_{\underline{n}'})}{(l^{(--)\underline{n}''}r^{(++)}_{\underline{n}''})}\right)H^{jl}_R. \tag{110}$$

To complete the description of the general solution of the string-inspired system of nonlinear equations (2) and (1), we have to present the expressions for the chiral vector fields $M^{(\mp\mp)i}_{(\pm\pm)}$ (3) in terms of the chiral harmonics. This can be easily done by applying the derivatives $\partial_{(++)}$ and $\partial_{(--)}$ to both sides of Eqs. (107) and (108) and contracting the results with the vectors $l^{(++)}$ and $r^{(--)}$, respectively. Then, using Eqs. (51) and (52), one obtains

$$M^{(++)i}_{(--)}=-H^{ij}_Rr^{(++)\underline{m}}\partial_{(--)}r^j_{\underline{m}}, \tag{111}$$

$$M^{(--)i}_{(++)}=-H^{ij}_Ll^{(--)\underline{m}}\partial_{(++)}l^j_{\underline{m}}. \tag{112}$$

Thus the chiral vectors $M^{(++)i}_{(--)}$ and $M^{(--)i}_{(++)}$ appearing in Eqs. (2) and (1) coincide with the covariant components $f^{(\mp\mp)i}_{(\pm\pm)}$ (86) and (91) of the chiral Cartan forms which constitute a basis on the ‘‘chiral spheres’’ S^{D-2}_R and S^{D-2}_L, respectively.

1. General solution of the string-inspired nonlinear equations

Equations (95), (110), (111), and (112) provide the *general solution* of the system of nonlinear equations (2), (1), and (3).

To be convinced of this, one has to take into account the following.

(i) When obtaining the expressions (95), (110), (111), and (112) for all the functions which enter Eqs. (2), (1), and (3), we started from the general solution of the bosonic string equations of motion in the standard Nambu–Goto approach and then used these solutions in the equations of the geometric approach.

(ii) As we demonstrated in Sec. I, the equations of the geometric approach describing the extrinsic geometry of the D-dimensional bosonic string worldsheet uniquely produce the system of nonlinear equations (2), (1), and (3).

(iii) The equations of geometric approach and, therefore, Eqs. (2), (1), and (3) specify the bosonic string worldsheet uniquely (up to symmetry transformations, see, e.g., Ref. 4), and thus describe exactly the same dynamical system as the ordinary (linear) string equations of motion.

(iv) The general solution (95), (110), (111), and (112) is written in terms of two sets of chiral spinor harmonics (78), which parameterize two copies of the compact coset (18)

$$S^{D-2}=\frac{SO(1,D-1)}{SO(1,1)\times SO(D-2)\circledS K_{D-2}}$$

isomorphic to the sphere S^{D-2}. Thus it contains $(D-2)$ right-moving and $(D-2)$ left-moving degrees of freedom, that is the same as the number of independent degrees of freedom of the general solution (74) of the standard string equations of motion (71).

With this reasoning in mind, we conclude that the expressions (95), (110), (111), and (112) for the functions W, G and $M_{(++)}^{(--)i}, M_{(--)}^{(++)i}$ obtained from the general solution of the standard string equations *have to provide the general solution* of the geometric approach equations (2), (1), and (3) (up to superfluous symmetry transformations).

IV. ON THE GROUP THEORETICAL AND GEOMETRICAL STRUCTURE OF THE SOLUTION

In the course of deriving the general solution (95), (110), (111), and (112) we have the expressions for the moving frame vectors (4) in terms of chiral harmonics

$$u_{\underline{m}}^{++}=e^{W-L}r_{\underline{m}}^{(++)},$$

$$u_{\underline{m}}^{i}=-(G_{\mathcal{L}}H_R)^{ij}(r_{\underline{m}}^{j}-V_{(++)}{}^{j}r_{\underline{m}}^{(++)}),$$

$$u_{\underline{m}}^{--}=e^{-(W-L)}(r_{\underline{m}}^{(--)}+r_{\underline{m}}^{(++)}V_{(++)}{}^{j}V_{(++)}{}^{j}-2r_{\underline{m}}^{i}V_{(++)}{}^{i}), \tag{113}$$

$$u_{\underline{m}}^{++}=e^{-(W+L)}(l_{\underline{m}}^{(++)}+l_{\underline{m}}^{(--)}V_{(--)}{}^{j}V_{(--)}{}^{j}-2l_{\underline{m}}^{i}V_{(--)}{}^{i}),$$

$$u_{\underline{m}}^{i}=-(G_{\mathcal{R}}H_L)^{ij}(l_{\underline{m}}^{j}-V_{(--)}{}^{j}l_{\underline{m}}^{(--)}),$$

$$u_{\underline{m}}^{--}=e^{W+L}l_{\underline{m}}^{(--)} \tag{114}$$

[more precisely, we have the first two equations in each set (113) and (114) while the third one can be restored from the orthonormality conditions (5), (79), and (80)].

If the functions W and $G_{\mathcal{L},\mathcal{R}}$ satisfy the nonlinear equations (2) and (1), then Eqs. (113) and (114) can be regarded as the solution of the corresponding associated linear system defined by Eqs. (11), with the Cartan forms (12) being specified by Eqs. (51), (52), (49), and (50).

The solution of the zero curvature representation given by the Maurer-Cartan equations (34) with the Cartan forms (12) can be obtained by differentiating (113) and (114). The solution is given by the following expressions for the generic Cartan 1-forms (13)–(16) in terms of chiral ones (86)–(89) and (90)–(93):

$$f^{++i}=e^{W-L}(G_{\mathcal{L}}H_R)^{ij}f_R^{(++)j}, \tag{115}$$

$$f^{--i}=-e^{-(W-L)}(G_{\mathcal{L}}H_R)^{ij}(f_R^{(--)j}-2\mathcal{D}_RV_{(++)}{}^{j}-2f_R^{(++)k}(V_{(++)}{}^{k}V_{(++)}{}^{j}-\tfrac{1}{2}\delta^{kj}V_{(++)}^{l}V_{(++)}{}^{l})),$$

$$\mathcal{D}_RV_{(++)}{}^{j}\equiv dV_{(++)}{}^{j}+\omega_RV_{(++)}{}^{j}-V_{(++)}{}^{k}A_R^{kj}, \tag{116}$$

$$\omega=\omega_R+f^{(++)i}V_{(++)}{}^{i}+d(W-L), \tag{117}$$

$$A^{ij}=(G_{\mathcal{L}}H_R)^{ik}(G_{\mathcal{L}}H_R)^{jl}(A_R^{kl}+(G_{\mathcal{L}}H_R)^{-1}d(G_{\mathcal{L}}H_R))^{kl}+2\,f^{(++)[k}V^{--l]}), \tag{118}$$

$$f^{--i}=e^{W+L}(G_{\mathcal{R}}H_L)^{ij}f_L^{(--)j}, \tag{119}$$

$$f^{++i}=-e^{-(W-L)}(G_{\mathcal{R}}H_L)^{ij}(f_L^{(++)j}-2\mathcal{D}_R V_{(--)}{}^j$$
$$-2\,f_L^{(--)k}(V_{(--)}{}^k V_{(--)}{}^j-\tfrac{1}{2}\delta^{kj}V_{(--)}{}^l V_{(--)}{}^l)),$$

$$\mathcal{D}_L V_{(--)}{}^j\equiv dV_{(--)}{}^j+\omega_R V_{(--)}{}^j-V_{(--)}{}^k A_L^{kj}, \tag{120}$$

$$\omega=\omega_R+f^{(--)i}V_{(--)}{}^i-d(W+L), \tag{121}$$

$$A^{ij}=(G_{\mathcal{L}}H_R)^{ik}(G_{\mathcal{L}}H_R)^{jl}(A_R^{kl}+(G_{\mathcal{L}}H_R)^{-1}d(G_{\mathcal{L}}H_R))^{kl}+2\,f^{++[k}V^{--l]}). \tag{122}$$

On the other hand, as follows from the consideration in the preceding section, the explicit form of the general solution (95) and (110), as well as the expressions (105) and (106) for the ''boost'' parameters, can be obtained algebraically from Eqs. (114) and (113) [with making use of the orthonormality constraints (5), (79), and (80) for the generic and chiral moving frame harmonics].

An intriguing point is that Eqs. (114) and (113) generating the general solution *have the form of the parabolic symmetry transformations* (83) and (82) of the chiral harmonics, but with nonchiral parameters.

Thus the prescription of how to solve the nonlinear equations (1) and (2) can be formulated as follows.

Let us introduce the two sets of chiral harmonics (78) and (79), which map the right (left) light-cone sectors of the worldsheet $\mathcal{M}^{(0,1)}\equiv\{(\xi^{(--)})\}$ $(\mathcal{M}^{(1,0)}\equiv\{(\xi^{(++)})\})$ onto the sphere S^{D-2},

$$r_{\underline{m}}^{(\underline{a})}:\ \ \mathcal{M}^{(0,1)}\equiv\{(\xi^{(--)})\}\rightarrow S^{D-2}=\frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times\mathrm{SO}(D-2)\circledS K_{D-2}},$$

$$r_{\underline{m}}^{(\underline{a})}:\ \ \mathcal{M}^{(1,0)}\equiv\{(\xi^{(++)})\}\rightarrow S^{D-2}=\frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times\mathrm{SO}(D-2)\circledS K_{D-2}}.$$

Further, let us assume that the generic harmonics (4) and (5)

$$u_{\underline{m}}^{\underline{a}}:\ \ \mathcal{M}^{(1,1)}\rightarrow\frac{\mathrm{SO}(1,D-1)}{\mathrm{SO}(1,1)\times\mathrm{SO}(D-2)}$$

are related to the chiral ones by the parabolic transformations (113) and (114) (with the chiral parameters H_L and H_R omitted for simplicity).

Then let us exploit the $\mathrm{SO}(D-2)$ gauge freedom to fix the $\mathrm{SO}(D-2)$ rotation matrix in (113) to be the unity one. The $\mathrm{SO}(D-2)$ rotation matrix in (114) taken in this gauge provides us with the solution (110) of the WZNW sigma-model-type equation (1) [with the chiral vectors $M_{(--)}^{(++)i}, M_{(--)}^{(++)i}$ determined by the homogeneously transforming components of chiral Cartan forms (86) and (91)

$$M_{(--)}^{(++)i}=f_{(--)R}^{(++)i},\quad M_{(--)}^{(++)i}=f_{(++)L}^{(--)i}$$

[cf. (112) and (111)]. The product of the SO(1,1) transformation factors from (113) and (114) produces the general solution of the Liouville-like equation (2).

Since the parabolic transformations (113) and (114) have a ''triangular'' form, one can expect that the described method of obtaining the general solution bears a tight relation to the known group-theoretical methods of solving nonlinear equations, like those developed and used in Refs.

14, 15, and 16. A detailed examination of such a relationship could provide a deeper insight into the nature of integrability and we consider it as an interesting problem for future study.

V. CONCLUSION

We have obtained the general solution of the string-inspired nonlinear equations (1), (2), and (3) describing the extrinsic geometry of the bosonic string worldsheet in the geometrical approach.[1,2,3,5]

The solution is given in terms of the two sets of chiral (left-moving and right-moving) Lorentz harmonic variables (78), (79), and (80) and has the form (after fixing a gauge with respect to some extra symmetries)

$$e^{-2W}=\tfrac{1}{2}r^{(++)}_{\underline{m}}l^{(--)\underline{m}}, \tag{123}$$

$$G^{kj}=-l^{k}_{\underline{m}}r^{j\underline{m}}+\frac{r^{(++)}_{\underline{m}}l^{k\underline{m}}l^{(--)}_{\underline{n}}r^{j\underline{n}}}{r^{(++)}_{\underline{p}}l^{(--)\underline{p}}}, \tag{124}$$

$$M^{(++)i}_{(--)}(\xi^{(--)})=r^{(++)\underline{m}}\partial_{(--)}r^{i}_{\underline{m}}, \tag{125}$$

$$M^{(--)i}_{(++)}(\xi^{(++)})=l^{(--)\underline{m}}\partial_{(++)}l^{i}_{\underline{m}}. \tag{126}$$

The analysis of the solution of the associated linear system demonstrates that the general solution we have found can be regarded as the parabolic subgroup $SO(1,1)\times SO(D-2)\otimes K_{D-2}$ transformations[11] of the chiral harmonics, the transformation parameters being nonchiral. As such transformations have the ''triangular'' form in the matrix representation, we can expect a close relation of our approach to the known group-theoretical methods of solving nonlinear equations.[14–16] It is an interesting task for further study to elaborate on the detailed form of such a relation.

A natural direction of extending our results is to look for the solution of a supersymmetric generalization of the considered nonlinear equations. Such a system describes the extrinsic geometry of the worldsheet superspace of $D=3,4,6,10$ superstring models. (Let us note that the explicit form of such supersymmetric equations are known at present only for the cases $D=3$, $N=1,2$.[17])

ACKNOWLEDGMENTS

The authors are grateful to D. Sorokin, S. Krivonos, B. Julia, and F. Toppan for interest in this work and useful remarks. One of the authors (I.B.) would like to thank Professor M. Virasoro and Professor S. Randjbar-Daemi for the hospitality at the Abdus Salam ICTP where part of this work was done. The work was supported in part by the INTAS Grants Nos. 96-0308, 96-0538, 93-127 ext the RFBR Grant No. 96-02-17634 and by the Fonds zur Förderung der Wissenschaftlichen Forschung under Project No. M472-TPH (I.B.).

APPENDIX: RELATION TO THE STANDARD FORM OF GENERAL SOLUTION OF THE LIOUVILLE EQUATION

Here we demonstrate that for the $D=3$ case Eq. (95) reproduces the general solution of the nonlinear Liouville equation. Using the well-known parametrization of the SO(1,2) matrices

$$l^{m}_{(\underline{a})}=\begin{pmatrix}\cosh A_L & \sinh A_L & 0\\ \sinh A_L & \cosh A_L & 0\\ 0 & 0 & 1\end{pmatrix},$$

$$r^{\underline{m}}_{(\underline{a})}=\begin{pmatrix}\cosh A_R & -\sinh A_R & 0\\ -\sinh A_R & \cosh A_R & 0\\ 0 & 0 & 1\end{pmatrix},$$

$$\partial_{(--)}A_L=\partial_{(++)}A_R=0, \tag{A1}$$

one finds the following form of chiral harmonics (78)

$$l^{(--)\underline{m}}=(\cosh A_L, \quad \sinh A_L, \quad 1),$$

$$l^{\perp\,\underline{m}}=(\sinh A_L, \quad \cosh A_L, \quad 0), \tag{A2}$$

$$l^{(++)\underline{m}}=(\cosh A_L, \quad \sinh A_L, \quad -1),$$

$$r^{(--)\underline{m}}=(\cosh A_R, \quad -\sinh A_R, \quad 1),$$

$$r^{\perp\,\underline{m}}=(-\sinh A_R, \quad \cosh A_R, \quad 0),$$

$$r^{(++)\underline{m}}=(\cosh A_R, \quad -\sinh A_R, \quad -1). \tag{A3}$$

Substituting these expressions into the Eqs. (95), (111), and (112) with $D=3$, one gets

$$e^{-W}=\cosh\frac{A_L+A_R}{2}, \tag{A4}$$

$$M^{(++)\perp}_{(--)}(\xi^{(--)})=r^{(++)\underline{m}}\partial_{(--)}r^{\perp}_{\underline{m}}=-\partial_{(--)}A_R\,, \tag{A5}$$

$$M^{(--)\perp}_{(++)}(\xi^{(++)})=l^{(--)\underline{m}}\partial_{(++)}l^{\perp}_{\underline{m}}=\partial_{(++)}A_L\,. \tag{A6}$$

The relation to the standard parametrization of the general solution of the Liouville equation (see, e.g., Ref. 18) is given by

$$f_{L,R}=\exp A_{L,R}\,.$$

[1] F. Lund and T. Regge, Phys. Rev. D **14**, 1524 (1976); R. Omnes, Nucl. Phys. B **149**, 269 (1979); B. M. Barbashov and V. V. Nesterenko, Commun. Math. Phys. **78**, 499 (1981). A. Zheltukhin, Yad. Fiz. **33**, 1723 (1981); Theor. Math. Phys. **52**, 666 (1982); Phys. Lett. B **116**, 147 (1982); Theor. Math. Phys. **56**, 785 (1984).
[2] B. M. Barbashov and V. V. Nesterenko, *Introduction to the Relativistic String Theory* (World Scientific, Singapore, 1990), and references therein.
[3] I. A. Bandos, P. Pasti, D. Sorokin, M. Tonin, and D. V. Volkov, Nucl. Phys. B **446**, 79 (1995).
[4] L. P. Eisenhart, *Riemannian Geometry* (Princeton Univ. Press, Princeton, NJ, 1949).
[5] I. A. Bandos, Phys. Lett. B **388**, 35 (1996).
[6] A. Galperin, E. Ivanov, S. Kalitzin, V. Ogievetsky, and E. Sokatchev, Class. Quantum Grav. **1**, 498 (1984); **2**, 155 (1985).
[7] E. Sokatchev, Phys. Lett. B **169**, 209 (1987); Class. Quantum Grav. **4**, 237 (1987).
[8] I. A. Bandos and A. A. Zheltukhin, JETP Lett. **54**, 421 (1991); Phys. Lett. B **288**, 77 (1992); Sov. J. Nucl. Phys. **56**, 198 (1993); Phys. Part. Nuclei **25**, 453 (1994); Class. Quantum Grav. **12**, 609 (1995).
[9] A. A. Kapustnikov and S. A. Ulanov, Constrained twistors principle of the string theory, hep-th/9704068.
[10] I. A. Bandos, Sov. J. Nucl. Phys. **51**, 906 (1990); JETP Lett. **52**, 205 (1990).
[11] A. Galperin, P. Howe, and K. Stelle, Nucl. Phys. B **368**, 248 (1992); A. Galperin, F. Delduc, and E. Sokatchev, *ibid.* **368**, 143 (1992); A. Galperin, K. Stelle, and P. Townsend, *ibid.* **402**, 531 (1993).
[12] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Approach in Theory of Solitons* (Springer, Berlin, 1987).
[13] M. Green, J. Schwarz, and E. Witten, *Superstring theory*, Vols. 1,2 (Cambridge Univ. Press, Cambridge, 1987).

[14] A. N. Leznov and M. V. Saveliev, Commun. Math. Phys. **74**, 111 (1980); A. N. Leznov, Theor. Math. Phys. **73**, 1233 (1987); A. N. Leznov and M. A. Mukhtarov, J. Math. Phys. **28**, 2574 (1987); Ch. Devchand and A. N. Leznov, Commun. Math. Phys. **160**, 551 (1994), and references therein.
[15] E. A. Ivanov and S. O. Krivonos, Theor. Math. Phys. **58**, 131 (1984); **66**, 60 (1984); E. A. Ivanov, S. O. Krivonos, and V. M. Leviant, Nucl. Phys. B **304**, 601 (1988).
[16] E. A. Ivanov, Theor. Math. Phys. **71**, 193 (1987); $N=4$ supersymmetric integrable systems, hep-th/9704133 and refs. therein.
[17] I. Bandos, D. Sorokin, and D. V. Volkov, Phys. Lett. B **372**, 77 (1996).
[18] M. Chaichian and P. P. Kulish, Phys. Lett. B **78**, 413 (1978).

A new perturbative technique for solving integro-partial differential equations

Peter A. Becker[a)]
Center for Earth Observing and Space Research, Institute for Computational Sciences and Informatics, and Department of Physics and Astronomy, George Mason University, Fairfax, Virginia 22030-4444

(Received 10 May 1999; accepted for publication 8 July 1999)

Integro-partial differential equations occur in many contexts in mathematical physics. Typical examples include time-dependent diffusion equations containing a parameter (e.g., the temperature) that depends on integrals of the unknown distribution function. The standard approach to solving the resulting nonlinear partial differential equation involves the use of predictor–corrector algorithms, which often require many iterations to achieve an acceptable level of convergence. In this paper we present an alternative procedure that allows us to separate a family of integro-partial differential equations into two related problems, namely (i) a perturbation equation for the temperature, and (ii) a linear partial differential equation for the distribution function. We demonstrate that the variation of the temperature can be determined by solving the perturbation equation *before* solving for the distribution function. Convergent results for the temperature are obtained by recasting the divergent perturbation expansion as a continued fraction. Once the temperature variation is determined, the self-consistent solution for the distribution function is obtained by solving the remaining, linear partial differential equation using standard techniques. The validity of the approach is confirmed by comparing the (input) continued-fraction temperature profile with the (output) temperature computed by integrating the resulting distribution function. © *1999 American Institute of Physics.* [S0022-2488(99)03410-6]

I. INTRODUCTION

Many of the time-dependent transport equations encountered in mathematical physics are nonlinear in nature due to the dependence of one or more of the coefficients on integrals of the unknown distribution function. In such cases, the transport equation becomes an integro-partial differential equation such as the Vlasov or Boltzmann equations. Physical applications include a large variety of diffusive and plasma phenomena[1,2] as well as nonlinear wave propagation,[3] the dynamics of self-gravitating mass distributions,[4] and the diffusion in energy space of photons due to Compton scattering.[5,6]

Integro-partial differential equations are usually solved by integrating forward in time from a given initial condition using a predictor–corrector algorithm[7,8] or a global relaxation method.[9,10] The convergence properties of such indirect methods are often difficult to predict in advance, and usually depend rather sensitively on both the governing equation and the nature of the initial conditions. In this paper we develop an alternative procedure that allows us to analyze the time variation of the integral function (in this case the temperature) using a *direct* method based upon the governing integro-partial differential equation. The temperature variation is determined by constructing a perturbation expansion via recursive application of the moment equation obtained by integrating the original nonlinear equation with respect to energy.

The coefficients of the perturbation expansion depend on the initial shape of the distribution

[a)]Electronic mail: pbecker@gmu.edu

function, and therefore the resulting temperature variation represents the self-consistent solution to the problem. With the temperature variation determined in advance, the equation governing the distribution function loses its integrodifferential character, and reduces to a linear partial differential equation which can be solved using a variety of standard techniques. The accuracy of the solution can be verified *a posteriori* by integrating the resulting distribution function over energy to obtain another (output) result for the variation of the temperature integral, which can be compared with the (input) temperature representation constructed using the information contained in the perturbation expansion.

In a certain sense, the method developed here allows us to ''separate'' the original integro-partial differential equation into two problems, the first being the determination of the self-consistent temperature variation and the second the solution of the remaining linear partial differential equation for the distribution function. The perturbation series for the temperature is divergent in general, but we demonstrate that it can be recast as a continued fraction that yields convergent results for a variety of initial distributions. In order to illustrate the technique, we focus here on a specific family of equations which is sufficiently general to admit a variety of interesting behaviors.

II. GOVERNING EQUATIONS

The primary motivation for this study is the analysis of the scattering of photons and electrons in the hot, tenuous plasma surrounding a compact astrophysical object such as a neutron star or black hole. This process, referred to as time-dependent Comptonization, is thought to be responsible for producing the variable x-ray emission observed from a variety of sources both within and outside our galaxy. The energy of the photons is modified as a result of multiple interactions with electrons, and consequently the photon energy distribution evolves over time. In this situation the photon distribution function $f(x,y)$ is governed by an integro-partial differential transport equation of the general form[5]

$$\frac{\partial f}{\partial y} = \frac{1}{x^i}\frac{\partial}{\partial x}\left\{ x^i \left[x^j \frac{f}{\theta(y)} + x^k \frac{\partial f}{\partial x} \right] \right\}, \tag{1}$$

where x represents the dimensionless photon energy, y measures the dimensionless time, and i, j, and k are constants. The function $\theta(y)$ represents the time-varying temperature, defined by the integral expression

$$\theta(y) \equiv \frac{I_\alpha(y)}{I_\alpha(0)}, \tag{2}$$

where α is a constant and the power moments of f are defined by

$$I_n(y) \equiv \int_0^\infty x^n f(x,y)\,dx. \tag{3}$$

Note that $\theta(0)=1$ by virtue of (2), and $\theta(y)>0$ for all y since $f(x,y)$ is non-negative. In the time-dependent Comptonization problem, we have $i=j=k=2$. However, we will develop the formalism for arbitrary values of i, j, and k in order to emphasize the generality of the mathematical method. For clarity in the discussion, we shall think of the test particles as ''photons'' and the scattering centers as ''electrons,'' although these identifications are arbitrary.

The total number density of the photons $N_r(y)$ is related to the photon distribution function $f(x,y)$ via

$$N_r(y) \equiv \int_0^\infty x^i f(x,y)\,dx, \tag{4}$$

so that $N_r(y)=I_i(y)$. Interpreting $x^{-i}(\partial/\partial x)x^i$ as the divergence operator in energy space, we observe that the transport equation (1) is written in explicit flux-conservation form, and therefore N_r remains constant since (1) contains no sources or sinks of photons. We seek to solve the transport equation for the distribution function $f(x,y)$ subject to the initial condition

$$f(x,0)\equiv f_0(x), \tag{5}$$

where $f_0(x)$ is a known function specified as part of the problem under consideration. The conserved number density is therefore given by $N_r=\int_0^\infty x^i f_0(x)dx$. Equation (1) drives $f(x,y)$ toward the steady state equilibrium solution given by the exponential spectrum

$$f_{\rm eq}(x)\equiv \frac{N_r p}{(p\,\theta_{\rm eq})^{(i+1)/p}\Gamma\left(\dfrac{i+1}{p}\right)}\exp\left(\frac{-x^p}{p\,\theta_{\rm eq}}\right), \quad p\equiv j-k+1, \tag{6}$$

where $\theta_{\rm eq}$ is the asymptotic equilibrium temperature and the normalization has been set so that the number density of $f_{\rm eq}$ is equal to N_r. In deriving (6) we have also assumed that $(i+1)/p>0$. Whether or not the solution $f(x,y)$ actually reaches $f_{\rm eq}(x)$ depends upon the rate at which the temperature varies in a given situation.

The underlying process modeled by (1) is a stochastic energization of the test particles due to the random motions of the scattering centers. This interpretation is made clear by using (1) to calculate the Fokker–Planck coefficients which express the rates of change of the mean energy $\langle x\rangle$ and the variance σ^2 for a monoenergetic distribution. The results obtained are

$$\frac{d\langle x\rangle}{dy}=(i+k)x^{k-1}-\frac{x^j}{\theta(y)}, \quad \frac{d\sigma^2}{dy}=2x^k, \tag{7}$$

which describe, respectively, the ''drifting'' and ''broadening'' of the distribution due to energy space diffusion.[11] In terms of these coefficients, (1) can be recast as the Fokker–Planck equation

$$\frac{\partial F}{\partial y}=-\frac{\partial}{\partial x}\left[F\frac{d\langle x\rangle}{dy}\right]+\frac{\partial^2}{\partial x^2}\left[F\frac{1}{2}\frac{d\sigma^2}{dy}\right], \tag{8}$$

where the photon number spectrum $F(x,y)$ is defined by

$$F(x,y)\equiv x^i f(x,y), \tag{9}$$

so that $N_r=\int_0^\infty F(x,y)dx$. It can be readily verified that (8) is equivalent to (1). Since i, j, and k are free parameters and $\theta(y)$ is an arbitrary power integral of F, we see that the Fokker–Planck coefficients associated with (1) encompass a large variety of microphysical scattering scenarios.

III. PERTURBATION EXPANSION FOR THE TEMPERATURE

We can obtain a relationship between the power moments by operating on (1) with $\int_0^\infty x^n\,dx$ and integrating by parts twice. This yields

$$\frac{dI_n}{dy}=(n-i)\left[(n+k-1)I_{n+k-2}(y)-\frac{I_{n+j-1}(y)}{\theta(y)}\right], \tag{10}$$

where we have assumed that the power moments I_n exist for all of the required values of n. The validity of this assumption depends on the asymptotic behavior of the initial distribution $f_0(x)$. By recursively applying (10), we can express all of the derivatives of any power moment with respect to y as closed functions of the moments.

One interesting consequence is that we can obtain the derivatives of the temperature integral function $\theta(y)$ as functions of the moments $I_n(y)$. Using (2), the zeroth derivative is given by

$$\theta(y)=\frac{I_\alpha(y)}{I_\alpha(0)}. \tag{11}$$

By making a single application of (10), we find that the first derivative can be expressed as

$$\theta^{(1)}(y)=\frac{\alpha-i}{I_\alpha(0)}\left[(\alpha+k-1)I_{\alpha+k-2}(y)-\frac{I_{\alpha+j-1}(y)}{\theta(y)}\right]. \tag{12}$$

Differentiation of (12) with respect to y yields

$$\theta^{(2)}(y)=\frac{\alpha-i}{I_\alpha(0)}\left[\frac{\theta^{(1)}(y)}{\theta^2(y)}I_{\alpha+j-1}(y)-\frac{1}{\theta(y)}\frac{dI_{\alpha+j-1}}{dy}+(\alpha+k-1)\frac{dI_{\alpha+k-2}}{dy}\right]. \tag{13}$$

Using (10) to eliminate the moment derivatives in (13), we obtain

$$\theta^{(2)}(y)=\frac{\alpha-i}{I_\alpha(0)}\left\{(\alpha+k-1)(\alpha+k-i-2)\left[(\alpha+2k-3)I_{\alpha+2k-4}(y)-\frac{I_{\alpha+k+j-3}(y)}{\theta(y)}\right]\right.$$
$$\left.+\frac{\theta^{(1)}(y)I_{\alpha+j-1}(y)}{\theta^2(y)}+(\alpha+j-i-1)\left[\frac{I_{\alpha+2j-2}(y)}{\theta^2(y)}-(\alpha+j+k-2)\frac{I_{\alpha+j+k-3}(y)}{\theta(y)}\right]\right\}. \tag{14}$$

Subsequent iterative applications of (10) can be used to derive expressions for the third and higher derivatives of $\theta(y)$.

By evaluating the derivatives sequentially starting with $\theta(y)$, we can calculate as many as desired if the required power moments $I_n(y)$ are known. Although we have no *a priori* means of evaluating the power moments $I_n(y)$ for general values of y, we *can* evaluate them for the special case $y=0$ since in this case they correspond to integrals of the known initial distribution $f_0(x)$, i.e.,

$$I_n(0)=\int_0^\infty x^n f(x,0)dx=\int_0^\infty x^n f_0(x)dx. \tag{15}$$

When $y=0$, the general expressions for the derivatives given by (11), (12), and (14) reduce to

$$\theta(0)=1, \tag{16}$$

$$\theta^{(1)}(0)=\frac{\alpha-i}{I_\alpha(0)}[(\alpha+k-1)I_{\alpha+k-2}(0)-I_{\alpha+j-1}(0)], \tag{17}$$

$$\theta^{(2)}(0)=\frac{\alpha-i}{I_\alpha(0)}\{(\alpha+k-1)(\alpha+k-i-2)[(\alpha+2k-3)I_{\alpha+2k-4}(0)-I_{\alpha+k+j-3}(0)]$$
$$+\theta^{(1)}(0)I_{\alpha+j-1}(0)+(\alpha+j-i-1)[I_{\alpha+2j-2}(0)-(\alpha+j+k-2)I_{\alpha+j+k-3}(0)]\}. \tag{18}$$

The method can be extended to evaluate the third and higher derivatives of $\theta(y)$ at $y=0$ for a given initial distribution $f_0(x)$.

Let us suppose that for some arbitrary value of M, all of the initial derivatives of $\theta(y)$ up to $\theta^{(M)}(0)$ have been determined using the method outlined above. We may then define the associated asymptotic perturbation (Taylor) series for $\theta(y)$ by writing

$$\Phi_N(y) \equiv \sum_{n=0}^{N} \frac{\theta^{(n)}(0)}{n!} y^n, \tag{19}$$

where $N \leqslant M$ indicates the truncation level of the series. Our expectation is that $\Phi_N(y)$ accurately approximates the time variation of the exact solution for $\theta(y)$ within some finite radius of convergence if N is sufficiently large. Based upon the existence of the Taylor series, we conclude that in principle the variation of $\theta(y)$ can be determined *before* solving for the unknown distribution $f(x,y)$. This accomplishes the formal ''separation'' of the original integro-partial differential equation (1) into two problems. The first problem is the determination of the variation of the integral function, which has been achieved (at least formally) by constructing the Taylor series (19). The second problem is the determination of the spectrum $f(x,y)$, which now reduces to the solution of a *linear* partial differential equation since the function $\theta(y)$ appearing in (1) can be approximated using (19). However, the convergence of the power series (19) introduces some potential complications which we address below.

IV. CONTINUED FRACTION REPRESENTATION

We have established that it is possible to develop a general computational scheme based on (10) that can be used to evaluate the initial derivatives of the self-consistent temperature integral function $\theta(y)$ in terms of the initial moments $I_n(0)$, which are easily computed using (15) once the initial distribution $f_0(x)$ is specified. From knowledge of the initial θ derivatives we are able to construct the formal Taylor series $\Phi_N(y)$ given by (19). However, a remaining difficulty centers on the convergence of this series. In many cases the radius of convergence turns out to be too small to be of any practical use, and in certain instances it may even vanish. We therefore seek an alternative means for utilizing the asymptotic information contained in the power series coefficients in order to extract a *global* representation of the function.

A global approximation can be constructed by recasting the data in the form of a continued fraction, which is equivalent to the process of Padé approximation.[12] In many instances this is a remarkably successful approach to the global modeling of an unknown function. We define the continued fraction representation using

$$\Psi_N(y) \equiv \cfrac{c_0}{1+\cfrac{c_1 y}{1+\cfrac{c_2 y}{1+\cdots\cfrac{c_{N-1} y}{1+c_N y}}}}, \tag{20}$$

where the constants $c_0,\ldots,c_N$ are the continued fraction coefficients, and the truncation level is indicated by the value of N. The continued fraction coefficients can be computed using the information contained in the θ derivatives by employing the standard two-dimensional algorithm described by Baker and Graves-Morris.[13]

To illustrate the flow of the algorithm, we assume that via successive applications of (10) we have evaluated all of the initial derivatives of the temperature integral function $\theta(y)$ up to $\theta^{(M)}(0)$ for some M. The algorithm is initialized by setting the zeroth column of the matrix $A_{n,m}$ using

$$A_{0,m} = \frac{\theta^{(m)}(0)}{m!}, \quad 0 \leqslant m \leqslant M, \tag{21}$$

and the first column of the matrix is calculated subsequently via

$$A_{1,m} = -\frac{A_{0,m+1}}{A_{0,0}}, \quad 0 \leqslant m \leqslant M-1. \tag{22}$$

The remaining elements in the matrix are obtained recursively using

$$A_{n,m}=\frac{A_{n-2,m+1}}{A_{n-2,0}}-\frac{A_{n-1,m+1}}{A_{n-1,0}}, \quad 0\leqslant m\leqslant M-n, \tag{23}$$

for $2\leqslant n\leqslant M$. The continued fraction coefficients occupy the zeroth row of the matrix, so that

$$c_n=A_{n,0}\ , \quad 0\leqslant n\leqslant M. \tag{24}$$

Note that the coefficient c_n is a function of the initial derivatives $\theta(0)$, $\theta^{(1)}(0),\ldots,\theta^{(n)}(0)$, and therefore the incorporation of higher-order derivatives into the scheme has no effect on the value of c_n.

As discussed in Sec. II, the exact solution for the temperature variation $\theta(y)$ must be positive for all $y>0$ because the distribution function f is non-negative. Hence $\theta(y)$ contains no poles in the domain $y>0$, and this must also be true of any acceptable continued fraction approximation. The singularity structure of the continued fraction $\Psi_N(y)$ can be determined by creating an equivalent rational function, and then solving for the zeros of the denominator. The presence of zeros in the domain $y>0$ causes the appearance of extraneous, unphysical poles in $\Psi_N(y)$. These unphysical poles migrate out of the computational domain as the truncation level N increases and $\Psi_N(y)$ approaches the exact solution $\theta(y)$. However, some of the lower order fractions may contain ''defects'' (poles for positive values of y), and if so they must be rejected. We consider the convergence properties of the continued fraction sequence $\Psi_0(y), \Psi_1(y),\ldots,\Psi_M(y)$ in Sec. VI, where we treat specific computational examples.

V. APPLICATION TO TIME-DEPENDENT COMPTONIZATION

In the problem of time-dependent Comptonization that serves as the primary motivation for this study, an intense distribution of radiation is scattered by hot electrons in a tenuous plasma. We suppose that the photons are injected impulsively into the plasma with a specified energy distribution at time $t=0$, and that the subsequent evolution of the distribution occurs as a result of photon–electron scattering. This process naturally leads to the production of time-variable (transient) x-ray emission. During the brightest observed x-ray transients, most of the energy is contained in the radiation field rather than in the gas, which implies that the average photon energy must remain constant even as the shape of the x-ray spectrum changes. In this case the electrons maintain a Maxwellian distribution and act mainly as catalysts in the evolution of the photon distribution, taking energy away from very energetic photons and giving it to low-energy photons until equilibrium is achieved. The electron temperature T_e depends on the shape of the radiation spectrum via an integral expression, and therefore T_e varies as a function of time as we demonstrate below. The equation governing the evolution of the radiation spectrum is consequently integro-partial differential in nature.

The time evolution of the photon spectrum under the influence of Compton scattering in an ionized, homogeneous hydrogen plasma is governed by the Kompaneets equation,[14]

$$\frac{\partial f}{\partial t}=\frac{n_e\sigma_T ckT_e(0)}{m_e c^2}\,\frac{1}{x^2}\,\frac{\partial}{\partial x}\left\{x^4\left[f+\frac{T_e(t)}{T_e(0)}\,\frac{\partial f}{\partial x}\right]\right\}, \tag{25}$$

where σ_T is the Thomson scattering cross section, k is Boltzmann's constant, c is the speed of light, and n_e and m_e denote the electron number density and mass, respectively. The dimensionless energy variable x is defined by

$$x\equiv\frac{\epsilon}{kT_e(0)}, \tag{26}$$

where ϵ is the photon energy and $T_e(0)$ denotes the electron temperature at the beginning of the transient ($t=0$). The initial spectrum $f_0(x)\equiv f(x,0)$ is assumed to be known, and f is normalized so that the total photon number density is given by

$$N_r(y)=\int_0^\infty x^2 f(x,y)\,dx=I_2(y), \tag{27}$$

where the power moments of f are defined by $I_n(y)\equiv\int_0^\infty x^n f(x,y)\,dx$ in accordance with (3). We remind the reader that N_r=constant according to the discussion in Sec. II. The associated total photon energy density is given by

$$U_r(y)=\int_0^\infty \epsilon\, x^2 f(x,y)\,dx=kT_e(0)I_3(y). \tag{28}$$

In this application, the explicit connection between the ''time parameter'' y and the true time t is established by making the definition

$$y(t)\equiv\int_0^t n_e(t)\sigma_T c\,\frac{kT_e(t')}{m_e c^2}\,dt', \tag{29}$$

where we have allowed for the possibility of a time dependence in the electron number density and we have set $y(0)=0$. Using the variables x and y our transport equation (25) can be reexpressed as

$$\frac{\partial f}{\partial y}=\frac{1}{x^2}\frac{\partial}{\partial x}\left\{x^4\left[\frac{f}{\theta(y)}+\frac{\partial f}{\partial x}\right]\right\}, \tag{30}$$

where the temperature function $\theta(y)$ is defined by

$$\theta(y)\equiv\frac{T_e(y)}{T_e(0)}. \tag{31}$$

In order to apply the method developed earlier in the paper to the current problem, we must demonstrate that the definitions for $\theta(y)$ in (2) and (31) are consistent, which can be established by showing that the right-hand side of (31) is proportional to one of the power moments of f in the Comptonization application. We begin by noting that (30) is formally equivalent to our prototype transport equation (1) if $i=j=k=2$. It follows that the power moments I_n must satisfy (10), which in this case reduces to

$$\frac{dI_n}{dy}=(n-2)\left[(n+1)I_n(y)-\frac{I_{n+1}(y)}{\theta(y)}\right]. \tag{32}$$

During a bright x-ray transient, most of the energy density is contained in the radiation field, and therefore the material gas cannot exchange a significant amount of energy with the photons. Consequently the integrated radiation energy density $U_r(y)=kT_e(0)I_3(y)$ should not change as a result of Comptonization, and therefore $I_3(y)$ must remain equal to its initial value $I_3(0)$. Setting $n=3$ in (32), we find that the condition $dI_3/dy=0$ is satisfied if

$$\theta(y)=\frac{I_4(y)}{4I_3(0)}, \tag{33}$$

which implies that T_e equals the inverse-Compton temperature of the radiation spectrum.[6] Since $\theta(0)=1$ by virtue of (31), we can rewrite (33) as

$$\theta(y)=\frac{I_4(y)}{I_4(0)}. \tag{34}$$

Note that the initial spectrum $f_0(x)$ must satisfy the condition $I_4(0)=4I_3(0)$ in order to be consistent with the requirement that $\theta(0)=1$. Equation (34) establishes the integro-partial differential nature of the governing equation (30) for this application. Hence the formal results obtained earlier in the paper can be applied to the problem of time-dependent Comptonization by setting $\alpha=4$ in (2).

The algorithm derived in Sec. III can be used to directly calculate the initial derivatives of $\theta(y)$ at $y=0$. However, in the Comptonization application under consideration here, it is more convenient to derive a differential recurrence relation between the successive moments I_n and I_{n+1} by rearranging (32) to obtain

$$I_{n+1}(y)=\frac{\theta(y)}{2-n}e^{(n+1)(n-2)y}\frac{d}{dy}[e^{-(n+1)(n-2)y}I_n(y)]. \tag{35}$$

Working in terms of the differential operator

$$\mathcal{D}_n\equiv\frac{1}{2-n}e^{(n+1)(n-2)y}\frac{d}{dy}e^{-(n+1)(n-2)y}, \tag{36}$$

we can apply (35) iteratively to find that

$$I_{n+1}(y)=\theta(y)\mathcal{D}_nI_n=\theta(y)\mathcal{D}_n\theta(y)\mathcal{D}_{n-1}\cdots\theta(y)\mathcal{D}_3I_3(y). \tag{37}$$

Since $I_3(y)=I_3(0)=\text{constant}$, we can carry out the differentiation to obtain the moments $I_n(y)$ as functions of the derivatives of $\theta(y)$. The first few results are

$$I_4(y)=4\,\theta(y)I_3(0), \tag{38}$$

$$I_5(y)=-\frac{\theta(y)}{2}[4\,\theta^{(1)}(y)-40\theta(y)]I_3(0), \tag{39}$$

$$I_6(y)=-\frac{\theta(y)}{3}[-360\theta^2(y)+76\theta(y)\,\theta^{(1)}(y)-2\,\theta^{(1)^2}(y)-2\,\theta(y)\,\theta^{(2)}(y)]I_3(0). \tag{40}$$

Similar results can be obtained for the higher moments. Note that $I_n(y)$ is a function of all of the derivatives of $\theta(y)$ up to $\theta^{(n-4)}(y)$. We have developed a computer algorithm based on (37) that efficiently derives expressions for the moments $I_n(y)$ in terms of the derivatives of $\theta(y)$. Since the values of the initial moments $I_n(0)$ are easily calculated using (15) for any initial spectrum $f_0(x)$, these expressions allow us to compute the corresponding initial derivatives $\theta^{(n)}(0)$ sequentially, beginning with the zeroth derivative which is set by (38). The development of the general expressions giving the initial moments as function of the derivatives $\theta^{(n)}(0)$ is the costliest part of the solution procedure. However, once these expressions have been established, they can be used to evaluate the initial derivatives for a variety of different initial spectra at very low cost. The initial derivatives are subsequently used to calculate the continued fraction coefficients using the algorithm discussed in Sec. IV, and the sequence of continued fraction approximations $\Psi_N(y)$ is evaluated using (20). This procedure forms the basis for the computational results presented in Sec. VI.

VI. COMPUTATIONAL EXAMPLES

In this section we apply our method to obtain quantitative results for the problem of time-dependent astrophysical Comptonization. Our computational procedure is as follows. First we calculate the initial θ derivatives and the associated continued fraction coefficients using the algorithms discussed in Secs. IV and V. In the second step we use this information to analyze the convergence properties of the sequence of continued fractions $\Psi_N(y)$ and compare them with the corresponding Taylor series $\Phi_N(y)$. In the third step we use a high-order continued fraction to approximate the temperature integral function $\theta(y)$ in the transport equation (30) and then we numerically solve the transport equation for the photon energy distribution $f(x,y)$. Finally, in the fourth step we compare the input temperature function $\Psi_N(y)$ with the result obtained for $\theta(y)$ by integrating the numerical solution for $f(x,y)$. Agreement between these two representations of the temperature confirms the accuracy of the method.

At the beginning of the x-ray transient, the radiation spectrum is given by the initial energy distribution $f_0(x)$ introduced in (5). According to the analysis presented in the preceding sections, knowledge of $f_0(x)$ is sufficient to determine the time variation of the temperature integral function $\theta(y)$. In typical astrophysical situations, the initial energy distributions of greatest interest are the optically thin electron–proton bremsstrahlung (free–free) spectrum

$$f_0(x)=x^{-3}e^{-x/4}, \tag{41}$$

and the monoenergetic spectrum

$$f_0(x)=N_0\, x_0^{-2}\,\delta(x-x_0), \tag{42}$$

where N_0 is the number density of the photons. In the latter case, we must set $x_0=4$ in order to satisfy the condition $\theta(0)=I_4(0)/4I_3(0)=1$ as required by (31) and (33).

The transport equation (30) drives the distribution $f(x,y)$ toward the steady state equilibrium solution given by the Wien spectrum

$$f_{\mathrm{eq}}(x)\equiv\frac{N_r}{2\,\theta_{\mathrm{eq}}^3}\,e^{-x/\theta_{\mathrm{eq}}}\ , \tag{43}$$

which has been obtained by setting $i=j=k=2$ in (6). Although f is always driven toward Wien form at all values of y, it may or may not reach equilibrium depending on the shape of the initial spectrum and the corresponding rate at which the temperature varies. This question can be resolved by calculating the asymptotic temperature θ_{eq} using information contained in the initial spectrum $f_0(x)$. In equilibrium, the dimensionless mean photon energy is given by

$$\overline{x}=\frac{\int_0^\infty x^3 f_{\mathrm{eq}}(x)\,dx}{\int_0^\infty x^2 f_{\mathrm{eq}}(x)\,dx}=3\,\theta_{\mathrm{eq}}, \tag{44}$$

where we have substituted for $f_{\mathrm{eq}}(x)$ using (43) to obtain the final result. Conservation of the photon number and energy densities implies that the value of $\overline{x}$ must be conserved, so that we can also write

$$\overline{x}=\frac{\int_0^\infty x^3 f_0(x)\,dx}{\int_0^\infty x^2 f_0(x)\,dx}=\frac{I_3(0)}{I_2(0)}. \tag{45}$$

Equations (44) and (45) can be combined to calculate the asymptotic temperature θ_{eq} for any initial spectrum $f_0(x)$. In the case of a monoenergetic initial spectrum with $x_0=4$, we obtain $\overline{x}=4$ and therefore $\theta_{\mathrm{eq}}=4/3$. Conversely, in the case of a bremsstrahlung initial spectrum, we obtain $\overline{x}=0$ because the number density of photons $N_r=I_2(0)=\int_0^\infty x^2 f_0(x)\,dx$ is formally infinite. This implies that in the bremsstrahlung case, $\theta_{\mathrm{eq}}=0$, and therefore no meaningful steady state exists according to (43). It is important to emphasize that our calculation of θ_{eq} has utilized an energy conservation principle that may not be available in all physical applications of the general transport equation (1).

TABLE I. Numerical results.

	Monoenergetic spectrum		Bremsstrahlung spectrum	
n	$\theta^{(n)}(0)$	c_n	$\theta^{(n)}(0)$	c_n
0	1.00×10^{0}	1.00	1.00×10^{0}	1.00
1	2.00×10^{0}	−2.00	-6.00×10^{0}	6.00
2	-1.20×10^{1}	5.00	1.32×10^{2}	5.00
3	8.00×10^{0}	−1.67	-6.36×10^{3}	11.13
4	1.87×10^{3}	3.59	5.29×10^{5}	8.99
5	-2.99×10^{4}	−2.56	-6.68×10^{7}	15.43
6	-6.85×10^{5}	4.69	1.18×10^{10}	12.28
7	4.07×10^{7}	−3.56	-2.76×10^{12}	19.62
8	3.65×10^{8}	4.13	8.24×10^{14}	15.72
9	-9.25×10^{10}	−3.33	-3.04×10^{17}	23.44
10	3.42×10^{11}	5.03	1.36×10^{20}	19.48
11	3.56×10^{14}	−4.77	-7.19×10^{22}	26.87
12	-6.36×10^{15}	4.53	4.47×10^{25}	23.51
13	-2.20×10^{18}	−4.30	-3.22×10^{28}	30.09
14	7.14×10^{19}	5.17	2.66×10^{31}	27.62
15	2.10×10^{22}	−5.67	-2.49×10^{34}	33.35
16	-9.31×10^{23}	4.97	2.64×10^{37}	31.59
17	-2.96×10^{26}	−5.33	-3.13×10^{40}	36.81
18	1.48×10^{28}	5.20	4.13×10^{43}	35.34
19	5.90×10^{30}	−6.36	-6.04×10^{46}	40.52
20	-2.70×10^{32}	6.06	9.73×10^{49}	38.88
21	-1.60×10^{35}	−10.63	-1.72×10^{53}	44.40
22	4.98×10^{36}	−7.69	3.32×10^{56}	42.26
23	5.87×10^{39}	−32.83	-6.99×10^{59}	48.45
24	9.03×10^{40}	23.42	1.59×10^{63}	45.34

The values of the initial θ derivatives and the associated continued fraction coefficients obtained for the bremsstrahlung and monoenergetic initial spectra are presented in Table I. Note the rapid divergence of the derivatives in each case, implying a limited radius of convergence for the Taylor series $\Phi_N(y)$ given by (19). We compare the convergence properties of the Taylor series and the continued fractions for the bremsstrahlung and monoenergetic initial spectra below.

A. Monoenergetic initial spectrum

The first and second columns of Table I contain, respectively, the results obtained in the monoenergetic case for the initial derivatives $\theta(0)$, $\theta^{(1)}(0),\ldots,\theta^{(M)}(0)$ and the continued fraction coefficients $c_0, c_1,\ldots,c_M$ for $M=24$. This choice for M is arbitrary and of no particular significance. The corresponding sequence of truncated Taylor series $\Phi_N(y)$ is plotted in Fig. 1. The radius of convergence of the Taylor series is limited to $y\lesssim0.15$ even for large N, which reflects the rapid increase in the absolute value of the derivatives of $\theta(y)$ at $y=0$. By contrast, the continued fraction coefficients obtained in the monoenergetic case grow much more slowly in absolute value. The continued fraction sequence $\Psi_N(y)$ contains defects (extraneous poles) for some odd values of N, but $\Psi_N(y)$ converges (albeit nonuniformly) for even values of N, as can be seen in Fig. 2, where we plot the sequence of continued fractions for $N=18$, 20, 22, 24. It is clear from Fig. 2 that the continued fractions converge even for values of y far outside the radius of convergence of the Taylor series. In our search for an accurate approximation to the exact solution for $\theta(y)$, we select the continued fraction $\Psi_N(y)$ that most closely approaches the correct asymptotic value $\theta_{\rm eq}=4/3$ for large y. According to Fig. 2, the best agreement is obtained by using the highest order fraction analyzed in this example, which is $\Psi_{24}(y)$.

With $\theta(y)$ approximated using the continued fraction $\Psi_{24}(y)$, the transport equation (30) reduces to a linear, second-order partial differential equation. We solve this equation using the

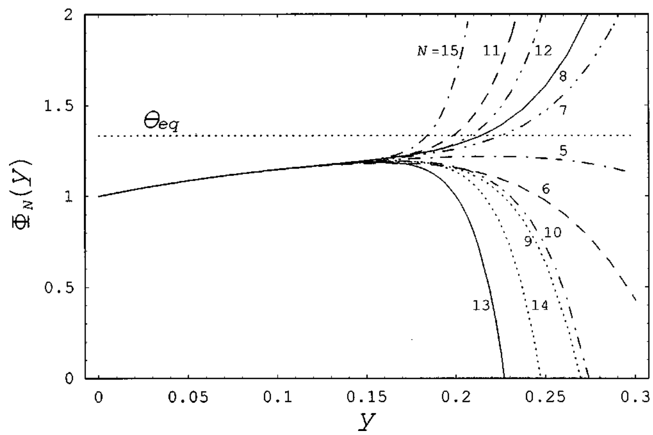

FIG. 1. Sequence of truncated Taylor series $\Phi_N(y)$ given by (19) obtained in the case of Comptonization of a monoenergetic initial spectrum (42) plotted as a function of y, with the truncation level N indicated for each curve. Note that $\Phi_N(y)$ diverges for $y \gtrsim 0.15$ even for large N, due to the rapid growth of the magnitude of the initial θ derivatives, as can be seen in Table I. The asymptotic equilibrium temperature $\theta_{eq} = 4/3$ is denoted by the dotted horizontal line.

IMSL routine DMOLCH over the range $0 \leqslant y \leqslant 2$ and $0 \leqslant x \leqslant 50$. The monoenergetic initial condition (42) imposed at $y=0$ is approximated using a Gaussian distribution with mean $\overline{x}=4$, variance $\sigma^2 = 0.01$, and photon number density $N_0 = 1$. The result obtained for the photon energy spectrum $G(x,y) \equiv x^3 f(x,y)$ is plotted in Fig. 3. It is convenient to plot G rather than f because energy conservation implies that $I_3 = \text{constant}$, and therefore $\int_0^\infty G\, dx$ is independent of y. In the monoenergetic case the initial condition (42) yields $I_3 = 4$. As y increases from zero, the distribution evolves away from the monoenergetic initial form and is well described by the equilibrium Wien distribution (43) for $y \gtrsim 1$.

B. Bremsstrahlung initial spectrum

The third and fourth columns of Table I contain, respectively, the results obtained in the bremsstrahlung case for the initial derivatives $\theta(0)$, $\theta^{(1)},\ldots,\theta^{(M)}(0)$ and the continued-fraction coefficients $c_0, c_1, \ldots, c_M$ for $M=24$. The corresponding sequence of truncated Taylor series

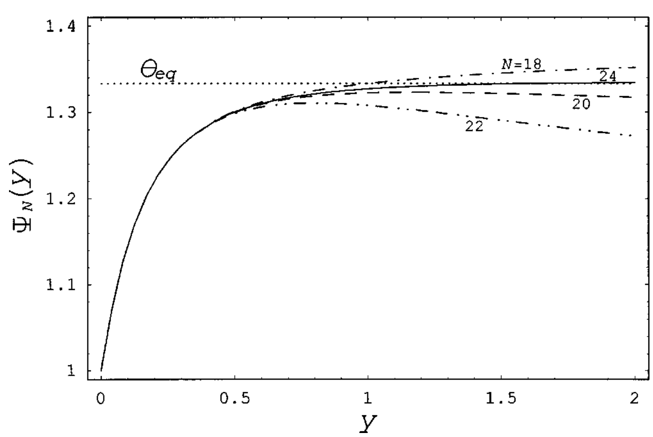

FIG. 2. Sequence of continued fractions $\Psi_N(y)$ given by (20) obtained in the case of Comptonization of a monoenergetic initial spectrum (42) plotted as a function of y, with the truncation level N indicated for each curve. We have plotted the results only for even values of N because in this example $\Psi_N(y)$ contains defects (extraneous poles) for some odd values of N. Note that $\Psi_N(y)$ converges (though nonuniformly) for even values of N throughout the entire computational domain, which extends well beyond the radius of convergence of the Taylor series depicted in Fig. 1. We require that acceptable approximations approach θ_{eq} as $y \to \infty$.

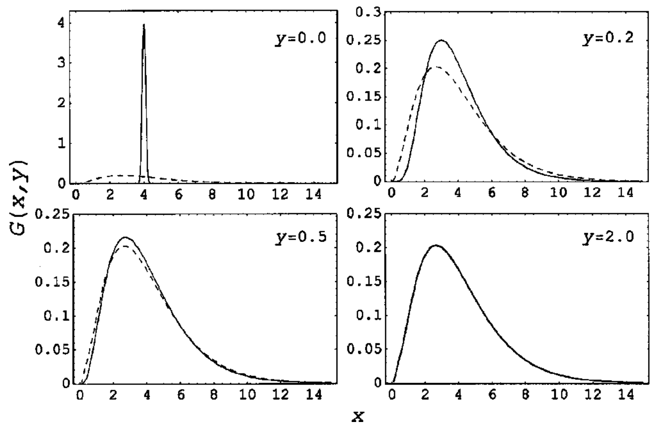

FIG. 3. Numerical result for the photon distribution $G(x,y)\equiv x^3 f(x,y)$ plotted as a function of x for the indicated values of y. In this case the initial spectrum imposed at $y=0$ corresponds to a Gaussian distribution with mean $\bar{x}=4$, variance $\sigma^2=0.01$, and total photon number density $N_0=1$, which approximates the monoenergetic initial spectrum (42). The solid lines denote G and the dashed lines represent the asymptotic equilibrium Wien spectrum given by (43). The photon spectrum is essentially given by the Wien form for $y\gtrsim 1$, and the two distributions are indistinguishable for $y=2$. The area under the curves $\int_0^\infty G\,dx=4$ due to energy conservation.

$\Phi_N(y)$ is plotted in Fig. 4, and the sequence of continued fractions $\Psi_N(y)$ is plotted in Fig. 5. Note that the radius of convergence of the Taylor series actually *decreases* with increasing truncation level N. Conversely, the sequence of continued fractions displays a pattern of uniform convergence with increasing N. The uniform convergence is a consequence of the fact that the computed continued fraction coefficients are all positive in this case, leading us to conjecture that the exact solution $\theta(y)$ is a Stieltjes function when the initial spectrum corresponds to optically thin bremsstrahlung.[12]

The pattern of uniform convergence of $\Psi_N(y)$ in the bremsstrahlung case suggests that we can obtain a reasonable approximation for the exact solution $\theta(y)$ using the continued fraction $\Psi_{24}(y)$. We impose the bremsstrahlung initial condition (41) at $y=0$ and solve the transport equation (30) using DMOLCH over the range $0\leqslant y\leqslant 2$ and $0\leqslant x\leqslant 50$. The result obtained for the function $G(x,y)\equiv x^3 f(x,y)$ is plotted in Fig. 6. The initial condition (41) combined with energy conservation implies that the area under the curve $\int_0^\infty G\,dx=4$ for all y. As y increases from zero,

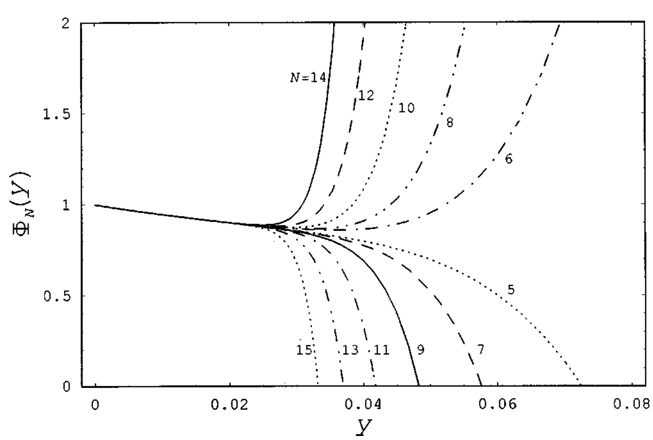

FIG. 4. Sequence of truncated Taylor series $\Phi_N(y)$ obtained in the case of Comptonization of a bremsstrahlung initial spectrum (41) plotted as a function of y, with the truncation level N indicated for each curve. Note that the radius of convergence decreases with increasing N due to the rapid growth of the initial θ derivatives (see Table I). In this case the asymptotic equilibrium temperature $\theta_{eq}=0$ due to the presence of an infinite number of zero-energy photons in the initial spectrum.

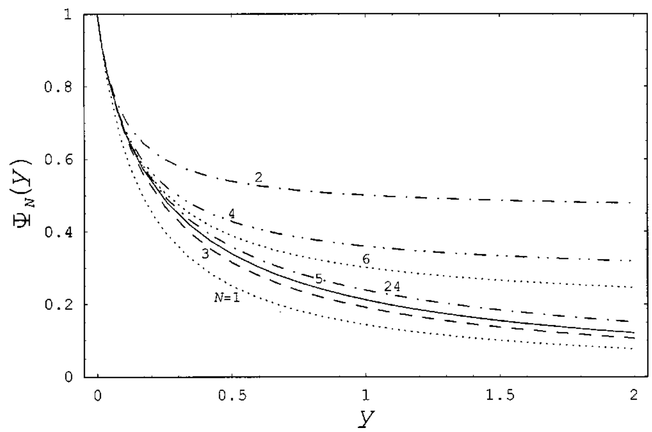

FIG. 5. Sequence of continued fractions $\Psi_N(y)$ obtained in the case of a bremsstrahlung initial spectrum (41) plotted as a function of y, with the truncation level N indicated for each curve. Note the pattern of uniform convergence for $N = 1,2,3,4,5,6$, which suggests that in this case the exact solution $\theta(y)$ is probably a Stieltjes function. For $N>6$, the results become strongly clustered around the $N=24$ curve.

the distribution evolves away from its initial bremsstrahlung form and attempts to approach the equilibrium Wien distribution (43). However, in this case the photon number density $N_r \to \infty$, and therefore the asymptotic value for the temperature $\theta_{\rm eq} = 0$ as discussed earlier. The divergence of the number density is due to the presence of an infinite number of zero-energy photons in the initial spectrum. Since these photons cannot all be upscattered to higher energies without violating overall energy conservation, the ''equilibrium'' solution for the distribution function given by (43) reduces to a pulse centered on zero energy, which is not a meaningful steady state solution. Hence equilibrium cannot be achieved in the bremsstrahlung case, in contrast to the result obtained when the initial spectrum contains a finite number density of photons, as in the monoenergetic example treated above. We analyze the self-consistency of the numerical solutions obtained in the monoenergetic and bremsstrahlung cases in Sec. VII.

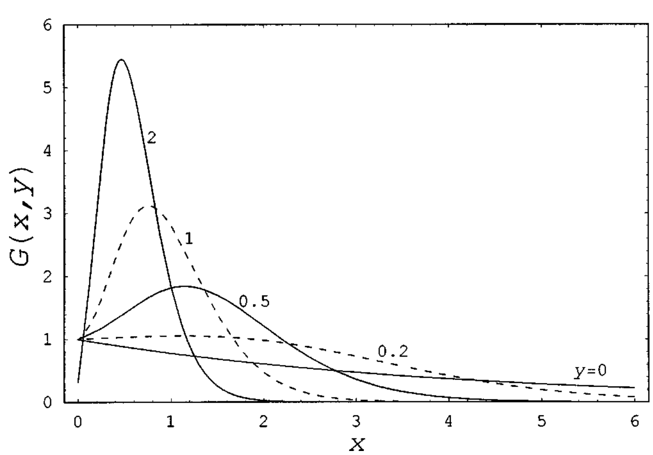

FIG. 6. Numerical results for the photon distribution $G(x,y) \equiv x^3 f(x,y)$ plotted as a function of x for the indicated values of y. In this case the initial condition is the optically thin bremsstrahlung spectrum (41). The distribution function approaches a pulse centered at zero energy as y increases due to the infinite number of zero-energy photons in the initial spectrum. The area under the curves $\int_0^\infty G\,dx = 4$ as a consequence of energy conservation.

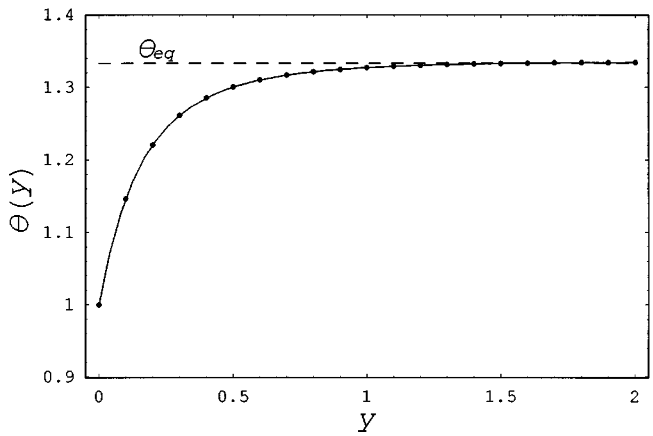

FIG. 7. Results obtained for the (input) continued fraction approximation $\Psi_{24}(y)$ (solid line) are compared with the (output) temperature calculated using the integral definition $\theta(y) \equiv I_4(y)/I_4(0)$ (closed dots) for the Comptonization of a monoenergetic initial spectrum. The dashed horizontal line indicates the asymptotic value of the temperature. The agreement between the two results confirms the self-consistency of the solution obtained for the spectrum $f(x,y)$, which validates the mathematical approach.

VII. EVALUATION OF THE METHOD

Our approach to the solution of the integro-partial differential transport equation (30) has been to approximate the exact solution for the temperature function $\theta(y)$ by using the available derivatives to construct the highest-order continued fraction that is consistent with the known asymptotic behavior of the temperature. With $\theta(y)$ approximated in this way, we have solved for the photon distribution $f(x,y)$ using a standard computer algorithm commonly available in the IMSL library. One may well ask whether there is any guarantee that the numerical solution so obtained is actually the correct physical solution. Interestingly, for the problem treated here there *is* a method that can be used to *guarantee* both the accuracy and the uniqueness of the solution. This is accomplished by integrating the numerically obtained distribution $f(x,y)$ to calculate the corresponding temperature distribution *a posteriori* using (2) and (3). In the case of astrophysical Comptonization that serves as our sample application in this paper, the corresponding expression is $\theta(y) = I_4(y)/I_4(0)$ as given by (34). A comparison between the result for $\theta(y)$ obtained in this manner and the continued fraction approximation $\Psi_N(y)$ used in the solution of the transport equation serves as the acid test of the entire mathematical and computational approach presented here. In Figs. 7 and 8 we perform this comparison for the bremsstrahlung and monoenergetic initial spectra, respectively. The agreement is clearly excellent, verifying the validity of the overall approach.

VIII. CONCLUSION

The technique developed here provides a powerful tool for determining the time dependence of the integral function in an integro-partial differential equation *before* solving for the unknown distribution. The numerical results we have obtained for the variation of the self-consistent temperature function $\theta(y)$ in the case of astrophysical Comptonization suggest that the method has acceptable accuracy and reliable convergence properties. The first step in the procedure is the determination of the initial derivatives of $\theta(y)$ at $y=0$ using the algorithm described in Sec. III, which is based on the differential equation (10) governing the power moments $I_n(y) = \int_0^\infty x^n f(x,y)\,dx$. The initial derivatives $\theta^{(n)}(0)$ are then used to calculate the continued fraction coefficients c_n appearing in the representation of the continued fraction $\Psi_N(y)$. The convergence of the continued fraction sequence is then analyzed and the highest-order fraction that can be constructed using the available set of coefficients is used to approximate the exact solution $\theta(y)$. Next, the transport equation (1) is solved numerically, using the continued fraction approximation

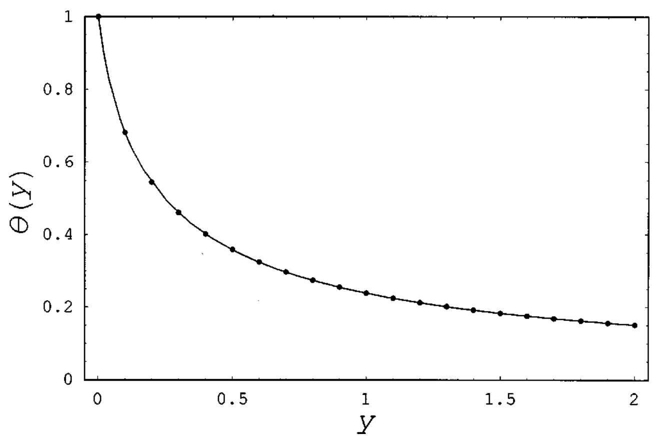

FIG. 8. Results obtained for the (input) continued fraction approximation $\Psi_{24}(y)$ (solid line) are compared with the (output) temperature calculated using the integral definition $\theta(y) \equiv I_4(y)/I_4(0)$ (closed dots) for the Comptonization of a bremsstrahlung initial spectrum. The agreement between the two results confirms the self-consistency of the solution.

as input to calculate the integral term. Finally, the self-consistency of the solution is evaluated by comparing the input continued fraction approximation $\Psi_N(y)$ with the output temperature calculated using the integral definition $\theta(y) \equiv I_\alpha(y)/I_\alpha(0)$. The level of agreement between $\Psi_N(y)$ and $\theta(y)$ provides a measure of the overall accuracy of the solution procedure.

It is interesting to contrast the behavior of the continued fraction sequence $\Psi_N(y)$ with that of the associated Taylor series $\Phi_N(y)$. Since the self-consistent solution for the temperature integral function $\theta(y)$ is positive for all real $y>0$, the limited radius of convergence of the Taylor series must reflect the presence of poles in $\theta(y)$ somewhere else in the complex plane. The existence of these poles essentially dooms any attempt to construct a useful perturbation series for the temperature. Conversely, the success of the continued fraction representation stems from its ability to produce a convergent *global* function that shares the same singularity structure as the exact solution $\theta(y)$. In general, $\Psi_N(y)$ converges toward the exact solution as the truncation level N increases, although the pattern of convergence can vary significantly from problem to problem. The cases for which $\theta(y)$ is a Stieltjes function are of particular importance because in these situations the convergence of the continued fraction sequence is uniform.[12]

In our computational examples, which focus on astrophysical Comptonization, we are able to utilize energy conservation to derive the asymptotic value of the temperature $\theta_{\rm eq}$ in the limit $y \to \infty$. This type of asymptotic information may not be available in every physical application of the general transport equation (1), but if $\theta_{\rm eq}$ can be calculated, we also have the option of including this information directly into the continued fraction using a two-point algorithm such as those given by Becker[15] and by Baker and Graves-Morris.[13] When this information is incorporated into the continued fraction, $\Psi_N(y)$ automatically approaches $\theta_{\rm eq}$ as $y \to \infty$. Note that in order to construct the two-point continued fraction we must first transform the time variable from the infinite domain $0 \leqslant y < \infty$ to an equivalent finite domain.

The method presented here bears some relation to techniques for solving partial differential equations proposed by Jumarie[16] and by Bender, Boettcher, and Milton.[17] In Jumarie's approach, a linear Fokker–Planck equation is used to generate moment equations similar to ours, which are solved using a maximum entropy principle to obtain the distribution function. Conversely, Bender, Boettcher, and Milton determine the distribution function governed by a nonlinear partial differential equation by employing a perturbation expansion followed by Padé summation, which resembles our approach to modeling the integral function $\theta(y)$. Although the procedures developed by these authors incorporate certain elements of the technique presented here, a crucial distinction is that their methods are not applicable to integro-partial differential equations.

In conclusion we point out that the procedure developed in this paper can be used as the basis for a new solution technique, or it can be incorporated into existing predictor–corrector or global iteration algorithms as a means of generating a trial solution for the integral function, which is subsequently improved upon using the standard methods. Our focus here has been upon the details of the method, and therefore we have not presented any comparisons between the efficiencies of the various algorithms available for solving integro-partial differential equations. Nonetheless, it is reasonable to expect that the method proposed here is likely to be quite efficient because the temperature variation is determined in advance of solving for the distribution function. In our computational examples we have treated the problem of astrophysical Comptonization, which is governed by the transport equation (30). However, we emphasize that the method is applicable to any equation of the form represented by (1), and that it can potentially be generalized to treat other transport equations, including those containing inhomogeneous, time-dependent source terms.

ACKNOWLEDGMENTS

The author is grateful to Prasad Subramanian for assistance in carrying out the computational work, and also acknowledges several useful conversations with Ben Crain and Todd Peterson.

[1] J. Satsuma and M. Mimura, ''Stability properties of spatially periodic solutions for nonlinear diffusion equations with a singular integral term,'' Jpn. J. Appl. Math. **4**, 59–72 (1987).
[2] E. Sonnendrücker, J. Roche, P. Bertrand, and A. Ghizzo, ''The semi-Lagrangian method for the numerical resolution of the Vlasov equation,'' J. Comput. Phys. **149**, 201–220 (1999).
[3] P. Knabner and L. von Wolfersdorf, ''Travelling wave solutions of a system of nonlinear diffusion equations with integral term,'' Z. Anal. Anwendungen **10**, 503–512 (1991).
[4] D. Wang, ''Global solutions and stability for self-gravitating isentropic gases,'' J. Math. Anal. Appl. **229**, 530–542 (1999).
[5] P. A. Becker and M. C. Begelman, ''Comptonization in supercritical winds. I. Spectral evolution,'' Astrophys. J. **310**, 534–551 (1986).
[6] G. B. Rybicki and A. P. Lightman, *Radiative Processes in Astrophysics* (Wiley, New York, 1979).
[7] R. M. Thomas and T. E. Simos, ''A family of hybrid exponentially fitted predictor–corrector methods for the numerical integration of the radial Schrödinger equation,'' J. Comput. Appl. Math. **87**, 215–226 (1997).
[8] P. K. Lamm, ''Approximation of ill-posed Volterra problems via predictor–corrector regularization methods,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **56**, 524–541 (1996).
[9] V. J. Ervin and W. J. Layton, ''A robust and parallel relaxation method based on algebraic splittings,'' Numer. Methods Partial Differential Equations **15**, 91–110 (1999).
[10] F. Liu, ''Convergence of approximate solutions for quasilinear hyperbolic conservation laws with relaxation,'' J. Partial Differential Eqs. **11**, 289–300 (1998).
[11] H. Risken, *The Fokker–Planck Equation*, 2nd ed. (Springer, Berlin, 1989).
[12] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw–Hill, New York, 1978).
[13] G. A. Baker and P. R. Graves-Morris, *Encyclopedia of Mathematics and its Applications* (Addison–Wesley, Reading, MA, 1981), Vols. 13 and 14.
[14] A. S. Kompaneets, ''The establishment of thermal equilibrium between quanta and electrons,'' Sov. Phys. JETP **4**, 730–737 (1957).
[15] P. A. Becker, ''A self-consistent theory of photohydrodynamical shocks,'' Astrophys. J. **327**, 772–793 (1988).
[16] G. Jumarie, ''Solution of the multivariate Fokker–Planck equation by using a maximum path entropy principle,'' J. Math. Phys. **31**, 2389–2392 (1990).
[17] C. Bender, S. Boettcher, and K. A. Milton, ''A new perturbative approach to nonlinear partial differential equations,'' J. Math. Phys. **32**, 3031–3038 (1991).

Integral representations of harmonic lattice sums

Jingfang Huang[a)]
Courant Institute of Mathematical Sciences, New York University, New York, New York 10012

(Received 7 May 1999; accepted for publication 22 June 1999)

We present new integral formulas for Coulombic lattice sums that arise in solid state physics, material science, and complex analysis. Unlike Ewald summation, which can yield approximate integral representations, the formulas described here are exact. Simple quadrature rules with modest numbers of nodes yield highly accurate results. © *1999 American Institute of Physics.* [S0022-2488(99)02910-2]

I. INTRODUCTION

Lattice sums arise in a variety of problems in mathematics, physics, biology, and chemistry. Examples include electromagnetic scattering by periodic arrays of obstacles, evaluation of the lattice energy of crystals, analysis of the thermodynamic and structural properties of electrolytes, and the computation of periodic solutions to partial differential equations. In this paper, we consider the calculation of Coulombic lattice sums. In two dimensions, these take the form

$$S_n=\sum_{\omega\in\Lambda_2}\frac{1}{\omega^n},\quad n\geqslant 3, \tag{1}$$

where $\Lambda_2=\{k_1+ik_2|k_1,k_2\in Z,k_1+ik_2\neq 0\}$. In three dimensions, the lattice sums are defined by

$$\mathcal{L}_n^m=\sum_{p\in\Lambda_3}Y_n^m(\theta_p,\phi_p)/r_p^{n+1},\quad n\geqslant 2,\quad m=0,...,n, \tag{2}$$

where $\Lambda_3=\{(k_1,k_2,k_3)|k_i\in Z,(k_1,k_2,k_3)\neq(0,0,0)\}$, (r_p,θ_p,ϕ_p) are the spherical coordinates of $p\in\Lambda_3$, and $Y_n^m(\theta_p,\phi_p)/r_p^{n+1}$ is a spherical harmonic function of degree $(-n-1)$. The spherical harmonics can be defined through the formula (Ref. 1)

$$\frac{Y_n^m(\theta,\phi)}{r^{n+1}}=A_n^m\cdot\left(\frac{\partial}{\partial x}+i\frac{\partial}{\partial y}\right)^m\left(\frac{\partial}{\partial z}\right)^{n-m}\left(\frac{1}{r}\right), \tag{3}$$

where

$$A_n^m=\frac{(-1)^n}{\sqrt{(n-m)!\cdot(n+m)!}}. \tag{4}$$

All of the lattice sums defined above are absolutely convergent. In many situations of interest, one also needs values for S_2, $\mathcal{L}_2^0$, $\mathcal{L}_2^1$, and $\mathcal{L}_2^2$. These quantities, however, are conditionally convergent and their valuation depends on the details of the physical problem.

There are several systematic techniques available for calculating such sums. The best-known method is due to Madelung and Ewald and is generally referred to as Ewald summation (Refs. 2–5). The method works by separating the lattice sum into two parts. Restricting our attention to S_n, we write

[a)]Electronic mail: huangjf@courant.nyu.edu

$$S_n = \sum_{\omega \in \Lambda_2} \frac{1}{\omega^n} \cdot F(\omega) + \sum_{\omega \in \Lambda_2} \frac{1}{\omega^n} \cdot (1 - F(\omega)),$$

where $F(\omega)$ is chosen so that the first sum converges rapidly in ''physical space,'' while the second sum decays rapidly in the Fourier domain, so that the Poisson summation formula (Ref. 6) can be used effectively. $F(\omega)$ is commonly chosen to be a simple Gaussian function. When $F(\omega)$ is the characteristic function of a subset of Λ_2, the method is occasionally referred to as the planewise summation method (see Refs. 7–9).

A rather different approach is through the use of recurrence formulas. In the two-dimensional case, there is a remarkable nonlinear relation (Ref. 10) of the form

$$(4m^2-1)(m-3)S_{2m} = 3\sum_{r=2}^{m-2} (2r-1)(2m-2r-1)S_{2r}S_{2m-2r}, \tag{5}$$

for $m \geqslant 4$. Given S_4 and S_6, all higher-order sums can be evaluated in this manner. (The odd lattice sums vanish.) Unfortunately, S_4 and S_6 are the most slowly converging, and are typically computed via Ewald summation. Furthermore, the formula (5) has no known analog in the three-dimensional case. More recently, Berman and Greengard (Ref. 11) used renormalization arguments to obtain a linear, infinite recurrence relation of the form

$$S_n - \sum_{k=n}^{\infty} A_{n-k}S_k = R_n, \tag{6}$$

where the terms A_n and R_n are simple finite sums. Their method extends to arbitrary dimensions and proceeds in the numerically attractive direction: i.e., it yields slowly converging sums in terms of rapidly converging ones. Another approach, that of Helsing and Lambert *et al.* (Refs. 12 and 13), relies on multipole expansions to accelerate the direct evaluation of the lattice sum to any desired precision and requires an amount of work that grows logarithmically with the number of lattice sites included.

In the present paper, we derive simple integral formulas (Theorems II.2 and IV.2) for the lattice sums S_n and $\mathcal{L}_n^m$. These formulas rely on ''plane-wave'' expansions of the fundamental singularity ($1/z$ in two dimensions and $1/r$ in three dimensions), which form the basis for modern versions of the fast multipole method (Refs. 14–17). We extend our analysis to the case of skewed lattices in two dimensions, where the corresponding sums play a fundamental role in the study of the Weierstrass $\wp$ function. A numerical scheme based on these formulas is extremely simple, easy to implement, and achieves very high accuracy.

II. TWO-DIMENSIONAL THEORY

Our starting point for the two-dimensional case is the plane-wave representation for a pole of degree n (Ref. 15).

Theorem II.1: *Let $z = x + iy$ denote a point in the complex plane:*

$$\textit{If}\ \ x>0,\ \ \textit{then}\ \ \frac{1}{z^n} = \frac{1}{(n-1)!}\int_0^\infty \lambda^{n-1} e^{-\lambda z}\, d\lambda;$$

$$\textit{if}\ \ x<0,\ \ \textit{then}\ \ \frac{1}{z^n} = \frac{1}{(n-1)!}\int_0^\infty (-1)^n \cdot \lambda^{n-1} e^{\lambda z}\, d\lambda;$$

$$\textit{if}\ \ y>0,\ \ \textit{then}\ \ \frac{1}{z^n} = \frac{1}{(n-1)!}\int_0^\infty (-i)^n \cdot \lambda^{n-1} e^{i\lambda z}\, d\lambda;$$

$$if \quad y<0, \quad then \quad \frac{1}{z^n}=\frac{1}{(n-1)!}\int_0^{\infty} i^n\cdot\lambda^{n-1}e^{-i\lambda z}\,d\lambda.$$

We will refer to these as the east, west, north, *and* south *formulas, respectively*.

Proof: For $n=1$, each formula is obvious by inspection. Differentiation yields the general case. ■

Theorem II.2: *Let $n\geqslant 3$. If n is not a multiple of 4, then $S_n=0$. Otherwise,*

$$S_n=\frac{4}{(n-1)!}\int_0^{\infty} e^{-\lambda}\cdot\lambda^{n-3}\cdot\left[\lambda^2\frac{1+\cos\lambda}{1-2e^{-\lambda}\cdot\cos\lambda+e^{-2\lambda}}\right]d\lambda. \tag{7}$$

Proof: If n is not a multiple of 4, simple symmetry considerations show that $S_n=0$. For $n=4l$, we divide the lattice points in the plane as follows. The *east list* is defined to be the set $\{(k,j)|k,j\in Z,k>0,-k\leqslant j\leqslant k\}$ and the *west list* is defined to be the set $\{(-k,j)|k,j\in Z,k>0,-k\leqslant j\leqslant k\}$. The *north list* is defined to be the set $\{(j,k)|j,k\in Z,k>0,-k<j<k\}$ and the *south list* is defined to be the set $\{(j,-k)|j,k\in Z,k>0,-k<j<k\}$. Note that lattice points along the diagonal lines $x=\pm y$ are preferentially assigned to the *east* and *west* lists. Making use of the integral representations in Theorem II.1, we can write

$$S_n=\frac{2}{(n-1)!}\int_0^{\infty}\lambda^{n-1}\sum_{k=1}^{\infty} e^{-\lambda k}\cdot\left(\sum_{j=-k}^{k} e^{-i\lambda j}+\sum_{j=-k+1}^{k-1} e^{-i\lambda j}\right)d\lambda.$$

The various geometric series can be summed analytically, yielding the desired results. ■

III. APPLICATION TO ELLIPTIC FUNCTIONS

The two-dimensional lattice sum S_n is a special case of the Eisenstein series G_n that arises in complex analysis and number theory, especially in the study of the Weierstrass $\wp$ function.

Definition III.1: Let $\tau=(\tau_1,\tau_2)$ be a complex number with $\tau_2>0$ and $-\frac{1}{2}<\tau_1\leqslant\frac{1}{2}$, and let Ω denote the set $\{k_1+k_2\tau|k_1,k_2\in Z,k_1+ik_2\neq 0\}$. The Weierstrass $\wp$ function is defined by

$$\wp(z)=\frac{1}{z^2}+\sum_{\omega\in\Omega}\left\{\frac{1}{(z-\omega)^2}-\frac{1}{\omega^2}\right\}.$$

If we expand the $\wp$ function as a Laurent series about the origin, we obtain

$$\wp(z)=\frac{1}{z^2}+\sum_{n=1}^{\infty}(2n+1)G_{2n+2}z^{2n},$$

where

$$G_n(\tau)=\sum_{\omega\in\Omega}\frac{1}{\omega^n}.$$

The series is valid for $0<|z|<\min\{|\omega|:\omega\in\Omega\}$.

Notice that the $G_n(\tau)$ is, in general, a skewed lattice sum. Our method, however, still applies.

Theorem III.1: *Let $n\geqslant 3$. If n is odd, then $G_n=0$. Otherwise, G_n has the integral representation*

$$G_n(\tau)=\frac{2}{(n-1)!}\int_0^{\infty}\lambda^{n-1}[f_1(\lambda,\tau)+i^n f_2(\lambda,\tau)]d\lambda, \tag{8}$$

where

$$f_1(\lambda,\tau)=\frac{e^{\lambda}-e^{\lambda\tau}+e^{\lambda+\lambda\tau}+e^{\lambda+2\lambda\tau}}{(e^{\lambda}-e^{\lambda\tau})(e^{\lambda+\lambda\tau}-1)}, \tag{9}$$

and

$$f_2(\lambda,\tau)=\frac{e^{i(\lambda+\lambda\tau)}(1+e^{i\lambda\tau})}{(e^{i\lambda\tau}-e^{i\lambda})(e^{i(\lambda+\lambda\tau)}-1)}. \tag{10}$$

Proof: Using the same decomposition as in the proof of Theorem II.2, we obtain

$$G_n(\tau)=\frac{1+(-1)^n}{(n-1)!}\int_0^\infty \lambda^{n-1}\sum_{k=1}^{\infty}\sum_{j=-k}^{k} e^{-\lambda k}e^{-\lambda j\tau}\,d\lambda$$
$$+\frac{i^n+(-i)^n}{(n-1)!}\int_0^\infty \lambda^{n-1}\sum_{j=1}^{\infty}\sum_{k=-j+1}^{j-1} e^{i\lambda k}e^{i\lambda j\tau}\,d\lambda.$$

If n is odd, it is clear that $G_n=0$. Explicit evaluation of the sums (all of which are absolutely convergent) completes the proof. ■

From the Laurent expansion for $\wp(z)$ and the fact that

$$e^z+e^{-z}=2\sum_{n=0}^{\infty}\frac{z^{2n}}{(2n)!},$$

we have the following corollary.

Corollary III.1: The Weierstrass $\wp(z)$ function has the integral representation

$$\wp(z)=\frac{1}{z^2}+\int_0^\infty \lambda[(e^{\lambda z}+e^{-\lambda z}-2)f_1(\lambda,\tau)-(e^{i\lambda z}+e^{-i\lambda z}-2)f_2(\lambda,\tau)]d\lambda. \tag{11}$$

IV. THREE-DIMENSIONAL THEORY

We turn our attention now to the lattice sums $\mathcal{L}_n^m$, with $n\geqslant 3$ and $0\leqslant m\leqslant n$. By symmetry considerations, it is easy to show that $\mathcal{L}_n^m=0$ unless n is even and m is a multiple of 4. We, therefore, restrict our attention to the nonzero cases.

In Refs. 16–17, the following plane-wave expansions are introduced and used to diagonalize certain translation operators. The cases with $n=0$ can be found in Ref. 18 and the higher-order results follow from (3).

Theorem IV.1: *Let (x,y,z) be a point with spherical coordinates (r,θ,ϕ) and let $C=1/(2\pi\sqrt{(n-m)!(n+m)!})$. If $z>0$, we have the* up *formula*

$$\frac{Y_n^m(\theta,\phi)}{r^{n+1}}=C\cdot i^m\int_0^\infty\int_0^{2\pi} e^{im\alpha}\lambda^n e^{-\lambda z+i\lambda(x\cos\alpha+y\sin\alpha)}\,d\alpha\,d\lambda. \tag{12}$$

For $z<0$, we have the down *formula*

$$\frac{Y_n^m(\theta,\phi)}{r^{n+1}}=C\cdot i^m(-1)^{n-m}\int_0^\infty\int_0^{2\pi} e^{im\alpha}\lambda^n e^{\lambda z+i\lambda(x\cos\alpha+y\sin\alpha)}\,d\alpha\,d\lambda. \tag{13}$$

For $y>0$, we have the north *formula*

$$\frac{Y_n^m(\theta,\phi)}{r^{n+1}}=C\cdot i^n(-1)^{n-m}\int_0^\infty\int_0^{2\pi}(\cos\alpha-1)^m(\sin\alpha)^{n-m}\lambda^n e^{-\lambda y+i\lambda(x\cos\alpha+z\sin\alpha)}\,d\alpha\,d\lambda. \tag{14}$$

TABLE I. The nonzero lattice sums S_n for $n \leqslant 100$.

n	Quadrature nodes	S_n
4	108	0.315 121 200 215 41D+01
8	104	0.425 577 303 536 49D+01
12	98	0.393 884 901 282 80D+01
16	76	0.401 569 503 302 50D+01
20	62	0.399 609 675 317 63D+01
24	52	0.400 097 680 530 38D+01
28	50	0.399 975 587 547 45D+01
32	50	0.400 006 103 605 38D+01
36	53	0.399 998 474 126 80D+01
40	59	0.400 000 381 470 10D+01
44	59	0.399 999 904 632 59D+01
48	61	0.400 000 023 841 86D+01
52	65	0.399 999 994 039 53D+01
56	69	0.400 000 001 490 12D+01
60	69	0.399 999 999 627 47D+01
64	69	0.400 000 000 093 13D+01
68	71	0.399 999 999 976 71D+01
72	72	0.400 000 000 005 85D+01
76	72	0.399 999 999 998 54D+01
80	76	0.400 000 000 000 39D+01
84	76	0.399 999 999 999 90D+01
88	80	0.400 000 000 000 02D+01
92	80	0.399 999 999 999 97D+01
96	86	0.400 000 000 000 00D+01
100	86	0.400 000 000 000 00D+01

For $y<0$, *we have the* south *formula*

$$\frac{Y_n^m(\theta,\phi)}{r^{n+1}} = C\cdot i^n(-1)^{n-m}\int_0^\infty\int_0^{2\pi}(\cos\alpha+1)^m(\sin\alpha)^{n-m}\lambda^n e^{\lambda y+i\lambda(x\cos\alpha+z\sin\alpha)}\,d\alpha\,d\lambda. \tag{15}$$

For $x>0$, *we have the* east *formula*

$$\frac{Y_n^m(\theta,\phi)}{r^{n+1}} = C\cdot i^{n-m}(-1)^n\int_0^\infty\int_0^{2\pi}(\cos\alpha+1)^m(\sin\alpha)^{n-m}\lambda^n e^{-\lambda x+i\lambda(y\cos\alpha+z\sin\alpha)}\,d\alpha\,d\lambda. \tag{16}$$

For $x<0$, *we have the* west *formula*

$$\frac{Y_n^m(\theta,\phi)}{r^{n+1}} = C\cdot i^{n-m}(-1)^n\int_0^\infty\int_0^{2\pi}(\cos\alpha-1)^m(\sin\alpha)^{n-m}\lambda^n e^{\lambda z+i\lambda(y\cos\alpha+z\sin\alpha)}\,d\alpha\,d\lambda. \tag{17}$$

Theorem IV.2: *Let n be an even integer and let m be a multiple of 4. Then*

$$\mathcal{L}_n^m = C\int_0^\infty \lambda^{n-3}e^{-\lambda}\int_0^{2\pi}\left\{\cos(m\alpha)\frac{f_{\mathrm{ud}}(\lambda,\alpha)}{d(\lambda,\alpha)} + i^n(\cos\alpha+1)^m(\sin\alpha)^{n-m}\frac{f_{\mathrm{nsew}}(\lambda,\alpha)}{d(\lambda,\alpha)}\right\}d\alpha\,d\lambda, \tag{18}$$

where

$$C = 1/(\pi\sqrt{(n-m)!(n+m)!}),$$

TABLE II. The nonzero lattice sums $\mathcal{L}_n^m$ for $n \leqslant 20$.

n	m	N_λ, N_α	$\mathcal{L}_n^m$
4	0	160, 60	0.310 822 668 269 93D+01
4	4	166, 66	0.185 752 072 772 93D+01
6	0	200, 100	0.573 329 289 434 56D+00
6	4	160, 66	−0.107 260 088 543 32D+01
8	0	154, 66	0.325 929 309 334 95D+01
8	4	154, 66	0.122 565 950 318 49D+01
8	8	146, 66	0.186 744 362 272 51D+01
10	0	146, 66	0.100 922 398 807 09D+01
10	4	146, 66	−0.101 695 761 831 70D+01
10	8	128, 66	−0.121 042 167 434 54D+01
12	0	128, 70	0.289 125 410 827 68D+01
12	4	132, 70	0.891 680 221 309 52D+00
12	8	106, 66	0.106 997 412 357 96D+01
12	12	100, 66	0.158 877 975 803 36D+01
14	0	116, 78	0.115 363 679 871 65D+01
14	4	120, 82	−0.848 339 903 867 39D+00
14	8	116, 78	−0.910 342 301 833 73D+00
14	12	110, 74	−0.110 536 661 309 21D+01
16	0	100, 82	0.279 235 629 890 90D+01
16	4	94, 82	0.800 719 357 862 68D+00
16	8	88, 78	0.841 131 590 059 37D+00
16	12	84, 74	0.947 939 589 167 13D+00
16	16	72, 70	0.150 058 477 228 41D+01
18	0	90, 86	0.125 802 435 502 05D+01
18	4	90, 86	−0.749 850 929 484 74D+00
18	8	86, 86	−0.784 927 978 206 34D+00
18	12	82, 82	−0.845 512 429 512 04D+00
18	16	76, 78	−0.104 366 495 400 43D+01
20	0	80, 90	0.270 422 478 070 66D+01
20	4	80, 90	0.709 961 282 759 62D+00
20	8	80, 90	0.734 552 158 130 00D+00
20	12	72, 86	0.783 312 574 984 17D+00
20	16	68, 82	0.891 188 099 869 61D+00
20	20	60, 78	0.141 533 291 104 07D+01

$$\begin{aligned} f_{\mathrm{ud}} = \lambda^3 \{ & 1 + 2\cos(\lambda\cos\alpha) + 2\cos(\lambda\sin\alpha) + 4\cos(\lambda\cos\alpha)\cos(\lambda\sin\alpha) \\ & + e^{-\lambda}(1 - 2\cos(\lambda\cos\alpha) - 2\cos(\lambda\sin\alpha) - 4\cos^2(\lambda\cos\alpha) - 4\cos^2(\lambda\sin\alpha)) \\ & + e^{-2\lambda}(-1 + 4\cos(\lambda\cos\alpha)\cos(\lambda\sin\alpha)) - e^{-3\lambda} \}, \end{aligned}$$

$$\begin{aligned} f_{\mathrm{nsew}} = 2\lambda^3 \{ & 1 + \cos(\lambda\cos\alpha) + 2e^{-\lambda}(\cos(\lambda\cos\alpha) + \cos^2(\lambda\cos\alpha)) - e^{-2\lambda}(1 + \cos(\lambda\cos\alpha) \\ & + 2\cos(\lambda\sin\alpha) + 2\cos(\lambda\cos\alpha)\cos(\lambda\sin\alpha)) \}, \end{aligned}$$

and

$$d(\lambda,\alpha) = (1 - 2e^{-\lambda}\cos(\lambda(\cos\alpha - \sin\alpha)) + e^{-2\lambda})(1 - 2e^{-\lambda}\cos(\lambda(\cos\alpha + \sin\alpha)) + e^{-2\lambda}).$$

Proof: The derivation of this formula is analogous to the two-dimensional case. We omit the details, but note that the term f_{ud} collects the contributions from the *up* and *down* directions, while the term f_{nsew} collects the contributions from the other four directions. The functions $f_{\mathrm{ud}}(\lambda,\alpha)/d(\lambda,\alpha)$ and $f_{\mathrm{nsew}}(\lambda,\alpha)/d(\lambda,\alpha)$ are well defined as $\lambda \to 0$. The zeros in the denominator are canceled by zeros in the numerator. ∎

V. NUMERICAL RESULTS

To illustrate the utility of our integral formulas, we evaluate the two- and three-dimensional lattice sums using generalized Laguerre quadrature with N_λ nodes for the integral on $[0,\infty]$ with the weight function $W(x)=e^{-x}\cdot x^{n-3}$, and the trapezoidal rule with N_α nodes for the inner integral on $[0, 2\pi]$ in the three-dimensional case. The results are listed in Tables I and II.

We compare our calculations to the results obtained in Ref. 11. They agree to at least 14 significant digits.

VI. CONCLUSION

In this paper, we have developed new integral formulas for harmonic lattice sums in two and three dimensions. Our method relies on the existence of plane-wave expansions for the fundamental solution, and can be used for rectangular or skewed lattices. The approach can be extended to a variety of kernels, including screened Coulomb (Yukawa) potentials and Helmholtz potentials (Ref. 19). Work along these lines is in progress and will be reported at a later date. Applications of the integral representations to problems in analytic number theory remain to be investigated.

ACKNOWLEDGMENTS

This research was supported by the Applied Mathematical Sciences Program of the U.S. Department of Energy under Contract No. DE-FG0288ER25053. The author is grateful to Professor Leslie Greengard for several helpful discussions.

[1] P. R. Wallace, *Mathematical Analysis of Physical Problems* (Dover, New York, 1984).
[2] P. Ewald, ''Die Berechnung optischer und elektrostatischer Gitterpotentiale,'' Ann. Phys. (Leipzig) **64**, 253 (1921).
[3] B. R. A. Nijboer and F. W. De Wette, ''On the calculation of lattice sums,'' Physica (Amsterdam) **23**, 309 (1957).
[4] D. Borwein and J. M. Borwein, ''Analysis of certain lattice sums,'' J. Math. Anal. Appl. **143**, 126 (1989).
[5] K. E. Schmidt and M. A. Lee, ''Implementing the fast multipole method in three dimensions,'' J. Stat. Phys. **63**, 1223 (1991).
[6] G. Barton, *Elements of Green's Functions and Propagation* (Oxford University Press, London, 1989).
[7] B. van der Hoff and G. C. Benson, Can. J. Phys. **31**, 1087 (1953); G. C. Benson and H. P. Schreiber, *ibid.* **33**, 529 (1955).
[8] C. A. Sholl, Proc. Phys. Soc. London **92**, 434 (1967).
[9] M. L. Glasser, J. Math. Phys. **14**, 409 (1973); **14**, 701 (1973); **15**, 188 (1974); **16**, 1237 (1974).
[10] T. M. Apostol, *Modular Functions and Dirichlet Series in Number Theory* (Springer-Verlag, New York, 1976).
[11] C. L. Berman and L. Greengard, ''A renormalization method for the evaluation of lattice sums,'' J. Math. Phys. **35**, 6036 (1994).
[12] J. Helsing, ''Bounds on the shear modulus of suspensions by interface methods,'' J. Mech. Phys. Solids **42**, 1123 (1994).
[13] C. G. Lambert, T. A. Darden, and J. A. Board, ''A multipole-based algorithm for efficient calculation of forces and potentials in macroscopic periodic assemblies of particles,'' J. Comput. Phys. **126**, 274 (1996).
[14] L. Greengard, J. F. Huang, V. Rokhlin, and S. Wandzura, ''Accelerating fast multipole methods for the Helmholtz equation at low frequencies,'' IEEE Comput. Sci. Eng. **5**, 32 (1998).
[15] T. Hrycak and V. Rokhlin, ''An improved fast multipole algorithm for potential fields,'' SIAM (Soc. Ind. Appl. Math.) J. Sci. Stat. Comput. **19**, 1804 (1998).
[16] H. Cheng, L. Greengard, and V. Rokhlin, ''A fast adaptive multipole algorithm in three dimensions,'' Research Report YALEU/DCS/RR-1158, 1998.
[17] L. Greengard and V. Rokhlin, ''A new version of the fast multipole method for the Laplace equation in three dimensions,'' in *Acta Numerica* (Cambridge University Press, Cambridge, 1997) Vol. 6, pp. 229–269.
[18] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953).
[19] S. K. Chin, N. A. Nicorovici, and R. C. McPhedran, ''Green's function and lattice sums for electromagnetic scattering by a square array of cylinders,'' Phys. Rev. E **49**, 4590 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Symmetries of boundary value problems in mathematical physics

M. Makai[a)]
KFKI Atomic Energy Research Institute, H-1525 Budapest 114 POB 49, Hungary

Y. Orechwa[b)]
751 East Boughton Road, Bolingbrook, Illinois 60440

(Received 26 February 1999; accepted for publication 19 May 1999)

The problem considered here is to find a function satisfying a linear elliptic differential or integral equation inside a finite simply region V and another linear first-order differential or integral equation on the ∂V boundary. The symmetries of the above problem form a point group. We show that if the homogeneous problem has only the trivial solution, then the symmetry of the solution inside V inherits the symmetry of the boundary value, given on ∂V. The boundary value is decomposed into irreducible components and the physical meaning of the irreducible components is highlighted. We then apply the results to investigate a widely utilized numerical solution technique that is based on a variational principle and utilizes two approximations. The first one approximates, the solution inside V by a polynomial, the second approximation assumes the solution on the boundary to be a low-order polynomial. By means of group representation theory, we show that the mentioned approximations may fail for certain combinations. The predicted problems have been observed in the VARIANT code, which is routinely used to solve the multigroup neutron diffusion equation. Our method is also applicable to the Schrödinger, and to heat conductance and wave equations. [S0022-2488(99)02610-9]

I. MOTIVATION

In the main fields of the authors, in neutron physics, we are looking for numerical techniques to solve the multigroup neutron diffusion equation in a large volume V of mosaic-like structure. One of the final goals is to work out real-time simulator models, where a computer program simulates a nuclear reactor. These days, the state of the art of reactor physics allows only for either a crude but real time or a fine but slow modeling of a power reactor core. Thus, acceleration methods are sought.

It is known that given a group G, the domain of which includes the solution space $\mathcal{L}$, we can split $\mathcal{L}$ into $|G|$ orthogonal subspace.[1] Since most numerical models arrive at a set of linear equations,[2] the obvious thing is to exploit the splitting offered by G. Furthermore, if G leaves the equation invariant, we can evaluate certain matrix elements easily.[3] In some cases we may hope to split the problem into independent subproblems.

In the investigations below, we focus on second-order equations (Schrödinger, heat conductance, diffusion, wave equation, etc), where the physical parameters (potential, heat conductance, diffusion constant, etc.) are regionwise constant and finite in V. In such cases[2] the solution of a boundary value problem is continuous at internal material interfaces, and the normal gradient of the solution is also continuous. The solution of the problem is unique. What is usually not unique is the solution to the associated eigenvalue problem, where we require some linear expression of the solution to become zero at the boundary. For some equations, however, the solution is unique

[a)]Electronic mail: makai@sunserv.kfki.hu
[b)]Electronic mail: YOrechwa@aol.com

up to a constant multiplication factor, and even positive inside the investigated volume.[4]

The symmetry group of the equation[5,6] can play the role of group G. In a finite volume V, the symmetries of the problem make a point group. The question is, how can we exploit that the solution space is split into orthogonal components? First, we find out, under what conditions the solution will possess the same symmetries as those of the boundary value. For a given class of operators, the solution inside V will be shown to inherit the symmetries of the boundary value on ∂V. Hence, for those operators the decomposition of the boundary value into irreducible components entails a decomposition of the solution inside V. In other words, V can be characterized by the solutions developing in response to some model boundary conditions. Those responses can be calculated beforehand and an effective iteration can be worked out in order to solve the boundary value problem.

When solving practical problems, we often resort to numerical methods.[7] A natural iteration is organized as follows. V is subdivided into smaller volumes called nodes. We sweep through the nodes one by one. Since the boundary sources determine the solution, first we collect the solution on the boundary using the continuity condition at internal boundaries and the boundary condition at an external boundary. We can approximate the solution on the boundary by low-order polynomials. The solution inside the node may again be approximated by some polynomials and, from the solution, we reevaluate the solution and its gradient at the boundaries and pass on to the next node. This procedure is often applied not to the solution, but to a linear combination of the solution and its normal gradient. This specific iterative method[8] is based on variational principles and runs into convergence problems. With the help of group theory, we can predict what kinds of approximations will lead to a convergence problem.

The structure of the paper is as follows. In Sec. II, the investigated boundary value problem is defined. The symmetries of a boundary problem[9] and their consequences are dealt with in Sec. III. A generally utilized iteration scheme is described in Sec. IV. Applications to neutron diffusion are provided in Sec. V. In Sec. VI we offer a number of concluding remarks.

II. THE PROBLEM

Throughout the present work, we consider the following boundary value problem:

$$\begin{aligned} \mathbf{A}\Psi(r) &= Q(r), \quad r \in V; \\ \mathbf{B}\Psi(x) &= q(x), \quad x \in \partial V, \end{aligned} \tag{1}$$

where V is a finite simply connected volume; its boundary is ∂V. $Q(r)$ is the external source in V, $q(x)$ is a given function on the boundary ∂V. Operator $\mathbf{A}$ is a linear, second-order differential operator in the space variable x, $\mathbf{B}$ is a linear, first-order differential operator. If $\mathbf{A}$ and $\mathbf{B}$ involve space-dependent functions, representing material properties, those functions are continuous, smooth functions of x. We assume that either $Q \neq 0$ or $q \neq 0$.

We assume, furthermore, that the homogeneous problem, with $Q=0$ and $q=0$, has no nontrivial solution. Under the stipulated conditions, the Fredholm alternative theorem ensures that the boundary value problem has a unique solution.[10] Volume V is often subdivided into parts that we call nodes, and inside a node the material properties are assumed as constant in x. In this case, the solution and its suitable normal derivative is continuous at internal material interfaces. Such problems are plentiful in mathematical physics.

Example 1: The first problem, which is mainly dealt with in this paper, is the multigroup neutron diffusion equation:[2]

$$\mathbf{A} = -\nabla^2 \underline{\Psi}(r) + S\underline{\Psi}(r), \tag{2}$$

where

$$S_{gg'} = (\Sigma_g - \chi_g \nu \Sigma_{fg'} - \Sigma_{sg'-g})/D_g. \tag{3}$$

Here subscript g refers to the neutron's energy (energy group index), D_g is the diffusion constant in energy group g, Σ_g is the removal cross section in group g, $\nu\Sigma_{fg}$ is the production cross section, Σ_{sgg} is the scattering cross section, and χ_g is the fission spectrum in energy group g. The solution $\underline{\Psi}$ is called neutron flux, it is a vector, its component Ψ_g gives the neutron flux in energy group g. The normal component of the net current is $J_g(x) = -D_g \boldsymbol{\nabla} \mathbf{n} \Psi_g(x)$, where $\mathbf{n}$ is the normal direction. The solution of problem (1) is positive provided $Q(r)$ and $q(x)$ are positive functions. Operator $\mathbf{B}$ may fix the flux, its normal gradient, or their linear combination.

Example 2: The Schrödinger equation.[11] A particle of mass m_0 moves in a potential field $V(x)$. The particle wave function is $\Psi(x)$, satisfying the Schrödinger equation:

$$\mathbf{A} = -\frac{\hbar^2}{2m_0}\nabla^2\Psi(r) + V(r)\Psi(r). \tag{4}$$

When a finite region is considered, the boundary condition may be $\Psi(x)=0$ on the boundary.

Analogous problems emerge in heat conductance, wave propagation, time-dependent diffusion, and in linear Boltzmann particle transport.[12]

Often we solve problem (1) iteratively. The solution and its normal derivatives are continuous at internal interfaces. With the help of the Green's function, the normal derivative on the boundary can be expressed by the solution on the boundary. The boundary currents can be expressed by boundary fluxes; this permits one to eliminate the gradients and to iterate only for the solution at internal boundaries. The continuity condition leads to a fix point problem; we have to find the fixed point of a set of equations.

III. SYMMETRIES OF BOUNDARY VALUE PROBLEMS

Consider a linear operator $\mathbf{A}$ acting on the function space $L_2(V)$, where V is a symmetric convex region. For a boundary condition problem, we assume linear operator $\mathbf{B}$ to form a function given on the boundary. The range of operator $\mathbf{A}$ includes V; the range of operator $\mathbf{B}$ includes the boundary ∂V. The operators commuting with $\mathbf{A}$ and $\mathbf{B}$ form the groups, G_A and G_B, respectively. Let G_V denote the symmetry group of V. The boundary value problem (BVP) is then determined by $(V, \mathbf{A}, \mathbf{B})$. We define the symmetry of the boundary value problem as follows.

Definition 1: A linear operator $\mathbf{P}$ *is said to be the symmetry of the boundary value problem* $(V, \mathbf{A}, \mathbf{B})$ *if (a)* $\mathbf{P}$ *transforms V into itself; (b)* $\mathbf{P}$ *and* $\mathbf{A}$ *commute:* $\mathbf{PA}=\mathbf{AP}$; *(c)* $\mathbf{P}$ *and* $\mathbf{B}$ *commute:* $\mathbf{PB}=\mathbf{BP}$. The symmetry group of the problem $(V, \mathbf{A}, \mathbf{B})$ is the intersection

$$G = G_V \cap G_A \cap G_B,$$

where G_V, G_A, and G_B are the symmetry groups of V, $\mathbf{A}$ and $\mathbf{B}$, respectively.

The symmetries of problem $(V, \mathbf{A}, \mathbf{B})$ form a group. This group may consist of a single element, the identity transformation, or of an infinite number of symmetries. This partly depends on operators $\mathbf{A}$ and $\mathbf{B}$. In a number of cases, it suffices to ensure that $\mathbf{P}$ commutes with $\mathbf{A}$; then it commutes with $\mathbf{B}$ as well. Often $\mathbf{B}$ is the identity transformation (Cauchy-type problem) or the normal gradient (Neumann-type problem), or a linear combination of these two operators. Most boundary value problems of reactor physics belong to one of these.

The symmetries of the boundary condition problem $(V, \mathbf{A}, \mathbf{B})$ suggest the application of point group theoretic techniques. Depending on the problem, the symmetries may form, for example, a discrete group or a Lie group. The symmetries can also be arranged into conjugacy classes[13] to which the results of group representation theory apply. Thus, by means of a projection operator (A6), any function can be decomposed into functions transforming according to the irreducible representations (irreps). First, we investigate some of the basic properties of the irreps.

According to projection operator (A6), the irrep f_i of a function f is a linear expression of f taken at different points in the range of f. Consequently, if f is n times differentiable, so is the irrep f_i. Let $f_i(\theta)$ be given in the interval $\theta \in [0, \pi/2n_F)$.

Definition 2: The $[0,\pi/2n_F)$ *interval is closed from the left and open from the right and is called the ground (where* n_F *is the number of faces of* V*).*

The direction $\theta=\pi/2n_F$ is a symmetry axis, and the character table tells us that f_i is an eigenvector of each symmetry operator, with the eigenvalue given in the ith row of the character table and in the column corresponding to the symmetry, thus, by applying a reflection through the face to f_i, we can obtain the function f_i in the range $[0,n_F)$. Here [,) denotes an interval closed from the left and open from the right. Now, applying the rotational symmetry of V, the $f_i(\theta)$ function is obtained for the entire $[0,2\pi]$ interval. When f_i is a component of a two-dimensional representation, more care is needed because the components may transform into each other. Thus, we arrive at the following statement.

Lemma 1: Let f be a given function on the boundary ∂V*. If* $f(\theta)$ *belongs to* C^n*, then its irreps* $f_i(\theta)$ *determined by (A6) also belong to* C^n*. The irrep* $f_i(\theta)$*,* $0\leqslant\theta\leqslant 2\pi$ *is uniquely given by its value in the ground* $[0,\pi/2n_F)$*.*

Thus, the ground is the quotient manifold V/G. The first question is the following: If we have a decomposition of the boundary condition, do we also have a decomposition of the solution? What can we gain by such a decomposition of the solution?

Basic Lemma. Let the problem

$$\begin{aligned} A\varphi&=0, \quad \text{in } V,\\ B\varphi&=f, \quad \text{on } \partial V, \end{aligned} \tag{5}$$

be given, and the linear operators **A** *and* **B** *be such that (a) when* $f=0$*, the only solution is* $\varphi=0$*, i.e., the homogeneous problem has only the identically zero function as the solution; (b) the null space of operator* **B** *is empty, i.e., if* $\mathbf{B}\varphi=0$ *then* $\varphi=0$*.*

Then, if f transforms according to the ith irreducible representation on the boundary, the solution transforms according to the ith irreducible representation inside V*.*

Proof: Let $f=f_i$ be a one-dimensional irreducible representation and φ_i the corresponding solution. Then, for any symmetry **P** we have $\mathbf{P}f_i=\alpha_i f_i$. Applying **P** to the second equation and making use of the commutation of **P** and **B**, we get $\mathbf{PB}\varphi_i=\mathbf{BP}\varphi_i=\alpha_i\mathbf{B}\varphi_i=\alpha_i f_i$. Applying **P** to the first equation, we get $\mathbf{PA}\varphi_i=\mathbf{AP}\varphi_i=0$. Multiplying the first equation by α_i, and because **A** is linear, we also have $\mathbf{A}(\alpha_i,\varphi_i)=0$. Thus, we have the following equations for $(\mathbf{P}\varphi_i-\alpha_i\varphi_i)$:

$$\begin{aligned} \mathbf{A}(\mathbf{P}\varphi_i-\alpha_i\varphi_i)&=0, \quad \text{in } V\\ \mathbf{B}(\mathbf{P}\varphi_i-\alpha_i\varphi_i)&=0, \quad \text{on } \partial V, \end{aligned} \tag{6}$$

but according to assumption (a), the only solution is $(\mathbf{P}\varphi_i-\alpha_i\varphi_i)=0$, which proves the statement for a one-dimensional representation. Note that the second equation above is not true unless assumption (b) is met.

Let $f=f_i$ be a component of a two- or three-dimensional representation, i.e., it transforms as

$$\mathbf{P}f_i=\sum_k P_{ik}f_k.$$

The boundary condition (BC) for each component k is given by $\mathbf{B}\varphi_k=f_k$. Multiplying that equation by P_{ik} and summing over k, **B** being linear, we have

$$\mathbf{B}\left(\sum_k P_{ik}\varphi_k\right)=\sum_k P_{ik}f_k.$$

Multiplying the second equation of the problem by **P**, and using the linearity of **B**, we have

$$\mathbf{PB}\varphi_i = \mathbf{BP}\varphi_i = \sum_k P_{ik} f_k .$$

Subtracting the last two equations, we get, by means of assumption (b),

$$\mathbf{B}\left(\mathbf{P}\varphi_i - \sum_k P_{ik}\varphi_k \right) = 0;$$

thus φ_i transforms as the ith column of a multidimensional representation on ∂V.

As to the transformation rules in V, if we multiply the first equation in (6) by P_{ik}, sum over k on one hand, and on the other we multiply the equation again by $\mathbf{P}$ and subtract the two. We have

$$\mathbf{A}\left(\mathbf{P}\varphi_i - \sum_k P_{ik}\varphi_k \right) = 0.$$

The last two equations form a BVP, the only solution of which is identically zero. This completes the proof. It should be noted that the second step of the proof also includes the first step.

Corollary: By virtue of the linearity of the problem, when a volumetric source Q is included, the solution to

$$\mathbf{A}\Phi_Q = Q, \quad in \ \ V,$$

$$\mathbf{B}\Phi_Q = f, \quad on \ \ \partial V,$$

is given by $\Phi_Q = \Phi + \Psi$, *where* Φ *is the solution of (5) and* Ψ *is the solution of*

$$\mathbf{A}\Psi = Q, \quad in \ \ V,$$

$$\mathbf{B}\Psi = 0, \quad on \ \ \partial V.$$

Applying the argument used in the proof of the Basic Lemma, we arrive at the following result. Let q_i be the ith irreducible component of Q. Then, an irreducible decomposition of the solution is $\Phi_Q = \Psi_i + \Phi_{i,}$ where

$$\begin{aligned} \mathbf{A}\Psi_i &= q_i \quad in \ \ V, \\ \mathbf{B}\Psi_i &= 0, \quad on \ \ \partial V, \end{aligned} \tag{5'}$$

and

$$\begin{aligned} \mathbf{A}\Phi_i &= 0, \quad in \ \ V, \\ \mathbf{B}\Phi_i &= f_i , \quad on \ \ \partial V. \end{aligned} \tag{5''}$$

The Basic Lemma allows one to find suitable representations for the boundary values on ∂V and for the solution in V. The general prescription is to find suitable representations with which a general function—given along the boundary—is decomposed into irreps, and then to solve the above equations for the components. To this end, projector (A6) is applied. Here we set out the following notation. A point x on the ∂V boundary of the simply connected volume V is characterized by an angle θ measured from a suitable center inside V. We assume, furthermore, that ∂V consists of n_F faces. When V is a regular triangle, square, hexagon, or pentagon, the faces are of equal length or area and there are $2n_F$ symmetry transformations leaving V invariant. Now the

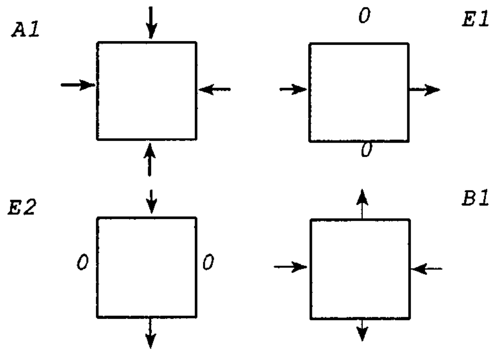

FIG. 1. Irreducible entering current patterns for a square node.

symmetries of the problem transform one point of the boundary into another boundary point. Thus, projector (A6) applied to a function $f(\theta)$ will give linear combinations of $f(\theta)$, $i=1, 2n_F$, and, the θ_i points belong to one orbit.

Lemma 2: The irreps of a boundary value $f(\theta)$ take the form

$$f_i(\theta)=m_i(\theta \bmod \pi/n_F)e_i\left(\left[\frac{\theta}{\pi}n_F\right]\right), \tag{7}$$

where [ ] denotes the entire part, and the function e_i ( ) takes only integer values, furthermore, the m_i function is not identically zero on the ground.

Proof: The irrep $f_i(\theta)$ is projected out according to Eq. (A6), therefore it is a linear combination of $P_{O_i}*f(\theta)=f(\mathbf{O}_i*\theta)$, $\mathbf{O}_i\in G$. The transformed points under the elements of group G form an orbit. The orbits crossing different points of the ground are equivalent, hence the entire ground can be treated as a single unit. $f_i(\theta)$ will transform as the ith column of an irreducible representation only when $f_i(\theta)$ takes the value in the ground multiplied by a constant for each transform of the ground. In other words, $f_i(\theta)$ is given by a nonzero distribution in the ground, and the amplitudes in each transform along the path V/G. This is precisely Eq. (7), thus, the lemma is proven. As a consequence of Lemma 2, we may characterize the irreps of a function given along the boundary by a vector e_i, which is composed of the $2n_F$ values taken by the function $e_i(\theta)$, $0\leqslant\theta\leqslant 2\pi$.

The irreps of the boundary condition have an important physical meaning: The irreps then depend on the shape of V. Since physical meaning is more transparent in a simple geometrical shape, let us consider the irreps of a square shaped V (see Fig. 1). They represent characteristic ambiances into which the volume under consideration is imbedded. The first one represents a homogeneous ambiance, the second and third one an x and a y directed gradient, respectively, and the fourth one a second derivative. To illustrate the impact of the geometrical shape of V on the irreducible boundary condition components, we introduce the irreps for a regular hexagon-shaped volume (see Fig. 2). Evidently, the six faces allow for a wider variety of ''gradients.''

We would just mention here that when we determine the energy of a charged body in the presence of an external electric potential field, we use an analogous decomposition.[14] The $\varphi(x)$ potential obeys the Poisson equation. In homogeneous space, the symmetry group of the Laplace operator is the rotation group; the irreducible basis functions are the spherical harmonics. If the charge distribution $e(x)$ and the potential $\varphi(x)$ are expanded in terms of the spherical harmonics Y_{lm}, we get

$$\varphi(x)=\sum_{l,m} a_{lm}Y_{lm}; \quad e(x)=\sum_{k,n} Q_{kn}Y_{kn},$$

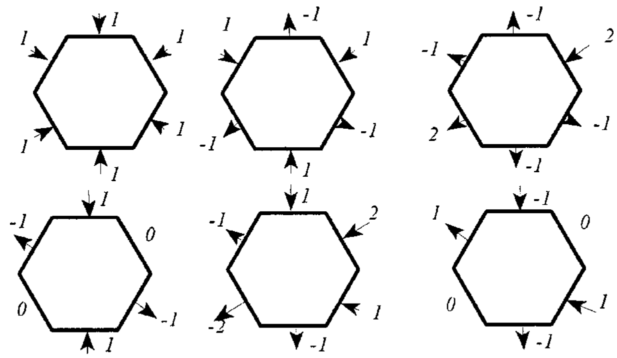

FIG. 2. Irreducible entering current patterns for a regular hexagon.

$$U=\int_V \varphi(x)e(x)dx=\sum_{l,m} a_{lm}Q_{lm}, \tag{8}$$

because the spherical harmonics form an orthonormal set. The terms in the sum correspond to the multipole (monopole, dipole, quadrupole, etc.) expansion. In Eqs. (5′) and (5″), the analogous symmetry components (irreducible representations) of the components are given. In those expressions, however, the external components (f_i and q_i) and the internal components (Ψ_i and Φ_i) are not separated explicitly.

IV. THE ITERATION

Now let us turn to the iterative solution of problem (1): more precisely, instead of problem (1), we deal with the associated eigenvalue problem, which has a distinguished role in neutron physics. Consider the following problem:

$$\begin{aligned} \mathbf{A}(k)\Phi(r)&=0, \quad r\in V,\\ \mathbf{B}\Phi(x)&=0, \quad x\in\partial V, \end{aligned} \tag{9}$$

where $\mathbf{A}(k)=\mathbf{A}^{(1)}+1/k^*\mathbf{A}^{(2)}$, and V is convex and composed of homogeneous regions:

$$V=\bigcup_{j=1}^{N} V_j,$$

where regions j and j' are disjoint except for the joint boundary $\partial V_{jj'}$. N is the number of homogeneous regions. Let ∂V_j denote the boundary of region (or node) j; then

$$V_j\cap V_{j'}=\partial V_{jj'}=(\partial V_j)\cap(\partial V_{j'}).$$

A boundary ∂V_j is called an external boundary, if

$$\partial V_j\cap\partial V=\partial V_j.$$

If ∂V_j is not external, then it is an internal boundary. $\mathbf{A}(k)$ and $\mathbf{B}$ are linear operators. Both $\mathbf{A}(k)$ and $\mathbf{B}$ may involve functions as coefficients.

In volume j, the equation to be solved is written as

$$\mathbf{A}_j(k)\Phi_j(r)=0.$$

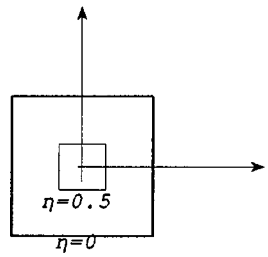

FIG. 3. Coordinate system in a node.

where subscript j refers to region j in which the coefficients associated with each $\mathbf{A}_j$ may be different. We assume, furthermore, the existence of a value of parameter k which turns the largest eigenvalue of $\mathbf{A}(k)$ into zero. The associated eigenfunction is non-negative.

On external boundaries, operator $\mathbf{B}$ involves only linear expressions:

$$\mathbf{B}\Phi = a_j \int_{\partial V_j} \Phi_j(x) dF_j + b_j \int_{\partial V_j} \partial_n \Phi_j(x) dF_j = 0.$$

Here a_j and b_j are constants. In problem (1), we have two types of approximations: the first is of the neutron distribution function inside the node, the second of the neutron distribution function on the surface of the node. Let the basis function set in node n be $[f_i(r),\ i=1,...,N_v]$ and on the surface ∂V_n $(h_j(x),\ j=1,...,N_s)$. Let us expand the solution inside a given node as

$$\Phi(r) = \sum_{i=1}^{N_v} \zeta_i f_i(r), \tag{10}$$

and on the surface of the node as

$$\Phi(x) = \sum_{i=1}^{N_s} \chi_j h_j(x). \tag{11}$$

Here the node index is suppressed; the N_v and N_s are the same for every node. The space variable inside the node is r; on the surface it is x. Thus, in general, we need a new coordinate (η) expressing how far a point is from the boundary. We use the coordinate system of Fig. 3. The boundary is given by $\eta=0$ and $\eta=1$ is the center of the node. Thus, in general, $r=(x,\eta)$ and on the boundary $\eta=0$. The $\eta=\text{const}$ lines are parallel to the volume's surface.

The two approximations [i.e., Eqs. (10) and (11)] are not independent, because when η tends to zero, the trial functions $f_i(r)$ must be expressible with the help of the trial functions $h_j(x)$:

$$f_i(x,0) = \sum_{j=1}^{N_s} b_{ij} h_j(x); \tag{12}$$

thus

$$\chi_j = \sum_{i=1}^{N_v} b_{ij} \zeta_i . \tag{13}$$

If condition (12) does not hold, the approximation is not self-consistent.

Let us express the volumetric source with the help of the basis functions as

$$S(r)=\sum_{i=1}^{N_v} f_i(r)s_i. \tag{14}$$

In this expansion and in the following we use the dot product, defined as

$$(f,g)_{V_n}=\int_{V_n} f(r)g(r)dV. \tag{15}$$

With the matrices

$$\mathbf{A}_{ij}=(f_i(r),\mathbf{A}f_j(r))_{V_n}, \tag{16}$$

and

$$\mathbf{B}_{ij}=(h_i(x),\mathbf{B}f_j(x,0))_{\partial V_n}=b_{ij}, \tag{17}$$

problem (1) is cast into the following form (see Ref. 8 for details):

$$\mathbf{A}\zeta=\mathbf{s}-\mathbf{M}\chi, \tag{18}$$

where ζ denotes the solution to the source problem,

$$\mathbf{A}\zeta=\mathbf{s}, \quad \textit{inside} \;\; V;$$

$$\mathbf{B}\zeta=0, \quad \textit{on} \;\; 2 \;\; \partial V, \tag{19}$$

and a particular solution to the nonhomogeneous problem is $\mathbf{M}_\chi$.

$$\mathbf{AM}\chi=0, \quad \mathbf{BM}\chi=\chi. \tag{20}$$

From the second line of Eq. (20), we get $\mathbf{M}=\mathbf{B}^{-1}$. This assumes that matrix b_{if} in expression (12) is invertible. In that case $\mathbf{M}_\chi$ is an interpolation: it interpolates the solution in the interior of the node from the values given on the boundary. The iteration may sweep through the nodes, starting out from a sensible initial guess. The first step in a new sweep is to solve Eq. (20) in the actual node, so we first collect χ on the boundary of the actual node using the continuity conditions. The new solution is $(\zeta+\mathbf{M}_\chi)$. From this, we determine the updated boundary values and pass on to the next node.

The iterative solution based on Eqs. (18)–(20) will be explored by symmetry considerations.[15] If there are transformations mapping the node into itself, these transformations form a group commuting with operators **A** and **B**. Such a group generates a splitting of the solution space into linearly independent subspaces. Operators **A** and **B** transform the elements of each subspace only among themselves, thus, the mentioned subspaces remain linearly independent when applying either operator to an element of the solution space. The main idea is as follows. Since the currents and fluxes are continuous at material interfaces, so is their arbitrary linear combination. We introduce the entering current $J^+=1/4(\Phi+2J)$ and the exiting current $J^-=1/4(\Phi-2J)$. We collect the fluxes, currents, and entering and exiting currents into vectors of n_F elements:

$$\underline{J}=(J_1,\ldots,J_{n_F}); \quad \underline{\Phi}=(\Phi_1,\ldots,\Phi_{n_F}),$$

$$\underline{J}^+=(J_1^+,\ldots,J_{n_F}^+); \quad \underline{J}^-=(J_1^-,\ldots,J_{n_F}^-).$$

If the iteration goes as

$$\underline{I}^{+}=\mathbf{T}\underline{I}^{-}=(1+2\mathbf{R})(1-2\mathbf{R})^{-1}\underline{I}^{-} \tag{21a}$$

or

$$\underline{J}=\mathbf{R}\underline{\Phi}, \tag{21b}$$

and there is a symmetrical matrix $\mathbf{M}$ such that $\mathbf{MR}=\mathbf{RM}$, then the eigenvectors of $\mathbf{M}$ form a suitable basis function set to expand the quantities involved in the iteration:

$$\mathbf{M}\underline{e}_i=\lambda_i\underline{e}_i,$$

$$\underline{\Phi}=\sum_i f_i\underline{e}_i; \quad \underline{J}=\sum_i g_i\underline{e}_i,$$

and on the new basis, $\mathbf{R}$ and $\mathbf{T}$ will be diagonal:

$$g_i=(\mathbf{MRM}^{-1})_i f_i.$$

The iteration proceeds as follows. From the flux on the boundary, the solution inside is determined. In the next step, the new current on the boundary is calculated from the inside solution. If there is an i such that $f_i\neq 0$ but $g_i=0$, then we get $\mathbf{R}=0$ and $\mathbf{T}=1$, indicating that in $(\underline{J}^{+},\underline{J}^{-})$ no convergence will occur.

Our purpose in the investigation is to find a relationship between two apparently independent approximations. What is the relationship between the approximation of the volumetric and surface source terms? The numerical method[16] implemented in VARIANT allows for two independent approximations for the two terms, but in certain cases the procedure did not converge. In these cases a nonvanishing, small error was observed[17] in the solution, furthermore, the phenomenon disappeared when a sufficiently high-order approximation was applied inside the node.

V. APPLICATION TO POLYNOMIAL APPROXIMATION

A. Square node

In a symmetric volume, the elements of the group G will play the role of matrix $\mathbf{M}$. The eigenvectors $\underline{e}_i$ are simultaneous eigenvectors of the elements in G. Eventually, the symmetries of the node permit one to classify the solution into different components. According to Eqs. (5′) and (5″) in Sec. II, the internal solution contains a given component only if that component is present either in the source or in the boundary condition. If a component is missing from the solution but is present in the boundary condition, it may cause numerical problems. This phenomenon is caused by solutions of the problem, which have zero average both over the node and over the entire node boundary. Such solutions do not contribute to the balance but may lead to an erroneous solution.

Before assessing the boundary value problems, a few general statements are set forth. The symmetry analysis of the solution of BVPs is given in Sec. III. We repeat briefly two things.

(i) An irreducible component is simultaneously an eigenfunction of every symmetry, and any function can be decomposed into irreducible components.

(ii) Applying the symmetries to a function, we get the same function at symmetric positions.

An irreducible component is a linear combination of the function under consideration taken at symmetric positions. Now we introduce the response matrix in order to eliminate the boundary currents and to set up an iteration for the boundary fluxes.

Definition 1: Let $\mathbf{u}=(u_1,\ldots,u_{nF})$ and $\boldsymbol{\nu}=(v_1,\ldots,v_{nF})$ be the respective values of function u and v on the n_F faces. Both functions are formed from the flux by linear operations commuting with the symmetries of the node. The matrix $\mathbf{R}$ relates $\mathbf{u}$ and $\boldsymbol{\nu}$ as

$$\mathbf{u}=\mathbf{R}\mathbf{v}, \tag{22}$$

TABLE I. Irreps of spatial polynomials in a square.

Vector	Polynomials
e_1	$1,(x^2+y^2),(x^4+y^4),x^2y^2$
e_2	(x^3y-y^3x)
e_3	$(x^2-y^2),(x^4-y^4)$
e_4	$xy,(x^3y+y^3x)$
e_5	x,x^3
e_6	xy^2
e_7	x^2y
e_8	y,y^3

and is called a response matrix.

Proposition 1: If V is a symmetric region, then any response matrix, as in Definition 1, is diagonalized by expressing both $\mathbf{u}$ and $\boldsymbol{\nu}$ by their irreducible components. If there are equivalent irreducible components, the response matrix will be block diagonal.

Below, it is shown that if a subspace is present in the approximation to the solution on the boundary, but not in the representation inside the node; the iteration scheme (18) will not converge because there is a subspace whose contribution does not decay during the iteration; its amplitude does not decrease with the number of iterations.

Proposition 2: Let us decompose the solution space $\mathcal{L}$ into irreducible subspaces as $\mathcal{L}=(\mathcal{L}^{(I)},\ldots,\mathcal{L}^{(K)})$. If there is a subspace $\mathcal{L}^{(m)}$, which is present in the boundary representation but missing from the internal representation, then the component corresponding to $\mathcal{L}^{(m)}$ represents a subspace of the entering current vector that will not decrease in the iteration (18).

Proof: The response matrix can be calculated in the following manner. The flux on the boundary is expressed in terms of the approximate polynomials as Eq. (4), and by the assumptions there is at least one nonzero coefficient in the expansion. According to the Basic Lemma, the solution inside the node will belong to the same subspace as the boundary condition. But, by the assumptions this subspace is empty inside the node, in the expansion (10) no $f_i(r)$ function belongs to this subspace. Consequently, the net current calculated from the flux inside will, according to Proposition 1, be zero. Thus, the corresponding response matrix $\mathbf{R}$ is also zero. The response matrix $\mathbf{T}$ [see Eq. (21a)] connecting entering currents to exiting currents and $\mathbf{R}$ are related as

$$\mathbf{T}=(1+2\mathbf{R})(1-2\mathbf{R})^{-1}, \tag{23}$$

so now $\|\mathbf{T}\|=1$, which means that applying n times $\mathbf{T}$ on an initial vector $\underline{I}^+$ we do not get smaller amplitudes if the initial vector has a component in the subspace m as stated.

Now we turn to the application of these observations to square-shaped nodes. The above cited statements have important consequences in the problem considered. First, a solution inside the node with given symmetry properties exists only if either the source or the boundary value has those symmetry properties. Second, the response matrix will be zero in a given irrep, if the solution has no matching component. This is because in that case the considered component of the solution is not zero on the boundary, but the solution inside V is zero. Consequently, the gradient calculated from it is also zero. Therefore, the corresponding element of the response matrix (21b) is zero, the response matrix (21a) is unity. When we determine the solution in a given subspace by means of polynomials, it should be verified whether the polynomials have at least one component in the subspace under consideration. Table I gives the symmetry components of the fourth-order polynomial in a square. There are six different subspaces; four of them are one-dimensional ones, and two, two-dimensional ones. According to Lemma 2, we can characterize their basis vectors by vectors $\underset{\sim}{e}_1,\ldots,\underset{\sim}{e}_8$. Four of them, viz. $\underset{\sim}{e}_1,\ldots,\underset{\sim}{e}_4$, are one-dimensional ones, and two, two-

TABLE II. Irreps of spatial moments on the boundary of a square.

Vector	Moment	Values at faces
e_1	0,2	(1,1,1,1)
e_2	1	(1,1,1,1)
e_3	0,2	(1,−1,1,1)
e_4	1	(1,−1,1,−1)
e_5	0,2	(0,1,0,−1)
e_6	1	(1,0,−1,0)
e_7	1	(0,1,0,−1)
e_8	0,2	(1,0,−1,0)

dimensional pairs (e_5, e_8) and $(\underline{e}_6, \underline{e}_7)$.
The irreps of the boundary conditions are given in Table II.
The response matrix will have the structure

$$I_i^- = \mathbf{R}_i I_i^+, \quad i = 1,...,4, \tag{24}$$

$$I_i^- = \mathbf{R}_5 I_i^+, \quad i = 5,...,8. \tag{25}$$

Here I_i^+ is the ith component of the entering current, proportional to vector $\underline{e}_i$. $\mathbf{R}_i$ denotes the ith block of the RM. Since $\underline{e}_5$ and $\underline{e}_6$ as well as $\underline{e}_7$ and $\underline{e}_8$ are equivalent, here the RMs are the same, according to Theorem A.6.

While the boundary condition has at least one component in each subspace when linear polynomials are used along the four faces, the linear approximation leaves four empty subspaces in Table I. In other words, to calculate the response matrix, a higher-order approximation is needed inside the node. To demonstrate the effect of an empty irrep inside, while the boundary value has that irrep, let us consider the following example.

With, at most, linear polynomials, on the boundary each subspace will have at least one component, (see Table II). At the same time inside the square, the first polynomial, where each subspace will have at least one component, is the fourth-order polynomial (see Table III).

This observation is general. A low-order approximation on the boundary is able to furnish a large number of overtones inside the node. This is because the elementary solutions to the diffusion equation are exponential functions and to fulfill the simplest boundary condition the solution will contain a number of exponential functions. Exploiting the connection between the Fourier transformation and the irreducible components, we remark here only that a polar coordinate system (r, α) can be introduced in which the irreducible component will contain $\cos(k\alpha)$ and $\sin(k\alpha)$, where k depends on the irrep and all $k(\mathrm{mod})n_F=\mathrm{const}$ modes belong to the same irrep. When the node is a circle, all modes form a separate irrep.

TABLE III. Irreducible vectors inside a square in increasing order of polynomials.

Vector\Order	0	1	2	3	4
e_1	1	1	$1,(x^2+y2)$	$1,(x^2+y^2)$	$1,(x^2+y^2),(x^4+y^4),x^2y^2$
e_2	⋯	⋯	⋯	⋯	(x^3y-y^3x)
e_3	⋯	⋯	(x^2-y^2)	(x^2-y^2)	(x^2-y^2)
e_4	⋯	⋯	xy	xy	$xy,(x^3y+y^3x)$
e_5	⋯	x	x	x,x^3	x,x^3
e_6	⋯	⋯	⋯	xy^2	xy^2
e_7	⋯	⋯	⋯	x^2y	x^2y
e_8	⋯	y	y	y,y^3	y,y^3

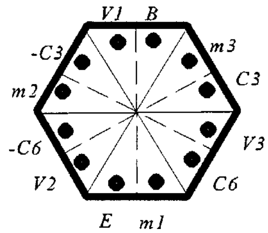

FIG. 4. Symmetries of a regular hexagon.

B. Hexagonal geometry

The hexagonal node is shown in Fig. 4. The regular representation is expressed by 12 element vectors. C_6 denotes a rotation by 60°, $-C_6$ by $-60°$. C_3 and $-C_3$ denote 120° rotations and its inverse, respectively. Let V_i denote reflections through planes connecting corners, m_i connecting midplane points. The irreps are projected out with the help of the character table. Table IV gives the irreducible vectors. Since there are two equivalent two dimensional subspaces in a regular representation, the two-dimensional representations E_1 and E_2 are associated with two subspaces distinguished by a superscript. The angular term gives the angular component of an analytical solution which transforms according to the given irrep.

Table IV has been derived with the help of the character table of group C_{6v} given in GAP.[13] The sectors belonging to subspace Γ_5 transform as coordinates (x,y), or in polar coordinates as $\cos(\alpha)$ and $\sin(\alpha)$. The subspace Γ_6 transforms as (x^2-y^2,xy), or, as $\cos(2\alpha)$ and $\sin(2\alpha)$. The vectors given in Table IV are the x and y components of the transforms of a point $(r,\ \alpha)$.

The irreps of the moments on the boundary are given in Table V. Here it is assumed that we are given a zeroth, first, or second-order power of the tangential coordinate on each face; the coefficients may be different on the six faces.

To assess the numerical method, we also need the irreps of the approximating polynomials inside the node. The irreducible components of the spatial moments are given in Table VI. It is seen that none of the interpolants contribute to the subspace Γ_2. This is because that subspace contains functions varying as $\sin(6\alpha)$ and that component appears first in sixth-order polynomials. Thus, any lower-order approximation will contain a nonvanishing subspace and cause a convergence problem.

IV. CONCLUSIONS

We investigated an iteration, Eqs. (18)–(20), which is widely used to solve the diffusion and transport equation. In the course of the iteration, the solution is iteratively improved in the nodes.

TABLE IV. Irreducible vectors of 12 elements in a regular hexagon.

Space	Dimension	Angular term	Notation	Elements
Γ_1	1	$\cos(6k\alpha)$	e_1	(1,1,1,1,1,1,1,1,1,1,1,1)
Γ_2	1	$\sin(6k\alpha)$	e_2	(1,−1,1,−1,1,−1,1,−1,1,−1,1,−1)
Γ_3	1	$\sin(6k+3)\alpha$	e_3	(1,−1,−1,1,1,−1,−1,1,1,−1,−1,1)
Γ_4	1	$\cos(6k+3)\alpha$	e_4	(1,1,−1,−1,1,1,−1,−1,1,1,−1,−1)
$\Gamma_5^{(1)}$	2	$\cos(6k+1)\alpha$	e_5	(2,2,1,1,−1,−1,−2,−2,−1,−1,1,1)
		$\cos(6k+4)\alpha$	e_6	(0,0,1,−1,−1,1,0,0,−1,1,1,−1)
$\Gamma_5^{(2)}$	2	$\cos(6k+2)\alpha$	e_7	(−2,2,−1,1,1,−1,2,−2,1,−1,−1,1)
		$\cos(6k+5)\alpha$	e_8	(0,0,1,1,1,1,0,0,−1,−1,−1,−1)
$\Gamma_6^{(1)}$	2	$\sin(6k+1)\alpha$	e_9	(2,2,−1,−1−1,−1,2,2,−1,−1,−1,−1)
		$\sin(6k+4)\alpha$	e_{10}	(0,0,1,−1,−1,1,0,0,1,−1,−1,1)
$\Gamma_6^{(2)}$	2	$\sin(6k+2)\alpha$	e_{11}	(−2,2,1,−1,1,−1,−2,2,1,−1,1,−1)
		$\sin(6k+5)\alpha$	e_{12}	(0,0,1,1,−1,−1,0,0,1,1,−1,−1)

TABLE V. Irreducible vectors on the boundary of a regular hexagon.

Vector	Moments	Values at faces
e_1	0,2	(1,1,1,1,1,1)
e_2	1	(1,1,1,1,1,1)
e_3	1	$(1,-1,1,-1,1,-1)$
e_4	0,2	$(1,-1,1,-1,1,-1)$
e_5	0,2	$(2,1,-1,-2,-1,1)$
e_6	0,2	$(0,1,1,0,-1,-1)$
e_7	1	$(2,1,-1,-2,-1,1)$
e_8	1	$(0,1,1,0,-1,-1)$
e_9	0,2	$(2,-1,-1,2,-1,-1)$
e_{10}	0,2	$(0,1,-1,0,1,-1)$
e_{11}	1	$(2,-1,-1,2,-1,-1)$
e_{12}	1	$(0,1,1,0,-1,-1)$

If the node has symmetries, which commute with the matrices of the iteration, we are able to decompose the solution vector into linearly independent components. Those components are not mixed by the response matrices. This situation is fairly common; the response matrix commutes with the symmetries of the node. The technique presented can be described as follows. The iteration endeavors to solve a linear set of equations with matrix **A**. Let us assume that a symmetric matrix, **M** is given and **M** commutes with **A**. We may use the eigenvectors $\underset{\sim}{e}_i$ of **M** to span the solution space. The iteration does not mix the the eigenvectors $\underset{\sim}{e}_i$, so the solution goes separately for each eigenvector. If the boundary condition has a component proportional to the eigenvectors $\underset{\sim}{e}_i$ but the solution inside the node does not, the iteration does not converge.

The solution of the diffusion equation often follows this setup. We have a value on the boundary from which we derive a solution inside the node, and, from the solution, a new boundary value is obtained. The symmetries permit us to decompose the solution vector and have a closer look at a step of the iteration. When the internal solution has no component in a subspace, but the boundary condition has, the iteration will not converge because any tiny component of the initial guess in that subspace will remain the same.

It should be emphasized, however, that the component under consideration may really be tiny. In such cases, the error will decrease until that component becomes dominant. The convergence difficulties predicted in the present paper have been observed in an actual program.[17] There, the boundary term and the source term were approximated by polynomials. Low-order polynomials of the boundary term cover every subspace, whereas inside the node higher-order polynomials should be used to avoid convergence problems.

TABLE VI. Irreducible vectors of interpolating polynomials inside a regular hexagon.

Vector\Order	0	1	2	3	4
e_1	1	⋯	(x^2+y^2)	⋯	$(x^2+y^2)^2$
e_2	⋯	⋯	⋯	⋯	⋯
e_3	⋯	⋯	⋯	$y(y^2-3x^2)$	⋯
e_4	⋯	⋯	⋯	$x(x^2-3y^2)$	⋯
e_5	⋯	x,y	⋯	$x(x^2+y^2)$	
e_6	⋯	x,y	⋯	$y(x^2+y^2)$	
e_7	⋯	x,y	⋯		
e_8	⋯	x,y	⋯		
e_9	⋯	⋯	(x^2-y^2)		$(5x^4-6x^2y^2-3y^4),x^3y$
e_{10}	⋯	⋯	xy		y^3x
e_{11}	⋯	⋯	⋯		$(-x^4+6x^2y^2-y^4)$
e_{12}	⋯	⋯	$(x^2-y^2),xy$		

APPENDIX A: GROUP THEORY PRIMER

In this appendix we summarize the results of group theory applied throughout the present work. The results cited below are available in standard textbooks.

Definition A.1: Let us consider function $f(x)$. The transformation operator $\mathbf{P}_O$ associated with matrix O is defined by the following identity in x:

$$\mathbf{P}_o f(x) \equiv f(O^{-1}x). \tag{A1}$$

With this definition, the set of symmetry operations (symmetries) is made isomorphic with the set of matrices associated with the symmetries.

Definition A.2: $\mathbf{P}_O$ is called a symmetry of region V if $\mathbf{O}$ maps V onto itself. The symmetries of V form a group; the group operation is $\mathbf{P}^*_{O1}\mathbf{P}_{O2}=$apply first $\mathbf{P}_{O2}$ then apply $\mathbf{P}_{O1}$.

Definition A.3: An orbit is a set of points g^*x, where $x\in V$ is given, $g\in G$. There is an induced equivalence relation between the points of V, with x being equivalent to y if they are in the same orbit of G.

Definition A.4: Let V/G denote the the set of orbits of G. V/G is referred to as the quotient (region). The projection π: $V\rightarrow V/G$ associates with each x in V its equivalent class $\pi(x)$, $x\in V/G$.

Definition A.5: A set of matrices under matrix multiplication $\{O_1,O_2,\ldots,O_n\}$, which is homomorphic with the group $\{O_1,O_2,\ldots,O_n\}$, is said to be a representation of the group.

Definition A.6: A regular representation associates matrices of order h with the group G.

Definition A.7: A representation is said to be reducible if an equivalent representation[18] exists in which each matrix O_i has the form

$$\mathbf{O}_i=\begin{pmatrix} \mathbf{A}_i & \mathbf{C}_i \\ 0 & \mathbf{B}_i \end{pmatrix}, \tag{A2}$$

If no such representation exists, the representation is said to be an irreducible representation of the group.

Definition A.8: A character table is associated with every group. The character table is a square table. It contains n_c rows and columns, where n_c is the number of conjugacy classes. The first column gives the dimension of the row, sometimes also known as representation.

Theorem A.1: (FALICOV) The h-dimensional space in which the regular representation acts by h-order matrices can be split the following manner. There are n_c subspaces transforming among themselves. If the dimension of the ith subspace is l_i then there are l_i equivalent subspaces. The matrices of a regular representation do not mix the equivalent subspaces.

Definition A.9: $\mathbf{O}$ is called a symmetry of operator $\mathbf{A}$ if $\mathbf{AO}=\mathbf{OA}$.

Definition A.10: Let $\{O_i\}$ be an irreducible representation and let $\{f_i^{(1)},f_i^{(2)},\ldots,f_i^{(li)}\}$ be l_i eigenfunctions of the symmetry operations for which

$$\mathbf{O}f_i^{(k)}=\sum_{j=1}^{l_i} O_{jk}f_i^{(j)}, \tag{A3}$$

holds for $\mathbf{O}=\mathbf{O}_l,\ldots,\mathbf{O}_h$. A function $f_i^{(k)}$ is said to belong to the kth row of the irreducible representation $\mathbf{O}_i$ if there exist partner functions $\{f_i^{(1)},f_i^{(2)},\ldots,f_i^{(k-1)}f_i^{(k+1)},\ldots,f_i^{(li)}\}$ such that the above equation is satisfied.

Theorem A.2: *Let* $\mathbf{S}$ *commute with all* $\mathbf{O}_i$. *Let* $f_i^{(i)}$ *and* $g_i^{(j')}$, *belong to different irreducible representations. Then*

$$(f_i^{(j)},g_{i'}^{(j')})=\frac{1}{l_i}\delta_{ii'}\delta_{jj'}\sum_j (f_i^{(j)},g_{i'}^{(j')}) \tag{A4}$$

and

$$(f_i^{(j)}, \mathbf{S}g_{i'}^{(j')}) = \frac{1}{l_i}\delta_{ii'}\delta_{jj'}\sum_j (f_i^{(j)}, \mathbf{S}g_{i'}^{(j')}). \tag{A5}$$

Proof: Ref. 19, pp. 115, 116.

The following theorem, taken from Ref. 1, pp. 52–55, is a summary of the relationship between the eigenspace of an operator (in our case **A**) and a group commuting with **A**. The relationship is twofold. The eigenfunctions of **A** may serve as basis functions of the irreducible representations. The symmetry operators on that basis are represented by matrices and the eigenvectors of those matrices are linear combinations of the eigenfunctions of **A**, furthermore, they are basis vectors of an irreducible representation. The reverse statement is the following: If irreducible functions are used to represent operator **A** by matrices, the resulting matrix will be diagonal; the elements belonging to partner functions of a given irrep are the same.

Theorem A.3:

(1) *The eigenfunctions of operator* **A** *generate a representation of G*.
(2) *Linear transformations to new eigenfunctions generates a representation equivalent to the original*.
(3) *If the eigenfunctions are orthonormal, then operator O is merely the matrix associated with O*.
(4) *The representation so generated is unitary*.
(5) *If the degeneracy is normal, the representation is irreducible*.
(6) *An arbitrary function* Φ *in the space of* **A** *can be used to construct an invariant subspace by forming the operation* $O\Phi$ *for all O in G*.
(7) *Functions that transform in accordance with two different irreducible representations of G are orthogonal*.
(8) *Any function in the space of* **A** *can be decomposed into a linear combination of functions transforming according to irreps of G*.

The operator,

$$\mathbf{P}^{(i)} \equiv \frac{l_i}{h}\sum_O \chi^{(i)}(O)\mathbf{P}_O, \tag{A6}$$

projects out the component transforming in accordance with the ith representation of group G.

Theorem A.4: *Let* Φ_i *be irreducible,* $i=1,...,n_c$ *(the number of classes). Then the matrix,*

$$\langle \Phi_i A \Phi_j \rangle,$$

is diagonal. The elements belonging to the components of a multi- (i.e., two- or three-) dimensional representation are equal.

Proof: This statement is an immediate consequence of the Wigner–Eckart theorem; see Ref. 20, pp. 129–134.

[1] L. M. Falicov, *Group Theory and Its Physical Applications* (The University of Chicago Press, Chicago, IL, 1966).
[2] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method* (Prentice–Hall, Englewood Cliffs, NJ, 1973).
[3] For the reader's convenience, a brief summary of the utilized group theoretic results is given as an Appendix.
[4] G. I. Habetler and M. A. Martino, ''The multigroup diffusion equation of reactor physics,'' Proc. Symp. Appl. Math. **11**, 127 (1961).
[5] P. J. Olver, *Application of Lie Groups to Differential Equations* (Academic, New York, NY, 1982).
[6] L. V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic, New York, NY, 1982).
[7] H. Greenspan C. N. Kolber, and D. Okrent, in *Computing Methods in Reactor Physics* (Gordon and Breach, New York, NY, 1968).
[8] G. Palmiotti *et al.*, VARIANT, Report No. ANL-95/40, Argonne National Laboratory, Argonne, 1995.
[9] M. Makai, ''Group theory applied to boundary value problems,'' Report No. ANL-FRA-1996-5, Argonne National Laboratory, Argonne, IL, 1996.
[10] E. Kreyszig, *Introductory Functional Analysis with Applications* (Wiley, New York, 1978).
[11] B. L. van der Waerden, ''From matrix mechanics and wave mechanics to unified quantum mechanics,'' No. AMS, **44**, 323 (1997); see also Ref. 12.

[12] V. S. Vladimirov, *Equations of Mathematical Physics* (Marcel Dekker, New York, 1971).
[13] Throughout the present work, all terms of group theory are used as defined in GAP; see M. Schönert *et al.*, *GAP—Groups, Algorithms and Programming* (Lehrstuhl für Mathematik, Rheinisch Westfälische Technische Hohschule, Aachen, Germany, 1995).
[14] L. D. Landau and E. M. Lifsitz, *Theoretical Physics* (Pergamon, Oxford, 1960), Vol. 2.
[15] As to reactor physical applications, a recent survey is Ref. 9.
[16] The convergence problem appears as ''matrix rank deficiency'' in Ref. 8. For details, see Chap. V of Ref. 8, and Ref. 17.
[17] C. B. Carrico, E. E. Lewis, and G. Palmiotti, ''Matrix rank in variational nodal approximations,'' Trans. Am. Nucl. Soc. **70**, 162 (1994).
[18] Representations $\{\mathbf{A}_i\}$ and $\{\mathbf{B}_i\}$ are said to be equivalent if there exists a matrix $\mathbf{X}$ such that $\mathbf{B}_i = \mathbf{X}\mathbf{A}_i\mathbf{X}^{-1}$ for all i.
[19] E. P. Wigner, *Group Theory, and Its Application to Quantum Mechanics of Atomic Spectra* (Academic, New York, NY, 1959).
[20] W. Ludwig and C. Falter, *Symmetries in Physics* (Springer-Verlag, Berlin, 1988).

Quasi-Hopf superalgebras and elliptic quantum supergroups

Yao-Zhong Zhang[a)] and Mark D. Gould
Department of Mathematics, University of Queensland, Brisbane, Qld 4072, Australia

(Received 6 October 1998; accepted for publication 11 February 1999)

We introduce the quasi-Hopf superalgebras which are $\mathbf{Z}_2$-graded versions of Drinfeld's quasi-Hopf algebras. We describe the realization of elliptic quantum supergroups as quasi-triangular quasi-Hopf superalgebras obtained from twisting the normal quantum supergroups by twistors which satisfy the graded shifted cocycle condition, thus generalizing the quasi-Hopf twisting procedure to the supersymmetric case. Two types of elliptic quantum supergroups are defined, that is, the face type $\mathcal{B}_{q,\lambda}(\mathcal{G})$ and the vertex type $\mathcal{A}_{q,p}[\mathrm{sl}(\widehat{n|n})]$ (and $\mathcal{A}_{q,p}[\mathrm{gl}(\widehat{n|n})]$), where $\mathcal{G}$ is any Kac–Moody superalgebra with symmetrizable generalized Cartan matrix. It appears that the vertex type twistor can be constructed only for $U_q[\mathrm{sl}(\widehat{n|n})$ in a nonstandard system of simple roots, all of which are fermionic. [S0022-2488(99)00210-8]

I. INTRODUCTION

One of the aims of this paper is to introduce $\mathbf{Z}_2$-graded versions of Drinfeld's quasi-Hopf algebras,[1] which are referred to as quasi-Hopf superalgebras. We then introduce elliptic quantum supergroups, which are defined as quasi-triangular quasi-Hopf superalgebras arising from twisting the normal quantum supergroups by twistors which satisfy the graded shifted cocycle condition, thus generalizing Drinfeld's quasi-Hopf twisting procedure[2–6] to the supersymmetric case. We adopt the approach in Ref. 4 and construct two types of twistors, i.e., the face-type twistor associated to any Kac–Moody superalgebra $\mathcal{G}$ with a symmetrizable generalized Cartan matrix and the vertex-type twistor associated to $\mathrm{sl}(\widehat{n|n})$ in a nonstandard simple root system in which all simple roots are odd (or fermionic). It should be pointed out that the face-type twistors for certain classes of *nonaffine* simple superalgebras were also constructed in Ref. 5.

The elliptic quantum groups[7,8] are believed to provide the underlying algebraic structures for integrable models based on elliptic solutions of the (dynamical) Yang–Baxter equation, such as Baxter's eight-vertex model,[9] the ABF (Andrews—Baxter–Forrester) model,[10] and their group theoretical generalizations.[11,12] The elliptic quantum supergroups described in this paper are expected to play a similar role in supersymmetric integrable models based on elliptic solutions[13,14] of the graded (dynamical) Yang–Baxter equation.

II. QUASI-HOPF SUPERALGEBRAS

Definition 1: A $\mathbf{Z}_2$-graded quasi-bialgebra is a $\mathbf{Z}_2$-graded unital associative algebra A over a field K which is equipped with algebra homomorphisms ϵ: $A\to K$ (counit), Δ: $A\to A\otimes A$ (coproduct), and an invertible homogeneous element $\Phi\in A\otimes A\otimes A$ (coassociator) satisfying

$$(1\otimes\Delta)\Delta(a)=\Phi^{-1}(\Delta\otimes 1)\Delta(a)\Phi,\quad \forall a\in A, \tag{II.1}$$

$$(\Delta\otimes 1\otimes 1)\Phi\cdot(1\otimes 1\otimes\Delta)\Phi=(\Phi\otimes 1)\cdot(1\otimes\Delta\otimes 1)\Phi\cdot(1\otimes\Phi), \tag{II.2}$$

[a)]Electronic mail: yzz@maths.uq.edu.au

$$(\epsilon\otimes 1)\Delta = 1 = (1\otimes\epsilon)\Delta, \tag{II.3}$$

$$(1\otimes\epsilon\otimes 1)\Phi = 1. \tag{II.4}$$

Equations (II.2)–(II.4) imply that Φ also obeys

$$(\epsilon\otimes 1\otimes 1)\Phi = 1 = (1\otimes 1\otimes\epsilon)\Phi. \tag{II.5}$$

The multiplication rule for the tensor products is $\mathbf{Z}_2$ graded and is defined for homogeneous elements $a,b,a',b'\in A$ by

$$(a\otimes b)(a'\otimes b') = (-1)^{[b][a']}(aa'\otimes bb'), \tag{II.6}$$

where $[a]\in\mathbf{Z}_2$ denotes the grading of the element a.

Definition 2: A quasi-Hopf superalgebra is a $\mathbf{Z}_2$*-graded quasi-bialgebra* (A,Δ,ϵ,Φ) *equipped with a* $\mathbf{Z}_2$*-graded algebra anti-homomorphism* $S: A\to A$ *(anti-pode) and canonical elements* α, $\beta\in A$ *such that*

$$m\cdot(1\otimes\alpha)(S\otimes 1)\Delta(a) = \epsilon(a)\alpha, \quad \forall a\in A, \tag{II.7}$$

$$m\cdot(1\otimes\beta)(1\otimes S)\Delta(a) = \epsilon(a)\beta, \quad \forall a\in A, \tag{II.8}$$

$$m\cdot(m\otimes 1)\cdot(1\otimes\beta\otimes\alpha)(1\otimes S\otimes 1)\Phi^{-1} = 1, \tag{II.9}$$

$$m\cdot(m\otimes 1)\cdot(S\otimes 1\otimes 1)(1\otimes\alpha\otimes\beta)(1\otimes 1\otimes S)\Phi = 1. \tag{II.10}$$

Here m denotes the usual product map on A: $m\cdot(a\otimes b) = ab$, $\forall a,b\in A$. Note that since A is associative, we have $m\cdot(m\otimes 1) = m\cdot(1\otimes m)$. For the homogeneous elements $a,b\in A$, the anti-pode satisfies

$$S(ab) = (-1)^{[a][b]}S(b)S(a), \tag{II.11}$$

which extends to inhomogeneous elements through linearity.

Applying ϵ to definitions (II.9) and (II.10) we obtain, in view of (II.4), $\epsilon(\alpha)\epsilon(\beta) = 1$. It follows that the canonical elements α and β are both even. By applying ϵ to (II.7), we have $\epsilon(S(a)) = \epsilon(a)$, $\forall a\in A$.

In the following we show that the category of quasi-Hopf superalgebras is invariant under a kind of gauge transformation. Let (A,Δ,ϵ,Φ) be a qausi-Hopf superalgebra, with α, β, and S satisfying (II.7)–(II.10), and let $F\in A\otimes A$ be an invertible homogeneous element satisfying the counit properties

$$(\epsilon\otimes 1)F = 1 = (1\otimes\epsilon)F. \tag{II.12}$$

It follows that F is even. Throughout we set

$$\Delta_F(a) = F\Delta(a)F^{-1}, \quad \forall a\in A, \tag{II.13}$$

$$\Phi_F = (F\otimes 1)(\Delta\otimes 1)F\cdot\Phi\cdot(1\otimes\Delta)F^{-1}(1\otimes F^{-1}). \tag{II.14}$$

Theorem 1: $(A,\Delta_F,\epsilon,\Phi_F)$, *defined by (II.13) and (II.14), together with* α_F,β_F, and S_F *given by*

$$S_F = S, \quad \alpha_F = m\cdot(1\otimes\alpha)(S\otimes 1)F^{-1}, \quad \beta_F = m\cdot(1\otimes\beta)(1\otimes S)F, \tag{II.15}$$

is also a quasi-Hopf superalgebra. The element F is referred to as a twistor, throughout.

The proof of this theorem is elementary. For demonstration we show in some detail the proof of the antipode properties. Care has to be taken of the gradings in tensor product multiplications and also in extending the antipode to the whole algebra. First of all let us state the following lemma.

Lemma 1: For any elements $\eta\in A\otimes A$ *and* $\xi\in A\otimes A\otimes A$,

$$m\cdot(1\otimes\alpha_F)(S\otimes 1)\eta=m\cdot(1\otimes\alpha)(S\otimes 1)(F^{-1}\eta), \tag{II.16}$$

$$m\cdot(1\otimes\beta_F)(1\otimes S)\eta=m\cdot(1\otimes\beta)(1\otimes S)(\eta F), \tag{II.17}$$

$$\begin{aligned}&m\cdot(m\otimes 1)\cdot(1\otimes\beta_F\otimes\alpha_F)(1\otimes S\otimes 1)\xi\\&\quad=m\cdot(m\otimes 1)\cdot(1\otimes\beta\otimes\alpha)(1\otimes S\otimes 1)[(1\otimes F^{-1})\cdot\xi\cdot(F\otimes 1)],\end{aligned} \tag{II.18}$$

$$\begin{aligned}&m\cdot(m\otimes 1)\cdot(S\otimes 1\otimes 1)(1\otimes\alpha_F\otimes\beta_F)(1\otimes 1\otimes S)\xi\\&\quad=m\cdot(m\otimes 1)\cdot(S\otimes 1\otimes 1)(1\otimes\alpha\otimes\beta)(1\otimes 1\otimes S)\cdot[(F^{-1}\otimes 1)\cdot\xi\cdot(1\otimes F)].\end{aligned} \tag{II.19}$$

Proof: Write $F=f_i\otimes f^i$ and $F^{-1}=\overline{f_i}\otimes\overline{f^i}$. Here and throughout, summation convention on repeated indices is assumed. Then (II.15) can be written as

$$\alpha_F=S(\overline{f_i})\alpha\overline{f^i},\quad \beta_F=f_i\beta S(f^i). \tag{II.20}$$

Further, write $\eta=\eta_k\otimes\eta^k$ and $\xi=\Sigma_i x_i\otimes y_i\otimes z_i$. Then

$$\begin{aligned}\text{lhs of (II.16)}&=m\cdot(1\otimes S(\overline{f_i})\alpha\overline{f^i})(S(\eta_k)\otimes\eta^k)\\&=m\cdot(S(\eta_k)\otimes S(\overline{f_i})\alpha\overline{f^i}\eta^k)\\&=S(\eta_k)S(\overline{f_i})\alpha\overline{f^i}\eta^k\\&=S(\overline{f_i}\eta_k)\alpha\overline{f^i}\eta^k\times(-1)^{[\eta_k][\overline{f_i}]},\end{aligned}$$

$$\begin{aligned}\text{rhs of (II.16)}&=m\cdot(1\otimes\alpha)(S\otimes 1)(\overline{f_i}\eta_k\otimes\overline{f^i}\eta^k)\times(-1)^{[\overline{f^i}][\eta_k]}\\&=S(\overline{f_i}\eta_k)\alpha\overline{f^i}\eta^k\times(-1)^{[\overline{f_i}][\eta_k]},\end{aligned}$$

thus proving (II.16). Equation (II.17) can be proved similarly. As for (II.18) we have

$$\begin{aligned}\text{lhs of (II.18)}&=\sum_i x_i\beta_F S(y_i)\alpha_F z_i\\&=\sum_i x_i f_j\beta S(f^j)S(y_i)S(\overline{f_k})\alpha\overline{f^k}z_i\\&=\sum_i x_i f_j\beta S(\overline{f_k}y_i f^j)\alpha\overline{f^k}z_i\times(-1)^{[y_i][f_j]+[\overline{f_k}])+[\overline{f_k}][f_j]},\end{aligned}$$

$$\begin{aligned}\text{rhs of (II.18)}&=m\cdot(m\otimes 1)\cdot(1\otimes\beta\otimes\alpha)(1\otimes S\otimes 1)\\&\quad\cdot\sum_i[x_i f_j\otimes\overline{f_k}y_i f^j\otimes\overline{f^k}z_i]\times(-1)^{[y_i]([f_j]+[\overline{f_k}])+[\overline{f_k}][f_j]},\end{aligned}$$

$$=\sum_i x_i f_j \beta S(\overline{f_k} y_i f^j)\alpha \overline{f^k} z_i \times(-1)^{[y_i]([f_j]+[\overline{f_k}])+[\overline{f_k}][f_j]},$$

where we have used the fact that the element F is even. Equation (II.19) is proved similarly.

Now let us prove the property (II.7) for α_F and Δ_F. We write, following Sweedler,

$$\Delta(a)=\sum_{(a)} a_{(1)}\otimes a_{(2)}. \tag{II.21}$$

Then, in view of Lemma 1,

$$\begin{aligned}
m\cdot(1\otimes\alpha_F)(S\otimes 1)\Delta_F(a) &= m\cdot(1\otimes\alpha)(S\otimes 1)(F^{-1}\Delta_F(a)\\
&= m\cdot(1\otimes\alpha)(S\otimes 1)(\Delta(a)F^{-1})\\
&= m\cdot(1\otimes\alpha)\sum_{(a)}(S(a_{(1)}\overline{f_i})\otimes a_{(2)}\overline{f^i})\times(-1)^{[\overline{f_i}][a_{(2)}]}\\
&= S(\overline{f_i})\sum_{(a)} S(a_{(1)})\alpha a_{(2)}\overline{f^i}\times(-1)^{[f_i]([a_{(1)}]+[a_{(2)}])}\\
&= S(\overline{f_i})\sum_{(a)} S(a_{(1)})\alpha a_{(2)}\overline{f^i}\times(-1)^{[f_i][a]}\\
&= (-1)^{[\overline{f_i}][a]}S(\overline{f_i})\sum_{(a)} S(a_{(1)})\alpha a_{(2)}\overline{f^i} \overset{(II.17)}{=} S(\overline{f_i})\epsilon(a)\alpha\overline{f^i}\times(-1)^{[\overline{f_i}][a]}\\
&= S(\overline{f_i})\epsilon(a)\alpha\overline{f^i} \overset{(II.20)}{=} \epsilon(a)\alpha_F,
\end{aligned} \tag{II.22}$$

where we have used the fact that

$$\epsilon(a)=0 \quad \text{if } [a]=1. \tag{II.23}$$

The property (II.8) for β_F and Δ_F is proved similarly. We then prove property (II.9), which reads in terms of the twisted objects

$$m\cdot(m\otimes 1)\cdot(1\otimes\beta_F\otimes\alpha_F)(1\otimes S\otimes 1)\Phi_F^{-1}=1. \tag{II.24}$$

Let us write

$$\Phi^{-1}=\sum_\nu \overline{X}_\nu\otimes\overline{Y}_\nu\otimes\overline{Z}_\nu. \tag{II.25}$$

Then, in view of (II.18),

$$\begin{aligned}
&\text{lhs of (II.24)}\\
&= m\cdot(m\otimes 1)\cdot(1\otimes\beta\otimes\alpha)(1\otimes S\otimes 1)[(1\otimes F^{-1})\Phi_F^{-1}(F\otimes 1)]\\
&= m\cdot(m\otimes 1)\cdot(1\otimes\beta\otimes\alpha)(1\otimes S\otimes 1)[(1\otimes\Delta)F\cdot\Phi^{-1}\cdot(\Delta\otimes 1)F^{-1}]\\
&= m\cdot(m\otimes 1)\cdot(1\otimes\beta\otimes\alpha)(1\otimes S\otimes 1)\\
&\quad\cdot\sum_{\nu,(f),(\overline{f})}[f_i\overline{X}_\nu\overline{f}_{j(1)}\otimes f^i_{(1)}\overline{Y}_\nu\overline{f}_{j(2)}\otimes f^i_{(2)}\overline{Z}_\nu\overline{f^j}]\\
&\quad\times(-1)^{([\overline{X}_\nu]+[\overline{f}_{j(1)}])([f^i_{(1)}]+[f^i_{(2)}])+[\overline{Z}_\nu]([\overline{f}_{j(1)}]+[\overline{f}_{j(2)}])+[\overline{Y}_\nu]([\overline{f}_{j(1)}]+[f^i_{(2)}])+[f^i_{(2)}][\overline{f}_{j(2)}]}
\end{aligned}$$

$$= \sum_{\nu} f_i \overline{X}_\nu \sum_{(\overline{f})} f_{j(1)} \beta S(\overline{f}_{j(2)}) S(\overline{Y}_\nu) \sum_{(f)} S(f^i_{(1)}) \alpha f^i_{(2)} \overline{Z}_\nu \overline{f^j}$$

$$\cdot (-1)^{([\overline{X}_\nu]+[\overline{Y}_\nu])([f^i_{(1)}]+[f^i_{(2)}])+([\overline{Y}_\nu]+[\overline{Z}_\nu])([\overline{f}_{j(1)}]+[\overline{f}_{j(2)}])+([f^i_{(1)}]+[f^i_{(2)}])([\overline{f}_{j(1)}]+[\overline{f}_{j(2)}])}$$

$$= \sum_{\nu} f_i \overline{X}_\nu \sum_{(f)} f_{j(1)} \beta S(\overline{f}_j(2)) S(\overline{Y}_\nu) \sum_{(f)} S(f^i_{(1)}) \alpha f^i_{(2)} \overline{Z}_\nu \overline{f^j}$$

$$\cdot (-1)^{([\overline{X}_\nu]+[Y_\nu])[f^i]+([Y_\nu]+[\overline{Z}_\nu])[\overline{f}_j]+[f^i][f_j]}$$

$$= \sum_{\nu} f_i \overline{X}_\nu \cdot (-1)^{([X_\nu]+[Y_\nu])[f^i]+([Y_\nu]+[\overline{Z}_\nu])[f_j]+[f^i][f_j]}$$

$$\cdot \sum_{(f)} f_{j(1)} \beta S(f_{j(2)}) S(\overline{Y}_\nu) \sum_{(f)} S(f^i_{(1)}) \alpha f^i_{(2)} \overline{Z}_\nu \overline{f^j}$$

$$\overset{(II.7),(II.8)}{=} \sum_{\nu} f_i \overline{X}_\nu \epsilon(\overline{f}_j) \beta S(\overline{Y}_\nu) \epsilon(f^i) \alpha Z_\nu f^j$$

$$\cdot (-1)^{([X_\nu]+[Y_\nu])[f^i]+([Y_\nu]+[Z_\nu])[f_j])+[f^i][f_j]}$$

$$\overset{(II.23)}{=} \sum_{\nu} f_i \overline{X}_\nu \epsilon(\overline{f}_j) \beta S(\overline{Y}_\nu) \epsilon(f^i) \alpha \overline{Z}_\nu \overline{f^j} = m \cdot (m \otimes 1) \cdot (1 \otimes \beta \otimes \alpha)(1 \otimes S \otimes 1)$$

$$\cdot [((1 \otimes \epsilon) F \otimes 1) \cdot \Phi^{-1} \cdot ((\epsilon \otimes 1) F^{-1} \otimes 1)]$$

$$\overset{(II.12)}{=} m \cdot (m \otimes 1) \cdot (1 \otimes \beta \otimes \alpha)(1 \otimes S \otimes 1) \Phi^{-1} \overset{(II.9)}{=} 1.$$

The property (II.10) for the twisted objects, which reads

$$m \cdot (m \otimes 1) \cdot (S \otimes 1 \otimes 1)(1 \otimes \alpha_F \otimes \beta_F)(1 \otimes 1 \otimes S) \Phi_F = 1, \tag{II.26}$$

is proved in a similar way.

Definition 3: A quasi-Hopf superalgebra (A,Δ,ϵ,Φ) *is called quasi-triangular if there exists an invertible homogeneous element* $\mathcal{R} \in A \otimes A$ *such that*

$$\Delta^T(a) \mathcal{R} = \mathcal{R} \Delta(a), \quad \forall a \in A, \tag{II.27}$$

$$(\Delta \otimes 1) \mathcal{R} = \Phi^{-1}_{231} \mathcal{R}_{13} \Phi_{132} \mathcal{R}_{23} \Phi^{-1}_{123}, \tag{II.28}$$

$$(1 \otimes \Delta) \mathcal{R} = \Phi_{312} \mathcal{R}_{13} \Phi^{-1}_{213} \mathcal{R}_{12} \Phi_{123}. \tag{II.29}$$

Throughout, $\Delta^T = T \cdot \Delta$ with T being the graded twist map which is defined, for homogeneous elements $a, b \in A$, by

$$T(a \otimes b) = (-1)^{[a][b]} b \otimes a; \tag{II.30}$$

and Φ_{132}, etc. are derived from $\Phi \equiv \Phi_{123}$ with the help of T:

$$\Phi_{132} = (1 \otimes T) \Phi_{123},$$

$$\Phi_{312} = (T \otimes 1) \Phi_{132} = (T \otimes 1)(1 \otimes T) \Phi_{123},$$

$$\Phi^{-1}_{231} = (1 \otimes T) \Phi^{-1}_{213} = (1 \otimes T)(T \otimes 1) \Phi^{-1}_{123},$$

and so on. We remark that our convention differs from the usual one which employs the inverse permutation on the positions (cf. Ref. 4).

It is easily shown that the properties (II.27)–(II.29) imply the graded Yang–Baxter-type equation,

$$\mathcal{R}_{12}\Phi_{231}^{-1}\mathcal{R}_{13}\Phi_{132}\mathcal{R}_{23}\Phi_{123}^{-1}=\Phi_{321}^{-1}\mathcal{R}_{23}\Phi_{312}\mathcal{R}_{13}\Phi_{213}^{-1}\mathcal{R}_{12}, \tag{II.31}$$

which is referred to as the graded quasi-Yang–Baxter equation, and the counit properties of $\mathcal{R}$:

$$(\epsilon\otimes 1)\mathcal{R}=1=(1\otimes\epsilon)\mathcal{R}. \tag{II.32}$$

Theorem 2: *Denoting by the set* $(A,\Delta,\epsilon,\Phi,\mathcal{R})$ *a quasi-triangular quasi-Hopf superalgebra, then* $(A,\Delta_F,\epsilon,\Phi_F,\mathcal{R}_F)$ *is also a quasi-triangular quasi-Hopf superalgebra, with the choice of* $\mathcal{R}_F$ *given by*

$$\mathcal{R}_F=F^T\mathcal{R}F^{-1}, \tag{II.33}$$

where $F^T=T\cdot F\equiv F_{21}$. *Here* Δ_F *and* Φ_F *are given by (II.13) and (II.14), respectively.*

The proof of this theorem is elementary computation. As an example, let us illustrate the proof of the property (II.28) for Δ_F, $\mathcal{R}_F$, and Φ_F. Applying the homomorphism $T\otimes 1$ to $(\Phi_F^{-1})_{123}$, one obtains

$$\begin{aligned}(\Phi_F^{-1})_{213}&=F_{13}(T\otimes 1)(1\otimes\Delta)F\cdot\Phi_{213}^{-1}\cdot(\Delta^T\otimes 1)F^{-1}\cdot(F^T)_{12}^{-1}\\&=F_{13}\sum_{(f)}(-1)^{[f^i_{(1)}][f_i]}(f^i_{(1)}\otimes f_i\otimes f^i_{(2)})\Phi_{213}^{-1}(\Delta^T\otimes 1)F^{-1}\cdot(F^T)_{12}^{-1},\end{aligned} \tag{II.34}$$

which gives rise to, by applying the homomorphism $1\otimes T$ to both sides,

$$\begin{aligned}(\Phi_F^{-1})_{231}&=F_{12}\sum_{(f)}(-1)^{([f^i_{(1)}]+[f^i_{(2)}])[f_i]}(f^i_{(1)}\otimes f^i_{(2)}\otimes f_i)\Phi_{231}^{-1}(1\otimes T)(\Delta^T\otimes 1)F^{-1}\cdot(F^T)_{13}^{-1}\\&=F_{12}(\Delta\otimes 1)F^T\cdot\Phi_{231}^{-1}(1\otimes T)(\Delta^T\otimes 1)F^{-1}\cdot(F^T)_{13}^{-1}.\end{aligned} \tag{II.35}$$

Then,

$$\begin{aligned}(\Delta_F\otimes 1)\mathcal{R}_F&=(F\otimes 1)(\Delta\otimes 1)\mathcal{R}_F\cdot(F^{-1}\otimes 1)\\&=F_{12}(\Delta\otimes 1)(F^T\mathcal{R}F^{-1})\cdot F_{12}^{-1}\\&=F_{12}(\Delta\otimes 1)F^T(\Delta\otimes 1)\mathcal{R}(\Delta\otimes 1)F^{-1}\cdot F_{12}^{-1}\\&\overset{\text{(II.28)}}{=}F_{12}(\Delta\otimes 1)F^T\cdot\Phi_{231}^{-1}\mathcal{R}_{13}\Phi_{132}\mathcal{R}_{23}\Phi_{123}^{-1}(\Delta\otimes 1)F^{-1}\cdot F_{12}^{-1}\\&\overset{\text{(II.35)}}{=}(\Phi_F^{-1})_{231}(F^T)_{13}(1\otimes T)(\Delta^T\otimes 1)F\cdot\mathcal{R}_{13}\Phi_{132}\mathcal{R}_{23}\Phi_{123}^{-1}(\Delta\otimes 1)F^{-1}\cdot F_{12}^{-1}\\&\overset{\text{(II.14)}}{=}(\Phi_F^{-1})_{231}(F^T)_{13}(1\otimes T)(\Delta^T\otimes 1)F\cdot\mathcal{R}_{13}\Phi_{132}\mathcal{R}_{23}(1\otimes\Delta)F^{-1}\cdot F_{23}^{-1}(\Phi_F^{-1})_{123}\end{aligned}$$

$$=(\Phi_F^{-1})_{231}(F^T)_{13}(1\otimes T)[(\Delta^T\otimes 1)F\cdot\mathcal{R}_{12}]\cdot\Phi_{132}\mathcal{R}_{23}(1\otimes\Delta)F^{-1}\cdot F_{23}^{-1}(\Phi_F^{-1})_{123}$$

$$\overset{\text{(II.27)}}{=}(\Phi_F^{-1})_{231}(F^T)_{13}(1\otimes T)[\mathcal{R}_{12}(\Delta\otimes 1)F]\cdot\Phi_{132}(1\otimes\Delta^T)F^{-1}\cdot\mathcal{R}_{23}F_{23}^{-1}(\Phi_F^{-1})_{123}$$

$$=(\Phi_F^{-1})_{231}(F^T)_{13}\mathcal{R}_{13}(1\otimes T)[(\Delta\otimes 1)F]\cdot\Phi_{132}(1\otimes\Delta^T)F^{-1}\cdot\mathcal{R}_{23}F_{23}^{-1}(\Phi_F^{-1})_{123}$$

$$\overset{\text{(II.33)}}{=}(\Phi_F^{-1})_{231}(\mathcal{R}_F)_{13}F_{13}^{-1}(1\otimes T)[(\Delta\otimes 1)F]\cdot\Phi_{132}(1\otimes\Delta^T)F^{-1}(F^T)_{23}^{-1}(\mathcal{R}_F)_{23}(\Phi_F^{-1})_{123}$$

$$=(\Phi_F^{-1})_{231}(\mathcal{R}_F)_{13}(1\otimes T)[F_{12}^{-1}(\Delta\otimes 1)F\Phi_{123}(1\otimes\Delta)F^{-1}\cdot F_{23}^{-1}]\cdot(\mathcal{R}_F)_{23}(\Phi_F^{-1})_{123}$$

$$\overset{\text{(II.14)}}{=}(\Phi_F^{-1})_{231}(\mathcal{R}_F)_{13}(1\otimes T)(\Phi_F)_{123}\cdot(\mathcal{R}_F)_{23}(\Phi_F^{-1})_{123}$$

$$=(\Phi_F^{-1})_{231}(\mathcal{R}_F)_{13}(\Phi_F)_{132}(\mathcal{R}_F)_{23}(\Phi_F^{-1})_{123}. \tag{II.36}$$

Let us now consider the special case that A arises from a normal quasi-triangular Hopf superalgebra via twisting with F. A quasi-triangular Hopf superalgebra is a quasi-triangular quasi-Hopf superalgebra with $\alpha=\beta=1$ and $\Phi=1\otimes 1\otimes 1$. Hence A has the following $\mathbf{Z}_2$ graded quasi-Hopf algebra structure:

$$\Delta_F(a)=F\Delta(a)F^{-1},\quad \forall a\in A,$$
$$\Phi_F=F_{12}\cdot(\Delta\otimes 1)F\cdot(1\otimes\Delta)F^{-1}\cdot F_{23}^{-1},$$
$$\alpha_F=m\cdot(S\otimes 1)F^{-1},\quad \beta_F=m\cdot(1\otimes S)F, \tag{II.37}$$
$$\mathcal{R}_F=F^T\mathcal{R}F^{-1}.$$

The twisting procedure is particularly interesting when the twistor $F\in A\otimes A$ depends on an element $\lambda\in A$, i.e., $F=F(\lambda)$, and is a shifted cocycle in the following sense. Here λ is assumed to depend on one (or possible several) parameters.

Definition 4: A twistor $F(\lambda)$ depending on $\lambda\in A$ is a shifted cocycle if it satisfies the graded shifted cocycle condition:

$$F_{12}(\lambda)\cdot(\Delta\otimes 1)F(\lambda)=F_{23}(\lambda+h^{(1)})\cdot(1\otimes\Delta)F(\lambda), \tag{II.38}$$

where $h^{(1)}=h\otimes 1\otimes 1$ and $h\in A$ is fixed.

Let $(A,\Delta_\lambda,\epsilon,\Phi(\lambda),\mathcal{R}(\lambda))$ be the quasi-triangular quasi-Hopf superalgebra obtained from twisting the quasi-triangular Hopf superalgebra by the twistor $F(\lambda)$. Then we have the following.

Proposition 1: We have

$$\Phi(\lambda)\equiv\Phi_F=F_{23}(\lambda+h^{(1)})F_{23}(\lambda)^{-1}, \tag{II.39}$$

$$\Delta_\lambda(a)^T\mathcal{R}(\lambda)=\mathcal{R}(\lambda)\Delta_\lambda(a),\quad \forall a\in A, \tag{II.40}$$

$$(\Delta_\lambda\otimes 1)\mathcal{R}(\lambda)=\Phi_{231}(\lambda)^{-1}\mathcal{R}_{13}(\lambda)\mathcal{R}_{23}(\lambda+h^{(1)}), \tag{II.41}$$

$$(1\otimes\Delta_\lambda)\mathcal{R}(\lambda)=\mathcal{R}_{13}(\lambda+h^{(2)})\mathcal{R}_{12}(\lambda)\Phi_{123}(\lambda). \tag{II.42}$$

As a corollary, $\mathcal{R}(\lambda)$ *satisfies the graded dynamical Yang–Baxter equation*

$$\mathcal{R}_{12}(\lambda+h^{(3)})\mathcal{R}_{13}(\lambda)\mathcal{R}_{23}(\lambda+h^{(1)})=\mathcal{R}_{23}(\lambda)\mathcal{R}_{13}(\lambda+h^{(2)})\mathcal{R}_{12}(\lambda). \tag{II.43}$$

III. QUANTUM SUPERGROUPS

Let $\mathcal{G}$ be a Kac–Moody superalgebra[15,16] with a symmetrizable generalized Cartan matrix $A=(a_{ij})_{i,j,\in I}$. As is well known, a given Kac–Moody superalgebra allows many inequivalent systems of simple roots. A system of simple roots is called distinguished if it has minimal odd roots. Let $\{\alpha_i, i\in I\}$ denote a chosen set of simple roots. Let (,) be a fixed invariant bilinear form on the root space of $\mathcal{G}$. Let $\mathcal{H}$ be the Cartan subalgebra and throughout we identify the dual $\mathcal{H}^*$ with $\mathcal{H}$ via (,). The generalized Cartan matrix $A=(a_{ij})_{i,j\in I}$ is defined from the simple roots by

$$a_{ij}=\begin{cases}\dfrac{2(\alpha_i,\alpha_j)}{(\alpha_i,\alpha_i)}, & \text{if } (\alpha_i,\alpha_i)\neq 0,\\ (\alpha_i,\alpha_j), & \text{if } (\alpha_i,\alpha_i)=0.\end{cases} \tag{III.1}$$

As we mentioned in the previous section, quantum Kac–Moody superalgebras are quasi-triangular quasi-Hopf superalgebras with $\alpha=\beta=1$ and $\Phi=1\otimes1\otimes1$. We shall not give the standard relations obeyed by the simple generators (or Chevalley generators) $\{h_i, e_i, f_i, i\in I\}$ of $U_q(\mathcal{G})$, but mention that for certain types of Dynkin diagrams extra q-Serre relations are needed in the defining relations. We adopt the following graded Hopf algebra structure,

$$\begin{gathered}\Delta(h)=h\otimes 1+1\otimes h,\\ \Delta(e_i)=e_i\otimes 1+t_i\otimes e_i,\quad \Delta(f_i)=f_i\otimes t_i^{-1}+1\otimes f_i,\\ \epsilon(e_i)=\epsilon(f_i)=\epsilon(h)=0,\\ S(e_i)=-t_i^{-1}e_i,\quad S(f_i)=-f_i t_i,\quad S(h)=-h,\end{gathered} \tag{III.2}$$

where $i\in I$, $t_i=q^{h_i}$ and $h\in\mathcal{H}$.

The canonical element $\mathcal{R}$ is called the universal R-matrix of $U_q(\mathcal{G})$, which satisfies the basic properties [e.g., (II.27)–(II.29) with $\Phi=1\otimes1\otimes1$ and (II.32)]

$$\begin{aligned}&\Delta^T(a)\mathcal{R}=\mathcal{R}\Delta(a),\quad \forall a\in U_q(\mathcal{G}),\\ &(\Delta\otimes 1)\mathcal{R}=\mathcal{R}_{13}\mathcal{R}_{23},\\ &(1\otimes\Delta)\mathcal{R}=\mathcal{R}_{13}\mathcal{R}_{12},\\ &(\epsilon\otimes 1)\mathcal{R}=(1\otimes\epsilon)\mathcal{R}=1,\end{aligned} \tag{III.3}$$

and the graded Yang–Baxter equation [cf. (II.31) with $\Phi=1\otimes1\otimes1$]

$$\mathcal{R}_{12}\mathcal{R}_{13}\mathcal{R}_{23}=\mathcal{R}_{23}\mathcal{R}_{13}\mathcal{R}_{12}. \tag{III.4}$$

The Hopf superalgebra $U_q(\mathcal{G})$ contains two important Hopf subalgebras U_q^+ and U_q^- which are generated by e_i and f_i, respectively. By Drinfeld's quantum double construction, the universal R-matrix $\mathcal{R}$ can be written in the form

$$\mathcal{R}=\left(1\otimes 1+\sum_t a^t\otimes a_t\right)\cdot q^{-\mathcal{T}}, \tag{III.5}$$

where $\{a^t\}\in U_q^+$ and $\{a_t\}\in U_q^-$. The element $\mathcal{T}$ is defined as follows. If the symmetrical Cartan matrix is nondegenerate, then $\mathcal{T}$ is the usual canonical element of $\mathcal{H}\otimes\mathcal{H}$. Let $\{h_l\}$ be a basis of $\mathcal{H}$ and $\{h^l\}$ be its dual basis. Then $\mathcal{T}$ can be written as

$$\mathcal{T}=\sum_l h_l\otimes h^l. \tag{III.6}$$

In the case of a degenerate symmetrical Cartan matrix, we extend the Cartan subalgebra $\mathcal{H}$ by adding some elements to it in such a way that the extended symmetrical Cartan matrix is nondegenerate.[17] Then $\mathcal{T}$ stands for the canonical element of the extended Cartan subalgebra. It still takes the form (III.6), but now $\{h_l\}$ ($\{h^l\}$) is understood to be the (dual) basis of the extended Cartan subalgebra. After such enlargement, one has $h=\Sigma_l(h^l,h)h_l=\Sigma_l(h_l,h)h^l$ for any given h in the enlarged Cartan subalgebra.

For later use, we work out the explicit form of the universal R-matrix for the simplest quantum affine superalgebra $U_q[\widehat{sl(1|1)}]$. This algebra is generated by Chevalley generators $\{e_i,f_i,h_i,d,i=0,1\}$ with e_i, f_i odd, and h_i, d even. Here and throughout d stands for the derivation operator. Let us write $h_i=\alpha_i$. Then we have $h_0=\delta-\varepsilon_1+\delta_1$ and $h_1=\varepsilon_1-\delta_1$, where $\{\varepsilon_1,\delta_1,\delta\}$ satisfy $(\varepsilon_1,\varepsilon_1)=1=-(\delta_1,\delta_1)$, $(\varepsilon_1,\delta_1)=(\delta,\delta)=(\delta,\varepsilon_1)=(\delta,\delta_1)=0$. We extend the Cartan subalgebra by adding to it the element $h_{\rm ex}=\varepsilon_1+\delta_1$. A basis for the enlarged Cartan subalgebra is thus $\{h_{\rm ex},h_0,h_1,d\}$. It is easily shown that the dual basis is $\{h^{\rm ex},h^0,h^1,c\}$, where $h^{\rm ex}=\frac{1}{2}(\varepsilon_1-\delta_1)=\frac{1}{2}h_1$, $h^0=d$, and $h^1=\varepsilon_1+d-\frac{1}{2}(\varepsilon_1-\delta_1)=d+\frac{1}{2}h_{\rm ex}$. As is well known, $U_q[\widehat{sl(1|1)}]$ can also be realized in terms of the Drinfeld generators[18] $\{X_n^\pm,H_n,H_n^{\rm ex},n\in\mathbf{Z},c,d\}$, where $X_n^\pm$ are odd and all other generators are even. The relations satisfied by the Drinfeld generators read[19]

$$\begin{gathered}
[c,a]=[H_0,a]=[d,d]=[H_n,H_m]=[H_n^{\rm ex},H_m^{\rm ex}]=0,\quad \forall a\in U_q[\widehat{sl(1|1)}],\\
q^{H_0^{\rm ex}}X_n^\pm q^{-H_0^{\rm ex}}=q^{\pm 2}X_n^\pm,\\
[d,X_n^\pm]=nX_n^\pm,\quad [d,H_n]=nH_n,\quad [d,H_n^{\rm ex}]=nH_n^{\rm ex},\\
[H_n,H_m^{\rm ex}]=\delta_{n+m,0}\frac{[2n]_q[nc]_q}{n},\\
[H_n^{\rm ex},X_m^\pm]=\pm\frac{[2n]_q}{n}X_{n+m}^\pm q^{\mp|n|c/2},\\
[H_n,X_m^\pm]=0=[X_n^\pm,X_m^\pm],\\
[X_n^+,X_m^-]=\frac{1}{q-q^{-1}}\left(q^{(c/2)(n-m)}\psi_{n+m}^+-q^{-(c/2)(n-m)}\psi_{n+m}^-\right),
\end{gathered} \tag{III.7}$$

where $[x]_q=(q^x-q^{-x})/(q-q^{-1})$, $[a,b]\equiv ab-(-1)^{[a][b]}ba$ denotes the supercommutator, and $\psi_{\pm n}^\pm$ are related to $H_{\pm n}$ by relations

$$\sum_{n\geqslant 0}\psi^{\pm}_{\pm n}z^{\mp n}=q^{\pm H_0}\exp\left(\pm(q-q^{-1})\sum_{n>0}H_{\pm n}z^{\mp n}\right). \tag{III.8}$$

The relationship between the Drinfeld generators and the Chevalley generators is

$$\begin{aligned}&e_1=X_0^+,\quad f_1=X_0^-,\quad h_1=H_0,\quad h_{\rm ex}=H_0^{\rm ex},\\&e_0=X_1^-q^{-H_0},\quad f_0=-q^{H_0}X_{-1}^+,\quad h_0=c-H_0.\end{aligned} \tag{III.9}$$

With the help of the Drinfeld generators, we find the following universal R-matrix,

$$\mathcal{R}=\mathcal{R}'\cdot q^{-\mathcal{T}}, \tag{III.10}$$

where

$$\begin{aligned}\mathcal{T}&=h_{\rm ex}\otimes h^{\rm ex}+h_0\otimes h^0+h_1\otimes h^1+d\otimes c\\&=\tfrac{1}{2}(H_0\otimes h_0^{\rm ex}+H_0^{\rm ex}\otimes H_0)+c\otimes d+d\otimes c,\end{aligned}$$

$$\mathcal{R}'=\mathcal{R}^{<}\mathcal{R}^0\mathcal{R}^{>},$$

$$\mathcal{R}^{<}=\prod_{n\geqslant 0}^{\rightarrow}\exp[(q-q^{-1})(q^{-nc/2}X_n^+\otimes q^{nc/2}X_{-n}^-)], \tag{III.11}$$

$$\mathcal{R}^0=\exp\left[-(q-q^{-1})\sum_{n=1}^{\infty}\frac{n}{[2n]_q}(H_n\otimes H_{-n}^{\rm ex}+H_n^{\rm ex}\otimes H_{-n})\right],$$

$$\mathcal{R}^{>}=\prod_{n\geqslant 0}^{\leftarrow}\exp[-(q-q^{-1})(X_{n+1}^-q^{nc/2-H_0}\otimes q^{-nc/2+H_0}X_{-n-1}^+)].$$

Here and throughout,

$$\prod_{k\geqslant 0}^{\rightarrow}A_k=A_0A_1A_2\cdots,\quad \prod_{k\geqslant 0}^{\leftarrow}A_k=\cdots A_2A_1A_0. \tag{III.12}$$

It seems to us that even for this simplest quantum affine superalgebra $U_q[\widehat{sl(1|1)}]$ the universal R-matrix has not been written down in its explicit form before.

Let us compute the image of $\mathcal{R}$ in the two-dimensional evaluation representation (π,V) of $U_q[\widehat{sl(1|1)}]$, where $V=\mathbf{C}^{1|1}=\mathbf{C}v_1\otimes\mathbf{C}v_2$ with v_1 even and v_2 odd. Let e_{ij} be the 2×2 matrix whose (i,j)-element is unity and zero otherwise. In the homogeneous gradation, the simple generators are represented by

$$\begin{aligned}&e_1=\sqrt{[\theta]_q}e_{12},\quad f_1=\sqrt{[\theta]_q}e_{21},\quad h_1=\theta(e_{11}+e_{22}),\quad h_{\rm ex}=2e_{11}+c_0(e_{11}+e_{22}),\\&e_0=z\sqrt{[\theta]_q}e_{21},\quad f_0=-z^{-1}\sqrt{[\theta]_q}e_{12},\quad h_0=-\theta(e_{11}+e_{22}),\end{aligned} \tag{III.13}$$

where θ and c_0 are arbitrary constants. Then it can be shown that the Drinfeld generators are represented by

$$\begin{aligned}&H_n=z^n\frac{[n\theta]_q}{n}(e_{11}+e_{22}),\quad H_n^{\rm ex}=z^n\frac{[2n]_q}{n}q^{n\theta}e_{11}+z^nc_n(e_{11}+e_{22}),\\&X_n^+=z^nq^{n\theta}\sqrt{[\theta]_q}e_{12},\quad X_n^-=z^nq^{n\theta}\sqrt{[\theta]_q}e_{21},\end{aligned} \tag{III.14}$$

where again c_n are arbitrary constants. In the following we set c_n to be zero. Then the image $R_{VV}(z;\theta,\theta')=(\pi_\theta\otimes\pi_{\theta'})\mathcal{R}$ depends on two extra nonadditive parameters θ, θ', and is given by

$$R_{VV}(z;\theta,\theta')=\frac{q^{-\theta-\theta'}-z}{1-zq^{-\theta-\theta'}}e_{11}\otimes e_{11}+e_{22}\otimes e_{22}+\frac{q^{-\theta'}-zq^{-\theta}}{1-zq^{-\theta-\theta'}}e_{11}\otimes e_{22}+\frac{q^{-\theta}-zq^{-\theta'}}{1-zq^{-\theta-\theta'}}e_{22}\otimes e_{11}$$
$$+\sqrt{[\theta]_q[\theta']_q}q^{-\theta}\frac{q-q^{-1}}{1-zq^{-\theta-\theta'}}e_{12}\otimes e_{21}-\sqrt{[\theta]_q[\theta']_q}q^{-\theta'}\frac{z(q-q^{-1})}{1-zq^{-\theta-\theta'}}e_{21}\otimes e_{12}. \tag{III.15}$$

Equation (III.15) is nothing but the R-matrix obtained in Ref. 20 by solving the Jimbo equation.

IV. ELLIPTIC QUANTUM SUPERGROUPS

Following Jimbo *et al.*,[4] we define elliptic quantum supergroups to be quasi-triangular quasi-Hopf superalgebras obtained from twisting the normal quantum supergroups (which are quasi-triangular quasi-Hopf superalgebras with $\alpha=\beta=1$ and $\Phi=1\otimes1\otimes1$) by twistors which satisfy the graded shifted cocycle condition.

A. Elliptic quantum supergroups of face type

Let ρ be an element in the (extended) Cartan subalgebra such that $(\rho,\alpha_i)=(\alpha_i,\alpha_i)/2$ for all $i\in I$, and

$$\phi=\mathrm{Ad}\,(q^{(1/2)\Sigma_l h_l h^l-\rho}), \tag{IV.1}$$

be an automorphism of $U_q(\mathcal{G})$. Here $\{h_l\}$ and $\{h^l\}$ are as in (III.6) and are the dual basis of the (extended) Cartan subalgebra. Namely,

$$\phi(e_i)=e_it_i,\quad \phi(f_i)=t_i^{-1}f_i,\quad \phi(q^h)=q^h. \tag{IV.2}$$

In the following we consider the special case in which the element λ introduced before belongs to the (extended) Cartan subalgebra. Let

$$\phi_\lambda=\phi^2\cdot\mathrm{Ad}\,(q^{2\lambda})=\mathrm{Ad}\,(q^{\Sigma_l h_l h^l-2\rho+2\lambda}) \tag{IV.3}$$

be an automorphism depending on the element λ and $\mathcal{R}$ be the universal R-matrix of $U_q(\mathcal{G})$. Following Jimbo *et al.*,[4] we define a twistor $F(\lambda)$ by the infinite product

$$F(\lambda)=\prod_{k\geq1}^{\leftarrow}(\phi_\lambda^k\otimes1)(q^{\mathcal{T}}\mathcal{R})^{-1}. \tag{IV.4}$$

It is easily seen that $F(\lambda)$ is a formal series in parameter(s) in λ with leading term 1. Therefore the infinite product makes sense. The twistor $F(\lambda)$ is referred to as a face-type twistor. It can be shown that $F(\lambda)$ satisfies the graded shifted cocycle condition

$$F_{12}(\lambda)(\Delta\otimes1)F(\lambda)=F_{23}(\lambda+h^{(1)})(1\otimes\Delta)F(\lambda), \tag{IV.5}$$

where, if $\lambda=\Sigma_l\lambda_l h^l$, then $\lambda+h^{(1)}=\Sigma_l(\lambda_l+h_l^{(1)})h^l$. The proof of (IV.5) is identical to the non-super case given by Jimbo *et al.*,[4] apart from the use of the graded tensor products. Moreover, it is easily seen that $F(\lambda)$ obeys the counit property

$$(\epsilon\otimes1)F(\lambda)=(1\otimes\epsilon)F(\lambda)=1. \tag{IV.6}$$

We have the following definition.

Definition 5 (Face-type elliptic quantum supergroup): We define elliptic quantum supergroup $\mathcal{B}_{q,\lambda}(\mathcal{G})$ of face type to be the quasi-triangular quasi-Hopf superalgebra $(U_q(\mathcal{G}),\Delta_\lambda,\epsilon,\Phi(\lambda),\mathcal{R}(\lambda))$ together with the graded algebra anti-homomorphism S defined by (III.2) and $\alpha_\lambda = m\cdot(S\otimes 1)F(\lambda)^{-1}$, $\beta_\lambda = m\cdot(1\otimes S)F(\lambda)$. Here ϵ is defined by (III.2), and

$$\begin{aligned}
&\Delta_\lambda(a)=F(\lambda)\Delta(a)F(\lambda)^{-1}, \quad \forall a\in U_q(\mathcal{G}),\\
&\mathcal{R}(\lambda)=F(\lambda)^T\mathcal{R}F(\lambda)^{-1},\\
&\Phi(\lambda)=F_{23}(\lambda+h^{(1)})F_{23}(\lambda)^{-1}.
\end{aligned} \tag{IV.7}$$

We now consider the particularly interesting case where $\mathcal{G}$ is of affine type. Then ρ contains two parts,

$$\rho=\bar{\rho}+gd, \tag{IV.8}$$

where $g=(\psi,\psi+2\bar{\rho})/2$, $\bar{\rho}$ is the graded half-sum of positive roots of the nonaffine part $\bar{\mathcal{G}}$, and ψ is highest root of $\mathcal{G}$; d is the derivation operator which gives the homogeneous gradation

$$[d,e_i]=\delta_{i0}e_i, \quad [d,f_i]=-\delta_{i0}f_i, \quad i\in I. \tag{IV.9}$$

We also set

$$\lambda=\bar{\lambda}+(r+g)d+s'c, \quad r,s'\in\mathbf{C}, \tag{IV.10}$$

where $\bar{\lambda}$ stands for the projection of λ onto the (extended) Cartan subalgebra of $\bar{\mathcal{G}}$. Denoting by $\{\bar{h}_j\}$ and $\{\bar{h}^j\}$ the dual basis of the (extended) Cartan subalgebra of $\bar{\mathcal{G}}$ and setting $p=q^{2r}$, we can decompose ϕ_λ into two parts,

$$\phi_\lambda=\mathrm{Ad}\,(p^dq^{2cd})\cdot\bar{\phi}_\lambda, \quad \bar{\phi}_\lambda=\mathrm{Ad}\,(q^{\Sigma_j\bar{h}_j\bar{h}^j+2(\bar{\lambda}-\bar{\rho})}). \tag{IV.11}$$

Introduce a formal parameter z (which will be identified with spectral parameter) into $\mathcal{R}$ and $F(\lambda)$ by setting

$$\begin{aligned}
&\mathcal{R}(z)=\mathrm{Ad}\,(z^d\otimes 1)\mathcal{R},\\
&F(z,\lambda)=\mathrm{Ad}\,(z^d\otimes 1)F(\lambda),\\
&\mathcal{R}(z,\lambda)=\mathrm{Ad}\,(z^d\otimes 1)\mathcal{R}(\lambda)=F(z^{-1},\lambda)^T\mathcal{R}(z)F(z,\lambda)^{-1}.
\end{aligned} \tag{IV.12}$$

Then it can be shown from the definition of $F(\lambda)$ that $F(z,\lambda)$ satisfies the difference equation

$$\begin{aligned}
&F(pq^{2c^{(1)}}z,\lambda)=(\bar{\phi}_\lambda\otimes 1)^{-1}(F(z,\lambda))\cdot q^{\mathcal{T}}\mathcal{R}(pq^{2c^{(1)}}z),\\
&F(0,\lambda)=F_{\bar{\mathcal{G}}}(\bar{\lambda}).
\end{aligned} \tag{IV.13}$$

The initial condition follows from the fact that $\mathcal{R}(z)q^{d\otimes c+c\otimes d}|_{z=0}$ reduces to the universal R-matrix of $U_q(\bar{\mathcal{G}})$.

Let us give some examples.

1. The case $\mathcal{B}q,\lambda[\mathrm{sl}(1|1)]$

In this case the universal R-matrix is given simply by

$$\begin{aligned}
&\mathcal{R}=\exp[(q-q^{-1})e\otimes f]q^{-\mathcal{T}}=[1+(q-q^{-1})e\otimes f]q^{-\mathcal{T}},\\
&\mathcal{T}=\tfrac{1}{2}(h\otimes h_{\mathrm{ex}}+h_{\mathrm{ex}}\otimes h).
\end{aligned} \tag{IV.14}$$

Let us write

$$\lambda=(s'+1)\tfrac{1}{2}h+s\tfrac{1}{2}h_{\rm ex},\quad s',s\in\mathbf{C}. \tag{IV.15}$$

Since h commutes with everything, ϕ_λ is independent of s'. Setting $w=q^{2(s+h)}$, we have

$$\phi_\lambda=\mathrm{Ad}(w^{1/2h_{\rm ex}}). \tag{IV.16}$$

The formula for the twistor becomes

$$\begin{aligned}F(w)&=\prod_{k\geqslant 1}\big(1-(q-q^{-1})w^kq^{-h}e\otimes fq^h\big)\\&=1-(q-q^{-1})\sum_{k=1}^{\infty}w^kq^{-h}e\otimes fq^h\\&=1-(q-q^{-1})\frac{w}{1-w}q^{-h}e\otimes fq^h.\end{aligned} \tag{IV.17}$$

2. The case $\mathcal{B}_{q,\lambda}[\mathrm{sl}(\hat{1}|1)$

Taking a basis $\{c,d,h,h_{\rm ex}\}$ of the enlarged Cartan subalgebra of $\mathrm{sl}(\widehat{1}|1)$, we write

$$\lambda=rd+s'c+(s''+1)\tfrac{1}{2}h+s\tfrac{1}{2}h_{\rm ex},\quad r,s',s'',s\in\mathbf{C}. \tag{IV.18}$$

Then ϕ_λ is independent of s' and s''. Set

$$p=q^{2r},\quad w=q^{2(s+h)}. \tag{IV.19}$$

Set $F(z;p,w)\equiv F(z,\lambda)$. Then (IV.13) take the form

$$F(pq^{2c^{(1)}}z;p,w)=(\overline{\phi_w}^{-1}\otimes 1)\big(F(z;p,w)\big)\cdot q^{\mathcal{T}}\mathcal{R}(pq^{2c^{(1)}}z), \tag{IV.20}$$

$$F(0;p,w)=F_{\mathrm{sl}(1|1)}(w), \tag{IV.21}$$

where $\overline{\phi_w}=\mathrm{Ad}(w^{(1/2h_{\rm ex}})$.

The image of (IV.20) in the two-dimensional representation (π,V) given by (III.13) (by setting $\theta=1$) yields a difference equation for $F_{VV}(z;p,w)=(\pi\otimes\pi)F(z;p,w)$. Noting that $\pi\cdot\overline{\phi_w}=\mathrm{Ad}\,(D_w^{-1})\cdot\pi$, where $D_w=e_{11}+we_{22}$, we find

$$F_{VV}(pz;p,w)=\mathrm{Ad}\,(D_w\otimes 1)\big(F_{VV}(z;p,w)\big)\cdot KR_{VV}(pz), \tag{IV.22}$$

where $K=(\pi\otimes\pi)q^{\mathcal{T}}=q^2e_{11}\otimes e_{11}+qe_{11}\otimes e_{22}+qe_{22}\otimes e_{11}+e_{22}\otimes e_{22}$ and $R_{VV}(pz)$ is given by (III.15) (with $\theta=\theta'=1$). Equation (IV.22) is a system of difference equations of q-KZ (Kaizhnik–Zamolodchikov) equation type,[21] and can be solved with the help of the q-hypergeometric series. The solution with the initial condition (IV.21) is given by

$$\begin{aligned}F_{VV}(z;p,w)=&{}_1\phi_0(z;p,w)e_{11}\otimes e_{11}+e_{22}\otimes e_{22}+f_{11}(z;p,w)e_{11}\otimes e_{22}+f_{22}(z;p,w)e_{22}\otimes e_{11}\\&+f_{12}(z;p,w)e_{12}\otimes e_{21}+f_{21}(z;p,w)e_{21}\otimes e_{12},\end{aligned} \tag{IV.23}$$

where

$${}_1\phi_0(z;p,w)=\frac{(pq^{-2}z;p)_\infty}{(pq^2z;p)_\infty},$$

$$f_{11}(z;p,w)={}_2\phi_1\left(\begin{matrix} wq^{-2} & q^{-2} \\ & w \end{matrix};p,pq^2z\right),$$

$$f_{12}(z;p,w)=-\frac{w(q-q^{-1})}{1-w}\,{}_2\phi_1\left(\begin{matrix} wq^{-2} & pq^{-2} \\ & pw \end{matrix};p,pq^2z\right), \tag{IV.24}$$

$$f_{21}(z;p,w)=\frac{zpw^{-1}(q-q^{-1})}{1-pw^{-1}}\,{}_2\phi_1\left(\begin{matrix} pw^{-1}q^{-2} & pq^{-2} \\ & p^2w^{-1} \end{matrix};p,pq^2z\right),$$

$$f_{22}(z;p,w)={}_2\phi_1\left(\begin{matrix} pw^{-1}q^{-2} & q^{-2} \\ & pw^{-1} \end{matrix};p,pq^2z\right).$$

Here

$$2\phi 1\left(\begin{matrix} q^a & q^b \\ & q^c \end{matrix};p,x\right)=\sum_{n=0}^{\infty}\frac{(q^a;p)_n(q^b;p)_n}{(p;p)_n(q^c;p)_n}x^n,$$
$$(a;p)_n=\prod_{k=0}^{n-1}(1-ap^k),\quad (a;p)_0=1. \tag{IV.25}$$

B. Elliptic quantum supergroups of vertex type

As we mentioned before, a given Kac–Moody superalgebras $\mathcal{G}$ allows many inequivalent simple root systems. By means of the ''extended'' Weyl transformation method introduced in Ref. 22, one can transform from one simple root system to another inequivalent one.[23] For $\mathcal{G}=\mathrm{sl}(\widehat{n|n})$, there exists a simple root system in which all simple roots are odd (or fermionic). This system can be constructed from the distinguished simple root system by using the ''extended'' Weyl operation repeatedly. We find the following simple roots, all of which are odd (or fermionic),

$$\alpha_0=\delta-\varepsilon_1+\delta_n,$$
$$\alpha_{2j}=\delta_j-\varepsilon_{j+1},\quad j=1,2,...,n-1, \tag{IV.26}$$
$$\alpha_{2i-1}=\varepsilon_i-\delta_i,\quad i=1,2,...,n$$

with δ, $\{\varepsilon_i\}_{i=1}^n$ and $\{\delta_i\}_{i=1}^n$ satisfying

$$(\delta,\delta)=(\delta,\varepsilon_i)=(\delta,\delta_i)=0,\quad (\varepsilon_i,\varepsilon_j)=\delta_{ij},$$
$$(\delta_i,\delta_j)=-\delta_{ij},\quad (\varepsilon_i,\delta_j)=0. \tag{IV.27}$$

Such a simple root system is usually called nonstandard. It seems to us that $\mathrm{sl}(\widehat{n|n})$ is the only nontwisted affine superalgebra which has a nonstandard system of simple roots, all of which are fermionic.

As will be shown below, for $\mathcal{G}=\mathrm{sl}(\widehat{n|}n)$ with the above fermionic simple roots, one can construct a different type of twistor. Following Jimbo *et al.*,[4] we say this twistor is of *vertex type*.

Let us write $h_i=\alpha_i\,(i=0,1,...,2n-1)$ with α_i given by (IV.26). We extend the Cartan subalgebra of $\mathrm{sl}(\widehat{n|}n)$ by adding to it the element $h_{\mathrm{ex}}=\Sigma_{i=1}^{n}(\varepsilon_i+\delta_i)$. A basis of the extended Cartan subalgebra is $\{h_{\mathrm{ex}},h_0,h_1,\ldots,h_{2n-1},d\}$. Denote by $\{h^{\mathrm{ex}},h^0,h^1,\ldots,h^{2n-1},c\}$ the dual basis. We have

$$h^{\mathrm{ex}}=\frac{1}{2n}\sum_{i=1}^{n}(\varepsilon_i-\delta_i),$$

$$h^{2k}=d+\sum_{i=1}^{k}(\varepsilon_i-\delta_i)-\frac{k}{n}\sum_{i=1}^{n}(\varepsilon_i-\delta_i), \tag{IV.28}$$

$$h^{2k+1}=d+\sum_{i=1}^{k+1}\varepsilon_i-\sum_{i=1}^{k}\delta_i-\frac{2k+1}{2n}\sum_{i=1}^{k}(\varepsilon_i-\delta_i),$$

where $k=0,1,...,n-1$. The canonical element $\mathcal{T}$ in the extended Cartan subalgebra reads

$$\mathcal{T}=h_{\mathrm{ex}}\otimes h^{\mathrm{ex}}+\sum_{i=0}^{2n-1}(h_i\otimes h^i)+d\otimes c. \tag{IV.29}$$

Let τ be the diagram automorphism of $U_q[\mathrm{sl}(\widehat{n|}n)]$ such that

$$\tau(e_i)=e_{i+1 \bmod 2n},\quad \tau(f_i)=f_{i+1 \bmod 2n},\quad \tau(h_i)=h_{i+1 \bmod 2n}. \tag{IV.30}$$

Obviously, the automorphism τ is nongraded since it preserves the grading of the generators and, moreover, $\tau^{2n}=1$. Then we can show

$$\tau(h_{\mathrm{ex}})=-h_{\mathrm{ex}}+\xi c,\quad \tau(c)=c,\quad \tau(h^{\mathrm{ex}})=-h^{\mathrm{ex}}+\frac{1}{2n}c,$$

$$\tau(h^{2k})=h^{2k+1 \bmod 2n}+\frac{\xi}{2n}\sum_{i=1}^{n}(\varepsilon_i-\delta_i)-\frac{\xi+n-2k-1}{2n}c, \tag{IV.31}$$

$$\tau(h^{2k+1})=h^{2k+2 \bmod 2n}+\frac{\xi}{2n}\sum_{i=1}^{n}(\varepsilon_i-\delta_i)-\frac{n-2k-1}{2n}c,$$

where $k=0,1,...,n-1$ and ξ is an arbitrary constant. Introduce element

$$\tilde{\rho}=\sum_{i=0}^{2n-1}h^i+\xi n h^{\mathrm{ex}}, \tag{IV.32}$$

which gives the principal gradation

$$[\tilde{\rho},e_i]=e_i,\quad [\tilde{\rho},f_i]=-f_i,\quad i=0,1,...,2n-1. \tag{IV.33}$$

It is easily shown that

$$\tau(\tilde{\rho})=\tilde{\rho},\quad (\tau\otimes\tau)\mathcal{T}=\mathcal{T}. \tag{IV.34}$$

Notice also that

$$(\tau\otimes\tau)\cdot\Delta=\Delta\cdot\tau, \tag{IV.35}$$

$$(\tau\otimes\tau)\mathcal{R}=\mathcal{R}.$$

Here the second relation is deduced from the uniqueness of the universal R-matrix of $U_q[sl(\widehat{n|n})]$. It can be shown that

$$\sum_{k=1}^{2n}(\tau^k\otimes 1)\mathcal{T}=\tilde{\rho}\otimes c+c\otimes\tilde{\rho}-\frac{2(n^2-1)-3\xi}{6}c\otimes c. \tag{IV.36}$$

Therefore, if we set

$$\tilde{\mathcal{T}}=\frac{1}{2n}\left(\tilde{\rho}\otimes c+c\otimes\tilde{\rho}-\frac{2(n^2-1)-3\xi}{6}c\otimes c\right), \tag{IV.37}$$

then we have

$$\sum_{k=1}^{2n}(\tau^k\otimes 1)(\mathcal{T}-\tilde{\mathcal{T}})=0. \tag{IV.38}$$

Introduce an automorphism

$$\tilde{\phi}_r=\tau\cdot \mathrm{Ad}\,(q^{[(r+c)/n]\tilde{\rho}}), \tag{IV.39}$$

which depends on a parameter $r\in\mathbf{C}$. Then the $2n$-fold product

$$\prod_{2n\geqslant k\geqslant 1}^{\leftarrow}(\tilde{\phi}_r^k\otimes 1)(q^{\tilde{\mathcal{T}}}\mathcal{R})^{-1}. \tag{IV.40}$$

is a formal power series in $p^{1/2n}$ where $p=q^{2r}$. Moreover, it has leading term 1 thanks to the relation (IV.38). Following Jimbo *et al.*,[4] we define the vertex-type twistor

$$E(r)=\lim_{N\to\infty}\prod_{2nN\geqslant k\geqslant 1}^{\leftarrow}(\tilde{\phi}_r^k\otimes 1)(q^{\tilde{\mathcal{T}}}\mathcal{R})^{-1}. \tag{IV.41}$$

Then one can show that $E(r)$ satisfies the graded shifted cocycle condition

$$E_{12}(r)(\Delta\otimes 1)E(r)=E_{23}(r+c^{(1)})(1\otimes\Delta)E(r). \tag{IV.42}$$

Moreover, $E(r)$ obeys the counit property

$$(\epsilon\otimes 1)E(r)=(1\otimes\epsilon)E(r)=1. \tag{IV.43}$$

We have the following.

Definition 6 (Vertex-type elliptic quantum supergroup): We define elliptic quantum supergroup $\mathcal{A}_{q,p}[sl(\widehat{n|n})]$ *of vertex type to be the quasi-triangular quasi-Hopf superalgebra* $(U_q[sl(\widehat{n|n})],\Delta_r,\epsilon,\Phi(r),\mathcal{R}(r))$ *together with the graded algebra anti-homomorphism S defined by (III.2) and* $\alpha_r=m\cdot(S\otimes 1)E(r)^{-1}$, $\beta_r=m\cdot(1\otimes S)E(r)$. *Here* ϵ *is defined by (III.2), and*

$$\Delta_r(a)=E(r)\Delta(a)E(r)^{-1},\quad \forall a\in U_q[sl(\widehat{n|n})],$$

$$\mathcal{R}(r)=E(r)^T\mathcal{R}E(r)^{-1}, \tag{IV.44}$$

$$\Phi(r)=E_{23}(r+c^{(1)})E_{23}(r)^{-1}.$$

Similar to the face-type case, introduce a formal parameter ζ (or spectral parameter) into $\mathcal{R}$ and $E(r)$ by the formulas

$$\tilde{\mathcal{R}}(\zeta)=\mathrm{Ad}\,(\zeta^{\tilde{\rho}}\otimes 1)\mathcal{R},$$
$$E(\zeta,r)=\mathrm{Ad}\,(\zeta^{\tilde{\rho}}\otimes 1)E(r), \tag{IV.45}$$
$$\tilde{\mathcal{R}}(\zeta,r)=\mathrm{Ad}\,(\zeta^{\tilde{\rho}}\otimes 1)\mathcal{R}(r)=E(\zeta^{-1},r)^T\tilde{\mathcal{R}}(\zeta)E(\zeta,r)^{-1}.$$

Then it can be shown from the definition of $E(r)$ that $E(\zeta,r)$ satisfies the difference equation

$$E(p^{1/2n}q^{(1/n)c^{(1)}}\zeta,r)=(\tau\otimes 1)^{-1}(E(\zeta,r))\cdot q^{\tilde{T}}\tilde{\mathcal{R}}(p^{1/2n}q^{(1/n)c^{(1)}}\zeta), \tag{IV.46}$$

$$E(0,r)=1. \tag{IV.47}$$

The initial condition follows from (IV.38) and the fact that we are working in the principal gradation. Equation (IV.46) implies that

$$E((p^{1/2n}q^{(1/n)c^{(1)}})^{2n}\zeta,r)=E(\zeta,r))\cdot\prod_{2n-1\geqslant k\geqslant 0}^{\leftarrow} q^{\tilde{T}}(\tau\otimes 1)^{2n-k}\tilde{\mathcal{R}}((p^{1/2n}q^{(1/n)c^{(1)}})^{2n-k}\zeta). \tag{IV.48}$$

Some remarks are in order. In nonsuper case,[4] π and τ are commutable in the sense that $\pi\cdot\tau=\mathrm{Ad}\,(h)\cdot\pi$ with h obeying $hv_i=v_{i+1\ \mathrm{mod}\ m}$, where $\{v_i\}$ are basis of the vector module $V=\mathbf{C}^m=\mathbf{C}v_1\oplus\cdots\oplus\mathbf{C}c_m$ of $\mathcal{A}_{q,p}(\hat{\mathrm{sl}}_m)$ and τ is the cyclic diagram automorphism of $\hat{\mathrm{sl}}_m$. In the super (or $\mathbf{Z}_2$ graded) case, however, π and τ are not ''commutable'' in the above sense. This is because τ is grading preserving while the $2n$-dimensional defining representation space $V=\mathbf{C}^{n|n}=\mathbf{C}v_1\oplus\cdots\oplus\mathbf{C}v_{2n}$ is graded. So to compute the image, one has to work out the action of τ at the universal level and then apply the representation π. Therefore, the knowledge of the universal R-matrix in its explicit form is required. This makes the image computation of the twistor more involved in the supersymmetric case.

As an example, consider the simplest case of elliptic quantum affine superalgebra $\mathcal{A}_{q,p}[\mathrm{sl}(\widehat{1|1})]$. Let us calculate the image in the two-dimensional representation (π,V), $V=\mathbf{C}^{1|1}$. As remarked above, we have to work at the universal level first and then apply the representation. We have the following.

Lemma 2: In the principal gradation, the action of τ on the Drinfeld generators is represented on V by

$$\tau(X_n^+)=(-1)^n z^{2n+1}q^{-n}e_{12},\quad \tau(X_n^-)=(-1)^{n+1}z^{2n-1}q^{-n}e_{21},$$
$$\tau(H_n)=(-1)^{n+1}z^{2n}\frac{[n]_q}{n}(e_{11}+e_{22}), \tag{IV.49}$$
$$\tau(H_n^{\mathrm{ex}})=(-1)^{n+1}z^{2n}\frac{[2n]_q}{n}\left(q^{-n}e_{11}+\frac{q-q^{-1}}{2}[n]_q(e_{11}+e_{22})\right).$$

Applying $\pi\otimes\pi$ to the both side of (IV.48) and writing $E_{VV}(\zeta;p)\equiv(\pi\otimes\pi)E(\zeta,r)$, where $p=q^{2r}$, we get

$$E_{VV}(p\zeta;p)=E_{VV}(\zeta;p)\cdot(\pi\otimes\pi)((\tau\otimes 1)\tilde{\mathcal{R}}(p^{1/2}\zeta))\cdot\tilde{\mathcal{R}}_{VV}(p\zeta), \tag{IV.50}$$

where $\tilde{R}_{VV}(\zeta)=(\pi\otimes\pi)\tilde{R}(\zeta)$. In view of (IV.49) and the explicit formula (III.11) of the universal R-matrix, (IV.50) is a system of eight difference equations.

We can also proceed directly. We have, with the help of Lemma 2,

$$(\pi\otimes\pi)(\tau^{2k}\otimes 1)(\mathrm{Ad}\,(p^k\zeta)^{\tilde\rho}\otimes 1)\mathcal{R}^{-1}q^{-\tilde{\mathcal{T}}}=K\cdot\overline{E}_{2k},$$
$$(\pi\otimes\pi)(\tau^{2k-1}\otimes 1)(\mathrm{Ad}\,(p^{k-1/2}\zeta)^{\tilde\rho}\otimes 1)\mathcal{R}^{-1}q^{-\tilde{\mathcal{T}}}=\rho_{2k-1}\cdot K^{-1}\cdot\overline{E}_{2k-1}, \tag{IV.51}$$

where $K=(\pi\otimes\pi)q^{\mathcal{T}}$ and

$$\rho_{2k-1}=\frac{(1+q^2p^{2k-1}\zeta^2)(1+q^{-2}p^{2k-1}\zeta^2)}{(1+p^{2k-1}\zeta^2)^2},$$

$$\begin{aligned}\overline{E}_{2k}=\frac{1}{1-q^2p^{2k}\zeta^2}&((1-q^{-2}p^{2k}\zeta^2)e_{11}\otimes e_{11}+(1-q^2p^{2k}\zeta^2)e_{22}\otimes e_{22}\\&+(1-p^{2k}\zeta^2)e_{11}\otimes e_{22}+(1-p^{2k}\zeta^2)e_{22}\otimes e_{11}\\&-(q-q^{-1})p^k\zeta e_{12}\otimes e_{21}+(q-q^{-1})p^k\zeta e_{21}\otimes e_{12}),\end{aligned} \tag{IV.52}$$

$$\begin{aligned}\overline{E}_{2k-1}=\frac{1}{1+q^{-2}p^{2k-1}\zeta^2}&((1+q^2p^{2k-1}\zeta^2)e_{11}\otimes e_{11}+(1+q^{-2}p^{2k-1}\zeta^2)e_{22}\otimes e_{22}\\&+(1+p^{2k-1}\zeta^2)e_{11}\otimes e_{22}+(1+p^{2k-1}\zeta^2)e_{22}\otimes e_{11}\\&+(q-q^{-1})p^{k-1/2}\zeta e_{12}\otimes e_{21}-(q-q^{-1})p^{k-1/2}\zeta e_{21}\otimes e_{12}).\end{aligned} \tag{IV.53}$$

Then

$$E_{VV}(\zeta;p)=\prod_{k\geqslant 1}^{\leftarrow}\rho_{2k-1}K\overline{E}_{2k}K^{-1}\overline{E}_{2k-1}=\rho(\zeta;p)(E_{VV}^1(\zeta;p)+E_{VV}^2(\zeta;p)), \tag{IV.54}$$

where

$$\rho(\zeta;p)=\frac{(-pq^2\zeta^2;p^2)_\infty}{(pq\zeta;p)_\infty(-pq\zeta;p)_\infty}, \tag{IV.55}$$

$$\begin{aligned}E_{VV}^1(\zeta;p)=\prod_{k\geqslant 1}^{\leftarrow}\frac{1}{(1+p^{2k-1}\zeta^2)^2}&((1-q^{-2}p^{2k}\zeta^2)(1+q^2p^{2k-1}\zeta^2)e_{11}\otimes e_{11}\\&+(1-q^2p^{2k}\zeta^2)(1+q^{-2}p^{2k-1}\zeta^2)e_{22}\otimes e_{22}\\&+(q-q^{-1})p^{k-1/2}\zeta(1-q^{-2}p^{2k}\zeta^2)e_{12}\otimes e_{12}\\&-(q-q^{-1})p^{k-1/2}\zeta(1-q^2p^{2k}\zeta^2)e_{21}\otimes e_{21}),\end{aligned} \tag{IV.56}$$

$$\begin{aligned}E_{VV}^2(\zeta;p)=\prod_{k\geqslant 1}^{\leftarrow}\frac{1}{1+p^{2k-1}\zeta^2}&((1-p^{2k}\zeta^2)e_{11}\otimes e_{22}+(1-p^{2k}\zeta^2)e_{22}\otimes e_{11}\\&-(q-q^{-1})p^k\zeta e_{12}\otimes e_{21}+(q-q^{-1})p^k\zeta e_{21}\otimes e_{12}).\end{aligned} \tag{IV.57}$$

The infinite product in $E_{VV}^2(\zeta;p)$ can be calculated directly and we find

$$E_{VV}^2(\zeta;p)=b_E(\zeta)(e_{11}\otimes e_{22}+e_{22}\otimes e_{11})+c_E(\zeta)(e_{12}\otimes e_{21}-e_{21}\otimes e_{12}), \tag{IV.58}$$

where

$$b_E(\zeta)\pm c_E(\zeta)=\frac{(pq^{\pm 1}\zeta;p)_\infty(-pq^{\mp 1}\zeta;p)_\infty}{(-p\zeta^2;p^2)_\infty}. \tag{IV.59}$$

As for $E^1_{VV}(\zeta;p)$, it can be written as

$$E^1_{VV}(\zeta;p)=X_{11}(\zeta;p)e_{11}\otimes e_{11}+X_{22}(\zeta;p)e_{22}\otimes e_{22}+X_{12}(\zeta;p)e_{12}\otimes e_{12}+X_{21}(\zeta;p)e_{21}\otimes e_{21}, \qquad \text{(IV.60)}$$

where $X_{ij}(\zeta;p)$ are the solution to the following system of four difference equations:

$$\begin{aligned}
X_{11}(p\zeta;p)&=\frac{1}{1-q^{-2}p^2\zeta^2}((1+q^{-2}p\zeta^2)X_{11}(\zeta;p)-p^{1/2}\zeta(q-q^{-1})X_{12}(\zeta;p)),\\
X_{12}(p\zeta;p)&=\frac{1}{1-q^{2}p^2\zeta^2}(-p^{1/2}\zeta(q-q^{-1})X_{11}(\zeta;p)+(1+q^{2}p\zeta^2)X_{12}(\zeta;p)),\\
X_{21}(p\zeta;p)&=\frac{1}{1-q^{-2}p^2\zeta^2}(p^{1/2}\zeta(q-q^{-1})X_{22}(\zeta;p)+(1+q^{-2}p\zeta^2)X_{21}(\zeta;p)),\\
X_{22}(p\zeta;p)&=\frac{1}{1-q^{2}p^2\zeta^2}((1+q^{2}p\zeta^2)X_{22}(\zeta;p)+p^{1/2}\zeta(q-q^{-1})X_{21}(\zeta;p)).
\end{aligned} \qquad \text{(IV.61)}$$

ACKNOWLEDGMENTS

Our interest in quasi-Hopf algebras was ignited by Bo-Yu Hou's beautiful lecture on the Jimbo *et al.* paper[4] when YZZ was visiting Northwest University, Xi'an, in December 1997. YZZ thanks Bo-Yu Hou for interesting him in the subject and for helpful discussions. The financial support from the Australian Research Council through a Queen Elizabeth II Fellowship Grant for YZZ is also gratefully acknowledged.

[1] V. G. Drinfeld, Leningrad Math. J. **1**, 1419 (1990).
[2] O. Babelon, D. Bernard, and E. Billey, Phys. Lett. B **375**, 89 (1996).
[3] C. Fronsdal, Lett. Math. Phys. **40**, 134 (1997).
[4] M. Jimbo, H. Konno, S. Odake, and J. Shiraishi, e-print q-alg/9712029.
[5] D. Arnaudon, E. Buffenoir, E. Ragoucy, and P. Roche, e-print q-alg/9712037.
[6] B. Enriquez and G. Felder, e-print q-alg/9703018.
[7] O. Foda, K. Iohara, M. Jimbo, R. Kedem, T. Miwa, and H. Yan, Lett. Math. Phys. **32**, 259 (1994).
[8] G. Felder, Elliptic quantum groups, in Proc. ICMP Paris 1994 (International Press, Boston), p. 211.
[9] R. J. Baxter, Ann. Phys. (N.Y.) **70**, 193 (1972).
[10] G. E. Andrews, R. J. Baxter, and P. J. Forrester, J. Stat. Phys. **35**, 193 (1984).
[11] A. Belavin, Nucl. Phys. B **180**, 189 (1981).
[12] M. Jimbo, T. Miwa, and M. Odake, Commun. Math. Phys. **116**, 507 (1988).
[13] V. V. Bazhanov and Y. G. Stroganov, Theor. Math. Phys. **62**, 253 (1985).
[14] T. Deguchi and A. Fujii, Mod. Phys. Lett. A **6**, 3413 (1991).
[15] V. G. Kac, Adv. Math. **26**, 8 (1977).
[16] V. G. Kac, Lect. Notes Math. **676**, 597 (1978).
[17] S. M. Khoroshkin and V. N. Tolstoy, Commun. Math. Phys. **141**, 599 (1991).
[18] V. G. Drinfeld, Sov. Math. Dokl. **36**, 212 (1988).
[19] Y.-Z. Zhang, J. Phys. A **30**, 8325 (1997), q-alg/9703020.
[20] A. J. Bracken, G. W. Delius, M. D. Gould, and Y.-Z. Zhang, J. Phys. A **27**, 6551 (1994).
[21] I. B. Frenkel and N. Yu. Reshetikhin, Commun. Math. Phys. **146**, 1 (1992).
[22] D. A. Leites, M. V. Saveliev, and V. V. Serganova, Serpukhov preprint 85–81 (1985).
[23] L. Frappat, A. Sciarrino, and P. Sorba, Commun. Math. Phys. **121**, 457 (1989).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 10 OCTOBER 1999

Comment on ‘‘Exact periodic solutions of the complex Ginzburg–Landau equation’’ [J. Math. Phys. 40, 884 (1999)]

Robert Conte
Service de Physique de l'État Condensé, CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France

Micheline Musette
Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, B-1050 Brussel, Belgium

(Received 17 March 1999; accepted for publication 17 May 1999)

[S0022-2488(99)03610-5]

In this recent paper, three solutions to the complex Ginzburg–Landau equation, numbered (18), (21), (24), were presented as new, while none of them are new, as now detailed. Solutions (18) and (21) are only defined in the case Im(p/q)=0, a case which should be listed as ‘‘degenerate’’ on page 884 since this is a confluence of the singularities of the solutions.[1,2]

The elliptic solution (18) was given earlier by Cariello and Tabor[1] [their formulas (3.7), (3.14a), and (3.14b), which in the case Im(p/q)=0 makes $|u|^2$ linear in the Weierstrass elliptic function].

Solutions (21) and (24) are neither elliptic, since the discriminant $\Delta \equiv g_2^3 - 27 g_3^2$ vanishes, nor rational, since $|u|^2$ reduces to a second degree polynomial in tanh($k\zeta$) with constant coefficients

$$|u|^2 = A_2[(k \tanh(k\zeta))^2 + c_1 k \tanh(k\zeta) + c_2], \quad A_2 k \neq 0, \quad \zeta = x - ct. \tag{1}$$

Such a class has been extensively investigated by many authors and there is no hope for any new result. Indeed, solution (21) is the front of Nozaki and Bekki,[3]

$$|u|^2 = A_2(k(\tanh(k\zeta) \pm 1))^2, \tag{2}$$

with however the constraint Im(p/q)=0. As to solution (24), it is either again the front of Nozaki and Bekki without any constraint [in the curious formula (34), the parameter S is zero], or the propagating hole of Bekki and Nozaki,[4]

$$|u|^2 = A_2[(k \tanh(k\zeta) + X)^2 + Y^2], \tag{3}$$

with however, the constraint $k^2 = X^2$ which gives to the velocity c a value only depending on (p, q, γ) while it is arbitrary in Ref. 4.

[1] F. Cariello and M. Tabor, Physica D **53**, 59 (1991).
[2] R. Conte and M. Musette, Physica D **69**, 1 (1993).
[3] K. Nozaki and N. Bekki, J. Phys. Soc. Jpn. **53**, 1581 (1984).
[4] N. Bekki and K. Nozaki, Phys. Lett. A **110**, 133 (1985).

Response to ''Comment on 'Exact periodic solutions of the complex Ginzburg–Landau equation' '' [J. Math. Phys. 40, 5283 (1999)]

Alexei V. Porubov and Manuel G. Velarde
Instituto Pluridisciplinar, UCM, Paseo Juan XXIII, 28040 Madrid, Spain

(Received 20 April 1999; accepted for publication 17 May 1999)

[S0022-2488(99)03710-X]

The condition $\mathrm{Im}(p/q)=0$ is insufficient for the existence of our solution (18) because we have to assume additionally $p_i \neq 0$, $q_i \neq 0$. There is no periodic solution in Sec. 3 of Ref. 1 to the complex Ginzburg–Landau equation. Case $\mathrm{Im}(p/q)=0$ is not studied there. Even the authors of Ref. 1 conclude in the paragraph below Eqs. (3.14), p. 63, that they have not been able to find significant new solutions.

Section V of our paper is devoted to the comparison of solutions (21) and (24) with those obtained by Bekki and Nozaki. However, our solutions admit representations different from bounded tanh solutions. Indeed, using Eq. (4) of our paper we can get in the limit $k \to 1$ the relationship $\wp \sim \coth^2(\kappa\zeta)$ giving an unbounded limit of our solutions (21), (24). Hence, our solutions do not coincide with those obtained by Bekki and Nozaki.

We think that the ''curious'' formula (34) is useful because of the two possibilities for the parameter S, one of them, see (27), is not zero.

[1] F. Cariello and M. Tabor, Physica D **53**, 59 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

On the Schrödinger equation with steplike potentials

Tuncay Aktosun[a)]
Department of Mathematics, North Dakota State University, Fargo, North Dakota 58105

(Received 19 July 1999; accepted for publication 2 August 1999)

The one-dimensional Schrödinger equation is considered when the potential is asymptotic to a positive constant on the right half line in a certain sense. The zero-energy limits of the scattering coefficients are obtained under weaker assumptions than used elsewhere, and the continuity of the scattering coefficients from the left are established. The scattering coefficients for the potential are expressed in terms of the corresponding coefficients for the pieces of the potential on the positive and negative half lines. The number of bound states for the whole potential is related to the number of bound states for the two pieces. Finally, an improved result is given on the small-energy asymptotics of reflection coefficients for potentials supported on a half line. © *1999 American Institute of Physics.*
[S0022-2488(99)03611-7]

I. INTRODUCTION

Consider the one-dimensional Schrödinger equation

$$\psi''(k,x)+k^2\psi(k,x)=V(x)\psi(k,x),\quad x\in\mathbf{R},\tag{1.1}$$

where the potential V is real valued and may be written as $V=V_1+V_2$ such that V_1 has support in $\mathbf{R}^-$, V_2 has support in $\mathbf{R}^+$, and

$$V_1\in L_1^1(\mathbf{R}^-),\quad V_2-c^2\in L_1^1(\mathbf{R}^+),\tag{1.2}$$

for some positive c. Here the prime denotes the derivative with respect to the spatial variable x, $\mathbf{R}^-=(-\infty,0)$, $\mathbf{R}^+=(0,+\infty)$, and $L_\alpha^1(I)$ is the set of measurable functions f on an interval I such that $\int_I dx(1+|x|)^\alpha|f(x)|$ is finite. We will use $\mathbf{C}^+$ to denote the upper half complex plane and $\overline{\mathbf{C}^+}=\mathbf{C}^+\cup\mathbf{R}$.

The scattering problem for (1.1) consists of the analysis of the scattering coefficients corresponding to the potential V. Such an analysis was given by Buslaev and Fomin[1] and by Cohen and Kappeler;[2] however, in Ref. 1 only the generic case was considered, and in Ref. 2 the exceptional case was considered under the stronger assumption of L_2^1 instead of L_1^1 in (1.2). The definition of an exceptional potential is given in Sec. II; informally speaking, an exceptional potential has a ''half-bound state'' at zero energy, or equivalently it is at the boundary of changing the number of its bound states by one. The bound states of (1.1) are its square-integrable solutions.

Our primary aim is to consider the small-k asymptotics of the scattering coefficients in the exceptional case and analyze their continuity at $k=0$ by assuming only (1.2). One consequence of our analysis is that under (1.2), the number of bound states is finite. We present a Levinson theorem relating the number of bound states to the zero-energy limit of the phase of the transmission coefficient, and relate the scattering coefficients corresponding to V, V_1, and V_2 to each other.

The inverse scattering problem for (1.1) is equivalent to the recovery of the potential in terms of an appropriate set of scattering data. Such problems were analyzed in Refs. 1–3. Our result in

[a)]Electronic mail: aktosun@plains.nodak.edu

Theorem 3.4 is expected to have an impact on the inverse scattering theory for (1.1) because it was used as a hypothesis in Ref. 2 to obtain various results in the analysis of the inverse scattering problem for (1.1).

Our results are also expected to have an impact on the phase recovery problem, a version of the inverse scattering problem for (1.1) with important applications[4–6] in the recovery of material properties of thin films. Mathematically speaking, one is interested in the recovery of V_2 by using only V_1 and the reflectivity measurements, i.e., the amplitudes of reflection coefficients without their phases. In reality, the phase of the complex-valued reflection coefficient cannot be measured, even though the reflectivity is easily measured[5–8] by using a device known as the reflectometer. Our analysis of the scattering coefficients for V in terms of those for V_1 and V_2 helps us to solve the phase recovery problem by the so-called two-layer method[9,10] using 33% less data than the so-called three-layer method.[9,11–13]

Our paper is organized as follows. In Sec. II we introduce the Jost solutions and scattering coefficients for (1.1), explain the distinction between the generic and exceptional cases, and obtain the small-k asymptotics of the Jost solution from the left and of its x derivative. In Sec. III, in the exceptional case, we prove that the Wronskian defined in (2.10) vanishes linearly as $k\to 0$ in $\overline{\mathbf{C}^+}$; the proof is nontrivial, but the result is significant and it enables us to obtain the small-k asymptotics of the scattering coefficients and establish the continuity at $k=0$ of the scattering coefficients from the left. In Sec. IV we present a Levinson theorem, relating the number of bound states to the zero-energy phase of the transmission coefficient. Section V explores the relation among the scattering coefficients for V, V_1, and V_2. In Sec. VI the small-k limits of the scattering coefficients for V_1 and V_2 are given, and in Sec. VII such limits are related to the corresponding limits for V. In Sec. VII it is also shown that, except for one special case, one can derive the small-k limits of the scattering coefficients for V in terms of the corresponding limits for V_1 and V_2; in the special case, namely when both V_1 and V_2 are generic and V is exceptional, for such a derivation one needs to know that (3.52) holds for some nonzero α even though the value of α is not needed. In Sec. VII we also relate the number of bound states for V to the corresponding numbers for V_1 and V_2, and show that the former number is one less than or equal to the sum of the numbers of bound states for V_1 and V_2. Finally, in Sec. VIII, when $c=0$ in (1.2), we present an improved result on the small-k asymptotics of the reflection coefficient from the right (left) for a potential supported on the left (right) half line in the generic case; this is done by reconsidering the special case in Sec. VI and relating the value of α in (3.52) to the parameters corresponding to V_1 and V_2. The small-energy expansions of the reflection coefficients given in Sec. VIII are expected to simplify various proofs in the direct and inverse scattering theory for the Schrödinger equation with potentials belonging to $L_1^1(\mathbf{R})$.

II. PRELIMINARIES

The scattering states of (1.1) correspond to solutions behaving like $e^{\pm ikx}$ as $x\to -\infty$ and like $e^{\pm i\gamma x}$ as $x\to +\infty$, where

$$\gamma := \sqrt{k^2-c^2}, \tag{2.1}$$

and the branch of the square root function is used with $\operatorname{Im}\gamma \geqslant 0$. Thus, when $k\in(-c,c)$, γ defined in (2.1) is purely imaginary and is given by $\gamma = i\sqrt{c^2-k^2}$. The mapping $k\mapsto\gamma$ is analytic from $\mathbf{C}^+$ to itself and is continuous on $\overline{\mathbf{C}^+}$. The inverse mapping $\gamma\mapsto k$ is analytic only in $\gamma\in\mathbf{C}^+\setminus i(0,c]$ and is continuous only in $\gamma\in\overline{\mathbf{C}^+}\setminus i[0,c)$.

The Jost solution from the left, $f_l(k,x)$, associated with V is the solution of (1.1) satisfying

$$e^{-i\gamma x}f_l(k,x)=1+o(1), \quad e^{-i\gamma x}f_l'(k,x)=i\gamma+o(1), \quad x\to+\infty. \tag{2.2}$$

Similarly, $f_r(k,x)$, the Jost solution from the right, is defined as the solution of (1.1) satisfying

$$e^{ikx}f_r(k,x)=1+o(1), \quad e^{ikx}f_r'(k,x)=-ik+o(1), \quad x\to-\infty. \tag{2.3}$$

The transmission and reflection coefficients from the left, T and L, can be defined in terms of the spatial asymptotics of $f_l(k,x)$ as

$$e^{-ikx}f_l(k,x)=\frac{1}{T(k)}+\frac{L(k)}{T(k)}e^{-2ikx}+o(1),\quad x\to-\infty.$$

Similarly, the transmission and reflection coefficients from the right can be defined by using the asymptotics of $f_r(k,x)$ as $x\to+\infty$; however, these coefficients can be expressed[1,2] in terms of T and L, and they are not essential in our analysis. We will never need the transmission coefficient from the right which is equal to $\gamma T(k)/k$, and the reflection coefficient from the right is used only in Theorem 3.5 and is given in (3.53). If $c=0$ then the transmission coefficients from the left and from the right are the same, but they are different if $c\neq 0$. Further properties of these coefficients can be found in Refs. 1 and 2.

In terms of the Jost solutions, we define the Faddeev functions $m_l(k,x)$ and $m_r(k,x)$:

$$m_l(k,x):=e^{-i\gamma x}f_l(k,x),\quad m_r(k,x):=e^{ikx}f_r(k,x).\tag{2.4}$$

From (2.2), (2.3), and (2.4) it follows that

$$m_l(k,x)=1+\frac{1}{2i\gamma}\int_x^\infty dy[e^{2i\gamma(y-x)}-1][V(y)-c^2]m_l(k,y),\tag{2.5}$$

$$m_l'(k,x)=-\int_x^\infty dy\, e^{2i\gamma(y-x)}[V(y)-c^2]m_l(k,y),\tag{2.6}$$

$$m_r(k,x)=1+\frac{1}{2ik}\int_{-\infty}^x dy[e^{2ik(x-y)}-1]V(y)m_r(k,y),\tag{2.7}$$

$$m_r'(k,x)=\int_{-\infty}^x dy\, e^{2ik(x-y)}V(y)m_r(k,y).\tag{2.8}$$

Proposition 2.1: Assume (1.2) is satisfied for some $c>0$. Then, for each fixed $x\in\mathbf{R}$, the functions $f_l(k,x)$ and $f_l'(k,x)$ are analytic in $\gamma\in\mathbf{C}^+$. Consequently, as $k\to 0$ in $\overline{\mathbf{C}^+}$, we have

$$f_l(k,0)=f_l(0,0)+O(k^2),\quad f_l'(k,0)=f_l'(0,0)+O(k^2).\tag{2.9}$$

Proof: The analyticity in $\gamma\in\mathbf{C}^+$ can be proved by iterating the Volterra integrals (2.5) and (2.6) and using (2.4). By (2.1), $k=0$ corresponds to $\gamma=ic$. Expanding $f_l(k,0)$ and $f_l'(k,0)$ in γ at $\gamma=ic$, we obtain (2.9). ■

Define

$$W(k):=\frac{2ik}{T(k)}=[f_r(k,x);f_l(k,x)],\tag{2.10}$$

where $[f;g]:=fg'-f'g$ denotes the Wronskian. Recall that the Wronskian of any two solutions of (1.1) is independent of x and depends only on k. Generically, $W(0)\neq 0$, and $f_l(0,x)$ and $f_r(0,x)$ are linearly independent. In the exceptional case, $f_l(0,x)$ and $f_r(0,x)$ are linearly dependent and hence $W(0)=0$. We will say that V is a generic (exceptional) potential if the generic (exceptional) case occurs. By (2.4), (2.5), and (2.7), both $f_l(0,x)$ and $f_r(0,x)$ are real valued. In the exceptional case, there exists a real nonzero constant α such that

$$\alpha=\frac{f_l(0,x)}{f_r(0,x)},\quad x\in\mathbf{R}.\tag{2.11}$$

III. ANALYSIS OF $W(k)$ IN THE EXCEPTIONAL CASE

In this section we analyze $W(k)$ in the exceptional case and show that $W(k)/k$ has a nonzero limit as $k\to 0$. The existence of such a limit was used as a hypothesis in many theorems in Ref. 2, and it was proved there only under the stronger assumption L_2^1 instead of L_1^1 used in (1.2). Our aim is to evaluate this limit under (1.2) alone. For the proof we proceed as in the Appendix of Ref. 14, where the method was first used in Ref. 15 for the Schrödinger equation with $c=0$.

As a first step, let us define the solutions of (1.1), $s(k,x)$, and $v(k,x)$, satisfying the boundary conditions

$$s(k,0)=1,\quad s'(k,0)=0;\quad v(k,0)=0,\quad v'(k,0)=1. \tag{3.1}$$

In fact, these solutions satisfy

$$s(k,x)=\begin{cases}\cos kx+\dfrac{1}{k}\displaystyle\int_x^0 dy\,\sin k(y-x)V(y)s(k,y), & x\leqslant 0,\\[2ex] \cos\gamma x+\dfrac{1}{\gamma}\displaystyle\int_0^x dy\,\sin\gamma(x-y)[V(y)-c^2]s(k,y), & x\geqslant 0,\end{cases} \tag{3.2}$$

$$v(k,x)=\begin{cases}\dfrac{\sin kx}{k}+\dfrac{1}{k}\displaystyle\int_x^0 dy\,\sin k(y-x)V(y)v(k,y), & x\leqslant 0,\\[2ex] \dfrac{\sin\gamma x}{\gamma}+\dfrac{1}{\gamma}\displaystyle\int_0^x dy\,\sin\gamma(x-y)[V(y)-c^2]v(k,y), & x\geqslant 0.\end{cases} \tag{3.3}$$

Note also that

$$s'(k,x)=\begin{cases}-k\sin kx-\displaystyle\int_x^0 dy\,\cos k(y-x)V(y)s(k,y), & x\leqslant 0,\\[2ex] -\gamma\sin\gamma x+\displaystyle\int_0^x dy\,\cos\gamma(x-y)[V(y)-c^2]s(k,y), & x\geqslant 0,\end{cases} \tag{3.4}$$

$$v'(k,x)=\begin{cases}\cos kx-\displaystyle\int_x^0 dy\,\cos k(y-x)V(y)v(k,y), & x\leqslant 0,\\[2ex] \cos\gamma x+\displaystyle\int_0^x dy\,\cos\gamma(x-y)[V(y)-c^2]v(k,y), & x\geqslant 0.\end{cases} \tag{3.5}$$

Using (3.1) we get

$$f_l(k,0)=[f_l(k,x);v(k,x)],\quad f_l'(k,0)=-[f_l(k,x);s(k,x)], \tag{3.6}$$

$$f_r(k,0)=[f_r(k,x);v(k,x)],\quad f_r'(k,0)=-[f_r(k,x);s(k,x)]. \tag{3.7}$$

From (3.2) and (3.4), as $x\to\pm\infty$ we obtain

$$s(k,x)=\frac{e^{ikx}A_1(k)}{2ik}+\frac{e^{-ikx}A_2(k)}{2ik}+o(1),\quad x\to-\infty, \tag{3.8}$$

$$s'(k,x)=\frac{e^{ikx}A_1(k)}{2}-\frac{e^{-ikx}A_2(k)}{2}+o(1),\quad x\to-\infty, \tag{3.9}$$

$$s(k,x)=\frac{e^{i\gamma x}A_3(k)}{2i\gamma}+\frac{e^{-i\gamma x}A_4(k)}{2i\gamma}+o(1),\quad x\to+\infty, \tag{3.10}$$

$$s'(k,x)=\frac{e^{i\gamma x}A_3(k)}{2}-\frac{e^{-i\gamma x}A_4(k)}{2}+o(1),\quad x\to+\infty, \tag{3.11}$$

where we have defined

$$A_1(k):=ik-\int_{-\infty}^{0}dy\, e^{-iky}V(y)s(k,y), \tag{3.12}$$

$$A_2(k):=ik+\int_{-\infty}^{0}dy\, e^{iky}V(y)s(k,y),$$

$$A_3(k):=i\gamma+\int_{0}^{\infty}dy\, e^{-i\gamma y}[V(y)-c^2]s(k,y),$$

$$A_4(k):=i\gamma-\int_{0}^{\infty}dy\, e^{i\gamma y}[V(y)-c^2]s(k,y). \tag{3.13}$$

Similarly, from (3.3) and (3.5), as $x\to\pm\infty$ we obtain

$$v(k,x)=\frac{e^{ikx}A_5(k)}{2ik}-\frac{e^{-ikx}A_6(k)}{2ik}+o(1),\quad x\to-\infty, \tag{3.14}$$

$$v'(k,x)=\frac{e^{ikx}A_5(k)}{2}+\frac{e^{-ikx}A_6(k)}{2}+o(1),\quad x\to-\infty, \tag{3.15}$$

$$v(k,x)=\frac{e^{i\gamma x}A_7(k)}{2i\gamma}-\frac{e^{-i\gamma x}A_8(k)}{2i\gamma}+o(1),\quad x\to+\infty, \tag{3.16}$$

$$v'(k,x)=\frac{e^{i\gamma x}A_7(k)}{2}+\frac{e^{-i\gamma x}A_8(k)}{2}+o(1),\quad x\to+\infty, \tag{3.17}$$

where we have defined

$$A_5(k):=1-\int_{-\infty}^{0}dy\, e^{-iky}V(y)v(k,y), \tag{3.18}$$

$$A_6(k):=1-\int_{-\infty}^{0}dy\, e^{iky}V(y)v(k,y),$$

$$A_7(k):=1+\int_{0}^{\infty}dy\, e^{-i\gamma y}[V(y)-c^2]v(k,y),$$

$$A_8(k):=1+\int_{0}^{\infty}dy\, e^{i\gamma y}[V(y)-c^2]v(k,y). \tag{3.19}$$

Evaluating the Wronskians in (3.6) as $x\to+\infty$ and by using (2.2), (3.10), (3.11), (3.16), and (3.17), we get

$$f_l(k,0)=A_8(k), \quad f_l'(k,0)=A_4(k). \tag{3.20}$$

Similarly, evaluating the Wronskians in (3.7) as $x\to-\infty$ and by using (2.3), (3.8), (3.9), (3.14), and (3.15), we have

$$f_r(k,0)=A_5(k), \quad f_r'(k,0)=-A_1(k). \tag{3.21}$$

Now let $\phi(k,x)$ be the solution of (1.1) satisfying

$$\phi(k,0)=f_l(0,0), \quad \phi'(k,0)=f_l'(0,0). \tag{3.22}$$

For the arguments in the rest of this section, there is no loss of generality in assuming that $f_l(0,0)\neq 0$; if $f_l(0,0)=0$, the proofs can be modified as in Ref. 14 to get the results given in Theorems 3.4 and 3.5. Because $\phi(0,x)$ and $f_l(0,x)$ are solutions of the same differential equation with the same initial conditions given in (3.22), we have

$$\phi(0,x)=f_l(0,x), \quad x\in\mathbf{R}. \tag{3.23}$$

Using (2.11) and (3.23) we see that in the exceptional case $\phi(0,x)$ remains bounded as $x\to\pm\infty$; this is because $f_l(0,x)$ and $f_r(0,x)$ remain bounded as $x\to+\infty$ and $x\to-\infty$, respectively. From (3.1) and (3.22) it follows that

$$\phi(k,x)=f_l(0,0)s(k,x)+f_l'(0,0)v(k,x). \tag{3.24}$$

Our aim is to express $W(k)$ defined in (2.10) in terms of $\phi(k,x)$. Evaluating the Wronskian in (2.10) at $x=0$ and using (3.20) and (3.21), we obtain

$$f_l(0,0)W(k)=f_r(k,0)f_l(0,0)A_4(k)+f_l(k,0)f_l(0,0)A_1(k). \tag{3.25}$$

Using (3.12), (3.18), (3.21), and (3.24), we have

$$f_l(0,0)A_1(k)=ikf_l(0,0)+f_l'(0,0)-f_l'(0,0)f_r(k,0)-\int_{-\infty}^0 dy\, e^{-iky}V(y)\phi(k,y). \tag{3.26}$$

Similarly, using (3.13), (3.19), (3.20), and (3.24), we get

$$f_l(0,0)A_4(k)=i\gamma f_l(0,0)-f_l'(0,0)+f_l'(0,0)f_l(k,0)-\int_0^\infty dy\, e^{i\gamma y}[V(y)-c^2]\phi(k,y). \tag{3.27}$$

Thus, using (3.26) and (3.27) in (3.25), we obtain

$$f_l(0,0)W(k)=-f_r(k,0)M_1(k)+f_l(k,0)M_2(k). \tag{3.28}$$

where

$$M_1(k):=-i\gamma f_l(0,0)+f_l'(0,0)+\int_0^\infty dy\, e^{i\gamma y}[V(y)-c^2]\phi(k,y), \tag{3.29}$$

$$M_2(k):=ikf_l(0,0)+f_l'(0,0)-\int_{-\infty}^0 dy\, e^{-iky}V(y)\phi(k,y). \tag{3.30}$$

Proposition 3.1: Assume (1.2) is satisfied for some $c>0$. Then, $M_1(k)$ defined in (3.29) is an analytic function of $\gamma\in\mathbf{C}^+$, $M_1(0)=0$, and

$$M_1(k)=O(k^2), \quad k\to 0 \ \text{ in } \overline{\mathbf{C}^+}. \tag{3.31}$$

Proof: Using (3.20), (3.27)–(3.29), we see that $M_1(k)$ is a linear combination of $f_l(k,0)$ and $f_l'(k,0)$, and in fact

$$M_1(k) = -f_l(0,0)f_l'(k,0) + f_l'(0,0)f_l(k,0). \tag{3.32}$$

Thus, by Proposition 2.1 and (3.32), $M_1(k)$ is analytic in $\gamma \in \mathbf{C}^+$. Using its Taylor series expansion around $\gamma = ic$, which corresponds to $k=0$, we obtain (3.31) and see that $M_1(0)=0$. ■

In the following proposition and elsewhere, we will use C to denote a generic positive constant whose value is not necessarily the same in different appearances.

Proposition 3.2: Assume that we are in the exceptional case and that (1.2) holds for some $c \geqslant 0$. Then,

$$|\phi(k,x) - \phi(0,x)| \leqslant C\left(\frac{|kx|}{1+|kx|}\right)^2, \quad x \leqslant 0, \tag{3.33}$$

with $k \in [-\epsilon, \epsilon]$ for any fixed positive ϵ.

Proof: Using (3.1), (3.22), and (3.24) we obtain

$$\phi(k,x) = f_l(0,0)\cos kx + f_l'(0,0)\,\frac{\sin kx}{k} + \frac{1}{k}\int_x^0 dy\, \sin k(y-x) V(y)\phi(k,y), \quad x \leqslant 0, \tag{3.34}$$

$$\phi(0,x) = f_l(0,0) + x f_l'(0,0) + \int_x^0 dy (y-x) V(y)\phi(0,y), \quad x \leqslant 0. \tag{3.35}$$

Let us write (3.35) as

$$\phi(0,x) = B_1 + xB_2 + B_3 + B_4,$$

where

$$B_1 := f_l(0,0) + \int_{-\infty}^0 dy\, yV(y)\phi(0,y),$$

$$B_2 := f_l'(0,0) - \int_{-\infty}^0 dy\, V(y)\phi(0,y),$$

$$B_3 := x\int_{-\infty}^x dy\, V(y)\phi(0,y),$$

$$B_4 := -\int_{-\infty}^x dy\, yV(y)\phi(0,y).$$

Because $\phi(0,y)$ is bounded on $\mathbf{R}^-$ and $V \in L_1^1(\mathbf{R}^-)$, we get $B_4 = o(1)$ as $x \to -\infty$. Again using $V \in L_1^1(\mathbf{R}^-)$ and the boundedness of $\phi(0,y)$ on $\mathbf{R}^-$, with the help of

$$|B_3| \leqslant C\int_{-\infty}^x dy |x||V(y)| \leqslant C\int_{-\infty}^x dy |y||V(y)|,$$

we get $B_3 = o(1)$ as $x \to -\infty$. Since $\phi(0,x)$ remains bounded as $x \to -\infty$, the linear growth in x in (3.35) as $x \to -\infty$ cannot happen and we must have $B_2 = 0$. Hence

$$f_l'(0,0) - \int_{-\infty}^0 dy\, V(y)\phi(0,y) = 0, \tag{3.36}$$

and this leads to

$$\phi(0,x)=f_l(0,0)+\int_{-\infty}^0 dy\, yV(y)\phi(0,y)+o(1),\quad x\to-\infty. \tag{3.37}$$

From (3.34)–(3.36), we get

$$\phi(k,x)-\phi(0,x)=I_1+I_2+I_3+I_4+I_5+\frac{1}{k}\int_x^0 dy\,\sin k(y-x)V(y)[\phi(k,y)-\phi(0,y)], \tag{3.38}$$

where we have defined

$$I_1:=x\left[\frac{\sin kx}{kx}-1\right]\int_{-\infty}^x dy\, V(y)\phi(0,y), \tag{3.39}$$

$$I_2:=[\cos kx-1]f_l(0,0),\quad I_3:=\frac{\sin kx}{k}\int_x^0 dy[1-\cos ky]V(y)\phi(0,y), \tag{3.40}$$

$$I_4:=-(1-\cos kx)\int_x^0 dy\,\frac{\sin ky}{k}V(y)\phi(0,y), \tag{3.41}$$

$$I_5:=\int_x^0 dy\, y\left[\frac{\sin ky}{ky}-1\right]V(y)\phi(0,y). \tag{3.42}$$

For $z\geqslant 0$, the function $z\mapsto z/(1+z)$ is monotone increasing and we have

$$|\sin z|\leqslant\frac{Cz}{1+z},\quad \left|1-\frac{\sin z}{z}\right|\leqslant\frac{Cz^2}{(1+z)^2},\quad |1-\cos z|\leqslant\frac{Cz^2}{(1+z)^2}. \tag{3.43}$$

Hence, for $x\leqslant 0$ and $k\in[-\epsilon,\epsilon]$, from (3.39)–(3.42) we get the estimates

$$|I_j|\leqslant\frac{C|kx|^2}{(1+|kx|)^2},\quad j=1,2,3,4,5. \tag{3.44}$$

Using (3.43) and (3.44) in (3.38), we obtain

$$|\phi(k,x)-\phi(0,x)|\leqslant\frac{C|kx|^2}{(1+|kx|)^2}+\frac{C|x|}{1+|kx|}\int_x^0 dy|V(y)||\phi(k,y)-\phi(0,y)|, \tag{3.45}$$

With the help of Gronwall's lemma, from (3.45) we get (3.33). ■

Proposition 3.3: Assume V is an exceptional potential and (1.2) holds for some $c>0$, and let $M_2(k)$ be the quantity defined in (3.30). Then, as $k\to 0$ on the real axis, we have

$$f_l(k,0)M_2(k)=ik\alpha f_l(0,0)+o(k), \tag{3.46}$$

where α is the real nonzero constant given in (2.11).

Proof: Using (3.30) and (3.36) we get

$$f_l(k,0)M_2(k)=f_l(k,0)[ikf_l(0,0)+J_1+J_2], \tag{3.47}$$

where

$$J_1 := -\int_{-\infty}^{0} dy\ [e^{-iky}-1]V(y)\phi(0,y), \tag{3.48}$$

$$J_2 := -\int_{-\infty}^{0} dy\ e^{-iky}V(y)[\phi(k,y)-\phi(0,y)]. \tag{3.49}$$

Because of (3.23), in the exceptional case $\phi(0,y)$ is bounded for $y\leqslant 0$. Using the inequality

$$|e^{iz}-iz-1|\leqslant \frac{Cz^2}{1+z}, \quad z\geqslant 0,$$

from (3.48), since $V\in L_1^1(\mathbf{R}^-)$, we get

$$J_1 = ik\int_{-\infty}^{0} dy\ yV(y)\phi(0,y)+o(k). \tag{3.50}$$

Moreover, using (3.33) in (3.49) we get

$$|J_2|\leqslant C|k|\int_{-\infty}^{0} dy\ \frac{|ky|}{1+|ky|}(-y)|V(y)|,$$

which gives us $J_2=o(k)$. Hence, using (2.9) and (3.50) in (3.47), we obtain

$$f_l(k,0)M_2(k)=ikf_l(0,0)^2+ikf_l(0,0)\int_{-\infty}^{0} dy\ yV(y)\phi(0,y)+o(k). \tag{3.51}$$

Using (3.37) we can explicitly evaluate the integral on the right-hand side of (3.51). Since $f_r(0,x)=1+o(1)$ as $x\to-\infty$, with the help of (2.11), (3.23), and (3.37), we get

$$\int_{-\infty}^{0} dy\ yV(y)\phi(0,y)=\alpha-f_l(0,0),$$

and hence (3.51) reduces to (3.46). ∎

Theorem 3.4: Assume that V is an exceptional potential and that (1.2) holds for some $c>0$. Then, the Wronskian $W(k)$ defined in (2.10) satisfies $W(0)=0$ and

$$W(k)=i\alpha k+o(k), \quad k\to 0 \ \text{ in } \ \overline{\mathbf{C}^+}, \tag{3.52}$$

where α is the real nonzero constant given by (2.11).

Proof: Using (3.28), (3.31), and (3.46), we see that (3.52) holds as $k\to 0$ through real values. However, using the Phragmén–Lindelöf theorems as on p. 2927 of Ref. 14, it follows that the limit is valid also when $k\to 0$ in $\overline{\mathbf{C}^+}$. ∎

The reflection coefficient from the right for (1.1), R, is related to T and L as[1,2]

$$R(k)=-\frac{L(k)^*T(k)}{T(k)^*}, \quad k\in\mathbf{R}\setminus\{0\}, \tag{3.53}$$

where the asterisk denotes complex conjugation. The continuity of T, L, and R at $k=0$ is already known[2] in the generic case under (1.2). Next, we show that their continuity holds also in the exceptional case.

Theorem 3.5: Assume that V is an exceptional potential and that (1.2) holds for some $c>0$. Then, the scattering coefficients T, L, and R are all continuous at $k=0$, and we have

$$T(k)=\frac{2}{\alpha}+o(1), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{3.54}$$

$$L(k)=1+o(1), \quad R(k)=-1+o(1), \quad k\to 0 \text{ in } \mathbf{R}, \tag{3.55}$$

where α is the real nonzero constant given in (2.11).

Proof: From (2.10) and Theorem 3.4, we get (3.54), which also proves the continuity of T at $k=0$. Using (3.53) and the identity[1,2]

$$L(k)=\frac{T(k)}{T(k)^*}, \quad k\in[-c,c]\setminus\{0\}, \tag{3.56}$$

and the fact that α is real and nonzero, we get (3.55) and the continuity of L and R at $k=0$. ■

IV. THE LEVINSON THEOREM

In Theorem 3.5 we have proved the continuity of T and L at $k=0$ in the exceptional case. In the generic case, the continuity of these functions is already known[2] and also follows from (2.10) and (3.56), which lead to

$$T(k)=\frac{2ik}{W(0)}+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{4.1}$$

$$L(k)=-1+o(1), \quad k\to 0 \text{ in } \mathbf{R}. \tag{4.2}$$

One consequence of the continuity of T at $k=0$ is the following analog of the Levinson theorem, which relates the number of bound states to the argument of T at $k=0$.

Theorem 4.1: Assume that V satisfics (1.2) for some $c\geqslant 0$. Then the number of bound states of (1.1) is finite and given by

$$N=\frac{d}{2}+\frac{1}{\pi}[\arg T(0^+)], \tag{4.3}$$

where $d=0$ in the exceptional case and $d=1$ in the generic case, and $\arg T(k)$ denotes the continuous branch of the argument of T normalized such that $\arg T(+\infty)=0$.

Proof: The continuity of T at $k=0$ is the additional assumption used in Corollary 1.5 of Ref. 2 in order to assure that $k=0$ cannot be an accumulation point for the poles of T in $\mathbf{C}^+$ and that the number of such poles is finite. It is already known[1,2] that such poles are simple and confined to the positive imaginary axis, $1/T$ is continuous in $\overline{\mathbf{C}^+}\setminus\{0\}$, and $T(k)=1+O(1/k)$ as $k\to\infty$ in $\overline{\mathbf{C}^+}$. Thus, we have all the ingredients to proceed as in the proof of Theorem 9.1 of Ref. 16. ■

Proposition 4.2: Assume that V satisfies (1.2) for some $c>0$. Then, the real nonzero constant α defined in (2.11) in the exceptional case has the same sign as that of $e^{iN\pi}$. The sign of $W(0)$ in the generic case is the same as the sign of $e^{i(N+1)\pi}$.

Proof: In the exceptional case, comparing (3.54) and (4.3) gives us the sign of α. In the generic case, comparing (4.1) and (4.3) we get the sign of $W(0)$. ■

V. FACTORIZATION

Let V_j denote V_1 and V_2 for $j=1$ and $j=2$, respectively. We will use T_j, L_j, and R_j for the transmission coefficient from the left, the reflection coefficient from the left, and the reflection coefficient from the right, respectively, for the potential V_j. Similarly, let $f_{l;j}(k,x)$ and $f_{r;j}(k,x)$ denote the Jost solutions from the left and from the right, respectively, for V_j. As in (2.10) we will use $W_j(k)$ to denote $2ik/T_j(k)$; in the exceptional case, as in (2.11) we will use α_j to denote the nonzero real constant $f_{l;j}(0,x)/f_{r;j}(0,x)$. We will also let N_j denote the number of bound states of V_j.

Proposition 5.1: Assume that V satisfies (1.2) for some $c \geqslant 0$. Then,

$$\frac{1}{T(k)} = \frac{1 - R_1(k) L_2(k)}{T_1(k) T_2(k)}, \quad k \in \mathbf{R} \setminus \{0\}, \tag{5.1}$$

$$\frac{L(k)}{T(k)} = \frac{L_2(k) - R_1(k)^*}{T_1(k)^* T_2(k)}, \quad k \in \mathbf{R} \setminus \{0\}. \tag{5.2}$$

The result stated in Proposition 5.1 holds when only L^1 is used instead of L_1^1 in (1.2); however, we will take the limit in (5.1) and (5.2) as $k \to 0$ and hence it is more convenient to have the result stated under (1.2). Proposition 5.1 is a special case of the following factorization result whose proof can be given as in Refs. 17 and 18. Let us partition the real axis $\mathbf{R}$ into p fragments as $\mathbf{R} = \cup_{j=1}^{p} (x_{j-1}, x_j)$, where $x_0 := -\infty$, $x_p := +\infty$, and $x_{j-1} < x_j$ for $j = 1,...,p$. We can then write the potential V in terms of its fragments $V_{j-1,j}$ as

$$V(x) = \sum_{j=1}^{p} V_{j-1,j}(x), \tag{5.3}$$

where we have defined

$$V_{j-1,j}(x) := \begin{cases} V(x), & x \in (x_{j-1}, x_j), \\ 0, & x \notin (x_{j-1}, x_j). \end{cases} \tag{5.4}$$

Note that $V_{j-1,j} \in L^1(\mathbf{R})$ for $j = 1,..., \ p-1$, and the rightmost fragment $V_{p-1,p}$ satisfies $V_{p-1,p} - c^2 \in L^1(\mathbf{R})$. Let $T_{j-1,j}$ and $L_{j-1,j}$ denote the transmission and reflection coefficients from the left, respectively, for $V_{j-1,j}$. Let us define the transition matrix Λ associated with V and $\Lambda_{j-1,j}$ associated with $V_{j-1,j}$ as

$$\Lambda(k) := \begin{bmatrix} \dfrac{1}{T(k)} & \dfrac{L(k)^*}{T(k)^*} \\ \dfrac{L(k)}{T(k)} & \dfrac{1}{T(k)^*} \end{bmatrix}, \quad \Lambda_{j-1,j}(k) := \begin{bmatrix} \dfrac{1}{T_{j-1,j}(k)} & \dfrac{L_{j-1,j}(k)^*}{T_{j-1,j}(k)^*} \\ \dfrac{L_{j-1,j}(k)}{T_{j-1,j}(k)} & \dfrac{1}{T_{j-1,j}(k)^*} \end{bmatrix}.$$

From (3.56) and the identity[2]

$$1 - |L(k)|^2 = \frac{\gamma}{k} |T(k)|^2, \quad k \in \mathbf{R} \setminus (-c, c),$$

it follows that the determinant of Λ is given by

$$\det \Lambda(k) = \begin{cases} \dfrac{\gamma}{k}, & k \in \mathbf{R} \setminus (-c, c) \\ 0, & 0 < |k| \leqslant c \end{cases}.$$

The two columns in each of Λ and $\Lambda_{p,p+1}$ are identical when $0 < |k| \leqslant c$.

Theorem 5.2: Assume V satisfies (1.2) for some $c \geqslant 0$, where L^1 is used instead of L_1^1. Let Λ be the transition matrix corresponding to the potential V and let $\Lambda_{j-1,j}$ correspond to the fragment $V_{j-1,j}$ defined in (5.4). Then, we have

$$\Lambda(k) = \Lambda_{0,1}(k) \Lambda_{1,2}(k) \cdots \Lambda_{p-1,p}(k), \quad k \in \mathbf{R} \setminus \{0\}. \tag{5.5}$$

The result in Proposition 5.1 corresponds to $p = 2$ in Theorem 5.2 by using the (1,1) and (2,1) entries in the matrix equality in (5.5) and that $R_1(k) = -L_1(k)^* T_1(k) / T_1(k)^*$ for $k \in \mathbf{R} \setminus \{0\}$.

VI. ASYMPTOTICS OF SCATTERING COEFFICIENTS FOR V_1 AND V_2

In considering the potential V_1, the analog of Theorem 4.1 states that

$$\arg T_1(0^+)=\left(N_1-\frac{d_1}{2}\right)\pi, \tag{6.1}$$

where $d_1=0$ if V_1 is exceptional and $d_1=1$ if V_1 is generic. Using the boundary conditions at $x=0$ based on the continuity of $f_{r;1}(k,x)$ and $f'_{r;1}(k,x)$, we have

$$\frac{1+R_1(k)}{T_1(k)}=f_{r;1}(k,0), \quad ik\frac{-1+R_1(k)}{T_1(k)}=f'_{r;1}(k,0), \tag{6.2}$$

where $f_{r;1}(k,x)$ is the Jost solution from the right for V_1. Thus, from (6.2) it follows that

$$W_1(k):=\frac{2ik}{T_1(k)}=ikf_{r;1}(k,0)-f'_{r;1}(k,0). \tag{6.3}$$

Using the general theory,[19–21] or with the help of (2.4), (2.7), and (2.8) we have

$$f_{r;1}(k,0)=f_{r;1}(0,0)+o(1), \quad f'_{r;1}(k,0)=f'_{r;1}(0,0)+o(1), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{6.4}$$

and hence (6.3) and (6.4) give us $W_1(0)=-f'_{r;1}(0,0)$. Generically $W_1(0)\neq 0$, and in the exceptional case we have $W_1(0)=0$. Thus, generically we obtain

$$W_1(k)=-f'_{r;1}(0,0)+o(1), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+},$$

and hence from (6.2) and (6.3) we get

$$T_1(k)=-\frac{2ik}{f'_{r;1}(0,0)}+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{6.5}$$

$$R_1(k)=-1+T_1(k)f_{r;1}(k,0)=-1-2ik\mu_1+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{6.6}$$

where we have defined

$$\mu_1:=\frac{f_{r;1}(0,0)}{f'_{r;1}(0,0)}. \tag{6.7}$$

Note that μ_1 is well defined because $f'_{r;1}(0,0)=-W_1(0)\neq 0$ when V_1 is generic. In Sec. VIII we will improve the result in (6.6) by evaluating the next term in the expansion. Comparing (6.1) and (6.5) we see that the sign of $f'_{r;1}(0,0)$ is the same as the sign of $e^{iN_1\pi}$. Moreover, with the help of (2.3) we get

$$f_{r;1}(k,x)=f_r(k,x), \quad f'_{r;1}(k,x)=f'_r(k,x), \quad x\leqslant 0. \tag{6.8}$$

Now let us turn to the exceptional case. In this case, under $V_1\in L^1_1(\mathbf{R}^-)$, it is known that $W_1(k)$ vanishes linearly in k as $k\to 0$ in $\overline{\mathbf{C}^+}$. We have[22]

$$W_1(k)=\frac{ik(\alpha_1^2+1)}{\alpha_1}+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+},$$

$$T_1(k)=\frac{2\alpha_1}{\alpha_1^2+1}+o(1), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{6.9}$$

$$R_1(k) = -\frac{\alpha_1^2-1}{\alpha_1^2+1} + o(1), \quad k \to 0 \ \text{in} \ \overline{\mathbf{C}^+}, \tag{6.10}$$

where α_1 is the real nonzero constant given by

$$\alpha_1 = \frac{f_{l;1}(0,x)}{f_{r;1}(0,x)}, \quad x \in \mathbf{R}.$$

Comparing (6.1) and (6.9) we see that the sign of α_1 is the same as the sign of $e^{iN_1\pi}$. Note that $(\alpha_1^2-1)/(\alpha_1^2+1)$ is an increasing function of α_1^2 and its values are confined to the interval $(-1,1)$. Thus, $R_1(0) \in (-1,1)$ in the exceptional case.

Let us now summarize some similar results for V_2, where (1.2) holds for some $c>0$. From Theorem 4.1 we have

$$\arg T_2(0^+) = \left(N_2 - \frac{d_2}{2} \right) \pi, \tag{6.11}$$

where $d_2=0$ for the exceptional case and $d_2=1$ in the generic case. Using the continuity of the Jost solution $f_{l;2}(k,x)$ and its derivative $f'_{l;2}(k,x)$ at $x=0$, we get

$$\frac{1+L_2(k)}{T_2(k)} = f_{l;2}(k,0), \quad ik\frac{1-L_2(k)}{T_2(k)} = f'_{l;2}(k,0). \tag{6.12}$$

As in Proposition 2.1 we have

$$f_{l;2}(k,0) = f_{l;2}(0,0) + O(k^2), \quad f'_{l;2}(k,0) = f'_{l;2}(0,0) + O(k^2), \quad k \to 0 \ \text{in} \ \overline{\mathbf{C}^+}, \tag{6.13}$$

where $f_{l;2}(0,0)$ and $f'_{l;2}(0,0)$ cannot simultaneously vanish because of (2.2). Moreover, from (2.2) we obtain

$$f_{l;2}(k,x) = f_l(k,x), \quad f'_{l;2}(k,x) = f'_l(k,x), \quad x \geqslant 0. \tag{6.14}$$

With the help of (6.12), defining

$$W_2(k) := \frac{2ik}{T_2(k)} = ikf_{l;2}(k,0) + f'_{l;2}(k,0), \tag{6.15}$$

from (6.13) we have

$$W_2(k) = f'_{l;2}(0,0) + ikf_{l;2}(0,0) + O(k^2), \quad k \to 0 \ \text{in} \ \overline{\mathbf{C}^+}. \tag{6.16}$$

The generic case occurs if $W_2(0) \neq 0$; therefore, generically we have $f'_{l;2}(0,0) \neq 0$, and in the exceptional case we have $f'_{l;2}(0,0)=0$. Thus, generically, from (6.15) and (6.16) we get

$$T_2(k) = \frac{2ik}{f'_{l;2}(0,0)} + \frac{2\mu_2 k^2}{f'_{l;2}(0,0)} + O(k^3), \quad k \to 0 \ \text{in} \ \overline{\mathbf{C}^+}, \tag{6.17}$$

where μ_2 is the real constant defined as

$$\mu_2 := \frac{f_{l;2}(0,0)}{f'_{l;2}(0,0)}. \tag{6.18}$$

Comparing (6.11) and (6.17), we conclude that the sign of $f'_{l;2}(0,0)$ is the same as the sign of $e^{i(N_2+1)\pi}$. With the help of (6.12), (6.13), and (6.17) we also get

$$L_2(k) = -1 + 2ik\mu_2 + 2k^2\mu_2^2 + O(k^3), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+}. \tag{6.19}$$

In the exceptional case we have $W_2(0)=0$, i.e., $f'_{l;2}(0,0)=0$ and $f_{l;2}(0,0)\neq 0$. In this case it follows from (6.16) that $W_2(k)$ vanishes linearly in k as $k\to 0$ in $\overline{\mathbf{C}^+}$. From (6.13) and (6.15) we get

$$W_2(k) = ikf_{l;2}(0,0) + O(k^2), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+},$$

$$T_2(k) = \frac{2}{f_{l;2}(0,0)} + O(k) = \frac{2}{\alpha_2} + O(k), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{6.20}$$

where α_2 is the real nonzero constant given by

$$\alpha_2 = \frac{f_{l;2}(0,x)}{f_{r;2}(0,x)}, \quad x \in \mathbf{R},$$

and we have used the fact that $f_{r;2}(0,x)=1$ for $x\leqslant 0$. Comparing (6.11) and (6.20), we see that the sign of α_2 is the same as the sign of $e^{iN_2\pi}$. Using (6.12) and (6.20), we obtain

$$L_2(k) = 1 + O(k), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+}. \tag{6.21}$$

VII. ASYMPTOTICS OF SCATTERING COEFFICIENTS FOR *V*

In this section, using the results in Sec. VI, with the help of (5.1) and (5.2), we will derive the small-k asymptotics of T and L and compare our results with those obtained in (4.1), (4.2), and Theorem 3.5.

With the help of (5.1), let

$$F(k) := \frac{T_1(k)T_2(k)}{T(k)} = 1 - R_1(k)L_2(k), \tag{7.1}$$

and let $\omega(k)$ denote the phase of $F(k)$ as normalized in Theorem 4.1. From (7.1) we get

$$\omega(0^+) = \arg T_1(0^+) + \arg T_2(0^+) - \arg T(0^+). \tag{7.2}$$

Using (4.3), (6.1), and (6.11) in (7.2), we obtain

$$\omega(0^+) = \left(N_1 + N_2 - N - \frac{d_1 + d_2 - d}{2} \right)\pi. \tag{7.3}$$

If both V_1 and V_2 are exceptional, from (6.10), (6.21), and (7.1) we get

$$F(k) = \frac{2\alpha_1^2}{\alpha_1^2 + 1} + o(1), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{7.4}$$

and hence $\omega(0^+)=0$. Using (6.9), (6.20), (7.1), and (7.4) we have

$$T(k) = \frac{2}{\alpha_1\alpha_2} + o(1), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+}. \tag{7.5}$$

Thus, V is exceptional, $N=N_1+N_2$, $T(0^+)$ is real and nonzero, and the sign of $T(0^+)$ is the same as that of $e^{i(N_1+N_2)\pi}$, where the latter fact is obtained by using (7.5) and the signs of α_1 and α_2 determined in Sec. VI. Using (5.2), (6.9), (6.10), (6.20), (6.21), and (7.5) we also get

$$L(k) = 1 + o(1), \quad k \to 0 \text{ in } \mathbf{R}.$$

If V_1 is exceptional and V_2 is generic, then using (6.9), (6.10), (6.17), (6.19), and (7.1), we obtain

$$F(k)=\frac{2}{\alpha_1^2+1}+o(1), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+},$$

$$T(k)=\frac{2i\alpha_1 k}{f'_{l;2}(0,0)}+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}.$$

In this case, V is generic and $\omega(0^+)=0$, and hence from (7.3) we get $N=N_1+N_2$. With the help of (5.2) we also obtain

$$L(k)=-1+o(1), \quad k\to 0 \text{ in } \mathbf{R}.$$

If V_1 is generic and V_2 is exceptional, then using (6.5), (6.6), (6.20), (6.21), and (7.1) we get

$$F(k)=2+O(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+},$$

$$T(k)=-\frac{2ik}{\alpha_2 f'_{r;1}(0,0)}+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+},$$

and hence V is generic, $\omega(0^+)=0$, and from (7.3) it follows that $N=N_1+N_2$. With the help of (5.2) we also obtain

$$L(k)=-1+o(1), \quad k\to 0 \text{ in } \mathbf{R}.$$

If both V_1 and V_2 are generic, using (6.6)–(6.8), (6.14), (6.18), (6.19), and (7.1) we get

$$F(k)=2ik(\mu_2-\mu_1)+o(k)=-\frac{2ikW(0)}{f'_{r;1}(0,0)f'_{l;2}(0,0)}+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{7.6}$$

where $W(k)$ is the Wronskian given in (2.10) and we have used

$$W(k)=[f_r(k,x);f_l(k,x)]=f_{r;1}(k,0)f'_{l;2}(k,0)-f'_{r;1}(k,0)f_{l;2}(k,0). \tag{7.7}$$

Thus, we have two possibilities, namely $W(0)\neq 0$ and $W(0)=0$. If $W(0)\neq 0$, then V is generic, and in this case using (5.2), (6.5), (6.6), (6.17), (6.19), (7.1), and (7.6) we get

$$T(k)=\frac{2ik}{W(0)}+o(k), \quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{7.8}$$

$$L(k)=-1+o(1), \quad k\to 0 \text{ in } \mathbf{R},$$

agreeing with (4.1) and (4.2). From (4.3) and (7.8) it is seen that $W(0)$ has the same sign as the sign of $e^{i(N+1)\pi}$. From (7.6), since $W(0)\neq 0$, we get $\omega(0^+)=\pm\pi/2$. In Sec. VI we have seen that the sign of $f'_{r;1}(0,0)$ is the same as that of $e^{iN_1\pi}$ and the sign of $f'_{l;2}(0,0)$ is the same as that of $e^{i(N_2+1)\pi}$. Thus, in the subcase $\mu_2>\mu_1$ with $\omega(0^+)=\pi/2$, from (7.3) we get $N=N_1+N_2-1$. Similarly, in the subcase $\mu_2<\mu_1$ with $\omega(0^+)=-\pi/2$, we obtain $N=N_1+N_2$.

If both V_1 and V_2 are generic and $W(0)=0$, then V is exceptional. In this case, without consulting Theorem 3.4, by using (6.5), (6.6), (6.17), and (6.19) we can only conclude that $F(k)=o(k)$ as $k\to 0$. If we knew the expansion in (6.6) up to $o(k^2)$, then we would have determined $F(k)$ up to $o(k^2)$ as well. If we use Theorem 3.4, with the help of (3.52), (3.56), (6.5), (6.17), and (7.1) we get

$$F(k)=\frac{2\alpha k^2}{f'_{r;1}(0,0)f'_{l;2}(0,0)}+o(k^2),\quad T(k)=\frac{2}{\alpha}+o(1),\quad k\to 0 \text{ in } \overline{\mathbf{C}^+},$$
$$L(k)=1+o(1),\quad k\to 0 \text{ in } \mathbf{R}, \tag{7.9}$$

where, by (4.3), the sign of α is the same as that of $e^{iN\pi}$. Since $|R_1(k)L_2(k)|<1$ for $k\in\mathbf{R}\setminus\{0\}$, we must have $\omega(0^+)=0$. In this case, with the help of (7.3), we get $N=N_1+N_2-1$.

The above analysis shows that $N=N_1+N_2$ or $N=N_1+N_2-1$. As in Theorem 2.1 of Ref. 22, using induction we obtain the following general result.

Theorem 7.1: Assume V satisfying (1.2) for some $c\geq 0$ is partitioned into p fragments as in (5.3), and let $N_{j\text{-}1,j}$ denote the number of bound states corresponding to $V_{j\text{-}1,j}$. Then

$$1-p+\sum_{j=1}^{p} N_{j-1,j}\leq N\leq \sum_{j=1}^{p} N_{j-1,j},\quad p=1,2,\ldots .$$

VIII. SMALL-ENERGY ASYMPTOTICS OF $R_1(k)$

In this section, we will improve the asymptotics in (6.6). We will obtain the small-k asymptotics of the reflection coefficients for potentials supported on a half line up to $o(k^2)$.

The results given here are expected to contribute to better understanding of the scattering and inverse scattering theory for the Schrödinger equation with $c=0$.

Theorem 8.1: Assume V_1 is real valued, is supported in $\mathbf{R}^-$, and $V_1\in L_1^1(\mathbf{R}^-)$. Then, in the generic case we have

$$R_1(k)=-1-2ik\mu_1+2k^2\left[\mu_1^2+\frac{1}{f'_{r;1}(0,0)^2}\right]+o(k^2),\quad k\to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{8.1}$$

where μ_1 is the quantity defined in (6.7).

Proof: Given the generic potential V_1, let us choose V_2 satisfying (1.2) with $c>0$ such that $\mu_2=\mu_1$, where μ_2 is the quantity defined in (6.18). As seen from (7.6), this corresponds to having V_2 generic and V exceptional. In this case, using (7.7) at $k=0$, with the help of (2.11), (6.8), and (6.14), we obtain $\alpha=f'_{l;2}(0,0)/f'_{r;1}(0,0)$. Thus, (7.9) gives us

$$F(k)=\frac{2k^2}{f'_{r;1}(0,0)^2}+o(k^2),\quad k\to 0 \text{ in } \overline{\mathbf{C}^+}. \tag{8.2}$$

On the other hand, using $\mu_2=\mu_1$ in (6.19) we get

$$L_2(k)=-1+2ik\mu_1+2k^2\mu_1^2+O(k^3),\quad k\to 0 \text{ in } \overline{\mathbf{C}^+}. \tag{8.3}$$

Because of (7.1) we have

$$R_1(k)=\frac{1-F(k)}{L_2(k)}, \tag{8.4}$$

and using (8.2) and (8.3) in (8.4) we get (8.1). ■

When $c=0$, the Taylor series expansion in (6.19) is no longer valid. However, we can use the analog of Theorem 8.1 and use the transformation $x\to -x$ to obtain the following result.

Corollary 8.2: Assume V_2 is real valued, is supported in $\mathbf{R}^+$, and $V_2\in L_1^1(\mathbf{R}^+)$, i.e., assume that $c=0$ in (1.2). Then, in the generic case we have

$$L_2(k) = -1 + 2ik\mu_2 + 2k^2\left[\mu_2^2 + \frac{1}{f'_{l;2}(0,0)^2}\right] + o(k^2), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+},$$

where μ_2 is the quantity defined in (6.18).

ACKNOWLEDGMENTS

The author is indebted to Martin Klaus for his comments. The research leading to this article was supported in part by the National Science Foundation under Grant No. DMS-9803219.

[1] V Buslaev and V. Fomin, ''An inverse scattering problem for the one-dimensional Schrödinger equation on the entire axis,'' Vestn. Leningr. Univ., Ser. 4: Fiz., Khim. **17**, 56–64 (1962).
[2] A. Cohen and T. Kappeler, ''Scattering and inverse scattering for steplike potentials in the Schrödinger equation,'' Indiana Univ. Math. J. **34**, 127–180 (1985).
[3] J. Legendre, ''Problème inverse de Schrödinger sur la ligne avec conditions dissymmétriques et applications,'' Ph.D. thesis, Acad. Montpellier, Univ. Sci. Tech. Languedoc, 1982.
[4] X.-L. Zhou and S.-H. Chen, ''Theoretical foundations of x-ray and neutron reflectometry,'' Phys. Rep. **257**, 223–348 (1995).
[5] *Proceedings of the Workshop on Methods of Analysis and Interpretation of Neutron Reflectivity Data*, edited by G. Felcher and T. Russell [Phys. B **173** (1991)].
[6] *Proceedings of the Fourth International Conference on Surface X-ray and Neutron Scattering*, edited by G. Felcher and H. You [Phys. B **221** (1996)].
[7] G. P. Felcher, ''Of butterflies and terraces,'' Neutron News **5**, 18–22 (1994).
[8] V. O. de Haan, A. A. van Well, S. Adenwalla, and G. P. Felcher, ''Retrieval of phase information in neutron reflectometry,'' Phys. Rev. B **52**, 10831–10833 (1995).
[9] T. Aktosun and P. Sacks, ''Inversion of reflectivity data for nondecaying potentials,'' SIAM J. Appl. Math (to appear).
[10] T. Aktosun and P. Sacks, ''Phase recovery with nondecaying potentials,'' Inverse. Probl. (submitted).
[11] C. F. Majkrzak and N. F. Berk, ''Exact determination of the phase in neutron reflectometry,'' Phys. Rev. B **52**, 10827–10830 (1995).
[12] C. F. Majkrzak and N. F. Berk, ''Exact determination of the neutron reflection amplitude or phase,'' Physica B **221**, 520–523 (1996).
[13] V. O. de Haan, A. A. van Well, P. E. Sacks, S. Adenwalla, and G. P. Felcher, ''Toward the solution of the inverse problem in neutron reflectometry,'' Physica B **221**, 524–532 (1996).
[14] T. Aktosun, M. Klaus, and C. van der Mee, ''Inverse wave scattering with discontinuous wave speed,'' J. Math. Phys. **36**, 2880–2928 (1995).
[15] M. Klaus, ''Low-energy behaviour of the scattering matrix for the Schrödinger equation on the line,'' Inverse Probl. **4**, 505–512 (1988).
[16] T. Aktosun, M. Klaus, and C. van der Mee, ''Wave scattering in one dimension with absorption,'' J. Math. Phys. **39**, 1957–1992 (1998).
[17] T. Aktosun, ''A factorization of the scattering matrix for the Schrödinger equation and for the wave equation in one dimension,'' J. Math. Phys. **33**, 3865–3869 (1992).
[18] T. Aktosun, M. Klaus, and C. van der Mee, ''Factorization of scattering matrices due to partitioning of potentials in one-dimensional Schrödinger-type equations,'' J. Math. Phys. **37**, 5897–5915 (1996).
[19] L. D. Faddeev, ''Properties of the *S*-matrix of the one-dimensional Schrödinger equation,'' Am. Math. Soc. Transl. (Ser. 2) **65**, 139–166 (1967) [Trudy Mat. Inst. Steklova **73**, 314–336 (1964) (Russian)].
[20] P. Deift and E. Trubowitz, ''Inverse scattering on the line,'' Commun. Pure Appl. Math. **32**, 121–251 (1979).
[21] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. (Springer, New York, 1989).
[22] T. Aktosun, M. Klaus, and C. van der Mee, ''On the number of bound states for the 1-D Schrödinger equation,'' J. Math. Phys. **39**, 4249–4256 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

On the Batalin, Fradkin, Fradkina, and Tyutin quantization of first order systems

Ricardo Amorim
Instituto de Física, Universidade Federal do Rio de Janeiro, RJ 21945-970—Caixa Postal 68528—Brazil

Ronaldo Thibes
Instituto de Física, Universidade Federal do Rio de Janeiro, RJ 21945-970—Caixa Postal 68528—Brazil and Instituto de Ciências Exatas e da Natureza, Universidade do Grande Rio Professor José de Souza Herdy, Rua Prof. José de Souza Herdy, 1160, 25 de Agosto—Duque de Caxias—RJ 25073-200—Brazil

(Received 24 May 1999; accepted for publication 9 August 1999)

By using the field–antifield formalism, we show that the method of Batalin, Fradkin, Fradkina, and Tyutin (BFFT) to convert Hamiltonian systems submitted to second class constraints introduces compensating fields which do not belong to the BRST cohomology at ghost number one. This assures that the gauge symmetries which arise from the BFFT procedure are not obstructed at quantum level. An example where massive electrodynamics is coupled to chiral fermions is considered. We solve the quantum master equation for the model and show that the respective counterterm has a decisive role in extracting anomalous expectation values associated with the divergence of the Noether chiral current. © *1999 American Institute of Physics.* [S0022-2488(99)02511-6]

I. INTRODUCTION

The seminal works of Dirac[1] on constraint Hamiltonian systems have been developed in several important research lines. One of these developments is due to Batalin, Fradkin, Fradkina, and Tyutin (BFFT),[2] where Hamiltonian systems submitted to second class constraints are conveniently considered. The method of BFFT consists in enlarging the original phase-space of the theory by adding compensating fields which permit to convert the second class constraints into first class ones. In doing so, it is possible to avoid Dirac brackets which can present severe problems when one follows the canonical approach to quantization.[3] As first class constraints are also necessarily associated with local gauge symmetries, a system converted by the BFFT procedure can be treated by using all the machinery associated with the Becchi–Rouet–Stora–Tyutin (BRST) formalism.[4]

The BRST approach for quantization of gauge theories appears with all its power in the field–antifield formalism.[5,3,6] This formalism gives an elegant and systematic way for constructing the functional generator of any general gauge theory, with possible reducible or open gauge algebras. At the same time, eventual obstructions to the gauge symmetries due to quantum effects are naturally taken in account inside the field–antifield formalism.

In this work we consider, by using some tools of the field–antifield formalism, the quantization of first order gauge theories which have been obtained from second class constrained systems by the process of conversion developed by BFFT. We show that the compensating fields introduced by the conversion procedure do not belong to the BRST cohomology[3] at ghost number one. So there is no possible term in the space of fields and antifields with ghost number one and BRST closed not being BRST exact. This means that the Wess–Zumino consistence condition[6] is solved in a trivial way; there is no gauge anomaly for such class of systems and the quantum master equation can always be solved with the inclusion of a proper counterterm in the quantum action. It is useful to observe that this counterterm, if it exists, can play a nontrivial role. We give an example where massive electrodynamics couples to chiral fermions. There we show that it is necessary to introduce a nontrivial counterterm in order to solve the quantum master equation. This counterterm permits us to extract an anomalous expectation value related to the divergence of the fermion Noether chiral current.

We would like to note that compensating fields have been largely employed directly inside Lagrangian descriptions.[7,8] There the purpose is not converting second class constraints, but to enlarge the symmetry content of a theory in such a way that the original description is recovered within some gauge choice. Under this last point of view, BFFT and Lagrangian compensating fields play similar roles. In several examples of Lagrangian descriptions it is proved that compensating fields also do not belong to the cohomology at ghost number one and can be used as well to extract anomalous expectation values of physically relevant quantities.[9]

We organized this work as follows: In Sec. II we present a brief review of the BFFT conversion of first order systems submitted to pure second class constraints. We display the local gauge invariance of the first order action which is introduced by the BFFT compensating fields. The functional quantization of such a system is described in Sec. III, by using the tools of the field–antifield formalism. We derive the BRST differential and explicitly show that the BFFT variables do not belong to the BRST cohomology at ghost number one. This assures that the quantum master equation can be solved for any system of this class. In Sec. IV the ideas presented in the first sections are applied to a model which describes massive electrodynamics coupled to chiral fermions in four space–time dimensions. By using a regularization that keeps the vector symmetry as a preferential one, the quantum master equation is solved with the introduction of an specific counterterm in the quantum action. A few different gauge fixing choices are explored and covariant actions are obtained. When the gauge freedom is fixed by identifying the compensating fields with external functions, we show that the independence of the path integral with respect to those external functions permit us to derive expectation values which are related to the anomalous divergence of the Noether chiral current. We reserve Sec. V to some general comments and concluding remarks.

II. FIRST ORDER SYSTEMS SUBMITTED TO SECOND CLASS CONSTRAINTS

In this section we will review a few topics on constrained Hamiltonian systems[3] and on the BFFT conversion procedure[2] in order to fix notations and to introduce some results that will be useful for further developments. Let us start by considering a generic first order system living in a (phase) space with discrete bosonic coordinates y^μ, $\mu=1,2,...,2N$. The extension to more general situations can be trivially done. Its action is written as

$$S_0=\int dt(B_\mu\dot{y}^\mu-\lambda^\alpha\chi_\alpha-H), \tag{2.1}$$

where B_μ, H, and χ_α are in principle arbitrary functions of the coordinates but do not depend on the velocities. The Lagrange multipliers λ^α are to be regarded as independent quantities. From the above expression one can read the symplectic form

$$f_{\mu\nu}=\frac{\partial B_\nu}{\partial y^\mu}-\frac{\partial B_\mu}{\partial y^\nu} \tag{2.2}$$

which has an inverse $f^{\mu\nu}$ if the system is well defined. With its aid, we can define the brackets between any two functions $A(y)$ and $B(y)$ as

$$\{A,B\}=\frac{\partial A}{\partial y^\mu}\,f^{\mu\nu}\frac{\partial B}{\partial y^\nu}. \tag{2.3}$$

It follows that

$$\{y^\mu,y^\nu\}=f^{\mu\nu}. \tag{2.4}$$

The brackets appearing in the above expressions can be interpreted as Poisson brackets only in a broad sense, since they take in account the primary second class constraints of the Dirac's scheme.[10] In this sense they are primary Dirac brackets. Let now H and χ_α, $\alpha=1,2,...,2n$, represent respectively, a first class Hamiltonian and a set of second class constraints. The Hamiltonian and the constraints then satisfy the structure

$$\begin{aligned}\{\chi_\alpha,\chi_\beta\}&=\Delta_{\alpha\beta},\\ \{H,\chi_\alpha\}&=V_\alpha^\beta\chi_\beta.\end{aligned}\tag{2.5}$$

As the χ's are second class, the constraint matrix $\Delta_{\alpha\beta}$ is regular.

It may be convenient to extend the phase-space by adding compensating variables ϕ^α, $\alpha=1,2,...,2n$, but at the same time converting the set of second class constraints into a first-class one. This assures that the number of degrees of freedom is not changed by the process, which also introduces local symmetries that permit one to quantize the theory by using the powerful tools of local gauge theories.

To perform this conversion through the BFFT procedure, it is assumed that the BFFT compensating variables ϕ^α satisfy fundamental brackets given by

$$\{\phi^\alpha,\phi^\beta\}=\omega^{\alpha\beta},\tag{2.6}$$

where ω is some constant, antisymmetric, and invertible matrix. In order to avoid the introduction of further second class constraints, it may be convenient to choose ω in such a way that the compensating variables form a set of canonical conjugated quantities. In any case, it follows that in the BFFT extended space, the brackets between any two quantities $A(y,\phi)$ and $B(y,\phi)$ are written as

$$\{A,B\}=\frac{\partial A}{\partial y^\mu}\,f^{\mu\nu}\frac{\partial B}{\partial y^\nu}+\frac{\partial A}{\partial\phi^\alpha}\,\omega^{\alpha\beta}\frac{\partial B}{\partial\phi^\beta}\tag{2.7}$$

as both sectors are independent.

The general idea of the BFFT algorithm is to replace the old set of second class constraints and the old Hamiltonian by a new set of first class constraints $\tilde\chi_\alpha=\tilde\chi_\alpha(y,\phi)$ and Hamiltonian $\tilde H=\tilde H(y,\phi)$ in such a way that they become involutive,

$$\begin{aligned}\{\tilde\chi_\alpha,\tilde\chi_\beta\}&=0,\\ \{\tilde H,\tilde\chi_\alpha\}&=0.\end{aligned}\tag{2.8}$$

By requiring that $\tilde A(y,0)=A(y)$ for any quantity A defined in the extended space, it is assured that the original formulation of the theory is recovered when the unitary gauge $\phi^\alpha=0$ is implemented. In Refs. 2 it is proven that Eqs. (2.8), submitted to the above condition, always have a power series solution in the compensating variables, with coefficients with only y^μ dependence. The second class constraints, for instance, can be extended to

$$\tilde\chi_\alpha(y,\phi)=\chi_\alpha(y)+X_{\alpha\beta}(y)\phi^\beta+X_{\alpha\beta\gamma}(y)\phi^\beta\phi^\gamma+\cdots.\tag{2.9}$$

Conditions (2.8) impose restrictions on the expansion coefficients. As an example, the regular matrices $X_{\alpha\beta}$ must satisfy the identity

$$X_{\alpha\beta}\omega^{\beta\gamma}X_{\delta\gamma}=-\Delta_{\alpha\delta}.\tag{2.10}$$

Even if some quantity $A(y)$ is not a second class constraint, it can also be extended to $\tilde A(y,\phi)$ in order to be involutive with the converted constraints $\tilde\chi_\alpha$. Following the BFFT procedure we can show that in this situation

$$\tilde{A}(y,\phi)=A(y)-\phi^{\alpha}\omega_{\alpha\beta}X^{\beta\gamma}\{\chi_{\gamma},A\}+\cdots, \tag{2.11}$$

where the dots represent at least second order corrections in ϕ to $A(y)$. In (2.11), the matrix X with contravariant indices is to be considered as the inverse of the corresponding covariant one. Now it is possible to prove that the first order action

$$S_0=\int dt[B_{\mu}\dot{y}^{\mu}+B_{\alpha}\dot{\phi}^{\alpha}-\lambda^{\alpha}\tilde{\chi}_{\alpha}-\tilde{H}] \tag{2.12}$$

is invariant under the gauge transformations

$$\begin{aligned}\delta y^{\mu}&=\{y^{\mu},\tilde{\chi}_{\alpha}\}\epsilon^{\alpha},\\ \delta\phi^{\alpha}&=\{\phi^{\alpha},\tilde{\chi}_{\beta}\}\epsilon^{\beta},\\ \delta\lambda^{\alpha}&=\dot{\epsilon}^{\alpha}.\end{aligned} \tag{2.13}$$

By using the Jacobi Identity and Eqs. (2.8) we see that (2.13) close in an Abelian algebra. As in (2.2), in (2.12) B_{α} is related to the inverse of $\omega^{\alpha\beta}$ through

$$\omega_{\alpha\beta}=\frac{\partial B_{\beta}}{\partial\phi^{\alpha}}-\frac{\partial B_{\alpha}}{\partial\phi^{\beta}}. \tag{2.14}$$

One can always choose $B_{\alpha}=\frac{1}{2}\omega_{\alpha\beta}\phi^{\beta}$ without loss of generality. By using some of the above equations, it is not difficult to show that

$$\delta[B_{\mu}\dot{y}^{\mu}+B_{\alpha}\dot{\phi}^{\alpha}-\lambda^{\alpha}\tilde{\chi}_{\alpha}-\tilde{H}]=\frac{d}{dt}\left\{\left[B_{\mu}f^{\mu\nu}\frac{\partial\tilde{\chi}_{\alpha}}{\partial y^{\nu}}+B_{\beta}\omega^{\beta\rho}\frac{\partial\tilde{\chi}_{\alpha}}{\partial\phi^{\rho}}-\tilde{\chi}_{\alpha}\right]\epsilon^{\alpha}\right\} \tag{2.15}$$

and consequently (2.12) is indeed invariant under the local gauge transformations (2.13), provided boundary terms can be discarded.

III. QUANTIZATION

Let us perform the quantization of the system described above along the field–antifield formalism.[5,3,6] To do so it is first necessary to introduce antifields $\Phi^*_A=(y^*_{\mu},\phi^*_{\alpha},\lambda^*_{\alpha},c^*_{\alpha})$ corresponding to the fields $\Phi^A=(y^{\mu},\phi^{\alpha},\lambda^{\alpha},c^{\alpha})$. In our case, y^{μ}, ϕ^{α}, and λ^{α} are bosonic and have ghost number zero. The ghosts c^{α} are fermionic and have ghost number one. The corresponding antifields have opposite grassmanian parity and ghost number given by minus the ghost number of the corresponding field minus one. One can verify that the field–antifield action

$$S=S_0+\int dt[y^*_{\mu}\{y^{\mu},\tilde{\chi}_{\alpha}\}c^{\alpha}+\phi^*_{\beta}\{\phi^{\beta},\tilde{\chi}_{\alpha}\}c^{\alpha}+\lambda^*_{\alpha}\dot{c}^{\alpha}] \tag{3.1}$$

satisfies then the classical master equation

$$\tfrac{1}{2}(S,S)=0, \tag{3.2}$$

where the antibracket between any two quantities $X[\Phi,\Phi^*]$ and $Y[\Phi,\Phi^*]$ is defined as

$$(X,Y)=\frac{\delta_r X}{\delta\Phi^A}\,\frac{\delta_l Y}{\delta\Phi^*_A}-\frac{\delta_r X}{\delta\Phi^*_A}\,\frac{\delta_l Y}{\delta\Phi^A}.$$

When pertinent, we are assuming the de Witt's notation of sum and integration over intermediary variables.

In the BV formalism, the BRST differential is introduced through

$$sX=(X,S) \tag{3.3}$$

for any local functional $X=X[\Phi,\Phi^*]$. As a consequence of the master equation (3.2) and Jacobi identity, s is nilpotent. So, saying that the BV action satisfies the master equation is equivalent to say that it is BRST invariant.

To fix a gauge we need to introduce trivial pairs $\overline{c}_\alpha$, $\overline{\pi}_\alpha$ as new fields. and the corresponding antifields $\overline{c}^{*\alpha}$, $\overline{\pi}^{*\alpha}$, as well as a gauge-fixing fermion Ψ. The antifields are eliminated by choosing $\Phi_A^*=\partial\Psi/\partial\Phi^A$. It is always possible to choose

$$\Psi=\overline{c}_\alpha\phi^\alpha \tag{3.4}$$

associated with the unitary gauge, but different choices can be done. It is also necessary to extend the field–antifield action to a nonminimal one,

$$S\rightarrow S_{nm}=S+\int dt\,\overline{\pi}_\alpha\overline{c}^{*\alpha} \tag{3.5}$$

in order to implement the gauge fixing introduced by Ψ. The gauge-fixed vacuum functional is then defined as

$$Z=\int [d\Phi^A][\det\omega]^{-(1/2)}[\det f]^{-(1/2)}\exp\left\{\frac{i}{\hbar}S_{nm}\left[\Phi^A,\Phi_A^*=\frac{\partial\Psi}{\partial\Phi^A}\right]\right\}. \tag{3.6}$$

In the unitary gauge, we observe that besides the identification $\overline{c}^{*\alpha}=\phi^\alpha$, $\phi_\alpha^*=\overline{c}_\alpha$, all the other antifields vanish. With this and the use of Eqs. (2.9)–(2.10), we see that formally (3.6) reduces to the Senjanovic[11] path integral

$$Z=\int [dy^\mu]|\det f|^{-(1/2)}\delta[\chi_\alpha]|\det\Delta|^{(1/2)}\exp\left\{\frac{i}{\hbar}\int dt[B_\mu\dot{y}^\mu-H]\right\}. \tag{3.7}$$

Actually this reduction can only be done if quantum effects do not obstruct the gauge symmetries. Possible obstructions are related to the dependence of the path integral with respect to redefinitions of the gauge-fixing fermion Ψ. In general, if the classical field–antifield action S can be replaced by some quantum action W expressed as a local functional of fields and antifields and satisfying the so-called quantum master equation

$$\tfrac{1}{2}(W,W)-i\hbar\Delta W=0, \tag{3.8}$$

then the gauge symmetries are not obstructed at quantum level. In expression (3.8) we have introduced the potentially singular operator $\Delta\equiv(\delta_r/\delta\Phi^A)(\delta_l/\delta\Phi_A^*)$ and it was assumed that W can be expanded in powers of $\hbar$ as

$$W[\Phi^A,\Phi_A^*]=S[\Phi^A,\Phi_A^*]+\sum_{p=1}^{\infty}\hbar^p M_p[\Phi^A,\Phi_A^*]. \tag{3.9}$$

The two first terms of the quantum master equation (3.8) are

$$(S,S)=0, \tag{3.10}$$

$$(M_1,S)=i\Delta S. \tag{3.11}$$

As expected, the tree approximation gives (3.2). Equation (3.11) is only formal, since the action of the operator Δ must be regularized. If it vanishes when applied on S, the quantum action

W can be identified with S. If ΔS gives a nontrivial result but there exists some M_1 expressed in terms of local fields such that (3.11) is satisfied, gauge symmetries are not obstructed at one loop order. Otherwise, the theory presents an anomaly

$$\mathcal{A}[\phi,\phi^*]=\Delta S+i(S,M_1)=a_\alpha c^\alpha+\cdots . \tag{3.12}$$

The nilpotency of the BRST operator implies that $s\mathcal{A}=0$, which is the Wess–Zumino consistence condition. So, looking for possible anomalies in any theory is the same as looking for local functionals with ghost number one that are BRST closed $(s\mathcal{A}=0)$ but not BRST exact $(\mathcal{A}\neq sB)$.

By using cohomological arguments, we can show that the quantum master equation, for first order systems with pure second class constraints converted with the use of the BFFT procedure, can always be solved. To prove this, let us first derive the BRST transformations of the fields and antifields for the converted system,

$$\begin{aligned}
sy^\mu&=\{y^\mu,\tilde{\chi}_\alpha\}c^\alpha,\\
s\phi^\beta&=\{\phi^\beta,\tilde{\chi}_\alpha\}c^\alpha,\\
s\lambda^\alpha&=\dot{c}^\alpha,\\
sc^\alpha&=0,\\
s\bar{c}_\alpha&=\bar{\pi}_\alpha,\\
s\bar{\pi}_\alpha&=0,\\
sy^*_\mu&=-\frac{\partial S}{\partial y^\mu},\\
s\phi^*_\alpha&=-\frac{\partial S}{\partial \phi^\alpha},\\
s\lambda^*_\alpha&=\tilde{\chi}_\alpha,\\
sc^*_\alpha&=-y^*_\mu\{y^\mu,\tilde{\chi}_\alpha\}-\phi^*_\beta\{\phi^\beta,\tilde{\chi}_\alpha\}-\dot{\lambda}^*,\\
s\bar{c}^*_\alpha&=0,\\
s\bar{\pi}^{*\alpha}&=\bar{c}^{*\alpha},
\end{aligned} \tag{3.13}$$

where S is given by (3.1). We see that $\bar{c}_\alpha$ and $\bar{\pi}_\alpha$ form BRST doublets $(sB=C,\ sC=0)$ and do not belong to the BRST cohomology.[3] The same is true for their antifields. To show that the other fields and antifields do not contribute to the cohomology at ghost number one, it is enough to study the cohomology of the linearized piece of s, which will be denoted by $s^{(1)}$.[12] If we assume that in the process of conversion of the constraints (see Eq. (2.9)), the invertible matrix $X(y)$ can be written as a power series in y (which will be the case for the example we are going to consider),

$$X(y)_{\alpha\beta}=X^{(0)}_{\alpha\beta}+X^{(1)}_{\alpha\beta\mu}y^\mu+X^{(2)}_{\alpha\beta\mu\nu}y^\mu y^\nu+\cdots \tag{3.14}$$

we see that

$$s^{(1)}\phi^{\alpha}=\omega^{\alpha\gamma}X^{(0)}_{\beta\gamma}c^{\beta},\qquad s^{(1)}c^{\alpha}=0. \tag{3.15}$$

The equations above imply that ϕ^{α} and $C^{\alpha}=\omega^{\alpha\gamma}X^{(0)}_{\beta\gamma}c^{\beta}$ form doublets under the action of $s^{(1)}$ and as a consequence they also do not belong to the cohomology. As c^{α} is trivially obtained from C^{α}, and since it is the only fundamental field (or antifield) with positive ghost number, it is not possible to construct a local functional with ghost number one that is BRST closed not being BRST exact. This means that any candidate to an anomaly can always be canceled by some counterterm M. So the situations found in Ref. 8 and later explored in Ref. 9 appear also here; enlarged symmetries due to compensating fields (here the BFFT variables) are not anomalous. This does not mean that they have a trivial role at the quantum level since the existence of a counterterm modify expectation values of relevant physical quantities.[9] In the next section we are going to show an example where all of these features are carefully taken in account in order to derive consistent quantum actions.

IV. MASSIVE VECTOR FIELDS COUPLED TO CHIRAL FERMIONS

We shall now apply the ideas discussed above to massive chiral electrodynamics. Although the fermions couple only one chirality to the connection A_{μ}, the second class system presents no gauge anomaly since it exhibits no gauge symmetry. When it is converted to a first class one, however, the fermions pass to transform in a chiral way and such a gauge transformation is known to lead to possible anomalies.[13] Accordingly to the ideas discussed in the last section, however, the BFFT variables play the role of Wess–Zumino fields and permit us to write the anomaly candidates as BRST exact functionals, solving in this way the quantum master equation at one loop order.

We start by considering the first order action

$$S_0=\int d^4x\{\dot{A}_{\mu}\pi^{\mu}+i\bar{\psi}\gamma^0\dot{\psi}-\mathcal{H}-\lambda^{\alpha}\chi_{\alpha}\}, \tag{4.1}$$

where the second class constraints

$$\chi_1=\pi^0,\qquad \chi_2=\partial_i\pi^i-m^2A^0+J^0, \tag{4.2}$$

and the first class Hamiltonian

$$H=\int d^3x\left\{\frac{1}{2}\,\pi_i^2+\frac{1}{4}F_{ij}^2+\frac{1}{2}m^2(A_0^2+A_i^2)-i\bar{\psi}\gamma^iD_i^+\psi+\partial_iA^i\chi_1-A_0\chi_2\right\} \tag{4.3}$$

have been introduced. In the above expressions we have defined the covariant derivatives D_{μ}^+ acting on the fermion ψ and the chiral projectors $P^{\pm}$, respectively, as

$$D_{\mu}^+=\partial_{\mu}-ieP^+A_{\mu},\qquad P^{\pm}=\tfrac{1}{2}(1\pm\gamma^5). \tag{4.4}$$

We have also adopted the metric convention $\eta^{\mu\nu}=\mathrm{diag}(-1,+1,+1,+1)$. Dirac matrices satisfy the usual anticommutation relation $\{\gamma^{\mu},\gamma^{\nu}\}=2\,\eta^{\mu\nu}$. As one can verify, action (4.1) is the first order version of

$$\mathcal{S}_{\mathrm{cov}}=\int d^4x\left[-\frac{1}{4}\,F_{\mu\nu}F^{\mu\nu}-\frac{1}{2}\,m^2A_{\mu}A^{\mu}+i\bar{\psi}\gamma^{\mu}D_{\mu}^+\psi\right]. \tag{4.5}$$

From (4.1) we extract the fundamental (equal time) brackets

$$\{\psi(x),\overline{\psi}(y)\}=i\gamma^0\delta^3(x-y) \tag{4.6}$$

for the fermionic sector and

$$\{A_\mu(x),\pi^\nu(y)\}=\delta_\mu^\nu\delta^3(x-y) \tag{4.7}$$

for the bosonic one. By using the above expressions, one can show, for instance, that the fermionic chiral current

$$J^\mu\equiv\overline{\psi}\gamma^\mu P^+\psi \tag{4.8}$$

has brackets between its components given by

$$\begin{aligned}\{J^\mu(x),J^\nu(y)\}&=ie^2\overline{\psi}M^{\mu\nu}P^+\psi\delta^3(x-y),\\ M^{\mu\nu}&=\gamma^\mu\gamma^0\gamma^\nu-\gamma^\nu\gamma^0\gamma^\mu.\end{aligned} \tag{4.9}$$

It is now easy to verify that the constraints and the Hamiltonian satisfy the bracket structure

$$\begin{aligned}\{\chi_1(x),\chi_2(y)\}&=-m^2\delta^3(x-y),\\ \{\chi_1(x),H\}&=\chi_2(x),\\ \{\chi_2(x),H\}&=\partial_i\partial^i\chi_1(x).\end{aligned} \tag{4.10}$$

Let us now use the BFFT algorithm for implementing the Abelian conversion of the above bracket structure. As we have two second class constraints, we introduce two BFFT variables ϕ^α, $\alpha=1,2$ and for simplicity demand that they satisfy

$$\{\phi^\alpha(x),\phi^\beta(y)\}=\epsilon^{\alpha\beta}\delta^3(x-y) \tag{4.11}$$

which gives the matrix $\omega^{\alpha\beta}$ as in Eq. (2.6). In (4.11) $\epsilon^{12}=-\epsilon^{21}=1$, $\epsilon^{11}=\epsilon^{22}=0$. A possible solution to Eqs. (2.8) via (2.9)–(2.11) is achieved with[2]

$$\begin{aligned}\tilde{\chi}_1&=\chi_1-m^2\phi^2,\\ \tilde{\chi}_2&=\chi_2+\phi^1,\\ \tilde{H}&=H+\int d^3x\left[\frac{1}{2m^2}(\phi^1)^2+\frac{1}{2}m^2(\partial_i\phi^2)^2-\frac{\phi^1}{m^2}\tilde{\chi}_2-\phi^2\nabla^2\tilde{\chi}_1\right]\\ &=\int d^3x\left[\frac{1}{2}\pi_i^2+\frac{1}{4}F_{ij}^2+\frac{1}{2}m^2(\tilde{A}_0^2+\tilde{A}_i^2)-i\overline{\psi}\gamma^iD_i^+\psi-\tilde{A}_0\tilde{\chi}_2+(\partial_i\tilde{A}^i)\tilde{\chi}_1\right],\end{aligned} \tag{4.12}$$

where we have defined the quantities

$$\begin{aligned}\tilde{A}_i&=A_i-\partial_i\phi^2,\\ \tilde{A}_0&=A_0+\frac{\phi^1}{m^2}.\end{aligned} \tag{4.13}$$

Correspondingly we have a first order action

$$S_0=\int d^4x\{\dot{A}_\mu\pi^\mu+\dot{\phi}^1\phi^2+i\bar{\psi}\gamma^0\dot{\psi}-\tilde{\mathcal{H}}-\lambda^\alpha\tilde{\chi}_\alpha\}, \tag{4.14}$$

which is invariant under the gauge transformations generated by $\tilde{\chi}_1$ and $\tilde{\chi}_2$ (see Eq. (2.13))

$$\begin{gathered}\delta\psi=-ie\epsilon^2P^+\psi,\quad \delta\bar{\psi}=ie\epsilon^2\bar{\psi}P^-,\\ \delta A_0=\epsilon^1,\quad \delta\pi^0=-m^2\epsilon^2,\\ \delta A_i=-\partial_i\epsilon^2,\quad \delta\pi^i=0,\\ \delta\phi^1=-m^2\epsilon^1,\quad \delta\phi^2=-\epsilon^2,\\ \delta\lambda^1=\dot{\epsilon}^1,\quad \delta\lambda^2=\dot{\epsilon}^2.\end{gathered} \tag{4.15}$$

In the expressions above ϵ^α are arbitrary space–time dependent parameters. We note that the variables $\tilde{A}_\mu$ are invariant under (4.15).

In order to quantize this system along the lines of the field–antifield formalism, associated with the parameters ϵ^α we introduce the ghosts c^α. We introduce also the trivial pairs $\bar{\pi}_\alpha$, $\bar{c}_\alpha$, and write down a gauge-fixed vacuum functional as in (3.6) with

$$\begin{aligned}S_{nm}=S_0+\int d^4x[&A^{0*}c^1-m^2\pi_0^*c^2-A^{i*}\partial_ic^2-m^2\phi_1^*c^1\\ &-\phi_2^*c^2+\lambda_1^*\dot{c}^1+\lambda_2^*\dot{c}^2-ie\psi^*P^+\psi c^2+ie\bar{\psi}P^-\bar{\psi}^*c^2+\bar{\pi}_\alpha\bar{c}^{\alpha*}],\end{aligned} \tag{4.16}$$

where some proper gauge-fixing fermion Ψ is assumed. Now observe that the terms in S_{nm} which involve the matter fields are

$$i\bar{\psi}[\gamma^0(\partial_0-ieP^+(\tilde{A}_0-\lambda^2))+\gamma^iD_i^+]\psi. \tag{4.17}$$

The quantities $\bar{A}_0=\tilde{A}_0-\lambda^2$ and $\bar{A}_i=A_i$ transform as $s\bar{A}_\mu=-\partial_\mu c^2$. As the fermions also transform consistently, as can be seen from (4.15), we obtain the action of the operator Δ over S_Ψ adopting canonical procedures. For instance, in a Pauli–Villars regularization scheme with a fermionic mass term with usual form, which means that the vector symmetry is taken as a preferential one, we see that

$$\Delta S_\Psi=-\frac{1}{96\pi}\int d^4x c^2\epsilon^{\mu\nu\rho\sigma}\bar{F}_{\mu\nu}\bar{F}_{\rho\sigma}, \tag{4.18}$$

where $\bar{F}_{\mu\nu}=\partial_\mu\bar{A}_\nu-\partial_\nu\bar{A}_\mu$ and possible normal parity terms in the original space of fields have been discarded. Equation (4.18) represents the essential candidate to the anomaly. It is easy to see, however, that

$$M_1=\frac{i}{96\pi}\int d^4x\phi^2\epsilon^{\mu\nu\rho\sigma}\bar{F}_{\mu\nu}\bar{F}_{\rho\sigma} \tag{4.19}$$

solves the one loop master equation, which means that we have achieved a consistent route for the quantization of the theory. The gauge fixed vacuum functional reads

$$Z=\int[d\Phi^A]\exp\left\{\frac{i}{\hbar}W\left[\Phi^A,\Phi_A^*=\frac{\partial\Psi}{\partial\Phi^A}\right]\right\} \tag{4.20}$$

with $[d\Phi^A]=(A_\mu,\pi^\mu,\phi^\alpha,\psi,\bar{\psi},\lambda^\alpha,c^\alpha,\bar{c}_\alpha,\bar{\pi}_\alpha)$, and all possible information about the system can be obtained from it. If we wish to write an effective quantum action in an explicitly covariant way we may eliminate the momenta through functional integrations in (4.20). Let us assume that the gauge fixing fermion Ψ does not depend on λ^1 or π^μ, consequently $\lambda_l^* = \pi_\mu^* = 0$. Suppose also that Ψ possibly depends on λ^2 only through an A_0 dependence. Integration in λ^1 and π^0 results in the substitution $\pi^0 \to m^2\phi^2$ in W. Under the redefinition

$$A_0 \to A_0 + \lambda^2 - \frac{\phi^1}{m^2} \tag{4.21}$$

we obtain the intermediate auxiliary quantum action

$$W_{\text{aux}} = \int d^4x \Big[(A_0 + \lambda^2)\dot{\phi} + \dot{A}_i \pi^i + i\bar{\psi}\gamma^0\dot{\psi} - \frac{1}{4}F_{ij}^2 - \frac{1}{2}\pi^{i^2} - \frac{1}{2}m^2(A_0 + \lambda_2)^2 - \frac{1}{2}m^2(A_i - \partial_i\phi)^2 + i\bar{\psi}\gamma^i D_i\psi + A_0(\partial_i\pi^i + J^0 + m^2(A_0 + \lambda^2)) \Big] + \hbar M_1 + S_{gf}, \tag{4.22}$$

where

$$S_{gf} = \int d^4x \Big[-\frac{\delta\Psi}{\delta A_\mu}\partial_\mu c^2 - m^2\frac{\delta\Psi}{\delta\phi^1}c^1 - \frac{\delta\Psi}{\delta\phi^2}c^2 - ie\frac{\delta\Psi}{\delta\psi}P^+\psi c^2 + ie\bar{\psi}P^-\frac{\delta\Psi}{\delta\psi}c^2 + \bar{\pi}_\alpha\frac{\delta\Psi}{\bar{c}_\alpha} \Big] \tag{4.23}$$

and M_1 is given by (4.19) without the bars in $F_{\mu\nu}$ because of (4.21). Further integration in λ_2 and π^i results in the effective quantum action

$$W_{\text{eff}} = \int d^4x \Big[-\frac{1}{4}F_{\mu\nu}^2 - \frac{1}{2}m^2(A_\mu - \partial_\mu\phi^2)^2 + i\bar{\psi}\gamma^\mu D_\mu^+\psi \Big] + \hbar M_1 + S_{gf}. \tag{4.24}$$

As we have already mentioned, a convenient choice of Ψ fixes all the gauge symmetry of the theory. We cite some possible choices for Ψ. The unitary gauge is achieved with $\Psi = \int d^4x \bar{c}_\alpha \phi^\alpha$ followed by functional integration on $\bar{\pi}_\alpha$ and ϕ^α. With this choice the quantum action reduces to the simple form (4.5) and the path integral presents the usual Liouville's measure for the pertinent fields. The choice

$$\Psi = \int d^4x \Big[\bar{c}_2\Big(\frac{\alpha\bar{\pi}^2}{2} + \partial_\mu A^\mu\Big) + \bar{c}_1\phi^1 \Big]$$

leads to the usual covariant Gaussian gauge fixing depending on the arbitrary parameter α. In this situation

$$S_{gf} = \int d^4x \Big[-\partial^\mu\bar{c}_2\partial_\mu c^2 + \bar{\pi}_2\Big(\frac{\alpha\bar{\pi}^2}{2} + \partial_\mu A^\mu\Big) + \bar{\pi}^1\phi^1 - m^2\bar{c}_1 c^1 \Big] \tag{4.25}$$

and the integration over c^1, $\bar{c}_1$, $\bar{\pi}_1$, ϕ^1 is trivial.

An interesting situation comes if we fix the compensating field ϕ^2 to some external value, say, $\phi^2 = \beta$. By choosing $\Psi = \bar{c}_1\phi^1 + \bar{c}_2(\phi^2 - \beta)$, we obtain, after a few trivial integrations and the absorption of some trivial normalization factors by the measure, that

$$Z[\beta] = \int [d\psi][d\bar{\psi}][dA^\mu]\exp\Big\{\frac{i}{\hbar}W_{\text{ext}}[\psi,\bar{\psi},A,\beta]\Big\}, \tag{4.26}$$

where

$$W_{\text{ext}}[\psi,\bar{\psi},A,\beta]=\int d^4x\left[-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}-\frac{1}{2}m^2(A_\mu-\partial_\mu\beta)^2+i\bar{\psi}\gamma^\mu D_\mu^+\psi+\frac{i\hbar}{96\pi}\beta\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\right]. \tag{4.27}$$

The condition that the path integral cannot depend on β, which comes from the Fradkin–Vilkoviski theorem, gives, for instance, that

$$i\hbar\frac{\delta Z[\beta]}{\delta\beta}\bigg|_{\beta=0}=\left\langle m^2\partial_\mu A^\mu+\frac{i\hbar}{96\pi}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\right\rangle\bigg|_{\beta=0}=0 \tag{4.28}$$

which is a surprising result. If we observe, however, that $\partial_\mu J^\mu=-m^2\partial_\mu A^\mu$ as a consequence of the equations of motion for the field A_μ in the unitary gauge, we can interpret Eq. (4.28) as the anomalous divergence of the Noether current (4.8) associated with the rigid chiral symmetry present in the original theory given by actions (4.1)–(4.5). This is an unexpected result derived from the quantum BFFT formalism. Similar results have recently been derived by using compensating fields at Lagrangian level.[9] In these last approaches, the compensating fields coupled directly to the chiral current in an extended QCD which presents not only vector but also chiral gauge symmetry.

V. CONCLUSIONS

In this work we have considered the BFFT quantization of first order systems submitted to pure second class constraints. We have shown that the gauge symmetries introduced by the BFFT procedure are not obstructed at the quantum level, since the compensating fields do not belong to the BRST cohomology at ghost number one. A specific example has been given, where massive electrodynamics couples to chiral fermions. The quantum master equation has been solved and the corresponding counterterm has played an essential role in extracting anomalous expectation values of physically relevant quantities. We would like to finish by commenting that a few generalizations could have been considered. We could have started from an already gauge invariant first order system with both first and second class constraints. Then it would be necessary to take care of both symmetry sectors, the original one and that introduced by the BFFT conversion procedure. Another possibility could be considering examples with more involving algebraic structure, as it occurs with some of the models cited in Ref. 2. We are now studying aspects of these subjects and results will be reported elsewhere.[14]

ACKNOWLEDGMENTS

We are indebted to N. R. F. Braga for a useful discussion. This work is supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq (Brazilian Research Agency).

[1] P. A. M. Dirac, Can. J. Math. **2**, 129 (1950); *Lectures on Quantum Mechanics* (Yeshiva University, New York, 1964).
[2] I. A. Batalin and E. S. Fradkin, Nucl. Phys. B **279**, 514 (1987); I. A. Batalin and L. V. Tyutin, Int. J. Mod. Phys. A **6**, 3255 (1991); I. A. Batalin, E. S. Fradkin, and T. E. Fradkina, Nucl. Phys. B **314**, 158 (1989), and references cited therein. See also some applications of the BFFT procedure in T. Fujiwara, Y. Igarashi, and J. Kubo, Nucl. Phys. B **341**, 695 (1990); Y. Kim, S. Kim, W. Kim, Y. Park, K. Kim, and Y. Kim, Phys. Rev. D **46**, 4574 (1992); R. Banerjee, *ibid.* **48**, R5467 (1993); N. Banerjee, S. Ghosh, and R. Banerjee, Nucl. Phys. B **417**, 257 (1994); R. Banerjee, H. J. Rothe, and K. D. Rothe, Phys. Rev. D **49**, 5438 (1994); N. Banerjee, R. Banerjee, and S. Ghosh, Ann. Phys. (N.Y.) **241**, 237 (1995); R. Amorim and J. Barcelos-Neto, Phys. Lett. B **333**, 413 (1994); Phys. Rev. D **53**, 7129 (1996); R. Amorim and A. Das, Mod. Phys. Lett. A **9**, 3543 (1994); R. Amorim, Z. Phys. C **67**, 695 (1995); J. Barcelos-Neto, Phys. Rev. D **55**, 2265 (1997); R. Banerjee and J. Barcelos-Neto, Nucl. Phys. B **499**, 453 (1997).
[3] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, 1992).
[4] C. Becchi, A. Rouet, and R. Stora, Commun. Math. Phys. **42**, 127 (1975); I. V. Tyutin, Preprint Lebedev Institute No. 39 (1975); C. Becchi, A. Rouet, and S. Stora, Ann. Phys. (N.Y.) **98**, 287 (1976).
[5] I. A. Batalin and G. A. Vilkovisky, Phys. Lett. B **120**, 27 (1981); I. A. Batalin and G. A. Vilkovisky, Phys. Rev. D **28**, 2567 (1983); **30**, 508(E) (1984); J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena*, 3rd ed. (Oxford Science, Oxford, 1996).

[6] J. Gomis, J. París, and S. Samuel, Phys. Rep. **259**, 1 (1995).
[7] E. C. G. Stuckelberg, Helv. Phys. Acta **30**, 209 (1957); J. Alfaro and P. H. Damgaard, Nucl. Phys. B **404**, 751 (1993).
[8] B. de Wit and M. T. Grisaru, ''*Compensating Fields and Anomalies*,'' in *Quantum Field Theory and Quantum Statistics*, edited by I. A. Batalin, C. J. Isham, and G. A. Vilkovisky (Adam Hilger, New York, 1987), Vol. 2.
[9] R. Amorim and N. R. F. Braga, Phys. Rev. D **57**, 1225 (1998); R. Amorim, N. R. F. Braga, and M. Henneaux, Phys. Lett. B **436**, 125 (1998); R. Amorim and N. F. R. Braga, ''Compensating fields, bosonization, and soldering in QCD_2,'' hep-th/9812058.
[10] L. Faddeev and R. Jackiw, Phys. Rev. Lett. **60**, 1692 (1988).
[11] P. Senjanovic, Ann. Phys. (N.Y.) **100**, 227 (1976).
[12] F. Brandt, N. Dragon, and M. Kreuzer, Phys. Lett. B **231**, 263 (1989); Nucl. Phys. B **332**, 224 (1990); O. Piguet and S. P. Sorella, ''*Algebraic Renormalization*,'' Lecture Notes in Physics (Springer-Verlag, Berlin, 1995), Vol. m28.
[13] S. Adler, Phys. Rev. **177**, 2426 (1969); J. Bell and R. Jackiw, Nuovo Cimento A **60**, 47 (1969).
[14] R. Amorim and R. Thibes (work in progress).

A path space formula for Gauss vectors in Chern–Simons quantum electrodynamics

John L. Challifour[a] and John P. Clancy
Department of Physics, Indiana University, Bloomington, Indiana 47405

(Received 1 June 1999; accepted for publication 2 July 1999)

Canonical quantization of a Chern–Simons gauge field minimally coupled to a spinor field is studied as an indefinite metric quantum field theory in the usual covariant gauges, by using a lattice cutoff. For this model, we show that positivity for the indefinite metric, Gauss' Law, gauge invariance and Osterwalder–Schrader positivity for a self-adjoint Hamiltonian are equivalent. In addition, the path-space formula for the Osterwalder–Schrader semigroup is constructed in terms of a Euclidean scalar, massive, Gaussian random field. © *1999 American Institute of Physics.* [S0022-2488(99)01011-7]

I. INTRODUCTION

Models containing Chern–Simons gauge fields have been of continuing interest in both topological and quantum field theory over the last 15 years. In particular, Chern–Simons gauge fields coupled to a Higgs field have been proposed as a microscopic framework for the quantum Hall effect.[1] Also, a lattice model of a Maxwell–Chern–Simons gauge field coupled to a Higgs field, with a modification of the Chern–Simons lattice action to provide reflection positivity, has been shown by Fröhlich and Marchetti[2] to lead to anyons[3] as the vortices for the soliton superselection sectors. This work uses a general theory of braid statistics in three-dimensional local quantum theory described in Ref. 4.

Within the framework of canonical quantization, several studies of models containing Chern–Simons terms in their action have appeared (see Ref. 5, and references therein.) In all of these papers, characterization of the physical states has been provided by means of Gauss' Law. However, there has been no attempt to connect this work with the framework of constructive quantum field theory. In this context, the primary notion must be that of Osterwalder–Schrader positivity for the Euclidean field theory.

In this work, we make explicit the indefinite metric formulation, which is implicit in the work of Haller and Lim-Lombridas,[5] for a Maxwell–Chern–Simons gauge field minimally coupled to a fermion field, hence Chern–Simons quantum electrodynamics (QED). Our cutoffs allow characterization of the physical subspace and representation of domains for the unbounded operators needed to characterize gauge invariance, Gauss' Law, and the physical Hamiltonian in terms of a Euclidean path-space formula on physical states. The Euclidean representation of the physical states provides a statement of Osterwalder–Schrader positivity for this model. Future tasks would be to implement the techniques of constructive quantum field theory so as to provide a continuum theory of a local, covariant gauge field minimally coupled to fermions in a reconstructed Hilbert space, where an indefinite metric characterizes the physical states.

The Maxwell–Chern–Simons gauge field in our model is described by the following Lagrangian:

$$\mathcal{L}_G = -\frac{1}{2}\,\partial_\mu A_\nu \partial^\mu A^\nu + \frac{\lambda}{2}\,(\partial A)^2 + \frac{m}{2}\,\epsilon^{\mu\nu\sigma}\partial_\mu A_\nu A_\sigma, \tag{1}$$

[a]Also at Department of Mathematics.

where the first term is the kinetic energy term, the second term fixes the gauge ($\lambda=0$ corresponds to the Feynman gauge and $\lambda=\infty$ corresponds to the Landau gauge), and the third part is the Chern–Simons term. Note that for these units, the parameter m must have dimensions of mass. The Euler–Lagrange equations then become

$$\Box A^{\mu}-\lambda\,\partial^{\mu}(\partial A)+m\,\epsilon^{\mu\alpha\beta}\partial_{\alpha}A_{\beta}=0 \tag{2}$$

for the field equations. Now, in terms of the electric and magnetic fields we have

$$\text{div }\mathbf{E}-mB=(\lambda-1)\,\partial_t(\partial A), \tag{3}$$

$$\text{curl }B-\partial_t\mathbf{E}+m\,\epsilon^{ij}\hat{e}_iE_j=-(\lambda-1)\,\text{grad}(\partial A), \tag{4}$$

with Faraday's Law being equivalent to the Bianchi identity. Note that in $2+1$ dimensions, the electric field lies in the plane defined by the spatial coordinates, while the magnetic field lies perpendicular to this plane. To be consistent with Ref. 5, the canonical momenta are transformed as

$$\Pi_{\mu}\rightarrow e^{-iF}\Pi_{\mu}e^{iF}, \tag{5}$$

where $F=\Sigma_{\mathbf{x}\in V}\delta^2A_0(\mathbf{x})\,\partial_lA_l(\mathbf{x})$. The effect of this transformation is a change in representation of the canonical commutation relations with

$$\Pi_0=(\lambda-1)(\partial A), \tag{6}$$

$$\Pi_l=E_l-\frac{m}{2}\,\epsilon_{ln}A_n \tag{7}$$

so that $\text{div }\mathbf{E}-mB=(m/2)\,\epsilon_{ln}\partial_lA_n-\partial_l\Pi_l$.

Deser, Jackiw, and Templeton[6] have described the electric and magnetic fields in terms of a massive scalar field ϕ having spin 1 excitations. This scalar field will become very useful in our construction of the physical subspace, so we make contact with it now. The scalar field ϕ is defined as

$$\partial_lE_l=-m\sqrt{-\Delta}\,\phi \tag{8}$$

from which E_l and B are expressed as

$$E_l=\left[\frac{m\,\partial_l}{\sqrt{-\Delta}}-\epsilon_{ln}\,\frac{\partial_n\partial_0}{\sqrt{-\Delta}}\right]\phi-\epsilon_{ln}\,\frac{\partial_n}{m}\,G, \tag{9}$$

$$B=\sqrt{-\Delta}\,\phi+\frac{\partial_0}{m}\,G, \tag{10}$$

where $G\equiv(1-\lambda)(\partial A)$ and the field equations give $\Box G=0$. From these equations, we can find particular solutions for the gauge field

$$A_0^p=-\frac{\sqrt{-\Delta}}{m}\,\phi+(1-2\xi)\,\partial_0\Box^{-1}G, \tag{11}$$

$$A_l^p=\frac{\partial_0\partial_l}{m\sqrt{-\Delta}}\,\phi-\frac{\epsilon_{ln}\partial_n}{\sqrt{-\Delta}}\,\phi+(1-2\xi)\,\partial_l\Box^{-1}G, \tag{12}$$

with $(1-2\xi)=(1-\lambda)^{-1}$, so that $\xi=0$ is the Feynman gauge and $\xi=1/2$ is the Landau gauge. As $\Box G=0$, we see that the gauge field is decomposed into a massive piece V_μ and a massless piece S_μ. As these are only particular solutions, the contribution of another massless scalar field, $R(x)$, which commutes with ϕ and satisfies the canonical commutation relations (CCR's) with G can be added to S_μ. Again, as only ϕ contributes to the physical fields, it alone should determine the physical sector of the theory.

In Sec. II, canonical quantization for Chern–Simons QED with spatial lattice cutoffs, is described in a Hilbert space with a Krein indefinite metric. The relation between the representation of the gauge field used here and that used by Haller and Lim-Lombridas[5] is given, as well as a proof that the free Hamiltonian is Krein essentially self-adjoint in all covariant gauges and essentially self-adjoint in Landau gauge. However, for the free Hamiltonian to generate a contraction semigroup, a gauge dependent shift must be implemented. In addition, it is shown that the interacting Hamiltonian is a Phillips perturbation of the free Hamiltonian, so that the full Hamiltonian generates a C_0 semigroup. Equivalence between gauge invariance, Gauss' Law, and positivity with respect to the Krein metric is demonstrated using a characterization of Gauss' Law used in Ref. 5 by means of the $\bar{\partial}$-Poincaré lemma. In Sec. III, the customary Gauss measure representation for this model is found by an extension of the Schwinger functions to coincident points which then leads to a path-space formula on the physical states for the physical self-adjoint Hamiltonian.

II. CANONCIAL QUANTIZATION

A. Indefinite metric

Ultimately, we wish to construct a gauge field A_μ (an operator valued distribution) on a Hilbert space $\mathcal{H}$, which contains a ground state, Ω_R, that is invariant under Poincaré and local gauge symmetries, and with A_μ both local and covariant. Several examples show that a modification of the usual Wightman axioms[7] is required such as suggested by Wightman and Gårding.[8] In particular, Strocchi and Wightman,[8] and Strocchi and others[9] have shown that if the gauge fields are coupled to a matter field which generates a local gauge symmetry, then an indefinite metric must be used for the existence of a nontrivial gauge field.

The indefinite metric is defined by a sesquilinear form $\{\cdot,\cdot\}$ related to the Hilbert space inner product $(\cdot,\cdot)$ by

$$\{\Phi,\Psi\}=(\Phi,\eta\Psi), \tag{13}$$

where η is required to be self-adjoint. The gauge field A_μ should be a symmetric operator with respect to the indefinite metric, i.e., $A_\mu\subset A_\mu^\dagger$, where † denotes the indefinite metric space adjoint, and * denotes the Hilbert space adjoint. In fact, the minimal closure $A_\mu^{\dagger\dagger}=A_\mu^{**}$ should be †-self-adjoint. Vacuum expectation values now satisfy

$$\{\mathcal{P}(A)\Omega_R,\mathcal{Q}(A)\Omega_R\}=(\Omega_R,\mathcal{P}(A)\mathcal{Q}(A)\Omega_R) \tag{14}$$

for polynomials in A_μ on the relativistic vacuum vector Ω_R, with $\eta\Omega_R=\Omega_R$. Physical states are defined in terms of the positive subspace

$$\mathcal{H}_+=\{\Phi\in\mathcal{H}|\{\Phi,\Phi\}\geqslant 0\} \tag{15}$$

with null vectors

$$\mathcal{H}_0=\{\Phi\in\mathcal{H}_+|\{\Phi,\Phi\}=0\} \tag{16}$$

as vectors in

$$\mathcal{H}_{\text{physical}}=\overline{\mathcal{H}_+/\mathcal{H}_0} \tag{17}$$

with the closure taken in the topology induced by the indefinite metric.

For a mathematically precise cutoff theory, we choose a spatially discrete periodic box V of side ''length'' L and lattice spacing δ. Fourier transforms are defined by

$$f_\mu(\mathbf{x}) = \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma} \hat{f}_\mu(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}}, \tag{18}$$

where Γ is the dual group to V. Our infrared cutoff is simply to restrict sums to $\mathbf{k}\in\Gamma_0 = \Gamma\backslash\{0\}$ with Γ chosen to be reflection symmetric about the origin in each variable. On this lattice, the time-zero gauge field may now be written as

$$A_\mu(\mathbf{x}) = V_\mu(\mathbf{x}) + S_\mu(\mathbf{x}), \tag{19}$$

where for $\mathbf{x}\in V$,

$$(\Omega_R, V_\mu(x)V_\nu(y)\Omega_R) = \frac{1}{V}\int_{-\infty}^{\infty} dk^0 \sum_{\mathbf{k}\in\Gamma_0} \theta(k^0)\,\delta(k^2-m^2) e^{-ik(x-y)}\left[\frac{k_\mu k_\nu}{m^2} - g_{\mu\nu} - i\epsilon_{\mu\nu\lambda}\frac{k^\lambda}{m}\right], \tag{20}$$

$$(\Omega_R, S_\mu(x)S_\nu(y)\Omega_R) = \frac{1}{V}\int_{-\infty}^{\infty} dk^0 \sum_{\mathbf{k}\in\Gamma_0} \theta(k^0) e^{-ik(x-y)}\left\{\delta(k^2)\left[\frac{-k_\mu k_\nu}{m^2} + i\epsilon_{\mu\nu\lambda}\frac{k^\lambda}{m}\right] + (1-2\xi)k_\mu k_\nu \delta'(k^2)\right\} \tag{21}$$

in which the midpoint lattice approximation

$$k_j = \frac{2\sin(k_j'\delta/2)}{\delta}, \qquad k'\in\Gamma_0 \tag{22}$$

for lattice derivatives is used. A parameterization using forward and backward approximations having a periodic Fourier transform is easily written down. Even though expression (22) is anti-periodic, it is more convenient for our purposes. The range of interest for ξ is $0\leqslant\xi\leqslant 1/2$, interpolating between the Feynman and Landau gauges.

Canonical quantization is implemented by means of the equal time commutation relations

$$[A_\mu(\mathbf{x}), \Pi_\nu(\mathbf{y})] = i g_{\mu\nu}\delta_{\mathbf{x},\mathbf{y}}. \tag{23}$$

A convenient representation for V_μ arises from the hermitian matrix appearing in (20), which has only one nonzero eigenvalue,

$$\lambda_2 = 2\left(\frac{\omega^2+m^2}{m^2}\right) \tag{24}$$

with

$$\omega^2 = 4\sum_{j=1}^{2} \frac{\sin^2(k_j\delta/2)}{\delta^2} \tag{25}$$

while the matrix for field S_μ in (21) in Landau gauge has two nonzero eigenvalues, denoted

$$\lambda_\pm = -\frac{\omega^2}{m^2} \pm \sqrt{\frac{\omega^4}{m^4} + \frac{2\omega^2}{m^2}} \tag{26}$$

and we define $\lambda_0 = -\lambda_-$ and $\lambda_1 = \lambda_+$. The presence of $\lambda_- < 0$ indicates an indefinite metric is required to recover the two point function for S_μ. The corresponding eigenvectors may readily be obtained as

$$M_2(\mathbf{k}) = \frac{1}{\sqrt{2}\omega\mu} \begin{pmatrix} \omega^2 \\ \mu k_1 - imk_2 \\ \mu k_2 + imk_1 \end{pmatrix} \tag{27}$$

and $M_0 = V_-$, $M_1 = V_+$, where

$$V_\pm(\mathbf{k}) = \frac{1}{N_\pm} \begin{pmatrix} \frac{\omega^2}{m^2}\left(\frac{\omega^2}{m^2}+1\right) \mp \frac{\omega^2}{m^2}\sqrt{\frac{\omega^2}{m^4}+\frac{2\omega^2}{m^2}} \\ \frac{\omega k_1}{m^2}\left(\frac{\omega^2}{m^2}+1\right) \mp \frac{\omega k_1}{m^2}\sqrt{\frac{\omega^4}{m^4}+\frac{2\omega^2}{m^2}} - i\frac{k_2}{m}\left(\frac{\omega^2}{m^2} \mp \sqrt{\frac{\omega^4}{m^4}+\frac{2\omega^2}{m^2}}\right) \\ \frac{\omega k_2}{m^2}\left(\frac{\omega^2}{m^2}+1\right) \mp \frac{\omega k_2}{m^2}\sqrt{\frac{\omega^4}{m^4}+\frac{2\omega^2}{m^2}} + i\frac{k_1}{m}\left(\frac{\omega^2}{m^2} \mp \sqrt{\frac{\omega^4}{m^4}+\frac{2\omega^2}{m^2}}\right) \end{pmatrix}, \tag{28}$$

$$N_\pm^2 = \left(\frac{2\omega^4}{m^4}\right)\left[2\left(\frac{\omega^2}{m^2}+1\right)\left(\frac{\omega^2}{m^2}+1 \mp \sqrt{\frac{\omega^4}{m^4}+\frac{2\omega^2}{m^2}}\right) + \left(\frac{\omega^2}{m^2} \mp \sqrt{\frac{\omega^4}{m^4}+\frac{2\omega^2}{m^2}}\right)\right]. \tag{29}$$

A representation of (23) for the time-zero gauge field may then be given by

$$A_\mu(\mathbf{x}) = \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \sum_{\alpha=0}^{2} \sqrt{\lambda_\alpha(\mathbf{k})}\left[M_{\mu\alpha}(\mathbf{k}) b_\alpha(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} + \overline{M_{\mu\alpha}(\mathbf{k}) b_\alpha^\dagger(\mathbf{k})} e^{-i\mathbf{k}\cdot\mathbf{x}}\right], \tag{30}$$

wherein

$$[\mathbf{b}_\mu(\mathbf{k}), b_\nu^\dagger(\mathbf{p})] = -g_{\mu\nu}\delta_{\mathbf{k},\mathbf{p}} \tag{31}$$

are the commutation relations with respect to the indefinite metric which make (30) formally †-symmetric. In terms of Fock space operators with

$$[c_\mu(\mathbf{k}), c_\nu^*(\mathbf{p})] = \delta_{\mathbf{k},\mathbf{p}} \tag{32}$$

place $b_\mu = c_\mu$, $b_\mu^\dagger = -g_{\mu\nu} c_\nu^*$.

The indefinite metric operator is the Gupta–Bleuler choice $\eta = -g$, where $g_{00} = 1$, $g_{ij} = -\delta_{ij}$, $j = 1,2$, and as such is a Krein metric operator, since $\eta^2 = 1$. As the eigenvectors in (27) and (28) are orthogonal to the eigenvectors for the zero eigenvalue, there is no loss of generality by working in the Landau gauge. As shown below, other gauge choices are related to each other by a Krein unitary transformation.

The massive field V_μ is given by $\alpha = 2$, while $\alpha = 0,1$ define the ghost field, S_μ. Clearly, on the Gårding domain obtained by applying polynomials in A_μ to the Fock vacuum, $A_\mu(\mathbf{x})$ is Krein symmetric and has a Krein self-adjoint closure $A_\mu(\mathbf{x})^{\dagger\dagger} = A_\mu(\mathbf{x})^{**}$. In the case of $V_\mu(\mathbf{x})$, it is readily seen that $V_\mu(\mathbf{x})$ defines an essentially self-adjoint operator on $\mathcal{H}$ and both metrics agree on the Fock space for V_μ. For the ghost piece S_μ, the situation is a bit more complicated, however we prove the following proposition.

Proposition II.1: $S_\mu = S_\mu^{(1)} + iS_\mu^{(0)}$, *where*

$$S_\mu^{(1)}(\mathbf{x}) = -\frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \sqrt{\lambda_1(\mathbf{k})}\left[M_{\mu 1}(\mathbf{k}) c_1(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} + \overline{M_{\mu 1}(\mathbf{k}) c_1^*(\mathbf{k})} e^{-i\mathbf{k}\cdot\mathbf{x}}\right], \tag{33}$$

$$S_\mu^{(0)}(\mathbf{x})=\frac{i}{\sqrt{V}}\sum_{\mathbf{k}\in\Gamma_0}\sqrt{\lambda_0(\mathbf{k})}[M_{\mu 0}(\mathbf{k})c_0(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}-\overline{M_{\mu 0}(\mathbf{k})c_0^*(\mathbf{k})}e^{-i\mathbf{k}\cdot\mathbf{x}} \tag{34}$$

defines an essentially normal operator on $\mathcal{H}$.

Proof: First, let $N=\Sigma_{\mathbf{k}\in\Gamma_0}c_\mu^*(\mathbf{k})c_\mu(\mathbf{k})$, where μ=0, 1, or 2, denote the number operator on $\mathcal{H}$, then $S_\mu^{(j)}(N+1)^{-1/2}$ are bounded operators. Consider the following set.

Definition II.2: Let $D(N,\alpha)$ *denote the dense set*

$$D(N,\alpha)=\{\Phi|\|e^{\alpha N^2\Phi}\Phi\|\equiv\|\Phi\|_{N,2,\alpha}<\infty,\alpha>0\} \tag{35}$$

which is a common core of analytic vectors for each $S_\mu^{(j)}$, since

$$\|S_\mu^{(j)^p}\Phi\|=\|(S_\mu^{(j)})^p(N+1)^{-p/2}(N+1)^{p/2}e^{-\alpha N^2}e^{\alpha N^2}\Phi\|\leqslant C_1^p p^{p/4}\|(S_\mu^{(j)})^p(N+1)^{-p/2}\|\|\Phi\|_{N,2,\alpha}.$$

By a Glimm–Jaffe bound,[10] after writing $S_\mu^{(j)^p}$ as a sum of 2^p Wick ordered monomials, we find

$$\|S_\mu^{(j)^p}\Phi\|\leqslant C_2^p p^{p/4}\|\Phi\|_{N,2,\alpha}$$

which means that $\Sigma_{p=0}^\infty(|t|^p/p!)\|S_\mu^{(j)^p}\Phi\|<\infty$ and each $S_\mu^{(j)}$ is essentially self-adjoint on D_N. Moreover,

$$\sum_{p_1=0}^{N_1}\frac{S_\mu^{(1)p_1}}{p_1!}\sum_{p_2=0}^{N_2}\frac{S_\mu^{(0)p_2}}{p_2!}\Phi=\sum_{p_2=0}^{N_2}\frac{S_\mu^{(0)p_2}}{p_2!}\sum_{p_1=0}^{N_1}\frac{S_\mu^{(1)p_1}}{p_1!}\Phi \tag{36}$$

with a bound uniform in N_1 and N_2,

$$\sum_{p_1=0}^{N_1}\sum_{p_2=0}^{N_2}\frac{(|t_1|c_4)^{p_1}(|t_2|c_5)^{p_2}}{p_1!p_2!}(p_1+p_2)^{(p_1+p_2)/4}\|\Phi\|_N\leqslant\sum_{p=0}^{N_1+N_2}\frac{p^{p/4}}{p!}(|t_1|c_4+|t_2|c_5)^p\|\Phi\|_N<\infty,$$

so that the double series in (36) converges uniformly in t_1 and t_2 in either order. Therefore,

$$e^{it_1S_\mu^{(1)}}e^{it_2S_\mu^{(0)}}\Phi=e^{it_2S_\mu^{(0)}}e^{it_1S_\mu^{(1)}}\Phi, \tag{37}$$

so each S_μ then defines an essentially normal operator on $\mathcal{H}$. □

Note, that since S_μ is a normal operator, it follows that $D(S_\mu)=D(S_\mu^*)=D(\eta S_\mu^\dagger\eta)$ and since $D(S_\mu)\subset D(S_\mu^\dagger)$, it follows that $D(S_\mu)=D(S_\mu^\dagger)$, that is, S_μ is Krein self-adjoint.

B. Haller and Lim-Lombridas representation

Canonical quantization of Chern–Simons QED in covariant gauges has been given by Haller and Lim-Lombridas[5] by using a representation for the ghost field S_μ which does not satisfy a standard representation for the CCR's. Since we wish to use their work for our analysis, we show how to obtain their representation from (30) and (31). This also clarifies the relation between the Krein space and Hilbert space in their work.

The scalar field R and G introduced in the introduction are described by b_0 and b_1 and should satisfy

$$[G(\mathbf{x}),R(\mathbf{y})]=\delta_{\mathbf{x},\mathbf{y}} \tag{38}$$

at equal times. The combination that relates Haller and Lim-Lombridas' a_Q and a_R to the b_0 and b_1 is

$$a_Q(\mathbf{k}) = \frac{[b_1(\mathbf{k}) - b_0(\mathbf{k})]}{\sqrt{2}\sigma(\mathbf{k})}, \tag{39}$$

$$a_R(\mathbf{k}) = \frac{\sigma(\mathbf{k})[b_1(\mathbf{k}) + b_0(\mathbf{k})]}{\sqrt{2}}, \tag{40}$$

$$a(\mathbf{k}) = -b_2(\mathbf{k}) \tag{41}$$

in which the scaling factor σ needs to be

$$\sigma = \frac{8\omega^{5/2}}{m^{5/2}} \left[\frac{\sqrt{1+\lambda_0} + \sqrt{1-\lambda_1}}{\sqrt{1+\lambda_0} - \sqrt{1-\lambda_1}} \right]^{1/2} = \frac{8\omega^2(\omega^2+2m^2)^{1/4}}{m^{5/2}} \tag{42}$$

for the two representations to agree. The operators a_Q and a_R then satisfy the Haller and Lim-Lombridas algebra

$$[a_Q(\mathbf{k}), a_Q^\dagger(\mathbf{k})] = 0 = [a_R(\mathbf{k}), a_R^\dagger(\mathbf{k})], \tag{43}$$

$$[a_Q(\mathbf{k}), a_R^\dagger(\mathbf{p})] = \delta_{\mathbf{k},\mathbf{p}} = [a_R(\mathbf{k}), a_Q^\dagger(\mathbf{p})]. \tag{44}$$

Inverting (39), (40), and (41) for b_μ in terms of a, a_Q, a_R and substituting into (30) reproduces (2.24) and (2.26) of Ref. 5 in Landau gauge. Throughout the remainder of this paper we shall often use these expressions for calculations (see Appendix A). Their utility lies in providing a convenient form for Gauss Law (Sec. III, pp. 12–16, of Ref. 5) and the fact that the Fock spaces $\mathcal{H}_Q$ and $\mathcal{H}_R$ generated by $a_Q^\dagger$ and $a_R^\dagger$, respectively, form null vectors in the Krein metric.

C. The Hamiltonian

With the representation of the gauge fields and their conjugate momenta, along with our understanding of the Krein and Hilbert spaces, we can form the Hamiltonian operator. The free Hamiltonian is decomposed as

$$H_0 = H_0^V + H_0^G(\xi), \tag{45}$$

where H_0^V is formed from the V_μ field, while $H_0^G(\xi)$ is formed from the S_μ field. In terms of Krein space operators, the Hamiltonian is given by

$$H_0^V = \sum_{\mathbf{k}\in\Gamma} \sqrt{\omega^2+m^2}\, b_2^\dagger(\mathbf{k}) b_2(\mathbf{k}), \tag{46}$$

$$\begin{aligned} H_0^G(\xi) = \sum_{\mathbf{k}\in\Gamma_0} \omega[b_0^\dagger(\mathbf{k})b_0(\mathbf{k}) + b_1^\dagger(\mathbf{k})b_1(\mathbf{k})] + \frac{(1-2\xi)}{2} \sum_{\mathbf{k}\in\Gamma_0} \frac{m^2}{\sqrt{\omega^2+2m^2}} [b_0^\dagger(\mathbf{k})b_0(\mathbf{k}) \\ + b_1^\dagger(\mathbf{k})b_1(\mathbf{k}) - b_1^\dagger(\mathbf{k})b_0(\mathbf{k}) - b_0^\dagger(\mathbf{k})b_1(\mathbf{k})], \end{aligned} \tag{47}$$

and we observe that $H_0 \subset H_0^\dagger$. However, on physical grounds for uniqueness, $\overline{H_0} = H_0^\dagger$ is necessary. To this end, the Hilbert space structure is used, and the Hamiltonian becomes

$$H_0^V = \sum_{\mathbf{k}\in\Gamma} \sqrt{\omega^2+m^2}\, c_2^*(\mathbf{k}) c_2(\mathbf{k}) \tag{48}$$

$$H_0^G(\xi)=\sum_{\mathbf{k}\in\Gamma_0}\omega[c_0^*(\mathbf{k})c_0(\mathbf{k})+c_1^*(\mathbf{k})c_1(\mathbf{k})]+\frac{(1-2\xi)}{2}\sum_{\mathbf{k}\in\Gamma_0}\frac{m^2}{\sqrt{\omega^2+2m^2}}[c_0^*(\mathbf{k})c_0(\mathbf{k})-c_1^*(\mathbf{k})c_1(\mathbf{k})+c_1^*(\mathbf{k})c_0(\mathbf{k})-c_0^*(\mathbf{k})c_1(\mathbf{k})] \tag{49}$$

and we observe that in Landau gauge this expression defines an essentially self-adjoint positive operator. However, in non-Landau gauges this expression is not positive definite.

To examine the free Hamiltonian further, recall the Fock representation relations

$$[c_\mu(\mathbf{k})\Phi]^{(n)}_{\mu_1,\dots,\mu_n}(\mathbf{k}_1,\dots,\mathbf{k}_n)=\sqrt{n+1}\,\Phi^{(n+1)}_{\mu\mu_1,\dots,\mu_n}(\mathbf{k},\mathbf{k}_1,\dots,\mathbf{k}_n), \tag{50}$$

$$[c_\mu^*(\mathbf{k})\Phi]^{(n)}_{\mu_1,\dots,\mu_n}(\mathbf{k}_1,\dots,\mathbf{k}_n)=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\delta_{\mu,\mu_j}\delta_{\mathbf{k},\mathbf{k}_j}\Phi^{(n-1)}_{\mu,\dots,\hat{\mu}_j,\dots,\mu_n}(\mathbf{k}_1\dots,\hat{\mathbf{k}}_j,\dots,\mathbf{k}_n), \tag{51}$$

so that we may determine the numerical range. A calculation gives

$$(\Phi,H_0^G(\xi)\Phi)=\sum_{\mathbf{k}\in\Gamma_0}\sum_{\{\mathbf{k}\}\in\Gamma_0}\sum_{\{\mu\}}\sum_{n}\omega\left(\left[1+\frac{(1-2\xi)m^2}{2\sqrt{2m^2+\omega^2}}\right](n+1)|\Phi^{(n+1)}_{0\{\mu\}}|^2+\left[1-\frac{(1-2\xi)m^2}{2\sqrt{2m^2+\omega^2}}\right]\times(n+1)|\Phi^{(n+1)}_{1\{\mu\}}|^2+\frac{(1-2\xi)m^2}{2\sqrt{2m^2+\omega^2}}[\overline{\Phi^{(n+1)}_{1\{\mu\}}}\Phi^{(n+1)}_{0\{\mu\}}-\overline{\Phi^{(n+1)}_{0\{\mu\}}}\Phi^{(n+1)}_{1\{\mu\}}]\right), \tag{52}$$

which is complex in general when $\xi\neq 1/2$ and symmetric with respect to the real axis. On the one particle subspace, let $h_0^G(\xi)=\overline{H_0^G(\xi)}_{|_{n=1}}$. Then both h_0^G and h_0^{G*} are quasiacretive since both $\mathrm{Re}(\phi,h_0^G\phi)$ and $\mathrm{Re}(\phi,h_0^{G*}\phi)$ are bounded below by $-[(1-2\xi)m/2\sqrt{2}]\|\phi\|^2$. Then, by Lemma 2.2 of Ref. 11, h_0^G is Krein self-adjoint. As shown below it follows easily that $\overline{H_0^G(\xi)}$ is also Krein self-adjoint.

The full Hamiltonian is expressed as a sum of unbounded operators,

$$H(\xi)=H_0^V+H_0^G(\xi)+H_0^D+H_{\mathrm{int}}, \tag{53}$$

where H_0^D is the free Dirac Hamiltonian and the interaction term is

$$H_{\mathrm{int}}=\sum_{\mathbf{x}\in V}\delta^2 j^\mu(\mathbf{x})A_\mu(\mathbf{x})=\sum_{\mathbf{x}\in V}\delta^2 j^\mu(\mathbf{x})(V_\mu(\mathbf{x})+S_\mu(\mathbf{x}))=H_{\mathrm{int}}^V+H_{\mathrm{int}}^G(\xi) \tag{54}$$

with the time-zero fermion current $j^\mu(\mathbf{x})=e{:}\,\psi^\dagger(\mathbf{x})\gamma^\mu\psi(\mathbf{x}){:}$. The following proposition summarizes the Hilbert and Krein space properties of these operators:

Proposition II.3: Let $\beta=[|(1-2\xi)|m]/2\sqrt{2}$. Then:

(a) *The closure $\overline{H_0^G(\xi)}$ is Krein self-adjoint for any real value of the gauge parameter ξ and self-adjoint for $\xi=1/2$.*

(b) *The closure $\overline{H_0^G(\xi)+\beta I}$ is maximal accretive and Krein self-adjoint.*

(c) *The boson interaction H_{int}^V is symmetric and Kato tiny relative to the positive self-adjoint operator $\overline{H_0^V}$, hence, $H_0^V+H_{\mathrm{int}}^V$ is essentially self-adjoint and bounded below.*

(d) *The ghost interaction, $H_{\mathrm{int}}^G(\xi)$ is Kato tiny relative to $H_0^G(\xi)+\beta I$, hence, $H_0^G(\xi)+H_{\mathrm{int}}^G+\beta I$ has a maximal quasiaccretive and Krein self-adjoint closure.*

(e) *The Hamiltonian $H(\xi)$ defines a unique Krein self-adjoint operator which is also maximal quasiaccretive by taking closures from any suitable core, for example, the finite particle vectors. The operators $\overline{H_0^V+H_{\mathrm{int}}^V}$ and $\overline{H_0^G(\xi)+H_{\mathrm{int}}^G+\beta I}$ commute as unbounded operators.*

Proof: (a) Suppose T is a closed operator on $\mathcal{H}$, then $T^\dagger = \eta T^* \eta$, so that T is Krein self-adjoint if and only if ηT is self-adjoint. Notice that for the one-particle operator $h_0^G = h_0^{G\dagger}$ is equivalent to $\eta h_0^G = (\eta h_0^G)^*$. For for the second quantized operator[12]

$$\overline{H_0^G =} d\Gamma(h_0^G) = d\Gamma(h_0^{G\dagger}) = \eta d\Gamma(h_0^G)^* \eta = H_0^{G\dagger}$$

with $\overline{\eta H_0^G}$ self-adjoint.

Consider the semigroup generated by $h_0(\xi)$ on the one partice space and let

$$M(\omega,t) = \begin{pmatrix} 1-\rho t & \rho t \\ -\rho t & 1+\rho t \end{pmatrix}, \tag{55}$$

where $\rho = [(1-2\xi)m^2]/(2\sqrt{\omega^2 + 2m^2})$, then

$$\|e^{-th_0(\xi)}\|^2 = \sum_{\mathbf{k}\in\Gamma_0} e^{-2t\omega}\overline{\phi_\alpha(\mathbf{k})(M^*M)}_{\alpha\beta}\phi_\beta(\mathbf{k}). \tag{56}$$

The Hermitian matrix $(M^*M)_{\alpha\beta}$ has eigenvalues $(\sqrt{1+\rho^2t^2} \pm \rho^2 t^2)^2$, and setting $|\rho|t = \sinh\theta$ leads us to bound terms in the sum above as

$$e^{-2t\omega_0}e^{\beta t}\exp\left[-\frac{\beta}{|\rho|}\left(\sinh\theta - \frac{|\rho|}{\beta}\theta\right)\right] \leqslant e^{\beta t} \tag{57}$$

since $|\rho|/\beta \leqslant 1$ uniformly in $\mathbf{k}$. Consequently, $h_0(\xi) + \beta I$ defines a C_0-contraction semigroup and by the Hille–Yosida theorem the closed operator $h_0(\xi) + \beta I$ is m-accretive. Part (b) now follows from

$$e^{-t(H_0^G(\xi)+\beta I)} = e^{-td\Gamma(h_0^G(\xi)+\beta I)} = \Gamma(e^{-t(h_0^G(\xi)+\beta I)}) \tag{58}$$

as the operator Γ preserves contractions.[12]

Proceeding as in Proposition (5.1) in Ref. 13, one easily obtains bounds

$$\|a(\mathbf{k})e^{-tH_0^V}\Phi\| \leqslant \frac{\text{const}}{\sqrt{t}}\|\Phi(\mathbf{k},\cdot)\|, \tag{59}$$

$$\|c_\mu(\mathbf{k})e^{-t(H_0^G(\xi)+\beta I)}\Phi\| \leqslant \frac{\text{const}}{\sqrt{t}}\|\Phi(\mathbf{k},\cdot)\|. \tag{60}$$

On the lattice, $j^\mu(\mathbf{x})$ are bounded operators so H_{int}^V and $H_{\text{int}}^G(\xi)$ are, respectively, Phillips perturbations of H_0^V and $H_0^G(\xi) + \beta I$ and are also Kato tiny. This proves (c) and (d) while (e) follows immediately. □

The Phillips nature of the perturbation for these semigroups is used in Sec. III by means of their Duhmael expansions.

D. Gauge transformations

The expressions for A_μ and Π_μ may be expressed entirely in terms of Landau gauge fields as

$$A_0(\mathbf{x};\xi) = A_0(\mathbf{x};1/2) - \frac{(1-2\xi)}{4}(-\Delta)^{-1}\left[\partial_l \Pi_l(\mathbf{x};1/2) - \frac{m}{2}\boldsymbol{\epsilon}_{ln}\partial_l A_n(\mathbf{x};1/2)\right], \tag{61}$$

$$A_l(\mathbf{x};\xi) = A_l(\mathbf{x};1/2) - \frac{(1-2\xi)}{4}(-\Delta)^{-1}\partial_l \Pi_0(\mathbf{x};1/2), \tag{62}$$

$$\Pi_0(\mathbf{x};\xi)=\Pi_0(\mathbf{x};1/2), \tag{63}$$

$$\Pi_l(\mathbf{x};\xi)=\Pi_l(\mathbf{x};1/2)+\frac{(1-2\xi)}{4}(-\Delta)^{-1}\frac{m}{2}\epsilon_{ln}\partial_n\Pi_0(\mathbf{x};1/2). \tag{64}$$

Formally, these relations may be implemented by an operator gauge transformation,

$$A_\mu(\mathbf{x};\xi)=e^T A_\mu(\mathbf{x};1/2)e^{-T}, \tag{65}$$

$$\Pi_\mu(\mathbf{x};\xi)=e^T\Pi_\mu(\mathbf{x};1/2)e^{-T} \tag{66}$$

in which

$$\begin{aligned} T&=\frac{-i}{4}\sum_{\mathbf{x}\in V}\delta^2(-\Delta)^{-1}(\partial A)[\operatorname{div}\mathbf{E}(\mathbf{x})-mB(\mathbf{x})]\\ &=\frac{-i(1-2\xi)}{4}\sum_{\mathbf{x}\in V}\delta^2(-\Delta)^{-1}\Pi_0(\mathbf{x};1/2)\left[\partial_l\Pi_l(\mathbf{x};1/2)-\frac{m}{2}\epsilon_{ln}\partial_l A_n(\mathbf{x};1/2)\right]\\ &=(1-2\xi)\sum_{\mathbf{k}\in\Gamma_0}\frac{16\omega^3}{m^3}[a_Q^\dagger(\mathbf{k})a_Q^\dagger(-\mathbf{k})-a_Q(\mathbf{k})a_Q(-\mathbf{k})] \end{aligned} \tag{67}$$

or in terms of Hilbert space adjoints

$$\begin{aligned} T=(1-2\xi)\sum_{\mathbf{k}\in V}\frac{m^2}{8\omega\sqrt{2m^2+\omega^2}}&[c_0(\mathbf{k})c_0(-\mathbf{k})+c_0^*(\mathbf{k})c_0^*(-\mathbf{k})\\ &+c_1(\mathbf{k})c_1(-\mathbf{k})+c_1^*(\mathbf{k})c_1^*(-\mathbf{k})+2c_0(\mathbf{k})c_1(-\mathbf{k})-2c_0^*(\mathbf{k})c_1^*(-\mathbf{k})]=T_1+T_2, \end{aligned} \tag{68}$$

where $T_1\subset T_1^*$ and $T_2\subset -T_2^*$. A calculation shows that $[T_1,T_2]\neq 0$, so that $\overline{T}$ cannot be realized as a normal operator as in Sec. II A. In fact, the closure of e^T is an unbounded Krein unitary operator.

Proposition II.4: (a) The operators $e^{\pm T}$ are densely defined with a Krein unitary closure on $\mathcal{H}$.

(b) The operator gauge transformations in (65) and (66) hold for the closures of $A_\mu(\mathbf{x})$ and $\Pi_\mu(\mathbf{x})$.

Proof: Vectors in $D(N,\alpha)$ form a core for $e^{\pm T}$, since noting that as a_Q, $a_Q^\dagger$ commute amongst themselves, $T^p=:T^p:$ is a sum of 2^p Wick monomials each of which is a sum of 4^p monomials in c and c^* with kernel $m^2/(8\omega\sqrt{2m^2+\omega^2})$. Repeating the calculation from Sec. II A gives

$$\|T^p\Phi\|\leqslant C_3(C_4)^p p!\,e^{-p^2/\alpha}\|e^{\alpha N^2}\Phi\|$$

which leads to

$$\sum_{p=0}^{\infty}\frac{|z|^p\|T^p\Phi\|}{p!}\leqslant C_3\sum_{p=0}^{\infty}(C_4|z|)^p e^{-p^2/\alpha}\|e^{\alpha N^2}\Phi\|\leqslant C\|e^{\alpha N^2}\Phi\|,$$

where C is an overall constant. Vectors in $D(N,\alpha)$ are then entire vectors for T.

Next, since $e^{\pm T}\subset e^{\pm(T_1-T_2)}$, $e^{\pm T}$ are closeable. Suppose Φ is in the domain of the closure with $\{\Phi_j\}\subset D(N,\alpha)$ such that $\Phi_j\xrightarrow{s}\Phi$, $e^{\pm T}\Phi_j\xrightarrow{s}\overline{e^{\pm T}}\Phi$. Clearly,

$$\|\{\overline{e^{\pm T}}\Phi,\overline{e^{\pm T}}\Phi\}-\{e^{\pm T}\Phi_j,e^{\pm T}\Phi_j\}\|\leqslant\|\overline{e^{\pm T}}\Phi-e^{\pm T}\Phi_j\|(\|\overline{e^{\pm T}}\Phi\|+\|e^{\pm T}\Phi_j\|)$$

as $\|\eta\|=1$. Using the relation, $T\subset -T^{\dagger}$,

$$|\{e^{\pm T}\Phi_j, e^{\pm T}\Phi_j\}-\{\Phi_j,\Phi_j\}|=\left|\lim_{N\to\infty}\sum_{\substack{p_1,p_2=1\\ N+1\leq p_1+p_2}}^{N}\frac{(-1)^p}{p_1!p_2!}\{\Phi_j, T^{p_1+p_2}\Phi_j\}\right|$$

$$\leq \lim_{N\to\infty}\sum_{p=N+1}^{2N}\frac{2^p}{p!}\|\Phi_j\|\|T^p\Phi_j\|$$

$$\leq\left(\lim_{N\to\infty}\sum_{p=N+1}^{2N} C_4^p e^{-p^2/\alpha}\right)(C_3\|\Phi_j\|\|e^{\alpha N^2}\Phi_j\|)=0.$$

Now letting $j\to\infty$ establishes (a).

The argument for A_μ and Π_μ is the same. The vectors in $D(N,\alpha)$ form a core for $\overline{A_\mu(\mathbf{x};\xi)}$ so we may again suppose a sequence Φ_j, A_μ convergent to Φ, whereby a simple calculation shows $e^T A_\mu(\mathbf{x})e^{-T}\Phi_j = A_\mu(\mathbf{x};\xi)\Phi_j$ since $A_\mu(\mathbf{x})e^{-T}\Phi_j\in D(e^T)$. Taking limits gives

$$\overline{e^T A_\mu(\mathbf{x})e^{-T}\Phi = A_\mu(\mathbf{x};\xi)\Phi}$$

and $e^T A_\mu(\mathbf{x})e^{-T}$ extends to the domain of the closure. Since $\overline{A_\mu(\mathbf{x};\xi)}$ is a normal operator, so is $\overline{e^T A_\mu(\mathbf{x})e^{-T}}$, establishing (b). □

Note, that similar observations to these apply to the electric and magnetic field operators

$$\overline{e^T E_l(\mathbf{x})e^{-T}=E_l(\mathbf{x}),}$$

$$\overline{e^T B(\mathbf{x})e^{-T}=B(\mathbf{x}).} \tag{69}$$

Consider the family of commuting unbounded operators[5]

$$\Omega(\mathbf{k})=a_Q(\mathbf{k})+\frac{m^{3/2}}{16\omega^3}j_0(\mathbf{k}) \tag{70}$$

which have $D(N^{1/2})$ as a common core and $D(N,\alpha)$ as common analytic vectors. Their closures define normal operators on $\mathcal{H}$. Form the unbounded operator[5]

$$\mathcal{G}=\frac{1}{\sqrt{V}}\sum_{\mathbf{k}\in\Gamma_0}\frac{8\omega^3}{m^{3/2}}[a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}+a_Q^{\dagger}(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}]+j_0(\mathbf{x}) \tag{71}$$

which is realized on $D(N^{1/2})$ as

$$\mathcal{G}(\mathbf{x})=\mathcal{G}_0(\mathbf{x})+\mathcal{G}_1(\mathbf{x})+j_0(\mathbf{x}) \tag{72}$$

and where $\mathcal{G}_0$ and j_0 are symmetric and $\mathcal{G}_1$ is skew-symmetric. These operators all commute so by the same argument as in Sec. II A, $\overline{\mathcal{G}}$ is realized as a commuting family of normal operators. One may then define a Krein unitary operator as

$$U(\lambda)=\exp\left[i\sum_{\mathbf{y}\in V}\delta^2\lambda(\mathbf{y})\mathcal{G}(\mathbf{y})\right]$$

$$=\exp\left[i\sum_{\mathbf{y}\in V}\delta^2\lambda(\mathbf{y})\mathcal{G}_0(\mathbf{y})\right]\exp\left[i\sum_{\mathbf{y}\in V}\delta^2\lambda(\mathbf{y})\mathcal{G}_1(\mathbf{y})\right]\exp\left[i\sum_{\mathbf{y}\in V}\delta^2\lambda(\mathbf{y})j_0(\mathbf{y})\right] \tag{73}$$

in which the first and third are unitary operators on $\mathcal{H}$ but the second is self-adjoint. By similar calculations to those in Proposition II.4 for λ a function in V, the transformations

$$U(\lambda)A_0(\mathbf{x})U(-\lambda)=A_0(\mathbf{x}), \tag{74}$$

$$U(\lambda)\Pi_0(\mathbf{x})U(-\lambda)=\Pi_0(\mathbf{x}), \tag{75}$$

$$U(\lambda)A_l(\mathbf{x})U(-\lambda)=A_l(\mathbf{x})-\partial_l\lambda(\mathbf{x}), \tag{76}$$

$$U(\lambda)\Pi_l(\mathbf{x})U(-\lambda)=\Pi_l(\mathbf{x})-\frac{m}{2}\,\epsilon_{ln}\partial_n\lambda(\mathbf{x}), \tag{77}$$

$$U(\lambda)\psi(\mathbf{x})U(-\lambda)=e^{-ie\lambda(\mathbf{x})}\psi(\mathbf{x}), \tag{78}$$

are valid on $D(\overline{A_\mu})$ and $D(\overline{\Pi_\mu})$ and $\mathcal{H}$, respectively. For any Gauss vector Φ_G, Ψ_G defined below one finds

$$\{\Phi_G,\mathcal{G}(\mathbf{x})\Psi_G\}=0 \tag{79}$$

as well as

$$\{U(\lambda)\Phi,U(\lambda)\Psi\}=\{\Phi,\Psi\} \tag{80}$$

on $\overline{D(U(\lambda))}\supset$Gauss vectors. The electric and magnetic fields are also invariant with respect to these c-number gauge transformations

$$\begin{aligned}\overline{U(\lambda)E_l(\mathbf{x})}U(-\lambda)&=\overline{E_l(\mathbf{x}),}\\ \overline{U(\lambda)B(\mathbf{x})}U(-\lambda)&=\overline{B(\mathbf{x})}.\end{aligned} \tag{81}$$

E. Gauss vectors

For precise statements about the domains of the unbounded operators appearing in (70), it is convenient to further exploit our choice of cutoffs by introducing harmonic oscillator coordinates (see Appendix B for definitions.) The ghost subspace is now represented as a direct sum of spaces $\mathcal{H}^{(n)}$, which are L_2 spaces. If $N=|\Gamma_0|$ is the number of nonzero momentum modes, then after elimination of redundant p's and q's, there are N-pairs of complex variables, z_1 and z_2, such that $\mathcal{H}^{(n)}$ is the nN-fold tensor product of $L_2(C^{2N},\mu_0;\mathcal{H}_V\otimes\mathcal{H}_F)$, where μ_0 is Lebesgue measure. The differential operators ∂_1, ∂_2, $\overline{\partial_1}$, $\overline{\partial_2}$ and their adjoints are defined in the distributional sense on their maximal domains. The advantage of this representation, is that it allows for a convenient characterization of the physical states, or Gauss vectors.

Definition II.5: The Gauss vectors are vectors in

$$\bigcap_{\mathbf{k}\in\Gamma_0}\text{Null}(\Omega(\mathbf{k}))=\mathcal{H}_G. \tag{82}$$

The subspace $\mathcal{H}_G$ is closed in $\mathcal{H}$ and can be related to the von Neumann algebra generated by the maximal commuting family of operators $\{\Omega(\mathbf{k})|\mathbf{k}\in\Gamma_0\}$. In terms of these harmonic oscillator coordinates, the Gauss vectors satisfy the following pair of equations:

$$[(\tfrac{1}{2}z_1+\overline{\partial_1})-i(\tfrac{1}{2}z_2+\overline{\partial_2})+\sqrt{2}\sigma\gamma j_0(\mathbf{k})]\Phi_G(\mathbf{k})=0, \tag{83}$$

$$[(\tfrac{1}{2}+\overline{\partial_1})+i(\tfrac{1}{2}z_2+\overline{\partial_2})+\sqrt{2}\sigma\gamma j_0(-\mathbf{k})]\Phi_G(\mathbf{k})=0, \tag{84}$$

where $\mathbf{k}\in\Gamma_0^+$ and $\gamma=(m^{3/2}/16\omega^3)$. As in Ref. 5, the following operator is defined:

$$D=\sum_{\mathbf{k}\in\Gamma_0}\gamma[a_R(\mathbf{k})j_0(-\mathbf{k})-a_R^\dagger(\mathbf{k})j_0(\mathbf{k})]$$
$$=\sum_{\mathbf{k}\in\Gamma_0^+}\frac{\gamma\sigma}{\sqrt{2}}[j_0(-\mathbf{k})(\overline{z_1}-i\overline{z_2})+j_0(\mathbf{k})(\overline{z_1}+i\overline{z_2})]. \tag{85}$$

A simple calculation then shows that

$$e^{-D}e^{|z|^2/2}[(\tfrac{1}{2}z_1+\overline{\partial}_1)-i(\tfrac{1}{2}z_2+\overline{\partial}_2)+\sqrt{2}\sigma\lambda j_0(\mathbf{k})]e^De^{-|z|^2/2}=\overline{\partial}_1-i\overline{\partial}_2, \tag{86}$$

$$e^{-D}e^{|z|^2/2}[(\tfrac{1}{2}z_1+\overline{\partial}_1)+i(\tfrac{1}{2}z_2+\overline{\partial}_2)+\sqrt{2}\sigma\lambda j_0(-\mathbf{k})]e^De^{-|z|^2/2}=\overline{\partial}_1+i\overline{\partial}_2, \tag{87}$$

therefore, the Gauss condition becomes a $\overline{\partial}$-equation. By applying techniques for the $L_2(C,\mu_0)-\overline{\partial}$-theory[14] one may prove the following theorem.

Theorem II.6: *The Gauss vectors have the form*

$$\Phi_G=e^De^{-|z|^2/2}F(z_1,z_2;\Phi_V,\Phi_F), \tag{88}$$

where F is analytic in z_1 and z_2 with coefficients in $\mathcal{H}_V\otimes\mathcal{H}_F$.

Corollary II.7: The operators $e^{\pm D}$ are densely defined with Krein unitary closure on $\mathcal{H}$. Moreover, for any two Gauss vectors Φ_G and Ψ_G,

$$\{\Phi_G,\Psi_G\}=(\Phi_V,\Psi_V), \tag{89}$$

where Φ_V,Ψ_V are, respectively, the projections of Φ_G, Ψ_G onto the subspace $\mathcal{H}_V\otimes\mathcal{H}_F$.

Proof: The proof follows that of Proposition II.4. From the definition of the harmonic oscillator coordinates, it follows that $D\subset -D^\dagger$, so $e^{\pm D}$ are closable operators. Suppose Φ is in the domain of closure and $\Phi_j\in C_0^\infty(\mathbb{R}^{nN},\mathcal{H}_V\otimes\mathcal{H}_F)$, such that $\Phi_j\overset{s}{\rightarrow}\Phi$, and $e^{\pm D}\Phi_j\overset{s}{\rightarrow}e^{\pm\overline{D}}\Phi$. Then, as in Proposition II.4,

$$|\{e^{\pm D}\Phi_j,e^{\pm D}\Phi_j\}-\{\Phi_j,\Phi_j\}|\leq\lim_{N\to\infty}\sum_{p=N+1}^{2N}\frac{2^p}{p!}\|\Phi_j\|\|D^p\Phi_j\|,$$

but since D is multiplication in $\overline{z_k}$, $D^p\Phi_j\in C_0^\infty$, and, due to the $p!$ in the denominator, the limit is 0.

The Gauss vector matrix elements are now of the form

$$\{\Phi_G,\Psi_G\}=\{e^{-|z|^2/2}F_\Phi,e^{-|z|^2/2}F_\Psi\}.$$

Since F is analytic, it may be expressed as a power series in z_1 and z_2. Using the definition of the harmonic oscillator coordinates, we have that

$$z_1(\mathbf{k})=\frac{\sigma}{\sqrt{2}}[a_Q^\dagger(\mathbf{k})+a_Q^\dagger(-\mathbf{k})-a_Q(\mathbf{k})-a_Q(-\mathbf{k})], \tag{90}$$

$$z_2(\mathbf{k})=\frac{i\sigma}{\sqrt{2}}[-a_Q^\dagger(\mathbf{k})+a_Q^\dagger(-\mathbf{k})-a_Q(\mathbf{k})+a_Q(-\mathbf{k})], \tag{91}$$

but, $\{a_Q e^{-|z|^2/2}, a_Q e^{-|z|^2/2}\}=0$ as well as the other combination of a_Q and $a_Q^\dagger$. Therefore, only the zeroth order term in the expansion survives, having coefficients in $\mathcal{H}_V \otimes \mathcal{H}_F$. □

A straightforward calculation shows that applying a gauge transformation to a Gauss vector does not leave the Gauss vector invariant. Rather, it produces another Gauss vector.

Proposition II.8: The Gauss subspace is an invariant subspace for the operator $e^{\pm T}$, *i.e.,* $e^{\pm T}\mathcal{H}_G \subset \mathcal{H}_G$.

The results of this section may be used to prove the following theorem.

Theorem II.9: *The following characterizations of the Gauss vectors are equivalent:*

(a) Positivity with respect to the indefinite metric;
(b) Gauge invariance;
(c) Gauss' Law.

Proof: Proposition II.8 and the remarks following (73), show that applying a gauge transformation to a Gauss vector produces another Gauss vector, so that shows that (c) implies (b). Corollary II.7 shows that the Gauss vectors are projections on $\mathcal{H}_V \otimes \mathcal{H}_F$, but, the metric η_R is positive on this subspace, so that, (c) implies (a). As gauge transformations act only on the ghost variables, it also follows from Corollary II.7 that (b) implies (a). Finally, states that are positive with respect to η are states formed from ϕ and the fermion variables, and since the ghost variables have zero norm, (a) implies (c). □

III. EUCLIDEAN FIELD

A. Two point function

The probability measure for the free Euclidean field is a complex Gauss measure and is uniquely characterized by its covariance, or two-point function. To calculate the covariance, we define the operators

$$\hat{A}_\mu(t,\mathbf{x}) = e^{-tH_0} A_\mu(0,\mathbf{x}) e^{tH_0} \quad t>0 \tag{92}$$

and their anti-time ordered product,

$$\langle \overline{T}\hat{A}_\mu(x)\hat{A}_\nu(y)\rangle = \begin{cases} (\Omega_R, \hat{A}_\mu(x)\hat{A}_\nu(y)\Omega_R) & \text{if } x_0<y_0 \\ (\Omega_R, \hat{A}_\mu(y)\hat{A}_\nu(x)\Omega_R) & \text{if } y_0<x_0 \end{cases}. \tag{93}$$

The decomposition $A_\mu = V_\mu + S_\mu$ leads to

$$\langle \overline{T}\hat{A}_\mu(x)\hat{A}_\nu(y)\rangle = \langle \overline{T}\hat{V}_\mu(x)\hat{V}_\nu(y)\rangle + \langle \overline{T}\hat{S}_\mu(x)\hat{S}_\nu(y)\rangle, \tag{94}$$

for which a short calculation shows that the anti-time ordered products for V_μ and S_μ are not continuous at equal time Schwinger points. However, the simple discontinuities for V_μ and S_μ exactly cancel, allowing the Schwinger functions for $\hat{A}_\mu$ to be extended by continuity to the Euclidean region. We shall choose this extension, whereupon

$$\langle \overline{T}(i\hat{A}_0(x))(i\hat{A}_0(y))\rangle = \langle B_0(x)B_0(y)\rangle, \tag{95}$$

$$\langle \overline{T}(i\hat{A}_0(x))(\hat{A}_l(y))\rangle = \langle B_0(x)B_l(y)\rangle, \tag{96}$$

$$\langle \overline{T}(\hat{A}_l(x))(i\hat{A}_0(y))\rangle = \langle B_l(x)B_0(y)\rangle, \tag{97}$$

$$\langle \overline{T}\hat{A}_l(x)\hat{A}_n(y)\rangle = \langle B_l(x)B_n(y)\rangle \tag{98}$$

defining the Euclidean field, B_μ, with covariance

$$\langle B_\mu(x)B_\nu(y)\rangle=\frac{1}{2\pi V}\sum_{\mathbf{k}\in\Gamma_0}\int_{-\infty}^{\infty}e^{ik(x-y)}\left[\frac{1}{k^2+m^2}\left(\delta_{\mu\nu}-\frac{k_\mu k_\nu}{k^2}-\frac{m\epsilon_{\mu\nu\lambda}k_\lambda}{k^2}\right)+(1-2\xi)\frac{k_\mu k_\nu}{(k^2)^2}\right]. \tag{99}$$

For f,g real test functions, B_μ may be realized as a real-valued Gaussian random field on a separable Hilbert space, but with a complex measure

$$\langle B(f)B(g)\rangle=\int_\Omega d\mu_{CS}(\omega)f(\omega)g(\omega) \tag{100}$$

where Ω may be chosen to be $\mathcal{S}'_R$, for example. A detailed discussion of this measure is given elsewhere.[15]

Here we deal with the semigroup representation on Gauss vectors in terms of a path-space integral. As we will show in the next section, this requires eliminating the ghost field S_μ, and defining an extension for the anti-time ordered products of V_μ to coincident Schwinger points. The extension we choose is one such that there exists a Euclidean field V_μ^E which can be expressed in terms of ϕ^E, a scalar Euclidean field, analogous to the field ϕ defined above. In this way, Osterwalder–Schrader positivity is a direct manifestation of the physical Hilbert space inner product for Gauss vectors.

The Minkowski expression for V_μ in terms of ϕ is

$$V_0(\mathbf{x})=-\frac{\sqrt{-\Delta_s}}{m}\phi(\mathbf{x}), \tag{101}$$

$$V_l(\mathbf{x})=\frac{\partial_l\pi(\mathbf{x})}{m\sqrt{-\Delta_s}}-\frac{\epsilon_{ln}\partial_n\phi(\mathbf{x})}{\sqrt{-\Delta_s}} \tag{102}$$

in which the presence of the canonical momentum $\pi(\mathbf{x})=\dot{\phi}(\mathbf{x})$ gives rise to a contact term in the anti-time ordered products through the relations

$$\langle\overline{T}(\hat{\phi}(x)\hat{\phi}(y))\rangle=\langle\phi^E(x)\phi^E(y)\rangle, \tag{103}$$

$$\langle\overline{T}(\hat{\phi}(x)\hat{\pi}(y))\rangle=\langle\phi^E(x)(-i\dot{\phi}^E(y))\rangle, \tag{104}$$

$$\langle\overline{T}(\hat{\pi}(x)\hat{\phi}(y))\rangle=\langle(-i\dot{\phi}^E(x))\phi^E(y)\rangle, \tag{105}$$

$$\langle\overline{T}(\hat{\pi}(x)\hat{\pi}(y))\rangle=\langle(-i\dot{\phi}^E(x))(-i\dot{\phi}^E(y))\rangle+\delta(x_0-y_0)\delta_{\mathbf{x},\mathbf{y}}. \tag{106}$$

In these relations, ϕ^E is the Euclidean scalar field with covariance

$$\langle\phi^E(x)\phi^E(y)\rangle=\frac{1}{2\pi V}\sum_{\mathbf{k}\in\Gamma_0}\int_{-\infty}^{\infty}dk\,\frac{e^{ik(x-y)}}{k^2+m^2} \tag{107}$$

and is valid with these cutoffs in the sense of generalized random processes. Define the Euclidean analog of (101) and (102) by

$$V_0^E(x)=i\,\frac{\sqrt{-\Delta_s}}{m}\phi^E(x), \tag{108}$$

$$V_l^E(x)=\frac{-i\partial_l\pi^E(x)}{m\sqrt{-\Delta_s}}-\frac{\epsilon_{ln}\partial_n\phi^E(x)}{\sqrt{-\Delta_s}}, \tag{109}$$

whereupon

$$\langle \overline{T}(iV_0(x))(iV_0(y))\rangle = \langle V_0^E(x)V_0^E(y)\rangle, \tag{110}$$

$$\langle \overline{T}(iV_0(x))V_l(y)\rangle = \langle V_0^E(x)V_l^E(y)\rangle, \tag{111}$$

$$\langle \overline{T}V_l(x)(iV_0(y))\rangle = \langle V_l^E(x)V_0^E(y)\rangle, \tag{112}$$

$$\langle \overline{T}V_l(x)V_n(y)\rangle = \langle V_l^E(x)V_n^E(y)\rangle + \frac{\partial_{l,x}\partial_{n,y}}{m^2(-\Delta_s)}\ \delta_{\mathbf{x},\mathbf{y}}\delta(x_0-y_0), \tag{113}$$

then up to contact terms, V_μ is given by the Euclidean field V_μ^E, constructed from ϕ^E. The expectation values for V_μ^E are to be calculated using the Gauss measure for ϕ^E and the relations (108) and (109).

B. Duhamel formula

As shown in Sec. II C, the Landau gauge Hamiltonian is the infinitesimal generator of a contraction semigroup. In addition, Theorem II.9 shows this choice of gauge to be sufficient to calculate matrix elements between Gauss states.

Proposition III.1: On the Gauss vectors we have for $t>0$,

$$\{\Phi_G, e^{-tH}\Psi_G\} = \{\Phi_V, e^{-t\tilde{H}}\Psi_V\} = (\Phi_V, e^{-t\tilde{H}}\Psi_V), \tag{114}$$

where

$$\tilde{H} = H_0^V + H_0^D + \sum_{\mathbf{x}\in V} \delta^2 j_\mu(\mathbf{x})V^\mu(\mathbf{x}) + \sum_{\mathbf{k}\in\Gamma_0} \frac{j_0^*(\mathbf{k})j_0(\mathbf{k})}{2m^2} + \sum_{\mathbf{k}\in\Gamma_0} \frac{i j_0^*(\mathbf{k})\epsilon_{ln}k_l j_n(\mathbf{k})}{m\omega^2} \tag{115}$$

and Φ_V, Ψ_V *are the projections of* Φ_G, Ψ_G *onto the physical subspace* $\mathcal{H}_V \otimes \mathcal{H}_F$.

Proof: First, we approximate the Gauss vectors with functions ϕ, $\psi \in C_0^\infty(\mathbb{R}^{nN}; \mathcal{H}_V\otimes\mathcal{H}_F)$, which are dense in L_2, but need not be Gauss vectors themselves. This approximation is necessary to handle the unbounded operators $e^{\pm D}$. Next, we approximate the exponential function as

$$e^{-tH} = s - \lim_{n\to\infty}\left(1+\frac{t}{n}H\right)^{-n} \tag{116}$$

since H is quasi m-accretive (c.f. Ref. 16, Chap. IX). Finally, we calculate

$$\{e^D\Phi_G, e^{-tH}e^D\Psi_G\} = \lim_{n\to\infty}\left\{\phi, e^{-D}\left(1+\frac{t}{n}H\right)^{-n} e^D\psi\right\} \tag{117}$$

using Corollary II.7.

Upon using the harmonic oscillator representation, the ghost part of the Hamiltonian becomes a second-order differential operator with polynomial coefficients. Therefore, $e^{-D}He^D\phi \in C_0^\infty$, and one has

$$e^{-D}\left(1+\frac{t}{n}H\right)^n e^D\phi = \left(1+\frac{t}{n}e^{-D}He^D\right)^n \phi \tag{118}$$

for Re $t>0$. The calculation of the commutators is now elementary, leading to the expression

$$e^{D}\left(1+\frac{t}{n}H\right)e^{-D}\phi=\left(1+\left[\frac{t}{n}H,D\right]+\frac{1}{2}\left[\left[\frac{t}{n}H,D\right],D\right]\right)\phi=\left(1+\frac{t}{n}[\tilde{H}+H_0^G+V_Q]\right)\phi \tag{119}$$

in which V_Q contains terms linear in a_Q and $a_Q^\dagger$, with coefficients that are bounded fermion currents (Ref. 5, Eq. (3.17)).

The arguments used for Proposition II.3 show that $\overline{\tilde{H}+H_0^G+V_Q}$ is a quasi-*m*-accretive operator, generating a C_0 semigroup, denoted $e^{-tH'}$. As a result, we have

$$e^{-D}\left(1+\frac{t}{n}H\right)^{n}e^{D}\psi=\left(1+\frac{t}{n}H'\right)^{n}\psi$$

again for $\mathrm{Re}\, t>0$. The Hille–Yosida theorem allows us to conclude that $(1+(t/n)H')^{-n}$ is a bounded operator on Fock space and agrees with $e^{-D}(1+(t/n)H)^{-n}e^{D}$ on C_0^∞, and therefore defines a unique, bounded extension for this expression. Furthermore, it follows that H' is quasi-*m*-accretive, so that

$$s-\lim_{n\to\infty} e^{-D}\left(1+\frac{t}{n}H'\right)^{-n}e^{D}=e^{-tH'}. \tag{120}$$

Therefore, the limit in (117) exists and we find

$$\{\Phi_G,e^{-tH}\Psi_G\}=\{\phi,e^{-tH'}\psi\} \tag{121}$$

for the matrix element.

The final step is to remove all the ghost terms from H' in the prior expression. This is accomplished by using the fact that V_Q is a Phillips perturbation of the free Hamiltonian, so that a norm convergent Duhamel series

$$e^{-tH'}=\sum_{n=0}^{\infty}(-1)^n\int_0^\infty du_1\cdots du_n\quad e^{-u_1(\tilde{H}+H_0^G)}V_Q e^{-u_2(\tilde{H}+H_0^G)}V_Q\cdots V_Q e^{-(t-u_n)(\tilde{H}+H_0^G)}, \tag{122}$$

where the integration variables are anti-time ordered, may be used. Operators in V_Q are of the form

$$\sum_{\mathbf{k}\in\Gamma_0}[\rho(\mathbf{k})a_Q(\mathbf{k})+\tau(\mathbf{k})a_Q^\dagger(\mathbf{k})], \tag{123}$$

where $\rho(\mathbf{k})$ and $\tau(\mathbf{k})$ are bounded fermion operators. These terms do not commute with $\tilde{H}$, but $\tilde{H}$ and H_0^G do commute. From

$$a_Q(\mathbf{k})e^{-sH_0^G}\psi=e^{-s\omega(\mathbf{k})}e^{-sH_0^G}a_Q(\mathbf{k})\psi=0,$$

where $\psi=e^{-|z|^2}/2F$, and the commutation of a_Q and $a_Q^\dagger$, successive use of this pull-through formula eliminates all a_Q terms in V_Q from (117). Only the $n=0$ term remains in (123) which produces

$$\{\phi,e^{-tH'}\psi\}=\{\phi,e^{-t\tilde{H}}e^{-tH_0^G}\psi\}.$$

The term $H_0^G(1/2)$ is given by

$$H_0^G(1/2)=\sum_{\mathbf{k}\in\Gamma_0}\omega(\mathbf{k})[a_Q^\dagger(\mathbf{k})a_R(\mathbf{k})+a_R^\dagger(\mathbf{k})a_Q(\mathbf{k})] \tag{124}$$

and one may easily show that in the sense of unbounded operators, each of the terms in the above equation commute with $\Sigma_{\mathbf{k}\in\Gamma_0}\omega(\mathbf{k})a_R^\dagger(\mathbf{k})a_Q(\mathbf{k})\psi=0$. This operator may be expressed in terms of Fock space operators as

$$\frac{1}{2}\sum_{\mathbf{k}\in\Gamma_0}\omega(\mathbf{k})[c_1(\mathbf{k})-c_0(\mathbf{k})]^*[c_1(\mathbf{k})-c_0(\mathbf{k})]$$

which defines a positive operator and generates a self-adjoint semigroup. In this manner, the two ghost parts in (124) may be separated in $e^{-tH_0^G}$, with the second term equal to one on ψ, and the first term gives one on ϕ as well, after taking Krein adjoints.

Now that all the ghost operators have been removed from the Gauss matrix elements in (117), Corollary II.7 further reduces the null vectors in ϕ and ψ to their projections onto $\mathcal{H}_V\otimes\mathcal{H}_F$. □

C. Path space formulas

Proposition III.1 showed that the Hamiltonian on the physical subspace is an essentially self-adjoint operator and that the interaction was a Phillips perturbation of the free Hamiltonian. As a result, the Hamiltonian is the infinitesimal generator of a C_0 semigroup and has a uniformly convergent Duhamel series expansion. The operator $\tilde{H}$ has additional fermion terms, but these are bounded operators, so they are also Phillips perturbations. For this operator, the Duhamel expansion is

$$\begin{aligned}(\Phi_V,e^{-t\tilde{H}}\Psi_V)&=\sum_{n=0}^{\infty}(-1)^n\int_0^\infty du_1\cdots du_n(\Phi_V,e^{-u_1H_0}\tilde{V}e^{-(u_2-u_1)H_0}\tilde{V}\cdots\tilde{V}e^{-(t-u_n)H_0})\\&=\sum_{n=0}^{\infty}(-1)^n\int_0^\infty du_1\cdots du_n(\Phi_v,\hat{\tilde{V}}(u_1)\cdots\hat{\tilde{V}}(u_n)\Psi_V),\end{aligned} \tag{125}$$

where the integration variable is anti-time ordered, $0\leqslant u_1\leqslant\cdots u_n\leqslant t$. Using a two dimensional version of Osterwalder–Schrader Euclidean fermions[17] sketched in Appendix C, the interaction terms

$$\tilde{V}=\sum_{\mathbf{x}\in V}\delta^2\left[j_\mu(\mathbf{x})V^\mu(\mathbf{x})+\frac{j_0(\mathbf{x})j_0(\mathbf{x})}{2m^2}+\frac{j_0(\mathbf{x})\epsilon_{nl}\partial_l j_n(\mathbf{x})}{m(-\Delta_s)}\right] \tag{126}$$

may be rewritten as

$$\hat{\tilde{V}}=e^{-t\tilde{H}_0}\tilde{V}e^{t\tilde{H}_0} \tag{127}$$

$$=W\sum_{\mathbf{x}\in V}\delta^2\left[j_\mu(\mathbf{x})V^\mu(\mathbf{x})+\frac{j_0(\mathbf{x})j_0(\mathbf{x})}{2m^2}+\frac{j_0(\mathbf{x})\epsilon_{nl}\partial_l j_n(\mathbf{x})}{m(-\Delta_s)}\right]. \tag{128}$$

Notice, that in the Wick expansion of the inner product in (125) when the anti-time ordered products are converted to expectations with respect to the Gauss measure for ϕ^E the extra terms on the RHS of (113) must be taken into account. It is easy to see that the net effect of this term is to add a further nonlocal interaction,

$$W\sum_{\mathbf{x}\in V}\left(\frac{\partial_l j_l^E(\mathbf{x})}{m(-\Delta)^{1/2}}\right)^2. \tag{129}$$

Consequently, removing the time ordering as all Euclidean expressions commute,

$$\begin{aligned}(\Phi_V, e^{-tH}\Psi_V) &= \left(I_0\Phi_V, W \exp\left[-\int_0^t du \sum_{\mathbf{x}\in V} \delta^2 V^E(u) \right] U(t)\Psi_V \right) \\ &= \int d\mu_0(\phi^E) \left(I_0\Phi_V, W \exp\left[-\int_0^t du \sum_{\mathbf{x}\in V} \delta^2 V^E(u) \right] I(t)\Psi_V \right), \end{aligned} \tag{130}$$

where ϕ^E is the scalar, Euclidean field defined above, and I is the corresponding Euclidean embedding map. In this way, the path-space formula on physical states satisfies Osterwalder–Schrader positivity for a Feynman–Kac–Nelson integral with a nonlocal fermion interaction density.

APPENDIX A: HALLER AND LIM-LOMBRIDAS' REPRESENTATION

$$\begin{aligned}A_0(\mathbf{x}) = &\frac{-1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{4\omega^3}{m^{7/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] - \frac{(1-2\xi)}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{2\omega}{m^{3/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} \\ &+ a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] + \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{m^{3/2}}{16\omega^2} [a_R(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a_R^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] \\ &- \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{\omega}{m\sqrt{2\mu}} [a(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}],\end{aligned} \tag{A1}$$

$$\begin{aligned}A_l(\mathbf{x}) = &\frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{8i\omega\epsilon_{ln}k_n}{m^{5/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} - a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] + \frac{(1-2\xi)}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{2k_l}{m^{3/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} \\ &+ a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] - \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{4\omega^2 k_l}{m^{7/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] \\ &+ \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{m^{3/2}k_l}{16\omega^3} [a_R(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a_R^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] - \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{\sqrt{\mu}k_l}{\sqrt{2}m\omega} [a(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} \\ &+ a^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] + \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{i\epsilon_{ln}k_n}{\omega\sqrt{2\mu}} [a(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} - a^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}],\end{aligned} \tag{A2}$$

$$\Pi_0(\mathbf{x}) = -\frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{8i\omega^2}{m^{3/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} - a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}], \tag{A3}$$

$$\begin{aligned}\Pi_l(\mathbf{x}) = &\frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{4i\omega k_l}{m^{3/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} - a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] - \frac{(1-2\xi)}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{\epsilon_{ln}k_n}{m^{1/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} \\ &+ a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] - \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{6\omega^2\epsilon_{ln}k_n}{m^{5/2}} [a_Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a_Q^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] \\ &- \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{m^{5/2}\epsilon_{ln}k_n}{32\omega^3} [a_R(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a_R^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] - \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{imk_l}{2\omega\sqrt{2\mu}} [a(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} \\ &- a^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}] - \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\in\Gamma_0} \frac{\sqrt{\mu}\epsilon_{ln}k_n}{2^{3/2}\omega} [a(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a^\dagger(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}].\end{aligned} \tag{A4}$$

APPENDIX B: HARMONIC OSCILLATOR COORDINATES

We define ghost space harmonic oscillator coordinates in the following manner:

$$
\begin{aligned}
q_{01}(\mathbf{k}) &= \tfrac{1}{2}[b_0(\mathbf{k})+b_0^\dagger(\mathbf{k})+b_0(-\mathbf{k})+b_0^\dagger(-\mathbf{k})],\\
q_{02}(\mathbf{k}) &= \frac{i}{2}[b_0(\mathbf{k})-b_0^\dagger(\mathbf{k})-b_0(-\mathbf{k})+b_0^\dagger(-\mathbf{k})],\\
p_{01}(\mathbf{k}) &= \frac{i}{2}[b_0(\mathbf{k})-b_0^\dagger(\mathbf{k})+b_0(-\mathbf{k})-b_0^\dagger(-\mathbf{k})],\\
p_{02}(\mathbf{k}) &= \tfrac{1}{2}[-b_0(\mathbf{k})-b_0^\dagger(\mathbf{k})+b_0(-\mathbf{k})+b_0^\dagger(-\mathbf{k})],
\end{aligned}
\tag{B1}
$$

$$
\begin{aligned}
q_{11}(\mathbf{k}) &= \tfrac{1}{2}[b_1(\mathbf{k})+b_1^\dagger(\mathbf{k})+b_1(-\mathbf{k})+b_1^\dagger(-\mathbf{k})],\\
q_{12}(\mathbf{k}) &= \frac{i}{2}[b_1(\mathbf{k})-b_1^\dagger(\mathbf{k})-b_1(-\mathbf{k})+b_1^\dagger(-\mathbf{k})],\\
p_{11}(\mathbf{k}) &= \frac{i}{2}[b_1(\mathbf{k})-b_1^\dagger(\mathbf{k})+b_1(-\mathbf{k})-b_1^\dagger(-\mathbf{k})],\\
p_{12}(\mathbf{k}) &= \tfrac{1}{2}[-b_1(\mathbf{k})-b_1^\dagger(\mathbf{k})+b_1(-\mathbf{k})+b_1^\dagger(-\mathbf{k})],
\end{aligned}
\tag{B2}
$$

one-half of which are redundant under $\mathbf{k}\leftrightarrow-\mathbf{k}$. One can verify that

$$[q_{\mu j}(\mathbf{k}),p_{\nu l}(\mathbf{k}')]=ig_{\mu\nu}\delta_{jl}\delta_{\mathbf{k},\mathbf{k}'} \tag{B3}$$

and that all other commutators vanish. From these definitions, suppressing the **k**-dependence, the following representation for the ghost operators arises:

$$a_Q(\mathbf{k})=\frac{1}{2\sqrt{2}\sigma}[(q_{11}-ip_{11})-(iq_{12}+p_{12})-(q_{01}-ip_{01})+(iq_{02}+p_{02})], \tag{B4}$$

$$a_Q^\dagger(\mathbf{k})=\frac{1}{2\sqrt{2}\sigma}[(q_{11}+ip_{11})+(iq_{12}-p_{12})-(q_{01}+ip_{01})+(-iq_{02}+p_{02})], \tag{B5}$$

$$a_R(\mathbf{k})=\frac{\sigma}{2\sqrt{2}}[(q_{11}-ip_{11})-(iq_{12}+p_{12})+(q_{01}-ip_{01})-(iq_{02}+p_{02})], \tag{B6}$$

$$a_R^\dagger(\mathbf{k})=\frac{\sigma}{2\sqrt{2}}[(q_{11}+ip_{11})+(iq_{12}-p_{12})+(q_{01}+ip_{01})-(-iq_{02}+p_{02})]. \tag{B7}$$

Next, we obtain a Schrödinger representation by means of von Neumann's theorem using the following:

$$q_{0j}=\frac{\partial}{\partial x_{0j}},\quad q_{1j}=-i\,\frac{\partial}{\partial x_{1j}},\quad p_{0j}=ix_{0j},\quad p_{1j}=x_{1j}. \tag{B8}$$

The following linear combinations:

$$z_1=x_{01}+ix_{11},\qquad z_2=x_{02}+ix_{12}, \tag{B9}$$

are useful and using them we can express the ghost operators as the following:

$$
\begin{aligned}
a_Q(\mathbf{k}) &= \frac{1}{\sqrt{2}\sigma}\left[-\left(\frac{1}{2}z_1+\overline{\partial_1}\right)+i\left(\frac{1}{2}z_2+\overline{\partial_2}\right)\right],\\
a_Q^\dagger(\mathbf{k}) &= \frac{1}{\sqrt{2}\sigma}\left[\left(\frac{1}{2}z_1-\overline{\partial_1}\right)+i\left(\frac{1}{2}z_2-\overline{\partial_2}\right)\right],\\
a_R(\mathbf{k}) &= \frac{\sigma}{\sqrt{2}}\left[\left(\frac{1}{2}\overline{z_1}+\partial_1\right)-i\left(\frac{1}{2}\overline{z_2}+\partial_2\right)\right],\\
a_R^\dagger(\mathbf{k}) &= \frac{\sigma}{\sqrt{2}}\left[-\left(\frac{1}{2}\overline{z_1}-\partial_1\right)-i\left(\frac{1}{2}\overline{z_2}-\partial_2\right)\right].
\end{aligned}
\tag{B10}
$$

APPENDIX C: EUCLIDEAN FERMIONS

In this appendix, the description of the two component Euclidean fermions as developed by Osterwalder and Schrader[17] is presented.

The time-zero fermion fields are defined as

$$\psi(0,\mathbf{x}) = \frac{1}{\sqrt{V}}\sum_{\mathbf{k}\in\Gamma}\sqrt{\frac{m_f}{E(\mathbf{k})}}\,[b(\mathbf{k})u(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}+d^*(\mathbf{k})v(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}], \tag{C1}$$

$$\psi^\dagger(0,\mathbf{x}) = \psi^*(0,\mathbf{x})\gamma_0 = \frac{1}{\sqrt{V}}\sum_{\mathbf{k}\in\Gamma}\sqrt{\frac{m_f}{E(\mathbf{k})}}\,[d(\mathbf{k})v^\dagger(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}+b^*(\mathbf{k})u^\dagger(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}], \tag{C2}$$

where $E(\mathbf{k}) = \sqrt{\mathbf{k}^2+m_f^2}$, and there are the standard anticommutation relations $[b(\mathbf{k}),b^*(\mathbf{p})]_+ = \delta_{\mathbf{k},\mathbf{p}} = [d(\mathbf{k}),d^*(\mathbf{p})]_+$. The two-dimensional gamma matrices that we use are

$$\gamma^0 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},\quad \gamma^1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},\quad \gamma^2 = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}, \tag{C3}$$

where $\gamma^\mu\gamma^\nu = g^{\mu\nu} - i\epsilon^{\mu\nu\alpha}\gamma_\alpha$ is one of the important features of the two-dimensional gamma matrices, the other being that there is no $\gamma.^5$ From this representation, the free Hamiltonian is obtained

$$H_0^D = \sum_{\mathbf{k}\in\Gamma} E(\mathbf{k})[b^*(\mathbf{k})b(\mathbf{k})+d^*(\mathbf{k})d(\mathbf{k})] \tag{C4}$$

in the usual manner.

The Dirac Hamiltonian can be used to define the following fields in the same manner as in Sec. III C, namely,

$$\hat{\psi}(t,\mathbf{x}) = e^{-tH_0^D}\psi(0,\mathbf{x})e^{tH_0^D}, \tag{C5}$$

$$\hat{\psi}^\dagger(t,\mathbf{x}) = e^{-tH_0^D}\psi^\dagger(0,\mathbf{x})e^{tH_0^D}, \tag{C6}$$

which are then used to compute the two-point function

$$(\Omega_R, \hat{\psi}(x), \hat{\psi}^\dagger(y)\Omega_R) = \frac{1}{2\pi V}\sum_{\mathbf{k}\in\Gamma}\int_{-\infty}^{\infty} dk^0 e^{ik(x-y)}\left(\frac{m_f+i\gamma^E\cdot k}{m_f^2+k^2}\right), \tag{C7}$$

where

$$\gamma_0^E=\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix},\quad \gamma_1^E=\begin{pmatrix}0 & i\\ -i & 0\end{pmatrix},\quad \gamma_2^E=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \tag{C8}$$

are the Euclidean gamma matrices. Diagonalizing the matrix in (C7),

$$m+i\gamma^E\cdot k=S_E(k)\begin{pmatrix}m_f+i|k| & 0\\ 0 & m_f-i|k|\end{pmatrix}S_E^{-1}(k) \tag{C9}$$

allows for the definition of the vectors

$$W_\pm^1=\begin{pmatrix}\sqrt{\pm(i|k|+m_f)}\\ 0\end{pmatrix},\quad W_\pm^2=\begin{pmatrix}0\\ \sqrt{\pm(i|k|-m_f)}\end{pmatrix} \tag{C10}$$

from which the Euclidean spinors may be defined,

$$\begin{aligned}u^j(k)&=S_E(k)W_+^j(k),\quad v^j(k)=S_E(-k)W_-^j(k),\\ \hat{u}^j(k)&=(W_+^j)^T(k)S_E(k)^{-1},\quad \hat{v}^j(k)=(W_-^j)^T(k)S_E(-k)^{-1}.\end{aligned} \tag{C11}$$

Note that in these expressions, $|k|=\sqrt{k^2}=\sqrt{k_0^2+k_1^2+k_2^2}$, that is, the Euclidean length of the vector.

The Euclidean fermi fields may now be defined as

$$\Psi_\alpha^{(1)}(x)=\frac{1}{\sqrt{2\pi V}}\sum_{\mathbf{k}\in\Gamma}\sum_{j=1,2}\int_{-\infty}^{\infty}dk_0\left[\frac{B(k,j)u_\alpha^j(k)e^{ikx}+D^*(k,j)v_\alpha^j(k)e^{-ikx}}{\sqrt{k^2+m_f^2}}\right], \tag{C12}$$

$$\Psi_\alpha^{(2)}(x)=\frac{1}{\sqrt{2\pi V}}\sum_{\mathbf{k}\in\Gamma}\sum_{j=1,2}\int_{-\infty}^{\infty}dk_0\left[\frac{D(k,j)\hat{v}_\alpha^j(k)e^{ikx}+B^*(k,j)\hat{u}_\alpha^j(k)e^{-ikx}}{\sqrt{k^2+m_f^2}}\right], \tag{C13}$$

where $[B(k,j),B^*(p,l)]_+=\delta_{\mathbf{k},\mathbf{p}}\delta(k_0-p_0)\delta_{jl}=[D(k,j),D^*(p,l)]_+$ and all other terms anticommute. The Euclidean fields have the following vacuum expectation value:

$$(\Omega_E,\Psi_\alpha^{(i)}\Psi_\beta^{(j)*}\Omega_E)=\delta_{ij}\delta_{\alpha\beta}(-\Delta+m^2)^{-1/2}, \tag{C14}$$

and anticommutator,

$$\{\Psi_\alpha^{(i)},\Psi_\beta^{(j)*}\}=2\delta_{ij}\delta_{\alpha\beta}(-\Delta+m^2)^{-1/2}. \tag{C15}$$

From these definitions, the current operator is then given by

$$j_\mu^E(x)=:\Psi^{(2)}\gamma_\mu^E\Psi^{(1)}(x):. \tag{C16}$$

For this paper, the reflection operator Θ is defined as

$$\Theta\Psi_\alpha^{(1)}\Theta^{-1}=(\Psi^{(2)}\gamma_0^E)_\alpha^*(\theta x),\quad \Theta\Psi_\alpha^{(2)}\Theta^{-1}=(\Psi^{(1)}\gamma_0^E)_\alpha^*(\theta x), \tag{C17}$$

where $\theta x=(-x_0,\mathbf{x})$. Let vectors dense in $\mathcal{H}_E$ be denoted by

$$:\Psi^{(1)}(x,\alpha)\Psi^{(2)}(y,\beta):=:\Psi_{\alpha_1}^{(1)}(x_1)\cdots\Psi_{\alpha_n}^{(1)}(x_n)\Psi_{\beta_1}^{(2)}(y_1)\cdots\Psi_{\beta_n}^{(2)}(y_n): \tag{C18}$$

and vectors which are dense in $\mathcal{H}_R$ by

$$:\hat{\psi}(x,\alpha)\hat{\psi}(y,\beta):=:\hat{\psi}_{\alpha_1}(x_1)\cdots\hat{\psi}_{\alpha_n}(x_n)\hat{\psi}_{\beta_1}(y_1)\cdots\hat{\psi}_{\beta_n}(y_n):. \tag{C19}$$

The embedding map is then defined by

Definition C.1: When $x_{j0}>0$, $y_{l_0}>0$ *and for* $j,l=1,2,...,n$ *there is a subspace* $\mathcal{E}_+^0 \subset \mathcal{H}_E$ *and a map* $W: \mathcal{E}_+^0 \to \mathcal{H}_R$ *defined by*

$$W{:}\Psi^{(1)}(x,\alpha)\Psi^{(2)}(y,\beta){:}\Omega_E = {:}\hat{\psi}(x,\alpha)\hat{\psi}(y,\beta){:}\Omega_R . \tag{C20}$$

The following lemmas directly transcribe from Ref. 17.

Lemma C.2: Let $X,Y \in \mathcal{E}_+^0$, *then* $(WX,WY)=(\Theta X,Y)$.

Lemma C.3: For $t \geqslant 0$ *and* $X \in \mathcal{E}_+^0$, *let* $U(t)$ *denote time translation in* $\mathcal{H}_E$. *Then,*

$$WU(t)X = e^{-tH_0^D}WX. \tag{C21}$$

[1] S. M. Girvin, ''The quantum Hall effect,'' in *The Quantum Hall Effect*, edited by S. M. Girvin and R. E. Prange (Springer-Verlag, Berlin, 1990).

[2] J. Fröhlich and P. A. Marchetti, ''Quantum field theories of vortices and anyons,'' Commun. Math. Phys. **121**, 177–223 (1989).

[3] F. Wilczec, ''Magnetic flux, angular momentum, and statistics,'' Phys. Rev. Lett. **48**, 1114–1146 (1982); ''Quantum mechanics of fractional-spin particles,'' *ibid.* **49**, 957–959 (1982).

[4] J. Fröhlich, G. Gabbiani, and P. A. Marchetti, ''Braid statistics in three-dimensional local quantum theory,'' in *Physics, Geometry, and Topology*, edited by H. C. Lee (Plenum, New York, 1990).

[5] K. Haller and E. Lim-Lombridas, ''Maxwell–Chern–Simons theory in covariant and Coulomb gauges,'' Ann. Phys. (N.Y.) **246**, 1–48 (1996).

[6] S. Deser, R. Jackiw, and S. Templeton, ''Topologically massive gauge theories,'' Ann. Phys. (N.Y.) **140**, 372–411 (1982).

[7] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Addison–Wesley, Reading, 1989).

[8] A. S. Wightman and L. Gårding, ''Fields as operator-valued distributions in relativistic quantum theory,'' Ark. Fys. **28**, 129–184 (1964); F. Strocchi and A. S. Wightman, ''Proof of the charge superslection rule in local relativistic quantum field theory,'' J. Math. Phys. **15**, 2198–2224 (1974); **17**, 1930–1931(E) (1976).

[9] F. Strocchi, ''Spontaneous symmetry breaking in local quantum gauge theory: The Higgs mechanism,'' Commun. Math. Phys. **56**, 57–78 (1977).

[10] J. Glimm and A. Jaffe, ''Quantum field theory models,'' in *Statistical Mechanics and Quantum Field Theory (Les Houches 1970)*, edited by C. DeWitt and R. Stora (Gordon and Breach, New York, 1971).

[11] J. L. Challifour, ''Self-adjointness of lattice Yang–Mills Hamiltonians and Kato's inequality with indefinite metric,'' Ann. Inst. Henri Poincaré **42**, 1–15 (1985).

[12] B. Simon, *The* $\mathcal{P}(\Phi)_2$ *Euclidean (Quantum) Field Theory* (Princeton University Press, Princeton, 1974).

[13] J. L. Challifour and W. A. Wood, ''Path-space formulas for covariant gauges in Krein spaces,'' J. Math. Phys. **28**, 2763–2774 (1987).

[14] L. Hørmander, *An Introduction to Complex Variables in Several Variables*, 2nd ed. (Von Nostrand, Princeton, 1973), Chap. 4.

[15] J. L. Challifour (in preparation).

[16] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed. (Springer-Verlag, Berlin, 1980).

[17] K. Osterwalder and R. Schrader, ''Euclidean Fermi fields and a Feynmac–Kac formula for Boson–Fermion models,'' Helv. Phys. Acta **46**, 277–302 (1973).

Multidimensional integrable Schrödinger operators with matrix potential

O. A. Chalykh[a)]
Department of Mathematics and Mechanics, Moscow State University, Moscow, 119899, Russia

V. M. Goncharenko[b)]
Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, LE 11 3TU, United Kingdom

A. P. Veselov[c)]
Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, LE 11 3TU, United Kingdom and Landau Institute for Theoretical Physics, Kosygina 2, Moscow, 117940, Russia

(Received 25 February 1999; accepted for publication 14 June 1999)

The Schrödinger operators with matrix rational potential, which are *D*-integrable, i.e., can be intertwined with the pure Laplacian, are investigated. Corresponding potentials are uniquely determined by their singular data which are a configuration of the hyperplanes in $\mathbf{C}^n$ with prescribed matrices. We describe some algebraic conditions (matrix locus equations) on these data, which are sufficient for *D*-integrability. As the examples some matrix generalizations of the Calogero–Moser operators are considered.
[S0022-2488(99)00911-1]

I. INTRODUCTION

Let us consider a Schrödinger operator

$$L=-\Delta+U(z),$$

where $z\in\mathbf{R}^n$ or $\mathbf{C}^n$ and $U(z)$ is a matrix-valued meromorphic function. We will call such an operator as *D-integrable* if there exists a differential operator $\mathcal{D}$ with meromorphic matrix coefficients and constant scalar highest term, such that

$$L\,\mathcal{D}=\mathcal{D}\,L_0, \tag{1}$$

where $L_0=-\Delta$ is the pure Laplacian acting on the vector-valued functions (cf. Refs. 1–3, where the scalar case $d=1$ has been considered).

In dimension $n=1$ all such operators with rational potentials can be described as the results of matrix Darboux transformations (see Ref. 4), which explains the terminology.

In dimension $n>1$ the situation is much more complicated even in the scalar case. For the review of the known results in this direction we refer to the recent paper.[5] In particular, as it has been shown in Refs. 1,2 the singularities of the potential $U(z)$ of any *D*-integrable Schrödinger operator have to be located on a union of the hyperplanes. The proof given in Refs. 1,2 works also in the matrix case under an additional assumption of the regularity (see below theorem 3), so in the rational case the potential $U(z)$ should have a form

a)Electronic mail: chalykh@mech.math.msu.su
b)Electronic mail: V.M.Gontcharenko@lboro.ac.uk
c)Electronic mail: A.P.Veselov@lboro.ac.uk

$$U(z)=\sum_{i=1}^{N}\frac{(\alpha_i,\alpha_i)A_i}{((\alpha_i,z)+c_i)^2}.$$

Such a potential is determined by a configuration of the hyperplanes Π_i in $\mathbf{C}^n$ given by the equations $(\alpha_i,z)+c_i=0$ with prescribed constant matrices A_i.

In the present paper we describe the conditions (so-called *matrix locus equations*) on these data which guarantee *D*-integrability. This generalizes to the matrix case the main result of the paper.[3] Locus equations can be interpreted as the conditions of the local trivial monodromy for the corresponding Schrödinger equations (cf. Refs. 4–6). This allows us to construct the examples of such configurations and related matrix *D*-integrable Schrödinger operators. Our proof of the existence of the intertwining operator $\mathcal{D}$ is effective; the corresponding formula is a matrix version of the Berest's formula.[7]

Some important examples of such operators were known in the theory of the generalized matrix Calogero–Moser systems (see Refs. 8–12), although the fact of their *D*-integrability seems to have not been emphasized. The corresponding operators have the form

$$L=-\Delta+\sum_{\alpha\in\mathcal{R}_+}\frac{m_\alpha(m_\alpha I-s_\alpha)(\alpha,\alpha)}{(\alpha,z)^2},$$

where $\mathcal{R}$ is a root system in $\mathbf{R}^n$ related to some Coxeter group G, m_α is an integer-valued G-invariant function on $\mathcal{R}$, s_α is the matrix of reflection with respect to the hyperplane $(\alpha,z)=0$.

We show that the operator L is *D*-integrable also for some non-Coxeter configurations $\mathcal{R}$ discovered in the scalar case in Refs. 13,14. Remarkably enough the matrix locus equations in this case turned out to coincide with the condition of the existence of the rational Baker–Akhiezer function in the so-called ''old axiomatics'' proposed in Refs. 15 and 16 (see Ref. 5 for the detailed discussion of this notion). This explains the appearance in the matrix case of the same configurations as in the scalar situation. Another interesting relation between the scalar and matrix generalizations of the Calogero–Moser system has been proposed recently by Bracken and Kamran[12] (see also Ref. 11).

We should mention that in the classical case matrix generalization of the Calogero–Moser system were introduced first by Gibbons and Hermsen in Ref. 17 (see Ref. 18 for further results in this direction). Our results show that the quantum situation is actually much richer than the classical one.

II. MONODROMY OF THE MATRIX SCHRÖDINGER EQUATIONS IN THE COMPLEX DOMAIN

Let us start with the one-dimensional case following essentially Ref. 4. Let

$$L=-D^2+U(z),\ z\in\mathbf{C},\quad D=\frac{d}{dz} \tag{2}$$

be a Schrödinger operator with meromorphic $d\times d$-matrix potential $U(z)$. Let $z=0$ be a regular singular point, i.e., a pole of the second order of $U(z)$. Consider a formal solution of the Schrödinger equation

$$L\psi=\lambda\psi \tag{3}$$

in the form

$$\psi=z^{-m}\sum_{i\geqslant 0}\psi_{-m+s}z^s. \tag{4}$$

Substituting (4) into the Schrödinger equation (3) with

$$U(z)=\frac{C_{-2}}{z^2}+\frac{C_{-1}}{z}+\sum_{r\geqslant 0}^{\infty} C_r z^r, \tag{5}$$

we obtain that ψ_{-m} is an eigenvector of C_{-2},

$$C_{-2}\psi_{-m}=m(m+1)\psi_{-m}.$$

If for any λ we can construct a basis of solutions of (3) with integer m (i.e., ψ is single-valued) then we say that the operator L has *local trivial monodromy* around $z=0$. This is equivalent to the fact that all the solutions of the corresponding matrix Schrödinger equation (3) are single-valued near $z=0$ for all λ. In this case one can prove that C_{-2} is diagonalizable with eigenvalues $\lambda_i=m_i(m_i+1)$, $i=1,2,\ldots,$ where $m_i\in\mathbf{Z}$ (see Ref. 4). Thus C_{-2} has a form

$$C_{-2}=\sum_{i=1}^{k} m_i(m_i+1)P_i, \tag{6}$$

where P_i are commuting projectors to the corresponding eigenspaces,

$$P_iP_j=\delta_{ij}P_i, \quad \sum_{i=1}^{k} P_i=I,$$

where I is identity operator. We assume that $0\leqslant m_1<m_2<\cdots<m_k=M$. The following result[4] gives the conditions on the coefficients C_j of the expansion of the potential (5) which are equivalent to the local trivial monodromy of L.

Theorem 1: *A matrix Schrödinger operator (2) with a meromorphic potential (5) has local trivial monodromy around* $z=0$ *if and only if* C_{-2} *has a form (6) and the coefficients* C_l *with* $l=-1,0,\ldots,2M-1$ *satisfy the relation*

$$P_iC_lP_j=0, \tag{7}$$

when $|m_i-m_j|\geqslant l+1$ *or* $m_i+m_j=l+1,l+3,\ldots,l+2k+1,\ldots$ *(i.e., when* m_i+m_j-l *is a positive odd number). In particular, the matrix residue* $C_{-1}=0$.

The coefficients $\psi_{-M},\psi_{-M+1},\ldots,\psi_{M-1}$ *of the corresponding expansions of the vector-eigenfunctions*

$$\psi=z^{-M}(\psi_{-M}+z\,\psi_{-M+1}+\cdots+z^k\psi_{-M+k}+\cdots)$$

satisfy the conditions

$$P_i\psi_l=0$$

if $m_i+l<0$ *or* $m_i+l=1,3,\ldots,2k+1,\ldots,2m_i-1$ *for* $m_i\geqslant 1$.

Notice that for $l=0$ conditions (7) are equivalent to the commutativity relation

$$[C_{-2},C_0]=0. \tag{8}$$

Now let us consider the multidimensional case. We will assume that the potential $U(z)$ of the Schrödinger operator

$$L=-\Delta+U(z), \tag{9}$$

where $z=(z_1,\ldots,z_n)\in\mathbf{C}^n$, $\Delta=(\partial^2/\partial z_1^2)+\cdots+(\partial^2/\partial z_n^2)$, is a meromorphic $d\times d$ matrix-valued function having a pole of the second order along the hyperplane $\Pi_\alpha:(\alpha,z)=0$, which is assumed to be nonisotropic: $(\alpha,\alpha)\neq 0$ (cf. Refs. 1,2,5). We will suppose for simplicity that $(\alpha,\alpha)=1$. The Laurent expansion of the potential in the normal direction α at the vicinity of Π_α can be written in the form

$$U(z)=\sum_{r\geqslant -2} C_r(\alpha,z)^r, \tag{10}$$

where $C_r=C_r(z^\perp)$ are some analytic $d\times d$ matrix-valued functions on the hyperplane Π_α and $z^\perp$ is orthogonal projection of z onto Π_α. Let us suppose that there exists a formal solution of the Schrödinger equation

$$L\psi=\lambda\psi \tag{11}$$

of the form

$$\psi(z)=(\alpha,z)^{-m}\sum_{s\geqslant 0} \psi_{-m+s}(\alpha,z)^s \tag{12}$$

for some m, where the coefficients $\psi_r=\psi_r(\lambda,z^\perp)$ are analytic vector-functions on Π_α. Substituting series (10) and (12) into Eq. (11) one can see that

$$C_{-2}\psi_{-m}=m(m+1)\psi_{-m},$$

i.e., ψ_{-m} is an eigenvector of C_{-2} with the eigenvalue $m(m+1)$.

Definition: We say that a Schrödinger operator (9) with the potential (10) has *local trivial monodromy around the hyperplane* Π_α if

(1) at any point of Π_α matrix C_{-2} is diagonalizable with eigenvalues having the form $m(m+1)$, $m\in\mathbf{Z}$,
(2) for any $m\in\mathbf{Z}$ such that $m(m+1)$ is an eigenvalue of C_{-2} and for any choice of the corresponding eigenvector $\psi_{-m}(z^\perp)$ there exists a formal solution (12) of Eq. (11) for any λ (notice that any such eigenvalue can be represented in the form $\mu(\mu+1)$ in two different ways: $\mu=m$ or $\mu=-m-1$).

In principle, C_{-2} might depend on the point at the hyperplane Π_α, but this is not the case.

Lemma 1: If the operator (9) has a local trivial monodromy around the hyperplane Π_α*, then* C_{-2} *is a constant matrix.*

To prove the lemma we can assume without loss of generality that $\alpha=(1,0,\ldots,0)$. Then the Schrödinger operator can be written in the form

$$L=-\frac{\partial^2}{\partial z_1^2}-\tilde{\Delta}+U(z), \tag{13}$$

where $\tilde{\Delta}=(\partial^2/\partial z_2^2)+\cdots+(\partial^n/\partial z_n^2)$. Thus we can consider (13) as a one-dimensional Schrödinger operator with the matrix operator-valued ''potential''

$$\tilde{U}(z)=-\tilde{\Delta}+U(z).$$

Applying formally theorem 1 and, in particular (8), we have

$$[C_0-\tilde{\Delta},C_{-2}]\equiv 0$$

or

$$[C_0,C_{-2}]-2\cdot\sum_{k=2}^{n}(\partial_k C_{-2})\partial_k-\tilde{\Delta}(C_{-2})\equiv 0.$$

Therefore, $\partial_k C_{-2}=0$ for all $k=2,\ldots,n$, i.e., C_{-2} is a constant. Alternative, more rigorous way to prove the lemma is to repeat the arguments of the proof of theorem 1 (see Refs. 4–6).

Now simularly to the one-dimensional case[4] (see theorem 1 above) one can prove the following:

Theorem 2: *A matrix Schrödinger operator (9) with a meromorphic potential (10) has local trivial monodromy around the hyperplane Π_α if and only if*

(1) C_{-2} is a constant diagonalizable matrix,

$$C_{-2}=\sum_{i=1}^{k} m_i(m_i+1)P_i\,,$$

where $0\leqslant m_1<m_2<\cdots<m_k=M$ are some integers, P_i are commuting projectors:

$$P_iP_j=\delta_{ij}P_i\,,\quad \sum_{i=1}^{k}P_i=I.$$

(2) The coefficients C_l with $l=-1,0,\ldots,2M-1$ satisfy the following relations:

$$P_iC_lP_j\equiv 0 \tag{14}$$

if $|m_i-m_j|\geqslant l+1$ or $m_i+m_j=l+1,l+3,\ldots,l+2k+1,\ldots$. In particular, $C_{-1}\equiv 0$ and $[C_0,C_{-2}]\equiv 0$.

The coefficients $\psi_{-M},\psi_{-M+1},\ldots,\psi_{M-1}$ of the corresponding expansions of the vector-eigenfunctions

$$\psi=(\alpha,z)^{-M}(\psi_{-M}+(\alpha,z)\psi_{-M+1}+\cdots+(\alpha,z)^k\psi_{-M+k}+\cdots)$$

satisfy the conditions

$$P_i\psi_l\equiv 0 \tag{15}$$

if $m_i+l<0$ or $m_i+l=1,3,\ldots,2k+1,\ldots,2m_i-1$ for $m_i\geqslant 1$.

III. MATRIX LOCUS EQUATIONS AND *D*-INTEGRABILITY

Let us consider a matrix Schrödinger operator (9) with a rational potential $U(z)$ decaying at infinity. We will assume that all the singularities are regular, i.e., $U(z)$ has the poles of the second order at most.

We would like to show that in this case the trivial monodromy property implies *D*-integrability, i.e., the existence of the intertwining operator (1). First of all, the potential must have the form

$$U(z)=\sum_{i=1}^{N}\frac{(\alpha_i,\alpha_i)A_i}{((\alpha_i,z)+c_i)^2} \tag{16}$$

due to the following result:

Theorem 3: *The regular singularities of the matrix potential of any D-integrable Schrödinger operator L are located on a union of nonisotropic hyperplanes. If such a potential is rational and decaying at infinity it should have a form (16).*

The proof essentially repeats the arguments of the scalar case investigated in Refs. 1,2. The coefficient (α_i,α_i) is written at the numerator of the expression (16) for the convenience, as this makes the matrices A_i independent on the choice of the equation of the corresponding hyperplane.

Let us assume now that the operator L with the potential (16) has local trivial monodromy around all the hyperplanes Π_i: $(\alpha_i,z)+c_i=0$. We will say in this case that L has *trivial monodromy*. The local trivial monodromy conditions (14) around all the hyperplanes form a highly-overdetermined algebraic system on the configuration of the hyperplanes with prescribed matrices A_i. We will call this system *matrix locus equations*.

Theorem 4: *Let L be a matrix Schrödinger operator (9) with a rational potential (16) satisfying the matrix locus equations. Then L is D-integrable.*

Proof: From the theorem 2 it follows that

$$A_s=\sum_{i=1}^{k_s} m_i^{(s)}(m_i^{(s)}+1)P_i^{(s)}, \quad 0\leqslant m_1^{(s)}<m_2^{(s)}<\cdots<m_{k_s}^{(s)}=M_s$$

with some projectors $P_i^{(s)}$: $P_i^{(s)}P_j^{(s)}=\delta_{ij}P_i^{(s)}$, $\Sigma_{i=1}^{k_s}P_i^{(s)}=I$.

Following the main idea of Ref. 3 let us introduce a linear space V consisting of the $d\times d$ matrix-valued functions $\Psi(z)$, $z\in\mathbf{C}^n$ which satisfy the conditions

(1) $\Psi(z)\Pi_{s=1}^N((\alpha_s,z)+c_s)^{M_s}$ is holomorphic in $\mathbf{C}^n$;

(2) the coefficients of the series expansion of $\Psi(z)$ at the vicinity of hyperplanes $(\alpha_s,z)+c_s=0$, $s=1,\ldots,N$ satisfy conditions (15) with $M=M_s$.

The crucial observation is that the matrix locus Eqs. (14) imply that the space V is invariant under L (cf. Refs. 3,4).

Let us consider the matrix function $\Psi_0=\Pi_{s=1}^N((\alpha_s,z)+c_s)^{M_s}e^{(k,z)}I$, where I is the identity matrix. Evidently, $\Psi_0\in V$ and, therefore, all the functions

$$\Psi_i=(L+k^2)^i\Psi_0, \quad i=1,2,\ldots$$

belong to V as well. These functions have the form

$$\Psi_i=\frac{P_i(k,z)e^{(k,z)}}{\Pi_{s=1}^N((\alpha_s,z)+c_s)^{M_s}},$$

where $P_i(k,z)$ are some matrix polynomials in k, z. Since

$$P_{i+1}=\phi\left(-\Delta-2\left(k,\frac{\partial}{\partial z}\right)+U(z)\right)\phi^{-1}P_i, \ \phi=\prod_{s=1}^N ((\alpha_s,z)+c_s)^{M_s},$$

the degrees of P_i in z are decreasing with i. So, there exists such j that $(L+k^2)\Psi_j=0$. It is easy to see that for $M=\Sigma_{s=1}^N M_s$,

$$\Psi_M=\left[(-2)^M M!\prod_{s=1}^N (\alpha_s,k)^{M_s}I+\cdots\right]e^{(k,z)}\neq 0, \tag{17}$$

where the dots mean the terms decaying while $z\to\infty$. We claim that $\Psi_{M+1}=(L+k^2)\Psi_M=0$. Indeed, assume that this is not true. Then for some $j>M$ we have

$$\Psi_{j+1}=(L+k^2)\Psi_j=0$$

with $\Psi_j\neq 0$. Since

$$P_{j+1}=\phi\left(-\Delta-2\left(k,\frac{\partial}{\partial z}\right)+U(z)\right)\phi^{-1}P_j=0$$

and P_j is polynomial in k its highest coefficient $P_j^{(0)}$ has to satisfy the condition

$$\left(k,\frac{\partial}{\partial z}\right)P_j^{(0)}=0.$$

One can show that this implies that $P_j^{(0)}$ must be polynomial in z (see Ref. 19, lemma 2.5). On the other hand one can see from (17) that Ψ_j for $j>M$ decays as $z\to\infty$. This contradiction means that $L\Psi_M=-k^2\Psi_M$. Presenting Ψ_M in the form $\Psi_M=\mathcal{D}e^{(k,z)}I$ for a proper matrix differential operator $\mathcal{D}(z,(\partial/\partial z))$ we have

$$L\Psi=L\mathcal{D}e^{(k,z)}I=-k^2\mathcal{D}e^{(k,z)}I=-\mathcal{D}k^2e^{(k,z)}I=\mathcal{D}L_0e^{(k,z)}I$$

and, therefore,

$$L\mathcal{D}=\mathcal{D}L_0.$$

The theorem is proved.

Remark: Notice that our proof gives an explicit formula for the intertwining operator

$$\mathcal{D}(e^{(k,z)}I)=(L+k^2)^M\left(\prod_{s=1}^{N}((\alpha_s,z)+c_s)^{M_s}e^{(k,z)}I\right).$$

Such a formula has been discovered in the scalar case by Berest.[7]

IV. GENERALIZED MATRIX CALOGERO–MOSER SYSTEM

Let us consider the following matrix Schrödinger operator:

$$L=-\Delta+\sum_{\alpha\in\mathcal{R}_+}\frac{m_\alpha(m_\alpha I-\hat{s}_\alpha)(\alpha,\alpha)}{(\alpha,z)^2}. \tag{18}$$

Here $\mathcal{R}$ is any Coxeter root system in $\mathbf{R}^n$, $\mathcal{R}_+$ is its positive part consisting of the normals to the reflection hyperplanes of the corresponding Coxeter group G, $m(\alpha)=m_\alpha$ is a G-invariant function on $\mathcal{A}$, $\hat{s}_\alpha$ stands for the reflection with respect to α in an arbitrary matrix representation π of the group G: $\hat{s}_\alpha=\pi(s_\alpha)$

For the trivial one-dimensional representation we have a scalar Schrödinger operator which is the well-known generalized Calogero–Moser operator related to the Coxeter group G (see Ref. 20). Thus (18) can be considered as a natural matrix generalization of these operators.

Cherednik[8] seems to be the first to consider such generalizations in the case when G is a Weyl group of any semisimple Lie algebra. He showed that the corresponding quantum system has n commuting quantum integrals and, therefore, it is integrable in a usual quantum mechanical sense.

Let us show that if all m_α are integers then the operator (18) is D-integrable. This implies the usual integrability and even more stronger property known as algebraic integrability (see theorem 7 below).

Let

$$s_\alpha(z)=z-2\frac{(\alpha,z)}{(\alpha,\alpha)}\alpha \tag{19}$$

be the orthogonal reflection with respect to the hyperplane $(\alpha,z)=0$. The matrix potential of the operator (18) has the following equivariance property for any $\alpha\in\mathcal{R}$:

$$\hat{s}_\alpha U(z)=U(s_\alpha(z))\hat{s}_\alpha. \tag{20}$$

This can be easily checked using G-invariance of m_α and the property

$$\hat{s}_\alpha \hat{s}_\beta = \hat{s}_{s_\alpha(\beta)} \hat{s}_\alpha .$$

From (20) it follows that the coefficients C_l of the Laurent expansion of the potential U near the hyperplane $(\alpha,z)=0$ satisfy the following relation:

$$\hat{s}_\alpha C_{2k} = C_{2k} \hat{s}_\alpha , \quad \hat{s}_\alpha C_{2k-1} + C_{2k-1} \hat{s}_\alpha = 0 \tag{21}$$

for any k. Comparing the formula (18) with (16) we see that the corresponding matrices A_i have two eigenvalues, $m_\alpha(m_\alpha+1)$ and $m_\alpha(m_\alpha-1)$. Using this it is easy to check that the relations (21) imply the local trivial monodromy conditions (14) and, therefore, D-integrability of the operator (18) due to the theorem 4. Thus, we have proved

Theorem 5: *The generalized matrix Calogero–Moser operator (18) with integer G-invariant* m_α *is D-integrable.*

In the scalar case the Calogero–Moser operator admits integrable deformations related to the non-Coxeter configurations of the hyperplanes.[13,14,5] It is interesting that these deformations admit a matrix generalization as well.

Let $\mathfrak{A}$ be a finite set of the hyperplanes Π_α in a complex Euclidean space $\mathbf{C}^n$ given by the equations $(\alpha,z)=0$, taken with some multiplicities $m_\alpha \in \mathbf{Z}_+$. Here $\alpha \in \mathcal{A}$, $\mathcal{A}$ is a finite set of noncollinear vectors. Consider the matrix Schrödinger operator

$$L = -\Delta + U(z)$$

with

$$U(z) = \sum_{\alpha \in \mathcal{A}} \frac{m_\alpha(m_\alpha - s_\alpha)(\alpha,\alpha)}{(\alpha,z)^2}, \tag{22}$$

where s_α is the $n\times n$ matrix of the reflection (19).

Theorem 6: *Operator L has trivial monodromy if and only if the following conditions for the configuration* $\mathcal{A}$ *hold for each* $\alpha \in \mathcal{A}$

$$A_j = \sum_{\beta \neq \alpha} \left. \frac{m_\beta(m_\beta+1)(\beta,\beta)(\alpha,\beta)^{2j-1}}{(\beta,z)^{2j+1}} \right|_{(\alpha,z)=0} \equiv 0, \quad j=1,2,\ldots,m_\alpha , \tag{23}$$

$$B_j = \sum_{\beta \neq \alpha} \left. \frac{m_\beta(\alpha,\beta)^{2j-1}}{(\beta,z)^{2j-1}} \right|_{(\alpha,z)=0} \equiv 0, \quad j=1,2,\ldots,m_\alpha . \tag{24}$$

Proof: Let us consider first the case $m_\alpha=1$. Then we have two locus conditions (see (14)) for L,

$$C_0 C_{-2} = C_{-2} C_0 \quad \text{or} \quad C_0 s_\alpha = s_\alpha C_0 \tag{25}$$

and

$$(C_1 \alpha, \alpha) = 0. \tag{26}$$

From (22) we can calculate

$$C_j = \sum_{\beta \neq \alpha} \left. \frac{m_\beta(m_\beta - s_\beta)(\beta,\beta)(\alpha,\beta)^j}{(\beta,z)^{2+j}} \right|_{(\alpha,z)=0} ,$$

and condition (25) reduces to

$$\sum_{\beta\neq\alpha}\frac{m_\beta(\beta,\beta)(s_\beta s_\alpha - s_\alpha s_\beta)}{(\beta,z)^2}\bigg|_{(\alpha,z)=0}\equiv 0. \tag{27}$$

Let us choose some $\gamma\neq\alpha$ and consider the subsum in (27) corresponding to the two-dimensional plane $\langle\alpha,\gamma\rangle$. Since s_β acts trivially on the orthogonal complement to the plane $\pi=\langle\alpha,\gamma\rangle$ for any $\beta\in\pi$ we may assume that $\alpha=(1,0),\ \beta=(\cos\phi_\beta,\sin\phi_\beta)$. Then,

$$s_\alpha=\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix},\quad s_\beta=\begin{pmatrix}-\cos 2\phi_\beta & -\sin 2\phi_\beta\\ -\sin 2\phi_\beta & \cos 2\phi_\beta\end{pmatrix}$$

and

$$s_\beta s_\alpha=\begin{pmatrix}\cos 2\phi_\beta & \sin 2\phi_\beta\\ -\sin 2\phi_\beta & \cos 2\phi_\beta\end{pmatrix},\quad s_\alpha s_\beta=(s_\beta s_\alpha)^{-1}.$$

Since $(\beta,z)|_{(\alpha,z)=0}=\beta_2 z_2=z_2\sin\phi_\beta$ we have to check that

$$\sum_{\beta\in\langle\alpha,\gamma\rangle,\beta\neq\alpha}\frac{\sin 2\phi_\beta}{\sin^2\phi_\beta}=0.$$

But it easily follows from the identity (24) $B_1\equiv 0$.

The second locus condition (26) reduces to

$$\sum_{\beta\neq\alpha}\frac{m_\beta(m_\beta+\cos 2\phi_\beta)\cos\phi_\beta}{\sin^3\phi_\beta}=0.$$

This is equivalent to the combination $A_1-2B_1=0$ of the identities (23)–(24).

Now let us consider the case $m_\alpha>1$. Locus equations (14) take the form

$$C_j s_\alpha=(-1)^j s_\alpha C_j,\quad j=0,1,\dots,2m_\alpha-2, \tag{28}$$

$$(C_{2m_\alpha-1}\alpha,\alpha)=0. \tag{29}$$

As above, everything reduces to the two-dimensional case and we will use the same notations. The relations (28) reduce to

$$\sum_{\beta\neq\alpha}\frac{m_\beta\cos^j\phi_\beta\sin 2\phi_\beta}{\sin^{2+j}\phi_\beta}=0, \tag{30}$$

for $j=2l$, $l=0,1,\dots,m_\alpha-1$, and

$$\sum_{\beta\neq\alpha}\frac{2\,s_\alpha m_\beta^2\cos^j\phi_\beta-(s_\alpha s_\beta+s_\beta s_\alpha)m_\beta\cos^j\phi_\beta}{\sin^{2+j}\phi_\beta}=0, \tag{31}$$

for $j=2l-1,\ l=1,\dots,m_\alpha-1$. Relation (30) is equivalent to

$$\sum_{\beta\neq\alpha}\frac{m_\beta\cos^{2l+1}\phi_\beta}{\sin^{2l+1}\phi_\beta}=0,$$

which coincide with $B_{l+1}\equiv 0$. Condition (31) is equivalent to

$$\begin{cases} \sum_{\beta\neq\alpha} \frac{m_\beta^2 \cos^{2l-1}\phi_\beta}{\sin^{2l+1}\phi_\beta}=0, \quad l=1,\dots,m_\alpha-1 \\ \sum_{\beta\neq\alpha} \frac{m_\beta \cos^{2l-1}\phi_\beta \cos 2\phi_\beta}{\sin^{2l+1}\phi_\beta}=0, \quad l=1,\dots,m_\alpha-1. \end{cases}$$

The first part of the last equations due to the identity $A_l\equiv 0$ reduces to

$$\sum_{\beta\neq\alpha} \frac{m_\beta \cos^{2l-1}\phi_\beta}{\sin^{2l+1}\phi_\beta}=0,$$

the left-hand side of which equals B_l+B_{l+1}. The second one is equivalent to $B_{l+1}-B_l\equiv 0$ and is satisfied for $l=1,\dots,m_\alpha-1$.

Finally, the condition $(C_{2m_\alpha-1}\alpha,\alpha)=0$ is equivalent to

$$\sum_{\beta\neq\alpha} \frac{m_\beta^2 \cos^{2m_\alpha-1}\phi_\beta + m_\beta \cos 2\phi_\beta \cos^{2m_\alpha-1}\phi_\beta}{\sin^{2m_\alpha+1}\phi_\beta}=0$$

or, using $A_{m_\alpha}\equiv 0$, to

$$\sum_{\beta\neq\alpha} \frac{m_\beta \cos^{2m_\alpha-1}\phi_\beta(\cos 2\phi_\beta-1)}{\sin^{2m_\alpha+1}\phi_\beta}=0,$$

which coincides with $B_{m_\alpha}\equiv 0$. Theorem 6 is proved.

Remark: It is interesting to note that conditions (23) and (24) are equivalent to the existence of the so-called Baker–Akhiezer function in ''old axiomatics'' (see Refs. 16 and 5). Indeed, the A-conditions (23) coincide with the locus equations for the scalar case and, therefore, guarantee the existence of the Baker–Akhiezer function in ''new axiomatics.''[5] The B-conditions (24) mean that the function $\phi=\Pi_{\beta\neq\alpha}(\beta,z)^{m_\beta}$ has zero odd normal derivatives at the hyperplane Π_α,

$$\left(\frac{\partial}{\partial\alpha}\right)^{2j-1} \prod (\beta,z)^{m_\beta}\Big|_{(\alpha,z)=0}=0,$$

which together with the new axiomatics provide the old one (see Sec. 1 in Ref. 5).

In particular, conditions (23) and (24) are satisfied for the following non-Coxeter configurations $A_n(m)$ and $C_{n+1}(m,l)$ discovered in Refs. 13,14,5.

Configuration $A_n(m)$ consists of the following vectors in $\mathbf{R}^{n+1}$: e_i-e_j with multiplicity m $(1\leqslant i<j\leqslant n)$ and $e_i-\sqrt{m}e_{n+1}$ with multiplicity 1 $(i=1,\dots,n)$ (for $m=1$ we have the root system A_n). Parameter m is also allowed to be negative. Then one should consider vectors e_i-e_j with the multiplicity $-1-m$. In the last case we have a complex configuration in $\mathbf{C}^{n+1}$.

Configuration $C_{n+1}(m,l)$ consists of the following set of vectors in $\mathbf{R}^{n+1}$:

$$C_{n+1}(m,l)=\begin{cases} e_i\pm e_j \quad \text{with multiplicity } k \\ 2e_i \quad \text{with multiplicity } m \\ 2\sqrt{k}e_{n+1} \quad \text{with multiplicity } l \\ e_i\pm\sqrt{k}e_{n+1} \quad \text{with multiplicity } 1, \end{cases}$$

where l and m are integer parameters such that $k=(2m+1)/(2l+1)\in\mathbf{Z}$, $1\leqslant i<j\leqslant n$. If $l=m=k=1$ the system $C_{n+1}(m,l)$ coincides with the classical root system C_{n+1}. As before, the parameters k,m,l may be negative, in that case the corresponding multiplicities should be $-1-k$, $-1-m$ or $-1-l$, respectively.

Corollary: The matrix Schrödinger operators with potentials (22) corresponding to the configurations $A_n(m)$ and $C_{n+1}(m,l)$ are D-integrable.

Let us prove now that in the considered cases D-integrability implies usual quantum integrability and even more—so-called algebraic integrability.

We say that a matrix Schrödinger operator L in $\mathbf{R}^n$ is *integrable* if there exists n pairwise commuting matrix differential operators $L_1=L, L_2,\ldots,L_n$ having the algebraically independent constant scalar highest symbols $P_j(k),\ j=1,\ldots,n$. If there exists one more commuting matrix differential operator L_{n+1} with the highest constant scalar symbol $P_{n+1}(k)$ such that $P_{n+1}(k)$ takes different values on the solutions of the system $P_i(k)=c_i\ (i=1,\ldots,n)$ for generic $c_1,\ldots,c_n$ the operator L is called *algebraically integrable* (see Refs. 21,15,9).

Let us assume that the matrix potential U is symmetric, $U=U^*$.

Theorem 7: *Any D-integrable matrix Schrödinger operator L with a rational symmetric potential (16) is algebraically integrable.*

Proof: We follow here the idea of the paper.[22] Let A^* denote a formal conjugate to a matrix differential operator A then taking a conjugation of the relation $L\mathcal{D}=\mathcal{D}L_0$ we have $\mathcal{D}^*L^*=L_0^*\mathcal{D}^*=L_0\mathcal{D}^*$. If $U=U^*$ then $L=L^*$ and we obtain $\mathcal{D}^*L=L_0\mathcal{D}^*$. Now define the operators $L_1=L, L_{1+i}=\mathcal{D}\partial_i\mathcal{D}^*\ (i=1,\ldots,n)$. We claim that they are pairwise commuting. Indeed, $LL_{1+i}=L\mathcal{D}\partial_i\mathcal{D}^*=\mathcal{D}L_0\partial_i\mathcal{D}^*=\mathcal{D}\partial_iL_0\mathcal{D}^*=\mathcal{D}\partial_i\mathcal{D}^*L=L_{1+i}L$, so $[L_1,L_k]=0$ for all $k=2,\ldots,n+1$. Consider now the commutator $[L_l,L_k]$, $l>1$. From the previous relations and Jacobi identity it follows that $[[L_k,L_l],L]=0$. Berezin's lemma (see lemma 2.5 in Ref. 19) says that the highest symbol of $[L_k,L_l]$ has to be polynomial in z, but from the definition of L_k and the construction of $\mathcal{D}$ it follows that it decays as $z\to\infty$. This means that $[L_k,L_l]=0$ for any $k,l=1,\ldots,n+1$. One can check that the highest symbols of L_k satisfies the property demanded at the definition of algebraic integrability. The theorem is proved.

Remark: The statement of the theorem seems to be true without the assumption of the symmetry of the potential.

V. TWO-DIMENSIONAL CASE

Let us consider the matrix locus configurations on the plane in the case when all the lines pass through the origin. In the scalar case essentially all such configurations have been described by Berest and Lutsenko[23] (see Ref. 5 for details).

In the matrix case in the polar coordinates (r,ϕ) the corresponding potential U has a form

$$U(r,\phi)=\frac{1}{r^2}V(\phi), \tag{32}$$

where

$$V(\phi)=\sum_{i=1}^{k}\frac{A_i}{\sin^2(\phi-\phi_i)}. \tag{33}$$

Here A_i are some matrices, ϕ_i are the angles corresponding to the lines of configurations. Strictly speaking this is true only on the real plane $\mathbf{R}^2$ but this can be easily generalized to $\mathbf{C}^2$ (see Ref. 5).

Proposition 1: Two-dimensional matrix Schrödinger operator

$$L=-\Delta+U \tag{34}$$

with the potential (32) has trivial monodromy if and only if the same is true for the one-dimensional Schrödinger operator

$$\mathcal{L}=-\frac{d^2}{d\phi^2}+V(\phi) \tag{35}$$

with trigonometric potential (33).

The proof is a simple check that the local trivial monodromy conditions for these two operators are equivalent.

Proposition 2: If the operator (35) is D-integrable then the same is true for the two-dimensional Schrödinger operator (34).

Indeed, in the one-dimensional case D-integrability implies the trivial monodromy for the operator (35) and, therefore, for the two-dimensional operator (34). According to the theorem 4 this guarantees the D-integrability of (34).

In dimension 1 D-integrability is equivalent to the fact that the operator $\mathcal{L}$ (35) is the result of so-called matrix Darboux transformation applied to $\mathcal{L}_0 = -(d^2/d\phi^2)$ (see, e.g., Ref. 4). All such operators can be described using the notion of quasideterminants introduced by Gelfand and Retakh (see Ref. 24).

Let k be the order of the intertwining operator $\mathcal{D}$. Consider any solution Φ of the simple matrix differential equation $-(d^2/d\phi^2)\Phi = \Phi C$ where Φ is $d \times kd$ matrix, C is any diagonalizable $kd \times kd$ matrix with the eigenvalues of the form $\lambda = p^2$ with $p \in \mathbf{Z}$. Let $\Phi = (\Psi_1, \ldots, \Psi_k)$ where Ψ_i are the corresponding $d \times d$ matrices. Then the intertwining operator $\mathcal{D}$ can be written as quasideterminant

$$\mathcal{D}(\Psi) = |W(\Psi_1, \ldots, \Psi_k, \Psi)|_{k+1,k+1},$$

where

$$W(\Psi_1, \ldots, \Psi_k, \Psi) = \begin{pmatrix} \Psi_1 & \ldots & \Psi_k & \Psi \\ \vdots & \ddots & \vdots & \vdots \\ \Psi_1^{(k-1)} & \ldots & \Psi_k^{(k-1)} & \Psi^{(k-1)} \\ \Psi_1^{(k)} & \ldots & \Psi_k^{(k)} & \Psi^{(k)} \end{pmatrix},$$

(see Ref. 4 for the details). The potential V has a form

$$V = 2a_1'(\phi), \tag{36}$$

where $a_1(\phi)$ is the first matrix coefficient of $\mathcal{D}$,

$$\mathcal{D} = D^k + a_1(\phi)D^{k-1} + \cdots + a_k(\phi),\ D = \frac{d}{d\phi}.$$

Under some assumptions on Φ one can give more explicit formula for the potential (see Ref. 4).

Theorem 8: *Two-dimensional matrix Schrödinger operator (34) with the potential of the form (32) related to any result (36) of the one-dimensional matrix Darboux transformation described above has trivial monodromy and therefore D-integrable. Conversely, for any D-integrable operator (34) the corresponding one-dimensional operator (35) is related to the operator* $\mathcal{L}_0 = -d^2/d\phi^2$ *by a matrix Darboux transformation.*

The proof of the inverse statement follows from

Lemma 2: Any one-dimensional Schrödinger operator (34) with trigonometric potential (33) which satisfies local trivial monodromy conditions at all the singularities is D-integrable.

Proof of the lemma essentially combines the arguments of the matrix rational case (see Ref. 4 or theorem 3 above) and the scalar trigonometric case investigated in Ref. 3.

It is worthy to derive the explicit formula for such operators in the simplest case of three lines with prescribed 2×2 matrices with the eigenvalues 0 and 2. In this case it seems to be more suitable to use matrix locus equations rather than Darboux transformation. Thus, let $V(\phi)$ be of the form

$$V(\phi)=\frac{2P_\alpha}{\sin^2(\phi-\alpha)}+\frac{2P_\beta}{\sin^2(\phi-\beta)}+\frac{2P_\gamma}{\sin^2(\phi-\gamma)}, \tag{37}$$

where $\phi,\alpha,\beta,\gamma\in\mathbf{C}$; $P_\alpha,P_\beta,P_\gamma$ are some projector matrices of rank 1. According to theorem 1 operator (35) has trivial monodromy if its Laurent expansion at pole $\phi=\phi_0$,

$$V(\phi)=\frac{C_{-2}}{(\phi-\phi_0)^2}+\frac{C_{-1}}{(\phi-\phi_0)}+C_0+C_1(\phi-\phi_0)+\cdots$$

satisfies the conditions

$$C_{-1}=0,$$

$$[C_{-2},C_0]=0, \tag{38}$$

$$C_{-2}C_1C_{-2}=0. \tag{39}$$

Conditions $C_{-1}=0$ are, obviously, fulfilled. Expanding $V(\phi)$ near $\phi=\alpha$,

$$V(\phi)=\frac{2P_\alpha}{(\phi-\alpha)^2}+\left(\frac{2P_\beta}{\sin^2(\alpha-\beta)}+\frac{2P_\gamma}{\sin^2(\alpha-\gamma)}+\frac{2P_\alpha}{3}\right)$$
$$+\left(\frac{-4P_\beta\cos(\alpha-\beta)}{\sin^3(\alpha-\beta)}+\frac{-4P_\gamma\cos(\alpha-\gamma)}{\sin^3(\alpha-\gamma)}\right)(\phi-\alpha)+\ \ldots,$$

and then near $\phi=\beta$ and $\phi=\gamma$ we get the following system of the equations using (38):

$$\left[P_\alpha,\frac{P_\beta}{\sin^2(\alpha-\beta)}+\frac{P_\gamma}{\sin^2(\alpha-\gamma)}\right]=0, \tag{40}$$

$$\left[P_\beta,\frac{P_\alpha}{\sin^2(\beta-\alpha)}+\frac{P_\gamma}{\sin^2(\beta-\gamma)}\right]=0, \tag{41}$$

$$\left[P_\gamma,\frac{P_\alpha}{\sin^2(\gamma-\alpha)}+\frac{P_\gamma}{\sin^2(\gamma-\beta)}\right]=0. \tag{42}$$

It is easy to see that (42) follows from (40) and (41). Conditions (39) give

$$P_\alpha\left(\frac{P_\beta\cos(\alpha-\beta)}{\sin^3(\alpha-\beta)}+\frac{P_\gamma\cos(\alpha-\gamma)}{\sin^3(\alpha-\gamma)}\right)P_\alpha=0, \tag{43}$$

$$P_\beta\left(\frac{P_\alpha\cos(\beta-\alpha)}{\sin^3(\beta-\alpha)}+\frac{P_\gamma\cos(\beta-\gamma)}{\sin^3(\beta-\gamma)}\right)P_\beta=0, \tag{44}$$

$$P_\gamma\left(\frac{P_\alpha\cos(\gamma-\alpha)}{\sin^3(\gamma-\alpha)}+\frac{P_\beta\cos(\gamma-\beta)}{\sin^3(\gamma-\beta)}\right)P_\gamma=0. \tag{45}$$

Solving of system of Eqs. (40)–(45) and making a suitable transformation

$$V(\phi)\rightarrow C\,V(\phi)C^{-1} \tag{46}$$

we arrive at the formula

$$P_\alpha = \frac{1}{\sin(\alpha-\beta)\sin(\alpha-\gamma)} \cdot \xi_\alpha^t \eta_\alpha, \tag{47}$$

where

$$\xi_\alpha = (-\cos\alpha, \sin\alpha), \eta_\alpha = (s(\alpha;\beta,\gamma), c(\alpha;\beta,\gamma)),$$

$$s(\alpha;\beta,\gamma) = \sin^2\alpha \cos(\beta+\gamma-\alpha) - \cos\alpha \sin\beta \sin\gamma,$$

$$c(\alpha;\beta,\gamma) = \cos^2\alpha \sin(\beta+\gamma-\alpha) - \sin\alpha \cos\beta \cos\gamma.$$

Projectors P_β and P_γ can be obtained by corresponding permutations of α, β, and γ.

Theorem 9: *Any three lines on the plane with prescribed matrices P_α, P_β, and P_γ (47) where α, β, γ are the corresponding angles form a matrix locus configuration. Modulo (46) this describes all three lines 2×2 matrix locus configurations with prescribed matrices having the eigenvalues* 0 *and* 2.

It is interesting to note that the potential (37), (47) is symmetric if and only if $\alpha=\vartheta$, $\beta=\vartheta+(\pi/3)$ and $\gamma=\vartheta+(2\pi/3)$ for some ϑ which corresponds to the matrix Calogero–Moser system (18) related to A_2 root system.

VI. CONCLUDING REMARKS

Similarly to the scalar case[5] one can introduce the notion of the multidimensional matrix Baker–Akhiezer function. This would lead to the proof of the algebraic integrability for the corresponding Schrödinger operators. The bispectral properties of these functions and the relations to the Huygens' Principle we are planning to discuss in a separate paper.

ACKNOWLEDGMENTS

This work was partially supported by INTAS (Grant No. 96-0770). O.A.C. was supported also by the Royal Society postdoctoral fellowship during 1998, which is highly acknowledged. He is grateful to Loughborough University, UK, for the hospitality during the period this work was completed. V.M.G. has been supported by ORS Award Scheme (UK) and Loughborough University. A.P.V. is grateful to Yu. Berest, P. Bracken, and N. Kamran for the fruitful discussions. We are grateful also to M. Feigin for the critical remarks.

[1] Yu. Yu. Berest and A. P. Veselov, ''On the singularities of the potentials of exactly solvable Schrödinger operators and Hadamard's problem,'' Russ. Math. Surveys **53**, n1, 208–209 (1998).
[2] Yu. Yu. Berest and A. P. Veselov, ''On the structure of singularities of integrable Schrödinger operators,'' Lett. Math. Phys. (submitted).
[3] O. A. Chalykh, ''Darboux transformations for multidimensional Schrödinger operators,'' Russ. Math. Surveys **53**, n2, 167–168 (1998).
[4] V. M. Goncharenko and A. P. Veselov, ''Monodromy of the matrix Schrödinger equations and Darboux transformations,'' J. Phys. A **31**, 5315–5326 (1998).
[5] O. A. Chalykh, M. V. Feigin, and A. P. Veselov, ''Multidimensional Baker–Akhiezer Functions and Huygens' Principle,'' Commun. Math. Phys. (accepted).
[6] J. J. Duistermaat and F. A. Grünbaum, ''Differential equations in the spectral parameter,'' Commun. Math. Phys. **103**, 177–240 (1986).
[7] Yu. Yu. Berest, ''Huygens' Principle and the bispectral problem,'' CRM Proceedings and Lecture Notes **14**, 11–30 (1998).
[8] I. Cherednik, ''Integration of quantum many-body problems by affine Knizhnik–Zamolodchikov equations,'' Adv. Math. **106**, 65–95 (1994).
[9] P. Etingof, ''Quantum integrable systems and representations of Lie algebras,'' J. Math. Phys. **36**, 2636–2651 (1995).
[10] P. Etingof and K. Styrkas, ''Algebraic integrability of Schrödinger operators and representations of Lie algebras,'' Compositio Mathematica **98**, 91–112 (1995).
[11] A. González–López, and N. Kamran, ''The multidimensional Darboux transformation,'' J. Geom. Phys. **26**, 202–226 (1998).

[12] P. Bracken and N. Kamran, ''Matrix Calogero-Moser Hamiltonians and the multidimensional Darboux transformation'' Preprint, 1998.
[13] A. P. Veselov, M. V. Feigin, and O. A. Chalykh, ''New integrable deformations of the quantum Calogero–Moser problem,'' Usp. Mat. Nauk **51**, n3, 185–186 (1996).
[14] O. A. Chalykh, M. V. Feigin, and A. P. Veselov, ''New integrable generalizations of the Calogero–Moser quantum problem,'' J. Math. Phys. **39**, 695–703 (1998).
[15] O. A. Chalykh and A. P. Veselov, ''Commutative rings of partial differential operators and Lie alrebras,'' Commun. Math. Phys. **126**, 597–611 (1990).
[16] A. P. Veselov, K. L. Styrkas, and O. A. Chalykh, ''Algebraic integrability for Schrödinger equation and finite reflection groups,'' Theor. Math. Phys. **94**, n2, 253–275 (1993).
[17] J. Gibbons and T. Hermsen, ''A generalization of the Calogero–Moser system,'' Physica D **11**, 337–348 (1984).
[18] I. Krichever, O. Babelon, E. Billey, and M. Talon, ''Spin generalization of the Calogero–Moser system and the matrix KP equation,'' Am. Math. Soc. Trans. **170**, 83–119 (1995).
[19] F. A. Berezin, ''Laplace operators on semisimple Lie groups,'' Proc. Moscow Math. Soc. **6**, 371–463 (1957).
[20] M. A. Olshanetsky and A. M. Perelomov, ''Quantum integrable systems related to Lie algebras,'' Phys. Rep. **94**, 313–404 (1983).
[21] I. M. Krichever, ''Methods of algebraic geometry in the theory of nonlinear equations,'' Russ. Math. Surveys **32**, n6, 185–213 (1977).
[22] O. A. Chalykh, ''Additional integrals of the generalized quantum Calogero–Moser problem,'' Theor. Math. Phys. **109**, 1269–1273 (1996).
[23] Yu. Yu. Berest and I. M. Lutsenko, ''Huygens' principle in Minkowski spaces and soliton solutions of the Korteweg–de Vries equation,'' Commun. Math. Phys. **190**, 113–132 (1997).
[24] I. Gelfand and V. Retakh, ''Determinants of matrices over noncommutative rings,'' Funct. Anal. Appl. **25**, n2, 91–102 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Evaluation of multiloop diagrams via lightcone integration

Y. J. Feng[a] and C. S. Lam[b]
Department of Physics, McGill University, Montreal, Quebec H3A 2T8, Canada

(Received 27 August 1998; accepted for publication 18 December 1998)

We present a systematic method to determine the dominant regions of internal momenta contributing to any two-body high-energy near-forward scattering diagram. Such a knowledge is used to evaluate leading high-energy dependences of loop diagrams. It also gives a good idea where dominant multiparticle cross sections occur. © *1999 American Institute of Physics.* [S0022-2488(99)00811-7]

I. INTRODUCTION

It is difficult to compute high-energy ($\sqrt{s}$) scattering amplitudes at small momentum transfers ($\sqrt{-t}$), even assuming the coupling constant g^2 to be small. This is so because each loop of a Feynman diagram is capable of producing a $\ln s$ factor, thus changing the effective expansion parameter from g^2 to $g^2 \ln s$. Even though the former may be small, the latter can become quite sizable at high energies, necessitating diagrams of high orders to be included. Such is for example the case when total cross section is computed in the framework of QCD.[1]

Usually such a daunting task of computed diagrams of many loops may be contemplated only in the leading-log approximation (LLA), though there are exceptions, especially for sets of diagrams with regular structures.[2] In LLA, only terms of the highest power of $\ln s$ are kept at each perturbative order, but even so the computation is far from being simple. For low-order diagrams, or diagrams with highly regular structures, the computation has been carried out and the results are well known.[3–6] However, for complicated diagrams, a systematic procedure to find even the leading-log contribution seems to be lacking. We shall discuss a method in the present paper. The available QCD result, via the exchange of the BFKL Pomeron,[5] violates the Froissart bound and needs to be improved.[7–9] Other diagrams must be included to restore unitarity so it would be useful to have a way to find out how the other diagrams behave at high energies. This can be achieved if the regions of internal momenta dominating the Feynman amplitude can be located, for then one simply integrates around them to obtain the LLA result.

For quark–quark scattering via the exchange of gluon ladders, the dominant region is known to be the multi-Regge region,[4,5] where gluons produced in the intermediate states are strongly ordered in rapidity, and the gluons being exchanged are dominantly spacelike. What we would like to discuss in the present paper is a general way to find such dominant regions for any diagram, and its associated high-energy dependence in LLA. We shall carry out the study for Feynman diagrams and for non-Abelian cut diagrams,[10–12] both because they are more general, and because there is already a considerable body of literature on the dispersion theoretic techniques.[5,6]

Such calculations of elastic amplitudes, besides giving the energy-dependence of total cross-sections via the optical theorem,[1] also tell us the kinematical regions where the dominant inelastic cross-sections come from, for via unitarity these are intimately related to the dominant internal momenta of the elastic amplitude. This knowledge would be of direct phenomenological interest as well.

The methods developed in this paper should also be useful in the study of two-dimensional effective QCD Lagrangians at high-energies.[13,14] A prerequisite needed to arrive at a reliable effective Lagrangian is to know which are the heavy modes that can be discarded, and which of

[a]Electronic mail: feng@physics.mcgill.ca
[b]Electronic mail: lam@physics.mcgill.ca

them must be integrated out to yield a new vertex in the effective Lagrangian. In perturbative language this is equivalent to finding the important regions of internal momenta around which to integrate. All others may simply be discarded.

In the rest of this section we shall describe what our method is based on, and provide a brief summary of the results.

At high energies it is convenient to use lightcone coordinates, $k_\pm = k^0 \pm k^3$. The components of a four-vector k^μ can then be written as $(k_+, k_-, k_\perp)$, and the loop integration expressed as $d^4k = dk_+ dk_- d^2k_\perp/2$. In the center-of-mass system, the momenta of the two incoming particles, with masses neglected, can be taken to be $p_2 = (\sqrt{s},0,0)$ and $p_1 = (0,\sqrt{s},0)$. The momentum transfer $\sqrt{-t}$ as well as all other transverse momenta $k_\perp$ are taken to be of order 1 as $s\to\infty$, so it is only the dominant regions in k_+ and k_- for every loop momentum k that have to be determined.

These regions are determined in the following way. First, observe that the inverse of the internal propagators are bilinear in the ''+'' and the ''−'' components of their line momenta, so the propagators give rise to simple poles in the ''+'' (or the ''−'') momenta which enable integrations in those variables to be carried out exactly by residue calculus.[4] Once this is done the locations of the ''+'' momenta are determined by the locations of the contributing poles and the ''−'' momenta. The ''−'' momenta are then fixed to be in the regions yielding the leading-log contributions to the amplitude.

We shall be able to do this both for Feynman diagrams and ''non-Abelian cut diagrams.''[10–12] Feynman diagrams are fundamental, but they often have the undesirable property that the LLA contributions of individual diagrams get cancelled in the sum.[4] To the extent that the usual technology only allows LLA to be computed, this cancellation is disastrous because it leaves no viable means to compute the leading high-energy behavior of the sum. Non-abelian cut diagrams are designed to combat this problem. The cancellation is actually a result of the destructive interference between the virtual gluons being exchanged. The non-Abelian cut diagrams allow the destructive interferences to take place before high energy approximations are taken. In this way the LLA contribution to the non-abelian cut diagrams will reflect directly the leading contributions to the sum. No further cancellation will occur.

In Sec. II we will review the *flow diagram* method of Cheng and Wu[4] for carrying out the ''+'' component integrations by residue calculus. This method is very effective in locating the poles and the dominant integration regions for relatively simple diagrams. In complicated diagrams one encounters the problem of *flow reversal* which will be discussed in Sec. III. This problem prevents a simple reading of the contributing poles directly from the flow diagrams. Nevertheless the locations of these poles can still be computed, but the complexity of computation grows quite fast with the number of loops of the diagram. This difficulty is then overcome by a ''path'' method to be discussed in Sec. IV. With this method the contributing poles can be located and the ''+'' momenta determined. What remains is to find the dominant ''−'' momenta that give rise to the LLA contribution. The recipe for doing so will be discussed in Sec. V. Finally, in Sec. VI, a number of examples are given to illustrate the procedure.

II. FLOW DIAGRAMS

Consider a diagram with n internal lines and l loops, whose line and loop momenta are denoted by $q_i (1 \leqslant i \leqslant n)$ and $k_b (1 \leqslant b \leqslant l)$, respectively. In lightcone coordinates, the denominator of the propagator for a line with momentum q is $d(q) = (q^2 - m^2 + i\epsilon) = (q_+ q_- - q_\perp^2 - m^2 + i\epsilon) \equiv (q_+ q_- - a + i\epsilon)$. These n propagators collectively define a set of poles for the integration variables k_{b+}, thus enabling these integrations to be performed with the help of residue calculus. To carry out this program we must identify, for each k_{b+}, which are the poles in the upper-half-plane and which are the poles in the lower-half-plane, for only the poles in one half-plane will be picked up by a contour integration. Their locations in turn depends on the sign of q_{i-}, the choice of loop momenta, as well as the order the k_{b+} integrations are carried out. With so many variables the problem is very complex indeed. *Flow diagram* was invented[4] to keep track of things and to determine the location of poles. We shall review its essence[4] in this section, and point out in the next section some of the complications hitherto overlooked. This complication makes it complicated to apply it to multiloop diagrams. In Sec. IV we shall propose a ''*path method*'' to bypass these complications, and enables the evaluation of the ''+'' integration to be carried out in a simple manner.

A flow diagram is a Feynman diagram (or a non-Abelian cut diagram) with arrows attached to each of its internal lines to indicate the direction of q_{i-}. Since the signs of the q_-'s vary over the integration region, generally more than one flow diagram is present for each Feynman or nonabelian cut diagram. Nevertheless, for a diagram with n internal lines, there are far fewer than 2^n flow diagrams that one might otherwise expect, for two reasons. First, momentum conservation forbids the arrows from a common vertex to point all inwards or all outwards. Second, for reasons to be explained below, one can reject flow diagrams in which arrows around any closed loop all point in the same (clockwise or counterclockwise) direction. With these two requirements, it is easy to see that the one-loop box diagram has only one flow diagram, rather than $2^4=16$.

In a flow diagram the signs of q_{i-} along the arrows are all positive, by definition. This allows the positions of the poles be located and the ''+'' integrations to be carried out, once the independent loops and their order of integrations are chosen. We shall now proceed to see how this is accomplished for the first integration, say k_{1+}.

k_{1+} flows through the lines of this first loop either in a clockwise or a counterclockwise direction. Its coefficient in $d(q_i)$ is $\pm q_{i-}$, depending on whether this direction is the same as the arrow or opposite. The pole of $1/d(q_i)$ in k_{1+} has an imaginary part $\mp i\epsilon/q_{i-}$, with all $q_{i-}>0$ by definition. Hence the lines with arrows pointing one way (clockwise or counterclockwise) have poles all in one half-plane, and those with arrows pointing the opposite way have poles in the other half-plane. Which is which does not matter because we can always define the loop momentum by reversing its sign.

It is now easy to understand the assertion made earlier in the section, that flow diagrams containing a closed loop with flow arrows all pointing in the same direction may be rejected. Taking this loop as the first loop of integration, this would imply all poles to be in the same half-plane. By closing the integration contour in the other half-plane, we get a zero integral so such a flow diagram can be ignored.

Sometimes pole locations for *subsequent integrations* can be located in the same way, i.e., by the direction of arrows in the flow diagram. In fact the explicit examples shown in Ref. 4 all seem to be of this type.

However, it is not guaranteed that pole locations for subsequent integrations can be located this way, as we shall now see. This is the complication mentioned in the section.

To make it easier to describe things later on, we shall call two momenta pointing in the same (opposite) direction around a loop to be *parallel* (*antiparallel*) in that loop.

III. FLOW REVERSAL

Suppose there are n_1 poles picked up by the k_{1+} integration, each contributing to a term in the integral. As a result of the integration, k_{1+} acquires an imaginary part $\mp i\epsilon/q_{i-}$ from the ith pole. The sign is $-/+$ if k_{1+} and q_{i-} are parallel/antiparallel. This imaginary part in turn imparts an imaginary part on every q_{j+} of the first loop, which is why the location of poles for the second and subsequent integrations may be altered. For simplicity, we shall assume from now on that ϵ is finite and positive, and has a common value in all the propagators.

This imaginary part of k_{1+} affects the location of poles in subsequent integrations only for lines j lying in loop 1. In that case, the imaginary part of $d(q_j)$ is changed from $i\epsilon$ to $i\epsilon(\mp q_{j-}/q_{i-}+1)$, with sign $-/+$ when lines j and i are parallel/antiparallel in loop 1. Unless the sign is $-$ and $q_j>q_i$, the imaginary part of $d(q_j)$ remains positive and the location of pole j in subsequent integrations is once again determined solely by the direction of its arrow around the integration loop, viz., it can be determined directly from the flow diagram. However, if lines j and i are parallel in the first loop, and that $q_j>q_i$, then the sign of the imaginary part of $d(q_j)$ becomes negative, and the pole location (upper or lower plane) will now be opposite to naive

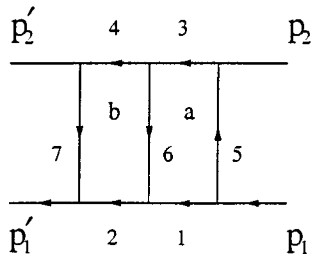

FIG. 1. A two-loop (Feynman) flow diagram.

expectations from the flow diagram. This situation can still be accommodated into the flow diagram if we simply reverse the arrow of this line by hand. This is *flow reversal*.

To summarize, here is how poles for the ''+'' integrations are computed for a given flow diagram, assuming a set of independent loops and a given order of k_{b+}-integrations have been chosen.

For the first loop, use the *naive rule* to read it off the flow diagram. This means that lines of this loop with arrows pointing in the same direction have their poles in the same half plane.

Assuming now k_{b+}-integrations have been carried out for $b=1,2,...,c$. We shall now proceed to do the $(c+1)$th integration for the term resulting from picking up poles located at line i_b for the bth loop, $b=1,2,...,c$.

First note that whatever loop $(c+1)$ is, it should not contain any of the lines $i_1,i_2,\ldots,i_c$. This is because the ''+'' momenta of these lines have been determined by previous integrations so they cannot be fixed again by the $(c+1)$th integration.

The naive rule can be used for lines j in loop $(c+1)$ if, (i) it is not in any one of the previous loops, 1,2,...,c, (ii) it is in a previous loop b but j is antiparallel to i_b in that loop, or (iii) j is parallel to i_b around loop b but $q_{j-}<q_{i_b-}$. In the remaining case, when j and i_b are parallel in loop b but $q_{j-}>q_{i_b-}$, we must reverse the arrow direction of line j before the naive rule is applied.

After all l ''+''-integrations are carried out, we obtain a number of terms, each of which is specified by a set of poles i_b for loop b. We shall call this collection of lines, $I=(i_1i_2\cdots i_l)$, a *contributing pole*.

Let us illustrate this recipe of obtaining contributing poles with two explicit examples, a two-loop diagram, and a four-loop diagram. In the process we will see how important it is to take flow reversals into account just to maintain consistency.

A. A two-loop example

Figure 1 is one of two possible flow diagrams for a two-loop Feynman diagram; the other has line 6 reversed.

Let a denote the loop with lines (1536) and b the loop with lines (2647). The big loop with lines (153472) is the union of these two loops and will be denoted by $a.b$. Only two of the three loop-momenta are independent.

There are three ways to start out the first loop integration, but the final results of their integrals must be the identical, and we must be able to pick the same contributing poles as well. We will illustrate here in detail how the latter can be achieved, iff proper flow reversals are taken into account.

Suppose we first integrate over loop a. In this loop the arrow of line 1 and the arrows of lines 5, 3, 6 are opposite, so their respective poles lie in opposite half planes of k_{a+}. We shall pick line 1 to be the relevant pole for further discussions. To simplify later descriptions we shall abbreviate this process of picking pole 1 from loop a simply by $a(1)$.

Now we are ready to tackle the second integration. Since line 1 is on $a.b$ we must not choose $a.b$ to be the second loop, so we are forced to choose it to be b. Line 6, which is in both loops b and a, is antiparallel to line 1 in loop a, so the naive rule once again applies to loop b. Lines 6 and

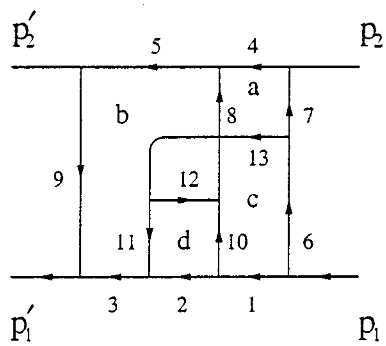

FIG. 2. A four-loop (Feynman) flow diagram.

2 are on one half-plane, and lines 4 and 7 on the other. We shall pick 2 and 6 to be the relevant poles. Consequently, we obtain two contributing poles, $I_1=(1,2)$ from $a(1)b(2)$, and $I_2=(1,6)$ from $a(1)b(6)$.

Next, let us start all over again but this time first carry out the integration around loop b to get $b(2)$ and $b(6)$. Now since line 2 lies in $a.b$, for the term $b(2)$ the second loop must be chosen to be a. In loop b line 6 is parallel to line 2, so there is a chance it might suffer a flow reversal. However, since $q_{2-}=q_{6-}+q_{1-}>q_{6-}$, flow reversal does not occur. Hence we have $b(2)a(1)$, so this contributing pole is $I_1=(1,2)$. For the term $b(6)$, since 6 is in a, the second loop must be chosen to be $a.b$. Now lines 2, 4, 7 are all in the first loop b, but 4 and 7 will not suffer flow reversal because they are antiparallel to 6. Line 2 is a different matter since $q_{2-}>q_{6-}$, so it would suffer a flow reversal. With this reversal, all lines in $a.b$ point in the same direction, with the sole exception of 1, so this yields $b(6)a.b(1)$, and the contributing pole is $I_2=(1,6)$. In this way we obtain the same set of contributing pole as before, as we should.

Finally suppose we carry out $a.b$ first, getting two terms $a.b(1)$ and $a.b(2)$. In the first case line 1 is in a so the second loop must be b. Lines 4 and 7 in b are antiparallel to 1 so they do not suffer from flow reversal. Line 2 is parallel to 1 and $q_{2-}>q_{1-}$ so it does suffer a flow reversal, thus leaving behind only line 6 of loop b in one direction. From $a.b(1)$ we therefore obtain $a.b(1)b(6)$ and the contributing pole $I_2=(1,6)$. Now consider the term $a.b(2)$. The second loop must now be a. Lines 5, 3 are antiparallel to 2 so they do not suffer flow reversal. Line 1 is parallel to 2 but $q_{1-}<q_{2-}$, so it does not suffer from flow reversal either. So no flow reversal occurs at all for lines in loop a, and this term yields $a.b(2)a(1)$, giving rise to the contributing pole $I_1=(1,2)$. The result is once again the same as the other two calculations. If flow reversals were not properly taken into account, the result would have been different and wrong.

The main lesson learned from this very simple example is that generally detailed loop-by-loop calculation must be performed, with proper flow reversals taken into account, in order to obtain the correct locations of the contributing poles. Also, the amount of calculations needed to determine the contributing poles may depend critically on the independent loops chosen and the order of integrations performed.

B. A four-loop example

The task of obtaining the contributing poles becomes more arduous for diagrams with a larger number of loops. The calculation must be carried out loop by loop, with more and more terms and flow reversals to keep track of. Besides, with multiloops there is a huge number of ways in choosing the independent loops and their order of integrations, each giving very different intermediate results though at the end they must all yield the same contributing poles. It is not known *a priori* how to make the best choice to maximally simplify the intermediate calculations.

To illustrate these points we shall work out in this subsection a four-loop example and obtain its contributing poles in two different ways.

Consider Fig. 2, with the following choice of independent loops: $a=(4,8,12,13,7)$, b

$=(5,9,3,11,12,8), c=(13,12,10,1,6)$, and $d=(10,12,11,2)$. Note that lines 8 and 13 are supposed not to intersect in the diagram. We shall carry out the integrations in the order *a,b,c,d* as much as possible.

The first integration over loop a yields $a(4)$ and $a(7)$.

We do the b integration next. The only lines common to loops a and b are 8 and 12, but since they are antiparallel to 4 and 7, no flow reversal takes place in carrying out the b integration. After the b-integration we get four terms, which for brevity shall be written together as additions: $[a(4)+a(7)][b(3)+b(11)]$.

Line 12 of loop c is also in loop a and loop b, and line 13 of loop c is in loop a. Since line 12 is antiparallel to 4 and 7 in loop a, and antiparallel to 3 and 11 in loop b, it suffers no flow reversal at loop c. Similarly line 13, being antiparallel to lines 4 and 7 in loop a, also has no flow reversal. Thus after the c integration, we get $[a(4)+a(7)][b(3)+b(1)][c(1)+c(10)]$.

The final d-integration is a bit complicated because loop d contains some of these poles from previously integrations so we are sometimes forced to take the loop *d.c* or the loop *d.b* instead of d itself. The final result contains 10 terms,

$$[a(4)+a(7)]b(3)c(1)[d(10)+d(11)]+[a(4)+a(7)]b(3)c(10)d.c(2) + [a(4)+a(7)]b(11)[c(1)d.b(2)+c(10)d.c(2)]. \tag{3.1}$$

To summarize, we have obtained ten contributing poles, (7,3,1,10), (7,3,1,11), (7,3,10,2), (7,11,2,1), (7,11,2,10), as well as another five with line 7 replaced by line 4.

Let us now illustrate another way to get the same result, by choosing this time the four independent loops to be $a=(4,8,12,13,7), b=(5,9,3,11,12,8), e=c.d=(1,6,13,11,2)$, and $d=(10,12,11,2)$, and try to carry out the integration in the order *a,b,c.d,d* as much as possible.

These loops are what we shall later call the *natural loops* for the contributing pole (7,3,1,10). They are obtained first by removing the lines 7,3,1,10 from the original diagram, and then inserting one of them back at a time to get the four loops.

The first two integrations are identical to those before, so we get $[a(4)+a(7)][b(3)+b(11)]$. Now $e=c.d$ contains the line 11 but not 3, so the next integration involving $b(3)$ gives $e(1)+e(2)$ but the next integration involving $b(11)$ gives $c(1)+c(10)$, as $c=d.(c.d)$. The last loop d contains lines 2 and 11, so for some terms the integration over d has to be changed into integration over *d.b* or $d.e=c$. The final answer is

$$[a(4)+a(7)]b(3)e(1)[d(10)+d(11)]+b(3)e(2)c(10) + b(11)c(1)d.b(2)+b(11|3)c(10)d.b(2). \tag{3.2}$$

This results in the same ten contributing poles as before, as it should.

The calculation could be even more complicated if we encounter a line j which is parallel to a pole line i of an earlier loop, but the relative magnitude of q_{j-} and q_{i-} can be either way. In that situation we must divide this flow diagram into two, one in which $q_{j-}<q_{i-}$ and line j is not reversed, and the other with $q_{j-}>q_{i-}$ where line j must be reversed.

IV. PATH METHOD FOR FINDING CONTRIBUTING POLES

In this section we propose a simple (path-) method to obtain the contributing poles. With this method there is no need to declare the independent loops and their order of integrations, so there is no need to keep track of the complicated flow reversals either. This makes the method most useful in the presence of a large number of loops.

We begin by choosing a path P in the flow diagram. By a path we mean a continuous line (no branches, no loops) running from beginning to end, with all the arrows on it pointing in the same direction. The thin solid lines in Figs. 3 and 4 are examples of such paths. By adding branches to

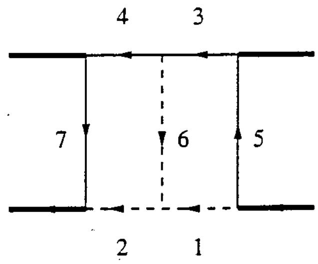

FIG. 3. The solid line is the path P used to obtain contributing poles for Fig. 1.

the path we can construct trees. A class of these trees, $T[P]$, turns out to be in one–one correspondence with the contributing poles. The path method of finding contributing poles is actually a method to construct the trees in $T[P]$.

From an l-loop diagram one can obtain trees by removing l lines. We shall refer to these removed lines as the *missing lines* for the tree. The set of all trees so obtained with path P as their common backbone will be denoted by $S[P]$. From $S[P]$ we select a subset $T[P]$ satisfying the following *directional rule*; when any one of the l missing lines is inserted into the tree, a loop is formed. If the inserted line around this loop is parallel to the lines along path P, this tree is rejected. If it is antiparallel, then this tree is retained to be a member of $T[P]$.

We assert that the missing lines of any tree in $T[P]$ is a contributing pole of the diagram, and there is actually a one–one correspondence between contributing poles and individual trees in $T[P]$. This is the essence of the *path method*.

This method does not restrict what path P one chooses, but the longer the path the fewer the number of contributing poles, and the easier the calculations. So in practice we often choose the longest path we can manage, though this is not a requirement of the method. In Sec. VII C an example will be shown in which computations based on two different paths are shown for comparison. The reason why one can get the same result by choosing different paths P, or equivalently different sets of contributing poles, is because of the freedom to choose poles from either half-plane each time we carry out any integration.

We have implicitly assumed in these discussions that a path P is chosen after we are given a flow diagram. This is not strictly necessary. We may start from a Feynman diagram or a non-Abelian cut diagram, without arrows attached, and start drawing a path on it. This can be taken as the starting point to determine possible flow diagrams consistent with this path; arrows on the path must all point in one direction, other arrows must be installed not to violate the direction rule to obtain contributing poles.

Before proceeding to prove the path method let us first see how it can be applied to obtain the contributing poles of Figs. 1 and 2 very simply.

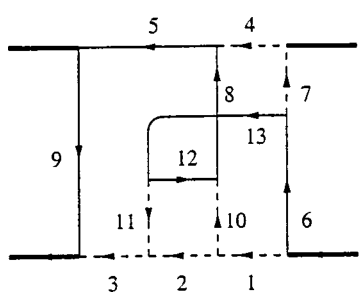

FIG. 4. The solid line is the path P used to obtain the contributing poles of Fig. 2.

A. Examples

For Fig. 1 let us choose the path P to be (5347), shown in Fig. 3 as thin solid lines. Then $S[P]=\{(P,6),(P,1),(P,2)\}$, and $T[P]=\{(P,6),(P,2)\}$. The tree $(P,1)$ violates the directional rule for the following reason so it is not in $T[P]$. When line 6 is inserted into $(P,1)$, it is parallel to P in the loop (6153), so it has to be rejected. With this $T[P]$, the contributing poles are the missing lines so they are (1,2) and (1,6), agreeing with the result obtained previously.

Let us next apply the method to obtain the contributing poles of Fig. 2, taking $P=(6,13,12,8,5,9)$ as the path (Fig. 4). Then,

$$T[P]=\{(P,7,2,11),(P,7,2,10),(P,7,1,11),(P,7,3,10),(P,7,1,3)\},$$

and five more with 7 replaced by 4. The contributing poles are therefore (7,3,1,10), (7,3,1,11), (7,3,10,2), (7,11,2,1), and (7,11,2,10), and another five with 7 replaced by 4, the same 10 terms as before. The trees in $S[P]/T[P]$ are $\{(P,7,1,2),(P,7,2,3),(P,7,10,11)\}$, and three more with 7 replaced by 4. $(P,7,1,2)$ violates the directional rule when the line 11 is inserted; $(P,7,2,4)$ violates the directional rule when line 10 is inserted; and $(P,7,10,11)$ violates the directional rule when 2 is inserted.

B. Proof

A tree $t\in S[P]$ defines a set of independent loops $\mathcal{N}[t]$ of the original diagram by filling in the missing lines one at a time. The special feature of $\mathcal{N}[t]$ is that the missing lines are never on the boundary of two loops. We shall later on refer to these loops as the *natural loops* for the missing lines.

Now we proceed to the proof of the path method. We assume we always close the integration contour in the half-plane in which poles reside on lines running in the opposite direction as those on P.

The proof makes use of the simple fact that the same set of contributing poles can be computed using any independent loops and any order of integration.

Removing the pole lines of a contributing pole from the original diagram gives rise to a tree in $S[P]$. We shall denote the set of all such tress as $T'[P]$. Our task is to show that $T[P]=T'[P]$.

Take any $t'\in T'[P]$. The removed pole lines clearly satisfy the directional rule when they are inserted back, because poles are always taken from those lines running in the opposite direction as P. Hence $t'\in T[P]$ and $T'[P]\subset T[P]$.

Conversely, take a $t\in T[P]$, and use the independent loops $\mathcal{N}[t]$ to compute the contributing poles. The missing lines of t are obviously one of the pole lines, for according to the directional rule they all run opposite to the path direction. Hence $t\in T'[P]$ and $T[P]\subset T'[P]$.

Putting the two together, we get $T[P]=T'[P]$, as desired.

V. NON-ABELIAN CUT DIAGRAMS

General methods found in the literature to compute high energy limits of Feynman diagrams[3,4] are by and large valid only in the leading-log approximation (LLA). They become virtually powerless if these leading-log contributions cancel when the Feynman diagrams are summed, a situation which unfortunately occurs quite frequently.[4] A method was developed recently to bypass this difficulty, by allowing the cancellations to occur before the high energy limit is taken. The cancellations are incorporated into the individual *non-Abelian cut diagrams*,[10,12] whose space–time amplitudes (for onshell diagrams) turn out to differ from the corresponding Feynman diagram only by having the denominators $(q_i^2-m^2+i\epsilon)^{-1}$ of certain propagators replaced by the corresponding Cutkosky propagators $-2\pi i\,\delta(q_i^2-m^2)$. The advantage of the non-Abelian cut diagrams is that the sum of Feynman diagrams is the same as the sum of non-Abelian cut diagrams, but in the latter cancellations took place before the high-energy limit is taken, so their leading-log contributions (LLA) survive the sum. For this to happen it is clearly necessary for the

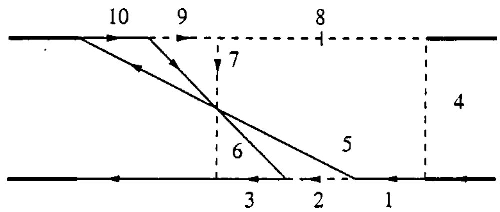

FIG. 5. A 3-loop (non-Abelian cut) flow diagram. The solid line represents the path P.

LLA of a non-Abelian cut diagram to have a smaller $\ln s$ power than the corresponding Feynman diagram, if the sum of the LLA contributions of the latter is to vanish. This is actually made possible by the presence of the Cutkosky propagators.

For high-energy two-body (e.g., *quark–quark*) scattering, the Cutkosky propagators occur only on the top quark lines. In the high energy limit, it can be shown that the combination $q_i^2 - m^2$ is actually proportional to the ''$-$'' momentum on that line, so a δ-function is that variable is a δ-function of the ''$-$'' momentum.[10,11] This has the effect of stopping the ''$-$'' momentum from flowing through this line, so as far as the flow diagram is concerned we may think of these lines as being absent. For the rest of the non-Abelian cut diagram the flows are constructed in exactly the same way as in a Feynman diagram, and contributing poles can be located the same way just as well.

As an example, consider the non-Abelian-cut (flow) diagram of Fig. 5, where the Cutkosky propagator is located at line 8, indicated there by a vertical bar ($|$). Hence the ''$-$'' momentum is absent from lines 8, and also from line 4 by continuity. We may therefore ignore these two lines in the rest of the discussions.

To obtain the contributing poles from the path method, we can choose the path to be $P=(1,5,10,6,3)$, then $T[P]=\{(P,9),(P,7)\}$, giving rise to the contributing poles (2,7) and (2,9).

VI. DOMINANT INTEGRATION REGIONS IN LLA

Contributing poles, extracted from the path method or otherwise, can be used to determine the internal momenta most important to the loop amplitude. The ''$+$'' momenta from the lines of a contributing pole $I=(i_1 i_2 \cdots i_l)$ are fixed by the pole condition to be $q_{i_k +}=(a_{i_k}-i\epsilon)/q_{i_k -}$, and those of any other line are fixed by momentum conservation. An easy way to read them out is to use the *natural loops* discussed before. These are simply the independent loops containing one and only one pole line each.

In LLA a number of simplifications emerge immediately. For quark-quark scattering in the c.m. system, quark 1 carries a ''$-$'' momentum $\sqrt{s}$ and quark 2 carries a ''$+$'' momentum $\sqrt{s}$. In LLA, where $|t|$ and squared masses are ignored compared to s, both quarks go straight through by carrying the full forward momenta with them. In other words, $q_{j-}\simeq\sqrt{s}$ for every line j of quark 1 (the ''bottom lines''), and $q_{j+}\simeq\sqrt{s}$ for every line j of quark 2 (the ''top lines''). This means that we can ignore the contributing poles with a pole line on top, for q_{j+} of a top line is $\sqrt{s}$ and not determined by the pole condition above. In other words, if we insist on taking a pole there, then this term will not contribute in the LLA.

The two-body amplitude for a flow diagram, after the ''$+$'' integration is performed, can be written as

$$M=\int\left(\prod_{b=1}^{l} d^2k_{b\perp}\right)F, \quad F=\int_R\left(\prod_{b=1}^{l} dx_b\right)G,$$
$$G\equiv\frac{N}{D}\equiv\frac{N}{\Pi_{j=1}^{n} d_j}, \tag{6.1}$$

where $x_b=k_{b-}/\sqrt{s}$ and $Q_j=q_{j-}/\sqrt{s}$ are the scaled ''$-$'' momenta. In practice x_b are chosen from the Q_j's of types (i) and (ii) below. The integration region R of x_b is determined by the n flow-diagram conditions $Q_j\geqslant 0$.

The denominator $D=\Pi_j d_j$ is derived from the denominators of the propagators $1/d(q_i)$, scaled in some convenient way as follows. (i) If line $j=i_k$ is part of the contributing pole I, then d_j is defined to be the scaled residue Q_j; (ii) if line j is a top line, then $d_j \equiv d(q_j)/s = \pm Q_j - a_j/s + i\epsilon \simeq \pm Q_j + i\epsilon$, where the sign in front of Q_j is $+/-$ if the arrow on line j is parallel/antiparallel to the ''+'' flow of quark 2; (iii) for any other line j, d_j is equal to $d(q_j)$ evaluated at the contributing pole, so

$$d_j = Q_j \sum \left(\pm \frac{a_{i_k}}{Q_{i_k}} \right) - a_j, \tag{6.2}$$

where the sum is taken over lines i_k in the same natural loops as line j, with an appropriate sign.

For convenience we will label lines of these three types by different indices; index p (for ''*p*ole'') for type (i), t (for ''*t*op'') for type (ii), and s (for ''side'') for type (iii). We shall retain the index j to denote any of them in general.

The numerator factor N consists of all the rest, including the vertex factors and factors of $\sqrt{s}$ discarded by D.

It should be noted that there are no explicit factors of ''i'' hidden in M, except those explicity contained in the vertices and those appearing as $i\epsilon$ in the propagators. An l-loop Feynman diagram has an explicit factor $(-i)^l$, and this is cancelled by the l factors of $2\pi i$ from contour integration, leaving behind no explicit factors of i. This observation is important in determining how the imaginary part of a scattering amplitude arises.

For nonabelian cut diagrams with c cuts, the Feynman propagator $1/d(q)$ at each cut line is replaced by the Cutkosky propagators $-2\pi i\delta(q^2-m^2)$,[10,11] so an explicit factor $(-i)^c$ will emerge.

From (6.1) and the rules for d_j, it would appear that the integral F diverges at the boundaries $Q_p=0$ and $Q_t=0$. Actually because of *obstructions* from the side lines s, the singularity in the Q_p variable is cancelled so there are no divergences at $Q_p=0$. This is so because as $Q_p \to 0$, the ''+'' momentum $q_{p+} \simeq a_p/Q_p$ becomes very large. At some point it will become much smaller than all the Q_s, whence $d_s \simeq (Q_s/Q_p)a_s$ for any line s in the natural loop of p. This washes out the factor Q_p in d_p, leaving behind no divergence at this boundary.

A divergence does occur at $Q_t=0$, but this divergence is an artifact of our high-energy approximation of dropping $\xi/s \equiv \mp(a_t - i\epsilon)/s$ compared Q_t, where ξ is of the order of the squared masses and the squared momentum transfer $-t$. If we restore it by installing a cutoff ξ/s at these boundaries, the divergences will be absent and they will be turned into enhancement factors of s. If the enhancement is logarithmic, the value of ξ does not matter in the LLA, and that will be the case in gauge theories. But if it is powerlike, then the coefficient of the power dependence would depend on $\xi = \mp(a_t - i\epsilon)$, and its effective value could be determined only after the transverse-momentum integrations.

The integral F, thus enhanced, receives contributions in the form

$$F \simeq \int_{(\xi/s)} \frac{dQ_1'}{Q_1'^{m_1}} F_1, \tag{6.3}$$

where Q_1' is either one of the Q_t's, or the radial variable of several of them that are linearly independent. As it will become clear shortly this will be the smallest of all the ''$-$'' variables in the dominant integration region R_0.

In the region $Q_1' \ll Q_p$, we may set the ratios $Q_1'/Q_p = 0$ in all remaining d_s. This removes obstructions from some of the side lines, so that the integrand of F_1 may now encounter singularity again in some variable Q_2', say like $1/Q_2'^{m_2}$, with $m_2 \geqslant 1$. This new singular variable Q_2' would be equal to some Q_t or Q_p, or the radial variable of several of them. Now we have

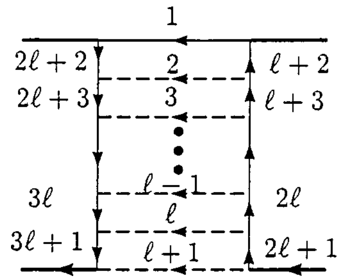

FIG. 6. Ladder diagram for scalar quarks and gluons. The path P is indicated by the light solid line and the poles indicated by the dotted lines. There is only one flow diagram and one contributing pole in this case.

$$F_1 \simeq \int_{B_1 Q_1'} \frac{dQ_2'}{Q_2'^{m_2}} F_2, \tag{6.4}$$

for some $B_1 \gg 1$. Similarly, in the region $Q_1' \ll Q_2' \ll Q_p$ for the remaining pole lines p, Q_2'/Q_p can also be set equal to zero, thus removing further obstructions from even more side lines. This enables another singular variable Q_3' to emerge, and so on. Continue this way until no further singularities are encountered, we get

$$F \simeq \int_{(\xi/s)} \frac{dQ_1'}{Q_1'^{m_1}} \int_{B_1 Q_1'} \frac{dQ_2'}{Q_2'^{m_2}} \int \cdots \int_{B_{\nu-1} Q_{\nu-1}'} \frac{dQ_\nu'}{Q_\nu'^{m_\nu}} F_{\nu+1}. \tag{6.5}$$

The integrand $F_{\nu+1}$ is assumed to be regular so its Q_i' dependencies can all be put equal to zero. All $B_i \gg 1$.

The dominant region of integration R_0 is then given by

$$R_0 = \{\xi/s \leqslant Q_1' \ll Q_2' \ll \cdots \ll Q_\nu' \ll 1\}, \tag{6.6}$$

from which we can work out where the ''+'' momenta are located as well. The transverse momenta $k_{b\perp}$ are all of the same order as the momentum transfer $\sqrt{-t}$.

In gauge theories only logarithmic enhancements occur. This means all $m_i = 1$, and

$$F \simeq \frac{F_{\nu+1}}{\nu!} (\ln s)^\nu. \tag{6.7}$$

For an l-loop Feynman diagram, the maximum enhancement is $\sim (\ln s)^l$. For an l-loop nonabelian cut diagrams with c cuts, the maximum enhancement is $\sim (\ln s)^{l-c}$. We shall refer to diagrams with these maximal enhancements as *saturated*, and these are the diagrams of most interestes to us in LLA. Diagrams with less enhancements will be called *unsaturated*. A number of saturated and unsaturated diagrams are considered in the next section as concrete examples to illustrate the procedures here. For saturated diagrams we will also work out the coefficient of the leading-log term.

VII. EXAMPLES

A. Scalar ladder diagram

Consider the ladder diagram Fig. 6 for scalar quarks and scalar gluons. There is only one nonzero flow diagram, as shown, and in it there is only one contributing pole, namely, $I = (2,3,...,l+1)$, indicated by the dotted lines. The path P from which this contributing pole is obtained is drawn as a light solid line in the diagram.

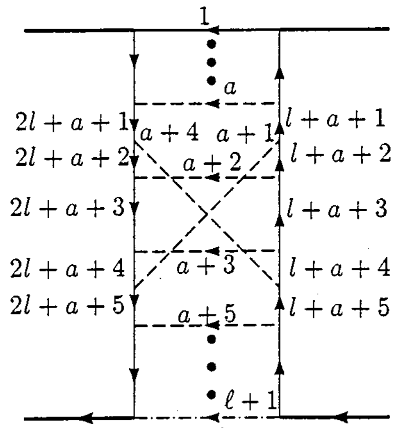

FIG. 7. A crossed ladder diagram, with path P given by the light solid line and the pole lines given by the dotted lines. The rungs above a and below $a+5$ are all uncrossed.

In the language of the last section, the pole lines are $2\leqslant p\leqslant l+1$, the top line is $t=1$, and the side lines are $l+2\leqslant s\leqslant 3l+1$.

The independent ''+'' momenta at the pole lines are given by

$$q_{p+}\sqrt{s}=(a_p-i\epsilon)/Q_p\,, \tag{7.1}$$

with $Q_{l+1}\simeq 1$ in LLA because it is a bottom line. Thus all these ''+'' momenta except q_{l+1} are capable of being large if the corresponding Q_p is small enough. The ''+'' momenta carried by the side lines can most easily be read off from the natural loops, which are rectangles bounded below by the line p and bounded above by the top line 1.

Following the discussions of last section, the top line 1 is the unique candidate for the first singular variable Q_1', and indeed it is with $m_1=1$. In the region $Q_1'\ll Q_j$ for $j>1$, obstructions from lines $l+2$ and $2l+2$ are removed, resulting in $d_{l+2}=-a_{l+2}$ and $d_{2l+2}=-a_{2l+2}$. This allows a new singular structure to emerge with $Q_2'=Q_2$ and $m_2=1$. This in turn removes the obstruction from lines $l+3$ and $2l+3$ in the region $Q_1'\ll Q_2'\ll Q_j$ for $j>2$, etc. Continuing this way, we obtain $Q_j'=Q_j$ and $m_j=1$ for $1\leqslant j\leqslant l$. Thus the diagram is saturated, and we obtain the amplitude to be

$$F=\frac{1}{l!}(\ln s)^l\prod_{i=l+2}^{3l+1}(-a_i). \tag{7.2}$$

In obtaining this expression, we have set the numerator N of the integrand to be 1.

The integration region is given by $R_0=\{\xi/s\leqslant Q_1\ll Q_2\ll\cdots\ll Q_l\ll 1\}$. According to (7.1), the ''+'' momenta are strongly ordered in the opposite way because the a_p's are all of the same order. The virtualities of the side lines s are all spacelike and of order 1, $q_s^2=-a_s=-q_{s\perp}^2$. In other words, the dominant momenta of the virtual gluons come from the multi-Regge region, the same as those used in the dispersion-relation approach.[5]

B. Crossed ladders

When the rungs of the ladders are crossed, the scalar diagram will no longer be saturated. This could be inferred from the example above and the s-channel dispersion relation, but let us see how to obtain this conclusion directly from the Feynman diagram, and how unsaturated it is.

Consider Fig. 7, which is obtained from Fig. 6 by crossing two rungs separated by $r=2$ horizontal rungs in between. The path P and the contributing pole remain unchanged. As before,

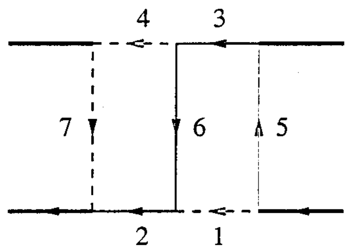

FIG. 8. The solid line is the path P' used to obtain the contributing poles from Fig. 1, indicated here by dotted lines. This is a different path than the one used in Fig. 3.

we let $Q_j'=Q_j$ for $1\leqslant j\leqslant a$, and let these Q_j''s to be strongly ordered as before. Then $m_j=1$ just as in the previous example. The question is what happens when we come to the region where the rungs are crossed.

Every ''+'' momentum $q_{i+}\sqrt{s}(1\leqslant i\leqslant 3l+1)$ is a linear combination of some $a_p/Q_p(2\leqslant p\leqslant l+1)$. We shall use the symbol $[p_1p_2\cdots p_k]$ to represent this ''+'' momentum if it receives contributions from $p=p_1,p_2,\ldots,p_k$ *in the crossed region*. Similarly, its ''−'' momenta are linear combinations of Q_1 and $Q_p(2\leqslant p\leqslant l)$, and those from the crossed region that contribute to the ''−'' momentum of a particular line will be enclosed between angular brackets $\langle\cdots\rangle$.

The ''−'' and ''+'' momenta contributions for the side line $s=l+a+k$ on the right $(1\leqslant k\leqslant 5)$ are $\langle a,\ldots,a+k-1\rangle[a+k,\ldots,a+5]$. For the side lines $s=2l+a+k$ on the left, they are $\langle a\rangle[a+4,a+2,a+3,a+1,a+5]$ for $k=1$, $\langle a,a+4\rangle[a+2,a+3,a+1,a+5]$ for $k=2$, $\langle a,a+4,a+2\rangle[a+3,a+1,a+5]$ for $k=3$, $\langle a,a+4,a+2,a+3\rangle[a+1,a+5]$ for $k=4$, and finally $\langle a,a+4,a+2,a+3,a+1\rangle[a+5]$ for $k=5$. There is no way to strongly order the variables $Q_{a+1},Q_{a+2},Q_{a+3},Q_{a+4}$ in the crossed region to get rid of all the obstructions. Whatever that works on the right-hand side will fail on the left-hand side, and vice versa. The only way out is to have these four to be of the same order, for then the ratio of any two of these four would be of order 1, and all the obstructions from the side lines would disappear. Their common radial variable $Q_{a+1}'=(\Sigma_{i=1}^4 Q_{a+i}^2)^{1/2}$ is singular, with $m_{a+1}=1$, because $Q_{a+1}'^3 dQ_{a+1}'/Q_{a+1}'^4=dQ_{a+1}'/Q_{a+1}'$. From there on, everything looks like the uncrossed ladder again, so $Q_j'=Q_{j+3}$ for $a+2\leqslant j\leqslant p=l-3$. The final integral F is proportional to $(\ln s)^{l-3}$, hence unsaturated. More generally, the same argument shows that if there are r uncrossed rungs between the two crossed rungs, then $F\sim(\ln s)^{l-r-1}$.

C. Two-loop QED diagram

Consider now the two-loop diagram Fig. 1 for electron–electron scattering by exchanging 3 photons.

We shall compute this in two ways. First, using the path P and the contributing poles of Fig. 3, we will obtain saturated contributions from each of these two contributing poles, but their sum vanishes so this diagram turns out to be unsaturated. To see this unsaturation directly, we will use another path P' shown in Fig. 8. This path has only one contributing pole so there can be no chance of a cancellation, and it gives rise to an unsaturated LLA amplitude. In this latter approach we would also be able to compute the coefficient of the leading log term by LLA calculation if we should want to.

The numerator N of (6.1) in this case comes from the vertices, and is proportional to s. For simplicity we will assume it to be simply s.

The path P from Fig. 3 gives two contributing poles, $I_1=(1,2)$ and $I_2=(1,6)$. First consider $I_1=(1,2)$. Since both poles lie on the bottom line, $Q_1\simeq Q_2\simeq 1$, there are no obstructions on the side lines. Since $Q_3>Q_4$, the integral is

$$F \simeq -\frac{s}{a_5 a_6 a_7}\int_{(\xi/s}\frac{dQ_4}{Q_4}\int_{Q_4}\frac{dQ_3}{Q_3} \simeq \frac{s}{(-)^3 a_5 a_6 a_7}\int_{\xi/s}\frac{dQ_4}{Q_4}\int_{Q_4}\frac{dQ_3}{Q_3} \simeq -\frac{s}{2a_5 a_5 a_6 a_7}(\ln s)^2. \tag{7.3}$$

Next consider $I_2=(1,6)$. The pole on 6 causes an obstruction from lines 2 and 7. By choosing $Q_1'=Q_4 \ll Q_2'=Q_6$, the obstruction from line 7 is removed but the obstruction from line 2 remains because $Q_2 \simeq 1$. However, since $Q_3 \simeq Q_6 = Q_2'$, the contribution from I_2 is

$$F \simeq \frac{s}{(-)^2 a_5 a_7}\int_{\xi/s}\frac{dQ_1'}{dQ_1'}\int_{B_1 Q_1'}\frac{dQ_2'}{Q_2'^2 (a_6/Q_2')} \simeq +\frac{s}{2a_5 a_6 a_7}(\ln s)^2. \tag{7.4}$$

The sum of the contributions from I_1 and I_2 vanishes in order $(\ln s)^2$ so the diagram is unsaturated.

To see this unsaturation directly, choose another path P' as shown in Fig. 8. The contributing pole is now $I'=(1,7)$. The obstruction induced by line 7 on lines 2 and 6 block out the factor $d_4 d_7 = Q_4 Q_7$, so to get a singular integrand for F we must enlist the help of Q_3. If Q_1' is the radial variable of Q_7 and Q_3, then the integrand of F is proportional to $Q_1' dQ_1'/Q_3 Q_4 Q_7 d_2 \sim dQ_1'/Q_1'$, so the leading contribution to this diagram is of the order $\ln s$.

D. Four-loop diagram

If the path P for the four-loop diagram Fig. 2 is chosen as in Fig. 4, then as we have seen there are 10 contributing poles. For illustration we will look at the contribution from a single one, (7,3,10,2). We shall see that there will be no $\ln s$ enhancement if the diagram is scalar, but if it is a QCD diagram then there will be a linear $\ln s$ enhancement from this contributing pole.

There are two top lines in this diagram, lines 4 and 5. Since $Q_5 > Q_4 = Q_7$, the only single-variable candidate for Q_1' is $Q_4 = Q_7$. However, the pole in 7 produces an obstruction on all the other lines in its natural loop, lines 8, 12, and 13. With three obstructing lines and only two singular factors, the resulting Q_1' dependence cannot be singular for a scalar diagram. One could go on and try to find a singular Q_1' among the radial variables of several Q_j's, and one would not succeed either. Consequently as a scalar diagram it has no $\ln s$ enhancement.

As a QCD diagram we must incorporate the vertex factors into the numerator N of the integrand of F. The vertex factor for a gluon connected to the top line is $2p_2$, and to the bottom line is $2p_1$. There are however also three triple-gluon vertices, at the junctions of lines (6,7,13) $=A$, $(11,12,13)=B$, and $(8,10,12)=C$. Each of them contains three terms, but one of the three terms of each is dotted into $2p_1$ and therefore produces an appropriate combination of $q_{i+j} g_{7,13}(q_{7+} - q_{13+})$ for A, $g_{12,13}(q_{12+} + q_{13+})$ for B, and $g_{8,12}(q_{8+} + q_{12+})$ for C. Since every line in the natural loop of 7 contains $\pm q_{7+}$, and hence a factor $1/Q_1'$, these three vertex factors can make the Q_1' variable much more singular. However, we may use only two out of the three, for otherwise the $g_{\alpha\beta}$ factors will lead to a dot product of the (7,4) vertex and the (4,5,8) vertex, thus producing an extra factor $2p_2 \cdot 2p_2 = 0$. With the help of two triple-gluon vertices, we get $m_1 = 1$ and a $\ln s$ enhancement from the Q_1' variable.

The remaining singular factors for the integrand come from lines 5 and 10. Since $Q_5 > Q_{10}$, if Q_2' comes from a single variable Q_j we must have $j=10$. This pole at 10 may produce obstructions on lines of its natural loop (10,12,13,6,1). Those on lines 12 and 13 have already been removed by Q_1', so this leaves obstructions from lines 1 and 6. The one on 1 is particularly troublesome because it is a bottom line, so $Q_1 \simeq 1$ and the obstruction can never be removed. For that reason Q_{10} is not a singular variable, and it can be checked that the radial variable of Q_{10} and one or two other Q_j's cannot be a singular variable either. The enhancement of the QCD diagram is therefore just $\ln s$.

E. Non-Abelian cut diagram

As a last example we consider the non-Abelian cut diagram, Fig. 5, treated as a scalar diagram with numerator factor $N=1$. The path is $P=(1,5,10,6,3)$ and the contributing poles are (2, 7) and (2, 9). Since in the LLA we would never have to consider any contributing pole on the top line, we can drop (2, 9) and consider only (2, 7).

In either case there is actually a hidden contributing pole at line 1. This does not show up explicitly in the path method because the cut line 8 reduces the flow diagram into a two-loop diagram, hence only two of the three poles show up explicitly. In any case, since $Q_4=Q_8=0$, we have $Q_1=1$, so the pole at 1 produces $q_{1+}=a_1$. This together with the q_{p+} obtained from the other two pole lines uniquely determine all the ''+'' momenta of all the lines. However, the contribution from q_{1+} is finite, it will never lead to an obstruction, so in some sense we can just forget about it.

Of the two uncut top lines, $Q_{10}>Q_9=Q_7$, so if Q_1' is given by a single Q_j, it would have to be $j=9$. The second singular variable is $Q_2'=Q_6=Q_{10}\gg Q_1'$, and the integral is

$$F=\int_{\xi/s}\frac{dQ_1'}{Q_1'^2(-a_7/Q_1')(-a_6)}\int_{m_1Q_1'}\frac{dQ_2'}{Q_2'(-a_5)(-a_4)}\int dQ_8(-2\pi i)\delta(Q_8)$$

$$\simeq-\frac{\pi i}{a_4a_5a_6a_7}(\ln s)^2. \tag{7.5}$$

ACKNOWLEDGMENTS

This research was supported in part by the Natural Science and Engineering Research Council of Canada and by the Québec Department of Education. Y.J.F. acknowledges the support of the Carl Reinhardt Major Foundation, and C.S.L. wants to thank Hung Cheng for a stimulating discussion.

[1] C. S. Lam, hep-ph/9804463.
[2] For example, see V. S. Fadin and L. N. Lipatov, Nucl. Phys. B **477**, 767 (1996); hep-ph/9802290.
[3] R. J. Eden, P. V. Landshoff, D. I. Olive, and J. C. Polkinghorne, *The Analytic S-Matrix* (Cambridge University Press, Cambridge, 1966).
[4] H. Cheng and T. T. Wu, *Expanding Protons: Scattering at High Energies* (MIT Press, Cambridge, 1987).
[5] L. N. Lipatov, Yad. Fiz. **23**, 642 (1976) [Sov. J. Nucl. Phys. **23**, 338 (1976)]; Ya. Ya. Balitskii and L. N. Lipatov, Yad. Fiz. **28**, 1597 (1978) [Sov. J. Nucl. Phys. **28**, 822 (1978)]; E. A. Kuraev, L. N. Lipatov, and V. S. Fadin, Zh. Eksp. Teor. Fiz. **71**, 840 (1976) [Sov. Phys. JETP **44**, 443 (1976)]; **72**, 377 (1977) [**45**, 199 (1977)]; V. Del Duca, hep-ph/9503226.
[6] J. Bartels, Nucl. Phys. B **151**, 293 (1979); **175**, 365 (1980).
[7] H1 Collaboration, hep-ex/ 9708016; ZEUS Collaboration, hep-ex/ 9709021.
[8] J. Bartels, Z. Phys. C **60**, 471 (1993); J. Bartels and M. Wuesthoff, *ibid.* **66**, 157 (1995); M. A. Braun, Nucl. Phys. B **175**, 365 (1980); hep-ph/9706373.
[9] See, for example, the references cited in Ref. 1.
[10] Y. J. Feng, O. Hamidi-Ravari, and C. S. Lam, Phys. Rev. D **54**, 3114 (1996).
[11] Y. J. Feng and C. S. Lam, Phys. Rev. D **55**, 4016 (1997).
[12] C. S. Lam and K. F. Liu, Nucl. Phys. B **483**, 514 (1997); Phys. Rev. Lett. **79**, 597 (1997).
[13] H. Verlinde and E. Verlinde, hep-th/9302104.
[14] L. N. Lipatov, Nucl. Phys. B **365**, 614 (1991); R. Kirschner, L. N. Lipatov, and L. Szymanowski, *ibid.* **425**, 579 (1994); Phys. Rev. D **51**, 838 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Quasispin graded-fermion formalism and $gl(m|n)\downarrow osp(m|n)$ branching rules

Mark D. Gould and Yao-Zhong Zhang[a)]
Department of Mathematics, University of Queensland, Brisbane, Queensland Qld 4072, Australia

(Received 11 May 1999; accepted for publication 13 July 1999)

The graded-fermion algebra and quasispin formalism are introduced and applied to obtain the $gl(m|n)\downarrow osp(m|n)$ branching rules for the ''two-column'' tensor irreducible representations of $gl(m|n)$, for the case $m\leqslant n(n>2)$. In the case $m<n$, all such irreducible representations of $gl(m|n)$ are shown to be completely reducible as representations of $osp(m|n)$. This is also shown to be true for the case $m=n$, except for the ''spin-singlet'' representations, which contain an indecomposable representation of $osp(m|n)$ with composition length 3. These branching rules are given in fully explicit form.
[S0022-2488(99)04410-2]

I. INTRODUCTION

It is well known that branching rules are of great importance in the study of representation theory. They also play an essential role in the determination of the parities for the components appearing in the twisted tensor product graphs and the construction of corresponding *R* matrices.[1,2]

There appear to be virtually no results in the literature on the branching rules for Lie superalgebras. The only exception is Ref. 3, in which the branching rules are determined for all typical and atypical irreducible representations of $osp(2|2n)$ with respect to its subalgebra $osp(1|2n)$. It is very interesting (and important) to investigate the branching rules for other Lie superalgebras.

In this paper we investigate the antisymmetric tensor irreducible representations of $gl(m|n)$. This class of representations is of interest since they are also irreducible under the fixed point subalgebra $osp(m|n)$. Moreover, their quantized versions can be shown to be affinizable to provide irreducible representations of the twisted quantum affine superalgebra $U_q[gl(m|n)^{(2)}]$ from which trigonometric *R* matrices with $U_q[osp(m|n)]$ invariance may be constructed.[4]

These *R* matrices determine new integrable models that have generated remarkable interest in physics recently,[5–7] particularly in condensed matter physics, where they give rise to new integrable models of strongly correlated electrons.

To explicitly construct such *R* matrices it is necessary to determine the reduction of the tensor product of two antisymmetric tensor irreducible representations into ''two column'' irreducible representations of $gl(m|n)$ which are then decomposed into irreducible representations of its fixed point subalgebra $osp(m|n)$.

We determine the $gl(m|n)\downarrow osp(m|n)$ branching rules for these two column irreducible tensor representations of $gl(m|n)$, for the case $m\leqslant n$, $n>2$. A natural framework for solving this problem is provided by the graded-fermion algebra and the quasispin formalism, which we introduce and develop in this paper. The Fock space for this graded-fermion algebra affords a convenient realization of the class of irreducible representations of $gl(m|n)$ concerned. The reduction to $osp(m|n)$, and thus the $gl(m|n)\downarrow osp(m|n)$ branching rules, can be achieved using the quasispin formalism.

[a)]Electronic mail: yzz@maths.uq.edu.au

II. $osp(m|n=2k)$ AS A SUBALGEBRA OF $gl(m|n)$

Throughout this paper, we assume $n=2k$ is even and set $h=[m/2]$ so that $m=2h$ for even m and $m=2h+1$ for odd m. For homogeneous operators A, B we use the notation $[A,B]=AB-(-1)^{[A][B]}BA$ to denote the usual graded commutator. Let E^a_b be the standard generators of $gl(m|n)$ obeying the graded commutation relations,

$$[E^a_b,E^c_d]=\delta^c_b E^a_d-(-1)^{([a]+[b])([c]+[d])}\delta^a_d E^c_b. \tag{II.1}$$

In order to introduce the subalgebra $osp(m|n)$, we first need a graded symmetric metric tensor $g_{ab}=(-1)^{[a][b]}g_{ba}$, which is assumed to be even. We shall make the convenient choice

$$g_{ab}=\xi_a\delta_{a\bar{b}}, \tag{II.2}$$

where

$$\bar{a}=\begin{cases} m+1-i, & a=i, \\ n+1-\mu, & a=\mu, \end{cases} \quad \xi_a=\begin{cases} 1, & a=1 \\ (-1)^{\mu}, & a=\mu \end{cases}. \tag{II.3}$$

In the above equations, $i=1,2,...,m$ and $\mu=1,2,...,n$. Note that

$$\xi_a^2=1, \quad \xi_a\xi_{\bar{a}}=(-1)^{[a]}, \quad g^{ab}=\xi_b\delta_{a\bar{b}}. \tag{II.4}$$

As generators of the subalgebra $osp(m|n=2k)$, we take

$$\sigma_{ab}=g_{ac}E^c_b-(-1)^{[a][b]}g_{ac}E^c_a=-(-1)^{[a][b]}\sigma_{ba}, \tag{II.5}$$

which satisfy the graded commutation relations,

$$[\sigma_{ab},\sigma_{cd}]=g_{cb}\sigma_{ad}-(-1)^{([a]+[b])([c]+[d])}g_{ad}\sigma_{cb}-(-1)^{[c][d]}(g_{bd}\sigma_{ac} \\ -(-1)^{([a]+[b])([c]+[d])}g_{ac}\sigma_{db}). \tag{II.6}$$

We have an $osp(m|n)$-module decomposition,

$$gl(m|n)=osp(m|n)+T, \quad [T,T]\subset osp(m|n), \tag{II.7}$$

where T is spanned by operators

$$T_{ab}=g_{ac}E^c_b+(-1)^{[a][b]}g_{bc}E^c_a=(-1)^{[a][b]}T_{ba}. \tag{II.8}$$

It is convenient to introduce the Cartan–Weyl generators,

$$\sigma^a_b=g^{ac}\sigma_{cb}=-(-1)^{[a]([a]+[b])}\xi_a\xi_b\sigma^{\bar{b}}_{\bar{a}}. \tag{II.9}$$

As a Cartan subalgebra we take the diagonal operators,

$$\sigma^a_a=E^a_a-E^{\bar{a}}_{\bar{a}}=-\sigma^{\bar{a}}_{\bar{a}}. \tag{II.10}$$

Note that for odd $m=2h+1$ we have $\overline{h+1}=h+1$, and thus $\sigma^{h+1}_{h+1}=E^{h+1}_{h+1}-E^{h+1}_{h+1}=0$.

The positive roots of $osp(m|n)$ are given by the even positive roots [usual positive roots for $o(m)\oplus sp(n)$] together with the odd positive roots $\delta_\mu+\epsilon_i$, $1\leqslant i\leqslant m$, $1\leqslant\mu\leqslant k=n/2$, where we have adopted the useful convention $\epsilon_{\bar{i}}=-\epsilon_i$, $i\leqslant h=[m/2]$ so that $\epsilon_{h+1}=0$ for odd $m=2h+1$. This is consistent with the **Z** gradation,

$$osp(m|n)=L_{-2}\oplus L_{-1}\oplus L_0\oplus L_1\oplus L_2. \tag{II.11}$$

Here $L_0=o(m)\oplus gl(k)$; the $gl(k)$ generators are given by

$$\sigma^{\mu}_{\nu}=E^{\mu}_{\nu}-(-1)^{\mu+\nu}E^{\bar{\nu}}_{\bar{\mu}},\quad 1\leqslant\mu,\nu\leqslant k, \tag{II.12}$$

and $L_{-2}\oplus L_0\oplus L_2=o(m)\oplus sp(n)$, where L_2 gives rise to an irreducible representation of L_0 with highest weight $(\dot{0}|2,\dot{0})$ spanned by the generators

$$\sigma^{\mu}_{\bar{\nu}}=E^{\mu}_{\bar{\nu}}-\xi_{\mu}\xi_{\bar{\nu}}E^{\nu}_{\bar{\mu}}=E^{\mu}_{\bar{\nu}}+(-1)^{\mu+\nu}E^{\nu}_{\bar{\mu}},\quad 1\leqslant\mu,\nu\leqslant k. \tag{II.13}$$

Finally, L_1 is spanned by odd root space generators,

$$\sigma^{\mu}_{i}=E^{\mu}_{i}+\xi_{\mu}E^{\bar{i}}_{\bar{\mu}}=E^{\mu}_{i}+(-1)^{\mu}E^{\bar{i}}_{\bar{\mu}},\quad 1\leqslant\mu\leqslant k,\quad 1\leqslant i\leqslant m, \tag{II.14}$$

and gives rise to an irreducible representation of L_0 with highest weight $(1,\dot{0}|1,\dot{0})$. L_{-1}, L_{-2} give rise to irreducible representations of L_0 dual to L_1, L_2, respectively.

The simple roots of $osp(m|n=2k)$ are thus given by the usual (even) simple roots of L_0 together with the odd simple root $\alpha_s=\delta_k-\epsilon_1$, which is the lowest weight of L_0-module L_1. Note that the simple roots of $o(m)$ depend on whether m is odd or even, and are given here for convenience: For $m=2h$, $\alpha_i=\epsilon_i-\epsilon_{i+1}$, $1\leqslant i<h$, $\alpha_h=\epsilon_{h-1}+\epsilon_h$. For $m=2h+1$, $\alpha_i=\epsilon_i-\epsilon_{i+1}$, $1\leqslant i<h$, $\alpha_h=\epsilon_h$. The simple roots of $gl(k)$ are given by

$$\alpha_{h+\mu}=\delta_{gm}-\delta_{\mu+1},\quad 1\leqslant\mu<k. \tag{II.15}$$

The graded half-sum of the positive roots of $osp(m|n=2k)$ is given by

$$\rho=\frac{1}{2}\sum_{i=1}^{h}(m-2i)\epsilon_i+\frac{1}{2}\sum_{\mu=1}^{k}(n-m+2-2\mu)\delta_{\mu}. \tag{II.16}$$

III. GRADED-FERMION REALIZATIONS

We introduce the graded anticommutator:

$$\{A,B\}\equiv AB+(-1)^{[A][B]}BA. \tag{III.1}$$

Note that $\{A,B\}\neq\{B,A\}$. To realize the antisymmetric tensor irreducible representations of $gl(m|n)$, we introduce graded fermions c_a and their adjoints $c_a^{\dagger}$ obeying the graded anticommutation relations,

$$\{c_a,c_b\}=\{c_a^{\dagger},c_b^{\dagger}\}=0,\quad \{c_a,c_b^{\dagger}\}=\delta_{ab}. \tag{III.2}$$

Thus, when $a=i$ is even c_i are fermions while for $a=\mu$ odd, c_{μ} are bosons that anticommute with the fermions.

To get a graded fermion realization of $gl(m|n)$, we set

$$E^a_b=c_a^{\dagger}c_b, \tag{III.3}$$

and note the graded commutation relations:

$$[E^a_b,c_d^{\dagger}]=\delta_{bd}c_a^{\dagger},\quad [E^a_b,c_d]=(-1)^{([a]+[b])[d]}\delta^a_d c_b. \tag{III.4}$$

Using these relations, it is easy to verify that the operators E^a_b given above indeed satisfy the $gl(m|n)$ graded commutation relations.

Thus, we obtain representations of $gl(m|n)$ on the graded fermion Fock space, which include the antisymmetric tensor representations. The Fock space can be shown to be completely reducible into type I unitary irreducible representations of $gl(m|n)$ according to

$$F=\bigoplus_{a=0}^{m}\hat{V}(\dot{1}_a,\dot{0}|\dot{0})\bigoplus_{b=1}^{\infty}\hat{V}(\dot{1}|b,\dot{0}). \tag{III.5}$$

Thus, for $N\leqslant m$, the space of N-particle states comprises the antisymmetric tensor representation of $gl(m|n)$ with highest weight $\Lambda_N=(\dot{1}_N,\dot{0}|\dot{0})$. For $N>m$ the space of N-particle states comprises the irreducible representations of $gl(m|n)$ with highest weights $\Lambda_N=(\dot{1}|N-m,\dot{0})$.

We introduce an extra ''spin'' index α and consider the family of graded fermions $c_{a\alpha}$ and their adjoints $c^\dagger_{a\alpha}$ obeying the graded anticommutation relations,

$$\{c_{a\alpha},c_{b\beta}\}=\{c^\dagger_{a\alpha},c^\dagger_{b\beta}\}=0,\quad \{c_{a\alpha},c^\dagger_{b\beta}\}=\delta_{ab}\delta_{\alpha\beta}. \tag{III.6}$$

Here all spin indices are understood to be even (so that the grading only depends on the orbital labels a, b, c, etc.).

We take, for our $gl(m|n)$ generators,

$$E^a_b=\sum_\alpha c^\dagger_{a\alpha}c_{b\alpha}, \tag{III.7}$$

which can be shown, as before, to satisfy the graded commutation relations

$$[E^a_b,c^\dagger_{d\alpha}]=\delta_{bd}c^\dagger_{a\alpha},\quad [E^a_b,c_{d\alpha}]=(-1)^{([a]+[b])[d]}\delta^a_d c_{b\alpha}, \tag{III.8}$$

from which we deduce that the E^a_b indeed obey the $gl(m|n)$ graded commutation relations. Thus, we may now construct more general irreducible representations of $gl(m|n)$ in the graded-fermion Fock space. In particular, for ''two-column'' irreducible representations, only two spin labels $\alpha=\pm$ are required.

IV. QUASISPIN (TWO SPIN LABELS)

We employ the above graded-fermion algebra with two spin labels $\alpha=\pm$. We set

$$Q_+=g_{dd'}c^\dagger_{d,+}c^\dagger_{d',-}=\sum_d \xi_d c^\dagger_{d,+}c^{\dagger}_{\overline{d},-},$$
$$Q_-=g^{dd'}c^\dagger_{d,-}c^\dagger_{d',+}=\sum_d \xi_d c_{d,-}c_{\overline{d},+}. \tag{IV.1}$$

Let $Q_0=\frac{1}{2}(\hat{N}-m+n)$, where $\hat{N}=\Sigma_{a=1}^{m+n}E^a_a$ is the first-order invariant of $gl(m|n)$ (i.e., the number operator). By straightforward computation, the following can be shown.

Proposition 1: *$Q_\pm$, Q_0 generate an $sl(2)$ Lie algebra, called the quasispin Lie algebra,*

$$[Q_+,Q_-]=2Q_0,\quad [Q_0,Q_\pm]=\pm Q_\pm. \tag{IV.2}$$

Moreover, $Q_\pm$, Q_0 commute with the generators of $osp(m|n=2k)$.

To see the significance of the graded fermion algebra for the construction of irreducible representations, we set

$$E^{a\alpha}_{b\beta}=c^\dagger_{a\alpha}c_{b\beta}, \tag{IV.3}$$

and note the graded commutation relations,

$$[E^{a\alpha}_{b\beta},c^\dagger_{c\gamma}]=\delta_{bc}\delta_{\beta\gamma}c^\dagger_{a\alpha},\quad [E^{a\alpha}_{b\beta},c_{c\gamma}]=-(-1)^{[c]([a]+[b])}\delta^a_c\delta^\alpha_\gamma c_{b\beta}, \tag{IV.4}$$

from which we deduce

$$[E^{a\alpha}_{b\beta}, E^{c\gamma}_{d\delta}] = \delta^c_b \delta^\gamma_\beta E^{a\alpha}_{d\delta} - (-1)^{([a]+[b])([c]+[d])} \delta^a_d \delta^\alpha_\delta E^{c\gamma}_{b\beta}, \tag{IV.5}$$

which are the defining relations of $gl(2m|2n)$. That is, $E^{a\alpha}_{b\beta}$ are the generators of $gl(2m|2n)$.

As we have seen, the spin-averaged operators,

$$E^a_b = \sum_{\alpha=\pm} E^{a\alpha}_{b\alpha}, \tag{IV.6}$$

form the generators of $gl(m|n)$. Similarly, the orbital averaged operators,

$$E^\alpha_\beta = \sum_\alpha E^{a\alpha}_{a\beta}, \quad \alpha,\beta = \pm, \tag{IV.7}$$

form the generators of the spin Lie algebra $gl(2)$, which commute with the $gl(m|n)$ generators. It is worth noting that the spin $sl(2)$ algebra with generators,

$$S_+ = E^+_-, \quad S_- = E^-_+, \quad S_0 = \tfrac{1}{2}(E^+_+ - E^-_-), \tag{IV.8}$$

also commute with the quasispin Lie algebra. Throughout, we denote the spin Lie algebra (IV.8) by $sl_S(2)$ and the quasispin Lie algebra by $sl_Q(2)$.

Then, the space of N-particle states gives rise to an irreducible representation of $gl(2m|2n)$ [and $osp(2m|2n)$] with highest weight,

$$\begin{cases} (\dot{1}_N, \overline{0}|\dot{0}), & N \leqslant 2m \\ (\dot{1}|N-2m, \dot{0}), & N > 2m. \end{cases} \tag{IV.9}$$

This N-particle space decomposes into a multiplicity-free direct sum of irreducible $gl(m|n) \oplus sl_S(2)$ modules,

$$\hat{V}(a,b) \otimes V_s, \tag{IV.10}$$

where V_s denotes the $(2s+1)$-dimensional irreducible representation of $sl_S(2)$, $b=2s$, $N=2a+b$ and $\hat{V}(a,b)$ denotes the irreducible representation of $gl(m|n)$ with highest weight,

$$\Lambda_{a,b} = \begin{cases} (\dot{2}_a, \dot{1}_b, \dot{0}|\dot{0}), & a+b \leqslant m, \\ (\dot{2}_a, \dot{1}|a+b-m, \dot{0}), & a \leqslant m, a+b > m, \\ (\dot{2}|a+b-m, a-m, \dot{0}), & a > m. \end{cases} \tag{IV.11}$$

In this way we may realize all required "two-column" irreducible representations of $gl(m|n)$, inside a given antisymmetric tensor irreducible representation of $gl(2m|2n)$ utilizing the graded-fermion calculus.

V. CASIMIR INVARIANTS AND CONNECTION WITH QUASISPIN

From now on we shall use the notation

$$\hat{L} \equiv gl(m|n), \quad L \equiv osp(m|n), \quad \hat{L}_0 \equiv gl(m) \oplus gl(n), \quad L_o \equiv o(m) \oplus sp(n). \tag{V.1}$$

Let $C_{\hat{L}}$, C_L denote the universal Casimir invariants of $\hat{L}$, L, respectively. Then for the two-column irreducible representations of $\hat{L}$ we are considering, a straightforward but tedious calculation shows that

$$C_{\hat{L}} - C_L = (m-n+2-\tfrac{1}{2}\hat{N})\hat{N} - \tfrac{1}{2}(n-m)(n-m-2) + 2Q^2, \tag{V.2}$$

where

$$Q^2=\mathbf{Q}\cdot\mathbf{Q}=Q_0(Q_0+1)+Q_-Q_+=Q_0(Q_0-1)+Q_+Q_- \tag{V.3}$$

is the square of the quasispin. Equation (V2) shows that Q^2 is expressible in terms of $C_{\hat{L}}$, C_L, and $\hat{N}$. It follows that Q^2, Q_-Q_+, Q_+Q_- must leave invariant (in fact, reduce to a scalar multiple of the identity on) a given irreducible representation of L inside a given (two-column) representation of $\hat{L}$. Given the highest weight of such an L module we may determine its quasispin $\overline{Q}$ [the lowest weight of the relevant $sl_Q(2)$ module] using (V2) and $Q^2=\overline{Q}(\overline{Q}-1)$.

It is worth noting that we may write, for our quasispin generators,

$$\mathbf{Q}=\mathbf{Q}^{(0)}+\mathbf{Q}^{(1)}, \tag{V.4}$$

where

$$Q_-^{(0)}=\sum_{i=1}^{m} c_{i,-}c_{\bar{i},+}, \quad Q_-^{(1)}=\sum_{\mu=1}^{n}(-1)^{\mu}c_{\mu,-}c_{\bar{\mu},+}, \tag{V.5}$$

and, similarly, for Q_+, while

$$Q_0^{(0)}=\tfrac{1}{2}(\hat{N}_0-m), \quad Q_0^{(1)}=\tfrac{1}{2}(\hat{N}_1+n), \tag{V.6}$$

with $\hat{N}_0=\Sigma_{i=1}^m E_i^i$ and $\hat{N}_1=\Sigma_{\mu=1}^n E_\mu^\mu$ being the number operators for even fermions and odd bosons, respectively. Then it can be shown that $\mathbf{Q}^{(0)}$, $\mathbf{Q}^{(1)}$ both determine $sl(2)$ algebra that commute, so that the quasispin $\mathbf{Q}$ may be interpreted as the total quasispin obtained by coupling the quasispins of the even and odd components, respectively.

Similar remarks apply to the total spin algebra. The total spin vector is a sum of even and odd components,

$$\mathbf{S}=\mathbf{S}^{(0)}+\mathbf{S}^{(1)}, \tag{V.7}$$

whose corresponding $sl(2)$ algebras [cf. (IV.8)] are generated by

$$E^{(0)\alpha}{}_{\beta}=\sum_{i=1}^{m} E^{i\alpha}_{i\beta}, \quad E^{(1)\alpha}{}_{\beta}=\sum_{\mu=1}^{n} E^{\mu\alpha}_{\mu\beta}, \tag{V.8}$$

respectively. We note that the quasispin and spin algebras $sl_Q^{(0)}(2)$, $sl_Q^{(1)}(2)$, $sl_S^{(0)}(2)$, $sl_S^{(1)}(2)$ all commute with each other.

We remark that the quasispin algebras $sl_Q^{(0)}(2)$, $sl_Q^{(1)}(2)$ play an important role in decomposing irreducible representations of $\hat{L}_0$ into irreducible representations of $L_{\bar{0}}$. They commute with the even subalgebra $L_{\bar{0}}$ of L, but not with L itself.

VI. QUASISPIN EIGENVALUES

Throughout, $\hat{V}(a,b)$ denotes the irreducible representation of $\hat{L}$ with highest weight $\Lambda_{a,b}$ given by (IV.11). Let $\hat{V}_{\bar{0}}(a,b)=\hat{V}_0(\dot{0}|a+b,a,\dot{0})$ be its minimal $\mathbf{Z}$-graded component. Note that $\hat{V}_{\bar{0}}(a,b)$ is an irreducible $gl(n)$ module and thus an irreducible $\hat{L}_0$ module. We have the following.

Proposition 2: $\hat{V}_{\bar{0}}(a,b)$ cyclically generates $\hat{V}(a,b)$ as an L module: viz.,

$$\hat{V}(a,b)=U(L)\hat{V}_{\bar{0}}(a,b). \tag{VI.1}$$

Proof: Set

$$W = U(L)\hat{V}_0^-(a,b) \subset \hat{V}(a,b), \tag{VI.2}$$

i.e., W is an L submodule. We show that equality holds. Obviously, $\hat{V}_0^-(a,b)$ is an L_0^- module (since $L_0^- = L_{-2} \oplus L_0 \oplus L_2 \subset \hat{L}_0$). Now, since $\hat{V}_0^-(a,b)$ is the minimal **Z**-graded component of $\hat{V}(a,b)$, we have, by the PBW theorem,

$$\hat{V}(a,b) = U(\hat{L}_+)\hat{V}_0^-(a,b). \tag{VI.3}$$

Using

$$\sigma^i_\mu = E^i_\mu - (-1)^\mu E^{\bar{\mu}}_i \in L_1^- \equiv L_1 \oplus L_{-1}, \tag{VI.4}$$

we have

$$E^i_\mu \hat{V}_0^-(a,b) = \sigma^i_\mu \hat{V}_0^-(a,b) + (-1)^\mu E^{\bar{\mu}}_i \hat{V}_0^-(a,b) = \sigma^i_\mu \hat{V}_0^-(a,b) \subset W, \tag{VI.5}$$

since $E^{\bar{\mu}}_i \hat{V}_0^-(a,b) \subset \hat{L}_- \hat{V}_0^-(a,b) = (0)$. It follows that

$$\hat{L}_+ \hat{V}_0^-(a,b) \subset W. \tag{VI.6}$$

Proceeding recursively, let us assume that

$$(\hat{L}_+)^i \hat{V}_0^-(a,b) \subset W, \quad \forall i \leqslant r. \tag{VI.7}$$

Then

$$\begin{aligned} E^i_\mu \hat{L}^r_+ \hat{V}_0^-(a,b) &= \sigma^i_\mu \hat{L}^r_+ \hat{V}_0^-(a,b) + (-1)^\mu E^{\bar{\mu}}_i \hat{V}_0^-(a,b) \\ &\subset L\hat{L}^r_+ \hat{V}_0^-(a,b) + \hat{L}_- \hat{L}^r_+ \hat{V}_0^-(a,b) \\ &\subset L\hat{L}^r_+ \hat{V}_0^-(a,b) + \hat{L}^{r-1}_+ \hat{V}_0^-(a,b) \subset W, \end{aligned} \tag{VI.8}$$

since $\hat{L}_- \hat{V}_0^-(a,b) = (0)$ and $\hat{L}^r_+ \hat{V}_0^-(a,b) \subset W$, $\hat{L}^{r-1}_+ \hat{V}_0^-(a,b) \subset W$ by the recursion hypothesis. Thus $\hat{L}^{r+1}_+ \hat{V}_0^-(a,b) \subset W$ so that, by induction, $\hat{L}^r_+ \hat{V}_0^-(a,b) \subset W$, $\forall r$. It follows that

$$\hat{V}(a,b) = U(\hat{L}_+)\hat{V}_0(a,b) \subset W. \tag{VI.9}$$

Thus, we must have $W = \hat{V}(a,b)$.

From the traditional quasispin formalism for $gl(n) \supset sp(n)$, we have a decomposition of L_0^- modules,

$$\hat{V}_0^-(a,b) = V_0(a,b) \oplus Q_+^{(1)} \hat{V}_0^-(a-1,b), \tag{VI.10}$$

where $V_0(a,b)$ is an irreducible L_0^- module with highest weight $(\dot{0}|a+b,a,\dot{0})$ and comprises quasispin minimal states with respect to quasispin algebra $\mathbf{Q}^{(1)}$ (and thus also $\mathbf{Q}$), so

$$Q_-^{(1)} V_0(a,b) = Q_- V_0(a,b) = 0. \tag{VI.11}$$

Note that for $n=2$, $\hat{V}_0^-(a,b) = V_0(a,b)$ is an irreducible L_0^- module, but not quasispin minimal. Thus, the case $n=2$ requires a separate treatment. However, for this case, $\hat{V}_0^-(a,b) = V_0(a,b)$ still has well-defined quasispin $\overline{Q}$ (the minimal weight of the quasispin algebra): in fact, $\overline{Q} = \frac{1}{2}(b-m+n)$ for this case.

Proceeding recursively, we arrive at the irreducible $sp(n)$ (and hence L_0^-) module decomposition,

$$\hat{V}_0^-(a,b)=\bigoplus_{c=0}^{a} Q^{(1)a-c}_{+}V_0(c,b), \tag{VI.12}$$

where

$$Q^{(1)a-c}_{+}V_0(c,b)\cong V_0(c,b)\subset \hat{V}_0^-(c,b) \tag{VI.13}$$

is the irreducible L_0^- module with highest weight $(\dot{0}|c+b,c,\dot{0})$. From the above remarks $V_0(c,b)$ in the decomposition (VI.13) is quasispin minimal with respect to $\mathbf{Q}^{(1)}$ (and $\mathbf{Q}$) so $Q_-^{a-c+1}Q^{(1)a-c}_{+}V_0(c,b)=(0)$. It follows that $Q_-^{a+1}\hat{V}_0^-(a,b)=(0)$. Thus, if $q_N=\frac{1}{2}(N-m+n)$ is the eigenvalue of Q_0 on $\hat{V}(a,b)$, $N=2a+b$, then we have the following.

Theorem 1: *The quasispin eigenvalues (i.e., quasispin minimal weights) occurring in $\hat{V}(a,b)$ lie in the range*

$$\overline{Q}=q_N,\ \ q_N-1,...,\ \ q_N-a, \tag{VI.14}$$

or $q_N\geqslant\overline{Q}\geqslant q_N-a$ (in integer steps).

In view of (V.2) and (V.3), the operator Q_-Q_+ must leave invariant an L submodule of $\hat{V}(a,b)$. In view of the above theorem, the (generalized) eigenvalues of Q_-Q_+ on $\hat{V}(a,b)$ must be of the form

$$Q_-Q_+\equiv\overline{Q}(\overline{Q}-1)-q_N(q_N+1)=(\overline{Q}+q_N)(\overline{Q}-q_N-1). \tag{VI.15}$$

This eigenvalue can only vanish if $\overline{Q}+q_N=0$, which would imply, from the above theorem, $q_N-k=-q_N$ for some $0\leqslant k\leqslant a$. Thus, $k=2q_N=N-m+n$ or, equivalently, $a\geqslant N-m+n\Leftrightarrow a\geqslant 2a+b-m+n\Leftrightarrow m-n\geqslant a+b$.

Thus, if $m\leqslant n$, the (generalized) eigenvalues of Q_-Q_+ are all nonzero, except for the trivial module $(a=b=0)$, which we ignore below. Thus, we have proved the following lemma.

Lemma 1: For $m\leqslant n$, Q_-Q_+ determines a nonsingular operator on $\hat{V}(a,b)$, except possibly for the trivial module corresponding to $m=n$, $a=b=0$.

Remarks: The above result is crucial in what follows and will not generally hold for $m>n$. Hence, throughout the remainder we assume $m\leqslant n$, $n>2$. Note that Q_-Q_+ is nonsingular even on the trivial module, except when $m=n$.

VII. INDUCED FORMS AND AN ORTHOGONAL DECOMPOSITION

We recall that the graded fermion calculus admits a grade-$*$ operation, defined by

$$(c^{\dagger}_{a,\alpha})^*=(-1)^{[a]}c_{a,\alpha},\quad c^*_{a,\alpha}=c^{\dagger}_{a,\alpha}, \tag{VII.1}$$

which we extend in the usual way with $(AB)^*=(-1)^{[A][B]}B^*A^*$. This induces a grade-$*$ operation on $\hat{L}$ and L. Explicitly,

$$(E^a_b)^*=(-1)^{[a]([a]+[b])}E^b_a,\quad (\sigma^a_b)^*=(-1)^{[a]([a]+[b])}\sigma^b_a. \tag{VII.2}$$

Moreover, the quasispin generators satisfy $Q_+^*=Q_-$, $Q_-^*=Q_+$, and $Q_0^*=Q_0$.

With this convention, the graded fermion Fock space admits a nondegenerate graded sesquilinear form $\langle,\rangle$. In particular, $\hat{V}(a,b)$ is equipped with such a form and is nondegenerate. Note that

$$\langle v,E^a_bw\rangle=(-1)^{[v]([a]+[b])}\langle(E^a_b)^*v,w\rangle, \tag{VII.3}$$

which is the invariance condition of the form. It is the unique (up to scalar multiples) invariant graded form on $\hat{V}(a,b)$.

We now note that $Q_+\hat{V}(a-1,b)$ is an L submodule of $\hat{V}(a,b)$. In view of Lemma 1 and Eqs. (V.2) and (V.3), we have the following.

Lemma 2: The form $\langle\, , \rangle$ *restricted to* $Q_+\hat{V}(a-1,b)\subset\hat{V}(a,b)$ *is nondegenerate, except for the case* $a=1$, $b=m-n=0$.

Proof: Under the above conditions, Q_-Q_+ is nonsingular on $\hat{V}(a-1,b)$, so $Q_-Q_+\hat{V}(a-1,b)=\hat{V}(a-1,b)$. Hence, for $v\in\hat{V}(a-1,b)$, we have $0=\langle Q_+\hat{V}(a-1,b),Q_+v\rangle\Rightarrow 0=\langle Q_-Q_+\hat{V}(a-1,b),v\rangle=\langle\hat{V}(a-1,b),v\rangle\Rightarrow v=0$ since $\langle\, , \rangle$ on $\hat{V}(a-1,b)$ is nondegenerate. This shows that the form $\langle\, , \rangle$ restricted to $Q_+\hat{V}(a-1,b)$ is nondegenerate, as required.

In view of Proposition 2, we have the following.

Proposition 3: $Q_-\hat{V}(a,b)=\hat{V}(a-1,b)$.

Proof: From Proposition 2, we have

$$Q_-\hat{V}(a,b)=Q_-U(L)\hat{V}_0^-(a,b)=U(L)Q_-\hat{V}_0^-(a,b)=U(L)Q_-^{(1)}\hat{V}_0^-(a,b)=U(L)\hat{V}_0^-(a-1,b), \tag{VII.4}$$

where the last step follows from a classical Lie algebra result. Again, utilizing Proposition 2, we have $U(L)\hat{V}_0^-(a-1,b)=\hat{V}(a-1,b)$, from which the result follows.

We are now in a position to prove the following.

Proposition 4: We have an L-module orthogonal decomposition;

$$\hat{V}(a,b)=\mathcal{K}\oplus Q_+\hat{V}(a-1,b), \tag{VII.5}$$

where $\mathcal{K}=\mathrm{Ker}\, Q_-\cap\hat{V}(a,b)$, *except for the case* $a=1$, $b=m-n=0$.

Proof: For $v\in\hat{V}(a,b)$, $\langle v,Q_+\hat{V}(a-1,b)\rangle=0\Leftrightarrow\langle Q_-v,\hat{V}(a-1,b)\rangle=0\Leftrightarrow Q_-v=0$ (by Proposition 3) $\Leftrightarrow v\in\mathcal{K}$. Since $\langle\, , \rangle$ restricted to $Q_+\hat{V}(a-1,b)$ is nondegenerate, the result follows.

Finally, in view of Theorem 1 we have Proposition 5.

Proposition 5: $\hat{V}(a=0,b)$ *is an irreducible L module.*

Proof: In such a case, $\hat{V}_0^-(0,b)=V_0(0,b)$ is an irreducible L_0^- module cyclically generated by an L maximal state. Thus, $\hat{V}(0,b)=U(L)V_0(0,b)$ must be an indecomposable L module. Since the form $\langle\, , \rangle$ on $\hat{V}(0,b)$ is nondegenerate, this forces $\hat{V}(0,b)$ to be an irreducible L module.

The result above shows that the minimal $\hat{L}$ irreducible representations are indeed irreducible under L.

VIII. PRELIMINARIES TO BRANCHING RULES

It is our aim below to prove, barring the exceptional case of Lemma 2, that $\mathcal{K}$ is an irreducible L module. Note that the maximal state of the L_0^- module $V_0(a,b)$ occurring in the decomposition (VI.10), in fact, coincides with the $\hat{L}_0$ maximal vector v_+^Λ of $\hat{V}_0^-(a,b)$: For $n>2$ it can be seen directly that

$$Q_-v_+^\Lambda=Q_-^{(1)}v_+^\Lambda=0, \tag{VIII.1}$$

for this maximal vector. Moreover, for $n>2$ we have

$$E_{\bar{\mu}}^{i}v_+^\Lambda=0, \quad 1\leqslant i\leqslant m, \quad 1\leqslant\mu\leqslant k; \tag{VIII.2}$$

otherwise, this vector would have weight $(\dot{0}|a+b,a,\dot{0})+\epsilon_i-\delta_{\bar{\mu}}$ $(\overline{\mu}>k=n/2)$, which is impossible since all $\hat{L}$ weight components are positive. Also, since v_+^Λ belongs to the $\hat{L}$ minimal **Z**-graded component, we must have

$$E_i^{\mu} v_+^{\Lambda} = 0, \quad \forall i, \mu. \tag{VIII.3}$$

Thus, for $\sigma_i^{\mu} \in L_1$, we have

$$\sigma_i^{\mu} v_+^{\Lambda} = (E_i^{\mu} + (-1)^{\mu} E_{\overline{\mu}}^{\overline{i}}) v_+^{\Lambda} = 0, \quad \forall i, 1 \leq \mu \leq k \Rightarrow L_1 v_+^{\Lambda} = (0). \tag{VIII.4}$$

It follows that the L_0^- module $V_0(a,b)$ must cyclically generate an indecomposable module over L:

$$V(a,b) = U(L) V_0(a,b), \tag{VIII.5}$$

with highest weight

$$\lambda_{a,b} \equiv (\dot{0} | a+b, a, \dot{0}). \tag{VIII.6}$$

Since

$$Q_- V_0(a,b) = Q_-^{(1)} V_0(a,b) = (0), \tag{VIII.7}$$

we have

$$Q_- V(a,b) = Q_- U(L) V_0(a,b) = U(L) Q_- V_0(a,b) = (0). \tag{VIII.8}$$

It follows that $V(a,b) \subset \mathcal{K}$.

We now show that $V(a,b) = \mathcal{K}$ is irreducible. First, in view of Proposition 3, we have the following lemma.

Lemma 3: $v \in \mathcal{K} \Leftrightarrow Q_+ Q_- v = 0$.

Proof: Obviously $v \in \mathcal{K} \Rightarrow Q_- v = 0 \Rightarrow Q_+ Q_- v = 0$. Conversely, $Q_+ Q_- v = 0 \Rightarrow$

$$0 = \langle Q_+ Q_- v, \hat{V}(a,b) \rangle = \langle Q_- v, Q_- \hat{V}(a,b) \rangle = \langle Q_- v, \hat{V}(a-1,b) \rangle \tag{VIII.9}$$

$\Rightarrow Q_- v = 0 \Rightarrow v \in \mathcal{K}$.

It follows that $\mathcal{K}$ consists of eigenstates of $Q_+ Q_-$ with a zero eigenvalue. Also, since $Q_- \mathcal{K} = (0)$ and $\mathcal{K} \subset \hat{V}(a,b)$, it follows that all states in $\mathcal{K}$ are eigenvectors of Q_0 with eigenvalue $q_N = \frac{1}{2}(N - m + n)$ and are, moreover, quasispin minimal states, and so have quasispin $\overline{Q} = q_N$. Thus, Q^2 reduces to a scalar multiple $\overline{Q}(\overline{Q} - 1) = q_N(q_N - 1)$ on $\mathcal{K}$. It then follows from (V.2) that the universal Casimir element C_L of L must reduce to a scalar multiple of the identity on $\mathcal{K}$. Since $V(a,b) \subset \mathcal{K}$ has highest weight $\lambda_{a,b}$, this eigenvalue can be shown to be given by

$$\chi_{\lambda_{a,b}}(C_L) = (\lambda_{a,b}, \lambda_{a,b} + 2\rho) = -(a+b)(a+b+n-m) - a(a+n-m-2). \tag{VIII.10}$$

Hence we have proved the following.

Lemma 4: C_L reduces to a scalar multiple of the identity on $\mathcal{K}$ with an eigenvalue given by (VIII.10).

Now $\mathcal{K}$ is a completely reducible L_0^- module. Hence we have the following.

Lemma 5: Suppose for any irreducible L_0^- module $V_0(\lambda)$ contained in an irreducible $\hat{L}_0$ module $\hat{V}_0(\Lambda) \subset \hat{V}(a,b)$ that $\chi_\lambda(C_L) = \chi_{\lambda_{a,b}}(C_L) \Leftrightarrow \Lambda = \Lambda_{a,b}$ and $\lambda = \lambda_{a,b}$. Then $\mathcal{K} = V(a,b)$ is irreducible.

Proof: Indeed, in such a case it follows from Lemma 4 that the highest weight vector of $V(a,b)$ must be the unique primitive vector in $\mathcal{K}$. This is enough to prove that $\mathcal{K}$ is irreducible.

Finally, we recall that $\hat{V}(a,b)$ comprises states with total spin $s = b/2$ and with particle number $N = 2a + b$. Then the possible irreducible representations of $\hat{L}_0$ occurring in $\hat{V}(a,b)$ must have highest weights of the form

$$\Lambda=(\dot{2}_{a'},\dot{1}_{b'},\dot{0}|c',d',\dot{0}). \tag{VIII.11}$$

Then we must have

$$2a'+b'+c'+d'=N=2a+b. \tag{VIII.12}$$

Moreover, the total spins for the even and odd components of this irreducible representation are $s_0=b'/2$ and $s_1=(c'-d')/2$, respectively. So, using the triangular rule for angular momenta, we have

$$s\leqslant s_0+s_1,\quad s_0\leqslant s+s_1,\quad s_1\leqslant s+s_0, \tag{VIII.13}$$

or

$$b\leqslant b'+c'-d',\quad b'\leqslant b+c'-d',\quad c'-d'\leqslant b+b'. \tag{VIII.14}$$

These inequalities turn out to be important below.

IX. $\hat{L}\downarrow L$ BRANCHING RULES

We start this section with some facts concerning $\hat{L}_0\downarrow L_0^-$. The possible $\hat{L}_0$ highest weights Λ occurring in $\hat{V}(a,b)$ are of the form of (VIII.11). The possible L_0^- highest weights λ in $\hat{V}(a,b)$ are obtained from such Λ by a classical contraction procedure and have the form

$$\lambda=(\dot{2}_c,\dot{1}_d,\dot{0}|e,f,\dot{0}),\quad c+d\leqslant h, \tag{IX.1}$$

where $d=b'\wedge(m-2c-b')$, $e-f=c'-d'$ [here and below $x\wedge y\equiv\min(x,y)$] and

$$c\leqslant a',\quad e+f\leqslant c'+d'=2a+b-2a'-2b'. \tag{IX.2}$$

Note that for $n>4$, there are additional restrictions on the allowed L_0^- dominant weights in order that they give rise to highest weights of L.[8] In the interests of a unified treatment of all cases, including $n=4$, we do not impose these supplementary conditions here.

Since $e-f=c'-d'$, the inequalities (VIII.14) lead to

$$b'\leqslant b+e-f,\quad b\leqslant b'+e-f,\quad e-f\leqslant b+b'. \tag{IX.3}$$

Hence, we have the following inequalities.

Lemma 6: $e\leqslant a+b-c$, $f\leqslant a-c$.

Proof: We have

$$e+f\leqslant 2a+b-2a'-b',\quad e-f\leqslant b+b'.$$

Adding these two inequalities gives $e\leqslant a+b-a'$. Thus, $e\leqslant a-c$ since $c\leqslant a'$. Similarly, adding

$$e+f\leqslant 2a+b-2a'-b',\quad f-e\leqslant b'-b$$

leads to $f\leqslant a-a'\leqslant a-c$.

We are now in a position to compute the eigenvalue $\chi_\lambda(C_L)$ compared with that of (VIII.10). By direct computation we have

$$\chi_\lambda(C_L)=(\lambda,\lambda+2\rho)=m(2c+d)-c(c+1)-(c+d)(c+d+1)$$
$$-(n-m)(e+f)+4c+d+2f-e^2-f^2, \tag{IX.4}$$

where we have used

$$\lambda=\sum_{i=1}^{c} 2\epsilon_i+\sum_{i=c+1}^{d+c}\epsilon_i+e\delta_1+f\delta_2, \tag{IX.5}$$

together with the expression for ρ of L. By a straightforward but tedious calculation, using (VIII.10) and (IX.4), we obtain

$$\chi_\lambda(C_L)-\chi_{\lambda_{a,b}}(C_L)=2cn+d(m-d)+2c(2a+b-2c-d)+(a+b-c-e)\times(a+b-c+e+n-m)+(a-c-f)(a-c+f+n-m-2) \tag{IX.6}$$

$$=[2c(n+1)+2f-2a]+d(m-d)+2c(2a+b-2c-d)+(a+b-c-e)(a+b-c+e+n-m)+(a-c-f)(a-c+f+n-m). \tag{IX.7}$$

All terms on the rhs of (IX.6) are positive, in view of the inequalities given above, except possibly the last due to the term $(a-c+f+n-m-2)$. Similarly, in (IX.7) all terms on the rhs are positive, except possibly the first.

We proceed stepwise.

(i) $c\geqslant 1$: Then the first term on the rhs of (IX.7) gives

$$2c(n+1)+2f-2a\geqslant 2(n+1+f-a).$$

This leads to two subclasses.

(i.1) $a\leqslant n+1$: The rhs terms are all non-negative, so (IX.7) can only vanish if $a=n+1$, $f=0=d$, $2a+b=2c+d$. But then, since $d=0$ this would imply $2c=2a+b\Rightarrow c>a=n+1$, which is impossible since $c\leqslant h\leqslant m\leqslant n$. Thus we conclude that the rhs must be strictly positive in this case.

(i.2) $a\geqslant n+2$: In this case all terms on the rhs of (IX.6) are non-negative, including the last term, since, for the case at hand,

$$a-c+f+n-m-2\geqslant n+2-c+f+n-m-2\geqslant n-c+f+n-m\geqslant 0,$$

since $n\geqslant m\geqslant h\geqslant c$. Since $c\geqslant 1$, the rhs of (IX.6) must be strictly positive in this case.

We thus conclude, for $c\geqslant 1$, that $\chi_\lambda(C_L)-\chi_{\lambda_{a,b}}(C_L)>0$. It remains then to consider the case $c=0$, in which case we have

$$\chi_\lambda(C_L)-\chi_{\lambda_{a,b}}(C_L)=d(m-d)+(a+b-e)(a+b+e+n-m)+(a-f)(a+f+n-m-2). \tag{IX.8}$$

Note that for the case $c=0$, the inequalities of Lemma 6 reduce to $e\leqslant a+b$, $f\leqslant a$ and for the case at hand we have

$$e-f=c'-d', \quad d=b'\wedge(m-b').$$

It is convenient to treat the cases $m=n$ and $m<n$ separately.

(ii) $c=0$, $n>m$: Here we assume $a\geqslant 1$, since when $a=0$, $\hat{V}(a=0,b)$ is already known to be an irreducible L module, so the branching rule is trivial.

Under these assumptions all terms on the rhs of (IX.8) are non-negative, including the last, since

$$a+f+n-m-2\geqslant f+n-m-1\geqslant 0.$$

Note that this factor can only vanish when $a=1$, $f=0$, $n=m+1$. There are thus two possibilities to consider for vanishing of the rhs of (IX.8):

(ii.1) $d=0$, $e=a+b$, $f=a$: Since $c'+d'=2a+b-2a'-b'\geqslant e+f=2a+b$ and $c'-d'=e-f=b$, this implies that $a'=b'=0$, $c'=a+b$, $d'=a$, and $\lambda=\lambda_{a,b}$. So in this case $\Lambda=(\dot{0}|a+b,a,\dot{0})=\Lambda_{a,b}$ and $\lambda=\lambda_{a,b}$.

(ii.2) $d=0$, $e=a+b$, $f=0$, $a=1$, $n=m+1$: Then $c'+d'\geqslant e+f=a+b$. Since $a=1$, we thus have

$$2+b=N=2a'+b'+c'+d'\geqslant 2a'+b'+a+b=2a'+b'+1+b$$

$\Rightarrow 1\geqslant 2a'+b'\Rightarrow a'=0$ and $b'\leqslant 1$. In such a case we must have $d=b'\wedge(m-b')$ and since $d=0\Rightarrow b'=0$, or $m=b'=1\Rightarrow n=2$, which we ignore. Then $\Lambda=(\dot{0}|c',b',\dot{0})$ with $c'-b'=e-f=a+b=1+b$, which corresponds to states with spin $(1+b)/2$, which is impossible since all states in $\hat{V}(a,b)$ have spin $b/2$. Thus, this latter case cannot occur.

Thus we have shown, for all cases, that when $n>m$, $\mathcal{K}=V(a,b)$ must be an irreducible module with highest weight $\lambda_{a,b}$, using Lemma 4.

In view of Proposition 3 we thus have the L module decomposition,

$$\hat{V}(a,b)=V(a,b)\oplus Q_+\hat{V}(a-1,b). \tag{IX.9}$$

Since Q_-Q_+ is nonsingular, $Q_+\hat{V}(a-1,b)\cong\hat{V}(a-1,b)$. By repeated application of (IX.9), we arrive at the irreducible L module decomposition,

$$\hat{V}(a,b)=\bigoplus_{c=0}^{a} Q_+^{a-c}V(c,b). \tag{IX.10}$$

Hence we have proved the following theorem.

Theorem 2: $(n>m,n>2)$**:** *We have the irreducible L-module decomposition,*

$$\hat{V}(a,b)=\bigoplus_{c=0}^{a} V(c,b). \tag{IX.11}$$

We emphasize that throughout $V(a,b)$ denotes the L module with highest weight $\lambda_{a,b}=(\dot{0}|a+b,a,\dot{0})$. It remains now to consider the case $m=n$, which is somewhat more interesting.

(iii) $c=0$, $m=n>2$: Again, we assume $a\geqslant 1$ since $\hat{V}(a=0,b)$ is an irreducible L module, as we have seen. We recall for the case at hand $e\leqslant a+b$, $f\leqslant a$, $a\geqslant 1$, $m=n>2$, $e-f=c'-d'$, $d=b'\wedge(m-b')$ and

$$\chi_\lambda(C_L)-\chi_{\lambda_{a,b}}(C_L)=d(m-d)+(a+b-e)(a+b+e)+(a-f)(a+f-2). \tag{IX.12}$$

There are now several cases to consider for the vanishing of (IX.12).

(iii.1) $a=f$: Then (IX.12) vanishes when $d=0$, $e=a+b$. Thus

$$c'+d'\geqslant e+f=2a+b=2a'+b'+c'+d'$$

$\Rightarrow a'=b'=0$, $c'+d'=2a+b$, and $c'-d'=e-f=b$. This corresponds to $\Lambda=\Lambda_{a,b}$ and $\lambda=\lambda_{a,b}$.

(iii.2) $f=2-a$: Then (IX.12) vanishes when $d=0$, $e=a+b$. Since $a\geqslant 1$ there are two cases.

(iii.2.1) $f=0$, $a=2$: This is only possible when $c'+d'\geqslant e+f=a+b\Rightarrow$

$$2a+b\geqslant 2a'+b'+c'+d'\geqslant 2a'+b'+a+b$$

$\Rightarrow a\geqslant 2a'+b'$ or $2\geqslant 2a'+b'$. This leads to two further cases.

(iii.2.1a) $f=0$, $a=2$, $a'=0$, $b'\leqslant 2$: In view of the contraction procedure, this is only consistent with $d=0$ if $b'=0$ (so $c'=a+b$, $d'=a$) or if $b=2$ and $m=n=2$. The latter case is being ignored and the former case cannot occur since then $c'-d'=e-f=a+b>b$ in contradiction to the fact that all states in $\hat{V}(a,b)$ have spin $b/2$.

(iii.2.1b) $f=0$, $a=2$, $a'=1$, $b'=d=0$: Then $c'-d'=e-f=a+b>b$, which again is impossible since all states have spin $b/2$.

(iii.2.2) $f=a=1$: Then $c'-d'=a+b-a=b$, $c'+b'\geqslant e+f=2a+b \Rightarrow a'=b'=0$, $c'=a+b$, $d'=a \Rightarrow \Lambda=\Lambda_{a,b}$, $\lambda=\lambda_{a,b}$.

(iii.3) $a+f-2<0$, $a>f$: This can only occur when $a=1$, $f=0$, in which case the rhs of (IX.12) becomes

$$d(m-d)+(a+b+e)(a+b-e)-1.$$

There are two cases for the vanishing of this.

(iii.3.1) $e=a+b$, $d=1$, $m=2$, which can occur, but we are ignoring since $n=m>2$.

(iii.3.2) $d=f=e=b=0$: Then $c'-d'=e-f=0$ and

$$N=2=2a+b=2a'+b'+c'+d'=2(a'+c')+b',$$

which can occur in the following cases:

$$a'=b'=0, \quad c'=d'=1 \Rightarrow \lambda=(\dot{0}|\dot{0}), \quad \Lambda=(\dot{0}|1,1,\dot{0});$$

$$b'=c'=d'=0, \quad a'=1 \Rightarrow \lambda=(\dot{0}|\dot{0}), \quad \Lambda=(2,\dot{0}|\dot{0}).$$

This exhausts all possibilities. It follows from the above that for $n=m>2$ the rhs of (IX.12) is always strictly positive and can only vanish in the last case, corresponding to $a=1$ and $b=0$. This is the irreducible representation $\hat{V}(2,\dot{0}|\dot{0})$ of $gl(n|n)$, which is known to give rise to an indecomposable $osp(n|n)$ module with a composition series of length 3 whose factors are isomorphic to the $osp(n|n)$ modules $V(1,0)$ and $V(0,0)$ (see Appendix).

Thus we have proved the decomposition

$$\hat{V}(a,b)=V(a,b)\oplus Q_+\hat{V}(a-1,b) \tag{IX.13}$$

with $V(a,b)$ an irreducible L-module of highest weight $\lambda_{a,b}$, provided $(a,b)\neq(1,0)$. Proceeding recursively we have the following theorem.

Theorem 3 (n=m>2): For $b>0$ we have the irreducible L-module decomposition,

$$\hat{V}(a,b)=\bigoplus_{c=0}^{a} V(c,b). \tag{IX.14}$$

For $b=0$ we have the L-module decomposition,

$$\hat{V}(a,0)=\bigoplus_{c=1}^{a} V(c,0), \tag{IX.15}$$

where $V(c,0)$ is irreducible for $c>1$ but $V(1,0)$ is indecomposable with a composition series of length 3 with composition factors isomorphic to irreducible L modules $V(1,0)$ and $V(0,0)$, the latter occurring twice.

Theorems 2 and 3 are our main results in this section concerning the $\hat{L}\downarrow L$ branching rules for the two-column tensor representations of $\hat{L}$. We remark that for the special case $n-m=0=b$, $a=1$, $\hat{V}(a-1,b)=\hat{V}(0,0)$ coincides with the identity module, which is the exceptional case of Lemma 2. For this case the form $\langle , \rangle$ on $\hat{V}(a,b)=\hat{V}(1,0)$ is degenerate on $Q_+\hat{V}(a-1,b)$

$=Q_+\hat{V}(0,0)$. Thus, Proposition 4 fails in this case (and only this case). This, of course, agrees with the result that $\hat{V}(a,b)=\hat{V}(1,0)\equiv\hat{V}(2,\dot{0}|\dot{0})$ is indecomposable for $m=n$.

ACKNOWLEDGMENTS

This paper was completed when YZZ visited Northwest University, China. He thanks the Australian Research Council IREX program for an Asia-Pacific Link Award and Institute of Modern Physics of the Northwest University for hospitality. The financial support from Australian Research Council large, small and QEII fellowship grants is also gratefully acknowledged.

APPENDIX: STRUCTURE OF $\hat{V}(z,\dot{0}|\dot{0})$ AS A $osp(n|n)$-MODULE

Here for completeness we determine the structure of the irreducible $\hat{L}=gl(n|n=2k)$ module $\hat{V}(2,\dot{0}|\dot{0})$ as a module over $L=osp(n|n)$, in fully explicit form.

First $\hat{V}(2,\dot{0}|\dot{0})$ admits the following **Z**-graded decomposition into irreducible $\hat{L}_0$ modules with highest weights shown:

$$\hat{V}(2,\dot{0}|\dot{0})=\hat{V}_0(2,\dot{0}|\hat{0})\oplus\hat{V}_1(1,\dot{0}|1,\dot{0})\oplus\hat{V}_2(\dot{0}|1,1,\dot{0}).$$

In the notation of the paper, the last space corresponds to the irreducible $\hat{L}_0$ module $\hat{V}_0^-(a=1,b=0)$. In terms of the graded fermion formalism, we have the following basis states:

$$\begin{aligned}
&\hat{V}_0(2,\dot{0}|\dot{0}):\ (c_{i,+}^\dagger c_{j,-}^\dagger+c_{j,+}^\dagger c_{i,-}^\dagger)|0\rangle,\quad 1\leqslant i,j\leqslant n,\\
&\hat{V}_1(1,\dot{0}|1,\dot{0}):\ (c_{i,+}^\dagger c_{\mu,-}^\dagger+c_{\mu,+}^\dagger c_{i,-}^\dagger)|0\rangle,\quad 1\leqslant i,\mu\leqslant n,\\
&\hat{V}_2(\dot{0}|1,1\dot{0}):\ (c_{\mu,+}^\dagger c_{\nu,-}^\dagger-c_{\nu,+}^\dagger c_{\mu,-}^\dagger)|0\rangle,\quad 1\leqslant\mu,\nu\leqslant n,
\end{aligned}\tag{A1}$$

where $|0\rangle$ is the vacuum state. The latter space decomposes into L_0^- modules according to

$$\hat{V}_2(\dot{0}|1,1,\dot{0})=V_0(\dot{0}|1,1\dot{0})\oplus V_0(\dot{0}|\dot{0}),$$

where $V_0(\dot{0}|\dot{0})$ is spanned by $Q_+^{(1)}|0\rangle$ (the trivial L_0^- module) and $V_0(\dot{0}|1,1,\dot{0})$ is an irreducible L_0^- module with the highest weight indicated and the following basis vectors:

$$(c_{\mu,+}^\dagger c_{\nu,-}^\dagger-c_{\nu,+}^\dagger c_{\mu,-}^\dagger)|0\rangle,\quad 1\leqslant\nu\neq\overline{\mu}\leqslant n,\tag{A2}$$

$$(\Omega_\mu^\dagger-\Omega_{\mu+1}^\dagger)|0\rangle,\quad 1\leqslant\mu<k,\tag{A3}$$

where

$$\Omega_\mu^\dagger\equiv c_{\mu,+}^\dagger c_{\overline{\mu},-}^\dagger-c_{\overline{\mu},+}^\dagger c_{\mu,-}^\dagger\,.$$

Note that this irreducible L_0^- module cyclically generates an indecomposable L module $\tilde{V}(\delta_1+\delta_2)$ with highest weight $\delta_1+\delta_2$ and highest weight vector given by (A2) with $\mu=1$, $\nu=2$.

Now $\hat{V}_1(1,\dot{0}|1,\dot{0})$ is also irreducible as an L_0^- module that is contained in $\tilde{V}(\delta_1+\delta_2)$. Then by applying the odd lowering generators $\sigma_\mu^i=E_\mu^i-(-1)^\mu E_i^{\overline{\mu}}(1\leqslant\mu\leqslant k,1\leqslant i\leqslant n)$ of L to the states (A1), the following states in $\hat{V}_0(2,\dot{0}|\dot{0})$ are easily seen to be in $\tilde{V}(\delta_1+\delta_2)$:

$$(c_{i,+}^\dagger c_{j,-}^\dagger+c_{j,+}^\dagger c_{i,-}^\dagger)|0\rangle,\quad 1\leqslant j\neq\overline{i}\leqslant n,\tag{A4}$$

$$(\Omega_i^\dagger-\Omega_{i+1}^\dagger)|0\rangle,\quad 1\leqslant i<k,\tag{A5}$$

where

$$\Omega_i^\dagger \equiv c_{i,+}^\dagger c_{,-}^\dagger + c_{\overline{i,+}}^\dagger c_{i,-}^\dagger .$$

Further, the following states are also seen to be in $\tilde{V}(\delta_1+\delta_2)$:

$$(\Omega_i^\dagger + (-1)^\mu \Omega_\mu^\dagger)|0\rangle, \quad 1 \leqslant i, \mu < k, \tag{A6}$$

which follows by applying $\sigma_i^{\overline{\mu}}$ to the states (A1) with $1 \leqslant \mu \leqslant k$. Summing (A6) on $\mu = i$ from 1 to k, we thus obtain

$$\left(\sum_{i=1}^{k} \Omega_i^\dagger + \sum_{\mu=1}^{k} (-1)^\mu \Omega_\mu^\dagger \right) |0\rangle = Q_+|0\rangle \in \tilde{V}(\delta_1+\delta_2). \tag{A7}$$

It is worth noting that the states (A6) are expressible in terms of the states (A3), (A5), and (A7).

The states (A1)–(A7) form a basis for the standard cyclic L module $\tilde{V}(\delta_1+\delta_2)$. We note that $\dim \tilde{V}(\delta_1+\delta_2) = \dim \hat{V}(2,\dot{0}|\dot{0}) - 1$ and $\tilde{V}(\delta_1+\delta_2)$ is the unique maximal L submodule of $\hat{V}(2,\dot{0}|\dot{0})$. In view of (A7), this module is not irreducible since it contains the trivial one-dimensional L module $V(\dot{0}|\dot{0})$ as a unique submodule.

The remaining state in $\hat{V}(2,\dot{0}|\dot{0})$, not in $\tilde{V}(\delta_1+\delta_2)$, is $Q_+^{(1)}|0\rangle$ (or $Q_+^{(0)}|0\rangle$), which thus generates the basis vector for the L factor module $\hat{V}(2,\dot{0}|\dot{0})/\tilde{V}(\delta_1+\delta_2)$, which is obviously isomorphic to the trivial L module $V(\dot{0}|\dot{0})$. We thus arrive at the L-module composition series $\hat{V}(2,\dot{0}|\dot{0}) \supset \tilde{V}(\delta_1+\delta_2) \supset V(\dot{0}|\dot{0}) \supset (0)$ with corresponding factors isomorphic to the irreducible L modules with highest weights $(\dot{0}|\dot{0})$, $\delta_1+\delta_2$, and $(\dot{0}|\dot{0})$, respectively.

This result is of importance to the explicit construction of new R matrices.[4] In particular, it gives rise to an L-invariant nilpotent contribution to the R matrices, a new effect not seen in the untwisted or nonsuper cases.

[1] G. W. Delius, M. D. Gould, and Y.-Z. Zhang, Int. J. Mod. Phys. A **11**, 3415 (1996).
[2] G. M. Gandenberger, N. J. MacKay, and G. M. T. Watts, Nucl. Phys. B **465**, 329 (1996).
[3] J. Van der Jeugt, J. Math. Phys. **37**, 4176 (1996).
[4] M. D. Gould and Y.-Z. Zhang, math-QA/9905021, Nucl. Phys. B, in press.
[5] M. D. Gould, J. R. Links, I. Tsohantjis, and Y.-Z. Zhang, J. Phys. A **30**, 4313 (1997).
[6] M. J. Martins and P. B. Ramos, Phys. Rev. B **56**, 6376 (1997).
[7] H. Saleur, ''The long delayed solution of the Bukhvostov–Lipatov model,'' e-print hepth/9811023.
[8] V. G. Kac, Lect. Notes Math. **676**, 597 (1978).

Construction of relativistic quantum fields in the framework of white noise analysis

Martin Grothaus[a)]
BiBoS, Universität Bielefeld, D 33615 Bielefeld, Germany, and Inst. Ang. Math., Universität Bonn, D 53115, Bonn, Germany

Ludwig Streit
BiBoS, Universität Bielefeld, D 33615 Bielefeld, Germany, and CCM, Universidade da Madeira, P 9000 Funchal, Portugal

(Received 9 February 1999; accepted for publication 6 May 1999)

We construct a class of Euclidean invariant distributions Φ_H indexed by a function H holomorphic at zero. These generalized functions can be considered as generalized densities w.r.t. the white noise measure, and their moments fulfill all Osterwalder–Schrader axioms, except for reflection positivity. The case where $F(s)=-(H(is)+\frac{1}{2}s^2)$, $s\in\mathbb{R}$, is a Lévy characteristic is considered in Rev. Math. Phys. **8**, 763 (1996). Under this assumption the moments of the Euclidean invariant distributions Φ_H can be represented as moments of a generalized white noise measure P_H. Here we enlarge this class by convolution with kernels G coming from Euclidean invariant operators $\mathcal{G}$. The moments of the resulting Euclidean invariant distributions Φ_H^G also fulfill all Osterwalder–Schrader axioms except for reflection positivity. For no nontrivial case we succeeded in proving reflection positivity. Nevertheless, an analytic extension to Wightman functions can be performed. These functions fulfill all Wightman axioms except for the positivity condition. Moreover, we can show that they fulfill the Hilbert space structure condition and therefore the modified Wightman axioms of indefinite metric quantum field theory [*Dynamics of Complex and Irregular Systems* (World Scientific, Singapore, 1993)].
 [S0022-2488(99)02608-0]

I. INTRODUCTION

This paper is motivated by the Euclidean strategy for constructing interacting field theories; see, e.g., Refs. 1 and 2 and the references therein. Formally, the interacting field theory with interaction V lives on the same measure space as the Euclidean free field μ_0 but has measure

$$d\mu_V=\frac{\exp(-\int_{\mathbb{R}^d}V(\phi(x))d^d(x))d\mu_0}{\int \exp(-\int_{\mathbb{R}^d}V(\phi(x))d^d(x))d\mu_0},\tag{1}$$

where $\phi(x)$ is a Gaussian random process at the point $x\in\mathbb{R}^d$, $d\in\mathbb{N}$. Since $\phi(x)$ in general is not an integrable function but rather a generalized function the question of how to define $V(\phi(x))$ leads to the problem of defining powers of $\phi(x)$. Furthermore, (1) is only formal if $\mu_V\neq\mu_0$, since the only probability measure absolutely continuous w.r.t. μ_0 and invariant under Euclidean translations is μ_0 itself.

A first step in the direction of giving sense to (1) is to construct Wick powers: $\phi(x)^m$: $m\in\mathbb{N}$, of the Gaussian random process ϕ at the point $x\in\mathbb{R}^d$. For $d\geqslant 2$ the Wick powers: $\phi(x)^m$: still are not integrable functions. Additionally, a so-called space cutoff is necessary, i.e., the integration in (1) is performed only over a bounded subset of $\mathbb{R}^d$, and sometimes also an ultraviolet cutoff, i.e., the Wick powers: $\phi(x)^m$: are smeared out with delta sequences. For certain

[a)]Electronic mail: grothaus@wiener.iam.uni.bonn.de

classes of interactions by these renormalizations integrable densities w.r.t. μ_0 have been constructed, and then some kind of limit that removes the cutoffs has been taken; a limit that does not require the output to be absolutely continuous. Examples are the following: the $P(\phi)_2$ model (the 2 stands for $d=2$), where the interaction is given by a Wick ordered polynomial $V=:P:$, semibounded from below, see, e.g., Ref. 1 and the references therein; the Høegh–Krohn model[3] in $d=2$ space–time dimensions, where the interaction is given by

$$V(s)=\int :\exp(as):d\nu(a),$$

where ν is a finite measure with compact support in the interval $(\sqrt{2\pi},\sqrt{2\pi})$; and the Albeverio Høegh–Krohn model[4] in d space–time dimensions, where the interaction is given by the Fourier transform of a measure ν with bounded support on the real line [and $d\nu(s)=\overline{d\nu(-s)}$], i.e.,

$$V(s)=\int \exp(ias)d\nu(a).$$

The Schwinger functions associated to the interacting field theory with interaction V are the moments of the measure ν_V. But moments one can also obtain from generalized functions considered as generalized densities w.r.t. a Gaussian measure μ, they only have to have the property that monomials are test functions. This is the basic idea of our approach. Motivated by the Euclidean strategy we consider the following generalized white noise functional:

$$\Phi_H=\exp^{\diamond}\left(-\int_{\mathbb{R}^d}H^{\diamond}(\phi(x))d^dx\right). \tag{2}$$

We assume that the function H is holomorphic at zero and $H(0)=0$. The Wick analytic function $H^{\diamond}(\phi(x))$ of the Gaussian process ϕ at the point $x\in\mathbb{R}^d$ coincides with the usual Wick ordered function $:H(\phi(x)):$. It turns out that $H^{\diamond}(\phi(x))$ is a generalized function from the Kondratiev space $(S)^{-1}$; see Sec. II B, and therefore also its integral, if it exists, is in $(S)^{-1}$. Thus, in general, we cannot take its exponential. But in the white noise distribution space $(S)^{-1}$ there exists the so-called Wick calculus; see Sec. II B; hence we can take its Wick exponential. In the case where H is linear and if we integrate only over $K\subset\mathbb{R}^2$, K compact (space cutoff), the function Φ_H is square integrable and we have a direct correspondence between (1) and (2), i.e.,

$$\Phi_H=\frac{\exp(-\int_K H(\phi(x))d^2(x))}{\int\exp(-\int_K H(\phi(x))d^2(x))d\mu},$$

where μ is the Gaussian white noise measure. In general, however, there is no need for the distribution Φ_H to be positive and for a large class of functions H there exists no measure that is representing Φ_H. It turns out that Φ_H can be represented by a measure if and only if the function $F(s)=-H(is)+\frac{1}{2}s^2$, $s\in\mathbb{R}$, is a Lévy characteristic; see Remark III.7 (ii). The associated measures are called generalized white noise measures.

Generalized white noise measures have been considered in Ref. 5. There are authors constructed Euclidean random fields over $\mathbb{R}^d$ by convoluting generalized white noise with integral kernels G coming from Euclidean invariant operators. The corresponding moments satisfy all Osterwalder–Schrader axioms[6] except for reflection positivity.

For all convoluted generalized white noise measures such that the Lévy characteristic of the generalized white noise measure has a holomorphic extension at zero, we can give an explicit formula for the generalized density w.r.t. the white noise measure; see Theorem III.9 below. Furthermore, there exists a large class of generalized functions Φ_H as in (2) that do not have an associated measure; see Remark III.16. In Theorem III.9 and Theorem III.15 we prove that the Schwinger functions corresponding to the convoluted generalized functions Φ_H^G also fulfill all Osterwalder–Schrader axioms except for reflection positivity.

For no nontrivial case we succeeded in proving reflection positivity. In Ref. 5 the authors present a partial negative result on reflection positivity for the Schwinger functions corresponding to moments of convoluted generalized white noise. We quote more details about their results in Sec. IV A.

Without reflection positivity we cannot perform the analytic continuation to Wightman functions via the reconstruction theorem.[6] Nevertheless, an analytic continuation can be performed. Using results from the theory of Laplace transforms in Ref. 5, the authors analytically continued the Schwinger functions, which are given as moments of convoluted generalized white noise to Wightman functions. In general, these functions only fulfill a part of the Wightman axioms, i.e., positivity (positive definiteness of the set of Wightman functions[7–9]) is missing. We generalized their idea to our case and in Theorem IV.1 we prove that the Schwinger functions corresponding to convoluted generalized functions Φ_H^G also have an analytic extension to Wightman functions. These Wightman functions fulfill all Wightman axioms, except for the positivity property. Furthermore, they fulfill the strong spectral condition with mass gap $m_0 > 0$, and their two-point functions admit a Källen–Lehmann representation. For the Fourier transform of the truncated Wightman functions in Ref. 5, the authors found explicit formulas. Using these formulas and the Jost–Schroer theorem, in Theorem IV.2 we prove a negative result concerning the positivity property; see also Remark IV.3.

Since the appearance of gauge theories, it has become natural to consider (local) quantum field theories (QFT) in which not all Wightman axioms are satisfied. Such a consideration has, in particular, been natural and also necessary for the study of ‘‘charged’’ fields interacting with gauge fields, because their description conflicts either with locality or with positivity. The physical reason for this is that in such theories one must use observables of the charged type that obey a Gaussian law; see, e.g., Morchio and Strocchi,[10] instead of using the usual local observables. Actually, from the study of fields such as, e.g., α-gauge-type Higgs models that do not satisfy positivity; see, e.g., Ref. 11 and references therein, it turns out that it is preferable to keep the locality condition and to give up the positivity condition. This leads to the so-called modified Wightman axioms of indefinite metric QFT.[12] The difference between indefinite metric QFT and standard QFT is that the axiom of positivity in the latter is replaced by the so-called Hilbert space structure condition in the former that permits the construction of a Hilbert space and a quantum field associated to a given collection of functions fulfilling the modified Wightman axioms.

In Ref. 13 the authors proved that the Wightman functions that are analytic continuations of the moments of convoluted generalized white noise fulfill the Hilbert space structure condition, and therefore the modified Wightman axioms. Again it was possible to generalize their proof to our case, and in Theorem IV.5 we prove that the Wightman functions that are analytic continuations of the moments of convoluted generalized functions Φ_H^G also fulfill the modified Wightman axioms.

The article is organized as follows. In Sec. II we introduce the concepts of Gaussian and white noise analysis[14–19] as far as necessary for our considerations. In the framework of white noise analysis, various aspects of QFT have been discussed.[20–23,16] Section III of this paper is intended to represent Euclidean QFT in the framework of white noise analysis. In Sec. III A we show how to check the Osterwalder–Schrader axioms (OS axioms) in terms of the T transform (the T transform is an infinite-dimensional generalization of the Fourier transform). The T transform of a generalized function is the generating functional of the corresponding Schwinger functions. Properties of generating functionals have also been discussed in Refs. 24 and 25. Having this tool in hands in Sec. III B, we construct the Euclidean invariant distributions Φ_H^G. In Sec. IV we discuss the reflection positivity, analytic continuation, and QFT with an indefinite metric.

II. GAUSSIAN ANALYSIS

A. Gaussian spaces

We start by considering the Gel’fand triple,

$$S(\mathbb{R}^d) \subset \mathcal{H} \subset S'(\mathbb{R}^d),$$

where $S(\mathbb{R}^d)$ is the space of rapidly decreasing, smooth test functions on $\mathbb{R}^d$. We assume $S(\mathbb{R}^d)$ to be equipped with its standard locally convex topology such that it is a nuclear space. $\mathcal{H}$ is a real separable Hilbert space containing $S(\mathbb{R}^d)$ as a dense and topological subspace. For instance, $\mathcal{H}$ can be chosen as the space of real-valued square integrable functions w.r.t. the Lebesgue measure on $\mathbb{R}^d$ or as a Sobolev space on $\mathbb{R}^d$. As is well known[26,27] $S=S(\mathbb{R}^d)$ is the projective limit of a family of Hilbert spaces $(\mathcal{H}_p)_{p\in\mathbb{N}_0}$, $\mathcal{H}_0=\mathcal{H}$, such that for all $p_1,p_2\in\mathbb{N}$ there exists $p\in\mathbb{N}$, such that $\mathcal{H}_p\subset\mathcal{H}_{p_1}$ and $\mathcal{H}_p\subset\mathcal{H}_{p_2}$ and the embeddings are of the Hilbert–Schmidt class. That is, S is a countably Hilbert space in the sense of Ref. 28. The dual space space S' is the space of tempered distributions. It is given as the inductive limit of the spaces $(\mathcal{H}_{-p})_{p\in\mathbb{N}_0}$ that are dual to the spaces $(\mathcal{H}_p)_{p\in\mathbb{N}}$ w.r.t. $\mathcal{H}$. We denote by $\langle\cdot,\cdot\rangle$ the dual pairings between $\mathcal{H}_p$ and $\mathcal{H}_{-p}$ and between S and S' given by the extension of the inner product $(\cdot,\cdot)$ on $\mathcal{H}$. Furthermore, $|\cdot|_p$ denote the norms on $\mathcal{H}_p$ and $\mathcal{H}_{-p}$, respectively, and we preserve this notation for the norms on the complexifications $\mathcal{H}_{p,\mathbb{C}}$ and $\mathcal{H}_{-p,\mathbb{C}}$ and tensor powers of these spaces.

Additionally, we introduce the notion of symmetric tensor power of the nuclear space S. The simplest way to do this is to start from usual symmetric tensor powers $\mathcal{H}_p^{\hat{\otimes}n}, n\in\mathbb{N}$, of Hilbert spaces. Using the definition

$$S^{\hat{\otimes}n}:=\underset{p\in\mathbb{N}}{\mathrm{prlim}}\ \mathcal{H}_p^{\hat{\otimes}n},$$

one can prove[26,27] that $S^{\hat{\otimes}n}$ is a nuclear space that is called the nth symmetric tensor power of S. The dual space $S'^{\hat{\otimes}n}$ can be written as

$$S'^{\hat{\otimes}n}=\underset{p\in\mathbb{N}}{\mathrm{indlim}}\ \mathcal{H}_{-p}^{\hat{\otimes}n}.$$

The space $S'(\mathbb{R}^d)^{\hat{\otimes}n}$ is canonically isomorphic to $S'(\mathbb{R}^{nd})$, the space of symmetric tempered distributions on $\mathbb{R}^{nd}$. All the results quoted above also hold for complex spaces.

In order to introduce a probability measure on the vector space S', we consider the σ-algebra $\mathcal{C}_\sigma(S')$ generated by cylinder sets. The canonical Gaussian measure μ on $(S',\mathcal{C}_\sigma(S'))$ is given by its characteristic function

$$\int_{S'}\exp(i\langle\omega,f\rangle)\,d\mu(\omega)=\exp\left(-\frac{1}{2}|f|^2\right),\quad f\in S,$$

via Minlos' theorem.[14–16] If we chose $\mathcal{H}=L^2(\mathbb{R}^d)$, the space of real-valued square-integrable functions w.r.t. the Lebesgue measure on $\mathbb{R}^d$, this is the Gaussian white noise measure. For $\mathcal{H}=H^{-1,2}(\mathbb{R}^d)$, the Sobolev space of order $(-1,\,2)$, this is the measure corresponding to the Euclidean free field with mass 1 in d dimensions.

The central space in our setup is the space of complex-valued functions that are square integrable w.r.t. this measure $L^2(\mu)=L^2(S',\mathcal{C}_\sigma(S'),\mu)$. An element of this space is the Wick exponential,

$$\begin{aligned}:\exp(\langle\omega,f\rangle):&:=\frac{\exp(\langle\omega,f\rangle)}{\mathbb{E}_\mu(\exp(\langle\cdot,f\rangle))},\quad \omega\in S',\ f\in S,\\ &=\sum_{n=0}^{\infty}\frac{1}{n!}\langle:\omega^{\otimes n}:,f^{\otimes n}\rangle.\end{aligned}\tag{3}$$

$\mathbb{E}_\mu$ denotes the expectation w.r.t. μ and the map $S'\ni\omega\mapsto:\omega^{\otimes n}:\in S'^{\hat{\otimes}n}$, $n\in\mathbb{N}$, is called the nth Wick power of $\omega\in S'$ $(\langle:\omega^{\otimes 0}:,f^{\otimes 0}\rangle:=f^{\otimes 0}:=1)$.[15,16] For any $\varphi^{(n)}\in S_{\mathbb{C}}^{\hat{\otimes}n}$, $n\in\mathbb{N}$, $\varphi^{(0)}\in\mathbb{C}$, we define the smooth Wick monomial of order n corresponding to the kernel $\varphi^{(n)}$ as follows:

$$I(\varphi^{(n)})(\omega) := \langle :\omega^{\otimes n}:, \varphi^{(n)} \rangle, \quad \omega \in S', n \in \mathbb{N}_0 .$$

Smooth Wick monomials of different order are orthogonal w.r.t. the standard inner product in $L^2(\mu)$. Furthermore, we can construct Wick monomials $I(f^{(n)})$ with kernels $f^{(n)} \in \mathcal{H}_{\mathbb{C}}^{\hat{\otimes} n}$ in the sense of measurable functions by using an approximation. More precisely, for any sequence $(\varphi_j^{(n)})_{j \in \mathbb{N}} \subset S_{\mathbb{C}}^{\hat{\otimes} n}$ that converges to $f^{(n)}$ in $\mathcal{H}_{\mathbb{C}}^{\hat{\otimes} n}$ we have the convergence of $I(\varphi^{(n)})$ to $I(f^{(n)})$ in any $L^p(\mu), p \geqslant 1$.[15] We use $I(f^{(n)}) = \langle :\omega^{\otimes n}:, f^{(n)} \rangle$ as a formal notation for the monomial introduced above. For Wick monomials associated to the kernels $f^{(n)} \in \mathcal{H}_{\mathbb{C}}^{\hat{\otimes} n}$ and $h^{(m)} \in \mathcal{H}_{\mathbb{C}}^{\hat{\otimes} m}$, $n,m \in \mathbb{N}_0$, we have the following orthogonality property:

$$(I(f^{(n)}), I(h^{(m)}))_{L^2(\mu)} = \int_{S'} \overline{\langle :\omega^{\otimes n}:, f^{(n)} \rangle} \langle :\omega^{\otimes m}:, h^{(m)} \rangle d\mu(\omega) = \delta_{n,m} n! (\overline{f^{(n)}}, h^{(n)}) \tag{4}$$

($\delta_{n,m}$ is the Kronecker delta).

Consider the space $\mathcal{P}(S')$ of smooth Wick polynomial on S':

$$\mathcal{P}(S') = \left\{ \varphi \,\middle|\, \varphi(\omega) = \sum_{n=0}^{N} \langle :\omega^{\otimes n}:, \varphi^{(n)} \rangle, \varphi^{(n)} \in S_{\mathbb{C}}^{\hat{\otimes} n}, \, \omega \in S', \, N \in \mathbb{N}_0 \right\}.$$

This space is dense in $L^2(\mu)$ and, as a consequence, for any $f \in L^2(\mu)$ we have the Itô–Segal–Wiener chaos decomposition,

$$f = \sum_{n=0}^{\infty} I(f^{(n)}), \quad f^{(n)} \in \mathcal{H}_{\mathbb{C}}^{\hat{\otimes} n}.$$

B. Generalized functions

For our considerations the space $L^2(\mu)$ is too small. A convenient way to solve this problem is to introduce a subspace of test functions in $L^2(\mu)$ and to use its larger dual space. In Gaussian analysis there exist various triples of test and generalized functions with $L^2(\mu)$ as a central space, here we choose the Kondratiev triple[29]

$$(S)^1 \subset L^2(\mu) \subset (S)^{-1}.$$

In order to construct these spaces of test and generalized functions, we define for any given $p,q \in \mathbb{Z}$ the following Hilbertian norm for the smooth Wick polynomials $\varphi(\omega) = \sum_{n=0}^{N} \langle :\omega^{\otimes n}:, \varphi^{(n)} \rangle$, $\omega \in S'$:

$$\|\varphi\|_{p,q,1}^2 := \sum_{n=0}^{\infty} (n!)^2 2^{nq} |\varphi^{(n)}|_p^2.$$

Then, for $p,q \in \mathbb{N}_0$, we define the Hilbert space $(\mathcal{H}_p)_q^1$ as the completion of $\mathcal{P}(S')$ w.r.t. $\|\cdot\|_{p,q,1}$. Or, equivalently,

$$(\mathcal{H}_p)_q^1 = \left\{ f \in L^2(\mu) \,\middle|\, f(\omega) = \sum_{n=0}^{\infty} \langle :\omega^{\otimes n}:, f^{(n)} \rangle, \, \|f\|_{p,q,1}^2 < \infty \right\}.$$

Finally, the space of test functions $(S)^1$ is defined as the projective limit of the spaces $(\mathcal{H}_p)_q^1$,

$$(S)^1 = \bigcap_{p,q \geqslant 0} (\mathcal{H}_p)_q^1.$$

Let $(\mathcal{H}_{-p})_{-q}^{-1}$ be the dual w.r.t. $L^2(\mu)$ of $(\mathcal{H}_p)_q^1$ and let $(S)^{-1}$ be the dual w.r.t. $L^2(\mu)$ of $(S)^1$. We know from general duality theory that[15]

$$(S)^{-1}=\bigcup_{p,q\geqslant 0}(\mathcal{H}_{-p})_{-q}^{-1}.$$

The bilinear dual pairing $\langle\langle\cdot,\cdot\rangle\rangle$ between $(S)^1$ and $(S)^{-1}$ is connected to the sesquilinear inner product on $L^2(\mu)$ by

$$\langle\langle f,\varphi\rangle\rangle=(\overline{f},\varphi)_{L^2(\mu)},\quad f\in L^2(\mu),\quad \varphi\in(S)^1. \tag{5}$$

Since the constant function 1 is in $(S)^1$, we may extend the notion of expectation from integrable functions to distributions $\Phi\in(S)^{-1}$:

$$\mathbb{E}_\mu(\Phi):=\langle\langle\Phi,1\rangle\rangle.$$

The chaos decomposition introduces the following natural decomposition of $\Phi\in(S)^{-1}$. Let $\Phi^{(n)}\in S_{\mathbb{C}}'^{\hat{\otimes}n}$ be given. Then there exists a distribution $I(\Phi^{(n)})$ acting on test functions $\varphi\in(S)^1$ as

$$\langle\langle I(\Phi^{(n)}),\varphi\rangle\rangle=n!\langle\Phi^{(n)},\varphi^{(n)}\rangle.$$

We use $I(\Phi^{(n)})=\langle:\omega^{\otimes n}:,\Phi^{(n)}\rangle$, as a formal notation for the distribution introduced above. Any $\Phi\in(S)^{-1}$ then has the unique decomposition

$$\Phi=\sum_{n=0}^{\infty}\langle:\omega^{\otimes n}:,\Phi^{(n)}\rangle, \tag{6}$$

where the sum converges in $(S)^{-1}$, and we have[29]

$$\langle\langle\Phi,\varphi\rangle\rangle=\sum_{n=0}^{\infty}n!\langle\Phi^{(n)},\varphi^{(n)}\rangle,\quad \varphi\in(S)^1.$$

Now it is not hard to see that $(\mathcal{H}_{-p})_{-q}^{-1}$ is a Hilbert space that can be described as follows:

$$(\mathcal{H}_{-p})_{-q}^{-1}=\{\Phi\in(S)^{-1}|\Phi^{(n)}\in S_{\mathbb{C}}'^{\hat{\otimes}n},\|\Phi\|_{-p,-q,-1}<\infty\}.$$

A useful tool in order to characterize $(S)^{-1}$ is the S transform. The S transform of elements from $(S)^{-1}$ is defined as the dual paring with the Wick exponential; see (3). Since the Wick exponential is not an element of $(S)^1$ the S transform of an element Φ from $(S)^{-1}$ is defined only locally, i.e.,

$$S\Phi(g):=\langle\langle\Phi,:\exp(\langle\cdot,g\rangle):\rangle\rangle,\quad g\in\mathcal{U}\subset S_{\mathbb{C}},$$

where $\mathcal{U}$ is an open neighborhood of zero depending on $\Phi\in(S)^{-1}$.

In order to characterize $(S)^{-1}$ we need to define holomorphic functions.

Definition II.1: Let $\mathcal{U}\subset S_{\mathbb{C}}$ be an open neighborhood of zero in $S_{\mathbb{C}}$. The map

$$F:\mathcal{U}\to\mathbb{C}$$

is holomorphic in $\mathcal{U}$ if it satisfies the following two properties.

(i) For each $g_0\in\mathcal{U}$, $g\in S_{\mathbb{C}}$, there exists a neighborhood $V_{g_0,g}$ around zero in $\mathbb{C}$, such that the map

$$z\mapsto F(g_0+zg)$$

is holomorphic in $V_{g_0,g}$.

(ii) For each $g\in\mathcal{U}$ *there exists an open set* $\mathcal{V}\subset\mathcal{U}$ *containing g such that* $F(\mathcal{V})$ *is bounded.*

Furthermore, if we identify two functions F_1 and F_2 coinciding on a neighborhood of zero, we can define $\mathrm{Hol}_0(S_{\mathbb{C}})$ as the space of germs of functions with the above properties.

The proof of the following characterization theorem is given in Ref. 29.

Theorem II.2: *(i) If* $\Phi\in(S)^{-1}$, *then* $S\Phi\in\mathrm{Hol}_0(S_{\mathbb{C}})$.

(ii) For any $F\in\mathrm{Hol}_0(S_{\mathbb{C}})$ *there exists a unique* $\Phi\in(S)^{-1}$ *such that* $S\Phi=F$.

As a consequence of this characterization we have the following corollary; for a proof we again refer to Ref. 29.

Corollary II.3: Let $(\Lambda, \mathcal{A}, \nu)$ *be a measure space and* $\lambda\mapsto\Phi_\lambda$ *a mapping from* Λ *to* $(S)^{-1}$. *Assume there exists an open neighborhood* $\mathcal{U}\subset S_{\mathbb{C}}$ *of zero such that (i)* $S\Phi_\lambda$, $\lambda\in\Lambda$, *is holomorphic on* $\mathcal{U}$; *(ii) the mapping* $\lambda\mapsto S\Phi_\lambda(g)$ *is measurable for every* $g\in\mathcal{U}$; *and (iii) there exists* $C\in L^1(\Lambda,\nu)$ *such that* $|S\Phi_\lambda(g)|\leqslant C(\lambda)$ *for all* $g\in\mathcal{U}$ *and for* ν*—almost all* $\lambda\in\Lambda$.

Then there are $p,q\in\mathbb{N}_0$ *such that* Φ *is Bochner integrable on* $(\mathcal{H}_{-p})^{-1}_{-q}$. *In particular,*

$$\int_\Lambda \Phi_\lambda\, d\nu(\lambda)\in(S)^{-1}.$$

Later on we also use the T transform of generalized functions, which is defined as

$$T\Phi(g):=\exp(-\tfrac{1}{2}|g|^2)\cdot S\Phi(ig),\quad \Phi\in(S)^{-1},\quad g\in\mathcal{U}. \tag{7}$$

An elementary calculation shows that the T transform is also given by

$$T\Phi(g)=\langle\langle\Phi,i\exp(\langle\cdot,g\rangle)\rangle\rangle,\quad g\in\mathcal{U}. \tag{8}$$

The characterization theorem and its corollary are also valid for the T transform.

For elements from $(S)^{-1}$ we can define the Wick product.

Definition II.4: Let $\Phi,\Psi\in(S)^{-1}$. *Then we define the Wick product by*

$$\Phi\diamond\Psi:=S^{-1}(S\Phi\cdot S\Psi).$$

This is well defined because $\mathrm{Hol}_0(S_{\mathbb{C}})$ is an algebra and thus by the characterization theorem in $(S)^{-1}$ there exists a unique element $\Phi\diamond\Psi$ such that $S(\Phi\diamond\Psi)=S\Phi\cdot S\Psi$. Clearly, this multiplication is associative.

By induction, we can define Wick powers,

$$\Phi^{\diamond n}=S^{-1}((S\Phi)^n)$$

in $(S)^{-1}$, and by taking finite linear combinations of them also Wick polynomials of finite order $\Sigma_{n=1}^N a_n\Phi^{\diamond n}$ can be defined in $(S)^{-1}$. Moreover, it is even possible to define Wick analytic functions in $(S)^{-1}$ under very general assumptions.

Theorem II.5: *Let F be analytic in a neighborhood of the point* $z_0=\mathbb{E}_\mu(\Phi)$ *in* $\mathbb{C}$, $\Phi\in(S)^{-1}$. *Then* $F^\diamond(\Phi)$ *defined as* $F^\diamond(\Phi):=S^{-1}(F(S\Phi))$ *exists in* $(S)^{-1}$.

For a proof we refer to Ref. 29.

Remark II.6: Let F be analytic at $z_0=\mathbb{E}_\mu(\Phi)$, $\Phi\in(S)^{-1}$, *i.e., F has the power series representation* $F(z)=\Sigma_n a_n(z-z_0)^n$, $z,a_n\in\mathbb{C}$. *Then the Wick series* $\Sigma_n a_n(\Phi-z_0)^{\diamond n}$ *converges in* S^{-1} *and*

$$F^\diamond(\Phi)=\sum_{n=1}^\infty a_n(\Phi-z_0)^{\diamond n}.$$

III. EUCLIDEAN QFT IN THE FRAMEWORK OF WHITE NOISE ANALYSIS

A. OS axioms in terms of the *T* transform

In 1973, Nelson[30] showed how to construct a relativistic QFT from a Euclidean Markov field. Inspired by this, Osterwalder and Schrader[6,31–34] gave a set of axioms, where Schwinger functions $(\mathcal{S}_n)_{n\in\mathbb{N}_0}$ defined on the Euclidean space–time can be analytically continued to Wightman distributions, i.e., to vacuum expectation values of a relativistic QFT. The OS axioms for a single, scalar field are the following.

OS1 (temperedness)**:** The sequence $(\mathcal{S}_n)_{n\in\mathbb{N}_0}$ is a sequence of tempered distributions, where $\mathcal{S}_n\in S'_{\mathbb{C}}(\mathbb{R}^{dn})$ and $S_0=1$. There exists $p\in\mathbb{N}$ and a sequence $(\sigma_n)_{n\in\mathbb{N}}$ of factorial growth such that for all $n\in\mathbb{N}$ the Schwinger functions fulfill the growth condition of the form

$$|\mathcal{S}_n(f_1\otimes\cdots\otimes f_n)|\leq\sigma_n\prod_{i=1}^{n}|f_i|_p,$$

where $f_1,\ldots,f_n\in S_{\mathbb{C}}(\mathbb{R}^d)$. A sequence $(\sigma_n)_{n\in\mathbb{N}}$ of positive numbers is said to be of factorial growth if the existent constants $\alpha,\beta\in\mathbb{R}^+$ such that

$$\sigma_n\leq\alpha(n!)^\beta,\quad\forall n\in\mathbb{N}.$$

OS2 (Euclidean invariance)**:** Each $\mathcal{S}_n$ is Euclidean invariant, i.e.,

$$\mathcal{S}_n(E_{(a,\Lambda)}f)=\mathcal{S}_n(f),\quad\forall f\in S_{\mathbb{C}}(\mathbb{R}^{dn}),$$

for all $(a,\Lambda)\in E^+(\mathbb{R}^d)$, the proper Euclidean group, where

$$E_{(a,\Lambda)}f(x_1,\ldots,x_n)=f(\Lambda^{-1}(x_1-a),\ldots,\Lambda^{-1}(x_n-a)),$$

for $a\in\mathbb{R}$, $\Lambda\in SO(d)$.

OS3 (reflection positivity)**:** For each sequence $(f_n)_{n\in\mathbb{N}_0}$, where $f_n\in S_{\mathbb{C}}(\mathbb{R}^{dn}_<)$, $f_0\in\mathbb{C}$, and for each $k\in\mathbb{N}_0$,

$$\sum_{n,m=0}^{k}S_{n+m}((\theta f_n)^*\otimes f_m)\geq0,$$

where $(\theta f_n)(t_1,\vec{x}_1;\ldots;t_n,\vec{x}_n)=f_n(-t_1,\vec{x}_1;\ldots;-t_n,\vec{x}_n)$, $t_i\in\mathbb{R}$, $\vec{x}_i\in\mathbb{R}^{d-1}$ (time reflection), $f_n^*(x_1,\ldots,x_n):=\overline{f_n(x_n,\ldots,x_1)}$, and the overbar denotes complex conjugation. The space $S_{\mathbb{C}}(\mathbb{R}^{dn}_<)$ is the space of Schwartz test functions having support in $\mathbb{R}^{dn}_<:=\{(t_1,\vec{x}_1;\ldots;t_n,\vec{x}_n)\in\mathbb{R}^{dn}|0<t_1<\cdots<t_n\}$.

OS4 (symmetry)**:** For $n\geq2$ and all $\pi\in\Sigma_n$, the permutation group,

$$\mathcal{S}_n(f_1\otimes\cdots\otimes f_n)=\mathcal{S}_n(f_{\pi(1)}\otimes\cdots\otimes f_{\pi(n)}),$$

where $f_1,\ldots,f_n\in S_{\mathbb{C}}(\mathbb{R}^d)$.

OS5 (cluster property)**:** For all $a\in\mathbb{R}$, $a\neq0$, and $m,n\geq1$,

$$\lim_{\lambda\to\infty}(\mathcal{S}_{m+n}(f_1\otimes\ldots\otimes f_m\otimes E_{(\lambda a,0)}(f_{m+1}\otimes\ldots\otimes f_{m+n}))-\mathcal{S}_m(f_1\otimes\ldots\otimes f_m)\mathcal{S}_n(f_{m+1}\otimes\ldots\otimes f_{m+n}))=0,$$

where $f_1,\ldots,f_{m+n}\in S_{\mathbb{C}}(\mathbb{R}^d)$.

Remark III.1 The assumptions in axiom (OS1) can be slightly weakened.[31] For technical reasons, by using this formulation it is convenient for us, and since the sequences of generalized functions we consider fulfill (OS1), we do not lose anything by this slightly stronger formulation.

In the case of Euclidean Markov fields, and also in the more general case of Euclidean reflection positivity fields,[35] Schwinger functions fulfilling (OS1)–(OS5) are obtained as the moments of the Euclidean field. In this section we construct Schwinger functions $(\mathcal{S}_n^{\Phi})_{n\in\mathbb{N}_0}$ that are moments of generalized functions $\Phi\in(S)^{-1}$ with $\mathbb{E}_{\mu}(\Phi)=1$. The moments $(\mathcal{S}_n^{\Phi})_{n\in\mathbb{N}_0}$, in general, do not satisfy all axioms (OS1)–(OS5). Nevertheless, we call them Schwinger functions because our aim is to work out a class of generalized functions $\Phi\in(S)^{-1}$ such that their moments fulfill all or a part of the OS axioms.

Definition III.2: Let $f_1,\dots,f_n\in S(\mathbb{R}^d)$, $n\in\mathbb{N}$. *The* nth *Schwinger function corresponding to* $\Phi\in(S)^{-1}$, $\mathbb{E}_{\mu}(\Phi)=1$, *is given as*

$$\mathcal{S}_n^{\Phi}(f_1\otimes\cdots\otimes f_n)=\langle\langle\Phi,\langle\omega,f_1\rangle\cdot\ \cdots\ \cdot\langle\omega,f_n\rangle\rangle\rangle,$$

and $\mathcal{S}_0^{\Phi}=\mathbb{E}_{\mu}(\Phi)=1$.

Since $\mathcal{P}(S')\subset(S)^1$, the dual paring in the above definition is well defined.

The Schwinger functions corresponding to $\Phi\in(S)^{-1}$ can be calculated via their T transform; see (8).

Proposition III.3 (Wick theorem)*:* Let $f_1,\dots,f_n\in S(\mathbb{R}^d)$, $n\in\mathbb{N}$. *Then the* nth *Schwinger functions corresponding to* $\Phi\in(S)^{-1}$ *is given by*

$$\mathcal{S}_n^{\Phi}(f_1\otimes\cdots\otimes f_n)=(-i)^n\frac{\partial^n}{\partial t_1\cdots\partial t_n}T\Phi(t_1f_1+\cdots+t_nf_n)\bigg|_{t_1=\cdots t_n=0}.$$

Proof: By construction, every distribution $\Phi\in(S)^{-1}$ is of finite order, i.e., for each $\Phi\in(S)^{-1}$ there exist $p,q\in\mathbb{N}_0$, such that $\Phi\in(\mathcal{H}_{-p})_{-q}^{-1}$. Furthermore, a straightforward calculation shows that for each $f\in S(\mathbb{R}^d)$ there exists $t_0>0$ such that $\exp(it\langle\cdot,f\rangle)\in(\mathcal{H}_p)_q^1$ for all $0\leqslant t<t_0$, and

$$-i\frac{d}{dt}\exp(it\langle\cdot,f\rangle)\bigg|_{t=0}=\langle\cdot,f\rangle,$$

w.r.t. the Hilbert space norm in $(\mathcal{H}_p)_q^1$. From this we can conclude that

$$\mathcal{S}_1^{\Phi}(f)=-i\frac{d}{dt}T\Phi(tf)\bigg|_{t=0}=-i\frac{d}{dt}\langle\langle\Phi,\exp(it\langle\cdot,f\rangle)\rangle\rangle\bigg|_{t=0}.$$

Since $(S)^1$ is an algebra under multiplication and this multiplication is continuous, we can define the pointwise product $\Phi\cdot\varphi\in(S)^{-1}$ of a distribution $\Phi\in(S)^{-1}$ with a test function $\varphi\in(S)^1$ via the dual paring. Utilizing this product, the proposition follows by an induction argument. ∎

Proposition III.4: For each generalized function $\Phi\in(S)^{-1}$ *with* $\mathbb{E}_{\mu}(\Phi)=1$, *the Schwinger functions* $(\mathcal{S}_n^{\Phi})_{n\in\mathbb{N}^0}$ *fulfill the axioms (OS1) and (OS4). Furthermore, (OS2) is fulfilled if* $T\Phi$ *is Euclidean invariant.*

Proof: The Schwinger functions $(\mathcal{S}_n^{\Phi})_{n\in\mathbb{N}_0}$ are symmetric by definition. Temperedness and factorial growth follows immediately from the fact that $\Phi\in(S)^{-1}$. Thus, (OS1) and (OS4) are fulfilled.

Assume that $T\Phi$ is Euclidean invariant. Then we apply Proposition III.3 to calculate the nth Schwinger function corresponding to Φ; of course, it is also Euclidean invariant. ∎

B. Euclidean-invariant distributions

In this section we construct a class of Euclidean-invariant generalized functions. We call generalized functions from $(S)^{-1}$ Euclidean invariant if their T transform is Euclidean invariant. Our construction is motivated by the Euclidean strategy for constructing interacting field theories; see, e.g., Refs. 1 and 2 and the references therein. In the framework of a white noise analysis, we can define the Gaussian random process indexed by $\mathcal{H}=L^2(\mathbb{R}^d)$ as

$$\phi(h):=\langle\cdot,h\rangle,\quad h\in L^2(\mathbb{R}^d).$$

As discussed in Sec. II A, $\phi(h)$ is an element of $L^2(\mu)$ for all $h\in L^2(\mathbb{R}^d)$. We are interested in the Gaussian random process ϕ at time that $t\in\mathbb{R}$ and at the point $\vec{x}\in\mathbb{R}^{d-1}$, where we write $x\in\mathbb{R}^d$ as $x=(t,\vec{x})$. $\phi(t,\vec{x})$ does not exist as a square-integrable function, but we can define

$$\phi(t,\vec{x}):=\langle\cdot,\delta_{t,\vec{x}}\rangle,\qquad \in(S)^{-1},$$

see (6) where $\delta_{t,\vec{x}}\in S'(\mathbb{R}^d)$ is the Dirac delta function at point $(t,\vec{x})\in\mathbb{R}^d$.

Assume that $H(z)=\Sigma_{k=0}^{\infty}(1/k!)H_k z^k$, $z\in U\subset\mathbb{C}$, is a holomorphic function in U, where U is an open neighborhood of $\mathbb{E}_\mu(\phi(t,\vec{x}))=0$. Then by using Theorem II.5, we can define

$$H^\diamond(\phi(t,\vec{x}))=\sum_{k=0}^{\infty}\frac{1}{k!}H_k\phi(t,\vec{x})^{\diamond k}\in(S)^{-1}=\sum_{k=0}^{\infty}\frac{1}{k!}H_k\langle:\omega^{\otimes k}:,\delta_{t,\vec{x}}^{\otimes k}\rangle;$$

also see Remark II.6.

Next, we want to define the integral

$$\int_{\mathbb{R}^d}H^\diamond(\phi(x))d^dx. \tag{9}$$

This can only be possible if we assume that $H_0=0$.

Theorem III.5: *Let H be holomorphic at zero such that* $H(0)=0$, *then (9) exists as a Bochner integral in a suitable subspace of* $(S)^{-1}$.

Proof: Our aim is to apply Corollary II.3. Let $r>0$ be in the radius of convergence of the Taylor expansion of H at the origin. We define

$$\mathcal{U}=\{g\in S_{\mathbb{C}}|\sup_{x\in\mathbb{R}^d}\{(1+|x|^2)^d|g(x)|\}<r\}.$$

It is easy to check that $\mathcal{U}$ is an open neighborhood of zero. For $g\in\mathcal{U}$ we have

$$\begin{aligned}S(H^\diamond(\phi(x)))(g)&=\sum_{k=1}^{\infty}\frac{1}{k!}H_k\langle g^{\otimes k},\delta_x^{\otimes k}\rangle=\sum_{k=1}^{\infty}\frac{1}{k!}H_k g(x)^k\leqslant\sum_{k=1}^{\infty}|r^{-1}g(x)|^k|H_k|\frac{r^k}{k!}\\&\leqslant\frac{1}{(1+|x|^2)^d}(1+|x|^2)^d|g(x)|r^{-1}\sum_{k=1}^{\infty}|H_k|\frac{r^k}{k!}\leqslant\frac{1}{(1+|x|^2)^d}\sum_{k=1}^{\infty}|H_k|\frac{r^k}{k!}.\end{aligned}\tag{10}$$

Obviously, $S(H^\diamond(\phi(\cdot)))(g)$ is measurable for all $g\in\mathcal{U}$. With the estimate (10) the holomorphy of $S(H^\diamond(\phi(x)))$ is clear. Since $(1+|x|^2)^{-d}\in L^1(\mathbb{R}^d)$, all assumption required in Corollary II.3 are fulfilled and the theorem is proved. ■

Corollary III.6: Let the function H be as in Theorem III.5. Then the generalized function,

$$\Phi_H:=\exp^\diamond\left(-\int_{\mathbb{R}^d}H^\diamond(\phi(x))d^dx\right),$$

is an elements of $(S)^{-1}$. *Its T transform is given by*

$$T\Phi_H(g)=\exp\left(-\int_{\mathbb{R}^d}H(ig(x))+\frac{1}{2}g(x)^2\,d^dx\right),$$

for all g in a neighborhood $\mathcal{U}\subset S_{\mathbb{C}}$ *of zero. In particular*, $\mathbb{E}_\mu(\Phi_H)=1$.

Proof: This corollary is an immediate consequence of Theorem II.5. For the calculation of the T transform, we used (7). Observe that $\mathbb{E}_\mu(\Phi_H)=T\Phi_H(0)$. ■

Remark III.7: (i) Since the Lebesgue measure on $\mathbb{R}^d$ *is Euclidean-invariant* $T\Phi_H$ *is Euclidean invariant.*

(ii) Consider the case in which the function $F(s) := -(H(is)+\frac{1}{2}s^2)$, $s\in\mathbb{R}$, *is a Lévy characteristic, i.e.,*

$$F(s)=ias-\frac{\sigma^2 s^2}{2}+\int_{\mathbb{R}\setminus\{0\}}\left(\exp(irs)-1-\frac{irs}{1+r^2}\right)d\nu(r),\quad s\in\mathbb{R},$$

where $a\in\mathbb{R}$, $\sigma\geqslant 0$, *and the measure* ν *satisfies the following condition:*

$$\int_{\mathbb{R}\setminus\{0\}}\min\{1,r^2\}d\nu(r)<\infty.$$

Then by the Lévy–Khinchine theorem[36] *we know that there exists a probability measure* P_H *on* $S'(\mathbb{R}^d)$ *such that*

$$T\Phi_H(f)=\int_{S'(\mathbb{R}^d)}\exp(i\langle\omega,f\rangle)dP_H(\omega),\quad f\in S(\mathbb{R}^d).$$

This implies that the Schwinger functions $(\mathcal{S}_n^{\Phi_H})_{n\in\mathbb{N}_0}$ *are the moments of the measure* P_H*. These measures are called generalized white noise measures.*

Next, we enlarge the class of Euclidean-invariant distributions. We do this by convolution with kernels associated to Euclidean-invariant operators. This idea is inspired by the method used in Ref. 5. There the authors started with Euclidean-invariant measures from the Lévy–Khinchine class and then they constructed image measures by convoluting the corresponding generalized white noise with kernels associated to Euclidean-invariant operators. These image measures are called convoluted generalized white noise measures.

Let $\mathcal{G}:S(\mathbb{R}^d)\to S(\mathbb{R}^d)$ be a linear continuous mapping. Then by the well-known kernels theorem there exists a distribution $K\in S'(\mathbb{R}^{2d})$, hereafter called the kernel of $\mathcal{G}$, such that

$$\mathcal{G}f(x)=\int_{\mathbb{R}^d}K(x,y)f(y)dy,\quad f\in S(\mathbb{R}^d),\quad x\in\mathbb{R}^d,$$

in the distributional sense. It is clear that the adjoint operator $\mathcal{G}^*:S'(\mathbb{R}^d)\to S'(\mathbb{R}^d)$ is a measurable transformation from $(S'(\mathbb{R}^d),\mathcal{C}_\sigma(S'(\mathbb{R}^d)))$ into itself. Furthermore, we assume that $\mathcal{G}$ is Euclidean invariant, i.e., $\mathcal{G}E_{(a,\wedge)}=E_{(a,\wedge)}\mathcal{G}$ for all $E_{(a,\wedge)}\in E^+(\mathbb{R}^d)$. This implies that $\mathcal{G}$ is translation invariant, thus, its kernel K has the form $K(x,y)=G(x-y)$.[7] The action of $\mathcal{G}$ on test functions from $S(\mathbb{R}^d)$ [and by duality on $S'(\mathbb{R}^d)$] is, by convolution,

$$\mathcal{G}f(x)=\int_{\mathbb{R}^d}G(x-y)f(y)dy,\quad f\in S(\mathbb{R}^d),\quad x\in\mathbb{R}^d.$$

We can also write $\mathcal{G}f$ as $G*f$. From now on we assume $\mathcal{G}$ to be essentially self-adjoint in $L^2(\mathbb{R}^d)$, with $S(\mathbb{R}^d)$ as a core. Then the convolution of the Gaussian random process ϕ with G is defined as

$$(G*\phi)(h)(\omega):=\langle G*\omega,f\rangle=\langle\omega,G*f\rangle,\quad \omega\in S'(\mathbb{R}^d),\quad f\in S(\mathbb{R}^d).$$

This definition can also be generalized from test functions f to tempered distributions. Then the process is in $(S)^{-1}$.

Example III.8: Let Δ *be the Laplace operator on* $\mathbb{R}^d$*. Let* $K(x,y)=G_\alpha(x-y)$ *be the Green's function of the pseudodifferential operator* $\mathcal{G}_\alpha=(-\Delta+m_0^2)^{-\alpha}$ *for some arbitrary* $m_0>0$ *and* $0<\alpha$*. It is given by*

$$G_\alpha(x)=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{\exp(ikx)}{(|k|^2+m_0^2)^\alpha}dk,\quad x\in\mathbb{R}^d,$$

where the integral has to be understood in the sense of a Fourier transform of a tempered distribution. One easily proves that $\mathcal{G}_\alpha:S(\mathbb{R}^d)\to S(\mathbb{R}^d)$ is continuous, essentially self-adjoint in $L^2(\mathbb{R}^d)$ with $S(\mathbb{R}^d)$ as a core, and Euclidean invariant.

Let $\Phi\in(S)^{-1}$ be a generalized function. We define its convolution with a Euclidean invariant kernel G by

$$T\Phi^G(g):=T\Phi(G*g),\quad g\in\mathcal{U}\subset S_{\mathbb{C}}(\mathbb{R}^d),$$

where $\mathcal{U}$ is an open neighborhood of zero. Since the operator $\mathcal{G}$ is linear and continuous, the characterization, Theorem II.2, implies that Φ^G is a well-defined and unique element in $(S)^{-1}$.

Theorem III.9: *Let H be as in Theorem III.5 and let the operator $\mathcal{G}$ be continuous in $S(\mathbb{R}^d)$, essentially self-adjoint in $L^2(\mathbb{R}^d)$ with $S(\mathbb{R}^d)$ as a core, and Euclidean invariant. Then the generalized function $\Phi_H^G\in(S)^{-1}$ is Euclidean invariant and can be written as*

$$\Phi_H^G=\exp^\diamond\left(-\int_{\mathbb{R}^d}H^\diamond(G*\phi(x))d^dx+\frac{1}{2}\langle :\cdot^{\otimes 2}:,(\mathcal{G}^{\otimes 2}-1)\mathrm{Tr}\rangle\right);\tag{11}$$

here $Tr\in S'(\mathbb{R}^d)^{\hat{\otimes}2}$ denotes the trace kernel defined by $\langle Tr,f\otimes g\rangle=(f,g)$, $f,g\in S(\mathbb{R}^d)$. The Schwinger functions $(\mathcal{S}_n^{H,G})_{n\in\mathbb{N}_0}$ (with the abbreviation $\mathcal{S}^{H,G}=\mathcal{S}^{\Phi_H^G}$) fulfill the axioms $(OS1)$, $(OS2)$, and $(OS4)$.

Remark III.10: In the case where F is a Lévy characteristic, see Remark III.7(ii), the measure P_H^G corresponding to the distribution Φ_H^G is an image measures of the measure P_H, more concretely, $P_H^G(A)=P_H(\mathcal{G}^{-1}A)$, $A\in\mathcal{C}_\sigma(S'(\mathbb{R}^d))$

Proof of Theorem III.9: Formula (11) is clear by taking its T transform. Euclidean invariance follows from the Euclidean invariance of $\mathcal{G}$ and Φ_H. Obviously, $\mathbb{E}_\mu(\Phi_H)=1$, thus the theorem follows by an application of Proposition III.3. ■

Example III.11: The choice $H\equiv 0$ and $\mathcal{G}_{1/2}=(-\Delta+m_0^2)^{-1/2}$ gives the free Euclidean field with mass $m_0>0$; see Example III.8. Theorem III.9 implies that the corresponding measure $P_0^{G_{1/2}}$ has the generalized density

$$\Phi_0^{G_{1/2}}=\exp^\diamond\left(\tfrac{1}{2}\langle :\cdot^{\otimes 2}:,(\mathcal{G}_{1/2}^{\otimes 2}-1)Tr\rangle\right),$$

w.r.t. the Gaussian white noise measure.

Remark III.12: Consider the Hilbert space N_{m_0}, which is defined as the closure of $S(\mathbb{R}^d)$ w.r.t. the Hilbert space norm $|\cdot|_{m_0}$ given by the scalar product

$$(f,g)_{m_0}:=\int_{\mathbb{R}^d}f(x)(-\Delta+m_0^2)^{-1}g(x)dx,\quad f,g\in S(\mathbb{R}^d),\quad m_0>0.$$

The random process indexed by $\mathcal{H}=N_{m_0}$:

$$\phi(h):=\langle\cdot,h\rangle_{m_0},\quad h\in N_{m_0},$$

is the free Euclidean field with mass m_0. Since N_{m_0} fulfills the assumptions on the Hilbert space $\mathcal{H}$ required in Sec. II A, it is also possible to take the measure corresponding to the free Euclidean field as a reference measure for the Euclidean-invariant distributions constructed above (this is the usual choice in constructive Euclidean QFT). Here we have chosen the white noise measure because it has the identity operator as a covariance operator. This is a reasonable choice for our approach, which involves a convolution with the operator kernel G.

In Ref. 5 the authors studied Schwinger functions $(\mathcal{S}_n^{H,G})_{n\in\mathbb{N}_0}$ that are moments of the measures P_H^G corresponding to the generalized functions Φ_H^G, where $F(s)=-(H(is)+\frac{1}{2}s^2)$, $s\in\mathbb{R}$, is a Lévy characteristic; see Remark III.7 (ii) and Remark III.10 For the truncated Schwinger functions the authors worked out explicit formulas. Before we give them let us recall the definition of truncated Schwinger functions.

A partition of the ordered set $\{1,...,n\}$ is a family of ordered subsets $I_1=\{i_1,\dots,i_{k(1)}\},\dots,I_l=\{i'_1,\dots,i'_{k(l)}\}$, so that $i_1<\cdots<i_{k(1)},\dots,i'_1<\cdots<i'_{k(l)}$ and so that $\cup_{1\leqslant j\leqslant l}I_j=\{1,...,n\}$ and $I_j\cap I_q=\emptyset$, $j\neq q$. The set of all partitions of $\{1,...,n\}$ we denote as $P^{(n)}$.

Definition III. 13: The truncated Schwinger functions $(\mathcal{S}_{n,T})_{n\in\mathbb{N}_0}$ corresponding to a given sequence of Schwinger functions $(\mathcal{S}_n)_{n\in\mathbb{N}_0}$ are defined recursively by the relation

$$\mathcal{S}_n(f_1\otimes\cdots\otimes f_n)=\sum_{P^{(n)}}\mathcal{S}_{k(1),T}(f_{i_1}\otimes\cdots\otimes f_{i_{k(1)}})\cdot\ \cdots\ \cdot\mathcal{S}_{k(l),T}(f_{i'_1}\otimes\cdots\otimes f_{i'_{k(l)}}),$$

where $f_1,\dots,f_n\in S(\mathbb{R}^d)$, $n\geqslant1$.

Proposition III.14: Let $H(z)=\Sigma_{n=0}^{\infty}(1/n!)H_nz^n$, $z\in U\subset\mathbb{C}$, and $\mathcal{G}$ be as in Theorem III.9 and $f_1,\dots,f_n\in S(\mathbb{R}^d)$, $n\geqslant1$. Then the truncated Schwinger functions $(\mathcal{S}_{n,T}^{H,G})_{n\in\mathbb{N}}$ are given by

$$\mathcal{S}_{n,T}^{H,G}(f_1\otimes\cdots\otimes f_n)=-H_n\int_{\mathbb{R}d}G*f_1(x)\cdot\ \cdots\ \cdot G*f_n(x)d^dx,\quad n\neq2,$$

$$\mathcal{S}_{2,T}^{H,G}(f_1\otimes f_2)=(-H_2+1)\int_{\mathbb{R}d}G*f_1(x)\cdot G*f_2(x)d^dx. \tag{12}$$

Proof: In the case where $F(s)=-(H(is)+\frac{1}{2}s^2)$, $s\in\mathbb{R}$ is a Lévy characteristic, this follows from Proposition 3.9 in Ref. 5 and the uniqueness of the truncated Schwinger functions. The coefficients in front of the integrals corresponding to the nth truncated Schwinger function in (12) are just the nth derivatives of the Lévy characteristic divided by i^n. Hence, for a general H as in Theorem III.9 these coefficients are given by the nth derivative of $-(H(iz)+\frac{1}{2}z^2)$, $z\in U$. ■

In Corollary 4.7. of Ref. 5, the authors have proved the cluster property of the Schwinger functions $(\mathcal{S}_n^{H,G})_{n\in\mathbb{N}_0}$ arising from measures. The proof given there easily generalizes to our case.

Theorem III.15: Let $\Phi_H^G\in(S)^{-1}$ *be as in Theorem III.9. Then the corresponding Schwinger functions $(\mathcal{S}_n^{H,G})_{n\in\mathbb{N}_0}$ fulfill the cluster property (OS5), i.e., for all $a\in\mathbb{R}$, $a\neq0$, and $m,n\geqslant1$,*

$$\lim_{\lambda\to\infty}(\mathcal{S}_{m+n}^{H,G}(f_1\otimes\cdots\otimes f_m\otimes E_{(\lambda a,0)}(f_{m+1}\otimes\cdots\otimes f_{m+n}))-\mathcal{S}_m^{H,G}(f_1\otimes\cdots\otimes f_m)$$
$$\times\mathcal{S}_n^{H,G}(f_{m+1}\otimes\cdots\otimes f_{m+n}))=0,$$

where $f_1,\dots,f_{m+n}\in S(\mathbb{R}^d)$.

Proof: See the proof of Corollary 4.7 in Ref. 5. There the authors proved the cluster property in the case where $F(s)=-(H(is)+\frac{1}{2}s^2)$, $s\in\mathbb{R}$ is a Lévy characteristic. The idea is to express the cluster property of the Schwinger functions as an equivalent property of the truncated Schwinger functions. Since their proof works independently of the choice of the coefficients in front of the integrals corresponding to the nth truncated Schwinger function, see (12), it easy generalizes to our case. ■

Remark III.16: The class of Schwinger functions $(\mathcal{S}_n^{H,G})_{n\in\mathbb{N}_0}$ corresponding to the distributions $\Phi_H^G\in(S)^{-1}$ as in Theorem III.9 differs from the class of Schwinger functions corresponding to the convoluted generalized white noise measures in Ref. 5. Let us compare the properties of the Lévy characteristics F that have been used in Ref. 5 with the properties of the functions H we employ, where $F(s)=-(H(is)+\frac{1}{2}s^2)$, $s\in\mathcal{O}\subset\mathbb{R}$.

We need that the function H is holomorphic at zero and $H(0)=0$. *This is our restriction in choosing the coefficients in front of the integrals corresponding to the* nth *truncated Schwinger function; see (12).*

In Ref. 5 the authors needed the condition that the measure ν *in the representation of the Lévy characteristic, see Remark III.7(ii), has finite moments to all orders. This implies that* $F \in C^{\infty}(\mathbb{R})$, *but F does not have to have a holomorphic extension. Furthermore, also* $F(0)=0$ *and F cannot be a polynomial of order larger than 2. That is, if only finite many* $H_n, n \in \mathbb{N}$, *are different from zero then all* H_n, $n \geqslant 3$, *have to be zero. Furthermore, the constant* $-H_n$ *is the* nth *moment of the measure* ν *for* $n \geqslant 3$.

IV. ON REFLECTION POSITIVITY, ANALYTIC CONTINUATION, AND QFT WITH INDEFINITE METRIC

A. Reflection positivity

In Sec. III B, we proved all OS axioms for Schwinger functions $(\mathcal{S}_n^{H,G})_{n \in \mathbb{N}_0}$ corresponding to the distributions $\Phi_H^G \in (S)^{-1}$, H,G as in Theorem III.9, except for reflection positivity.

In Ref. 5 the authors present a partial negative result on reflection positivity of Schwinger functions $(\mathcal{S}_n^{H,G})_{n \in \mathbb{N}_0}$, which are moments of convoluted generalized white noise P_H^G. Consider a Lévy characteristic represented as in Remark III 7(ii). The part arising from the measure ν is called the Poisson part and the other part is called the Gaussian part (the reason for these names and decomposition lies in the properties of the corresponding measures). For the Schwinger functions $(\mathcal{S}_n^{H,G})_{n \in \mathbb{N}_0}$ that are moments of convoluted generalized white noises P_H^G with the nonzero Poisson in part in Ref. 5, some examples have been constructed that do not have the reflection positivity property. Roughly speaking, the Schwinger functions $(\mathcal{S}_n^{H,G})_{n \in \mathbb{N}_0}$ do not have the reflection positivity property, if the terms in $\mathcal{S}_n^{H,G}$ emerging from the ''interaction'' (Poisson part) are large in comparison with the ''free'' terms (the Gaussian part). More details on this considerations can be found in Ref. 5, Remark 5.12.

We discuss the question of whether reflection positivity holds or does not hold in the next section in terms of the Wightman functions; see Theorem IV.2 and Remark IV.3.

B. Analytic continuation to Wightman functions

If a sequence of Schwinger functions fulfills all OS axioms one can perform the analytic continuation to Wightman functions via the reconstruction theorem.[6] These Wightman functions fulfill the Wightman axioms

W1 (temperedness)**:** The sequence $(W_n)_{n \in \mathbb{N}_0}$ is a sequence of tempered distributions, where $W_n \in S'_{\mathbb{C}}(\mathbb{R}^{dn})$ and $W_0=1$. These functions fulfill the Hermiticity condition

$$\overline{W_n(f)} = W_n(f^{\star}).$$

W2 (Poincaré invariance)**:** Each W_n is Poincaré invariant, i.e.,

$$W_n(P_{(a,\Lambda)}f) = W_n(f), \quad \forall f \in S_{\mathbb{C}}(\mathbb{R}^{dn}),$$

for all $(a,\Lambda) \in P_+^{\uparrow}(\mathbb{R}^d)$, where $P_+^{\uparrow}(\mathbb{R}^d)$ is the proper, orthochronous Poincaré group. The definition of $P_{(a,\Lambda)}f$ is analog to the definition of Euclidean transformations in $S_{\mathbb{C}}(\mathbb{R}^{dn})$; see (OS2).

W3 (positivity)**:** For each sequence $(f_n)_{n \in \mathbb{N}_0}$, where $f_n \in S_{\mathbb{C}}(\mathbb{R}^{dn})$, $f_0 \in \mathbb{C}$, and each $k \in \mathbb{N}_0$,

$$\sum_{n,m=0}^{k} W_{n+m}(f_n^{\star} \otimes f_m) \geqslant 0.$$

W4 (locality)**:** If for $n \geqslant 2$ for some $1 \leqslant j \leqslant n-1$: $\langle x_{j+1}-x_j, x_{j+1}-x_j \rangle_M < 0$, then

$$W_n(x_1,\ldots,x_j,x_{j+1},\ldots,x_n) = W_n(x_1,\ldots,x_{j+1},x_j,\ldots,x_n),$$

where $\langle x,x\rangle_M = t^2 - |\vec{x}|^2$, $x=(t,\vec{x})\in\mathbb{R}^d$, is the Minkowski inner product.

We remark that by (W2) every W_n is actually a distribution in the difference variables, i.e., there is a tempered distribution $w_n \in S'_{\mathbb{C}}(\mathbb{R}^{d(n-1)})$, defined as

$$w_n(x_1-x_2,\dots,x_{n-1}-x_n) := W_n(x_1,\dots,x_n).$$

The Fourier transform on $S_{\mathbb{C}}(\mathbb{R}^{dn})$ and $S'_{\mathbb{C}}(\mathbb{R}^{dn})$, respectively, we denote by $\mathcal{F}$ or $\hat{}$ and is taken w.r.t. the Euclidean inner product. The forward mass cone of mass m_0 is defined as

$$V^+_{m_0} := \{p\in\mathbb{R}^d | p^2 > m_0^2, \quad p^0 = \langle p, e_0\rangle_M > 0\}, \quad m_0 \geqslant 0,$$

where $e_0=(1,0,0,0)$. By $V^{\star,+}_{m_0}$ we denote its closure and V^+_0 is called a forward light cone. The backward mass cone is defined by $V^-_{m_0} := \theta V^+_{m_0}$ where θ again denotes the time reflection.

W5 (spectral condition)**:** For any $n\geqslant 2$ the Fourier transform $\hat{w}_n$ has support in the $(n+1)$-fold product of the forward light cone $(V^{\star +}_0)^{n-1}$.

W6 (cluster property)**:** For any $n,m\in\mathbb{N}$ and any space like $a\in\mathbb{R}^d$, i.e., $\langle a,a\rangle_M<0$,

$$\lim_{\lambda\to\infty} \big(W_{m+n}(f_1\otimes\cdots\otimes f_m\otimes T_{\lambda a}(f_{m+1}\otimes\cdots\otimes f_{m+n})) - W_m(f_1\otimes\cdots\otimes f_m)W_n(f_{m+1}\otimes\cdots\otimes f_{m+n})\big) = 0,$$

for $f_1,\dots,f_{m+n}\in S_{\mathbb{C}}(\mathbb{R}^d)$, where $T_{\lambda a}$ denotes the translation by λa.

Without reflection positivity we cannot perform the analytic continuation to Wightman functions via the standard reconstruction theorem. Nevertheless, an analytic continuation can be done. Using the results from the theory of Laplace transforms in Ref. 5, the authors analytically continued the truncated Schwinger functions $(\mathcal{S}^{H,G_\alpha}_{n,T})_{n\in\mathbb{N}_0}$, which are moments of convoluted generalized white noise, to truncated Wightman functions $(W^{H,G_\alpha}_{n,T})_{n\in\mathbb{N}_0}$ for $\alpha\in(0,\frac{1}{2}]$; see Example III 8. The truncated Wightman functions are related to the Wightman functions in the same way as truncated Schwinger functions are related to Schwinger functions; see Definition III.13. In particular, the authors found an explicit formula for $\hat{W}^{H,G_\alpha}_{n,T}$, the Fourier transform of the nth truncated Wightman function. In order to give these formulas we introduce the notations

$$\mu^+_\alpha(p) = (2\pi)^{-d/2}\sin(\pi\alpha)\mathbf{1}_{\{p^2>m_0^2, p^0>0\}}(p)\frac{1}{(p^2-m_0^2)^\alpha}, \quad p\in\mathbb{R}^d, \quad m_0>0,$$

$$\mu^-_\alpha(p) = (2\pi)^{-d/2}\sin(\pi\alpha)\mathbf{1}_{\{p^2>m_0^2, p^0<0\}}(p)\frac{1}{(p^2-m_0^2)^\alpha}, \quad \alpha\in(0,\tfrac{1}{2}],$$

$$\mu_\alpha(p) = (2\pi)^{-d/2}(\cos(\pi\alpha)\mathbf{1}_{\{p^2>m_0^2\}}(p) + \mathbf{1}_{\{p^2<m_0^2\}}(p))\frac{1}{|p^2-m_0^2|^\alpha},$$

where $\mathbf{1}_A$ is the indicator function of the subset $A\subset R^d$. In Proposition 7.12. and Corollary 7.13 in Ref. 5, it is proved that in the case when $F(s) = -(H(is)+\frac{1}{2}s^2)$, $s\in\mathbb{R}$ is a Lévy characteristic and $\alpha\in(0,\frac{1}{2}]$, the Fourier transform of the nth truncated Wightman function for $n\geqslant 3$ is given by

$$\hat{W}^{H,G_\alpha}_{n,T} = -H_n(2\pi)^d 2^{(n-1)}\left(\sum_{j=1}^n \prod_{l=1}^{j-1}\mu^+_\alpha(p_l)\mu_\alpha(p_j)\prod_{l=j+1}^n \mu^-_\alpha(p_l)\right)\delta\left(\sum_{l=1}^n p_l\right). \tag{13}$$

In the case $n=2$ one has to distinguish between the two cases $\alpha\in(0,\frac{1}{2})$ and $\alpha=\frac{1}{2}$. For $\alpha\in(0,\frac{1}{2})$ the two-point function is given by

$$\hat{W}_{2,T}^{H,G_\alpha}=(-H_2+1)(2\pi)^d 2(\mu_\alpha(p_1)\mu_\alpha^-(p_2)+\mu_\alpha^+(p_1)\mu_\alpha(p_2))\delta(p_1+p_2) \tag{14}$$

and

$$\hat{W}_{2,T}^{H,G_{1/2}}=(-H_2+1)(2\pi)^{d+1}\mathbf{1}_{\{p_1^0>0\}}(p_1)\,\delta(p_1^2-m_0^2)\,\delta(p_1+p_2) \tag{15}$$

is the Fourier transform of the well-known two-point function of the relativistic free field.

The truncated Wightman function $W_{n,T}^{H,G_\alpha}$ is an analytic continuation of the truncated Schwinger function $\mathcal{S}_{n,T}^{H,G_\alpha}$ in the sense that

$$\mathcal{S}_{n,T}^{H,G_\alpha}(\mathfrak{I}(z_1^0,\mathfrak{R}(\vec{z_1}),\dots,\mathfrak{I}(z_n^0),\mathfrak{R}(\vec{z_n}))=\mathcal{F}\hat{W}_{n,T}^{H,G_\alpha}(z), \quad z\in\mathbb{C}_<^{dn}, \tag{16}$$

where

$$\mathbb{C}_<^{dn}:=\{(z_1^0,\vec{z_1};\dots;z_n^0,\vec{z_n})\in\mathbb{C}^{dn}|\mathfrak{I}(z_{j+1}^0-z_j^0)>0, j=1,...,n-1, \mathfrak{I}(\vec{z_j})=0, \mathfrak{R}(z_j^0)=0, j=1,...,n\}$$

[$\mathfrak{R}(z)$ is the real part and $\mathfrak{I}(z)$ is the imaginary part of a (vector-valued) complex variable z]. The function $\hat{W}_{n,T}^{H,G_\alpha}$ is determined uniquely by this requirement. Furthermore, $W_{n,T}^{H,G_\alpha}(\mathfrak{R}(z))$ is the boundary value of $\mathcal{F}\hat{W}_{n,T}^{H,G_\alpha}(z)$ for $\mathfrak{I}(z_{j+1}-z_j)\to 0$ inside T^n, i.e., the relation

$$\lim_{\Gamma\ni\mathfrak{I}(z_{j+1}-z_j)\to 0}\mathcal{F}\hat{W}_{n,T}^{H,G_\alpha}(z)=W_{n,T}^{H,G_\alpha}(\mathfrak{R}(z)), \tag{17}$$

holds in the sense of tempered distributions in the argument $\mathfrak{R}(z)\in\mathbb{R}^d$. Here T^n is the tubular domain in $\mathbb{C}^{dn}$ with base V_0^+, i.e.,

$$T^n:=\{(z_1,\dots,z_n)\in\mathbb{C}^{dn}|z_{j+1}-z_j\in\mathbb{R}^d+iV_0^+, j=1,...,n-1\},$$

and $\Gamma\subset V_0^+$ is a subcone of V_0^+ such that $\Gamma\cup\{0\}$ is closed in $\mathbb{R}^d$.

Theorem IV.1: *Let H be as in Theorem III.5 and $\mathcal{G}_\alpha$ as in Example III.8, $\alpha\in(0,\frac{1}{2}]$.*

(i) The Schwinger functions $(\mathcal{S}_n^{H,G_\alpha})_{n\in\mathbb{N}_0}$ can be analytically extended to Wightman functions $(W_n^{H,G_\alpha})_{n\in\mathbb{N}_0}$ in the sense of (16) and (17).

(ii) The sequence $(W_n^{H,G\alpha})_{n\in\mathbb{N}_0}$ satisfies the axioms (W1), (W2), and (W4)–(W6).

(iii) The Fourier transform of the truncated Wightman functions are given by the formulas (13), (14), and (15), respectively.

(iv) For $0<\alpha<\frac{1}{2}$, $H_1=0$, $H_2<1$, $\hat{W}_2^{H,G_\alpha}=\hat{W}_{2,T}^{H,G_\alpha}$ admits a Källen–Lehmann representation. Therefore, the corresponding Gaussian Euclidean field with covariance function $\mathcal{S}_2^{H_\alpha}$ is reflection positive but not Markov. For $\alpha=\frac{1}{2}$ the corresponding Gaussian Euclidean field is the Markov free field of mass m_0.

(v) The sequence $(W_n^{H,G_\alpha})_{n\in\mathbb{N}_0}$ fulfills the strong spectral condition with mass gap m_0, i.e., $\hat{w}_{n,T}^{H,G_\alpha}$ is supported in the forward mass cones $(V_{m_0}^{\star+})^{n-1}$ for any $n\geq 2$.

Proof: (i) Let us consider the Fourier transformed truncated Wightman functions in (13), (14), and (15). If we now chose coefficients H_n corresponding to a general function H as assumed in the theorem, then the corresponding truncated Wightman functions $W_{n,T}^{H,G_\alpha}$ are analytic continuations of the truncated Schwinger functions $\mathcal{S}_{n,T}^{H,G_\alpha}$ in the sense of (16) and (17). Of course, the corresponding Wightman functions W_n^{H,G_α} are analytic continuations of the truncated Schwinger functions $\mathcal{S}_n^{H,G_\alpha}$ in the same sense.

(ii)–(v) In the case where the Wightman functions correspond to Schwinger functions obtained from convoluted generalized white noise this was proved in Ref. 5, Sec. 7.5. Since in our case we only have a different coefficient H_n, the same is true for a general function H as assumed in the theorem. ■

Now let us return to the question of whether positivity holds or not. In terms of the Schwinger functions, this question has been discussed in Ref. 5, see Sec. IV A. The following theorem is an immediate consequence of the Jost–Schroer theorem.[37–40]

Theorem IV.2: *Let H be as in Theorem III.5,* $H_1=0$, $H_2<1$. *Then the following statements are equivalent*

(i) The sequence of Wightman functions $(W_n^{H,G_{1/2}})_{n\in\mathbb{N}_0}$ *fulfills the positivity condition (W3).*

(ii) For $n\geqslant 3$ *vanish the truncated Wightman functions, i.e.,*

$$W_{n,T}^{H,G_{1/2}}=0, \quad n\geqslant 3.$$

Proof: Since the two-point function $W_2^{H,G_{1/2}}$ is the two-point function of the relativistic free field with mass m_0 and the sequence of Wightman functions $(W_n^{H,G_{1/2}})_{n\in\mathbb{N}_0}$ fulfills (W1), (W2), and (W4)–(W6), see Theorem IV.1, the statement of Theorem IV.2 is just the statement of the Jost–Schroer theorem. ■

Remark IV.3: (i) Theorem IV.2 implies together with the explicit formulas for the Fourier transform of the truncated Wightman functions, see (13), that in the case $\alpha=\frac{1}{2}$ *positivity holds if and only if* $H_n=0$, $n\geqslant 3$.

(ii) For $\alpha=\frac{1}{2}$, Theorem IV.2 implies the negative result on reflection positivity of the Schwinger functions derived in Ref. 5, see Sec. II A.

(iii) Under certain assumptions on the measure in the Källen–Lehmann representation, one can also prove a Jost–Schroer theorem for generalized free fields. It is still an open question, whether this generalization of the Jost–Schroer theorem can be applied to the sequence of Wightman functions $(W_n^{H,G_\alpha})_{n\in\mathbb{N}_0}$, $\alpha\in(0,\frac{1}{2})$; see Theorem IV.1. One has to check whether one can prove a Jost–Schroer theorem for generalized free fields having a Källen–Lehmann representation as the two-point functions $\hat{W}_{2,T}^{H,G_\alpha}$, $\alpha\in(0,\frac{1}{2})$; see Theorem IV.1 (iv).

C. QFT with indefinite metric

In Sec. IV B we performed the analytic continuation from Schwinger functions to Wightman functions. The main interesting object, however, is the underlying quantum field theory. Given a family $(W_n)_{n\in\mathbb{N}}$ obeying (W1)–(W6) by the Wightman reconstruction theorem,[41] there exists an essentially unique field theory obeying the Gårding–Wightman axioms for a single Hermitian scalar field. Since for no nontrivial cases we proved positivity of the sequence of tempered distributions $(W_n^{H,G_\alpha})_{n\in\mathbb{N}_0}$ as in Theorem IV.1, we cannot reconstruct the field theory by the Wightman reconstruction theorem. The positivity conditions is used in order to construct a physical Hilbert space as the closure of the Borchers algebra. This is not possible without the positivity condition.

Morchio and Strocchi[10,12] considered quantum field theories in which not all Wightman axioms are satisfied. For Wightman functions not fulfilling the positivity condition Morchio and Strocchi introduced the so-called modified Wightman axioms of indefinite metric QFT. In their set of axioms, the positivity condition is substituted by the weaker Hilbert space structure condition (HSSC):

W′3 (HSSC): There exists a sequence $(p_n)_{n\in\mathbb{N}}$, where for all $n\in\mathbb{N}, p_n: S(\mathbb{R}^{nd})\to[0,\infty)$ is a Hilbert seminorm, such that

$$|W_{m+n}(f^\star\otimes g)|\leqslant p_m(f)p_n(g),$$

for all $f\in S_{\mathbb{C}}(\mathbb{R}^{dm})$ and $g\in S_{\mathbb{C}}(\mathbb{R}^{dn}), n,m\in\mathbb{N}$.

The HSSC permits the construction of a Hilbert space $\mathcal{K}$ and a scalar, local quantum field ϕ associated to a given collection of tempered distributions $(W_n)_{n\in\mathbb{N}_0}$ fulfilling the modified Wightman axioms (W1), (W2), (W′3), (W4), and (W5). Moreover, in Ref. 10, the following theorem is proved.

Theorem IV.4: *Let* $(W_n)_{n\in\mathbb{N}_0}$ *be a sequence of Wightman functions that fulfill (W1), (W2),* $(W'3)$*, (W4), and (W5). Then there exists:*

(i) a Hilbert space $\mathcal{K}$ *with scalar product* $(\cdot,\cdot)_{\mathcal{K}}$*, a distinguished vacuum vector* $\Omega\in\mathcal{K}$*, and an indefinite inner product* $(\cdot,\cdot)_T$*, which differs from* $(\cdot,\cdot)_{\mathcal{K}}$ *only by a self-adjoint metric operator T with* $T^2=1$*, i.e.,* $(\cdot,\cdot)_T=(\cdot,T\cdot)_{\mathcal{K}}$*;*

(ii) a T-symmetric and local quantum field ϕ*, which is a distribution valued field operator* $\phi(x)$ *acting on a dense core* $\mathcal{D}\subset\mathcal{K}$ *with adjoint* $\phi^{\star}(x)=T\phi(x)T$ *and the commutator*

$$[\phi(x),\phi(y)]=0,$$

for x and y space-like separated. Furthermore, ϕ *is connected with the Wightman functions of the theory by*

$$W_n(x_1,\ldots,x_n)=(\Omega,\phi(x_1)\cdots\phi(x_n)\Omega)_T;\quad \text{and}$$

(iii) a T unitary representation $\mathcal{U}$ *of the orthochonous Poincaré group on* $\mathcal{K}$*, i.e., a representation with* $T\mathcal{U}^{\star}T=\mathcal{U}^{-1}$*, such that* Ω *is invariant under* $\mathcal{U}$ *and* $\phi(x)$ *transforms covariantly,*

$$\mathcal{U}(a,\Lambda)\phi(x)\mathcal{U}(a,\Lambda)^{-1}=\phi(\Lambda^{-1}(x-a)),\quad (a,\Lambda)\in P_+^{\uparrow}(\mathbb{R}^d).$$

Furthermore, $\mathcal{U}$ *fulfills the following spectral condition:*

$$\int_{\mathbb{R}^d}(\Psi_1,\mathcal{U}(a,1)\Psi_2)_T\exp(-iqa)\,da=0,\quad \forall\Psi_1,\Psi_2\in\mathcal{D},$$

if $q\notin V_0^{\star +}$.

A quadruple $((\mathcal{K},(\cdot,\cdot)_{\mathcal{K}},\Omega),T,\phi,\mathcal{U})$ is called a QFT with indefinite metric.

Theorem IV.5: *The Wightman functions* $(W_n^{H,G_\alpha})_{n\in\mathbb{N}_0}$ *as obtained in Theorem IV.1 fulfill the modified Wightman axioms (W1), (W2),* $(W'3)$*, (W4), and (W5) (of Morchio and Strocchi).*

Proof: In Ref. 13, Theorem 4.1, this is proved for the Wightman functions corresponding to the moments of convoluted generalized white noise. The proof is done under the use of explicit formulas for the Fourier transform of the truncated Wightman functions and works for an arbitrary sequence of coefficients $(H_n)_{n\in\mathbb{N}}$; see (13), (14), and (15). Thus, also in our case. ■

In general the seminorms in the HSSC are not invariant under transformations of the orthochonous Poincaré group. Hence, in general, the metric operator T does not commute with $\mathcal{U}(a,\Lambda),(a,\Lambda)\in P_+^{\uparrow}(\mathbb{R}^d)$, and the representation of the orthochonous Poincaré group on $\mathcal{H}$ is not unitary. In our case, however, the seminorms in the HSSC at least can be chosen translations invariant.

Theorem IV.6: For the sequence of Wightman functions $(W_n^{H,G_\alpha})_{n\in\mathbb{N}_0}$ *as in Theorem IV.5, the Hilbert seminorms in the HSSC can be chosen translations invariant. Thus, there exists a Hilbert space structure such that* $[\mathcal{U}(a,1),T]=0$*,* $a\in\mathbb{R}^d$*, and the representation of the translation group* $\mathcal{U}(a,1)$ *is unitary. If P denotes the generator of* $\mathcal{U}(a,1)$*, then spec*$(P)\subset V_0^{\star +}$*.*

Proof: In Ref. 42, Theorem 4.3, this is proved for the Wightman functions corresponding to the moments of convoluted generalized white noise. By the same arguments as in the proof of Theorem IV.5, their proof generalizes to our case. ■

Remark IV.7: (i) We remark that the cluster property of Wightman functions is not an item of the modified Wightman axioms, since, in general, it does not imply the uniqueness of the vacuum and irreducibility of the field algebra as it does in the standard QFT.

(ii) The uniqueness of the vacuum cannot hold if T commutes with $\mathcal{U}(a,1),a\in\mathbb{R}^d$*, and* $T\Omega\notin\mathbb{C}\Omega$*, since in this case* $T\Omega$ *is translations invariant.*

(iii) We observe that there exist sequences of Wightman functions associated to sequences of Schwinger functions that are not moments of measures, fulfilling the modified Wightman axioms of Morchio and Strocchi.

ACKNOWLEDGMENTS

We would like to thank Sergio Albeverio and Yuri Kondratiev for valuable discussions. Financial support of the DFG through Project No. STR 88/6-1 is gratefully acknowledged. Part of this work was done at CCM under the auspices of PRAXIS XXI (2/2.1/MAT/175/94) with support from FEDER through CITMA.

[1] B. Simon, *The* $P(\Phi)_2$ *Euclidean (Quantum) Field Theory* (Princeton University Press, Princeton, 1974).
[2] J. Glimm and A. Jaffe, *Quantum Physics: A Functional Integral point of View* (Springer-Verlag, Berlin, 1987).
[3] R. Høegh-Krohn, Commun. Math. Phys. **21**, 244 (1971).
[4] S. Albeverio and R. Høegh-Krohn, Commun. Math. Phys. **30**, 171 (1973).
[5] S. Albeverio, H. Gottschalk, and J.-L. Wu, Rev. Mod. Phys. **8**, 763 (1996).
[6] K. Osterwalder and R. Schrader, Commun. Math. Phys. **31**, 83 (1973).
[7] R. Streater and A. Wightman, *PCT, Spin, Statistics and All That*, The Mathematical Physics Monograph Series (Benjamin, Reading, 1964).
[8] R. Jost, *The General Theory of Quantized Fields*, Number IV, in Lectures in Applied Mathematics (American Mathematical Society, City, 1965).
[9] N. Bogoliubov, A. Logunov, and R. Todorov, *Introduction to Axiomatic Quantum Field Theory* (Benjamin, Reading, 1975).
[10] G. Morchio and F. Strocchi, Ann. Inst. Henri Poincaré Phys. Theor. A **33**, 251 (1980).
[11] L. Jakobczyk and F. Strocchi, Commun. Math. Phys. **119**, 529 (1988).
[12] F. Strocchi, *Selected Topics on the General Properties of Quantum Field Theory*, Number 51 in Lecture Notes in Physics (World Scientific, Singapore, 1993).
[13] S. Albeverio, H. Gottschalk, and J.-L. Wu, Commun. Math. Phys. **184**, 509 (1997).
[14] T. Hida, Brownian Motion (Springer-Verlag, Berlin, 1980).
[15] Y. Berezansky and Y. Kondratiev, *Spectral Methods in Infinite-Dimensional Analysis* (Kluwer Academic, Dordrecht, 1995), originally in Russian (Naukova Dumka, Kiev, 1988).
[16] T. Hida, H.-H. Kuo, J. Potthoff, and L. Streit, *White Noise. An Infinite Dimensional Calculus* (Kluwer, Dordrecht, 1993).
[17] N. Obata, *White Noise Calculus and Fock Space*, Vol. 1577 of Lecture Notes in Mathematics (Springer-Verlag, Berlin, 1994).
[18] H. Holden, B. Øksendal, J. Ubøe, and T. Zhang, *Stochastic Partial Differential Equations. A Modeling White Noise Functional Approach* (Birkhäuser, Boston, 1996).
[19] H.-H. Kuo, *White Noise Distribution Theory* (CRC Press, Boca Raton, 1996).
[20] S. Albeverio, T. Hida, J. Potthoff, M. Röckner, and L. Streit, Rev. Mod. Phys. **1**, 291 (1990).
[21] S. Albeverio, T. Hida, J. Potthoff, M. Röckner, and L. Streit, Rev. Mod. Phys. **1**, 313 (1990).
[22] S. Albeverio, T. Hida, J. Potthoff, and L. Streit, Phys. Lett. B **217**, 511 (1989).
[23] J. Potthoff and L. Streit, J. Funct. Anal. **111**, 295 (1990).
[24] J. Fröhlich, Helv. Phys. Acta **47**, 265 (1974).
[25] J. Fröhlich, Adv. Math. **23**, 119 (1977).
[26] A. Pietsch, *Nuclear Locally Convex Spaces* (Springer-Verlag, Berlin, 1972).
[27] H. Schaefer, *Topological Vector Spaces* (Springer-Verlag, Berlin, 1971).
[28] I. Gel'fand and N. Vilenkin, *Generalized Functions* (Academic, New York, 1968).
[29] Y. Kondratiev, P. Leukert, and L. Streit, Acta Appl. Math. **44**, 269 (1996).
[30] E. Nelson, J. Funct. Anal. **12**, 97 (1973).
[31] K. Osterwalder and R. Schrader, Commun. Math. Phys. **42**, 281 (1975).
[32] V. Glaser, Commun. Math. Phys. **37**, 257 (1974).
[33] G. Hegerfeldt, Commun. Math. Phys. **35**, 155 (1974).
[34] M. Zinoniev, Commun. Math. Phys. **174**, 1 (1995).
[35] J. Fröhlich, Ann. Inst. Henri Poincaré Phys. Theor. **21**, 271 (1974).
[36] Lukacs, *Characteristic Functions* (Griffin, London, 1970).
[37] R. Jost, ''Properties of Wightman functions,'' in *Lectures on Field Theory and the Many-Body Problem*, edited by E. Caianiello (Academic, New York, 1961).
[38] P. Federbush and K. Johnson, Phys. Rev. **120**, 1926 (1960).
[39] K. Pohlmeyer, Commun. Math. Phys. **12**, 204 (1969).
[40] O. Steinmann, Commun. Math. Phys. **87**, 259 (1982).
[41] A. Wightman, Phys. Rev. **101**, 860 (1956).
[42] S. Albeverio, H. Gottschalk, and J.-L. Wu, Rep. Math. Phys. **40**, 385 (1997).

Derivative expansion of the effective action for quantum electrodynamics in 2+1 and 3+1 dimensions

V. P. Gusynin
Bogolyubov Institute for Theoretical Physics, 252143 Kiev, Ukraine and Institute for Theoretical Physics, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland

I. A. Shovkovy[a)]
Physics Department, University of Cincinnati, Cincinnati, Ohio 45221-0011

(Received 10 May 1999; accepted for publication 1 July 1999)

The derivative expansion of the one-loop effective action in QED_3 and QED_4 (quantum electrodynamics) is considered. The first term in such an expansion is the effective action for a constant electromagnetic field. An explicit expression for the next term containing two derivatives of the field strength $F_{\mu\nu}$, but exact in the magnitude of the field strength, is obtained. The general results for both fermion and scalar electrodynamics are presented. The cases of pure electric and pure magnetic external fields are considered in detail. The Feynman technique for the perturbative expansion of the one-loop effective action in the number of derivatives is developed. © *1999 American Institute of Physics.* [S0022-2488(99)00711-2]

I. INTRODUCTION

Quantum electrodynamics (QED) is known to be the best studied example of quantum field theory. Mainly, this is due to the weakness of the fine structure (coupling) constant, $\alpha \approx 1/137$, which allows us to perform many perturbative calculations as power series in α with an incredibly high accuracy. Despite the smallness of α, even in the realm of quantum electrodynamics, there are some questions that theory has not answered yet. In this paper, in particular, we address the problem of derivation of the low-energy effective action which at present is solved only partially for QED.

The low-energy effective action in quantum electrodynamics describes the dynamics of the electromagnetic field, assuming that the production of the on shell fermions is absent or negligible. Apparently, such a description is self-consistent only if the fermions are massive and the characteristic photon energies are sufficiently small. The mentioned two conditions, as is clear, are necessary to suppress the process of the particle–antiparticle pair creation (on-shell).

Intuitively, the low-energy effective theory is obtained from quantum electrodymanics by ''integrating out'' the fermion field. After doing so, one arrives at a nonlinear theory that involves only the electromagnetic field degrees of freedom. In terms of the S-matrix language, one considers just those processes in QED which contain only photons among the asymptotic scattering states. The fermions, on the other hand, appear only through the internal loops by producing all kinds of photon vertices.

The problem of deriving the effective action is an old one. Its roots go back to the well known papers of Heisenberg and Euler,[1] and Weisskopf.[2] There, for the first time, the effective action in QED (for the case of a constant electromagnetic field) was derived. From the viewpoint of application, the derived effective action contains, for example, the information on the photon–photon scattering at the tree level. It was this scattering process, in fact, that motivated consideration of the problem in Refs. 1,2, in the first place. Later, some further progress was achieved by Schwinger[3] who, by using the proper time technique, rederived the result of Refs. 1,2 and, in

[a)]On leave of absence from Bogolyubov Institute for Theoretical Physics, 252143 Kiev, Ukraine.

addition, gave a nice interpretation to the imaginary part of the effective action in the case of a constant electric field.

Obviously, the next most natural step in deriving the low-energy effective action in QED would be to take into account the effect of small deviations from the constant configuration of the field. In other words, the problem is to obtain the effective action as an expansion in powers of derivatives of the field strength. It turns out, however, that the latter is very difficult to accomplish (see Refs. 4,5 for some early attempts in this direction) unless the weak field approximation is used. In this connection it is appropriate to mention that, in the weak field limit, the expansion is known up to four derivatives with respect to the field strength.[6] Our approach, on the other hand, does not involve any assumptions about the weakness of the background field.

A real progress in solving the problem started with the result of Ref. 7, where an elaborated method, which, in principle, leads to a general result for the derivative expansion in QED, was presented. Because of the complicated character of the method, however, the explicit expression applicable to the most general case of the electromagnetic field background was not presented there. Recently, the derivative expansion of the effective action was obtained in the case of (2+1)-dimensional QED.[8] This latter is a quite general result, containing all the terms quadratic in derivatives of the field strength with respect to the space–time coordinates. Finally, in our previous paper,[9] we obtained a similar result for the effective action but in (3+1)-dimensional QED. As in the (2+1)-dimensional case, it was given in a covariant form valid for the most general constant component of the electromagnetic field background what, as we will see later, is a much more complicated problem than that in 2+1 dimensions.

For completeness, we mention that some related interesting results were obtained in Ref. 10 for QED and in Refs. 11,12 for non-Abelian gauge theories.

In this paper we extend our method, which was originally presented for the case of (3+1)-dimensional QED,[9] to QED in 2+1 dimensions. In particular, we obtain the derivative expansion of the effective action which includes up to two space–time derivatives of the electromagnetic field and, further, we formulate the Feynman rules for the perturbative expansion of the one-loop effective action in the number of derivatives. We also derive the explicit expressions for the derivative corrections to the imaginary part of the effective action in an external electric field. And finally, as a by-product, we resolve the controversy posed in Ref. 13 where a result different from that of Ref. 8 was presented.

The paper is organized as follows. In Sec. II we outline the general method developed in our previous paper.[9] Section III is devoted to solving some technical problems in dealing with functions of the matrix argument $F_{\mu\nu}$. Then, in Secs. IV and VII, we present the main results of our paper, namely, the derivative expansions for spinor and scalar QED, respectively. In Secs. V–VI and Secs. VIII–IX we calculate the derivative expansions for two particular cases of the external electromagnetic field, the purely magnetic and purely electric backgrounds, in both 2+1 and 3+1 dimensions. Finally, in Sec. X, we develop the Feynman diagram technique for generating the perturbative expansion in the number of derivatives. Four appendices contain different formulas used throughout the main text.

II. DERIVATIVE EXPANSION OF THE ONE-LOOP EFFECTIVE ACTION IN QED

Let us start from the general formalism which was originally developed in Ref. 9 for (3+1)-dimensional quantum electrodynamics. While doing so, we will notice that, to a great extent, the method does not depend on the dimension of the space–time. We will pay special attention to all those places where it does depend.

In this paper we restrict ourselves to the one-loop effective action. This is the same approximation which was used by Schwinger[3] in the case of a constant external electromagnetic field.

As is known, the one-loop effective action in QED reduces to computing the fermion determinant

$$W^{(1)}(A)\equiv\int d^n x\mathcal{L}^{(1)}=-i\ln\det(i\hat{\mathcal{D}}-m)=-\frac{i}{2}\ln\det\left(\mathcal{D}_\mu^2+\frac{e}{2}\sigma_{\mu\nu}F^{\mu\nu}+m^2\right)$$
$$=-\frac{i}{2}\int d^n x\langle x|tr\ln\left(\mathcal{D}_\mu^2+\frac{e}{2}\sigma_{\mu\nu}F^{\mu\nu}+m^2\right)|x\rangle. \tag{1}$$

Here $\hat{\mathcal{D}}=\gamma^\mu\mathcal{D}_\mu$ and the covariant derivative is $\mathcal{D}_\mu=\partial_\mu+ieA_\mu$. By definition, $\sigma_{\mu\nu}=i[\gamma_\mu,\gamma_\nu]/2$ and tr refers to the spinor indices of the Dirac matrices γ_μ. States $|x\rangle$ are the eigenstates of a self-conjugate coordinate operator x_μ. Throughout the paper we use the Minkowski metric, i.e., $\eta_{\mu\nu}=(1,-1,-1)$ or $\eta_{\mu\nu}=(1,-1,-1,-1)$, depending on the actual space-time dimension. And in both 2+1 and 3+1 dimensions, we work with the 4×4 representation of the Dirac γ-matrices.

For calculating the effective action in Eq. (1), we employ a version of the so-called worldline (or string-inspired) formalism developed in Refs. 14–16. Such an approach to an ordinary field theory, based on the path integral over one-dimensional world lines, was extended to the evaluation of Feynman diagrams for Green functions in higher loop orders.[17–19] It has demonstrated its power reproducing known theoretical results in QED while allowing one to invoke new technique to study the theory's behavior in strong coupling regime.[20] For some recent applications of the worldline formalism as well as for an extensive list of references, see Refs. 21 and 22. Note, however, that our method differs from the one commonly used in the literature by a choice of the worldline propagators, and is closer in spirit to the method used in Refs. 17 and 23.

With use of the formal identity $\ln(H+m^2)=-\int_0^\infty\exp[-i\tau(H+m^2)]d\tau/\tau$ for introducing the proper-time coordinate τ, the effective Lagrangian can be represented through the diagonal matrix elements of the operator $U(\tau)=\exp(-i\tau H)$,

$$\mathcal{L}^{(1)}(A)=\frac{i}{2}\int_0^\infty\frac{d\tau}{\tau}e^{-im^2\tau}tr\langle x|\exp(-i\tau H)|x\rangle, \tag{2}$$

where the second order differential operator H is given by

$$H=-\Pi_\mu\Pi^\mu+\frac{e}{2}\sigma_{\mu\nu}F^{\mu\nu}(x),\quad \Pi_\mu=-i\mathcal{D}_\mu. \tag{3}$$

The matrix elements $\langle x|\exp(-i\tau H)|x\rangle$ entering the right-hand side of Eq. (2) may be interpreted as the matrix elements of the evolution operator of a spinning particle with τ and H being the proper time and the Hamiltonian of the particle. The corresponding canonical momenta are P_μ's which obey the commutation relations $[x_\mu,P^\nu]=i\delta_\mu^\nu$ and are defined by $\langle x|P_\mu|y\rangle=-i\partial_\mu\delta(x-y)$ in coordinate representation. Following the standard approach,[24] we represent the transition amplitude $\langle z|U(\tau)|y\rangle$ between points $x(0)=y$ and $x(\tau)=z$ in terms of a path integral over the real and Grassmann coordinates, $x_\mu(t)$ and $\psi_\mu(t)$, as

$$tr\langle z|U(\tau)|y\rangle=N^{-1}\int\mathcal{D}[x(t),\psi(t)]\exp\left\{i\int_0^\tau dt[L_{\rm bos}(x(t))+L_{\rm fer}(\psi(t),x(t))]\right\}, \tag{4}$$

where N is a normalization factor, and

$$L_{\rm bos}(x)=-\frac{1}{4}\frac{dx_\nu}{dt}\frac{dx^\nu}{dt}-eA_\nu(x)\frac{dx^\nu}{dt}, \tag{5}$$

$$L_{\rm fer}(\psi,x)=\frac{i}{2}\psi_\nu\frac{d\psi^\nu}{dt}-ie\psi^\nu\psi^\lambda F_{\nu\lambda}(x). \tag{6}$$

The integration in Eq. (4) goes over trajectories $x^{\mu}(t)$ and $\psi^{\mu}(t)$ parameterized by $t\in[0,\tau]$. The definition of the integration measure assumes the following boundary conditions:

$$x(0)=y,\quad x(\tau)=z,\quad \psi(0)=-\psi(\tau). \tag{7}$$

We choose a special gauge condition for the vector potential $A_{\mu}(x)$, namely, the Fock–Schwinger gauge[25]

$$(x^{\nu}-y^{\nu})A_{\nu}(x)=0, \tag{8}$$

which leads to the series

$$\begin{aligned}A_{\nu}(x)&=\tfrac{1}{2}(x^{\lambda}-y^{\lambda})F_{\lambda\nu}(y)+\tfrac{1}{3}(x^{\lambda}-y^{\lambda})(x^{\sigma}-y^{\sigma})\partial_{\sigma}F_{\lambda\nu}(y)\\&\quad+\tfrac{1}{8}(x^{\lambda}-y^{\lambda})(x^{\sigma}-y^{\sigma})(x^{\mu}-y^{\mu})\partial_{\sigma}\partial_{\mu}F_{\lambda\nu}(y)+\cdots\\&=\sum_{n=0}^{\infty}\frac{(x^{\lambda}-y^{\lambda})(x^{\nu_1}-y^{\nu_1})\cdots(x^{\nu_n}-y^{\nu_n})}{n!(n+2)}\partial_{\nu_1}\partial_{\nu_2}\cdots\partial_{\nu_n}F_{\lambda\nu}(y).\end{aligned} \tag{9}$$

This choice of the gauge for the vector potential turns out to be very convenient for developing a perturbative theory in the number of the derivatives of the electromagnetic field with respect to the space–time coordinates.

Carrying out the change of the variable $x(t)$ for $x'(t)=x(t)-y$ in the path integral in Eq. (4) (henceforth we omit the prime) and substituting Eq. (9) into Eq. (4), we obtain

$$\begin{aligned}tr\langle z|U(\tau)|y\rangle&=N^{-1}\int D[x(t),\psi(t)]\exp\left[i\int_0^{\tau}dt\left(-\frac{1}{4}\frac{dx_{\nu}}{dt}\frac{dx^{\nu}}{dt}-\frac{e}{2}x^{\lambda}F_{\lambda\nu}(y)\frac{dx^{\nu}}{dt}+L_{\text{bos}}^{\text{int}}(x)\right)\right]\\&\quad\times\exp\left[i\int_0^{\tau}dt\left(\frac{i}{2}\psi_{\nu}\frac{d\psi^{\nu}}{dt}-ie\psi^{\nu}\psi^{\lambda}F_{\nu\lambda}(y)+L_{\text{fer}}^{\text{int}}(x,\psi)\right)\right].\end{aligned} \tag{10}$$

The new boundary conditions for $x(t)$ are $x(0)=0$ and $x(\tau)=z-y$. Notice, that $F_{\mu\nu}$ in Eq. (10) does not depend on $x(t)$. As follows from Eqs. (5), (6), and (9), the expressions for the interacting terms, $L_{\text{bos}}^{\text{int}}(x)$ and $L_{\text{fer}}^{\text{int}}(x,\psi)$, containing derivatives of $F_{\mu\nu}$ with respect to coordinates, take the form

$$\begin{aligned}L_{\text{bos}}^{\text{int}}(x)&=\sum_{n=1}^{\infty}\frac{eF_{\nu_0\nu_1,\nu_2\cdots\nu_{n+1}}}{n!(n+2)}\frac{dx^{\nu_0}}{dt}x^{\nu_1}(t)\cdots x^{\nu_{n+1}}(t)\\&=\frac{e}{3}F_{\nu\lambda,\sigma}\frac{dx^{\nu}}{dt}x^{\lambda}x^{\sigma}+\frac{e}{8}F_{\nu\lambda,\sigma\kappa}\frac{dx^{\nu}}{dt}x^{\lambda}x^{\sigma}x^{\kappa}+\cdots,\end{aligned} \tag{11}$$

$$\begin{aligned}L_{\text{fer}}^{\text{int}}(x,\psi)&=-\sum_{n=1}^{\infty}\frac{i}{n!}eF_{\lambda\mu,\nu_1\cdots\nu_n}\psi^{\lambda}(t)\psi^{\mu}(t)x^{\nu_1}(t)\cdots x^{\nu_n}(t)\\&=-ieF_{\nu\lambda,\sigma}\psi^{\nu}\psi^{\lambda}x^{\sigma}-\frac{ie}{2}F_{\nu\lambda,\sigma\kappa}\psi^{\nu}\psi^{\lambda}x^{\sigma}x^{\kappa}+\cdots.\end{aligned} \tag{12}$$

Here we use the conventional notation for the partial derivatives

$$F_{\lambda\mu,\nu_1\nu_2\cdots\nu_n}(x)=\partial_{\nu_1}\partial_{\nu_2}\cdots\partial_{\nu_n}F_{\lambda\mu}(x). \tag{13}$$

Now we see that the problem of obtaining the derivative expansion reduces to the evaluation of the path integral in Eq. (10) in the framework of the perturbative theory with an infinite number of interacting terms given in Eqs. (11) and (12). Fortunately, for computing the effective action that includes only a finite number of the derivatives, it is sufficient to consider only a finite number of the interacting terms. Later, we shall restrict ourselves to obtaining only the two-derivative terms in the action. So far, we continue developing the scheme for the most general case.

As usual, introducing real and Grassmann external sources, the matrix elements of the evolution operator can be represented as follows:

$$tr\langle z|U(\tau)|y\rangle=\exp\left\{ i\int_0^\tau dt\left[L_{\rm bos}^{\rm int}\left(\frac{1}{i}\,\frac{\delta}{\delta\eta(t)}\right)+L_{\rm fer}^{\rm int}\left(\frac{1}{i}\,\frac{\delta}{\delta\eta(t)},-\frac{\delta}{\delta\xi(t)}\right)\right]\right\} \times Z_\tau[\eta,\xi](z;y)\Big|_{\eta=0,\xi=0}, \tag{14}$$

where the generating functional is just the Gaussian path integral,

$$Z_\tau[\eta,\xi](z;y)=N^{-1}\int D[x(t),\psi(t)]\exp\left[\frac{i}{2}\int_0^\tau dt\left(-\frac{1}{2}\,\frac{dx_\nu}{dt}\,\frac{dx^\nu}{dt}-ex^\lambda F_{\lambda\nu}(y)\,\frac{dx^\nu}{dt}+2\,\eta_\nu x^\nu\right)\right] \times\exp\left[-\frac{1}{2}\int_0^\tau dt\left(\psi_\nu\frac{d\psi^\nu}{dt}-2e\,\psi^\nu\psi^\lambda F_{\nu\lambda}(y)+2\,\xi_\nu\psi^\nu\right)\right]. \tag{15}$$

The calculation of this generating functional reduces to obtaining the ''classical'' trajectories for $x_\nu(t)$ and $\psi_\nu(t)$, satisfying the appropriate boundary conditions, and to computing the determinants of the one-dimensional differential operators,

$$O_1=\frac{\eta_{\mu\nu}}{2}\,\frac{d^2}{dt^2}-eF_{\mu\nu}\frac{d}{dt},\quad \text{and}\ \ O_2=i\,\eta_{\mu\nu}\frac{d}{dt}-2ieF_{\mu\nu}, \tag{16}$$

defined on the interval $[0,\tau]$ with the periodic and antiperiodic boundary conditions for their eigenstates, respectively.

The ''classical'' trajectories are easily obtained by solving the equations of motion that the bosonic and Grassmanian worldline actions in Eq. (15) require. So, we arrive at

$$x_{\rm cl}^\mu(t)=\left(\frac{e^{2eFt}-1}{e^{2eF\tau}-1}\right)^{\mu\nu}(z-y)_\nu +\int_0^\tau dt'\left(\frac{e^{2eFt}-1}{e^{2eF\tau}-1}\,\frac{(e^{2eF(\tau-t')}-1)}{eF}-\theta(t-t')\,\frac{(e^{2eF(t-t')}-1)}{eF}\right)^{\mu\nu}\eta_\nu(t'), \tag{17}$$

and

$$\psi_{\rm cl}^\mu(t)=\int_0^\tau dt'\left(e^{2eF(t-t')}\left(\theta(t-t')-\frac{1}{1+e^{-2eF\tau}}\right)\right)^{\mu\nu}\xi_\nu(t'). \tag{18}$$

Then, the result of the path integration in Eq. (15) for the case of the coincident arguments $z=y=x$ reads

$$Z_\tau[\eta,\xi](x;x)=C_0\sqrt{\frac{\det(O_2)}{\det'(O_1)}}\exp\left(\frac{i}{2}S_{\rm cl}^{\rm bos}[\eta]-\frac{1}{2}S_{\rm cl}^{\rm fer}[\xi]\right), \tag{19}$$

where the normalization constant C_0 should be determined by comparing the result with the Schwinger's one, or by satisfying the normalization condition

$$Z_{r=0}[\eta,\xi](z;y)=\delta(z-y), \tag{20}$$

which is equivalent to the operator equality $U(0)=1$. The prime in Eq. (19) denotes skipping a zero mode in the definition of the determinant. With our normalization convention for the determinants (see the next section), it is easy to check that the overall factor $C_0=-i/(2\pi\tau)^2$ in 3+1 dimensions and $C_0=\exp[-i\pi/4]/[2(\pi\tau)^{3/2}]$ in 2+1 dimensions.

The expressions for $S_{\rm cl}^{\rm bos}$ and $S_{\rm cl}^{\rm fer}$ are quadratic forms in the external sources

$$S_{\rm cl}^{\rm bos}[\eta]=\int_0^\tau dt_1\int_0^\tau dt_2\,\eta_\nu(t_1)D_\lambda^\nu(t_1,t_2)\,\eta^\lambda(t_2), \tag{21}$$

$$S_{\rm cl}^{\rm fer}[\xi]=\int_0^\tau dt_1\int_0^\tau dt_2\,\xi_\nu(t_1)S_\lambda^\nu(t_1,t_2)\,\xi^\lambda(t_2), \tag{22}$$

where the Green functions are given in terms of functions of the matrix argument $F_{\mu\nu}$,

$$D(t_1,t_2)=\frac{1}{2eF}\left[\epsilon(t_1-t_2)(1-e^{2eF(t_1-t_2)})+\coth(eF\tau)(1+e^{2eF(t_1-t_2)})-\frac{e^{eF(\tau-2t_2)}+e^{eF(2t_1-\tau)}}{\sinh(eF\tau)}\right], \tag{23}$$

$$S(t_1,t_2)=\tfrac{1}{2}[\epsilon(t_1-t_2)-\tanh(eF\tau)]e^{2eF(t_1-t_2)}. \tag{24}$$

Substitution of Eqs. (19), (23), and (24) into Eq. (14) leads to the expression for $tr\langle x|U|x\rangle$. After expanding the exponent in powers of the operator valued interacting terms, $L_{\rm bos}^{\rm int}$ and $L_{\rm fer}^{\rm int}$ (containing functional derivatives with respect to the sources $\eta_\mu(t)$ and $\xi_\mu(t)$), one has to calculate the result of the derivative action on the generating functional. Starting from this point, we have to restrict ourselves to a specific finite number of the derivatives in the effective action. As we mentioned before, in this paper we are interested in the two-derivative terms (see Sec. X for some discussions on computing the higher order approximations). Therefore, we obtain

$$tr\langle x|U(\tau)|x\rangle=\left(1+i\int_0^\tau dt[V_2(t)+W_2(t)]-\frac{1}{2}\int_0^\tau\int_0^\tau dt_1dt_2[V_1(t_1)V_1(t_2)+W_1(t_1)W_1(t_2)]-\int_0^\tau\int_0^\tau dt_1dt_2V_1(t_1)W_1(t_2)\right)Z_\tau[\eta,\xi](x,x)\Big|_{\eta=0,\xi=0}, \tag{25}$$

where, as follows from Eqs. (11), (12), and (14), the vertex generating operators are

$$\begin{aligned}
V_1(t)&=\frac{i}{3}eF_{\nu\lambda,\mu}\lim_{t_0\to t}\frac{d}{dt_0}\,\frac{\delta^3}{\delta\eta_\nu(t_0)\,\delta\eta_\lambda(t)\,\delta\eta_\mu(t)},\\
V_2(t)&=\frac{1}{8}eF_{\nu\lambda,\mu\kappa}\lim_{t_0\to t}\frac{d}{dt_0}\,\frac{\delta^4}{\delta\eta_\nu(t_0)\,\delta\eta_\lambda(t)\,\delta\eta_\mu(t)\,\delta\eta_\kappa(t)},\\
W_1(t)&=-eF_{\nu\lambda,\mu}\frac{\delta^2}{\delta\xi_\nu(t)\,\delta\xi_\lambda(t)}\,\frac{\delta}{\delta\eta_\mu(t)},\\
W_2(t)&=\frac{i}{2}eF_{\nu\lambda,\mu\kappa}\frac{\delta^2}{\delta\xi_\nu(t)\,\delta\xi_\lambda(t)}\,\frac{\delta^2}{\delta\eta_\mu(t)\,\delta\eta_\kappa(t)}.
\end{aligned} \tag{26}$$

Substituting the generating functional (19) which depends on the Green functions (23) and (24), we rewrite Eq. (25) in the form

$$
\begin{aligned}
tr\langle x|U(\tau)|x\rangle = C_0\sqrt{\frac{\det(O_2)}{\det'(O_1)}}\Bigg\{ 1-\frac{i}{8}eF_{\nu\lambda,\mu\kappa}\int_0^\tau dt[\dot{D}^{\nu\lambda}(t,t)D^{\mu\kappa}(t,t)+\dot{D}^{\nu\mu}(t,t)D^{\lambda\kappa}(t,t) \\
+\dot{D}^{\nu\kappa}(t,t)D^{\lambda\mu}(t,t)+4S^{\nu\lambda}(t,t)D^{\mu\kappa}(t,t)] \\
-\frac{i}{18}e^2F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\int_0^\tau\int_0^\tau dt_1dt_2[9D^{\mu\rho}(1,2)(S^{\kappa\sigma}(2,2)S^{\lambda\nu}(1,1) \\
-2S^{\kappa\lambda}(2,1)S^{\sigma\nu}(2,1))+6S^{\sigma\kappa}(2,2)(\dot{D}^{\nu\lambda}(1,1)D^{\mu\rho}(1,2)+\dot{D}^{\nu\mu}(1,1)D^{\lambda\rho}(1,2) \\
+\dot{D}^{\nu\rho}(1,2)D^{\lambda\mu}(1,1))+\dot{D}^{\nu\lambda}(1,1)\dot{D}^{\sigma\kappa}(2,2)D^{\mu\rho}(1,2)+2\dot{D}^{\nu\lambda}(1,1) \\
\times(\dot{D}^{\sigma\rho}(2,2)D^{\mu\kappa}(1,2)+\dot{D}^{\sigma\mu}(2,1)D^{\kappa\rho}(2,2))+\dot{D}^{\nu\mu}(1,1)\dot{D}^{\sigma\rho}(2,2)D^{\lambda\kappa}(1,2) \\
+2\dot{D}^{\nu\kappa}(1,2)(\dot{D}^{\sigma\rho}(2,2)D^{\lambda\mu}(1,1)+\dot{D}^{\sigma\mu}(2,1)D^{\lambda\rho}(1,2)) \\
+\dot{D}^{\nu\kappa}(1,2)\dot{D}^{\sigma\lambda}(2,1)D^{\mu\rho}(1,2)+\dot{D}^{\nu\rho}(1,2)\dot{D}^{\sigma\mu}(2,1)D^{\lambda\kappa}(1,2) \\
+\ddot{D}^{\nu\sigma}(1,2)(D^{\lambda\mu}(1,1)D^{\kappa\rho}(2,2)+D^{\lambda\kappa}(1,2)D^{\mu\rho}(1,2)+D^{\lambda\rho}(1,2)D^{\mu\kappa}(1,2))]\Bigg\}.
\end{aligned}
\tag{27}
$$

Here the dotted functions are defined by the expressions

$$\dot{D}^{\mu\nu}(1,2)\stackrel{\text{def}}{=}\frac{\partial}{\partial t_1}D^{\mu\nu}(t_1,t_2), \tag{28}$$

$$\ddot{D}^{\mu\nu}(1,2)\stackrel{\text{def}}{=}\frac{\partial^2}{\partial t_1\partial t_2}D^{\mu\nu}(t_1,t_2), \tag{29}$$

$$\dot{D}^{\mu\nu}(t,t)\stackrel{\text{def}}{=}\lim_{t_0\to t}\frac{\partial}{\partial t_0}D^{\mu\nu}(t_0,t). \tag{30}$$

Having the representation (27) together with the Green functions (23) and (24), one is left with a need to perform the integrations over the proper time. This latter, however, may look like a rather complicated problem due to the necessity to disentangle the Lorentz indices while doing the integration. In the next section, we show how this problem can be solved.

III. HOW TO DEAL WITH FUNCTIONS OF THE MATRIX ARGUMENT $F_{\mu\nu}$

In the previous section we developed the general method for calculation the derivative expansion in QED. However, there was not given an explicit final expression, since we needed a technique dealing with functions of the matrix argument $F_{\mu\nu}$. Below we show, following the method of Ref. 26, how to deal with those functions as well as how to calculate the determinants of the differential operators in Eq. (16).

Let us begin by introducing notations that we are going to use below. When working with the electromagnetic field strength tensor, it is usually very convenient to introduce the invariants built of the field strength. In (3+1)-dimensional theory, the standard choice of the two independent invariants reads

$$\mathcal{F}=-\tfrac{1}{4}F^{\mu\nu}F_{\mu\nu}, \quad \mathcal{G}=\tfrac{1}{8}\epsilon^{\mu\nu\lambda\kappa}F_{\lambda\kappa}F_{\mu\nu}. \tag{31}$$

In our calculations, though, it will be more convenient to work with the following couple of invariants

$$K_+ = \sqrt{\sqrt{\mathcal{F}^2+\mathcal{G}^2}+\mathcal{F}}, \quad K_- = \sqrt{\sqrt{\mathcal{F}^2+\mathcal{G}^2}-\mathcal{F}}. \tag{32}$$

As for the (2+1)-dimensional theory, there exists only one independent invariant built of the electromagnetic field strength, and it is given by the expression analogous to $\mathcal{F}$ in Eq. (31).

Now we proceed to the case of (3+1)-dimensional QED. It is this case that was considered in Ref. 26. The authors of that paper introduced the set of matrices $A^{\nu\lambda}_{(j)}$ with $j \in \{1,2,3,4\}$,

$$A_{(j)\mu\nu} = \frac{-\overline{f}_j^2 \eta_{\mu\nu} + f_j F_{\mu\nu} + F^2_{\mu\nu} - i \overline{f}_j \overset{*}{F}_{\mu\nu}}{2(f_j^2 - \overline{f}_j^2)}, \tag{33}$$

where

$$f_{1,2} = \pm i K_-, \quad f_{3,4} = \pm K_+; \tag{34}$$

$$\overline{f}_{1,2} = \mp K_+, \quad \overline{f}_{3,4} = \mp i K_-. \tag{35}$$

The main property of the matrices (33) that we are interested in are their (left and right) contractions with the field strength tensor,

$$F^{\nu\lambda} A_{(i)\lambda\mu} = A^{\nu\kappa}_{(i)} F_{\kappa\mu} = f_i A^{\nu}_{(i)\mu}. \tag{36}$$

Other useful properties of these matrices that will be used below are

$$\sum_j A^{\mu\nu}_{(j)} = \eta^{\mu\nu}, \quad A^{\mu}_{(j)\mu} = 1, \quad A^{\mu\nu}_{(k)} A_{(j)\nu\lambda} = \delta_{kj} A^{\mu}_{(j)\lambda}. \tag{37}$$

As follows from the property in Eq. (36), for any function $\Phi(F)$ of the tensor argument $F_{\mu\nu}$, we get

$$\Phi(F)_{\mu\nu} = \sum_j A_{(j)\mu\nu} \Phi(f_{(j)}). \tag{38}$$

Matrices with similar properties can also be introduced for (2+1)-dimensional tensor $F_{\mu\nu}$ as well. Indeed, the following set of matrices:

$$A^{\mu\nu}_{(\pm 1)} = \frac{1}{2}\left(\frac{(F^2)^{\mu\nu}}{2\mathcal{F}} \pm \frac{F^{\mu\nu}}{\sqrt{2\mathcal{F}}} \right), \quad A^{\mu\nu}_{(0)} = \eta^{\mu\nu} - \frac{(F^2)^{\mu\nu}}{2\mathcal{F}} \tag{39}$$

in the (2+1)-dimensional case have properties similar to those in Eqs. (36) and (37). As is easy to check directly, their eigenvalues are

$$f_{\pm 1} = \pm \sqrt{2\mathcal{F}}, \quad f_0 = 0. \tag{40}$$

In particular, for the Green functions (23) and (24), which are functions of the tensor argument $F_{\mu\nu}$, we obtain the following representations:

$$D^{\nu\lambda}(t_1,t_2)=\sum_j A^{\nu\lambda}_{(j)}\frac{1}{2ef_j}\Bigg[\epsilon(t_1-t_2)(1-e^{2ef_j(t_1-t_2)})+\coth(ef_j\tau)(1+e^{2ef_j(t_1-t_2)}) -\frac{e^{ef_j(\tau-2t_2)}+e^{ef_j(2t_1-\tau)}}{\sinh(ef_j\tau)}\Bigg], \tag{41}$$

$$S^{\nu\lambda}(t_1,t_2)=\sum_j A^{\nu\lambda}_{(j)}\frac{1}{2}(\epsilon(t_1-t_2)-\tanh(ef_j\tau))\exp[2ef_j(t_1-t_2)]. \tag{42}$$

As is seen, in the case of vanishing field, the propagators $D^{\nu\lambda}(t_1,t_2)$ and $S^{\nu\lambda}(t_1,t_2)$ coincide with those used in Refs. 17 and 23.

Another problem is related to calculating the determinants of the operators (16). The latter are nothing else but products of all eigenvalues of the operators. Once again, making use of the matrices in Eq. (33) or in Eq. (39) for (3+1)- or (2+1)-dimensional cases, respectively, we look for the eigenvectors of the operators O_1 and O_2 in the form

$$x^{\nu}_{(j)}(t)=A^{\nu}_{(j)\lambda}a^{\lambda}\phi(t), \tag{43}$$

$$\psi^{\nu}_{(j)}(t)=A^{\nu}_{(j)\lambda}\xi^{\lambda}\eta(t), \tag{44}$$

where a^{λ} and ξ^{λ} are constant nonzero vectors, ϕ and η are scalar functions of t. As a result, the problem of obtaining eigenvalues reduces to solving ordinary differential equations for the scalar functions ϕ and η with appropriate boundary conditions.

Now, it is easy to check that, up to an unimportant constant, the corresponding determinants read (note that we skip a zero mode of the operator O_1),

$$\det{}'^{(3+1)}(O_1)=\frac{\sinh^2(e\tau K_+)}{(e\tau K_+)^2}\frac{\sin^2(e\tau K_-)}{(e\tau K_-)^2}, \tag{45}$$

$$\det{}^{(3+1)}(O_2)=\cosh^2(e\tau K_+)\cos^2(e\tau K_-), \tag{46}$$

in the case of QED in 3+1 dimensions, and

$$\det{}'^{(2+1)}(O_1)=\frac{\sinh^2(e\tau\sqrt{2\mathcal{F}})}{(e\tau\sqrt{2\mathcal{F}})^2}, \tag{47}$$

$$\det{}^{(2+1)}(O_2)=\cosh^2(e\tau\sqrt{2\mathcal{F}}), \tag{48}$$

in the case of QED in 2+1 dimensions. To obtain these results we used the following formulas for infinite products:[27]

$$\prod_{n=1}^{\infty}\left(1+\frac{x^2}{\pi^2n^2}\right)=\frac{\sinh x}{x},\quad \prod_{n=1}^{\infty}\left(1+\frac{4x^2}{\pi^2(2n+1)^2}\right)=\cosh x, \tag{49}$$

and similar ones with replacement $x\to iy$.

IV. GENERAL RESULTS IN THE SPINOR QED

By making use of the results from the previous section, we can proceed with the calculation of (27).

After substituting the Green functions (41) and (42), as well as the explicit expressions for the determinants of the operators O_1 and O_2, a straightforward, though tedious computation gives the result for the diagonal matrix element of the $U(\tau)$,

$$
\begin{aligned}
tr\langle x|U(\tau)|x\rangle = tr\langle x|U(\tau)|x\rangle_0 \\
\times\Bigg[1 - \frac{i}{8} eF_{\nu\lambda,\mu\kappa} \sum_{j,l} \big(C^V(f_j,f_l)(A_{(j)}^{\nu\lambda}A_{(l)}^{\mu\kappa} + 2A_{(j)}^{\nu\mu}A_{(l)}^{\lambda\kappa}) + 2C^W(f_j,f_l)A_{(j)}^{\lambda\nu}A_{(l)}^{\mu\kappa} \big) \\
- \frac{i}{18} e^2 F_{\nu\lambda,\mu} F_{\sigma\kappa,\rho} \sum_{j,l,k} \big(9C_1^{WW}(f_j,f_l,f_k)A_{(j)}^{\kappa\sigma}A_{(l)}^{\lambda\nu}A_{(k)}^{\mu\rho} \\
+ 9C_2^{WW}(f_j,f_l,f_k)A_{(j)}^{\kappa\lambda}A_{(l)}^{\sigma\nu}A_{(k)}^{\mu\rho} + 6C_1^{VW}(f_j,f_l,f_k)A_{(j)}^{\sigma\kappa}(A_{(l)}^{\nu\lambda}A_{(k)}^{\mu\rho} + A_{(l)}^{\nu\mu}A_{(k)}^{\lambda\rho}) \\
+ 6C_2^{VW}(f_j,f_l,f_k)A_{(j)}^{\sigma\kappa}A_{(l)}^{\nu\rho}A_{(k)}^{\lambda\mu} - C_1^{VV}(f_j,f_l,f_k)(A_{(j)}^{\nu\lambda}A_{(l)}^{\kappa\sigma}A_{(k)}^{\mu\rho} + A_{(j)}^{\nu\mu}A_{(l)}^{\kappa\rho}A_{(k)}^{\lambda\sigma} \\
+ 2A_{(j)}^{\nu\lambda}A_{(l)}^{\kappa\rho}A_{(k)}^{\mu\sigma}) - C_2^{VV}(f_j,f_l,f_k)(A_{(j)}^{\nu\sigma}A_{(l)}^{\kappa\lambda}A_{(k)}^{\mu\rho} + A_{(j)}^{\nu\rho}A_{(l)}^{\kappa\mu}A_{(k)}^{\lambda\sigma} + 2A_{(j)}^{\nu\sigma}A_{(l)}^{\kappa\mu}A_{(k)}^{\lambda\rho}) \\
- 2C_3^{VV}(f_j,f_l,f_k)(A_{(j)}^{\nu\lambda}A_{(l)}^{\kappa\mu}A_{(k)}^{\sigma\rho} + A_{(j)}^{\kappa\rho}A_{(l)}^{\nu\sigma}A_{(k)}^{\lambda\mu}) - C_4^{VV}(f_j,f_l,f_k)A_{(j)}^{\nu\kappa}A_{(l)}^{\lambda\mu}A_{(k)}^{\sigma\rho} \\
- C_5^{VV}(f_j,f_l,f_k)A_{(j)}^{\nu\kappa}(A_{(l)}^{\lambda\sigma}A_{(k)}^{\mu\rho} + A_{(l)}^{\lambda\rho}A_{(k)}^{\mu\sigma}) \big) \Bigg],
\end{aligned}
\tag{50}
$$

where the explicit expressions for the coefficients $C_i^{XY}(\alpha,\beta,\gamma)$ (with $X,Y \in \{V,W\}$) are given in Appendix A and the diagonal matrix elements $tr\langle x|U(\tau)|x\rangle_0$ correspond to the nonderivative case,

$$tr\langle x|U(\tau)|x\rangle_0^{(3+1)} = -\frac{i}{4\pi^2\tau^2}(e\tau K_-)(e\tau K_+)\cot(e\tau K_-)\coth(e\tau K_+) \tag{51}$$

in 3+1 dimensions, and

$$tr\langle x|U_0(\tau)|x\rangle_0^{(2+1)} = \frac{\exp(-i\pi/4)}{2(\pi\tau)^{3/2}}(e\tau\sqrt{2\mathcal{F}})\coth(e\tau\sqrt{2\mathcal{F}}) \tag{52}$$

in 2+1 dimensions.

Equation (50) (along with a similar one for scalar QED) is the main result of our paper. Note that the renormalization of the effective action (2) formally reduces to (i) performing a subtraction (precisely the same as in the original Schwinger's paper[3]) of a term containing no derivatives of field strength with respect to coordinates, and (ii) changing all bare quantities for the renormalized ones, $e \to e_R$ and $A_\mu \to A_\mu^R$, defined as follows:

$$e_R = Z_3^{1/2} e, \quad A_\mu^R = Z_3^{-1/2} A_\mu, \quad Z_3^{-1} = 1 + \mathbf{C} e^2, \tag{53}$$

where

$$\mathbf{C}^{(3+1)} = \frac{1}{12\pi^2}\int_{1/\Lambda^2}^{\infty}\frac{ds}{s}\exp(-sm^2), \tag{54}$$

$$\mathbf{C}^{(2+1)} = \frac{1}{6\pi^{3/2}}\int_0^{\infty}\frac{ds}{\sqrt{s}}\exp(-sm^2) = \frac{1}{6\pi m}, \tag{55}$$

and Λ is an ultraviolet cutoff in (3+1)-dimensional QED.

After subtraction and conversion to the renormalized quantities the effective action becomes finite in the limit $\Lambda \to \infty$. Since the derivative part of the effective action depends on e and A_μ only through the product $eA_\mu = e_R A_\mu^R$ it does not change its form and no further renormalization is required to make the derivative part well defined (below we use only renormalized quantities, although we always omit the script ''R'' in their notation).

By using the asymptotic behavior of the coefficient functions (given in Appendix A), one easily finds the following expansion of $tr\langle x|U(\tau)|x\rangle$ in powers of τ:

$$tr\langle x|U(\tau)|x\rangle = tr\langle x|U(\tau)|x\rangle_0 \left[1 + \frac{ie^2\tau^3}{20} F^{\nu\lambda} F_{\nu\lambda,\mu}{}^{\mu} + \frac{ie^2\tau^3}{180} \left(\frac{7}{2} F^{\nu\lambda,\mu} F_{\nu\lambda,\mu} - F^{\nu\lambda,}{}_{\lambda} F_{\nu\mu,}{}^{\mu} \right) + \cdots \right]. \tag{56}$$

As is clear, this is the weak field limit of our general result in spinor QED. In the effective action, the given order in τ results in the two-derivative corrections of the order $1/m^2$,

$$\mathcal{L}_{1/m^2}^{(3+1)\mathrm{spin}} = \frac{\alpha}{720\pi m^2} [18F^{\nu\lambda} F_{\nu\lambda,\mu}{}^{\mu} + 7F^{\nu\lambda,\mu} F_{\nu\lambda,\mu} - 2F^{\nu\lambda,}{}_{\lambda} F_{\nu\mu,}{}^{\mu}], \tag{57}$$

in 3+1 dimensions, and of the order $1/m^3$,

$$\mathcal{L}_{1/m^3}^{(2+1)\mathrm{spin}} = \frac{\alpha}{720 m^3} [18F^{\nu\lambda} F_{\nu\lambda,\mu}{}^{\mu} + 7F^{\nu\lambda,\mu} F_{\nu\lambda,\mu} - 2F^{\nu\lambda,}{}_{\lambda} F_{\nu\mu,}{}^{\mu}], \tag{58}$$

in 2+1 dimensions.

The expansion in Eq. (56) was obtained earlier in the heat kernel approach.[4] While the latter is a perfect tool for deriving the effective action in the weak field limit, it is not very useful when the field becomes strong. Our approach here, on the other hand, is free from such a limitation and the general result in Eq. (50) contains all the two derivative terms like $\partial F \partial F (F/m^2)^n$, where n is an arbitrary positive integer and the Lorentz indices (not shown) are contracted in all possible ways. To substantiate this claim, we present the next to leading terms of the weak field expansion in Eq. (B1) in Appendix B.

As we saw above, the formal expansion in τ corresponds to an expansion of the effective action in the inverse powers of the mass parameter. This means that, while making use of such an expansion, one cannot get any reliable results in the limit of the vanishing fermion mass. This, in particular, is the main reason why the authors of Ref. 13, who used an expression like (56), came to a wrong conclusion about the absence of corrections to the one-loop effective action coming from inhomogeneities of a static magnetic field when $m \to 0$. Such a conclusion ''contradicts'' the result of Ref. 8. The latter, as we will see, completely agrees with our result for the derivative expansion.

V. SPINOR QED IN 2+1 DIMENSIONS

Let us consider the case of the purely magnetic field background to which a special attention was paid in Ref. 8. To proceed with analyzing this case, note that the electromagnetic field strength tensor takes the following form:

$$F^{\mu\nu}(x) = B(x) \mathbf{F}^{\mu\nu}, \tag{59}$$

where $B(x)$ is a pseudoscalar function coinciding with the magnetic field strength and $\mathbf{F}^{\mu\nu}$ is a constant matrix with the only nonzero components $\mathbf{F}^{12} = -\mathbf{F}^{21} = 1$. As is seen it satisfies the following normalization condition: $\mathbf{F}^{\mu\nu}\mathbf{F}_{\mu\nu} = 2$.

To reduce the general result presented in Eq. (50) for the particular choice of the field given in Eq. (59), we have to use the properties of $A_{(j)}^{\mu\nu}$'s presented in Sec. III. Just to get a feeling of how they work, let us consider an example,

$$F_{\nu\lambda,\mu\kappa}\sum_{j,l} C^W(f_j,f_l)A^{\lambda\nu}_{(j)}A^{\mu\kappa}_{(l)}=\frac{\partial_\mu\partial_\kappa B}{B}\sum_{j,l} C^W(f_j,f_l)f_{(j)}A^{\mu\kappa}_{(l)}$$
$$=2\frac{\partial_\mu\partial_\kappa B}{B}\sqrt{2\mathcal{F}}[C^W(\sqrt{2\mathcal{F}},0)A^{\mu\kappa}_{(0)}$$
$$+C^W(\sqrt{2\mathcal{F}},\sqrt{2\mathcal{F}})(A^{\mu\kappa}_{(-1)}+A^{\mu\kappa}_{(+1)})]$$
$$=-2iC^W(\sqrt{2\mathcal{F}},\sqrt{2\mathcal{F}})(\mathbf{F}^2)^{\mu\kappa}\partial_\mu\partial_\kappa B$$
$$=-2iC^W(\sqrt{2\mathcal{F}},\sqrt{2\mathcal{F}})\sum_{i=1}^{2}\partial_i\partial_i B. \tag{60}$$

In this derivation, we made use of the Bianchi identity. We recall that the latter should be satisfied since the electromagnetic field was introduced in the theory through the vector potential by minimal coupling. The identity itself reads $A^{\mu\nu}_{(0)}\partial_\nu B\equiv 0$. The direct consequence of it is the independence of the magnetic field, for the particular choice (59), on the time coordinate. By noticing that the matrix $A^{\mu\nu}_{(0)}$, as well as any other from the set, does not depend on $B(x)$ we obtain the secondary identity, $A^{\mu\nu}_{(0)}\partial_\mu\partial_\nu B\equiv 0$, by differentiating the original one. It is this last form of the Bianchi identity that was actually used in our derivation in Eq. (60).

The other expressions, similar to that in Eq. (60), along with the functions like $C^W(\sqrt{2\mathcal{F}},\sqrt{2\mathcal{F}})$ are listed in Appendix C.

The final result for the derivative part of the diagonal matrix element (50), for the particular choice of the field configuration in Eq. (59), reads

$$tr\langle x|U(\tau)|x\rangle^{(2+1)}_{\rm der}=-\frac{ie^2(\partial_i B)^2}{(4\pi|eB|)^{3/2}}\frac{1}{\sqrt{\omega}}(3\omega^2Y^4-3\omega Y^3-4\omega^2Y^2+3\omega Y+\omega^2)$$
$$=\frac{ie^2(\partial_i B)^2}{(4\pi|eB|)^{3/2}}\frac{\sqrt{\omega}}{2}\frac{d^3}{d\omega^3}(\omega\coth\omega), \tag{61}$$

where $\omega=i\tau|eB|$, $Y=\coth\omega$, and $(\partial_i B)^2\equiv\Sigma_{i=1}^2\partial_i B\partial_i B$. Substituting the last expression into Eq. (2), we come to the integral representation for the derivative part of the effective Lagrangian (we perform the change of the integration variable τ for $\omega=i\tau|eB|$),

$$\mathcal{L}^{(2+1)\rm spin}_{\rm der}(B)=-\frac{e^2(\partial_i B)^2}{4(4\pi|eB|)^{3/2}}\int_0^\infty\frac{d\omega}{\sqrt{\omega}}\exp\left(-\frac{m^2}{|eB|}\omega\right)\frac{d^3}{d\omega^3}(\omega\coth\omega). \tag{62}$$

The last expression coincides with the result presented in Ref. 8 (note that in notation of Ref. 8, $\partial_i B\partial_i B=4\partial B\bar{\partial}B$). One can be convinced that the integrand in (62) is a negative function that means that inhomogeneities of the magnetic field background, in approximation under consideration (one-loop and two derivatives), lead to the reduction of vacuum energy density for any value of the ratio $m^2/|eB|$. The latter situation does not, however, prove that a spontaneous generation of a nonhomogeneous magnetic field happens in QED since the sign of the two derivative term in the expansion of the effective action is not a sufficient argument for making a conclusion of that kind.[28]

We would like also to give another representation for the derivative part of the Lagrangian in terms of special functions. To get it, we need to perform the integration in (62) by parts (see Eq. (D5) in Appendix D). Here is such a representation,

$$\mathcal{L}_{\rm der}^{(2+1){\rm spin}}(B) = -\frac{e^2(\partial_i B)^2}{\sqrt{2}\pi(4|eB|)^{3/2}}\left[5\zeta\left(-\frac{3}{2},1+\frac{m^2}{2|eB|}\right) - 9\frac{m^2}{2|eB|}\zeta\left(-\frac{1}{2},1+\frac{m^2}{2|eB|}\right) + 3\left(\frac{m^2}{2|eB|}\right)^2\zeta\left(\frac{1}{2},1+\frac{m^2}{2|eB|}\right) + \left(\frac{m^2}{2|eB|}\right)^3\zeta\left(\frac{3}{2},1+\frac{m^2}{2|eB|}\right)\right]. \tag{63}$$

Often, in the limit of large or small values of the external field, it is more convenient to work with the asymptotic expansions of the effective action rather than the exact expression as in Eq. (63). First, let us consider the case $m^2 \ll |eB|$. Then, using the last representation, we easily derive the following asymptotic expansion:

$$\mathcal{L}_{\rm der}^{(2+1){\rm spin}}(B) \simeq -\frac{e^2(\partial_i B)^2}{\sqrt{2}(4\pi|eB|)^{3/2}}\sum_{k=0}^{\infty}\frac{5-2k}{k!}\Gamma\left(k+\frac{1}{2}\right)\zeta\left(k-\frac{3}{2}\right)\left(-\frac{m^2}{2|eB|}\right)^k. \tag{64}$$

In order to get the asymptotic expansion for $m^2 \gg |eB|$, we make use of the integral representation in Eq. (62) and obtain

$$\mathcal{L}_{\rm der}^{(2+1){\rm spin}}(B) \simeq -\frac{e^2(\partial_i B)^2}{2\pi^{3/2}m^3}\sum_{k=0}^{\infty}\frac{B_{2k+4}}{(2k+1)!}\Gamma\left(2k+\frac{3}{2}\right)\left(\frac{2|eB|}{m^2}\right)^{2k}, \tag{65}$$

where B_k are the Bernoulli numbers.

Now, let us consider the case of the purely electric field background. Without losing the generality, we assume that the field is directed along the first axis of the two-dimensional space. Again the field strength tensor is factored similar to (59),

$$F^{\mu\nu}(x) = E(x)\mathbf{F}^{\mu\nu}, \tag{66}$$

where $E(x)$ is the magnitude of the electric field. Now the constant matrix $\mathbf{F}^{\mu\nu}$ has nonzero components $\mathbf{F}^{10} = -\mathbf{F}^{01} = 1$, and satisfies the normalization condition, $\mathbf{F}^{\mu\nu}\mathbf{F}_{\mu\nu} = -2$. The general expression (50) simplifies considerably for our choice of the background field. And the derivative part of that expression now reads

$$tr\langle x|U(\tau)|x\rangle_{\rm der}^{(2+1)}(E) = \frac{i\exp(-i\pi/4)}{(4\pi|eE|)^{3/2}}\frac{e^2(\partial_\parallel E)^2}{\sqrt{\omega}}(3\omega^2Y^4 - 3\omega Y^3 - 4\omega^2Y^2 + 3\omega Y + \omega^2)$$
$$= -\frac{i\exp(-i\pi/4)}{(4\pi|eE|)^{3/2}}\frac{e^2(\partial_\parallel E)^2}{2}\sqrt{\omega}\frac{d^3}{d\omega^3}(\omega\coth\omega), \tag{67}$$

where now $\omega = \tau|eE|$, $Y = \coth\omega$, and $(\partial_\parallel E)^2 \equiv (\partial_0 E\partial_0 E - \partial_1 E\partial_1 E)$. Here we used the Bianchi identity again to show that the electric field does not depend on the second spatial coordinate. Substituting this expression into Eq. (2), we come to the integral representation for the derivative part of the effective Lagrangian,

$$\mathcal{L}_{\rm der}^{(2+1){\rm spin}}(E) = \frac{\exp(-i\pi/4)e^2(\partial_\parallel E)^2}{4(4\pi|eE|)^{3/2}}\int_0^\infty\frac{d\omega}{\sqrt{\omega}}\exp\left(-i\frac{m^2}{|eE|}\omega\right)\frac{d^3}{d\omega^3}(\omega\coth\omega). \tag{68}$$

As expected in the case of an electric field background, this derivative correction to the effective action contains a nonzero imaginary contribution. A convenient representation of the latter can be obtained in the following way. First, in Eq. (68), we switch to a new variable, $z = i\omega$, so that the integration runs along the imaginary axis of z from zero to $i\infty$. Then, we move the integration contour to the real axis of z. As is easy to check, the integrand has poles at $z = \pi n$ $(n = 1,2,...)$. As a result, the real and the imaginary contributions get naturally separated. Indeed, the real part of $\mathcal{L}_{\rm der}^{(2+1){\rm spin}}$ is given by the principal value of the integral along the Re(z) axis, while the imaginary part appears due to the integration along the infinite set of the vanishingly small semi-circles above the poles, $z=\pi n+\epsilon\exp[i(\pi-\phi)]$ (where $0<\phi<\pi$ and $\epsilon\to 0$ at the end). In this way, we easily obtain the imaginary part of the right-hand side in Eq. (68),

$$\mathrm{Im}\,\mathcal{L}_{\mathrm{der}}^{(2+1)\mathrm{spin}}(E)=-\frac{e^2(\partial_\parallel E)^2}{2^8\pi^3|eE|^{3/2}}\sum_{n=1}^{\infty}\frac{1}{n^{5/2}}\exp\left(-\frac{\pi m^2 n}{|eE|}\right)\times\left[15+18\frac{\pi m^2 n}{|eE|}+12\left(\frac{\pi m^2 n}{|eE|}\right)^2+8\left(\frac{\pi m^2 n}{|eE|}\right)^3\right]. \tag{69}$$

We note that the result of the summation in the last expression (as well as in similar formulas later on) can be given in terms of the polylogarithmic function $\mathrm{Li}_\nu(x)$.[29] Equation (69) determines the correction to the probability of the particle-antiparticle pair creation (by definition, the probability density is $\mathcal{W}=2\,\mathrm{Im}\,\mathcal{L}$) in an external electric field due to small inhomogeneities in space–time. We emphasize that the correction due to a time derivative of the field has the ''wrong'' sign, i.e., it works against the particle creation. The gradient in the space direction parallel to the field strength, on the other hand, amplifies the process.

As is known, in the case of constant electric field, the imaginary part of the effective Lagrangian is given by

$$\mathrm{Im}\,\mathcal{L}^{(2+1)\mathrm{spin}}(E)=\frac{|eE|^{3/2}}{4\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^{3/2}}\exp\left(-\frac{\pi m^2}{|eE|}n\right)=\frac{|eE|^{3/2}}{4\pi^2}\mathrm{Li}_{3/2}\left[\exp\left(-\frac{\pi m^2}{|eE|}\right)\right]. \tag{70}$$

This as well as the first correction due to the derivatives remain finite even in the limit of zero fermion mass. Despite of this fact, we still expect that the derivative expansion (with the electric field background) may fail in the limit of vanishingly small mass due to higher orders in the number of derivatives. Below we shall see that the same is true in the spinor QED in 3+1 dimensions as well.

VI. SPINOR QED IN 3+1 DIMENSIONS

As was mentioned at the beginning of the paper, the derivative expansion in QED_4 was also studied in Ref. 7. The result of that paper was presented in an explicit form for the special class of the electromagnetic field configurations,

$$\mathcal{G}=0,\quad F^{\mu\nu}(x)=\Phi(x)\mathbf{F}^{\mu\nu}, \tag{71}$$

where $\Phi(x)$ is a slowly varying function that defines the magnitude of the field, and $\mathbf{F}^{\mu\nu}$ is a constant matrix. For convenience, let us normalize the matrix $\mathbf{F}^{\mu\nu}$ by the condition $\mathbf{F}^{\mu\nu}\mathbf{F}_{\mu\nu}=2$. Then the scalar function $\Phi(x)$ is nothing else but $\sqrt{(-2\mathcal{F})}$. As was shown in our previous paper,[9] the general result for the diagonal matrix element (50) in the case of field (71) reduces to the same result as was presented in Ref. 7,

$$\begin{aligned} tr\langle x|U(\tau)|x\rangle_{\mathrm{der}}^{(3+1)}(\Phi)&=\frac{1}{(4\pi)^2\tau}\frac{\partial_\mu\Phi\,\partial^\mu\Phi}{\Phi^2}(3\omega^2Y^4-3\omega Y^3-4\omega^2Y^2+3\omega Y+\omega^2),\\ &=-\frac{1}{(4\pi)^2\tau}\frac{\partial_\mu\Phi\,\partial^\mu\Phi}{\Phi^2}\frac{\omega}{2}\frac{d^3}{d\omega^3}(\omega\coth\omega),\end{aligned} \tag{72}$$

where $\omega=\tau e\Phi$, $Y=\coth\omega$. As in the (2+1)-dimensional theory, here we used the Bianchi identity, which this time reads

$$(\eta^{\mu\nu}+(\mathbf{F}^2)^{\mu\nu})\partial_\nu\Phi\equiv 0. \tag{73}$$

In the case of magnetic field along the third axis, for example, this condition means that the specified field cannot depend on the time and the third spatial coordinates, while in the case of electric field along the first axis, it cannot depend on the second and third spatial coordinates.

Now, let us consider two particular cases of external field that we studied in 2+1 dimensions; purely magnetic and purely electric field backgrounds. Both of them are just different possibilities of that given in Eq. (71).

Thus, in the case of magnetic field (along the third axis in space) we come to the following integral representation for the derivative part of the effective Lagrangian:

$$\mathcal{L}_{\rm der}^{(3+1)\rm spin}(B) = -\frac{e^2(\partial_i B)^2}{(8\pi)^2|eB|}\int_0^\infty \frac{d\omega}{\omega}\exp\left(-\frac{m^2}{|eB|}\omega\right)\frac{d^3}{d\omega^3}(\omega\coth\omega). \tag{74}$$

Resembling the situation in 2+1 dimensions, inhomogeneities of the external magnetic field tend to reduce vacuum energy density for any value of the ratio $m^2/|eB|$.

Performing integration in the right-hand side of Eq. (74) by parts (see Eq. (D6) in Appendix D), we find the following representation (for the representation of the part of the effective action without derivatives in terms of special functions, see Ref. 30):

$$\begin{aligned}\mathcal{L}_{\rm der}^{(3+1)\rm spin}(B) = &-\frac{e^2(\partial_i B)^2}{(8\pi)^2|eB|}\left[\frac{11}{6}\left(\frac{m^2}{|eB|}\right)^3 + \left(\frac{m^2}{|eB|}\right)^2 - \frac{1}{3}\frac{m^2}{|eB|} - \left(\frac{m^2}{|eB|}\right)^3\psi\left(1+\frac{m^2}{2|eB|}\right)\right.\\ &+24\zeta'\left(-2,1+\frac{m^2}{2|eB|}\right) - 24\frac{m^2}{|eB|}\zeta'\left(-1,1+\frac{m^2}{2|eB|}\right)\\ &\left.+6\left(\frac{m^2}{|eB|}\right)^2\left[\ln\Gamma\left(1+\frac{m^2}{2|eB|}\right)-\ln\sqrt{2\pi}\right]\right].\end{aligned} \tag{75}$$

As $m^2 \ll |eB|$, this expression allows the following asymptotic expansion:

$$\begin{aligned}\mathcal{L}_{\rm der}^{(3+1)\rm spin}(B) \simeq &-\frac{e^2(\partial_i B)^2}{(8\pi)^2|eB|}\left[24\zeta'(-2) + \frac{2m^2}{3|eB|} - \frac{m^4}{2|eB|^2} + \frac{m^6}{3|eB|^3}\right.\\ &\left.-\frac{m^8}{2|eB|^4}\sum_{k=0}^{\infty}\frac{k+1}{k+4}\zeta(k+2)\left(-\frac{m^2}{2|eB|}\right)^k\right],\end{aligned} \tag{76}$$

where $\zeta'(-2) \approx -0.030$. As $m^2 \gg |eB|$, on the other hand, we obtain

$$\mathcal{L}_{\rm der}^{(3+1)\rm spin}(B) \simeq -\frac{e^2(\partial_i B)^2}{(2\pi)^2 m^2}\sum_{k=0}^{\infty}\frac{B_{2k+4}}{2k+1}\left(\frac{2|eB|}{m^2}\right)^{2k}. \tag{77}$$

In case of the electric field along the first axis, on the other hand, we obtain the following expression for the derivative part of the effective action:

$$\mathcal{L}_{\rm der}^{(3+1)\rm spin}(E) = -\frac{ie^2(\partial_\parallel E)^2}{(8\pi)^2|eE|}\int_0^\infty \frac{d\omega}{\omega}\exp\left(-i\frac{m^2}{|eE|}\omega\right)\frac{d^3}{d\omega^3}(\omega\coth\omega). \tag{78}$$

This expression has both real and imaginary part, as always happens in the case of an external electric field. Another representation for it is obtained by analytical continuation of (75) according to the rule $|eB| \to -i|eE|$. The imaginary part though is easily extracted from (78) in a standard way,

$$\mathrm{Im}\,\mathcal{L}^{(3+1)\mathrm{spin}}_{\mathrm{der}}(E)=\frac{e^2(\partial_\parallel E)^2}{2^6\pi^4|eE|}\sum_{n=1}^{\infty}\frac{1}{n^3}\exp\left(-\frac{\pi m^2 n}{|eE|}\right)\left[6+6\frac{\pi m^2 n}{|eE|}+3\left(\frac{\pi m^2 n}{|eE|}\right)^2+\left(\frac{\pi m^2 n}{|eE|}\right)^3\right], \tag{79}$$

which determines a correction to the Schwinger result[3] for the imaginary part of the effective action in a constant electric field,

$$\mathrm{Im}\,\mathcal{L}^{(3+1)\mathrm{spin}}(E)=\frac{(eE)^2}{8\pi^3}\sum_{n=1}^{\infty}\frac{1}{n^2}\exp\left(-\frac{\pi m^2}{|eE|}n\right)=\frac{(eE)^2}{8\pi^3}\mathrm{Li}_2\left[\exp\left(-\frac{\pi m^2}{|eE|}\right)\right]. \tag{80}$$

The result in Eq. (79) is in agreement with that of Ref. 10.

As is easy to establish, both the Schwinger result for a constant field and the first correction due to derivatives are finite in the limit of the vanishing fermion mass. As we argued in the case of the (2+1)-dimensional spinor QED, this may not be the case in higher orders of the perturbative expansion in the number of derivatives.

VII. GENERAL RESULT IN THE SCALAR QED

Now turning to the calculation of the derivative expansion for the scalar electrodynamics, one does not need to repeat all the calculations similar to those done in Sec. IV. In order to see this, we recall that the effective one-loop Lagrangian in this case reads

$$L^{(1)\mathrm{scal}}(x)=-i\int_0^\infty\frac{d\tau}{\tau}\langle x|U_{\mathrm{bos}}(\tau)|x\rangle e^{-im^2\tau}. \tag{81}$$

The evolution connected with the transition amplitude, $\langle z|U_{\mathrm{bos}}(\tau)|y\rangle$, is described now by the Hamiltonian (compare with Eqs. (2) and (3)),

$$H_{\mathrm{bos}}=-\Pi_\mu\Pi^\mu,\quad \Pi_\mu=-i\partial_\mu+eA_\mu(x). \tag{82}$$

Thus, omitting all terms originating from the fermion part in the expression (4), i.e., putting $L^{\mathrm{int}}_{\mathrm{fer}}=0$ in Eqs. (10), (14) and $S^{\mathrm{fer}}_{\mathrm{cl}}=0$ in Eq. (19), we come to the following expression:

$$\begin{aligned}\langle x|U_{\mathrm{bos}}(\tau)|x\rangle=&\langle x|U_{\mathrm{bos}}(\tau)|x\rangle_0\\&\times\Bigg[1-\frac{i}{8}eF_{\nu\lambda,\mu\kappa}\sum_{j,l}C^V(f_j,f_l)(A^{\nu\lambda}_{(j)}A^{\mu\kappa}_{(l)}+2A^{\nu\mu}_{(j)}A^{\lambda\kappa}_{(l)})+\frac{i}{18}e^2F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\\&\times\sum_{j,l,k}(C^{VV}_1(f_j,f_l,f_k)(A^{\nu\lambda}_{(j)}A^{\kappa\sigma}_{(l)}A^{\mu\rho}_{(k)}+A^{\nu\mu}_{(j)}A^{\kappa\rho}_{(l)}A^{\lambda\sigma}_{(k)}+2A^{\nu\lambda}_{(j)}A^{\kappa\rho}_{(l)}A^{\mu\sigma}_{(k)})\\&+C^{VV}_2(f_j,f_l,f_k)(A^{\nu\sigma}_{(j)}A^{\kappa\lambda}_{(l)}A^{\mu\rho}_{(k)}+A^{\nu\rho}_{(j)}A^{\kappa\mu}_{(l)}A^{\lambda\sigma}_{(k)}+2A^{\nu\sigma}_{(j)}A^{\kappa\mu}_{(l)}A^{\lambda\rho}_{(k)})\\&+2C^{VV}_3(f_j,f_l,f_k)(A^{\nu\lambda}_{(j)}A^{\kappa\mu}_{(l)}A^{\sigma\rho}_{(k)}+A^{\kappa\rho}_{(j)}A^{\nu\sigma}_{(l)}A^{\lambda\mu}_{(k)})+C^{VV}_4(f_j,f_l,f_k)A^{\nu\kappa}_{(j)}A^{\lambda\mu}_{(l)}A^{\sigma\rho}_{(k)}\\&+C^{VV}_5(f_j,f_l,f_k)A^{\nu\kappa}_{(j)}(A^{\lambda\sigma}_{(l)}A^{\mu\rho}_{(k)}+A^{\lambda\rho}_{(l)}A^{\mu\sigma}_{(k)})\Bigg].\end{aligned} \tag{83}$$

The coefficients used here are the same as in Eq. (50). As for the nonderivative factors, they have the standard form,

$$\langle x|U_{\mathrm{bos}}(\tau)|x\rangle_0^{(3+1)}=-\frac{i}{(4\pi\tau)^2}\frac{(e\tau K_-)(e\tau K_+)}{\sin(e\tau K_-)\sinh(e\tau K_+)} \tag{84}$$

in 3+1 dimensions, and

$$\langle x|U_{\rm bos}(\tau)|x\rangle_0^{(2+1)} = -\frac{\exp(-i\pi/4)}{(4\pi\tau)^{3/2}} \frac{(e\tau\sqrt{2\mathcal{F}})}{\sinh(e\tau\sqrt{2\mathcal{F}})} \tag{85}$$

in 2+1 dimensions, as can be easily checked by using the expressions for the determinants given in Sec. III and by taking into account the fact that, because of spin degrees of freedom, we had the additional factor 4 for fermions.

In the case of scalar theory, the renormalization of the electromagnetic field and charge is given by the same formulas (53) but this time the corresponding constants read

$$\mathbf{C}^{(3+1)} = \frac{1}{48\pi^2}\int_{1/\Lambda^2}^{\infty} \frac{ds}{s}\exp(-sm^2), \tag{86}$$

$$\mathbf{C}^{(2+1)} = \frac{1}{24\pi^{3/2}}\int_0^{\infty} \frac{ds}{\sqrt{s}}\exp(-sm^2) = \frac{1}{24\pi m}. \tag{87}$$

To get a result of the type as in Ref. 4, one has to expand the coefficient functions in powers of proper time. Thus the expansion for $\langle x|U_{\rm bos}(\tau)|x\rangle$ (weak field limit) reads

$$\langle x|U_{\rm bos}(\tau)|x\rangle = \langle x|U_{\rm bos}(\tau)|x\rangle_0\left[1 - \frac{ie^2\tau^3}{30}F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu} - \frac{ie^2\tau^3}{180}(4F^{\nu\lambda,\mu}F_{\nu\lambda,\mu} + F^{\nu\lambda,}{}_{\lambda}F_{\nu\mu,}{}^{\mu}) + \cdots\right]. \tag{88}$$

This expansion up to the order τ^5 is given in Eq. (B26) in Appendix B. As in the spinor QED, it is useful only in the case of heavy scalar particles (weak fields), when the mass scale is much larger than all other scales in the theory.

In the effective action of scalar QED, the expansion in Eq. (88) corresponds to the following leading two derivative terms

$$\mathcal{L}_{1/m^2}^{(3+1)\rm scal} = \frac{\alpha}{720\pi m^2}[6F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu} + 4F^{\nu\lambda,\mu}F_{\nu\lambda,\mu} + F^{\nu\lambda,}{}_{\lambda}F_{\nu\mu,}{}^{\mu}], \tag{89}$$

in 3+1 dimensions, and

$$\mathcal{L}_{1/m^3}^{(2+1)\rm scal} = -\frac{\alpha}{720m^3}[6F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu} + 4F^{\nu\lambda,\mu}F_{\nu\lambda,\mu} + F^{\nu\lambda,}{}_{\lambda}F_{\nu\mu,}{}^{\mu}], \tag{90}$$

in 2+1 dimensions.

VIII. SCALAR QED IN 2+1 DIMENSIONS

Let us start by considering the case of an external magnetic field as in Eq. (59). This time the derivative part of the general expression (83) reduces to

$$\begin{aligned}\langle x|U_{\rm bos}(\tau)|x\rangle_{\rm der}^{(2+1)} &= -\frac{ie^2(\partial_i B)^2}{4(4\pi|eB|)^{3/2}}\frac{\sqrt{\omega}}{\sinh\omega}(3\omega Y^3 - 3Y^2 - 2\omega Y + 1)\\ &= \frac{ie^2(\partial_i B)^2}{4(4\pi|eB|)^{3/2}}\frac{\sqrt{\omega}}{2}\left(\frac{d^3}{d\omega^3} + \frac{d}{d\omega}\right)\left(\frac{\omega}{\sinh\omega}\right),\end{aligned} \tag{91}$$

where $\omega = i\tau|eB|$, $Y = \coth\omega$. After substituting the last expression into Eq. (81), we come to the integral representation for the derivative part of the effective Lagrangian (after performing the change of integration variable $\tau\to\omega = i\tau|eB|$),

$$\mathcal{L}_{\rm der}^{(2+1){\rm scal}}(B)=\frac{e^2(\partial_i B)^2}{(16\pi|eB|)^{3/2}}\int_0^\infty \frac{d\omega}{\sqrt{\omega}}\exp\left(-\frac{m^2}{|eB|}\omega\right)\left(\frac{d^3}{d\omega^3}+\frac{d}{d\omega}\right)\left(\frac{\omega}{\sinh\omega}\right), \tag{92}$$

which coincides with the result presented in Ref. 8. As in the case of spinor QED, there exists another representation of (92) given in terms of special functions (see Eq. (D11) in Appendix D),

$$\mathcal{L}_{\rm der}^{(2+1){\rm scal}}(B)=\frac{e^2(\partial_i B)^2}{\sqrt{2}\pi(16|eB|)^{3/2}}\left[20\zeta\left(-\frac{3}{2},\frac{1}{2}+\frac{m^2}{2|eB|}\right)-18\frac{m^2}{|eB|}\zeta\left(-\frac{1}{2},\frac{1}{2}+\frac{m^2}{2|eB|}\right)\right.$$
$$\left.+\left(1+3\left(\frac{m^2}{|eB|}\right)^2\right)\zeta\left(\frac{1}{2},\frac{1}{2}+\frac{m^2}{2|eB|}\right)+\frac{1}{2}\left(\frac{m^2}{|eB|}+\left(\frac{m^2}{|eB|}\right)^3\right)\zeta\left(\frac{3}{2},\frac{1}{2}+\frac{m^2}{2|eB|}\right)\right]. \tag{93}$$

Numerical study of the integral in (92) shows that inhomogeneities of magnetic field background, in approximation under consideration (one-loop and two derivatives), lead to decreasing the vacuum energy density for $m^2/|eB|\gtrsim 0.927$ and to increasing that density for $m^2/|eB|\lesssim 0.927$, in accordance with Ref. 8.

Analytically, we can obtain only the limiting cases as we did in spinor electrodynamics. In particular, for $m^2\ll|eB|$, the effective action takes the following asymptotic form:

$$\mathcal{L}_{\rm der}^{(2+1){\rm scal}}(B)\simeq\frac{e^2(\partial_i B)^2}{(16\pi|eB|)^{3/2}}\sum_{k=0}^{\infty}\frac{1}{k!}\left[(2^k-2\sqrt{2})(5-2k)\zeta\left(k-\frac{3}{2}\right)\right.$$
$$\left.+\left(2^k-\frac{1}{\sqrt{2}}\right)(1-2k)\zeta\left(k+\frac{1}{2}\right)\right]\Gamma\left(k+\frac{1}{2}\right)\left(-\frac{m^2}{2|eB|}\right)^k, \tag{94}$$

while for $m^2\gg|eB|$, the expansion reads

$$\mathcal{L}_{\rm der}^{(2+1){\rm scal}}(B)\simeq-\frac{e^2(\partial_i B)^2}{32\pi^{3/2}m^3}\sum_{k=0}^{\infty}\frac{(2^{2k+3}-1)B_{2k+4}+(2^{2k+1}-1)B_{2k+2}}{(2k+1)!}\Gamma\left(2k+\frac{3}{2}\right)\left(\frac{|eB|}{m^2}\right)^{2k}. \tag{95}$$

Now, let us consider the case of electric field background. Without losing the generality, we assume that the field is directed along the first axis of space. We obtain

$$\langle x|U_{\rm bos}(\tau)|x\rangle_{\rm der}^{(2+1)}(E)=\frac{i\exp(-i\pi/4)}{4(4\pi|eE|)^{3/2}}\frac{e^2(\partial_\parallel E)^2\sqrt{\omega}}{\sinh\omega}(3\omega Y^3-3Y^2-2\omega Y+1)$$
$$=-\frac{i\exp(-i\pi/4)}{(16\pi|eE|)^{3/2}}e^2(\partial_\parallel E)^2\sqrt{\omega}\left(\frac{d^3}{d\omega^3}+\frac{d}{d\omega}\right)\left(\frac{\omega}{\sinh\omega}\right), \tag{96}$$

where now $\omega=\tau|eE|$ and $Y=\coth\omega$. Substituting this expression into (81), we come to the integral representation for the derivative part of the effective Lagrangian,

$$\mathcal{L}_{\rm der}^{(2+1){\rm scal}}(E)=-\frac{\exp(-i\pi/4)e^2(\partial_\parallel E)^2}{(16\pi|eE|)^{3/2}}\int_0^\infty\frac{d\omega}{\sqrt{\omega}}\exp\left(-i\frac{m^2}{|eE|}\omega\right)\left(\frac{d^3}{d\omega^3}+\frac{d}{d\omega}\right)\left(\frac{\omega}{\sinh\omega}\right). \tag{97}$$

And we easily find the imaginary part of the expression,

$$\mathrm{Im}\,\mathcal{L}_{\mathrm{der}}^{(2+1)\mathrm{scal}}(E)=\frac{e^2(\partial_{\parallel}E)^2}{2^9\pi^3|eE|^{3/2}}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^{5/2}}\exp\left(-\frac{\pi m^2 n}{|eE|}\right)$$
$$\times\left[15+18\frac{\pi m^2 n}{|eE|}+4\pi^2n^2\left(3\frac{m^4}{|eE|^2}-1\right)+8\frac{m^2\pi^3n^3}{|eE|}\left(\frac{m^4}{|eE|^2}-1\right)\right], \quad (98)$$

which determines the correction to the corresponding result for case of constant electric field,

$$\mathrm{Im}\,\mathcal{L}^{(2+1)\mathrm{scal}}(E)=\frac{|eE|^{3/2}}{8\pi^2}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^{3/2}}\exp\left(-\frac{\pi m^2}{|eE|}n\right)=-\frac{|eE|^{3/2}}{8\pi^2}\mathrm{Li}_{3/2}\left[-\exp\left(-\frac{\pi m^2}{|eE|}\right)\right]. \quad (99)$$

A simple numerical analysis of the derivative correction in Eq. (98) shows that the sum in the right-hand side, being positive for large values of the mass (or small values of the electric field), changes its sign at $m^2\approx 0.721|eE|$. Therefore, unlike the case of spinor QED, the time derivative of the field increases (while the gradient in space decreases) the probability of particle–antiparticle pair creation only for $m^2 \gtrsim 0.721|eE|$.

As in spinor QED, the two-derivative correction to the process of the pair production in scalar QED is convergent even in the limit of the vanishing mass. This observation, of course, is not enough to prove that the derivative expansion is well defined to all orders in the massless theory.

IX. SCALAR QED IN 3+1 DIMENSIONS

The derivative expansion for the electromagnetic field of the form (71) was presented in our previous paper[9] (we just rewrite it in different form),

$$\langle x|U_{\mathrm{bos}}(\tau)|x\rangle_{\mathrm{der}}^{(3+1)}=\frac{1}{(8\pi)^2\tau}\frac{\partial_\mu\Phi\partial^\mu\Phi}{\Phi^2}\frac{\omega}{\sinh\omega}(3\omega Y^3-3Y^2-2\omega Y+1)$$
$$=-\frac{1}{(8\pi)^2\tau}\frac{\partial_\mu\Phi\partial^\mu\Phi}{\Phi^2}\frac{\omega}{2}\left(\frac{d^3}{d\omega^3}+\frac{d}{d\omega}\right)\left(\frac{\omega}{\sinh\omega}\right), \quad (100)$$

with $\omega=\tau e\Phi$, $Y=\coth\omega$.

Now, let us consider the two most interesting particular cases as before. As in the case of the fermion theory presented in Sec. VI, in the case of scalar QED, the derivative part of the effective Lagrangian is easily obtained by using (100) with $\Phi=iB$,

$$\mathcal{L}_{\mathrm{der}}^{(3+1)\mathrm{scal}}(B)=\frac{e^2(\partial_i B)^2}{2(8\pi)^2|eB|}\int_0^\infty\frac{d\omega}{\omega}\exp\left(-\frac{m^2}{|eB|}\omega\right)\left(\frac{d^3}{d\omega^3}+\frac{d}{d\omega}\right)\left(\frac{\omega}{\sinh\omega}\right). \quad (101)$$

And again, as is easy to check, the situation with (101) resembles that in (2+1)-dimensional scalar QED; inhomogeneities of the external magnetic field lead to decreasing the vacuum energy density for large values of the ratio $m^2/|eB|$ $(m^2/|eB|\gtrsim 0.41)$ and to increasing for small values $(m^2/|eB|\lesssim 0.41)$.

In addition to the representation (101), we find the following one (see Eq. (D12) in Appendix D):

$$\mathcal{L}_{\rm der}^{(3+1){\rm scal}}(B)=\frac{e^2(\partial_i B)^2}{2(8\pi)^2|eB|}\left[\frac{11}{6}\left(\frac{m^2}{|eB|}\right)^3-\frac{m^2}{|eB|}\left(1+\left(\frac{m^2}{|eB|}\right)^2\right)\psi\left(\frac{1}{2}+\frac{m^2}{2|eB|}\right)\right.$$
$$+\frac{7}{6}\frac{m^2}{|eB|}+2\left(1+3\left(\frac{m^2}{|eB|}\right)^2\right)\left[\ln\Gamma\left(\frac{1}{2}+\frac{m^2}{2|eB|}\right)-\ln\sqrt{2\pi}\right]$$
$$\left.+24\zeta'\left(-2,\frac{1}{2}+\frac{m^2}{2|eB|}\right)-24\frac{m^2}{|eB|}\zeta'\left(-1,\frac{1}{2}+\frac{m^2}{2|eB|}\right)\right]. \tag{102}$$

In the limit $m^2 \ll |eB|$, this expression allows the following asymptotic expansion:

$$\mathcal{L}_{\rm der}^{(3+1){\rm scal}}(B)\simeq\frac{e^2(\partial_i B)^2}{2(8\pi)^2|eB|}\left[-18\zeta'(-2)-\ln 2+\frac{2m^2}{3|eB|}+\frac{m^6}{3|eB|^3}\right.$$
$$-\frac{m^4}{2|eB|^2}\sum_{k=0}^{\infty}\frac{k+1}{k+2}(2^{k+2}-1)\zeta(k+2)\left(-\frac{m^2}{2|eB|}\right)^k$$
$$\left.-\frac{m^8}{2|eB|^4}\sum_{k=0}^{\infty}\frac{k+1}{k+4}(2^{k+2}-1)\zeta(k+2)\left(-\frac{m^2}{2|eB|}\right)^k\right]. \tag{103}$$

In the limit $m^2 \gg |eB|$, on the other hand, we obtain

$$\mathcal{L}_{\rm der}^{(3+1){\rm scal}}(B)\simeq\frac{e^2(\partial_i B)^2}{(8\pi)^2 m^2}\sum_{k=0}^{\infty}\frac{(2^{2k+3}-1)B_{2k+4}+(2^{2k+1}-1)B_{2k+2}}{2k+1}\left(\frac{|eB|}{m^2}\right)^{2k}. \tag{104}$$

In the case of the electric field directed along the first axis, we obtain

$$\mathcal{L}_{\rm der}^{(3+1){\rm scal}}(E)=\frac{ie^2(\partial_{\|}E)^2}{2(8\pi)^2|eE|}\int_0^{\infty}\frac{d\omega}{\omega}\exp\left(-i\frac{m^2}{|eE|}\omega\right)\left(\frac{d^3}{d\omega^3}+\frac{d}{d\omega}\right)\left(\frac{\omega}{\sinh\omega}\right). \tag{105}$$

Thus, the imaginary part of derivative part of the Lagrangian reads

$$\operatorname{Im}\mathcal{L}_{\rm der}^{(3+1){\rm scal}}(E)=\frac{e^2(\partial_{\|}E)^2}{2^7\pi^4|eE|}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^3}\exp\left(-\frac{\pi m^2 n}{|eE|}\right)$$
$$\times\left[6+6\frac{\pi m^2 n}{|eE|}+\pi^2 n^2\left(3\frac{m^4}{|eE|^2}-1\right)+\frac{m^2\pi^3 n^3}{|eE|}\left(\frac{m^4}{|eE|^2}-1\right)\right], \tag{106}$$

which determines the correction to the probability of particle-antiparticle creation in a constant electric field expressed through

$$\operatorname{Im}\mathcal{L}^{(3+1){\rm scal}}(E)=\frac{(eE)^2}{16\pi^3}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^2}\exp\left(-\frac{\pi m^2}{|eE|}n\right)=-\frac{(eE)^2}{16\pi^3}\operatorname{Li}_2\left[-\exp\left(-\frac{\pi m^2}{|eE|}\right)\right]. \tag{107}$$

As in (2+1)-dimensional case, we observe that the sum in the right-hand side of Eq. (106) is positive only for the large enough values of the mass ($m^2 \gtrsim 0.388|eE|$).

The expression (106) concludes the list of our results describing the influence of slowly varying external electromagnetic fields on the spinor and scalar QED vacuum in the two-derivative approximation.

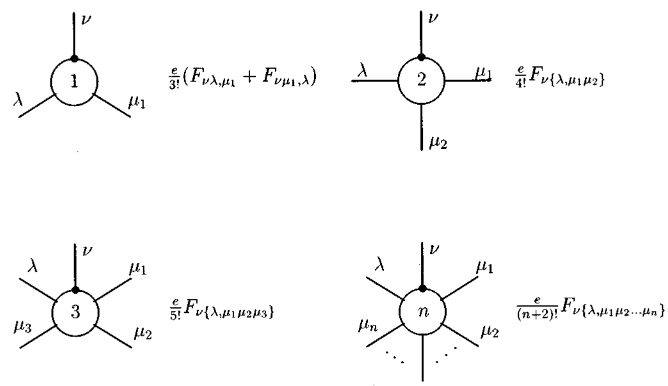

FIG. 1. Diagrammatic notations for the boson interaction vertices. The curly brackets denote symmetrization of the type, $F_{\nu\{\lambda,\mu_1,\ldots,\mu_n\}}=F_{\nu\lambda,\mu_1,\ldots,\mu_n}+F_{\nu\mu_1,\lambda,\ldots,\mu_n}+\cdots+F_{\nu\mu_n,\mu_1,\ldots,\lambda}$.

X. HOW TO GET HIGHER DERIVATIVE TERMS

Obviously, the method of the present paper can be applied for calculating the higher derivative terms (with their total number equal to four or higher) of the low energy effective action in QED. However, the computational work with increasing the total number of derivatives is getting so hard that obtaining already all the four derivative terms seems to be impossible without use of a computer. Just to get feeling how difficult this problem is, let us consider the classification of all the relevant Feynman diagrams in four derivative approximation.

To facilitate the calculation of the perturbative expansion in number of derivatives in the problem at hand, it is appropriate to develop the Feynman diagram technique. Our starting point will be the system of Eqs. (14) and (19). Then, as is seen, the derivative expansion results from all (connected as well as disconnected) vacuum diagrams produced by (14). A somewhat disappointing feature of our Lagrangian is an infinite number of local interactions. Nevertheless, as will become clear in a moment, while working at any finite order of the perturbative theory, one requires only a finite number of those interactions.

We observe that there are two different types of local interactions in (14). The first (bosonic) type contains only the bosonic fields, $x_\mu(t)$. The corresponding vertices are shown in Fig. 1. The other interactions involve both the boson, $x_\mu(t)$, and the spinor fields, $\psi_\mu(t)$. These latter produce the vertices given in Fig. 2. The integers in the vertices denote the number of derivatives (later called the weights of vertices) of the electromagnetic field with respect to space–time. Some legs in the diagrams are marked by circles and bullets. The circles correspond to legs related to the first Lorentz index (ν) of the tensor weight, $F_{\nu\lambda,\mu_1,\ldots,\mu_n}$, assigned to the vertex, while the bullets, on the other hand, mark legs which contain the derivatives with respect to the proper time. The latter act on the (bosonic) propagators attached to the marked legs.

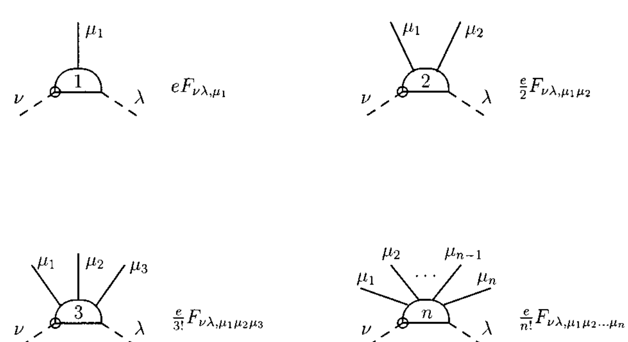

FIG. 2. Diagrammatic notations for the fermion–boson interaction vertices.

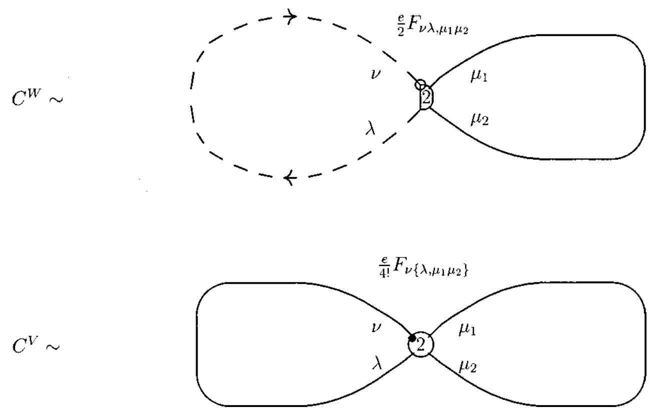

FIG. 3. Two simplest examples of diagrams related to the two-derivative terms C^W and C^V in our general expression for spinor QED.

The Feynman rules for writing expressions corresponding to Feynman diagrams are more or less standard. One has to use the propagators given in (41) and (42) for connecting the bosonic (solid) and the fermion (dashed) legs, respectively. The combinatoric factors can be straightforwardly derived. Two simplest diagrams giving a nonzero contribution to two-derivative terms in the effective action are represented in Fig. 3.

Let us mention the most general rules. To start with, we classify all diagrams leading to terms with a given finite number (called the weight of the corresponding diagrams from now on) of derivatives in the expansion. First of all, we see that diagrams with weight N may contain different number of vertices. We denote by der(N) the set of all diagrams of a given weight N. By marking the bosonic (Fig. 1) and the fermion (Fig. 2) vertices with n derivatives just by $[n]$ and $\overline{[n]}$, respectively, we see that the set der(N) contains a finite number of elements: der(N) $=\{\overline{[N]},[N],\overline{[N-1]}\oplus\overline{[1]},\overline{[N-1]}\oplus[1],[N-1]\oplus\overline{[1]},[N-1]\oplus[1],\ldots\}$. Each of the elements in der(N) produces in its turn a (finite) number of Feynman diagrams differing from one another by all possible connections (by means of propagators) between all legs of the vertices. Thus, the diagrams of weight two in Fig. 3, related to C^W and C^V in the general expression (50), correspond to elements $\overline{[2]}$ and $[2]$ in the set der(2), respectively.

Any element of der(N) specifies the number of different vertices as well as their separate weights. If the number of different vertices in a diagram is given by integers $\{V_1,V_2,\ldots,V_k\}$ then the overall factor in front of the corresponding expression is $1/(V_1!V_2!\ldots V_k!)$. Next, let the total number of bosonic and the fermion vertices be k_B and k_F (so that $k_B+k_F=k$), respectively. Then the total number of bosonic legs of all the vertices in such a diagram is $2k_B+N)$, while the number of the fermion legs is $2k_F$. Since, we are interested in vacuum diagrams (with all legs being connected) only, the diagrams of an odd weight N are not relevant for our derivative expansion. So, we put $N=2n$. As is easy to count, the total number of all possible connections (by means of k_B+n bosonic and k_F the fermion propagators) between these vertices is $(N+2k_B-1)!!(2k_F-1)!!$, where we assume that $(-1)!!\equiv 1$. This is an upper bound for the number of different diagrams with the given vertex set corresponding to the given element $\overline{[V_1]}\oplus\cdots\oplus\overline{[V_{k_F}]}\oplus[V_{k_F+1}]\oplus\cdots\oplus[V_k]\in\text{der}(N)$. However, due to the symmetry of the vertices with respect to permutations of their nonmarked legs as well as with respect to permutations of identical vertices, some of the diagrams are in fact equivalent. For example, the naive number of all relevant diagrams for the two-derivative terms in the expansion of the effective action is 25. On the other hand, as is seen from our general result (50), the actual number of nonequivalent terms is 11.

Now let us say several words about the sign factors of diagrams. First, all diagrams of weight $N=2n$ have an overall factor $(-i)^n$. To get the right sign resulting from the fermion loops, one preliminary has to assign the direction of the fermion flow in the diagram by adding arrows on the fermion (dashed) lines. Then the overall sign factor is obtained by multiplying sign factors for each the fermion loop of the diagram. Each of the loop factors is defined by the formula, $(-1)^{N_0+1}$, where N_0 is the number of arrows running into circles of loop vertices. This rule takes into account the fact that the fermion propagators are antisymmetric with respect to the simultaneous permutation of their Lorentz indices and proper time coordinates as well as the fact that tensor weight at the fermion vertices feels the order of first two indices.

Concluding this section, we would like to express a hope that the brief description of the Feynman technique given here would be enough for writing a code in some of the languages used for analytical computations if such a need appears.

XI. CONCLUSIONS

In conclusion, here we further develop the method of our previous paper[9] and generalize it to quantum electrodynamics in $2+1$ dimensions. The distinctive feature of our approach is the use of a special matrix basis (in Lorentz indices) in order to deal with functions of antisymmetric tensors such as the (background) field strength tensor in QED. In Sec. III, we give the explicit representation for these matrices as well as demonstrate how they facilitate the calculation. It is also the use of these matrices that allowed us to obtain the derivative expansion in the fully covariant form.

Then, in this paper, we derived explicit expression of the two-derivative term in the derivative expansion of the effective action in QED in both fermion and scalar QED in $2+1$ and $3+1$ dimensions. In addition, we also calculated the leading order corrections to the probability of the particle–antiparticle creation rate produced by space–time gradients of the electric field background. The latter gives a nontrivial generalization of the famous Schwinger result in a constant electric field.[3]

Among other results, here we derived the Feynman rules for generating the perturbative expansion of the effective action in the number of derivatives. This means that, in principle, an arbitrary finite order of the derivative expansion is calculable in our approach. For obvious reasons, the complexity of calculation explodes at higher orders and, in the case of the four-derivative approximation, the computational work already becomes so hard that it is almost impossible to get a result in the closed form without using a computer. By making use of the Feynman rules, derived in this paper, one can write a computer code in order to calculate higher order approximations.

At the end, let us also make a few remarks about possible tests and applications of derivative expansion obtained in this paper.

As in the case of the Euler–Heisenberg action, the derivative corrections will affect, among other things, the photon–photon scattering amplitude. For a vanishing background field, the latter is discussed in detail in Ref. 6. Obviously, when the background field is nonzero the corresponding amplitude and the energy dependence of the cross section are going to change. As for the explicit form of the result, it will be given elsewhere.

Besides that, it is likely that the explicit dependence of the photon–photon cross section would be of great interest in studies of some real systems which exist under extremely large magnetic fields. The vicinity of the neutron stars and the early Universe[31] are the most natural candidates of such systems.

The formal derivative expansion might also be useful in other problems, such as the generalization of the theory of magnetic catalysis of chiral symmetry breaking in QED_4 (Ref. 32) and QED_3 (Ref. 33) to the case of inhomogeneous external fields.

ACKNOWLEDGMENTS

We would like to thank Theodore Hall for pointing out mistakes in the expressions for the imaginary part of the effective action in the first version of the paper. V.P.G. is grateful to the members of the Institute for Theoretical Physics, University of Bern, Switzerland, for hospitality during his stay there. The work of V.P.G. was supported by Grant No. INTAS-93-2058-ext ''East–West network in constrained dynamical systems,'' by the Swiss National Science Foundation Grant No. CEEC/NIS/96-98/7 IP 051219 and by the Foundation of Fundamental Researches of Ministry of Sciences of Ukraine under Grant No. 2.5.1/003. The work of I.A.S. was supported by the U.S. Department of Energy Grant No. DE-FG02-84ER40153.

APPENDIX A: COEFFICIENT FUNCTIONS WHICH APPEAR IN THE DERIVATIVE EXPANSION

Here we give the functions used in Eqs. (50) and (83). [Here we corrected the typos which appeared in Ref. 9, namely we (i) omitted an extra term in the expression for C_4^{VV} that was mistakenly present; (ii) replaced the wrong factor $H(\tau\beta)$ in the last term of C_4^{VV} by $H(\tau\gamma)$, and (iii) added the third term in C_5^{VV} which was originally missing. In addition, we rewrote C_5^{VV} in a slightly different form.]

$$C^W(\overline{\alpha},\overline{\beta})=\tau^2\tanh(\alpha\tau)H(\beta\tau), \tag{A1}$$

$$C^V(\overline{\alpha},\overline{\beta})=\alpha\tau^3H(\alpha\tau)H(\beta\tau)-\frac{\alpha\tau}{\beta^2-\alpha^2}[H(\beta\tau)-H(\alpha\tau)], \tag{A2}$$

$$C_1^{WW}(\overline{\alpha},\overline{\beta},\overline{\gamma})=\frac{\tau^3}{8}\tanh(\alpha\tau)\tanh(\beta\tau)H(\gamma\tau), \tag{A3}$$

$$C_2^{WW}(\overline{\alpha},\overline{\beta},\overline{\gamma})=\frac{\tau^2}{4}[\tanh(\alpha\tau)+\tanh(\beta\tau)]\left(\frac{H(\alpha\tau+\beta\tau)-H(\gamma\tau)}{\alpha+\beta-\gamma}-\frac{H(\gamma\tau)}{\alpha+\beta}\right), \tag{A4}$$

$$C_1^{VW}(\overline{\alpha},\overline{\beta},\gamma)=-\frac{\tau^3}{4}\tanh(\alpha\tau)\left(\beta\tau H(\beta\tau)H(\gamma\tau)-\frac{H(\beta\tau)-H(\gamma\tau)}{\tau(\beta+\gamma)}\right), \tag{A5}$$

$$C_2^{VW}(\overline{\alpha},\overline{\beta},\overline{\gamma})=\frac{\tau^2\beta\tanh(\alpha\tau)}{2(\beta^2-\gamma^2)}[H(\beta\tau)-H(\gamma\tau)], \tag{A6}$$

$$\begin{aligned}C_1^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})=&\frac{\tau^5\alpha\beta}{2}H(\alpha\tau)H(\beta\tau)H(\gamma\tau)-\frac{\tau^3\alpha H(\alpha\tau)}{2(\beta-\gamma)}(H(\beta\tau)-H(\gamma\tau))\\&-\frac{\tau^3\beta H(\beta\tau)}{2(\alpha+\gamma)}(H(\alpha\tau)-H(\gamma\tau))-\frac{\tau}{2}\frac{H(\alpha\tau)}{(\alpha+\gamma)(\alpha+\beta)}\\&-\frac{\tau}{2}\frac{H(\beta\tau)}{(\alpha+\beta)(\beta-\gamma)}+\frac{\tau}{2}\frac{H(\gamma\tau)}{(\alpha+\gamma)(\beta-\gamma)},\end{aligned} \tag{A7}$$

$$\begin{aligned}C_2^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})=&-\frac{\tau^3\alpha\beta H(\alpha\tau)H(\beta\tau)}{2(\alpha-\beta)(\alpha-\beta+\gamma)}+\frac{\tau^3[2(\alpha-\beta)+\gamma]H(\gamma\tau)[\beta H(\beta\tau)-\alpha H(\alpha\tau)]}{2(\alpha-\beta)(\alpha-\beta+\gamma)}\\&+\frac{\alpha\tau}{2}H(\alpha\tau)\left(\frac{2(\beta+\gamma)}{(\alpha^2-\beta^2)(\alpha^2-\gamma^2)}-\frac{2\alpha-\beta+\gamma}{(\alpha-\beta)^2(\alpha+\gamma)(\alpha-\beta+\gamma)}\right)\\&+\frac{\beta\tau}{2}H(\beta\tau)\left(\frac{2(\gamma-\alpha)}{(\alpha^2-\beta^2)(\beta^2-\gamma^2)}+\frac{2\beta-\alpha-\gamma}{(\alpha-\beta)^2(\beta-\gamma)(\alpha-\beta+\gamma)}\right)\\&+\frac{\tau}{2}H(\gamma\tau)\left(\frac{2(\alpha\beta+\gamma^2)}{(\alpha^2-\gamma^2)(\beta^2-\gamma^2)}-\frac{\gamma}{(\alpha+\gamma)(\beta-\gamma)(\alpha-\beta+\gamma)}\right)\\&+\frac{\tau}{2(\alpha-\beta)(\alpha-\beta+\gamma)},\end{aligned} \tag{A8}$$

$$C_3^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})=-\frac{\tau^3\alpha\beta H(\alpha\tau)}{\beta^2-\gamma^2}[H(\beta\tau)-H(\gamma\tau)]+\frac{\alpha\tau H(\alpha\tau)}{(\alpha-\beta)(\alpha^2-\gamma^2)}-\frac{\tau}{\beta^2-\gamma^2}\left(\frac{\beta}{\alpha-\beta}H(\beta\tau)-\frac{\alpha\beta+\gamma^2}{\alpha^2-\gamma^2}H(\gamma\tau)\right), \tag{A9}$$

$$C_4^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})=-\frac{2\tau\alpha^2H(\alpha\tau)}{(\alpha^2-\beta^2)(\alpha^2-\gamma^2)}+\frac{2\tau\beta^2H(\beta\tau)}{(\alpha^2-\beta^2)(\beta^2-\gamma^2)}+\frac{2\tau\gamma^2H(\gamma\tau)}{(\alpha^2-\gamma^2)(\gamma^2-\beta^2)}, \tag{A10}$$

$$\begin{aligned}C_5^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})=&\frac{\tau^3\alpha}{2}\left(\frac{\alpha}{(\alpha+\beta)(\alpha+\beta+\gamma)}+\frac{1}{\alpha+\gamma}\right)H(\alpha\tau)H(\beta\tau)\\&+\frac{\tau^3\alpha}{2}\left(\frac{\alpha}{(\alpha+\gamma)(\alpha+\beta+\gamma)}+\frac{1}{\alpha+\beta}\right)H(\alpha\tau)H(\gamma\tau)+\tau^3H(\beta\tau)H(\gamma\tau)\\&+\frac{\tau^3}{2}\left(\frac{\beta\gamma(\beta+\gamma)}{\alpha+\beta+\gamma}-2\alpha^2\right)\frac{H(\beta\tau)H(\gamma\tau)}{(\alpha+\beta)(\alpha+\gamma)}+\frac{\alpha\tau H(\alpha\tau)\left(2+\frac{\alpha+\beta}{\alpha+\gamma}+\frac{\alpha+\gamma}{\alpha+\beta}\right)}{2(\alpha+\beta)(\alpha+\gamma)(\alpha+\beta+\gamma)}\\&+\frac{\tau}{2}\left(\frac{2H(\gamma\tau)}{(\alpha+\gamma)(\beta-\gamma)}-\frac{2H(\beta\tau)}{(\alpha+\beta)(\beta-\gamma)}+\frac{\gamma H(\gamma\tau)}{(\alpha+\gamma)^2(\alpha+\beta+\gamma)}\right.\\&\left.+\frac{\beta H(\beta\tau)}{(\alpha+\beta)^2(\alpha+\beta+\gamma)}-\frac{1}{(\alpha+\beta)(\alpha+\beta+\gamma)}-\frac{1}{(\alpha+\gamma)(\alpha+\beta+\gamma)}\right).\end{aligned} \tag{A11}$$

Here we used the following notation:

$$H(x)=\frac{x\coth x-1}{x^2}, \tag{A12}$$

and the letters with bars differ from the letters without those only in a factor of the electric charge, $\alpha=e\overline{\alpha}$. Note that in Ref. 9 we ignored this difference.

As $\tau\rightarrow 0$, these coefficient functions have the following asymptotic behavior:

$$C^W(\overline{\alpha},\overline{\beta})\simeq\frac{\alpha\tau^3}{3}-\frac{\alpha\tau^5}{45}(5\alpha^2+\beta^2)+O(\tau^7), \tag{A13}$$

$$C^V(\overline{\alpha},\overline{\beta})\simeq\frac{2\alpha\tau^3}{15}-\frac{\alpha\tau^5}{105}(\alpha^2+\beta^2)+O(\tau^7), \tag{A14}$$

$$C_1^{WW}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq\frac{\alpha\beta\tau^5}{24}+O(\tau^7), \tag{A15}$$

$$C_2^{WW}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq-\frac{\tau^3}{12}+\frac{\tau^5}{180}(4\alpha^2+4\beta^2+\gamma^2-7\alpha\beta-\alpha\gamma-\beta\gamma)+O(\tau^7), \tag{A16}$$

$$C_1^{VW}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq\frac{\alpha\tau^5}{180}(\gamma-6\beta)+O(\tau^7), \tag{A17}$$

$$C_2^{VW}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq\frac{\alpha\beta\tau^5}{90}+O(\tau^7), \tag{A18}$$

$$C_1^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq\frac{\tau^3}{90}-\frac{\tau^5}{1890}(2\alpha^2+2\beta^2+2\gamma^2-51\alpha\beta-9\alpha\gamma+9\beta\gamma)+O(\tau^7), \tag{A19}$$

$$C_2^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq-\frac{4\tau^3}{45}+\frac{\tau^5}{1890}(10\alpha^2+10\beta^2+13\gamma^2+15\alpha\beta+3\alpha\gamma-3\beta\gamma)+O(\tau^7), \tag{A20}$$

$$C_3^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq-\frac{\tau^3}{45}+\frac{\tau^5}{945}(2\alpha^2+2\beta^2+2\gamma^2+9\alpha\beta)+O(\tau^7), \tag{A21}$$

$$C_4^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq\frac{2\tau^3}{45}-\frac{4\tau^5}{945}(\alpha^2+\beta^2+\gamma^2)+O(\tau^7), \tag{A22}$$

$$C_5^{VV}(\overline{\alpha},\overline{\beta},\overline{\gamma})\simeq\frac{8\tau^3}{45}-\frac{\tau^5}{1890}(20\alpha^2+23\beta^2+23\gamma^2-12\alpha\beta-12\alpha\gamma+6\beta\gamma)+O(\tau^7). \tag{A23}$$

APPENDIX B: EXPANSION OF THE DERIVATIVE TERMS IN POWERS OF THE PROPER TIME

In this Appendix we give the proper time expansion of the derivative terms, as in Eqs. (56) and (88), up to the order τ^5.

In case of spinor QED, from Eq. (50) we derive the expansion

$$\begin{aligned}tr\langle x|U(\tau)|x\rangle\simeq tr\langle x|U(\tau)|x\rangle_0\Bigg[&1+\frac{ie^2\tau^3}{20}F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu}\\&+\frac{ie^2\tau^3}{180}\left(\frac{7}{2}F^{\nu\lambda,\mu}F_{\nu\lambda,\mu}-F^{\nu\lambda,}{}_{\lambda}F_{\nu\mu,}{}^{\mu}\right)-\frac{ie^4\tau^5}{315}F_{\nu\lambda,\mu\kappa}(16\mathcal{F}\eta^{\mu\kappa}F^{\nu\lambda}+F^{\nu\lambda}(F^2)^{\mu\kappa})\\&+\frac{ie^4\tau^5}{1890}(2F_{\nu\lambda,\mu}F^{\nu\rho,}{}_{\rho}(F^2)^{\lambda\mu}-2F_{\nu\lambda,\mu}F^{\nu\lambda,\rho}(F^2)^{\mu}{}_{\rho}-37F_{\nu\lambda,\mu}F^{\nu\sigma,\mu}(F^2)^{\lambda}{}_{\sigma}\\&+F_{\nu\mu,}{}^{\mu}F_{\sigma\rho,}{}^{\rho}(F^2)^{\nu\sigma})-\frac{ie^4\tau^5}{2520}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}(38\eta^{\mu\rho}F^{\lambda\sigma}F^{\nu\kappa}-12\eta^{\nu\kappa}F^{\lambda\rho}F^{\mu\sigma}\\&+47\eta^{\mu\rho}F^{\nu\lambda}F^{\sigma\kappa}+16\eta^{\kappa\rho}F^{\lambda\sigma}F^{\nu\mu})\Bigg].\end{aligned} \tag{B1}$$

Notice that despite the difference between the two sets of matrices $A_{(j)}^{\mu\nu}$ in 2+1 and 3+1 dimensions, the expression in square brackets is independent of the dimension up to this order in the expansion. In calculation, we took into account the Bianchi identity to show that many seemingly different terms appearing in the expansion reduce to the same structures. In particular, the following relations are the identities that we needed:

$$F_{\nu\lambda,\mu\kappa}\eta^{\lambda\kappa}F^{\nu\mu}=\tfrac{1}{2}F_{\nu\lambda,\mu\kappa}\eta^{\mu\kappa}F^{\nu\lambda}, \tag{B2}$$

$$F_{\nu\lambda,\mu\kappa}F^{\nu\mu}(F^2)^{\lambda\kappa}=\tfrac{1}{2}F_{\nu\lambda,\mu\kappa}F^{\nu\lambda}(F^2)^{\mu\kappa}, \tag{B3}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\nu\rho}F^{\lambda\sigma}F^{\mu\kappa}=\tfrac{1}{2}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\rho}F^{\lambda\sigma}F^{\nu\kappa}, \tag{B4}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\nu\kappa}F^{\lambda\sigma}F^{\mu\rho}=F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\nu\kappa}F^{\lambda\rho}F^{\mu\sigma}+\tfrac{1}{2}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\rho}F^{\lambda\sigma}F^{\nu\kappa}, \tag{B5}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\sigma}F^{\kappa\rho}F^{\nu\lambda}=-\tfrac{1}{2}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\rho}F^{\nu\lambda}F^{\sigma\kappa}, \tag{B6}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\sigma\rho}F^{\mu\kappa}F^{\nu\lambda}=-2F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\kappa\rho}F^{\lambda\sigma}F^{\nu\mu}, \tag{B7}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\lambda\sigma}F^{\kappa\rho}F^{\nu\mu}=-\tfrac{1}{4}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\rho}F^{\nu\lambda}F^{\sigma\kappa}, \tag{B8}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\nu\kappa}\eta^{\sigma\rho}(F^2)^{\lambda\mu}=-F_{\nu\lambda,\mu}F_{\rho}{}^{\nu\rho,}(F^2)^{\lambda\mu}, \tag{B9}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\sigma}\eta^{\nu\kappa}(F^2)^{\lambda\rho}=-\tfrac{1}{2}F_{\nu\lambda,\mu}F^{\nu\lambda,\rho}(F^2)_{\rho}{}^{\mu}, \tag{B10}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\rho}\eta^{\nu\kappa}(F^2)^{\lambda\sigma}=-F_{\nu\lambda,\mu}F^{\nu\sigma,\mu}(F^2)_{\sigma}{}^{\lambda}, \tag{B11}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\mu\kappa}\eta^{\nu\rho}(F^2)^{\lambda\sigma}=-F_{\nu\lambda,\mu}F^{\nu\sigma,\mu}(F^2)^{\lambda}{}_{\sigma}+\tfrac{1}{2}F_{\nu\lambda,\mu}F^{\nu\lambda,\rho}(F^2)^{\mu}{}_{\rho}, \tag{B12}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\nu\sigma}\eta^{\lambda\kappa}(F^2)^{\mu\rho}=F_{\nu\lambda,\mu}F^{\nu\lambda,\rho}(F^2)^{\mu}{}_{\rho}, \tag{B13}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\eta^{\nu\mu}\eta^{\kappa\rho}(F^2)^{\lambda\sigma}=-F_{\nu\mu,}{}^{\mu}F_{\sigma\rho,}{}^{\rho}(F^2)^{\nu\sigma}. \tag{B14}$$

After expanding $tr\langle x|U(\tau)|x\rangle_0$ in Eq. (B1) in powers of τ up to the terms of order τ^3 and substituting the obtained expression in the definition of the effective action, we arrive at the following two-derivative correction of the order $1/m^6$:

$$\begin{aligned}\mathcal{L}_{1/m^6}^{(3+1)\mathrm{spin}}=&-\frac{11\alpha^2}{630m^6}F^{\beta\gamma}F_{\beta\gamma}F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu}+\frac{4\alpha^2}{315m^6}(F^2)^{\mu\kappa}F^{\nu\lambda}F_{\nu\lambda,\mu\kappa}\\&+\frac{\alpha^2}{270m^6}F^{\beta\gamma}F_{\beta\gamma}\left(\frac{7}{2}F^{\nu\lambda,\mu}F_{\nu\lambda,\mu}-F^{\nu\lambda}{}_{\lambda}F_{\nu\mu,}{}^{\mu}\right)\\&-\frac{2\alpha^2}{945m^6}(2F_{\nu\lambda,\mu}F^{\nu\rho,}{}_{\rho}(F^2)^{\lambda\mu}-2F_{\nu\lambda,\mu}F^{\nu\lambda,\rho}(F^2)^{\mu}{}_{\rho}-37F_{\nu\lambda,\mu}F^{\nu\sigma,\mu}(F^2)^{\lambda}{}_{\sigma}\\&+F_{\nu\mu,}{}^{\mu}F_{\sigma\rho,}{}^{\rho}(F^2)^{\nu\sigma})+\frac{\alpha^2}{630m^6}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}(38\eta^{\mu\rho}F^{\lambda\sigma}F^{\nu\kappa}-12\eta^{\nu\kappa}F^{\lambda\rho}F^{\mu\sigma}\\&+47\eta^{\mu\rho}F^{\nu\lambda}F^{\sigma\kappa}+16\eta^{\kappa\rho}F^{\lambda\sigma}F^{\nu\mu}),\end{aligned} \tag{B15}$$

to the one-loop effective action in spinor QED in 3+1 dimensions, and the correction of the order $1/m^7$,

$$\begin{aligned}\mathcal{L}_{1/m^7}^{(2+1)\mathrm{spin}}=&-\frac{11\alpha^2\pi}{336m^7}F^{\beta\gamma}F_{\beta\gamma}F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu}+\frac{\alpha^2\pi}{42m^7}(F^2)^{\mu\kappa}F^{\nu\lambda}F_{\nu\lambda,\mu\kappa}\\&+\frac{\alpha^2\pi}{144m^7}F^{\beta\gamma}F_{\beta\gamma}\left(\frac{7}{2}F^{\nu\lambda,\mu}F_{\nu\lambda,\mu}-F^{\lambda}{}_{\nu\lambda,}F_{\nu\mu,}{}^{\mu}\right)\\&-\frac{\alpha^2\pi}{252m^7}(2F_{\nu\lambda,\mu}F^{\nu\rho,}{}_{\rho}(F^2)^{\lambda\mu}-2F_{\nu\lambda,\mu}F^{\nu\lambda,\rho}(F^2)^{\mu}{}_{\rho}-37F_{\nu\lambda,\mu}F^{\nu\sigma,\mu}(F^2)^{\lambda}{}_{\sigma}\\&+F_{\nu\mu,}{}^{\mu}F_{\sigma\rho,}{}^{\rho}(F^2)^{\nu\sigma})+\frac{\alpha^2\pi}{336m^7}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}(38\eta^{\mu\rho}F^{\lambda\sigma}F^{\nu\kappa}-12\eta^{\nu\kappa}F^{\lambda\rho}F^{\mu\sigma}\\&+47\eta^{\mu\rho}F^{\nu\lambda}F^{\sigma\kappa}+16\eta^{\kappa\rho}F^{\lambda\sigma}F^{\nu\mu}),\end{aligned} \tag{B16}$$

to the effective action in 2+1 dimensions. It turns out that these latter can be further simplified. Indeed, after integrating by parts, the results can be expressed through the following seven Lorentz scalars,

$$L_1=F^{\beta\gamma}F_{\beta\gamma}F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu}, \tag{B17}$$

$$L_2 = F^{\beta\gamma} F_{\beta\gamma} F^{\nu\lambda,\mu} F_{\nu\lambda,\mu}, \tag{B18}$$

$$L_3 = F^{\beta\gamma} F_{\beta\gamma} F^{\nu\lambda,}{}_{\lambda} F_{\nu\mu}{}^{\mu}, \tag{B19}$$

$$L_4 = F_{\nu\lambda,\mu\kappa} F^{\nu\lambda} (F^2)^{\mu\kappa}, \tag{B20}$$

$$L_5 = F^{\kappa\nu} F_{\nu\lambda,\mu} F^{\lambda\sigma} F_{\sigma\kappa}{}^{\mu}, \tag{B21}$$

$$L_6 = F_{\nu\lambda,\mu}{}^{\mu} (F^3)^{\nu\lambda} \tag{B22}$$

$$L_7 = F_{\nu\mu,}{}^{\mu} F_{\lambda\rho,}{}^{\rho} (F^2)^{\nu\lambda}. \tag{B23}$$

Thus, the final results in 3+1 and in 2+1 dimensions read

$$\mathcal{L}^{(3+1)\mathrm{spin}}_{1/m^6} = -\frac{16\alpha^2}{315 m^6} L_1 - \frac{8\alpha^2}{315 m^6} L_2 + \frac{2\alpha^2}{315 m^6} L_3 - \frac{\alpha^2}{945 m^6} L_4 - \frac{11\alpha^2}{945 m^6} L_5 - \frac{26\alpha^2}{945 m^6} L_6 + \frac{4\alpha^2}{189 m^6} L_7, \tag{B24}$$

$$\mathcal{L}^{(2+1)\mathrm{spin}}_{1/m^7} = -\frac{2\alpha^2\pi}{21 m^6} L_1 - \frac{\alpha^2\pi}{21 m^6} L_2 + \frac{\alpha^2\pi}{84 m^6} L_3 - \frac{\alpha^2\pi}{504 m^6} L_4 - \frac{11\alpha^2\pi}{504 m^6} L_5 - \frac{13\alpha^2\pi}{252 m^6} L_6 + \frac{5\alpha^2\pi}{126 m^6} L_7, \tag{B25}$$

respectively. This should be compared with the result of Ref. 6 (see Eq. (14) there). Notice that the photon field in Ref. 6 desribes on-shell quanta, and, as a result, the terms containing L_1, L_3, L_6, and L_7 do not appear (they are proportional to $k^2=0$).

In a similar way, in the case of scalar QED we obtain the following expression for the expansion of Eq. (83):

$$\begin{aligned}\langle x|U_{\mathrm{bos}}(\tau)|x\rangle \simeq \langle x|U_{\mathrm{bos}}(\tau)|x\rangle_0 \Big[1 &- \frac{ie^2\tau^3}{30} F^{\nu\lambda} F_{\nu\lambda,\mu}{}^{\mu} - \frac{ie^2\tau^3}{180} (4F^{\nu\lambda,\mu} F_{\nu\lambda,\mu} + F^{\nu\lambda,}{}_{\lambda} F_{\nu\mu,}{}^{\mu}) \\ &+ \frac{ie^4\tau^5}{840} F_{\nu\lambda,\mu\kappa} (4\mathcal{F}\eta^{\mu\kappa} F^{\nu\lambda} + 2F^{\nu\lambda}(F^2)^{\mu\kappa}) + \frac{ie^4\tau^5}{7560} (8F_{\nu\lambda,\mu} F^{\nu\rho,}{}_{\rho} (F^2)^{\lambda\mu} \\ &+ 13F_{\nu\lambda,\mu} F^{\nu\lambda,\rho} (F^2)^{\mu}{}_{\rho} + 20F_{\nu\lambda,\mu} F^{\nu\sigma,\mu} (F^2)^{\lambda}{}_{\sigma} + 4F_{\nu\mu,}{}^{\mu} F_{\sigma\rho,}{}^{\rho} (F^2)^{\nu\sigma}) \\ &+ \frac{ie^4\tau^5}{5040} F_{\nu\lambda,\mu} F_{\sigma\kappa,\rho} (8\eta^{\mu\rho} F^{\lambda\sigma} F^{\nu\kappa} - 4\eta^{\nu\kappa} F^{\lambda\rho} F^{\mu\sigma} - 17\eta^{\mu\rho} F^{\nu\lambda} F^{\sigma\kappa} \\ &+ 24\eta^{\kappa\rho} F^{\lambda\sigma} F^{\nu\mu}) \Big],\end{aligned} \tag{B26}$$

leading to the $1/m^6$ correction to the effective Lagrangian density,

$$\begin{aligned}\mathcal{L}^{(3+1)\mathrm{scal}}_{1/m^6} = &-\frac{\alpha^2}{126 m^6} F^{\beta\gamma} F_{\beta\gamma} F^{\nu\lambda} F_{\nu\lambda,\mu}{}^{\mu} + \frac{\alpha^2}{210 m^6} (F^2)^{\mu\kappa} F^{\nu\lambda} F_{\nu\lambda,\mu\kappa} \\ &- \frac{\alpha^2}{1080 m^6} F^{\beta\gamma} F_{\beta\gamma} \left(4F^{\nu\lambda,\mu} F_{\nu\lambda,\mu} + F^{\nu\lambda,}{}_{\lambda} F_{\nu\mu,}{}^{\mu} \right. \\ &+ \frac{\alpha^2}{3780 m^6} \left(8F_{\nu\lambda,\mu} F^{\nu\rho,}{}_{\rho} (F^2)^{\lambda\mu} + 13F_{\nu\lambda,\mu} F^{\nu\lambda,\rho} (F^2)^{\mu}{}_{\rho} + 20F_{\nu\lambda,\mu} F^{\nu\sigma,\mu} (F^2)^{\lambda}{}_{\sigma} \right. \\ &+ 4F_{\nu\mu,}{}^{\mu} F_{\sigma\rho,}{}^{\rho} (F^2)^{\nu\sigma} + \frac{\alpha^2}{2520 m^6} F_{\nu\lambda,\mu} F_{\sigma\kappa,\rho} (8\eta^{\mu\rho} F^{\lambda\sigma} F^{\nu\kappa} - 4\eta^{\nu\kappa} F^{\lambda\rho} F^{\mu\sigma} \\ &- 17\eta^{\mu\rho} F^{\nu\lambda} F^{\sigma\kappa} + 24\eta^{\kappa\rho} F^{\lambda\sigma} F^{\nu\mu}),\end{aligned} \tag{B27}$$

in 3+1 dimensions, and the $1/m^7$ correction,

$$\mathcal{L}^{(2+1)\mathrm{scal}}_{1/m^7}=\frac{5\alpha^2\pi}{336m^7}F^{\beta\gamma}F_{\beta\gamma}F^{\nu\lambda}F_{\nu\lambda,\mu}{}^{\mu}-\frac{\alpha^2\pi}{112m^7}(F^2)^{\mu\kappa}F^{\nu\lambda}F_{\nu\lambda,\mu\kappa}$$
$$+\frac{\alpha^2\pi}{576m^7}F^{\beta\gamma}F_{\beta\gamma}\left(4F^{\nu\lambda,\mu}F_{\nu\lambda,\mu}+F^{\nu\lambda,}{}_{\lambda}F_{\nu\mu,}{}^{\mu}-\frac{\alpha^2\pi}{2016m^7}\left(8F_{\nu\lambda,\mu}F^{\nu\rho,}{}_{\rho}(F^2)^{\lambda\mu}\right.\right.$$
$$+13F_{\nu\lambda,\mu}F^{\nu\lambda,\rho}(F^2)^{\mu}{}_{\rho}+20F_{\nu\lambda,\mu}F^{\nu\sigma,\mu}(F^2)^{\lambda}{}_{\sigma}+4F_{\nu\mu,}{}^{\mu}F_{\sigma\rho,}{}^{\rho}(F^2)^{\nu\sigma}$$
$$-\frac{\alpha^2\pi}{1344m^7}F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}(8\eta^{\mu\rho}F^{\lambda\sigma}F^{\nu\kappa}-4\eta^{\nu\kappa}F^{\lambda\rho}F^{\mu\sigma}-17\eta^{\mu\rho}F^{\nu\lambda}F^{\sigma\kappa}$$
$$+24\eta^{\kappa\rho}F^{\lambda\sigma}F^{\nu\mu}), \tag{B28}$$

in 2+1 dimensions.

Up to a divergence, the derived corrections to the effective action are equivalent to

$$\mathcal{L}^{(3+1)\mathrm{scal}}_{1/m^6}=-\frac{13\alpha^2}{2520m^6}L_1-\frac{\alpha^2}{840m^6}L_2-\frac{\alpha^2}{2520m^6}L_3+\frac{\alpha^2}{1890m^6}L_4$$
$$+\frac{\alpha^2}{3780m^6}L_5-\frac{11\alpha^2}{3780m^6}L_6+\frac{\alpha^2}{1890m^6}L_7, \tag{B29}$$

$$\mathcal{L}^{(2+1)\mathrm{scal}}_{1/m^7}=\frac{13\alpha^2}{1344m^6}L_1+\frac{\alpha^2}{448m^6}L_2+\frac{\alpha^2}{1344m^6}L_3-\frac{\alpha^2}{1008m^6}L_4$$
$$-\frac{\alpha^2}{2016m^6}L_5+\frac{11\alpha^2}{2016m^6}L_6-\frac{\alpha^2}{1008m^6}L_7, \tag{B30}$$

in 3+1 and 2+1 dimensions, respectively. Here we used the same seven scalars as in the case of spinor QED above.

APPENDIX C: COEFFICIENT FUNCTIONS WHICH APPEAR IN PURELY ELECTRIC AND PURELY MAGNETIC CASES

In this Appendix we list the formulas, similar to that in Eq. (60), which appear in the course of reduction the general expression for the derivative contribution to the case of a pure magnetic (electric) field background. These are

$$F_{\nu\lambda,\mu\kappa}\sum_{j,l}C^W(f_j,f_l)A^{\lambda\nu}_{(j)}A^{\mu\kappa}_{(l)}=-2iC^W(\overline{\alpha},\overline{\alpha})\sum_{i=1}^{2}\partial_i\partial_iB, \tag{C1}$$

$$F_{\nu\lambda,\mu\kappa}\sum_{j,l}C^V(f_j,f_l)(A^{\nu\lambda}_{(j)}A^{\mu\kappa}_{(l)}+2A^{\nu\mu}_{(j)}A^{\lambda\kappa}_{(l)})=4iC^V(\overline{\alpha},\overline{\alpha})\sum_{i=1}^{2}\partial_i\partial_iB, \tag{C2}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\sum_{j,l,k}C_1^{WW}(f_j,f_l,f_k)A^{\kappa\sigma}_{(j)}A^{\lambda\nu}_{(l)}A^{\mu\rho}_{(k)}=4C_1^{WW}(\overline{\alpha},\overline{\alpha},\overline{\alpha})\sum_{i=1}^{2}(\partial_iB)^2, \tag{C3}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\sum_{j,l,k} C_2^{WW}(f_j,f_l,f_k)A^{\kappa\lambda}_{(j)}A^{\sigma\nu}_{(l)}A^{\mu\rho}_{(k)}=-2C_2^{WW}(\bar{\alpha},-\bar{\alpha},\bar{\alpha})\sum_{i=1}^{2}(\partial_i B)^2, \tag{C4}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\sum_{j,l,k} C_1^{VW}(f_j,f_l,f_k)A^{\sigma\kappa}_{(j)}(A^{\nu\lambda}_{(l)}A^{\mu\rho}_{(k)}+A^{\nu\mu}_{(l)}A^{\lambda\rho}_{(k)})$$
$$=2(C_1^{VW}(\bar{\alpha},\bar{\alpha},\bar{\alpha})+2C_1^{VW}(\bar{\alpha},\bar{\alpha},-\bar{\alpha}))\sum_{i=1}^{2}(\partial_i B)^2, \tag{C5}$$

$$F_{\nu\lambda,\mu}F_{\sigma k,\rho}\sum_{j,l,k} C_2^{VW}(f_j,f_l,f_k)A^{\sigma\kappa}_{(j)}A^{\nu\rho}_{(l)}A^{\lambda\mu}_{(k)}=2C_2^{VW}(\bar{\alpha},\bar{\alpha},\bar{\alpha})\sum_{i=1}^{2}(\partial_i B)^2, \tag{C6}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\sum_{j,l,k} C_1^{VV}(f_j,f_l,f_k)(A^{\nu\lambda}_{(j)}A^{\kappa\sigma}_{(l)}A^{\mu\rho}_{(k)}+A^{\nu\mu}_{(j)}A^{\kappa\rho}_{(l)}A^{\lambda\sigma}_{(k)}+2A^{\nu\lambda}_{(j)}A^{\kappa\rho}_{(l)}A^{\mu\sigma}_{(k)})$$
$$=(4C_1^{VV}(\bar{\alpha},\bar{\alpha},\bar{\alpha})-4C_1^{VV}(-\bar{\alpha},\bar{\alpha},\bar{\alpha})-C_1^{VV}(\bar{\alpha},-\bar{\alpha},\bar{\alpha}))\sum_{i=1}^{2}(\partial_i B)^2, \tag{C7}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\sum_{j,l,k} C_2^{VV}(f_j,f_l,f_k)(A^{\nu\sigma}_{(j)}A^{\kappa\lambda}_{(l)}A^{\mu\rho}_{(k)}+A^{\nu\rho}_{(j)}A^{\kappa\mu}_{(l)}A^{\lambda\sigma}_{(k)}+2A^{\nu\sigma}_{(j)}A^{\kappa\mu}_{(l)}A^{\lambda\rho}_{(k)})$$
$$=-(4C_2^{VV}(\bar{\alpha},\bar{\alpha},\bar{\alpha})-C_2^{VV}(-\bar{\alpha},\bar{\alpha},\bar{\alpha}))\sum_{i=1}^{2}(\partial_i B)^2, \tag{C8}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\sum_{j,l,k} C_3^{VV}(f_j,f_l,f_k)(A^{\nu\lambda}_{(j)}A^{\kappa\mu}_{(l)}A^{\sigma\rho}_{(k)}+A^{\kappa\rho}_{(j)}A^{\nu\sigma}_{(l)}A^{\lambda\mu}_{(k)})$$
$$=-(4C_3^{VV}(\bar{\alpha},\bar{\alpha},\bar{\alpha})-C_3^{VV}(\bar{\alpha},-\bar{\alpha},\bar{\alpha}))\sum_{i=1}^{2}(\partial_i B)^2, \tag{C9}$$

$$C_4^{VV}(f_j,f_l,f_k)A^{\nu\kappa}_{(j)}A^{\lambda\mu}_{(l)}A^{\sigma\rho}_{(k)}=C_4^{VV}(\bar{\alpha},\bar{\alpha},\bar{\alpha})\sum_{i=1}^{2}(\partial_i B)^2, \tag{C10}$$

$$F_{\nu\lambda,\mu}F_{\sigma\kappa,\rho}\sum_{j,l,k} C_5^{VV}(f_j,f_l,f_k)A^{\nu\kappa}_{(j)}(A^{\lambda\sigma}_{(l)}A^{\mu\rho}_{(k)}+A^{\lambda\rho}_{(l)}A^{\mu\sigma}_{(k)})$$
$$=(C_5^{VV}(\bar{\alpha},-\bar{\alpha},\bar{\alpha})+2C_5^{VV}(-\bar{\alpha},\bar{\alpha},\bar{\alpha}))\sum_{i=1}^{2}(\partial_i B)^2, \tag{C11}$$

where $\bar{\alpha}=\sqrt{2\mathcal{F}}$.

The latter expressions contain the coefficient functions from Appendix A calculated for a particular value of their arguments. The convenient representation for them reads

$$C^V(\bar{\alpha},\bar{\alpha})=\frac{\tau^2}{2\omega}(3\omega^2H^2+3H-1), \tag{C12}$$

$$C^W(\bar{\alpha},\bar{\alpha})=\tau^2\tanh(\omega)H, \tag{C13}$$

$$C_1^{WW}(\bar{\alpha},\bar{\alpha},\bar{\alpha})=\frac{\tau^3}{8}\tanh^2(\omega)H, \tag{C14}$$

$$C_2^{WW}(\overline{\alpha},-\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{4}(1-\tanh^2(\omega))H, \tag{C15}$$

$$C_1^{VW}(\overline{\alpha},\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{4}\tanh(\omega)\omega H^2, \tag{C16}$$

$$C_1^{VW}(\overline{\alpha},\overline{\alpha},-\overline{\alpha}) = -\frac{\tau^3}{4\omega}\tanh(\omega)(2\omega^2H^2+3H-1), \tag{C17}$$

$$C_2^{VW}(\overline{\alpha},\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{4\omega}\tanh(\omega)(\omega^2H^2+3H-1), \tag{C18}$$

$$C_1^{VV}(\overline{\alpha},\overline{\alpha},\overline{\alpha}) = \frac{\tau^3}{4\omega^2}(4\omega^4H^3+7\omega^2H^2-2\omega^2H+3H-1), \tag{C19}$$

$$C_1^{VV}(-\overline{\alpha},\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{2\omega^2}(4\omega^4H^3+10\omega^2H^2-3\omega^2H+6H-2), \tag{C20}$$

$$C_1^{VV}(\overline{\alpha},-\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{4\omega^2}(2\omega^4H^3-\omega^2H^2-3H+1), \tag{C21}$$

$$C_2^{VV}(\overline{\alpha},\overline{\alpha},\overline{\alpha}) = \frac{\tau^3}{2\omega^2}(2\omega^4H^3+5\omega^2H^2-2\omega^2H+3H-1), \tag{C22}$$

$$C_2^{VV}(-\overline{\alpha},\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{4\omega^2}(2\omega^4H^3+11\omega^2H^2-2\omega^2H+9H-3), \tag{C23}$$

$$C_3^{VV}(\overline{\alpha},\overline{\alpha},\overline{\alpha}) = \frac{\tau^3}{4\omega^2}(4\omega^4H^3+13\omega^2H^2-4\omega^2H+9H-3), \tag{C24}$$

$$C_3^{VV}(\overline{\alpha},-\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{2\omega^2}(\omega^4H^3+4\omega^2H^2-\omega^2H+3H-1), \tag{C25}$$

$$C_4^{VV}(\overline{\alpha},\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{4\omega^2}(2\omega^4H^3+5\omega^2H^2-2\omega^2H+3H-1), \tag{C26}$$

$$C_5^{VV}(\overline{\alpha},-\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{4\omega^2}(2\omega^4H^3-\omega^2H^2-2\omega^2H-3H+1), \tag{C27}$$

$$C_5^{VV}(-\overline{\alpha},\overline{\alpha},\overline{\alpha}) = -\frac{\tau^3}{\omega^2}(2\omega^4H^3+5\omega^2H^2-2\omega^2H+3H-1), \tag{C28}$$

where, by definition, $\omega = e\overline{\alpha}\tau$ and $H = H(\omega)$.

APPENDIX D: SPECIAL FUNCTION REPRESENTATION FOR THE INTEGRALS WHICH APPEAR IN THE PURELY ELECTRIC AND PURELY MAGNETIC CASES

In the main text, we saw that the calculation of the effective action for spinor QED in an external magnetic field reduces to evaluating the following integral (with $\mu=1/2$ in 2+1 dimensions, and $\mu=0$ in 3+1 dimensions):

$$I^{(\mathrm{spin})}(\sigma;\mu)=\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}\frac{d^3}{d\omega^3}(\omega\coth\omega)=\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}\frac{d^3}{d\omega^3}\left(\omega\coth\omega-1-\frac{\omega^2}{3}\right)$$
$$=-\int_0^\infty d\omega\left(\omega\coth\omega-1-\frac{\omega^2}{3}\right)\frac{d^3}{d\omega^3}(\omega^{\mu-1}e^{-\sigma\omega})$$
$$=\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}\left(\coth\omega-\frac{1}{\omega}-\frac{\omega}{3}\right)\left(\frac{(3-\mu)(2-\mu)(1-\mu)}{\omega^2}+\frac{3\sigma(2-\mu)(1-\mu)}{\omega}+3\sigma^2(1-\mu)+\sigma^3\omega\right),\tag{D1}$$

where we integrated by parts (to avoid divergences as $\omega\to 0$ we subtracted the first two terms of the hyperbolic cotangent asymptotes). For large enough values of the parameter μ, one can apply the following table integrals:[27]

$$\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}\coth\omega=\Gamma(\mu)\left[2^{1-\mu}\xi\left(\mu,1+\frac{\sigma}{2}\right)+\sigma^{-\mu}\right],\tag{D2}$$

$$\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}=\sigma^{-\mu}\Gamma(\mu).\tag{D3}$$

Thus, the integral in Eq. (D1) yields

$$I^{(\mathrm{spin})}(\sigma;\mu)=2^{-\mu}\Gamma(\mu+1)\left[\sigma^3\zeta\left(\mu+1,1+\frac{\sigma}{2}\right)+6\sigma^2\frac{1-\mu}{\mu}\zeta\left(\mu,1+\frac{\sigma}{2}\right)-12\sigma\frac{2-\mu}{\mu}\zeta\left(\mu-1,1+\frac{\sigma}{2}\right)+8\frac{3-\mu}{\mu}\zeta\left(\mu-2,1+\frac{\sigma}{2}\right)\right].\tag{D4}$$

As one can easily check, the original integral in Eq. (D1) is well defined for $\mu>-1$. Therefore, the last expression should allow a well defined analytical continuation to the whole that range of values of μ. Notice that this should be true even despite the fact that the intermediate integrals, as in Eqs. (D2) and (D3), may not be well defined for all values $\mu>-1$. In particular, by an analytical continuation, we obtain the results for the values of μ which are of interest,

$$I^{(\mathrm{spin})}\left(\sigma;\frac{1}{2}\right)=2\sqrt{2\pi}\left[5\zeta\left(-\frac{3}{2},1+\frac{\sigma}{2}\right)-9\frac{\sigma}{2}\zeta\left(-\frac{1}{2},1+\frac{\sigma}{2}\right)+3\left(\frac{\sigma}{2}\right)^2\zeta\left(\frac{1}{2},1+\frac{\sigma}{2}\right)+\left(\frac{\sigma}{2}\right)^3\zeta\left(\frac{3}{2},1+\frac{\sigma}{2}\right)\right],\tag{D5}$$

$$I^{(\mathrm{spin})}(\sigma;0)=\frac{11}{6}\sigma^3+\sigma^2-\frac{1}{3}\sigma-\sigma^3\psi\left(1+\frac{\sigma}{2}\right)+6\sigma^2\left[\ln\Gamma\left(1+\frac{\sigma}{2}\right)-\ln\sqrt{2\pi}\right]-24\sigma\zeta'\left(-1,1+\frac{\sigma}{2}\right)+24\zeta'\left(-2,1+\frac{\sigma}{2}\right),\tag{D6}$$

where the prime denotes the derivative of zeta function with respect to its first argument. In derivation of the second expression we used the following identities:[27]

$$\zeta(-1,q)=-\frac{q^2}{2}+\frac{q}{2}-\frac{1}{12},\qquad \zeta(0,q)=\frac{1}{2}-q,$$

$$\zeta(-2,q)=-\frac{q^3}{3}+\frac{q^2}{2}-\frac{q}{6}, \quad \zeta'(0,q)\equiv \left.\frac{\partial\zeta(z,q)}{\partial z}\right|_{z=0}=\ln\Gamma(q)-\ln\sqrt{2\pi}, \tag{D7}$$

$$\lim_{z\to 1}\left(\zeta(z,q)-\frac{1}{z-1}\right)=-\psi(q).$$

In the case of scalar QED, we come to the integral (again, with $\mu=1/2$ in 2+1 dimensions, and $\mu=0$ in 3+1 dimensions),

$$\begin{aligned}
I^{(\mathrm{scal})}(\sigma;\mu)&=\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}\left(\frac{d^3}{d\omega^3}+\frac{d}{d\omega}\right)\frac{\omega}{\sinh\omega}\\
&=\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}\left[\frac{d^3}{d\omega^3}\left(\frac{\omega}{\sinh\omega}-1+\frac{\omega^2}{6}\right)+\frac{d}{d\omega}\left(\frac{\omega}{\sinh\omega}-1\right)\right]\\
&=-\int_0^\infty d\omega\left[\left(\frac{\omega}{\sinh\omega}-1+\frac{\omega^2}{6}\right)\frac{d^3}{d\omega^3}+\left(\frac{\omega}{\sinh\omega}-1\right)\frac{d}{d\omega}\right](\omega^{\mu-1}e^{-\sigma\omega})\\
&=\int_0^\infty d\omega\,\omega^{\mu-1}e^{-\sigma\omega}\left[\left(\frac{1}{\sinh\omega}-\frac{1}{\omega}+\frac{\omega}{6}\right)\left(\frac{(3-\mu)(2-\mu)(1-\mu)}{\omega^2}+\frac{3\sigma(2-\mu)(1-\mu)}{\omega}\right.\right.\\
&\quad\left.\left.+3\sigma^2(1-\mu)+\sigma^3\omega\right)+\left(\frac{1}{\sinh\omega}-\frac{1}{\omega}\right)(1-\mu+\sigma\omega)\right],
\end{aligned}\tag{D8}$$

where we integrated by parts as in the spinor case. In addition to the table integral in (D3), we need also the following one:

$$\int_0^\infty \frac{d\omega\,\omega^{\mu-1}e^{-\sigma\omega}}{\sinh\omega}=2^{1-\mu}\Gamma(\mu)\zeta\left(\mu,\frac{1+\sigma}{2}\right). \tag{D9}$$

Thus, we obtain

$$\begin{aligned}
I^{(\mathrm{scal})}(\sigma;\mu)&=2^{-\mu}\Gamma(\mu+1)\left[\sigma(1+\sigma^2)\zeta\left(\mu+1,\frac{1+\sigma}{2}\right)+2(1+3\sigma^2)\frac{1-\mu}{\mu}\zeta\left(\mu,\frac{1+\sigma}{2}\right)\right.\\
&\quad\left.-12\sigma\frac{2-\mu}{\mu}\zeta\left(\mu-1,\frac{1+\sigma}{2}\right)+8\frac{3-\mu}{\mu}\zeta\left(\mu-2,\frac{1+\sigma}{2}\right)\right].
\end{aligned}\tag{D10}$$

And, finally, by analytical continuation, we obtain the results for two values of μ that are of interest,

$$\begin{aligned}
I^{(\mathrm{scal})}\left(\sigma;\frac{1}{2}\right)&=\sqrt{\frac{\pi}{2}}\left[20\zeta\left(-\frac{3}{2},\frac{1+\sigma}{2}\right)-18\sigma\zeta\left(-\frac{1}{2},\frac{1+\sigma}{2}\right)\right.\\
&\quad\left.+(1+3\sigma^2)\zeta\left(\frac{1}{2},\frac{1+\sigma}{2}\right)+\frac{\sigma}{2}(1+\sigma^2)\zeta\left(\frac{3}{2},\frac{1+\sigma}{2}\right)\right],
\end{aligned}\tag{D11}$$

$$\begin{aligned}
I^{(\mathrm{scal})}(\sigma;0)&=\frac{11}{6}\sigma^3+\frac{7}{6}\sigma-\sigma(1+\sigma^2)\psi\left(\frac{1+\sigma}{2}\right)+2(1+3\sigma^2)\left[\ln\Gamma\left(\frac{1+\sigma}{2}\right)-\ln\sqrt{2\pi}\right]\\
&\quad-24\sigma\zeta'\left(-1,\frac{1+\sigma}{2}\right)+24\zeta'\left(-2,\frac{1+\sigma}{2}\right).
\end{aligned}\tag{D12}$$

[1] W. Heisenberg and H. Euler, Z. Phys. **98**, 714 (1936).
[2] V. Weisskopf, K. Dan. Vidensk. Selsk. Mat. Fys. Medd. **14**, 6 (1936).
[3] J. Schwinger, Phys. Rev. **82**, 664 (1951).
[4] J. Hauknes, Ann. Phys. (NY) **156**, 303 (1984).
[5] C. Martin and D. Vautherin, Phys. Rev. D **38**, 3593 (1988).
[6] D. A. Dicus, C. Kao, and W. W. Repko, Phys. Rev. D **57**, 2443 (1998).
[7] H. W. Lee, P. Y. Pac, and H. K. Shin, Phys. Rev. D **40**, 4202 (1989).
[8] D. Cangemi, E. D'Hoker, and G. Dunne, Phys. Rev. D **51**, R2513 (1995).
[9] V. P. Gusynin and I. A. Shovkovy, Can. J. Phys. **74**, 282 (1996).
[10] G. Dunne and T. Hall, Phys. Lett. B **419**, 322 (1998); Phys. Rev. D **58**, 105022 (1998); *ibid.* **60**, 065002 (1999).
[11] D. G. C. McKeon, Phys. Rev. D **55**, 7989 (1997).
[12] T. D. Gargett and I. N. McArthur, J. Math. Phys. **39**, 4430 (1998).
[13] W. Dittrich and H. Gies, Phys. Lett. B **392**, 182 (1997).
[14] Z. Bern and D. A. Kosower, Phys. Rev. Lett. **66**, 1669 (1991).
[15] Z. Bern and D. A. Kosower, Nucl. Phys. B **379**, 451 (1992).
[16] M. J. Strassler, Nucl. Phys. B **385**, 145 (1992).
[17] D. G. C. McKeon, Can. J. Phys. **71**, 334 (1993); Ann. Phys. **224**, 139 (1993).
[18] M. G. Schmidt and C. Schubert, Phys. Lett. B **331**, 69 (1994); Phys. Rev. D **53**, 2150 (1996).
[19] B. Körs and M. G. Schmidt, Eur. Phys. J. C **6**, 175 (1999).
[20] A. W. Schreiber, C. Alexandrou, and R. Rosenfelder, Nucl. Phys. A **631**, 635c (1998).
[21] J. W. van Holten, Nucl. Phys. B **457**, 375 (1995).
[22] M. Reuter, M. G. Schmidt, and C. Schubert, Ann. Phys. (NY) **259**, 313 (1997).
[23] D. G. C. McKeon and T. Sherry, Mod. Phys. Lett. A **9**, 2167 (1994).
[24] E. S. Fradkin, Nucl. Phys. **76**, 588 (1966); E. S. Fradkin and D. M. Gitman, Phys. Rev. D **44**, 3230 (1991).
[25] V. A. Fock, Sov. Phys. **12**, 404 (1937); P. Pascual and R. Tarrach, *QCD Renormalization for the Practitioner, Lecture Notes in Physics* (Springer, Berlin, 1984), Vol. 194.
[26] I. A. Batalin and A. E. Shabad, Sov. Phys. JETP **33**, 483 (1971).
[27] I. S. Gradshtein and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, Orlando, 1980).
[28] D. Cangemi, E. D'Hoker, and G. Dunne, Phys. Rev. D **52**, R3163 (1995).
[29] L. Lewin, *Polylogarithms and Associated Functions* (Elsevier, North–Holland, Amsterdam, 1981).
[30] W. Dittrich, W.-Y. Tsai, and K.-H. Zimmermann, Phys. Rev. D **19**, 2929 (1979).
[31] J. Ambjørn and P. Olesen, Nucl. Phys. B **330**, 193 (1990); T. Vachaspati, Phys. Lett. B **265**, 258 (1991).
[32] V. P. Gusynin, V. A. Miransky, and I. A. Shovkovy, Phys. Rev. D **52**, 4747 (1995); Nucl. Phys. B **462**, 249 (1996).
[33] A. V. Shpagin, hep-ph/9611412; K. Farakos, G. Koutsoumbas, and N. E. Mavromatos, Phys. Lett. B **431**, 147 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Monopoles and harmonic maps

Theodora Ioannidou and Paul M. Sutcliffe
Institute of Mathematics, University of Kent at Canterbury, Canterbury, CT2 7NZ, United Kingdom

(Received 1 March 1999; accepted for publication 16 June 1999)

Recently Jarvis has proved a correspondence between SU(N) monopoles and rational maps of the Riemann sphere into flag manifolds. Furthermore, he has outlined a construction to obtain the monopole fields from the rational map. In this paper we examine this construction in some detail and provide explicit examples for spherically symmetric SU(N) monopoles with various symmetry breakings. In particular we show how to obtain these monopoles from harmonic maps into complex projective spaces. The approach extends in a natural way to monopoles in hyperbolic space and we use it to construct new spherically symmetric SU(N) hyperbolic monopoles. © *1999 American Institute of Physics.* [S0022-2488(99)03009-1]

I. INTRODUCTION

This paper is concerned with static SU(N) BPS monopoles, which are topological solitons in a Yang–Mills–Higgs gauge theory. The Bogomolny equation, which describes all static monopoles, is integrable and so a variety of techniques are available for studying monopoles, including twistor methods. Despite this fact there are still only a limited number of known explicit monopole solutions, though the integrability of the Bogomolny equation allows many features of monopoles, such as the dimensions of their moduli spaces, to be determined.

An example where the integrability of the Bogomolny equation can be used to prove results on monopoles is the correspondence proved by Jarvis,[1] between monopoles and rational maps from the Riemann sphere into flag manifolds. The rational map arises as the scattering data, along half-lines from the origin, of a linear operator constructed from the monopole fields. Furthermore, in proving the correspondence Jarvis outlines an ''inverse scattering'' procedure whereby the monopole fields can be reconstructed from the rational map. It is this construction which is the focus of this paper. The construction involves solving a nonlinear partial differential equation which is equivalent to the Bogomolny equation, but for which the boundary conditions are given in terms of the rational map. This is the main point of the construction, since for the original Bogomolny equation it is not at all clear how to specify boundary conditions on the fields so as to obtain a unique monopole solution. We perform the construction explicitly for several examples of SU(N) monopoles with spherical symmetry and a variety of symmetry breakings. The solutions are obtained from harmonic maps of the plane into $\mathbb{CP}^{N-1}$, with the degrees of the harmonic maps related to the topological charges of the monopoles.

Perhaps we should make it clear at this point that there are several approaches to studying spherically symmetric monopoles[2–5] and the main aim of this paper is not the construction of new monopole solutions, but rather to gain a better understanding of the correspondence between monopoles and rational maps. In particular we study the construction of monopole solutions from the rational map data, and SU(N) monopoles with spherical symmetry are a good vehicle for this.

The construction of monopoles from rational maps has a natural generalization to monopoles in hyperbolic space. Using this approach we construct explicit solutions for spherically symmetric SU(N) hyperbolic monopoles. As far as we are aware these multimonopole solutions are new. As we shall see, the construction of hyperbolic monopoles has a simplifying feature in comparison to the Euclidean case, and therefore a useful way to obtain the Euclidean solutions is as the zero curvature limit of the hyperbolic ones.

II. SU(N) MONOPOLES

BPS monopoles are finite energy solutions to the Bogomolny equation

$$D_i\Phi=-\tfrac{1}{2}\epsilon_{ijk}F^{jk}, \tag{1}$$

where $D_i=\partial_i+[A_i,]$ is the covariant derivative with A_i, for $i=1,2,3$, an su(N)-valued gauge potential with gauge field $F_{jk}=\partial_j A_k-\partial_k A_j+[A_j,A_k]$. The Higgs field Φ, is an su(N)-valued scalar field for which nontrivial asymptotic boundary conditions are imposed so that topological solitons exist. More precisely, there is a choice of gauge such that in a given direction the Higgs field for large radius r is given by

$$\Phi=i\Phi_0-\frac{i}{r}\Phi_1+O\left(\frac{1}{r^2}\right), \tag{2}$$

where $\Phi_0=\text{diag}(\lambda_1,\lambda_2,...,\lambda_N)$, with the trace-free condition requiring that $\lambda_1+\lambda_2+\cdots+\lambda_N=0$, and we choose the ordering such that $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_N$. Φ_1 is another diagonal matrix, $\Phi_1=\frac{1}{2}\text{diag}(n_1,n_2-n_1,...,n_{N-1}-n_{N-2},-n_{N-1})$, and it can be shown that the numbers $n_1,n_2,\dots,n_{N-1}$ are always integers.

Φ_0 is the vacuum expectation value of Φ and it breaks the symmetry group from SU(N) to a residual symmetry group J, given by the isotropy group of Φ_0. The Higgs field on the two-sphere at infinity defines a map from S^2 to the coset space of vacua SU(N)/J, so that when $\pi_2(\text{SU}(N)/J)$ is nontrivial then all solutions have a topological characterization.

If the λ_i are all distinct then the residual symmetry group is the maximal torus, that is, $J=\text{U}(1)^{N-1}$, and this is known as maximal symmetry breaking. In this case

$$\pi_2\left(\frac{\text{SU}(N)}{\text{U}(1)^{N-1}}\right)=\pi_1(\text{U}(1)^{N-1})=\mathbb{Z}^{N-1} \tag{3}$$

so the monopoles are associated with $N-1$ integers, which are called the topological charges, and are precisely the integers $n_1,n_2,\dots,n_{N-1}$ appearing in Φ_1.

In contrast the case of minimal symmetry breaking is when all but one of the λ_i coincide, so the residual symmetry group is U($N-1$). Since

$$\pi_2\left(\frac{\text{SU}(N)}{\text{U}(N-1)}\right)=\mathbb{Z} \tag{4}$$

there is only one topological charge in this case and the remaining integers are called magnetic weights. The simplest way to distinguish the topological charges from the magnetic weights is to examine the expression for the energy of the monopole.

The condition (2) guarantees that the configuration has finite energy

$$E=\frac{1}{4\pi}\int-\text{tr}\left(\frac{1}{2}F_{ij}^2+(D_i\Phi)^2\right)d^3x. \tag{5}$$

The energy depends only on the topological charges and the asymptotic eigenvalues of Φ, in fact

$$E=(\lambda_1-\lambda_2)n_1+(\lambda_2-\lambda_3)n_2+\cdots+(\lambda_{N-1}-\lambda_N)n_{N-1}. \tag{6}$$

From this expression it can be seen that the difference $\lambda_j-\lambda_{j+1}$ determines the mass of the monopole of type j, of which there are n_j in the given solution. In the minimal symmetry breaking case, where $\lambda_2=\lambda_3=\cdots=\lambda_N$, then all but the first type of monopole becomes massless, so that n_1 remains a topological charge but the remaining integers become magnetic weights and do not contribute to the value of the energy. Note that we cannot distinguish in a gauge invariant way between $n_2 - n_1, \ldots, n_{N-1} - n_{N-2}$ and $-n_{N-1}$, and so when we refer to values of magnetic weights it is understood that this equivalence should be applied.

For intermediate cases of symmetry breaking the residual symmetry group is $J = \mathrm{U}(1)^r \times K$, where K is a rank $N - r - 1$ semisimple Lie group, the exact form of which depends on how the λ_i coincide with each other. Such monopoles have r topological charges.

III. RATIONAL MAPS

In this section we briefly review the recent correspondence proved by Jarvis,[1] between SU(N) monopoles and rational maps from the Riemann sphere into flag manifolds. Actually the correspondence proved in Ref. 1 is more general than this and is valid for all compact semisimple gauge groups G, but in this paper we shall only be concerned with the simplest case of $G = \mathrm{SU}(N)$.

The first step is to introduce polar coordinates, so that a point of $\mathbb{R}^3$ is given by a distance r from the origin and a direction determined by a point z on the Riemann sphere around the origin. In terms of the usual angular coordinates θ, φ this is simply $z = e^{i\varphi} \tan(\theta/2)$.

The Jarvis map is obtained by considering Hitchin's equation

$$(D_r - i\Phi)s = 0 \tag{7}$$

for the complex N-vector s, along each radial half-line from the origin out to infinity, with the direction of the half-line determined by the value of z.

For the moment we shall assume that we are dealing with maximal symmetry breaking. From the boundary conditions (2) we see that since at spatial infinity Φ is in the gauge orbit of Φ_0 then in the N-dimensional solution space there is a one-dimensional subspace generated by the solution which decays at the fastest rate as $r \to \infty$. Now evaluate this solution at $r = 0$. This procedure has thus determined a line in $\mathbb{C}^N$ for each value of z. The next step is to consider how this line varies with the direction z, and the analysis shows that it varies holomorphically. The crucial ingredient here is that the Bogomolny equation (1) implies that $[D_r - i\Phi, D_{\bar{z}}] = 0$, so that the operator in Eq. (7) commutes with the covariant derivative in the angular direction $D_{\bar{z}}$. It can be shown that the degree of this holomorphic map into $\mathbb{CP}^{N-1}$ is precisely the topological charge n_1, and hence the map is rational. Note that if we apply a gauge transformation then the map will be transformed by multiplication by a constant element of SU(N), corresponding to the gauge transformation evaluated at the origin, so that we consider only the equivalence classes of such maps.

Now we repeat the above process but this time we consider the two-dimensional solution space generated by the solution which decays fastest and the solution which decays the next fastest. In the same way as above this will now define a holomorphic plane in $\mathbb{C}^N$ (i.e., a space spanned by two holomorphic lines), which of course will contain the holomorphic line we have already described. The degree of this plane is equal to the topological charge n_2, and so again the map is rational. Proceeding in this way we finally arrive at the rational map $R: \mathbb{CP}^1 \mapsto F(\mathbb{C}^N)$, where $F(\mathbb{C}^N)$ denotes the space of total flags in $\mathbb{C}^N$. This is a series of vector subspaces $0 \subset V_1 \subset V_2 \subset \cdots \subset V_{N-1} \subset \mathbb{C}^N$, where V_i has dimension i, which is clearly the structure we have just described.

In the above discussion the degrees refer to the elements of the homotopy group $\pi_2(F(\mathbb{C}^N)) = \mathbb{Z}^{N-1}$, and are given by the highest powers which occur in some holomorphic polynomials, as described later. For a detailed discussion of rational maps into flag manifolds and their relationship to monopoles the interested reader may find it useful to consult Refs. 6 and 7.

For symmetry breaking which is not maximal the picture is similar, except that now the rational map will not be into the space of total flags, since the exponential decay of some of the solutions will be the same and hence some of the subspaces V_i will be missing from the flag. This of course corresponds to the fact that there will now be fewer topological charges, and these correspond to the degrees of the maps into the vector spaces which remain in the flag.

Because the construction of the rational map from the monopole does not break the rotational symmetry of $\mathbb{R}^3$ it is a very useful approach for studying monopoles with symmetries. For the case of SU(2) there is only one vector subspace; the space of lines in $\mathbb{C}^2$. Thus the rational map is $R:\mathbb{CP}^1 \mapsto \mathbb{CP}^1$, that is, a rational map between Riemann spheres, and its degree is the sole topological charge. By explicit construction of some symmetric maps the existence of various SU(2) monopoles with special symmetries has been proved.[8]

IV. CONSTRUCTING THE MONOPOLE

The proof of the correspondence between monopoles and rational maps[1] involves constructing the monopole from the rational map. The starting point is to write the Bogomolny equation (1) in terms of the coordinates $r,z,\bar{z}$ and observe that a (complex) gauge can always be chosen so that

$$\Phi = -iA_r = -\frac{i}{2}H^{-1}\partial_r H, \quad A_z = H^{-1}\partial_z H, \quad A_{\bar{z}} = 0, \tag{8}$$

where $H \in SL(N,\mathbb{C})$ is a Hermitian matrix.

The Bogomolny equation is then equivalent to the single equation for H

$$\partial_r(H^{-1}\partial_r H) + \frac{(1+|z|^2)^2}{r^2}\partial_{\bar{z}}(H^{-1}\partial_z H) = 0 \tag{9}$$

which we shall refer to as the Jarvis equation. Jarvis[1] then proves that solutions of this equation are determined by the rational map, which specifies the asymptotic boundary conditions on H for large r. The analysis presented in Ref. 1 is complicated and is not very suitable for attempting to implement the construction explicitly, so in this section we shall present a more explicit prescription for determining the boundary conditions on H in terms of the rational map.

For simplicity in this section we shall restrict to the case of SU(2) monopoles. With a choice of normalization for the Higgs field we have the boundary conditions on the monopole as

$$\Phi = \Phi_\infty\left[1 - \frac{n}{2r} + O\left(\frac{1}{r^2}\right)\right] \tag{10}$$

where Φ_∞ is in the gauge orbit of $i\sigma_3 = \text{diag}(i,-i)$, and n is the topological charge.

Any 2×2 Hermitian matrix H, which has unit determinant, can always be written in the form

$$H = \exp\{g(P - \tfrac{1}{2})\} \tag{11}$$

where g is a real function and P is a 2×2 Hermitian projector, that is, $P^\dagger = P = P^2$. A motivation for introducing projectors is that it is a useful formulation for dealing with similar equations that arise in the context of Skyrmions.[9] Examining the asymptotic boundary condition (10) for large r we see that the magnitude of the Higgs field at infinity is a constant and moreover the direction of the Higgs field in the SU(2) algebra is independent of the radius to leading order in $1/r$. Comparing this behavior with Eq. (8) for the Higgs field in terms of H, we find that the leading order behavior for large r is that the profile function g is independent of the angular coordinates $z,\bar{z}$ and the projector P is a function only of the angular coordinates. We are now going to examine the large r behavior of the solution, so we use the above leading order result and set $g(r)$ and $P(z,\bar{z})$.

Computing the Higgs field we obtain

$$\Phi = -\frac{i}{2}H^{-1}\partial_r H = -\frac{i}{2}g'\left(P - \frac{1}{2}\right) \tag{12}$$

with magnitude

$$\|\Phi\|^2=-\tfrac{1}{2}\,\mathrm{tr}(\Phi^2)=\frac{g'^2}{16}=1-\frac{n}{r}+O\left(\frac{1}{r^2}\right). \tag{13}$$

Integrating this equation for g we obtain (there is a choice of sign here that we shall discuss below)

$$g=-4r+2n\log r+O(1). \tag{14}$$

On substituting the form (11) into Eq. (9) and using the asymptotic expression (14) we obtain the result that

$$e^{4r}r^{-2(n+1)}(1+|z|^2)^2[PP_{z\bar{z}}+P_{\bar{z}}P_z]+O\left(\frac{1}{r^2}\right)=0, \tag{15}$$

where subscripts denote partial differentiation. The coefficient of the growing term in (15) must therefore vanish and we find the equation satisfied by P is

$$(PP_z)_{\bar{z}}=0. \tag{16}$$

The equation $PP_z=0$ gives the instanton solutions of the $\mathbb{CP}^1$ σ-model in the plane (see e.g., Ref. 10) and clearly these will satisfy Eq. (16). Furthermore, as we prove in the Appendix, this gives all solutions of Eq. (16).

All instanton solutions of the $\mathbb{CP}^1$ σ-model are given by

$$P=\frac{ff^\dagger}{|f|^2}, \tag{17}$$

where f is a 2-component column vector whose entries are holomorphic functions of z. Note that the multiplication of f by an overall factor does not change the projector P, so that f is an element of $\mathbb{CP}^1$.

Substituting the asymptotic behavior (14) into equation (12) we obtain the expression for the Higgs field on the two-sphere at infinity

$$\Phi_\infty=i(2P-1). \tag{18}$$

The topological charge, n, is the winding number of this map, which is equal to the degree of the holomorphic vector $f(z)$ which is used to construct the projector via (17). Thus we conclude that the boundary condition on H is determined in this simple and explicit way in terms of the degree n rational map $f(z)\colon\mathbb{CP}^1\mapsto\mathbb{CP}^1$.

Note that (17) and (18) give us an explicit expression for the Higgs field at infinity in terms of the rational map. Naively one may think that this does not contain very much information, since for example it is always possible to choose a (singular) gauge in which the Higgs field at infinity is diagonal and constant. However, the important point is that our expression is given in an explicit *known* gauge, and therefore we have removed the gauge freedom and are left with the physical information in the Higgs field: the fact that it is rational.

To be precise, we have not yet proved an equivalence between the rational map f and the one introduced by Jarvis.[1] To prove this equivalence we shall now show that f is indeed the map obtained as the scattering data. (We thank Nick Manton for suggesting this analysis.)

In a unitary gauge there is a basis of solutions to Hitchin's equation (7) which have the leading order large r behavior

$$s\sim e^{-\lambda_j r}v_j,$$

where λ_j is an eigenvalue of $-i\Phi_\infty$ and v_j is the corresponding eigenvector. In the SU(2) case, where $\lambda_1=-\lambda_2=1$, the scattering map is determined by the decaying solution or more fundamentally by the solution associated with the $\lambda_1=1$ eigenspace. Recall that the scattering map is obtained by evaluating this solution at the origin $r=0$. Now, the gauge (8) Hitchin's equation is trivialized to $\partial_r s=0$, so the solutions are r independent and hence the scattering map is the eigenvector of $-i\Phi_\infty$ with eigenvalue one. Thus all that remains to be shown is that f is the eigenvector of $-i\Phi_\infty$ with eigenvalue one. Using the explicit expression (18) and the definition of the projector (17) this is elementary as

$$-i\Phi_\infty f=(2P-1)f=\left(\frac{2ff^\dagger}{|f|^2}-1\right)f=f.$$

At this point it is worth while making a comment about the choice of sign made in Eq. (14) when taking the square root and integrating Eq. (13). If the opposite choice of sign is made then following through the analysis we find that the boundary condition is determined by an antiholomorphic map. Thus this choice of sign is merely an orientation and determines whether we wish monopoles to correspond to holomorphic or antiholomorphic rational maps.

The construction of a monopole from its rational map is now clear. Choose a rational map $f(z)$ and calculate the associated projector (17). Then compute the solution of the Jarvis equation (9) satisfying the boundary condition that for large r

$$H\sim\exp(r(2-4P)). \tag{19}$$

Obviously this construction is not easy to implement explicitly in practice, since it still requires the solution of a nonlinear partial differential equation. In this sense it is not as powerful as say the ADHMN construction,[11] which reduces the problem to solving a set of nonlinear matrix ordinary differential equations plus a further linear system of ordinary differential equations. The advantage is that for the construction discussed here the data is free, in that any rational map is allowed, whereas in the ADHMN construction the Nahm data must satisfy complicated constraints (including the aforementioned set of nonlinear ordinary differential equations). Thus even using the ADHMN construction very few explicit examples of monopole solutions are known. There is always an inherent difficulty associated with solving the monopole equations and the difference between these two alternative constructions is whether the main difficulty resides in performing the construction or specifying the data upon which the construction is performed.

There are simplifying special cases for which we are able to perform the construction explicitly, the easiest example being the rational map $f=(1,z)^t$, which corresponds to the spherically symmetric SU(2) 1-monopole. In this case the asymptotic behavior, $g(r)$ and $P(z,\overline{z})$, is valid for all r and substituting (11) into the Jarvis equation gives the following ordinary differential equation for the profile function

$$g''+\frac{2}{r^2}(1-e^g)=0. \tag{20}$$

The large r boundary condition $g\sim-4r$, together with the condition $g(0)=0$, which is required for H to be well defined at the origin, determines the unique solution of (20) as

$$g=2\log(2r/\sinh 2r). \tag{21}$$

This gives the well-known 1-monopole solution and comparing the asymptotic expansion of (21) with Eq. (14) we verify that $n=1$, so we see explicitly that the topological charge is determined as the degree of the rational map and there is no freedom in the profile function once the map has been specified.

In the following section we provide some explicit examples of solutions to the Jarvis equation, corresponding to spherically symmetric SU(N) monopoles with various symmetry breakings. We present the rational maps and describe how the solutions of the Jarvis equation are obtained from these in terms of harmonic maps into $\mathbb{CP}^{N-1}$.

V. HARMONIC MAPS AND SPHERICAL MONOPOLES

In the first part of this section we briefly review some facts that we shall need about harmonic maps of the $\mathbb{CP}^{N-1}$ σ-model in the plane. These results can be found in, for example, Ref. 10.

A. Harmonic maps

The harmonic map (or σ-model) equations for the $\mathbb{CP}^{N-1}$ model are given by

$$[P_{z\bar{z}},P]=0, \tag{22}$$

where P is an $N\times N$ Hermitian projector.

As stated earlier, one set of solutions to these equations are the instantons given by

$$P(f)=\frac{ff^{\dagger}}{|f|^2} \tag{23}$$

where $f(z)$ is an N-component column vector which is a holomorphic function of z and whose degree is equal to the topological charge of the σ-model. Another set of solutions are the anti-instantons, which have the same form but this time f is an antiholomorphic function, and then the σ-model topological charge is minus the degree of f.

For $N=2$ these are all the finite action solutions, but for $N>2$ there are other noninstanton solutions. These can be described by introducing the operator Δ defined by its action on any vector $f\in\mathbb{C}^N$ as

$$\Delta f=\partial_z f-\frac{f(f^{\dagger}\partial_z f)}{|f|^2} \tag{24}$$

and then define further vectors $\Delta^k f$ by induction: $\Delta^k f=\Delta(\Delta^{k-1}f)$.

To proceed further we note the following useful properties of $\Delta^k f$ when f is holomorphic:

$$(\Delta^k f)^{\dagger}\Delta^l f=0, \quad k\neq l \tag{25}$$

$$\partial_{\bar{z}}(\Delta^k f)=-\Delta^{k-1}f\frac{|\Delta^k f|^2}{|\Delta^{k-1}f|^2}, \quad \partial_z\left(\frac{\Delta^{k-1}f}{|\Delta^{k-1}f|^2}\right)=\frac{\Delta^k f}{|\Delta^{k-1}f|^2}. \tag{26}$$

These properties either follow directly from the definition of Δ or are easy to prove.[10] It is also convenient to define projectors P_k corresponding to the family of vectors $\Delta^k f$ as

$$P_k=P(\Delta^k f), \quad k=0,...,N-1. \tag{27}$$

Applying Δ a total of $N-1$ times to a holomorphic vector gives an antiholomorphic vector, so that a further application of Δ gives the zero vector and hence no corresponding projector.

The projectors P_k are solutions of the harmonic map equations (22) and all solutions can be found in this way by starting with an appropriate holomorphic vector f. In the $\mathbb{CP}^1$ case the operator Δ converts a holomorphic vector to an antiholomorphic vector, that is, instantons to anti-instantons and these are all the solutions in this case.

Note that the projectors obtained from this sequence always satisfy the relation $\Sigma_{k=0}^{N-1}P_k=1$.

For connecting harmonic maps with monopoles it is useful to recall the following interpretation of the noninstanton solutions.[10] From a holomorphic vector f form the exterior product of f and its derivatives as

$$h^k=f\wedge\partial_z f\wedge\cdots\wedge\partial_z^k f, \quad k=0,...,N-1. \tag{28}$$

Thus h^k is holomorphic, though it is an element of a larger dimensional space; it may be represented as a totally antisymmetric tensor with $k+1$ indices. With this notation it may then be shown that

$$\overline{h^{k-1}}\cdot h^k \simeq \Delta^k f \tag{29}$$

where $\cdot$ denotes the summation over all the indices of h^{k-1} and all but the first index of h^k. Here $\simeq$ denotes that two vectors are equal up to an overall factor, which is the important equivalence since we are dealing with elements of projective spaces. Equation (29) leads to the relation

$$\deg(\Delta^k f) = \deg(h^k) - \deg(h^{k-1}) \tag{30}$$

where the left-hand side is defined as the σ-model topological charge of the projector $P_k = P(\Delta^k f)$, and $\deg(h^k)$ is the highest power of z which occurs in the holomorphic tensor h^k. Thus the noninstanton solutions may be interpreted as special mixtures of instantons and anti-instantons.

B. Spherical monopoles

In Sec. IV we saw that the rational map for the spherically symmetric SU(2) 1-monopole is given by $f(z) = (1,z)^t$. This map is spherically symmetric in the sense that a rotation in $\mathbb{R}^3$, which is realized as an SU(2) Möbius transformation

$$z \mapsto \tilde{z} = \frac{\alpha z + \beta}{-\bar{\beta} z + \bar{\alpha}} \quad \text{with } |\alpha|^2 + |\beta|^2 = 1 \tag{31}$$

can be compensated by a constant SU(2) gauge transformation. Explicitly

$$f(\tilde{z}) \simeq \begin{pmatrix} \bar{\alpha} & -\bar{\beta} \\ \beta & \alpha \end{pmatrix} f(z). \tag{32}$$

This is the only spherically symmetric map into $\mathbb{CP}^1$ which has positive degree and hence there are no more spherically symmetric SU(2) monopoles. For SU(N) we first require spherically symmetric maps into $\mathbb{CP}^{N-1}$ and these are given by

$$f = (f_0, \ldots, f_j, \ldots, f_{N-1})^t, \quad \text{where } f_j = z^j \sqrt{\binom{N-1}{j}} \tag{33}$$

and $\binom{N-1}{j}$ denote the binomial coefficients. It can be shown that these maps are spherically symmetric by an explicit presentation of the compensating transformation as in (32). Of course there are other spherically symmetric maps which are obtained by embedding the above maps and setting all other entries to be zero.

For a spherically symmetric SU(N) monopole we require a rational map into the space of total flags $F(\mathbb{C}^N)$, which has spherical symmetry. Thus we need an explicit representation of the holomorphic line, the holomorphic plane (which contains the line), etc. As we shall see in more detail below, we take each k-dimensional subspace to be the space spanned by the vectors $f, \partial_z f, \ldots, \partial_z^{k-1} f$, where f is the spherical map (33). Note that these are precisely the spaces h^{k-1} defined in (28). Thus the topological charges of the monopole, n_k, are given by

$$n_k = \deg(h^{k-1}), \quad k = 1,\ldots,N-1. \tag{34}$$

Hence from (30) it is clear that the monopole topological charges are therefore not equal to the σ-model topological charges of the harmonic maps from which we shall create them. The exception to this statement is the case $N=2$, where all the harmonic maps are instantons and then the only degree is $\deg(h^0)$ which in this case is equal to the σ-model topological charge.

The degree of the map (33) is $N-1$ and hence it is easy to calculate the degree of h^k from (28) which, after taking into account the antisymmetry, gives

$$n_k = \deg(h^{k-1}) = k(N-k), \quad k = 1,\ldots,N-1. \tag{35}$$

Thus we have computed the monopole topological charges and now it remains to construct the corresponding solution of the Jarvis equation. The SU(N) generalization of the SU(2) form given in (11) is to take a sum of the $N-1$ projectors

$$H=\exp\left[g_0\left(P_0-\frac{1}{N}\right)+g_1\left(P_1-\frac{1}{N}\right)+\cdots+g_{N-2}\left(P_{N-2}-\frac{1}{N}\right)\right], \tag{36}$$

where $g_k(r)$ for $k=0,\ldots,N-2$, are profile functions. Recall that the projector P_{N-1} is a linear combination of the other projectors plus the identity matrix, which is why it is not included in the above formula. The profile functions satisfy the regularity condition $g_k(0)=0$, and have a linear growth in r for large r, the coefficients of which determine the symmetry breaking pattern. Once the symmetry breaking is specified the profile functions are, of course, uniquely determined; since there is a one-to-one correspondence between monopoles and rational maps. We shall illustrate this explicitly in the following with some examples.

C. SU(3) examples

For $N=3$, with symmetry breaking to $U(1)\times U(1)$, the charges (35) are $(n_1,n_2)=(2,2)$. From (33) the holomorphic line is given by $f=(1,\sqrt{2}z,z^2)^t$ and the plane is spanned by f and f_z. The SU(3) case has a simplifying feature, in that the holomorphic plane in $\mathbb{C}^3$ can be specified by giving a line orthogonal to the plane; which will then be antiholomorphic. This line is given by

$$f_\perp=\overline{f\times f_z}=\sqrt{2}(\bar{z}^2,-\sqrt{2}\bar{z},1)^t \tag{37}$$

which is clearly antiholomorphic and by construction is orthogonal to the holomorphic plane, that is, $f_\perp^+ f=f_\perp^+ f_z=0$. By inspection of (37) the plane has degree two and clearly has spherical symmetry (compare the structure of $f_\perp$ and f). Hence in this case it is simple to see that the charge is (2, 2). However, as an illustration of the general formalism we shall also present this example in terms of the notation described above. Thus we find

$$h^0=\begin{pmatrix}1\\ \sqrt{2}z\\ z^2\end{pmatrix},\quad h^1=\begin{pmatrix}0 & \sqrt{2} & 2z\\ -\sqrt{2} & 0 & \sqrt{2}z^2\\ -2z & -\sqrt{2}z^2 & 0\end{pmatrix} \tag{38}$$

giving $(n_1,n_2)=(\deg(h^0),\ \deg(h^1))=(2,2)$.

Taking the h^k from (38) we construct the associated projectors, using (29), and insert these into the form for H given in (36). This gives a solution of the Jarvis equation provided the profile functions satisfy the ordinary differential equations

$$-g_0''+\frac{2}{r^2}(e^{g_0-g_1}-1)+\frac{2}{r^2}(e^{g_1}-1)=0,$$

$$-g_1''-\frac{2}{r^2}(e^{g_0-g_1}-1)+\frac{4}{r^2}(e^{g_1}-1)=0. \tag{39}$$

The Higgs field is given in terms of the solution of the Jarvis equation by (8) and the eigenvalues and topological charges can simply be read off by restricting to a given radial line, say $z=0$, which gives

$$\Phi=\frac{i}{6}\,\mathrm{diag}(g_1'-2g_0',g_0'-2g_1',g_0'+g_1').$$

Each profile function has an asymptotic expansion of the form

$$g_k = -\alpha_k r + \beta_k \log r + \log \gamma_k + O\left(\frac{1}{r}\right) \tag{40}$$

with the α_k determined by the vacuum expectation value of the Higgs field. Comparing with (2) for this case we have that

$$\lambda_1 = \frac{2\alpha_0 - \alpha_1}{6}, \quad \lambda_2 = \frac{2\alpha_1 - \alpha_0}{6}, \quad \lambda_3 = -\frac{\alpha_0 + \alpha_1}{6}, \quad n_1 = \frac{2\beta_0 - \beta_1}{3}, \quad n_2 = \frac{\beta_0 + \beta_1}{3}. \tag{41}$$

It is simple to verify that the topological charge is (2, 2) without resorting to an explicit solution of the profile function equations (39). Maximal symmetry breaking implies that $\alpha_0 > \alpha_1 > 0$, so that the terms in (39) which contain exponentials of profile functions do not contribute to the leading order behavior which is $O(1/r^2)$. The coefficients of this leading order term then simply give that $\beta_0 = 4$, $\beta_1 = 2$, which when substituted into (41) confirms that $(n_1, n_2) = (2,2)$.

The explicit solutions for the profile functions can be obtained, for example, if we choose $\Phi_0 = \mathrm{diag}(2,0,-2)$, then the solution is $g_0 = 2g_1 = 2g$, where g is the 1-monopole profile function defined in (21).

If we now consider the case of minimal symmetry breaking then the topological charges which survive will be unchanged, but the magnetic weights will not be given by the topological charges which do not survive. As an example consider the symmetry breaking to U(1)×SU(2) given by $\Phi_0 = \mathrm{diag}(1, -\frac{1}{2}, -\frac{1}{2})$. From (41) this corresponds to setting $\alpha_0 = 3$, $\alpha_1 = 0$. The previous analysis of the profile function equations must now be modified to take into account the fact that exponentials of profile functions may now contribute to leading order (this happens whenever any of the α_k coincide or are zero, and corresponds to changing the symmetry breaking pattern). In this case it is easy to see that (39) requires that $\beta_0 = 3$, $\beta_1 = 0$, $\gamma_1 = \frac{1}{2}$ which gives the values $(n_1, [n_2]) = (2,[1])$, where we have used the notation that square brackets denote magnetic weights rather than topological charges.

The profile function equations that we obtain are related to those derived from the ansatz based approach of Bais *et al.*[3,2] and the methods employed there can be adapted to solve for the profile functions explicitly. This method requires a careful limiting procedure to be taken to deal with nonmaximal symmetry breaking. In Sec. VI we shall see that the solutions for monopoles in hyperbolic space are obtained without the need for this limiting procedure and the Euclidean case can then be obtained from the natural limit in which the curvature of hyperbolic space tends to zero.

For this example the solution is (see Sec. VI)

$$g_0 = \log \frac{81 r^4}{4[(-3r-1)e^{-r} + e^{2r}][(3r-1)e^{r} + e^{-2r}]},$$

$$g_1 = \log \frac{9r^2[(-3r-1)e^{-r} + e^{2r}]}{2[(3r-1)e^{r} + e^{-2r}]^2} \tag{42}$$

and it can be checked that the asymptotic properties are as stated above.

Spherically symmetric monopoles of lower charge, such as the (1, 1) monopole, can be obtained in a similar way by embedding the spherically symmetric maps (33) of lower degree.

D. SU(4) examples

For maximally broken SU(4) the charge, from (35), is (3, 4, 3) and the associated profile function equations are

$$-g_0'' + \frac{3}{r^2}(e^{g_0 - g_1} - 1) + \frac{3}{r^2}(e^{g_2} - 1) = 0,$$

$$-g_1''-\frac{3}{r^2}(e^{g_0-g_1}-1)+\frac{4}{r^2}(e^{g_1-g_2}-1)+\frac{3}{r^2}(e^{g_2}-1)=0,$$

$$-g_2''-\frac{4}{r^2}(e^{g_1-g_2}-1)+\frac{6}{r^2}(e^{g_2}-1)=0. \tag{43}$$

As for the SU(3) case it is a simple task to confirm the topological charge by a leading order analysis of this set of equations. For the choice $\Phi_0=\mathrm{diag}(3,1,-1,-3)$, corresponding to equal monopole masses, the explicit solution is $g_0/3=g_1/2=g_2=g$, where g is given by (21).

There are several possible symmetry breakings and in each case it is a simple matter to determine both the topological charges and magnetic weights by an analysis of equations (43).

For $\Phi_0=\mathrm{diag}(1,\frac{1}{2},\frac{1}{2},-2)$ the symmetry breaking is U(1)×SU(2)×U(1) and the charge is (3, [3], 3). The corresponding explicit solution is

$$g_0=\log\frac{625r^6}{[-25e^{-2r}-(30r-24)e^{-r}+e^{4r}][25e^{2r}-(30r+24)e^{r}-e^{-4r}]},$$

$$g_1=\log\frac{25r^4[25e^{2r}-(30r+24)e^{r}-e^{-4r}]}{2[-25e^{-2r}-(30r-24)e^{-r}+e^{4r}][6e^{-2r}+(5r-6)e^{3r}-(5r+6)e^{-3r}+6e^{2r}]},$$

$$g_2=\log\frac{50r^2[6e^{-2r}+(5r-6)e^{3r}-(5r+6)e^{-3r}+6e^{2r}]}{[-25e^{-2r}-(30r-24)e^{-r}+e^{4r}]^2}. \tag{44}$$

Choosing $\Phi_0=\mathrm{diag}(\frac{3}{4},\frac{3}{4},-\frac{1}{4},-\frac{5}{4})$ gives the symmetry breaking SU(2)×U(1)×U(1) with charge ([2], 4, 3) and solution

$$g_0=\log\frac{256r^6}{9[(4r+3)e^{-3r/2}+e^{5r/2}-4e^{r/2}][(4r-3)e^{3r/2}-e^{-5r/2}+4e^{-r/2}]},$$

$$g_1=\log\frac{16r^4[(4r-3)e^{3r/2}-e^{-5r/2}+4e^{-r/2}]}{3[(4r+3)e^{-3r/2}+e^{5r/2}-4e^{r/2}][(-4r-1)e^{r}+(4r-1)e^{-r}+e^{-3r}+e^{3r}]},$$

$$g_2=\log\frac{16r^2[(-4r-1)e^{r}+(4r-1)e^{-r}+e^{-3r}+e^{3r}]}{3[(4r+3)e^{-3r/2}+e^{5r/2}-4e^{r/2}]^2}. \tag{45}$$

By taking two pairs of eigenvalues to be equal, for example $\Phi_0=\mathrm{diag}(\frac{1}{2},\frac{1}{2},-\frac{1}{2},-\frac{1}{2})$, the symmetry is broken to SU(2)×U(1)×SU(2). In this case the charge is ([2], 4, [2]) and the profile functions are given by

$$g_0=\log\frac{r^6}{9(r\cosh r-\sinh r)^2},\quad g_1=\log\frac{r^4}{3(\sinh^2 r-r^2)},\quad g_2=\log\frac{r^2(\sinh^2 r-r^2)}{3(r\cosh r-\sinh r)^2}. \tag{46}$$

Finally, minimal symmetry breaking to U(1)×SU(3) occurs when three eigenvalues coincide, say $\Phi_0=\mathrm{diag}(3,-1,-1,-1)$ and this gives a charge (3, [2], [1]) with solution

$$g_0=\log\frac{1024r^6}{9[e^{r}(8r^2-4r+1)-e^{-3r}][e^{3r}-e^{-r}(8r^2+4r+1)]},$$

$$g_1=\log\frac{16r^4[e^{3r}-e^{-r}(8r^2+4r+1)]}{3[e^{r}(8r^2-4r+1)-e^{-3r}][e^{2r}(2r^2-r)+e^{-2r}(2r^2+r)]},$$

$$g_2 = \log \frac{64r^2[e^{2r}(2r^2-r)+e^{-2r}(2r^2+r)]}{3[e^r(8r^2-4r+1)-e^{-3r}]^2}. \tag{47}$$

VI. HYPERBOLIC MONOPOLES

Hyperbolic monopoles are solutions of the Bogomolny equation (1) in which Euclidean space $\mathbb{R}^3$ is replaced by hyperbolic 3-space, which we denote by $\mathbb{H}^3_\kappa$, where $-\kappa^2$ is the curvature of hyperbolic space. They were first studied by Atiyah,[12] who observed that S^1 invariant instantons can be interpreted as hyperbolic monopoles. Often hyperbolic monopoles turn out to be easier to study than the Euclidean case and we shall see this explicitly in the following. It has long been expected that in the limit as the curvature of hyperbolic space tends to zero then Euclidean monopoles are recovered, but only recently has this been rigorously established.[13] In this section we shall adapt the methods of Sec. V to the hyperbolic case to obtain spherically symmetric SU(N) monopoles. The SU(2) 1-monopole solution has been obtained before,[14] as a circle invariant instanton, but we believe that all our multi-monopole solutions are new. By explicitly taking the zero curvature limit we recover the Euclidean monopole solutions and explain why this is a simpler way to obtain these solutions than to consider the Euclidean case from the beginning.

Perhaps the most familiar description of $\mathbb{H}^3_\kappa$ is as the interior of the unit ball. In terms of angular coordinates $z,\bar{z}$ and radial coordinate $\rho \in [0,1)$ the metric is

$$ds^2 = \frac{4}{\kappa^2(1-\rho^2)^2}\left(d\rho^2 + \rho^2 \frac{4dzd\bar{z}}{(1+|z|^2)^2}\right) = dr^2 + \frac{\sinh^2(\kappa r)}{\kappa^2}\frac{4\,dz\,d\bar{z}}{(1+|z|^2)^2}, \tag{48}$$

where we have introduced r, the hyperbolic distance from the origin, through the relation $\rho = \tanh(\kappa r/2)$.

The radial scattering analysis proceeds as in the Euclidean case with the upshot that the Jarvis equation (9) in hyperbolic space becomes[13,1]

$$\partial_r(H^{-1}\partial_r H) + \frac{\kappa^2(1+|z|^2)^2}{\sinh^2(\kappa r)}\partial_{\bar{z}}(H^{-1}\partial_z H) = 0. \tag{49}$$

Note that in the zero curvature limit, $\kappa \to 0$, the Euclidean equation (9) is recovered.

Solutions of (49) can be obtained using the form (36), with the same harmonic maps, but leading to modified equations for the profile functions. The equations for the monopole fields in terms of H are still given by (8), but with r now being hyperbolic distance. Hence the asymptotic boundary conditions remain the same as in the Euclidean case and together with the requirement that the profile functions vanish at the origin this determines a unique solution for any given choice of vacuum expectation value Φ_0.

For the SU(2) 1-monopole there is just one profile function, which must satisfy the equation

$$g'' + \frac{2\kappa^2}{\sinh^2(\kappa r)}(1-e^g) = 0. \tag{50}$$

If we again normalize the Higgs field to have unit magnitude then the boundary conditions on the profile function are $g(0)=0$ and $g(r) \sim -4r$ for large r. The solution is

$$g = 2\log \frac{(2+\kappa)\sinh(\kappa r)}{\kappa \sinh((2+\kappa)r)} \tag{51}$$

which gives the known SU(2) hyperbolic 1-monopole.[14]

A. SU(3) examples

The profile function equations for the SU(3) charge (2, 2) hyperbolic monopole are

$$-g_0''\frac{\sinh^2(\kappa r)}{\kappa^2}+2(e^{g_0-g_1}-1)+2(e^{g_1}-1)=0,$$

$$-g_1''\frac{\sinh^2(\kappa r)}{\kappa^2}-2(e^{g_0-g_1}-1)+4(e^{g_1}-1)=0. \tag{52}$$

For equal monopole masses, with $\Phi_0=\mathrm{diag}(2,0,-2)$, the solution is $g_0=2g_1=2g$, with g given by (51). For general Φ_0, including minimal symmetry breaking, we now describe how the solution to (52) can be obtained using Hirota's method.

Introducing the tau-functions τ_0,τ_1 via the transformation

$$g_0=\log\frac{\sinh^4(\kappa r)}{\tau_0\tau_1\kappa^4},\quad g_1=\log\frac{\tau_0\sinh^2(\kappa r)}{\tau_1^2\kappa^2} \tag{53}$$

converts Eq. (52) into Hirota bilinear form

$$\mathcal{D}^2\tau_i\cdot\tau_i+4\tau_{i+1}\tau_{i-1}=0,\quad i=0,1 \tag{54}$$

where we have defined $\tau_{-1}=\tau_2=1$, and $\mathcal{D}$ is the Hirota derivative defined by[15]

$$\mathcal{D}^m\alpha\cdot\beta=(\partial_r-\partial_{\tilde{r}})^m\alpha(r)\beta(\tilde{r})|_{\tilde{r}=r}. \tag{55}$$

The Hirota derivative has many special properties which make the construction of solutions to bilinear equations such as (54) an elegant procedure. In particular from (55) it is clear that its action on exponential functions takes the simple form

$$\mathcal{D}^m e^{\alpha_1 r}\cdot e^{\alpha_2 r}=(\alpha_1-\alpha_2)^m e^{(\alpha_1+\alpha_2)r}. \tag{56}$$

Using this property, together with the bilinear form of the equation, means that it is a simple task to find solutions which are finite sums of exponential functions; in the context of integrable soliton equations, such as the KdV equation, solutions of bilinear equations which are a finite sum of exponentials correspond to multi-solitons.[15]

Note that the bilinear equations (54) are independent of the curvature, $-\kappa^2$, so in particular these equations are the ones which also arise for Euclidean monopoles. However, the transformation (53) involves κ which means that the boundary conditions on the tau-functions are κ dependent, and this is of crucial importance. For hyperbolic monopoles, i.e., $\kappa\neq 0$, the solutions satisfying the required boundary conditions are always given as a simple sum of exponentials, whereas for Euclidean monopoles the boundary conditions (except for the case of maximal symmetry breaking) mean that the solutions are not so simple and involve a sum of products of exponentials and polynomials. By taking the limit $\kappa\to 0$ of a hyperbolic solution the more complicated Euclidean solutions are obtained and this is perhaps the most natural method to construct Euclidean monopoles.

As an example, consider the minimal symmetry breaking of SU(3) obtained from $\Phi_0=\mathrm{diag}(1,-\frac{1}{2},-\frac{1}{2})$. As discussed in Sec. V this choice of Φ_0 corresponds to the large r boundary conditions $g_i\sim-\alpha_i r$, with $\alpha_0=3$, $\alpha_1=0$. Comparing this with the transformation (53) gives the large r boundary conditions

$$\tau_0\sim A_0 e^{(2+2\kappa)r},\quad \tau_1\sim A_1 e^{(1+2\kappa)r} \tag{57}$$

for some constants A_i. The requirement that $g_i(0)=0$ gives the conditions at the origin that $\tau_0=r^2+\cdots$, $\tau_1=r^2+\cdots$ as $r\to 0$. Using the leading order behavior (57) together with the properties of the Hirota derivative it is a simple task to find the following explicit solution

$$\tau_0=\frac{2\kappa e^{(2+2\kappa)r}-(3+4\kappa)e^{-r}+(3+2\kappa)e^{-(1+2\kappa)r}}{(3+2\kappa)(3+4\kappa)\kappa},$$

$$\tau_1=\frac{(3+2\kappa)e^{(1+2\kappa)r}-(3+4\kappa)e^{r}+2\kappa e^{-(2+2\kappa)r}}{(3+2\kappa)(3+4\kappa)\kappa}. \tag{58}$$

As claimed above, we see that there is only an exponential dependence on r; this corresponds to the fact that hyperbolic monopoles approach the vacuum value exponentially, rather than algebraically like Euclidean monopoles.

Taking the limit $\kappa\to 0$ this solution becomes

$$\tau_0=\tfrac{2}{9}(e^{2r}-(3r+1)e^{-r}),\qquad \tau_1=\tfrac{2}{9}((3r-1)e^{r}+e^{-2r}) \tag{59}$$

so we see the emergence of the algebraic factors. Substituting (59) into (53) we obtain the Euclidean monopole solution given by (42).

B. SU(4) examples

The SU(4) equations are

$$-g_0''\frac{\sinh^2(\kappa r)}{\kappa^2}+3(e^{g_0-g_1}-1)+3(e^{g_2}-1)=0,$$

$$-g_1''\frac{\sinh^2(\kappa r)}{\kappa^2}-3(e^{g_0-g_1}-1)+4(e^{g_1-g_2}-1)+3(e^{g_2}-1)=0,$$

$$-g_2''\frac{\sinh^2(\kappa r)}{\kappa^2}-4(e^{g_1-g_2}-1)+6(e^{g_2}-1)=0. \tag{60}$$

The solution $g_0/3=g_1/2=g_2=g$, with g given by (51), corresponds to maximal symmetry breaking with $\Phi_0=\mathrm{diag}(3,1,-1,-3)$.

To obtain the solution for arbitrary Φ_0 we introduce the tau-functions as

$$g_0=\log\frac{\sinh^6(\kappa r)}{\tau_0\tau_2\kappa^6},\quad g_1=\log\frac{\tau_0\sinh^4(\kappa r)}{\tau_1\tau_2\kappa^4},\quad g_2=\log\frac{\tau_1\sinh^2(\kappa r)}{\tau_2^2\kappa^2} \tag{61}$$

which transforms the equation into the Hirota form

$$\mathcal{D}^2\tau_i\cdot\tau_i+2(1+i)(3-i)\tau_{i+1}\tau_{i-1}=0,\quad i=0,1,2 \tag{62}$$

where $\tau_{-1}=\tau_3=1$.

As an example we give the solution for minimal symmetry breaking with $\Phi_0=\mathrm{diag}(3,-1,-1,-1)$, which is

$$\tau_0=3\,\frac{\kappa^2e^{(3+3\kappa)r}-(3\kappa^2+5\kappa+2)e^{(-1+\kappa)r}+(3\kappa^2+8\kappa+4)e^{(-1-\kappa)r}-(\kappa^2+3\kappa+2)e^{(-1-3\kappa)r}}{8(1+\kappa)(2+\kappa)(2+3\kappa)\kappa^2},$$

$$\tau_1=3\,\frac{(2+\kappa)\cosh((2+4\kappa)r)-(4+4\kappa)\cosh((2+2\kappa)r)+(2+3\kappa)\cosh(2r)}{8(1+\kappa)(2+\kappa)(2+3\kappa)\kappa^2},$$

$$\tau_2=3\,\frac{(\kappa^2+3\kappa+2)e^{(1+3\kappa)r}-(3\kappa^2+8\kappa+4)e^{(1+\kappa)r}+(3\kappa^2+5\kappa+2)e^{(1-\kappa)r}-\kappa^2e^{-(3+3\kappa)r}}{8(1+\kappa)(2+\kappa)(2+3\kappa)\kappa^2}.$$

Taking the zero curvature limit results in

$$\tau_0=\frac{3}{32}(e^{3r}-(8r^2+4r+1)e^{-r}),$$

$$\tau_1=\frac{3r}{16}((2r-1)e^{2r}+(2r+1)e^{-2r}), \tag{63}$$

$$\tau_2=\frac{3}{32}((8r^2-4r+1)e^{r}-e^{-3r}),$$

which is the Euclidean monopole solution given in (47).

VII. CONCLUSION

We have studied in some detail the construction of SU(N) monopoles from scattering data which consists of a rational map of the Riemann sphere into a flag manifold. Explicit solutions have been obtained in the case of spherical symmetry and we have shown how these solutions involve harmonic maps of the plane into $\mathbb{CP}^{N-1}$. This approach was generalized to the case of hyperbolic monopoles and new spherically symmetric solutions found, whose zero curvature limit was investigated explicitly.

The Jarvis equation is integrable, but in this paper we have made no use of the Lax pair. The precise description of the boundary conditions in terms of the rational map makes this a very convenient formulation of the Bogomolny equation and it may prove useful to undertake a classical inverse scattering study. An alternative, which is currently under investigation, is the numerical solution of the Jarvis equation, which is more promising than a numerical solution of the Bogomolny equation since the boundary conditions can be specified in a simple manner to ensure the existence of a unique solution.

ACKNOWLEDGMENTS

Many thanks to Conor Houghton, Stuart Jarvis, Nick Manton, Michael Singer, and Wojtek Zakrzewski for useful discussions. P.M.S. acknowledges the EPSRC for an Advanced Fellowship and the Grant No. GR/L88320.

APPENDIX

Let $P(z,\overline{z})$ be a 2×2 Hermitian projector. In this appendix we prove that the only solutions of the equation

$$(PP_z)_{\overline{z}}=0 \tag{A1}$$

are the σ-model instantons given by

$$PP_z=0. \tag{A2}$$

Let $F=PP_z$, then using the fact that P is a projector, which is a solution of (A1), it is clear that F satisfies the following properties:

$$F_{\overline{z}}=0, \tag{A3}$$

$$PF=F, \tag{A4}$$

$$FP=0. \tag{A5}$$

Taking (A4) with (A5) gives that $F^2=0$, so that it has at most rank one and can be written as

$$F=uw^{\dagger}, \tag{A6}$$

where u and w are two orthogonal column vectors, that is $w^{\dagger}u=0$. Substituting this expression for F into (A3) and multiplying both sides by w leads to the result that $w^{\dagger}u_{\bar{z}}=0$, that is, $u_{\bar{z}}$ is orthogonal to w. But since we already know that u is orthogonal to w and they are elements of a two-dimensional vector space then this implies that u and $u_{\bar{z}}$ are parallel. Thus u must have the form $u(z,\overline{z})=g(z,\overline{z})\tilde{u}(z)$, where $\tilde{u}$ is a holomorphic vector and g is some function.

P is a Hermitian projector so it may be written as

$$P=\frac{vv^{\dagger}}{v^{\dagger}v} \tag{A7}$$

for some 2-component column vector v. Substituting the expressions (A6) and (A7) into property (A4) shows that u and v are parallel, that is, $v=\lambda u$ for some function λ. Now using the earlier factorization of u we obtain $v=\lambda g\tilde{u}$, so that finally we arrive at the result that

$$P=\frac{\tilde{u}\tilde{u}^{\dagger}}{\tilde{u}^{\dagger}\tilde{u}}. \tag{A8}$$

As we have already shown that $\tilde{u}$ is a holomorphic vector then this is an instanton solution of Eq. (A2) (see, for example, Ref. 10) and the required result is proved.

[1] S. Jarvis, ''A rational map for Euclidean monopoles via radial scattering,'' Oxford, preprint, 1996.
[2] F. A. Bais and D. Wilkinson, Phys. Rev. D **19**, 2410 (1979).
[3] F. A. Bais and H. A. Weldon, Phys. Rev. Lett. **41**, 601 (1978).
[4] N. Ganoulis, P. Goddard, and D. Olive, Nucl. Phys. B **205**, 601 (1982).
[5] M. C. Bowman, E. Corrigan, P. Goddard, A. Puaca, and A. Soper, Phys. Rev. D **28**, 3100 (1983).
[6] J. Hurtubise, Commun. Math. Phys. **120**, 613 (1989).
[7] M. K. Murray, Commun. Math. Phys. **125**, 661 (1989).
[8] C. J. Houghton, N. S. Manton, and P. M. Sutcliffe, Nucl. Phys. B **510**, 507 (1998).
[9] T. Ioannidou, B. Piette, and W. J. Zakrzewski, hep-th/9811071, 1998.
[10] W. J. Zakrzewski, *Low Dimensional Sigma Models* (IOP, 1989).
[11] W. Nahm, ''The construction of all self-dual multimonopoles by the ADHM method,'' in *Monopoles in Quantum Field Theory*, edited by N. S. Craigie, P. Goddard, and W. Nahm (World Scientific, Singapore, 1982).
[12] M. F. Atiyah, ''Magnetic monopoles in hyperbolic spaces,'' in *M. F. Atiyah Collected Works*, Vol. 5, 579.
[13] S. Jarvis and P. Norbury, Bull. London Math. Soc. **29**, 737 (1997).
[14] A. Chakrabarti, J. Math. Phys. **27**, 340 (1986).
[15] R. Hirota, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey, Topics in Current Physics, Vol. 17 (Springer–Verlag, New York, 1980), p. 157.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Adiabatic evolution for systems with infinitely many eigenvalue crossings

A. Joye
Institut Fourier, Unité Mixte de Recherche CNRS-UJF 5582, Université de Grenoble I, BP 74, 38402 Saint Martin d'Hères Cedex, France

F. Monti
Laboratoire de Physique, CNRS, Université de Bourgogne, BP 400, 21011 Dijon, France and École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

S. Guérin
Laboratoire de Physique, CNRS, Université de Bourgogne, BP 400, 21011 Dijon, France

H. R. Jauslin
Laboratoire de Physique, CNRS, Université de Bourgogne, BP 400, 21011 Dijon, France

(Received 31 December 1998; accepted for publication 22 July 1999)

We formulate an adiabatic theorem adapted to models that present an instantaneous eigenvalue experiencing an infinite number of crossings with the rest of the spectrum. We give an upper bound on the leading correction terms with respect to the adiabatic limit. The result requires only differentiability of the considered projector, and some geometric hypothesis on the local behavior of the eigenvalues at the crossings. © *1999 American Institute of Physics.* [S0022-2488(99)00511-3]

I. INTRODUCTION

The availability of intense pulsed laser sources has opened a large field of possibilities to control atomic and molecular dynamical processes. One of the main theoretical tools to analyze these processes is adiabatic Floquet theory[1] and references therein. The general setup can be described as follows. One considers a molecule described by a Hamiltonian H_0 acting on a Hilbert space $\mathcal{H}$, in interaction with one radiation mode of frequency ω. (The description of the interaction with several modes of different frequencies can be formulated along similar lines.) Since the intensity of the field is quite large, the field is treated as a classical field. The Hamiltonian of the molecule perturbed by the electromagnetic field can be written, for example, as

$$H=H_0+EMF(\omega t+\theta_0), \tag{1}$$

where M is the dipole moment operator of the molecule, $E\in\mathbb{R}$ is a parameter representing the amplitude of the radiation field, F is a real valued 2π-periodic function and θ_0 the initial phase. We assume that H_0 has a discrete spectrum. In order to describe a laser pulse the amplitude is taken as a slowly varying time dependent function $E(\epsilon t)$, where one takes, e.g., $\epsilon=1/T_p$ with T_p the duration of the pulse. A new technique that provides an efficient method for complete transfer of population is based on frequency chirping: within the pulse duration the frequency is also slowly modulated $\omega=\omega(\epsilon t)$.

This model has thus two kinds of time dependencies in the Hamiltonian: one that is periodic and another one that is slowly varying. The periodic part can be treated by Floquet methods, and the slowly varying part by adiabatic theory. Adiabatic Floquet theory is based on the following statement: Assume that in the Hamiltonian (1) the parameter E and the frequency ω are made time dependent, $E(t)$, $\omega(t)$, and M stays time independent. Consider the propagator $U(t,t_0;\theta_0)$, solution of the Schrödinger equation

$$i\frac{\partial}{\partial t}U(t,t_0;\theta_0)=H(\omega t+\theta_0)U(t,t_0;\theta_0),\quad U(t,t;\theta_0)=\mathbb{1} \tag{2}$$

acting on the Hilbert space $\mathcal{H}$. We consider an enlarged Hilbert space by tensoring $\mathcal{H}$ with the space of square integrable functions on the unit circle: $L^2(\mathbb{S}^1,\mathcal{H})$. The operator $U(t,t_0;\theta)$ can be lifted into the enlarged space, interpreting the θ-dependence as a multiplication operator. We can then define

$$U_K(t,t_0)=e^{-t\omega(t)\partial}U(t,t_0;\theta)e^{t_0\omega(t_0)\partial}$$

where $\partial=\partial/\partial\theta$. The statement is that Eq. (2) is equivalent to

$$i\frac{\partial}{\partial t}U_K(t,t_0)=K(t)U_K(t,t_0) \tag{3}$$

with

$$K(t)=-i\varpi(t)\frac{\partial}{\partial\theta}+H_0+E(t)MF(\theta)$$

and $\varpi(t)$ denotes an effective instantaneous frequency defined by $\varpi(t)=\omega(t)+t\,d\omega(t)/dt$. Assuming that the time dependence of $E(t)$, $\omega(t)$ is slow one can develop adiabatic techniques for the evolution of (3). When K has pure point spectrum, the first ingredients are the instantaneous eigenvalues and eigenvectors. They always can be written and labeled in the form

$$\begin{aligned}\lambda_{j,k}&=\lambda_{j,0}+k\varpi,\quad k\in\mathbb{Z}\\ \psi_{j,k}(\theta)&=\psi_{j,0}(\theta)e^{ik\theta}.\end{aligned} \tag{4}$$

The index j has the same cardinality as the dimension of the Hilbert space $\mathcal{H}$. Thus, even if we take simple models with finite dimensional $\mathcal{H}$, the Floquet spectrum has infinitely many eigenvalues. As functions of E and ϖ, these eigenvalues may exhibit crossings, which the adiabatic approximation can accommodate in case there is a finite number of them, see Refs. 2 and 3. The structure (4) of the eigenvalues is such that if we consider a slowly varying *effective* frequency $\varpi(t)$ that goes through 0 at some time t_0, the nature of the spectrum becomes quite different. One can encounter situations in which a branch of instantaneous eigenvalues undergoes an infinite number of crossings with other branches, or the spectrum may become suddenly continuous. Hence it becomes necessary to investigate the validity of the adiabatic theorem in such situations. Let us stress that a strictly positive time dependent frequency $\omega(t)$ may give rise quite naturally to an effective frequency $\varpi(t)$ that goes to zero.[4] Indeed, consider a linear variation of ω of the form

$$\omega(t)=\omega_0-at,$$

with ω_0, $a>0$ on the time interval $[0,\omega_0/a)$, which is far from exotic. Then

$$\varpi(t)=\omega_0-2at$$

goes through zero at $t_0=\omega_0/(2a)\in[0,\omega_0/a)$. As it has been shown in Refs. 4 and 5, the possibility to vary the frequency is a powerful method to enhance the control of molecular processes driven by laser.

We will confine ourselves to the case where a branch of eigenvalue undergoes an infinity of crossings with other branches. As this situation is not generic, as actual crossings are more the exception than the rule, we give below a whole class of systems for which this situation is true. Moreover, it is probably the only case in which we get enough regularity to prove an adiabatic theorem. Note also that in case $\varpi(t)$ passes through 0, the domain of $K(t)$ becomes time dependent, so that technical issues regarding regularity of the evolution operator have to be addressed. This is done in the Appendix A.

The goal of the present paper is to formulate an adiabatic theorem that can be applied to such situations with an estimate on the corrections to the adiabatic limit. Adiabatic Theorems without gap conditions are known to be true, see Ref. 6, however, in general, no estimates on the error terms are available.

While this work was motivated by the physical situation described above and discussed below in the examples, our analysis of the adiabatic approximation is model independent and can be applied to more general situations.

II. ADIABATIC THEOREM

A. Context

The adiabatic approximation in quantum mechanics has a long history which we will not attempt to retrace here. We refer the reader to the recent surveys[7,8] and references therein. Let us simply recall here that the works following that of Born and Fock[2] by Kato,[9] Nenciu,[10] and Avron, Seiler, and Yaffe[11] have led to a formulation of the adiabatic theorem under the usual gap assumption that is general and where the error term is well controlled and of order ϵ. In case the gap assumption is modified, the situation is less explicit. In this section, we switch back to the notation $H(\epsilon t)$ for the slowly varying time-dependent Hamiltonian. Assume $H(s)$ is smooth in $s\in[0,1]$ and there exists a spectral projector $P(s)$ of $H(s)$ which is strongly C^2 on $[0,1]$. Avron and Elgart have shown in Ref. 6 that the adiabatic theorem holds under these conditions, provided $P(s)$ is of finite rank, independently of any spectral considerations. A similar result was proven by Bornemann[18] for discrete hamiltonians in case the set of eigenvalue crossings is of measure zero in time. The limitation of these approaches is that, in general, no estimate can be made on the rate at which the adiabatic regime is attained. In certain specific situations, an estimate on this rate is available. In the case where the spectral measure μ_φ is α-Hölder continuous, with $\varphi=P'(s)\psi(s)$, ψ such that $P=|\psi\rangle\langle\psi|$, the rate of convergence was shown in Ref. 6 to be of order $\epsilon^{\alpha/(2+\alpha)}$. A case where the spectrum of $H(s)$ is assumed to be dense pure point is dealt with in Ref. 12. Another situation, considered in Ref. 13, where the gap hypothesis is not necessarily fulfilled occurs when $H(s)=H_0(s)+\epsilon H_1(s)$, where the domain of $H_1(s)$ is smaller than that of $H_0(s)$. In both cases, the error term remains of order ϵ. In the present article, we consider another situation in which the usual gap assumption is modified and the error made in the adiabatic approximation can be estimated. We make the hypothesis that the projector $P(s)$ is associated with an eigenvalue $\lambda(s)$, in the sense that $H(s)P(s)=\lambda(s)P(s)$, for all $s\in[0,1]$. We assume that $\lambda(s)$ is isolated in the spectrum except at a series of times $\{o_k\}_{k\in\mathbb{N}}$ accumulating at $a\in(0,1)$ where it experiences crossings with the rest of the spectrum. Requiring some conditions on the local behavior of the gap between $\lambda(s)$ and the rest of the spectrum near the crossing points o_k, we estimate the error term in the theorem without *a priori* knowledge on the nature of the rest of the spectrum. Note that for $s=o_k$ such that $\lambda(o_k)$ is not isolated in the spectrum, $P(o_k)$ does not represent the entire spectral projector associated with the eigenvalue $\lambda(o_k)$.

B. One crossing

Let us make more precise the regularity hypotheses under which we shall work. In order to deal with the application described above, we will assume the Hamiltonian is unbounded. This causes technical difficulties motivating the part (ii) of the hypothesis below which justifies our manipulations. We show in the appendix that this assumption is verified for our models. In case $H(s)$ is bounded, this part of the assumption is automatically verified.

$(H0)$ (i) We assume that for all $s\in[0,1]\setminus\{a\}$, $H(s)$ is a strongly C^1 self-adjoint operator defined on a dense domain $\mathcal{D}$ independent of s in a separable Hilbert space $\mathcal{K}$, where $0<a<1$.

Whereas $H(a)$ is bounded self-adjoint on $\mathcal{K}$. We also assume the existence of a projector $P(s)$ of $H(s)$ which is strongly C^2 on $[0,1]$ and such that $H(s)P(s)=P(s)H(s)=\lambda(s)P(s)$, for all $s \in [0,1]$.

(ii) Further assume that the unitary evolution operators $U(s)=U(s,0)$ and $A(s)=A(s,0)$ generated by $H(s)$, respectively $H(s)+\epsilon i[P'(s),P(s)]$ (see (5), (6)) are well defined for all $s \in [0,1]$ and possess the properties (i) to (v) listed in Theorem A.1. Note that $P(s)$ needs not be finite dimensional and λ is continuous.

We start by considering one crossing of λ with the rest of the spectrum by revisiting the strategy proposed in Ref. 2, making use of the general analysis presented in Ref. 11.

Let $g(s)$ be the gap between $\lambda(s)$ and the rest of the spectrum of $H(s)$: $g(s) = \mathrm{dist}(\lambda(s),\sigma(s)\setminus\{\lambda(s)\})\geq 0$, $s\in[0,1]$. We also introduce the bounded, strongly C^1 operator $L(s)=i[P'(s),P(s)]$. We assume that $g^{-1}\{0\}=\{o\}$ and consider the strong differential equations on $\mathcal{D}$

$$i\epsilon U'(s)=H(s)U(s), \quad U(0)=\mathbb{1}, \tag{5}$$

$$i\epsilon A'(s)=(H(s)+\epsilon L(s))A(s), \quad A(0)=\mathbb{1}. \tag{6}$$

The unitary A is the so called *adiabatic evolution* which possesses the well known intertwining relation $A(s)P(0)=P(s)A(s)$.[9,14] Finally, let $W(s)$ be defined by $W(s)=A^{-1}(s)U(s)$. We have on $\mathcal{D}$

$$iW'(s)=-A^{-1}(s)L(s)A(s)W(s), \quad W(0)=\mathbb{1}, \tag{7}$$

in the strong sense. To compare the adiabatic and actual evolutions, we need to compute the size of the difference of the unitary $W(s)$ at two times surrounding the crossing. This is the aim of the next result.

Lemma 2.1: Under the above assumptions, we have for any $0\leq u_0\leq t<o<s\leq u_1\leq 1$,

$$\|W(u_0)-W(u_1)\|\leq C(\epsilon|u_0-t|/g_t^2+\epsilon|u_1-s|/g_s^2+\epsilon/g_t+\epsilon/g_s+|s-t|) \tag{8}$$

where $g_t=\inf_{u\in[u_0,t]}g(u)$, $g_s=\inf_{u\in[s,u_1]g(u)}$ and the constant C is uniform in u_0, u_1, s, and t (see Fig. 1).

Remark: On the basis of the classical paper by Born and Fock,[2] and the detailed analysis of crossings by Hagedorn,[3] one would expect the corresponding estimate without the first two terms.

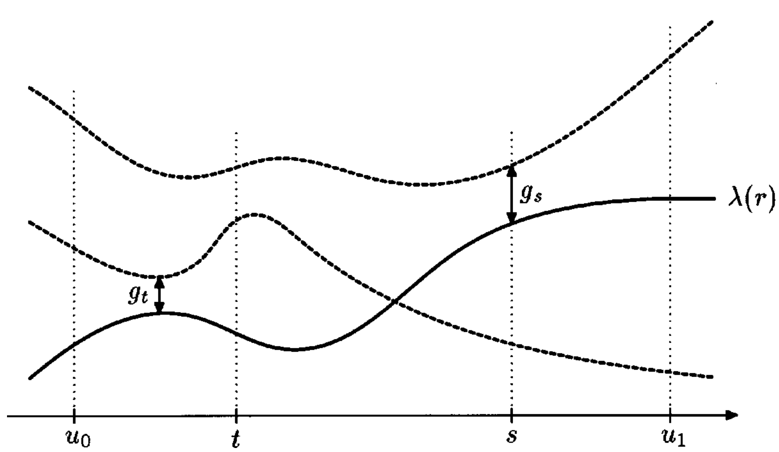

FIG. 1. The various quantities defined in Lemma 2.1.

However, such an estimate requires more detailed knowledge of the structure of spectrum, e.g., that the gap is given by the distance between two eigenvalues, than what we assume in our general setting.

The proof of Lemma 2.1 is presented in Appendix B. The idea of the proof is to integrate Eq. (7) over the interval $[u_0,u_1]$ and then to get estimates of the sizes on each subintervals $[u_0,t]$, $[t,s]$, and $[s,u_1]$ which involves only the gaps.

Lemma 2.1 can be used to treat two standard situations:

(1) If there is a gap G between $\lambda(s)$ and the rest of the spectrum, this lemma implies that the adiabatic approximation holds with an error term bounded by $C\epsilon/G^2$.

(2) If one starts the evolution on a crossing point which splits like s^α near 0, we can use this lemma to show that the adiabatic approximation is valid with an error bounded by

$$\|U(1)-A(1)\|\leqslant C\epsilon^{1/(1+2\alpha)}$$

if ϵ is small enough. This is precisely the situation encountered at the beginning of the interaction of a laser pulse with frequency that is in resonance with the difference between two energy levels of the molecule.[15,16]

To get this estimate, we can consider only half of the problem by letting aside all the terms containing a t and setting $u_1=1$:

$$\|W(1)-W(0)\|\leqslant C(\epsilon|1-s|/g_s^2+\epsilon/g_s+s). \tag{9}$$

This is indeed fully justified by the proof of the lemma (see Appendix B). Next, we have by hypothesis that $g(s)\geqslant g_s=Gs^\alpha$ if s is small. Introducing this behavior in Eq. (9), we obtain $\|W(1)-W(0)\|\leqslant C(\epsilon/s^{2\alpha}+s)$. The result follows now by balancing the two contributions by choosing $s=s(\epsilon)=\epsilon^{1/(1+2\alpha)}$. Again, with more information on the spectrum, as in Refs. 2 and 3, one should be able to improve the above estimate to order $\epsilon^{1/(1+\alpha)}$.

C. Infinite number of crossings

We now have all the information required to proceed to the case of an infinite number of crossings. We make the following hypotheses describing what happens in the neighborhood of each crossing (see Fig. 2).

Spectral hypotheses: There exist two partitions $\{u_k^\pm\}_{k\in\mathbb{N}}$ of $[0,a)$ and $(a,1]$ respectively:

$$0=u_0^-<\cdots<u_{k-1}^-<u_k^-\cdots\to u_\infty^-=a=u_\infty^+\leftarrow\cdots u_k^+<u_{k-1}^+<\cdots<u_0^+=1$$

such that for each $k\in\mathbb{N}^*$,

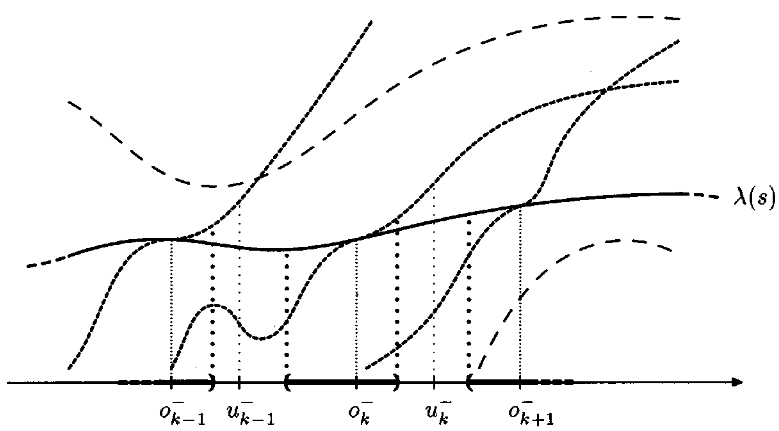

FIG. 2. Illustration of the spectral hypotheses $H1-H2$ on the interval $(0,a)$. The intervals V_l^- are represented by (——).

$(H1)$ one can find non empty open intervals $V_k^{\pm}$, which satisfy $V_k^- \subset [u_{k-1}^-, u_k^-]$, $V_k^+ \subset [u_k^+, u_{k-1}^+]$ and

$$\sup_{s \in V_k^{\pm}} g(s) \leq \inf_{t \in I_k^{\pm}} g(t), \tag{10}$$

where $I_k^- = [u_{k-1}^-, u_k^-] \setminus V_k^-$ and $I_k^+ = [u_k^+, u_{k-1}^+] \setminus V_k^+$.

$(H2)$ there are constants $G_{\pm}(k) > 0$ and a k-independent positive constant α such that for all $s \in V_k^{\pm}$:

$$G_{\pm}(k)|s - o_k^{\pm}|^{\alpha} \leq g(s), \tag{11}$$

for some points $o_k^{\pm} \in V_k^{\pm}$.

Comments: (1) These Spectral Hypotheses mean that the crossings are well separated and that they behave as power of order at most α. Hypothesis $(H1)$ tells us that outside the crossing regions $(V_k^{\pm})$ the gaps are relatively ''large.'' This means that the only accumulation point of small gaps is a.

(2) The choice of a constant exponent α is not as restrictive as it might look at first. Indeed, we are interested in an upper bound, so it is the greatest α that will determine the global behavior.

(3) In the applications, we will consider examples where $g^{-1}\{0\} = \{o_k^{\pm}\}$: the set of crossing points of $\lambda(s)$ with the rest of the spectrum. This implies $\alpha > 0$. But, the case of an infinite number of avoided crossings can be treated by taking $\alpha = 0$ in Hypothesis $(H2)$.

To obtain an estimate for the difference between the real evolution $U(1)$ and the adiabatic one $A(1)$, the idea is to apply Lemma 2.1 on a finite number of crossings and to take a simple integral bound [as in (B2)] over the rest of the interval surrounding a. The choice of the number of crossings will be optimized with respect to ϵ in order to get a simple form for the bound of the remainder term. To state the corresponding result, we need to introduce some notations. Let $\Delta_{\pm}(k) = \max\{|u_k^{\pm} - o_k^{\pm}|, |u_{k-1}^{\pm} - o_k^{\pm}|\}$ and $\tau_{\pm}(k) = \max\{\Delta_{\pm}(k)/G_{\pm}^2(k), \Delta_{\pm}^{\alpha}(k)/G_{\pm}(k)\}$. The functions $K \mapsto |u_K^{\pm} - a| / \sum_{k=1}^{K} \tau_{\pm}(k)^{1/(1+2\alpha)}$ are monotonically decreasing to zero, so, if ϵ is small enough, we define $K_{\pm}(\epsilon) \in \mathbb{N}^*$ as the greatest integer satisfying

$$\frac{|u_K^{\pm} - a|}{\sum_{k=1}^{K} \tau_{\pm}(k)^{1/(1+2\alpha)}} \geq \epsilon^{1/(1+2\alpha)}. \tag{12}$$

This integer always exists if ϵ is sufficiently small and, by construction, $K_{\pm}(\epsilon) \to \infty$ as $\epsilon \to 0$.

Theorem 2.1: For ϵ small enough, under $(H0)$ and the spectral hypotheses $(H1)$, $(H2)$ and provided that

$$\varsigma(\epsilon \tau_{\pm}(k))^{1/(1+2\alpha)} \leq |V_k^{\pm}|/2 \quad \text{for all} \quad 1 \leq k \leq K_{\pm}(\epsilon), \tag{13}$$

for some constant $\varsigma > 0$, we have that

$$U(1) = A(1) + O(\max\{|u_{K_-(\epsilon)}^- - a|, |u_{K_+(\epsilon)}^+ - a|\}).$$

Hence, as $\lim_{\epsilon \to 0} K_{\pm}(\epsilon) = \infty$, $\|U(1) - A(1)\|$ goes to zero for $\epsilon \to 0$ as fast as $\max\{|u_{K_-(\epsilon)}^- - a|, |u_{K_+(\epsilon)}^+ - a|\}$.

Remarks: (1) The theorem states that the error can be estimated provided we can compute the critical value $K_{\pm}(\epsilon)$. Further considerations on the practical aspects of this computation are given in the next section.

(2) Condition (13) implies that the size of the intervals $V_k^{\pm}$ cannot be too small with respect to $\epsilon \tau_{\pm}(k)$.

(3) While we shall apply the theorem in a situation where the spectrum is simple and pure point, the theorem remains valid under the sole existence of an eigenvalue separated from the rest of the spectrum by gaps with the properties stated in $(H1)-(H2)$, without any knowledge on the rest of the spectrum or restriction on the dimension of $P(s)$.

(4) The introduction of an adjustable constant ς is necessary in the following application to satisfy the hypothesis of the theorem.

III. APPLICATION

We can obtain more explicit estimates on the rest by considering some specific behavior at the crossings.

Let us introduce the following notation: $F_k \sim f(k)$ means that there exist two constants $0<c_1<c_2<\infty$ such that $c_1 f(k) \leqslant F_k \leqslant c_2 f(k)$ for $k\in\mathbb{N}^*$ large enough. We have the

Proposition 3.1: Assume the hypothesis of Theorem 2.1 and the following behavior for the relevant quantities:

$$|u_k^{\pm}-a| = C_1/k^{\beta}+C_2/k^{\beta+1}+o(1/k^{\beta+1}), \quad \beta>0,\ C_1\neq 0$$

$$G_{\pm}(k)\sim k^{\gamma},$$

$$|V_k^{\pm}|\sim 1/k^{\delta}, \quad \delta>0.$$

We set $\mu=\min\{\beta+1+2\gamma,\alpha(\beta+1)+\gamma\}$. Then $\|U(1)-A(1)\|=O(\epsilon^p)$ where the exponent p is given by

$$p=\begin{cases} \dfrac{1}{1+2\alpha} & \text{if } \mu>(1+2\alpha) \\ \dfrac{1}{1+2\alpha}-\nu \quad \forall \nu>0 & \text{if } \mu=(1+2\alpha) \\ \dfrac{\beta}{(\beta+1)(1+2\alpha)-\mu} & \text{if } \mu<(1+2\alpha) \end{cases}$$

provided that δ satisfy the following constraints: $\beta+1\leqslant\delta\leqslant\beta+\max\{1,\mu/(1+2\alpha)\}$.

Remark: Let us mention that it can be shown that in case $\alpha=\beta=\gamma=1$ and $\delta=2$, we can take $p=1/3$, instead of $p=1/3-\nu$, for all $\nu>0$. Now, if in Lemma 2.1, the right member were missing the terms $\epsilon|u_0-t|/g_t^2+\epsilon|u_1-s|/g_s^2$, as one would expect with a little more information on the spectrum, an analysis similar to the one provided above leads to an error term of order $\epsilon^{1/3}$. This makes it reasonable to expect that in such a situation the error actually is of that order, as it was the case in the corresponding analysis of one crossing performed in Ref. 2, see Ref. 3. Finally, it is shown in the examples below that the values $\alpha=\beta=\gamma=1$ and $\delta=2$ are generic in some sense.

IV. EXAMPLES

We now consider a family of models for which the situation just described takes place as the effective frequency ϖ takes the value zero. We start by considering the most general model for a two level system driven by a periodic field. The model can be characterized by choosing freely the eigenvalues $\lambda_{+,m}=\lambda_+ + m\omega$ and $\lambda_{-,k}=\lambda_- + k\omega$ and the corresponding eigenfunctions of the form:

$$\psi_{+,m}(\theta)=\begin{pmatrix} e^{ix(\theta)}\cos z(\theta) \\ e^{iy(\theta)}\sin z(\theta)\end{pmatrix} e^{im\theta} \quad \text{and} \quad \psi_{-,k}(\theta)=\begin{pmatrix} -e^{-iy(\theta)}\sin z(\theta) \\ e^{-ix(\theta)}\cos z(\theta)\end{pmatrix} e^{ik\theta}, \tag{14}$$

in which the functions x, y, and z are periodic modulo an integer multiple of θ.

Defining the unitary matrix

$$\Upsilon(\theta)=\begin{pmatrix} e^{ix(\theta)}\cos z(\theta) & -e^{iy(\theta)}\sin z(\theta) \\ e^{iy(\theta)}\sin z(\theta) & e^{-ix(\theta)}\cos z(\theta) \end{pmatrix}$$

the corresponding Floquet Hamiltonian can be written as (dropping the θ dependence in the notation)

$$K=-i\varpi\partial-i\varpi\Upsilon(\partial\Upsilon^{-1})+\Upsilon D\Upsilon^{-1},$$

where $D=\mathrm{diag}(\lambda_+,\lambda_-)$. Using the notation $2\wp=x+y$, $2\vartheta=y-x$ and choosing, without loss of generality, $\lambda_+=-\lambda_-=\lambda$, the Floquet Hamiltonian can be expressed as

$$K=-i\varpi\partial+\begin{pmatrix} \varpi\partial\vartheta+(\lambda-\varpi\partial\wp)\cos(2z) & (-i\varpi\partial z+(\lambda-\varpi\partial\wp)\sin(2z))e^{-2i\vartheta} \\ (i\varpi\partial z+(\lambda-\varpi\partial\wp)\sin(2z))e^{2i\vartheta} & -\varpi\partial\vartheta-(\lambda-\varpi\partial\wp)\cos(2z) \end{pmatrix} \tag{15}$$

where ∂f denotes the derivative with respect to θ. Note that when $\varpi=0$ the operator K reduces to the (matrix) multiplication operator by $\Upsilon(\theta)D\Upsilon^{-1}(\theta)$ on $L^2(\mathbb{S}^1,\mathbb{C}^2)$, whose spectrum consists of two eigenvalues $\pm\lambda$ which are infinitely degenerate. This is to be compared with the general situation where K for $\varpi=0$ becomes a multiplication operator by an arbitrary 2π periodic 2×2 matrix $H(\theta)$. In that case, the spectrum of K is continuous and given by two band functions which are the instantaneous (in θ) eigenvalues of $H(\theta)$.

We will consider two different models with the same eigenvalues but with different eigenfunctions. We remark that since the validity of the adiabatic theorem depends only on the properties of the eigenvalues (and regularity properties of the projectors), it gives the same upper bound for the correction for all the models (15) with equal spectrum. However, it is clear that the theorem is useful if the couplings between considered levels are nonzero. With this regard, we discuss below two examples that have the same spectrum, with an infinite number of crossings. For the first one, which is the widely used RWA (rotating wave approximation) model of quantum optics, the couplings are all equal to zero, except one (see below). The second model is a perturbation of the first one that yields nonzero couplings between the levels.

We choose, for example, the following eigenvalues:

$$\lambda_{\pm,k}(\varpi)=k\varpi\pm(\eta(\varpi)+\varpi)/2, \quad \text{where} \quad \eta(\varpi)=\sqrt{(\varpi-\omega_0)^2+\Omega^2} \tag{16}$$

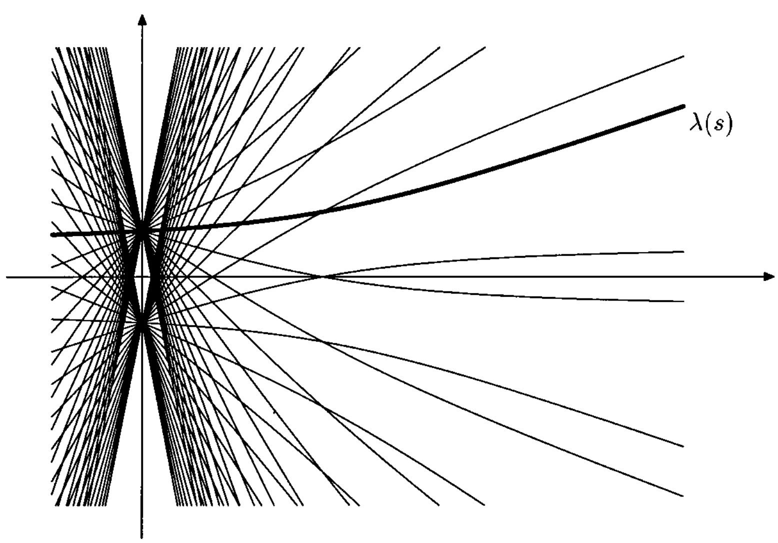

FIG. 3. The first eigenvalues of the RWA and modified RWA models.

and ω_0, Ω are constants. The first model is defined by choosing $x(\theta)=0$, $y(\theta)=\theta$, i.e., $2\vartheta(\theta)=2\wp(\theta)=\theta$ and $\cos(2z)=-(\varpi-\omega_0)/\eta(\varpi)$, $\sin(2z)=\Omega/\eta(\varpi)$, hence z is independent of θ. The corresponding Floquet Hamiltonian is given by

$$K_{\mathrm{RWA}}(\theta)=-i\varpi\frac{\partial}{\partial\theta}+\frac{1}{2}\begin{pmatrix}\omega_0 & \Omega e^{-i\theta}\\ \Omega e^{i\theta} & -\omega_0\end{pmatrix}.$$

The second model is defined by the choice $x(\theta)=-\varrho(\theta)/2$, $y(\theta)=\theta-\varrho(\theta)/2$, i.e., $2\vartheta(\theta)=\theta$, $2\wp(\theta)=\theta-\varrho(\theta)$ and the same z as for the RWA case. This leads to

$$K_M(\theta)=K_{\mathrm{RWA}}(\theta)+\frac{\varpi}{2\eta}\partial\varrho\begin{pmatrix}\omega_0-\varpi & \Omega e^{-i\theta}\\ \Omega e^{i\theta} & \varpi-\omega_0\end{pmatrix}.$$

We consider now a supplementary smooth slow time dependence in the parameter $\varpi=\varpi(s)$ and in $\varrho=\varrho(\theta,s)$. This implies that the eigenvalues, the eigenvectors and the corresponding eigenprojectors are smooth functions of s, so that the regularity Hypothesis $(H0)$(i) is satisfied. We show in appendix that $(H0)$(ii) is satisfied as well for any choice of smooth functions x, y, z, and λ.

We assume, for simplicity, that $\varpi(s)=s$ (but any other smooth monotonic function of s would equally do). This choice corresponds to the chirping that is most often realized in experiments. We select the eigenvalue $\lambda(s)=\lambda_{+,0}(s)=(\eta(s)+s)/2$ and denote by ψ the associated eigenvector (see Fig. 3). The only crossings that λ experiences are with the $\lambda_{-,k+1}$'s and they take place at times s such that

$$\eta(s)=ks, \quad k\in\mathbb{Z}^*. \tag{17}$$

We remark however that these crossings can lead to corrections to adiabaticity, or not, depending on whether the corresponding eigenvectors are coupled. The nonadiabatic coupling among the branches is measured by the following scalar product:

$$\begin{aligned}\langle\psi(s)|\partial_s\psi_{-,k+1}(s)\rangle&=-\frac{1}{2\pi}\int_0^{2\pi}e^{i(k+1)\theta-2i\wp(\theta,s)}(z'(s)-i\sin(2z(s))\vartheta'(\theta,s))d\theta\\&=-\frac{z'(s)}{2\pi}\int_0^{2\pi}e^{ik\theta+i\varrho(\theta,s)}d\theta,\end{aligned}$$

where the $'$ denotes the derivative with respect to s.

Recall that the couplings between the eigenstate $\psi(s)$ associated with the level $\lambda(s)$ and its orthogonal complement in the Hilbert space is given by the operator $L(s)=i[P'(s),P(s)]$, see (7), since the adiabatic evolution $A(s)$ follows the instantaneous eigenspaces. A direct computation of the matrix elements $\langle\psi_{-,k+1}(s)|L(s)\psi(s)\rangle$ with $P'(s)=|\psi'(s)\rangle\langle\psi(s)|+|\psi(s)\rangle\langle\psi'(s)|$ shows that the above scalar product is proportional to the couplings responsible for the nonadiabatic transitions.

For the RWA model, as $\varrho=0$ the nonadiabatic couplings are given by

$$\langle\psi(s)|\partial_s\psi_{-,k+1}(s)\rangle=-z'(s)\,\delta_{k,0}.$$

Thus, the level $\lambda(s)$ is *not* coupled to the infinitely many other levels it crosses. Hence we are led in this case to an effective problem displaying no crossing, so that the error is of order ϵ in this case.

For the other model, we will obtain nonzero couplings at all the crossings, if we choose $\varrho(\theta,s)$ such that $\exp(i\varrho(\theta,s))$ has infinitely many nonzero Fourier components. For example, one can take $\varrho(\theta,s)=\rho(s)\sin(\theta)$ (in particular ρ can be chosen constant). This coupling is then given by

$$\langle \psi(s)|\partial_s \psi_{-,k+1}(s)\rangle = (-1)^{k+1} z'(s) J_k(\rho(s)),$$

where J_k is a Bessel function.

We will now verify that the assumptions of Proposition 3.1 are satisfied. Let us focus on the interval $(0,S]$, for S small enough. The interval $[-S,0)$ can be treated similarly. Again to simplify the notations we will not explicit the $+$ sub/superscripts.

Remark: The preceding two examples have been chosen for their simplicity and explicit complete analytical solvability. However, we emphasize that the following analysis is valid for all the models (15) under the sole assumption that the eigenvalues can be written as $\lambda_{\pm,m}(s) = ms \pm \aleph(s)/2$, where $\aleph$ is a C^2 function with bounded derivatives such that $\aleph(0) > 0$. In particular they are satisfied for the eigenvalues given in (16). The hypotheses imply that the function $f_\zeta(s) = \aleph(s) - \zeta s$ is strictly decreasing for any ζ greater than, say, some ζ_0. Under these conditions the following assertion shows that the crossings that $\lambda(s) = \aleph(s)/2$ experiences with the rest of the spectrum take place at times such that $\aleph(s) = ks$, $k \in \mathbb{N}$ large enough. Again, the actual corrections to adiabaticity will depend on the particular properties of the associated eigenvectors which are measured by the scalar product $\langle \psi(s)|\partial_s \psi_{-,k+1}(s)\rangle$, which generically will not be zero for an infinite number of crossings.

Assertion 1: For $\zeta \geqslant \zeta_0$, the function $f_\zeta(s) = \aleph(s) - \zeta s$ has a unique positive zero o_ζ and if $\zeta < \xi$ we have $o_\zeta > o_\xi$.

From the expansion

$$f_\zeta(s) = \aleph(0) + (\aleph'(0) - \zeta)s + O(s^2),$$

we obtain the behavior of o_ζ:

$$o_\zeta = \frac{\aleph(0)}{\zeta - \aleph'(0)} + O(1/\zeta^3). \tag{18}$$

We define the sequence $u_k > 0$ by the equation:

$$\aleph(u_k) - k u_k = (k+1)u_k - \aleph(u_k), \quad \text{i.e.,} \quad \aleph(u_k) = (k+1/2)u_k. \tag{19}$$

Assertion 1 implies that $u_k < o_k < u_{k-1}$ and, from Eq. (18) and the fact that $u_k = o_{k+1/2}$, we obtain

$$u_k = \frac{\aleph(0)}{k + 1/2 - \aleph'(0)} + O(1/k^3). \tag{20}$$

Next, we have

Assertion 2: On the interval $[u_k, u_{k-1}]$, the spectral gap is given by

$$g(s) = \text{dist}(\lambda(s), \sigma(s) \backslash \{\lambda(s)\}) = |\aleph(s) - ks| \leqslant u_{k-1}/2.$$

More precisely, for $u_k \leqslant s \leqslant o_k$ we have that $g(s) = \aleph(s) - ks \leqslant u_k/2$ and for $o_k \leqslant s \leqslant u_{k-1}$ we have that $g(s) = ks - \aleph(s) \leqslant u_{k-1}/2$.

This assertion is easily proven by considering the different cases.

We now prove that the spectral hypothesis $(H1)$–$(H2)$ are verified. Assertion 1 and Equation (20) show that the sequence $\{u_k\}$ is (for k large enough) monotonically decreasing to $a = 0$. To define the intervals V_k, we choose any point r_k in (o_k, u_{k-1}) such that $g(r_k) = kr_k - \aleph(r_k) \leqslant u_k/2$ and set $V_k = (u_k, r_k)$. The V_k's are disjoint and $I_k = \{u_k\} \cup [r_k, u_{k-1}]$. By definition of V_k, we have that $g(s) \leqslant u_k/2 = g(u_k)$ and for $r_k \leqslant s \leqslant u_{k-1}$ the gap is given by $g(s) = ks - \aleph(s) \geqslant u_k/2$. Whence, hypothesis $(H1)$ is satisfied. Finally to prove that $(H2)$ holds, we need to estimate the behavior of $g(s)$ on V_k: the mean value theorem implies that for each $s \in V_k \backslash \{o_k\}$, there is a q_s, in the interval joining s and o_k, such that

$$g(s) = |\aleph(s) - ks| = |k - \aleph'(q_s)||s - o_k| \sim k|s - o_k|,$$

which shows that $(H2)$ is satisfied with $\alpha=1$ and $G(k)\sim k$.

It remains to check the conditions given in the statement of Proposition 3.1. We have

$$|u_k-0|=u_k=\aleph(0)/k+\aleph(0)(\aleph'(0)-1/2)/k^2+O(1/k^3),\quad \text{i.e.,}\quad \beta=1$$

$$G(k)\sim k \quad \text{i.e.,}\quad \gamma=1, \tag{21}$$

$$|V_k|\sim 1/k^2 \quad \text{i.e.,}\quad \delta=2.$$

To get the estimate for $|V_k|$, we have used that $(u_k,o_k]\subset V_k\subset(u_k,u_{k-1}]$ and the expressions for o_k, and u_k in Eqs. (18) and (20). This implies that, $\mu=\alpha(\beta+1)+\gamma=1+2\alpha$ and $\delta=\beta+1$. So, we can use the second case of Proposition 3.1 to prove that the adiabatic approximation holds for the models:

$$\|U(1)-A(1)\|\leqslant c\epsilon^p,\quad \text{for any}\quad p<\tfrac{1}{3}. \tag{22}$$

In keeping with the first remark of Sec. III, we recall that a more careful analysis yields $p=1/3$.

ACKNOWLEDGMENTS

Supports from the EC contract ERBCHRXCT94-0460 for the project ''Stability and universality in classical mechanics'', from the CNRS and from grants of the Fonds National Suisse de la Recherche Scientifique are acknowledged.

APPENDIX A: TECHNICALITIES

In this appendix, we show that an operator on $L^2(S^1,\mathcal{H})$ of the form

$$K(s,\theta)=-i\varpi(s)\frac{\partial}{\partial\theta}+H(s,\theta), \tag{A1}$$

where $H(s,\theta)$ is a bounded operator in $\mathcal{H}$ such that $s\mapsto H(s,\theta)$ and $s\mapsto\partial/\partial\theta H(s,\theta)$ are norm continuous and $s\mapsto\varpi(s)$ is continuous, admits a strongly continuous unitary propagator $U(s)=U(s,0)$ with all expected regularity properties, even if there is a value a for which $\varpi(a)=0$. Notice that the assumptions on H will be satisfied if, for example, $(s,\theta)\mapsto H(s,\theta)$ is strongly C^1.

The proof relies on a theorem of Kato,[17] which we will restate in a more suitable form for our purpose.

Theorem A.1 (Kato). Let $\mathcal{K}$ and $\mathcal{D}$ be Hilbert spaces such that $\mathcal{D}$ is densely and continuously embedded in $\mathcal{K}$ and let $K(t)$, $0\leqslant t\leqslant T$, be a family of self-adjoint operators in $\mathcal{K}$. Suppose that

(1) $\mathcal{D}\subset\operatorname{dom}K(t)$ for all $0\leqslant t\leqslant T$, whence the $K(t)$ are bounded operators from $\mathcal{D}$ to $\mathcal{K}$, and the application $t\mapsto K(t)$ is norm continuous from $\mathcal{D}$ to $\mathcal{K}$;

(2) there exists a family of isomorphisms $S(t)$ from $\mathcal{D}$ to $\mathcal{K}$ which is strongly continuously differentiable and such that

$$S(t)K(t)S(t)^{-1}=K(t)+B(t)$$

where $B(t)$ is a strongly continuous bounded operator on $\mathcal{K}$.

Under those conditions, there exists a unique family of unitary operators $U(t,s)$ on $\mathcal{K}$ defined for $0\leqslant s,t\leqslant T$ with the following properties:

(i) $U(t,s)$ is strongly continuous on $\mathcal{K}$ in s,t with $U(s,s)=\mathbb{1}$;

(ii) $U(t,r)=U(t,s)U(s,r)$;

(iii) $U(t,s)\mathcal{D}\subset\mathcal{D}$, $\|U(t,s)\|_{\mathcal{D}}\leqslant Ne^{c|t-s|}$ and is strongly continuous on $\mathcal{D}$ in s,t simultaneously;

(iv) $(d/ds)\,U(t,s)\psi=iU(t,s)K(s)\psi$ for any $\psi\in\mathcal{D}$, for $0\leqslant s,t\leqslant T$;

(v) for each $\psi\in\mathcal{D}$ and fixed s, $(d/dt)\,U(t,s)\psi$ exists and is equal to $-iK(t)U(t,s)\psi$ and strongly continuous in $\mathcal{K}$ in t.

To prove this theorem, we apply Theorem 6.1 in Ref. 17 to the operator $A(t)=iK(t)$, which is stable with constants of stability $c=0$ and $N=1$ (see Definition 3.1 and Theorem 4.1 therein). The fact that $U(t,s)$ is unitary follows from the self-adjointness of $K(t)$, the construction of $U(t,s)$ by unitary approximants given in the proofs of Theorem 4.1 and 6.1 in Ref. 17 and the invertibility of $U(t,s)$, which is a consequence of the fact that $A^o(t)=-iK(T-t)$ satifies also the hypothesis of Theorem 6.1 in Ref. 17. See also Remark 5.3 therein.

We now prove that the family of self-adjoint operators defined by Eq. (A1) satisfies the hypothesis of Theorem A.1. To simplify the notation, we will not explicit the θ-dependence and write ∂ for $\partial/\partial\theta$.

Proof: For $\mathcal{D}$, we choose $\mathrm{dom}(-iw_*\partial)$ for some $w_*>0$, and we notice that for any t such that $\varpi(t)\neq 0$, we have that $\mathrm{dom}\,K(t)=\mathcal{D}$ and if $\varpi(t)=0$, then $\mathrm{dom}\,K(t)=\mathcal{K}$. For the norm on $\mathcal{D}$, we choose the graph norm associated to $-iw_*\partial$:

$$\|\psi\|_{\mathcal{D}}^2=\|\psi\|^2+\|-iw_*\partial\psi\|^2\geq\|\psi\|^2.$$

Whence, $\mathcal{D}$ is a dense continuously embedded subspace of $\mathcal{K}$. For any s,t and any $\psi\in\mathcal{D}$, we have

$$\|(K(t)-K(s))\psi\|^2\leq 2\,\frac{|\varpi(t)-\varpi(s)|^2}{w_*^2}\|-iw_*\partial\psi\|^2+2\|H(t)-H(s))\psi\|^2$$
$$\leq 2\,\max\left\{\frac{|\varpi(t)-\varpi(s)|^2}{w_*^2};\left\|H(t)-H(s)\right\|^2\right\}\|\psi\|_{\mathcal{D}}^2.$$

which shows the norm continuity of $K(t)$.

We set $S(t)=S=-iw_*\partial+i$. S is an isomorphism between $\mathcal{D}$ and $\mathcal{K}$ which is strongly differentiable (by t independence). It remains to show that S satisfies Hypothesis (2) of Theorem A.1. For this, we first notice that for any $\psi\in\mathrm{dom}\,K(t)$, we have that $S^{-1}\psi\in\mathcal{D}\subset\mathrm{dom}\,K(t)$ and

$$K(t)S^{-1}\psi=S^{-1}K(t)\psi+H(t)S^{-1}\psi-S^{-1}H(t)\psi=S^{-1}K(t)\psi+S^{-1}SH(t)S^{-1}\psi-S^{-1}H(t)\psi$$
$$=S^{-1}(K(t)-iw_*\partial H(t)S^{-1})\psi. \tag{A2}$$

Whence, for any $\psi\in\mathrm{dom}\,K(t)$, we have that the left-hand side of Eq. (A2) belongs to $\mathcal{D}$. So we can write,

$$SK(t)S^{-1}\psi=K(t)\psi-iw_*\partial H(t)S^{-1}\psi,\quad \text{for all}\quad \psi\epsilon\mathrm{dom}\,K(t).$$

Setting $B(t)=-iw_*\partial H(t)S^{-1}$, we have a strongly continuous bounded operator (by the assumptions on H) which satisfies $SK(t)S^{-1}\supset K(t)+B(t)$. To show the reverse inclusion, we can consider any $b\geq 2\,\sup_t\|B(t)\|$ which implies that ib belongs to the resolvent set of both $K(t)+B(t)$ and $SK(t)S^{-1}$. It follows that $(K(t)+B(t)+ib)^{-1}\subset S(K(t)+ib)^{-1}S^{-1}$. But since the left hand side has domain $\mathcal{K}$, we must have equality between $K(t)+B(t)$ and $SK(t)S^{-1}$ instead of inclusion. □

In the examples of Sec. IV, both $H(s,\theta)$ defined through (15) by means of smooth functions x,y,z,λ of (s,θ), and $H(s,\theta)+\epsilon i[P'(s,\theta),P(s,\theta)]$ where $P(s,\theta)=|\psi(s,\theta)\rangle\langle\psi(s,\theta)|$ with $\psi(s,\theta)$ given by one of the vectors (14) satisfy the hypotheses of the theorem. Hence assumption $(H0)$ (ii) is satisfied for these models.

APPENDIX B: PROOF OF LEMMA 2.1, THEOREM 2.1, AND PROPOSITION 3.1

Proof of Lemma 2.1: The idea of the proof is to integrate Eq. (7) over the interval $[u_0,u_1]$ and then to get ''nice'' estimates of the sizes on each subintervals $[u_0,t]$, $[t,s]$, and $[s,u_1]$. By integrating Eq. (7), we get

$$i(W(u_1)-W(u_0))=-\int_{u_0}^{t}A^{-1}(u)L(u)A(u)W(u)du-\int_{t}^{s}A^{-1}(u)L(u)A(u)W(u)du$$
$$-\int_{s}^{u_1}A^{-1}(u)L(u)A(u)W(u)du. \tag{B1}$$

For the middle term, we simply use the properties of the operator norm and the fact that $A(u)$ and $W(u)$ are unitary to obtain

$$\|W(s)-W(t)\|\leqslant\int_{t}^{s}\|L(u)\|du\leqslant\sup_{u\in[0,1]}\|L(u)\||s-t|, \tag{B2}$$

i.e., we do not care about the behavior of $g(u)$ inside the subinterval $[t,s]$. To estimate the first integral, let $Q(u)=\mathbb{1}-P(u)$. A simple computation, using $P(s)P'(s)P(s)\equiv 0$, shows that

$$P(u)L(u)P(u)=Q(u)L(u)Q(u)=0, \tag{B3}$$

and due to the intertwining property of $A(u)$, we can write

$$W(t)-W(u_0)=i\int_{u_0}^{t}(P(0)A^{-1}(u)L(u)A(u)Q(0)+Q(0)A^{-1}(u)L(u)A(u)P(0))W(u)du. \tag{B4}$$

Now, we need to extract an explicit ϵ dependence from this equality in order to obtain the estimates stated in the lemma. To do this, we follow Ref. 11 and introduce the bounded operator $\mathcal{R}_L(u)$ defined by

$$\mathcal{R}_L(u)=\frac{1}{2i\pi}\oint_{\Gamma(u)}R(u,\lambda)L(u)R(u,\lambda)d\lambda,$$

where $R(u,\lambda)=(H(u)-\lambda)^{-1}$ is the resolvent of $H(u)$ at λ and where the loop $\Gamma(u)$ is a circle centered at $\lambda(u)$ of radius $g(u)/2$. It has the properties (see Refs. 11 and 13)

$$[\mathcal{R}_L(u),H(u)]=[L(u),P(u)], \tag{B5}$$

$$P(u)\mathcal{R}_L(u)P(u)=Q(u)\mathcal{R}_L(u)Q(u)=0. \tag{B6}$$

Standard arguments show that $\mathcal{R}_L(u)$ is strongly C^1 and that

$$\mathcal{R}_L'(u)=\frac{1}{2i\pi}\oint_{\Gamma(u)}(R(u,\lambda)L'(u)R(u,\lambda)-R(u,\lambda)H'(u)R(u,\lambda)L(u)R(u,\lambda)$$
$$-R(u,\lambda)L(u)R(u,\lambda)H'(u)R(u,\lambda))d\lambda, \tag{B7}$$

where $H'(u)R(u,\lambda)$ is to be understood as the bounded operator

$$H'(u)R(u,\lambda)=H'(u)R(u,i)(\mathbb{1}+(\lambda-i)R(u,\lambda)). \tag{B8}$$

Hence, we get the following estimates:

$$\|\mathcal{R}_L(u)\| \leq \frac{|\Gamma(u)|}{2\pi}\|L(u)\|(g(u)/2)^{-2} = 2\|L(u)\|/g(u), \tag{B9}$$

$$\|\mathcal{R}_L'(u)\| \leq c \max\{\|H'(u)R(u,i)\|\|L(u)\|,\|L'(u)\|\}/g^2(u). \tag{B10}$$

The main property of $\mathcal{R}_L(u)$ (see Ref. 11) is that it satisfies for any $\psi\in\mathcal{D}$ the following equalities, as verified by means of (B5):

$$P(0)A^{-1}(u)L(u)A(u)Q(0)\psi = -i\epsilon\frac{d}{du}(P(0)A^{-1}(u)\mathcal{R}_L(u)A(u)Q(0)\psi) + i\epsilon P(0)A^{-1}(u)\mathcal{R}_L'(u)A(u)Q(0)\psi \tag{B11}$$

and

$$Q(0)A^{-1}(u)L(u)A(u)P(0)\psi = i\epsilon\frac{d}{du}(Q(0)A^{-1}(u)\mathcal{R}_L(u)A(u)P(0)\psi) - i\epsilon Q(0)A^{-1}(u)\mathcal{R}_L'(u)A(u)P(0)\psi. \tag{B12}$$

These equations imply that $\int_{u_0}^{t}A^{-1}(u)L(u)A(u)W(u)du$ is proportional to ϵ. Indeed, Equalities (B3) and the intertwining property of $A(u)$ show that the diagonal blocks are 0.

Introducing Equalities (B11) and (B12) in Eq. (B4), we get

$$\begin{aligned} W(t)-W(u_0) = &-\epsilon\int_{u_0}^{t}\frac{d}{du}(Q(0)A^{-1}(u)\mathcal{R}_L(u)A(u)P(0) \\ &- P(0)A^{-1}(u)\mathcal{R}_L(u)A(u)Q(0))W(u)du - \epsilon\int_{u_0}^{t}(P(0)A^{-1}(u)\mathcal{R}_L'(u)A(u)Q(0) \\ &- Q(0)A^{-1}(u)\mathcal{R}_L'(u)A(u)P(0))W(u)du. \end{aligned} \tag{B13}$$

Performing an integration by part in the first integral, using the differential equation (7) for $W(u)$ and taking into account that $A(u)$, $W(u)$ are unitary and $P(0)$, $Q(0)$ are projectors, gives us the following bound for the norm of the difference $W(t)-W(u_0)$:

$$\begin{aligned} \|W(t)-W(u_0)\| \leq 2\epsilon(&\|\mathcal{R}_L(t)\| + \|\mathcal{R}_L(u_0)\| + \sup_{u\in[u_0,t]}\|\mathcal{R}_L(u)\|\|L(u)\|(t-u_0) \\ &+ \sup_{u\in[u_0,t]}\|\mathcal{R}_L'(u)\|(t-u_0)). \end{aligned} \tag{B14}$$

Next, we use first Estimates (B9) and (B10) and then the fact that $0\leq u_0<t\leq 1$ to obtain the desired bound:

$$\begin{aligned} \|W(t)-W(u_0)\| &\leq \frac{8\epsilon}{g_t}\sup_{u\in[u_0,t]}\|L(u)\| + \frac{4\epsilon}{g_t}\sup_{u\in[u_0,t]}\|L(u)\|^2(t-u_0) + c\,\frac{2\epsilon}{g_t^2} \\ &\quad\times\sup_{u\in[u_0,t]}\{\|H'(u)R(u,i)\|\|L(u)\|,\|L'(u)\|\}(t-u_0) \\ &\leq 12\frac{\epsilon}{g_t}\sup_{u\in[0,1]}\{\|L(u)\|,\|L(u)\|^2\} + 2\,\frac{\epsilon|t-u_0|}{g_t^2} \\ &\quad\times\sup_{u\in[0,1]}\{\|H'(u)R(u,i)\|\|L(u)\|,\|L'(u)\|\} \\ &\leq c_2\left(\frac{\epsilon}{g_t} + \frac{\epsilon|t-u_0|}{g_t^2}\right). \end{aligned} \tag{B15}$$

Using the same kind of arguments, shows that on the subinterval $[s,u_1]$, we have

$$\|W(u_1)-W(s)\|\leq c_2\left(\frac{\epsilon|s-u_1|}{g_s^2}+\frac{\epsilon}{g_s}\right). \tag{B16}$$

Combining estimates (B2), (B15), and (B16) gives the announced bound for $\|W(u_1)-W(u_0)\|$.□

Proof of Theorem 2.1: In the sequel, we will denote by the same symbol c all inessential constants. Let us consider the interval $[0;a)$. In order to simplify the notations, we will not write the subscripts/superscripts $-$. Picking some $t,s\in V_k$ such that $t<o_k<s$ and $|t-o_k|=|s-o_k|$, we get

$$\begin{aligned}\|W(u_k)-W(u_{k-1})\|&\leq c(\epsilon|t-u_{k-1}|/g_t^2+\epsilon|s-u_k|/g_s^2+\epsilon/g_t+\epsilon/g_s+|t-s|)\\&\leq c\left(\epsilon\frac{\Delta(k)}{G(k)^2}|t-o_k|^{-2\alpha}+\epsilon\frac{1}{G(k)}|t-o_k|^{-\alpha}+|t-o_k|\right)\\&\leq c\left(\epsilon\frac{\Delta(k)}{G(k)^2}|t-o_k|^{-2\alpha}+\epsilon\frac{\Delta^{\alpha}(k)}{G(k)}|t-o_k|^{-2\alpha}+|t-o_k|\right) \qquad (B17)\\&\leq c(\epsilon\tau(k)|t-o_k|^{-2\alpha}+|t-o_k|) \qquad (B18)\end{aligned}$$

by the preceding section. Indeed, we have that $g_t=\inf_{u\in[u_{k-1},t]}g(u)=g(r_t)$ for some $r_t\in[u_{k-1},t]$. Now, by Hypothesis $(H1)$, $r_t\in V_k$. Whence, we have that

$$g_t=g(r_t)\geq G(k)|r_t-o_k|^{\alpha}\geq G(k)|t-o_k|^{\alpha}$$

as $r_t\leq t\leq o_k$. Using the same kind of arguments, we can show that $g_s=\inf_{u\in[s,u_k]}g(u)\geq G(k)|s-o_k|^{\alpha}$. Finally to obtain the bound (B17), it remains to notice that $|s-t|=|t-o_k|+|s-o_k|=2|t-o_k|$ together with $|t-o_k|$, $|t-u_{k-1}|\leq\Delta(k)$ and $|s-o_k|$, $|s-u_k|\leq\Delta(k)$.

We now get an estimate by choosing $t=t(\epsilon,k)$ in order to balance the two contributions appearing in the last term of Eq. (B17) above: for some constant $\varsigma>0$, we set

$$\frac{\varsigma^{1+2\alpha}\epsilon\tau(k)}{|t(\epsilon,k)-o_k|^{2\alpha}}=|t(\epsilon,k)-o_k|, \tag{B19}$$

i.e.,

$$|t(\epsilon,k)-o_k|=\varsigma(\epsilon\tau(k))^{1/(1+2\alpha)}. \tag{B20}$$

By definition, $t(\epsilon,k)\in V_k$, hence, as k will eventually be bounded from above by $K(\epsilon)$, this imposes Condition (13) in the statement of the theorem. Replacing t by $t(\epsilon,k)$ in (B17) and summing over k, we get for any $K\leq K(\epsilon)$,

$$\|W(0)-W(u_K)\|\leq c(\varsigma+\varsigma^{-2\alpha})\sum_{k=1}^{K}(\epsilon\tau(k))^{1/(1+2\alpha)}. \tag{B21}$$

On the other hand, using the differential Eq. (7), we obtain

$$\|W(u_K)-W(a)\|\leq\int_{u_K}^{a}\|L(u)\|du\leq c|u_K-a|. \tag{B22}$$

Again, we balance the two right-hand sides in (B21) and (B22) by setting the integer $K=K(\epsilon)$, which has been defined in Eq. (12). Consequently,

$$\|W(0)-W(a)\| \leqslant c\left((\varsigma+\varsigma^{-2\alpha})\epsilon^{1/(1+2\alpha)}\sum_{k=1}^{K(\epsilon)} \tau(k)^{1/(1+2\alpha)}+|u_{K(\epsilon)}-a|\right)$$
$$\leqslant c(\varsigma+\varsigma^{-2\alpha}+1)|u_{K(\epsilon)}-a| \equiv C(\varsigma)|u_{K(\epsilon)}-a|, \tag{B23}$$

where $C(\varsigma)$ is independent of ϵ. Proceeding similarly on $(a,1]$ completes the proof. □

Remark: In the step (B17) we deliberately lost a little in the estimate by using $|t-o_k|^{-\alpha} \leqslant \Delta_k^{\alpha}|t-o_k|^{-2\alpha}$ in order to simplify the subsequent arguments. It is nevertheless possible to get slightly sharper results by not adopting this simplification, however the analysis gets more involved and less transparent. We simply note here that in the examples discussed in this paper, this more careful analysis yields, for the generic situation, an error term of order ϵ^p with an exponent $p=1/3$, instead of the value $p=1/3-\nu$, for any $\nu>0$ obtained there.

Proof of Proposition 3.1: The idea of the proof is to explicit conditions on the different exponents ensuring the validity of Theorem 2.1. We will only consider the interval $[0,a)$, the same kind of arguments will apply on $(a,1]$. Again, in order to simplify the notations we will let aside the subscripts/superscripts $-$.

First, we have that $2\Delta(k)=u_k-u_{k-1}=C_1\beta/k^{\beta+1}+o(1/k^{\beta+1})\sim 1/k^{\beta+1}$, which implies that

$$\delta \geqslant \beta+1>0, \tag{B24}$$

since $2\Delta(k)\geqslant |V_k|\sim 1/k^{\delta}$. Notice that the length of the V_k can be rescaled by a uniform constant if $\delta=\beta+1$.

Next, $\Delta(k)/G^2(k)\sim 1/k^{\beta+1+2\gamma}$ and $\Delta^{\alpha}(k)/G(k)\sim 1/k^{\alpha(\beta+1)+\gamma}$. So, if we denote by $\mu=\min\{\beta+1+2\gamma,\alpha(\beta+1)+\gamma\}$ then $\tau(k)=\max\{\Delta(k)/G^2(k), \Delta^{\alpha}(k)/G(k)\}\sim 1/k^{\mu}$ by increasing the overall constant in Theorem 2.1 if necessary. Whence,

$$\sum_{k=1}^{K}\tau(k)^{1/(1+2\alpha)}\sim\sum_{k=1}^{K}k^{-\mu/(1+2\alpha)}\sim\begin{cases}K^0 & \text{if } \mu>1+2\alpha\\ \log K & \text{if } \mu=1+2\alpha\\ K^{1-\mu/(1+2\alpha)} & \text{if } \mu<1+2\alpha\end{cases} \tag{B25}$$

and considering the definition of $K(\epsilon)$ [see Eq. (12)], we obtain

$$\epsilon^{1/(1+2\alpha)}\sim\frac{|u_{K(\epsilon)}-a|}{\sum_{k=1}^{K(\epsilon)}\tau(k)^{1/(1+2\alpha)}}\sim\begin{cases}K(\epsilon)^{-\beta} & \text{if } \mu>1+2\alpha\\ K(\epsilon)^{-\beta}/\log K(\epsilon) & \text{if } \mu=1+2\alpha\\ K(\epsilon)^{-\beta-1+\mu/(1+2\alpha)} & \text{if } \mu<1+2\alpha\end{cases}. \tag{B26}$$

Condition (13) stated in Theorem 2.1 reads

$$\varsigma(\epsilon\tau(k))^{1/(1+2\alpha)}\leqslant |V_k|/2 \tag{B27}$$

for all $1\leqslant k\leqslant K(\epsilon)$. Notice that this condition is automatically satisfied if $\delta<\mu/(1+2\alpha)$. In general, it will be satisfied for a sufficiently small ς, if

$$F(\epsilon)\equiv\epsilon^{1/(1+2\alpha)}K(\epsilon)^{\delta-\mu/(1+2\alpha)} \tag{B28}$$

remains bounded as $\epsilon\to 0$. Using (B26), we have

$$F(\epsilon)\sim\begin{cases}K(\epsilon)^{\delta-\beta-\mu/(1+2\alpha)} & \text{if } \mu>1+2\alpha,\\ K(\epsilon)^{\delta-1-\beta}/\log K(\epsilon) & \text{if } \mu=1+2\alpha,\\ K(\epsilon)^{\delta-\beta-1} & \text{if } \mu<1+2\alpha.\end{cases} \tag{B29}$$

As $K(\epsilon)\to\infty$ for $\epsilon\to 0$, Eq. (B29) implies that $F(\epsilon)$ will remain bounded if $\delta\leqslant\beta+\max\{1;\mu/(1+2\alpha)\}$.

Hence, using (B23) and (B26), we get that the adiabatic theorem (2.1) holds with a remainder term on $[0,a)$,

$$O(|u_{K_-(\epsilon)}-\alpha|)=O(K(\epsilon)^{-\beta})=O(\epsilon^p)$$

where the exponent p is given by

$$p=\begin{cases}\dfrac{1}{1+2\alpha} & \text{if } \mu>(1+2\alpha)\\[2mm] \dfrac{1}{1+2\alpha}-\nu\ \forall\nu>0 & \text{if } \mu=(1+2\alpha)\\[2mm] \dfrac{\beta}{(\beta+1)(1+2\alpha)-\mu} & \text{if } \mu<(1+2\alpha)\end{cases}$$

provided that $\beta+1\leqslant\delta\leqslant\beta+\max\{1;\mu/(1+2\alpha)\}$. To determine p in case $\mu=1+2\alpha$ and $\delta=\beta+1$, we have used the estimate $\epsilon^{-1/(1+2\alpha)}\sim K(\epsilon)^{\beta}\log K(\epsilon)<K(\epsilon)^{\beta+\nu'}$ for all $\nu'>0$. This ends the proof of the proposition. □

[1] S. Guérin, F. Monti, J.-M. Dupont, and H. R. Jauslin, ''On the Relation Between Cavity-Dressed States, Floquet States, RWA and Semiclassical Models,'' J. Phys. A **30**, 7193–7215 (1997).
[2] M. Born and V. Fock, ''Beweis des Adiabatensatzes,'' Z. Phys. **51**, 165–180 (1928).
[3] G. Hagedorn, ''Adiabatic Expansions near Eigenvalue Crossings,'' Ann. Phys. (N.Y.) **196**, 278–295 (1989).
[4] S. Guérin, ''Complete dissociation by chirped laser pulses designed by adiabatic Floquet analysis,'' Phys. Rev. A **56**, 1458–1462 (1997).
[5] S. Chelkowski, A. D. Bandrauk, and P. B. Corkum, ''Efficient molecular dissociation by a chirped ultrashort infrared laser pulse,'' Phys. Rev. Lett. **65**, 2355–2358 (1990).
[6] J. E. Avron and A. Elgart ''Adiabatic Theorem without a Gap Condition,'' Commun. Math. Phys. **203**, 445–463 (1999).
[7] J. E. Avron and A. Elgart, ''An Adiabatic Theorem without a Gap Condition,'' to appear in *Proceedings of Qmath7*, Prague (1998).
[8] A. Joye and Ch.-Ed. Pfister, ''Exponential Estimates in Adiabatic Quantum Evolution,'' to appear in: *Proceedings of the XIIth International Congress of Mathematical Physics, Brisbane, 13–19 July 1997*.
[9] T. Kato, ''On the Adiabatic Theorem of Quantum Mechanics,'' J. Phys. Soc. Jpn. **5**, 435–439 (1950).
[10] G. Nenciu, ''On the Adiabatic Theorem of Quantum Mechanics,'' J. Phys. A **13**, L15–L18 (1980).
[11] J. E. Avron, R. Seiler, and L. G. Yaffe, ''Adiabatic Theorems and Applications to the Quantum Hall Effect,'' Commun. Math. Phys. **110**, 33–49 (1987).
[12] J. Avron, J. Howland, and B. Simon, ''Adiabatic Theorems for Dense Point Spectra,'' Commun. Math. Phys. **128**, 497–507 (1990).
[13] A. Joye, ''An Adiabatic Theorem for Singularly Perturbed Hamiltonian,'' Ann. Inst. Henri Poincaré, Sect. A **63**, 231–250 (1995).
[14] S. G. Krein, *Linear Differential Equations in Banach Space* (American Mathematical Society, Providence, 1971).
[15] M. Holthaus, ''Pulse-Shape-controlled Tunnelling in a Laser Field,'' Phys. Rev. Lett. **69**, 1596 (1992).
[16] S. Guérin and H. R. Jauslin, ''Laser-Enhanced Tunnelling Through Resonant Intermediate Levels,'' Phys. Rev. A **55**, 1262–1275 (1997).
[17] T. Kato, ''Linear Evolution Equation of ''Hyperbolic'' Type,'' J. Fac. Sci., Univ. Tokyo, Sect. 1 **17**, 241–258 (1970).
[18] F. Bornemann, ''Homogenization in time of singularly perturbed mechanical systems,'' Lec. Notes Math. **1687** Springer (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

New solvable and quasiexactly solvable periodic potentials

Avinash Khare[a)]
Institute of Physics, Sachivalaya Marg, Bhubaneswar 751005, Orissa, India

Uday Sukhatme[b)]
Department of Physics, University of Illinois at Chicago, Chicago, Illinois 60607-7059

(Received 22 July 1999; accepted for publication 11 August 1999)

Using the formalism of supersymmetric quantum mechanics, we obtain a large number of new analytically solvable one-dimensional periodic potentials and study their properties. More specifically, the supersymmetric partners of the Lamé potentials $ma(a+1)\,\mathrm{sn}^2(x,m)$ are computed for integer values $a=1,2,3,\ldots$. For all cases (except $a=1$), we show that the partner potential is distinctly different from the original Lamé potential, even though they both have the same energy band structure. We also derive and discuss the energy band edges of the associated Lamé potentials $pm\,\mathrm{sn}^2(x,m)+qm\,\mathrm{cn}^2(x,m)/\mathrm{dn}^2(x,m)$, which constitute a much richer class of periodic problems. Computation of their supersymmetric partners yields many additional new solvable and quasiexactly solvable periodic potentials. © *1999 American Institute of Physics.* [S0022-2488(99)03511-2]

I. INTRODUCTION

The energy spectrum of electrons on a lattice is of central importance in condensed matter physics. In particular, knowledge of the existence and locations of band edges and band gaps determines many physical properties. Unfortunately, even in one dimension, there are very few analytically solvable periodic potential problems in quantum mechanics. The aim of this paper is to extend the small currently known set of analytically solvable periodic potentials.

For a potential with period L, one is seeking solutions of the Schrödinger equation subject to the Bloch condition

$$\psi(x)=e^{ikL}\psi(x+L), \tag{1}$$

where k denotes the crystal momentum. The spectrum shows energy bands whose edges correspond to $kL=0,\ \pi$, that is the wave functions at the band edges satisfy $\psi(x)=\pm\psi(x+L)$. For periodic potentials, the band edge energies and wave functions are often called eigenvalues and eigenfunctions, and we will also use this terminology. The classic textbook example which is used to demonstrate band structure is the Kronig–Penney model,

$$V(x)=\sum_{n=-\infty}^{\infty} V_0\,\delta(x-nL).$$

It should be noted that the band edges for the Kronig–Penney model can only be computed by solving a transcendental equation. Another well-studied class of periodic potentials is

$$V(x)=pm\,\mathrm{sn}^2(x,m), \quad p\equiv a(a+1). \tag{2}$$

[a)]Electronic mail: khare@iopb.res.in
[b)]Electronic mail: sukhatme@uic.edu

Here $\mathrm{sn}(x,m)$ is a Jacobi elliptic function of real elliptic modulus parameter $m(0\leqslant m\leqslant 1)$ with period $4K(m)$. For simplicity, from now on, we will not explicitly display the modulus parameter m as an argument of Jacobi elliptic functions.[1] The elliptic function potentials of Eq. (2) have a period $L=2K(m)$, and will be referred to as Lamé potentials, since the corresponding Schrödinger equation is called Lamé's equation.[2,3] It is well known that for any integer value $a=1,2,3,...$, the corresponding Lamé potential (2) has a bound bands followed by a continuum band.[2,3] All band edge energies and wave functions are analytically known.

At this point it is worth recalling that supersymmetric quantum mechanics (SUSYQM) has proved useful in discovering many, new, analytically solvable potentials on both the full as well as the half line.[4] It is then natural to enquire if one can also use similar techniques to discover new solvable periodic potentials. In this paper, we demonstrate that this is indeed possible.

Our work is inspired by several recent papers,[5–8] which discuss various general aspects of SUSYQM for periodic potentials. In particular, Dunne and Feinberg[5] defined and developed the concept of ''self-isospectral'' periodic potentials in detail. A one-dimensional potential $V_-(x)$ of period L is said to be self-isospectral if its supersymmetric partner potential $V_+(x)$ is just the original potential up to a discrete transformation—a translation by any constant amount, a reflection, or both. A common example is translation by half a period, in which case the condition for self-isospectrality is

$$V_+(x)=V_-(x-L/2). \tag{3}$$

It is easily checked that if the superpotential W satisfies

$$W(x)=-W(x-L/2), \tag{4}$$

then condition (3) immediately follows. In this sense, any self-isospectral potential is rather uninteresting, since application of the SUSYQM formalism[4] to it just yields a discrete transformation and basically nothing new. We have recently pointed out[9] that the Lamé potentials given in Eq. (2) are not self-isospectral for $a\geqslant 2$, and hence SUSYQM generates new exactly solvable periodic problems. This point is further developed in detail in this paper.

We expand our discussion to the band edges and wave functions of a much richer class of periodic potentials given by

$$V(x)=pm\,\mathrm{sn}^2(x)+qm\frac{\mathrm{cn}^2(x)}{\mathrm{dn}^2(x)},\quad p\equiv a(a+1),\quad q\equiv b(b+1), \tag{5}$$

where, like $\mathrm{sn}(x)$, the Jacobi elliptic functions $\mathrm{cn}(x)$ and $\mathrm{dn}(x)$ also have a modulus parameter m which, for notational convenience, is not explicitly displayed. The potentials of Eq. (5) are called associated Lamé potentials, since the corresponding Schrödinger equation is called the associated Lamé equation.[3] More precisely, we often refer to the associated Lamé potential of Eq. (5) as the (p,q) potential and note that $(p,0)$ potentials are just the ordinary Lamé potentials. Although some results for (p,q) potentials are available in scattered form in the mathematical literature, many of our results are new. In particular, we obtain all band edge energies and wave functions for the special case $p=q=a(a+1)$ for $a=1,2,3,...$. We study many (p,q) potentials and check whether they are self-isospectral by constructing and examining the supersymmetric partner potentials. In most cases, $V_-(x)$ is not self-isospectral, and consequently $V_+(x)$ is a new, exactly, or quasiexactly solvable periodic potential.

The associated Lamé potentials given by Eq. (5) can also be rewritten in the alternative form

$$V(x)=pm\,\mathrm{sn}^2(x)+qm\,\mathrm{sn}^2(x+K(m)), \tag{6}$$

since[1]

$$\mathrm{sn}(x+K)=\mathrm{cn}(x)/\mathrm{dn}(x),\quad \mathrm{cn}(x+K)=-\sqrt{1-m}\,\mathrm{sn}(x)/\mathrm{dn}(x),\quad \mathrm{dn}(x+K)=\sqrt{1-m}/\mathrm{dn}(x).$$

It is clear from (6) that potentials (p,q) and (q,p) have the same energy spectra with wave functions shifted by $K(m)$. Therefore, it is sufficient to restrict our attention to $p \geqslant q$.

Before actually solving the Schrödinger equation for the associated Lamé potential (5), let us make a few general comments. Throughout this paper, we have chosen units with $\hbar = 1$, and taken the particle mass in the Schrödinger equation to be 1/2. Note that in the limit when the elliptic modulus parameter $m = 0$, the potential vanishes and one has a rigid rotator problem of period $2K(0) = \pi$, whose energy eigenvalues are at $E = 0,1,4,9,...$, with all the nonzero values being twofold degenerate. On the other hand, the limit $m \rightarrow 1$ is much trickier since $K(m)$ tends to infinity and the periodic nature of the potential is obscured. The Schrödinger equation for finding the eigenstates for an arbitrary periodic potential is called Hill's equation in the mathematics literature.[3] A general property of Hill's equation is the oscillation theorem which states that for a potential with period L, the band edge wave functions arranged in order of increasing energy $E_0 \leqslant E_1 \leqslant E_2 \leqslant E_3 \leqslant E_4 \leqslant E_5 \leqslant E_6 \leqslant \cdots$ are of period $L,2L,2L,L,L,2L,2L,\ldots$. The corresponding number of wave function nodes in the interval L are 0,1,1,2,2,3,3,..., and the energy band gaps are given by $\Delta_1 \equiv E_2 - E_1$, $\Delta_2 \equiv E_4 - E_3$, $\Delta_3 \equiv E_6 - E_5, \ldots$. We shall see that the expected $m = 0$ limit and the oscillation theorem are very useful in identifying if all band edge eigenstates have been properly determined or if some have been missed.

The plan of the paper is as follows. In Sec. II, we briefly review the basic ideas of SUSYQM. A detailed discussion of Lamé potentials and their supersymmetric partners is given in Sec. III. Solutions of the Schrödinger equation for the associated Lamé potentials are presented in Sec. IV. Many key new results are summarized in Table III. It is shown that the locus of quasiexactly solvable problems[10,11] in the (p,q) plane are parabolas about the line $p = q$. Our solutions are valid for any real choice of the parameters a, b [recall $p = a(a+1)$, $q = b(b+1)$]. Integer and half-integer values of a, b, including the very interesting special case $a = b =$ integer, are treated in detail in Sec. V. In most cases, the application of SUSYQM gives new solvable periodic potentials, many of which are illustrated in the figures. Finally, Sec. VI contains some concluding remarks.

II. SUPERSYMMETRIC QUANTUM MECHANICS FORMALISM

The supersymmetric partner potentials $V_{\pm}(x)$ are defined in terms of the superpotential $W(x)$ by

$$V_{\pm}(x) = W^2(x) \pm W'(x). \tag{7}$$

The corresponding Hamiltonians $H_{\pm}$ can be factorized as

$$H_{-} = A^{+}A, \quad H_{+} = AA^{+}, \tag{8}$$

where

$$A = \frac{d}{dx} + W(x), \quad A^{+} = -\frac{d}{dx} + W(x), \tag{9}$$

so that the spectra of $H_{\pm}$ are non-negative. It is also clear that on the full line, both $H_{\pm}$ cannot have zero energy modes since both $\psi_0^{(\pm)}$ given by

$$\psi_0^{(\pm)}(x) = \exp\left(\pm \int^x W(y)\, dy \right), \tag{10}$$

cannot be simultaneously normalized.

On the other hand, when the superpotential $W(x)$ is periodic $[W(x+L) = W(x)]$ then the potentials $V_{-}(x)$ and $V_{+}(x)$ are isospectral—their spectra match completely, including the zero modes, and SUSY is unbroken provided

TABLE I. The eigenvalues and eigenfunctions for the five band edges coresponding to the $a=2$ Lamé potential V_- which gives $(p,q)=(6,0)$ and its SUSY partner V_+. Here $B\equiv 1+m+\delta$ and $\delta\equiv\sqrt{1-m+m^2}$. The potentials $V_\pm$ have period $L=2K(m)$ and their analytic forms are given by Eqs. (14) and (17), respectively. The periods of various eigenfunctions and the number of nodes in the interval L are tabulated.

E	$\psi^{(-)}$	$(B-3m\,\mathrm{sn}^2(x))\psi^{(+)}$	Period	Nodes
0	$m+1+\delta-3m\,\mathrm{sn}^2(x)$	1	$2K$	0
$2\delta-1-m$	$\mathrm{cn}(x)\mathrm{dn}(x)$	$\mathrm{sn}(x)[6m-(m+1)B+m\,\mathrm{sn}^2(x)(2B-3-3m)]$	$4K$	1
$2\delta-1+2m$	$\mathrm{sn}(x)\mathrm{dn}(x)$	$\mathrm{cn}(x)[B+m\,\mathrm{sn}^2(x)(3-2B)]$	$4K$	1
$2\delta+2-m$	$\mathrm{sn}(x)\mathrm{cn}(x)$	$\mathrm{dn}(x)[B+\mathrm{sn}^2(x)(3m-2B)]$	$2K$	2
4δ	$m+1-\delta-3m\,\mathrm{sn}^2(x)$	$\mathrm{sn}(x)\mathrm{cn}(x)\mathrm{dn}(x)$	$2K$	2

$$\int_0^L W(y)\,dy=0. \tag{11}$$

It is worth noting that in this case both $\psi_0^{(\pm)}$ belong to the Hilbert space. Thus in this case even though SUSY is unbroken, the Witten index is zero.[5] The condition (11) is trivially satisfied in case $W(x)$ is an odd function of x and throughout this paper we shall only consider superpotentials W which are odd function of x. Further, using the known eigenfunctions $\psi_n^{(-)}(x)$ of $V_-(x)$ one can immediately write down the corresponding un-normalized eigenfunctions $\psi_n^{(+)}(x)$ of $V_+(x)$. In particular, from Eq. (10) it follows that the ground state of $V_+(x)$ is given by

$$\psi_0^{(+)}(x)=\frac{1}{\psi_0^{(-)}(x)}, \tag{12}$$

while the excited states $\psi_n^{(+)}(x)$ are obtained from $\psi_n^{(-)}(x)$ by using the relation

$$\psi_n^{(+)}(x)=\left[\frac{d}{dx}+W(x)\right]\psi_n^{(-)}(x) \quad (n\geqslant 1). \tag{13}$$

Thus by starting from an exactly solvable periodic potential $V_-(x)$, one gets another isospectral periodic potential $V_+(x)$. As emphasized previously, if $V_-(x)$ is not self-isospectral, then $V_+(x)$ is a new solvable periodic potential!

III. LAMÉ POTENTIALS $(p,0)$ AND THEIR SUPERSYMMETRIC PARTNERS

The supersymmetric quantum mechanics formalism of Sec. II will now be applied to the Lamé potentials $ma(a+1)\mathrm{sn}^2(x,m)$. Analytic solutions are known for integer values of a,[2] and the supersymmetric partner potentials can be readily computed. We first discuss the results for small integer values of a, and then present some eigenstate results for arbitrary integer values of a.

A. Lamé potentials with $a=1,2,3$

1. $a=1$

The $a=1$ Lamé potential $V_-=2m\,\mathrm{sn}^2(x)-m$ is known to be self-isospectral[5] since its SUSY partner satisfies $V_+(x)=V_-(x-K(m))$. Both $V_+(x)$ and $V_-(x)$ have one energy band ranging from energy 0 to energy $1-m$, with a continuum starting at energy 1.[2] Note that at $m=0$ one has energy eigenvalues at 0, 1 as expected for a rigid rotator and as $m\to 1$, one gets $V_-(x)\to 1-2\,\mathrm{sech}^2 x$, the bandwidth $1-m$ vanishes as expected, and one has an energy level at $E=0$.

2. $a=2$

For the $a=2$ case, the Lamé potential (2) has two bound bands and a continuum band. The energies and wave functions of the five band edges are well known.[2,3] The lowest energy band ranges from $2+2m-2\delta$ to $1+m$, the second energy band ranges from $1+4m$ to $4+m$, and the continuum starts at energy $2+2m+2\delta$, where $\delta=\sqrt{1-m+m^2}$. The wave functions of all the band edges are given in Table I. Note that in the interval $2K(m)$ corresponding to the period of the Lamé potential, the number of nodes increases with energy. In order to use the SUSYQM formalism, we must shift the Lamé potential by a constant to ensure that the ground state, i.e., the lower edge of the lowest band, has energy $E=0$. As a result, the potential

$$V_-(x)=-2-2m+2\delta+6m\,\mathrm{sn}^2(x) \tag{14}$$

has its ground state energy at zero with a corresponding un-normalized wave function[2]

$$\psi_0^{(-)}(x)=1+m+\delta-3m\,\mathrm{sn}^2(x). \tag{15}$$

The corresponding superpotential is

$$W=-\frac{d}{dx}\log\psi_0^{(-)}(x)=\frac{6m\,\mathrm{sn}(x)\mathrm{cn}(x)\mathrm{dn}(x)}{\psi_0^{(-)}(x)}, \tag{16}$$

and hence the partner potential $V_+(x)$ for the potential $V_-(x)$ given in Eq. (14) is

$$V_+(x)=-V_-(x)+\frac{72m^2\,\mathrm{sn}^2(x)\mathrm{cn}^2(x)\mathrm{dn}^2(x)}{[1+m+\delta-3m\,\mathrm{sn}^2(x)]^2}. \tag{17}$$

Although the SUSYQM formalism guarantees that the potentials $V_\pm$ are isospectral, they are not self-isospectral, since they do not satisfy Eq. (3).[9] Therefore, $V_+(x)$ as given by Eq. (17) is a new periodic potential which is strictly isospectral to the potential (14) and hence it also has two bound bands and a continuum band. In Fig. 1 we have plotted the potentials $V_\pm(x)$ corresponding to $a=2$ for three different values of the parameter m. The values are $m=0.5$, 0.8, 0.998. The difference in shape between $V_-(x)$ and $V_+(x)$ is manifest from the figures, especially for large m. Using Eqs. (12) and (13) and the known eigenstates of $V_-(x)$, we can immediately compute all the band-edge Bloch wave functions for $V_+(x)$. In Table I we have given the energy eigenvalues and wave functions for the isospectral partner potentials $V_\pm(x)$. At $m=0$ one has energy eigenvalues 0, 1, 4 as expected for a rigid rotator. As $m\to 1$, one gets $V_-(x)\to 4-6\,\mathrm{sech}^2 x$, the bandwidths vanish as expected, and one has two energy levels at $E=0$, 3, with a continuum above $E=4$.

3. $a=3$

For the $a=3$ Lamé potential, the ground state wave function is

$$\psi_0^{(-)}(x)=\mathrm{dn}(x)[2m+\delta_1+1-5m\,\mathrm{sn}^2(x)],$$

the corresponding superpotential is[9]

$$W=\frac{m\,\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}(x)}\,\frac{[2m+\delta_1+11-15m\,\mathrm{sn}^2(x)]}{[2m+\delta_1+1-5m\,\mathrm{sn}^2(x)]}, \tag{18}$$

and the partner potentials $V_\pm(x)$ are[9]

$$V_-(x)=-2-5m+2\delta_1+12m\,\mathrm{sn}^2(x),\qquad \delta_1\equiv\sqrt{1-m+4m^2}, \tag{19}$$

and

$$V_+(x)=-V_-(x)+\frac{2m^2\,\mathrm{sn}^2(x)\mathrm{cn}^2(x)}{\mathrm{dn}^2(x)}\,\frac{[2m+\delta_1+11-15m\,\mathrm{sn}^2(x)]^2}{[2m+\delta_1+1-5m\,\mathrm{sn}^2(x)]^2}. \tag{20}$$

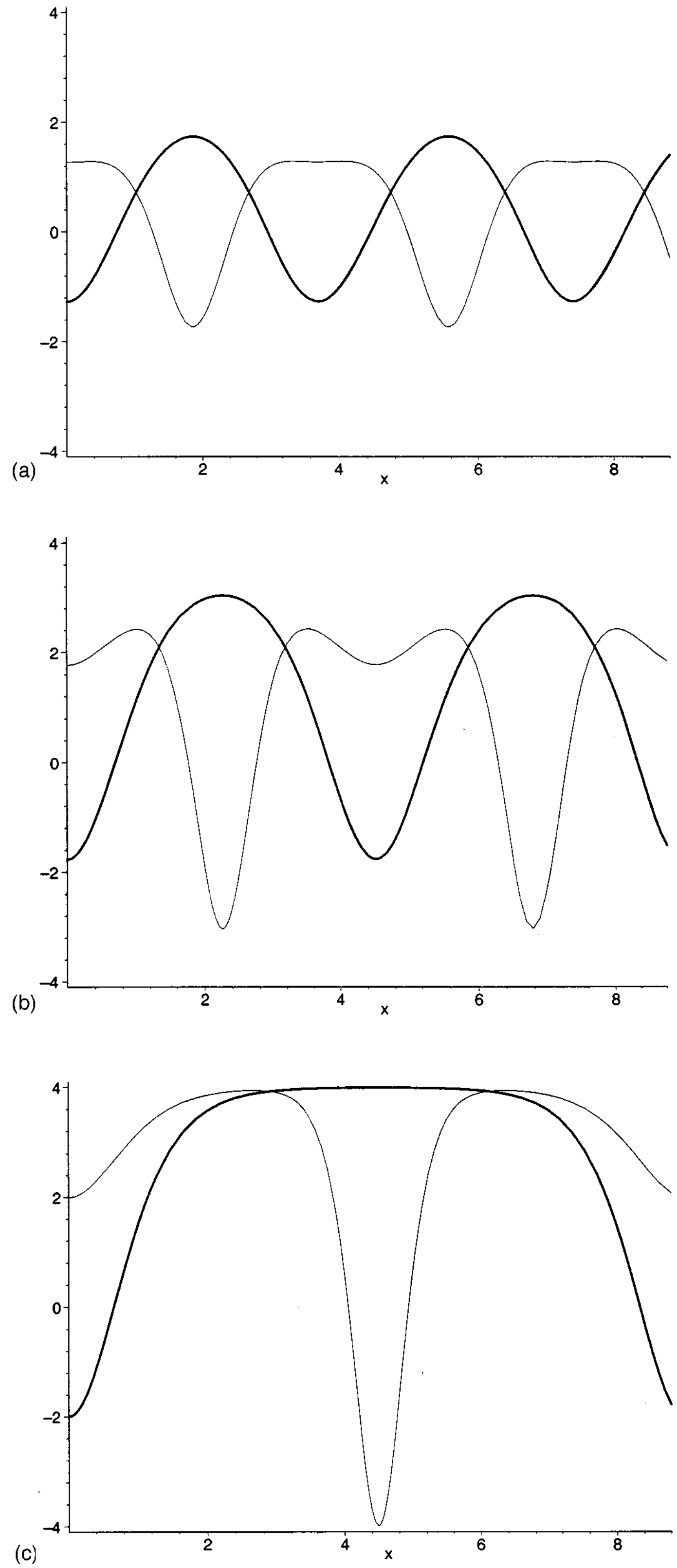

FIG. 1. The (6,0) Lamé potential $V_-(x)$ corresponding to $a=2$ (the thick line) as given by Eq. (14) and its supersymmetric partner potential $V_+(x)$ (the thin line) as given by Eq. (17) for three choices of m: (a) 0.5, (b) 0.8, (c) 0.998.

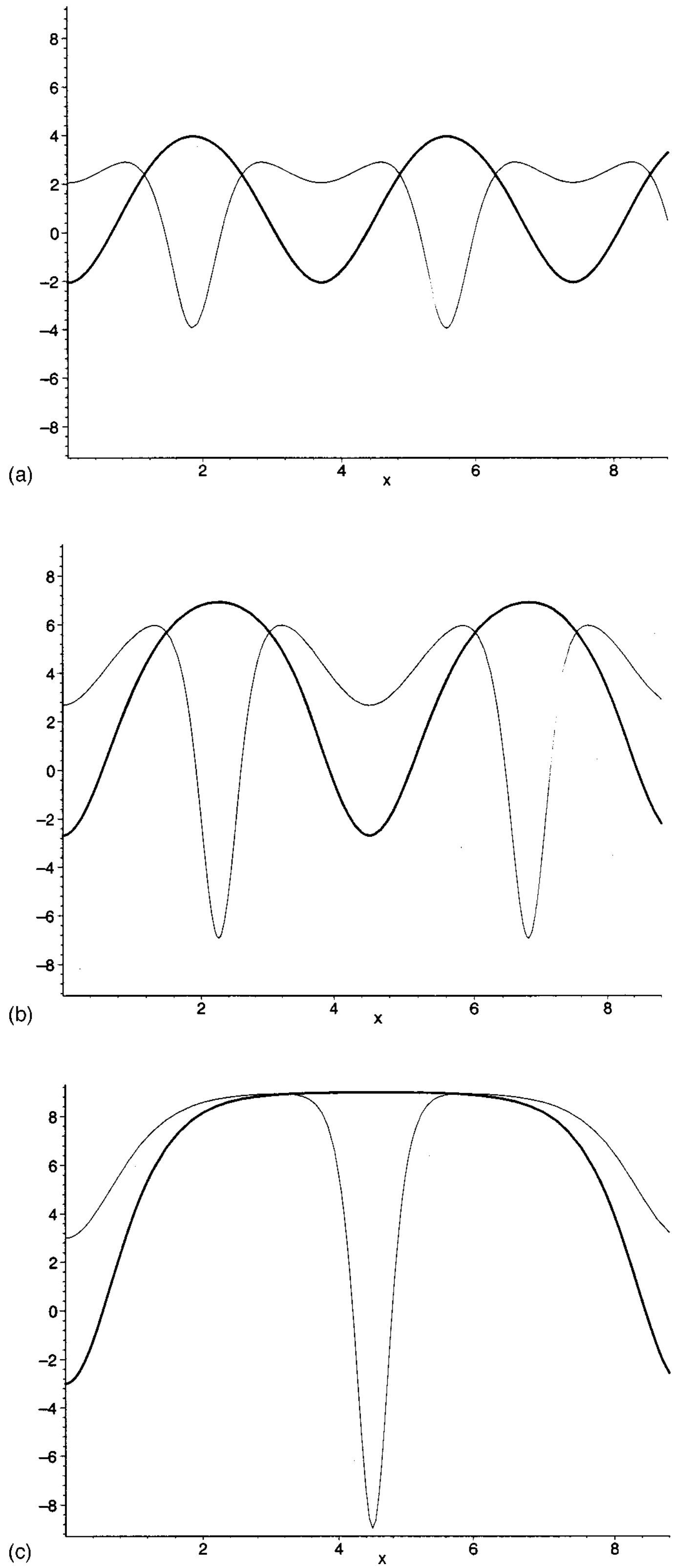

FIG. 2. The (12, 0) Lamé potential $V_{-}(x)$ corresponding to $a=3$ (the thick line) as given by Eq. (19) and its supersymmetric partner potential $V_{+}(x)$ (the thin line) as given by Eq. (20) for three choices of m: (a) 0.5, (b) 0.8, (c) 0.998.

TABLE II. The eigenvalues and eigenfunctions for the seven band edges corresponding to the $a=3$ Lamé potential V_- which gives $(p,q)=(12,0)$ and its SUSY partner V_+. Here $\delta_1 \equiv \sqrt{1-m+4m^2}$; $\delta_2 \equiv \sqrt{4-m+m^2}$; $\delta_3 \equiv \sqrt{4-7m+4m^2}$. The potentials $V_\pm$ have period $L=2K(m)$ and their analytic forms are given by Eqs. (19) and (20), respectively. The periods of various eigenfunctions and the number of nodes in the interval L are tabulated.

E	$\psi^{(-)}$	$\psi_0^- \psi^{(+)}$	Period	Nodes
0	$\mathrm{dn}(x)[1+2m+\delta_1-5m\,\mathrm{sn}^2(x)]$	1	$2K$	0
$3-3m+2\delta_1-2\delta_2$	$\mathrm{cn}(x)[2+m+\delta_2-5m\,\mathrm{sn}^2(x)]$	$10m(1-m+\delta_2-\delta_1)\mathrm{sn}(x)\mathrm{cn}^2(x) \times \mathrm{dn}^2(x)-(1-m)\dfrac{\mathrm{sn}(x)\psi_0^-\psi^-}{\mathrm{cn}(x)\mathrm{dn}(x)}$	$4K$	1
$3+2\delta_1-2\delta_3$	$\mathrm{sn}(x)[2+2m+\delta_3-5m\,\mathrm{sn}^2(x)]$	$10m(1+\delta_3-\delta_1)\mathrm{cn}(x)\mathrm{sn}^2(x)\mathrm{dn}^2(x) -(1-2m\,\mathrm{sn}^2(x))\dfrac{\mathrm{cn}(x)\psi_0^-\psi^-}{\mathrm{sn}(x)\mathrm{dn}(x)}$	$4K$	1
$2-m+2\delta_1$	$\mathrm{sn}(x)\mathrm{cn}(x)\mathrm{dn}(x)$	$\mathrm{dn}^3(x)[1+2m+\delta_1+(m-2-2\delta_1) \times \mathrm{sn}^2(x)]$	$2K$	2
$4\delta_1$	$\mathrm{dn}(x)[1+2m-\delta_1-5m\,\mathrm{sn}^2(x)]$	$\mathrm{sn}(x)\mathrm{cn}(x)\mathrm{dn}^3(x)$	$2K$	2
$3-3m+2\delta_1+2\delta_2$	$\mathrm{cn}(x)[2+m-\delta_2-5m\,\mathrm{sn}^2(x)]$	$10m(1-m-\delta_2-\delta_1)\mathrm{sn}(x)\mathrm{cn}^2(x) \times \mathrm{dn}^2(x)-(1-m)\dfrac{\mathrm{sn}(x)\psi_0^-\psi^-}{\mathrm{cn}(x)\mathrm{dn}(x)}$	$4K$	3
$3+2\delta_1+2\delta_3$	$\mathrm{sn}(x)[2+2m-\delta_3-5m\,\mathrm{sn}^2(x)]$	$10m(1-\delta_3-\delta_1)\mathrm{cn}(x)\mathrm{sn}^2(x) \times \mathrm{dn}^2(x)-(1-2m\,\mathrm{sn}^2(x))\dfrac{\mathrm{cn}(x)\psi_0^-\psi^-}{\mathrm{sn}(x)\mathrm{dn}(x)}$	$4K$	3

Clearly, the potential $V_-(x)$ is not self-isospectral. In fact, $V_-(x)$ and $V_+(x)$ are distinctly different periodic potentials which have the same seven band edges corresponding to three bound bands and a continuum band.[2] In Fig. 2 we have plotted the potentials $V_\pm(x)$ corresponding to $a=3$ for several different values of the parameter m. The values of m are 0.5, 0.8, 0.998. It is clear from Fig. 2 that the potentials $V_+(x)$ and $V_-(x)$ have different shapes and are far from being self-isospectral. Using Eqs. (12) and (13) and the known eigenstates of $V_-(x)$, we can immediately compute all seven band edges corresponding to the known three bound bands and a continuum band.[2,3] For example, the ground state $\psi_0^{(+)}$ is given by

$$\psi_0^{(+)}(x)=\frac{1}{\psi_0^{(-)}(x)}=\frac{1}{\mathrm{dn}(x)[1+2m+\delta_1-5m\,\mathrm{sn}^2(x)]}. \tag{21}$$

The wave functions for the remaining six states are similarly written down by using Eq. (13). These are shown in Table II. The band edge energies for the $a=3$ Lamé potential (12,0) as a function of the elliptic modulus parameter m are plotted in Fig. 3. Note that at $m=0$ one has energy eigenvalues at 0,1,4,9 as expected for a rigid rotator and as $m\to 1$, one gets $V_-(x)\to 9 - 12\,\mathrm{sech}^2 x$, the bandwidths vanish as expected, and one has three energy levels at $E=0$, 5, 8 with a continuum above $E=9$.

B. Results for general integer values of *a*

The extension to higher values of a is straightforward. It is possible to make several general comments about the form of the band edge wave functions for the partner potentials $V_+(x)$. This is most conveniently done by separately discussing the cases of even and odd values of a.

1. a=even integer

For a even, say $a=2N$, it is known[2] that there are $N+1$ solutions of the form $F_N(\mathrm{sn}^2 x)$, and N solutions each of the three forms

$$\mathrm{sn}\,x\,\mathrm{cn}\,x\,F_{N-1}(\mathrm{sn}^2 x), \quad \mathrm{sn}\,x\,\mathrm{dn}\,x\,F_{N-1}(\mathrm{sn}^2 x), \quad \mathrm{cn}\,x\,\mathrm{dn}\,x\,F_{N-1}(\mathrm{sn}^2 x).$$

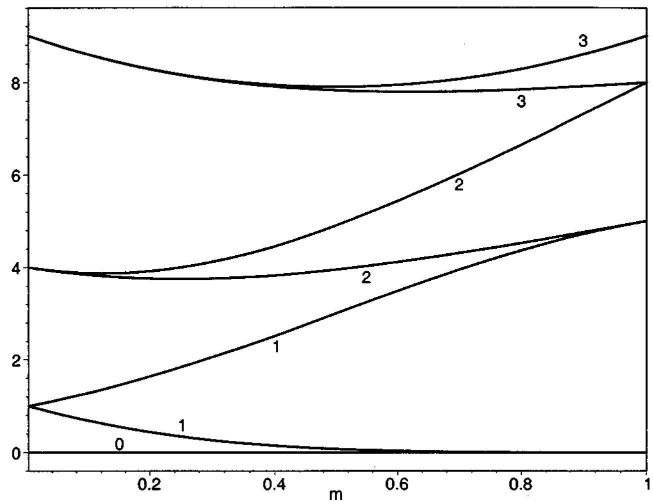

FIG. 3. Band edge energies for the (12,0) Lamé potential corresponding to $a=3$ as a function of the elliptic modulus parameter m. This figure is drawn using the eigenvalues given in Table II. The band edges are labeled by the number of wave function nodes in the interval $2K(m)$.

Here F_r denotes a polynomial of degree r in its argument. The ground state $\psi_0^-(x)$ (which is the lower edge of the lowest band) is of the form $F_N(\mathrm{sn}^2 x)$. It is easily checked using Eq. (13) that the corresponding partner potential $V_+(x)$ has N solutions each of the four forms

$$\frac{\mathrm{dn}\,x\,G_N(\mathrm{sn}^2 x)}{\psi_0^-(x)},\quad \frac{\mathrm{sn}\,x\,G_N(\mathrm{sn}^2 x)}{\psi_0^-(x)},\quad \frac{\mathrm{cn}\,x\,G_N(\mathrm{sn}^2 x)}{\psi_0^-(x)},\quad \frac{\mathrm{sn}\,x\,\mathrm{cn}\,x\,\mathrm{dn}\,x\,G_{N-1}(\mathrm{sn}^2 x)}{\psi_0^-(x)},$$

while the ground state is given by $\psi_0^+(x)=1/\psi_0^-(x)$.

2. a=odd integer

For a odd, say $a=2N+1$, it is known[2] that the Lamé potentials have $N+1$ solutions each of the three forms

$$\mathrm{sn}\,x\,F_N(\mathrm{sn}^2 x),\quad \mathrm{cn}\,x\,F_N(\mathrm{sn}^2 x),\quad \mathrm{dn}\,x\,F_N(\mathrm{sn}^2 x)$$

and N solutions of the form

$$\mathrm{sn}\,x\,\mathrm{cn}\,x\,\mathrm{dn}\,x\,F_{N-1}(\mathrm{sn}^2 x).$$

The ground state $\psi_0^-(x)$ is of the form $\mathrm{dn}\,x\,F_N(\mathrm{sn}^2 x)$. We can then easily deduce that the corresponding partner potentials $V_+(x)$ will have $N+1$ solutions each of the two forms

$$\frac{\mathrm{sn}\,x\,G_{N+1}(\mathrm{sn}^2 x)}{\psi_0^-(x)},\quad \frac{\mathrm{cn}\,x\,G_{N+1}(\mathrm{sn}^2 x)}{\psi_0^-(x)},$$

and N solutions each of the two forms

$$\frac{\mathrm{dn}\,x\,G_{N+1}(\mathrm{sn}^2 x)}{\psi_0^-(x)},\quad \frac{\mathrm{sn}\,x\,\mathrm{cn}\,x\,\mathrm{dn}\,x\,G_N(\mathrm{sn}^2 x)}{\psi_0^-(x)},$$

while as usual, the ground state is given by $\psi_0^+(x)=1/\psi_0^-(x)$.

In summary, for integral a, Lamé potentials with $a \geqslant 2$ are not self-isospectral. They have distinct supersymmetric partner potentials even though both potentials have the same $(2a+1)$ band edge eigenvalues.

IV. ASSOCIATED LAMÉ POTENTIALS (p,q) AND THEIR SUPERSYMMETRIC PARTNERS

In contrast to the Lamé potentials discussed previously, there seems to be no systematic treatment of associated Lamé potentials in the literature. Therefore, we will first devote some time to discussing the properties of associated Lamé potentials, show that they are quasiexactly solvable, and then proceed to construct and study their isospectral supersymmetric partner potentials.

A. Description of associated Lamé potentials

As mentioned before, we will refer to the associated Lamé potentials given by Eq. (5) or equivalently Eq. (6) as the (p,q) potential. The special cases $q=0$, as well as $p=0$, correspond to ordinary Lamé potentials.

In general, for any value of p and q, the associated Lamé potentials have a period $2K(m)$ since

$$\mathrm{sn}(x+2K) = -\mathrm{sn}(x), \quad \mathrm{cn}(x+2K) = -\mathrm{cn}(x), \quad \mathrm{dn}(x+2K) = \mathrm{dn}(x).$$

However, for the special case $p=q$, Eq. (6) shows that the period is $K(m)$. From a physical viewpoint, if one thinks of a Lamé potential $(p,0)$ as due to a one-dimensional regular array of atoms with spacing $2K(m)$, and ''strength'' p, then the associated Lamé potential (p,q) results from two alternating types of atoms spaced by $K(m)$ with ''strengths'' p and q, respectively. If the two types of atoms are identical [which makes $p=q$], one expects a potential of period $K(m)$.

Extrema (defined for this discussion as either local or global maxima and minima) of associated Lamé potentials are easily found by setting $dV(x)/dx=0$. This gives

$$\mathrm{sn}(x)\mathrm{cn}(x)[p\,\mathrm{dn}^4(x) - q(1-m)] = 0.$$

Extrema occur when (i) $\mathrm{sn}(x)=0$, that is $x=0, \pm 2K(m), \pm 4K(m), \ldots$; (ii) $\mathrm{cn}(x)=0$, that is $x=\pm K(m), \pm 3K(m), \ldots$; (iii) $\mathrm{dn}^4(x)=(1-m)q/p$. At the points specified by (i) and (ii), one always has extrema and $V(x)$ has values pm and qm. In addition, since $\mathrm{dn}^4(x)$ has a minimum value $(1-m)^2$ and a maximum value unity,[1] condition (iii) also yields extrema provided

$$(1-m)^2 \leqslant (1-m)q/p \leqslant 1.$$

For given fixed values of q and m, this condition has a solution provided p lies in the critical range

$$q(1-m) \leqslant p \leqslant q/(1-m).$$

Alternatively, for given fixed values of p and q with $p \geqslant q$, condition (iii) has a solution provided m is greater than the critical value $1-q/p$.

The associated Lamé potentials for $q=2$, $m=0.5$, and several values of p are plotted in Fig. 4(a). In the critical range of p values $1 \leqslant p \leqslant 4$, one expects extrema coming from condition (iii), and these are clearly seen in Fig. 4(a). In general the period is $2K(0.5)=3.708$, but for $p=q=2$, the period $K(0.5)$ is evident. Note that as p increases, any given extremum changes character. For example, at $x=0$, as p increases, one goes from a maximum to a local minimum to an absolute minimum. In Fig. 4(b) we have plotted associated Lamé potentials for $p=4$, $q=2$, and several values of m. As expected from the previous discussion, one always sees extrema at the points specified by conditions (i) and (ii), and additional extrema coming from condition (iii) are evident for $m \geqslant 1/2$.

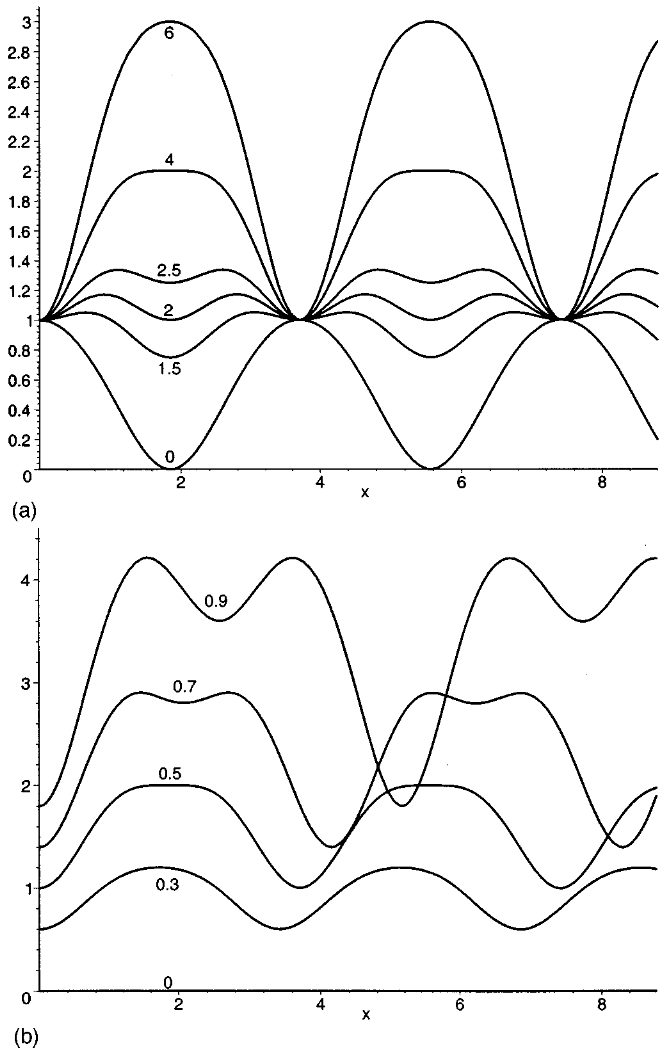

FIG. 4. (a) Plots of the (p,q) associated Lamé potentials for $q=2$, $m=0.5$, and several values of p. The curves are labeled by the value of p. (b) Plots of the (p,q) associated Lamé potentials for $p=4$, $q=2$, and several values of m. The curves are labeled by the value of m.

B. Solutions of the associated Lamé equation: Parabolas of solvability

The associated Lamé equation is just the Schrödinger equation for the potential in Eq. (5),

$$-\frac{d^2\psi}{dx^2}+\left[pm\,\mathrm{sn}^2(x)+qm\,\frac{\mathrm{cn}^2(x)}{\mathrm{dn}^2(x)}-E\right]\psi=0. \tag{22}$$

On substituting

$$\psi(x)=[\mathrm{dn}(x)]^{-b}y(x), \tag{23}$$

it is easily shown that $y(x)$ satisfies the Hermite elliptic equation[3]

$$y''(x)+2bm\,\frac{\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}(x)}\,y'(x)+[\lambda-(a+1-b)(a+b)m\,\mathrm{sn}^2(x)]y(x)=0, \tag{24}$$

where

TABLE III. Eigenvalues and eigenfunctions for various associated Lamé potentials (p,q) with $p=a(a+1)$ and $q=(a-n+1)(a-n)$ for $n=1,2,3,\ldots$. The periods of various eigenfunctions and the number of nodes in the interval $2K(m)$ are tabulated. Here $\delta_4 \equiv \sqrt{1-m+m^2(a-1)^2}$; $\delta_5 \equiv \sqrt{4-7m+2ma+m^2(a-2)^2}$; $\delta_6 \equiv \sqrt{4-m-2ma+m^2(a-1)^2}$; $\delta_7 \equiv \sqrt{9-9m+m^2(a-2)^2}$.

q	E	$\mathrm{dn}^{-a}(x)\psi$	Period	Nodes
$a(a-1)$	ma^2	1	$2K$	0
$(a-1)(a-2)$	$1+m(a-1)^2$	$\dfrac{\mathrm{cn}(x)}{\mathrm{dn}(x)}$	$4K$	1
$(a-1)(a-2)$	$1+ma^2$	$\dfrac{\mathrm{sn}(x)}{\mathrm{dn}(x)}$	$4K$	1
$(a-2)(a-3)$	$2+m(a^2-2a+2)\pm 2\delta_4$	$\dfrac{[m(2a-1)\mathrm{sn}^2(x)-1+m-ma\pm\delta_4]}{\mathrm{dn}^2(x)}$	$2K$	2,0
$(a-2)(a-3)$	$4+m(a-1)^2$	$\dfrac{\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}^2(x)}$	$2K$	2
$(a-3)(a-4)$	$5+m(a^2-4a+5)\pm 2\delta_5$	$\dfrac{\mathrm{cn}(x)[m(2a-1)\mathrm{sn}^2(x)-2+2m-ma\pm\delta_5]}{\mathrm{dn}^3(x)}$	$4K$	3,1
$(a-3)(a-4)$	$5+m(a^2-2a+2)\pm 2\delta_6$	$\dfrac{\mathrm{sn}(x)[m(2a-1)\mathrm{sn}^2(x)-2+m-ma\pm\delta_6]}{\mathrm{dn}^3(x)}$	$4K$	3,1
$(a-4)(a-5)$	$10+m(a^2-4a+5)\pm 2\delta_7$	$\dfrac{\mathrm{sn}(x)\mathrm{cn}(x)[m(2a-1)\mathrm{sn}^2(x)-3+2m-ma\pm\delta_7]}{\mathrm{dn}^4(x)}$	$2K$	4,2

$$p=a(a+1), \quad q=b(b+1), \quad E=\lambda+mb^2. \tag{25}$$

On further substituting $\mathrm{sn}(x)=\sin t$, $y(x)\equiv z(t)$, one obtains Ince's equation,

$$(1-m\sin^2 t)z''(t)+(2b-1)m\sin t\cos t\, z'(t)+[\lambda-(a+1-b)(a+b)m\sin^2 t]z(t)=0, \tag{26}$$

which is a well-known quasiexactly solvable (QES) equation.[3] In particular, on substituting

$$\cos t=u, \quad z(t)\equiv w(u)=\sum_{n=0}^{\infty}\frac{u^n R_n}{n!}, \tag{27}$$

it is easily shown that R_n satisfies a three-term recursion relation. In particular, if $a+b+1=n$ $(n=1,2,3,\ldots)$, then one obtains n QES solutions. Actually n QES solutions are also obtained in case $b-a=-n(n=1,2,3,\ldots)$, but since q is unchanged under $b\to-b-1$, no really new solutions are obtained in this case. The QES solutions for $n=1,2,3,4,5$ are given in Table III. In particular, for any given choice of $p=a(a+1)$, Table III lists the eigenstates of the associated Lamé equation for various values of q.

For $q=a(a-1)$, there is just one eigenstate with energy ma^2 and wave function $\psi=\mathrm{dn}^a(x)$. Since the wave function has period $2K(m)$ and is nodeless, this is clearly the ground state wave function of the $(a(a+1),a(a-1))$ potential for any real choice of the parameter a. The equations $p=a(a+1)$ and $q=a(a-1)$ are the parametric forms of the equation of the parabola $(p-q)^2=2(p+q)$, which is plotted in Fig. 5 and labeled $P1$. For any point on the parabola, one knows the ground state wave function and energy $E_0=ma^2$. The parabola $P1$ includes the points (2,0) and (6,2).

For $q=(a-1)(a-2)$, we see from Table III that two eigenstates with energies $1+m(a-1)^2$ and $1+ma^2$ are known. Since they have period $4K(m)$ and just one node in the interval $L=2K(m)$, they must correspond to the first and second band edge energies E_1 and E_2 of the $(a(a+1),(a-1)(a-2))$ potential. Eliminating a from the equations $p=a(a+1)$ and $q=(a-1)(a-2)$ gives the ''parabola of solvability'' $(p-q)^2=8(p+q)-12$, which is plotted in Fig. 5 and labeled $P2$. This parabola includes the points (2,0) and (6,0), which correspond to Lamé

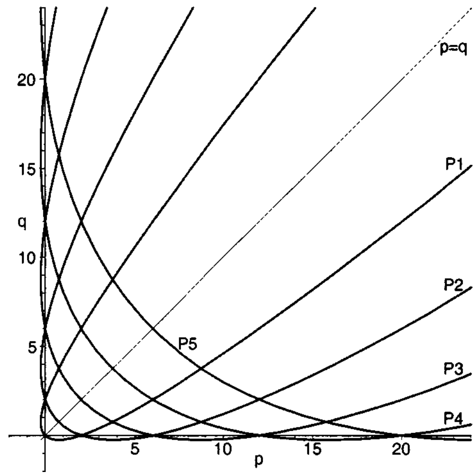

FIG. 5. Parabolas of solvability. All associated Lamé potentials (p,q) which are quasisolvable are illustrated. Each parabola corresponds to a choice of q in Table III. Parabola Pn is for $p=a(a+1)$, $q=(a-n+1)(a-n)$ for $n=1,2,3,...$, and one knows n eigenstates for any point on it from Table III.

potentials. Similarly, the parabolas of solvability Pn $(n=0,1,2,...)$ corresponding to $q=(a-n+1)(a-n)$ in Table III are plotted. n eigenstates are known for any point on the parabola of solvability Pn.

C. Supersymmetric partner potentials

It is easily checked from Table III that the solution corresponding to $q=a(a-1)$ as well as one of the $q=(a-2)(a-3)$ solutions are nodeless and correspond to the ground state. Hence, for these cases, one can obtain the superpotential and hence the partner potential V_+ and enquire if V_- is self-isospectral. For example, consider the case of $p=a(a+1), q=a(a-1)$ in which case W is given by

$$W \equiv -\frac{\psi_0'(x)}{\psi_0(x)} = am\,\frac{\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}(x)}, \tag{28}$$

so that the corresponding partner potentials are

$$\begin{aligned} V_- &= (a-1)am\,\frac{\mathrm{cn}^2(x)}{\mathrm{dn}^2(x)} + ma(a+1)\mathrm{sn}^2(x) - ma^2, \\ V_+ &= a(a+1)m\,\frac{\mathrm{cn}^2(x)}{\mathrm{dn}^2(x)} + m(a-1)a\,\mathrm{sn}^2(x) - ma^2. \end{aligned} \tag{29}$$

It is easily seen that these partner potentials satisfy Eq. (3), are consequently self-isospectral, and SUSY gives nothing new in this case. It is amusing to note that the superpotential W obtained here was in fact discussed in Ref. 5 [see their Eq. (32)].

Let us now consider the SUSY partner potential computed from the ground state for the $p=a(a+1), q=(a-2)(a-3)$ case. It is given by (see Table III)

$$\psi_0(x) = [m(a-1) - 1 - \delta_1 + m(2a-1)\mathrm{sn}^2(x)](\mathrm{dn}(x))^{a-2}, \tag{30}$$

where $\delta_1 = \sqrt{1-m+m^2(a-1)^2}$. The corresponding superpotential W turns out to be

TABLE IV. The five eigenvalues and eigenfunctions for the self-isospectral associated Lamé potential corresponding to $a=2$, $b=1$ which gives $(p,q)=(6,2)$. The potential is $V_-(x)=6m\,\mathrm{sn}^2(x)+2m\,\mathrm{cn}^2(x)/\mathrm{dn}^2(x)-4m$, and has period $2K(m)$. The number of nodes in the interval $2K(m)$ is tabulated.

E	$\psi^{(-)}$	Period	Nodes
0	$\mathrm{dn}^2(x)$	$2K$	0
$5-3m-2\sqrt{4-3m}$	$\frac{\mathrm{cn}(x)}{\mathrm{dn}(x)}[3m\,\mathrm{sn}^2(x)-2-\sqrt{4-3m}]$	$4K$	1
$5-2m-2\sqrt{4-5m+m^2}$	$\frac{\mathrm{sn}(x)}{\mathrm{dn}(x)}[3m\,\mathrm{sn}^2(x)-2-m-\sqrt{4-5m+m^2}]$	$4K$	1
$5-2m+2\sqrt{4-5m+m^2}$	$\frac{\mathrm{sn}(x)}{\mathrm{dn}(x)}[3m\,\mathrm{sn}^2(x)-2-m+\sqrt{4-5m+m^2}]$	$4K$	3
$5-3m+2\sqrt{4-3m}$	$\frac{\mathrm{cn}(x)}{\mathrm{dn}(x)}[3m\,\mathrm{sn}^2(x)-2+\sqrt{4-3m}]$	$4K$	3

$$W=\frac{m(a-2)\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}(x)}-\frac{2m(2a-1)\mathrm{sn}(x)\mathrm{cn}(x)\mathrm{dn}(x)}{[m(1-a)-1-\delta_1+m(2a-1)\mathrm{sn}^2(x)]}. \tag{31}$$

Hence the corresponding partner potentials are

$$V_-(x)=ma(a+1)\mathrm{sn}^2(x)+m(a-3)(a-2)\frac{\mathrm{cn}^2(x)}{\mathrm{dn}^2(x)}-2-m(a^2-2a+2)+2\delta_1, \tag{32}$$

$$V_+(x)=-V_-(x)+2W^2(x). \tag{33}$$

It is easily checked that these potentials are not self-isospectral since they do not satisfy the condition (3). Thus one has discovered a whole class of new elliptic periodic potentials $V_+(x)$ as given by Eq. (33) for which three states are analytically known no matter what a is. In particular, the energy eigenfunctions for V_+ of these three states are easily obtained by using the corresponding energy eigenstates of V_- as given in Table III and using Eqs. (12) and (13).

V. ASSOCIATED LAMÉ POTENTIALS WITH SPECIAL VALUES OF *p* AND *q*

We shall now discuss associated Lamé potentials $(a(a+1),b(b+1))$, where a and b are either both positive integers or half-integers. These values are of special interest since they correspond to intersections of two parabolas of solvability. Consequently, one has greater knowledge about the eigenstates, especially if states of period $2K$ are known for one parabola and states of period $4K$ are known for the other. In most cases, we show that although several band edge energies are exactly known from Table III, one usually does not know all the band edge energies, that is one has a quasiexactly solvable problem. However, in the special case of $p=q$ $(a=b=\text{integer})$, we show that all the band edge eigenstates can be obtained and one has an exactly solvable periodic problem.

A. $a,b=$ integer, $a\neq b$

First, let us note that the Lamé potentials $(a(a+1),0)$ are in this category when $a=$ integer and $b=0$. For example, when $a=3$, one has the (12,0) potential. We see from Fig. 5 that two parabolas of solvability pass through the point (12,0). From Table III it follows that three band edges of period $2K(m)$ are obtained from $q=(a-2)(a-3)$ and four band edges of period $4K(m)$ are obtained from $q=(a-3)(a-4)$. Altogether, arranging in order of increasing nodes, one has seven band edges with periods $2K$, $4K$, $4K$, $2K$, $2K$, $4K$, $4K$ with 0, 1, 1, 2, 2, 3, 3 nodes, respectively. There are no missing states, and as discussed in Sec. III A, this gives three bound bands and a continuum band.

TABLE V. The three eigenvalues and eigenfunctions for the associated Lamé potential corresponding to $a=b=1$ which gives $(p,q)=(2,2)$. The potential has period $K(m)$ and the number of nodes in the interval $K(m)$ is tabulated.

E	$\text{dn}(x)\psi^{(-)}$	Period	Nodes
0	$\text{dn}^2(x)+\sqrt{1-m}$	K	0
$4\sqrt{1-m}$	$\text{dn}^2(x)-\sqrt{1-m}$	$2K$	1
$2-m+2\sqrt{1-m}$	$\text{sn}(x)\text{cn}(x)$	$2K$	1

As a second example with $q\neq 0$, consider the (6,2) associated Lamé potential, that is $p=6$, $q=2$. In this case, taking $a=2$, one can get five band edges from Table III—one solution of period $2K$ is obtained from $q=a(a-1)$, while the remaining four solutions of period $4K$ are obtained from $q=(a-3)(a-4)$. The eigenvalues and eigenfunctions are given in Table IV along with the number of nodes in one period $2K$. It is clear that there are two solutions of period $2K$ with two nodes in the interval $2K$ which have to be present but have not been obtained. This is also clear from the $m=0$ limit, since the energies from Table V are 0,1,1,9,9 and the states at 4,4 are missing. Thus, this is a QES problem. Figure 6 illustrates the (6,2) associated Lamé potential and its supersymmetric partner for three choices of m. The self-isospectral nature of the (6,2) potential is evident from Fig. 6—it also follows from Eq. (29) with $a=2$. The band edge energies for the (6,2) associated Lamé potential as a function of the elliptic modulus parameter m are shown in Fig. 7. The two unobtained band edges of period $2K$ will have energies $E=4$ at $m=0$.

Let us now discuss the general associated Lamé potential $(a(a+1),b(b+1))$, with $a>b$. Using Table III, we obtain $(a-b)$ states of period $2K(4K)$ for $q=[a-(a-b)][a-(a-b-1)]$ for $(a-b)$ odd (even), and $(a+b+1)$ states of period $4K(2K)$ for $q=[a-(a+b+1)]\times[a-(a+b)]$ for $(a-b)$ odd (even). It can be established that some states are missing by looking at the node structure as well as the $m=0$ limit. Hence, we again have a QES problem.

B. $a=b=$ integer

Let us now discuss the special case of $p=q=a(a+1)$, $a=1,2,\ldots$. In this case the associated Lamé potential (5) has period K, rather than $2K$. It then follows from the oscillation theorem that with increasing energy, the band edges must have periods $K,2K,2K,K,K,\ldots$ and in the $m=0$ limit the eigenvalues must go to $E=0,4,16,36,\ldots$, with all nonzero eigenvalues being doubly degenerate. It is easy to check from Table V that one case for which we already have exact results is when $p=q=2$. In particular, consider the special case $a=1$, for which $V_-(x)$ of Eq. (32) takes the form

$$V_-(x)=2m\,\text{sn}^2(x)+2m\frac{\text{cn}^2(x)}{\text{dn}^2(x)}-2-m+2\sqrt{1-m}. \tag{34}$$

Using Table V, we can calculate three energy eigenvalues and eigenfunctions of V_- taking $a=1$ in $q=(a-2)(a-3)$. These are given in Table V. Whereas the ground state is of period K, the next two states in Table V indeed have period $2K$. Using $a=1$ in Eqs. (30)–(33), we find that the corresponding SUSY partner potential is

$$V_+(x)=2-m-2\sqrt{1-m}-\frac{8\sqrt{1-m}m^2\,\text{sn}^2(x)\text{cn}^2(x)}{[\text{dn}^2(x)+\sqrt{1-m}]^2}. \tag{35}$$

Are the potentials $V_\pm(x)$ self-isospectral? Using the relations

$$\text{sn}(x+K(m)/2)=(1+\sqrt{1-m})^{1/2}\left[\frac{\sqrt{1-m}\,\text{sn}(x)+\text{cn}(x)\text{dn}(x)}{\text{dn}^2(x)+\sqrt{1-m}}\right], \tag{36}$$

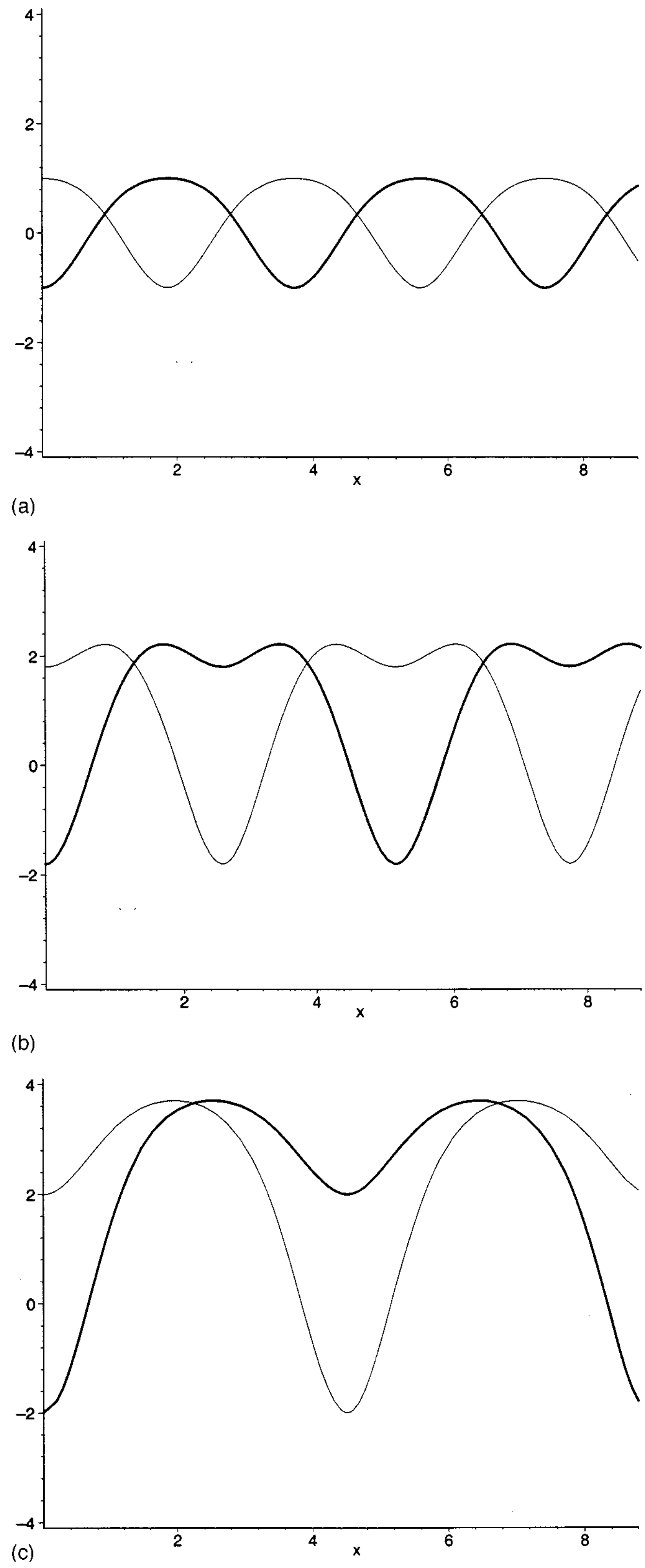

FIG. 6. The (6, 2) associated Lamé potential $V_-(x)$ (the thick line) and its supersymmetric partner potential $V_+(x)$ (the thin line) for three choices of m: (a) 0.5, (b) 0.9, (c) 0.998.

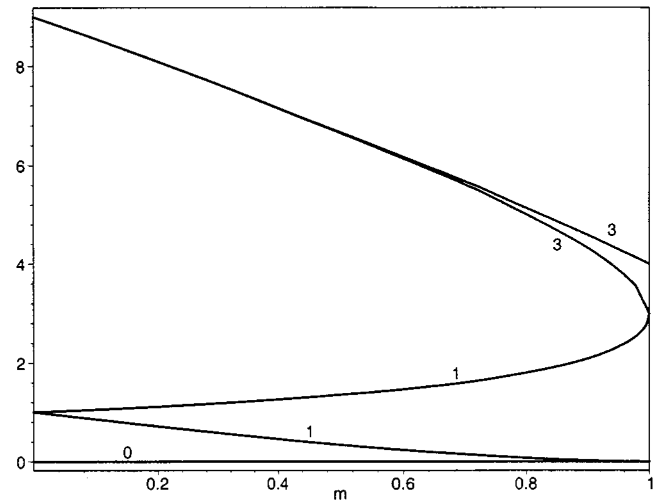

FIG. 7. Band edge energies for the associated Lamé potential (6, 2) as a function of the elliptic modulus parameter m. This figure corresponds to Table IV. The band edges are labeled by the number of wave function nodes in the interval $2K(m)$.

$$\mathrm{cn}(x+K(m)/2)=(1+\sqrt{1-m})^{1/2}(1-m)^{1/4}\left[\frac{(1+\sqrt{1-m})^{1/2}\,\mathrm{cn}(x)-\mathrm{sn}(x)\mathrm{dn}(x)}{\mathrm{dn}^2(x)+\sqrt{1-m}}\right], \quad (37)$$

$$\mathrm{dn}(x+K(m)/2)=(1-m)^{1/4}\left[\frac{(1+\sqrt{1-m})\mathrm{dn}(x)-m\,\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}^2(x)+\sqrt{1-m}}\right], \quad (38)$$

a little algebra reveals that indeed $V_\pm$ are self-isospectral and satisfy Eq. (3).

Are the higher members of the $p=q$ family (i.e., $p=q=6,12,20,...$) also self-isospectral? If our experience with the Lamé case is any guide then we would doubt it. Indeed, we will now show that the (6,6) associated Lamé potential is not self-isospectral. We get five band edges analytically from Table III. In particular, take $a=2$ and consider the case of $q=(a-4)(a-5)$, for which we know two eigenstates as given in Table III. In fact, in this case three more eigenstates can be analytically obtained but the corresponding eigenvalues and eigenfunctions have not been given in Table III since the energy eigenvalues are solutions of a cubic equation whose exact solution for arbitrary a cannot be written in a compact form. However, for $a=2$, we are able to solve the cubic equation and obtain the three eigenvalues in a closed simple form. In particular consider an ansatz of the form

$$y=A+B\,\mathrm{sn}^2 x+D\,\mathrm{sn}^4 x. \quad (39)$$

On substituting this ansatz in Eq. (24) it is easy to show that the energy eigenvalue $\lambda(=E-m(a-4)^2)$ must obey the cubic equation

TABLE VI. The five eigenvalues and eigenfunctions for the associated Lamé potential corresponding to $a=b=2$ which gives $(p,q)=(6,6)$. Here $\delta_8\equiv\sqrt{16-16m+m^2}$. The number of nodes in one period $K(m)$ of the potential is tabulated.

E	$\mathrm{dn}^2(x)\psi^{(-)}$	Period	Nodes
0	$1-(4-m-\delta_8)\mathrm{sn}^2(x)+(4-2m-\delta_8)\mathrm{sn}^4(x)$	K	0
$-4+2m+2\delta_8$	$1-2\,\mathrm{sn}^2(x)+m\,\mathrm{sn}^4(x)$	$2K$	1
$2-m-6\sqrt{1-m}+2\delta_8$	$\mathrm{sn}(x)\mathrm{cn}(x)[1-(1-\sqrt{1-m})\mathrm{sn}^2(x)]$	$2K$	1
$2-m+6\sqrt{1-m}+2\delta_8$	$\mathrm{sn}(x)\mathrm{cn}(x)[1-(1+\sqrt{1-m})\mathrm{sn}^2(x)]$	K	2
$4\delta_8$	$1-(4-m+\delta_8)\mathrm{sn}^2(x)+(4-2m+\delta_8)\mathrm{sn}^4(x)$	K	2

$$\lambda^3+[28m-20-12am]\lambda^2+[64-304m+160ma+32m^2(a-2)(a-3)]\lambda$$
$$-64m(2a-3)(2-2m+ma)=0. \tag{40}$$

The solution of Eq. (40) is in general quite lengthy but in the special case of $a=2$ this cubic equation is easily solved yielding three eigenvalues in a compact form. On combining them with the two levels given in Table III, we obtain the eigenvalues and eigenfunctions of all the five band edges for the case $p=q=6$. These are given in Table VI. We have also verified that these five eigenstates in ascending order of energy indeed have periods K, $2K$, $2K$, K, K, respectively, and that the energy eigenvalues have expected limits at $m=0$. In particular the associated Lamé potential $V_-(x)$ is

$$V_-(x)=6m\,\mathrm{sn}^2(x)+6m\,\frac{\mathrm{cn}^2(x)}{\mathrm{dn}^2(x)}-8-2m+2\delta_8, \tag{41}$$

whose ground state energy is zero while the corresponding eigenfunction ψ_0^- is

$$\psi_0^-(x)=\frac{[1-(4-m-\delta_8)\mathrm{sn}^2(x)+(4-2m-\delta_8)\mathrm{sn}^4(x)]}{\mathrm{dn}^2(x)},\qquad \delta_8=\sqrt{16-16m+m^2}. \tag{42}$$

Hence the corresponding superpotential is

$$W(x)=\frac{-2m\,\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}(x)}+\frac{2\,\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}(x)\psi_0^-(x)}[(4-m-\delta_8)-2(4-2m-\delta_8)\mathrm{sn}^2(x)], \tag{43}$$

and the partner potential $V_+(x)$ which is isospectral to $V_-(x)$ is

$$V_+(x)=-V_-(x)+2W^2(x). \tag{44}$$

It is not difficult to see that the W as given by Eq. (43) does not satisfy the self-isospectral condition (4) and hence unlike the $p=q=2$ case, the $p=q=6$ potential is *not* self-isospectral. In Fig. 8, we have plotted the potentials $V_\pm(x)$ corresponding to $p=q=6$ for several different values of the parameter m. The figures confirm that the potentials are far from being self-isospectral. Thus we have obtained a new exactly solvable periodic potential (44) which has two bound bands and a continuum band, with five band edges and the corresponding eigenfunctions being exactly known using Table VI and Eqs. (12) and (13). In Fig. 9, we plot the band edge energies for the (6,6) potential as a function of the elliptic modulus parameter m.

It is also clear from here that even the higher associated Lamé potentials with $p=q=12$, 20,..., which have 7, 9,...band edges are also exactly solvable in principle and none of them will be self-isospectral, so that in each case one obtains a new exactly solvable periodic potential. In particular, for $p=q=n(n+1)$ there will be $(2n+1)$ band edges in both $V_\pm(x)$ whose energy eigenvalues can be obtained from Table III when q has the form $[n-2n][n-(2n+1)]$. Out of the $(2n+1)$ band edges in $V_-(x),(n+1)$ solutions (including the ground state) have the form $F_n(\mathrm{sn}^2 x)/[\mathrm{dn}^n x]$ while n solutions have the form $F_{n-1}(\mathrm{sn}^2 x)\mathrm{sn}\,x\,\mathrm{cn}\,x/[\mathrm{dn}^n x]$. On the other hand, as far as the $(2n+1)$ solutions of the partner potential V_+ are concerned, there are n states each of the two forms

$$\frac{\mathrm{sn}\,x\,\mathrm{cn}\,x\,G_n(\mathrm{sn}^2 x)}{\mathrm{dn}^{2n-1}x\,\psi_0^-(x)},\ \frac{G_{n+1}(\mathrm{sn}^2 x)}{\mathrm{dn}^{2n-1}x\,\psi_0^-(x)},$$

while the ground state (i.e., the lower edge of the lowest band) is given by $\psi_0^+(x)=1/\psi_0^-(x)$.

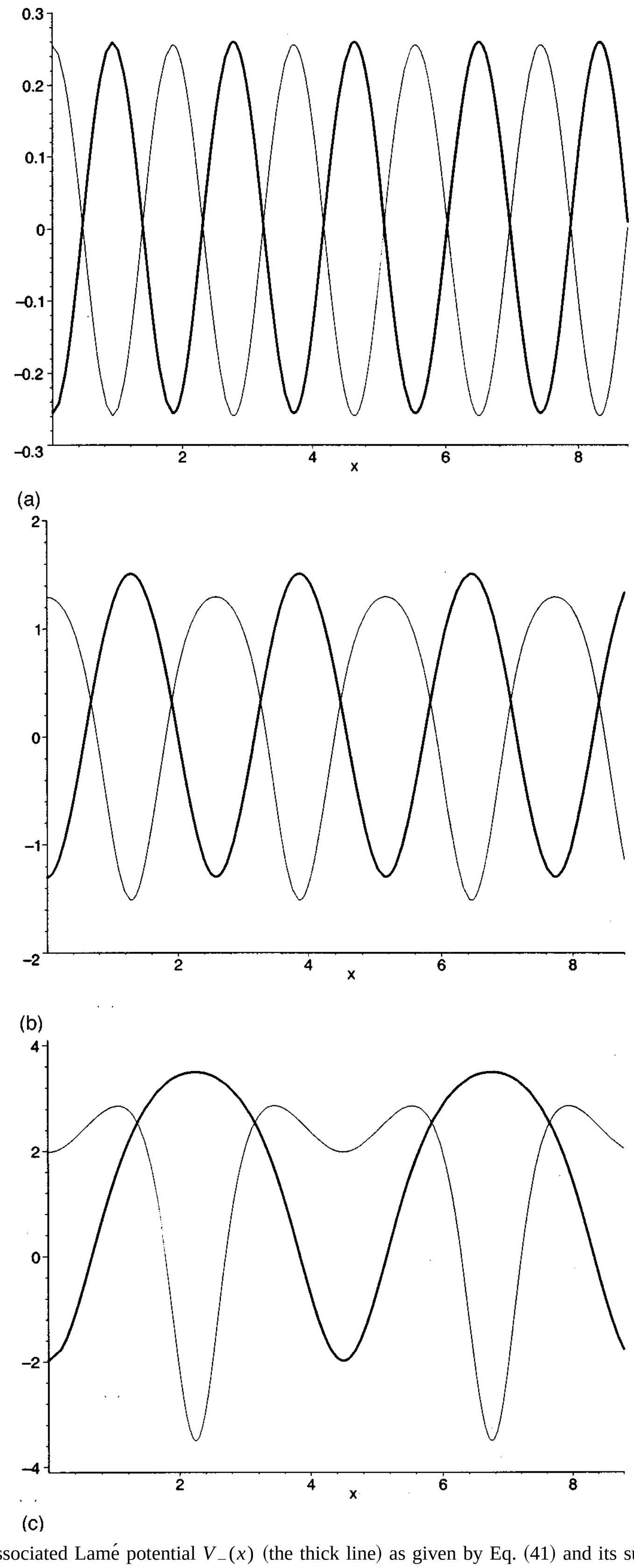

FIG. 8. The (6, 6) associated Lamé potential $V_{-}(x)$ (the thick line) as given by Eq. (41) and its supersymmetric partner potential $V_{+}(x)$ (the thin line) as given by Eq. (44) for three choices of m: (a) 0.5, (b) 0.9, (c) 0.998.

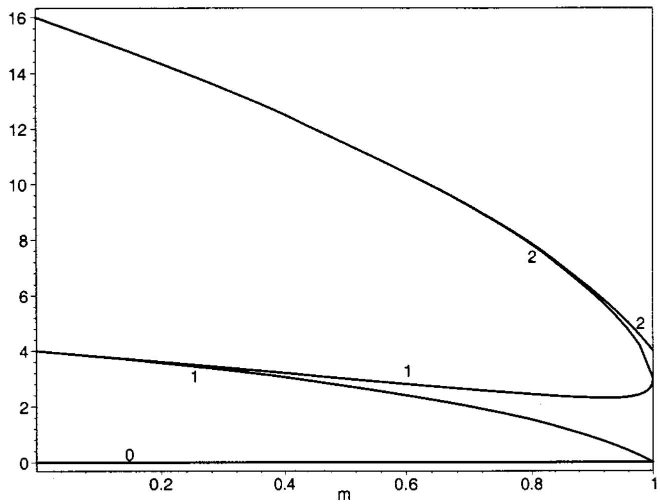

FIG. 9. Band edge energies for the associated Lamé potential (6, 6) as a function of the elliptic modulus parameter m. This figure corresponds to Table VI. The band edges are labeled by the number of wave function nodes in the interval $K(m)$.

C. a,b = half-integer

Let us now specialize to the case when both a, b are half-integral with $a>b$. As an illustration, let us first consider the case of $a=3/2$, $b=1/2$ so that $p=15/4$, $q=3/4$. In this case, the oscillation theorem requires band edges with periods $2K,4K,4K,2K,2K,\ldots$. Using Table III and Fig. 5, we see one gets three eigenstates when $q=(a-2)(a-3)$ with $a=3/2$, all with period $2K$. The ground state is at $E_0=9m/4$ while there are two degenerate levels at $E_3=E_4=4+m/4$. To understand this degeneracy better, let us go along the parabola of solvability P2 given by $q=(a-2)(a-3)$. The band gap is given by $\Delta_2\equiv|-2+m+2\sqrt{1-m+m^2(a-1)^2}|$ and is plotted in Fig. 10. It vanishes at $a=3/2$ (15/4, 3/4) potential, and has the correct values $\Delta_2=2\sqrt{1-m+m^2}-2+m$ for $a=2$ and $\Delta_2=2\sqrt{1-m+4m^2}-2+m$ for $a=3$, which correspond to the (6,0) and (12,0) Lamé potentials. The vanishing of Δ_2 at $a=3/2$ occurs because the eigenfunctions corresponding to E_3 and E_4 cross over as one goes along the parabola P2.

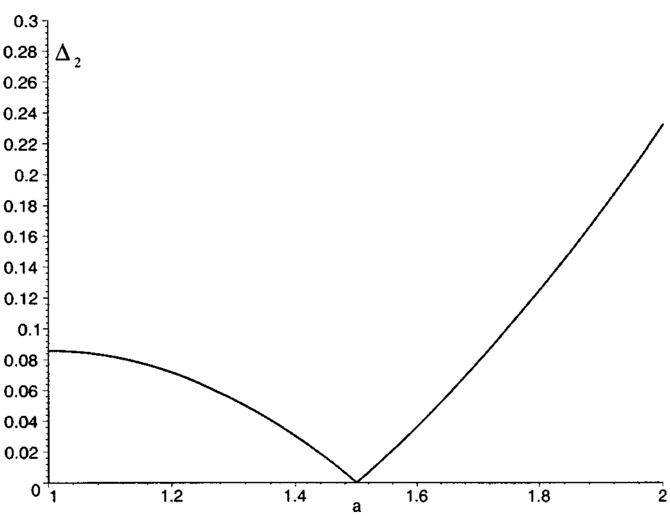

FIG. 10. Energy gap $\Delta_2\equiv|E_4-E_3|$ as one moves along the parabola of solvability $P2$ corresponding to $q=(a-2)\times(a-3)$ and $p=a(a+1)$.

TABLE VII. Energy eigenvalues and eigenfunctions for the associated Lamé potential corresponding to $a=7/2$, $b=1/2$ which gives $(p,q)=(63/4,3/4)$. Here $\delta_9 \equiv \sqrt{4-4m+25m^2}$; $V_-(x)=\frac{63}{4}m\,\mathrm{sn}^2(x)+\frac{3}{4}m\,\mathrm{cn}^2(x)/\mathrm{dn}^2(x)-2-29m/4+\delta_9$. The last column gives the number of eigenfunction nodes in one period $2K(m)$ of the potential.

E	$\mathrm{dn}^{1/2}(x)\psi^{(-)}$	Period	Nodes
0	$[12m\,\mathrm{sn}^2(x)-2-5m-\delta_9]\mathrm{dn}^2(x)$	$2K$	0
$2-m+\delta_9$	$\mathrm{sn}(x)\mathrm{cn}(x)\mathrm{dn}^2(x)$	$2K$	2
$2\delta_9$	$[12m\,\mathrm{sn}^2(x)-2-5m+\delta_9]\mathrm{dn}^2(x)$	$2K$	2
$14-7m+\delta_9$	$\mathrm{sn}(x)\mathrm{cn}(x)[1-2\,\mathrm{sn}^2(x)]$	$2K$	4
$14-7m+\delta_9$	$[1-8\,\mathrm{sn}^2(x)\mathrm{cn}^2(x)]$	$2K$	4

These arguments are easily generalized in case $p=(n+1/2)(n+3/2)$, $q=(k+1/2)(k+3/2)$ with $n>k$. The energy eigenvalues of $(n-k)$ states can be obtained by using Table III in case q is of the form $q=[n+1/2-(n-k)][n+1/2-(n-k-1)]$ and the corresponding eigenstates have period $2K(4K)$ depending on whether $(n-k)$ is odd (even). On the other hand, the energy of $(n+k+2)$ states is obtained when q is of the form $q=[n+1/2-(n+k+2)][n+1/2-(n+k+1)]$ and these states have the same period $2K(4K)$ as the $n-k$ states when $n-k$ is odd (even). It turns out that the $n-k$ solutions are in fact common in both and so we only obtain the energy of the $n+k+2$ band edges and all of them have the same period $2K(4K)$ depending on whether $n-k$ is odd (even), so that it is only a QES problem and not an exactly solvable problem since one is unable to obtain a single eigenstate with period $4K(2K)$ in case $n-k$ is odd (even).

We would like to point out some of the pecularities of the spectrum in these cases. For example, in case $(p,q)=(35/4,3/4),(63/4,3/4),(99/4,3/4)\ldots$ then one finds that several QES energy levels of period $4K,2K,4K,\ldots$, respectively, are analytically known of which the two at the highest energy are doubly degenerate. As an illustration, in Table VII we have given several QES energy eigenstates all of period $2K$ for the (63/4, 3/4) potential. The interesting point about this case is that the partner potentials $V_\pm(x)$ are not self-isospectral and hence one has discovered a new QES potential where several band edges of period $2K$ and the corresponding eigenfunctions are explicitly known. Of these, the band edges with four nodes are doubly degenerate, again due to crossover of energy levels. Using the ground state wave function, the superpotential is computed to be

$$W=\frac{3m}{2}\frac{\mathrm{sn}(x)\mathrm{cn}(x)}{\mathrm{dn}(x)}-\frac{24m\,\mathrm{sn}(x)\mathrm{cn}(x)\mathrm{dn}(x)}{[12m\,\mathrm{sn}^2(x)-2-5m-\sqrt{4-4m+25m^2}]}. \tag{45}$$

Using Eqs. (12) and (13) the eigenstates of the SUSY partner potential V_+ are then determined.

VI. COMMENTS AND CONCLUSIONS

In this paper, we have discussed solutions of the type given in Table III, which correspond to the parabolas of solvability shown in Fig. 5. Lamé potentials $(p,0)$ with $p=a(a+1)$ and integer a, always have two parabolas of solvability passing through—one parabola gives all states of period $2K$ and the other gives all states of period $4K$. This provides a deeper understanding of why such Lamé potentials are fully solvable.[12] Similarly, we have obtained eigenstates for a large class of associated Lamé potentials (p,q). Further, using the formalism of supersymmetric quantum mechanics, we have been able to discover many new exactly solvable and quasiexactly solvable periodic potentials involving Jacobi elliptic functions. This is a very substantial improvement over the currently known small number of exactly solvable periodic problems.

ACKNOWLEDGMENTS

One of us (U.S.) acknowledges the hospitality of the Institute of Physics, Bhubaneswar and partial financial support from the U.S. Department of Energy. Both authors thank the Mehta Research Institute, Allahabad for hospitality.

[1] For the properties of Jacobi elliptic functions, see, e.g., I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York, 1980). The modulus parameter m is often called k^2 in the mathematics literature. The related complementary quantity $(1-m)$ is often called k'^2.
[2] F. M. Arscott, *Periodic Differential Equations* (Pergamon, Oxford, 1981); E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, Cambridge, 1980).
[3] W. Magnus and S. Winkler, *Hill's Equation* (Wiley, New York, 1966).
[4] See, e.g., F. Cooper, A. Khare, and U. P. Sukhatme, Phys. Rep. **251**, 267 (1995).
[5] G. Dunne and J. Feinberg, Phys. Rev. D **57**, 1271 (1998).
[6] G. Dunne and J. Mannix, Phys. Lett. B **428**, 115 (1998).
[7] H. Braden and A. Macfarlane, J. Phys. A **18**, 3151 (1995).
[8] Y. Brihaye and M. Godart, J. Math. Phys. **34**, 5283 (1993); Y. Brihaye and S. Braibant, *ibid.* **34**, 2107 (1993).
[9] U. Sukhatme and A. Khare, E-print quant-ph/9902072 (1999).
[10] A. Turbiner, Commun. Math. Phys. **118**, 467 (1988).
[11] A. G. Ushveridze, *Quasi Exactly Solvable Models in Quantum Mechanics* (IOP, Bristol, 1993).
[12] Some isolated solutions of the Schrödinger equation for Lamé potentials with half-integer values of a are also known (Ref. 2), but they have certain peculiar features which require further study.

Random Schrödinger operators arising from lattice gauge fields. I. Existence and examples

Oliver Knill[a)]
Department of Mathematics, University of Texas, Austin, Texas 78712

(Received 3 August 1998; accepted for publication 19 July 1999)

We consider new models of ergodic Schrödinger operators whose existence relies on a cohomological theorem of Feldman and Moore in ergodic theory. These operators generalize the Harper operator which describes the case of a constant magnetic field. An example is the case when the magnetic field is given by independent random variables attached to the lattice plaquettes. A generalization of the Feldman–Moore theorem by Lind to non-Abelian groups also allows us to consider Schrödinger operators obtained from non-Abelian lattice gauge fields. The existence result extends to more general graphs like to operators on tilings and to higher dimensions. We compute some moment expansions for the density of states. For example, for independent, identically and uniformly distributed magnetic fields, a model which has been studied at least since 1970, and whose existence can also be seen without involving the above-mentioned existence theorem, we show that the *n*th moment is the number of closed paths in the two-dimensional lattice starting at the origin for which the winding number vanishes at each plaquette point. This goes beyond the Brinkman–Rice self-retracing path approximation. Other examples are a higher dimensional example, a one-dimensional Anderson model which can be treated in this framework, as well as the Hofstadter model with constant magnetic field, where one averages over all possible magnetic fields. We also reprove a result of Jitomirskaya–Mandelshtam stating that the deterministic Aharonov–Bohm model is a compact perturbation of the free Laplacian. © *1999 American Institute of Physics.* [S0022-2488(99)03911-0]

I. INTRODUCTION

In this article, we consider a class of ergodic discrete Schrödinger operators which we call discrete random electromagnetic Laplacians. An example in two dimensions is the bounded ergodic self-adjoint operator $L=A+A^*$ on $l^2(\mathbf{Z}^2,\mathbf{C}^N)$, where

$$(Au)_n=A_1(n)u_{n+e_1}+A_2(n)u_{n+e_2}$$

and where the unitary matrices $A_i(n)\in \mathrm{U}(N)$ have the property that the magnetic fields

$$B(n)=A_2(n)^*A_1(n+e_2)^*A_2(n+e_1)A_1(n)$$

on different plaquettes are identically distributed $\mathrm{U}(N)$-valued random variables with law μ (Fig. 1).

The question arises: Given a magnetic field B determined by an arbitrary stochastic process, can we find A such that $B=dA$? If the $B(n)$ are invariant under translation in one direction, the answer is no in general: For a measurable circle-valued map B, there is in general no measurable circle-valued map A such that $B=dA=A(T)A^{-1}$. In a probabilistic or ergodic theoretical setup, nontrivial cohomological constraints appear. Feldman and Moore noticed, however, in Ref. 1 that in dimensions $d=2$, these constraints are absent. This stays true even if the Abelian group $\mathrm{U}(1)$ is

[a)]Electronic mail: knill@math.utexas.edu

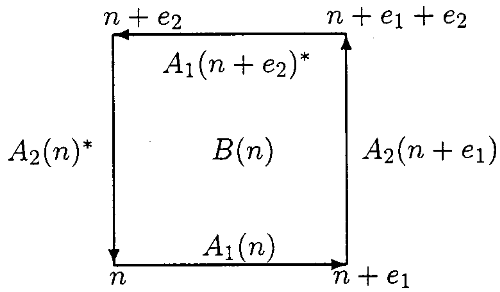

FIG. 1. The magnetic field $B=dA$ at a plaquette n is obtained by integrating up A along the boundary $B(n) = A_2(n)^* A_1(n+e_2)^* A_2(n+e_1) A_1(n)$.

replaced by a non-Abelian group like $U(N)$.[2] For Abelian groups, we generalized this theorem in Ref. 3 to higher dimensions. The existence theorem furthermore extends to discrete ergodic Laplacians on more general graphs like graphs defined by aperiodic tilings.

While in special cases the independent magnetic fields can be obtained directly by choosing the random variables $A_i(n)$ to be independent, this is not true in general. Already in the case of independent identically distributed magnetic fields, there are counterexamples: independent identically distributed random variables $A_i(n)$ lead in general to correlations between the magnetic field variables $B(n)$: If, for example, $A_i(n)$ take randomly the two values 1 and $e^{i\pi/4}$, then two adjacent plaquettes P_n, P_m cannot have magnetic fields $B(n)=1$ and $B(m)=-1$. In other words, vector potentials A, which give independent identically distributed magnetic fields B, are not independent in general.

There are several motivations to study such operators.

(1) Discrete magnetic Laplacians in two dimensions are tight binding approximations for the quantum mechanical model of an electron in the plane exposed to an ergodic magnetic field. This generalizes the Harper operator for which the magnetic field is constant in the plane. Independent identically distributed magnetic fields are models for which the spectral type[4–6] is still unclear. One of the questions is whether such models have eigenvalues.[7] Another problem is to describe the density of states for which approximations have been known since Ref. 8 and for which numerical investigations[9–12] have been done.

(2) Each random operator is an element of a C^* algebra $(\mathcal{X},\mathrm{tr})$ which is determined entirely by the field F. This generalizes the case when $B=F_{12}=e^{2\pi i\alpha}$ is constant and where $\mathcal{X}$ is the rotation algebra.[13,14] The relatively abstract Feldman–Moore theorem in ergodic theory allows us to so define subalgebras in the crossed products of $L^\infty(X)$ with a $\mathbf{Z}^2$ action which is generated by two unitaries U,V which have a prescribed commutator $UVU^{-1}V^{-1}$ in $L^\infty(X)$.

(3) In the partition function of one matrix models appears a van der Monde determinant (see, e.g., Ref. 15). Using the potential theoretical energy $I(L)=-\int\int \log|E-E'|dk(E)dk(E')$ of the density of states dk of an ergodic Schrödinger operator L one can define an infinite-dimensional van der Monde determinant $e^{-I(L)}$. This leads to the variational problem $L \mapsto e^{-I(L)}$ which is the topic of Ref. 16.

We now give an overview over the results of this paper. We first define electromagnetic Laplacians and observe that the Feldman–Moore–Lind results imply that ergodic magnetic Laplacians exist. The formalism can be considered as an ergodic version of differential forms. To a field $F=dA$ is attached a current $j=d^*F$ which is divergence-free $d^*j=0$. In two dimensions, not every current is given by a field. The equivalence classes of currents j, modulo currents of the form d^*F is the cohomology group $\mathcal{H}^1(U)$, a group with the cardinality of the continuum.[3] However, in dimensions $d>2$, it is again a consequence of the triviality of higher dimensional cohomology groups that every one-form j is of the form $j=d^*F$.

In Sec. II, we compute moments of the density of states for independent identically distributed magnetic Laplacians in two dimensions. If the law of the magnetic field B is the Haar measure μ_{Haar} on $U(1)$, then the density of states is determined by a random walk in $\mathbf{Z}^2$ having the following global geometrical constraints: the nth moment of the density of states, $\mathrm{tr}(L^n)$ is the number of closed paths in $\mathbf{Z}^2$ which have length n and give zero winding number to every plaquette. The combinatorial problem to compute the moments of the density of states was considered first in Ref. 8, where the approximation was used that the paths should have no loops. This is now called the ''Brinkman–Rice approximation.'' The exact expression which we give in this paper for the number of paths is new and was not mentioned in Ref. 8 or in subsequent works on the problem that we are aware of. We show also that random magnetic fields with law μ_{Haar} can be generated by taking μ_{Haar}-distributed vector potentials, so that in this special case, Feldman–Moore's existence theorem is not needed. We notice then that all the spectral properties of the operators in the Abelian as well as non-Abelian case depend only on the field $F=dA$ and not on the specific realization of the vector potential A. The explicit calculation of the moments of the density of states for independent identically distributed fields leads to an Aubry duality for the deformed operators $L_\lambda = A_1 + A_1^* + \lambda(A_2 + A_2^*)$: the density of states of L_λ is related to the density of states of $L_{1/\lambda}$ in the same way as for the Harper case.[17]

Some other examples follow in Sec. III. We review a result of Jitomirskaya and Mandelshtam[18] stating that a change of the field on a finite set of cells is a compact perturbation of the operator. A special case is the magnetic Aharonov–Bohm operator with magnetic flux $B \in \mathrm{U}(1)$ different from 1 only in one cell. This result stays true for aperiodic lattices like the Penrose lattice. It seems to be unknown, whether the Aharonov–Bohm perturbation from the free operator is trace class. Also the existence result generalizes to other periodic graphs or aperiodic tilings. We notice for example that to any measure μ on U(1), there exists a measurable vector potential on a Penrose lattice such that the magnetic fields in the plaquettes are independent, identically distributed U(1)-valued random variables with law μ. Because a Penrose graph is not a Cayley graph of a group, the more abstract setup of countable ergodic equivalence relations developed in Ref. 1 is needed.

II. NOTATION AND EXISTENCE

We consider first the two-dimensional case. Let $(X,\mathcal{F},m)$ be a probability space. Two commuting measure-preserving invertible transformations T_1, T_2 on X define a dynamical system with time $\mathbf{Z}^2$. Let U be a Polish (=complete separable metrisable) group. Examples are subgroups of Lie groups like the unitary groups U(N) or the special linear group SL(N,**C**). A two-form $B\tau_{12}$ is defined by a measurable map $B \in \mathcal{U} = \mathcal{L}(X,U) = \{B | X \to U, \text{measurable}\}$. Two measurable U-valued maps $A_1, A_2 \in \mathcal{U}$ define a one-form or vector potential $A = A_1\tau_1 + A_2\tau_2$. Define the curvature of A as the two-form $dA\tau_{12}$ with

$$dA(x) = A_2^{-1}(x) A_1^{-1}(T_2 x) A_2(T_1 x) A_1(x).$$

Not every two-form B can be written as $B = dA$ with a one-form A. For example, if T_1 is the identity map and $T_2 = T$ is ergodic, then not every measurable map $B \in \mathcal{U} = \mathcal{L}(X,U)$ can be written as $B = A^{-1}A(T)$ with $A \in \mathcal{U}$ because the cohomology group

$$\mathcal{H}^1(U) = \mathcal{U} / \{A \in \mathcal{U} | B = A^{-1}A(T)\}$$

of cocycles modulo coboundaries is nontrivial. We know that this group has the cardinality of the continuum.[3] The following result of Feldman and Moore[1] was extended by Lind[2] to non-Abelian groups. A dynamical system given by a group T^g of automorphisms on $(X,\mathcal{F},m)$ is called free, if $m(\{T^g(x) = x\}) > 0$ implies $g = 0$.

Theorem II.1 (Feldman–Moore–Lind): *Assume that the $\mathbf{Z}^2$-dynamical system is free. Let U be a not necessarily Abelian Polish group. For any magnetic field distribution $B\tau_{12}$ with $B \in \mathcal{U}$, there is a vector potential $A = A_1\tau_1 + A_2\tau_2$, which satisfies $dA = B$.*

Example. A magnetic field B taking values in $\{1,-1\}$ is determined by the measurable set $Y = B^{-1}(-1) = \{x | B(x) = -1\}$. Feldman–Moore's result implies that there exist two measurable sets Z_1, Z_2 such that

$$Y=Z_1+T_1(Z_1)+Z_2+T_2(Z_2),$$

where $+$ is the symmetric difference, the addition in the group $\mathcal{F}$.

Assume now that U is a subgroup of the unitary group $U(N)$ of $n\times n$ matrices. Given a one-form $A=A_1\tau_1+A_2\tau_2$, we define a discrete self-adjoint random Schrödinger operator $L=A+A^*$ as follows: For almost all $x\in X$, consider the operator $L(x)$ on $l^2(\mathbf{Z}^2,\mathbf{C}^N)$ given by $(L(x)u)=(A(x)+A(x)^*)u$, where

$$(A(x)u)_n=A_1(x)u_{n+e_1}+A_2(x)u_{n+e_2}$$

and where $e_1=(1,0),e_2=(0,1)$ are the basis vectors in $\mathbf{Z}^2$. We call $L=A+A^*$ a discrete random magnetic Laplacian. Such operators are discrete versions of the continuous operators $L=(\nabla-iA)^2$ (see, e.g., Ref. 5). We also call them ''random magnetic Laplacians'' if $U=U(1)$ or ''random Yang–Mills Laplacians'' if $U=U(N)$. Associated with L is a one-parameter family of operators $L=A_1+A_1^*+\lambda(A_2+A_2^*)$, $\lambda\in\mathbf{R}$ in which we will mainly concentrate on the case $\lambda=1$. The field of a random Laplacian $L=A+A^*$ is defined for $d\geqslant 2$ as $F_{ij}=dA_{ij}=A_j^{-1}A_i(T_j)^{-1}A_j(T_i)A_i$. If $U=U(1)$, one says that the magnetic field $B=A_2^{-1}A_1(T_2)^{-1}A_2(T_1)A_1$ has the magnetic flux $\arg(B)$. The phases of L are the functions $\arg(A_i)$.

The operator L is not uniquely defined by the field B. With a zero-form $C\in\mathcal{U}$, the gauge transformed operator CLC^{-1} is also a discrete random Laplacian with the same magnetic field B but the gauge potential A has changed to $CAC^*\tau=\Sigma_i C_iA_iC_i(T_i)^*\tau_i$. The choice of the gauge is not the only source of nonuniqueness. The nontrivial nonuniqueness is measured by the moduli space of flat fields $\{(A_1,A_2)|dA=0\}/\{A=dC\}$ which is for Abelian U as a group isomorphic to the first cohomology group $\mathcal{H}^1(U)$.

The above-mentioned definitions generalize to the higher dimensional case. Take d automorphisms $T_1,\ldots,T_d$ on a probability space $(X,\mathcal{F},m)$. A one-form $A=\Sigma_{i=1}^d A_i\tau_i$ is given by d functions $A_i\in\mathcal{U}=\mathcal{L}(X,U(N))$ and defines a field

$$dA=F=\sum_{i<j}F_{ij}\tau_{ij},$$

where $F_{ij}=A_iA_j(T_i)A_j(T_i)^{-1}A_j^{-1}$. This gives random self-adjoint operator $L=A+A^*$ where each $L(x)$ acts on the Hilbert space $l^2(\mathbf{Z}^d,\mathbf{C}^N)$ by

$$(L(x)u)=\sum_{i=1}^{d}A_i(x)u+A_i^*u.$$

Theorem II.2 (Triviality of cohomology groups in higher dimensions): *If U is Abelian and* $\mathbf{Z}^d$ *acts freely, then* $dF=0$ *implies that there exists A such that* $dA=F$.

The proof of this theorem which was included in Ref. 19 is now the subject of a separate article.[3] We call a random operator $L=A+A^*$ determined by F a random discrete electromagnetic Laplacian with field F. Examples are given in the following.

(1) A Harper operator is obtained when the field B takes a constant value in $U(N)$. Diagonalization reduces all questions to the case $N=1$. For reviews see Refs. 20–22.

(2) An example of a quasiperiodic magnetic Laplacian is defined by the dynamical system $(U(1),T_1:\theta\mapsto\theta+\alpha,T_2:\theta\mapsto\theta+\beta,d\theta)$ and the magnetic field $B(n)(\theta)=\exp(i(\theta+n_1\alpha+n_2\beta))\in U(1)$. The spectrum of the Laplacian is determined by the two real numbers α,β. Non-Abelian versions, where T_i are translations on the unitary group $U(N)$ are defined similarly. Note however that the vector potential (A_1,A_2) is only measurable and we cannot expect it to be almost periodic.

(3) An example of a limit periodic magnetic field is $B(n)=\Sigma_{j=0}^{\infty} b_j \cos(2\pi(n_1 2^{-j}+n_2 3^{-j}))$ with $\Sigma_j |b_j|<\infty$. As in the previous example, the operator L is not limit periodic and only the physically relevant field B is.

(4) Given a probability measure μ on U(1) and independent, identically distributed random variables $B(n)$, $n\in\mathbf{Z}^2$ with law μ. Even so $B(n)$ are independent, the vector potential A is in general not given by independent identically distributed random variables.

(5) An example of an aperiodic, strictly ergodic field taking only finitely many values is $B(n)=1-2\cdot 1_{[0,\gamma)}(\theta+n_1\alpha+n_2\beta)\in\{-1,1\}$, where γ,α,β are rationally independent.

(6) In all these examples, one obtains one-parameter families of deformed operators $L=A_1+A_1^*+\lambda(A_2+A_2^*)$ with $\lambda\in\mathbf{R}$. In the stationary independent magnetic field case the almost sure spectral properties of L depend only on λ and the field $B=dA$.

Higher dimensional cohomology groups[23–25] are defined as follows (see Refs. 23–25).

Let $I=\{1,\ldots,d\}$ and let $\mathcal{I}_p$ be the set of sets $J=\{j_1<j_2<\ldots<j_p\}\subset I$. Let $\mathcal{C}^p$ be the set of maps $A:\mathcal{I}_p\to\mathcal{U}$ which becomes a group by pointwise addition. Extend this map to the set of all p-tuples $J=(j_1,j_2\ldots,j_p)$ with $j_k\in I$ by requiring $A_{\pi(J)}=\mathrm{sign}(\pi)A_J$ for any permutation π of J. We write $A=\Sigma_J A_J\tau_J$. Define $d_p:\mathcal{C}^p\to\mathcal{C}^{p+1}$ by

$$d_pA=\sum_{i,J}(A_J(T_i)-A_J)\tau_{iJ}.$$

The kernel of d_p contains cocycles of degree p, whereas the image of d_{p-1} consists of coboundaries of degree p. Because for $A=\Sigma_J A_J\tau_J$,

$$d_p\circ d_{p-1}A=\sum_{i,j,J}[A_J(T_iT_j)-A_J(T_i)-A_J(T_j)+A_J]\tau_{ijJ}$$

is both symmetric and antisymmetric in i,j, it must vanish and $d_p\circ d_{p-1}=0$ gives rise to geometric cohomology groups $\mathcal{H}^p_{\mathrm{geom}}(G,\mathcal{U})=\ker(d_p)/\mathrm{im}(d_{p-1})$.

Cohomology of currents. Given a $\mathbf{Z}^d$ action on the group $\mathcal{U}=\mathcal{L}(X,U)$. A one-form $A=\Sigma_i A_i\tau_i\in\mathcal{C}^1$ defines an electromagnetic field $F=dA\in\mathcal{C}^2$ and so a current $j=d^*F=*d*F\in\mathcal{C}^1$, where an asterisk $(*)$ is the Hodge operation $*:\mathcal{C}^n\to\mathcal{C}^{d-n},A_I\mapsto(-1)^{n(d-n)}A_{I*}$, where $I^*=\{1,\ldots,n\}\setminus I\}$. A current is defined even if the group $\mathcal{U}$ is non-Abelian.

Proposition II.3: Assume $N\geq 1,d=2$ *or* $N=1,d\geq 2$. *Every current* $d^*F=j$ *is divergence free:* $d^*j=0$.

Proof: If $d=2$, the Hodge involution for one-forms is given by $A_1\tau_1+A_2\tau_2=(A_1,A_2)\mapsto(A_2,A_1^{-1})$. The divergence of j is given by

$$d^*j=*d*(j_1,j_2)=(*d)(j_2,j_1^*)=j_1j_2(T_2)^*j_1^*(T_1)j_2.$$

If we plug in $j=(j_1,j_2)=d^*F=(F^*F(T_2),F(T_1)^*F)$, we get

$$d^*j=F^*F(T_2)F(T_2)^*F(T_1T_2)F(T_2T_1)^*F(T_1)F(T_1)^*F=1.$$

In the Abelian case, $d^*j=0$ follows in any dimension from $d^*d^*=0$. □

One can ask whether every current j which is divergence free $d^*j=0$ does come from a field F satisfying $d^*F=j$ (Fig. 2). The answer is ''no'' in two dimensions and ''yes'' in dimensions three or higher. There are uncountably many equivalence classes of currents in two dimensions because of the following Proposition.

Proposition II.4: Assume $d=2$, *let* U *be a Polish group and let* G *be a free* $\mathbf{Z}^2$ *action. The moduli space of all divergence free currents* j *modulo currents* j *coming from fields* $j=d^*F$ *is isomorphic to the first cohomology group* $\mathcal{H}^1(U)$. *On the other hand, for* $d\geq 3$ *and Abelian* U, *every divergence free current* j *is of the form* d^*F.

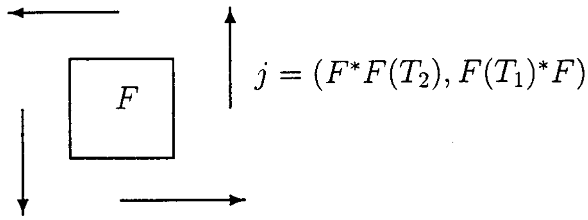

FIG. 2. Illustration of the current of a not necessarily Abelian Aharonov–Bohm field, where the field F is constant different from 1 on one plaquette only.

Proof: Assume first $d=2$. j is a cocycle if $j_2 j_1^*(T_1)=j_1^* j_2(T_2)$ and a coboundary if there exists a solution F of $j_1=F(T_2)F^*$, $j_2=FF(T_1)^*$. If j is a cocycle, then the Hodge dual $\tilde{j}=*j$ satisfies a zero curvature equation. Also, $d^*F=j$ if and only if $\tilde{j}$ is a gradient $d(*F)=\tilde{j}$. The moduli space of zero curvature fields modulo gradient fields is $\mathcal{H}^1(U)$.

Assume now $d\geq 3$ and that U is Abelian. Given the one-form j, define the $(d-1)$-form $\tilde{j}=*j$. Since $\mathcal{H}^{d-1}(U)$ is trivial for $d\geq 3$, there exists a $(d-2)$-form $\tilde{F}$ satisfying $d\tilde{F}=\tilde{j}$. Let $F=*\tilde{F}$. Then $d^*F=j$. □

Question. We do not know whether in dimensions $d\geq 3$, every current j can be written as d^*F with a field F satisfying additionally $dF=0$. If this were true and F were unique, an interesting class of higher dimensional operators $L=A+A^*$ were defined by taking independent identically distributed random variables j and taking A satisfying $d^*dA=j$.

III. GENERAL REMARKS ON THE SPECTRUM

If U is a subgroup of the unitary group $\mathrm{U}(N)$, the Laplacian L is an element of a von Neumann algebra $\mathcal{X}$ which is the crossed product of $\mathcal{A}=L^\infty(X,M(N,\mathbf{C}))$ with the $\mathbf{Z}^d$-action generated by automorphisms $f\mapsto f(T^n)$, where $f(T^n)(x)=f(T^n x)$. The algebra $\mathcal{X}$ is obtained by completing the algebra of all polynomials in the variables $\tau_1,\ldots,\tau_d$ with coefficients in $\mathcal{A}$,

$$K=\sum_{n\in F\subset\mathbf{Z}^d} K_n\tau^n,\quad (KL)_n=\sum_{l+m=n} K_l L_m(T^l)\tau^n$$

with respect to the norm $|||K|||=|\ ||K(x)||\ |_\infty$. Here, $K(x)$ is the bounded linear operator on $l^2(\mathbf{Z}^d,\mathbf{C}^N)$ defined by $(K(x)u)(n)=\Sigma_m K_m(x)u(n+m)$ and $||\cdot||$ is the operator norm on $\mathcal{B}(l^2(\mathbf{Z}^d))$ and $|\cdot|_\infty$ the essential supremum norm. The involution in $\mathcal{X}$ is $(\Sigma_n K_n\tau^n)^*=\Sigma_n K_n^*(T^{-n})\tau^{-n}$ and the *trace* is $\mathrm{tr}(K)=\int_X \mathrm{Tr}(L_0(x))d\mu(x)$, where Tr denotes the usual trace on the finite dimensional matrix algebra $M(N,\mathbf{C})$. This construction of Murray and von Neumann works in the same way, when $\mathbf{Z}^d$ is replaced by a more general discrete group.

For any self-adjoint $L\in\mathcal{X}$, and if f is a continuous, bounded function on $\mathbf{R}$, the element $f(L)\in\mathcal{X}$ is defined through the functional calculus. The functional $f\mapsto\mathrm{tr}(f(L))$ on $C(R)$ defines a measure on $\mathbf{R}$, called the density of states of L. If the $\mathbf{Z}^2$ action is ergodic, then the spectrum of $L(x)$ is constant almost everywhere and coincides with the support of the density of states.

The magnetic field does not determine L because of gauge ambiguity. However, all the information about the spectrum is determined from B only:

Proposition III.1: Assume $N\geq 1, d=2$ *or* $N=1, d\geq 2$. *Given two one-forms* $A,\tilde{A}$ *which satisfy* $d\tilde{A}=dA=B$. *Define* $L=A+A^*, \tilde{L}=\tilde{A}+\tilde{A}^*$. *There exists for every x a unitary operator* $\mathrm{U}(x)$ *on* $l^2(\mathbf{Z}^d,\mathbf{C}^N)$ *such that* $\mathrm{U}^*(x)L(x)\mathrm{U}(x)=\tilde{L}$. *Especially, the density of states, the (in the ergodic case almost everywhere constant) spectral types and the spectrum depend only on the field B.*

Proof: We consider first a setup, where measurability is discarded. Let $\mathcal{U}=\mathrm{U}(N)^{\mathbf{Z}^2}$ be the set of possible fields $n\mapsto B(n)$. For every $B\in\mathcal{U}$, we can find a bounded self-adjoint operator $L=A+A^*$ on $l^2(\mathbf{Z}^2,\mathbf{C}^N)$ which has the field $B=dA$. The special gauge $A_2(n,m)=A_1(0,m)=1$ for

$n,m \in \mathbf{Z}$, determines A and makes $L = L_B$ unique. For $N=1$, a canonical gauge can also be defined for $d \geqslant 2$: let $E_i \subset \mathbf{Z}^d$ be the vector space spanned by $(e_1, e_2, \ldots, e_i)$. Put $A_i = 1$ on E_1, then on all lines orthogonal to E_1 in E_2 and inductively on all lines orthogonal to E_j in E_{j+1}. This determines A as can be seen by induction: first construct A on E_2 then on edges orthogonal to E_2 in E_3, etc., always using $dA = B$. The condition $dB = 0$ assures that the definition is consistent. The map $B \mapsto L_B$ is continuous if $\mathcal{U}$ has the product topology and $\mathcal{B}(l^2(\mathbf{Z}^d))$ has the strong operator topology. However a change of the magnetic field of one single plaquette changes L_B globally.

Take the diagonal operator $U_{n,m} = G(n) \delta_{n,m}$, where $\overline{G(n)} A_i(n) G(n-e_i) = \tilde{A}_i$. This function $n \mapsto G(n)$ exists since any vector potential can be gauged to the canonical gauge.

For ergodic operators L, the density of states of L and $\tilde{L}$ exists. We do not need the measurability of the conjugating operator $\mathrm{U}(x)$ to get $L^n(x)_{00} = \tilde{L}^n(x)_{00}$. It follows that $\mathrm{tr}(L^n) = \int_X \mathrm{Tr}(L^n(x)_{00}) dm(x) = \int_X \mathrm{Tr}(\tilde{L}^n(x)_{00}) dm(x) = \mathrm{tr}(\tilde{L}^n)$.

Similarly, $(\phi, L^n(x), \phi) = (U(x)\phi, \tilde{L}^n(x), U(x)\phi)$ for every $\phi \in l^2(\mathbf{Z}^d, \mathbf{C}^N)$, so that also the spectral types are the same. □

Remark: The map $x \mapsto \mathrm{U}(x)$ is not measurable in general. Examples with $T_1 = \mathrm{Id}$ show this. In other words, while the operators $L(x)$ and $\tilde{L}(x)$ are conjugated in $\mathcal{B}(l^2(\mathbf{Z}^d, \mathbf{C}^N))$, the conjugation is in general not possible in the algebra $\mathcal{X}$.

Corollary III.2: Given a sequence of operators $L^{(n)} = A^{(n)} + (A^{(n)})^*$ and an operator $L = A + A^*$ defined over the same $\mathbf{Z}^d$ action. Assume that the fields $B^{(n)} = dA^{(n)}$ converge to $B = dA$ in $L^\infty(X,U)^d$. If there exists an interval $I \subset \mathbf{R}$ such that $\sigma(L^{(n)}) \cap I = \emptyset, \forall n \in \mathbf{N}$ then also $\sigma(L) \cap I = \emptyset$.

Proof: In the canonical gauge, the operators converge pointwise in the strong operator topology and so in the resolvent sense. The claim follows from general principles (Ref. 26 Theorem VIII.24). □

IV. INDEPENDENT IDENTICALLY DISTRIBUTED MAGNETIC FIELDS

We concentrate in this paragraph on the case of magnetic Laplacians, where the magnetic fields $\{B(n)\}_{n \in \mathbf{Z}^2}$ are independent, identically distributed U(1)-valued random variables B with law μ. This means that the probability that B takes a value in some interval $I \subset \mathbf{T}$ is $\mu(I)$. Denote by $\hat{\mu}_n = \int_{\mathrm{U}(1)} z^n \, d\mu$ the nth moment of μ. The sequence $\{\hat{\mu}\}_n$ is the Fourier transform of the measure μ. Denote by Γ_n the set of oriented closed paths in $\mathbf{Z}^2$ of length n. Let $n(\gamma,P)$ be the winding number of the path γ with respect to the plaquette P. The moments of the density of states can be computed with a random walk expansion as it is used in statistical physics (see, e.g., Ref. 27).

Proposition IV.1: Let μ be a Borel measure on the circle $\mathrm{U}(1)$. *The* nth *moment of the density of states* $\mathrm{tr}(L^n)$ *of an independent identically distributed magnetic Laplacian with law μ is*

$$\sum_{\gamma \in \Gamma_n} \prod_P \left(\int_{\mathrm{U}(1)} z^{n(\gamma,P)} \, d\mu(z) \right) = \sum_{\gamma \in \Gamma_n} \prod_P \hat{\mu}_{n(\gamma,P)} .$$

Proof: Given A satisfying $dA = B$. Write $\int_\gamma A$ for the product of the A_i along the path γ. The path γ encloses a region $\Omega(\gamma)$ which is a collection $\{P\}$ of plaquettes. We use the discrete Green formula $\int_\gamma A = \prod_P B^{n(\gamma,P)}$ to compute

$$\begin{aligned}
\mathrm{tr}(L^n) &= \sum_{\gamma \in \Gamma_n} \int_X \int_\gamma A(x) \, dm(x) \\
&= \sum_{\gamma \in \Gamma_n} \int_X \prod_P B_P(x)^{n(\gamma,P)} \, dm(x) \\
&= \sum_{\gamma \in \Gamma_n} \prod_P \int_X B_P(x)^{n(\gamma,P)} \, dm(x)
\end{aligned}$$

$$= \sum_{\gamma \in \Gamma_n} \prod_P \left(\int_{U(1)} z^{n(\gamma,P)} \, d\mu(z) \right).$$

In this calculation, we used that the expectation of a product of independent random variables is the product of the expectation and that X^n is independent of Y^n if X is independent of Y. □

Remarks: (1) The random walk expansion breaks down if the group U(1) is replaced by a non-Abelian group U. While it is then still true that $\mathrm{tr}(L^n) = \Sigma_{\gamma \in \Gamma_n} \int_X \int_\gamma A(x)\, dm(x)$, where $\int_\gamma A(x)$ is an ordered product, the Green formula $\int_\gamma A = \Pi_P B^{n(\gamma,P)}$ no longer makes sense.

(2) The moments of the spectral measures of the unit vectors δ_k, $k \in \mathbf{Z}^2$ can also be computed with a random walk expansion $(\delta_k, L^n(x)\delta_k) = \Sigma_{\gamma \in \Gamma_n} \Pi_\gamma A(T^k x) = \Sigma_{\gamma \in \Gamma_n} \Pi_P B_P(T^k x)^{n(\gamma,P)}$.

(3) It follows from Proposition IV.1 that the density of states dk depends continuously on the law μ.

Also for the deformed operators $L_\lambda = A_1 + A_1^* + \lambda(A_2 + A_2^*)$, there is a similar formula for the density of states. Let $y(\gamma) \in 2\mathbf{N}$ be the number of steps a path makes in the y direction.

Corollary IV.2 (Aubry duality): The moments of the density of states of L_λ depend only on μ and λ:

$$\mathrm{tr}(L_\lambda^n) = \sum_{\gamma \in \Gamma_n} \lambda^{y(\gamma)} \prod_P \hat{\mu}_{n(\gamma,P)}.$$

Furthermore, the duality $dk(\mu,\lambda,E) = dk(\mu,1/\lambda,E/\lambda)$ holds.

Proof: The random walk expansion is proven in the same way as in Proposition IV.1. It follows that $L_\lambda^{(1)} = A_1 + A_1^* + \lambda(A_2 + A_2^*)$ and that $L_\lambda^{(2)} = \lambda(A_1 + A_1^*) + A_2 + A_2^*$ have the same density of states. The duality follows from $L_\lambda^{(1)}/\lambda = L_{1/\lambda}^{(2)}$. □

It follows that the ''Lyapunov exponent'' $\lambda(\mu,\lambda,E) := \int \log|E-E'| dk(\mu,\lambda,E')$ satisfies $\lambda(\mu,\lambda,E) \geq \log(\lambda/2)$.

Remark: If μ is the Haar measure and $\{A_1(n), A_2(n)\}_{n \in \mathbf{Z}^2}$ are independent Haar distributed random variables, the duality is stronger: The obvious symmetry $A_1 \leftrightarrow A_2$ implies that the operators $L_\lambda^{(1)} = A_1 + A_1^* + \lambda(A_2 + A_2^*)$ and $L_\lambda^{(2)} = \lambda(A_1 + A_1^*) + A_2 + A_2^*$ are isospectral. Especially, L_λ and $\lambda L_{1/\lambda}$ are isospectral.

The formula for the moments of the density of states becomes especially simple if μ is the Haar measure on $\mathbf{T}$.

Corollary IV.3: Let μ be the Haar measure on U(1). *Then* $\mathrm{tr}(L^n)$ *is the number of closed paths of length n starting at* $0 \in \mathbf{Z}^2$ *for which every plaquette has zero winding number.*

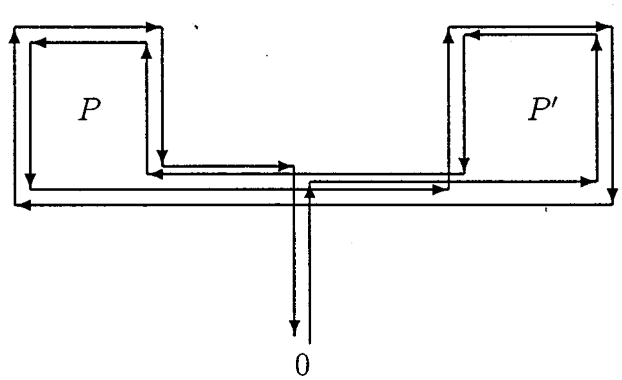

FIG. 3. A noncontractible path in the two-dimensional lattice which gives zero winding number to all plaquettes. These paths are neglected in the Brinkman–Rice ''self-retracting path approximation.''

Proof: In this case, $\hat{\mu}_k = \delta_{k,0}, \forall k \in \mathbf{Z}$. This implies $\Pi_P(\int_{\mathrm{U}(1)} z^{n(\gamma,P)}\, d\mu(z)) = 0$ if there exists a plaquette P for which the path γ has positive winding number $n(\gamma,P)$. If the winding number is zero for all P, then $\Pi_P(\int_{\mathrm{U}(1)} 1\, d\mu(z)) = 1$. □

Remarks: (1) The number of closed paths γ of length n in $\mathbf{Z}^2$ for which $n(\gamma,P)=0$ for all plaquettes P is in general strictly larger than the number of closed contractible paths of length n. A path can visit different plaquettes at different times without being contractible. See Fig. 3.

The additional paths not treated in the Brinkman–Rice approximation are so numerous that the radius of convergence for the Green function changes. Random walks with the stronger topological constraint of being contractible were investigated in Ref. 28.

(2) For the deformed operator $L = A_1 + A_1^* + \lambda(A_2 + A_2^*)$ with independent Haar distributed functions A_i, we get

$$\mathrm{tr}(L^n) = \sum_{\gamma \in \Gamma_n^{(0)}} \lambda^{y(\gamma)},$$

where $\Gamma_n^{(0)}$ is the set of paths in Γ_n which give zero winding number to every plaquette and $y(\gamma)$ is the number of steps the path γ makes in the y direction.

Corollary IV.4: Let μ be the Haar measure on a finite cyclic subgroup $\mathbf{Z}_p$ of U(1). *Then* $\mathrm{tr}(L^n)$ *is the number of closed paths of length n beginning at $0 \in \mathbf{Z}^2$ for which the winding numbers satisfy $n(\gamma,P) = 0 (\mathrm{mod}\, p)$ for all plaquettes P.*

Proof: $\Pi_P(\int_{\mathrm{U}(1)} z^{n(\gamma,P)}\, d\mu(z)) = 0$, if there exists a plaquette P which has a winding number $n(\gamma,P)$ which is not zero modulo p. □

Remarks: (1) The independence of the magnetic fields is essential in Proposition IV.1 and its Corollaries. Independent vector potentials would not be enough in general.

(2) The random walk expansion in Proposition IV.1 shows that $\mathrm{tr}(L^n)$ is a polynomial in the infinite set of variables $\hat{\mu}_k, \hat{\mu}_k^{-1}$ with integer coefficients.

Let us illustrate the random walk expansion in the almost Mathieu–Harper case, where μ is a point measure so that the magnetic field is constant:

Corollary IV.5: If μ is the Dirac measure on $e^{2\pi i\alpha} \in$ U(1), *then*

$$\mathrm{tr}(L^n) = \sum_{\gamma \in \Gamma_n} \prod_P e^{2\pi i \alpha n(\gamma,P)}.$$

(Especially, if μ is the Dirac measure on $1 \in$ U(1), *then* $\mathrm{tr}(L^n) = ((2n)!)^2 (n!)^{-2}$ *is the number of closed paths of length n beginning at $0 \in \mathbf{Z}^2$.) For the Harper magnetic Laplacian $L = A_1 + A_1^* + \lambda(A_2 + A_2^*)$,*

$$\mathrm{tr}(L^n) = \sum_{\gamma \in \Gamma_n} \lambda^{y(\gamma)} \prod_P e^{2\pi i \alpha n(\gamma,P)},$$

where $y(\gamma) \in 2\mathbf{N}$ is the number of steps of γ in the y direction.

For another illustration, let L be the Harper operator over the (nonergodic) integrable ''twist map'' $T:(x,y) \mapsto (x+y,y)$ on the two-dimensional torus $\mathbf{T}^2$ with invariant Lebesgue measure. We consider the potential $V(x,y) = 2\cos(x)$. Call the random (now nonergodic) operator the ''Hofstadter operator.'' The density of states has zero Lebesgue measure.[29] The following illustration should be compared with the case of a random magnetic field:

Corollary IV.6: If L is the Hofstadter operator, then $\mathrm{tr}(L^n)$ *is the number of closed paths of length n beginning at $0 \in \mathbf{Z}^2$ for which the sum of all winding numbers $\Sigma_P n(\gamma,P)$ over all plaquettes P vanishes.*

Proof: $\int_{\mathbf{T}} \exp(\Sigma_P 2\pi i y n(\gamma,P)) dy = 1$ if and only if $\Sigma_P n(\gamma,P) = 0$. □

Remarks: (1) Random walk expansions would work on any planar graph generalizing $\mathbf{Z}^2$ (see also Sec. VI). A solvable case is the Bethe lattice $\mathbf{B}_{2h}$ with degree $2h$, where Kesten determined the number of closed paths of length n.[30] His calculation leads to $\mathrm{tr}(L-E)^{-1} = (2h-1)/(E(h$

$-1)+h\sqrt{E^2-4(2h-1)})$. The imaginary part of this divided by π is the density of states which has support on $[-2\sqrt{2h-1},2\sqrt{2h-1}]$. (For a modern calculation see Ref. 31). Since there are no closed loops on the Bethe lattice, there is no magnetic field and the density of states of any magnetic Laplacian on the Bethe lattice is the same.

(2) Corollary IV.5 shows that $\mathrm{tr}(L^4)=8+4(\hat{\mu}_0+\hat{\mu}_0^{-1})$, so that discrete random magnetic Laplacians with different values of $\mathrm{Re}(\hat{\mu}_0)$ have different density of states. Especially, almost Mathieu operators with different $\cos(\alpha)$ cannot be isospectral (see also Ref. 32).

(3) By changing the orientation of the plaquettes, it becomes obvious that the laws μ and $\overline{\mu}(A)=\mu(\overline{A})$ give isospectral Laplacians.

In some cases, the theorem of Feldman–Moore is not needed for constructing the Laplacian.

Proposition IV.7: Let U be a compact subgroup of U(1). *If the law μ of the independent identically distributed U*-valued random variables $\{A_i(n)\}_{n\in\mathbf{Z},i=1,2}$ *of the vector potential* $A_1\tau_1+A_2\tau_2$ *is the Haar measure on U, then $B(n)=dA(n)$ are independent identically distributed U-valued random variables with law μ.*

Proof: Given measurable subsets $Y_n\subset U$, $n\in\mathbf{Z}^2$ of positive measure, define $Z_n=B(n)^{-1}(Y_n)\subset X=U^{(\mathbf{Z}^2)}$. Let $m=\mu^{(\mathbf{Z}^2)}$ be the product measure on X. The claim is that

$$m\Big(\bigcap_{n\in F} Z_n\Big)=m(Z_k)\cdot m\Big(\bigcap_{n\in F\setminus\{k\}} Z_n\Big) \tag{1}$$

for any finite set $F\subset\mathbf{Z}^2$ and that the law of $B(n)$ is the Haar measure μ.

(i) $m(Z_n)=\mu(Y_n)$, for all $n\in\mathbf{Z}^2$.

Proof: A product of Haar distributed U-valued random variables is again Haar distributed because it must be U invariant. It follows that the law of $B(n)=A_2^*(n)A_1^*(n+1)A_2(n+1)A_1(n)$ is the Haar measure μ and therefore $m(Z_n)=\mu(Y_n)$.

(ii) For any finite set F of sets $\{Y_n\}_{n\in F}\subset U$ with $m(Y_n)>0$, one has $m(\cap_{n\in F}B(n)^{-1}(Y_n))>0$.

Proof: We can realize one element in $\cap_{n\in F}B(n)^{-1}(Y_n)$ using the canonical gauge. There exists then an open neighborhood of this point in $U^{(\mathbf{Z}^2)}$ which is in $\cap_{n\in F}B(n)^{-1}(Y_n)$. An open set has positive measure.

(iii) For $k\in F$, the measure $\tilde{\mu}(Y_k)=m(B(k)^{-1}(Y_k)|\cap_{n\in F\setminus\{k\}}Z_n)$ is equal to $\mu(Y_k)=m(Z_k)=m(B(k)^{-1}(Y_k))$.

Proof: By the uniqueness of the Haar measure, we have only to show that $\tilde{\mu}$ is translational invariant. By multiplying $A_1(k+l\cdot e_i),l=1,\ldots,|F|$ with some constant $C=e^{2\pi\alpha}\in U$, we change the field $B(k)\mapsto B(k)C$ without affecting $\{B(n)\}_{n\in F\setminus\{k\}}$. Therefore $\tilde{\mu}(Y_k)=\tilde{\mu}(Y_k+\alpha)$ and $\tilde{\mu}=\mu$.

Proof of the claim. By (ii), Eq. (1) can be written as

$$m\Big(Z_k\Big|\bigcap_{n\in F\setminus\{k\}} Z_n\Big)=m(Z_k).$$

The left-hand side of this is by (iii) equal to $\tilde{\mu}(Y_k)=\mu(Y_k)$ and the right-hand side is by (i) also equal to $\mu(Y_k)$. □

Remarks: (1) There are other ways to get independent magnetic fields, if μ is the Haar measure: Define $A_2(n)=1$ for all $n\in\mathbf{Z}^2$ and a family $\{A_1(n)\}_{n\in\mathbf{Z}^2}$ of independent Haar distributed random variables. An argument similar to the proof of Proposition IV.7 shows that $\{dA(n)=B(n)\}_{n\in\mathbf{Z}^2}$ are independent Haar distributed random variables.

(2) We do not know whether a generalization of Proposition IV.7 holds when U is non-Abelian.

(3) In dimensions $d>2$, there is no hope to get a result analogous to Proposition IV.7, because there are then more plaquettes than bonds so that a single bond influences several plaquettes and prevents independent, identically distributed fields.

(4) Another open question is whether one has some or even pure point spectrum almost everywhere in the case of magnetic Laplacians with Haar distributed magnetic vector potentials. One would at least expect to have pure point spectrum for L_λ with λ large or small enough. For numerical calculations see Ref. 7.

(5) Proposition IV.7 shows that for those specific operators, there is more symmetry as in the Mathieu case. Aubry-duality goes deeper: the operators L_λ and $L_{1/\lambda}$ have the same spectral type because a multiplication of L_λ with $1/\lambda$ gives $L_{1/\lambda}$.

V. OTHER EXAMPLES

A. Laplacians in higher dimensions

We turn now to independent identically distributed magnetic Laplacians in higher dimensions. We restrict the discussion to the case $d=3$. As indicated already, we cannot realize independent, identically distributed electromagnetic fields F by a vector potential, since such fields do not satisfy the Maxwell equation $dF=0$, which is required if $F=dA$. Consider now time-dependent magnetic fields in the plane together with an electric field changing in time. Given a vector potential $A=(A_1,A_2,A_3)\in\mathcal{C}^1$, we think of A_1 as the electrostatic potential and of (A_2,A_3) as the magnetic vector potential. Then $dA=F$ is a three-dimensional field. $E_1=F_{12}$ and $E_2=F_{13}$ are the coordinates of an ''electric'' vector field in the plane and $B=F_{23}$ is a ''magnetic'' field in the plane. For fixed $k\in\mathbf{Z}$, denote by $L^{(k)}$ the magnetic Laplacian in the plane, given by the vector potential $(n,m)\mapsto(A_2(k,n,m),A_3(k,n,m))$. The operator $L^{(k)}$ is a two-dimensional magnetic Laplacian at time k.

The existence theorem in Ref. 3 assures that a field F satisfying $dF=0$ defines an electromagnetic Laplacian L determined by a one-form A satisfying $F=dA$. By prescribing the electric fields E_1,E_2 and the magnetic field $B^{(k_0)}$ at some time k_0, the Maxwell equation $dF=0$ determines the whole field F.

The next proposition which follows from the central limit theorem for circle-valued random variables, emphases why IID magnetic distributed operators with Haar distribution are natural.

Proposition V.1: Let F be determined by the electric fields and the magnetic field at some time k_0. *Assume that the electric fields* $\{E_1(n),E_2(n)\}_{n\in\mathbf{Z}^3}$ *are independent identically distributed random variables with the same distribution* μ *which is not a Haar distribution of a subgroup of* U(1). *Let* $B(k_0,n)$, $n\in\mathbf{Z}^2$ *be any set of random variables. Then the distribution of the magnetic field of the two-dimensional operators* $L^{(k)}$ *converges in law to the uniform Haar distribution of* U(1) *for* $|k|\to\infty$.

Proof: The Maxwell equation $dF=0$ (which follows from $F=dA$), *implies that*

$$B^{(k+1)}(n)^*=B^{(k)}(n)E_1^{(k)}(n+e_1)E_1^{(k)}(n)^*E_2^{(k)}(n)E_2^{(k)}(n+e_1)^*.$$

The proof of Proposition IV.7 shows that the random variables

$$\{C(n)=E_2^{(k_0)}(n+e_2)^*E_2^{(k_0)}(n)E_1^{(k_0)}(n)^*E_1^{(k_0)}(n+e_1)\}_{n\in\mathbf{Z}^2}$$

are all independent so that also $\{B^{(k_0\pm l)}(n)\}_{n\in\mathbf{Z}^2}$ is obtained from $\{B^{(k)}(n)\}_{n\in\mathbf{Z}^2}$ by multiplying it with independent identically distributed random variables. The claim follows now from the central limit theorem for independent identically distributed U(1)-valued random variables.[33] (On compact topological groups, the Haar measure plays the role of the Gaussian measure in $\mathbf{R}$) (Fig. 4). □

Proposition V.1 has the following interpretation: a time-dependent random electric field [which might be arbitrarily small but which does not take values in a subgroup of U(1)] turns an initially arbitrary magnetic field for time $|k|\to\infty$ into an independent identically distributed Haar distributed magnetic field.

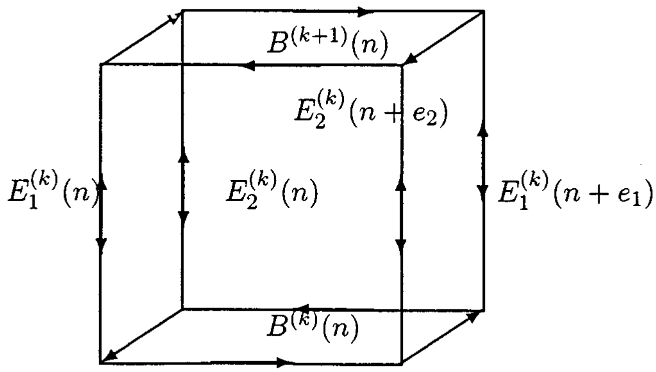

FIG. 4. The Maxwell equation $dF=0$ determines the magnetic field $B^{(k+1)}(n)^*=B^{(k)}(n)E_1^{(k)}(n+e_1)E_1^{(k)}(n)^*E_2^{(k)}(n)E_2^{(k)}(n+e_1)^*$ at time $(k+1)$ from the magnetic field $B^{(k)}$ and the electric field $(E_1^{(k)},E_2^{(k)})$ at time k.

B. One-dimensional operators

Take an electromagnetic Laplacian $L=A+A^*$ in d dimensions, where the electromagnetic field $dA=F$ has only electric components $F_{1k}(n)=E_k(n)$ which are constant in space ($\mathbf{Z}^d=\mathbf{Z}\oplus\mathbf{Z}^{d-1}=$space$\oplus$time) and depend therefore only on the first (=time) coordinate $n=n_1$. The restriction of L to the invariant Hilbert space of functions which are constant in space gives a one-dimensional operator $(Hu)_n=u_{n+1}+u_{n-1}+V(n)u_n$, where

$$V(n)=\sum_{k=1}^{d} E_k(n)+E_k(n)^*=\sum_{k=1}^{d} 2\cos(\arg(E_k(n))).$$

Every one-dimensional operator can be written like this. The number of dimensions which are needed depends on the norm. Since $dF=0$, Feldman–Moore's existence theorem shows that if V is an ergodic potential, then the equation $F=dA$ can be solved with a measurable vector potential A leading to an ergodic electromagnetic Laplacian. The one-dimensional potential $\Sigma_{k=1}^{d} 2\cos(\arg(E_k(n)))$ is ergodic, if T_1 was ergodic.

Some Anderson models can be treated as random magnetic Laplacians and allow a combinatorial calculation of the density of states: given independent identically distributed random variables $V(n)$ $n\in\mathbf{Z}^d$ with law μ, define the $\mathcal{B}(l^2(\mathbf{Z}^d))$-valued random variable $(Lu)_n=\Sigma_{|m-n|=1}u_m+V(n)u_n$ which is an Anderson model. By adding to each vertex of $\mathbf{Z}^d$ an oriented loop, one obtains a new lattice $\mathbf{L}^d$. Denote by Γ_n the set of paths γ in $\mathbf{L}^d$ which have length n. (Each loop has length 1 and we distinguish paths which pass in different directions through the loop) (Fig. 5).

Corollary V.2: (a) Given the discrete d-dimensional Anderson Schrödinger operator with independent identically distributed potential $V(n)=2\cos(\alpha(n))$, *where* $\alpha(n)$ *are uniformly distributed in* $[0,2\pi]$. *The* n'th *moment of the density of states is the number of closed paths of length n in* $\mathbf{L}^d$, *for which every loop has vanishing winding number.*

(b) If $V(n)=\pm 2$, *where* $V(n)$ *are uniformly distributed in* $\{0,2\}$, *the* n'th *moment of the density of states is the number of closed paths of length n* in $\mathbf{L}^d$ *for which every loop has an even winding number.*

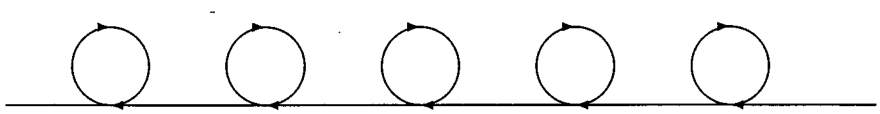

FIG. 5. The graph $\mathbf{L}$ in the case $d=1$. At each vertex is attached an oriented loop.

Proof: Write $L=A+A^*$ as a $(d+1)$-dimensional magnetic Laplacian, where $A_i=1,i=1,\ldots,d$ and $A_{d+1}(n)=\exp(i\alpha(n))$ are independent identically distributed U(1)-valued random variables with uniform Haar distribution μ. This is equivalent to taking real-valued random variables $\alpha(n)$ with uniform distribution on $[0,1]$ and to form the independent identically distributed potential $V(n)=2\cos(2\pi\alpha(n))$ which has an absolutely continuous law $4(2\pi)^{-1}\sqrt{1-x^2}$. As before, we compute with the random walk expansion

$$\mathrm{tr}(L^n)=\sum_{\gamma\in\Gamma_n}\prod_P \hat{\mu}_{n(\gamma,P)}.$$

Since all nonzero moments of ν are zero, $\mathrm{tr}(L^n)$ is the number of closed paths in the lattice $\mathbf{L}^d$ which give in case (a) zero and in case (b) zero $(\mathrm{mod}\,2)$ winding number to every loop. □

Remark: Relations between two- and one-dimensional operators are prototyped by the Harper–Mathieu case $A_1=\tau$, $A_2=e^{2\pi i\alpha}$ which give the one-dimensional operator $\tau+\tau^*+2\cos(2\pi\alpha)$. For more examples with constant magnetic field, see Ref. 34. Other, not constant magnetic fields can be obtained as follows: let $A_1\tau_1=\tau$ be the unitary Koopman operator for a transformation T on a probability space Ω and let $A_2(x)=e^{2\pi f(x)}$, where f is a su(N)-valued random variable. Then $UV=VUe^{2\pi i(f(Tx)-f(x))}$ and we get a one-dimensional operator $L=\tau+\tau^*+2\cos(f(x))$ on $l^2(\mathbf{Z},\mathbf{C}^N)$.

C. Deterministic Aharonov–Bohm Laplacians

It is illustrative to see what deterministic perturbations of the magnetic field does on the operator. We denote by L_F the d-dimensional Laplacian with field F in the special gauge.

Proposition V.3 (Jitomirskaya–Mandelshtam Ref. 18): Assume U is Abelian. A change of $F\in U^{\mathbf{Z}^d}$ *on a finite set of plaquettes leads to a compact perturbation* L_F *of the free Laplacian* L_1.

Proof: Assume first $d=2$. If B is multiplied by $C\in U^{\mathbf{Z}^2}$ such that $C_n\neq 1$ only for finitely many n and $\Pi_n C_n=1$, we call $\tilde{B}=BC$ a zero flux perturbation of 1. It is enough to show the claim for a perturbation of the field B of one single plaquette. By construction, if $\tilde{B}$ is a zero flux perturbation of B, then $L_{\tilde{B}}$ is a finite rank perturbation of L_B.

Let $L=L_B$ be the original operator and let $\tilde{L}=L_{\tilde{B}}$ be the operator belonging to $\tilde{B}$ satisfying $\tilde{B}(n)=B(n)$ for all $n\in\mathbf{Z}^2$ except one n_0, where $\tilde{B}(n_0)=B(n_0)C$ with $C=e^{i\alpha}\in U$. Define for each $k\in\mathbf{N}$ a zero flux perturbation B_k of B by changing $\tilde{B}$ on k^2 plaquettes in a box of size $k\times k$ to $\tilde{B}_kC_k^{-1}$ with $C_k=e^{-i\alpha/n^2}$. Then, $L_{B_k}\to L_{\tilde{B}}$ in norm so that $L_{\tilde{B}}$ is a limit of finite rank operators L_{B_k}.

For general d, we can build any perturbation by composing finitely many perturbations lying in two-dimensional planes and for which the previous argument applies. □

Remarks: (1) The Aharonov–Bohm operator [the situation when the field $B(n)$ is different from 1 exactly at one plaquette] shows that one has never a finite rank perturbation $L_B\mapsto L_{\tilde{B}}$, if $B\tilde{B}^{-1}$ has compact support and nonzero flux. It would be interesting to know if the Aharonov–Bohm operator is a trace class perturbation of the free Laplacian.

(2) There is the following formula for the Fourier transform of the spectral measure $dk_l=dk_{e_l}$, where $e_l(n)=\delta_{ln}$ is a unit vector in $l^2(\mathbf{Z}^2)$:

$$\widehat{dk}_{ln}=\sum_{\gamma\in\Gamma_n} B^{n(\gamma)},$$

where $n(\gamma)$ is the winding number of the path with respect to a point in the plaquette, where B is different from 1.

(3) A similar argument shows that the Jitomirskaya–Mandelstam result is also true for some aperiodic tilings like the Penrose tiling.

(4) Beside the Abelian or non-Abelian Aharonov–Bohm operators (for which a complete spectral analysis is not yet done), other deterministic operators would be interesting to study. An example is a discrete version of the Iwatsuka operator L in $d=2$ (see Ref. 5), where the magnetic field B is translational invariant in one direction and asymptotically constant in the other direction. Then, L is a direct product of one-dimensional operators $(Lu)_n = u_{n+1} + u_{n-1} + \cos(n\alpha(n))u(n)$, where $\alpha(n) \to \alpha^{\pm}$ for constants $\alpha^{\pm}$. If α^- or α^+ is rational, then also L has some absolutely continuous spectrum. If both $\alpha^{\pm}$ are irrational, Last's results[29] allow us to prove that L has no absolutely continuous spectrum. This is different from the continuous case, where the corresponding operator has purely absolutely continuous spectrum.

VI. MAGNETIC LAPLACIANS ON TILINGS AND OTHER LATTICES

A. Magnetic Laplacians on the triangular lattice

The triangular lattice is the Cayley graph of the group $G=\mathbf{Z}^2$ with the three generators e_1, e_2, e_1+e_2. A situation with two different fluxes has been considered in Ref. 35 (see also Ref. 14). A magnetic field is a cocycle which assigns to each triangle $\Delta(g_1, g_2, g_3)$, $g_i \in \mathbf{Z}^2$ a group element in U. This cocycle is determined by the value of $B_d(n)$ on $\Delta(n, n+e_1, n+e_2)$ and $B_u(n)$ on $\Delta(n+e_1, n+e_1+e_2, n+e_2)$ for each $n \in \mathbf{Z}^2$. The two measurable maps $B_d, B_u \in L^{\infty}(X,U)$ and an ergodic $\mathbf{Z}^2$ action so determine the magnetic field.

Proposition VI.1: Every stationary $\mathrm{U}(N)$*-valued field B on a triangular lattice in* $\mathbf{Z}^2$ *is given by a vector potential A so that* $B=dA$. *The spectral properties of L depend only on B. If* $\{B(n)\}_{n \in \mathbf{Z}^2}$ *are independent identically distributed random variables with Haar distribution on* $U=\mathrm{U}(1)$, *then* $\mathrm{tr}(L^n)$ *is the number of closed paths in the triangular lattice which give zero winding number to all triangles.*

Proof: In order to get the vector potential A, we form $B(x)=B_u(x)B_d(x)$, which is the field on the quadratic plaquette $P(x)$. Feldman–Moore–Lind's theorem gives the existence of the first two coordinates (A_1, A_2) of the vector potential. We define then A_3 through $A_3A_2(T_1)A_1=B_d$.

In the Abelian case, a second proof is obtained directly from the algebraic group cohomology for the group $G=\mathbf{Z}^2$ acting on $\mathcal{U}=\mathcal{L}(X,U)$: the magnetic field B with law μ is an algebraic 2-cocycle. Since the second cohomology group is trivial, it is of the form dA, where A is a one-form.

For Abelian U, the random walk expansion is done in the same way as for the square lattice by putting A_2 identically zero.

In order to see that all the spectral properties depend only on the field B, we take the same special gauge as in the square lattice case. □

Remarks: (1) Discrete magnetic Laplacians on more general graphs with uniform magnetic field with values in U(1) have been considered by Sunada.[36]

(2) If the graph G is the Cayley graph of an infinite Abelian group with finitely many generators and $U \subset \mathrm{U}(1)$, the magnetic Laplacians are elements in a hyperfinite von Neumann algebra $\mathcal{X}$. The second group cohomology vanishes and every algebraic cocycle B is of the form $B=dA$.

B. Magnetic Laplacians on aperiodic tilings

Aperiodic tilings in $\mathbf{R}^2$ define a plane graph and one can ask if it is possible to assign to the edges of the graph U(1) random variables in such a way that the magnetic fields in the pieces of the tiling are independent identically distributed U(1)-valued random variables. For simplicity, we consider only the case of the Penrose tiling with plaquettes built by Robinson triangles. The case when the plaquettes are Penrose rhombs can be reduced to that by multiplying the field values of the triangles building the rhomb.

Proposition VI.2: Given a measurable $U=\mathrm{U}(1)$*-valued field distribution B on the Penrose lattice. There exists a vector potential A such that* $dA=B$.

For a proof see Ref. 3. One can deduce from this:

Corollary VI.3: Given an independent identically distributed U(1)*-valued field B on the Robinson triangles of a Penrose tiling. There exists a measurable vector potential A on the edges of the Penrose graph such that* $dA=B$.

Remarks: (1) For more general tilings, where all pieces of the tiling are composed of the same number k of triangles (which is the case in the Penrose tiling where each Penrose rhomb is a union of two Robinson triangles) we can also realize independent identically distributed magnetic field configurations, where the law $\mu=\nu\oplus\cdots\oplus\nu$ is the kth convolution of a measure ν. This is for example the case if μ is the Haar measure on a closed subgroup of U(1).

(2) The existence of the density of states of a magnetic Laplacian L on the tiling follows from the fact L in a finite type von Neumann algebra. Hof[37] has given a direct proof of the existence and proven that the density of states and spectrum is constant on the space of tilings.

(3) For independent identically distributed magnetic fields with Haar measure of U(1), we get that $\mathrm{tr}(L^n)$ is the number of closed paths in the tiling graph such that the winding number is zero for each tile. The computation of the density of states is already nontrivial for the free Laplacian with zero magnetic field. There are some numerical results about the random walk on Penrose lattice.[38]

ACKNOWLEDGMENTS

This paper with a cohomological part was written at Caltech and was posted in *mp_arc* in 1995.[19] The present version is reorganized, updated and the proofs of the cohomological results are subject to a separate paper.[3] I owe valuable remarks or literature information to Y. Avron, A. Barrelli, J.Bellisard, J.-P. Conze, J. Feldman, G. Hjorth, A. Hof, S. Jitomirskaya, Y. Last, D. Lind, Y. Peres, and B. Simon. While preparing the final version, I acknowledge the support of the Swiss National Science Foundation.

[1] J. Feldman and C. Moore, ''Ergodic equivalence relations, cohomology and von Neumann algebras I,II,'' Trans. Am. Math. Soc. **234**, 289–359 (1977).

[2] D. A. Lind, ''Products of coboundaries for commuting nonsingular automorphisms,'' Z. Wahrscheinlichkeitstheor. Verwandte Geb. **43**, 135–139 (1978).

[3] O. Knill, ''On the cohomology of discrete Abelian group actions,'' part is contained in Mathematical Physics Preprint Archive, *http://www.ma.utexas.edu/mp_arc/papers/*95−195, 1995.

[4] R. Carmona and J. Lacroix, *Spectral Theory of Random Schrödinger Operators* (Birkhäuser, Boston, 1990).

[5] H. L. Cycon, R. G. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators—with Application to Quantum Mechanics and Global Geometry* (Springer, Berlin, 1987).

[6] L. Pastur and A. Figotin, *Spectra of Random and Almost-Periodic Operators*, Lecture Notes in Mathematics (Springer, Verlag, 1992), p. 1294.

[7] O. Knill, ''A remark on quantum dynamics,'' Helv. Phys. Acta **71**, 233–241 (1998).

[8] W. F. Brinkman and T. M. Rice, ''Single particle excitations in magnetic insulators,'' Phys. Rev. B **2**, 1324–1338 (1970).

[9] S. V. Meshkov and D. V. Berkov, ''Density of states for one hole in a half filled infinite two dimensional hubbard model,'' Phys. Lett. A **170**, 448–453 (1992).

[10] G. Gavazzi, J. M. Wheatley, and A. J. Schofield, ''Single-particle motion in a random magnetic flux,'' Phys. Rev. B **47**, 15170–15176 (1993).

[11] A. Barelli, R. Fleckinger, and T. Ziman, ''Two-dimensional electron in a random magnetic field,'' Phys. Rev. B **49**, 3340–3349 (1994).

[12] A. Barelli, R. Fleckinger, and T. Ziman, ''Localisation of two dimensional electrons with uniform and random fluxes,'' Europhys. Lett. **27**, 531–536 (1994).

[13] M. A. Shubin, ''Discrete magnetic Laplacians,'' Commun. Math. Phys. **164**, 259–275 (1994).

[14] J. Bellissard, ''Non commutative methods in semiclassical analysis,'' in *Transition to Chaos in Classical and Quantum Mechanics*, edited by S. Graffi (Springer, Berlin, 1994), pp. 1–640.

[15] E. J. Martinec, ''On the origin of integrability in matrix models,'' Commun. Math. Phys. **138**, 437–449 (1991).

[16] O. Knill, ''Random Schrödinger operators arising from lattice gauge fields. II. Determinants'' (unpublished).

[17] J. Avron and B. Simon, ''Almost periodic Schrödinger operators. II. The integrated density of states,'' Duke Math. J. **50**, 369–391 (1983).

[18] S. Ya. Zhitomirskaya and V. A. Mandelshtam, ''Aharonov-Bohm problem on a square lattice,'' Theor. Math. Phys. **86**, 241–604 (1991).

[19] O. Knill, ''Discrete random electromagnetic Laplacians'' (unpublished).

[20] J. Bellissard, ''Le papillon de Hofstadter (d' après B. Helffer and J. Sjöstrand),'' Séminaire Bourbaki, 44 ème année **745**, 7–39 (1991).

[21] Y. Last, ''Almost everything about the almost Mathieu operator. I.,'' in *Proceedings of the 11th International Congress of Mathematical Physics*, Paris, 18–23 July 1994 (International Press, Cambridge, MA, 1995), pp. 366–372.
[22] S. Ya. Jitomirskaya, ''Anderson localization for the almost mathieu equation II, point spectrum for $\lambda \geqslant 2$,'' Commun. Math. Phys. **168**, 201–205 (1995).
[23] A. Katok and S. Katok, ''Higher cohomology for Abelian groups of toral automorphisms,'' Ergod. Th. Dyn. Sys. **15**, 569–592 (1995).
[24] J. DePauw, ''Théorèmes ergodiques pour cocycle de degré 2,'' Thèse de doctorat, Université de Bretagne Occidentale, 1994.
[25] O. Knill, ''Spectral, ergodic and cohomological problems in dynamical systems,'' Ph.D. theses, ETH Zürich, 1993.
[26] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, Orlando, 1980), Vol. I.
[27] F. Constantinecu, J. Fröhlich, and T. Spencer, ''Analyticity of the density of states and replica method for random Schrödinger operators on the lattice,'' J. Stat. Phys. **34**, 571–596 (1984).
[28] L. B. Koralov, S. K. Nechaev, and G. Sinai Ya, ''Asymptotic behavior of a two-dimensional random walk with topological constraints,'' Theor. Probab. Appl. **38**, 296–306 (1993).
[29] Y. Last, ''Zero measure spectrum for the almost Mathieu operator,'' Commun. Math. Phys. **164**, 421–432 (1994).
[30] H. Kesten, ''Symmetric random walks on groups,'' Trans. Am. Math. Soc. **92**, 336–356 (1959).
[31] H. Acosta and A. Klein, ''Analyticity of the density of states in the Anderson model on the Bethe lattice,'' J. Stat. Phys. **65**, 277–305 (1992).
[32] O. Knill, ''Isospectral deformation of discrete random Laplacians,'' in *On Three Levels*, edited by M. Fannes *et al.* (Plenum, New York, 1994), pp. 321–330.
[33] K. V. Mardia, *Statistics of Directional Data* (Academic, London, 1972).
[34] V. A. Mandelshtam and S. Ya. Jitomirskaya, ''1D-quasiperiodic operators. Latent symmetries,'' Commun. Math. Phys. **139**, 589–604 (1991).
[35] J. Bellissard, C. Kreft, and R. Seiler, ''Analysis of the spectrum of a particle on a triangular lattice with two magnetic fluxes by algebraic and numerical methods,'' J. Phys. A **24**, 2339–2353 (1991).
[36] T. Sunada, ''A discrete analogue of periodic magnetic Schrödinger operators,'' in *Geometry of the Spectrum*, edited by R. Brooks and C. Gordon (Amer. Math. Soc., Providence, RI, 1994), *Contemporary Mathematics* Vol. 173, pp. 283–299.
[37] A. Hof, ''A remark on Schrödinger operators on aperiodic tilings,'' J. Stat. Phys. **81**, 851–855 (1995).
[38] B. Sutherland, ''Self-similar ground state wave function for electrons on a two-dimensional Penrose lattice,'' Phys. Rev. B **34**, 3904–3909 (1986).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Nonstandard Feynman path integral for the harmonic oscillator

Ken Loo[a)]
PO Box 9160, Portland, Oregon 97207

(Received 8 April 1998; accepted for publication 13 July 1999)

Using Nonstandard Analysis, we will provide a rigorous computation for the harmonic oscillator Feynman path integral. The computation will be done without having prior knowledge of the classical path. We will see that properties of classical physics falls out naturally from a purely quantum mechanical point of view. We will assume that the reader is familiar with Nonstandard Analysis. © *1999 American Institute of Physics.* [S0022-2488(99)01711-9]

I. INTRODUCTION

In quantum mechanics, we are interested in finding the wave function that satisfies Schrodinger's equation. Equivalently, we can find the propagator or integral kernel $K(q,q_0,t)$ which satisfies

$$i\hbar\frac{\partial K(q,q_0,t)}{\partial t}=\left[\frac{-\hbar^2}{2m}\Delta_q+V(q)\right]K(q,q_0,t),$$
$$K(q,q_0,0)=\delta(q-q_o),\quad q,q_0\in\mathbb{R}^d. \tag{1.1}$$

Formally, the wave function is related to the propagator via

$$\varphi(q,t)=\int_{-\infty}^{+\infty}K(q,q_0,t)\phi(q_0)dq_0=\langle K(q,q_0,t),\phi(q_0)\rangle, \tag{1.2}$$

with boundary condition $\varphi(q,0)=\phi(q)$. In Feynman's formulation of quantum mechanics, he proposed that the propagator is given by a functional integral, also referred to as Feynman path integral or just path integral in physics literature,

$$K(q,q_0,t)=\int_{x(0)=q_0}^{x(t)=q}\exp\left\{\frac{iS[x(s)]}{\hbar}\right\}dx(s)=\lim_{n\to\infty}\int_{R^{dn}}w_{d,n}\exp\left[\frac{i\epsilon}{\hbar}S\{\mathbf{x}_{n+1}\ldots,\mathbf{x}_0\}\right]d\mathbf{x}_1\cdots d\mathbf{x}_n, \tag{1.3}$$

where

$$x_0=q_0,x_{n+1}=q,\epsilon=\frac{t}{n},$$

$$w_{d,n}=\left(\frac{m}{2i\pi\hbar\epsilon}\right)^{d(n+1)/2},$$

$$S\{\mathbf{x}_{n+1}\cdots\mathbf{x}_0\}=\sum_{j=1}^{n+1}\left[\frac{m}{2}\left(\frac{\mathbf{x}_j-\mathbf{x}_{j-1}}{\epsilon}\right)^2-V(\mathbf{x}_j)\right].$$

[a)]Electronic mail: look@sdf.lonestar.org

The integrals in the second line of (1.3) are d-dimensional improper Riemann integrals. The integral in the first equality is the purely formal path integral that integrates over all paths $x(s)$, $0\leqslant s\leqslant t$, with $x(0)=q_0$, and $x(t)=q$. The quantity $S[x(s)]$ is the action integral,

$$S[x(s)]=\int_0^t L(x,\dot{x})ds, \quad L(x,\dot{x})=\frac{m}{2}\dot{x}-V(x). \tag{1.4}$$

The motivation for the notation of the formal path integral is that as n goes to infinity, the sum in the exponent becomes the action integral. In this manner, Feynman was able to deduce classical mechanics from quantum mechanics through the action integral and $\hbar\to 0$. In this paper, we will compute the harmonic oscillator path integral without prior knowledge of classical physics. It turns out that properties of classical physics naturally falls out of the computation without taking $\hbar\to 0$. The computation shows that the path integral separates into the product of two quantities: one independent of $\hbar$, the other dependent on $\hbar$. The quantity that is independent of $\hbar$ contains properties of classical physics. Thus, in some sense we are deviating from the standard interpretation of the path integral being a bridge between quantum mechanics and classical mechanics via the action integral and $\hbar$. We are considering quantum mechanics as purely quantum mechanics and extracting properties of classical mechanics without prior knowledge of classical physics.

Mathematically, the formal integration over paths can not be a rigorously well-defined measure theoretic integration because of the oscillatory nature of the integrand (see Refs. 1 and 2). A popular technique to make sense of the (1.3) is to replace t by $-it$ and use the Wiener integral (see Refs. 3 and 4).

In nonstandard analysis, we have that $\lim_{n\to\infty} a_n=a$ iff ${}^*a_\omega\approx a$ for any infinite natural number $\omega\in{}^*\mathbb{N}-\mathbb{N}$ with $\{{}^*a_m\}_{m\in{}^*\mathbb{N}}$ being the $*$ extension of $\{a_n\}_{n\in\mathbb{N}}$, and $\approx$ means that $a_\omega+h_\omega=a$ where h_ω an infinitesimal. We can use nonstandard analysis to define the path integral; the standard part can replace the limit in (1.3). Using nonstandard analysis to replace the limit in (1.3) is not a new concept (see Refs. 5, 6 and references within), doing so partially solves the problems of the Feynman path integral on the propagator.

We can redefine (1.3) in the following manner: let $\omega\in{}^*\mathbb{N}, m,t\in\mathbb{R}^+$, $\epsilon=t/\omega$, ${}^*V(x):{}^*\mathbb{R}^d\to{}^*\mathbb{R}$ be an internal function, and $x_0=q_0$, $x_{\omega+1}=q$ be fixed points in $\mathbb{R}^d$. We call the expression

$$\int_{{}^*\mathbb{R}^{d\omega}}\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{(d/2)(\omega+1)}\exp\left[\frac{i\epsilon}{\hbar}\sum_{j=1}^{\omega+1}\left[\frac{m}{2}\left(\frac{x_j-x_{j-1}}{\epsilon}\right)^2-{}^*V(x_j)\right]\right]dx_1\cdots dx_\omega, \tag{1.5}$$

an internal functional integral. In (1.5), all integrals are $*$-transformed improper Riemann integrals. In particular, if for all $\omega\in{}^*\mathbb{N}-\mathbb{N}$, the standard part of the internal functional integral exists and it is independent of the choice of ω, we call the standard part a standard functional integral or Feynman path integral and denote it by

$$st\int_{{}^*\mathbb{R}^{d\omega}}\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{(d/2)(\omega+1)}\exp\left[\frac{i\epsilon}{\hbar}\sum_{j=1}^{\omega+1}\left[\frac{m}{2}\left(\frac{x_j-x_{j-1}}{\epsilon}\right)^2-{}^*V(x_j)\right]\right]dx_1\cdots dx_\omega. \tag{1.6}$$

Equation (1.6) is just a nonstandard analysis way of saying that the limit in (1.3) exists. There still remain the problem of for which class of potentials V the expression (1.5) exists and whether (1.6) actually produces the propagator.

We will demonstrate the usage of (1.5) and (1.6) on the d-dimensional harmonic oscillator path integral. The harmonic oscillator plays a major role in quantum field theory and the recent advances due to Duru and Kleinert in the Coulomb potential path integral (see Ref. 7 and references within). The harmonic oscillator carries the potential $V(x)=m\lambda^2/2x^2$, its internal functional integral is:

$$\int_{{}^*\mathbb{R}^{d\omega}}\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{(d/2)(\omega+1)}\exp\left[\frac{i\epsilon}{\hbar}\sum_{j=1}^{\omega+1}\left[\frac{m}{2}\left(\frac{x_j-x_{j-1}}{\epsilon}\right)^2-\frac{m}{2}\lambda^2x_j^2\right]\right]dx_1\ldots dx_\omega. \tag{1.7}$$

Notice that each *d*-dimensional integral factors into *d* products of one-dimensional integrals, thous we shall compute the one-dimensional harmonic oscillator internal functional integral.

II. THE HARMONIC OSCILLATOR

There are many ways to compute the harmonic oscillator path integral (see Refs. 7–9 and references within), we will use a method of computation similar to that of Ref. 8. We differ from the popular techniques in that we will do the computation without prior knowledge of the classical path of the harmonic oscillator and we will rigorously do the computation with nonstandard analysis. As pointed out earlier, it turns out that properties of classical physics falls out naturally from a purely quantum mechanical derivation.

It is well known that for $0<t<\pi/\lambda$, the propagator for the one-dimensional harmonic oscillator is

$$K(q,q_0,t)=\left(\frac{m}{2\pi i\hbar}\right)^{1/2}\sqrt{\frac{\lambda}{\sin\lambda t}}\exp\left\{\frac{im}{\hbar}\frac{\lambda}{\sin\lambda t}[(q_0^2+q^2)\cos\lambda t-2qq_0]\right\}. \tag{2.1}$$

Equation (2.1) carries a singularity at $t=\pi/\lambda$. A singularity like that of (2.1) can, in fact, be given rigorous mathematical meaning if $K(q,q_0,t)$ is interpreted as a distribution (see Ref. 10). Due to the form of (2.1), we would not expect the standard functional integral to exist at $t=\pi/\lambda$ as a function.

The popular method to compute the one-dimensional harmonic oscillator Feynman path integral is by writing $S[x(s)]=S[x^{\mathrm{cl}}(s)+\delta x(s)]$, where $x^{\mathrm{cl}}(s)$ is the classical path of the harmonic oscillator that satisfies the equation of motions $\ddot{x}^{\mathrm{cl}}(s)=-\lambda^2x^{\mathrm{cl}}(s)$, with the boundary condition $x^{\mathrm{cl}}(0)=q_0$, $x^{\mathrm{cl}}(t)=q$. Namely, $x^{\mathrm{cl}}(s)=[q\sin\lambda s+q_0\sin\lambda(t-s)]/\sin\lambda t$. The action integral becomes

$$\int_0^t\frac{m}{2}[(\dot{x}^{\mathrm{cl}})^2-\lambda^2(x^{\mathrm{cl}})^2]ds+\int_0^t\frac{m}{2}[(\delta\dot{x})^2-\lambda^2(\delta x)^2]ds+\int_0^t m(\dot{x}^{\mathrm{cl}}\,\delta\dot{x}-\lambda^2x^{\mathrm{cl}}\,\delta x)ds. \tag{2.2}$$

Using $\delta x(0)=0=\delta x(t)$ and the equation of motions of x^{cl}, the last integral is 0 after an integration by parts. Integrating by parts on the first integral and using the equation of motions of x^{cl}, the first integral becomes $(m/2)x^{\mathrm{cl}}\dot{x}^{\mathrm{cl}}|_0^t=(m\lambda/2\sin\lambda t)[(q^2+q_0^2)\cos t-2qq_0]$. Without much concern on the existence and meaning of the path integral, we can write

$$\begin{aligned}&\int_{x(0)=q_0}^{x(t)=q}\exp\left\{\frac{iS[x(s)]}{\hbar}\right\}dx(s)\\&\quad=\exp\left\{\frac{im\lambda}{2\hbar\sin\lambda t}[(q^2+q_0^2)\cos t-2qq_0]\right\}\int_{\delta x(0)=0}^{\delta x(t)=0}\exp\left\{\frac{iS[\delta x(s)]}{\hbar}\right\}d\,\delta x(s).\end{aligned} \tag{2.3}$$

The path integral on the right-hand side is the quantum fluctuation; the integral is over all paths, which starts from 0 at $s=0$, and ends at 0 at $s=t$. We leave it to the reader to look up the computation of the quantum fluctuation in the literature.

We will give a rigorous treatment of (2.3) by using the time-sliced internal path integral in (1.5) and (1.6). In our work, we do not start with having knowledge of the classical path. We will start with an arbitrary bounded path $w(s)$ which satisfies $w(0)=q_0$, $w(t)=q$, and separate the propagator into a product of classical and quantum amplitudes. In this approach, we will see that the classical contribution actually comes from quantum mechanics without prior knowledge of classical mechanics.

To shorten the notation, we write

$$
\begin{aligned}
&\frac{i\epsilon}{\hbar}\sum_{j=1}^{n+1}\left[\frac{m}{2}\left(\frac{x_j-x_{j-1}}{\epsilon}\right)^2-\frac{m}{2}\lambda^2x_j^2\right]\\
&=\left(\frac{im}{2\hbar\epsilon}\right)\left[x_0^2-2x_0x_1+x_{n+1}^2-2x_nx_{n+1}+\sum_{j=1}^{n}2x_j^2-\sum_{j=1}^{n}2x_jx_{j-1}-\epsilon^2\lambda^2\sum_{j=1}^{n+1}x_j^2\right]\\
&=\left(\frac{im}{2\hbar\epsilon}\right)x^t\left\{\begin{pmatrix}
1 & -1 & 0 & \cdots & 0\\
-1 & & & & \vdots\\
0 & & 0 & & 0\\
\vdots & & & & -1\\
0 & \cdots & & -1 & 1
\end{pmatrix}\right.\\
&\quad-\epsilon^2\lambda^2\begin{pmatrix}
0 & \cdots & & & & \cdots & \cdots & 0\\
0 & 1 & 0 & \cdots & & & \cdots & 0\\
\vdots & \ddots & \ddots & \ddots & \ddots & & & \vdots\\
 & & \ddots & \ddots & \ddots & & & \\
 & & & \ddots & \ddots & \ddots & & \\
\vdots & & & & \ddots & \ddots & & \vdots\\
0 & \cdots & & & \cdots & 0 & 1 & 0\\
0 & \cdots & & & & \cdots & 0 & 1
\end{pmatrix}\\
&\quad\left.+\begin{pmatrix}
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0\\
0 & 2 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0\\
\vdots & -1 & 2 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & 0\\
\vdots & 0 & -1 & 2 & -1 & 0 & \cdots & \cdots & \cdots & 0\\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots\\
\vdots & \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots\\
0 & & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 2 & -1 & 0\\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 2 & 0\\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0
\end{pmatrix}\right\}x=\left(\frac{im}{2\hbar\epsilon}\right)(x^tT_nx),
\end{aligned}
\tag{2.4}
$$

where T_n is the $(n+2)$ by $(n+2)$ symmetric matrix,

$$
T_n=\begin{pmatrix}
1 & -1 & 0 & \cdots & 0\\
-1 & & & & \vdots\\
0 & & S_n & & 0\\
\vdots & & & & -1\\
0 & \cdots & & -1 & 1-\epsilon^2\lambda^2
\end{pmatrix},
\tag{2.5}
$$

with S_n being the n by n symmetric matrix $S_n = A_n - \epsilon^2 \lambda^2 B_n$, where

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix},$$

$$B_n = \begin{pmatrix} 1 & 0 & \cdots & & & & \cdots & 0 \\ 0 & 1 & 0 & \cdots & & & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ \vdots & & & & \ddots & \ddots & & \vdots \\ 0 & \cdots & & & \cdots & 0 & 1 & 0 \\ 0 & \cdots & & & & \cdots & 0 & 1 \end{pmatrix}, \tag{2.6}$$

and x is the column vector,

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \\ x_{n+1} \end{pmatrix}. \tag{2.7}$$

For notation convenience, we will use a bar instead of $*$ to indicate the $*$ transforms of matrices, determinate of matrices, and vectors.

We are interested in knowing when the internal functional integral of the harmonic oscillator exists. From (2.1), it would be reasonable to postulate the existence of the functional integral for $t < \pi/\lambda$. Indeed, this turns out to be the case. For $t < \pi/\lambda$, $\overline{S_\omega}$ (the $*$ transform of S_n with $n = \omega \in {}^*\mathbb{N} - \mathbb{N}$) turns out to be $*$-positive definite, which allows us to actually compute the integrals.

Proposition 2.1: For $0 < t < \sqrt{[n^2\pi^2/\lambda^2(n+1)^2][1-\pi^2/12(n+1)^2]}$, S_n *is positive definite.*

Proof: An elementary computation shows that for $k = 1,2,...,n$,

$$A_n \begin{pmatrix} \sin\frac{k\pi}{n+1} \\ \sin\frac{2k\pi}{n+1} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \sin\frac{nk\pi}{n+1} \end{pmatrix} = \left(2 - 2\cos\frac{k\pi}{n+1}\right) \begin{pmatrix} \sin\frac{k\pi}{n+1} \\ \sin\frac{2k\pi}{n+1} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \sin\frac{nk\pi}{n+1} \end{pmatrix}. \tag{2.8}$$

Hence, the n distinct eigenvalues of S_n are $2-2\cos(k\pi/(n+1))-\lambda^2(t/n)^2$. To show that S_n is positive definite, it is enough to find the values of t for which the eigenvalues are positive, or $\cos(k\pi/(n+1))<1-\lambda^2t^2/2n^2$. Since $\cos(k\pi/(n+1))\leqslant\cos(\pi/(n+1))$ for $k=1,2,...,n$, it is enough to find t for which $\cos(\pi/(n+1))<1-\lambda^2t^2/2n^2$. By Taylor expanding $\cos(\pi/(n+1))$ to about 0 for the first three nonzero terms, we have

$$\cos\left(\frac{\pi}{n+1}\right)<1-\frac{\lambda^2t^2}{2n^2}\Leftrightarrow 1-\frac{\pi^2}{2(n+1)^2}+\frac{\pi^4\cos\eta}{4!(n+1)^4}$$
$$<1-\frac{\lambda^2t^2}{2n^2}\Leftrightarrow t^2<\frac{\pi^2n^2}{\lambda^2(n+1)^2}\left[1-\frac{\pi^2\cos\eta}{12(n+1)^2}\right], \tag{2.9}$$

where $0<\eta\leqslant\pi/(n+1)$. When $t<\sqrt{[n^2\pi^2/\lambda^2(n+1)^2][1-\pi^2/12(n+1)^2]}$, we have $t^2<[\pi^2n^2/\lambda^2(n+1)^2][1-\pi^2\cos\eta/12(n+1)^2]$. □

Theorem 2.2: *Let $t\in\mathbb{R}$ and $0<t<\pi/\lambda$. For any $\omega\in{}^*\mathbb{N}-\mathbb{N}$, $\overline{S_\omega}$ is positive definite in the $*$ transformed sense.*

Proof: $*$ transforming Proposition 2.1 and setting $n=\omega$, we have that $\overline{S_\omega}$ is positive definite when

$$0<t<\sqrt{\frac{\omega^2\pi^2}{\lambda^2(\omega+1)^2}\left[1-\frac{\pi}{12(\omega+1)^2}\right]}=\frac{\pi}{\lambda}+h, \tag{2.10}$$

where h is infinitesimal. When t is standard and $0<t<\pi/\lambda$, (2.10) holds. □

From here on, let us take $0<t<\pi/\lambda$. We will now proceed to separate the functional integral into a classical part and a quantum fluctuation part. Suppose $w(s)$ is an arbitrary path with $|w(s)|<\infty$ for $0\leqslant s\leqslant t$. Furthermore, let $w(0)=q_0$, and $w(t)=q$. We make the substitution $x_j=w(jt/(n+1))+y_j=w_j+y_j$ [notice that $y_0=0=y_{n+1}$ since $w(0)=x_0=q_0$ and $w(t)=x_{n+1}=q$]. Using the fact that T_n is symmetric, we have

$$\begin{aligned}x^tT_nx&=(y+w)^tT_n(y+w)\\&=w^tT_nw+y^tT_ny+w^tT_ny+y^tT_nw\\&=w^tT_nw+y^tT_ny+(T_nw)^ty+(w^tT_ny)^t\\&=w^tT_nw+y^tT_ny+(T_nw)^ty+w^tT_ny\\&=w^tT_nw+y^tT_ny+2(T_nw)^ty.\end{aligned} \tag{2.11}$$

By using $y_0=0=y_{n+1}$ and writing T_n as

$$T_n=\begin{pmatrix}0&0&0&\cdots&0\\0&&&&\vdots\\0&&S_n&&0\\\vdots&&&&0\\0&\cdots&&0&0\end{pmatrix}+\begin{pmatrix}1&-1&0&\cdots&&0\\-1&&&&&\vdots\\0&&0&&&0\\\vdots&&&&&-1\\0&\cdots&&&-1&1-\epsilon^2\lambda^2\end{pmatrix}, \tag{2.12}$$

we obtain

$$x^tT_nx=w^tT_nw+y^tT_ny+2(T_nw)^ty=w^tT_nw+\hat{y}^tS_n\hat{y}+2\rho^t\hat{y}, \tag{2.13}$$

where

$$y=\begin{pmatrix}0\\ y_1\\ \vdots\\ y_n\\ 0\end{pmatrix},\quad \hat{y}=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix},\quad w=\begin{pmatrix}w_0\\ w_1\\ \vdots\\ w_n\\ w_{n+1}\end{pmatrix},\tag{2.14}$$

and

$$\rho=S_n\begin{pmatrix}w_1\\ w_2\\ \vdots\\ w_{n-1}\\ w_n\end{pmatrix}-\begin{pmatrix}w_0\\ 0\\ \vdots\\ 0\\ w_{n+1}\end{pmatrix}=S_n\hat{w}-\hat{\hat{w}}.$$

We then have the following.

Lemma 2.3: Under the assumption on t, let $\epsilon=t/n$, *then*

$$\int_{\mathbb{R}^n}\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{(1/2)(n+1)}\exp\left[\frac{i\epsilon}{\hbar}\sum_{j=1}^{n+1}\left[\frac{m}{2}\left(\frac{x_j-x_{j-1}}{\epsilon}\right)^2-\frac{m}{2}\lambda^2x_j^2\right]\right]dx_1\cdots dx_n$$
$$=\exp\left[\frac{im}{\hbar\epsilon}(w^tT_nw-\rho^tS_n^{-1}\rho)\right]\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{1/2}\sqrt{\frac{1}{\det S_n}}.\tag{2.15}$$

Proof: Since S_n is positive definite, it is invertible. Since S_n is symmetric, the following holds:

$$\hat{y}^tS_n\hat{y}+2\rho^t\hat{y}=(\hat{y}+S_n^{-1}\rho)^tS_n(\hat{y}+S_n^{-1}\rho)-\rho^tS^{-1}\rho.\tag{2.16}$$

Using our shortened notation in (2.13) and (2.16), the integrals in (2.15) is equivalent to

$$\exp\left[\frac{im}{\hbar\epsilon}(w^tT_nw-\rho^tS_n^{-1}\rho)\right]\int_{\mathbb{R}^n}\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{(1/2)(n+1)}\exp\left(\frac{im}{2\hbar\epsilon}z^tS_nz\right)dz_1\cdots dz_n.\tag{2.17}$$

In obtaining (2.17), we performed the change of variables $x_j=w_j+y_j$, and then from $y_j+(S_n^{-1}\rho)_j=z_j$. We get (2.15) after diagonalizing S_n and doing the decoupled integrals. □

Corollary 2.4: Under the previous definition of $w(s)$, *The one-dimensional harmonic oscillator internal functional integral is well defined, and it is equal to*

$$\exp\left[\frac{im}{\hbar\epsilon}(\overline{w}^t\overline{T}_\omega\overline{w}-\overline{\rho}^t\overline{S}_\omega^{-1}\overline{\rho})\right]\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{1/2}\sqrt{\frac{1}{\det S_\omega}}.\tag{2.18}$$

where the bars denote a $*$ *transform.*

Proof: This is just the $*$ transform, Lemma 2.3. □

Remark 2.1: There is no restriction on the choice of the path w except that it starts at q_0 and ends at q. We will show that the exponential part of (2.18) turns out to be the classical amplitude of the previous formal calculation of the propagator and the other factor is the quantum fluctuation. If we choose $\overline{w}$ to be the $*$ transform of the classical path, it can be shown that each entry in $\overline{\rho}$ is infinitesimal, and $\overline{w}^t\overline{T}_\omega\overline{w}/\epsilon$ is infinitesimally close to $\lambda/\sin\lambda t[(q_0^2+q^2)\cos\lambda t-2qq_0]$.

There are many techniques to compute the quantum fluctuation $\lim_{n\to\infty}\sqrt{1/\epsilon\det S_n}$ in the literature; we present a rigorous method to compute the limit with Nonstandard Analysis by computing $\mathrm{st}\sqrt{1/\epsilon\det S_\omega}$, for $\omega\epsilon{}^*\mathbb{N}-\mathbb{N}$.

Proposition 2.5: Let $\epsilon=t/n$. *Denote* $A_{j,n}$ *and* $C_{j,n}$ *with* $0<j\leqslant n$ *to be the j by j matrices given by*

$$A_{j,n}=\begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}, \tag{2.19}$$

and

$$C_{j,n}=\begin{pmatrix} 1 & 0 & \cdots & & & & \cdots & 0 \\ 0 & 1 & 0 & \cdots & & & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ \vdots & & & & \ddots & \ddots & & \vdots \\ 0 & \cdots & & & \cdots & 0 & 1 & 0 \\ 0 & \cdots & & & & \cdots & 0 & 0 \end{pmatrix}.$$

Define $D_{j,n}=\det\|A_{j,n}-\epsilon^2\lambda^2 C_n\|$. *After* $*$ *transforming, we have that for* $\omega\in{}^*\mathbb{N}-\mathbb{N}$, $\overline{D}_{k,\omega}\approx{}^*\cos(kt\lambda/\omega)={}^*\cos(k\epsilon\lambda)$.

Proof: For $k=1$, $A_{1,\omega}=(1)$, $C_{1,\omega}=(0)$, and $D_{1,\omega}=1$. For $k=2$,

$$\overline{A}_{2,\omega}=\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \quad \overline{C}_{2,\omega}=\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.20}$$

and $\overline{D}_{2,\omega}=1-\epsilon^2\lambda^2$. Hence, the claim is true for $k=1,2$. We expand $D_{j,n}$ on the top row, and get the recursion relation

$$D_{j,n}=(2-\epsilon^2\lambda^2)D_{j-1,n}-D_{j-2,n}, \quad 2<j\leqslant n. \tag{2.21}$$

Since we are interested in $\omega\in{}^*\mathbb{N}-\mathbb{N}$, we will consider the cases for which $4-(2-\lambda^2\epsilon^2)^2>0$. To solve the difference equation, we substitute $D_{j,n}=Aa^{j-1}$ into (2.21) and get $a^2-(2-\lambda^2\epsilon^2)a+1=0$ with solutions

$$a_{\pm}=\frac{(2-\lambda^2\epsilon^2)\pm i\sqrt{4-(2-\lambda^2\epsilon^2)^2}}{2}. \tag{2.22}$$

Both solutions $a_{\pm}$ have norm 1. Thus, we can denote $a_{\pm}=e^{i\pm\theta}$, where $\theta=\arg(a_+)>0$. Hence,

$$D_{j,n}=A^+\exp\{i(j-1)\theta\}+A^-\exp\{-i(j-1)\theta\}. \tag{2.23}$$

Solving for the initial conditions,

$$D_1=1=A^++A^-, \quad D_2=1-\epsilon^2\lambda^2=A^+a_++A^-a_-, \tag{2.24}$$

we get

$$A^{\pm}=\frac{1}{2}\pm i\left(\frac{\lambda\epsilon}{2\sqrt{4-\lambda^2\epsilon^2}}\right). \tag{2.25}$$

We now proceed to get an estimate for θ. By definition of θ, $\cos\theta=1-\lambda^2\epsilon^2/2$. After expanding the left-hand side about 0, we get $1-(\theta^2/2)\cos\eta=1-\lambda^2\epsilon^2/2$, $0<\eta\leq\theta$. To estimate θ by $\lambda\epsilon$, we write $\theta=\lambda\epsilon+\phi$, and obtain

$$1-\frac{(\lambda\epsilon+\phi)^2}{2}\cos\eta=1-\frac{\lambda^2\epsilon^2}{2},\quad \Rightarrow\phi=-\lambda\epsilon\pm\frac{\lambda\epsilon}{\sqrt{\cos\eta}},\Rightarrow\phi=-\lambda\epsilon+\frac{\lambda\epsilon}{\sqrt{\cos\eta}}, \tag{2.26}$$

where the last implication is due to $\theta>0$. Thus, $\theta=\lambda\epsilon-\lambda\epsilon(1-1/\sqrt{\cos\eta})$.

By $*$ transforming the above and setting $j=k$, $n=\omega$, $2<k\leq\omega$, $\omega\in{}^*\mathbb{N}-\mathbb{N}$, we get

$$\epsilon=\frac{t}{\omega}\approx 0,\quad \theta,\eta\approx 0,\quad 1-\frac{1}{\sqrt{{}^*\cos\eta}}\approx 0,\quad A^{+}\approx A^{-}\approx\frac{1}{2}. \tag{2.27}$$

Equations (2.27) and (2.23) imply that

$$\overline{D}_{k,\omega}\approx{}^*\cos\{(k-1)\theta\}={}^*\cos\left\{\left(k\lambda\epsilon-\theta-k\lambda\epsilon\left(1-\frac{1}{\sqrt{{}^*\cos\eta}}\right)\right)\right\}\approx{}^*\cos(k\lambda\epsilon). \tag{2.28}$$

In the last $\approx$ in (2.28), we used the fact that the cosine function is uniformly continuous, which translates into $*\cos x\approx{}^*\cos y$ whenever $x\approx y$ in the language of Nonstandard Analysis. □

With the aid of Proposition 2.5, we can show the following.

Theorem 2.6: $\epsilon\det\overline{S_\omega}$ *is infinitesimally close to* $\sin(\lambda t)/\lambda$.

Proof: As before, we will use bars to denote the $*$ transform of the determinant of matrices. Let $S_{j,n}=\det\|A_{j,n}-\epsilon^2\lambda^2B_{j,n}\|$, where $A_{j,n}$ is a j by j matrix as defined in (2.19), $B_{j,n}$ is the j by j identity matrix, $\epsilon=t/n$, and $1\leq j\leq n$. Notice that $S_{n,n}=\det\|S_n\|$. Expanding $S_{j,n}$ and $D_{j,n}$ on the bottom row, we get the recursion relation $S_{j,n}=D_{j,n}+(1-\epsilon^2\lambda^2)S_{j-1,n}$, or, equivalently, $S_{j,n}-S_{j-1,n}=D_{j,n}-\epsilon^2\lambda^2S_{j-1,n}$. Summing the last equality gives

$$S_{n,n}-S_{1,n}=\sum_{j=2}^{n}D_{j,n}-\epsilon^2\lambda^2\sum_{j=2}^{n}S_{j-1,n},$$

$$\Rightarrow\epsilon S_{n,n}=\epsilon S_{1,n}+\sum_{j=2}^{n}\epsilon D_{j,n}-\epsilon^2\lambda^2\sum_{j=2}^{n}\epsilon S_{j-1,n}. \tag{2.29}$$

From the recursion relation, we also get that for $j\geq 3$, $|S_{j-1,n}|\leq(\Sigma_{k=2}^{j-1}|D_{k,n}|)+|S_{1,n}|$.

We now $*$ transform the second equation in (2.29) and write from Proposition (2.5) $\overline{D}_{m,\omega}={}^*\cos(m\epsilon\lambda)+h_m$, where h_m is infinitesimal. We get

$$\epsilon\det\overline{S}_\omega=\epsilon\overline{S}_{1,\omega}+\sum_{m=1}^{\omega}\epsilon\,{}^*\cos\left(\frac{mt\lambda}{\omega}\right)-\epsilon\,{}^*\cos(\epsilon\lambda)+\epsilon\sum_{m=2}^{\omega}h_m-\epsilon^2\lambda^2\sum_{m=2}^{\omega}\epsilon\overline{S}_{m-1,\omega}. \tag{2.30}$$

The set $\{h_m|1\leq m\leq\omega\}$ is an internal set, so $\max_{1\leq m\leq\omega}h_m\in\{h_m|1\leq m\leq\omega\}$, and it is an infinitesimal. Thus, $\epsilon\Sigma_{m=2}^{\omega}h_m\leq(t/\omega)\omega\max_{1\leq m\leq\omega}h_m\approx 0$. From the bound on $|S_{j-1,n}|$, we get

$$\left|\epsilon^2\lambda^2\sum_{m=2}^{\omega}\epsilon\overline{S}_{m-1,\omega}\right|\leqslant\epsilon^3\lambda^2|\overline{S}_{1,\omega}|+\epsilon^2\lambda^2\left\{\sum_{m=3}^{\omega}\epsilon\left[\left(\sum_{k=2}^{m-1}|\overline{D}_{k,\omega}|\right)+|\overline{S}_{1,\omega}|\right]\right\}$$
$$<\epsilon^3\lambda^2|\overline{S}_{1,\omega}|+\epsilon^2\lambda^2\left[\left(\sum_{m=2}^{\omega}2t\right)+t|\overline{S}_{1,\omega}|\right]\approx 0. \tag{2.31}$$

Finally, using the limit of Riemann sums in the language of nonstandard analysis, we have $\Sigma_{m=1}^{\omega}\epsilon*\cos(mt\lambda/\omega)\approx\int_0^t\cos(\lambda s)ds=\sin(\lambda t)/\lambda$. Since the other two terms in (2.30) are also infinitesimals, the result follows. □

We are now ready to derive the classical amplitude from the exponential in (2.18) by using results from Theorem 2.6.

Proposition 2.7: The exponential in (2.18) satisfies the following:

$$\exp\left[\frac{im}{\hbar\epsilon}(\overline{\omega}^t\overline{T}_\omega\overline{\omega}-\overline{\rho}^t\overline{S}_\omega^{-1}\overline{\rho})\right]\approx\exp\left\{\frac{im}{\hbar}\frac{\lambda}{\sin\lambda t}[(q_0^2+q^2)\cos\lambda t-2qq_0]\right\}, \tag{2.32}$$

where $q_0=x_0$, *and* $q=x_{\omega+1}$.

Proof: From the definition of ρ, w, T_n, and S_n^{-1} in (2.12)–(2.14), we multiply out $\rho^tS_n^{-1}\rho$ and express it in terms of w^tT_nw,

$$\begin{aligned}\rho^tS_n^{-1}\rho&=(\hat{w}^tS_n-\hat{\hat{w}}^t)S_n^{-1}(S_n\hat{w}-\hat{\hat{w}})\\&=(\hat{w}^t-\hat{\hat{w}}^tS_n^{-1})(S_n\hat{w}-\hat{\hat{w}})\\&=\hat{w}^tS_n\hat{w}-\hat{w}^t\hat{\hat{w}}-\hat{\hat{w}}^t\hat{w}+\hat{\hat{w}}^tS_n^{-1}\hat{\hat{w}}\\&=w^tT_nw-w_0^2-(1-\epsilon^2\lambda^2)w_{n+1}^2\\&\quad+w_0^2(\overline{S_n^{-1}})_{11}+w_0w_{n+1}(\overline{S_n^{-1}})_{1n}\\&\quad+w_0w_{n+1}(\overline{S_n^{-1}})_{n1}+w_{n+1}^2(\overline{S_n^{-1}})_{nn}.\end{aligned} \tag{2.33}$$

After $*$ transforming (2.33), we get

$$\begin{aligned}\frac{1}{\epsilon}(\overline{w}^t\overline{T}_\omega\overline{w}-\overline{\rho}^t\overline{S}_\omega^{-1}\overline{\rho})&=\frac{1}{\epsilon}[\overline{w}_0^2+(1-\epsilon^2\lambda^2)\overline{w}_{\omega+1}^2-\overline{w}_0^2(\overline{S_\omega^{-1}})_{11}\\&\quad-\overline{w}_{\omega+1}^2(\overline{S_\omega^{-1}})_{\omega\omega}-\overline{w}_0\overline{w}_{\omega+1}(\overline{S_\omega^{-1}})_{1\omega}-\overline{w}_0\overline{w}_{\omega+1}(\overline{S_\omega^{-1}})_{\omega1}]\\&=\frac{1}{\epsilon}\left[q_0^2\left(1-\frac{\overline{S}_{\omega-1,\omega}}{\overline{S}_{\omega,\omega}}\right)+q^2\left(1-\frac{\overline{S}_{\omega-1,\omega}}{\overline{S}_{\omega,\omega}}\right)\right.\\&\quad\left.-2qq_0\frac{(-1)^{\omega+1}(-1)^{\omega-1}}{\overline{S}_{\omega,\omega}}-\epsilon^2\lambda^2q^2\right]\\&=\frac{1}{\epsilon\det\overline{S}_\omega}[q_0^2(\overline{S}_{\omega,\omega}-\overline{S}_{\omega-1,\omega})+q^2(\overline{S}_{\omega,\omega}-\overline{S}_{\omega-1,\omega})-2qq_0]-\epsilon\lambda^2q^2\\&\approx\frac{\lambda}{\sin\lambda t}[(q_0^2+q^2)\cos\lambda t-2qq_0].\end{aligned} \tag{2.34}$$

In the third line above, we used $*$-Cramer's rule and our previous definition of $\overline{S}_{k,\omega}$. The factor $(-1)^{\omega+1}$ comes from a cofactor expansion in the numerator of Cramer's rule for $(S_\omega^{-1})_{1\omega}$ and $(S_\omega^{-1})_{\omega1}$; the factor $(-1)^{\omega-1}$ comes from the determinant of a triangular matrix with -1's along the diagonal after the latter expansion. In the fourth line, we used results from Theorem 2; namely, $\overline{S}_{\omega,\omega}=\det\overline{S}_{\omega}$, $\epsilon\det\overline{S}_{\omega}\approx\sin\lambda t/\lambda$, and $\overline{S}_{\omega,\omega}-\overline{S}_{\omega-1,\omega}\approx\overline{D}_{\omega,\omega}\approx\cos\lambda t$. □

Notice that in (2.34), the end result depends only on the end points of the path w. It does not matter which path is chosen as long as it starts at q and ends at q_0. Hence, it is not necessary to use the classical path x^{cl} to do the computation.

Theorem 2.8: *For $t<\pi/\lambda$, the one-dimensional harmonic oscillator standard functional integral is given by*

$$\left(\frac{m}{2\pi i\hbar}\right)^{1/2}\sqrt{\frac{\lambda}{\sin\lambda t}}\exp\left\{\frac{im}{\hbar}\,\frac{\lambda}{\sin\lambda t}\left[(q_0^2+q^2)\cos\lambda t-2qq_0\right]\right\}. \tag{2.35}$$

Proof: This follows from Corollary 2.4, Theorem 2.6, and Proposition 2.7. □

Corollary 2.9: For the d-dimensional harmonic oscillator standard functional integral, we have

$$st\left\{\int_{*\mathbb{R}^{d\omega}}\left(\frac{m}{2\pi i\hbar\epsilon}\right)^{(d/2)(\omega+1)}*\exp\left[\frac{i\epsilon}{\hbar}\sum_{j=1}^{\omega+1}\left[\frac{m}{2}\left(\frac{x_j-x_{j-1}}{\epsilon}\right)^2-\frac{m}{2}\lambda^2x_j^2\right]\right]dx_1\cdots dx_\omega\right\}$$

$$=\left(\frac{m}{2\pi i\hbar}\right)^{d/2}\left(\frac{\lambda}{\sin\lambda t}\right)^{d/2}\exp\left\{\frac{im}{\hbar}\,\frac{\lambda}{\sin\lambda t}\left[(\mathbf{q}_0^2+\mathbf{q}^{-2})\cos\lambda t-2\mathbf{q}\mathbf{q}_0\right]\right\}. \tag{2.36}$$

Proof: Follows from factoring (2.36) into products of one-dimensional harmonic oscillator standard functional integrals and Theorem 2.8. □

ACKNOWLEDGMENT

The author would like to give thanks to Photine Tsoukalas.

[1] R. Cameron, ''A family of integrals serving to connect the Wiener and Feynman integrals,'' J. Math. Phys. **39**, 126–141 (1961).

[2] R. Cameron, ''The Ilstow and Feynman integrals,'' J. Anal. Math **10**, 287–361 (1962/1963).

[3] J. Glimm and A. Jaffe, *Quantum Physics, A Functional Integral Point of View*, 2nd. ed. (Springer-Verlag, Berlin, 1987).

[4] M. Reed and B. Simon, *Functional Analysis I and II* (Academic, New York, 1980).

[5] S. Albeverio, J. Fenstat, R. Høegh-Krohn, and T. Lindstrøm, *Nonstandard Methods in Stochastic Analysis and Mathematical Physics* (Academic, New York, 1986).

[6] K. Loo, ''A rigorous real time Feynman path integral,'' J. Math. Phys. **40**, 1,64–70 (1999).

[7] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, and Polymer Physics* (World Scientific, Singapore, 1995).

[8] I. M. Gelfand and A. M. Yaglom, ''Integration in functional spaces,'' J. Math. Phys. **1**, 48–49 (1960).

[9] L. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981).

[10] S. Zelditch, ''Reconstruction of singularities for solutions of Schrodinger's equation,'' Commun. Math. Phys. **93**, 1–26 (1984).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Bose–Einstein condensation in an external potential at zero temperature: Solitary-wave theory

Dionisios Margetis[a)]
Gordon McKay Laboratory, Harvard University, Cambridge, Massachusetts 02138-2901

(Received 28 June 1999; accepted for publication 26 July 1999)

For a trapped, dilute atomic gas of short-range, repulsive interactions at extremely low temperatures, when Bose–Einstein condensation is nearly complete, some special forms of the time-dependent condensate wave function and the pair-excitation function, the latter being responsible for phonon creation, are investigated. Specifically, (i) a class of external potentials $V_e(\mathbf{r},t)$ that allow for localized, shape-preserving solutions to the nonlinear Schrödinger equation for the condensate wave function, each recognized as a solitary wave moving along an arbitrary trajectory, is derived and analyzed in any number of space dimensions; and (ii) for any such external potential and condensate wave function, the nonlinear integro-differential equation for the pair-excitation function is shown to admit solutions of the same nature. Approximate analytical results are presented for a sufficiently slowly varying trapping potential. Numerical results are obtained for the condensate wave function when V_e is a time-independent, spherically symmetric harmonic potential. © *1999 American Institute of Physics.* [S0022-2488(99)03211-9]

I. INTRODUCTION

The first successful experiments on Bose–Einstein condensation in dilute atomic gases were reported recently by the groups at JILA,[1] Rice University,[2] and MIT.[3] In their respective experiments, vapors of ^{87}Rb, ^{7}Li, and ^{23}Na atoms were confined by traps of inhomogeneous magnetic fields acting on the spin of the unpaired electron of each atom. A combination of laser and evaporative cooling techniques were employed to cool each gas below the phase transition point. Many similar experiments followed soon after these pioneering works. These experimental observations have, in turn, stimulated theoretical interest, with emphasis on the study of the effect on condensation of parameters that can be controlled externally, aiming at new predictions or designs of future experiments. Major problems related to Bose–Einstein condensation in a trap include equilibrium and nonequilibrium properties of the boson gas, such as collective excitations and vortices, and description of time evolution under the influence of time-dependent trapping potentials.

An entirely quantum mechanical treatment of Bose–Einstein condensation in dilute systems of hard spheres lacking translational symmetry at extremely low temperatures, when condensation into a single-particle state is nearly complete, was given in 1961 by Wu.[4] In his approach, the two crucial quantities for the minimal description of the Bose system are: (i) the condensate wave function $\Phi(\mathbf{r},t)$, which, to the lowest approximation in the particle density, satisfies a Schrödinger equation with a self-coupling term of third order, also derived by Gross[5] and Pitaevskii[6] by other methods, and (ii) the pair-excitation function $K_0(\mathbf{r},\mathbf{r}';t)$, which describes the scattering of two atoms from the condensate to other states at positions $\mathbf{r}$ and $\mathbf{r}'$, offers a systematic treatment of physical effects such as sound vibrations, and provides corrections to higher orders in the particle density; $K_0(\mathbf{r},\mathbf{r}';t)$ was shown to satisfy a nonlinear integro-differential equation. To the lowest approximation, this analysis has recently been extended both for zero and finite temperatures to incorporate the effect of a sufficiently smooth external potential that increases rapidly at large

[a)]Electronic mail: dmarget@fas.harvard.edu

distances.[7–9] Of particular significance is the underlying ansatz for the many-body Schrödinger state vector at zero temperature:[4,7,8]

$$\mathbf{\Psi}(t)=\mathcal{N}(t)e^{\mathcal{P}(t)}(N!)^{-1/2}a_0^*(t)^N|\mathrm{vac}\rangle, \tag{1.1}$$

where $\mathcal{P}(t)$ describes the creation of pairs from the condensate

$$\mathcal{P}(t)=(2N)^{-1}\int d\mathbf{r}\,d\mathbf{r}'\,\psi_1^*(\mathbf{r},t)\psi_1^*(\mathbf{r}',t)K_0(\mathbf{r},\mathbf{r}';t)a_0(t)^2, \tag{1.2}$$

$\mathcal{N}(t)$ is the normalization constant, which is immaterial for present purposes, N is the total number of atoms, $a_0^*(t)$ and $a_0(t)$ are the creation and annihilation operators for the condensate, respectively, and $\psi_1^*(\mathbf{r},t)$ is the boson creation field operator corresponding to the space orthogonal to the condensate wave function. Formula (1.1), being combined with a consistent approximation for the N-body Hamiltonian, is a nontrivial generalization of the many-body wave function of Lee, Huang, and Yang[10] for the case with translational invariance and periodic boundary conditions, where the main effect of particle interactions is the creation and annihilation of pairs of opposite momenta. The inclusion of pair excitation according to Eq. (1.1) necessarily modifies the equation of motion for $\Phi(\mathbf{r},t)$. Some physically interesting implications of this second-order approximation without any external potential, such as the difference between a compressional wave and a phonon, are discussed in Ref. 4.

Recent numerical or analytical studies of properties of nonuniform atomic gases undergoing Bose–Einstein condensation at extremely low temperatures have focused on the nonlinear Schrödinger equation for the condensate wave function either in its time-independent[8,11–14] or its time-dependent form.[15] A different approach by Benjamin, Quiroga, and Johnson[16] deals with the relative motion of the atoms in a hyperspherical coordinate system, with application to two-dimensional harmonic traps. In other contexts, several types of nonlinear Schrödinger equations are examined in the light of soliton theory,[17] often with emphasis on the description and conditions of existence of a pulselike solution—from now on referred to as a solitary wave—whose main feature is the preservation of its shape during propagation. A summary and discussion of some of these approaches can be found in the very recent comprehensive paper by Morgan *et al.*,[18] whose terminology is mainly adopted here.

Soliton theory usually describes nonlinear waves that interact like classical elastic particles, in the sense that the initial shape and velocity of the waves are regained asymptotically, yet possibly with a phase shift. Studies of such a behavior are believed to have been motivated from some unusual findings in a computation by Fermi, Pasta, and Ulam in 1955.[19,20] Significant advances toward the understanding of solutions to the underlying Korteweg–deVries (or KdV) equation were made ten years later by Zabusky and Kruskal,[21] followed by systematic investigations of Gardner *et al.*[22] A good list of references and exposition of methods or concepts germane to widely known types of evolution equations are given in Ref. 23. It has been realized that a central role in soliton theory is played by the ''Bäcklund transformations,'' which have provided a test for solitonic behavior and led to higher soliton solutions to some equations. (For a review of the mathematically advanced theory, see Ref. 19 and the references therein.)

It is well-known that the Schrödinger equation with a self-coupling term of third order and zero external potential admits soliton solutions in the sense of Ref. 24. In general, the inclusion of a term accounting for an external potential modifies the nature of the associated solutions, as is pointed out in Ref. 18. Specifically, Morgan *et al.*[18] examine conditions on nonlinear terms and accompanying external potentials that allow for localized solitary-wave solutions, and provide a physical interpretation of their results. They justifiably conclude that (i) such nonlinearities should not explicitly depend on the space variable x in $(1+1)$ dimensions, and (ii) the change in the potential experienced by the wave must be linear in x. They subsequently attempt to extend their results to higher dimensions, with restriction to motion along fixed axes in space. This in turn imposes conditions on the external potential, which they briefly describe. Notably, one-dimensional motion of shape-preserving pulses of the condensate wave function is also studied in

Refs. 25 and 26 for positive and negative scattering lengths, respectively, with restriction to time-independent parabolic potentials of weak confinement along one specified axis (cigar-shaped traps).

It should be emphasized, however, that, although it simplifies the treatment, the assumption of rectilinear motion in a space of dimensions higher than one is not necessary for the existence of solitary-wave solutions: motion of the solitary wave along an arbitrary trajectory in any number of space dimensions is possible, provided the external potential is consistently chosen. Furthermore, in Refs. 18, 25, and 26 the effects of scattering processes due to atomic interactions are ignored. Such a simplified approach, though adequate for some cases of experimental relevance, is certainly physically incomplete and needs improvement. It has been argued by others, for instance, that predictions based on the usual nonlinear Schrödinger equation become, in general, questionable for time-dependent systems, when the number of noncondensed particles may grow in time.[27] In the present paper, scattering processes are minimally taken into account through the *joint* consideration of the condensate wave function and the pair-excitation function.[4,7–9] The purpose of this work is to study solitary-wave motion by addressing the aforementioned issues in some detail, complementing, therefore, the analysis in Ref. 18, as a step toward an understanding of more complicated nonequilibrium properties of the trapped Bose gas. An outline of the paper is provided below.

In Sec. II, external potentials $V_e(\mathbf{r},t)$ in $(d+1)$ dimensions $(d\geqslant 1)$ are analyzed under the assumption that they sustain a condensate wave function identified with a single pulse that preserves its shape while moving along an arbitrarily prescribed trajectory in the d-dimensional Euclidean space. Focus is on the Schrödinger equation containing a cubic self-coupling term and positive scattering length a. The analysis starts with $d=1$, but with a perspective different from Ref. 18, and proceeds to generalizing to $d\geqslant 2$. Given a consistent V_e, the initial condition for the condensate wave function, when the nonlinearity plays an important role, is discussed. An argument is sketched to verify that, as a consequence of the requisite decomposition for the potential, the harmonic potentials constitute the sole class of admissible time-independent potentials that allow for solitary-wave solutions.[28] Furthermore, the assumption of nonuniqueness of the derived decomposition for the potential furnishes a class of time-dependent harmonic potentials. In Sec. III, it is demonstrated that the corresponding lowest-order nonlinear integro-differential equation for the pair-excitation function admits solitary waves in $(2d+1)$ dimensions. Section IV proceeds to determine approximately the initial amplitudes for the condensate wave function and the pair-excitation function corresponding to the lowest state of the condensate in a case of experimental interest, namely, when the trapping potential is slowly varying in space. In Sec. V, both analytical and numerical results are obtained for the lowest-energy condensate wave function under a three-dimensional, spherically symmetric harmonic potential.

II. THE CONDENSATE WAVE FUNCTION

The time-dependent nonlinear Schrödinger equation for the condensate wave function $\Phi(\mathbf{r},t)$ in an external potential $V_e(\mathbf{r},t)$ is $(\hbar=2m=1)$[7,8]

$$i(\partial/\partial t)\Phi(\mathbf{r},t)=[-\nabla^2+V_e(\mathbf{r},t)+8\pi aN\Omega^{-1}|\Phi(\mathbf{r},t)|^2-4\pi aN\Omega^{-1}\zeta(t)]\Phi(\mathbf{r},t), \tag{2.1}$$

where

$$\Omega^{-1}\int d\mathbf{r}|\Phi(\mathbf{r},t)|^2=1, \tag{2.2}$$

$$\zeta(t)=\Omega^{-1}\int d\mathbf{r}|\Phi(\mathbf{r},t)|^4, \tag{2.3}$$

a is the scattering length, assumed to be positive, N is the number of particles, and Ω is the volume of the system.

For mathematical convenience, Eq. (2.1) is cast in the form

$$i(\partial/\partial t)F(\mathbf{r},t)=[-\nabla^2+V_e(\mathbf{r},t)+|F(\mathbf{r},t)|^2]F(\mathbf{r},t), \tag{2.4}$$

where

$$F(\mathbf{r},t)=(8\pi a\rho_0)^{1/2}e^{-i4\pi\alpha\rho_0\sigma(t)}\Phi(\mathbf{r},t), \quad \rho_0=N/\Omega, \tag{2.5}$$

provided that

$$\sigma(t)=\int^t dt\,\zeta(t)+\text{const}, \tag{2.6}$$

where $\int^t$ denotes an indefinite integral. The normalization condition (2.2) now reads

$$\int d\mathbf{r}|F(\mathbf{r},t)|^2=8\pi aN. \tag{2.7}$$

A. The one-dimensional nonlinear Schrödinger equation

In the one-dimensional case, both the external potential and the condensate wave function depend on one space variable, say x. Equation (2.4) then becomes

$$i\,\frac{\partial F(x,t)}{\partial t}=\left[-\frac{\partial^2}{\partial x^2}+V_e(x,t)+|F(x,t)|^2\right]F(x,t). \tag{2.8}$$

For $V_e=0$, this reduces to the more or less standard form of the nonlinear Schrödinger equation.[24] Solitary-wave solutions of this equation are assumed to be of the form (see the Appendix):

$$F(x,t)=f(x-\alpha(t))e^{-i\theta(x,t)}, \tag{2.9}$$

where $f(x)$ and $\theta(x,t)$ are real functions, sufficiently smooth in x and t, and $\alpha(t)$ is a continuously differentiable function of time. Under the assumption of a potential $V_e(x,t)$ increasing sufficiently rapidly for $x\to\pm\infty$, it is necessary to require that

$$f(x)\to 0 \ \text{ rapidly } \quad \text{as } |x|\to\infty. \tag{2.10}$$

The example of the one-dimensional harmonic oscillator (briefly reviewed in the Appendix) suggests that f should decrease faster than exponentially in $|x|$ for large values of $|x|$. The same conclusion can be reached by employing the Wentzel–Kramers–Brillouin method.

The substitution of Eq. (2.9) into Eq. (2.8), and separation of real and imaginary parts, yield a system of coupled differential equations for f and θ:

$$f(x-\alpha(t))\,\frac{\partial^2\theta}{\partial x^2}+2f'(x-\alpha(t))\,\frac{\partial\theta}{\partial x}=-\alpha'(t)f'(x-\alpha(t)), \tag{2.11}$$

$$-f''(x-\alpha(t))+\left(\frac{\partial\theta}{\partial x}\right)^2 f(x-\alpha(t))+[V_e(x,t)+f(x-\alpha(t))^2]f(x-\alpha(t))=\frac{\partial\theta}{\partial t}f(x-\alpha(t)), \tag{2.12}$$

where the prime denotes differentiation with respect to argument. Equation (2.11) can be rewritten as

$$\frac{\partial}{\partial x}\left[f(x-\alpha(t))^2\frac{\partial\theta}{\partial x}\right]=-\alpha'(t)f'(x-\alpha(t))f(x-\alpha(t)). \tag{2.13}$$

This is explicitly integrated to give

$$\frac{\partial\theta}{\partial x}=-\frac{1}{2}\alpha'(t)+\frac{A_1(t)}{f(x-\alpha(t))^2}, \tag{2.14}$$

except at points $x=x(t)$ where $f(x-\alpha(t))$ vanishes. It immediately follows that

$$\theta(x,t)=\int^x dx\,\frac{\partial\theta}{\partial x}+A(t)=-\frac{1}{2}\alpha'(t)x+\int^x dx\,\frac{A_1(t)}{f(x-\alpha(t))^2}+A(t), \tag{2.15}$$

where x lies between consecutive zeros of $f(x-\alpha(t))$, calling for the possible use of different corresponding A_1's and A's. Consider the simplest case where f has no zeros. According to the preceding formula, for nonzero $A_1(t)$, the limiting behavior of f at large distances x and fixed time t gives rise to increasingly rapid oscillations in x of the real and imaginary parts of the condensate wave function. This in turn implies an infinite expectation value of the kinetic energy term $-\partial^2/\partial x^2$ in the Hamiltonian of the system. To eliminate this unphysical possibility, it is necessary to set $A_1(t)$ equal to zero:

$$A_1(t)\equiv 0. \tag{2.16}$$

To put this argument on a firm foundation, it is expedient to invoke the following conditions.

(i) Normalizability of $F(x,t)$ from Eq. (2.7), viz.

$$\int dx|F(x,t)|^2=\int dx\, f(x-\alpha(t))^2<\infty. \tag{2.17}$$

(ii) Finite kinetic energy of the condensate, viz.

$$\int dx\,F^*(x,t)\left(-\frac{\partial^2}{\partial x^2}\right)F(x,t)=\int dx\left|\frac{\partial F}{\partial x}\right|^2<\infty. \tag{2.18}$$

The last condition entails

$$\int dx\, f'(x-\alpha(t))^2<\infty, \tag{2.19a}$$

$$\int dx\left(\frac{\partial\theta}{\partial x}\right)^2 f(x-\alpha(t))^2<\infty. \tag{2.19b}$$

The use of Eqs. (2.14) and (2.17) in Eq. (2.19b) gives

$$\int dx\,A_1(t)\left[-\alpha'(t)+\frac{A_1(t)}{f(x-\alpha(t))^2}\right]<\infty, \tag{2.20}$$

which is impossible unless identity (2.16) holds. A similar argument can be applied to the case where f has any number of zeros.

For smooth real f, the resulting phase $\theta(x,t)$ is

$$\theta(x,t)=-\tfrac{1}{2}\alpha'(t)x+A(t), \tag{2.21}$$

in agreement with Eq. (9) of Ref. 18. The substitution of Eq. (2.21) into Eq. (2.12) yields a consistency equation for $V_e(x,t)$:

$$f''(x-\alpha(t))=[V_e(x,t)+f(x-\alpha(t))^2+\tfrac{1}{2}\alpha''(t)x+\tfrac{1}{4}\alpha'(t)^2-A'(t)]f(x-\alpha(t)). \tag{2.22}$$

It is inferred that $V_e(x,t)$ must be expressed as

$$V_e(x,t)=\mathcal{V}_1(x-\alpha(t))+x\mathcal{V}_2(t)+\mathcal{V}_3(t), \tag{2.23}$$

where

$$\mathcal{V}_1(x)=\frac{f''(x)}{f(x)}-f(x)^2, \tag{2.24}$$

$$\mathcal{V}_2(t)=-\tfrac{1}{2}\alpha''(t), \tag{2.25}$$

$$\mathcal{V}_3(t)=-\tfrac{1}{4}\alpha'(t)^2+A'(t). \tag{2.26}$$

Equation (2.23) gives the requisite form of potentials for given $f(x)$, $\alpha(t)$, and $A(t)$. Note that some of the inflection points of $f(x)$ need to coincide with its zeros. By close examination of Eqs. (2.23)–(2.26), the following should be pointed out.

(1) Given a $\mathcal{V}_1(x)$, the differential equation (2.24) suggests, in some sense, an eigenvalue problem. More particularly, when $|x|$ is sufficiently large, condition (2.10) becomes effective, indicating that $f^2 \ll |f''/f|$. Under this approximation, Eq. (2.24) becomes

$$f''(x)\sim\mathcal{V}_1(x)f(x), \tag{2.27a}$$

which is a linear equation. Hence, only discrete shifts ϵ_m of $\mathcal{V}_1(x)=\mathcal{V}_{1m}(x)$ are permissible, corresponding to ''eigenfunctions'' $f=f_m$ (m=non-negative integer). These shifts in turn induce discrete amounts of shift in $A'(t)$ through Eqs. (2.23) and (2.26). Accordingly, $F(x,t)$ exhibits a behavior of the form $e^{-i\epsilon_m t}f_m(x-\alpha(t))$ in the fixed trapping potential

$$\mathcal{V}_e(x)=\mathcal{V}_{1m}(x)+\sum_{l\leqslant m-1}\epsilon_l+C_0 \tag{2.27b}$$

experienced by the pulse, where C_0 is a constant.[29]

(2) For $\alpha(t)$ different from a constant, the only class of time-independent potentials $V_e(x,t)=V_e(x)$ of the form (2.23) consists of the harmonic potentials. Indeed, differentiation in x of both sides of Eq. (2.23) twice yields

$$V_e''(x)=\mathcal{V}_l''(x-\alpha(t))=K=\text{const}>0. \tag{2.28}$$

Hence,

$$V_e(x)=\tfrac{1}{2}Kx^2+\overline{K}x+\mathcal{C}. \tag{2.29}$$

(3) If $V_e(x,t)$ admits a second decomposition

$$V_e(x,t)=\mathcal{U}_1(x-\beta(t))+x\mathcal{U}_2(t)+\mathcal{U}_3(t), \tag{2.30}$$

where

$$\mathcal{U}_1(x)=\frac{\breve{f}''(x)}{\breve{f}(x)}-\breve{f}(x)^2, \tag{2.31}$$

$$\mathcal{U}_2(t)=-\tfrac{1}{2}\beta''(t), \tag{2.32}$$

$$\mathcal{U}_3(t)=-\tfrac{1}{4}\beta'(t)^2+B'(t), \tag{2.33}$$

and $\mathcal{U}_1(x)\neq\mathcal{V}_1(x)$, $\mathcal{U}_3(t)\neq\mathcal{V}_3(t)$, two cases for $\alpha(t)$ and $\beta(t)$ need to be distinguished.

(i) $\alpha(t)-\beta(t)\neq\text{const}$. Differentiation of Eqs. (2.23) and (2.30) with respect to x twice yields

$$\mathcal{V}_1''(x-\alpha(t))=\mathcal{U}_1''(x-\beta(t))=K. \tag{2.34}$$

Therefore,

$$V_e(x,t)=\tfrac{1}{2}K[x-\alpha(t)]^2+K_1[x-\alpha(t)]+K_2+x\mathcal{V}_2(t)+\mathcal{V}_3(t) \tag{2.35a}$$

$$=\tfrac{1}{2}K[x-\beta(t)]^2+M_1[x-\beta(t)]+M_2+x\mathcal{U}_2(t)+\mathcal{U}_3(t), \tag{2.35b}$$

i.e., $V_e(x,t)$ is the *time-dependent* harmonic potential

$$V_e(x,t)=\tfrac{1}{2}Kx^2+\overline{K}(t)x+\mathcal{C}(t) \quad (K>0). \tag{2.36}$$

A comparison of Eqs. (2.35a) and (2.35b) furnishes the consistency equations

$$-K\alpha(t)+K_1+\mathcal{V}_2(t)=-K\beta(t)+M_1+\mathcal{U}_2(t), \tag{2.37a}$$

$$\tfrac{1}{2}K\alpha(t)^2-K_1\alpha(t)+K_2+\mathcal{V}_3(t)=\tfrac{1}{2}K\beta(t)^2-M_1\beta(t)+M_2+\mathcal{U}_3(t). \tag{2.37b}$$

(ii) $\alpha(t)-\beta(t)=\mathcal{C}_1=\text{const}$. From Eqs. (2.25) and (2.32),

$$\mathcal{V}_2(t)=\mathcal{U}_2(t). \tag{2.38}$$

Equations (2.23) and (2.30) combined give

$$\mathcal{V}_3(t)-\mathcal{U}_3(t)=\mathcal{U}_1(x-\mathcal{C}_1)-\mathcal{V}_1(x)=\epsilon=\text{const}. \tag{2.39}$$

In view of (2.26) and (2.33),

$$A(t)=B(t)+\epsilon t+\text{const}. \tag{2.40}$$

The meaning of this ϵ becomes apparent from Eqs. (2.27): it is the discrete amount of shift in $\mathcal{V}_1(x)$ corresponding to a shift from the ''eigenfunction'' $f(x)$ to another ''eigenfunction'' $\breve{f}(x)$ under the same trapping potential $\mathcal{V}_e$ experienced by the solitary wave.

B. The nonlinear Schrödinger equation in d space dimensions, $d\geqslant 2$

The foregoing analysis in one dimension can be extended to higher dimensions. For definiteness, consider $d=3$. In accord with the conditions in the recent experiments,[1–3] it is assumed that

$$V_e(\mathbf{r},t)\to+\infty, \quad \text{uniformly in } \hat{\mathbf{r}}=\mathbf{r}/|\mathbf{r}| \quad \text{as } r=|\mathbf{r}|\to\infty. \tag{2.41}$$

Instead of assuming motion of the solitary wave along a fixed axis, as is the case in Ref. 18, let

$$F(\mathbf{r},t)=f(\mathbf{r}-\boldsymbol{\alpha}(t))e^{-i\theta(\mathbf{r},t)}, \tag{2.42}$$

where $\boldsymbol{\alpha}(t)$ is a twice differentiable vector function of time, $f(\mathbf{r})$ and $\theta(\mathbf{r},t)$ are real and sufficiently smooth, and from Eq. (2.7),

$$\int d\mathbf{r}\, f(r-\boldsymbol{\alpha}(t))^2=8\pi aN. \tag{2.43}$$

In view of condition (2.41), it is reasonable to assume that

$$f\to 0 \quad \text{rapidly}, \quad \text{uniformly in } \hat{\mathbf{r}}=\mathbf{r}/|\mathbf{r}| \quad \text{as } r\to\infty, \tag{2.44}$$

ensuring that the condensate is localized and has a finite kinetic energy, as indicated in the Appendix. The substitution of Eq. (2.42) in Eq. (2.4) gives

$$-i\boldsymbol{\alpha}'(t)\cdot\nabla f(\mathbf{r}-\boldsymbol{\alpha}(t))+f(\mathbf{r}-\boldsymbol{\alpha}(t))\frac{\partial\theta(\mathbf{r},t)}{\partial t}=-\nabla^2 f(\mathbf{r}-\boldsymbol{\alpha}(t))+2i\nabla f(\mathbf{r}-\boldsymbol{\alpha}(t))\cdot\nabla\theta(\mathbf{r},t)$$
$$+if(\mathbf{r}-\boldsymbol{\alpha}(t))\nabla^2\theta(\mathbf{r},t)+f(\mathbf{r}-\boldsymbol{\alpha}(t))|\nabla\theta(\mathbf{r},t)|^2$$
$$+[V_e(\mathbf{r},t)+f(\mathbf{r}-\boldsymbol{\alpha}(t))^2]f(\mathbf{r}-\boldsymbol{\alpha}(t)). \tag{2.45}$$

Upon separation of real and imaginary parts, the preceding equation decomposes into

$$-\boldsymbol{\alpha}'(t)\cdot\nabla f(\mathbf{r}-\boldsymbol{\alpha}(t))=2\nabla f(\mathbf{r}-\boldsymbol{\alpha}(t))\cdot\nabla\theta(\mathbf{r},t)+f(\mathbf{r}-\boldsymbol{\alpha}(t))\nabla^2\theta(\mathbf{r},t), \tag{2.46}$$

$$f(\mathbf{r}-\boldsymbol{\alpha}(t))\frac{\partial\theta(\mathbf{r},t)}{\partial t}=-\nabla^2 f(\mathbf{r}-\boldsymbol{\alpha}(t))+|\nabla\theta(\mathbf{r},t)|^2 f(\mathbf{r}-\boldsymbol{\alpha}(t))$$
$$+[V_e(\mathbf{r},t)+f(\mathbf{r}-\boldsymbol{\alpha}(t))^2]f(\mathbf{r}-\boldsymbol{\alpha}(t)). \tag{2.47}$$

Equation (2.46) is recast in the form

$$\nabla\cdot(f^2\nabla\theta)=-f\boldsymbol{\alpha}'(t)\cdot\nabla f, \quad f=f(\mathbf{r}-\boldsymbol{\alpha}(t)), \tag{2.48}$$

which holds regardless of the specific form for the shape $f=f(\mathbf{r},t)$ of $F(\mathbf{r},t)$. A particular solution to this equation is

$$\theta_p(\mathbf{r},t)=-\tfrac{1}{2}\boldsymbol{\alpha}'(t)\cdot\mathbf{r}+A(t). \tag{2.49}$$

With $\theta=\theta_p+\theta_1$, $\theta_1(\mathbf{r},t)$ satisfies the homogeneous equation

$$\nabla\cdot(f^2\nabla\theta_1)=0. \tag{2.50}$$

Integration by parts over a finite region $\mathcal{R}$ bounded by a surface $\mathcal{S}$ yields

$$0=\int d\mathbf{r}\,\theta_1\nabla\cdot(f^2\nabla\theta_1)=\oint_{\mathcal{S}}d\mathcal{S}\,f^2\theta_1\hat{\mathbf{n}}\cdot\nabla\theta_1-\int d\mathbf{r}\,f^2|\nabla\theta_1|^2, \tag{2.51}$$

where $\hat{\mathbf{n}}$ is the unit vector normal to $\mathcal{S}$ pointing outward. When $\mathcal{R}$ extends to infinity, the surface integral becomes arbitrarily small because of the condition (2.44), in analogy with the one-dimensional case. Consequently,

$$f^2|\nabla\theta_1|^2=0 \quad \text{almost everywhere}, \tag{2.52}$$

i.e., except for a set of points of measure zero. When $f\neq 0$, this in turn entails

$$\theta_1(\mathbf{r},t)=C_1(t) \quad \text{almost everywhere}. \tag{2.53}$$

At the zeros of f, $|\nabla\theta_1|$ seems to be indeterminate, calling for the use of different C_1's in Eq. (2.53). However, for a sufficiently smooth $\theta(\mathbf{r},t)$, $C_1(t)$ can be taken to be zero everywhere without loss of generality. Accordingly, $\theta(\mathbf{r},t)$ reads

$$\theta(\mathbf{r},t)=\theta_p(\mathbf{r},t)=-\tfrac{1}{2}\boldsymbol{\alpha}'(t)\cdot\mathbf{r}+A(t), \tag{2.54}$$

which is a generalization of Eq. (2.21).

The external potential consistent with Eqs. (2.47) and (2.54) is

$$V_e(\mathbf{r},t)=\mathcal{V}_1(\mathbf{r}-\boldsymbol{\alpha}(t))+\boldsymbol{\mathcal{V}}_2(t)\cdot\mathbf{r}+\mathcal{V}_3(t), \tag{2.55}$$

where

$$\mathcal{V}_1(\mathbf{r})=\frac{\nabla^2 f(\mathbf{r})}{f(\mathbf{r})}-f(\mathbf{r})^2, \tag{2.56}$$

$$\boldsymbol{\mathcal{V}}_2(t)=-\tfrac{1}{2}\boldsymbol{\alpha}''(t), \tag{2.57}$$

$$\mathcal{V}_3(t)=-\tfrac{1}{4}|\boldsymbol{\alpha}'(t)|^2+A'(t). \tag{2.58}$$

Notably, $\nabla^2 f(\mathbf{r})$ needs to vanish at any surface where $f(\mathbf{r})$ vanishes.

A few important remarks are in order.

(1) For an external potential increasing in $|\mathbf{r}|$, Eq. (2.56) bears the features of an eigenvalue problem. Specifically, for $|\mathbf{r}|\to\infty$, a linear equation is recovered approximately:

$$\nabla^2 f(\mathbf{r})\sim\mathcal{V}_1(\mathbf{r})f(\mathbf{r}). \tag{2.59}$$

Analogies with the one-dimensional case are easily drawn from this equation.

(2) When $\boldsymbol{\alpha}(t)$ is not a constant, the only time-independent potential of the form (2.55) that satisfies condition (2.41) is the d-dimensional harmonic potential. The justification for this is somewhat more demanding than for the one-dimensional case. With $V_e(\mathbf{r},t)=V_e(\mathbf{r})$, the application of the Laplacian to both sides of Eq. (2.55) gives

$$\nabla^2 V_e(\mathbf{r})=\nabla^2\mathcal{V}_1(\mathbf{r}-\boldsymbol{\alpha}(t))=K=\text{const}>0. \tag{2.60}$$

In three dimensions, a solution to Eq. (2.60) for $V_e(\mathbf{r})$ is:

$$V_p(\mathbf{r})=\frac{1}{2}\sum_{i,j=1,2,3}K_{ij}x_i x_j+\sum_{j=1,2,3}\overline{K}_j x_j+\mathcal{C}, \tag{2.61}$$

where $(x_1,x_2,x_3)=\mathbf{r}=(x,y,z)$,

$$\text{Tr}[K_{ij}]=K, \tag{2.62}$$

and the matrix $[K_{ij}]$ is symmetric and positive definite. Every admissible solution to Eq. (2.60) can be written as

$$V_e(\mathbf{r})=V_p(\mathbf{r})+V_1(\mathbf{r}), \tag{2.63}$$

where $V_1(\mathbf{r})$ is a smooth function satisfying Laplace's equation:

$$\nabla^2 V_1(\mathbf{r})=0 \quad \text{everywhere}. \tag{2.64}$$

If $\mathcal{S}$ is now a spherical surface with center $\mathbf{r}$ and radius R, then according to Gauss' mean value theorem[30]

$$V_1(\mathbf{r})=\frac{1}{4\pi R^2}\oint_{\mathcal{S}}dS'\,V_1(\mathbf{r}'). \tag{2.65}$$

Since R can be taken to be arbitrarily large, it follows that V_1 cannot be forced to comply with condition (2.41). Consequently, $V_1(\mathbf{r})$ is equal to a constant. Without loss of generality,

$$V_1(\mathbf{r})\equiv 0. \tag{2.66}$$

(3) Let $V_e(\mathbf{r},t)$ admit an alternative decomposition,

$$V_e(\mathbf{r},t)=\mathcal{U}_1(\mathbf{r}-\boldsymbol{\beta}(t))+\mathbf{r}\cdot\boldsymbol{\mathcal{U}}_2(t)+\mathcal{U}_3(t), \tag{2.67}$$

where

$$\mathcal{U}_1(\mathbf{r}) = \frac{\nabla^2 \breve{f}(\mathbf{r})}{\breve{f}(\mathbf{r})} - \breve{f}(\mathbf{r})^2, \tag{2.68}$$

$$\boldsymbol{\mathcal{U}}_2(t) = -\tfrac{1}{2}\boldsymbol{\beta}''(t), \tag{2.69}$$

$$\mathcal{U}_3(t) = -\tfrac{1}{4}|\boldsymbol{\beta}'(t)|^2 + B'(t), \tag{2.70}$$

and $\mathcal{U}_1(\mathbf{r}) \neq \mathcal{V}_1(\mathbf{r})$, $\mathcal{U}_3(t) \neq \mathcal{V}_3(t)$. In analogy with the one-dimensional case, there are two distinct possibilities.

(i) $\boldsymbol{\alpha}(t) - \boldsymbol{\beta}(t) \neq \text{const}$. Then,

$$\nabla^2 \mathcal{V}_1(\mathbf{r} - \boldsymbol{\alpha}(t)) = \nabla^2 \mathcal{U}_1(\mathbf{r} - \boldsymbol{\beta}(t)) = K, \tag{2.71}$$

which in turn implies that

$$\begin{aligned} V_e(\mathbf{r},t) &= \frac{1}{2} \sum_{i,j=1,2,3} K_{ij}[x_i - \alpha_i(t)][x_j - \alpha_j(t)] + [\mathbf{r} - \boldsymbol{\alpha}(t)] \cdot \mathbf{K}_1 + K_2 + \mathbf{r} \cdot \boldsymbol{\mathcal{V}}_2(t) + \mathcal{V}_3(t) \\ &= \frac{1}{2} \sum_{i,j=1,2,3} M_{ij}[x_i - \beta_i(t)][x_j - \beta_j(t)] + [\mathbf{r} - \boldsymbol{\beta}(t)] \cdot \mathbf{M}_1 + M_2 + \mathbf{r} \cdot \boldsymbol{\mathcal{U}}_2(t) + \mathcal{U}_3(t), \end{aligned} \tag{2.72}$$

where

$$\mathrm{Tr}[K_{ij}] = \mathrm{Tr}[M_{ij}] = K, \tag{2.73}$$

and K_2, M_2 are immaterial constants. Therefore, $V_e(\mathbf{r},t)$ is the time-dependent harmonic potential

$$V_e(\mathbf{r},t) = \frac{1}{2} \sum_{i,j=1,2,3} K_{ij} x_i x_j + \mathbf{r} \cdot \overline{\mathbf{K}}(t) + \mathcal{C}(t). \tag{2.74}$$

(ii) $\boldsymbol{\alpha}(t) - \boldsymbol{\beta}(t) = \mathbf{C}_1 = \text{const}$. Without loss of generality, $\mathbf{C}_1 = 0$. It is easily found that

$$\boldsymbol{\mathcal{V}}_2(t) = \boldsymbol{\mathcal{U}}_2(t), \tag{2.75}$$

$$\mathcal{V}_3(t) - \mathcal{U}_3(t) = \mathcal{U}_1(\mathbf{r}) - \mathcal{V}_1(\mathbf{r}) = \epsilon = \text{const}. \tag{2.76}$$

Equation (2.76) implies that

$$A(t) = B(t) + \epsilon t + \text{const}. \tag{2.77}$$

Therefore, $\breve{f}(\mathbf{r})$ is just another ''eigenfunction'' of Eq. (2.56) under the same trapping potential $\mathcal{V}_e$ seen by the pulse.

III. THE PAIR-EXCITATION FUNCTION

The pair-excitation function $K_0(\mathbf{r},\mathbf{r}';t)$ satisfies the integro-differential equation[8]

$$\left[i\frac{\partial}{\partial t}-2E(t)\right]K_0(\mathbf{r},\mathbf{r}';t)=-\nabla^2K_0(\mathbf{r},\mathbf{r}';t)-\nabla'^2K_0(\mathbf{r},\mathbf{r}';t)+8\pi a\rho_0\Phi(\mathbf{r},t)^2\delta(\mathbf{r}-\mathbf{r}')$$
$$+\{-2\overline{\zeta}(t)-16\pi a\rho_0\zeta(t)-2\zeta_e(t)+V_e(\mathbf{r},t)+V_e(\mathbf{r}',t)$$
$$+16\pi a\rho_0[|\Phi(\mathbf{r},t)|^2+|\Phi(\mathbf{r}',t)|^2]\}K_0(\mathbf{r},\mathbf{r}';t)$$
$$+8\pi a\rho_0\int d\mathbf{r}''\,\Phi^*(\mathbf{r}'',t)^2K_0(\mathbf{r},\mathbf{r}'';t)K_0(\mathbf{r}',\mathbf{r}'';t)$$
$$-8\pi a\rho_0\Omega^{-1}\Big\{\Phi(\mathbf{r},t)\Phi(\mathbf{r}',t)[|\Phi(\mathbf{r},t)|^2+|\Phi(\mathbf{r}',t)|^2-\zeta(t)]$$
$$+\Phi(\mathbf{r},t)\int d\mathbf{r}''\,K_0(\mathbf{r}',\mathbf{r}'';t)|\Phi(\mathbf{r}'',t)|^2\Phi^*(\mathbf{r}'',t)$$
$$+\Phi(\mathbf{r}',t)\int d\mathbf{r}''\,K_0(\mathbf{r},\mathbf{r}'';t)|\Phi(\mathbf{r}'',t)|^2\Phi^*(\mathbf{r}'',t)\Big\}, \tag{3.1}$$

where

$$E(t)=i\Omega^{-1}\int d\mathbf{r}\,\frac{\partial\Phi(\mathbf{r},t)}{\partial t}\Phi^*(\mathbf{r},t), \tag{3.2}$$

$$\overline{\zeta}(t)=\Omega^{-1}\int d\mathbf{r}|\nabla\Phi(\mathbf{r},t)|^2,\quad \zeta_e(t)=\Omega^{-1}\int d\mathbf{r}\,V_e(\mathbf{r},t)|\Phi(\mathbf{r},t)|^2, \tag{3.3}$$

and $\nabla\equiv\nabla_{\mathbf{r}}$, $\nabla'\equiv\nabla_{\mathbf{r}'}$. Without loss of generality, $K_0(\mathbf{r},\mathbf{r}';t)$ has been chosen to satisfy

$$K_0(\mathbf{r},\mathbf{r}';t)=K_0(\mathbf{r}',\mathbf{r};t), \tag{3.4}$$

$$\int d\mathbf{r}\,\Phi^*(\mathbf{r},t)K_0(\mathbf{r},\mathbf{r}';t)=0. \tag{3.5}$$

In order to investigate the possibility for solitary-wave solutions to Eq. (3.1), the following preliminary steps are taken:

(i) By virtue of Eq. (2.5), $\Phi(\mathbf{r},t)$ is replaced by $(8\pi a\rho_0)^{-1/2}e^{i4\pi a\rho_0\sigma(t)}F(\mathbf{r},t)$.
(ii) To balance out the exponential factor introduced above, $K_0(\mathbf{r},\mathbf{r}';t)$ is written as

$$K_0(\mathbf{r},\mathbf{r}';t)=e^{i8\pi a\rho_0\sigma(t)}\mathcal{K}_0(\mathbf{r},\mathbf{r}';t). \tag{3.6}$$

The resulting equation for this $\mathcal{K}_0(\mathbf{r},\mathbf{r}';t)$ is

$$i\frac{\partial\mathcal{K}_0(\mathbf{r},\mathbf{r}';t)}{\partial t}=-\nabla^2\mathcal{K}_0(\mathbf{r},\mathbf{r}';t)-\nabla'^2\mathcal{K}_0(\mathbf{r},\mathbf{r}';t)+F(\mathbf{r},t)^2\delta(\mathbf{r}-\mathbf{r}')+\{V_e(\mathbf{r},t)+V_e(\mathbf{r}',t)$$
$$+2[|F(\mathbf{r},t)|^2+|F(\mathbf{r}',t)|^2]\}\mathcal{K}_0(\mathbf{r},\mathbf{r}';t)+\int d\mathbf{r}''\,F^*(\mathbf{r}'',t)^2\mathcal{K}_0(\mathbf{r},\mathbf{r}'';t)\mathcal{K}_0(\mathbf{r}',\mathbf{r}'';t)$$
$$-(8\pi aN)^{-1}\Big\{F(\mathbf{r},t)F(\mathbf{r}',t)[|F(\mathbf{r},t)|^2+|F(\mathbf{r}',t)|^2-\hat{\zeta}(t)]$$
$$+F(\mathbf{r},t)\int d\mathbf{r}''\,\mathcal{K}_0(\mathbf{r}',\mathbf{r}'';t)|F(\mathbf{r}'',t)|^2F^*(\mathbf{r}'',t)$$
$$+F(\mathbf{r}',t)\int d\mathbf{r}''\,\mathcal{K}_0(\mathbf{r},\mathbf{r}'';t)|F(\mathbf{r}'',t)|^2F^*(\mathbf{r}'',t)\Big\}, \tag{3.7}$$

where

$$\hat{\zeta}(t)=8\pi a\rho_0\zeta(t), \tag{3.8}$$

and $E(t)$ was replaced by

$$E(t)=\overline{\zeta(t)}+\zeta_e(t)+4\pi a\rho_0\zeta(t), \tag{3.9}$$

by employing Eq. (2.1).

Given Eqs. (2.42) and (3.4), solitary-wave solutions are sought in the form

$$\mathcal{K}_0(\mathbf{r},\mathbf{r}';t)=\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))e^{-i\chi(\mathbf{r},\mathbf{r}';t)}, \tag{3.10}$$

where $\kappa_0(\mathbf{r},\mathbf{r}')$, $\chi(\mathbf{r},\mathbf{r}';t)$, and $\boldsymbol{\gamma}(t)$ are sufficiently smooth real functions satisfying

$$\kappa_0(\mathbf{r},\mathbf{r}')=\kappa_0(\mathbf{r}',\mathbf{r}),\qquad \chi(\mathbf{r},\mathbf{r}';t)=\chi(\mathbf{r}',\mathbf{r};t), \tag{3.11}$$

$$\int d\mathbf{r}\, f(\mathbf{r}-\boldsymbol{\alpha}(t))\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))e^{i[\theta(\mathbf{r},t)-\chi(\mathbf{r},\mathbf{r}';t)]}=0. \tag{3.12}$$

The substitution of Eq. (3.10) into Eq. (3.7) by virtue of Eqs. (2.42) and (2.55) yields

$$\begin{aligned}
&-i\boldsymbol{\gamma}'(t)\cdot(\nabla\kappa_0+\nabla'\kappa_0)+\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))\frac{\partial\chi(\mathbf{r},\mathbf{r}';t)}{\partial t}\\
&\quad=-\nabla^2\kappa_0-\nabla'^2\kappa_0+2i(\nabla\kappa_0\cdot\nabla\chi+\nabla'\kappa_0\cdot\nabla'\chi)+i\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))(\nabla^2\chi+\nabla'^2\chi)\\
&\qquad+\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))(|\nabla\chi|^2+|\nabla'\chi|^2)+f(\mathbf{r}-\boldsymbol{\alpha}(t))^2\delta(\mathbf{r}-\mathbf{r}')e^{i\chi(\mathbf{r},\mathbf{r}';t)-i2\theta(\mathbf{r},t)}\\
&\qquad+\{\mathcal{V}_1(\mathbf{r}-\boldsymbol{\alpha}(t))+\mathcal{V}_1(\mathbf{r}'-\boldsymbol{\alpha}(t))+(\mathbf{r}+\mathbf{r}')\cdot\boldsymbol{\mathcal{V}}_2(t)+2\mathcal{V}_3(t)\\
&\qquad+2[f(\mathbf{r}-\boldsymbol{\alpha}(t))^2+f(\mathbf{r}'-\boldsymbol{\alpha}(t))^2]\}\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))\\
&\qquad+\int d\mathbf{r}''\, f(\mathbf{r}''-\boldsymbol{\alpha}(t))^2\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}''-\boldsymbol{\gamma}(t))\kappa_0(\mathbf{r}'-\boldsymbol{\gamma}(t),\mathbf{r}''-\boldsymbol{\gamma}(t))\\
&\qquad\times\exp\{2i\theta(\mathbf{r}'',t)-i[\chi(\mathbf{r},\mathbf{r}'';t)+\chi(\mathbf{r}',\mathbf{r}'';t)-\chi(\mathbf{r},\mathbf{r}';t)]\}\\
&\qquad-(8\pi aN)^{-1}\Big\{f(\mathbf{r}-\boldsymbol{\alpha}(t))f(\mathbf{r}'-\boldsymbol{\alpha}(t))[f(\mathbf{r}-\boldsymbol{\alpha}(t))^2+f(\mathbf{r}'-\boldsymbol{\alpha}(t))^2-\hat{\zeta}]\\
&\qquad\times\exp\{i\chi(\mathbf{r},\mathbf{r}';t)-i[\theta(\mathbf{r},t)+\theta(\mathbf{r}',t)]\}\\
&\qquad+f(\mathbf{r}-\boldsymbol{\alpha}(t))\int d\mathbf{r}''\,\kappa_0(\mathbf{r}'-\boldsymbol{\gamma}(t),\mathbf{r}''-\boldsymbol{\gamma}(t))f(\mathbf{r}''-\boldsymbol{\alpha}(t))^3\\
&\qquad\times\exp\{i[\theta(\mathbf{r}'',t)-\theta(\mathbf{r},t)]+i[\chi(\mathbf{r},\mathbf{r}';t)-\chi(\mathbf{r}',\mathbf{r}'';t)]\}\\
&\qquad+f(\mathbf{r}'-\boldsymbol{\alpha}(t))\int d\mathbf{r}''\,\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t)),\mathbf{r}''-\boldsymbol{\gamma}(t))f(\mathbf{r}''-\boldsymbol{\alpha}(t))^3\\
&\qquad\times\exp\{i[\theta(\mathbf{r}'',t)-\theta(\mathbf{r}',t)]+i[\chi(\mathbf{r},\mathbf{r}';t)-\chi(\mathbf{r},\mathbf{r}'';t)]\}\Big\},
\end{aligned} \tag{3.13}$$

where it is understood that $\kappa_0=\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))$ and $\chi=\chi(\mathbf{r},\mathbf{r}';t)$, and $\hat{\zeta}$ is now time independent. Elimination of the above phase factors succeeds if χ is taken equal to

$$\chi(\mathbf{r},\mathbf{r}';t)=\theta(\mathbf{r},t)+\theta(\mathbf{r}',t)=-\tfrac{1}{2}\boldsymbol{\alpha}'(t)\cdot(\mathbf{r}+\mathbf{r}')+2A(t). \tag{3.14}$$

In view of Eq. (3.14), separation of the real and imaginary parts in Eq. (3.13) leads to

$$-\boldsymbol{\gamma}'(t)\cdot(\nabla\kappa_0+\nabla'\kappa_0)=2(\nabla\kappa_0\cdot\nabla\chi+\nabla'\kappa_0\cdot\nabla'\chi)+\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))(\nabla^2\chi+\nabla'^2\chi), \tag{3.15}$$

$$\begin{aligned}
&-\tfrac{1}{2}\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))\boldsymbol{\alpha}''(t)\cdot(\mathbf{r}+\mathbf{r}')+2\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))A'(t)\\
&\quad=-\nabla^2\kappa_0-\nabla'^2\kappa_0+\tfrac{1}{2}\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))|\boldsymbol{\alpha}'(t)|^2+f(\mathbf{r}-\boldsymbol{\alpha}(t))^2\delta(\mathbf{r}-\mathbf{r}')\\
&\qquad+\{\mathcal{V}_1(\mathbf{r}-\boldsymbol{\alpha}(t))+\mathcal{V}_1(\mathbf{r}'-\boldsymbol{\alpha}(t))+(\mathbf{r}+\mathbf{r}')\cdot\boldsymbol{\mathcal{V}}_2(t)+2\mathcal{V}_3(t)\\
&\qquad+2[f(\mathbf{r}-\boldsymbol{\alpha}(t))^2+f(\mathbf{r}'-\boldsymbol{\alpha}(t))^2]\}\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}'-\boldsymbol{\gamma}(t))\\
&\qquad+\int d\mathbf{r}''\,f(\mathbf{r}''-\boldsymbol{\alpha}(t))^2\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t)),\mathbf{r}''-\boldsymbol{\gamma}(t))\kappa_0(\mathbf{r}'-\boldsymbol{\gamma}(t),\mathbf{r}''-\boldsymbol{\gamma}(t))\\
&\qquad-(8\pi aN)^{-1}\Big\{f(\mathbf{r}-\boldsymbol{\alpha}(t))f(\mathbf{r}'-\boldsymbol{\alpha}(t))[f(\mathbf{r}-\boldsymbol{\alpha}(t))^2+f(\mathbf{r}'-\boldsymbol{\alpha}(t))^2-\hat{\zeta}]\\
&\qquad+f(\mathbf{r}-\boldsymbol{\alpha}(t))\int d\mathbf{r}''\,\kappa_0(\mathbf{r}'-\boldsymbol{\gamma}(t),\mathbf{r}''-\boldsymbol{\gamma}(t))f(\mathbf{r}''-\boldsymbol{\alpha}(t))^3\\
&\qquad+f(\mathbf{r}'-\boldsymbol{\alpha}(t))\int d\mathbf{r}''\,\kappa_0(\mathbf{r}-\boldsymbol{\gamma}(t),\mathbf{r}''-\boldsymbol{\gamma}(t))f(\mathbf{r}''-\boldsymbol{\alpha}(t))^3\Big\},
\end{aligned} \tag{3.16}$$

of which the first one is satisfied if

$$\boldsymbol{\gamma}(t)=\boldsymbol{\alpha}(t)+\boldsymbol{\alpha}_0, \tag{3.17}$$

where $\boldsymbol{\alpha}_0$ is a vector constant. Without loss of generality, this $\boldsymbol{\alpha}_0$ is set equal to zero.

In Eq. (3.16), $\boldsymbol{\mathcal{V}}_2(t)$ and $\mathcal{V}_3(t)$ are replaced by $-\frac{1}{2}\boldsymbol{\alpha}''(t)$ and $-\frac{1}{4}|\boldsymbol{\alpha}'(t)|^2+A'(t)$ from Eqs. (2.57) and (2.58), respectively. With a subsequent shift both of $\mathbf{r}$ and $\mathbf{r}'$ by $\boldsymbol{\alpha}(t)$, all time dependencies are eliminated and an equation for $\kappa_0(\mathbf{r},\mathbf{r}')$ is obtained:

$$\begin{aligned}
&-\nabla^2\kappa_0(\mathbf{r},\mathbf{r}')-\nabla'^2\kappa_0(\mathbf{r},\mathbf{r}')+f(\mathbf{r})^2\delta(\mathbf{r}-\mathbf{r}')+\{\mathcal{V}_1(\mathbf{r})+\mathcal{V}_1(\mathbf{r}')+2[f(\mathbf{r})^2+f(\mathbf{r}')^2]\}\kappa_0(\mathbf{r},\mathbf{r}')\\
&\quad+\int d\mathbf{r}''\,f(\mathbf{r}'')^2\kappa_0(\mathbf{r},\mathbf{r}'')\kappa_0(\mathbf{r}',\mathbf{r}'')-(8\pi aN)^{-1}\Big\{f(\mathbf{r})f(\mathbf{r}')[f(\mathbf{r})^2+f(\mathbf{r}')^2-\hat{\zeta}]\\
&\quad+f(\mathbf{r})\int d\mathbf{r}''\,\kappa_0(\mathbf{r}',\mathbf{r}'')f(\mathbf{r}'')^3+f(\mathbf{r}')\int d\mathbf{r}''\,\kappa_0(\mathbf{r},\mathbf{r}'')f(\mathbf{r}'')^3\Big\}=0,
\end{aligned} \tag{3.18}$$

where

$$\int d\mathbf{r}\,f(\mathbf{r})\kappa_0(\mathbf{r},\mathbf{r}')=0. \tag{3.19}$$

When the number of particles, N, is sufficiently large, Eq. (3.18) is approximated by

$$\begin{aligned}
&-\nabla^2\kappa_0(\mathbf{r},\mathbf{r}')-\nabla'^2\kappa_0(\mathbf{r},\mathbf{r}')+f(\mathbf{r})^2\delta(\mathbf{r}-\mathbf{r}')+\{\mathcal{V}_1(\mathbf{r})+\mathcal{V}_1(\mathbf{r}')+2[f(\mathbf{r})^2+f(\mathbf{r}')^2]\}\kappa_0(\mathbf{r},\mathbf{r}')\\
&\quad+\int d\mathbf{r}''\,f(\mathbf{r}'')^2\kappa_0(\mathbf{r},\mathbf{r}'')\kappa_0(\mathbf{r}',\mathbf{r}'')=0.
\end{aligned} \tag{3.20}$$

IV. SLOWLY VARYING TRAPPING POTENTIAL

In order to elucidate the dependence on the physical parameters of the problem, let

$$\tilde{\Phi}(\mathbf{r},t)=\sqrt{\rho_0}\Phi(\mathbf{r},t), \quad \rho_0=N/\Omega. \tag{4.1}$$

$\tilde{\Phi}(\mathbf{r},t)$ satisfies

$$i(\partial/\partial t)\tilde{\Phi}(\mathbf{r},t)=[-\nabla^2+V_e(\mathbf{r},t)+8\pi a|\tilde{\Phi}(\mathbf{r},t)|^2-4\pi a\tilde{\zeta}(t)]\tilde{\Phi}(\mathbf{r},t), \tag{4.2}$$

and the normalization condition

$$N^{-1}\int d\mathbf{r}|\tilde{\Phi}(\mathbf{r},t)|^2=1. \tag{4.3}$$

In the above,

$$\tilde{\zeta}(t)=N^{-1}\int d\mathbf{r}|\tilde{\Phi}(\mathbf{r},t)|^4. \tag{4.4}$$

Equation (2.42) reads

$$\tilde{\Phi}(\mathbf{r},t)=\tilde{f}(\mathbf{r}-\boldsymbol{\alpha}(t))\exp\{i\tfrac{1}{2}\boldsymbol{\alpha}'(t)\cdot\mathbf{r}-iA(t)\}, \tag{4.5}$$

where

$$N^{-1}\int d\mathbf{r}\,\tilde{f}(\mathbf{r})^2=1. \tag{4.6}$$

The external potential is

$$V_e(\mathbf{r},t)=\tilde{\mathcal{V}}_1(\mathbf{r}-\boldsymbol{\alpha}(t))+\mathbf{r}\cdot\tilde{\boldsymbol{\mathcal{V}}}_2(t)+\tilde{\mathcal{V}}_3(t), \tag{4.7}$$

where

$$\tilde{\mathcal{V}}_1(\mathbf{r})=\frac{\nabla^2\tilde{f}(\mathbf{r})}{\tilde{f}(\mathbf{r})}-8\pi a\tilde{f}(\mathbf{r})^2+4\pi a\tilde{\zeta}, \quad \tilde{\zeta}=N^{-1}\int d\mathbf{r}\,\tilde{f}(\mathbf{r})^4, \tag{4.8a}$$

and $\tilde{\boldsymbol{\mathcal{V}}}_2(t)$, $\tilde{\mathcal{V}}_3(t)$ are given by equations similar to Eqs. (2.57) and (2.58). Therefore, $\tilde{f}(\mathbf{r})=\tilde{f}_m(\mathbf{r})(m=0,1,...,)$ correspond to states of the condensate with energies $\mathcal{E}_m$ under the external potential

$$\mathcal{V}_e=\tilde{\mathcal{V}}_{1m}+\mathcal{E}_m \quad (\tilde{\mathcal{V}}_1=\tilde{\mathcal{V}}_{1m}). \tag{4.8b}$$

Given a $\mathcal{V}_e(\mathbf{r})$, Eq. (4.2) can be solved approximately for the lowest state of the condensate when $\mathcal{V}_e(\mathbf{r})$ is sufficiently slowly varying. This is the case in the recent experiments on Bose–Einstein condensation, where the trap is of macroscopic dimensions. By applying the procedure of Refs. 8 and 13, neglect of the Laplacian furnishes

$$[\mathcal{V}_e(\mathbf{r})+8\pi a\tilde{f}(\mathbf{r})^2-4\pi a\tilde{\zeta}-\mathcal{E}]\tilde{f}(\mathbf{r})=0, \tag{4.9}$$

where $\mathcal{E}=\mathcal{E}_0$, or,

$$\tilde{f}(\mathbf{r})\sim\begin{cases}(8\pi a)^{-1/2}[\mathcal{E}+4\pi a\tilde{\zeta}-\mathcal{V}_e(\mathbf{r})]^{1/2}, & \mathbf{r}\text{ inside }\mathcal{R}_0\\ 0, & \mathbf{r}\text{ outside }\mathcal{R}_0\end{cases}, \tag{4.10}$$

since $\tilde{f}(\mathbf{r})$ can be chosen to be non-negative. The region $\mathcal{R}_0$ is determined by

$$\mathcal{V}_e(\mathbf{r})<\mathcal{E}+4\pi a\tilde{\zeta}, \quad \mathbf{r}\in\mathcal{R}_0. \tag{4.11}$$

At the boundary $\partial\mathcal{R}_0$ of $\mathcal{R}_0$,

$$\mathcal{E}+4\pi a\tilde{\zeta}=\mathcal{V}_e(\mathbf{r}), \quad \mathbf{r}\in\partial\mathcal{R}_0. \tag{4.12}$$

Under this approximation, an expression for $\mathcal{E}$ is obtained via multiplication of Eq. (4.9) by $\tilde{f}(\mathbf{r})$ and integration over $\mathbf{r}$:

$$\mathcal{E}\sim 4\pi a\tilde{\zeta}+\tilde{\zeta}_e, \tag{4.13}$$

where

$$\tilde{\zeta}_e=N^{-1}\int d\mathbf{r}\,\mathcal{V}_e(\mathbf{r})|\tilde{f}(\mathbf{r})|^2. \tag{4.14}$$

Formula (4.10) breaks down in the vicinity of $\partial\mathcal{R}_0$. A remedy to this problem is provided in Refs. 8 and 13.

It remains to discuss the pair-excitation function $K_0(\mathbf{r},\mathbf{r}';t)$. With

$$K_0(\mathbf{r},\mathbf{r}';t)=\tilde{\kappa}_0(\mathbf{r}-\boldsymbol{\alpha}(t),\mathbf{r}'-\boldsymbol{\alpha}(t))e^{-i\chi(\mathbf{r},\mathbf{r}';t)}, \tag{4.15}$$

and use of Eq. (3.14), $\tilde{\kappa}_0(\mathbf{r},\mathbf{r}')$ should satisfy

$$\begin{aligned}&-\nabla^2\tilde{\kappa}_0(\mathbf{r},\mathbf{r}')-\nabla'^2\tilde{\kappa}_0(\mathbf{r},\mathbf{r}')+8\pi a\tilde{f}(\mathbf{r})^2\delta(\mathbf{r}-\mathbf{r}')+\{-2\check{\zeta}-16\pi a\tilde{\zeta}-2\tilde{\zeta}_e+\mathcal{V}_e(\mathbf{r})+\mathcal{V}_e(\mathbf{r}')\\&\quad+16\pi a[\tilde{f}(\mathbf{r})^2+\tilde{f}(\mathbf{r}')^2]\}\tilde{\kappa}_0(\mathbf{r},\mathbf{r}')+8\pi a\int d\mathbf{r}''\,\tilde{f}(\mathbf{r}'')^2\tilde{\kappa}_0(\mathbf{r},\mathbf{r}'')\tilde{\kappa}_0(\mathbf{r}',\mathbf{r}'')=0,\end{aligned} \tag{4.16}$$

where

$$\check{\zeta}=N^{-1}\int d\mathbf{r}|\nabla\tilde{f}(\mathbf{r})|^2. \tag{4.17}$$

Note that shifting $\mathcal{V}_e$ by a constant does not affect the equation of motion.

Following Ref. 8, let

$$p_0(\mathbf{R},\mathbf{r})=\tilde{\kappa}_0(\mathbf{r}_1,\mathbf{r}_2), \tag{4.18}$$

where

$$\mathbf{R}=\tfrac{1}{2}(\mathbf{r}_1+\mathbf{r}_2), \quad \mathbf{r}=\mathbf{r}_1-\mathbf{r}_2. \tag{4.19}$$

Hence,

$$p_0(\mathbf{R},-\mathbf{r})=p_0(\mathbf{R},\mathbf{r}). \tag{4.20}$$

The integro-differential equation for $p_0(\mathbf{R},\mathbf{r})$ reads

$$\begin{aligned}&-\tfrac{1}{2}\nabla_{\mathbf{R}}^2p_0(\mathbf{R},\mathbf{r})-2\nabla_{\mathbf{r}}^2p_0(\mathbf{R},\mathbf{r})+8\pi a\tilde{f}(\mathbf{R})^2\delta(\mathbf{r})+\{-2\check{\zeta}-16\pi a\tilde{\zeta}-2\tilde{\zeta}_e+\mathcal{V}_e(\mathbf{R}+\tfrac{1}{2}\mathbf{r})\\&\quad+\mathcal{V}_e(\mathbf{R}-\tfrac{1}{2}\mathbf{r})+16\pi a[\tilde{f}(\mathbf{R}+\tfrac{1}{2}\mathbf{r})^2+\tilde{f}(\mathbf{R}-\tfrac{1}{2}\mathbf{r})^2]\}p_0(\mathbf{R},\mathbf{r})\\&\quad+8\pi a\int d\mathbf{r}'\,\tilde{f}(\mathbf{R}+\mathbf{r}')^2p_0(\mathbf{R}+\tfrac{1}{4}\mathbf{r}+\tfrac{1}{2}\mathbf{r}',\tfrac{1}{2}\mathbf{r}-\mathbf{r}')p_0(\mathbf{R}-\tfrac{1}{4}\mathbf{r}+\tfrac{1}{2}\mathbf{r}',-\tfrac{1}{2}\mathbf{r}-\mathbf{r}')=0.\end{aligned} \tag{4.21}$$

In the spirit of Eq. (4.9), $\nabla_{\mathbf{R}}^2$ is neglected, while

$$\mathcal{V}_e(\mathbf{R}+\tfrac{1}{2}\mathbf{r})\sim\mathcal{V}_e(\mathbf{R})\sim\mathcal{V}_e(\mathbf{R}-\tfrac{1}{2}\mathbf{r}), \tag{4.22}$$

$$\tilde{f}(\mathbf{R}+\tfrac{1}{2}\mathbf{r})\sim\tilde{f}(\mathbf{R})\sim\tilde{f}(\mathbf{R}-\tfrac{1}{2}\mathbf{r}), \tag{4.23}$$

$$p_0(\mathbf{R}+\tfrac{1}{4}\mathbf{r}+\tfrac{1}{2}\mathbf{r}',\tfrac{1}{2}\mathbf{r}-\mathbf{r}')\sim p_0(\mathbf{R},\tfrac{1}{2}\mathbf{r}-\mathbf{r}'), \tag{4.24a}$$

$$p_0(\mathbf{R}-\tfrac{1}{4}\mathbf{r}+\tfrac{1}{2}\mathbf{r}',-\tfrac{1}{2}\mathbf{r}-\mathbf{r}')\sim p_0(\mathbf{R},-\tfrac{1}{2}\mathbf{r}-\mathbf{r}'). \tag{4.24b}$$

Equation (4.21) then reduces to

$$-\nabla_{\mathbf{r}}^2 p_0(\mathbf{R},\mathbf{r})+4\pi a\tilde{f}(\mathbf{R})^2\delta(\mathbf{r})+\{-\check{\zeta}-8\pi a\tilde{\zeta}-\tilde{\zeta}_e+\mathcal{V}_e(\mathbf{R})+16\pi a\tilde{f}(\mathbf{R})^2\}p_0(\mathbf{R},\mathbf{r})$$
$$+4\pi a\tilde{f}(\mathbf{R})^2\int d\mathbf{r}'\, p_0(\mathbf{R},\mathbf{r}')p_0(\mathbf{R},\mathbf{r}-\mathbf{r}')=0. \tag{4.25}$$

Because the nonlinear term is a convolution integral, the equation of motion can be solved *exactly* with recourse to the Fourier transform in $\mathbf{r}$ of $p_0(\mathbf{R},\mathbf{r})$:

$$\overline{p_0}(\mathbf{R},\mathbf{k})=\int d\mathbf{r}e^{i\mathbf{k}\cdot\mathbf{r}}p_0(\mathbf{R},\mathbf{r}), \tag{4.26}$$

which transforms Eq. (4.25) into

$$4\pi a\tilde{f}(\mathbf{R})^2\overline{p_0}(\mathbf{R},\mathbf{k})^2+[k^2+k_0(\mathbf{R})^2]\overline{p_0}(\mathbf{R},\mathbf{k})+4\pi a\tilde{f}(\mathbf{R})^2=0, \tag{4.27}$$

where

$$k_0(\mathbf{R})^2=-\check{\zeta}-8\pi a\tilde{\zeta}-\tilde{\zeta}_e+\mathcal{V}_e(\mathbf{R})+16\pi a\tilde{f}(\mathbf{R})^2. \tag{4.28}$$

Equation (4.27) is solved explicitly to give

$$\overline{p_0}(\mathbf{R},\mathbf{k})=[8\pi a\tilde{f}(\mathbf{R})^2]^{-1}\{-k^2-k_0(\mathbf{R})^2+\sqrt{[k^2+k_0(\mathbf{R})^2]^2-(8\pi a)^2\tilde{f}(\mathbf{R})^4}\}. \tag{4.29}$$

In view of formula (4.10),

$$\overline{p_0}(\mathbf{R},\mathbf{k})\sim\begin{cases}-k_0(\mathbf{R})^{-2}\{k^2+k_0(\mathbf{R})^2-k[k^2+2k_0(\mathbf{R})^2]^{1/2}\}, & \mathbf{R}\ \text{inside}\ \mathcal{R}_0\\ 0, & \mathbf{R}\ \text{outside}\ \mathcal{R}_0\end{cases}, \tag{4.30}$$

by neglecting $\check{\zeta}$ since $|\nabla\tilde{f}(\mathbf{r})|\simeq 0$ unless $\mathbf{r}$ is sufficiently close to $\partial\mathcal{R}_0$, so that

$$k_0(\mathbf{R})^2=8\pi a\tilde{f}(\mathbf{R})^2. \tag{4.31}$$

Inversion of $\overline{p_0}(\mathbf{R},\mathbf{k})$ is carried out as follows. For $\mathbf{R}$ outside $\mathcal{R}_0$,

$$p_0(\mathbf{R},\mathbf{r})=0. \tag{4.32}$$

If $\mathbf{R}$ lies inside $\mathcal{R}_0$,

$$
\begin{aligned}
p_0(\mathbf{R},\mathbf{r}) &= \frac{1}{(2\pi)^3} \lim_{\delta\to 0^+} \int d\mathbf{k}\, e^{-i\mathbf{k}\cdot\mathbf{r}} e^{-\delta k} \overline{p_0}(\mathbf{R},\mathbf{k}) \\
&= -\frac{1}{k_0^2} \frac{1}{(2\pi)^3} \lim_{\delta\to 0^+} \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \int_0^{\infty} dk\, k^2 e^{-ik|\mathbf{r}|\cos\theta} e^{-\delta k} \\
&\quad \times [k^2 + k_0^2 - k(k^2 + 2k_0^2)^{1/2}] \\
&= \frac{2}{k_0^2|\mathbf{r}|} \frac{1}{(2\pi)^2} \lim_{\delta\to 0^+} \mathrm{Im} \int_0^{\infty} dk\, e^{ik|\mathbf{r}|} e^{-\delta k} k^2 (k^2 + 2k_0^2)^{1/2} \quad (k = \sqrt{2} k_0 \sinh t) \\
&= \frac{k_0^2}{2\pi^2|\mathbf{r}|} \lim_{\delta\to 0^+} \mathrm{Im} \int_0^{\infty} dt\, e^{i\sqrt{2}k_0|\mathbf{r}|\sinh t} e^{-\delta\sqrt{2}k_0 \sinh t} (\sinh 2t)^2 \\
&= \pi^{-2}(4\pi a)^{3/2} \tilde{f}(\mathbf{R})^3 \frac{\mathrm{Im}\{S_{0,4}(iw) - S_{0,0}(iw)\}}{w},
\end{aligned}
\tag{4.33}
$$

where $k_0 = k_0(\mathbf{R})$ and

$$
w = (16\pi a)^{1/2} \tilde{f}(\mathbf{R}) |\mathbf{r}|, \tag{4.34}
$$

and $S_{0,4}$ and $S_{0,0}$ are Lommel's functions.[31]

V. $\tilde{f}(\mathbf{r})$ IN A THREE-DIMENSIONAL SPHERICALLY SYMMETRIC HARMONIC POTENTIAL

In the actual experiments on Bose–Einstein condensation, the trapping potential is of complicated form. This is usually modeled as an anisotropic harmonic potential. In this section, $\tilde{f}(\mathbf{r})$ for the lowest state of the condensate is examined in some detail in the simplifying case of a spherically symmetric harmonic potential. A similar task is undertaken in Ref. 12, where the nonlinear Schrödinger equation is given in terms of the chemical potential.

With an external potential $V_e(\mathbf{r},t) = \frac{1}{4}\omega_0^2 r^2$, $\tilde{\mathcal{V}}_1(\mathbf{r})$ is taken to be

$$
\tilde{\mathcal{V}}_1(\mathbf{r}) = \tfrac{1}{4}\omega_0^2 r^2 - \mathcal{E}, \tag{5.1}
$$

as is suggested by the eigenvalue problem associated with Eq. (2.56). Terms linear in x, y, and z are omitted. It follows that

$$
\tilde{\boldsymbol{\mathcal{V}}}_2(t) = \tfrac{1}{2}\omega_0^2 \boldsymbol{\alpha}(t), \quad \tilde{\mathcal{V}}_3(t) = \mathcal{E} - \tfrac{1}{4}\omega_0^2 |\boldsymbol{\alpha}(t)|^2, \tag{5.2}
$$

yielding

$$
\boldsymbol{\alpha}(t) = \mathbf{r}_0 \cos\omega_0 t + \frac{\mathbf{v}_0}{\omega_0} \sin\omega_0 t, \tag{5.3}
$$

$$
A(t) = \mathcal{E}t + \frac{1}{8}\left(\frac{|\mathbf{v}_0|^2 - \omega_0^2|\mathbf{r}_0|^2}{\omega_0} \sin 2\omega_0 t + 2\mathbf{v}_0\cdot\mathbf{r}_0 \cos 2\omega_0 t \right) + \text{const}, \tag{5.4}
$$

where $\mathbf{r}_0$ and $\mathbf{v}_0$ are determined by the initial conditions and the constant is real.

For the state of lowest energy $\mathcal{E} = \mathcal{E}_0$, $\tilde{f}(\mathbf{r}) = \tilde{f}_0(r)$ is spherically symmetric.[32] Let

$$
q(\xi) = (4\pi)^{1/2} (N^2\omega_0/2)^{-1/4} r \tilde{f}_0(r), \quad \xi = (\omega_0/2)^{1/2} r. \tag{5.5}
$$

From Eq. (4.8a), this $q(\xi)$ satisfies

$$-\frac{d^2q(\xi)}{d\xi^2}+\xi^2q(\xi)+\Lambda^2\frac{q(\xi)^3}{\xi^2}=\lambda^2q(\xi)\quad(\xi>0),\tag{5.6}$$

supplemented with the boundary conditions

$$q(0)=0,\tag{5.7}$$

$$\lim_{\xi\to\infty}q(\xi)=0,\tag{5.8}$$

and the normalization condition

$$\int_0^\infty d\xi\, q(\xi)^2=1.\tag{5.9}$$

In the above,

$$\Lambda=(2a^2N^2\omega_0)^{1/4},\tag{5.10}$$

$$\lambda^2=\frac{2\mathcal{E}}{\omega_0}+\frac{1}{2}\Lambda^2\int_0^\infty\frac{d\xi}{\xi^2}q(\xi)^4,\quad \lambda>0.\tag{5.11}$$

Note that, for $\xi\to\infty$, the nonlinear term in Eq. (5.6) can be neglected, and the asymptotic behavior of $q(\xi)$ is found via the direct application of the Wentzel–Kramers–Brillouin method:

$$q(\xi)\sim C(\xi^2-\lambda^2)^{-1/4}\exp\{-(\lambda^2/2)[(\xi/\lambda)\sqrt{(\xi/\lambda)^2-1}-\cosh^{-1}(\xi/\lambda)]\},\tag{5.12}$$

where C is independent of ξ. Compare with Ref. 12. For a discussion on the determination of this C see Ref. 33.

Some insight into the solution to Eqs. (5.6)–(5.9) can be obtained by considering the following cases.

(i) $\Lambda\gg1$. To leading order in Λ, neglect of the second derivative of $q(\xi)$ results in

$$q(\xi)\sim q^{(0)}(\xi)=\begin{cases}(\xi/\Lambda)\sqrt{\lambda^2-\xi^2}, & 0\leqslant\xi<\lambda\\ 0, & \xi>\lambda,\end{cases}\tag{5.13}$$

which trivially satisfies Eqs. (5.7) and (5.8). $q^{(0)}(\xi)$ satisfies Eq. (5.9) provided that λ is

$$\lambda\sim\lambda^{(0)}=(\tfrac{15}{2}\Lambda^2)^{1/5}.\tag{5.14}$$

A similar calculation for an anisotropic potential can be found in Ref. 14, where the chemical potential is employed. From Eq. (5.11),

$$\mathcal{E}^{(0)}=\tfrac{5}{14}(\tfrac{15}{2})^{2/5}(2a^2N^2\omega_0)^{1/5}\omega_0=\tfrac{5}{21}(\tfrac{15}{2})^{2/5}\Lambda^{4/5}\epsilon_0^{\text{ho}},\tag{5.15}$$

where $\epsilon_0^{\text{ho}}=\frac{3}{2}\omega_0$ is the ground-state energy of the three-dimensional harmonic oscillator. Approximation (5.13) starts to break down at a distance of the order of $\Lambda^{-2/15}$ from inside the ''boundary'' $\xi=\lambda$, and then needs to be modified according to the procedure in Refs. 8 and 13. This procedure provides a smooth connection to asymptotic formula (5.12) when $0<\xi-\lambda\ll1$ while $\xi-\lambda\gg O(\Lambda^{-2/15})$.[33]

(ii) $\Lambda\ll1$. To zeroth order in Λ, the known solution for the ground-state wave function of the three-dimensional harmonic oscillator is obtained:

$$q^{(0)}(\xi)=2\pi^{-1/4}\xi e^{-\xi^2/2},\tag{5.16}$$

with energy $\mathcal{E}^{(0)}=(\omega_0/2)\lambda^{(0)^2}=\epsilon_0^{\text{ho}}$. The first-order energy correction $\mathcal{E}^{(1)}$ can be obtained

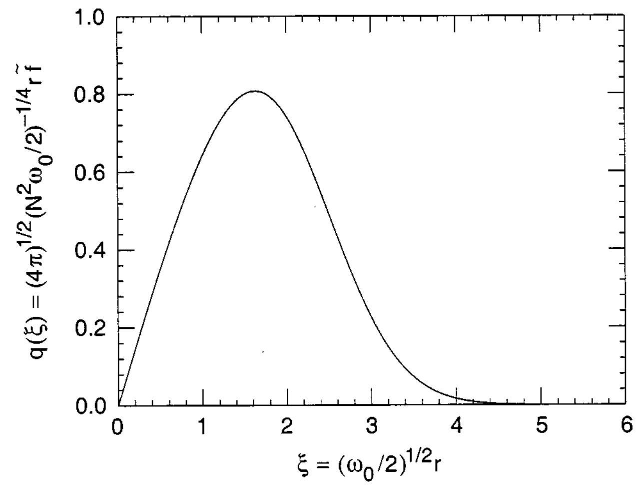

FIG. 1. Solution to Eqs. (5.6)–(5.9) for $\Lambda^2=12.1$ (^{87}Rb atoms, $a\simeq 110a_0$, $N=10^3$, and $\omega_0=(2\pi\times 120)/\sqrt{8}$ rad/s). Numerically computed eigenvalue is $\lambda^2=6.8$.

through the standard perturbation methods, by treating $V^{(0)}(\xi)=\Lambda^2\xi^{-2}q^{(0)}(\xi)^2$ as the perturbing potential. Therefore, $\lambda^{(1)^2}$ equals the matrix element

$$\lambda^{(1)^2}=\int_0^\infty d\xi\, q^{(0)}(\xi)V^{(0)}(\xi)q^{(0)}(\xi). \tag{5.17}$$

By virtue of Eq. (5.11),

$$\mathcal{E}^{(1)}=\frac{\omega_0\Lambda^2}{4}\int_0^\infty \frac{d\xi}{\xi^2}q^{(0)}(\xi)^4=\sqrt{\frac{2}{\pi}}\frac{\omega_0\Lambda^2}{4},\quad \Lambda\ll 1, \tag{5.18}$$

or

$$\mathcal{E}\sim\mathcal{E}^{(0)}+\mathcal{E}^{(1)}=\frac{3}{2}\omega_0+\sqrt{\frac{2}{\pi}}\frac{\omega_0\Lambda^2}{4}. \tag{5.19}$$

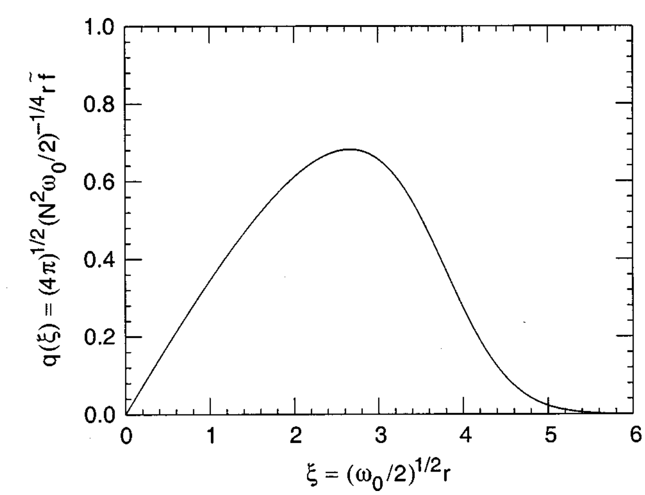

FIG. 2. Solution to Eqs. (5.6)–(5.9) for $\Lambda^2=121$ (^{87}Rb atoms, $N=10^4$, and $\omega_0=(2\pi\times 120)/\sqrt{8}$ rad/s). Eigenvalue is $\lambda^2=15.6$.

Numerical results. In order to make some contact with recent experimental situations, plots of $q(\xi)$ are presented in Figs. 1 and 2 for two different values of Λ, in close relation to the JILA experiments, where ^{87}Rb atoms were used ($a=110a_0$, a_0: the Bohr radius).[1] Specifically, in Fig. 1, $\Lambda^2=12.1$, corresponding, for instance, to $N=10^3$ and $\omega_0=(2\pi\times 120)/\sqrt{8}$ rad/s. The numerically computed eigenvalue there is $\lambda^2=6.8$, giving $\mathcal{E}=1.8\epsilon_0^{\text{ho}}$. Compare with $\mathcal{E}^{(0)}=1.45\epsilon_0^{\text{ho}}$ provided by Eq. (5.15). In Fig. 2, $\Lambda^2=121$. The corresponding eigenvalue is found to be $\lambda^2=15.6$, giving $\mathcal{E}=3.8\epsilon_0^{\text{ho}}$. Compare with $\mathcal{E}^{(0)}=3.63\epsilon_0^{\text{ho}}$ from Eq. (5.15).

VI. CONCLUSIONS AND DISCUSSION

In the theoretical treatment of Bose–Einstein condensation in dilute atomic gases with repulsive interactions, the trap is replaced by a sufficiently smooth external potential $V_e(\mathbf{r},t)$ that acts simultaneously on each atom and increases sufficiently rapidly at large distances. As a consequence, the boson system is no longer translationally invariant. Work carried out 38 years ago[4] turns out to be a suitable starting point. An important element introduced there was the systematic consideration of scattering processes, such as pair creation, with a study of some of their physical consequences. In the presence of a trapping potential, pair creation plays a significant role, being described mathematically by the pair-excitation function $K_0(\mathbf{r},\mathbf{r}';t)$. On the basis of the ansatz (1.1), a nonlinear integro-differential equation is satisfied by $K_0(\mathbf{r},\mathbf{r}';t)$.

Solitary-wave solutions to the nonlinear evolution equations for the condensate wave function $\Phi(\mathbf{r},t)$ and the pair-excitation function are uncovered in any number of space dimensions, if $V_e(\mathbf{r},t)$ can properly be decomposed into (i) a trapping potential $\mathcal{V}_e$ translated by the position vector $\mathbf{r}(t)=\boldsymbol{\alpha}(t)$ of the pulse ''center of mass,'' and (ii) a potential linear in the space coordinates, according to (2.55)–(2.58). It is somewhat tempting to put these statements in the language of classical mechanics, recognizing, for instance, the second term mentioned above as the potential associated with a uniform force. The conclusions here are the natural generalization of results obtained for the one-dimensional case, without any restriction to motion along fixed axes in space.[18] Given an external potential that meets the aforementioned conditions, the initial amplitudes are obtained by solving a nonlinear ''eigenvalue problem'' for $\Phi(\mathbf{r},t=0)$ under $\mathcal{V}_e$, and a nonlinear integro-differential equation for $K_0(\mathbf{r},\mathbf{r}';t=0)$. The motion of the solitary wave in space, i.e., the vector $\boldsymbol{\alpha}(t)$, is determined by the uniform force. In this sense, the solitary wave is expected to behave like a classical particle. Conversely, given an admissible $\Phi(\mathbf{r},t=0)$, i.e., sufficiently smooth and rapidly decreasing to zero as $\mathbf{r}\rightarrow\infty$, it is possible to construct an external potential that permits solitary-wave behavior for both $\Phi(\mathbf{r},t>0)$ and $K_0(\mathbf{r},\mathbf{r}';t>0)$. Of course, in real experimental situations, the form of the external potential may deviate from the one given by Eq. (2.55). The question of the stability of the solitary-wave solutions under variations of $V_e(\mathbf{r},t)$ is not addressed in this paper.

As is also pointed out in Ref. 8, the approximate Hamiltonian that furnishes the equation of motion for K_0 does not include, for instance, the scattering of phonons and the decay of a single phonon into two or three phonons. In other words, under the present approximation, the phonons have infinite lifetimes and remain stable. This in turn implies that the ansatz (1.1) and the existing equations of motion are of rather special forms, being valid only over some moderate time scale. The problem of shorter or longer time scales is not touched upon in this paper; this time limitation may depend on the higher-order terms in the Hamiltonian or the initial condition for the condensate wave function. It is believed that the ansatz for the many-body wave function can be generalized. A challenging open problem is to obtain such generalizations, which must satisfy many consistency conditions.

ACKNOWLEDGMENTS

The author is greatly indebted to Professor Tai Tsun Wu for his unwavering guidance and encouragement. He also wishes to thank Professor Wolfgang Ketterle and Professor Bertrand

Halperin for useful discussions, and Margaret Owens for her valuable assistance throughout the preparation of the manuscript. This work was supported in part by the United States Department of Energy under Grant No. DE-FG02-84ER40158.

APPENDIX

Consider the one-dimensional linear Schrödinger equation in a harmonic-oscillator potential ($\hbar=2m=1,\ \omega_0=2$):

$$i\,\frac{\partial\varphi(x,t)}{\partial t}=\left(-\frac{\partial^2}{\partial x^2}+x^2\right)\varphi(x,t). \tag{A1}$$

It is well known[34] that an initial displacement of the ground-state wave function $\varphi_0(x)=\pi^{-1/4}e^{-x^2/2}$ at $t=0$ by $x=x_0$ produces the wave packet

$$\varphi(x,t)=\varphi_0(x-x_0e^{-2it})e^{-i\nu(t)},\quad t>0, \tag{A2}$$

where, for definiteness, $\nu(t=0)=0$. Substitution into Eq. (A1) furnishes

$$\nu(t)=t+\frac{ix_0^2}{4}e^{-i4t}-\frac{ix_0^2}{4}. \tag{A3}$$

$\varphi(x,t)$ is subsequently recast in a form where magnitude and phase are separated:

$$\begin{aligned}\varphi(x,t)&=\pi^{-1/4}\exp\left\{-\frac{1}{2}(x-x_0e^{-2it})^2-it+\frac{x_0^2}{4}e^{-i4t}-\frac{x_0^2}{4}\right\}\\&=\pi^{-1/4}\exp\left[-\frac{1}{2}(x-x_0\cos 2t)^2\right]\exp\left[-i\left(t+x_0x\sin 2t-\frac{x_0^2}{4}\sin 4t\right)\right],\end{aligned} \tag{A4}$$

which is a one-dimensional solitary wave. Note that with the units of Eq. (A1) the eigenvalue corresponding to $\varphi_0(x)$ is equal to 1.

The preceding analysis can be extended to the d-dimensional Schrödinger equation

$$i\,\frac{\partial\varphi(\mathbf{r},t)}{\partial t}=\left(-\nabla^2+\sum_{j=1}^{d}x_j^2\right)\varphi(\mathbf{r},t), \tag{A5}$$

where $\mathbf{r}=(x_1,\ldots,x_d)$, $d\geqslant 2$. With an initial displacement of the ground-state wave function $\varphi_0(\mathbf{r})=\pi^{-d/4}e^{-\mathbf{r}\cdot\mathbf{r}/2}$ by $\mathbf{r}=\mathbf{r}_0$, at later times $\varphi(\mathbf{r},t)$ becomes

$$\varphi(\mathbf{r},t)=\varphi_0(\mathbf{r}-\mathbf{r}_0e^{-2it})e^{-i\nu(t)},\quad t>0. \tag{A6}$$

After some straightforward algebra,

$$\nu(t)=d\cdot t+\frac{i|\mathbf{r}_0|^2}{4}e^{-i4t}-\frac{i|\mathbf{r}_0|^2}{4}, \tag{A7}$$

$$\begin{aligned}\varphi(\mathbf{r},t)&=\pi^{-d/4}\left(\prod_{j=1}^{d}\exp\left[-\frac{1}{2}(x_j-x_{j0}\cos 2t)^2\right]\right)\exp\left[-i\left(d\cdot t+\mathbf{r}_0\cdot\mathbf{r}\sin 2t-\frac{|\mathbf{r}_0|^2}{4}\sin 4t\right)\right]\\&=\varphi_0(\mathbf{r}-\mathbf{r}_0\cos 2t)\exp\left[-i\left(d\cdot t+\mathbf{r}_0\cdot\mathbf{r}\sin 2t-\frac{|\mathbf{r}_0|^2}{4}\sin 4t\right)\right].\end{aligned} \tag{A8}$$

This is a solitary wave in d space dimensions.

[1] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell, Science **269**, 198 (1995).

[2] C. C. Bradley, C. A. Sackett, J. J. Tollett, and R. G. Hulet, Phys. Rev. Lett. **75**, 1687 (1995); C. C. Bradley, C. A. Sackett, and R. G. Hulet, *ibid.* **78**, 985 (1997).
[3] K. B. Davis, M.-O. Mewes, M. R. Andrews, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle, Phys. Rev. Lett. **75**, 3969 (1995).
[4] T. T. Wu, J. Math. Phys. **2**, 105 (1961).
[5] E. P. Gross, Ann. Phys. (N.Y.) **4**, 57 (1958); **9**, 292 (1960); Nuovo Cimento **20**, 454 (1961); J. Math. Phys. **4**, 195 (1963).
[6] L. P. Pitaevskii, Zh. Eksp. Teor. Fiz. **40**, 646 (1961) [Sov. Phys. JETP **13**, 451 (1961)].
[7] T. T. Wu, in *Quantum Chromodynamics: Collisions, Confinement, and Chaos*, edited by H. Fried and B. Müller (World Scientific, Singapore, 1997), pp. 76–81.
[8] T. T. Wu, Phys. Rev. A **58**, 1465 (1998).
[9] D. Margetis and T. T. Wu, Phys. Rev. A (submitted).
[10] T. D. Lee, K. Huang, and C. N. Yang, Phys. Rev. **106**, 1135 (1957).
[11] M. Edwards, R. J. Dodd, C. W. Clark, and K. Burnett, J. Res. Natl. Inst. Stand. Technol. **101**, 553 (1996); R. J. Dodd, *ibid.* **101**, 545 (1996); M. Edwards, R. J. Dodd, C. W. Clark, P. A. Ruprecht, and K. Burnett, Phys. Rev. A **53**, R1950 (1996); G. Baym and C. Pethick, Phys. Rev. Lett. **76**, 6 (1996).
[12] M. Edwards and K. Burnett, Phys. Rev. A **51**, 1382 (1995).
[13] F. Dalfovo, L. Pitaevskii, and S. Stringari, Phys. Rev. A **54**, 4213 (1996); J. Res. Natl. Inst. Stand. Technol. **101**, 537 (1996).
[14] F. Dalfovo and S. Stringari, Phys. Rev. A **53**, 2477 (1996).
[15] Some recent papers dealing with time-dependent solutions to the nonlinear Schrödinger equation are: P. A. Ruprecht, M. J. Holland, K. Burnett, and M. Edwards, Phys. Rev. A **51**, 4704 (1995); M. Holland and J. Cooper, *ibid.* **53**, R1954 (1996); Yu. Kagan, E. L. Surkov, and G. V. Shlyapnikov, *ibid.* **54**, R1753 (1996); Y. Castin and R. Dum, Phys. Rev. Lett. **77**, 5315 (1996); A. Smerzi and S. Fantoni, *ibid.* **78**, 3589 (1997); V. M. Pérez-García, H. Michinel, J. I. Chirac, M. Lewenstein, and P. Zoller, Phys. Rev. A **56**, 1424 (1997); R. Dum, A. Sanpera, K.-A. Suominen, M. Brewczyk, M. Kuś, K. Rzążewski, and M. Lewenstein, Preprint quant-ph/9708009, xxx.lanl.gov.
[16] S. C. Benjamin, L. Quiroga, and N. F. Johnson, Phys. Rev. A **54**, 4309 (1996). It is worthy of notice that the interatomic interactions there are assumed to be of the form const/$|r_i - r_j|^2 (i \neq j)$. However, even approximate analytical solutions to the resulting hyperangular equation seem to be intractable.
[17] Some papers concerned with soliton solutions to nonlinear Schrödinger equations containing an external potential are the following: H.-H. Chen and C.-S. Liu, Phys. Rev. Lett. **37**, 693 (1976); Phys. Fluids **21**, 377 (1978); A. C. Newell, J. Math. Phys. **19**, 1126 (1978); D. J. Kaup and A. C. Newell, *ibid.* **19**, 798 (1978); Proc. R. Soc. London, Ser. A **361**, 413 (1978); R. W. Hasse, Phys. Rev. A **25**, 583 (1982); J. E. Herrera, J. Phys. A **17**, 95 (1984); R. Balakrishnan, Phys. Rev. A **32**, 1144 (1985); A. B. Nassar, *ibid.* **33**, 3502 (1986); M. A. de Moura, *ibid.* **37**, 4998 (1988); Y. Nogami and F. M. Toyama, Phys. Rev. E **49**, 4497 (1994); M. A. de Moura, J. Phys. A **27**, 7157 (1994).
[18] S. A. Morgan, R. J. Ballagh, and K. Burnett, Phys. Rev. A **55**, 4338 (1997).
[19] R. S. Palais, Bull. Am. Math. Soc. **34**, 339 (1997).
[20] An article about the original computation can be found in E. Fermi, J. Pasta, and S. Ulam, in *Nonlinear Wave Motion*, Lectures in Applied Mathematics Vol. 15, edited by A. C. Newell (American Mathematical Society, Providence, RI, 1974), pp. 143–155.
[21] N. J. Zabusky and M. D. Kruskal, Phys. Rev. Lett. **15**, 240 (1965).
[22] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Phys. Rev. Lett. **19**, 1095 (1967); Commun. Pure Appl. Math. **27**, 97 (1974).
[23] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, New York, 1991).
[24] For example, see the classical papers V. E. Zakharov and A. B. Shabat, Zh. Eksp. Teor. Fiz. **61**, 118 (1971) [Sov. Phys. JETP **34**, 62 (1972)]; Zh. Eksp. Teor. Fiz. **64**, 1627 (1973) [Sov. Phys. JETP **37**, 823 (1973)]; R. Hirota, J. Math. Phys. **14**, 805 (1973).
[25] A. D. Jackson, G. M. Kavoulakis, and C. J. Pethick, Preprint cond-mat/9803116, xxx.lanl.gov.
[26] V. M. Pérez-García, H. Michinel, and H. Herrero, Phys. Rev. A **57**, 3837 (1998).
[27] Y. Castin and R. Dum, Phys. Rev. Lett. **79**, 3553 (1997); C. W. Gardiner, Phys. Rev. A **56**, 1414 (1997); Y. Castin and R. Dum, *ibid.* **57**, 3008 (1998).
[28] J. M. Irvine, *Nuclear Structure Theory* (Pergamon, Oxford, 1972), Chap. 13.
[29] Of course, C_0 can be determined by prescribing the value of the potential at some fixed point x_0.
[30] For three dimensions, see the classical book O. D. Kellogg, *Foundations of Potential Theory* (Dover, New York, 1953), p. 223; for two dimensions, see L. V. Ahlfors, *Complex Analysis* (McGraw–Hill, Singapore, 1979), p. 165; for generalization to higher space dimensions, see, for instance, D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order* (Springer, New York, 1983), p. 21.
[31] Bateman Manuscript Project, *Higher Transcendental Functions*, edited by A. Erdèlyi (McGraw-Hill, New York, 1953), Vol. II, pp. 40, 41, 84.
[32] S. Coleman, V. Glaser, and A. Martin, Commun. Math. Phys. **58**, 211 (1978).
[33] D. Margetis, Ph.D. thesis, Harvard University, 1999.
[34] This is an elementary result in quantum mechanics. For generalizations, see K. Husimi, Prog. Theor. Phys. **9**, 381 (1953); E. H. Kerner, Can. J. Phys. **36**, 371 (1958).

A generalization of Wigner's unitary–antiunitary theorem to Hilbert modules

Lajos Molnár
Institute of Mathematics and Informatics, Lajos Kossuth University, 4010 Debrecen, P.O. Box 12, Hungary

(Received 4 May 1999; accepted for publication 30 June 1999)

Let $\mathcal{H}$ be a Hilbert C^*-module over a matrix algebra A. It is proved that any function $T:\mathcal{H}\to\mathcal{H}$ which preserves the absolute value of the (generalized) inner product is of the form $Tf=\varphi(f)Uf$ $(f\in\mathcal{H})$, where φ is a phase-function and U is an A-linear isometry. The result gives a natural extension of Wigner's classical unitary–antiunitary theorem for Hilbert modules. © *1999 American Institute of Physics.* [S0022-2488(99)01611-4]

I. INTRODUCTION AND STATEMENT OF THE RESULT

Wigner's unitary–antiunitary theorem reads as follows. Let H be a complex Hilbert space and let $T:H\to H$ be a bijective function (linearity or continuity is not assumed) with the property that

$$|\langle Tx,Ty\rangle|=|\langle x,y\rangle| \quad (x,y\in H).$$

Then T is of the form

$$Tx=\varphi(x)Ux \quad (x\in H),$$

where $U:H\to H$ is either a unitary or an antiunitary operator and $\varphi:H\to\mathbb{C}$ is a so-called phase-function which means that its values are of modulus 1. This celebrated result plays a very important role in quantum mechanics and in representation theory in physics.

In our recent paper[1] we presented a new, algebraic approach to this theorem. Our idea turned out to be strong enough to give a natural generalization of Wigner's theorem for Hilbert C^*-modules over matrix algebras. However, in the main result [Ref. 1, Theorem 1] we supposed that our map is surjective and, in addition, a condition was imposed on the underlying module which was proved to be equivalent to that its so-called modular dimension is high enough. In the present paper, refining and modifying our argument quite significantly, we obtain our Wigner-type result in full generality, that is, neither the surjectivity of the transformation in question nor the high dimensionality of the Hilbert module is assumed.

First, we clarify the concepts and notation that we are going to use throughout. For a bit more detailed discussion we refer to the introduction of Ref. 1. Let A be a C^*-algebra. Let $\mathcal{H}$ be a left A-module with a map $[.,.]:\mathcal{H}\times\mathcal{H}\to A$ satisfying

(i) $[f+g,h]=[f,h]+[g,h]$;
(ii) $[af,g]=a[f,g]$;
(iii) $[g,f]=[f,g]^*$;
(iv) $[f,f]\geqslant 0$ and $[f,f]=0$ if and only if $f=0$

for every $f,g,h\in\mathcal{H}$ and $a\in A$. If $\mathcal{H}$ is complete with respect to the norm $f\mapsto\|[f,f]\|^{1/2}$, then we say that $\mathcal{H}$ is a Hilbert A-module or a Hilbert C^*-module over A with generalized inner product [.,.]. Nowadays, Hilbert modules over C^*-algebras play a very important role in many parts of functional analysis such as, for example, in the K-theory of C^*-algebras. There is another concept of Hilbert modules due to Saworotnow.[2] These are modules over H^*-algebras. The only formal difference in the definition is that in the case of Saworotnow's modules, the generalized inner product takes its values in the trace-class of the underlying H^*-algebra and the norm with respect to which we require completeness is $f \mapsto (\mathrm{tr}[f,f])^{1/2}$. Saworotnow's modules appear naturally when dealing with multivariate stochastic processes and they have applications in Clifford analysis and hence in some parts of mathematical physics.

If the underlying C^*-algebra A is the algebra $M_d(\mathbb{C})$ of all $d \times d$ complex matrices, then, A being finite dimensional, the norms on A are all equivalent. Therefore, the Hilbert C^*-modules over the C^*-algebra $M_d(\mathbb{C})$ are the same as Saworotnow's Hilbert modules over the H^*-algebra $M_d(\mathbb{C})$. We emphasize this fact since, in general, the behavior of Saworotnow's Hilbert modules is much nicer and we shall use several results concerning them. Finally, we note that it seems to be more common to use right modules instead of left ones. Of course, this is not a real difference, only a question of taste.

Now we are in a position to formulate the main result of the paper. Recall that in any C^*-algebra A, the element $|a|$ denotes the square root of a^*a $(a \in A)$.

Theorem: *Let $\mathcal{H}$ be a Hilbert C^*-module over the matrix algebra $A = M_d(\mathbb{C})$, $d > 1$. Let $T:\mathcal{H} \to \mathcal{H}$ be a function with the property that*

$$|[Tf,Tf']| = |[f,f']| \quad (f,f' \in \mathcal{H}). \tag{1}$$

Then there exists an A-isometry $U:\mathcal{H} \to \mathcal{H}$ and a phase-function $\varphi:\mathcal{H} \to \mathbb{C}$ such that

$$Tf = \varphi(f)Uf \quad (f \in \mathcal{H}).$$

Here, A-isometry means that $U:\mathcal{H} \to \mathcal{H}$ is a linear map with $U(af) = aUf$ and $[Uf,Uf'] = [f,f']$ $(a \in A,\ f,\ f' \in \mathcal{H})$.

The corresponding result for the case $d = 1$, that is, when $\mathcal{H}$ is a Hilbert space, can be found in Refs. 3 and 4 (for a recent paper also see Ref. 5). As we shall see in the proof, the nonappearance of A-anti-isometries in the above result is the consequence of the noncommutativity of the underlying algebra A.

Hilbert spaces over algebras different from $\mathbb{R}$ and $\mathbb{C}$ do appear in mathematical physics (see, for example, Ref. 6 for a Wigner-type theorem concerning Hilbert spaces over the skew-field of quaternions). We believe that our present result may also have physical interpretation.

II. PROOF

We give some additional definitions and notation that we shall use in the proof of our theorem. As mentioned in the introduction, Saworotnow's modules have many convenient properties which are familiar in the theory of Hilbert spaces (we refer to Ref. 2). First of all, if $\mathcal{H}$ is a Hilbert module over an H^*-algebra, then $\mathcal{H}$ is a Hilbert space with the inner product $\langle .,. \rangle = \mathrm{tr}[.,.]$. If $\mathcal{M} \subset \mathcal{H}$ is a closed submodule, then its orthogonal complement with respect to $\langle .,. \rangle$ and $[.,.]$ are the same. A linear operator T on $\mathcal{H}$ which is bounded with respect to the Hilbert space norm defined above is called an A-linear operator if $T(af) = aTf$ holds true for every $f \in \mathcal{H}$ and $a \in A$. Every A-linear operator T is adjointable, namely, the adjoint T^* of T in the Hilbert space sense is A-linear and we have $[Tf,g] = [f,T^*g]$ $(f,g \in \mathcal{H})$. Consequently, the collection of all A-linear operators forms a C^*-subalgebra in the full operator algebra on the Hilbert space $\mathcal{H}$. This will be denoted by $\mathcal{B}(\mathcal{H})$ while the notation of the full operator algebra over a Hilbert space H is $B(H)$.

In the case of a Hilbert module $\mathcal{H}$ over an H^*-algebra, the natural equivalent of the Hilbert basis is the so-called modular basis.[7] An element $f \in \mathcal{H}$ is called a modular unit vector, if $[f,f]$ is a nonzero minimal projection in A. A family $\{f_\alpha\}_\alpha \subset \mathcal{H}$ is said to be modular orthonormal if

(a) $[f_\alpha, f_\beta] = 0$ if $\alpha \neq \beta$,
(b) f_α is a modular unit vector for every α.

A maximal modular orthonormal family of vectors in $\mathcal{H}$ is called a modular basis. The common cardinality of modular bases in $\mathcal{H}$ is called the modular dimension of $\mathcal{H}$ (see Ref. 7, Theorem 2).

Now, we define operators which are the natural equivalent of the finite rank operators in the case of Hilbert spaces. If $f,g \in \mathcal{H}$, then let $f \odot g$ denote the A-linear operator defined by

$$(f \odot g)h = [h,g]f \quad (h \in \mathcal{H}).$$

It is easy to see that for every A-linear operator S we have

$$S(f \odot g) = (Sf) \odot g, \quad (f \odot g)S = f \odot (S^* g)$$

and

$$(f \odot g)(f' \odot g') = ([f',g]f) \odot g' = f \odot ([g,f']g').$$

Define

$$\mathcal{F}(\mathcal{H}) = \left\{ \sum_{k=1}^{n} f_k \odot g_k : f_k, g_k \in \mathcal{H} (k=1,...,n), n \in \mathbb{N} \right\}$$

which is a *-ideal in the C^*-algebra of all A-linear operators. Observe that if $\mathcal{H}$ is a Hilbert module over $M_d(\mathbb{C})$, then the range of every element of $\mathcal{F}(\mathcal{H})$ has finite linear dimension, but there can be finite rank operators on the Hilbert space $\mathcal{H}$ which do not belong to $\mathcal{F}(\mathcal{H})$. In general, if the underlying H^*-algebra is infinite dimensional, then these two classes of operators have nothing to do with each other.

We begin with some auxiliary results that we shall need in the proof of our theorem.

Lemma 1: Let $A = M_d(\mathbb{C})$, $d \in \mathbb{N}$. If $\mathcal{H}$ is a Hilbert A-module, then every projection in $\mathcal{B}(\mathcal{H})$ is of the form $P = \Sigma_\alpha f_\alpha \odot f_\alpha$, where $\{f_\alpha\}_\alpha \subset \mathcal{H}$ is a modular orthonormal basis in the range of P (the range of an A-linear projection is a closed submodule).

If $\{f_\alpha\}_\alpha \subset \mathcal{H}$ is a modular orthonormal set, then for the orthogonal projection onto the closed submodule generated by $\{f_\alpha\}_\alpha$ (which is an A-linear projection) we have $P = \Sigma_\alpha f_\alpha \odot f_\alpha$.

Proof: Let first $P \in \mathcal{B}(\mathcal{H})$ be a projection and let $\{f_\alpha\}_\alpha$ denote a modular orthonormal basis in the closed submodule $\operatorname{rng} P$. By Ref. 7 Theorem 1, we have

$$f = \sum_\alpha [f,f_\alpha] f_\alpha \quad (f \in \operatorname{rng} P).$$

Since $Pf = 0$ and $[f,f_\alpha] = 0$ for $f \in \operatorname{rng} P^\perp$, we obtain $P = \Sigma_\alpha f_\alpha \odot f_\alpha$.

Now, let $\{f_\alpha\}_\alpha \subset \mathcal{H}$ be a modular orthonormal set and denote $\mathcal{M}$ the closed submodule generated by this set. We show that $\{f_\alpha\}_\alpha$ is a modular basis in $\mathcal{M}$. Since this collection is a modular orthonormal family, if this was not maximal, then we could find a nonzero element $f \in \mathcal{M}$ which is modular orthogonal to $\{f_\alpha\}_\alpha$, that is, $[f,f_\alpha] = 0$ for every α. But this is a contradiction, since every element of $\mathcal{M}$ can be approximated by finite sums of the form $a_1 f_{\alpha_1} + \cdots + a_n f_{\alpha_n}$ $(a_i \in A)$ and hence we would obtain that f is modular orthogonal to itself. By the first part of the proof we obtain that the orthogonal projection onto $\mathcal{M}$ is equal to $\Sigma_\alpha f_\alpha \odot f_\alpha$, so this operator is an A-linear projection. □

Lemma 2: Let $A = M_d(\mathbb{C})$, $d \in \mathbb{N}$ and let $\mathcal{H}$ be a Hilbert A-module. Suppose that $\mathcal{M} \subset \mathcal{H}$ is a closed submodule and $\{f_\alpha\}_\alpha$ is a modular orthonormal system generating $\mathcal{M}$. Then for every $g,h \in \mathcal{M}$ we have

(i) $g = \Sigma_\alpha [g,f_\alpha] f_\alpha$,
(ii) $[g,h] = \Sigma_\alpha [g,f_\alpha][f_\alpha,h]$.

Moreover, the vector $k \in \mathcal{H}$ belongs to $\mathcal{M}$ if and only if

$$[k,k]=\sum_{\alpha}[k,f_\alpha][f_\alpha,k].$$

Proof: See Ref. 7, Theorem 2, and its proof. □

Proposition 3: Let $A=M_d(\mathbb{C})$, $d\in\mathbb{N}$. If $\mathcal{H}$ is a Hilbert A-module, then $\mathcal{B}(\mathcal{H})$ is a type I von Neumann factor. If the modular dimension of $\mathcal{H}$ is greater than 2, then $\mathcal{B}(\mathcal{H})$ is not isomorphic to $M_2(\mathbb{C})$.

Proof: It is clear that $\mathcal{B}(\mathcal{H})$ is a von Neumann algebra since it is the commutant of the set $\{L_a : L_a f=af(f\in\mathcal{H},a\in A)\}$ in the full operator algebra over $\mathcal{H}$ as a Hilbert space. To show that $\mathcal{B}(\mathcal{H})$ is a factor, it is sufficient to verify that the central projections in $\mathcal{B}(\mathcal{H})$ are all trivial. Let $P\in\mathcal{B}(\mathcal{H})$ be a nonzero central projection. Let f be a modular unit vector in rng P. For any $a,b\in A$ we have

$$P\cdot g\odot(af)=g\odot(af)\cdot P=g\odot P(af)=g\odot(af).$$

This implies that

$$P(b[f,f]a^*g)=P((g\odot(af))bf)=(g\odot(af))bf=b[f,f]a^*g.$$

The element $[f,f]$ is a rank-one projection. Hence, every element of A is the sum of $b[f,f]a^*$-type elements and hence we obtain that $Pg=g$ for every $g\in\mathcal{H}$. Thus $P=I$. So, $\mathcal{B}(\mathcal{H})$ is a factor. We next prove that $\mathcal{B}(\mathcal{H})$ is type I. Let $f\in\mathcal{H}$ be a modular unit vector. Since $[f,f]f=f$ (see Ref. 7, Lemma 1), for any $A\in\mathcal{B}(\mathcal{H})$ we compute

$$f\odot f\cdot A\cdot f\odot f=([Af,f]f)\odot f=([A([f,f]f),[f,f]f]f\odot f)=([f,f][Af,f][f,f]f)\odot f=\lambda(f\odot f),$$

where λ is scalar such that $[f,f][Af,f][f,f]f=\lambda f$ (the existence of such a scalar follows from the fact that $[f,f]$ is a rank-one matrix). This shows that the projection $f\odot f$ is Abelian. So, every nonzero central projection in $\mathcal{B}(\mathcal{H})$ contains a nonzero Abelian projection which means that $\mathcal{B}(\mathcal{H})$ is type I.

Suppose that the modular dimension of $\mathcal{H}$ is greater than 2. To see that $\mathcal{B}(\mathcal{H})$ is not isomorphic to $M_2(\mathbb{C})$ it is now enough to show that the linear dimension of $\mathcal{B}(\mathcal{H})$ is greater then 4. Let $\{f_1,f_2,f_3\}$ be a modular orthonormal set in $\mathcal{H}$. Denote $[f_i,f_i]=e_i$. If $d\geqslant 2$, then there are elements $a_i,b_i\in A$ such that $\{e_ia_i,e_ib_i\}$ is independent for every $i=1,2,3$. It is easy to check that $\{(a_if_i)\odot f_i,(b_if_i)\odot f_i : i=1,2,3\}$ is linearly independent. Therefore, the algebraic dimension of $\mathcal{B}(\mathcal{H})$ is at least 6. If $d=1$, then the statement is trivial. □

Let H be a Hilbert space. Recall that if $x,y\in H$, then $x\otimes y$ stands for the operator defined by $(x\otimes y)(z)=\langle z,y\rangle x(z\in H)$. The ideal of all finite rank operators in $B(H)$ is denoted by $F(H)$.

*Lemma 4: Let H be a Hilbert space. If $\phi:F(H)\to B(H)$ is a *-homomorphism which preserves the rank-one projections, then there is an isometry $U\in B(H)$ such that ϕ is of the form*

$$\phi(A)=UAU^*\quad (A\in F(H)).$$

*Similarly, if $\psi:F(H)\to B(H)$ is a *-antihomomorphism preserving the rank-one projections, then ψ is of the form*

$$\psi(A)=VA^{\mathrm{tr}}V^*\quad (A\in F(H)),$$

where V is an isometry and tr *denotes the transpose with respect to a fixed orthonormal basis in H.*

Proof: Let $y,z\in H$ be such that $\langle\phi(y\otimes y)z,z\rangle=1$. Define

$$Ux=\phi(x\otimes y)z\quad (x\in H).$$

It is easy to see that U is an isometry and $UA=\phi(A)U(A\in F(H))$. Let $x\in H$ be an arbitrary unit vector. Then $\phi(x\otimes x)$ is a rank-one projection, so it is of the form $\phi(x\otimes x)=x'\otimes x'$ with some unit vector $x'\in H$. Since

$$Ux\otimes x=\phi(x\otimes x)U=x'\otimes U^*x',$$

we obtain that x' is equal to Ux multiplied by a scalar of modulus 1. Therefore, $\phi(x\otimes x)=Ux\otimes Ux=U\cdot x\otimes x\cdot U^*$. Since this holds true for every unit vector $x\in H$, by linearity we have the first assertion of the lemma.

As for the second statement, we can apply a similar argument. Choosing $y,z\in H$ such that $\langle\psi(y\otimes y)z,z\rangle=1$, define

$$\tilde{V}x=\psi(y\otimes x)z \quad (x\in H).$$

One can verify that $\tilde{V}$ is an anti-isometry (that is, a conjugate-linear isometry), and then prove that $\psi(A)=\tilde{V}A^*\tilde{V}^*(A\in F(H))$. Considering an antiunitary operator J for which $JA^*J^*=A^{\mathrm{tr}}$ and defining $V=\tilde{V}J$, we conclude the proof. □

Lemma 5: Let (a_n) *be a sequence in the Hilbert space H and let* $b\in H$ *be such that* $\Sigma_n a_n\otimes a_n=b\otimes b$ *in the trace norm. Then for every n there exists a scalar* λ_n *such that* $a_n=\lambda_n b$.

Proof: Clearly, we may assume that $\|b\|=1$. Taking traces on both sides of the equality $\Sigma_n a_n\otimes a_n=b\otimes b$, we obtain $\Sigma_n\|a_n\|^2=1$. On the other hand, we also have

$$\sum_n |\langle b,a_n\rangle|^2=\left\langle\left(\sum_n a_n\otimes a_n\right)b,b\right\rangle=1.$$

By the Schwarz inequality,

$$1=\sum_n |\langle b,a_n\rangle|^2\leqslant\sum_n \|a_n\|^2=1.$$

So, there are equalities in the Schwarz inequalities $|\langle b,a_n\rangle|\leqslant\|a_n\|$. This implies the assertion. □

Proof of Theorem: We define an orthoadditive projection-valued measure μ on the lattice $\mathcal{P}(\mathcal{H})$ of all A-linear projections as follows. If $\{f_\alpha\}_\alpha$ is a modular orthonormal set, then let

$$\mu\left(\sum_\alpha f_\alpha\odot f_\alpha\right)=\sum_\alpha Tf_\alpha\odot Tf_\alpha.$$

Observe that by (1), $\{Tf_\alpha\}_\alpha$ is also modular orthonormal and, hence, by Lemma 1 $\Sigma_\alpha Tf_\alpha\odot Tf_\alpha$ belongs to $\mathcal{P}(\mathcal{H})$. We show that μ is well-defined. Let $\{f_\alpha\}_\alpha$ and $\{g_\beta\}_\beta$ generate the same closed submodule $\mathcal{M}$. We claim that the same holds true for $\{Tf_\alpha\}_\alpha$ and $\{Tg_\beta\}_\beta$. Indeed, if $g\in\mathcal{M}$, then due to the fact that $\{f_\alpha\}_\alpha$ is a modular basis in $\mathcal{M}$ we see that $g=\Sigma_\alpha[g,f_\alpha]f_\alpha$. This implies that

$$[Tg,Tg]=[g,g]=\sum_\alpha [g,f_\alpha][f_\alpha,g]=\sum_\alpha [Tg,Tf_\alpha][Tf_\alpha,Tg],$$

which, by Lemma 2, gives us that Tg belongs to the closed submodule generated by $\{Tf_\alpha\}_\alpha$. It is now obvious that μ is an orthoadditive $\mathcal{P}(\mathcal{H})$-valued measure on $\mathcal{P}(\mathcal{H})$.

Let us suppose that the modular dimension of $\mathcal{H}$ is greater than 2. By Proposition 3 we can apply a deep result of Bunce and Wright [Ref. 8, Theorem A]. It states that every bounded finitely orthoadditive, Banach space valued measure on the set of all projections in a von Neumann algebra without a summand isomorphic to $M_2(\mathbb{C})$ can be uniquely extended to a bounded linear transformation defined on the whole algebra. Let $\phi:\mathcal{B}(\mathcal{H})\to\mathcal{B}(\mathcal{H})$ denote the transformation corresponding to μ. Since it sends projections to projections, it is a standard argument to verify that ϕ is a Jordan *-endomorphism of $\mathcal{B}(\mathcal{H})$, that is, we have $\phi(T)^2=\phi(T^2)$, $\phi(T)^*=\phi(T^*)(T\in\mathcal{B}(\mathcal{H}))$ (see, for example, the proof of Ref. 9, Theorem 2).

We prove that $\phi(f\odot f)=Tf\odot Tf$ for every $f\in\mathcal{H}$. Let $[f,f]=\Sigma_i\lambda_i^2 e_i$, where λ_i's are non-negative real numbers and e_i's are pairwise orthogonal rank-one projections. Define $f_i=(1/\lambda_i)e_i f$. We have $[f_i,f_i]=e_i$ and $[f_i,f_j]=0$ if $i\neq j$, that is, $\{f_i\}_i$ is modular orthonormal. Then $f=\Sigma_i\lambda_i f_i=\Sigma_i e_i f$ since

$$\sum_i [f,f_i][f_i,f]=\sum_i \lambda_i^2 e_i=[f,f]$$

implies that $f=\Sigma_i[f,f_i]f_i=\Sigma_i e_i f$ (see Lemma 2). So, we have

$$\phi(f\odot f)=\sum_{i,j} \phi(e_i f\odot e_j f).$$

But $(e_i f)\odot(e_j f)=0$ if $i\neq j$. Indeed, we compute $[g,e_j f]e_i f=[g,f]e_j e_i f=0$ for every $g\in\mathcal{H}$. Hence,

$$\phi(f\odot f)=\sum_i \phi(e_i f\odot e_i f)=\sum_i \lambda_i^2\phi(f_i\odot f_i)=\sum_i \lambda_i^2\mu(f_i\odot f_i)=\sum_i \lambda_i^2 Tf_i\odot Tf_i.$$

So, the question is that whether the equality $Tf\odot Tf=\Sigma_i\lambda_i^2 Tf_i\odot Tf_i$ holds true. Clearly, $\{Tf_i\}$ is modular orthonormal. We compute

$$[Tf,Tf]=[f,f]=\sum_i [f,f_i][f_i,f]=\sum_i [Tf,Tf_i][Tf_i,Tf]$$

which, by Lemma 2, implies that $Tf=\Sigma_i[Tf,Tf_i]Tf_i$. We know that $|[Tf,Tf_i]|=|[f,f_i]|=\lambda_i e_i$. Similarly, $|[Tf_i,Tf]|=|[f_i,f]|=\lambda_i e_i$. Since e_i is a rank-one projection, we obtain that $[Tf,Tf_i]$ is also rank-one. Furthermore, as $|[Tf,Tf_i]|=|[Tf_i,Tf]|$ is a scalar multiple of e_i we can infer that $[Tf,Tf_i]=\mu_i\lambda_i e_i$, where μ_i is a scalar of modulus 1. Therefore, we have

$$Tf\odot Tf=\sum_{i,j} \mu_i\overline{\mu_j}(\lambda_i e_i Tf_i\odot\lambda_j e_j Tf_j).$$

But similarly as above, for $i\neq j$ we have

$$(e_i Tf_i\odot e_j Tf_j)g=[g,e_j Tf_j]e_i Tf_i=[g,Tf_j]e_j e_i Tf_i=0.$$

Therefore

$$\begin{aligned}Tf\odot Tf&=\sum_{i,j} \mu_i\overline{\mu_j}(\lambda_i e_i Tf_i\odot\lambda_j e_j Tf_j)\\&=\sum_i \mu_i\overline{\mu_i}(\lambda_i e_i Tf_i\odot\lambda_i e_i Tf_i)=\sum_i \lambda_i e_i Tf_i\odot\lambda_i e_i Tf_i=\sum_i \lambda_i^2(e_i Tf_i\odot e_i Tf_i).\end{aligned}$$

But $(e_i Tf_i\odot e_i Tf_i)=Tf_i\odot Tf_i$. Indeed, since Tf_i is a modular unit vector, we have $e_i Tf_i=[f_i,f_i]Tf_i=[Tf_i,Tf_i]Tf_i=Tf_i$ (see Ref. 7, Lemma 1). Consequently, we obtain $Tf\odot Tf=\Sigma_i\lambda_i^2 Tf_i\odot Tf_i$ and this was to be proved. So, we get $\phi(f\odot f)=Tf\odot Tf$ for every $f\in\mathcal{H}$.

We assert that ϕ is either a *-homomorphism or a *-antihomomorphism. By Lemma 1 the minimal projections in $\mathcal{H}$ are exactly the operators of the form $f\odot f$, where $f\in\mathcal{H}$ is a modular unit vector. Clearly, ϕ sends minimal projections to minimal projections. By Ref. 1, Lemma 2, the linear space generated by the minimal projections in $\mathcal{B}(\mathcal{H})$ is $\mathcal{F}(\mathcal{H})$. Since $\mathcal{B}(\mathcal{H})$ is a type I factor, it is isomorphic to the full operator algebra $B(H)$ on a Hilbert space H. Since *-isomorphisms preserve the minimal projections, $\mathcal{F}(\mathcal{H})$ corresponds to the ideal $F(H)$ of all finite rank operators in $B(H)$. Under this identification, we obtain a Jordan *-homomorphism $\tilde{\phi}$ on $F(H)$ corresponding to $\phi_{|\mathcal{F}(\mathcal{H})}$ which sends rank-one projections to rank-one projections. Since $F(H)$ is a local matrix algebra, by Ref. 10, Theorem 8, we obtain that $\tilde{\phi}$ is the sum of a *-homomorphism and a *-antihomomorphism. As $\tilde{\phi}$ preserves the rank-one projections, from the simplicity of the ring $F(H)$ it follows that $\tilde{\phi}$ is either a *-homomorphism or a *-antihomomorphism. Obviously, the same holds for $\phi_{|\mathcal{F}(\mathcal{H})}$.

Let us suppose that the modular dimension of $\mathcal{H}$ is greater than d. By Ref. 1, Remark 2, there are vectors $g,h\in\mathcal{H}$ such that $[g,h]=I$. The map $\phi_{|\mathcal{F}(\mathcal{H})}$ is either a *-homomorphism or a *-antihomomorphism. First consider this latter case. Referring to Lemma 4 we have an operator $U\in\mathcal{B}(\mathcal{H})$ with $U^*U=I$ and a *-antiautomorphism ψ of $\mathcal{F}(\mathcal{H})$ such that $\phi(A)=U\psi(A)U^*(A\in\mathcal{F}(\mathcal{H}))$.

We define

$$Vf=\psi(g\odot f)U^*Th \quad (f\in\mathcal{H}),$$

where $g,h\in\mathcal{H}$ are fixed and such that $[g,h]=I$. Clearly, V is a conjugate-linear operator. We have

$$VAf=\psi(g\odot(Af))U^*Th=\psi(g\odot fA^*)U^*Th=\psi(A)^*\psi(g\odot f)U^*Th=\psi(A)^*Vf,$$

that is, $VA=\psi(A)^*V(A\in\mathcal{F}(\mathcal{H}))$. We compute

$$\begin{aligned}[Vf,Vf]&=[\psi(g\odot f)U^*Th,\psi(g\odot f)U^*Th]\\&=[\psi(g\odot f\cdot f\odot g)U^*Th,U^*Th]\\&=[U\psi(g\odot f\cdot f\odot g)U^*Th,Th]\\&=[\phi(g\odot f\cdot f\odot g)Th,Th]\\&=[\phi(\sqrt{[f,f]}g\odot\sqrt{[f,f]}g)Th,Th]\\&=[(T(\sqrt{[f,f]}g)\odot T(\sqrt{[f,f]}g))Th,Th]\\&=[Th,T(\sqrt{[f,f]}g)][T(\sqrt{[f,f]}g),Th]\\&=[h,\sqrt{[f,f]}g][\sqrt{[f,f]}g,h]=[h,g][f,f][g,h]=[f,f].\end{aligned}$$

Since V is conjugate-linear, by polarization we obtain

$$[Vf,Vf']=[f',f] \quad (f,f'\in\mathcal{H}).$$

We show that $\operatorname{rng}T\subset\operatorname{rng}U$ which will imply $UU^*T=T$ (UU^* is the projection onto the range of U). Let $f\in\mathcal{H}$. In the previous part of the proof we have learned that $Tf\odot Tf$ is a linear combination of operators of the form $Tf_b\odot Tf_b$, where f_b's are modular unit vectors. We have

$$Tf_b\odot Tf_b=\phi(f_b\odot f_b)=U\psi(f_b\odot f_b)U^*$$

and, ψ being a *-antiautomorphism, $\psi(f_b\odot f_b)$ is a minimal projection. Therefore, $\psi(f_b\odot f_b)=f'_b\odot f'_b$ with some modular unit vector f'_b and hence $Tf_b\odot Tf_b=Uf'_b\odot Uf'_b$. Now let $Tf=g'+g''$, where $g'\in\operatorname{rng}U$ and $g''\in\operatorname{rng}U^\perp$. We have

$$[g'',g'']^2=[g'',Tf][Tf,g'']=[(Tf\odot Tf)g'',g'']=0.$$

This gives us that $g''=0$ which shows that $Tf\in \operatorname{rng} U$.

We next prove that V is surjective. Let $f\in\mathcal{H}$ be arbitrary. Since ψ is a *-antiautomorphism of $\mathcal{F}(\mathcal{H})$, we can find an operator $R\in\mathcal{F}(\mathcal{H})$ such that $\psi(R)^*=f\odot U^*Th$. We compute

$$\begin{aligned}VRg&=\psi(R)^*Vg\\&=\psi(R)^*\psi(g\odot g)U^*Th\\&=\psi(R)^*U^*\phi(g\odot g)Th\\&=\psi(R)^*U^*(Tg\odot Tg)Th\\&=[Th,Tg]\psi(R)^*U^*Tg\\&=[Th,Tg][U^*Tg,U^*Th]f\\&=[Th,Tg][UU^*Tg,Th]f\\&=[Th,Tg][Tg,Th]f=[h,g][g,h]f=f.\end{aligned}$$

Since f was arbitrary, we have the surjectivity of V.

We compute

$$\begin{aligned}[][UVf',Tf][Tf,UVf']&=[(Tf\odot Tf)UVf',UVf']\\&=[U^*(Tf\odot Tf)UVf',Vf']\\&=[U^*\phi(f\odot f)UVf',Vf']\\&=[\psi(f\odot f)Vf',Vf']\\&=[(V\cdot f\odot f)f',Vf']=[f',(f\odot f)f']=[f',f][f,f'].\end{aligned}$$

This gives us that

$$\begin{aligned}[][V^{-1}U^*Tf',f][f,V^{-1}U^*Tf']&=[UVV^{-1}U^*Tf',Tf][Tf,UVV^{-1}U^*Tf']\\&=[UU^*Tf',Tf][Tf,UU^*Tf']\\&=[Tf',Tf][Tf,Tf']=[f',f][f,f'].\end{aligned}$$

Replacing f by $xf(x\in A)$, we obtain

$$[V^{-1}U^*Tf',f]x^*x[f,V^{-1}U^*Tf']=[f',f]x^*x[f,f'].$$

Since every element of A is a linear combination of elements of the form x^*x, it follows that

$$[V^{-1}U^*Tf',f]y[f,V^{-1}U^*Tf']=[f',f]y[f,f']$$

holds for every $y\in A$. This implies that for every $f\in\mathcal{H}$, the matrices $[f,V^{-1}U^*Tf']$ and $[f,f']$ are linearly dependent. It requires only elementary linear algebra to verify the following assertion. If X,Y are vector spaces and $A,B:X\to Y$ are linear operators such that for every $x\in X$, the set $\{Ax,Bx\}$ is linearly dependent, then either A and B have rank at most one or $\{A,B\}$ is linearly dependent. Since the rank of the linear operator $f\mapsto[f,f']$ is clearly greater than 1 if $f'\neq 0$, we have a scalar $\lambda_{f'}$ (depending only on f') such that $[f,V^{-1}U^*Tf']=\lambda_{f'}[f,f'](f,f'\in\mathcal{H})$. This gives us that there is a function $\varphi:\mathcal{H}\to\mathbb{C}$ such that $V^{-1}U^*Tf'=\varphi(f')f'$ which results in $Tf'=\varphi(f')UVf'$. It follows from the properties of T,U,V that φ is of modulus 1. Finally, we have

$$|[f,f']|=|[Tf,Tf']|=|[UVf,UVf']|=|[Vf,Vf']|=|[f',f]|.$$

Since this must hold true for every $f,f' \in \mathcal{H}$, it follows that for every rank-one matrix $a \in A$ we have $|a| = |a^*|$. But this is an obvious contradiction. Since we have started with assuming that $\phi_{|\mathcal{F}(\mathcal{H})}$ is a *-antihomomorphism, we thus obtain that it is in fact a *-homomorphism.

Pushing the problem from $\mathcal{B}(\mathcal{H})$ to the full operator algebra $B(H)(\cong \mathcal{B}(\mathcal{H}))$, we see that there is an A-isometry $U \in \mathcal{B}(\mathcal{H})$ such that $\phi(A) = UAU^*(A \in \mathcal{F}(\mathcal{H}))$. This gives us that $Tf \odot Tf = Uf \odot Uf$ for every $f \in \mathcal{H}$. Similarly as before, this implies that $\operatorname{rng} T \subset \operatorname{rng} U$ which yields $UU^*Tf = Tf(f \in \mathcal{H})$. We next compute

$$[Uf',Tf][Tf,Uf'] = [(Tf \odot Tf)Uf',Uf'] \\ = [(Uf \odot Uf)Uf',Uf'] = [Uf',Uf][Uf,Uf'] = [f',f][f,f'],$$

which gives us that

$$[U^*Tf',f][f,U^*Tf'] = [UU^*Tf',Tf][Tf,UU^*Tf'] = [Tf',Tf][Tf,Tf'] = [f',f][f,f'].$$

Just as above, it follows that U^*Tf' is a scalar multiple of f'. Therefore, there exists an A-isometry U and a phase-function $\varphi:\mathcal{H} \to \mathbb{C}$ such that

$$Tf = \varphi(f)Uf \quad (f \in \mathcal{H}).$$

This completes the proof in the case when the modular dimension n of $\mathcal{H}$ is greater than d.

We now treat the low dimensional cases, that is, when $n \leqslant d$. Let H_d denote the d-dimensional complex Euclidean space. Then H_d can be considered as a Hilbert A-module. Here, the module operation is $(a,\xi) \mapsto a(\xi)$ and the generalized inner product is defined by $[\xi,\zeta] = \xi \otimes \zeta$. Clearly, the modular dimension of this module is 1. It now follows from the structure of our Hilbert A-modules (see, for example, Ref. 11) that $\mathcal{H}$ is isomorphic to the n-fold direct sum of H_d with itself. So, we may assume that $\mathcal{H} = \Sigma_{i=1}^n \oplus H_d$. The definition of the module operation and that of the inner product on this direct sum is defined as follows:

$$a[\xi_i]_i = [a\xi_i]_i, \quad [[\xi_i]_i,[\zeta_i]_i] = \sum_i \xi_i \otimes \zeta_i .$$

Let us describe the elements of $\mathcal{B}(\mathcal{H})$. Since every element of $\mathcal{B}(\mathcal{H})$ is a linear operator on the direct sum of vector spaces, it can represented by a matrix

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix},$$

where a_{ij}'s are linear operators acting on H_d. Now, A-linearity means that

$$\begin{bmatrix} a_{11}a\xi_1 + \cdots + a_{1n}a\xi_n \\ \vdots \\ a_{n1}a\xi_1 + \cdots + a_{nn}a\xi_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} a\xi_1 \\ \vdots \\ a\xi_n \end{bmatrix} = \begin{bmatrix} a(a_{11}\xi_1 + \cdots + a_{1n}\xi_n) \\ \vdots \\ a(a_{n1}\xi_1 + \cdots + a_{nn}\xi_n) \end{bmatrix}$$

holds for every $a \in A$ and $\xi_i \in H_d$. It is easy to see that this is equivalent to $a_{ij}a = aa_{ij}(a \in A)$ which means that a_{ij}'s are scalars. Consequently, $\mathcal{B}(\mathcal{H})$ is isomorphic to $M_n(\mathbb{C})$.

Suppose that $n > 1$. If ζ is any vector in H_d, then let ζ^k denote the element of $\mathcal{H}$ whose coordinates are all 0 except for the kth one which is ζ. Fix a unit vector $\xi \in H_d$. We have

$$\sum_i (T\xi^k)_i \otimes (T\xi^k)_i = [T\xi^k, T\xi^k] = [\xi^k, \xi^k] = \xi \otimes \xi.$$

From Lemma 5 we infer that for every $i=1,...,n$, there is a scalar α_{ik} such that $(T\xi^k)_i=\alpha_{ik}\xi$. Clearly, the columns of the matrix (α_{ik}) are unit vectors. Since $[T\xi^k,T\xi^l]=0$ for $k\neq l$, it follows that the columns of our matrix are pairwise orthogonal as well. So (α_{ik}) is a unitary matrix and hence it defines an A-unitary operator U on $\mathcal{H}$. Considering U^*T instead of T, we can assume that $T\xi^k$ is equal to ξ^k for every $k=1,...,n$. If f is any vector in $\mathcal{H}$, then considering the equality

$$|\xi\otimes(Tf)_k|=|[T\xi^k,Tf]|=|[\xi^k,f]|=|\xi\otimes f_k|,$$

we obtain

$$(Tf)_k=\mu_k f_k \quad (k=1,...,n) \tag{2}$$

with some scalars μ_k of modulus 1. We claim that all the μ_k's are equal. Fix a $g\in\mathcal{H}$ whose coordinates are pairwise orthogonal unit vectors in H_d (recall that $n\leqslant d$). It is apparent that if we multiply T from the left by an A-unitary operator whose matrix is diagonal, then the so obtained transformation still has the property (2). So we may assume that $Tg=g$. Let $f\in\mathcal{H}$ be arbitrary. We have

$$\left|\sum_i \mu_i f_i\otimes g_i\right|=|[Tf,Tg]|=|[f,g]|=\left|\sum_i f_i\otimes g_i\right|.$$

This implies that

$$\sum_{i,j}\langle\mu_j f_j,\mu_i f_i\rangle g_i\otimes g_j=\sum_{i,j}\langle f_j,f_i\rangle g_i\otimes g_j$$

which gives that

$$\langle\mu_j f_j,\mu_i f_i\rangle=\langle f_j,f_i\rangle.$$

So, if $\langle f_i,f_j\rangle\neq 0$, then we have $\mu_i=\mu_j$. Suppose now that $\langle f_i,f_j\rangle=0$ but $f_i,f_j\neq 0$. Let $\zeta\in H_d$ be any nonzero vector and consider $\zeta^i+\zeta^j$. By what we have just proved, it follows that $T(\zeta^i+\zeta^j)$ is a scalar multiple of $\zeta^i+\zeta^j$. We compute

$$|\zeta\otimes(\mu_i f_i+\mu_j f_j)|=|[\zeta^i+\zeta^j,Tf]|=|[T(\zeta^i+\zeta^j),Tf]|=|[\zeta^i+\zeta^j,f]|=|\zeta\otimes(f_i+f_j)|$$

which clearly gives us that $\mu_i=\mu_j$. Therefore, we obtain that for any vector $f\in\mathcal{H}$, Tf is equal to f multiplied by a complex number of modulus 1. The assertion of the theorem now follows for the case $1<n\leqslant d$.

Finally, suppose that $n=1$, which means that $\mathcal{H}=H_d$. Our problem is to describe those maps $T:H_d\rightarrow H_d$ for which $|T\xi\otimes T\zeta|=|\xi\otimes\zeta|(\xi,\zeta\in H_d)$. But this equality clearly implies that $T\zeta$ is equal to ζ multiplied by a scalar of modulus 1.

The proof of the theorem is now complete. □

ACKNOWLEDGMENTS

This research was supported from the following sources: (1) Hungarian National Foundation for Scientific Research (OTKA), Grant No. T-030082 F-019322; (2) a grant from the Ministry of Education, Hungary, Reg. No. FKFP 0304/1997.

[1] L. Molnár, ''An algebraic approach to Wigner's unitary–antiunitary theorem,'' J. Aust. Math. Soc. A, Pure Math. Stat. **65**, 354–369 (1998).
[2] P. P. Saworotnow, ''A generalized Hilbert space,'' Duke Math. J. **35**, 191–197 (1968).
[3] V. Bargmann, ''Note on Wigner's theorem on symmetry operations,'' J. Math. Phys. **5**, 862–868 (1964).
[4] C. S. Sharma and D. F. Almeida, ''A direct proof of Wigner's theorem on maps which preserve transition probabilities between pure states of quantum systems,'' Ann. Phys. **197**, 300–309 (1990).

[5] J. Rätz, ''On Wigner's theorem: Remarks, complements, comments, and corollaries,'' Aequationes Math. **52**, 1–9 (1996).
[6] C. S. Sharma and D. F. Almeida, ''Additive isometries on a quaternionic Hilbert space,'' J. Math. Phys. **31**, 1035–1041 (1990).
[7] L. Molnár, ''Modular bases in a Hilbert *A*-module,'' Czech. Math. J. **42**, 649–656 (1992).
[8] L. J. Bunce and D. M. Wright, ''The Mackey–Gleason problem,'' Bull. Am. Math. Soc. **26**, 288–293 (1992).
[9] L. Molnár, ''The set of automorphisms of $B(H)$ is topologically reflexive in $B(B(H))$,'' Stud. Math. **122**, 183–193 (1997).
[10] N. Jacobson and C. Rickart, ''Jordan homomorphisms of rings,'' Trans. Am. Math. Soc. **69**, 479–502 (1950).
[11] M. Ozawa, ''Hilbert $B(H)$-modules and stationary processes,'' Kodai Math. J. **3**, 26–39 (1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Two new potentials for the free particle model

J. Morales[a)] and J. J. Peña
Universidad Autónoma Metropolitana, Azcapotzalco, CBI-Area de Física, Av. San Pablo 180, 02200 México, D.F., Mexico

(Received 18 January 1999; accepted for publication 11 August 1999)

In this work, is proposed a very simple method for obtaining the generalized potential associated with a known standard potential. The procedure is straightforward because it only uses two Ricatti-type relationships as enough condition to find a generalized potential; one particular equation is needed to identify the specific potential under study and one general Ricatti relationship is used to find the corresponding generalized potential. Moreover, the method is completely general due to the fact that an arbitrary potential has been considered in its development for which the procedure can also be used to find new potentials which could be needed in the modeling of important quantum interactions. The usefulness of the proposed approach, is shown with the treatment of the three- and one-dimensional potential for the free particle model. This work example leads to two new potentials whose Hamiltonians are isospectral when they are compared with the former Hamiltonian.
 [S0022-2488(99)03311-3]

I. INTRODUCTION

At the origin of the quantum theory, the proposition of interaction models to explain different experimental data has played a role of extraordinary importance in the theoretical treatment of atoms and molecules. In this regard, in search of solutions for the second order differential equation involved, the factorization of the Schrödinger relationship is a powerful operational method which has its roots in the Dirac's second quantization treatment of the one-dimensional harmonic oscillator potential.[1] Today, for obtaining the algebraic representation of any other standard potential, the following procedures are available: Infeld and Hull's (IH) factorization method,[2] quantizing classical dynamical variables,[3] using the algebraic representation of the orthogonal polynomials directly involved[4] and by an alternative approach[5] to the IH procedure. However, the above procedures do not allow extending the scope of the so-called isospectral potentials,[6,7] generalized or modified, whose treatment has been considered only for some particular cases.[8–10] In the present work we consider a general formalism for obtaining any modified and generalized potential associated with a known or unknown specific potential. The proposed approach is based on a procedure that instead of factoring the Schrödinger equation uses the equivalent of supersymmetry techniques applied to quantum mechanics.[11] As it can be seen later, as a result of the factorization of the Hamiltonian, which incorporates an arbitrary potential, we propose a general method to find isospectral potentials by means of the use of two Ricatti-type relationships: a specific one to identify the potential under study and a general one to obtain the isospectral potentials. Finally, in spite that the proposed approach is totally general, due to the fact that it considers know and unknown potentials, in Sec. III we show the usefulness of the proposed algorithm by obtaining the generalized potential associated with the free particle model.

II. GENERAL PROCEDURE FOR ANY $V(r)$ POTENTIAL

As mentioned in the previous section, the principal idea behind this work is concerned with the factorization of a general Hamiltonian H, with an arbitrary potential $V(r)$, instead of the factorization of the corresponding Schrödinger relationship. That is, when one consider the first order operational equations

[a)]Electronic mail: jmr@hp9000a1.uam.mx

$$a^{\pm}=\frac{\hbar}{\sqrt{2m}}\left(\beta(r)-\frac{l}{r}\mp\frac{d}{dr}\right), \tag{2.1}$$

where $\beta(r)$ is any function, it becomes evident that

$$a^{\mp}a^{\pm}=H^{\pm}-V^{\pm}(r)+\frac{\hbar^2}{2m}\left(\beta^2(r)\pm\beta'(r)-\frac{2l\beta(r)}{r}-\frac{(l\mp l)}{r^2}\right), \tag{2.2}$$

where

$$H^{\pm}=-\frac{\hbar^2}{2m}\left(\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}\right)+V^{\pm}(r). \tag{2.3}$$

That is, there are two Hamiltonians (named $H^{\pm}$) related with two potentials (named $V^{\pm}(r)$), which can be exactly factorized by means of

$$a^{\mp}a^{\pm}=H^{\pm}+C^{\pm} \tag{2.4}$$

with the condition to fulfill the relationship

$$C^{\pm}=-V^{\pm}(r)+\frac{\hbar^2}{2m}\left(\beta^2(r)\pm\beta'(r)-\frac{2l\beta(r)}{r}-\frac{(l\mp l)}{r^2}\right), \tag{2.5}$$

where $C^{\pm}$ is a constant parameter. Without loss of generality, one can always choose $C^{\pm}=C$ in order to obtain

$$V^{\pm}(r)-V^{\mp}(r)=\pm\frac{\hbar^2}{m}\left(\beta'(r)+\frac{l}{r^2}\right) \tag{2.6}$$

which is equivalent to

$$V^{+}(r)-V^{-}(r)=\frac{\hbar^2}{m}\left(\beta'(r)+\frac{l}{r^2}\right), \tag{2.7}$$

where, from Eq. (2.5),

$$V^{+}(r)=\frac{\hbar^2}{2m}\left(\beta^2(r)+\beta'(r)-\frac{2l\beta(r)}{r}\right)-C. \tag{2.8}$$

Consequently, this last equation can be arranged in order to have the form of the Ricatti relationship,

$$\beta'(r)=-\beta^2(r)+\frac{2l}{r}\beta(r)+\frac{2m}{\hbar^2}(V^{+}(r)+C). \tag{2.9}$$

In fact, the identification of $\beta(r)=y$, $Q(r)=-1$, $P(r)=2l/r$, and $R(r)=(2m/\hbar^2)(V^{+}(r)+C)$, allows us to obtain

$$y'=Q(r)y^2+P(r)y+R(r), \tag{2.10}$$

which has as general solution

$$y=y_p+\frac{b}{\mu}, \tag{2.11}$$

where y_p is a particular solution,

$$\mu=e^{-\int^r[2Q(r)y_p+P(r)]dr}\left(\gamma-b\int^r e^{\int^r[2Q(r)y_p+P(r)]dr}Q(r)dr\right), \tag{2.12}$$

and where γ and b are constant. In other words, if we consider a specific $\beta_p(r)$ as a particular solution of the specific Ricatti relationship,

$$\beta_p'(r)=-\beta_p^2(r)+\frac{2l}{r}\beta_p(r)+\frac{2m}{\hbar^2}(V_p^+(r)+C), \tag{2.13}$$

two events take place; first, the choice of an $\beta_p(r)$ ansatz lets us identify the $V_p^+(r)$ particular potential under study and, second, it is always possible to construct $\beta_g(r)$, the general solution of Eq. (2.9), by means of

$$\beta(r)=\beta_g(r)=\beta_p(r)+\frac{b}{\rho(r)}, \tag{2.14}$$

where the subindex g is used to denote the general solution and

$$\rho(r)=\frac{1}{r^{2l}}e^{2\int^r\beta_p(r)dr}\left(\gamma+b\int^r r^{2l}e^{-2\int^r\beta_p(r)dr}dr\right). \tag{2.15}$$

Thus, according to Eq. (2.7), the identification of the corresponding $V_p^-(r)$ particular potential under study occurs by means of

$$V_p^-(r)=V_p^+(r)-\frac{\hbar^2}{m}\left(\beta_p'(r)+\frac{l}{r^2}\right) \tag{2.16}$$

in such a way that using Eq. (2.8) and Eqs. (2.13)–(2.15) one finds that the generalized potential $V_g^+(x)$ is given by

$$V_g^+(r)=V_p^+(r)+\frac{\hbar^2 b}{m}\left(\frac{b}{\rho^2(r)}-\frac{\rho'(r)}{\rho^2(r)}+\frac{2\beta_p(r)}{\rho(r)}-\frac{2l}{r\rho(r)}\right). \tag{2.17}$$

At this point, it should be noticed that

$$\rho'(r)=2\beta_p(r)\rho(r)+b-\frac{2l\rho(r)}{r}, \tag{2.18}$$

thereby reducing Eq. (2.17) to

$$V_g^+(r)=V_p^+(r) \tag{2.19}$$

as well as

$$H_g^+=H_p^+=-\frac{\hbar^2}{2m}\left(\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}\right)+V_p^+(r) \tag{2.20}$$

which means that there does not exist a special $V_g^+(r)$ potential. This important result indicates that the generalized operators

$$a_g^{\pm}=A^{\pm}=\frac{\hbar}{\sqrt{2m}}\left(\beta_p(r)+\frac{b}{\rho(r)}-\frac{l}{r}\mp\frac{d}{dr}\right) \tag{2.21}$$

factorize, both, the particular Hamiltonian H_p^+ or H_g^+ according to

$$A^-A^+=H_{p,g}^+ +C. \tag{2.22}$$

Similarly, the generalized Hamiltonian H_g^- is factorized by means of

$$A^+A^-=H_g^- +C, \tag{2.23}$$

where H_g^- is associated with the generalized potential $V_g^-(r)$ which is given, after using Eq. (2.7), Eq. (2.14), and Eq. (2.16), by

$$V_g^-(r)=V_p^-(r)-\frac{\hbar^2}{m}\frac{d}{dr}\left(\frac{b}{\rho(r)}\right). \tag{2.24}$$

Finally, the existence of this new potential, which is generalized, is particularly important due to the fact that Eq. (2.22) and Eq. (2.23) lead to

$$H_g^-A^+=A^+H_{p,g}^+ \tag{2.25}$$

or

$$H_g^-A^+\psi=E_nA^+\psi, \tag{2.26}$$

where E_n and ψ are, respectively, the energy and eigenfunctions of $H_{p,g}^+$. Therefore, from the above equation it becomes evident that the Hamiltonian H_g^- has eigenfunctions $\phi=A^+\psi$ with the same energy spectra of $H_{p,g}^+$ and consequently the corresponding $V_g^-(r)$ generalized potential has the energy spectra of the particular potential $V_p^+(r)$. This result clearly justifies why the new potential $V_g^-(r)$ is called isospectral, which is in good agreement with already published results on certain particular Hamiltonians.[6–8,12–14] Evidently, the same occurs with the modified potential $V_p^-(r)$ given in Eq. (2.16) because the Hamiltonians H_p^+ or H_g^+ are related to H_p^- through the operators $a^{\pm}$ by means of

$$H_{p,g}^+a^-=a^-H_p^- \quad \text{and} \quad H_p^-a^+=a^+H_{p,g}^+, \tag{2.27}$$

which lead to

$$H_p^-(a^+\psi)=E_p^+(a^+\psi), \tag{2.28}$$

where $E_p^+=E_n$, proving that the modified Hamiltonian H_p^- is also isospectral.

III. TWO NEW POTENTIALS FOR THE FREE PARTICLE MODEL

The previously described procedure for obtaining generalized potentials, is straightforward and reduces to the use of three relationships; the master Eq. (2.8) with a specific solution $\beta_p(r)$, which can be taken *ad hoc* to match with the $V_p^+(r)$ particular potential under study, Eq. (2.16) used to find the $V_p^-(r)$ modified partner potential and finally the use of Eq. (2.24) to get the corresponding $V_g^-(r)$ generalized potential. In this section, we will consider the case of the well known free particle model in order to generalize it as well as to find the corresponding associated potential and its reduction to the particular one-dimensional situation.

A. A new three-dimensional potential for the free particle model $(0<r<\infty)$

According to Eq. (2.8), any particular potential is given by

$$V_p^+(r)=\frac{\hbar^2}{2m}\left(\beta_p^2(r)+\beta_p'(r)-\frac{2l\beta_p(r)}{r}\right)-C_p. \tag{3.1}$$

Thus, with the purpose of identifying the three-dimensional potential for the free-particle interaction model we use in the previous equation the ansatz $\beta_p(r)=0$ which gives rise to

$$V_p^+(r)=C_p=0. \tag{3.2}$$

Consequently, according to Eq. (2.16), the corresponding modified potential is then

$$V_p^-(r)=-\frac{\hbar^2 l}{mr^2}. \tag{3.3}$$

That is, the Hamiltonians $H_p^\pm$ associated with these potentials are

$$H_p^\pm=-\frac{\hbar^2}{2m}\left(\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}\right)-\frac{\hbar^2(l\mp l)}{2mr^2}, \tag{3.4}$$

and they are factorized by means of $a^\mp a^\pm=H_p^\pm$, where $a^\pm$ are the first order differential operators

$$a^\pm=\frac{\hbar}{\sqrt{2m}}\left(-\frac{l}{r}\mp\frac{d}{dr}\right). \tag{3.5}$$

On the other hand, from Eq. (2.15) and by considering the fact that $\beta_p(r)=0$ one gets

$$\rho(r)=\frac{\gamma}{r^{2l}}+\frac{br}{2l+1}. \tag{3.6}$$

Finally, according to Eq. (2.24), the generalized potential that corresponds to this case is given by

$$V_g^-(r)=-\frac{\hbar^2 l}{mr^2}-\frac{\hbar^2 b(2l(\gamma+R(r))r^{2l-1}-br^{4l})}{m(\gamma+R(r))^2}, \tag{3.7}$$

where $R(r)=(br^{2l+1}/2l+1)$. Consequently, the Hamiltonian H_g^- associated with this generalized potential becomes

$$H_g^-=-\frac{\hbar^2}{2m}\left(\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}\right)-\frac{\hbar^2 l}{mr^2}-\frac{\hbar^2 b(2l(\gamma+R(r))r^{2l-1}-br^{4l})}{m(\gamma+R(r))^2} \tag{3.8}$$

in such a way that it is factorized by means of $A^+A^-=H_g^-$, where $A^\pm$ are given by

$$A^\pm=\frac{\hbar}{\sqrt{2m}}\left(\frac{br^{2l}}{\gamma+R(r)}-\frac{l}{r}\mp\frac{d}{dr}\right). \tag{3.9}$$

On the other hand, due to the fact that wave functions for this case of the free particle model are given, according to Buckingham,[15] by $\psi_p(l,r)=krj_l(kr)$, where $j_l(kr)$ are the spherical Bessel functions, it becomes that normalized eigenfunctions of the corresponding generalized H_g^- Hamiltonian are given, from Eq. (2.26), by

$$\phi_g(l,z)=\phi_g(z)=A^+\psi_p(l,z)=\frac{\hbar k}{\sqrt{2m}}\left(\frac{bz^{2l+1}j_l(z)}{\gamma_k+R(z)}-(l+1)j_l(z)-zj_l'(z)\right), \tag{3.10}$$

where $z=kr$, $\gamma_k=k^{2l+1}\gamma$, and $\psi_p(l,z)=zj_l(z)$. Thus, the normalized Hamiltonian $H_g^-(z)$ of Eq. (3.8) becomes

$$H_g^-(z)=-\frac{\hbar^2k^2}{2m}\left(\frac{d^2}{dz^2}-\frac{l(l-1)}{z^2}+\frac{4blz^{2l-1}}{\gamma_k+R(z)}-\frac{2b^2z^{4l}}{(\gamma_k+R(z))^2}\right) \tag{3.11}$$

for which it follows that

$$H_g^-(z)\phi_g(z)=-\frac{\hbar^3k^3}{(2m)^{3/2}}\left(\frac{bz^{2l+1}j_l''(z)}{\gamma_k+R(z)}+\frac{2bz^{2l}j_l'(z)}{\gamma_k+R(z)}-\frac{l(l+1)bz^{2l-1}j_l(z)}{\gamma_k+R(z)}\right.$$
$$\left.-(l+3)j_l''(z)-zj_l'''(z)+\frac{l(l+1)(l-1)j_l(z)}{z^2}+\frac{l(l-1)j_l'(z)}{z}\right). \tag{3.12}$$

Finally, with the aim of simplifying the third and second order derivatives of the spherical Bessel functions that appeared in the above relationship, we use repeatedly the differential equation[16]

$$j_l''(z)+\frac{2j_l'(z)}{z}+\left(1-\frac{l(l+1)}{z^2}\right)j_l(z)=0 \tag{3.13}$$

in order to get

$$H_g^-(z)\phi_g(z)=-\frac{\hbar^2k^2}{2m}\left(\frac{\hbar k}{\sqrt{2m}}\right)\left\{zj_l'(z)-\frac{bz^{2l+1}j_l(z)}{\gamma_k+R(z)}+(l+1)j_l(z)\right\}. \tag{3.14}$$

That is, by identifying the terms inside the two parentheses with the wave function $\phi_g(z)$ of Eq. (3.10) it becomes that

$$H_g^-(z)\phi_g(z)=E_g\phi_g(z), \tag{3.15}$$

where $E_g=E_p=(\hbar^2k^2/2m)$, which proves that the generalized Hamiltonian $H_g^-(z)$ is isospectral with respect to the particular Hamiltonians $H_p^{\pm}(z)$ as expected for the free particle interaction model.

Due to fact that we are involved with a scattering potential, it is important to find the phase shift existing between the generalized potential and the specific one of the free particle model. In this case, according to Eq. (3.10) the asymptotic wave functions of the generalized potential are given by

$$\phi_g(z)\rightarrow\sin\left(z-\frac{\pi(l+1)}{2}\right), \tag{3.16}$$

$$z\rightarrow\infty$$

which indicates that these wave functions are displaced in one unit in the angular momentum l with respect to the corresponding free particle wave functions given by[17]

$$\phi_g(z)\rightarrow \sin\left(z-\frac{\pi l}{2}\right), \tag{3.17}$$

$$z\rightarrow\infty.$$

Finally, concerning the specific case of the free particle one-dimensional potential, it is important to notice that $l=0$ in Eq. (3.3) leads to the modified potential $V_p^-=0$ in such a way that the use of Eq. (3.7) gives origin to the corresponding generalized potential

$$V_g^-(x)=\frac{\hbar^2}{m}\left(\frac{b}{\gamma+bx}\right)^2 \tag{3.18}$$

in good agreement with already published results.[14] At this point, it is important to notice that there is the possibility of finding another, different from the above, one-dimensional potential for the same free particle model if, instead of using the mentioned reductions in the general solutions, one can consider an alternative ansatz in the corresponding Ricatti relationship for $l=0$ as shown in the following section.

B. A new one-dimensional potential for the free particle model

As mentioned in the last paragraph, when considering the Ricatti relationship given in Eq. (2.9) for the one-dimensional potential situation, $r=x, -\infty<x<\infty$ and $l=0$, one have

$$\beta'(x)=-\beta^2(x)+\frac{2m}{\hbar^2}(V^+(x)+C) \tag{3.19}$$

which it has as its general solution, Eq. (2.11). It should be pointed out, that in this new situation the μ that appears in Eq. (2.12) is given by

$$\mu=e^{-\int^x 2Q(x)y_p dx}\left(\gamma-b\int^x e^{\int^x 2Q(x)y_p dx}Q(x)dx\right), \tag{3.20}$$

where γ and b are constant. Similarly to the three-dimensional case, we can consider $\beta_p(x)$ as a particular solution of the specific Ricatti relationship

$$V_p^+(x)=\frac{\hbar^2}{2m}(\beta_p^2(x)+\beta_p'(x))-C_p \tag{3.21}$$

in such a way that the corresponding solution of Eq. (3.19) is

$$\beta(x)=\beta_g(x)=\beta_p(x)+\frac{b}{\rho(x)}, \tag{3.22}$$

where the subindex g is used to denote the general solution and

$$\rho(x)=e^{2\int\beta_p(x)dx}\left(\gamma+b\int e^{-2\int\beta_p(x)dx}dx\right). \tag{3.23}$$

Thus, according to Eq. (2.16) and Eq. (2.24), the one-dimensional modified potential is given by

$$V_p^-(x) = V_p^+(x) - \frac{\hbar^2}{m}\beta_p'(x) \tag{3.24}$$

and

$$V_g^-(x) = V_p^-(x) - \frac{\hbar^2}{m}\frac{d}{dx}\left(\frac{b}{\rho(x)}\right) \tag{3.25}$$

is the corresponding generalized potential.

In consequence, if in Eq. (3.21) we choose the ansatz $\beta_p(x) = B$, where B is any constant, it obtained the specific potential

$$V_p^+(x) = \frac{\hbar^2 B^2}{2m} - C_p\,. \tag{3.26}$$

Clearly, without loss of generality we can put in the previous equation $C_p = (\hbar^2 B^2/2m)$ in order to shown that $V_p^+(x)$ and the corresponding modified $V_p^-(x)$ potentials are null which means that both are the well known one-dimensional potential for the free particle model. In that case, the first order operational equations

$$a_p^{\pm} = \frac{\hbar}{\sqrt{2m}}\left(B \mp \frac{d}{dx}\right) \tag{3.27}$$

factorize the corresponding Hamiltonians $H_p^{\pm}$ by means of

$$a_p^{\mp} a_p^{\pm} = H_p^{\pm} + C_p\,, \tag{3.28}$$

where $H_p^- = H_p^+ = -(\hbar^2/2m)(d^2/dx^2)$.

Finally, with the purpose of identifying the generalized potential that corresponds to this new situation of $\beta_p(x) = B$, it is necessary to consider Eq. (3.23) in order to find

$$\rho(x) = \gamma e^{2Bx} - \frac{b}{2B}. \tag{3.29}$$

In consequence, putting $b = 1$ in the above relationship and using Eq. (3.25), we found that the new generalized or partner potential associated with the one-dimensional free particle system is given by

$$V_g^-(x) = \frac{(2B)^3 \hbar^2 \gamma e^{2Bx}}{m(2B\gamma e^{2Bx} - 1)^2} \tag{3.30}$$

in such a way that the respective generalized Hamiltonian

$$H_g^- = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{(2B)^3 \hbar^2 \gamma e^{2Bx}}{m(2B\gamma e^{2Bx} - 1)^2}, \tag{3.31}$$

as well as $H_p^{\pm}$, are factorized according to the rule

$$A^+ A^- = H_g^- + C \text{ and } A^- A^+ = H_p^{\pm} + C, \tag{3.32}$$

where $C = C_p$ and

$$A^{\pm}=\frac{\hbar}{\sqrt{2m}}\left(B+\frac{2B}{2B\gamma e^{2Bx}-1}\mp\frac{d}{dx}\right). \tag{3.33}$$

Concerning the generalized wave functions, if we consider the most simple form $\psi_p=\sin(kx)$ for the eigenfunctions of the particular one-dimensional free particle Hamiltonian, it becomes that the ϕ_g eigenfunctions of the corresponding generalized Hamiltonian are given, according to Eq. (2.26), by

$$\phi_g=A^{+}\psi_p=\frac{\hbar}{\sqrt{2m}}\left(k\sin(kx)+\frac{2k\sin(kx)}{2k\gamma e^{2kx}-1}-k\cos(kx)\right), \tag{3.34}$$

where, without loss of generality, we have used $B=k$ in order to normalize the exponential function.

This means that, in a similar way to the three-dimensional case, we have

$$H_g^{-}\phi_g=\frac{\hbar^3k^3}{(2m)^{3/2}}\left(\sin(kx)+\frac{2\sin(kx)}{2k\gamma e^{2kx}-1}-\cos(kx)\right). \tag{3.35}$$

Thus, by identifying in the previous equation those terms defining the wave function ϕ_g, it obtained the eigenvalue relationship

$$H_g^{-}\phi_g=E_g\phi_g=\frac{\hbar^2k^2}{2m}\phi_g, \tag{3.36}$$

which means that $E_g=E_p$, as should be fulfilled in order to have a generalized Hamiltonian H_g^{-} which should be isospectral when compared with the particular Hamiltonians $H_p^{\pm}$.

Before closing this work, with the Concluding Remarks, it is important to point out that in the above development it was considered a one-dimensional free particle Hamiltonian which does not have any negative energy eigenvalue. In fact, the positive energy spectrum results from the selection of $\psi_p=\sin(kx)$ as eigenfunctions of $H_p^{\pm}$. That means that the existence of negative energy eigenvalues for the same Hamiltonian occurs only on the condition of having an appropriate wave function. That is, for $V(x)=0$ exists a spectrum with negative energy E_{p-} when $\psi_{p-}=\sinh(kx)$ is taken as an eigenfunction of the one-dimensional free particle Hamiltonian although it is well known that this form of the wave function is unnormalizable. In spite of the above, in this new situation it becomes that the eigenfunctions for the generalized Hamiltonian H_g^{-} are given by

$$\phi_{g-}=A^{+}\psi_{p-}=\frac{\hbar k}{\sqrt{2m}}\left(\sinh(kx)+\frac{2\sinh(kx)}{2k\gamma e^{2kx}-1}-\cosh(kx)\right), \tag{3.37}$$

where the down sign minus in the wave functions is used to denote negative energy eigenvalues, for which

$$H_g^{-}\phi_{g-}=-\frac{\hbar^2k^2}{2m}\left(\frac{\hbar k}{\sqrt{2m}}\right)\left(\sinh(kx)+\frac{2\sinh(kx)}{2k\gamma e^{2kx}-1}-\cosh(kx)\right). \tag{3.38}$$

Clearly, this last relationship is rewritten as

$$H_g^{-}\phi_{g-}=E_{g-}\phi_{g-}=-\frac{\hbar^2k^2}{2m}\phi_{g-}, \tag{3.39}$$

which means, as before in the case of unbounded states, that the generalized Hamiltonian is also isospectral with negative energy spectrum $E_{g-}=-(\hbar^2k^2/2m)=E_{p-}$ in good agreement with Sukumar.[8]

IV. CONCLUDING REMARKS

This work is concerned with the fact that quantum mechanical problems can be solved from the factorization of the Hamiltonian instead of, as usual, the factorization of the Schrödinger equation for any arbitrary potential. For that, in this paper we have shown how two Ricatti-type relationships can be used in order to identify standard potentials as well as to find the corresponding modified and generalized potentials which, as far as we know, are so-called isospectral potentials. It is important to emphasize that the proposed procedure is general, due to the fact that in its development the potential involved is always considered anyone. For that reason, it becomes evident that the displayed method can be used to obtain new potentials that can be incorporated in the treatment of different quantum mechanical applications. This new potentials come from a simple proposition of an ansatz which is necessary to start the procedure for obtaining the corresponding generalized potential. In order to show the usefulness of the proposed approach, we have obtained the generalized three-dimensional potential associated with the free particle model when an ansatz null was used. In this respect, it is important to point out that the generalized potential that corresponds to the one-dimensional case it is not necessarily unique. For example, we have shown that a generalized potential for the one-dimensional free particle model is derived from the use of the three-dimensional treatment for the particular case $l=0$. However, another different generalized potential for the same situation results as a consequence of the use of the one-dimensional version of the proposed approach with a different ansatz. In any case, as expected, we explicitly prove that in all cases the new Hamiltonians are isospectral due to the fact that they lead to the same energy spectra which is obtained when using specific or standard Hamiltonians.

ACKNOWLEDGMENTS

We want to acknowledge the referee for having pointed out to us the convenience of including the treatment of a bounded state which has improved the results presented here. This work was supported by CONACYT-México, under Scientific Project No. 5-4840E.

[1] P. A. M. Dirac, Z. Phys. **3**, 289 (1920).
[2] L. Infeld and E. Hull, Rev. Mod. Phys. **23**, 21 (1951).
[3] O. L. de Lange and R. E. Raab, Phys. Rev. A **37**, 1858 (1988).
[4] J. Morales, J. J. Peña, M. Sánchez, and J. López-Bonilla, Int. J. Quantum Chem., Symp. **25**, 155 (1991).
[5] J. Morales, G. Arreaga, J. J. Peña, and J. López-Bonilla, Int. J. Quantum Chem., Symp. **26**, 171 (1992).
[6] B. Mielnik, J. Math. Phys. **25**, 3387 (1984).
[7] Z. Dongpei, J. Phys. A **20**, 4331 (1987).
[8] C. V. Sukumar, J. Phys. A **18**, 2917 (1985).
[9] S. Seshadri, V. Balakrishnan, and S. Lakahimibala, J. Math. Phys. **39**, 838 (1998).
[10] V. G. Korolev, J. Phys. A **31**, 9297 (1998).
[11] R. Dutt, A. Share, and U. P. Sukhatme, Am. J. Phys. **56**, 163 (1988).
[12] D. J. Fernández, Lett. Math. Phys. **8**, 337 (1984).
[13] E. Drigo Filho J. Phys. A **21**, L1025 (1988).
[14] J. Morales, J. J. Peña, G. Ovando, and V. Gaftoi, Int. J. Quantum Chem. **71**, 465 (1999).
[15] R. A. Buckingham, ''4. The Continuum'' in *Quantum Theory, I. Elements*, edited by D. R. Bates (Academic, New York, 1961).
[16] R. A. Buckingham, ''3. Exactly Soluble Bound State Problems,'' in *Quantum Theory, I. Elements*, edited by D. R. Bates (Academic, New York, 1961).
[17] L. I. Schiff, *Quantum Mechanics* (McGraw-Hill, Kogokusha, tokyo, 1968).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Derivation of the wave function collapse in the context of Nelson's stochastic mechanics

Michele Pavon[a)]
Dipartimento di Elettronica e Informatica, Università di Padova, via Gradenigo 6/A, and LADSEB-CNR, 35131 Padova, Italy

(Received 20 January 1999; accepted 25 February 1999)

The von Neumann collapse of the quantum mechanical wave function after a position measurement is derived by a purely probabilistic mechanism in the context of Nelson's stochastic mechanics. © *1999 American Institute of Physics.* [S0022-2488(99)01109-3]

I. INTRODUCTION

Nelson's stochastic mechanics[1–5] is a quantization procedure for classical dynamical systems based on stochastic processes of the diffusion type. This theory leads to predictions that agree with those of standard quantum mechanics and are confirmed by experiment. The fundamental assumption is that interaction with a background field causes the system to undergo a diffusion process with diffusion coefficient $\hbar/m$. A fascinating hypothesis concerning the origin of the underlying Brownian motion has been recently advanced by Calogero in Ref. 6. Namely, that this ''tremor'' may be caused by the interaction of every particle with the gravitational force due to all other particles of the Universe. Following this idea, he obtains a formula for Planck's action constant h. The latter yields the correct order of magnitude for h when current cosmological data are employed.

It is hardly surprising that the most controversial issue in stochastic mechanics is the measurement problem. Indeed, in Ref. 7, Francesco Guerra writes: ''Therefore, we see that the basic problem in the interpretation of stochastic mechanics is related to the basic problem in the interpretation of quantum mechanics: To evaluate the effects of the measurement and explain the mechanism of the wave packet reduction.''

Our purpose in this paper is to show that, in the frame of Nelson's stochastic mechanics, *the wave function reduction after a position measurement may be obtained through a purely probabilistic mechanism*, namely a stochastic variational principle. The latter has the appealing interpretation of changing the pair of forward and backward drifts of the reference process as little as possible given the result of the measurement. This variational principle is quite similar to the one that yields the new stochastic model after measurement for nonequilibrium thermodynamical systems, see Sec. V, the only difference being that, in view of the time reversibility of stochastic mechanics, a time-symmetric kinematics has to be employed. As we have shown elsewhere,[8–10] this kinematics also permits us to develop in a natural way a Lagrangian and a Hamiltonian formalism in stochastic mechanics. In particular, it permits us to define a *momentum process* having the same first and second moment of the corresponding quantum momentum operator. It is then possible to derive a stochastic counterpart of Hamilton's canonical equations, and to obtain a simple probabilistic interpretation of the uncertainty principle[9] along the lines of Refs. 1, 11–13.

II. KINEMATICS OF FINITE-ENERGY DIFFUSIONS

In this section, we review some essential concepts and results of the kinematics of the diffusion processes. We refer the reader to Refs. 2–5 and 14–18 for a thorough account. Let

[a)]Electronic mail: pavon@dei.unipd.it

$(\Omega,\mathcal{E},\mathbf{P})$ be a probability space, and let I_n denote the $n\times n$ identity matrix. A stochastic process $\{\xi(t);t_0\leqslant t\leqslant t_1\}$ mapping $[t_0,t_1]$ into $L_n^2(\Omega,\mathcal{E},\mathbf{P})$ is called a *finite-energy diffusion* with constant diffusion coefficient $I_n\sigma^2$ if the increments admit the representation

$$\xi(t)-\xi(s)=\int_s^t\beta(\tau)d\tau+\sigma[w_+(t)-w_+(s)],\quad t_0\leqslant s<t\leqslant t_1, \tag{II.1}$$

where the *forward drift* $\beta(t)$ is at each time t a measurable function of the past $\{\xi(\tau);0\leqslant\tau\leqslant t\}$, and $w_+(\cdot)$ is a standard, n-dimensional *Wiener process* with the property that $w_+(t)-w_+(s)$ is independent of $\{\xi(\tau);0\leqslant\tau\leqslant s\}$. Moreover, β must satisfy the finite-energy condition

$$E\left\{\int_{t_0}^{t_1}\beta(t)\cdot\beta(t)dt\right\}<\infty. \tag{II.2}$$

In Ref. 16, Föllmer has shown that a finite-energy diffusion also admits a reverse-time differential. Namely, there exists a measurable function $\gamma(t)$ of the future $\{\xi(\tau);t\leqslant\tau\leqslant t_1\}$ called *backward drift*, and another Wiener process w_- such that

$$\xi(t)-\xi(s)=\int_s^t\gamma(\tau)d\tau+\sigma[w_-(t)-w_-(s)],\quad t_0\leqslant s<t\leqslant t_1. \tag{II.3}$$

Moreover, γ satisfies

$$E\left\{\int_{t_0}^{t_1}\gamma(t)\cdot\gamma(t)dt\right\}<\infty, \tag{II.4}$$

and $w_-(t)-w_-(s)$ is independent of $\{\xi(\tau);t\leqslant\tau\leqslant t_1\}$. Let us agree that dt always indicates a strictly positive variable. For any function f defined on $[t_0,t_1]$, let

$$d_+f(t):=f(t+dt)-f(t)$$

be the *forward increment* at time t, and

$$d_-f(t)=f(t)-f(t-dt)$$

be the *backward increment* at time t. For a finite-energy diffusion, Föllmer has also shown in Ref. 16 that the forward and backward drifts may be obtained as Nelson's conditional derivatives, namely

$$\beta(t)=\lim_{dt\searrow 0}E\left\{\frac{d_+\xi(t)}{dt}\middle|\xi(\tau),t_0\leqslant\tau\leqslant t\right\}, \tag{II.5}$$

and

$$\gamma(t)=\lim_{dt\searrow 0}E\left\{\frac{d_-\xi(t)}{dt}\middle|\xi(\tau),t\leqslant\tau\leqslant t_1\right\}, \tag{II.6}$$

the limits being taken in $L_n^2(\Omega,\mathcal{B},P)$. It was finally shown in Ref. 16 that the one-time probability density $\rho(\cdot,t)$ of $\xi(t)$ (which exists for every $t>t_0$) is absolutely continuous on $\mathbb{R}^n$ and the following relation holds a.s. $\forall t>0$

$$E\{\beta(t)-\gamma(t)|\xi(t)\}=\sigma^2\nabla\log\rho(\xi(t),t). \tag{II.7}$$

Let ξ be a finite-energy diffusion satisfying (II.1) and (II.3). Let $f:\mathbb{R}^n\times[t_0,t_1]\to\mathbb{R}$ be twice continuously differentiable with respect to the spatial variable and once with respect to time. Then, we have the following change of variables formulas:

$$f(\xi(t),t)-f(\xi(s),s)=\int_s^t\left(\frac{\partial}{\partial\tau}+\beta(\tau)\cdot\nabla+\frac{\sigma^2}{2}\Delta\right)f(\xi(\tau),\tau)d\tau \tag{II.8}$$

$$+\int_s^t\sigma\nabla f(\xi(\tau),\tau)\cdot d_+w_+(\tau), \tag{II.9}$$

$$f(\xi(t),t)-f(\xi(s),s)=\int_s^t\left(\frac{\partial}{\partial\tau}+\gamma(\tau)\cdot\nabla-\frac{\sigma^2}{2}\Delta\right)f(\xi(\tau),\tau)d\tau \tag{II.10}$$

$$+\int_s^t\sigma\nabla f(\xi(\tau),\tau)\cdot d_-w_-(\tau). \tag{II.11}$$

The stochastic integrals appearing in (II.9) and (II.11) are a (forward) Ito integral and a backward Ito integral, respectively; see Ref. 15 for details. Let us introduce the *current drift* $v(t):=(\beta(t)+\gamma(t))/2$ and the *osmotic drift* $u(t):=(\beta(t)-\gamma(t))/2$. Notice that, when σ tends to zero, v tends to $\dot{\xi}$, and u tends to zero. The semisum and the semidifference of (II.9) and (II.11) give two more useful formulas:

$$f(\xi(t),t)-f(\xi(s),s)=\int_s^t\left(\frac{\partial}{\partial\tau}+v(\tau)\cdot\nabla\right)f(\xi(\tau),\tau)d\tau$$

$$+\frac{\sigma}{2}\left[\int_s^t\nabla f(\xi(\tau),\tau)\cdot d_+w_++\int_s^t\nabla f(\xi(\tau),\tau)\cdot d_-w_-\right], \tag{II.12}$$

$$0=\int_s^t\left(u(\tau)\cdot\nabla+\frac{\sigma^2}{2}\Delta\right)f(\xi(\tau),\tau)d\tau+\frac{\sigma}{2}\left[\int_s^t\nabla f(\xi(\tau),\tau)\cdot d_+w_+-\int_s^t\nabla f(\xi(\tau),\tau)\cdot d_-w_-\right]. \tag{II.13}$$

Specializing (II.12) and (II.13) to $f(x,t)=x$, we get

$$\xi(t)-\xi(s)=\int_s^t v(\tau)d\tau+\frac{\sigma}{2}[w_+(t)-w_+(s)+w_-(t)-w_-(s)], \tag{II.14}$$

$$0=\int_s^t u(\tau)d\tau+\frac{\sigma}{2}[w_+(t)-w_+(s)-w_-(t)+w_-(s)]. \tag{II.15}$$

The finite-energy diffusion $\xi(\cdot)$ is called *Markovian* if there exist two measurable functions $b_+(\cdot,\cdot)$, and $b_-(\cdot,\cdot)$, such that $\beta(t)=b_+(\xi(t),t)$ a.s. and $\gamma(t)=b_-(\xi(t),t)$ a.s., for all t in $[t_0,t_1]$. The duality relation (II.7) now reads as

$$b_+(\xi(t),t)-b_-(\xi(t),t)=\sigma^2\nabla\log\rho(\xi(t),t). \tag{II.16}$$

This immediately gives the *osmotic equation*,

$$u(x,t)=\frac{\sigma^2}{2}\nabla\log\rho(x,t), \tag{II.17}$$

where $u(x,t):=(b_+(x,t)-b_-(x,t))/2$. The probability density $\rho(\cdot,\cdot)$ of $\xi(t)$ satisfies (at least weakly) the *Fokker–Planck equation*,

$$\frac{\partial\rho}{\partial t}+\nabla\cdot(b_+\rho)=\frac{\sigma^2}{2}\Delta\rho.$$

The latter can also be rewritten, in view of (II.16), as the *equation of continuity* of hydrodynamics,

$$\frac{\partial\rho}{\partial t}+\nabla\cdot(v\rho)=0, \tag{II.18}$$

where $v(x,t):=(b_+(x,t)+b_-(x,t))/2$.

III. A TIME-SYMMETRIC KINEMATICS FOR DIFFUSION PROCESSES

We recall here the basic facts from the time-symmetric kinematics developed in Refs. 8, 19. Let us multiply (II.15) by $-i$, and add it to (II.14). We get

$$\xi(t)-\xi(s)=\int_s^t[v(\tau)-iu(\tau)]d\tau+\frac{\sigma}{2}[(1-i)(w_+(t)-w_+(s))+(1+i)(w_-(t)-w_-(s))]. \tag{III.19}$$

We call $v_q(t):=v(t)-iu(t)$ the *quantum drift*, and

$$w_q(t):=\frac{1-i}{2}w_+(t)+\frac{1+i}{2}w_-(t) \tag{III.20}$$

the *quantum noise*. Hence, we can rewrite (III.19) as

$$\xi(t)-\xi(s)=\int_s^t v_q(\tau)d\tau+\sigma[w_q(t)-w_q(s)]. \tag{III.21}$$

At first sight, this decomposition of the *real-valued* increments of ξ into the sum of two *complex* quantities might look somewhat odd. Nevertheless, this representation enjoys several important properties:

(1) When σ^2 tends to zero, $v-iu$ tends to $\dot{\xi}$.

(2) The quantum drift $v_q(t)$ contains at each time t precisely the same information as the pair $(v(t),u(t))$ [or, equivalently, the pair $(\beta(t),\gamma(t))$].

(3) The representation (III.21), differently from (II.1) and (II.3), enjoys an important symmetry with respect to time. Indeed, under time reversal, (III.21) transforms into

$$\xi(t)-\xi(s)=\int_s^t\overline{v_q(\tau)}d\tau+\sigma[\overline{w_q(t)-w_q(s)}], \tag{III.22}$$

where an overbar indicates conjugation; see Ref. 9, p. 145.

The representation (III.21) has proven to be crucial in order to develop a Lagrangian and Hamiltonian dynamics formalism in the context of Nelson's stochastic mechanics; see Refs. 8–10. In particular, to develop the second form of Hamilton's principle, the key tool has been a change of variables formula related to representation (III.21). In order to recall such a formula, we need first to define stochastic integrals with respect to the quantum noise w_q. Let us denote by $d_bf(t):=[(1-i)/2]\,d_+f(t)+[(1+i)/2]\,d_-f(t)$ the *bilateral increment* of f at time t. Then, from (III.20) and (II.15), we get

$$d_+w_q(t)=\frac{1+i}{\sigma}u(x(t),t)dt+d_+w_++o(dt),$$

$$d_- w_q(t) = \frac{-1+i}{\sigma} u(x(t),t)dt + d_+ w_- + o(dt).$$

These, in turn, give immediately

$$d_b w_q(t) := \frac{1-i}{2} d_+ w_+(t) + \frac{1+i}{2} d_- w_-(t) + o(dt). \tag{III.23}$$

Let $f(x,t)$ be a measurable, $\mathbb{C}^n$-valued function, such that

$$P\left\{ \omega: \int_0^T f(\xi(t),t) \cdot \overline{f(\xi(t),t)} dt < \infty \right\} = 1.$$

In view of (III.23), we define

$$\int_s^t f(\xi(\tau),\tau) \cdot d_b w_q(\tau) := \frac{1-i}{2} \int_s^t f(\xi(\tau),\tau) \cdot d_+ w_+(\tau) + \frac{1+i}{2} \int_s^t f(\xi(\tau),\tau) \cdot d_- w_-(\tau).$$

Thus, integration with respect to the bilateral increments of w_q is defined through a linear combination with complex coefficients of a forward and a backward Ito integral. Let $f(x,t)$ be a complex-valued function with real and imaginary parts of class $C^{2,1}$. Then, multiplying (II.13) by $-i$, and then adding it to (II.12), we get the change of variables formula

$$f(\xi(t),t) - f(\xi(s),s) = \int_s^t \left(\frac{\partial}{\partial \tau} + v_q(\tau) \cdot \nabla - \frac{i\sigma^2}{2} \Delta \right) f(\xi(\tau),\tau) d\tau + \int_s^t \sigma \nabla f(\xi(\tau),\tau) \cdot d_b w_q(\tau). \tag{III.24}$$

It is important to understand that this formula, and, in particular, the coefficient of the Laplacian term, follows from basic probabilistic arguments.

IV. THE QUANTUM HAMILTON PRINCIPLE

Stochastic mechanics may be based, since the fundamental paper by Guerra and Morato,[20] on stochastic variational principles of a hydrodynamic type. Other versions of the variational principle have been proposed in Refs. 4, 14, and in Ref. 8. We outline here the quantum Hamilton principle of Ref. 8, since it employs the time-symmetric kinematics of Sec. III that we shall need to derive the wave function collapse.

Let $\mathcal{X}_{\rho_1}$ denote the family of all finite-energy, $\mathbb{R}^n$-valued diffusions on $[t_0,t_1]$ with diffusion coefficient $I_n(\hbar/m)$, and having marginal probability density ρ_1 at time t_1. Let $\mathcal{V}$ denote the family of finite-energy, C^n-valued stochastic processes on $[t_0,t_1]$. Let $L(x,v) := \frac{1}{2} m v \cdot v - V(x)$ be defined on $R^n \times C^n$. Also, let S_0 be a complex-valued function on R^n. Consider the problem of extremizing on $(x,v_q) \in (\mathcal{X}_{\rho_1} \times \mathcal{V})$,

$$E\left\{ \int_{t_0}^{t_1} L(x(t),v_q(t))dt + S_0(x(t_0)) \right\}, \tag{IV.25}$$

subject to the constraint that

$$x \text{ has quantum drift (velocity) } v_q. \tag{IV.26}$$

Notice that the quadratic term in the Lagrangian may be rewritten in terms of the forward and backward drifts as follows:

$$\frac{m}{2}v_q(t)\cdot v_q(t)=\frac{m}{2}\left[\frac{1-i}{2}\beta(t)+\frac{1+i}{2}\gamma(t)\right]\cdot\left[\frac{1-i}{2}\beta(t)+\frac{1+i}{2}\gamma(t)\right]$$

$$=\frac{-im}{4}[\beta(t)\cdot\beta(t)+2i\beta(t)\cdot\gamma(t)-\gamma(t)\cdot\gamma(t)]$$

$$=\frac{-im}{4}[(\beta(t)+i\gamma(t))\cdot(\beta(t)+i\gamma(t))] \tag{IV.27}$$

In Ref. 8, Sec. VIII, the following result was established.

Theorem IV.1: *Suppose that $S_q(x,t)$ of class $C^{2,1}$ solves on $[t_0,t_1]$ the initial value problem,*

$$\frac{\partial S_q}{\partial t}+\frac{1}{2m}\nabla S_q\cdot\nabla S_q+V(x)-\frac{i\hbar}{2m}\Delta S_q=0, \tag{IV.28}$$

$$S_q(x,t_0)=S_0(x), \tag{IV.29}$$

and satisfies the technical condition

$$E\left\{\int_{t_0}^{t_1}\nabla S_q(x(t),t)\cdot\overline{\nabla S_q(x(t),t)}dt\right\}<\infty,\quad \forall x\in\mathcal{X}_{\rho_1}. \tag{IV.30}$$

Then, any $x\in\mathcal{X}_{\rho_1}$ having quantum drift $(1/m)\nabla S(x(t),t)$ solves the extremization problem.

A crucial role in the proof is played by the change of variables formula (III.24) that here reads as

$$f(\xi(t),t)-f(\xi(s),s)=\int_s^t\left(\frac{\partial}{\partial\tau}+v_q(\tau)\cdot\nabla-\frac{i\hbar}{2m}\Delta\right)f(\xi(\tau),\tau)d\tau+\int_s^t\sqrt{\frac{\hbar}{m}}\nabla f(\xi(\tau),\tau)\cdot d_b w_q(\tau). \tag{IV.31}$$

The existence of a solution for the apparently complicated nonlinear, complex Cauchy problem (IV.28)–(IV.29) is dealt with as follows. Let $\{\psi(x,t);t_0\leq t\leq t_1\}$ be the solution of the *Schrödinger equation*,

$$\frac{\partial\psi}{\partial t}=\frac{i\hbar}{2m}\Delta\psi-\frac{i}{\hbar}V(x)\psi, \tag{IV.32}$$

with initial condition $\psi_0(x):=\exp(i/\hbar)S_0(x)$. If $\psi(x,t)$ never vanishes on $\mathbb{R}^n\times[t_0,t_1]$, and satisfies the condition

$$E\left\{\int_{t_0}^{t_1}\nabla\log\psi(x(t),t)\cdot\overline{\nabla\log\psi(x(t),t)}dt\right\}<\infty,\quad\forall x\in\mathcal{X}_{\rho_1}; \tag{IV.33}$$

then $S_q(x,t):=(\hbar/i)\log\psi(x,t)$ satisfies (IV.28)–(IV.29) and (IV.30). If, moreover, $\psi_0(x)$ has L^2 norm 1, and the terminal density satisfies $\rho_1(x,t)=|\psi(x,t_1)|^2$, then there does exist a Markov diffusion having the required quantum drift, namely, the *Nelson process* associated to $\{\psi(x,t);t_0\leq t\leq t_1\}$, and Born's relation $\rho(x,t)=|\psi(x,t)|^2$ holds; see Ref. 8 for the details. The construction of the Nelson process corresponding to $\psi(x,t)$ in the case where $\psi(x,t)$ vanishes requires considerable care. It is discussed in Ref. 21 and Ref. 5, Chap. IV, and references therein.

V. MEASUREMENT IN NONEQUILIBRIUM THERMODYNAMICS

In this section, we discuss the measurement for nonequilibrium thermodynamical systems. This serves as an introduction to measurement in stochastic mechanics to be discussed in the following section. Consider an open thermodynamical system whose macroscopic evolution is modeled by an n-dimensional Markov diffusion process $\{x(t);t_0 \leqslant t\}$ with a forward Ito differential,

$$d_+ x(t) = b_+(x(t))dt + \sigma\, d_+ w_+ .$$

Let $\rho(x,t)$ denote the probability density of $x(t)$ satisfying the Fokker–Planck equation,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (b_+ \rho) = \frac{\sigma^2}{2} \Delta \rho. \tag{V.34}$$

The *equilibrium state* is given by the Maxwell–Boltzmann distribution law,

$$\overline{\rho}(x) = C \exp\left[-\frac{H(x)}{kT} \right],$$

where H is the Hamiltonian function, and we have the relation

$$b_+(x) = -\frac{\sigma^2}{2kT} \nabla H(x),$$

where k is Boltzmann's constant and T is the absolute temperature. Suppose that at time t_1 a measurement is made that yields the new probability density $\tilde{\rho}(x,t_1)$. Let $\mathcal{X}_{\tilde{\rho}(t_1)}$ denote the class of finite-energy diffusions on $[t_1,t_2]$ with diffusion coefficient σ^2 and having marginal $\tilde{\rho}(x,t_1)$ at time t_1. Let us pose the following question: Among all processes in $\mathcal{X}_{\tilde{\rho}(t_1)}$, which one should we use to model the macroscopic evolution of the system from t_1 up to t_2? Everybody agrees that we should employ the stochastic process $\{\tilde{x}(t);t_1 \leqslant t \leqslant t_2\}$ that has the same forward drift field $b_+(x)$ of the ''reference'' process x. This is supported by the observation that the new process must have the same equilibrium distribution of the previous one. Let us show that the new process $\{\tilde{x}(t);t_1 \leqslant t \leqslant t_2\}$ may be obtained as the solution of a variational problem. Assume that the Kullback–Leibler pseudodistance between $\tilde{\rho}(t_1)$ and $\rho(t_1)$ is finite, namely,

$$H(\tilde{\rho}(t_1),\rho(t_1)) := E\left\{ \log \frac{\tilde{\rho}(\tilde{x}(t_1),t_1)}{\rho(\tilde{x}(t_1),t_1)} \right\} = \int_{\mathbb{R}^n} \log \frac{\tilde{\rho}(\tilde{x},t_1)}{\rho(\tilde{x},t_1)} \tilde{\rho}(x,t_1)dx < \infty.$$

Let $\mathbb{D}_{\tilde{\rho}(t_1)}$ denote the class of probability measures on $\Omega = C([t_1,t_2])$ that are equivalent to the measure P induced by the reference process $\{x(t);t_1 \leqslant t \leqslant t_2\}$. For $Q \in \mathbb{D}_{\tilde{\rho}(t_1)}$, let

$$H(Q,P) = E_Q\left[\log \frac{dQ}{dP} \right]$$

denote the *relative entropy* of Q with respect to P. It then follows from Girsanov's theorem that[16,17]

$$H(Q,P) = H(\tilde{\rho}(t_1),\rho(t_1)) + E_Q\left[\int_{t_1}^{t_2} \frac{1}{2\sigma^2} [b_+(\tilde{x}(t)) - \beta^Q(t)] \cdot [b_+(\tilde{x}(t)) - \beta^Q(t)] dt \right].$$

Since $H(\tilde{\rho}(t_1),\rho(t_1))$ is constant over $\mathbb{D}_{\tilde{\rho}(t_1)}$, it trivially follows that the probability measure $\tilde{Q}$ corresponding to the process $\tilde{x}$ having forward drift b_+ minimizes $H(Q,P)$ over $\mathbb{D}_{\tilde{\rho}(t_1)}$. This problem may be interpreted as a problem of large deviation of the empirical distribution according to Schrödinger's original motivation.[22,17] We consider now an apparently different variational problem that has the same solution as the previous one. We do so because it is this second form, which, in a suitably modified form, applies to the quantum case. Let $\mathcal{X}_{\tilde{\rho}_2}$ denote the family of finite-energy diffusions on $[t_1,t_2]$ with diffusion coefficient σ^2 and having marginal density $\tilde{\rho}_2$ at time t_2. Consider the problem of minimizing with respect to the pair $(\tilde{x},\gamma)$ the functional

$$E\left\{\int_{t_1}^{t_2}\frac{1}{2\sigma^2}[b_-(\tilde{x}(t))-\gamma(t)]\cdot[b_-(\tilde{x}(t))-\gamma(t)]dt-\log\frac{\tilde{\rho}(\tilde{x}(t_1),t_1)}{\rho(\tilde{x}(t_1),t_1)}\right\},$$

subject to the constraint that γ be the backward drift of $\tilde{x}$ on $[t_1,t_2]$. This problem is a variant of the one first considered and solved in Ref. 23, Theorem 2. The connection between the two variational problems, and their relation to the theory of Schrödinger processes and bridges, has been thoroughly investigated in Ref. 24. In order to solve this problem, rather than reproducing the arguments in Refs. 23, 24, we take the opportunity to introduce the variational method based on nonlinear Lagrange functionals.[25] This method permits us to solve also the more complicated quantum case. Suppose that we wish to minimize $J:Y\to\overline{\mathbb{R}}$, where $\overline{\mathbb{R}}$ denotes the extended reals, over the nonempty subset S of Y.

Lemma V.1: (Lagrange Lemma): *Let* $\Lambda:Y\to\overline{\mathbb{R}}$ *and let* $y_0\in S$ *minimize* $J+\Lambda$ *over* Y. *Assume that* $\Lambda(\cdot)$ *is* finite *and* constant *over* S. *Then* y_0 *minimizes* J *over* S.

Proof: For any $y\in S$, we have $J(y_0)+\Lambda(y_0)\leqslant J(y)+\Lambda(y)=J(y)+\Lambda(y_0)$. Hence $J(y_0)\leqslant J(y)$. □

A functional Λ that is constant and finite on S is called a *Lagrange functional*. Obviously, a similar result holds if the problem is an extremization problem. Let us apply this simple idea to the above problem. Let $\varphi(x,t)$ be a real-valued function of class $C^{2,1}$ defined on $\mathbb{R}^n\times[t_1,t_2]$, and satisfying the technical condition

$$E\left\{\int_{t_1}^{t_2}\nabla\varphi(x(t),t)\cdot\nabla\varphi(x(t),t)dt\right\}<\infty,\quad \forall x\in\mathcal{X}_{\tilde{\rho}_2}. \tag{V.35}$$

Corresponding to such a φ, we introduce the functional,

$$\Lambda^{\varphi}(\tilde{x},\gamma):=E\Bigg\{\varphi(\tilde{x}(t_2),t_2)-\varphi(\tilde{x}(t_1),t_1) + \int_{t_1}^{t_2}\left[-\frac{\partial\varphi}{\partial t}(\tilde{x}(t),t)-\gamma(t)\cdot\nabla\varphi(\tilde{x}(t),t)+\frac{\sigma^2}{2}\Delta\varphi(\tilde{x}(t),t)\right]dt\Bigg\}.$$

In view of (II.11) and (V.35), we have that $\Lambda^{\varphi}(\tilde{x},\gamma)=0$ whenever the pair $(\tilde{x},\gamma)$ satisfies the constraint, since the stochastic integral has zero expectation. Thus, it is a *Lagrange functional* for the problem. Consider next the *unconstrained* minimization of the functional $J+\Lambda^{\varphi}$. For a fixed $\tilde{x}\in\mathcal{X}_{\tilde{\rho}_2}$, and a fixed time $t\in[t_1,t_2]$, we consider the *pointwise* minimization of the integrand of $J+\Lambda^{\varphi}$ with respect to γ,

$$\text{minimize}_{\gamma\in R^n}\left\{\frac{1}{2\sigma^2}(b_-(\tilde{x}(t),t)-\gamma)\cdot(b_-(\tilde{x}(t),t)-\gamma)-\gamma\cdot\nabla\varphi(\tilde{x}(t),t)\right\}.$$

We get

$$\gamma^o(\tilde{x})(t)=b_-(\tilde{x}(t),t)+\sigma^2\,\nabla\varphi(\tilde{x}(t),t). \tag{V.36}$$

Substituting back expression (V.36) into $J+\Lambda^{\varphi}$, we get the following functional of $\tilde{x}$:

$$(J+\Lambda^{\varphi})(\tilde{x},\gamma^{o}(\tilde{x})):=E\left\{\varphi(\tilde{x}(t_2),t_2)-\varphi(\tilde{x}(t_1),t_1)-\log\frac{\tilde{\rho}(\tilde{x}(t_1),t_1)}{\rho(\tilde{x}(t_1),t_1)}\right.$$
$$+\int_{t_1}^{t_2}\left[-\frac{\sigma^2}{2}\nabla\varphi(\tilde{x}(t),t)\cdot\nabla\varphi(\tilde{x}(t),t)-\frac{\partial\varphi}{\partial t}(\tilde{x}(t),t)\right.$$
$$\left.\left.-b_-(\tilde{x}(t),t)\cdot\nabla\varphi(\tilde{x}(t),t)+\frac{\sigma^2}{2}\Delta\varphi(\tilde{x}(t),t)\right]dt\right\}. \tag{V.37}$$

Next, we seek to find a function φ such that the functional $(J+\Lambda^{\varphi})(\tilde{x},\gamma^{o}(\tilde{x}))$ becomes constant over $\mathcal{X}_{\tilde{\rho}_2}$. Suppose φ solves on $[t_1,t_2]$ the initial value problem,

$$\frac{\partial\varphi}{\partial t}+b_-(x,t)\cdot\nabla\varphi(x,t)-\frac{\sigma^2}{2}\Delta\varphi(x,t)=-\frac{\sigma^2}{2}\nabla\varphi(x,t)\cdot\nabla\varphi(x,t), \tag{V.38}$$

$$\varphi(x,t_1)=-\log\frac{\tilde{\rho}(x,t_1)}{\rho(x,t_1)}. \tag{V.39}$$

Then $(J+\Lambda^{\varphi})(\tilde{x},\gamma^{o}(x))=E\{\varphi(\tilde{x}(t_2),t_2)\}$ is constant over $\mathcal{X}_{\tilde{\rho}_2}$ since such processes have the same marginal density at time t_2. Hence, any $x\in\mathcal{X}_{\tilde{\rho}_2}$ solves the unconstrained minimization of $J+\Lambda^{\varphi}$. To solve the original constrained problem, we need to find $\tilde{x}\in\mathcal{X}_{\tilde{\rho}_2}$ that has backward drift given by (V.36). In order to do that, we first proceed to find the solution of (V.38)–(V.39). Define $\tilde{\rho}(x,t):=\exp[-\varphi(x,t)]\rho(x,t)$. Then, if φ satisfies (V.38), using the Fokker–Planck equation satisfied by ρ, we get

$$\frac{\partial\tilde{\rho}}{\partial t}=\exp[-\varphi]\left(-\frac{\partial\varphi}{\partial t}\rho+\frac{\partial\rho}{\partial t}\right)$$
$$=\left(b_-\cdot\nabla\varphi-\frac{\sigma^2}{2}\Delta\varphi+\frac{\sigma^2}{2}\nabla\varphi\cdot\nabla\varphi\right)\tilde{\rho}-\exp[-\varphi]\nabla\cdot(b_+\rho)+\exp[-\varphi]\frac{\sigma^2}{2}\Delta\rho$$
$$=\frac{\sigma^2}{2}\Delta\tilde{\rho}+b_+\cdot\nabla\varphi\tilde{\rho}-\exp[-\varphi]\nabla\rho\cdot b_+-\exp[-\varphi]\rho\nabla\cdot b_+=-\nabla\cdot(\tilde{\rho}b_+)+\frac{\sigma^2}{2}\Delta\tilde{\rho}.$$

We conclude that if $\tilde{\rho}$ is the solution of the Fokker–Planck equation (V.34) on $[t_1,t_2]$ with an initial condition at time t_1 given by $\tilde{\rho}(x,t_1)$, then $\varphi:=-\log(\tilde{\rho}/\rho)$ solves the initial value problem (V.38)–(V.39). Thus, we have the following result.

Theorem V.2: *Let $\tilde{\rho}$ be the solution of the Fokker–Planck equation (V.34) on $[t_1,t_2]$ with initial condition given by $\tilde{\rho}(x,t_1)$. Then $\varphi:=-\log(\tilde{\rho}/\rho)$ solves the initial value problem (V.38)–(V.39). Suppose that φ satisfies (V.35), and that $\tilde{\rho}_2(x)=\tilde{\rho}(x,t_2)$. Then the stochastic process $\tilde{x}\in\mathcal{X}_{\tilde{\rho}_2}$ having backward drift field $\tilde{b}_-(x,t)=b_-(x,t)-\sigma^2\nabla\log(\tilde{\rho}/\rho)(x,t)=b_+-\sigma^2\nabla\log\tilde{\rho}(x,t)$ solves the constrained minimization problem.*

In view of (II.16), we see that the solution process has forward drift $b_+(\cdot)$, and therefore coincides with the solution of the previous variational problem. Consider the same problem on the interval $[t_1,t_3]$, where $t_3>t_2$. If we impose the density $\tilde{\rho}(x,t_3)$ at the final time, the solution process coincides with the previous solution process up to time t_2. This may be viewed as a form of coherence with respect to the terminal time. It is also important to observe that the new process $\{\tilde{x}(t);t_1\leq t\leq t_2\}$ has the same forward drift of the reference process $\{x(t);t_1\leq t\leq t_2\}$, but a *different backward drift*. Hence, while the forward transition probabilities have been preserved, *the reverse-time transition probabilities have changed*. Thus, we see that it is impossible, even in principle, to estimate the reverse-time transition probabilities by repeated measurement. In Refs. 14, 7, Nelson and Guerra regard as a serious drawback of stochastic mechanics the fact that transition probabilities of the Nelson process are not open to experimental verification if we accept that transition probabilities are associated to a definite quantum state. We shall come back to this crucial point in the next section.

VI. A STOCHASTIC DERIVATION OF WAVE FUNCTION COLLAPSE

In Sec. IV, we have seen that the Schrödinger equation is obtained through a simple exponential transformation from the Hamilton–Jacobi equation (IV.28) of an appropriate stochastic variational principle. Suppose now that a position measurement of the quantum system is made at time t_1, and we ask the following: What should be the new stochastic process on $[t_1,t_2]$? First of all, we consider the situation without measurement up to time t_2. In this case, the variational principle of Sec. IV would have as a solution the Nelson process $\{x(t);t_0\leqslant t\leqslant t_2\}$ extended up to time t_2 with quantum drift $v_q(t)=(\hbar/im)\,\nabla\log\psi(x(t),t)$, where $\{\psi(x,t):t_0\leqslant t\leqslant t_2\}$ is the solution of the Schrödinger equation (IV.32). The Nelson process $\{x(t);t_1\leqslant t\leqslant t_2\}$ will play the role of a ''reference process.'' Suppose that the measurement at time t_1 yields the new probability density $\tilde{\rho}(x,t_1)$. For instance, if we assume that the measurement at time t_1 only gives the information that x lies in a certain subset D of the configuration space of the system, the density $\tilde{\rho}(x,t_1)$ just after the measurement is given, according to Bayes' theorem, by

$$\tilde{\rho}(x,t_1)=\frac{\chi_D(x)\rho(x,t_1)}{\int_D\rho(x',t_1)dx'},$$

where $\rho(x,t_1)$ is the probability density of the Nelson reference process right before the measurement is made. We need now to find an appropriate variational mechanism that, employing the Nelson reference process and the probability density $\tilde{\rho}(x,t_1)$, produces the new process $\{\tilde{x}(t);t_1\leqslant t\leqslant t_2\}$. It is apparent that the variational mechanism of the previous section is not suitable here. Indeed, as observed before, that mechanism preserves completely the *forward* drift and transition probabilities, but changes, possibly in a dramatic way, the backward drift and transition probabilities. This is not acceptable in stochastic mechanics, were forward and backward drifts and transition probabilities *must always be granted the same status*. In other words, the time reversibility of the theory must be reflected also by the theory of measurement. On the other hand, preserving both drifts, or equivalently both transition probabilities, amounts to preserving the process $\{x(t);t_0\leqslant t\leqslant t_2\}$, which is impossible since the probability density at time t_1 has changed. Thus, we need to find a variational mechanism that *changes both drifts as little as possible, given the new density at time* t_1. It should be apparent that, at this point, the time-symmetric kinematics of Sec. III is called for. Given that kinematics, and by analogy with the variational principle of the previous section, we are then led to the following formulation.

In the notation of Sec. IV, we consider the problem of extremizing on $(\tilde{x},\tilde{v}_q)\in(\mathcal{X}_{\tilde{\rho}_2}\times\mathcal{V})$ the functional

$$J(\tilde{x},\tilde{v}_q):=E\left\{\int_{t_1}^{t_2}\frac{mi}{2\hbar}(v_q(\tilde{x}(t),t)-\tilde{v}_q(t))\cdot(v_q(\tilde{x}(t),t)-\tilde{v}_q(t))\,dt+\frac{1}{2}\log\frac{\tilde{\rho}(\tilde{x}(t_1),t_1)}{\rho(\tilde{x}(t_1),t_1)}\right\},\tag{VI.40}$$

subject to the constraint that

$$\tilde{x}\ \text{ has quantum drift (velocity) }\ \tilde{v}_q.\tag{VI.41}$$

Here $v_q(x,t)=(\hbar/im)\,\nabla\log\psi(x,t)$ is the quantum drift field of the Nelson reference process, and $\mathcal{X}_{\tilde{\rho}_2}$ is the family of all finite-energy, $\mathbb{R}^n$-valued diffusions on $[t_1,t_2]$ with diffusion coefficient

$I_n(\hbar/m)$, and having probability density $\tilde{\rho}_2$ at time t_2. The structure of the functional is quite similar to the one of the previous section. Here, $\hbar/mi$ replaces σ^2 in view of formula (IV.31). The $\frac{1}{2}$ in the boundary term is justified by the following relation; see (IV.27),

$$\frac{mi}{2\hbar}(v_q(x,t)-\tilde{v}_q(t))\cdot(v_q(x,t)-\tilde{v}_q(t))$$
$$=\frac{m}{4\hbar}[(b_+(x,t)-\tilde{b}_+(t))+i(b_-(x,t)-\tilde{b}_-(t))]\cdot[(b_+(x,t)-\tilde{b}_+(t))+i(b_-(x,t)-\tilde{b}_-(t))],$$

which shows that a $\frac{1}{4}$ appears on the right-hand side. To solve this variational problem, we employ the same strategy as in the previous section. Let $\varphi(x,t)$ be a complex-valued function of class $C^{2,1}$ defined on $\mathbb{R}^n\times[t_1,t_2]$, and satisfying the technical condition

$$E\left\{\int_{t_1}^{t_2}\nabla\varphi(x(t),t)\cdot\overline{\nabla\varphi(x(t),t)}dt\right\}<\infty,\quad \forall x\in\mathcal{X}_{\rho_2}. \tag{VI.42}$$

Corresponding to such a φ, we introduce the functional

$$\Lambda^{\varphi}(\tilde{x},\tilde{v}_q):=E\left\{\varphi(\tilde{x}(t_2),t_2)-\varphi(\tilde{x}(t_1),t_1)\right.$$
$$\left.+\int_{t_1}^{t_2}\left[-\frac{\partial\varphi}{\partial t}(\tilde{x}(t),t)-\tilde{v}_q(t)\cdot\nabla\varphi(\tilde{x}(t),t)+\frac{i\hbar}{2m}\Delta\varphi(\tilde{x}(t),t)\right]dt\right\}.$$

In view of (III.24), and of property (VI.42), we see that $\Lambda^{\varphi}(\tilde{x},\tilde{v}_q)=0$ whenever the pair $(\tilde{x},\tilde{v}_q)$ satisfies the constraint. Thus, it is a *Lagrange functional* for the problem. Consider next the *unconstrained* extremization of the functional $J+\Lambda^{\varphi}$. For a fixed $\tilde{x}\in\mathcal{X}_{\tilde{\rho}_2}$, and a fixed time $t\in[t_1,t_2]$, we consider the *pointwise* extremization of the integrand of $J+\Lambda^{\varphi}$ with respect to $\tilde{v}_q$,

$$\text{extremize}_{\tilde{v}\in C^n}\left\{\frac{mi}{2\hbar}(v_q(\tilde{x}(t),t)-\tilde{v})\cdot(v_q(\tilde{x}(t),t)-\tilde{v})-\tilde{v}\cdot\nabla\varphi(\tilde{x}(t),t)\right\}.$$

We get

$$\tilde{v}_q^o(\tilde{x})(t)=v_q(\tilde{x}(t),t)+\frac{\hbar}{mi}\nabla\varphi(\tilde{x}(t),t). \tag{VI.43}$$

Substituting back expression (VI.43) into $J+\Lambda^{\varphi}$, we get the following functional of $\tilde{x}$:

$$(J+\Lambda^{\varphi})(\tilde{x},\tilde{v}_q^o(x)):=E\left\{\varphi(\tilde{x}(t_2),t_2)-\varphi(\tilde{x}(t_1),t_1)+\int_{t_1}^{t_2}\left[\frac{i\hbar}{2m}\nabla\varphi(\tilde{x}(t),t)\cdot\nabla\varphi(\tilde{x}(t),t)-\frac{\partial\varphi}{\partial t}(\tilde{x}(t),t)\right.\right.$$
$$\left.\left.-v_q(\tilde{x}(t),t)\cdot\nabla\varphi(\tilde{x}(t),t)+\frac{i\hbar}{2m}\Delta\varphi(\tilde{x}(t),t)\right]dt\right\}. \tag{VI.44}$$

We seek next to choose the function φ so that the functional $(J+\Lambda^{\varphi})(\tilde{x},\tilde{v}_q^o(x))$ becomes constant over $\mathcal{X}_{\tilde{\rho}_2}$. Suppose φ solves on $[t_1,t_2]$ the initial value problem,

$$\frac{\partial\varphi}{\partial t}+v_q(x,t)\cdot\nabla\varphi(x,t)-\frac{i\hbar}{2m}\Delta\varphi(x,t)=\frac{i\hbar}{2m}\nabla\varphi(x,t)\cdot\nabla\varphi(x,t), \tag{VI.45}$$

$$\varphi(x,t_1)=\frac{1}{2}\log\frac{\tilde{\rho}(x,t_1)}{\rho(x,t_1)}. \tag{VI.46}$$

Then $(J+\Lambda^{\varphi})(\tilde{x},\tilde{v}_q^o(x))=E\{\varphi(\tilde{x}(t_2),t_2)\}$ is constant over $\mathcal{X}_{\tilde{\rho}_2}$ since such processes have the same marginal density at time t_2. Hence, any $x\in\mathcal{X}_{\tilde{\rho}_2}$ solves the unconstrained extremization of $J+\Lambda^{\varphi}$. To solve the original constrained extremization problem, we need to find the $\tilde{x}\in\mathcal{X}_{\tilde{\rho}_2}$ that has quantum drift given by (VI.43). In order to do that, we first proceed to find the solution of (VI.45)–(VI.46). Write $\psi(x,t_1)=\rho(x,t_1)^{1/2}\exp[(i/\hbar)S(x,t_1)]$, and define $\tilde{\psi}(x,t):=\exp[\varphi(x,t)]\psi(x,t)$. Then, if φ satisfies (VI.45), using the Schrödinger equation (IV.32) satisfied by ψ, we get

$$\frac{\partial\tilde{\psi}}{\partial t}=\exp[\varphi]\left(\frac{\partial\varphi}{\partial t}\psi+\frac{\partial\psi}{\partial t}\right)$$
$$=\frac{i}{\hbar}V(x)\tilde{\psi}+\frac{i\hbar}{2m}\exp\varphi(\Delta\psi+2\nabla\psi\cdot\nabla\varphi+\nabla\varphi\cdot\nabla\varphi\psi+\Delta\varphi\psi)=\frac{i\hbar}{2m}\Delta\tilde{\psi}-\frac{i}{\hbar}V(x)\tilde{\psi}.$$

Observing that $\tilde{\psi}(x,t_1)=\tilde{\rho}(x,t_1)^{1/2}\exp[(i/\hbar)S(x,t_1)]$, we conclude that if $\tilde{\psi}$ is the solution of the Schrödinger equation (IV.32) on $[t_1,t_2]$ with initial condition at time t_1 given by $\tilde{\rho}(x,t_1)^{1/2}\exp[(i/\hbar)S(x,t_1)]$; then $\varphi:=\log(\tilde{\psi}/\psi)$ solves the initial value problem (VI.45)–(VI.46). Thus, we get the following result.

Theorem VI.1: *Suppose that $\tilde{\psi}$ is the solution of the Schrödinger equation (IV.32) on $[t_1,t_2]$ with initial condition at time t_1 given by $\tilde{\rho}(x,t_1)^{1/2}\exp[(i/\hbar)S(x,t_1)]$. Then $\varphi:=\log(\tilde{\psi}/\psi)$ solves the initial value problem (VI.45)–(VI.46). Suppose that φ satisfies (VI.42), and that $\tilde{\rho}_2(x)=|\tilde{\psi}(x,t_2)|^2$. Then the stochastic process $\tilde{x}\in\mathcal{X}_{\tilde{\rho}_2}$ having quantum drift $(\hbar/mi)\nabla\log\tilde{\psi}(\tilde{x}(t),t)$ solves the constrained extremization problem.*

Thus, by a purely probabilistic argument, we have shown that the new process after the measurement at time t_1 is associated to another solution $\tilde{\psi}$ of the same Schrödinger equation (IV.32). The association is precisely as before, namely, the quantum drift is proportional to the gradient of the logarithm of $\tilde{\psi}$. In other words, the new process is just the Nelson process associated to the solution $\{\tilde{\psi}(x,t);t_1\leq t\leq t_2\}$. It is important to observe that the new wave function has the same phase at time t_1 as the old one before measurement. This agrees with standard quantum mechanics when it is assumed that immediate repetition of the measurement yields the same result and does not change the wave function except for an arbitrary phase factor; see, e.g., Refs. 26, 27. Here, however, no further assumption is needed: *The invariance of the phase follows from the variational principle*. This is a crucial point. Indeed, if we assume the invariance of the phase after a position measurement in stochastic mechanics, then the variational principle of Sec. IV suffices to produce the new Nelson process (associated to the solution $\{\tilde{\psi}(x,t)\}$ of the Schrödinger equation). Also notice that the solution process possesses the same coherence property with respect to the time interval as the solution process of the previous section.

VII. DISCUSSION

In this paper we have shown that, in the frame of Nelson's stochastic mechanics, the wave function reduction does not need to be *postulated*, but may be *derived* from the standard rules of probability (Bayes' theorem) and a stochastic variational principle of transparent significance. It seems to us that this result lends support to the point of view of Blanchard, Golin, and Serva in Ref. 28, where it was shown that some apparent paradoxes of stochastic mechanics related to repeated measurements could be removed by introducing an appropriate new process after each measurement. The new process, indeed, is the Nelson process associated to the new solution $\tilde{\psi}$ of the Schrödinger equation. A general comparison between standard quantum mechanics and stochastic mechanics is beyond the aims of this paper, and anyway beyond the knowledge and the understanding of the present author. We refer the reader to Refs. 4, 14, as well as to a series of recent papers by Guerra,[7,29] for a thorough and deep analysis on the possibility of regarding Nelson's stochastic mechanics as a complete physical theory.

Nevertheless, it seems legitimate to us to stress that stochastic mechanics, including the elements of a theory of measurement outlined in Ref. 28 and here, can simply be based on the hypothesis of universal Brownian motion and on stochastic variational principles. Thus, stochastic mechanics appears as a generalization of classical mechanics whose foundations are completely independent from standard quantum mechanics. Moreover, this theory is now capable of providing a transparent probabilistic derivation of the two most mysterious features of standard quantum mechanics, namely the uncertainty principle and the wave function collapse.

ACKNOWLEDGMENTS

We wish to thank all the participants in the Fall 98 seminar on''Control of quantum systems,'' and expecially P. Marchetti for stimulating me through very lively discussions to (hopefully) improve my understanding on the measurement problem in stochastic mechanics.

[1] I. Fenyes, Z. Phys. **132**, 81 (1952).
[2] E. Nelson, *Dynamical Theories of Brownian Motion* (Princeton University Press, Princeton, 1967).
[3] F. Guerra, Phys. Rep. **77**, 263 (1981).
[4] E. Nelson, *Quantum Fluctuations* (Princeton University Press, Princeton, 1985).
[5] Ph. Blanchard, Ph. Combe, and W. Zheng, *Math. and Physical Aspects of Stochastic Mechanics*, Lecture Notes in Physics (Springer-Verlag, New York, 1987), Vol. 281.
[6] F. Calogero, Phys. Lett. A **228**, 335 (1997).
[7] F. Guerra, in *The Foundation of Quantum Mechanics* (Kluwer, Amsterdam, 1995).
[8] M. Pavon, J. Math. Phys. **36**, 6774 (1995).
[9] M. Pavon, Phys. Lett. A **209**, 143 (1995).
[10] M. Pavon, J. Math. Phys. **37**, 3375 (1996).
[11] L. de la Pena-Auerbach and M. Cetto, Phys. Lett. A **39**, 65 (1972).
[12] D. De Falco, S. De Martino, and S. De Siena, Phys. Rev. Lett. **49**, 181 (1982).
[13] S. Golin, J. Math. Phys. **26**, 2781 (1985).
[14] E. Nelson, in *Stochastic Processes in Classical and Quantum Systems*, Lecture Notes in Physics, edited by S. Albeverio, G. Casati, and D. Merlini (Springer-Verlag, New York, 1986), Vol. 262, pp. 438–469.
[15] E. Nelson, in *Ècole d'Ètè de Probabilitès de Saint-Flour XV–XVII*, Lecture Notes in Mathematics, edited by P. L. Hennequin (Springer-Verlag, New York, 1988), Vol. 1362, pp. 428–450.
[16] H. Föllmer, in *Stochastic Processes — Mathematics and Physics*, Lecture Notes in Mathematics (Springer-Verlag, New York, 1986), Vol. 1158, pp. 119–129.
[17] H. Föllmer, in Ref. 15, pp. 102–203.
[18] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus* (Springer-Verlag, New York, 1988).
[19] M. Pavon, in *Proceedings of the Conference on Stochastic Differential and Difference Equations*, edited by I. Csiszár and Gy. Michaletzky (Birkhäuser, Boston, 1997), pp. 150–155.
[20] F. Guerra and L. Morato, Phys. Rev. D **27**, 1774 (1983).
[21] E. Carlen, Commun. Math. Phys. **94**, 293 (1984).
[22] E. Schrödinger, Sitzungsber. K. Preuss. Akad. Wiss., Phys. Math. K1. 144 (1931).
[23] F. Guerra and M. Pavon, in *Analysis and Control of Nonlinear Systems*, edited by C. I. Byrnes, C. F. Martin, and R. E. Saeks (Elsevier Science, Amsterdam, 1988), pp. 571–578.
[24] M. Pavon and A. Wakolbinger, in *Modeling, Estimation and Control of Systems With Uncertainty*, edited by G. B. Di Masi, A. Gombani, and A. Kurzhanski (Birkhäuser, Boston, 1991), pp. 334–348.
[25] P. Kosmol and M. Pavon, Syst. Control Lett. **24**, 215 (1995).
[26] J. S. Bell and M. Nauenberg, *The Moral Aspects of Quantum Mechanics*, Preludes in Theoretical Physics, edited by A. De-Shalit, H. Feshbach, and L. van Hove (North-Holland, Amsterdam, 1966), pp. 279–286.
[27] J. B. Keller, Am. J. Phys. **58**, 768 (1990).
[28] Ph. Blanchard, S. Golin, and M. Serva, Phys. Rev. D **34**, 3732 (1986).
[29] F. Guerra, ''The problem of the physical interpretation of Nelson stochastic mechanics as a model for quantum mechanics,'' preprint, March 1997.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Estimates for the spectral shift function of the polyharmonic operator

Alexander Pushnitski[a)]
Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire LE11 3TU, United Kingdom

(Received 19 May 1999; accepted for publication 10 August 1999)

The Lifshits–Krein spectral shift function is considered for the pair of operators $H_0=(-\Delta)^l$, $l>0$ and $H=H_0+V$ in $L_2(\mathbb{R}^d)$, $d\geqslant 1$; here V is a multiplication operator. The estimates for this spectral shift function $\xi(\lambda;H,H_0)$ are obtained in terms of the spectral parameter $\lambda>0$ and the integral norms of V. These estimates are in a good agreement with the ones predicted by the classical phase space volume considerations.
[S0022-2488(99)02311-7]

I. INTRODUCTION

The main object of study of this paper is the Lifshits–Krein *spectral shift function* (SSF). For an exposition of the SSF theory, see, e.g., Ref. 1 or 2. For a general pair of self-adjoint operators H_1, H_2 in a Hilbert space $\mathcal{H}$, satisfying some trace class condition (see Sec. II C), the SSF $\xi(\lambda;H_2,H_1)$ appears in connection with the *trace formula:*

$$\mathrm{Tr}(\psi(H_2)-\psi(H_1))=\int_{-\infty}^{\infty}\xi(\lambda;H_2,H_1)\psi'(\lambda)d\lambda,\quad \psi\in C_0^\infty(\mathbb{R}). \tag{1.1}$$

Besides, the SSF is related to the scattering matrix $S(\lambda)$ for the pair H_1, H_2 by the *Birman–Krein formula*:

$$\det S(\lambda)=e^{-2\pi i\xi(\lambda;H_2,H_1)},$$

for almost every λ on the absolutely continuous spectrum of H_1. This formula allows one to interpret the SSF as the scattering phase. Moreover, sometimes it is treated as the definition of the SSF. See Ref. 1 for references and a discussion.

Let $H_0=(-\Delta)^l$, $l>0$ in $L_2(\mathbb{R}^d)$, $d\geqslant 1$, and let $V=V(x)$, $x\in\mathbb{R}^d$, be a (real-valued) perturbation potential, which decays sufficiently fast as $|x|\to\infty$. In this paper we obtain bounds on $\xi(\lambda;H_0+V,H_0)$ in terms of λ and integral norms of V. The most interesting case is $l=1$ (Schrödinger operator). Nevertheless, the technique of this paper allows us to obtain bounds on the SSF in a uniform way for all $l\in(0,\infty)$ (and all dimensions d), and thus we consider the problem in its natural generality. The main result of this paper (Theorem 2.2) is formulated in Sec. II and its corollaries—in Sec. III. Our estimates are very close to the ones predicted by the semiclassical intuition—see the discussion in Sec. III. Our results extend the estimates of Ref. 3 and are closely related to the results of Ref. 4—see the discussion at the end of Sec. III. Our technique is based on the new representation for the SSF obtained in Ref. 5.

[a)]Electronic mail: a.b.pushnitski@lboro.ac.uk

II. PRELIMINARIES: STATEMENT OF THE MAIN RESULT

A. Notation

For a closable linear operator T in a Hilbert space $\mathcal{H}$, by $\overline{T}$ we denote the closure of T. For a self-adjoint operator A, the symbols $\sigma(A)$, $\rho(A)$ denote its spectrum and resolvent set; $R(z,A) = (A-zI)^{-1}$ [for $z\in\rho(A)$] and $E_A(\delta)$ is the spectral projection associated to a Borel set $\delta\subset\mathbb{R}$. By $\mathbf{S}_\infty(\mathcal{H})$ we denote the space of all compact operators in $\mathcal{H}$. For $T=T^*\in\mathbf{S}_\infty(\mathcal{H})$ we introduce the counting functions of the spectrum by $n_\pm(s,T) := \operatorname{rank} E_{\pm T}((s,+\infty))$, $s>0$. Note that (see, e.g., Ref. 6)

$$n_\pm(s_1+s_2,T_1+T_2)\leqslant n_\pm(s_1,T_1)+n_\pm(s_2,T_2), \quad s_1,s_2>0. \tag{2.1}$$

For $p\geqslant 1$ the Neumann–Schatten classes $\mathbf{S}_p$ are defined in a usual way:

$$T\in\mathbf{S}_p \quad \text{if } \|T\|_{\mathbf{S}_p}^p := \sum_n s_n^p(T)<\infty,$$

where $\{s_n(T)\}$ is the sequence of singular numbers of T.

An integral without the domain of integration explicitly specified implies integration over $\mathbb{R}^d$. Formulas and statements with double indices ($\pm$ and $\mp$) should be read as pairs of statements, in one of which all the indices take upper values and in another—the lower ones. By $C(d), C(l)$, etc. (possibly with sub- and superscripts) we denote various constants that depend only on d, l, etc. and whose particular values are of no importance. A constant that first appears in formula $(i.j)$ is denoted by $C_{i.j}$.

We remind the reader of the definition of the lattice space $l_1(L_r)\subset L_r(\mathbb{R}^d), r\geqslant 1$:

$$u\in l_1(L_r) \quad \text{if } \|u\|_{l_1(L_r)} := \sum_{j\in\mathbb{Z}^d}\left(\int_{\mathbf{Q}^d+j}|u|^r\,dx\right)^{1/r}<\infty, \quad \mathbf{Q}^d=(0,1)^d\subset\mathbb{R}^d.$$

Everywhere $\mathcal{H}=L_2(\mathbb{R}^d)$, $d\geqslant 1$ and $H_0=(-\Delta)^l$, $l>0$; $\varkappa = d/(2l)$. We shall need a notation for a logarithmic weight function. Namely, for $x\in\mathbb{R}^d$ and $\gamma>0$ let

$$F_\gamma(x)=1+(\log_+|x|)^\gamma. \tag{2.2}$$

B. Assumptions on *V*

Below by V we denote both a function $V:\mathbb{R}^d\to\mathbb{R}$ and the (self-adjoint) operator of multiplication by $V(x)$ in $L_2(\mathbb{R}^d)$. In various places we shall use some of the following assumptions on V:

$$V \text{ is } H_0\text{-form compact;} \tag{2.3}$$

$$V\in L_1(\mathbb{R}^d); \tag{2.4}$$

$$V\in L_\varkappa(\mathbb{R}^d); \tag{2.5}$$

$$\int |V(x)|F_\gamma(x)\,dx<\infty, \quad \gamma>2. \tag{2.6}$$

Note that for $\varkappa<1$ (2.3) follows from (2.4), and for $\varkappa>1$—from (2.5). Clearly, under the assumption (2.3) the operators H_0+V, H_0+V_+, H_0-V_- are well defined via the corresponding quadratic forms.

The following assumption will appear only for $\varkappa\geqslant 2$:

$$\text{if } \varkappa\geqslant 2, \quad \text{then } V\in l_1(L_2). \tag{2.7}$$

C. Existence of the SSF

Let $H_1,\ H_2$ be a pair of self-adjoint operators in a Hilbert space $\mathcal{H}$, such that

$$H_2 - H_1 \in \mathbf{S}_1(\mathcal{H}). \tag{2.8}$$

Then the SSF $\xi(\lambda;H_2,H_1)$ exists and is given by the *Krein formula*,[7]

$$\xi(\lambda;H_2,H_1) = \frac{1}{\pi} \lim_{\epsilon\to+0} \arg \Delta_{H_2/H_1}(\lambda + i\epsilon), \quad \text{a.e. } \lambda \in \mathbb{R}, \tag{2.9}$$

where $\Delta_{H_2/H_1}(z) = \det(H_2 - zI)(H_1 - zI)^{-1}$ and the branch of the argument is fixed by

$$\arg \Delta_{H_2/H_1}(z) \to 0 \quad \text{as } \operatorname{Im} z \to +\infty.$$

The SSF obeys the monotonicity property:[7]

$$\pm((H_2 - H_1)\chi,\chi) \geqslant 0 \quad \forall \chi \in \mathcal{H} \Rightarrow \pm \xi(\lambda;H_2,H_1) \geqslant 0. \tag{2.10}$$

For a triple $H_0,\ H_1,\ H_2$ of operators, such that $H_1 - H_0 \in \mathbf{S}_1$ and $H_2 - H_0 \in \mathbf{S}_1$, one has

$$\xi(\lambda;H_2,H_0) = \xi(\lambda;H_2;H_1) + \xi(\lambda;H_1,H_0). \tag{2.11}$$

In applications, instead of (2.8), it is usually possible to check the inclusion $f(H_2) - f(H_1) \in \mathbf{S}_1$, where f is some monotone smooth enough function. In this case one can first take $\tilde{H}_1 = f(H_1),\ \tilde{H}_2 = f(H_2)$ and define $\xi(\lambda;\tilde{H}_2,\tilde{H}_1)$ according to (2.9), and then put

$$\xi(\lambda;H_2,H_1) := \operatorname{sign} f' \cdot \xi(f(\lambda);f(H_2),f(H_1)). \tag{2.12}$$

Thus defined, $\xi(\lambda;H_2,H_1)$ still obeys the trace formula (1.1). See Refs. 1 and 2 for details.

Proposition 2.1: (i) *Let $\varkappa < 2$ and let V obey (2.3), (2.4). Then, for any $\lambda_0 < \inf(\sigma(H_0) \cup \sigma(H_0 + V))$, the relation*

$$R(\lambda_0,H_0) - R(\lambda_0,H_0 + V) \in \mathbf{S}_1 \tag{2.13}$$

holds. Thus, the spectral shift functions $\xi(\lambda;H_0 + V,H_0)$, $\xi(\lambda;H_0 + V_+,H_0)$, $\xi(\lambda;H_0 - V_-,H_0)$ are well defined by (2.12) with $f(\lambda) = (\lambda - \lambda_0)^{-1}$. The following inequalities hold:

$$\xi(\lambda;H_0 - V_-,H_0) \leqslant 0 \leqslant \xi(\lambda;H_0 + V_+,H_0), \tag{2.14}$$

$$\xi(\lambda;H_0 - V_-,H_0) \leqslant \xi(\lambda;H_0 + V,H_0) \leqslant \xi(\lambda;H_0 + V_+,H_0). \tag{2.15}$$

(ii) *Let $\varkappa \geqslant 2$ and let V obey (2.3), (2.7). Then, for any $\lambda_0 < \inf(\sigma(H_0) \cup \sigma(H_0 + V))$ with a large enough absolute value, and for any integer $k > \varkappa - \frac{1}{2}$, the relation*

$$R^k(\lambda_0,H_0) - R^k(\lambda_0,H_0 + V) \in \mathbf{S}_1 \tag{2.16}$$

holds. Thus, the spectral shift functions $\xi(\lambda;H_0 + V,H_0)$, $\xi(\lambda;H_0 + V_+,H_0)$, $\xi(\lambda;H_0 - V_-,H_0)$ are well defined by (2.12) with $f(\lambda) = (\lambda - \lambda_0)^{-k}$. The inequalities (2.14), (2.15) hold.

Similar statements appeared in the literature in many different versions in connection with the trace class scattering theory. Nevertheless, for the sake of completeness, we give the proof of Proposition 2.1 at the end of this section.

The inequalities (2.15) reduce the problem of estimating the SSF to the case of perturbations of a definite sign. Thus, below we shall always assume that $V \geqslant 0$ and consider the pair of functions $\xi(\lambda;H_0 \pm V,H_0)$.

Finally, note that, since $\sigma(H_0) = [0,\infty)$, one has

$$\xi(-\lambda;H_0-V,H_0)=-\operatorname{rank} E_{H_0-V}((-\infty,-\lambda)),\quad \lambda>0. \tag{2.17}$$

D. Estimates for the SSF

The main result of this paper is the following.

Theorem 2.2: *Let $V>0$. Under the assumptions (2.3), (2.6), (2.7), the following estimates hold for $\lambda>0$ and $\gamma>2$:*

$$\xi(\lambda;H_0+V,H_0)\leqslant C(d,l,\gamma)\lambda^{\varkappa-1}\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}+C(d,l)\lambda^{\varkappa-1}(\log_+\lambda)\|V\|_{L_1}, \tag{2.18}$$

$$|\xi(\lambda;H_0-V,H_0)|\leqslant C(d,l,\gamma)\lambda^{\varkappa-1}\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}+C(d,l)\lambda^{\varkappa-1}(\log_+\lambda)\|V\|_{L_1}$$
$$+|\xi(-\lambda;H_0-6V,H_0)|. \tag{2.19}$$

The proof is given in Secs. IV–VIII. The factor 6 in (2.19) is chosen this way in order to simplify the constants appearing in the proof. Actually, this factor can be replaced by any number greater than 1; at the same time, the constants $C(d,l,\gamma)$ and $C(d,l)$ may have to be increased. Using bounds for $|\xi(-\lambda;H_0-6V,H_0)|$ [which, by (2.17), reduce to the bounds on the number of eigenvalues of H_0-6V], one can estimate the SSF $\xi(\lambda;H_0-V,H_0)$ entirely in terms of λ and integral norms of V. This will be done in Sec. III. In Sec. III we also discuss Theorem 2.2 and related results.

E. Proof of Proposition 2.1

(i) By (2.4), one has $\sqrt{|V|}R(\lambda_0,H_0)\in\mathbf{S}_2$. Therefore, (2.13) follows from the identity

$$R(\lambda_0,H_0+V)=R(\lambda_0,H_0)-(\operatorname{sign}V\sqrt{|V|}R(\lambda_0,H_0))^*(I+\sqrt{|V|}R(\lambda_0,H_0)\sqrt{|V|}\operatorname{sign}V)^{-1}$$
$$\times(\sqrt{|V|}R_0(\lambda_0,H_0)).$$

Clearly,

$$H_0-V_-\leqslant H_0\leqslant H_0+V_+,\quad H_0-V_-\leqslant H_0+V\leqslant H_0+V_+,$$

in the quadratic form sense. Therefore,

$$R(\lambda_0,H_0-V_-)\geqslant R(\lambda_0,H_0)\geqslant R(\lambda_0,H_0+V_+), \tag{2.20}$$

$$R(\lambda_0,H_0-V_-)\geqslant R(\lambda_0,H_0+V)\geqslant R(\lambda_0,H_0+V_+). \tag{2.21}$$

From here, by (2.10) and (2.11), we get

$$\xi((\lambda-\lambda_0)^{-1};R(\lambda_0,H_0-V_-),R(\lambda_0,H_0))\geqslant 0\geqslant\xi((\lambda-\lambda_0)^{-1};R(\lambda_0,H_0+V_+),R(\lambda_0,H_0)),$$

$$\xi((\lambda-\lambda_0)^{-1};R(\lambda_0,H_0-V_-),R(\lambda_0,H_0))\geqslant\xi((\lambda-\lambda_0)^{-1};R(\lambda_0,H_0+V),R(\lambda_0,H_0))$$
$$\geqslant\xi((\lambda-\lambda_0)^{-1};R(\lambda_0,H_0+V_+),R(\lambda_0,H_0)).$$

From here, by the definition (2.12), the relations (2.14), (2.15) follow.

(ii) By (2.7), one has

$$VR^{k+1/2}(\lambda_0,H_0)\in\mathbf{S}_1,\quad k>\varkappa-\tfrac{1}{2}$$

(see, e.g., Ref. 8, Theorem 11.1). Therefore, by Ref. 9, Theorem XI.12, the inclusion (2.16) holds. Due to the results of Ref. 10 (see also Ref. 2, Sec. 8.10), the inequalities (2.20), (2.21) together with (2.16), (2.11) imply

$$\xi((\lambda-\lambda_0)^{-k};R^k(\lambda_0,H_0-V_-),R^k(\lambda_0,H_0))\geqslant 0\geqslant \xi((\lambda-\lambda_0)^{-k};R^k(\lambda_0,H_0+V_+),R^k(\lambda_0,H_0)),$$

$$\xi((\lambda-\lambda_0)^{-k};R^k(\lambda_0,H_0-V_-),R^k(\lambda_0,H_0))\geqslant \xi((\lambda-\lambda_0)^{-k};R^k(\lambda_0,H_0+V),R^k(\lambda_0,H_0))$$
$$\geqslant \xi((\lambda-\lambda_0)^{-k};R^k(\lambda_0,H_0+V_+),R^k(\lambda_0,H_0)).$$

From here, by the definition (2.12), the relations (2.14), (2.15) follow. ■

III. COROLLARIES AND DISCUSSION

A. Semiclassical considerations

Let us consider a ''classical analog'' of the SSF, expressed in terms of the phase space volumes, corresponding to the systems with the Hamiltonians $h_0(p,x)=p^{2l}$ and $h_\pm(p,x)=p^{2l}\pm V(x)$, where $V\geqslant 0$. Let ω_d be the volume of a unit ball in $\mathbb{R}^d$; define

$$\xi^{\mathrm{cl}}(\lambda;H_0+V,H_0)=(2\pi)^{-d}\,\mathrm{vol}\{(p,x)\in\mathbb{R}^{2d}|h_0(p,x)<\lambda<h_+(p,x)\}$$
$$=(2\pi)^{-d}\omega_d\int_{\mathbb{R}^d}(\lambda^{\varkappa}-(\lambda-V(x))_+^{\varkappa})dx,\quad \lambda>0, \tag{3.1}$$

$$\xi^{\mathrm{cl}}(\lambda;H_0-V,H_0)=-(2\pi)^{-d}\,\mathrm{vol}\{(p,x)\in\mathbb{R}^{2d}|h_-(p,x)<\lambda<h_0(p,x)\}$$
$$=-(2\pi)^{-d}\omega_d\int_{\mathbb{R}^d}((\lambda+V(x))^{\varkappa}-\lambda^{\varkappa})dx,\quad \lambda>0. \tag{3.2}$$

It is well known that $\xi^{\mathrm{cl}}(\lambda;H_0\pm V,H_0)$ behaves in many respects like $\xi(\lambda;H_0\pm V,H_0)$; for example, it has the same asymptotics in most asymptotical regimes—see the review in Ref. 11 and references therein. The integrands in (3.1), (3.2) admit the following elementary bounds:

$$\lambda^{\varkappa}-(\lambda-V)_+^{\varkappa}\leqslant\max\{\varkappa,1\}\lambda^{\varkappa-1}V,\quad \varkappa>0; \tag{3.3}$$

$$(\lambda+V)^{\varkappa}-\lambda^{\varkappa}\leqslant\varkappa\lambda^{\varkappa-1}V,\quad \varkappa\leqslant 1; \tag{3.4}$$

$$(\lambda+V)^{\varkappa}-\lambda^{\varkappa}\leqslant C_1V^{\varkappa}+C_2\lambda^{\varkappa-1}V,\quad \varkappa>1. \tag{3.5}$$

It is easy to give concrete explicit values for C_1, C_2 in (3.5); for example, $C_1=2^{\varkappa}$, $C_2=2^{\varkappa}-1$. Substituting (3.3)–(3.5) into (3.1), (3.2), we get the following bounds for ξ^{cl}:

$$\xi^{\mathrm{cl}}(\lambda;H_0+V,H_0)\leqslant(2\pi)^{-d}\omega_d\max\{\varkappa,1\}\lambda^{\varkappa-1}\|V\|_{L_1},\quad \varkappa>0; \tag{3.6}$$

$$|\xi^{\mathrm{cl}}(\lambda;H_0-V,H_0)|\leqslant(2\pi)^{-d}\omega_d\varkappa\lambda^{\varkappa-1}\|V\|_{L_1},\quad \varkappa\leqslant 1; \tag{3.7}$$

$$|\xi^{\mathrm{cl}}(\lambda;H_0-V,H_0)|\leqslant C_1(2\pi)^{-d}\omega_d\|V\|_{L_\varkappa}^{\varkappa}+C_2(2\pi)^{-d}\omega_d\lambda^{\varkappa-1}\|V\|_{L_1},\quad \varkappa>1. \tag{3.8}$$

Estimates (3.6)–(3.8) are in good agreement with the asymptotics of ξ^{cl} for high energy and a large coupling constant:

$$\xi^{\mathrm{cl}}(\lambda;H_0\pm V,H_0)\sim\pm(2\pi)^{-d}\omega_d\varkappa\lambda^{\varkappa-1}\int V(x)dx,\quad \lambda\to\infty, \tag{3.9}$$

$$\xi^{\mathrm{cl}}(\lambda;H_0-gV,H_0)\sim-(2\pi)^{-d}\omega_d g^{\varkappa}\int V^{\varkappa}(x)dx,\quad g\to\infty. \tag{3.10}$$

We consider (3.6)–(3.8) as the model estimates. Clearly, (2.18) is in agreement with (3.6) up to the constants, the logarithmic weight F_γ and the term $\log_+ \lambda$. In order to compare (2.19) with (3.7), (3.8), one has to estimate the term $|\xi(-\lambda;H_0-6V,H_0)|$. This will be done below differently for $\varkappa>1$, $\varkappa<1$ and $\varkappa=1$.

B. The case $\varkappa>1$

In this case, we use the Cwikel–Lieb–Rozenblum bound:[12–14]

$$|\xi(-\lambda;H_0-V,H_0)|\leqslant C(d,l)\|V\|_{L_\varkappa}^{\varkappa}. \tag{3.11}$$

Substituting (3.11) into (2.19), we get the following.

Corollary 3.1: Let $\varkappa>1$. Assume the hypothesis of Theorem 2.2 and the inclusion (2.5); then, for any $\lambda>0$ and $\gamma>2$,

$$|\xi(\lambda;H_0-V,H_0)|\leqslant C(d,l,\gamma)\lambda^{\varkappa-1}\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}+C(d,l)\lambda^{\varkappa-1}(\log_+\lambda)\|V\|_{L_1}+C(d,l)\|V\|_{L_\varkappa}^{\varkappa}. \tag{3.12}$$

C. The case $\varkappa<1$

For $\varkappa<1$ let us use the Birman–Schwinger principle and estimate $|\xi(-\lambda;H_0-V,H_0)|$ (for $V\geqslant 0$) in the following way:

$$|\xi(-\lambda;H_0-V,H_0)|\leqslant \mathrm{Tr}(\sqrt{V}R(-\lambda,H_0)\sqrt{V})=C(d,l)\lambda^{\varkappa-1}\|V\|_{L_1}. \tag{3.13}$$

Substituting (3.13) into (2.19) and taking into account the obvious estimate $\|V\|_{L_1}\leqslant\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}$, we get the following corollary.

Corollary 3.2: Let $V\geqslant 0$ and $\varkappa<1$; assume (2.6). Then, for any $\lambda>0$ and $\gamma>2$,

$$|\xi(\lambda;H_0-V,H_0)|\leqslant C(d,l,\gamma)\lambda^{\varkappa-1}\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}+C(d,l)\lambda^{\varkappa-1}(\log_+\lambda)\|V\|_{L_1}. \tag{3.14}$$

D. The case $\varkappa=1$

Now

$$\xi^{\mathrm{cl}}(-\lambda;H_0-V,H_0)=-(2\pi)^{-d}\omega_d\int V(x)dx,\quad \forall\lambda\in\mathbb{R}.$$

Nevertheless, as it is well known, the ''naive'' estimate,

$$|\xi(-\lambda;H_0-V,H_0)|\leqslant C\|V\|_{L_1},\quad \lambda>0, \tag{3.15}$$

is wrong; see, e.g., Ref. 15 for the discussion. Instead, there are numerous estimates of $\xi(-\lambda;H_0-V,H_0)$, which are worse than (3.15) by a logarithmic term of some kind; see Refs. 16 and 17. Such estimates are a bit cumbersome as compared to (3.11), (3.13). Instead of discussing them, we consider two rough but simple estimates. The first one is (see, e.g., Ref. 18, Proposition 5.5)

$$|\xi(-\lambda;H_0-V,H_0)|\leqslant\|\sqrt{V}R(-\lambda,H_0)\sqrt{V}\|_{\mathbf{S}_q}^q\leqslant C(d,q)\lambda^{1-q}\|V\|_{L_q}^q,\quad q>1; \tag{3.16}$$

it is of an ''almost correct'' order in V and λ. Substituting (3.16) into (2.19), we get the following corollary.

Corollary 3.3: Let $\varkappa=1$; assume (2.3), (2.6) and let $V\in L_q(\mathbb{R}^d)$ for some $q>1$. Then, for any $\lambda>0$ and $\gamma>2$,

$$|\xi(\lambda;H_0-V,H_0)|\leqslant C(d,\gamma)\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}+C(d)(\log_+\lambda)\|V\|_{L_1}+C(d,q)\lambda^{1-q}\|V\|_{L_q}^q. \tag{3.17}$$

Another simple estimate is valid (see, e.g., Ref. 18, Proposition 5.4) for λ bounded away from zero:

$$|\xi(-\lambda;H_0-V,H_0)|\leqslant C(d,r)\|V\|_{l_1(L_r)},\quad r>1,\quad \lambda\geqslant 1. \tag{3.18}$$

Note that $V\in l_1(L_r)$, $r>1$, implies (2.3). Thus, substituting (3.18) into (2.19), we obtain the following.

Corollary 3.4: Let $\varkappa=1$; *assume (2.6) and let* $V\in l_1(L_r)$ *for some* $r>1$. *Then, for any* $\lambda\geqslant 1$ *and* $\gamma>2$,

$$|\xi(\lambda;H_0-V,H_0)|\leqslant C(d,\gamma)\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}+C(d)(\log_+\lambda)\|V\|_{L_1}+C(d,r)\|V\|_{l_1(L_r)}. \tag{3.19}$$

Substituting the estimates of Refs. 16, 17 into (2.19), one can obtain more precise statements.

E. A comparison with the results of Ref. 3

In Ref. 3, the following result has been obtained. Let $d\geqslant 2$, $l=1$ and let $V=V(x)\geqslant 0$ satisfy the estimate

$$V(x)\leqslant C_{3.20}(1+|x|)^{-\rho},\quad \rho>d. \tag{3.20}$$

For all $\lambda\geqslant C>0$ and all coupling constants $g>0$ the following bounds have been established (See Ref. 3, Theorem 4.2):

$$\xi(\lambda;H_0+gV,H_0)\leqslant C_{3.21}g\lambda^{\varkappa-1}(|\log\lambda|+1), \tag{3.21}$$

$$|\xi(\lambda;H_0-gV,H_0)|\leqslant C_{3.22}(g\lambda^{\varkappa-1}(|\log\lambda|+1)+g^{\varkappa}). \tag{3.22}$$

The constants $C_{3.21},C_{3.22}$ may depend on d and $C_{3.20}$. Clearly, (3.21) follows from (2.18) and (3.22)—from (3.12), (3.19). The basic difference between our results and the ones of Ref. 3 is in the fact that the dependence on V is explicit in (2.18), (3.12), (3.19) but not in (3.21), (3.22). Besides, Theorem 2.2 and Corollaries 3.1–3.4 extend the results of Ref. 3 in some other respects.

(1) Theorem 2.2 and Corollary 3.2 deal with the case $\varkappa<1$, which has not been considered in Ref. 3.
(2) The class of potentials V in Theorem 2.2 and Corollaries 3.1–3.4 is broader than the one given by (3.20).
(3) Theorem 2.2 and Corollaries 3.1–3.3 concern all $\lambda>0$, whereas in Ref. 3 λ is assumed to be bounded away from zero.

F. Integral estimates for the SSF

In Ref. 4, the *integral* estimates for the SSF $\xi(\lambda,H_0\pm V,H_0)$ have been obtained (for the same operators $H_0,V\geqslant 0$ as in the present paper). Before discussing them, let us write down the estimates for ξ^{cl}, which easily follow from the definition (3.1), (3.2):

$$\int_0^R\xi^{\mathrm{cl}}(\lambda;H_0+V,H_0)d\lambda\leqslant(2\pi)^{-d}\omega_dR^{\varkappa}\|V\|_{L_1}\quad R>0,\quad \varkappa>0, \tag{3.23}$$

$$\int_0^R|\xi^{\mathrm{cl}}(\lambda;H_0-V,H_0)|d\lambda\leqslant(2\pi)^{-d}\omega_dR^{\varkappa}\|V\|_{L_1}\quad R>0,\quad \varkappa\leqslant 1, \tag{3.24}$$

$$\int_0^R |\xi^{\mathrm{cl}}(\lambda;H_0-V,H_0)|d\lambda \leqslant C_1(d,l)R^{\varkappa}\|V\|_{L_1}+C_2(d,l)R\|V\|_{L_\varkappa}^{\varkappa}, \quad R>0, \ \ \varkappa>1. \tag{3.25}$$

It appears that the estimates (3.23), (3.25) can be carried over to the ''real'' SSF (see estimates (6.2) and (6.10), respectively, in Ref. 4); it is interesting that even the constant in [Ref. 4, Eq. (6.2)] coincides with the classical one, given by (3.23). For $\varkappa<1$, the estimate

$$\int_0^R |\xi(\lambda;H_0-V,H_0)|d\lambda \leqslant C(d,l)R^{\varkappa}\|V\|_{L_1} \quad R>0, \quad \varkappa<1$$

[with the constant $C(d,l)$ different from the classical one, given by (3.24)] follows from [Ref. 4, Eq. (6.7)] and (3.13). For $\varkappa=1$ the estimate [Ref. 4, Eq. (6.7)] together with (3.16) implies

$$\int_0^R |\xi(\lambda;H_0-V,H_0)|d\lambda \leqslant C(d)R\|V\|_{L_1}+C(d,q)R^{2-q}\|V\|_{L_q}^{q}, \quad R>0, \quad \varkappa=1,$$

for any $q>1$, and together with (3.18) it implies

$$\int_0^R |\xi(\lambda;H_0-V,H_0)|d\lambda \leqslant C(d)R\|V\|_{L_1}+C(d,r)R\|V\|_{l_1(L_r)}, \quad R>1, \quad \varkappa=1,$$

for any $r>1$.

G. Remarks

(1) Note that in Ref. 3 some estimates for the SSF have been found that have better order in λ and g, than (3.21), (3.22), depending on the exponent ρ in (3.20). These estimates, however, are of a conditional character, since they depend on some hypothesis on the boundary values of the resolvent of H_0, which has not been proved yet.

(2) The proof of Theorem 2.2 borrows some elements of Ref. 3. But the operator theoretic part of our approach (Propositions 4.2, 4.3) is completely different.

(3) It is natural to compare the estimates for the SSF with its asymptotics. Note that formula (3.9) for the ''real'' SSF and its various extensions is well known (see, e.g., Ref. 11 and references therein). The relation (3.10) for the ''real'' SSF is a well-known fact for $\lambda<0$ (see, e.g., Ref. 18) and has been proved in Ref. 19 for $\lambda>0$ and $l=1$. Comparing the estimates (2.18), (3.12), (3.14), (3.19) with the asymptotics (3.9), (3.10), we see that the estimates are of a correct order in the coupling constant g (as $g\to\infty$) and of an almost correct (up to the logarithmic terms) order in λ as $\lambda\to\infty$.

(4) In the case $d=l=1$, a pointwise estimate on the SSF, which is somewhat different from (3.14), has been obtained in Ref. 20. See the end of Sec. VIII for the discussion of this estimate.

IV. REPRESENTATION FOR THE SSF

Assume (2.3), (2.4) and denote $W:=\sqrt{V}$. Assumption (2.3) means that

$$W(H_0+I)^{-1/2}\in \mathbf{S}_\infty. \tag{4.1}$$

For $\operatorname{Im} z>0$ consider the ''sandwiched resolvent,''

$$T(z):=(W(H_0+I)^{-1/2})(H_0+I)R_0(z)(W(H_0+I)^{-1/2})^*=\overline{WR_0(z)W};$$

by (4.1), $T(z)\in\mathbf{S}_\infty$. Denote

$$A(z)=\operatorname{Re} T(z), \quad K(z)=\operatorname{Im} T(z).$$

By (2.4), for any bounded interval $\delta\subset\mathbb{R}$, one has

$$WE_{H_0}(\delta)\in\mathbf{S}_2. \tag{4.2}$$

From here we get the following statement.

Proposition 4.1: Assume (2.3), (2.4). For a.e. $\lambda\in\mathbb{R}$,

$$\exists\lim_{\epsilon\to 0+} T(\lambda+i\epsilon)=:T(\lambda+i0)\in\mathbf{S}_\infty \quad\text{and}\quad K(\lambda+i0)\in\mathbf{S}_1. \tag{4.3}$$

Proof: For any $\delta\subset\mathbb{R}$, denote

$$T_\delta(z):=(W(H_0+I)^{-1/2}E_{H_0}(\delta))(H_0+I)R_0(z)(W(H_0+I)^{-1/2}E_{H_0}(\delta))^*,$$

$$A_\delta(z)=\operatorname{Re}T_\delta(z),\quad K_\delta(z)=\operatorname{Im}T_\delta(z). \tag{4.4}$$

Now let $\delta\subset\mathbb{R}$ be some open *bounded* interval. It is one of the fundamental results of the trace class scattering theory (see Ref. 21 or 2) that (4.2) implies

$$\exists T_\delta(\lambda+i0)\in\mathbf{S}_2 \quad\text{and}\quad K_\delta(\lambda+i0)\in\mathbf{S}_1,\quad a.e.\ \lambda\in\mathbb{R}. \tag{4.5}$$

On the other hand, the operator $T_{\mathbb{R}\setminus\delta}(z)$ is analytically extendable through δ; obviously,

$$T_{\mathbb{R}\setminus\delta}(\lambda)\in\mathbf{S}_\infty \quad\text{and}\quad K_{\mathbb{R}\setminus\delta}(\lambda)=0,\quad \lambda\in\delta. \tag{4.6}$$

Finally, writing $T(z)=T_\delta(z)+T_{\mathbb{R}\setminus\delta}(z)$ and taking into account (4.5), (4.6), we get (4.3) for *a.e.* $\lambda\in\delta$. Since $\delta\subset\mathbb{R}$ is arbitrary, this implies the required statement. ■

It will follow from the reasoning of Sec. VII that under the additional assumption (2.6), the condition (4.3) actually holds for all $\lambda\neq 0$ and $T(\lambda+i0)$ depends continuously on λ in the operator norm and $K(\lambda+i0)$—in the trace norm.

Let λ be such that (4.3) holds; denote

$$\mathcal{N}_\pm(\lambda):=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{dt}{1+t^2}n_\pm(1,A(\lambda+i0)+tK(\lambda+i0)). \tag{4.7}$$

One easily checks that (4.3) implies convergence of the integral in (4.7).

Proposition 4.2 (Refs. 5 and 4): Assume (2.3), (2.4), (2.7); then for a.e. $\lambda\in\mathbb{R}$,

$$\xi(\lambda;H_0\pm V,H_0)=\pm\mathcal{N}_\mp(\lambda). \tag{4.8}$$

Representation (4.7), (4.8) for the SSF has been established in Ref. 4 (see also Ref. 22 for a generalization) as an abstract operator theoretic fact; application to the polyharmonic operator was considered in Ref. 4, Theorems 6.2, 6.3. Formula (4.8) can be considered as the *Birman–Schwinger principle on the continuous spectrum*; see Ref. 5 for the discussion.

A straightforward analysis of the r.h.s. of (4.7) gives the following.

Proposition 4.3 (Ref. 5): Let, for some $\lambda>0$, *the condition (4.3) hold. Then for any* $\theta\in(0,1)$ *the following estimate holds:*

$$\mathcal{N}_\pm(\lambda)\leq n_\pm(1-\theta,A(\lambda+i0))+\theta^{-1}\pi^{-1}\|K(\lambda+i0)\|_{\mathbf{S}_1}. \tag{4.9}$$

In the proof of Theorem 2.2, we shall use (4.9) together with the following decomposition:

$$A(\lambda+i0)=A_\delta(\lambda+i0)+A_{\mathbb{R}\setminus\delta}(\lambda),\quad \lambda>0,\quad \delta=(0,2\lambda), \tag{4.10}$$

where the operators on the r.h.s. are defined by (4.4). Substituting (4.10) into (4.9), fixing $\theta=1/4$, and using (2.1), we get the following inequality:

$$\mathcal{N}_\pm(\lambda)\leqslant n_\pm(\tfrac{1}{2},A_{\mathbb{R}\setminus\delta}(\lambda))+n_\pm(\tfrac{1}{4},A_\delta(\lambda+i0))+4\pi^{-1}\|K(\lambda+i0)\|_{\mathbf{S}_1}. \tag{4.11}$$

The relation (4.11) plays the key role in the proof of Theorem 2.2. In what follows we estimate each of the three terms on the r.h.s. of (4.11). Note that $A_{\mathbb{R}\setminus\delta}(\lambda)\geqslant 0$; thus, $n_-(\tfrac{1}{2},A_{\mathbb{R}\setminus\delta}(\lambda))=0$. Certainly, there is much freedom in choosing the constants θ in (4.9) and $s_{1,2}$ in (2.1), but this choice affects only the constants in the resulting formulas (2.18), (3.12), (3.14), (3.17), (3.19).

V. ESTIMATE FOR $A_{\mathbb{R}\setminus\delta}$

Assume (2.3); as in the end of the previous section, let us fix some $\lambda>0$, denote $\delta=(0,2\lambda)$, and define the operator $A_{\mathbb{R}\setminus\delta}$ according to (4.4).

Proposition 5.1: Under the above assumptions, for any $s>0$,

$$n_+(s,A_{\mathbb{R}\setminus\delta}(\lambda+i0))\leqslant|\xi(-\lambda;H_0-3s^{-1}V,H_0)|. \tag{5.1}$$

Proof: A straightforward calculation shows that

$$R(\lambda,H_0)E_{H_0}(\mathbb{R}\setminus\delta)\leqslant 3R(-\lambda,H_0),$$

in the quadratic form sense. It follows that $A_{\mathbb{R}\setminus\delta}(\lambda+i0)\leqslant 3T(-\lambda)$. From here, using the Birman–Schwinger principle, we get

$$\begin{aligned} n_+(s,A_{\mathbb{R}\setminus\delta}(\lambda+i0))&\leqslant n_+(s,3T(-\lambda))\\ &=\operatorname{rank}E_{H_0-3s^{-1}V}((-\infty,-\lambda))\\ &=|\xi(-\lambda;H_0-3s^{-1}V,H_0)|. \end{aligned}$$

■

VI. ESTIMATE FOR K

Assume (2.3), (2.4). For $\lambda>0$ consider the operator

$$(WE((0,\lambda)))(WE((0,\lambda)))^*. \tag{6.1}$$

By (4.2), the operator (6.1) belongs to the trace class. We will need a well-known representation for the derivative of (6.1) with respect to λ. In order to write down this representation, for every $t>0$ define the operator $J(t):L_2(\mathbb{R}^d)\to L_2(\mathbb{S}^{d-1})$, which acts according to the formula

$$J(t):\ \ f(x)\mapsto(2\pi)^{-d/2}\int_{\mathbb{R}^d}e^{it\langle x,\nu\rangle}W(x)f(x)dx,\quad \nu\in\mathbb{S}^{d-1}.$$

For $d=1$ by $\mathbb{S}^{d-1}$ we mean the set $\{-1,\ 1\}$. Clearly, $J(t)\in\mathbf{S}_2$ and

$$\|J(t)\|^2_{\mathbf{S}_2}=d\omega_d(2\pi)^{-d}\|V\|_{L_1}. \tag{6.2}$$

A straightforward calculation shows that the operator-valued function (6.1) is differentiable in λ in the trace class and

$$\mathcal{F}(\lambda):=\frac{d}{d\lambda}(WE((0,\lambda)))(WE((0,\lambda)))^*=(2l)^{-1}\lambda^{\varkappa-1}J^*(\lambda^{1/(2l)})J(\lambda^{1/(2l)})\geqslant 0. \tag{6.3}$$

If for some $\lambda>0$ the limit $T(\lambda+i0)$ exists, then, clearly,

$$K(\lambda+i0)=\pi\mathcal{F}(\lambda). \tag{6.4}$$

Thus,

$$\|K(\lambda+i0)\|_{\mathbf{s}_1}=\pi\,\mathrm{Tr}\,\mathcal{F}(\lambda)=\pi(2l)^{-1}\lambda^{\varkappa-1}\|J(\lambda^{1/(2l)})\|_{\mathbf{s}_2}^2=\pi\varkappa\omega_d(2\pi)^{-d}\lambda^{\varkappa-1}\|V\|_{L_1}. \quad (6.5)$$

VII. ESTIMATE FOR A_δ

A. Preliminary estimates

Define the function $\varphi(t)$, $t>0$, by

$$\varphi(t)=\begin{cases}|e^{it}-1|^2, & t\leqslant\pi,\\ 4, & t\geqslant\pi.\end{cases} \quad (7.1)$$

Proposition 7.1: Assume (2.4). For any $t_1>0$, $t_2>0$ *the following estimate holds:*

$$\|J(t_1)-J(t_2)\|_{\mathbf{s}_2}^2\leqslant d\,\omega_d(2\pi)^{-d}\int_{\mathbb{R}^d}V(x)\varphi(|x||t_1-t_2|)dx. \quad (7.2)$$

Proof:

$$\|J(t_1)-J(t_2)\|_{\mathbf{s}_2}^2\leqslant(2\pi)^{-d}\int_{\mathbb{R}^d}V(x)\int_{\mathbb{S}^{d-1}}|e^{it_1\langle\nu,x\rangle}-e^{it_2\langle\nu,x\rangle}|^2\,d\nu\,dx$$

$$\leqslant(2\pi)^{-d}\int_{\mathbb{R}^d}V(x)\int_{\mathbb{S}^{d-1}}\varphi(|x||t_1-t_2|)d\nu\,dx$$

$$=(2\pi)^{-d}\,d\,\omega_d\int_{\mathbb{R}^d}V(x)\varphi(|x||t_1-t_2|)dx.$$

■

We are reminded that $\mathcal{F}(\lambda)$ is defined by (6.3).

Proposition 7.2: Assume (2.4). For any $\lambda_1>0$, $\lambda_2>0$*:*

$$\|\mathcal{F}(\lambda_1)-\mathcal{F}(\lambda_2)\|_{\mathbf{s}_1}\leqslant\varkappa\omega_d(2\pi)^{-d}|\lambda_1^{\varkappa-1}-\lambda_2^{\varkappa-1}|\|V\|_{L_1}+2\varkappa\omega_d(2\pi)^{-d}\lambda_2^{\varkappa-1}\|V\|_{L_1}^{1/2}$$

$$\times\left(\int V(x)\varphi(|x||\lambda_1^{1/(2l)}-\lambda_2^{1/(2l)}|)dx\right)^{1/2}. \quad (7.3)$$

Proof: By (6.3),

$$\|\mathcal{F}(\lambda_1)-\mathcal{F}(\lambda_2)\|_{\mathbf{s}_1}=(2l)^{-1}\|\lambda_1^{\varkappa-1}J^*(\lambda_1^{1/(2l)})J(\lambda_1^{1/(2l)})-\lambda_2^{\varkappa-1}J^*(\lambda_2^{1/(2l)})J(\lambda_2^{1/(2l)})\|_{\mathbf{s}_1}$$

$$\leqslant(2l)^{-1}|\lambda_1^{\varkappa-1}-\lambda_2^{\varkappa-1}|\|J^*(\lambda_1^{1/(2l)})J(\lambda_1^{1/(2l)})\|_{\mathbf{s}_1}$$

$$+(2l)^{-1}\lambda_2^{\varkappa-1}\|(J^*(\lambda_1^{1/(2l)})-J^*(\lambda_2^{1/(2l)}))J(\lambda_1^{1/(2l)})\|_{\mathbf{s}_1}$$

$$+(2l)^{-1}\lambda_2^{\varkappa-1}\|J^*(\lambda_2^{1/(2l)})(J(\lambda_1^{1/(2l)})-J(\lambda_2^{1/(2l)}))\|_{\mathbf{s}_1}.$$

Substituting (6.2) and (7.2) onto the r.h.s. of the last estimate, we arrive at (7.3). ■

Proposition 7.3: Assume (2.6); then there exist such constants $0<C_{7.4}^{(1)}(\gamma)$ *and* $0<C_{7.4}^{(2)}(\gamma)<1$, *that*

$$\int V(x)\varphi(|x|s)dx\leqslant C_{7.4}^{(1)}(\gamma)|\log s|^{-\gamma}\|VF_\gamma\|_{L_1},\quad \forall s\in(0,C_{7.4}^{(2)}(\gamma)). \quad (7.4)$$

Proof: One has

$$\int V(x)\varphi(|x|s)dx \le \sup_{x\in\mathbb{R}^d}\frac{\varphi(|x|s)}{F_\gamma(x)}\int V(x)F_\gamma(x)dx.$$

It is a straightforward calculation to check that

$$\sup_{x\in\mathbb{R}^d}\frac{\varphi(|x|s)}{F_\gamma(x)} \le C_{7.4}^{(1)}(\gamma)|\log s|^{-\gamma},$$

for some constant $C_{7.4}^{(1)}(\gamma)$ and for all small enough $s>0$. ∎

B. Existence of $A_\delta(\lambda+i0)$

Proposition 7.4: Assume (2.6), fix $\lambda>0$ and let $\delta=(0,2\lambda)$. Then the limit $A_\delta(\lambda+i0)$ exists in the trace norm and

$$A_\delta(\lambda+i0)=\int_0^{2\lambda}\frac{\mathcal{F}(t)-\mathcal{F}(\lambda)}{t-\lambda}dt; \tag{7.5}$$

the last integral is absolutely convergent in the trace norm.

Proof: We start from the obvious formula, which is a consequence of the spectral theorem:

$$A_\delta(\lambda+i\epsilon)=\int_0^{2\lambda}\frac{t-\lambda}{(t-\lambda)^2+\epsilon^2}\mathcal{F}(t)dt, \quad \epsilon>0.$$

Clearly,

$$\int_0^{2\lambda}\frac{t-\lambda}{(t-\lambda)^2+\epsilon^2}\mathcal{F}(t)dt=\int_0^{2\lambda}\frac{t-\lambda}{(t-\lambda)^2+\epsilon^2}(\mathcal{F}(t)-\mathcal{F}(\lambda))dt. \tag{7.6}$$

From Propositions 7.2, 7.3, it follows that

$$\|\mathcal{F}(\lambda+s)-\mathcal{F}(\lambda)\|_{\mathbf{S}_1}\le C(\lambda,V)|\log|s\|^{-\gamma/2}, \tag{7.7}$$

for all small enough $s\in\mathbb{R}$. Thus, the integrands both in (7.6) and in (7.5) are dominated (in the trace norm) by an integrable function $C(\lambda,V)|t-\lambda|^{-1}|\log|t-\lambda\|^{-\gamma/2}$ in the neighborhood of $t=\lambda$. It follows that the r.h.s. of (7.6) converges to the r.h.s. of (7.5) as $\epsilon\to+0$. ∎

Remark 7.5: Note that it follows from the estimate (7.7) that under the assumptions (2.3), (2.6) the condition (4.3) holds for all $\lambda>0$ and the operator $A_\delta(\lambda+i0)$ is continuous in $\lambda>0$ in the trace norm. Thus, $T(\lambda+i0)$ is continuous in $\lambda>0$ in the operator norm. Next, if instead of (2.6) one assumes a stronger condition,

$$\int V(x)(1+|x|)^\gamma dx<\infty, \quad \gamma\in(0,2),$$

then the same reasoning leads to the estimate

$$\|\mathcal{F}(\lambda+s)-\mathcal{F}(\lambda)\|_{\mathbf{S}_1}\le C(\lambda,V)|s|^{\gamma/2},$$

for all small enough $s\in\mathbb{R}$. This estimate implies a Hölder continuity of $A_\delta(\lambda+i0)$ in the trace norm with the exponent $\gamma/2$ and thus the Hölder continuity of $T(\lambda+i0)$ in the operator norm. However, in what follows we shall not need these facts.

C. Estimates for $A_\delta(\lambda+i0)$

First we consider the case $\lambda=1$.

Proposition 7.6: Assume (2.6); let $\delta=(0,2)$. *One has*

$$\|A_\delta(1+i0)\|_{\mathbf{s}_1}\leqslant C(d,l)\|V\|_{L_1}+C(d,l)\|V\|_{L_1}^{1/2}\int_0^1\frac{ds}{s}\left(\int V(x)\varphi(|x|s)dx\right)^{1/2}. \tag{7.8}$$

Proof: Let us use (7.5) and (7.3):

$$\begin{aligned}\|A_\delta(1+i0)\|_{\mathbf{s}_1}&\leqslant\int_0^2\frac{\|\mathcal{F}(\lambda)-\mathcal{F}(1)\|_{\mathbf{s}_1}}{|\lambda-1|}d\lambda\\&\leqslant\varkappa\omega_d(2\pi)^{-d}\|V\|_{L_1}\int_0^2\frac{|\lambda^{\varkappa-1}-1|}{|\lambda-1|}d\lambda\\&\quad+2\varkappa\omega_d(2\pi)^{-d}\|V\|_{L_1}^{1/2}\int_0^2\frac{d\lambda}{|\lambda-1|}\left(\int V(x)\varphi(|x||\lambda^{1/(2l)}-1|)dx\right)^{1/2}.\end{aligned} \tag{7.9}$$

Changing the variable in the last integral, we obtain

$$\begin{aligned}&\int_0^2\frac{1}{|\lambda-1|}\left(\int V(x)\varphi(|x||\lambda^{1/(2l)}-1|)dx\right)^{1/2}d\lambda\\&=\int_{-1}^{2^{1/(2l)}-1}\frac{2l(s+1)^{2l-1}}{|(s+1)^{2l}-1|}\left(\int V(x)\varphi(|x|s)dx\right)^{1/2}ds\\&\leqslant C(l)\int_{-1}^1\frac{ds}{|s|}\left(\int V(x)\varphi(|x|s)dx\right)^{1/2}+\int_1^{2^{1/(2l)}-1}\frac{2l(s+1)^{2l-1}}{|(s+1)^{2l}-1|}\left(\int V(x)\varphi(|x|s)dx\right)^{1/2}ds\\&\leqslant C(l)\int_0^1\frac{ds}{s}\left(\int V(x)\varphi(|x|s)dx\right)^{1/2}+C(l)\|V\|_{L_1}^{1/2}.\end{aligned} \tag{7.10}$$

Note that the integral over $(1,2^{1/(2l)}-1)$ enters the last calculation only if $2^{1/(2l)}-1>1$, i.e., if $l<\frac{1}{2}$. In order to estimate this integral, we use the bound $\varphi(t)\leqslant 4$. The estimates (7.9) and (7.10) together give (7.8). ■

Proposition 7.7: Assume (2.6), fix $\lambda>0$ *and let* $\delta=(0,2\lambda)$. *The following estimate holds:*

$$\|A_\delta(\lambda+i0)\|_{\mathbf{s}_1}\leqslant C(d,l,\gamma)\lambda^{\varkappa-1}\|V\|_{L_1}^{1/2}\|VF_\gamma\|_{L_1}^{1/2}+C(d,l)\lambda^{\varkappa-1}(\log_+\lambda)\|V\|_{L_1}. \tag{7.11}$$

Proof: We start from the standard dilatation argument. Namely, let U_r, $r>0$ be the unitary dilatation operator in $L_2(\mathbb{R}^d)$: $(U_rf)(x)=r^{d/2}f(rx)$. Then

$$U_rA(z)U_r^*=r^{2l}A^{(r)}(r^{2l}z),$$

where $A^{(r)}$ corresponds to the perturbation potential $V^{(r)}(x)=V(xr)$. Thus, taking $r=\lambda^{-1/(2l)}$, using (7.8), and changing variable in the resulting integrals, we obtain

$$\begin{aligned}\|A(\lambda+i0)\|_{\mathbf{s}_1}&=\|U_rA(\lambda+i0)U_r^*\|_{\mathbf{s}_1}\\&=\lambda^{-1}\|A^{(r)}(1+i0)\|_{\mathbf{s}_1}\\&\leqslant\lambda^{\varkappa-1}C(d,l)\|V\|_{L_1}+C(d,l)\lambda^{\varkappa-1}\|V\|_{L_1}^{1/2}\int_0^{\lambda^{1/(2l)}}\frac{ds}{s}\left(\int V(x)\varphi(|x|s)dx\right)^{1/2}.\end{aligned}$$

It remains to estimate the last integral. First let $\lambda \geqslant 1$. Using (7.4) and the bound $\varphi(t) \leqslant 4$, we get

$$\int_0^{\lambda^{1/(2l)}} \frac{ds}{s} \left(\int V(x)\varphi(|x|s)dx \right)^{1/2} \leqslant (C_{7.4}^{(1)}(\gamma))^{1/2} \|VF_\gamma\|_{L_1}^{1/2} \int_0^{C_{7.4}^{(2)}} \frac{ds}{s} |\log s|^{-\gamma/2} + 2\|V\|_{L_1}^{1/2} \int_{C_{7.4}^{(2)}}^{\lambda^{1/(2l)}} \frac{ds}{s}$$
$$\leqslant C(\gamma)\|VF_\gamma\|_{L_1}^{1/2} + \|V\|_{L_1}^{1/2} l^{-1} \log_+ \lambda - 2\|V\|_{L_1}^{1/2} \log C_{7.4}^{(2)}$$
$$\leqslant (C(\gamma) - 2\log C_{7.4}^{(2)})\|VF_\gamma\|_{L_1}^{1/2} + \|V\|_{L_1}^{1/2} l^{-1} \log_+ \lambda$$
$$\leqslant C(\gamma)\|VF_\gamma\|_{L_1}^{1/2} + \|V\|_{L_1}^{1/2} l^{-1} \log_+ \lambda.$$

Finally, if $\lambda < 1$, we merely replace the integration interval $(0,\lambda^{1/(2l)})$ by (0, 1), thus getting the upper bound. ■

VIII. PROOF OF THEOREM 2.2. CONCLUDING REMARKS

A. Proof of Theorem 2.2

Obviously,

$$n_\pm(\tfrac{1}{4}, A_\delta(\lambda + i0)) \leqslant 4\|A_\delta(\lambda + i0)\|_{\mathbf{s}_1}.$$

It remains to substitute the last estimate together with (5.1), (6.5), (7.11) into (4.11) and take into account Proposition 4.2. ■

B. Remarks

(1) Our Proposition 7.4 is fairly close to the ''limiting absorption principle in the trace class'' of Ref. 3, though these two statements are proved by using a very different technique.

(2) One can exploit (4.11) by using some other (different from ours) estimates for $K(\lambda + i0)$, $A_\delta(\lambda + i0)$, $A_{\mathbb{R}\setminus\delta}(\lambda + i0)$, thus obtaining new estimates for the SSF. Let us give an example.

In Ref. 23, the following estimate has been proved for $l = 1$, $d \geqslant 2$:

$$\|A_\delta(\lambda + i0)\|_{\mathbf{s}_2}^2 \leqslant C(d)\lambda^{d-2-(q/2)} \int\int \frac{V(x)V(x')}{|x-x'|^q} dx\, dx', \quad q \in [0, d-1]. \tag{8.1}$$

For $d = 1$ this estimate is also true (which follows from the explicit formula for the integral kernel of the resolvent of H_0). Observing that

$$n_\pm(s, A_\delta(\lambda + i0)) \leqslant \tfrac{1}{2} s^{-2} \|A_\delta(\lambda + i0)\|_{\mathbf{s}_2}^2, \tag{8.2}$$

and substituting (8.1), (8.2) into (4.11), one obtains the following estimates for $l = 1$, $\lambda > 0$, $q \in [0, d-1]$:

$$\xi(\lambda; H_0 + V, H_0) \leqslant C(d)\lambda^{d-2-(q/2)} \int\int \frac{V(x)V(x')}{|x-x'|^q} dx\, dx' + C(d)\lambda^{(d/2)-1}\|V\|_{L_1},$$

$$|\xi(\lambda; H_0 - V, H_0)| \leqslant C(d)\lambda^{d-2-(q/2)} \int\int \frac{V(x)V(x')}{|x-x'|^q} dx\, dx' + C(d)\lambda^{(d/2)-1}\|V\|_{L_1}$$
$$+ |\xi(-\lambda; H_0 - 6V, H_0)|.$$

One can combine the last estimate with (3.11), (3.13), (3.16), (3.18) or similar bounds in an obvious way. Note that the estimate (2.10) of Ref. 20 is contained in this series of estimates (for $d=1$).

ACKNOWLEDGMENTS

The author is grateful to M. Sh. Birman and A. V. Sobolev for useful discussions. The research was partially supported by EPSRC Grant No. GR/M00549.

[1] M. Sh. Birman and D. R. Yafaev, ''The spectral shift function. The papers of M. G. Krein and their further development,'' Alg. Anal. **4**, 1–44 (1992) (in Russian); English translation in St. Petersburg Math. J. **4**, 833–870 (1993).
[2] D. R. Yafaev, *Mathematical Scattering Theory. General Theory* (American Mathematical Society, Providence, RI, 1992).
[3] A. V. Sobolev, ''Efficient bounds for the spectral shift function,'' Ann. Inst. Henri Poincaré Phys. Theor. **58**, 55–83 (1993).
[4] A. B. Pushnitski, ''Integral estimates for the spectral shift function,'' Alg. Anal. **10**, 198–234 (1998) (in Russian); English translation to appear in St. Petersburg Math. J. **10** (1999).
[5] A. B. Pushnitski, ''Representation for the spectral shift function for perturbations of a definite sign,'' Alg. Anal. **9**, 197–213 (1997) (in Russian); English translation in St. Petersburg Math. J. **9**, 1181–1194 (1998).
[6] M. Sh. Birman and M. Z. Solomyak, ''Spectral theory of selfadjoint operators in Hilbert space,'' translated from Russian. Mathematics and its Applications (Soviet Series) (Reidel, Dordrecht, 1987).
[7] M. G. Krein, ''On the trace formula in perturbation theory,'' Mat. Sb. **33**, 597–626 (1953) (in Russian).
[8] M. Sh. Birman and M. Z. Solomyak, ''Estimates for the singular numbers of integral operators,'' Usp. Mat. Nauk **32**, 17–84 (1977); English translation Russi. Math. Surv. **32**, 15–89 (1977).
[9] M. Reed and B. Simon, ''Methods of modern mathematical physics,'' *Scattering Theory* (Academic, New York, 1979), Vol. 3.
[10] L. S. Koplienko, ''On the theory of the spectral shift function,'' Probl. Mat. Fiz. **9**, 62–72 (1971) (in Russian).
[11] D. Robert, ''Semi-classical asymptotics of the spectral shift function,'' *AMS Translation Series 2*, Vol. 189 Advances in Mathematical Sciences 1999, Vol. 41, pp. 187–203.
[12] M. Cwikel, ''Weak type estimates for singular values and the number of bound states of Schrödinger operators,'' Ann. Math. **206**, 93–100 (1977).
[13] E. Lieb, ''The number of bound states of one-body Schrödinger operators and the Weyl problem,'' Bull. Am. Math. Soc. **82**, 751–753 (1976).
[14] G. V. Rozenblum, ''Distribution of the discrete spectrum of singular differential operators,'' Dokl. Akad. Nauk SSSR **202**, 1012–1015 (1972) (in Russian); English translation in Sov. Math. Dokl. **13**, 245–249 (1972).
[15] M. Sh. Birman and A. Laptev, ''The negative discrete spectrum of a two-dimensional Schrödinger operator,'' Commun. Pure Appl. Math. **49**, 967–997 (1996).
[16] M. Z. Solomyak, ''Piecewise-polynomial approximation of functions from $H^l((0,1)^d)$, $2l=d$, and applications to the spectral theory of the Schrödinger operator,'' Isr.l J. Math. **86**, 253–276 (1994).
[17] T. Weidl, ''Another Look at Cwikel's inequality,'' *AMS Transl. Series 2, Vol. 189*, Advances in Math. Sciences,1999, Vol.**41**, pp. 247–254.
[18] M. Sh. Birman and M. Z. Solomyak, ''Estimates for the number of negative eigenvalues of the Schrödinger operator and its generalizations,'' *Estimates and Asymptotics For Discrete Spectra of Integral and Differential Equations* (Leningrad, 1989–90) Advances in Soviet Mathematics (American Mathematical Society, Providence, RI, 1991).
[19] A. B. Pushnitski, ''Spectral shift function of the Scrödinger operator in the large coupling constant limit,'' to appear in Commun. Partial Diff. Eqns.
[20] V. Kostrykin and R. Schrader, ''Scattering theory approach to random Schrödinger operators in one dimension,'' Rev. Math. Phys. Vol. **11,** 187–242 (1999).
[21] M. Sh. Birman and S. B. Entina, ''The stationary approach in abstract scattering theory,'' Izv. Akad. Nauk SSSR, Met. **31**, 401–430 (1967); English translation in Math. USSR Izv. **1**, 391–420 (1967).
[22] F. Gesztesy and K. A. Makarov, ''The Ξ operator and its relation to Krein's spectral shift function,'' preprint.
[23] A. V. Sobolev and D. R. Yafaev, ''Spectral properties of the abstract scattering matrix, '' Trudy MIAN **188**, 125–149 (1990) (in Russian); English translation in Proc. Steklov Inst. Math. **3**, 159–189 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Symplectic Dirac–Kähler fields

M. Reuter[a)]
Institut für Physik, Universität Mainz, Staudingerweg 7, D-55099 Mainz, Germany

(Received 26 March 1999; accepted for publication 12 July 1999)

For the description of space–time fermions, Dirac–Kähler fields (inhomogeneous differential forms) provide an interesting alternative to the Dirac spinor fields. In this paper we develop a similar concept within the symplectic geometry of phase spaces. Rather than on space–time, symplectic Dirac–Kähler fields can be defined on the classical phase space of any Hamiltonian system. They are equivalent to an infinite family of metaplectic spinor fields, i.e., spinors of $\mathrm{Sp}(2N)$, in the same way an ordinary Dirac–Kähler field is equivalent to a (finite) multiplet of Dirac spinors. The results are interpreted in the framework of the gauge theory formulation of quantum mechanics which was proposed recently. An intriguing analogy is found between the lattice fermion problem (species doubling) and the problem of quantization in general.
[S0022-2488(99)01411-5]

I. INTRODUCTION

In a classic paper[1] Kähler proposed a description of fermions in terms of inhomogeneous differential forms. Rather than by spinor fields, the fermions are represented by a set of antisymmetric tensors in this approach. The role of the Dirac equation is taken over by the so-called Dirac–Kähler equation which involves only tensor manipulations. It imitates the γ-matrix algebra with the help of the Clifford product for forms.

At first sight it seems puzzling how a family of tensor fields carrying integer spin can describe a particle of half-integer spin. This paradox is resolved if one notes that (in 4 space–time dimensions) a single Dirac–Kähler field actually corresponds to a multiplet of 4 ordinary Dirac spinors which mix under Lorentz transformations in a nontrivial way (''flavor mixing'').

The Dirac–Kähler fermions have attracted a lot of attention both from the physics[2–9] and the mathematics[10,11] point of view. In particular they have made their appearance in lattice field theory.[12] It is a well-known problem that a straightforward lattice discretization of the ordinary Dirac action does not describe one but rather 16 fermions in the continuum limit. The reason for this replication of fermionic states (usually referred to as the species ''doubling'' problem) is that the lattice propagator in momentum space has poles at all 16 corners of the Brillouin zone. The Kogut–Susskind[13] or staggered lattice fermions were proposed as an attempt to solve this problem. They are based on a more sophisticated lattice action which reduces the number of fermion species from 16 to 4. Later on it turned out[2,3] that the Kogut–Susskind fermions are nothing but Dirac–Kähler fields discretized on a hypercubic lattice. As it deals with differential forms only, Dirac–Kähler theory on the lattice can take advantage of all the mathematical tools provided by the algebraic topology of cell complexes. In particular, by a standard procedure, the differential forms of the continuum formulation can be replaced by appropriate cochains on the lattice. These cochains are functions defined on the lattice points, links, plaquettes, cubes, and hypercubes of the underlying lattice. In this manner it becomes obvious that the extra fermion species implied by the Kogut–Susskind lattice action and the fact that a Dirac–Kähler field contains 4 ordinary Dirac fermions have a common origin.

We only mention that the species doublers on the lattice can be avoided completely by using

[a)]Electronic mail: reuter@thep.physik.uni-mainz.de

Wilson fermions or the nonlocal ''SLAC derivative,''[12] for instance. Alternatively one can regard the 4 Dirac fermions contained in one Kogut–Susskind field as 4 different physical ''flavors.'' As we are interested in Dirac–Kähler fermions here we shall adopt this latter point of view in the following.

Dirac–Kähler (DK) fields can be defined on any Riemannian manifold $(\mathcal{M}_n, g)$, i.e., on any smooth n-dimensional manifold equipped with a metric g. [The pseudo-Riemannian case (Lorentzian space–times) can be dealt with in a completely analogous fashion.] From the physics point of view this manifold represents *space–time*.

The main purpose of the present paper is to propose an analog of the DK fields which ''live'' on symplectic rather than Riemannian manifolds. This means that we are going to study DK fields not over space–time but rather over a *phase space*.

A symplectic manifold $(\mathcal{M}_{2N}, \omega)$ is a smooth $2N$-dimensional manifold which is endowed with a closed, nondegenerate 2-form $\omega = \frac{1}{2}\omega_{ab} d\phi^a \wedge d\phi^b$. (The ϕ^a, $a=1,...,2N$, are local coordinates on $\mathcal{M}_{2N}$.) This manifold should be thought of as the phase space of a Hamiltonian system with N degrees of freedom. The corresponding Poisson bracket is given by $\{\phi^a, \phi^b\} = \omega^{ab}$ where the matrix (ω^{ab}) is the inverse of (ω_{ab}). Using local Darboux coordinates $\phi^a \equiv (p^i, q^i)$, $i = 1,...,N$, this matrix is independent of ϕ^a, $\omega^{qp} = -\omega^{pq} = I$, and the only nonvanishing brackets are $\{q^i, p^j\} = \delta^{ij}$. If ϕ^a and $\tilde{\phi}^a$ are local coordinates belonging to two overlapping charts of an atlas covering $\mathcal{M}_{2N}$ then, by the very definition of a symplectic manifold, the coordinate transformation $\phi \to \tilde{\phi}$ is symplectic, i.e., the Jacobian matrix $(\partial \tilde{\phi}^a / \partial \phi^b)$ is an element of Sp$(2N)$ at every point of the overlap region. Sp$(2N)$, the group of linear canonical transformations, plays the same role for phase space which the Lorentz group plays for space–time. In particular, it is the structure group of the frame bundle over $\mathcal{M}_{2N}$.

As for introducing DK fields on symplectic manifolds the first question which we must answer is what kind of spinor field should be used in place of the ordinary Dirac spinors of relativistic field theory. The only natural choice here is to employ the so-called metaplectic spinors,[14] i.e., the spinors of the metaplectic group Mp$(2N)$. Basically Mp$(2N)$ is related to Sp$(2N)$ in the same way Spin(n) is related to SO(n). In particular, there exists a two-to-one homomorphism between the two groups, i.e., Mp$(2N)$ covers Sp$(2N)$ twice. The construction of metaplectic spin bundles and spinor fields over a symplectic manifold proceeds almost literally along the same lines as in the case of space–time spinors, the main difference being that it is Mp$(2N)$ now which serves as the structure group. For a detailed exposition we must refer to the literature.[14,15]

Metaplectic spinors have been used in many different contexts including geometric quantization,[15] semiclassical approximations,[16] Parisi–Sourlas supersymmetry,[17] string theory,[18,19] and anyon superconductivity.[20] Most recently they played an important role in an approach to quantization[21] which is based upon a Yang–Mills theory on phase space with metaplectic ''matter'' fields. This new formulation of quantum mechanics is one of the main motivations for the present work. We shall come back to it later on.

Let us briefly describe how one can construct representations of Mp$(2N)$.[22] One has to associate an operator $M(S)$ to every matrix $S \equiv (S^a{}_b) \in \text{Sp}(2N)$ in such a way that $M(S_1)M(S_2) = \pm M(S_1 S_2)$. These operators can be built up from a kind of ''γ matrices'' which constitute a symplectic Clifford algebra:

$$\gamma^a \gamma^b - \gamma^b \gamma^a = 2i\omega^{ab}. \tag{1.1}$$

We require $M(S)$ to satisfy the usual compatibility condition between the vector and the spinor representation:

$$M(S)^{-1} \gamma^a M(S) = S^a{}_b \gamma^b. \tag{1.2}$$

Every infinitesimal Sp$(2N)$ transformation is of the form $S^a{}_b = \delta^a_b + \omega^{ac}\kappa_{cb}$ with symmetric coefficients κ_{ab}. Inserting this together with the ansatz $M(S) = 1 - (i/2)\kappa_{ab}\Sigma^{ab}_{\text{meta}}$ into the compatibility condition it is easy to show that the latter is solved by

$$\Sigma^{ab}_{\text{meta}} = \tfrac{1}{4}(\gamma^a\gamma^b + \gamma^b\gamma^a) \tag{1.3}$$

and that these generators satisfy the $\text{Sp}(2N)$ commutator relations.[22] Thus every representation of the symplectic Clifford algebra gives rise to a representation of $\text{Mp}(2N)$.

The most obvious difference between the metaplectic and the space–time spinors is that the symplectic Clifford algebra involves a commutator rather than an anticommutator. As an immediate consequence, this algebra has no finite dimensional matrix representations, and metaplectic spinors are necessarily infinite component objects. What is meant by a ''metaplectic representation'' is a representation in which γ^a is a Hermitian operator on an infinite dimensional Hilbert space $\mathcal{V}$. Hence the operators M obtained by exponentiating the generators (1.3) give rise to a unitary representation. (See Refs. 22 and 23 for further details.)

The symplectic Clifford algebra (1.1) admits a rather intriguing reinterpretation which is also at the heart of the new approach to quantization[21] mentioned above. Assume we are given a quantum mechanical system with a Hilbert space $\mathcal{V}$ along with N position and momentum operators $\hat{x}^i$ and $\hat{\pi}^i$ acting on it. They satisfy the canonical commutator relations $[\hat{x}^i,\hat{\pi}^j] = i\hbar\,\delta^{ij}$. By virtue of the identification $\gamma^i = \kappa\hat{\pi}^i$, $\gamma^{N+i} = \kappa\hat{x}^i$ for $i=1,...,N$ and with the constant $\kappa \equiv \sqrt{2/\hbar}$ it is obvious that the ''symplectic Clifford algebra'' (1.1) is actually nothing but the canonical commutation relations for the $\hat{x}-\hat{\pi}$ auxiliary quantum system. We call it an ''auxiliary'' system because it should not be confused with the actual physical system under consideration, the one whose (curved) phase space is $\mathcal{M}_{2N}$. (The classical phase space pertaining to the auxiliary system is simply $\mathbf{R}^{2N}$ equipped with the standard symplectic structure.)

The metaplectic spin bundles are bundles over $\mathcal{M}_{2N}$ with the typical fiber $\mathcal{V}$ and the structure group $\text{Mp}(2N)$.[14] At each point ϕ of $\mathcal{M}_{2N}$ a local copy of $\mathcal{V}$, denoted $\mathcal{V}_\phi$, is attached. Metaplectic spinor fields are sections through these bundles. Locally they are simply functions which assume values in $\mathcal{V}$:

$$\psi: \quad \mathcal{M}_{2N} \to \mathcal{V}, \quad \phi \mapsto |\psi\rangle_\phi \in \mathcal{V}_\phi. \tag{1.4}$$

The notation $|\psi\rangle_\phi$ means that the spinor $|\psi\rangle \in \mathcal{V}$, ''lives'' in the local Hilbert space at ϕ. Upon introducing a basis $\{|\alpha\rangle\}$ in $\mathcal{V}$ we write $\psi^\alpha(\phi) \equiv \langle\dot{\alpha}|\psi\rangle_\phi$ for its components. Here α is an infinite dimensional generalization of a spinor index. If we take $\{|\alpha\rangle\}$ to be the $\hat{x}$ eigenbasis, for instance, then $\alpha \equiv (\alpha^1,\dots,\alpha^N) \in \mathbf{R}^N$. (See Refs. 22 and 23 for details.)

In the present paper we shall focus on the local aspects of the bundles involved. We only mention that on certain manifolds there are topological obstructions which prevent them from carrying globally well-defined metaplectic spinor fields.[14] In Ref. 23 we characterized these obstructions using methods from quantum field theory.

Let us come back to the main question which we are trying to answer in this paper: *Do there exist ''symplectic Dirac-Kähler fields'' which are related to the metaplectic spinors in the same way the ordinary Dirac-Kähler fields are related to Dirac spinors*?

Apart from being interesting in its own right, this question is of obvious physical relevance. The fascinating property of metaplectic spinor fields is that, *on a purely group theoretical basis*, they introduce aspects of quantum mechanics into the geometry of classical phase spaces. By pure representation theory one is led to the auxiliary quantum system in the local Hilbert spaces $\mathcal{V}_\phi$. In Ref. 21 we explained in detail how these auxiliary systems relate to the actual physical quantum system with the classical phase space $\mathcal{M}_{2N}$. Using this as our starting point, we showed that it is possible to replace conventional canonical quantization by two new rules with a more transparent physical and geometrical meaning.

Classical mechanics and classical statistical mechanics are geometric theories which are conveniently described in the language of symplectic geometry. Only tensor fields are needed to formulate them. Quantum mechanics, on the other hand, has a natural interpretation in terms of spinor fields on phase space. Thus, in a sense, the very process of quantization is tantamount to a transition from tensors to spinors. But this is precisely what Dirac–Kähler theory is about: its basic fields are tensors which, however, are equivalent to a multiplet of spinors.

Before embarking on the detailed constructions let us briefly outline the strategy for finding the ''symplectic DK fields'' which we shall follow in this paper.

Our main tools are two types of auxiliary quantum systems with Hilbert spaces $\mathcal{V}$ and $\mathcal{V}^F$, respectively. We mentioned already the (bosonic) $\hat{x}-\hat{\pi}$ system on $\mathcal{V}$ whose canonical operators realize the metaplectic γ-matrices γ^a. We also need a similar fermionic system with a (finite-dimensional) Hilbert space $\mathcal{V}^F$ and a set of operators $\hat{\chi}^\mu$, $\mu=1,...,n$, satisfying the canonical anticommutator relations $\hat{\chi}^\mu\hat{\chi}^\nu+\hat{\chi}^\nu\hat{\chi}^\mu=\hbar\,\delta^{\mu\nu}$. The SO($n$)-Dirac matrices γ^μ are treated as a special realization of this algebra.

An important technical ingredient is the Weyl symbol calculus.[24–28] Let $\mathcal{L}(\mathcal{V})$ and $\mathcal{L}(\mathcal{V}^F)$ be the spaces of linear operators on $\mathcal{V}$ and $\mathcal{V}^F$, respectively. It is possible to uniquely characterize every operator $\hat{b}\in\mathcal{L}(\mathcal{V})$ and $\hat{f}\in\mathcal{L}(\mathcal{V}^F)$ in terms of classical phase functions (symbols) $b(y)$ and $f(\theta)$. Here y and θ are coordinates on the (flat) classical phase spaces which belong to the auxiliary systems. In the bosonic case, $y\equiv(y^a)\in\mathbf{R}^{2N}$ is a vector with commuting entries, while $\theta\equiv(\theta^\mu)$ is a set of n anticommuting Grassmann numbers. The space of all bosonic (fermionic) symbol functions, equipped with certain algebraic structures, is referred to as the bosonic (fermionic) Weyl algebra $\mathcal{W}(\mathcal{W}^F)$.

Given a space–time manifold $\mathcal{M}_n$, we consider fields on this manifold which assume values in $\mathcal{V}^F$, $\mathcal{L}(\mathcal{V}^F)$ and $\mathcal{W}^F$, respectively. In an obvious notation, we denote them $\psi^\alpha(x)$, $\hat{F}(x)$, and $F(x,\theta)$.

Similarly, given a phase-space manifold $\mathcal{M}_{2N}$, we define fields $\psi^\alpha(\phi)$, $\hat{B}(\phi)$, and $B(\phi,y)$ which assume values in $\mathcal{V}$, $\mathcal{L}(\mathcal{V})$ and $\mathcal{W}$, respectively.

In the first part of this paper we shall reformulate standard Dirac–Kähler theory in terms of the fermionic Weyl symbol calculus. We shall see that $\psi^\alpha(x)$ is an ordinary Dirac spinor and that $F(x,\theta)$ can be identified with a DK field. The Grassmann variables θ^μ will play the role of the basis differentials dx^μ.

This first part of the investigation is quite interesting in its own right. For instance, we shall discover that the Clifford product which is at the heart of DK theory is basically the same thing as the star product of the fermionic Weyl symbol calculus. As a consequence, $\mathcal{W}^F$ turns out to be an Atiyah–Kähler algebra.[10,11]

In the second part of this paper we investigate in detail what happens to the standard DK theory, reformulated in terms of fermionic Weyl symbols, when we replace fermionic symbols by bosonic ones everywhere. This means that we switch from the $\hat{\chi}^\mu$ to the $\hat{x}-\hat{\pi}$ system. Then $\psi^\alpha(\phi)$ is a metaplectic spinor field, and by analogy with the fermionic setting we shall argue that $B(\phi,y)$ is the ''symplectic DK field'' which we are looking for. Schematically our approach can be summarized as follows:

$$\begin{array}{ccc} \text{DK fields} & \Leftrightarrow & \text{fermionic symbols} \\ & & \downarrow \\ \text{symplectic DK fields} & \Leftrightarrow & \text{bosonic symbols} \end{array} \tag{1.5}$$

The rest of this paper is organized as follows: In the second half of this introduction we discuss some aspects of standard DK theory which will be important later on. Then, in Sec. II, we reformulate this theory in terms of fermionic Weyl symbols. Particular attention is paid to the decomposition of DK fields as a set of Dirac spinors. The construction of the symplectic DK fields is performed in Sec. III. We investigate in detail which properties of SO(n) DK fields can be translated to the Sp($2N$) case and which cannot. Section IV contains a summary and various remarks on the quantization problem in the light of the present work. Some material needed as a background for Sec. II is relegated to the Appendix.

As for its mathematical rigor, the style of this paper is informal. Occasionally the language of fiber bundles is used as a convenient tool but we are mostly interested in the local properties of the bundles involved and no pretense is made as for a rigorous and complete discussion of the global aspects.

DK fields on space–time

Let us start with an arbitrary (curved) n-dimensional Riemannian manifold $(\mathcal{M}_n,g)$. Upon introducing local coordinates x^μ, the tangent space $T_x\mathcal{M}_n$ and the cotangent space $T_x^*\mathcal{M}_n$ at the point x of $\mathcal{M}_n$ are spanned by the basis vectors $\partial_\mu\equiv\partial/\partial x^\mu$ and dx^μ, $\mu=1,...,n$, respectively. These spaces constitute the fibers of the (co-)tangent bundle over $\mathcal{M}_n$. Replacing $T^*\mathcal{M}_n$ by its p-fold tensor power we obtain the bundle of (covariant) tensors of rank p. Restricting ourselves to completely antisymmetric tensors we are led to the exterior algebra $\bigwedge(T_x^*\mathcal{M}_n)=\oplus_{p=0}^n \bigwedge^p(T_x^*\mathcal{M}_n)$. Its elements are the inhomogeneous differential forms

$$\Phi(x)=\sum_{p=0}^{n}\Phi^{(p)}(x),\quad \Phi^{(p)}(x)\in\bigwedge^p(T_x^*\mathcal{M}_n),$$
$$\Phi^{(p)}(x)=\frac{1}{p!}F^{(p)}_{\mu_1\cdots\mu_p}(x)dx^{\mu_1}\wedge\cdots\wedge dx^{\mu_p} \tag{1.6}$$

where $F^{(p)}_{\mu_1\cdots\mu_p}$ are completely antisymmetric coefficients. The corresponding algebra multiplication is the wedge product ''$\wedge$.''

Since we have a metric $g=g_{\mu\nu}(x)dx^\mu\otimes dx^\nu$ at our disposal which gives rise to an analogous bilinear form $g'=g^{\mu\nu}(x)\partial_\mu\otimes\partial_\nu$ for the cotangent bundle we can promote the fibers $\bigwedge(T_x^*\mathcal{M}_n)$ of the exterior algebra bundle to an Atiyah–Kähler algebra $\mathbf{AK}(T_x^*\mathcal{M}_n,g')$.[1,10,11]

Quite generally, the Atiyah–Kähler algebra $\mathbf{AK}(V,Q)$ corresponding to an arbitrary vector space V equipped with a quadratic form Q consists of the elements of the exterior algebra over V, $\bigwedge(V)=\oplus_p\bigwedge^p(V)$, for which the following three products are defined:

(1) the exterior product ''$\wedge$,''
(2) the inner product $(\cdot,\cdot)$ induced by Q,
(3) the Clifford product ''$\vee$.''

The three products are required to be distributive with respect to the addition and to satisfy the relation

$$a\vee b=a\wedge b+(a,b) \tag{1.7}$$

for all $a,b\in\bigwedge^1(V)$. The Clifford product is associative by definition. Hence the basic rule (1.7) is sufficient in order to work out the $\vee$ product of two arbitrary elements in $\bigwedge(V)$. Below we shall give a closed formula for this product.

The Atiyah–Kähler algebra combines the notions of an exterior algebra, a Grassmann algebra, and a Clifford algebra in an consistent manner. If we omit the Clifford product it reduces to the Grassmann algebra $\bigwedge(V,Q)$, while omitting both $\vee$ and $(\cdot,\cdot)$ yields the exterior algebra $\bigwedge(V)$. Without the structure of the $\wedge$ product it becomes a Clifford algebra because (1.7) entails $a\vee b+b\vee a=2(a,b)$.[11]

In the case at hand, $V=\bigwedge(T_x^*\mathcal{M}_n)$ and $Q=g'$. This means that for two basis one-forms the inner product is given by $(dx^\mu,dx^\nu)=g'(dx^\mu,dx^\nu)=g^{\mu\nu}$ and similarly for higher forms; for instance, $(dx^\mu\wedge dx^\nu,dx^\rho\wedge dx^\sigma)=g^{\mu\rho}g^{\nu\sigma}-g^{\mu\sigma}g^{\nu\rho}$.

A bundle over $\mathcal{M}_n$ with typical fiber $\mathbf{AK}(T_x^*\mathcal{M}_n,g')$ is called an Atiyah–Kähler bundle and sections through such bundles are referred to as Dirac–Kähler fields. Locally they are described by a collection of antisymmetric tensor fields $\{F^{(p)}_{\mu_1\cdots\mu_p},\ p=0,...,n\}$. The three products defined in the fiber give rise to analogous products on the space of sections, for instance $(\Phi_1\vee\Phi_2)(x)\equiv\Phi_1(x)\vee\Phi_2(x)$. Of course, also all the other operations of the conventional exterior calculus can be applied to Dirac–Kähler fields: the exterior derivative d, the coderivative $d^\dagger$, or the contraction with a vector field v, $\mathbf{i}(v)$, to mention just a few.

In our case the relations defining the Clifford product assume the following form when expressed in terms of the generating elements:

$$1 \vee 1 = 1, \quad 1 \vee dx^\mu = dx^\mu \vee 1 = dx^\mu,$$
$$dx^\mu \vee dx^\nu = dx^\mu \wedge dx^\nu + g^{\mu\nu}. \tag{1.8}$$

By virtue of the postulated associativity of the $\vee$ product, these relations are sufficient in order to determine the Clifford product of two arbitrary differential forms. One finds[1,3]

$$\Phi_1 \vee \Phi_2 = \sum_{p=0}^{n} \frac{(-1)^{p(p-1)/2}}{p!} (\mathcal{A}^p e_{\mu_1} \neg \cdots e_{\mu_p} \neg \Phi_1) \wedge (e^{\mu_1} \neg \cdots e^{\mu_p} \neg \Phi_2) \tag{1.9}$$

with $e_\mu \neg \equiv \mathbf{i}(\partial_\mu)$, $e^\mu \neg \equiv g^{\mu\nu} \mathbf{i}(\partial_\nu)$ where $\mathbf{i}(\partial_\mu)$ denotes the contraction with the basis vector ∂_μ. It is an antiderivation with the properties

$$\mathbf{i}(\partial_\mu) 1 = 0, \quad \mathbf{i}(\partial_\mu) dx^\nu = \delta_\mu^\nu,$$
$$\mathbf{i}(\partial_\mu)(\Phi_1 \wedge \Phi_2) = (\mathbf{i}(\partial_\mu)\Phi_1) \wedge \Phi_2 + (\mathcal{A}\Phi_1) \wedge \mathbf{i}(\partial_\mu)\Phi_2.$$

In writing down Eqs. (1.9) and (1.10) we used the ''main automorphism'' $\mathcal{A}$, a linear map whose action on the DK-field (1.6) is defined as

$$\mathcal{A}\Phi = \sum_{p=0}^{n} (-1)^p \Phi^{(p)}. \tag{1.10}$$

Later on we shall also need the ''main antiautomorphism'' $\mathcal{B}$ which acts according to

$$\mathcal{B}\Phi = \sum_{p=0}^{n} (-1)^{p(p-1)/2} \Phi^{(p)}. \tag{1.11}$$

Obviously, $\mathcal{A}^2 = \mathcal{B}^2 = 1$, $\mathcal{AB} = \mathcal{BA}$, and also

$$\mathcal{A}(\Phi_1 \wedge \Phi_2) = (\mathcal{A}\Phi_1) \wedge (\mathcal{A}\Phi_2),$$
$$\mathcal{B}(\Phi_1 \wedge \Phi_2) = (\mathcal{B}\Phi_2) \wedge (\mathcal{B}\Phi_1), \tag{1.12}$$

for any pair of DK fields.

As an important special case of (1.9) we note for later use that

$$dx^\mu \vee \Phi = dx^\mu \wedge \Phi + e^\mu \neg \Phi. \tag{1.13}$$

Let us look at the physical interpretation of the DK fields now. From now on we shall specialize the discussion to a flat space–time $\mathcal{M}_n = \mathbf{R}^n$ with the metric $g_{\mu\nu} = \delta_{\mu\nu}$. The generalization of a curved manifold and/or a manifold with Lorentzian signature would be straightforward, but we shall avoid these technical complications here since they are not important for the point we would like to make.

The interpretation of a DK field as a multiplet of Dirac spinors is based upon the following two logically independent observations.

(i) From (1.8) we obtain for the antisymmetrized Clifford product of two basis differentials

$$dx^\mu \vee dx^\nu + dx^\nu \vee dx^\mu = 2\delta^{\mu\nu}. \tag{1.14}$$

This relation should be compared to the one satisfied by the Euclidean Dirac matrices γ^μ:

$$\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2\delta^{\mu\nu}. \tag{1.15}$$

We conclude that the Clifford left multiplication with dx^μ defines a representation of the algebra of γ matrices in the space of (complex) inhomogeneous differential forms: $\gamma^\mu \triangleq dx^\mu \vee$. This representation is reducible though. Assuming n even from now on, a Dirac spinor has $2^{n/2}$ complex components, and an irreducible representation of the algebra (1.15) is in terms of $2^{n/2} \times 2^{n/2}$ matrices. On the other hand, the dimension of the exterior algebra is 2^n, i.e., a DK field Φ has 2^n independent complex component fields. We shall see in a moment that the space $\mathcal{K}$ of all DK fields Φ can be decomposed into $2^{n/2}$ subspaces $\mathcal{K}^{(\alpha)}$ which are invariant under Clifford left multiplication, $\mathcal{K} = \oplus_{\alpha=1}^{k} \mathcal{K}^{(\alpha)}$, $k \equiv 2^{n/2}$. On $\mathcal{K}^{(\alpha)}$, $dx^\mu \vee$ gives rise to an irreducible representation of the algebra (1.15).

(ii) From the exterior derivative d and its adjoint, the coderivative $d^\dagger$, we can form the so-called Dirac–Kähler operator $d-d^\dagger$ which has the property that it squares to the Laplacian:

$$(d-d^\dagger)^2 = -(dd^\dagger + d^\dagger d) = \partial_\mu \partial^\mu. \tag{1.16}$$

It shares this property with the Dirac operator $\gamma^\mu \partial_\mu$ and hence some relationship among the two might be expected. In fact, it turns out that the Dirac–Kähler operator can be expressed in terms of a Clifford multiplication from the left:

$$(d-d^\dagger)\Phi(x) = dx^\mu \vee \partial_\mu \Phi(x). \tag{1.17}$$

Since we know already that $dx^\mu \vee$ corresponds to a γ matrix and leaves the spaces $\mathcal{K}^{(\alpha)}$ invariant, we see that the Dirac–Kähler equation

$$(d-d^\dagger + m)\Phi = 0 \tag{1.18}$$

decomposes to a set of equations $(d-d^\dagger+m)\Phi^{(\alpha)}=0$, $\Phi^{(\alpha)} \in \mathcal{K}^{(\alpha)}$, each of which is equivalent to an ordinary Dirac equation $(\gamma^\mu \partial_\mu + m)\psi = 0$.

Following Becher and Joos[3] we can construct the invariant subspaces $\mathcal{K}^{(\alpha)}$ as follows. We introduce a new basis $\{Z_{\alpha\beta}\}$ in $\mathcal{K}$ whose elements are labeled by a pair of indices $\alpha,\beta = 1,...,2^{n/2}$ and which are required to satisfy

$$dx^\mu \vee Z_{\alpha\beta} = \sum_{\gamma=1}^{2^{n/2}} (\gamma^{\mu T})_{\alpha\gamma} Z_{\gamma\beta} \tag{1.19}$$

where the Euclidean Dirac matrices γ^μ are in the irreducible $2^{n/2}$-dimensional representation. (We use the notation $\mu,\nu,\ldots = 1,...,n$ for Lorentz indices and $\alpha,\beta,\gamma,\ldots = 1,...,2^{n/2}$ for spinor indices.) They satisfy (1.15) and are assumed to be Hermitian, $\gamma_\mu = \gamma_\mu^\dagger$. Frequently we shall regard $Z \equiv (Z_{\alpha\beta})$ as a matrix or, more precisely, as an inhomogeneous differential form which assumes values in the space of spinor matrices. Then (1.19) reads

$$dx^\mu \vee Z = \gamma^{\mu T} Z. \tag{1.20}$$

This equation is satisfied by

$$Z = \sum_{p=0}^{n} \frac{1}{p!} \gamma_{\mu_1}^T \cdots \gamma_{\mu_p}^T \, dx^{\mu_1} \wedge \cdots \wedge dx^{\mu_p}. \tag{1.21}$$

Every DK field Φ can be expanded in the basis $\{Z_{\alpha\beta}\}$:

$$\Phi(x) = \sum_{\alpha,\beta} \psi_\alpha^{(\beta)}(x) Z_{\alpha\beta}. \tag{1.22}$$

Hence it follows immediately from (1.19) that the invariant subspaces $\mathcal{K}^{(\alpha)}$ are spanned by

$$\Phi^{(\beta)} \equiv \sum_\alpha \psi_\alpha^{(\beta)} Z_{\alpha\beta} \in \mathcal{K}^{(\beta)}, \quad \beta \text{ fixed.} \tag{1.23}$$

In fact, one has

$$dx^\mu \vee \Phi^{(\beta)} = \sum_\alpha \left(\sum_\delta \gamma^\mu_{\alpha\delta} \psi_\delta^{(\beta)} \right) Z_{\alpha\beta}, \tag{1.24}$$

which shows that on $\mathcal{K}^{(\beta)}$ Clifford left-multiplication with dx^μ is equivalent to acting with the Dirac matrix γ^μ on the spinor $\psi^{(\beta)} \equiv \{\psi_\alpha^{(\beta)}; \alpha = 1,...,2^{n/2}\}$. For every fixed value of β, $\psi^{(\beta)}$ is an ordinary $2^{n/2}$-component Dirac field. By virtue of the orthogonal decomposition $\Phi = \Sigma_\beta \Phi^{(\beta)}$, a DK field describes a multiplet of $2^{n/2}$ Dirac fields.

It is convenient to combine the expansion coefficients $\psi_\alpha^{(\beta)}$ into a spinor matrix $\hat{\psi}$,

$$(\hat{\psi})_{\alpha\beta} \equiv \psi_\alpha^{(\beta)}, \tag{1.25}$$

so that (1.22) reads

$$\Phi(x) = \mathrm{Tr}[\hat{\psi}(x) Z^T]. \tag{1.26}$$

Writing $\hat{\psi}[\Phi]$ for the matrix related to a given DK field Φ, Eq. (1.24) amounts to

$$\hat{\psi}[dx^\mu \vee \Phi] = \gamma^\mu \hat{\psi}[\Phi]. \tag{1.27}$$

Occasionally one finds a slightly different approach in the literature.[2] One assumes that the inhomogeneous form (1.6) is given and one uses its coefficient functions $F^{(p)}_{\mu_1 \cdots \mu_p}$ in order to construct a spinor matrix $\hat{F}$ by simply replacing $dx^\mu \to \gamma^\mu$ everywhere:

$$\hat{F} \equiv \hat{F}[\Phi] \equiv \sum_{p=0}^{n} \frac{1}{p!} F^{(p)}_{\mu_1 \cdots \mu_p} \gamma^{\mu_1} \cdots \gamma^{\mu_p}. \tag{1.28}$$

Then one verifies that the map $\Phi \mapsto \hat{F}[\Phi]$ satisfies

$$\hat{F}[dx^\mu \vee \Phi] = \gamma^\mu \hat{F}[\Phi], \tag{1.29}$$

a property it has in common with $\hat{\psi}$. Hence we might expect that these two matrix-valued fields are related. Indeed, it turns out that they coincide up to a constant factor. To see this, one inserts the expansions (1.6) and (1.21) into (1.26) and obtains the following formula for the coefficients of Φ, $F^{(p)}_{\mu_1 \cdots \mu_p}$, as a function of $\hat{\psi}$:

$$F^{(p)}_{\mu_1 \cdots \mu_p}(x) = (-1)^{p(p-1)/2} \mathrm{Tr}[\hat{\psi}(x) \gamma_{[\mu_1} \cdots \gamma_{\mu_p]}]. \tag{1.30}$$

Because of the orthogonality and completeness relations enjoyed by the Dirac matrices, Eq. (1.30) has a unique solution for $\hat{\psi}$ as a function of the coefficients $F^{(p)}_{\mu_1 \cdots \mu_p}$ which define Φ. One finds that $\hat{\psi}$ and $\hat{F}$ are essentially the same thing:

$$\hat{\psi}(x) = 2^{-n/2} \sum_{p=0}^{n} \frac{1}{p!} F^{(p)}_{\mu_1 \cdots \mu_p}(x) \gamma^{\mu_1} \cdots \gamma^{\mu_p} \tag{1.31}$$

$$= 2^{-n/2} \hat{F}(x). \tag{1.32}$$

This formula together with (1.25) gives us a practical tool to compute the projection of Φ on the invariant subspaces $\mathcal{K}^{(\alpha)}$.

In standard discussions of Dirac–Kähler theory, because of the simple proportionality of $\hat{\psi}$ and $\hat{F}$, there is no need for a conceptual distinction between the two matrices. In order to establish their equivalence only familiar identities involving γ matrices such as

$$\mathrm{Tr}[\gamma^{[\mu_1}\cdots\gamma^{\mu_p]}(\gamma_{[\nu_1}\cdots\gamma_{\nu_q]})^\dagger]=2^{n/2}p!\,\delta^{pq}\delta^{\mu_1}_{[\nu_1}\cdots\delta^{\mu_p}_{\nu_q]} \tag{1.33}$$

are needed. In the symplectic case, the situation will be more complicated and we have to distinguish more carefully $\hat{\psi}$ which arises from the construction of left-invariant subspaces and $\hat{F}$ which obtains by replacing $dx^\mu\to\gamma^\mu$ in Φ. *A priori* it is not clear that the two objects can easily be related to each other since the metaplectic γ ''matrices'' are infinite dimensional. Hence the question whether there are trace identities analogous to (1.33) is a nontrivial issue.

II. DIRAC–KÄHLER FIELDS AND FERMIONIC WEYL SYMBOLS

In this section we describe the relation between the conventional Dirac–Kähler fermions and the Weyl symbol calculus. In Sec. II A we summarize various properties of the fermionic Weyl symbol calculus and discuss a number of special aspects and applications which will be relevant. In Sec. II B we show that the fermionic Weyl algebra $\mathcal{W}^F$ is an Atiyah–Kähler algebra, and in Sec. II C we introduce $\mathcal{W}^F$-valued fields over space–time. In Sec. II D we demonstrate that they can be identified with Dirac–Kähler fields. They carry a reducible representation of the Clifford algebra. The decomposition of $\mathcal{W}^F$ into invariant subspaces which carry an irreducible representation is performed in Sec. II E.

A. The fermionic Weyl algebra

We consider a set of operators $\hat{\chi}^\mu$, $\mu=1,...,n$, which satisfy the canonical anticommutation relations

$$\hat{\chi}^\mu\hat{\chi}^\nu+\hat{\chi}^\nu\hat{\chi}^\mu=\hbar\,\delta^{\mu\nu}. \tag{2.1}$$

We could think of the $\hat{\chi}$'s as world line fermions which represent the spin of a relativistic particle, for instance.[17,26] The most general operator we can construct by forming linear combinations of products of $\hat{\chi}$'s has the structure

$$\hat{f}=\sum_{p=0}^{n}\frac{1}{p!}f^{(p)}_{\mu_1\cdots\mu_p}\hat{\chi}^{\mu_1}\hat{\chi}^{\mu_2}\cdots\hat{\chi}^{\mu_p} \tag{2.2}$$

with arbitrary (complex-valued) constants $f^{(p)}_{\mu_1\cdots\mu_p}$.

We would like to establish a linear one-to-one correspondence between the operators (2.2) and functions f depending on Grassmann numbers $\theta^1,\theta^2,\ldots,\theta^n$ with $\theta^\mu\theta^\nu+\theta^\nu\theta^\mu=0$. The function $f(\theta)$ which characterizes the operator $\hat{f}$ is called the symbol of $\hat{f}$: $f=\mathrm{symb}(\hat{f})$. There are many ''symbol maps'' which relate operators to classical functions. Here we are interested in the Weyl symbol which is defined as follows. Given the operator (2.2) we define

$$f(\theta)=[\mathrm{symb}(\hat{f})](\theta)=\sum_{p=0}^{n}\frac{1}{p!}f^{(p)}_{\mu_1\cdots\mu_p}\theta^{\mu_1}\theta^{\mu_2}\cdots\theta^{\mu_p} \tag{2.3}$$

which for a given ordering is a well-defined map from operators to functions. In particular, $\mathrm{symb}(\hat{\chi}^\mu)=\theta^\mu$ and $\mathrm{symb}(I)=1$ where I is the unit operator. The inverse mapping is not well defined yet, because in (2.3) we can add to $f^{(p)}_{\mu_1\cdots\mu_p}$ arbitrary tensors which are symmetric in at least one index pair. This does not change $f(\theta)$, but it does change $\hat{f}$. Specifying a unique operator $\hat{f}$ for a given $f(\theta)$ amounts to picking a particular operator ordering prescription. In fact,

$f(\theta)$ can be regarded as a classical phase function of a mechanical system with Grassmann-odd phase-space coordinates θ^μ, and the $\hat{\chi}^\mu$'s are the corresponding quantum operators. We shall employ the Weyl correspondence rule which means that $\hat{f}$ follows from $f(\theta)$ by substituting $\theta^\mu \to \hat{\chi}^\mu$ in (2.3) and writing all operator products in Weyl ordered, i.e., completely antisymmetrized form. For instance, the product $\theta^\mu \theta^\nu$ yields the operator $[\hat{\chi}^\mu \hat{\chi}^\nu]_{\text{Weyl}} = \frac{1}{2}(\hat{\chi}^\mu \hat{\chi}^\nu - \hat{\chi}^\nu \hat{\chi}^\mu) \equiv \hat{\chi}^{[\mu} \hat{\chi}^{\mu]}$. For an arbitrary monomial,

$$\text{symb}^{-1}(\theta^{\mu_1} \cdots \theta^{\mu_p}) = \hat{\chi}^{[\mu_1} \cdots \hat{\chi}^{\mu_p]} \tag{2.4}$$

where the square brackets indicate complete antisymmetrization.

From now on we shall require the constants $f^{(p)}_{\mu_1 \cdots \mu_p}$ appearing in the series expansion of the symbol $f(\theta)$ to be completely antisymmetric tensors. Then the operator $\hat{f}$ associated with the series (2.3) is obtained by simply replacing $\theta^\mu \to \hat{\chi}^\mu$ in this series, and this leads us back to the operator (2.2).

If n is odd, the inverse symbol map is still not uniquely defined, because in this case the operator $\hat{\chi}^1 \hat{\chi}^2 \cdots \hat{\chi}^n$ commutes with all operators and is proportional to the identity therefore. By multiplying any operator by $\hat{\chi}^1 \hat{\chi}^2 \cdots \hat{\chi}^\nu$ if necessary one can represent all operators by even symbols. This prescription makes the correspondence between operators and symbols bijective. (See Refs. 26 and 27 for further details.)

If a string of operators is not contracted with an antisymmetric tensor we must reorder it before we can use

$$\text{symb}(\hat{\chi}^{[\mu_1} \cdots \hat{\chi}^{\mu_p]}) = \theta^{\mu_1} \cdots \theta^{\mu_p} \tag{2.5}$$

in order to read off its symbol. For instance,

$$\text{symb}(\hat{\chi}^\mu \hat{\chi}^\nu) = \text{symb}\left[\hat{\chi}^{[\mu} \hat{\chi}^{\nu]} + \frac{\hbar}{2} \delta^{\mu\nu} \right] = \theta^\mu \theta^\nu + \frac{\hbar}{2} \delta^{\mu\nu}. \tag{2.6}$$

The symbols $f(\theta)$ are functions of the same type as those considered in Appendix A, to which the reader might turn at this point. Among other things, various linear operations on such functions are discussed there which are particularly useful in the context of the symbol calculus. This includes the ''main automorphism'' $\mathcal{A}$, the ''main antiautomorphism'' $\mathcal{B}$, the Hodge operator $*$, and the modified Hodge operator $\star$.

While we allow for complex coefficients $f^{(p)}_{\mu_1 \cdots \mu_p}$, we assume that the operators $\hat{\chi}^\mu$ are Hermitian, $\hat{\chi}^\mu = (\hat{\chi}^\mu)^\dagger$, and that θ^μ is real, $\overline{\theta^\mu} = \theta^\mu$. Hence it follows that

$$\text{symb}(\hat{f}^\dagger) = \overline{\text{symb}(\hat{f})}, \tag{2.7}$$

where the overbar means complex conjugation.

There is a simple integral formula for the operator $\hat{f}$ associated with a given Weyl symbol $f(\theta)$:

$$\hat{f} = \int \hat{\Omega}(\rho) \tilde{f}(\rho) d^n \rho. \tag{2.8}$$

Here $\tilde{f}(\rho)$ is the Fourier transform of $f(\theta)$ as defined in Eq. (A22) of the Appendix, and

$$\hat{\Omega}(\rho) \equiv \exp(-i \hat{\chi}^\mu \rho_\mu) \tag{2.9}$$

is the fermionic analogue of the Weyl operators which implement translations on phase space. Using the identities of Appendix A one can verify that Eqs. (2.4) and (2.8) are indeed equivalent.

An important concept is the ''star product''[3] or ''twisted product'' which mimics the multiplication of operators at the level of symbols. [Both in the fermionic and the bosonic case we keep using the traditional name ''star product'' even though we write ''$\circ$'' instead of the usual symbol ''$*$.'' Following Refs. 29, 30, and 21 this notation indicates that we are dealing with a *fiberwise* twisted product which should not be confused with the $*\equiv *_{\mathcal{M}}$ product which would refer to the *base* of the Weyl algebra bundles we are going to construct in Sec. II B. It is the $*_{\mathcal{M}}$ product rather than the $\circ$ product which is needed for the deformation quantization[24] of physical systems on the phase-space $\mathcal{M}$. In the present paper, the $*_{\mathcal{M}}$ product plays no role, however. (Note also that ''$*$'' stands for the Hodge operator in our case.)] It satisfies

$$\mathrm{symb}(\hat{f}\hat{g})=\mathrm{symb}(\hat{f})\circ\mathrm{symb}(\hat{g}) \tag{2.10}$$

for all operators $\hat{f}$ and $\hat{g}$. As a consequence, the $\circ$ product is associative, distributive with respect to $+$, but not commutative. It is a deformation of the pointwise product of functions to which it reduces in the limit $\hbar\to 0$. From

$$\mathrm{symb}(I)=1, \quad \mathrm{symb}(\hat{\chi}^\mu)=\theta^\mu \tag{2.11}$$

and Eq. (2.6) it follows that

$$\begin{aligned} 1\circ 1=1, \quad 1\circ\theta^\mu=\theta^\mu\circ 1=\theta^\mu, \\ \theta^\mu\circ\theta^\nu=\theta^\mu\theta^\nu+\frac{\hbar}{2}\delta^{\mu\nu}. \end{aligned} \tag{2.12}$$

By virtue of its postulated distributivity and associativity, the relations (2.12) characterize the $\circ$ product uniquely. They are sufficient to work out the product $f\circ g$ of arbitrary functions f and g.

Equation (2.7) implies that complex conjugation changes the order of the factors in a star product:

$$\overline{f_1\circ f_2}=\overline{f_2}\circ\overline{f_1}. \tag{2.13}$$

The space of functions $f(\theta)$ equipped with the $\circ$ product will be referred to as the *fermionic Weyl algebra* $\mathcal{W}^F$.

In the literature[26,27] one finds the following integral representation for the $\circ$ product of two arbitrary functions:

$$(f_1\circ f_2)(\theta)=\epsilon_n\left(\frac{\hbar}{2i}\right)^n\int \exp\left[-\frac{2}{\hbar}(\theta_1\theta+\theta\theta_2+\theta_2\theta_1)\right]f_1(\theta_1)f_2(\theta_2)d^n\theta_1 d^n\theta_2, \tag{2.14}$$

where (indices are raised and lowered with $g_{\mu\nu}=\delta_{\mu\nu}$) $\theta_1\theta\equiv\theta_1^\mu\theta_\mu$, etc., and with ϵ_n defined as in Eq. (A23). For our purposes, various alternative representations of the star product are needed. They are derived in appendix B. The first one reads

$$(f_1\circ f_2)(\theta)=\sum_{p=0}^{n}\left(\frac{\hbar}{2}\right)^p\frac{(-1)^{p(p-1)/2}}{p!}\left[\mathcal{A}^p\frac{\partial}{\partial\theta^{\mu_1}}\frac{\partial}{\partial\theta^{\mu_2}}\cdots\frac{\partial}{\partial\theta^{\mu_p}}f_1(\theta)\right]\left[\frac{\partial}{\partial\theta_{\mu_1}}\frac{\partial}{\partial\theta_{\mu_2}}\cdots\frac{\partial}{\partial\theta_{\mu_p}}f_2(\theta)\right] \tag{2.15}$$

with the automorphism $\mathcal{A}:\mathcal{W}^F\to\mathcal{W}^F$ defined in Appendix A. An equivalent form involving both left and right derivatives is

$$(f_1\circ f_2)(\theta)=\sum_{p=0}^{n}\left(\frac{\hbar}{2}\right)^p\frac{1}{p!}f_1(\theta)\frac{\overleftarrow{\partial}}{\partial\theta^{\mu_p}}\frac{\overleftarrow{\partial}}{\partial\theta^{\mu_{p-1}}}\cdots\frac{\overleftarrow{\partial}}{\partial\theta^{\mu_1}}\frac{\partial}{\partial\theta_{\mu_1}}\frac{\partial}{\partial\theta_{\mu_2}}\cdots\frac{\partial}{\partial\theta_{\mu_p}}f_2(\theta). \tag{2.16}$$

The most compact representation reads

$$(f_1 \circ f_2)(\theta) = f_1(\theta) \exp\left[\frac{\hbar}{2} \frac{\overleftarrow{\partial}}{\partial\theta^\mu} \frac{\vec{\partial}}{\partial\theta_\mu}\right] f_2(\theta), \tag{2.17}$$

where $\overleftarrow{\partial}/\partial\theta^\mu$ is a right derivative acting on f_1 and $\vec{\partial}/\partial\theta^\mu$ a left derivative acting on f_2. The result (2.17) looks surprisingly simple and is completely analogous to its bosonic counterpart. All the complicated sign factors which appeared during the calculation, either explicitly or hidden in the $\mathcal{A}$ automorphism, conspired to disappear from the final result.

Depending on the problem at hand one or another of the above representations is the most convenient one. Equation (2.15) we shall relate to Kähler's formula for the Clifford product shortly. From Eq. (2.16) one immediately reads off the important special cases

$$\theta^\mu \circ f(\theta) = \theta^\mu f(\theta) + \frac{\hbar}{2} \frac{\partial}{\partial\theta_\mu} f(\theta), \tag{2.18}$$

$$f(\theta) \circ \theta^\mu = f(\theta)\theta^\mu + \frac{\hbar}{2} f(\theta) \frac{\overleftarrow{\partial}}{\partial\theta_\mu}. \tag{2.19}$$

In order to calculate the product of two δ functions, which we shall need later on, Eq. (2.14) is most suitable:

$$(\delta \circ \delta)(\theta) = (-1)^{n(n-1)/2} \left(\frac{\hbar}{2}\right)^n. \tag{2.20}$$

Up to now we regarded the $\hat{\chi}^\mu$'s as abstract operators. Let us look at concrete representations on some finite dimensional vector space $\mathcal{V}^F$. If Γ_μ is a set of Hermitian matrices which satisfy the Clifford algebra relations

$$\Gamma^\mu \Gamma^\nu + \Gamma^\nu \Gamma^\mu = 2\delta^{\mu\nu}, \tag{2.21}$$

then

$$\hat{\chi}^\mu = \sqrt{\frac{\hbar}{2}} \Gamma^\mu \tag{2.22}$$

satisfies the canonical anticommutation relations (2.1). Here Γ^μ denotes the Dirac matrices in an arbitrary, possibly reducible representation. The notation γ^μ is reserved for the (essentially unique) irreducible representation on $\mathcal{V}^F = \mathbf{C}^k$, $k \equiv 2^{n/2}$, if n is even. The operators $\hat{f}: \mathcal{W}^F \to \mathcal{W}^F$ of Eq. (2.2) are $k \times k$ matrices then. The space of these operators will be denoted by $\mathcal{L}(\mathcal{W}^F)$.

The identification (2.22) must be interpreted with some care. In setting up the symbol calculus one adopts the rule that the operators $\hat{\chi}^\mu$ anticommute with numbers of odd Grassmann parity. On the other hand, the entries of the matrices Γ^μ are ordinary complex numbers, so Γ^μ commutes with all elements of the Grassmann algebra.

Next we list a few properties of the operators $\hat{\Omega}$ which we shall need shortly. These operators are reminiscent of the (bosonic) Weyl operators. However, as they stand, they are not unitary but rather Hermitian, $\Omega(\rho) = \Omega(\rho)^\dagger$. This is due to the fact that $\hat{\chi}^\mu$ anticommutes with the Grassmann-odd ρ_μ's. Actually it is the operators $\hat{\Omega}(i\rho/\hbar) = \exp(\hat{\chi}^\mu \rho_\mu/\hbar)$ which play the role of the Weyl operators on a fermionic phase space. They are unitary, $\hat{\Omega}(i\rho/\hbar)^\dagger = \hat{\Omega}(i\rho/\hbar)^{-1}$, and they shift $\hat{\chi}^\mu$ by ρ^μ times the unit operator:

$$\hat{\Omega}(i\rho/\hbar)^\dagger \hat{\chi}^\mu \hat{\Omega}(i\rho/\hbar) = \hat{\chi}^\mu + \rho^\mu. \tag{2.23}$$

This leads to a projective representation of the translation group since

$$\hat{\Omega}(\rho_1)\hat{\Omega}(\rho_2)=\exp\left(\frac{\hbar}{2}\rho_1^\mu\rho_{2\mu}\right)\hat{\Omega}(\rho_1+\rho_2). \tag{2.24}$$

The derivative of $\hat{\Omega}(\rho)$ can be written in either of the two forms

$$\frac{\partial}{\partial\rho_\mu}\hat{\Omega}(\rho)=\hat{\Omega}(\rho)\left[i\hat{\chi}^\mu+\frac{\hbar}{2}\rho^\mu\right] \tag{2.25}$$

$$=\left[i\hat{\chi}^\mu-\frac{\hbar}{2}\rho^\mu\right]\hat{\Omega}(\rho). \tag{2.26}$$

When we replace in $\hat{\Omega}$ the operators $\hat{\chi}^\mu$ by the Dirac matrices via (2.22) we are led to

$$\check{\Omega}(\rho)=\exp(-i\sqrt{\hbar/2}\Gamma^\mu\rho_\mu). \tag{2.27}$$

The properties of $\check{\Omega}$ are slightly different from those of $\hat{\Omega}$ because Γ^μ commutes with ρ_μ. The $\check{\Omega}$'s are unitary matrices,

$$\check{\Omega}(\rho)^\dagger=\check{\Omega}(-\rho)=\check{\Omega}(\rho)^{-1} \tag{2.28}$$

with the composition law

$$\check{\Omega}(\rho_1)\check{\Omega}(\rho_2)=\exp\left(-\frac{\hbar}{2}\rho_1^\mu\rho_{2\mu}\right)\check{\Omega}(\rho_1+\rho_2). \tag{2.29}$$

The expressions for their derivative are

$$\frac{\partial}{\partial\rho_\mu}\check{\Omega}(\rho)=\check{\Omega}(-\rho)\left[-i\sqrt{\frac{\hbar}{2}}\Gamma^\mu-\frac{\hbar}{2}\rho^\mu\right] \tag{2.30}$$

$$=\left[-i\sqrt{\frac{\hbar}{2}}\Gamma^\mu+\frac{\hbar}{2}\rho^\mu\right]\check{\Omega}(\rho). \tag{2.31}$$

We shall need these relations when we decompose the reducible Dirac–Kähler representation.

An interesting example where one can see the symbol calculus at work is the generalization of the chirality matrix γ_5 in four dimensions. We assume that n is even in the remainder of this section and employ the Dirac matrices γ^μ in the $2^{n/2}$-dimensional representation. From (1.15) and $(\gamma^\mu)^\dagger=\gamma^\mu$ it follows that the matrix

$$\gamma_{n+1}\equiv-i^{n(n-1)/2}\gamma^1\gamma^2\cdots\gamma^n \tag{2.32}$$

satisfies $\gamma^\mu\gamma_{n+1}=-\gamma_{n+1}\gamma^\mu$,

$$\gamma_{n+1}^2=1,\quad \gamma_{n+1}^\dagger=\gamma_{n+1} \tag{2.33}$$

in all even dimensions. The sign of (2.32) is chosen such that for $n=4$

$$\gamma_5=\gamma^1\gamma^2\gamma^3\gamma^4. \tag{2.34}$$

We identify

$$\gamma^\mu=\kappa\hat{\chi}^\mu, \tag{2.35}$$

where

$$\kappa \equiv \sqrt{\frac{2}{\hbar}} \tag{2.36}$$

and interpret γ_{n+1} as the matrix representation of the abstract operator

$$\hat{G}_{n+1} = -i^{n(n-1)/2} \kappa^n \hat{\chi}^1 \hat{\chi}^2 \cdots \hat{\chi}^n. \tag{2.37}$$

Its symbol $\text{symb}(\hat{G}_{n+1}) \equiv G_{n+1}$ follows directly from (2.5) if we note that $\hat{\chi}^1 \cdots \hat{\chi}^n = \epsilon_{\mu_1 \cdots \mu_n} \hat{\chi}^{[\mu_1} \cdots \hat{\chi}^{\mu_n]}/(n!)$:

$$\begin{aligned} G_{n+1}(\theta) &= -i^{n(n-1)/2} \kappa^n \theta^1 \theta^2 \cdots \theta^n \\ &= -(-i)^{n(n-1)/2} \kappa^n \theta^n \theta^{n-1} \cdots \theta^1. \end{aligned} \tag{2.38}$$

Hence, up to a constant, γ_{n+1} is represented by the δ function:

$$G_{n+1}(\theta) = -(-i)^{n(n-1)/2} \kappa^n \delta(\theta). \tag{2.39}$$

As a consequence of Eqs. (2.20) and (A21), this function satisfies

$$G_{n+1} \circ G_{n+1} = 1, \quad \overline{G_{n+1}} = G_{n+1}, \tag{2.40}$$

which reflects the properties (2.33) of γ_{n+1}. By virtue of (A26) the Fourier transform of G_{n+1} is the constant function

$$\tilde{G}_{n+1}(\rho) = -(-i)^{n(n-1)/2} \kappa^n. \tag{2.41}$$

In Appendix A we defined the modified Hodge operator $\star$ for a general Grassmann algebra and we showed that it is related to the Fourier transformation via Eq. (A41). Using the latter equation together with the integral representation (2.14) for the star product it is not difficult to see that the application of $\star$ to some $f \in \mathcal{W}^F$ is essentially equivalent to a star-multiplication with G_{n+1} from the right. For a homogeneous function of degree p,

$$\star f^{(p)}(\theta) = -(-i)^{n(n-1)/2} \kappa^{2p-n} (f^{(p)} \circ G_{n+1})(\theta). \tag{2.42}$$

If we rescale θ we can write down a similar equation for inhomogeneous functions even:

$$\star f(\theta/\kappa) = -(-i)^{n(n-1)/2} (f \circ G_{n+1})(\theta/\kappa). \tag{2.43}$$

Finally we remark that the chirality operator $\hat{G}_{n+1}$ can be expressed as an integral over the Weyl operators:

$$\hat{G}_{n+1} = -(-i)^{n(n-1)/2} \kappa^n \int d^n\rho\, \hat{\Omega}(\rho). \tag{2.44}$$

This is a remarkable relation because contrary to the original definition of γ_{n+1} as the product of all Dirac matrices it carries over to the symplectic case almost literally.

B. $\mathcal{W}^F$ as an Atiyah–Kähler algebra

Let us come back to the abstract Atiyah–Kähler algebra $\mathbf{AK}(V,Q)$ discussed in Sec. I. It is important to observe that the Weyl algebra $\mathcal{W}^F$ which we reviewed in the previous section contains all the ingredients which make up an Atiyah–Kähler algebra.

(i) The vector space V is spanned by the basis elements $\theta^1, \ldots, \theta^n$ and the exterior algebra over this space, $\bigwedge(V) = \oplus_{p=0}^n \bigwedge^p(V)$, consists of monomials $\theta^{\mu_1} \cdots \theta^{\mu_p} \in \bigwedge^p(V)$. The exterior product "$\wedge$" on $\bigwedge(V)$ is the pointwise product of (inhomogeneous) functions $f(\theta) \in \bigwedge(V)$.

(ii) By starting from the canonical anticommutation relations (2.1) we have tacitly decided for an inner product $(\cdot|\cdot)$ on $\bigwedge(V)$. The quadratic form Q is induced by $g_{\mu\nu}\equiv\delta_{\mu\nu}$, regarded as an inner product of V. On $\bigwedge^1(V)$ we have

$$(\theta^\mu|\theta^\nu)=\kappa^{-2}\delta^{\mu\nu} \tag{2.45}$$

and similarly for $p>1$ (see below).

(iii) The star product on $\mathcal{W}^F$ provides a concrete realization of the abstract Clifford product. The Clifford product is associative and distributive over "+," and so is the star product. Moreover, $\vee$, $\wedge$, and $(\cdot,\cdot)$ have to satisfy the consistency condition (1.7). From Eq. (2.12) it follows that this relation is indeed satisfied by the star multiplication together with the pointwise multiplication and the inner product $(\cdot|\cdot)$:

$$\theta^\mu\circ\theta^\nu=\theta^\mu\theta^\nu+(\theta^\mu|\theta^\nu). \tag{2.46}$$

In particular, upon symmetrization, $\theta^\mu\circ\theta^\nu+\theta^\nu\circ\theta^\mu=2(\theta^\mu|\theta^\nu)$.

Thus we may conclude that the *fermionic Weyl algebra* $\mathcal{W}^F$ *is a concrete realization of an Atiyah–Kähler algebra*.

Let us be more explicit about the inner product on $\bigwedge(V)$. Within the symbol calculus, the standard inner product of DK theory[11] admits a very natural representation in terms of the star product:

$$(f_1|f_2)=[\overline{f_1}\circ f_2](\theta=0). \tag{2.47}$$

Here f_1 and f_2 are two arbitrary inhomogeneous functions. We allow their expansion coefficients $f^{(p)}_{\mu_1\cdots\mu_p}$ to become complex. Note, however, that the complex conjugation in (2.47) is necessary even if the coefficients are taken to be real, see Eq. (A10). Using the integral representation

$$(f_1|f_2)=(i\kappa^2)^{-n}\epsilon_n\int\overline{f_1(\theta_1)}\exp(\kappa^2\theta_1^\mu\theta_{2\mu})f_2(\theta_2)d^n\theta_1d^n\theta_2 \tag{2.48}$$

and expanding f_1 and f_2 according to

$$f(\theta)=\sum_{p=0}^{n}\frac{1}{p!}\kappa^p f^{(p)}_{\mu_1\cdots\mu_p}\theta^{\mu_1}\cdots\theta^{\mu_p} \tag{2.49}$$

with appropriate powers of κ separated off from the expansion coefficients, it is easy to derive that

$$(f_1|f_2)=\sum_{p=0}^{n}\frac{1}{p!}\overline{f^{(p)}_{1\mu_1\cdots\mu_p}}f_2^{(p)\mu_1\cdots\mu_p}. \tag{2.50}$$

The inner products among the basis elements of $\bigwedge^p(V)$ (homogeneous functions of degree p) can be written down similarly. For $p=1$ one recovers (2.45), and for $p=2$ one has, for instance

$$(\theta^\mu\theta^\nu|\theta^\rho\theta^\sigma)=\kappa^{-4}(\delta^{\mu\rho}\delta^{\nu\sigma}-\delta^{\mu\sigma}\delta^{\nu\rho}). \tag{2.51}$$

We note that $(\cdot|\cdot)$ has the important property of making the star multiplication with θ^μ a self-adjoint operator. If we define

$$C^\mu:\mathcal{W}^F\rightarrow\mathcal{W}^F,\quad (C^\mu f)(\theta)=\kappa\theta^\mu\circ f(\theta) \tag{2.52}$$

then Eq. (2.18) tells us that C^μ is given by the first-order differential operator

$$C^\mu=\kappa\theta^\mu+\frac{1}{\kappa}\frac{\partial}{\partial\theta_\mu}. \tag{2.53}$$

If one writes the inner product as in (2.47), the self-adjointness of C^μ is obvious:

$$(C^\mu f_1|f_2)=\kappa[\overline{(\theta^\mu\circ f_1)}\circ f_2](0)=\kappa[(\overline{f_1}\circ\theta^\mu)\circ f_2](0)=\kappa[\overline{f_1}\circ(\theta^\mu\circ f_2)](0)=(f_1|C^\mu f_2). \tag{2.54}$$

Here we exploited (2.13) and the associativity of the star product.

C. Symbol-valued fields on space–time

The most familiar application of the above symbol calculus is the deformation theory approach[24,26] to the quantization of fermionic systems. In this context, the variables θ^μ are coordinates on the *phase space* of the physical system under consideration. If there are additional bosonic degrees of freedom (such as the position of a spinning particle, say) this fermionic phase space is embedded in a larger graded phase space, a supermanifold with both commuting and anticommuting coordinates.[17]

In the present paper we are investigating a different setting. Rather than phase space, the physical arena here is *space–time*, an ordinary Riemannian manifold $(\mathcal{M}_n,g)$, not a supermanifold. The fermionic Weyl algebra $\mathcal{W}^F$ enters the construction as the fiber of certain bundles over space–time which we shall refer to as ''Weyl algebra bundles.''[30]

By definition, the base of a Weyl algebra bundle is $(\mathcal{M}_n,g)$ and the typical fiber is $\mathcal{W}^F$, i.e., at each space–time point x we attach a local copy $\mathcal{W}_x^F$ of $\mathcal{W}^F$. The quadratic form Q on $\mathcal{W}_x^F$ is provided by the metric g evaluated at the point x. Local coordinates on the total space are pairs (x,f) where $x\equiv(x^\mu)$ are coordinates referring to some chart of $\mathcal{M}_n$, and f is a function of the Grassmann variables $\theta^1,\ldots,\theta^n$. The transition functions are defined in close analogy with the exterior algebra bundle. A coordinate transformation $x\to\tilde{x}(x)$ on $\mathcal{M}_n$ is accompanied by $f\to\tilde{f}$ with $\tilde{f}$ such that $\tilde{f}(\tilde{\theta})=f(\theta)$ where $\tilde{\theta}^\mu\equiv(\partial\tilde{x}^\mu/\partial x^\nu)\theta^\nu$, i.e., θ^μ transforms in the same manner as dx^μ.

Sections through a Weyl algebra bundle are locally represented by functions

$$x\mapsto F(x,\cdot)\in\mathcal{W}_x^F, \tag{2.55}$$

where

$$F(x,\cdot):\theta\mapsto F(x,\theta) \tag{2.56}$$

is a function of n commuting and n anticommuting variables. We define a fiberwise star product of two such sections by

$$(F_1\circ F_2)(x,\theta)=(F_1(x,\cdot)\circ F_2(x,\cdot))(\theta) \tag{2.57}$$

for each point x.

We can apply the inverse symbol map to $F(x,\cdot)$ and thus obtain a family of operators labeled by the space–time points x:

$$\hat{F}(x)\equiv\mathrm{symb}^{-1}F(x,\cdot). \tag{2.58}$$

If we fix a concrete matrix representation of the fermionic operators on some representation space $\mathcal{V}^F$, then $\hat{F}(x)$ acts on a local copy $\mathcal{V}_x^F$ of $\mathcal{V}^F$, i.e., $\hat{F}(x)\in\mathcal{L}(\mathcal{V}_x^F)$. We are particularly interested in the case where $\mathcal{V}^F$ carries the irreducible $2^{n/2}$-dimensional representation of the Clifford algebra (for n even). Then $\mathcal{V}_x^F$ is a fiber of the usual spin bundle over $\mathcal{M}_n$ whose sections are the familiar Dirac spinor fields.

In the present paper we shall not be concerned with the global properties of Weyl algebra bundles. Our main interest is in the metaplectic analog of the Dirac–Kähler construction, and for this purpose it is sufficient to compare to the topologically trivial bundles over the flat space–time $\mathcal{M}_n=\mathbf{R}^n$. An analogous discussion could be given for arbitrary curved space–times as well, but we shall avoid the necessary technical complications here. Thus, in our case, sections can be represented globally by functions $F(x,\theta)$. We remark that there exists a natural inner product on the space of these functions:

$$\langle F_1|F_2\rangle = \int d^n x (F_1(x,\cdot)|F_2(x,\cdot)). \tag{2.59}$$

D. Dirac–Kähler fields and symbol calculus

Let us assume we are given an arbitrary DK field Φ on flat Euclidean space–time $\mathbf{R}^n$. It possesses an expansion

$$\Phi(x) = \sum_{p=0}^{n} \frac{1}{p!} F^{(p)}_{\mu_1\cdots\mu_p}(x) dx^{\mu_1}\wedge\cdots\wedge dx^{\mu_p}. \tag{2.60}$$

The (complex) coefficient functions are taken to be completely antisymmetric in all p indices so that there is a bijective correspondence between forms Φ and sets $\{F^{(p)}_{\mu_1\cdots\mu_p}\}$ of antisymmetric tensors. Given these tensors, we form the following matrix-valued field:

$$\hat{F}(x) = \sum_{p=0}^{n} \frac{1}{p!} F^{(p)}_{\mu_1\cdots\mu_p}(x) \gamma^{\mu_1}\gamma^{\mu_2}\cdots\gamma^{\mu_p}. \tag{2.61}$$

From now on we assume that n is even and that the Dirac matrices are in their irreducible representation. Hence $\hat{F}(x)$ acts on a local copy $\mathcal{V}^F_x$ of the representation space $\mathcal{V}^F = \mathbf{C}^k$, $k = 2^{n/2}$. By virtue of (2.35) we may regard $\hat{F}(x)$ as a matrix realization of the abstract operator

$$\hat{F}(x) = \sum_{p=0}^{n} \frac{\kappa^p}{p!} F^{(p)}_{\mu_1\cdots\mu_p}(x) \hat{\chi}^{\mu_1}\hat{\chi}^{\mu_2}\cdots\hat{\chi}^{\mu_p}. \tag{2.62}$$

For every point x, the symbol of this operator is $F(x,\theta) = [\mathrm{symb}\,\hat{F}(x)](\theta)$, or

$$F(x,\theta) = \sum_{p=0}^{n} \frac{\kappa^p}{p!} F^{(p)}_{\mu_1\cdots\mu_p}(x) \theta^{\mu_1}\theta^{\mu_2}\cdots\theta^{\mu_p}. \tag{2.63}$$

Thus we have set up a linear one-to-one correspondence between differential forms $\Phi(x)$ and symbol functions $F(x,\theta)$. Schematically,

$$\Phi(x) \in \bigwedge(T^*_x\mathcal{M}_n) \leftrightarrows \hat{F}(x) \in \mathcal{L}(\mathcal{V}^F_x) \leftrightarrows F(x,\cdot) \in \mathcal{W}^F_x. \tag{2.64}$$

The first one of the two bijections in (2.64) is the usual ''Dirac–Kähler correspondence'' $dx^\mu \leftrightarrows \gamma^\mu$ which we mentioned already in Sec. I, while the second one is the Weyl symbol map. Taken in conjunction, these maps relate DK fields to symbols. In particular,

$$dx^\mu \leftrightarrows \kappa\theta^\mu. \tag{2.65}$$

We shall use the notation $\Phi: F \mapsto \Phi[F]$ for the linear map which yields the differential form belonging to a given symbol. For instance,

$$\Phi[\kappa\theta^\mu] = dx^\mu. \tag{2.66}$$

What makes the above construction particularly useful is that under the map Φ many of the familiar linear and bilinear operations involving differential forms naturally pass over to the symbol functions and vice versa. This is immediately obvious for the automorphism $\mathcal{A}$, the antiautomorphism $\mathcal{B}$, the Hodge operator $*$, and the modified Hodge operator $\star$. Comparing their definition for symbols in Appendix A to their standard definition in terms of differential forms one sees that

$$\mathcal{A}\Phi[F]=\Phi[\mathcal{A}F], \quad \mathcal{B}\Phi[F]=\Phi[\mathcal{B}F],$$
$$*\Phi[F]=\Phi[*F], \quad \star\Phi[F]=\Phi[\star F]. \tag{2.67}$$

The exterior derivative $d=dx^\mu\partial_\mu$ translates into $\kappa\theta^\mu\partial_\mu$,

$$d\Phi[F]=\Phi[\kappa\theta^\mu\partial_\mu F] \tag{2.68}$$

while the contraction $\mathbf{i}(v)$ with a vector field $v=v^\mu\partial_\mu$ becomes a derivative with respect to θ:

$$\mathbf{i}(v)\Phi[F]=\Phi\left[\kappa^{-1}v^\mu\frac{\partial}{\partial\theta^\mu}F\right]. \tag{2.69}$$

In particular,

$$e_\mu\neg\Phi[F]=\Phi\left[\kappa^{-1}\frac{\partial}{\partial\theta^\mu}F\right]. \tag{2.70}$$

The natural inner product on the space of DK fields is[5]

$$\langle\Phi_1,\Phi_2\rangle=\int\tilde{\Phi}_1\wedge*\Phi_2. \tag{2.71}$$

[All terms which are not of degree n are supposed to be discarded from the integrand in (2.71).] Its counterpart at the symbol level is (2.59) with (2.47):

$$\langle\Phi[F_1],\Phi[F_2]\rangle=\langle F_1|F_2\rangle. \tag{2.72}$$

The coderivative $d^\dagger$ is the formal adjoint of d with respect to $\langle\cdot,\cdot\rangle$. On flat space one has

$$d^\dagger\Phi=-e^\mu\neg\partial_\mu\Phi \tag{2.73}$$

whence

$$d^\dagger\Phi[F]=\Phi\left[-\kappa^{-1}\partial_\mu\frac{\partial}{\partial\theta_\mu}F\right]. \tag{2.74}$$

The wedge product of differential forms is mapped onto the pointwise product of symbol functions:

$$\Phi[F_1]\wedge\Phi[F_2]=\Phi[F_1F_2]. \tag{2.75}$$

The most important aspect of the form/symbol correspondence is that the image of the Clifford product is precisely the fiberwise star product (2.57):

$$\Phi[F_1]\vee\Phi[F_2]=\Phi[F_1\circ F_2]. \tag{2.76}$$

This can be seen for instance by mapping Kähler's formula (1.9) for the Clifford product on our representation (2.15) of the fermionic Weyl star product:

$$\begin{aligned}\Phi[F_1]\vee\Phi[F_2]&=\sum_{p=0}^{n}\frac{(-1)^{p(p-1)/2}}{p!}(\mathcal{A}^p e_{\mu_1}\neg\cdots e_{\mu_p}\neg\Phi[F_1])\wedge(e^{\mu_1}\neg\cdots e^{\mu_p}\neg\Phi[F_2])\\&=\sum_{p=0}^{n}\frac{(-1)^{p(p-1)/2}}{p!}\Phi\left[\kappa^{-p}\mathcal{A}^p\frac{\partial}{\partial\theta^{\mu_1}}\cdots\frac{\partial}{\partial\theta^{\mu_p}}F_1\right]\wedge\Phi\left[\kappa^{-p}\frac{\partial}{\partial\theta_{\mu_1}}\cdots\frac{\partial}{\partial\theta_{\mu_p}}F_2\right]\\&=\Phi\left[\sum_{p=0}^{n}\kappa^{-2p}\frac{(-1)^{p(p-1)/2}}{p!}\left(\mathcal{A}^p\frac{\partial}{\theta^{\mu_1}}\cdots\frac{\partial}{\partial\theta^{\mu_p}}F_1\right)\left(\frac{\partial}{\partial\theta_{\mu_1}}\cdots\frac{\partial}{\partial\theta_{\mu_p}}F_2\right)\right]\\&=\Phi[F_1\circ F_2].\end{aligned}\tag{2.77}$$

Here we used (2.67), (2.70), and (2.75).

One also could prove Eq. (2.76) inductively. If we replace dx^μ by $\kappa\theta^\mu$ and ''$\vee$'' by the star product in the relations (1.8) which define the Clifford product we obtain exactly Eq. (2.12) for the star product. Therefore Eq. (2.76) is correct for zero- and one-forms. Its generalization for arbitrary p forms makes essential use of the associativity of both the Clifford and the star product.

By virtue of our rules for the form/symbol correspondence also Eqs. (1.13) and (2.18) are now seen to be completely equivalent.

In the DK equation we need the Clifford product of dx^μ with an arbitrary form:

$$dx^\mu\vee\Phi[F]=\Phi[\kappa\theta^\mu]\vee\Phi[F]=\Phi[\kappa\theta^\mu\circ F]=\Phi[C^\mu F].\tag{2.78}$$

Here C^μ is the first-order differential operator (2.53). In Sec. I we discussed already that $dx^\mu\vee$, regarded as an operator on the space of DK fields, gives rise to the Clifford algebra (1.14). For consistency the same should be true for the star multiplication with $\kappa\theta^\mu$ and for C^μ on the space of symbols. In fact, it is easy to see that

$$(\kappa\theta^\mu\circ)(\kappa\theta^\nu\circ)+(\kappa\theta^\nu\circ)(\kappa\theta^\mu\circ)=2\,\delta^{\mu\nu}\tag{2.79}$$

and

$$C^\mu C^\nu+C^\nu C^\mu=2\,\delta^{\mu\nu}.\tag{2.80}$$

The DK operator acting on forms reads

$$(d-d^\dagger)\Phi=dx^\mu\wedge\partial_\mu\Phi+e^\mu\neg\partial_\mu\Phi=dx^\mu\vee\partial_\mu\Phi,\tag{2.81}$$

where the second equality follows from (1.13). Therefore $d-d^\dagger$ becomes $\kappa\theta^\mu\circ\partial_\mu$ or $C^\mu\partial_\mu$ at the symbol level:

$$(d-d^\dagger)\Phi[F]=\Phi[\kappa\theta^\mu\circ\partial_\mu F]=\Phi[C^\mu\partial_\mu F].\tag{2.82}$$

This converts the DK equation to

$$\left[\left(\frac{1}{\kappa}\frac{\partial}{\partial\theta_\mu}+\kappa\theta^\mu\right)\partial_\mu+m\right]F(x,\theta)=0.\tag{2.83}$$

In closing we return to the chirality operator γ_{n+1}. Under the map Φ, the image of the delta function is essentially the volume form $\mathrm{Vol}\equiv dx^1\wedge dx^2\wedge\cdots\wedge dx^n$:

$$\Phi[\kappa^n\delta(\cdot)]=(-1)^{n(n-1)/2}\,\mathrm{Vol}.\tag{2.84}$$

For the chirality operator this means that

$$\Phi[G_{n+1}\circ F]=-i^{n(n-1)/2}\,\mathrm{Vol}\vee\Phi[F]\tag{2.85}$$

which, at the operator level, corresponds to

$$\widehat{G_{n+1}\circ F}=\gamma_{n+1}\hat{F}. \tag{2.86}$$

Thus we see that (up to unimportant constants) the fermionic δ function, the volume form, and the chirality matrix γ_{n+1} are simply different variants of the same object. Furthermore, by Eq. (2.43), star multiplication of F by G_{n+1} *from the right* amounts to applying the modified Hodge operator $\star$.

E. Invariant subspaces of $\mathcal{W}^F$

The differential operators C^μ or the star left-multiplication by $\kappa\theta^\mu$ define a representation of the Clifford algebra (2.21) in the space of symbols $f(\theta)$. As $f(\theta)$ has 2^n independent (complex) components, this representation is reducible. It can be decomposed into $2^{n/2}$ representations each of which is isomorphic to the $2^{n/2}$-dimensional irreducible representation provided by the matrices γ^μ. (We assume n even in this section.) As a consequence, a symbol-valued field $F(x,\theta)$ describes $2^{n/2}$ ordinary Dirac spinor fields.

In the light of the form/symbol correspondence which we developed in Sec. II D it is clear that the representation carried by $\mathcal{W}^F$ could be decomposed simply by invoking the standard discussion at the level of differential forms. However, as our main motivation for studying the symbol formulation of DK fields is to get some understanding of their symplectic analogs we shall reformulate the method of Becher and Joos[3] in symbol language and use this as a guide in the symplectic case. As a by-product we shall find a very elegant derivation of their matrix-valued form Z which puts it in a more general perspective.

We have to decompose the Weyl algebra in orthogonal subspaces,

$$\mathcal{W}^F=\bigoplus_{\alpha=1}^{k}\mathcal{W}^F_{(\alpha)},\quad k=2^{n/2} \tag{2.87}$$

such that $\mathcal{W}^F_{(\alpha)}$ is invariant under star left-multiplication by θ^μ. Following a strategy similar to the one described in Sec. I we look for a $k\times k$-matrix valued function $Z(\theta)$ with the property

$$\kappa\theta^\mu\circ Z_{\alpha\beta}(\theta)=\sum_{\gamma=1}^{k}(\gamma^{\mu T})_{\alpha\gamma}Z_{\gamma\beta}(\theta). \tag{2.88}$$

The function Z is readily found in our formalism. Since the star product with θ^μ involves first derivatives at most, Eq. (2.88) is reminiscent of the formulas for the derivative of the Weyl operators $\hat{\Omega}$ and $\check{\Omega}$ which we displayed in Sec. II A. In fact, using those formulas together with (2.18) it is easy to show that there exists a rescaling of the arguments of $\hat{\Omega}$ and $\check{\Omega}$ in such a way that the star multiplication by θ^μ corresponds to an operator multiplication by $\hat{\chi}^\mu$ or Γ^μ:

$$\theta^\mu\circ\hat{\Omega}(\pm i\kappa^2\theta)=\mp\hat{\Omega}(\pm i\kappa^2\theta)\hat{\chi}^\mu, \tag{2.89}$$

$$\theta^\mu\circ\hat{\Omega}(\pm\kappa^2\theta)=\pm i\hat{\chi}^\mu\hat{\Omega}(\pm\kappa^2\theta), \tag{2.90}$$

$$\kappa\theta^\mu\circ\check{\Omega}(\pm i\kappa^2\theta)=\pm\Gamma^\mu\check{\Omega}(\pm i\kappa^2\theta). \tag{2.91}$$

For the problem at hand, Eq. (2.91) is precisely what we need. If Γ_μ constitutes a Clifford algebra, so does Γ_μ^T. Hence we may set $Z(\theta)=\check{\Omega}(i\kappa^2\theta)$ with $\Gamma_\mu=\gamma_\mu^T$. Thus

$$Z(\theta)=\exp[\kappa\theta^\mu\gamma_\mu^T] \tag{2.92}$$

or in expanded form

$$Z(\theta)=\sum_{p=0}^{n}\frac{\kappa^p}{p!}\gamma_{\mu_1}^T\cdots\gamma_{\mu_p}^T\theta^{\mu_1}\cdots\theta^{\mu_p}. \tag{2.93}$$

Clearly (2.93) is precisely the symbol corresponding to the form (1.21) which was found by Becher and Joos[3] using different techniques. In the context of the present investigation it is important to keep in mind that Z is nothing but a rescaled fermionic Weyl operator since the latter has a well-known bosonic analog.

Because of the completeness properties of the γ matrices, $\{Z_{\alpha\beta};\quad \alpha,\beta=1,...,k\}$ is a basis for $\mathcal{W}^F$ and we may expand any symbol as

$$F(x,\theta)=\sum_{\beta=1}^{k}\sum_{\alpha=1}^{k}\psi_\alpha^{(\beta)}(x)Z_{\alpha\beta}(\theta)=\sum_{\beta=1}^{k}F^{(\beta)}(x,\theta). \tag{2.94}$$

(Here we use already the notation appropriate for the role of $\mathcal{W}^F$ as a fiber at the point x.) The rest of the argument parallels our discussion in Sec. I. We obtain $k\equiv 2^{n/2}$ invariant subspaces $\mathcal{W}_{(\alpha)}^F$ (left ideals) which are spanned by

$$F^{(\beta)}(x,\cdot)=\sum_{\alpha=1}^{k}\psi_\alpha^{(\beta)}(x)Z_{\alpha\beta}(\cdot)\in\mathcal{W}_{x(\beta)}^F. \tag{2.95}$$

For every fixed value of β, the expansion coefficients $\psi^{(\beta)}\equiv\{\psi_\alpha^{(\beta)};\alpha=1,...,k\}$ can be interpreted as an ordinary Dirac spinor. Equation (2.88) shows that acting with $\kappa\theta^\mu\circ$ on $F^{(\beta)}$ is equivalent to applying γ^μ on $\psi^{(\beta)}$:

$$\kappa\theta^\mu\circ F^{(\beta)}=\sum_\alpha\left(\sum_\delta\gamma^\mu_{\alpha\delta}\psi_\delta^{(\beta)}\right)Z_{\alpha\beta}=\sum_\alpha[\gamma^\mu\psi^{(\beta)}]_\alpha Z_{\alpha\beta}. \tag{2.96}$$

Let us arrange the expansion coefficients $\psi_\alpha^{(\beta)}$ as a $k\times k$-matrix: $(\hat{\psi})_{\alpha\beta}\equiv\psi_\alpha^{(\beta)}$. Then,

$$F(x,\theta)=\mathrm{Tr}[\hat{\psi}(x)Z(\theta)^T]. \tag{2.97}$$

Denoting the $\hat{\psi}$ matrix which belongs to a given section F by $\hat{\psi}[F]$ we obtain from (2.96)

$$\hat{\psi}[\kappa\theta^\mu\circ F]=\gamma^\mu\hat{\psi}[F], \tag{2.98}$$

which mirrors (1.27) at the symbol level.

Given a symbol-valued field $F(x,\theta)$ we can immediately construct the associated spinor matrix-valued field $\hat{F}(x)$ of (2.61) by replacing $\theta^\mu\rightarrow\kappa^{-1}\gamma^\mu$ in its series expansion (2.63). In the process of decomposing the reducible representation carried by F we discovered a second spinor-matrix, $\hat{\psi}$, which is related to F in a canonical way, too. By essentially the same argument as in Sec. I it follows that the two matrices are equal up to a constant:

$$\hat{\psi}[F](x)=2^{-n/2}\hat{F}(x). \tag{2.99}$$

If we insert the expansions (2.63) and (2.93) for $F(x,\theta)$ and $Z(\theta)$, respectively, into Eq. (2.97), we obtain Eq. (1.30) for the set $\{F^{(p)}_{\mu_1\cdots\mu_p}\}$ expressed in terms of $\hat{\psi}$. Making an ansatz for $\hat{\psi}$ in terms of antisymmetrized products of γ matrices and using the trace identity (1.33), one finds that the expansion coefficients of $\hat{\psi}$ and $\hat{F}$ differ by an overall constant only.

While this last step was straightforward for the SO(n) spinors, it will be much less trivial for metaplectic spinors where the representation space is infinite-dimensional and trace identities such as (1.33) are not likely to exist. It will be interesting to see how (2.99) is modified then.

III. SYMPLECTIC DIRAC–KÄHLER FIELDS

In Sec. II we reformulated the theory of standard DK fermions over space–time in terms of fields F which assume values in the fermionic Weyl algebra $\mathcal{W}^F$. Now we are going to ask what happens if we replace $\mathcal{W}^F$ by its (actually much more familiar) bosonic counterpart, the bosonic Weyl algebra $\mathcal{W}$. Rather than space–time it is now a phase-space $(\mathcal{M}_{2N},\omega)$ which plays the role of the base manifold. As we shall argue, replacing the Riemannian structure by a symplectic one, the structure group SO(n) by Sp$(2N)$, and, most importantly, fermionic Weyl symbols by bosonic ones, we are led to the notion of a ''symplectic DK field'' in a very natural way.

In Sec. III A we begin by working out some special properties of bosonic Weyl symbols which will become important in our construction. In this context, we are basically discussing the conventional quantum mechanics of the auxiliary quantum system with canonical operators $\hat{x}^i$ and $\hat{\pi}^i$ which results from quantizing the flat ''auxiliary phase-space'' $\mathbf{R}^{2N}$. [Later on the auxiliary phase space will be identified with the tangent space to the true (physical) phase-space $\mathcal{M}_{2N}$.] The operators $\hat{x}^i$ and $\hat{\pi}^i$ take over the role previously played by $\hat{\chi}^\mu$.

Section III B is devoted to the metaplectic γ matrices. In particular, we propose a symplectic analog of the chirality matrix γ_5 there. The actual construction of the symplectic DK fields is performed in Sec. III C, and in Sec. III D it is shown how they relate to the metaplectic spinor fields.

A. Bosonic Weyl symbols

We consider a Hamiltonian system with N degrees of freedom whose classical phase space is the symplectic plane $(\mathbf{R}^{2N},\omega)$. The associated quantum mechanical Hilbert space is $\mathcal{V}$ and $\mathcal{L}(\mathcal{V})$ denotes the space of linear operators on $\mathcal{V}$. The Hilbert space $\mathcal{V}$ carries a representation of the canonical commutation relations

$$[\hat{\varphi}^a,\hat{\varphi}^b]=i\hbar\,\omega^{ab},\quad a,b=1,...,2N. \tag{3.1}$$

In a canonical operator basis we split $\hat{\varphi}^a\equiv(\hat{\pi}^i,\hat{x}^i)$, $i=1,...,N$, so that the only nonvanishing commutator is between the momenta $\hat{\pi}^i$ and the positions $\hat{x}^i$:$[\hat{\pi}^i,\hat{x}^j]=-i\hbar\,\delta^{ij}$. The matrix (ω^{ab}) is the inverse of the constant matrix (ω_{ab}) formed from the coefficients of the symplectic two-form ω: $\omega_{ab}\omega^{bc}=\delta_a^c$. On $(\mathbf{R}^{2N},\omega)$ we use canonical coordinates $y^a\equiv(y_p^i,y_q^i)$ such that

$$(\omega_{ab})=\begin{pmatrix}0 & +I\\ -I & 0\end{pmatrix},\quad (\omega^{ab})=\begin{pmatrix}0 & -I\\ +I & 0\end{pmatrix}.$$

For the natural skew-symmetric inner product on the symplectic plane we write

$$\omega(y_1,y_2)\equiv y_1^a\omega_{ab}y_2^b. \tag{3.2}$$

The Weyl (or Heisenberg) operators[16]

$$\hat{T}(y)=\exp\left(\frac{i}{\hbar}y^a\omega_{ab}\hat{\varphi}^b\right) \tag{3.3}$$

implement the translations on phase space in the Hilbert space $\mathcal{V}$:

$$\hat{T}(y)^\dagger\hat{\varphi}^a\hat{T}(y)=\hat{\varphi}^a+y^a. \tag{3.4}$$

This is a projective representation of the translation group since

$$\hat{T}(y_1)\hat{T}(y_2)=\exp\left[\frac{i}{2\hbar}\,\omega(y_1,y_2)\right]\hat{T}(y_1+y_2). \tag{3.5}$$

The Weyl operators are orthogonal and complete in the sense that

$$\mathrm{Tr}[\hat{T}(y_1)^\dagger \hat{T}(y_2)] = (2\pi\hbar)^N \delta^{(2N)}(y_1 - y_2) \tag{3.6}$$

$$\int d^{2N}y \langle \alpha | \hat{T}(y)^\dagger | \alpha' \rangle \langle \beta | \hat{T}(y) | \beta' \rangle = (2\pi\hbar)^N \delta^{(N)}(\alpha - \beta') \delta^{(N)}(\beta - \alpha'). \tag{3.7}$$

Here $\{|\alpha\rangle\}$ is the basis which diagonalizes the position operators:

$$\hat{x}^i |\alpha\rangle = \alpha^i |\alpha\rangle, \quad \alpha \equiv (\alpha^1, \dots, \alpha^N). \tag{3.8}$$

Sometimes it will be more suggestive to use a tensor notation instead of the bra-ket formalism; for instance, one writes $\hat{b}^\alpha{}_\beta \equiv \langle \alpha | \hat{b} | \beta \rangle$ for the matrix elements of some arbitrary $\hat{b} \in \mathcal{L}(\mathcal{V})$ or $\delta^\alpha{}_\beta \equiv \delta^{(N)}(\alpha - \beta)$ for the identity operator. The eigenvalues $\alpha \in \mathbb{R}^N$ should be thought of as a continuous analog of a spinor index. In the $\hat{x}$ eigenbasis, the Weyl operators are given by

$$\hat{T}(y)^\alpha{}_\beta = \exp\left[\frac{i}{\hbar} \left(y_p \alpha - \frac{1}{2} y_p y_q \right) \right] \delta^{(N)}(\alpha - \beta - y_q) \tag{3.9}$$

with $y_p \alpha \equiv y_p^i \alpha^i$, etc., where the summation over $i = 1,...,N$ is understood.

From the completeness relation (3.7) it follows that every operator $\hat{b}$ can be represented as

$$\hat{b} = (2\pi\hbar)^{-N} \int d^{2N}y \, \tilde{b}(y) \hat{T}(y) \tag{3.10}$$

with the complex-valued function $\tilde{b}$ given by

$$\tilde{b}(y) = \mathrm{Tr}[\hat{T}(y)^\dagger \hat{b}]. \tag{3.11}$$

The function $\tilde{b}$ (referred to as the *alternative* Weyl symbol[16]) is closely related to the Weyl symbol of $\hat{b}$. In fact, $b(y) \equiv [\mathrm{symb}(\hat{b})](y)$ is the Fourier transform of $\tilde{b}$:

$$b(y) = (2\pi\hbar)^{-N} \int d^{2N}y_0 \, \tilde{b}(y_0) \exp\left[\frac{i}{\hbar} \omega(y_0, y) \right] \tag{3.12}$$

The inverse transformation reads

$$\tilde{b}(y) = (2\pi\hbar)^{-N} \int d^{2N}y_0 \, b(y_0) \exp\left[\frac{i}{\hbar} \omega(y_0, y) \right], \tag{3.13}$$

i.e., the symplectic Fourier transformation is an exact involution, $\tilde{\tilde{b}}(y) = b(y)$ (and not only an involution up to a reflection of the argument).

Equations (3.10)–(3.12) define the (bosonic) Weyl symbol map ''symb'' from $\mathcal{L}(\mathcal{V})$ to the space of (generalized) functions over the symplectic plane, as well as its inverse. The classical phase function $b(y)$ uniquely represents an operator $\hat{b}$ which is Weyl ordered. In particular, the monomial $y^{a_1} y^{a_2} \cdots y^{a_p}$ stands for the completely symmetrized operator product $\hat{\varphi}^{(a_1} \hat{\varphi}^{a_2} \cdots \hat{\varphi}^{a_p)}$. Conversely,

$$[\mathrm{symb}\{\hat{\varphi}^{(a_1} \cdots \hat{\varphi}^{a_p)}\}](y) = y^{a_1} y^{a_2} \cdots y^{a_p}. \tag{3.14}$$

The symmetrization in (3.14) is crucial, otherwise commutator terms would occur. For instance,

$$[\mathrm{symb}\{\hat{\varphi}^a \hat{\varphi}^b\}](y) = y^a y^b + i \frac{\hbar}{2} \omega^{ab}. \tag{3.15}$$

An important special class of symbols are those which admit a power series expansion

$$b(y)=\sum_{p=0}^{\infty}\frac{\kappa^p}{p!}b^{(p)}_{a_1\cdots a_p}y^{a_1}y^{a_2}\cdots y^{a_p}. \tag{3.16}$$

By the symbol map, they are bijectively related to the operators

$$\hat{b}=\sum_{p=0}^{\infty}\frac{\kappa^p}{p!}b^{(p)}_{a_1\cdots a_p}\hat{\varphi}^{a_1}\cdots\hat{\varphi}^{a_p} \tag{3.17}$$

provided the tensors $b^{(p)}_{a_1\cdots a_p}$ are completely symmetric. If b is a power series, the ''alternative Weyl symbol'' $\tilde{b}$ is a sum of derivatives of δ functions:

$$\tilde{b}(y)=(2\pi\hbar)^N b\left(i\hbar\,\omega^{ac}\frac{\partial}{\partial y^c}\right)\delta^{(2N)}(y). \tag{3.18}$$

As in every symbol calculus, the pertinent star product is required to satisfy symb$(\hat{b}_1\hat{b}_2)=b_1\circ b_2$ where b_1 and b_2 are the symbols of $\hat{b}_1$ and $\hat{b}_2$, respectively. At least for power series, the bosonic Weyl star product is uniquely determined by its associativity, the distributivity over ''+,'' and the basic relations

$$\begin{gathered}1\circ 1=1,\quad 1\circ y^a=y^a\circ 1=y^a,\\ y^a\circ y^b=y^a y^b+i\frac{\hbar}{2}\omega^{ab},\end{gathered} \tag{3.19}$$

which follow from (3.14), (3.15) and symb$(I)=1$. Explicit formulas for the star product[25,28] of arbitrary symbols include

$$(b_1\circ b_2)(y)=b_1(y)\exp\left[i\frac{\hbar}{2}\frac{\overleftarrow{\partial}}{\partial y^a}\omega^{ac}\frac{\overrightarrow{\partial}}{\partial y^c}\right]b_2(y) \tag{3.20}$$

and

$$\begin{aligned}(b_1\circ b_2)(y)=(\pi\hbar)^{-2N}\int d^{2N}y_1 d^{2N}y_2\exp[-2i\{\omega(y,y_1)+\omega(y_1,y_2)\\ +\omega(y_2,y)\}/\hbar]b_1(y_1)b_2(y_2).\end{aligned} \tag{3.21}$$

The differential operators which effect the star left-multiplication with κy^a,

$$(C^a b)(y)=\kappa y^a\circ b(y), \tag{3.22}$$

are easily read off from Eq. (3.20):

$$C^a=\kappa y^a+\frac{i}{\kappa}\omega^{ab}\frac{\partial}{\partial y^b}. \tag{3.23}$$

On the space of symbols with an appropriate fall-off behavior we would like to introduce a sesquilinear inner product $(\cdot|\cdot)$ with respect to which C^a is self-adjoint,

$$(C^a b_1|b_2)=(b_1|C^a b_2). \tag{3.24}$$

It is clear from our earlier discussion that the choice

$$(b_1|b_2)=[\overline{b_1}\circ b_2](y=0) \tag{3.25}$$

meets this requirement. Since $\overline{b_1 \circ b_2} = \overline{b_2} \circ \overline{b_1}$ also here, the proof is the same as in (2.54). If b_1 and b_2 are power series of the type (3.16), Eq. (3.25) boils down to

$$(b_1|b_2) = \sum_{p=0}^{\infty} \frac{i^p}{p!} \overline{b_{1,a_1 \cdots a_p}^{(p)}} \omega^{a_1 c_1} \cdots \omega^{a_p c_p} b_{2,c_1 \cdots c_p}^{(p)}. \tag{3.26}$$

It is instructive to look at various alternative ways of representing this inner product. There exists the integral representation

$$(b_1|b_2) = (2\pi)^{-2N} \int d^{2N}y_1 d^{2N}y_2 \overline{b_1}(y_1/\kappa) e^{-i\omega(y_1,y_2)} b_2(y_2/\kappa), \tag{3.27}$$

which can be reexpressed in terms of a symplectic Fourier transform:

$$(b_1|b_2) = (2\pi\hbar)^{-N} \int d^{2N}y \overline{b_1}(\tfrac{1}{2}y) \tilde{b}_2(y). \tag{3.28}$$

Furthermore, if (3.18) can be applied,

$$(b_1|b_2) = b_2\left(-i\kappa^{-2}\omega^{ac} \frac{\partial}{\partial y^c} \right) \overline{b_1}(y)|_{y=0}. \tag{3.29}$$

The above formulas should be compared to their counterparts in the fermionic symbol calculus. Bosonic symbols admitting a power series expansion are characterized by sets $\{b_{a_1 \cdots a_p}^{(p)}, p = 0,1,2,...\}$ consisting of infinitely many *symmetric* tensors. Fermionic symbol functions are equivalent to a finite set $\{f_{\mu_1 \cdots \mu_p}^{(p)}, p = 0,1,...,n\}$ of *antisymmetric* tensors instead.

We saw that the (modified) Hodge operator is essentially the same operation as the Grassmannian Fourier transformation. Omitting all sign factors (which anyhow have no bosonic analog) we have, schematically,

$$*f(\theta) \propto \star f(\theta) \propto \tilde{f}(\theta) \propto f\left(\frac{\partial}{\partial\theta} \right) \delta(\theta). \tag{3.30}$$

Thus one is tempted to define a bosonic version of the Hodge operator simply by setting $(*b)(y) = \tilde{b}(y)$ so that $** = 1$ on any b. If b is a power series, Eq. (3.18) is indeed formally analogous to (A24) for the fermionic Fourier transformation. However, the difference is that the derivative of the fermionic delta function, $f(\partial/\partial\theta)\delta(\theta)$, again is a powers series in the θ's, while this is of course not true for the derivatives of the bosonic delta function, $\delta^{(2N)}(y)$. In the former case, the monomials $\theta^{\mu_1} \cdots \theta^{\mu_p}$ are mapped onto monomials of the same type. Therefore one set of antisymmetric tensors $\{f_{a_1 \cdots a_p}^{(p)}\}$ is mapped onto another set of such tensors. In the latter case, the space of symmetric tensors $b_{a_1 \cdots a_p}^{(p)}$ is not mapped onto itself. The image of $y^{a_1} y^{a_2} \cdots y^{a_p}$ is a singular symbol $\propto \partial_y^{a_1} \cdots \partial_y^{a_p} \delta^{(2N)}(y)$, $\partial_y^a \equiv \omega^{ab} \partial/\partial y^b$.

Nevertheless it will be helpful to think of the symplectic Fourier transformation as the bosonic (symmetric tensor) analog of the Hodge operator. For instance, by (2.48) with (A22) and (A40) the fermionic inner product has the same general structure as (3.28):

$$(f_1|f_2) \propto \int \overline{f_1}(\theta) \tilde{f}_2(\theta) d^n\theta \propto \int \overline{f_1}(\theta)(*f_2)(\theta) d^n\theta. \tag{3.31}$$

In the language of differential forms this is nothing but the familiar inner product $*(\overline{\Phi_1} \wedge *\Phi_2)$ in disguise. The product $(b_1|b_2)$ introduced above is analogous to it, but refers to symmetric rather than antisymmetric tensors.

The space of symbols $b(y)$ equipped with the pointwise product of functions, the star product, and the inner product constitutes the bosonic Weyl algebra $\mathcal{W}$. It is the counterpart of the algebra $\mathcal{W}^{\mathrm{F}}$ which, endowed with analogous structures, had turned out to be an Atiyah–Kähler algebra. Because $(y^a|y^b)=i\hbar\omega^{ab}/2$ we see that the three product structures on $\mathcal{W}$ satisfy the consistency condition

$$y^a\circ y^b=y^a y^b+(y^a|y^b). \tag{3.32}$$

This relation is completely analogous to Eq. (2.46) which had been identified with the defining property of an Atiyah–Kähler algebra, Eq. (1.7). This supports our point of view that *the bosonic Weyl algebra is the natural analog of an Atiyah–Kähler algebra if one works in a symplectic rather than a Riemannian setting*.

B. γ matrices for Mp$(2N)$ and the analog of γ_5

The generators of Mp$(2N)$ in the spinor representation are obtained as symmetrized bilinears $\Sigma^{ab}_{\mathrm{meta}}=(\gamma^a\gamma^b+\gamma^b\gamma^a)/4$ built from $2N$ ''γ matrices'' satisfying

$$\gamma^a\gamma^b-\gamma^b\gamma^a=2i\omega^{ab}. \tag{3.33}$$

Upon identifying

$$\gamma^a=\kappa\hat{\varphi}^a \tag{3.34}$$

it is clear that the relations (3.33) coincide precisely with (3.1). Hence, what in the language of group theory is called a ''symplectic Clifford algebra'' is nothing but the canonical commutation relations of a bosonic quantum system with the canonical operators $\hat{\varphi}^a=(\hat{\pi}^j,\hat{x}^j)$. For N finite, all irreducible representations of the canonical commutation relations are unitarily equivalent, so the same is true for the symplectic Clifford algebra. All these representations are infinite dimensional.

We consider representations where γ^a is a Hermitian operator on the Hilbert space $\mathcal{V}$. Frequently $\mathcal{V}$ is taken to be the Fock space of N independent harmonic oscillators.[17,31] Then the γ^a's are linear combinations of the corresponding creation and annihilation operators. Here we shall employ another representation which is particularly natural in the gauge theory approach to quantum mechanics.[21] We pick the $\hat{x}$-eigenbasis (3.8) with respect to which $\langle\alpha|\hat{x}^j|\beta\rangle=\alpha^j\delta^{(N)}(\alpha-\beta)$ and $\langle\alpha|\hat{\pi}^j|\beta\rangle=-i\hbar\partial_j\delta^{(N)}(\alpha-\beta)$. Therefore, in a symbolic matrix notation with $\langle\alpha|\gamma^a|\beta\rangle\equiv(\gamma^a)^\alpha{}_\beta$,

$$\begin{aligned}(\gamma^j)^\alpha{}_\beta&=-(2i/\kappa)\partial^j\delta^{(N)}(\alpha-\beta),\\ (\gamma^{N+j})^\alpha{}_\beta&=\kappa\alpha^j\delta^{(N)}(\alpha-\beta),\quad j=1,...,N.\end{aligned} \tag{3.35}$$

The Hilbert space $\mathcal{V}$ is the space of square integrable functions $\psi(\alpha)\equiv\langle\alpha|\psi\rangle\equiv\psi^\alpha$ with its usual inner product. The generators $\Sigma^{ab}_{\mathrm{meta}}$ act on $\mathcal{V}$ as second-order differential operators (Schrödinger Hamiltonians with a quadratic potential; see Refs. 22 and 23 for further details).

Any attempt at putting metaplectic spinors on a similar footing as the SO(n) spinors faces the problem that $\mathcal{V}$ is infinite dimensional and that a metaplectic spinor formally is an object $\psi^\alpha\equiv\psi(\alpha)$ with infinitely many components. As an immediate consequence, trace identities such as (1.33) have no direct counterpart for the metaplectic γ ''matrices''. In the $\hat{x}$ basis, for instance, the trace of an operator $\hat{b}\in\mathcal{L}(\mathcal{V})$ reads $\mathrm{Tr}(\hat{b})=\int d^N\alpha\langle\alpha|\hat{b}|\alpha\rangle$, and it is clear that monomials such as $\gamma^{a_1}\cdots\gamma^{a_p}$ do not possess a trace. Remarkably enough, it turns out that there exist identities similar to (1.33) even in the infinite dimensional case which, however, involve the Sp$(2N)$ analog of γ_{n+1}.

We are familiar with the fact that when we are dealing with spinors on an even-dimensional space–time there exists a chirality matrix γ_{n+1}, a generalization of γ_5 in four dimension, which anticommutes with any γ^μ. Its eigenvalues are -1 and $+1$, and the corresponding eigenspaces are the left- and right-handed Weyl spinors, respectively. It is quite interesting that we can introduce an analogous concept for metaplectic spinors and that the pertinent ''chirality operator'' has a very natural interpretation even. Let us try to find an operator $\gamma_P \in \mathcal{L}(\mathcal{V})$ which anticommutes with all γ^a's,

$$\gamma_P \gamma^a + \gamma^a \gamma_P = 0 \tag{3.36}$$

and satisfies

$$\gamma_P^\dagger = \gamma_P^{-1} = \gamma_P . \tag{3.37}$$

Thus γ_P has the same algebraic properties as γ_5, its eigenvalues are ± 1 and, provided it actually exists, we can use it to form the ''chiral'' projections

$$\psi_\pm = \Pi_\pm \psi, \quad \Pi_\pm \equiv \tfrac{1}{2}(1 \pm \gamma_P) \tag{3.38}$$

of any $\psi \in \mathcal{V}$. Since $\Sigma^{ab}_{\rm meta}$ commutes with γ_P, the M$p(2N)$ transformations leave the subspaces with $\gamma_P = +1$ and $\gamma_P = -1$ invariant, so that the representation of Mp$(2N)$ implied by the γ matrices (3.35) decomposes accordingly.

Looking at the ''metaplectic γ_5 matrix'' from the point of view of the auxiliary quantum mechanics with the $\hat{\varphi}$ degrees of freedom it becomes clear that we may identify γ_P with the standard parity operator P in this context. By definition, P changes the sign of both the positions $\hat{x}^i$ and the momenta $\hat{\pi}^i$: $P\hat{x}^i P = -\hat{x}^i$, $P\hat{\pi}^i P = -\hat{\pi}^i$. Hence $P\gamma^a P = -\gamma^a$ for $\gamma^a = \kappa(\hat{\pi}^i, \hat{x}^i)$, which is exactly (3.36) with $\gamma_P \equiv P$. The operator P acts on the wave functions $\psi \in \mathcal{V}$ as $(P\psi)(\alpha) \equiv (\gamma_P \psi)(\alpha) = \psi(-\alpha)$. This means that in the $\hat{x}$ representation

$$\gamma_P |\alpha\rangle = |-\alpha\rangle \tag{3.39}$$

so that the matrix elements of γ_P are given by

$$(\gamma_P)^\alpha{}_\beta \equiv \langle \alpha | \gamma_P | \beta \rangle = \delta^{(N)}(\alpha + \beta). \tag{3.40}$$

Thus, ''metaplectic chirality'' is nothing but ''fiberwise parity,'' and the projections $\Pi_\pm \mathcal{V}$ are simply the subspaces of even and odd wave functions, respectively.

The operator γ_P can be written in a manifestly basis independent way:

$$\gamma_P = (4\pi\hbar)^{-N} \int d^{2N} y \, \hat{T}(y). \tag{3.41}$$

[Equation (3.41) shows that γ_P belongs to the family of parity-type operators discussed by Grossmann[32] and Royer.[33]] The general properties of the Weyl operators imply that (3.41) has the desired properties (3.36), (3.37) and using the matrix elements (3.9) one finds that (3.41) coincides with (3.40). Equation (3.41) is strikingly similar to Eq. (2.44) for $\hat{G}_{n+1}$ which confirms our interpretation that the *fiberwise parity transformation is the analog of* γ_5.

The operator γ_P has a well-defined finite trace:

$$\mathrm{Tr}[\gamma_P] = 2^{-N}. \tag{3.42}$$

This follows from (3.41) with (3.6) or simply by noting that

$$\mathrm{Tr}[\gamma_P] = \int d^N \alpha (\gamma_P)^\alpha{}_\alpha = \int d^N \alpha \, \delta^{(N)}(2\alpha) = 2^{-N}. \tag{3.43}$$

While the very existence of this trace is remarkable, we see the first major difference between the bosonic and the fermionic case here. Both γ_5 and γ_P have eigenvalues ± 1, but the pairing of positive and negative eigenvalues which leads to $\mathrm{Tr}(\gamma_5) = 0$ does not happen for γ_P.

Finite products of γ^a matrices and in particular the unit operator do not possess a well-defined trace. On the other hand, traces with a γ_P insertion,

$$\mathrm{Tr}[\hat{b}\,\gamma_P]=\int d^N\alpha\langle\alpha|\hat{b}|-\alpha\rangle \tag{3.44}$$

are much better behaved because the reflection $\alpha\mapsto-\alpha$ removes possible ''short distance singularities'' (reminiscent of ultraviolet divergences in field theory) which would plague $\langle\alpha|\hat{b}|\alpha\rangle$.

This situation is quite similar to what one encounters in quantum field theory in the computation of chiral anomalies or, from a mathematical point of view, of the analytical index of the Dirac operator.[34,35] There one considers $\mathrm{Tr}(I)$ and $\mathrm{Tr}(\gamma_5)$ where the trace is over the infinite dimensional Hilbert space of Dirac spinor fields. While $\mathrm{Tr}(I)$ does not exist, $\mathrm{Tr}(\gamma_5)$ can be interpreted as the index of the Dirac operator.

An important trace of the type (3.44) is

$$\mathrm{Tr}[\gamma^{(a_1}\cdots\gamma^{a_p)}\gamma_P\gamma_{(b_1}\cdots\gamma_{b_q)}]=i^p2^{-N}p!\,\delta^{pq}\delta^{a_1}_{(b_1}\delta^{a_2}_{b_2}\cdots\delta^{a_p}_{b_p)} \tag{3.45}$$

with the convenient abbreviation $\gamma_a\equiv\omega_{ab}\gamma^b$, $\omega^{ab}\gamma_b=\gamma^a$. Equation (3.45) is similar to (1.33) for the $\mathrm{SO}(n)$ γ matrices, but contains an additional factor of γ_P without which the trace would not exist. Equation (3.45) follows from the properties for the $\hat{T}$ operators. First one uses (3.41) with (3.6) to show that

$$\mathrm{Tr}[\hat{T}(y)\gamma_p]=2^{-N} \tag{3.46}$$

is independent of y. Next one writes

$$\begin{aligned}\mathrm{Tr}[\hat{T}(y_1)\gamma_P\hat{T}(y_2)]&=\mathrm{Tr}[\hat{T}(y_1)\hat{T}(-y_2)\gamma_P]\\&=\exp\left[\frac{i}{2\hbar}\,\omega(y_1,-y_2)\right]\mathrm{Tr}[\hat{T}(y_1-y_2)\gamma_P]=2^{-N}\exp\left[-\frac{i}{2\hbar}y_1^a\omega_{ab}y_2^b\right].\end{aligned} \tag{3.47}$$

If one now expands the first and the last expression of (3.47) in powers of y_1 and y_2 and equates equal powers, the result is precisely Eq. (3.45).

Some important special cases of (3.45) include

$$\begin{aligned}\mathrm{Tr}[\gamma^a\gamma_P]&=0,\\\mathrm{Tr}[\gamma^{(a_1}\cdots\gamma^{a_p)}\gamma_P]&=0,\\\mathrm{Tr}[\gamma^a\gamma^b\gamma_P]&=2^{-N}i\,\omega^{ab}.\end{aligned} \tag{3.48}$$

The reader is invited to check some of these relations by using the matrix elements of γ^a in the $|\alpha\rangle$ basis. It is instructive to see that these calculations involve only well-defined manipulations of distributions and that no additional ad hoc regularization is needed. This is different from the derivation of the closely related dimension-counting formulas for the spinors of $\mathrm{OSp}(n|2N)$ which appear in certain approaches to the covariant quantization of superstrings,[18] for instance.

C. The Dirac–Kähler construction on phase space

Let $(\mathcal{M}_{2N},\omega)$ denote an arbitrary $2N$-dimensional symplectic manifold which serves as the phase space of some Hamiltonian system. Let us consider the Weyl algebra bundle[30,31] over $\mathcal{M}_{2N}$. Its typical fiber is the bosonic Weyl algebra $\mathcal{W}$, i.e., the space of symbols $b(\cdot)$ equipped with the pointwise product of functions, the star product, and the inner product $(\cdot|\cdot)$. At each point ϕ of

$\mathcal{M}_{2N}$ we attach a local copy $\mathcal{W}_\phi$ of $\mathcal{W}$. The matrix (ω_{ab}) which enters the definition of the Weyl algebra $\mathcal{W}_\phi$ are the coefficients of the symplectic two-form ω evaluated at the point ϕ. By virtue of Darboux's theorem, there exist local coordinates (ϕ^a) such that those coefficients assume their canonical form on the entire (ϕ^a) chart. Local coordinates on the total space are pairs (ϕ,b) with b a function $b:\mathbf{R}^{2N}\to\mathbf{C}$, $y\mapsto b(y)$. The transition functions of the bundle are defined in such a way that the variables $(y^1,\ldots,y^{2N})$ on which b depends are the components of a vector $y \in T_\phi\mathcal{M}_{2N}$, i.e., $y^a=d\phi^a(y)$. A symplectic change of coordinates $\phi^a\to\tilde{\phi}^a(\phi)$ (canonical transformation) is to be combined with a transformation in the fiber, $b\to\tilde{b}$, such that $\tilde{b}(\tilde{y})=b(y)$ with $\tilde{y}^a=(\partial\tilde{\phi}^a/\partial\phi^b)y^b$.

Along with the Weyl algebra bundle we also consider the metaplectic spinor bundle over $(\mathcal{M}_{2N},\omega)$ which we described in Sec. I. Its fiber at ϕ, $\mathcal{V}_\phi$, is a copy of the Hilbert space $\mathcal{V}$ on which we already constructed a representation of the metaplectic Clifford algebra and, as a consequence, of the structure group $\mathrm{Mp}(2N)$.

Let us look at sections through the Weyl algebra bundle. Locally they are specified by functions $\phi\mapsto B(\phi,\cdot)\in\mathcal{W}_\phi$ where $B(\phi,\cdot):\mathbf{R}^{2N}\to\mathbf{C}$, $y\mapsto B(\phi,y)$ is a Weyl symbol ''living'' in the fiber at ϕ. In this context, the flat ''auxiliary phase-space'' $\mathbf{R}^{2N}$ is identified with the tangent space $T_\phi\mathcal{M}_{2N}$. Hence the function $B(\cdot,\cdot)$ is a map from (a part of) the total space of the tangent bundle into $\mathbf{C}$.

Many of the concepts which we developed in Sec. III A for symbols $b\in\mathcal{W}$ naturally pass over to the sections B. At every point ϕ of $\mathcal{M}_{2N}$ we can apply the inverse symbol map to $B(\phi,\cdot)\in\mathcal{W}_\phi$ and obtain a unique operator $\hat{B}(\phi)=\mathrm{symb}^{-1}B(\phi,\cdot)$ which acts on the local copy $\mathcal{V}_\phi$ of the Hilbert space $\mathcal{V}$. Thus a section B gives rise to a family of operators $\hat{B}(\phi)\in\mathcal{L}(\mathcal{V}_\phi)$ labeled by the points of phase space. Its matrix elements with respect to a given basis in $\mathcal{V}$ will be denoted $\hat{B}(\phi)^\alpha{}_\beta\equiv\langle\alpha|\hat{B}(\phi)|\beta\rangle$. Globally speaking, $\hat{B}$ is a section through the bundle of (1,1) multispinors.[22,21]

The fiberwise star product of two sections is defined by

$$(B_1\circ B_2)(\phi,y)=B_1(\phi,y)\exp\left[\frac{i\hbar}{2}\frac{\overleftarrow{\partial}}{\partial y^a}\omega^{ab}\frac{\overrightarrow{\partial}}{\partial y^b}\right]B_2(\phi,y). \tag{3.49}$$

This star product has to be carefully distinguished from the $*_\mathcal{M}$ product whose associated Moyal bracket $\{f,g\}_M=(f*_\mathcal{M}g-g*_\mathcal{M}f)/i\hbar$ replaces the classical Poisson bracket in the deformation quantization approach,[24,36] and which involves derivatives with respect to ϕ^a rather than y^a. In general, the $*_\mathcal{M}$ product is much more complicated than the $\circ$ product. It can be constructed iteratively by Fedosov's method,[29,37–40] but we shall not need it in the present context.

The fiberwise inner product of two sections is given by $(B_1|B_2)(\phi)=(\overline{B_1}\circ B_2)(\phi,0)$. The natural sesquilinear form on the space of sections is $\langle B_1|B_2\rangle=\int d\mu_L(B_1|B_2)$ where $d\mu_L$ is the Liouville measure.

After these preparations we are now able to construct an analog of the Dirac–Kähler fields on phase spaces.

Let $\bigotimes^p_{\mathrm{sym}}(T^*\mathcal{M}2N)$ denote the p-fold symmetrized tensor power of the cotangent bundle. A section $\mathbf{\Sigma}^{(p)}$ through this bundle is a symmetric tensor field of rank p. We shall also consider the direct sum

$$\bigotimes_{\mathrm{sym}}(T^*\mathcal{M}_{2N})=\bigoplus_{p=0}^{\infty}\bigotimes^p_{\mathrm{sym}}(T^*\mathcal{M}_{2N}). \tag{3.50}$$

Its sections $\mathbf{\Sigma}=\Sigma_{p=0}^\infty\mathbf{\Sigma}^{(p)}$ are analogous to the inhomogeneous differential forms, but with symmetric rather than antisymmetric tensor fields. In local (Darboux) coordinates ϕ^a, $\mathbf{\Sigma}$ can be expanded as

$$\boldsymbol{\Sigma}(\phi)=\sum_{p=0}^{\infty}\frac{1}{p!}B^{(p)}_{a_1\cdots a_1}(\phi)d\phi^{a_1}\otimes_{\text{sym}}d\phi^{a_2}\otimes_{\text{sym}}\cdots\otimes_{\text{sym}}d\phi^{a_p} \tag{3.51}$$

with $\otimes_{\text{sym}}$ denoting the symmetric counterpart of the wedge product; for instance, $d\phi^a\otimes_{\text{sym}}d\phi^b = d\phi^a\otimes d\phi^b+d\phi^b\otimes d\phi^a$. The complex-valued coefficients $B^{(p)}_{a_1\cdots a_p}$ are taken to be completely symmetric in all p indices. We shall refer to $\boldsymbol{\Sigma}$ as an ''inhomogeneous symmetric tensor'' (IST).

Guided by the corresponding construction in the fermionic case we shall now associate an operator $\hat{B}(\phi)\in\mathcal{L}(\mathcal{V}_\phi)$ to $\boldsymbol{\Sigma}(\phi)$ by replacing in Eq. (3.51) the differentials $d\phi^a$ with the Gamma matrices γ^a:

$$\hat{B}(\phi)=\sum_{p=0}^{\infty}\frac{1}{p!}B^{(p)}_{a_1\cdots a_p}(\phi)\gamma^{a_1}\gamma^{a_2}\cdots\gamma^{a_p}=\sum_{p=0}^{\infty}\frac{\kappa^p}{p!}B^{(p)}_{a_1\cdots a_p}(\phi)\hat{\varphi}^{a_1}\hat{\varphi}^{a_2}\cdots\hat{\varphi}^{a_p}. \tag{3.52}$$

(As discussed earlier, we interpret the γ matrices as κ times the canonical operators $\hat{\varphi}^a$ of the auxiliary quantum system in the fiber.) Conversely, every Weyl ordered operator on $\mathcal{V}_\phi$ which admits a power series expansion gives rise to a unique IST. The operator can be expanded in the symmetrized monomials $\hat{\varphi}^{(a_1}\ldots\hat{\varphi}^{a_p}$ with coefficients which are symmetric tensors and define an IST therefore.

Now we form the symbol of $\hat{B}$: $B(\phi,y)=[\text{symb}\,\hat{B}(\phi)](y)$, i.e.,

$$B(\phi,y)=\sum_{p=0}^{\infty}\frac{\kappa^p}{p!}B^{(p)}_{a_1\cdots a_p}(\phi)y^{a_1}y^{a_2}\cdots y^{a_p}. \tag{3.53}$$

Taking both steps together we arrive at a one-to-one correspondence between ISTs and symbols with a power series expansion in y:

$$\boldsymbol{\Sigma}(\phi)\in\otimes_{\text{sym}}(T^*_\phi\mathcal{M}_{2N})\leftrightarrows\hat{B}(\phi)\in\tilde{\mathcal{L}}(\mathcal{V}_\phi)\leftrightarrows B(\phi,\cdot)\in\tilde{\mathcal{W}}_\phi. \tag{3.54}$$

This chain of bijections is similar to (2.64). However, the difference is that in the fermionic setting every symbol or every Weyl ordered operator gives rise to an inhomogeneous tensor field. This is not true in the bosonic case. We have to explicitly restrict the symbols and operators to those which allow for a power series expansion in y^a or $\hat{\varphi}^a$, respectively. [This is indicated by the notation $\tilde{\mathcal{L}}(\mathcal{V}_\phi)$ and $\tilde{\mathcal{W}}_\phi$.] Nevertheless we shall continue to consider also symbols B which are not analytic in y because they will play a central role in the reduction of symplectic DK fields.

Now let us look at the rules of the symbol/tensor correspondence in the bosonic case. Clearly the differentials $d\phi^a$ correspond to κy^a: $d\phi^a\leftrightarrows\kappa y^a$. Hence, if we write the (linear, invertible) map from the symbols to the ISTs as $B\mapsto\boldsymbol{\Sigma}[B]$, we have $\boldsymbol{\Sigma}[\kappa y^a]=d\phi^a$ or more generally

$$\boldsymbol{\Sigma}[\kappa^p y^{a_1}\cdots y^{a_1}]=d\phi^{a_1}\otimes_{\text{sym}}\cdots\otimes_{\text{sym}}d\phi^{a_p}. \tag{3.55}$$

Up to this point the situation is the same as in Sec. II D with the commuting y's replacing the anticommuting θ's. This converts the wedge product to the symmetric tensor product. Differences become manifest when we look at the list of natural operations for ISTs and their realization at the symbol level.

The automorphism $\mathcal{A}$ and the antiautomorphism $\mathcal{B}$, while important for dealing with the ubiquitous sign factors in exterior algebra computations, are unnecessary for symmetric tensors. As we argued already, the Hodge operator has a natural bosonic translation, the symplectic Fourier transformation. However it does not leave the space $\tilde{\mathcal{W}}$ invariant and, as a consequence, does not induce a map of one IST onto another. Furthermore, the exterior derivative is a derivation on the exterior algebra which does not require a connection for its definition. Also this concept has no analog on the bosonic side.

However, every vector field $v=v^a(\phi)\partial_a$ on $\mathcal{M}_{2N}$ gives rise to a contraction operator $\mathbf{i}(v)$. By definition, it is a linear operator on the space of ISTs, depending linearly on v, and satisfying $\mathbf{i}(\partial_a)1=0$, $\mathbf{i}(\partial_a)d\phi^b=\delta_a^b$ as well as

$$\mathbf{i}(v)[\boldsymbol{\Sigma}_1\otimes_{\text{sym}}\boldsymbol{\Sigma}_2]=[\mathbf{i}(v)\boldsymbol{\Sigma}_1]\otimes_{\text{sym}}\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_1\otimes_{\text{sym}}[\mathbf{i}(v)\boldsymbol{\Sigma}_2]. \tag{3.56}$$

Its realization on $\tilde{\mathcal{W}}$ reads

$$\mathbf{i}(v)\boldsymbol{\Sigma}[B]=\boldsymbol{\Sigma}\left[\kappa^{-1}v^a\frac{\partial}{\partial y^a}B\right]. \tag{3.57}$$

We also define the operators

$$e_a\neg\equiv\mathbf{i}(\partial_a),\quad e^a\neg\equiv\omega^{ab}\mathbf{i}(\partial_b) \tag{3.58}$$

with the basis vectors $\partial_a\equiv\partial/\partial\phi^a$ referring to a system of Darboux local coordinates.

The most important properties of the fermionic Weyl algebra $\mathcal{W}^F$ were the three different product structures with which it is endowed and which make it an Atiyah–Kähler algebra. The bosonic Weyl algebra $\mathcal{W}$ is equipped with three analogous products (pointwise multiplication, star product, inner product) which satisfy the basic consistency condition (3.32). At the end of Sec. III A this led us to the conclusion that $\mathcal{W}$ is the symplectic counterpart of an Atiyah–Kähler algebra. In the same sense the ISTs $\boldsymbol{\Sigma}$ are analogous to the Dirac–Kähler fields Φ.

The product structures on $\mathcal{W}_\phi$ give rise to related products on the space of symmetric tensor fields. One easily verifies that the pointwise product of bosonic symbols is tantamount to the symmetric tensor product:

$$\boldsymbol{\Sigma}[B_1]\otimes_{\text{sym}}\boldsymbol{\Sigma}[B_2]=\boldsymbol{\Sigma}[B_1B_2]. \tag{3.59}$$

Furthermore, guided by our experience with the fermionic case, we now *define* the Clifford product for symmetric tensor field as the image of the bosonic star product under the symbol/tensor correspondence (3.54):

$$\boldsymbol{\Sigma}[B_1]\vee\boldsymbol{\Sigma}[B_2]=\boldsymbol{\Sigma}[B_1\circ B_2]. \tag{3.60}$$

By construction, the ''symplectic Clifford product,'' also denoted ''$\vee$,'' is associative and distributive (but not commutative). From Eqs. (3.49), (3.57), and (3.58) one obtains the following explicit representation for the product of two ISTs:

$$\boldsymbol{\Sigma}_1\vee\boldsymbol{\Sigma}_2=\sum_{p=0}^{\infty}\frac{i^p}{p!}[e_{a_1}\neg e_{a_2}\neg\cdots e_{a_p}\neg\boldsymbol{\Sigma}_1]\otimes_{\text{sym}}[e^{a_1}\neg e^{a_2}\neg\cdots e^{a_p}\neg\boldsymbol{\Sigma}_2]. \tag{3.61}$$

This equation is strikingly similar to Kähler's formula (1.9) for the ordinary Clifford product. We emphasize that while Eq. (3.61) might look complicated it is uniquely determined by the fundamental relations

$$\begin{gathered}1\vee 1=1,\quad 1\vee d\phi^a=d\phi^a\vee 1=d\phi^a,\\ d\phi^a\vee d\phi^b=d\phi^a\otimes_{\text{sym}}d\phi^b+i\omega^{ab}\end{gathered} \tag{3.62}$$

if associativity and distributivity are imposed.

Turning to the last product structure on $\mathcal{W}$, there is an obvious choice for a fiberwise inner product $(\cdot,\cdot)$ of symmetric tensor fields: $(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)=(B_1|B_2)$ where $\boldsymbol{\Sigma}_{1,2}$ is related to $B_{1,2}$ via (3.54). Thus it is clear that the ISTs may be regarded as sections through a ''symplectic Atiyah–Kähler bundle.''

The left-multiplication by the basis element $d\phi^a$ reads explicitly

$$d\phi^a \vee \mathbf{\Sigma} = d\phi^a \otimes_{\rm sym} \mathbf{\Sigma} + i e^a \neg \mathbf{\Sigma}. \tag{3.63}$$

It defines a representation of the symplectic Clifford algebra in the space of inhomogeneous symmetric tensor fields:

$$d\phi^a \vee d\phi^b - d\phi^b \vee d\phi^a = 2i\omega^{ab}. \tag{3.64}$$

Comparing (3.64) to (3.33), $d\phi^a \vee$ takes the place of the metaplectic Dirac matrix γ^a. Since $d\phi^a = \mathbf{\Sigma}[\kappa y^a]$, $d\phi^a \vee$ applied to tensors is the same as $\kappa y^a \circ$ applied to symbols:

$$d\phi^a \vee \mathbf{\Sigma}[B] = \mathbf{\Sigma}[\kappa y^a \circ B] = \mathbf{\Sigma}[C^a B]. \tag{3.65}$$

The differential operators C^a were introduced in Eq. (3.23). They are formally self-adjoint with respect to the inner product $(\cdot|\cdot)$. They constitute a representation of the symplectic Clifford algebra in space of bosonic Weyl symbols:

$$C^a C^b - C^b C^a = 2i\omega^{ab}. \tag{3.66}$$

Since κy^a is the symbol of $\kappa \hat{\varphi}^a = \gamma^a$, the operator associated to $C^a B$ is $\gamma^a \hat{B}$ with $\hat{B} = \mathrm{symb}^{-1}(B)$. In summary, we have the chain of correspondences

$$d\phi^a \vee \mathbf{\Sigma} \leftrightarrows \gamma^a \hat{B} \leftrightarrows C^a B. \tag{3.67}$$

Thus we managed to implement the essence of the Dirac–Kähler idea in a symplectic rather than a Riemannian setting. We constructed a representation of the corresponding Clifford algebra on the space of symmetric tensor fields over a phase-space manifold rather than on the exterior algebra over space–time.

Up to this point our considerations focused on the kinematic aspects of the theory. We have not yet found an analog of the DK equation. Since d and $d^\dagger$ do not exist for symmetric tensors, the DK operator $d - d^\dagger$ has no direct counterpart. Still it is possible to write down a ''symplectic DK equation'' with the necessary covariance properties:

$$[d\phi^a \vee \nabla_\alpha + m]\mathbf{\Sigma} = 0. \tag{3.68}$$

(Here ∇ is a symplectic connection.) Equation (3.68) could be rewritten as a set of ''metaplectic Dirac equations'' in the same way as the ordinary DK equation can be decomposed into a set of ordinary Dirac equations. Metaplectic Dirac operators have been investigated in the mathematical literature recently[41] but no physical application has emerged so far. In Sec. IV we shall see that from a kinematical and representation theory point of view the symplectic DK fields indeed do play an important role in the gauge theory approach to quantization. The interpretation of field equations such as Eq. (3.68), if any, will remain an open problem though.

We close this section with a few comments on the ''metaplectic γ_5 matrix'' in relation to the DK fields. In the SO(n) case we saw that γ_{n+1}, the volume form, and the δ function are different guises of the same object. Some properties of γ_{n+1} are similar in the symplectic case, others are quite different. The symbol of γ_P, too, is proportional to a δ function,

$$G_P \equiv \mathrm{symb}(\gamma_P), \quad G_P(y) = (\pi\hbar)^N \delta^{(2N)}(y). \tag{3.69}$$

This symbol is completely unrelated to the volume form, however. In the SO(n) case we know that the Clifford right multiplication by G_{n+1} is equivalent to the modified Hodge operator ($\star f \propto f \circ G_{n+1} \propto f \circ \delta^{(n)}$). This property has a partial analog since by virtue of (3.21) the symplectic Fourier transformation which corresponds to $\star$ is essentially the same operation as the star multiplication by G_P from the right:

$$\tilde{b}(y) = 2^{-N}(b \circ G_P)(y/2). \tag{3.70}$$

However, this statement on the space of symbol functions (including distributions) does not imply a corresponding relation for symmetric tensors. The symbol G_P has no IST associated to it.

The matrix γ_P makes its appearance also in the natural inner product on $\mathcal{L}(\mathcal{V}_\phi)$. By virtue of the identity

$$(B_1|B_2)=2^N \operatorname{Tr}[\hat{B}_1^\dagger \hat{B}_2 \gamma_P] \tag{3.71}$$

the inner product on $\mathcal{W}_\phi$ induces a corresponding product for the operators. The latter differs form the familiar Hilbert–Schmidt inner product by the additional γ_P matrix which tends to improve the regularity properties of the trace. Equation (3.71) is most easily proven as follows:

$$\begin{aligned}(B_1|B_2)&=(\overline{B_1}\circ B_2)(y=0)\\ &=(\pi\hbar)^{-N}\int d^{2N}y(\overline{B_1}\circ B_2)(y)G_P(y)\\ &=(\pi\hbar)^{-N}\int d^{2N}y(\overline{B_1}\circ B_2\circ G_P)(y)\\ &=(\pi\hbar)^{-N}\int d^{2N}y[\operatorname{symb}\{\hat{B}_1^\dagger\hat{B}_2\gamma_P\}](y)=2^N\operatorname{Tr}[\hat{B}_1^\dagger\hat{B}_2\gamma_P].\end{aligned} \tag{3.72}$$

Here (3.69) was used along with the standard results[24] $\int d^{2N}y\, b_1(y) b_2(y)=\int d^{2N}y(b_1\circ b_2)(y)$ and $\operatorname{Tr}(\hat{b})=(2\pi\hbar)^{-N}\int d^{2N}y\, b(y)$.

D. Decomposition of the symplectic DK representation

We have seen that $d\phi^a\vee$ and $\kappa y^a\circ$ induce a representation of the symplectic Clifford algebra on the space of symmetric tensors and their symbols, respectively. We also saw that the corresponding representations in the SO(n) case are reducible, so it is natural to ask if the same is true in the symplectic setting. We shall demonstrate that *at the level of the symbols* the representation is indeed reducible. However, in contradistinction to the SO(n) case, the decomposition of $\mathcal{W}$ does not induce a concomitant decomposition of the (symmetric) tensor algebra.

We shall see that the representation of the symplectic Clifford algebra carried by the symbol-valued fields $B(\phi,y)$ can be decomposed into infinitely many irreducible representations each of which is equivalent to the one defined by the metaplectic γ matrices (3.35). (We recall that this is the representation of the Heisenberg algebra used in conventional canonical quantization.) As a consequence, every field $B(\phi,y)$ amounts to a collection of infinitely many metaplectic spinor fields $\psi^\alpha(\phi)$. Now we discuss the question of the (ir)reducibility for the symbols B, the operators $\hat{B}$ and the tensors Σ separately.

1. Symbols

We are going to show that the bosonic Weyl algebra $\mathcal{W}$ admits an orthogonal decomposition

$$\mathcal{W}=\bigoplus_{\alpha\in\mathbf{R}^N}\mathcal{W}_{(\alpha)} \tag{3.73}$$

such that the subspaces $\mathcal{W}_{(\alpha)}$ are invariant under star-left multiplication by y^a, i.e., $y^a\circ b\in\mathcal{W}_{(\alpha)}$ if $b\in\mathcal{W}_{(\alpha)}$. To this end we use an infinite dimensional generalization of the Becher–Joos method.[3] We look for a $2N$-parameter family of operators $\hat{Z}(y)$, $y\in\mathbf{R}^{2N}$, with the property

$$y^a\circ\hat{Z}(y)=\hat{Z}(y)\hat{\varphi}^a. \tag{3.74}$$

One should think of $\hat{Z}(\cdot)$ as an operator-valued symbol, i.e., the ''$y^a \circ$'' in (3.74) is given by $\kappa^{-1}C^a$ as if $\hat{Z}$ was an ordinary symbol. With our experience from the fermionic case we suspect that $\hat{Z}$ should be closely related to the Weyl operators. It turns out that this is indeed the case. The derivative of the Weyl operators reads

$$\frac{\partial}{\partial y^a}\hat{T}(y)=\frac{i}{\hbar}\,\omega_{ab}\left[\hat{\varphi}^b-\frac{1}{2}y^b\right]\hat{T}(y)=\frac{i}{\hbar}\,\omega_{ab}\hat{T}(y)\left[\hat{\varphi}^b+\frac{1}{2}y^b\right]. \tag{3.75}$$

Equation (3.75) entails that the argument of $\hat{T}$ can be rescaled in such a way that left multiplication with y^a is equivalent to the operator multiplication by $\hat{\varphi}^a$, either from the left or from the right:

$$\begin{aligned} y^a\circ\hat{T}(\pm 2iy)&=\mp i\hat{\varphi}^a\hat{T}(\pm 2iy),\\ y^a\circ\hat{T}(\pm 2y)&=\mp\hat{T}(\pm 2y)\hat{\varphi}^a. \end{aligned} \tag{3.76}$$

Hence

$$\hat{Z}(y)=\hat{T}(-2y)=\exp(-i\kappa y^a\omega_{ab}\gamma^b) \tag{3.77}$$

is a solution to our problem. In the $\hat{x}$ eigenbasis the matrix elements $\hat{Z}(y)^\alpha{}_\beta=\langle\alpha|\hat{Z}|\beta\rangle$ are given by

$$\hat{Z}(y)^\alpha{}_\beta=\exp\left[-\frac{i}{\hbar}y_p(\alpha+\beta)\right]\delta^{(N)}(\alpha-\beta+2y_q). \tag{3.78}$$

They can be used in order to verify that

$$\begin{aligned} \langle\alpha|y_q\circ\hat{Z}(y)|\beta\rangle&=\beta\langle\alpha|\hat{Z}(y)|\beta\rangle,\\ \langle\alpha|y_p\circ\hat{Z}(y)|\beta\rangle&=i\hbar\frac{\partial}{\partial\beta}\langle\alpha|\hat{Z}(y)|\beta\rangle, \end{aligned} \tag{3.79}$$

which is (3.74) in the ''position representation.''

We shall need the star product of two different $\hat{Z}$ matrix elements. After some algebra one finds the remarkably simple result

$$\hat{Z}(y)^\alpha{}_\beta\circ\hat{Z}(y)^{\bar{\alpha}}{}_{\bar{\beta}}=2^{-N}(\gamma_P)^{\alpha}{}_{\bar{\beta}}\hat{Z}(y)^{\bar{\alpha}}{}_\beta. \tag{3.80}$$

When combined with the identity $\hat{Z}^\dagger=\gamma_P\hat{Z}\gamma_P$ Eq. (3.80) at $y=0$ gives rise to the inner product

$$(\hat{Z}^\alpha{}_\beta|\hat{Z}^{\bar{\alpha}}{}_{\bar{\beta}})=2^{-N}(\gamma_p)^{\alpha\bar{\alpha}}\delta_{\beta\bar{\beta}}. \tag{3.81}$$

The orthogonality and completeness relations (3.6) and (3.7) for $\hat{T}(y)$ imply similar relations for $\hat{Z}(y)$. As a consequence, $\{\hat{Z}(\cdot)^\alpha{}_\beta|\alpha,\beta\in\mathbf{R}^N\}$ is a basis in the space of symbol functions $b(\cdot)$. Every $b\in\mathcal{W}$ has an expansion of the form $b(y)=\int d^N\alpha d^N\beta\psi^\alpha_{(\beta)}\hat{Z}(y)^\beta{}_\alpha$ where the ''coefficients'' $\psi^\alpha_{(\beta)}$ are actually functions $\mathbf{R}^N\times\mathbf{R}^N\to\mathbf{C}$.

We continue the discussion directly for the case when $\mathcal{W}$ is the fiber $\mathcal{W}_\phi$ and the symbols $b(\cdot)$ are the $\mathcal{W}$-valued fields $B(\cdot,\phi)$ evaluated at a given point ϕ. Equations (3.6) and (3.7) imply that B can be expanded as

$$B(\phi,y)=\int d^N\alpha\int d^N\beta\psi^\alpha_{(\beta)}(\phi)\hat{Z}(y)^\beta{}_\alpha \tag{3.82}$$

and that the expansion coefficients are given by

$$\psi^{\alpha}_{(\beta)}(\phi)=(\pi\hbar/2)^{-N}\int d^{2N}y B(\phi,y)\hat{Z}^{\dagger}(y)^{\alpha}{}_{\beta}. \tag{3.83}$$

In a sense which we shall make precise later on, $\psi_{(\beta)}\equiv\{\psi^{\alpha}_{(\beta)};\alpha\in\mathbf{R}^N\}$ are the components of infinitely many metaplectic spinors labeled by the ''index'' β. If we define

$$B_{(\beta)}(\phi,y)\equiv\int d^N\alpha\,\psi^{\alpha}_{(\beta)}(\phi)\hat{Z}(y)^{\beta}{}_{\alpha} \tag{3.84}$$

so that $B(\phi,y)=\int d^N\beta B_{(\beta)}(\phi,y)$ then Eq. (3.74) implies that the invariant subspace $\mathcal{W}_{(\beta)}$ is spanned by precisely the symbols of the type (3.84):

$$y^a\circ B_{(\beta)}=\int d^N\alpha\int d^N\overline{\alpha}\psi^{\alpha}_{(\beta)}\hat{Z}^{\beta}{}_{\overline{\alpha}}(\hat{\varphi}^a)^{\overline{\alpha}}{}_{\alpha}=\int d^N\overline{\alpha}(\hat{\varphi}^a\psi_{(\beta)})^{\overline{\alpha}}\hat{Z}^{\beta}{}_{\overline{\alpha}}. \tag{3.85}$$

Here $(\hat{\varphi}^a\psi_{(\beta)})^{\overline{\alpha}}\equiv\int d^N\alpha(\hat{\varphi}^a)^{\overline{\alpha}}{}_{\alpha}\psi^{\alpha}_{(\beta)}$. We see that if the symbol $B_{(\beta)}$ is related to the spinor $\psi_{(\beta)}$ by (3.84) then $y^a\circ B_{(\beta)}$ and $\hat{\varphi}^a\psi_{(\beta)}$ are related in the same way. Likewise $\kappa y^a\circ$ corresponds to a multiplication by γ^a.

Given an arbitrary symbol in $\mathcal{W}$ we can project it on any of the subspaces $\mathcal{W}_{(\beta)}$. We introduce projection operators $\mathcal{P}_{(\beta)}$ by $B_{(\beta)}=\mathcal{P}_{(\beta)}B$. If we combine Eqs. (3.83) and (3.84) it follows that

$$B_{(\beta)}(\phi,y)=\int d^{2N}y'\,\mathcal{P}_{(\beta)}(y,y')B(\phi,y'), \tag{3.86}$$

where the integral kernel of the projector is given by

$$\mathcal{P}_{(\beta)}(y,y')=(\pi\hbar/2)^{-N}\langle\beta|\hat{Z}(y)\hat{Z}^{\dagger}(y')|\beta\rangle. \tag{3.87}$$

Upon using (3.77), (3.5), and (3.9) we obtain explicitly

$$\mathcal{P}_{(\beta)}(y,y')=(\pi\hbar)^{-N}\exp\left[-\frac{2i}{\hbar}(\beta+y_q)(y_p-y'_p)\right]\delta^{(N)}(y_q-y'_q). \tag{3.88}$$

The projectors $\{\mathcal{P}_{(\beta)};\beta\in\mathbf{R}^N\}$ are orthogonal and complete in the sense that

$$\begin{gathered}\int d^{2N}y'\,\mathcal{P}_{(\beta)}(y,y')\mathcal{P}_{(\overline{\beta})}(y',y'')=\delta^{(N)}(\beta-\overline{\beta})\mathcal{P}_{(\beta)}(y,y''),\\ \int d^N\beta\,\mathcal{P}_{(\beta)}(y,y')=\delta^{(2N)}(y-y').\end{gathered} \tag{3.89}$$

Furthermore, as a consequence of Eq. (3.80), the inner product of two different projections reads

$$(B_{(-\beta_1)}|B_{(\beta_2)})=2^{-N}\delta^{(N)}(\beta_1-\beta_2)\int d^N\alpha\,\overline{\psi}^{\alpha}_{(\beta_1)}\psi^{\alpha}_{(\beta_2)}. \tag{3.90}$$

Note the sign flip on the left-hand side of Eq. (3.90). Obviously $B_{(-\beta)}$ is the natural dual of $B_{(\beta)}$ (similar to a spinor adjoint).

To summarize: Every symbol-valued field $B(\phi,y)$ gives rise to infinitely many projections $B_{(\beta)}(\phi,y)$ each of which is equivalent to a metaplectic spinor field $\psi_{(\beta)}(\phi)$ with components $\psi^{\alpha}_{(\beta)}(\phi)$ given by (3.83). This is to mean that the fields $\psi_{(\beta)}$ carry an irreducible representation of the Clifford algebra: $\kappa y^a\circ B_{(\beta)}$ corresponds to the spinor multiplied by a γ matrix, $\gamma^a\psi_{(\beta)}$.

Up to this point the situation is similar to the SO(n) case, but differences will show up shortly.

2. Operators

As in the fermionic case, it proves advantageous to combine the expansion coefficients $\psi^{\alpha}_{(\beta)}$ as a matrix $\hat{\psi}$:

$$\hat{\psi}^{\alpha}{}_{\beta} \equiv \psi^{\alpha}_{(\beta)} \equiv \langle \alpha | \hat{\psi} | \beta \rangle. \tag{3.91}$$

(We suppress the argument ϕ for the time being.) We shall need some properties of the linear, invertible map $B \mapsto \hat{\psi}[B]$ which relates the symbols to the new operator $\hat{\psi}$.

By definition, $B(y)$ is the *ordinary* Weyl symbol of the operator $\hat{B}$ introduced earlier. Remarkably enough, this symbol plays a dual role: the same function but with its argument rescaled, $B(\frac{1}{2}y)$, turns out to be the *alternative* Weyl symbol of the new operator $\hat{\psi}$. This is most easily seen if one uses $\hat{Z}^{\dagger}(y) = \hat{T}(2y)$ in

$$\hat{\psi} = (\pi\hbar/2)^{-N} \int d^{2N}y B(y) \hat{Z}^{\dagger}(y), \tag{3.92}$$

$$B(y) = \mathrm{Tr}[\hat{Z}(y)\hat{\psi}], \tag{3.93}$$

which follows from the equations in Sec. III D 1, and then compares (3.92), (3.93) to Eqs. (3.10) and (3.11). Thus,

$$[\mathrm{symb}\{\hat{B}\}](y) = B(y) \Leftrightarrow [\text{alt-symb}\{\hat{\psi}\}](y) = B(\tfrac{1}{2}y). \tag{3.94}$$

This dual role played by B is another hint at the very natural relationship between the Dirac–Kähler idea and the Weyl symbol calculus.

Regarding $\hat{\psi}$ as a functional of B it is not difficult to establish that

$$\hat{\psi}[1] = 2^N \gamma_P, \tag{3.95}$$

$$\hat{\psi}[\kappa y^a] = 2^N \gamma^a \gamma_P, \tag{3.96}$$

$$\hat{\psi}[\kappa y^a \circ B] = \gamma^a \hat{\psi}[B], \tag{3.97}$$

$$\hat{\psi}[B_1 \circ B_2] = 2^{-N} \hat{\psi}[B_1] \gamma_P \hat{\psi}[B_2], \tag{3.98}$$

$$\hat{\psi}[\kappa^p y^{a_1} \circ y^{a_2} \circ \cdots y^{a_p}] = 2^N \gamma^{a_1} \gamma^{a_2} \cdots \gamma^{a_p} \gamma_P, \tag{3.99}$$

$$\hat{\psi}[\kappa^2 y^a y^b] = 2^N \gamma^a \gamma^b \gamma_P - 2^N i \omega^{ab} \gamma_P. \tag{3.100}$$

Equation (3.95) follows directly from the definition of γ_P and Eq. (3.97) is our earlier result (3.85), Eq. (3.96) being a special case. The most important relation is (3.98). It can be proven by using (3.80) and (3.93) in order to show that $B_1 \circ B_2 = 2^{-N} \mathrm{Tr}\{\hat{Z}\hat{\psi}[B_1]\gamma_P \hat{\psi}[B_2]\}$. When compared to Eq. (3.93), this equation implies (3.98).

Above we had introduced the projectors $\mathcal{P}_{(\beta)}$ which project any symbol on the invariant subspaces $\mathcal{W}_{(\beta)}$. The map $B \mapsto \hat{\psi}[B]$ given by (3.92) induces a corresponding projection on the space of operators $\hat{\psi}$. In the language of our auxiliary quantum mechanical system this projection has a very natural interpretation: it is simply the projection on the position eigenstate $|\beta\rangle$. From Eq. (3.84) we can read off that $B_{(\beta)}$ has the structure of an expectation value in the state $|\beta\rangle$,

$$B_{(\beta)}(y)=\langle\beta|\hat{Z}(y)\hat{\psi}[B]|\beta\rangle=\mathrm{Tr}[\hat{Z}(y)\hat{\psi}[B]\hat{P}_{(\beta)}]. \tag{3.101}$$

Here $\hat{P}_{(\beta)}\equiv|\beta\rangle\langle\beta|$ is the corresponding projector on the Hilbert space. It follows from (3.101) that symbols $B\in\mathcal{W}_{(\beta)}$ are associated to operators of the form $\hat{\psi}P_{(\beta)}$:

$$\hat{\psi}[\mathcal{P}_{(\beta)}B]=\hat{\psi}[B]\hat{P}_{(\beta)}. \tag{3.102}$$

Finally we have to address the important question of how the operator $\hat{\psi}$ is related to the operator $\hat{B}$ which was the central building block in the Dirac–Kähler construction. Imitating the SO(n) case, we had obtained $\hat{B}$ in Eq. (3.52) by replacing $d\phi^a\rightarrow\gamma^a$ in the tensor field $\mathbf{\Sigma}$. In Sec. II we have seen that for ordinary DK fields $\hat{\psi}$ and $\hat{F}$ coincide up to a constant factor. It is quite remarkable that, with a minor modification, the same identification is possible in the symplectic situation where $\mathcal{V}$ is infinite dimensional. It turns out that

$$\hat{\psi}[B]=2^N\hat{B}\gamma_P$$

or

$$\hat{B}=2^{-N}\hat{\psi}[B]\gamma_P. \tag{3.103}$$

This relationship can be proven in a variety of ways. For instance, we can take advantage of the following very compact representation of operators $\hat{b}$ in terms of their symbols b:[32]

$$\hat{b}=2^{-N}(2\pi\hbar)^{-N}\int d^{2N}y\,b(\tfrac{1}{2}y)\hat{T}(y)\gamma_P. \tag{3.104}$$

The advantage of (3.104) as compared to the old representation (3.10) is that no Fourier transformation is involved any longer. Equation (3.104) is easily established by inserting the integral representation for γ_P on its right-hand side and then combining the two Weyl operators with the help of (3.5). From Eq. (3.104) we infer that if $B(y)$ is the ordinary symbol of $\hat{B}$ then $B(\tfrac{1}{2}y)$ is the alternative Weyl symbol of $2^N\hat{B}\gamma_P$. Moreover, we saw already that $B(\tfrac{1}{2}y)$ is the alternative Weyl symbol of $\hat{\psi}$. As a consequence, $\hat{\psi}$ must coincide with $2^N\hat{B}\gamma_P$.

It is instructive to give a different proof when $B(y)$ is a power series. This is the case for instance when the symbol originates from an IST via the DK construction. For $\hat{B}$ or $B(y)$ given, the task is to solve $B(y)=\mathrm{Tr}[\hat{Z}(y)\hat{\psi}]$ for the unknown operator $\hat{\psi}$. Using the expansion (3.53) for B and the (expanded) exponential (3.77) for $\hat{Z}$, Eq. (3.104) turns into

$$B^{(p)}_{a_1\cdots a_p}=i^{-p}\,\mathrm{Tr}[\gamma_{(a_1}\cdots\gamma_{a_p)}\hat{\psi}]. \tag{3.105}$$

In the corresponding calculation for the SO(n) case we made an ansatz for $\hat{\psi}$ as a power series in γ^μ and used the γ^μ-trace identities in order to project on its coefficients. Because of the additional matrix γ_P in the analogous identities (3.45) for the metaplectic γ matrices the appropriate ansatz for the symplectic $\hat{\psi}$ is a power series in γ^a (with coefficients $\psi^{(p)}_{a_1\cdots a_p}$) times an explicit factor of γ_P. With this ansatz in (3.105), the trace identities imply $\psi^{(p)}_{a_1\cdots a_p}=2^NB^{(p)}_{a_1\cdots a_p}$ which proves (3.103). In this manner we see that the factor of γ_P connecting $\hat{\psi}$ to $\hat{B}$ is simply a reflection of the corresponding factor in the trace identities. We discussed already that in the infinite dimensional situation the γ_P under the traces is crucial in order to make them well defined.

The γ_P matrix in (3.103) has the consequence that $\hat{\psi}$ does not admit a power series expansion even if $\hat{B}$ does so. This has important implications for the Dirac–Kähler program. As we are going to discuss next it means that the decomposition of $\mathcal{W}$ into subspaces $\mathcal{W}_{(\alpha)}$ which are invariant under star left multiplication does *not* translate into a corresponding decomposition of the symmetric tensor algebra into subspaces invariant under (symplectic) Clifford left multiplication. In this respect the SO(n) and the Sp($2N$) cases are quite different.

Let us first look at how the space of operators $\hat{B}$ decomposes under $\mathcal{W}=\oplus\mathcal{W}_{(\beta)}$. Equations (3.102) and (3.103) imply that

$$\widehat{\mathcal{P}_{(\beta)}B}=\hat{B}\,\gamma_P\hat{P}_{(\beta)}\gamma_P=\hat{B}\hat{P}_{(-\beta)}\,. \tag{3.106}$$

Hence, at the level of the $\hat{B}$ operators, the projection $\mathcal{P}_{(\beta)}$ amounts to a right multiplication by $\hat{P}_{(-\beta)}$.

From Eq. (3.106) we can obtain a very useful by-product. If we take the symbol on both sides of this equation and abbreviate $P_{(\alpha)}\equiv\mathrm{symb}[\hat{P}_{(\alpha)}]$ then the result is the following compact formula for the projection $B_{(\beta)}$:

$$\mathcal{P}_{(\beta)}B\equiv B_{(\beta)}=B\circ P_{(-\beta)}\,. \tag{3.107}$$

More explicitly, because $P_{(\alpha)}(y)=\delta^{(N)}(y_q-\alpha)$, this means that

$$B_{(\beta)}(y)=B(y)\circ\delta^{(N)}(y_q+\beta). \tag{3.108}$$

By virtue of (3.21) the latter equation can be brought to the following form which is the most convenient one for practical calculations:

$$B_{(\beta)}(y_p,y_q)=B\left(y_p-\frac{i\hbar}{2}\frac{\partial}{\partial\overline{y_q}},y_q\right)\delta^{(N)}(\overline{y_q}+\beta)\bigg|_{\overline{y_q}=y_q}. \tag{3.109}$$

As usual, $y\equiv(y_p,y_q)$ consists of N-component momentum- and position-type variables y_p and y_q.

The structure of $B_{(\beta)}$ is particularly transparent if $B(y)\equiv B(y_q)$ does not depend on the momenta. Then its projection on $\mathcal{W}_{(\beta)}$ reads

$$B_{(\beta)}(y)=B(y_q)\,\delta^{(N)}(y_q+\beta), \tag{3.110}$$

i.e., it is sharply localized at $y_q=-\beta$. If B depends also on y_p there are additional terms involving derivatives of $\delta^{(N)}(y_q+\beta)$. Nevertheless, *as long as B depends on y polynomially, the projected symbol* $B_{(\beta)}$ *has support only on the hyperplane* $y_q=-\beta$. This localization of the symbols makes it very easy to visualize the β subspace of $\mathcal{W}$. In fact, this intuitive interpretation of $\mathcal{W}_{(\beta)}$ is the reason why we are using the $\hat{x}$ eigenbasis on $\mathcal{V}$ rather than the harmonic oscillator (Fock space) basis which yields the traditional representation of the γ^a matrices.

3. Inhomogeneous symmetric tensors

We know that every symbol-valued field $B(\phi,y)$ gives rise to infinitely many projections $B_{(\beta)}\in\mathcal{W}_{(\beta)}$ each of which is equivalent to a spinor $\psi_{(\beta)}$. On $\mathcal{W}_{(\beta)}$, $\kappa y^a\circ B_{(\beta)}$ corresponds to $\gamma^a\psi_{(\beta)}$ and it represents the Clifford algebra irreducibly. On the other hand, in Sec. III C we defined the symplectic Clifford product as the image of the star product under the symbol/tensor-correspondence (3.54). It is a natural question therefore whether the representation of the Clifford algebra provided by ''$d\phi^a\vee$'' on the space of symmetric tensors is reducible as well.

At this point we have to recall that the symbol/tensor-correspondence (3.54) is a bijection between tensors $\mathbf{\Sigma}(\phi)$ and symbols $B(\phi,y)$ which are *analytic in y*. Only if B allows for a power series expansion in y the substitution $\kappa y^a\to d\phi^a$ yields a tensor field. As for the question of the reducibility, the crucial observation is that *even if* $B(\phi,y)$ *is analytic in y, the projections* $B_{(\beta)}(\phi,y)$ *are not in general*. This is obvious from Eq. (3.109) which shows that $B_{(\beta)}$ is typically a distribution with a sharp localization (in the auxiliary phase space) on the plane $y_q=-\beta$.

Therefore we must conclude that the decomposition of the bosonic Weyl algebra $\mathcal{W} = \oplus \mathcal{W}_{(\beta)}$ does not imply a corresponding decomposition of the space of ISTs. This was different in the fermionic case where the analyticity of $F(x,\theta)$ comes for free and where ''symbol-valued fields'' and ''inhomogeneous differential forms'' are two completely equivalent concepts.

From these observations we can learn what the correct notion of a ''symplectic Dirac–Kähler field'' actually is. Traditionally, in the SO(N) case, a DK field meant a set of (antisymmetric) tensor fields. This is a historic accident, however, and one could have talked equally well about $\mathcal{W}^F$-valued fields over space–time. When we go from space–time to phase space and from SO(n) to Sp($2N$) we see that the notion which generalizes is not that of a collection (of now symmetric) tensor fields but rather the idea of Weyl symbol-valued fields. On phase space the fields $B(\phi,y)$, with a not necessarily analytic dependence on y, play a role which is completely analogous to that of $F(x,\theta)$ on space–time. The former is equivalent to a set of Mp($2N$) spinors in very much the same way as the latter gives rise to a multiplet of Spin(n) spinors.

IV. SUMMARY AND DISCUSSION

In the first part of this paper we have shown that the theory of space–time DK fermions allows for a remarkably simple and natural reinterpretation in the framework of the symbol calculus. More precisely, it is the fermionic Weyl symbol which is to be used here. This symbol was employed in the context of first quantized particle and string theory occasionally, but so far it has not reached the popularity of the Wick symbol which is commonly chosen for fermionic systems.

We have set up a one-to-one correspondence between DK fields $\Phi(x)$ and symbol-valued fields $F(x,\theta)$ by associating a family of auxiliary quantum systems, with canonical operators $\hat{\chi}^\mu$ and anticommuting phase-space coordinates θ^μ, to each point x of space–time. The fermionic operators $\hat{\chi}^\mu$ and Grassmann variables θ^μ replace the Dirac matrices γ^μ and the differentials dx^μ, respectively. The nontrivial aspect of this correspondence is that it maps all the natural operations which we know for differential forms onto equally natural and well-known operations for symbols. For instance, the star product which is at the heart of every symbol calculus turned out to be related to the Clifford product, a pivotal concept in standard DK theory, in precisely this manner. More generally, we were able to identify all the defining structures of an Atiyah–Kähler algebra on the space of fermionic Weyl symbols.

Our approach provides some new computational tools for calculations involving DK fields, an integral representation of the Clifford product, for example. More important, it sheds new light on the geometrical meaning of various constructions in the standard approach. For instance, the matrix-valued form Z has turned out to be nothing but a fermionic Weyl operator.

In the second part of this paper we developed a symplectic analog of DK theory. We replaced space–time by phase space, the ''Lorentz group'' SO(n) by Sp($2N$), Dirac fields by metaplectic spinors, and we then asked if there exists a corresponding notion of a DK field. The answer turned out to be in the affirmative, but with some qualifications. The crucial step in our construction was switching from the fermionic auxiliary quantum system to a bosonic one whose basic operators $\hat{\varphi}^a$ satisfy canonical commutation relations and thus realize the symplectic Clifford algebra. Using the Riemannian situation as a guideline we formulated the auxiliary quantum theory in terms of (now bosonic) Weyl symbols. We argued that it is the symbol-valued fields $B(\phi,y)$ which deserve the name of a ''symplectic Dirac–Kähler field.'' The fields with an analytic dependence on y are equivalent to a set of symmetric tensor fields, the symplectic counterpart of an inhomogeneous differential form. We described in detail which properties of the standard DK fields pass over to the symplectic case and which do not. We discovered for example that all the defining structures of an Atiyah–Kähler algebra have analogs in the symplectic setting. In particular, the bosonic Weyl star product gives rise to a ''Clifford product.''

It is an interesting feature of this method that both the ordinary and the symplectic Clifford product arise as a quantum deformation (in the sense of Ref. 24) of the corresponding tensor product (wedge product and $\otimes_{\text{sym}}$), the deformation parameter being $\hbar$ or κ^{-2}. (In order to make this explicit at the tensor level one should refrain from the convenient rescaling of the tensor components by factors of κ.)

The most important differences between the Riemannian and the symplectic case occur when it comes to decomposing the representation of the Clifford algebra carried by the symbol-valued fields. While the decomposition of the bosonic Weyl algebra into left invariant subspaces can be carried out along the same lines as for the fermionic algebra, it does not induce a corresponding decomposition of the space of inhomogeneous symmetric tensor fields. The reason is that the projection on the invariant subspaces does not respect the analyticity of $B(\phi,y)$ which is necessary for a tensor interpretation. We take this as a hint that it is actually the concept of a Weyl-algebra-valued field which is at the heart of DK theory, both on space–time and on phase space, rather than the idea of inhomogeneous (anti)symmetric tensors. The fields $B(\phi,y)$ are equivalent to a multiplet of metaplectic spinors in the same way an ordinary DK field is equivalent to a multiplet of Dirac spinors.

Let us close with a few additional comments.

We begin with a remark on what it precisely means that a DK field is ''equivalent'' to a set of spinor fields. This remark applies to SO(n) and Sp($2N$) DK fields alike. Strictly speaking, a metaplectic spinor is defined by its transformation properties under local Sp($2N$) transformations, the phase-space analog of the local Lorentz transformations. (See Ref. 21 for a detailed discussion of those transformation properties and of the vielbein formalism for phase spaces.) Let us fix some point ϕ of $\mathcal{M}_{2N}$ and let us change the basis in its tangent space $T_\phi\mathcal{M}_{2N}$ by means of a symplectic matrix $S(\phi)\equiv[S(\phi)^a{}_b]$. This induces a corresponding unitary transformation $M(S)\in \mathrm{Mp}(2N)$ in the local Hilbert space $\mathcal{V}_\phi$. The components of a vector and a spinor transform as $y^a \to (S^{-1})^a{}_b y^b$ and $\psi^\alpha \to M(S)^\alpha{}_\beta \psi^\beta$, respectively. It is important to observe that the spinors contained in a DK field $B(\phi,y)$ do not individually transform in this manner. In fact, as a direct consequence of (1.2) the $\hat{Z}$ operator transforms according to

$$M(S)^\dagger \hat{Z}(y) M(S) = \hat{Z}(S^{-1}y). \tag{4.1}$$

Therefore Eq. (3.93) reads in the rotated basis

$$B(\phi,S^{-1}y) = \mathrm{Tr}[\hat{Z}(y) M(S) \hat{\psi}(\phi) M(S)^\dagger]. \tag{4.2}$$

This means that $\hat{\psi}\equiv(\psi^\alpha_{(\beta)})$ does not transform as a set of independent spinors labeled by the index β. The index β, too, is acted upon by a spin matrix:

$$\psi^\alpha_{(\beta)} \to M(S)^\alpha{}_\gamma \psi^\gamma_{(\delta)} M^\dagger(S)^\delta{}_\beta. \tag{4.3}$$

For space–time DK fields this is a well-known phenomenon which is referred to as ''flavor mixing.''[3] Among other things it implies that DK fermions have a nonstandard coupling to gravity.[2,42,6] The curved-space Dirac equation for a massless DK field reads

$$\gamma^\mu(\partial_\mu \hat{\psi} - i\omega_\mu^{IJ}[\sigma_{IJ},\hat{\psi}]) = 0. \tag{4.4}$$

If $\hat{\psi}$ was a set of independent spinors the spin connection $\omega_\mu^{IJ}\sigma_{IJ}$ multiplied $\hat{\psi}$ from the left only. The flavor mixing caused by the commutator is weak in the Newtonian limit of gravity and most probably cannot be excluded on experimental grounds.[42]

Finally let us comment on the relation of the symplectic DK fields to the gauge theory formulation of quantum mechanics[21] which was proposed recently. Its basic ingredient is a family of local Hilbert spaces $\mathcal{V}_\phi$ attached to the points of phase space. (For different formulations of quantum (field) theory using local Hilbert spaces see Refs. 43, 31, and 44.) This theory resulted from an attempt at understanding the principles of canonical quantization at a perhaps deeper or at least physically and geometrically more natural level.

The theory is a Yang–Mills-type gauge theory on phase space. Its ''matter fields'' are metaplectic spinors $\psi^\alpha(\phi)$. Canonical quantization is replaced by two new rules. The first one is that in order to go from classical mechanics to semiclassical quantum mechanics we must switch from the vector representation of $\mathrm{Sp}(2N)$ to its spinor representation. The second rule is a consistency condition which tells us how to sew together local semiclassical expansions so as to recover exact quantum mechanics. It is formulated as symmetry principle: the Yang–Mills theory must be invariant under a new type of background-quantum split symmetry. As it turns out, this implies that the gauge field is a universal, nondynamical Abelian connection $\tilde{\Gamma}$. (The gauge group is the group of all unitary transformations on $\mathcal{V}$ and the connection components $\tilde{\Gamma}_a$ are Hermitian operators. Being Abelian means that the curvature of $\tilde{\Gamma}$ is proportional to the unit operator on $\mathcal{V}$.)

The upshot of this construction is the following two-step procedure for the quantization of physical systems on arbitrary curved phase spaces $\mathcal{M}_{2N}$

(1) Find an Abelian spin connection $\tilde{\Gamma}$ on $\mathcal{M}_{2N}$. It is guaranteed to exist on any symplectic manifold and can be constructed iteratively by Fedosov's method.[29,30,38]

(2) Construct (multi) spinor fields which are covariantly constant (possibly up to a phase) with respect to the connection $\tilde{\Gamma}$. They are local generalizations of states and observables.

In particular, states are represented by a covariantly constant spinor field $\psi^\alpha(\phi)$. If the value of this field is known at a fixed reference point ϕ_0 it is known everywhere in phase space (up to a physically irrelevant phase). The wave function $\mathbf{\Psi}$ of conventional quantum mechanics is identified with $\psi^\alpha(\phi_0)\equiv\mathbf{\Psi}(\alpha)$. For further details we refer to Ref. 21.

This approach reveals that, in a sense, classical mechanics is related to quantum mechanics in the same way tensor fields (integer spin) relate to spinor fields (half-integer spin) or space–time bosons relate to fermions. What is at the heart of the quantization process is changing the representation of $\mathrm{Sp}(2N)$, the ''Lorentz group'' of phase space.

According to the proposal of Ref. 21 this change of representation, while very natural from a particle physics point of view, still has to be done ''by hand'' in the same sense as in the standard approach the canonical commutation relations are imposed ''by hand.'' One might wonder if there is a more natural way of describing this change of representation, and it is here that Dirac–Kähler theory comes into play. DK theory certainly cannot *explain* why nature has decided to pick the spinor representation of $\mathrm{Sp}(2N)$ but it can put this question into a novel and perhaps somewhat unexpected perspective.

The symplectic DK fields give a precise meaning to the idea that classical mechanics ''contains'' the basic building blocks of quantum mechanics, namely the metaplectic spinor fields. On the one hand, the DK fields $B(\phi,y)$ belong to the realm of classical mechanics in the sense that they are c-number functions on the classical tangent bundle. On the other hand, $B(\phi,y)$ is equivalent to a family of spinor fields $\psi_{(\beta)}=(\psi^\alpha_{(\beta)})$ whose members are labeled by the ''flavor index'' β. Quantum mechanics is a theory whose basic ingredient is a *single* metaplectic spinor field. This leads us to conclude *that the process of quantization can be understood as the elimination of all but one flavor of metaplectic spinors, i.e., as a projection on a fixed* β*-subspace*. [Note, however, that covariantly constant DK fields do not amount to covariantly constant projected spinor fields. The reason is the flavor mixing: the condition $\nabla(\tilde{\Gamma})B=0$ involves a commutator of $\tilde{\Gamma}$ with B, while $\nabla(\tilde{\Gamma})\psi=0$ contains only a left-multiplication by $\tilde{\Gamma}$.]

In the same sense as above, this projection has to be done ''by hand.'' [1] (We mention that also the approach of Ref. 45 constructs quantum mechanics from functions on the classical tangent bundle by imposing certain constraints. This approach does not involve metaplectic spinors, however, and the DK construction seems not to answer the questions raised there.) However, with this interpretation, there is an almost perfect analogy between the following two problems which are usually thought of as belonging to rather different branches of physics: the construction of a lattice theory which describes a single species of fermions, and the quantization of physical systems in general. On the Riemannian (or space–time) side, the question is how to avoid the fermion replication which results from the Kogut–Susskind action, and the corresponding symplectic (or phase-space) problem is how to obtain a quantum theory from classical structures. At a heuristic level, the solution to both problems is exactly the same: one must project out a single spinor from a Dirac–Kähler field. Whether this is merely a formal similarity or whether space-time fermions can teach us something about the general structure of quantum mechanics remains to be seen.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor H. Joos for introducing me to the Dirac–Kähler equation long ago and for many interesting discussions ever since. I am also grateful to E. Gozzi for numerous helpful conversations and for the hospitality at the Department of Theoretical Physics, University of Trieste, while this work was in progress. This project was supported in part by a NATO traveling grant.

APPENDIX A: GRASSMANN ALGEBRAS

In this appendix we collect a number of definitions and identities related to Grassmann algebras which are needed in the main body of the paper. In particular, the main automorphism $\mathcal{A}$, the main antiautomorphism $\mathcal{B}$, the Hodge operator $*$, and the modified Hodge operator $\star$ are discussed and our conventions are specified.

We consider a Grassmann algebra with the real generators $\theta^1,\dots,\theta^n$, i.e., $\theta^\mu\theta^\nu+\theta^\nu\theta^\mu=0$ for all $\mu,\nu=1,...,n$, and introduce functions

$$f(\theta)\equiv f(\theta^1,\dots,\theta^n)=\sum_{p=0}^{n} f^{(p)}(\theta), \tag{A1}$$

where $f^{(p)}$ is homogeneous of degree p:

$$f^{(p)}(\theta)=\frac{1}{p!}f^{(p)}_{\mu_1\cdots\mu_p}\theta^{\mu_1}\cdots\theta^{\mu_p}. \tag{A2}$$

The complex-valued constants $f_{\mu_1\cdots\mu_p}$ are completely antisymmetric in all indices. By definition, the main automorphism $\mathcal{A}$ and the main antiautomorphism $\mathcal{B}$ act on these functions according to

$$(\mathcal{A}f)(\theta)=\sum_{p=0}^{n}(-1)^p f^{(p)}(\theta), \tag{A3}$$

$$(\mathcal{B}f)(\theta)=\sum_{p=0}^{n}(-1)^{p(p-1)/2} f^{(p)}(\theta). \tag{A4}$$

Their main properties are

$$\begin{gathered}\mathcal{A}^2=\mathcal{B}^2=1,\quad \mathcal{A}\mathcal{B}=\mathcal{B}\mathcal{A},\\ \mathcal{A}(fg)=(\mathcal{A}f)(\mathcal{A}g),\\ \mathcal{B}(fg)=(\mathcal{B}g)(\mathcal{B}f),\end{gathered} \tag{A5}$$

where $(fg)(\theta)\equiv f(\theta)g(\theta)$ is the pointwise product. Some useful identities involving $\mathcal{A}$ and $\mathcal{B}$ include

$$\theta^\mu f(\theta)=(\mathcal{A}f)(\theta)\,\theta^\mu, \tag{A6}$$

$$\theta^{\mu_1}\cdots\theta^{\mu_p}f(\theta)=(\mathcal{A}^p f)(\theta)\,\theta^{\mu_1}\cdots\theta^{\mu_p}, \tag{A7}$$

$$\theta^{\mu_p}\theta^{\mu_{p-1}}\cdots\theta^{\mu_1}=(-1)^{p(p-1)/2}\theta^{\mu_1}\theta^{\mu_2}\cdots\theta^{\mu_p}=\mathcal{B}\theta^{\mu_1}\theta^{\mu_2}\cdots\theta^{\mu_p}. \tag{A8}$$

Denoting complex conjugation by an overbar we assume $\overline{\theta^\mu}=\theta^\mu$ and set

$$\overline{fg}=\overline{g}\,\overline{f} \tag{A9}$$

for any two functions f and g. If one makes the additional assumption that the coefficients $f^{(p)}_{\mu_1\cdots\mu_p}$ are real, then Eq. (A8) shows that

$$\overline{f(\theta)}=(\mathcal{B}f)(\theta). \tag{A10}$$

Usually we shall allow the coefficients to be complex though. The automorphism $\mathcal{A}$ can be used in order to convert right-derivatives $\overleftarrow{\partial}/\partial\theta^\mu$ to left-derivatives $\overrightarrow{\partial}/\partial\theta^\mu\equiv\partial/\partial\theta^\mu$:

$$f(\theta)\frac{\overleftarrow{\partial}}{\partial\theta^\mu}=\mathcal{A}\frac{\partial}{\partial\theta^\mu}f(\theta). \tag{A11}$$

More generally, one has

$$f(\theta)\frac{\overleftarrow{\partial}}{\partial\theta^{\mu_p}}\cdots\frac{\overleftarrow{\partial}}{\partial\theta^{\mu_1}}=\mathcal{A}^p\frac{\partial}{\partial\theta^{\mu_p}}\cdots\frac{\partial}{\partial\theta^{\mu_1}}f(\theta), \tag{A12}$$

which is easily proven by induction. Since $\mathcal{A}$ anticommutes with $\partial/\partial\theta^\mu$, it follows that ($p,q=0,1,2,...$)

$$\mathcal{A}^q\frac{\partial}{\partial\theta^{\mu_1}}\cdots\frac{\partial}{\partial\theta^{\mu_p}}f(\theta)=(-1)^{pq}\frac{\partial}{\partial\theta^{\mu_1}}\cdots\frac{\partial}{\partial\theta^{\mu_p}}\mathcal{A}^qf(\theta). \tag{A13}$$

In particular,

$$\mathcal{A}^p\frac{\partial}{\partial\theta^{\mu_1}}\cdots\frac{\partial}{\partial\theta^{\mu_p}}f(\theta)=(-1)^{p}\frac{\partial}{\partial\theta^{\mu_1}}\cdots\frac{\partial}{\partial\theta^{\mu_p}}\mathcal{A}^pf(\theta). \tag{A14}$$

These identities will be needed in order to establish the equivalence of the Clifford product and the fermionic star product.

Our conventions for the integration are $\int d\theta^\mu=0$ and $\int\theta^\mu\,d\theta^\mu=1$ (μ not summed). We define

$$d^n\theta\equiv d\theta^1 d\theta^2\cdots d\theta^n \tag{A15}$$

so that

$$\int\theta^{\mu_n}\theta^{\mu_{n-1}}\cdots\theta^{\mu_1}d^n\theta=\epsilon^{\mu_1\mu_2\cdots\mu_n} \tag{A16}$$

with $\epsilon^{12\cdots n}=+1$. Using (A11) one can show that

$$\int f(\theta)\frac{\partial}{\partial\theta^\mu}g(\theta)d^n\theta=\int f(\theta)\frac{\overleftarrow{\partial}}{\partial\theta^\mu}g(\theta)d^n\theta \tag{A17}$$

for arbitrary inhomogeneous functions f and g.

In our conventions, the delta function is defined to satisfy

$$\int f(\theta)\delta(\theta-\xi)d^n\theta=f(\xi). \tag{A18}$$

(Note the order of the factors.) It is given by

$$\delta(\theta-\xi)=(\theta^n-\xi^n)(\theta^{n-1}-\xi^{n-1})\cdots(\theta^1-\xi^1) \tag{A19}$$

or by the Fourier representation

$$\delta(\theta-\xi)=(-1)^{n(n-1)/2}\int e^{(\theta^\mu-\xi^\mu)\rho_\mu}d^n\rho. \tag{A20}$$

Here $\{\xi^1,\ldots,\xi^n\}$ and $\{\rho_1,\ldots,\rho_n\}$ are two additional sets of real Grassmann variables which anticommute among themselves and with the θ's. (Indices are raised and lowered with the flat metric $g_{\mu\nu}=\delta_{\mu\nu}$.) Depending on the value of n, δ is either Grassmann real or purely imaginary:

$$\overline{\delta(\theta)}=(-1)^{n(n-1)/2}\delta(\theta). \tag{A21}$$

The Fourier transform $\tilde{f}$ is defined according to

$$\tilde{f}(\rho)=\epsilon_n^{-1}\int e^{i\theta^\mu\rho_\mu}f(\theta)d^n\theta \tag{A22}$$

with (all formulas given in this appendix are valid for both n even and n odd)

$$\epsilon_n\equiv\begin{cases}1: & \text{for } n \text{ even}\\ -i: & \text{for } n \text{ odd.}\end{cases} \tag{A23}$$

The advantage of our conventions is that they give rise to a simple formula for $\tilde{f}$ in terms of multiple derivatives of the δ function which is free from explicit sign factors and powers of i. One obtains

$$\tilde{f}(\rho)=f\left(i\frac{\partial}{\partial\rho}\right)\delta(\rho) \tag{A24}$$

because with (A20)

$$\tilde{f}(\rho)=\epsilon_n^{-1}\int f\left(i\frac{\partial}{\partial\rho}\right)e^{i\theta^\mu\rho_\mu}d^n\theta=f\left(i\frac{\partial}{\partial\rho}\right)\epsilon_n^{-1}(-1)^{n(n-1)/2}\delta(-i\rho)=f\left(i\frac{\partial}{\partial\rho}\right)\delta(\rho). \tag{A25}$$

In particular,

$$f(\theta)=1\Rightarrow\tilde{f}(\rho)=\delta(\rho). \tag{A26}$$

The inverse transformation reads

$$f(\theta)=\int e^{-i\theta^\mu\rho_\mu}\tilde{f}(\rho)d^n\rho. \tag{A27}$$

The Grassmann Fourier transformation has the involutive property

$$\tilde{\tilde{f}}(\theta)=\epsilon_n^{-1}f(\theta), \tag{A28}$$

i.e., for n even it is an exact involution. Derivative and multiplication operators are conjugate in the sense that

$$[\widetilde{\theta^\mu f(\theta)}](\rho)=i\frac{\partial}{\partial\rho_\mu}\tilde{f}(\rho), \tag{A29}$$

$$\left[i \widetilde{\frac{\partial}{\partial \theta^\mu} f(\theta)} \right](\rho) = \rho_\mu \tilde{f}(\rho). \tag{A30}$$

Using either (A22) or (A24) one can work out the Fourier transform of a product of θ's. The result is

$$[\theta^{\mu_1} \widetilde{\theta^{\mu_2} \cdots} \theta^{\mu_p}](\rho) = \frac{C_{np}}{(n-p)!} \epsilon^{\mu_1 \cdots \mu_p \nu_1 \cdots \nu_{n-p}} \rho_{\nu_1} \rho_{\nu_2} \cdots \rho_{\nu_{n-p}} \tag{A31}$$

with the constants

$$C_{np} \equiv i^p (-1)^{p(p-1)/2} (-1)^{n(n-1)/2}. \tag{A32}$$

Identifying $dx^\mu \equiv \theta^\mu$, the exterior algebra $\bigwedge(T_x^* \mathbf{R}^n)$ endowed with the inner product coming from $g_{\mu\nu} = \delta_{\mu\nu}$ provides a special example of a Grassmann algebra. In this context we are familiar with the notion of a Hodge star operator which maps p forms onto $(n-p)$ forms. In the case at hand we introduce a corresponding linear map $*: f(\theta) \mapsto (*f)(\theta)$ which generalizes this concept. On the basis elements, the Hodge operator acts according to

$$*(\theta^{\mu_1} \cdots \theta^{\mu_p}) = \frac{1}{(n-p)!} \epsilon^{\mu_1 \cdots \mu_p}{}_{\nu_1 \cdots \nu_{n-p}} \theta^{\nu_1} \cdots \theta^{\nu_{n-p}} \tag{A33}$$

and it is extended to arbitrary functions $f(\theta)$ by linearity. Writing $(*f)(\theta) = \Sigma_{p=0}^n (1/p!) \times [*f]^{(p)}_{\mu_1 \cdots \mu_p} \theta^{\mu_1} \cdots \theta^{\mu_p}$ one finds for the components

$$[*f]^{(n-p)}_{\mu_1 \cdots \mu_{n-p}} = \frac{1}{p!} f^{(p)}_{\nu_1 \cdots \nu_p} \epsilon^{\nu_1 \cdots \nu_p}{}_{\mu_1 \cdots \mu_{n-p}}. \tag{A34}$$

(Note that in parts of the literature a different definition of $*$ is used which amounts to interchanging the transformation laws of the basis vectors and the components, respectively.) Acting twice with $*$ on a homogeneous function of degree p the result is

$$** f^{(p)} = (-1)^{p(n-p)} f^{(p)}. \tag{A35}$$

Because of the p-dependent sign factor on the right-hand side of (A35) the star operator does not give rise to an involution on the space of all (i.e., inhomogeneous) functions. This motivates us to introduce the operator

$$\star \equiv * \mathcal{B}, \tag{A36}$$

which we shall refer to as the ''modified Hodge operator.'' For homogeneous functions,

$$\star f^{(p)} = (-1)^{p(p-1)/2} * f^{(p)}, \tag{A37}$$

which implies

$$\star\star f^{(p)} = (-1)^{n(n-1)/2} f^{(p)} \tag{A38}$$

with a sign factor independent of p. Hence, for any inhomogeneous function f,

$$\star\star f = (-1)^{n(n-1)/2} f. \tag{A39}$$

For $n=4$, say, $\star\star = 1$ so that $\star$ is an exact involution.

The (modified) Hodge operator is closely related to the Grassmann Fourier transformation. Comparing (A31) to (A33) shows that for homogeneous functions

$$*f^{(p)}=(-i)^{p}(-1)^{p(p-1)/2}(-1)^{n(n-1)/2}\widetilde{f^{(p)}},\tag{A40}$$

$$\star f^{(p)}=(-i)^{p}(-1)^{n(n-1)/2}\widetilde{f^{(p)}}.\tag{A41}$$

Using (A24) we may express the Fourier transform by the derivative of a δ function:

$$*f^{(p)}(\theta)=(-1)^{p(p-1)/2}(-1)^{n(n-1)/2}f^{(p)}\left(\frac{\partial}{\partial\theta}\right)\delta(\theta),\tag{A42}$$

$$\star f^{(p)}(\theta)=(-1)^{n(n-1)/2}f^{(p)}\left(\frac{\partial}{\partial\theta}\right)\delta(\theta).\tag{A43}$$

Note that the sign factor on the right-hand side of (A43) is independent of p. Hence it follows that for arbitrary inhomogeneous functions

$$\star f(\theta)=(-1)^{n(n-1)/2}f\left(\frac{\partial}{\partial\theta}\right)\delta(\theta).\tag{A44}$$

This is an interesting representation of the Hodge operator, because in contrast to (A33), Eq. (A44) continues to be meaningful if we regard θ^{μ} as a *commuting* variable. This fact will become important in the construction of the metaplectic DK fields.

APPENDIX B: REPRESENTATIONS OF THE FERMIONIC WEHL STAR PRODUCT

In this appendix we derive several important representations of the fermionic star product, Eqs. (2.15), (2.16), and (2.17), from the integral representation (2.14).

We start by shifting θ_1 and θ_2 in Eq. (2.14):

$$(f_1\circ f_2)(\theta)=\epsilon_n\left(\frac{\hbar}{2i}\right)^n\int\exp\left(\frac{2}{\hbar}\theta_1\theta_2\right)f_1(\theta_1+\theta)f_2(\theta_2+\theta)d^n\theta_1 d^n\theta_2.\tag{B1}$$

Next we Taylor-expand f_1 and f_2 with respect to θ_1 and θ_2. Because the exponential produces only terms with equal numbers of θ_1's and θ_2's, only those terms in the product of the two Taylor series survive the integration which contain equal numbers as well:

$$\begin{aligned}(f_1\circ f_2)(\theta)=\epsilon_n\left(\frac{\hbar}{2i}\right)^n\sum_{p=0}^{n}\left(\frac{1}{p!}\right)^2\int\exp\left(\frac{2}{\hbar}\theta_1\theta_2\right)[\theta_1^{\mu_1}\cdots\theta_1^{\mu_p}\tilde{\partial}_{\mu_1}\cdots\tilde{\partial}_{\mu_p}f_1(\theta)]\\ \times[\theta_2^{\nu_1}\cdots\theta_2^{\nu_p}\tilde{\partial}_{\nu_1}\cdots\tilde{\partial}_{\nu_p}f_2(\theta)]d^n\theta_1 d^n\theta_2.\end{aligned}\tag{B2}$$

Here $\tilde{\partial}_\mu\equiv\partial/\partial\theta^\mu$. Because

$$(\theta^1\theta^2\cdots\theta^p)(\xi^1\xi^2\cdots\xi^p)=(-1)^p(\xi^1\xi^2\cdots\xi^p)(\theta^1\theta^2\cdots\theta^p)\tag{B3}$$

for two arbitrary sets of mutually anticommuting Grassmann-odd objects, we may use Eqs. (A7) and (A14) to write

$$\begin{aligned}\theta_1^{\mu_1}\cdots\theta_1^{\mu_p}\tilde{\partial}_{\mu_1}\cdots\tilde{\partial}_{\mu_p}f_1(\theta)&=(-1)^p\tilde{\partial}_{\mu_1}\cdots\tilde{\partial}_{\mu_p}\theta_1^{\mu_1}\cdots\theta_1^{\mu_p}f_1(\theta)\\ &=(-1)^p\tilde{\partial}_{\mu_1}\cdots\tilde{\partial}_{\mu_p}(\mathcal{A}^pf_1)(\theta)\theta_1^{\mu_1}\cdots\theta_1^{\mu_p}\\ &=[\mathcal{A}^p\tilde{\partial}_{\mu_1}\cdots\tilde{\partial}_{\mu_p}f_1(\theta)]\theta_1^{\mu_1}\cdots\theta_1^{\mu_p}.\end{aligned}\tag{B4}$$

Thus we arrive at

$$(f_1\circ f_2)(\theta)=\epsilon_n\left(\frac{\hbar}{2i}\right)^n\sum_{p=0}^{n}\left(\frac{1}{p!}\right)^2[\mathcal{A}^p\tilde{\partial}_{\mu_1}\cdots\tilde{\partial}_{\mu_p}f_1(\theta)]I^{\mu_1\cdots\mu_p}_{\nu_1\cdots\nu_p}[\tilde{\partial}^{\nu_1}\cdots\tilde{\partial}^{\nu_p}f_2(\theta)] \tag{B5}$$

with

$$I^{\mu_1\cdots\mu_p}_{\nu_1\cdots\nu_p}=\int\exp\left(\frac{2}{\hbar}\theta_1\theta_2\right)\theta_1^{\mu_1}\cdots\theta_1^{\mu_p}\theta_{2\nu_1}\cdots\theta_{2\nu_p}d^n\theta_1 d^n\theta_2. \tag{B6}$$

Upon expanding the exponential, only the term of order $n-p$ can contribute to the integral:

$$I^{\mu_1\cdots\mu_p}_{\nu_1\cdots\nu_p}=\frac{1}{(n-p)!}\left(\frac{2}{\hbar}\right)^{n-p}J^{\mu_1\cdots\mu_p}_{\nu_1\cdots\nu_p}, \tag{B7}$$

$$J^{\mu_1\cdots\mu_p}_{\nu_1\cdots\nu_p}\equiv\int(\theta_1^\alpha\theta_{2\alpha})^{n-p}\theta_1^{\mu_1}\cdots\theta_1^{\mu_p}\theta_{2\nu_1}\cdots\theta_{2\nu_p}d^n\theta_1 d^n\theta_2. \tag{B8}$$

For symmetry reasons the tensor J must have the structure

$$J^{\mu_1\cdots\mu_p}_{\nu_1\cdots\nu_p}=\lambda(n,p)\delta^{[\mu_1}_{\nu_1}\cdots\delta^{\mu_p]}_{\nu_p}=\frac{\lambda(n,p)}{p!}\sum_{\pi\in\mathcal{S}_p}\mathrm{sign}(\pi)\delta^{\pi(\mu_1)}_{\nu_1}\delta^{\pi(\mu_2)}_{\nu_2}\cdots\delta^{\pi(\mu_p)}_{\nu_p}, \tag{B9}$$

where $\mathcal{S}_p$ is the symmetric group of p objects. The constants $\lambda(n,p)$ are most easily determined by choosing the special index combination $J^{12\cdots p}_{12\cdots p}$ for which only the identical permutation contributes in (B9). Furthermore, the summation over α in (B8) is restricted to $\alpha>p$ then:

$$\begin{aligned}\lambda(n,p)&=p!\int[\theta_1^{p+1}\theta_2^{p+1}+\cdots+\theta_1^n\theta_2^n]^{n-p}(\theta_1^1\theta_1^2\cdots\theta_1^p)(\theta_2^1\theta_2^2\cdots\theta_2^p)d^n\theta_1 d^n\theta_2\\&=p!(n-p)!\int(\theta_1^1\theta_1^2\cdots\theta_1^p)[\theta_1^{p+1}\theta_2^{p+1}\theta_1^{p+2}\theta_2^{p+2}\cdots\theta_1^n\theta_2^n](\theta_2^1\theta_2^2\cdots\theta_2^p)d^n\theta_1 d^n\theta_2.\end{aligned} \tag{B10}$$

Commuting the θ's next to the corresponding $d\theta$'s produces various sign factors so that finally

$$\lambda(n,p)=(-1)^n(-1)^{n(n-1)/2}(-1)^{p(p-1)/2}p!(n-p)! \tag{B11}$$

If we note that $\epsilon_n i^n(-1)^{n(n-1)/2}=1$ both for n even and n odd, we see that

$$\epsilon_n\left(\frac{\hbar}{2i}\right)^n\frac{1}{p!}I^{\mu_1\cdots\mu_p}_{\nu_1\cdots\nu_p}=(-1)^{p(p-1)/2}\left(\frac{\hbar}{2}\right)^p\delta^{[\mu_1}_{\nu_1}\cdots\delta^{\mu_p]}_{\nu_p}. \tag{B12}$$

Inserting (B12) into (B5) we obtain precisely the final result given in Eq. (2.15) of the main text.

The representation (2.16) follows from (2.15) by using (A12) in order to convert the left derivatives which act on f_1 to right derivatives. One also needs (A8) to switch from the index sequence $(\mu_1,\mu_2,\ldots,\mu_p)$ to $(\mu_p,\mu_{p-1},\ldots,\mu_1)$.

The last representation, Eq. (2.17), follows from (2.16) by commuting left and right derivatives with the same index next to each other. No sign factor is picked up during this reshuffling.

[1] E. Kähler, Rend. Math. Ser. V **21**, 425 (1962).
[2] T. Banks, Y. Dothan, and D. Horn, Phys. Lett. B **117**, 413 (1982).
[3] P. Becher and H. Joos, Z. Phys. C **15**, 343 (1982).
[4] H. Joos, Helv. Phys. Acta **63**, 670 (1990); Nucl. Phys. B (Proc. Suppl.) **17**, 704 (1990); H. Joos and S. I. Azakov, Helv. Phys. Acta **67**, 723 (1994).
[5] H. Aratyn and A. H. Zimerman, Phys. Lett. B **137**, 392 (1984).
[6] I. M. Benn and R. W. Tucker, Commun. Math. Phys. **89**, 341 (1983).

[7] C. A. Linhares, J. A. Mignaco, and M. A. Rego Monteiro, Lett. Math. Phys. **10**, 79 (1985).
[8] M. Shimono, Prog. Theor. Phys. **84**, 331 (1990).
[9] H. Dilger and H. Joos, Nucl. Phys. B (Proc. Suppl.) **34**, 195 (1994).
[10] M. Atiyah, *Vector Fields on Manifolds* (Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen, Heft, 1970), p. 200.
[11] W. Graf, Ann. Inst. Henri Poincaré, Sect. A **29**, 85 (1978).
[12] I. Montvay and G. Münster, *Quantum Fields on a Lattice* (Cambridge University Press, Cambridge, 1994).
[13] J. Kogut and L. Susskind, Phys. Rev. D **11**, 395 (1975); L. Susskind, *ibid.* **16**, 3031 (1977).
[14] B. Kostant, Symposia Mathematica **14**, 139 (1974).
[15] N. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1992).
[16] R. G. Littlejohn, Phys. Rep. **138**, 193 (1986).
[17] E. Gozzi and M. Reuter, Nucl. Phys. B **325**, 356 (1989).
[18] M. B. Green and C. M. Hull, Phys. Lett. B **225**, 57 (1989); **229**, 215 (1989); I. Bars and R. Kallosh, *ibid.* **233**, 117 (1989).
[19] H. García-Compeán, Preprint IASSNS-HEP-98/36; hep-th/9804188.
[20] D. J. Castano, Phys. Lett. B **269**, 345 (1991).
[21] M. Reuter, Int. J. Mod. Phys. A **13**, 3835 (1998); hep-th/9804036.
[22] E. Gozzi and M. Reuter, J. Mod. Phys. A: Math. Gen. **26**, 6319 (1993).
[23] M. Reuter, Int. J. Mod. Phys. A **10**, 65 (1995).
[24] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, Ann. Phys. (N.Y.) **111**, 64 (1978); **111**, 111 (1978).
[25] F. A. Berezin, Sov. Phys. Usp. **23**, 763 (1980).
[26] F. A. Berezin and M. S. Marinov, Ann. Phys. (N.Y.) **104**, 336 (1977).
[27] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, 1992).
[28] E. Gozzi and M. Reuter, Int. J. Mod. Phys. A **9**, 5801 (1994).
[29] B. Fedosov, J. Diff. Geom. **40**, 213 (1994).
[30] B. Fedosov, *Deformation Quantization and Index Theory* (Akademie, Berlin, 1996).
[31] E. S. Fradkin and V. Y. Linetsky, Nucl. Phys. B **431**, 569 (1994); **444**, 577 (1995).
[32] A. Grossmann, Commun. Math. Phys. **48**, 191 (1976).
[33] A. Royer, Phys. Rev. A **15**, 449 (1977).
[34] N. Berline, E. Getzler, and M. Vergne, *Heat Kernels and Dirac Operators* (Springer, New York, 1992).
[35] R. A. Bertlmann, *Anomalies in Quantum Field Theory* (Oxford University Press, New York, 1996).
[36] M. Kontsevich, q-alg/9709040.
[37] C. Emmrich and A. Weinstein, hep-th/9311094.
[38] C. Emmrich and H. Römer, Acta Phys. Pol. B **27**, 2393 (1996); hep-th/9701111.
[39] M. Bordemann and S. Waldmann, q-alg/9605012; q-alg/9605038; M. Bordemann, N. Neumaier, and S. Waldmann, Commun. Math. Phys. **198**, 363 (1998); q-alg/9711016.
[40] I. Gelfand, V. Retakh, and M. Shubin, dg-ga/9707024.
[41] K. Habermann, Commun. Math. Phys. **184**, 629 (1997), and references therein.
[42] J. Krüger, Diploma thesis, University of Hamburg, 1985.
[43] T. Wyrozumski, Phys. Rev. D **42**, 1152 (1990) E. Prugovecki, Class. Quantum Grav. **13**, 1007 (1996), and references therein; W. Drechsler and P. Tuckey, *ibid.* **13**, 611 (1996); D. Graudenz, gr-qc/9412013; hep-th/9604180; B. Iliev, quant-ph/9804062, and references therein.
[44] G. Mack and V. Schomerus, in *Infinite Dimensional Geometry, Noncommutative Geometry, Operator Algebras, Fundamental Interactions*, edited by R. Coquereaux, M. Dubois-Violette, and P. Flad (World Scientific, Singapore, 1995); G. Mack, in *Salamfestschrift*, edited by A. Ali, J. Ellis, and S. Randjbar-Daemi (World Scientific, Singapore, 1994).
[45] E. Gozzi, Phys. Lett. A **202**, 330 (1995); Nucl. Phys. (Proc. Suppl.) **57**, 223 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

World-line Green functions with momentum and source conservations

Haru-Tada Sato[a)]
Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, D-69120 Heidelberg, Germany

(Received 11 May 1999; accepted for publication 4 August 1999)

Based on the generating functional method with an external source function, a useful constraint on the source function is proposed for analyzing the one- and two-loop world-line Green functions. The constraint plays the same role as the momentum conservation law of a certain nontrivial form, and transforms ambiguous Green functions into the uniquely defined Green functions. We also argue reparametrizations of the Green functions defined on differently parameterized world-line diagrams. © *1999 American Institute of Physics.*
[S0022-2488(99)01511-X]

I. INTRODUCTION

String theory organizes the scattering amplitudes in a very compact form (in the infinite string tension limit), and this fact makes the investigation of field theory scattering amplitudes very nontrivial and potentially useful.[1–5] In this spirit, multiloop scattering amplitudes have also been studied both from the string theory viewpoint[6–9] and the field theory based on the first quantization formalism (world-line formalism).[10–14] The general structure of field theory amplitudes (with N external momenta $p_1, p_2, \ldots, p_N$) at the one-loop order is described as[1,2]

$$\Gamma_N = \int_0^\infty \frac{dT}{T}\left(\frac{1}{4\pi T}\right)^{D/2}\left(\prod_{n=1}^{N}\int_0^T d\tau_n\right)K(\tau_1,\tau_2,\ldots,\tau_n;T)\exp\left[\frac{1}{2}\sum_{j,k=1}^{N} p_j\cdot p_k G_B(\tau_j,\tau_k)\right], \tag{1.1}$$

where K is a certain function which depends on the detail of theory in question (it can be determined systematically). The exponent including the function G_B is sometimes called the generating kinematical factor, and is a theory independent object. G_B is the (world-line) Green function between two points on a loop of length T. One can similarly write down the generalized formulas for certain multiloop cases[9,13] with using multiloop Green functions.[8,11] In this sense, determinations of multiloop Green functions are important factors in the world-line formalism.

On the whole, there are three kinds of computation methods to obtain the world-line Green functions: (i) invert kinematic terms (solve differential equations); (ii) define as Gaussian determinants or compute two-point correlators in the path integral method; (iii) similar as the second method, but reducing the path integrals into ordinary integrals (thus no appearance of determinants). Depending on the computation methods, different forms for Green functions have been obtained (we show several examples in the main text), and a natural question raised from this observation is whether or not they all describe the same amplitudes when applied to the amplitude formulas such as (1.1). The answer should be yes if the formalism is self-consistent. We call this problem the ambiguity problem of Green functions, and there must be a reason why we are allowed to have a variety of Green functions for a unique amplitude. Once we find a prescription for this problem, we shall be able to ignore the existence of various ambiguous Green functions, and rather understand them all as a family of Green functions.

[a)]Electronic mail: sato@thphys.uni-heidelberg.de

In this paper, we focus on the ambiguity problem of the multiloop Green functions. This is an important problem from the above viewpoint: All of ambiguous Green functions should be reduced to the uniquely defined ones under the constraint of a vanishing identity, without changing the value of a kinematical factor. The problem is trivial in the one-loop case, and summarized as follows. In the original definition

$$\frac{1}{2}\partial^2 G_B(\tau)=\delta(\tau)-\frac{1}{T} \tag{1.2}$$

with imposing rotational invariance and periodicity, G_B is uniquely determined as the rotational symmetric form

$$G_B(\tau_1,\tau_2)=G_B(\tau_1-\tau_2)=|\tau_1-\tau_2|-\frac{(\tau_1-\tau_2)^2}{T}. \tag{1.3}$$

However, we do not necessarily use this functional form as concerns the kinematical factor itself, which actually does not change if we add a polynomial in τ_j to the $G_B(\tau_j,\tau_k)$ in (1.1) because of the conservation law $\Sigma_k p_k=0$. This ambiguity is very easy to verify in the one-loop case, where the defining equation of G_B is simple and its rotational invariance should be clear. On the other hand, in multiloop cases, the situation is much too complicated to identify the ambiguity which can be canceled by a certain condition such as conservation law, because the definitions and calculations of multiloop Green functions are in general complicated—in addition that the rotational invariance is unclear in certain cases.

As the simplest nontrivial example, we discuss the two-loop Green functions.[11,12] We refer to the Green functions containing the ambiguity as the *wide sense* Green functions. As suggested above, the value of a kinematical factor should be invariant for any set of wide sense Green functions. Various wide sense Green functions can be obtained depending on how to define and evaluate, and we shall verify that all of them can be identified with each other in the sense of keeping the kinematical factor invariant by way of examples. To this end, we obviously need a constraint such as the total momentum conservation law. However, in the generic multiloop cases, a useful form of the conservation law is in practice not a simple summation (along the single loop as mentioned in the one-loop case), because of the presence of additional internal lines. We generalize the conservation law into a more suitable form to our purpose. In addition to the momentum conservation law, we also present a continuous analog of the conservation law; that is a constraint on the integrals of external source functions along the two-loop world-line vacuum diagram. This continuous version is very simple and useful to apply practical computations, and is nontrivial since such a constraint does not exist in the source term of the usual formulation of field theory.

In Sec. II, for notational conveniences, we briefly review the two-loop kinematical factor and the Green functions in ϕ^3 theory. In Sec. III, we present a useful two-loop momentum conservation formula, and demonstrate how to apply the formula to the identification of different wide sense Green functions. In Sec. IV, employing the generating functional method with external source functions,[14] we consider another derivation of the Green functions. In this case, we show that there also exists a similar constraint formula on the source functions, and verify that it plays the same role as the momentum conservation method of Sec. III. In Sec. V, we further confirm the validity and usefulness of the source constraint in more specific cases (one-loop QED). Section VI is a short note on the previous work,[12] concerning new reparametrization transformations of the two-loop Green functions. Section VII contains conclusions.

II. NOTATIONS

For the purpose of setting our notations, we briefly review the world-line Green functions and the master amplitude formulas corresponding to Eq. (1.1) in the two-loop ϕ^3 theory.[9,11,13] The (two-loop) master formula is a fundamental quantity which contains all necessary Feynman diagrams belonging to a certain class of diagrams. The classes are labeled by two or three integers (N',N_3) or (N_1,N_2,N_3), and amplitudes are certain combinations of these classes.[13] We call the first labeling the loop type, and the latter the symmetric type. The general form of the master formula is as follows:

$$\Gamma_M^{\text{2-loop}}=\frac{1}{12}(-g)^{N+2}\int dM(4\pi)^{-D}\Delta^{D/2}\exp[E_G]. \tag{2.1}$$

For the loop type parameterization $(N=N'+N_3)$, the integration measure dM is

$$dM=\frac{dT}{T}dT_3 d\tau_\alpha d\tau_\beta \prod_{n=1}^{N'} d\tau_n \prod_{l=1}^{N_3} d\tau_l^{(3)}, \tag{2.2}$$

and Δ is the determinant factor

$$\Delta=(TT_3+TG_B(\tau_\alpha,\tau_\beta))^{-1}. \tag{2.3}$$

The exponential part E_G, the generating kinematical factor, takes the following bilinear form in N external momenta $(p_j, p_k^{(3)};\ j=1,...,N';\ k=1,...,N_3)$:

$$E_G=\frac{1}{2}\sum_{j,k=1}^{N'} p_j p_k G_{00}^{(1)}(\tau_j,\tau_k)+\frac{1}{2}\sum_{j,k=1}^{N_3} p_j^{(3)} p_k^{(3)} G_{33}^{(1)}(\tau_j^{(3)},\tau_k^{(3)})+\sum_{j=1}^{N'}\sum_{k=1}^{N_3} p_j p_k^{(3)} G_{03}^{(1)}(\tau_j,\tau_k^{(3)}), \tag{2.4}$$

where the bilinear momenta should be understood as Lorentz contracted forms (hereafter as well). The explicit forms of these Green functions are[12]

$$G_{00}^{(1)}(\tau,\tau')=G_B(\tau,\tau')-\frac{1}{4}\frac{(G_B(\tau,\tau_\alpha)-G_B(\tau,\tau_\beta)-G_B(\tau',\tau_\alpha)+G_B(\tau',\tau_\beta))^2}{T_3+G_B(\tau_\alpha,\tau_\beta)}, \tag{2.5}$$

$$G_{33}^{(1)}(z_1,z_2)=G_{33}^{(1)}(z_1-z_2)==|z_1-z_2|-\frac{(z_1-z_2)^2}{T_3+G_B(\tau_\alpha,\tau_\beta)}, \tag{2.6}$$

$$G_{03}^{(1)}(\tau,z)=\begin{cases} G_{00}^{(1)}(\tau,\tau_\alpha)+\dfrac{1}{T_3+G_B(\tau_\alpha,\tau_\beta)}(T_3 z-z^2+z[G_B(\tau,\tau_\beta)-G_B(\tau,\tau_\alpha)]) & \text{for } \tau_\beta<\tau_\alpha \\ G_{00}^{(1)}(\tau,\tau_\beta)+\dfrac{1}{T_3+G_B(\tau_\alpha,\tau_\beta)}(T_3 z-z^2+z[G_B(\tau,\tau_\alpha)-G_B(\tau,\tau_\beta)]) & \text{for } \tau_\alpha<\tau_\beta. \end{cases} \tag{2.7}$$

The τ parameters $\{\tau_\alpha,\tau_\beta,\tau_n|n=1,...,N'\}$ run from zero to T, which stands for the length of a loop (fundamental loop), and $\tau_n^{(3)}$, $n=1,...,N_3$ run from zero to T_3, the length of the internal line (the rest part of the vacuum diagram). T and T_3 are to be integrated from zero to infinity. In Ref. 9, we pointed out that one may fix and eliminate one of the parameters $\{\tau_\alpha,\tau_\beta,\tau_n|n=1,...,N'\}$ because of the rotational symmetry of the fundamental loop. This means that we can set one of these parameters to be zero which corresponds to the origin of world-line coordinate along the fundamental loop. Obviously, $G_{00}^{(1)}$ is invariant under this rotation, and does not receive any serious change, however $G_{03}^{(1)}$ does not even possess any translational symmetry such as seen in $G_{33}^{(1)}$. Hence the explicit form of $G_{03}^{(1)}$ depends on which parameter will be set zero. For example, if we choose τ_β as such origin, $G_{03}^{(1)}$ should follow the form for $\tau_\beta<\tau_\alpha$. Similarly, if τ_α, then take for $\tau_\alpha<\tau_\beta$. There is also a different Green function[11] from Eq. (2.5). However, both coincide under the same momentum conservation constraint (for $N_3=0$) as the one-loop type.

Using the transformation obtained in Ref. 12, we can transform the above quantities to the other version (symmetric parameterization). It is done by dividing the fundamental loop into two pieces $T=T_1+T_2$ with $N'=N_1+N_2$ and $\{\tau_n\}\rightarrow\{\tau_n^{(1)},\tau_n^{(2)}\}$. In this case, we have[11]

$$dM=dT_1dT_2dT_3\prod_{i=1}^{3}\prod_{n=1}^{N_i}d\tau_n^{(i)}, \tag{2.8}$$

$$\Delta=(T_1T_2+T_2T_3+T_3T_1)^{-1}, \tag{2.9}$$

and

$$E_G=\frac{1}{2}\sum_{a=1}^{3}\sum_{j,k}^{N_a}p_j^{(a)}G_{aa}^{\text{sym}}(\tau_j^{(a)},\tau_k^{(a)})+\sum_{a=1}^{3}\sum_{j}^{N_a}\sum_{k}^{N_{a+1}}p_j^{(a)}p_k^{(a+1)}G_{aa+1}^{\text{sym}}(\tau_j^{(a)},\tau_k^{(a+1)}), \tag{2.10}$$

where we set $\tau^{(4)}=\tau^{(1)}$ and $N_4=N_1$, etc. in accord with the cyclic expression. The Green functions are[11]

$$G_{aa}^{\text{sym}}(\tau,\tau')=G_{aa}^{\text{sym}}(\tau-\tau')=|\tau-\tau'|-\frac{T_{a+1}+T_{a+2}}{T_1T_2+T_2T_3+T_3T_1}(\tau-\tau')^2, \tag{2.11}$$

$$G_{aa+1}^{\text{sym}}(\tau,\tau')=\tau+\tau'-\frac{\tau^2T_{a+1}+\tau'^2T_a+(\tau+\tau')^2T_{a+2}}{T_1T_2+T_2T_3+T_3T_1}. \tag{2.12}$$

All the formulas in this section are reproduced from string theory,[8,9] and in this sense, we refer to these Green functions (2.5)–(2.7), (2.11), and (2.12) as the standard forms.

III. MOMENTUM CONSERVATION CONSTRAINT

In this section, we encounter the (wide sense) Green functions of different forms, depending on calculation methods (in the symmetric parameterization). However, the value of E_G should be shown to be invariant under the constraint of total momentum conservation. In the one-loop case, as mentioned in the introduction, the constraint is expressed by the identity

$$\sum_{j=1}^{N}\sum_{k=1}^{N}p_j\cdot p_k\tau_k^m=0. \tag{3.1}$$

The single summation over all momenta p_j is nothing but the summation over the fundamental loop. However, the same structure cannot be seen in (2.4) or (2.10) for the $N_3\neq 0$ case. Hence, we shall derive a suitable two-loop generalization of this identity, and explain how it works. To illustrate the idea clearly, we need a couple of examples of different Green functions in the first place.

As explained in Ref. 13, Eq. (2.1) is obtained from the path integral

$$\Gamma_M^{\text{2-loop}}=\frac{(-g)^{N+2}}{2\cdot 3!}\int d^Dx_1d^Dx_2\prod_{a=1}^{3}\int_0^\infty dT_ae^{-m^2T_a}\int_{y_a(T_a)=x_1}^{y_a(0)=x_2}\mathcal{D}y_a(\tau)$$
$$\times\exp\left[-\int_0^{T_a}\frac{1}{4}\dot{y}_a^2d\tau^{(a)}\right]\prod_{n=1}^{N_a}\int_0^{T_a}d\tau_n^{(a)}e^{ip_n^{(a)}y(\tau_n^{(a)})} \tag{3.2}$$

by using the mode expansion

$$y_a(\tau)=x_1+\frac{\tau}{T_a}(x_2-x_1)+\sum_{m=1}^{\infty}y_m\sin\left(\frac{m\pi\tau}{T_a}\right). \tag{3.3}$$

A straightforward computation in this case shows that the E_G part is composed of the following Green functions instead of $G_{ab}^{\rm sym}$:

$$G_{aa}^M(\tau,\tau')=|\tau-\tau'|-(\tau+\tau')+2\,\frac{\tau\tau'}{T_a}\left(1-\Delta\,\frac{T_1T_2T_3}{T_a}\right), \tag{3.4}$$

$$G_{aa+1}^M(\tau,\tau')=-2\Delta T_1T_2T_3\;\frac{\tau\tau'}{T_aT_{a+1}}. \tag{3.5}$$

Note that the x_1 integration generates the total momentum conservation factor

$$(2\pi)^D\delta\left(\sum_a\sum_n^{N_a}p_n^{(a)}\right). \tag{3.6}$$

A second example is from Ref. 14. The world-line Green function should also be derived as a two-point function in the sense of ordinary field theory,

$$\mathcal{G}_{\mu\nu}(\tau_1^{(a)},\tau_2^{(b)})=\langle x_\mu(\tau_1^{(a)})x_\nu(\tau_2^{(b)})\rangle=\frac{\delta}{\delta J_a^\mu(\tau_1^{(a)})}\,\frac{\delta}{\delta J_c^\nu(\tau_2^{(b)})}\ln Z[J]\bigg|_{J=0}, \tag{3.7}$$

where the generating functional is given by

$$Z[J]\equiv\int d^Dy_1d^Dy_2\left(\prod_{a=1}^3\int_{x_a(0)=y_1}^{x_a(T_a)=y_2}\mathcal{D}x_a\right)\exp\left[-\frac{1}{4}\sum_a\int_0^{T_a}\dot{x}_a^2(\tau)d\tau+\sum_a\int_0^{T_a}J_a^\mu(\tau)x_a^\mu(\tau)d\tau\right]. \tag{3.8}$$

For later convenience, we here write the intermediate expression (putting $w=(y_1+y_2)/2$, $z=y_2-y_1$)

$$Z[J]=(\Pi_{a=1}^3(4\pi T_a)^{-D/2})\exp\left[-\frac{1}{2}\sum_{a=1}^3\int_0^{T_a}\int_0^{T_a}J_\mu^a(\tau)\tilde{G}_{\mu\nu}^{(a)}(\tau,\tau')J_\nu^a(\tau')d\tau d\tau'\right]$$
$$\times\int dzdw\prod_{a=1}^3\;\exp\left[w\int_0^{T_a}J^a(\tau)d\tau-\frac{1}{4}z^\mu A_{\mu\nu}^a z^\nu+z^\nu\int_0^{T_a}J_\mu^aR_{\mu\nu}^a\right] \tag{3.9}$$

as well as the final expression

$$Z[J]=i\,\delta^D\left(\sum_{a=1}^3\int_0^{T_a}J_a^\mu(\tau)d\tau\right)(4\pi)^{(D/2)}\left(\prod_{a=1}^3(4\pi T_a)^{(-D/2)}\right)\det\nolimits_L^{-(1/2)}\left(\sum_a A^a\right)$$
$$\times\exp\left[-\frac{1}{2}\sum_a\int_0^{T_a}\int_0^{T_a}J_\mu^a(\tau)\tilde{G}_{\mu\nu}^{(a)}(\tau,\tau')J_\nu^a(\tau')d\tau d\tau'\right]$$
$$\times\exp\left[\left(\sum_a A^a\right)_{\rho\sigma}^{-1}\left(\sum_a\int_0^{T_a}R_{\rho\mu}^aJ_\mu^a(\tau)d\tau\right)\left(\sum_c\int_0^{T_c}R_{\sigma\nu}^cJ_\nu^c(\tau)d\tau\right)\right], \tag{3.10}$$

where we take

$$A_{\mu\nu}^a=\delta_{\mu\nu}T_a^{-1},\quad R_{\mu\nu}^a=\left(\frac{\tau^{(a)}}{T_a}-\frac{1}{2}\right)\delta_{\mu\nu}, \tag{3.11}$$

and

$$\tilde{G}^{(a)}_{\mu\nu}(\tau_1,\tau_2)=\delta_{\mu\nu}\left(|\tau_1-\tau_2|-(\tau_1+\tau_2)+2\,\frac{\tau_1\tau_2}{T_a}\right). \tag{3.12}$$

A main difference from the first example is the existence of nonconstant source terms, where further τ integrations are formally impossible. This is the reason for having a different form of the Green function. Further decoupling the metric factor ($g_{\mu\nu}=-\delta_{\mu\nu}$),

$$\mathcal{G}_{\mu\nu}(\tau^{(a)},\tau^{(b)})=-g_{\mu\nu}G^J_{ab}(\tau^{(a)},\tau^{(b)}), \tag{3.13}$$

we actually derive the second different form

$$G^J_{ab}(\tau_1^{(a)},\tau_2^{(b)})=\delta_{ab}\left(|\tau_1^{(a)}-\tau_2^{(b)}|-(\tau_1^{(a)}+\tau_2^{(b)})+2\,\frac{\tau_1^{(a)}\tau_2^{(b)}}{T_a}\right)$$
$$-\frac{1}{2}(2\tau_1^{(a)}-T_a)(2\tau_2^{(b)}-T_b)\,\frac{T_1T_2T_3}{T_aT_b}\,\Delta. \tag{3.14}$$

Now, let us derive the two-loop version of the constraint (3.1). It can be derived by combining the trivial identities

$$\left(\sum_{a=1}^{3}\sum_{j=1}^{N_a}p_j^{(a)}\right)\sum_{k=1}^{N_b}p_k^{(b)}(\tau_k^{(b)})^m=0;\quad b=1,2,3 \tag{3.15}$$

with multiplying weight coefficients $C_m^{(b)}$. The result is arranged in the form suitable to E_G,

$$0=\sum_{a=1}^{3}\sum_{j,k}^{N_a}p_j^{(a)}p_k^{(a)}C_m^{(a)}(\tau_j^{(a)})^m+\sum_{a=1}^{3}\sum_{j}^{N_a}\sum_{k}^{N_{a+1}}p_j^{(a)}p_k^{(a+1)}(C_m^{(a)}(\tau_j^{(a)})^m+C_m^{(a+1)}(\tau_k^{(a+1)})^m), \tag{3.16}$$

where m is an arbitrary integer and the mth coefficient $C_m^{(a)}$ may depend only on T_a. One can add an arbitrary number of copies of this identity to E_G with different choices of $C_m^{(a)}$. Consider the E_G where G^M_{ab} is substituted for $G^{\rm sym}$ in (2.10), and add the identity (3.16) to the E_G. Then a new set of Green functions can be read from the modified E_G as

$$G'_{aa}(\tau_1^{(a)},\tau_2^{(a)})=G^M_{aa}(\tau_1^{(a)},\tau_2^{(a)})+2C_m^{(a)}(\tau_1^{(a)})^m+\cdots, \tag{3.17}$$

$$G'_{aa+1}(\tau_1^{(a)},\tau_2^{(a+1)})=G^M_{aa+1}(\tau_1^{(a)},\tau_2^{(a+2)})+C_m^{(a)}(\tau_1^{(a)})^m+C_m^{(a+1)}(\tau_2^{(a+1)})^m+\cdots, \tag{3.18}$$

where $\cdots$ means the additions of further different copies mentioned above. These relations suggest that a variety of Green function's representations can be derived starting from G^M_{ab}. In practice, the three representations listed here ($G^{\rm sym}$,$G^M G^J$) are connected to the following choices of the $C_m^{(a)}$ coefficients. We obtain $G'_{ab}=G^J_{ab}$, if we choose

$$C_0^{(a)}=-\frac{1}{2}\Delta T_1T_2T_3,\quad C_1^{(a)}=\Delta\,\frac{T_1T_2T_3}{T_a},\ \text{ others}=0, \tag{3.19}$$

and we obtain $G'_{ab}=G^{\rm sym}_{ab}$, if we choose

$$C_1^{(a)}=1,\ \ C_2^{(a)}=-\frac{1}{T_a}\left(1-\Delta\,\frac{T_1T_2T_3}{T_a}\right),\ \text{ others}=0. \tag{3.20}$$

In this way, every possible form is related to the standard form by the transformation rules (3.17) and (3.18), or in other words, by the two-loop momentum constraint formula (3.16).

IV. SOURCE CONSTRAINT

In this section, we discuss what identity in the generating functional method should play the same role as the momentum conservation constraint (3.16).

Let us first recall the computation process from (3.9) to (3.10). The w integration in (3.9) gives rise to the similar δ-function divergence as before (cf. Eq. (3.6)) in the sense of the Minkowski formulation, and we then have

$$Z[J]=i\delta\left(\sum_{a=1}^{3}\int_0^{T_a}J^a_\mu(\tau)d\tau\right)\prod_{a=1}^{3}(4\pi T_a)^{-D/2}$$
$$\times\exp\left[-\frac{1}{2}\sum_{a=1}^{3}\int_0^{T_a}\int_0^{T_a}J^a_\mu(\tau)\tilde{G}^{(a)}_{\mu\nu}(\tau,\tau')J^a_\nu(\tau)d\tau d\tau'\right]I[J] \tag{4.1}$$

with

$$I[J]=\int dz\prod_{a=1}^{3}\exp\left[-\frac{1}{4}z^\mu A^a_{\mu\nu}z^\nu+z^\nu\int_0^{T_a}J^a_\mu(\tau)R^a_{\mu\nu}(\tau)d\tau\right]. \tag{4.2}$$

Here, the $R^a_{\mu\nu}$ given in (3.11) is a symmetric tensor, however it is not a general property. Rather, the following reflection antisymmetry is general and important,

$$R^a_{\mu\nu}(\tau)=-R^a_{\mu\nu}(T^a-\tau). \tag{4.3}$$

Suppose that J^a_μ behaves as an even or odd function with respect to the center point $T^a/2$ for the interval $0\leqslant\tau^{(a)}\leqslant T_a$; i.e.,

$$J^a_\mu(\tau^{(a)})=\pm J^a_\mu(T_a-\tau^{(a)}). \tag{4.4}$$

Using these properties, we have

$$z^\nu\int_0^{T_a}J^a_\mu(\tau)R^a_{\mu\nu}(\tau)d\tau=\mp z^\nu\int_0^{T_a}R^a_{\nu\mu}(\tau)J^a_\mu(\tau)d\tau. \tag{4.5}$$

By this formula, we perform the Gaussian integral in (4.2),

$$I[J]=(4\pi)^{(D/2)}\det{}_L^{-(1/2)}\left(\sum_a A^a\right)$$
$$\times\exp\left[\mp\left(\sum_a\int_0^{T_a}J^a_\mu R^a_{\mu\rho}d\tau\right)\left(\sum_a A^a\right)^{-1}_{\rho\sigma}\left(\sum_a\int_0^{T_a}R^a_{\sigma\nu}J^a_\nu d\tau\right)\right]. \tag{4.6}$$

Again using (4.5), we can eliminate the complex signature symbol

$$I[J]=(4\pi)^{(D/2)}\det{}_L^{-(1/2)}\left(\sum_a A^a\right)$$
$$\times\exp\left[\left(\sum_a\int_0^{T_a}R^a_{\rho\mu}J^a_\mu d\tau\right)\left(\sum_a A^a\right)^{-1}_{\rho\sigma}\left(\sum_a\int_0^{T_a}R^a_{\sigma\nu}J^a_\nu d\tau\right)\right]. \tag{4.7}$$

This result (4.7) holds for any linear combination of even and odd J^a_μ functions. It should be noted that the odd source case implies

$$\int_0^{T_a} J_\mu^a(\tau^{(a)})d\tau^{(a)}=0 \quad (a=1,2,3). \tag{4.8}$$

It means a strong sense ''momentum'' conservation which is subjected to only one of three internal lines, and trivially satisfies

$$\sum_{a=1}^{3}\int_0^{T_a} J_\mu^a(\tau^{(a)})d\tau^{(a)}=0. \tag{4.9}$$

Mimicking this property, we in general impose this identity as the total ''momentum'' conservation (sum of three lines), as advocated by the δ-function in (4.1).

Now, let us compare the roles of discrete and continuous constraints (3.16) and (4.9) in an example. We notice the following term in G_{ab}^J (q.v. (3.14)):

$$-\frac{1}{2}(2\tau_1^{(a)}-T_a)(2\tau_2^{(b)}-T_b)\frac{T_1T_2T_3}{T_aT_b}\Delta \tag{4.10}$$

and its corresponding term in the generating functional (4.7),

$$\ln I[J]=\delta_{\mu\nu}T_1T_2T_3\Delta\sum_a\sum_b\int_0^{T_a}\left(\frac{\tau^{(a)}}{T_a}-\frac{1}{2}\right)J_\mu^a d\tau^{(a)}\int_0^{T_b}\left(\frac{\tau^{(b)}}{T_b}-\frac{1}{2}\right)J_\nu^b d\tau^{(b)}+\cdots. \tag{4.11}$$

If we subtract the $C_m^{(a)}$ terms from (4.10) with the choice (3.19), we obtain G^M as understood from (3.17) and (3.18). On the other hand, using the source constraint (4.9), we can remove from (4.11) the linear terms in $\tau^{(a)}$ and $\tau^{(b)}$ as well as the constant term,

$$\ln I[J]\approx\delta_{\mu\nu}T_1T_2T_3\Delta\sum_a\sum_b\int_0^{T_a}\int_0^{T_b}\frac{\tau^{(a)}\tau^{(b)}}{T_aT_b}J_\mu^a(\tau^{(a)})J_\nu^b(\tau^{(a)})d\tau^{(a)}d\tau^{(b)}+\cdots. \tag{4.12}$$

This manipulation leads to the Green function G_{ab}^M as expected; i.e., the removal of the $\tau^{(a)}$ and $\tau^{(b)}$ linear terms corresponds to the subtraction of $C_1^{(a)}$ given in (3.19), and the constant term removal to $C_0^{(a)}$. In this way, the source constraint (4.9) plays the same role as the actual momentum conservation constraint (3.16) on the kinematical factor E_G, thus on the wide sense world-line Green functions. It is worth noting that the constraint (4.9) is simpler than (3.16).

V. EXAMPLES IN QED

The idea of the source constraint gives a family of equivalent world-line Green functions as seen in the previous section. This property is useful for identifying different (wide sense) Green functions obtained by various computations. In this section, we verify its usefulness in more specific cases. The examples discussed here is the one-loop photon scatterings in the scalar and spinor QED cases. First, we discuss the scalar case, and then the spinor case.

The N-point function for a complex boson loop is known to be given[2] by the closed path integral of one-dimensional bosonic field $x^\mu(\tau)$,

$$\Gamma_N(p_1,\ldots,p_N)\equiv\int_0^\infty\frac{dT}{T}\oint\mathcal{D}x\left(\int_0^T\prod_{j=1}^N d\tau_j d\theta_j\overline{\theta_j}\right)\exp\left[\int_0^T\left(-\frac{1}{4}\dot{x}^2+J\cdot x\right)d\tau\right], \tag{5.1}$$

with the following specific source function;

$$J^\mu(\tau)=\sum_{j=1}^N\delta(\tau-\tau_j)\left(\overline{\theta_j}\theta_j\epsilon_j^\mu\frac{\partial}{\partial\tau_j}+ip_j^\mu\right), \tag{5.2}$$

where ϵ_j^μ are photon polarization vectors, and θ_j and $\overline{\theta_j}$ are the Grassmann variables. This source is neither an even function nor an odd one in τ, and we assume the one-loop version of the constraint (4.9) to be

$$\int_0^T J^\mu(\tau)d\tau=0. \tag{5.3}$$

This leads to the constraint similar to the momentum conservation law

$$\sum_{j=1}^{N} J_j^\mu=0; \quad J_j^\mu=\overline{\theta_j}\theta_j\epsilon_j^\mu\frac{\partial}{\partial\tau_j}+ip_j^\mu. \tag{5.4}$$

The second term in J_j exactly corresponds to the momentum conservation law, while the first term does not vanish in the sum at all. In this sense, the present constraint (5.4) assumes a nontrivial conservation law. Let us see how our idea works in the following. We perform the path integral (5.1) as the mode integrations with expanding

$$x^\mu(\tau)=x_0^\mu+\sum_{n>0} x_n^\mu \sin\left(\frac{n\pi\tau}{T}\right) \quad (-\infty\leqslant x_n\leqslant\infty). \tag{5.5}$$

Note that x_0 integration diverges as the δ-function corresponding to the constraint (5.3). (This is similar to the δ-function in Eq. (4.1)). Remember that this kind of divergence is usually removed by hand (so-called zero mode divergence). The resulting expression is then

$$\Gamma_N(p_1,\ldots,p_N)=\int \frac{dT}{T}\left(\frac{1}{4\pi T}\right)^{D/2}\left(\prod_{j=1}^{N} d\tau_j d\theta_j d\overline{\theta_j}\right)\exp\left[\frac{1}{2}g_{\mu\nu}\sum_{j,l=1}^{N} J_j^\mu J_l^\nu \widetilde{G}_B(\tau_j,\tau_l)\right], \tag{5.6}$$

where we formally put $g_{\mu\nu}=-\delta_{\mu\nu}$ and

$$\widetilde{G}_B(\tau_i,\tau_j)=\sum_{m=1}^{\infty}\frac{4T}{m^2\pi^2}\sin\left(\frac{\pi m\tau_i}{T}\right)\sin\left(\frac{\pi m\tau_j}{T}\right) \tag{5.7}$$

$$=|\tau_i-\tau_j|-(\tau_i+\tau_j)+2\frac{\tau_i\tau_j}{T}. \tag{5.8}$$

Here we have used the following formula at the second line of the above,

$$\sum_{m=1}^{\infty}\frac{\cos(mx)}{m^2}=\frac{1}{4}(|x|-\pi)^2-\frac{\pi^2}{12}. \tag{5.9}$$

Under the constraint (5.4), we realize that $\widetilde{G}_B$ in the exponent (the generating kinematical factor) in Eq. (5.6) behaves as the one-loop Green function (1.3) exactly, and we thus rederive the same result[2]

$$\Gamma_N(p_1,\ldots,p_N)=\int \frac{dT}{T}\left(\frac{1}{4\pi T}\right)^{D/2}\left(\prod_{j=1}^{N} d\tau_j d\theta_j d\overline{\theta_j}\right)\exp\left[-\frac{1}{2}\sum_{j,l=1}^{N} J_j\cdot J_l G_B(\tau_j,\tau_l)\right]. \tag{5.10}$$

In this example, it is clear that the source constraint helps us obtain a correct kinematical factor even if a different (wide sense) Green function appears in an intermediate step.

The similar argument applies to the fermion loop case as well. For simplicity, we discuss only the spin part (world-line fermion $\psi^\mu(\tau)$), since the bosonic part is essentially the same as the above case. The world-line fermion part of the N-point amplitude is given by[2]

$$\tilde{\Gamma}_N \equiv \oint \mathcal{D}\psi \left(\prod_{j=1}^{N} \int_0^T d\tau_j d\theta_j d\overline{\theta_j} \right) \exp\left[\int_0^T \left(-\frac{1}{2} \psi^\mu \partial_\tau \psi_\mu + \eta^\mu \psi_\mu \right) d\tau \right] \tag{5.11}$$

with the source function

$$\eta^\mu(\tau) = \sum_{j=1}^{N} \sqrt{2}(\theta_j \epsilon_j^\mu + i\overline{\theta}_j p_j^\mu)\delta(\tau - \tau_j). \tag{5.12}$$

Assuming the source constraint

$$\int_0^T \eta^\mu(\tau) d\tau = 0 \tag{5.13}$$

or equivalently

$$\sum_{j=1}^{N} K_j = 0, \quad K_j^\mu = \sqrt{2}(\theta_j \epsilon_j^\mu + i\overline{\theta}_j p_j^\mu), \tag{5.14}$$

and performing the path integral with the mode expansion

$$\psi^\mu(\tau) = \sum_{r \in \mathbf{Z} + \frac{1}{2}} b_r^\mu \cos\left(\frac{2\pi r \tau}{T} \right), \tag{5.15}$$

we obtain (the detail is in Appendix A),

$$\overline{\Gamma}_N = \prod_{j=1}^{N} \int d\tau_j d\theta_j d\overline{\theta_j} \exp\left[\frac{1}{4} \sum_{j,l=1}^{N} K_j \cdot K_l \tilde{G}_F(\tau_j, \tau_l) \right] \tag{5.16}$$

with

$$\tilde{G}_F(\tau_i, \tau_j) = \mathrm{sign}(\tau_j - \tau_i) + \frac{2}{T}(\tau_i - \tau_j). \tag{5.17}$$

Under the constraint (5.14), $\tilde{G}_F$ in the exponent plays the same role as the standard fermion Green function

$$G_F(\tau_i, \tau_j) = \mathrm{sign}(\tau_j - \tau_i), \tag{5.18}$$

and we reproduce the correct answer[2]

$$\tilde{\Gamma}_N = \prod_{j=1}^{N} \int d\tau_j d\theta_j d\overline{\theta_j} \exp\left[\frac{1}{4} \sum_{j,l=1}^{N} K_j \cdot K_l G_F(\tau_j, \tau_l) \right]. \tag{5.19}$$

VI. REPARAMETRIZATION-TYPE TRANSFORMATION

This section is independent of the previous sections. In this section, we consider reparametrizations and transformations between the standard two-loop Green functions. As mentioned in Sec. II, it is known that G^{sym} and $G^{(1)}$ are connected by a certain transformation.[12] Here, we point out another transformation between them, using periodicities of the Green functions, and also discuss reparametrizations of $G^{(1)}$ through exchanging two of internal lines.

The symmetric Green functions G^{sym} and also $G_{33}^{(1)}$ satisfy the following properties of periodicity:

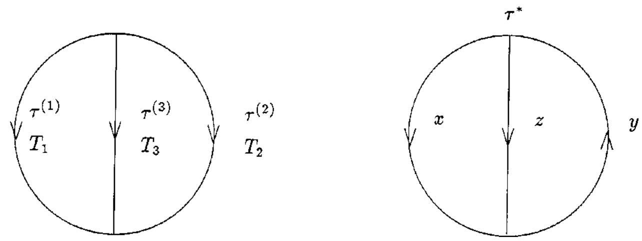

FIG. 1. The directions of τ parameters.

$$G_{ab}^{\rm sym}(T_a-\tau^{(a)},T_b-\tau^{(b)})=G_{ab}^{\rm sym}(\tau^{(a)},\tau^{(b)})\quad (a,b=1,2,3), \tag{6.1}$$

$$G_{aa}^{\rm sym}(\tau-P_a)=G_{aa}^{\rm sym}(\tau),\quad G_{33}^{(1)}(\tau-P_{11})=G_{33}^{(1)}(\tau), \tag{6.2}$$

where

$$P_a=T_a+\frac{T_{a+1}T_{a+2}}{T_{a+1}+T_{a+2}},\quad P_{11}=T_3+G_B(\tau_\alpha,\tau_\beta)=\frac{1}{T\Delta}. \tag{6.3}$$

Putting $P_3=P_{11}$ with identifying $T_1=T(1-u)$ and $T_2=Tu$, we easily find

$$u=\frac{|\tau_\alpha-\tau_\beta|}{T}, \tag{6.4}$$

and the necessary relations for the transformation between $G^{\rm sym}$ and $G^{(1)}$,[12]

$$T_1=T-|\tau_\alpha-\tau_\beta|,\quad T_2=|\tau_\alpha-\tau_\beta|. \tag{6.5}$$

For later convenience, we assign more concrete notations to τ_n in $G^{(1)}$ on the loop-type parameterization:

$$\tau_n=\begin{cases} x_n & (\tau^*<\tau_n)\\ y_n & (\tau_n<\tau^*)\\ z_n & \text{on the internal line } T_3,\end{cases} \tag{6.6}$$

where τ^* is given below (see Eq. (6.8) and Fig. 1).

With these relations, the transformation rule between $G^{\rm sym}$ and $G^{(1)}$ is allowed to be expressed as

$$\begin{cases} \tau_n^{(1)}=x_n-\tau^*,\\ \tau_n^{(2)}=\tau^*-y_n,\\ \tau_n^{(3)}=z_n,\end{cases} \tag{6.7}$$

where

$$\tau^*=\tau_\alpha\theta(\tau_\alpha-\tau_\beta)+\tau_\beta\theta(\tau_\beta-\tau_\alpha), \tag{6.8}$$

$$\tau^*-T_2\leqslant y_n\leqslant \tau^*\leqslant x_n\leqslant T_1+\tau^*. \tag{6.9}$$

We can obtain another transformation by combining the property (6.1) with (6.7); i.e., replacing $\tau^{(a)}\to T_a-\tau^{(a)}$ on the right-hand side in (6.7),

$$\begin{cases} \tau_n^{(1)}=T_1-x_n+\tau^*, \\ \tau_n^{(2)}=T_2+y_n-\tau^*, \\ \tau_n^{(3)}=T_3-z_n. \end{cases} \tag{6.10}$$

From these (two-loop) transformations, we can generate an infinite number of transformations for the *one-loop* case, since τ^* is reduced to an arbitrary number when the both edges of the z line approach each other along the fundamental loop (of course, with vanishing T_3); for example,

$$\text{for} \quad \tau^*=0: \quad \begin{cases} \tau^{(1)}=x \\ \tau^{(2)}=-y \end{cases}, \quad \text{or} \quad \begin{cases} \tau^{(1)}=T_1-x \\ \tau^{(2)}=T_2+y \end{cases}, \tag{6.11}$$

$$\text{for} \quad \tau^*=T_2: \quad \begin{cases} \tau^{(1)}=x-T_2 \\ \tau^{(2)}=T_2-y \end{cases}, \quad \text{or} \quad \begin{cases} \tau^{(1)}=T-x \\ \tau^{(2)}=y \end{cases}. \tag{6.12}$$

Although these one-loop transformations are certainly trivial by themselves, an interesting deduction is that one can generate a set of transformations of h-loop Green functions from $(h+1)$-loop transformations by setting one of h copies of τ^* to be an arbitrary value.

As a second application of (6.7), let us consider some reparametrizations of $G_{ab}^{(1)}$; $a,b=0,3$. We show that the transformation of the Green functions $G_{ab}^{(1)}(\tau_1,\tau_2)$ living on the zx-loop (loop made of internal lines where the z and x variables are defined) into $G_{00}^{(1)}(\tau_1',\tau_2')$ on the xy-loop can be found through the cyclic permutation symmetry of G^{sym} (exchanging z-line and y-line). Namely, transforming $G^{(1)}\to G^{\mathrm{sym}}\to G^{(1)}$ successively, we can read how to transform like

$$G_{00}^{(1)}(x_1,x_2)\to G_{00}^{(1)}(x_1',x_2'), \tag{6.13}$$

$$G_{03}^{(1)}(x,z)\to G_{00}^{(1)}(x',y'), \tag{6.14}$$

$$G_{33}^{(1)}(z_1,z_2)\to G_{00}^{(1)}(y_1',y_2'). \tag{6.15}$$

Suppose that each of $G_{ab}^{(1)}$ on the xz-loop is related to $G_{ij}^{\mathrm{sym}}(\tau^{(i)},\tau^{(j)})$ by the rule (6.7), and that each of $G_{00}^{(1)}$ on the xy-loop is related to $G_{ij}^{\mathrm{sym}}(\tau^{(i)},\tau'^{(j)})$ by the same rule as (6.7), $\tau'^{(1)}=x'-\tau^*$, $\tau'^{(2)}=\tau^*-y'$, $\tau'^{(3)}=z'$. Putting $\tau'^{(2)}=\tau^{(3)}$ and $\tau'^{(3)}=\tau^{(2)}$ (corresponding to the exchange of z- and y-lines), and eliminating $\tau^{(a)}$ and $\tau'^{(a)}$ from these transformation rules, we find the following transformation rule attributed from the exchange between z- and y-lines:

$$\begin{cases} x'=x \\ y'=\tau^*-z \quad \text{and} \quad T_3\leftrightarrow T_2 \\ z'=\tau^*-y. \end{cases} \tag{6.16}$$

Remember that Δ^{-1} is invariant in any exchange of T_a. The simplest check of this rule is the following case:

$$\begin{aligned} G_{33}^{(1)}(z_1,z_2)&=|y_1'-y_2'|-\frac{(y_1'-y_2')^2}{(T_3+G_B(\tau_\alpha,\tau_\beta))T}(T_1+T_2)|_{T_2\leftrightarrow T_3} \\ &=|y_1'-y_2'|-\frac{(y_1'-y_2')^2}{(T_3+G_B(\tau_\alpha,\tau_\beta))T}(T_3+T(1-u))=G_{00}^{(1)}(y_1',y_2'). \end{aligned} \tag{6.17}$$

Similarly, we derive another transformation rule from the yz-loop to the xy-loop (exchange of z-line and x-line),

$$\begin{cases} x' = \tau^* + z \\ y' = y \quad \text{and} \quad T_3 \leftrightarrow T_1 \\ z' = x - \tau^* . \end{cases} \tag{6.18}$$

One can organize these two sets of transformations in a unified way: Let us express the untransforming (identical) variables in (6.16) and (6.18) as

$$\tau = x\,\theta(\tau - \tau^*) + y\,\theta(\tau^* - \tau), \tag{6.19}$$

and assign $\tilde{\tau}$ to be the parameter transforming to the z' variable,

$$\tilde{\tau} = y\,\Theta(\tau - \tau^*) + x\,\theta(\tau^* - \tau), \tag{6.20}$$

while considering the transformation

$$G_{00}^{(1)}(\tau_1, \tau_2) \to G_{00}^{(1)}(\tau_1', \tau_2'), \tag{6.21}$$

$$G_{03}^{(1)}(\tau, z) \to G_{00}^{(1)}(\tau', \tilde{\tau}'), \tag{6.22}$$

$$G_{33}^{(1)}(z_1, z_2) \to G_{00}^{(1)}(\tilde{\tau}_1', \tilde{\tau}_2'). \tag{6.23}$$

The above two sets of rules (6.16) and (6.18) are now expressed in the compact form

$$\begin{cases} \tau' = \tau \\ \tilde{\tau}' = \tau^* - z\,\text{sign}(\tau - \tau^*) \quad \text{and} \quad T_3 \leftrightarrow T^* \\ z' = (\tau^* - \tilde{\tau})\text{sign}(\tau - \tau^*) , \end{cases} \tag{6.24}$$

where

$$T^* = T_2\,\theta(\tau - \tau^*) + T_1\,\theta(\tau^* - \tau). \tag{6.25}$$

In order to verify these relations, one should note that TP_{11} $(=\Delta^{-1})$ is invariant under this transformation rule, and also that T transforms as

$$T = T^* + \tilde{T}^* \to T_3 + \tilde{T}^*, \tag{6.26}$$

where

$$\tilde{T}^* = T - T^* = T_1\,\theta(\tau - \tau^*) + T_2\,\theta(\tau^* - \tau). \tag{6.27}$$

It is rather convenient to rewrite the Green function (2.7) as

$$G_{03}^{(1)}(\tau, z) = z + |\tau - \tau^*| - \frac{1}{TP_{11}}[z^2 T + 2z|\tau - \tau^*|T^* + (\tau - \tau^*)^2 (T_3 + T^*)], \tag{6.28}$$

than considering the original form

$$G_{03}^{(1)}(\tau, z) = G_B^{(1)}(\tau, \tau^*) + \frac{1}{P_{11}}\{T_3 z - z^2 - \text{sign}(\tau_\alpha - \tau_\beta)[G_B(\tau, \tau_\alpha) - G_B(\tau, \tau_\beta)]\}. \tag{6.29}$$

Applying the transformation (6.24) to Eqs. (6.21), (6.28), and (6.23), we obtain

$$G_{00}^{(1)}(\tau_1, \tau_2) = G_B(\tau_1, \tau_2) - \frac{\Delta}{T} T^{*2} (\tau_1 - \tau_2)^2, \tag{6.30}$$

$$G_{00}^{(1)}(\tau,\tilde{\tau}) = G_B(\tau,\tilde{\tau}) - \frac{\Delta}{T}[T(\tau^* - \tilde{\tau}) - T^*(\tau - \tilde{\tau})]^2, \tag{6.31}$$

$$G_{00}^{(1)}(\tilde{\tau}_1,\tilde{\tau}_2) = G_B(\tilde{\tau}_1,\tilde{\tau}_2) - \frac{\Delta}{T}\tilde{T}^{*2}(\tilde{\tau}_1 - \tilde{\tau}_2)^2. \tag{6.32}$$

These representations are independent of the choice of either $\tau^* = \tau_\alpha$ or τ_β, and reproduce Eq. (23) of Ref. 12 for the particular choice $\tau^* = \tau_\beta$ (correcting an error in the literature).

VII. CONCLUSIONS

In this paper, we have investigated two types of the constraints on the two-loop kinematical factor and the world-line Green functions. One is nothing but the momentum conservation law on external legs, and the other is the vanishing constraint on the source term integrals along the whole of world-line. Although there is no direct connection between two of them, the latter can be regarded as a continuous version of the former. Because of the ambiguity raised by the constraints, an infinite number of wide sense Green functions are in fact possible to take part in the kinematical factor exponent. However, we have verified that all these Green functions can be identified with the standard (restricted) Green functions, all of which are reduced from a world-sheet Green function,[8] and some of which are related to actual solutions of defining differential equations with possessing the rotational invariance along the fundamental loop.[11] Conversely, this fact means that the conservation constraints loosen some imposed restrictions on the standard Green functions, and eventually make various evaluations and approaches possible. Since this ambiguity is a reflection of the conservation laws, it is easy to infer that the ignorance of it can lead to the breaking of unitarity and gauge invariance. Therefore we always have to take account of either of two versions of the constraints when dealing with the world-line formalism.

Once having the constraints, one can ignore the differences among different Green functions when applying the Green functions to amplitude formulas. In this sense, these constraints will be useful for analyzing higher loop's world-line Green functions. Especially it is clear that the source constraint is much easier to apply than the momentum conservation constraint in the multiloop cases. In two-loop Yang–Mills theory, there arises a different Green function in the calculation in a constant background field,[14] and the source constraint is actually useful to identify the Green function with the standard one in the vanishing limit of constant background field (as demonstrated in Secs. III and IV). Obviously, the similar thing is expected to occur in the multiloop cases. Since expressions of multiloop Green functions are complicated, these constraints will be useful for simplifying and transforming the expressions into convenient forms together with the transformation property (suggested below (6.12). It might be interesting to speculate a usefulness of our techniques in the thermal world-line cases.[15]

In the final part of the paper, we have considered the transformations among the Green functions of standard forms, associated with the reparametrizations of the two-loop world-line diagram. On the one hand, the form of world-sheet Green function is independent of the orderings of two vertices, which join the internal line and the fundamental loop. On the other hand, the crossing type Green functions (2.7) and (2.12), which belong to the type of a correlation between the fundamental loop and the internal line, are neither translational invariant nor ordering independent. It might be that this gap will be filled in some way around by taking account of the discussed transformations into the loop-type Green function $G_{00}^{(1)}$. The crossing-type Green functions are necessary in non-Abelian gauge theory, and a complexity in the combinatorics problem will be caused by this type (similarly to the ϕ^3 theory case[14]). We hope for a useful parameterization or a transformation to overcome these problems.

APPENDIX: FERMION MODE INTEGRATION

We show the derivation of (5.16) in this Appendix. The fermion field (5.15) is an expansion which satisfies $\psi(0) = -\psi(T) \neq 0$ and $\int_0^T \psi^\mu(\tau)d\tau = 0$. First, we rewrite

$$H\equiv\int_0^T\left(-\frac{1}{2}\psi^\mu\partial_\tau\psi_\mu+\eta^\mu\psi_\mu\right)d\tau=I-J, \tag{A1}$$

where

$$I\equiv\int_0^T d\tau_1 d\tau_2 d\tau d\tau'\left(\psi(\tau_1)-\frac{1}{2}\eta(\tau)G_I(\tau-\tau_1)\right)\left(\delta(\tau_1-\tau_2)\frac{-1}{2}\frac{\partial}{\partial\tau_2}\right)\times\left(\psi(\tau_2)-\frac{1}{2}\eta(\tau')G_I(\tau'-\tau_2)\right), \tag{A2}$$

$$J\equiv-\left(\frac{1}{2}\right)^3\int_0^T d\tau_1 d\tau_2 d\tau d\tau'\,\eta(\tau)\eta(\tau')G_I(\tau-\tau_1)\delta(\tau_1-\tau_2)\frac{\partial}{\partial\tau_2}G_I(\tau'-\tau_2), \tag{A3}$$

with introducing the function

$$G_I(\tau)=\frac{2}{\pi}\sum_{m\geqslant 1}\frac{1}{m}\sin\left(\frac{2\pi m\tau}{T}\right), \tag{A4}$$

which satisfies

$$\frac{1}{2}\partial_\tau G_I(\tau)=\delta(\tau)-\frac{1}{T}, \tag{A5}$$

and

$$G_I(\tau_1-\tau_2)=\frac{\partial}{\partial\tau_1}G_B(\tau_1,\tau_2). \tag{A6}$$

Putting (5.12) and (A4) into (A3), and performing the integrals, we obtain

$$J=-\left(\frac{1}{2}\right)^3\sum_{j,l}K_jK_l\left(-\frac{\partial}{\partial\tau_j}\right)\sum_{m\geqslant 1}\frac{2T}{\pi^2m^2}\cos\left(\frac{2\pi m(\tau_i-\tau_j)}{T}\right). \tag{A7}$$

Using the summation formula (5.9), we have

$$J=-\left(\frac{1}{2}\right)^2\sum_{j,l}K_jK_l\left(\frac{2}{T}(\tau_j-\tau_l)-\mathrm{sign}(\tau_j-\tau_l)\right). \tag{A8}$$

Shifting $\psi\to\psi+\frac{1}{2}\eta G_I$ in the path integral (5.11), the quantity I is reduced to the free integral

$$I\to-\frac{1}{2}\int_0^T\psi\cdot\dot\psi, \tag{A9}$$

and this yields nothing but the path integral normalization

$$\oint\mathcal{D}\psi e^{-(1/2)\int_0^T\psi\cdot\dot\psi d\tau}=1. \tag{A10}$$

This can be checked by integrating the modes (5.15). Therefore we derive (5.16) owing to (A8) and (A10),

$$\tilde{\Gamma}_N = \oint \mathcal{D}\psi\left(\prod_{j=1}^{N}\int d\tau_j d\theta_j d\overline{\theta_j}\right)e^H = \prod_{j=1}^{N}\int d\tau_j d\theta_j d\overline{\theta_j} e^J. \tag{A11}$$

[1] Z. Bern and D. A. Kosower, Nucl. Phys. B **379**, 451 (1992).

[2] M. J. Strassler, Nucl. Phys. B **385**, 145 (1992).

[3] A. A. Tseytlin, Phys. Lett. B **168**, 63 (1986); R. R. Metsaev and A. A. Tseytlin, Nucl. Phys. B **298**, 109 (1988); O. D. Andreev and A. A. Tseytlin, Phys. Lett. B **207**, 157 (1988).

[4] P. Di Vecchia, A. Lerda, L. Magnea, and R. Marotta, Phys. Lett. B **351**, 445 (1995); P. Di Vecchia, A. Lerda, L. Magnea, R. Marotta, and R. Russo, Nucl. Phys. B **469**, 235 (1996).

[5] C. Schubert, Acta Phys. Pol. B **27**, 3965 (1996); C. Schubert, PRINT-97-273; M. G. Schmidt and C. Schubert, LAPTH-703-98 (hep-th/9810161).

[6] K. Roland, Phys. Lett. B **289**, 148 (1992); SISSA/ISAS 131-93-EP.

[7] P. Di Vecchia, A. Lerda, L. Magnea, R. Marotta, and R. Russo, Phys. Lett. B **388**, 65 (1996).

[8] K. Roland and H.-T. Sato, Nucl. Phys. B **480**, 99 (1996).

[9] K. Roland and H.-T. Sato, Nucl. Phys. B **515**, 488 (1998).

[10] M. G. Schmidt and C. Schubert, Phys. Rev. D **53**, 2150 (1996); K. Daikouji, M. Shino, and Y. Sumino, *ibid.* **53**, 4598 (1996); M. Reuter, M. G. Schmidt, and C. Schubert, Ann. Phys. (N.Y.) **259**, 313 (1997); H. T. Sato, Nucl. Phys. B **491**, 477 (1997).

[11] M. G. Schmidt and C. Schubert, Phys. Lett. B **331**, 69 (1994).

[12] H.-T. Sato, Phys. Lett. B **371**, 270 (1996).

[13] M. G. Schmidt and H.-T. Sato, Nucl. Phys. B **524**, 742 (1998).

[14] M. G. Schmidt and H.-T. Sato, Nucl. Phys. B (1999) (hep-th/9812229) (to be published).

[15] H. T. Sato, J. Math. Phys. (1999) (hep-th/9809053) (to be published); M. Haack and M. G. Schmidt, Eur. Phys. J. C **7**, 149 (1999); I. A. Shovkovy, Phys. Lett. B **441**, 313 (1998).

The action operator for continuous-time histories

K. Savvidou[a)]
Theoretical Physics Group, Blackett Laboratory, Imperial College of Science, Technology & Medicine, London SW7 2BZ, United Kingdom

(Received 12 March 1999; accepted for publication 29 March 1999)

We define the action operator for the consistent histories formalism, as the quantum analog of the classical action functional, for the simple harmonic oscillator case. We conclude that the action operator is the generator of time transformations, and is associated with the two types of time evolution of the standard quantum theory: the wave-packet reduction and the unitary time evolution. We construct the corresponding classical histories and demonstrate the relevance with the quantum histories. Finally, we show the relation of the action operator to the decoherence functional. © *1999 American Institute of Physics.* [S0022-2488(99)02510-4]

I. INTRODUCTION

One of the basic elements in the consistent histories formalism is the idea of a ''homogeneous history.'' This is a time-ordered sequence of propositions about the system and, in the original approaches to the formalism, is represented by a class operator $\widetilde{C}$,

$$\widetilde{C} := U(t_0,t_1)\alpha_{t_1}U(t_1,t_2)\alpha_{t_2}\cdots U(t_{n-1},t_n)\alpha_{t_n}U(t_n,t_0), \tag{I.1}$$

where α_{t_i} is a single-time projection operator representing a property of the system at time t_i, and $U(t,t')=e^{-(i/h)H(t-t')}$ is the unitary time evolution operator.[1–4]

In the ''History Projection Operator'' (HPO) approach developed by Isham and collaborators,[5–8] a homogeneous history ''α_{t_1} is true at time t_1 and α_{t_2} is true at time $t_2\dots$ and α_{t_n} is true at time t_n'' is represented by a *projection operator* α, defined as the tensor product of projection operators $\alpha := \alpha_{t_1}\otimes\alpha_{t_2}\otimes\cdots\otimes\alpha_{t_n}$ on the n-fold tensor product of copies of the standard Hilbert space $\mathcal{V}_n := \mathcal{H}_{t_1}\otimes\mathcal{H}_{t_2}\otimes\cdots\otimes\mathcal{H}_{t_n}$. This approach reestablishes the logical nature of propositions about a physical system since these projection operators (and their disjunctions) represent a type of *temporal* quantum logic.

Most discussions of the consistent-histories formalism have involved histories defined at a finite set of time points. However, it is important to extend this to include a continuous-time variable (especially for potential applications to quantum field theory), and in order to construct continuous-time histories on the continuous tensor product of copies of the Hilbert space $\mathcal{V}_{\rm cts} := \otimes\mathcal{H}_t$, Isham and Linden defined the History Group[6] as an analog of the canonical group of normal quantum theory. This group plays a crucial role in the physical interpretation of the theory: the spectral projectors of the generators of its Lie algebra represent history propositions about the system.

For the example of a nonrelativistic point particle moving on a line, the history group was defined as a generalized Weyl group with Lie algebra,

$$[x_t,x_{t'}]=0, \tag{I.2}$$

$$[p_t,p_{t'}]=0, \tag{I.3}$$

[a)]Electronic mail: k.savvidou@ic.ac.uk

$$[x_t, p_{t'}] = i\hbar\tau\delta(t-t'), \tag{I.4}$$

where $-\infty < t, t' < +\infty$, and τ is a constant with dimensions of time. It is important to emphasize that the generators of the history algebra x_t and p_t, $t \in \mathbf{R}$, are Schrödinger-picture operators. After being properly smeared, they correspond (actually, their spectral projectors), to propositions about the time-averaged values of the position and the momentum of the system, respectively. The evident resemblance of the history algebra to the algebra of a quantum field theory meant that the one-dimensional quantum mechanics history theory could be treated mathematically in some respects as a $1+1$ dimension quantum *field* theory.

In a previous paper,[5] the requirement of the existence of the Hamiltonian operator H that represents propositions about the time-averaged values of the energy of the system—in particular, for the example of a simple harmonic oscillator in one dimension—together with the explicit relation between the Hamiltonian and the creation and annihilation operators, selected uniquely a Fock space as the representation space of the history algebra (1.2–1.4) on the history space $\mathcal{V}_n$. We shall return to this representation in more detail shortly.

The history algebra generators x_t and p_t can be seen heuristically as operators (actually they are operator-valued distributions on $\mathcal{V}_n$), that for each time label t, are defined on the Hilbert space $\mathcal{H}_t$. The question then arises if, and how, these Schrödinger-picture objects with different time labels are related: in particular, is there a transformation law ''from one Hilbert space to another?'' One anticipates that the analog of this question in the context of a histories treatment of a relativistic quantum field theory would be crucial to showing the Poincaré invariance of the system.

In the Hamilton–Jacobi formulation of Classical Mechanics, it is the *action functional* that plays the role of the generator of a canonical transformation of the system from one time to another.[9] Indeed, the Hamilton–Jacobi functional S, evaluated for the realized path of the system—i.e., for a solution of the classical equations of motion, under some initial conditions—is the generating function of a canonical transformation, which transforms the system variables position x and momentum p from an initial time $t=0$ to another time t. It is therefore natural to investigate whether a quantum analog of the action functional exists for the HPO theory.

Indeed, in Ref. 5, where we explored the quantum field theory case for the continuous-time histories, we were not able to show the manifest covariance of the theory under the ''external'' Poincaré group. However, we did not consider the action as an operator, our main goal in the present paper is to enhance the theory in this direction so as to have a clearer view of the time-transformation issue. This will ultimately allow us to readdress the problem of the Poincaré covariance of the quantum field theory.[10]

In what follows, we first prove the existence of the action operator S_κ, using the same type of quantum field theory methods that were used to prove the existence of the Hamiltonian operator H_κ.[5] We will show that, constructed as a quantum analog of the classical action functional, S_κ does indeed act as a generator of time transformations in the HPO theory. Furthermore—and more speculatively—this is arguably related to the two laws of time-evolution in standard quantum theory: namely, wave-packet reduction and the unitary time evolution between measurements.

A comparison with the classical theory case seems appropriate at this point, and thus, in Secs. III and IV, we present a classical analog of the HPO, where the continuous-time classical histories can be seen to be an analog of the continuous-time quantum histories.

In Sec. V, we further exploit the classical analogy to discuss the ''classical'' behavior of the history quantum scheme. In particular, we expect the action operator to be involved in some way with the dynamics of the theory. To this end, we show how it appears in the expression for the decoherence functional expression, with operators acting on coherent states, as used earlier by Isham and Linden.[6]

II. THE ACTION OPERATOR DEFINED

In the generalized consistent histories theory by Gell-Mann and Hartle[1,2] and others, a homogeneous history α is a time-ordered sequence of propositions about the system, and is represented by a class operator $\tilde{C}$,

$$\tilde{C} := U(t_0,t_1)\alpha_{t_1} U(t_1,t_2)\alpha_{t_2} \cdots U(t_{n-1},t_n)\alpha_{t_n} U(t_n,t_0), \tag{II.1}$$

where α_t, is a single-time projection operator representing a proposition about the system at time t_i. If a particular history α belongs to a consistent set, then the probability for the history to be realized is

$$\mathrm{Prob}(\alpha) = \mathrm{tr}_H \tilde{C}^\dagger_\alpha \rho_{t_0} \tilde{C}_\alpha, \tag{II.2}$$

where ρ_{t_0} is the density matrix of the initial state. Deciding whether or not a particular set of histories is consistent involves evaluating the decoherence functional,

$$d(\alpha,\beta) := \mathrm{tr}_H \tilde{C}^\dagger_\alpha \rho_{t_0} \tilde{C}_\beta, \tag{II.3}$$

which is a complex-valued function of a pair of histories α and β. In particular, if α and β are disjoint propositions belonging to a consistent set, then they satisfy the ''decoherence'' condition,

$$d(\alpha,\beta) = 0. \tag{II.4}$$

We note that, as a product of projectors, the class operator $\tilde{C}_\alpha$ is generally not itself a projector, and hence the temporal logic structure of quantum mechanics is lost. This is remedied in the HPO theory, in which the history proposition ''α_{t_1} is true at time t_1 and α_{t_2} is true at time t_2... and α_{t_n} is true at time t_n'' is represented by the tensor product $\alpha := \alpha_{t_1} \otimes \alpha_{t_2} \otimes \cdots \otimes \alpha_{t_n}$.[8] This is a genuine projection operator on the n-fold tensor product $\mathcal{V}_n := \mathcal{H}_{t_1} \otimes \mathcal{H}_{t_2} \otimes \cdots \otimes \mathcal{H}_{t_n}$. Each constituent proposition α_t, labeled by the time parameter t, is defined on a copy of the standard quantum theory Hilbert space, with the same t-label $\mathcal{H}_t$.

This is a straightforward idea for a discrete set of times $(t_1,t_2,\ldots,t_n)$, but, for reasons given in the Introduction, it is important to extend these ideas to continuous-time histories that are to be defined on some sort of continuous tensor product of copies of the Hilbert space $\mathcal{V}_{\mathrm{cts}} := \otimes \mathcal{H}_t$.

A key technical tool is the history group, constructed as an analog of the canonical group[3] of normal quantum theory. For a particle moving in one dimension, the standard canonical commutation relation,

$$[x,p] = i\hbar, \tag{II.5}$$

is replaced by the ''history algebra,''

$$[x_t, x_{t'}] = 0, \tag{II.6}$$

$$[p_t, p_{t'}] = 0, \tag{II.7}$$

$$[x_t, p_{t'}] = i\hbar\tau\delta(t-t'), \tag{II.8}$$

where $-\infty < t,t' < +\infty$. The constant τ has dimensions of time[11] and, in what follows, for convenience we shall choose units in which $\tau=1$. These operators are written in the Schrödinger picture: t labels the Hilbert space—it is *not* the time parameter that appears in the Heisenberg picture for normal quantum theory.

To be mathematically precise, Eqs. (II.6)–(II.7) must be smeared,

$$[x_f, x_g] = 0, \tag{II.9}$$

$$[p_f, p_g] = 0, \tag{II.10}$$

$$[x_f,p_g]=i\hbar\int_{-\infty}^{+\infty}f(t)g(t)dt, \tag{II.11}$$

where f and g belong to some appropriate subset of the space $\mathcal{L}^2(\mathbf{R},dt)$ of square integrable functions on $\mathbf{R}$.

The evident resemblance of the above with the canonical commutation algebra of a quantum field theory in 1+1 dimensions leads to the treatment of the history algebra using mathematical ideas drawn from the former. In particular, a unique representation of the history algebra can be selected by the requirement that a representation of the (analog of the) Hamiltonian operator exists:[12] physically, this operator represents history propositions about the time-averaged values of the energy.

In previous work,[5] we explored the familiar example of a simple harmonic oscillator in one dimension. In this case, the history algebra is extended to include the commutators,

$$[H_\kappa,x_f]=-\frac{i\hbar}{m}p_{\kappa f}, \tag{II.12}$$

$$[H_\kappa,p_f]=i\hbar\omega^2x_{\kappa f}, \tag{II.13}$$

$$[H_\kappa,H_{\kappa'}]=0, \tag{II.14}$$

where H_κ is the time-averaged history energy operator $H_\kappa:=\int_{-\infty}^{\infty}\kappa(t)H_t\,dt$. The smearing function $\kappa(t)$ belongs to some subset of the space $\mathcal{L}^2(\mathbf{R},dt)$, in general, not the same as the subset on which the test functions of the x_t and p_t are defined. The specific choice of test functions is partly determined by the physical situations to which the formalism is to be applied.

The Fock representation of the history algebra is based on the definition of the ''annihilation'' operator,

$$b_t:=\sqrt{\frac{m\omega}{2\hbar}}x_t+i\sqrt{\frac{1}{2m\omega\hbar}}p_t, \tag{II.15}$$

with commutation relations

$$[b_t,b_{t'}]=0, \tag{II.16}$$

$$[b_t,b_{t'}^\dagger]=\delta(t-t'), \tag{II.17}$$

and is uniquely selected by the requirement that the time-averaged Hamiltonian operator exists in this representation; heuristically, H_t is connected with the operator $b^\dagger$ by the expression $H_t=\hbar\omega b_t^\dagger b_t$.

In the Hamiltonian formalism for a classical system, the action functional is defined as

$$S_{\rm cl}:=\int_{-\infty}^{+\infty}(p\dot{q}-H)dt, \tag{II.18}$$

where q is the position, p is the momentum, and H the Hamiltonian of the system. Following the same line of thought as for the definition of the Hamiltonian algebra, we want to find a representation of the history algebra in which their exists a one-parameter family of operators S_t—or better their smeared form $S_{\lambda,\kappa}$. Heuristically, we have

$$S_t:=(p_t\dot{x}_t-H_t), \tag{II.19}$$

$$S_{\lambda,\kappa}:=\int_{-\infty}^{+\infty}(\lambda(t)p_t\dot{x}_t-\kappa(t)H_t)dt, \tag{II.20}$$

where $S_{\lambda,\kappa}$ is the smeared action operator with smearing functions $\lambda(t),\kappa(t)$. In order to discuss the existence of an operator $S_{\lambda,\kappa}$, we note that, if this operator exists, the Hamiltonian algebra, Eqs. (II.12)–(II.14), would be augmented in the form

$$[S_{\lambda,\kappa},x_f]=i\hbar\left(x_{(d/dt)(\lambda f)}+\frac{p_{\kappa f}}{m}\right), \tag{II.21}$$

$$[S_{\lambda,\kappa},p_f]=i\hbar(p_{\lambda f}+m\omega x_{\kappa f}), \tag{II.22}$$

$$[S_{\lambda,\kappa},H_{\kappa'}]=i\hbar H_{(d/dt)(\lambda\kappa')}-\frac{i\hbar}{m}\int_{-\infty}^{\infty}(\kappa'(t)\dot{\lambda}(t)\dot{p}_t^2)dt, \tag{II.23}$$

$$[S_{\lambda,\kappa},S'_{\lambda,\kappa}]=i\hbar H_{(d/dt)(\lambda'\kappa)}-i\hbar H_{(d/dt)(\lambda\kappa')}-i\hbar\int_{-\infty}^{\infty}\left([(\kappa(t)\dot{\lambda}'(t))-(\kappa'(t)\dot{\lambda}(t))]\frac{\dot{p}_t^2}{m}\right)dt. \tag{II.24}$$

Although we have defined the action operator in a general smeared form, in what follows we will mainly employ only the case $\lambda(t)=1$ and $\kappa(t)=1$ that accords with the expression for the classical action functional. This choice of smearing functions poses no technical problems restrictions, provided we keep to the requirement that the smearing functions for the position and momentum operators are square-integrable functions. In particular, the products of the smearing functions f and g in Eqs (II.21)–(II.24) with the test functions $\lambda(t)=1$ and $\kappa(t)=1$, are still square integrable.

The existence of the action operator in HPO. We now examine whether the action operator actually exists in the Fock representation of the history algebra employed in our earlier work.[5] Henceforward we choose $\lambda(t)=1$. Then the formal commutation relations are

$$S_\kappa:=\int_{-\infty}^{+\infty}(p_t\dot{x}_t-\kappa(t)H_t)dt, \tag{II.25}$$

$$[S_\kappa,x_f]=i\hbar\left(x_{\dot{f}}+\frac{p_{\kappa f}}{m}\right), \tag{II.26}$$

$$[S_\kappa,p_f]=i\hbar(p_{\dot{f}}+m\omega x_{\kappa f}), \tag{II.27}$$

$$[S_\kappa,H_{\kappa'}]=i\hbar H_{\dot{\kappa}'}, \tag{II.28}$$

$$[S_\kappa,S_{\kappa'}]=i\hbar H_{\dot{\kappa}}-i\hbar H_{\dot{\kappa}'}. \tag{II.29}$$

A key observation is that *if* the operators $e^{(i/\hbar)S_\kappa}$ existed, they would produce the history algebra automorphism,

$$e^{(i/\hbar)sS_\kappa}b_te^{-(i/\hbar)sS_\kappa}=e^{-i\omega\int_t^{t+s}\kappa(t+s')ds'+s(d/dt)}b_t, \tag{II.30}$$

or, in the more rigorous smeared form,

$$e^{(i/\hbar)sS_\kappa}b_fe^{-(i/\hbar)sS_\kappa}=b_{\Sigma_s f}, \tag{II.31}$$

where the unitary operator Σ_s is defined on $\mathcal{L}^2(\mathbf{R})$ by

$$(\Sigma_s\psi)(t):=e^{-i\omega\int_t^{t+s}\kappa(t+s')ds'}\psi(t+s). \tag{II.32}$$

However, an important property of the Fock construction states that when there exists a unitary operator e^{isA} acting on $\mathcal{L}^2(\mathbf{R})$, there exists a unitary operator $\Gamma(e^{isA})$ that acts on the exponential Fock space [a general expression for a Fock space is $e^{\mathcal{H}}=\oplus_{n=0}^{\infty}(\otimes_n \mathcal{H})$, where $\mathcal{H}$ is called the base Hilbert space of the Fock space $e^{\mathcal{H}}$] $\mathcal{F}(\mathcal{L}^2(\mathbf{R}))$ in such a way that

$$\Gamma(e^{isA})b_f^\dagger\Gamma(e^{isA})^{-1}=b^\dagger_{e^{isA}f}, \tag{II.33}$$

then the operator $d\Gamma(A)$ on $\mathcal{F}(\mathcal{L}^2(\mathbf{R}))$ can also be defined as

$$\Gamma(e^{isA})=e^{is\,d\Gamma(A)}, \tag{II.34}$$

in terms of A, a self-adjoint operator that acts on $\mathcal{L}^2(\mathbf{R})$. In particular, it follows that the representation of the history algebra on the Fock space $\mathcal{F}(\mathcal{L}^2(\mathbf{R}))$ carries a (weakly continuous) representation of the one-parameter family of unitary operators $s\mapsto e^{(i/\hbar)sS_\kappa}=\Gamma(\Sigma_s)$. Therefore, the generator S_κ also exists on $\mathcal{F}(\mathcal{L}^2(\mathbf{R}))$ and $S=d\Gamma(-\hbar\sigma_\kappa)$, where σ_κ is a self-adjoint operator that acts on $\mathcal{L}^2(\mathbf{R})$ and is defined as

$$\sigma_\kappa\psi(t):=\left(-\omega\kappa(t)-i\frac{d}{dt}\right)\psi(t). \tag{II.35}$$

In what follows, we will restrict our attention to the particular case $\kappa(t)=1$ for the simple harmonic oscillator action operator S,

$$S:=\int_{-\infty}^{+\infty}\langle p_t\dot{x}_t-H_t\rangle dt. \tag{II.36}$$

The Liouville operator definition. The first term of the action operator, Eq. (II.36), is identical to the kinematical part of the classical action functional, Eq. (II.18). For reasons that will become apparent later, we write S_κ as the difference between two operators: the Liouville operator and the Hamiltonian operator. The Liouville operator is formally written as

$$V:=\int_{-\infty}^{\infty}(p_t\dot{q}_t)dt, \tag{II.37}$$

where

$$S_\kappa=V-H_\kappa. \tag{II.38}$$

We prove the existence of V on $\mathcal{F}(\mathcal{L}^2(\mathbf{R}))$ using the same technique as before. Namely, we can see at once that the history algebra automorphism,

$$e^{(i/\hbar)sV}b_f e^{-(i/\hbar)sV}=b_{B_s f}, \tag{II.39}$$

is unitarily implementable. Here, the unitary operator B_s, $s\in\mathbf{R}$, acting on $\mathcal{L}^2(\mathbf{R})$ is defined by

$$(B_s f)(t):=e^{s(d/dt)}f(t)=e^{isD}f(t)=f(t+s), \tag{II.40}$$

where $D:=-i(d/dt)$. The Liouville operator V has some interesting commutation relations with the generators of the history algebra:

$$[V,x_f]=-i\hbar x_{\dot{f}}, \tag{II.41}$$

$$[V,p_f]=-i\hbar p_{\dot{f}}, \tag{II.42}$$

$$[V,H_\kappa]=-i\hbar H_{\dot{\kappa}}, \tag{II.43}$$

$$[V,S_k]=i\hbar H_{\dot\kappa}, \tag{II.44}$$

$$[V,H]=0, \tag{II.45}$$

$$[V,S]=0, \tag{II.46}$$

where we have defined $H:=\int_{-\infty}^{\infty}H_t\,dt$.

We notice that V transforms, for example, b_t from one time t—that refers to the Hilbert space $\mathcal{H}_t$—to another time $t+s$, that refers to $\mathcal{H}_{t+s}$. More precisely, V transforms the support of the operator-valued distribution b_t from t to $t+s$:

$$e^{(i/\hbar)sV}b_f e^{-(i/\hbar)sV}=b_{f_s}, \tag{II.47}$$

where $f_s(t):=f(s+t)$. We shall return to the significance of this later.

The Fourier-transformed ''n-particle'' history propositions. An interesting family of history propositions emerges from the representation space $\mathcal{F}[\mathcal{L}^2(\mathbf{R},dt)]$, acting on the δ-function normalized basis of states $|0\rangle$, $|t_1\rangle:=b^\dagger_{t_1}|0\rangle$, $|t_1,t_2\rangle:=b^\dagger_{t_1}b^\dagger_{t_2}|0\rangle$, etc.; or, in smeared form, $|\phi\rangle:=b^\dagger_\phi|0\rangle$, etc. The projection operator $|t\rangle\langle t|$ corresponds to the history proposition ''there is a unit energy $\hbar\omega$ concentrated at the time point t.'' The physical interpretation for this family of propositions, was deduced from the action of the Hamiltonian operator on the family of $|t\rangle$ states:

$$H_t|0\rangle=0, \tag{II.48}$$

$$H_t|t_1\rangle=\hbar\omega\,\delta(t-t_1)|t_1\rangle, \tag{II.49}$$

$$H_t|t_1,t_2\rangle=\hbar\omega[\delta(t-t_1)+\delta(t-t_2)]|t_1,t_2\rangle.$$
$$\vdots \tag{II.50}$$

To study the behavior of the S operator, a particularly useful basis for $\mathcal{F}[\mathcal{L}^2(\mathbf{R},dt)]$ is the Fourier transforms of the $|t\rangle$ states. Indeed, if we consider the Fourier transformations,

$$|\nu\rangle=\int_{-\infty}^{+\infty}e^{i\nu t}b_t^\dagger|0\rangle dt, \tag{II.51}$$

$$|\nu_1,\nu_2\rangle=\int_{-\infty}^{+\infty}e^{i\nu_1t_1}e^{i\nu_2t_2}b^\dagger_{t_1}b^\dagger_{t_2}|0\rangle dt_1\,dt_2, \tag{II.52}$$

$$b_\nu=\int_{-\infty}^{+\infty}e^{i\nu t}b_t\,dt, \tag{II.53}$$

$$b_\nu^\dagger=\int_{-\infty}^{+\infty}e^{-i\nu t}b_t^\dagger\,dt, \tag{II.54}$$

the Fourier transformed $|\nu\rangle$ states are defined by $|\nu\rangle:=b^\dagger_\nu|0\rangle$, $|\nu_1,\nu_2\rangle:=b^\dagger_{\nu_1}b^\dagger_{\nu_2}|0\rangle$, etc. The eigenvectors of the operator S are calculated to be

$$S|0\rangle=0, \tag{II.55}$$

$$S|\nu\rangle=\hbar(\nu-\omega)|\nu\rangle, \tag{II.56}$$

$$S|\nu_1,\nu_2\rangle=\hbar[(\nu_1-\omega)+(\nu_2-\omega)]|\nu_1,\nu_2\rangle, \tag{II.57}$$

and we note, in particular, that $e^{(i/\hbar)sS}|0\rangle=|0\rangle$.

The $|\nu\rangle$ states are also eigenstates of the Hamiltonian operator:

$$H|0\rangle=0, \tag{II.58}$$

$$H|\nu\rangle=\hbar\omega|\nu\rangle, \tag{II.59}$$

$$H|\nu_1,\nu_2\rangle=2\hbar\omega|\nu_1,\nu_2\rangle. \\ \vdots \tag{II.60}$$

Again, as for the case of the $|t\rangle$ states, for the special case of $H:=\int_{-\infty}^{\infty}H_t\,dt$, and for the simple harmonic oscillator example, we see how the integer-spaced spectrum of the standard quantum field theory appears in the HPO theory. The $|\nu\rangle\langle\nu|$ history propositions give the spectrum of the action operator and they have an interesting connection with the $|t\rangle\langle t|$ propositions.

A. The velocity operator

In Ref. 5, we emphasized the existence of the operator $\dot{x}_t:=(d/dt)x_t$, that corresponds to history propositions about the velocity of the system. The velocity operator is better defined in its smeared form using the familiar quantum field theory procedure,

$$\dot{x}_f=-x_{\dot{f}}. \tag{II.61}$$

In analogy with quantum field theory, this requires the function f to be differentiable and to ''vanish at infinity'' so that the implicit integration by parts in Eq. (II.61) is valid. We note that, in this HPO theory, the velocity operator commutes with the position

$$[x_t,\dot{x}_{t'}]=0, \tag{II.62}$$

and therefore there exist history propositions about the position and the velocity at the same time.

Furthermore, the existence of the Liouville operator in the HPO scheme, allows an interesting comparison between the velocity $\dot{x}_f$ and the momentum p_f operators: namely, the momentum operator is defined by the history commutation relation of the position with the Hamiltonian, while we can define the velocity operator from the history commutation relation of the position with the Liouville operator:

$$[x_f,H]=i\hbar\frac{p_f}{m}, \tag{II.63}$$

$$[x_f,V]=i\hbar\dot{x}_f. \tag{II.64}$$

These relations signify the different nature of the momentum p_f from the velocity $\dot{x}_f$ concerning the dynamical behavior of the momentum (related to the Hamiltonian operator), as opposed to the kinematical behavior of the velocity (related to the Liouville operator).

B. The Heisenberg picture

In standard quantum theory, a Heisenberg-picture operator $A(s)$ is defined as

$$A_{\mathrm{H}}(s):=e^{(i/\hbar)sH}Ae^{-(i/\hbar)sH}. \tag{II.65}$$

In particular, for the case of a simple harmonic oscillator, the equation of motion is

$$\frac{d^2}{ds^2}x(s)+\omega^2x(s)=0, \tag{II.66}$$

from which we obtain the solution

$$x(s)=\cos(s\omega)x+\frac{1}{m\omega}\sin(s\omega)p, \tag{II.67}$$

$$p(s) = -m\omega \sin(s\omega)x + \cos(s\omega)p, \tag{II.68}$$

where we have used the classical equation,

$$p := m \left.\frac{dx(s)}{ds}\right|_{s=0}. \tag{II.69}$$

The commutation relations between these operators is

$$[x(s_1), x(s_2)] = \frac{i\hbar}{m\omega} \sin[\omega(s_1 - s_2)]. \tag{II.70}$$

In formulating a history analog of the Heisenberg picture,[5] we adopted a ''time-averaged'' Heisenberg picture defined by

$$x_{\kappa,t} := e^{i/\hbar H_\kappa} x_t e^{-i/\hbar H_\kappa} = \cos[\omega\kappa(t)]x_t + \frac{1}{m\omega}\sin[\omega\kappa(t)]p_t, \tag{II.71}$$

for suitable test functions κ. The analog of the equations of motion is the functional differential equation,

$$\frac{\delta^2 x_{\kappa,t}}{\delta\kappa(s_1)\,\delta\kappa(s_2)} + \delta(t - s_1)\,\delta(t - s_2)\,\omega^2 x_{\kappa,t} = 0 \tag{II.72}$$

and

$$\delta(t - s)p_t = m \left.\frac{\delta x_{\kappa,t}}{\delta\kappa(s)}\right|_{\kappa=0} \tag{II.73}$$

is the history analog of the classical equation, $p := m[dx(s)/ds]|_{s=0}$.

We noted then that the Heisenberg-picture in a HPO theory involves two time labels: an ''external'' label t—that specifies the time the proposition is asserted—and an ''internal'' label s that, for a fixed time t, is the time parameter of the Heisenberg picture associated with the copy $\mathcal{H}_t$ of the standard Hilbert space. Using our new results, the two labels appear naturally in a new version of the Heisenberg picture: they are related to the groups that produce the two types of time transformations. In addition, the analogy with the classical expressions is regained.

To see this explicitly, we define a Heisenberg picture analog of x_t as

$$x_{\kappa,t,s} := e^{(i/\hbar)sH_\kappa} x_t e^{-(i/\hbar)sH_\kappa} = \cos[\omega s\kappa(t)]x_t + \frac{1}{m\omega}\sin[\omega s\kappa(t)]p_t \tag{II.74}$$

$$p_{\kappa,t,s} := e^{(i/\hbar)sH_\kappa} p_t e^{-(i/\hbar)sH_\kappa} = -m\omega \sin[\omega s\kappa(t)]x_t + \cos[\omega s\kappa(t)]p_t. \tag{II.75}$$

The commutation relations for these operators are

$$[x_{\kappa,t}(s), x_{\kappa',t'}(s')] = \frac{i\hbar}{m\omega}\sin[\omega\kappa(s' - s)]\,\delta(t - t'), \tag{II.76}$$

$$[x_{\kappa,t}(s), S_{\kappa'}] = i\hbar\left[\cos[s\omega\kappa(t)]\dot{x}_t + \frac{1}{m\omega}\sin[s\omega\kappa(t)]\dot{p}_t - \frac{\kappa'}{m}p_{\kappa,t,s}\right], \tag{II.77}$$

$$[p_{\kappa,t}(s), S_{\kappa'}] = i\hbar\left[\cos[s\omega\kappa(t)]\dot{p}_t - m\omega\sin[s\omega\kappa(t)]\dot{x}_t + \kappa'(t)x_{\kappa,t,s}\right], \tag{II.78}$$

and from these commutators we obtain the HPO analog of the equations of motion,

$$\frac{d^2}{ds^2}x_{\kappa,t,s}+\omega^2\kappa(t)^2 x_{\kappa,t,s}=0. \tag{II.79}$$

We notice the strong resemblance with standard quantum theory; for the case $\kappa(t)=1$, the classical expressions are fully recovered.

In the HPO formalism, the Heisenberg picture objects appear time averaged with respect to the "external" time label t. On the other hand, the "internal" time label s is the time-evolution parameter of the standard Heisenberg picture, as viewed in the Hilbert space $\mathcal{H}_t$. In what follows, we will show how the Heisenberg picture operators evolve in time under the action of the groups of time transformations.

III. TIME TRANSFORMATION IN THE HPO FORMALISM

In classical theory, the Hamiltonian H is the generator of time transformations. In terms of Poisson brackets, the generalized equation of motion for an arbitrary function u is given by

$$\frac{du}{dt}=\{u,H\}+\frac{\partial u}{\partial t}. \tag{III.1}$$

In a HPO theory, the Hamiltonian operator H_t produces phase changes in time, preserving the time label t of the Hilbert space on which, at least formally, H_t is defined. On the other, it is the Liouville operator V that assigns, analogous to the classical case, history commutation relations, and produces time transformations "from one Hilbert space to another." The action operator generates a combination of these two types of time transformation. If we use the notation $x_f(s)$ for the *history* Heisenberg-picture operators smeared with respect to the time label t, we observe that they behave as standard Heisenberg-picture operators, with time parameter s. Furthermore, their history commutation relations strongly resemble the classical expressions:

$$[x_f(s),V]=i\hbar\dot{x}_f(s), \tag{III.2}$$

$$[x_f(s),H]=\frac{i\hbar}{m}p_f(s), \tag{III.3}$$

$$[x_f(s),S]=i\hbar\left(\dot{x}_f(s)-\frac{1}{m}p_f(s)\right). \tag{III.4}$$

We define a one-parameter group of transformations $T_V(\tau)$, with elements $e^{(i/\hbar)\tau V}$, $\tau\in\mathbf{R}$, where V is the Liouville operator and we consider its action on the b_t operator; for simplicity, we write the unsmeared expressions

$$e^{(i/\hbar)\tau V}b_{t,s}e^{-(i/\hbar)\tau V}=b_{t+\tau,s}. \tag{III.5}$$

The Liouville operator is the generator of transformations of the time parameter t labeling the Hilbert spaces $\mathcal{H}_t$.

Then, we define a one-parameter group of transformations T_{H}, with elements $e^{(i/\hbar)\tau H}$, where H is the time-averaged Hamiltonian operator,

$$e^{(i/\hbar)\tau H}b_{t,s}e^{-(i/\hbar)\tau H}=b_{t,s+\tau}. \tag{III.6}$$

The Hamiltonian operator is the generator of phase changes of the time parameter s, produced only on one Hilbert space $\mathcal{H}_t$, for a fixed value of the t parameter.

Finally, we define the one-parameter group of transformations T_S, with elements $e^{(i/\hbar)\tau S}$, where S is the action operator,

$$e^{(i/\hbar)\tau S}b_{t,s}e^{-(i/\hbar)\tau S}=b_{t+\tau,s+\tau}. \tag{III.7}$$

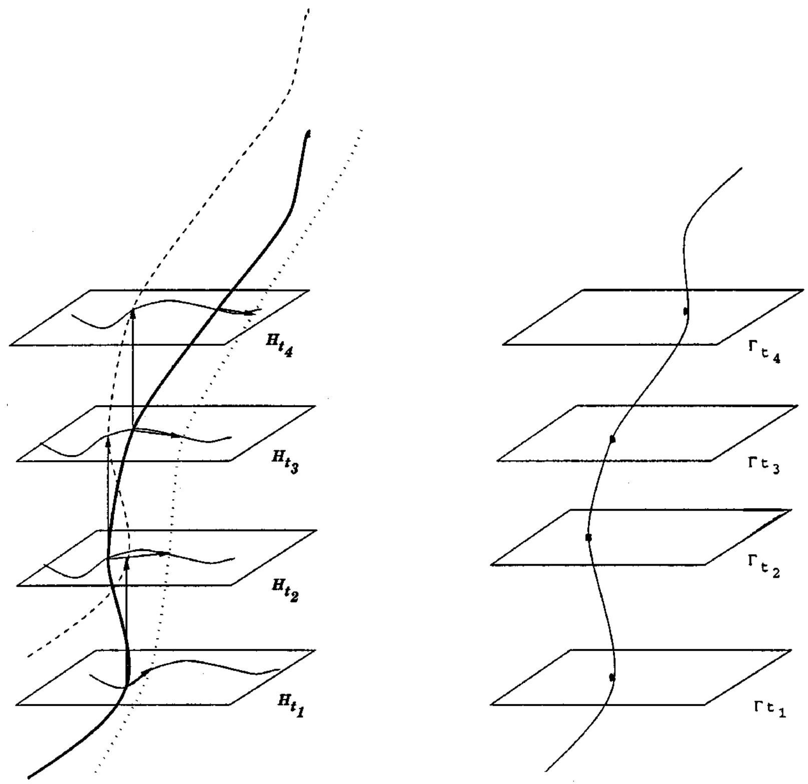

FIG. 1. Quantum and classical history curves. In (a) the transformation of the history curves generated by V is represented by the dashed line, while the transformation generated by H are represented by the dotted line. The curves drawn on each ''Hilbert plane'' correspond to the Hamiltonian transformations, as affected on the corresponding Hilbert space. In (b) the classical history remains invariant under the corresponding time transformations.

The action operator generates both types of time transformations—a feature that appears only in the HPO scheme.

In Fig. 1(a) and 1(b), we denote a quantum continuous-time history as a curve and the tensor product of Hilbert spaces as a sequence of planes, each one representing a copy of the standard Hilbert space. Each plane is labeled by the time label t that the corresponding Hilbert space H_t carries. We depict then a history, as a curve along an n-fold sequence of ''Hilbert planes'' $\mathcal{H}_{t_i}$. In analogy to this, we symbolize a classical history, as a curve along an n-fold sequence of planes corresponding to copies of the standard phase space Γ_{t_i}, as we will explain later. The time transformations generated by the Liouville operator shift the path in the direction of the ''Hilbert planes.'' On the other hand, the Hamiltonian operator generates time transformations that move the history curve in the direction of the path, as represented on one ''Hilbert plane.''

The duality in time evolution. In standard quantum theory, time evolution is described by two different laws: the wave-packet reduction that occurs at a measurement, and the unitary time evolution that takes place between measurements. Thus, according to von Neumann, one has to augment the Schrödinger equation with the collapse of the wave function associated with a measurement.[13]

It seems that the two types of time transformations observed in the HPO theory correspond to the two dynamical processes in standard quantum theory: the time transformations generated by the Liouville operator V are (argueably) related to the wave-packet reduction (the time ordering implied by the wave-packet reduction to be precise), while the time transformations produced by the Hamiltonian operator H are related to the unitary time evolution between measurements.

The argument in support of this assertion is as follows. Keeping in mind the description of the History space as a tensor product of single-time Hilbert spaces $\mathcal{H}_t$, the V operator acts on the

Schrödinger-picture projection operators, translating them in time from one Hilbert space to another. These time-ordered projectors appear in the expression for the decoherence functional that defines probabilities. In history theory, the expression for probabilities in a consistent set is the same as that derived in standard quantum theory using the projection postulate on a time-ordered sequence of measurements.[1,2] It is this that suggests a relation of the Liouville operator to ''wave-packet reduction.'' To strengthen this claim, in what follows we will show the analogy of V with the S_{cts} operator (an approximation of the derivative operator), that appears in the decoherence function and is implicitly related to the wave-packet reduction by specifying the time ordering of the action of the single-time projectors. The action of V as a generator of time translations depends on the partial (in fact, total) ordering of the time parameter treated as the causal structure in the underlying spacetime. Hence, the V-time translations illustrate the purely kinematical function of the Liouville operator.

The Hamiltonian operator producing transformations, with an evident reference to the Heisenberg time evolution, appears as the ''clock'' of the theory. As such, it depends on the particular physical system that the Hamiltonian describes. Indeed, we would expect the definition of a ''clock'' for the evolution in time of a physical system to be connected with the dynamics of the system concerned. We note that the idea of reparametrizing time depends on the smearing function $\kappa(t)$ used in the definition of the Hamiltonian operator; $\kappa(t)$ is kept fixed for a particular physical system.

The coexistence of the two types of time evolution, as reflected in the action operator identified as the generator of such time transformations, is a striking result. In particular, its definition is in accord with its classical analog, namely, the Hamilton action functional. In classical theory, a distinction between a kinematical and a dynamical part of the action functional also arises, in the sense that the first part corresponds to the symplectic structure and the second to the Hamiltonian.

IV. THE CLASSICAL SIGNATURE OF THE HPO FORMALISM

Let us now consider more closely the relation of the classical and the quantum histories. We have shown above how the action operator generates time translations from one Hilbert space to another through the Liouville operator; and on each labeled Hilbert space $\mathcal{H}_t$, through the Hamiltonian operator. We now wish to discuss in more detail the analog of these transformations in the classical case.

We recall that a history is a time-ordered sequence of propositions about the system. The continuous-time quantum history in the HPO system makes assertions about the values of the position or the momentum of the system, or a linear combination of them, at each moment of time, and is represented by a projection operator on the continuous tensor product of copies of the standard Hilbert space.

One expects that a continuous-time classical history should reflect the underlying temporal logic of the situation. Thus, the assertions about the position and the momentum of the system at each moment of time should be represented on an analogous history space: this can be achieved by using the Cartesian product of a continuous family (labeled by the time t) of copies of the standard classical state space.

In classical mechanics, a (fine-grained) classical history is represented by a path in the state space. Indeed, a path γ is defined as a map on the standard phase space,

$$\gamma:\mathbf{R}\rightarrow\Gamma,$$

$$t\mapsto(q(\gamma(t)),p(\gamma(t))), \tag{IV.1}$$

where $q(\gamma(t))$ and $p(\gamma(t))$ are the position and momentum coordinates of the path γ, at the time t. For our purposes, we shall consider the path $t\mapsto\gamma(t)$ to defined for t in some finite time interval $[t_1,t_2]$. We shall denote the set of such paths by Π.

The key idea of this new approach to classical histories is contained in the symplectic structure of the theory: the choice of the Poisson bracket must be such that it includes entries at different moments of time. Thus, we suppose that the space of functions on Π is equipped with the ''history Poisson bracket'' defined by

$$\{q_t, p_{t'}\} = \delta(t - t'), \tag{IV.2}$$

where we defined the functions q_t on Π as

$$q_t : \Pi \rightarrow \mathbf{R}$$

$$\gamma \mapsto q_t(\gamma) := q(\gamma(t)),$$

and similarly for p_t.

We now define the history action functional $S_h(\gamma)$ on Π as

$$S_h(\gamma) := \int_{t_1}^{t_2} [p_t \dot{q}_t - H_t(p_t, q_t)](\gamma) dt, \tag{IV.3}$$

where $q_t(\gamma)$ is the position coordinate q at the time point $t \in [t_1, t_2]$ of the path γ, and $\dot{q}_t(\gamma)$ is the velocity coordinate at the time point $t \in [t_1, t_2]$ of the path γ.

We also define the history classical analogs for the Liouville and time-averaged Hamiltonian operators as

$$V_h(\gamma) := \int_{t_1}^{t_2} [p_t \dot{q}_t](\gamma) dt, \tag{IV.4}$$

$$H_h(\gamma) := \int_{t_1}^{t_2} [H_t(p_t, q_t)](\gamma) dt, \tag{IV.5}$$

$$S_h(\gamma) = V_h(\gamma) - H_h(\gamma). \tag{IV.6}$$

In classical mechanics, the least action principle states that there exists a functional $S(\gamma) = \int_{t_1}^{t_2} [p\dot{q} - H(p,q)](\gamma) dt$ such that the physically realized path is the curve in state space, γ_0, with respect to which the condition $\delta S(\gamma_0) = 0$ holds, when we consider variations around this curve. From this, the Hamilton equations are deduced to be

$$\dot{q} = \{q, H\}, \tag{IV.7}$$

$$\dot{p} = \{p, H\}, \tag{IV.8}$$

where q and p—the coordinates of the realized path γ_0—are the solutions of the classical equations of motion. For any function $F(q,p)$ of the classical solutions, it is also true that

$$\{F, H\} = \dot{F}. \tag{IV.9}$$

In the case of the classical continuous-time histories, one can formulate the above variational principal in terms of the Hamilton equations with the statement: A classical history $\gamma_{\rm cl}$ is the realized path of the system—i.e., a solution of the equations of motion of the system—if it satisfies the equations

$$\{q_t, V_h\}(\gamma_{\rm cl}) = \{q_t, H_h\}(\gamma_{\rm cl}), \tag{IV.10}$$

$$\{p_t, V_h\}(\gamma_{\rm cl}) = \{p_t, H_h\}(\gamma_{\rm cl}), \tag{IV.11}$$

where $\gamma_{\rm cl}=(q_t(\gamma_{\rm cl}),p_t(\gamma_{\rm cl}))$, and $q_t(\gamma_{\rm cl})$ is the position coordinate of the realized path $\gamma_{\rm cl}$ at the time point t. The equations (IV.10)–(IV.11) are the history equivalent of the Hamilton equations of motion. Indeed, for the case of the simple harmonic oscillator in one dimension, Eqs. (IV.10)–(IV.11) become

$$\dot{q}_t(\gamma_{\rm cl})=\frac{p_t}{m}(\gamma_{\rm cl}), \tag{IV.12}$$

$$\dot{p}_t(\gamma_{\rm cl})=-m\omega^2 q_t(\gamma_{\rm cl}), \tag{IV.13}$$

where $\dot{q}_t(\gamma_{\rm cl})=\dot{q}(\gamma_{\rm cl}(t))|$ is the value of the velocity of the system at time t. One would have expected the result in Eqs. (IV.10)–(IV.11) for the classical analog of the histories formalism, as it shows that the classical analog of the two types of time transformation in the quantum theory coincide.

From Eqs. (IV.10)–(IV.11), we also conclude that the canonical transformation generated by the history action functional $S_h(\gamma_{\rm cl})$ leaves invariant the paths that are classical solutions of the system:

$$\{q_t,S_h\}(\gamma_{\rm cl})=0, \tag{IV.14}$$

$$\{p_t,S_h\}(\gamma_{\rm cl})=0. \tag{IV.15}$$

It also holds that any function F on Π satisfies the equation

$$\{F,S_h\}(\gamma_{\rm cl})=0. \tag{IV.16}$$

Some of these statements are implicit in a previous work by Anastopoulos;[14] an interesting application of a similar extended Poisson bracket using a different formulation has been done by Kouletsis.[15]

''Classical'' coherent states for the simple harmonic oscillator. The relation between the classical and the quantum theories can be further exemplified by using coherent states. This special class of states was used in Ref. 6 to represent certain continuous-time history propositions in the history space. Coherent states are particularly useful for this purpose since they form a natural (overcomplete) base for the Fock space representation of the history algebra.

A class of coherent states in the relevant Fock space is generated by unitary transformations on the cyclic vacuum state:

$$|f,h\rangle := U[f,h]|0\rangle, \tag{IV.17}$$

where $U[f,h]$ is the Weyl operator defined as

$$U[f,h]:=e^{(i/\hbar)(x_f-p_h)}, \tag{IV.18}$$

where f and h are test functions in $L^2(\mathbf{R})$. The Weyl generator,

$$\alpha(f,h):=x(f)-p(h), \tag{IV.19}$$

can alternatively be written as

$$\alpha(f,h)=\frac{\hbar}{i}(b^\dagger(w)-b(w^*)), \tag{IV.20}$$

where $w:=f+ih$.

Suppose now that, for a pair of functions (f,h), the operator $\alpha(f,h)$ commutes with the action operator S,

$$[S,\alpha(f,h)]=0. \tag{IV.21}$$

Then any pair (*f*,*h*) satisfying this equation is necessarily a solution of the system of differential equations obtained from Eq. (IV.21):

$$\dot{f}+m\omega^2 h=0, \tag{IV.22}$$

$$\dot{h}-\frac{f}{m}=0. \tag{IV.23}$$

We see that if we identify f with the classical momentum $p_{\rm cl}$ and h with the classical position $x_{\rm cl}$, then Eqs. (IV.22)–(IV.23) are precisely the classical equations of motion for the simple harmonic oscillator:

$$\ddot{x}_{\rm cl}+\omega^2 x_{\rm cl}=0. \tag{IV.24}$$

The classical solutions (*f*,*h*) distinguish a special class of Weyl operators $\alpha_{\rm cl}(f,h)$, and, hence, a special class of coherent states:

$$|\exp z_{\rm cl}\rangle := U_{\alpha_{\rm cl}(f,h)}|0\rangle, \tag{IV.25}$$

where $z_{\rm cl}:=f+ih$.

These classical-like features stem from the following relation with S:

$$[S,U_{\alpha_{\rm cl}}]=0, \tag{IV.26}$$

$$[S,P_{|\exp z_{\rm cl}\rangle}]=0, \tag{IV.27}$$

where $P_{|\exp z_{\rm cl}\rangle}$ is the projection operator onto the (non-normalized) coherent state $|\exp z_{\rm cl}\rangle$:

$$P_{|\exp z_{\rm cl}\rangle} := \frac{|\exp z_{\rm cl}\rangle\langle \exp z_{\rm cl}|}{\langle \exp z_{\rm cl}|\exp z_{\rm cl}\rangle}. \tag{IV.28}$$

We note that there exists an analogy between Eqs. (IV.14)–(IV.15) and Eq. (IV.27), if we consider (*f*,*h*) to be the classical solution: $t\mapsto(q_t,p_t)(\gamma_{\rm cl})$. In classical histories, the canonical transformation, Eqs. (IV.14)–(IV.15), generated by the history action functional, vanishes on a solution to the equations of motion. On the other hand, when we deal with quantum histories, the action operator produces the classical equations of motion, Eqs. (IV.23)–(IV.22), when we require that it commutes with the projector [as in Eq. (IV.28)], which corresponds to a classical solution (*f*,*h*) of the system. However, we do not imply from this appearance of the classical limit: to make any such physical predictions, we must involve the decoherence functional and the coarse graining operation.

Notice that the construction above holds for a generic potential, as long as there exists a representation on $\mathcal{V}_{\rm cts}$ of the history algebra on which the action operator is defined.

V. THE DECOHERENCE FUNCTIONAL ARGUMENT

In the consistent histories quantum theory, the dynamics of a system is described by the decoherence functional. In a classical theory it is the action functional that plays a similar role in regard to the dynamics of the system. It is only natural, then, to seek for the appearance of the action operator in the decoherence functional. The aim is to write the HPO expression for the decoherence functional, with respect to an operator that includes S, and to compare this operator (i.e., its matrix elements), with the operator $\mathcal{S}_{\rm cts}\mathcal{U}$ that appears in the decoherence functional.[6]

In the HPO formalism, the decoherence functional d has been constructed for the special case of continuous-time projection operators corresponding to coherent states.[6] To this end, a continuous product of projectors $\otimes_t P_{|\exp \lambda(t)\rangle}$ is identified with $P_{\otimes_t|\exp \lambda(\cdot)\rangle}$: the projector onto the (non-normalized) coherent states $\otimes_t|\exp \lambda(t)\rangle$ in the continuous tensor product $\otimes_t L^2_t(\mathcal{R})$. More precisely, this continuous tensor product is isomorphic to Fock space:

$$\mathcal{V}_{\mathrm{cts}} := \otimes_{t\in \mathbf{R}} \mathcal{L}^2{}_t(\mathbf{R}) \approx \exp(\mathcal{L}^2(\mathbf{R}, dt)), \tag{V.1}$$

and we can identify a projector on the Hilbert space $\mathcal{V}_{\mathrm{cts}}$ as

$$\otimes_{t\in \mathbf{R}} P_{|\exp(\lambda(t))\rangle} = P_{|\exp \lambda(\cdot)\rangle}, \tag{V.2}$$

with $P_{|\exp \lambda(\cdot)\rangle} = e^{-\langle \lambda,\lambda\rangle}|\exp \lambda(\cdot)\rangle\langle \exp \lambda(\cdot)|$. The action of the continuous-time histories projectors on the non-normalized coherent states $|\exp(\lambda(\cdot))\rangle$ is denoted by

$$P_{|\exp \lambda(\cdot)\rangle}|\exp(\mu(\cdot))\rangle = e^{-(\lambda,\mu-\lambda)}|\exp \lambda(\cdot)\rangle. \tag{V.3}$$

The decoherence functional $d(\mu,\nu)$ for two continuous-time histories is denoted by

$$d(\mu,\nu) = \mathrm{tr}_{\mathcal{V}_{\mathrm{cts}}\otimes \mathcal{V}_{\mathrm{cts}}}(P_{|\exp \mu(\cdot)\rangle} \otimes P_{|\exp(\nu(\cdot))\rangle} X), \tag{V.4}$$

where

$$X := \langle 0|\rho_{-\infty}|0\rangle (\mathcal{S}_{\mathrm{cts}}\mathcal{U})^{\dagger} \otimes (\mathcal{S}_{\mathrm{cts}}\mathcal{U}). \tag{V.5}$$

The operator $\mathcal{S}_{\mathrm{cts}}$ that appears in this expression for the $d(\mu,\nu)$ was defined as an approximation of the derivative operator, in the sense that

$$\mathcal{S}_{\mathrm{cts}}|\exp \nu(\cdot)\rangle = |\exp(\nu(\cdot) + \dot{\nu}(\cdot))\rangle, \tag{V.6}$$

while the dynamics was introduced by the operator $\mathcal{U}$, defined in such a way that the notion of time evolution is encoded,

$$e^{\langle \lambda,\dot{\lambda}\rangle} e^{(i/\hbar)H[\lambda]} = \mathrm{tr}_{\mathcal{V}_{\mathrm{cts}}}(\mathcal{S}_{\mathrm{cts}}\mathcal{U}\mathcal{P}_{|\exp \lambda(\cdot)\rangle}). \tag{V.7}$$

We expect V and H to play a similar role to that of $\mathcal{S}_{\mathrm{cts}}$ and $\mathcal{U}$, respectively, inside an expression for the decoherence functional. To demonstrate this, we will use the type of Fock space construction given in Eqs. (II.33)–(II.34). In particular, we use the property

$$\Gamma(A)|\exp \nu(\cdot)\rangle = |\exp(A\nu(\cdot))\rangle, \tag{V.8}$$

where A is an operator that acts on the elements $\nu(\cdot)$ of the base Hilbert space $\mathcal{H}$, while the operator $\Gamma(A)$, defined by Eq. (II.33), acts on the coherent states $|\exp \nu(\cdot)\rangle$ of the Fock space $e^{\mathcal{H}}$.

We notice that $\mathcal{U}$ is related to the unitary time evolution, Eq. (V.7), in a similar way to that of the Hamiltonian operator H,

$$e^{isH}|\exp \nu(\cdot)\rangle = \Gamma(e^{is\omega I})|\exp \nu(\cdot)\rangle = |\exp(e^{is\omega}\nu(\cdot))\rangle, \tag{V.9}$$

where I is the unit operator. We also notice that the action of the operator e^{isH} produces phase changes, as reflected on the right-hand side of Eq. (V.9) (which has been calculated for the special case of the simple harmonic oscillator). Furthermore, when the operator $\mathcal{S}_{\mathrm{cts}}$ acts on a coherent state, Eq. (V.6), it transforms it to another coherent state, which involves the addition to the defining function $\nu(\cdot)$ in a way that involves the time derivative of ν; and it is noteworthy that the Liouville operator V acts in a similar way:

$$e^{isV}|\exp \nu(\cdot)\rangle = \Gamma(e^{isD})|\exp \nu(\cdot)\rangle = |\exp(e^{isD}\nu(\cdot))\rangle, \tag{IV.10}$$

where

$$(e^{isD}\nu)(t)=\nu(t+s), \tag{V.11}$$

where $D:=-i(d/dt)$. The operator e^{isD} acts on the base Hilbert space, and corresponds to the operator e^{isV} under the Γ construction on the Fock space; that is, it acts on the vector $\nu(t)$ and transforms it to another one $\nu(t+s)$, which, for each time t, is translation by the time interval s.

This suggests that we define the operator $\mathcal{A}_s:=e^{isS}$, where $\mathcal{S}:=\int_{-\infty}^{+\infty}(p_t\dot{x}_t-H_t)dt$ is the action operator for the simple harmonic oscillator, which one expects to be related to the operator $\mathcal{S}_{\rm cts}\mathcal{U}$. For this reason, we write the matrix elements of both operators and compare them.

The general formula for the matrix elements of an arbitrary operator $\mathcal{T}$ with respect to the coherent states basis in the history space that was used in Ref. 6 is

$$\langle \exp \mu(\cdot)|\mathcal{T}|\exp \nu(\cdot)\rangle = e^{(\langle \mu,\delta/\delta\overline{\lambda}\rangle+\langle \delta/\delta\lambda,\nu\rangle)}\langle \exp \lambda(\cdot)|\mathcal{T}|\exp \lambda(\cdot)\rangle|_{\lambda=\overline{\lambda}=0}, \tag{V.12}$$

hence we need only compare the diagonal matrix elements of the two operators $\mathcal{S}_{\rm cts}\mathcal{U}$ and $\mathcal{A}_s$. Thus, we have

$$\langle \exp(\lambda(\cdot)|\mathcal{S}_{\rm cts}\mathcal{U}|\exp(\lambda(\cdot)\rangle = e^{\langle \lambda,\lambda+\dot{\lambda}\rangle}e^{(i/\hbar)H[\lambda]}, \tag{V.13}$$

where $H[\lambda]:=\int_{-\infty}^{\infty}H(\lambda(t))dt$ and $H(\lambda):=H(\lambda,\lambda)=\langle\lambda|H|\lambda\rangle/\langle\lambda|\lambda\rangle$; and

$$\langle \exp \lambda(\cdot)|\mathcal{A}_s|\exp \lambda(\cdot)\rangle = e^{\langle \lambda, e^{is(\omega I+D)}\lambda\rangle}, \tag{V.14}$$

with

$$(e^{is(\omega I+D)}\lambda)(t)=e^{is\omega}\lambda(t+s). \tag{V.15}$$

We can also write both of the above operators on the history space $\mathcal{F}(\mathcal{L}^2(\mathbf{R}))$ using their corresponding operators on the Hilbert space $\mathcal{L}^2(\mathbf{R})$. The Γ construction shows that

$$\mathcal{S}_{\rm cts}\mathcal{U}=\Gamma(1+i\sigma), \tag{V.16}$$

$$\mathcal{A}_s=\Gamma(e^{is\sigma})=e^{is\, d\Gamma(\sigma)}, \tag{V.17}$$

where $\sigma=\omega I+iD$, and I is the unit operator. As expressions of the same function σ, the operators $\mathcal{S}_{\rm cts}\mathcal{U}$ and $\mathcal{A}_s$ commute. However, we cannot readily compute their common spectrum because the operator $\mathcal{S}_{\rm cts}\mathcal{U}$ is not self-adjoint.

We might speculate that the value of the decoherence functional is maximized for a continuous-time projector that corresponds to a coarse graining around the classical path. Indeed, if we take such a generic projection operator P, we expect that it should commute with the operator $\mathcal{S}_{\rm cts}\mathcal{U}$. In this context, we noticed earlier that the projection operator that corresponds to a classical solution (f,h) commutes with the action operator

$$[\mathcal{S}_{\rm cts}\mathcal{U},P_{(f,h)}]=0. \tag{V.18}$$

Finally, this argument should be compared with the similar condition for classical histories:

$$\{S_h,F_C\}(\gamma_{\rm cl})=0. \tag{V.19}$$

VI. CONCLUSIONS

We have examined the example of the simple harmonic oscillator, in one dimension, within the History Projection Operator formulation of the consistent-histories scheme. We defined the action operator as the quantum analog of the classical Hamilton action functional and we have proved its existence by finding a representation on the $\mathcal{F}(\mathcal{L}^2(\mathbf{R}))$ space of the history algebra. We have shown that the action operator is the generator of two types of time transformations: translations in time from one Hilbert space $\mathcal{H}_t$, labeled by the time parameter t, to another Hilbert space with a different label t, and phase changes in time with respect to the time parameter s of the standard Heisenberg-time evolution that acts in each individual Hilbert space $\mathcal{H}_t$. We have expressed the action operator in terms of the Liouville and Hamiltonian operators—which are the generators of the two types of time transformation— and that correspond to the kinematics and the dynamics of the theory, respectively.

We have constructed continuous-time classical histories defined on the continuous Cartesian product of copies of the phase space and demonstrated an analogous expression to the classical Hamilton's equations.

Finally, we have shown that the action operator commutes with the defining operator of the decoherence functional, thus appearing in the expression for the dynamics of the theory, as would have been expected.

One of the major reasons for undertaking this study was to provide new tools for tackling the recalcitrant problem of constructing a manifestly covariant quantum field theory in the consistent histories formalism. Work on this problem is now in progress, with the expectation that the Hamiltonian and Liouville operators will play a central role in the proof of explicit Poincaré invariance of the theory.

ACKNOWLEDGMENTS

I would like to thank Chris Isham for the help and support during this work. I would also like to thank Charis Anastopoulos and Nikos Plakitsis. I gratefully acknowledge support from the NATO Research Fellowships department of the Greek Ministry of Economy.

[1] J. Hartle, ''Spacetime quantum mechanics and the quantum mechanics of spacetime,'' in *Proceedings on the 1992 Les Houches School, Gravitation and Quantisation*, 1992.
[2] M. Gell-Mann and J. Hartle, ''Quantum mechanics in the light of quantum cosmology,'' in *Proceedings of the 25th International Conference on High Energy Physics*, Singapore, 2–8 August 1990, edited by K. K. Phua and Y. Yamaguchi (World Scientific, Singapore, 1990).
[3] R. B. Griffiths, ''Consistent histories and the interpretation of quantum mechanics,'' J. Stat. Phys. **36**, 219–272 (1984).
[4] R. Omnés, ''Logical reformulation of quantum mechanics. I. Foundations,'' J. Stat. Phys. **53**, 893–932 (1988).
[5] C. Isham, N. Linden, K. Savvidou, and S. Schreckenberg, ''Continuous time and consistent histories,'' J. Math. Phys. **39**, 1818–1834 (1998).
[6] C. Isham and N. Linden, ''Continuous histories and the history group in generalised quantum theory,'' J. Math. Phys. **36**, 5392–5408 (1995).
[7] C. J. Isham, ''Quantum logic and the histories approach to quantum theory,'' J. Math. Phys. **23**, 2157 (1994).
[8] C. J. Isham, N. Linden, and S. Schreckenberg, ''The classification of decoherence functionals: an analogue of Gleason's theorem,'' J. Math. Phys. **35**, 6360 (1994).
[9] J. J. Halliwell, ''Notes on quantum cosmology lectures,'' M.Sc. course in quantum fields and fundamental forces, Imperial College, 1995.
[10] K. Savvidou, ''The continuous-time histories approach to quantum field theory: The Poincare symmetry,'' work in preparation.
[11] From a private communication of Professor Tulsi Dass with Professor Chris Isham.
[12] H. Araki, ''Hamiltonian formalism and the canonical commutation relations in quantum field theory,'' J. Math. Phys. **1**, 492–504 (1960).
[13] H. D. Zeh, *The Physical Basis of the Direction of Time* (Springer-Verlag, Berlin, 1992).
[14] C. Anastopoulos, ''On the selection of preferred consistent sets,'' Int. J. Theor. Phys. **37**, 2261 (1998).
[15] I. Kouletsis, ''A classical history theory: geometrodynamics and general field dynamics regained,'' gr-qc 9801019.

Entropic integrals of hyperspherical harmonics and spatial entropy of *D*-dimensional central potentials

R. J. Yáñez
Departamento de Matemática Aplicada and Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, 18071 Granada, Spain

W. Van Assche
Departement Wiskunde, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Heverlee, Belgium

R. González-Férez, and Jesús S. Dehesa
Departamento de Física Moderna and Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, 18071 Granada, Spain

(Received 24 March 1999; accepted for publication 22 July 1999)

The information entropy of a single particle in a quantum-mechanical *D*-dimensional central potential is separated in two parts. One depends only on the specific form of the potential (radial entropy) and the other depends on the angular distribution (spatial entropy). The latter is given by an entropic-like integral of the hyperspherical harmonics, which is expressed in terms of the entropy of the Gegenbauer polynomials. This entropy is expressed in terms of the values of the quadratic logarithmic potential of Gegenbauer polynomials $C_n^\lambda(t)$ at the zeros of these polynomials. Then this potential for integer λ is given as a finite expansion of Chebyshev polynomials of even order, whose coefficients are shown to be Wilson polynomials. [S0022-2488(99)00111-5]

I. INTRODUCTION

The Boltzmann–Shannon information entropy[1,2] of a many particle system cannot have a closed form, although tight rigorous bounds in terms of radial expectation values have been derived.[3] Even for single-particle systems with prototypic central potentials (e.g., Coulomb and harmonic oscillator) this quantity, which measures the spread or extent of the associated quantum-mechanical probability density, has been fully determined in an analytical way only recently for quantum-mechanical states located at the two extremes of the energy spectrum.[4–9] That is, for the ground state and the first few lowest-lying excited states (where the wave function has a very simple form) and for high-lying (Rydberg) excited states (where the physical entropies are controlled by the asymptotics of L^p-like norms of the classical orthogonal polynomials[4,6,7,8,10,11])

Let us consider the information entropy of a *D*-dimensional single-particle system in a central potential $V(r)$, defined by

$$S_\rho = -\int \rho(\mathbf{r}) \log \rho(\mathbf{r}) d\mathbf{r}, \tag{1}$$

where $\mathbf{r}=(x_1,\ldots,x_D)$ and $\rho(\mathbf{r})=|\Psi(\mathbf{r})|^2$. The wave function Ψ is given by the Schrödinger equation of the system which in atomic units is

$$[-\tfrac{1}{2}\nabla^2+V(r)]\Psi(\mathbf{r})=E\Psi(\mathbf{r}).$$

Here, the information entropy S_ρ is shown to decompose into two parts; the radial entropy, which depends on the specific form of the potential $V(r)$, and the angular or ''spatial'' entropy $S(Y,D)$, which is an integral involving hyperspherical harmonics $Y_{l,\{\mu\}}(\Omega_D)$, as it is shown later on in this section. The rest of the paper is devoted to the determination of the spatial entropy, what is carried out in three steps. First, in Sec. II, this entity is explicitly expressed in terms of the entropy of the orthonormal Gegenbauer or ultraspherical polynomials $\{\hat{C}_n^\lambda(x);\ n=0,1,...\}$. Then, in Sec. III, this Gegenbauer entropy $E_n^\lambda \equiv E_n(\hat{C}_n^\lambda)$ is given in terms of the logarithmic potential of the squares of Gegenbauer polynomials; and, finally, this quadratic logarithmic potential with integer parameter λ is expressed in the form of a finite Chebyshev expansion, whose coefficients turn out to be Wilson polynomials.

To show the entropy decomposition we start writing the kinetic energy operator $-\frac{1}{2}\boldsymbol{\nabla}$ in hyperspherical coordinates,[12] this equation is transformed in

$$\left[-\frac{1}{2}\left(\frac{d^2}{dr^2}+\frac{D-1}{r}\frac{d}{dr}-\frac{\Lambda^2}{r^2}\right)+V(r)\right]\Psi(\mathbf{r})=E\Psi(\mathbf{r}).$$

The solution of this equation takes the form

$$\Psi(\mathbf{r})=R_{nl}(r)Y_{l,\{\mu\}}(\Omega_D), \tag{2}$$

where $R_{nl}(r)$ is the radial eigenfunction, i.e., solution of the radial Schrödinger equation

$$\left[-\frac{1}{2}\left(\frac{d^2}{dr^2}+\frac{D-1}{r}\frac{d}{dr}+\frac{l(l+D-2)}{r^2}\right)+V(r)\right]R_{nl}(r)=E_{nl}R_{nl}(r) \tag{3}$$

and $Y_{l,\{\mu\}}(\Omega_D)$ are the eigenfunctions of the nonradial part of the Hamiltonian, Λ^2, i.e.,[12]

$$\Lambda^2 Y_{l,\{\mu\}}(\Omega_D)=l(l+D-2)Y_{l,\{\mu\}}(\Omega_D), \tag{4}$$

where the orbital quantum number l and the magnetic quantum numbers $\{\mu\}$, are integers verifying

$$l=\mu_1\geqslant\mu_2\geqslant\cdots\geqslant\mu_{D-1},$$

with $\mu_{D-1}=|m|$. The functions $Y_{l,\{\mu\}}(\Omega_D)$ are the hyperspherical harmonics, given by[12–14]

$$Y_{l,\{\mu\}}(\Omega_D)=N_{l,\{\mu\}}e^{im\varphi}\prod_{j=1}^{D-2} C_{\mu_j-\mu_{j+1}}^{\alpha_j+\mu_{j+1}}(\cos\theta_j)(\sin\theta_j)^{\mu_{j+1}}, \tag{5a}$$

with the normalizing constant

$$N_{l,\{\mu\}}^2=\frac{1}{2\pi}\prod_{j=1}^{D-2}\frac{(\alpha_j+\mu_j)(\mu_j-\mu_{j+1})![\Gamma(\alpha_j+\mu_{j+1})]^2}{\pi 2^{1-2\alpha_j-2\mu_{j+1}}\Gamma(2\alpha_j+\mu_j+\mu_{j+1})}=\frac{1}{2\pi}\prod_{j=1}^{D-2}N_{l,\{\mu\}}^{(j)}. \tag{5b}$$

Here $2\alpha_j=D-j-1$, $C_n^\lambda(t)$ is a Gegenbauer polynomial of degree n and parameter λ, and the angles $\theta_1,\theta_2,\ldots,\theta_{D-2}$, ϕ are defined by

$$x_1=r\sin\theta_1\sin\theta_2\cdots\sin\theta_{D-2}\cos\varphi,$$

$$x_2=r\sin\theta_1\sin\theta_2\cdots\sin\theta_{D-2}\sin\varphi,$$

$$x_3=r\sin\theta_1\sin\theta_2\cdots\cos\theta_{D-2},$$

$$\vdots\ \vdots$$

$$x_{D-1}=r\sin\theta_1\cos\theta_2,$$

$$x_D=r\cos\theta_1,$$

with $0\leqslant\theta_j\leqslant\pi$, $j=1,...,D-2$, and $0\leqslant\varphi<2\pi$.

As a particular case, in two dimensions, we have

$$Y_m(\Omega_2)=\frac{e^{im\varphi}}{2\pi} \tag{6a}$$

with $|m|=0,1...$, and in the three-dimensional case ($D=3$),

$$Y_{l,m}(\Omega_3)=\left(\frac{(l+\frac{1}{2})(l-|m|)![\Gamma(|m|+\frac{1}{2})]^2}{2^{1-2|m|}\pi^2(l+|m|)!}\right)^{1/2} e^{im\varphi}(\sin\theta)^{|m|}C_{l-|m|}^{|m|+1/2}(\cos\theta), \tag{6b}$$

with $l=0,1,...y$ $|m|=0,1,...,l$. The expression (6b) can be written in terms of Legendre functions using the relationship between these functions and the Gegenbauer polynomials[15]

$$C_{l-m}^{m+1/2}(t)=(-1)^m\frac{(1-t^2)^{-m/2}m!2^m}{(2m)!}P_l^m(t). \tag{7}$$

Then, taking into account (2) and that the volume element in a D-dimensional space is

$$d\mathbf{r}=r^{D-1}drd\Omega_D, \quad d\Omega_D=\left(\prod_{j=1}^{D-2}\sin^{2\alpha_j}\theta_j d\theta_j\right)d\phi,$$

we have from (1) that the Boltzmann–Shannon entropy of a D-dimensional particle in a central potential $V(r)$ can be decomposed into two parts,

$$S_\rho=S(R;D)+S(Y;D), \tag{8}$$

where

$$S(R;D)=-\int r^{D-1}R_{nl}^2(r)\log R_{nl}^2(r)dr \tag{9}$$

is the contribution from the radial part of the wave function to the entropy (to be called radial entropy heretoforth), and

$$S(Y;D)=-\int |Y_{l,\{\mu\}}(\Omega_D)|^2\log|Y_{l,\{\mu\}}(\Omega_D)|^2 d\Omega_D \tag{10}$$

is the contribution from the angular part of the density, to be called spatial entropy heretoforth.

The radial entropy $S(R;D)$ cannot be calculated without knowing the specific form of the potential $V(r)$. Moreover, even when this form is known such as for Coulomb and harmonic oscillator potentials, its closed form has not yet been derived; despite it, its asymptotic behavior has been rigorously found as said before.

On the contrary, the spatial entropy $S(Y;D)$, also called entropy of the hyperspherical harmonic $Y_{l,\{\mu\}}(\Omega_D)$, does not depend on the potential so that it can be evaluated independently of it. This is done in the following sections.

II. SPATIAL ENTROPY AND GEGENBAUER ENTROPY

Taking into account expression (5a) of the hyperspherical harmonics, we find that the spatial entropy of a D-dimensional particle is

$$S(Y;D) = -\int |Y_{l,\{\mu\}}(\Omega_D)|^2 \log|Y_{l,\{\mu\}}(\Omega_D)|^2 d\Omega_D$$

$$= -\log N^2_{l,\{\mu\}} + \sum_{j=1}^{D-2} N^{(j)}_{l,\{\mu\}} (S(C^{\alpha_j+\mu_j+1}_{\mu_j-\mu_{j+1}}) - \mu_{j+1} I^{(j)}_{\{\mu\}}), \tag{11a}$$

where

$$S(C^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}}) \equiv -\int_{-1}^{+1} (1-t^2)^{\alpha_j+\mu_{j+1}-(1/2)} [C^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}}(t)]^2 \log[C^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}}(t)]^2 dt, \tag{11b}$$

and

$$I^{(j)}_{\{\mu\}} = \int_{-1}^{+1} (1-t^2)^{\alpha_j+\mu_{j+1}-(1/2)} [C^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}}(t)]^2 \log(1-t^2) dt, \tag{11c}$$

where $S(C^\alpha_n)$ is the entropy of the Gegenbauer polynomials of degree n and parameter α. The integral $I^{(j)}_{\{\mu\}}$ can be calculated by use of the following result:[16]

$$\int_{-1}^{1} (1-x^2)^{\lambda-1/2} [C^\lambda_n(x)]^2 \log(1-x^2) dx$$

$$= \frac{\Gamma(2\lambda+n) 2^{1-2\lambda}}{n! \Gamma^2(\lambda)} \frac{\pi}{\lambda+n} \left[2\psi(2\lambda+n) - 2\psi(\lambda+n) - 2\log 2 - \frac{1}{\lambda+n} \right],$$

where $\psi(x)$ is the digamma or psi function, i.e., $\psi(x) = \Gamma'(x)/\Gamma(x)$. This value can also be obtained by differentiation of the Gegenbauer orthogonality relation

$$\int_{-1}^{+1} (1-x^2)^{\lambda-1/2} [C^\lambda_n(x)/(\lambda)_n]^2 dx = \frac{\pi 2^{1-2\lambda} \Gamma(n+2\lambda)}{n! [\Gamma(n+\lambda+1)]^2}$$

with respect to λ and noting that $(\partial/\partial\lambda)[C^\lambda_n(x)/(\lambda)_n]$ is a polynomial of degree $n-1$.

Taking into account this value and substituting it into (11a), one has

$$S(Y;D) = \log 2\pi - \sum_{j=1}^{D-2} \log N^{(j)}_{l,\{\mu\}} + \sum_{j=1}^{D-2} N^{(j)}_{l,\{\mu\}} S(C^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}})$$

$$-\sum_{j=1}^{D-2} \mu_{j+1} \left[2\psi(2\alpha_j+\mu_j+\mu_{j+1}) - 2\psi(\alpha_j+\mu_j) - 2\log 2 - \frac{1}{\alpha_j+\mu_j} \right]. \tag{12}$$

Finally, since the orthonormal Gegenbauer polynomial $\hat{C}^\lambda_n(x) = h_n^{-1} C^\lambda_n(x)$, with

$$h_n^2 = \pi \frac{2^{1-2\lambda} \Gamma(n+2\lambda)}{[\Gamma(\lambda)]^2 (n+\lambda) n!},$$

one has from (11b),

$$S(Y;D) = \log 2\pi + \sum_{j=1}^{D-2} E(\hat{C}^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}})$$

$$-\sum_{j=1}^{D-2} \mu_{j+1} \left[2\psi(2\alpha_j+\mu_j+\mu_{j+1}) - 2\psi(\alpha_j+\mu_j) - 2\log 2 - \frac{1}{\alpha_j+\mu_j} \right], \tag{13a}$$

where

$$E(\hat{C}^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}})=-\int_{-1}^{+1}(1-t^2)^{\alpha_j+\mu_{j+1}-(1/2)}[\hat{C}^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}}(t)]^2\log[\hat{C}^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}}(t)]^2dt \tag{13b}$$

is the entropy of orthonormal Gegenbauer polynomials of degree $\mu_j-\mu_{j+1}$ and parameter $\alpha_j+\mu_{j+1}$, or simply, Gegenbauer entropy.

So, it is observed that the calculation of the angular entropy of a particle under an arbitrary central potential gets reduced to the computation of the entropy of Gegenbauer polynomials. In particular, we have

(i) For $D=2$, the entropy $S(Y;2)$ is

$$S(Y;2)=\log 2\pi, \tag{14}$$

which does not depend on the magnetic quantum number m.

(ii) For $D=3$, Eq. (12) reduces to

$$\begin{aligned}S(Y;3)=&-\int|Y_{lm}(\Omega_3)|^2\log|Y_{lm}(\Omega_3)|^2d\Omega_3\\ &=\log 2\pi-\log\left[\frac{(l+\frac{1}{2})(l-|m|)![\Gamma(|m|+\frac{1}{2})]^2}{\pi 2^{-2|m|}(l+|m|)!}\right]+\frac{(l+\frac{1}{2})(l-|m|)![\Gamma(|m|+\frac{1}{2})]^2}{\pi 2^{-2|m|}(l+|m|)!}S(C^{|m|+(1/2)}_{l-|m|})\\ &-|m|\left[2\psi(l+|m|+1)-2\psi\left(l+\frac{1}{2}\right)-2\log 2-\frac{1}{l+\frac{1}{2}}\right].\end{aligned} \tag{15}$$

Also, from Eq. (13b) one has

$$S(Y;3)=\log 2\pi+E(\hat{C}^{|m|+(1/2)}_{l-|m|})-|m|\left[2\psi(l+|m|+1)-2\psi\left(l+\frac{1}{2}\right)-2\log 2-\frac{1}{l+\frac{1}{2}}\right]. \tag{16}$$

It is interesting to note that, using the relations (7) and (15), we have the alternative expression

$$S(Y;3)=-\log\left(\frac{(2l+1)(l-m)!}{4\pi(l+m)!}\right)+\left(\frac{(2l+1)(l-m)!}{2(l+m)!}\right)S(P_l^m), \tag{17a}$$

where

$$S(P_l^m)=-\int_{-1}^{+1}[P_l^m(t)]^2\log[P_l^m(t)]^2dt \tag{17b}$$

denotes the entropy of the Legendre function $P_l^m(t)$.

III. ENTROPY AND LOGARITHMIC POTENTIAL OF GEGENBAUER POLYNOMIALS

Aptekarev *et al.*[4] have recently shown that the asymptotic behavior of the entropy of the Gegenbauer polynomials $E_n^\lambda=E(\hat{C}_n^\lambda)$ defined by

$$E_n^\lambda:=-\int_{-1}^{+1}\omega_\lambda(t)[\hat{C}_n^\lambda(t)]^2\log[\hat{C}_n^\lambda(t)]^2dt \tag{18}$$

is given by

$$E_n^\lambda\approx E_\infty^\lambda=-1+\log 2-I(\rho_0,\omega_\lambda)=-1+(1-2\lambda)\log 2+\log\pi+o(1),$$

where the relative entropy between the equilibrium measure

$$\rho_0(x)=(1/\pi)(1/\sqrt{1-x^2})$$

and the weight function $\omega_\lambda(x)=(1-x^2)^{\lambda-(1/2)}$ is

$$I(\rho_0,\omega_\lambda)=\int_{-1}^{+1}\rho_0(t)\log\frac{\rho_0(t)}{\omega_\lambda(t)}\,dt=2\lambda\log 2-\log\pi.$$

Also, they observed[4] that the rate of convergence of E_n^λ is

$$E_n^\lambda=E_\infty^\lambda+\frac{k_\lambda}{n}+o\left(\frac{1}{n}\right) \tag{19}$$

and studied the numerical dependence of k_λ on λ. Moreover, they conjectured a similar rate of convergence for the entropy of general orthogonal polynomials with a constant k depending only on the weight function of the polynomials.

Moreover, the consideration of the probability weight function

$$\rho_\lambda(x)=\frac{\lambda\Gamma^2(\lambda)}{2^{1-2\lambda}\Gamma(2\lambda)\pi}(1-x^2)^{\lambda-(1/2)}$$

and its corresponding orthonormal polynomials

$$p_n^\lambda(x)=\left[\frac{\Gamma(2\lambda)(n+\lambda)n!}{\lambda\Gamma(n+2\lambda)}\right]^{1/2}\qquad C_n^\lambda(x)=\frac{2^{1-2\lambda}\Gamma(2\lambda)\pi}{\lambda\Gamma^2(\lambda)}\hat{C}_n^\lambda(x)$$

has led to the authors[5] to obtain the following asymptotic expression:

$$\begin{aligned}\mathcal{E}_n^\lambda&=-\int_{-1}^{+1}\rho_\lambda(x)[p_n^\lambda(x)]^2\log[p_n^\lambda(x)]^2dx\\&=-\frac{n}{n+\lambda}-2\log\frac{\Gamma(n+\lambda)}{\Gamma(\lambda)n!}[1+o(1)]-\log\frac{\Gamma(2\lambda)(n+\lambda)n!}{\lambda\Gamma(n+2\lambda)}.\end{aligned}$$

Here we will use a two-step method which allow us to calculate the exact value of the Gegenbauer entropy $\mathcal{E}_n^\lambda$ for any n. First, we use the expression[7]

$$\mathcal{E}_n^\lambda=2\log\gamma_n-2\sum_{j=1}^{n}V_n^\lambda(x_{j,n};\nu_n), \tag{20}$$

where $x_{j,n}$ $(j=1,2,...,n)$ and γ_n are the zeros and the leading coefficient of the orthonormal Gegenbauer polynomial $p_n^\lambda(x)$, respectively, ν_n is the probability measure given by

$$d\nu_n(t)=(1-t^2)^{\lambda-(1/2)}[p_n^\lambda(t)]^2,$$

and the logarithmic potential $\hat{V}_n^\lambda(x,\nu_n)$ of the measure ν_n is given by

$$\hat{V}_n^\lambda=\hat{V}_n^\lambda(z,\nu_n)=\int_{-1}^{+1}(1-t^2)^{\lambda-(1/2)}[p_n^\lambda(t)]^2\log\frac{1}{|z-t|}\,dt,$$

which is related to the logarithmic potential of the Gegenbauer polynomial C_n^λ defined as

$$V_n^\lambda(z,\nu_n)=\int_{-1}^{+1}(1-t^2)^{\lambda-1/2}[C_n^\lambda(t)]^2\log\frac{1}{|z-t|}\,dt$$

by the relationship

$$\hat{V}_n^{\lambda}(z)=\frac{n+\lambda}{n}\frac{\Gamma(2\lambda)}{\Gamma(n+2\lambda)}\frac{n!\Gamma(\lambda+1)}{\sqrt{\pi}\Gamma(\lambda+\frac{1}{2})}V_n^{\lambda}(z).$$

Taking into account that

$$\gamma_n=\frac{2^n\Gamma(\lambda+n+1)\Gamma(2\lambda)}{\Gamma(\lambda+1)\Gamma(n+2\lambda)},$$

we remark from (20) that the Gegenbauer entropy $\mathcal{E}_n^{\lambda}$ can be computed once we will have a precise knowledge of the logarithmic potential V_n^{λ} at the zeros of C_n^{λ}. The use of the recursive approach recently developed by the authors[7] to determine this logarithmic potential does not allow to find an explicit closed expression for the Gegenbauer entropy $\mathcal{E}_n^{\lambda}$ in a simple way save for the first two integers values $\lambda=0,1,$[5,7]

$$\mathcal{E}_n^0=\log 2-1,\quad \mathcal{E}_n^1=-\frac{n}{n+1}.$$

To go further we need a procedure to calculate the logarithmic potential other than the aforementioned recursive approach.[7] This is done in the following by expanding it in terms of Chebyshev polynomials of the first kind $T_k(x)$; this is the second step of our method.

If P_n $(n=0,1,2,...)$ are orthogonal polynomials with respect to the measure μ on $[-1,1]$, then we can use that[7]

$$\log|z-t|=\log\frac{1}{2}-2\sum_{k=1}^{\infty}\frac{1}{k}T_k(z)T_k(t)$$

to find

$$-\int_{-1}^{1}\log|x-t|P_n^2(x)d\mu(x)=\log 2\int_{-1}^{1}P_n^2(x)d\mu(x)+2\sum_{k=1}^{\infty}\frac{T_k(t)}{k}\int_{-1}^{1}T_k(x)P_n^2(x)d\mu(x),$$

which gives an expansion in terms of Chebyshev polynomials of the first kind, with expansion coefficients in terms of

$$\int_{-1}^{1}T_k(x)P_n^2(x)d\mu(x),$$

and these integrals still need to be computed explicitly. Let us apply this idea for Gegenbauer polynomials $P_n(x)=C_n^{\lambda}(x)$. The use of Clausen's formula[17]

$$[C_n^{\lambda}(x)]^2=\left[\frac{(2\lambda)_n}{n!}\right]^2{}_3F_2\left(\begin{matrix}-n,n+2\lambda,\lambda\\2\lambda,\lambda+\frac{1}{2}\end{matrix};1-x^2\right),$$

allows us to find

$$\int_{-1}^{1}T_k(x)[C_n^{\lambda}(x)]^2(1-x^2)^{\lambda-(1/2)}dx$$
$$=\left[\frac{(2\lambda)_n}{n!}\right]^2\sum_{j=0}^{n}\frac{(-n)_j(n+2\lambda)_j(\lambda)_j}{(2\lambda)_j(\lambda+\frac{1}{2})_j j!}\int_{-1}^{1}(1-x^2)^{j+\lambda-(1/2)}T_k(x)dx.$$

If λ is an integer, then $[C_n^\lambda(x)]^2(1-x^2)^\lambda$ is a polynomial of degree $2n+2\lambda$, and by the orthogonality of the Chebyshev polynomials the integral will vanish for $k>2n+2\lambda$. Therefore we only need to consider $k\leqslant 2n+2\lambda$. By symmetry the integral for k odd vanishes and for even indices we have[18]

$$\int_{-1}^{1}(1-x^2)^{\lambda+j-(1/2)}T_{2k}(x)dx=(-1)^k\frac{\pi\Gamma(2j+2\lambda+1)}{2^{2j+2\lambda}\Gamma(j+k+\lambda+1)\Gamma(j+\lambda+1-k)}.$$

Using Legendre's duplication formula for the gamma function gives $\Gamma(2j+2\lambda+1)=2^{2j+2\lambda}\pi^{-1/2}\Gamma(j+\lambda+\frac{1}{2})\Gamma(j+\lambda+1)$ so that

$$\int_{-1}^{1}[C_n^\lambda(x)]^2(1-x^2)^{\lambda-(1/2)}T_{2k}(x)dx$$
$$=\left[\frac{(2\lambda)_n}{n!}\right]^2\frac{(-1)^k\sqrt{\pi}\Gamma(\lambda+\frac{1}{2})\Gamma(\lambda+1)}{\Gamma(k+\lambda+1)\Gamma(\lambda+1-k)}\,{}_4F_3\left(\begin{matrix}-n,n+2\lambda,\lambda,\lambda+1\\2\lambda,\lambda+1+k,\lambda+1-k\end{matrix};1\right). \quad (21)$$

Hence, we have

$$V_n^\lambda(t)=\left[\frac{(2\lambda)_n}{n!}\right]^2\sqrt{\pi}\left[\log 2\,\frac{\Gamma(\lambda+\frac{1}{2})}{\Gamma(\lambda+1)}\,{}_3F_2\left(\begin{matrix}-n,n+2\lambda,\lambda\\2\lambda,\lambda+1\end{matrix};1\right)+\Gamma\left(\lambda+\frac{1}{2}\right)\Gamma(\lambda+1)\right.$$
$$\left.\times\sum_{k=1}^{n+\lambda}(-1)^kT_{2k}(t)\frac{1}{k\Gamma(k+\lambda+1)\Gamma(\lambda+1-k)}\,{}_4F_3\left(\begin{matrix}-n,n+2\lambda,\lambda,\lambda+1\\2\lambda,\lambda+1+k,\lambda+1-k\end{matrix};1\right)\right]. \quad (22)$$

Observe that this shows that on the interval $[-1,1]$ the logarithmic potential V_n^λ is a polynomial of degree $2n+2\lambda$ when λ is an integer, which is indeed compatible with the explicit formulas

$$\hat{V}_n^{(0)}(t)=\log 2+\frac{T_{2n}(t)}{2n};\quad n\geqslant 1$$

$$\hat{V}_n^{1}(t)=\log 2-\frac{T_{2n+2}(t)}{2n+2};\quad n\geqslant 0$$

$$\hat{V}_n^{2}(t)=\log 2+\frac{n+1}{n+3}\,\frac{T_{2n+4}(t)}{2n+4}-\frac{U_{2n+2}(t)}{(n+1)(n+3)}+\frac{1}{(n+1)(n+3)},$$

already found by us.[7]

The ${}_3F_2$ series is terminating and balanced, hence Saalschütz' theorem[19] gives

$${}_3F_2\left(\begin{matrix}-n,n+2\lambda,\lambda\\2\lambda,\lambda+1\end{matrix};1\right)=\frac{\lambda\Gamma(2\lambda)n!}{(n+\lambda)\Gamma(2\lambda+n)}.$$

Also the ${}_4F_3$ series is balanced, but its evaluation is more complicated. A balanced and terminating ${}_4F_3$ at unit argument is however a Wilson polynomial or a Racah polynomial taken at some point. The Wilson polynomial W_n is given by

$$\frac{W_n(x^2;a,b,c,d)}{(a+b)_n(a+c)_n(a+d)_n}={}_4F_3\left(\begin{matrix}-n,n+a+b+c+d-1,a+ix,a-ix\\a+b,a+c,a+d\end{matrix};1\right)$$

(see, e.g., Koekoek and Swarttouw[20] or Wilson[21,22]). Identifying parameters shows that

$$ {}_4F_3\left(\begin{matrix} -n,n+2\lambda,\lambda,\lambda+1 \\ 2\lambda,\lambda+1+k,\lambda+1-k \end{matrix};1\right) = \frac{W_n(-\frac{1}{4};\lambda+\frac{1}{2},\lambda-\frac{1}{2},k+\frac{1}{2},-k+\frac{1}{2})}{(2\lambda)_n(\lambda+1+k)_n(\lambda+1-k)_n}. $$

Using this and the normalization

$$ \frac{n+\lambda}{\lambda}\frac{\Gamma(2\lambda)}{\Gamma(n+2\lambda)}\frac{n!\Gamma(\lambda+1)}{\sqrt{\pi}\Gamma(\lambda+\frac{1}{2})}V_n^\lambda(t)=\hat{V}_n^\lambda(t), $$

this gives

$$ \hat{V}_n^\lambda(t)=\log 2+\frac{2(n+\lambda)\Gamma^2(\lambda+1)}{\lambda n!}\sum_{k=1}^{n+\lambda}(-1)^k\frac{T_{2k}(t)}{2k}\frac{W_n(-\frac{1}{4};\lambda+\frac{1}{2},\lambda-\frac{1}{2},k+\frac{1}{2},-k+\frac{1}{2})}{\Gamma(\lambda+k+n+1)\Gamma(\lambda-k+n+1)}. \tag{23} $$

A few comments about Wilson polynomials, which are pertinent to illustrate the effectiveness of Eq. (23), are contained in the Appendix. Expression (23) for the logarithmic potential with integer λ generalizes our expressions for $\lambda=0$, 1, and 2 already mentioned.

For completeness and comparison, it is worth to mention here that recently Buyarov *et al.*[23] have developed an alternative, fully different method to work out a formula for the logarithmic potential (with integer λ) evaluated at the zeros of the polynomial $C_n^\lambda(x)$. This formula, which does not use Wilson polynomials, together with Eq. (20) has allowed them to obtain an explicit expression for the entropy $\mathcal{E}_n^\lambda$ for λ, $n\in\mathbb{N}$, $\lambda\geqslant 2$. In doing so, they corrected and extended the value obtained previously by one of them[24] for $\mathcal{E}_n^2$, which is given by

$$ \begin{aligned} \mathcal{E}_n^2&=\log\frac{n+3}{3(n+1)}-\frac{n(n^2+2n-1)}{(n+1)(n+2)(n+3)}-\frac{2}{\sqrt{(n+1)^3(n+3)^3}}\frac{T'''_{n+2}(z_n)}{T''_{n+2}(z_n)} \\ &=\log\frac{n+3}{3(n+1)}-\frac{n^3-5n^2-29n-27}{(n+1)(n+2)(n+3)}-\frac{1}{n+2}\left(\frac{n+3}{n+1}\right)^{n+2}, \end{aligned} $$

where $z_n=n+2/\sqrt{(n+1)(n+3)}$. Also, they were able to obtain analytically[23] that the rate of convergence of $\mathcal{E}_n^\lambda$, $\lambda\in\mathbb{N}$, $\lambda\geqslant 2$, is given by

$$ \mathcal{E}_n^\lambda=\mathcal{E}_\infty^\lambda+\frac{\gamma_\lambda}{n}+O(n^{-2}), $$

where the dominant term is

$$ \mathcal{E}_\infty^\lambda=1+\log\frac{\Gamma(2\lambda)}{\Gamma(\lambda)\Gamma(\lambda+1)} $$

and the second term is controlled by the constant

$$ \gamma_\lambda=-2\lambda^2+\lambda-2\sum_{j=1}^{\lambda-1}\sqrt{\xi_j}\,\frac{R}{S'}\,\frac{J_{\lambda+\frac{1}{2}}}{J_{\lambda-\frac{1}{2}}}(\xi_j). $$

Here, $J_\lambda(x)$ denotes the Bessel function of order λ and ξ_j, $j=1,...,\lambda-1$ are the zeros of the polynomial $S\equiv S_{2\lambda-2}(x)$. The polynomials $\{S_{-1}=0,\ S_0=1,...,S_{2\lambda-2}\}$ are generated by the recurrence relation

$$ S_{j+1}(x)=(2\lambda-2j-3)S_j(x)-xS_{j-1}(x), $$

and the polynomial $R(x)$ is given by

$$R(x)=\sum_{j=0}^{2\lambda-2}(-1)^{j}S_{j-1}(x)S_{2\lambda-j-3}(x).$$

Finally, let us mention that there exists another alternative to be explored which replaces the second step of our method for the following procedure. Instead of expanding the quadratic Gegenbauer potential V_n^λ in Chebyshev polynomials, one can express it as a sum of linear Gegenbauer potentials by the linearization formula of $[C_n^\lambda(t)]^2$ in terms of $C_n^\lambda(t)$; this can be done by use of the Dougall's linearization formula of products of two $C_n^\lambda(x)$ as a sum involving $C_k^\lambda(x)$. Then, the linear Gegenbauer potential, which is a definite integral of $(1-z^2)^{\lambda-1/2}Q_n^\lambda(z)$, can be evaluated by use of the explicit hypergeometric representation for the Gegenbauer function of the second kind $Q_n^\lambda(z)$.

IV. SUMMARY AND CONCLUSIONS

The spread of the quantum-mechanical probability density $\rho(\mathbf{r})$ for a D-dimensional single-particle system moving in a potential $V(\mathbf{r})$ is controlled by the Boltzmann–Shannon information entropy S_ρ. For central potentials, this quantity can be decomposed in two parts: the radial entropy and the angular or spatial entropy. The former is given by Eq. (9) and it depends on the analytic form of the potential. The latter, denoted by $S(Y;D)$ and associated to the physical form of the system, is described in Eq. (10) by the entropic integral or just the ''entropy'' of the known hyperspherical harmonics $Y_{l,\{\mu\}}(\Omega_D)$. Moreover, the entropy $S(Y;D)$ is explicitly expressed by Eq. (13b) in terms of the entropic integral of the orthonormal Gegenbauer polynomial $\hat{C}_n^\lambda$, where n and λ depends on D and the quantum numbers $l,\{\mu\}$. This mathematical notion $E_n^\lambda \equiv E(\hat{C}_n^\lambda)$, defined by Eq. (18) and called Gegenbauer entropy, was naturally encountered in the study of the position and momentum information entropies of two specific physical systems: the harmonic oscillator and the hydrogen atom.[5,7] Its asymptotic ($n\to\infty$) behavior has been thoroughly investigated specifically[4,23] and in the more general context of entropic integrals for general orthogonal polynomials.[10]

Finally, this paper and Ref. 23 describe methods to calculate the Gegenbauer entropy for arbitrary integer values of λ and n. However we have not been able yet to extend our results to real or, at least, half-integer values of the parameter λ, what is necessary [see, e.g., Eq. (16)] for a complete description of the entropy of an arbitrary (hyper)spherical harmonic, which describes the spatial entropy of single-particle systems in arbitrary quantum-mechanical states of any (D-dimensional) central potential. This would require a generalization of our methods or the design of a new mathematical strategy to evaluate the Gegenbauer entropy E_n^λ and/or $\mathcal{E}_n^\lambda$ for real λ.

ACKNOWLEDGMENTS

This work has been partially supported by the European project INTAS-93-219-Ext as well as by the Spanish project DGES PB 95-1205 and the research grant FQM-207 of the Junta de Andalucia. W. V. A. is a Research Director of the Belgian Fund for Scientific Research (FWO).

APPENDIX: SOME NOTES ABOUT WILSON POLYNOMIALS

Wilson polynomials are invariant under a permutation of the parameters. They are orthogonal on $[0,\infty)$ when all the parameters a,b,c,d are such that $\Re(a,b,c,d)>0$ and non-real parameters occur in conjugate pairs. In our case, one of the parameters $-k+\frac{1}{2}$ may be negative (also $\lambda-\frac{1}{2}$ when $\lambda<1/2$). If some of the parameters are negative, then Wilson polynomials can still be orthogonal on $[0,\infty)\cup E$, where E contains mass points of the orthogonality measure on $(-\infty,0]$. The first mass point in this case turns out to be $-1/4$, which is exactly the point where we evaluate the Wilson polynomial.

We will use a generating function for the Wilson polynomials

$$ {}_2F_1\left(\begin{matrix} a+ix, b+ix \\ a+b \end{matrix}; t\right) {}_2F_1\left(\begin{matrix} c-ix, d-ix \\ c+d \end{matrix}; t\right) = \sum_{n=0}^{\infty} \frac{W_n(x^2;a,b,c,d)}{(a+b)^n(c+d)^n} \frac{t^n}{n!}, $$

(see Refs. 20 or 21). In our case the parameters are $a=\lambda+\frac{1}{2}$, $b=\lambda-\frac{1}{2}$, $c=k+\frac{1}{2}$, $d=-k+\frac{1}{2}$, and the variable is $x^2=-\frac{1}{4}$. The generating function then becomes

$$ {}_2F_1\left(\begin{matrix} \lambda, \lambda-1 \\ 2\lambda \end{matrix}; t\right) {}_2F_1\left(\begin{matrix} 1-k, k+1 \\ 1 \end{matrix}; t\right) = \sum_{n=0}^{\infty} W_n\left(-\frac{1}{4};\lambda+\frac{1}{2},\lambda-\frac{1}{2},k+\frac{1}{2},-k+\frac{1}{2}\right) \frac{t^n}{(2\lambda)_n(n!)^2}. \tag{A1} $$

Observe that one of the hypergeometric functions on the left is a polynomial of degree $k-1$. From this generating function we can now obtain the coefficient of t^n as the convolution of the coefficients of the two hypergeometric series on the left, giving

$$ W_n\left(-\frac{1}{4};\lambda+\frac{1}{2},\lambda-\frac{1}{2},k+\frac{1}{2},-k+\frac{1}{2}\right) = (2\lambda)_n(n!)^2 \sum_{j=0}^{k-1} \frac{(1-k)_j(k+1)_j}{(j!)^2} \frac{(\lambda)_{n-j}(\lambda-1)_{n-j}}{(2\lambda)_{n-j}(n-j)!}, \quad n \geqslant k-1. \tag{A2} $$

[1] L. Boltzmann, *Vorlesungen über Gastheorie* (Barth, Leipzig, 1896).

[2] C. E. Shannon, ''A mathematical theory of information,'' Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948).

[3] J. C. Angulo and J. S. Dehesa, ''Tight rigorous bounds to atomic information entropies,'' J. Chem. Phys. **97**, 6485–6495 (1992).

[4] A. I. Aptekarev, J. S. Dehesa, and R. J. Yáñez, ''Spatial entropy of central potentials and strong asymptotics of orthogonal polynomials,'' J. Math. Phys. **35**, 4423–4428 (1994).

[5] R. J. Yáñez, W. van Assche, and J. S. Dehesa, ''Position and momentum information entropies of the *D*-dimensional harmonic oscillator and hydrogen atom,'' Phys. Rev. A **50**, 3065–3079 (1994).

[6] W. van Assche, R. J. Yáñez, and J. S. Dehesa, ''Entropy of orthogonal polynomials with Freud weights and information entropies of the harmonic oscillator potential,'' J. Math. Phys. **36**, 1–13 (1995).

[7] J. S. Dehesa, W. van Assche, and R. J. Yáñez, ''Information entropy of classical orthogonal polynomials and their application to the harmonic oscillator and Coulomb potentials,'' Meth. Appl. Anal. **4**, 91–110 (1997).

[8] J. S. Dehesa, R. J. Yáñez, A. I. Aptekarev, and V. S. Buyarov, ''Strong asymptotics of Laguerre polynomials and information entropies of 2D-harmonic oscillator and 1D-Coulomb potentials,'' J. Math. Phys. **39**, 3050–3060 (1999).

[9] J. Sánchez-Ruiz, ''Logarithmic potential of Hermite polynomials and information entropies of the harmonic oscillator eigenstates,'' J. Math. Phys. **38**, 5031–5043 (1997).

[10] A. I. Aptekarev, V. S. Buyarov, and J. S. Dehesa, ''Asymptotic behavior of the L^p-norms and the entropy for general orthogonal polynomials,'' R. Akad. Nauk. Mat. Sb. **185**, 3–30 (1994) (in Russian); Russ. Acad. Sci. Sb. Math. **82**, 373–395 (1995).

[11] A. I. Aptekarev, V. S. Buyarov, W. van Assche, and J. S. Dehesa, ''Asymptotics for entropy integrals of orthogonal polynomials,'' Dokl. Akad. Nauk **346**, 439–441 (1996) (in Russian); Dokl. Math. **53**, 47–49 (1996).

[12] J. Avery, *Hyperspherical Harmonics: Applications in Quantum Theory* (Kluwer Academic, Dordrecht, 1989).

[13] J. D. Louck, ''Generalized orbital angular momentum and the *n*-fold degenerate quantum-mechanical oscillator. Part III. Radial integrals,'' J. Mol. Spectrosc. **4**, 334–341 (1960).

[14] A. F. Nikiforov, S. K. Suslov, and V. B. Uvarov, *Classical Orthogonal Polynomials of a Discrete Variable* (Springer-Verlag, Berlin, 1991).

[15] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1970).

[16] W. Van Assche, R. J. Yáñez, R. González-Férez, and J. S. Dehesa, ''Functionals of Gegenbauer polynomials and *D*-dimensional hydrogenic momentum expectation values,'' J. Math. Phys. (submitted).

[17] G. Gasper, ''Using computer algebraic systems to derive formulas involving orthogonal polynomials and other special functions,'' in *Orthogonal Polynomials: Theory and Practice*, NATO ASI series C 294, edited by P. Nevai (Kluwer, Dordrecht, 1990), pp. 163–179.

[18] I. S. Gradsteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, San Diego, 1980).

[19] L. J. Slater, *Generalized Hypergeometric Functions* (Cambridge University Press, Cambridge, 1966).

[20] R. Koekoek and R. Swarttouw, ''The Askey-scheme of hypergeometric orthogonal polynomials and its *q*-analoq,'' Delft University of Technology, Faculty of Technical Mathematics and Informatics, Report No. 98-17, 1998. See also the online version at http://aw.twi.tudelft.nl/~koekoek/askey

[21] J. A. Wilson, ''Some hypergeometric orthogonal polynomials,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **11**, 690–701 (1980).

[22] J. A. Wilson, ''Asymptotics for the ${}_4F_3$ polynomials,'' J. Approx. Theory **66**, 58–71 (1991).
[23] V. S. Buyarov, P. López Artés, A. Martínez-Finkelshtein, and W. Van Assche, ''Information entropy of Gegenbauer polynomials'' (preprint).
[24] V. S. Buyarov, ''On information entropy of Gegenbauer polynomials,'' Vestn. Mosk. Univ., Ser. 1: Mat., Mekh. **6**, 8–11 (1997) (in Russian).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Bethe Ansatz and thermodynamic limit of affine quantum group invariant extensions of the $t-J$ model

J. Ambjørn[a)]
Niels Bohr Institute, Blegdamsvej 17, Copenhagen, Denmark

A. Avakyan[b)] and T. Hakobyan[c)]
Yerevan Physics Institute, Br. Alikhanian st.2, 375036, Yerevan, Armenia

A. Sedrakyan[d)]
Niels Bohr Institute, Blegdamsvej 17, Copenhagen, Denmark

(Received 20 October 1998; accepted for publication 26 May 1999)

We have constructed a one-dimensional exactly solvable model, which is based on the $t-J$ model of strongly correlated electrons, but which has additional quantum group symmetry, ensuring the degeneration of states. We use Bethe Ansatz technique to investigate this model. The thermodynamic limit of the model is considered and equations for different density functions written down. These equations demonstrate that the additional color degrees of freedom of the model behave as in a gauge theory, namely, an arbitrary distribution of color indices over particles leave invariant the energy of the ground state and the excitations. The *S*-matrix of the model is shown to be the product of the ordinary $t-J$ model *S*-matrix and the unity matrix in the color space.
[S0022-2488(99)02411-1]

I. INTRODUCTION

Since the discovery of high-T_c cuprate superconductivity the one-dimensional physics of strongly correlated electrons has been in focus in many publications.[1] The Hubbard[2] and $t-J$ models[3] are such examples, motivated in part by the fact that high-T_c compounds display antiferromagnetism in the absence of doping. The $t-J$ model was proposed by Zhang and Rice[3] and describes strongly correlated electrons with anntiferromagnetic exchange interaction.

The interest in one-dimensional physics grew after Andersons claim[4] that two-dimensional systems may have features in common with one-dimensional systems. In addition it should be mentioned that powerful methods in 1D such as bosonization, 2D conformal field theory, and in particular the Bethe Ansatz technique allow the detailed study of such systems.

The $t-J$ model may be used as well for the heavy fermion system.[5,6]

At the supersymmetric point $J=2t$ the $t-J$ model becomes exactly integrable[5,7–12] because the Hamiltonian can be represented as a graded permutation in a superalgebra of two fermions and one boson.

In Refs. 13–15 we developed the technique for construction of a family of spin chain Hamiltonians and their fermionic representations, which have the same energy levels as some basic model (XXZ, Hubbard, $t-J$ or others) but with huge degeneracy as a result of an affine quantum symmetry added to the basic model. We called this procedure an affinization of the model.

The first example of this type of model was constructed in Ref. 16, giving rise to the Hubbard Hamiltonian in the infinite repulsion limit.

[a)]Electronic mail: ambjorn@nbivms.nbi.dk
[b)]Electronic mail: avakyan@lx2.yerphi.am
[c)]Electronic mail: hakob@lx2.yerphi.am
[d)]Electronic mail: sedrak@nbivms.nbi.dk; Permanent address: Yerevan Physics Institute, Br. Alikhanian st. 2, 375036, Yerevan, Armenia.

In Ref. 14 we have fermionized the simplest examples of this newly defined family of models and have shown that it leads to extensions of one-band Hubbard Hamiltonians. The η-pairing mechanism introduced by Yang[17,18] was found in one of examples in addition to other exactly solvable Hubbard models with the superconducting ground state.[19–23] The essential property of this extensions is the fact that, besides the ordinary electron hopping and Hubbard interaction terms, they contain also bond-charge interaction, pair-hopping, and nearest-neighbor interaction terms. In Ref. 15 the SU(N) affinization of the $t-J$ model was carried out, giving rise to a model where the spin–spin coupling term consists of interaction between the total spins (i.e., the sum of the spins of all band) at nearest-neighbor sites. The presence of the affine symmetry, which ensures the degeneracy of levels exponentially proportional to the length (area) of the space, might lead to a new type of string theory.

In this article we define an extension of $t-J$ model such that an affine quantum group symmetry is present, and we use the Bethe Ansatz technique to solve the model. We find the *S*-matrix excitations on empty background, the ground state and construct the thermodynamic limit of the model. As one might expect, the *S*-matrix of the excitations on empty background consists of the *S*-matrix of the ordinary $t-J$ model multiplied by the unity matrix in the additional space of ''colors.'' Therefore the Bethe equations are not different from ones for the $t-J$ model, but the rapidities presented in equations correspond to particles with the arbitrary colors. The degeneracy of the corresponding n-particle states come from arbitrary partitions of the color indices over particles. The same is true for the ground state. The situation is exactly as in gauge theories if we distinguish the states which differ by pure gauge transformations. All these results are presented in Sec. V.

The thermodynamic limit of the model with corresponding equations are represented in Sec. VI, where we also shown that the *S*-matrix of our model in an arbitrary background is equal to ordinary $t-J$ model *S*-matrix multiplied by the Kronecker symbols over the additional color indices.

II. QUANTUM GROUP INVARIANT HAMILTONIANS FOR REDUCIBLE REPRESENTATIONS

Let $\mathbf{V}=\oplus_{i=1}^{N}\mathbf{V}_{\lambda_i}$ be a direct sum of finite dimensional irreducible representations $\mathbf{V}_{\lambda_i}$ of quantum group $U_q g$.[24–26] We denote by $\mathbf{V}(x_1,\ldots,x_N)$ the representation with spectral parameters x_i of the corresponding affine quantum group $U_q\hat{g}$,[26]

$$\mathbf{V}(x_1,\ldots,x_N)=\bigoplus_{i=1}^{M}\mathbf{N}_{\lambda_i}\hat{\otimes}\mathbf{V}_{\lambda_i}(x_i), \tag{1}$$

where all the $\mathbf{V}_{\lambda_i}(x_i)$ are M nonequivalent irreps and $\mathbf{N}_{\lambda_i}\simeq\mathbf{C}^{N_i}$ have dimensions equal to the multiplicity of $\mathbf{V}_{\lambda_i}(x_i)$ in $\mathbf{V}(x_1,\ldots,x_N)$. Note that $\Sigma_{i=1}^{M}N_i=N$. The $\hat{}$ over the tensor product signifies that $U_q\hat{g}$ does not act on $\mathbf{N}_{\lambda_i}\hat{\otimes}\mathbf{V}_{\lambda_i}(x_i)$ by means of comultiplication Δ but instead acts as $\mathrm{id}\otimes g$.

In Ref. 14 the general matrix form of the intertwining operator

$$H(x_1,\ldots,x_N):$$

$$\mathbf{V}(x_1,\ldots,x_N)\otimes\mathbf{V}(x_1,\ldots,x_N)\rightarrow\mathbf{V}(x_1,\ldots,x_N)\otimes\mathbf{V}(x_1,\ldots,x_N), \tag{2}$$

$$[H(x_1,\ldots,x_N),\ \Delta(a)]=0,\quad \forall a\in U_q\hat{g}$$

had been written using the projection operators

$$X_b^a=|a\rangle\langle b|. \tag{3}$$

Here the vectors $|a\rangle$ span the space $\mathbf{V}$. In accordance with the decomposition (1) we will use the double index $a=(n_i,a_i)$, $i=1,...,M$ where the first index $n_i=1,...,N_i$ characterizes the multiplicity of $\mathbf{V}_{\lambda_i}$ and the second one $a_i=1,...,\dim \mathbf{V}_{\lambda_i}$ is the vector index of $\mathbf{V}_{\lambda_i}$. Then the intertwining operator (2) is

$$H(A,B)=\sum_{i,j=1}^{M}\left(\sum_{n_i,n_j,m_i,m_j} A_{ij}{}^{m_i m_j}_{n_i n_j}\sum_{a_i,a_j} X^{(n_i,a_i)}_{(m_i,a_i)}\otimes X^{(n_j,a_j)}_{(m_j,a_j)}\right.$$
$$\left.+\sum_{n_i,n_j,m_i,m_j} B_{ij}{}^{m_i m_j}_{n_i n_j}\sum_{a_i,a_j,a_i',a_j'} R_{ij}{}^{a_i a_j}_{a_i' a_j'}(x_i/x_j) X^{(n_j,a_j')}_{(m_i,a_i)}\otimes X^{(n_i,a_i')}_{(m_j,a_j)}\right), \tag{4}$$

where the R-matrix

$$R_{\mathbf{V}_{\lambda_i}\otimes \mathbf{V}_{\lambda_j}}(x_i/x_j)|a_i\rangle\otimes|a_j\rangle=\sum_{a_i',a_j'} R_{ij}{}^{a_i a_j}_{a_i' a_j'}(x_i/x_j)|a_i'\rangle\otimes|a_j'\rangle$$

is the intertwining operator between two actions of affine quantum group $U_q\hat{g}$ on $\mathbf{V}_{\lambda_i}\otimes\mathbf{V}_{\lambda_j}$, which are induced correspondingly by comultiplication Δ and opposite comultiplication $\bar{\Delta}$,[24–26]

$$R_{\mathbf{V}_{\lambda_i}\otimes \mathbf{V}_{\lambda_j}}(x_i/x_j)\Delta(g)=\bar{\Delta}(g)R_{\mathbf{V}_{\lambda_i}\otimes \mathbf{V}_{\lambda_j}}(x_i/x_j).$$

A_{ij} and B_{ij}, $B_{ii}=0$ in (4) are arbitrary matrices. In general, $H(A,B)$ depends on deformation parameter q of quantum group, which is included in the R-matrix. Note that $R_{\mathbf{V}_{\lambda_i}\otimes \mathbf{V}_{\lambda_j}}(x_i/x_j)$ does not depend on q and is identity only if λ_i or λ_j are trivial one-dimensional representations. So, in the special case, when the only nontrivial R-matrixes in (4) are between representations, one of which is trivial representation, the expression of $H(A,B)$ does not depend on q. Then $H(A,B)$ commutes with the quantum group action for all values of deformation parameter. In the following we consider only this case.

Following Refs. 13 and 16 we can from the operator H construct the following Hamiltonian acting on $\mathbf{W}=\mathbf{V}^{\otimes L}$ (here and in the following we omit the dependence on x_i)

$$\mathcal{H}=\sum_{i=1}^{L-1} H_{ii+1}, \tag{5}$$

where the indices i and $i+1$ denote the sites where H acts nontrivially. By the construction, $\mathcal{H}$ is quantum group invariant,

$$[\mathcal{H},\Delta^{L-1}(g)]=0, \quad \forall g\in U_q\hat{g}.$$

Let us define the projection operators $\mathcal{Q}^i$ on $\mathbf{V}$ for each class of equivalent irreps (λ_i,x_i), $i=1,...,M$,

$$\mathcal{Q}^i v_j=\delta_{ij}v_j, \quad \forall v_j\in V_{\lambda_j}(x_j),$$

$$\sum_{i=1}^{M}\mathcal{Q}^i=\mathrm{id}, \quad (\mathcal{Q}^i)^2=\mathcal{Q}^i.$$

Their action on $\mathbf{W}$ is given by

$$\mathcal{Q}^i=\sum_{l=1}^{L}\mathcal{Q}^i_l.$$

It is easy to see that these projections commute with Hamiltonian $\mathcal{H}$ and quantum group $U_q\hat{g}$,

$$[\mathcal{Q}^i,\mathcal{H}]=0,\ \ [\mathcal{Q}^i,U_q\hat{g}]=0. \tag{6}$$

Denoted by $\mathbf{W}_{p_1\cdots p_M}$ the subspace of $\mathbf{W}$ with values p_i of $\mathcal{Q}^i$ on it. Then we have the decomposition

$$\mathbf{W}=\bigoplus_{\substack{p_1,\ldots,p_M\\ p_1+\cdots+p_M=L}}\mathbf{W}_{p_1\cdots p_M}. \tag{7}$$

Let $\mathbf{V}^0$ be the linear space, spanned by the highest weight vectors in V,

$$\mathbf{V}^0:=\oplus_{i=1}^N v^0_{\lambda_i},$$

where $v^0_{\lambda_i}\in\mathbf{V}_{\lambda_i}$ is a highest weight vector. We also define $\mathbf{W}^0:=\mathbf{V}^{0\otimes L}$. The space $\mathbf{W}^0$ is $\mathcal{H}$-invariant. For general q the action of $U_q\hat{g}$ on $\mathbf{W}^0$ generate the whole space $\mathbf{W}$. Indeed, the $U_q\hat{g}$-action on each state of type $v^0_{\lambda_{i_1}}\otimes\cdots\otimes v^0_{\lambda_{i_L}}$ generates the space $\mathbf{V}_{\lambda_{i_1}}\otimes\cdots\otimes\mathbf{V}_{\lambda_{i_L}}$, because the tensor product of finite dimensional irreducible representations of an affine quantum group is irreducible.[27]

Consider now the subspace $\mathbf{W}^0_{p_1\cdots p_M}=\mathbf{W}^0\cap\mathbf{W}_{p_1\cdots p_M}$. According to (7) we have the decomposition

$$\mathbf{W}^0=\bigoplus_{\substack{p_1,\ldots,p_M\\ p_1+\cdots+p_M=L}}\mathbf{W}^0_{p_1\cdots p_M}. \tag{8}$$

Note that

$$d_{p_1\cdots p_M}:=\dim\mathbf{W}^0_{p_1\cdots p_M}=\binom{L}{p_1\cdots p_M}N_1^{p_1}\cdots N_M^{p_M}.$$

Let us define by $\mathcal{H}_0$ the restriction of $\mathcal{H}$ on $\mathbf{W}_0$: $\mathcal{H}_0:=\mathcal{H}|_{\mathbf{W}_0}$. It follows from (6) that Hamiltonians $\mathcal{H}$ and $\mathcal{H}_0$ have block diagonal form with respect to the decompositions (7) and (8), respectively. Every eigenvector $w^0_{\alpha_{p_1\cdots p_M}}\in\mathbf{W}^0_{p_1\cdots p_M}$ with energy value $E_{\alpha_{p_1\cdots p_M}}$ gives rise to an irreducible $U_q\hat{g}$-multiplet $\mathbf{W}_{\alpha_{p_1\cdots p_M}}$ of dimension

$$\dim\mathbf{W}_{\alpha_{p_1\cdots p_M}}=\prod_{k=1}^{M}(\dim\mathbf{V}_{\lambda_k})^{p_k}. \tag{9}$$

On $\mathbf{W}_{\alpha_{p_1\cdots p_M}}$ the Hamiltonian $\mathcal{H}$ is diagonal with eigenvalue $E_{\alpha_{p_1\cdots p_M}}$. In particular, in the case when all $\mathbf{V}_{\lambda_i}$ are equivalent, the degeneracy levels are the same for all $E_{\alpha_{p_1\cdots p_M}}$ and are equal to $(\dim\mathbf{V}_\lambda)^L$.

Now, let us assume we know the energy spectrum $E_{\alpha_{p_1\cdots p_M}}$ for $\mathcal{H}_0$. Then the statistical sum is given by

$$Z_{\mathcal{H}_0}(\beta)=\sum_{\substack{p_1,\ldots,p_M\\ p_1+\cdots+p_M=L}}\ \sum_{\alpha_{p_1\cdots p_M}=1}^{d_{p_1\cdots p_M}}\exp(-\beta E_{\alpha_{p_1\cdots p_M}}), \tag{10}$$

and it follows that the statistical sum of $\mathcal{H}$ has the following form:

$$Z_{\mathcal{H}}(\beta)=\sum_{\substack{p_1,\ldots,p_M\\ p_1+\cdots+p_M=L}}\prod_{k=1}^{M}(\dim \mathbf{V}_{\lambda_k})^{p_k}\sum_{\alpha_{p_1\cdots p_M}=1}^{d_{p_1\cdots p_M}}\exp(-\beta E_{\alpha_{p_1\cdots p_M}}). \tag{11}$$

So, if the underlying Hamiltonian $\mathcal{H}_0$ is integrable and its eigenvectors and eigenvalues can be found, then we know these for $\mathcal{H}$ too. Acting with the quantum group on all eigenvectors of an energy level of $\mathcal{H}_0$ one obtains the whole eigenspace of $\mathcal{H}$ for this level.

III. MULTIBAND $t-J$ MODEL WITH VANISHING SPIN–SPIN COUPLING $J=0$

Let us consider here the quantum group $U_q\widehat{sl}_2$. We choose $\mathbf{V}=\mathbf{V}_0\oplus\mathbf{V}_j$ for decomposition (1) i.e., we take a direct sum of the trivial spin-0 and the $2j+1$-dimensional spin-j representation of $U_q sl_2$. The R-matrix in the second term in (4) does not depend on q and spectral parameters x_i and coincides with the identity, as it was mentioned above. So, using (4) and (5), we obtain the following Hamiltonian:

$$\mathcal{H}(t,V_1,V_2)=\sum_{i=1}^{L-1}\left[-t\sum_{p=1}^{2j+1}(X_{i0}^{\,p}X_{i+1p}^{\,0}+X_{i+10}^{\,p}X_{ip}^{\,0})+V_1X_{i0}^{\,0}X_{i+10}^{\,0}+V_2\sum_{p,p'=1}^{2j+1}X_{ip}^{\,p}X_{i+1p'}^{\,p'}\right]. \tag{12}$$

The Hamiltonian $\mathcal{H}=\Sigma_i H_{ii+1}$ was constructed from the operator $H=H_{ii+1}$, where H can be written in the matrix form

$$H=\begin{pmatrix} V_1 & 0 & 0 & 0\\ 0 & 0 & -t\cdot\mathrm{id} & 0\\ 0 & -t\cdot\mathrm{id} & 0 & 0\\ 0 & 0 & 0 & V_2\cdot\mathrm{id}\end{pmatrix}. \tag{13}$$

The projection on the highest weight space coincides with the constructing block of the XXZ Hamiltonian in an external magnetic field. This implies that the restriction of (12) to the space $\mathbf{W}^0$ is

$$\mathcal{H}_0(t,W_1,W_2)=\mathcal{H}_{XXZ}(t,\Delta,B)=-\frac{t}{2}\sum_{i=1}^{L-1}\left(\sigma_i^x\sigma_{i+1}^x+\sigma_i^y\sigma_{i+1}^y+\Delta\sigma_i^z\sigma_{i+1}^z+\frac{B}{2}\sigma_i^z\right), \tag{14}$$

where

$$\Delta=-\frac{V_1+V_2}{2t},\qquad B=\frac{2}{t}(V_1-V_2). \tag{15}$$

For the special case $V_1+V_2=0$ $\mathcal{H}_0$ gives rise to the free fermionic (or equivalently XY) Hamiltonian ($\Delta=0$).

The projection operators X_b^a are expressed through the fermionic creation-annihilation operators as follows:

$$\begin{gathered} X_{i0}^{\,p}=\mathcal{P}c_{i,p}^{+},\qquad X_{ip}^{\,0}=c_{i,p}\mathcal{P},\\ X_{ip}^{\,p}=n_{i,p}\mathcal{P}=\mathcal{P}n_{i,p},\qquad X_{i0}^{\,0}=(1-n_i)\mathcal{P}=\mathcal{P}(1-n_i). \end{gathered} \tag{16}$$

Here we introduced the projection operator which forbids double occupation on all sites

$$\mathcal{P}=\prod_{i=1}^{L}\mathcal{P}_i,\qquad \mathcal{P}_i=\prod_{p\neq p'}(1-n_{i,p}n_{i,p'})$$

and the total particle number $n_i=\Sigma_p n_{i,p}$ at site i.

After the substitution of the fermionic representation (16) into (12) we obtain

$$\mathcal{H}(t,V_1,V_2)=\mathcal{P}\sum_{i=1}^{L-1}\left[-t\sum_{p=1}^{2j+1}(c_{i,p}^{+}c_{i+1,p}+c_{i+1,p}^{+}c_{i,p})+Vn_in_{i+1}-V_1(n_i+n_{i+1})+V_1\right]\mathcal{P}, \tag{17}$$

where $V=V_1+V_2$. The chemical potential term $-V_1\Sigma_{i=1}^{L-1}(n_i+n_{i+1})$ commutes with $\mathcal{H}$ and can be omitted. So, up to unessential boundary and constant terms (17) is a multicomponent $t-J$ model with vanishing spin–spin coupling $(J=0)$,

$$\mathcal{H}(t,V)=\sum_{i=1}^{L-1}\left[-t\sum_{p=1}^{2j+1}(c_{i,p}^{+}c_{i+1,p}+c_{i+1,p}^{+}c_{i,p})+Vn_in_{i+1}\right]+\sum_{i=1}^{L}\sum_{\substack{p\neq p',\\ p,p'=1}}^{2j+1}U_{p,p'}n_{i,p}n_{i,p'}, \tag{18}$$

where the infinite Hubbard interaction amplitude $U_{p,p'}=+\infty$ between p and p' bands excludes sites with double and more occupations. It follows from the above considerations that this model has energy levels which coincide with the levels of XXZ Heisenberg model, but that the degeneracy of the levels is different.

For vanishing density–density interaction $V=0$ the Hamiltonian (18) describes the infinite repulsion limit of the multiband Hubbard model. Thus, according to (15) $\Delta=0$ and it has the energy levels of the free fermionic model.

IV. MULTIBAND EXTENSION OF THE $t-J$ MODEL WITH AFFINE QUANTUM GROUP SYMMETRY

In this section we consider Hamiltonians which have the same energy levels as the $t-J$ model but have affine quantum group symmetry. Because each site in the ordinary $t-J$ model has three states, one should for this purpose take the direct sum of three spaces. Let

$$\mathbf{V}=\mathbf{V}_0\oplus\mathbf{V}_j\oplus\mathbf{V}_j. \tag{19}$$

Recall that the $t-J$ model is given by

$$\mathcal{H}_{t-J}(t,J,V)-\mathcal{P}\sum_{i=1}^{L-1}\left[-t\sum_{\sigma=\pm\frac{1}{2}}(c_{i,\sigma}^{+}c_{i+1,\sigma}+c_{i+1,\sigma}^{+}c_{i,\sigma})+J\mathbf{S}_i\mathbf{S}_{i+1}+Vn_in_{i+1}\right]\mathcal{P}, \tag{20}$$

where c_σ^+, c_σ are creation–annihilation operators of spin-$\frac{1}{2}$ fermion, $\mathbf{S}=\Sigma_{\sigma,\sigma'}c_\sigma^+\sigma_{\sigma\sigma'}c_{\sigma'}$ is the fermionic spin operator and $\mathcal{P}=\Pi_{i=1}^{L}(1-n_{i,\uparrow}n_{i,\downarrow})$ forbids double occupation of sites.

We rewrite it in terms of Hubbard operators X_b^a, where $a,b=0,\pm\frac{1}{2}$,

$$\mathcal{H}(t,J,V)=\sum_{i=1}^{L-1}\left[\sum_{\sigma=\pm\frac{1}{2}}\left(-t(X_{i0}^{\ \sigma}X_{i+1\sigma}^{\ 0}+X_{i+10}^{\ \sigma}X_{i\sigma}^{\ 0})+\frac{1}{2}J\cdot X_{i-\sigma}^{\ \sigma}X_{i+1\sigma}^{\ -\sigma}\right)\right.$$
$$\left.+\sum_{\sigma,\sigma'=\pm\frac{1}{2}}(\sigma\sigma' J+V)X_{i\sigma}^{\ \sigma}X_{i+1\sigma'}^{\ \sigma'}\right]. \tag{21}$$

Let us now look at the general expression (4) of intertwining operators H_{ij} acting on the space (19). For convenience we make index change in the following way. The two spin-j representations we use are denoted by $\sigma=\pm\frac{1}{2}$. The intrinsic index in each $V_j^{(\sigma)}$ is denoted by $k,k=1,...,2j+1$. So, instead of (n_i,a_i) in (4) we have (σ,k), if i corresponds to the spin-j multiplet. Because the spin-0 singlet is one dimensional and single, we just use for it the index 0. The nonequivalent irreps in

(19) are $V_j^{(\sigma)}$ and V_0 and, as mentioned above, the R-matrix for two such representations is the identity. After performing the first sum in (4) over nonequivalent multiplets we obtain

$$H(A,a,b_1,b_2)=\sum_{\sigma_1,\sigma_2,\sigma_1',\sigma_2'}\left(A_{\sigma_1\sigma_1'}^{\sigma_2\sigma_2'}\sum_{k,k'}X_{(\sigma_2,k)}^{(\sigma_1,k)}\otimes X_{(\sigma_2',k')}^{(\sigma_1',k')}\right)+a\cdot X_0^0\otimes X_0^0$$
$$+\sum_{k,\sigma}(b_1\cdot X_0^{(\sigma,k)}\otimes X_{(\sigma,k)}^0+b_2\cdot X_{(\sigma,k)}^0\otimes X_0^{(\sigma,k)}). \tag{22}$$

To implement the restriction $H_0(A,a,b_1,b_2)$ of this operator on the highest weight space one just should eliminate the sum over k,k' and put $k=k'=0$. Comparing (22) and (21) it follows that the expressions coincide if one chooses

$$a=0,\quad b_1=b_2=-t,\quad A_{\sigma-\sigma}^{-\sigma\sigma}=J/2,\quad A_{\sigma\sigma}^{\sigma'\sigma'}=(\sigma\sigma')\cdot J+V,$$

and choose the other values of $A_{\sigma_1\sigma_2}^{\sigma_1'\sigma_2'}$ equal to zero.

So, the Hamiltonian $\mathcal{H}(A,a,b_1,b_2)$ corresponding to (22) with these values of parameters gives rise to a $t-J$ model (20) on the highest weight space. According to the previous considerations it will have the same energy levels as the $t-J$ model, but with different degeneracy. Recall that for $J=2t$ the $t-J$ model is ''supersymmetric'' and integrable.

We express the Hubbard operators in terms of multiband fermionic creation–annihilation operators as follows:

$$X_{i0}^{(\sigma,k)}=\mathcal{P}c_{i,\sigma}^{k+},\quad X_{i(\sigma,k)}^{0}=c_{i,\sigma}^{k}\mathcal{P},$$
$$X_{i(-\sigma,k)}^{(\sigma,k)}=c_{i,\sigma}^{k+}c_{i,-\sigma}^{k}\mathcal{P}=\mathcal{P}c_{i,\sigma}^{k+}c_{i,-\sigma}^{k},\quad X_{i(\sigma,k)}^{(\sigma,k)}=n_{i,\sigma}^{k}\mathcal{P}=\mathcal{P}n_{i,\sigma}^{k}. \tag{23}$$

Here as before we used the projection operator, which forbids double occupation on all sites

$$\mathcal{P}=\prod_{i=1}^{L}\mathcal{P}_i,\quad \mathcal{P}_i=\prod_{(\sigma,k)\neq(\sigma',k')}(1-n_{i,\sigma}^{k}n_{i,\sigma'}^{k'}).$$

Now, we can write down the Hamiltonian (5) in terms of multiband fermions, substituting (23) into (22). We obtain in this way the multiband generalization of (20),

$$\mathcal{H}(t,J,V)=\mathcal{P}\sum_{i=1}^{L-1}\left[-t\sum_{k=1}^{2j+1}\sum_{\sigma=\pm\frac{1}{2}}(c_{i,\sigma}^{k+}c_{i+1,\sigma}^{k}+c_{i+1,\sigma}^{k+}c_{i,\sigma}^{k})+J\mathbf{S}_i\mathbf{S}_{i+1}+Vn_in_{i+1}\right]\mathcal{P}. \tag{24}$$

Here k is the band index, and $\mathbf{S}=\Sigma_k\mathbf{S}^k$, $n=\Sigma_k n^k$ are total spin and total particle number operators. It is easy to see that we have conservation of the particle number operators $\Sigma_{i,\sigma}^{k}=n^k$ for the all colors k.

V. BETHE ANSATZ FOR THE $t-J$ MODEL WITH AFFINE QUANTUM GROUP SYMMETRY

The goal of this section is to apply the Bethe Ansatz technique to our model and derive the corresponding Bethe equations for the excitations.

After making some trivial Pauli matrix calculations one can represent Hamiltonian (24) as

$$\mathcal{H}(t,J,V)=\mathcal{P}\sum_{i=1}^{L-1}\left[-t\sum_{k=1}^{2j+1}\sum_{\sigma=\pm\frac{1}{2}}(c_{i,\sigma}^{k+}c_{i+1,\sigma}^{k}+c_{i+1,\sigma}^{k+}c_{i,\sigma}^{k})\right.$$

$$\left.+2J\sum_{k,k',\sigma\neq\tau}c_{i,\sigma}^{k+}c_{i,\tau}^{k}c_{i+1,\tau}^{k'+}c_{i+1,\sigma}^{k'}+(V-J)n_in_{i+1}\right]\mathcal{P}. \tag{25}$$

Due to the conservation of the particle number operator n^k, and according to the coordinate Bethe Ansatz we look for eigenvectors of (25), corresponding to N fermions of $2j+1$ bands in the following form:

$$|\Psi\rangle=\sum_{k_1\cdots k_N=1}^{2j+1}\sum_{x_1\sigma_1}\cdots\sum_{x_N\sigma_N}\psi^{k_1\cdots k_N}(x_1\sigma_1,\ldots,x_N\sigma_N)c_{x_1\sigma_1}^{k_1+}\cdots c_{x_N\sigma_N}^{k_N+}|0\rangle, \tag{26}$$

where $|0\rangle$ is the empty vacuum state.

The eigenvalue equation $H|\Psi\rangle=E|\Psi\rangle$ in the one particle sector

$$-t(\Psi^k(x-1,\sigma)+\Psi^k(x,\sigma))=E\Psi^k(x,\sigma) \tag{27}$$

gives us

$$E=-2t\cos p \tag{28}$$

after substituting of the plane wave function with momentum p into (27).

The eigenvalue equations in the two particle sector allows us to fix the energy of the state as a sum of two one particle energies, as well as the two two particle scattering matrix $S_{\sigma_1\sigma_2,k_1k_2}^{\sigma_1'\sigma_2',k_1'k_2'}(p_1,p_2)$. We choose the antisymmetric wave function $\Psi^{k_1k_2}(x_1\sigma_1,x_2\sigma_2)$ as

$$\psi^{k_1,k_2}(x_1\sigma_1,x_2\sigma_2)=A^{k_1k_2}(p_1\sigma_1,p_2\sigma_2)e^{i(p_1x_1+p_2x_2)}-A^{k_2k_1}(p_2\sigma_1,p_1\sigma_2)e^{i(p_2x_1+p_1x_2)} \tag{29}$$

for $x_2\leqslant x_1$ and

$$\psi^{k_1,k_2}(x_1\sigma_1,x_2\sigma_2)=A^{k_2k_1}(p_2\sigma_2,p_1\sigma_1)e^{i(p_1x_1+p_2x_2)}-A^{k_1k_2}(p_1\sigma_2,p_2\sigma_1)e^{i(p_2x_1+p_1x_2)} \tag{30}$$

for $x_1\leqslant x_2$.

The continuity condition at $x_1\approx x_2$ should be imposed,

$$A^{k_1k_2}(p_1\sigma_1,p_2\sigma_2)-A^{k_1k_2}(p_2\sigma_1,p_1\sigma_2)=A^{k_2k_1}(p_2\sigma_2,p_1\sigma_1)-A^{k_2k_1}(p_1\sigma_2,p_2\sigma_1). \tag{31}$$

Use of the Hamiltonian (24) and the most general form (26) of the eigenfunctions ψ, the eigenvalue equations can be written as

$$\begin{aligned}-t[\psi^{k_1k_2}&(x_1+1\sigma_1,x_2\sigma_2)(1-\delta_{x_1+1,x_2})+\psi^{k_1k_2}(x_1-1\sigma_1,x_2\sigma_2)(1-\delta_{x_1,x_2+1})\\&+\psi^{k_1k_2}(x_1\sigma_1,x_2+1\sigma_2)(1-\delta_{x_1,x_2-1})+t+\psi^{k_1k_2}(x_1\sigma_1,x_2-1\sigma_2)(1-\delta_{x_1-1,x_2})]\\&+2J\delta_{|x_1,x_2|,1}\psi^{k_1k_2}(x_1\sigma_2,x_2\sigma_1)++(V-J)\delta_{|x_1-x_2|,1}\psi^{k_1k_2}(x_1\sigma_1,x_2\sigma_2)\\&=E\psi^{k_1k_2}(x_1\sigma_1,x_2\sigma_2).\end{aligned} \tag{32}$$

The terms $(1-\delta)$ appeared near the hopping terms as a result of projective operator $\mathcal{P}$, preventing double occupancy of the sites.

Two different cases can be considered:

(i) $|x_1-x_2|>1$. In this case eigenvalue Eq. (32) is reduced to

$$-t[\psi^{k_1k_2}(x_1+1\sigma_1,x_2\sigma_2)+\psi^{k_1k_2}(x_1-1\sigma_1,x_2\sigma_2)+\psi^{k_1k_2}(x_1\sigma_1,x_2+1\sigma_2)$$
$$+\psi^{k_1k_2}(x_1\sigma_1,x_2-1\sigma_2)]=E\psi^{k_1k_2}(x_1\sigma_1,x_2\sigma_2). \tag{33}$$

After substitution of expressions (29)–(30) for the plane waves into (33) and some simple calculations, the spectrum of two particle state can be found,

$$E=-2t(\cos p_1+\cos p_2). \tag{34}$$

(ii) $|x_1-x_2|=1$. Without loss of generality one can take $x_2=x_1+1$. Then the eigenvalue equation reduces to

$$-t[\psi^{k_1k_2}(x_1-1\sigma_1,x_2\sigma_2)+\psi^{k_1k_2}(x_1\sigma_1,x_2+1\sigma_2)+2J\psi^{k_1k_2}(x_1\sigma_2,x_2\sigma_1)$$
$$+(V-J)\psi^{k_1k_2}(x_1\sigma_1,x_2\sigma_2)=E\psi^{k_1k_2}(x_1\sigma_1,x_2\sigma_2). \tag{35}$$

By substitution of (29) and use of continuity conditions (31) one can express the amplitude $A^{k_2k_1}(p_2\sigma_1,p_1\sigma_2)$ after scattering through $A^{k_1k_2}(p_1\sigma_1,p_2\sigma_2)$ and $A^{k_1k_2}(p_1\sigma_2,p_2\sigma_1)$ before, and, therefore get the R-matrix of the model,

$$A^{k_2k_1}(p_2\sigma_1,p_1\sigma_2)=R^{\sigma_1'\sigma_2',k_2k_1}_{\sigma_1\sigma_2,k_1'k_2'}A^{k_1'k_2'}(p_1\sigma_1',p_2\sigma_2'), \tag{36}$$

where the R-matrix is the product of the ordinary $t-J$ model R-matrix multiplied by the permutation operator in the k-index space,

$$R^{\sigma_1'\sigma_2',k_2k_1}_{\sigma_1\sigma_2,k_1'k_2'}=R^{\sigma_1'\sigma_2'}_{\sigma_1\sigma_2}(t-J)\cdot\delta^{k_2}_{k_1'}\cdot\delta^{k_1}_{k_2'}. \tag{37}$$

At the supersymmetric point $2J=t,\ V=-J/4$,

$$R^{\sigma_1'\sigma_2'}_{\sigma_1\sigma_2}(t-J)=\frac{(\lambda_1-\lambda_2)\hat{P}+i\hat{I}}{\lambda_1-\lambda_2+i}. \tag{38}$$

In (38) $\lambda=\frac{1}{2}\cot(p/2)$ is the rapidity, $\hat{I}=\delta^{\sigma_1'}_{\sigma_1}\delta^{\sigma_2'}_{\sigma_2}$, and $\hat{P}=\delta^{\sigma_2'}_{\sigma_1}\delta^{\sigma_1'}_{\sigma_2}$.

The scattering matrix $S^{\sigma_1'\sigma_2',k_1'k_2'}_{\sigma_1\sigma_2,k_1k_2}$ will be defined multiplying the R-matrix by permutation operator $\hat{P}$ in the spin (σ) and color (k) spaces.

The exact integrability of the model is connected with the fact that the S-matrix should fulfill the Yang–Baxter triangular relations

$$S^{\sigma_1'\sigma_2',k_1'k_2'}_{\sigma_1\sigma_2,k_1k_2}(\lambda_1-\lambda_2)\cdot S^{\sigma_1''\sigma_3',k_1''k_3'}_{\sigma_1'\sigma_3,k_1'k_3}(\lambda_1-\lambda_3)\cdot S^{\sigma_2''\sigma_3'',k_2''k_3''}_{\sigma_2'\sigma_3',k_2'k_3'}(\lambda_2-\lambda_3)$$
$$=S^{\sigma_2'\sigma_3',k_2'k_3'}_{\sigma_2\sigma_3,k_2k_3}(\lambda_2-\lambda_3)\cdot S^{\sigma_1'\sigma_3'',k_1'k_3''}_{\sigma_1\sigma_3',k_1k_3'}(\lambda_1-\lambda_3)\cdot S^{\sigma_1''\sigma_3'',k_1''k_2''}_{\sigma_1'\sigma_2',k_1'k_2'}(\lambda_1-\lambda_2) \tag{39}$$

and constraints on t,J,V are imposed just by these equations.

Consider now N itinerant electrons in a box of L sites with periodic boundary conditions. If we successively make a change of position of an electron and its neighboring electron in a chain, each interchange produces a scattering matrix and when the particle comes back to its starting position from the other side, we will have the cyclic product of S-matrices, which is called the transfer matrix,

$$\hat{T}_j(\lambda_j)=\hat{S}_{j,j+1}(\lambda_j-\lambda_{j+1})\cdots\hat{S}_{j,N}(\lambda_j-\lambda_N)\cdot\hat{S}_{j,1}(\lambda_j-\lambda_1)\cdots\hat{S}_{j,j-1}(\lambda_j-\lambda_{j-1}). \tag{40}$$

Here we skip the matrix indexes σ and k while the hat on S means the operator in that space. The periodicity means that the transfer matrix has to be diagonal for all $j=1,...,N$ with the eigenvalues $\exp(ip_jN)$, or, on terms of rapidity λ_j,

$$e^{ip_jN}=\left(\frac{\lambda_j+i/2}{\lambda_j-i/2}\right)^L. \tag{41}$$

The matrix (40) is the trace of the so-called monodromy matrix, which by definition is the product of the S-matrices without taking trace. Hence, the monodromy matrix can be considered as a $(2\times 2)\otimes((2j+1)\times(2j+1))$ matrix in the spin σ and color k spaces. Let us remember now that this operator is a unity operator in the color space. We will not describe here the details of the algebraic Bethe Ansatz (see, e.g., Ref. 28), but already now it is clear, that because the monodromy matrix of our model is the monodromy matrix of the ordinary $t-J$ model multiplied by the unity matrix in the k-space, the generalization of the algebraic Bethe Ansatz to the present model gives rise to the same equations as the ordinary $t-J$ model.

Specifically we get

$$\left(\frac{\lambda_j+i/2}{\lambda_j-i/2}\right)^L=\prod_{\beta=1}^{M}\frac{\lambda_j-\Lambda_\beta+i/2}{\lambda_j-\Lambda_\beta-i/2},\quad j=1,...,N, \tag{42}$$

$$\prod_{j=1}^{N}\frac{\Lambda_\alpha-\lambda_j+i/2}{\Lambda_\alpha-\lambda_j-i/2}=-\prod_{\beta=1}^{M}\frac{\Lambda_\alpha-\Lambda_\beta+i}{\Lambda_\alpha-\Lambda_\beta-i},\quad \alpha=1,...,M, \tag{43}$$

where L is the number of lattice sites, N is the number of electrons, and M is the number of spin down electrons.

We see that color disappeared from the equations, which means that we can make an arbitrary partition of color charges on the state of N particles in a chain and all wave vectors will become eigenvalues of the Hamiltonian, provided that their λ's fulfill the Bethe equations.

Equations (42)–(43) are Lai's[8] form of Bethe equations, written on a basis of empty background.

The total energy is given by

$$E=-2N+2\sum_{j=1}^{N}\frac{1/2}{\lambda_j^2+1/4}. \tag{44}$$

However for the construction of the thermodynamic limit of our model it is more convenient to use the Sutherland's form [7] of the Bethe equations, which is equivalent to Lai's equations.[11,29] In Sutherland's representation one start with a ferromagnetic pseudovacuum state with all spins up and consider excitations as N_h-holes(empty sites) and M spin down electrons.

In Sutherland's representation we have

$$\begin{aligned}\left(\frac{\lambda_j+i/2}{\lambda_j-i/2}\right)^L&=\prod_{\beta=1}^{M}\frac{\lambda_j-\Lambda_\beta+i/2}{\lambda_j-\Lambda_\beta-i/2},\quad j=1,...,M+N_h,\\ 1&=\prod_{j=1}^{M+N_h}\frac{\lambda_j-\Lambda_\beta+i/2}{\lambda_j-\Lambda_\beta-i/2},\quad \beta=1,...,M.\end{aligned} \tag{45}$$

Equations (45) have real and complex solutions. The complex solutions are in a form known as strings, which may be found by fixing N_h and M (they are conserved quantities) and letting the lattice size $L\to\infty$. Following Refs. 11 and 30 one finds complex solutions in the form

$$\lambda_j=\lambda+i(n+1-2j),\quad j=1,...,n,$$
$$\Lambda_\tau=\lambda+i(n-2\tau),\quad \tau=1,...,n-1, \tag{46}$$

for arbitrary n. In a finite box these string solutions are not exact but as in Refs. 8, 31, and 32, we will assume that the corrections are small in the power of L and hence vanish in the thermodynamic limit.

We would like to write the equation for the centrum of the strings of length n, λ_α^n (which are the real variables), let their number is M' and rapidities Λ_β of spin down electrons.

Substituting (45) into Lai's form of Bethe equations and taking the logarithms from the left- and right-hand sides, one can get

$$L\theta\left(\frac{\lambda_\alpha^n}{n}\right)=2\pi I_\alpha^n+\sum_{m=1}^{\infty}\sum_{\gamma}^{N_m}\Theta_{nm}(\lambda_\alpha^n-\lambda_\gamma^m)+\sum_{\beta=1}^{M}\theta\left(\frac{\lambda_\alpha^n-\Lambda_\beta}{n}\right) \tag{47}$$

and

$$\sum_{n=1}^{\infty}\sum_{\alpha=1}^{N_n}\theta\left(\frac{\Lambda_\beta-\lambda_{pha}^n}{n}\right)=2\pi J_\beta, \tag{48}$$

where N_m is the number of strings of length m, $\theta(x)=2\tan^{-1}(x)$, and

$$\Theta_{mn}=\begin{cases}\theta\left(\frac{x}{|n-m|}\right)+2\theta\left(\frac{x}{|n-m|+2}\right)+\cdots+2\theta\left(\frac{x}{n+m-2}\right) & \text{if } n\neq m\\ 2\theta\left(\frac{x}{2}\right)+2\theta\left(\frac{x}{4}\right)+\cdots+2\theta\left(\frac{x}{n+m-2}\right) & \text{if } n=m\end{cases}. \tag{49}$$

I_α^n and J_β are integers (half-integers) appearing after the choice of the branch of the logarithm.

The integers M, N_m, and M' satisfy the relation

$$M'=M+\sum_{m=1}^{\infty}mN_m. \tag{50}$$

Any solution is defined by a set of I_α^n and J_β. All possible integer (half-integer) values define the states called vacancies. The vacancies may be occupied by particles of color k or not occupied at all. In a case of occupied vacancies one can mark the corresponding I_α^n and J_β by the color index k as for particles, $I_\alpha^{n,k}$ and J_β^k. When the state is empty, the corresponding integer will be mentioned as $\underline{I}_\alpha^n$ and $\underline{J}_\beta$.

One may calculate the number of string states (46) taking account also color degrees of freedom and will get $(2j+1)^L$.

VI. THE THERMODYNAMIC LIMIT OF THE BETHE ANSATZ EQUATIONS

In the thermodynamic limit $L\to\infty$ the solution becomes densely packed (the difference of two neighbor solutions is of order $1/L$, $\lambda_{j+1}^n-\lambda_j^n=O(1/L)$) and one can introduce density functions of the states and pass from sums to integrals in Eqs. (47)–(48).

Let us define the particle and hole densities as follows:

$$\rho_p^{n,k}(\lambda)=\lim_{L\to\infty}\frac{1}{L(\lambda_{I_{j+1}^{n,k}}-\lambda_{I_j^{n,k}})},\quad \rho_h^n(\lambda)=\lim_{L\to\infty}\frac{1}{L(\lambda_{\bar{I}_{j+1}^n}-\lambda_{\bar{I}_j^n})},$$
$$\sigma_p^k(\lambda)=\lim_{L\to\infty}\frac{1}{L(\Lambda_{J_{j+1}^k}-\Lambda_{J_j^k})},\quad \sigma_h(\lambda)=\lim_{L\to\infty}\frac{1}{L(\Lambda_{\bar{J}_{j+1}}-\Lambda_{\bar{J}_j})}, \tag{51}$$

and the sum of the hole and particle densities defines the density functions of vacancies,

$$\rho_t^n=\sum_{k=1}^{2j+1}\rho_p^{n,k}+\rho_h^n,\quad \sigma_t=\sum_{k=1}^{2j+1}\sigma_p^k+\sigma_h. \tag{52}$$

The passage from sum to integral is straightforward and in the thermodynamic limit the Bethe equations (47)–(48) become integral equations,

$$\rho_t^n(\lambda)=f_n(\lambda)-\sum_{m=1}^{\infty}\int_{-\infty}^{\infty}d\lambda' A_{n,m}(\lambda-\lambda')\rho_p^m(\lambda')-\int_{-\infty}^{\infty}d\Lambda f_n(\lambda-\Lambda)\sigma_p(\Lambda) \tag{53}$$

and

$$\sigma_t(\Lambda)=\sum_{n=1}^{\infty}\int_{-\infty}^{\infty}d\lambda f_n(\Lambda-\lambda)\rho_p^n(\lambda). \tag{54}$$

$f_n(\lambda)$ and $A_{n,m}(\lambda)$ in Eqs. (53)–(54) are defined by

$$f_n(\lambda)=\frac{1}{2\pi}\frac{d\theta(\lambda)}{d\lambda}=\frac{1}{\pi}\frac{n}{\lambda^2+n^2} \tag{55}$$

and

$$A_{n,m}=\begin{cases} f_{|n-m|}(\lambda)+2f_{|n-m|+2}(\lambda)+\cdots+2f_{n+m-2}(\lambda), & \text{if } n\neq m\\ 2f_2(\lambda)+2f_4(\lambda)+\cdots+2f_{2n-2}(\lambda), & \text{if } n=m\end{cases}. \tag{56}$$

Following Yang and Yang,[33] Takahashi,[31] and Gaudin,[32] we can write down the thermodynamic equilibrium equations. For the ordinary $t-J$ model it was done in Refs. 5,30, and 34.

The conserved quantities of our model are the following:

(1) *The energy* (the expression in ferromagnetic background slightly differs from Lai's expression (44)),

$$E=L\left(1-4\pi\sum_{n=1}^{\infty}\sum_{k=1}^{2j+1}\int d\lambda f_n(\lambda)\rho_p^{n,k}(\lambda)\right); \tag{57}$$

(2) *The number of different particles*,

$$N^k=L\left(1-\int d\lambda\sum_{n=1}^{\infty}\rho_p^{n,k}(\lambda)+\int d\Lambda\sigma_p^k(\Lambda)\right); \tag{58}$$

(3) *And the magnetization*,

$$S_z=\frac{L}{2}\left(1-\sum_{n=1}^{\infty}\sum_{k=1}^{2j+1}(2n-1)\int d\lambda\rho_p^{n,k}(\lambda)-\sum_{k=1}^{2j+1}\int d\Lambda\sigma_p^k(\Lambda)\right). \tag{59}$$

Correspondingly, for the free energy F one can write

$$F=E-\sum_{k=1}^{2j+1}\mu_k N^k - BS_z - TS, \tag{60}$$

where μ_k are the chemical potentials of particles, B is the external magnetic field, and T is the temperature. We will calculate the entropy S in a standard way, as the logarithm of the number of possible states of the model in the interval $d\lambda$,

$$N(\lambda,d\lambda)=e^{S(\lambda)d\lambda}=\frac{L(\Sigma_{k=1}^{2j+1}\rho_p^k d\lambda)!}{\Pi_{k=1}^{2j+1}(L\rho_p^k d\lambda)!(L\rho_h d\lambda)!}. \tag{61}$$

By Stirling's formula one finds from (61),

$$S=L\left\{\int d\lambda \sum_{m=1}^{\infty}\left[\rho_t^n \log \rho_t^n - \sum_{k=1}^{2j+1}\rho_p^{n,k}\log \rho_p^{n,k} - p_h^n \log \rho_h^n\right]\right.$$
$$\left.+\int d\Lambda\left[\sigma_t \log \sigma_t - \sum_{k=1}^{2j+1}\sigma_p^k \log \sigma_p^k - \sigma_h \log \sigma_h\right]\right\}. \tag{62}$$

To obtain the equilibrium equations one should minimize the free energy over density functions of the particles and holes. Puting $\delta F=0$ for variations of the density functions $\rho_p^{n,k}$, ρ_h^n, σ_p^k, σ_h one gets after some algebraic transformations an infinite set of integral equations for the densities. If one defines the excited energies as usual

$$\frac{\rho_p^{n,k}}{\rho_t^n-\rho_p^{n,k}}=e^{-(\epsilon^{n,k}/T)}, \qquad \frac{\sigma_p^k}{\sigma_t-\sigma_p^k}=e^{-(\epsilon_\Lambda^k/T)} \tag{63}$$

and

$$\epsilon_0^{n,k}=-4\pi f_n(\lambda)-(2n-1)B+\mu_k, \tag{64}$$

we obtain

$$\epsilon^{n,k}(\lambda)=\epsilon_0^{n,k}+T\sum_{m=1}^{\infty}\int d\lambda' A_{n,m}(\lambda-\lambda')\log(1+e^{[-\epsilon^{m,k}(\lambda')]/T})$$
$$-T\int d\Lambda f_n(\lambda-\Lambda)\log(1+e^{[-\epsilon_\Lambda^k(\Lambda)]/T}) \tag{65}$$

for $n=1,...,\infty$, and

$$\epsilon_\Lambda^k=-\mu_k-B+T\sum_{n=1}^{\infty}\int d\lambda f_n(\lambda-\Lambda)\log(1+e^{[-e^{n,k}(\lambda)]/T}). \tag{66}$$

It is seen from these equations for densities and excitation energies that if all particles have the same chemical potential $\mu=\mu_k$, the solutions will be independent of k. Correspondingly, for minimum value of the free energy one can has

$$\frac{F}{L}=(1+B-\mu)-T\sum_{m=1}^{\infty}\int d\lambda f_m(\lambda)\log(1+e^{[-\epsilon^m(\lambda)]/T})=(1-2\mu)-\epsilon_\Lambda(\Lambda=0) \tag{67}$$

in correspondence with Refs. 5 and 30.

In order to analyze the ground state further one should take the limit $T\to 0$ Eqs. (65)–(66) and then put the solution obtained into the expression for the energy (57). If we suppose that $\epsilon^n>0$ for

$n > 1$ (which can be checked after all by analysing the corresponding equations for the ϵ^n's) in the sum of the integral Eqs. (65)–(66), only one term will contribute in the limit $T \to 0$, and we have

$$\epsilon^{1,k}(\lambda) = \epsilon_0^{1,k} + \int d\Lambda f_1(\lambda - \Lambda) \overline{\epsilon_\Lambda^k}(\Lambda),$$
$$\epsilon_\Lambda^k(\Lambda) = -(\mu_k + B) - \int d\lambda f_1(\Lambda - \lambda) \overline{\epsilon^{1,k}}(\lambda), \tag{68}$$

where

$$\overline{\epsilon_\Lambda^k}(\Lambda) = \begin{cases} \epsilon_\Lambda^k(\Lambda), & \text{if } \epsilon_\Lambda^k(\Lambda) < 0 \\ 0, & \text{if } \epsilon_\Lambda^k(\Lambda) > 0 \end{cases}, \quad \overline{\epsilon^{1,k}}(\lambda) = \begin{cases} \epsilon^{1,k}(\lambda), & \text{if } \epsilon^{1,k}(\lambda) < 0 \\ 0, & \text{if } \epsilon_\Lambda^k(\Lambda) > 0 \end{cases}. \tag{69}$$

Hence, in Sutherland's approach we are lead to the concept of two seas; one for real(nonstring) solutions, another for rapidity Λ (spin-down electrons).

Suppose that $\epsilon^{1,k}(\lambda) = 0$ at some point $\lambda = Q$ and $\epsilon_\Lambda^k = 0$ at $\Lambda = Q_\Lambda$, which are the Fermi rapidities. Then we can write the equations for the ground state energy as

$$\frac{E_0}{L} = 1 - 4\pi \sum_{k=1}^{2j+1} \int_{-Q}^{Q} d\lambda f_1(\lambda) \rho_p^{1,k}(\lambda). \tag{70}$$

The equations for the density functions also simplify

$$\sum_{k=1}^{2j+1} \rho_p^{1,k}(\lambda) = f_1(\lambda) - \sum_{k=1}^{2j+1} \left(\int_{-\infty}^{-Q_\Lambda} + \int_{Q_\Lambda}^{\infty} \right) d\Lambda f_1(\lambda - \Lambda) \sigma_p^k(\Lambda) \tag{71}$$

and

$$\sum_{k=1}^{2j+1} \sigma_p(\Lambda) = \sum_{k=1}^{2j+1} \int_{-Q}^{Q} d\lambda f_1(\Lambda - \lambda) \rho_p^{1,k}(\lambda). \tag{72}$$

The Fermi boundaries Q and Q_Λ are determined from the particle number and magnetization equations

$$\frac{N}{L} = 1 - \sum_{k=1}^{2j+1} \left[\int_{-Q}^{Q} d\lambda \rho_p^{1,k}(\lambda) + \left(\int_{-\infty}^{-Q_\Lambda} + \int_{Q_\Lambda}^{\infty} \right) d\Lambda \sigma_p^k(\Lambda) \right],$$
$$\frac{2S_z}{L} = 1 - \sum_{k=1}^{2j+1} \left[\int_{-Q}^{Q} d\lambda \rho_p^{1,k}(\lambda) + \left(\int_{-\infty}^{-Q_\Lambda} - \int_{Q_\Lambda}^{\infty} \right) d\Lambda \sigma_p^k(\Lambda) \right]. \tag{73}$$

Due to a theorem by Lieb and Mattis,[35] in zero magnetic field S_z should be zero, which means that spin-down electrons constitute half of all electrons $M = N/2$ and there is no string state in the ground state.

We see that the equations for the ground state are invariant under transformation of the index k, which means that we can arbitrarily distribute the color over the particles. They are as gauge modes in the vacuum of gauge theories.

In the half-filled case $Q_\Lambda = 0$ and $Q = \infty$ (as easy to see in Lai's form of equations[5]) and following Refs. 5 and 9, one can obtain $E_0 = 1 - 2 \log 2$.

It is also possible, following Ref. 30, to consider the excitations of the model, introduce the so-called shift functions, and write down equations for them. After that, following Ref. 36, one can find the S-matrix for the excitations of the model. It follows as above, that the S-matrix of our model is equal to the S-matrix of ordinary $t-J$ model multiplied by unity operators in an additional color space.

VII. CONCLUSIONS

We have constructed a one-dimensional model which is based on the $t-J$ model of strongly correlated electrons, but which has an additional quantum group symmetry, ensuring the degeneration of the states. We use the Bethe Ansatz technique to investigate this model. The equations for density functions, written in the thermodynamic limit, demonstrate that the additional degrees of freedom of the model behave as gauge modes. The presence of these modes, in our opinion, gives rise to the possibility of constructing a new type of integrable models, if one has an interaction between two models of this kind. Also, different topological properties, usually appearing in gauge theories, could be present in our model, and would be interesting to investigate.

ACKNOWLEDGMENT

This work was supported in part by INTAS Grant No. 96-524.

[1] V. Korepin and F. H. L. Essler, *Exactly Solvable Models of Strongly Correlated Electrons* (World Scientific, Singapore, 1994).
[2] J. Hubbard, Proc. R. Soc. London, Ser. A **276**, 238 (1963).
[3] F. C. Zhang and T. M. Rice, Phys. Rev. B **37**, 3759 (1988).
[4] P. W. Anderson, Phys. Rev. Lett. **65**, 2306 (1990).
[5] P. Schlottmann, Phys. Rev. B **37**, 5177 (1987).
[6] P. Schlottmann, Physica B **171**, 30 (1991).
[7] B. Sutherland, Phys. Rev. B **12**, 3795 (1975).
[8] C. K. Lai, J. Math. Phys. **15**, 167 (1974).
[9] P. A. Bares, G. Blatter, and M. Ogata, Phys. Rev. B **44**, 130 (1991).
[10] S. Sarkar, J. Phys. A **24**, 5775 (1991).
[11] F. H. L. Essler and V. Korepin, Phys. Rev. B **46**, 9147 (1992).
[12] D. Forster, Phys. Rev. Lett. **63**, 2140 (1989).
[13] T. Hakobyan and A. Sedrakyan, Phys. Lett. B **377**, 250 (1996).
[14] A. Avakyan, T. Hakobyan, and A. Sedrakyan, Nucl. Phys. B **490**, [FS],633 (1997).
[15] J. Ambjorn, A. Avakyan, T. Hakobyan, and A. Sedrakyan, cond-mat/9702128.
[16] V. Rittenberg, F. Alcaraz, D. Arnaudon, and M. Scheunert. Int. J. Mod. Phys. A **9**, 19 (1994).
[17] C. N. Yang, Phys. Rev. Lett. **63**, 2144 (1989).
[18] C. N. Yang and S. Zhang, Mod. Phys. Lett. B **4**, 759 (1990).
[19] R. Strack, Phys. Rev. Lett. **70**, 833 (1993); R. Strack and D. Vollhardt, *ibid.* **72**, 3425 (1994).
[20] A. A. Ovchinnikov, Mod. Phys. Lett. B **7**, 1397 (1993).
[21] V. Korepin, J. de Boer, and A. Schadschneider, Phys. Rev. Lett. **74**, 789 (1995).
[22] A. Montorsi and D. K. Campbell, Phys. Rev. B **53**, 5153 (1996).
[23] A. Schadschneider, Phys. Rev. B **51**, 10386 (1995).
[24] V. G. Drinfeld, ''Quantum groups,'' ICM Proceedings, 1986, pp. 798–820 (unpublished).
[25] M. Jimbo, Lett. Math. Phys. **10**, 63 (1985).
[26] M. Jimbo, Lett. Math. Phys. **11**, 247 (1986).
[27] V. Chari and A. Pressley, Commun. Math. Phys. **142**, 261 (1991).
[28] V. Korepin, N. M. Bogoliubov, and A. Izergin, *Quantum Inverse Scattering Method and Correlation Functions* (Cambridge University Press, Cambridge, 1993).
[29] P. A. Bares, J. Carmelo, J. Ferrara, and P. Horsch, Phys. Rev. B **46**, 14624 (1992).
[30] E. Williams, cond-mat/9304009.
[31] M. Takahashi, Prog. Theor. Phys. **46**, 401 (1971).
[32] M. Gaudin, Phys. Rev. Lett. **26**, 1301 (1971).
[33] C. N. Yang and C. P. Yang, J. Math. Phys. **10**, 1115 (1969).
[34] G. Juttner and A. Klumper, Europhys. Lett. **37**, 335 (1997); G. Juttner, A. Klumper, and J. Suzuki, cond-mat/9611058.
[35] E. H. Lieb and D. Mattis, Phys. Rev. **125**, 164 (1962).
[36] F. H. L. Essler and V. Korepin, cond-mat/9310056.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Monotone Riemannian metrics and relative entropy on noncommutative probability spaces

Andrew Lesniewski
Paribas Capital Markets, The Equitable Tower, 787 Seventh Avenue, New York, New York 10019

Mary Beth Ruskai[a)]
Department of Mathematics, University of Massachusetts-Lowell, Lowell, MA 01854

(Received 22 February 1999; accepted for publication 24 May 1999)

We use the relative modular operator to define a generalized relative entropy for any convex operator function g on $(0,\infty)$ satisfying $g(1)=0$. We show that these convex operator functions can be partitioned into convex subsets, each of which defines a unique symmetrized relative entropy, a unique family (parametrized by density matrices) of continuous monotone Riemannian metrics, a unique geodesic distance on the space of density matrices, and a unique monotone operator function satisfying certain symmetry and normalization conditions. We describe these objects explicitly in several important special cases, including $g(w)=-\log w$, which yields the familiar logarithmic relative entropy. The relative entropies, Riemannian metrics, and geodesic distances obtained by our procedure all contract under completely positive, trace-preserving maps. We then define and study the maximal contraction associated with these quantities. © *1999 American Institute of Physics.* [S0022-2488(99)01410-3]

I. INTRODUCTION

For quantum systems, a state is described by a density matrix P, i.e., a positive semidefinite operator with trace one. We will let $\overline{\mathcal{D}}$ denote the set of *density matrices*. For classical discrete or commutative systems we can identify the states with the subset of diagonal density matrices, each of which defines a probability vector $p\in\mathbf{R}^n$. For commutative systems the usual logarithmic relative entropy,

$$H_{\log}(p,q)=\sum_k p_k\log(p_k/q_k), \tag{1}$$

can be generalized to

$$h_g(p,q)=\sum_k p_k g(q_k/p_k), \tag{2}$$

where g is a convex function on $(0,\infty)$ with $g(1)=0$. It is well known that any such H_g contracts under stochastic mappings, i.e., $H_g(Ap,Aq)\leqslant H_g(p,q)$ when A is a column stochastic matrix. Cohen *et al.*[1] defined the entropy contraction coefficient as

$$\eta_g(A)=\sup_{p\neq q}\frac{H_g(Ap,Aq)}{H_g(p,q)}. \tag{3}$$

[a)]Electronic mail: bruskai@cs.uml.edu

In the pair of papers,[1,2] it was shown that for each fixed A all the contraction coefficients associated with those g that are also *operator* convex are equivalent, more precisely the following.

Theorem I.1: *If g is operator convex, then*

$$\eta_g(A)=\eta_{\log}(A)=\eta_{(w-1)^2}(A)\leqslant\eta_{|w-1|}(A). \tag{4}$$

A summary of these results is given in Ref. 3. It suffices to mention here that the observation,

$$\frac{d^2}{dt^2}H_g(p,p+tv)\bigg|_{t=0}=g''(0)\sum_k (v_k)^2/p_k=H_{(w-1)^2}(p,p+v), \tag{5}$$

plays a critical role. The quantity $\Sigma_k(v_k)^2/p_k$ can also be written as $M_p(v,v)$, where

$$M_p(u,v)=-\frac{\partial^2}{\partial\alpha\,\partial\beta}H_g(p+\alpha u,p+\beta v)\bigg|_{\alpha=\beta=0} \tag{6}$$

is the Riemannian metric corresponding to the Fisher information. Čencov[4,5] showed that, for commutative systems, this is the *only* Riemannian metric that satisfies the monotonicity condition $M_{Ap}(Av,Av)\leqslant M_p(v,v)$. Thus, we can regard Theorem I.1 as stating that for operator convex g the maximal contraction of the relative entropy and its associated Riemannian metric are the same. Since there is only one Riemannian metric, all the contraction coefficients must be equal.

For quantum systems, the usual logarithmic relative entropy is given by

$$H_{\log}(P,Q)=\operatorname{Tr}P(\log P-\log Q) \tag{7}$$

$$=\int_0^\infty \operatorname{Tr}P\left[\frac{1}{Q+tI}(P-Q)\frac{1}{P+tI}\right]dt, \tag{8}$$

with P,Q in $\mathcal{D}$, the set of invertible density matrices. The integral representation (8) can be used to show that

$$M_P^{\log}(A,B)\equiv-\frac{\partial^2}{\partial\alpha\,\partial\beta}H_{\log}(P+\alpha A,Q+\beta B)\bigg|_{\alpha=\beta=0}$$

$$=\int_0^\infty \operatorname{Tr}A\left[\frac{1}{P+tI}B\frac{1}{P+tI}\right]dt. \tag{9}$$

Although $M_P^{\log}(A,B)$ is a monotone Riemannian metric, it is not the only possibility; $M_P(A,B)=\operatorname{Tr}A^*P^{-1}B$ is also monotone under completely positive, trace-preserving maps. The study of monotone Riemannian metrics on noncommutative probability spaces was initiated by Morozova and Čencov,[6] who did not, however, provide any explicit examples. A complete characterization of monotone Riemannian metrics (which includes the examples above) was given recently by Petz.[7–10] The quantum structure is much richer because left and right multiplications by P^{-1} are not equivalent. We will see that $M_P(A,B)$ can always be written in the form $\operatorname{Tr}A^*\Omega_P(B)$, where Ω_P reduces to multiplication by P^{-1} when P and B commute. Thus, for example, (9) above gives $\Omega_P(B)=\int_0^\infty[1/(P+tI)]B[1/(P+tI)]dt$, which becomes $P^{-1}B$ when P and B commute.

Earlier, Ruskai[3] tried to extend the entropy contraction coefficient results of Cohen *et al.* to noncommutative situations, but obtained only a few preliminary results. Although one can formally define $H_g(P,Q)=\operatorname{Tr}Pg(Q/P)$ the expression Q/P is ambiguous in the quantum case. Using the nonstandard definition $Q/P=P^{-1/2}QP^{-1/2}$ [which yields $H_g(P,Q)=\operatorname{Tr}P\log P^{-1/2}QP^{-1/2}$ rather than (7) when $g(w)=-\log w$] Ruskai and Petz[11] were able to prove an analog of Theorem I.1, using the fact that

$$\frac{d^2}{dt^2} H_g(P,P+tA)\bigg|_{t=0} = g''(0)\,\mathrm{Tr}\,AP^{-1}A, \tag{10}$$

for all g. In essence, their convention for Q/P always yields the Riemannian metric $M_P(A,B) = \mathrm{Tr}\,A^*P^{-1}B$.

A better alternative is to use the relative modular operator introduced by Araki[12–16,7] to define Q/P. This yields the usual logarithmic entropy (7) and a rich family of generalized relative entropies. Moreover, differentiation then yields the entire family of monotone Riemannian metrics found by Petz.[7–10]

In this paper we use the relative modular operator to study both the relative entropies and Riemannian metrics associated with convex operator functions. For simplicity, we restrict ourselves to the matrix algebras associated with finite dimension systems. Although we do not believe this restriction is essential, it avoids many technical complications. [The most serious arises when the condition $\mathrm{Tr}\,P=1$ is not compatible with the requirement that P be invertible (in the sense of having a bounded inverse in the relevant operator algebra). In that case, one must restrict the domain of $H_g(P,Q)$ to those pairs P,Q that have comparable approximate null spaces in some suitable sense.] We show that each convex operator function defines a convex family of relative entropies, a unique symmetrized relative entropy, a unique family (parametrized by density matrices) of continuous monotone Riemannian metrics, a unique geodesic distance on the space of density matrices, and a unique monotone operator function. We describe these objects explicitly in several important special cases, including $g(w) = -\log w$. We then define and study the contraction coefficient associated with the relative entropy, Riemannian metrics, and metrics. Finally, we present examples showing that these contraction coefficients can have any value in [0,1] for a suitable stochastic map.

The paper is organized as follows. In Sec. II, we give some basic definitions and results for relative entropy and Riemannian metrics. In Sec. III, we define the corresponding geodesic distance, including the Bures metric as a special case. Finally, in Sec. IV we study the contraction of all the quantities under stochastic maps and give bounds on the maximal contraction.

II. RELATIVE ENTROPY AND RIEMANNIAN METRICS

A. Definitions

We begin by describing the relative modular operator that was originally introduced by Araki to generalize the logarithmic relative entropy to type III von Neumann algebras.[12–16,7] Later, Petz[16] used it to generalize relative entropy itself. Let $\mathcal{D}$ denote the subset of invertible operators in $\mathcal{D}$. Let $P,Q \in \mathcal{D}$, i.e., P and Q are positive definite matrices with $\mathrm{Tr}(P) = \mathrm{Tr}(Q) = 1$. For matrix algebras, the relative modular operator associated with the pair of states $\rho_P(A) = \mathrm{Tr}(AP)$ and $\rho_Q(A) = \mathrm{Tr}(AQ)$ reduces to

$$\Delta_{Q,P} = L_Q R_P^{-1}, \tag{11}$$

where L_Q and R_P are the left and right multiplication operators, respectively. Thus, $\Delta_{Q,P}(A) = QAP^{-1}$. It is easy to verify directly that $\Delta_{Q,P}$ is a positive Hermitian operator with respect to the Hilbert–Schmidt inner product.

Definition II.1: Let g be an operator convex function defined on (0,∞) such that $g(1) = 0$. *The relative g entropy of P and Q is*

$$H_g(P,Q) = \mathrm{Tr}(P^{1/2} g(\Delta_{Q,P}) P^{1/2}). \tag{12}$$

We will let $\mathcal{G}$ denote the set of functions satisfying these conditions. Note, however, that the argument of g, as defined here, is shifted from that (which we here denote g_C) in Refs. 1 and 2 so that $g_C(w) = g(w+1)$. Using standard results from the theory of monotone and convex operator functions, one can show that $\mathcal{G}$ is the class of functions that can be written in the form

$$g(w)=a(w-1)+b(w-1)^2+c\,\frac{(w-1)^2}{w}+\int_0^\infty \frac{(w-1)^2}{w+s}\,d\nu(s), \tag{13}$$

where $b,c>0$ and ν is a positive measure on $(0,\infty)$ with finite mass $\int_0^\infty d\nu(s)$. The term $(w-1)^2/w$ may seem unfamiliar, as it is usually included implicitly in the integral. However, writing it separately will be convenient later and is necessary to ensure that the measure has finite mass. The function $g(w)=-\log w$ yields the usual logarithmic relative entropy (7), which we continue to denote as $H_{\log}(P,Q)$. The function $g(w)=(w-1)^2$ yields

$$H_{(w-1)^2}=\mathrm{Tr}(P-Q)P^{-1}(P-Q), \tag{14}$$

which we call the ''quadratic relative entropy;'' it plays an extremely important role in our development. The function $g(w)=(w-1)^2/(w+1)$ yields the equally important, but less familiar, $H_{\mathrm{Bures}}(P,Q)=\mathrm{Tr}(P-Q)[L_Q+R_P]^{-1}(P-Q)$, where we use the subscript Bures because (as will be explained in Sec. III) it eventually leads to a geodesic on $\mathcal{D}$ referred to as the ''metric of Bures.''

We will study the properties of relative entropy and related quantities under a class of maps referred to as ''stochastic'' based on the concept of a completely positive, trace-preserving map $\phi:\mathcal{A}_1\to\mathcal{A}_2$ from one operator algebra to another. Such maps are linear and positivity preserving, not only on the original algebra, but when lifted to tensors products. A precise definition and useful representation theorems can be found in Stinespring,[17] Arveson,[18] Choi,[19] Kraus,[20] and Lindblad.[21] In particular, ϕ is a completely positive map if and only if there exist operators $\{V_k\}$ with $V_k:\mathcal{A}_1\to\mathcal{A}_2$, such that

$$\phi(A)=\sum_{k=1}^{N} V_k A V_k^* . \tag{15}$$

It then follows that ϕ is trace preserving if and only if $\Sigma_k V_k^* V_k=I$; whereas ϕ is unital [i.e., $\phi(I_1)=I_2$] if and only if $\Sigma_k V_k V_k^*=I$. For von Neumann algebras with finite trace (as is the case here) one can use the Hilbert–Schmidt inner product $\langle A,B\rangle=\mathrm{Tr}\,A^*B$ to define the adjoint $\hat{\phi}$ of any completely positive map so that $\mathrm{Tr}\,A^*\phi(B)=\mathrm{Tr}\,\hat{\phi}(A)^*B$. It is then easy to see that $\hat{\phi}(A)=\Sigma_k V_k^* A V_k$ and that ϕ is trace preserving if and only if $\hat{\phi}$ is unital. Then, for the finite-dimensional algebras considered here, we can use the following.

Definition II.2: A stochastic map $\phi:\mathcal{A}_1\to\mathcal{A}_2$ *is a completely positive, trace-preserving map.*

(For general von Neumann algebras, a stochastic map should be defined as the dual of a unital completely positive map.) For commutative systems, a stochastic map always corresponds to a column stochastic matrix, as discussed in the Introduction above and in Refs. 1–3. For noncommutative systems, a partial trace (see Sec. IV D or, e.g., Refs. 22, 21) is an example of a stochastic map.

B. Relative entropy

We begin by defining a relative entropy distance as a bilinear function on $\mathcal{D}$ with the properties we expect of the relative g entropy $H_g(P,Q)$. It is sometimes convenient to extend our definition from $\mathcal{D}\times\mathcal{D}$ to the somewhat larger set of pairs P,Q of positive definite matrices with $\mathrm{Tr}\,P=\mathrm{Tr}\,Q$.

Definition II.3: By a relative entropy distance *we mean a function* $H(P,Q)$ *satisfying the following.*

(a) $H(P,Q)\geq 0$ *with* $H(P,Q)=0\Leftrightarrow P=Q$.
(b) $H(\lambda P,\lambda Q)=\lambda H(P,Q)$ *for* $\lambda>0$.
(c) $H(P,Q)$ *is jointly convex in* P *and* Q.

In addition, we say that the relative entropy is monotone *if*

(d) *$H(P,Q)$ decreases under stochastic maps ϕ,*
that it is symmetric *if*
(e) *$H(P,Q)=H(Q,P)$,*
and that it is differentiable *if*
(f) *the function $g(x,y)=H(P+xA,Q+yB)$ is differentiable.*

Conditions (b), (c), and (d) are not independent. It is well known that by embedding $\mathbf{C}^{n\times n}$ in $\mathbf{C}^{n\times n}\otimes\mathbf{C}^{2\times 2}$ and choosing ϕ to correspond to the partial trace over $\mathbf{C}^2$, one can show that (d) implies the subadditivity relation,

$$H(P_1+P_2,Q_1+Q_2)\leq H(P_1,Q_1)+H(P_2,Q_2). \tag{16}$$

But for functions satisfying the homogeneity condition (b) this is equivalent to joint convexity. Because any stochastic map can be represented as a partial trace,[21] it follows that when (a) and (b) hold, then $(c)\Leftrightarrow(d)$. Nevertheless, the properties of convexity and monotonicity are each of sufficient importance to justify explicitly stating them separately.

A relative entropy distance (even if symmetric) is not a metric in the usual sense, because it need not satisfy the triangle inequality. Nevertheless, such quantities have been widely used[23–25] to measure the difference between P and Q. Later, we shall show that every relative g entropy defines a relative entropy distance that then defines a Riemannian metric and an associated geodesic distance.

Theorem II.4: *Every relative g entropy of the form given in Definition II.1 is a differentiable monotone relative entropy distance in the sense of Definition II.3.*

Proof: Properties (a), (b), and (f) are straightforward; (d) is due to Petz[16] and implies (c) by the above remarks. A simple new proof of (d) is given in Sec. II F.

Theorem II.5: *For each operator convex function $g\in\mathcal{G}$,*

$$H_g(P,Q)=\mathrm{Tr}(Q-P)[b_gP^{-1}+c_gQ^{-1}](Q-P), \tag{17}$$

$$\int_0^\infty \mathrm{Tr}\left((Q-P)\frac{1}{L_Q+sR_P}(Q-P)\right)d\nu_g(s)$$

$$=\mathrm{Tr}[(Q-P)R_P^{-1}g(\Delta_{QP})(Q-P)], \tag{18}$$

where b_g, c_g, and ν_g are as in (13).

Proof: We first observe that

$$(\Delta_{Q,P}-I)(P^{1/2})=(Q-P)P^{-1/2}=R_{P^{-1/2}}(Q-P), \tag{19}$$

so that

$$H_{w-1}(P,Q)=\mathrm{Tr}[P^{1/2}(Q-P)P^{-1/2}]=0, \tag{20}$$

and the linear term in (13) does not contribute. We also find using (19) again,

$$\begin{aligned}H_g(P,Q)&=\langle(\Omega_{Q,P}-I)(P^{1/2}),(\Delta_{Q,P}+sI)^{-1}(\Delta_{Q,P}-I)(P^{1/2})\rangle\\&=\mathrm{Tr}[(Q-P)(\Delta_{Q,P}+sI)^{-1}R_{P^{-1}}(Q-P)]\\&=\mathrm{Tr}(Q-P)\frac{1}{L_Q+sR_p}(Q-P).\end{aligned} \tag{21}$$

Letting $s=0$ yields

$$H_{(w-1)^2/w}(P,Q)=\mathrm{Tr}[(Q-P)Q^{-1}(Q-P)]=H_{(w-1)^2}(Q,P), \tag{22}$$

and one easily verifies that

$$H_{(w-1)^2}(P,Q)=\mathrm{Tr}((Q-P)P^{-1}(Q-P)). \tag{23}$$

Using these results in (13) gives the desired result (17).

It is worth pointing out that the cyclicity of the trace implies that

$$\mathrm{Tr}(Q-P)\frac{1}{R_P+sL_Q}(Q-P)=\mathrm{Tr}(Q-P)\frac{1}{L_P+sR_Q}(Q-P), \tag{24}$$

although

$$\mathrm{Tr}(Q-P)\frac{1}{R_P+sL_Q}(Q-P)\neq\mathrm{Tr}(Q-P)\frac{1}{R_Q+sL_P}(Q-P),$$

in general.

One can also use the heat kernel representation,

$$(\Delta_{Q,P}+sI)^{-1}=\int_0^\infty e^{-u(\Delta_{Q,P}+sI)}\,du, \tag{25}$$

to obtain another integral representation of $H_g(P,Q)$.

Theorem II.6: *Let $m_g(u)=\int_0^\infty e^{-us}d\nu(s)$ denote the Laplace transform of the measure ν_g. Then*

$$H_g(P,Q)=b_gH_{(w-1)^2}(P,Q)+c_gH_{(w-1)^2}(Q,P)+\int_0^\infty H_{(w-1)^2e^{-uw}}(P,Q)m_g(u)du,$$

where we formally extend our definition of $H_g(P,Q)$ to the nonconvex function $g(w)=(w-1)^2e^{-uw}$.

Proof: We use (25) in (13) to get

$$\begin{aligned}&\int_0^\infty\langle(\Delta_{Q,P}-I)(P^{1/2}),(\Delta_{Q,P}+sI)^{-1}(\Delta_{Q,P}-I)(P^{1/2})\rangle d\nu_g(s)\\&=\int_0^\infty\langle(\Delta_{Q,P}-I)(P^{1/2}),e^{-u\Delta_{Q,P}}(\Delta_{Q,P}-I)(P^{1/2})\rangle m_g(u)du\\&=\int_0^\infty\mathrm{Tr}((Q-P)(R_{P^{-1}}e^{-u\Delta_{Q,P}})(Q-P))m_g(u)du\\&=\int_0^\infty H_{(w-1)^2e^{-uw}}(P,Q)m_g(u)du,\end{aligned}$$

where we have interchanged the order of integration and then used (19) again.

C. Monotone Riemannian metrics

We now consider the relation between relative g entropy and Riemannian metrics. Note that the set of density matrices $\mathcal{D}$ has a natural structure as a smooth manifold, so that we can define a Riemannian metric on its tangent bundle $T_*\mathcal{D}$, whose fibers consist of traceless, self-adjoint matrices or

$$T_P\mathcal{D}=\{A=A^*:\mathrm{Tr}\,A=0\}. \tag{26}$$

Definition II.7: By a Riemannian metric on $\mathcal{D}$, we mean a positive definite bilinear form $M_P(A,B)$ on $T_P\mathcal{D}$ such that the map $P\rightarrow M_P(A,A)$ is smooth for each fixed $A\in T_*\mathcal{D}$. The metric is monotone if it contracts under stochastic maps in the sense

$$m_{\phi(P)}[\phi(A),\phi(B)]\leqslant M_P(A,B), \tag{27}$$

when ϕ is a stochastic map.

Note that this definition of monotone requires that the stochastic map ϕ act on the base point (i.e., the indexing density matrix P) as well as the arguments of the bilinear form.

Theorem II.8: *For each $g\in\mathcal{G}$ and density matrix $P\in\mathcal{D}$,*

$$M_P^g(A,B)=-\frac{\partial^2}{\partial\alpha\,\partial\beta}H_g(P+\alpha A,P+\beta B)\bigg|_{\alpha=\beta=0}, \tag{28}$$

$$=\langle A,\Omega_P^g(B)\rangle=\mathrm{Tr}\,A_P^g(B), \tag{29}$$

defines a Riemannian metric on $T_P\mathcal{D}$, and a positive linear operator Ω_P^g on $T_P\mathcal{D}$.

The theorem follows easily from the fact that R_P,L_P and their inverses are positive semidefinite operators with respect to the Hilbert–Schmidt inner product, e.g., $\mathrm{Tr}\,A^*R_PA>0$, and the integral representation in Theorem II.5. We find

$$\begin{aligned}\langle A,\Omega_P^g(B)\rangle&=(b_g+c_g)\mathrm{Tr}[AL_P^{-1}(B)+BL_P^{-1}(A)]\\&\quad+\int_0^\infty \mathrm{Tr}[A(L_P+sR_P)^{-1}(B)+B(L_P+sR_P)^{-1}(A)]d\nu_g(s)\\&=(b_g+c_g)\mathrm{Tr}\,A[L_P^{-1}+R_P^{-1}](B)\\&\quad+\int_0^\infty \mathrm{Tr}\,A[(L_P+sR_P)^{-1}+(R_P+sL_P)^{-1}](B)d\nu_g(s)\\&=\int_0^\infty \mathrm{Tr}\,A[(L_P+sR_P)^{-1}+(R_P+sL_P)^{-1}](B)N_g(s)ds\\&=\left\langle A,\int_0^\infty[(L_P+sR_P)^{-1}+(R_P+sL_P)^{-1}](B)N_g(s)ds\right\rangle,\end{aligned} \tag{30}$$

where, for simplicity, we temporarily subsume the quadratic terms into the integral by defining N_g so that $N_g(s)ds=(b_g+c_g)\delta(s)ds+d\nu_g(s)$. It is critical that A and B are self-adjoint so that we can interchange A and B by replacing L_P by R_P as in

$$\mathrm{Tr}\,BL_P^{-1}(A)=\mathrm{Tr}\,BP^{-1}A=\mathrm{Tr}\,ABP^{-1}=\mathrm{Tr}\,AR_P^{-1}(B). \tag{31}$$

This result would not hold if we did not require the perturbations of P and Q to be self-adjoint. Given that requirement, the result is necessarily symmetric in the sense that we get the same result from both $H_g(P,Q)$ and $H_g(Q,P)$. This is already evident in the quadratic term, whose coefficient depends only on the sum $b+c$, and will be discussed further below.

We can now use (30) to obtain several explicit formulas for Ω_P^g:

$$\Omega_P^g=\int_0^\infty\left(\frac{1}{sR_P+L_P}+\frac{1}{sL_P+R_P}\right)N_g(s)ds \tag{32}$$

$$=\int_0^\infty\frac{1}{sR_P+L_P}(N_g(s)+s^{-1}N_g(s^{-1}))ds \tag{33}$$

$$=R_P^{-1}\int_0^{\infty}\frac{1}{s+\Delta_{P,P}}\,\sigma_g(s)ds \tag{34}$$

$$=\int_0^1\left(\frac{1}{sR_P+L_P}+\frac{1}{sL_P+R_P}\right)\sigma_g(s)ds, \tag{35}$$

where we have used the change of variable $s\to s^{-1}$ and

$$\sigma_g(s)=N_g(s)+s^{-1}N_g(s^{-1}).$$

Note that $\sigma_g(s^{-1})=s\sigma_g(s)$. Then, if we define

$$k(\lambda)=\int_0^{\infty}\frac{1}{s+\lambda}\,\sigma_g(s)ds=\int_0^1\left[\frac{1}{s+\lambda}+\frac{1}{s\lambda+1}\right]\sigma_g(s)ds, \tag{36}$$

we find that $k(\lambda^{-1})=\lambda k(\lambda)$, $\Omega_P^g=R_P^{-1}k(\Delta_{P,P})$, and that k can be expressed in terms of g as

$$k(w)=\frac{g(w)+wg(w^{-1})}{(w-1)^2}. \tag{37}$$

We will let $\mathcal{K}$ denote this set of functions, i.e.,

$$\mathcal{K}=\{k:-k \text{ is operator monotone, } k(w^{-1})=wk(w), \quad\text{and}\quad k(1)=1\}. \tag{38}$$

We have recovered half of Petz's result[7–10] that there is a one-to-one correspondence between symmetric Riemannian metrics and functions of the form (36) that satisfy the normalization condition $k(1)=1$. (But note that our k corresponds to $1/f$ in Petz's notation.) Our approach also easily yields an explicit expression for both Ω_P^g and its inverse.

Theorem II.9: *For each $g\in\mathcal{G}$ and $P\in\mathcal{D}$, the operator Ω_P^g as defined in Theorem II.8 satisfies $\Omega_P^g=R_P^{-1}k(L_PR_P^{-1})$ and $[\Omega_P^g]^{-1}=R_Pf(L_PR_P^{-1})$, where $k(w)$ is given by (37) and $f(w)=1/k(w)$.*

Although Ω_P^g is initially defined only on $T_*\mathcal{D}$, it can easily be extended to all traceless matrices using the natural complexification $\mathrm{Tr}\,A=0\Rightarrow A=A_1+iA_2$ with $A_1,A_2\in T_P\mathcal{D}$ and then to all of $\mathbf{C}^{n\times n}$ using linearity and $\Omega_P^g(I)=P^{-1}I$. The result is equivalent to using any of the formulas for Ω_P^g above together with the obvious extension of L_P and R_P to all of $\mathbf{C}^{n\times n}$. We can summarize this discussion as follows.

Theorem II.10: *For each $g\in\mathcal{G}$ and $P\in\mathcal{D}$, the operator Ω_P^g as defined in Theorem II.8 can be extended to a positive linear operator on $\mathbf{C}^{n\times n}$ so that $M_P^g(A,B)=\mathrm{Tr}\,A^*\Omega_P^g(B)$ defines an inner product on $\mathbf{C}^{n\times n}$. On the other hand, for each $g\in\mathcal{G}$ and $P\in\mathcal{D}$, Eq. (34) defines a positive linear operator Ω_P^g on all of $\mathbf{C}^{n\times n}$, and the bilinear form $M_P^g(A,B)=\mathrm{Tr}\,A^*\Omega_P^g(B)$ extends to a monotone Riemannian metric satisfying the symmetry condition $M_P^g(A,B)=M_P^g(B^*,A^*)$.*

This result is essentially due to Petz,[7–10] who also showed the converse result that every symmetric monotone Riemannian metric is of this form. We give an independent proof of monotonicity at the end of this section. That the metric is symmetric is a consequence of the cyclicity of the trace.

The following result is essentially due to Kubo and Ando,[26] who developed a theory of operator means.

Theorem II.11: *If k given by (36) satisfies $k(1)=1$, then for all $P,Q\in\mathcal{D}$,*

$$R_P^{-1}+L_Q^{-1}\geqslant R_P^{-1}k(\Delta_{Q,P})\geqslant(R_P+L_Q)^{-1}. \tag{39}$$

Proof: This follows easily from (36), the elementary inequality,

$$\frac{w+1}{2w} \geqslant \frac{1+t}{2}\left[\frac{1}{t+w}+\frac{1}{tw+1}\right] \geqslant \frac{2}{w+1}, \tag{40}$$

and the fact that the normalization $k(1)=1$ implies that $2\sigma_g(t)/(t+1)$ is a probability measure on $[0,1]$.

As immediate corollaries, we find

$$\Omega_P^{(w-1)^2} = L_P^{-1} + R_P^{-1} \geqslant \Omega_P^{\sigma} \geqslant (R_P + L_P)^{-1} = \Omega_p^{\text{Bures}}, \tag{41}$$

$$M_P^{(w-1)^2}(A,A) \geqslant M_P^g(A,A) \geqslant M_P^{\text{Bures}}(A,A), \tag{42}$$

$$H_{(w-1)^2}^{\text{sym}}(P,Q) \geqslant H_g^{\text{sym}}(P,Q) \geqslant H_{\text{Bures}}(P,Q), \tag{43}$$

where the superscript indicates the symmetric relative entropy associated with g. Thus $k(w)=2/(w+1)$ corresponds to the minimum symmetric relative entropy and minimum Riemannian metric among the class studied here. By contrast, we will see that $g(w)=(w-1)^2$ corresponds to $k(w)=(w+1)/(2w)$ so that the quadratic relative entropy is maximal.

The operators Ω_P^g and $[\Omega_P^g]^{-1}$ are noncommutative versions of multiplication by P^{-1} and P, respectively. Hence, in view of the cyclicity of the trace, the following result is not surprising.

Theorem II.12: *The operator* Ω_P^g *given by (34) satisfies* $\mathrm{Tr}\,\Omega_P^g(A) = \mathrm{Tr}\,AP^{-1}$ *and* $\mathrm{Tr}[\Omega_P^g]^{-1}(A) = \mathrm{Tr}\,AP$.

Proof: We first observe that in a basis in which P is diagonal with eigenvalues p_k,

$$\left[R_P^{-1}\frac{1}{s+\Delta_{P,P}}(A)\right]_{jk} = \left[\frac{1}{sR_P+L_P}(A)\right]_{jk} = \frac{1}{sp_k+p_j}a_{jk}, \tag{44}$$

so that

$$[\Omega_P^g(A)]_{jk} = \int_0^\infty \frac{a_{jk}}{sp_k+p_j}\sigma_g(s)ds. \tag{45}$$

Then for every $g\in\mathcal{G}$, $P\in\mathcal{D}$, and $A\in T_P\mathcal{D}$,

$$\begin{aligned}\mathrm{Tr}\,\Omega_P^g(A) &= \sum_j \int_0^\infty \frac{a_{jj}}{sp_j+p_j}\sigma_g(s)ds\\ &= \sum_j p_j^{-1}a_{jj}\int_0^\infty \frac{1}{s+1}\sigma_g(s)ds = k(1)\,\mathrm{Tr}\,P^{-1}A = \mathrm{Tr}\,P^{-1}A.\end{aligned}$$

The proof for the inverse is similar. Since $1/k$ is also operator monotone, we can use Theorem II.9 to conclude that $[\Omega_P^g]^{-1}$ can be written in the form

$$[\Omega_P^g]^{-1} = aR_P + bL_P - \int_0^\infty \frac{R_P^2}{sR_P+L_P}d_\mu(s),$$

for some positive measure μ.

D. Correspondence between defining functions

We now make some remarks on the relation between $g(w)$, $wg(w^{-1})$, and $k(w)$. It should be clear from the development above that every function $g\in\mathcal{G}$ defines a Riemannian metric and a function k as in (36) or (37). If we now consider $\hat{g}(w)=wg(w^{-1})$, it is easy to verify that $\hat{g}(w)\in\mathcal{G}$ as well and that $H_{\hat{g}}(P,Q)=H_g(Q,P)$. Thus, the map $g(w)\to wg(w^{-1})$ has the effect of switching the arguments of the relative entropy and the function $g(w)+wg(w^{-1})$ yields the symmetrized relative entropy $H_g(P,Q)+H_g(Q,P)$. Now, if we begin with a function g and relative entropy $H_g(P,Q)$, the differentiation in (28) automatically yields a symmetric result. Thus, all convex combinations $ag(w)+(1-a)\hat{g}(w)$ of g and $\hat{g}(w)$ yield the same Riemannian metric and the same function $k\in\mathcal{K}$.

Conversely, every $k\in\mathcal{K}$ defines a unique *symmetric* relative entropy via the function $g^{\rm sym}(w)=(w-1)^2k(w)$. It follows immediately from the integral representation (37) and (13) that $g^{\rm sym}(w)$ is also in $\mathcal{G}$ and that $wg^{\rm sym}(w^{-1})=g^{\rm sym}(w)$. Thus, k selects from the convex set of relative entropies associated with a given $g\in\mathcal{G}$, the symmetric one. If we observe that the integral representation (36) is equivalent to $-k$ being an operator monotone function, we can summarize the discussion above as follows.

Theorem II.13: *There is a one-to-one correspondence between each of the following.*

(a) *Monotone Riemannian metrics extended to bilinear forms via the symmetry condition* $M_P^g(A,B)=M_P^g(B^*,A^*)$.
(b) *Monotone (decreasing) operator functions satisfying* $k(w^{-1})=wk(w)$ *with the normalization* $k(1)=1$.
(c) *Convex operator functions in* $\mathcal{G}$ *that satisfy the symmetry relation* $wg(w^{-1})=g(w)$.

The relations between these are given by (34), (36), and (37). In view of this theorem, it would be appropriate to identify a given operator Ω_P^g by using the (unique) symmetric function $g^{\rm sym}$. However, we will continue to use the asymmetric g for such familiar cases as the logarithm. One might expect the one-to-one correspondence to extend to twice-differentiable symmetric monotone relative entropies. However, Petz and Ruskai[11] consider relative entropies of the form $\tilde{H}_g(P,Q)=\mathrm{Tr}\,Pg(P^{-1/2}QP^{-1/2})$. This class of monotone relative entropies can be symmetrized, however, differentiation of H_g yields the Riemannian metric $M_P^{(w-1)^2}(A,B)=\mathrm{Tr}\,A^*[P^{-1}B+BP^{-1}]$ for all $g\in\mathcal{G}$. Thus, in particular $\tilde{H}_{\log}(P,Q)=\mathrm{Tr}\,P\log(P^{-1/2}QP^{-1/2})$ is an example of a relative entropy distance (in the sense of Definition II.3) that is *not* a relative g entropy. Another class of distinct relative entropy distances is given by squares of the geodesic distances introduced in Sec. III. Thus, the properties in Definition 1.3 are *not* sufficient to completely characterize the relative g entropy and allow us to extend the one-to-one correspondence in Theorem 2.4 to a class of relative entropies. Although we believe that such an additional condition must exist, we have not found it.

E. Examples

We now give explicit expressions for the relative entropy, Ω_P^g and related quantities in several important special cases. These examples will also illustrate the relation between the functions g, $\hat{g}$, $g^{\rm sym}$, and k discussed above.

Example 1: Take $g(w)=-\log w$. Then $\hat{g}(w)=w\log w$, $g^{\rm sym}=(w-1)\log w$, $k(w)=(w-1)^{-1}\log w$, $N_g(s)=(s+1)^{-2}$ and $\sigma_g(s)=1/(s+1)$. Then, $H_{\log}(P,Q)$ is given by (7), and

$$H_{\log}^{\rm sym}=H_{(w-1)\log w}=\mathrm{Tr}(P-Q)[\log P-\log Q] \tag{46}$$

$$=\int_0^\infty \mathrm{Tr}(P-Q)\,\frac{1}{Q+xI}\,(P-Q)\,\frac{1}{P+xI}\,dx \tag{47}$$

and

$$\Omega_P^{\log}=\int_0^\infty\left(\frac{1}{sR_P+L_P}+\frac{1}{sL_P+R_P}\right)\frac{1}{(s+1)^2}\,ds=\int_0^\infty\frac{1}{s+1}\,\frac{1}{L_P+sR_P}\,ds.$$

Making the change of variables $s\to sR_P$ in the last integral yields

$$\Omega_P^{\log}=\int_0^\infty \frac{1}{s+L_P}\,\frac{1}{s+R_P}\,ds, \tag{48}$$

so that

$$\langle A,\Omega_P^{\log}(B)\rangle=\mathrm{Tr}\int_0^\infty A^*\frac{1}{sI+P}\,B\,\frac{1}{sI+P}\,ds, \tag{49}$$

a result that we obtained earlier (9) using the integral representation (8), or

$$\log P-\log Q=\int_0^\infty\left[\frac{1}{Q+xI}-\frac{1}{P+xI}\right]dx. \tag{50}$$

In this case, it is also well known[27,15] that the inverse operator can be written as

$$[\Omega_P^{\log}]^{-1}(B)=\int_0^1 P^tBP^{1-t}\,dt. \tag{51}$$

Example 2: Take $g(w)=(w-1)^2$. Then $\hat{g}(w)=(w-1)^2/w$, $g^{\mathrm{sym}}(w)=(w-1)^2(w+1)/w$ and $k(w)=(w+1)/(2w)$. Then $H_{(w-1)^2}(P,Q)$ is given by (14),

$$H^{\mathrm{sym}}_{(w-1)^2}(P,Q)=\mathrm{Tr}(Q-P)[P^{-1}+Q^{-1}](Q-P),$$

$$\Omega_P^{(w-1)^2}=R_P^{-1}+L_P^{-1},$$

and

$$\langle A,\Omega_P^{(w-1)^2}(A)\rangle=H_{(w-1)^2}(P,P+A)=\mathrm{Tr}\,AP^{-1}A.$$

The associated function is the maximal function satisfying the prescribed conditions. The operator $\Omega_P^{(w-1)^2}(B)=P^{-1}B+BP^{-1}$, so that

$$\Omega_P^{(w-1)^2}=R_P^{-1}+L_P^{-1}=R_P^{-1}[R_P+L_P]L_P^{-1}. \tag{52}$$

Example 3: For $s_0>0$ take $g_{s_0}(w)=(w-1)^2/(w+s_0)$. Then $\hat{g}_{s_0}(w)=(w-1)^2/(1+ws_0)$, $g^{\mathrm{sym}}(w)=(w-1)^2(w+1)(1+s_0)/(1+ws_0)(w+s_0)$, $k(w)=(w+1)(1+s_0)/(1+ws_0)(w+s_0)$, and $N_g(s)=\delta(s-s_0)$. Thus

$$\Omega_P^{g_{s_0}}=\frac{1}{s_0R_P+L_P}+\frac{1}{s_0L_P+R_P}=(s_0+1)[s_0R_P+L_P]^{-1}[R_P+L_P][R_P+s_0L_P]^{-1}. \tag{53}$$

When no confusion will result, it will be convenient to employ a slight abuse of notation and write $\Omega_P^{s_0}$ for $\Omega_P^{g_{s_0}}$. The case $s_0=1$ is particularly important; we have already seen that it yields the minimal $k\in\mathcal{K}$. Then $k(w)=2/(1+w)$, $g(w)=g^{\mathrm{sym}}(w)=(w-1)^2/(w+2)$, and $\Omega_P^{g_{s_0=1}}\equiv\Omega_P^{\mathrm{Bures}}=[R_P+L_P]^{-1}$. The corresponding Riemannian metric is $\langle A,\Omega_P^{\mathrm{Bures}}(B)\rangle=\mathrm{Tr}\,A^*[R_P+L_P]^{-1}(B)$ and the corresponding relative entropy,

$$H_{\mathrm{Bures}}(P,Q)=\mathrm{Tr}(Q-P)[R_P+L_Q]^{-1}(Q-P)=\mathrm{Tr}\,QXPX, \tag{54}$$

where $X=[R_P+L_Q]^{-1}(Q-P)$. Because of the cyclicity of the trace, $H_{\mathrm{Bures}}(P,Q)$ is already symmetric and $[R_Q+L_P]^{-1}$ would have given the same result.

Example 4: Take $g(w)=1-w^\alpha$. Then $k(w)=(1-w^\alpha)(1-w^{1-\alpha})/\alpha(1-\alpha)(1-w)^2$ and $N_g(s)=(\sin\pi s/\pi)(1+s)^{\alpha-2}$. Thus

$$H_{1-w^\alpha}(P,Q)=1-\mathrm{Tr}\,Q^\alpha P^{1-\alpha},$$

$$\Omega_P^g=R_P^{-1}\int_0^\infty \frac{1}{sI+\Delta_{P,P}}\,\frac{\sin \pi s}{\pi}\,(1+s)^{\alpha-2}\,ds.$$

After the change of variables $s\rightarrow sR_P$, this becomes

$$\Omega_P^g=\frac{\sin \pi s}{\pi}\int_0^\infty \frac{1}{L_P+s}\,\frac{R_P^{1-\alpha}+s^{1-\alpha}}{(R_P+s)^{2-\alpha}}\,ds. \tag{55}$$

F. Monotonicity proof

We now present a new proof of the monotonicity of the relative entropies and Riemannian metrics associated with convex operator functions.

Theorem II.14: *For every convex operator function g of the type considered here, both the relative entropy $H_g(P,Q)$ and the corresponding Riemannian metric are monotone, i.e.,*

$$H_g(P,Q)\geqslant H_g[\phi(P),\phi(Q)], \tag{56}$$

$$\langle A\Omega_P^g A\rangle \geqslant \langle \phi(A)\Omega_{\phi(P)}^g\phi(A)\rangle. \tag{57}$$

This result is essentially due to Petz.[16] We given an independent proof as an immediate corollary of the following theorem and the integral representations (17) and (34).

Theorem II.15: If ϕ is stochastic,

$$\mathrm{Tr}\,A^*\frac{1}{R_P+sL_Q}A=\mathrm{Tr}\,\phi\left(A^*\frac{1}{R_P+sL_Q}A\right)\geqslant \mathrm{Tr}\,\phi(A^*)\,\frac{1}{R_{\phi(P)}+sL_{\phi(Q)}}\,\phi(A). \tag{58}$$

Proof: If $P>0$, then $\mathrm{Tr}\,A^*PA\geqslant 0$, and $\mathrm{Tr}\,A^*AP\geqslant 0$, so that both L_P and R_P are positive as operators on the Hilbert–Schmidt space. Thus, for $Q>0$, the operator R_P+sL_Q is also positive. Let $X=[R_P+sL_Q]^{-1/2}(A)-[R_P+sL_Q]^{1/2}\hat{\phi}(B)$ with $B=[R_{\phi(P)}+sL_{\phi(Q)}]^{-1}\phi(A)$. Then $\mathrm{Tr}\,X^*X\geqslant 0$, so that

$$\mathrm{Tr}\,A^*\frac{1}{R_P+sL_Q}A-\mathrm{Tr}\,A^*\hat{\phi}(B)-\mathrm{Tr}\,\hat{\phi}(B^*)A+\mathrm{Tr}\,\hat{\phi}(B^*)[R_P+sL_Q]\hat{\phi}(B)\geqslant 0. \tag{59}$$

Since it is easy to see that

$$-\mathrm{Tr}\,A^*\hat{\phi}(B)-\mathrm{Tr}\,\phi(B^*)A=-2\,\mathrm{Tr}\,\phi(A^*)\,\frac{1}{R_{\phi(P)}+sL_{\phi(Q)}}\,\phi(A),$$

the desired result will follow if we can show that the last term in (59) is bounded above by the right side of (58). We find

$$\begin{aligned}\mathrm{Tr}\,\hat{\phi}(B^*)[R_P+sL_Q]\hat{\phi}(B)&=\mathrm{Tr}\,\hat{\phi}(B^*)\hat{\phi}(B)P+\hat{\phi}(B^*)sQ\hat{\phi}(B)\\&=\mathrm{Tr}\,\hat{\phi}(B^*)\hat{\phi}(B)P+\hat{\phi}(B)\hat{\phi}(B^*)sQ\\&\leqslant \mathrm{Tr}\,\hat{\phi}(B^*B)P+\hat{\phi}(BB^*)sQ,\end{aligned}$$

where the inequality follows from the positivity of P and Q and the operator inequality:

$$\hat{\phi}(B^*)\hat{\phi}(B)\leqslant \hat{\phi}(B^*B), \tag{60}$$

which holds for any B because the trace-preserving condition on ϕ gives $\hat{\phi}(I_2)=I_1$. Then using, e.g., $\mathrm{Tr}\,\hat{\phi}(B^*B)P=\mathrm{Tr}\,B^*B\phi(P)$, we find

$$\begin{aligned}\mathrm{Tr}\,\hat{\phi}(B^*)[R_P+sL_Q]\hat{\phi}(B) &\leqslant \mathrm{Tr}\,B^*B\phi(P)+BB^*s\phi(Q)\\ &=\mathrm{Tr}\,B^*[B\phi(P)+s\phi(Q)B]\\ &=\mathrm{Tr}\,B^*[R_{\phi(P)}+sL_{\phi(Q)}]B=\mathrm{Tr}\,B^*\phi(A)\\ &=\mathrm{Tr}\,\phi(A^*)\frac{1}{R_{\phi(P)}+sL_{\phi(Q)}}\phi(A).\end{aligned}$$

It is interesting to observe that the strategy used here is very similar to that used by Lieb and Ruskai[22] to prove a Schwarz inequality for completely positive mappings and, as a special case, the monotonicity of the quadratic relative entropy. At that time, Lieb and Ruskai could use these Schwarz inequalities to prove many special cases of the strong subadditivity of the logarithmic relative entropy, but not the general case. A complete proof of strong subadditivity[28] (see also Refs. 3, 29) seemed to require one of the convex trace function theorems of Lieb.[27] It is therefore curious that now, some 25 years later, we have finally found a way to recover strong subadditivity directly from the Schwarz strategy of Lieb and Ruskai.[22]

It should also be noted that Uhlmann had earlier[30] used a very different approach (based on interpolation theory) to show the logarithmic relative entropy was monotone under a related class of mappings that are Schwarz in the sense $\phi(A^*A)\geqslant\phi(A^*)\phi(A)$ and Petz[16] extended this to other relative entropies.

III. GEODESIC DISTANCE

We now wish to consider the contraction of the relative entropy and corresponding Riemannian metric under stochastic mappings. Before doing so, it will be useful to consider the geodesic distance that arises from the Riemannian metrics considered here.

Definition III.1: Associated with every Riemannian metric $\langle A,\Omega_P^g(B)\rangle$ *of the form (28) is a geodesic distance* $D_g(P,Q)$, *which is defined as*

$$D_g(P,Q)\equiv\inf\int_0^1\sqrt{\langle\dot{S}(t),\Omega_{S(t)}^g\dot{S}(t)\rangle dt},$$

where the infimum is taken over all smooth paths $S(t)$ *with* $S(0)=P$ *and* $S(1)=Q$.

Theorem III.2: *The square* $[D_g(P,Q)]^2$ *of every geodesic distance of the form given in Definition III.1 is a differentiable monotone relative entropy distance in the sense of Definition II.3. In addition,* $D_g(P,Q)$ *satisfies the triangle inequality* $D_g(P,R)\leqslant D_g(P,Q)+D_g(Q,R)$.

Proof: Properties (a), (b), and (e) of Definition II.3 are readily verified. Property (d), i.e., the monotonicity $D_g[\phi(P),\phi(Q)[\leqslant D_g(P,Q)$, can be proven directly, but also follows easily as a corollary to Theorem IV.2 below. The triangle inequality is standard. That $D_g(P+xA,Q+yB)$ is differentiable in the sense of Definition II.3(f) follows from standard results (see, e.g., Theorem 3.6, part (2) of Ref. 31. Q.E.D.

It is well known (see, e.g., Refs. 32–36) that the metric associated with the minimal function $k(w)=2/(1+w)$, discussed in Example 3, is (except for normalization) the metric of Bures, i.e., $D_{2(w-1)^2/(1+w)}(P,Q)=4D^{\mathrm{Bures}}(P,Q)$, where

$$\begin{aligned}[D^{\mathrm{Bures}}(P,Q)]^2&=\inf\{\mathrm{Tr}(W-X)(W-X)^*:WW^*=P,XX^*=Q\}\\ &=2[1-\mathrm{Tr}(\sqrt{P}Q\sqrt{P})^{1/2}] \qquad (61)\\ &\leqslant\mathrm{Tr}[\sqrt{P}-\sqrt{Q}]^2=2[1-\mathrm{Tr}\sqrt{P}\sqrt{Q}]=H_{1-\sqrt{w}}(P,Q). \qquad (62)\end{aligned}$$

It follows immediately from (41) that

$$D^{\mathrm{sym}}_{(w-1)^2}(P,Q) \geq D_g(P,Q) \geq 4D^{\mathrm{Bures}}(P,Q), \tag{63}$$

so that $4D^{\mathrm{Bures}}(P,Q)$ gives the minimal geodesic distance of this type.

IV. CONTRACTION UNDER STOCHASTIC MAPS

A. Contraction coefficients

Because the relative entropies, Riemannian metrics, and geodesic distances all contract under stochastic maps, their maximal contraction is a well-defined quantity in the following sense.

Definition IV.1: For each fixed convex operator function g of the form given in Def. II.1 and stochastic map ϕ, *we define three entropy contraction coefficients:*

$$\eta_g^{\mathrm{RelEnt}}(\phi) = \sup_{P\neq Q} \frac{H_g[\phi(P),\phi(Q)]}{H_g[P,Q]}, \tag{64}$$

$$\eta_g^{\mathrm{Riem}}(\phi) = \sup_{P} \sup_{A\in T_P\mathcal{D}} \frac{\langle \phi(A), \Omega^g_{\phi(P)}[\phi(A)]\rangle}{\langle A, \Omega^g_P[A]\rangle}, \tag{65}$$

$$\eta_g^{\mathrm{geod}}(\phi) = \sup_{P\neq Q} \frac{[D_g(\phi(P),\phi(Q))]^2}{[D_g(P,Q)]^2}. \tag{66}$$

In Refs. 1, 2 it was shown that for commutative systems, $\eta_g^{\mathrm{RelEnt}}(\phi) = \eta_g^{\mathrm{Riem}}(\phi) = \eta_{(w-1)^2}(\phi)$. Here, we will prove some relations between these various η.

Theorem IV.2: *The three contraction coefficients defined above satisfy*

$$1 \geq \eta_g^{\mathrm{RelEnt}}(\phi) \geq \eta_g^{\mathrm{Riem}}(\phi) \geq \eta_g^{\mathrm{geod}}(\phi). \tag{67}$$

The intuition behind the second inequality can be seen by letting $A=B=Q-P$ in the integral representations of Theorems 2.2 and 2.3. Then the only difference between the ratios in (64) and (65) is that the modular operator in the former is $\Delta_{Q,P}$ while that in the latter is $\Delta_{P,P}$. This would seem to indicate that the first supremum is taken over a larger set. However, the two are not directly comparable because the condition $P\neq Q$ in the first case precludes the choice $\Delta_{P,P}$. Hence, we consider $Q=P+\epsilon A$.

Proof: The upper bound of 1 follows immediately from Theorem II.14. To prove the second inequality $\eta_g^{\mathrm{RelEnt}}(\phi) \geq \eta_g^{\mathrm{Riem}}(\phi)$, we consider, as suggested above, $H_g(P,P+\epsilon A) = \mathrm{Tr}\, P^{1/2} g(\Delta_{P,P+\epsilon A})(P^{1/2})$. Proceeding as in the proof of Theorem II.5, but with the shorthand $dN_g(s) = (b_g + c_g)\delta(s)ds + d\nu_g(s)$, we obtain

$$\begin{aligned} H_g(P,P+\epsilon A) &= \epsilon^2 \int_0^\infty \mathrm{Tr}\left[A \frac{1}{L_{P+\epsilon A} + sR_P}(A)\right] dN_g(s) \\ &= \epsilon^2 \int_0^\infty \mathrm{Tr}\left[A \frac{1}{L_P + sR_P}(A)\right] dN_g(s) + O(\epsilon^2) \\ &= \epsilon^2 \langle \phi(A), \Omega^g_{\phi(P)}(A))\rangle + O(\epsilon^3). \end{aligned}$$

Thus

$$\eta_g^{\mathrm{RelEnt}}(\phi) = \sup_{P\neq Q} \frac{H_g[\phi(P),\phi(Q)]}{H_g[P,Q]} \geq \sup_{P} \sup_{A\in T_*\mathcal{D}} \frac{H_g[\phi(P),\phi(P+\epsilon A)]}{H_g(P,P+\epsilon A)}.$$

However,

$$\frac{H_g[\phi(P),\phi(P+\epsilon A)]}{H_g(P,P+\epsilon A)}=\frac{\langle\phi(A),\Omega^g_{\phi(P)}[\phi(A)]\rangle+O(\epsilon)}{\langle A,\Omega^g_P[A]\rangle+O(\epsilon)}.$$

Since the quantity on the right can be made arbitrarily close to $\eta_g^{\rm Riem}(\phi)$, we conclude that $\eta_g^{\rm RelEnt}(\phi)\geq\eta_g^{\rm Riem}(\phi)$. Finally, to prove the third inequality we first choose $S_0(t)$ to be a minimizing path for $D_g(P,Q)$, i.e.,

$$D_g(P,Q)=\int_0^1\sqrt{\langle\dot{S}_0(t),\Omega^g_{S_0(t)}\dot{S}_0(t)\rangle}dt.$$

Then, $\phi\circ S_0$ is a smooth path from $\phi(P)$ to $\phi(Q)$. Moreover, the linearity of ϕ implies that $(d/dt)\phi\circ S_0(t)=\phi\circ\dot{S}_0(t)$. Thus

$$\begin{aligned}D_g[\phi(Q),\phi(Q)]&\leq\int_0^1\sqrt{\langle\phi\circ\dot{S}_0(t),\Omega^g_{\phi\circ S_0(t)}\phi\circ\dot{S}_0(t)\rangle}dt\\&\leq[\eta_g^{\rm Riem}(\phi)]^{1/2}\int_0^1\sqrt{\langle\dot{S}_0(t),\Omega^g_{S_0(t)}\dot{S}_0(t)\rangle}dt\\&=[\eta_g^{\rm Riem}(\phi)]^{1/2}D_g(P,Q).\end{aligned}$$

Dividing both sides by $D_g(P,Q)$ and taking the supremum of the left-hand side, gives the desired result. Q.E.D.

In this case of the first inequality $\eta_g^{\rm RelEnt}(\phi)\geq\eta_g^{\rm Riem}(\phi)$, we proved slightly more, namely, that either equality holds or the supremum in $\eta_g^{\rm RelEnt}(\phi)$ is actually attained for some non-negative (but not necessarily strictly positive) density matrices P,Q, i.e., strict inequality implies that there exists $P\neq Q\in\overline{\mathcal{D}}$ such that

$$H_g[\phi(P),\phi(Q)]=\eta_g^{\rm RelEnt}(\phi)H_g(P,Q). \tag{68}$$

This follows from the fact that we can always find a maximizing sequence (P_k,Q_k) such that

$$\lim_{k\to\infty}\frac{H_g[\phi(P_k),\phi(Q)_k]}{H_g(P_k,Q_k)}=\eta_g^{\rm RelEnt}(\phi).$$

Since we are in a finite-dimensional space, the space of non-negative density matrices is compact so that we can find a convergent subsequence $(P_{k_j},Q_{k_j})\rightarrow(P,Q)$. Then either $P=Q$, in which case we necessarily have $\eta_g^{\rm RelEnt}(\phi)=\eta_g^{\rm Riem}(\phi)$, or (68) holds. (Strictly speaking, we must also exclude the possibility that both $H_g(P_k,Q_k)$ and $H_g[\phi(P_k),\phi(Q_k)]$ diverge to ∞.) We expect that for most choices of g, Eq. (68) holds only in very special cases [see, e.g., the partial trace example in Sec. IV D, which yield $\eta_g^{\rm RelEnt}(\phi)=1=\eta_g^{\rm Riem}(\phi)$.] Indeed, even for commutative systems, early proofs[37,1] that equality holds for $\eta_{\log}(A)=\eta_{(w-1)^2}(A)$ depended on a demonstration that (68) could not hold in general.

Another special situation occurs for the minimal g, which yields the Bures metric. If P,Q commute, then

$$\begin{aligned}H_{\rm Bures}(P,Q)&\equiv{\rm Tr}(P-Q)([L_P+R_Q]^{-1}+[L_Q+R_P]^{-1})(P-Q)\\&=2\,{\rm Tr}(P-Q)(P+Q)^{-1}(P-Q)=2\langle(P-Q),\Omega^{\rm Bures}_{P+Q}[(P-Q)]\rangle.\end{aligned}$$

Thus, if the supremum for $\eta^{\rm Riem}_{\rm Bures}(\phi)$ happens to be attained for a commuting pair R,A (with $R\in\mathcal{D}$ and $A\in T_P\mathcal{D}$) whose images $\phi(R),\phi(A)$ also commute, then

$$H_{\text{Bures}}[\phi(R+A),\phi(R-A)]=\eta_{\text{Bures}}^{\text{Riem}}(\phi)H_{\text{Bures}}(R+A,R-A). \tag{69}$$

If $\eta_{\text{Bures}}^{\text{RelEnt}}(\phi)=\eta_{\text{Bures}}^{\text{Riem}}(\phi)$, then this also yields equality in (68); however, it does *not* give a strict inequality for $\eta_{\text{Bures}}^{\text{RelEnt}}(\phi)\geqslant\eta_{\text{Bures}}^{\text{Riem}}(\phi)$. On the contrary, it seems to offer some heuristic support for equality.

We expect that in those exceptional situations in which the supremum $\eta_g^{\text{RelEnt}}(\phi)$ is attained the result is equal to $\eta_g^{\text{Riem}}(\phi)$ so that equality always holds, at least for the first inequality in Theorem IV.2.

Recall that many common choices for g [e.g., $g(w)=(w-1)^2$ or $g(w)=-\log w$] do not yield a symmetric relative entropy, i.e., $H_g(P,Q)\neq H_g(Q,P)$. This raises the question of whether or not the entropy contraction coefficient [which we denote $\eta_g^{\text{sym}}(\phi)\equiv\eta_{g(w)+wg(w^{-1})}^{\text{RelEnt}}(\phi)$] for the symmetrized relative entropy,

$$H_g^{\text{sym}}(P,Q)=H_g(P,Q)+H_g(Q,P)=H_{g(w)+wg(w^{-1})}(P,Q), \tag{70}$$

is the same as $\eta_g^{\text{RelEnt}}(\phi)$. Although we believe equality holds, we can only prove that

$$\eta_g^{\text{sym}}(\phi)\leqslant\eta_g^{\text{RelEnt}}(\phi). \tag{71}$$

Nevertheless, Theorem IV.2 holds for any g. In fact, since there is a unique Riemannian metric associated with all g, which yield the same symmetrized relative entropy, we have $\eta_g^{\text{RelEnt}}(\phi)\geqslant\eta_g^{\text{sym}}(\phi)\geqslant\eta_g^{\text{Riem}}(\phi)$. To prove (71) it suffices to observe that

$$H_{g(w)+wg(w^{-1})}(P,Q)=H_g^{\text{sym}}(P,Q)=H_g(P,Q)+H_g(Q,P),$$

so that

$$\begin{aligned}H_g^{\text{sym}}[\phi(P),\phi(Q)]&=H_g[\phi(P),\phi(Q)]+H_g[\phi(Q),\phi(P)]\\&\leqslant\eta_g^{\text{RelEnt}}(\phi)H_g(P,Q)+\eta_g^{\text{RelEnt}}(\phi)H_g(Q,P)\\&=\eta_g^{\text{RelEnt}}(\phi)H_g^{\text{sym}}(P,Q).\end{aligned}$$

In the case of the quadratic entropy, it easily follows that $\eta_{(w-1)^2}^{\text{Riem}}(\phi)=\eta_{(w-1)^2}^{\text{RelEnt}}(\phi)=\eta_{(w-1)^2}^{\text{sym}}(\phi)$.

Finally, we note that the joint convexity of relative entropy, Riemannian metrics, and $[D_g(P,Q)]^2$ imply that the corresponding contraction coefficients are convex in ϕ. [Although we did not explicitly state the joint convexity for $M_P(A,A)$ it is an easy consequence of homogeneity and contraction under partial traces.]

Theorem IV.3: *For each fixed $g\in\mathcal{G}$, each of the contraction coefficients $\eta_g^{\text{RelEnt}}(\phi)$, $\eta_g^{\text{Riem}}(\phi)$, and $\eta_g^{\text{geod}(\phi)}$ is convex in ϕ.*

Proof: Since the argument is straightforward, we give details only for the relative entropy. Let $\phi=x\phi_1+(1-x)\phi_2$:

$$\begin{aligned}H_g[\phi(P),\phi(Q)]&=H_g[x\phi_1(P)+(1-x)\phi_2(P),x\phi_1(Q)+(1-x)\phi_2(Q)]\\&\leqslant xH_g[\phi_1(P),\phi_1(Q)]+(1-x)H_g[\phi_2(P),\phi_2(Q)]\\&\leqslant x\eta_g^{\text{RelEnt}}(\phi_1)H_g(P,Q)+(1-x)\eta_g^{\text{RelEnt}}(\phi_2)H_g(P,Q)\\&=[x\eta_g^{\text{RelEnt}}(\phi_1)+(1-x)\eta_g^{\text{RelEnt}}(\phi_2)]H_g(P,Q).\end{aligned}$$

Dividing both sides by $H_g(P,Q)$ implies

$$\eta_g^{\text{RelEnt}}(\phi)\leqslant x\eta_g^{\text{RelEnt}}(\phi_1)+(1-x)\eta_g^{\text{RelEnt}}(\phi_2).$$ Q.E.D.

B. Eigenvalue formulation of $\eta_g^{\text{Riem}}(\phi)$

We now show how $\eta_g^{\text{Riem}}(\phi)$ is related to the following set of eigenvalue problems:

$$[\hat{\phi}\circ\Omega_{\phi(P)}^g\circ\phi](A)=\lambda\,\Omega_P^g(A). \tag{72}$$

In view of Theorem II.10, this is a well-defined linear eigenvalue problem on $\mathbf{C}^{n\times n}$ for each fixed pair ϕ and P. The following remarks are easily verified.

(a) The eigenvalue problem (72) can be rewritten as $\Phi_P^g\circ\phi(B)=\lambda B$, where

$$\Phi_P^g\equiv(\Omega_P^g)^{-1}\circ\hat{\phi}\circ\Omega_{\phi(P)}^g. \tag{73}$$

Furthermore, Φ_P^g is trace preserving. This follows from Theorem II.12 and

$$\begin{aligned}\operatorname{Tr}\Phi_P^g(B)&=\operatorname{Tr}P\,\hat{\phi}\circ\Omega_{\phi(P)}^g(B)=\langle P,\hat{\phi}\circ\Omega_{\phi(P)}^g(B)\rangle\\&=\langle\phi(P),\Omega_{\phi(P)}^g(B)\rangle=\langle\Omega_{\phi(P)}^g[\phi(P)],B\rangle\\&=\langle I,B\rangle=\operatorname{Tr}B.\end{aligned}$$

(b) We can assume without loss of generality that the matrices that are eigenvectors in (72) are self-adjoint, i.e., that $A=A^*$. Indeed, it is easy to check that the operator $\Omega_P^{g_{s_0}}(A)=(sR_P+L_P)\times[R_P+L_P]^{-1}(R_P+sL_P)(A)$ satisfies $[\Omega_P^{g_{s_0}}(A)]^*=\Omega_P^{g_{s_0}}(A^*)$. Therefore, the operators Ω_P^g, $\Omega_{\phi(P)}^g$, ϕ, $\hat{\phi}$ and Φ_P^g all map adjoints to adjoints.

(c) For each fixed P, the eigenvalue equation is satisfied with $A=P$ and eigenvalue $\lambda=1$, which is the largest eigenvalue. The operators on both sides of (72) are self-adjoint (in fact, positive definite) with respect to the Hilbert–Schmidt inner product and the corresponding orthogonality condition for the other eigenvectors reduces to $\operatorname{Tr}A=0$.

In view of these observations, it is easy to conclude from the max–min principle that the second-largest eigenvalue $\lambda_2^g(\phi,P)$ satisfies

$$\lambda_2^g(\phi,P)=\sup_{A\in T_P\mathcal{D}}\frac{\langle\phi(A),\Omega_{\phi(P)}^g[\phi(A)]\rangle}{\langle A,\Omega_P^g[A]\rangle}, \tag{74}$$

for each fixed P. Then taking the supremum over $\mathcal{D}$ yields

Theorem IV.4: *For each $g\in\mathcal{G}$ and stochastic map ϕ,*

$$\eta_g^{\text{Riem}}(\phi)=\sup_{P\in\mathcal{D}}\lambda_2^g(\phi,P). \tag{75}$$

We have already observed that every Ω_P^g can be regarded as a noncommutative variant of multiplication by P^{-1}. Indeed, if both pairs of operators P,A and $\phi(P),\phi(A)$ associated with a particular eigenvalue commute for some g, then $\Omega_P^g(A)=R_{P^{-1}}(A)=L_{P^{-1}}(A)$ for *all* g and the corresponding eigenvalue equations are the same. It may be tempting to conjecture that the eigenvalue equations for different g are related by a similarity transform, which would then imply that all $\lambda_2(\phi,P)$ are equal so that all $\eta_g^{\text{Riem}}(\phi)$ are identical. However, for a given fixed P, R_P, and L_P commute, which implies that Ω_P^g and Ω_P^h commute for any pair of functions g and h. Since commuting operators are simultaneously diagonalizable and similar operators have the same eigenvalues, this would imply that all of the eigenvalue operators $B\to[(\Omega_P^g)^{-1}\circ\hat{\phi}\circ\Omega_{\phi(P)}^g\circ\phi](B)$ are identical. This is easily seen to be false in specific examples. Moreover, as discussed at the end of Sec. IV D, one can find examples of nonunital ϕ for which different $\eta_g^{Riem}(\phi)$ are *not* identical.

Theorem IV.5: *We can rewrite the eigenvalue problem (72) so that*

$$\lambda_2^g(\phi,P)=\sup_\alpha \frac{\langle \hat{\phi}(\alpha),(\Omega_P^g)^{-1}[\hat{\phi}(\alpha)]\rangle}{\langle \alpha,(\Omega_{\phi(P)}^g)^{-1}[\alpha]\rangle},$$

where the supremum is now taken over $\{\alpha\in \mathrm{Range}(\phi):\mathrm{Tr}[\Omega_{\phi(P)}^g]^{-1}(\alpha)=0\}$.

Proof:

$$\lambda_2^g(\phi,P)=\sup_{A:\mathrm{Tr}(A)=0} \frac{\langle \phi(A),\Omega_{\phi(P)}^g[\phi(A)]\rangle}{\langle A,\Omega_P^g[A]\rangle}$$

$$=\sup_{B:\mathrm{Tr}[\Omega_P^g]^{-1/2}(B)=0} \frac{\langle B[\Omega_P^g]^{-1/2}\circ\hat{\phi}\circ\Omega_{\phi(P)}^g\circ\phi\circ[\Omega_P^g]^{-1/2}B]\rangle}{\langle B,B\rangle}.$$

If we now write $\Gamma=[\Omega_{\phi(P)}^g]^{1/2}\circ\phi\circ[\Omega_P^g]^{-1/2}$, we see that $\lambda_2^g(\phi,P)$ is the largest eigenvalue of $\Gamma^*\Gamma$, where Γ maps

$$\{B:\mathrm{Tr}[\Omega_P^g]^{-1/2}(B)=0\}\to\{\beta\in\mathrm{Range}(\phi):\mathrm{Tr}[\Omega_{\phi(P)}^g]^{-1/2}(\beta)=0\}.$$

Since $\Gamma\Gamma^*$ and $\Gamma^*\Gamma$ have the same nonzero eigenvalues,

$$\lambda_2^g(\phi,P)=\sup_{\beta:\mathrm{Tr}[\Omega_{\phi(P)}^g]^{-1/2}(\beta)=0} \frac{\langle \beta[\Omega_{\phi(P)}^g]^{1/2}\circ\phi\circ[\Omega_P^g]^{-1}\circ\hat{\phi}\circ[\Omega_{\phi(P)}^g]^{1/2}]\beta\rangle}{\langle \beta,\beta\rangle}$$

$$=\sup_{\alpha:\mathrm{Tr}[\Omega_{\phi(P)}^g]^{-1}(\alpha)=0} \frac{\langle \hat{\phi}(\alpha)[\Omega_P^g]^{-1}\hat{\phi}(\alpha)\rangle}{\langle \alpha,[\Omega_{\phi(P)}^g]^{-1}\alpha\rangle}.$$

If we apply this result with $\Omega_P^{\mathrm{Bures}}=[R_P+L_p]^{-1}$, it follows easily from the theorem above that

$$\lambda_2^{\mathrm{Bures}}(\phi,P)=\sup_{\alpha:\mathrm{Tr}\ \phi(P)\alpha=0} \frac{\mathrm{Tr}\ \phi(\alpha)P\hat{\phi}(\alpha)}{\mathrm{Tr}\ \alpha\phi(P)\alpha}. \tag{76}$$

It is tempting to write $\phi(P)\alpha=\beta=\phi(B)$ and replace the constraint $\mathrm{Tr}\ \phi(P)\alpha=0$ by $\mathrm{Tr}\,B=0$. The denominator would then become $\langle \phi(B)[\phi(P)]^{-1}\phi(B)\rangle$, which has the same form as the numerator in (65) when $k(w)=(w+1)/2w$ [corresponding to $g=(w-1)^2$]. However, there is no guarantee that $\phi(\alpha)=\hat{\phi}([\phi(P)]^{-1}]B)$. On the contrary, this cannot possibly hold because we would then have that the λ (and hence η) for the two extremal functions $k(w)=2/(1+w)$ and $k(w)=(w+1)/2w$ are inverses, which is inconsistent with $\lambda^g(\phi,P)\leqslant\eta_g^{\mathrm{Riem}}(\phi)\leqslant 1$ (except in the case $\lambda=1$, which is not generic). There is, however, a sense in which the operators associated with these two extremal functions are inverses since $\Omega_P^{(w-1)^2}=R_P^{-1}+L_P^{-1}=R_P^{-1}[R_P+L_P]L_P^{-1}=R_P^{-1}[\Omega_P^{\mathrm{Bures}}]^{-1}L_P^{-1}$. It seems likely that if the η_g^{Riem} for these two extremal functions are equal, then all of them are.

Unlike the case of $\eta_g^{\mathrm{RelEnt}}(\phi)$, we do expect that the supremum for $\eta_g^{\mathrm{Riem}}(\phi)$ is actually attained. Indeed, we know that for each fixed P the supremum in (74) is attained for some $A\neq 0$, which satisfies the eigenvalue problem (72). As before, we can find a maximizing sequence of density matrices P_k for (75) so that $\eta_g^{\mathrm{Riem}}(\phi)=\lim_{k\to\infty}\lambda_2^g(\phi,P_k)$. For each P_k, let A_k be the solution to the eigenvalue problem (72) for $\lambda_2^g(\phi,P_k)$ normalized so that $\mathrm{Tr}|A_k|=1$. Then we can find a convergent subsequence for which $P_k\to P\in\mathcal{D}$ and $A_k\to A\neq 0$, since $\mathrm{Tr}|A|\neq 0$. It then follows that (72) holds for this P,A with $\lambda=\eta_g^{\mathrm{Riem}}(\phi)$ (although P is only non-negative), which implies that

$$\langle \phi(A),\Omega^{g}_{\phi(P)}[\phi(A)]\rangle = \eta^{\mathrm{Riem}}_{g}(\phi)\langle A,\Omega^{g}_{P}(A)\rangle .$$

C. Bounds on contraction coefficients

We first give an upper bound for $\eta^{\mathrm{Riem}}_{\log}$ using

$$\eta^{\mathrm{Dobrushin}}(\phi)\equiv \sup_{A\in T_{*}\mathcal{D}} \frac{\mathrm{Tr}|\phi(A)|}{\mathrm{Tr}|A|}. \tag{77}$$

This can be interpreted as the norm of ϕ regarded as an operator on the Banach space of traceless matrices with norm $\mathrm{Tr}|A|$. Although the function $g(w)=|w-1|$ is *not* operator convex, $\eta^{\mathrm{Dobrushin}}(\phi)$ is analogous to the contraction coefficient of the (nondifferentiable) symmetric relative g entropy $H_{|w-1|}(P,Q)=\mathrm{Tr}|P-Q|$, which, however, is *not* the relative g entropy obtained by using $g(w)=|w-1|$ in Definition II.1. Nevertheless, $\eta^{\mathrm{Dobrushin}}(\phi)$ is a natural and useful object to consider. It was shown in Ref. 3 (see Theorem 2) that

$$\eta^{\mathrm{Dobrushin}}(\phi)=\tfrac{1}{2}\sup\{\mathrm{Tr}|\phi(E-F)|:E,F \text{ 1-dim proj s};EF=0\}, \tag{78}$$

where ''1-dim projs'' means that E,F are one-dimensional projections in $\overline{\mathcal{D}}$. The expression on the right in (78) shows that we are justified in interpreting $\eta^{\mathrm{Dobrushin}}(\phi)$ as a noncommutative analog of Dobrushin's coefficient of ergodicity.

Theorem IV.6: *If ϕ is stochastic,*

$$\eta^{\mathrm{Riem}}_{\log}(\phi)\le \eta^{\mathrm{Dobrushin}}(\phi)\equiv \sup_{A\in T_{*}\mathcal{D}} \frac{\mathrm{Tr}|\phi(A)|}{\mathrm{Tr}|A|}. \tag{79}$$

Proof: The map $B\to(\Omega^{\log}_{P})^{-1}\circ\hat{\phi}\circ\Omega^{\log}_{\phi(P)}(B)\equiv\Phi_{\log}(B)$ is positivity preserving, as well as trace preserving. The former follows from the integral representations (49) and (51) for $\Omega^{\log}_{P}$ and its inverse, together with the fact that the composition of positivity-preserving maps is positive preserving. Then, taking the trace of the absolute value of both sides of the eigenvalue problem $\Phi[\phi(A)]=\lambda A$ and using Theorem 1 of Ref. 3 yields

$$\lambda\,\mathrm{Tr}|A|=\mathrm{Tr}|\Phi[\phi(A)]|\le \mathrm{Tr}|\phi(A)|. \quad \text{Q.E.D.} \tag{80}$$

Although we believe that this result holds for any g, we do not have a proof except for the log. Our proof depended on the observation that in the case of the log the map $\Phi_{g}(B)=(\Omega^{g}_{P})^{-1}\circ\hat{\phi}\circ\Omega^{g}_{\phi(P)}(B)$ is positivity preserving. However, explicit examples can be found to show that Φ_{g} is *not* positivity preserving in general. Indeed, although both $\Omega^{\mathrm{Bures}}_{P}=[R_{P}+L_{P}]^{-1}$ and $\Omega^{(w-1)^{2}}_{P}=R_{P}^{-1}+L_{P}^{-1}$ are positive semidefinite with respect to the Hilbert–Schmidt inner product, they are not positivity preserving in the sense of mapping positive operators to positive operators. The difference is analogous to the difference between an ordinary matrix being positive semidefinite and having positive elements.

We now consider lower bounds on $\eta^{\mathrm{Riem}}_{g}(\phi)$. In Ref. 3 it was shown that

$$\eta^{\mathrm{Dobrushin}}(\phi)\le\sqrt{\eta^{\mathrm{Riem}}_{(w-1)^{2}}(\phi)}. \tag{81}$$

We now give a lower bound that holds for all $\eta^{\mathrm{Riem}}_{g}(\phi)$ when the map ϕ is unital, i.e., $\phi(I)=I$.

Theorem IV.7: *If ϕ is unital,*

$$\eta^{\mathrm{Riem}}_{g}(\phi)\ge \sup_{\mathrm{Tr}\,A=0}\frac{\mathrm{Tr}|\phi(A)|^{2}}{\mathrm{Tr}|A|^{2}}. \tag{82}$$

This is an immediate consequence of the definition (65); it also follows from Theorem IV.4 and the fact that the right side of (82) is just $\lambda_2(\phi,I)$ when ϕ is unital. The right side of (82) can also be interpreted as the square of the norm of ϕ regarded as an operator on the Banach space of traceless matrices with Hilbert–Schmidt norm $\sqrt{\operatorname{Tr} A^*A}$. When ϕ is self-adjoint in the sense $\hat{\phi}=\phi$, every trace-preserving map is unital.

If ϕ maps $\mathbf{C}^{n\times n}$ to itself, then the results of this section can be restated in terms of the eigenvalues and singular values of ϕ. Since ϕ is trace preserving, $\phi(B)=\Lambda B$ implies that either $\Lambda=1$ or $\operatorname{Tr} B=0$. If we restrict ϕ to the matrices with trace zero, then $\eta^{\text{Dobrushin}}(\phi)$ is the largest magnitude of an eigenvalue and for unital ϕ, $\lambda_2(\phi,I)$ is the largest eigenvalue of $\hat{\phi}\phi$. Thus, for unital stochastic maps, $\lambda_2(\phi,I)=\Lambda_2(\hat{\phi}\phi)$, where we have continued our convention of using the subscript 2 for eigenvalues of maps restricted to $T_*\overline{\mathcal{D}}$. If ϕ is self-adjoint, the two lower bounds (81) and (82) coincide and $\lambda_2(\phi,I)=\Lambda_2(\hat{\phi}\phi)=[\Lambda_2(\phi)]^2$ in the usual sense of second largest eigenvalue of. For general unital ϕ, (82) is stronger since

$$\eta^{\text{Riem}}_{(w-1)^2}(\phi)\geqslant\lambda_2(\phi,I)=\Lambda_2(\hat{\phi}\phi)\geqslant[\eta^{\text{Dobrushin}}(\phi)]^2. \tag{83}$$

We now explicitly state some conjectures that have already been discussed.

Conjecture IV.8: For each fixed $g\in\mathcal{G}$,

$$\eta_g^{\text{RelEnt}}(\phi)=\eta_g^{\text{Riem}}(\phi)=\eta_g^{\text{geod}}(\phi)\leqslant\eta_g^{\text{Dobrushin}}(\phi). \tag{84}$$

Conjecture IV.9: If ϕ *is unital, then*

$$\eta_g^{\text{Riem}}=\Lambda_2(\hat{\phi}\phi)\equiv\sup_{\operatorname{Tr} A=0}\frac{\operatorname{Tr}|\phi(A)|^2}{\operatorname{Tr}|A|^2}, \tag{85}$$

for all $g\in\mathcal{G}$.

If this conjecture holds, then for unital ϕ the contraction coefficient η_g^{Riem} is independent of g. Theorem IV.13 at the end of the next section contains an explicit example of a nonunital stochastic map for which η_g^{Riem} depends nontrivially on g; therefore, the hypothesis that ϕ be unital is essential. In view of (82) it would suffice to show that $\eta_g^{\text{Riem}}\leqslant\Lambda_2(\hat{\phi}\phi)$.

D. Examples

We now consider some special classes of stochastic maps $\phi:\mathcal{A}_1\to\mathcal{A}_2$. We begin by looking at some maps for which all contraction coefficients are easily seen to be zero or one. We then consider maps from $\mathbf{C}^{2\times2}$ to $\mathbf{C}^{2\times2}$ that provide support for the conjectures above.

We first consider the case in which $\mathcal{A}_2$ is one dimensional, e.g., ϕ projects onto a one-dimensional subalgebra (which need not have an identity) of $\mathcal{A}_1$. Then, since ϕ is trace preserving and maps density matrices to density matrices, we must have $\phi(P)=\phi(Q)$, $\forall P,Q$ with $\operatorname{Tr}\varphi(P)=1$ so that $\phi(P)\neq0$. Thus, $H_g[\phi(P),\phi(Q)]=D_g[\phi(P),\phi(Q)]=0$, $\forall P,Q$, which implies $\eta_g^{\text{RelEnt}}(\phi)=\eta_g^{\text{geod}}(\phi)=0$. If $\operatorname{Tr} B=0$, then $\phi(B)=0$. [To see this note that one can find a,b such that $P=(aI+bB)$ is a density matrix.] Thus, $\langle\phi(B)\Omega^g_{\phi(P)}\phi(B)\rangle=0$ and $\operatorname{Tr}|\phi(B)|=0$ for all B in $T_*\overline{\mathcal{D}}$, which implies that $\eta_g^{\text{Riem}}(\phi)=\eta^{\text{Dobrushin}}(\phi)=0$. We can summarize this as the following.

Theorem IV.10: *If the image of the stochastic map* ϕ *is one dimensional, then* $\eta_g^{\text{RelEnt}}(\phi)=\eta_g^{\text{Riem}}(\phi)=\eta_g^{\text{geod}}(\phi)=\eta^{\text{Dobrushin}}(\phi)=0$ *for all* $g\in\mathcal{G}$.

We next consider the important special case in which ϕ is a partial trace τ. In the simplest case, let $\tau:\mathbf{C}^{2n\times2n}\to\mathbf{C}^{n\times n}$ be the map that takes

$$\mathbf{M}=\begin{pmatrix}A & B\\ C & D\end{pmatrix}\to\tau(\mathbf{M})=A+D, \tag{86}$$

where $\mathbf{M}\in\mathbf{C}^{2n\times 2n}$ has been written in block form and $A,B,C,D\in\mathbf{C}^{n\times n}$. Then the homogeneity of relative entropy (see Definition II.3b) implies that for

$$\mathbf{P}=\begin{pmatrix} P & 0 \\ 0 & P \end{pmatrix} \quad \text{and} \quad \mathbf{Q}=\begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix},$$

$$H_g(\mathbf{P},\mathbf{Q})=H_g(2P,2Q)=H_g(\tau(\mathbf{P}),\tau(\mathbf{Q})),$$

for any g, and, similarly,

$$\langle \mathbf{A},\Omega_P^g(\mathbf{A})\rangle=\langle 2A,\Omega_{2P}^g(2A)\rangle=\langle \tau(\mathbf{A}),\Omega_{\tau(P)}^g(\tau(\mathbf{A}))\rangle,$$

when

$$\mathbf{A}=\begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}.$$

From this, we easily see that

$$\eta_g^{\text{RelEnt}}(\phi)=\eta_g^{\text{Riem}}(\phi)=\eta_g^{\text{geod}}(\phi)=\eta^{\text{Dobrushin}}(\phi)=1, \tag{87}$$

where we have assumed implicitly that τ acts on the full algebra of all $2n\times 2n$ matrices.

The partial trace described above is similar to a conditional expectation, i.e., a map for which $\mathcal{A}_2$ is a subalgebra (with identity) of $\mathcal{A}_1$ and $\phi(A)=A$, $\forall A\in\mathcal{A}_2$. Both partial traces and conditional expectations are included in the following.

Theorem IV.11: *If the stochastic map ϕ is also an isomorphism from a nontrivial subalgebra (with identity) of $\mathcal{A}_1$ to $\mathcal{A}_2$, then $\eta_g^{\text{RelEnt}}(\phi)=\eta_g^{\text{Riem}}(\phi)=\eta_g^{\text{geod}}(\phi)=\eta^{\text{Dobrushin}}(\phi)=1$ for all $g\in\mathcal{G}$.*

Since every completely positive map can be represented as a partial trace,[21] this might seem to suggest that $\eta=1$ always holds. However, these representations involve multiple copies of the algebra, so that the partial trace is not acting on the full algebra in the higher-dimensional space. Thus, the representation of $\mathcal{A}_1$ need necessarily not contain a subalgebra with the desired isomorphism property. Examples of maps with $\eta<1$ were already found in Ref. 1 for commutative algebras, and two different noncommutative examples are given below.

We now state two results for maps $\phi:\mathbf{C}^{2\times 2}\to\mathbf{C}^{2\times 2}$. The proofs are postponed to a subsequent paper.[38] Recall that any density matrix in $\mathbf{C}^{2\times 2}$ can be written in the form $\frac{1}{2}[I+\mathbf{w}\cdot\sigma]$, where $\mathbf{w}\in\mathbf{R}^3$ and σ denotes the vector of Pauli matrices. The first theorem provides evidence for the two conjectures at the end of the previous section.

Theorem IV.12: *For the unital map $\phi_T:I+\mathbf{w}\cdot\sigma\to I+T\mathbf{w}\cdot\sigma$,*

$$\eta_g^{\text{RelEnt}}(\phi_T)=\eta_g^{\text{Riem}}(\phi_T)=\eta_g^{\text{geod}}(\phi_T)=\|T\|^2, \quad \forall g\in\mathcal{G},$$

and $\eta^{\text{Dobrushin}}(\phi_T)=\|T\|$.

The next example gives a nonunital stochastic map for which $\eta_g^{\text{Riem}}(\phi)$ varies with g. For $\alpha,\tau>0$ with $\alpha+\tau\leq 1$, define

$$\phi_{\alpha,\tau}[I+\mathbf{w}\cdot\sigma]=I+\alpha w_1\sigma_1+\tau\sigma_2. \tag{88}$$

It is easily seen to be stochastic because the condition $\alpha+\tau\leq 1$ ensures that it is a convex combination of stochastic maps. For $g_{s_0}(w)=(w-1)^2/(w+s_0)$ as in Example 3 of Sec. II E,

$$\eta_{g_{s_0}}^{\text{Riem}}(\phi)=\sup_{0\leqslant\omega\leqslant1}\frac{[(1-\tau^2+(\rho-\alpha^2)\omega^2][1-\omega^2]}{[1-\tau^2-\alpha^2\omega^2][1-\tau^2(1-\rho)-(1-\rho)\alpha^2\omega^2]}$$

$$\geqslant\frac{\alpha^2}{1-\left(\frac{1-s_0}{1+s_0}\right)^2\tau^2},$$

where $1-\rho=(1-s_0)/(1+s_0)$ and equality holds for $s_0\approx0$. In particular, we can conclude the following.

Theorem IV.13: *For the nonunital stochastic map ϕ given by (88), there is an $S>0$ such that for $s_0\in[0,S)$,*

$$\eta_{s_0}^{\text{Riem}}(\phi)=\frac{\alpha^2}{1-\left(\frac{1-s_0}{1+s_0}\right)^2\tau^2}.$$

Furthermore,

$$\eta_{(w-1)^2}^{\text{Riem}}(\phi)=\frac{\alpha^2}{1-\tau^2}<\alpha=\eta^{\text{Dobrushin}}(\phi).$$

If $s_1\in(0,S)$, we have $\eta_{s_1}^{\text{Riem}}(\phi)>\eta_{s_0}^{\text{Riem}}(\phi)=\alpha^2/1-\tau^2$.

ACKNOWLEDGMENTS

Portions of the this work were done while the second author was an Affiliate of Department of Physics at Harvard University; she is grateful to Professor Arthur Jaffe for his hospitality. A.L.W. was partially supported by the National Science Foundation, Grant No. DMS-94-24344 while at the Department of Physics, Harvard University, Cambridge, Massachusetts, and M.B.R. was supported by the National Science Foundation, Grants No. DMS-94-08903 and No. DMS-97-06981.

[1] J. E. Cohen, Y. Iwasa, Gh. Rautu, M. B. Ruskai, E. Seneta, and Gh. Zbaganu, ''Relative entropy under mappings by stochastic matrices,'' Linear Algebr. Appl. **179**, 211–235 (1993).
[2] M. D. Choi, M. B. Ruskai, and E. Seneta, ''Equivalence of certain entropy contraction coefficients,'' Linear Algebr. Appl. **208/209**, 29–36 (1993).
[3] M. B. Ruskai, ''Beyond strong subadditivity? Improved bounds on the contraction of generalized relative entropy,'' Rev. Math. Phys. **6**, 1147–1161 (1994).
[4] L. L. Campbell, ''An extended Čencov characterization of the information metric,'' Proc. Am. Math. Soc. **98**, 135–141 (1986).
[5] N. N. Čencov, *Statistical Decision Rules and Optimal Inferences*, Transl. Math. Monogr. **53** (American Mathematical Society, Providence, 1982).
[6] E. A. Morozova and N. N. Čencov, ''Markov invariant geometry on state manifolds,'' Itogi Nauk Tekhn **36**, 69–102 (1990) (in Russian).
[7] D. Petz, ''Information geometry on quantum states,'' Hungarian Academy of Sciences preprint No. 39, 1995, in *Quantum Probability Communications*, edited by R. L. Hudson *et al.* (World Scientific, Singapore, 1998.), Vol. 10, pp. 135–158.
[8] D. Petz, ''Monotone metrics on matrix spaces,'' Linear Algebr. Appl. **244**, 81–96 (1996).
[9] D. Petz and Cs. Sudar, ''Geometries of quantum states,'' J. Math. Phys. **37**, 2662–2673 (1996).
[10] D. Petz and Cs. Sudar, ''Extending the Fisher metric to density matrices,'' preprint.
[11] D. Petz and M. B. Ruskai, ''Contraction of generalized relative entropy under stochastic mappings,'' Quantum Probab. Infinite Dimensional Anal. **1**, 83–89 (1998).
[12] H. Araki, ''Relative entropy of state of von Neumann algebras,'' Publ. RIMS Kyoto Univ. **9**, 809–833 (1976).
[13] H. Araki, ''Recent progress in entropy and relative entropy,'' *Proceedings of the VIIIth International Congress on Mathematical Physics*, edited by R. Seneor and M. Mebkhout (World Scientific, Singapore, 1987), pp. 354–365.
[14] O. Bratteli and D. Robinson, *Operator Algebras and Quantum Statistical Mechanics* (Springer-Verlag, Berlin, 1981).
[15] M. Ohya and D. Petz, *Quantum Entropy and Its Use* (Springer-Verlag, Berlin, 1993).
[16] D. Petz, ''Quasi-entropies for finite quantum systems'' Rep. Math. Phys. **23**, 57–65 (1986).

[17] W. F. Stinespring, ''Positive functions on C^*-algebras,'' Proc. Am. Math. Soc. **6**, 211–216 (1955).
[18] W. B. Arveson, ''Subalgebras of C^*-algebras,'' Acta Math. **123**, 141–224 (1969).
[19] M.-D. Choi, ''Completely positive linear maps on complex matrices,'' Linear Algebr. Appl. **10**, 285–290 (1975).
[20] K. Kraus, *States, Effects and Operations: Fundamental Notions of Quantum Theory* (Springer-Verlag, Berlin, 1983).
[21] G. Lindblad, ''Completely positive maps and entropy inequalities,'' Commun. Math. Phys. **40**, 147–151 (1975).
[22] E. Lieb and M. B. Ruskai, ''Some operator inequalities of the Schwarz type,'' Adv. Math. **12**, 269–273 (1974).
[23] J. E. Cohen and Gh. Zbaganu, *Comparison of Stochastic Matrices* (Birkhauser, Basel, 1998).
[24] L. K. Jones and C. L. Byrne, ''General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis,'' IEEE Trans. Inf. Theory **36**, 23–30 (1990).
[25] V. Vedral and M. B. Plenio, ''Entanglement measures and purification procedures,'' Phys. Rev. A **57**, 1619–1633 (1998).
[26] F. Kubo and T. Ando, ''Means of positive linear operators,'' Math. Ann. **246**, 205–224 (1980).
[27] E. Lieb, ''Convex trace functions and the Wigner–Yanase–Dyson conjecture,'' Adv. Math. **11**, 267–288 (1973).
[28] E. Lieb and M. B. Ruskai, ''Proof of the strong subadditivity of quantum mechanical entropy,'' J. Math. Phys. **14**, 1938–1941 (1973).
[29] A. Wehrl, ''General properties of entropy,'' Rev. Mod. Phys. **50**, 221–260 (1978).
[30] A. Uhlmann, ''Relative entropy and the Wigner–Yanase–Dyson–Lieb concavity in an interpolation theory,'' Rev. Math. Phys. **54**, 21–32 (1977).
[31] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. 1.
[32] J. Dittman and A. Uhlmann, ''Connections and matrics respecting standard purification'' preprint.
[33] A. Uhlmann, ''The metric of Bures and the geometric phase,'' in *Groups and Related Topics*, edited by R. Gielerak *et al.* (Kluwer Academic, City, 1992).
[34] A. Uhlmann, ''Density operators as an arena for differential geometry,'' Rep. Math. Phys. **33**, 253–263 (1993).
[35] A. Uhlmann, ''Spheres and hemispheres as quantum state spaces,'' Erwin Schrödinger Institute preprint No. 102, Vienna, 1994.
[36] A. Uhlmann, ''Geometric phases and related structures,'' Rep. Math. Phys. **36**, 461–481 (1995).
[37] R. Ahlswede and P. Gacs, ''Spreading of sets in product spaces and hypercontraction of the Markov operator,'' Ann. Prob. **4**, 925–939 (1976).
[38] M. B. Ruskai, ''Contraction of Riemannian metrics and relative entropy on the algebra of 2×2 matrices,'' preprint.

Theory of separability of multi-Hamiltonian chains

Maciej Błaszak
Physics Department, A. Mickiewicz University, Umultowska 85, 61-614 Poznan, Poland

(Received 12 February 1999; accepted for publication 21 July 1999)

The theory of separability of one-Casimir bi-Hamiltonian chains is extended onto unsplit multi-Casimir bi-Hamiltonian chains. Multi-Casimir extensions of the known one-Casimir chains are constructed. © *1999 American Institute of Physics.* [S0022-2488(99)01811-3]

I. INTRODUCTION

In the last decade considerable progress has been made in construction of new integrable finite dimensional dynamical systems showing bi-Hamiltonian property. The majority of them originate from stationary flows, restricted flows or nonlinearization of Lax equations of underlying soliton systems (see Ref. 1 and the literature quoted there). Quite recently a fundamental property of such systems has been discovered, i.e., their separability. It was proved[1,2,3] that most bi-Hamiltonian finite dimensional chains, which start with a Casimir of the first Poisson structure and terminate with a Casimir of the second Poisson structure, are integrable by quadratures, through the solutions of the appropriate Hamilton–Jacobi equation. In this paper we develop a theory generalizing them into multi-Hamiltonian systems with more than one Casimir variable. The results are illustrated by several examples of the known and new multi-Hamiltonian systems whose integrability by quadratures is demonstrated.

II. PRELIMINARY CONSIDERATIONS

Let us reexamine some facts about bi-Hamiltonian systems, both finite and infinite dimensional. We recall some definitions. Let M be a differentiable manifold, TM and T^*M its tangent and cotangent bundle. At any point $u \in M$, the tangent and cotangent spaces are denoted by T_uM and T_u^*M, respectively. The pairing between them is given by the map $\langle \cdot , \cdot \rangle : T_u^*M \times T_uM \to R$. For each smooth function $F \in C^\infty(M)$, dF denotes the differential of F (gradient ∇F for finite systems and variation δF for field systems). M is said to be a Poisson manifold if it is endowed with a Poisson bracket $\{\cdot,\cdot\}: C^\infty(M) \times C^\infty(M) \to C^\infty(M)$, in general degenerated. The related Poisson tensor π is defined by $\{F,G\}_\pi(u) := \langle dG, \pi \circ dF \rangle(u) = \langle dG(u), \pi(u) dF(u) \rangle$. So, at each point u, $\pi(u)$ is a linear map $\pi(u)$: $T_u^*M \to T_uM$ which is skew-symmetric and fulfills the Jacobi identity. Any function $c \in C^\infty(M)$, such that $dc \in \ker \pi$, is called a Casimir of π. Let π_0, π_1: $T^*M \to TM$ be two Poisson tensors on M. A vector field K is said to be a bi-Hamiltonian with respect to π_0 and π_1 if there exist two smooth functions $H, F \in C^\infty(M)$ such that

$$K = \pi_0 \circ dH = \pi_1 \circ dF. \tag{1}$$

Poisson tensors π_0 and π_1 are said to be compatible if the associated pencil $\pi_\lambda = \pi_1 - \epsilon \pi_0$ is itself a Poisson tensor for any λ. Moreover, if π_0 is invertible, the tensor $N = \pi_1 \circ \pi_0^{-1}$, called a recursion operator, is a Nijenhuis (hereditary) tensor of such a property that when it acts on a given bi-Hamiltonian vector field K, it produces another bi-Hamiltonian vector field being a symmetry generator of K. Hence, having the invariant Nijenhuis tensor, one can construct a hierarchy of Hamiltonian symmetries and related hierarchy of constants of motion for underlying system, so important for its integrability.

Unfortunately, for majority of bi-Hamiltonian finite dimensional systems, both Poisson structures are degenerated, so one cannot construct the recursion Nijenhuis tensor inverting one of the Poisson structures. Nevertheless, due to the nonuniqueness of Hamiltonian functions, determined up to an appropriate Casimir function, it is always possible to construct a finite bi-Hamiltonian chain starting and terminating with Casimirs of π_0 and π_1, respectively. Actually, as was first shown by Gel'fand and Zakharevich,[4] having Poisson manifold M of $\dim M = 2n+1$ with a linear Poisson pencil $\pi_\lambda = \pi_1 - \lambda \pi_0$ of maximal rank, a Casimir of the pencil is a polynomial in λ of order n

$$h_\lambda = h_0 \lambda^n + h_1 \lambda^{n-1} + \cdots + h_n \tag{2}$$

and generates a bi-Hamiltonian chain

$$\pi_\lambda \nabla h_\lambda = 0 \Leftrightarrow \begin{array}{c} \pi_0 \circ \nabla h_0 = 0 \\ \pi_0 \circ \nabla h_1 = K_1 = \pi_1 \circ \nabla h_0 \\ \pi_0 \circ \nabla h_2 = K_2 = \pi_1 \circ \nabla h_1 \\ \vdots \\ \pi_0 \circ \nabla h_n = K_n = \pi_1 \circ \nabla h_{n-1} \\ 0 = \pi_1 \circ \nabla h_n, \end{array} \tag{3}$$

where $K \equiv K_1$, $H \equiv h_1$, $F \equiv h_0$, and $d \equiv \nabla$. References 1–3 demonstrate that in canonical coordinates (q,p,c), where $q = (q_1,\ldots,q_n)^T$ and $p = (p_1,\ldots,p_n)^T$ are generalized coordinates and c is a Casimir coordinate, a canonical π_0 and noncanonical π_1 Poisson structure (both degenerated) take the general form

$$\pi_0 = \begin{pmatrix} \theta_0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \pi_1 \begin{pmatrix} \theta_1 & \overline{K} \\ -\overline{K}^T & 0 \end{pmatrix}, \tag{4}$$

where

$$\theta_0 = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \theta_1 = \begin{pmatrix} D & A \\ -A^T & B \end{pmatrix} \tag{5}$$

are nondegenerated Poisson matrices from the space $\overline{M} \ni (q,p)$, $\overline{K} = \overline{K}(q,p,c)$ is related to K_1 vector field from the hierarchy (5) in the following way: $K_1 = (\overline{K},0)^T$ and A,B,D are entries depending on the generalized coordinates (q,p). Moreover, the operator

$$N = \theta_1 \circ \theta_0^{-1} = \begin{pmatrix} A & -D \\ B & A^T \end{pmatrix} \tag{6}$$

is a Nijenhuis tensor on $\overline{M}$ but not a recursion operator for the projected vector fields $\overline{K}_r \in T\overline{M}$.

Then, it was shown that each hierarchy (3), where θ_1 takes the form

$$\theta_1 = \begin{pmatrix} 0 & A(q) \\ -A^T(q) & B(q,p) \end{pmatrix}, \tag{7}$$

or can be transformed into the form (7) by an appropriate canonical transformation, with Hamiltonian functions

$$h_k(q,p,c) = h_k(q,p) + c b_k(q), \tag{8}$$

is separable. The expression for θ_1 given by (7) is the most general form of the second Poisson tensor when the change from natural coordinates to separation coordinates is given by a point transformation. Actually, first one has to pass to the so-called Darboux–Nijenhuis coordinates (λ_i,μ_i) via a canonical transformation $\phi:(\lambda,\mu)\rightarrow(q,p)$. The transformation ϕ is constructed by a generating function $S=\Sigma_{i=1}^{n}p_i\sigma_i(\lambda)$ from the equations

$$q_i=\frac{\partial S}{\partial p_i},\quad \mu_i=\frac{\partial S}{\partial \lambda_i}, \tag{9}$$

where $q_k=\sigma_k(\lambda)$ are calculated from

$$b_k(q)=(-1)^k\sum_{\substack{j_1,\ldots,j_k\\ j_1<..<j_k}}\lambda_{j_1}\cdot\cdots\cdot\lambda_{j_k}:=\rho_k(\lambda),\quad k=1,\ldots,n, \tag{10}$$

and $b_0(q)=1:=\rho_0(\lambda)$, where $\rho_k(\lambda)$ are the so-called Viete polynomials (symmetric polynomials). In Darboux–Nijenhuis coordinates the chain (3) transforms into a bi-Hamiltonian Nijenhuis chain (3) where now

$$\pi_0=\begin{pmatrix}0 & I & 0\\ -I & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\quad \pi_1=\begin{pmatrix}0 & \Lambda & \dfrac{\partial h_1}{\partial \mu}\\ -\Lambda & 0 & -\dfrac{\partial h_1}{\partial \lambda}\\ -\left(\dfrac{\partial h_1}{\partial \mu}\right)^T & \left(\dfrac{\partial h_1}{\partial \lambda}\right)^T & 0\end{pmatrix},\quad \Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_n),$$
$$h_k(\lambda,\mu,c)=-\sum_{i=1}^{n}\frac{\partial\rho_k(\lambda)}{\partial\lambda_i}\frac{f_i(\lambda_i,\mu_i)}{\Delta_i(\lambda)}+c\rho_k(\lambda)=h_k(q,p,c)\cdot\phi, \tag{11}$$

$\rho_k(\lambda)$ are given by (10),

$$\Delta_i(\lambda):=\prod_{j\neq i}(\lambda_i-\lambda_j) \tag{12}$$

and $f_i(\lambda_i,\mu_i)$ are arbitrary smooth functions.

Then, it was demonstrated that for each $h_k(\lambda,\mu,c)$ there exists a canonical transformation $(\lambda,\mu)\rightarrow(b,a)$ in the form $b_i=\partial W/\partial a_i$, $\mu_i=\partial W/\partial\lambda_i$, where $W(\lambda,a)$ is the generating function, such that the related Hamilton–Jacobi equations

$$h_r\left(\lambda,\frac{\partial W}{\partial\lambda},c\right)=\mathrm{const} \tag{13}$$

can be solved and hence, the implicit solutions for the trajectories $\lambda_i(t_r)$, with respect to the evolution parameter t_r, can be constructed.

In the next section we extend the formalism onto unsplit multi-Casimir chains, i.e., the case with more then one Casimir variable, when each bi-Hamiltonian chain conserves the form (3) and does not split into two or more sub-chains. So, the theory presented does not cover such examples as stationary Boussinesq flows.[5] The following properties of Viete $\rho_r(\lambda)$ and $\Delta_i(\lambda)$ polynomials will be useful[1,2]

$$\frac{\partial\rho_r}{\partial\lambda_i}:=-\rho_{r-1}^{i},\quad \lambda_i\frac{\partial\rho_r}{\partial\lambda_i}=\rho_r-\rho_r^{i},\quad \lambda_i\rho_{r-1}^{i}=-\rho_r+\rho_r^{i}, \tag{14}$$

$$\frac{\partial\rho_r^k}{\partial\lambda_i}:=-\rho_{r-1}^{ki}, \quad \lambda_i\frac{\partial\rho_r^k}{\partial\lambda_i}=\rho_r^k-\rho_r^{ki}, \quad \lambda_i\rho_{r-1}^{ki}=-\rho_r^k+\rho_r^{ki},$$
$$\frac{\partial}{\partial\lambda_i}\frac{1}{\Delta_i}=-\frac{1}{\Delta_i}\sum_{\alpha\neq i}\frac{1}{\lambda_{i\alpha}}, \quad \frac{\partial}{\partial\lambda_\beta}\frac{1}{\Delta_i}=\frac{1}{\Delta_i}\frac{1}{\lambda_{i\beta}}, \quad \beta\neq i. \tag{15}$$

III. MULTI-HAMILTONIAN NIJENHUIS CHAINS AND THEIR INTEGRABILITY BY QUADRATURES

Consider the set of Hamiltonian functions

$$h_r(\lambda,\mu,c)=-\sum_{i=1}^n\frac{\partial\rho_r}{\partial\lambda_i}\frac{f_i(\lambda_i,\mu_i)}{\Delta_i}+\sum_{j=1}^n c_j\beta_{j,r}(\lambda), \quad r=1,...,n, \tag{16}$$

where $\beta_{1,r}(\lambda)\equiv\rho_r(\lambda)$ and $\beta_{m,r}$, $m=2,...,n$, are defined by the recursive formula

$$\beta_{m,r}=\beta_{m-1,r+1}-\beta_{m-1,1}\cdot\beta_{1,r}. \tag{17}$$

The explicit forms of $\beta_{j,r}$ functions for $n=2,3$ are given in Appendix A.

In this section we prove that functions (16) form $\binom{n+1}{2}$ bi-Hamiltonian chains with respect to $(n+1)$ appropriate Poisson matrices, and then we solve the related Hamilton–Jacobi equations, confirming the separability of dynamical systems generated by Hamiltonians (16). We start with proofs of some properties of h_r functions and their components.

Lemma 1: (i) The following representation of $\beta_{k,r}$ is valid:

$$\beta_{k,r}(\lambda)=\sum_{i=1}^n\frac{\partial\rho_r}{\partial\lambda_i}\frac{\lambda_i^{n+k-1}}{\Delta_i}=-\sum_{i=1}^n\rho_{r-1}^i\frac{\lambda_i^{n+k-1}}{\Delta_i}. \tag{18}$$

(ii) The following property of $\beta_{k,r}$ is fulfilled:

$$\frac{\partial\beta_{r,k+1}}{\partial\lambda_i}=-\frac{\partial\rho_{k+1}}{\partial\lambda_i}\frac{\partial\beta_{r,1}}{\partial\lambda_i}=\rho_k^i\frac{\partial\beta_{r,1}}{\partial\lambda_i}. \tag{19}$$

The proof is given in Appendix B.

Lemma 2: Let $R=\Sigma_{j=1}^n[\lambda_j^2(\partial/\partial\lambda_j)+\lambda_j]$, then the following relations hold for functions $\beta_{k,1}$: (i)

$$R\cdot\beta_{k-1,1}=k\beta_{k,1}\Rightarrow\frac{1}{k!}R^k\cdot\beta_{0,1}=\beta_{k,1}, \quad \beta_{0,1}=1, \tag{20}$$

(ii)

$$\beta_{k,1}(\lambda)=-\sum_{\substack{i_1,...,i_k\\ i_1\leq...\leq i_k}}\lambda_{i_1}\lambda_{i_2}\dots\lambda_{i_k}, \quad k\leq n. \tag{21}$$

The proof is inductive, using the results of Lemma 1.

Notice that according to Lemma 1 (i), functions h_r can be put into the form

$$h_r(\lambda,\mu,c)=-\sum_{j=1}^n\frac{\partial\rho_r}{\partial\lambda_j}\frac{\overline{f}_j(\mu_j,\lambda_j,c)}{\Delta_j}, \quad \overline{f}_j(\mu_j,\lambda_j,c)=f_j(\mu_j,\lambda_j)-\sum_{l=1}^n c_l\lambda_j^{n-1+l}. \tag{22}$$

Lemma 3: The respective derivatives of h_r read

$$
\frac{\partial h_r}{\partial \mu_i} = -\frac{\partial \rho_r}{\partial \lambda_i}\frac{\partial h_1}{\partial \mu_i} = \rho^i_{r-1}\frac{\partial h_1}{\partial \mu_i},
$$

$$
\frac{\partial h_r}{\partial \lambda_i} = -\frac{\partial \rho_r}{\partial \lambda_i}\frac{\partial h_1}{\partial \lambda_i} = \rho^i_{r-1}\frac{\partial h_1}{\partial \lambda_i}, \tag{23}
$$

$$
\frac{\partial^2 h_r}{\partial \lambda_i \partial \lambda_j} = -\frac{1}{\lambda_i-\lambda_j}\frac{\partial \rho_r}{\partial \lambda_j}\frac{\partial h_1}{\partial \lambda_i} - \frac{1}{\lambda_j-\lambda_i}\frac{\partial \rho_r}{\partial \lambda_i}\frac{\partial h_1}{\partial \lambda_j} = \frac{\rho^j_{r-1}}{\lambda_i-\lambda_j}\frac{\partial h_1}{\partial \lambda_i} + \frac{\rho^i_{r-1}}{\lambda_j-\lambda_i}\frac{\partial h_1}{\partial \lambda_j}.
$$

The proof is given in Appendix C.

Now, let us consider $n+1$ matrices $3n\times 3n$ of rank $2n$ in the following form:

$$
\pi_0 = \left(\begin{array}{c|c}
\begin{array}{cc} 0 & I \\ -I & 0 \end{array} & \begin{array}{c|c|c} 0 & \cdots & 0 \end{array} \\
\hline
\begin{array}{c} 0 \\ \hline \vdots \\ \hline 0 \end{array} & 0
\end{array}\right),
$$

$$
\pi_1 = \left(\begin{array}{c|c}
\begin{array}{cc} 0 & \Lambda \\ -\Lambda & 0 \end{array} & \begin{array}{c|c|c|c} K_1 & 0 & \cdots & 0 \end{array} \\
\hline
\begin{array}{c} -K_1^T \\ \hline 0 \\ \hline \vdots \\ \hline 0 \end{array} & 0
\end{array}\right)
$$

$$
\pi_2 = \left(\begin{array}{c|c}
\begin{array}{cc} 0 & \Lambda^2 \\ -\Lambda^2 & 0 \end{array} & \begin{array}{c|c|c|c|c} K_2 & K_1 & 0 & \cdots & 0 \end{array} \\
\hline
\begin{array}{c} -K_2^T \\ \hline -K_1^T \\ \hline 0 \\ \hline \vdots \\ \hline 0 \end{array} & 0
\end{array}\right)
$$

$$
\vdots \tag{24}
$$

$$
\pi_n = \left(\begin{array}{c|c}
\begin{array}{cc} 0 & \Lambda^n \\ -\Lambda^n & 0 \end{array} & \begin{array}{c|c|c|c} K_n & K_{n-1} & \cdots & K_1 \end{array} \\
\hline
\begin{array}{c} -K_n^T \\ \hline -K_{n-1}^T \\ \hline \vdots \\ \hline -K_1^T \end{array} & 0
\end{array}\right),
$$

where I is $n\times n$ unit matrix, $\Lambda^m=\operatorname{diag}(\lambda_1^m,\lambda_2^m,...,\lambda_n^m)$ where all λ_i are different and $K_m=(\partial h_m/\partial\mu_1,\ldots,\partial h_m/\partial\mu_n,-\partial h_m/\partial\lambda_1,\ldots,-\partial h_m/\partial\lambda_n)^T$ are Hamiltonian vector fields with Hamiltonians h_r (16).

Lemma 4: All functions $h_r(\lambda,\mu,c)$ are in involution with respect to canonical Poisson tensor π_0.

Proof: According to Lemma 1, for a given r, all terms $c_j\beta_{j,r}(\lambda)$ $j=1,...,n$ can be absorbed by f_i $i=1,...,n$ terms of h_r. However, according to the results from Refs. 1 and 2, functions of the form (22) are in involution with respect to π_0. □

Theorem 1: (i) Consider two Poisson tensors $3n\times 3n$ of rank $2n$

$$\pi_0=\left(\begin{array}{cc|c|c|c} 0 & I & 0 & \cdots & 0 \\ -I & 0 & & & \\ \hline \multicolumn{2}{c|}{0} & & & \\ \multicolumn{2}{c|}{\vdots} & & 0 & \\ \multicolumn{2}{c|}{0} & & & \end{array}\right), \tag{25}$$

$$\pi_m=\left(\begin{array}{cc|c|c|c|c|c|c} 0 & \Lambda^m & X_m & \cdots & X_1 & 0 & \cdots & 0 \\ -\Lambda^m & 0 & Y_m & & Y_1 & & & \\ \hline -X_m^T & -Y_m^T & & & & & & \\ \multicolumn{2}{c|}{\vdots} & & & & & & \\ -X_1^T & -Y_1^T & & & 0 & & & \\ \multicolumn{2}{c|}{0} & & & & & & \\ \multicolumn{2}{c|}{\vdots} & & & & & & \\ \multicolumn{2}{c|}{0} & & & & & & \end{array}\right).$$

The compatibility condition between π_0 and π_m is given by

$$\frac{\partial Y_{rj}}{\partial\mu_i}+\frac{\partial X_{ri}}{\partial\lambda_j}=0,\quad i,j=1,...,n,\quad r=1,...,m \tag{26}$$

with admissible solution

$$X_{ri}=\frac{\partial f_r}{\partial\mu_i},\quad Y_{ri}=-\frac{\partial f_r}{\partial\lambda_i},\quad f_r=f_r(\lambda,\mu,c). \tag{27}$$

(ii) Conditions for f_r functions to make π_m Poissonian follow from Jacobi identity and read

$$0=\{\{\lambda_i,\lambda_j\}_{\pi_m},c_k\}_{\pi_m}+\text{c.p.}$$

$$\Updownarrow \tag{28}$$

$$0=(\lambda_j^m-\lambda_i^m)\frac{\partial^2 f_{m+1-k}}{\partial\mu_i\partial\mu_j}+\sum_{r=1}^m\left(\frac{\partial^2 f_{m+1-k}}{\partial\mu_i\partial c_r}\frac{\partial f_{m+1-r}}{\partial\mu_j}-\frac{\partial^2 f_{m+1-k}}{\partial\mu_j\partial c_r}\frac{\partial f_{m+1-r}}{\partial\mu_i}\right),$$

$$0=\{\{\mu_i,\mu_j\}_{\pi_m},c_k\}_{\pi_m}+\text{c.p.}$$

$$\Updownarrow \tag{29}$$

$$0=(\lambda_j^m-\lambda_i^m)\frac{\partial^2 f_{m+1-k}}{\partial\lambda_i\partial\lambda_j}+\sum_{r=1}^{m}\left(\frac{\partial^2 f_{m+1-k}}{\partial\lambda_i\partial c_r}\frac{\partial f_{m+1-r}}{\partial\lambda_j}-\frac{\partial^2 f_{m+1-k}}{\partial\lambda_j\partial c_r}\frac{\partial f_{m+1-r}}{\partial\lambda_i}\right),$$

$$0=\{\{\lambda_i,\mu_j\}_{\pi_m},c_k\}_{\pi_m}+\text{c.p.}$$

$$\Updownarrow \tag{30}$$

$$0=(\lambda_j^m-\lambda_i^m)\frac{\partial^2 f_{m+1-k}}{\partial\mu_i\partial\lambda_j}-m\lambda_j^{m-1}\delta_{ij}\frac{\partial f_{m+1-k}}{\partial\mu_i}$$

$$+\sum_{r=1}^{m}\left(\frac{\partial^2 f_{m+1-k}}{\partial\mu_i\partial c_r}\frac{\partial f_{m+1-r}}{\partial\lambda_j}-\frac{\partial^2 f_{m+1-k}}{\partial\lambda_j\partial c_r}\frac{\partial f_{m+1-r}}{\partial\mu_i}\right),$$

$$0=\{\{c_i,c_j\}_{\pi_m},\lambda_k(\mu_k)\}_{\pi_m}+\text{c.p.}$$

$$\Updownarrow$$

$$0=\frac{\partial}{\partial\mu_k(\lambda_k)}\sum_{l=1}^{m}\left(\frac{\partial f_{m+1-i}}{\partial\lambda_l}\frac{\partial f_{m+1-j}}{\partial\mu_l}-\frac{\partial f_{m+1-i}}{\partial\mu_l}\frac{\partial f_{m+1-j}}{\partial\lambda_l}\right),$$

$$\Updownarrow \tag{31}$$

$$0=\{f_i,f_j\}_{\pi_m},\quad i,j=1,...,m.$$

Lemma 5: Functions $f_r=h_r(\lambda,\mu,c)$ (16) fulfill the conditions of Theorem 1. The proof is given in Appendix D.

Consequently, all matrices (24) are Poissonian and compatible with the canonical one.

Theorem 2: On the extended phase space $M\ni(\lambda_1,\ldots,\lambda_n,\mu_1,\ldots,\mu_n,c_1,\ldots,c_n)$, the functions

$$h_r(\lambda,\mu,c)=-\sum_{i=1}^{n}\frac{\partial\rho_r}{\partial\lambda_i}\frac{f_i(\lambda_i,\mu_i)}{\Delta_i}+\sum_{j=1}^{n}c_j\beta_{j,r}(\lambda),$$

$$h_{1-r}=c_r,\quad r=1,...,n, \tag{32}$$

form $\binom{n+1}{2}$ bi-Hamiltonian chains

$$\begin{array}{l}\pi_i\nabla h_{-i}=0\\ \pi_i\nabla h_{-i+1}=K_1=\pi_k\nabla h_{-k+1}\\ \pi_i\nabla h_{-i+2}=K_2=\pi_k\nabla h_{-k+2}\\ \qquad\vdots\\ \pi_i\nabla h_{-i+j}=K_j=\pi_k\nabla h_{-k+j}\\ \qquad\vdots\\ \pi_i\nabla h_{-i+n}=K_n=\pi_k\nabla h_{-k+n}\\ \qquad 0=\pi_k\nabla h_{-k+n+1}\end{array}\qquad 0\leqslant i<k\leqslant n, \tag{33}$$

with respect to $(n+1)$ Poisson tensors (24).

The proof is given in Appendix E.

From the fact that $h_{n+1}=0$ and the validity of the chain (33) we find that the Poisson structure π_m has the following Casimir functions: $c_{m+1}, c_{m+2}, \ldots, c_n, h_n, \ldots, h_{n-m+1}$.

Lemma 6: Functions $h_r(\lambda,\mu,c)$, $r=1,\ldots,n$ are in involution with respect to an arbitrary Poisson tensor π_k, $k=1,\ldots,n$.

Proof: As all h_r are in involution with respect to π_0 and belong to chains (33), hence

$$\{h_i,h_j\}_{\pi_k}=\langle\nabla h_j,\pi_k\nabla h_i\rangle=\langle\nabla h_j,\pi_0\nabla h_{i+k}\rangle=\{h_{i+k},h_j\}_{\pi_0}=0. \tag{34}$$

□

Now, we integrate equations of motion from the hierarchy (33) solving the Hamilton–Jacobi equation for Hamiltonians (16). According to the Hamilton–Jacobi theory, we look for a canonical transformation $(\lambda,\mu)\rightarrow(b,a)$ in the form $b_i=\partial W/\partial a_i$, $\mu_i=\partial W/\partial\lambda_i$, where $W(\lambda,a)$ is the generating function, satisfying the Hamilton–Jacobi equation

$$h_r\left(\lambda,\frac{\partial W}{\partial\lambda},c\right)=\sum_{k=1}^{n}\frac{\rho_{r-1}^k(\lambda)f_k(\lambda_k,\partial W/\partial\lambda_k)}{\Delta_k}+\sum_{i=1}^{n}c_i\beta_{i,r}(\lambda)=\text{const}_r. \tag{35}$$

Now we demonstrate the separability of Eq. (35). Take the generating $W(\lambda,a)$ function in the form

$$W(\lambda,a)=\sum_{i=1}^{n}W_i(\lambda_i,a), \tag{36}$$

apply the result of Lemma 1 and hence, Eq. (35) turns into the form

$$\sum_{k=1}^{n}\frac{\rho_{r-1}^k(\lambda)[f_k(\lambda_k,\partial W_k/\partial\lambda_k)-\Sigma_{i=1}^n c_i\lambda_k^{n-1+i}]}{\Delta_k}=\text{const}_r. \tag{37}$$

Applying the relation

$$-\sum_{i=1}^{n}\frac{\partial\rho_r}{\partial\lambda_i}\frac{\lambda_i^m}{\Delta_i}=\sum_{i=1}^{n}\frac{\rho_{r-1}^i(\lambda)\lambda_i^m}{\Delta_i}=\begin{Bmatrix}1, & m=n-r\\ 0, & m\neq n-r\end{Bmatrix},\quad r=1,\ldots,n, \tag{38}$$

the solution of Eq. (37) reads

$$f_k(\lambda_k,\partial W_k/\partial\lambda_k)=g(\lambda_k),\quad \text{const}_r=a_{n+1-r}, \tag{39}$$

where

$$g(\xi)=a_1+a_2\xi+\cdots+a_n\xi^{n-1}+c_1\xi^n+\cdots+c_n\xi^{2n-1}. \tag{40}$$

Hence, $W(\lambda,a)$ can be obtained by solving n decouples first-order ODE (39). For example, if

$$f_i(\lambda_i,\mu_i)=\varphi(\lambda_i)f(\mu_i)+\psi(\lambda_i), \tag{41}$$

then we obtain

$$W(\lambda,a)=\sum_{k=1}^{n}\int^{\lambda_k}f^{-1}\left(\frac{g(\xi)-\psi(\xi)}{\varphi(\xi)}\right)d\xi. \tag{42}$$

In new canonical variables a_i, $b_i=\partial W/\partial a_i$, the Hamiltonians $h_r(\lambda,\mu,c)$ become

$$h_r = a_{n+1-r} \tag{43}$$

with

$$b_i = \frac{\partial W}{\partial a_i} = \sum_{k=1}^{n} \int^{\lambda_k} (f^{-1})' \frac{\xi^{i-1}}{\varphi(\xi)} d\xi, \tag{44}$$

where $(f^{-1})'$ means derivative of f^{-1}. As in the new coordinates each h_r generates a trivial flow

$$(a_j)_{t_r} = -\frac{\partial h_r}{\partial b_j} = 0, \quad (b_j)_{t_r} = \frac{\partial h_r}{\partial a_j} = \delta_{j,n+1-r}, \quad (c_j)_{t_r} = 0, \tag{45}$$

hence

$$b_i = t_{n+1-i} + \text{const.} \tag{46}$$

Combining (44) with (46) we arrive at implicit solutions for the trajectories $\lambda_i(t_k)$, with respect to the evolution parameter t_k in the form

$$\sum_{k=1}^{n} \int^{\lambda_k} (f^{-1})' \frac{\xi^{i}}{\varphi(\xi)} d\xi = \delta_{i,n-k} t_k + \text{const}, \quad i = 0,...,n-1. \tag{47}$$

IV. MULTI-HAMILTONIAN CHAINS IN ARBITRARY CANONICAL COORDINATES AND THEIR TRANSFORMATION TO DARBOUX–NIJENHUIS REPRESENTATION

Let us introduce arbitrary coordinates (q,p,c) related to the Darboux–Nijenhuis coordinates (λ,μ,c) by a point transformation

$$q_k = \zeta_k(\lambda), \quad k = 1,...,n, \tag{48}$$

which is canonical, i.e., coming from the generating function $S = \Sigma_{i=1}^{n} p_i \zeta_k(\lambda)$. Then

$$q_i = \frac{\partial S}{\partial p_i}, \quad \mu_i = \frac{\partial S}{\partial \lambda_i}, \tag{49}$$

so the first equation reconstructs $q_i = \zeta_k(\lambda)$ and solving the second one with respect to p_k we get the missing part of the canonical transformation

$$p_k = \eta_k(\lambda,\mu). \tag{50}$$

Applying the inverse of the transformation (48), (50) to Hamiltonian functions (16) and Poisson matrices (24) one finds that

$$h_r(q,p,c) = h_r(q,p) + \sum_{i=1}^{n} c_i b_{i,r}(q) \tag{51}$$

and the nondegenerated part θ_m of rank $2n$ of each π_m (also implectic)

$$\theta_m = N^m \circ \theta_0 = \begin{pmatrix} 0 & \Lambda^m \\ -\Lambda^m & 0 \end{pmatrix} \tag{52}$$

takes now the form

$$\theta_m = N^m \circ \theta_0 = \begin{pmatrix} 0 & A_m(q) \\ -A_m^T(q) & B_m(q,p) \end{pmatrix}, \quad m=1,...,n, \tag{53}$$

where matrix elements $(B_m)_{ij}$ are at most linear in p coordinates.

Conversely, if we have a multi-Hamiltonian chain in (q,p,c) coordinates, and the set of equations

$$\beta_{1,r}(\lambda) \equiv \rho_r(\lambda) = b_{1,r}(q), \quad r=1,...,n \tag{54}$$

has the solution $\lambda_i = \lambda_i(q)$, $\quad i=1,...,n$ such that $\lambda_i(q) \neq \lambda_j(q)$ for $i \neq j$, which is equivalent to the statement that the Nijenhuis (but not recursion) tensor

$$N(q,p) = \begin{pmatrix} 0 & A_1(q) \\ -A_1^T(q) & B_1(q,p) \end{pmatrix} \cdot \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}^{-1} = \begin{pmatrix} A_1(q) & 0 \\ B_1(q,p) & A_1^T(q) \end{pmatrix} \tag{55}$$

is nondegenerated and has n distinct eigenvalues λ_i each with multiplicity 2, then the canonical transformation (48)–(50) transforms a given chain to the one considered in the preceding section.

The admissible reductions of the number of Casimir variables are the following. For arbitrary $1 \leqslant m < n$, let $c_i \neq 0$, $1 \leqslant i \leqslant m$ and $c_i = 0$ for $m < i \leqslant n$. The first m Poisson structures π_i, $i \leqslant m$ survey the projection $(\lambda,\mu,c_1,\ldots,c_n) \in M \to \bar{M} \ni (\lambda,\mu,c_1,\ldots,c_m)$ and we have still $\binom{m+1}{2}$ bi-Hamiltonian chains (33).

Remark: In the limit $c_1 = \cdots = c_n = 0$, the systems considered lose the bi-Hamiltonian property, turning into the so-called quasi-bi-Hamiltonian systems[1–3,6,7] on a symplectic manifold $M \ni (q,p)$, being still separable and integrable by quadratures. Moreover, because of the property (22), each of the multi-Hamiltonian systems considered on a Poisson manifold $M \ni (q,p,c)$ has a quasi-bi-Hamiltonian representation on a symplectic manifold $M \ni (q,p)$.

V. EXAMPLES

The theory developed in the preceding sections will be illustrated by several representative examples of already known as well as new multi-Hamiltonian systems.

Example 1: Stationary t_2—flow of dispersive water waves.[8]

The Hamiltonian functions and Poisson structures in Ostrogradsky variables are as follows:

$$h_1(q,p,c) = -4p_1p_2 + 5q_2p_1^2 - \tfrac{5}{8}q_1q_2^3 - \tfrac{3}{4}q_1^2q_2 - \tfrac{7}{64}q_2^5 + \tfrac{1}{2}q_2c_1 + (\tfrac{1}{2}q_1 + \tfrac{1}{8}q_2^2)c_2,$$

$$h_2(q,p,c) = q_1p_1^2 + 4q_2p_1p_2 - \tfrac{5}{4}q_2^2p_1^2 - 2p_2^2 + \tfrac{5}{64}q_1q_2^4 - \tfrac{3}{16}q_1^2q_2^2 - \tfrac{1}{4}q_1^3 + \frac{45}{6\times 128}q_2^6$$
$$+ (\tfrac{1}{2}q_1 + \tfrac{3}{8}q_2^2)c_1 - (\tfrac{1}{4}q_1q_2 - \tfrac{3}{16}q_2^3)c_2,$$

$$\pi_0 = \left(\begin{array}{cccc|cc} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$\pi_1=\left(\begin{array}{cccc|cc} 0 & 0 & -\frac{3}{2}q_2 & -\frac{1}{2}q_1-\frac{15}{8}q_2^2 & \frac{\partial h_1}{\partial p_1} & 0 \\ 0 & 0 & 1 & q_2 & \frac{\partial h_1}{\partial p_2} & 0 \\ \frac{3}{2}q_2 & -1 & 0 & -p_1 & -\frac{\partial h_1}{\partial q_1} & 0 \\ \frac{1}{2}q_1+\frac{15}{8}q_2^2 & -q_2 & p_1 & 0 & -\frac{\partial h_1}{\partial q_2} & 0 \\ \hline -\frac{\partial h_1}{\partial p_1} & -\frac{\partial h_1}{\partial p_2} & \frac{\partial h_1}{\partial q_1} & \frac{\partial h_1}{\partial q_2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$\pi_2=\left(\begin{array}{cccc|cc} 0 & 0 & \frac{3}{8}q_2^2-\frac{1}{2}q_1 & -\frac{1}{4}q_1q_2-\frac{15}{16}q_2^3 & \frac{\partial h_2}{\partial p_1} & \frac{\partial h_1}{\partial p_1} \\ 0 & 0 & -\frac{1}{2}q_2 & -\frac{1}{2}q_1-\frac{7}{8}q_2^2 & \frac{\partial h_2}{\partial p_2} & \frac{\partial h_1}{\partial p_2} \\ -\frac{3}{8}q_2^2+\frac{1}{2}q_1 & \frac{1}{2}q_2 & 0 & \frac{1}{2}q_2p_1 & -\frac{\partial h_2}{\partial q_1} & -\frac{\partial h_1}{\partial q_1} \\ \frac{1}{4}q_1q_2+\frac{15}{16}q_2^3 & \frac{1}{2}q_1+\frac{7}{8}q_2^2 & -\frac{1}{2}q_2p_1 & 0 & -\frac{\partial h_2}{\partial q_2} & -\frac{\partial h_1}{\partial q_2} \\ \hline -\frac{\partial h_2}{\partial p_1} & -\frac{\partial h_2}{\partial p_2} & \frac{\partial h_2}{\partial q_1} & \frac{\partial h_2}{\partial q_2} & 0 & 0 \\ -\frac{\partial h_1}{\partial p_1} & -\frac{\partial h_1}{\partial p_2} & \frac{\partial h_1}{\partial q_1} & \frac{\partial h_1}{\partial q_2} & 0 & 0 \end{array}\right).$$

Hence, we have three bi-Hamiltonian chains (33)

$$\pi_0\nabla c_1=0 \qquad \pi_0\nabla c_1=0$$

$$\pi_0\nabla h_1=K_1=\pi_1\nabla c_1 \qquad \pi_0\nabla h_1=K_1=\pi_2\nabla c_2$$

$$\pi_0\nabla h_2=K_2=\pi_1\nabla h_1\pi_0 \qquad \nabla h_2=K_2=\pi_2\nabla c_1$$

$$0=\pi_1\nabla h_2, \qquad\qquad 0=\pi_2\nabla h_1,$$

$$\pi_1\nabla c_2=0$$

$$\pi_1\nabla c_1=K_1=\pi_2\nabla c_2$$

$$\pi_1\nabla h_1=K_2=\pi_2\nabla c_1$$

$$0=\pi_2\nabla h_1.$$

The canonical transformation to the Darboux–Nijenhuis coordinates reads

$$q_1=-(3\lambda_1^2+3\lambda_2^2+4\lambda_1\lambda_2),$$

$$q_2=-2(\lambda_1+\lambda_2),$$

$$p_1=\frac{1}{2}\,\frac{\mu_2-\mu_1}{\lambda_1-\lambda_2},$$

$$p_2=-\frac{1}{2}\,\frac{\lambda_1(3\mu_2-2\mu_1)-\lambda_2(3\mu_1-2\mu_2)}{\lambda_1-\lambda_2},$$

where now h_r, $r=1,2$ take the form (16) with

$$f_i(\lambda_i,\mu_i)=2\lambda_i^6-\tfrac{1}{2}\mu_i^2.$$

Example 2: The first two-Casimir extension of the Henon–Heiles system.

In Ref. 1 the first one-Casimir extension of the Henon–Heiles system was considered in the form

$$(q_1)_{tt}=-3q_1^2-\tfrac{1}{2}q_2^2+c,\quad (q_2)_{tt}=-q_1q_2,$$

with two constants of the motion

$$h_1=H=\tfrac{1}{2}p_1^2+\tfrac{1}{2}p_2^2+q_1^3+\tfrac{1}{2}q_1q_2^2-cq_1,$$

$$h_2=\tfrac{1}{2}q_2p_1p_2-\tfrac{1}{2}q_1p_2^2+\tfrac{1}{16}q_2^4+\tfrac{1}{4}q_1^2q_2^2-\tfrac{1}{4}q_2^2c.$$

The transformation to Darboux–Nijenhuis coordinates was found in the form

$$q_1=\lambda_1+\lambda_2,\quad p_1=\frac{\lambda_1\mu_1}{\lambda_1-\lambda_2}+\frac{\lambda_2\mu_2}{\lambda_2-\lambda_1},$$

$$q_2=2\sqrt{-\lambda_1\lambda_2},\quad p_2=\sqrt{-\lambda_1\lambda_2}\left(\frac{\mu_1}{\lambda_1-\lambda_2}+\frac{\mu_2}{\lambda_2-\lambda_1}\right),$$

$$f_i(\lambda_i,\mu_i)=\lambda_i^4+\tfrac{1}{2}\lambda_i\mu_i^2. \tag{56}$$

For a two-Casimir extension we get immediately

$$\beta_{2,1}=-[(\lambda_1+\lambda_2)^2-\lambda_1\lambda_2]=-(q_1^2+\tfrac{1}{4}q_2^2),$$

$$\beta_{2,2}=\lambda_1\lambda_2(\lambda_1+\lambda_2)=-\tfrac{1}{4}q_1q_2^2,$$

hence, new constants of the motion read

$$h_1=H=\tfrac{1}{2}p_1^2+\tfrac{1}{2}p_2^2+q_1^3+\tfrac{1}{2}q_1q_2^2-c_1q_1-(q_1^2+\tfrac{1}{4}q_2^2)c_2,$$

$$h_2=\tfrac{1}{2}q_2p_1p_2-\tfrac{1}{2}q_1p_2^2+\tfrac{1}{16}q_2^4+\tfrac{1}{4}q_1^2q_2^2-\tfrac{1}{4}q_2^2c_1-\tfrac{1}{4}q_1qc_2,$$

where the Newton equations related to the energy H are

$$(q_1)_{tt}=-3q_1^2-\tfrac{1}{2}q_2^2+c_1+2q_1c_2,\quad (q_2)_{tt}=-q_1q_2+\tfrac{1}{2}q_2c_2.$$

This is tri-Hamiltonian system with the following Poisson structures

$$\pi_0=\left(\begin{array}{cccc|cc} 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ -1 & 0 & 0 & 0 & 0 & 0\\ 0 & -1 & 0 & 0 & 0 & 0\\ \hline 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$\pi_1=\left(\begin{array}{cccc|cc} 0 & 0 & q_1 & \frac{1}{2}q_2 & \frac{\partial h_1}{\partial p_1} & 0\\ 0 & 0 & \frac{1}{2}q_2 & 0 & \frac{\partial h_1}{\partial p_2} & 0\\ -q_1 & -\frac{1}{2}q_2 & 0 & \frac{1}{2}p_2 & -\frac{\partial h_1}{\partial q_1} & 0\\ -\frac{1}{2}q_2 & 0 & -\frac{1}{2}p_2 & 0 & -\frac{\partial h_1}{\partial q_2} & 0\\ \hline -\frac{\partial h_1}{\partial p_1} & -\frac{\partial h_1}{\partial p_2} & \frac{\partial h_1}{\partial q_1} & \frac{\partial h_1}{\partial q_2} & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$\pi_2=\left(\begin{array}{cccc|cc} 0 & 0 & q_1^2+\frac{1}{4}q_2^2 & \frac{1}{2}q_1q_2 & \frac{\partial h_2}{\partial p_1} & \frac{\partial h_1}{\partial p_1}\\ 0 & 0 & \frac{1}{2}q_1q_2 & \frac{1}{4}q_2^2 & \frac{\partial h_2}{\partial p_2} & \frac{\partial h_1}{\partial p_2}\\ -q_1^2-\frac{1}{4}q_2^2 & -\frac{1}{2}q_1q_2 & 0 & \frac{1}{2}q_1p_2 & -\frac{\partial h_2}{\partial q_1} & -\frac{\partial h_1}{\partial q_1}\\ -\frac{1}{2}q_1q_2 & -\frac{1}{4}q_2^2 & -\frac{1}{2}q_1p_2 & 0 & -\frac{\partial h_2}{\partial q_2} & -\frac{\partial h_1}{\partial q_2}\\ \hline -\frac{\partial h_2}{\partial p_1} & -\frac{\partial h_2}{\partial p_2} & \frac{\partial h_2}{\partial q_1} & \frac{\partial h_2}{\partial q_2} & 0 & 0\\ -\frac{\partial h_1}{\partial p_1} & -\frac{\partial h_1}{\partial p_2} & \frac{\partial h_1}{\partial q_1} & \frac{\partial h_1}{\partial q_2} & 0 & 0 \end{array}\right).$$

The first two of them come from one-Casimir extension and the last one was constructed according to formula (53). Notice that again we have three bi-Hamiltonian chains (33).

Example 3: The second two-Casimir extension of the Henon–Heiles system.

In Ref. 1 the second one-Casimir extension of the Henon–Heiles system was considered in the form

$$(q_1)_{tt}=-3q_1^2-\frac{1}{2}q_2^2,\quad (q_2)_{tt}=-q_1q_2+\frac{c}{q_2^3},$$

with two constants of motion

$$\overline{h_1}=-2q_2p_1p_2+2q_1p_2^2-\frac{1}{4}q_2^4-q_1^2q_2^2+\frac{2q_1}{q_2^2}c,$$

$$\overline{h_2}=-2H=-p_1^2-p_2^2-2q_1^3-q_1q_2^2-\frac{1}{q_2^2}c.$$

Moreover, a transformation to the inverse Darboux–Nijenhuis coordinates was found. Nevertheless, according to the result of Ref. 3, the admissible transformation to the Darboux–Nijenhuis coordinates reads

$$q_1=\frac{1}{2}\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right),\quad p_1=2\lambda_1\lambda_2\left(\frac{\lambda_1\mu_1}{\lambda_1-\lambda_2}+\frac{\lambda_2\mu_2}{\lambda_2-\lambda_1}\right),$$

$$q_2=\frac{1}{\sqrt{-\lambda_1\lambda_2}},\quad p_2=-2\sqrt{-\lambda_1\lambda_2}\left(\frac{\lambda_1^2\mu_1}{\lambda_1-\lambda_2}+\frac{\lambda_2^2\mu_2}{\lambda_2-\lambda_1}\right),$$

$$f_i(\lambda_i,\mu_i)=\frac{1}{4}\lambda_i^{-3}+4\lambda_i^4\mu_i^2.$$

The two-Casimir extension we get by adding new terms to the constants of motion

$$h_1(q,p,c_1,c_2)=\overline{h_1}(q,p,c=c_1)+b_{2,1}(q)c_2=\overline{h_1}(q,p,c_1)-\left(\frac{4q_1^2}{q_2^4}+\frac{1}{q_2^2}\right)c_2,$$

$$h_2(q,p,c_1,c_2)=\overline{h_2}(q,p,c=c_1)+b_{2,2}(q)c_2=\overline{h_2}(q,p,c_1)+\frac{2q_1}{q_2^4}c_2,$$

where the Newton equations related to the energy $H=-1/2h_2$ are

$$(q_1)_{tt}=-3q_1^2-\frac{1}{2}q_2^2-\frac{1}{q_2^4}c_2,\quad (q_2)_{tt}=-q_1q_2+\frac{c_1}{q_2^3}-\frac{4q_1}{q_2^5}c_2.$$

Again this is tri-Hamiltonian system, where the first two Poisson structures are given in Ref. 1 (with additional last row and column with zeros), while the new third Poisson structure reads

$$\pi_2=\left(\begin{array}{cccc|cc}
0 & 0 & \frac{1}{q_2^2} & -\frac{2q_1}{q_2^3} & \frac{\partial h_2}{\partial p_1} & \frac{\partial h_1}{\partial p_1}\\
0 & 0 & -\frac{2q_1}{q_2^3} & \frac{4q_1^2+q_2^2}{q_2^4} & \frac{\partial h_2}{\partial p_2} & \frac{\partial h_1}{\partial p_2}\\
-\frac{1}{q_2^2} & \frac{2q_1}{q_2^3} & 0 & -\frac{2q_1p_2}{q_2^4} & -\frac{\partial h_2}{\partial q_1} & -\frac{\partial h_1}{\partial q_1}\\
\frac{2q_1}{q_2^3} & -\frac{4q_1^2+q_2^2}{q_2^4} & \frac{2q_1p_2}{q_2^4} & 0 & -\frac{\partial h_2}{\partial q_2} & -\frac{\partial h_1}{\partial q_2}\\
\hline
-\frac{\partial h_2}{\partial p_1} & -\frac{\partial h_2}{\partial p_2} & \frac{\partial h_2}{\partial q_1} & \frac{\partial h_2}{\partial q_2} & 0 & 0\\
-\frac{\partial h_1}{\partial p_1} & -\frac{\partial h_1}{\partial p_2} & \frac{\partial h_1}{\partial q_1} & \frac{\partial h_1}{\partial q_2} & 0 & 0
\end{array}\right).$$

Example 4: m-Casimir extension of the relativistic *n*-body problem.

Consider the Hamiltonian dynamical system with the Hamiltonian given by

$$H=\sum_{i=1}^{n}\frac{\varphi_i(\lambda_i)}{\Delta_i}e^{a\mu_i}, \tag{57}$$

where φ_i are arbitrary smooth functions and a is an arbitrary constant. The corresponding dynamical system takes the Newtonian form

$$(\lambda_i)_{tt}=2\sum_{k\neq i}\frac{(\lambda_i)_t(\lambda_k)_t}{\lambda_i-\lambda_k}, \quad i=1,...,n, \tag{58}$$

which depends explicitly on velocities. The derivative of the formula (58) is given in Appendix F. Notice that Eqs. (58) do not depend on φ_i functions, hence the dynamics is not influenced by φ_i terms.

Newtonian dynamics (58) is a special case of the integrable relativistic n-body problems introduced by Ruijsenaars and Schneider.[9] Now, comparing (57) with (22) one immediately concludes that (λ,μ) is a Darboux–Nijenhuis chart for the Hamiltonian H, and as a consequence, system (58) is quasi-bi-Hamiltonian and separable, with solution given by the implicit formulas (47)

$$\frac{1}{a}\sum_{k=1}^{n}\int^{\lambda_k}\frac{\xi^i}{g(\xi)}d\xi=\delta_{i,n-1}t+\text{const}, \quad i=0,...,n-1, \tag{59}$$

where $g(\xi)=a_1+a_2\xi+\cdots+a_n\xi^{n-1}$. This fact was noticed for the first time by Morosi and Tondo.[10] Notice, that trajectories $\lambda_i(t)$ do not depend on the $\varphi_i(\lambda_i)$ factors, as expected, so without a loss of generality one can put $\varphi_i(\lambda_i)=1$.

The system (58) can be naturally extended to an m-Casimir one with the Hamiltonian

$$H=\sum_{i=1}^{n}\frac{\varphi_i(\lambda_i)}{\Delta_i}e^{a\mu_i}+\sum_{j=1}^{m}c_j\beta_{j,1}(\lambda), \quad 1\leqslant m\leqslant n \tag{60}$$

and the related Newton equations of motion

$$(\lambda_i)_{tt}=2\sum_{k\neq i}\frac{(\lambda_i)_t(\lambda_k)_t}{\lambda_i-\lambda_k}-a\sum_{j=1}^{m}c_j\frac{\partial\beta_{j,1}}{\partial\lambda_i}(\lambda_i)_t. \tag{61}$$

The dynamical system (61) has n constants of motion

$$h_r(\lambda,\mu,c)=-\sum_{i=1}^{n}\frac{\partial\rho_r(\lambda)}{\partial\lambda_i}\frac{e^{a\mu_i}}{\Delta_i}+\sum_{j=1}^{m}c_j\beta_{j,r}(\lambda), \quad r=1,...,n,$$

$(m+1)$ Poisson structures (24) and the solution given by implicit formula (59), where now

$$g(\xi)=a_1+a_2\xi+\cdots+a_n\xi^{n-1}+c_1\xi^n+\cdots+c_m\xi^{n+m-1}.$$

The one-Casimir extension, which is bi-Hamiltonian, has the following Newton equations of motion:

$$(\lambda_i)_{tt}=2\sum_{k\neq i}\frac{(\lambda_i)_t(\lambda_k)_t}{\lambda_i-\lambda_k}+ac(\lambda_i)_t, \quad i=1,...,n.$$

The two-Casimir extension, which is tri-Hamiltonian, has the Newton equations in the form

$$(\lambda_i)_{tt}=2\sum_{k\neq i}\frac{(\lambda_i)_t(\lambda_k)_t}{\lambda_i-\lambda_k}+ac_1(\lambda_i)_t+ac_2\left(\lambda_i+\sum_{k=1}^{n}\lambda_k\right)(\lambda_i)_t, \quad i=1,...,n.$$

Example 5: $(m+1)$-Hamiltonian formulation for elliptic separable potentials.

In Ref. 11 a theorem has been proved which says that every natural Hamiltonian system

$$H(q,p,c)=\tfrac{1}{2}(p,p)+V(q)+\tfrac{1}{2}c(q,q) \tag{62}$$

separable in generalized elliptic coordinates, in the extended phase space of variables (q,p,c), where $p=(q_1,\dots,q_n)^T$, $p=(p_1,\dots,p_n)^T$ and c is additional Casimir variable, admits the bi-Hamiltonian formulation

$$\begin{pmatrix} q \\ p \\ c \end{pmatrix}_t = \pi_0 \nabla h_1 = \pi_1 \nabla h_0,$$

where

$$\pi_0 = \left(\begin{array}{cc|c} 0 & I & 0 \\ -I & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right), \tag{63}$$

$$\pi_1 = \left(\begin{array}{cc|c} 0 & A-\tfrac{1}{2}q\otimes q & \partial h_1/\partial p \\ -A+\tfrac{1}{2}q\otimes q & \tfrac{1}{2}p\otimes q - \tfrac{1}{2}q\otimes p & -\partial h_1/\partial q \\ \hline -(\partial h_1/\partial p)^T & (\partial h_1/\partial q)^T & 0 \end{array}\right),$$

$A=\mathrm{diag}(\alpha_1,\dots,\alpha_n)$, α_i are different positive constants, $h_0=c$, $h_1=H-c\Sigma_i\alpha_i$, $(.,.)$ stands for the scalar product and $\otimes-$ the respective tensor product. The potential $V(q)$ satisfies the equations

$$0=(\alpha_i-\alpha_k)\frac{\partial^2 V}{\partial q_i \partial q_k}+\frac{3}{2}\left(q_k\frac{\partial V}{\partial q_i}-q_i\frac{\partial V}{\partial q_k}\right)+\frac{1}{2}\sum_{j=1}^{n}\left(q_j q_k\frac{\partial^2 V}{\partial q_i\partial q_j}-q_i q_j\frac{\partial^2 V}{\partial q_j \partial q_k}\right), \tag{64}$$

$i,k=1,\dots,n$, which are the iff conditions for π_1 to fulfill the Jacobi identity. This bi-Hamiltonian formulation generates the chain of commuting bi-Hamiltonian vector fields

$$\pi_0\nabla h_0=0, \quad \pi_0\nabla h_1=\pi_1\nabla h_0,\dots, \quad \pi_0\nabla h_n=\pi_1\nabla h_{n-1}, \quad 0=\pi_1\nabla h_n, \tag{65}$$

starting with the Casimir h_0 of π_0 and terminating with the Casimir h_n of π_1, where

$$h_r=\sum_{k=0}^{r}\rho_k(\alpha)\overline{h}_{r-k}, \quad \overline{h}_r=\frac{1}{2}\sum_{i=1}^{n}\alpha_i^{r-1}K_i+\frac{1}{2}c(q,A^{r-1}q),$$

$$K_i=\sum_{j\neq i}\frac{q_i p_j-q_j p_i}{\alpha_i-\alpha_j}+p_i^2+V_i(q), \tag{66}$$

$\rho_k(\alpha)$ are Viete polynomials of α_i, $V_i(q)$ are functions of q such that $\Sigma_i V_i(q)=V(q)$ and the Poisson brackets $\{H,K_i\}_{\pi_0}=0$, $i=1,\dots,n$.

Then, in Refs. 1 and 3 the inverse statement was proved, i.e., every natural Hamiltonian system (62) admitting the bi-Hamiltonian formulation (63)–(66) is separable in the generalized elliptic coordinates, which are just the Darboux–Nijenhuis coordinates $\lambda_1,\dots,\lambda_n$.

Now, according to the theory presented, let us generalize the Hamiltonian system (62) to the form

$$H(q,p,c)=\frac{1}{2}(p,p)+V(q)+\sum_{k=1}^{m}c_k b_{k,1}(q), \quad m=1,\dots,n \tag{67}$$

being multi-Hamiltonian and separable. The few first $b_{k,1}(q)$ functions are the following:

$$b_{1,1}(q)=\tfrac{1}{2}(q,q),$$

$$b_{2,1}(q)=\tfrac{1}{2}(q,Aq)-\tfrac{1}{4}(q,q)^2,$$

$$b_{3,1}(q)=\tfrac{1}{2}(q,q)(q,Aq)-\tfrac{1}{2}(q,A^2q)-\tfrac{1}{8}(q,q)^3,\ldots \tag{68}$$

being particular solutions of Eq. (64).

For example, the three Poisson structures of the system (67) with two Casimirs read

$$\pi_0=\left(\begin{array}{cc|cc} 0 & I & 0 & 0\\ -I & 0 & 0 & 0\\ \hline 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{array}\right),$$

$$\pi_1=\left(\begin{array}{cc|cc} 0 & A-\tfrac{1}{2}q\otimes q & \partial h_1/\partial p & 0\\ -A+\tfrac{1}{2}q\otimes q & \tfrac{1}{2}p\otimes q-\tfrac{1}{2}q\otimes p & -\partial h_1/\partial q & 0\\ \hline -(\partial h_1/\partial p)^T & (\partial h_1/\partial q)^T & 0 & 0\\ 0 & 0 & 0 & 0\end{array}\right),$$

$$\pi_2=\left(\begin{array}{cc|cc} 0 & (A-\tfrac{1}{2}q\otimes q)^2 & \partial h_2/\partial p & \partial h_1/\partial p\\ -(A-\tfrac{1}{2}q\otimes q)^2 & \begin{array}{c}(A-\tfrac{1}{2}q\otimes q)(\tfrac{1}{2}p\otimes q-\tfrac{1}{2}q\otimes p)\\ +(\tfrac{1}{2}p\otimes q-\tfrac{1}{2}q\otimes p)(A-\tfrac{1}{2}q\otimes q)\end{array} & -\partial h_2/\partial q & -\partial h_1/\partial q\\ \hline -(\partial h_2/\partial p)^T & (\partial h_2/\partial q)^T & 0 & 0\\ -(\partial h_1/\partial p)^T & (\partial h_1/\partial q)^T & 0 & 0\end{array}\right).$$

The functions $h_r(q,p,c)$, forming three admissible bi-Hamiltonian chains, are given by formulas (66), where now

$$\overline{h_r}=\frac{1}{2}\sum_{i=1}^{n}\alpha_i^{r-1}K_i+c_1b_{1,r}(q)+c_2b_{2,r}(q).$$

VI. SUMMARY AND FINAL COMMENTS

In this paper we have discussed the separability theory of multi-Hamiltonian systems in the case when each Poisson structure has more then one Casimir. In fact, we started from the system written in separated coordinates $(\mu_i,\lambda_i)_{i=1}^n$. Such a system can be represented by n-point dynamics on some curve. Actually, let us consider a curve in (λ,μ) plane, in the particular form

$$f(\lambda,\mu)=h_\lambda,\quad h_\lambda=c\lambda^n+h_1\lambda^{n-1}+\cdots+h_n, \tag{69}$$

where $f(\lambda,\mu)$ is an arbitrary smooth function. Then, let us take n different points (μ_i,λ_i) from the curve:

$$f(\lambda_i,\mu_i)=c\lambda_i^n+h_1\lambda_i^{n-1}+\cdots+h_n,\quad i=1,\ldots,n, \tag{70}$$

which will define our separated coordinates. The explicit dependence of h_k on $(\mu_i,\lambda_i,c)_{i=1}^n$ is given by the solution of n linear equations (70), while for fixed values of $h_k=a_{n+1-k}$ and $\mu_i=\partial W_i/\partial\lambda_i$ the system (70) allows us to solve the appropriate Hamilton–Jacobi equations.

In previous papers[1–3] the bi-Hamiltonian chain was constructed for a *separation curve* in the form (69). In this paper multi-Hamiltonian chains was constructed for a separation curve in the form

$$f(\lambda,\mu)=c_n\lambda^{2n-1}+\cdots+c_1\lambda^n+h_1\lambda^{n-1}+\cdots+h_n. \tag{71}$$

Here we have to mension that the idea to relate of multi-Hamiltonian property with an m-parameter family of curves comes from Vanhaecke.[12]

While in the case of one Casimir the admissible form of a separation curve for which one can construct a Poisson pencil is rather restrictive, in multi-Casimir case the freedom of the choice is much bigger. Here we consider the simplest multi-Casimir extension leading to unsplit chains, like in the one-Casimir case. Choosing other admissible forms of the separation curve with more then one Casimir, one can construct split bi(multi)-Hamiltonian chain, i.e., the chain which splits onto few bi(multi)-Hamiltonian sub-chains, each starting and terminating with some Casimir of the appropriate Poisson structure. The work is still in progress and the results will be published in a separate paper.

The structure of degenerated Poisson pencils π_λ expressed in separated coordinates $(\mu_i,\lambda_i)_{i=1}^n$ is particularly clear, i.e., rows and columns related with Casimir variables are appropriate vector fields from the chain. Hence, the reduction of the pencil onto a symplectic leaf of π_0 (fixing the values of all c_i) is immediate as obviously $\theta_\lambda=\theta_1-\lambda\theta_0$, where

$$\theta_0=\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \theta_1=\begin{pmatrix} 0 & \Lambda \\ -\Lambda & 0 \end{pmatrix}, \tag{72}$$

is a nondegenerated Poisson pencil. Hence, the Marsden–Ratiu bi-Hamiltonian reduction[13,14] of the pair π_0, π_1 is trivial in separated coordinates.

Now, let us consider an arbitrary canonical transformation

$$(q,p)\to(\lambda,\mu)$$

independent of Casimir coordinates. The advantage of staying inside such class of transformations is that the clear structure of the pencil is preserved and the Marsden–Ratiu reduction of the Poisson pencil is still trivial. Of course the most general case of multi-Hamiltonian separability theory takes place when one goes beyond the set of canonical coordinates. But then the simple structure of degenerated Poisson pencil is lost and the nontrivial problem of the Marsden–Ratiu reduction for such pencil appears.

Although for all examples from previous section the related Nijenhuis tensor N takes the form (55), i.e., is diagonalizable by an appropriate point transformation, nevertheless, also the more general case, when a diagonalizing transformation is of nonpoint nature, is also admissible. One of the systematic method of construction of separated coordinates in such a case leads through the so-called Hankel–Fröbenius coordinates. Some details the reader can find in Ref. 5.

Concluding, it seems that the developing nowadays multi-Hamiltonian separability theory, which bases on a Poisson pencil and separation curve, is enough general and efficient to be considered as an alternative to the well-known Sklyanin one,[15] bases on Lax operator and a spectral curve.

APPENDIX A

$$n=2$$

$$\beta_{1,1}=-\lambda_1-\lambda_2, \quad \beta_{2,1}=-\lambda_1^2-\lambda_2^2-\lambda_1\lambda_2,$$

$$\beta_{1,2}=\lambda_1\lambda_2, \quad \beta_{2,2}=\lambda_1\lambda_2(\lambda_1+\lambda_2),$$

$$n=3$$

$$\beta_{1,1}=-\lambda_1-\lambda_2-\lambda_3,$$

$$\beta_{1,2}=\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3,$$

$$\beta_{1,3}=-\lambda_1\lambda_2\lambda_3,$$

$$\beta_{2,1}=-\lambda_1^2-\lambda_2^2-\lambda_3^2-\lambda_1\lambda_2-\lambda_1\lambda_3-\lambda_2\lambda_3,$$

$$\beta_{2,2}=2\lambda_1\lambda_2\lambda_3+\lambda_1\lambda_2(\lambda_1+\lambda_2)+\lambda_1\lambda_3(\lambda_1+\lambda_3)+\lambda_2\lambda_3(\lambda_2+\lambda_3),$$

$$\beta_{2,3}=-\lambda_1\lambda_2\lambda_3(\lambda_1+\lambda_2+\lambda_3),$$

$$\beta_{3,1}=-\lambda_1^3-\lambda_2^3-\lambda_3^3-\lambda_1\lambda_2(\lambda_1+\lambda_2)-\lambda_1\lambda_3(\lambda_1+\lambda_3)-\lambda_2\lambda_3(\lambda_2+\lambda_3)-\lambda_1\lambda_2\lambda_3,$$

$$\beta_{3,2}=\lambda_1\lambda_2(\lambda_1^2+\lambda_1\lambda_2+\lambda_2^2)+\lambda_1\lambda_3(\lambda_1^2+\lambda_1\lambda_3+\lambda_3^2)$$
$$+\lambda_2\lambda_3(\lambda_2^2+\lambda_2\lambda_3+\lambda_3^2)+2\lambda_1\lambda_2\lambda_3(\lambda_1+\lambda_2+\lambda_3),$$

$$\beta_{3,3}=-\lambda_1\lambda_2\lambda_3(\lambda_1^2+\lambda_2^2+\lambda_3^2+\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3).$$

APPENDIX B: THE PROOF OF LEMMA 1

(i) The proof is inductive. As

$$\beta_{1,r}(\lambda)\equiv\rho_r(\lambda)=-\sum_{i=1}^{n}\frac{\rho_{r-1}^i\lambda_i^n}{\Delta_i},$$

then

$$\begin{aligned}\beta_{k+1,r}(\lambda)&=\beta_{k,r+1}-\beta_{k,1}\cdot\beta_{1,r}\\&=-\sum_{i=1}^{n}\left[\frac{\rho_r^i\lambda_i^{n+k-1}}{\Delta_i}-\frac{\rho_r\lambda_i^{n+k-1}}{\Delta_i}\right]\\&=-\sum_{i=1}^{n}\left[\frac{(\rho_r+\lambda_i\rho_{r-1}^i)\lambda_i^{n+k-1}}{\Delta_i}-\frac{\rho_r\lambda_i^{n+k-1}}{\Delta_i}\right]\\&=-\sum_{i=1}^{n}\frac{\rho_{r-1}^i\lambda_i^{n+k}}{\Delta_i},\end{aligned}$$

where we used the relation $\lambda_i\rho_{r-1}^i=-\rho_r+\rho_r^i$. □

(ii)

$$\begin{aligned}\frac{\partial\beta_{r,k+1}}{\partial\lambda_i}&=-\frac{\partial}{\partial\lambda_i}\sum_{l=1}^{n}\frac{\rho_k^l\lambda_l^{n+r-1}}{\Delta_l}\\&=-\rho_k^i\frac{\partial}{\partial\lambda_i}\frac{\lambda_i^{n+r-1}}{\Delta_i}-\sum_{l\neq i}\lambda_l^{n+r-1}\frac{\partial}{\partial\lambda_i}\frac{\rho_k^l}{\Delta_l}\\&=-\rho_k^i(n+r-1)\frac{\lambda_i^{n+r-2}}{\Delta_i}+\rho_k^i\frac{\lambda_i^{n+r-1}}{\Delta_i}\sum_{l\neq i}\frac{1}{\lambda_i-\lambda_l}+\sum_{l\neq i}\rho_{k-1}^{il}\frac{\lambda_l^{n+r-1}}{\Delta_l}\\&\quad-\sum_{l\neq i}\rho_k^l\frac{\lambda_l^{n+r-1}}{\Delta_l}\frac{1}{\lambda_l-\lambda_i}.\end{aligned}$$

Applying formulas (14), (15), the last two terms can be transformed into the form

$$\sum_{l\neq i}\left(\rho_{k-1}^{il}-\frac{\rho_k^l}{\lambda_l-\lambda_i}\right)\frac{\lambda_l^{n+r-1}}{\Delta_l}=-\sum_{l\neq i}\frac{\rho_k^i}{\lambda_l-\lambda_i}\frac{\lambda_l^{n+r-1}}{\Delta_l}$$

$$=-\rho_k^i\sum_{l\neq i}\frac{\lambda_l^{n+r-1}}{\Delta_l}\frac{1}{\lambda_l-\lambda_i},$$

hence

$$\frac{\partial\beta_{r,k+1}}{\partial\lambda_i}=-\rho_k^i(n+r-1)\frac{\lambda_i^{n+r-2}}{\Delta_i}+\rho_k^i\left(\frac{\lambda_i^{n+r-1}}{\Delta_i}\sum_{l\neq i}\frac{1}{\lambda_i-\lambda_l}+\sum_{l\neq i}\frac{\lambda_l^{n+r-1}}{\Delta_l}\frac{1}{\lambda_i-\lambda_l}\right)$$

$$=-\rho_k^i\frac{\partial}{\partial\lambda_i}\sum_{l=1}^{n}\frac{\lambda_l^{n+r-1}}{\Delta_l}=\rho_k^i\frac{\partial\beta_{r,1}}{\partial\lambda_i}.$$

□

APPENDIX C: THE PROOF OF LEMMA 3

Applying relations (14), (15) and representation (22) we have

$$\frac{\partial h_r}{\partial\mu_i}=\rho_{r-1}^i\left(\frac{\partial}{\partial\mu_i}\frac{\overline{f_i}}{\Delta_i}\right)$$

$$=\rho_{r-1}^i\frac{\partial h_1}{\partial\mu_i},$$

$$\frac{\partial h_r}{\partial\lambda_i}=\sum_{j=1}^{n}\left(\frac{\partial\rho_{r-1}^j}{\partial\lambda_i}\frac{\overline{f_j}}{\Delta_j}+\rho_{r-1}^j\frac{\partial}{\partial\lambda_i}\frac{\overline{f_j}}{\Delta_j}\right)$$

$$=\rho_{r-1}^i\left(\frac{\partial}{\partial\lambda_i}\frac{\overline{f_i}}{\Delta_i}\right)+\sum_{j\neq i}\left(\frac{\partial\rho_{r-1}^j}{\partial\lambda_i}\frac{\overline{f_j}}{\Delta_j}+\rho_{r-1}^j\overline{f_j}\frac{\partial}{\partial\lambda_i}\frac{1}{\Delta_j}\right)$$

$$=\rho_{r-1}^i\left(\frac{\partial}{\partial\lambda_i}\frac{\overline{f_i}}{\Delta_i}\right)+\sum_{j\neq i}\left(-\rho_{r-2}^{ji}+\frac{\rho_{r-1}^j}{\lambda_j-\lambda_i}\right)\frac{\overline{f_j}}{\Delta_j}$$

$$=\rho_{r-1}^i\left(\frac{\partial}{\partial\lambda_i}\frac{\overline{f_i}}{\Delta_i}\right)+\sum_{j\neq i}\frac{\rho_{r-1}^i}{\lambda_j-\lambda_i}\frac{\overline{f_j}}{\Delta_j}$$

$$=\rho_{r-1}^i\left(\frac{\partial}{\partial\lambda_i}\frac{\overline{f_i}}{\Delta_i}+\sum_{j\neq i}\frac{1}{\lambda_j-\lambda_i}\frac{\overline{f_j}}{\Delta_j}\right)=\rho_{r-1}^i\frac{\partial h_1}{\partial\lambda_i}.$$

□

Analogously one can prove the last derivative.

APPENDIX D: THE PROOF OF LEMMA 5

Vector fields $K_r=(X_r,Y_r)^T$ are Hamiltonian, i.e., $K_r=\pi_0\nabla H_r$. Conditions (28) are fulfilled as

$$\frac{\partial}{\partial p_i}\frac{\partial}{\partial p_j}h_r=\frac{\partial}{\partial p_k}\frac{\partial}{\partial c_l}h_r=0,\quad i\neq j.$$

Conditions (31) follow from involutivity of h_r with respect to π_0. Taking into account the results of Lemma 1 (ii) and Lemma 3, conditions (29) and (30) reduce to the following three:

$$m\lambda_i^{m-1}+\sum_{k=1}^{m}\frac{\partial\beta_{k,1}}{\partial\lambda_i}\rho_{m-k}^i=0,\quad j=i, \tag{D1}$$

$$(\lambda_j^m-\lambda_i^m)\frac{1}{\lambda_i-\lambda_j}+\sum_{k=1}^{m}\frac{\partial\beta_{k,1}}{\partial\lambda_j}\rho_{m-k}^i=0,\quad j\neq i, \tag{D2}$$

$$(\lambda_j^m-\lambda_i^m)\frac{1}{\lambda_i-\lambda_j}+\sum_{k=1}^{m}\frac{\partial\beta_{k,1}}{\partial\lambda_i}\rho_{m-k}^j=0,\quad j\neq i. \tag{D3}$$

First we prove the relation (D1). From $\pi_0\nabla h_r=\pi_m\nabla h_{r-m}$ we get

$$\frac{\partial h_r}{\partial\lambda_i}=\lambda_i^m\frac{\partial h_{r-m}}{\partial\lambda_i}+\sum_{k=1}^{m}\frac{\partial h_{r-m-k}}{\partial\lambda_i}\beta_{k,r-m}$$

$$\Downarrow$$

$$\frac{\partial\beta_{l,r}}{\partial\lambda_i}=\lambda_i^m\frac{\partial\beta_{l,r-m}}{\partial\lambda_i}+\sum_{k=1}^{m}\frac{\partial\beta_{l,r-m-k}}{\partial\lambda_i}\beta_{k,r-m}$$

$$\Downarrow \text{ Lemma1(ii)}$$

$$\rho_{r-1}^i\frac{\partial\beta_{l,1}}{\partial\lambda_i}=\lambda_i^m\rho_{r-m-1}^i\frac{\partial\beta_{l,1}}{\partial\lambda_i}+\sum_{k=1}^{m}\rho_{m-k}^i\frac{\partial\beta_{l,1}}{\partial\lambda_i}\beta_{k,r-m}$$

$$\Downarrow$$

$$\rho_{r-1}^i=\lambda_i^m\rho_{r-m-1}^i+\sum_{k=1}^{m}\rho_{m-k}^i\beta_{k,r-m} \tag{D4}$$

$$\Downarrow\frac{\partial}{\partial\lambda_i}$$

$$0=m\lambda_i^{m-1}\rho_{r-m-1}^i+\sum_{k=1}^{m}\frac{\partial\beta_{k,r-m}}{\partial\lambda_i}\rho_{m-k}^i$$

$$\Downarrow \text{ Lemma1(i)}$$

$$0=m\lambda_i^{m-1}\rho_{r-m-1}^i+\sum_{k=1}^{m}\frac{\partial\beta_{k,1}}{\partial\lambda_i}\rho_{r-m-1}^i\rho_{m-k}^i$$

$$\Downarrow$$

$$0=m\lambda_i^{m-1}+\sum_{k=1}^{m}\frac{\partial\beta_{k,1}}{\partial\lambda_i}\rho_{m-k}^{i}.$$

□

To prove relations (D2) and (D3) substitute (D4) to equality

$$\frac{\partial}{\partial\lambda_j}\rho_{2m-l}^{i}-\frac{\partial}{\partial\lambda_i}\rho_{2m-l}^{j}=0,$$

hence we get

$$0=(\lambda_j^m-\lambda_i^m)\rho_{m-l-1}^{ij}+\sum_{k=1}^{m}\left(\rho_{m-k}^{i}\rho_{m-l}^{j}\frac{\partial\beta_{k,1}}{\partial\lambda_j}-\rho_{m-k}^{j}\rho_{m-l}^{i}\frac{\partial\beta_{k,1}}{\partial\lambda_i}\right).$$

Applying the relation $\rho_r^i-\rho_r^j=(\lambda_i-\lambda_j)\rho_{r-1}^{ij}$ we have

$$0=\left[(\lambda_j^m-\lambda_i^m)\frac{1}{\lambda_j-\lambda_i}+\sum_{k=1}^{m}\rho_{m-k}^{i}\frac{\partial\beta_{k,1}}{\partial\lambda_j}\right]\rho_{m-l}^{j}+\left[(\lambda_j^m-\lambda_i^m)\frac{1}{\lambda_i-\lambda_j}+\sum_{k=1}^{m}\rho_{m-k}^{j}\frac{\partial\beta_{k,1}}{\partial\lambda_i}\right]\rho_{m-l}^{i}.$$

From arbitrariness of l one gets equalities (D2) and (D3). □

APPENDIX E: THE PROOF OF THEOREM 2

The basic bi-Hamiltonian chains read

$$\pi_{m-1}\nabla h_{r+1}=\pi_m\nabla h_r,\quad m=1,...,n. \tag{E1}$$

All the remaining chains can be constructed from (E1) ones. The validity of (E1) can be proved by induction. For $m=1$ we get

$$\pi_0\nabla h_{r+1}=\pi_1\nabla h_r$$

$$\Updownarrow$$

$$\frac{\partial h_{r+1}}{\partial\mu_i}=\lambda_i\frac{\partial h_r}{\partial\mu_i}+\beta_{1,r}\frac{\partial h_1}{\partial\mu_i}, \tag{E2}$$

$$\frac{\partial h_{r+1}}{\partial\lambda_i}=\lambda_i\frac{\partial h_r}{\partial\lambda_i}+\beta_{1,r}\frac{\partial h_1}{\partial\lambda_i},\quad i=1,...,n, \tag{E3}$$

$$\left(\frac{\partial h_1}{\partial\lambda}\right)^T\frac{\partial h_r}{\partial\mu}-\left(\frac{\partial h_1}{\partial\mu}\right)^T\frac{\partial h_r}{\partial\lambda}=0. \tag{E4}$$

This was proved for $c_1\neq 0$ in Ref. 2. Additional terms with $c_j\neq 0$, $j>1$ do not influence (E2) and from (E3) have to fulfill the conditions

$$\frac{\partial\beta_{j,r+1}}{\partial\lambda_i}=\lambda_i\frac{\partial\beta_{j,r}}{\partial\lambda_i}+\beta_{1,r}\frac{\partial\beta_{j,1}}{\partial\lambda_i},\quad j=2,...,n. \tag{E5}$$

Equations (E5) are true according to Lemma 1 (i) and the relation $\rho_r^i=\lambda_i\rho_{r-1}^i+\rho_r$. Assuming the validity of the chain

$$\pi_{m-2}\nabla h_{r+1}=\pi_{m-1}\nabla h_r \tag{E6}$$

we prove the validity of

$$\pi_{m-1}\nabla h_{r+1}=\pi_m\nabla h_r \tag{E7}$$

$$\Updownarrow$$

$$\lambda_i^{m-1}\frac{\partial h_{r+1}}{\partial\mu_i(\lambda_i)}-\lambda_i^m\frac{\partial h_r}{\partial\mu_i(\lambda_i)}=(\beta_{m,r}-\beta_{m-1,r+1})\frac{\partial h_1}{\partial\mu_i(\lambda_i)}+(\beta_{m-1,r}-\beta_{m-2,r+1})\frac{\partial h_2}{\partial\mu_i(\lambda_i)}+\cdots$$
$$+(\beta_{m+1-j,r}-\beta_{m-j,r+1})\frac{\partial h_j}{\partial\mu_i(\lambda_i)}+\cdots+(\beta_{2,r}-\beta_{1,r+1})\frac{\partial h_{m-1}}{\partial\mu_i(\lambda_i)}$$
$$+\beta_{1,r}\frac{\partial h_m}{\partial\mu_i(\lambda_i)},\quad i=1,...,n, \tag{E8}$$

$$\sum_{i=1}^{m-1}\left[\left(\frac{\partial h_i}{\partial\lambda}\right)^T\frac{\partial h_{r+1}}{\partial\mu}-\left(\frac{\partial h_i}{\partial\mu}\right)^T\frac{\partial h_{r+1}}{\partial\lambda}\right]=\sum_{j=1}^{m}\left[\left(\frac{\partial h_j}{\partial\lambda}\right)^T\frac{\partial h_r}{\partial\mu}-\left(\frac{\partial h_j}{\partial\mu}\right)^T\frac{\partial h_r}{\partial\lambda}\right]$$
$$\Leftrightarrow\sum_{i=1}^{m-1}\{h_i,h_{r+1}\}_{\pi_0}$$
$$=\sum_{j=1}^{m}\{h_j,h_r\}_{\pi_0}. \tag{E9}$$

Equality (E9) is fulfilled according to involutivity of h_r with respect to π_0. Now, let us consider (E8) with respect to $\mu_i(\lambda_i)$ derivatives. From the analogous ith component of the chain (E6), multiplied by λ_i we get

$$\lambda_i^{m-1}\frac{\partial h_{r+1}}{\partial\mu_i(\lambda_i)}-\lambda_i^m\frac{\partial h_r}{\partial\mu_i(\lambda_i)}=(\beta_{m-1,r}-\beta_{m-2,r+1})\lambda_i\frac{\partial h_1}{\partial\mu_i(\lambda_i)}+\cdots$$
$$+(\beta_{m-j,r}-\beta_{m-1-j,r+1})\lambda_i\frac{\partial h_j}{\partial\mu_i(\lambda_i)}+\cdots+\beta_{1,r}\lambda_i\frac{\partial h_{m-1}}{\partial\mu_i(\lambda_i)}. \tag{E10}$$

Comparing (E8) with (E10) and using relations from Lemma 3 we get

$$0=(\beta_{m,r}-\beta_{m-1,r+1})+(\beta_{m-1,r}-\beta_{m-2,r+1})(\rho_1^i-\lambda_i\rho_0^i)+\cdots$$
$$+(\beta_{m-j,r}-\beta_{m-1-j,r+1})(\rho_j^i-\lambda_i\rho_{j-1}^i)+\cdots+\beta_{1,r}(\rho_{m-1}^i-\lambda_i\rho_{m-2}^i). \tag{E11}$$

Taking into account the relation $\rho_j^i-\lambda_i\rho_{j-1}^i=\rho_j\equiv\beta_{1,j}$, (E11) turns into

$$0=(\beta_{m,r}-\beta_{m-1,r+1})+(\beta_{m-1,r}-\beta_{m-2,r+1})\beta_{1,1}+\cdots+(\beta_{m-j,r}-\beta_{m-1-j,r+1})\beta_{1,j}+\cdots$$
$$+(\beta_{2,r}-\beta_{1,r+1})\beta_{1,m-2}+\beta_{1,r}\beta_{1,m-1}. \tag{E12}$$

To verify the equality (E12) we apply the relations

$$\beta_{m-j,r}-\beta_{m-1-j,r+1}=-\beta_{m-1-j,1}\cdot\beta_{1,r},$$

$$\beta_{m,1}=\beta_{1,m}-\beta_{m-1,1}\cdot\beta_{1,1}-\beta_{m-2,1}\cdot\beta_{1,2}-\cdots-\beta_{2,1}\cdot\beta_{1,m-2}-\beta_{1,1}\cdot\beta_{1,m-1},$$

following from (17). □

APPENDIX F: THE DERIVATION OF EQUATION OF MOTION (58)

$$(\lambda_i)_t=\frac{\partial H}{\partial \mu_i}=\frac{a\varphi_i(\lambda_i)}{\Delta_i}e^{a\mu_i},$$

$$(\mu_i)_t=-\frac{\partial H}{\partial \lambda_i}=-\frac{\partial \varphi_i}{\partial \lambda_i}\frac{1}{\Delta_i}e^{a\mu_i}+\frac{\varphi_i}{\lambda_i}e^{a\mu_i}\sum_{k\neq i}\frac{1}{\lambda_i-\lambda_k}-\sum_{k\neq i}\frac{\varphi_k}{\lambda_k}e^{a\mu_k}\frac{1}{\lambda_k-\lambda_i}$$

$$=-\frac{\partial \varphi_i}{\partial \lambda_i}\frac{1}{\varphi_i}\frac{1}{a}(\lambda_i)_t+\frac{1}{a}(\lambda_i)_t\sum_{k\neq i}\frac{1}{\lambda_i-\lambda_k}-\frac{1}{a}\sum_{k\neq 1}(\lambda_k)_t\frac{1}{\lambda_k-\lambda_i},$$

$$\ln(\lambda_i)_t=\ln a+\ln \varphi_i+a\mu_i-\ln \Delta_i\,,$$

$$\frac{(\lambda_i)_{tt}}{(\lambda_i)_t}=[\ln(\lambda_i)_t]_t=\frac{(\varphi_i)_t}{\varphi_i}+a(\mu_i)_t-\frac{(\Delta_i)_t}{\Delta_i}$$

$$=\frac{\partial \varphi_i}{\partial \lambda_i}\frac{1}{\varphi_i}(\lambda_i)_t+a(\mu_i)_t-\sum_{k\neq i}\frac{(\lambda_k)_t}{\lambda_k-\lambda_i}-(\lambda_i)_t\sum_{k\neq i}\frac{1}{\lambda_i\lambda_k}$$

$$=-2\sum_{k\neq i}\frac{(\lambda_k)_t}{\lambda_k-\lambda_i},$$

as

$$\frac{\partial}{\partial \lambda_k}\Delta_i=\frac{\Delta_i}{\lambda_k-\lambda_i},\quad k\neq i,\quad \frac{\partial}{\partial \lambda_i}\Delta_i=\sum_{k\neq i}\frac{\Delta_i}{\lambda_i-\lambda_k}.$$

□

[1] M. Błaszak, ''On separability of bi-Hamiltonian chain with degenerated Poisson structures,'' J. Math. Phys. **39**, 3213 (1998).
[2] M. Błaszak, ''Bi-Hamiltonian separable chains on Riemannian manifolds,'' Phys. Lett. A **243**, 25 (1998).
[3] M. Błaszak, *Multi-Hamiltonian Theory of Dynamical Systems* (Springer-Verlag, Berlin, 1998).
[4] I. M. Gel'fand and I. Zakharevich, ''On the local geometry of a bi-Hamiltonian structure,'' in *The Gelfand Mathematical Seminars 1990-1992*, edited by L. Corwin *et al.* (Birkauser, Boston, 1993), pp. 51–112.
[5] G. Falqui, F. Magri, and G. Tondo, ''Reduction of bihamiltonian systems and separation of variables: an example from the Boussinesq hierarchy,'' Theor. Math. Phys. (in press).
[6] R. Brouzet, R. Caboz, J. Rabenivo, and V. Ravoson, ''Two degrees of freedom quasi-bi-Hamiltonian systems,'' J. Phys. A **29**, 2069 (1996).
[7] C. Morosi and G. Tondo, ''Quasi-bi-Hamiltonian systems and separability,'' J. Phys. A **30**, 2799 (1997).
[8] M. Błaszak, ''On a non-standard algebraic description of integrable nonlinear systems,'' Physica (Utrecht) **198A**, 637 (1993).
[9] S. N. Ruijsenaars and H. Schneider, ''A new class of integrable systems and its relation to solitons,'' Ann. Phys. (N.Y.) **170**, 370 (1986).
[10] C. Morosi and G. Tondo, ''On a class of systems both Quasi-Bi-Hamiltonian and Bi-Hamiltonian,'' Phys. Lett. A **247**, 59 (1998).
[11] S. Rauch-Wojciechowski, ''A bi-Hamiltonian formulation for separable potentials and its application to the Kepler problem and the Euler problem of two centers of gravitation,'' Phys. Lett. A **160**, 149 (1991).
[12] P. Vanhaecke, ''Integrable systems in the realm of algebraic geometry,'' *Lecture Notes in Mathematics*, Vol. 1638 (Springer-Verlag, New York, 1996).
[13] J. E. Marsden and T. Ratiu, ''Reduction of Poisson manifolds,'' Lett. Math. Phys. **11**, 161 (1986).
[14] P. Casati, F. Magri, and M. Pedroni, ''Bihamiltonian manifolds and τ—function,'' in *Mathematical Aspects of Classical Field Theory 1991*, edited by M. J. Gotay *et al.*, Contemporary Mathematics, Vol. 132, pp. 213–234 (American Mathematical Society, Washington, DC, 1992).
[15] E. K. Sklyanin, ''Separation of variables. New Trends,'' Prog. Theor. Phys. Suppl. **118**, 35 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Nonisospectral scattering problems: A key to integrable hierarchies

Pilar R. Gordoa and Andrew Pickering
Area de Fisica Teórica, Facultad de Ciencias, Edificio de Fisica, Universidad de Salamanca, 37008 Salamanca, Spain

(Received 12 February 1999; accepted for publication 19 July 1999)

We show that certain partial differential equations associated to nonisospectral scattering problems in $2+1$ dimensions provide a key to associated integrable hierarchies of both ordinary and partial differential equations. This is illustrated using (an extension of) a known second-order and two new third-order nonisospectral scattering problems. These scattering problems allow us to derive new hierarchies of integrable partial differential equations, in both $1+1$ and $2+1$ dimensions, together with their underlying linear problems (isospectral and nonisospectral); and also new hierarchies of integrable ordinary differential equations, again with their underlying linear problems. [S0022-2488(99)04210-3]

I. INTRODUCTION

The introduction of the inverse scattering transform (IST) for the Korteweg–de Vries (KdV) equation[1] and the nonlinear Schrödinger equation[2] marked the birth of a whole new area of mathematics. Since then, a great deal of effort has been spent on finding new completely integrable systems, i.e., similarly solvable through an underlying linear problem. The extension of these ideas to partial differential equations (PDEs) in multidimensions[3,4] and, through the development of the inverse monodromy transform (IMT), to ordinary differential equations (ODEs),[5] has led to the huge variety of completely integrable systems that we know today.

In this paper we will be concerned with a particular class of PDEs in multidimensions, namely those arising as the compatibility condition of a *nonisospectral* scattering problem. The first example of such a PDE is in fact due to Calogero,[6] and has as a subcase the equation

$$u_{xt}=u_{xxxy}+4u_xu_{xy}+2u_{xx}u_y\,. \tag{1.1}$$

This equation arises as the compatibility condition of the Lax pair,

$$\psi_{xx}+(u_x-\lambda)\psi=0,\quad \psi_t=4\lambda\psi_y+2u_y\psi_x-u_{xy}\psi, \tag{1.2}$$

where the spectral parameter $\lambda=\lambda(y,t)$ satisfies the constraint[7,8]

$$\lambda_t=4\lambda\lambda_y\,. \tag{1.3}$$

Such a construction can be carried out with any number of additional spatial variables, i.e., y could be taken to be a vector $y=(y_1,y_2,\ldots,y_n)$. For the purposes of the present work we will, however, be taking y to be a scalar.

The application of the inverse scattering transform to Eq. (1.1) has been discussed by Calogero and Degasperis.[6,7] More recently, Bogoyavlenskii has shown that this equation admits ''breaking soliton'' solutions.[8] Other examples of PDEs associated with nonisospectral scattering problems in $2+1$ dimensions can be found in Refs. 9, 10, 11. The $2+1$-dimensional extension of the nonlinear Schrödinger equation given in Ref. 9 has also been discussed in Refs. 12 and 13. It is related to the $2+1$-dimensional extension of the classical Boussinesq system given in Ref. 10

in the same way as the $1+1$-dimensional systems.[14] The systems constructed in Ref. 11 are multicomponent generalizations of Eq. (1.1) above; more recently we have looked at modifications of such systems.

The aim of the present paper is to show that PDEs such as (1.1) above, or in fact generalizations thereof, play a key role in the construction of hierarchies of integrable equations and their associated linear problems. This is one of the reasons why we believe that the study of PDEs having nonisospectral scattering problems is of importance. Another is the connection that exists between nonisospectral scattering problems and linear problems for ODEs,[15] which we exploit (and generalize) below. Other reasons include the information they give about the Painlevé analysis (including truncation) of whole hierarchies of differential equations, and also the new nontrivial deformations of well-known PDEs that can arise from such scattering problems.[16]

In what follows we derive a wide variety of new integrable PDEs and ODEs. In their most general formulation these include nonlocal terms. We also obtain a wide variety of corresponding linear problems; nonisospectral scattering problems for hierarchies of PDEs in $2+1$ dimensions; nonisospectral and isospectral scattering problems for hierarchies of PDEs in $1+1$ dimensions; and also monodromy problems for hierarchies of ODEs.

The layout of the paper is as follows. In Sec. II we introduce the main ideas underlying our approach within the context of the KdV and nonisospectral KdV hierarchies. We show how these hierarchies can be characterized using a single equation. These ideas are then further developed in Sec. III, again within the context of KdV-type scattering problems. We use a single PDE in $2+1$ dimensions, and its associated nonisospectral scattering problem, to construct whole hierarchies of both PDEs and ODEs, together with their corresponding hierarchies of underlying linear problems. The extension of our ideas to third-order scattering problems is made in Secs. IV and V, which deal, respectively, with linear problems based on those for the Kaup–Kupershmidt (KK) (Refs. 17, 18), and Sawada–Kotera (SK) (Refs. 19, 20) hierarchies. In these sections we introduce new hierarchies of PDEs in $2+1$ and $1+1$ dimensions, together with their underlying linear problems, and also new hierarchies of ODEs, again with their underlying linear problems. That is, in addition to new hierarchies of PDEs, and the known hierarchies of ODEs which arise directly as similarity reductions of the KK and SK hierarchies,[21,22] we also obtain four new hierarchies of ODEs. These hierarchies of ODEs govern special integrals of similarity reductions of the modified KK/SK hierarchy, and we believe them to be analagous to the hierarchy of the first Painlevé equation P_I.[22] The third order linear problems given here for these hierarchies of ODEs are all new. Our final section is devoted to conclusions.

II. HIERARCHIES FROM NONISOSPECTRAL SCATTERING

In this section we describe our main ideas within the context of the KdV hierarchy. Our claim is that instead of considering Eq. (1.1) as simply an extension of the KdV equation to two spatial dimensions, that in fact this equation embodies information about the entire KdV hierarchy, and also about the entire nonisospectral KdV hierarchy. This is done by considering the extra variable not as a spatial variable, but as a temporal variable, say τ, and recalling that (1.1) can be written as

$$U_t = \mathcal{R} U_\tau, \tag{2.1}$$

where

$$\mathcal{R} = \partial_x^2 + 4U + 2U_x \partial_x^{-1} \tag{2.2}$$

is the recursion operator of the KdV hierarchy.[6,23] [Equation (1.1) is then recovered by the change of variables $U = u_x$, and relabeling τ as y.] In (2.2) $\partial_x \equiv \partial/\partial x$. We use such partial differential operators in what follows, and without further comment, also in the case of ODEs; similarly for the notation U_x.

Given Eq. (2.1) together with its Lax pair, we can then write down the Lax pair of any member of the KdV or nonisospectral KdV hierarchies simply by specifying t and τ to be appropriate flow times. Equation (2.1) gives both the iteration required and also our base equation (the starting point for our iteration). Thus, taking $t=t_5$ and $\tau=t_3$, the fifth order flow of each of these hierarchies can be written as $U_{t_5}=\mathcal{R}U_{t_3}$, for which we have the Lax pair,

$$\psi_{xx}+(U-\lambda)\psi=0,\quad \psi_{t_5}=4\lambda\psi_{t_3}+2[\partial_x^{-1}U_{t_3}]\psi_x-U_{t_3}\psi, \tag{2.3}$$

where

$$\lambda_{t_5}=4\lambda\lambda_{t_3}. \tag{2.4}$$

This gives the iteration between the fifth order flow and the third order flow. It is this third order flow which is our base equation, and which then gives the evolution of U and ψ with respect to t_3 (we could also specify an evolution with respect to t_1 as our base equation). For the nonisospectral KdV hierarchy we take $t=t_3$ and $\tau=y$ in (2.1), which then gives $U_{t_3}=\mathcal{R}U_y$, $\psi_{t_3}=4\lambda\psi_y+2[\partial_x^{-1}U_y]\psi_x-U_y\psi$, and $\lambda_{t_3}=4\lambda\lambda_y$. Combining these last with Eqs. (2.3), (2.4) then gives the fifth order nonisospectral flow $U_{t_5}=\mathcal{R}^2U_y$ and its Lax pair

$$\psi_{xx}+(U-\lambda)\psi=0,\quad \psi_{t_5}=(4\lambda)^2\psi_y+2[\partial_x^{-1}(4\lambda U_y+\mathcal{R}U_y)]\psi_x-[4\lambda U_y+\mathcal{R}U_y]\psi, \tag{2.5}$$

where λ now satisfies the constraint $\lambda_{t_5}=(4\lambda)^2\lambda_y$. For the isospectral KdV hierarchy we set $t=t_3$ in (2.1) and take the reduction $\partial_\tau=\partial_x$. This then gives $U_{t_3}=\mathcal{R}U_x$, $\psi_{t_3}=(4\lambda+2U)\psi_x-U_x\psi$, and $\lambda_{t_3}=0$. Thus we obtain the well known fifth order KdV flow, $U_{t_5}=\mathcal{R}^2U_x$, together with its Lax pair (isospectral since $\lambda_{t_5}=0$).

Now, it is a basic technique within the theory of integrable systems to specify a spectral problem of a certain type, and to expand the coefficients as polynomials in the spectral parameter, thus generating hierarchies of PDEs. This was first done by Ablowitz, Kaup, Newell, and Segur in their seminal paper of 1974,[24] and was in fact the technique we used in Ref. 11 to generate multicomponent hierarchies of PDEs in $2+1$ dimensions associated with nonisospectral scattering problems. Here we have essentially reversed this process; given a recursion operator $\mathcal{R}$, we characterize the corresponding hierarchy using a single equation of the form (2.1), and in this way generate the corresponding hierarchy of linear problems. It is interesting that this connection between relations of the form $U_{t_5}=\mathcal{R}U_{t_3}$, implicit whenever we write down higher order members of a hierarchy, and PDEs of the form $U_t=\mathcal{R}U_\tau$, having nonisospectral scattering problems, has not been exploited before.

In fact the equations we will be using are more general than (2.1), since we are interested not only in obtaining new hierarchies of PDEs, but also in their reductions to ODEs. The advantage of our approach is that it gives immediately a recursion relation between the coefficients of scattering problems corresponding to successive members of a hierarchy; this makes itself felt when dealing with scattering problems of order three or higher. In what follows we do not take different choices of base equation as we did above, but rather we take one base equation in $2+1$ dimensions. We then look at reductions or special cases of the resulting hierarchy.

III. THE KORTEWEG–de VRIES CASE

In this section we illustrate our approach using a generalization of the scattering problem for Eq. (2.1) above. That is, instead of (2.1) we consider the equation

$$U_t=\mathcal{R}U_\tau+g, \tag{3.1}$$

where $g=g(\tau,t)$, i.e., g is a function of all possible times and y, but not of x. This PDE has the Lax pair,

$$\psi_{xx}+(U-\lambda)\psi=0, \quad \psi_t=4\lambda\psi_\tau+[2\partial_x^{-1}U_\tau]\psi_x-U_\tau\psi, \tag{3.2}$$

where the spectral parameter $\lambda=\lambda(\tau,t)$, again to be thought of as a function of all possible times and y, but not of x, satisfies

$$\lambda_t=4\lambda\lambda_\tau+g. \tag{3.3}$$

We also note that this PDE admits the Darboux transformation (DT),

$$\tilde{U}=U+2(\ln\psi)_{xx} \tag{3.4}$$

between two solutions U and $\tilde{U}$. Given this DT and the Lax pair, it is a simple matter to write down a Bäcklund transformation (BT) for (3.1).

We now consider the form taken by higher members of a hierarchy, and their corresponding linear problems, if we use (3.1) both as a starting point and also to iterate between successive flows. The reasons why we are interested in the particular sequence generated by (3.1) will become clear shortly.

We label the times of our hierarchy as t_{2n+1}, and write a generic member of the hierarchy as

$$U_{t_{2n+1}}=Q_n, \tag{3.5}$$

for some (in general nonlocal) functional Q_n of U, and we write corresponding generic evolutions of ψ and λ as

$$\psi_{t_{2n+1}}=\Gamma_n\psi_y+2P_n\psi_x-P_{n,x}\psi, \quad \lambda_{t_{2n+1}}=\Lambda_n. \tag{3.6}$$

It is then easy to see that (3.1), (3.2), and (3.3) yield the recursion relations

$$Q_n=\mathcal{R}Q_{n-1}+g_n, \tag{3.7}$$

$$\Gamma_n=4\lambda\Gamma_{n-1}, \tag{3.8}$$

$$P_n=4\lambda P_{n-1}+\partial_x^{-1}Q_{n-1}, \tag{3.9}$$

$$\Lambda_n=4\lambda\Lambda_{n-1}+g_n \tag{3.10}$$

(note that we also iterate on the function g in (3.1)).

These are the recursion relations between successive flows of the hierarchy and the coefficients of their corresponding linear problems. Since we take as our starting point the PDE,

$$U_{t_3}=\mathcal{R}U_y+g_1, \tag{3.11}$$

along with its corresponding scattering problem (which gives the evolution of ψ and λ with t_3), we obtain in this way the hierarchy of evolution equations

$$U_{t_{2n+1}}=Q_n=\mathcal{R}^nU_y+\sum_{i=1}^{n}g_i\mathcal{R}^{n-i}1. \tag{3.12}$$

This has the corresponding hierarchy of nonisospectral scattering problems

$$\psi_{xx}+(U-\lambda)\psi=0, \quad \psi_{t_{2n+1}}=(4\lambda)^n\psi_y+2P_n\psi_x-P_{n,x}\psi, \tag{3.13}$$

with

$$P_n = \partial_x^{-1} \sum_{i=0}^{n-1} (4\lambda)^{n-1-i} Q_i, \tag{3.14}$$

where we have set $Q_0 = U_y$, and where λ satisfies

$$\lambda_{t_{2n+1}} = (4\lambda)^n \lambda_y + \sum_{i=1}^{n} (4\lambda)^{n-i} g_i. \tag{3.15}$$

The flows of the hierarchy (3.12) are nonlocal, because of the action of powers of $\mathcal{R}$ on U_y and also on 1. This last yields nonautonomous terms, whose inclusion here is new. These flows can be made local by introducing suitable auxiliary dependent variables. The DT (3.4) holds, of course, for every member of this hierarchy, and thus we can easily write down a BT. Allowing τ in (3.1) to be a vector would allow the construction of a hierarchy consisting of linear combinations of the above flows.

In the special case where all $g_i \equiv 0$, the above hierarchy is the one-component breaking soliton hierarchy. The DT (3.4) had previously been shown to hold for this special case in Ref. 25. Making the reduction $\partial_y = \partial_x$ in this special case just obtains the KdV hierarchy, for which the hierarchy of linear problems and DT is of course well-known. We now consider other subcases and reductions.

In addition to nonisospectral scattering problems, our interest here is in obtaining hierarchies of ODEs and their linear problems. From (3.12) we consider two alternative routes to a hierarchy of ODEs, via two different hierarchies of PDEs in $1+1$ dimensions. Our first reduction is via $\partial_y = \partial_x$, which requires $g_{i,y} = \lambda_y = 0$, and which leads to a nonisospectral deformation of the KdV hierarchy,

$$U_{t_{2n+1}} = \mathcal{R}^n U_x + \sum_{i=1}^{n} g_i \mathcal{R}^{n-i} 1. \tag{3.16}$$

This hierarchy could have been obtained immediately through a suitable choice of base equation (in $1+1$ dimensions) and corresponding nonisospectral Lax pair, i.e., $U_{t_3} = \mathcal{R} U_x + g_1$, $\psi_{xx} + (U - \lambda)\psi = 0$, $\psi_{t_3} = (4\lambda + 2U)\psi_x - U_x \psi$, and $\lambda_{t_3} = g_1$. We note that this hierarchy of evolution equations in $1+1$ dimensions can also be found in Ref. 26.

Our second reduction is via $\partial_{t_{2n+1}} = 0$, which then requires $g_{i,t_{2n+1}} = \lambda_{t_{2n+1}} = 0$, and leads to a nonisospectral deformation of inverse KdV flows,

$$\mathcal{R}^n U_y + \sum_{i=1}^{n} g_i \mathcal{R}^{n-i} 1 = 0. \tag{3.17}$$

The Lax pairs of (3.16) and (3.17) follow immediately from those for our $2+1$ hierarchy.

We now consider reductions to ODEs. For the case where an ODE arises as a stationary reduction of an evolution equation having a nonisospectral scattering problem, the connection between the linear problem for that ODE and the nonisospectral scattering problem was discussed in Ref. 15 (e.g., for the case $n=1$ of (3.16)). However in this paper we will be interested in more than just stationary reductions, and in more than just reductions of evolution equations. In Secs. IV and V we deal with hierarchies of ODEs that apparently do not arise simply as the (integrated) stationary flows of completely integrable evolution equations. The simplest examples of such ODEs are fourth order.

For the PDEs (3.16) and (3.17) we consider respectively the reductions $\partial_{t_{2n+1}} = 0$ and $\partial_y = \partial_x$. All g_i are now of course just constant parameters. Each of these reductions obtains the sequence of ODEs,

$$\mathcal{R}^n U_x + \sum_{i=1}^{n} g_i \mathcal{R}^{n-i} 1 = 0. \tag{3.18}$$

A corresponding sequence of linear problems for these ODEs is obtained from the Lax pairs for (3.16) and (3.17); the reduction from (3.17) requires a generalization of the approach in Ref. 15. This sequence of linear problems is

$$\psi_{xx} + (U-\lambda)\psi = 0, \quad \left(\sum_{i=1}^{n} (4\lambda)^{n-i} g_i \right) \psi_\lambda = [(4\lambda)^n + 2P_n]\psi_x - P_{n,x}\psi, \tag{3.19}$$

where we must assume of course that not all g_i are zero, and where

$$P_n = \partial_x^{-1} \left(\sum_{i=0}^{n-1} (4\lambda)^{n-1-i} \left[\mathcal{R}^i U_x + \sum_{j=1}^{i} g_j \mathcal{R}^{i-j} 1 \right] \right). \tag{3.20}$$

It is then straightforward to write down corresponding matrix linear problems for the hierarchy (3.18).

The first of these ODEs and associated linear problems is equivalent, under a simple change of variables, to the first Painlevé equation and its (scalar or matrix) linear problem, so we do not consider this case further here. For $n>1$, the ODEs (3.18) represent an extension of those discussed in Refs. 27, 28, 21, 29 (for $n=2$ simply by an additive constant).

We note that for $n>1$ the ODEs (3.18) are in the general case nonautonomous, and that for $n>2$ they are nonlocal. This last means that for $n>2$ these ODEs could not be obtained using a standard Painlevé classification at that order. All of these ODEs can of course be written locally by introducing suitable auxiliary dependent variables.

Let us now consider the case where the ODEs (3.18), for $n>1$, are local, i.e., where $g_i=0$ for $i=1,2,...,n-2$. In this case the hierarchy of ODEs becomes

$$\mathcal{R}^n U_x + g_{n-1}(4U + 2xU_x) + g_n = 0. \tag{3.21}$$

For $g_{n-1} \neq 0$ this hierarchy has the first integral

$$F_n F_{n,xx} - \frac{1}{2}(F_{n,x})^2 + \left(2U + \frac{g_n}{2g_{n-1}} \right) F_n^2 = -\frac{1}{2}\left(\frac{1}{2} - \alpha_n \right)^2, \tag{3.22}$$

where the right-hand side is the constant of integration, and where

$$F_n = (\partial_x^{-1}\mathcal{R}^{n-1}U_x) + g_{n-1}x + \left(\sum_{i=0}^{n-2} \left(-\frac{g_n}{g_{n-1}} \right)^{n-i-1} \partial_x^{-1}\mathcal{R}^i U_x \right) + \frac{1}{2}\left(-\frac{g_n}{g_{n-1}} \right)^n. \tag{3.23}$$

The hierarchy (3.22), (3.23) is equivalent under the Bäcklund transformation,

$$U = V_x - V^2 - \frac{1}{4}\left(\frac{g_n}{g_{n-1}} \right), \tag{3.24}$$

$$V = -\frac{1}{2}\left(\frac{F_{n,x} + (1/2) - \alpha_n}{F_n} \right), \tag{3.25}$$

to the generalized P_{II} hierarchy

$$(\partial_x + 2V)F_n \Big|_{U=V_x - V^2 - (g_n/(4g_{n-1}))} + (1/2) - \alpha_n = 0. \tag{3.26}$$

For $g_n=0$ the hierarchy (3.22), (3.23) is the P_{34} hierarchy as defined in Ref. 29, (3.24), (3.25) is the Bäcklund transformation[21,29] onto the corresponding P_{II} hierarchy,[27,5,28] and the hierarchy (3.21) is a similarity reduction of the standard KdV hierarchy $U_{t_{2n+1}}=\mathcal{R}^n U_x$ (Refs. 28, 21, 29) (for which in the case $n=2$ a matrix linear problem is known to the author of Ref. 21). Thus for $g_n\neq 0$ the above results give a generalization of this P_{34} hierarchy, Bäcklund transformation, corresponding P_{II} hierarchy, and KdV similarity reduction. For the special case $n=2$ these results can be found in Ref. 30. Further generalizations are possible (e.g., see Remark One below).

When $g_{n-1}=0$ we can integrate (3.21) immediately to obtain the hierarchy of the first Painlevé equation as defined by Kudryashov,[28] i.e.,

$$(\partial_x^{-1}\mathcal{R}^n U_x)+g_n x=0, \tag{3.27}$$

where our requirement now that $g_n\neq 0$ allows us to set any constant of integration to zero. Again it is straightforward to also give matrix linear problems for this hierarchy (with compatibility condition (3.27) rather than its derivative). We note that matrix linear problems for the first three members of this hierarchy have been obtained in Ref. 31.

Remark one: It is possible to generalize the above results if we allow functions of integration to be introduced under the application of the recursion operator. A simple way to do this is to include an extra term in (3.1) which explicitly represents such an integration, i.e., we consider $U_t=\mathcal{R}U_\tau+aU_x+g$, where $a=a(\tau,t)$. The corresponding nonisospectral Lax pair is $\psi_{xx}+(U-\lambda)\psi=0$, $\psi_t=4\lambda\psi_\tau+[(2\partial_x^{-1}U_\tau)+a]\psi_x-U_\tau\psi$, where λ satisfies the same equation (3.3). The DT (3.4) also holds for this PDE.

Iterating also on a, we obtain corresponding to (3.12) above the hierarchy of PDEs,

$$U_{t_{2n+1}}=\mathcal{R}^n U_y+\sum_{i=1}^{n} a_i\mathcal{R}^{n-i}U_x+\sum_{i=1}^{n} g_i\mathcal{R}^{n-i}1. \tag{3.28}$$

The extra terms correspond of course to the addition of lower order KdV flows. Taking reductions as before, we obtain corresponding generalizations of our hierarchies of PDEs in $1+1$ dimensions, and also of our hierarchy of ODEs. In the local case this generalized hierarchy of ODEs reads

$$\mathcal{R}^n U_x+\sum_{i=2}^{n} a_i\mathcal{R}^{n-i}U_x+g_{n-1}(4U+2xU_x)+g_n=0, \tag{3.29}$$

where g_{n-1}, g_n (not both zero) and all a_i are now constant, and where without loss of generality we have taken $a_1=0$. For $g_n=0$ these ODEs are just similarity reductions of sums of KdV flows. The above results on first integrals of (3.21) are easily generalized to (3.29). The case $n=2$ of (3.29) has recently been obtained by Cosgrove,[30] and is known to have hyperelliptic asymptotics.[32] Linear problems for the hierarchy (3.29) are obtained in the same way as described above.

Remark two: So far we have insisted that not all g_i are zero. If we allow all g_i to vanish, in which case the hierarchy of ODEs (3.29) is autonomous, then, following the approach in Ref. 33, we can use the linear problems to obtain constants of motion. Equivalently, we could also use the corresponding matrix linear problem. For Eq. (3.29) with $g_{n-1}=g_n=0$, this approach then yields $n+1$ constants of integration. These are easily identified as the fluxes corresponding to conserved densities consisting of appropriate linear combinations of the Hamiltonian densities of the KdV hierarchy. We note that in the case $n=2$ our third constant of motion vanishes when rewritten in the coordinates of the corresponding Hénon–Heiles system, which then leaves the two constants of motion obtained in Ref. 33.

IV. THE KAUP–KUPERSHMIDT CASE

In this section we consider a PDE associated with a nonisospectral scattering problem based on a member of the KK hierarchy. We write this PDE as the system

$$U_t = \theta[U]W, \tag{4.1}$$

$$W = K[U]U_\tau - qx, \tag{4.2}$$

where $q = q(\tau,t)$ is a function of all possible times and y, but not of x, and where the operators $\theta[U]$ and $K[U]$ are defined by[34,35,36]

$$\theta[U] = \partial_x^3 + (U\partial_x + \partial_x U), \tag{4.3}$$

$$K[U] = \partial_x^{-1}[\partial_x^5 + 3(\partial_x U\partial_x^2 + \partial_x^2 U\partial_x) + 2(\partial_x^3 U + U\partial_x^3) + 8(\partial_x U^2 + U^2\partial_x)]\partial_x^{-1}. \tag{4.4}$$

The recursion operator of the KK hierarchy is given by $\mathcal{R} = \theta[U]K[U]$, so the above system can be written

$$U_t = \mathcal{R}U_\tau - q(2U + xU_x). \tag{4.5}$$

The Lax pair for the system (4.1), (4.2) is

$$\psi_{xxx} = -2U\psi_x - (U_x - \lambda)\psi, \tag{4.6}$$

$$\psi_t = -27\lambda^2\psi_\tau - 9\lambda[2(\partial_x^{-1}U_\tau)]\psi_{xx} + [9\lambda U_\tau + W]\psi_x - [3\lambda U_{x\tau} + 24\lambda U(\partial_x^{-1}U_\tau) + W_x]\psi, \tag{4.7}$$

where $\lambda = \lambda(\tau,t)$ is a function of all possible times and y, but not of x, and satisfies the equation

$$\lambda_t = -27\lambda^2\lambda_\tau - 3\lambda q. \tag{4.8}$$

The compatibility condition of this Lax pair reads

$$(\psi_{xxx})_t - (\psi_t)_{xxx} \equiv X_1\psi_x + X_0\psi = 0, \tag{4.9}$$

where

$$X_1 \equiv 2(-U_t + \theta[U]W) = 0, \tag{4.10}$$

$$X_0 \equiv (-U_t + \theta[U]W)_x + 3\lambda(-W_x + \partial_x K[U]U_\tau - q) = 0. \tag{4.11}$$

This nonisospectral scattering problem, and the corresponding system of PDEs (4.1), (4.2), appear to be new. This system admits the Darboux transformation

$$\tilde{U} = U + \tfrac{3}{2}[\log(\psi\psi_{xx} - \tfrac{1}{2}\psi_x^2 + U\psi^2)]_{xx}, \tag{4.12}$$

where we also have a corresponding mapping on W, too long to be included here. This DT is of course the same as that which holds for the $1+1$ KK hierarchy.[26]

In this section we will construct a new hierarchy of PDEs in $2+1$ dimensions having nonisospectral scattering problems. By taking various reductions to $1+1$ dimensions we will obtain new hierarchies of PDEs having isospectral or nonisospectral scattering problems. By taking further reductions we will obtain new hierarchies of integrable ODEs, together with their underlying linear problems. Explicit examples are given in the case $n=1$ and also in Appendix C for $n=2$.

It is the system (4.1), (4.2) which provides the iteration we use in order to construct our hierarchy of PDEs in $2+1$ dimensions. However we should remember that the KK hierarchy is a double sequence of PDEs, defined as[34,35,36]

$$U_{\tau_m}=\theta[U]H_m[U], \quad H_{m+2}[U]=K[U]\theta[U]H_m[U], \tag{4.13}$$

where the two starting points

$$H_0[U]=1 \quad \text{and} \quad H_1[U]=U_{xx}+4U^2 \tag{4.14}$$

give one sequence beginning with the trivial flow $U_{\tau_0}=U_x$, and another beginning with the fifth order KK equation itself, respectively. However if we use (4.1), (4.2) also as our base equation we will obtain a hierarchy which when we take reductions to $1+1$ dimensions will allow the recovery of only one of the KK sequences (the first mentioned above, i.e., that which includes the first and seventh order flows). This problem can be overcome by making a shift in the system (4.1), (4.2). To do this locally we set $U=u_x$; we then send $u\to u+\frac{1}{2}\partial_y^{-1}a$, where a is a function of all possible times and y, but not of x. This means that as base equation we take the more general $2+1$ system,

$$U_{t_1}=\theta[U]W_1, \tag{4.15}$$

$$W_1=K[U]U_y+a(U_{xx}+4U^2)-q_1x. \tag{4.16}$$

This PDE has the Lax pair,

$$\psi_{xxx}=-2U\psi_x-(U_x-\lambda)\psi, \tag{4.17}$$

$$\begin{aligned}\psi_{t_1}=&-27\lambda^2\psi_y-9\lambda[2(\partial_x^{-1}U_y)+a]\psi_{xx}+[9\lambda U_y+W_1]\psi_x\\&-[3\lambda U_{xy}+24\lambda U(\partial_x^{-1}U_y)+12a\lambda U+W_{1,x}]\psi,\end{aligned} \tag{4.18}$$

where λ satisfies the equation

$$\lambda_{t_1}=-27\lambda^2\lambda_y-3\lambda q_1. \tag{4.19}$$

The extra term included in (4.16) is of course just a copy of $H_1[U]$. This means that reductions to $1+1$ dimensions of this t_1 flow will include both the τ_2 and τ_1 (seventh and fifth order) flows of the standard KK hierarchy (4.13), (4.14). Similarly, the t_n flow of the $2+1$ nonisospectral hierarchy that we now construct will include amongst its reductions to $1+1$ dimensions both the τ_{2n} and τ_{2n-1} flows of the standard KK hierarchy. The DT (4.12) holds also for the system (4.15), (4.16).

Following the approach developed in the last section, we consider a generic member of this $2+1$ hierarchy,

$$U_{t_n}=\theta[U]W_n, \tag{4.20}$$

and write corresponding generic evolutions of ψ and λ as

$$\psi_{t_n}=\Gamma_n\psi_y+A_n\psi_{xx}+B_n\psi_x+C_n\psi, \tag{4.21}$$

$$\lambda_{t_n}=\Lambda_n \tag{4.22}$$

where the compatibility condition of (4.21) with (4.6) tells us that we may take

$$C_n=\tfrac{1}{3}(4UA_n-A_{n,xx}-3B_{n,x}). \tag{4.23}$$

We then obtain from (4.1), (4.2), (4.7), and (4.8) the recursion relations

$$W_n = K[U]\theta[U]W_{n-1} - q_n x, \tag{4.24}$$

$$\Gamma_n = -27\lambda^2\Gamma_{n-1}, \tag{4.25}$$

$$A_n = -27\lambda^2 A_{n-1} - 18\lambda\,\partial_x^{-1}\theta[U]W_{n-1}, \tag{4.26}$$

$$B_n = -27\lambda^2 B_{n-1} + 9\lambda\,\theta[U]W_{n-1} + W_n, \tag{4.27}$$

$$\Lambda_n = -27\lambda^2\Lambda_{n-1} - 3\lambda q_n \tag{4.28}$$

(where we have also iterated on the function q in (4.2)). These recursion relations, together with the base Eq. (4.15), (4.16), and its scattering problem, then yield a hierarchy of evolution equations, which for future convenience we note can be written as

$$U_{t_n} = \theta[U]W_n, \tag{4.29}$$

$$W_n = (K[U]\theta[U])^{n-1}K[U]U_y + a(K[U]\theta[U])^{n-1}H_1[U] - \sum_{i=1}^{n} q_i (K[U]\theta[U])^{n-i}x, \tag{4.30}$$

or equivalently as

$$U_{t_n} = Q_n = \theta[U]W_n = \mathcal{R}^n U_y + a\mathcal{R}^{n-1}\theta[U]H_1[U] - \sum_{i=1}^{n} q_i \mathcal{R}^{n-i}\theta[U]x. \tag{4.31}$$

This hierarchy has the scattering problem

$$\psi_{xxx} = -2U\psi_x - (U_x - \lambda)\psi, \tag{4.32}$$

$$\psi_{t_n} = (-27\lambda^2)^n\psi_y + A_n\psi_{xx} + B_n\psi_x + \tfrac{1}{3}(4UA_n - A_{n,xx} - 3B_{n,x})\psi, \tag{4.33}$$

where

$$A_n = -9a\lambda(-27\lambda^2)^{n-1} - 18\lambda\,\partial_x^{-1}\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i, \tag{4.34}$$

$$B_n = 9\lambda\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i + \sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i, \tag{4.35}$$

where Q_i is as given by (4.31) and we have set $Q_0 = U_y$. The spectral parameter λ satisfies

$$\lambda_{t_n} = (-27\lambda^2)^n\lambda_y - 3\lambda\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i. \tag{4.36}$$

Note that in the above

$$B_n = -\frac{1}{2}A_{n,x} + \sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i. \tag{4.37}$$

The first term on the right-hand side of (4.31) represents a nonisospectral extension to $2+1$ dimensions of the standard KK sequence which includes the seventh order flow. The second term is the standard $1+1$-dimensional KK sequence based on the fifth order flow. The last term represents a nonisospectral deformation which gives rise to nonautonomous terms; for $n>1$, these are in the general case nonlocal. This $2+1$ hierarchy can be written locally using suitable auxiliary dependent variables. The DT (4.12) holds, of course, for every member of this hierarchy. Allowing τ in (4.1), (4.2) to be a vector would allow us to construct a hierarchy consisting of linear combinations of the above flows.

This hierarchy of PDEs in $2+1$ dimensions is new. We now consider reductions of this hierarchy to $1+1$ dimensions and also to ODEs. In both cases we again obtain new hierarchies of equations together with their underlying linear problems.

A. Reductions to hierarchies of PDEs in 1+1 dimensions

In this subsection we discuss reductions to $1+1$ dimensions of our $2+1$ hierarchy. We begin by looking at reductions which give nonisospectral deformations of the standard KK hierarchy (4.13), (4.14).

Our first reduction is via $\partial_y=\partial_x$ and $a=0$. This then requires $q_{i,y}=\lambda_y=0$ and leads to the sequence of equations

$$U_{t_n}=\theta[U]W_n, \tag{4.38}$$

$$W_n=H_{2n}[U]-\sum_{i=1}^{n} q_i(K[U]\theta[U])^{n-i}x, \tag{4.39}$$

which we can also write as

$$U_{t_n}=\mathcal{R}^n U_x-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x. \tag{4.40}$$

This is a nonisospectral deformation of the standard KK sequence which includes the trivial and seventh order flows ($n=0$ and $n=1$, respectively). The Lax pair follows easily from that for our $2+1$ system.

Our second reduction is via $\partial_y=0$ and $a=1$. This requires $q_{i,y}=\lambda_y=0$ and leads to

$$U_{t_n}=\theta[U]W_n, \tag{4.41}$$

$$W_n=H_{2n-1}[U]-\sum_{i=1}^{n} q_i(K[U]\theta[U])^{n-i}x, \tag{4.42}$$

which we can also write as

$$U_{t_n}=\mathcal{R}^{n-1}\theta[U]H_1[U]-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x. \tag{4.43}$$

This is a nonisospectral deformation of the standard KK sequence which includes fifth order KK itself $(n=1)$. Again the Lax pair follows easily from that for our $2+1$ system.

We see from the above that the t_n flow has as reductions nonisospectral deformations of both the τ_{2n} and τ_{2n-1} flows of the KK hierarchy. For $n>1$ the additional nonautonomous terms are in the general case nonlocal. In the special case where all $q_i=0$ we just recover the standard $1+1$-dimensional KK hierarchy.

We now look at alternative reductions of our $2+1$ system to hierarchies in $1+1$ dimensions. We begin by taking the reduction $\partial_{t_n}=0$, which requires $a_{t_n}=q_{i,t_n}=\lambda_{t_n}=0$, and which leads to the system,

$$0=\theta[U]W_n\,, \tag{4.44}$$

$$W_n=(K[U]\theta[U])^{n-1}K[U]U_y+a(K[U]\theta[U])^{n-1}H_1[U]-\sum_{i=1}^{n} q_i(K[U]\theta[U])^{n-i}x. \tag{4.45}$$

This corresponds to a nonisospectral deformation of inverse KK flows,

$$\mathcal{R}^nU_y+a\mathcal{R}^{n-1}\theta[U]H_1[U]-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x=0. \tag{4.46}$$

If we make the further reduction $W_n=0$ in (4.44), (4.45) then we obtain the hierarchy

$$(K[U]\theta[U])^{n-1}K[U]U_y+a(K[U]\theta[U])^{n-1}H_1[U]-\sum_{i=1}^{n} q_i(K[U]\theta[U])^{n-i}x=0, \tag{4.47}$$

which again in the general case is nonisospectral. The Lax pairs for Eqs. (4.46) and (4.47) follow from the Lax pairs for our $2+1$-dimensional system. In the special case where all $q_i=0$ these Lax pairs are isospectral.

B. Reductions to hierarchies of ODEs

We now consider reductions from the $1+1$-dimensional systems given in the previous section to hierarchies of integrable ODEs. In what follows we consider reductions to four different hierarchies of ODEs. The local cases of two of these just correspond to similarity reductions of the two sequences of the KK hierarchy, and are known.[21,22] The local cases of the other two hierarchies appear to be analogous to the P_{I} hierarchy and are new. Only the first member of each of these, which are P_{I} and a fourth order equation found by Cosgrove,[30] was known previously. The results presented here for this fourth order ODE, namely, integrable PDEs from which it can be obtained as a reduction, and the underlying linear problems for those PDEs and also for the ODE, are all new. Similarly for every member of these two new sequences of ODEs (except of course for P_{I}).

We begin by considering the reduction $\partial_{t_n}=0$ of the hierarchies (4.40) and (4.43), which gives

$$\mathcal{R}^nU_x-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x=0 \tag{4.48}$$

and

$$\mathcal{R}^{n-1}\theta[U]H_1[U]-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x=0, \tag{4.49}$$

respectively. All q_i are now of course just constant parameters. The inclusion of the nonlocal terms in the above two hierarchies is new. The same two sequences of ODEs can also be obtained from the reductions $\partial_y=\partial_x$ and $a=0$, and $\partial_y=0$ and $a=1$, respectively, of (4.46). Both of these reductions constitute a generalization of the approach in Ref. 15.

The linear problem for (4.48) is given by

$$\psi_{xxx}=-2U\psi_x-(U_x-\lambda)\psi, \tag{4.50}$$

$$-3\lambda\left(\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i\right)\psi_\lambda=A_n\psi_{xx}+[(-27\lambda^2)^n+B_n]\psi_x+\frac{1}{3}(4UA_n-A_{n,xx}-3B_{n,x})\psi, \tag{4.51}$$

where we must assume of course that not all q_i are zero, and where

$$A_n=-18\lambda\,\partial_x^{-1}\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i\,, \tag{4.52}$$

$$B_n=9\lambda\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i+\sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i\,, \tag{4.53}$$

W_i is as given by (4.39) and $Q_i=\theta[U]W_i$ $(Q_0=U_x)$.

The linear problem for (4.49) is given by

$$\psi_{xxx}=-2U\psi_x-(U_x-\lambda)\psi, \tag{4.54}$$

$$-3\lambda\left(\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i\right)\psi_\lambda=A_n\psi_{xx}+B_n\psi_x+\frac{1}{3}(4UA_n-A_{n,xx}-3B_{n,x})\psi, \tag{4.55}$$

where again we must assume that not all q_i are zero, and where

$$A_n=-9\lambda(-27\lambda^2)^{n-1}-18\lambda\,\partial_x^{-1}\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}Q_i\,, \tag{4.56}$$

$$B_n=9\lambda\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}Q_i+\sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i\,, \tag{4.57}$$

W_i is as given by (4.42) and $Q_i=\theta[U]W_i$ (and we have used the fact that $Q_0=0$). It is then straightforward to give corresponding matrix linear problems for the hierarchies (4.48) and (4.49).

In the local case $q_i=0$, $i=1,...,n-1$, the sequences of ODEs (4.48) and (4.49) become

$$\mathcal{R}^nU_x-q_n(2U+xU_x)=0, \tag{4.58}$$

and

$$\mathcal{R}^{n-1}\theta[U]H_1[U]-q_n(2U+xU_x)=0, \tag{4.59}$$

where our requirement now is that $q_n\neq 0$. These ODEs are just similarity reductions of the KK hierarchy.[21,22] First integrals for every member of these hierarchies are given in Refs. 21, 22, and also in Ref. 30 for the case $n=1$ of (4.59). This case $n=1$ of (4.59) was derived in Ref. 30 using Painlevé classification.

Let us now consider reductions to what are arguably more interesting sequences of ODEs. Instead of simply taking the reduction $\partial_{t_n}=0$ in (4.38), (4.39) and (4.41), (4.42), we take $\partial_{t_n}=0$ and also $W_n=0$ to obtain

$$H_{2n}[U]-\sum_{i=1}^{n}q_i(K[U]\theta[U])^{n-i}x=0 \tag{4.60}$$

and

$$H_{2n-1}[U]-\sum_{i=1}^{n} q_i(K[U]\theta[U])^{n-i}x=0, \tag{4.61}$$

respectively, where again all q_i are just constant parameters. The sequences (4.60) and (4.61) can also be obtained from the reductions $\partial_y=\partial_x$ and $a=0$, and $\partial_y=0$ and $a=1$, respectively, of (4.47). These sequences of ODEs, except for the case $n=1$ of (4.61) which is just the first Painlevé equation P_I, apparently do not arise simply as the (integrated) stationary flow of hierarchies of completely integrable evolution equations. The derivation of the corresponding linear problems is therefore nontrivial, and requires a generalization of the approach in Ref. 15.

The linear problem for (4.60) is given by that for (4.48) but with $W_n=0$. This we can write in the form (here A_n and B_n have been divided by -3λ, and so are different from those defined above)

$$\psi_{xxx}=-2U\psi_x-(U_x-\lambda)\psi, \tag{4.62}$$

$$\left(\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i\right)\psi_\lambda=A_n\psi_{xx}+[9\lambda(-27\lambda^2)^{n-1}+B_n]\psi_x+\frac{1}{3}(4UA_n-A_{n,xx}-3B_{n,x})\psi, \tag{4.63}$$

where we assume that not all q_i are zero, and where

$$A_n=6\partial_x^{-1}\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i\,, \tag{4.64}$$

$$B_n=-3\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i+9\lambda\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}W_i\,, \tag{4.65}$$

W_i is as given by (4.39) and $Q_i=\theta[U]W_i$ (and $Q_0=U_x$).

The linear problem for (4.61) is given by that for (4.49) but again with $W_n=0$. This we rewrite as (again A_n and B_n have been divided by -3λ)

$$\psi_{xxx}=-2U\psi_x-(U_x-\lambda)\psi, \tag{4.66}$$

$$\left(\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i\right)\psi_\lambda=A_n\psi_{xx}+B_n\psi_x+\frac{1}{3}(4UA_n-A_{n,xx}-3B_{n,x})\psi, \tag{4.67}$$

where again we must assume that not all q_i are zero, and where

$$A_n=3(-27\lambda^2)^{n-1}+6\partial_x^{-1}\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}Q_i\,, \tag{4.68}$$

$$B_n=-3\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}Q_i+9\lambda\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}W_i\,, \tag{4.69}$$

W_i is as given by (4.42) and $Q_i=\theta[U]W_i$ (and $Q_0=0$).

In the local case $q_i=0$, $i=1,...,n-1$, the sequences of ODEs (4.60) and (4.61) become

$$H_{2n}[U]-q_nx=0, \tag{4.70}$$

and

$$H_{2n-1}[U]-q_nx=0, \tag{4.71}$$

where our requirement now is that $q_n \neq 0$. Corresponding to the above linear problems we can give matrix linear problems with compatibility conditions Eqs. (4.70) and (4.71), rather than their derivatives. For example, that for the case $n=1$ of (4.70) is given in Appendix A. These hierarchies of ODEs (and of course the nonlocal extensions given above) are new. They govern special integrals of similarity reductions of the modified KK/SK hierarchy, and we believe them to be analogous to the P_{I} hierarchy.[22] A further two such sequences can be derived from the SK hierarchy (see Sec. V). Only two of the above ODEs, namely, the case $n=1$ of each sequence, are known. The case $n=1$ of (4.70) is a special case of an ODE obtained by Cosgrove using Painlevé classification.[30] The linear problem given above is new. The case $n=1$ of (4.71) is just the first Painlevé equation P_{I}. This ODE does of course have a second order linear problem. However here it appears as the first member of a sequence of ODEs for which we can give third order linear problems.

Remark one: It is possible to generalize the above results if we allow functions of integration to be included under the application of the recursion operator. This can be done by including extra terms in (4.1), (4.2) which explicitly represent the two integrations in the operator $K[U]$ in $\mathcal{R}$. Thus instead of (4.1), (4.2) we could consider the system

$$U_t = \theta[U]W + bU_x, \tag{4.72}$$

$$W = K[U]U_\tau + a(U_{xx} + 4U^2) - qx, \tag{4.73}$$

where $a=a(\tau,t)$ and $b=b(\tau,t)$ are functions of all possible times and y, but not of x. This system has the Lax pair

$$\psi_{xxx} = -2U\psi_x - (U_x - \lambda)\psi, \tag{4.74}$$

$$\begin{aligned}\psi_t = &-27\lambda^2\psi_\tau - 9\lambda[2(\partial_x^{-1}U_\tau) + a]\psi_{xx} + [9\lambda U_\tau + W + b]\psi_x \\ &- [3\lambda U_{x\tau} + 24\lambda U(\partial_x^{-1}U_\tau) + 12a\lambda U + W_x]\psi,\end{aligned} \tag{4.75}$$

where λ satisfies the same Eq. (4.8). The DT (4.12) also holds for this system.

Iterating also on a and b we obtain corresponding to (4.29), (4.30) the hierarchy

$$U_{t_n} = \theta[U]W_n + b_n U_x, \tag{4.76}$$

$$\begin{aligned}W_n = &(K[U]\theta[U])^{n-1}K[U]U_y + \sum_{i=1}^{n} a_i H_{2n-2i+1}[U] \\ &+ \sum_{i=1}^{n-1} b_i H_{2n-2i}[U] - \sum_{i=1}^{n} q_i (K[U]\theta[U])^{n-i}x,\end{aligned} \tag{4.77}$$

and so we see that a_1 here is just the function a appearing in (4.29), (4.30). Corresponding to (4.31) we obtain

$$U_{t_n} = \mathcal{R}^n U_y + \sum_{i=1}^{n} b_i \mathcal{R}^{n-i} U_x + \sum_{i=1}^{n} a_i \mathcal{R}^{n-i}\theta[U]H_1[U] - \sum_{i=1}^{n} q_i \mathcal{R}^{n-i}\theta[U]x, \tag{4.78}$$

and so we see that the extra terms correspond of course to the addition of lower order KK flows. Taking reductions as before we obtain corresponding generalizations of our hierarchies of PDEs in $1+1$ dimensions and also of our hierarchies of ODEs, together with their associated linear problems. In the local case (corresponding, respectively, to (4.58), (4.59), (4.70), (4.71)) these generalized hierarchies of ODEs are

$$\mathcal{R}^n U_x + \sum_{i=1}^{n} a_i \mathcal{R}^{n-i}\theta[U]H_1[U] + \sum_{i=1}^{n} b_i \mathcal{R}^{n-i}U_x - q_n(2U + xU_x) = 0, \tag{4.79}$$

$$\mathcal{R}^{n-1}\theta[U]H_1[U] + \sum_{i=2}^{n} a_i \mathcal{R}^{n-i}\theta[U]H_1[U] + \sum_{i=1}^{n} b_i \mathcal{R}^{n-i}U_x - q_n(2U + xU_x) = 0, \tag{4.80}$$

$$H_{2n}[U] + \sum_{i=1}^{n} a_i H_{2n-2i+1}[U] + \sum_{i=1}^{n} b_i H_{2n-2i}[U] - q_n x = 0, \tag{4.81}$$

$$H_{2n-1}[U] + \sum_{i=2}^{n} a_i H_{2n-2i+1}[U] + \sum_{i=1}^{n} b_i H_{2n-2i}[U] - q_n x = 0, \tag{4.82}$$

where in the last two equations we have taken the reduction $W_n = -b_n$ instead of $W_n = 0$. (In fact our requirement that $q_n \neq 0$ allows us to take $b_n = 0$ in all of the above.) The hierarchies (4.79) and (4.80) are just similarity reductions of sums of KK flows; the results of Refs. 21, 22 on first integrals of (4.58) and (4.59) are easily generalized to this case. For $n=1$ (4.81) is equivalent to Cosgrove's equation,[30] including nondominant terms (see Appendix A). Linear problems for the above four hierarchies follow in the same way as described above.

Remark two: So far we have insisted that not all q_i are zero. If we allow all q_i to vanish, in which case the hierarchies of ODEs (4.79), (4.80), (4.81), and (4.82) are autonomous, then, following the approach in Ref. 33, we can use the linear problems to obtain constants of motion. Equivalently, we could also use the corresponding matrix linear problems. For the case $n=1$ of (4.79) and (4.80) with $q_n = 0$ (ODEs of orders 7 and 5 respectively), we obtain in this way two constants of motion; for (4.80) these are as obtained in Ref. 33 for the corresponding Hénon–Heiles system, the first being just the trivial first integral of (4.80), which corresponds to the energy. For the case $n=1$ of (4.81) with $q_n = 0$ (an ODE of order 4) this approach yields one constant of motion, which in addition to a simple first integral then leads to two constants of motion. These are as found in Ref. 30. For (4.82) with $q_n = 0$, the case $n=1$ is trivial. However the case $n=2$ (an ODE of order eight) is less so, and we are able to obtain three constants of motion.

C. Examples: $n=1$

In the case $n=1$ we can set $U = u_x$, and write the system (4.29), (4.30) locally as

$$u_{xt_1} = W_{1,xxx} + 2u_x W_{1,x} + u_{xx} W_1, \tag{4.83}$$

$$W_{1,x} = u_{xxxxxy} + 10u_{xxxy}u_x + 2u_{xxxx}u_y + 15u_{xxy}u_{xx} + 9u_{xxx}u_{xy} + 16u_{xy}u_x^2 + 16u_{xx}u_x u_y + a(u_{xxxx} + 8u_x u_{xx}) - q_1, \tag{4.84}$$

where $a = a(y,t)$ and $q_1 = q_1(y,t)$. This system is of course just our base equation (4.15), (4.16), which has Lax pair (4.17), (4.18) with $\lambda = \lambda(y,t)$ satisfying (4.19). We now consider reductions of this system to PDEs in $1+1$ dimensions, and also to ODEs.

Our first two reductions are to the PDEs,

$$U_{t_1} = W_{1,xxx} + 2UW_{1,x} + U_x W_1, \tag{4.85}$$

$$W_1 = H_2[U] - q_1 x = U_{xxxx} + 12UU_{xx} + 6U_x^2 + \tfrac{32}{3}U^3 - q_1 x, \tag{4.86}$$

and

$$U_{t_1} = W_{1,xxx} + 2UW_{1,x} + U_x W_1, \tag{4.87}$$

$$W_1 = H_1[U] - q_1 x = U_{xx} + 4U^2 - q_1 x, \tag{4.88}$$

where in each case q_1 is a function of t_1 only. These have temporal part of the Lax pair given, respectively, by

$$\psi_{t_1} = -18\lambda U \psi_{xx} + [-27\lambda^2 + 9\lambda U_x + W_1]\psi_x - [3\lambda U_{xx} + 24\lambda U^2 + W_{1,x}]\psi, \tag{4.89}$$

and

$$\psi_{t_1} = -9\lambda \psi_{xx} + W_1 \psi_x - [12\lambda U + W_{1,x}]\psi, \tag{4.90}$$

where in each case λ is a function of t_1 only and $\lambda_{t_1} = -3\lambda q_1$. This last means that for $q_1 \neq 0$ the above Lax pairs are nonisospectral. When $q_1 = 0$ they become isospectral, in which case (4.85), (4.86) and (4.87), (4.88) reduce to seventh and fifth order KK, respectively.

Our second two reductions to PDEs in $1+1$ dimensions are to

$$0 = W_{1,xxx} + 2u_x W_{1,x} + u_{xx} W_1, \tag{4.91}$$

$$\begin{aligned} W_{1,x} = {} & u_{xxxxxy} + 10u_{xxxy}u_x + 2u_{xxxx}u_y + 15u_{xxy}u_{xx} + 9u_{xxx}u_{xy} \\ & + 16u_{xy}u_x^2 + 16u_{xx}u_x u_y + a(u_{xxxx} + 8u_x u_{xx}) - q_1, \end{aligned} \tag{4.92}$$

and to its further reduction

$$\begin{aligned} 0 = {} & u_{xxxxxy} + 10u_{xxxy}u_x + 2u_{xxxx}u_y + 15u_{xxy}u_{xx} + 9u_{xxx}u_{xy} \\ & + 16u_{xy}u_x^2 + 16u_{xx}u_x u_y + a(u_{xxxx} + 8u_x u_{xx}) - q_1, \end{aligned} \tag{4.93}$$

where in each case a and q_1 are functions of y only. The temporal (now y) part of the Lax pair for (4.91), (4.92) is

$$27\lambda^2 \psi_y = -9\lambda[2u_y + a]\psi_{xx} + [9\lambda u_{xy} + W_1]\psi_x - [3\lambda u_{xxy} + 24\lambda u_x u_y + 12a\lambda u_x + W_{1,x}]\psi, \tag{4.94}$$

and that for (4.93) is

$$9\lambda \psi_y = -3[2u_y + a]\psi_{xx} + 3u_{xy}\psi_x - [u_{xxy} + 8u_x u_y + 4au_x]\psi, \tag{4.95}$$

where in each case λ is a function of y only and $\lambda_y = -q_1/(9\lambda)$. For $q_1 \neq 0$ (4.91), (4.92) is a nonisospectral deformation of an inverse KK flow $\mathcal{R}U_y + a\theta[U]H_1[U] - q_1(2U + xU_x) = 0$ and appears to be new. The case $q_1 = 0$ has isospectral Lax pair. Equation (4.93) also appears to be new and again has isospectral or nonisospectral Lax pair according as to whether $q_1 = 0$ or $q_1 \neq 0$, respectively. Details of the Painlevé analysis of (4.93), and also of the system (4.83), (4.84), can be found in Appendix B.

Our first two reductions to ODEs are to $\mathcal{R}U_x - q_1\theta[U]x = 0$, i.e.,

$$\left(U_{xxxxxx} + 14UU_{xxxx} + 35U_x U_{xxx} + \tfrac{49}{2}U_{xx}^2 + 56U^2 U_{xx} + 70UU_x^2 + \tfrac{56}{3}U^4 \right)_x - q_1(2U + xU_x) = 0, \tag{4.96}$$

and to $\theta[U]H_1[U] - q_1\theta[U]x = 0$, i.e.,

$$(U_{xxxx} + 10UU_{xx} + \tfrac{15}{2}U_x^2 + \tfrac{20}{3}U^3)_x - q_1(2U + xU_x) = 0. \tag{4.97}$$

These ODEs are similarity reductions of seventh and fifth order KK, respectively. They have linear problems with ψ_λ given, respectively, by

$$3\lambda q_1 \psi_\lambda = 18\lambda U \psi_{xx} + [27\lambda^2 - 9\lambda U_x - W_1]\psi_x + [3\lambda(U_{xx} + 8U^2) + W_{1,x}]\psi, \tag{4.98}$$

where W_1 is given by (4.86), and

$$3\lambda q_1\psi_\lambda=9\lambda\psi_{xx}-W_1\psi_x+[12\lambda U+W_{1,x}]\psi, \tag{4.99}$$

where W_1 is given by (4.88), and where in each case $q_1\neq 0$. We note that a matrix linear problem for the ODE (4.97) is known to the author of Ref. 21.

Our second two reductions to ODEs are to $H_2[U]-q_1x=0$, i.e.,

$$U_{xxxx}+12UU_{xx}+6U_x^2+\tfrac{32}{3}U^3-q_1x=0, \tag{4.100}$$

and to $H_1[U]-q_1x=0$, i.e.,

$$U_{xx}+4U^2-q_1x=0. \tag{4.101}$$

These ODEs have linear problems with ψ_λ given by

$$q_1\psi_\lambda=6U\psi_{xx}+[9\lambda-3U_x]\psi_x+[U_{xx}+8U^2]\psi, \tag{4.102}$$

and

$$q_1\psi_\lambda=3\psi_{xx}+4U\psi, \tag{4.103}$$

respectively, where in each case $q_1\neq 0$. Equation (4.101) is of course just P_{I}, which has a well-known second order linear problem. Equation (4.100) is a special case of an ODE found by Cosgrove using Painlevé classification,[30] although the dominant terms are in fact due to Harada and Oishi[37] (see also Ito[38]). The full equation is (4.82) with $n=1$; a matrix linear problem for this ODE is given in Appendix A.

V. THE SAWADA–KOTERA CASE

In this Section we consider a PDE associated with a nonisospectral scattering problem based on a member of the SK hierarchy. We note that the structure of the SK hierarchy bears a close resemblence to that of the KK hierarchy. Moreover, these two hierarchies have the same modified hierarchy, and so are related by a Bäcklund transformation.[18,34]

We write our $2+1$-dimensional member of the SK hierarchy as the system

$$U_t=\theta[U]W, \tag{5.1}$$

$$W=J[U]U_\tau-qx, \tag{5.2}$$

where $q=q(\tau,t)$ is a function of all possible times and y, but not of x, and where the operators $\theta[U]$ and $J[U]$ are defined by[34,35,36]

$$\theta[U]=\partial_x^3+(U\partial_x+\partial_xU), \tag{5.3}$$

$$J[U]=\partial_x^{-1}[\partial_x^5+\tfrac{1}{2}(\partial_x^3U+U\partial_x^3)+\tfrac{1}{8}(\partial_xU^2+U^2\partial_x)]\partial_x^{-1}. \tag{5.4}$$

The recursion operator of the SK hierarchy is given by $\mathcal{R}=\theta[U]J[U]$, so the above system can be written

$$U_t=\mathcal{R}U_\tau-q(2U+xU_x). \tag{5.5}$$

The Lax pair for the system (5.1), (5.2) is

$$\psi_{xxx}=-\tfrac{1}{2}U\psi_x+\lambda\psi, \tag{5.6}$$

$$\psi_t = -27\lambda^2\psi_\tau - \tfrac{3}{4}[6\lambda(\partial_x^{-1}U_\tau) - UU_\tau - U_x(\partial_x^{-1}U_\tau) - 2U_{xx\tau}]\psi_{xx}$$
$$+[\tfrac{3}{4}(6\lambda U_\tau - 2U_xU_\tau - UU_{x\tau} - U_{xx}(\partial_x^{-1}U_\tau) - 2U_{xxx\tau}) + W]\psi_x$$
$$-\tfrac{3}{2}\lambda[2U_{x\tau} + U(\partial_x^{-1}U_\tau)]\psi, \tag{5.7}$$

where the spectral parameter λ, again a function of all possible times and y, but not of x, satisfies the equation

$$\lambda_t = -27\lambda^2\lambda_\tau - 3\lambda q. \tag{5.8}$$

The compatibility condition of this Lax pair reads

$$(\psi_{xxx})_t - (\psi_t)_{xxx} \equiv X_2\psi_{xx} + X_1\psi_x + X_0\psi = 0, \tag{5.9}$$

where

$$X_2 \equiv 3(-W_x + \partial_x J[U]U_\tau - q)_x = 0, \tag{5.10}$$

$$X_1 \equiv \tfrac{3}{2}(-W_x + \partial_x J[U]U_\tau - q)_{xx} + \tfrac{1}{2}(-U_t + \theta[U]W) = 0, \tag{5.11}$$

$$X_0 \equiv 3\lambda(-W_x + \partial_x J[U]U_\tau - q) = 0. \tag{5.12}$$

We note that the spatial part of the above Lax pair and that of the KK hierarchy are related as arising from different scalar representations of the scattering problem for the aforementioned common modified hierarchy.[34]

The above nonisospectral scattering problem, and the corresponding system of PDEs (5.1), (5.2), appear to be new. This system admits the Darboux transformation

$$\tilde{U} = U + 12(\log\psi)_{xx}, \tag{5.13}$$

where we also have a corresponding mapping on W, too long to be included here. This DT is of course the same as that which holds for the $1+1$ SK hierarchy.[26]

As in the previous section, we construct a new hierarchy of PDEs in $2+1$ dimensions having nonisospectral scattering problems; by taking various reductions we obtain new hierarchies of PDEs and ODEs together with their underlying linear problems. Once again we give explicit examples in the cases $n=1$ and $n=2$ (these last can be found in Appendix D).

It is the system (5.1), (5.2) which provides the iteration we use in order to construct our hierarchy of PDEs in $2+1$ dimensions. However the SK hierarchy, just like the KK hierarchy, is a double sequence of PDEs. This double sequence is defined as[34,35,36]

$$U_{\tau_m} = \theta[U]G_m[U], \quad G_{m+2}[U] = J[U]\theta[U]G_m[U], \tag{5.14}$$

where the two starting points

$$G_0[U] = 1 \quad \text{and} \quad G_1[U] = U_{xx} + \tfrac{1}{4}U^2 \tag{5.15}$$

give one sequence beginning with the trivial flow $U_{\tau_0} = U_x$, and another beginning with the fifth order SK equation itself, respectively. In order to include reductions to each of the above sequences in our $2+1$ hierarchy, we make, in the same way as we did in the KK case, a shift in the system (5.1), (5.2). This then leads to the more general base equation,

$$U_{t_1} = \theta[U]W_1, \tag{5.16}$$

$$W_1 = J[U]U_y + a(U_{xx} + \tfrac{1}{4}U^2) - q_1 x, \tag{5.17}$$

where a is a function of all possible times and y, but not of x. This PDE has the Lax pair,

$$\psi_{xxx}=-\tfrac{1}{2}U\psi_x+\lambda\psi, \tag{5.18}$$

$$\begin{aligned}\psi_{t_1}=&-27\lambda^2\psi_y\\&-\tfrac{3}{4}[6\lambda(\partial_x^{-1}U_y)-UU_y-U_x(\partial_x^{-1}U_y)-2U_{xxy}+2a(6\lambda-U_x)]\psi_{xx}\\&+[\tfrac{3}{4}(6\lambda U_y-2U_xU_y-UU_{xy}-U_{xx}(\partial_x^{-1}U_y)-2U_{xxxy}-2aU_{xx})\\&+W_1]\psi_x-\tfrac{3}{2}\lambda[2U_{xy}+U(\partial_x^{-1}U_y)+2aU]\psi,\end{aligned} \tag{5.19}$$

where λ satisfies the equation

$$\lambda_{t_1}=-27\lambda^2\lambda_y-3\lambda q_1. \tag{5.20}$$

The extra term included in (5.17) is just a copy of $G_1[U]$. This means that the $2+1$ nonisospectral hierarchy that we now construct will include amongst its reductions to $1+1$ dimensions both the τ_{2n} and τ_{2n-1} flows of the standard SK hierarchy. We also note that the DT (5.13) holds also for the system (5.16), (5.17).

Following the approach developed in Sec. III, we consider a generic member of this $2+1$ hierarchy,

$$U_{t_n}=\theta[U]W_n, \tag{5.21}$$

and write corresponding generic evolutions of ψ and λ as

$$\psi_{t_n}=\Gamma_n\psi_y+A_n\psi_{xx}+B_n\psi_x+C_n\psi, \tag{5.22}$$

$$\lambda_{t_n}=\Lambda_n, \tag{5.23}$$

where the compatibility condition of (5.22) with (5.6) tells us that we may take

$$C_n=\tfrac{1}{3}(UA_n-A_{n,xx}-3B_{n,x}). \tag{5.24}$$

We then obtain from (5.1), (5.2), (5.7), and (5.8) the recursion relations

$$W_n=J[U]\theta[U]W_{n-1}-q_nx, \tag{5.25}$$

$$\Gamma_n=-27\lambda^2\Gamma_{n-1}, \tag{5.26}$$

$$A_n=-27\lambda^2A_{n-1}-\tfrac{9}{2}\lambda\partial_x^{-1}\theta[U]W_{n-1}+\tfrac{3}{4}[U\theta[U]W_{n-1}+U_x\partial_x^{-1}\theta[U]W_{n-1}+2\partial_x^2\theta[U]W_{n-1}], \tag{5.27}$$

$$\begin{aligned}B_n=&-27\lambda^2B_{n-1}+\tfrac{9}{2}\lambda\theta[U]W_{n-1}\\&-\tfrac{3}{4}\partial_x[U\theta[U]W_{n-1}+U_x\partial_x^{-1}\theta[U]W_{n-1}+2\partial_x^2\theta[U]W_{n-1}]+W_n,\end{aligned} \tag{5.28}$$

$$\Lambda_n=-27\lambda^2\Lambda_{n-1}-3\lambda q_n \tag{5.29}$$

(where we have also iterated on the function q in (5.2)). These recursion relations, together with the base Eq. (5.16), (5.17) and its scattering problem, then yield a hierarchy of evolution equations, which for future convenience we note can be written as

$$U_{t_n}=\theta[U]W_n, \tag{5.30}$$

$$W_n=(J[U]\theta[U])^{n-1}J[U]U_y+a(J[U]\theta[U])^{n-1}G_1[U]-\sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x, \tag{5.31}$$

or equivalently as

$$U_{t_n}=Q_n=\theta[U]W_n=\mathcal{R}^nU_y+a\mathcal{R}^{n-1}\theta[U]G_1[U]-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x. \tag{5.32}$$

This hierarchy has the scattering problem

$$\psi_{xxx}=-\tfrac{1}{2}U\psi_x+\lambda\psi, \tag{5.33}$$

$$\psi_{t_n}=(-27\lambda^2)^n\psi_y+A_n\psi_{xx}+B_n\psi_x+C_n\psi, \tag{5.34}$$

where

$$A_n=-\frac{3}{2}a(-27\lambda^2)^{n-1}(6\lambda-U_x)-\frac{9}{2}\lambda\partial_x^{-1}\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i + \frac{3}{4}\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}[UQ_i+U_x\partial_x^{-1}Q_i+2\partial_x^2Q_i], \tag{5.35}$$

$$B_n=-\frac{3}{2}a(-27\lambda^2)^{n-1}U_{xx}+\frac{9}{2}\lambda\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i -\frac{3}{4}\partial_x\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}[UQ_i+U_x\partial_x^{-1}Q_i+2\partial_x^2Q_i] +\sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i, \tag{5.36}$$

$$C_n=-3a\lambda(-27\lambda^2)^{n-1}U-\frac{3}{2}\lambda\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}[U\partial_x^{-1}Q_i+2Q_{i,x}], \tag{5.37}$$

C_n being calculated either directly (by iteration) or from (5.24) modulo equations (5.17) and (5.25) and any additive functions of *y* and *t*, and where Q_i is as given by (5.32), and we have set $Q_0=U_y$. The spectral parameter λ satisfies

$$\lambda_{t_n}=(-27\lambda^2)^n\lambda_y-3\lambda\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i. \tag{5.38}$$

Note that in the above

$$B_n=-A_{n,x}+\sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i. \tag{5.39}$$

The first term on the right-hand side of (5.32) represents a nonisospectral extension to $2+1$ dimensions of the standard SK sequence which includes the seventh order flow. The second term is the standard $1+1$-dimensional SK sequence based on the fifth order flow. The last term represents a nonisospectral deformation which gives rise to nonautonomous terms; for $n>1$, these are in the general case nonlocal. This $2+1$ hierarchy can be written locally using suitable auxiliary dependent variables. The DT (5.13) holds, of course, for every member of this hierarchy. Allowing τ in (5.1), (5.2) to be a vector would allow us to construct a hierarchy consisting of linear combinations of the above flows.

This hierarchy of PDEs in $2+1$ dimensions is new. We now consider reductions of this hierarchy to $1+1$ dimensions and also to ODEs. In both cases we again obtain new hierarchies of equations together with their underlying linear problems.

A. Reductions to hierarchies of PDEs in 1+1 dimensions

As we did above in the KK case, we now discuss reductions to $1+1$ dimensions of our $2+1$ hierarchy. Once again, we begin by looking at reductions which give nonisospectral deformations of the standard flows in $1+1$ dimensions (5.14), (5.15).

Our first reduction is via $\partial_y=\partial_x$ and $a=0$. This then requires $q_{i,y}=\lambda_y=0$ and leads to the sequence of equations

$$U_{t_n}=\theta[U]W_n\,, \tag{5.40}$$

$$W_n=G_{2n}[U]-\sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x, \tag{5.41}$$

which we can also write as

$$U_{t_n}=\mathcal{R}^n U_x-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x. \tag{5.42}$$

This is a nonisospectral deformation of the standard SK sequence which includes the trivial and seventh order flows ($n=0$ and $n=1$, respectively). The Lax pair follows easily from that for our $2+1$ system.

Our second reduction is via $\partial_y=0$ and $a=1$. This requires $q_{i,y}=\lambda_y=0$ and leads to

$$U_{t_n}=\theta[U]W_n\,, \tag{5.43}$$

$$W_n=G_{2n-1}[U]-\sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x, \tag{5.44}$$

which we can also write as

$$U_{t_n}=\mathcal{R}^{n-1}\theta[U]G_1[U]-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x. \tag{5.45}$$

This is a nonisospectral deformation of the standard SK sequence which includes fifth order SK itself $(n=1)$. Again the Lax pair follows easily from that for our $2+1$ system.

We see from the above that the t_n flow has as reductions nonisospectral deformations of both the τ_{2n} and τ_{2n-1} flows of the SK hierarchy. For $n>1$ the additional nonautonomous terms are in the general case nonlocal. In the special case where all $q_i=0$ we just recover the standard $1+1$-dimensional SK hierarchy.

We now look at alternative reductions of our $2+1$ system to hierarchies in $1+1$ dimensions. We begin by taking the reduction $\partial_{t_n}=0$, which requires $a_{t_n}=q_{i,t_n}=\lambda_{t_n}=0$, and which leads to the system,

$$0=\theta[U]W_n\,, \tag{5.46}$$

$$W_n=(J[U]\theta[U])^{n-1}J[U]U_y+a(J[U]\theta[U])^{n-1}G_1[U]-\sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x. \tag{5.47}$$

This corresponds to a nonisospectral deformation of inverse SK flows,

$$\mathcal{R}^nU_y+a\mathcal{R}^{n-1}\theta[U]G_1[U]-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x=0. \tag{5.48}$$

If we make the further reduction $W_n=0$ in (5.46), (5.47) then we obtain the hierarchy

$$(J[U]\theta[U])^{n-1}J[U]U_y+a(J[U]\theta[U])^{n-1}G_1[U]-\sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x=0, \tag{5.49}$$

which again in the general case is nonisospectral. The Lax pairs for Eqs. (5.48) and (5.49) follow from the Lax pairs for our $2+1$-dimensional system. In the special case where all $q_i=0$ these Lax pairs are isospectral.

B. Reductions to hierarchies of ODEs

We now consider reductions from the $1+1$-dimensional systems given in the previous section to hierarchies of integrable ODEs. In what follows we consider reductions to four different hierarchies of ODEs. The local cases of two of these just correspond to similarity reductions of the two sequences of the SK hierarchy, and are known.[21,22] The local cases of the other two hierarchies again appear to be analogous to the P_{I} hierarchy, and are new. The first member of each of these is in fact equivalent under a simple rescaling to the corresponding first member of each of the two new hierarchies of ODEs obtained in Sec. IV. The linear problems and other results given here for these two hierarchies are new (except again for P_{I}).

We begin by considering the reduction $\partial_{t_n}=0$ of the hierarchies (5.42) and (5.45), which gives

$$\mathcal{R}^nU_x-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x=0 \tag{5.50}$$

and

$$\mathcal{R}^{n-1}\theta[U]G_1[U]-\sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x=0, \tag{5.51}$$

respectively, where all q_i are now just constant parameters. The inclusion of the nonlocal terms in the above two hierarchies is new. The same two sequences of ODEs can also be obtained from the reductions $\partial_y=\partial_x$ and $a=0$, and $\partial_y=0$ and $a=1$, respectively, of (5.48). Again these last two reductions constitute a generalization of the approach in Ref. 15.

The linear problem for (5.50) is given by

$$\psi_{xxx}=-\tfrac{1}{2}U\psi_x+\lambda\psi, \tag{5.52}$$

$$-3\lambda\left(\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i\right)\psi_\lambda=A_n\psi_{xx}+[(-27\lambda^2)^n+B_n]\psi_x+C_n\psi, \tag{5.53}$$

where we must assume of course that not all q_i are zero, and where

$$A_n=-\frac{9}{2}\lambda\partial_x^{-1}\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i+\frac{3}{4}\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}[UQ_i+U_x\partial_x^{-1}Q_i+2\partial_x^2Q_i], \tag{5.54}$$

$$B_n=-A_{n,x}+\sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i, \tag{5.55}$$

$$C_n=-\frac{3}{2}\lambda\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}[U\partial_x^{-1}Q_i+2Q_{i,x}], \tag{5.56}$$

W_i is as given by (5.41) and $Q_i=\theta[U]W_i$ ($Q_0=U_x$).

The linear problem for (5.51) is given by

$$\psi_{xxx}=-\tfrac{1}{2}U\psi_x+\lambda\psi, \tag{5.57}$$

$$-3\lambda\left(\sum_{i=1}^{n}(-27\lambda^2)^{n-i}q_i\right)\psi_\lambda=A_n\psi_{xx}+B_n\psi_x+C_n\psi, \tag{5.58}$$

where again we must assume that not all q_i are zero, and where

$$A_n=-\frac{3}{2}(-27\lambda^2)^{n-1}(6\lambda-U_x)-\frac{9}{2}\lambda\partial_x^{-1}\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}Q_i$$
$$+\frac{3}{4}\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}[UQ_i+U_x\partial_x^{-1}Q_i+2\partial_x^2Q_i], \tag{5.59}$$

$$B_n=-A_{n,x}+\sum_{i=1}^{n}(-27\lambda^2)^{n-i}W_i, \tag{5.60}$$

$$C_n=-3\lambda(-27\lambda^2)^{n-1}U-\frac{3}{2}\lambda\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}[U\partial_x^{-1}Q_i+2Q_{i,x}], \tag{5.61}$$

W_i is as given by (5.44) and $Q_i=\theta[U]W_i$ (and we have used the fact that $Q_0=0$). Again it is straightforward to give corresponding matrix linear problems for the hierarchies (5.50) and (5.51).

In the local case $q_i=0$, $i=1,...,n-1$, the sequences of ODEs (5.50) and (5.51) become

$$\mathcal{R}^nU_x-q_n(2U+xU_x)=0, \tag{5.62}$$

and

$$\mathcal{R}^{n-1}\theta[U]G_1[U]-q_n(2U+xU_x)=0, \tag{5.63}$$

where our requirement now is that $q_n\neq 0$. These ODEs are similarity reductions of the SK hierarchy.[21,22] First integrals for every member of these hierarchies are given in Refs. 21, 22, and also in Ref. 30 for the case $n=1$ of (5.63). These first integrals of the above hierarchies are related by a Bäcklund transformation to those of the hierarchies (4.58) and (4.59).[22] The case $n=1$ of (5.63) was derived in Ref. 30 using Painlevé classification.

We now consider further reductions, as we did in Sec. IV, to what we again consider to be perhaps more interesting sequences of ODEs. Thus, instead of simply taking the reduction $\partial_{t_n}=0$ in (5.40), (5.41) and (5.43), (5.44), we take $\partial_{t_n}=0$ and also $W_n=0$ to obtain

$$G_{2n}[U]-\sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x=0 \tag{5.64}$$

and

$$G_{2n-1}[U]-\sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x=0, \tag{5.65}$$

respectively, where again all q_i are constant parameters. The sequences (5.64) and (5.65) can also be obtained from the reductions $\partial_y=\partial_x$, and $a=0$, and $\partial_y=0$ and $a=1$, respectively, of (5.49). As with the two sequences (4.60) and (4.61), it seems that except for the case $n=1$ of (5.65) which is again just the first Painlevé equation P_{I}, these two sequences do not arise as the (integrated) stationary flow of hierarchies of completely integrable evolution equations.

The linear problem for (5.64) is given by that for (5.50) but with $W_n=0$. That is, it is (5.52), (5.53) (where we assume that not all q_i are zero) with A_n given by (5.54), $B_n=-A_{n,x}+\Sigma_{i=1}^{n-1}(-27\lambda^2)^{n-i}W_i$, i.e.,

$$B_n=\frac{9}{2}\lambda\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}Q_i$$
$$-\frac{3}{4}\partial_x\sum_{i=0}^{n-1}(-27\lambda^2)^{n-i-1}[UQ_i+U_x\partial_x^{-1}Q_i+2\partial_x^2Q_i]+\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i}W_i, \tag{5.66}$$

and C_n by (5.56), and where W_i is as given by (5.41) and $Q_i=\theta[U]W_i$ (and $Q_0=U_x$).

The linear problem for (5.65) is given by that for (5.51) but again with $W_n=0$. That is, it is (5.57), (5.58) (where we assume that not all q_i are zero) with A_n given by (5.59), $B_n=-A_{n,x}+\Sigma_{i=1}^{n-1}(-27\lambda^2)^{n-i}W_i$, i.e.,

$$B_n=-\frac{3}{2}(-27\lambda^2)^{n-1}U_{xx}+\frac{9}{2}\lambda\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}Q_i$$
$$-\frac{3}{4}\partial_x\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i-1}[UQ_i+U_x\partial_x^{-1}Q_i+2\partial_x^2Q_i]+\sum_{i=1}^{n-1}(-27\lambda^2)^{n-i}W_i, \tag{5.67}$$

and C_n by (5.61), and where W_i is as given by (5.44) and $Q_i=\theta[U]W_i$ (and $Q_0=0$).

Note that the order of the highest derivative of U appearing in the above two expressions for B_n (in the term $\partial_x^3Q_{n-1}$) is the same as the order of the corresponding ODE ((5.64) or (5.65), respectively), and so this derivative of U can be replaced. In general, the coefficients of W_n and of the highest derivative of U appearing in B_n are inherited from our $2+1$-dimensional scattering problem. In certain of our reductions to PDEs in $1+1$ dimensions and to ODEs, where we can express W_n locally, we can then eliminate either W_n or this highest derivative of U, one in favor of the other, if we so wish. Of the reductions dealt with in the current paper, it is only in the case of those to the ODEs (5.64) and (5.65) that the linear problem contains derivatives of U of the same order as the ODE, which can then be replaced (see our examples later). Doing so then gives a scalar linear problem with compatibility condition (5.64) or (5.65), rather than their derivatives.

In the local case $q_i=0$, $i=1,...,n-1$, the sequences of ODEs (5.64) and (5.65) become

$$G_{2n}[U]-q_nx=0, \tag{5.68}$$

and

$$G_{2n-1}[U]-q_nx=0, \tag{5.69}$$

where our requirement now is that $q_n \neq 0$. Corresponding to the above linear problems we can give matrix linear problems with compatibility conditions (5.68) and (5.69). These hierarchies of ODEs are new, as also are the nonlocal extensions given above. They govern special integrals of similarity reductions of the modified KK/SK hierarchy, and we believe them to be analogous to the P_{I} hierarchy.[22] The first member of each of the above hierarchies is equivalent under a simple rescaling to the corresponding first member of (4.70) and (4.71).

Remark one: Again we can generalize the above results if we allow functions of integration to be included under the application of the recursion operator. Thus instead of (5.1), (5.2) we could consider the system

$$U_t = \theta[U]W + bU_x, \tag{5.70}$$

$$W = J[U]U_\tau + a(U_{xx} + \tfrac{1}{4}U^2) - qx, \tag{5.71}$$

where $a = a(\tau,t)$ and $b = b(\tau,t)$ are functions of all possible times and y, but not of x. This system has the Lax pair

$$\psi_{xxx} = -\tfrac{1}{2}U\psi_x + \lambda\psi, \tag{5.72}$$

$$\begin{aligned}\psi_t = &-27\lambda^2\psi_\tau \\ &- \tfrac{3}{4}[6\lambda(\partial_x^{-1}U_\tau) - UU_\tau - U_x(\partial_x^{-1}U_\tau) - 2U_{xx\tau} + 2a(6\lambda - U_x)]\psi_{xx} \\ &+ [\tfrac{3}{4}(6\lambda U_\tau - 2U_xU_\tau - UU_{x\tau} - U_{xx}(\partial_x^{-1}U_\tau) - 2U_{xxx\tau} - 2aU_{xx}) \\ &+ W + b]\psi_x - \tfrac{3}{2}\lambda[2U_{x\tau} + U(\partial_x^{-1}U_\tau) + 2aU]\psi,\end{aligned} \tag{5.73}$$

where λ satisfies the same equation (5.8). The DT (5.13) also holds for this system.

Iterating also on a and b we obtain corresponding to (5.30), (5.31) the hierarchy

$$U_{t_n} = \theta[U]W_n + b_nU_x, \tag{5.74}$$

$$\begin{aligned}W_n = &(J[U]\theta[U])^{n-1}J[U]U_y + \sum_{i=1}^{n} a_iG_{2n-2i+1}[U] \\ &+ \sum_{i=1}^{n-1} b_iG_{2n-2i}[U] - \sum_{i=1}^{n} q_i(J[U]\theta[U])^{n-i}x,\end{aligned} \tag{5.75}$$

and so we see that a_1 here is the function a appearing in (5.30), (5.31). Corresponding to (5.32) we obtain

$$U_{t_n} = \mathcal{R}^nU_y + \sum_{i=1}^{n} b_i\mathcal{R}^{n-i}U_x + \sum_{i=1}^{n} a_i\mathcal{R}^{n-i}\theta[U]G_1[U] - \sum_{i=1}^{n} q_i\mathcal{R}^{n-i}\theta[U]x, \tag{5.76}$$

and so we see that the extra terms correspond to the addition of lower order SK flows. Taking reductions as before we obtain corresponding generalizations of our hierarchies of PDEs in 1 + 1 dimensions and also of our hierarchies of ODEs, together with their associated linear problems. In the local case (corresponding, respectively, to (5.62), (5.63), (5.68), (5.69)) these generalized hierarchies of ODEs read

$$\mathcal{R}^nU_x + \sum_{i=1}^{n} a_i\mathcal{R}^{n-i}\theta[U]G_1[U] + \sum_{i=1}^{n} b_i\mathcal{R}^{n-i}U_x - q_n(2U + xU_x) = 0, \tag{5.77}$$

$$\mathcal{R}^{n-1}\theta[U]G_1[U]+\sum_{i=2}^{n} a_i\mathcal{R}^{n-i}\theta[U]G_1[U]+\sum_{i=1}^{n} b_i\mathcal{R}^{n-i}U_x-q_n(2U+xU_x)=0, \quad (5.78)$$

$$G_{2n}[U]+\sum_{i=1}^{n} a_iG_{2n-2i+1}[U]+\sum_{i=1}^{n} b_iG_{2n-2i}[U]-q_nx=0, \quad (5.79)$$

$$G_{2n-1}[U]+\sum_{i=2}^{n} a_iG_{2n-2i+1}[U]+\sum_{i=1}^{n} b_iG_{2n-2i}[U]-q_nx=0, \quad (5.80)$$

where in the last two equations we have taken the reduction $W_n=-b_n$ instead of $W_n=0$. (Again, our requirement that $q_n\neq 0$ allows us to take $b_n=0$ in all of the above.) The hierarchies (5.77) and (5.78) are just similarity reductions of sums of SK flows. It is straightforward to give first integrals for these two hierarchies and thus generalize the results of Refs. 21,22 for (5.62) and (5.63). For $n=1$ (5.79) is equivalent to (4.81) (see Appendix A). Linear problems for the above four hierarchies follow in the same way as described above.

Remark two: Once again, we have so far asked that not all q_i are zero. If we allow all q_i to vanish, in which case the hierarchies of ODEs (5.77), (5.78), (5.79), and (5.80) are autonomous, then we are able to use the linear problems to obtain constants of motion. For the case $n=1$ of (5.77) and (5.78) with $q_n=0$, we obtain two constants of motion; for (5.78) these are as obtained in Ref. 33 for the corresponding Hénon–Heiles system, the first being just the trivial first integral of (5.78), which corresponds to the energy. The case $n=1$ of (5.79) with $q_n=0$ is equivalent to the corresponding case of (4.81), and the first integral obtained here is equivalent (modulo the additional simple first integral) to the first integral obtained earlier for the latter. The case $n=1$ of (5.80) with $q_n=0$ is trivial; for the case $n=2$ we obtain three constants of motion.

C. Examples: $n=1$

In the case $n=1$ we can set $U=u_x$, and write the system (5.30), (5.31) locally as

$$u_{xt_1}=W_{1,xxx}+2u_xW_{1,x}+u_{xx}W_1, \quad (5.81)$$

$$W_{1,x}=u_{xxxxxy}+u_{xxxy}u_x+\tfrac{1}{2}u_{xxxx}u_y+\tfrac{3}{2}(u_{xxy}u_{xx}+u_{xxx}u_{xy})$$
$$+\tfrac{1}{4}(u_{xy}u_x^2+u_{xx}u_xu_y)+a(u_{xxxx}+\tfrac{1}{2}u_xu_{xx})-q_1, \quad (5.82)$$

where $a=a(y,t)$ and $q_1=q_1(y,t)$. This system is just our base equation (5.16), (5.17), which has Lax pair (5.18), (5.19) with $\lambda=\lambda(y,t)$ satisfying (5.20). We now consider reductions of this system to PDEs in $1+1$ dimensions, and also to ODEs.

Our first two reductions are to the PDEs,

$$U_{t_1}=W_{1,xxx}+2UW_{1,x}+U_xW_1, \quad (5.83)$$

$$W_1=G_2[U]-q_1x=U_{xxxx}+\tfrac{3}{2}UU_{xx}+\tfrac{3}{4}U_x^2+\tfrac{1}{6}U^3-q_1x \quad (5.84)$$

and

$$U_{t_1}=W_{1,xxx}+2UW_{1,x}+U_xW_1, \quad (5.85)$$

$$W_1=G_1[U]-q_1x=U_{xx}+\tfrac{1}{4}U^2-q_1x, \quad (5.86)$$

where in each case q_1 is a function of t_1 only. These have temporal part of the Lax pair given, respectively, by

$$\psi_{t_1} = -\tfrac{3}{2}[3\lambda U - UU_x - U_{xxx}]\psi_{xx} + [-27\lambda^2 + \tfrac{3}{2}(3\lambda U_x - U_x^2 - UU_{xx} - U_{xxxx}) + W_1]\psi_x - \tfrac{3}{2}\lambda[2U_{xx} + U^2]\psi \tag{5.87}$$

and

$$\psi_{t_1} = -\tfrac{3}{2}(6\lambda - U_x)\psi_{xx} + (W_1 - \tfrac{3}{2}U_{xx})\psi_x - 3\lambda U\psi, \tag{5.88}$$

where in each case λ is a function of t_1 only and $\lambda_{t_1} = -3\lambda q_1$. This last means that for $q_1 \neq 0$ the above Lax pairs are nonisospectral. When $q_1 = 0$ they become isospectral, in which case (5.83), (5.84) and (5.85), (5.86) reduce to seventh and fifth order SK, respectively.

Our second two reductions to PDEs in $1+1$ dimensions are to

$$0 = W_{1,xxx} + 2u_x W_{1,x} + u_{xx} W_1, \tag{5.89}$$

$$W_{1,x} = u_{xxxxxy} + u_{xxxy}u_x + \tfrac{1}{2}u_{xxxx}u_y + \tfrac{3}{2}(u_{xxy}u_{xx} + u_{xxx}u_{xy}) + \tfrac{1}{4}(u_{xy}u_x^2 + u_{xx}u_x u_y) + a(u_{xxxx} + \tfrac{1}{2}u_x u_{xx}) - q_1 \tag{5.90}$$

and to its further reduction

$$0 = u_{xxxxxy} + u_{xxxy}u_x + \tfrac{1}{2}u_{xxxx}u_y + \tfrac{3}{2}(u_{xxy}u_{xx} + u_{xxx}u_{xy}) + \tfrac{1}{4}(u_{xy}u_x^2 + u_{xx}u_x u_y) + a(u_{xxxx} + \tfrac{1}{2}u_x u_{xx}) - q_1, \tag{5.91}$$

where in each case a and q_1 are functions of y only. The temporal (now y) part of the Lax pair for (5.89), (5.90) is

$$27\lambda^2\psi_y = -\tfrac{3}{4}[6\lambda u_y - u_x u_{xy} - u_{xx}u_y - 2u_{xxxy} + 2a(6\lambda - u_{xx})]\psi_{xx} + [\tfrac{3}{4}(6\lambda u_{xy} - 2u_{xx}u_{xy} - u_x u_{xxy} - u_{xxx}u_y - 2u_{xxxxy} - 2au_{xxx}) + W_1] \times \psi_x - \tfrac{3}{2}\lambda[2u_{xxy} + u_x u_y + 2au_x]\psi, \tag{5.92}$$

and that for (5.91) is

$$27\lambda^2\psi_y = -\tfrac{3}{4}[6\lambda u_y - u_x u_{xy} - u_{xx}u_y - 2u_{xxxy} + 2a(6\lambda - u_{xx})]\psi_{xx} + [\tfrac{3}{4}(6\lambda u_{xy} - 2u_{xx}u_{xy} - u_x u_{xxy} - u_{xxx}u_y - 2u_{xxxxy} - 2au_{xxx})]\psi_x - \tfrac{3}{2}\lambda[2u_{xxy} + u_x u_y + 2au_x]\psi, \tag{5.93}$$

where in each case λ is a function of y only and $\lambda_y = -q_1/(9\lambda)$. For $q_1 \neq 0$ (5.89), (5.90) is a nonisospectral deformation of an inverse SK flow $\mathcal{R}U_y + a\theta[U]G_1[U] - q_1(2U + xU_x) = 0$ and appears to be new. The case $q_1 = 0$ has isospectral Lax pair. Equation (5.91) also appears to be new and again has isospectral or nonisospectral Lax pair according as to whether $q_1 = 0$ or $q_1 \neq 0$, respectively. In Appendix B we give details of the Painlevé analysis of (5.91), and also of (5.81), (5.82).

Our first two reductions to ODEs are to $\mathcal{R}U_x - q_1\theta[U]x = 0$, i.e.,

$$(U_{xxxxxx} + \tfrac{7}{2}UU_{xxxx} + \tfrac{7}{2}U_x U_{xxx} + \tfrac{7}{2}U_{xx}^2 + \tfrac{7}{2}U^2 U_{xx} + \tfrac{7}{4}UU_x^2 + \tfrac{7}{24}U^4)_x - q_1(2U + xU_x) = 0, \tag{5.94}$$

and to $\theta[U]G_1[U] - q_1\theta[U]x = 0$, i.e.,

$$(U_{xxxx} + \tfrac{5}{2}UU_{xx} + \tfrac{5}{12}U^3)_x - q_1(2U + xU_x) = 0. \tag{5.95}$$

These ODEs are similarity reductions of seventh and fifth order SK, respectively. They have linear problems with ψ_λ given, respectively, by

$$\begin{aligned}3\lambda q_1\psi_\lambda=&\tfrac{3}{2}[3\lambda U-UU_x-U_{xxx}]\psi_{xx}\\&+[27\lambda^2-\tfrac{3}{2}(3\lambda U_x-U_x^2-UU_{xx}-U_{xxxx})-W_1]\psi_x\\&+\tfrac{3}{2}\lambda[2U_{xx}+U^2]\psi,\end{aligned}\tag{5.96}$$

where W_1 is given by (5.84), and

$$3\lambda q_1\psi_\lambda=\tfrac{3}{2}(6\lambda-U_x)\psi_{xx}+(\tfrac{3}{2}U_{xx}-W_1)\psi_x+3\lambda U\psi,\tag{5.97}$$

where W_1 is given by (5.86), and where in each case $q_1\neq0$. We note that a matrix linear problem for the ODE (5.95) is known to the author of Ref. 21.

Our second two reductions to ODEs are to $G_2[U]-q_1x=0$, i.e.,

$$U_{xxxx}+\tfrac{3}{2}UU_{xx}+\tfrac{3}{4}U_x^2+\tfrac{1}{6}U^3-q_1x=0,\tag{5.98}$$

and to $G_1[U]-q_1x=0$, i.e.,

$$U_{xx}+\tfrac{1}{4}U^2-q_1x=0.\tag{5.99}$$

These ODEs have linear problems with ψ_λ given by

$$\begin{aligned}\lambda q_1\psi_\lambda=&\tfrac{1}{2}[3\lambda U-UU_x-U_{xxx}]\psi_{xx}\\&+[9\lambda^2-\tfrac{1}{2}(3\lambda U_x-\tfrac{1}{4}U_x^2+\tfrac{1}{2}UU_{xx}+\tfrac{1}{6}U^3-q_1x)]\psi_x+\tfrac{1}{2}\lambda[2U_{xx}+U^2]\psi\end{aligned}\tag{5.100}$$

and

$$\lambda q_1\psi_\lambda=\tfrac{1}{2}(6\lambda-U_x)\psi_{xx}+\tfrac{1}{2}(q_1x-\tfrac{1}{4}U^2)\psi_x+\lambda U\psi,\tag{5.101}$$

respectively, where we have used the ODEs (5.98) and (5.99) to replace the highest order derivatives of U, and where in each case $q_1\neq0$. Equation (5.99) is just P_{I}, which again we have placed at the base of a hierarchy of ODEs having third order linear problems. Equation (5.98) can be rescaled onto (4.100) (see Appendix A).

VI. CONCLUSIONS

We have shown that certain PDEs associated with nonisospectral scattering problems in $2+1$ dimensions play an important role in the construction of hierarchies of integrable PDEs and ODEs, together with their underlying linear problems. This approach makes use of a characterization of hierarchies using a single equation. We have thus obtained a wide variety of new hierarchies of integrable PDEs and ODEs. These include new PDEs in $2+1$ dimensions having nonisospectral scattering problems, new PDEs in $1+1$ dimensions having nonisospectral or isospectral scattering problems, and also new ODEs and their corresponding monodromy problems. Certain examples of these last appear to be of particular interest. In addition, we have considered the use of linear problems to derive constants of motion. The examples in this paper are based on KdV, KK, and SK type scattering problems. Further extensions and examples can be found in Ref. 39.

ACKNOWLEDGMENTS

A.P. is grateful to Professor J. M. Cervero for his invitation to visit the University of Salamanca in June 1998, where the main part of this work was completed. A.P. thanks the Ministry of

Education and Culture of Spain for a postdoctoral fellowship. The research in this paper was supported in part by the DGICYT under Contract No. PB95-0947. A.P. and P.R.G. thank Chris Cosgrove for providing them with a copy of Ref. 30.

APPENDIX A: SCALAR AND MATRIX LINEAR PROBLEMS FOR COSGROVE'S EQUATION

Here we give scalar and matrix spectral problems for the fourth order ODEs corresponding to the case $n=1$ of (4.81) and (5.79), that is, for

$$U_{xxxx}+12UU_{xx}+6U_x^2+\tfrac{32}{3}U^3+a_1(U_{xx}+4U^2)-q_1x=0, \tag{A1}$$

and

$$V_{yyyy}+\tfrac{3}{2}VV_{yy}+\tfrac{3}{4}V_y^2+\tfrac{1}{6}V^3+\tilde{a}_1(V_{yy}+\tfrac{1}{4}V^2)-\tilde{q}_1y=0, \tag{A2}$$

where in each case we have used the fact that $q_n\neq 0$ and $\tilde{q}_n\neq 0$ to set the additional additive constant equal to zero. The question of how to find the linear problem for such ODEs would appear to be nontrivial since they apparently do not arise, for example, as the (integrated) stationary flow of an integrable evolution equation. The two PDEs which perhaps most naturally give (A1) and (A2) as reductions, and from which we can most easily obtain the corresponding linear problems, are (4.93) and (5.91), respectively.

The ODEs (A1) and (A2) are related by the transformation

$$U(x)=(V(y)+\tilde{a}_1)/8,\quad x=y-(\tilde{a}_1^3/(12\tilde{q}_1)),\quad a_1=-\tilde{a}_1/2,\quad q_1=\tilde{q}_1/8, \tag{A3}$$

and in turn to the equation found by Cosgrove using Painlevé classification[30] (see also Ref. 37),

$$W_{zzzz}-18WW_{zz}-9W_z^2+24W^3-\alpha W^2-(\alpha^2/9)W-\beta z=0, \tag{A4}$$

by the transformation

$$U(x)=-(3W(z)/2)-(\alpha/24),\quad x=z-(7\alpha^3)/(3888\beta),\quad a_1=\alpha/2,\quad q_1=-3\beta/2. \tag{A5}$$

Here we assume $\beta\neq 0$ and so again set the additive constant to zero. Note that in the gauge chosen by Cosgrove, which removes the W_{zz} term, the coefficient α appears nonlinearly, whereas in our corresponding generalized hierarchies such constants appear linearly (since these consist of a linear combination of such equations).

Our scalar linear problem for (A1) is

$$\psi_{xxx}=-2U\psi_x-(U_x-\lambda)\psi, \tag{A6}$$

$$q_1\psi_\lambda=(6U+3a_1)\psi_{xx}+(9\lambda-3U_x)\psi_x+(U_{xx}+8U^2+4a_1U)\psi, \tag{A7}$$

and that for (A2) is

$$\psi_{yyy}=-\tfrac{1}{2}V\psi_y+\lambda\psi, \tag{A8}$$

$$\begin{aligned}\lambda\tilde{q}_1\psi_\lambda=&\tfrac{1}{2}[3\lambda V-VV_y-V_{yyy}+\tilde{\alpha}_1(6\lambda-V_y)]\psi_{yy}\\&+[9\lambda^2-\tfrac{1}{2}(3\lambda V_y-\tfrac{1}{4}V_y^2+\tfrac{1}{2}VV_{yy}+\tfrac{1}{6}V^3+\tfrac{1}{4}\tilde{a}_1V^2-\tilde{q}_1y)]\psi_y\\&+\tfrac{1}{2}\lambda[2V_{yy}+V^2+2\tilde{a}_1V]\psi.\end{aligned} \tag{A9}$$

We recall that there is a well-known relationship between the two third order operators (A.6) and (A.8).[34]

The compatibility condition of the linear problem (A6), (A7) gives the derivative of (A1). That for (A8), (A9) gives (A2). Corresponding to each of these scalar linear problems we have matrix linear problems with compatibility conditions (A1) and (A2), respectively. The simplest of these is that corresponding to (A6), (A7), i.e.,

$$\mathbf{\Psi}_x=\begin{pmatrix} 0 & 1 & 0\\ -U & 0 & 1\\ -\lambda & -U & 0\end{pmatrix}\mathbf{\Psi}, \tag{A10}$$

$$q_1\mathbf{\Psi}_\lambda=\begin{pmatrix} -(U_{xx}+2U^2+a_1U) & 3(U_x+3\lambda) & -3(2U+a_1)\\ \begin{matrix}-U_{xxx}-7UU_x\\ -3\lambda(U-a_1)-a_1U_x\end{matrix} & 2(U_{xx}+2U^2+a_1U) & 3(3\lambda-U_x)\\ \begin{matrix}2UU_{xx}-U_x^2+\frac{14}{3}U^3\\ +a_1U^2-9\lambda^2-q_1x\end{matrix} & \begin{matrix}U_{xxx}+7UU_x+a_1U_x\\ +3\lambda(a_1-U)\end{matrix} & \begin{matrix}-U_{xx}-2U^2\\ -a_1U\end{matrix}\end{pmatrix}\mathbf{\Psi}. \tag{A11}$$

Corresponding to (A8), (A9) we have a matrix linear problem with

$$\mathbf{\Psi}_y=\begin{pmatrix} 0 & 1 & 0\\ -V/4 & 0 & 1\\ -\lambda-(V_y/4) & -V/4 & 0\end{pmatrix}\mathbf{\Psi}. \tag{A12}$$

The second half of this matrix linear problem is too long to reproduce here.

APPENDIX B: PAINLEVÉ ANALYSIS

In this Appendix we give brief details of the WTC Painlevé test[40] for some of the more fundamental equations presented in this paper. Thus for example for the $2+1$-dimensional system (4.83), (4.84) we seek expansions about a noncharacteristic movable singular manifold $\varphi=0$, assuming a leading order behavior $u\sim u_0\varphi^\alpha$, $W\sim W_0\varphi^\beta$. Without loss of generality we use Kruskal's "reduced ansatz" and take $\varphi=x+\xi(y,t)$ and all coefficients in the expansion to be functions of y and t only. For the system (4.83), (4.84) we thus obtain the families

$$\alpha=-1\quad u_0=3/2\quad \beta=-1\quad W_0\ \text{arbitrary}\quad \{-1,0,1,2,3,4,4,6,8\} \tag{B1}$$

$$\alpha=-1\quad u_0=15/2\quad \beta=-3\quad W_0\ \text{arbitrary}\quad \{-5,-1,0,1,4,6,8,8,12\} \tag{B2}$$

$$\alpha=-1\quad u_0=24\quad \beta=-6\quad W_0=-95040\xi_y\quad \{-11,-5,-1,1,6,8,12,14,18\}, \tag{B3}$$

where for each leading order behavior the set of numbers in braces is the set of indices or "resonances" where arbitrary data is to be introduced into the expansion. The PDE (4.83), (4.84) is then found to pass the WTC Painlevé test.

The system (4.83), (4.84) has as a subequation the $1+1$-dimensional PDE (4.93). This has the families

$$\alpha=-1\quad u_0=3/2\quad \{-1,1,3,4,6,8\}, \tag{B4}$$

$$\alpha=-1\quad u_0=15/2\quad \{-5-1,1,6,8,12\}, \tag{B5}$$

whose leading orders and indices are easily seen to be subsets of those of the above families of (4.83), (4.84) (and so to be consistent under an appropriate reduction). The PDE (4.93) is again found to pass the WTC Painlevé test.

The system (5.81), (5.82) also passes the WTC Painlevé test, with families

$$\alpha=-1 \quad u_0=12 \quad \beta=-4 \quad W_0 \text{ arbitrary} \quad \{-1,0,1,3,4,5,6,8,10\}, \tag{B6}$$

$$\alpha=-1 \quad u_0=60 \quad \beta=1 \quad W_0 \text{ arbitrary} \quad \{-11,-5,-1,0,1,6,8,11,12\}, \tag{B7}$$

$$\alpha=-1 \quad u_0=24 \quad \beta=-6 \quad W_0=1728\xi_y \quad \{-2,-1,1,3,4,6,8,9,14\}. \tag{B8}$$

Again this system has a $1+1$-dimensional subequation, namely, (5.91). This has the families

$$\alpha=-1 \quad u_0=12 \quad \{-1,1,3,4,6,8\}, \tag{B9}$$

$$\alpha=-1 \quad u_0=60 \quad \{-5,-1,1,6,8,12\}, \tag{B10}$$

whose leading orders and indices are clearly subsets of those of the families of (5.81), (5.82). Equation (5.91) also passes the WTC Painlevé test.

The families of (4.83), (4.84) and (5.81), (5.82) correspond to those of seventh order KK and seventh order SK, respectively. Here we have two additional indices (1 and 6) which correspond to having taken a potential and an extra derivative, and the remaining seven indices are then identified with those of the corresponding families of seventh order KK and seventh order SK if we also take account of any shift on the leading order behavior of W (any such shift affects three of the remaining indices). We also note that the actual expansion for W for the family (B7) is $W\sim\xi_t+W_0\varphi$, where the extra lower order term ξ_t restores the dominance in (5.81) (see Ref. 41).

The families of the ODE (A1) (equivalently (A2) or (A4)) correspond in exactly the same way to those of the PDEs (4.93) or (5.91); again we have the two additional indices 1 and 6. Details of the Painlevé analysis of this ODE can be found in Ref. 30 (the dominant terms and corresponding families were first given in Ref. 37).

APPENDIX C: THE KAUP–KUPERSHMIDT CASE: $n=2$

Here we consider reductions of our $2+1$ system (4.29), (4.30) in the special case $n=2$. As described in Sec. IV, we have reductions to PDEs in $1+1$ dimensions, with isospectral and nonisospectral Lax pairs. However here, for reasons of brevity, we restrict ourselves to giving explicitly only the further reductions of these PDEs down to ODEs.

Our first two reductions are to the ODEs $\mathcal{R}^2U_x-q_1\mathcal{R}\theta[U]x-q_2\theta[U]x=0$ and $\mathcal{R}\theta[U]H_1[U]-q_1\mathcal{R}\theta[U]x-q_2\theta[U]x=0$, which we write (here $U_{(4x)}$ means U_{xxxx} and $U^2_{(4x)}$ means $(U_{xxxx})^2$, etc.) as

$$\begin{aligned}
&(U_{(12x)}+26UU_{(10x)}+143U_xU_{(9x)}+494U_{xx}U_{(8x)}+260U^2U_{(8x)}\\
&\quad+1144U_{xxx}U_{(7x)}+2288UU_xU_{(7x)}+1872U_{(4x)}U_{(6x)}+6292UU_{xx}U_{(6x)}\\
&\quad+4459U_x^2U_{(6x)}+\tfrac{3848}{3}U^3U_{(6x)}+\tfrac{2197}{2}U^2_{(5x)}+11102UU_{xxx}U_{(5x)}\\
&\quad+20787U_xU_{xx}U_{(5x)}+12636U^2U_xU_{(5x)}+6374UU^2_{(4x)}+31200U_xU_{xxx}U_{(4x)}\\
&\quad+20618U_{xx}^2U_{(4x)}+26312U^2U_{xx}U_{(4x)}+36634UU_x^2U_{(4x)}+\tfrac{10192}{3}U^4U_{(4x)}\\
&\quad+25480U_{xx}U_{xxx}^2+16146U^2U_{xxx}^2+122694UU_xU_{xx}U_{xxx}+30225U_x^3U_{xxx}\\
&\quad+\tfrac{88816}{3}U^3U_xU_{xxx}+\tfrac{83876}{3}UU_{xx}^3+\tfrac{120315}{2}U_x^2U_{xx}^2+\tfrac{64792}{3}U^3U_{xx}^2\\
&\quad+85176U^2U_x^2U_{xx}+\tfrac{14560}{3}U^5U_{xx}+21021UU_x^4+\tfrac{40040}{3}U^4U_x^2\\
&\quad+\tfrac{4160}{9}U^7)_x-q_1(g_{xxx}+2Ug_x+U_xg)-q_2(2U+xU_x)=0
\end{aligned} \tag{C1}$$

and

$$
\begin{aligned}
&(U_{(10x)}+22U_{(8x)}+99U_xU_{(7x)}+275U_{xx}U_{(6x)}+176U^2U_{(6x)}\\
&\quad+495U_{xxx}U_{(5x)}+1188UU_xU_{(5x)}+\tfrac{605}{2}U_{(4x)}^2+2464UU_{xx}U_{(4x)}\\
&\quad+1716U_x^2U_{(4x)}+\tfrac{1936}{3}U^3U_{(4x)}+1518UU_{xxx}^2+5566U_xU_{xx}U_{xxx}\\
&\quad+4312U^2U_xU_{xxx}+\tfrac{3674}{3}U_{xx}^3+3124U^2U_{xx}^2+8360UU_x^2U_{xx}\\
&\quad+\tfrac{3520}{3}U^4U_{xx}+1089U_x^4+2640U^3U_x^2+\tfrac{1408}{9}U^6)_x\\
&\quad-q_1(g_{xxx}+2Ug_x+U_xg)-q_2(2U+xU_x)=0,
\end{aligned}
\tag{C2}
$$

respectively, where in order to write these ODEs and their corresponding linear problems locally we have defined

$$
g=K[U]f_x \quad \text{and} \quad f_x=\theta[U]x=(2U+xU_x). \tag{C3}
$$

Equation (C1) has linear problem with ψ_λ given by

$$
-3\lambda(q_2-27\lambda^2q_1)\psi_\lambda=A_2\psi_{xx}+[(-27\lambda^2)^2+B_2]\psi_x+\tfrac{1}{3}[4UA_2-A_{2,xx}-3B_{2,x}]\psi, \tag{C4}
$$

where

$$
A_2=-18\lambda\partial_x^{-1}(Q_1-27\lambda^2Q_0), \tag{C5}
$$

$$
B_2=9\lambda(Q_1-27\lambda^2Q_0)+(W_2-27\lambda^2W_1), \tag{C6}
$$

$\partial_x^{-1}Q_0=U,$ W_1 is given by (4.86),

$$
\begin{aligned}
\partial_x^{-1}Q_1&=\partial_x^{-1}\theta[U]W_1\\
&=U_{xxxxxx}+14UU_{xxxx}+35U_xU_{xxx}+\tfrac{49}{2}U_{xx}^2\\
&\quad+56U^2U_{xx}+70UU_x^2+\tfrac{56}{3}U^4-q_1f
\end{aligned}
\tag{C7}
$$

and

$$
\begin{aligned}
W_2&=H_4[U]-(q_1K[U]\theta[U]x+q_2x)\\
&=U_{(10x)}+24UU_{(8x)}+96U_xU_{(7x)}+277U_{xx}U_{(6x)}+212U^2U_{(6x)}\\
&\quad+495U_{xxx}U_{(5x)}+1272UU_xU_{(5x)}+302U_{(4x)}^2+2746UU_{xx}U_{(4x)}\\
&\quad+1563U_x^2U_{(4x)}+\tfrac{2576}{3}U^3U_{(4x)}+1686UU_{xxx}^2+5384U_xU_{xx}U_{xxx}\\
&\quad+5152U^2U_xU_{xxx}+\tfrac{3821}{3}U_{xx}^3+3864U^2U_{xx}^2+8596UU_x^2U_{xx}\\
&\quad+1680U^4U_{xx}+861U_x^4+3360U^3U_x^2+\tfrac{2240}{9}U^6-(q_1g+q_2x).
\end{aligned}
\tag{C8}
$$

Equation (C2) has a linear problem with ψ_λ given by

$$
-3\lambda(q_2-27\lambda^2q_1)\psi_\lambda=A_2\psi_{xx}+B_2\psi_x+\tfrac{1}{3}[4UA_2-A_{2,xx}-3B_{2,x}]\psi, \tag{C9}
$$

where

$$
A_2=-9\lambda(-27\lambda^2)-18\lambda\partial_x^{-1}Q_1, \tag{C10}
$$

$$
B_2=9\lambda Q_1+(W_2-27\lambda^2W_1), \tag{C11}
$$

W_1 is given by (4.88),

$$\partial_x^{-1}Q_1=\partial_x^{-1}\theta[U]W_1=U_{xxxx}+10UU_{xx}+\tfrac{15}{2}U_x^2+\tfrac{20}{3}U^3-q_1f, \tag{C12}$$

and

$$\begin{aligned}W_2&=H_3[U]-(q_1K[U]\theta[U]x+q_2x)\\&=U_{(8x)}+20UU_{(6x)}+60U_xU_{(5x)}+134U_{xx}U_{(4x)}+136U^2U_{(4x)}\\&\quad+84U_{xxx}^2+544UU_xU_{xxx}+408UU_{xx}^2+396U_x^2U_{xx}+\tfrac{1120}{3}U^3U_{xx}\\&\quad+560U^2U_x^2+\tfrac{256}{3}U^5-(q_1g+q_2x).\end{aligned} \tag{C13}$$

In each of the above examples q_1 and q_2 are not both zero. The nonlocal terms included in the above ODEs are new. When $q_1=0$, Eqs. (C1) and (C2) are similarity reductions of thirteenth and eleventh order KK, respectively.

Our second two reductions to ODEs are to $H_4[U]-q_1K[U]\theta[U]x-q_2x=0$ and $H_3[U]-q_1K[U]\theta[U]x-q_2x=0$, which we can write as

$$\begin{aligned}&U_{(10x)}+24UU_{(8x)}+96U_xU_{(7x)}+277U_{xx}U_{(6x)}+212U^2U_{(6x)}\\&\quad+495U_{xxx}U_{(5x)}+1272UU_xU_{(5x)}+302U_{(4x)}^2+2746UU_{xx}U_{(4x)}\\&\quad+1563U_x^2U_{(4x)}+\tfrac{2576}{3}U^3U_{(4x)}+1686UU_{xxx}^2+5384U_xU_{xx}U_{xxx}\\&\quad+5152U^2U_xU_{xxx}+\tfrac{3821}{3}U_{xx}^3+3864U^2U_{xx}^2+8596UU_x^2U_{xx}\\&\quad+1680U^4U_{xx}+861U_x^4+3360U^3U_x^2+\tfrac{2240}{9}U^6-(q_1g+q_2x)=0\end{aligned} \tag{C14}$$

and

$$\begin{aligned}&U_{(8x)}+20UU_{(6x)}+60U_xU_{(5x)}+134U_{xx}U_{(4x)}+136U^2U_{(4x)}\\&\quad+84U_{xxx}^2+544UU_xU_{xxx}+408UU_{xx}^2+396U_x^2U_{xx}+\tfrac{1120}{3}U^3U_{xx}\\&\quad+560U^2U_x^2+\tfrac{256}{3}U^5-(q_1g+q_2x)=0,\end{aligned} \tag{C15}$$

where f and g are as defined in (C3). Both of these ODEs are new.

Equation (C14) has a linear problem with ψ_λ given by (C4) with $W_2=0$, i.e.,

$$(q_2-27\lambda^2q_1)\psi_\lambda=A_2\psi_{xx}+[9\lambda(-27\lambda^2)+B_2]\psi_x+\tfrac{1}{3}[4UA_2-A_{2,xx}-3B_{2,x}]\psi, \tag{C16}$$

where now

$$A_2=6\partial_x^{-1}(Q_1-27\lambda^2Q_0), \tag{C17}$$

$$B_2=-3(Q_1-27\lambda^2Q_0)+9\lambda W_1, \tag{C18}$$

$\partial_x^{-1}Q_0=U$, and W_1 and $\partial_x^{-1}Q_1$ are given by (4.86) and (C7), respectively.

Equation (C15) has a linear problem with ψ_λ given by (C9) with $W_2=0$, i.e.,

$$(q_2-27\lambda^2q_1)\psi_\lambda=A_2\psi_{xx}+B_2\psi_x+\tfrac{1}{3}[4UA_2-A_{2,xx}-3B_{2,x}]\psi, \tag{C19}$$

where now

$$A_2=3(-27\lambda^2)+6\partial_x^{-1}Q_1, \tag{C20}$$

$$B_2=-3Q_1+9\lambda W_1, \tag{C21}$$

and W_1 and $\partial_x^{-1}Q_1$ are given by (4.88) and (C12), respectively.

Once again in each of the above examples q_1 and q_2 are not both zero.

APPENDIX D: THE SAWADA–KOTERA CASE: $n=2$

Here we consider reductions of our $2+1$ system (5.30), (5.31) in the special case $n=2$. As described in Sec. V, we have reductions to PDEs in $1+1$ dimensions, with isospectral and nonisospectral Lax pairs. However as in Appendix C, for reasons of brevity, we restrict ourselves here to giving explicitly only the further reductions of these PDEs down to ODEs.

Our first two reductions are to the ODEs $\mathcal{R}^2U_x-q_1\mathcal{R}\theta[U]x-q_2\theta[U]x=0$ and $\mathcal{R}\theta[U]G_1[U]-q_1\mathcal{R}\theta[U]x-q_2\theta[U]x=0$, which we write as

$$\begin{aligned}(U_{(12x)}&+\tfrac{13}{2}UU_{(10x)}+26U_xU_{(9x)}+\tfrac{169}{2}U_{xx}U_{(8x)}+\tfrac{65}{4}U^2U_{(8x)}\\&+169U_{xxx}U_{(7x)}+104UU_xU_{(7x)}+\tfrac{507}{2}U_{(4x)}U_{(6x)}+\tfrac{1105}{4}UU_{xx}U_{(6x)}\\&+\tfrac{299}{2}U_x^2U_{(6x)}+\tfrac{481}{24}U^3U_{(6x)}+143U_{(5x)}^2+\tfrac{871}{2}UU_{xxx}U_{(5x)}\\&+\tfrac{1443}{2}U_xU_{xx}U_{(5x)}+\tfrac{585}{4}U^2U_xU_{(5x)}+260UU_{(4x)}^2+\tfrac{2067}{2}U_xU_{xxx}U_{(4x)}\\&+\tfrac{3029}{4}U_{xx}^2U_{(4x)}+\tfrac{1235}{4}U^2U_{xx}U_{(4x)}+325UU_x^2U_{(4x)}+\tfrac{637}{48}U^4U_{(4x)}\\&+910U_{xx}U_{xxx}^2+\tfrac{351}{2}U^2U_{xxx}^2+\tfrac{4563}{4}UU_xU_{xx}U_{xxx}+195U_x^3U_{xxx}\\&+\tfrac{2093}{24}U^3U_xU_{xxx}+\tfrac{6799}{24}UU_{xx}^3+\tfrac{1755}{4}U_x^2U_{xx}^2+\tfrac{3367}{48}U^3U_{xx}^2\\&+\tfrac{819}{4}U^2U_x^2U_{xx}+\tfrac{455}{96}U^5U_{xx}+\tfrac{273}{8}UU_x^4+\tfrac{455}{48}U^4U_x^2\\&+\tfrac{65}{576}U^7)_x-q_1(g_{xxx}+2Ug_x+U_xg)-q_2(2U+xU_x)=0\end{aligned}\tag{D1}$$

and

$$\begin{aligned}(U_{(10x)}&+\tfrac{11}{2}UU_{(8x)}+\tfrac{33}{2}U_xU_{(7x)}+44U_{xx}U_{(6x)}+11U^2U_{(6x)}\\&+66U_{xxx}U_{(5x)}+\tfrac{99}{2}UU_xU_{(5x)}+\tfrac{77}{2}U_{(4x)}^2+\tfrac{209}{2}UU_{xx}U_{(4x)}\\&+\tfrac{99}{2}U_x^2U_{(4x)}+\tfrac{121}{12}U^3U_{(4x)}+\tfrac{231}{4}UU_{xxx}^2+187U_xU_{xx}U_{xxx}\\&+\tfrac{187}{4}U^2U_xU_{xxx}+\tfrac{143}{3}U_{xx}^3+\tfrac{77}{2}U^2U_{xx}^2+\tfrac{275}{4}UU_x^2U_{xx}\\&+\tfrac{55}{12}U^4U_{xx}+\tfrac{33}{8}U_x^4+\tfrac{55}{8}U^3U_x^2+\tfrac{11}{72}U^6)_x\\&-q_1(q_{xxx}+2Ug_x+U_xg)-q_2(2U+xU_x)=0,\end{aligned}\tag{D2}$$

respectively, where in order to write these ODEs and their corresponding linear problems locally we have defined

$$g=J[U]f_x\quad\text{and}\quad f_x=\theta[U]x=(2U+xU_x).\tag{D3}$$

Equation (D1) has a linear problem with ψ_λ given by

$$-3\lambda(q_2-27\lambda^2q_1)\psi_\lambda=A_2\psi_{xx}+[(-27\lambda^2)^2+B_2]\psi_x+C_2\psi,\tag{D4}$$

where

$$\begin{aligned}A_2=&-\tfrac{9}{2}\lambda\partial_x^{-1}(Q_1-27\lambda^2Q_0)+\tfrac{3}{4}[(UQ_1+U_x\partial_x^{-1}Q_1+2Q_{1,xx})\\&-27\lambda^2(UQ_0+U_x\partial_x^{-1}Q_0+2Q_{0,xx})],\end{aligned}\tag{D5}$$

$$B_2=-A_{2,x}+(W_2-27\lambda^2W_1),\tag{D6}$$

$$C_2=-\tfrac{3}{2}\lambda[(U\partial_x^{-1}Q_1+2Q_{1,x})-27\lambda^2(U\partial_x^{-1}Q_0+2Q_{0,x})], \tag{D7}$$

$\partial_x^{-1}Q_0=U, W_1$ is given by (5.84),

$$\begin{aligned}\partial_x^{-1}Q_1&=\partial_x^{-1}\theta[U]W_1\\&=U_{xxxxxx}+\tfrac{7}{2}UU_{xxxx}+\tfrac{7}{2}U_xU_{xxx}\\&\quad+\tfrac{7}{2}U_{xx}^2+\tfrac{7}{2}U^2U_{xx}+\tfrac{7}{4}UU_x^2+\tfrac{7}{24}U^4-q_1f\end{aligned} \tag{D8}$$

and

$$\begin{aligned}W_2&=G_4[U]-(q_1J[U]\theta[U]x+q_2x)\\&=U_{(10x)}+\tfrac{9}{2}UU_{(8x)}+18U_xU_{(7x)}+43U_{xx}U_{(6x)}+\tfrac{29}{4}U^2U_{(6x)}\\&\quad+66U_{xxx}U_{(5x)}+\tfrac{87}{2}UU_xU_{(5x)}+\tfrac{155}{4}U_{(4x)}^2+\tfrac{337}{4}UU_{xx}U_{(4x)}\\&\quad+\tfrac{123}{2}U_x^2U_{(4x)}+\tfrac{133}{24}U^3U_{(4x)}+48UU_{xxx}^2+197U_xU_{xx}U_{xxx}\\&\quad+\tfrac{133}{4}U^2U_xU_{xxx}+\tfrac{271}{6}U_{xx}^3+\tfrac{399}{16}U^2U_{xx}^2+\tfrac{259}{4}UU_x^2U_{xx}+\tfrac{35}{16}U^4U_{xx}+\tfrac{63}{8}U_x^4\\&\quad+\tfrac{35}{8}U^3U_x^2+\tfrac{35}{576}U^6-(q_1g+q_2x).\end{aligned} \tag{D9}$$

Equation (D2) has a linear problem with ψ_λ given by

$$-3\lambda(q_2-27\lambda^2q_1)\psi_\lambda=A_2\psi_{xx}+B_2\psi_x+C_2\psi, \tag{D10}$$

where

$$A_2=-\tfrac{3}{2}(-27\lambda^2)(6\lambda-U_x)-\tfrac{9}{2}\lambda\partial_x^{-1}Q_1+\tfrac{3}{4}(UQ_1+U_x\partial_x^{-1}Q_1+2Q_{1,xx}), \tag{D11}$$

$$B_2=-A_{2,x}+(W_2-27\lambda^2W_1), \tag{D12}$$

$$C_2=-3\lambda(-27\lambda^2)U-\tfrac{3}{2}\lambda[U\partial_x^{-1}Q_1+2Q_{1,x}], \tag{D13}$$

W_1 is given by (5.86),

$$\partial_x^{-1}Q_1=\partial_x^{-1}\theta[U]W_1=U_{xxxx}+\tfrac{5}{2}UU_{xx}+\tfrac{5}{12}U^3-q_1f \tag{D14}$$

and

$$\begin{aligned}W_2&=G_3[U]-(q_1J[U]\theta[U]x+q_2x)\\&=U_{(8x)}+\tfrac{7}{2}UU_{(6x)}+\tfrac{21}{2}U_xU_{(5x)}+\tfrac{37}{2}U_{xx}U_{(4x)}+4U^2U_{(4x)}\\&\quad+\tfrac{39}{4}U_{xxx}^2+16UU_xU_{xxx}+12UU_{xx}^2+\tfrac{33}{2}U_x^2U_{xx}+\tfrac{25}{12}U^3U_{xx}\\&\quad+\tfrac{25}{8}U^2U_x^2+\tfrac{1}{12}U^5-(q_1g+q_2x).\end{aligned} \tag{D15}$$

In each of the above examples q_1 and q_2 are not both zero. The nonlocal terms included in the above ODEs are new. In the case $q_1=0$, Eqs. (D1) and (D2) are similarity reductions of thirteenth and eleventh order SK, respectively.

Our second two reductions to ODEs are to $G_4[U]-q_1J[U]\theta[U]x-q_2x=0$ and $G_3[U]-q_1J[U]\theta[U]x-q_2x=0$, which we can write as

$$
\begin{aligned}
&U_{(10x)}+\tfrac{9}{2}UU_{(8x)}+18U_xU_{(7x)}+43U_{xx}U_{(6x)}+\tfrac{29}{4}U^2U_{(6x)}\\
&\quad+66U_{xxx}U_{(5x)}+\tfrac{87}{2}UU_xU_{(5x)}+\tfrac{155}{4}U_{(4x)}^2+\tfrac{337}{4}UU_{xx}U_{(4x)}\\
&\quad+\tfrac{123}{2}U_x^2U_{(4x)}+\tfrac{133}{24}U^3U_{(4x)}+48UU_{xxx}^2+197U_xU_{xx}U_{xxx}\\
&\quad+\tfrac{133}{4}U^2U_xU_{xxx}+\tfrac{271}{6}U_{xx}^3+\tfrac{399}{16}U^2U_{xx}^2+\tfrac{259}{4}UU_x^2U_{xx}\\
&\quad+\tfrac{35}{16}U^4U_{xx}+\tfrac{63}{8}U_x^4+\tfrac{35}{8}U^3U_x^2+\tfrac{35}{576}U^6-(q_1g+q_2x)=0
\end{aligned}
\tag{D16}
$$

and

$$
\begin{aligned}
&U_{(8x)}+\tfrac{7}{2}UU_{(6x)}+\tfrac{21}{2}U_xU_{(5x)}+\tfrac{37}{2}U_{xx}U_{(4x)}+4U^2U_{(4x)}\\
&\quad+\tfrac{39}{4}U_{xxx}^2+16UU_xU_{xxx}+12UU_{xx}^2+\tfrac{33}{2}U_x^2U_{xx}+\tfrac{25}{12}U^3U_{xx}\\
&\quad+\tfrac{25}{8}U^2U_x^2+\tfrac{1}{12}U^5-(q_1g+q_2x)=0,
\end{aligned}
\tag{D17}
$$

where f and g are as defined in (D3). Both of these ODEs are new.

Equation (D16) has a linear problem with ψ_λ given by (D4) with $W_2=0$, i.e., A_2 is as given by (D5), $B_2=-A_{2,x}-27\lambda^2W_1$, and C_2 is as given by (D7). In these expressions for A_2, B_2, and C_2, $\partial_x^{-1}Q_0=U$, W_1 and $\partial_x^{-1}Q_1$ are given by (5.84) and (D8), respectively, and the tenth order derivative of U appearing in B_2 can be replaced using Eq. (D16).

Equation (D17) has a linear problem with ψ_λ given by (D10) with $W_2=0$, i.e., A_2 is as given by (D11), $B_2=-A_{2,x}-27\lambda^2W_1$, and C_2 is as given by (D13). In these expressions for A_2, B_2, and C_2, W_1 and $\partial_x^{-1}Q_1$ are given by (5.86) and (D14), respectively, and the eighth order derivative of U appearing in B_2 can be replaced using Eq. (D17).

Once again in each of the above examples q_1 and q_2 are not both zero.

[1] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, ''Method for solving the Korteweg–de Vries equation,'' Phys. Rev. Lett. **19**, 1095–1097 (1967).
[2] V. E. Zakharov and A. B. Shabat, ''Exact theory of the two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media,'' Zh. Eksp. Teor. Fiz. **61**, 118–134 (1971); Sov. Phys. JETP **34**, 62–69 (1972).
[3] V. E. Zakharov and A. B. Shabat, ''A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem, I,'' Funk. Anal. Pril. **8**, 43–53 (1974); Funct. Anal. Appl. **8**, 266–235 (1974).
[4] V. S. Dryuma, ''Analytic solution of the two-dimensional Korteweg–de Vries equation,'' Zh. Eksp. Teor. Fiz. Pis'ma Red. **19**, 753–755 (1974); Sov. Phys. JETP Lett. **19**, 387–388 (1974).
[5] H. Flaschka and A. C. Newell, ''Monodromy and spectrum preserving deformations, I,'' Commun. Math. Phys. **76**, 65–116 (1980).
[6] F. Calogero, ''A method to generate solvable nonlinear evolution equations,'' Lett. Nuovo Cimento **14**, 443–447 (1975).
[7] F. Calogero and A. Degasperis, *Spectral Transform and Solitons I* (North–Holland, Amsterdam, 1982).
[8] O. I. Bogoyavlenskii, ''Overturning solitons in two-dimensional integrable equations,'' Usp. Mat. Nauk **45**, 17–77 (1990); Russ. Math. Surveys **45**, 1–86 (1990).
[9] V. E. Zakharov, ''The inverse scattering method,'' in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer, Berlin, 1980).
[10] A. Pickering, ''The singular manifold method revisited,'' J. Math. Phys. **37**, 1894–1927 (1996).
[11] P. A. Clarkson, P. R. Gordoa, and A. Pickering, ''Multicomponent equations associated to nonisospectral scattering problems,'' Inverse Probl. **13**, 1463–1476 (1997).
[12] I. A. B. Strachan, ''Wave solutions of a 2+1-dimensional generalization of the nonlinear Schrödinger equation,'' Inverse Probl. **8**, L21–L27 (1992).
[13] Yi-shen Li and You-jin Zhang, ''Symmetries of a 2+1-dimensional breaking soliton equation,'' J. Phys. A **26**, 7487–7494 (1993).
[14] R. L. Sachs, ''On the integrable variant of the Boussinesq system: Painlevé property, rational solutions, a related many-body system, and equivalence with the AKNS hierarchy,'' Physica D **30**, 1–27 (1988).
[15] D. Levi, O. Ragnisco, and M. A. Rodriguez, ''On nonisospectral flows, Painlevé equations, and symmetries of differential and difference equations,'' Teor. Mat. Fiz. **93**, 473–480 (1992); Theor. Math. Phys. **93**, 1409–1414 (1993).
[16] S. P. Burtsev, V. E. Zakharov, and A. V. Mikhailov, ''Inverse scattering method with variable spectral parameter,'' Teor. Mat. Fiz. **70**, 323–341 (1987).
[17] D. J. Kaup, ''On the inverse scattering problem for cubic eigenvalue problems of the class $\psi_{xxx}+6Q\psi_x+6R\psi=\lambda\psi$,'' Stud. Appl. Math. **62**, 189–216 (1980).
[18] A. P. Fordy and J. Gibbons, ''Some remarkable nonlinear transformations,'' Phys. Lett. A **75**, 325–325 (1980).

[19] K. Sawada and T. Kotera, ''A method for finding *N*-soliton solutions of the KdV equation and KdV-type equation,'' Prog. Theor. Phys. **51**, 1355–1367 (1974).
[20] P. J. Caudrey, R. K. Dodd, and J. D. Gibbon, ''A new hierarchy of Korteweg–de Vries equations,'' Proc. R. Soc. London, Ser. A **351**, 407–422 (1976).
[21] A. N. W. Hone, ''Nonautonomous Hénon–Heiles systems,'' Physica D **118**, 1–16 (1998).
[22] N. Joshi and A. Pickering, ''Bäcklund transformations for similarity reductions of the modified Sawada-Kotera / Kaup-Kupershmidt hierarchy'' (submitted).
[23] P. J. Olver, ''Evolution equations possessing infinitely many symmetries,'' J. Math. Phys. **18**, 1212–1215 (1977).
[24] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, ''The inverse scattering transform: Fourier analysis for nonlinear problems,'' Stud. Appl. Math. **53**, 249–315 (1974).
[25] Yi-shen Li, ''Remarks on the breaking soliton equations,'' preprint, 1993.
[26] D. Levi and O. Ragnisco, ''Nonisospectral deformations and Darboux transformations for the third order spectral problem,'' Inverse Probl. **4**, 815–828 (1988).
[27] H. Airault, ''Rational solutions of Painlevé equations,'' Stud. Appl. Math. **61**, 31–53 (1979).
[28] N. A. Kudryashov, ''The first and second Painlevé equations of higher order and some relations between them,'' Phys. Lett. A **224**, 353–360 (1997).
[29] P. A. Clarkson, N. Joshi, and A. Pickering, ''Bäcklund transformations for the second Painlevé hierarchy: A modified truncation approach,'' Inverse Probl. **15**, 175–187 (1999).
[30] C. M. Cosgrove, ''Higher order Painlevé equations in the polynomial class I. Bureau symbol P2,'' Stud. Appl. Math. (to be published); http://www.maths.usyd.edu.au:8000/res/Nonlinear/Cos/1998-22.html
[31] N. A. Kudryashov and M. B. Soukharev, ''Uniformization and transcendence of solutions for the first and second Painlevé hierarchies,'' Phys. Lett. A **237**, 206–216 (1998).
[32] R. Littlewood, ''Hyperelliptic asymptotics of Painlevé-type equations,'' preprint, 1998. (Contact address: School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia).
[33] A. P. Fordy, ''The Hénon–Heiles system revisited,'' Physica D **52**, 204–210 (1991).
[34] A. P. Fordy and J. Gibbons, ''Factorization of operators I. Miura transformations,'' J. Math. Phys. **21**, 2508–2510 (1980).
[35] B. Fuchssteiner and W. Oevel, ''The bi-Hamiltonian structure of some nonlinear fifth- and seventh-order differential equations and recursion formulas for their symmetries and conserved covariants,'' J. Math. Phys. **23**, 358–363 (1982).
[36] J. Weiss, ''On classes of integrable systems and the Painlevé property,'' J. Math. Phys. **25**, 13–24 (1984).
[37] H. Harada and S. Oishi, ''A new approach to completely integrable partial differential equations by means of the singularity analysis,'' J. Phys. Soc. Jpn. **54**, 51–56 (1985).
[38] M. Ito, ''An extension of nonlinear evolution equations of the KdV (mK-dV) type to higher orders,'' J. Phys. Soc. Jpn. **49**, 771–778 (1980).
[39] P. R. Gordoa and A. Pickering, ''On a new nonisospectral variant of the Boussinesq hierarchy'' (submitted).
[40] J. Weiss, M. Tabor, and G. Carnevale, ''The Painlevé property for partial differential equations,'' J. Math. Phys. **24**, 522–526 (1983).
[41] C. Gilson and A. Pickering, ''Factorization and Painlevé analysis of a class of nonlinear third-order partial differential equations,'' J. Phys. A **28**, 2871–2888 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Elliptic Calogero–Moser systems and isomonodromic deformations

Kanehisa Takasaki[a)]
Department of Fundamental Sciences, Kyoto University, Yoshida, Sakyo-ku, Kyoto 606-8501, Japan

(Received 27 May 1999; accepted for publication 21 July 1999)

We show that various models of the elliptic Calogero–Moser systems are accompanied with an isomonodromic system on a torus. The isomonodromic partner is a nonautonomous Hamiltonian system defined by the same Hamiltonian. The role of the time variable is played by the modulus of the base torus. A suitably chosen Lax pair (with an elliptic spectral parameter) of the elliptic Calogero–Moser system turns out to give a Lax representation of the nonautonomous system as well. This Lax representation ensures that the nonautonomous system describes isomonodromic deformations of a linear ordinary differential equation on the torus on which the spectral parameter of the Lax pair is defined. A particularly interesting example is the ''extended twisted BC_l model'' recently introduced along with some other models by Bordner and Sasaki, who remarked that this system is equivalent to Inozemtsev's generalized elliptic Calogero–Moser system. We use the ''root-type'' Lax pair developed by Bordner *et al.* to formulate the associated isomonodromic system on the torus.
[S0022-2488(99)02911-4]

I. INTRODUCTION

In 1996, Manin[1] proposed a new expression of the sixth Painlevé equation. This is a differential equation of the form

$$(2\pi i)^2 \frac{d^2 q}{d\tau^2} = \sum_{a=0}^{3} \alpha_a \wp'(q+\omega_a), \tag{1}$$

where $\wp'(u)$ is the derivative of the Weierstrass $\wp$ function with primitive periods 1 and τ,

$$\wp(u) = \wp(u|1,\tau) = \frac{1}{u^2} + \sum_{(m,n)\neq(0,0)} \left(\frac{1}{(u+m+n\tau)^2} - \frac{1}{(m+n\tau)^2} \right), \tag{2}$$

ω_a $(a=0,1,2,3)$ are the origin and the three half-periods of the torus $E_\tau = \mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$,

$$\omega_0 = 0, \quad \omega_1 = \frac{1}{2}, \quad \omega_2 = \frac{1}{2} + \frac{\tau}{2}, \quad \omega_3 = \frac{\tau}{2}, \tag{3}$$

and α_a $(a=0,1,2,3)$ are the simple linear combinations $(\alpha_0,\alpha_1,\alpha_2,\alpha_3) = (\alpha, -\beta, \gamma, 1/2-\delta)$ of the four parameters α, β, γ, and δ of the sixth Painlevé equation

a)Electronic mail: takasaki@yukawa.kyoto-u.ac.jp

$$\frac{dy^2}{dx^2}=\frac{1}{2}\left(\frac{1}{y}+\frac{1}{y-1}+\frac{1}{y-x}\right)\left(\frac{dy}{dx}\right)^2-\left(\frac{1}{x}+\frac{1}{x-1}+\frac{1}{y-x}\right)\frac{dy}{dx}$$
$$+\frac{y(y-1)(y-x)}{x^2(x-1)^2}\left(\alpha+\beta\frac{x}{y^2}+\gamma\frac{x-1}{(y-1)^2}+\delta\frac{x(x-1)}{(y-x)^2}\right). \tag{4}$$

Manin's equation can be written in the Hamiltonian form

$$2\pi i\frac{dq}{d\tau}=p,\quad 2\pi i\frac{dp}{d\tau}=-\frac{\partial\mathcal{H}}{\partial q}, \tag{5}$$

with the Hamiltonian

$$\mathcal{H}=\frac{1}{2}p^2-\sum_{a=0}^{3}\alpha_a\wp(q+\omega_a). \tag{6}$$

Since the Hamiltonian depends on the modulus τ explicitly, this is a nonautonomous Hamiltonian system. In this new framework, Manin reconsidered the affine Weyl group symmetries of the sixth Painlevé equation discovered by Okamoto,[2] solutions for special values of α, β, γ, and δ constructed by Hitchin,[3] etc.

Manin's equation reveals an unexpected link between the Painlevé equation and the elliptic Calogero–Moser systems, i.e., the Calogero–Moser systems[4] with elliptic potentials. In order to see this relation, we introduce a new variable t and formally replace $2\pi i d/d\tau\to d/dt$ in the aforementioned equations. The outcome is the autonomous equation

$$\frac{d^2q}{dt^2}=\sum_{a=0}^{3}\alpha_a\wp'(q+\omega_a) \tag{7}$$

and its Hamiltonian form

$$\frac{dq}{dt}=p,\quad \frac{dp}{dt}=-\frac{\partial\mathcal{H}}{\partial q}. \tag{8}$$

If all α_n's take the same value $-g^2/8$, one can use an identity of the $\wp$ function to rewrite the above equation as

$$\frac{d^2q}{dt^2}=-\frac{g^2}{8}\sum_{a=0}^{3}\wp'(q+\omega_a)=-g^2\wp'(2q). \tag{9}$$

This is exactly the two-body elliptic Calogero–Moser system; the l-body elliptic Calogero–Moser system (A_{l-1} model) is defined by the Hamiltonian

$$\mathcal{H}=\frac{1}{2}\sum_{j=1}^{l}p_j^2+\frac{g^2}{2}\sum_{j\neq h}\wp(q_j-q_k). \tag{10}$$

As Krichever[5] demonstrated, this elliptic Calogero–Moser system is an isospectral integrable system with a Lax representation

$$\frac{\partial L(z)}{\partial t}=[L(z),M(z)], \tag{11}$$

where the Lax pair $L(z)$ and $M(z)$ are matrix-valued functions of a spectral parameter z on the torus E_τ. Furthermore, the general case falls into Inozemtsev's generalization of the elliptic Calogero–Moser system[6] defined by the Hamiltonian

$$\mathcal{H}=\frac{1}{2}\sum_{j=1}^{l} p_j^2+\frac{g_m^2}{2}\sum_{\epsilon,\epsilon'=\pm 1}\sum_{j\neq k}\wp(\epsilon q_j+\epsilon' q_k)+\frac{1}{2}\sum_{j=1}^{l}\sum_{a=0}^{3} g_a^2\wp(q_j+\omega_a). \tag{12}$$

Levin and Olshanetsky[7] developed a geometric formulation of isomonodromic systems on a general Riemann surface, and characterized Manin's equation as an isomonodromic system on the torus E_τ. Their interpretation of isomonodromic deformations is based on the notion of the Hitchin systems.[8] According to this interpretation, the coordinates q_j of Calogero–Moser particles are identified with the moduli of an SU(l) flat bundle on the torus E_τ, and the L-matrix $L(z)$ is nothing but the Higgs field on this bundle. (Such a link between the elliptic Calogero–Moser systems and the Hitchin systems was already pointed out before their work by Nekrasov[9] and Enriquez and Rubtsov.[10]) Isomonodromic deformations are special deformations of these geometric data as the complex structure of the base torus (or, equivalently, the modulus τ) varies. This geometric picture suggests a wide range of generalizations of isomonodromic deformations (see, e.g., the recent work of Levin and Olshanetsky[11]).

Unfortunately, however, it is only the special case with $\alpha_0=\alpha_1=\alpha_2=\alpha_3$ that was successfully treated in the formulation of Levin and Olshanetsky. This is simply because no suitable Lax representation was available for the Inozemtsev system. Inozemtsev[6] presented a Lax representation, but it is not suited for that purpose.

Recently, a new type of Lax pair—the root type Lax pair—was proposed by Bordner *et al.*[12,13,14] for various models of the elliptic Calogero–Moser systems including the Inozemtsev system. This is a Lax pair constructed on the basis of an underlying root system (e.g., the A_{l-1} root system for the aforementioned elliptic Calogero–Moser system, and the BC_l root system for the Inozemtsev system). The construction covers not only the ordinary elliptic Calogero–Moser systems (the ''untwisted models'') but also the ''twisted models'' introduced by D'Hoker and Phong[15] and their generalizations (the ''extended twisted models''). The Inozemtsev system coincides with the extended twisted BC_l model in the classification of Bordner and Sasaki.[14] In particular, the root type Lax pair for the extended twisted BC_l model gives a Lax representation to the aforementioned isospectral analogue of Manin's equation.

One of the goals of this paper is to show, using the root type Lax pair, that each of these elliptic Calogero–Moser systems are accompanied with an isomonodromic system on a torus. The first step of the construction is simply to replace the equations of motions

$$\frac{dq}{dt}=\{q,\mathcal{H}\},\quad \frac{dp}{dt}=\{p,\mathcal{H}\} \tag{13}$$

of the elliptic Calogero–Moser system by the nonautonomous system

$$2\pi i\frac{dq}{d\tau}=\{q,\mathcal{H}\},\quad 2\pi i\frac{dp}{d\tau}=\{p,\mathcal{H}\} \tag{14}$$

with the same Hamiltonian $\mathcal{H}$. We then rewrite this nonautonomous system into a Lax equation of the form

$$2\pi i\frac{\partial L(z)}{\partial\tau}+\frac{\partial M(z)}{\partial z}=[L(z),M(z)] \tag{15}$$

using a root type Lax pair $L(z)$ and $M(z)$. This Lax equation implies the Frobenius integrability of the linear system

$$\frac{\partial Y(z)}{\partial z}=L(z)Y(z),\quad 2\pi i\frac{\partial L(z)}{\partial \tau}+M(z)Y(z)=0, \tag{16}$$

from which one can deduce that the non-autonomous system is an isomonodromic system on the torus E_τ.

Actually, we shall use the root type Lax pair made of slightly different building blocks. The root type Lax pairs, like the previously known Lax pairs, contain complex analytic functions $x(u,z)$, $y(u,z)$, etc. that satisfy special functional equations (called the ''Calogero functional equations''[16]). Bordner *et al.* use the Weierstrass sigma function to construct those functions. We use the Jacobi theta function θ_1 instead. This is inspired by the work of Levin and Olshanetsky, who used substantially the same function to construct the *L*-matrix (i.e., the Higgs field in their framework) for isomonodromic systems on a torus. This minuscule difference is rather crucial for deriving an isomonodromic Lax equation as above.

The functions $x(u,z)$ and $y(u,z)$ that we use are, in fact, identical to the functions that Felder and Wieczerkowski[17] used in their study on the Knizhnik–Zamolodchikov–Bernard (KZB) equation.[18] This is by no means a coincidence. As Levin and Olshanetsky stressed, the KZB equation and the Hitchin system (or, rather, its isomonodromic version) are closely related.

In order to illustrate that our method also works for some other cases, we show a construction of an isomonodromic analogue for the ''spin generalization''[19] of the elliptic Calogero–Moser system. Actually, a multispin generalization of this construction is also possible, which is nothing but the genus-one case of Levin and Olshanetsky's framework.

This paper is organized as follows: In Sec. II, we illustrate our construction of isomonodromic systems in the case of the most classical A_{l-1} model. This will serve as a prototype of the subsequent discussion. Section III is devoted to the models treated by the root type Lax pairs, and Sec. IV to the spin generalization. Section V is for concluding remarks. Technically complicated calculations are collected in the Appendices.

II. ISOMONODROMIC SYSTEMS ON THE TORUS: A PROTOTYPE

We start with illustrating our construction for the most fundamental case—the A_{l-1} model and its Lax pair in the vector representation of $\mathrm{SU}(l)$.

A. A_{l-1} model of elliptic Calogero–Moser systems

The A_{l-1} model is defined by the Hamiltonian

$$\mathcal{H}=\frac{1}{2}\sum_{j=1}^{l} p_j^2+\frac{g^2}{2}\sum_{j\neq k}\wp(q_j-q_k). \tag{17}$$

Here q_j and p_j $(j=1,...,l)$ are the coordinates and momenta of the particles with the canonical Poisson brackets

$$\{q_j,p_k\}=\delta_{jk},\quad \{q_j,q_k\}=\{p_j,p_k\}=0. \tag{18}$$

Following Manin's equation, we normalize the primitive periods as

$$2\omega_1=1,\quad 2\omega_3=\tau. \tag{19}$$

The equations of motion are give by the canonical equations

$$\begin{aligned}\frac{dq_j}{dt}&=\{q_j,\mathcal{H}\}=p_j,\\ \frac{dp_j}{dt}&=\{p_j,\mathcal{H}\}=-g^2\sum_{k\neq j}\wp'(q_j-q_k).\end{aligned} \tag{20}$$

This elliptic Calogero–Moser system has a Lax pair of the form

$$L(z)=\sum_{j=1}^{l} p_j E_{jj}+ig\sum_{j\neq k} x(q_j-q_k,z)E_{jk},$$
$$M(z)=\sum_{j=1}^{l} D_j E_{jj}+ig\sum_{j\neq k} y(q_j-q_k,z)E_{jk}, \tag{21}$$

where E_{jk} is the matrix unit, $(E_{jk})_{mn}=\delta_{mj}\delta_{nk}$. The diagonal elements D_j of $M(z)$ are given by

$$D_j=ig\sum_{k\neq j}\wp(q_j-q_k), \tag{22}$$

and $x(u,z)$ is a function that satisfies, along with its u-derivative

$$y(u,z)=\frac{\partial x(u,z)}{\partial u}, \tag{23}$$

the functional equations,

$$x(u,z)y(v,z)-y(u,z)x(v,z)=x(u+v,z)(\wp(u)-\wp(v)), \tag{24}$$

$$x(u,z)y(-u,z)-y(u,z)x(-u,z)=\wp'(u), \tag{25}$$

$$x(u,z)x(-u,z)=\wp(z)-\wp(u). \tag{26}$$

Using these functional equations, one can easily prove the following well known result:[5]

Proposition 1: The matrices L(z) and M(z) satisfy the Lax equation

$$\frac{\partial L(u)}{\partial t}=[L(z),M(z)]. \tag{27}$$

As far as the elliptic Calogero–Moser system is concerned, the choice of $x(u,z)$ and $y(u,y)$ is rather irrelevant. A standard choice is the function

$$x(u,z)=\frac{\sigma(z-u)}{\sigma(z)\sigma(u)}, \tag{28}$$

where $\sigma(u)=\sigma(u|1,\tau)$ is the Weierstrass sigma function with primitive periods 1 and τ.

Thus, the elliptic Calogero–Moser system is an isospectral integrable system. An involutive set of conserved quantities can be extracted from the traces $\mathrm{Tr}\,L(z)^k$, $k=2,3,...$ of powers of the L-matrix. The quadratic trace is substantially the Hamiltonian itself,

$$\mathrm{Tr}\frac{L(z)^2}{2}=\mathcal{H}+(\text{independent of } p \text{ and } q). \tag{29}$$

The functions $x(u,z)$ and $y(u,z)$ based on the sigma function, however, are not very suited for constructing an isomonodromic system. We shall show an alternative in the next subsection.

B. Our choice of $x(u,z)$ and $y(u,z)$

Inspired by the work of Levin and Olshanetsky,[7] we take the following function $x(u,z)$ and its u-derivative $y(u,z)$ for constructing an isomonodromic Lax pair:

$$x(u,z)=\frac{\theta_1(z-u)\theta_1'(0)}{\theta_1(z)\theta_1(u)}. \tag{30}$$

Here $\theta_1(u)$ is one of Jacobi's elliptic theta functions,

$$\theta_1(u)=\theta_1(u|\tau)=-\sum_{n=-\infty}^{\infty}\exp\left(\pi i\tau\left(n+\frac{1}{2}\right)^2+2\pi i\left(n+\frac{1}{2}\right)\left(u+\frac{1}{2}\right)\right), \tag{31}$$

and $\theta_1'(u)$ its derivative. Accordingly, the partner $y(u,z)$ can be written

$$y(u,z)=-x(u,z)(\rho(u)+\rho(z-u)), \tag{32}$$

where $\rho(u)$ denotes the logarithmic derivative of $\theta_1(u)$,

$$\rho(u)=\frac{\theta_1'(u)}{\theta_1(u)}. \tag{33}$$

The function $\rho(u)$, too, plays an important role throughout this paper.

Proposition 2: These functions x(u,z) and y(u,z) satisfy the functional equations (24)–(26) and the differential equation

$$2\pi i\frac{\partial x(u,z)}{\partial\tau}+\frac{\partial^2 x(u,z)}{\partial u\partial z}=0. \tag{34}$$

The last differential equation (a kind of 1+2-dimensional ''heat equation'') is a characteristic of our (x,y) pair, and plays a key role in our construction of isomonodromic systems.

We give a proof of these properties in Appendix A. The following are supplementary remarks on these functions:

(a) The proof of (24–26) is based on the following analytical properties of $x(u,z)$:
(1) $x(u,z)$ is a meromorphic function of u and z. The poles on the u plane and the z plane are both located at the lattice points $u=m+n\tau$ and $z=m+n\tau(m,n\in\mathbb{Z})$.
(2) $x(u,z)$ has the following quasiperiodicity:

$$\begin{gathered}x(u+1,z)=x(u,z),\ \ x(u+\tau,z)=e^{2\pi iz}x(u,z),\\ x(u,z+1)=x(u,z),\ \ x(u,z+\tau)=e^{2\pi iu}x(u,z).\end{gathered} \tag{35}$$

(3) At the origin of the u and z planes, $x(u,z)$ exhibits the following singular behavior:

$$\begin{gathered}x(u,z)=\frac{1}{u}-\rho(z)+O(u)\ \ (u\to 0),\\ x(u,z)=-\frac{1}{z}+\rho(u)+O(z)\ \ (z\to 0).\end{gathered} \tag{36}$$

(b) These properties are an immediate consequence of the following well known facts:
(1) $\theta_1(u)$ is an entire function with simple zeros at the lattice points $u=m+n\tau(m,n\in\mathbb{Z})$.
(2) $\theta_1(u)$ is an odd and quasiperiodic function,

$$\begin{gathered}\theta_1(-u)=\theta_1(u+1)=-\theta_1(u),\\ \theta_1(u+\tau)=-e^{-\pi i\tau-2\pi iu}\theta_1(u).\end{gathered} \tag{37}$$

(c) One can similarly see the following analytical properties of $\rho(u)$:
(1) $\rho(u)$ is a meromorphic function with poles at the lattice points $u=m+n\tau(m,n\in\mathbb{Z})$.
(2) $\rho(u)$ is an odd function with additive quasiperiodicity,

$$\rho(-u)=-\rho(u),\ \ \rho(u+1)=\rho(u),\ \ \rho(u+\tau)=\rho(u)-2\pi i. \tag{38}$$

(3) At the origin $u=0$, $\rho(u)$ exhibits the following singular behavior:

$$\rho(u)=\frac{1}{u}+\frac{\theta_1'''(0)}{3\theta_1'(0)}u+O(u^3)\ \ (u\to 0). \tag{39}$$

(d) The proof of (34) is based on the well known ‘‘heat equation’’

$$4\pi i\frac{\partial\theta_1(u)}{\partial\tau}=\frac{\partial^2\theta_1(u)}{\partial u^2} \tag{40}$$

of the Jacobi theta function.

C. Isomonodromic deformations

Replacing $d/dt\to 2\pi i d/d\tau$, one obtains a nonautonomous Hamiltonian system,

$$\begin{aligned} 2\pi i\frac{dq_j}{d\tau}&=\{q_j,\mathcal{H}\}=p_j,\\ 2\pi i\frac{dp_j}{d\tau}&=\{p_j,\mathcal{H}\}=-g^2\sum_{k\neq j}\wp'(q_j-q_k). \end{aligned} \tag{41}$$

We now demonstrate that this gives an isomonodromic system on the torus E_τ. A key is the following Lax equation:

Proposition 3: L(z) and M(z) satisfy the Lax equation

$$2\pi i\frac{\partial L(z)}{\partial\tau}+\frac{\partial M(z)}{\partial z}=[L(z),M(z)]. \tag{42}$$

Proof: Let us notice that the right-hand side of the isospectral Lax equation is in fact the Poisson bracket of $L(z)$ and the Hamiltonian,

$$[L(z),M(z)]=\frac{\partial L(z)}{\partial t}=\{L(z),\mathcal{H}\}. \tag{43}$$

Since the phase space and the Hamiltonian are the same as those of the original system, the relation $[L(z),M(z)]=\{L(z),\mathcal{H}\}$ persists in the present setup. Thus the right-hand side of the Lax equation can be written

$$[L(z),M(z)]=\{L(z),\mathcal{H}\}=\sum_{j=1}^{l}\{p_j,\mathcal{H}\}E_{jj}+ig\sum_{j\neq k}\{q_j-q_k,\mathcal{H}\}y(q_j-q_k,z)E_{jk}. \tag{44}$$

On the other hand,

$$\begin{aligned} 2\pi i\frac{\partial L(z)}{\partial\tau}+\frac{\partial M(z)}{\partial z}&=\sum_{j=1}^{l}2\pi i\frac{dp_j}{d\tau}E_{jj}+ig\sum_{j\neq k}2\pi i\left(\frac{dq_j}{d\tau}-\frac{dq_k}{d\tau}\right)y(q_j-q_k,z)E_{jk}\\ &\quad+ig\sum_{j\neq k}\left(2\pi i\frac{\partial x(u,z)}{\partial\tau}+\frac{\partial y(u,z)}{\partial z}\right)_{u=q_j-q_k}E_{jk}. \end{aligned} \tag{45}$$

The last sum vanishes because of the ‘‘heat equation’’ (34). The other part coincides, term-by-term, with the above expression of the commutator $[L(z),M(z)]$. Q.E.D.

This Lax equation enables us to interpret the nonautonomous Hamiltonian system as an isomonodromic system on the torus E_τ. The Lax equation is nothing but the Frobenius integrability condition of a linear system of the form

$$\frac{\partial Y(z)}{\partial z}=L(z)Y(z),\quad 2\pi i\frac{\partial Y(z)}{\partial \tau}+M(z)Y(z)=0. \tag{46}$$

The first equation is an ordinary differential equation on the torus E_τ, and has a regular singular point at $z=0$. Analytic continuation of the solution along this singular point yields a monodromy matrix Γ_0. Besides this local monodromy matrix, there are global monodromy matrices Γ_α and Γ_β that arise in analytic continuation along the $\alpha(z\to z+1)$ and $\beta(z\to z+\tau)$ cycles. The second equation of the above linear system implies that these monodromy matrices are left invariant as τ varies.

Let us specify this observation in more detail. The situation is more complicated than isomonodromic systems on the Riemann sphere. The monodromy of $L(z)$ and $M(z)$ themselves are nontrivial,

$$L(z+1)=L(z),\quad M(z+1)=M(z),$$
$$L(z+\tau)=e^{2\pi iQ}L(z)e^{-2\pi iQ}, \tag{47}$$
$$M(z+\tau)=e^{2\pi iQ}(M(z)+2\pi iL(z))e^{-2\pi iQ}-2\pi iP,$$

where $Q=\Sigma_{j=1}^{l}q_jE_{jj}$ and $P=\Sigma_{j=1}^{l}p_jE_{jj}$. These relations are a consequence of the quasi-periodicity of $x(u,z)$, $y(u,z)$, and $\rho(z)$. The monodromy of $L(z)$ implies that $Y(z)$ has to be treated as a section of a nontrivial $GL(l,\mathbb{C})$-bundle (or $SL(l,\mathbb{C})$-bundle, if we take the center of mass frame with $\Sigma_{j=1}^{l}p_j=0$) on the torus E_τ. The monodromy matrices Γ_0, Γ_α, and Γ_β thus arise as follows:

$$Y(ze^{2\pi i})=Y(z)\Gamma_0,\quad Y(z+1)=Y(z)\Gamma_\alpha,\quad Y(z+\tau)=e^{2\pi iQ}Y(z)\Gamma_\beta. \tag{48}$$

Note that the exponential factor in the last relation reflects the nontrivial monodromy of $L(z)$ along the β-cycle. Having this monodromy structure of $Y(z)$, one can deduce the following fundamental observation:

Proposition 4: The monodromy matrices do not depend on τ, *i.e.,*

$$\frac{d\Gamma_0}{d\tau}=\frac{d\Gamma_\alpha}{d\tau}=\frac{d\Gamma_\beta}{d\tau}=0. \tag{49}$$

Proof: Let us rewrite the second equation of the linear system as

$$M(z)=-2\pi i\frac{\partial Y(z)}{\partial \tau}Y(z)^{-1}, \tag{50}$$

and examine the implication of the monodromy structure of $Y(z)$ noted above. This leads to the following relations:

$$M(ze^{2\pi i}) = M(z) - 2\pi i Y(z)\frac{\partial \Gamma_0}{\partial \tau}\Gamma_0^{-1}Y(z)^{-1},$$

$$M(z+1) = M(z) - 2\pi i Y(z)\frac{\partial \Gamma_\alpha}{\partial \tau}\Gamma_\alpha^{-1}Y(z)^{-1}, \tag{51}$$

$$M(z+\tau) = e^{2\pi i Q}(M(z)+2\pi i L(z))e^{-2\pi i Q} - 2\pi i P - 2\pi i Y(z)\frac{\partial \Gamma_\beta}{\partial \tau}\Gamma_\beta^{-1}Y(z)^{-1}.$$

(We have used the relation $2\pi i dQ/d\tau = P$.) These relations are consistent with the aforementioned monodromy structure of $M(z)$ if and only if the monodromy matrices of $Y(z)$ are independent of τ. Q.E.D.

III. ELLIPTIC CALOGERO–MOSER SYSTEMS BASED ON ROOT SYSTEMS

Here we consider the elliptic Calogero–Moser systems associated with a general irreducible (but not necessary reduced) root system Δ.

In the following, the root system Δ is assumed to be realized in an l-dimensional Euclidean space $M = \mathbb{R}^l$. Let $x \cdot y$ denote the inner product of two vectors in M and its bilinear extension to the complexification $M^{\mathbb{C}} = M \otimes_{\mathbb{R}} \mathbb{C}$. The dual space $M^* = \mathrm{Hom}(M, \mathbb{R})$ of M is identified with M by this inner product. Each element $\alpha \in \Delta$ induces a reflection (the Weyl reflection) $s_\alpha(x) = x - (2\alpha \cdot x/\alpha \cdot \alpha)\alpha$. This gives a representation of the Weyl group $W(\Delta)$ on M. The root system Δ is invariant under the action of this Weyl group.

The elliptic Calogero–Moser system associated with the root system Δ is a Hamiltonian system on $M \times M$ (or its complexification $M^{\mathbb{C}} \times M^{\mathbb{C}}$). The orthogonal coordinates $(q,p) = (q_1,\ldots,q_l,p_1,\ldots,p_l)$ of $M \times M$ give canonical coordinates and momenta with the Poisson brackets

$$\{q_j,p_k\} = \delta_{jk}, \quad \{p_j,p_k\} = \{q_j,q_k\} = 0. \tag{52}$$

A. Simply laced models

We first consider the case of simply laced (A_{l-1}, D_l, and E_l) root systems. The associated elliptic Calogero–Moser system is defined by the Hamiltonian

$$\mathcal{H} = \frac{1}{2}p \cdot p + \frac{g^2}{2}\sum_{\alpha \in \Delta} \wp(\alpha \cdot q). \tag{53}$$

Here g is a coupling constant, and $\wp(u)$ the Weierstrass $\wp$ function with primitive periods 1 and τ. The equations of motion can be written

$$\frac{dq}{dt} = p, \quad \frac{dp}{dt} = -\frac{g^2}{2}\sum_{\alpha \in \Delta} \wp'(\alpha \cdot q)\alpha. \tag{54}$$

We first review the "root type" Lax pair of Bordner *et al.* for these models,[12] then explain how to convert these isospectral systems to isomonodromic systems.

1. Root-type Lax pair

The "root-type" Lax pair for these simply laced models are $\Delta \times \Delta$ matrices, i.e., matrices whose rows and columns are indexed by the root system Δ. They are made of three parts,

$$L(z) = P + X_1(z) + X_2(z), \quad M(z) = D + Y_1(z) + Y_2(z). \tag{55}$$

P and D are diagonal matrices,

$$P_{\beta\gamma} = p \cdot \beta \delta_{\beta\gamma}, \quad D_{\beta\gamma} = D_\beta \delta_{\beta\gamma} \quad (\beta, \gamma \in \Delta), \tag{56}$$

and the diagonal elements D_β of D are given by

$$D_\beta = ig\wp(\beta\cdot q) + ig\sum_{\gamma\in\Delta,\beta\cdot\gamma=1}\wp(\gamma\cdot q). \tag{57}$$

$X_1(z)$, etc. are diagonal-free matrices of the form

$$\begin{aligned}
X_1(z) &= ig\sum_{\alpha\in\Delta} x(\alpha\cdot q,z)E(\alpha),\\
X_2(z) &= 2ig\sum_{\alpha\in\Delta} x(\alpha\cdot q,2z)E(2\alpha),\\
Y_1(z) &= ig\sum_{\alpha\in\Delta} y(\alpha\cdot q,z)E(\alpha),\\
Y_2(z) &= ig\sum_{\alpha\in\Delta} y(\alpha\cdot q,2z)E(2\alpha),
\end{aligned} \tag{58}$$

where $x(u,z)$ and $y(u,z)$ are the same as the functions used in the previous section, and $E(\alpha)$ and $E(2\alpha)$ are $\Delta\times\Delta$ matrices of the form

$$E(\alpha)_{\beta\gamma} = \delta_{\alpha,\beta-\gamma},\quad E(2\alpha)_{\beta\gamma} = \delta_{2\alpha,\beta-\gamma}\quad (\beta,\gamma\in\Delta). \tag{59}$$

(We have slightly modified the notation of Bordner *et al.*; $x(u,2z)$, $y(u,2z)$, and $E(2\alpha)$ amount to $x_d(u,z)$, $y_d(u,z)$, and $E_d(\alpha)$ in their notation.)

These matrices satisfy the Lax equation

$$\frac{\partial L(z)}{\partial t} = [L(z),M(z)] \tag{60}$$

under the equations of motions. The traces $\mathrm{Tr}\,L(z)^k$, $k=2,3,\ldots$, of powers of $L(z)$ are conserved, and an involutive set of conserved quantities can be extracted from these traces. The Hamiltonian itself can be reproduced from the quadratic trace $\mathrm{Tr}\,L(z)^2$. We refer the details of these results to the paper of Bordner *et al.*[12] The choice of $x(u,z)$ and $y(u,z)$ is irrelevant in this case, too.

Thus, in particular, the A_{l-1} model turns out to have at least two distinct Lax pairs—the Lax pair of $l\times l$ matrices realized in the vector representation of $sl(l)$, and the Lax pair of $l(l-1)\times l(l-1)$ matrices based on the A_{l-1} root system. This is also the case for the other simply laced root systems. Bordner *et al.* call the Lax pairs of the first type the ''minimal type,'' because they are realized in a minimal representation of the associated (not necessary simply laced) Lie algebra. It should be noted that the ''root type'' Lax pairs do not possess a Lie algebraic structure; unlike the usual root basis of simple Lie algebras, the matrices $E(\alpha)$ and $E(2\alpha)$ are not closed under the Lie bracket.

2. Isomonodromic system

The prescription for constructing an isomonodromic analogue is the same as the previous case, namely, to replace $d/dt\to 2\pi i d/d\tau$. This converts the equations of motion of the elliptic Calogero–Moser system to the nonantonomous system

$$2\pi i\frac{dq}{dt} = p,\quad 2\pi i\frac{dp}{dt} = -\frac{g^2}{2}\sum_{\alpha\in\Delta}\wp'(\alpha\cdot q)\alpha. \tag{61}$$

Let $x(u,z)$ be the function defined in (30), and $y(u,z)$ its u-derivative. The following are the keys to an isomonodromic interpretation.

Proposition 5: 1. L(z) and K(z) satisfy the Lax equation

$$2\pi i\frac{\partial L(z)}{\partial\tau}+\frac{\partial M(z)}{\partial z}=[L(z),M(z)]. \tag{62}$$

2. L(z) and M(z) have the following monodromy properties:

$$\begin{gathered}L(z+1)=L(z),\ \ M(z+1)=M(z),\\ L(z+\tau)=e^{2\pi iQ}L(z)e^{-2\pi iQ},\\ M(z+\tau)=e^{2\pi iQ}(M(z)+2\pi iL(z))e^{-2\pi iQ}-2\pi iP,\end{gathered} \tag{63}$$

where Q is the diagonal matrix with matrix elements $Q_{\beta\gamma}=q\cdot\beta\delta_{\beta\gamma}$.

Proof: The proof is almost the same as the proof for the isomonodromic Lax pair of the A_{l-1} model in the vector representation. Let us first verify the Lax equation. The right-hand side of the Lax equation can be written

$$[L(z),M(z)]=\{P,\mathcal{H}\}+ig\sum_{\alpha\in\Delta}\{\alpha\cdot q,\mathcal{H}\}y(\alpha\cdot q,z)E(\alpha)+2ig\sum_{\alpha\in\Delta}\{\alpha\cdot q,\mathcal{H}\}y(\alpha\cdot q,2z)E(2\alpha). \tag{64}$$

On the other hand,

$$\begin{aligned}2\pi i\frac{\partial L(z)}{\partial\tau}+\frac{\partial M(z)}{\partial z}=&\,2\pi i\frac{\partial P}{\partial\tau}+ig\sum_{\alpha\in\Delta}2\pi i\frac{\partial\alpha\cdot q}{\partial\tau}y(\alpha\cdot q,z)E(\alpha)\\ &+2ig\sum_{\alpha\in\Delta}2\pi i\frac{\partial\alpha\cdot q}{\partial\tau}y(\alpha\cdot q,2z)E(2\alpha)\\ &+ig\sum_{\alpha\in\Delta}\left(2\pi i\frac{\partial x(u,z)}{\partial\tau}+\frac{\partial y(u,z)}{\partial z}\right)_{u=\alpha\cdot q}E(\alpha)\\ &+2ig\sum_{\alpha\in\Delta}\left(4\pi i\frac{\partial x(u,2z)}{\partial\tau}+\frac{\partial y(u,2z)}{\partial z}\right)_{u=\alpha\cdot q}E(2\alpha).\end{aligned} \tag{65}$$

The last two sums vanish because of (34). The other part coincides by the equations of motion. Thus we obtain the Lax equation. Let us next consider the monodromy of $L(z)$ and $M(z)$. Note the commutation relations

$$[Q,E(\alpha)]=q\cdot\alpha E(\alpha),\ \ [Q,E(2\alpha)]=2q\cdot\alpha E(2\alpha), \tag{66}$$

which can be exponentiated as follows:

$$e^{2\pi iQ}E(\alpha)e^{-2\pi iQ}=e^{2\pi iq\cdot\alpha}E(\alpha),\ \ e^{2\pi iQ}E(2\alpha)e^{-2\pi iQ}=e^{4\pi iq\cdot\alpha}E(2\alpha). \tag{67}$$

The monodromy property of $L(z)$ and $M(z)$ can be derived from these relations and the quasiperiodicity of $x(u,z)$ and $y(u,z)$. Q.E.D.

The rest is parallel to the case in the previous section. The only difference is that the ordinary differential equation

$$\frac{dY(z)}{dz}=L(z)Y(z) \tag{68}$$

on the torus E_τ has *four* regular singular points at $z=0,\omega_1,\omega_2,\omega_3$. The latter three singular points originates in $X_2(z)$. Let Γ_a $(a=0,1,2,3)$ denote the monodromy matrices in analytic continuation of $Y(z)$ around these four points. The Lax equation implies that these local monodromy matrices and the two global ones Γ_α and Γ_β are independent of τ,

$$\frac{\partial\Gamma_0}{\partial\tau}=\cdots=\frac{\partial\Gamma_3}{\partial\tau}=\frac{\partial\Gamma_\alpha}{\partial\tau}=\frac{\partial\Gamma_\beta}{\partial\tau}=0. \tag{69}$$

B. Nonsimply laced models

The elliptic Calogero–Moser system associated with a nonsimply laced (B_l, C_l, F_4, G_2, and BC_l) root systems can have several independent coupling constants, one for each Weyl group orbit in the root system. The root type Lax pairs are extended to the nonsimply laced cases by Bordner *et al.*[13] As they pointed out, one can construct a different root type Lax pair for each Weyl group orbit of the root system. Thus the B_l, C_l, F_4, and G_2 models have, respectively, two distinct Lax pairs based on the orbits of long and short roots, whereas the BC_l model has three based on the orbits of long, middle, and short roots. Note that each Weyl group orbit consists of roots of the same length.

Although all the nonsimply laced models can be treated in the same way, let us illustrate our construction of isomonodromic systems for the BC_l model. This is also intended to be a prototype of the case that we shall consider in the next subsection.

1. BC_l model

The BC_l root system can be realized in $M=\mathbb{R}^l$,

$$\begin{aligned}
&\Delta(BC_l)=\Delta_l\cup\Delta_m\cup\Delta_s,\\
&\Delta_l=\{\pm 2e_j|1\leq j\leq l\}\quad(\text{long roots}),\\
&\Delta_m=\{\pm e_j\pm e_k|j\neq k\}\quad(\text{middle roots}),\\
&\Delta_s=\{\pm e_j|1\leq j\leq l\}\quad(\text{short roots}),
\end{aligned} \tag{70}$$

where $e_1,\ldots,e_l$ are the standard orthonormal basis of $\mathbb{R}^l$. Δ_l, Δ_m, and Δ_s give the three Weyl group orbits.

The Hamiltonian of the BC_l model takes the form

$$\mathcal{H}=\frac{1}{2}p\cdot p+\frac{g_m^2}{2}\sum_{\alpha\in\Delta_m}\wp(\alpha\cdot q)+\frac{g_l^2}{4}\sum_{\alpha\in\Delta_l}\wp(\alpha\cdot q)+\tilde{g}_s^2\sum_{\alpha\in\Delta_s}\wp(\alpha\cdot q). \tag{71}$$

The equations of motion can be written

$$\begin{aligned}
&\frac{dq}{d\tau}=p,\\
&\frac{dp}{d\tau}=-\frac{g_m^2}{2}\sum_{\alpha\in\Delta_m}\wp'(\alpha\cdot q)\alpha-\frac{g_l^2}{4}\sum_{\alpha\in\Delta_l}\wp'(\alpha\cdot q)\alpha-\tilde{g}_s^2\sum_{\alpha\in\Delta_s}\wp'(\alpha\cdot q)\alpha.
\end{aligned} \tag{72}$$

g_m, g_l, and $\tilde{g}_s$ are three independent coupling constants. $\tilde{g}_s$ is a modified (''renormalized'' in the terminology of Bordner *et al.*) coupling constant connected with a more fundamental (''bare,'' so to speak) coupling constant g_s as

$$\tilde{g}_s^2 = g_s^2 + \frac{g_s g_l}{2}. \tag{73}$$

The ''bare'' coupling constant appears in the construction of a Lax pair.

2. Root-type Lax pair for the BC_l model

As mentioned above, there are at least three root type Lax pairs based on the three Weyl group orbits Δ_m, Δ_l, and Δ_s. Bordner *et al.* constructed only one of them, namely, a Lax pair based on Δ_m. Here we present a Lax pair based on Δ_s. This is a $2l \times 2l$ system, much smaller than the Lax pair based on Δ_m, and presumably more suitable for studying the associated isomonodromic deformations.

The Lax pair are indexed by Δ_s and take the following form:

$$\begin{aligned} L(z) &= P + X_1(z) + X_2(z) + X_3(z), \\ M(z) &= D + Y_1(z) + Y_2(z) + Y_3(z). \end{aligned} \tag{74}$$

P and D are diagonal matrices,

$$P_{\beta\gamma} = p \cdot \beta \delta_{\beta\gamma}, \quad D_{\beta\gamma} = D_\beta \delta_{\beta\gamma} \quad (\beta, \gamma \in \Delta_s), \tag{75}$$

and the diagonal elements of D are given by

$$D_\beta = i g_m \sum_{\gamma \in \Delta_s, \beta \cdot \gamma = 1} \wp(\gamma \cdot q) + i g_l \wp(2\beta \cdot q) + i g_s \wp(\beta \cdot q). \tag{76}$$

$X_1(z)$, etc. are diagonal-free matrices of the form

$$\begin{aligned} X_1(z) &= i g_m \sum_{\alpha \in \Delta_m} x(\alpha \cdot q, z) E(\alpha), \quad X_2(z) = i g_l \sum_{\alpha \in \Delta_l} x(\alpha \cdot q, z) E(\alpha), \\ X_3(z) &= 2 i g_s \sum_{\alpha \in \Delta_s} x(\alpha \cdot q, 2z) E(2\alpha), \quad Y_1(z) = i g_m \sum_{\alpha \in \Delta_m} y(\alpha \cdot q, z) E(\alpha), \\ Y_2(z) &= i g_l \sum_{\alpha \in \Delta_l} y(\alpha \cdot q, z) E(\alpha), \quad Y_3(z) = i g_s \sum_{\alpha \in \Delta_s} y(\alpha \cdot q, 2z) E(2\alpha), \end{aligned} \tag{77}$$

where

$$E(\alpha)_{\beta\gamma} = \delta_{\alpha, \beta - \gamma}, \quad E(2\alpha)_{\beta\gamma} = \delta_{2\alpha, \beta - \gamma} \quad (\beta, \gamma \in \Delta_s). \tag{78}$$

This Lax pair is a specialization of the Lax pair for the extended twisted model that we shall present in the next subsection.

3. Isomonodromic system

This system, too, can be converted to an isomonodromic system by replacing $d/dt \to 2\pi i d/d\tau$. The equations of motion are a nonautonomous system of the form

$$\begin{aligned} 2\pi i \frac{dq}{d\tau} &= p, \\ 2\pi i \frac{dp}{d\tau} &= -\frac{g_m^2}{2} \sum_{\alpha \in \Delta_m} \wp'(\alpha \cdot q)\alpha - \frac{g_l^2}{4} \sum_{\alpha \in \Delta_l} \wp'(\alpha \cdot q)\alpha - \tilde{g}_s^2 \sum_{\alpha \in \Delta_s} \wp'(\alpha \cdot q)\alpha. \end{aligned} \tag{79}$$

The following can be verified just as in the case of simply lased models:

(1) $L(z)$ and $M(z)$ satisfy the Lax equation

$$2\pi i \frac{\partial L(z)}{\partial \tau} + \frac{\partial M(z)}{\partial z} = [L(z),M(z)]. \tag{80}$$

(2) $L(z)$ and $M(z)$ have the following monodromy property:

$$L(z+1)=L(z), \quad M(z+1)=M(z),$$

$$L(z+\tau)=\epsilon^{2\pi iQ}L(z)e^{-2\pi iQ}, \tag{81}$$

$$M(z+\tau)=e^{2\pi iQ}(M(z)+2\pi iL(z))e^{-2\pi iQ}-2\pi iP.$$

The interpretation of this Lax equation, too, is parallel to the simply laced models. The ordinary differential equation

$$\frac{dY(z)}{dz}=L(z)Y(z) \tag{82}$$

on the torus E_τ has four regular singular points at $z=0,\omega_1,\omega_2,\omega_3$. The local monodromy matrices Γ_a $(a=0,1,2,3)$ at these points and the global monodromy matrices Γ_α and Γ_β are invariant as τ varies.

C. Twisted and extended twisted models

We now proceed to the ''twisted'' and ''extended twisted'' models. The Hamiltonian of the untwisted models can be generally written

$$\mathcal{H}=\frac{1}{2}p\cdot p+\frac{1}{2}\sum_{\alpha\in\Delta} g_{|\alpha|}^2\wp(\alpha\cdot q). \tag{83}$$

The twisted models, introduced by D'Hoker and Phong[15] for nonsimply laced root systems, are defined by a Hamiltonian of the form

$$\mathcal{H}=\frac{1}{2}p\cdot p+\frac{1}{2}\sum_{\alpha\in\Delta} g_{|\alpha|}^2\wp_{\nu(\alpha)}(\alpha\cdot q), \tag{84}$$

where $\wp_{\nu(\alpha)}(u)$ are the $\wp$-functions with suitably rescaled primitive periods. D'Hoker and Phong proved the integrability of those twisted models by constructing a Lax pair in a representation of the associated Lie algebra. Bordner and Sasaki[14] proposed an alternative approach based on root systems rather than Lie algebras, and pointed out that the twisted model of the B_l, C_l, and BC_l types can be further extended. The extended twisted models have one (for the B_l and C_l models) or two (for the BC_l model) extra types of elliptic potentials.

Our construction of isomonodromic systems can be extended to the twisted and extended twisted models. We illustrate this result, just as in the previous subsection, for the BC_l model. As Bordner and Sasaki noted, the extended twisted BC_l model is made of five different types of elliptic potentials, and coincides with the Inozemtsev system.[6]

1. Extended twisted BC_l model

The extended twisted BC_l model is defined by the Hamiltonian

$$\mathcal{H}=\frac{1}{2}p\cdot p+\frac{g_m^2}{2}\sum_{\alpha\in\Delta_m}\wp(\alpha\cdot q)+\frac{g_{l1}^2}{4}\sum_{\alpha\in\Delta_l}\wp(\alpha\cdot q)+\frac{\tilde{g}_{l2}^2}{4}\sum_{\alpha\in\Delta_l}\wp^{(2)}(\alpha\cdot q)$$
$$+\tilde{g}_{s1}^2\sum_{\alpha\in\Delta_s}\wp(\alpha\cdot q)+\tilde{g}_{s2}^2\sum_{\alpha\in\Delta_s}\wp^{(1/2)}(\alpha\cdot q). \tag{85}$$

$\tilde{g}_{l2}$, $\tilde{g}_{s1}$, and $\tilde{g}_{s2}$ are ''renormalized'' coupling constants, which are related to unrenormalized coupling constants g_{l2}, g_{s1}, and g_{s2} as follows:

$$\tilde{g}_{l2}^2=g_{l2}^2+2g_{l1}g_{l2},$$
$$\tilde{g}_{s1}^2=g_{s1}^2+2g_{s1}g_{s2}+\tfrac{1}{2}(g_{s1}g_{l1}+g_{s1}g_{l2}+g_{s2}g_{l2}), \tag{86}$$
$$\tilde{g}_{s2}^2=g_{s2}^2+\frac{g_{s2}g_{l1}}{2}.$$

$\wp^{(1/2)}$ and $\wp^{(2)}$ are the $\wp$ functions with rescaled primitive periods

$$\wp^{(1/2)}(u)=\wp(u|\tfrac{1}{2},\tau),\quad \wp^{(2)}(u)=\wp(u|2,\tau). \tag{87}$$

(This Hamiltonian is slightly different from the Hamiltonian of Bordner and Sasaki, though the contents are essentially the same. With this modification, this model reduces to the untwisted BC_l model as $g_{l2}\rightarrow 0$ and $g_{s2}\rightarrow 0$.)

2. Root-type Lax pair for the extended twisted BC_l model

One can construct, like the untwisted model, three different root type Lax pairs can be constructed based on the three Weyl group orbits Δ_m, Δ_l, and Δ_s. The Lax pair based on Δ_m is presented by Bordner and Sasaki. The Lax pair based on Δ_s can be obtained by modifying the Lax pair for the untwisted BC_l model as follows.

The Lax pair $L(z)$ and $M(z)$ are indexed by Δ_s and made of four parts,

$$L(z)=P+X_1(z)+X_2(z)+X_3(z),$$
$$M(z)=D+Y_1(z)+Y_2(z)+Y_3(z). \tag{88}$$

The diagonal matrix P is the same as the P in the untwisted model. The diagonal matrices of D are given by

$$D_\beta=ig_m\sum_{\gamma\in\Delta_m,\beta\cdot\gamma=1}\wp(\gamma\cdot q)+ig_{l1}\wp(2\beta\cdot q)+ig_{l2}\wp^{(2)}(2\beta\cdot q)+ig_{s1}\wp(\beta\cdot q)+ig_{s2}\wp^{(1/2)}(\beta\cdot q). \tag{89}$$

$X_1(z)$ and $Y_1(z)$ are the same as those for the untwisted model. The other matrices take the following form:

$$
\begin{aligned}
X_2(z) &= \sum_{\alpha\in\Delta_l} (ig_{l1}x(\alpha\cdot q,z)+ig_{l2}x^{(2)}(\alpha\cdot q,z))E(\alpha),\\
X_3(z) &= \sum_{\alpha\in\Delta_s} (2ig_{s1}x(\alpha\cdot q,2z)+2ig_{s2}x^{(1/2)}(\alpha\cdot q,2z))E(2\alpha),\\
Y_2(z) &= \sum_{\alpha\in\Delta_l} (ig_{l1}y(\alpha\cdot q,z)+ig_{l2}y^{(2)}(\alpha\cdot q,z))E(\alpha),\\
Y_3(z) &= \sum_{\alpha\in\Delta_s} (ig_{s1}y(\alpha\cdot q,2z)+ig_{s2}y^{(1/2)}(\alpha\cdot q,2z))E(2\alpha).
\end{aligned}
\tag{90}
$$

This Lax pair reduces to the Lax pair of the untwisted model if $g_{l2}=0$ and $g_{s2}=0$.

The new objects arising here are the functions $x^{(1/2)}(u,z)$, $x^{(2)}(u,z)$ and their u-derivatives

$$
y^{(1/2)}(u,z)=\frac{\partial x^{(1/2)}(u,z)}{\partial u},\quad y^{(2)}(u,z)=\frac{\partial x^{(2)}(u,z)}{\partial u}.
\tag{91}
$$

For the consistency of the Lax equation

$$
\frac{\partial L(z)}{\partial t}=[L(z),M(z)],
\tag{92}
$$

these functions have to satisfy several functional equations. D'Hoker and Phong[15] and Bordner and Sasaki[14] use a set of functions based on the Weierstrass sigma functions. We use the function $x(u,z)=x(u,z|\tau)$ defined in (30) and its modifications

$$
\begin{aligned}
x^{(1/2)}(u,z) &= 2x(2u,z|2\tau)=\frac{2\theta_1(z-2u|2\tau)\theta_1'(0|2\tau)}{\theta_1(z|2\tau)\theta_1(2u|2\tau)},\\
x^{(2)}(u,z) &= \frac{1}{2}x\left(\frac{u}{2},z\,\middle|\,\frac{\tau}{2}\right)=\frac{\theta_1\left(z-\frac{u}{2}\,\middle|\,\frac{\tau}{2}\right)\theta_1'\left(0\,\middle|\,\frac{\tau}{2}\right)}{2\theta_1\left(z\,\middle|\,\frac{\tau}{2}\right)\theta_1\left(\frac{u}{2}\,\middle|\,\frac{\tau}{2}\right)}.
\end{aligned}
\tag{93}
$$

These functions $x^{(1/2)}(u,z)$ and $x^{(2)}(u,z)$, too, satisfy 1+2-dimensional ''heat equations'' of the form

$$
\begin{aligned}
2\pi i\frac{\partial x^{(1/2)}(u,z)}{\partial\tau}+\frac{\partial^2 x^{(1/2)}(u,z)}{\partial u\,\partial z} &= 0,\\
2\pi i\frac{\partial x^{(2)}(u,z)}{\partial\tau}+\frac{\partial^2 x^{(2)}(u,z)}{\partial u\,\partial z} &= 0.
\end{aligned}
\tag{94}
$$

The functional identities for these functions and the proof of the Lax equation are presented in Appendices B and C.

3. Isomonodromic system

Replacing $d/dt\to 2\pi i d/d\tau$, we obtain a nonautonomous Hamiltonian system with the same Hamiltonian. The isomonodromic interpretation of this nonautonomous system is again based on the following two observations:

(1) $L(z)$ and $M(z)$ satisfy the Lax equation

$$2\pi i\frac{\partial L(z)}{\partial\tau}+\frac{\partial M(z)}{\partial z}=[L(z),M(z)]. \tag{95}$$

(2) The monodromy of $L(z)$ and $M(z)$ is the same as the monodromy of the Lax pair for the untwisted model,

$$\begin{aligned}L(z+1)&=L(z),\quad M(z+1)=M(z),\\ L(z+\tau)&=e^{2\pi iQ}L(z)e^{-2\pi iQ},\\ M(z+\tau)&=e^{2\pi iQ}(M(z)+2\pi iL(z))e^{-2\pi iQ}-2\pi iP.\end{aligned} \tag{96}$$

The ordinary differential equation defined on the torus E_τ by the matrix $L(z)$ has four regular singular points at $u=0,\omega_1,\omega_2,\omega_3$. The Lax equation and the monodromy of $L(z)$ and $M(z)$ ensure that the local monodromy matrices Γ_α $(a=0,1,2,3)$ and the global monodromy matrices Γ_α and Γ_β are independent of τ.

4. Relation to the Inozemtsev system

The final task is to clarify the relation to the Inozemtsev system. In terms of the orthogonal coordinates $q_j=q\cdot e_j$ and $p_j=p\cdot e_j$ $(j=1,...,l)$, the aforementioned Hamiltonian can be written

$$\begin{aligned}\mathcal{H}=&\frac{1}{2}\sum_{j=1}^{l}p_j^2+\frac{g_m^2}{2}\sum_{\epsilon,\epsilon'=\pm1}\sum_{j\neq k}\wp(\epsilon q_j+\epsilon' q_k)+\frac{g_{l1}^2}{2}\sum_{j=1}^{l}\wp(2q_j)\\ &+\frac{\tilde g_{l2}^2}{2}\sum_{j=1}^{l}\wp^{(2)}(2q_j)+2\tilde g_{s1}^2\sum_{j=1}^{l}\wp(q_j)+2\tilde g_{s2}^2\sum_{j=1}^{l}\wp^{(1/2)}(q_j).\end{aligned} \tag{97}$$

One can rewrite this Hamiltonian using the identities

$$\begin{aligned}\wp(2u)&=\tfrac{1}{4}\wp(u)+\tfrac{1}{4}\wp(u+\omega_1)+\tfrac{1}{4}\wp(u+\omega_2)+\tfrac{1}{4}\wp(u+\omega_3),\\ \wp^{(1/2)}(u)&=\wp(u)+\wp(u+\omega_1)-\wp(\omega_1),\\ \wp^{(2)}(2u)&=\tfrac{1}{4}\wp(u)+\tfrac{1}{4}\wp(u+\omega_3)-\tfrac{1}{4}\wp(\omega_3).\end{aligned} \tag{98}$$

The outcome is, up to a term $h(\tau)$ depending on τ only, the Inozemtsev Hamiltonian

$$\mathcal{H}=\frac{1}{2}\sum_{j=1}^{l}p_j^2+\frac{g_m^2}{2}\sum_{\epsilon,\epsilon'=\pm1}\sum_{j\neq k}\wp(\epsilon q_j+\epsilon' q_k)+\sum_{j=1}^{l}\sum_{a=0}^{3}g_a^2\wp(q_j+\omega_a)+h(\tau). \tag{99}$$

The coupling constants g_a $(a=0,1,2,3)$ are given by

$$g_0^2=\frac{1}{8}(g_{l1}^2+\tilde g_{l2}^2)+2(\tilde g_{s1}^2+\tilde g_{s2}^2),\quad g_1^2=\frac{g_{l1}^2}{8}+2\tilde g_{s2}^2,\quad g_2^2=\frac{g_{l1}^2}{8},\quad g_3^2=\frac{1}{8}(g_{l1}^2+\tilde g_{l2}^2). \tag{100}$$

IV. SPIN GENERALIZATION OF ELLIPTIC CALOGERO–MOSER SYSTEMS

''Spin generalization'' is a generalization of the elliptic Calogero–Moser systems coupled to spin degrees of freedom. Such a spin generalization is characterized by a simple Lie algebra rather than a root system. The (classical) spin variables take values in the dual space $\mathfrak{g}^*$, or a coadjoint orbit therein, of the Lie algebra $\mathfrak{g}$. We shall first examine the $sl(l)$ model as a prototype, then proceed to the models based on a general simple Lie algebra.

A. Spin generalization for $sl(l)$

The $sl(l)$ spin generalization was first introduced by Krichever *et al.*[19] They obtained the spin generalization, just like the spinless case,[5] via the pole dynamics of the matrix KP hierarchy.

1. Hamiltonian formalism

This model is a constrained Hamiltonian system. The Hamiltonian is given by

$$\mathcal{H}=\frac{1}{2}\sum_{j=1}^{l} p_j^2-\frac{1}{2}\sum_{j\neq k}\wp(q_j-q_k)F_{jk}F_{kj}. \tag{101}$$

Here q_j and p_j $(j=1,...,l)$ are the canonical coordinates and momenta of the Calogero–Moser particles, and F_{jk} $(j,k=1,...,l)$ a set of classical $sl(l)$ spin variables, whose Poisson brackets are determined by the Kostant–Kirillov Poisson structure on the dual space of $sl(l)$,

$$\{F_{jk},F_{mn}\}=\delta_{mk}F_{jn}-\delta_{jn}F_{mk}. \tag{102}$$

The equations of motion can be written

$$\begin{gathered}\frac{dq_j}{dt}=p_j,\quad \frac{dp_j}{dt}=\sum_{k\neq j}\wp'(q_j-q_k)F_{jk}F_{kj},\\ \frac{dF_{jk}}{dt}=-\sum_{m\neq j}\wp(q_j-q_m)F_{jm}+\sum_{m\neq k}\wp(q_m-q_k)F_{mk}-\wp(q_j-q_k)(F_{jj}-F_{kk}).\end{gathered} \tag{103}$$

In particular, the diagonal elements F_{jj} of the spin variables are conserved quantities, $dF_{jj}/dt=0$. Although the Hamiltonian does not contain the diagonal elements explicitly, they do appear in the equations of motion. We now put the constraints

$$F_{jj}=0\quad (j=1,...,l). \tag{104}$$

These constraints ensure the integrability. (Actually, the integrability is retained if the constraints are replaced by $F_{jj}=c$, $j=1,...,l$, where c is a constant.)

2. Lax pair in vector representation

The Lax pair of the spinless A_{l-1} model in the vector representation of $sl(l)$ can be readily extended to the spin generalization as follows:

$$\begin{gathered}L(z)=\sum_{j=1}^{l} p_jE_{jj}+\sum_{j\neq k}\sigma(q_j-q_k,z)F_{kj}E_{jk},\\ M(z)=-\sum_{j\neq k}\sigma(q_j-q_k,z)(\rho(q_j-q_k)+\rho(z-q_j+w_k))F_{kj}E_{jk},\end{gathered} \tag{105}$$

where

$$\rho(u)=\frac{\theta_1'(u)}{\theta_1(u)},\quad \sigma(u,z)=\frac{\theta_1(u-z)\,\theta_1'(0)}{\theta_1(z)\,\theta_1(u)}. \tag{106}$$

It is these functions that Felder and Wieczerkowski used in the KZB equation.[17] The function $\rho(u)$ is already familiar to us. The function $\sigma(u,z)$ is also just a disguise of the function $x(u,z)$ that we have used in the preceding sections,

$$\sigma(u,z)=-x(u,z). \tag{107}$$

We however dare to retain the notation of Felder and Wieczerkowski so as to stress the similarity with their work. In these notations, the aforementioned functional identities of $x(u,z)$ and $y(u,z)$ can be rewritten

$$\sigma(u,z)\sigma(v,z)(\rho(v)+\rho(z-v)-\rho(u)-\rho(z-u))=\sigma(u+v,z)(\wp(u)-\wp(v)), \tag{108}$$

$$2\sigma(u,z)\sigma(-u,z)(\rho(u)+\rho(z-u))=-\wp'(u), \tag{109}$$

$$\sigma(u,z)\sigma(-u,z)=\wp(z)-\wp(u). \tag{110}$$

Using these functional identities, one can derive the Lax equation

$$\frac{\partial L(z)}{\partial t}=[L(z),M(z)]. \tag{111}$$

Note that the constraints (104) are always assumed when we consider the Lax equation. Thus the spin generalization, too, is an isospectral integrable system. An involutive set of conserved quantities obtained from the traces $\mathrm{Tr}\,L(z)^k$, $k=2,3,\ldots$. The Hamiltonian itself can be reproduced from the quadratic trace.

The matrix $F=\Sigma_{j\neq k}F_{kj}E_{jk}$, which is the residue of $L(z)$ at $z=0$, stays on a coadjoint orbit of $sl(l)$ as t varies. The phase space of the spin generalization can be thereby restricted to the direct product of the phase space of Calogero–Moser particles and a coadjoint orbit of various dimensions in the dual space of $sl(l)$. The lowest dimensional non-trivial coadjoint orbit can be parametrized by $2l$ variables a_j, b_j $(j=1,\ldots,l)$ as

$$F_{jk}=igb_ja_k \quad (j\neq k), \tag{112}$$

where g is a constant. These reduced spin degrees of freedom, however, can be eliminated by a diagonal gauge transformation of the Lax equations. (This does not mean that a_j and b_j are nondynamical. The elimination procedure is done by partially solving the equations of motion for those variables.) This gauge transformation in turn gives rise to non-zero diagonal elements in $M(z)$, and the outcome is nothing but the Lax equation of the spinless elliptic Calogero–Moser system with coupling constant g. The spinless system is thus embedded in the spin generalization.

3. Isomonodromic system

There is no substantial difference in the construction of an isomonodromic system. The equations of motion are given by

$$\begin{gathered}2\pi i\frac{dq_j}{d\tau}=p_j, \quad 2\pi i\frac{dp_j}{d\tau}=\sum_{k\neq j}\wp'(q_j-q_k)F_{jk}F_{kj}, \\ 2\pi i\frac{dF_{jk}}{d\tau}=\sum_{m\neq j}\wp(q_j-q_m)F_{jm}-\sum_{m\neq k}\wp(q_m-q_k)F_{mk}.\end{gathered} \tag{113}$$

(Terms including F_{jj}'s have been eliminated by the constraints.) The Lax equation, too, can be written in the same form

$$2\pi i\frac{\partial L(z)}{\partial\tau}+\frac{\partial M(z)}{\partial z}=[L(z),M(z)]. \tag{114}$$

Behind this Lax equation is the ‘‘heat equation’’

$$2\pi i\frac{\partial\sigma(u,z)}{\partial\tau}+\frac{\partial^2\sigma(u,z)}{\partial u\,\partial z}=0 \tag{115}$$

satisfied by $\sigma(u,z)$. The final piece of the ring is the monodromy of $L(z)$ and $M(z)$,

$$L(z+1)=L(z), \quad M(z+1)=M(z),$$
$$L(z+\tau)=e^{2\pi iQ}L(z)e^{-2\pi iQ}, \tag{116}$$
$$M(z+\tau)=e^{2\pi iQ}(M(z)+2\pi iL(z))e^{-2\pi iQ}-2\pi iP.$$

As opposed to the root type Lax pairs, the ordinary differential equation

$$\frac{dY(z)}{dz}=L(z)Y(z) \tag{117}$$

on the torus E_τ has only one regular singularity at $z=0$. Thus the local monodromy matrix Γ_0 and the global monodromy matrices Γ_α and Γ_β are all that are invariant under the deformations.

B. Preliminaries for the general simple Lie algebra

Let $\mathfrak{g}$ be a (complex) simple Lie algebra of rank l, $\mathfrak{h}$ a Cartan subalgebra, and Δ the associated root system. The Cartan subalgebra induces a root space decomposition of $\mathfrak{g}$,

$$\mathfrak{g}=\mathfrak{h}\oplus\bigoplus_{\alpha\in\Delta}\mathfrak{g}_\alpha. \tag{118}$$

We choose a basis $\{e_\alpha,h_\mu|\alpha\in\Delta,\mu=1,...,l\}$ of $\mathfrak{g}$ as follows:

(1) h_μ, $\mu=1,...,l$, are an orthonormal basis of $\mathfrak{h}$ with respect to the Killing form $B:\mathfrak{h}\times\mathfrak{h}\to\mathbb{C}$, i.e.,

$$B(h_\mu,h_\nu)=\delta_{\mu\nu}. \tag{119}$$

(2) The Killing form induces an isomorphism $\mathfrak{h}^*=\mathrm{Hom}(\mathfrak{h},\mathbb{C})\simeq\mathfrak{h}$, which determines an element h_α for each $\alpha\in\mathfrak{h}^*$. In terms of the basis h_μ of $\mathfrak{h}$, this map can be written explicitly,

$$\alpha\mapsto h_\alpha=\sum_{\mu=1}^{l}\alpha(h_\mu)h_\mu, \tag{120}$$

The root subspace $\mathfrak{g}_\alpha$ is one dimensional. e_α is a basis of $\mathfrak{g}_\alpha$ such that

$$[e_\alpha,e_{-\alpha}]=h_\alpha. \tag{121}$$

This choice of e_α amounts to the normalization

$$B(e_\alpha,\epsilon_{-\alpha})=1. \tag{122}$$

The Lie brackets of the basis elements other than $[e_\alpha,e_{-\alpha}]$ now takes the form

$$[e_\alpha,e_\beta]=N_{\alpha,\beta}e_{\alpha+\beta} \quad (\alpha+\beta\neq 0),$$
$$[h_\mu,e_\alpha]=\alpha(h_\mu)e_\alpha, \tag{123}$$
$$[h_\mu,h_\nu]=0.$$

The structure constants $N_{\alpha,\beta}$ are antisymmetric with respect to the indices, and vanish if $\alpha+\beta\notin\Delta$. The following general relation among the structure constants will be used in the course of the proof of a Lax equation.

Lemma 1:

$$N_{-\beta,\alpha+\beta}=N_{-\alpha,-\beta}=N_{\alpha+\beta,-\alpha}. \tag{124}$$

Proof: If $\alpha=\beta$, this relation is trivially satisfied, because all the structure constants vanish. Let us consider the case where $\alpha\neq\beta$. By the Jacobi identity, we have

$$[e_{\alpha+\beta},[e_{-\alpha},e_{-\beta}]]=[[e_{\alpha+\beta},e_{-\alpha}],e_{-\beta}]+[e_{-\alpha},[e_{\alpha+\beta},e_{-\beta}]].$$

This implies the identity

$$N_{-\alpha,-\beta}h_{\alpha+\beta}=N_{\alpha+\beta,-\alpha}h_{\beta}-N_{\alpha+\beta,-\beta}h_{\alpha},$$

which, by the relation $h_{\alpha+\beta}=h_\alpha+h_\beta$, can be rewritten

$$(N_{-\alpha,-\beta}+N_{\alpha+\beta,-\beta})h_\alpha+(N_{-\alpha,-\beta}-N_{\alpha+\beta,-\alpha})h_\beta=0.$$

Since we have assumed that $\alpha\neq\beta$, h_α and h_β are linearly independent, so that the two coefficients in this linear retion are equal to zero. Q.E.D.

We can now specify the classical spin variables for a general simple Lie algebra. Those spin variables, by definition, are coordinates of the dual space $\mathfrak{g}^*=\mathrm{Hom}(\mathfrak{g},\mathbb{C})$. Let F_α and G_μ be the coordinates dual to the above basis e_α and h_μ. In other words, they are the coefficients of e_α and h_μ in the linear combination

$$\sum_{\alpha\in\Delta}F_{-\alpha}e_\alpha+\sum_{\mu=1}^{l}G_\mu h_\mu \tag{125}$$

that realizes the isomorphism $\mathfrak{g}^*\simeq\mathfrak{g}$ induced by the Killing form. The Kostant–Kirillov Poisson structure on $\mathfrak{g}^*$ determine the Poisson brackets of these spin variables, which take the same form as the Lie brackets of the Lie algebra basis,

$$\begin{gathered}\{F_\alpha,F_{-\alpha}\}=G_\alpha=\sum_{\mu=1}^{l}\alpha(h_\mu)G_\mu,\\ \{F_\alpha,F_\beta\}=N_{\alpha,\beta}F_{\alpha+\beta}\quad(\alpha+\beta\neq0),\\ \{G_\mu,F_\alpha\}=\alpha(h_\mu)F_\alpha,\\ \{G_\mu,G_\nu\}=0.\end{gathered}\tag{126}$$

C. Spin generalization for the general simple Lie algebra

1. Hamiltonian formalism

The spin generalization based on $\mathfrak{g}$, too, is a constrained Hamiltonian system defined on $\mathfrak{h}\times\mathfrak{h}\times\mathfrak{g}^*$ by the Hamiltonian

$$\mathcal{H}=\frac{1}{2}B(p,p)-\frac{1}{2}\sum_{\alpha\in\Delta}\wp(\alpha(q))F_{-\alpha}F_\alpha \tag{127}$$

and the constraints

$$G_\mu=0\quad(\mu=1,...,l). \tag{128}$$

Here q and p are understood to take values in $\mathfrak{h}$. $B(p,q)$ and $\alpha(q)$ amount to $p\cdot p$ and $\alpha\cdot q$ in the models based on root systems. Let us use the same ''dot notation'' for the Killing form $\mathfrak{h}\times\mathfrak{h}\to\mathbb{C}$ and the pairing $\mathfrak{h}^*\times\mathfrak{h}\to\mathbb{C}$. The Hamiltonian then takes a more familiar form

$$\mathcal{H}=\frac{1}{2}p\cdot p-\frac{1}{2}\sum_{\alpha\in\Delta}\wp(\alpha\cdot q)F_{-\alpha}F_\alpha. \tag{129}$$

The equations of motion can be readily written down in the language of the coordinates $q_\mu = q\cdot h_\mu$ and momenta $p_\mu = p\cdot h_\mu$ of Calogero–Moser particles and the spin variables F_α and G_μ on $\mathfrak{g}^*$,

$$\frac{dq_\mu}{dt} = p_\mu,$$
$$\frac{dp_\mu}{dt} = -\frac{1}{2}\sum_{\alpha\in\Delta} \alpha\cdot h_\mu \wp'(\alpha\cdot q)F_{-\alpha}F_\alpha,$$
$$\frac{dF_\alpha}{dt} = -\sum_{\beta\in\Delta,\alpha-\beta\in\Delta} \wp(\beta\cdot q)F_{\alpha-\beta}F_\beta N_{\alpha,-\beta} - \wp(\alpha\cdot q)G_\alpha F_\alpha, \tag{130}$$
$$\frac{dG_\mu}{dt} = 0.$$

In particular, the diagonal elements G_μ of the spin variables are conserved quantities. One can thereby safely put the aforementioned constraints.

2. Lax pair

The integrability of our spin generalization is ensured by the existence of a Lax pair as follows:

Proposition 6: Let V be any finite dimensional representation of $\mathfrak{g}$, *and* E_α *and* H_μ *the endomorphisms on V that represent* e_α *and* h_μ. *Then the endomorphisms*

$$L(z) = P + \sum_{\alpha\in\Delta} \sigma(\alpha\cdot q,z)F_{-\alpha}E_\alpha \quad P = \sum_{\mu=1}^{l} p_\mu H_\mu,$$
$$M(z) = -\sum_{\alpha\in\Delta} \sigma(\alpha\cdot q,z)(\rho(\alpha\cdot q)+\rho(z-\alpha\cdot q))F_{-\alpha}E_\alpha \tag{131}$$

on V satisfy the Lax equation

$$\frac{\partial L(z)}{\partial t} = [L(z),M(z)]. \tag{132}$$

Proof: Using the equations of motion and the constraints, one can express the t-derivative of the L-matrix as

$$\frac{\partial L(z)}{\partial t} = \mathrm{I}+\mathrm{II}+\mathrm{III}, \tag{133}$$

where

$$\mathrm{I} = \sum_{\mu=1}^{l} \frac{dp_\mu}{dt} H_\mu = -\frac{1}{2}\sum_{\alpha\in\Delta} \wp'(\alpha\cdot q)F_{-\alpha}F_\alpha H_\alpha,$$
$$\mathrm{II} = \sum_{\alpha\in\Delta}\sum_{\mu=1}^{l} \frac{d\alpha\cdot q}{dt} \left.\frac{\partial\sigma(u,z)}{\partial u}\right|_{u=\alpha\cdot q} F_{-\alpha}E_\alpha$$
$$= -\sum_{\alpha\in\Delta} \alpha\cdot\alpha\sigma(\alpha\cdot q,z)(\rho(\alpha\cdot q)+\rho(z-\alpha\cdot q))F_{-\alpha}E_\alpha,$$

$$\mathrm{III}=\sum_{\alpha\in\Delta}\sigma(\alpha\cdot q,z)\frac{dF_{-\alpha}}{dt}E_{\alpha}=-\sum_{\alpha,\beta\in\Delta,\alpha+\beta\neq 0}\sigma(\alpha\cdot q,z)\wp(\beta\cdot q)F_{-\alpha-\beta}F_{\beta}N_{-\alpha,-\beta}E_{\alpha}.$$

Similarly, the commutator of the Lax pair can be written

$$[L(z),M(z)]=\mathrm{IV}+\mathrm{V}+\mathrm{VI}, \tag{134}$$

where VI stands for terms from the commutator $[P,M(z)]$,

$$\mathrm{IV}=-\sum_{\alpha\in\Delta}\sigma(\alpha\cdot q,z)(\rho(\alpha\cdot q)+\rho(z-\alpha\cdot q))F_{-\alpha}[P,E_{\alpha}]$$

$$=-\sum_{\alpha\in\Delta}\sigma(\alpha\cdot q,z)(\rho(\alpha\cdot q)+\rho(z-\alpha\cdot q))\alpha\cdot pF_{-\alpha}E_{\alpha},$$

and V+VI are the the other terms grouped into the Cartan part (V) and the off-Cartan part (VI),

$$\mathrm{V}=-\sum_{\alpha\in\Delta}\sigma(-\alpha\cdot q,z)\sigma(\alpha\cdot q,z)(\rho(\alpha\cdot q)+\rho(z-\alpha\cdot q))F_{-\alpha}F_{\alpha}[E_{\alpha},E_{-\alpha}]$$

$$=-\sum_{\alpha\in\Delta}\sigma(-\alpha\cdot q,z)\sigma(\alpha\cdot q,z)(\rho(\alpha\cdot q)+\rho(z-\alpha\cdot q))F_{-\alpha}F_{\alpha}H_{\alpha},$$

$$\mathrm{VI}=-\sum_{\alpha,\beta\in\Delta,\alpha+\beta\neq 0}\sigma(\alpha\cdot q,q)\sigma(\beta\cdot q,z)(\rho(\beta\cdot q)+\rho(z-\beta\cdot q))F_{-\alpha}F_{\alpha}[E_{\alpha},E_{\beta}]$$

$$=-\sum_{\alpha,\beta\in\Delta,\alpha+\beta\neq 0}\sigma(\alpha\cdot q,z)\sigma(\beta\cdot q,z)(\rho(\beta\cdot q)+\rho(z-\beta\cdot q))F_{-\alpha}F_{-\beta}N_{\alpha,\beta}E_{\alpha+\beta}.$$

It is obvious that IV=II. Using (109), we can readily see that V=I. Thus it remains to prove that VI=III. This is achieved as follows:

$$\mathrm{VI}=-\frac{1}{2}\sum_{\alpha,\beta\in\Delta,\alpha+\beta\neq 0}\sigma(\alpha\cdot q,z)\sigma(\beta\cdot q,z)(\rho(\beta\cdot q)+\rho(z-\beta\cdot q)-\rho(z-\alpha\cdot q)$$
$$-\rho(\alpha\cdot q))F_{-\alpha}F_{-\beta}N_{\alpha,\beta}E_{\alpha+\beta}\ \text{[symmetrized with respect to } \alpha \text{ and } \beta\text{]},$$
$$=-\frac{1}{2}\sum_{\alpha,\beta\in\Delta,\alpha+\beta\neq 0}\sigma((\alpha+\beta)\cdot q,z)(\wp(\alpha\cdot q)-\wp(\beta\cdot q))F_{-\alpha}F_{-\beta}N_{\alpha,\beta}E_{\alpha+\beta}\cdot\text{[(108) is used]},$$
$$=\sum_{\alpha,\beta\in\Delta,\alpha+\beta\neq 0}\sigma((\alpha+\beta)\cdot q,z)\wp(\beta\cdot q)F_{-\alpha}F_{-\beta}N_{\alpha,\beta}E_{\alpha+\beta}$$
$$\text{[asymmetrized with respect to } \alpha \text{ and } \beta\text{]},$$
$$=\sum_{\alpha,\beta\in\Delta,\alpha+\beta\neq 0}\sigma(\alpha\cdot q,z)\wp(\beta\cdot q)F_{-\alpha,-\beta}F_{\beta}N_{\alpha+\beta-\beta}E_{\alpha}.$$
$$\text{[substituting } \beta\rightarrow-\beta \text{ and } \alpha\rightarrow\alpha+\beta\text{]}.$$

Finally using the identity $N_{\alpha+\beta,-\beta}=-N_{-\alpha,-\beta}$, cf. (124), we find that the last sum is equal to III. Q.E.D.

Note that the above proof persists to be meaningful if E_{α} and H_{μ} are replaced by the Lie algebra elements e_{α} and h_{μ}. In other words, the Lax equation actually lives in the Lie algebra $\mathfrak{g}$ itself rather than its representations. This resembles the case of the Toda systems.

3. Isomonodromic system

The passage to an isomonodromic analog is straightforward. Replacing $d/dt \to 2\pi i d/d\tau$, one obtains the nonautonomous system,

$$2\pi i \frac{dq_\mu}{d\tau} = p_\mu ,$$

$$2\pi i \frac{dp_\mu}{d\tau} = -\frac{1}{2} \sum_{\alpha \in \Delta} \alpha \cdot h_\mu \wp'(\alpha \cdot q) F_{-\alpha} F_\alpha , \tag{135}$$

$$2\pi i \frac{dF_\alpha}{d\tau} = -\sum_{\beta \in \Delta, \alpha - \beta \in \Delta} \wp(\beta \cdot q) F_{\alpha - \beta} F_\beta N_{\alpha, -\beta} .$$

(Terms including G_μ's have been eliminated by the constraints.) These equations can be converted to the Lax equation

$$2\pi i \frac{\partial L(z)}{\partial \tau} + \frac{\partial M(z)}{\partial z} = [L(z), M(z)]. \tag{136}$$

The monodromy of $L(z)$ and $M(z)$, too, takes the same form,

$$L(z+1) - L(z), \quad M(z+1) = M(z),$$

$$L(z+\tau) = e^{2\pi i Q} L(z) e^{-2\pi i Q}, \tag{137}$$

$$M(z+\tau) = e^{2\pi i Q}(M(z) + 2\pi i L(z)) e^{-2\pi i Q} - 2\pi i P,$$

where $Q = \Sigma_{\mu=1}^{l} q_\mu H_\mu$. The Lax equation implies that the monodromy data of the ordinary differential equation

$$\frac{dY(z)}{dz} = L(z) Y(z) \tag{138}$$

on the torus E_τ is invariant as τ varies. $Y(z)$ now take values in the representation space V; the monodromy around a singular point or of a cycle of E_τ is represented by a linear transformation on V. The ordinary differential equation has a regular singularity at $z=0$ only. The local monodromy around this singular point is a linear transformation $\Gamma_0 \in GL(V)$. Similarly, the global monodromy along the α and β cycles give Γ_α, $\Gamma_\beta \in GL(V)$. These linear transformations Γ_0, Γ_α, and Γ_β are the monodromy data that are left invariant.

V. CONCLUSION

We have thus demonstrated that various models of the elliptic Calogero–Moser systems are accompanied with an isomonodromic partner. A technical clue is the choice of fundamental functions $x(u,z)$, $y(u,z)$, etc. in the Lax pair $L(z)$ and $M(z)$. For $L(z)$ and $M(z)$ to give an isomonodromic Lax pair, these functions are required to satisfy a kind of ''heat equation'' besides the functional equations. We have illustrated the construction of the isomonodromic Lax pair for several typical cases—the Lax pair of the A_{l-1} mode in the vector representation, the root type Lax pair for various untwisted and twisted models, and the Lax pair of the spin generalizations.

The most interesting case in the context of Manin's equation is the root type Lax pair for the extended twisted BC_l model (or, equivalently, the Inozemtsev system). The root type Lax pair based on short roots of the BC_l root system consists of $2l \times 2l$ matrices.

The construction of a Lax pair, however, is merely the first step towards a full understanding of Manin's equation and its possible generalizations. The next issue is to elucidate the meaning of the affine Weyl group symmetries, various special solutions, etc. in this framework. Recent works by Noumi and Yamada,[20] Deift, Its, Kapaev, and Zhou[21] and Kitaev and Korotkin[22] are very suggestive in this respect.

The spin generalization that we have discussed is a special case of a more general multispin system, i.e., the elliptic Calogero–Moser systems coupled to ''Gaudin spins'' sitting at the punctures of a punctured torus.[9,10] This is the Hitchin system on a punctured torus; we have considered the case with only one puncture located at $z=0$. It is rather straightforward, though more complicated, to generalize our Lax pair to the multispin generalization. This gives a generalization, to other simple Lie groups, of the SU(2) isomonodromic system of Korotkin and Samtleben.[23] The dynamical r-matrix in the work of Felder and Wieczerkowski[17] plays a central role here. We shall report this result elsewhere.

ACKNOWLEDGMENTS

I am indebted to Ryu Sasaki for a number of crucial ideas on the elliptic Calogero–Moser systems. I am also grateful to Shingo Kawai and Kazuo Okamoto for valuable comments on isomonodromic systems on the torus. Finally, I would like to thank Koji Hasegawa, Gen Kuroki, Takashi Takebe, and Yasuhiko Yamada for discussions on many aspects of isospectral and isomonodromic systems. This work is partly supported by the Grant-in-Aid for Scientific Research (No. 10640165) from the Ministry of Education, Science, and Culture.

APPENDIX A: PROOF OF FUNCTIONAL IDENTITIES AND HEAT EQUATION FOR UNTWISTED MODELS

1. Proof of (24)

Let $f(u,v,z)$ denote the difference of both hand sides of (24),

$$f(u,v,z)=x(u,z)y(v,z)-y(u,z)x(v,z)-x(u+v,z)(\wp(u)-\wp(v)). \tag{A1}$$

This function turns out to have the following analytical properties:

(1) $f(u,v,z)$ has the same quasiperiodicity as $x(u,z)$ on the u plane, i.e.,

$$f(u+1,v,z)=f(u,v,z),\quad f(u+\tau,v,z)=e^{2\pi i z}f(u,v,z). \tag{A2}$$

(2) $f(u,v,z)$ is an entire function on the u plane.

The first property is obvious from the quasiperiodicity of $x(u,z)$ and the periodicity of $\wp(u)$. Furthermore, poles of $f(u,v,z)$ can appear only at the lattice points $u=m+n\tau\,(m,n\in\mathbb{Z})$ on the u plane. Therefore, in order to verify the second property, we have only to show that $f(u,v,z)$ is nonsingular at these points. Actually, because of the quasiperiodicity, it is sufficient to consider the point $u=0$ only. As $u\to 0$, the singular terms $x(u,z)$, $y(u,z)$, and $\wp(u)$ in $f(u,v,z)$ behave as

$$x(u,z)=\frac{1}{u}+O(1),\quad y(u,z)=-\frac{1}{u^2}+O(1),\quad \wp(u)=\frac{1}{u^2}+O(u^2) \tag{A3}$$

so that

$$\begin{aligned} f(u,v,z)&=\left(\frac{1}{u}+O(1)\right)y(v,z)-\left(-\frac{1}{u^2}+O(1)\right)x(v,z)\\ &\quad-(x(u,z)+y(u,z)u+O(u^2))\left(\frac{1}{u^2}-\wp(v)+O(u^2)\right)\\ &=O(1). \end{aligned} \tag{A4}$$

We can thus verify the above two properties of $f(u,v,x)$.

Actually, any function with these two properties should vanish identically. This can be seen in several different ways. The shortest will be to resort to algebraic geometry of line bundles on the torus E_τ. A more elementary proof is to consider the quotient $f(u,v,z)/x(u,z)$. This quotient is a doubly-periodic meromorphic function, and all possible poles are located at the lattice points $u=m+n\tau\,(m,n\in\mathbb{Z})$, and at most of first order. In other words, $f(u,v,z)/x(u,z)$ is a meromorphic function on the torus with the only possible pole at $u=0$, but the order of pole cannot be greater than one. Such a function has to be a constant. On the other hand, because of the pole of $x(u,z)$ at $u=0$, $f(u,v,z)/x(u,z)$ has a zero at $u=0$. Therefore the constant should be equal to zero.

2. Proof of (25) and (26)

(25) can be readily derived from (24) by letting $v\to -u$. Let us consider (26). By (25),

$$\frac{\partial}{\partial u}(x(u,z)x(-u,z)) = -x(u,z)y(-u,z)+y(u,z)x(-u,z) = -\wp'(u). \tag{A5}$$

Consequently,

$$x(u,z)x(-u,z) = -\wp(u)+(\text{independent of } u). \tag{A6}$$

Since $x(u,z)=-x(z,u)=-x(-u,-z)$, the left-hand side of the last relation is in fact an anti-symmetric function of u and z. Therefore,

$$x(u,z)x(-u,z) = \wp(z)-\wp(u)+\text{const}. \tag{A7}$$

Now consider the limit as $u\to z$. Both $x(u,z)x(-u,z)$ and $\wp(z)-\wp(u)$ tend to zero in this limit. Thus the constant on the right-hand side has to be zero.

3. Proof of (34)

Let us rewrite the both hand sides of (34) into a more accessible form. Differentiating $x(u,z)$ by τ gives

$$\frac{\partial x(u,z)}{\partial \tau} = x(u,z)\frac{\partial}{\partial \tau}(\log\theta_1(z-u)+\log\theta_1'(0)-\log\theta_1(z)-\log\theta_1(u)). \tag{A8}$$

By the heat equation (40) of the Jacobi theta function,

$$4\pi i\frac{\partial}{\partial\tau}\theta_1(u) = \frac{\theta_1''(u)}{\theta_1(u)} = \frac{\partial}{\partial u}\left(\frac{\theta_1'(u)}{\theta_1(u)}\right)+\left(\frac{\theta_1'(u)}{\theta_1(u)}\right)^2 = \rho'(u)+\rho(u)^2. \tag{A9}$$

Letting $u\to 0$ and recalling the singular behavior of $\rho(u)$ at $u=0$, we obtain

$$4\pi i\frac{\partial}{\partial\tau}\log\theta_1'(0) = \lim_{u\to 0}(\rho'(u)+\rho(u)) = \frac{\theta_1'''(0)}{\theta_1'(0)}. \tag{A10}$$

Plugging these formulas into the above expression of $\partial x(u,z)/\tau$ gives

$$4\pi i\frac{\partial x(u,z)}{\partial\tau} = x(u,z)f(u,z), \tag{A11}$$

where

$$f(u,z)=\rho'(z-u)+\rho(z-u)^2+\frac{\theta_1'''(0)}{\theta_1'(0)}-\rho'(z)-\rho(z)^2-\rho'(u)-\rho(u)^2. \tag{A12}$$

On the other hand, we have

$$\frac{\partial x(u,z)}{\partial u\partial z}=-\frac{\partial}{\partial z}(x(u,z)(\rho(u)+\rho(z-u)))=-x(u,z)g(u,z), \tag{A13}$$

where

$$g(u,z)=(\rho(z-u)-\rho(z))(\rho(u)+\rho(z-u))+\rho'(z-u). \tag{A14}$$

The goal is to verify that $f(u,z)=2g(u,z)$. It is sufficient to prove the following two properties of $f(u,z)-2g(u,z)$, because such a function has to be identically zero.

(1) $f(u,z)-2g(u,z)$ is a doubly-periodic function on the u plane with primitive periods 1 and τ.
(2) $f(u,z)-2g(u,z)$ is an entire function, and has a zero at $u=0$.

The first property is obvious if one notices the following quasiperiodicity of $f(u,z)$ and $g(u,z)$:

$$\begin{aligned} f(u+1,z)&=f(u,z), \quad f(u+\tau,z)=f(u,z)+4\pi i(\rho(u)+\rho(z-u)),\\ g(u+1,z)&=g(u,z), \quad g(u+\tau,z)=g(u,z)+2\pi i(\rho(u)+\rho(z-u)). \end{aligned} \tag{A15}$$

Let us check the second property. Possible poles of $f(u,z)$ and $g(u,z)$ are located at the two points $u=0$ and $u=z$ of the fundamental domain of the period lattice $\mathbb{Z}+\tau\mathbb{Z}$. Again recalling the singular behavior of $\rho(u)$ at $u=0$, one can confirm by straightforward calculations that

$$f(u,z)=O(u), \quad g(u,z)=O(u) \quad (u\to 0). \tag{A16}$$

Thus $f(u,z)-2g(u,z)$ turns out to be nonsingular and have a zero at $u=0$. Similarly, one can see that $f(u,z)-2g(u,z)$ is nonsingular at $u=z$.

APPENDIX B: VERIFICATION OF THE LAX PAIR FOR EXTENDED TWISTED BC_l MODEL

To prove the Lax equation, it is sufficient to derive the following three equations:

$$\frac{\partial X_a(z)}{\partial t}=[P,X_a(z)] \quad (a=1,2,3), \tag{B1}$$

$$\frac{dp\cdot\mu}{dt}=[X_1(z)+X_2(z)+X_3(z),Y_1(z)+Y_2(z)+Y_3(z)]_{\mu\mu}, \tag{B2}$$

$$0=[X_1(z)+X_2(z)+X_3(z),D+Y_1(z)+Y_2(z)+Y_3(z)]_{\mu\nu} \quad (\mu\neq\nu). \tag{B3}$$

μ and ν run over the set Δ_s of short roots.

The proof of (B1) is quite easy. Let us consider the case of $a=1$. The t-derivative of $X_1(z)$ can be written

$$\frac{\partial X_1(z)}{\partial t}=ig_m\sum_{\alpha\in\Delta_m}\alpha\cdot py(\alpha\cdot q,z)E(\alpha). \tag{B4}$$

Using the commutation relation $[P,E(\alpha)]=\alpha\cdot pE(\alpha)$, one can readily see that the right-hand side is equal to $[P,X_1(z)]$. The other two in (B1) can be similarly derived.

The rest of this Appendix is devoted to the other two equations (B2) and (B3).

1. Proof of (B2)

We calculate the diagonal elements

$$[X_a(z),Y_b(z)]_{\mu\mu}=\sum_{\nu\in\Delta_s}(X_{a,\mu\nu}(z)Y_{b,\nu\mu}(z)-Y_{b,\mu\nu}(z)X_{a,\nu\mu}(z)) \tag{B5}$$

of the nine commutators one-by-one.

a. Vanishing terms

Some part of the matrix elements of $X_a(z)$ and $Y_b(z)$ turn out to vanish by the nature of the BC_l root system,

$$X_{1,\mu,-\mu}(z)=Y_{1,\mu,-\mu}(z)=0, \tag{B6}$$

$$X_{2,\mu\nu}(z)=Y_{2,\mu\nu}(z)=0 \quad (\mu\neq-\nu), \tag{B7}$$

$$X_{3,\mu\nu}(z)=Y_{3,\mu\nu}(z)=0 \quad (\mu\neq-\nu). \tag{B8}$$

The first relation is due to the fact that $\mu-(-\mu)=2\mu$ can never be a middle root. The second and third relations are obvious if one notices that $\mu-\nu$ is a long root (or, equivalently, twice a short root) if and only if $\mu=-\nu$.

In particular,

$$[X_1(z),Y_2(z)]_{\mu\mu}=[X_1(z),Y_3(z)]_{\mu\mu}=[X_2(z),Y_1(z)]_{\mu\mu}=[X_3(z),Y_1(z)]_{\mu\mu}=0. \tag{B9}$$

b. Calculation of $[X_1(z),Y_1(z)]_{\mu\mu}$

By definition,

$$[X_1(z),Y_1(z)]_{\mu\mu}=-g_m^2\sum_{\nu\in\Delta_s,\mu-\nu\in\Delta_m}(x((\mu-\nu)\cdot q,z)y((\nu-\mu)\cdot q,z)$$
$$-y((\mu-\nu)\cdot q,z)x((\nu-\mu)\cdot q,z)). \tag{B10}$$

We rewrite this sum to a sum over the middle root $\alpha=\mu-\nu$. Since the middle roots α of this form are characterized by the condition that $\alpha\cdot\mu=1$, the right-hand side can be rewritten

$$-g_m^2\sum_{\alpha\in\Delta_m,\alpha\cdot\mu=1}(x(\alpha\cdot q,z)y(-\alpha\cdot q,z)-y(\alpha\cdot q,z)x(-\alpha\cdot q,z)).$$

Actually, the possible values of $\alpha\cdot\mu$ are limited to 0 and ±1 only. Therefore this sum is equal to

$$-\frac{g_m^2}{2}\sum_{\alpha\in\Delta_m}\alpha\cdot\mu(x(\alpha\cdot q,z)y(-\alpha\cdot q,z)-y(\alpha\cdot q,z)x(-\alpha\cdot q,z)).$$

(The factor 1/2 compensates the contributions from $\alpha\cdot\mu=1$ and $\alpha\cdot\mu=-1$.) Noting that $\alpha\cdot\mu=\{p\cdot\mu,\alpha\cdot q\}$, we can express $[X_1(z),Y_1(z)]$ as a Poisson bracket of the form

$$[X_1(z),Y_1(z)]_{\mu\mu}=\{p\cdot\mu,V_{11}\}, \tag{B11}$$

where

$$V_{11}=\frac{g_m^2}{2}\sum_{\alpha\in\Delta_m}x(\alpha\cdot q,z)x(-\alpha\cdot q,z). \tag{B12}$$

c. Contributions of other commutators

By (B7) and (B8), the diagonal elements of the other commutators are a sum of just two terms,

$$[X_a(z),Y_b(z)]_{\mu\mu}=X_{a,\mu,-\mu}Y_{b,-\mu,\mu}-Y_{b,\mu,-\mu}X_{a,-\mu,\mu}. \tag{B13}$$

Let us consider the case of $a=2$ and $b=2$ in some detail. By definition,

$$\begin{aligned}[X_2(z),Y_2(z)]_{\mu\mu}=&-(g_{l1}x(2\mu\cdot q,z)+g_{l2}x^{(2)}(2\mu\cdot q,z))(g_{l1}y(-2\mu\cdot q,z)+g_{l2}y^{(2)}(-2\mu\cdot q,z))\\ &+(g_{l1}y(2\mu\cdot q,z)+g_{l2}y^{(2)}(2\mu\cdot q,z))(g_{l1}x(-2\mu\cdot q,z)+g_{l2}x^{(2)}(-2\mu\cdot q,z)).\end{aligned}$$

Since $\alpha=2\mu$ is a long root, and long roots with nonvanishing inner product with μ are 2μ and -2μ only, the right-hand side can be rewritten

$$\begin{aligned}&-\frac{1}{4}\sum_{\alpha\in\Delta_l}\alpha\cdot\mu(g_{l1}x(\alpha\cdot q,z)+g_{l2}x^{(2)}(\alpha\cdot q,z))(g_{l1}y(-\alpha\cdot q,z)+g_{l2}y^{(2)}(-\alpha\cdot q,z))\\ &+\frac{1}{4}\sum_{\alpha\in\Delta_l}\alpha\cdot\mu(g_{l1}y(\alpha\cdot q,z)+g_{l2}y^{(2)}(\alpha\cdot q,z))(g_{l1}x(-\alpha\cdot q,z)+g_{l2}x^{(2)}(-\alpha\cdot q,z)).\end{aligned}$$

(The factor 1/4 compensates the contributions from $\alpha\cdot\mu=2$ and $\alpha\cdot\mu=-2$.) We can again cast this into a Poisson bracket,

$$[X_2(z),Y_2(z)]_{\mu\mu}=\{p\cdot\mu,V_{22}\}, \tag{B14}$$

where

$$V_{22}=\frac{1}{4}\sum_{\alpha\in\Delta_l}(g_{l1}x(\alpha\cdot q,z)+g_{l2}x^{(2)}(\alpha\cdot q,z))(g_{l1}x(-\alpha\cdot q,z)+g_{l2}x^{(2)}(-\alpha\cdot q,z)). \tag{B15}$$

Similarly, one can obtain

$$\begin{gathered}[X_2(z),Y_3(z)]_{\mu\mu}=\{p\cdot\mu,V_{23}\},\quad [X_3(z),Y_2(z)]_{\mu\mu}=\{p\cdot\mu,V_{32}\},\\ [X_3(z),Y_3(z)]_{\mu\mu}=\{p\cdot\mu,V_{33}\},\end{gathered} \tag{B16}$$

where

$$\begin{aligned}V_{23}&=\frac{1}{2}\sum_{\alpha\in\Delta_s}(g_{l1}x(2\alpha\cdot q,z)+g_{l2}x^{(2)}(2\alpha\cdot q,z))(g_{s1}x(-\alpha\cdot q,2z)+g_{s2}x^{(1/2)}(-\alpha\cdot q,2z)),\\ V_{32}&=\frac{1}{2}\sum_{\alpha\in\Delta_s}(g_{s1}x(\alpha\cdot q,2z)+g_{s2}x^{(1/2)}(\alpha\cdot q,2z))(g_{l1}x(-2\alpha\cdot q,z)+g_{l2}x^{(2)}(-2\alpha\cdot q,z)),\\ V_{33}&=\sum_{\alpha\in\Delta_s}(g_{s1}x(\alpha\cdot q,2z)+g_{s2}x^{(1/2)}(\alpha\cdot q,2z))(g_{s1}x(-\alpha\cdot q,2z)+g_{s2}x^{(1/2)}(-\alpha\cdot q,2z)).\end{aligned} \tag{B17}$$

Collecting the results of these calculations, we find that the right-hand side of (B2) takes the form of the Poisson bracket $\{p\cdot\mu,V\}$, where

$$V=V_{11}+V_{22}+V_{23}+V_{32}+V_{33}. \tag{B18}$$

d. Writing V in terms of $\wp$ functions

The final step is to rewrite V in terms of the Weierstrass $\wp$ functions. For V_{11}, this can be done by use of (26). The other parts are due to the following functional identities:

$$x^{(1/2)}(u,z)x^{(1/2)}(-u,z) = -\wp^{(1/2)}(u) + \wp^{(1/2)}\left(\frac{z}{2}\right), \tag{B19}$$

$$x^{(2)}(u,z)x^{(2)}(-u,z) = -\wp^{(2)}(u) + \wp^{(2)}(2z), \tag{B20}$$

$$x(u,2z)x^{(1/2)}(-u,2z) + x^{(1/2)}(u,2z)x(-u,2z) = -2\wp(u) + \text{const.}, \tag{B21}$$

$$x(u,2z)x(-2u,z) + x(2u,z)x(-u,2z) = -\wp(u) + \text{const.}, \tag{B22}$$

$$x(u,2z)x^{(2)}(-2\mu,z) + x^{(2)}(2u,z)x(-u,2z) = -\wp(u) + \text{const.}, \tag{B23}$$

$$x^{(1/2)}(u,2z)x(-2u,z) + x(2u,z)x^{(1/2)}(-u,2z) = -\wp^{(1/2)}(u) + \text{const.}, \tag{B24}$$

$$x^{(1/2)}(u,2z)x^{(2)}(-2u,z) + x^{(2)}(2u,z)x^{(1/2)}(-u,2z) = -\wp(u) + \text{const.}, \tag{B25}$$

$$x(u,z)x^{(2)}(-u,z) + x^{(2)}(u,z)x(-u,z) = -2\wp^{(2)}(u) + \text{const.} \tag{B26}$$

The first two are substantially the same as (26) except that the variables and the primitive periods are rescaled. ''const.'' in the other identities stand for terms that are independent of u, thereby negligible in the Poisson bracket with $p\cdot\mu$; remember that they are not absolute constants, but functions of z and τ. We shall prove these identities in Appendix C. Using these functional identities, one can see that V is equal to the potential part of the Hamiltonian $\mathcal{H}$, up to nondynamical terms independent of p and q.

To summarize, we have shown that the sum of the (μ,μ) elements of the nine commutators coincides with the Poisson bracket $\{p\cdot\mu,V\}$, which is equal to $dp\cdot\mu/dt$ by the equations of motion of the model.

2. Proof of (B3)

The proof can be separated into the cases where $\nu=-\mu$ and $\nu\neq\pm\mu$.

a. $\nu=-\mu$

The vanishing of the $(\mu,-\mu)$ elements of the commutators other than $[X_a(z),D]$ $(a=1,2,3)$ and $[X_1(z),D]$ is immediate from (B7) and (B8). $[X_a(z),D]_{\mu,-\mu}$ vanishes because of the symmetry $D_{-\mu}=D_\mu$. As for $[X_1(z),Y_1(z)]_{\mu,-\mu}$, we have

$$[X_1(z),Y_1(z)]_{\mu,-\mu} = -g_m^2 \sum_{\nu\in\Delta_s\setminus\{\pm\mu\}} x((\mu-\nu)\cdot q,z)y((\nu+\mu)\cdot q,z) + g_m^2 \sum_{\nu\in\Delta_s\setminus\{\pm\mu\}} y((\mu-\nu)\cdot q,z)x((\nu+\mu)\cdot q,z). \tag{B27}$$

By substituting $\nu\to-\nu$, the second sum on the right-hand side turns out to be identical to the first sum. The two sums thus cancel with each other.

b. $\nu\neq\pm\mu$

The following can be readily seen by using (B7) and (B8):

$$\begin{gathered}[X_2(z),D]_{\mu\nu} = [X_3(z),D]_{\mu\nu} = 0,\\ [X_2(z),Y_2(z)]_{\mu\nu} = [X_2(z),Y_3(z)]_{\mu\nu} = [X_3(z),Y_3(z)]_{\mu\nu} = 0.\end{gathered} \tag{B28}$$

The (μ,ν) elements of other commutators can be calculated as follows:

$$\begin{aligned}[X_1(z),D]_{\mu\nu} &= -X_{1,\mu\mu}(z)(D_\mu - D_\nu)\\ &= g_m x((\mu-\nu)\cdot q,z)\\ &\quad\times\Big(g_{s1}\wp(\mu\cdot q)+g_{s2}\wp^{(1/2)}(\mu\cdot q)+g_{l1}\wp(2\mu\cdot q)+g_{l2}\wp^{(2)}(2\mu\cdot q)\\ &\quad -g_{s1}\wp(\nu\cdot q)-g_{s2}\wp^{(1/2)}(\nu\cdot q)-g_{l1}\wp(2\nu\cdot q)-g_{l2}\wp^{(2)}(2\nu\cdot q)\\ &\quad +\sum_{\lambda\in\Delta_m,\alpha\cdot\mu=1}\wp(\alpha\cdot q)-\sum_{\alpha\in\Delta_m,\alpha\cdot\nu=1}\wp(\alpha\cdot q)\Big),\end{aligned} \tag{B29}$$

$$\begin{aligned}[X_1(z),Y_1(z)]_{\mu\nu} &= \sum_{\lambda\in\Delta_s}(X_{1,\mu\lambda}(z)Y_{1,\lambda\nu}(z)-Y_{1,\mu\lambda}(z)X_{1,\lambda\nu}(z))\\ &= -g_m^2\sum_{\lambda\in\Delta_s\setminus\{\mu,\nu\}}(x((\mu-\lambda)\cdot q,z)y((\lambda-\nu)\cdot q,z)\\ &\quad -y((\mu-\lambda)\cdot q,z)x((\lambda-\nu)\cdot q,z)),\end{aligned} \tag{B30}$$

$$\begin{aligned}[X_1(z),Y_2(z)]_{\mu\nu} &= X_{1,\mu,-\nu}(z)Y_{2,-\nu,\nu}(z)-Y_{2,\mu,-\mu}(z)X_{1,-\mu,\nu}(z)\\ &= -g_m x((\mu+\nu)\cdot q,z)(g_{l1}y(-2\nu\cdot q,z)+g_{l2}y^{(2)}(-2\nu\cdot q,z))\\ &\quad +(g_{l1}y(2\mu\cdot q,z)+g_{l2}y^{(2)}(2\mu\cdot q,z))g_m x(-(\mu+\nu)\cdot q,z),\end{aligned} \tag{B31}$$

$$\begin{aligned}[X_1(z),Y_3(z)]_{\mu\nu} &= X_{1,\mu,-\nu}(z)Y_{3,-\nu,\nu}(z)-Y_{3,\mu,-\mu}(z)X_{1,-\mu,\nu}(z)\\ &= -g_m x((\mu+\nu)\cdot q,z)(g_{s1}y(-\nu\cdot q,2z)+g_{s2}y^{(1/2)}(-\nu\cdot q,2z))\\ &\quad +(g_{s1}y(\mu\cdot q,2z)+g_{s2}y^{(1/2)}(\mu\cdot q,2z))g_m x(-(\mu+\nu)\cdot q,z),\end{aligned} \tag{B32}$$

$$\begin{aligned}[X_2(z),Y_1(z)]_{\mu\nu} &= X_{2,\mu,-\mu}(z)Y_{1,-\mu,\nu}(z)-Y_{1,\mu,-\nu}(z)X_{2,-\nu,\nu}(z)\\ &= -(g_{l1}x(2\mu\cdot q,z)+g_{l2}x^{(2)}(2\mu\cdot q,z))g_m y(-(\mu+\nu)\cdot q,z)\\ &\quad +g_m y((\mu+\nu)\cdot q,z)(g_{l1}x(-2\nu\cdot q,z)+g_{l2}x^{(2)}(-2\nu\cdot q,z)),\end{aligned} \tag{B33}$$

$$\begin{aligned}[X_3(z),Y_1(z)]_{\mu\nu} &= X_{3,\mu,-\nu}(z)Y_{1,-\nu,\nu}(z)-Y_{1,\nu,-\nu}(z)X_{3,-\nu,\nu}(z)\\ &= -2(g_{s1}x(\mu\cdot q,2z)+g_{s2}x^{(1/2)}(\mu\cdot q,2z))g_m y(-(\mu+\nu)\cdot q,z)\\ &\quad +2g_m y((\mu+\nu)\cdot q,z)(g_{s1}x(-\nu\cdot q,2z)+g_{s2}x^{(1/2)}(-\nu\cdot q,2z)).\end{aligned} \tag{B34}$$

We now sum up all these quantities, regroup terms into those multiplied by the same monomial of coupling constants, and show the cancellation in each partial sum. There are six monomials of coupling constants that can occur, i.e., g_m^2, $g_m g_{l1}$, $g_m g_{l2}$, $g_m g_{s1}$, and $g_m g_{s2}$.

Let us consider the terms multiplied by g_m^2. This is a sum of the following two quantities:

$$\mathrm{I}=x((\mu-\nu)\cdot q,z)\Big(\sum_{\alpha\in\Delta_m,\alpha\cdot\mu=1}\wp(\alpha\cdot q)-\sum_{\alpha\in\Delta_m,\alpha\cdot\nu=1}\wp(\alpha\cdot q)\Big)$$

$$\mathrm{II}=-\sum_{\lambda\in\Delta_s\setminus\{\mu,\nu\}}(x((\mu-\lambda)\cdot q,z)y((\lambda-\nu)\cdot q,z)-y((\mu-\lambda)\cdot q,z)x((\lambda-\nu)\cdot q,z)).$$

By the functional identity (24), we can rewrite II into a sum over middle roots,

$$\mathrm{II}=-\sum_{\lambda\in\Delta_s\setminus\{\mu,\nu\}} x((\mu-\nu)\cdot q,z)(\wp((\mu-\lambda)\cdot q)-\wp((\nu-\lambda)\cdot q))$$

$$=-x((\mu-\nu)\cdot q,z)\left(\sum_{\alpha\in\Delta_m,\alpha\cdot\mu=1}\wp(\alpha\cdot q)-\sum_{\alpha\in\Delta_m,\alpha\cdot\nu=1}\wp(\alpha\cdot q)\right).$$

Here the sum over λ has been converted to a sum over α by putting $\alpha=\mu-\lambda$ and $\alpha=\nu-\lambda$ in the two $\wp$ function in the first line. Note that μ, ν, and λ are all orthogonal to each other. We thus find that $\mathrm{I}+\mathrm{II}=0$.

For the other partial sums, we use the following functional identities, which we shall prove in Appendix C:

$$x(2u,z)y(-u-v,z)-y(2u,z)x(-u-v,z)+x(u+v,z)y(-2v,z)$$
$$-y(u+v,z)x(-2v,z)-x(u-v,z)(\wp(2u)-\wp(2v))=0, \tag{B35}$$

$$x^{(2)}(2u,z)y(-u-v,z)-y^{(2)}(2u,z)x(-u-v,z)+x(u+v,z)y^{(2)}(-2v,z)$$
$$-y(u+v,z)x^{(2)}(-2v,z)-x(u-v,z)(\wp^{(2)}(2u)-\wp^{(2)}(2v))=0, \tag{B36}$$

$$2x(u,2z)y(-u-v,z)-y(u,2z)x(-u-v,z)+x(u+v,z)y(-v,2z)$$
$$-2y(u+v,z)x(-v,2z)-x(u-v,z)(\wp(u)-\wp(v))=0, \tag{B37}$$

$$2x^{(1/2)}(u,2z)y(-u-v,z)-y^{(1/2)}(u,2z)x(-u-v,z)+x(u+v,z)y^{(1/2)}(-v,2z)$$
$$-2y(u+v,z)x^{(1/2)}(-v,2z)-x(u-v,z)(\wp^{(1/2)}(u)-\wp^{(1/2)}(v))=0. \tag{B38}$$

By these functional identities, we can confirm that all the partial sums regrouped by $g_m g_{l1}$, $g_m g_{l2}$, $g_m g_{s1}$, and $g_m g_{s2}$, respectively, cancel out.

APPENDIX C: PROOF OF FUNCTIONAL IDENTITIES FOR TWISTED MODELS

We here prove the functional identities that we have encountered in Appendix B. Although the proof is optimized to our choice of $x(u,z)$, $x^{(1/2)}(u,z)$, and $x^{(2)}(u,z)$, the same method can in principle apply to other solutions of the functional equations, such as the functions used by D'Hoker and Phong[15] and Bordner and Sasaki.[14]

1. Analytical properties of $x^{(1/2)}(u,z)$ and $x^{(2)}(u,z)$

The proof of the identities including $x^{(1/2)}(u,z)$ and $x^{(2)}(u,z)$, like the proof in Appendix A, is based on the analytical properties of those functions.

(a) $x^{(1/2)}(u,z)$ has the following analytical properties:

(1) $x^{(1/2)}(u,z)$ is a meromorphic function of u and z. The poles on the u plane and the z plane are located at the lattice points $u=m/2+n\tau$ and $z=m+2n\tau\,(m,n\in\mathbb{Z})$.

(2) $x^{(1/2)}(u,z)$ has the following quasiperiodicity:

$$x^{(1/2)}(u+\tfrac{1}{2},z)=x^{(1/2)}(u,z),\quad x^{(1/2)}(u+\tau,z)=e^{2\pi iz}x^{(1/2)}(u,z),$$
$$x^{(1/2)}(u,z+1)=x^{(1/2)}(u,z),\quad x^{(1/2)}(u,z+2\tau)=e^{4\pi iz}x^{(1/2)}(u,z). \tag{C1}$$

(3) At the origin of the u and z planes, this function exhibits the following singular behavior:

$$x^{(1/2)}(u,z)=\frac{1}{u}-2\rho(z|2\tau)+O(u)\quad(u\rightarrow 0),$$
$$x^{(1/2)}(u,z)=-\frac{2}{z}+2\rho(2u|2\tau)+O(z)\quad(z\rightarrow 0).\tag{C2}$$

(b) $x^{(2)}(u,z)$ has the following analytical properties:

(1) $x^{(2)}(u,z)$ is a meromorphic function of u and z. The poles on the u plane and the z plane are located at the lattice points $u=2m+n\tau$ and $z=m+n\tau/2\,(m,n\in\mathbb{Z})$.

(2) $x^{(2)}(u,z)$ has the following quasiperiodicity:

$$x^{(2)}(u+2,z)=x^{(2)}(u,z),\quad x^{(2)}(u+\tau,z)=e^{2\pi iz}x^{(2)}(u,z),$$
$$x^{(2)}(u,z+1)=x^{(2)}(u,z),\quad x^{(2)}\left(u,z+\frac{\tau}{2}\right)=e^{\pi iu}x^{(2)}(u,z).\tag{C3}$$

(3) At the origin of the u and z planes, this function exhibits the following singular behavior:

$$x^{(2)}(u,z)=\frac{1}{u}-\frac{1}{2}\rho\left(z\,\middle|\,\frac{\tau}{2}\right)+O(u)\quad(u\rightarrow 0),$$
$$x^{(2)}(u,z)=-\frac{1}{2z}+\frac{1}{2}\rho\left(\frac{u}{2}\,\middle|\,\frac{\tau}{2}\right)+O(z)\quad(z\rightarrow 0).\tag{C4}$$

2. Proof of (B35)–(B38)

These four identities can be treated in much the same way. Let us illustrate the proof for (B35) only. Since the line of the proof is almost the same as the proof of (24), we show an outline of the proof and leave the details to the reader.

Let $f(u,v,z)$ denote the left-hand side of (B35),

$$f(u,v,z)=x(2u,z)y(-u-v,z)-y(2u,z)x(-u-v,z)+x(u+v,z)y(-2v,z)$$
$$-y(u+v,z)x(-2v,z)-x(u-v,z)(\wp(2u)-\wp(2v)).\tag{C5}$$

Our task is to show the following analytic properties of $f(u,v,z)$, which imply that this function is identically zero:

(1) $f(u,v,z)$ has the quasiperiodicity as follows:

$$f(u+1,v,z)=f(u,v,z),\quad f(u+\tau,v,z)=e^{2\pi iz}f(u,v,z).\tag{C6}$$

(2) $f(u,v,z)$ is an entire function on the u plane.

The first property is immediate from the quasiperiodicity of $x(u,z)$, etc. Furthermore, it is obvious from the definition that all possible poles of $f(u,v,z)$ on the u plane are limited to the lattice points $u=m/2+n\tau/2$ and $u=-v+m+n\tau\,(m,n\in\mathbb{Z})$. In view of the quasiperiodicity, therefore, we have only to verify that $f(u,v,z)$ is nonsingular at $u=0,1/2,\tau/2,1/2+\tau/2$, and $-v$.

The absence of poles at $u=0,1/2$ and $-v$ can be verified by straightforward calculations on the basis of the singular behavior of $x(u,z)$, $x^{(1/2)}(u,z)$, and $x^{(2)}(u,z)$ as $u\rightarrow 0$.

In order to examine the points $u=\tau/2$ and $u=1/2+\tau/2$, one has to examine the singular behavior of $x(2u,z)$ and $y(2u,z)$ as $u\rightarrow\tau/2,1/2+\tau/2$. This can be worked out by combining the quasiperiodicity of $x(u,z)$ and $y(u,z)$ and their singular behavior as $u\rightarrow 0$,

(1) As $u\rightarrow\tau/2$,

$$x(2u,z)=e^{2\pi iz}x(2u-\tau,z)=e^{2\pi iz}\left(\frac{1}{2u-\tau}+O(1)\right),$$
$$y(2u,z)=2^{2\pi iz}y(2u-\tau,z)=e^{2\pi iz}\left(-\frac{1}{(2u-\tau)^2}+O(1)\right). \tag{C7}$$

(2) As $u\to 1/2+\tau/2$,

$$x(2u,z)=e^{2\pi iz}x(2u-1-\tau,z)=e^{2\pi iz}\left(\frac{1}{2u-1-\tau}+O(1)\right),$$
$$y(2u,z)=e^{2\pi iz}y(2u-1-\tau,z)=e^{2\pi iz}\left(-\frac{1}{(2u-1-\tau)^2}+O(1)\right). \tag{C8}$$

Using these observations, one can confirm the absence of poles of $f(u,v,z)$ at $u=\tau/2$ and $1/2+\tau/2$ by direct calculations.

We can thus verify that $f(u,v,z)$ is indeed an entire function on the u plane.

3. Proof of (B21)–(B26)

Rather than directly proving these identities, let us prove them in a differentiated form. For illustration, we consider the first identity (B21). Differentiating this identity by u gives

$$x(u,2z)y^{(1/2)}(-u,2z)-y(u,2z)z^{(1/2)}(-u,2z)+x^{(1/2)}(u,2z)y(-u,2z)-y^{(1/2)}(u,2z)x(-u,2z)$$
$$=2\wp'(u). \tag{C9}$$

One can prove it directly, repeating the complex analytic reasoning that we have presented in other cases. An alternative way is to take the limit, as $v\to u$, of the functional identity

$$x(2u,2z)y^{(1/2)}(-u-v,2z)-y(2u,2z)x^{(1/2)}(-u-v,2z)+x^{(1/2)}(u+v,2z)y(-2v,2z)$$
$$-y^{(1/2)}(u+v,2z)x(-2v,2z)-x(u-v,z)(\wp(2u)-\wp(2v))=0. \tag{C10}$$

(This yields the above identity upon substituting $u\to u/2$ and $v\to v/2$.) This functional identity can be derived by the same method as the proof of (B35)–(B38).

Similarly, the third and fifth of (B21)–(B26) are obtained from the following functional identities:

$$2x(u,2z)y^{(2)}(-u-v,z)-y(u,2z)x^{(2)}(-u-v,z)+x^{(2)}(u+v,z)y(-2v,2z)$$
$$-2y^{(2)}(u+v,z)x(-2v,2z)-x(u-v,z)(\wp(u)-\wp(v))=0, \tag{C11}$$

$$2x^{(1/2)}(u,2z)y^{(2)}(-u-v,z)-y^{(1/2)}(u,2z)x^{(2)}(-u-v,z)+x^{(2)}(u+v,z)y^{(1/2)}(-2v,2z)$$
$$-2y^{(2)}(u+v,z)x^{(1/2)}(-2v,z)-x(u-v,z)(\wp(u)-\wp(v))=0. \tag{C12}$$

The second, fourth, and sixth of (B21)–(B26) can be similarly derived from the last three of (B35)–(B38). This completes the proof of the functional identities.

We conclude this Appendix with a comment on the ''const.'' terms of these identities. In principle, these terms can be determined by examining the identities at a special point of the u plane. Let us consider, e.g., (B21). At $u=z$, the first term on the left-hand side vanishes. Evaluating the other terms at this point, therefore, one finds that

$$\text{const.}=2\wp(z)-x^{(1/2)}(z,2z)x(-z,2z). \tag{C13}$$

The same formula can be reproduced by substituting $u=-z$. One can similarly derive an explicit expression for the other identities.

[1] Yu. I. Manin, ''Sixth Painlevé equation, universal elliptic curve, and mirror of $\mathbb{P}^2$,'' alg-geom/9605010, AMS Transl. **186**, 131–151 (1998).
[2] K. Okamoto, ''Studies on the Painlevé equations I: Sixth Painlevé equation P_{VI},'' Ann. Mat. Pura Appl. **146**, 337–381 (1987).
[3] N. J. Hitchin, ''Twistor spaces, Einstein metrics, and isomonodromic deformations,'' J. Diff. Geom. **42**, 30–112 (1995).
[4] F. Calogero, ''Solution of the one-dimensional *N*-body problem with quadratic and/or inversely quadratic pair potentials,'' J. Math. Phys. **12**, 419–436 (1971); B. Sutherland, ''Exact results for a quantum many-body problem in one-dimension. II,'' Phys. Rev. A **5**, 1372–1376 (1972); J. Moser, ''Three integrable Hamiltonian systems connected with isospectral deformations,'' Adv. Math. **16**, 197–220 (1975); ''Integrable systems of nonlinear evolution equations,'' in *Dynamical Systems, Theory and Applications*, Lecture Notes in Physics, edited by J. Moser (Springer-Verlag, Berlin, 1975); Vol. 38; F. Calogero, C. Marchioro, and O. Ragnisco, ''Exact solution of the classical and quantal one-dimensional many body problems with the two body potential $V_a(x)=g^2a^2/\sinh^2 ax$,'' Lett. Nuovo Cimento **13**, 383–387 (1975); F. Calogero, ''Exactly solvable one-dimensional many body problems,'' *ibid.* **13**, 411–416 (1975).
[5] I. M. Krichever, ''Elliptic solutions of the Kadomtsev–Petviashvili equation and integrable systems of particles,'' Funct. Anal. Appl. **14**, 282–290 (1980).
[6] V. I. Inozemtsev, ''Lax representation with spectral parameter on a torus for integrable particle systems,'' Lett. Math. Phys. **17**, 11–17 (1989).
[7] A. M. Levin and M. A. Olshanetsky, ''Painlevé–Calogero correspondence,'' alg-geom/9706012, AMS Transl. (2) **191** (1999) (to be published); ''Classical limit of the Knizhnik–Zamolodchikov–Bernard equations as hierarchy of isomonodromic deformations, hep-th/9709207.
[8] N. J. Hitchin, ''Stable bundles and integrable systems,'' Duke Math. J. **54**, 91–144 (1987).
[9] N. Nekrasov, ''Holomorphic bundles and many body problems,'' hep-th/9503157, Commun. Math. Phys. **180**, 587–603 (1996).
[10] B. Enriquez and V. Rubtsov, ''Hitchin systems, higher Gaudin operators and *r*-matrices,'' alg-geom/9503010, Math. Res. Lett. **3**, 343–357 (1996).
[11] A. M. Levin and M. A. Olshanetsky, ''Nonautonomous Hamiltonian systems related to highest Hitchin integrals,'' math-ph/9901019.
[12] A. J. Bordner, E. Corrigan, and R. Sasaki, ''Calogero–Moser models I: A new formulation,'' hep-th/9805106, Prog. Theor. Phys. **100**, 1107–1129 (1998).
[13] A. J. Bordner, R. Sasaki, and K. Takasaki, ''Calogero–Moser models II: Symmetries and foldings,'' hep-th/9809068.
[14] A. J. Bordner and R. Sasaki, ''Calogero–Moser models III: Elliptic potentials and twisting,'' hep-th/9812232.
[15] E. D'Hoker and D. H. Phong, ''Calogero–Moser Lax pairs with spectral parameter,'' hep-th/9804124, Nucl. Phys. B **530**, 537–610 (1998); ''Calogero–Moser and Toda systems for twisted and untwisted affine Lie algebras,'' hep-th/9804125, Nucl. Phys. B **530**, 611–640 (1998).
[16] F. Calogero, ''On a functional equation connected with integrable many-body problems,'' Lett. Nuovo Cimento **16**, 77–80 (1976).
[17] G. Felder and C. Wieczerkowski, ''Conformal blocks on elliptic curves and the Knizhnik–Zamolodchikov–Bernard equations,'' hep-th/941104, Commun. Math. Phys. **176**, 133–161 (1996).
[18] D. Bernard, ''On the Wess–Zumino–Witten model on the torus,'' Nucl. Phys. B **303**, 77–93 (1988); ''On the Wess–Zumino–Witten model on Riemann surfaces,'' *ibid.* **309**, 145–174 (1988).
[19] I. Krichever, O. Babelon, E. Billey, and M. Talon, ''Spin generalization of the Calogero–Moser systems and the matrix KP equation,'' hep-th/9411160, AMS Transl. **170**, 83–119 (1995).
[20] M. Noumi and Y. Yamada, ''Affine Weyl groups, discrete dynamical systems, and Painlevé equation,'' math.QA/9804132, Commun. Math. Phys. **199**, 281–295 (1998); ''Higher order Painlevé equations of type $A_l^{(1)}$,'' math.QA/9808003, Funkcialaj Ekvacioj **41**, 483–503 (1998).
[21] P. Deift, A. Its, A. Kapaev, and X. Zhou, ''On the algebro-geometric integration of the Schlesinger equations and on elliptic solutions of the Painlevé VI equation,'' Commun. Math. Phys. **203**, 613–633 (1999).
[22] A. V. Kitaev and D. A. Korotkin, ''On solutions of the Schlesinger equations in terms of Θ-functions,'' math-ph/9810007, Intern. Math. Res. Notes **17**, 877–905 (1988).
[23] D. A. Korotkin and J. A. H. Samtleben, ''On the quantization of isomonodromic deformations on the torus,'' hep-th/9511087, Int. J. Mod. Phys. A **12**, 2013–2030 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

A Sato formula for reflectionless finite difference operators

J. F. van Diejen
Universidad de Chile, Departamento de Matemáticas, Facultad de Ciencias, Casilla 653, Santiago 1, Chile

(Received 29 March 1999; accepted for publication 1 June 1999)

An explicit parametrization of the Jost functions of the reflectionless Jacobi operators in terms of their spectral data is presented. Interpolation produces a Sato formula for the eigenfunctions of a class of reflectionless finite difference operators on the line. © *1999 American Institute of Physics.* [S0022-2488(99)00311-4]

I. INTRODUCTION

The works on the underlying algebraic structure of infinite-dimensional integrable systems, first initiated by Sato and then further developed (mainly) by the Kyoto school, stand out as one of the highlights of soliton theory.[1–8] Central in Sato's scheme is the idea that a large class of integrable nonlinear evolution equations, among which, e.g., archetype equations as the Korteweg–de Vries (KdV) and the Kadomtsev–Petviashvili (KP) equations, may be understood from a conceptual point of view within the framework of the geometry of infinite-dimensional Grassmann manifolds and the representation theory of Kac–Moody algebras. An important practical feature of the Sato construction is that it produces a closed formula for the eigenfunctions of the Lax operator of the associated linear problem. For instance, in the case of the (soliton regime of the) KdV equation one ends up with an explicit formula for the eigenfunctions of the one-dimensional Schrödinger operators with reflectionless rapidly decreasing potentials.[1–6,8,9] (The potentials of interest are also commonly referred to as soliton potentials or Bargmann potentials.)

The present paper aims at exhibiting an analogous Sato formula for the eigenfunctions of the reflectionless Jacobi operators. To this end we employ inverse-scattering techniques for these operators that have their origin in the work of Flaschka on the soliton dynamics of the infinite Toda chain.[10–14] (The Jacobi operator in turn arises as the Lax operator in the linear problem associated to the infinite Toda chain.)

The structure of the paper is as follows. Sections II and III serve to prepare the grounds by recalling briefly some preliminaries regarding the scattering and inverse scattering theory of Jacobi operators. The inverse scattering theory is then applied in Sec. IV so as to produce a determinantal formula for the *Jost function* of the reflectionless Jacobi operator. The method followed here is based on Flaschka's approach towards corresponding determinantal formulas for the *coefficients* of this Jacobi operator.[10–12] In Sec. V we evaluate our determinantal representation for the reflectionless Jost function with the aid of the Cauchy determinant formula; this entails an explicit Sato type formula for the Jost function under consideration. The transition from the determinantal formula for the Jost function to the Sato type formula is to be compared with the passage from Flaschka's determinantal expressions for the coefficients of the reflectionless Jacobi operator to their explicit representation in terms of tau functions due to Hirota (who, incidentally, used direct methods rather than inverse scattering theory).[15,16,12] Next, in Sec. VI, our Sato formula is interpolated from $\mathbb{Z}$ to $\mathbb{R}$. The result is a Sato formula for reflectionless finite difference operators living on the whole line rather than just on the integer lattice. (In other words, we pass from discrete difference operators acting on functions over $\mathbb{Z}$ to analytic difference operators acting on functions over $\mathbb{R}$.) Finally, the paper is concluded with a list of miscellaneous remarks in Sec. VII.

II. SCATTERING

We will start out by recalling some standard notions from the scattering theory for Jacobi operators (see, e.g., Refs. 10–12). Let

$$D=a_n T+a_{n-1} T^{-1}+b_n \quad (n\in\mathbb{Z}) \tag{II.1}$$

be a Jacobi operator with $a_n>0$ and $b_n\in\mathbb{R}$ for $n\in\mathbb{Z}$. Here T represents a shift operator that acts on lattice functions $\psi:\mathbb{Z}\to\mathbb{C}$ via $(T\psi)(n)=\psi(n+1)$. For the purpose of scattering we are interested in the situation of a Jacobi operator that tends asymptotically to a discrete Laplacian when $|n|$ becomes large. More precisely, from now on it will be assumed that for $|n|\to\infty$ the coefficients a_n and b_n converge rapidly (say exponentially) to 1 and 0, respectively. We then have that our Jacobi operator constitutes a self-adjoint operator in the Hilbert space $l^2(\mathbb{Z})$ of square-summable functions over the integer lattice $\mathbb{Z}$.

The spectral problem associated to D (II.1) is governed by a discrete difference equation of the form

$$a_n\psi(n+1,z)+b_n\psi(n,z)+a_{n-1}\psi(n-1,z)=(z+z^{-1})\psi(n,z), \quad n\in\mathbb{Z}. \tag{II.2}$$

Here z denotes a (possibly complex) spectral parameter. The conditions on the coefficients guarantee that asymptotically (for $|n|\to\infty$) the solutions of the difference equation in Eq. (II.2) decompose into a linear combination of the plane waves z^n and z^{-n}. The Jost function (of the first kind) $\psi_{\text{jost}}(n,z)$ is the solution to Eq. (II.2) that is characterized by an asymptotics of the form z^n for $n\to\infty$. More precisely, one has that

$$\psi_{\text{jost}}(n,z)\to\begin{cases} z^n & \text{for } n\to+\infty \\ \alpha(z)z^n+\beta(z)z^{-n} & \text{for } n\to-\infty, \end{cases} \tag{II.3}$$

where $\alpha(z)$ and $\beta(z)$ represent two z-dependent coefficients that describe the asymptotics of the Jost function at minus infinity. For generic z, the solution $\psi_{\text{jost}}(n,z)$ is obtained from a fundamental system consisting of two independent solutions to the second-order difference Eq. (II.2) by taking an appropriate linear combination. The coefficients $\alpha(z)$ and $\beta(z)$ contain important information regarding the spectrum of the Jacobi operator under consideration. Specifically, the zeros of $\alpha(z)$—which with our conditions on the coefficients a_n and b_n are simple, finite in number, and lie inside the punctured interval $]-1,1[\setminus\{0\}$—determine the discrete spectrum. (Notice in this connection that at such spectral values the Jost function decays exponentially for $|n|\to\infty$.) Let us denote the zeros in question by $1>z_1>z_2>\cdots>z_N>-1$ (with $z_j\neq 0$) and let the numbers $\nu_1,\ldots,\nu_N>0$ be the corresponding normalization constants,

$$\nu_j=\left(\sum_{n\in\mathbb{Z}}\psi^2_{\text{jost}}(n,z_j)\right)^{-1}, \quad j=1,\ldots,N. \tag{II.4}$$

The discrete spectral values $z_1,\ldots,z_N$ and normalization constants $\nu_1,\ldots,\nu_N$, together with the *reflection coefficient* $r(z):=-\beta(z^{-1})/\alpha(z)$, are referred to as the *spectral data* of the Jacobi operator. (Our definition of $\beta(z)$ differs from the standard conventions (cf. e.g., Refs. 10–12) by the change $\beta(z)\to-\beta(z^{-1})$.) When $r(z)=\beta(z)\equiv 0$ (identically in z) the Jacobi operator is said to be *reflectionless*.

III. INVERSE SCATTERING

It is known from the inverse scattering theory for Jacobi operators (with $a_n>0$, $b_n\in\mathbb{R}$ and $a_n\to 1$, $b_n\to 0$ rapidly for $|n|\to\infty$) that both the coefficients a_n, b_n and the Jost function $\psi_{\text{jost}}(n,z)$ can be completely recovered from the spectral data.[10–12] Specifically, one has that

$$a_n=\frac{K(n+1,n+1)}{K(n,n)}, \tag{III.1a}$$

$$b_n=\frac{K(n,n+1)}{K(n,n)}-\frac{K(n-1,n)}{K(n-1,n-1)}, \tag{III.1b}$$

and

$$\psi_{\text{jost}}(n,z)=\sum_{m=n}^{\infty} K(n,m)z^m, \quad 0<|z|\leqslant 1 \tag{III.2}$$

$(n\in\mathbb{Z})$, where $K(n,m)$ is a kernel of the form

$$K(n,m)=\begin{cases}\left(1+F(2n)+\sum_{l=n+1}^{\infty} k(n,l)F(l+n)\right)^{-1/2} & \text{for } m=n,\\ K(n,n)\,k(n,m) & \text{for } m>n.\end{cases} \tag{III.3}$$

Here the kernel $F(l+n)$, which is determined by the function

$$F(m)=\frac{1}{2\pi i}\oint r(z)z^{m-1}\mathrm{d}z+\sum_{j=1}^{N} \nu_j z_j^m \tag{III.4}$$

(where the integration is along the unit circle in the positive direction), encodes the dependence on the spectral data, and the kernel $k(n,m)$, $m>n$ is obtained from $F(m)$ (III.4) as the (unique) solution of the *discrete Gelfand–Levitan–Marchenko equation*

$$k(n,m)+F(n+m)+\sum_{l=n+1}^{\infty} k(n,l)F(l+m)=0, \;\; m>n. \tag{III.5}$$

IV. THE REFLECTIONLESS CASE: DETERMINANTAL REPRESENTATIONS

In the reflectionless situation, i.e., with $r(z)\equiv 0$, the kernel associated to $F(m)$ (III.4) becomes seperable and of finite rank,

$$F(m)=\sum_{j=1}^{N} \nu_j z_j^m. \tag{IV.1}$$

It is known from the work of Flaschka that the question of solving the discrete Gelfand–Levitan–Marchenko equation (III.5) reduces in this situation to a finite-dimensional problem that can be solved explicitly in closed form.[10–12] In this section we will use the exact solution in question to arrive at a determinantal formula for the reflectionless Jost function.

To this end we first recall Flaschka's solution of the Gelfand–Levitan–Marchenko equation with $r(z)\equiv 0$. The key step lies in the substitution of the separable ansatz,

$$k(n,m)=\sum_{j=1}^{N} k_j(n)z_j^m \tag{IV.2}$$

for the unknown kernel $k(n,m)$. The discrete Gelfand–Levitan–Marchenko equation of Eq. (III.5)—with the function $F(m)$ is taken from Eq. (IV.1)—produces the relation

$$\sum_{j=1}^{N} (\nu_j z_j^n+k_j(n))z_j^m+\sum_{i,j=1}^{N} \left(k_i(n)\,\nu_j z_j^m \sum_{l=n+1}^{\infty} (z_i z_j)^l\right)=0. \tag{IV.3}$$

Summation of the geometric series gives rise to the following N-dimensional linear system for the quantities $k_1(n),\ldots,k_N(n)$,

$$k_j(n)+\nu_j\sum_{i=1}^{N} k_i(n)\frac{z_i^{n+1}z_j^{n+1}}{1-z_iz_j}=-\nu_j z_j^n,\quad j=1,\ldots,N. \tag{IV.4}$$

The linear system is subsequently solved in the standard way by means of Cramer's Rule,

$$k_j(n)=\frac{\det \mathbf{A}^{(j)}(n)}{\det \mathbf{A}(n)},\quad j=1,\ldots,N, \tag{IV.5}$$

where

$$\mathbf{A}(n)=\begin{bmatrix} 1+\nu_1\frac{z_1^{2n+2}}{1-z_1^2} & \cdots & \nu_1\frac{z_1^{n+1}z_j^{n+1}}{1-z_1z_j} & \cdots & \nu_1\frac{z_1^{n+1}z_N^{n+1}}{1-z_1z_N} \\ \vdots & \ddots & \vdots & & \vdots \\ \nu_j\frac{z_j^{n+1}z_1^{n+1}}{1-z_jz_1} & \cdots & 1+\nu_j\frac{z_j^{2n+2}}{1-z_j^2} & \cdots & \nu_j\frac{z_j^{n+1}z_N^{n+1}}{1-z_jz_N} \\ \vdots & & \vdots & \ddots & \vdots \\ \nu_N\frac{z_N^{n+1}z_1^{n+1}}{1-z_Nz_1} & \cdots & \nu_N\frac{z_N^{n+1}z_j^{n+1}}{1-z_Nz_j} & \cdots & 1+\nu_N\frac{z_N^{2n+2}}{1-z_N^2} \end{bmatrix} \tag{IV.6}$$

and

$$\mathbf{A}^{(j)}(n)=\begin{bmatrix} 1+\nu_1\frac{z_1^{2n+2}}{1-z_1^2} & \cdots & -\nu_1 z_1^n & \cdots & \nu_1\frac{z_1^{n+1}z_N^{n+1}}{1-z_1z_N} \\ \vdots & \ddots & \vdots & & \vdots \\ \nu_j\frac{z_j^{n+1}z_1^{n+1}}{1-z_jz_1} & \cdots & -\nu_j z_j^n & \cdots & \nu_j\frac{z_j^{n+1}z_N^{n+1}}{1-z_jz_N} \\ \vdots & & \vdots & \ddots & \vdots \\ \nu_N\frac{z_N^{n+1}z_1^{n+1}}{1-z_Nz_1} & \cdots & -\nu_N z_N^n & \cdots & 1+\nu_N\frac{z_N^{2n+2}}{1-z_N^2} \end{bmatrix}. \tag{IV.7}$$

$\uparrow$
(jth column)

(The matrix $\mathbf{A}^{(j)}(n)$ is obtained from $\mathbf{A}(n)$ (IV.6) by replacing its jth column by the transpose of the vector $(-\nu_1 z_1^n,\ldots,-\nu_N z_N^n)$.) The (inverse square of the) diagonal of the kernel $K(n,m)$ (III.3) is now computed as

$$K^{-2}(n,n)\overset{\text{Eqs. (IV.1),(IV.2)}}{=}1+\sum_{j=1}^{N}\nu_j z_j^{2n}+\sum_{i,j=1}^{N}k_i(n)\nu_j\frac{z_i^{n+1}z_j^{2n+1}}{1-z_iz_j}$$

$$\overset{\text{Eq. (IV.4)}}{=}1-\sum_{j=1}^{N}k_j(n)z_j^n$$

$$\stackrel{\text{Eq. (IV.5)}}{=} \frac{\det \mathbf{A}(n) - \Sigma_{j=1}^{N} z_j^n \det \mathbf{A}^{(j)}(n)}{\det \mathbf{A}(n)}$$

$$= \det \mathbf{A}(n-1)/\det \mathbf{A}(n). \tag{IV.8}$$

(In the last step one uses that the difference $\mathbf{A}(n-1)-\mathbf{A}(n)$ is a rank-one matrix of the form $[\nu_i z_i^n z_j^n]_{1\leqslant i,j\leqslant N}$, whence $\det \mathbf{A}(n-1) - \det \mathbf{A}(n) = -\Sigma_{j=1}^{N} z_j^n \det \mathbf{A}^{(j)}(n)$.)

One thus ends up with the following determinantal representation for the coefficients a_n (III.1a):

$$a_n = \frac{\sqrt{\det \mathbf{A}(n+1)\det \mathbf{A}(n-1)}}{\det \mathbf{A}(n)}, \quad n\in\mathbb{Z}. \tag{IV.9}$$

In a similar way one arrives at a determinantal formula for the coefficients b_n (III.1b),

$$\begin{aligned} b_n &\stackrel{\text{Eqs. (III.1b),(III.3)}}{=} k(n,n+1) - k(n-1,n) \\ &\stackrel{\text{Eq. (IV.2)}}{=} \sum_{j=1}^{N} (k_j(n) z_j^{n+1} - k_j(n-1) z_j^n) \\ &\stackrel{\text{Eq. (IV.5)}}{=} \sum_{j=1}^{N} \left(\frac{\det \mathbf{A}^{(j)}(n)}{\det \mathbf{A}(n)} z_j^{n+1} - \frac{\det \mathbf{A}^{(j)}(n-1)}{\det \mathbf{A}(n-1)} z_j^n \right) \\ &= \sum_{j=1}^{N} \left(\frac{\det \mathbf{B}^{(j)}(n)}{\det \mathbf{A}(n)} - \frac{\det \mathbf{B}^{(j)}(n-1)}{\det \mathbf{A}(n-1)} \right), \end{aligned} \tag{IV.10}$$

where

$$\mathbf{B}^{(j)}(n) = \begin{bmatrix} 1+\nu_1 \frac{z_1^{2n+2}}{1-z_1^2} & \cdots & \nu_1(z_1-z_1^{-1})\frac{z_1^{n+1}z_j^{n+1}}{1-z_1 z_j} & \cdots & \nu_1 \frac{z_1^{n+1} z_N^{n+1}}{1-z_1 z_N} \\ \vdots & \ddots & \vdots & & \vdots \\ \nu_j \frac{z_j^{n+1} z_1^{n+1}}{1-z_j z_1} & \cdots & \nu_j(z_j - z_j^{-1}) \frac{z_j^{2n+2}}{1-z_j^2} & \cdots & \nu_j \frac{z_j^{n+1} z_N^{n+1}}{1-z_j z_N} \\ \vdots & & \vdots & \ddots & \vdots \\ \nu_N \frac{z_N^{n+1} z_1^{n+1}}{1-z_N z_1} & \cdots & \nu_N(z_N - z_N^{-1}) \frac{z_N^{n+1} z_j^{n+1}}{1-z_N z_j} & \cdots & 1+\nu_N \frac{z_N^{2n+2}}{1-z_N^2} \end{bmatrix}. \tag{IV.11}$$

↑
(jth column)

(So the matrix $\mathbf{B}^{(j)}(n)$ is obtained from $\mathbf{A}(n)$ (IV.6) by replacing the jth column by the transpose of the vector $(\nu_1(z_1-z_1^{-1})z_1^n z_j^n/(1-z_1z_j), \ldots, \nu_N(z_N-z_N^{-1})z_N^n z_j^n/(1-z_N z_j))$.) In passing from the third to the fourth line of Eq. (IV.10) it was used that

$$\sum_{j=1}^{N} \left(\det \mathbf{A}^{(j)}(n) z_j^{n+1} - \det \mathbf{B}^{(j)}(n) \right)$$

$$= \sum_{j=1}^{N} \det \begin{bmatrix} 1+\nu_1 \frac{z_1^{2n+2}}{1-z_1^2} & \cdots & \nu_1 \frac{z_1^{n+1} z_j^{n+1}}{1-z_1 z_j}(z_j - z_1) & \cdots & \nu_1 \frac{z_1^{n+1} z_N^{n+1}}{1-z_1 z_N} \\ \vdots & \ddots & \vdots & & \vdots \\ \nu_j \frac{z_j^{n+1} z_1^{n+1}}{1-z_j z_1} & \cdots & 0 & \cdots & \nu_j \frac{z_j^{n+1} z_N^{n+1}}{1-z_j z_N} \\ \vdots & & \vdots & \ddots & \vdots \\ \nu_N \frac{z_N^{n+1} z_1^{n+1}}{1-z_N z_1} & \cdots & \nu_N \frac{z_N^{n+1} z_j^{n+1}}{1-z_N z_j}(z_j - z_N) & \cdots & 1+\nu_N \frac{z_N^{2n+2}}{1-z_N^2} \end{bmatrix} = 0.$$

$$\uparrow$$
$$(j\text{th column})$$

(To see that the determinants on the second line of the above formula indeed sum up to yield zero, one develops the determinant of jth term to the jth column; this way it is not difficult to see that the subdeterminant in the jth term originating from the minor corresponding to the ith row cancels against the subdeterminant in the ith term originating from the minor corresponding to the jth row.) It is immediate from the last line of Eq. (IV.10) that we may write

$$b_n = \partial_t \log\left(\frac{\det \mathbf{A}(n,t)}{\det \mathbf{A}(n-1,t)} \right) \bigg|_{t=0}, \quad n \in \mathbb{Z}, \tag{IV.12}$$

where $\mathbf{A}(n,t)$ denotes the matrix of the form in Eq. (IV.6) with $\nu_j \to \nu_j e^{t(z_j - z_j^{-1})}$, $j = 1, \ldots, N$.

The above method for computing the coefficients of the reflectionless Jacobi operator is due to Flaschka,[10] who used it to arrive at the determinantal representation for a_n given by Eq. (IV.9). The determinantal representation for b_n in Eq. (IV.12) can moreover be found e.g., in the book by Toda[12] (who provides a somewhat different proof). More recently an alternative approach towards the derivation of the determinantal formulas for the coefficients a_n and b_n was presented in Ref. 17 (see also Refs. 13,14).

After these preparations we are now finally in the position to present the corresponding determinantal formula for the Jost function $\psi_{\text{jost}}(n,z)$. Indeed, starting from Eq. (III.2) we get by successive manipulations,

$$\begin{aligned} \psi_{\text{jost}}(n,z) &\overset{\text{Eqs. (III.3),(IV.2)}}{=} K(n,n)\left(z^n + \sum_{j=1}^{N} k_j(n) \sum_{l=n+1}^{\infty} (z_j z)^l \right) \\ &= K(n,n)\left(z^n + \sum_{j=1}^{N} k_j(n) \frac{(z_j z)^{n+1}}{1-z_j z} \right) \\ &\overset{\text{Eqs. (IV.5),(IV.8)}}{=} \frac{z^n \left(\det \mathbf{A}(n) + \sum_{j=1}^{N} z_j^n \det \mathbf{A}^{(j)}(n) \frac{z_j z}{1-z_j z} \right)}{\sqrt{\det \mathbf{A}(n) \det \mathbf{A}(n-1)}} \\ &= \frac{z^n \det \mathbf{C}(n,z)}{\sqrt{\det \mathbf{A}(n) \det \mathbf{A}(n-1)}}, \end{aligned} \tag{IV.13}$$

where $\mathbf{C}(n,z)$ denotes the matrix obtained from $\mathbf{A}(n)$ (IV.6) by the substitution $\nu_j \to \nu_j(1-zz_j^{-1})/(1-zz_j)$, $j=1,\ldots,N$. More explicitly, the matrix $\mathbf{C}(n,z)$ is given by

$$\mathbf{C}(n,z)=\begin{bmatrix} 1+\nu_1\frac{z_1^{2n+2}}{1-z_1^2}\left(\frac{1-zz_1^{-1}}{1-zz_1}\right) & \cdots & \nu_1\frac{z_1^{n+1}z_N^{n+1}}{1-z_1z_N}\left(\frac{1-zz_1^{-1}}{1-zz_1}\right) \\ \vdots & \ddots & \vdots \\ \nu_N\frac{z_N^{n+1}z_1^{n+1}}{1-z_Nz_1}\left(\frac{1-zz_N^{-1}}{1-zz_N}\right) & \cdots & 1+\nu_N\frac{z_N^{2n+2}}{1-z_N^2}\left(\frac{1-zz_N^{-1}}{1-zz_N}\right) \end{bmatrix}. \quad \text{(IV.14)}$$

In order to verify that the expressions on the third and fourth line of Eq. (IV.13) are indeed equivalent, we employ the following pole expansion for the determinant of the matrix $\mathbf{C}(n,z)$ (IV.14):

$$\det \mathbf{C}(n,z)=\det \mathbf{C}_0(n)+\sum_{j=1}^{N} \det \mathbf{C}_j(n)\,\frac{zz_j}{1-zz_j}, \quad \text{(IV.15)}$$

where $\mathbf{C}_0(n)=\lim_{z\to 0}\mathbf{C}(n,z)=\mathbf{A}(n)$ and $\mathbf{C}_j(n)$ is the matrix obtained from $\mathbf{C}(n,z)$ via the substitution $z=z_j^{-1}$ after having multiplied the jth row by a factor $1-zz_j$ to compensate the singularity. Subtraction of z_k^{n+1}/z_j^{n+1} times the jth column of $\mathbf{C}_j(n)$ from its kth column (for $k=1,\ldots,N,\ k\neq j$) and multiplication of the resulting matrix from the left by $\mathrm{diag}(1-z_1/z_j,\ldots,1,\ldots,1-z_N/z_j)$ and from the right by $\mathrm{diag}(1/(1-z_1/z_j),\ldots,1,\ldots,1/(1-z_N/z_j))$ (where the units are in the jth slot), leads one to a matrix that is equal to the matrix $\mathbf{A}^{(j)}(n)$ with its jth column multiplied by z_j^n. Hence, we conclude that the expressions on the third and fourth line of Eq. (IV.13) are equal.

V. THE DISCRETE SATO FORMULA

We will now evaluate the determinantal representations of the previous section with the aid of the Cauchy determinant formula [see, e.g., Macdonald (Ref. 18, p. 67)],

$$\det\left[\frac{1}{1-x_jy_k}\right]_{1\leq j,k\leq N}=\frac{\prod_{1\leq j<k\leq N}(x_j-x_k)(y_j-y_k)}{\prod_{1\leq j,k\leq N}(1-x_jy_k)}. \quad \text{(V.1)}$$

This leads us to an explicit parametrization of the Jost function for the reflectionless Jacobi operator in terms of the spectral data. To describe the result some notation is needed,

$$\tau(n)=\sum_{J\in\{1,\ldots,N\}}\prod_{j\in J}\frac{\nu_jz_j^{2n+2}}{1-z_j^2}\prod_{\substack{j,k\in J\\ j<k}}\left(\frac{z_j-z_k}{1-z_jz_k}\right)^2, \quad \text{(V.2a)}$$

$$\sigma(n)=\sum_{J\in\{1,\ldots,N\}}\sum_{j\in J}(z_j-z_j^{-1})\prod_{j\in J}\frac{\nu_jz_j^{2n+2}}{1-z_j^2}\prod_{\substack{j,k\in J\\ j<k}}\left(\frac{z_j-z_k}{1-z_jz_k}\right)^2, \quad \text{(V.2b)}$$

$$\chi(n,z)=\sum_{J\in\{1,\ldots,N\}}\prod_{j\in J}\frac{\nu_jz_j^{2n+2}}{1-z_j^2}\left(\frac{1-zz_j^{-1}}{1-zz_j}\right)\prod_{\substack{j,k\in J\\ j<k}}\left(\frac{z_j-z_k}{1-z_jz_k}\right)^2. \quad \text{(V.2c)}$$

The following theorem provides the parametrization of the reflectionless Jost function in terms of the spectral data.

Theorem 1 (Spectral Parametrization): *Let $D=a_nT+a_{n-1}T^{-1}+b_n$ be a (self-adjoint) Jacobi operator in $l^2(\mathbb{Z})$ with $a_n>0$, $b_n\in\mathbb{R}$ such that $a_n\to 1$, $b_n\to 0$ rapidly (say exponentially)* *for* $|n|\to\infty$. *Furthermore, let D be reflectionless* $(r(z)\equiv 0)$ *and characterized by spectral data of the form* $1>z_1>\cdots>z_N>-1$, $z_j\neq 0$ *(corresponding to the discrete spectrum (more precisely, the discrete spectrum consists of N eigenvalues of the form* $(z_j+z_j^{-1})$, $j=1,\ldots,N$.*)) and* $\nu_1,\ldots,\nu_N>0$ *(corresponding to the associated normalization constants, cf. Eq. (II.4)).*

Then the coefficients of D may be written as

$$a_n=\frac{\sqrt{\tau(n+1)\,\tau(n-1)}}{\tau(n)},\qquad b_n=\frac{\sigma(n)}{\tau(n)}-\frac{\sigma(n-1)}{\tau(n-1)},$$

and the Jost solution $\psi_{\mathrm{jost}}(n,z)$ *of the eigenvalue equation* $D\psi=(z+z^{-1})\psi$ *with asymptotics of the form* $\psi_{\mathrm{jost}}(n,z)\to z^n$ *for* $n\to+\infty$ *takes the form*

$$\psi_{\mathrm{jost}}(n,z)=\frac{z^n\,\chi(n,z)}{\sqrt{\tau(n)\,\tau(n-1)}}.$$

Here the functions $\tau(n)$, $\sigma(n)$, *and* $\chi(n,z)$ *are defined by Eqs. (V.2a), (V.2b) and (V.2c), respectively. Moreover, the spatial variable n lives on the integer lattice* $\mathbb{Z}$ *and the spectral variable z is assumed to take values in the punctured unit disk* $0<|z|\leqslant 1$.

Proof: The reconstruction from the spectral data given in Sec. IV provides us with determinantal representations for the coefficients and Jost function of the reflectionless Jacobi operator (cf. Eqs. (IV.9), (IV.12), and (IV.13).

It is not very difficult to infer—with the aid of the Cauchy determinant formula (V.1)—that the determinant of the matrix $\mathbf{A}(n)$ (IV.6) is given by $\tau(n)$ (V.2a). Indeed, the matrix $\mathbf{A}(n)$ (IV.6) has the structure of an identity matrix plus a matrix of the form **NDCD**, where $\mathbf{N}=\mathrm{diag}(\nu_1,\ldots,\nu_N)$, $\mathbf{D}=\mathrm{diag}(z_1^{n+1},\ldots,z_N^{n+1})$, and $\mathbf{C}=[(1-z_jz_k)^{-1}]_{1\leqslant j,k\leqslant N}$. We thus have that the determinant of $\mathbf{A}(n)$ amounts to the sum of all principal minors of the matrix **NDCD**, which are readily evaluated by means of the Cauchy determinant formula (with $x_j=y_j=z_j$, $j=1,\ldots,N$) so as to produce $\det\mathbf{A}(n)=\tau(n)$. The stated expressions for a_n, b_n, and $\psi_{\mathrm{jost}}(n,z)$ then follow from Eqs. (IV.9), (IV.12), and (IV.13). In this connection it is helpful to recall that the matrices $\mathbf{A}(n,t)$ and $\mathbf{C}(n,z)$, appearing in Eqs. (IV.12) and (IV.13), are obtained from the matrix $\mathbf{A}(n)$ (IV.6) via the substitutions $\nu_j\to\nu_j e^{t(z_j-z_j^{-1})}$ and $\nu_j\to\nu_j(1-zz_j^{-1})/(1-zz_j)$, respectively. (Thus one has that $\partial_t\log(\det\mathbf{A}(n,t))|_{t=0}=\sigma(n)/\tau(n)$ and $\det\mathbf{C}(n,z)=\chi(n,z)$.) □

Theorem V immediately implies the following Sato-type formula for (the wave functions of) the reflectionless discrete difference operators of Jacobi-type.

Corollary 1. (Discrete Sato Formula): Let $0<|z|\leqslant 1$, $1>z_1>\cdots>z_N>-1$ *with* $z_j\neq 0$, $\nu_1,\ldots,\nu_N>0$, *and let* $\tau(n)$, $\sigma(n)$, *and* $\chi(n,z)$ *be given by Eqs. (V.2a), (V.2b), and (V.2c). Then we have that the discrete Sato function of the form*

$$\psi_{\mathrm{jost}}(n,z)=\frac{z^n\,\chi(n,z)}{\sqrt{\tau(n)\,\tau(n-1)}}$$

solves the discrete difference equation,

$$a_n\,\psi(n+1,z)+b_n\,\psi(n,z)+a_{n-1}\,\psi(n-1,z)=(z+z^{-1})\,\psi(n,z),\quad n\in\mathbb{Z},$$

with coefficients given by

$$a_n=\frac{\sqrt{\tau(n+1)\,\tau(n-1)}}{\tau(n)},\qquad b_n=\frac{\sigma(n)}{\tau(n)}-\frac{\sigma(n-1)}{\tau(n-1)}.$$

VI. INTERPOLATION: FROM $\mathbb{Z}$ TO $\mathbb{R}$

Next we interpolate the discrete Sato formula of Corollary V so as to arrive at a corresponding Sato formula for certain reflectionless finite difference operators living on the whole line instead of merely on the integer lattice. For this purpose it is convenient to perform a trigonometric change of variables/parameters,

$$z=e^{y},\quad n=x,\ z_j=e^{-\kappa_j}\quad (j=1,\ldots,N). \tag{VI.1}$$

The functions τ, σ, and χ of Eqs. (V.2a), (V.2b), and (V.2c) pass with these substitutions over into

$$\tau(x)=\sum_{J\in\{1,\ldots,N\}}\prod_{j\in J}\frac{\nu_j e^{-\kappa_j}}{2\sinh(\kappa_j)}\prod_{\substack{j,k\in J\\ j<k}}\left(\frac{\sinh\frac{1}{2}(\kappa_j-\kappa_k)}{\sinh\frac{1}{2}(\kappa_j+\kappa_k)}\right)^2\exp\left(-2x\sum_{j\in J}\kappa_j\right), \tag{VI.2a}$$

$$\sigma(x)=-2\sum_{J\in\{1,\ldots,N\}}\left[\sum_{j\in J}\sinh(\kappa_j)\prod_{j\in J}\frac{\nu_j e^{-\kappa_j}}{2\sinh(\kappa_j)}\prod_{\substack{j,k\in J\\ j<k}}\left(\frac{\sinh\frac{1}{2}(\kappa_j-\kappa_k)}{\sinh\frac{1}{2}(\kappa_j+\kappa_k)}\right)^2\right.$$
$$\left.\times\exp\left(-2x\sum_{j\in J}\kappa_j\right)\right], \tag{VI.2b}$$

$$\chi(x,y)=\sum_{J\in\{1,\ldots,N\}}\left[\prod_{j\in J}\frac{\nu_j}{2\sinh(\kappa_j)}\frac{\sinh\frac{1}{2}(y+\kappa_j)}{\sinh\frac{1}{2}(y-\kappa_j)}\prod_{\substack{j,k\in J\\ j<k}}\left(\frac{\sinh\frac{1}{2}(\kappa_j-\kappa_k)}{\sinh\frac{1}{2}(\kappa_j+\kappa_k)}\right)^2\exp\left(-2x\sum_{j\in J}\kappa_j\right)\right]. \tag{VI.2c}$$

The following theorem provides a Sato formula for the wave functions of analytic difference operators of the form $\mathcal{D}=a(x)\,T+a(x-1)\,T^{-1}+b(x)$ acting on functions $\psi:\mathbb{R}\to\mathbb{C}$. Here $(T\psi)(x)=\psi(x+1)$ and the coefficients $a(x)$, $b(x)$ are obtained from the coefficients a_n, b_n of Theorem 1/Corollary 2 via the substitutions in Eq. (VI.1).

Theorem 3. (Difference Sato Formula): *Let* $0<\kappa_1<\cdots<\kappa_N$, $\nu_1,\ldots,\nu_N>0$, *and let* $\tau(x)$, $\sigma(x)$. *and* $\chi(x,z)$ *be given by Eqs. (VI.2a), (VI.2b), and (VI.2c). Then we have that the difference Sato function of the form,*

$$\psi(x,y)=\frac{\exp(xy)\ \chi(x,y)}{\sqrt{\tau(x)\ \tau(x-1)}},$$

solves the analytic difference equation

$$a(x)\ \psi(x+1,y)+b(x)\ \psi(x,y)+a(x-1)\ \psi(x-1,y)=2\cosh(y)\ \psi(x,y),\quad x\in\mathbb{R},$$

with coefficients given by

$$a(x)=\frac{\sqrt{\tau(x+1)\ \tau(x-1)}}{\tau(x)},\quad b(x)=\frac{\sigma(x)}{\tau(x)}-\frac{\sigma(x-1)}{\tau(x-1)}.$$

Here it is assumed that the spectral parameter $y\in\mathbb{C}$ *is not equal to* $\kappa_j \bmod 2\pi i$ *for* $j=1,\ldots,N$.

Proof: It is clear that, with the stated restrictions on the parameters, the coefficients $a(x)$, $b(x)$, and the wave function $\psi(x,y)$ are well-defined and regular for $x\in\mathbb{R}$. Indeed, the tau functions in the denominators stay away from zero for $x\in\mathbb{R}$, since $\tau(x)$ is positive on the real line as a sum of positive terms. For $x\in\mathbb{Z}$ and $\mathrm{Re}(y)\leqslant 0$, the statement of the theorem is immediate from Corollary V. Hence, we have that at such integer values for x the difference equation with Sato wave function holds as an identity in the parameters $\kappa_1,\ldots,\kappa_N$ and $\nu_1,\ldots,\nu_N$. The generalization to general $x\in\mathbb{R}$ then follows from the observation that a translation $x\to x+\Delta$, $\Delta\in\mathbb{R}$ is equivalent to a reparametrization of the form $\nu_j\to\nu_j e^{2\Delta}$, $j=1,\ldots,N$, together with a multiplication of the wave function by the (x-independent) overall factor $e^{\Delta y}$. The passage from y in the left half-plane to general complex $y\neq\kappa_j \bmod 2\pi i$, $j=1,\ldots,N$ is clear by analyticity. □

VII. MISCELLANEOUS REMARKS

A. Asymptotics

From the formula of Theorem 1/Corollary 2 we read-off that our Jost function $\psi_{\mathrm{jost}}(n,z)$ has an asymptotics at infinity of the form

$$\psi_{\mathrm{jost}}(n,z)\to\begin{cases} z^n & \text{for } n\to+\infty \\ z^n\prod_{j=1}^{N}\left(\frac{z_j-z}{1-z_jz}\right) & \text{for } n\to-\infty. \end{cases} \tag{VII.1}$$

This asymptotics is thus of the general form given by Eq. (II.3) with $\alpha(z)=\Pi_{j=1}^{N}[(z_j-z/1-z_jz)]$ and $\beta(z)=0$. The vanishing of $\beta(z)$ is in agreement with the fact that our Jacobi operator is reflectionless. The formulas moreover confirm that the zeros of $\alpha(z)$ do indeed correspond to the values $z=z_j$, $j=1,\ldots,N$. For the wave function of Theorem 3 the asymptotics at infinity becomes ($n\to x$, $z\to e^y$ and $z_j\to e^{-\kappa_j}$),

$$\psi(x,y)\to\begin{cases} \exp(xy) & \text{for } x\to+\infty \\ \exp(xy)\prod_{j=1}^{N}\frac{\sinh\frac{1}{2}(y+\kappa_j)}{\sinh\frac{1}{2}(y-\kappa_j)} & \text{for } x\to-\infty. \end{cases} \tag{VII.2}$$

B. Summation and integration formulas

It is clear from the explicit formula of Theorem 1/Corollary 2 that the reflectionless Jost function $\psi_{\mathrm{jost}}(n,z)$ decays exponentially for $|n|\to\infty$ at the discrete spectral values $z=z_j$, $j=1,\ldots,N$. Hence, the wave function is indeed square-summable at these spectral values. (As it should be, because the spectral values $z_1,\ldots,z_N$ constitute the discrete spectrum.) The interpretation of the parameters $\nu_1,\ldots,\nu_N$ as the corresponding normalization constants (cf. Eq. (II.4)) moreover implies that the wave function of Theorem 1/Corollary 2 satisfies the following summation identity;

$$\sum_{n\in\mathbb{Z}}\psi_{\mathrm{jost}}^2(n,z_j)=1/\nu_j,\quad j=1,\ldots,N. \tag{VII.3}$$

For the wave function of Theorem 3 this translates into the summation formula

$$\sum_{n\in\mathbb{Z}}\psi^2(n+x,-\kappa_j)=1/\nu_j,\quad j=1,\ldots,N \tag{VII.4a}$$

($x\in\mathbb{R}$). Here it was again used (cf. the proof of Theorem 3) that the shifted function $\psi(n+x,-\kappa_j)$ differs from $\psi(n,-\kappa_j)$ by a factor $e^{-\kappa_j x}$ combined with a reparametrization of the form $\nu_k\to\nu_k e^{-2\kappa_k x}$, $k=1,\ldots,N$. Multiplication of Eq. (VII.4a) by $e^{2\pi i m x}$ ($m\in\mathbb{Z}$), and integration in x over a unit interval, entails upon interchanging the summation and integration on the left-hand side,

$$\int_{-\infty}^{\infty} \psi^2(x,-\kappa_j)\, e^{2\pi i m x} \mathrm{d}x = \begin{cases} 1/\nu_j & \text{for } m=0 \\ 0 & \text{for } m\in\mathbb{Z}\setminus\{0\}, \end{cases} \tag{VII.4b}$$

$j=1,\ldots,N$. The integration formula of Eq. (VII.4b) not only follows from the summation formula of Eq. (VII.4a) but is in fact *equivalent* to it in view of the Poisson summation formula.

C. Analyticity

The restrictions on the domains of the variables and the parameters in Theorem 3 are not very essential. In fact, we have that for the given wave function $\psi(x,y)$ our difference equation holds as an identity between analytic expressions (with singularities) in the variables x, y and the parameters $\kappa_1,\ldots,\kappa_N$, $\nu_1,\ldots,\nu_N$. Within the parameter and variable regimes stated by the theorem, however, both our wave function and the coefficients of the difference equation are regular.

D. Relation to the infinite Toda chain

The expressions for the coefficients a_n, b_n in the case of a reflectionless Jacobi operator, given by Eqs. (IV.9), (IV.12) (determinantal representations) or by Theorem 1/Corollary 2 (explicit representations in terms of the tau function), induce formulas for the N-soliton solutions of the inifinite Toda chain that are due to Flaschka[10] and Hirota,[15] respectively.

It is well-known that the Jacobi operator $D=a_nT+a_{n-1}T^{-1}+b_n$ may serve as a Lax operator for the infinite Toda chain. In fact, this observation formed the starting point for Flaschka's solution of the equations of motion for this dynamical particle system by means of the *Inverse Scattering Transform*.[10–12] In a nutshell, the state of affairs is the following. Let us assume that the Jacobi operator D is characterized by the spectral data $1>z_1>\cdots>z_N>-1$ $(z_j\neq 0)$, $\nu_1,\ldots,\nu_n>0$ and $r(z)$. Furthermore, let us denote by $a_n(t)$, $b_n(t)$ the coefficients of the Jacobi operator characterized by the time-dependent spectral data $z_j(t)=z_j$, $\nu_j(t)=\nu_j e^{(z_j-z_j^{-1})t}$ ($j=1,\ldots,N$) and $r(z;t)=r(z)e^{(z-z^{-1})t}$, with $t\in\mathbb{R}$. Then the quantities $a_n(t)$, $b_n(t)$ satisfy (Flaschka's version of) the equations of motion for the infinite Toda chain,

$$\dot{a}_n(t)=a_n(t)(b_{n+1}(t)-b_n(t))/2,\quad \dot{b}_n(t)=a_n^2(t)-a_{n-1}^2(t), \tag{VII.5}$$

with initial conditions given by $a_n(0)=a_n$, $b_n(0)=b_n$ $(n\in\mathbb{Z})$.[10–12] In other words, the complicated nonlinear Toda dynamics for the coefficients $a_n(t)$, $b_n(t)$ is related via the (inverse) scattering transform to a simple linear evolution of the corresponding spectral data $z_j(t)$, $\nu_j(t)$, and $r(z;t)$. The fact that the discrete spectral values $z_1,\ldots,z_N$ actually turn out to be time-independent is by no means a coincidence. It is a reflection of a more general phenomenon referred to as *iso-spectrality*, which says that the whole spectrum of the Jacobi operator with coefficients evolving in accordance with the Toda flow (VII.5) does not depend on t.

In the above picture *reflectionless* Jacobi operators correspond to *soliton* solutions of the infinite Toda chain. The formulas for the coefficients a_n, b_n in Eqs. (IV.9), (IV.12) thus give rise—upon plugging in the time-dependence for the normalization constants $\nu_j\to\nu_j e^{(z_j-z_j^{-1})t}$—to a determinantal representation for the N-soliton solution of the Toda chain. These are the well-known solutions found by Flaschka.[10–12] (To be precise, Flaschka actually wrote down only the determinantal formula for $a_n(t)$; the formula for $b_n(t)$ may be found in Ref. 12.) In exactly the same way the closed expressions for the coefficients a_n, b_n in terms of the tau function $\tau(n)$ (V.2a) (cf. Theorem 1 and Corollary 2) are converted into completely explicit formulas for the N-soliton solutions of the infinite Toda chain. The explicit soliton formulas of this type are also well-known and were in fact first introduced by Hirota using direct methods rather than inverse scattering theory.[15,16,12] For more recent developments pertaining to the integration of the Toda equations of motion via inverse scattering techniques we refer, e.g., to Refs. 17,14, and references therein.

In the present paper we concentrated on the solution of the eigenvalue problem for the reflectionless Jacobi operator by means of inverse scattering theory rather than on the solution of the equations of motion for the Toda system. One could nevertheless plug in the time dependence $\nu_j \to \nu_j e^{(z_j - z_j^{-1})t}$ also in the Sato formula for the Jost function of Theorem 1/Corollary 2. This way one arrives at a closed expression for the so-called *Baker function* (see, e.g., Ref. 5) of (the linear problem associated to) the infinite Toda chain. Our original Jost function (i.e., with time-independent spectral data) is in this context often referred to as the *stationary Baker function*. From the point of view of integrable sytems the main result of this paper therefore amounts to an explicit (Sato type) formula for the soliton Baker function of the infinite Toda chain.

E. Bidiagonal reduction: $b_n = 0$

For the special case of a Jacobi operator with zeros on the main diagonal, i.e., with $b_n = 0$ for all $n \in \mathbb{Z}$, the inverse scattering theory leading up to the discrete Gelfand–Levitan–Marchenko equation (cf. Eq. (III.5)) was developed by Case and Chui.[19,20] From the point of view of integrable systems (cf. the previous remark), the specialization to $b_n = 0$ amounts to a reduction from the Toda chain to the Kac–van Moerbeke chain.[21,12,17,14]

It is not difficult to see that the Jost function for the relevant bidiagonal Jacobi operator of the type $D = a_n T + a_{n-1} T^{-1}$ enjoys the symmetry property $\psi_{\text{jost}}(n,-z) = (-1)^n \psi_{\text{jost}}(n,z)$. (This is immediate from the definition of the Jost function, cf. Sec. II.) As a consequence, the spectrum of the Jacobi operator is now evenly distributed around the origin. In particular, the discrete eigenvalues always occur in even pairs. More precisely, in the bidiagonal situation the spectral data possess the following symmetry properties: (i) $r(-z) = r(z)$, (ii) the discrete spectral values are of the form $1 > z_1 > \cdots > z_M > 0 > z_{M+1} > \cdots > z_{2M} > -1$ with $z_{2M+1-j} = -z_j$, and (iii) the associated normalization constants $\nu_1, \ldots, \nu_{2M} > 0$ satisfy the symmetry property $\nu_{2M+1-j} = \nu_j$.

Reversely, if we assume that the spectral data of a reflectionless Jacobi operator $D = a_n T + a_{n-1} T^{-1} + b_n$ possess the above symmetry properties (ii) and (iii) (the first property is now of course trivial since $r(z) \equiv 0$ by assumption), then it follows that the main diagonal vanishes: $b_n = 0$ for all $n \in \mathbb{Z}$. To see this it is convenient to employ the representation for b_n in the second line of Eq. (IV.10) (with $N = 2M$). Indeed, we have from (ii) and (iii) that $k_{2M+1-j}(n) = (-1)^n k_j(n)$, whence the sums on the second line of Eq. (IV.10) vanish (since the jth term cancels against the $(2M+1-j)$th term). A simple way to deduce the symmetry property of the kernel functions $k_j(n)$, is to convince oneself that the original linear system for $k_1(n), \ldots, k_{2M}(n)$ in Eq. (IV.4) $(N = 2M)$ admits a solution with the symmetry property $k_{2M+1-j}(n) = (-1)^n k_j(n)$ when the spectral data meet conditions (ii) and (iii). Indeed, substitution of a solution of the type $(k_1(n), \ldots, k_M(n), (-1)^n k_M(n), \ldots, (-1)^n k_1(n))$ into Eq. (IV.4) reduces the linear system of $2M$ equations in $2M$ unknowns (viz., $k_1(n), \ldots, k_{2M}(n)$) to (two copies of) a linear system of M equations in M unknowns (viz., $k_1(n), \ldots, k_M(n)$).

The upshot is that for the case of a Jacobi operator with zero diagonal the dimension of the determinantal formulas for the coefficients a_n and the Jost function $\psi_{\text{jost}}(n,z)$ given by Eqs. (IV.9) and (IV.13) may be reduced from $2M$ to M. Furthermore, the explicit expressions for the coefficients a_n and the Jost function $\psi_{\text{jost}}(n,z)$ from Theorem 1/Corollary 2 admit in this bidiagonal situation a corresponding simplification. (Specifically, the sums in the relevant tau and chi functions become over subsets of $\{1, \ldots, M\}$ rather than $\{1, \ldots, 2M\}$ (cf. Eqs. (V.2a)–(V.2c)) and the spectral parameters z and $z_1, \ldots, z_M, -z_M, \ldots, -z_1$ enter the formulas via their squares z^2 and $z_1^2, \ldots, z_M^2$). This gives rise to a corresponding reduction of the Sato formula of Theorem 3 for reflectionless difference operators of the form $\mathcal{D} = a(x)\, T + a(x-1)\, T^{-1}$.

For the precise details regarding the resulting parametrization of the coefficients and Jost function for the bidiagonal reflectionless Jacobi operators, as well as the associated Sato formula for reflectionless analytic difference operators, the reader is referred to Ref. 22.

ACKNOWLEDGMENT

Work supported in part by the Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) through Grant No. 1980832.

[1] M. Sato, ''Soliton equations as dynamical systems on an infinite dimensional Grassmann manifolds,'' RIMS Kôkyûroku **439**, 30–46 (1981).
[2] M. Sato and Y. Sato, ''Soliton equations as dynamical systems on infinite dimensional Grassmann manifold,'' in *Nonlinear Partial Differential Equations in Applied Science*, edited by H. Fujita, P. D. Lax, and G. Strang, in *Proceedings of the U.S.–Japan Seminar*, Tokyo, 1982, Lecture Notes in Num. Appl. Anal. (North–Holland, Amsterdam, 1983), Vol. 5, pp. 259–271.
[3] E. Date, M. Kashiwara, M. Jimbo, and T. Miwa, ''Transformation groups for soliton equations,'' in *Nonlinear Integrable Systems: Classical Theory and Quantum Theory*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983), pp. 39–119.
[4] M. Jimbo and T. Miwa, ''Solitons and infinite dimensional Lie algebras,'' Publ. Res. Inst. Math. Sci. **19**, 943–1001 (1983).
[5] G. Segal and G. Wilson, ''Loop groups and equations of KdV type,'' Publ. Math. IHES **61**, 5–65 (1985).
[6] Y. Ohta, J. Satsuma, D. Takahashi, and T. Tokihiro, ''An elementary introduction to Sato theory,'' Prog. Theor. Phys. Suppl. **94**, 210–214 (1988).
[7] V. G. Kac, *Infinite Dimensional Lie Algebras*, 3rd ed. (Cambridge University Press, Cambridge, 1990).
[8] P. van Moerbeke, ''Integrable foundations of string theory,'' in *Lectures on Integrable Systems*, edited by O. Babelon, P. Cartier, and Y. Kosmann-Schwarzbach, in Proceedings of the CIMPA School of Integrable Systems dedicated to the memory of Jean-Louis Verdier, Sophia-Antipolis, 1991 (World Scientific, Singapore, 1994), pp. 163–267.
[9] J. F. van Diejen and A. N. Kirillov, ''A combinatorial formula for the associated Legendre functions of integer degree,'' Adv. Math. (to be published).
[10] H. Flaschka, ''On the Toda lattice. II.,'' Prog. Theor. Phys. **51**, 703–716 (1974).
[11] M. Toda, ''Studies of a nonlinear lattice,'' Phys. Rep. **18**, 1–124 (1975).
[12] M. Toda, *Theory of Nonlinear Lattices*, 2nd ed. (Springer-Verlag, Berlin, 1989).
[13] G. Teschl, ''Trace formulas and inverse spectral theory for Jacobi operators,'' Commun. Math. Phys. **196**, 175–202 (1998).
[14] G. Teschl, ''On the Toda and Kac–Van Moerbeke hierarchies,'' Math. Z. **231**, 325–344 (1999).
[15] R. Hirota, ''Exact *N*-soliton solutions of a nonlinear lumped network equation,'' J. Phys. Soc. Jpn. **35**, 286–288 (1973).
[16] R. Hirota, ''Direct methods in soliton theory,'' in *Solitons*, Topics in Current Physics 17, edited by R. K. Bullough and P. J. Caudry (Springer-Verlag, Berlin, 1980), pp. 157–176.
[17] F. Gesztesy, H. Holden, B. Simon, and Z. Zhao, ''On the Toda and Kac–Van Moerbeke systems,'' Trans. Am. Math. Soc. **339**, 849–868 (1993).
[18] I. G. Macdonald, *Symmetric Foundations and Hall Polynomials*, 2nd ed. (Oxford University Press, Oxford, 1995).
[19] K. M. Case and S. C. Chui, ''The discrete version of the Marchenko equation in the inverse scattering problem,'' J. Math. Phys. **14**, 1643–1650 (1973).
[20] K. M. Case, ''The discrete inverse scattering problem in one dimension,'' J. Math. Phys. **15**, 143–146 (1974).
[21] M. Kac and P. van Moerbeke, ''On an explicitly soluble system of nonlinear differential equations related to certain Toda lattices,'' Adv. Math. **16**, 160–169 (1975).
[22] J. F. van Diejen and A. N. Kirillov, ''Formulas for *q*-spherical functions using inverse scattering theory of reflectionless Jacobi operators,'' Commun. Math. Phys. (to be published).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Existence and multiplicity results for massive particles trajectories in a universe with boundary

Anna Germinario[a)]
Dipartimento Interuniversitario di Matematica, Università degli Studi di Bari, Via E. Orabona, 4, 70125 Bari BA, Italy

Fabio Giannoni[b)]
Dipartimento di Matematica e Fisica, Università di Camerino, Via Madonna delle Carceri, 20, 62032 Camerino MC, Italy

(Received 8 March 1999; accepted for publication 14 April 1999)

In an open time-convex region Λ of a strongly causal Lorentzian manifold $(\mathcal{M},g)$, we consider an event p and a timelike, injective curve γ. We look for geodesics connecting p and γ in Λ and satisfying the conservation law $g(z)[\dot{z},\dot{z}]=-E$ for a fixed $E>0$. It is already known that such geodesics are the stationary points of the arrival time functional τ. Our main result is to prove the existence of a decreasing flow for τ, by means of a shortening procedure. This makes possible to apply to τ global variational methods obtaining existence and multiplicity results (using the Ljusternik–Schnirelmann category theory) and also to develop a Morse theory. [S0022-2488(99)03210-7]

I. INTRODUCTION AND STATEMENT OF THE RESULTS

This paper deals with the study of timelike geodesics, having a prescribed parameterization proportional to the arc length, on a noncomplete, strongly causal Lorentzian manifold representing a universe with boundary.

In recent years variational methods have been applied to the study of Lorentzian geodesics. In particular, a large number of papers concerns variational principles which allow us to get existence and multiplicity results for geodesics connecting a point with a curve (see, e.g., Refs. 1, 2 and 3).

In Ref. 4, Kovner stated a variational principle for timelike and lightlike geodesics joining a source with a receiver in a space–time $(\mathcal{M},g)$. The source is represented by a timelike curve γ and the observer by an event p of $\mathcal{M}$. The principle states that among the future pointing curves $z:[0,1]\rightarrow\mathcal{M}$ joining p and γ and satisfying

$$g(z)[\dot{z},\dot{z}]=-E \tag{1}$$

for some $E\geqslant 0$, geodesics are the critical points of the *arrival time* functional τ defined (assuming that γ is injective) in the following way:

$$\tau(z)=\gamma^{-1}(z(1)). \tag{2}$$

A rigorous mathematical proof of this principle can be found in Ref. 3 for the lightlike case and in Ref. 2 for the timelike case. In Ref. 5, assuming that $\mathcal{M}$ admits a time function T the authors develop a Ljusternik–Schnirelmann theory for light rays (namely they take $E=0$) obtaining some multiplicity results depending on the topology of the space of the lightlike curves joining p and γ in a manifold with boundary. They look for the critical points of the functional

[a)]Electronic mail: germinar@pascal.dm.uniba.it
[b)]Electronic mail: giannoni@campus.unicam.it

$$Q(z)=\int_0^1 \langle \dot{z},\nabla T(z)\rangle^2 ds$$

that are lightlike pregeodesics and they assume that $\mathcal{M}$ is not complete which is equivalent to search geodesics lying in some open subset Λ of a complete manifold. Moreover in Ref. 6 a Morse theory has been developed for the same problem. In Ref. 2 the variational principle for timelike curves involving the functional τ is examined. Assuming that $\mathcal{M}$ is a manifold without boundary, existence and multiplicity results are obtained. This is done after proving that the critical points of τ on the space of the future pointing curves satisfying (1) are timelike geodesics.

The approach based on the use of the functional τ has some advantages; for example, the critical points of Q have not a clear physical meaning and their Euler–Lagrange equation is rather complicated. This makes the proof of the Morse theory in Ref. 6 quite involved. For this reason in Ref. 7, using τ, the Morse relations are stated for timelike geodesics and, with a limit process, also for lightlike geodesics.

Our aim in this paper is to prove the existence and the multiplicity (using Ljusternik–Schnirelmann category theory and Morse theory) of timelike geodesics z joining a point p and a curve γ, satisfying (1) for some $E>0$, when the space–time has a boundary. More precisely, we look for geodesics z lying in an open subset Λ of $\mathcal{M}$ with topological boundary $\partial\Lambda$. We are motivated in this study by the fact that some relevant physical examples of space–times have a boundary (e.g., Schwarzchild, Reissner–Nordström, Kerr space–times).

We point out that, as it will be clear in the sequel, the presence of $\partial\Lambda$ makes the problem more difficult. Indeed in Refs. 2 and 7, where the case of manifolds without boundary is studied, the functional τ is defined on a manifold of curves of class $H^{1,1}$, in order to prove that it satisfies the Palais–Smale condition. On the other hand, when Λ has a boundary, the manifold of the curves joining p and γ is not complete and τ does not satisfy the Palais–Smale condition because of the existence of sequences of curves approaching $\partial\Lambda$. A technique based on a penalization argument (used, for example, in Refs. 5 and 6 with curves of class $H^{1,2}$) fails with curves of class $H^{1,1}$. More precisely, if f is a functional defined on a set of curves of a manifold with boundary, we can add to f a term (depending on a parameter δ) which goes to infinity on each sequence of curves approaching $\partial\Lambda$ (see, e.g., Ref. 5, Lemma 7.2). Then a limit process allows us to get the critical points of f.

Unfortunately, this approach works when we deal with curves of class $H^{1,m}$ with $m>1$. When, as in our case, the natural space for the problem is $H^{1,1}$, it is not possible to prove that the penalized functionals satisfy useful completeness and compactness properties because we are not able to prove that the penalization term goes to infinity near the boundary.

In this paper we shall use a completely different approach to the problem. Indeed, in critical point theorems and in the Ljusternik–Schnirelmann theory for a functional f, the Palais–Smale condition is used for the construction of a flow along which f is strictly decreasing. Here we shall directly prove the existence of a decreasing flow for the functional τ, by means of a shortening method.

This procedure has been firstly introduced for studying Riemannian geodesics connecting two fixed points (see, e.g., Ref. 8). Recently, in Ref. 9, the authors have used a similar technique in stating a Morse theory for lightlike geodesics connecting a point with a curve. Here we shall deal with the timelike case. The main difference with respect to Ref. 9 is a different way to prove the existence of minimizers between a point and an integral curve of the vector field giving the time orientation on $\mathcal{M}$.

We wish to point out that the shortening method has a further advantage; it allows us to work in a very general context. Indeed, (a) the Lorentzian manifold is not assumed to be stably causal but only time orientable; (b) the boundary $\partial\Lambda$ of the open region Λ where we search timelike geodesics is not assumed neither smooth nor having a spacelike normal vector at any point; (c) the curve γ is not assumed to be a closed embedding of $\mathbf{R}$. Moreover, the shortening method (far from the critical points) makes simpler the proof of the Morse relations (see Ref. 7 for the case without boundary).

We consider a Lorentzian manifold $(\mathcal{M},g)$ whose metric tensor g will be denoted by

$$g(z)[\cdot,\cdot]=\langle \cdot,\cdot \rangle$$

for any $z\in\mathcal{M}$. It is not restrictive to assume that there exists a time orientation W on $\mathcal{M}$ namely a smooth timelike vector field W on $\mathcal{M}$. We recall that a vector $v\in T_p\mathcal{M}$, $p\in\mathcal{M}$ is said *future pointing* if

$$\langle W(z),v\rangle<0,$$

it is called *causal* if

$$\langle W(z),v\rangle\leq 0.$$

A smooth curve $y:I\rightarrow\mathcal{M}$ is *future pointing* (respectively *causal*) if $\dot{y}(s)$ is future pointing (respectively causal) for any $s\in I$. (We refer to Refs. 10 and 11 for the basic notions of Lorentzian geometry.)

The main results of this paper will be stated assuming that $\mathcal{M}$ is *strongly causal*. This means that, for any $q\in\mathcal{M}$, each future pointing causal curve starting arbitrarily close to q and leaving some fixed neighborhood of q, can not return arbitrarily close to q (for more details see Refs. 10 and 11).

Let Λ be an open, connected subset of $\mathcal{M}$, $p\in\Lambda$, $\gamma:\mathbf{R}\rightarrow\Lambda$ a smooth, timelike curve such that $p\notin\gamma(\mathbf{R})$. With respect to the orientation W, we assume that

$$\gamma \text{ is future pointing.} \tag{3}$$

From now on, we fix $E>0$. To prove the existence of future pointing, timelike curves joining p and γ we need the following assumption:

$$\begin{gathered}\text{there exists a } C^1\text{-piecewise, future pointing, curve } z:[0,1]\rightarrow\Lambda\\ \text{joining } p \text{ and } \gamma \text{ and satisfying } \langle\dot{z},\dot{z}\rangle=-E\\ \text{on any interval where } z \text{ is } C^1.\end{gathered} \tag{4}$$

Moreover we need a time-convexity assumption on the closure $\overline{\Lambda}$ of Λ. We assume that

$$\begin{gathered}\overline{\Lambda} \text{ is time-convex, i.e., all the timelike geodesics in } \Lambda\cup\partial\Lambda,\\ \text{whose end points are in } \Lambda, \text{ are entirely contained in } \Lambda.\end{gathered} \tag{5}$$

Here $\partial\Lambda$ denotes the topological boundary of Λ. To be sure that the arrival time functional τ defined in (2) is well defined we need that

$$\gamma \text{ is injective.} \tag{6}$$

In the following, for any $q\in\Lambda$ we shall denote by $\gamma_q:I\rightarrow\mathcal{M}$, (where $I\subset\mathbf{R}$ is an interval) the maximal integral curve of W starting at q, i.e., the maximal solution of the Cauchy problem

$$\begin{cases}\dot{\eta}=W(\eta)\\ \eta(0)=q.\end{cases}$$

In order to get a decreasing flow for the functional τ, we shall prove the existence of minimizers in Λ between events and timelike curves. The following hypothesis will be necessary:

$$\gamma \text{ is an integral curve of } W, \tag{7}$$

$$\forall q \in \Lambda, \quad \gamma_q(s) \in \Lambda, \quad \forall s \leqslant 0. \tag{8}$$

We introduce now a space of broken timelike geodesics which will be used as a space of trial curves. We set

$$\mathcal{B}^+_{p,\gamma,E}(\Lambda) = \{z:[0,1] \to \Lambda\} | z \text{ is a } C^2\text{-piecewise curve, such that}$$

$$z(0)=p, \ z(1) \in \gamma(\mathbf{R}) \text{ and, on any interval } I \text{ where } z \text{ is } C^2, \ z \text{ is}$$

$$\text{a future pointing geodesic satisfying } \langle \dot{z}, \dot{z} \rangle = -E\}.$$

We want to relate the number of future pointing timelike geodesics satisfying (1) with the Ljusternik–Schnirelmann category of $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$. In next sections we shall prove the homotopy equivalence between $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ (endowed with the uniform topology) and the Sobolev space of future pointing geodesics satisfying (1) and joining p and γ. Moreover a timelike extention of the Fermat principle holds (see Ref. 2).

Now we can state our last assumption which is equivalent to the ones formulated in Refs. 2 and 7 (as shown in Sec. II). For any $c \in \mathbf{R}$ we set

$$\tau^c = \{z \in \mathcal{B}^+_{p,\gamma,E}(\Lambda) | \tau(z) \leqslant c\}.$$

Definition I.1: Let c be a real number. We say that $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ *is c-precompact if any sequence* $(z_n)_{n \in \mathbf{N}} \subset \tau^c$ *has a uniformly convergent subsequence in* Λ*, up to reparameterization. The functional* τ *is said pseudocoercive on* $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ *if* $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ *is c-precompact for any* $c \in \mathbf{R}$.

We note that, if τ is pseudocoercive on $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ and if γ is an integral curve of W, then τ is bounded from below on $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$.

We shall prove the following existence result:

Theorem I.2: *Let* $(\mathcal{M},g)$ *be a strongly causal Lorentzian manifold. Let* $E>0,p,\gamma,\Lambda$ *satisfy assumptions (3)–(8) and* τ *be pseudocoercive on* $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$*. Then there exists at least one future pointing, timelike geodesic z joining p and* γ *in* Λ*, satisfying (1) and minimizing* τ *on the set of the future pointing curves joining p and* γ *and satisfying (1).*

We shall also prove a multiplicity result by using the Ljusternik–Schnirelmann theory. We refer to Ref. 12 for its main properties. Here we only recall that, if X is a topological space and Y is a subspace of X, the *Ljusternik–Schnirelmann category* of Y in X, denoted by $\mathrm{cat}(Y,X)$, is the minimal integer number (possibly infinite) of closed, contractible subsets of X covering Y. We shall denote by $\mathrm{cat}(X)$ the category of X in itself.

Theorem I.3: *Let* $\mathcal{M},E>0,p,\gamma,\Lambda$ *as in Theorem I.2. Assume that* τ *is pseudocoercive on* $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$*. Then there exists at least* $\mathrm{cat}(\mathcal{B}^+_{p,\gamma,E}(\Lambda))$ *future pointing, timelike geodesics z joining p and* γ *in* Λ *and satisfying (1).*

The shortening method allows us to state also a Morse theory for this problem. To this aim, we recall some definitions.

Definition I.4: Let $(\mathcal{M},g)$ *be a Lorentzian manifold, and z:* $[0,1] \to \mathcal{M}$ *be a geodesic. A smooth vector field* ζ *along z is called Jacobi field if it satisfies the equation*

$$D_s^2 \zeta + R(\zeta,\dot{z})\dot{z} = 0,$$

where R is the curvature tensor of the metric g. A point $z(s)$*,* $s \in]0,1]$ *is said to be conjugate to* $z(0)$ *along z if there exists a nonvanishing Jacobi field* ζ *along* $z_{|[0,s]}$ *such that*

$$\zeta(0) = \zeta(s) = 0. \tag{9}$$

The multiplicity of the conjugate point $z(s)$ *is the maximal number of linearly independent Jacobi fields satisfying (9).*

Definition I.5: Let $z:[0,1]\to\mathcal{M}$ *be a geodesic. The index* $\mu(z)$ *of* z *is the number of conjugate points* $z(s)$, $s\in]0,1]$ *to* $z(0)$, *counted with their multiplicity.*

Let X be a topological space and $\mathcal{K}$ a field. For any $l\in\mathbf{N}$ let $H_l(X;\mathcal{K})$ be the lth homology group of X with coefficients in $\mathcal{K}$. Since $\mathcal{K}$ is a field, then $H_l(X;\mathcal{K})$ is a vector space whose dimension $\beta_l(X;\mathcal{K})$ (eventually $+\infty$) is called the lth *Betti number of* X (with coefficients in $\mathcal{K}$). The *Poincaré polynomial* $\mathcal{P}(X;\mathcal{K})$ is defined as the following formal series:

$$\mathcal{P}(X;\mathcal{K})(\kappa)=\sum_{l\in\mathbf{N}}\beta_l(X;\mathcal{K})\kappa^l.$$

Let $\mathcal{G}^+_{p,\gamma,E}(\Lambda)$ be the set of the future pointing geodesics satisfying (1), joining p and γ and whose image is contained in Λ.

Theorem I.6: *Let* $(\mathcal{M},g)$ *be a strongly causal Lorentzian manifold. Let* $E>0,p,\gamma,\Lambda$ *satisfy (3)–(8). Moreover, assume that*

(1) *for any geodesic* $z\in\mathcal{L}^+_{p,\gamma,E}$, $z(1)$ *is nonconjugate to* $z(0)=p$;
(2) τ *is pseudocoercive on* $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$.

Then, for any field $\mathcal{K}$, *there exists a formal series* $S(\kappa)$ *with coefficients in* $\mathbf{N}\cup\{+\infty\}$, *such that*

$$\sum_{z\in\mathcal{G}^+_{p,\gamma,E}(\Lambda)}\kappa^{\mu(z)}=\mathcal{P}(\mathcal{B}^+_{p,\gamma}(\Lambda);\mathcal{K})(\kappa)+(1+\kappa)S(\kappa). \tag{10}$$

We point out that Theorem I.6 has been proved in Ref. 7 when $\mathcal{M}$ is a manifold without boundary. Moreover, we refer to Ref. 9 for a discussion and some consequences of the previous theorem, which, up to obvious changes hold also for the timelike case. Here we point out that relations (10) link the number card $(\mathcal{G}^+_{p,\gamma,E}(\Lambda))$ of geodesics in $\mathcal{G}^+_{p,\gamma,E}(\Lambda)$ to the topology of the space $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$. Indeed, setting $\kappa=1$ in (10) we get

$$\text{card}(\mathcal{G}^+_{p,\gamma,E}(\Lambda))=\sum_{l=0}^{\infty}\beta_l(\mathcal{B}^+_{p,\gamma,E}(\Lambda);\mathcal{K})+2S(1). \tag{11}$$

As $S(1)$ is non-negative, by (11) we can deduce the classical Morse inequalities (see Ref. 8) and information about the oddity of card$(\mathcal{G}^+_{p,\gamma,E}(\Lambda))$.

In Sec. II we shall discuss the existence of minimizers between a point and a future pointing timelike curve in $\mathcal{M}$. In Sec. III we shall describe the shortening procedure and prove the main theorems.

II. MINIMIZERS FOR THE ARRIVAL TIME

Let W be the smooth vector field giving an orientation on $\mathcal{M}$. Then, we can equip $\mathcal{M}$ with the Riemannian structure, given by

$$\langle\zeta,\zeta\rangle_R=\langle\zeta,\zeta\rangle-2\,\frac{\langle\zeta,W(z)\rangle^2}{\langle W(z),W(z)\rangle} \tag{12}$$

for any $z\in\mathcal{M}$ and $\zeta\in T_z\mathcal{M}$. It is not difficult to prove that (12) defines a bilinear form on $\mathcal{M}$ whose positivity follows from the wrong way Schwartz inequality (see Ref. 11). In the sequel we shall denote by d_R the distance induced by (12). The metric (12) allows us to define in an intrinsic way the Sobolev space,

$$H^{1,2}([0,1],\Lambda)=\left\{z\in\text{Ac}([0,1],\Lambda)\,\middle|\,\int_0^1\langle\dot z,\dot z\rangle_R ds<+\infty\right\},$$

where Ac([0,1],Λ) is the set of absolutely continuous curves from [0,1] to Λ (with respect to d_R). It is well known that $H^{1,2}([0,1],\Lambda)$ is a smooth Hilbert manifold. Now we set

$$\Omega^{1,2}_{p,\gamma}=\Omega^{1,2}_{p,\gamma}(\Lambda)=\{z\in H^{1,2}([0,1],\Lambda)\,|\,z(0)=p\quad z(1)\in\gamma(\mathbf{R})\}.$$

It is not difficult to see that $\Omega^{1,2}_{p,\gamma}$ is a smooth manifold. For any $z\in\Omega^{1,2}_{p,\gamma}$ the tangent space at z can be identified with

$$T_z\Omega^{1,2}_{p,\gamma}=\{\zeta\in H^{1,2}([0,1],T\Lambda)\,|\,\zeta(s)\in T_{z(s)}\mathcal{M}\quad \zeta(0)=0\quad \zeta(1)\parallel\dot\gamma(\tau(z))\}.$$

As we look for future pointing timelike geodesics satisfying (1), we set

$$\mathcal{L}^+_{p,\gamma,E}=\{z\in\Omega^{1,2}_{p,\gamma}\,|\,\langle\dot z,\dot z\rangle=-E,\quad \langle W(z),\dot z\rangle<0\ \text{ a.e.}\}.$$

By (4)

$$\mathcal{L}^+_{p,\gamma,E}\neq\emptyset$$

and following the results of Refs. 2 and 7, it can be proved that, for any $E>0$, $\mathcal{L}^+_{p,\gamma,E}$ is a C^1 Hilbert manifold whose tangent space at $z\in\mathcal{L}^+_{p,\gamma,E}$ can be identified with

$$T_z\mathcal{L}^+_{p,\gamma,E}=\{\zeta\in T_z\Omega^{1,2}_{p,\gamma}\,|\,\langle\dot z,D_s\zeta\rangle=0\ \text{ a.e.}\}.$$

The natural scalar product on $\mathcal{L}^+_{p,\gamma,E}$ is

$$\langle\zeta,\zeta\rangle_1=\int_0^1\langle D^R_s\zeta,D^R_s\zeta\rangle_R ds+\int_0^1\langle\zeta,\zeta\rangle_R ds \tag{13}$$

for any $z\in\mathcal{L}^+_{p,\gamma,E}$ and $\zeta\in T_z\mathcal{L}^+_{p,\gamma,E}$, where D^R_s denotes the covariant derivative with respect to (12) evaluated along $\dot z$.

It has been proven in Refs. 2 and 7 the following timelike extension of the Fermat principle.

Proposition II.1: Let τ *be as in (2),* $E>0$ *and* $z\in\mathcal{L}^+_{p,\gamma,E}$*. Then z is a critical point for* τ *in* $\mathcal{L}^+_{p,\gamma,E}$ *if and only if z is a geodesic.*

In the sequel, for any $p_*\in\mathcal{M}$ and any future pointing, injective, timelike curve $\gamma_*:]\alpha_*,\beta_*[\to\Lambda$ we shall use the following notation:

$$\mathcal{L}^+_{p_*,\gamma_*,E}([a_1,a_2],\Lambda)=\{z\in\Omega^1_{p_*,\gamma_*}([a_1,a_2],\Lambda)\,|\,\langle\dot z,\dot z\rangle=-E,\ \langle W(z),\dot z\rangle<0\ \text{ a.e.}\}.$$

We shall again denote by τ the functional defined by

$$\gamma_*^{-1}(z(a_2))$$

on $\mathcal{L}^+_{p_*,\gamma_*,E}([a_1,a_2],\Lambda)$.

Remark II.2: For any $z\in\mathcal{M}$ there exists a neighborhood $\mathcal{U}_z$ of z and a coordinate system $\varphi=(x_1,\dots,x_{N-1},t)$ ($N=\dim\mathcal{M}$) on $\mathcal{U}_z$ such that

$$W=\frac{\partial}{\partial t},\quad \mathcal{U}_z=\Sigma\times]a_1,a_2[,$$

where Σ is a spacelike hypersurface parameterized by $x_1,\dots,x_{N-1}$. Moreover in the coordinate $x=(x_1,\dots,x_{N-1})$ and $t\in]a_1,a_2[$ the metric g is given by

$$g(x,t)[(\xi,\theta)(\xi,\theta)]=\langle\alpha(x,t)\xi,\xi\rangle_0+2\langle\delta(x,t),\xi\rangle_0\theta-\beta(x,t)\theta^2, \tag{14}$$

where $(\xi,\theta)\in T_{(x,(a_1+a_2)/2)}\Sigma\times\mathbf{R}$, $\langle\cdot,\cdot\rangle_0$ is the (positive definite) restriction of g to Σ, α is a smooth, symmetric positive definite operator, δ is a smooth vector field on Σ and β is a smooth positive real function such that

$$\beta(x,t)=-g(x,t)[W,W] \tag{15}$$

(see Ref. 9 for further details).

Remark II.3: From now on, unless it is to normalize the vector field W and to take a reparameterization of γ we can assume that

$$\langle W(z),W(z)\rangle=-1 \quad \forall z\in\mathcal{M}, \tag{16}$$

so that by (7)

$$\langle\dot\gamma,\dot\gamma\rangle=-1 \quad \forall s\in\mathbf{R}.$$

Therefore in the coordinate systems $(x_1,\ldots,x_{N-1},t)$ (see Remark II.2), a curve $z=(x,t)\in\mathcal{L}^+_{p,\gamma,E}$ if and only if

$$\dot t=\langle\delta(x,t),\dot x\rangle_0+\sqrt{\langle\delta(x,t),\dot x\rangle_0^2+\langle\alpha(x,t)\dot x,\dot x\rangle_0+E} \tag{17}$$

as (15) becomes

$$\beta(x,t)=1. \tag{18}$$

Moreover, in such a coordinate system, any integral curve of W can be written as

$$s\mapsto(\overline{x},s+\overline{t})$$

for some $(\overline{x},\overline{t})$.

Proposition II.4: Let $q\in\mathcal{M}$. For any sufficiently small compact neighborhood U of q there exists $\rho=\rho(q,U)>0$ and $l=l(q,U)>0$ such that, for any integral curve γ_* of W such that

$$0<d_R(q,\mathrm{Im}\,\gamma_*)\leq\rho$$

and for any interval $[a_1,a_2]$ such that

$$a_2-a_1<l,$$

there exists a future pointing geodesic $z:[a_1,a_2]\to U$ joining q and γ_* and minimizing τ on $\mathcal{L}^+_{q,\gamma_*,E}([a_1,a_2],\mathcal{M})$.

Proof: Let U be a compact neighborhood of q where the metric g can be written as in (14) of Remark II.2. Assume that

$$U=V\times]b_1,b_2[,\quad q=(x_1,0)\in U$$

for some $x_1\in V$. By Remark II.3, there exists $\alpha_*,\beta_*\in\mathbf{R}$ such that

$$\gamma_*(s)=(x_2,t_2+s)\in U \quad s\in]\alpha_*,\beta_*[$$

for some $(x_2,t_2)\in U$. By Proposition II.1, we have to minimize the functional

$$F(x)=\int_{a_1}^{a_2}[\langle\delta(x,t_x),\dot x\rangle_0+\sqrt{\langle\delta(x,t_x),\dot x\rangle_0^2+\langle\alpha(x,t)\dot x,\dot x\rangle_0+E}]ds$$

on the manifold $\Omega^{1,2}([a_1,a_2],V)$ of the curves of class $H^{1,2}([a_1,a_2],V)$ joining x_1 and x_2, where t_x is the solution of the Cauchy problem

$$\begin{cases} \dot{t}=\langle \delta(x,t),\dot{x}\rangle_0+\sqrt{\langle \delta(x,t),\dot{x}\rangle_0^2+\langle \alpha(x,t)\dot{x},\dot{x}\rangle_0+E} \\ t(0)=0. \end{cases} \tag{19}$$

As the support of any curve $z=(x,t_x)$ is contained in the compact subset U, thanks to the Schwartz inequality

$$\frac{|\langle \delta(x,t_x),\dot{x}\rangle_0|}{\sqrt{\langle \alpha(x,t_x)\dot{x},\dot{x}\rangle_0}}\leqslant c_1$$

for some $c_1>0$, so that it is not difficult to show that there exists $c_2,c_3>0$ such that

$$c_2\int_{a_1}^{a_2}\sqrt{\langle \dot{x},\dot{x}\rangle_0}ds\leqslant F(x)\leqslant c_3\int_{a_1}^{a_2}\sqrt{\langle \dot{x},\dot{x}\rangle_0}ds+\sqrt{E}(a_2-a_1) \tag{20}$$

for any curve $x\in\Omega^{1,2}([a_1,a_2],V)$. Then, by the first inequality of (20), F is bounded from below.

Now let $(x_n)_{n\in\mathbf{N}}$ be a minimizing sequence for F. We prove that there exists $M>0$ such that for any $n\in\mathbf{N}$,

$$d_0(x_n(s_n),\partial V)\geqslant M, \tag{21}$$

where s_n is defined by

$$d_0(x_n(s_n),\partial V)=\min_{s\in[a_1,a_2]} d_0(x_n(s),\partial V)$$

and d_0 denotes the distance induced by $\langle\cdot,\cdot\rangle_0$. Indeed if (21) were not true, up to a subsequence,

$$\lim_{n\to+\infty} d_0(x_n(s_n),\partial V)=0. \tag{22}$$

By the first inequality of (20), there results

$$F(x_n)\geqslant c_2\int_{a_1}^{a_2}\sqrt{\langle \dot{x}_n,\dot{x}_n\rangle_0}ds\geqslant c_2 d_0(x_1,x_n(s_n)). \tag{23}$$

By (22), for n sufficiently large,

$$d_0(x_1,x_n(s_n))\geqslant\frac{d_0(x_1,\partial V)}{2}.$$

Then in (23) we get

$$\inf F\geqslant c_2\frac{d_0(x_1,\partial V)}{2}. \tag{24}$$

On the other hand, let $y:[a_1,a_2]\to V$ the geodesic chord joining x_1 and x_2 with respect to $\langle\cdot,\cdot\rangle_0$. By the second inequality of (20), there results

$$F(y)\leqslant c_3\int_{a_1}^{a_2}\sqrt{\langle \dot{y},\dot{y}\rangle_0}+\sqrt{E}(a_2-a_1)=c_3 d_0(x_1,x_2)+\sqrt{E}(a_2-a_1). \tag{25}$$

By opportunely choosing ρ and l we get by (25)

$$\inf F \leqslant c_2 \frac{d_0(x_1, \partial V)}{4}$$

in contradiction with (24). This proves that each minimizing sequence is uniformly far from ∂V and, by the Ekeland variational principle (see Ref. 13), the existence of a minimizing Palais–Smale sequence (with respect to the $H^{1,1}$-norm). Since F satisfies the Palais–Smale condition with respect to the $H^{1,1,}$-norm (see Ref. 2), we get the existence of a minimum point of F whose support is contained in V. □

As a first consequence of Proposition II.4 we can prove the equivalence of Definition I.1 with the corresponding ones of Refs. 2 and 7 (where the pseudocoercivity is assumed on $\mathcal{L}^+_{p,\gamma,E}$). Indeed any curve $z \in \mathcal{L}^+_{p,\gamma,E}$ can by uniformly approximated by a sequence $(z_n)_{n\in\mathbf{N}} \subset \mathcal{B}^+_{p,\gamma,E}(\Lambda)$ (see Ref. 9). Then it is not difficult to prove the following Proposition.

Proposition II.5: The functional τ is pseudocoercive on $\mathcal{L}^+_{p,\gamma,E}$ if and only if it is pseudocoercive on $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$.

The following proposition has already been proven in Ref. 7. However, we give here a sketch of the proof. We shall denote by $l(z)$ the length of z with respect to the Riemann structure (12) for any $z \in H^{1,2}([0,1],\Lambda)$.

Proposition II.6: Let $\mathcal{M}$ be strongly causal and τ be pseudocoercive on $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$. Then for any $c \in \mathbf{R}$ there exists $D(c)>0$ such that for any $z \in \mathcal{L}^+_{p,\gamma,E}$,

$$\tau(z) \leqslant c \Rightarrow l(z) \leqslant D(c).$$

Proof: Assume by contradiction the existence of a sequence $(z_m)_{m\in\mathbf{N}}$ in $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ such that

$$\tau(z_m) \leqslant c$$

and

$$\lim_{m\to+\infty} l(z_m) = +\infty. \tag{26}$$

As τ is pseudocoercive, there exists a curve $z:[0,1]\to\mathcal{M}$ such that

$$\lim_{m\to+\infty} \hat{z}_m = z \quad \text{uniformly in } [0,1], \tag{27}$$

where $\hat{z}_m$ is a reparameterization of z_m. Since $\mathcal{M}$ is strongly causal and z is the uniform limit of a sequence of causal curves, z is injective or constant. As $p \notin \gamma(\mathbf{R})$, z is nonconstant so it is injective. Moreover $z([0,1])$ is compact, then it can be covered by a finite number of local charts where the metric can be written as in (14). This makes it possible, to construct a smooth map T on a relatively compact neighborhood $\mathcal{U}$ of $z([0,1])$ such that for any $q \in \mathcal{U}$,

$$\langle \nabla T(q), \nabla T(q)\rangle < 0, \quad \langle \nabla T(q), W(q)\rangle < 0,$$

and, for m sufficiently large and $s \in [0,1]$,

$$\langle \nabla T(z_m), \dot{z}_m\rangle > 0.$$

By (27),

$$T(z_m(1)) \quad \text{is bounded.} \tag{28}$$

Moreover,

$$T(z_m(1)) - T(p) = \int_0^1 \langle \nabla T(z_m), \dot{z}_m\rangle ds, \tag{29}$$

and, by the choice of the orientation of ∇T, there exists $\nu_0 > 0$ such that

$$\langle \nabla T(z_m), \dot{z}_m \rangle \geq \nu_0 \sqrt{\langle \dot{z}_m, \dot{z}_m \rangle_R} \tag{30}$$

for any $s \in [0,1]$. Then, by (28), (29), (30), there results that $l(z_m)$ is bounded, in contradiction with (26). □

Proposition II.7: Let $E, W, \Lambda, p, \gamma: \mathbf{R} \to \Lambda$ satisfy assumptions (3)–(8), τ be pseudocoercive on $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ and fix $c \in \mathbf{R}$. Then there exists $\rho^* = \rho^*(c) > 0$ and $l^* = l^*(c) > 0$ satisfying the following property: for any $z \in \tau^c \cap \mathcal{L}^+_{p,\gamma,E}(\Lambda)$, $[a_1, a_2] \subset [0,1]$ with $a_2 - a_1 \leq l^*$, $z_1 = z(a_1)$, $z_2 = z(a_2) \in z([0,1])$ with $d_R(z_1, z_2) \leq \rho^*$ and for any $\hat{z}_1 \in \gamma_1(]\alpha_1, 0])$ with $d_R(\hat{z}_1, z_1) \leq \rho^*$, there exists a unique future pointing geodesic w satisfying (1) and such that

(1) $w(a_1) = \hat{z}_1$;
(2) $w(a_2) \in \gamma_2(]\alpha_2, \beta_2[)$;
(3) $w(s) \in \Lambda$ *for any* $s \in [a_1, a_2]$;
(4) $\tau(w) = \inf\{\tau(y) | y \in \mathcal{L}^+_{\hat{z}_1, \gamma_2, E}([a_1, a_2], \Lambda)\}$;

where $\gamma_i :]\alpha_i, \beta_i[\to \Lambda$, $i = 1,2$, denotes the maximal integral curve W in Λ starting at z_i.

Proof: As τ is pseudocoercive, there exists a compact subset K of $\bar{\Lambda}$ such that

$$z([0,1]) \subset K \quad \forall z \in \tau^c \cup \mathcal{L}^+_{p,\gamma,E}(\Lambda).$$

We can consider a finite number of open subsets $U_1, \ldots, U_m$ of $\mathcal{M}$ covering K, such that $\bar{U}_i$ is compact and each $\bar{U}_i$ satisfies the properties of Remarks II.2 and II.3. If ρ_* is sufficiently small, $z_1, z_2, \hat{z}_1 \in U_i \cap \Lambda$ for some $i = 1,...,m$. Then, by Proposition II.4 there exists a minimizer w in $\mathcal{L}^+_{\hat{z}_1, \gamma_2, E}([a_1, a_2], U_i)$. Since $\cup_{i=1}^m \bar{U}_i$ is compact, as the exponential map is locally invertible and as w is the minimum point of τ, the minimizing geodesic is unique provided that ρ^* is sufficiently small. It remains to be proved that the minimizer w is included in Λ. Note that, by Remarks II.2, II.3, setting $z(s) = (x(s), t(s))$ in $[a_1, a_2]$, it is possible to prove the existence of a curve $\hat{z} \in \mathcal{L}^+_{\hat{z}_1, \gamma_2, E}([a_1, a_2], U_i)$ having the same spatial component x of z and such that (using comparison theorems in ordinary differential equations) $\tau(\hat{z}) \leq \tau(z) = 0$. Then as w is the minimum point of τ, also $\tau(w) \leq 0$, so that by (8), $w(a_2) \in \Lambda$. Moreover there exists two continuous maps $\theta_1, \theta_2 : [0,1] \to U_i \cap \Lambda$ such that

(a) for any $\lambda \in [0,1]$, $\theta_2(\lambda)$ is in the future of $\theta_1(\lambda)$;
(b) $\theta_1(0) = \hat{z}_1$, $\theta_2(0) = w(a_2)$;
(c) $\theta_1(\lambda) \neq \theta_2(\lambda)$ for any $\lambda \neq 1$;
(d) $\theta_1(1) = \theta_2(1)$;
(e) for any $\lambda \in [0,1]$, denoting by $\gamma_2(\lambda)$ the maximal integral curve of W starting at $\theta_2(\lambda)$, there exists a unique minimizer of τ on $\mathcal{L}^+_{\theta_1(\lambda), \gamma_2(\lambda), E}([a_1, a_2], U_i)$.

We set

$$A = \{\lambda \in [0,1] \,|\, \text{the geodesic minimizing } \tau \text{ on } \mathcal{L}^+_{\theta_1(\lambda), \gamma_2(\lambda), E}([a_1, a_2], U_i) \text{ does not intersect } \partial\Lambda\}.$$

If we choose ρ^* and l^* sufficiently small, by Proposition II.4, $1 \in A$ so A is not empty. Then we consider

$$\lambda_0 = \inf A \geq 0.$$

It is sufficient to prove that $\lambda_0 \in A$ and $\lambda_0 = 0$. Consider a sequence $(\lambda_n)_{n \in \mathbf{N}} \subset A$ such that

$$\lim_{n\to\infty} \lambda_n = \lambda_0$$

and let $(w_n)_{n\in\mathbf{N}}$ be the corresponding sequence of geodesics such that, for any $n\in\mathbf{N}, w_n$ is the minimum point of τ on $\mathcal{L}^+_{\theta_1(\lambda_n),\gamma_2(\lambda_n),E}([a_1,a_2],U_i)$ and

$$w_n([a_1,a_2])\subset\Lambda \quad \forall n\in\mathbf{N}.$$

Up to a subsequence, there exists a geodesic $\overline{w}$ satisfying (1) such that

$$\lim_{n\to\infty} w_n = \overline{w}$$

with respect to the C^2-norm and

$$\overline{w}(a_1)=\theta_1(\lambda_0),\;\; \overline{w}(a_2)\in\Lambda,\;\; \overline{w}([a_1,a_2])\subset\overline{\Lambda}.$$

By the time-convexity of $\overline{\Lambda}$, we get

$$\overline{w}([a_1,a_2])\subset\Lambda,$$

then $\lambda_0\in A$ and by continuity, $\lambda_0=0$. □

III. DEFORMATION RESULTS

Here we shall state the existence of a shortening flow for τ, adapting to our case the ideas of Ref. 8. This is the crucial step for developing a Ljusternik–Schnirelmann category theory and a Morse theory.

Fix $c>\inf\{\tau(z)|z\in\mathcal{L}^+_{p,\gamma,E}\}$. Let $K=K(c)$ be the compact subset of $\overline{\Lambda}$ such that

$$z([0,1])\subset K(c) \quad \forall z\in\tau^c\cup\mathcal{L}^+_{p,\gamma,E}$$

whose existence is assured by the pseudocoercivity. We can cover $K(c)$ by a finite number of U_i, open subsets of $\mathcal{M}$, where the metric can be written as in Remarks II.2 and II.3. Consider $D(c)$ as in Proposition II.6, $\rho^*(c)$ and $l^*(c)$ as in Proposition II.7, and choose $N=N(c)$ such that

$$\frac{D(c)}{N}<\rho^*, \quad \frac{1}{N}<l^*.$$

Choose a partition

$$0=s_0<s_1<\cdots<s_{N-1}<s_N=1$$

of [0,1], such that for any $i=1,...,N$,

$$s_i-s_{i-1}=\frac{1}{N}.$$

Consider a curve $z\in\tau^c\cap\mathcal{L}^+_{p,\gamma,E}$ and choose $N+1$ points $z_0,\ldots,z_N$ on $z([0,1])$ such that for any $i=1,...,N$,

$$z_0=z(0)=p,\;\; z_N=z(1),\;\; d_R(z_i,z_{i-1})=\frac{l(z)}{N}.$$

Denote by $\gamma_i (i=1,...,N)$ the maximal integral curve of W such that $\gamma_i(0)=z_i$. Note that $\gamma_N=\gamma(s+\tau(z))$ for any s. By Proposition II.7, let w_1 be the future pointing geodesic minimizing τ

on $\mathcal{L}^+_{p,\gamma_1,E}([s_0,s_1],\Lambda)$ and, for any $i=2,...,N$, let w_i be the future pointing geodesic minimizing τ on $\mathcal{L}^+_{w_{i-1}(s_{i-1}),\gamma_i,E}([s_{i-1},s_i],\Lambda)$. Note that, if ρ^* is opportunely chosen and N is large enough

$$d_R(w_i(s_i),z_i)\leqslant\rho^* \quad i=1,...,N.$$

We define a curve $\eta_1\in\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ by setting

$$\eta_1([s_{i-1},s_i])=w_i \quad i=1,...,N.$$

By elementary comparison theorems for ordinary differential equations,

$$\tau(\eta_1)\leqslant\tau(z)\leqslant c,$$

then

$$\eta_1([0,1])\subset K(c).$$

Starting from η_1, a second curve η_2 can be constructed. On any minimizer $w_i(i=1,...,N)$ consider the point m_i such that

$$d_R(w_i(s_{i-1}),m_i)=d_R(m_i,w_i(s_i)).$$

Denote by λ_i the maximal integral curve of w starting at m_i for any $i=1,...,N$ and denote by $\lambda_{N+1}(s)=\gamma(s+\tau(\eta_1))$. Consider the partition of $[0,1]$ given by

$$\sigma_0=0,\ \ \sigma_i=\frac{2i-1}{2N} \quad \forall i=1,...,N,\ \ \sigma_{N+1}=1.$$

Denote by u_1 the minimizer of τ on $\mathcal{L}^+_{p,\lambda_1,E}([\sigma_0,\sigma_1],\Lambda)$ and, by induction, for any $i=1,2,...,N+1$ let u_i be the minimizer of τ on $\mathcal{L}^+_{u_{i-1}(\sigma_{i-1}),\lambda_i,E}([\sigma_{i-1},\sigma_i],\Lambda)$. The curve $\eta_2\in\mathcal{B}^+_{p,\gamma,E}(\Lambda)$ is such that

$$\eta_2([\sigma_{i-1},\sigma_i])=u_i\,.$$

Again by using comparison theorems in ordinary differential equations, it is possible to prove that

$$\tau(\eta_2)\leqslant\tau(\eta_1).$$

We can prove now that

$$\tau(\eta_2)<\tau(\eta_1). \tag{31}$$

Indeed, if $\tau(\eta_2)=\tau(\eta_1)$, by comparison theorems in ordinary differential equations, there results that η_1 is a minimizer on an interval $[\sigma_i,\sigma_{i+1}]$. By the above construction, η_1 consists of two timelike geodesics satisfying (1). If it is not a geodesic, $\dot{\eta}_1$ has a discontinuity at $s_{i+1}=(\sigma_{i+1}+\sigma_i)/2$. Denote by U_{η_1} the parallel transport of $\dot{\gamma}(\tau(\eta_1))$ along η_1. Since η_1 is a minimizer, it satisfies (see Ref. 2)

$$\int_{\sigma_i}^{\sigma_{i+1}}\frac{\langle\dot{\eta}_1,D_s\zeta\rangle}{\langle U_{\eta_1},\dot{\eta}_1\rangle}ds=0 \tag{32}$$

for any C^∞-vector field ζ along η_1 such that $\zeta(\sigma_i)=0=\zeta(\sigma_{i+1})$. By (32),

$$g=\frac{\dot{\eta}_1}{\langle U_{\eta_1},\dot{\eta}_1\rangle}\in C^1([\sigma_i,\sigma_{i+1}],T\mathcal{M}), \tag{33}$$

then also

$$-\frac{E}{\langle U_{\eta_1}\dot{\eta}_1\rangle^2}=\langle g,g\rangle\in C^1([\sigma_i,\sigma_{i+1}],\mathbf{R}).$$

As $\langle U_{\eta_1},\dot{\eta}_1\rangle^2$ is strictly positive,

$$\langle U_{\eta_1},\dot{\eta}_1\rangle\in C^1([\sigma_i,\sigma_{i+1}],\mathbf{R}),$$

then, by (33), $\dot{\eta}_1$ is of class C^1. Then, whenever we are far from timelike geodesics satisfying (1), (31) holds.

Now we can construct a continuous flow $\eta(\sigma,z)$ for any $\sigma\in[0,1]$, $z\in\tau^c$. For any $\sigma\in[0,1]$ we can define $\eta(\sigma,z)$ on the interval $[s_0,s_1]$ as follows. Assume that $z([s_0,s_1])$ is contained in U_i. Thanks to Remarks II.2, II.3, we set $p=(x_0,0)$, $\gamma_1=(x_1,t_1+s)$ and $z(s)=(x(s),t(s))$ in $[s_0,s_1]$, where x is a curve joining x_0 and x_1 and t satisfies the Cauchy problem (19). Let α_i and δ_i be the coefficients of the metric g in U_i. Let $y(\sigma)$ be the minimizer of the functional

$$F_\sigma(y)=\int_{s_0}^{\sigma s_1}\langle\delta_i(y,t_y),\dot{y}\rangle_0 ds+\int_{s_0}^{\sigma s_1}\sqrt{\langle\alpha_i(y,t_y)\dot{y},\dot{y}\rangle_0^2+\langle\delta_i(y,t_y),\dot{y}\rangle_0+E}\,ds \tag{34}$$

with boundary conditions $y(0)=x_0$, $y(\sigma s_1)=x(\sigma s_1)$, where t_y is the solution of (19) in $[0,\sigma s_1]$. Denote by $\hat{y}(\sigma)$ the extention of $y(\sigma)$ to $[s_0,s_1]$ obtained taking $\hat{y}(\sigma)(s)=x(s)$ for any $s\in[\sigma s_1,s_1]$. Finally denote by $\hat{t}_y(\sigma)$ the corresponding solution of (19) in $[s_0,s_1]$. We define

$$\eta(\sigma,z)_{|[s_0,s_1]}=(\hat{y}(\sigma),\hat{t}_y(\sigma)).$$

In the same way, η can be defined on the other intervals $[s_{i-1},s_i]$. Note that

$$\eta(1,z)=\eta_1.$$

Similarly, we can extend the flow η to a map on $[0,2]\times\tau^c$ in such a way that

$$\eta(2,z)=\eta_2.$$

Iterating the previous arguments, replacing the original curve z by η_2, we obtain a flow $\eta(\sigma,z)$ defined on $\mathbf{R}^+\times\tau^c$.

As a first consequence of the existence of the flow η, since

$$\tau(\eta(\sigma,z))\leqslant\tau(z)$$

for any σ and for any z, we immediately deduce the following proposition.

Proposition III.1: $\mathcal{L}^+_{p,\gamma,E}\cap\tau^c$ is homotopically equivalent to $\mathcal{B}^+_{p,\gamma,E}(\Lambda)\cap\tau^c$, for any $c\in\mathbf{R}$. Moreover, choosing a suitable continuous map $\rho^*(c)$ and arguing as in Sec. IX of Ref. 5, we also obtain

Proposition III.2: $\mathcal{L}^+_{p,\gamma,E}$ is homotopically equivalent to $\mathcal{B}^+_{p,\gamma,E}(\Lambda)$.

Classical deformation results (see, e.g., Ref. 12) can be reproved for the functional τ on $\mathcal{L}^+_{p,\gamma,E}$ by using compactness arguments close to the ones used for the shortening method for Riemannian geodesics. More precisely the following propositions hold.

Proposition III.3: Let c be a regular value for τ on $\mathcal{L}^+_{p,\gamma,E}$ *(namely $\tau^{-1}(c)$ does not contain geodesics). Then, there exists a positive number $\delta=\delta(c)$ and a continuous map $H\in C^0([0,1]\times\tau^{c+\delta},\tau^{c+\delta})$, such that*

(1) *$H(0,z)=z$, for every $z\in\tau^{c+\delta}$;*

(2) $\mathrm{H}(1,\tau^{c+\delta})\subset\tau^{c-\delta}$;
(3) *$H(\sigma,z)\in\tau^{c-\delta}$, for any $\sigma\in[0,1]$ and $z\in\tau^{c-\delta}$.*

Proposition III.4: Let K_c be the set of the geodesics in $\tau^{-1}(c)\cap\mathcal{L}^+_{p,\gamma,E}$. Then for any open neighborhood U of K_c, there exists a positive number $\delta=\delta(U,c)$ and a homotopy $H\in C^0([0,1]\times\tau^{c+\delta},\tau^{c+\delta})$, such that

(1) *$H(0,z)=z$, for any $z\in\tau^{c+\delta}$.*
(2) *$H(1,\tau^{c+\delta}\setminus U)\subset\tau^{c-\delta}$;*
(3) *$H(\sigma,z)\in\tau^{c-\delta}$, for any $\sigma\in[0,1]$ and $z\in\tau^{c-\delta}$.*

Proof of Theorem I.2: As, by Proposition II.5, τ is pseudocoercive on $\mathcal{L}^+_{p,\gamma,E}$ and γ is an integral curve of W, it is not difficult to show that τ is bounded from below on $\mathcal{L}^+_{p,\gamma,E}$. Then, if

$$d=\inf_{z\in\mathcal{L}^+_{p,\gamma,E}}\tau(z)\in\mathbf{R}$$

were not a critical value for τ, by Proposition III.3, there should exist $\delta>0$ and a homotopy between $\tau^{d+\delta}$ and $\tau^{d-\delta}$. But this is a contradiction, as $\tau^{d+\delta}\neq\emptyset$ and $\tau^{d-\delta}=\emptyset$. □

Proof of Theorem I.3: As $\mathcal{L}^+_{p,\gamma,E}$ is not empty, $\mathrm{cat}(\mathcal{L}^+_{p,\gamma,E})\geqslant 1$. Then, for any $k\in\mathbf{N}$, $0<k\leqslant\mathrm{cat}(\mathcal{L}^+_{p,\gamma,E})$ we can define

$$\Gamma_k=\{B\subset\mathcal{L}^+_{p,\gamma,E}|\mathrm{cat}(B,\mathcal{L}^+_{p,\gamma,E})\geqslant k\},\qquad c_k=\inf_{B\in\Gamma_k}\sup_{z\in B}\tau(z).$$

By using Propositions III.3 and III.4 and by classical arguments in critical point theory (see, e.g., Ref. 12), each c_k is well defined and it is a critical value of τ. Moreover if for some k, $c_k=c_{k+1}$ there are infinitely many critical points of τ at the level c_k, so, by Proposition III.2, the proof is complete. □

Proof of Theorem I.6: Far from geodesics of $\mathcal{L}^+_{p,\gamma,E}$, we can use the shortening flow described in this section for studying the homotopy type of the sublevels of τ. Near geodesics we can proceed as in Ref. 7. This makes it possible to write the classical Morse relations using, for any geodesic $z\in\mathcal{L}^+_{p,\gamma,E}$, the Morse index $m(z,\tau)$ of z as a critical point of τ. As in Ref. 7 and using the nondegeneracy assumption 1., it can be proved that

$$m(z,\tau)=\mu(z)$$

giving the proof of (10). □

[1] F. Antonacci and P. Piccione, ''A Fermat principle on Lorentzian manifolds and applications,'' Appl. Math. Lett. **9**, 91–96 (1996).
[2] F. Giannoni, A. Masiello, and P. Piccione, ''A timelike extension of Fermat's principle in general relativity and applications,'' Calc. Var. **6**, 263–283 (1998).
[3] V. Perlick, ''On Fermat's principle in general relativity. I. The general case,'' Class. Quantum Grav. **7**, 1319–1331 (1990).
[4] I. Kovner, ''Fermat principle in arbitrary gravitational fields,'' Astrophys. J. **351**, 114–120 (1990).
[5] F. Giannoni, A. Masiello, and P. Piccione, ''A variational theory for light rays in stably causal Lorentzian manifolds: Regularity and multiplicity results,'' Commun. Math. Phys. **187**, 375–415 (1997).
[6] F. Giannoni, A. Masiello, and P. Piccione, ''A Morse theory for light rays on stably causal Lorentzian manifolds,'' Ann. Inst. H. Poin., Physique Theorique **69**, 359–412 (1998).
[7] F. Giannoni, A. Masiello, and P. Piccione, ''A Morse theory for massive particles and photons in general relativity'' (unpublished).
[8] J. Milnor, *Morse Theory*, Ann. Math. Stud. (Princeton University Press, Princeton, NJ, 1963), Vol. 51.
[9] F. Giannoni, A. Masiello, and P. Piccione, ''The Fermat principle in general relativity and applications'' (unpublished).
[10] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian Geometry* (Marcel Dekker, New York, 1996).
[11] B. O'Neill, *Semi-Riemannian Geometry with Applications to Relativity* (Academic, New York, 1983).
[12] J. Mawhin and M. Willem, *Critical Points Theorems and Hamiltonian Systems* (Springer-Verlag, New York, 1988).
[13] I. Ekeland, ''On the variational principle,'' J. Math. Anal. Appl. **47**, 324–353 (1974).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Algebraically special, expanding and twisting gravitational fields with vanishing Newman–Unti–Tanbyrubo parameter

B. V. Ivanov
Institute for Nuclear Research and Nuclear Energy, Tzarigradsko Shausse 72, Sofia 1784, Bulgaria

(Received 10 June 1999; accepted for publication 22 July 1999)

The system of Einstein–Maxwell equations for fields mentioned in the title is simplified. Known pure radiation solutions are systematized and new solutions are given by separating the variables. © *1999 American Institute of Physics.* [S0022-2488(99)00611-8]

I. INTRODUCTION

There exists many papers dealing with algebraically special, expanding and twisting vacuum, pure radiation or Einstein–Maxwell fields. An extensive bibliography up to 1980 exists in Ref. 1. Further results on Einstein–Maxwell fields may be found in Ref. 2. Pure radiation fields have been studied extensively in Refs. 3–8. Petrov type II fields are characterized by the nonvanishing of the second Weyl scalar Ψ_2,

$$\Psi_2=(m+iM)\rho^3. \tag{1}$$

Here m is the mass parameter, M is the Newman–Unit–Tamburino (NUT) parameter, and ρ is given by

$$\rho=-\frac{1}{r+i\Sigma}, \tag{2}$$

where r is the coordinate along the null congruence of geodesics and Σ is the twist.[1]

It has been noticed in different contexts that the condition $M=0$ simplifies the equations.[2,4,5,7,9] In the present paper we explore this condition with the help of the method proposed in Ref. 3, putting the emphasis on intrinsically time-dependent solutions. We clarify the general structure of the solutions obtained by separation of the traditional variables r, u, ζ, $\bar{\zeta}$, where u is the retarded time and ζ, $\bar{\zeta}$ span a two-dimensional surface. Known solutions are systematized and new solutions are given.

Twisting gravitational fields with $M=0$ generalize the classes of Robinson–Trautman (RT) (Ref. 10) and Kerr–Schild (KS) fields[11] and are physically realistic, their simplest representatives being the Schwarzschild, Kerr, and Vaidya solutions.

In Sec. II the method of Stephani is applied to simplify the Einstein–Maxwell equations. In Sec. III the main field equation is reduced to a linear second order equation. Its general solution with separated variables is found and studied in Sec. IV. Section V is devoted to solutions linear in u, while in Sec. VI solutions with exponential behavior are discussed. Section VII contains some conclusions.

II. THE METRIC AND FIELD EQUATIONS

The standard form of the metric for expanding and twisting fields is[1]

$$ds^2=\frac{2d\zeta d\bar{\zeta}}{\rho\bar{\rho}P^2}-2\Omega[dr+Wd\zeta+\bar{W}d\bar{\zeta}+H\Omega], \tag{3}$$

$$\Omega = du + L d\zeta + \overline{L} d\overline{\zeta}.$$

It is determined by the r-independent real functions P, m, M and the complex function L,

$$2i\Sigma = P^2(\overline{\partial} L - \partial \overline{L}), \tag{4}$$

$$W = \rho^{-1} L_u + i\partial\Sigma, \tag{5}$$

$$H = -r(\ln P)_u - (mr + M\Sigma - \kappa \Phi_1^0 \overline{\Phi}_1^0)\rho\overline{\rho} + \frac{K}{2}, \tag{6}$$

$$K = 2P^2 Re[\partial(\overline{\partial} \ln P - \overline{L}_u)], \tag{7}$$

where $\partial = \partial_\zeta - L\partial_u$ and κ is the Newton constant. The basic functions satisfy the following system of equations:

$$(\partial - 3L_u)(m + iM) = -2\kappa P^{-1} \Phi_1^0 \overline{\Phi}_2^0, \tag{8}$$

$$P^{-3} M = \mathrm{Im}\, \partial\partial\overline{\partial\partial} V, \tag{9}$$

$$\frac{n^2}{2P^2} = -P[P^{-3}(m + iM)]_u + P(\partial\partial\overline{\partial\partial} V)_u - (\partial\partial V)_u (\overline{\partial\partial} V)_u, \tag{10}$$

$$(\partial - 2L_u)\Phi_1^0 = 0, \tag{11}$$

$$(\partial - L_u)(P^{-1}\Phi_2^0) + (P^{-2}\Phi_1^0)_u = 0, \tag{12}$$

where $V_u = P$, Φ_1^0, and Φ_2^0 are the essential parts of the Maxwell scalars

$$\Phi_1 = \rho^2 \Phi_1^0, \tag{13}$$

$$\Phi_2 = \rho\Phi_2^0 + \rho^2 P(2\overline{L}_u - \overline{\partial})\Phi_1^0 + 2i\rho^3 P(\Sigma \overline{L}_u - \overline{\partial}\Sigma)\Phi_1^0, \tag{14}$$

and n is either the energy density of pure radiation (then $\Phi_1 = \Phi_2 = 0$) or is given in the Einstein–Maxwell case by

$$n^2 = 2\kappa \Phi_2^0 \overline{\Phi}_2^0. \tag{15}$$

Vacuum solutions have $\Phi_1 = \Phi_2 = n = 0$.

The form of the metric is preserved by certain coordinate transformations, one of which is the change of the retarded time

$$u' = F(u, \zeta, \overline{\zeta}), \tag{16}$$

$$P' = F_u^{-1} P, \tag{17}$$

$$L' = F_u L - F_\zeta, \tag{18}$$

$$\Sigma' = F_u^{-1}\Sigma, \tag{19}$$

$$(m + iM)' = F_u^{-3}(m + iM). \tag{20}$$

Let us apply now to the system (8)–(12) the method of Stephani.[3] It works when either Φ_1^0 or Φ_2^0 vanishes. Then Eq. (8) is solved by

$$m+iM=\phi_u^3. \tag{21}$$

The complex field ϕ is invariant under Eq. (16). It satisfies $\partial\phi=0$, which stated differently gives

$$L=\frac{\phi_\zeta}{\phi_u}. \tag{22}$$

When $M=0$ we can use Eq. (16) to transform m to a positive or negative constant m_0 so that

$$\phi=m_0^{1/3}[u+iq(\zeta,\bar{\zeta})], \tag{23}$$

$$L=iq_\zeta. \tag{24}$$

Obviously $L_u=0$. This gauge differs from the most popular Kerr's gauge when $P_u=0$, but is very suitable when the NUT parameter vanishes. Equations (9) and (10) become

$$\partial\partial\overline{\partial\partial}V=\overline{\partial\partial}\partial\partial V, \tag{25}$$

$$\frac{n^2}{2P^2}=3m_0P^{-3}P_u+P\partial\partial\overline{\partial\partial}P-\partial\partial P\overline{\partial\partial}P, \tag{26}$$

with $\partial=\partial_\zeta-iq_\zeta\partial_u$. When $P_u\neq 0$, n^2 can be made positive by the choice of m_0 at least for some region of space–time.[1,4,9]

If $\Phi_1^0=0$, Eq. (12) gives

$$\Phi_2^0=PG(\phi,\bar{\zeta}), \tag{27}$$

where G is an arbitrary function. If $\Phi_2^0=0$, Eqs. (11) and (12) give

$$\Phi_1^0=\overline{C}(\bar{\zeta})\exp\left[2i\int(\ln P)_u q_\zeta d\zeta\right], \tag{28}$$

with $C(\zeta)$ being an arbitrary analytic function. This is a particular case of Theorem 26.2 from Ref. 1. The expressions for the other functions in the metric simplify

$$\Sigma=P^2Q, \tag{29}$$

$$W=i\partial(P^2Q), \tag{30}$$

$$K=P^2(\bar{\partial}\partial+\partial\bar{\partial})\ln P, \tag{31}$$

$$H=-r(\ln P)_u-(m_0r-\kappa\Phi_1^0\overline{\Phi}_1^0)\rho\bar{\rho}+\frac{K}{2}. \tag{32}$$

We have introduced the real, invariant under (16), function

$$Q(\zeta,\bar{\zeta})=q_{\zeta\bar{\zeta}}. \tag{33}$$

The Weyl scalars[12] contain a lot of terms with L_u and also simplify in the gauge $L_u=0$. We quote only the leading terms:

$$\Psi_2=m_0\rho^3, \tag{34}$$

$$\Psi_3=-\rho^2P^3\partial I+O(\rho^3), \tag{35}$$

$$\Psi_4 = \rho P^2 I_u + O(\rho^2), \tag{36}$$

$$I = P^{-1}\overline{\partial\partial P}. \tag{37}$$

What remains to be solved is the couple of Eqs. (25) and (26) for P and q. For pure radiation Eq. (26) is just an inequality.

When $m+iM=0$ (Petrov types III and N) Eq. (8) holds identically, but still a potential ϕ may be introduced with the property $\partial\phi=0$ and we can study the subclass of solutions satisfying Eq. (23) (with $m_0=1$). Then in all other equations we can set $m_0=0$.

III. THE FIELD EQUATION (25)

Equation (25) is a linear equation of fourth order with respect to V. However, in the gauge $L_u=0$ it becomes a linear equation of second order for P. This can be established with the help of the commutator

$$[\partial,\overline{\partial}] = 2iQ\partial_u. \tag{38}$$

Then Eq. (25) becomes

$$Q_{\zeta\bar\zeta}P + 2Q_{\bar\zeta}P_\zeta + 2Q_\zeta P_{\bar\zeta} + 4QP_{\zeta\bar\zeta} + 2i(q_{\bar\zeta}Q_\zeta - q_\zeta Q_{\bar\zeta})P_u + 4iQ(q_{\bar\zeta}P_{\zeta u} - q_\zeta P_{\bar\zeta u}) + 4Qq_\zeta q_{\bar\zeta}P_{uu} = 0. \tag{39}$$

This equation characterizes the twisting solutions because when $\Sigma=0$ it is trivial. It is nonlinear in q except for time-independent solutions when only the first four terms remain. In this case Eq. (26) turns into the equation for type III RT solutions, containing only P. It is logical to solve first Eq. (26) for P and plug the result in Eq. (39) to find q. In the present paper time-dependent solutions will be discussed mainly. In this case Eq. (39) should be solved for P when q is given and the result placed in Eq. (26) to find n^2 for pure radiation solutions.

Similar equations have been derived for time-independent fields in Ref. 1 [see Eq. (25.46)] and 2 [Eq. (5.12)]. For time-dependent fields and concrete expressions for q such equations can be found in Refs. 7 and 9. A linear equation of second order for $(-2\Sigma)^{1/2}$ is found in Ref. 5 in terms of Cauchy–Riemann structures admitting Lie groups of symmetries and used in Refs. 4–6 to obtain pure radiation solutions.

If P and q depend on ζ and $\bar\zeta$ via a single function $\alpha(\zeta,\bar\zeta)$ the terms in brackets in Eq. (39) cancel and it yields

$$(Q_{\alpha\alpha}\alpha_\zeta\alpha_{\bar\zeta} + Q_\alpha\alpha_{\zeta\bar\zeta})P + 4(Q_\alpha\alpha_\zeta\alpha_{\bar\zeta} + Q\alpha_{\zeta\bar\zeta})P_\alpha + 4Q\alpha_\zeta\alpha_{\bar\zeta}(P_{\alpha\alpha} + q_\alpha^2 P_{uu}) = 0. \tag{40}$$

Let $\zeta=(1/\sqrt{2})(x+iy)$. Suppose that $\alpha=x$ and ∂_y is a Killing vector. Then Eq. (40) becomes

$$q_x^2 QP_{uu} + QP_{xx} + Q_xP_x + \tfrac{1}{4}Q_{xx}P = 0 \tag{41}$$

and $Q=\frac{1}{2}q_{xx}$. In the case of axial symmetry $\alpha=\sigma=\zeta\bar\zeta$ and Eq. (40) reads

$$q_\sigma^2 P_{uu} + P_{\sigma\sigma} + \frac{(\sigma Q)_\sigma}{\sigma Q}P_\sigma + \frac{(\sigma Q_\sigma)_\sigma}{4\sigma Q}P = 0, \tag{42}$$

$$Q = (\sigma q_\sigma)_\sigma. \tag{43}$$

The introduction of $z=\ln\sigma$ as a variable greatly simplifies Eqs. (42) and (43) making them an analog of Eq. (41),

$$q_z^2 QP_{uu} + QP_{zz} + Q_zP_z + \tfrac{1}{4}Q_{zz}P = 0, \tag{44}$$

$$Q=e^{-z}q_{zz}. \tag{45}$$

Thus every solution with y-symmetry has an axisymmetric mirror with the same P and Q but as functions of z.

IV. SEPARATION OF VARIABLES

Let us search for twisting solutions by separating the variables $P=p(\zeta,\bar{\zeta})f(u)$. Equation (39) transforms into

$$f_{uu}+Af_u+Cf=0, \tag{46}$$

$$A=\frac{i}{2}\left[\frac{Q_\zeta}{q_\zeta Q}-\frac{Q_{\bar{\zeta}}}{q_{\bar{\zeta}}Q}+2\left(\frac{p_\zeta}{q_\zeta p}-\frac{p_{\bar{\zeta}}}{q_{\bar{\zeta}}p}\right)\right], \tag{47}$$

$$C=(4q_\zeta q_{\bar{\zeta}}Qp)^{-1}(Q_{\zeta\bar{\zeta}}p+2Q_{\bar{\zeta}}p_\zeta+2Q_\zeta p_{\bar{\zeta}}+4Qp_{\zeta\bar{\zeta}}). \tag{48}$$

Equation (46) is a well-known linear equation. It requires that A and C must be constant and possesses three types of solutions according to the sign of $\lambda^2=4C-A^2$,

$$f=e^{-(Au/2)}(C_1e^{(\beta u/2)}+C_2e^{-(\beta u/2)}), \tag{49}$$

$$f=e^{-(Au/2)}\left(C_1\sin\frac{\lambda u}{2}+C_2\cos\frac{\lambda u}{2}\right), \tag{50}$$

$$f=e^{-(Au/2)}(C_1u+C_2), \tag{51}$$

where $\beta^2\equiv-\lambda^2>0$, $\lambda^2>0$ and $\lambda=0$, respectively. Notice the common exponential factor if $A\neq0$. It is quite interesting that damping exponential behavior is generic for RT solutions, but it stems from the analog of Eq. (26) in the nontwisting case.[13,14] When Eq. (40) holds, A vanishes. There are still three types of solutions, one of them with exponential behavior. When there is y-symmetry Eqs. (46) and (41) become

$$f_{uu}+Cf=0, \tag{52}$$

$$Qp_{xx}+Q_xp_x+(\tfrac{1}{4}Q_{xx}-Cq_x^2Q)p=0. \tag{53}$$

In the case of axial symmetry Eq. (52) still holds, but Eq. (53) is replaced either by the same equation with x changed to $z=\ln\sigma$ or by

$$p_{\sigma\sigma}+(\ln\sigma Q)_\sigma p_\sigma+\left[\frac{(\sigma Q_\sigma)_\sigma}{4\sigma Q}-Cq_\sigma^2\right]p=0. \tag{54}$$

Let us discuss next Eq. (26) for two of the types of u-behavior allowed by Eq. (39). Suppose first that $P=pe^{cu}$. Introduce the function $B=pe^{icq}$. Then Eq. (26) reads

$$n^2=6m_0c+2p^2(\psi e^{-4icq})_\zeta e^{2icq+4cu}, \tag{55}$$

$$\psi=BB_{\zeta\bar{\zeta}\bar{\zeta}}-B_\zeta B_{\bar{\zeta}\bar{\zeta}}. \tag{56}$$

When $m_0=0$ the solution is of type III only when $\psi\neq0$ because $\partial I=\psi/B^2$. When $m_0\neq0$ and $\psi=0$ the solution is of KS type because $\partial I=I_{\bar{\zeta}}=0$ and I can be nullified by a coordinate transformation $\bar{\zeta}'=\bar{g}(\bar{\zeta})$. This means $\partial\partial P=0$ [see Eq. (37)], which can be lifted to $\partial\partial V=0$. The latter is exactly the Kerr–Schild condition.[1,3]

Suppose next that $P=pu$. When there is y-symmetry Eq. (26) yields

$$n^2=\frac{6m_0}{u}+\frac{1}{2}p^2(pp_{xxxx}-p_{xx}^2)u^4-2p^2(q_xp_x+Qp)^2u^2. \tag{57}$$

When there is axial symmetry Eq. (26) changes into

$$n^2=\frac{6m_0}{u}+2p^2[2pp_{\sigma\sigma}+4\sigma pp_{\sigma\sigma\sigma}+\sigma^2(pp_{\sigma\sigma\sigma\sigma}-p_{\sigma\sigma}^2)]u^4-2\sigma^2[(q_\sigma p^2)_\sigma]^2u^2. \tag{58}$$

Finally, let us perform the separation of variables in the Maxwell equations. Equation (28) simplifies

$$\Phi_1^0=\overline{C(\zeta)}\exp[2iq(\ln f)_u]. \tag{59}$$

When $f=u$ and $f=e^{cu}$ this becomes, respectively,

$$\Phi_1^0=\overline{C(\zeta)}e^{(2iq/u)}, \tag{60}$$

$$\Phi_1^0=\overline{C(\zeta)}e^{2icq}. \tag{61}$$

Equation (12) or (27) separates as follows:

$$P^{-1}\Phi_2^0=\overline{N(\zeta)}e^{a(u+iq)}, \tag{62}$$

$N(\zeta)$ being another arbitrary analytic function and a is a constant different from c in general. The energy density of a null Maxwell field (15) becomes

$$\frac{n^2}{2P^2}=\kappa|N(\zeta)|^2e^{2au}. \tag{63}$$

Obviously polynomial dependence of P in Eq. (26) is not allowed by (63). Consequently, null Maxwell fields cannot induce solutions with $P=pu$. It is seen also from Eqs. (55), (57), and (58) that vacuum solutions of type II are not possible because the different terms cannot cancel each other.

V. SOLUTIONS LINEAR IN *U*

These are solutions with $C=0$ in Eq. (46) or Eqs. (52), (53), and (54). Then these equations are identical to the equations for a u-independent P. Therefore we can use any time-independent solution of Eqs. (25) and (26) found in the literature which has $L_u=0$ and axial or y-symmetry. We can also use time-independent solutions of Eq. (53) or Eq. (54) which are not solutions of Eq. (26).

For example, let us find polynomial solutions with y-symmetry which capitalize on the only known vacuum RT solution for type III fields.[1,10] If we take $L=ix^{b-1}$, $b\neq 1$ then Eq. (53) becomes the Euler equation

$$x^2p_{xx}+(b-2)xp_x+\tfrac{1}{4}(b-2)(b-3)p=0 \tag{64}$$

and its solutions are

$$p=|x|^a, \tag{65}$$

$$p=\ln|x|, \tag{66}$$

$$p=|x|^{(3-b)/2}[C_1\sin(\mu\ln|x|)+C_2\cos(\mu\ln|x|)], \tag{67}$$

when $b<3$, $b=3$, and $b>3$, respectively. Here $a=\frac{1}{2}(3-b\pm\sqrt{3-b})$ and $\mu=\frac{1}{2}\sqrt{b-3}$. The first of these solutions has been obtained essentially in Ref. 4 by a rather intricate procedure. Plugging Eq. (65) and $q_x=x^{b-1}$ into Eq. (57) yields the inequality to be satisfied,

$$\frac{12m_0}{u}|x|^{2(b-1\pm\sqrt{3-b})}+2a(a-1)(3-2a)u^4-(2-\sqrt{3-b})^2|x|^{2b}u^2\geqslant 0. \tag{68}$$

Throughout the paper we consider $u>0$. The first term is always positive (we accept that $m_0>0$ unless stated otherwise), the third is always negative, while the second changes sign and is positive for $a<0$ and $1<a<3/2$. The third term has negative poles in x for any $b\neq 0$ and $b\neq -1$ which cannot be compensated by the second term. Hence, pure radiation solutions of type III are not allowed and the first term should be present to compensate the negative poles with positive ones. This is achieved when $b<2$. The plus sign in the formula for a must be taken in both cases. For growing u, however, the first term diminishes to zero. Thus the energy density is positive for $b<2$ and small enough u. When $b=0$ the third term is x-independent but the second is negative, excluding again a type III solution. Type II solution still exists. Finally, when $b=-1$ the third term vanishes but the second is still nonpositive because $a=1$ or $a=3$. When $a=1$ only the first term remains and n is x-independent, positive and bounded. This is a KS solution because $\partial\bar{\partial}V=-xq_x^2/2$ and can be turned into zero by a u-independent transformation of V.

The second solution, given by Eq. (66), has no free parameters and leads to incurable negative poles and logarithmic singularities in n and is therefore unphysical. As a last comment, we may start with $p=x^{b-1}$ in Eq. (53) and obtain a Euler equation for Q with the corresponding three types of solutions.

The simplest solution of Eq. (53) is probably when P is u-dependent only, i.e., $p=p_0=\text{const}$. Then $Q_{xx}=0$ and $P=p_0u$, $L=(i/\sqrt{2})x(c_1x+2c_2)$, where c_i are arbitrary constants and

$$n^2=\frac{6m_0}{u}-2p_0^4(c_1x+c_2)^2u^6. \tag{69}$$

The energy density is positive for any x when $c_1=0$ and u is small enough. This is not a KS field.

Another simple solution is obtained when $p=e^{ax}$, a being some constant. Then Eq. (53) has constant coefficients and one of the solutions is $q_x=\exp(-2ax)$. The energy density is positive, $n^2=6m_0/u$ and $I=a^2/2$. Consequently the solution is of type II and is equivalent to a KS solution.

Let us investigate next the axisymmetric solutions. As has been mentioned in the previous sections, Eq. (53) has a counterpart for axisymmetric fields with the same Q and P but with x replaced by z. However, q and n are different, being given by Eqs. (45) and (58) instead of $q_{xx}=2Q$ and Eq. (57). The solution with $p=p_0$ has a mirror solution

$$L=i\bar{\zeta}(c_3\ln\sigma+c_4), \tag{70}$$

$$n^2=\frac{6m_0}{u}-2c_3^2p_0^4u^6. \tag{71}$$

The energy density is positive for small u. When $c_3=0$ it is positive everywhere and the solution is a KS field.

Another solution has $q=\sigma$, $L=i\bar{\zeta}$, $Q=1$. Then Eq. (54) yields

$$p_{\sigma\sigma}+\frac{1}{\sigma}p_\sigma=0 \tag{72}$$

and $p=c_5\ln\sigma+c_6$. It was found by a different method in Ref. 9. When $c_5=0$ it coincides with the previous solution with $c_3=0$.

A well-known solution is given by $p=1+\sigma/2$. Equation (54) becomes linear with respect to Q,

$$(2+\sigma)\sigma Q_{\sigma\sigma}+(2+5\sigma)Q_{\sigma}+4Q=0 \tag{73}$$

and its solution is the hypergeometric function $F(2,2,1,-\sigma/2)$. It degenerates into a rational function for these values of its parameters and we have

$$Q=\frac{4k(2-\sigma)}{(2+\sigma)^3}, \tag{74}$$

$$L=-\frac{ik\bar{\zeta}}{(1+\sigma/2)^2}, \tag{75}$$

where k is an arbitrary constant, characterizing the magnitude of the twist (rotation). This is the Kramer solution[1,7,8,9] which is of KS type and has $n^2=6m_0/u$. It represents a radiating Kerr metric. This is seen best in the gauge $P_u=0$.

It must be stressed that any solution of the type $P=pu$ can be transformed from the $L_u=0$ gauge to the $P_u=0$ gauge by choosing $F=u^2/2$ in Eq. (16). Then $P'=p$, $L'=(2u)^{1/2}L$, $m'=(2u)^{-3/2}m$ which transfers the u-dependence to the mass parameter and the twist.

A generalization of the Kramer solution was undertaken in Ref. 7 for the same L and the ansatz $P=A(\sigma,u)(1+\sigma/2)$. It was shown that A satisfies a second order linear equation. The results of this paper guarantee that P also satisfies a second order linear equation.

The mirror of the solution with $p=e^{ax}$ described above is $p=e^{az}$ and $Q=k(1-2a)e^{-2az}$, $a\neq 1/2$. Hence,

$$P=\sigma^a u, \tag{76}$$

$$L=ik\bar{\zeta}\sigma^{-2a}. \tag{77}$$

It is very similar to the Kramer solution for $a=1$. For $a=1/2$, P is still given by Eq. (76) while $L=ik\bar{\zeta}\sigma^{-1}\ln\sigma$. The energy density is

$$n^2=\frac{6m_0}{u}-2\delta_{1,2a}u^2. \tag{78}$$

VI. SOLUTIONS WITH EXPONENTIAL BEHAVIOR

These are solutions with $P=pe^{cu}$ and $C=-c^2$ in Eq. (46) or Eqs. (52), (53), and (54). We shall give two solutions with y-symmetry. First we take $q_x=2(1-b)x^{b-1}$, where the constant $b\neq 1$ and put this expression in Eq. (53),

$$x^2p_{xx}+(b-2)xp_x+[\tfrac{1}{4}(b-2)(b-3)+4c^2(1-b)^2x^{2b}]p=0. \tag{79}$$

When $b=0$ this is the Euler equation, similar to Eq. (64) and has three types of solutions. The first two of them read

$$p=|x|^{(3/2)\pm\gamma}, \tag{80}$$

$$p=|x|^{(3/2)}\ln|x|, \tag{81}$$

when $c^2<3/16$ or $c^2=3/16$, respectively. Here $\gamma=\frac{1}{2}(3-16c^2)^{1/2}$. When $b\neq 0$ the solution is given by Bessel functions,

$$p=x^{(3-b)/2}[C_1J_\nu(\epsilon x^b)+C_2Y_\nu(\epsilon x^b)], \tag{82}$$

where $\nu = \sqrt{3-b}/2b$, $\epsilon = 2|b-1|c/b$. In all cases

$$L = i\sqrt{2}(1-b)x^{b-1}. \tag{83}$$

These are essentially the solutions found in Ref. 4, where x-symmetry was used instead together with a special transformation between a u-independent and a u-dependent solution.

The second solution to be presented is an analog of Eq. (69), i.e., $P = e^{cu}$ depends only on u. Then Eq. (53) may be integrated to

$$Q = \tfrac{1}{2}(b^2 - \tfrac{2}{3}c^2 q_x^4 + 2dq_x)^{1/2}, \tag{84}$$

where $b > 0$ and d are constants. When $d = 0$, Eq. (84) is solved by elliptic functions,

$$L = i\frac{b}{\lambda} cn(\lambda x), \tag{85}$$

where $\lambda = (\frac{8}{3}b^2c^2)^{1/4}$ and the modulus of cn is 1/2. It is clear from Eq. (84) that $Q \leqslant b/2$. Equation (55) yields

$$n^2 = 6cm_0 - 4c^2Q^2e^{4cu}. \tag{86}$$

The second term is definitely negative and if $m_0 > 0$, $c > 0$ dominates for large u, no matter how big m_0 is. However, if $m_0 < 0$, $c < 0$ the first term remains positive while the second is damped exponentially for large retarded times. Choosing $|m_0|$ big enough we can arrange for $n^2 > 0$.

Let us give next an example of an axisymmetric solution. Let us substitute in Eq. (54) $q_\sigma = Q = a = \text{const}$ to obtain the generalization of Eq. (72),

$$p_{\sigma\sigma} + \frac{1}{\sigma}p_\sigma + a^2c^2p = 0. \tag{87}$$

It is solved by Bessel functions, e.g.,

$$P = e^{cu}J_0(ac\sigma), \tag{88}$$

$$L = ia\bar{\zeta}. \tag{89}$$

This solution was found and discussed in Ref. 9.

Finally, we give an example of a solution without y-symmetry or axial symmetry. We shall obtain the most general KS pure radiation field with $P = pe^{cu}$. The results of Sec. IV and Eq. (56) show that

$$q = \frac{1}{2ic}\ln\frac{B}{\bar{B}}, \tag{90}$$

$$P = (B\bar{B})^{1/2}e^{cu}, \tag{91}$$

where B satisfies $\psi = 0$, namely,

$$B^2(\ln B)_{\zeta\bar{\zeta}} = S(\zeta). \tag{92}$$

S is an arbitrary analytic function. If $S \neq 0$ we introduce the new variable $\omega = \int^\zeta S(\zeta')d\zeta'$ and Eq. (92) transforms into the Liouville equation,

$$B^2(\ln B)_{\omega\bar{\zeta}} = 1. \tag{93}$$

Returning to the original variables, its general solution reads

$$B = S(\zeta)^{1/2} \frac{f(\zeta) + g(\bar{\zeta})}{\sqrt{f(\zeta)_\zeta g(\bar{\zeta})_{\bar{\zeta}}}}. \tag{94}$$

The functions f and g are arbitrary. When $S=0$ the solution of Eq. (92) is

$$B = f(\zeta) g(\bar{\zeta}). \tag{95}$$

It has zero twist, i.e., it is a RT solution. Then we may set $L=q=0$ which means $f(\bar{\zeta}) \equiv g(\zeta)$ and $P = f\bar{f}e^{cu}$. A coordinate transformation sets $f=1$. Therefore solution (95) is equivalent to $B=1$.

In Refs. 3 and 15 the general axisymmetric pure radiation KS fields were found. Interestingly enough, their u-dependence covers the three types (49), (50), and (51) with $A=0$. The exponential solution was generalized in Ref. 3 to a nonsymmetric one with

$$B = G(\zeta) + \bar{\zeta} K(\zeta), \tag{96}$$

G and K being arbitrary. Equation (96) may be obtained from the general formula (94) by specializing to $g(\bar{\zeta}) = \bar{\zeta}$,

$$B = f(\zeta) \left[\frac{S(\zeta)}{f(\zeta)_\zeta} \right]^{1/2} + \bar{\zeta} \left[\frac{S(\zeta)}{f(\zeta)_\zeta} \right]^{1/2}. \tag{97}$$

The general KS field with nonradiative Maxwell field has been given in Ref. 11. It was remarked in Ref. 1 that no solution with a null Maxwell field has been found. Formulas (94) and (95) allow the study of this question in the context of solutions with separated variables. The energy density (63) tolerates exponential dependence. Equation (55) reads in this case,

$$3 m_0 c = \kappa N(\zeta) \bar{N}(\bar{\zeta}) B \bar{B}. \tag{98}$$

Obviously B cannot be given by Eq. (94) but a nontwisting solution with $B=1$ is possible. Then $P = e^{cu}$ and $\Phi_2^0 = (3 m_0 c/\kappa)^{1/2}$. The mass parameter is constant, as usual, $m = m_0$. This is nothing but a special case of the Einstein–Maxwell solution of Robinson and Trautman[10] in disguise [see also Eq. (24.41) in Ref. 1]. The conclusion is that there is no twisting KS solution with a null Maxwell field and separated variables.

VII. CONCLUSIONS

In this paper we have carried out the idea of Stephani based on the introduction of an invariant potential for algebraically special, twisting and expanding gravitational fields. It works best in the important subclass of fields with vanishing NUT parameter. The system of Einstein–Maxwell equations is reduced to the couple of Eqs. (25) and (26) for P and L where L is represented by a real function of two variables $q(\zeta,\bar{\zeta})$ (the $L_u = 0$ gauge). We have studied mainly pure radiational solutions for which Eq. (26) is an inequality. The main equation (25), originally of fourth order, becomes a linear second order equation for P (39), which further simplifies for the most common symmetries. In some cases it is linear for q_x, too. Of course, it is not possible to enumerate all solutions of Eq. (39). The method of separation of variables was used to systematize most of the known solutions obtained in the past by a variety of different approaches. We have also investigated the region where the energy density of pure radiation is positive. Samples of new solutions were given to illustrate the application of Eqs. (39), (53), (54), and (55). Three types of behavior with respect to the retarded time have been found; exponential, trigonometric, and linear. Among the new solutions is the most general KS pure radiation field with exponential time behavior. It was shown that the radiation field cannot be a null Maxwell field unless the twist vanishes. Finally, it should be mentioned that the case $m=0,\ M\neq 0$ is much more difficult because Eq. (9) remains nonlinear in P.

ACKNOWLEDGMENT

This work was supported by the Bulgarian National Fund for Scientific Research under Contract No. F-632.

[1] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).
[2] E. Herlt and H. Stephani, Class. Quantum Grav. **1**, 95 (1984).
[3] H. Stephani, Gen. Relativ. Gravit. **15**, 173 (1983).
[4] J. Tafel, P. Nurowski, and J. Lewandowski, Class. Quantum Grav. **8**, L83 (1991).
[5] J. Lewandowski P. Nurowski, and J. Tafel, Class. Quantum Grav. **8**, 493 (1991).
[6] A. Grundland and J. Tafel, Class. Quantum Grav. **10**, 2337 (1993).
[7] D. Kramer and U. Hähner, Class. Quantum Grav. **12**, 2287 (1995).
[8] U. von der Gönna and D. Kramer, Class. Quantum Grav. **15**, 2017 (1998).
[9] H. Stephani, J. Phys. A **12**, 1045 (1979).
[10] I. Robinson and A. Trautman, Proc. R. Soc. London, Ser. A **265**, 463 (1962).
[11] G. C. Debney, R. P. Kerr, and A. Schild, J. Math. Phys. **10**, 1842 (1969).
[12] D. W. Trim and J. Wainwright, J. Math. Phys. **15**, 535 (1974).
[13] A. D. Rendall, Class. Quantum Grav. **5**, 1339 (1988).
[14] D. Singleton, Class. Quantum Grav. **7**, 1330 (1990).
[15] E. Herlt, Gen. Relativ. Gravit. **12**, 1 (1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Noncanonical quantization of gravity. I. Foundations of affine quantum gravity

John R. Klauder[a)]
Departments of Physics and Mathematics, University of Florida, Gainesville, Florida 32611

(Received 11 June 1999; accepted for publication 14 July 1999)

The nature of the classical canonical phase-space variables for gravity suggests that the associated quantum field operators should obey affine commutation relations rather than canonical commutation relations. Prior to the introduction of constraints, a primary kinematical representation is derived in the form of a reproducing kernel and its associated reproducing kernel Hilbert space. Constraints are introduced following the projection operator method, which involves no gauge fixing, no complicated moduli space, nor any auxiliary fields. The result, which is only qualitatively sketched in the present paper, involves another reproducing kernel with which inner products are defined for the physical Hilbert space and which is obtained through a reduction of the original reproducing kernel. Several of the steps involved in this general analysis are illustrated by means of analogous steps applied to one-dimensional quantum mechanical models. These toy models help in motivating and understanding the analysis in the case of gravity. © *1999 American Institute of Physics.* [S0022-2488(99)02011-3]

I. INTRODUCTION

General relativity is, in certain ways, fundamentally different than most other physically relevant classical field theories, and the same remark applies to attempts to provide associated quantum formulations. The space–time metric $g_{\mu\nu}(x)$, $x\in\mathbb{R}^4$, $\mu,\nu=0,1,2,3$, possesses a signature requirement that is incompatible with the space of metrics being a linear vector space. (Although we assume a $3+1$ theory of gravity for illustrative purposes, it is straightforward to generalize to an $s+1$ theory as well, $s\geqslant 1$.) The inverse metric $g^{\sigma\mu}(x)$, defined so that $g^{\sigma\mu}(x)g_{\mu\nu}(x)=\delta^\sigma_\nu$, is classically trivial but it is quantum mechanically challenged since the left-hand side involves the product of two operator-valued distributions. Moreover, the spatial and temporal constraints that hold at each space–time point classically close algebraically, but they exhibit an anomaly (more commonly called a factor-ordering problem) when quantized. In effect, this fact changes the constraints from first class (classically) to second class (quantum mechanically). And, of course, there is the well-known fact that unlike other theories that take place on a fixed space–time stage, the theory of gravity involves the dynamics of the space–time stage itself. Our purpose in this article is to discuss some basic issues surrounding quantum gravity from a viewpoint different than traditional ones. (*Remark*: The closest work in spirit to that discussed in this paper is that of the author,[1] Isham and Kakas,[2] and especially Pilati.[3] See Sec. V for an extensive discussion. We do not directly comment on current schemes for quantizing gravity.)

Let us outline the general approach we shall adopt. First, we focus on basic kinematics and the quantum theory of positive definite, 3×3 matrix-valued field variables and associated noncanonical "conjugate" field variables, designed to offer an initial class of coherent states and coherent-state induced Hilbert space representations. In this step it is noteworthy that, besides the signature issue for the metric, the existence of an operator $g^{jk}(x)$, $j,k=1,2,3$, inverse to $g_{kl}(x)$ is shown, such that, when suitably defined, the equation $g^{jk}(x)g_{kl}(x)=\delta^j_k$ is fulfilled. Second, we introduce

[a)]Electronic mail: klauder@phys.ufl.edu

the spatial and temporal constraints in the projection operator approach recently developed by the author and others.[4–7] This procedure has the advantage of working entirely with the classical degrees of freedom, including the *c*-number Lagrange multipliers—specifically, the lapse and shift functions. It is not necessary to introduce additional fields (e.g., ghosts with false statistics), nor choose gauges, nor pass to moduli spaces, etc. Initially, the constraints are imposed in a regularized fashion. Subsequently, the removal of the regularization is analyzed, a process that often involves an automatic change of Hilbert-space representation. Assuming that the limit removing the regularization exists, the physical Hilbert space that arises is then, generally speaking, best described as a reproducing kernel Hilbert space that emerges from the reduction of the original reproducing kernel.

Before we undertake any discussion of gravity, however, we sketch in Sec. II the key concepts as applied to some simple, few degree-of-freedom systems. In Sec. III we construct a suitable kinematical framework for quantum gravity, while in Sec. IV we analyze the introduction of constraints. Section V contains a general discussion about operator representations and constraints in relation to reparametrization invariance. In Part II of this work, the analysis of gravitational constraints is discussed in detail. In addition, the classical limit of the affine gravitational quantum theory developed in Secs. III–IV will be discussed and compared with classical gravity.

II. ELEMENTARY ILLUSTRATION OF KEY CONCEPTS

As is generally well known, there is only one irreducible representation up to unitary equivalence of the canonical, self-adjoint operators P and Q satisfying the Weyl form of the canonical commutation relations. This representation, equivalent to the Schrödinger representation, implies that the spectrum of both P and Q cover the whole real line. Such operator degrees of freedom are appropriate for many systems with a finite or an infinite number of degrees of freedom, but they are inappropriate for gravity. The reason for this is that the classical 3×3 metric is strictly positive definite and the associated quantum field operator cannot be represented by an operator whose spectrum is unbounded above and below. Instead of the usual relation $[Q,P]=i$, with $\hbar=1$, one is led to consider an *affine commutation relation*,[8] which for a single degree of freedom takes the form

$$[Q,D]=iQ. \tag{1}$$

Here $D\equiv(PQ+QP)/2$ denotes the dilation operator, and it follows[9,10] that solutions of the affine commutation relations exist with irreducible, self-adjoint operators D and Q for which—and this is the important part—$Q>0$. (There are two other inequivalent self-adjoint solutions: one where $Q<0$, which is rather like the representation of interest, and another for which $Q=0$.[9] Neither of these representations will be of interest in this article.) Even though the operator P is only a symmetric operator that has no self-adjoint extension, the introduction of the self-adjoint operator D provides the substitute commutation relation given above. These two commutation relations are not in conflict since the affine commutation relation follows directly from the Heisenberg commutation relation simply by multiplication of the latter by Q. We note that an analog of the affine variables will be used in the case of the gravitational field to maintain the positivity of the local quantum field operator for the 3×3 spatial metric.

Continuing with the one-dimensional example, and based on self-adjoint operators that satisfy the affine commutation relation, let us introduce *affine coherent states*,[11] $|p,q\rangle\in\mathfrak{H}$, defined by the expression

$$|p,q\rangle\equiv e^{ipQ}e^{-i\ln(q)D}|\eta\rangle, \quad -\infty<p<\infty, \quad 0<q<\infty. \tag{2}$$

Here, the fiducial vector $|\eta\rangle$ is chosen to satisfy several conditions, which, using the shorthand $\langle(\cdot)\rangle\equiv\langle\eta|(\cdot)|\eta\rangle$, are specifically given by

$$\langle Q^{-1}\rangle\equiv C<\infty, \quad \langle\mathbf{1}\rangle=1, \quad \langle Q\rangle=1, \quad \langle D\rangle=0, \quad \langle P\rangle=0. \tag{3}$$

The first condition is required, while the remaining conditions are chosen for convenience. The coherent states also admit a resolution of unity[11] expressed in the form

$$\mathbf{1}=\int |p,q\rangle\langle p,q|\,d\tau(p,q), \quad d\tau(p,q)=dp\,dq/2\pi C, \tag{4}$$

integrated over the half-plane $\mathbb{R}\times\mathbb{R}^+$.

In particular, diagonalizing the self-adjoint operator $Q=\int_0^\infty x|x\rangle\langle x|\,dx$ in terms of standard Dirac-normalized eigenvectors, leads to a representation for the coherent-state overlap given by

$$\begin{aligned}\langle p,q|r,s\rangle &\equiv \langle \eta|e^{i\ln(q)D}e^{-ipQ}\,e^{irQ}e^{-i\ln(s)D}|\eta\rangle\\ &=(qs)^{-1/2}\int_0^\infty \eta(x/q)^* e^{-ix(p-r)}\eta(x/s)\,dx,\end{aligned} \tag{5}$$

where the fiducial function $\eta(x)=\langle x|\eta\rangle$ denotes the Schrödinger representation of the fiducial vector $|\eta\rangle$. It is important to observe, for some suitable function F, that

$$\langle p,q|r,s\rangle=F(q,p-r,s), \tag{6}$$

namely, that p and r universally enter in the form $p-r$. It is also clear that $\langle p,q|r,s\rangle$ defines a *continuous, positive-definite function*, which, apart from the continuity, means that

$$\sum_{n,m=1}^{N} \alpha_n^*\alpha_m\,\langle p_n,q_n|p_m,q_m\rangle \geqslant 0, \tag{7}$$

for arbitrary complex $\{\alpha_n\}_{n=1}^N$ and real $\{p_n,q_n\}_{n=1}^N$ sequences, with $N<\infty$. The function (5) may be taken as the *reproducing kernel for a reproducing kernel Hilbert space*.[12] Note that the information in $\langle p,q|r,s\rangle$ is enough to recover $\eta(x)$ apart from an overall constant phase factor. Thus, different fiducial functions (not related by a constant phase factor) generate distinct reproducing kernels. Since each reproducing kernel Hilbert space has one and only one reproducing kernel,[12] it follows for different $\eta(x)$ that the Hilbert space functional realizations are completely disjoint, except for the zero element. Basic elements of a dense set of vectors in each such Hilbert space are given by continuous functions of the form

$$\psi(p,q)\equiv\sum_{n=1}^{N} \alpha_n\langle p,q|p_n,q_n\rangle, \tag{8}$$

defined for arbitrary complex $\{\alpha_n\}_{n=1}^N$ and real $\{p_n,q_n\}_{n=1}^N$ sequences, with $N<\infty$. Let a second such function be given by

$$\phi(p,q)\equiv\sum_{j=1}^{J} \beta_j\langle p,q|\overline{p_j},\overline{q_j}\rangle, \tag{9}$$

defined for arbitrary complex $\{\beta_j\}_{j=1}^J$ and real $\{\overline{p_j},\overline{q_j}\}_{j=1}^J$ sequences, with $J<\infty$. The inner product of two such vectors is then *defined*[12] to be

$$\langle\psi|\phi\rangle\equiv(\psi(\cdot,\cdot)_J\,\phi(\cdot,\cdot))\equiv\sum_{n=1}^{N}\sum_{j=1}^{J} \alpha_n^*\beta_j\langle p_n,q_n|\overline{p_j},\overline{q_j}\rangle, \tag{10}$$

which when $|\phi\rangle=|\psi\rangle$ is, by definition, non-negative. The resultant pre-Hilbert space is completed to a (reproducing kernel) Hilbert space $\mathcal{C}$ by including all Cauchy sequences in the norm $\||\psi\rangle\| \equiv\sqrt{\langle\psi|\psi\rangle}$ as N tends to infinity. Finally, we note that the space of functions appropriate to one reproducing kernel is *identical* to the space of functions appropriate to a second reproducing kernel that is just a constant multiple of the first reproducing kernel. This fact does not contradict the uniqueness of the reproducing kernel for each Hilbert space because strictly different inner products are assigned in the two cases. Of course, the foregoing discussion applies quite generally and is not limited to any one sort of reproducing kernel.

When the states $|p,q\rangle$ form a set of coherent states—as we assume in the present case—the inner product has an *alternative representation* given by a local integral of the form

$$\langle\psi|\phi\rangle=\int \psi(p,q)^*\phi(p,q)\,d\tau(p,q), \tag{11}$$

expressed in terms of $\psi(p,q)\equiv\langle p,q|\psi\rangle$ and $\phi(p,q)\equiv\langle p,q|\phi\rangle$. It follows that this formula holds for all elements of the completed Hilbert space $\mathcal{C}$, and, moreover, every element of the so-completed space is a *bounded and continuous function*, the collection of which forms a rather special closed subspace of $L^2(\mathbb{R}^2,d\tau)$.

Thanks to the coherent-state resolution of unity, it follows that the coherent state overlap function satisfies the integral equation

$$\langle p'',q''|p',q'\rangle=\int \langle p'',q''|p,q\rangle\langle p,q|p',q'\rangle\,d\tau(p,q), \tag{12}$$

a basic relation, which, if it was established as a first step for the continuous function $\langle p'',q''|p',q'\rangle$ $[=\langle p',q'|p'',q''\rangle^*]$, guarantees the existence of a local integral representation for the inner product of two arbitrary elements in the associated reproducing kernel Hilbert space. Several useful properties follow from this reproducing property. For example, repeated use of the resolution of unity leads to the fact that

$$\langle p'',q''|p',q'\rangle=\lim_{L\to\infty}\int\cdots\int\prod_{l=0}^{L}\langle p_{l+1},q_{l+1}|p_l,q_l\rangle\prod_{l=1}^{L}\,d\tau(p_l,q_l), \tag{13}$$

in which we have identified $p'',q''=p_{L+1},q_{L+1}$ and $p',q'=p_0,q_0$. In turn, making an (unjustified!) interchange of the (continuum) limit with the integrations, and writing for the integrand the form it would assume for continuous and differentiable paths, gives rise to the suggestive but strictly formal expression[13]

$$\langle p'',q''|p',q'\rangle=\int e^{-i\int_0^T q(t)\,\dot p(t)\,dt}\,\mathcal{D}\tau(p,q), \tag{14}$$

which determines a formal path integral representation for the kinematics that applies for any $T>0$. Thus, the existence of a coherent state resolution of unity is the necessary and sufficient condition to introduce a traditional coherent state phase-space path integral representation for the kinematics, which specifically leads to the reproducing kernel. A path integral for the kinematics is a necessary prerequisite to obtain a path integral for the dynamics. While less than ideal, the formal path integral itself may be used as a starting point for quantization. Even though the formal nature of the path integral renders it basically undefined, one may always (re)introduce a regularization by a lattice-limit formulation (as above), using suitable ingenuity to choose an acceptable integrand. This procedure is more or less standard by now.

However, it must be appreciated that reproducing kernels for reproducing kernel Hilbert spaces are *not required* to fulfill a (positive) local integral representation for the inner product. In cases where the integral for the resolution of unity does not exist, one must accept the inner product that is given directly by the reproducing kernel, which means that the integral relation (12) is replaced by

$$(\langle\cdot,\cdot|p'',q''\rangle_J\langle\cdot,\cdot|p',q'\rangle)\equiv\langle p'',q''|p',q'\rangle. \tag{15}$$

When this is the case, we say that $\{|p,q\rangle\}$ forms a set of *weak coherent states*,[14] i.e., the elements of $\{|p,q\rangle\}$ span the Hilbert space $\mathfrak{H}$, but do not admit a local integral representation for the inner product of elements in the associated reproducing kernel Hilbert space.

A simple example of a reproducing kernel Hilbert space without a local integral representation for the inner product is determined, for $u'',u'\in\mathbb{R}$, by the reproducing kernel $\langle u''|u'\rangle \equiv\exp[-(u''-u')^2]$; here, one must use $(\langle\cdot|u''\rangle_J\langle\cdot|u'\rangle)\equiv\langle u''|u'\rangle$, which is then extended by linearity and continuity to all Hilbert space vectors.

A more relevant set of examples is given by the following discussion applied to our simple model. Let $\alpha>-\frac{1}{2}$, and choose $\eta(x)\equiv N\,x^{\alpha}\exp(-\beta x)$. Here the factor N is fixed by requiring $\int_0^\infty|\eta(x)|^2\,dx=1$. The two conditions $\langle Q\rangle=1$ and $\langle Q^{-1}\rangle=C<\infty$, lead to $\beta-\frac{1}{2}=\alpha>0$; in this case $C=1-1/(2\beta)$. In turn, the reproducing kernel is given explicitly[15] by

$$\langle p,q|r,s\rangle=\left[\frac{(qs)^{-1/2}}{\frac{1}{2}(q^{-1}+s^{-1})+i\frac{1}{2}\beta^{-1}(p-r)}\right]^{2\beta}$$

$$=\exp(-2\beta\ln\{[\tfrac{1}{2}(q^{-1}+s^{-1})+i\tfrac{1}{2}\beta^{-1}(p-r)]/(qs)^{-1/2}\}); \tag{16}$$

the second form is given for comparison purposes to the gravitational case. As long as $\beta>\frac{1}{2}$, it follows that the states $|p,q\rangle$ form a set of coherent states with a proper resolution of unity, and therefore a local integral representation for the inner product exists. In this case, path integrals exist as lattice limits, and the whole situation seems familiar. On the other hand, if $0<\beta\leqslant\frac{1}{2}$, the overlap function $\langle p,q|r,s\rangle$ defined above is still a positive-definite function and, therefore, it is a valid reproducing kernel that leads to an associated reproducing kernel Hilbert space; however, *such a Hilbert space does not admit a local integral representation for the inner product in terms of the given representatives*. Therefore, there is *no* conventional coherent state path integral for the kinematics, and thus also for the dynamics, in a reproducing kernel Hilbert space representation when $0<\beta\leqslant\frac{1}{2}$. This lack of a conventional coherent state path integral representation may appear to be detrimental to any program to introduce quantization, dynamics, etc.—but there is hope.

There is another way to generate the reproducing kernel for the given family of fiducial vectors that may be applied for all $\beta>0$. Let us first focus on $\beta>\frac{1}{2}$. In that case, observe, by construction and using $\partial_p\equiv\partial/\partial p$, etc., that for every $|\psi\rangle\in\mathfrak{H}$,

$$B\,\psi(p,q)\equiv\{-iq^{-1}\partial_p+1+\beta^{-1}q\,\partial_q\}\,\psi(p,q)=0, \tag{17}$$

an equation that represents a (complex) *polarization*[16] of $L^2(\mathbb{R}^2,d\tau)$. It follows that the second-order differential operator $A\equiv\frac{1}{2}\beta B^\dagger B\geqslant 0$, and therefore A can be used to generate a semigroup. In particular, for any $T>0$ and as $\nu\rightarrow\infty$, the expression $e^{-\nu TA}$ becomes a *projection operator* onto the subspace $\mathcal{C}$ of solutions to the polarization equation.[17] In a two degree of freedom Schrödinger representation—symbolized by $|p,q)$, where $(p,q)\in\mathbb{R}\times\mathbb{R}^+$ and $(p,q|r,s)=\delta(p-r)\delta(q-s)$—it follows, from a two-variable Feynman–Kac–Stratonovich path integral formula,[18] that

$$\langle p'',q''|p',q'\rangle\equiv\lim_{\nu\to\infty}(p'',q''|e^{-\nu TA}|p',q')$$

$$=\lim_{\nu\to\infty}\mathcal{N}\int e^{-i\int_0^T q(t)\,\dot p(t)\,dt-(1/2\nu)\int_0^T[\beta^{-1}q(t)^2\dot p(t)^2+\beta q(t)^{-2}\dot q(t)^2]\,dt}\,\mathcal{D}p\,\mathcal{D}q$$

$$\equiv\lim_{\nu\to\infty}e^{\nu T/2}\int e^{-i\int_0^T q(t)\,dp(t)}\,dW^{\nu}(p,q), \tag{18}$$

where W^ν denotes a two-dimensional Wiener measure with diffusion constant ν, pinned at $t=0$ to p',q' and at time $t=T$ to p'',q'', which is supported on a space of constant negative curvature $R=-2/\beta$. It is noteworthy in (18) that the variable $p(t)$ enters only in the form $\dot p(t)$, a fact which leads to the result depending only on the difference, $p''-p'$. For every $\nu<\infty$, and with probability one, all Wiener paths in the given path integral are *continuous*, and, for purposes of coordinate transformations, it is convenient to adopt the (midpoint) Stratonovich rule to define the stochastic integral $-\int q(t)\,dp(t)$ since, in that case, the rules of the ordinary calculus hold. Such a representation is said to involve a *continuous-time regularization*.[19]

Now we consider the case where $0<\beta\leqslant\frac{1}{2}$. The solutions to (17) are, up to a factor, analytic functions, but they are no longer square integrable, as is clear from the fact that the large q behavior of (16) is controlled only by the factor $q^{-\beta}$. As a consequence, the operator $A\geqslant 0$ has only a continuous spectrum. The family of operators $e^{-\nu TA}$ is still a semigroup, and the expression $(p'',q'|e^{-\nu TA}|p',q')$ still has the formal path integral representation given in the middle line of (18). However, as $\nu\to\infty$, $e^{-\nu TA}$ does not lead to a projection operator; instead we need to extract the germ of that semigroup as $\nu\to\infty$. If we let $E\geqslant 0$ denote continuum eigenvalues for the operator $\frac{1}{2}B^{\dagger}B$, we can write

$$(p'',q''|e^{-\nu TA}|p',q')=\int (p'',q''|E,v)\,e^{-\nu T\beta E}(E,v|p',q')\,\rho(E,v)\,dE\,dv, \tag{19}$$

for some density of states $\rho(E,v)$. Here, the variable v labels degeneracy for $\frac{1}{2}B^{\dagger}B$. For the sake of illustration, let us assume that $\rho(E,v)\simeq\overline{C}E^{w}\,\overline{\rho}(v)$, $w>-1$, for $E\ll 1$. We base this assumption on the fact that there is no reason for an E-dependent degeneracy for very tiny E. Thus, we consider the expression $J(\nu)\equiv(\nu\beta T)^{w+1}/\overline{C}\,\Gamma(w+1)$, and are led to the fact that

$$\begin{aligned}\lim_{\nu\to\infty} J(\nu)\,(p'',q''|e^{-\nu TA}|p',q')\\ &=\lim_{\nu\to\infty}\int (p'',q''|E,v)\,e^{-\nu T\beta E}(E,v|p',q')\,J(\nu)\,\rho(E,v)\,dE\\ &=\int (p'',q''|0,v)(0,v|p',q')\,\overline{\rho}(v)\,dv.\end{aligned} \tag{20}$$

In effect, this procedure has enabled us to pass to the germ of the semigroup. Observe that the rescaling factor is independent of the coherent state labels, and thus we are only making a ν-dependent rescaling before the limit $\nu\to\infty$ is taken. If necessary, we can rescale our expression by letting $J(\nu)\to\overline{J}(\nu)=M(p'',q'')\,M(p',q')\,J(\nu)$ to achieve normalization without affecting its positive-definite character. In summary, for $0<\beta\leqslant\frac{1}{2}$, we claim that instead of (18) we can write

$$\begin{aligned}\langle p'',q''|p',q'\rangle&\equiv\lim_{\nu\to\infty}\overline{J}(\nu)(p'',q''|e^{-\nu TA}|p',q')\\ &=\lim_{\nu\to\infty}\overline{\mathcal{N}}\int e^{-i\int_0^T q(t)\,\dot{p}(t)\,dt-(1/2\nu)\int_0^T[\beta^{-1}q(t)^2\dot{p}(t)^2+\beta q(t)^{-2}\dot{q}(t)^2]\,dt}\,\mathcal{D}p\,\mathcal{D}q\\ &\equiv\lim_{\nu\to\infty}\overline{J}(\nu)\,e^{\nu T/2}\int e^{-i\int_0^T q(t)\,dp(t)}\,dW^{\nu}(p,q).\end{aligned} \tag{21}$$

Convergence in this case is initially regarded in the sense of distributions. To lead to the desired result, we appeal to analyticity (up to a specific factor) of the result in $q^{-1}+i\beta^{-1}p$, analyticity (up to another factor) in the variable $s^{-1}-i\beta^{-1}r$, and dependence on $p-r$ [cf. (16)].Note that $\overline{J}(\nu)$ can always be determined self-consistently by insisting that $\langle p,q|p,q\rangle=1$ for all (p,q). In simpler terms, we can always regard $\overline{J}(\nu)$ as part of the needed normalization coded into $\overline{\mathcal{N}}$ in the formal path integral expression.

As will become evident later, various features of reproducing kernels, reproducing kernel Hilbert spaces, and associated rules for defining inner products illustrated above will carry over into the quantum gravity case as well.

A. Operators and symbols

In addition to the properties of the reproducing kernel Hilbert space, certain *symbols* associated with operators are important. Let us introduce the upper symbol $H(p,q)$ associated to the operator $\mathcal{H}(P,Q)$ and defined, modulo suitable domain conditions, by the expression

$$H(p,q)\equiv\langle p,q|\mathcal{H}(P,Q)|p,q\rangle=\langle\mathcal{H}(p+P/q,qQ)\rangle. \tag{22}$$

For example, if $\mathcal{H}(P,Q)=P^2-Q^{-1}$ denotes a quantum Hamiltonian, then $H(p,q)=p^2+\langle P^2\rangle/q^2-C/q$. Since $C=O(1)$ (e.g., for large β) and $\langle Q\rangle=1$, it follows that $\langle P^2\rangle=O(\hbar^2)$. Observe that H basically agrees with the expected classical Hamiltonian in the limit that $\hbar\rightarrow 0$, but prior to that limit H includes a quantum-induced barrier to singularities in solutions of the usual classical equations of motion. We adopt the expression $H(p,q)$ as the ($\hbar$-augmented) classical Hamiltonian and refer to the connection between the quantum generator $\mathcal{H}$ and the classical generator H as the *weak correspondence principle*.[20] In this way, the classical and quantum theories may both *coexist*, as they do in Nature.

There is also another set of symbols that are important. We introduce the lower symbol $h(p,q)$ that is related to the operator $\mathcal{H}(P,Q)$ by the expression

$$\mathcal{H}(P,Q)=\int h(p,q)\,|p,q\rangle\langle p,q|\,d\tau(p,q). \tag{23}$$

For the one-parameter class of fiducial vectors leading to (16), it follows that a dense set of operators admit such a symbol for a reasonable set of functions.

In the quantum gravity case, there are twin goals: (i) to ensure that the field operators of interest are well defined and locally self-adjoint in the given field operator representation; and (ii) to choose locally self-adjoint constraint operators that have a weak correspondence principle which connects them with the desired form of the classical constraint generators (possibly $\hbar$ augmented).

B. Imposition of constraints

We adopt the projection operator approach to the quantization of systems with constraints.[4–7] Let $\{\Phi_\alpha(P,Q)\}_{\alpha=1}^{A}$, $A<\infty$, denote a set of constraints, each element given by a self-adjoint operator. Further, assume that $\Phi\cdot\Phi\equiv\Sigma_{\alpha=1}^{A}(\Phi_\alpha)^2$ is also self adjoint. We define the (provisional) physical Hilbert space $\mathfrak{H}_{\text{phys}}\equiv\mathbb{E}\mathfrak{H}$, where $\mathbb{E}=\mathbb{E}^\dagger=\mathbb{E}^2$ is a uniquely defined projection operator, and in turn choose

$$\mathbb{E}\equiv\mathbb{E}(\Phi\cdot\Phi\leqslant\delta(\hbar)^2), \tag{24}$$

where $\delta(\hbar)$ is *not* a δ-function but a small, positive, possibly $\hbar$-dependent, *regularization parameter* for the set of constraints. As shown below, $\delta(\hbar)$ is chosen so that $\mathbb{E}$ is the desired projection operator. This choice may entail a specific representation of the Hilbert space and a suitable limit as $\delta\rightarrow 0$ to extract the germ of the projection operator.

We may illustrate this latter situation for the constraint $\Phi\equiv\mathrm{Q}-1$, assuming initially that $\delta<1$. In this case

$$\langle\psi|\mathbb{E}((Q-1)^2\leqslant\delta^2)|\phi\rangle=\int_{1-\delta}^{1+\delta}dx\int d\sigma(y)\,\psi(x,y)^*\phi(x,y), \tag{25}$$

where σ accounts for any degeneracy that may be present. When restricted to functions ψ_0 and ϕ_0 in the dense set $\mathfrak{D}$, where (say)

$$\mathfrak{D}\equiv\{\text{ polynomial }(x,y)\,e^{-x^2-y^2}\}, \tag{26}$$

and rescaled by a suitable factor, the projection operator matrix elements lead to the expression

$$(2\delta)^{-1}\langle\psi_0|\mathbb{E}((Q-1)^2\leq\delta^2)|\phi_0\rangle=(2\delta)^{-1}\int_{1-\delta}^{1+\delta}dx\int d\sigma(y)\ \psi_0(x,y)^*\phi_0(x,y). \tag{27}$$

Now, as $\delta\to 0$, this expression becomes

$$\int \psi_0(1,y)^*\phi_0(1,y)\,d\sigma(y)\equiv((\psi_0,\phi_0)). \tag{28}$$

Interpreting this final expression as a sequilinear form, one completes the desired Hilbert space by adding all Cauchy sequences in the associated norm $\||\psi_0))\|\equiv\sqrt{((\psi_0,\psi_0))}$. The result is the true physical Hilbert space in which the constraint $Q-1=0$ is fulfilled, and this example illustrates how constraints are to be treated when $\Phi\cdot\Phi$ has its zero in the continuous spectrum.

A second example of an imposition of constraints is given by $\Phi_1=Q-1$, as before, along with $\Phi_2=D$. This situation corresponds to second class constraints, and serves as a simple qualitative model of what occurs in the gravitational case. In this case

$$\mathbb{E}=\mathbb{E}(D^2+(Q-1)^2\leq\delta(\hbar)^2). \tag{29}$$

Here, the left-hand side of the argument can be regarded as a ''Hamiltonian,'' and the ground state $|0'\rangle$ for such a system (nondegenerate, in the present example) can be sought. In particular, there are two positive parameters, δ' and δ'', such that, for all δ with $\delta'\leq\delta<\delta''$, then

$$\mathbb{E}=\mathbb{E}(D^2+(Q-1)^2\leq\delta(\hbar)^2)\equiv|0'\rangle\langle 0'|. \tag{30}$$

This is the desired choice to make for $\mathbb{E}$ in the case where $\Phi\cdot\Phi$ has a discrete spectrum near zero that does not include zero.

For completeness, if $\Phi\cdot\Phi$ has a discrete spectrum including zero, then it suffices to choose

$$\mathbb{E}=\mathbb{E}(\Phi\cdot\Phi=0)=\mathbb{E}(\Phi\cdot\Phi\leq\delta(\hbar)^2), \tag{31}$$

where in the present case $\delta(\hbar)>0$ is chosen small enough to include only the subspace for which $\Phi\cdot\Phi=0$. This simple case does not seem to arise in quantum gravity. See Refs. 4,7 for a discussion of gauge invariance.

When the number of constraints is infinite, $A=\infty$, as will be the case for a field theory, then a slightly different approach is appropriate. One form this takes is dealt with in Sec. IV.

C. Appearance of time

In a reparametrization-invariant problem in quantum mechanics it is typical that dynamics is cast in the guise of kinematics at the expense of introducing an additional degree of freedom plus a first-class constraint; see, e.g., Ref. 21. Let the resultant kinematical reproducing kernel with the extra degree of freedom be given by $\langle p'',q'',s'',t''|\mathbb{E}|p',q',s',t'\rangle$, where $\mathbb{E}$ is the projection operator enforcing the constraint. Next, reduce this expression, for example, as in the procedure

$$\langle p'',q'',t''|p',q',t'\rangle\equiv\int\int\langle p'',q'',s'',t''|\mathbb{E}|p',q',s',t'\rangle\,ds''\,ds'. \tag{32}$$

The result is a new positive-definite function that can be used to define a reproducing kernel Hilbert space. However, it may well happen that the following identity holds:

$$\begin{aligned}\langle p'',q'',t''|p',q',t'\rangle&=(\langle\cdot,\cdot,\cdot|p'',q'',t''\rangle_J\langle\cdot,\cdot,\cdot|p',q',t'\rangle)\\&=(\langle\cdot,\cdot,t|p'',q'',t''\rangle_J\langle\cdot,\cdot,t|p',q',t'\rangle).\end{aligned} \tag{33}$$

This equation means that the space spanned by the states $|p,q,t\rangle$ by varying p, q, *and t* is the *same space* spanned, by the same states, by varying p and q but with t held *fixed* at some value (e.g.,

$t=0$). This situation implies that the states $|p,q,t\rangle$ are *extended coherent states* in the sense of Ref. 22, and, in particular, thanks to using canonical group coordinates in the coherent-state parametrization, that $|p,q,t\rangle=\exp(-i\mathcal{H}t)|p,q,0\rangle$ for some self-adjoint ''Hamiltonian'' $\mathcal{H}$. The parameter t is then recognized as the ''time'' t. For an explicit example of how this procedure works in detail, see Ref. 23.

It is expected that a suitable time parameter will emerge in the extension of these ideas to the gravitational case.

D. Matrix generalization

Our preceding analysis has been confined to a single p and q and the associated affine quantum operators. In any generalization to the gravitational case, it will first be necessary to generalize the preceding discussion to 3×3 (or more generally to $s\times s$) matrix degrees of freedom and repeat an analysis similar to that of the present section. We do not include this discussion here since that is the subject of a separate work.[24] It is safe to say, apart from some technical details, that there are no special surprises in this generalization, and the basic concepts that we shall need are already present in the simplest case on which we have concentrated.

III. GRAVITATIONAL KINEMATICS

A. Preliminaries

Let us start with the introduction of a three-dimensional topological space $\mathcal{S}$ that is locally isomorphic to a subset of $\mathbb{R}^3$. Locally, we generally use three ''spatial'' coordinates, say x^j, $j=1,2,3$, to label a point in $\mathcal{S}$. This labeling is nonsingular and thus one-to-one. Whether a single coordinate chart covers $\mathcal{S}$ depends on the global topological structure of $\mathcal{S}$. Let us fix this global topological structure from the outset—for example, topologically equivalent to $\mathbb{R}^3$, S^3, T^3, etc. The theory of quantum gravity developed here does not engender topological changes of the underlying topological space $\mathcal{S}$. Note this lack of topological change applies only to the space $\mathcal{S}$. It is unrelated to any presumed ''space'' and/or ''topology'' associated with any quantum metric tensor, which, after all, is typically distributional in character.

If the space $\mathcal{S}$ is such that more than one coordinate patch is required we arrange for the necessary matching conditions and rename the coordinates within each patch by x^j, $j=1,2,3$, for some domain. We can also consider alternative coordinates, say $\overline{x^j}$, $j=1,2,3$, which are also nonsingular. We admit only differentiable coordinate transformations such that the Jacobian $[\partial x/\partial\overline{x}]\equiv\det(\partial x^j/\partial\overline{x^k})\neq0$ everywhere. The group composed of such invertible coordinate transformations is the *diffeomorphism group*.

We can also introduce functions on the space $\mathcal{S}$ which in coordinate form may be denoted by $f(x)$. A scalar function is one for which $\overline{f}(\overline{x})=f(x)$, while a scalar density of weight one satisfies $\overline{b}(\overline{x})=[\partial x/\partial\overline{x}]\,b(x)$, or stated as a volume form, $dV=\overline{b}(\overline{x})\,d^3\overline{x}=b(x)\,d^3x$. Observe that it is not necessary to have a metric in order to have a volume form. Whether $\mathcal{S}$ is compact or noncompact, we assume that $0<b(x)<\infty$ for all x and therefore $0<b(x)^{-1}<\infty$ for all x as well. These properties are still valid after a nonsingular coordinate change. Integrals of a scalar function take the form $\int f\,dV=\int f(x)\,b(x)\,d^3x$ and are invariant under any coordinate transformation in the diffeomorphism group.

In the ADM (Arnowitt, Deser, Misner)[25] canonical formulation of classical gravity there are two fundamental fields $g_{kl}(x)$ $[=g_{lk}(x)]$ and $\pi^{kl}(x)$ $[=\pi^{lk}(x)]$. The metric $g_{kl}(x)$ transforms as a (two-valent covariant) tensor, while the momentum $\pi^{kl}(x)$ transforms as a (two-valent contravariant) tensor density of weight one. Thus $\int g_{kl}(x)\,\pi^{kl}(x)\,d^3x$ [or even $\int g_{kl}(x,t)\,\dot{\pi}^{kl}(x,t)\,d^3x\,dt$] is an invariant under diffeomorphism group transformations (on the spatial hyperspace, of course). Note well: The latter example pertains to a generalization of the former one, including an additional independent variable t that possibly could be identified with (coordinate) ''time.''

A metric is not an arbitrary tensor but is restricted to be positive definite. Specifically, for any real α^j, $j=1,2,3$, where $\Sigma_{j=1}^{3}(\alpha^j)^2>0$, it follows that $\alpha^k g_{kl}(x)\,\alpha^l>0$ for all x. As a consequence, the positive-definite (two-valent contravariant) tensor $g^{kl}(x)$ exists at each point and is defined so that $g^{kl}(x)\,g_{lm}(x)=\delta^k_m$. In addition, $\sqrt{g(x)}\equiv\sqrt{\det[g_{kl}(x)]}>0$ transforms as a scalar density of weight one. Thus, as is well known, $\sqrt{g(x)}\,d^3x$ characterizes a volume form, but this choice ties the volume form to a specific metric, or at least to a specific class of metrics. This close association to specific metrics is something we would like to avoid, and it leads us to choose $b(x)\,d^3x$ as the preferred volume form. Of course, if $b(x)=\sqrt{g(x)}$ everywhere in any coordinate system, then the volume form $b(x)\,d^3x$ is identical to the one based on a metric space and given by $\sqrt{g(x)}\,d^3x$.

B. Reproducing kernel—Original Hilbert space

A study of canonical quantum gravity begins with the introduction of metric and momentum local quantum field operators, which we denote by $\sigma_{kl}(x)$ $[=\sigma_{lk}(x)]$ and $\mu^{kl}(x)$ $[=\mu^{lk}(x)]$, respectively. For such fields one postulates the canonical commutation relations,

$$[\sigma_{kl}(x),\,\mu^{rs}(y)]=i\,\delta^{rs}_{kl}\,\delta(x,y),$$
$$[\sigma_{kl}(x),\,\sigma_{rs}(y)]=0, \tag{34}$$
$$[\mu^{kl}(x),\,\mu^{rs}(y)]=0,$$

with $\delta^{rs}_{kl}\equiv(\delta^r_k\delta^s_l+\delta^r_l\delta^s_k)/2$. Since the right-hand side of the first equation is a tensor density of weight one, it is consistent that we define $\sigma_{kl}(x)$ to be a tensor and $\mu^{rs}(x)$ to be a tensor density of weight one. However, just as its one-dimensional counterpart, there are no local self-adjoint field and momentum operators that satisfy the canonical commutation relations as well as the requirement that $\{\sigma_{kl}(x)\}>0$. To arrive at a suitable substitute set of commutation relations, we introduce, along with the local metric field operator $\sigma_{kl}(x)$, the local ''scale'' field operator $\kappa^r_k(x)$, which together obey the *affine commutation relations*[1–3]

$$[\kappa^r_k(x),\,\kappa^s_l(y)]=i\,\tfrac{1}{2}[\delta^s_k\,\kappa^r_l(x)-\delta^r_l\,\kappa^s_k(x)]\,\delta(x,y),$$
$$[\sigma_{kl}(x),\,\kappa^r_s(y)]=i\,\tfrac{1}{2}[\delta^r_k\,\sigma_{ls}(x)+\delta^r_l\,\sigma_{ks}(x)]\,\delta(x,y), \tag{35}$$
$$[\sigma_{kl}(x),\,\sigma_{rs}(y)]=0.$$

In these relations, $\sigma_{kl}(x)$ remains a tensor, while $\kappa^r_s(x)$ is a tensor density of weight one under coordinate transformations. The local operators $\kappa^r_s(x)$ are generators of the $GL(3,R)^\infty$ group,[2,3] while the local operators $\sigma_{kl}(x)$ are commuting ''translations'' coupled with the $GL(3,R)^\infty$ group by a semidirect product. The given affine commutation relations are the natural generalization of the one-dimensional affine commutation relation presented in (1). In the case of one degree of freedom, the affine commutation relations follow from the canonical ones, while in the case of fields this is, strictly speaking, incorrect. It is true that

$$\kappa^r_k(x)=\tfrac{1}{2}[\sigma_{kl}(x)\,\mu^{lr}(x)+\mu^{rl}(x)\sigma_{lk}(x)]_R, \tag{36}$$

where the subscript R denotes an infinite multiplicative renormalized product to be defined later. However, the presence of an infinite rescaling means that either the canonical or the affine set of commutation relations can hold, but not both at the same time. Since it is the affine commutation relations that are consistent with local self-adjoint operator solutions enjoying metric positivity, we shall adopt the noncanonical affine commutation relations. The choice of the affine commutation relations means that the canonical commutation relations do *not* hold, and therefore we are dealing with a *non*canonical quantization of the gravitational field.

Accepting the affine field operators as generators, we introduce a primary set of normalized affine coherent states each of which—in a deliberate abuse of notation—is defined by

$$|\pi,g\rangle \equiv e^{i\int \pi^{kl}(x)\,\sigma_{kl}(x)\,d^3x}\,e^{-i\int \gamma^r_s(x)\,\kappa^s_r(x)\,d^3x}\,|\eta\rangle, \tag{37}$$

for a suitable fiducial vector $|\eta\rangle$ characterized below. In $|\pi,g\rangle$ the argument ''π'' denotes the momentum matrix field π^{ab} while ''g'' denotes the metric matrix field g_{ab}. By all rights, the states in question should have been called $|\pi,\gamma\rangle$, but as we shall see, the overlap of two such states, *for the featured choice of* $|\eta\rangle$ [cf. (70)], depends only on the matrix (for each point x) $g \equiv \exp(\gamma^T/2)\exp(\gamma/2)\equiv\{g_{ab}\}$, where T means ''transpose.'' [Observe by this parametrization that $g>0$, as opposed to a traditional triad for which $g\geqslant 0$. To see that this distinction may possibly make a real difference, see Ref. 15. We remark that the nonsymmetric matrix $\exp(\gamma/2)$ would have relevance for spinor fields.] If the space $\mathcal{S}$ is noncompact, then, as smooth c-number fields, both π and γ should go to zero sufficiently fast so that the indicated smeared field operators are indeed self-adjoint operators and generate unitary transformations as required. On the other hand, as we shall shortly see, this asymptotic behavior can, effectively, be significantly relaxed.

The overlap of two such coherent states leads to an expression of the form

$$\langle \pi'',g''|\pi',g'\rangle = F(g'',\pi''-\pi',g'), \tag{38}$$

for some continuous functional F that depends only on the difference of the fields, $\pi''(x)-\pi'(x)$, an analog of which already occurred for the one-dimensional example. Whatever choice is made for the fiducial vector $|\eta\rangle$, the coherent state overlap function defines a continuous, positive-definite functional, which, therefore, defines a reproducing kernel and its associated (separable) reproducing kernel Hilbert space $\mathcal{C}$. By construction, therefore, the set of coherent states $\{|\pi,g\rangle\}$ span the Hilbert space $\mathfrak{H}$. As such they form a basis (overcomplete to be sure!) for $\mathfrak{H}$. Based on arguments to follow, we are led to the proposal [cf., (16)] that

$$\langle \pi'',g''|\pi',g'\rangle = \exp\left(-2\int b(x)\,d^3x\,\ln \left\{\frac{\det\{\tfrac{1}{2}[g''^{kl}(x)+g'^{kl}(x)]+i\tfrac{1}{2}b(x)^{-1}[\pi''^{kl}(x)-\pi'^{kl}(x)]\}}{(\det[g''^{kl}(x)])^{1/2}(\det[g'^{kl}(x)])^{1/2}}\right\}\right). \tag{39}$$

This equation is central to our analysis of quantum gravity.

The coherent-state overlap (39) may be read in two qualitatively different ways. Although arrived at on the basis that $\pi''(x),\pi'(x)\to 0$ and $\gamma''(x),\gamma'(x)\to 0$ as $|x|\to\infty$, the given expression exists for a far wider limiting behavior. In particular, suppose there is a *fixed asymptotic behavior* such that for both $\pi=\pi''$ and $\pi=\pi'$, $\pi^{ab}(x)-\tilde{\pi}^{ab}(x)\to 0$ and for both $g=g''$ and $g=g'$, $g_{ab}(x)-\tilde{g}_{ab}(x)\to 0$, all terms vanishing sufficiently fast as $|x|\to\infty$. In this case $g\equiv \exp(\gamma^T/2)\tilde{g}\exp(\gamma/2)$. Note that the asymptotic fields can depend on x. In this case the coherent-state overlap still holds in the form given. This kind of asymptotic behavior reflects a change of the fiducial vector $|\eta\rangle$, which now depends on the explicitly chosen asymptotic form for the momentum and metric—or, equivalently, as we effectively do, one can hold $|\eta\rangle$ fixed and change the representations of the operator [cf. (71)]. By choosing a suitable asymptotic momentum and metric one can, in effect, redefine the topology of the underlying space $\mathcal{S}$. However, for simplicity, we shall assume simple Euclidean-like asymptotic behavior of the momentum and metric [$\tilde{\pi}^{ab}(x)\equiv 0$ and $\tilde{\gamma}^r_s(x)\equiv 0$, i.e., $\tilde{g}_{ab}(x)\equiv\delta_{ab}$].

A second way to study the coherent-state overlap is under coordinate transformations. Observe that $\langle \pi'',g''|\pi',g'\rangle$ is *invariant* if, everywhere, we make the replacements

(i) $b(x)\,d^3x$, by $\overline{b(\overline{x})}\,d^3\overline{x} = b(x)\,d^3x$,

(ii) $g^{kl}(x)$, by $\overline{g}^{kl}(\overline{x}) = M^k_r(x)\,g^{rs}(x)\,M^l_s(x)$,

(iii) $b(x)^{-1}\pi^{kl}(x)$, by $\overline{b(\overline{x})}^{-1}\overline{\pi}^{kl}(\overline{x})=b(x)^{-1}M_r^k(x)\pi^{rs}(x)M_s^l(x)$,

all for an arbitrary nonsingular matrix $M\equiv\{M_r^k\}$, $M_r^k(x)\equiv(\partial\overline{x}^k/\partial x^r)(x)$, which arises from a nonsingular coordinate transformation $x\rightarrow\overline{x}=\overline{x}(x)$. It suffices to restrict attention to those coordinate transformations continuously connected to the identity. When $\mathcal{S}$ is compact, a wide class of M is allowed; when $\mathcal{S}$ is noncompact, the allowed elements M must also map coherent states into coherent states for the same fiducial vector. This restriction excludes any connection by coordinate transformations of two field sets with fundamentally different asymptotic behavior; such fields live in disjoint sets. The invariance under admissible coordinate transformations is symbolized by the statement that

$$\langle\overline{\pi''},\overline{g''}|\overline{\pi'},\overline{g'}\rangle=\langle\pi'',g''|\pi',g'\rangle, \tag{40}$$

for all suitable M. Since the allowed M form a representation of the connected component of the diffeomorphism group, it follows from this identity, and suitable continuity, that for sufficiently restricted M, the transformation $|\pi,g\rangle\rightarrow|\overline{\pi},\overline{g}\rangle$ is induced by a *unitary transformation*, specifically that

$$|\overline{\pi},\overline{g}\rangle\equiv U(M)|\pi,g\rangle, \tag{41}$$

$$U(M)\equiv\exp\left[-i\int N^j(x)\mathcal{H}_j(x)d^3x\right]. \tag{42}$$

Here $\mathcal{H}_j(x)$ denotes a local operator tensor density of weight one while $N^j(x)$ denotes a *c*-number tensor with sufficiently rapid decay at spatial infinity. Furthermore, using the shorthand that $\int N^jH_j\equiv\int N^j(y)H_j(y)d^3y$, the connection between M and N^j is implicitly given by

$$M_r^a g^{rs} M_s^b=g^{ab}-\left\{\int N^jH_j,g^{ab}\right\}+\frac{1}{2!}\left\{\int N^kH_k,\left\{\int N^jH_j,g^{ab}\right\}\right\}+\ldots\equiv e^{-\{\int N^jH_j,\cdot\}}g^{ab}, \tag{43}$$

where $\{\cdot,\cdot\}$ denotes the classical Poisson brackets, and specifically, e.g., $\{g_{ab}(x),\pi^{rs}(y)\}=\delta_{ab}^{rs}\delta(x,y)$. In this expression, $H_j(x)=-2g_{jk}(x)\pi^{kl}_{|l}(x)$, $j=1,2,3$, where $(\ )_{|l}$ is the covariant derivative with respect to the 3×3 metric, denotes the classical generators of the diffeomorphism group.[26] The relationship of $H_j(y)$ and $\mathcal{H}_j(y)$ may be determined as follows. Expansion of the relation

$$\langle\pi'',g''|e^{-i\int N^j(x)\mathcal{H}_j(x)d^3x}|\pi',g'\rangle=\langle\pi'',g''|\overline{\pi'},\overline{g'}\rangle, \tag{44}$$

to first order in N^j, leads to

$$\begin{aligned}\langle\pi'',g''|&\int N^j(x)\mathcal{H}_j(x)d^3x|\pi',g'\rangle/\langle\pi'',g''|\pi',g'\rangle\\&=-i2\int b(x)d^3x([g''^{kl}(x)+g'^{kl}(x)]+ib(x)^{-1}[\pi''^{kl}(x)-\pi'^{kl}(x)])^{-1}\\&\quad\times[\delta g'^{kl}(x)-ib(x)^{-1}\delta\pi'^{kl}(x)]+i\int b(x)d^3x\ g'_{kl}(x)\delta g'^{kl}(x),\end{aligned} \tag{45}$$

where

$$\delta g'^{kl}(x)\equiv g'^{kl}_{,j}(x)N^j(x)-g'^{jl}(x)N^k_{,j}(x)-g'^{kj}(x)\,N^l_{,j}(x), \tag{46}$$

and likewise for $\delta\pi'^{kl}(x)$. This relation determines the coherent state matrix elements of $\mathcal{H}_j(x)$. Finally, observe that the diagonal coherent state matrix elements read as

$$\langle \pi,g|\mathcal{H}_j(x)|\pi,g\rangle = -2g_{jk}(x)\pi^{kl}_{|l}(x), \tag{47}$$

in conformity with the weak correspondence principle.

C. Path integral construction

If the given coherent states $|\pi,g\rangle$ possessed a resolution of unity, namely a non-negative measure $\rho(\pi,g)$ (countably or even finitely additive), such that

$$\int \langle \pi'',g''|\pi,g\rangle\langle \pi,g|\pi',g'\rangle d\rho(\pi,g) = \langle \pi'',g''|\pi',g'\rangle, \tag{48}$$

then the construction of a path integral for the reproducing kernel would be straightforward and would follow the pattern illustrated in Sec. II for a single degree of freedom. However, for the proposed reproducing kernel $\langle \pi'',g''|\pi',g'\rangle$ given in (39), no such measure exists and thus the traditional resolution of unity is unavailable. Consequently, as defined, $\{|\pi,g\rangle\}$ is a set of weak coherent states.

A similar kind of problem arose in the simple model discussed in Sec. II (when $0<\beta\leqslant\frac{1}{2}$). In that case, the construction of a path integral representation proceeded in an alternative manner, beginning first with a polarization. We assert that each of the given Hilbert space representatives,

$$\psi(\pi,g) \equiv \langle \pi,g|\psi\rangle = \sum_{n=1}^{N} \alpha_n \langle \pi,g|\pi_n,g_n\rangle \in \mathcal{C}, \tag{49}$$

satisfies the functional differential equation [cf. (17)],

$$B^r_s(x)\psi(\pi,g) \equiv \left[-ig^{rt}(x)\frac{\delta}{\delta\pi^{ts}(x)} + \delta^r_s + b(x)^{-1}g_{st}(x)\frac{\delta}{\delta g_{tr}(x)}\right]\psi(\pi,g) = 0, \tag{50}$$

for all spatial points x. Next, let us introduce the operator

$$\mathcal{A} \equiv \tfrac{1}{2}\int B^s_r(x)^\dagger B^r_s(x) b(x) d^3x, \tag{51}$$

and observe that $\mathcal{A}\geqslant 0$. Thus, with $T>0$ and as $\nu\to\infty$, it follows that $\overline{\mathcal{J}}(\nu)e^{-\nu T\mathcal{A}}$, for some $\overline{\mathcal{J}}(\nu)$, serves to select out the subspace where (50) is fulfilled. Just as in the toy model of Sec. II, the operator $\mathcal{A}$ is a second-order (functional) differential operator, and, as a consequence, a Feynman–Kac–Stratonovich path (i.e., functional) integral representation may be introduced. In particular, we obtain the formal expression [cf. (21)],

$$\langle \pi'',g''|\pi',g'\rangle = \lim_{\nu\to\infty}\overline{\mathcal{N}}\int \exp\left[-i\int g_{ab}\dot{\pi}^{ab}\,d^3x\,dt\right]\exp\left\{-(1/2\nu)\int [b(x)^{-1}g_{ab}g_{cd}\dot{\pi}^{bc}\dot{\pi}^{da}\right.$$
$$\left. + b(x)g^{ab}g^{cd}\dot{g}_{bc}\dot{g}_{da}]d^3x\,dt\right\}\prod_{x,t}\prod_{a\leqslant b} d\pi^{ab}(x,t)dg_{ab}(x,t). \tag{52}$$

Here, let us interpret t, $0\leqslant t\leqslant T$, as coordinate ''time.'' On the right-hand side the canonical fields are functions of space and time, that is,

$$g_{ab} = g_{ab}(x,t), \quad \pi^{ab} = \pi^{ab}(x,t); \tag{53}$$

the overdot ($\dot{}$) denotes a partial derivative with respect to t, and the integration is subject to the boundary conditions that $\pi(x,0),g(x,0)=\pi'(x),g'(x)$ and $\pi(x,T),g(x,T)=\pi''(x),g''(x)$. Observe that the field π enters this path integral expression only in the form $\dot{\pi}$; this fact is responsible for the result of the path integral depending only on $\pi''-\pi'$. It is important to note, for any $\nu<\infty$, that underlying the formal measure given above, there is a genuine, countably additive measure on (generalized) functions g_{kl} and π^{rs}. Loosely speaking, such functions have Wiener-like behavior with respect to time and δ-correlated, generalized Poisson-like behavior with respect to space.

While (52) is invariant under spatial diffeomorphisms, it is less evident that it is also invariant under transformations of the time coordinate (by itself). (The author thanks A. Ashtekar for raising the question of temporal transformation properties.) Formally speaking, the role of the limit $\nu\to\infty$ is to remove the effects of the continuous-time regularization. It is clear, however, that there is no need that removing those effects must be done in a *uniform* way independent of x. Thus, we may replace ν by $\nu N(x)$—now under the integral sign—where $N(x)$, $0<N(x)<\infty$, is smooth and reflects the relative rate at which the regularization is removed at different spatial points. The end result is invariant under such a change. Moreover, at each point x we can run the process with different ''clock'' rates, i.e., $N(x)dt\to N(x,t)dt$ as long as the elapsed time is qualitatively unchanged. This remark means that we can choose any smooth *lapse function* $N(x,t)$, $0<N(x,t)<\infty$, with the consequence that

$$\langle \pi'',g''|\pi',g'\rangle=\lim_{\nu\to\infty}\mathcal{N}\int \exp\left[-i\int g_{ab}\dot{\pi}^{ab}\,d^3x\,dt\right]\exp\left\{-(1/2\nu)\int [b(x)^{-1}g_{ab}g_{cd}\dot{\pi}^{bc}\dot{\pi}^{da}\right.$$

$$\left.+b(x)g^{ab}g^{cd}\dot{g}_{bc}\dot{g}_{da}]N(x,t)^{-1}\,d^3x\,dt\right\}\prod_{x,t}\prod_{a\leq b}d\pi^{ab}(x,t)dg_{ab}(x,t). \tag{54}$$

The necessary conditions for this more general expression to hold are, for all T, $0<T<\infty$, and at all x, that

$$\int_0^T N(x,t)dt<\infty, \tag{55}$$

$$\int_0^\infty N(x,t)dt=\infty. \tag{56}$$

In this sense we observe that our formal path integral representation (52) for the coherent-state overlap is actually *invariant* under transformations of the time coordinate.

D. Metrical quantization

The formal, ν-dependent, weighting factor in the path integral expression (52) involves a *metric* $d\Sigma^2$ on the classical phase space, which may be read out of the expression

$$d\Sigma^2/dt^2=\int [b^{-1}g_{kl}g_{rs}\dot{\pi}^{lr}\dot{\pi}^{sk}+bg^{kl}g^{rs}\dot{g}_{lr}\dot{g}_{sk}]d^3x. \tag{57}$$

As presented, this expression for $d\Sigma^2$ is a *derived* quantity. Alternatively, it is clear that one could *start* the analysis by *postulating* a specific functional form for $d\Sigma^2$ to be used in a continuous-time regularization in the path integral construction of the reproducing kernel, and, finally, by appealing to the GNS (Gel'fand, Naimark, Segal) Theorem,[27] to recover the representation of the local field operators $\sigma_{kl}(x)$ and $\kappa^r_s(x)$. Adopting a metric on the classical phase space as the first step in a quantization procedure is called *metrical quantization*.[28] To carry out such a scheme for gravity, it is necessary that any postulated $d\Sigma^2$ satisfy several properties. First, it must be diffeomorphism *invariant* and, second, on physical grounds, it should only depend on $d\pi^{kl}$ (or $\dot{\pi}^{kl}$ for $d\Sigma^2/dt^2$) and not on π^{kl} itself. Hence, we are initially led to consider

$$d\Sigma^2/dt^2=\int [b^{-2}L_{abcd}\dot{\pi}^{bc}\dot{\pi}^{da}+M^{abcd}\dot{g}_{bc}\dot{g}_{da}]b(x)d^3x, \tag{58}$$

for suitable, positive-definite tensors L and M constructed just from g_{kl}. The given choice for L and M, i.e., $L_{abcd}=\frac{1}{2}[g_{ab}g_{cd}+g_{ac}g_{bd}]$ and $M^{abcd}=\frac{1}{2}[g^{ab}g^{cd}+g^{ac}g^{bd}]$ satisfy $M=L^{-1}$ as matrices. This choice is very natural and moreover is identical to the form suggested by the study of certain GL(3,R) coherent states for a 3×3 positive-definite matrix degree of freedom.[24]

Nevertheless, in a metric-first quantization scheme, it is appropriate to examine other choices as well. For example, a term such as $b(x)^{-1}\dot{g}_{ab}(x)\dot{\pi}^{ab}(x)$ might be included, but this term may be eliminated by a translation of the momentum. Additionally, one may consider nonlocal contributions involving, for example, the term $\dot{g}_{bc}(x)\dot{g}_{da}(y)$, together with a kernel $K(x,y)$ specifying the interrelationship of the field at y to the field at x. However, no satisfactory solution for $K(x,y)$ other than one proportional to $\delta(x,y)$ will lead to an expression for $d\Sigma^2$ that is *invariant* under all diffeomorphisms. In point of fact, the possible choices for L and M are rather limited, especially when one requires that the (formal) integration measure at each point is canonical and thus has the form $\Pi_{a\leqslant b}\, d\pi^{ab}\, dg_{ab}$. For example, for $\lambda>0$, let us consider the proposal that

$$L_{abcd}(\lambda)\equiv\tfrac{1}{2}[g_{ab}g_{cd}+g_{ac}g_{bd}+(\lambda-1)g_{bc}g_{da}]. \tag{59}$$

Then in order to lead to a canonical integration measure, it would be necessary that

$$M^{abcd}(\lambda)\equiv\tfrac{1}{2}[g^{ab}g^{cd}+g^{ac}g^{bd}+(\lambda^{-1}-1)g^{bc}g^{da}]. \tag{60}$$

Only for $\lambda=1$ is $M=L^{-1}$, which is just the choice we have made. (The form for the DeWitt metric,[29] where $\lambda=-1$, is excluded because we require that L and M be positive definite.)

The preceeding discussion has rather convincingly suggested the specifically chosen functional form for $d\Sigma^2$—apart from one issue. It may seem even more natural to choose[30]

$$d\Sigma^2/dt^2=\int [g^{-1/2}g_{ab}g_{cd}\dot{\pi}^{bc}\dot{\pi}^{da}+g^{1/2}g^{ab}g^{cd}\dot{g}_{bc}\dot{g}_{da}]d^3x, \tag{61}$$

rather than the choice we have made. This is a natural choice from a classical point of view, but it is less satisfactory from a quantum point of view. In either case, observe that a path integral such as (52) involves fields with $3+1$ independent variables; however, there are no *space derivatives* involved, only *time derivatives*. Such a model is known as an *ultralocal quantum field theory*, and by now there is much that is known about the rigorous construction and evaluation of such nontrivial (i.e., non-Gaussian) functional integrals through the study, for example, of ultralocal scalar quantum fields.[31] It is through the analysis of the gravitational models as ultralocal quantum field theories that the metric (61) is ruled out; for a simple reason, see Sec. V.

E. Operator realization

In order to realize the metric and scale fields as quantum operators in a Hilbert space, $\mathfrak{H}$, it is expedient to introduce a set of conventional local *annihilation and creation operators*, $A(x,k)$ and $A(x,k)^\dagger$, respectively, with the only nonvanishing commutator given by

$$[A(x,k),A(x',k')^\dagger]=\delta(x,x')\,\delta(k,k')\mathbf{1}, \tag{62}$$

where $\mathbf{1}$ denotes the unit operator. Here, $x\in\mathbb{R}^3$, while $k\equiv\{k_{rs}\}$ denotes a positive-definite, 3×3 matrix degree of freedom confined to the domain where $\{k_{rs}\}>0$. We introduce a ''no-particle'' state $|0\rangle$ such that $A(x,k)|0\rangle=0$ for all arguments. Additional states are determined by suitably smeared linear combinations of

$$A(x_1,k_1)^\dagger A(x_2,k_2)^\dagger\cdots A(x_p,k_p)^\dagger|0\rangle, \tag{63}$$

for all $p \geqslant 1$, and the span of all such states is $\mathfrak{H}$ provided, apart from constant multiples, that $|0\rangle$ is the only state annihilated by all the A operators. Thus, we are led to a conventional Fock representation for the A and $A^\dagger$ operators. Note that the Fock operators are irreducible, and thus all operators acting in $\mathfrak{H}$ are given as suitable functions of them.

Next, let $c(x,k)$ be a possibly complex, c-number function and introduce the translated Fock operators,

$$B(x,k) \equiv A(x,k) + c(x,k)\mathbf{1}, \tag{64}$$

$$B(x,k)^\dagger \equiv A(x.k)^\dagger + c(x,k)^* \mathbf{1}. \tag{65}$$

Evidently, the only nonvanishing commutator of the B and $B^\dagger$ operators is

$$[B(x,k), B(x',k')^\dagger] = \delta(x,x')\,\delta(k,k')\mathbf{1}, \tag{66}$$

the same as the A and $A^\dagger$ operators. With regard to transformations of the coordinate x, it is clear that $c(x,k)$ (just like the local operators A and B) should transform as a scalar density of weight one-half. Thus, we set

$$c(x,k) \equiv b(x)^{1/2} d(x,k), \tag{67}$$

where $d(x,k)$ transforms as a scalar. The criteria for acceptable $d(x,k)$ are, for each x, that

$$\int_+ |\, d(x,k)|^2\, dk = \infty, \tag{68}$$

$$\int_+ k_{rs} |d(x,k)|^2 dk = 2\,\delta_{rs}, \tag{69}$$

the latter assuming [cf. the discussion following (39)] that $\tilde{g}_{kl}(x) = \delta_{kl}$. In (68) and (69) we have introduced $dk \equiv \Pi_{a \leqslant b}\, dk_{ab}$, and the symbol ''+'' signifies an integration over only those k values for which $\{k_{ab}\} > 0$.

We shall focus on only one particular choice for d, specifically,

$$d(x,k) \equiv \frac{K e^{-\operatorname{tr}(k)}}{\det(k)}, \tag{70}$$

which is everywhere independent of x; K denotes a positive constant to be fixed later. The given choice for d corresponds to the case where the asymptotic fields $\tilde{\pi}^{kl}(x) \equiv 0$ and $\tilde{g}_{kl}(x) \equiv \delta_{kl}$. [*Remark:* For different choices of asymptotic fields it suffices to choose

$$d(x,k) \rightarrow \tilde{d}(x,k) \equiv \frac{K\, e^{-ib(x)^{-1}\tilde{\pi}^{ab}(x)k_{ab}}\, e^{-\tilde{g}^{ab}(x)k_{ab}}}{\det(k)}. \tag{71}$$

We shall not explicitly discuss this case further.]

In terms of these quantities, the local metric operator is defined by

$$\sigma_{ab}(x) \equiv b(x)^{-1} \int_+ B(x,k)^\dagger k_{ab} B(x,k)\, dk, \tag{72}$$

and the local scale operator is defined by

$$\kappa_s^r(x) \equiv -i\,\frac{1}{2}\int_+ B(x,k)^\dagger (k_{st}\vec{\partial}^{tr} - \overleftarrow{\partial}^{rt} k_{ts}) B(x,k)\, dk. \tag{73}$$

Here $\vec{\partial}^{st} \equiv \partial/\partial k_{st}$, $\overleftarrow{\partial}^{rt} \equiv \partial/\partial k_{rt}$ *acting to the left*, and $\sigma_{ab}(x)$ transforms as a tensor while $\kappa_s^r(x)$ transforms as a tensor density of weight one. It is straightforward to show that these operators satisfy the required affine commutation relations, and, moreover, that[31,32]

$$\langle 0| e^{i\int \pi^{ab}(x)\sigma_{ab}(x)d^3x} e^{-i\int \gamma_r^s(x)\kappa_s^r(x)d^3x}|0\rangle$$
$$= \exp\left\{ -K^2 \int b(x)d^3x \int [e^{-2\delta^{ab}k_{ab}} - e^{-i\pi^{ab}(x)k_{ab}/b(x)} e^{-[(\delta^{ab}+g^{ab}(x))k_{ab}]}]dk/(\det k)^2 \right\}$$
$$= \exp\left[-2\int b(x)d^3x \ln\left([\det(g_{ab}(x))]^{1/2} \det\left\{ \frac{1}{2}[\delta^{ab}+g^{ab}(x)] - i\frac{1}{2} b(x)^{-1}\pi^{ab}(x) \right\} \right) \right], \tag{74}$$

where K has been chosen so that

$$K^2 \int_+ k_{rs} e^{-2\,\mathrm{tr}(k)}\, dk/(\det k)^2 = 2\,\delta_{rs}\,. \tag{75}$$

An obvious extension of this calculation leads to (39).

F. Local operator products

Basically, local products for the gravitational field operators follow the pattern for other ultralocal quantum field theories.[31,32] As motivation, consider the product

$$\sigma_{ab}(x)\sigma_{cd}(y) = b(x)^{-2} \int_+ \int_+ B(x,k)^\dagger k_{ab}[B(x,k),B(y,k')^\dagger]k'_{cd}B(y,k')\,dk\,dk' + :\sigma_{ab}(x)\sigma_{cd}(y):$$
$$= b(x)^{-2}\delta(x,y)\int_+ B(x,k)^\dagger k_{ab}k_{cd}B(x,k)\,dk + :\sigma_{ab}(x)\sigma_{cd}(y):, \tag{76}$$

where : : denotes normal ordering with respect to A and $A^\dagger$. When $y=x$, this relation formally becomes

$$\sigma_{ab}(x)\sigma_{cd}(x) = b(x)^{-2}\delta(x,x)\int_+ B(x,k)^\dagger k_{ab}k_{cd}B(x,k)\,dk + :\sigma_{ab}(x)\sigma_{cd}(x):. \tag{77}$$

We define the renormalized (subscript ''R'') local product,

$$[\sigma_{ab}(x)\sigma_{cd}(x)]_R \equiv b(x)^{-1}\int_+ B(x,k)^\dagger\, k_{ab}\, k_{cd}\, B(x,k)\, dk, \tag{78}$$

after formally dividing both sides by the divergent dimensionless ''scalar'' $b(x)^{-1}\delta(x,x)$. [*Remark*: For scalar ultralocal theories, the formal dividing factor is the divergent dimensionless ''number'' $b^{-1}\delta(0)$, where $b>0$ is an arbitrary factor with suitable dimensions. For gravity, $b \to b(x)$, our scalar density of weight one. Note that limits involving test functions offer a rigorous definition of the renormalized product.[31,32]] Higher-order local products exist as well, for example,

$$[\sigma_{a_1b_1}(x)\sigma^{a_2b_2}(x)\sigma_{a_3b_3}(x)\cdots\sigma_{a_pb_p}(x)]_R$$
$$\equiv b(x)^{-1}\int_+ B(x,k)^\dagger\,(k_{a_1b_1}k^{a_2b_2}k_{a_3b_3}\cdots k_{a_pb_p})\,B(x,k)\,dk\,, \tag{79}$$

which, after contracting on b_1 and b_2, implies that

$$[\sigma_{a_1b}(x)\sigma^{a_2b}(x)\sigma_{a_3b_3}(x)\cdots\sigma_{a_pb_p}(x)]_R=\delta^{a_2}_{a_1}[\sigma_{a_3b_3}(x)\cdots\sigma_{a_pb_p}(x)]_R\,. \tag{80}$$

It is in this sense that $[\sigma_{ab}(x)\sigma^{bc}(x)]_R=\delta^c_a$.

We take up only one further point regarding local products. It is rather natural[31,32] to try to define the local momentum ''operator'' by

$$\mu^{rs}(y)=-i\frac{1}{2}\int_+ B(y,k)^\dagger\,(\vec{\partial}^{rs}-\overleftarrow{\partial}^{rs})\,B(y,k)\,dk, \tag{81}$$

but this expression only leads to a form and not a local operator. Furthermore, the putative canonical commutation relation becomes

$$\begin{aligned}[\sigma_{ab}(x),\,\mu^{rs}(y)]&=i\,\delta^{rs}_{ab}\,\delta(x,y)\,b(x)^{-1}\int_+ B(x,k)\,B(x,k)\,dk\\ &=i\,\delta^{rs}_{ab}\,\delta(x,y)\,[\textstyle\int_+|d(x,k)|^2\,dk+\cdots],\end{aligned} \tag{82}$$

which has a divergent multiplier and is, therefore, not even a form. On the other hand, it is true that

$$\begin{aligned}\tfrac{1}{2}[\sigma_{rl}(x)\mu^{ls}(x)+\mu^{sl}(x)\sigma_{lr}(x)]_R&=-i\tfrac{1}{2}\int_+ B(x,k)\,(k_{rl}\vec{\partial}^{ls}-\overleftarrow{\partial}^{sl}k_{lr})\,B(x,k)\,dk\\ &=\kappa^s_r(x),\end{aligned} \tag{83}$$

as claimed.

IV. IMPOSITION OF CONSTRAINTS

Gravity has four constraints at every point $x\in\mathcal{S}$, and, when expressed in suitable units, they are the familiar spatial and temporal constraints, all densities of weight one, given by[26]

$$H_a(x)=-2g_{ab}(x)\pi^{bc}{}_{|c}(x), \tag{84}$$

$$\begin{aligned}H(x)=\tfrac{1}{2}g(x)^{-1/2}&[g_{ab}(x)g_{cd}(x)+g_{ad}(x)g_{cb}(x)-2g_{ac}(x)g_{bd}(x)]\\ &\times\pi^{ac}(x)\pi^{bd}(x)+g(x)^{1/2}\;\;{}^{(3)}R(x).\end{aligned} \tag{85}$$

The spatial constraints are comparatively easy to incorporate since their generators serve as generators of the diffeomorphism group acting on functions of the canonical variables. Stated otherwise, finite spatial diffeomorphism transformations map any coherent state onto another coherent state as in (41) and (42). However, this is decidely not the case for the temporal constraint. What follows is an account of what to do about these constraints *in principle*; in Part II on this subject, we will discuss how to accomplish these goals.

One satisfactory procedure to incorporate all the necessary constraints is as follows. Let $\{h_p(x)\}_{p=1}^\infty$ denote a complete, orthonormal set of real functions on $\mathcal{S}$ relative to the weight $b(x)$. In particular, we suppose that

$$\int\;h_p(x)\,h_n(x)\,b(x)\,d^3x=\delta_{pn}\,, \tag{86}$$

$$b(x)\,\Sigma_{p=1}^\infty\,h_p(x)\,h_p(y)=\delta(x,y). \tag{87}$$

Based on this orthonormal set of functions, we next introduce four infinite sequences of constraints,

$$H_{(p)\,a} \equiv \int h_p(x)\, H_a(x)\, d^3x, \tag{88}$$

$$H_{(p)} \equiv \int h_p(x)\, H(x)\, d^3x, \tag{89}$$

$1 \leq p < \infty$, all of which vanish in the classical theory.

For the quantum theory let us assume, for each p, that $\mathcal{H}_{(p)\,a}$ and $\mathcal{H}_{(p)}$ are selfadjoint, and even stronger that

$$X_P^2 \equiv \Sigma_{p=1}^{P}\, 2^{-p}\, [\Sigma_{a=1}^{3} (\mathcal{H}_{(p)\,a})^2 + (\mathcal{H}_{(p)})^2], \tag{90}$$

is selfadjoint for all $P < \infty$. Note well, as one potential example, the factor 2^{-p} introduced as part of a regulator as $P \rightarrow \infty$; we comment on this regulator in the next section. For each $\delta \equiv \delta(\hbar) > 0$, let

$$\mathbb{E}_P \equiv \mathbb{E}(X_P^2 \leq \delta^2) \tag{91}$$

denote a projection operator depending on X_P and δ, as indicated. How such projection operators may be constructed is discussed in Ref. 7 and will be dealt with in Part II. Let

$$S_P \equiv \limsup_{\pi,\, g}\, \langle \pi, g | \mathbb{E}_P | \pi, g \rangle, \tag{92}$$

which satisfies $S_P > 0$ since $\mathbb{E}_P \not\equiv 0$ when restricted to sufficiently large δ. Finally, we define

$$\langle\langle \pi'', g'' | \pi', g' \rangle\rangle \equiv \limsup_{P \rightarrow \infty} S_P^{-1}\, \langle \pi'', g'' | \mathbb{E}_P | \pi', g' \rangle, \tag{93}$$

as a reduction of the original reproducing kernel. The result is either trivial, say if δ is too small, or it leads to a continuous, positive-definite functional on the original phase space variables. We focus on the latter case.

To obtain the final physical Hilbert space, one must study $\langle\langle \pi'', g'' | \pi', g' \rangle\rangle$ as a function of the regularization parameter δ. Since gravity has an anomaly,[33] there should be a minimum value of δ, which is still positive, that defines the proper theory, rather like the example in (30). Assuming we can find and then use that value, $\langle\langle \pi'', g'' | \pi', g' \rangle\rangle$ becomes the reproducing kernel for the physical Hilbert space $\mathfrak{H}_{\rm phys}$. Attaining this goal would then permit the real work of extracting the physics to begin.

Our discussion regarding constraints in this paper has indeed been brief. Although the program we have in mind is not simple, it has the virtue of being realizable, at least in principle. After all, before any calculational scheme is developed it is always wise to ensure that the object under study has a good chance of existing!

V. DISCUSSION

In the preceding sections we have outlined an approach to quantum gravity that it is somewhat different than currently considered. As a background to our philosophy, let us briefly review some of the common weak points in the standard ways of quantizing gravity, and use these comments as motivation for our approach.

A. Traditional viewpoints and commentary

(i) Viewed by way of conventional perturbation theory, quantum gravity has two main difficulties of principle. On the one hand, the perturbative split of the metric in the form $g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\mu\nu}(x)$ (or any other background metric), with canonical quantization of the ''small'' deviation $h_{\mu\nu}(x)$ violates signature properties, since in that case the spectrum of $h_{\mu\nu}(x)$ is unbounded above and below. On the other hand, as an asymptotically nonfree theory, gravity is nonrenormalizable and poorly described by a perturbation theory that needs an unending addition of distinct counterterms with divergent coefficients.

To address these obstacles, we first note that the affine approach guarantees a proper metric signature from the very beginning, and, second, we remind the reader that certain asymptotically nonfree, nonrenormalizable models have indeed been solved,[31] and their solution procedures form the core of the present approach to quantize gravity.

(ii) While the constraints of classical gravity are first class, there is an anomaly in the quantum constraints and thus they are effectively second class. Usual views toward second-class constraints involve solving and eliminating them, introducing and then quantizing Dirac brackets, or the conversion of second-class constraints into first-class constraints. Each of these methods is often complicated and not all are guaranteed to be valid beyond a semiclassical treatment if the classical constraint hypersurface has a non-Euclidean geometry.[34] These difficulties have stimulated searches to get around the second-class character altogether, either by introducing non-Hermitian constraint operators that may close algebraically,[35] or by introducing additional fields and space–time dimensions until the anomaly cancels.

Regarding these comments, we accept the anomaly and the second-class constraints that it implies. Giving up a classical symmetry is not so heretical as it may seem. For example, Hamiltonian classical mechanics enjoys a full covariance under general canonical coordinate transformations, but that invariance is *not* preserved *in its classical form* when we go to the quantum theory. For example, consider the classical Poisson brackets for a set of generator elements,

$$\{e^{ap+bq}, e^{cp+dq}\} = (bc-ad)\, e^{(a+c)p+(b+d)q}, \tag{94}$$

where a,b,c, and d are parameters, while in canonical quantum mechanics we have the corresponding commutator algebra,

$$[e^{aP+bQ}, e^{cP+dQ}] = (2i)\sin[(bc-ad)/2]\, e^{(a+c)P+(b+d)Q}. \tag{95}$$

These expressions agree in their algebraic structure for selected elements, but not for the whole algebra. Equivalence begins to break down at the quadratic level, which is exactly the case for the temporal constraint in gravity. In particle mechanics there are sound physical reasons[36] for this ‘‘breakdown’’ of symmetry, and attempts to restore the symmetry—as in geometric quantization[37]—go counter to such sound physical principles. There is no reason that a similar scenario does not hold for gravity. The breakdown of the classical symmetry and the appearance of a quantum ‘‘anomaly’’ (better called a ‘‘quantum mechanical symmetry breaking’’[38]) could, just as in the quantum mechanics case, carry real physics.

Accepting the second-class nature of (part of) the constraints of quantum gravity means a different approach must be taken. As already noted, earlier approaches required solving for and eliminating the unphysical variables, the introduction of Dirac brackets, etc., all of which are rather technical and may be extremely complicated. In the present view, afforded by the projection operator approach to constraints—*second-class constraints included*—none of these particular complications arise. Instead, one projects onto the state (or, with degeneracy, states) for which the sum of the square of the constraints is bounded. Why the square and not the fourth power? Using the fourth power would not be wrong; the only change would involve a unitary transformation of the original result, which maps one set of ‘‘ground’’ states onto another set of ‘‘ground’’ states. The square is chosen for simplicity, not for any reasons of exclusivity.

B. Relation to earlier work

Pilati, in a series of papers[3] (see also Ref. 2), analyzed a strong coupling model of quantum gravity in which the temporal constraint given in (85) was modified to read as

$$H'(x) = \tfrac{1}{2} g(x)^{-1/2} [g_{ab}(x) g_{cd}(x) + g_{ad}(x) g_{cb}(x) - 2 g_{ac}(x) g_{bd}(x)]\, \pi^{ac}(x) \pi^{bd}(x), \tag{96}$$

namely, the second term involving the scalar curvature ${}^{(3)}R(x)$ based on the metric $g_{ab}(x)$ was dropped. The reason for doing so was to achieve a theory in which the temporal constraint $H'(x)$ itself was patterned after the Hamiltonian density of an ultralocal theory. This modification was thought to be advantageous because then all the machinery developed for ultralocal quantum field theory could be used for the strong coupling gravitational model. Once that model was under control, it was the hope to reintroduce the dropped term by a perturbation theory analysis. Unfortunately, the reintroduction of dropped terms involving spatial gradients has never been successfully accomplished by a perturbation analysis about a non-Gaussian ultralocal model. This failure is most likely because such ''interaction terms'' generally amount to nonrenormalizable perturbations of the unperturbed (ultralocal) models.

The program advanced in the present paper takes a different view toward these issues.

First, we focus on kinematics with the knowledge that for pure gravity the Hamiltonian operator vanishes, as it does in any situation that is reparametrization invariant. In its place we find constraints, and the real physical content of the theory lies in the particular constraints. However, before the constraints can be introduced, there must be a ''primary container'' to receive them. In our case, this primary container is the Hilbert space and set of relevant operators prior to the introduction of *any* form of the constraints, and which is based on the fundamental physical nature of the variables, i.e., positive-definite, 3×3 matrix-valued, local field operators, etc. This is the preferred procedure: *Quantization before the introduction of any constraints*. At this primary level, there is no coupling of one degree of freedom with another—any coupling comes through the enforcement of specific constraints. Hence, in the primary container the degrees of freedom are mutually independent of each other. For finitely many kinematical degrees of freedom this means that the Hilbert space is a product over spaces for each of the separate degrees of freedom; in a field theory, this independence means that the kinematical operators enter as ultralocal field operators. Consequently, even though the several constraint operators waiting to be introduced may themselves *not* be ultralocal in nature, the primary container itself, which has been prepared to receive them, is ultralocal.

At this point the reader may wish to reexamine (39)—in essence our ''primary container''—to recall the appearance of an ultralocal state on a set of field operators. Apart from the 3×3 matrix character, the functional form of (39) emerges from (i) the product of N expressions of the form (16) for independent arguments p_n, q_n, r_n, s_n, $1\leq n\leq N$; (ii) the replacement of β by $b_n\Delta$ and (p_n-r_n) by $(p_n-r_n)\Delta$; and (iii) the limit as $N\rightarrow\infty$, $\Delta\rightarrow0$ such that $\Sigma(\cdot)\, b_n\,\Delta\rightarrow\int(\cdot)\, b(x)\, dx$. In this way we have created, from a collection of independent single affine degrees of freedom, the reproducing kernel for affine gravity in $1+1$-dimensional space. In a similar manner, a set of independent 3×3 affine degrees of freedom can be (and were) used to build an ultralocal representation for 3×3 metric and momentum fields in (39); moreover, this type of construction does not favor the ''natural'' phase-space metric (61). In summary, we emphasize that whenever the ''dynamics'' appears through constraints, the primary container should be ultralocal in character. We next turn our attention to the introduction of the constraints.

In the projection operator approach it is recognized from the outset that the physical Hilbert space—or better the *regularized* physical Hilbert space—is a *subspace* of the original Hilbert space that is uniquely determined by an associated projection operator $\mathbb{E}$. Whatever form the constraints may take, they are ''encoded'' into the projection operator $\mathbb{E}$, and a regularization means that the constraints are satisfied to a certain level of precision determined by a regularization parameter δ. How to turn constraint operators into projection operators in general has been discussed in Ref. 7; as regards the gravitational case, that project will be discussed in Part II .

Continuing still in a general framework, let us consider an expression that may be used either to generate dynamics or to enforce constraints. From a classical point of view, and especially from a path integral point of view, it may seem that quantities used in either of these ways may be rather similar. However, it is important to already understand that there is a fundamental distinction between the use of a quantum operator either (i) to generate unitary transformations or (ii) to serve as a constraint operator in a given system. In the first case, the operator must be self-adjoint and thus densely defined, while in the second case the operator may be defined on only the zero vector!

This fact has profound consequences. In particular, to have a self-adjoint generator requires that the operator representation in the primary container must already be finely ''tuned'' to ensure that the generator that will be introduced is self-adjoint (as in Haag's Theorem[39]). For constraint operator imposition this need not be the case, and the reason this is so is because we allow for changes, i.e., adaptions, of the primary container representations through the process of reduction of the reproducing kernel. As an example, let us consider only the local temporal operator $\mathcal{H}(x)$ for gravity. On the one hand, to generate unitary time evolutions it may be necessary that $\int\mathcal{H}(x)\,d^3x$ be self-adjoint. On the other hand, to enforce constraints, it is only necessary that $\mathcal{H}_{(p)}$ [cf. (89)] be ''small,'' but there is no requirement that these operators must be *uniformly* ''small.'' Instead they can be ''small'' in the sense that $X^2 \equiv \Sigma_{p=1}^{\infty} s_p(\mathcal{H}_{(p)})^2$ is ''small,'' where the set of positive constants $\{s_p\}$ serve as regulators to control convergence of the series. The example chosen was that $s_p = 2^{-p}$, but there is nothing special about that choice. Any reasonable choice that leads to a self-adjoint operator X should lead to the same reproducing kernel in the final analysis when the regularization parameter δ attains its final value for the problem at hand.

A rather simple example of the general procedure discussed above can be seen in studies involving product representations.[40] Additionally, it is instructive to reanalyze the relativistic free field by this procedure to see how ultralocal representations turn into nonultralocal representations.

Suffice it to say, it is this vast difference between the required nature of constraint operators and unitary generators that permits us to start with ultralocal field operator representations and emerge with *non*ultralocal operators in the physical Hilbert space.

ACKNOWLEDGMENTS

The author thanks B. Bodmann, J. Govaerts, A. Kempf, S. V. Shabanov, and G. Watson for useful comments.

[1] J. R. Klauder, in *Relativity*, edited by M. S. Carmeli, S. I. Fickler, and L. Witten (Plenum, New York, 1970), p. 1.
[2] C. J. Isham and A. C. Kakas, Class. Quantum Grav. **1**, 621,633 (1984).
[3] M. Pilati, Phys. Rev. D **26**, 2645 (1982); **28**, 729 (1983); G. Francisco and M. Pilati, *ibid.* **31**, 241 (1985).
[4] J. R. Klauder, Ann. Phys. (N.Y.) **254**, 419 (1997).
[5] S. V. Shabanov, in *Path Integrals: Dubna '96* (JINR, Dubna, Russia, 1996), p. 133.
[6] J. Govaerts, J. Phys. A **30**, 603 (1997).
[7] J. R. Klauder, Nucl. Phys. B **547**, 397 (1999).
[8] J. R. Klauder and E. W. Aslaksen, Phys. Rev. D **2**, 272 (1970).
[9] I. M. Gel'fand and M. A. Naimark, Dokl. Akad. Nauk SSSR **55**, 570 (1947).
[10] E. W. Aslaksen and J. R. Klauder, J. Math. Phys. **9**, 206 (1968).
[11] E. W. Aslaksen and J. R. Klauder, J. Math. Phys. **10**, 2267 (1969).
[12] N. Aronszajn, Proc. Cambridge Philos. Soc. **39**, 133 (1943); Trans. Am. Math. Soc. **68**, 337 (1950); H. Meschkowski, *Hilbertische Raum Mit Kern-Funktion* (Springer-Verlag, Berlin, 1962).
[13] J. R. Klauder and B.-S. Skagerstam, *Coherent States* (World Scientific, Singapore, 1985), p. 60.
[14] J. R. Klauder and B.-S. Skagerstam, in Ref. 13, p. 6.
[15] J. R. Klauder, in *Functional Integration*, edited by J.-P. Antoine and E. Tirapegui (Plenum, New York, 1980), p. 101.
[16] J. Sniatycki, *Geometric Quantization and Quantum Mechanics* (Springer-Verlag, New York, 1980).
[17] I. Daubechies, J. R. Klauder, and T. Paul, J. Math. Phys. **28**, 85 (1987).
[18] G. Roepstorff, *Path Integral Approach to Quantum Physics* (Springer-Verlag, Berlin, 1996), pp. 195–199.
[19] I. Daubechies and J. R. Klauder, J. Math. Phys. **26**, 2239 (1985).
[20] J. R. Klauder, J. Math. Phys. **8**, 2392 (1967).
[21] J. Govaerts, *Hamiltonian Quantisation and Constrained Dynamics* (Leuven University Press, Leuven, Belgium, 1991).
[22] J. R. Klauder and B. F. Whiting, J. Phys. A **26**, 1697 (1993).
[23] J. R. Klauder, ''New measures for the quantization of systems with constraints,'' *Proceedings of the 1998 Bialowieza Workshop* (in press), revised version of quant-ph/9607020.
[24] G. Watson and J. R. Klauder (in preparation).
[25] R. Arnowitt, S. Deser, and C. Misner, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), p. 227.
[26] C. Misner, K. Thorne, and J .A. Wheeler, *Gravitation* (Freeman, San Francisco, 1971).
[27] G. G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley-Interscience, New York, 1972).
[28] J. R. Klauder, in *Quantum Future*, edited by P. Blanchard and A. Jadczyk (Springer-Verlag, Berlin, 1999), p. 129.
[29] B. S. DeWitt, Phys. Rev. D **5**, 2458 (1967).
[30] J. R. Klauder, in *Probabilistic Methods in Quantum Field Theory and Quantum Gravity*, edited by P. H. Damgaard, H. Hüffel, and A. Rosenblum (North-Holland, New York, 1990), p. 73.

[31] J. R. Klauder, *Beyond Conventional Quantization* (Cambridge University Press, Cambridge, UK, 1999).
[32] J. R. Klauder, Commun. Math. Phys. **18**, 307 (1970); Acta Physica Austriaca, Suppl. VIII, p. 227, 1971; in *Mathematical Methods in Theoretical Physics*, edited by W. E. Brittin (Colorado Assoc. Univ. Press, 1973), p. 329.
[33] J. Anderson, in *Proceedings of the 1st Eastern Theoretical Physics Conference, University of Virginia, 1962*, edited by M. E. Rose (Gordon and Berach, New York, 1963), p. 387.
[34] S. V. Shabanov and J. R. Klauder, Nucl. Phys. B **511**, 713 (1998); see also S. V. Shabanov, ''Geometry of the physical phase space in quantum gauge theories,'' Phys. Rep. (in press).
[35] A. Komar, Phys. Rev. D **20**, 380 (1979).
[36] J. R. Klauder, Found. Phys. **27**, 1467 (1997); in *Quantization, Coherent States, and Poisson Structures*, edited by A. Strasburger *et al.* (Polish Scientific Publishers PWN, Warsaw, 1998), p. 59.
[37] A. Echeverria-Enriquez, M. C. Munoz-Lecanda, N. Roman-Roy, C. Victoria-Monge, Extr. Math. **13**, 135 (1998).
[38] R. Jackiw, in *Conceptual Foundations of Quantum Field Theory*, edited by T. Y. Cao (Cambridge University Press, Cambridge, UK, 1999).
[39] R. Haag, *Local Quantum Physics* (Springer-Verlag, Berlin, 1992).
[40] J. R. Klauder, ''Product representations and the quantization of constrained systems,'' Proceedings of the Steklov Mathematical Institute (in press), quant-ph/9811051.

K slicing the Schwarzschild and the Reissner–Nordstrom spacetimes

Asghar Qadir[a)] and Azad A. Siddiqui[b)]
Department of Mathematics, Quaid-i-Azam University, Islamabad, Pakistan

(Received 17 August 1998; accepted for publication 20 May 1999)

A K slicing of the Reissner–Nordstrom (RN) spacetime is shown to provide a complete foliation of the region up to the inner horizon, and the horizon to correspond to the limit as $K\rightarrow\pm\infty$. The implications of this foliation in the context of York time are discussed.
[S0022-2488(99)00909-3]

I. INTRODUCTION

In an earlier paper[1] we had followed up the proposal of Brill *et al.*[2] (hereafter referred to as BCI) for foliating the Schwarzschild spacetime by a sequence of spacelike hypersurfaces of constant mean extrinsic curvature, K. They used a variational principle for minimizing the (3−) area of the hypersurface for a given (4−) volume of the world tube traced out by it. The corresponding Lagrange multiplier turns out to be K. Using Kruskal–Szekres (KS) coordinates (v,u),[3] BCI obtained the K-slicing equation,

$$\frac{dv}{du}=\frac{Av+Eu}{Au+Ev}, \tag{1}$$

$$E=H-Kr^3/3, \quad A=[E^2+r^3(r-r_s)]^{1/2}, \tag{2}$$

where r_s is the Schwarzschild radius ($2m$ in gravitational units) and H is an arbitrary parameter that measures how much the intrinsic and extrinsic curvatures vary on the K surface. We converted to the compactified KS coordinates (ψ,ξ), given by[4]

$$\begin{aligned}\psi&=\tan^{-1}(v+u)+\tan^{-1}(v-u),\\ \xi&=\tan^{-1}(v+u)-\tan^{-1}(v-u),\end{aligned} \tag{3}$$

so as to be able to see the foliation in a Penrose diagram. In these coordinates, Eq. (1) becomes

$$\frac{d\psi}{d\xi}=\frac{A\sin\psi\cos\xi+E\sin\xi\cos\psi}{A\sin\xi\cos\psi+E\sin\psi\cos\xi}. \tag{4}$$

Our procedure was to take a trial value of the initial r for a given K, r_i, and put

$$H=Kr_i^3/3\pm\sqrt{r_i^3(r_s-r_i)}. \tag{5}$$

We found that the hypersurfaces would rise, very slowly at first. Taking r_i such that it rose least allowed r_f (the final value of r) to come at $I°$ ($\psi=0$, $\xi=\pi$). Thus, our ansatz required that the hypersurfaces have $d\psi/d\xi|_{\xi=0}=0$ and that they end at $I°$. The procedure worked but we had

[a)]Senior Associate Member of the International Center for Theoretical Physics, Trieste, Italy. Electronic mail: qadirs@isb.pol.com.pk
[b)]Present address: College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Rawalpindi, Pakistan.

provided no proof that we had a complete foliation. In this paper we provide a proof that a complete foliation was achieved by the ansatz for the Schwarzschild spacetime and extend it to the Reissner–Nördstrom geometry. It is shown that a complete K foliation of the spacetime up to the inner horizon is achieved, but it does not proceed beyond it.

II. PROOF OF COMPLETENESS OF THE SCHWARZSCHILD FOLIATION

To prove completeness we need to verify that no region of the spacetime is left out from the foliation and that the foliating hypersurfaces do not intersect anywhere. We first discuss the hypersurfaces for limiting values of K. $K=0$ corresponds to the hypersurface, $\psi=0$. For this hypersurface the value of r at $\xi=0$, r_i, is r_s. We now consider the limit as $|K|\to\infty$. It had been seen that as we increased $|K|$ we had to choose smaller values for r_i. Let us suppose that for some very large $|K|$, $r_i=\epsilon r_s>0$ is such that $\epsilon\ll 1$. Then, from Eq. (5),

$$H=Kr_s^3\epsilon^3/3\pm r_s^2\epsilon^{3/2}(1-\epsilon)^{1/2}. \tag{6}$$

Now E and A cannot remain small for sufficiently large $|K|$ unless $r=r_i$ throughout. Thus, considering Eqs. (6) and (4), we require that as $|K|\to\infty$, either: (i) $\epsilon\to 0$ and hence $r-\epsilon r_s\to 0$; or (ii) $\epsilon\nrightarrow 0$ and hence r, A, $|E|\to\infty$. In fact, for case (i) $\epsilon\sim|K|^{-1/3}$ so that H, E, and A remain constant. Hence, the limit $|K|\to\infty$ corresponds to the limiting hypersurface $r=0$, which is $v=\infty (u\neq\infty)$ or $\psi=\pi/2$, $|\xi|<\pi/2$. In case (ii) $A\sim\pm E$. In that case Eq. (4) gives $d\psi/d\xi=\pm 1$. Thus, the rest of the hypersurface is at $\psi=\pm\xi+\pi$ and $\psi=\pm\xi-\pi$ with r at infinity. Hence the complete hypersurfaces are $\mathscr{I}_L^-\cup\{r=0\}\cup\mathscr{I}_R^-$ for $K=-\infty$ and $\mathscr{I}_L^+\cup\{r=0\}\cup\mathscr{I}_R^+$ for $K=+\infty$, where the subscript L refers to the left and R to the right side of the Penrose diagram. Notice that these hypersurfaces are not spacelike in themselves ($\mathscr{I}^\pm$ being null). This does not reflect a problem with the foliation as that consists of spacelike hypersurfaces that tend to these null hypersurfaces as $|K|\to\infty$. Also notice that the family of hypersurfaces is symmetric under ψ reflection, with $K\leftrightarrow -K$.

It remains to verify that the hypersurfaces before the limit do cover the whole space and do not intersect. The key point is the existence and uniqueness of the solution of differential equation (4). It can be written in (t,r) coordinates as

$$\frac{dt}{dr}=f(r)=\frac{rE}{(r-r_s)A}, \tag{7}$$

which is a function of r only. It is continuous and $\lim_{r\to\infty}f(r)=1$ while $\lim_{r\to 0}f(r)=0$. The boundedness of the function at $r=r_s$, which is the Schwarzschild coordinate singularity, can be seen in (ψ,ξ) coordinates from Eq. (4), $\lim_{r\to r_s}f(r)=\lim_{\psi\to\pm\xi}f(\psi,\xi)=\pm 1$. Further, writing $f(r)$ as $f(t,r)$, it clearly satisfies the Lipshitz condition:

$$|f(t_1,r)-f(t_2,r)|=|f(r)-f(r)|=0\leqslant N|t_1-t_2|, \tag{8}$$

where N is a $+ve$ constant. Therefore $\exists!$ solution of Eq. (4) for the given conditions. There are two free parameters H and K in Eq. (4). These two parameters are fixed by our requirements $d\psi/d\xi|_{\xi=0}=0$ and $\psi(\pi)=0$. Let us consider two different hypersurfaces: $\psi_1(\xi)$ and $\psi_2(\xi)$, for two different values of the curvature $K_1\neq K_2$ satisfying the requirements. Suppose that these hypersurfaces intersect at some point $\xi=\xi_0$. Considering $\xi=\xi_0$ as our starting point, $\psi_1(\xi_0)=\psi_2(\xi_0)$ fixes one of the parameters, say H. Now, by the uniqueness of the solution of Eq. (4) two different values of K, K_1 and K_2, cannot give the same end point $\psi_1(\pi)=0=\psi_2(\pi)$. Hence, our assumption that the two hypersurfaces intersect must be wrong. This proves that the hypersurfaces do not intersect.

To show that the hypersurfaces cover the whole space, we write Eq. (4) as the integral equation,

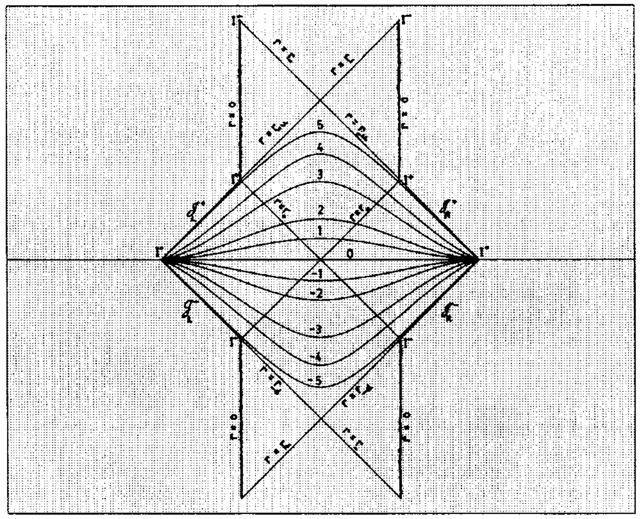

FIG. 1. Foliation of the compactified RN spacetime, with $Q/m=0.8$, by spacelike K slices. The hypersurfaces are labeled according to their serial numbers in Table I. Notice that only the region $r_-<r<\infty$ is covered by them. In the limit of $K\rightarrow+\infty$ we get the upper boundary of the central block, $\mathscr{T}_L^+\cup\{r=r_{-u}\}\cup\mathscr{T}_R^+$ and in the limit of $K\rightarrow-\infty$ we get the lower boundary, $\mathscr{T}_L^-\cup\{r=r_{-d}\}\cup\mathscr{T}_R^-$.

$$\psi(\xi)=\int_{z=0}^{\xi} f(\psi(z),z,K,H)dz+C. \tag{9}$$

Using the requirement $d\psi/d\xi|_{\xi=0}=0$, in principle, we can eliminate H from Eq. (9).

Now, requiring that (ψ_0,ξ_0) lies on the hypersurface, and that $\psi(\pi)=0$, gives

$$\psi(\xi)=-\int_{\xi}^{\pi} f(\psi(z),z,K)dz. \tag{10}$$

This proves the existence of hypersurfaces passing through every pair (ψ_0,ξ_0). Hence, the hypersurfaces cover the whole space. Note that for a given (ψ_0,ξ_0) there will be a specific K that satisfies Eq. (10).

III. FOLIATION OF THE REISSNER–NÖRDSTROM SPACETIME

We now turn our attention to the RN geometry. Following the BCI procedure again leads to Eq. (4) with v and u in Eq. (3) being the generalized KS (GKS) coordinates,[5]

TABLE I. Eleven K slices for different values of the mean extrinsic curvature, K, for an RN black hole with $Q/m=0.8$, are described by the corresponding values for the initial value of r, r_i, the constant H, and the initial value of ψ in the Penrose diagram, ψ_i. The maximum (minimum for $K<0$) value, ψ^*, in these cases is identical to ψ_i.

No.	K	$r_i/2m$	H	ψ_i
0	0.0	0.8	0.0	0.0
±1	0.01	0.799	∓0.0179	±0.418
±2	0.02	0.796	∓0.0335	±0.801
±3	0.05	0.779	∓0.0780	±1.534
±4	0.1	0.732	∓0.1062	±2.078
±5	0.8	0.538	∓0.1186	±2.516

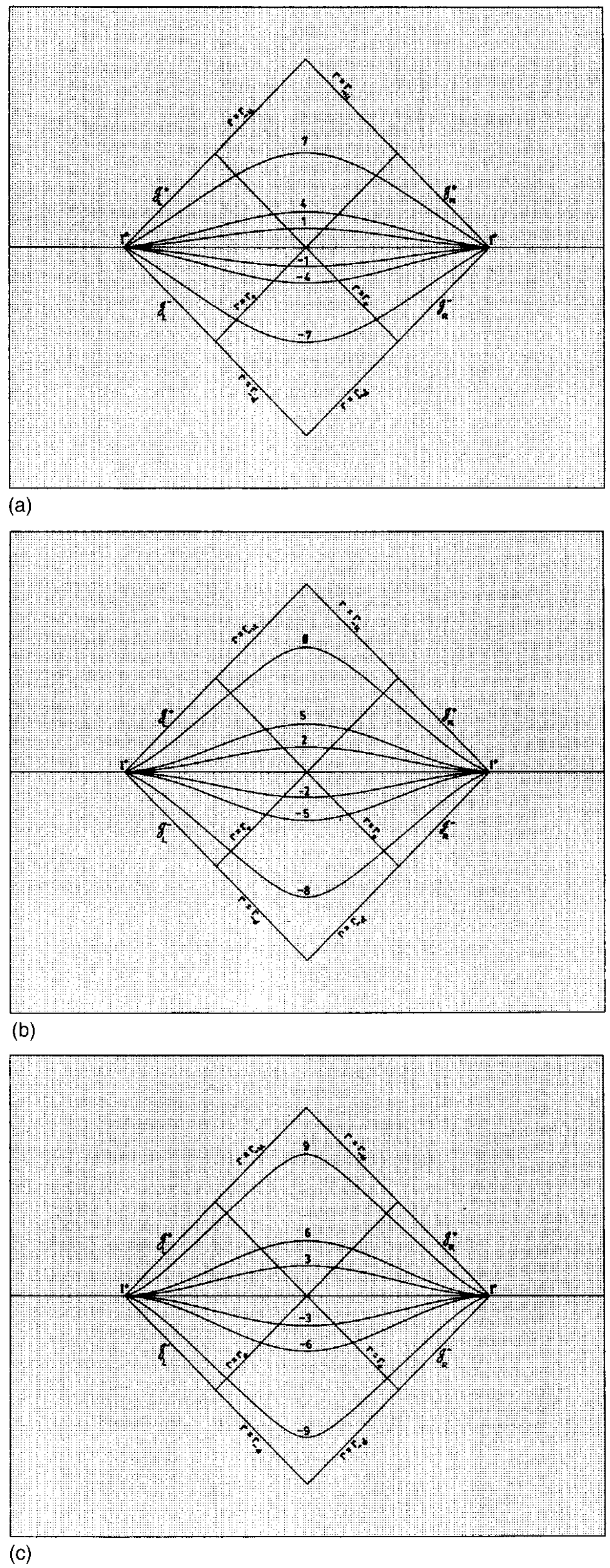

FIG. 2. Hypersurfaces for $|K|=0.01$, 0.02, and 0.1 in an RN black hole, with (a) $Q/m=0.6$, (b) $Q/m=0.8$, and (c): $Q/m=0.95$ are shown. The hypersurfaces are labeled according to their serial numbers in Table II.

TABLE II. Eighteen K slices for different values of ''K'' and Q/m are described by the corresponding values of ψ_i. The maximum (minimum for $K<0$) value, ψ^*, in these cases is identical to ψ_i.

No.	K	Q/m	ψ_i
± 1		3.0/5	± 0.317
± 2	± 0.01	4/5	± 0.418
± 3		4.75/5	± 0.537
± 4		3.0/5	± 0.592
± 5	± 0.02	4/5	± 0.801
± 6		4.75/5	± 0.918
± 7		3.0/5	± 1.576
± 8	± 0.1	4/5	± 2.078
± 9		4.75/5	± 2.353

$$
\begin{aligned}
v &= \alpha \exp(r/\beta_+)\left|\frac{r}{r_+}-1\right|^{r_+^2/\beta_+(r_+-r_-)}\left|\frac{r}{r_-}-1\right|^{-r_-^2/\beta_+(r_+-r_-)}\cosh(t/\beta_+),\\
u &= \alpha \exp(r/\beta_+)\left|\frac{r}{r_+}-1\right|^{r_+^2/\beta_+(r_+-r_-)}\left|\frac{r}{r_-}-1\right|^{-r_-^2/\beta_+(r_+-r_-)}\sinh(t/\beta_+),
\end{aligned}
\tag{11}
$$

for $r_+<r<\infty$ and v, u interchanged for $r_-<r<r_+$, where

$$r_\pm = m \pm \sqrt{m^2-Q^2}, \tag{12}$$

are the outer and inner horizons (Q being the RN black hole charge) and α, β_+ are the constants

$$\alpha = \frac{16 r_+ r_-}{r_+ - r_-}, \quad \beta_+ = \frac{2r_+^2}{r_+ - r_i}. \tag{13}$$

The entire spacetime cannot be covered simply by these coordinates. For $0<r<r_-$ they are not applicable. There we must replace β_+ by β_- with the obvious change in Eq. (13). The two patches together are adequate for the entire spacetime, the latter being valid for $0<r<r_+$. In the region $r_-<r<r_+$ we can patch the coordinate systems together by providing a coordinate transformation.

We now follow the rest of the procedure as before. Now A of Eq. (2) is modified to

$$A = [E^2 + r^2(r-r_+)(r-r_-)]^{1/2}, \tag{14}$$

while E is unaltered. For the ansatz, Eq. (5) is altered to

$$H = K r_i^3/3 \pm r_i \sqrt{(r_+-r_i)(r_i-r_-)}, \tag{15}$$

in the region $r_-<r<r_+$. Note that $r_i \leqslant r_+$ always, as is obvious from the Penrose diagram (since we take $\xi_i=0$). We again require that the hypersurfaces have $d\psi/d\xi|_{\xi=0}=0$ and ask that the hypersurfaces extend to $I°$.

The foliation proceeds smoothly for medium values of $|K|$ (about 0.01–0.8) but runs into problems of numerical instability at very high or very low values of $|K|$. By using more computer time one can extend the range of stability. As such, there is no inherent problem with the numerics but merely a matter of economizing the use of the computer. Typical hypersurfaces are given for six values of $|K|$ for an RN black hole with $Q/m=0.8$, in the Penrose diagram given in Fig. 1.

Table I gives the values of K with r_i and the corresponding ψ_i. We note that it is only the region $r_-<r<\infty$ that is covered by this foliation. This is not surprising as the requirements, that the hypersurfaces be spacelike and reach $I°$, could not be met by hypersurfaces that enter the region $r<r_-$. We will consider the physical implications of this observation shortly. At present we need to verify that our ansatz provides a complete foliation of the central block of the Penrose diagram, shown in Fig. 1.

IV. PROOF OF COMPLETENESS OF THE REISSNER–NÖRDSTROM FOLIATION

To see this we note that Eqs. (2) and (5) as modified by Eqs. (14) and (15) allow two possible solutions as $|K|\rightarrow\infty$: (i) either $r\rightarrow r_i$ throughout, A, $E\rightarrow 0$ and $d\psi/d\xi\rightarrow\pm 1$ or $|\psi|=-\pi\pm\xi$; or (ii) r, A, $|E|\rightarrow\infty$ and again $d\psi/d\xi\rightarrow\pm 1$ or $|\psi|=\pi\pm\xi$. Thus, the hypersurfaces $K=\pm\infty$ correspond to $\mathscr{T}_L^+\cup\{r=r_{-u}\}\cup\mathscr{T}_R^+$ and $\mathscr{T}_L^-\cup\{r=r_{-d}\}\cup\mathscr{T}_R^-$ in Fig. 1. Thus, our ansatz works and gives the boundary of the central block as the limit as $K\rightarrow\pm\infty$. Again the differential equation (4) with E unaltered and A now given by Eq. (14), satisfies all the conditions for the existence of the unique solution. The same argument that is used for the Schwarzschild case can be applied here to prove that the hypersurfaces do not intersect and cover the whole space *up to the inner horizon*.

V. DISCUSSION OF THE RESULTS

We now turn to the discussion of the implications of our foliation. It had been argued[6] that there is a special significance of the York time in a cosmological context. Namely, in terms of such a time parameter the formation of a black hole singularity is simultaneous with the ‘‘big crunch’’ singularity of a closed universe containing the black hole. In the sense of a compactified (‘Schwarzschild universe’) the same point was noted in our foliation of the Schwarzschild spacetime by K surfaces.[1] Of course, our foliation procedure is not the only one for obtaining K slices. Since the singularity, $r=0$, is timelike, here it does not appear to be possible to obtain spacelike K slices in the region $0<r<r_-$. In this foliation it appears that we should regard the inner horizon as simultaneous with the end of the ‘‘compactified RN universe.’’ This is, of course, consistent with the fact that $r=r_-$ is a blue-shift horizon and the barrier that forms at this place due to backreaction that was noticed by Matzner *et al.*[7] and the infinitely blue-shifted radiation from the outside universe pointed out by Penrose and Simpson.[8] We conjecture that there may be a physical barrier at $r=r_-$. This would remove the enormous difference between the topology of the Schwarzschild and RN–Penrose diagrams. This argument would be stronger if the above foliation could be proved to be unique, but we have not been able to prove it so far.

It is now necessary to investigate how a variation of Q/m changes the behavior of the hypersurfaces. In the case $Q/m=0$ we have the Schwarzschild spacetime[1] and some hypersurfaces first rise then fall to I° (for $K>0$). For the RN spacetime with $Q/m=0.8$, we found that the hypersurfaces started flat (i.e., $d\psi/d\xi=0$) at the highest point and then dropped to I° (again for $K>0$). For comparison we considered hypersurfaces for $|K|=0.01$, 0.02 and 0.1 with $Q/m=0.6$, 0.8, and 0.95 in Fig. 2. It is seen that for a given K the ψ_i is raised or lowered by raising or lowering Q/m.

Finally, it is worthwhile to comment on wider applications of the BCI procedure to obtain K-slice foliations. It is natural to limit the attempt to static spacetimes, where there can be greater expectations of achieving the foliation. However, one could try to look for stationary but nonstatic solutions of the Einstein equations, such as the Kerr metric. We find that the procedure is implementable to the extent of providing the K-slicing equation. Work is in progress to check if our ansatz can be appropriately modified for the purpose. Other such investigations would be valuable in view of the importance of K slicing.[6]

ACKNOWLEDGMENTS

One of us (AQ) is grateful to the King Fahad University of Petroleum and Minerals, Saudi Arabia for the provision of research facilities. Both of us are grateful to TWAS for the provision of a computer at the Mathematics Department, Quaid-i-Azam University, Islamabad. We are very grateful for extremely useful comments to an unknown referee.

[1] A. Pervez, A. Qadir, and A. A. Siddiqui, Phys. Rev. D **51**, 4598 (1995).

[2] D. R. Brill, J. M. Cavallo, and J. A. Isenberg, J. Math. Phys. **21**, 2789 (1980).
[3] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, New York, 1973), pp. 827–835.
[4] See Ref. 3, pp. 917–920.
[5] A. Qadir, *Einstein's General Theory of Relativity: An Introduction* (under preparation).
[6] A. Qadir and J. A. Weeler, in *Form* Su(3) *to Gravity: Yuval Neeman Festchrift*, edited by E. S. Gotsman and G. Tauber (Cambridge University Press, Cambridge, England, 1985); in *Spacetime Symmetries*, Proceedings of the Workshop, College Park, Maryland, 1988 edited by Y. S. Kim and W. W. Zachary [Nucl. Phys. B (Proc. Suppl.) **6**, 345 (1989)].
[7] R. A. Matzner, N. Zamorano, and V. D. Sandberg, Phys. Rev. D **19**, 2821 (1979).
[8] M. Simpson and R. Penrose, Int. J. Theor. Phys. **7**, 183 (1973).

Exact solutions in Einstein–Yang–Mills–Dirac systems

Gerd Rudolph and Torsten Tok
Institute of Theoretical Physics, University Leipzig, 04109 Leipzig, Augustusplatz 10, Germany

Igor Volobuev[a)]
Institute of Nuclear Physics, Moscow State University, 119899 Moscow, Russia

(Received 11 June 1998; accepted for publication 13 July 1999)

For the first time, exact solutions for the full self-consistent Einstein–Yang–Mills–Dirac systems with gauge groups $SU(2)$ and $SU(4)$ in Robertson–Walker space–time $R\times S^3$ are presented, which are symmetric under the action of the group $SO(4)$ of spatial rotations. The approach is based on the dimensional reduction method for gauge, spinor, and gravitational fields and relates symmetric solutions in EYMD theory to certain solutions of an effective dynamic system. The solutions are interpreted as solutions of the cosmological type with an oscillating Yang–Mills field, which exchanges energy with a spinor field. The explicit form of the solution for the spinor field shows that its energy changes the sign in the process of the evolution of the Yang–Mills field from one vacuum to another. © *1999 American Institute of Physics.* [S0022-2488(99)04710-6]

I. INTRODUCTION

Exact solutions in gravity coupled to fields of different types have always attracted much attention. In particular, the last few years witnessed a great interest in solutions to EYM systems. There were found both numerical[1] and exact solutions with $SO(3)$ and $SO(4)$ groups of spatial symmetry.[2,3] The exact $SO(3)$-symmetric solutions turn out to be static and singular and are, in fact, a generalization of Reissner–Nordström solutions to non-Abelian gauge theories. The $SO(4)$-symmetric solutions correspond to the Robertson–Walker ansatz for the metric and are interpreted either as wormhole solutions in the Euclidean domain[4] or as cosmological solutions for the radiation dominated universe.[5]

The study of exact solutions in Einstein–Dirac systems also attracted attention.[6,7] It was found that the Robertson-Walker ansatz in such a system leads to the so-called ''ghost solutions'' in ED systems, for which the Dirac field has a vanishing energy–momentum tensor.

In the present paper we consider a self-consistent EYMD system and find exact solutions for the case of the gauge groups $SU(2)$ and $SU(4)$. The group $SU(4)$ is the simplest gauge group, which gives qualitatively new results in comparison with the group $SU(2)$. We would like to note that the obtained solutions are interesting in their own right as the first example of exact self-consistent solutions in the Einstein–Yang–Mills–Dirac system. Our solutions are of the radiation-dominated type and describe an exchange of energy between the YM and Dirac fields. They can be considered as a first step toward finding cosmological solutions with the energy–momentum tensors derived from the fundamental Lagrangians of particle physics, rather than with phenomenological ones, which is important for treating the dynamics of the early Universe. We also show that our solutions explain the phenomenon of the ''ghost solutions'' in ED systems.

To find these solutions we employ the dimensional reduction method for gravitational, gauge, and spinor fields,[8,9] which enables us to relate symmetric solutions in EYMD system to certain solutions of an effective dynamic system and essentially simplifies the problem of finding symmetric solutions.

[a)]Electronic mail: volobuev@theory.npi.msu.su

II. THE EFFECTIVE ACTION

We consider an $SO(4)$-symmetric Einstein–Yang–Mills–Dirac system with the standard action in space–time $M=R\times S^3$:

$$S=S_{\rm E}+S_{\rm YM}+S_{\rm D}, \tag{1}$$

where

$$S_{\rm E}=\frac{1}{16\pi\kappa}\int_M (R-\Lambda)\mathrm{d}v, \tag{2}$$

$$S_{\rm YM}=\frac{1}{8g^2}\int_M \mathrm{tr}(F_{\alpha\beta}F^{\alpha\beta})\mathrm{d}v, \tag{3}$$

$$S_{\rm D}=\int_M \frac{i}{2}\bar{\psi}\gamma^\alpha E_\alpha{}^\mu\left(\frac{\partial}{\partial x^\mu}-\frac{1}{8}\,\omega_{\mu\alpha\beta}[\gamma^\alpha,\gamma^\beta]+\delta'(A_\mu)\right)\psi\,\mathrm{d}v+\mathrm{h.c.} \tag{4}$$

Here Λ denotes the cosmological constant (which is assumed to be positive and very small, in accordance with the existing experimental data), R is the scalar curvature, and $dv=\sqrt{|\det(\mathbf{g})|}\mathrm{d}^4x$ is the canonical volume form corresponding to a metric $\mathbf{g}$ on $M=R\times S^3$. F and A are the Yang–Mills field strength and potential, respectively, and the trace is taken in the adjoint representation of the Lie algebra of the gauge group. In the Dirac action, ψ denotes the spinor field, $\{E_\alpha=E_\alpha{}^\mu\partial_\mu\}$ is an orthonormal frame on $M=R\times S^3$, $\omega_{\mu\alpha\beta}=\omega_{\alpha\beta}(\partial_\mu)=\mathbf{g}(E_\alpha,\boldsymbol{\nabla}_\mu E_\beta)$ is the spin connection, $\{\gamma^\alpha\}$ are the gamma matrices, δ is the representation of the gauge group in the spinor space, and δ' denotes the induced representation of its Lie algebra. Finally, h.c. stands for Hermitian conjugate.

We identify S^3 with the group manifold of $SU(2)$. Then the action σ of the symmetry group, $K=SO(4)\equiv(SU(2)\times SU(2))/\{\mathbf{1},-\mathbf{1}\}$ on $SU(2)$, is given by

$$\sigma((k_1,k_2),x)=k_1xk_2{}^{-1},\quad (k_1,k_2)\in SU(2)\times SU(2),\quad x\in SU(2). \tag{5}$$

The isotropy subgroup H at $x=\mathbf{1}_{SU(2)}$ is isomorphic to $SU(2)$, and is given by

$$H=\{(k,k)\in K,\ \ k\in SU(2)\}. \tag{6}$$

Further, we have the reductive decomposition $\mathfrak{K}=\mathfrak{H}\oplus\mathfrak{M}$ of the Lie algebra $\mathfrak{K}$ of K, where

$$\mathfrak{H}=\{(X,X);\quad X\in su(2)\}, \tag{7}$$

$$\mathfrak{M}=\{(X,-X);\quad X\in su(2)\}. \tag{8}$$

Obviously $S^3=(SU(2)\times SU(2))/H$ is a symmetric space, and $SO(4)$ invariance is equivalent to invariance under $SU(2)\times SU(2)$. The isotropy representation of $H=SU(2)$ in $\mathfrak{M}$ is $\underset{\sim}{3}$, i.e., it is equivalent to the adjoint representation.

Now we have to reduce the action of the original theory due to the $SO(4)$ symmetry. We begin with the action of the gravitational field.

The most general form of an $SO(4)$-invariant metric $\mathbf{g}$ on M is

$$\mathbf{g}=-N^2(t)\mathrm{d}t^2+a^2(t)\mathrm{d}\Omega^2_{S^3}, \tag{9}$$

where $\mathrm{d}\Omega^2_{S^3}$ is the standard line element on a 3-sphere of radius 2 corresponding to the Killing metric on $SU(2)\equiv S^3$ multiplied by $-\frac{1}{2}$, and $N(t),a(t)$ are arbitrary functions of time. For the sake of the future convenience, we assume $N(t)$ and $a(t)$ to have the dimension of length and t and $\mathrm{d}\Omega^2_{S^3}$ to be dimensionless. Though the metric (9) can be brought to the conformal form with

$N(t)=a(t)$ by an appropriate reparametrization of time, in the Lagrangian formalism $N(t)$ must be kept arbitrary, until the equations of motion are obtained with the help of the variational principle, because it plays the role of a Lagrange multiplier in the gravitational action (2).

An orthonormal coframe $\{\theta^\mu\}$ on M is given by

$$\theta^0=N\,\mathrm{d}t, \quad \theta^i=a\,\vartheta^i, \quad i=1,2,3, \tag{10}$$

where $\vartheta^i\sigma_i/(2i)\equiv\vartheta$ is the canonical left-invariant 1-form on $SU(2)$, σ_i being the Pauli matrices. In what follows we put $\tau_k=\sigma_k/(2i)$ and denote by $\eta_{\alpha\beta}$ the Minkowski metric $\eta_{\alpha\beta}=\mathrm{diag}(-1,1,1,1)$. It is a matter of simple calculations to find the components of the spin connection on M:

$$\begin{aligned}\omega_{0i}&=-\omega_{i0}=-\dot{a}\frac{1}{aN}\,\delta_{il}\theta^l,\\ \omega_{ik}&=-\frac{1}{2a}\,\epsilon_{ikl}\theta^l, \quad i=1,2,3.\end{aligned} \tag{11}$$

Using the standard formulas for the curvature, substituting it into (2), and integrating over S^3, one easily finds the reduced gravitational action,

$$S_{\mathrm{E}}=\frac{16\pi^2}{16\pi\kappa}\int_R a^3N\left(\frac{3}{2a^2}+6\left(\dot{a}\frac{1}{aN}\right)^2-\dot{a}\dot{N}\frac{6}{aN^3}+\ddot{a}\frac{6}{aN^2}-\Lambda\right)\mathrm{d}t, \tag{12}$$

where $16\pi^2$ is the volume of $SU(2)=S^3$ with the standard metric $\mathrm{d}\Omega^2_{S^3}$.

In accordance with the principle of the least action, to obtain the corresponding equations of motion we have to vary this action, keeping all the variations equal to zero at the boundaries. Therefore, it is possible to omit a complete divergence in this action without altering the equations of motion (Ref. 10, Sec. 93). Doing so, we get the effective action

$$S_{\mathrm{E}}^{\mathrm{eff}}=\frac{16\pi^2}{16\pi\kappa}\int_R\left(\frac{3}{2}aN-\Lambda a^3N-6\,\frac{1}{N}a\dot{a}^2\right)\mathrm{d}t, \tag{13}$$

which we consider as the reduced action of the gravitational field.

Next, we turn to the gauge field action (3). It is common knowledge that symmetric gauge fields are pull-backs to the base manifold of the invariant connections in the corresponding principal fiber bundles. The classification of invariant connections for the case of fiber-transitive group action is given by Wang's theorem (Ref. 11, Chap. 2, Sec. 11). A generalization to the case of nontransitive symmetry group action was given in Ref. 12, and the problem of lifting the symmetry group action to the bundle was considered in Ref. 13. Using the results of these papers, we get that an $SO(4)$-symmetric gauge potential A on $R\times S^3$ is in one-to-one correspondence with a triplet $\{\tau, \mathcal{A}, \Phi\}$, where τ is a homomorphism from the isotropy group H into the gauge group G, defined up to conjugacy,

$$\tau: H\to G, \tag{14}$$

$\mathcal{A}$ is a gauge potential on R with values in the centralizer $\mathfrak{C}_{\mathfrak{G}}(\tau'(\mathfrak{H}))$ of $\tau'(\mathfrak{H})$ in $\mathfrak{G}=\mathrm{Lie}(G)$ and Φ is a linear mapping,

$$\Phi: R\to\mathfrak{M}^*\otimes\mathfrak{G}, \tag{15}$$

satisfying

$$\Phi\circ\mathrm{Ad}(h)=\mathrm{Ad}(\tau(h))\circ\Phi, \quad \forall h\in H. \tag{16}$$

Here $\mathrm{Ad}(h)$ is the restriction to H of the adjoint representation of K applied to $\mathfrak{M}$ and $\mathrm{Ad}(\tau(h))$ is the restriction to $\tau(H)$ of the adjoint representation of G in $\mathfrak{G}$. In fact, Eq. (16) means that the mapping Φ intertwines these representations.

As we have already mentioned, we will consider gauge groups $SU(2)$ and $SU(4)$. It turns out that the corresponding principal bundles are trivial, because they are classified by the second Chern number, which is equal to zero due to $H^4(R\times S^3)=0$. Therefore, the gauge potential A is defined on $R\times S^3$ globally for both gauge groups.

Further, since the isotropy subgroup is simple, the homomorphism τ can be either injective or trivial. In the latter case the intertwining operator $\Phi=0$, and it is easy to see that the spinor field decouples from the EYM system, which leads to the known ''ghost solutions'' for this field. The same is valid, when the representation δ is trivial. Therefore, we discard these possibilities in the sequel.

If the homomorphism τ is injective, the intertwining operator Φ is nontrivial for any gauge group G, because the isotropy representation in the case under consideration is equivalent to the adjoint one and the restriction to $\tau(H)$ of the adjoint representation of G always contains at least one adjoint representation of H. Thus, we see that the gauge fields on $R\times S^3$ always possess symmetric degrees of freedom for injective homomorphisms τ.

As for the spinor field, the situation is more complicated: it is quite possible that there are no symmetric spinor fields on $R\times S^3$ for some injective homomorphisms τ and some representations δ of the gauge group carried by the spinor field. A sufficient condition for the existence of symmetric spinor fields on $R\times S^3$ is that the restriction of δ to $\tau(H)$ contains a representation $\underline{2}$ of $H=SU(2)$ (see, for example, Ref. 9). If we restrict ourselves to gauge groups of the $SU(n)$ series and to the fundamental representations δ, this condition is fulfilled, in particular, for regular embeddings of $SU(2)$ into $SU(n)$, but it turns out that the resulting effective dynamic systems are the same, as for the gauge group $SU(2)$. The first qualitatively new case is that of the gauge group $SU(4)$ with the nonregular embedding $SU(2)\rightarrow SU(4)$ defined by the decomposition of the fundamental representation of $SU(4)$, $\underline{4}\rightarrow\underline{2}\oplus\underline{2}$. Effective dynamic systems for other gauge groups and similar nontrivial embedding look very much like this one. It is for these reasons that in the present paper we restrict ourselves to the consideration of the gauge groups $SU(2)$ and $SU(4)$.

It is easy to see that in the case of the gauge group $SU(2)$ the centralizer is trivial, and there is no reduced gauge potential $\mathcal{A}$, whereas in the second case there can be a nontrivial centralizer. Therefore, the case of the group $SU(4)$ is more general, and we will carry out all the calculations for this case and then explain the difference from the case of $SU(2)$.

Thus, we take the gauge group $G=SU(4)$ and define the homomorphism $\tau:H\rightarrow G$ by the decomposition of the fundamental representation $\underline{4}$ of $SU(4)$:

$$\underline{4}\rightarrow(\underline{2},\underline{2}), \tag{17}$$

i.e., we represent the space C^4 as the tensor product $C^2\otimes C^2$ and let the fundamental representation of $\tau(H)$ act on the first factor and the fundamental representation of the centralizer $C_G(\tau(H))=SU(2)$ act on the second factor.

It is known that the adjoint representation of $sl(n)$ can be expressed in terms of its fundamental representation by[14]

$$\mathrm{ad}\, sl(n)=\underline{n}^*\tilde{\otimes}\,\underline{n}, \tag{18}$$

where an overtilde means dropping a one-dimensional trivial representation and $\underline{n}^*$ denotes the contragradient representation, $t(x)^*=-t(x)^T$. Therefore, we get

$$\mathrm{ad}\, su(4)\rightarrow(\underline{2},\underline{2})^*\tilde{\otimes}(\underline{2},\underline{2})=(\underline{3},\underline{1})\oplus(\underline{3},\underline{3})\oplus(\underline{1},\underline{3}). \tag{19}$$

A basis in $su(4)$ adapted to this decomposition is

$$H_k = \tau_k \otimes \mathbf{1} = \begin{pmatrix} \tau_k & 0 \\ 0 & \tau_k \end{pmatrix}, \tag{20}$$

$$P_{kA} = \tau_k \otimes \sigma_A, \tag{21}$$

$$H_A = \mathbf{1} \otimes \tau_A, \tag{22}$$

where Eq. (20) explains our rule for evaluating the tensor product of matrices. It is easy to calculate the commutators of these generators, for instance,

$$[P_{kA}, P_{lB}] = \epsilon_{klm} H_m \delta_{AB} + \epsilon_{ABC} H_C \delta_{kl}. \tag{23}$$

The constraint (16) means that the mapping Φ is an intertwining operator, which intertwines the isotropy representation $\underline{3}$ of $H = SU(2)$ in $\mathfrak{M}$ with the representation (19). We introduce basic intertwining operators from $\mathfrak{M}$ into $\mathfrak{G}$ defined by the relations

$$I((\tau_k, -\tau_k)) = H_k, \tag{24}$$

$$I_A((\tau_k, -\tau_k)) = P_{kA}. \tag{25}$$

Then for any $X = (X^k \tau_k) \in su(2)$, we have

$$\Phi((X, -X)) = \xi I(X, -X) + \xi^A I_A(X, -X), \tag{26}$$

$$= X^k(\xi H_k + \xi^A P_{kA}) = X \otimes (\xi \mathbf{1} + \xi^A \sigma_A), \tag{27}$$

where ξ and ξ^A, $A = 1,2,3$, are real-valued functions on R. In what follows we denote the vector (ξ^1, ξ^2, ξ^3) by $\vec{\xi}$ and put $\hat{\xi} = \xi \mathbf{1} + \xi^A \sigma_A$ and $\tilde{\xi} = \xi^A \sigma_A$. The centralizer $\mathfrak{C}_{su(4)}(\tau'(su(2)))$ of $\tau'(su(2))$ in $\mathfrak{G} = su(4)$ is $su(2)$ and is spanned by the Lie algebra elements H_A, $A = 1,2,3$. The matrix $\hat{\xi}$ is in the representation $(\underline{1} + \underline{3})$ of the centralizer $\mathfrak{C}$ and $i\hat{\xi} \in u(2)$.

The symmetric gauge potential A on $R \times S^3$ can be easily expressed in terms of the matrix $\hat{\xi}$ and the canonical left-invariant 1-form $\vartheta = \vartheta^i \tau_i$ on $SU(2) \equiv S^3$:

$$A = \frac{\xi + 1}{2} H_i \vartheta^i + \frac{1}{2} \xi^A P_{kA} \vartheta^k + A_0^A H_A \, dt = \frac{1}{2} \vartheta \otimes (\hat{\xi} + \mathbf{1}) + \mathbf{1} \otimes A_0 \, dt. \tag{28}$$

Here A_0 is a function on R with values in $su(2)$, i.e., $A_0 = A_0^A \tau_A$. The term $\mathbf{1} \otimes A_0 \, dt$ is the reduced gauge potential on R, which can be gauged out, but we keep it for the moment, because it is necessary for deriving the equations of motion. It is not difficult to calculate the corresponding field strength F and to obtain the reduced Yang–Mills action, which takes the simple form

$$S_{\mathrm{YM}} = \frac{16\pi^2}{8g^2} \int_R 24 \left(\frac{a}{2N} \dot{\xi}^2 + \frac{a}{2N} (\dot{\vec{\xi}} + \vec{A}_0 \times \vec{\xi})^2 - \frac{N}{8a} ((\xi^2 + (\vec{\xi})^2 - 1)^2 + 4\xi^2 (\vec{\xi})^2) \right) dt, \tag{29}$$

where $\vec{A}_0$ is the vector (A_0^1, A_0^2, A_0^3) and $\vec{A}_0 \times \vec{\xi}$ is the vector product of $\vec{A}_0$ and $\vec{\xi}$.

If the gauge group is $G = SU(2)$, the unique (up to conjugacy) nontrivial homomorphism $\tau: H \to G$ can be defined by the identity mapping, i.e.,

$$\tau((k,k)) = k, \quad k \in SU(2). \tag{30}$$

In this case, the centralizer $\mathfrak{C}_{\mathfrak{G}}(\tau'(\mathfrak{H}))$ is trivial, and the intertwining operator is

$$\Phi((X, -X)) = \xi(t) X, \quad X \in su(2), \quad \xi(t) \in R. \tag{31}$$

It is clear that the gauge potential A on $R\times S^3$ still has the form (28) with $\vec{\xi}=0$, $A_0=0$. The reduced Yang–Mills action is also given by (29), provided one puts $\vec{\xi}=0$ and rescales the coupling constant $g\rightarrow 2g$.

Next we have to reduce the action (4) for a symmetric spinor field. We choose δ to be the fundamental representation, i.e., we can write the spinor ψ as a 4×4, resp., 4×2 matrix on which an element g of $SU(4)$, resp., $SU(2)$ acts via right multiplication by g^{-1}.

In accordance with our choice of the metric signature (9), we have

$$\{\gamma^\mu,\gamma^\nu\}=-2\,\eta^{\mu\nu},\quad \eta^{\mu\nu}=\mathrm{diag}(-1,1,1,1),\quad \gamma^5=i\gamma^0\gamma^1\gamma^2\gamma^3, \tag{32}$$

$$\gamma^0=\begin{pmatrix}0 & \mathbf{1}\\ \mathbf{1} & 0\end{pmatrix},\quad \gamma^i=\begin{pmatrix}0 & -\sigma^i\\ \sigma^i & 0\end{pmatrix},\quad i=1,2,3,\quad \gamma^5=\begin{pmatrix}\mathbf{1} & 0\\ 0 & -\mathbf{1}\end{pmatrix}. \tag{33}$$

The bispinor representation $\Delta(s)$ is defined by

$$\Delta(s)^{-1}\gamma^\mu\Delta(s)=\Lambda(s)^\mu{}_\rho\gamma^\rho, \tag{34}$$

where s is an element of the group Spin(1,3) and Λ is the covering homomorphism from Spin(1,3) onto $SO(1,3)$. On the Lie algebra level we get

$$[\Delta'(A),\gamma^\mu]=-A^\mu{}_\nu\gamma^\nu,\quad \Delta'(A)=-\tfrac{1}{8}A^\mu{}_\nu[\gamma_\mu,\gamma^\nu], \tag{35}$$

where $A_{\mu\nu}=-A_{\nu\mu}$ is an element of $so(1,3)\equiv\mathrm{spin}(1,3)\equiv sl(2,C)$.

An $SO(4)$-symmetric spinor field ψ on $R\times S^3$ is in one-to-one correspondence with a matrix-valued function ρ on R, which satisfies the condition[9]

$$(\delta'(\tau'(h))+\Delta'(\lambda'(h)))\rho=0,\quad \forall h\in\mathfrak{H}. \tag{36}$$

Here $\lambda':\mathfrak{H}\rightarrow so(1,3)$ is the homomorphism induced by the isotropy representation, which can be calculated explicitly: $\lambda'(\tau_a)^b_c=-\epsilon^b_{ac}$. Therefore, if the gauge group G is $SU(4)$, Eq. (36) reads as

$$(\tfrac{1}{4}\gamma^i\gamma^j\epsilon_{ijk}+\delta'(H_k))\rho=((\tau_k\otimes\mathbf{1})\rho-\rho(\tau_k\otimes\mathbf{1}))=0. \tag{37}$$

The general solution of this constraint equation is

$$\rho=\begin{pmatrix}u_1\mathbf{1} & u_2\mathbf{1}\\ v_1\mathbf{1} & v_2\mathbf{1}\end{pmatrix}=\mathbf{1}\otimes\begin{pmatrix}u_1 & u_2\\ v_1 & v_2\end{pmatrix},\quad u_1,u_2,v_1,v_2\in C^\infty(R), \tag{38}$$

i.e., a symmetric spinor on $R\times S^3$ is parametrized by two complex doublets $u=(u_1,u_2)^T$ and $v=(v_1,v_2)^T$, one for each chirality. We see from (38) that the reduced gauge group $C=SU(2)$ acts on both doublets by the fundamental representation. Taking into account Eqs. (11) and (28), it is a matter of simple calculations to get the reduced action,

$$S_{\mathrm{D}}=16\pi^2\int_R a^3N\left(\frac{i}{N}(\bar{u}\dot{u}-\dot{\bar{u}}u+\bar{v}\dot{v}-\dot{\bar{v}}v)+\frac{1}{N}A_0^B(\bar{u}\sigma_Bu+\bar{v}\sigma_Bv)-\frac{3}{2a}(\bar{u}\hat{\xi}u-\bar{v}\hat{\xi}v)\right)\mathrm{d}t. \tag{39}$$

If the gauge group G is $SU(2)$, Eq. (36) reads as

$$((\tau_k\otimes\mathbf{1})\rho-\rho\tau_k)=0, \tag{40}$$

and we get

$$\rho = \begin{pmatrix} u\mathbf{1} \\ v\mathbf{1} \end{pmatrix} = \mathbf{1} \otimes \begin{pmatrix} u \\ v \end{pmatrix}, \quad u,v \in C^{\infty}(R), \tag{41}$$

i.e., a symmetric spinor on $R \times S^3$ for $G = SU(2)$ depends on two arbitrary complex functions u and v, one for each chirality. The reduced action has the same form (39), if we put there $\vec{\xi} = 0$, $A_0 = 0$.

Now we can write down the reduced action of the coupled EYMD system. In the case of the gauge group $SU(4)$ it has the form

$$S = S_E^{\text{eff}} + S_{\text{YM}} + S_{\text{D}} = 16\pi^2 \int_R \Bigg\{ \frac{1}{16\pi\kappa} \left(\frac{3}{2} aN - \Lambda a^3 N - 6 \frac{a}{N} \dot{a}^2 \right)$$
$$+ \frac{24}{8g^2} \left(\frac{a}{2N} \dot{\xi}^2 + \frac{a}{2N} (\dot{\vec{\xi}} + \vec{A}_0 \times \vec{\xi})^2 - \frac{N}{8a} ((\xi^2 + (\vec{\xi})^2 - 1)^2 + 4\xi^2 (\vec{\xi})^2) \right)$$
$$+ \left(i a^3 (\bar{u}\dot{u} - \dot{\bar{u}} u + \bar{v}\dot{v} - \dot{\bar{v}} v) + a^3 A_0^B (\bar{u}\sigma_B u + \bar{v}\sigma_B v) - \frac{3a^2 N}{2} (\bar{u}\hat{\xi} u - \bar{v}\hat{\xi} v) \right) \Bigg\} \mathrm{d}t. \tag{42}$$

If we choose the gauge group to be $SU(2)$, we have to put $\vec{\xi} = 0$, $A_0 = 0$ in this action, to rescale the coupling constant $g \to 2g$, and to take into account that the variables u and v are no longer isospinors, but ordinary functions.

At this point, we have to comment on the meaning of our classical spinor field. It is clear that it can describe the symmetric energy levels of the Dirac field in the EYM background: either the lowest one with the positive energy, or the highest one with the negative energy. Therefore, we have to assume that the spinor field of our solutions is a one-particle Dirac wave function and should be normalized to unity, which, in the case of symmetric fields, amounts to

$$16\pi^2 a^3 (\bar{u}u + \bar{v}v) = 1. \tag{43}$$

It turns out that this constraint is compatible with the equations of motion derived from the action (42).

III. THE FIELD EQUATIONS AND SOLUTIONS

Variation of this action with respect to a, N, ξ, $\vec{\xi}$, u, v and $\bar{u}$, $\bar{v}$ gives us the field equations. When taken in the special gauge $A_0 = 0$ and in the conformal time [the latter condition means that now $N(t) = a(t)$], they have the form

$$a \frac{\delta}{\delta a} S = \frac{1}{16\pi\kappa} \left(\frac{3}{2} a^2 - 3\Lambda a^4 - 6\dot{a}^2 + 12 a\ddot{a} \right) + \frac{3}{2g^2} \left(\dot{\xi}^2 + (\dot{\vec{\xi}})^2 + \frac{1}{4} ((\xi^2 + (\vec{\xi})^2 - 1)^2 + 4\xi^2 (\vec{\xi})^2) \right)$$
$$+ 3a^3 (i(\bar{u}\dot{u} - \dot{\bar{u}} u + \bar{v}\dot{v} - \dot{\bar{v}} v) - (\bar{u}\hat{\xi} u - \bar{v}\hat{\xi} v)) = 0, \tag{44}$$

$$a \frac{\delta}{\delta N} S = \frac{1}{16\pi\kappa} \left(\frac{3}{2} a^2 - \Lambda a^4 + 6\dot{a}^2 \right) - \frac{3}{2g^2} \left(\dot{\xi}^2 + (\dot{\vec{\xi}})^2 + \frac{1}{4} ((\xi^2 + (\vec{\xi})^2 - 1)^2 + 4\xi^2 (\vec{\xi})^2) \right)$$
$$- \frac{3}{2} a^3 (\bar{u}\hat{\xi} u - \bar{v}\hat{\xi} v) = 0, \tag{45}$$

$$\frac{\delta}{\delta \xi} S = -\frac{3}{2g^2} (2\ddot{\xi} + (\xi(\xi^2 + (\vec{\xi})^2 - 1) + 2\xi(\vec{\xi})^2)) - \frac{3}{2} a^3 (\bar{u}u - \bar{v}v) = 0, \tag{46}$$

$$\frac{\delta}{\delta \xi^A} S = -\frac{3}{g^2} \left(\ddot{\xi}^A + \frac{1}{2} (\xi^A (\xi^2 + (\vec{\xi})^2 - 1) + 2\xi^2 \xi^A) \right) - \frac{3}{2} a^3 (\bar{u}\sigma^A u - \bar{v}\sigma^A v) = 0, \tag{47}$$

$$\frac{1}{2}a^{-3/2}\frac{\delta}{\delta\overline{u}}S=i\frac{d}{dt}a^{3/2}u-\frac{3}{4}\hat{\xi}a^{3/2}u=0,$$
$$\frac{1}{2}a^{-3/2}\frac{\delta}{\delta u}S=i\frac{d}{dt}a^{3/2}\overline{u}+\frac{3}{4}a^{3/2}\overline{u}\hat{\xi}=0, \tag{48}$$

$$\frac{1}{2}a^{-3/2}\frac{\delta}{\delta\overline{v}}S=i\frac{d}{dt}a^{3/2}v+\frac{3}{4}\hat{\xi}a^{3/2}v=0,$$
$$\frac{1}{2}a^{-3/2}\frac{\delta}{\delta v}S=i\frac{d}{dt}a^{3/2}\overline{v}-\frac{3}{4}a^{3/2}\overline{v}\hat{\xi}=0. \tag{49}$$

Variation with respect to A_0 gives a constraint,

$$\frac{\delta S}{\delta A_0^B}=\frac{3}{g^2}(\epsilon_{BCD}\xi^C\dot{\xi}^D)+a^3(\overline{u}\sigma_B u+\overline{v}\sigma_B v)=0, \tag{50}$$

which means that the total isospin of the gauge and the spinor fields equals zero.

Now we will show that it is possible to find exact solutions to this system of equations. We begin with the simpler case of the gauge group $SU(2)$. There is no constraint (50) in this case, and we also have to drop the equation (47), to put $\vec{\xi}=0$ in the others and to rescale $g\to 2g$.

The Dirac equations (48) and (49) give

$$\overline{u}u=\frac{C_u}{a^3},\quad \overline{v}v=\frac{C_v}{a^3} \tag{51}$$

and

$$\overline{u}\dot{u}-\dot{\overline{u}}u=-\frac{3i}{2}\xi\overline{u}u, \tag{52}$$

$$\overline{v}\dot{v}-\dot{\overline{v}}v=\frac{3i}{2}\xi\overline{v}v, \tag{53}$$

where, in accordance with (43), the positive constants C_u and C_v fulfill the normalization condition

$$16\pi^2(C_u+C_v)=1. \tag{54}$$

We will discuss the explicit choice of these constants later.

Now Eqs. (44) and (45) simplify considerably. Their sum gives

$$\frac{1}{16\pi\kappa}(3a^2-4\Lambda a^4+12a\ddot{a})=0. \tag{55}$$

Multiplying this equation by $\dot{a}/a$ and integrating, we see that

$$\tfrac{3}{2}a^2-\Lambda a^4+6\dot{a}^2=E, \tag{56}$$

where E is an arbitrary constant, which has the meaning of the total energy of the system. Equation (56) is the standard Friedman equation for the radiation-dominated universe and has a simple analog in mechanics. We can consider a as the coordinate of a particle with unit mass and energy $E/12$, which moves in an inverted double-well potential (recall that $\Lambda>0$),

$$W(a)=\tfrac{1}{8}a^2-\tfrac{1}{12}\Lambda a^4. \tag{57}$$

If $\Lambda<\Lambda_E=9/16E$, then the motion will be periodic. This means that our solution describes a universe that first expands and then contracts, where $a=0$ corresponds to a singular metric in the beginning and the end. If $\Lambda\geqslant\Lambda_E$, then the solution can be either static or can describe an expanding or contracting universe, but such solutions are of no interest from the physical point of view, and we will not discuss them in more detail.

Next we turn to the YM equation (46) (we recall that we have rescaled $g\to 2g$ in the case under consideration). Due to Eq. (51), it decouples from the equations for u and v:

$$\frac{3}{8g^2}(2\ddot{\xi}+(\xi(\xi^2-1)))+\frac{3}{2}(C_u-C_v)=0. \tag{58}$$

The first integral of this equation is

$$\frac{3}{8g^2}\left(\dot{\xi}^2+\frac{1}{4}(\xi^2-1)^2\right)+\frac{3}{2}\xi(C_u-C_v)=\frac{E}{16\pi k}, \tag{59}$$

where the integration constant is due to Eqs. (45) and (56).

Equation (59) also has an analog in mechanics. A point particle with unit mass, energy $g^2E/(12\pi\kappa)$ and coordinate ξ, moves in a double-well potential,

$$V(\xi)=\tfrac{1}{8}(\xi^2-1)^2+2g^2\xi(C_u-C_v). \tag{60}$$

We can interpret the first term in (59) as the energy of the Yang–Mills field and the second term as the energy of the Dirac field due to the interaction with the gauge field, the equation describing an exchange of energy between the two fields. The exact solution of Eq. (59) is possible in terms of elliptic functions of the first kind.[15]

It is not difficult to see that for sufficiently small values of the coupling constant, for example, $g^2/4\pi<1$, potential (60) always has three extrema: the absolute minimum, a local minimum, and a local maximum. We consider, for instance, the case where $C_u-C_v<0$ and the energy of the system $g^2E/(12\pi\kappa)$ lies between the local minimum and the local maximum of the potential $V(\xi)$. Then the system will move between the turning points defined by the real zeros of the polynomial,

$$\frac{g^2E}{12\pi\kappa}-V(\xi). \tag{61}$$

In the case under consideration we have four real zeros $d<c<b<a$, i.e.,

$$\frac{1}{2}\dot{\xi}^2=\frac{g^2E}{12\pi\kappa}-V(\xi)=\frac{1}{8}(a-\xi)(b-\xi)(c-\xi)(\xi-d). \tag{62}$$

Moreover, the form of the potential (60) stipulates that in the case $C_u-C_v<0$ we have $d<c<0$.

The equations of motion have two solutions with either $b\leqslant\xi\leqslant a$ or $d\leqslant\xi\leqslant c$. The solution for the latter case is given by the integral

$$t(\xi)=\int_d^\xi\frac{dx}{\sqrt{\dfrac{g^2E}{6\pi\kappa}-\dfrac{1}{4}(x^2-1)^2-4g^2x(C_u-C_v)}}, \tag{63}$$

which can be solved explicitly,[15]

$$t(\xi)=\frac{2}{\sqrt{(a-a)(b-d)}}F\left(\arcsin\sqrt{\frac{(a-c)(\xi-d)}{(c-d)(a-\xi)}},\sqrt{\frac{(a-b)(c-d)}{(a-c)(b-d)}}\right), \tag{64}$$

where F is the elliptic integral of the first kind. This function can be inverted, and we can get $\xi(t)$ expressed in terms of the Jacobi elliptic function *sinus amplitudinis*. Solutions for other values of E and C_u-C_v can be written out in a similar way with the help of the formulas available in Ref. 15.

With a given solution ξ we can solve the Dirac equations (48) and (49),

$$u=\sqrt{\frac{C_u}{a^3}}\exp\left\{-i\int\frac{3}{4}\xi\,\mathrm{d}t\right\}, \tag{65}$$

$$v=\sqrt{\frac{C_v}{a^3}}\exp\left\{i\int\frac{3}{4}\xi\mathrm{d}t\right\}. \tag{66}$$

One easily checks that our solution fulfills the whole system of field equations.

Now we consider the case of gauge group $SU(4)$. The Dirac equations (48) and (49) again lead to (51), but instead of (52) and (53) we now have

$$\bar{u}\dot{u}-\dot{\bar{u}}u=-\frac{3i}{2}\bar{u}\hat{\xi}u, \tag{67}$$

$$\bar{v}\dot{v}-\dot{\bar{v}}v=\frac{3i}{2}\bar{v}\hat{\xi}v. \tag{68}$$

These equations are also sufficient to decouple the Friedman equation from the Yang–Mills–Dirac equations, and we get again the equation (56) for the scale factor a.

Now we have to solve the Yang–Mills equations (46) and (47). We start with the discussion of the constraint (50). In what follows we restrict ourselves to the case where

$$\epsilon_{BCD}\xi^C\dot{\xi}^D=0,\quad \text{for } B=1,2,3, \tag{69}$$

i.e., the isospins of the gauge field and the Dirac field vanish separately. This equation means that the angular momentum of the motion in the $\vec{\xi}$ space equals zero, that is, the motion goes along a straight line passing through the origin, and the vector $\vec{\xi}$ is always proportional to a fixed vector $\vec{\xi}_0$. We use the remaining gauge freedom to choose $\vec{\xi}_0=(0,0,1)$, i.e., $\tilde{\xi}=\zeta\sigma_3$, $\zeta\in R$. In this gauge we obtain from the Yang–Mills equations (47),

$$\bar{u}\sigma_A u-\bar{v}\sigma_A v=0,\quad \text{for } A=1,2. \tag{70}$$

Further, we get from Eqs. (50) and (69),

$$\bar{u}\sigma_B u+\bar{v}\sigma_B v=0,\quad B=1,2,3. \tag{71}$$

Hence, we have $\bar{u}\sigma_1 u=\bar{u}\sigma_2 u=0$. These equations have two solutions:

$$\text{1st case:}\quad u=a^{-3/2}\alpha_1 w_+,\quad v=a^{-3/2}\beta_1 w_- \tag{72}$$

$$\text{2nd case:}\quad u=a^{-3/2}\alpha_2 w_-,\quad v=a^{-3/2}\beta_2 w_+, \tag{73}$$

where

$$w_+ = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad w_- = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{74}$$

and $\alpha_i, \beta_i \in C$, $|\alpha_i| = |\beta_i|$, $i = 1,2$. In particular, we obtain $C_u = C_v = |\alpha_i|^2 = 1/(32\pi^2)$ in the ith case, because the constants also must fulfill the normalization condition (54).

If we put $\tilde{\xi} = \zeta\sigma_3 \equiv 0$, then the Yang–Mills equations (47) demand $u = v = 0$. Therefore, we consider only the nontrivial case, when $\tilde{\xi} \not\equiv 0$. Hence, we have two Yang–Mills equations (46) and (47), which now take the form

$$\frac{3}{2g^2}(2\ddot{\xi} + (\xi(\xi^2 + \zeta^2 - 1) + 2\xi\zeta^2)) = 0, \tag{75}$$

$$\frac{3}{2g^2}(2\ddot{\zeta} + (\zeta(\xi^2 + \zeta^2 - 1) + 2\xi^2\zeta)) + 3S = 0, \tag{76}$$

with $S := C_u = 1/(32\pi^2)$ in the first and $S := -C_u = -1/(32\pi^2)$ in the second case.

To solve Eqs. (75) and (76), we pass to new variables, in accordance with

$$\begin{aligned} \xi &= \tfrac{1}{2}(x + y), \\ \zeta &= \tfrac{1}{2}(x - y). \end{aligned} \tag{77}$$

It is easy to check that the equations for x and y decouple and take the form

$$\frac{3}{2g^2}(2\ddot{x} + x(x^2 - 1)) + 3S = 0, \tag{78}$$

$$\frac{3}{2g^2}(2\ddot{y} + y(y^2 - 1)) - 3S = 0. \tag{79}$$

The first integrals of these equations are

$$\frac{3}{2g^2}\left(\dot{x}^2 + \frac{1}{4}(x^2 - 1)^2\right) + 3Sx = \frac{E_1}{8\pi\kappa}, \tag{80}$$

$$\frac{3}{2g^2}\left(\dot{y}^2 + \frac{1}{4}(y^2 - 1)^2\right) - 3Sy = \frac{E_2}{8\pi\kappa}, \tag{81}$$

where due to (45) and (56) the constants E_1 and E_2 fulfill

$$E_1 + E_2 = E. \tag{82}$$

These equations can also be solved exactly in terms of Jacobi elliptic functions, but unlike the case of the group $SU(2)$, there will be two different periods of motion in x and y.

We would like to note here that equations of the type (75), (76) with $S = 0$ for the Euclidean EYM system were first found in Ref. 4, but they were solved there only for the case of either $\xi = 0$ or $\zeta = 0$. It is likely that the existence of two periods can modify the equations for the energy quantization of the wormholes.[4,16]

Substituting the solutions for ξ and ζ into Eq. (28), we get

$$A=\frac{1}{2}\,\theta\otimes(\hat{\xi}+\mathbf{1})=\begin{pmatrix}\frac{1+x}{2}\,\theta & 0\\ 0 & \frac{1+y}{2}\,\theta\end{pmatrix}, \tag{83}$$

i.e., the gauge potential A takes values only in the $su(2)\oplus su(2)$ subalgebra of $su(4)$, each $su(2)$ part of the gauge potential being coupled to only one of the spinor fields u, resp., v.

If we have a solution to the system of equations (80) and (81), it is easy to integrate the Dirac equations (48) and (49). With given solutions ξ and ζ we obtain

$$u=w+\sqrt{\frac{C_u}{a^3}}\exp\left\{-i\int\frac{3}{4}x\,\mathrm{d}t\right\}, \tag{84}$$

$$v=w_-\sqrt{\frac{C_u}{a^3}}\exp\left\{i\int\frac{3}{4}y\,\mathrm{d}t+i\varphi_1\right\} \tag{85}$$

in the first and

$$u=w_-\sqrt{\frac{C_u}{a^3}}\exp\left\{-i\int\frac{3}{4}y\,\mathrm{d}t\right\}, \tag{86}$$

$$v=w+\sqrt{\frac{C_u}{a^3}}\exp\left\{i\int\frac{3}{4}x\mathrm{d}t+i\varphi_2\right\}, \tag{87}$$

in the second case, φ_1 and φ_2 being arbitrary constant phases.

IV. DISCUSSION

In studying the self-consistent EYMD system we have found that the evolution of the metric decouples from the remaining system and is described by the Friedman equation for the radiation-dominated universe. The same result for the case of EYM systems was obtained earlier in Refs. 4, 5, and solutions for the spinor field in this background were studied in Ref. 17. Unlike the latter solutions, our solutions take into account the back-reaction of the spinor field on the YM and gravitational fields.

As we have already mentioned, our solutions describe just one energy level of the Dirac field in the EYM background, which is the symmetric one and therefore exactly the lowest one. It is easy to see from the Dirac equations (48), (49) that the symmetric lowest level has a nonzero energy, when $\xi\neq 0$ in the case $G=SU(2)$ and $\det\hat{\xi}=\xi^2-\tilde{\xi}^2\neq 0$ in the case $G=SU(4)$. It makes clear that, in the absence of gauge fields, the symmetric level of the Dirac field always has zero energy. Therefore, the corresponding energy–momentum tensor vanishes, and the $SO(4)$-invariant solution for the Dirac field becomes a ''ghost solution.''

Now we will discuss our solutions in more detail. We start with the simpler case of the gauge group $SU(2)$. In this case the Yang–Mills equations admit three static solutions: two minima and one local maximum of the potential,

$$V(\xi)=\tfrac{1}{8}(\xi^2-1)^2+2g^2\xi(C_u-C_v). \tag{88}$$

Of course, these are solutions of the whole system only if the constant E is equal to the value of $V(\xi)$ in these extrema. In the absence of the spinor field, these extrema are two vacua $\xi=-1$, $\xi=+1$ of the YM field with Chern–Simons numbers $q=0$ and $q=1$ and a sphaleronlike solution $\xi=0$ with Chern–Simons number $q=\frac{1}{2}$ lying on top of the potential barrier between the vacua, where the Chern–Simons number is defined for any t as

$$q(t)=\frac{1}{8\pi^2}\int_{\{t\}\times S^3}\mathrm{tr}_\delta\left(A\wedge \mathrm{d}A+\frac{1}{3}A\wedge[A,A]\right),$$

the trace being taken in the representation δ of the gauge group in the spinor space. In the case of symmetric gauge fields this number can be calculated explicitly to be $q(t)=(\frac{1}{4})(2+3\xi(t)-(\xi(t))^3)$. Chiral spinor field configuration shifts the position of the extrema and the Chern–Simons index of the corresponding gauge field configurations, which is quite natural in the presence of matter fields.[18–22]

By fine tuning the cosmological constant Λ and the energy E, we can get a static sphaleronlike solution of the whole system. This solution corresponds to the local maxima of $V(\xi)$ (60) and $W(a)$ (57). Obviously, this solution has two unstable modes—one in the gravitational and one in the gauge field sector. This is another indication that the static solution is a ''cosmological'' analog of the first Bartnik–McKinnon solution.[23]

Next, we observe that Eq. (45) is the (0,0) component of the Einstein equations, and therefore we can interpret the constant E, see Eq. (56), as the total energy of the YMD system. Then Eq. (59) describes, in particular, the exchange of energy between Yang–Mills and spinor fields. This interpretation is in agreement with the solutions (65) and (66) for the spinor field in the case of the gauge group $SU(2)$: the momentary frequency, resp., energy of the spinor field is given by the integrand in the exponent of the solutions, and this is up to a factor ξ.

At this point we encounter the problem of positive definiteness of the energy, which is well known in classical Dirac theory. We see that the left-handed and the right-handed spinors make contributions of opposite sign to the energy of the system (59). Therefore, a physically meaningful solution in our case can only be a solution with either left-handed or right-handed spinors and the Yang–Mills field ξ taking values in the interval, where the energy of the corresponding spinor field is positive. Thus, the solution (64) with $C_u=0$, $C_v=1/(16\pi^2)$ is physically meaningful, because $\xi<0$ for this solution and the energy of the left-handed spinor field v is positive definite, which is also seen from Eq. (59).

If the total energy of the YMD system is larger than the local maximum of the potential $V(\xi)$, the motion in the variable ξ, stemming from the YM field, will go over an interval, including both minima of the potential $V(\xi)$. When ξ crosses zero, the energy of the spinor field also changes its sign. This is in accordance with the observations in Refs. 17 and 24, where it was shown that the spinor field has zero modes in the sphaleron background and that moving between neighboring vacua of the gauge field results in a shift of the energy levels of the spinor field. We also obtain from our solution that the effect is opposite for the spinor fields with opposite chirality.

However, when ξ crosses zero, our solution describes spinor matter with negative energy, which is unphysical. Furthermore, because we are dealing with a one-particle system, we cannot consistently describe the process of creation and annihilation of fermions. This is also easily seen from the fact that the classical axial current of the Dirac field, given in the case under consideration by

$$j=\overline{\psi}\gamma_\alpha\gamma^5\psi\theta^a=-\overline{\psi}\gamma_a\psi\theta^\alpha=-\frac{C_v}{a^3}\theta^0, \tag{89}$$

is classically conserved, i.e., $d(*j)=0$, whereas the anomaly equation,[18,19,25,22]

$$d(*j)=-(1/8\pi^2)\mathrm{tr}_\delta(F\wedge F), \tag{90}$$

implies that the evolution of the YM field between topologically distinct vacua must result in a change of the fermion number,

$$N_L(t)=-\int_{\{t\}\times S^3}*j. \tag{91}$$

Thus, within our interpretation of the solutions, only the motion in the vicinity of a certain vacuum is allowed.

Finally, we comment on the case of gauge group $SU(4)$. Under assumption (69), we can completely solve the field equations. Equation (83) means that the Yang–Mills potential takes values only in an $su(2)\oplus su(2)$ subalgebra of $su(4)$ and, therefore, splits into two parts. Each part is coupled to one spinor field of a definite chirality. Thus, in some sense, we simply have a doubling of the solution for $SU(2)$.

Similar to (59), Eqs. (80) and (81) describe the exchange of energy between the spinor and gauge field. In the absence of the spinor field, the YM field has the following extrema:
a local maximum,

$$\xi=0, \quad \zeta=0, \quad q=1;$$

four minima,

$$\xi=1, \quad \zeta=0, \quad q=2,$$

$$\xi=0, \quad \zeta=\pm 1, \quad q=1,$$

$$\xi=-1, \quad \zeta=0, \quad q=0;$$

and four saddle points,

$$\xi=\tfrac{1}{2}, \quad \zeta=\pm\tfrac{1}{2}, \quad q=\tfrac{3}{2},$$

$$\xi=-\tfrac{1}{2}, \quad \zeta=\pm\tfrac{1}{2}, \quad q=\tfrac{1}{2},$$

the Chern–Simons number now being given by $q(t)=(\frac{1}{2})(2+3\xi(t)-(\xi(t))^3-3\xi(t)(\zeta(t))^2)$.

For sufficiently small values of the coupling constant [approximately, $g^2/(16\pi)<1$], the inclusion of the spinor field lifts the degeneracy so that three of the four absolute minima become local minima. Again, by fine tuning the constants E_1 and E_2 and the cosmological constant Λ we can obtain the corresponding static solutions, some of them being stable and some unstable.

If the constants E_1 and E_2 are such that the variables x and y describe a motion in the vicinity of the local minimum, where the energy of both left-handed and right-handed spinors is positive, the corresponding solutions are physically meaningful. Such solutions exist for any sign of the constant S (76), but, in contrast to the $SU(2)$ case, we have no chiral solutions.

If the energies E_1 and E_2 are large enough, in the case of gauge group $SU(4)$ we also have the phenomenon of energy level crossing in the evolution of the spinor field. But similar to the case of the gauge group $SU(2)$, these solutions become unphysical, because of negative energy of the Dirac field. Therefore, they are unable to describe transitions between the topologically distinct vacua, which the system has in this case.

In the end we would like to emphasize once again that the obtained solutions are interesting in their own right as the first example of exact self-consistent solutions in the Einstein–Yang–Mills–Dirac system. They can be of use for constructing cosmological solutions in EYMD theories, which would take into account the excited levels of the Dirac field as well. Their physical interpretation is not clear yet. For example, they can be considered as describing baby universes,[26] but it is quite possible that there can exist other interpretations of these solutions, which differ from the one proposed here.

ACKNOWLEDGMENTS

The authors are grateful to Yu. Kubyshin for fruitful discussions. One of the authors (I.V.) is grateful to the Center of Natural Sciences of the University of Leipzig for the warm hospitality extended to him during his stay in Leipzig.
Igor Volobuev also acknowledges partial support under the INTAS-93-1630-EXT project.

[1] R. Bartnik and J. McKinnon, Phys. Rev. Lett. **61**, 141 (1988); P. Bizon, *ibid.* **64**, 2844 (1990).
[2] D. V. Gal'tsov and M. S. Volkov, Phys. Lett. B **274**, 173 (1991).
[3] G. Rudolph, T. Tok, and I. P. Volobuev, Int. J. Mod. Phys. A **11**, 4999 (1996).
[4] O. Bertolami, J. M. Mourão, R. F. Picken, and I. P. Volobuev, Int. J. Mod. Phys. A **6**, 4149 (1991).
[5] O. Bertolami and J. M. Mourão, Class. Quantum Grav. **8**, 1271 (1991).
[6] M. Henneaux, Phys. Rev. D **21**, 857 (1980).
[7] R. T. Jantzen, J. Math. Phys. **23**, 1137 (1982).
[8] N. S. Manton, Nucl. Phys. B **193**, 502 (1981).
[9] Yu. A. Kubyshin, J. M. Mourão, G. Rudolph, and I. P. Volobuev, *Dimensional Reduction of Gauge Theories, Spontaneous Compactification and Model Building*, Springer Lecture Notes in Physics 349, 1989.
[10] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, Oxford, 1975).
[11] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. 1.
[12] J. Harnad, S. Shnider, and L. Vinet, J. Math. Phys. **21**, 2719 (1980).
[13] G. Rudolph, Lett. Math. Phys. **14**, 133 (1987).
[14] E. B. Dynkin, Trudy Mosk. Matem. Obshchestva **1**, 39 (1952).
[15] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1995), Formulas (3.145) and (3.147).
[16] S. J. Rey, Nucl. Phys. B **336**, 146 (1990).
[17] G. W. Gibbons and A. R. Steif, Phys. Lett. B **314**, 13 (1993); **320**, 245 (1994).
[18] V. A. Rubakov, Rep. Prog. Phys. **51**, 189 (1988).
[19] M. E. Shaposhnikov, Nucl. Phys. B **299**, 797 (1988).
[20] N. Christ, Phys. Rev. D **21**, 1591 (1980).
[21] M. Hellmund and J. Kripfganz, Nucl. Phys. B **373**, 749 (1992).
[22] V. A. Rubakov and M. E. Shaposhnikov, preprint CERN-TH-96-13, March 1996; Phys. Usp. **39**, 461 (1996).
[23] M. S. Volkov, O. Brodbeck, G. Lavrelashvili, and N. Straumann, Phys. Lett. B **349**, 438 (1995).
[24] M. S. Volkov, hep-th/9401071, 1994; Phys. Lett. B **328**, 89 (1994); **334**, 40 (1994).
[25] A. Y. Morosov, Usp. Phys. Nauk **150**, 337 (1986).
[26] S. Hawking, Mod. Phys. Lett. A **5**, 453 (1990).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

The identities of the algebraic invariants of the four-dimensional Riemann tensor. III

G. E. Sneddon[a)]
School of Computer Science, Mathematics, and Physics, James Cook University, Townsville, 4811, Australia

(Received 6 March 1998; accepted for publication 13 July 1999)

This paper extends the investigation of the invariants of the Riemann tensor to include the invariants that are of odd degree in the trace-free Ricci tensor. It is shown that these invariants can be expressed in terms of 15 such invariants that are irreducible. As a consequence, it is possible to write down a complete set of invariants of the Riemann tensor. Several syzygies for these invariants have been found and examples of these are given. These syzygies suggest there may be several new syzygies of invariants with even degree in the trace-free Ricci tensor. A large number of these have also been found and are discussed in the paper.
 [S0022-2488(99)04110-9]

I. INTRODUCTION

As the title suggests, the aim of this paper is to extend the recent work of the author[1,2] on the invariants of the Riemann tensor. The main result is to include those invariants that are of odd degree in the trace-free Ricci tensor and thereby find a complete set of Riemann invariants.

An *invariant* of a set of tensors in n dimensions is a polynomial function of the components of those tensors that is invariant under some group of transformations of the tensors. It is known[3] that any invariant of a set of tensors can be expressed as a linear combination of complete contractions of products of those tensors (together with the metric or $\epsilon_{i_1 \cdots i_n}$ as appropriate). In our case, the components of the Riemann tensor transform under the proper Lorentz group. A *complete* set of invariants, $\{I_1,\ldots,I_m\}$, is one for which any other invariant I can be written as a polynomial function of $\{I_1,\ldots,I_m\}$, but no element of the set can be written as a polynomial function of the others. An *independent* set of invariants, $\{I_1,\ldots,I_m\}$, is one for which $f(I_1,\ldots,I_m)=0$ implies that f is the trivial function, $f\equiv 0$. An invariant, T, is *reducible* if it can be written as a polynomial whose arguments are invariants of lower degree than T. This shall be written as $T\approx 0$. Two invariants will be said to be *equivalent* if their difference is reducible.

In the past 50 years there have been several attempts to understand the invariants of the Riemann tensor and the relations between them. Initial attempts were directed towards finding a suitable set of independent invariants[4–8] and, more recently, the physical interpretation of some of the invariants.[9,10] Carminati and McLenaghan[9] pointed out that in many cases an independent set is not sufficient, and that it is important to have a complete set of invariants. Such a set can be used to find all other invariants in any algebraically special case, as well as in the general case. In this sense, a complete set will contain all possible information about the invariants.

However, a complete set of invariants typically will contain more elements than an independent set, and so there must be relationships (syzygies) between them. Carminati and McLenaghan, McIntosh and Zachary,[11] and Zachary and McIntosh[12] have made some preliminary investigations into these relationships in algebraically special cases. Sneddon[1,2] showed how the complex bivector representation, together with some results from invariant theory, could be used to obtain a large number of the identities connecting the invariants of the Riemann tensor. These papers shall be referred to as I and II, respectively. The bivector formalism used is that of Buchdahl.[13] (A

a)Electronic mail: Graeme.Sneddon@jcu.edu.au

summary is given in I.) In this formalism, the Riemann tensor is represented by two complex 3×3 matrices together with the Ricci scalar, R. The matrix, Ψ, that corresponds to the Weyl tensor, is symmetric and trace-free. The other matrix, Γ, corresponds to the trace-free part of the Ricci tensor and is Hermitian. Under a proper Lorentz transformation, the bivectors undergo a complex orthogonal transformation.[13]

It was observed in II that there are two subsets of the invariants for which the problem of finding a complete set simplifies somewhat. The first subset is the set of invariants in which Γ appears only in the combination $K=\Gamma\bar{\Gamma}$. The second is the set of invariants of even degree in Γ. For simplicity, it is convenient to refer to invariants that are of even degree in the trace-free Ricci tensor as ''even'' invariants, and those with odd degree as ''odd'' invariants.

Since the product of any two odd invariants will be an even invariant, the view was expressed in II that, in most cases, it may be sufficient to consider the even invariants only. However, there are a number of reasons why the odd invariants should also be investigated. First, various odd invariants have been used previously. The invariant $S^i{}_j S^j{}_k S^k{}_i$, where $S^i{}_j$ is the trace-free Ricci tensor, is the most elementary of these. This is proportional to the invariant r_2 of Carminati and McLenaghan. There is also the invariant, D^-D^+S, of Géhéniau and Debever[5] (as expressed by Campbell and Wainwright[14]). This is proportional to the odd invariant m_4 introduced by Carminati and McLenaghan. It is important to understand how these quantities relate to the even invariants described in II. Second, there is the hope that it may be possible to simplify the rather large set of invariants found in II if the odd invariants are included. Third, it may be possible to simplify the syzygies of the even invariants if they can be written in terms of the odd invariants. At the very least, it may be possible to gain further insight into the nature of these syzygies. Unfortunately, these hopes are realized to a limited extent only. While the inclusion of the odd invariants means that some even invariants are no longer needed in a complete set, there is a greater number of odd invariants that need to be added to the set. Also, there does not seem to be any great simplification of the syzygies of even invariants beyond the observation that some can be obtained from corresponding syzygies of odd invariants. However, one consequence of this work is that several more syzygies of even invariants have been found. Some of these have total degree that is less than the total degree of those in II and so, in this sense, are more fundamental.

The initial goal is to find a set, $\mathcal{K}_0$, of odd invariants which are irreducible and for which any other odd invariant can be expressed as a polynomial function of the elements of $\mathcal{K}_0$ together with the even invariants. Section II summarizes the existing results for even invariants. In Sec. III, the nature of the odd invariants is discussed and some preliminary results are obtained. It is shown that any odd invariant can be written as a linear function of a small number of odd invariants with the coefficients of these terms being even invariants. In Sec. IV, a suitable set of odd invariants is obtained. Since these invariants are irreducible, they will form the set $\mathcal{K}_0$. A complete set, $\mathcal{K}$, of invariants of Ψ and Γ can then be given. Section V describes some of the syzygies of the odd invariants. Many of these syzygies are obtained by the method of skew-symmetrizing over $n+1$ indices that was discussed in II. Others are obtained by a ''method of undetermined coefficients,'' a numerical procedure developed by Ouchterlony[15] for finding syzygies. In Sec. VI, it is noted that several of these syzygies can be used to obtain syzygies of even invariants. The method of undetermined coefficients is then used to find those syzygies of even invariants that have the lowest possible degree. Examples of some of the new syzygies are given in the appendices. They are also available in an electronic form from http://www.jcu.edu.au/~mages/.

Throughout, the notation used is that of I and II. Indices $i,j,\ldots$ will range from 1 to 4, while bivector indices $A,B,\ldots$ will range from 1 to 3. Square brackets will be used to denote the trace of a matrix. Thus $[AB]$ is the trace of the matrix AB. Details of the symbols used are in Appendix D.

II. THE EVEN INVARIANTS

For those invariants where Γ appears only in the combination $\Gamma\bar{\Gamma}$, the invariants are best expressed in terms of the matrices $A=\Psi$ and $B=\Gamma\bar{\Gamma}$. It was demonstrated in I that a complete set for these invariants is the set, $\mathcal{I}$, given by

TABLE I. This is the set $\mathcal{J}$. Any invariant of Ψ and Γ that is of even degree in Γ can be written as a polynomial function of these invariants.

Real invariants	Complex invariants
$[B]=[K]$	
$[B^2]=[K^2]$	$[A^2]=[\Psi^2]$
$\det B=\det K$	$\det A=\det \Psi$
	$[AB]=[\Psi K]=\overline{[C]}$
	$[AB^2]=[\Psi K^2]=\overline{[BC]}$
	$[A^2B]=[\Psi^2 K]=\overline{[D]}$
	$[A^2B^2]=[\Psi^2 K^2]=\overline{[BD]}$
$[AC]=[\Psi\Gamma\bar{\Psi}\bar{\Gamma}]$	$[AD]=[\Psi\Gamma\bar{\Psi}^2\bar{\Gamma}]=\overline{[A^2C]}$
$[ABC]=[\Psi K\Gamma\bar{\Psi}\bar{\Gamma}]$	
$[AB^2C]=[\Psi K^2\Gamma\bar{\Psi}\bar{\Gamma}]$	$[ABD]=[\Psi K\Gamma\bar{\Psi}^2\bar{\Gamma}]=\overline{[A^2BC]}$
$[A^2D]=[\Psi^2\Gamma\bar{\Psi}^2\bar{\Gamma}]$	
$[A^2BD]=[\Psi^2 K\Gamma\bar{\Psi}^2\bar{\Gamma}]$	$[AB^2D]=[\Psi K^2\Gamma\bar{\Psi}^2\bar{\Gamma}]=\overline{[A^2B^2C]}$
$[A^2B^2D]=[\Psi^2K^2\Gamma\bar{\Psi}^2\bar{\Gamma}]$	
$[A^2BACD]=[\Psi^2K\Psi\Gamma\bar{\Psi}\bar{K}\bar{\Psi}^2\bar{\Gamma}]$	

$$\begin{array}{cccc} & & [A^2] & [A^3] \\ [B] & [AB] & [A^2B] & \\ [B^2] & [AB^2] & [A^2B^2] & \\ [B^3]. & & & \end{array} \tag{1}$$

Since there will be 15 real invariants in $\mathcal{I}$, and there can only be 13 independent invariants of Ψ and Γ, there must be two identities connecting these invariants. In fact, $[A^2B^2]$ satisfies a cubic equation whose coefficients are functions of the other invariants in (1). This equation was given in I and II, but in an incorrect form.[16] The corrected version is given in Appendix A.

For the set of even invariants, the fact that Γ is Hermitian means that it transforms as $\Gamma \to S\Gamma\bar{S}^{-1}$ and this complicates the invariants and their relationships quite considerably. In II, it was claimed that the set, $\mathcal{J}$, given in Table I, is a complete set for these invariants. This table gives the invariants in terms of the symmetric matrices $C=\Gamma\Psi\bar{\Gamma}$ and $D=\Gamma\Psi^2\bar{\Gamma}$. Also, the invariants $\det A$ and $\det B$ have been used instead of $[A^3]$ and $[B^3]$. While it was not proved that every element of $\mathcal{J}$ is irreducible, it seemed unlikely that there would be any identity to express any of them in term of even invariants of lower degree. In fact, this is confirmed in Sec. VI.

Also in II, a large number of syzygies were found by making use of the fact that skew-symmetrizing over $n+1$ indices will give an expression which is zero. For example, for any 3×3 matrices, X, Y, and Z,

$$X^{[A}{}_A Y^B{}_B Z^C{}_C \delta^{D]}{}_E=0. \tag{2}$$

When expanded, this equation gives a matrix identity satisfied by 3×3 matrices. Several syzygies of even invariants were obtained by multiplying such matrix identities by C, D or CD and then taking the trace. These syzygies are linear in the invariants

$$\mathbf{u}=([AC],[ABC],[AB^2C])^T,$$

$$\mathbf{v}=([A^2C],[A^2BC],[A^2B^2C])^T,$$

$$\mathbf{w}=([A^2D],[A^2BD],[A^2B^2D])^T.$$

Other syzygies were found that were quadratic in these invariants.

III. SOME PRELIMINARY RESULTS

Any invariant of Ψ and Γ can be expressed in terms of fully contracted products of $\Psi^A{}_B$, $\Gamma^A{}_{\dot{B}}$, ϵ^{ABC} and their conjugates. Since dotted and undotted indices must both occur in pairs, each term in an invariant of odd degree in Γ must have an odd number of factors of ϵ^{ABC} and an odd number of factors of $\epsilon^{\dot{A}\dot{B}\dot{C}}$. Furthermore, since any two factors of ϵ^{ABC} (and any two of $\epsilon^{\dot{A}\dot{B}\dot{C}}$) can be written in terms of Kronecker delta functions, it will be sufficient to consider odd invariants that have one factor of ϵ^{ABC} and one factor of $\epsilon^{\dot{A}\dot{B}\dot{C}}$. Note that each of these factors will change sign under an orthogonal transformation whose determinant is -1, but the product, $\epsilon^{ABC}\epsilon^{\dot{A}\dot{B}\dot{C}}$, will transform as a tensor under any orthogonal transformation. The simplest odd invariant is $\det(\Gamma^A{}_{\dot{B}})=\frac{1}{6}\epsilon^{\dot{A}\dot{B}\dot{C}}\epsilon_{ABC}\Gamma^A{}_{\dot{A}}\Gamma^B{}_{\dot{B}}\Gamma^C{}_{\dot{C}}$. However, this quantity is complex unless α, the determinant of the metric of the space of bivectors, is real. For this reason, it is preferable to express the odd invariants in terms of $e_{ABC}=\sqrt{\alpha}\epsilon_{ABC}$. Then,

$$\Delta=\frac{1}{6}e^{\dot{A}\dot{B}\dot{C}}e_{ABC}\Gamma^A{}_{\dot{A}}\Gamma^B{}_{\dot{B}}\Gamma^C{}_{\dot{C}}=\sqrt{\frac{\alpha}{\bar{\alpha}}}\det(\Gamma^A{}_{\dot{B}}) \tag{3}$$

is real. This is the same quantity, Δ, that was introduced in I and it is proportional to the invariant, r_2, of Carminati and McLenaghan.

Any odd invariant will be a sum of terms, each of which is an even invariant (or a constant) multiplied by an invariant of the form

$$(L,M,N)=e^{\dot{A}\dot{B}\dot{C}}e_{ABC}L^A{}_{\dot{A}}M^B{}_{\dot{B}}N^C{}_{\dot{C}}, \tag{4}$$

where L, M, and N are each products of the matrices Ψ, $\bar{\Psi}$, Γ, and $\bar{\Gamma}$. Therefore, the elements of $\mathcal{K}_0$ can be assumed to have this form. The following theorem further restricts the possible form of the elements of $\mathcal{K}_0$. In the statement of the theorem, Ψ and Ψ^2 are treated as distinct quantities.

Theorem: *The elements of* $\mathcal{K}_0$ *can be chosen to have the form of either* (Γ,Γ,N), *where N has at most one factor of each of* Ψ, Ψ^2, $\bar{\Psi}$ *and* $\bar{\Psi}^2$, *or* (Γ,M,N), *where M and N between them have exactly one factor of each of* Ψ, Ψ^2, $\bar{\Psi}$, *and* $\bar{\Psi}^2$.

In order to prove this theorem, some intermediate results are needed.

Lemma 1: Any invariant of the form (L,M,N) is equivalent to an expression where the odd invariants all have the form (Γ,M,N).

Proof: It follows from the identity, $e^{\dot{A}\dot{B}\dot{C}}e_{[ABC}P^A{}_{D]}Q^D{}_{\dot{A}}M^B{}_{\dot{B}}N^C{}_{\dot{C}}=0$, that, if $L^A{}_{\dot{A}}=P^A{}_DQ^D{}_{\dot{A}}$,

$$\begin{aligned}(PQ,M,N)&=[P](Q,M,N)-(Q,PM,N)-(Q,M,PN)\\&\approx-(Q,PM,N)-(Q,M,PN).\end{aligned} \tag{5}$$

Similarly, if $L^A{}_{\dot{A}}=P^A{}_{\dot{D}}Q^{\dot{D}}{}_{\dot{A}}$, then $(PQ,M,N)\approx-(P,MQ,N)-(P,M,NQ)$. Therefore, if $L=P\Gamma Q$,

$$(P\Gamma Q,M,N)\approx(\Gamma,PMQ,N)+(\Gamma,PM,NQ)+(\Gamma,MQ,PN)+(\Gamma,M,PNQ),$$

and hence the result. □

Lemma 2: Any invariant of the type $(\Gamma,\Gamma M,\Gamma N)$ *can be factorized and so is reducible.*

Proof: In the expression, $e^{\dot{A}\dot{B}\dot{C}}e_{ABC}\Gamma^A{}_{\dot{A}}\Gamma^B{}_{\dot{D}}M^{\dot{D}}{}_{\dot{B}}\Gamma^C{}_{\dot{E}}N^{\dot{E}}{}_{\dot{C}}$, the factors $e_{ABC}\Gamma^A{}_{\dot{A}}\Gamma^B{}_{\dot{D}}\Gamma^C{}_{\dot{E}}$ can be replaced by $\Delta e_{\dot{A}\dot{D}\dot{E}}$ and it follows that $(\Gamma,\Gamma M,\Gamma N)=\Delta([M][N]-[MN])$. This result is a generalization of the factorization of the invariant D^-D^-S (of Géhéniau and Debever[5]) first noted by Carminati and McLenaghan and also discussed in I.

In I it is shown that, if any matrix product has two separate factors of Ψ, Eq. (2) can be used to combine these two factors into a single factor of Ψ^2. A similar result holds for (Γ,M,N).

Lemma 3: Any invariant, (Γ,M,N), *for which M* and *N* between them contain two separate factors of Ψ can be expressed in terms of invariants with the factors of Ψ combined into a single factor of Ψ^2.

Proof: If both factors of Ψ were in either *M* or *N*, Eq. (2) could be used to combine them into Ψ^2. Therefore, Ψ would need to be a factor of both *M* and *N*. If Ψ was not the first factor of either *M* or *N* then, by Lemma 2, (Γ,M,N) could be factorized.

If Ψ was the first factor of *M* but not of *N*, $(\Gamma,M,N)=(\Gamma,\Psi P,\Gamma Q)$ where *Q* has a factor of Ψ. Making use of the result following Eq. (5), $(\Gamma,\Psi P,\Gamma Q)\approx-(\Gamma Q,\Psi P,\Gamma)-(\Gamma,\Psi PQ,\Gamma)$, so $(\Gamma,\Psi P,\Gamma Q)\approx-\frac{1}{2}(\Gamma,\Psi PQ,\Gamma)$. Then ΨPQ can be written in terms of matrix products where the two factors of Ψ have been combined.

If Ψ was the first factor of both *M* and *N* then, making use of Eq. (5),

$$\begin{aligned}(\Gamma,M,N)&=(\Gamma,\Psi P,\Psi Q)\\&\approx-(\Gamma,\Psi^2P,Q)-(\Psi\Gamma,\Psi P,Q)\\&\approx-(\Gamma,\Psi^2P,Q)+(\Psi^2\Gamma,P,Q)+(\Psi\Gamma,P,\Psi Q)\\&\approx-(\Gamma,\Psi^2P,Q)+(\Psi^2\Gamma,P,Q)-(\Gamma,P,\Psi^2Q)-(\Gamma,\Psi P,\Psi Q),\end{aligned}$$

and, once again, (Γ,M,N) can be expressed in terms of invariants where the two factors of Ψ have been combined.

This result can be extended to factors of Ψ^2, $\overline{\Psi}$, and $\overline{\Psi}^2$. Therefore, the elements of $\mathcal{K}_0$ can be chosen so that *M* and *N* between them have at most one factor of Ψ, Ψ^2, $\overline{\Psi}$, and $\overline{\Psi}^2$.

Proof of Theorem: In order to obtain the main theorem it remains to show that, if *M* and *N* do not have exactly one factor of each of Ψ, Ψ^2, $\overline{\Psi}$, and $\overline{\Psi}^2$, then (Γ,M,N) can be reduced to (Γ,Γ,P).

If neither *M* or *N* contains a factor of $\overline{\Psi}$, for example, one of them (*M* say) must have Γ as its last factor. Then $(\Gamma,Q\Gamma,N)\approx-(Q\Gamma,\Gamma,N)-(\Gamma,\Gamma,QN)$. Therefore, $(\Gamma,Q\Gamma,N)\approx-\frac{1}{2}\times(\Gamma,\Gamma,QN)$. The result for Ψ, Ψ^2, and $\overline{\Psi}^2$ can be obtained similarly. □

IV. THE ODD INVARIANTS

It turns out that any invariants with a given degree in Γ, Ψ, and $\overline{\Psi}$ will be equivalent to each other to within a constant multiple. Thus it is sufficient to take one representative of each case. The odd invariants of degree three in Γ will be considered first. The invariant, Δ, with no factors of Ψ, etc. has already been defined in Eq. (3). Θ_{mn} will denote an invariant of degree three in Γ and degrees *m* and *n* in Ψ and $\overline{\Psi}$, respectively. It follows from Lemma 2 that these invariants must contain at least one factor of Ψ or Ψ^2 and one of $\overline{\Psi}$ or $\overline{\Psi}^2$. Thus the next simplest odd invariant after Δ is

$$\Theta_{11}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi\Gamma\overline{\Psi}).$$

Θ_{11} is real and is proportional to the invariant m_4 defined by Carminati and McLenaghan.[9] By multiplying out the expressions for Δ and Θ_{11}, it is straightforward to show that

$$\Delta\Theta_{11}=[AB^2C]-[B][ABC]-\tfrac{1}{2}([B^2]-[B]^2)[AC]. \tag{6}$$

This is actually a modified version of the equation for m_4 given in I. It also shows that, when the odd invariants are included, $[AB^2C]$ is reducible.

Other invariants of this type are

$$\Theta_{21}=\overline{\Theta}_{12}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2\Gamma\overline{\Psi}),$$

$$\Theta_{22}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2\Gamma\overline{\Psi}^2),$$

$$\Theta_{33}=\tfrac{1}{2}(\Gamma,\Psi\Gamma\overline{\Psi},\Psi^2\Gamma\overline{\Psi}^2).$$

Note that there are no invariants with degree three in Γ that have degree three in either Ψ or $\overline{\Psi}$, but not both. Equations similar to Eq. (6) for $\Delta\Theta_{mn}$ can also be found using straightforward index manipulations, except for the equation for $\Delta\Theta_{33}$. This equation is longer than the others and the manipulations needed would be quite complicated. It was actually obtained by the method of undetermined coefficients developed by Ouchterlony[15] and described in the next section. The equation is given in the Appendix. The equations for $\Delta\Theta_{mn}$ show that $[AB^2C],[A^2B^2C]$, $[A^2B^2D]$, and $[A^2BACD]$ are reducible and so would no longer be in the complete set. They would be replaced by Θ_{11}, Θ_{12}, Θ_{22}, and Θ_{33}. Similarly, $\det B$ would be replaced by Δ. Since the product of any two odd invariants must be of degree six in Γ, it is clear that these are the only elements of $\mathcal{J}$ that are reducible in this way. The invariants, Θ_{mn}, are irreducible and so must be in the set $\mathcal{K}_0$. They would also be contained in a complete set of invariants of the Riemann tensor.

The idea that some elements of $\mathcal{J}$ can simply be replaced by an element of $\mathcal{K}_0$ to form a complete set for all invariants is rather appealing. Unfortunately, there are several other odd invariants that are irreducible that also need to be included in a complete set. These are invariants of degree five in Γ and can be defined as follows:

$$\Lambda_{11}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi K\Gamma\overline{\Psi}),$$

$$\Lambda_{21}=\overline{\Lambda}_{12}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2K\Gamma\overline{\Psi}),$$

$$\Lambda_{22}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2K\Gamma\overline{\Psi}^2),$$

$$\Lambda_{31}=\overline{\Lambda}_{13}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2K\Psi\Gamma\overline{\Psi}),$$

$$\Lambda_{32}=\overline{\Lambda}_{23}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2K\Psi\Gamma\overline{\Psi}),$$

$$\Lambda_{33}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi\Gamma\overline{\Psi}^2\overline{\Gamma}\Psi^2\Gamma\overline{\Psi}).$$

Note that there are several possible forms for the invariant of degree three in Ψ and $\overline{\Psi}$. For example, $(\Gamma,K\Psi\Gamma\overline{\Psi},\Psi^2\Gamma\overline{\Psi}^2)$ and $(\Gamma,\Psi K\Gamma\overline{\Psi},\Psi^2\Gamma\overline{\Psi}^2)$ are two possibilities. The first is equivalent to $-\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2\Gamma\overline{\Psi}^2\overline{\Gamma\Psi}\Gamma\overline{\Psi})$. The second is equivalent to $(\Psi\Gamma\overline{\Psi},K\Gamma,\Psi^2\Gamma\overline{\Psi}^2)$ which, in turn, is equivalent to invariants like the first. In all cases, the invariant can be put into the form (Γ,Γ,N). Different orderings of the factors in N result in invariants that are equivalent to Λ_{33} above.

It can be shown that none of these invariants is reducible. One way to see this is as follows. If Λ_{11} is reducible, it must be a linear combination of odd invariants of lower degree. The only possibility would be $\Lambda_{11}=E_1\Theta_{11}+E_2\Delta$, where E_1 and E_2 are even invariants of the appropriate degree. If this equation is multiplied by Δ, it should be possible to identify E_1 and E_2 so that either the equation is identically satisfied, or the result is syzygy of even invariants. Neither of these situations is the case, so Λ_{11} is irreducible. This method can be extended to the other invariants, Λ_{mn}.

Invariants of degree greater than five in Γ can be shown to be reducible. For example, for the invariant, $(\Gamma,\Gamma,\Psi K^2\Gamma\overline{\Psi})$, $\Psi K^2\Gamma\overline{\Psi}\approx -K\Psi K\Gamma\overline{\Psi}-K^2\Psi K\Gamma\overline{\Psi}$. Therefore $(\Gamma,\Gamma,\Psi K^2\Gamma\overline{\Psi})$ can be reduced to invariants for which the first factor of each matrix is Γ. These can be factorized, so the invariant is reducible. Similar arguments can be applied to most other invariants of degree greater than 5 in Γ. The exception is those invariants of degree three in both Ψ and $\bar{\Psi}$ and degree 7 in Γ. It is not too difficult to show that each such invariant can be expressed in terms of invariants of the type (Γ,Γ,N). These invariants can then be reduced to one of the following:

$$\begin{gathered}(\Gamma,\Gamma,\Psi K\Gamma\bar{\Psi}^2\bar{\Gamma}\Psi^2\Gamma\bar{\Psi}),\\(\Gamma,\Gamma,\Psi\Gamma\bar{\Psi}^2\bar{\Gamma}K\Psi^2\Gamma\bar{\Psi}),\\(\Gamma,\Gamma,\Psi\Gamma\bar{\Psi}^2\bar{\Gamma}\Psi^2K\Gamma\bar{\Psi}),\\(\Gamma,\Gamma,\Psi K\Psi^2\Gamma\bar{\Psi}^2\bar{K}\bar{\Psi}).\end{gathered}$$

To show that the first three are equivalent, use the expanded form of Eq. (2) and Lemma 2. Thus, the first few factors of N in the first invariant can be written as $(\Psi)(K)(\Gamma\bar{\Psi}^2\bar{\Gamma})$ which is equivalent to $-(\Psi)(\Gamma\bar{\Psi}^2\bar{\Gamma})(K)$ + ''terms whose first factor is Γ.'' Therefore, the first invariant is equivalent to a multiple of the second. Similarly, it is also equivalent to a multiple of the third. The equivalence to the final invariant follows from $\Psi(K)(\Gamma\bar{\Psi}^2\bar{\Gamma})(\Psi^2)\approx-\Psi(K)(\Psi^2)(\Gamma\bar{\Psi}^2\bar{\Gamma})-\Psi(\Gamma\bar{\Psi}^2\bar{\Gamma})(\Psi^2)(K)-\Psi(\Gamma\bar{\Psi}^2\bar{\Gamma})(K)(\Psi^2)$. There is a slight advantage in choosing

$$\tilde{\Lambda}_{33}=\tfrac{1}{2}(\Gamma,\Gamma,\Psi^2K\Psi\Gamma\bar{\Psi}\bar{K}\bar{\Psi}^2)$$

as the representative of these invariants. We should now investigate whether or not $\tilde{\Lambda}_{33}$ is reducible. In fact it is quite difficult (though presumably not impossible) to show this one way or the other by using the current techniques. However, the fact that it is reducible can be demonstrated by once again using the method of undetermined coefficients. This method, and the result for $\tilde{\Lambda}_{33}$, will be discussed in the next section.

Taking this work into account, it is clear that the elements of the set $\mathcal{K}_0$, are

$$\begin{array}{llll}\Delta & & & \\ & \Theta_{11},\Lambda_{11}, & \Theta_{12},\Lambda_{12}, & \Lambda_{13}\\ & \Theta_{21},\Lambda_{21}, & \Theta_{22},\Lambda_{22}, & \Lambda_{23}\\ & \Lambda_{31}, & \Lambda_{32}, & \Theta_{33},\Lambda_{33}.\end{array}\tag{7}$$

There are 15 real invariants and they would all need to be included in a complete set of the invariants of the Riemann tensor. On the other hand, some of the elements of $\mathcal{J}$ will not appear in the complete set. These are the six invariants

$$\begin{array}{llll}\det B & & & \\ & [AB^2C], & [AB^2D], & \\ & [A^2B^2C], & [A^2B^2D], & \\ & & & [A^2BACD].\end{array}\tag{8}$$

Therefore, including the odd invariants will increase the size of the complete set by 9. The complete set, $\mathcal{K}$, is given in Table II. If the Ricci scalar, R, is included, the set will be a complete set of Riemann invariants.

TABLE II. This is the set $\mathcal{K}$. Any invariant of Ψ and Γ can be written as a polynomial function of these invariants.

Real invariants	Complex invariants
$[B]=[K]$	
$[B^2]=[K^2]$	$[A^2]=[\Psi^2]$
$\Delta=\det\Gamma$	$\det A=\det\Psi$
	$[AB]=[\Psi K]=\overline{[C]}$
	$[AB^2]=[\Psi K^2]=\overline{[BC]}$
	$[A^2B]=[\Psi^2 K]=\overline{[D]}$
	$[A^2B^2]=[\Psi^2 K^2]=\overline{[BD]}$
$[AC]=[\Psi\Gamma\overline{\Psi\Gamma}]$	$[AD]=[\Psi\Gamma\bar{\Psi}^2\bar{\Gamma}]$
$[ABC]=[\Psi K\Gamma\bar{\Psi}\bar{\Gamma}]$	$=\overline{[A^2C]}$
$\Theta_{11}=\frac{1}{2}(\Gamma,\Gamma,\Psi\Gamma\bar{\Psi})$	$[ABD]=[\Psi K\Gamma\bar{\Psi}^2\bar{\Gamma}]$
$[A^2D]=[\Psi^2\Gamma\bar{\Psi}^2\bar{\Gamma}]$	$=\overline{[A^2BC]}$
$[A^2BD]=[\Psi^2 K\Gamma\bar{\Psi}^2\bar{\Gamma}]$	$\Theta_{21}=\frac{1}{2}(\Gamma,\Gamma,\Psi^2\Gamma\bar{\Psi})$
$\Theta_{22}=\frac{1}{2}(\Gamma,\Gamma,\Psi^2\Gamma\bar{\Psi}^2)$	$=\bar{\Theta}_{12}$
$\Theta_{33}=\frac{1}{2}(\Gamma,\Psi\Gamma\bar{\Psi},\Psi^2\Gamma\bar{\Psi}^2)$	
	$\Lambda_{21}=\frac{1}{2}(\Gamma,\Gamma,\Psi^2 K\Gamma\bar{\Psi})$
$\Lambda_{11}=\frac{1}{2}(\Gamma,\Gamma,\Psi K\Gamma\bar{\Psi})$	$=\bar{\Lambda}_{12}$
$\Lambda_{22}=\frac{1}{2}(\Gamma,\Gamma,\Psi^2 K\Gamma\bar{\Psi}^2)$	$\Lambda_{31}=\frac{1}{2}(\Gamma,\Gamma,\Psi^2 K\Psi\Gamma\bar{\Psi})$
$\Lambda_{33}=\frac{1}{2}(\Gamma,\Gamma,\Psi\Gamma\bar{\Psi}^2\bar{\Gamma}\Psi^2\Gamma\bar{\Psi})$	$=\bar{\Lambda}_{13}$
	$\Lambda_{32}=\frac{1}{2}(\Gamma,\Gamma,\Psi^2 K\Psi\Gamma\bar{\Psi}^2)$
	$=\bar{\Lambda}_{23}$

V. SYZYGIES OF THE ODD INVARIANTS

One identity that remains to be found for the work in Sec. IV is the one that gives an expression for $\tilde{\Lambda}_{33}$ in terms of the other invariants. The Second Fundamental Theorem states that it must be possible to obtain any such identity from the property that skew-symmetrizing over four indices must annihilate any tensor in three dimensions. Unfortunately, the details of how this might be achieved in this case remain unclear. Therefore, we resort to other means. The identity must have the form,

$$a_1\tilde{\Lambda}_{33}=\sum_{i=1}^{15} E_i O_i\,, \tag{9}$$

where the O_i are the elements of $\mathcal{K}_0$ and each of the E_i is an even invariant of the appropriate degree. For example, the coefficient of Θ_{22} in Eq. (9) will be of first degree in Ψ and $\bar{\Psi}$ and fourth degree in Γ (or second degree in K). It must have the form

$$a[B][AC]+b[ABC]+c[AB][C],$$

where a, b, and c are undetermined coefficients. Altogether there are 65 undetermined coefficients, a_i, (including a_1) in Eq. (9).

Ouchterlony[15] has shown that if the form of the syzygy is known as in this case, the coefficients can often be determined by a numerical procedure. Using the procedure, he independently found the syzygy for nonsymmetric 3×3 matrices [Eq. (6) of II] as well as several syzygies for sets of 2×2 matrices. The coefficients, a_i, have the property that they will satisfy Eq. (9) for any choice of the components of Ψ and Γ. Hence, each numerical choice for these components will give a homogeneous equation for the a_i. A set of 65 or more numerical values of Ψ and Γ chosen at random should give a complete set of equations for the a_i. If the coefficient matrix of this set of homogeneous equations is M, the null space of M will contain the possible solutions for a_i. If the values of the components and the subsequent calculations have finite precision, these solutions and the resulting syzygies will only be approximate. The exact, rational coefficients can be found by taking linear combinations so that the first nonzero component of each vector is 1 and, if this is the ith component, $a_i = 0$ for each of the other vectors. Once this is done, the remaining components will be close to rational numbers, and it should be possible to identify the correct values. Finally, the syzygies can often be simplified further by taking appropriate linear combinations. This method is used here to obtain some of the syzygies of odd invariants, and is used in Sec. VI to find some more syzygies of even invariants.

The application of the method to find the syzygies (9) is straightforward and it turns out that there are six syzygies of this type. One of these [Eq. (B1)] is an equation for $\tilde{\Lambda}_{33}$ so this invariant is reducible. The other five express homogeneous linear relations between the elements of $\mathcal{K}_0$, where the coefficients are even invariants. One linear combination of these syzygies does not contain any of the Λ_{mn}. This is also given in the Appendix. The same method was used to search for any identities of lower degree that were linear in the odd invariants, but none were found. This confirms that the invariants Θ_{mn} and Λ_{mn} are irreducible. It is possible there will be syzygies of higher degree, but no attempt has been made to find these.

There are also syzygies that are quadratic in the odd invariants. The equations for $\Delta\Theta_{mn}$ mentioned in the previous section are examples of these. However, these identities can be solved for $[AB^2C]$, etc. and so are best seen as ''reducing equations'' for these invariants rather than as syzygies of invariants of a complete set. Expressions can also be found for $\Delta\Lambda_{mn}$. The expression for $\Delta\Lambda_{11}$ is given in Appendix B. There will be similar equations to express the product of any two odd invariants in terms of even invariants. These equations can be found by writing the product of the two odd invariants in index notation and then expressing $e^{ABC}e_{DEF}$ and $e^{\dot{A}\dot{B}\dot{C}}e_{\dot{D}\dot{E}\dot{F}}$ in terms of Kronecker delta functions. They can also be found by using the numerical procedure described above. The equation for Θ_{11}^2 is in Appendix B.

VI. SYZYGIES OF THE EVEN INVARIANTS

Now the syzygies of odd invariants in Sec. V can be used to obtain some new syzygies of even invariants. As in II, each identity will be characterized by its degree in Ψ, K, and $\bar{\Psi}$, even though some identities may have a fractional degree in K. A syzygy whose degrees in Ψ, K, and $\bar{\Psi}$ are l, m, and n, respectively will be said to have degree (l,m,n). If the five syzygies that are linear in the odd invariants are multiplied by Δ, and the expressions for $\Delta\Theta_{mn}$ and $\Delta\Lambda_{mn}$ are used, the result is some syzygies of even invariants that have degree (3,5,3). A check with the degrees of the identities found in II shows that these must be new identities. The appearance of these new identities suggests that there may be more. These could be found by an exhaustive use of the property that skew-symmetrizing over $n+1$ indices gives an expression that is zero. However, once again, it is simplest to use the method of undetermined coefficients to obtain the required relations. The idea is, for any given degree, construct all products, E_i, of invariants in $\mathcal{J}$ that have that degree and use the method to find all sets of coefficients a_i such that

$$\sum a_i E_i = 0$$

for any numerical choice of Ψ and Γ.

The method of construction of the identities in II indicates that all syzygies of first degree in $\bar{\Psi}$ have already have been found. Consequently, all syzygies of first degree in Ψ will also be known. The next group to look at is those with degree $(2,m,2)$. The first one to be found had degree (2,4,2). Of all the syzygies of invariants of even degree, this one has the lowest degree. As the degree of Ψ, K, or $\bar{\Psi}$ is increased, the number of syzygies also increases. Incidently, this procedure also showed that there were no syzygies of degree (3,3,3). This fact rules out the possibility that the invariant, $[A^2BACD]$, can be expressed as a polynomial function of even invariants of lower degree. In fact, all the elements of $\mathcal{J}$ can similarly be shown to be irreducible (in terms of even invariants), thereby justifying the claim that $\mathcal{J}$ is a complete set for the even invariants Ψ and Γ.

The degrees of the syzygies found, and the number of new independent syzygies of each degree are

$$\begin{array}{ccc} (2,4,2) & (3,4,2) & (3,4,3) \\ 1 \text{ syzygy} & 1 \text{ syzygy} & 2 \text{ syzygies} \\ (2,5,2) & (3,5,2) & (3,5,3) \\ 1 \text{ syzygy} & 2 \text{ syzygies} & 5 \text{ syzygies} \\ (2,6,2) & (3,6,2) & (3,6,3) \\ 2 \text{ syzygies} & 3 \text{ syzygies} & 3 \text{ syzygies} \end{array}.$$

The complex conjugates of syzygies of degree $(3,m,2)$ are syzygies of degree $(2,m,3)$. The rest are real or, in some cases, purely imaginary. Also, the five syzygies of degree (3,5,3) are simply independent linear combinations of those obtained from the syzygies that are linear in the odd invariants. While this list is exhaustive for the given degrees, there may be syzygies of still higher degree. Of course, syzygies of a given degree can include multiples of syzygies of lower degree. There were 28 syzygies of degree (3,6,3). However, 21 of these could be obtained as multiples of other syzygies, leaving 7 new syzygies.

Since there is only one syzygy of degree (2,4,2), it can be labeled by T^{242}. The terms that are quadratic in **u**, **v**, and **w** are given by

$$T^{242}=\mathbf{u}^T F_1 \mathbf{u}+\cdots, \tag{10}$$

where

$$F_1=\begin{pmatrix} -\frac{1}{2}[B]^2+\frac{1}{2}[B^2], & [B], & -1 \\ [B], & -1, & 0 \\ -1, & 0, & 0 \end{pmatrix}.$$

The syzygy of degree (3,4,2) has the form

$$T^{342}=\mathbf{v}^T F_1 \mathbf{u}+\cdots. \tag{11}$$

There are two syzygies of degree (3,4,3). These can be arbitrarily labeled T^{343a} and T^{343b}. They are given by

$$T^{343a}=\overline{\mathbf{v}}^T F_1 \mathbf{v}+\mathbf{w}^T F_1 \mathbf{u}+\cdots,$$

$$T^{343b}=\overline{\mathbf{v}}^T F_2 \mathbf{v}-\mathbf{w}^T F_2 \mathbf{u}-4[B][A^2BACD]+\cdots, \tag{12}$$

where

$$F_2=\begin{pmatrix} [B]^2-5[B^2], & 2[B], & 6 \\ 2[B], & -18, & 0 \\ 6, & 0, & 0 \end{pmatrix},$$

and the term involving $[A^2BACD]$ has been included. By taking the appropriate linear combinations of all these syzygies, there does appear to be some common elements in some of them. This can be seen in those terms of the syzygies that are quadratic in **u**, **v**, and **w**. These terms are given in Appendix C. The full equation for T^{242} is also given in the Appendix. The complete set of equations is available from the web site, http://www.jcu.edu.au/˜mages/ in a form suitable for input into most computer algebra packages. Alternatively, they may be obtained directly from the author.

VII. CONCLUSION

This work substantially completes the program, begun in Ref. 1, of finding a complete set of invariants of the Riemann tensor. The complete set consists of the elements of $\mathcal{K}$ together with the Ricci scalar, R, a total of 38 real invariants. Incorporating the odd invariants has made the complete set more cumbersome, rather than enabling some simplification. Clearly, such a large complete set will be unwieldy in most situations. However, there are now a large number of syzygies available that can be used to express the elements of this set in terms of some smaller set. The possibilities seem to separate into two cases, depending on whether Δ is zero or nonzero. In either case, the syzygies contain sufficient information to calculate all elements of the complete set from a relatively small number of them.

If Δ is known and $\Delta \neq 0$, the odd invariants (which have been the focus of this paper) can readily be obtained from the set of even invariants, $\mathcal{J}$. In this case, apart from Δ itself, the odd invariants may assume less importance. On the other hand, it was noted in II that the set $\mathcal{I}$, together with **u**, will be sufficient to determine all elements of $\mathcal{J}$ by the solution of linear equations only. (Indeed, it may be possible to achieve this starting with just $\mathcal{I}$ and $[AC]$.) Therefore, for most purposes, if $\Delta \neq 0$, it will be sufficient to have a knowledge of the set $\mathcal{I}$ (with $\det B$ replaced by Δ) and the invariants **u** and R.

There are several Segre-types of the Ricci tensor for which $\Delta = 0$ and so this case should not be overlooked. It might be expected that the odd invariants are more important in this case. If Δ is equal to zero, the equations for $\Delta\Theta_{mn}$ and $\Delta\Lambda_{mn}$ show that there will be many more relationships between the elements of $\mathcal{J}$, and it may be possible to use these to determine $\mathcal{J}$ from a knowledge of $\mathcal{I}$ only. At the same time, it should also be possible to start with one of the odd invariants, Θ_{11} say, and use the equations for their products to calculate the remaining odd invariants. For the specific case of metrics generated by Maxwell fields this task is even easier. For these metrics, $\Gamma_{A\dot{B}} \propto \Phi_A \bar{\Phi}_{\dot{B}}$ and, of all the odd invariants defined, the only one that is nonzero is Θ_{33}.

There remains the task of relating algebraic properties of elements of the complete set to geometric and physical properties of space–time. It is expected that the syzygies found in this paper and the preceding two papers will play an important role in that work.

ACKNOWLEDGMENTS

I would like to thank Brian Edgar and Erik Ouchterlony for informing me of their work on using numerical methods to find syzygies, and for responding to my many queries on their method. I would also like to thank S. Bonanos for bringing the error in the cubic equation for $[A^2B^2]$ to my attention. I am grateful to Ian Whittingham who has, once again, made a number of helpful suggestions on the final manuscript.

APPENDIX A: SYZYGY FOR SYMMETRIC MATRICES

This is the corrected version of the syzygy connecting the elements of a complete set for two symmetric 3×3 matrices. If both matrices are trace-free,

$$c_3[A^2B^2]^3 + c_2[A^2B^2]^2 + c_1[A^2B^2] + c_0 = 0,$$

where

$$c_3=-4,$$

$$c_2=[AB]^2+5[A^2][B^2],$$

$$c_1=2[B^2][A^2B]^2+2[A^2][AB^2]^2-2[AB][AB^2][A^2B]-2[A^2][B^3][A^2B]$$
$$-2[A^3][B^2][AB^2]-[A^2][B^2][AB]^2+2[A^3][B^3][AB]-2[A^2]^2[B^2]^2,$$

$$c_0=-\tfrac{4}{3}([B^3][A^2B]^3+[A^3][AB^2]^3)+[A^2B]^2[AB^2]^2-\tfrac{1}{2}[A^2][B^2]([B^2][A^2B]^2+[A^2][AB^2]^2)$$
$$+2[A^3][B^3][A^2B][AB^2]-\tfrac{4}{9}[A^3][B^3][AB]^3+\tfrac{2}{3}[AB]^2([A^3][B^2][AB^2]+[A^2][B^3][A^2B])$$
$$+\tfrac{1}{4}[A^2]^2[B^2]^2[AB]^2-\tfrac{1}{3}[AB]([A^2]^2[B^3][AB^2]+[A^3][B^2]^2[A^2B])+\tfrac{2}{3}[A^2][B^2]([A^3][B^2]$$
$$\times[AB^2]+[A^2][B^3][A^2B])-\tfrac{2}{3}[A^2][A^3][B^2][B^3][AB]+\tfrac{1}{4}[A^2]^3[B^2]^3+\tfrac{1}{18}([A^3]^2[B^2]^3$$
$$+[A^2]^3[B^3]^2)-\tfrac{1}{3}[A^3]^2[B^3]^2.$$

APPENDIX B: SYZYGIES OF ODD INVARIANTS

This Appendix contains some of the syzygies of odd invariants. The first is the equation for $\tilde{\Lambda}_{33}$ in terms of the other invariants,

$$8\tilde{\Lambda}_{33}=\Delta(2[\overline{A^2}][A^2C][AB]-2\det\bar{A}[A^2][AB^2]-4[A^2][\overline{A^2}][ABC]-2[A^2B][\overline{A^2}][AC]$$
$$+2\det\bar{A}[A^2][AB][B]+4[A^2][\overline{A^2}][AC][B]+4\det A\det\bar{A}[B]^2-4\det A\det\bar{A}[B^2]$$
$$-2\det A[\overline{A^2}][BC]+2[A^2][AD][C]+2\det A[\overline{A^2}][B][C]-2[A^2][AC][D])$$
$$+(4[A^2B^2][\overline{A^2}]-16[A^2BD]-4[A^2B][\overline{A^2}][B]+8[A^2D][B]+[A^2][\overline{A^2}][B]^2$$
$$-[A^2][\overline{A^2}][B^2]+4[A^2][BD]+8[A^2B][D]-4[A^2][B][D])\Theta_{11}+(8[ABC]$$
$$-8[AC][B])\Theta_{22}+8[AC]\Lambda_{22}. \tag{B1}$$

There are five other syzygies of the same degree that are linear in the odd invariants. One of these is real, and does not involve any of the Λ_{mn},

$$0=\Delta(2[\overline{A^2}][A^2C][AB]-4[A^2][\overline{A^2}][ABC]-2[A^2B][\overline{A^2}][AC]+6[A^2D][AC]-6[A^2C][AD]$$
$$+2[A^2][\overline{A^2}][AC][B]+3\det A\det\bar{A}[B]^2-3\det A\det\bar{A}[B^2]+2[A^2][\overline{A^2}][AB][C]$$
$$+2[A^2][AD][C]-2[A^2][AC][D])+(-4[B]^2+4[B^2])\Theta_{33}+(-4[A^2B^2][\overline{A^2}]$$
$$+12[A^2BD]+4[A^2B][\overline{A^2}][B]-4[A^2D][B]-[A^2][\overline{A^2}][B]^2+[A^2][\overline{A^2}][B^2]$$
$$-4[A^2][BD]-8[A^2B][D]+4[A^2][B][D])\Theta_{11}+(4[\overline{A^2}][AB^2]-12[ABD]$$
$$-4[\overline{A^2}][AB][B]+4[AD][B]+8[AB][D])\Theta_{21}+(-12[A^2BC]+4[A^2C][B]$$
$$+4[A^2][BC]+8[A^2B][C]-4[A^2][B][C])\Theta_{12}+(12[ABC]-4[AC][B]$$
$$-8[AB][C])\Theta_{22}. \tag{B2}$$

The following is the equation for $\Delta\Theta_{33}$. This equation shows that $[A^2BACD]$ is reducible. Also given is the equation for $\Delta\Lambda_{11}$. Other expressions for $\Delta\Lambda_{mn}$ are a bit longer, with the equation for $\Delta\Lambda_{33}$ having 61 terms,

$$\begin{aligned}8\Delta\Theta_{33}=&-8[A^2BACD]+12\det A\det\bar{A}\det B-2\det\bar{A}[A^2B^2][AB]-2\det\bar{A}[A^2B][AB^2]\\&+2[A^2][\overline{A^2}][AB^2C]+4[A^2D][ABC]-4[A^2C][ABD]+4[A^2BD][AC]\\&-4[A^2BC][AD]+4\det\bar{A}[A^2B][AB][B]-2[A^2][\overline{A^2}][ABC][B]-2[A^2D][AC][B]\\&+2[A^2C][AD][B]+[A^2][\overline{A^2}][AC][B]^2+\det A\det\bar{A}[B]^3-[A^2][\overline{A^2}][AC][B^2]\\&-\det A\det\bar{A}[B][B^2]+4[A^2B][AD][C]-2\det A[BD][C]+4[A^2C][AB][D]\\&-2\det A[BC][D]+4\det A[B][C][D].\end{aligned}\tag{B3}$$

$$\begin{aligned}4\Delta\Lambda_{11}=&-2\det B[AC]+4[AB^2C][B]-2[ABC][B]^2+[AC][B]^3-2[ABC][B^2]\\&-[AC][B][B^2]+2[AB^2][BC]-2[AB][B][BC]-2[AB^2][B][C]+[AB][B]^2[C]\\&+[AB][B^2][C].\end{aligned}\tag{B4}$$

The following is the equation for Θ_{11}^2:

$$\begin{aligned}8\Theta_{11}{}^2=&14\Delta^2[A^2][\overline{A^2}]+4[\overline{A^2}][AB][AB^2]-16[AB][ABD]-16[ABC][AC]+8[AB^2][AD]\\&+4[A^2B^2][\overline{A^2}][B]-16[A^2BD][B]-4[\overline{A^2}][AB]^2[B]+8[AC]^2[B]-2[A^2B][\overline{A^2}][B]^2\\&+12[A^2D][B]^2+[A^2][\overline{A^2}][B]^3-2[A^2B][\overline{A^2}][B^2]-4[A^2D][B^2]-[A^2][\overline{A^2}][B][B^2]\\&+8[A^2C][BC]+8[A^2B][BD]+4[A^2][B][BD]-16[A^2BC][C]+8[AB][AC][C]\\&+4[A^2][BC][C]+8[A^2B][C]^2-4[A^2][B][C]^2+8[A^2B^2][D]+8[AB]^2[D]\\&-8[A^2B][B][D]-2[A^2][B]^2[D]-2[A^2][B^2][D]-48\Delta\Theta_{22}.\end{aligned}\tag{B5}$$

APPENDIX C: SYZYGIES OF EVEN INVARIANTS

The quadratic terms (as well as the coefficient of $[A^2BACD]$) of the remaining syzygies of even invariants are given below,

$$\begin{aligned}T^{252}&=\mathbf{u}^TG_1\mathbf{u}+\cdots,\\T^{352a}&=\mathbf{u}^TG_1\mathbf{v}+\cdots,\\T^{352b}&=\mathbf{u}^TG_2\mathbf{v}+\cdots,\\T^{353a}&=\bar{\mathbf{v}}^TG_1\mathbf{v}+\mathbf{w}^TG_1\mathbf{u}+\cdots,\\T^{353b}&=\mathbf{w}^TG_2\mathbf{u}+\cdots,\\T^{353c}&=\bar{\mathbf{v}}^TG_2\mathbf{v}+\cdots,\\T^{353d}&=\bar{\mathbf{v}}^TG_3\mathbf{v}-\mathbf{w}^TG_3\mathbf{u}-2([B]^2+[B^2])[A^2BACD]+\cdots,\\T^{353e}&=\bar{\mathbf{v}}^TG_4\mathbf{v}-\mathbf{w}^TG_4\mathbf{u}-(2[B]^2-8[B^2])[A^2BACD]+\cdots,\end{aligned}\tag{C1}$$

where

$$G_1=\begin{pmatrix}\det B, & 0, & 0\\ 0, & -[B], & 1\\ 0, & 1, & 0\end{pmatrix},\quad G_2=\begin{pmatrix}0, & [B^2], & -[B]\\ -[B^2], & 0, & 3\\ [B], & -3, & 0\end{pmatrix},$$

$$G_3=\begin{pmatrix} -2[B][B^2], & 3[B]^2-2[B^2], & 3[B] \\ 3[B]^2-2[B^2], & -16[B], & 3 \\ 3[B], & 3, & 0 \end{pmatrix},$$

$$G_4=\begin{pmatrix} 3\det B, & 4[B^2]-2[B]^2, & -2[B] \\ 4[B^2]-2[B]^2, & 9[B], & -3 \\ -2[B], & -3, & 0 \end{pmatrix},$$

$$T^{262a}=\mathbf{u}^T H_1 \mathbf{u}+\cdots,$$

$$T^{262b}=\mathbf{u}^T H_2 \mathbf{u}+\cdots,$$

$$T^{362a}=\mathbf{u}^T H_1 \mathbf{v}+\cdots,$$

$$T^{362b}=\mathbf{u}^T H_2 \mathbf{v}+\cdots,$$

$$T^{362c}=\mathbf{u}^T H_3 \mathbf{v}+\cdots,$$

$$T^{363a}=-\overline{\mathbf{v}}^T(H_1+H_2)\mathbf{v}+(-32\det B+4[B]^3-8[B][B^2])[A^2BACD]+\cdots,$$

$$T^{363b}=\mathbf{w}^T(H_1+H_2)\mathbf{u}+(-32\det B+4[B]^3-8[B][B^2])[A^2BACD]+\cdots, \tag{C2}$$

$$T^{363c}=\overline{\mathbf{v}}^T H_1 \mathbf{v}+\mathbf{w}^T H_1 \mathbf{u}+\cdots,$$

$$T^{363d}=\overline{\mathbf{v}}^T H_3 \mathbf{v}+\cdots,$$

$$T^{363e}=\mathbf{w}^T H_3 \mathbf{u}+\cdots,$$

$$T^{363f}=\overline{\mathbf{v}}^T H_4 \mathbf{v}-\mathbf{w}^T H_4 \mathbf{u}+(-24\det B+4[B]^3-8[B][B^2])[A^2BACD]+\cdots,$$

$$T^{363g}=\overline{\mathbf{v}}^T H_5 \mathbf{v}-\mathbf{w}^T H_5 \mathbf{u}+(42\det B-5[B]^3+9[B][B^2])[A^2BACD]+\cdots,$$

where

$$H_1=\begin{pmatrix} 0, & 2\det B, & 0 \\ 2\det B, & -[B]^2+[B^2], & 0 \\ 0, & 0, & 2 \end{pmatrix},$$

$$H_2=\begin{pmatrix} 0, & -[B]^3+[B][B^2], & 2[B]^2-2[B^2] \\ -[B]^3+[B][B^2], & 7[B]^2+[B^2], & -12[B] \\ 2[B]^2-2[B^2], & -12[B], & 18 \end{pmatrix},$$

$$H_3=\begin{pmatrix} 0, & 6\det B-[B]^3-[B][B^2], & 4[B]^2-2[B^2] \\ -6\det B+[B]^3+[B][B^2], & 0, & -10[B] \\ -4[B]^2+2[B^2], & 10[B], & 0 \end{pmatrix},$$

$$H_4=\begin{pmatrix} 0, & -[B]^3+3[B][B]^2, & -2[B^2] \\ -[B]^3+3[B][B]^2, & 5[B]^2-9[B^2], & -2[B] \\ -2[B^2], & -2[B], & 6 \end{pmatrix},$$

$$H_5=\begin{pmatrix} 0, & -[B][B]^2, & 2[B]^2-[B^2] \\ -[B][B]^2, & [B]^2+3[B^2], & -7[B] \\ 2[B]^2-[B^2], & -7[B], & 12 \end{pmatrix}.$$

$T^{242}=0$ is the syzygy of lowest degree. It is given by

$$\begin{aligned} 8T^{242}=&-4\det B[A^2B][\overline{A^2}]+24\det B[A^2D]+4[\overline{A^2}][AB^2]^2-16[AB][AB^2D]-8[ABC]^2\\ &-16[AB^2][ABD]-16[AB^2C][AC]-16[A^2B^2D][B]+6\det B[A^2][\overline{A^2}][B]\\ &-4[\overline{A^2}][AB][AB^2][B]+16[AB][ABD][B]+16[ABC][AC][B]+8[AB^2][AD][B]\\ &-4[A^2B^2][\overline{A^2}][B]^2+16[A^2BD][B]^2+2[\overline{A^2}][AB]^2[B]^2-4[AC]^2[B]^2\\ &-8[AB][AD][B]^2+4[A^2B][\overline{A^2}][B]^3-8[A^2D][B]^3-[A^2][\overline{A^2}][B]^4\\ &+8[A^2B^2][\overline{A^2}][B^2]-16[A^2BD][B^2]-2[\overline{A^2}][AB]^2[B^2]+4[AC]^2[B^2]\\ &+8[AB][AD][B^2]-8[A^2B][\overline{A^2}][B][B^2]+16[A^2D][B][B^2]+3[A^2][\overline{A^2}][B]^2[B^2]\\ &-2[A^2][\overline{A^2}][B^2]^2-16[A^2BC][BC]+8[AB][AC][BC]+8[A^2C][B][BC]\\ &+4[A^2][BC]^2-16[A^2B^2][BD]+8[AB]^2[BD]+16[A^2B][B][BD]-4[A^2][B]^2[BD]\\ &+8[A^2][B^2][BD]-16[A^2B^2C][C]+8[AB^2][AC][C]+16[A^2BC][B][C]\\ &-8[AB][AC][B][C]-8[A^2C][B]^2[C]+8[A^2C][B^2][C]+8[A^2B][BC][C]\\ &-4[A^2][B][BC][C]+8[A^2B^2][C]^2-8[A^2B][B][C]^2+2[A^2][B]^2[C]^2\\ &-2[A^2][B^2][C]^2-4\det B[A^2][D]+8[AB][AB^2][D]+16[A^2B^2][B][D]\\ &-8[AB]^2[B][D]-16[A^2B][B]^2[D]+4[A^2][B]^3[D]+8[A^2B][B^2][D]\\ &-8[A^2][B][B^2][D]. \end{aligned} \tag{C3}$$

APPENDIX D: NOTATION

TABLE III. This is a list of the main symbols used in this paper.

Ψ, Γ	3×3 complex matrices that represent the Weyl tensor and the trace-free Ricci tensor
A	Ψ
$B=K$	$\Gamma\bar{\Gamma}$
C	$\Gamma\bar{\Psi}\bar{\Gamma}$
D	$\Gamma\bar{\Psi}^2\bar{\Gamma}$
$[A]$	Square brackets are used to denote the trace of a matrix
$\mathbf{u}$	$([AC],[ABC],[AB^2C])^T$
$\mathbf{v}$	$([A^2C],[A^2BC],[A^2B^2C])^T$
$\mathbf{w}$	$([A^2D],[A^2BD],[A^2B^2D])^T$
Δ	$\frac{1}{6}e^{\dot{A}\dot{B}\dot{C}}e_{ABC}\Gamma^{A}{}_{\dot{A}}\Gamma^{B}{}_{\dot{B}}\Gamma^{C}{}_{\dot{C}}$ $=\sqrt{\alpha/\bar{\alpha}}\,\det(\Gamma^{A}{}_{\dot{B}})$
(L,M,N)	$e^{\dot{A}\dot{B}\dot{C}}e_{ABC}L^{A}{}_{\dot{A}}M^{B}{}_{\dot{B}}N^{C}{}_{\dot{C}}$
$\mathcal{I}$	A complete set for invariants of Ψ and K
$\mathcal{J}$	A complete set for even invariants of Ψ and Γ
$\mathcal{K}_0$	A set of irreducible odd invariants
$\mathcal{K}$	A complete set for invariants of Ψ and Γ

[1] G. E. Sneddon, ''The identities of the algebraic invariants of the four-dimensional Riemann tensor,'' J. Math. Phys. **37**, 1059–1075 (1996).
[2] G. E. Sneddon, ''The identities of the algebraic invariants of the four-dimensional Riemann tensor. II'' J. Math. Phys. **39**, 1659–1679 (1998).
[3] G. B. Gurevich, *Foundations of the Theory of Algebraic Invariants* (Noordhoff, Groningen, 1964).
[4] V. V. Narlikar and K. R. Karmarkar, ''The scalar invariants of a general gravitational metric,'' Proc. Ind. Acad. Sci. **29**, 91–97 (1948).
[5] J. Géhéniau and R. Debever, ''Les invariants de courboure de l'space de Riemann à quatre dimensions,'' Bull. Acad. Roy. Belg., Cl. des Sc. **XLII**, 114–123 (1956); J. Géhéniau, ''Les invariants de courboure des espaces Riemanniens de la relativité,'' *ibid.* **XLII**, 252–255 (1956); R. Debever, ''Étude géométrique du tenseur de Riemann–Christoffel des espaces de Riemann à quatre dimensions,'' *ibid.* **XLII**, 313–327 (1956); **XLII**, 608–621 (1956).
[6] P. J. Greenberg, ''The algebra of the Riemann curvature tensor in general relativity: Preliminaries,'' Stud. Appl. Math. **51**, 277–308 (1972).
[7] G. Sobczyk, ''Space–time algebra approach to curvature,'' J. Math. Phys. **22**, 333–342 (1981).
[8] G. E. Sneddon, ''On the algebraic invariants of the four-dimensional Riemann tensor,'' Class. Quantum Grav. **3**, 1031–1032 (1986).
[9] J. Carminati and R. G. McLenaghan, ''Algebraic invariants of the Riemann tensor in a four-dimensional Lorentzian space,'' J. Math. Phys. **32**, 3135–3140 (1991).
[10] B. M. Haddow, ''Geometric interpretation of the mixed invariants of the Riemann spinor,'' Gen. Relativ. Gravit. **28**, 481–509 (1996).
[11] C. B. G. McIntosh and E. Zakhary, ''Mixed invariants of the Riemann tensor,'' in *Relativity Today*, in *Proceedings of the Fourth Hungarian Relativity Workshop*, edited by R. P. Kerr and Z. Perjés (Akadémiai Kiadó, Budapest, 1994), pp. 3–12.
[12] E. Zakhary and C. B. G. McIntosh, ''A complete set of Riemann invariants,'' Gen. Relativ. Gravit. **29**, 539–581 (1997).
[13] H. A. Buchdal, ''On rotor calculus I,'' J. Aust. Math. Soc. B, Appl. Math. **6**, 402–423 (1966); ''On rotor calculus II,'' *ibid.* **6**, 424–448 (1966).
[14] S. J. Campbell and J. Wainwright, ''Algebraic computing and the Newman–Penrose formalism in general relativity,'' Gen. Relativ. Gravit. **8**, 987–1001 (1977).
[15] E. Ouchterlony, ''A systematic approach to matrix invariants,'' M.Sc. thesis, Linköping University, 1997.
[16] S. Bonanos, ''The algebra of two symmetric matrices: Proving completeness and deriving syzygies for a set of invariants of the Riemann tensor,'' J. Math. Phys. **40**, 2064–2073 (1999).

Symplecton for $\mathcal{U}_h(sl(2))$ and representations of $SL_h(2)$

N. Aizawa
Department of Applied Mathematics, Osaka Women's University, Sakai, Osaka 590-0035, Japan

(Received 22 March 1999; accepted for publication 14 June 1999)

Polynomials of boson creation and annihilation operators which form irreducible tensor operators for Jordanian quantum algebra $\mathcal{U}_h(sl(2))$, called h-symplecton, are introduced and their properties are investigated. It is shown that many properties of symplecton for Lie algebra $sl(2)$ are extended to h-symplecton. The h-symplecton is also a basis of irreducible representation of $SL_h(2)$ dual to $\mathcal{U}_h(sl(2))$. As an application of the procedure used to construct h-symplecton, we construct the representation bases of $SL_h(2)$ on the quantum h-plane. © *1999 American Institute of Physics.* [S0022-2488(99)01211-6]

I. INTRODUCTION

It is no doubt that well-developed representation theories are necessary when we apply algebraic objects to physics. The simplest examples in quantum physics are angular momentum algebra $su(2)$ and rotation matrices in three-dimensional space SO(3). Their algebraic structure is simple but contents of representation theories are quite rich.[1] To investigate these algebraic objects or their complexification could be a foundation for further investigation of higher dimensional objects.

As for deformation of Lie groups and Lie algebras, q-deformation of Lie algebra $sl(2)$ and Lie group $SL(2)$ (and their real form) is studied quite well. Their representation theories have attracted much interest in both physics and mathematics and give a way to higher dimensional cases.[2] There exists, however, some other deformation of Lie groups and algebras and these are generally called nonstandard deformation. The most studied one may be the so-called Jordanian deformation obtained by Drinfeld twist from a Lie algebra or a known quantum algebra. The simplest examples are, of course, the Jordanian deformation of Lie algebra $sl(2)$ and its dual. The Jordanian deformation of Lie group $SL(2)$, denoted by $SL_h(2)$, is studied in Refs. 3, 4, 5 and then Ohn introduced its dual algebra, namely, Jordanian deformation of $sl(2)$ denoted by $\mathcal{U}_h(sl(2))$.[6] The Jordanian quantum algebra $\mathcal{U}_h(sl(2))$ is more natural than the q-deformed $sl(2)$ in the sense that it is regarded as the angular momentum algebra with nonstandard coproduct (Sec. III) and we can use ordinary boson operators to represent $\mathcal{U}_h(sl(2))$, while it is hard to regard the q-deformed $sl(2)$ as angular momentum and q-deformed boson algebras are used for representations.[7] However, the representation theories of $\mathcal{U}_h(sl(2))$ and $SL_h(2)$ have not been developed yet. We do not know, for example, the Racha coefficients and matrix elements of the universal R-matrix for $\mathcal{U}_h(sl(2))$. As for $SL_h(2)$, even its representation matrices are not obtained.

In this article, in order to develop representation theories for Jordanian deformed algebras, we study symplecton for $\mathcal{U}_h(sl(2))$ and apply it to investigate representation matrices of $SL_h(2)$. The use of symplecton could be legitimated by recalling the properties of symplecton and q-deformed case. The symplecton, introduced by Biedenharn and Louck,[8,9] is a polynomial of boson creation and annihilation operators which form an irreducible tensor operator of $sl(2)$, that is, symplecton is a basis of irreducible representation (irrep.) for both $sl(2)$ and $SL(2)$. It is known that the symplecton is written in terms of Gauss hypergeometric function and product of two symplecton is reduced to a series of symplecton with Racha coefficients. In Ref. 8, application of symplecton to the Elliot model for nuclei is discussed, then it is found that Weyl-ordered polynomials for position and momentum operators are equivalent to symplecton.[10] Many properties of symplecton are inherited from $sl(2)$ to the q-deformed case.[2,11,12] The q-deformed symplecton, called

q-symplecton, is a irreducible tensor operator so that it is a irrep. basis for q-deformed $sl(2)$ and $SL(2)$. The q-symplecton is written in terms of q-hypergeometric function and product of two q-symplecton is reduced to a series of q-symplecton with q-Racha coefficients. q-Deformation of the Weyl-ordered polynomial[13] is formulated with q-symplecton. These facts show that symplecton is a powerful tool to investigate representation.

The plan of this article is as follows. Next three sections are mainly preparation for symplecton of $\mathcal{U}_h(sl(2))$. We often call the symplecton for $\mathcal{U}_h(sl(2))$ h-symplecton. The next section is a review of symplecton for $sl(2)$. Some of the properties of symplecton listed in Sec. II will be extended to h-symplecton. Section III is devoted to the Jordanian quantum algebra $\mathcal{U}_h(sl(2))$ and Jordanian quantum group $SL_h(2)$. We give new results on the twist element and Racha coefficients for $\mathcal{U}_h(sl(2))$. In Sec. IV, tensor operators for a Hopf algebra is introduced according to Ref. 14 and the relation between tensor operators for a Lie algebra and a Hopf algebra obtained by Drinfeld twist is discussed. Applying the result in Sec. IV, the h-symplecton is constructed from the $sl(2)$ symplecton in Sec. V. The properties of h-symplecton are studied in Secs. V and VI. We shall consider another irreducible tensor operators obtained from the quantum h-plane for $\mathcal{U}_h(sl(2))$ in Sec. VII and using these tensor operators, as well as h-symplecton, irreps. of $SL_h(2)$ are considered. Section VIII is concluding remarks.

II. SYMPLECTON FOR $sl(2)$

The symplecton realization of $sl(2)$ is said to be ''minimal,'' since only one kind of boson operator is used. It is in marked contrast to the well-known Jordan–Schwinger realization where two kinds of bosons are necessary. Let us first review the definition and important properties of the $sl(2)$ symplecton.[8,9]

Let $\bar{a},a$ be boson operators satisfying $[\bar{a},a]=1$, and define

$$J_+=-\tfrac{1}{2}a^2,\quad J_-=\tfrac{1}{2}\bar{a}^2,\quad J_0=\tfrac{1}{2}(a\bar{a}+\bar{a}a). \tag{II.1}$$

It is easy to verify that (II.1) satisfies the $sl(2)$ commutation relations

$$[J_0,J_\pm]=\pm 2J_\pm,\quad [J_+,J_-]=J_0. \tag{II.2}$$

The symplecton is a polynomial in $\bar{a}$ and a and form a irreducible tensor operator of $sl(2)$ belonging to the spin j representation $(j=\tfrac{1}{2},1,\tfrac{3}{2},\ldots)$. Namely the symplecton, denoted by $P_j^m(a,\bar{a})$, is defined by

$$\begin{aligned}[][J_\pm,P_j^m]&=\sqrt{(j\mp m)(j\pm m+1)}P_j^{m\pm 1},\\ [J_0,P_j^m]&=2mP_j^m.\end{aligned} \tag{II.3}$$

The basic idea of symplecton is to treat $\bar{a}$ and a in a symmetric way. To this end, the usual ''boson calculus'' is replaced with the so-called ''symplecton calculus,'' that is, instead of the boson vacuum $|0\rangle$ satisfying $\bar{a}|0\rangle=0$, the formal ket $|\,\rangle$ which is not annihilated by both $\bar{a}$ and a is introduced. The representation bases in the realization (II.1) are formed by letting P_j^m act on $|\,\rangle$, and the action of generators on the bases is defined by $J_\alpha|jm\rangle=[J_\alpha,P_j^m]|\,\rangle$. There exists an appropriate definition of an inner product for these $|jm\rangle$, so that we obtain the usual unitary representations of $sl(2)$ with spin j.

The explicit form of the polynomials $P_j^m(a,\bar{a})$ is found by solving $[J_+,P_j^j]=0$ to obtain $P_j^j=a^{2j}$, and then using the action of J_- to calculate P_j^m,

$$P_j^m(a,\bar{a})=\frac{1}{2^{j-m}}\left[\frac{(2j)!(j-m)!}{(j+m)!}\right]^{1/2}\sum_{s=0}^{j-m}\frac{\bar{a}^{j-m-s}a^{j+m}\bar{a}^s}{s!(j-m-s)!}. \tag{II.4}$$

An alternative form for P_j^m is obtained by starting with $P_j^{-j}=\bar{a}^{2j}$ and then using the action of J_+,

$$P_j^m(a,\overline{a})=\frac{1}{2^{j+m}}\left[\frac{(2j!)!(j+m)!}{(j-m)!}\right]^{1/2}\sum_{s=0}^{j+m}\frac{a^s\overline{a}^{j-m}a^{j+m-s}}{s!(j+m-s)!}. \tag{II.5}$$

We would like to list some properties of $sl(2)$ symplecton. For their proof or detail, we refer the reader to Refs. 8 and 9.

(1) A set of polynomials $\{P_j^m(a,\overline{a})|m=-j,-j+1,...,j\}$ forms representation bases for the Lie group $SL(2)$ as well as the Lie algebra $sl(2)$. The boson commutation relation is covariant under the action of $SL(2)$ defined by

$$(a',\overline{a}')=(a,\overline{a})\begin{pmatrix}x & u\\ v & y\end{pmatrix}, \tag{II.6}$$

where the 2×2 matrix is an element of $SL(2)$. The transformed polynomial $P_j^m(a',\overline{a}')$ is decomposed into $P_j^m(a,\overline{a})$ multiplied by polynomials in the entries of $SL(2)$ matrix,

$$P_j^m(a',\overline{a}')=\sum_n P_j^n(a,\overline{a})d_{nm}^j(g)\quad g\in SL(2). \tag{II.7}$$

The $(2j+1)\times(2j+1)$ matrix $d_{nm}^j(g)$ gives an irrep. of $SL(2)$ and is called Wigner's *d*-function in physics terminology.

(2) The polynomials $P_j^m(a,\overline{a})$ have a generating function. Let ξ, η be ordinary *c*-numbers commuting with $a,\overline{a}$. Then

$$(\xi a+\eta\overline{a})^{2j}=\sqrt{(2j)!}\sum_{m=-j}^{j}\Phi_{jm}(\xi,\eta)P_j^m(a,\overline{a}), \tag{II.8}$$

where Φ_{jm} are well-known representation bases of both $sl(2)$ and $SL(2)$,

$$\Phi_{jm}(\xi,\eta)=\frac{\xi^{j+m}\eta^{j-m}}{\sqrt{(j+m)!(j-m)!}}. \tag{II.9}$$

Irreps. of $sl(2)$ are constructed on (II.9) by the realization

$$J_+=\xi\frac{d}{d\eta},\ J_-=\eta\frac{d}{d\xi},\ J_0=\xi\frac{d}{d\xi}-\eta\frac{d}{d\eta}, \tag{II.10}$$

while irreps. of $SL(2)$ are obtained by the following transformation:

$$(\xi',\eta')=(\xi,\eta)\begin{pmatrix}x & u\\ v & y\end{pmatrix}, \tag{II.11}$$

it follows that

$$\Phi_j^m(\xi',\eta')=\sum_n \Phi_j^n(\xi,\eta)\,d_{nm}^{\,j}(g), \tag{II.12}$$

where we have obtained the same *d*-function as (II.7).

(3) The symplecton polynomials can be expressed in terms of Gauss hypergeometric function ${}_2F_1(a,b;c;z)$. The polynomial ${}_2F_1(a,b;c;z)$ is defined by

$${}_2F_1(a,b;c;z)=\sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{n!(c)_n}z^n, \tag{II.13}$$

where $(a)_n$ stands for the sifted factorial

$$(a)_n=\begin{cases}1 & n=0\\ a(a+1)\cdots(a+n-1) & n=1,2,...\end{cases}. \tag{II.14}$$

Now define the operator $N=a\overline{a}$, then the symplecton $P_j^m(a,\overline{a})$ is written in terms of ${}_2F_1(a,b;c;z)$ with $z=-1$ and the parameters *a*, *c* become functions of operator *N*. The expression (II.4) becomes

$$P_j^m=\frac{1}{2^{j+m}}\left[\frac{(2j)!}{(j+m)!(j-m)!}\right]^{1/2}\frac{(N+j-m)!}{(N-2m)!}\,{}_2F_1(-N+2m,-j+m;-N-j+m;-1)(\overline{a})^{-2m}. \tag{II.15}$$

In this way, properties of P_j^m are reduced to properties of the hypergeometric function. Especially, the equivalence of two form (II.4) and (II.5) is explained by the formula

$$ {}_2F_1(a,b;c;z)=(1-z)^{c-a-b}\,{}_2F_1(c-a,c-b;c;z). \tag{II.16}$$

(4) The polynomials $P_j^m(a,\overline{a})$ are transformed under the action $a\to\overline{a},\ \overline{a}\to -a$,

$$P_j^m(\overline{a},-a)=(-1)^{j-m}P_j^{-m}(a,\overline{a}). \tag{II.17}$$

To define an inner product for the bases $|jm\rangle = P_j^m|\ \rangle$, the property (II.17) and the product formula discussed below play a crucial role.

(5) Let P_j^m and $P_{j'}^{m'}$ be the symplecton polynomials, then they obey the product law

$$P_j^m P_{j'}^{m'}=\sum_{k=|j-j'|}^{j+j'} \langle k|j|j'\rangle C_{m',m,m+m'}^{j',j,k} P_k^{m+m'}, \tag{II.18}$$

where

$$\langle k|j|j'\rangle = 2^{k-j-j'}(2k+1)^{-1/2}\nabla(kjj'),$$

$$\nabla(abc)=\left[\frac{(a+b+c+1)!}{(a+b-c)(a-b+c)!(-a+b+c)!}\right]^{1/2}, \tag{II.19}$$

and $C_{m',m,m+m'}^{j',j,k}$ is the Clebsch–Gordan coefficient (CGC) for $sl(2)$. The associativity of the products $(P_a^\alpha P_b^\beta)P_c^\gamma = P_a^\alpha(P_b^\beta P_c^\gamma)$ gives a relation between ‘‘triangle functions,’’

$$\nabla(acf)\nabla(bdf)=(2f+1)\sum_e W(abcd;ef)\nabla(abe)\nabla(cde), \tag{II.20}$$

where $W(abcd;ef)$ is the Racha coefficient.

The inner product for $|jm\rangle$ is defined by

$$\langle jm|j'm'\rangle = \langle|(-1)^{j-m}P_j^{-m}\cdot P_{j'}^{m'}|\rangle, \tag{II.21}$$

and the operation $\langle|(\cdots)|\rangle$ means to take only the $j=0$ part of the expression $(\cdots)$. Applying the product law (II.18) to the RHS of (II.21), we see that that the $j=0$ part is given by the CGC $C_{m,m',0}^{j,j',0}$, so that the bases $|jm\rangle$ are orthonormal.

III. JORDANIAN DEFORMATION OF $sl(2)$ AND $SL(2)$

The Jordanian quantum algebras $\mathcal{U}_h(\mathbf{g})$ are obtained from the (universal enveloping algebra $\mathcal{U}(\mathbf{g})$ of) Lie algebras $\mathbf{g}$ from Drinfeld twist.[15] We denote the coproduct, conuit and antipode for $\mathcal{U}(\mathbf{g})$, when it is regarded as a Hopf algebra, by Δ, ϵ, S, respectively. With the invertible element $\mathcal{F}\in\mathcal{U}(\mathbf{g})\otimes\mathcal{U}(\mathbf{g})$ satisfying

$$(\epsilon\otimes id)(\mathcal{F})=(id\otimes\epsilon)(\mathcal{F})=1, \tag{III.1}$$

$$\mathcal{F}_{12}(\Delta\otimes id)(\mathcal{F})=\mathcal{F}_{23}(id\Delta)(\mathcal{F}), \tag{III.2}$$

the algebra $\mathcal{U}_h(\mathbf{g})$ is defined by the same commutation relations as $\mathbf{g}$ and the following Hopf algebra mappings:

$$\tilde{\Delta}=\mathcal{F}\Delta\mathcal{F}^{-1},\quad \tilde{\epsilon}=\epsilon,\quad \tilde{S}=uSu^{-1}, \tag{III.3}$$

where $u=m(id\otimes S)(\mathcal{F})$, $u^{-1}=m(S\otimes id)(\mathcal{F}^{-1})$, m denotes the usual product in $\mathbf{g}$. This is a triangular Hopf algebra whose universal R-matrix is given by $\mathcal{R}=\mathcal{F}_{21}\mathcal{F}^{-1}$.

For the case of $\mathbf{g}=sl(2)$, $\mathcal{F}$ is given by[16]

$$\mathcal{F}=\exp(-\tfrac{1}{2}J_0\otimes\sigma),\quad \sigma=-\ln(1-2hJ_+). \tag{III.4}$$

The twist element $\mathcal{F}$ used here gives different form of $\mathcal{U}_h(sl(2))$ from the one in Ref. 6. The relationship between these two forms is given in Appendix A. The explicit form of Hopf algebra mappings for $\mathcal{U}_h(sl(2))$ is summarized in Appendix B (some of them will be used in the later computation). An application of the $\mathcal{U}_h(sl(2))$ to the Heisenberg spin chain is found in Ref. 16. The finite dimensional highest weight irreps. for $\mathcal{U}_h(sl(2))$ are same as $sl(2)$, because of the same commutation relations. We shall use the following lemmas on tensor product representations in subsequent sections.

Lemma III.1 (Ref. 16): *Let V^{j_1}, V^{j_2} be the representation space with the highest weight j_1, j_2. Then the tensor product of them is completely reducible, i.e.,*

$$V^{j_1}\otimes V^{j_2}=\bigoplus_{j=|j_1-j_2|}^{j_1+j_2} V^j,$$

and the bases of V^j are given by

$$e_m^{(j_1j_2)j}=\sum C^{j_1,j_2,j}_{m_1,m_2,m}F^{j_1,j_2}_{k_1,k_2,m_1m_2}e^{j_1}_{k_1}\otimes e^{j_2}_{k_2}, \tag{III.5}$$

where $C^{j_1,j_2,j}_{m_1,m_2,m}$ is the CGC of $sl(2)$ and $F^{j_1,j_2}_{k_1,k_2,m_1,m_2}$ is the matrix element of $\mathcal{F}$ on $V^{j_1}\otimes V^{j_2}$. The explicit form of matrix elements $F^{j_1,j_2}_{k_1,k_2,m_1,m_2}$ is given in Appendix C. (It seems to be the first time to show the explicit form of $F^{j_1,j_2}_{k_1,k_2,m_1,m_2}$ in the literature, and this also gives the explicit form of the R-matrix for $\mathcal{U}_h(sl(2))$.)

Lemma III.2: The Racha coefficients for $sl(2)$ and $\mathcal{U}_h(sl(2))$ coincide.
Lemma III.2 is proved in Appendix D.

The matrix quantum group dual to $\mathcal{U}_h(sl(2))$ is called the Jordanian quantum group $SL_h(2)$. It is generated by four elements x, y, u and v subject to the relations[3,4,5]

$$[v,x]=hv^2,\quad [u,x]=h(1-x^2),$$
$$[v,y]=hv^2,\quad [u,y]=h(1-y^2), \tag{III.6}$$
$$[x,y]=h(xv-yv),\quad [v,u]=h(xv+vy).$$

It follows that the central element of $SL_h(2)$ which gives the determinant of the quantum matrix

$$T=\begin{pmatrix} x & u \\ v & y \end{pmatrix} \tag{III.7}$$

is defined by

$$\det T=xy-uv-hxv=1. \tag{III.8}$$

The $SL_h(2)$ has a Hopf algebra structure. The relations (III.6) and Hopf algebra mappings are summarized in the FRT-formalism[17] with the R-matrix

$$R=\begin{pmatrix} 1 & h & -h & h^2 \\ 0 & 1 & 0 & h \\ 0 & 0 & 1 & -h \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{III.9}$$

The coproduct, the counit, and the antipode are given by

$$\Delta(T)=T\otimes T,$$

$$\epsilon(T)=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},$$

$$S(T)=T^{-1}=\begin{pmatrix}y-hv & -u-h(y-x)+h^2v\\ -v & x+hv\end{pmatrix}.$$

Let us define the d-function for $SL_h(2)$ using the notion of comodule. A vector space M is called right $SL_h(2)$ comodule if there is a map $\rho: M\to M\otimes SL_h(2)$ such that the following relations are satisfied

$$(\rho\otimes id)\circ\rho=(id_M\otimes\Delta)\circ\rho, \quad (id_M\otimes\epsilon)\circ\rho=id_M, \tag{III.10}$$

where id_M stands for the identity map in M. Using bases e_i of M, the map ρ is written as

$$\rho(e_i)=\sum_j e_j\otimes\tilde{d}_{ji}, \tag{III.11}$$

it follows that the relations (III.10) are rewritten as

$$\Delta(\tilde{d}_{ij})=\sum_k \tilde{d}_{ik}\otimes\tilde{d}_{kj}, \quad \epsilon(\tilde{d}_{ij})=\delta_{ij}. \tag{III.12}$$

We call the $\tilde{d}_{ij}$ satisfying (III.11) and (III.12) the d-function for $SL_h(2)$. In the following sections, we deal with the case in which the vector space M has an algebraic structure. It is natural, in this case, to require that the map ρ should respect the extra structure on M.

IV. TENSOR OPERATORS AND TWIST

To define the symplecton for $\mathcal{U}_h(sl(2))$, it is necessary to extend the notion of tensor operators to Hopf algebra. This has been carried out by Rittenberg and Scheunert.[14] Tensor operators are defined for each realization of the Hopf algebra $\mathcal{H}$ under consideration. Assuming that we have a realization of $\mathcal{H}$, we first define the adjoint action.

Definition IV.1: Let W,W' be a representation space of $\mathcal{H}$, and let t be an operator which carries W into W'. Then the adjoint action of $X\in\mathcal{H}$ on t is defined by

$$\mathrm{ad}\,X(t)=m(id\otimes S)(\Delta(X)(t\otimes 1)). \tag{IV.1}$$

The adjoint action has two important properties

$$\mathrm{ad}\,XX'(t)=\mathrm{ad}\,X\circ\mathrm{ad}\,X'(t), \quad \mathrm{ad}\,X(t\otimes s)=\sum_i \mathrm{ad}\,X_i(t)\otimes\mathrm{ad}\,X_i'(s), \tag{IV.2}$$

where the coproduct for X is written as $\Delta(X)=\Sigma_i X_i\otimes X_i'$. From these properties, we see that the adjoint action gives a representation of $\mathcal{H}$,

$$\mathrm{ad}[X,X'](t)=[\mathrm{ad}\,X,\mathrm{ad}\,X'](t). \tag{IV.3}$$

Tensor operators for $\mathcal{H}$ are defined as operators which form representation bases of $\mathcal{H}$ under the adjoint action.

Definition IV.2: Let $D(X)$ be a representation matrix of $X\in\mathcal{H}$. The operators t_α are called the tensor operator, if they satisfy the relation

$$\mathrm{ad}\, X(t_\alpha)=\sum_\beta D(X)_{\beta\alpha}t_\beta. \tag{IV.4}$$

If the representation is irreducible, the tensor operators are called irreducible tensor operators.

The explicit form of the adjoint action for $\mathcal{U}_h(sl(2))$ reads

$$\begin{aligned}\mathrm{ad}\, J_0(t)&=[J_0,t]e^{-\sigma},\\ \mathrm{ad}\, J_+(t)&=e^{-\sigma}[J_+e^{\sigma},t],\\ \mathrm{ad}\, J_-(t)&=\left[J_-+hJ_0+\frac{h}{2}J_0^2,t\right]e^{-\sigma}-h[J_0,t]e^{-2\sigma}-\frac{h}{2}[J_0,[J_0,t]]e^{-2\sigma}.\end{aligned} \tag{IV.5}$$

Some examples of the $\mathcal{U}_h(sl(2))$ tensor operators are considered in Ref. 18 and they are applied to construct boson algebra which is covariant under the action of Jordanian matrix quantum groups.[19]

Since the coproduct for the Lie algebra and Jordanian quantum algebra is related via the twist element (III.3), tensor operators for these algebras are also related by twisting via $\mathcal{F}$.[20]

Lemma IV.1: Let t_α be tensor operators for the Lie algebra $\mathbf{g}$ *and $\tilde{t}_\alpha$ be corresponding ones for Jordanian quantum algebra $\mathcal{U}_h(\mathbf{g})$. Then these tensor operators are related via the twist element $\mathcal{F}$,*

$$\tilde{t}_\alpha=m(id\otimes\tilde{S})(\mathcal{F}(t_\alpha\otimes 1)\mathcal{F}^{-1}), \tag{IV.6}$$

$$t_\alpha=m(id\otimes S)(\mathcal{F}^{-1}(\tilde{t}_\alpha\otimes 1)\mathcal{F}). \tag{IV.7}$$

Proof: The first relation (IV.6) is derived in Ref. 20 (Proposition 3). The second one (IV.7) is its inverse. The expression used in Lemma IV.1 is different from Ref. 20, it may be good to show the second relation as an example of the proof. It is proved by showing the substitution of (IV.7) into (IV.6) gives the identity map.

Let us write the twist element and its inverse as

$$\mathcal{F}=\sum f^a\otimes f_a,\quad \mathcal{F}^{-1}=\sum g^a\otimes g_a,$$

then

$$u=\sum f^aS(f_a),\quad u^{-1}=\sum S(g^a)g_a,$$

and the relation (IV.6) becomes

$$\tilde{t}_\alpha=\sum f^at_\alpha g^b\tilde{S}(f_ag_b)=\sum f^at_\alpha g^buS(f_ag_b)u^{-1}=\sum f^at_\alpha S(f_a)u^{-1}, \tag{IV.8}$$

where we used

$$\sum g^buS(g_b)=\sum g^bf^aS(g_bf_a)=m(id\otimes S)(\mathcal{F}^{-1}\mathcal{F})=1.$$

On the other hand, the relation (IV.7) is rewritten

$$t_\alpha=\sum g^a\tilde{t}_\alpha f^bS(g_af_b)=\sum g^a\tilde{t}_\alpha uS(g_a). \tag{IV.9}$$

Substituting (IV.9) into (IV.8),

$$\tilde{t}_\alpha = \sum f^a g^b \tilde{t}_\alpha u S(f_a g_b) u^{-1} = \sum f^a g^b \tilde{t}_\alpha \tilde{S}(f_a g_b) = m(id \otimes \tilde{S})(\mathcal{F}\mathcal{F}^{-1}(\tilde{t}_\alpha \otimes 1)) = \tilde{t}_\alpha .$$

This proves the second relation in Lemma IV.1. □

V. SYMPLECTON POLYNOMIALS FOR $\mathcal{U}_h(sl(2))$

In this section, we derive the explicit form of the symplecton for $\mathcal{U}_h(sl(2))$ and investigate its properties. Since $\mathcal{U}_h(sl(2))$ has the same commutation relations as $sl(2)$, $\mathcal{U}_h(sl(2))$ and $sl(2)$ have the same realizations. Therefore the symplecton realization for $\mathcal{U}_h(sl(2))$, which is identical to the one for $sl(2)$, is the realization in terms of the usual boson operators. This is a contrast to the q-symplecton where the q-deformed boson operators are used.

Let $\bar{a}$ and a be boson operators satisfying $[\bar{a},a]=1$, then the generators of $\mathcal{U}_h(sl(2))$ are realized by

$$J_+ = -\tfrac{1}{2}a^2, \quad J_- = \tfrac{1}{2}\bar{a}^2, \quad J_0 = \tfrac{1}{2}(a\bar{a} + \bar{a}a). \tag{V.1}$$

The h-symplecton, denoted by $\tilde{P}_j^m(a,\bar{a})$, is defined as a polynomial in $\bar{a},a$ satisfying

$$\begin{aligned} \mathrm{ad}\, J_\pm(\tilde{P}_j^m) &= \sqrt{(j \mp m)(j \pm m + 1)}\,\tilde{P}_j^{m\pm 1}, \\ \mathrm{ad}\, J_0(\tilde{P}_j^m) &= 2m\tilde{P}_j^m , \end{aligned} \tag{V.2}$$

where the adjoint action on the LHS is given by (IV.5). Using Lemma IV.1, the explicit form of h-symplecton is obtained from the corresponding one for $sl(2)$.

Proposition V.1: The explicit form of the h-symplecton defined by (V.2) is given by

$$\tilde{P}_j^m(a,\bar{a}) = P_j^m(a,\bar{a}) e^{m\sigma}, \tag{V.3}$$

where σ is given in (III.4) and $P_j^m(a,\bar{a})$ denotes $sl(2)$ symplecton.

Proof: By definition of $sl(2)$ symplecton, it holds that

$$(J_0 - 2m)P_j^m = P_j^m J_0 .$$

Using this and the RHS of (IV.8),

$$\tilde{P}_j^m = \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{1}{2}\right)^n P_j^m (J_0 + 2m)^n S(\sigma)^n u^{-1} = P_j^m \sum_{n=0}^{\infty}\sum_{s=0}^{n} \frac{(-1)^n (2m)^s}{2^n (n-s)!\, s!} J_0^{n-s} (\sigma)^n u^{-1}.$$

Changing the order of sum, then replacing $n-s$ with n, we obtain

$$\tilde{P}_j^m = P_j^m \sum_{s,n=0}^{\infty} \frac{(-1)^{n+s}(2m)^s}{2^{n+s} n!\, s!} J_0^n S(\sigma)^{n+s} u^{-1}. \tag{V.4}$$

Note that

$$u = \sum_{n=0}^{\infty}\left(-\frac{1}{2}\right)^n \frac{1}{n!} J_0^n S(\sigma)^n,$$

and (III.3), it follows that (V.4) is rewritten as

$$\tilde{P}_j^m = P_j^m \sum_{s=0}^{\infty} \left(-\frac{1}{2}\right)^s \frac{(2m)^s}{s!} \tilde{S}(\sigma)^s = P_j^m e^{m\sigma},$$

where (B4) is used in the last equality. □

We would like to show some explicit form of h-symplecton. For $j=1/2$,

$$\tilde{P}_{1/2}^{-1/2} = \bar{a} e^{-\sigma/2} \equiv \bar{a}_h, \quad \tilde{P}_{1/2}^{1/2} = \bar{a} e^{\sigma/2} \equiv a_h, \tag{V.5}$$

and for $j=1$,

$$\tilde{P}_1^{-1} = \bar{a}^2 e^{-\sigma} = \bar{a}_h^2 + h \bar{a}_h a_h,$$

$$\tilde{P}_1^0 = (\bar{a} a + a \bar{a})/\sqrt{2} = (\bar{a}_h a_h + a_h \bar{a}_h - h a_h^2)/\sqrt{2}, \tag{V.6}$$

$$\tilde{P}_1^1 = a^2 e^{\sigma} = a_h^2.$$

The $j=1/2$ h-symplecton forms covariant h-deformed oscillator algebra

$$[\bar{a}_h, a_h] = 1 - h a_h^2, \tag{V.7}$$

i.e., the commutation relation (V.7) is preserved under the action of $SL_h(2)$,

$$(a_h', \bar{a}_h') = (a_h, \bar{a}_h) \begin{pmatrix} x & u \\ v & y \end{pmatrix}. \tag{V.8}$$

This shows that it is possible to construct representations of $\mathrm{SL}_h(2)$ on h-symplecton. We shall discuss it later. It may be worth noting that the action (V.8) is different from the ones in Refs. 19 and 20, where a and $\bar{a}$ are *not* mixed by the action of quantum groups.

The $j=1$ h-symplecton forms an algebra isomorphic to $sl(2)$. Its commutation relations are

$$[P_0^1, P_1^1] = 2\sqrt{2} P_1^1 (1 - h P_1^1),$$

$$[P_1^0, P_1^{-1}] = -2\sqrt{2} P_1^{-1} (1 - h P_1^1), \tag{V.9}$$

$$[P_1^1 P_1^{-1}] = -2\sqrt{2} (1 - h P_1^1) P_1^0.$$

The generators of $sl(2)$ are written in terms of P_1^m,

$$J_+ = -\tfrac{1}{2} P_1^1 (1 - h P_1^1), \quad J_0 = \frac{1}{\sqrt{2}} P_1^0, \quad J_- = \tfrac{1}{2} P_1^{-1} (1 - h P_1^1). \tag{V.10}$$

We see, from the explicit form of h-symplecton (V3), that the h dependence of polynomial $\tilde{P}_j^m(a,\bar{a})$ is absorbed in σ which is an infinite polynomial in a^2. Recall that the relationship between $sl(2)$ symplecton and Gauss hypergeometric function ${}_2F_1$ is given in terms of the operator $N = a\bar{a}$, then we see that the factor $e^{m\sigma}$ in (V.3) does not affect this relationship. Therefore the specific hypergeometric function for h-symplecton may be again ${}_2F_1$.

The fact that the $j=1/2$ h-symplecton forms covariant h-oscillator algebra may suggest that it is useful to write h-symplecton in terms of covariant h-oscillators (V.5).

Proposition V.2: The h-symplecton is written in terms of covariant h-oscillators as follows. The corresponding expression for (II.4) is

$$\tilde{P}_j^m(a_h,\overline{a_h})=\frac{1}{2^{j-m}}\left[\frac{(2j)!(j-m)!}{(j+m)!}\right]^{1/2}\sum_{s=0}^{j-m}\frac{1}{s!(j-m-s)!}$$
$$\times\overline{a_h}(\overline{a_h}+ha_h)\cdots\{\overline{a_h}+(j-m-s-1)ha_h\}a_h^{j+m}$$
$$\times\{\overline{a_h}-(2m+s)ha_h\}\{\overline{a_h}-(2m+s-1)ha_h\}\cdots\{\overline{a_h}-(2m+1)ha_h\}, \quad \text{(V.11)}$$

and for (II.5) is

$$\tilde{P}_j^m(a_h,\overline{a_h})=\frac{1}{2^{j+m}}\left[\frac{(2j)!(j+m)!}{(j-m)!}\right]^{1/2}\sum_{s=0}^{j+m}\frac{1}{s!(j+m-s)!}$$
$$\times a_h^s(\overline{a_h}-hsa_h)\{\overline{a_h}+h(1-s)a_h\}\cdots\{\overline{a_h}+h(j-m-1-s)a_h\}a_h^{j+m-s}. \quad \text{(V.12)}$$

Proof: From (V.5),

$$\overline{a}=\overline{a_h}e^{\sigma/2}, \quad a=a_he^{-\sigma/2}.$$

Substituting these into (II.4) and (II.5), straightforward calculation proves the proposition. □

In order to discuss generating functions for *h*-symplecton, it is possible to apply Lemma IV.1 to the generating function (II.8) for $sl(2)$ symplecton, since the RHS of (II.8) is a sum of tensor operators of $sl(2)$. It follows that the RHS of (II.8) becomes the sum of *h*-symplecton; $\sqrt{(2j)!}\sum_{m=-j}^{j}\Phi_{jm}\tilde{P}_j^m$. However the LHS may be quite complicated and may not be in closed form. Another way to obtain generating functions for *h*-symplecton is to substitute (V.3) and (V.5) into (II.8),

$$(\xi a_he^{-\sigma/2}+\eta\overline{a_h}e^{\sigma/2})^{2j}=\sqrt{(2j)!}\sum_{m=-j}^{j}\Phi_{jm}(\xi,\eta)\tilde{P}_j^m(a_h,\overline{a_h})e^{-m\sigma}. \quad \text{(V.13)}$$

It is possible to remove σ from (V.13) by using the relation $e^{\sigma}+ha_h^2=1$, however, the obtained relation is quite complicated. Therefore the simplest generating function for *h*-symplecton may be (V.13), where the σ is regarded as a independent quantity subject to the relations

$$[\sigma,a_h]=0, \ [\sigma,\overline{a_h}]=2ha_h, \quad \text{(V.14)}$$

and $\lim_{h\to 0}\sigma=0$.

VI. PRODUCT LAW FOR *h*-SYMPLECTON

It is possible to extend the product law (II.18) for $sl(2)$ symplecton to *h*-symplecton. The product law plays a crucial role when the symplecton calculus is considered. In this section, we first prove the product law for *h*-symplecton by using the one for $sl(2)$ symplecton, then consider the symplecton calculus for $\mathcal{U}_h(sl(2))$.

Theorem VI.1: *Let* $\tilde{P}_j^m$ *and* $\tilde{P}_{j'}^{m'}$ *be h-symplecton, then these obey the product law,*

$$\tilde{P}_j^m\tilde{P}_{j'}^{m'}=\sum_{k=|j-j'|}^{j+j'}\sum_{n,n'}\langle k|j|j'\rangle(F^{-1})_{n,n'm,m'}^{j,j'}C_{n',m,n'+m}^{j',j,k}\tilde{P}_k^{n'+m}, \quad \text{(VI.1)}$$

where $C_{m_1,m_2,m}^{j_1,j_2,j}$ *is the CGC for* $sl(2)$ *and*

$$\langle k|j|j'\rangle=2^{k-j-j'}(2k+1)^{-1/2}\nabla(kjj'),$$

$$\nabla(abc)=\left[\frac{(a+b+c+1)!}{(a+b-c)!(a-b+c)!(-a+b+c)!}\right]^{1/2}. \tag{VI.2}$$

Proof: From Proposition V.1,

$$\tilde{P}_j^m\tilde{P}_{j'}^{m'}=P_j^m e^{m\sigma}\tilde{P}_{j'}^{m'}e^{-m\sigma}e^{m\sigma}.$$

Using the Hopf algebra mappings for σ given in (B4), we see that the adjoint action of $e^{m\sigma}$ is given by

$$\mathrm{ad}e^{m\sigma}(t)=e^{m\sigma}te^{-m\sigma}. \tag{VI.3}$$

h-Symplecton is an irreducible tensor operator of $\mathcal{U}_h(sl(2))$, and it follows that

$$e^{m\sigma}\tilde{P}_{j'}^{m'}e^{-m\sigma}=\mathrm{ad}e^{m\sigma}(\tilde{P}_{j'}^{m'})=\sum_{n'}(e^{m\sigma})_{n',m'}^{j'}\tilde{P}_{j'}^{n'}$$
$$=\sum_{n,n'}\delta_{n,m}(e^{m\sigma})_{n',m'}^{j'}\tilde{P}_{j'}^{n'}=\sum_{n,n'}(F^{-1})_{n,n'm,m'}^{j,j'}\tilde{P}_{j'}^{n'}, \tag{VI.4}$$

where the matrix elements of $\mathcal{F}$ (C4) is used in the last equality. Therefore, we have

$$\tilde{P}_j^m\tilde{P}_{j'}^{m'}=\sum_{n,n'}(F^{-1})_{n,n'm,m'}^{j,j'}P_j^m\tilde{P}_{j'}^{n'}e^{m\sigma}=\sum_{n,n'}(F^{-1})_{n,n'm,m'}^{j,j'}P_j^mP_{j'}^{n'}e^{(n'+m)\sigma}.$$

Applying the product law (II.18) for $sl(2)$ symplecton, Theorem VI.1 is proved. □

Corollary VI.1: The associativity of the products $(\tilde{P}_a^\alpha\tilde{P}_b^\beta)\tilde{P}_c^\gamma=\tilde{P}_a^\alpha(\tilde{P}_b^\beta\tilde{P}_c^\gamma)$ *gives the same relation as (II.20) for the triangle function* $\nabla(abc)$ *that appeared in Theorem VI.1.*

Proof: The associativity gives the same relation as (II.20), but the Racha coefficients are replaced with the ones for $\mathcal{U}_h(sl(2))$. From Lemma III.2, these two kinds of Racha coefficients coincide. □

Let us now consider the h-symplecton calculus. We assume the formal ket$|\,\rangle$ and that both $\bar{a}|\,\rangle$ and $a|\,\rangle$ are nonvanishing vectors. Then the vectors defined by $|jm\rangle=\tilde{P}_j^m|\,\rangle$ are irrep. bases of $\mathcal{U}_h(sl(2))$ provided that the action of $X\in\mathcal{U}_h(sl(2))$ is defined by $X|jm\rangle=\mathrm{ad}X(\tilde{P}_j^m)|\,\rangle$. The dual bases are defined by $\langle jm|=\langle\,|\tilde{P}_j^{-m}(-1)^{j-m}$ in order to keep the correspondence with the $h=0$ case. The action of $X\in\mathcal{U}_h(sl(2))$ is, of course, given by $\langle jm|=\langle\,|\mathrm{ad}X(\tilde{P}_j^{-m})(-1)^{j-m}$. The inner product is defined in the same manner as $h=0$ case, namely,

$$\langle jm|j'm'\rangle=\langle\,|(-1)^{j-m}\tilde{P}_j^{-m}\cdot\tilde{P}_{j'}^{m'}|\,\rangle,$$

the operation $\langle|(\cdots)|\rangle$ means to take only the $j=0$ part of the expression $(\cdots)$. Applying the product law (VI.1) for h-symplecton, we obtain

$$\langle jm|j'm'\rangle=\delta_{j,j'}2^{-2j}F_{-m,m\ -m,m'}^{j,j}. \tag{VI.5}$$

Therefore the vectors $|jm\rangle$ and $|j'm'\rangle$ are orthonormal if they belong to different irreps. but *not* orthonormal if they belong to a same irrep. The nonvanishing part on the RHS of (VI.5) depends on only the twist element $\mathcal{F}$.

From the product law, we can show the following relations for h-symplecton:

Proposition VI.1: The following relations hold for h-symplecton:

$$\sum_{m,m'}\tilde{P}_j^m\tilde{P}_{j'}^{m'}F_{m,m'l,l'}^{j,j'}=\sum_k\langle k|j|j'\rangle C_{l',l,l'+k}^{j',j,k}\tilde{P}_k^{l+l'}, \tag{VI.6}$$

$$\tilde{P}^{m'}_{j'}(a_h,\overline{a_h}+2hma_h)=\sum_{l=0}(F^{-1})^{j,j'}_{n,n'm,m'}\tilde{P}^{m'+l}_{j'}(a_h,\overline{a_h}). \tag{VI.7}$$

Proof: The relation (VI.6) is easily proved by multiplying the product law (VI.1) by $F^{j,j'}_{n,n'm,m'}$ and summing over m,m'. The relation (VI.7) is derived by moving $e^{m\sigma}$ to the right of $P^{m'}_{j'}$ in (VI.4). One can do that by using the relations

$$e^{m\sigma}a_h=a_he^{m\sigma},\quad e^{m\sigma}\overline{a_h}=(\overline{a_h}+2hma_h)e^{m\sigma}.$$

□

VII. QUANTUM *h*-PLANE AND REPRESENTATIONS OF $SL_h(2)$

The Jordanian quantum algebra $\mathcal{U}_h(sl(2))$ and Jordanian quantum group $SL_h(2)$ are dual each other. It follows that any representation basis of $\mathcal{U}_h(sl(2))$ is also representation basis for $SL_h(2)$ belonging to the same representation. Since *h*-symplecton is an irrep. basis of $\mathcal{U}_h(sl(2))$, it is also an irrep. basis of $SL_h(2)$. We have seen this for $j=1/2$ in Sec. V. The relation (V.8) can be generalized to arbitrary *j*,

$$\tilde{P}^m_j(a'_h,\overline{a'_h})=\sum_n \tilde{P}^n_j(a_h,\overline{a_h})\tilde{d}^j_{nm}(g)\qquad g\in SL_h(2). \tag{VII.1}$$

We can obtain *d*-functions for $\mathcal{U}_h(sl(2))$ by substituting (V.8) into the explicit form of *h*-symplecton given in Proposition V.2. However, as is seen from the explicit form, the actual computation seems to be complicated.

The use of quantum *h*-plane[21] provides us a procedure which is a little bit simpler in computation. In this section, we shall find irrep. bases for $SL_h(2)$ in terms of the quantum *h*-plane which give the same irreps. as *h*-symplecton by using the tensor operator approach. Recall that the functions $\Phi_{jm}(\xi,\eta)$ defined by (II.9) are irrep. bases of $sl(2)$ in the realization (II.10) and irrep. bases of SL(2) under (II.11) as well. We can regard $\Phi_{jm}(\xi,\eta)$ as an irreducible tensor operator of $sl(2)$, since it is easy to verify that

$$\begin{aligned}[J_\pm,\Phi_{jm}]&=\sqrt{(j\mp m)(j\pm m+1)}\Phi_{jm\pm1},\\ [J_0,\Phi_{jm}]&=2m\Phi_{jm}.\end{aligned} \tag{VII.2}$$

From Lemma IV.1, it is easy to find the corresponding irreducible tensor operators for $\mathcal{U}_h(sl(2))$.

Proposition VII.1: Let ξ, η be commutative numbers, then the following are irreducible tensor operators for $\mathcal{U}_h(sl(2))$:

$$\tilde{\Phi}_{jm}(\xi,\eta)=\Phi_{jm}(\xi,\eta)e^{m\sigma}, \tag{VII.3}$$

where

$$\sigma=-\ln\left(1-2h\xi\frac{d}{g\eta}\right).$$

For $j=1/2$, we have

$$\tilde{\Phi}_{(1/2)(1/2)}=\xi e^{\sigma/2}\equiv\xi_h,\quad \tilde{\Phi}_{(1/2)(-1/2)}=\eta e^{-\sigma/2}\equiv\eta_h, \tag{VII.4}$$

and they satisfy the commutation relation

$$[\xi_h,\eta_h]=h\xi_h^2, \tag{VII.5}$$

this corresponds to the commutation relation of quantum h-plane in Ref. 21. It is easily verified that the commutation relation (VII.5) is preserved under the action of $SL_h(2)$,

$$(\xi_h', \eta_h') = (\xi_h, \eta_h)\begin{pmatrix} x & u \\ v & y \end{pmatrix}. \tag{VII.6}$$

It is an easy exercise to write $\tilde{\Phi}_{jm}$ in terms of ξ_h and η_h. Then $\tilde{\Phi}_{jm}(\xi_h, \eta_h)$ forms irrep. bases of $SL(2)$, that is, the d-functions for $SL_h(2)$ are obtained by substituting (VII.6) into $\tilde{\Phi}_{jm}(\xi_h, \eta_h)$.

Proposition VII.2: Irreps. of $SL_h(2)$ on the quantum h-plane are obtained by

$$\tilde{\Phi}_{jm}(\xi_h', \eta_h') = \sum_k \tilde{\Phi}_{jk}(\xi_h, \eta_h)\tilde{d}^j_{km}, \tag{VII.7}$$

where the irrep. bases are given by

$$\begin{aligned}\tilde{\Phi}_{jm} &= c_{jm}\xi_h^{j+m}(\eta_h - h(j+m)\xi_h)(\eta_h - h(j+m-1)\xi_h)\cdots(\eta_h - h(2m+1)\xi_h),\\ &= c_{jm}\eta_h(\eta_h + h\xi_h)\cdots(\eta_h + (j-m-1)h\xi_h)\xi_h^{j+m},\end{aligned} \tag{VII.8}$$

with

$$c_{jm} = \frac{1}{\sqrt{(j+m)!(j-m)!}}.$$

Since $\tilde{P}_{jm}$ and $\tilde{\Phi}_{jm}$ give the same irreps. of $\mathcal{U}_h(sl(2))$, they also give the same d-functions of $SL_h(2)$. Indeed, the explicit computation shows that we obtain the same d-functions for $j=1/2$ and $j=1$. The $j=1/2$ case gives the 2×2 quantum matrix T (III.7) itself, while $j=1$ d-function reads

$$d^1 = \begin{pmatrix} x^2 + hxv & \sqrt{2}(ux + huv) & u^2 + hu(x+y+hv) \\ \sqrt{2}xv & 1 + 2uv & \sqrt{2}(uy + huv) \\ v^2 & \sqrt{2}yv & y^2 + hyv \end{pmatrix}. \tag{VII.9}$$

The d-functions for $SL_h(2)$ are also discussed in Ref. 22, where the authors assert that the d-functions can be obtained from the q-deformed ones via a contraction method and show some explicit examples. Another way to obtain the d-functions is to use the recurrence relations for d-functions. This will be discussed in a separate publication.

VIII. CONCLUDING REMARKS

We have constructed h-symplecton in this article and investigated some of its properties. It has been seen that many properties of $sl(2)$ symplecton are inherited to h-symplecton. Unfortunately, h-dependence of h-symplecton is absorbed in σ, namely, twist element $\mathcal{F}$, so that we cannot see specific hypergeometric function for h-deformation. It will become clear what kind of hypergeometric functions are specific to h-deformed quantities if we obtain explicit form of d-function for $SL_h(2)$ as in the case of q-deformed $SU(2)$.[23] The $sl(2)$ symplecton has a simple generating function. We presented (V.13) as a generating function for h-symplecton. However, this may be one of possible choices, we might find simpler generating function. The use of quantum h-plane ξ_h, η_h instead of ξ, η is one of the possibilities. We have done some calculation to find simpler form of generating function in terms of ξ_h and η_h, however, all what we obtained have more complicated form.

We would like to emphasize the usefulness of Lemma IV.1. This provides us with a much simpler procedure to obtain h-symplecton than starting with the definition (V.2) and using the lemma, we could easily find another irrep. bases (VII.8) for $SL_h(2)$. This lemma is, of course, applicable to any Jordanian quantum algebra, since we usually know the explicit form of twist element. Furthermore, the lemma is extended to quasitriangular Hopf algebras.[24] For quasitriangular Hopf algebras, the twist elements are usually not known, they are known up to certain order of the deformation parameters. It is expected that many properties of tensor operators for quasitriangular Hopf algebras are studied based on the present knowledge of the tensor operators for Lie algebras via Lemma IV.1, even if the explicit form of tensor operators is not obtained. It may also be possible to apply Lemma IV.1 to the investigation of q-symplecton.

APPENDIX A: RELATION TO OHN's $\mathcal{U}_h(sl(2))$

Ohn defined in Ref. 6 $\mathcal{U}_h(sl(2))$ as an algebra generated by H, X, and Y subject to

$$\begin{aligned} &[X,Y]=H, \quad [H,X]=2\,\frac{\sinh hX}{h}, \\ &[H,Y]=-Y(\cosh hX)-(\cosh hX)Y. \end{aligned} \tag{A1}$$

Meanwhile, the commutation relations of $J_\pm$, J_0, which are generators of $\mathcal{U}_h(sl(2))$ in this article, are same as $sl(2)$. These two kinds of generators are related by

$$\begin{aligned} &H=e^{-\sigma/2}J_0, \quad X=\frac{\sigma}{2h}, \\ &Y=e^{-\sigma/2}\left(J_-+\frac{h}{2}J_0^2\right)-\frac{h}{8}e^{\sigma/2}(e^{-\sigma}-1). \end{aligned} \tag{A2}$$

By this relation, not only the commutation relations but also the Hopf algebra mappings are transformed each other. The relation (A2) corresponds to the one parameter case discussed in Ref. 25, where the two parameter Jordanian deformation of $gl(2)$ is considered.

APPENDIX B: HOPF ALGEBRA STRUCTURE OF $\mathcal{U}_h(sl(2))$

We here give explicit formulas for the coproduct, counit, and antipode of $\mathcal{U}_h(sl(2))$ calculated from (III.3).

(i) Coproduct,

$$\begin{aligned} &\tilde{\Delta}(J_0)=J_0\otimes e^{\sigma}+1\otimes J_0, \\ &\tilde{\Delta}(J_+)=J_+\otimes 1+e^{-\sigma}\otimes J_+, \\ &\tilde{\Delta}(J_-)=J_-\otimes e^{\sigma}+1\otimes J_- -hJ_0\otimes e^{\sigma}J_0-\frac{h}{2}J_0(J_0+2)\otimes e^{\sigma}(e^{\sigma}-1). \end{aligned} \tag{B1}$$

(ii) Counit,

$$\tilde{\epsilon}(X)=0, \quad X=J_\pm, J_0. \tag{B2}$$

(iii) Antipode,

$$\begin{aligned} &\tilde{S}(J_0)=-J_0e^{-\sigma}, \quad \tilde{S}(J_+)=-J_+e^{\sigma}, \\ &\tilde{S}(J_-)=-J_-e^{-\sigma}-\frac{h}{2}J_0^2(e^{-\sigma}+1)e^{-\sigma}+hJ_0(e^{-\sigma}-1)e^{-\sigma}. \end{aligned} \tag{B3}$$

All of these are reduced to the ones for $sl(2)$ in the limit of $h=0$. The Hopf algebra mappings for σ have simple form

$$\tilde{\Delta}(\sigma)=\sigma\otimes 1+1\otimes\sigma,\quad \tilde{\epsilon}(\sigma)=0,\quad \tilde{S}(\sigma)=-\sigma. \tag{B4}$$

APPENDIX C: MATRIX ELEMENTS OF $\mathcal{F}$

In this Appendix, we show the explicit formula of matrix elements of the twist element $\mathcal{F}$ (III.4) and some of their properties. We denote an irrep. basis of $\mathcal{U}_h(sl(2))$ by the bracket notation $|jm\rangle$ for the sake of simplicity.

It is easily verified the following relations from (III.3):

$$\begin{aligned}\tilde{\Delta}(J_\pm)\mathcal{F}|j_1m_1\rangle\otimes|j_2m_2\rangle&=\sqrt{(j_1\mp m_1)(j_1\pm m_1+1)}\mathcal{F}|j_1m_1\pm 1\rangle\otimes|j_2m_2\rangle\\&\quad+\sqrt{(j_2\mp m_2)(j_2\pm m_2+1)}\mathcal{F}|j_1m_1\rangle\otimes|j_2m_2\pm 1\rangle,\\ \tilde{\Delta}(J_0)\mathcal{F}|j_1m_1\rangle\otimes j_2m_2\rangle&=2(m_1+m_2)\mathcal{F}|j_1m_1\rangle\otimes|j_2m_2\rangle.\end{aligned} \tag{C1}$$

It shows that the vectors $\mathcal{F}|j_1m_1\rangle\otimes|j_2m_2\rangle$ for $\mathcal{U}_h(sl(2))$ play the same role as $|j_1m_1\rangle\otimes|j_2m_2\rangle$ for $sl(2)$. Equation (III.5) is readily obtained from this. Another proof of Lemma III.1 with the bases of $\mathcal{U}_h(sl(2))$ in Ref. 6 is found in Refs. 26 and 27.

In the bracket notation, matrix elements of $\mathcal{F}$ are defined by

$$F^{j_1,j_2}_{k_1,k_2\ \ m_1,m_2}=\langle j_1k_1|\otimes\langle j_2k_2|\mathcal{F}|j_1m_1\rangle\otimes|j_2m_2\rangle.$$

We first show a relationship between the matrix elements of $\mathcal{F}$ and its inverse

$$F^{j_1,j_2}_{-n_1,-n_2\ \ -m_1,-m_2}=(F^{-1})^{j_1,j_2}_{m_1,m_2\ \ n_1,n_2}. \tag{C2}$$

The LHS of (C2) is calculated as

$$\begin{aligned}\text{LHS}&=\langle j_1-n_1|\otimes\langle j_2-n_2|\sum_{l=0}^{\infty}\frac{(-J_0)^l\otimes\sigma^l}{2^l l!}|j_1-m_1\rangle\otimes|j_2-m_2\rangle\\&=\delta_{n_1,m_1}\langle j_2-n_2|\exp(n_1\sigma)|j_2-m_2\rangle,\end{aligned} \tag{C3}$$

where $\langle j_1-n_1|J_0=-2n_1\langle j_1-n_1|$ is used. While the RHS is

$$\text{RHS}=\langle j_1m_1|\otimes\langle j_2m_2|\sum_{l=0}^{\infty}\frac{J_0^l\otimes\sigma^l}{2^l l!}|j_1n_1\rangle\otimes|j_2n_2\rangle=\delta_{m_1,n_1}\langle j_2m_2|\exp(n_1\sigma)|j_2n_2\rangle. \tag{C4}$$

Note that

$$J_+|jm\rangle=\sqrt{(j-m)(j+m+1)}|jm+1\rangle,$$

$$\langle j-m|J_+=\sqrt{(j-m)(j+m+1)}\langle j-m-1|.$$

It follows that any polynomials in J_+, denoted by $f(J_+)$, satisfies

$$\langle jm|f(J_+)|jn\rangle=\langle j-n|f(J_+)|j-m\rangle. \tag{C5}$$

Since σ is a polynomial in J_+, we see that (C3) is equal to (C4). Thus (C2) has been proved.

We next show that the matrix elements of $\mathcal{F}$ are given by

$$F^{j_1,j_2}_{k_1,k_2\ m_1,m_2}=\delta_{k_1,m_1}\theta(m_2\leqslant k_2\leqslant j_2)$$
$$\times S^{j_2}_{k_2,m_2}\begin{cases}\dfrac{(2k_2-2m_1-2m_2-2)!!}{(k_2-m_2)!(-2m_1-2)!!}h^{k_2-m_2}, & \text{for } m_1\leqslant 0,\\ (-2h)^{k_2-m_2}\displaystyle\sum_{l=0}^{j_2-m_2}(-1)^l\binom{2m_1}{k_2-m_2-l}\dfrac{(2l+2m_1-2)!!}{2^l l!(2m_1-2)!!}, & \text{for } m_1>0,\end{cases} \tag{C6}$$

where $n!!=1$ for $n\leqslant 0$ and $\theta(m_2\leqslant k_2\leqslant j_2)=1$ if and only if the inequality in the parentheses holds, otherwise θ vanishes. $S^{j_2}_{k_2,m_2}$ is defined by

$$S^{j_2}_{k_2,m_2}=\left\{\frac{(j_2-m_2)!(j_2+k_2)!}{(j_2+m_2)!(j_2-k_2)!}\right\}^{1/2}.$$

To prove (C6), note that similar to (C4), we have

$$F^{j_1,j_2}_{k_1,k_2\ m_1,m_2}=\delta_{k_1,m_1}\langle j_2k_2|e^{-m_1\sigma}|j_2m_2\rangle. \tag{C7}$$

One can use the power series expansion in order to compute the RHS of (C7),

$$(1-X)^{-l/2}=\sum_{n=0}^{\infty}\frac{(2n+l-2)!!}{2^n n!(l-2)!!}X^n,\quad l\in\mathbf{Z}_+. \tag{C8}$$

(i) For $m_1\leqslant 0$.
Let $m_1=-l/2$ ($l\in\mathbf{Z}_+$) and using (C8),

$$e^{-m_1\sigma}|j_2m_2\rangle=(1-2hJ_+)^{-l/2}|j_2m_2\rangle$$
$$=\sum_{n=0}^{j_2-m_2}\frac{(2n+l-2)!!}{2^n n!(l-2)!!}(2h)^n\left\{\frac{(j_2-m_2)!(j_2+m_2+n)!}{(j_2+m_2)!(j_2-m_2-n)!}\right\}^{1/2}|j_2m_2+n\rangle.$$

Therefore $\langle j_2k_2|e^{-m_1\sigma}|j_2m_2\rangle$ takes values if and only if $k_2=m_2+n$. This proves the first part of (C6).

(ii) For $m_1>0$.
Let $m_1=l/2\,(l\in\mathbf{Z}_+)$. Since

$$e^{-m_1\sigma}|j_2m_2\rangle=e^{-l\sigma}e^{l\sigma/2}|j_2m_2\rangle,$$

we can apply the previous result to compute $e^{l\sigma/2}|j_2m_2\rangle$ and then applying the binomial expansion to $e^{-l\sigma}=(1-2hJ_+)^l_-$,

$$e^{-m_1\sigma}|j_2m_2\rangle=\sum_{t=0}^{j_2-m_2}\frac{(2t+l-2)!!}{2^t t!(l-2)!!}(2h)^t\left\{\frac{(j_2-m_2)!(j_2+m_2+t)!}{(j_2+m_2)!(j_2-m_2-t)!}\right\}^{1/2}e^{-l\sigma}|j_2m_2+t\rangle$$
$$=\sum_{n=0}^{l}\sum_{t=0}^{j_2-m_2}\binom{l}{n}\frac{(2t+l-2)!!}{2^t t!(l-2)!!}(-1)^n(2h)^{t+n}\left\{\frac{(j_2-m_2)!(j_2+m_2+t+n)!}{(j_2+m_2)!(j_2-m_2-t-n)!}\right\}^{1/2}$$
$$\times|j_2m_2+t+n\rangle.$$

Replacing $t+n$ with n, we obtain

$$e^{-m_1\sigma}|j_2m_2\rangle=\sum_{t,n}(-1)^t(-2h)^n\binom{l}{n-t}\times\frac{(2t+l-2)!!}{2^t t!(l-2)!!}\left\{\frac{(j_2-m_2)!(j_2+m_2+n)!}{(j_2+m_2)!(j_2-m_2-n)!}\right\}^{1/2}\quad |j_2m_2+n\rangle.$$

Again $\langle j_2k_2|e^{-m_1\sigma}|j_2m_2\rangle$ takes values if and only if $k_2=m_2+n$. This completes the proof of (C6).

We can obtain the explicit formula for the universal *R*-matrix in the irreps. with highest weight j_1 and j_2 by combining (C6) and relation (C2), since the universal *R*-matrix for $\mathcal{U}_h(sl(2))$ is given by $\mathcal{R}=\mathcal{F}_{21}\mathcal{F}^{-1}$.

APPENDIX D: PROOF OF LEMMA III.2

Let V^a, and V^b, and V^c be representation spaces of $\mathcal{U}_h(sl(2))$ with highest weight *a*, *b*, and *c*, respectively. Bases of each space are denoted as e^a_α, $-a\leqslant\alpha\leqslant a$. We would like to construct irrep. bases in the space $V^a\otimes V^b\otimes V^c$ in two ways, namely, $(V^a\otimes V^b)\otimes V^c$ and $V^a\otimes(V^b\otimes V^c)$. According to the discussion in Appendix B, irrep. bases in the space $V^a\otimes V^b$ are given by

$$e^{(ab)d}_\delta=\sum C^{a,b,d}_{\alpha,\beta,\delta}\,\mathcal{F}\,e^a_\alpha\otimes e^b_\beta.$$

Then we couple these with the bases in V^c to obtain

$$\psi^e_\epsilon=\sum C^{d,c,e}_{\delta,\gamma,\epsilon}(\tilde{\Delta}\otimes id)(\mathcal{F})e^{(ab)d}_\delta\otimes c^c_\gamma=\sum C^{d,c,e}_{\delta,\gamma,\epsilon}C^{a,b,d}_{\alpha,\beta,\delta}(\tilde{\Delta}\otimes id)(\mathcal{F})\mathcal{F}_{12}e^a_\alpha\otimes e^b_\beta\otimes e^c_\gamma.$$

Similarly we obtain the following bases when we couple V^b and V^c first:

$$\psi'^e_\epsilon=\sum C^{a,f,e}_{\alpha,\rho,\epsilon}C^{b,c,f}_{\beta,\gamma,\rho}(id\otimes\tilde{\Delta})(\mathcal{F})\mathcal{F}_{23}e^a_\alpha\otimes e^b_\beta\otimes e^c_\gamma.$$

From (B4),

$$(id\otimes\tilde{\Delta})(\mathcal{F})=\exp(-\tfrac{1}{2}J_0\otimes\tilde{\Delta}(\sigma))=\mathcal{F}_{12}\mathcal{F}_{13}.$$

Using the relations (III.3), (III.2) and above

$$(\tilde{\Delta}\otimes id)(\mathcal{F})=\mathcal{F}_{12}(\Delta\otimes id)(\mathcal{F})\mathcal{F}_{12}^{-1}=\mathcal{F}_{23}(id\otimes\Delta)(\mathcal{F})\mathcal{F}_{12}^{-1}=(id\otimes\tilde{\Delta})(\mathcal{F})\mathcal{F}_{23}\mathcal{F}_{12}^{-1}=\mathcal{F}_{12}\mathcal{F}_{13}\mathcal{F}_{23}\mathcal{F}_{12}^{-1}.$$

It follows that ψ^e_ϵ and ψ'^e_ϵ are rewritten as

$$\psi^e_\epsilon=\sum C^{d,c,e}_{\delta,\gamma,\epsilon}C^{a,b,d}_{\alpha,\beta,\delta}\mathcal{F}_{12}\mathcal{F}_{13}\mathcal{F}_{23}e^a_\alpha\otimes e^b_\beta\otimes e^c_\gamma,$$

$$\psi'^e_\epsilon=\sum C^{a,f,e}_{\alpha,\rho,\epsilon}C^{b,c,f}_{\beta,\gamma,\rho}\mathcal{F}_{12}\mathcal{F}_{13}\mathcal{F}_{23}e^a_\alpha\otimes e^b_\beta\otimes e^c_\gamma.$$

The Racha coefficients $W_h(abce;df)$ for $\mathcal{U}_h(sl(2))$ is defined by

$$\psi_{\epsilon}^{e}=\sum_{f} \sqrt{(2d+1)(2f+1)}W_h(abce;df)\psi_{\epsilon}'^{e}. \tag{D1}$$

It is now obvious that the Racha coefficients for $\mathcal{U}_h(sl(2))$ satisfy the relation

$$\sum_{\delta} C_{\delta,\gamma,\epsilon}^{d,c,e}C_{\alpha,\beta,\delta}^{a,b,d}=\sum_{f,\rho} C_{\alpha,\rho,\epsilon}^{a,f,e}C_{\beta,\gamma,\rho}^{b,c,f}\sqrt{(2d+1)(2f+1)}W_h(abce;df). \tag{D2}$$

This is the same relation for the Racha coefficients for $sl(2)$. This proves Lemma III.2.

[1] See, for example, L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics: Theory and Application, Encyclopedia of Mathematics and Its Applications* (Addison–Wesley, Reading, 1981), Vol. 8.
[2] See, for example, L. C. Biedenharn and M. A. Lohe, *Quantum Group Symmetry and q-Tensor Algebras* (World Scientific, Singapore, 1995).
[3] E. E. Demidov *et al.*, Prog. Theor. Phys. Suppl. **102**, 203 (1990).
[4] S. Zakrewski, Lett. Math. Phys. **22**, 287 (1991).
[5] H. Ewen, O. Ogievetsky, and J. Wess, Lett. Math. Phys. **22**, 297 (1991).
[6] Ch. Ohn, Lett. Math. Phys. **25**, 85 (1992).
[7] There exists mappings from the ordinary boson algebra to *q*-deformed ones (Refs. 20, 24, 28). It is, however, simpler to use the *q*-deformed boson algebras for representation theories.
[8] L. C. Biedenharn and J. D. Louck, Ann. Phys. (N.Y.) **63**, 459 (1971).
[9] L. C. Biedenharn and J. D. Louck, *The Racha–Wigner Algebra in Quantum Theory, Encyclopedia of Mathematics and Its Applications* (Addison–Wesley, Reading, 1981), Vol. 9.
[10] M. A. Lohe, L. C. Biedenharn, and J. D. Louck, Phys. Rev. D **43**, 617 (1991).
[11] L. C. Biedenharn and M. A. Lohe, in *Quantum Groups*, in *Proceedings of the Argonne Workshop*, edited by T. Curtright, D. Fairlie, and C. Zachos (World Scientific, Singapore, 1991).
[12] M. Nomura, J. Math. Phys. **33**, 3636 (1992).
[13] I. M. Gelfand and D. B. Fairlie, Commun. Math. Phys. **136**, 487 (1991).
[14] V. Rittenberg and M. Scheunert, J. Math. Phys. **33**, 436 (1992).
[15] V. G. Drinfeld, Leningrad Math. J. **1**, 1419 (1990); See also, N. Yu. Reshetikhin, Lett. Math. Phys. **20**, 331 (1990).
[16] P. P. Kulish and A. A. Stolin, Czech. J. Phys. **47**, 123 (1997).
[17] L. D. Faddeev, N. Yu. Reshetikhin, and L. A. Takhtajan, Leningrad Math. J. **1**, 193 (1990).
[18] N. Aizawa, J. Phys. A **31**, 5467 (1998).
[19] C. Quesne, Czech. J. Phys. **48**, 1471 (1998).
[20] G. Fiore, J. Math. Phys. **39**, 3437 (1998).
[21] V. Karimipour, Lett. Math. Phys. **30**, 87 (1994).
[22] R. Chakrabarti and C. Quesne, preprint, math.QA/9811064.
[23] T. Masuda *et al.*, J. Funct. Anal. **99**, 357 (1991), and references therein.
[24] G. Fiore, preprint, q-alg/9708017.
[25] N. Aizawa, Czech. J. Phys. **48**, 1273 (1998).
[26] N. Aizawa, J. Phys. A **30**, 5981 (1997).
[27] J. Van der Jeugt, J. Phys. A **31**, 1495 (1998); Czech. J. Phys. **47**, 1283 (1997).
[28] P. P. Kulish and E. V. Damaskinsky, J. Phys. A **23**, L415 (1990); M. Chaichian, P. P. Kulish, and J. Lukierski, Phys. Lett. B **262**, 43 (1991).

Towards the canonical tensor operators of $u_q(3)$. II. The denominator function problem

Sigitas Drobažis Ališauskas[a)]
Institute of Theoretical Physics and Astronomy, A. Goštauto 12, Vilnius 2600, Lithuania

(Received 17 May 1999; accepted for publication 21 July 1999)

The explicit denominator (normalization) function of the canonical tensor operators of the quantum algebra $u_q(3)$, corresponding to the maximal null space case is derived *ab initio* in terms of double basic hypergeometric series, which cannot be obtained as any q-extension of the SU(3) denominator polynomial $G^1_{b''}(\Delta,x)$ in terms of multiple (double or triple) balanced hypergeometric series, introduced by Biedenharn, Louck, and their collaborators (although their $q=1$ versions are shown being equivalent). The corresponding orthonormal seed isoscalar factors of the coupling (Wigner–Clebsch–Gordan) coefficients of $u_q(3)$ and SU(3) with multiple irreducible representations are presented. Conjectured expression of the q-polynomials [which ratios appear in the $u_q(3)$ and (new) SU(3) denominator functions for an arbitrary value of the canonical multiplicity label t of the repeating irreducible representations] in terms of multiple partition dependent q-series (extension of the maximal and minimal null space versions) is presented and considered. © *1999 American Institute of Physics.* [S0022-2488(99)01311-0]

I. INTRODUCTION AND PRELIMINARIES

The matrix elements of unit SU(3) canonical tensor operators,[1–4] which are characterized by the null space inclusion property, together with their Hermitian and conjugation properties and vanishing conditions of certain SU(3):U(2) maximal shift isoscalar factors (isofactors or matrix elements of projective[3] canonical tensor operators), reveal themselves as the most universal complete algebraic system for the orthonormal coupling (Wigner or Clebsch–Gordan) coefficients of the SU(3) group with the repeating irreducible representations (irreps) in the direct product decomposition. (For their numerical applications, see Ref. 5.) It is noticeable that the normalization problem of the coupling coefficients in the canonical splitting of the multiplicity for SU(3) (emphasized in many papers by Biedenharn, Louck and their collaborators) has been solved explicitly.[4,6] with the normalization coefficients or denominator functions $D^2(\Gamma_t,x)$ expressed in terms of the ratio of the preliminary guessed remarkable polynomials $G^t_{b''}(\Delta;x)$, which for the fixed multiplicity label t are completely determined by their null space (the weight space distribution of zeros), polynomial, and symmetry properties.[3] (We are using here and throughout an integer b'' for the irrep parameter instead of q in order to escape mixing with a deformation parameter.) Particularly, the investigation of the new class of special functions was started from the polynomials $G^1_{b''}(\Delta;x)$ of the maximal null space case;[7,8] their multiple well-poised series version of Holman *et al.*[9] has been extended to SU(n) by Milne.[10] However, the role of q-extension[11] of this multiple well-poised series for the quantum algebra $u_q(n)$ is not clear and the rather complicated generating and denominator function technique of Refs. 3, 4 hardly may be extended from SU(3) to the quantum algebra $u_q(3)$ with generic $q\neq 1$.

In a previous paper,[12] some important constructive elements of the explicit matrix elements of the $u_q(3)$ canonical tensor operators were derived using the generalized projection operator technique [extended from SU(3) canonical tensor operators as presented in Ref. 13] and some distinc-

[a)]Electronic mail: sigal@itpa.lt

tive and symmetry properties of the SU(3) and $u_q(3)$ canonical isofactors were considered. Particularly, explicit normalized isofactors of $u_q(3)$ characterized by the maximal null space were presented. However, the expressions of normalization coefficients (3.7b) and (3.14b) of Ref. 12 in terms of double sums are not universal (in some regions they are indefinite), although Eq. (3.14b), together with the symmetry relations (3.7a) and (3.14a), always can be helpful. Both Eqs. (3.7b) and (3.14b) turn into finite series only for the definite couples of fixed integers, related to restriction of the multiplicity of repeating irreps, in contrast with the SU(3) case,[4,7] when a single of such fixed integer parameters is sufficient.

In our paper,[14] mainly devoted to the overlaps of biorthogonal coupling coefficients of SU(3) and $u_q(3)$ (expressed in terms of classic and basic hypergeometric series), some matrix elements of the self-adjoint unit canonical tensor operators with the minimal null space are also derived (including their explicit normalization) and some boundary (seed) isofactors with the maximal null space[12] are specified as single sums (related to basic hypergeometric series, in contrast with multiple series,[15,16] cf. Ref. 17). In our last paper,[18] the expansion of the matrix elements of SU(3) and $u_q(3)$ canonical tensor operators in terms of the biorthogonal (dual) coupling coefficients[19,20] is considered. The composition of these dual expansions gives the explicit overlap coefficients for the Draayer–Akiyama construction[5,13] and, particularly, the normalization of explicit canonical seed isofactors, specified for the minimal null space cases (more general as the self-adjoint ones) and also presented as single sums (again related to the classic or basic hypergeometric series, respectively). In this minimal null space case, the normalization coefficients or denominator functions are expressed in terms of double sums and turn into finite series for a fixed integer, related to the multiplicity of repeating irreps. In this aspect they are similar to the denominator functions in the SU(3) case,[4,7,21] although they are rather different in their structure.

In the present paper, the solution of problems of Ref. 12 is under continuation; the normalization coefficients (denominator functions) of the canonical tensor operators are considered for the quantum algebra $u_q(3)$ and reconsidered for the SU(3) group. Some previously derived results are reviewed below in this section. In Sec. II, new expressions of the normalization coefficients (denominator functions) for the matrix elements of the SU(3) and $u_q(3)$ unit canonical tensor operators with the maximal null space are derived *ab initio*, using the dual expansion technique.[14,18] The extreme canonical seed isofactors and their normalization coefficients are expressed in terms of the finite series (single and double, respectively) for fixed multiplicity of repeating irreps. The symmetry properties [which were essential for the proof[3,8] of the uniqueness of solutions[4,7] for the denominator functions in the SU(3) case] are also demonstrated for the new denominator functions. In Sec. III, a conjecture about a q-generalization of the polynomial $G^t_{b''}(\Delta;x)$ in terms of partition dependent multiple q-factorial series for an arbitrary value of the multiplicity label t (valid also for the maximal t) is presented and discussed. Some symmetries and other properties of these new multiple q-factorial series are compared with the SU(3) case.[4]

Here and in what follows we use the same notations for irreps and basis states of SU(3) and $u_q(3)$ as were used in Refs. 12–14, 18, 19, with $(a\,b)$ for the mixed tensor irreps and $a=m_{13}-m_{23}$, $b=m_{23}-m_{33}$ where $[m_{13},m_{23},m_{33}]$ is a partition and m_{ij} are the Gelfand–Tsetlin parameters. The basis states are labeled by the hypercharge $y=m_{12}+m_{22}-\frac{2}{3}(m_{13}+m_{23}+m_{33})$ [or $z=\frac{1}{3}(b-a)-\frac{1}{2}y=m_{23}-\frac{1}{2}(m_{12}+m_{22})$], the isospin $i=\frac{1}{2}(m_{12}-m_{22})$ and its projection $i_z=m_{11}-\frac{1}{2}(m_{12}+m_{22})$. The parameter z usually is more convenient in explicit expressions than y, because linear combinations $i\pm z\geqslant 0$, $a+z-i\geqslant 0$, $b-z-i\geqslant 0$ are integers. For the state of irrep $(a\,b)$ in the coproduct $(a'b')\otimes(a''b'')$ decomposition, $z=z'+z''+v$, where again $v=\frac{1}{3}(a'-b'+a''-b''-a+b)$ is an integer. The parameters of the highest weight state take on the values $y_0=\frac{1}{3}(a+2b)$, $i_0=\frac{1}{2}a=-z_0$, while for the lowest weight state $\overline{y_0}=-\frac{1}{3}(2a+b)$, $\overline{i}_0=\frac{1}{2}b=\overline{z_0}$ and for the maximal isospin state $y_m=\frac{1}{3}(a-b)$, $i_m=\frac{1}{2}(a+b)$, $z_m=\frac{1}{2}(b-a)$.

The multiplicity r of irrep (ab) in the coproduct $(a'b')\otimes(a''b'')$ decomposition (intertwining number) is equal to $r=\min r_{\alpha\beta\gamma}+1$ $(\alpha=1,2,3;\beta=1,2,3,;\gamma=1,2)$ where integers $r_{\alpha\beta\gamma}$ form the $3\times 3\times 2$ array (1.3b) of Ref. 12 with equidistant parameters in the layers, rows, and columns. The multiplicity $\mathcal{M}$ of the canonical tensor operators $T^{(a''b'')t}_{y''_0 i''_0 i''_z}$ with fixed shifts $a-a'$ and $b-b'$ may exceed the external multiplicity r (see Refs. 12–14) and in this case the lowest values of the canonical multiplicity label $t=1,2,...,\mathcal{M}$ may be eliminated by the null space inclusion property.[3]

We use the Cartan–Weyl generators $E_{ik}(i,j,k=1,2,3)$ of the unitary quantum algebra $u_q(3)=U_q(u(3))$, with generic q and composite generators expressed in terms of q-deformed commutators, which satisfy the commutation relations[12,22,23] and the corresponding coproduct expansion rules. Here and in what follows $[x]$ and $[x]!$ are, respectively, the q-numbers and q-factorials,

$$[x]=(q^x-q^{-x})/(q-q^{-1}),\quad [x]!=[x][x-1]\ldots[2][1],$$

$$(\alpha|q)_n=\prod_{k=0}^{n-1}[a+k],\quad [a|q]_n=\prod_{k=0}^{n-1}[a-k], \tag{1.1}$$

$$[1]!=[0]!=(\alpha|q)_0=[\alpha|q]_0=1,$$

which are invariant under substitution $q\leftrightarrow q^{-1}$.

Solution of the equation system (2.14) of Ref. 12 for the bilinear combinations of canonical isofactors under conditions $t\leqslant\hat{I}'-i_m+i''_m+1$ is not always possible, in contrast with the results of our last papers,[14,18] where using the recursive constructions [a recoupling technique analogous with (2.13) and (5.1) of Ref. 13] the following superpositions of the canonical boundary (asymmetric seed) isofactors were derived:

$$(\tilde{T}^k|\eta_{-,+,\tilde{I}})_q=\sum_{t\geqslant k+1}U_3\left\{\begin{matrix}(a'b') & (a''-k,\overset{1}{b''}-k) & (a\,b)\\ \\ \overset{k+1}{(k\,k)} & (a\,b) & \overset{t}{(a''b'')}\end{matrix}\right\}_q\left[\begin{matrix}(a'b') & \overset{t}{(a''b'')} & (ab)\\ \\ y_0'i_0' & \overline{y_0''i_0''} & \tilde{y}\tilde{I}\end{matrix}\right]_q^{(3)} \tag{1.2a}$$

$$=\frac{([a+1][b+1][a+b+2][b']![a'+b'+1]!)^{1/2}\Gamma[ab\tilde{I}\tilde{z}]\nabla[\overline{i_0''}i_0'\tilde{I}]}{([b'']!)^{1/2}\mathcal{N}\binom{q,k}{a''b''}[a'b';ab]}$$

$$\times\sum_{j,j'}\frac{(-1)^{(b''-a''-k)/2-\tilde{I}+j}q^{Q_6+j(j+1)-j'(j'+1)+3\tilde{I}/(\tilde{I}+1)/2}}{\nabla^2[\frac{1}{2}(a''-k),\frac{1}{2}a',j]\nabla^2[\frac{1}{2}(a'+a''-a-k)-v,\frac{1}{2}a,j]}$$

$$\times\frac{[2j+1][2j'+1]}{[b-v+\frac{1}{2}(a'+a''-k)-j+1]![b-v+\frac{1}{2}(a'+a''-k)+j+2]!}$$

$$\times\frac{\nabla^2[\frac{1}{2}(a''+b'')-k,k',j]}{\nabla^2[\frac{1}{2}(b''-k),\frac{1}{2}a',j']\nabla^2[\frac{1}{2}k,j',\tilde{I}]\Gamma^2[abj',\tilde{z}-\frac{1}{2}k]}, \tag{1.2b}$$

(which can be also used for expansion of more universal coupling coefficients in terms of biorthogonal isofactors with superscript $-$, $+$, $\tilde{I}$). Here and in what follows $\tilde{z}=\frac{1}{2}(b''-a')+v)$, $\nabla[abc]$ and $\Gamma[abiz]$ denote, respectively, the functions

$$\nabla[abc]=\left(\frac{[a+b-c]![a-b+c]![a+b+c+1]!}{[b+c-a]!}\right)^{1/2}, \tag{1.3}$$

and

$$\Gamma[abiz]=\left(\frac{[i+z]![a+z-i]![a+z+i+1]!}{[i-z]![b-z-i]![b-z+i+1]!}\right)^{1/2}. \tag{1.4}$$

The renormalization factor

$$\mathcal{N}\begin{pmatrix} q,k \\ a''b'' \end{pmatrix}[a'b';ab]=\frac{\mathcal{D}\begin{pmatrix} q,t=1 \\ a''-k,b''-k \end{pmatrix}[a'b';ab]\mathcal{D}\begin{pmatrix} q,t=k+1 \\ k\,k \end{pmatrix}[a\,b;a\,b]}{[a''-k]![b''-k]!([k]!)^{1/2}} \tag{1.5}$$

is expressed in terms of the denominator functions[12,14] $\mathcal{D}\binom{q,t}{\cdots}[\cdots]$ of the $u_q(3)$ canonical tensor operators with maximal and minimal null space, respectively, and the corresponding q-phase Q_6 is expressed as follows:

$$\begin{aligned} Q_6=&(b'-v+1)(b'+b''-b+v-k)-\tfrac{1}{8}(a'+b''-k)(a'+b''-k+2)\\ &-\tfrac{1}{2}(a''-k)(a+b'-a''+k+v)+\tfrac{1}{2}(b''-k)(b''-b+v-k)\\ &-\tfrac{1}{4}(a+b-b'-b''-v+k)(a+b-b'-b''-v+k+2)-\tfrac{1}{8}k^2+\tfrac{3}{4}k+\tfrac{1}{2}k\tilde{z}. \end{aligned} \tag{1.6}$$

Particular canonical boundary (seed) isofactors may be derived from (1.2b) by means of a Gram–Schmidt procedure, beginning from $k=\mathcal{M}-1$, and using the overlaps $(\tilde{T}^k|\tilde{T}^{k'})_q$ equivalent to bilinear combination of special recoupling coefficients (3.16) of Ref. 18,

$$\sum_{t>\max(k,k')} U_3\begin{Bmatrix} (a'b') & \overset{1}{(a''-k,b''-k)} & (ab) \\ \overset{k+1}{(kk)} & (ab) & (a''b'') \end{Bmatrix}_q \times U_3\begin{Bmatrix} (a'b') & \overset{1}{(a''-k',b''-k')} & (ab) \\ \overset{k'+1}{(k'k')} & (ab) & \overset{t}{(a''b'')} \end{Bmatrix}_q \tag{1.7a}$$

$$=(\tilde{T}^k|\tilde{T}^{k'})_q=\sum_{\tilde{j}',\tilde{I}}(\tilde{T}^k|\eta^{+,\tilde{j}'',+})_q(\eta_{+,\tilde{j}'',+}|\eta^{-,+,\tilde{I}})_q(\eta_{-,+,\tilde{I}}|\tilde{T}^{k'})_q. \tag{1.7b}$$

Here the expansion coefficients of the twisted tensor operators $\tilde{T}^{(a''b'')t=k+1,q}_{y''i''i''_z}$ in terms of nonorthonormal tensor operators, which correspond to the dual coupled states $|\eta_{+,\tilde{j}'',+})_q$, is expressed [see (3.13b) of Ref. 18] as follows:

$$(\tilde{T}^k|\eta^{+,\tilde{j}'',+})_q=\frac{(-1)^{\tilde{j}'',\tilde{z}''}[a+b'-a''+v+k+1]!}{\mathcal{N}\begin{pmatrix}q,k\\a''b''\end{pmatrix}[a'b';ab]\nabla[\tfrac{1}{2}b',\tfrac{1}{2}b,\tilde{j}'']([\tilde{j}''+\tilde{z}'']!)^{1/2}}$$

$$\times\frac{H[a''b''\tilde{j}''\tilde{z}'']([2\tilde{j}''+1][\tilde{j}'-\tilde{z}']!)^{1/2}}{([a+1]![a+b+2]![a']![a'+b'+1]![a'']![b'']![a''+b''+1]!)^{1/2}}$$

$$\times\sum_{n,s',s''}\frac{(-1)^{n+s'}q^{R+(s''-s')(b'+b''+a+v+3)+k(s'-b')+n(a+b'-a''+v+k+2)}}{[n]![b''+v-k-n]![a'+a''-a-v-k-n]![a-a'+v+n]!}$$

$$\times\times\frac{[a'-n]![b-b''-v+k+n]![\tilde{j}''+\tilde{z}''+s']![s'']![b-b'-v+s'']!}{[s']![\tilde{j}''-\tilde{z}''-s']![b-b'-v+s']![s''-s']![k+s'-s'']!}$$

$$\times\frac{[a-a'+b''+v-k+s''+n+1]![a+b''+v+s'+2]!}{[a-a'+b''+v-k+s''+1]![b'+v-s'']![b''-k+s''+1]!}$$

$$\times\frac{[b'+v-s']!}{[a-a'-a''+v+ks''+n]![s+b''+v-k+s''+2]!},\tag{1.8}$$

where $\tilde{z}''=\frac{1}{2}(b-b')-v$, $\tilde{z}'=\frac{1}{2}(a''-a)-v$ and $H[abiz]$ denotes the function

$$H[abiz]=([a+z-i]![a+z+i+1]![b-z-i]![b-z+i+1]!)^{1/2}.\tag{1.9}$$

In Eq. (1.8) the same renormalization factor (1.5) and the q-phase,

$$\begin{aligned}R=&-Q_1(b'a'aba''b'';\tilde{j}''\tilde{z}'')-\tfrac{1}{2}\tilde{z}'(3\tilde{z}'+2a'-2b')\\&-\tfrac{1}{8}(a^2-a''^2)+\tfrac{1}{2}a'b'+b'-\tfrac{1}{2}(a+b-3a'-a''-b'')\\&+a''(b''+v-k)+\tfrac{1}{2}(b''-k)(a'-b'-b''+b-v+k)+\tfrac{1}{2}(a''-k)\\&\times(-a-2b-a''+v+k-4)+(a-a'+v)(a+b'+v+2)\\&+\tfrac{1}{2}k(k-2a''+v-3)-\tfrac{1}{4}(a-a''+2v)(a''+2b''),\end{aligned}\tag{1.10}$$

with

$$\begin{aligned}Q_1(a'b'a''b''a\,b;\tilde{J}\tilde{z})&=Q_1(b''a''b'a'b\,a;\tilde{J},-\tilde{z})\\&=\tfrac{1}{2}\{\tilde{J}(\tilde{J}+1)+\tilde{z}(3\tilde{z}+2a-2b)-ab+\tfrac{1}{2}(a'+b'')+a''+b'-a-b\}\end{aligned}\tag{1.11}$$

are used.

The auxiliary triangle overlap matrix in Eq. (1.7b) is expressed as follows:

$$(\eta_{+,\tilde{j}'',+}|\eta^{-,+,\tilde{I}})_q=q^{Q_1(baa'b'b''a'';\tilde{j}'',-\tilde{z}''}-Q_1(b''a''b'a'ba;\tilde{I},-\tilde{z})-b'/2-a'$$

$$\times(-1)^{\tilde{j}''-\tilde{z}''}\frac{[2\tilde{I}+1]\nabla[\tfrac{1}{2}b',\tfrac{1}{2}b,\tilde{j}'']H[a''b''\tilde{j}''\tilde{z}'']}{\nabla[\tfrac{1}{2}a',\tfrac{1}{2}b'',\tilde{I}]H[ab\tilde{I}\tilde{z}]}$$

$$\times\left(\frac{[a]![a+b+1]![a']![\tilde{j}''-\tilde{z}'']![\tilde{I}+\tilde{z}]![2\tilde{j}''+1]}{[a'']![a''+b''+1]![b']![\tilde{j}''+\tilde{z}'']![\tilde{I}-\tilde{z}]![b+1]}\right)^{1/2}$$

$$\frac{[\tilde{I}+\tilde{j}''+\frac{1}{2}(a-a''+v)]!(a-a''+v|q)_{\tilde{I}-\tilde{j}''+(a''-a-v)/2}}{[\tilde{I}-\tilde{j}''+\frac{1}{2}(a''-a-v)]![\tilde{I}+\tilde{j}''+\frac{1}{2}(a''-a-v)+1]!}$$

(see corrected Eq. (3.17) of Ref. 18).

II. DENOMINATOR FUNCTIONS OF CANONICAL TENSOR OPERATORS WITH THE MAXIMAL NULL SPACE

Using (3.1) of Ref. 12 (as a double sum, with fixed summation parameters $j'=\frac{1}{2}b'$ and m') and q-version of Minton's summation formula we derived (cf. Ref. 14) the boundary (seed) canonical isofactors with the multiplicity label, which corresponds to the maximal null space,

$$\begin{bmatrix} (a'b') & \overset{t=1}{(a''b'')} & (ab) \\ \\ \overline{y_0'}\,\overline{i_0'} & \tilde{y}''\tilde{j}'' & \overline{y_0}\,\overline{i_0} \end{bmatrix}^{(3)}_q$$

$$=(-1)^{a'+a''-a-v}\frac{([a+1]![a+b+2]![a']![a'+b'+1]!)^{1/2}}{\mathcal{D}\begin{pmatrix} q,t=1 \\ a''b''\end{pmatrix}[a'b';ab]\nabla[\frac{1}{2}b,\frac{1}{2}b',\tilde{j}'']}$$
$$\times\frac{([a'']![b'']![a''+b''+1]![2\tilde{j}''+1][\tilde{j}''-\tilde{z}'']!)^{1/2}}{([\tilde{j}''+\tilde{z}'']!)^{1/2}H[a''b''\tilde{j}''\tilde{z}''][a'+a''-v+1]!}$$
$$\times\sum_s\frac{(-1)^s q^{Q_5-s(a'+a''-v+1)}[\tilde{j}''+\tilde{z}''+s]![b'+v-s]!}{[s]![\tilde{j}''-\tilde{z}''-s]![-v+s]![a'+a''+b-v-s+2]!}, \tag{2.1a}$$

$$=(-1)^{b''-\tilde{z}''-\tilde{j}''}\frac{\Delta[\frac{1}{2}b,\frac{1}{2}b',\tilde{j}'']([a+1]![a+b+2]![a']![a'+b'+1]!)^{1/2}}{\mathcal{D}\begin{pmatrix} q,t=1 \\ a''b''\end{pmatrix}[a'b';a\,b]H[a''b''\tilde{j}''\tilde{z}'']}$$
$$\times\frac{([a'']![b'']![a''+b''+1]![2\tilde{j}''+1][\tilde{j}''-\tilde{z}'']![\tilde{j}''+\tilde{z}'']!)^{1/2}}{[a'+a''+b-v+2]![b'+b''+a+v+2]!}$$
$$\times\sum_s\frac{q^{Q_5-(\tilde{j}''-\tilde{z}'')(a+b'+b''+v-\tilde{j}''-\tilde{z}''+2)+s\{(b'+b)/2-\tilde{j}''+1\}}}{[s]![\tilde{j}''-\tilde{z}''-s]![\tilde{j}''+\frac{1}{2}(b-b')-s]![-v+s]!}$$
$$\times\frac{[b'+b''+a+v+s+2]!}{[a'+a''+v+\tilde{z}''-\tilde{j}''+s+1]!}, \tag{2.1b}$$

where $\Delta[a,b,c]$ denotes the triangle coefficient

$$\Delta[a,b,c]=\left(\frac{[a+b-c]![a-b+c]![-a+b+c]!}{[a+b+c+1]!}\right)^{1/2}, \tag{2.2}$$

and Q_5 denotes the q-phase,

$$Q_5=(a'+a''-a-v)^2+\tfrac{1}{2}\{(b'+b''-b+v)(a+b-2b''-2v)+av+\tilde{i}''(\tilde{i}''+1)\}-\tfrac{1}{8}(b-b')(b-b'+2). \tag{2.3}$$

The sums that appeared in (2.1a) and (2.1b) can be expressed in terms of the basic hypergeometric series ${}_3\phi_2$ related to the Clebsch–Gordan coefficient of the quantum algebra $u_q(1,1)$. Second version (2.1b) is more convenient for numerical calculations[24] (because it includes a minimal number of terms in the sum, and all its terms are of the same sign) and is derived using the following rearrangements of series:

$$\sum_s \frac{q^{-s\{(b'+b)/2+\tilde{j}''+1\}}}{[s]![\tilde{j}''-\tilde{z}''-s]![\tilde{j}''+\frac{1}{2}(b'-b)-s]![b-b'-v+s]!} \times\frac{[a+b''+v+s+1]!}{[\frac{1}{2}(b'+b)-\tilde{j}''+b''+a+v+s+2]!}$$
$$=\sum_{s,s',u'}\frac{q^{-s(\tilde{j}''+\tilde{z}''+u')-s'(b'+v-u')+u'-\tilde{j}''+\tilde{z}''}}{[s-\tilde{j}''+\tilde{z}''+u']![\tilde{j}''-\tilde{z}''-s]![\tilde{j}''+\frac{1}{2}(b'-b)-s]![b-b'-v+s]!}\times\frac{(-1)^{\tilde{j}''-\tilde{z}''-s'-u'}[a+b''+v+s'+1]!}{[s']![\tilde{j}''-\tilde{z}''-s'-u']![\frac{1}{2}(b'+b)-\tilde{j}''+b''+a+v+s'+2]!} \tag{2.4a}$$

$$=\sum_{s,s',u}\frac{q^{-s(b'+b''+a+v+u+2)+s'(a+b''-\tilde{z}''-\tilde{j}''+u+2)}}{[s-\tilde{j}''+\tilde{z}''+u]![\tilde{j}''-\tilde{z}''-s]![\tilde{j}''+\frac{1}{2}(b'-b)-s]![b-b'-v+s']!}\times\frac{(-1)^{\tilde{j}''-\tilde{z}''-s'-u}[a+b''+v+s'+1]!}{[s']![\tilde{j}''-\tilde{z}''-s'-u]![\frac{1}{2}(b'+b)-\tilde{j}''+b''+a+v+s+2]!} \tag{2.4b}$$

in analogy with Refs. 25, 26. Summation over u' or u in (2.4a) or (2.4b) is related to $\delta_{s,s'}$ and gives the l.h.s., when the Chu–Vandermonde summation formulas (cf. Refs. 25–27) lead either to the sum in (2.1a) or to the sum in (2.1b).

Denominator function $\mathcal{D}\binom{q,t=1}{a''b''}[a'b';ab]$ may be derived from overlap

$$(\tilde{T}^0|\tilde{T}^{k'})_q=\delta_{0,k'}=\sum_{\tilde{j}'',\tilde{I}}(\tilde{T}^0|\eta^{+,\tilde{j}'',+})_q(\eta_{+,\tilde{j}'',+}|\eta^{-,+,\tilde{I}})_q(\eta_{-,+\tilde{I}}|\tilde{T}^0)_q$$
$$=\sum_{\tilde{j}''}(\tilde{T}^0|\eta^{+,\tilde{j}'',+})_q\left[\begin{array}{ccc}(a'b') & \overset{t=1}{(a''b'')} & (a\,b)\\ & & \\ \overline{y_0'}\,\overline{i_0'} & \tilde{y}''\tilde{j}'' & \overline{y_0}\,\overline{i_0}\end{array}\right]^{(3)}_q, \tag{2.5}$$

with the boundary (seed) canonical isofactors expressed by means of (2.1a) and overlaps (expansion coefficients)

$$(\widetilde{T}^0|\eta^{+,\overline{\tilde{j}''},+})_q=(-1)^{b''-\tilde{z}''-\tilde{j}''}\frac{q^{R+(a'+a''-a-v)(a+b'-a''+v+2)}[a+b'-a''+v+1]!}{\mathcal{D}\begin{pmatrix} q,t=1 \\ a''b'' \end{pmatrix}[a'b';a\,b]H[a''b''\tilde{j}''\tilde{z}'']\nabla[\tfrac{1}{2}b',\tfrac{1}{2}b,\tilde{j}'']}$$

$$\times\left(\frac{[a'']![b'']![a''+b''+1]![2\tilde{j}''+1][\tilde{j}''-\tilde{z}'']!}{[a+1]![a+b+2]![a']![a'+b'+1]![\tilde{j}''+\tilde{z}'']!}\right)^{1/2}$$

$$\times\sum_n \frac{q^{-n(a+b'-a''+v+2)}[\tilde{j}''+\tilde{z}''+n]![b'+v-n]![a-a''+v+n]!}{[n]![\tilde{j}''-\tilde{z}''-n]![b-b'-v+n]!} \quad (2.6)$$

derived as a special case of (1.8) after summation over $s=s'$ of the balanced basic hypergeometric series[27] ${}_3\phi_2$, which appeared for $k=0$. Hence we write

$$\mathcal{D}^2\begin{pmatrix} q,t=1 \\ a''b'' \end{pmatrix}[a'b';a\,b]$$

$$=q^{Q_5+R}\frac{[a'']![b'']![a''+b''+1]![a+b'-a''+v+1]!}{[a'+a''-v+1]!}$$

$$\times\sum_{\tilde{j}'',n,s'}\frac{(-1)^{\tilde{j}''-\tilde{z}''+s'}[2\tilde{j}''+1][\tilde{j}''-\tilde{z}'']![\tilde{j}''+\tilde{z}''+n]!}{[\tilde{j}''+\tilde{z}'']![\tfrac{1}{2}(b'+b)-\tilde{j}'']![\tfrac{1}{2}(b'+b)+\tilde{j}''+1]!}$$

$$\times\frac{q^{(a'+a''-a-v-n)(a+b'-a''+v+2)}[b'+v-n]![a-a''+v+n]!}{H^2[a''b''\tilde{j}''\tilde{z}''][n]![\tilde{j}''-\tilde{z}''-n]![b-b'-v+n]!}$$

$$\times\frac{q^{-s(a'+a''+v+1)}[j''+\tilde{z}''+s']![b'+v-s']!}{[s']![\tilde{j}''-\tilde{z}''-s']![-v+s']![a'+a''+b-v-s'+2]!} \quad (2.7a)$$

$$=\frac{q^{Q_5+R}[a'']![b'']![a+b'-a''+v+1]!}{[a'+a''-v+1]![b'+b''+v+1]![a''+b-v+1]!}$$

$$\times\sum_{s,s',n}\frac{(-1)^{b-b'-s+s'}[b'+v-s']![a-a''+v+n]![b'+v-n]!}{[-v+s']![a'+a''+b-v-s'+2]![a-a'+b''+v-s]!}$$

$$\times\frac{q^{(a'+a''-a-v-n)(a+b'-a''+v+2)-s(a'+a''-v+1)}}{[s]![b''-s]![b-v-s]![b-b'-2v+s'+n-s]!}$$

$$\times\frac{[a''+b''+b-v-s+1]![b-b'-2v+s'+n]!}{[b-b'-v+n]![b'-b+2v-s'+s]![b'-b+2v-n+s]!}. \quad (2.7b)$$

We rearranged the sum over $\tilde{j}''$ as a very well-poised ${}_8\phi_7$ series of (2.7a) into the balanced ${}_4\phi_3$ series which corresponds to the sum over s in (2.7b) (related to the most symmetric expression for $q-6j$ coefficient) in analogy with (4.8) of Ref. 18 using Watson's formula (2.5.1) of Ref. 27 as presented by (6.10) of Ref. 28 with parameters

$$a\to b'-b+2v+1,\ \ b\to s'+1,\ \ c\to -b+v+2,$$

$$d\to n+1,\ \ e\to a'-b''-a-v,\ \ N\to b'',\ \ s\to\tilde{j}''+\tilde{z}''.$$

Further we replace some factors in (2.7b) as follows:

$$\frac{[b-b'-2v+s'+n]!}{[-v+s']![b-b'-v+n]![s]![b-b'-2v+s'+n-s]!}$$
$$=\sum_{s''}\frac{q^{s''(b-b'-2v+s'+n)-s(s'-v)}}{[s'']![s'-v-s'']![s-s'']![b-b'-v+n-s+s'']!}$$

and use the Chu–Vandermonde formulas[25–27] for summation over s,s'.

Finally after change of summation parameters we obtain the new expression for the denominator function

$$\mathcal{D}^2\binom{q,t=1}{a''b''}[a'b';a\,b]$$
$$=q^{-(a'+b-2v+2)(a'+a''-a-v)-v(b'+v+1)+(b''+v)(b'+b+2)}$$
$$\times\frac{[a'']![b'']![a+b'-a''+v+1]![a'-b''+b-v+1]!}{[a'+a''-v+1]![b'+b''+v+1]![a+b''+v+1]![a''+b-v+1]!}$$
$$\times\sum_{n,n'}\frac{q^{n(a'+a+2)-n'(b'+b+2)}[b-b''-v+n]![b+a''-v+n+1]!}{[n]![a-a'+v+n]![n']![b'-b+v+n']![b''-n-n']![b''+v-n-n']!}$$
$$\times\frac{[a'-b''-v+n']![a'+a''-v+n'+1]!}{[a+b'-a''-b''+v+n+n'+1]![a+b'+v+n+n'+2]!},\tag{2.8}$$

which is indefinite only if both

$$b-b''-v<0\quad\text{and}\quad b'-a''+v<0,$$

or

$$a'-b''-v<0\quad\text{and}\quad a-a''+v<0,$$

i.e., in the null space situation, when isofactors (2.1b) are vanishing.

We see that the expression for denominator function (2.8) satisfies the symmetry relations

$$\mathcal{D}^2\binom{q,t=1}{a''b''}[a'b';a\,b]=\mathcal{D}^2\binom{q^{-1},t=1}{b''a''}[b'a';b\,a]\tag{2.9a}$$

$$=q^{b'(a'-b''+b-v+2)-a''(b'-a''+a+v+2)}\mathcal{D}^2\binom{q^{-1},t=1}{a''b''}[ba;b'a']\tag{2.9b}$$

$$=\frac{q^{v(3v-a'-b')}[a'']![b'']!}{[a''-v]![b''+v]!}\mathcal{D}^2\binom{q,t=1}{a''-v,b''+v}[a'-v,b'+v;a+v,b-v]\tag{2.9c}$$

$$=\frac{q^{(a'-a-v)(a+b'-a''-b''+2)}[a'']![b'']!}{[a'+a''-a-v]![a-a'+b''+v]!}$$
$$\times\mathcal{D}^2\binom{q^{-1},t=1}{a'+a''-a-v,a-a'+b''+v}[a+v,b';a'-v,b]\tag{2.9d}$$

$$=\frac{q^{(b-b'-v)(a+b'-a''+v+2)}[a'']![b'']!}{[a''-b'+b-v]![b'+b''-b+v]!}$$
$$\times\mathcal{D}^2\left(\begin{matrix}q^{-1},t=1\\ a''-b'+b-v,b'+b''-b+v\end{matrix}\right)[a',b-v;a,b'+v]. \tag{2.9e}$$

Relations (2.9a) and (2.9b) coincide with (3.7a) and (3.14a) of Ref. 12, when the Regge-type symmetry relations (2.9c), (2.9d), and (2.9e) correspond to transposition of layers, rows, and columns, respectively, in subarray (1.5b) of Ref. 12, characterizing the multiplicity of the canonical tensor operators. Substitution $q\to q^{-1}$ in (2.9a) and (2.9b) is correlated with transposition of the whole array (1.3b) (interchange of its rows and columns). Hence, the symmetry group of denominator function (2.8) includes 16 elements. In analogy with Refs. 4 and 21 for fixed integer b'' we may re-express the denominator function,

$$\mathcal{D}^2\left(\begin{matrix}q,t=1\\ a''b''\end{matrix}\right)[a'b';a\,b]=\frac{q^{-(a'+b-2v+2)(a'+a''-a-v)-v(b'+v+1)+(b''+v)(b'+b+2)}}{[b'+b''+v+1]![a+b''+v+1]![b''+v]!}$$
$$\times\frac{[a'']![a'-b''+b-v+1]![b-b''-v]![a'-b''-v]!}{[a-a'+b''+v]![b'+b''-b+v]![a+b'+b''+v+2]!}\mathbf{G}^1_{b''}(\Delta;x)_q \tag{2.10}$$

in terms of q-polynomials,

$$\mathbf{G}^1_{b''}(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_q$$
$$=\sum_{n,n'}\frac{q^{n(a'+a+2)-n'(b'+b+2)}}{[n]![n']![b''-n-n']!}$$
$$\times[a-a'+b''+v|q]_{b''-n}(b-b''-v+1|q)_n(b+a''-v+2|q)_n$$
$$\times[b''+v|q]_{n+n'}[a+b'+b''+v+2|q]_{b''-n-n'}$$
$$\times[a+b'-a''+v+1|q]_{b''-n-n'}$$
$$\times[b'+b''-b+v|q]_{b''-n'}(a'-b''-v+1|q)_{n'}(a'+a''-v+2|q)_{n'} \tag{2.11a}$$

$$=[a-a'+b''+v|q]_{b''}[b''+v|q]_{b''}[b'+b''-b+v|q]_{b''}$$
$$\times\sum_{n,n'}q^{n(a'+a+2)-n'(b'+b+2)}\frac{(b-b''-v+1|q)_n(b+a''-v+2|q)_n}{[n]!(a-a'+v+1|q)_n}$$
$$\times\frac{(-a-b'-b''-v-2|q)_{b''-n-n'}(-a-b'+a''-v-1|q)_{b''-n-n'}}{[b''-n-n']!(v+1|q)_{b''-n-n'}}$$
$$\times\frac{(a'-b''-v+1|q)_{n'}(a'+a''-v+2|q)_{n'}}{[n']!(b'-b+v+1|q)_{n'}}, \tag{2.11b}$$

where

$$x_1=b'+1,\;\; x_2=-a'-b'-2,\;\; x_3=a'+1,$$
$$\Delta_1=a-a'+b''+v,\;\; \Delta_2=b''+v,\;\; \Delta_3=b'+b''-b+v. \tag{2.12}$$

The symmetry properties of the polynomials $G^{t}_{b''}(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)$ were quite important for the proof of the uniqueness of solution in Ref. 4. Let us consider the symmetries of $\mathbf{G}^{1}_{b''}(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_q$. The evident symmetry properties

$$\mathbf{G}^{1}_{b''}(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_q=\mathbf{G}^{1}_{b''}(\Delta_3,\Delta_2,\Delta_1;x_3,x_2,x_1)_{q^{-1}} \tag{2.13a}$$

$$=q^{b''(b'+b+2)}\mathbf{G}^{1}_{b''}(\Delta_1,\Delta_3,\Delta_2;x_1,x_3,x_2)_{q^{-1}} \tag{2.13b}$$

$$=q^{-b''(a'+a+2)}\mathbf{G}^{1}_{b''}(\Delta_2,\Delta_1,\Delta_3;x_2,x_1,x_3)_{q^{-1}} \tag{2.13c}$$

$$=q^{-b''(b'+b+2)}\mathbf{G}^{1}_{b''}(\Delta_2,\Delta_3,\Delta_1;x_2,x_3,x_1)_{q} \tag{2.13d}$$

$$=q^{b''(a'+a+2)}\mathbf{G}^{1}_{b''}(\Delta_3,\Delta_1,\Delta_2;x_3,x_1,x_2)_{q} \tag{2.13e}$$

(with a'',b'' fixed) correspond to substitutions

$$a'\leftrightarrow b,\ \ b'\leftrightarrow a;$$

$$a'\to -a-b-3,\ \ b'\leftrightarrow b,\ \ a\to -a'-b'-3,\ \ v\to b'-b+v;$$

$$a'\leftrightarrow a,\ \ b'\to -a-b-3,\ \ b\to -a'-b'-3,\ \ v\to a-a'+v;$$

$$a'\to b'\to -a'-b'-3,\ \ a\to b\to -a-b-3,\ \ v\to b'-b+v;$$

$$b'\to a'\to -a'-b'-3,\ \ b\to a\to -a-b-3,\ \ v\to a-a'+v;$$

respectively, or to the row permutations in array A_1,

$$A_1(\Delta;x)=\begin{bmatrix}\Delta_1 & \Delta_2+x_1 & \Delta_3-x_1\\ \Delta_2 & \Delta_3+x_2 & \Delta_1-x_2\\ \Delta_3 & \Delta_1+x_3 & \Delta_2-x_3\end{bmatrix} \tag{2.14a}$$

$$=\begin{bmatrix}a-a'+b''+v & b'+b''+v+1 & b''-b+v-1\\ b''+v & b''-a'-b+v-2 & b'+b''+a+v+2\\ b'+b''-b+v & a+b''+v+1 & b''-a'+v-1\end{bmatrix} \tag{2.14b}$$

[see Eq. (1.6) of Ref. 4]. Relation (2.13a) follows directly from (2.9b), but we failed to find such definition of q-polynomial which allowed to escape additional q-phases in other symmetry relations.

In order to check the symmetry of q-polynomial (2.11a) with respect to the column permutations and transpositions of array A_1, (and to find the corresponding q-phases), we use in (2.7b) the rearranged sums over s,

$$\sum_s \frac{[a''+b''+b-v-s+1]!}{[s]![b''-s]![a-a'+b''+v-s]![b-b'-2v+s'+n-s]!}$$
$$\times\frac{(-1)^{b-b'+s}}{[b-v-s]![b'-b+2v-s'+s]![b'-b+2v-n+s]!}$$
$$=\frac{[a''+b''+b'+v-s'-n+1]!}{[b'+v-s']![b'+v-n]![a'+a''-a-v-s']!}$$

$$\times\frac{[b'+b''+v+1]!}{[a'+a''-a-v-n]![a-a'+b''+v]![b-b'-2v+s'+n]!}$$

$$\times\sum_{s}(-1)^{s'+n+s}\frac{[a'+a''-a-v-s]![b'+v-s]![a-a'+b''+v+s]!}{[s]![n-s]![s'-s]![a''-s'-n+s]![b'+b''+v-s+1]!} \tag{2.15a}$$

$$=\frac{[b'+b''+v+1]![a''+b-v+1]!}{[b-b'-2v+s'+n]![b-v]![n]![b'+v-s']![a''-n]![a'+a''-a-v-n]!}$$

$$\times\sum_{s}(-1)^{a''-s}\frac{[a''+b-v-s'-s]![b-b'+a''-2v+n-s]!}{[s]![a''-s'-s]![a-a'+b''+v-s]!}$$

$$\times\frac{[a'+a''-a-v-n+s]!}{[a'-a-v+s]![a''+b-v-s+1]!}, \tag{2.15b}$$

in accordance with alternative expressions and the Regge symmetry of $q-6j$ coefficients.[25,26] Hence instead of (2.8) we obtain

$$\mathcal{D}^2\begin{pmatrix} q,t=1 \\ a''b'' \end{pmatrix}[a'b';a\,b]$$

$$=\frac{q^{(a+b'+2v+2)(a'+a''-a-v)-v(b'+v+1)-a''(a+b+2)}[a'']![b'']!}{[b''+v]![b'+b''-b+v]![a-a'+b''+v]![a'+a''-v+1]![a''+b-v+1]!}$$

$$\times\sum_{s,s''}(-1)^{a''-s''}\frac{q^{s''(a+b-a'+1)-s(a+b'+2)}[b'+v-s]![a-a'+b''+v+s]!}{[s]![a'+a''-a-v-s]![s''-s]![a''-b'+b-v+s-s'']!}$$

$$\times\frac{[a+v+s-s'']![b'+b''-b+v+s''-s']![a''+b''+v-s'']!}{[-v+s'']![a''-s'']![a'+a''+b-v-s''+2]!} \tag{2.16a}$$

$$=q^{(a+b'+v+2)(a'+a''-a-v)+vb''-a''(b'+b''+a+v+2)}$$

$$\times\frac{[a'']![b'']![a+b'-a''+v+1]!}{[b'+b''+a+v+2]!}$$

$$\times\sum_{s,n}\frac{q^{s(b+b''+a+v+2)-n(a+b'-a''+v+2)}[b'+v-n]![a-a''+v+n]!}{[n]![a''-n]![a'+a''-a-v-n]![b-b'-v+n]![s]![a''-v-s]!}$$

$$\times\frac{[a-a'+b''+v+n-s]![a'+a''-a-v-n+s]!}{[a-a'+b''+v-s]![a'-a-v+s]![a'+s+1]![a''+b-v-s+1]!}, \tag{2.16b}$$

in the first case replacing some factors in (2.7b) with inserted (2.15a) by

$$\sum_{s''}(-1)^{a''-n+s-s''}\frac{q^{(s''-s')(b'+a''+b''+v-s'-n+1)-(a''-s'-n+s)(a'+a''+b-v-s'+2)}[a+v+s-s'']!}{[s''-s']![a''-n+s-s'']![a'+a''+b-v-s''+2]!}$$

and using the Chu–Vandermonde formulas[25–27] for summation over s,s', and in the second case, using the Chu–Vandermonde formula, after inserting (2.15b), for summation over s'. Expression (2.16a) [with factor $q^{b''(a-b')}$] may be also derived from (2.8), applying (2.9a) and using interchange of the first and third column of array (2.14b), together with transition $q\to q^{-1}$.

Otherwise, expression (2.16b) is an analog of (3.14b) or (3.7b) of Ref. 12, but in contrast with these two formulas, all its terms are positive and (in all non-null-space situations) definite. Besides it is invariant (up to an elementary factor) with respect to transposition of array A (2.14b). In the SU(3) $(q=1)$ case (2.16b) may be rearranged replacing some ratios of factorials by the Chu–Vandermonde formulas in analogy with (3.17) of Ref. 13. This way the SU(3) denominator function may be written as follows:

$$\mathcal{D}^2\begin{pmatrix} t=1 \\ a''b'' \end{pmatrix}(a'b';a\,b)$$
$$=\frac{a''!b''!(a+b'-a''+v+1)!(a'-b''+b-v+1)!}{(a'+a''+b-v+2)!(b'+b''+a+v+2)!}\sum_{n,n'}\frac{(b-b''-v+n)!}{n!(a-a'+v+n)!}$$
$$\times\frac{(a'-b''-v+n')!}{(b'+b''+v-n+1)!n'!(b'-b+v+n')!(b''-n-n')!(b''+v-n-n')!}$$
$$\times\frac{(a'+a''+b''+b-v-n-n'+2]!}{(a+b''+v-n'+1)!(a+b'-a''-b''+v+n+n'+1)!} \tag{2.17}$$

accepting the form related to less symmetric form (5.6) of $\mathcal{G}^1_{b''}(A)$ function.[4] Although expression (2.17) resembles Eq. (2.8) (the majority of factorial arguments, dependent on summation parameters, coinciding in the both versions), the separate sums in (2.17) are related to the balanced hypergeometric functions, which q-extension cannot include q-phases dependent on summation parameters, in contrast with (2.8).

We may apply also the q-Saalschütz identity,

$$\sum_r \frac{(a|q)_r(b|q)_r}{[r]!(c|q)_r}\frac{(c-a-b|q)_{s-r}}{[s-r]!}=\frac{(c-a|q)_s(c-b|q)_s}{(c|q)_s[s]!}, \tag{2.18}$$

as presented in Refs. 29, 30, separately to each row of the q-polynomial (2.11b) depending on n, $b''-n-n'$, and n', respectively, and perform the summation over n,n', using the Chu–Vandermonde formulas. This way we obtain a new expression

$$\mathbf{G}^1_{b''}(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_q$$
$$=[a-a'+b''+v|q]_{b''}[b''+v|q]_{b''}[b'+b''-b+v|q]_{b''}$$
$$\times\sum_{s,s'} q^{s(a'+a+2)-s'(b'+b+2)}\frac{(-b'-b''-v-1|q)_s(a''-b'-v|q)_s}{[s]!(a-a'+v+1|q)_s}$$
$$\times\frac{(a'-b''+b-v+2|q)_{b''-s-s'}(a'+a''+b-v+3|q)_{b''-s-s'}}{[b''-s-s']!(v+1|q)_{b''-s-s'}}$$
$$\times\frac{(-a-b''-v-1|q)_{s'}(a''-a-v|q)_{s'}}{[s']!(b'-b+v+1|q)_{s'}}, \tag{2.19}$$

which corresponds to permutation of the second and third column of array A_1 (2.14b), together with transition $q\to q^{-1}$. Thus, the determinantal symmetry of the q-polynomial $\mathbf{G}^1_{b''}(\Delta;x)_q$ with 72 elements (in terms of the transpositions of the elements of array A_1) is demonstrated.

Otherwise, the symmetry relations (2.9c), (2.9d), and (2.9e) with 8 elements [in terms of the transpositions of the elements of subarray (1.5b) of Ref. 12] correspond to the reduction formulas of G functions.[3,4]

III. CONJECTURE ABOUT GENERIC DENOMINATOR FUNCTIONS

In Refs. 3 and 4, it was demonstrated, that the ratio $G^{t-1}_{b''}(\Delta;x)/G^t_{b''}(\Delta;x)$ of subsequent polynomials (4.2) or (5.6) of Ref. 4 (with $G^0_{b''}=G^{b''+1}_{b''}=1$) appears, together with the definite linear factors, under the square root sign in the denominator (normalization) function $D(\Gamma_t;x)$ of the SU(3) canonical tensor operators. For the denominator function of the $u_q(3)$ canonical tensor operators, the corresponding linear factors turn into the definite q-numbers, together with the definite q-phase factors, but q-extension of expression (4.2) or (5.6) of Ref. 4 is meaningless. We introduce here the following q-analog of the denominator polynomial,

$$
\begin{aligned}
&\mathbf{G}^t_{b''}(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_q\\
&=\prod_{s=1}^{t}\frac{[b''-s+1]!}{[s-1]!}\sum_{\lambda,\mu,\nu}\frac{g(\lambda,\mu,\nu;[b''-t+1]^t)}{M[\lambda]M[\mu]M[\nu]}q^{(a'+a+2)\Sigma_s\lambda_s-(b'+b+2)\Sigma_s\mu_s}\\
&\quad\times\prod_{s=1}^{t}(x_1-\Delta_3+t-s|q)_{\lambda_s}(x_1-\Delta_3+a''+b''+2-s|q)_{\lambda_s}[\Delta_1+1-s|q]_{b''-t+1-\lambda_s}\\
&\quad\times[\Delta_1-x_2-t+s|q]_{\nu_s}[\Delta_1-x_2-a''-b''-2+s|q]_{\nu_s}[\Delta_2+1-s|q]_{b''-t+1-\nu_s}\\
&\quad\times(x_3-\Delta_2+t-s|q)_{\mu_s}(x_3-\Delta_2+a''+b''+2-s|q)_{\mu_s}[\Delta_3+1-s|q]_{b'-t+1-\mu_s}
\end{aligned}\tag{3.1a}
$$

$$
\begin{aligned}
&=\prod_{s=1}^{t}\frac{[b''-s+1]!}{[s-1]!}\prod_{i=1}^{3}\prod_{s=1}^{t}[\Delta_i+1-s|q]_{b''-t+1}\\
&\quad\times\sum_{\lambda,\mu,\nu}g(\lambda,\mu,\nu;[b''-t+1])q^{(a'+a+2)\Sigma_s\lambda_s-(b'+b+2)\Sigma_s\mu_s}\\
&\quad\times\langle{}_2\mathcal{F}_1[b-b''-v+t,a''+b-v+2;a-a'+v+t]|\lambda\rangle_q\\
&\quad\times\langle{}_2\mathcal{F}_1[-a-b'-b''-v+t-3,-a-b'+a''-v-1;v+t]|\nu\rangle_q\\
&\quad\times\langle{}_2\mathcal{F}_1[a'-b''-v+t,a'+a''-v+2;b'-b+v+t]|\mu\rangle_q
\end{aligned}\tag{3.1b}
$$

in terms of variables (2.12), where $\lambda=[\lambda_1,\lambda_2,\ldots,\lambda_t]$, μ, ν are the partitions which may denote irreps of $\mathrm{U}(t)$. Coefficients $g(\lambda,\mu,\nu;[b''-t+1]^t)$ are the Littlewood–Richardson numbers, i.e., they are equal to the multiplicity of the irrep $[b''-t+1,...,b''-t+1]$ ($b''-t+1$ repeated t times) in the direct product $\lambda\times\mu\times\nu$ decomposition. Factor $M[\lambda]$ corresponds to a q-analog of the measure of the Young frame λ,

$$
\begin{aligned}
M[\lambda]&=\frac{\Pi_{s=1}^t[\lambda_s+t-s]!}{\Pi_{r<s}[\lambda_r-\lambda_s+s-r]}=(\dim[\lambda])^{-1}\prod_{s=1}^{t}(t-s+1|q)_{\lambda_s},\\
\dim[\lambda]&=\frac{\Pi_{r<s}[\lambda_r-\lambda_s+s-r]}{[1]![2]!\cdots[t-1]!}.
\end{aligned}\tag{3.2}
$$

Generalized q-hypergeometric coefficients in (3.1b) are defined by

$$
\langle{}_2\mathcal{F}_1[a,b;c]|\lambda\rangle_q=M^{-1}[\lambda]\prod_{s=1}^{t}\frac{(a-s+1|q)_{\lambda_s}(b-s+1|q)_{\lambda_s}}{(c-s+1|q)_{\lambda_s}}\tag{3.3}
$$

(cf. Refs. 4, 29–31) and appear as expansion coefficients of generalized q-hypergeometric (Gauss) series

$$
{}_2\mathcal{F}_1[a,b;c;q;z]_{\pm}=\sum_{\lambda}\langle{}_2\mathcal{F}_1[a,b;c]|\lambda\rangle_q q^{\pm(a+b-c-t)\Sigma\lambda_s}e_\lambda(z)\tag{3.4}
$$

in terms of symmetric (Schur) functions $e_\lambda(z)$ with $z=(z_1,z_2,\dots,z_t)$.

Expression (3.1a) turn into (2.11a) for $t=1$ with $g(\lambda,\mu,\nu;[b''-t+1])=\delta_{\lambda_1+\mu_1+\nu_1,b''}$. Otherwise, for $t=b''$ we obtain

$$\begin{aligned}\mathbf{G}_{b''}^{b''}&(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_q\\&=\sum_{n,n'}\frac{q^{n(a'+a+2)-n'(b'+b+2)}}{[n]![n']![b''-n-n']!}\\&\quad\times(a-a'+v+1|q)_{b''-n}[b-v|q]_n[b+a''-v+2|q]_n\\&\quad\times(v+1|q)_{n+n'}(a+b'+v+3|q)_{b''-n-n'}(b'-a''+a+v+1|q)_{b''-n-n'}\\&\quad\times(b'-b+v+1|q)_{b''-n'}[a'-v|q]_{n'}[a'+a''-v+2|q]_{n'},\end{aligned}\tag{3.5}$$

since $g(\lambda,\mu,\nu;[1^{b''}])=1$ only if $\lambda=[1^n]$, $\mu=[1^{n'}]$, $\nu=[1^{b''-n-n'}]$ and vanishes otherwise. Function $\mathbf{G}_{b''}^{b''}$ is equivalent to the q-polynomial factor of the denominator function $\mathbf{D}^2\left({q,t=b''+1 \atop a''b''}\right)[a'b';ab]$, which may be derived from (4.8c) of Ref. 18 after interchange of layers of array (1.3b) of Ref. 12, i.e., applying to $\mathbf{D}^2\left({q,t=b''+v+1 \atop a''b''}\right)[a'b';ab]$ the substitution

$$\begin{aligned}&a'\leftrightarrow a'-v,\quad b'\leftrightarrow b'+v,\quad a''\leftrightarrow a''-v,\\&b''\leftrightarrow b''+v,\quad a\leftrightarrow a+v,\quad b\leftrightarrow b-v.\end{aligned}\tag{3.6}$$

In a similar manner as in Sec. IV of Ref. 4 we may derive for (3.1a) the reduction formulas and find the weight space $\mathbf{W}_{b''}^t(\Delta)$ of zeros. For example, when $v<0$ the separate factors

$$\begin{aligned}\prod_{s=1}^{t}&[a-a'+b''+v+1-s|q]_{-v}[b'+b''-b+v+1-s|q]_{-v}\\&\times[a+b'+s+2|q]_{-v}[a+b'-a''+v+s|q]_{-v}\end{aligned}\tag{3.7}$$

may be extracted from (3.1a), since the factor $[b''+v+1-s|q]_{b''-t+1-\nu_s}$ vanish unless $\nu_t\geqslant -v$. In this case the separate blocks of (3.1a) may be rearranged into a close form as (4.9) and (4.10) of Ref. 4, with new parameters, using the new summation partition $\nu_t'=[\nu_1+v,\nu_2+v,\dots,\nu_t+v]$, and taking into account the identity

$$\begin{aligned}&\frac{g(\lambda,\mu,\nu;[b''-t+1]^t)}{M[\nu]}\prod_{s=1}^{t}[b''-s+1]![b''+v+1-s|q]_{b''-t+1-\nu_s}\\&\quad=\frac{g(\lambda,\mu,\nu';[b''+v-t+1]^t)}{M[\nu']}\prod_{s=1}^{t}[b''+v-s+1]![b''+1-s|q]_{b''-t+1-\nu_s'}.\end{aligned}\tag{3.8}$$

Hence, for $t=b''+v$, $v<0$ the q-polynomial factor of (3.1a) accepts after rearrangement the form related to the corresponding q-polynomial factor of $\mathbf{D}^2\left({q,t=b''+v+1 \atop a''b''}\right)[a'b';ab]$, as presented by (4.8c) of Ref. 18. Analogical rearrangement of (3.1a) is also possible when $a-a'+v<0$ or $b'-b+v<0$.

We are still unable to prove the general expression (3.1a) for $1<t<\mathcal{M}$ in the generic $u_q(3)$, $q\neq 1$, case. As it is shown by Theorem 3.2 of Ref. 4 the polynomial $G_{b''}^t(\Delta;x)$ is the unique polynomial that possesses the following properties: (i) total degree $2t(b''-t+1)$ in x_1,x_2,x_3, (ii) determinantal symmetry, and (iii) the weight space $\mathbb{W}_{b''}^t(\Delta)$ of zeros. The properties (i) and (iii) for $\mathbf{G}_{b''}^t(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_{q=1}$ may be checked in a similar manner as it was done in Sec. IV of Ref. 4. Evidently, this polynomial is invariant under row interchange in array

$$A_t(\Delta;x)=\begin{bmatrix}\Delta_1-t+1 & \Delta_2+x_1-t+1 & \Delta_3-x_1-t+1\\ \Delta_2-t+1 & \Delta_3+x_2-t+1 & \Delta_1-x_2-t+1\\ \Delta_3-t+1 & \Delta_1+x_3-t+1 & \Delta_2-x_3-t+1\end{bmatrix}. \tag{3.9}$$

Otherwise, the generalized hypergeometric coefficients $\langle {}_2\mathcal{F}_1(a,b;c)|\lambda\rangle$, $\langle {}_2\mathcal{F}_1(a',b';c')|\mu\rangle$, and $\langle {}_2\mathcal{F}_1(a'',b'';c'')|\nu\rangle$ in (3.1b) for $q=1$ either may be expanded using the generalization of Saalschütz identity[29–31] separately, or their product may be rearranged using the following generalized version of (2.2) of Ref. 30,

$$\sum_{\lambda,\mu,\nu} g(\lambda\mu\nu;\Lambda)\langle {}_2\mathcal{F}_1(c-a,c-b;c)|\lambda\rangle\langle {}_2\mathcal{F}_1(c'-a',c'-b';c')|\mu\rangle$$
$$\times\langle {}_2\mathcal{F}_1(c''-a'',c''-b'';c'')|\nu\rangle$$
$$=\sum_{\lambda,\mu,\nu,\kappa} g(\lambda\mu\nu\kappa;\Lambda)\langle {}_2\mathcal{F}_1(a,b;c)|\lambda\rangle\langle {}_2\mathcal{F}_1(a',b';c')|\mu\rangle\langle {}_2\mathcal{F}_1(a'',b'';c'')|\nu\rangle$$
$$\times\langle {}_1\mathcal{F}_0(c-a-b+c'-a'-b'+c''-a''-b'')|\kappa\rangle, \tag{3.10}$$

where

$$\langle {}_1\mathcal{F}_0[d]\kappa\rangle_q=M^{-1}[\kappa]\prod_{s=1}^{t}(d-s+1|q)_{\kappa_s} \tag{3.11}$$

[taking the three factor version of generalized Euler identity (2.1) of Ref. 30 as generating function for relation (3.10)]. Since in the (3.1b) case $d=c-a-b+c'-a'-b'+c''-a''-b''=0$, factor $\langle {}_1\mathcal{F}_0(0)|\kappa\rangle=0$, unless κ is a trivial partition of 0. Hence Eq. (3.10) corresponds to the column 2–column 3 interchange of the array A_t. In the generic $q\neq 1$ case for this purpose it would be reasonable to use the q-generalized partition dependent Saalschütz identity,

$$\sum_{\mu,\nu} g(\mu\nu;\lambda)\langle {}_2\mathcal{F}_1[a,b;c]|\mu\rangle_q\langle {}_1\mathcal{F}_0[c-a-b]|\nu\rangle_q=\langle {}_2\mathcal{F}_1[c-a,c-b;c]|\lambda\rangle_q, \tag{3.12}$$

and the q-generalized partition dependent binomial identity (a multiple q-analog of the Chu–Vandermonde formula),

$$\sum_{\mu,\nu} g(\mu\nu;\lambda)q^{y\Sigma\mu_s-x\Sigma\nu_s}\langle {}_1\mathcal{F}_0[x]|\mu\rangle_q\langle {}_1\mathcal{F}_0[y]|\nu\rangle_q=\langle {}_1\mathcal{F}_0[x+y]|\lambda\rangle_q, \tag{3.13}$$

(cf. Refs. 29–31). Although we did not prove Eqs. (3.12) and (3.13) for generic λ and $q\neq 1$, it was demonstrated that coefficients $g(\mu\nu;\lambda)$ appear here and in the multiple series (3.1a)–(3.1b) as usual Littlewood–Richardson numbers rather than special q-dependent analogs of Littlewood–Richardson coefficients.[32–35]

Unfortunately, we had not proved the symmetry of polynomial

$$\mathbf{G}^{t}_{b''}(\Delta_1,\Delta_2,\Delta_3;x_1,x_2,x_3)_{q=1}$$

with respect of the transposition of array A. The attempts to perform the transformations which lead from (2.8) to (2.16b) in the generic $t>1$ case were unsuccessful.

IV. CONCLUDING REMARKS

We obtained completed and universal solutions for the normalized canonical seed isofactors of SU(3) and $u_q(3)$ with extreme values of multiplicity label, which may replace the conjectured algorithm (20) and other initial steps of computation algorithm[5] of the normalized SU(3) coupling coefficients. New expressions for the denominator function of the maximal null space case of the $u_q(3)$ canonical tensor operators, which cannot be derived in frames of the traditional SU(3) technique,[7–10] give also the new expressions and relations for the denominator function of the SU(3) canonical tensor operators as special function accepting the different forms. Nevertheless, similarly as algorithm (20) of Ref. 5 and extended for more than 20 years (beginning in 1972–1975[7,8,21,36]) resolution of the SU(3) denominator function problem before the final result[4,31] was derived, the present study of the normalization coefficients for the $u_q(3)$ canonical tensor operators in the generic multiplicity label case is inconclusive and search for generating functions related to multiple q-version of Euler identity and other techniques for investigation of multiple partition dependent q-factorial series need to be developed.

[1] L. C. Biedenharn, A. Giovannini, and J. D. Louck, J. Math. Phys. **8**, 691 (1967).
[2] J. D. Louck, Am. J. Phys. **38**, 3 (1970).
[3] L. C. Biedenharn, M. A. Lohe, and J. D. Louck, J. Math. Phys. **26**, 1458 (1985).
[4] J. D. Louck, L. C. Biedenharn, and M. A. Lohe, J. Math. Phys. **29**, 1106 (1988).
[5] J. P. Draayer and Y. Akiyama, J. Math. Phys. **14**, 1904 (1973).
[6] L. C. Biedenharn, R. A. Gustafson, M. A. Lohe, J. D. Louck, and S. Milne, in *Special Functions: Group Theoretical Aspects and Applications*, edited by R. A. Askey, T. H. Koornwinder, and W. Schempp (Reidel, Dordrecht, 1984), p. 129.
[7] L. C. Biedenharn, J. D. Louck, E. Chacón, and M. Ciftan, J. Math. Phys. **13**, 1957 (1972).
[8] L. C. Biedenharn and J. D. Louck, J. Math. Phys. **13**, 1985 (1972).
[9] W. J. Holman III, L. C. Biedenharn, and J. D. Louck, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **7**, 529 (1976).
[10] S. C. Milne, Adv. Math. **36**, 169 (1980).
[11] S. C. Milne, Adv. Math. **58**, 1 (1985).
[12] S. Ališauskas, J. Math. Phys. **37**, 5719 (1996).
[13] S. Ališauskas, J. Math. Phys. **33**, 1983 (1992).
[14] S. Ališauskas, J. Phys. A **30**, 4615 (1997).
[15] R. M. Asherova, J. P. Draayer, Yu. I. Kharitonov, and Yu. F. Smirnov, in *Quantum Group Symposium at Group 21* (Goslar 1996), edited by H.-D. Doebner and V. K. Dobrev (Heron, Sofia, 1996), p. 323.
[16] R. M. Asherova, J. P. Draayer, Yu. I. Kharitonov, and Yu. F. Smirnov, Found. Phys. **27**, 1035 (1997).
[17] R. M. Asherova, J. P. Draayer, Yu. I. Kharitonov, and Yu. F. Smirnov, Yad. Fiz. **61**, 1930 (1998).
[18] S. Ališauskas and J. P. Draayer, J. Phys. A **31**, 7461 (1998). Note that the phase factor in the r.h.s. of Eq. (3.17) should be corrected to $(-1)^{\tilde{j}''-\tilde{z}''}$.
[19] S. Ališauskas, J. Math. Phys. **29**, 2351 (1988).
[20] S. Ališauskas, J. Phys. A **28**, 985 (1995).
[21] J. D. Louck, M. A. Lohe, and L. C. Biedenharn, J. Math. Phys. **16**, 2408 (1975).
[22] Yu. F. Smirnov, V. N. Tolstoy, and Yu. I. Kharitonov, Yad. Fiz. **54**, 721 (1991); Sov. J. Nucl. Phys. **54**, 437 (1991).
[23] S. Ališauskas and Yu. F. Smirnov, J. Phys. A **27**, 5925 (1994).
[24] Note the several corrections of Ref. 14 presented in Ref. 18, especially of Eqs. (5.11)–(5.16). Besides the last a in Eq. (5.16) should be changed to a'', the phase factor in the r.h.s. of Eq. (3.1b) should be corrected to $(-1)^{\tilde{j}'+\tilde{z}'}$, and degree of q in the r.h.s. of Eq. (4.11) should be changed to opposite.
[25] A. P. Jucys and A. A. Bandzaitis, *Theory of Angular Momentum in Quantum Mechanics*, 2nd ed. (Mokslas, Vilnius, 1977) (in Russian).
[26] R. M. Asherova, Yu. F. Smirnov, and V. N. Tolstoy, Yad. Fiz. **59**, 1859 (1996); Czech. J. Phys. **46**, 127 (1996).
[27] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Vol. 35 in *Encyclopedia of Mathematics and Its Applications*, edited by G.-C. Rota (Cambridge University Press, Cambridge, 1990).
[28] M. A. Lohe and L. C. Biedenharn, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **25**, 218 (1994).
[29] J. D. Louck and L. C. Biedenharn, J. Math. Anal. Appl. **59**, 423 (1977).
[30] L. C. Biedenharn and J. D. Louck, Adv. in Appl. Math. **9**, 447 (1988).
[31] L. C. Biedenharn, A. M. Bincer, M. A. Lohe, and J. D. Louck, Adv. in Appl. Math. **13**, 62 (1992).
[32] A. Lascoux, B. Leclerc, and J.-Y. Thibon, J. Math. Phys. **38**, 1041 (1998).
[33] B. Leclerc and J.-Y. Thibon, math.QA/9809122.
[34] A. N. Kirillov and M. Shimozono, math.QA/9803062.
[35] M. Shimozono and J. Weyman, math.QA/9804036.
[36] L. C. Biedenharn, M. A. Lohe, and J. D. Louck, in *Proceedings of the Fourth International Colloquium on Group Theoretical Methods in Physics*, Nijmegen, The Netherlands, 1975, edited by A. Janner, T. Jansen, and M. Boon (Springer, New York, 1976), p. 395.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Galilean wavelets: Coherent states of the affine Galilei group

J.-P. Antoine[a)]
Institut de Physique Théorique, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

I. Mahara[b)]
Institut de Physique Théorique, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium, and CTSPS, Clark Atlanta University, Atlanta, Georgia 30314

(Received 21 April 1999; accepted for publication 20 July 1999)

We derive Galilean wavelets, by which we mean coherent states of the affine Galilei group, that is, the Galilei group extended by independent space and time dilations. The construction follows a general method based on square integrable group representations, possibly modulo a subgroup, i.e., on a homogeneous space of the underlying group. We also examine the restriction to the Schrödinger subgroup, which contains only dilations that leave invariant the Schrödinger and the heat equations. © *1999 American Institute of Physics.* [S0022-2488(99)01111-1]

I. INTRODUCTION: SPACE–TIME WAVELETS

Wavelet analysis has become by now a widely used technique in signal and image processing. It is a particular time-frequency, or rather time-scale, method, which is specially efficient at detecting and analyzing nonstationary effects in signals, such as discontinuities, transients, etc. Mathematically speaking, wavelets are simply coherent states associated to affine groups of various dimensions.

However, for analyzing a signal with such a technique, which privileges prior knowledge, it is advisable to choose a wavelet that matches the characteristics of the signal as well as possible. For instance, analyzing a static signal requires only a static wavelet, that is, a coherent state associated to some symmetry group of space, containing dilations of some kind. On the contrary, in the case of a time-dependent signal (such as a movie or video sequence), one needs a *time-dependent* wavelet, in other words, a coherent state associated to a symmetry group of space–time.

Then the question arises, how does one construct wavelets adapted to a given setup? A possible answer is given by the general theory of coherent states developed in Refs. 1 and 2 and thoroughly described in Refs. 3 and 4. The construction proceeds in two steps. First, one chooses an appropriate group G of transformations applicable to the signal under consideration. Next one selects a square integrable representation U of G, possibly modulo a subgroup, if any. Then the corresponding coherent states (CS) are simply the elements of the orbit under U of a fixed vector η in the representation space.

In this paper, we will apply this approach to the construction of time-dependent wavelets. However, there are several possibilities for the group G, depending on the kinematics (i.e., relativity group) one chooses.

(i) *Kinematical or Euclidean wavelets:* This is the simplest case. G consists of space–time translations, space–time dilations (denoted a, a_0, respectively), space rotations (which reduce to reflections in one space dimension), and time reflection.

(ii) *Galilean wavelets:* G is the *affine Galilei group*, which combines the (extended) Galilei group with independent space and time dilations (a,a_0).

[a)]Electronic mail: antoine@fyma.ucl.ac.be
[b)]Present address: Faculté des Sciences, Université du Burundi, Bujumbura, Burundi.

(iii) *Schrödinger wavelets:* Here G consists of the Galilei group plus Schrödinger dilations, i.e., with the constraint $a_0 = a^2$ (that is, dilations that leave invariant the Schrödinger and the heat equations).[5,6]

(iv) *Relativistic wavelets: G* is the *Weyl–Poincaré group*, combining the Poincaré group with relativistic space–time dilations, characterized by the relation $a_0 = a$.

Kinematical wavelets have been introduced by Duval-Destin and Murenzi[7] for the analysis of simple motion, and later applied by Murenzi *et al.*[8] to various practical problems of motion tracking. Relativistic (Poincaré) wavelets have been constructed by Bohnké,[9] using mathematical work of Unterberger,[10] and also (in 1+1 dimensions), by Bertrand and Bertrand.[11] But the Galilean case has been largely undocumented up to now, maybe because of the complicated structure of the Galilei group.[12,13] Some results were obtained in Ref. 14, but a comprehensive study is still missing, except for the brief overview given in Ref. 15. This paper aims at filling this gap.

The general affine Galilei group and the Schrödinger subgroup are treated in Secs. II and III, respectively. For the convenience of the reader, we have briefly summarized in Appendices A and B the general construction method of coherent states associated to square integrable group representations, following Refs. 1–4, and the Mackey method of induced representations.

II. THE EXTENDED AFFINE GALILEI GROUP AND ITS UNITARY REPRESENTATIONS

A. The affine Galilei group

The kinematics of a free nonrelativistic physical particle is governed by its invariance under the action of the Galilei group,[12] which is a ten parameter group $\mathcal{G}_0$ of transformations of Newtonian space–time (we work in 3 space dimensions, although the whole discussion extends to n dimensions, for any $n \geqslant 3$). An element $g \in \mathcal{G}_0$ is of the form

$$g = (b, \mathbf{a}, \mathbf{v}, R), \quad b \in \mathbb{R}, \quad \mathbf{a}, \mathbf{v} \in \mathbb{R}^3, \quad R \in \mathrm{SO}(3), \tag{2.1}$$

where b is a time translation and $\mathbf{a}$ a spatial translation, $\mathbf{v}$ a velocity boost, and R a spatial rotation. The action of g on a space–time point $(\mathbf{x}, t)$ is given by $g(\mathbf{x}, t) = (\mathbf{x}', t')$, where

$$\mathbf{x}' = R\mathbf{x} + \mathbf{v}t + \mathbf{a}, \quad t' = t + b. \tag{2.2}$$

The group law of $\mathcal{G}_0$ is

$$(b, \mathbf{a}, \mathbf{v}, R)(b', \mathbf{a}', \mathbf{v}', R') = (b + b', \mathbf{a} + b'\mathbf{v} + R\mathbf{a}', \mathbf{v} + R\mathbf{v}', RR'). \tag{2.3}$$

The next step is to construct projective representations of $\mathcal{G}_0$ by the standard method of central extensions.[12]

In order to get Galilean wavelets, we have to replace $\mathcal{G}_0$ by the affine group obtained by combining $\mathcal{G}_0$ with independent space and time dilations. Thus our program should be, first, to define the semidirect product $\mathcal{G}_0 \rtimes \mathcal{D}_2$ of the pure Galilei group $\mathcal{G}_0$ with a two-dimensional dilation group $\mathcal{D}_2 \equiv (\mathbb{R}_*^+)^2 \simeq \mathbb{R}^2$, and then to construct projective representations of the resulting group as usual. However, a straightforward computation, given in Appendix C, shows that the only central extensions of $\mathcal{G}_0 \rtimes \mathcal{D}_2$ by $\mathbb{R}$ are of the form $\mathcal{G}_0 \rtimes G_{\mathrm{WH}}$, where G_{WH} is the Weyl–Heisenberg group, itself a central extension of $\mathcal{D}_2$. In particular, the central extension procedure fails to generate mass, as it does in the usual situation, where no dilations are considered.

The only alternative is to reverse the order of the two constructions:

(1) Take first a central extension of $\mathcal{G}_0$ by $\mathbb{R}$, which yields the extended Galilei group $\mathcal{G}^M$ corresponding to the extension parameter $M > 0$.

(2) Construct the semidirect product $\mathcal{G}_{\mathrm{aff}}^M = \mathcal{G}^M \rtimes \mathcal{D}_2$, the extended affine Galilei group.

However, as we shall see below, M should *not* be interpreted as mass, but rather as a mass unit or mass scale (the physical mass varies under dilations, whereas M is fixed).

As for the first step, the group $\mathcal{G}_0$ admits a one-parameter family of central extensions $\mathcal{G}^M$, indexed by $M>0$.[12,13] Writing a generic element of $\mathcal{G}^M$ as $g=(\theta,b,\mathbf{a};\mathbf{v};R)$, with $\theta\in\mathbb{R}$, the structure of $\mathcal{G}^M$ is defined by the multiplication law

$$
\begin{aligned}
gg' &= (\theta,b,\mathbf{a};\mathbf{v};R)(\theta',b',\mathbf{a}';\mathbf{v}';R') \\
&= (\theta+\theta'+\xi(g,g'),b+b',\mathbf{a}+b'\mathbf{v}+R\mathbf{a}';\mathbf{v}+R\mathbf{v}';RR'). \qquad (2.4)
\end{aligned}
$$

The quantity ξ: $\mathcal{G}_0\times\mathcal{G}_0\to\mathbb{R}$ is a *multiplier*, which up to equivalence (in the sense of cocycles), can be taken as

$$\xi(g,g')=M[\mathbf{v}\cdot R\mathbf{a}'+\tfrac{1}{2}b'\mathbf{v}^2],\quad M=\text{const}>0. \qquad (2.5)$$

Next we introduce separate space ($a\equiv e^{\sigma}$) and time ($a_0\equiv e^{\tau}$) dilations, corresponding to the following action on space–time, which replaces (2.2),

$$\mathbf{x}'=e^{\sigma}R\mathbf{x}+e^{\tau}\mathbf{v}t+\mathbf{a},\quad t'=e^{\tau}t+b,\quad \sigma,\tau\in\mathbb{R}. \qquad (2.6)$$

Then the multiplication law of $\mathcal{G}^M_{\text{aff}}$ reads, with $g\equiv(\theta,b,\mathbf{a};\mathbf{v};R,\sigma,\tau)$,

$$
\begin{aligned}
gg' &= (\theta+e^{2\sigma-\tau}\theta'+M[e^{\sigma}\mathbf{v}\cdot R\mathbf{a}'+\tfrac{1}{2}e^{\tau}\mathbf{v}^2b'],b+e^{\tau}b',\mathbf{a}+e^{\tau}b'\mathbf{v}+e^{\sigma}R\mathbf{a}'; \\
&\times \mathbf{v}+e^{\sigma-\tau}R\mathbf{v}';RR',\sigma+\sigma',\tau+\tau'). \qquad (2.7)
\end{aligned}
$$

The inverse of an element $g\in\mathcal{G}^M_{\text{aff}}$ is given by

$$
\begin{aligned}
(\theta,b,\mathbf{a};\mathbf{v};R;\sigma,\tau)^{-1} &= (-e^{\tau-2\sigma}\theta+Me^{\tau-2\sigma}[\mathbf{v}\cdot\mathbf{a}-\tfrac{1}{2}b\mathbf{v}^2],-e^{-\tau}b,e^{-\sigma}R^{-1}(b\mathbf{v}-\mathbf{a}); \\
&-e^{\tau-\sigma}R^{-1}\mathbf{v};R^{-1},-\sigma,-\tau). \qquad (2.8)
\end{aligned}
$$

From this law, we note that $\mathcal{G}^M_{\text{aff}}$ is a semidirect product,

$$\mathcal{G}^M_{\text{aff}}=\mathcal{T}\rtimes(\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2)), \qquad (2.9)$$

where $\mathcal{T}=\{(\theta,b,\mathbf{a};\mathbf{0};I,0,0)\}\sim\mathbb{R}^5$ and $\mathcal{V}=\{(0,0,\mathbf{0};\mathbf{v};I,0,0)\}\sim\mathbb{R}^3$ are two Abelian subgroups of $\mathcal{G}^M_{\text{aff}}$.

In addition, the center of $\mathcal{G}^M_{\text{aff}}$ is trivial, which shows indeed that the latter cannot be a central extension. However, it is in fact an extension of $\mathcal{G}_0\rtimes\mathcal{D}_2$, but a noncentral one.

B. Orbits under $\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2)$ in $\mathcal{T}^*$

Since $\mathcal{G}^M_{\text{aff}}$ is a semidirect product of the form $G=\mathcal{T}\rtimes H$, with $\mathcal{T}$ Abelian and $H=\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2)$, its unitary irreducible representations (UIR) may be constructed by the Mackey method of induced representations. As described in Appendix B, the first step of the scheme is to compute the orbits under H in $\mathcal{T}^*$, the dual of $\mathcal{T}$.

According to (2.7), the action of $(\mathbf{v};R,\sigma,\tau)\in\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2)$ on the element $(\theta,b,\mathbf{a})\in\mathcal{T}$ is given by

$$(\mathbf{v};R,\sigma,\tau)(\theta,b,\mathbf{a})=(e^{2\sigma-\tau}\theta+M[e^{\sigma}\mathbf{v}\cdot R\mathbf{a}+\tfrac{1}{2}e^{\tau}b\mathbf{v}^2],e^{\tau}b,e^{\tau}b\mathbf{v}+e^{\sigma}R\mathbf{a}). \qquad (2.10)$$

Let us denote by $(q,E,\mathbf{p})$ the generic element of $\mathcal{T}^*$. The action of $(q,E,\mathbf{p})$ on $(\theta,b,\mathbf{a})$ is defined by

$$\langle q,E,\mathbf{p}|\theta,b,\mathbf{a}\rangle=q\theta+Eb-\mathbf{p}\cdot\mathbf{a}, \qquad (2.11)$$

where $\mathbf{p}\cdot\mathbf{a}$ denotes the Euclidean scalar product in $\mathbb{R}^3$. According to (B1) and (2.10), the action of $(\mathbf{v};R,\sigma,\tau)\in H$ on $(q,E,\mathbf{p})$ is defined by

$$\langle(\mathbf{v};R,\sigma,\tau)(\theta,b,\mathbf{a})|(q,E,\mathbf{p})\rangle=\langle(\theta,b,\mathbf{a})|(\mathbf{v};R,\sigma,\tau)^{-1}(q,E,\mathbf{p})\rangle. \tag{2.12}$$

A straightforward calculation then gives

$$(\mathbf{v};R,\sigma,\tau)(q,E,\mathbf{p})=(e^{\tau-2\sigma}q,e^{-\tau}E+e^{-\sigma}R\mathbf{p}\cdot\mathbf{v}+\tfrac{1}{2}qMe^{\tau-2\sigma}\mathbf{v}^2,e^{-\sigma}R\mathbf{p}+e^{\tau-2\sigma}qM\mathbf{v}). \tag{2.13}$$

These relations, which are a part of the coadjoint action of $\mathcal{G}_{\rm aff}^M$,[15] show that the real mass parameter is not M, but rather the combination qM, and it varies under dilations, as it should.

The determination of the orbits under $H=\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2)$ in $\mathcal{T}^*$ is equivalent to solving the following problem. Given two elements $(q,E,\mathbf{p})$, $(q',E',\mathbf{p}')$ in $\mathcal{T}^*$, does there exist an element $(\mathbf{v};R,\sigma,\tau)$ in $\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2)$ such that $(\mathbf{v};R,\sigma,\tau)(q,E,\mathbf{p})=(q',E',\mathbf{p}')$? From (2.13), answering this question amounts to solving the system

$$\begin{aligned} q'&=e^{\tau-2\sigma}q,\\ E'&=e^{-\tau}E+e^{-\sigma}\mathbf{v}\cdot R\mathbf{p}+\tfrac{1}{2}e^{\tau-2\sigma}qM\mathbf{v}^2,\\ \mathbf{p}'&=e^{-\sigma}R\mathbf{p}+e^{\tau-2\sigma}qM\mathbf{v}, \end{aligned} \tag{2.14}$$

in the unknown $(\mathbf{v};R,\sigma,\tau)$. From (2.14) it is clear that the sign of q is invariant on each orbit. Then a straightforward calculation yields

$$e^{\tau}=\frac{q'}{q}e^{2\sigma},\quad q\left(E-\frac{\mathbf{p}^2}{2qM}\right)=e^{2\sigma}q'\left(E'-\frac{\mathbf{p}'^2}{2q'M}\right),\quad \mathbf{v}=\frac{1}{q'M}(\mathbf{p}'-e^{-\sigma}R\mathbf{p}). \tag{2.15}$$

Therefore, the sign of $E-(\mathbf{p}^2/2qM)$ is also invariant on each orbit. All the elements $(q,E,\mathbf{p})$ such that $E-(\mathbf{p}^2/2qM)>0$ (resp. $E-(\mathbf{p}^2/2qM)<0$) are on the same orbit, which has the same dimension as $\mathcal{T}$, namely 5. Similarly, the elements $(q,E,\mathbf{p})$ such that $E-(\mathbf{p}^2/2qM)=0$ are on the same orbit, of dimension 4.

C. The UIR associated with the orbit $q>0$, $E-(\mathbf{p}^2/2qM)>0$

Let us consider in $\mathcal{T}^*$ the point $p_0\equiv(q_0,E_0,\mathbf{p}_0)=(1,1,\mathbf{0})$. The isotropy group of p_0 is SO(3). Thus the inducing subgroup is $\mathcal{K}=\mathcal{T}\rtimes\mathrm{SO}(3)$ and there is a one-to-one correspondence between $X\equiv\mathcal{G}_{\rm aff}^M/\mathcal{K}$ and the set $\{(e^{\tau-2\sigma},e^{-\tau}+\tfrac{1}{2}e^{\tau-2\sigma}M\mathbf{v}^2,e^{\tau-2\sigma}M\mathbf{v})\}$ with $(\mathbf{v};I,\sigma,\tau)\in\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2)$.

Since the isotropy group of p_0 is SO(3), the element Λ_p defined in (B3), that is,

$$\Lambda_p p_0=p,\ \text{ with }\ p\equiv(q,E,\mathbf{p}), \tag{2.16}$$

may be taken in the form $\Lambda_p=(\mathbf{v};I,\sigma,\tau)$. Thus we obtain the system

$$q=e^{\tau-2\sigma},\ \ E=e^{-\tau}+\tfrac{1}{2}e^{\tau-2\sigma}M\mathbf{v}^2,\ \ \mathbf{p}=e^{\tau-2\sigma}M\mathbf{v}, \tag{2.17}$$

whose solution reads

$$\mathbf{v}=\frac{\mathbf{p}}{qM},\ \ \tau=-\ln\left(E-\frac{\mathbf{p}^2}{2qM}\right),\ \ \sigma=-\tfrac{1}{2}\ln q\left(E-\frac{\mathbf{p}^2}{2qM}\right). \tag{2.18}$$

In this way, we have defined a section s: $X\to\mathcal{G}_{\rm aff}^M$ as required in (B3). Let us consider the element $(a,\Lambda)^{-1}(0,\Lambda_p)$ and denote by $(0,\Lambda_{p'})$ the representative of its class (here we have put $a\equiv(\theta,b,\mathbf{a})\in\mathcal{T}$ and $\Lambda\equiv(\mathbf{v};R,\sigma,\tau)\in H$). A straightforward calculation then yields

$$(a,\Lambda)^{-1}(0,\Lambda_p)=(0,\Lambda_{p'})\cdot k, \tag{2.19}$$

with

$$k^{-1}=(\Lambda_p^{-1}a,I)(0,R), \tag{2.20}$$

and

$$\begin{aligned} p'\equiv(q',E',\mathbf{p}')=\Lambda^{-1}p=(\mathbf{v};R,\sigma,\tau)^{-1}p \\ =(e^{2\sigma-\tau}q,e^{\tau}[E-\mathbf{p}\cdot\mathbf{v}+\tfrac{1}{2}qM\mathbf{v}^2],e^{\sigma}R^{-1}[\mathbf{p}-qM\mathbf{v}]). \end{aligned} \tag{2.21}$$

We may define a UIR of $\mathcal{K}$ by

$$L(k^{-1})=e^{i\langle p_0|\Lambda_p^{-1}a\rangle}D_j(R)=e^{i\langle p|a\rangle}D_j(R) \tag{2.22}$$

with D_j a UIR of SO(3). From (2.21) it is easy to check that

$$dq'dE'd^3\mathbf{p}'=e^{5\sigma}dqdEd^3\mathbf{p}. \tag{2.23}$$

The representation (2.22) of $\mathcal{K}$ then induces a UIR of $\mathcal{G}_{\text{aff}}^M$ defined according to (B6) by the following formula:

$$\begin{aligned} (U_{++}(g)\psi)(\Lambda_p)\equiv(U_{++}(\theta,b,\mathbf{a};\mathbf{v};R,\sigma,\tau)\psi)(\Lambda_p) \\ =e^{i(q\theta+Eb-\mathbf{p}\cdot\mathbf{a})}e^{5\sigma/2}D_j(R)\psi(\Lambda_{p'}), \end{aligned} \tag{2.24}$$

with p' given by (2.21). The representation space of U_{++} is $\mathfrak{H}_L\simeq L^2(X_{++},dqdEd^3\mathbf{p};\mathfrak{H}_j)$, where $X_{++}=\{(q,E,\mathbf{p}):q>0,E-\mathbf{p}^2/2qM>0\}$ and $\mathfrak{H}_j$ is the representation space of D_j.

D. Square integrability of the representation U_{++}

The left invariant measure on $\mathcal{G}_{\text{aff}}^M$ is given by the formula

$$d\mu_l(g)=e^{3\tau-8\sigma}d\theta\ db\ d^3\mathbf{a}\ d^3\mathbf{v}\ dm(R)d\sigma d\tau,\quad g\equiv(\theta,b,\mathbf{a};\mathbf{v};R,\sigma,\tau), \tag{2.25}$$

with $dm(R)$ the usual invariant measure on SO(3). The square integrability condition (A2) is satisfied if there exists a nonzero $\psi\in\mathfrak{H}_L$ such that

$$d(\psi)\equiv\int_{\mathcal{G}_{\text{aff}}^M}|\langle U_{++}(g)\psi|\psi\rangle|^2d\mu_l(g)<\infty. \tag{2.26}$$

In the sequel we consider for SO(3) only the trivial, spin 0, representation $D_j=I$, the generalization to others being straightforward. The square integrability condition (2.28) then reads

$$\begin{aligned} 0<\int_{\mathcal{G}_{\text{aff}}^M}\left\{\int_{X_{++}}e^{5\sigma/2}e^{-i(q\theta+Eb-\mathbf{p}\cdot\mathbf{a})}\overline{\psi((\mathbf{v};R,\sigma,\tau)^{-1}(q,E,\mathbf{p}))}\psi(q,E,\mathbf{p})dqdEd^3\mathbf{p}\right\} \\ \times\left\{\int_{X_{++}}e^{5\sigma/2}e^{i(\hat{q}\theta+\hat{E}b-\hat{\mathbf{p}}\cdot\mathbf{a})}\psi((\mathbf{v};R,\sigma,\tau)^{-1}(\hat{q},\hat{E},\hat{\mathbf{p}}))\overline{\psi(\hat{q},\hat{E},\hat{\mathbf{p}})}d\hat{q}d\hat{E}d^3\hat{\mathbf{p}}\right\} \\ \times e^{3\tau-8\sigma}d\theta\ db\ d^3\mathbf{a}\ d^3\ \mathbf{v}dm(R)d\sigma d\tau<\infty. \end{aligned} \tag{2.27}$$

After integration over θ, b, $\mathbf{a}$, condition (2.27) reduces to

$$0<\int e^{3(\tau-\sigma)}|\psi(q,E,\mathbf{p})|^2|\psi((\mathbf{v};R,\sigma,\tau)^{-1}(q,E,\mathbf{p}))|^2dqdEd^3\mathbf{p}\ d^3\mathbf{v}\ d\sigma d\tau dm(R)<\infty, \tag{2.28}$$

and, since SO(3) is compact, (2.28) holds if

$$\int e^{3(\tau-\sigma)}|\psi(q,E,\mathbf{p})|^2|\psi((\mathbf{v};R,\sigma,\tau)^{-1}(q,E,\mathbf{p}))|^2 dq dE d^3\mathbf{p}\ d^3\mathbf{v}\ d\sigma d\tau<\infty. \tag{2.29}$$

For the analysis of (2.29), let us perform the coordinate transformation $(q,E,\mathbf{p})\mapsto(\boldsymbol{\xi},I,\eta,\zeta)\in\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R}^2$ defined by

$$(q,E,\mathbf{p})=(\boldsymbol{\xi};I,\eta,\zeta)(1,1,\mathbf{0})=(e^{\zeta-2\eta},e^{-\zeta}+\tfrac{1}{2}Me^{\zeta-2\eta}\boldsymbol{\xi}^2,e^{\zeta-2\eta}M\boldsymbol{\xi}). \tag{2.30}$$

Since $(\mathbf{v};R,\sigma,\tau)^{-1}(\boldsymbol{\xi};I,\eta,\zeta)=(e^{\zeta-\eta}R^{-1}(\mathbf{v}-\boldsymbol{\xi});I,\sigma-\eta,\tau-\zeta)^{-1}(\mathbf{0};R^{-1},0,0)$, we may also define new coordinates $\mathbf{v}'$, σ', τ' by the formula

$$(e^{\zeta-\eta}R^{-1}(\mathbf{v}-\boldsymbol{\xi});I,\sigma-\eta,\tau-\zeta)^{-1}(\mathbf{0};R,0,0)^{-1}=(\mathbf{v}';I,\sigma',\tau')(\mathbf{0};R^{-1},0,0). \tag{2.31}$$

Taking inverses, we can write (2.31) in the form

$$e^{\zeta-\eta}(\mathbf{v}-\boldsymbol{\xi})=-e^{\tau'-\sigma'}\mathbf{v}',\quad \sigma=-\sigma'+\eta,\quad \tau=-\tau'+\zeta. \tag{2.32}$$

From the coordinate transformations defined by (2.30) and (2.32), we deduce directly

$$dq dE d^3\mathbf{p}=2M^3e^{3\zeta-8\eta}d\eta\ d\zeta\ d^3\boldsymbol{\xi}, \tag{2.33}$$

$$e^{3(\tau-\sigma)}d^3\mathbf{v}d\sigma d\tau=d^3\mathbf{v}'d\sigma'd\tau'. \tag{2.34}$$

The square integrability condition (2.29) then reads

$$\int e^{3\tau-8\eta}|\tilde{\psi}(\boldsymbol{\xi},\eta,\zeta)|^2 d^3\boldsymbol{\xi}\ d\eta\ d\zeta\int|\tilde{\psi}(\mathbf{v},\sigma,\tau)|^2 d^3\mathbf{v}\ d\sigma d\tau<\infty, \tag{2.35}$$

where we have written $\tilde{\psi}(\boldsymbol{\xi},\eta,\zeta)\equiv\psi(q,E,\mathbf{p})=\psi((\boldsymbol{\xi};I,\eta,\zeta)(1,1,\mathbf{0}))$. More generally, the square integrability condition

$$d(\psi,\phi)\equiv\int_{\mathcal{G}_{\mathrm{aff}}^M}|\langle U_{++}(g)\psi|\phi\rangle|^2 d\mu_l(g)<\infty,\quad \forall\phi\in\mathfrak{H}_L \tag{2.36}$$

is easily shown, using (2.33) and (2.18) again, to be equivalent to

$$d(\psi,\phi)\sim\|\phi\|^2\int|\tilde{\psi}(\mathbf{v};\sigma,\tau)|^2 d^3\mathbf{v}d\sigma d\tau=\|\phi\|^2\int_{X_{++}}\frac{|\psi(q,E,\mathbf{p})|^2}{q^4\left(E-\dfrac{\mathbf{p}^2}{2qM}\right)}dq dE d^3\mathbf{p}<\infty,\quad\forall\phi\in\mathfrak{H}_L. \tag{2.37}$$

There is obviously a dense set of functions ψ satisfying this condition, so that we obtain the following:

Proposition 1: The representation U_{++} defined by (2.26) is square integrable over the group $\mathcal{G}_{\mathrm{aff}}^M$, a vector $\psi\in\mathfrak{h}_L$ being admissible iff the last integral in (2.37) converges.

Of course, the same property holds for the irreducible representations associated to the other five-dimensional orbits, corresponding to $(E-(\mathbf{p}^2/2qM))<0$ and/or $q<0$. Since these are all the orbits which have nonzero Lebesgue measure in $\mathcal{T}^*$, we could have inferred the result from the general study of square integrability made by Aniello *et al.*[16] We refer the reader to the forthcoming monograph, Ref. 4, for a systematic analysis.

E. Galilean wavelets

According to Proposition 1, any admissible vector $\eta \in \mathfrak{H}_L$ generates a tight frame of Galilean wavelets, indexed by $\mathcal{G}_{\rm aff}^M$ itself and defined, as usual (see Appendix A) by $\eta_g = U_{++}(g)\eta$, $g \in \mathcal{G}_{\rm aff}^M$, with the expression (2.24) for the representation. These are of course CS of the Gilmore–Perelomov-type.

In addition, it is possible to construct sets of wavelets indexed by fewer parameters by taking quotients $Y=\mathcal{G}_{\rm aff}^M/K$ by various closed subgroups K and appropriate sections $s:Y\to\mathcal{G}_{\rm aff}^M$. The simplest example is obtained if we take for K the subgroup $\mathcal{T}_0 \simeq \mathbb{R}$ of time dilations $e^{-\tau}$. The corresponding coset space $Y=\mathcal{G}_{\rm aff}^M/\mathcal{T}_0$ is parametrized by points $y=(\theta,b,\mathbf{a};\mathbf{v};R,\sigma)$. We consider first the basic or Galilean section $s_0:Y\to\mathcal{G}_{\rm aff}^M$,

$$s_0(y)=(\theta,b,\mathbf{a};\mathbf{v};R,\sigma,0). \tag{2.38}$$

A straightforward calculation shows that the admissibility condition (A11) reduces to setting $E-(\mathbf{p}^2/2qM)=1$ (or a constant different from 0) in the integral (2.37), that is, a vector $\eta\in\mathfrak{H}_L$ is admissible mod$(\mathcal{T}_0,s_0)$ iff the following integral converges:

$$\int\int_{q>0} q^{-4}\left|\eta\left(q,1+\frac{\mathbf{p}^2}{2qM},\mathbf{p}\right)\right|^2 dq\,d^3\mathbf{p}<\infty. \tag{2.39}$$

We consider now a general section,

$$s_\beta(y)=s_0(y)(0,0,\mathbf{0};\mathbf{0};I;0,\beta(y))=(\theta,b,\mathbf{a};\mathbf{v};R,\sigma,\beta(y)), \tag{2.40}$$

where $\beta:Y\to\mathcal{T}_0\equiv\mathbb{R}$ is a Borel function and $\beta(y)\in\mathcal{T}_0$ represents a time dilation. Again this corresponds to a relation between the variables $(q,E,\mathbf{p})$, which may conveniently be written as $f_\beta(q,E,\mathbf{p})=0$. Thus the admissibility condition mod$(\mathcal{T}_0,s_\beta)$ reads

$$\int_{X_{++}} \frac{|\eta(q,E,\mathbf{p})|^2}{q^4\left(E-\dfrac{\mathbf{p}^2}{2qM}\right)}\,\delta(f_\beta(q,E,\mathbf{p}))\,dq\,dE\,d^3\mathbf{p}<\infty. \tag{2.41}$$

For any such admissible vector η, one obtains a dense set of CS, indexed by $Y=\mathcal{G}_{\rm aff}^M/\mathcal{T}_0$ and given by $\eta_{s_\beta(y)}=U_{++}(s_\beta(y))\eta$, where $U_{++}(s_\beta(y))$ is the representation (2.24) with τ replaced by $\beta(y)$ in (2.21). Of course, for $\beta(y)\equiv 0$, one recovers the Galilean section s_0.

An interesting example is given by $\beta(y)=2\sigma$, i.e., the Schrödinger case. The corresponding constraint relation is simply $q=1$, so that we get the admissibility condition,

$$\int\int_{E-\mathbf{p}^2/2M>0} \frac{|\eta(1,E,\mathbf{p})|^2}{E-\dfrac{\mathbf{p}^2}{2M}}\,dE\,d^3\mathbf{p}<\infty. \tag{2.42}$$

Notice that the q-integration has disappeared in (2.42), because inserting the factor $\delta(q-1)$ in (2.41) is equivalent to quotienting out the subgroup Θ of phase factors, in addition to $\mathcal{T}_0$. We will study this example in detail in the next section.

III. PROJECTIVE REPRESENTATIONS OF THE SCHRÖDINGER GROUP

Another possibility is to impose from the beginning the relation $\tau=2\sigma$, that is, to start from the Galilei–Schrödinger group $\mathcal{G}_S=\mathcal{G}_0\rtimes\mathcal{D}_S$, where $\mathcal{D}_S$ is the corresponding one-dimensional subgroup of $\mathcal{D}_2$, and then to construct projective representations of it, that is, to determine central extensions $\mathcal{G}_S^M$ of $\mathcal{G}_S$ by $\mathbb{R}$. The motivation for this restriction is that $\mathcal{G}_S$ is a natural invariance group of both the Schrödinger and the heat equations[5,6] (the full invariance group, called the

Schrödinger group, contains in addition the so-called *expansions*, the nonrelativistic analogs of pure conformal transformations). This restriction leads to several drastic modifications with respect to the case of $\mathcal{G}_{\text{aff}}^M$.

The group law of $\mathcal{G}_S^M$ reads

$$(\theta,b,\mathbf{a};\mathbf{v};R,\sigma)(\theta',b',\mathbf{a}';\mathbf{v}';R',\sigma') = (\theta+\theta'+M[e^{\sigma}\mathbf{v}\cdot R\mathbf{a}'+\tfrac{1}{2}e^{2\sigma}\mathbf{v}^2],b+e^{2\sigma}b', \mathbf{a}+e^{2\sigma}b'\mathbf{v}+e^{\sigma}R\mathbf{a}';\mathbf{v}+e^{\sigma}R\mathbf{v}';RR',\sigma+\sigma'). \tag{3.1}$$

As a consequence, $\mathcal{G}_S^M$ has a nontrivial center, namely the one-dimensional subgroup Θ of phase factors, isomorphic to $\mathbb{R}$. The group $\mathcal{G}_S^M$ is a semidirect product, with structure $\mathcal{G}_S^M = \mathcal{T}\rtimes(\mathcal{V}\rtimes \mathrm{SO}(3)\times\mathbb{R})$.

As shown in Appendix C, the extended Galilei–Schrödinger group $\mathcal{G}_S^M$ coincides with the subgroup of $\mathcal{G}_{\text{aff}}^M$ defined by the equation $\tau=2\sigma$. Therefore, $\mathcal{G}_S^M$ is also a semidirect product of $\mathcal{G}^M$ by $\mathcal{D}_S$, and we have, with $\lhd_M$ denoting a central extension with parameter M,

$$\mathcal{G}_S^M = \mathbb{R}\lhd_M\mathcal{G}_S \equiv \mathbb{R}\lhd_M(\mathcal{G}_0\rtimes\mathcal{D}_S) = \mathcal{G}^M\rtimes\mathcal{D}_S \equiv (\mathbb{R}\lhd_M\mathcal{G}_0)\rtimes\mathcal{D}_S\,.$$

Thus, in the Schrödinger case, the two operations $\lhd_M$ and $\rtimes$ commute.

Next we proceed with the construction of the induced UIRs of $\mathcal{G}_S^M$, exactly as in the previous sections. According to (2.15), the automorphisms induced in $\mathcal{T}^*$ by $\mathcal{V}\rtimes(\mathrm{SO}(3)\times\mathbb{R})$ are given by

$$(\mathbf{v};R,\sigma)(q,E,\mathbf{p}) = (q,e^{-2\sigma}E+e^{-\sigma}\mathbf{v}\cdot R\mathbf{p}+\tfrac{1}{2}qM\mathbf{v}^2,e^{-\sigma}R\mathbf{p}+qM\mathbf{v}). \tag{3.2}$$

Thus q is a constant on each orbit. Let us fix $q=q_0>0$. Then we have three cases to analyze,

$$E-\frac{\mathbf{p}^2}{2q_0M}<0,\quad E-\frac{\mathbf{p}^2}{2q_0M}=0,\quad E-\frac{\mathbf{p}^2}{2q_0M}>0, \tag{3.3}$$

the first and the last ones corresponding to generic orbits, the second one to a degenerate, lower dimensional orbit. As before, we consider only the last case.

It is easy to check that all the points $(E,\mathbf{p})$ such that $E-(\mathbf{p}^2/2q_0M)>0$ are equivalent to the point $(1,\mathbf{0})$, whose isotropy group is SO(3). The inducing group is then $K=\mathcal{T}\rtimes SO(3)$. The section $s:X=\mathcal{G}_S^M/K\to\mathcal{G}_S^M$ is defined as for $\mathcal{G}_{\text{aff}}^M$ and we obtain the following UIR of $\mathcal{G}_S^M$:

$$(U_+^{(q_0)}(\theta,b,\mathbf{a};\mathbf{v};R,\sigma)\psi)(E,\mathbf{p}) = e^{5\sigma/2}e^{i(q_0\theta+Eb-\mathbf{p}\cdot\mathbf{a})}D_j(R)\,\psi((\mathbf{v};R,\sigma)^{-1}(E,\mathbf{p})), \tag{3.4}$$

with D_j a UIR of SO(3) and

$$(\mathbf{v};R,\sigma)^{-1}(E,\mathbf{p}) = (e^{2\sigma}[E-\mathbf{p}\cdot\mathbf{v}+\tfrac{1}{2}q_0M\mathbf{v}^2],e^{\sigma}R^{-1}[\mathbf{p}-q_0M\mathbf{v}]). \tag{3.5}$$

The representation space of $U_+^{(q_0)}$ is $\mathfrak{H}_L^{(q_0)}\simeq L^2(X_+^{(q_0)},dE\,d^3\mathbf{p};\mathfrak{H}_j)$, where $X_+^{(q_0)}=\{(E,\mathbf{p}): E-\mathbf{p}^2/2q_0M>0\}$ and $\mathfrak{H}_j$ is the representation space of D_j.

For the purpose of the analysis of the square integrability of the representation (3.3), we consider again the trivial, spin zero, representation D_0 of SO(3). From (2.25), the left invariant measure on $\mathcal{G}_S^M$ is

$$d\mu_l(g) = e^{-2\sigma}d\theta\;\;db\;\;d^3\mathbf{a}\;\;d^3\mathbf{v}\;\;dm(R)\,d\sigma,\quad g\equiv(\theta,b,\mathbf{a};\mathbf{v};R,\sigma), \tag{3.6}$$

with $dm(R)$ the usual invariant measure on SO(3). Of course, the representation $U_+^{(q_0)}$ cannot be square integrable on the whole group $\mathcal{G}_S^M$, since the center $\Theta\simeq\mathbb{R}$ of the latter is noncompact. Thus we consider the square integrability on the quotient $Y=\mathcal{G}_S^M/\Theta$. We define a section $s:Y\to\mathcal{G}_S^M$ as follows:

$$s(b,\mathbf{a};\mathbf{v};R,\sigma)=(0,b,\mathbf{a};\mathbf{v};R,\sigma). \tag{3.7}$$

The square integrability mod(Θ,s) amounts to

$$\int |\psi(E,\mathbf{p})|^2 |\psi((\mathbf{v};R,\sigma)^{-1}(E,\mathbf{p}))|^2 e^{3\sigma} dE d^3\mathbf{p} \ \ d^3\mathbf{v} \ \ d\sigma<\infty. \tag{3.8}$$

In order to verify the validity of this relation, we proceed exactly as in Sec. D above. First, we perform the following coordinate transformation $(E,\mathbf{p})\mapsto(\boldsymbol{\xi},I,\eta)$ defined by

$$(E,\mathbf{p})=(\boldsymbol{\xi},I,\eta)(1,\mathbf{0})=(e^{-2\eta}+\tfrac{1}{2}q_0M\boldsymbol{\xi}^2,q_0M\boldsymbol{\xi}). \tag{3.9}$$

Since $(\mathbf{v};R,\sigma)^{-1}(\boldsymbol{\xi},I,\eta)=(e^{\eta}R^{-1}[\mathbf{v}-\boldsymbol{\xi}],I,\sigma-\eta)^{-1}(\mathbf{0},R,0)^{-1}$, we may also define new coordinates by the formula

$$(e^{\eta}R^{-1}[\mathbf{v}-\boldsymbol{\xi}],I,\sigma-\eta)^{-1}(\mathbf{0},R,0)^{-1}=(\mathbf{v}',I,\sigma')(\mathbf{0},R,0)^{-1}, \tag{3.10}$$

or, equivalently,

$$e^{\eta}(\mathbf{v}-\boldsymbol{\xi})=-e^{\sigma'}R\mathbf{v}', \quad \sigma'=-\sigma+\eta. \tag{3.11}$$

From the coordinate transformations (3.9) and (3.11), we deduce

$$dE d^3\mathbf{p}=2M^3e^{-2\eta}d\eta d^3\boldsymbol{\xi}, \ \ e^{3\sigma}d^3\mathbf{v} \ \ d\sigma=d^3\mathbf{v}'d\sigma'. \tag{3.12}$$

The square integrability condition (3.8) then reads

$$\|\psi\|^2\int |\tilde{\psi}(\mathbf{v};\sigma)|^2 d^3\mathbf{v} \ \ d\sigma=\frac{\|\psi\|^2}{q_0^4}\int_{X_+^{(q_0)}} \frac{|\psi(E,\mathbf{p})|^2}{E-\dfrac{\mathbf{p}^2}{2q_0M}} dE d^3\mathbf{p}<\infty. \tag{3.13}$$

In conclusion, we may state

Proposition 2: The representation $U_+^{(q_0)}$ defined by (3.4) is square integrable over the group $\mathcal{G}_S^M$ modulo its center Θ, the phase subgroup, and the section s given in (3.7), a vector $\psi \in \mathfrak{H}_L^{(q_0)}$ being admissible iff the integral in (3.13) converges.

A similar result may be obtained for the $(n+1)$-dimensional Schrödinger group, for any $n \geqslant 1$.

Clearly there is a dense set of admissible vectors ψ, and each of them generates a tight frame of CS, of the Gilmore–Perelomov-type. Typical wavelets of this kind are, for instance, in space–time:

(a) the Schrödinger–Marr wavelet,

$$\check{\psi}_{\mathrm{SM}}(\mathbf{x},t)=\left(i\partial_t+\frac{\Delta}{2m}\right)e^{-(\mathbf{x}^2+t^2)/2};$$

(b) the Schrödinger–Cauchy wavelet,

$$\check{\psi}_{\mathrm{SC}}(\mathbf{x},t)=\left(i\partial_t+\frac{\Delta}{2m}\right)\frac{1}{(t+i)\Pi_{j=1}^3(x_j+i)}.$$

Here we have put $m=q_0M$, so that m is the true mass, as it appears in the Schrödinger equation, as observed in Ref. 13. Preliminary numerical analysis of these functions reveal a rather complicated structure. Yet the wavelet transform they generate is, by construction, well adapted to quantum situations governed by the Schrödinger equation, and one may hope that they will prove useful in problems such as the description of Rydberg states in atomic physics, or in the description of laser–atom interactions.

Our last remark is to make the connection between the analysis of the extended Galilei–Schrödinger group $\mathcal{G}_S^M$ and the special subgroup discussed at the end of Sec. E. Mathematically, imposing the constraint $q=1$ in (2.41), inserting a factor $\delta(q-1)$, is equivalent to quotienting out the subgroup Θ. In technical terms, the original representation space $\mathfrak{H}_L \simeq L^2(X_{++}, dq\,dE\,d^3\mathbf{p};\mathfrak{H}_j)$ is a direct integral over q,

$$\mathfrak{H}_L = \int_{q>0}^{\oplus} \mathfrak{H}_L^{(q)} dq, \tag{3.14}$$

and we are taking the restriction to a single component, corresponding to $q=1$, or, more generally, $q=q_0$. As a result, the phase factor in $U_+^{(q_0)}$ is trivial, and may be factored out, exactly as for the Weyl–Heisenberg group.

We also note, in conclusion, that the same analysis may be done for the other nondegenerate orbit, corresponding to $E-(\mathbf{p}^2/2q_0M)<0$. One gets another UIR of $\mathcal{G}_S^M$, called $U_-^{(q_0)}$, and acting in the Hilbert space $\mathfrak{H}_{L-}^{(q_0)} \equiv L^2(X_-^{(q_0)}, dE\,d^3\mathbf{p};\mathfrak{H}_j)$. The two spaces $\mathfrak{H}_{L\pm}^{(q_0)} \equiv L^2(X_\pm^{(q_0)}, dE\,d^3\mathbf{p};\mathfrak{H}_j)$ deserve to be called Schrödinger–Hardy spaces, because they are the genuine analogs of the usual Hardy spaces on $\mathbb{R}$, i.e., the subspaces of progressive, resp. antiprogressive, wavelets.

ACKNOWLEDGMENTS

During the early stages of this work, we enjoyed the collaboration of C. Kalisa and M. N. Hounkonnou. In addition we benefited from interesting discussions with S. T. Ali and R. Murenzi, and some of the calculations have been checked and corrected by St. Liénart. It is a pleasure to thank them all.

APPENDIX A: GROUP RELATED COHERENT STATES: A GENERAL FORMALISM

In this section, we present a quick survey of the main results of the coherent state formalism. Besides the standard texts such as Ref. 17 or Ref. 18, a systematic discussion may be found in the review paper, Ref. 3, or the original papers.[1,2]

1. Coherent states on a locally compact group

Let G be a locally compact group, dg the left invariant Haar measure on G and $g \mapsto U(g)$ a continuous, unitary irreducible representation of G in a Hilbert space $\mathfrak{H}$.

A vector $\eta \in \mathfrak{H}$ is said to be *admissible* if

$$d(\eta) = \int_G |\langle U(g)\eta|\eta\rangle|^2 dg < \infty, \tag{A1}$$

or, equivalently, if

$$0 < \int_G |\langle U(g)\eta|\phi\rangle|^2 dg < \infty, \quad \forall \phi \in \mathfrak{H}. \tag{A2}$$

Let $\mathcal{A}$ be the set of all admissible vectors in $\mathfrak{H}$. If $\mathcal{A} \neq \{0\}$, then it is dense in $\mathfrak{H}$. In particular, $\mathcal{A} = \mathfrak{H}$ if and only if G is unimodular, i.e., it has a Haar measure which is both left and right invariant. If $\mathcal{A} \neq \{0\}$, the representation U of G in $\mathfrak{H}$ is said to be *square integrable*. The positive constant $d(\eta)$ is often called the formal dimension of U.

Let $\eta \in \mathcal{A}$ be fixed. Then the map $W_\eta : \mathfrak{H} \to L^2(G, dg)$, defined by

$$(W_\eta \phi)(g) = d(\eta)^{1/2} \langle U(g)\eta|\phi\rangle, \quad \phi \in \mathfrak{H}, \tag{A3}$$

is isometric. Equivalently, the vectors $\eta_g \equiv U(g)\eta$ generate a a resolution of the identity

$$d(\eta)^{-1}\int_G |\eta_g\rangle\langle\eta_g| dg = I_{\mathfrak{H}}, \tag{A4}$$

where $I_{\mathfrak{H}}$ is the identity operator in $\mathfrak{H}$. Then the orbit $\mathcal{S}(\eta)=\{\eta_g\equiv U(g)\eta, g\in G\}$ of η under the action of G is an overcomplete family of vectors in $\mathfrak{H}$, indexed by the points of G. These vectors are called *coherent states* (CS) associated with the representation U and the map W_η the *coherent state map*, or the *wavelet transform*.[3]

In addition, the map W_η intertwines U with the left regular representation U_L of G,

$$W_\eta U(g)=U_L(g)W_\eta, \quad \forall g\in G. \tag{A5}$$

This means that every square integrable representation is unitarily equivalent to a subrepresentation of the left regular representation of G. The set of UIR's with this property is called the *discrete series* of representations of G, since they correspond to the discrete (or atomic) part of the Plancherel measure that governs the decomposition of the left regular representation into irreducible components. However, the square integrability property itself is sometimes taken as a definition of the discrete series representations. For further information, see, for instance, Ref. 19.

The isometry property of W_η has interesting consequences.

(1) The range $\mathfrak{H}_\eta$ of W_η is a closed subspace of $L^2(G,dg)$. Denote by P_η the corresponding projection operator, $\mathfrak{H}_\eta = W_\eta\mathfrak{H} = P_\eta L^2(G,dg)$, with

$$W_\eta W_\eta^* = P_\eta, \quad W_\eta^* W_\eta = I_{\mathfrak{H}}. \tag{A6}$$

Then the projection operator P_η is an integral operator, with kernel $K_\eta : G\times G\to\mathbb{C}$ given by $K_\eta(g,g')=d(\eta)^{-1}\langle\eta_g|\eta_{g'}\rangle$, that is,

$$(P_\eta\phi)(g)=\int_G K(g,g')\phi(g')dg', \quad \forall\phi\in L^2(G,dg). \tag{A7}$$

(2) Since $P_\eta = P_\eta^2$, the kernel satisfies the relation

$$K_\eta(g,g')=\int_G K_\eta(g,g'')K_\eta(g'',g')dg''. \tag{A8}$$

In particular, Eq. (A8) gives the reproducing property of the kernel K_η,

$$\phi_\eta(g)=\int_G K_\eta(g,g')\phi_\eta(g')dg', \quad \forall g\in G, \quad \forall\phi_\eta\in\mathfrak{H}_\eta. \tag{A9}$$

Thus K_η is called a *reproducing kernel* and the space $\mathfrak{H}_\eta$ is called a reproducing kernel Hilbert space.

(3) Finally, again by the isometry property (A6), the map W_η may be inverted on its range by the adjoint operator, and one obtains a *reconstruction formula*,

$$W_\eta^{-1}\Phi = \int_G \Phi(g)\,\eta_g dg, \quad \Phi\in\mathfrak{H}_\eta. \tag{A10}$$

Typical examples of square integrable representations are the natural representations of the (connected) affine groups of the line or the plane, namely, the ''$ax+b$'' group (translations and dilations of $\mathbb{R}$) and the similitude group of the plane, $\mathrm{SIM}(2)=\mathbb{R}^2\rtimes(\mathbb{R}_*^+\times\mathrm{SO}(2))$ (translations, dilations, and rotations of $\mathbb{R}^2$). The corresponding coherent states are the familiar *wavelets*, in one and two dimensions, respectively.

2. Generalization: CS on homogeneous spaces

However, most continuous UIRs of G are not square integrable. For instance, there is none if the center of G is noncompact or, more generally, if G has no discrete series, like the Poincaré group. Moreover, it might be desirable to have a family of coherent states which are not indexed by the points of G itself, but rather by the points of some homogeneous space $X=G/H$, where H is a closed subgroup of G. A typical example is $H=H_\eta$, the isotropy subgroup (up to a phase) of η: this is the familiar case of the Gilmore–Perelomov coherent states.[18,20,21]

In all such cases, one may recover most of the previous theory under additional conditions, as follows. Let U be a UIR of G on a Hilbert space $\mathfrak{H}$, ν be a (quasi-)invariant measure on $X=G/H$ and $\sigma:X\to G$ a Borel section. The representation U is then said to be *square integrable mod* (H,σ) if there exists a nonzero $\eta\in\mathfrak{H}$ such that

$$0<\int_X |\langle U(\sigma(x))\eta|\phi\rangle|^2 d\nu(x)=\langle\phi|A_\sigma\phi\rangle<\infty,\quad \forall\phi\in\mathfrak{H}, \tag{A11}$$

where A_σ is a bounded invertible operator. For such a vector η, which is said to be admissible mod(H,σ), one defines as *(covariant) coherent states* the vectors $\eta_{\sigma(x)}=U(\sigma(x))\eta$, $x\in X$. In this way one obtains a total set $\mathcal{S}_\sigma=\mathcal{S}_\sigma(\eta)$ of vectors in $\mathfrak{H}$, indexed by the points of X.

If the inverse operator A_σ^{-1} is also bounded, the set $\mathcal{S}_\sigma$ is called a *frame*, and a *tight frame* if A_σ is a multiple of the identity, $A_\sigma=\lambda I_{\mathfrak{H}}$. This terminology is borrowed from the theory of nonorthogonal expansions,[22,23] to which the present construction reduces in the where X is a purely discrete space, with ν a counting measure.

The covariant coherent states have essentially the same properties as the ordinary ones. Assume for simplicity that we have a frame and that the measure ν is invariant. Then one has successively

(1) *A coherent state map* W_η: $\mathfrak{H}\to L^2(X,d\nu)$, given by

$$\phi\mapsto(W_\eta\phi)(x)=\langle\eta_{\sigma(x)}|\phi\rangle. \tag{A12}$$

Then $\mathfrak{H}_\eta=\mathrm{Ran}\,W_\eta$ is complete with respect to the scalar product $\langle\Phi|\Psi\rangle_\eta\equiv\langle\Phi|W_\eta A_\sigma^{-1}W_\eta^{-1}\Psi\rangle$, and W_η: $\mathfrak{H}\to\mathfrak{H}_\eta$ is an isometry.

(2) *A resolution of the frame operator* A_σ:

$$\int_X |\eta_{\sigma(x)}\rangle\langle\eta_{\sigma(x)}|d\nu(x)=A_\sigma. \tag{A13}$$

(3) *A reproducing kernel:* The orthogonal projector $P_\eta:L^2(X,d\nu)\to\mathfrak{H}_\eta$ is an integral operator K_σ with reproducing kernel $K_\sigma(x,y)=\langle\eta_{\sigma(x)}|A_\sigma^{-1}\eta_{\sigma(y)}\rangle$.

(4) *A reconstruction formula:* W_η is invertible on its range and

$$W_\eta^{-1}\Phi=\int_X \Phi(x)A_\sigma^{-1}\eta_{\sigma(x)}d\nu(x),\quad \Phi\in\mathfrak{H}_\eta. \tag{A14}$$

This general construction encompasses, of course, the standard Gilmore–Perelomov CS, such as[18,20,21]

(a) CS of semisimple groups, such as SU(2), which yields spin CS, SU(1,1), used for defining path integrals, or SO(3,2) and Sp(4,$\mathbb{R}$), familiar in nuclear physics.

(b) The Weyl–Heisenberg group G_{WH}, which yields the canonical (oscillator) CS (here H is the center of G_{WH}); the map W_η is called the Windowed Fourier Transform or Gabor Transform.

(c) The similitude group SIM(n) of $\mathbb{R}^n(n\geqslant 3)$: for $H=\mathrm{SO}(n-1)$, one gets the axisymmetric n-dimensional wavelets.

In addition, it yields CS for many groups which have no square integrable representations, and are thus inaccessible to the Gilmore–Perelomov method. A notable example is that of the relativity groups, such as the Euclidean, the Galilei or the Poincaré groups, or more generally, the semidirect product $G=V\rtimes S$ of a vector space V by a semisimple group S of automorphisms of V. In that case, one usually takes for $X=G/H$ a coadjoint orbit of G, i.e., a natural phase space for the system at hand. It follows that $H\neq H_\eta$, for any η, since H contains translations. Thus it is imperative to use the general construction in order to get CS.[2,3]

The interesting point is that, when one combines these groups with dilations, thus getting the corresponding affine groups, square integrability is often regained. This is true in the Poincaré case, as shown by Unterberger[10] and Bohnké.[9] It is also true for the (extended) Galilei group, as shown in this paper. But it is *not* true in the Weyl–Heisenberg case; the affine Weyl–Heisenberg group has no square integrable representations, and one has to go to appropriate homogeneous spaces.[24]

APPENDIX B: THE MACKEY METHOD OF INDUCED REPRESENTATIONS

For the convenience of the reader, we recall briefly in this Appendix the essential points of the Mackey method of induced representations. Let G be a Lie group of the form $G=\mathcal{T}\rtimes H$, a semidirect product of the Abelian group $\mathcal{T}\sim\mathbb{R}^n$ $(n>0)$ with the semisimple group H. We denote by $\mathcal{T}^*$ the dual of $\mathcal{T}$ and by $\langle x|a\rangle$ the action of the element $x\in\mathcal{T}^*$ on the element $a\in\mathcal{T}$. Every element $h\in H$ induces an automorphism in $\mathcal{T}$, and thus by duality induces the automorphism in $\mathcal{T}^*$ defined as follows:

$$\langle hx|a\rangle=\langle x|h^{-1}a\rangle,\quad \forall(a,x)\in\mathcal{T}\times\mathcal{T}^*. \tag{B1}$$

The unitary irreducible representations of $\mathcal{T}$ are one-dimensional, that is, characters, and are given by

$$\chi_{x_0}(a)=\exp\, i\langle x_0|a\rangle,\quad \forall a\in\mathcal{T}, \tag{B2}$$

with x_0 a fixed element of $\mathcal{T}^*$. Given $x_0\in\mathcal{T}^*$, let $H_0\subset H$ be its isotropy subgroup, $K=\mathcal{T}\rtimes H_0$ and $X=G/K\simeq H/H_0$. Then there is a one-to-one correspondence between X and the set Hx_0, and we may define a section

$$s:X\to G:s(x)=(0,\Lambda_x),\quad \text{where}\ \ \Lambda_x x_0=x. \tag{B3}$$

Let L be the UIR of K which coincides with χ_{x_0} on $\mathcal{T}$ and with the UIR D of H_0 on H_0. Then the Hilbert space $\mathfrak{H}_L$ of the UIR of G induced by L is defined as follows.

We consider functions f defined on G with values in $\mathfrak{K}_L$, the Hilbert space of the representation L, and satisfying the following covariance property:

$$f(gk)=L(k^{-1})f(g),\quad g\in G,\quad k\in K. \tag{B4}$$

Therefore, f is completely determined if we define its value for an arbitrary element of every left coset modulo K. Since left cosets are labeled by elements of Hx_0 according to (B3), we may choose a quasi-invariant measure $d\mu(x)$ on $X=G/K$ and then consider elements $f:G\to\mathfrak{K}_L$, whose restriction $f(\Lambda_x)\equiv f(0,\Lambda_x)$ to the representatives $(0,\Lambda_x)$ form a Hilbert space $\mathfrak{H}_L\simeq L^2(X,d\mu;\mathfrak{K}_L)$ with respect to the scalar product

$$\langle f'|f\rangle=\int_X\langle f'(\Lambda_x)|f(\Lambda_x)\rangle_L\, d\mu(x), \tag{B5}$$

where $\langle\cdot|\cdot\rangle_L$ denotes the scalar product in $\mathfrak{K}_L$. The representation U of G in $\mathfrak{H}_L$ induced by the representation L of K is then defined as follows:

$$(U(a,\Lambda)f)(\Lambda_x)=\left|\frac{d\mu(x')}{d\mu(x)}\right|^{1/2} e^{i\langle x|a\rangle}D(\Lambda_x^{-1}\Lambda\Lambda_{x'})f(\Lambda_{x'}), \tag{B6}$$

with $(0,\Lambda_{x'})$ the representative of the class of the element $(a,\Lambda)^{-1}(0,\Lambda_x)$. It is easy to check that $\Lambda_x^{-1}\Lambda\Lambda_{x'}\in H_0$.

Remark: If $x_0'\in Hx_0$, its isotropy group H_0' is conjugate to H_0. The point is that the inducing procedure applied to x_0' gives the same representation U (up to equivalence), provided the equivalence class of D does not change. Therefore the Mackey method implies the determination of all the orbits under H in $\mathcal{T}^*$ and their respective isotropy groups (up to conjugation).

APPENDIX C: CENTRAL EXTENSIONS OF $\mathcal{G}_0 \rtimes D_2$ BY $\mathbb{R}$

1. The structure of $\mathcal{G}_0 \rtimes \mathcal{D}_2$

The generic element $g\in\mathcal{G}_0\rtimes\mathcal{D}_2$ is parametrized as $g=(b,\mathbf{a};\mathbf{v};R,\sigma,\tau)$ and the multiplication law is defined by

$$\begin{aligned}&(b,\mathbf{a};\mathbf{v};R,\sigma,\tau)(b',\mathbf{a}';\mathbf{v}';R',\sigma',\tau')\\&\quad=(b+e^{\tau}b',\mathbf{a}+e^{\sigma}[b'\mathbf{v}+R\mathbf{a}'];\mathbf{v}+e^{\sigma-\tau}R\mathbf{v}';RR',\sigma+\sigma',\tau+\tau'),\end{aligned} \tag{C1}$$

with $b,\mathbf{a},\mathbf{v},R,\sigma,\tau$ representing, respectively, time translations, space translations, pure Galilei transformations, rotations, space dilations, and time dilations. Their respective infinitesimal generators are given by H for the subgroup of time translations, P_i $(i=1,2,3)$ for translations along the coordinate axis i, K_i $(i=1,2,3)$ for pure Galilei transformations along the axis i, J_i $(i=1,2,3)$ for rotations around the coordinate axis i, D_S for space dilations, D_T for time dilations. The only nonvanishing commutators between these generators are the following, where ϵ_{ijk} is the totally antisymmetric tensor of order three:

$$\begin{aligned}&[J_i,J_j]=\epsilon_{ijk}J_k,\ \ [H,D_T]=-H,\\&[J_i,P_j]=\epsilon_{ijk}P_k,\ \ [P_i,D_S]=-P_i,\\&[J_i,K_j]=\epsilon_{ijk}K_k,\ \ [K_i,D_S]=-K_i,\\&[H,K_i]=-P_i,\ \ [K_i,D_T]=K_i.\end{aligned} \tag{C2}$$

In particular, one has

$$[D_S,D_T]=0. \tag{C3}$$

2. Central extensions of $\mathcal{G}_0 \rtimes \mathcal{D}_2$

In order to construct central extensions of Lie groups, it is easier to deal with the Lie algebra of the group than with the group itself. Therefore, instead of searching for the central extensions of the group $\mathcal{G}_0\rtimes\mathcal{D}_2$, we will determine the equivalence classes of central extensions of its Lie algebra.

Let $\mathfrak{g}$ be a real Lie algebra. Then every central extension of $\mathfrak{g}$ is determined by a skew-symmetric bilinear form $B:\mathfrak{g}\to\mathbb{R}$ such that

$$B([X,Y],Z)+B([Y,Z],X)+B([Z,X],Y)=0,\quad \forall X,Y,Z\in\mathfrak{g}, \tag{C4}$$

with $[\cdot,\cdot]$ denoting the Lie bracket in $\mathfrak{g}$. The extension associated to B is trivial if there exists a linear form $f:\mathfrak{g}\to\mathbb{R}$ such that

$$B(X,Y)=f([X,Y]),\quad \forall X,Y\in\mathfrak{g}, \tag{C5}$$

and two extensions are equivalent if their bilinear forms differ by a trivial one.

For the Lie algebra of $\mathcal{G}_0$ the existence of a one-parameter family of nontrivial extensions is due to the fact that $B(K_i, P_j) = \gamma\delta_{ij}$, with $\gamma \in \mathbb{R}$ an arbitrary number, which is related to the usual physical concept of mass.[12] In the case of $\mathcal{G}_0 \rtimes \mathcal{D}_2$, however, this is impossible. Indeed, we have

$$B([D_S, K_i], P_i) + B([K_i, P_i], D_S) + B([P_i, D_S], K_i) = 0, \tag{C6}$$

which implies, using (C2), that

$$B(K_i, P_i) = 0. \tag{C7}$$

The relation (C7) can also be obtained using D_T,

$$B([D_T, K_i], P_i) + B([K_i, P_i], D_T) + B([P_i, D_T], K_i) = 0. \tag{C8}$$

On the other hand, we may write

$$B(D_S, D_T) = \gamma, \tag{C9}$$

with $\gamma \in \mathbb{R}$ an arbitrary number. It is then easy to check that the commutation relations of the central extensions of the Lie algebra (C2) coincide with the latter, except for (C3), which now reads

$$[D_S, D_T] = \gamma I, \tag{C10}$$

where I is the generator of the center of the extension. Since (C10) means that $\{D_S, D_T, I\}$ is a Lie algebra isomorphic to the Weyl–Heisenberg algebra, the central extensions so obtained will be semidirect products of $\mathcal{G}_0$ with the Weyl–Heisenberg group, and none of them yields a mass observable.

Alternatively, one may ask whether the Lie algebra (C2) contain a subalgebra, whose central extensions generate the mass in the usual sense. Since (C7) can be derived from (C6) as well as from (C8), such a subalgebra, if it exists, must contain an element of the form

$$D_{ST} = D_S + \lambda D_T \tag{C11}$$

with $\lambda \neq 0$. Replacing D_S by D_{ST} in (C6) gives

$$(2 - \lambda) B(K_i, P_i) = 0. \tag{C12}$$

Thus the only solution generating a mass operator is given by $\lambda = 2$, namely,

$$B(K_i, P_i) = \gamma, \quad [K_i, P_i] = \gamma I, \tag{C13}$$

with $\gamma \in \mathbb{R}$ and I the generator of the center of the extension. Hence we obtain precisely the commutation relations of the central extensions of the Schrödinger group. It is then easy to check that these extensions coincide with the subgroup of $\mathcal{G}_{\text{aff}}^M$ defined by the equation

$$\tau = 2\sigma. \tag{C14}$$

In other words, in the Schrödinger case, the two operations of central extension and of taking a semidirect product with the dilation group commute.

[1] S. T. Ali, J-P. Antoine, and J-P. Gazeau, ''De Sitter to Poincaré contraction and relativistic coherent states,'' Ann. Inst. Henri Poincaré **52**, 83–111 (1990); ''Square integrability of group representations on homogeneous spaces. I, II,'' *ibid.* **55**, 829–855 (1991); **55**, 857–890 (1991).
[2] S. T. Ali, J-P. Antoine, and J-P. Gazeau, ''Continuous frames in Hilbert space,'' Ann. Phys. (N.Y.) **222**, 1–37 (1993); ''Relativistic quantum frames,'' *ibid.* **222**, 38–88 (1993).

[3] S. T. Ali, J-P. Antoine, J-P. Gazeau, and U. A. Mueller, ''Coherent states and their generalizations: A mathematical overview,'' Rev. Math. Phys. **7**, 1013–1104 (1995).
[4] S. T. Ali, J-P. Antoine, and J-P. Gazeau, *Coherent States, Wavelets, and their Generalizations* (Springer, New York, 1999) (to be published).
[5] U. Niederer, ''The maximal kinematical invariance group of the free Schrödinger equation,'' Helv. Phys. Acta **45**, 802–810 (1972).
[6] M. Perroud, ''Projective representations of the Schrödinger group,'' Helv. Phys. Acta **50**, 233–252 (1977).
[7] M. Duval-Destin and R. Murenzi, ''Spatio-temporal wavelets: Application to the analysis of moving patterns,'' in *Progress in Wavelet Analysis and Applications (Proc. Toulouse 1992)*, edited by Y. Meyer and S. Roques (Editions Frontières, Gif-sur-Yvette, 1993), pp. 399–408.
[8] J-P. Leduc, F. Mujica, R. Murenzi, and M. J. T. Smith, ''Missile-tracking algorithm using target-adapted spatio-temporal wavelets,'' *Automatic Object Recognition VII*, SPIE Proc. **3069**, 400–411 (1997); ''Spatio-temporal wavelet transforms for motion tracking,'' in *IEEE ICASSP '97* (IEEE Computer Soc. Press, Los Alamitos, CA, 1997), Vol. 4, pp. 3013–3016; ''Spatio-temporal continuous wavelets applied to missile warhead detection and tracking,'' *Visual Communications and Image Processing '97*, SPIE Proc. **3024**, 787–798 (1997); F. Mujica, R. Murenzi, M. J. T. Smith, and J-P. Leduc, ''Robust tracking in compressed image sequences,'' J. Electron. Imaging **7**, 746–754 (1998).
[9] G. Bohnké, ''Treillis d'ondelettes associés aux groupes de Lorentz,'' Ann. Inst. Henri Poincaré **54**, 245–259 (1991).
[10] A. Unterberger, *Analyse Harmonique et Analyse Pseudo-Différentielle du Cône de Lumière, Astérisque* (SMF, Paris, 1987), Vol. 156.
[11] J. Bertrand and P. Bertrand, ''Reflectivity study of time-varying targets,'' in *Signal Processing VI: Theory and Applications* (*Proc. EUSIPCO 92, Brussels*), edited by J. Vandewalle *et al.* (Elsevier, Amsterdam, 1992), pp. 1813–1816.
[12] J-M. Lévy-Leblond, ''Galilei group and nonrelativistic quantum mechanics,'' J. Math. Phys. **4**, 453–507 (1963); ''Galilei group and Galilean invariance,'' in *Group Theory and its Applications*, edited by E. M. Loebl (Academic, New York, 1971), Vol. II, pp. 221–299.
[13] J. Voisin, ''On some unitary representations of the Galilei group. I. Irreducible representations,'' J. Math. Phys. **6**, 1519–1529 (1965).
[14] C. Kalisa, ''Etats cohérents affines: canoniques, galiléens et relativistes,'' Thèse de doctorat, Université Catholique de Louvain, 1993.
[15] J-P. Antoine, ''Time-dependent wavelets, alias space–time affine coherent states,'' in *Quantization, Coherent States, and Poisson Structures (Proc. Bialowieza XIV, 1995)*, edited by A. Strasburger, S. T. Ali, J-P. Antoine, J-P. Gazeau, and A. Odzijewicz (PWN, Warsaw, 1998), pp. 123–132.
[16] P. Aniello, G. Cassinelli, E. De Vito, and A. Levrero, ''Square-integrability of induced representations of semidirect products,'' Rev. Math. Phys. **10**, 301–313 (1998); ''Wavelet transforms and discrete frames associated to semidirect products,'' J. Math. Phys. **39**, 3965–3973 (1998).
[17] J. R. Klauder and B. S. Skagerstam, *Coherent States: Applications in Physics and Mathematical Physics* (World Scientific, Singapore, 1985).
[18] A. M. Perelomov, *Generalized Coherent States and their Applications* (Springer-Verlag, Berlin, 1986).
[19] G. Warner, *Harmonic Analysis on Semi-Simple Lie Groups. I* (Springer-Verlag, Berlin, 1972).
[20] R. Gilmore, ''Geometry of symmetrized states,'' Ann. Phys. (N.Y.) **74**, 391–463 (1972); ''On properties of coherent states,'' Rev. Mex. Fis. **23**, 143–187 (1974).
[21] A. M. Perelomov, ''Coherent states for arbitrary Lie group,'' Commun. Math. Phys. **26**, 222–236 (1972).
[22] R. J. Duffin and A. C. Schaeffer, ''A class of nonharmonic Fourier series,'' Trans. Am. Math. Soc. **72**, 341–366 (1952).
[23] I. Daubechies, A. Grossmann, and Y. Meyer, ''Painless nonorthogonal expansions,'' J. Math. Phys. **27**, 1271–1283 (1986).
[24] C. Kalisa and B. Torrésani, ''*N*-dimensional affine Weyl–Heisenberg wavelets,'' Ann. Inst. Henri Poincaré **59**, 201–236 (1993).

On the braided Fourier transform on the *n*-dimensional quantum space

Giovanna Carnovale[a)]
Mathematisch Instituut, P.O. Box 80.010, 3508 TA, Utrecht, The Netherlands

(Received 19 November 1998; accepted for publication 27 March 1999)

We work out in detail a theory of integrability on the braided covector Hopf algebra and the braided vector Hopf algebra of type A_n introduced by Majid. Using a braided Fourier transform very similar to the one defined by Kempf and Majid we obtain n-dimensional analogs of results by Koornwinder expressing the correspondence between products of the q^2-Gaussian $g_{q^2}(\underline{x})$ times monomials, and products of the q^2-Gaussian $G_{q^2}(\underline{\partial})$ times q^2-Hermite polynomials under the transform. We invert the correspondence by finding a suitable inversion, different from the one of Kempf and Majid. We show that with this transforms, whenever $n\geqslant 2$, the Plancherel measure will depend on the parity of the power series that we are transforming. [S0022-2488(99)00410-7]

I. INTRODUCTION

Recently Majid (see Ref. 1 and references therein) has defined a generalization of the concept of Hopf algebra, namely braided groups. Hopf superalgebras and genuine Hopf algebras are examples of these objects, but there are more examples, associated to quantum groups, since the category of representations of a quasitriangular Hopf algebra is braided. These objects appear also in Ref. 2 with different terminology.

Kempf and Majid introduced[3] an integration theory for a class of braided groups arising from matrix solutions of the quantum Yang–Baxter equation as ‘‘braided covector algebras’’ (see also Ref. 1). They used this theory to define a formal braided Fourier transform and its inverse on these algebras. In their paper they also present the case of the braided line as an example, and the n-dimensional case in less detail.

The main problem in their theory is that it is very difficult to find an explicit integral that behaves well enough. They provide powerful general results in a theoretical way, but the description in specific cases is often hard to handle. Besides, they do not provide a definition of convergence of an integral nor do they treat in their article the case when a generalized function may be called integrable.

The purpose of this paper is to work out as far as we can the example of the n-dimensional quantum space of type A_n viewed as a braided group. We will provide different definitions of integrability, with examples and counterexamples, with respect to an integral similar to that in Ref. 3. Our definitions are based on extensions of representations of the braided covector and vector algebras. Using these facts, one can show in a more rigorous way the translation invariance of the integral proved in Ref. 3. We define different types of Fourier transforms, all based on but different from Ref. 3. One of them looks more like that in Ref. 4 since the integral does not have trivial braiding with elements in the braided group, and because the braided antipode appears in the definition. We also took inspiration from Ref. 5, where an analog of the Fourier transform for the case of the braided line is also studied, although the transform in Ref. 5 goes from an algebra to itself, while we are looking for a transform going from an algebra to its braided dual, as in Refs. 3, 4, and 6.

We find an n-dimensional analog of the correspondence between products of q^2-Gaussians

[a)]Electronic mail address: carnoval@math.u-cergy.fr

times monomials and other q^2-Gaussians times q^2-Hermite polynomials, similar to the classical case and to the results for the braided line in Ref. 5. We give also inverses for our transforms that invert the correspondence mentioned above, similarly to what appears in Ref. 5. The main tool for this inversion formula is the symmetry between the braided vector algebra and the braided covector algebra. Kempf and Majid had already defined an inversion formula in their article, but they used properties that our integral does not have. Other inversion formulas for the braided line are to be found in Ref. 6 where the case of distributions is also treated.

One of the most interesting results is that whenever $n \geqslant 2$, there is a loss of symmetry, so that the Plancherel measure will no longer be the same in the whole space. Indeed there is a action of $\mathbf{Z}_2^n$ associated to the parity of the power series we are working with, and the Plancherel measure will be constant only on the subspaces of power series with constant parity. Therefore, the transforms we define can also be seen as sine and cosine transform. A phenomenon of break of symmetry for q-integrals was also noted in Ref. 7, where the authors were defining a calculus associated to a q-deformed Heisenberg algebra.

Other definitions of analogs of the Fourier transform on genuine Hopf algebras, quantum spaces, or commutative algebras appeared before Ref. 5 in Refs. 8–10.

II. NOTATION AND PRELIMINARIES

In this paper a complex algebra has, unless otherwise stated, always a unit and q is a real number in (0,1). For a positive integer m, and for any $q \neq 1$ we write $[m]_q = (q^m - 1)/(q-1)$ and $[m]_q! = \Pi_{j=1}^m [j]_q$.

For any $a \in \mathbf{R}$ and for any $k \in \mathbf{Z}_{\geqslant 0}$, we will put $(a;q)_k = \Pi_{l=0}^{k-1}(1-aq^l)$. We will also write $(a;q)_\infty = \lim_{k\to\infty}(a;q)_k$ and for r real numbers $a_1,\ldots,a_r$ we will put $(a_1,\ldots,a_r;q)_\infty = \Pi_{j=1}^r (a_j;q)_\infty$. Finally, for $a \geqslant b$ with a and b both in $\mathbf{Z}_{\geqslant 0}$ in we will use the q-binomial coefficient $\left[{a \atop b}\right]_q = [a]_q!/[b]_q![a-b]_q! = (q;q)_a/(q;q)_b(q;q)_{a-b}$.

Whenever for any capital character E we have a multi-index $E = (e_1,\ldots,e_n)$ we will put $E_i = \Sigma_{j=1}^{i-1} e_j$ and $E^i = \Sigma_{j=i+1}^n e_j$. Hence $|E| = e_i + E_i + E^i$ for every $i.$

We identify the set $\{+,-\}$ with $\mathbf{Z}_2$, letting $+$ correspond to $\bar{0}$ and $-$ correspond to $\bar{1}$, so that n-tuples in $\{+,-\}^n$ can be identified with n-tuples in $\mathbf{Z}_2^n$. By means of this identification we define the map $A:\mathbf{Z}^n \to \mathbf{Z}_2^n \to \{+,-\}^n$ by reducing modulo 2 first, i.e., $A(b) = +$ if b is even and $-$ otherwise. We will also denote by $B:\{+,-\}^n \to \{0,1\}^n \subset \mathbf{Z}^n$ the map sending each ''even'' entry to 0 and each ''odd'' entry to 1.

Given an operator on the twofold tensor product of an n-dimensional vector space V, we identify this operator with the $n^2 \times n^2$ matrix R and we denote its entries by R^{ab}_{cd} where a, b are the row entries and c, d are the column entries. For such an R, the operator acting on the p-fold tensor product of V (for $p \geqslant 2$) as R on the ith and jth components and as the identity elsewhere will be denoted by R_{ij}. For summation we will use Einstein convention.

We recall that a braided group over a field K is an associative algebra A with multiplication m and a coassociative coalgebra with comultiplication Δ and counit ε together with an invertible linear map $\Psi: A\otimes A \to A \otimes A$ called braiding, and a linear map $S:A\to A$ called braided antipode such that the following properties hold: $\Psi(m\otimes \mathrm{id}) = (\mathrm{id}\otimes m)(\Psi\otimes\mathrm{id})(\mathrm{id}\otimes\Psi)$; $\Psi(\mathrm{id}\otimes m) = (m\otimes\mathrm{id})(\mathrm{id}\otimes\Psi)(\Psi\otimes\mathrm{id})$; $(\mathrm{id}\otimes\Delta)\circ\Psi = (\Psi\otimes\mathrm{id})(\mathrm{id}\otimes\Psi)(\Delta\otimes\mathrm{id})$; $(\Delta\otimes\mathrm{id})\circ\Psi = (\mathrm{id}\otimes\Psi)(\Psi\otimes\mathrm{id})(\mathrm{id}\otimes\Delta)$; $1\varepsilon = m(\mathrm{id}\otimes S)\Delta = m(S\otimes\mathrm{id})\Delta$; $\Delta m = (m\otimes m)(\mathrm{id}\otimes\Psi\otimes\mathrm{id})(\Delta\otimes\Delta)$; $\varepsilon\circ m = \varepsilon\otimes\varepsilon$; and $\Delta(1) = 1\otimes 1$. If there is no braided antipode, A is called a ''braided bialgebra.''

We will work with two particular braided groups, namely the braided covector Hopf algebra $\hat{V}(R)$ and the braided vector algebra $V(R)$ associated to the standard matrix solution R of the quantum Yang–Baxter equation of type A_n [i.e., defining the quantum group $\mathrm{SL}_q(n)$]. For a general definition of braided covector and vector algebra, see Ref. 1, Chap. 10. One sees there that these algebras are comodule algebras for $\mathrm{SL}_q(n)$, and that all possible vector spaces obtained by tensoring $\hat{V}(R)$ and $V(R)$ can be provided by an algebra structure such that they are again comodule algebras for $\mathrm{SL}_q(n)$. The product is then defined by means of the braidings and the product in $\hat{V}(R)$ and $V(R)$.

In this particular case $\hat{V}(R)$ is the unital associative algebra generated by $x_1,\ldots,x_n$, with relations given by $x_ix_j=qx_jx_i$ if $i>j$, i.e., $\hat{V}(R)$ is the n-dimensional quantum space. The counit is given by $\varepsilon(x_j)=0$ for every j. We know that this algebra has a basis given by (increasing) ordered monomials $x_1^{e_1}\ldots x_n^{e_n}$. The general formulas for the braiding Ψ, the comultiplication Δ, and the antipode S in Ref. 1 reduce as follows:

$$\Psi(x_i\otimes x_j)=\begin{cases} qx_j\otimes x_i, & \text{if } i<j,\\ q^2x_i\otimes x_i, & \text{if } i=j,\\ (q^2-1)x_i\otimes x_j+qx_j\otimes x_i, & \text{if } i>j,\end{cases}$$

so that, for $i<j$, $\Psi(x_i^a\otimes x_j^b)=q^{ab}x_j^b\otimes x_i^a$,

$$\Delta(x_1^{e_1}\cdots x_n^{e_n})=\sum_{j_1=0}^{e_1}\cdots\sum_{j_n=0}^{e_n}\left(\prod_{i=1}^{n}\begin{bmatrix} e_i\\ j_i\end{bmatrix}_{q^2}\right)q^{\Sigma_v J_v(e_v-j_v)}x_1^{e_1-j_1}\cdots x_n^{e_n-j_n}\otimes x_1^{j_1}\cdots x_n^{j_n},$$

and $S(x_1^{e_1}\cdots x_n^{e_n})=(-1)^{|E|}q^{|E|^2-|E|}x_1^{e_1}\cdots x_n^{e_n}$.

Here $V(R)$ is the associative unital algebra generated by $\partial_1,\ldots,\partial_n$, with relations given by $\partial_i\partial_j=q\partial_j\partial_i$ if $i<j$. The ordered monomials provide a basis for $V(R)$. In this case we fix the basis given by ordered monomials with decreasing order. The braiding, counit, comultiplication, and antipode are given by

$$\Psi(\partial_i\otimes\partial_j)=\begin{cases} q\partial_j\otimes\partial_i, & \text{if } i>j,\\ q^2\partial_i\otimes\partial_i, & \text{if } i=j,\\ (q^2-1)\partial_i\otimes\partial_j+q\partial_j\otimes\partial_i, & \text{if } i<j,\end{cases}$$

$\varepsilon(\partial_j)=0$ for every j, $S(\partial_n^{e_n}\cdots\partial_1^{e_1})=(-1)^{|E|}q^{|E|^2-|E|}\partial_n^{e_n}\cdots\partial_1^{e_1}$, and

$$\Delta(\partial_n^{e_n}\cdots\partial_1^{e_1})=\sum_{j_n=0}^{e_n}\cdots\sum_{j_1=0}^{e_1}\left(\prod_{i=1}^{n}\begin{bmatrix} e_i\\ j_i\end{bmatrix}_{q^2}\right)q^{\Sigma_u J^u(e_u-j_u)}\partial_n^{e_n-j_n}\cdots\partial_1^{e_1-j_1}\otimes\partial_n^{j_n}\cdots\partial_1^{j_1}.$$

By Majid's theory (see Corollary 9.2.14 and Proposition 10.3.6 in Ref. 1) we recover the braiding between $\hat{V}(R)$ and $V(R)$ and between $V(R)$ and $\hat{V}(R)$. They are given by

$$\Psi_{V(R),\hat{V}(R)}(\partial_i\otimes x_j)=\begin{cases} q^{-1}x_j\otimes\partial_i, & \text{if } i\neq j,\\ q^{-2}x_j\otimes\partial_j+\sum_{r>j}(q^{-2}-1)x_r\otimes\partial_r, & \text{if } i=j,\end{cases}$$

$$\Psi_{\hat{V}(R),V(R)}(x_i\otimes\partial_j)=\begin{cases} q^{-1}\partial_j\otimes x_i, & \text{if } i\neq j,\\ \sum_{r<j}(q^{-2}-1)q^{-2(j-r)}\partial_r\otimes x_r+q^{-2}\partial_j\otimes x_j, & \text{if } i=j.\end{cases}$$

For every choice of nonzero constants c_j, for $j=1,\ldots,n$, there is an algebra isomorphism ψ between $\hat{V}(R)$ and $V(R)$ mapping x_j to $c_j\partial_{n+1-j}$, such that $\Delta_{V(R)}\circ\psi=(\psi\otimes\psi)\circ\Delta_{\hat{V}(R)}$ and $S_{V(R)}\circ\psi=S_{\hat{V}(R)}$. In particular, for $c_j=q^{-j+(1/2)(n-1)}$ for every j, then we also have $\Psi_{V(R),V(R)}(\psi\otimes\psi)=\Psi_{\hat{V}(R),\hat{V}(R)}$.

For this choice of the c_j's we have

$$(\psi^{-1}\otimes\psi)\circ\Psi_{\hat{V}(R),V(R)}=\Psi_{V(R),V(R)}\circ(\psi\otimes\psi^{-1}),$$

$$(\psi\otimes\psi^{-1})\circ\Psi_{V(R),\hat{V}(R)}=\Psi_{\hat{V}(R),V(R)}\circ(\psi^{-1}\otimes\psi),$$

however, ψ is not a morphism in the braided category since it is *not* true that $(\psi\otimes\mathrm{id})\circ\Psi_{\hat{V}(R),V(R)}=\Psi_{\hat{V}(R),\hat{V}(R)}\circ(\mathrm{id}\otimes\psi)$, as one can easily see by computing the actions of the left-hand side and of the right-hand side on $(\partial_2\otimes x_1)$ for $n=2$. This has to do with the fact that $\hat{V}(R)$ and $V(R)$ are not dual as braided groups in the sense that there is no invariant quantum metric (see Ref. 1 and references therein).

It is also well known that there is a left action of $V(R)$ on $\hat{V}(R)$ where each ∂_j acts by means of braided partial differentiation with respect to x_j. In the A_n case, this turns out to be, for $f(\underline{x})\in\hat{V}(R)$,

$$\partial_j f(\underline{x})=x_j^{-1}\left[\frac{f(q^2x_1,\ldots,q^2x_j,x_{j+1},\ldots,x_n)-f(q^2x_1,\ldots,q^2x_{j-1},x_j,\ldots x_n)}{(q^2-1)}\right],$$

where the inverse of x_j is only formal, and ''apparent.'' In particular, one has

$$\partial_j(x_1^{e_1}\cdots x_n^{e_n})=[e_j]_{q^2}q^{E_j}x_1^{e_1}\cdots x_{j-1}^{e_{j-1}}x_j^{e_j-1}x_{j+1}^{e_{j+1}}\cdots x_n^{e_n}.$$

Formally we can repeat the same constructions with $\hat{V}(R)^{\mathrm{ext}}$ [resp. $V(R)^{\mathrm{ext}}$], the algebra of formal power series in the x_j's (resp. ∂_j's) with the given defining relations. In this case everything that we have described above is defined as in $\hat{V}(R)$ and $V(R)$.

Specializing the results about the braided exponential map in Ref. 1 and references therein, one has

$$\exp(x|\partial):=\sum_{e_1,\ldots,e_n\geq 0} x_1^{e_1}\cdots x_n^{e_n}\otimes\frac{\partial_n^{e_n}}{[e_n]_{q^2}!}\cdots\frac{\partial_1^{e_1}}{[e_1]_{q^2}!}.$$

By Example 10.4.16 in Ref. 1 this is equal to $e_{q^{-2}}((1-q^{-2})\Sigma_{i=1}^n x_i\otimes\partial_i)$, where $e_q(z)=\Sigma_{k=0}^\infty z^k/(q;q)_k$ (see Ref. 5 for further details). It follows by straightforward computation that $\exp(x|\partial)$ is also equal to $E_{q^2}((1-q^2)\Sigma_{i=1}^n x_i\otimes\partial_i)$, where $E_{q^2}(z)=\Sigma_{k=0}^\infty q^{(1/2)k(k-1)}z^k/(q;q)_k$. Corollary 10.4.17 in Ref. 1, which appeared first in Ref. 11, gives us also a braided version of the Taylor formula. This is given by

$$\begin{aligned}f(\Delta(x_1),\ldots,\Delta(x_n))&=f(x_1+y_1,\ldots,x_n+y_n)\\&=\sum_{e_1,\ldots,e_n\geq 0} y_1^{e_1}\cdots y_n^{e_n}\left(\frac{\partial_n^{e_n}}{[e_n]_{q^2}!}\cdots\frac{\partial_1^{e_1}}{[e_1]_{q^2}!}f(x_1,\ldots,x_n)\right)\\&=\exp(y|\partial)f(\underline{x})\end{aligned}$$

where $y_j=x_j\otimes 1$ and x_i stands for $1\otimes x_i$ after the second equality sign.

III. INTEGRATION ON $\hat{V}(R)^{\mathrm{ext}}$

We start with the ''indefinite'' integral with respect to x_i. We repeat shortly the definition in Ref. 3, where the integral is viewed as an operator on $\hat{V}(R)^{\mathrm{ext}}$.

Definition 3.1: The braided partial integral with respect to x_i acting on $f(\underline{x})\in\hat{V}(R)^{\mathrm{ext}}$ is given by

$$\int_0^{x_i} f:=(1-q^2)\sum_{k=0}^{\infty} q^{2k}x_i f(q^{-2}x_1,\ldots,q^{-2}x_{i-1},q^{2k}x_i,x_{i+1},\ldots,x_n).$$

It is easy to see that the operator defined above acts as a pseudo inverse for the partial differential operator ∂_i. It is indeed only a right inverse, since it acts as a left inverse for ∂_i only on series containing x_i in every monomial of its expansion (see remark in Ref. 3, p. 6815). Each

$\int_0^{x_i}$ is a well-defined operator from $\hat{V}(R)^{\mathrm{ext}}$ to $\hat{V}(R)^{\mathrm{ext}}$ since for every basis monomial $x_1^{e_1}\cdots x_n^{e_n}$ we can write $\int_0^{x_i}x_1^{e_1}\cdots x_n^{e_n}$ as a monomial in the x_j's with a coefficient that is a *convergent* series of complex numbers. Since one can read $\int_0^{x_i}f$ as a ''function'' of the x_j's, it makes sense to consider $\int_0^{ax_i}$ for a nonzero constant a. In particular we can define $\int_0^{(-1)^r q^1 x_i} f$ for every integer r and l as $(-1)^r(1-q^2)\Sigma_{k=0}^{\infty}q^{2k+l}x_i f(q^{-2}x_1,\ldots q^{-2}x_{i-1},(-1)^r q^{2k+l}x_i,x_{i+1},\ldots x_n)$.

Then, for every $f\in\hat{V}(R)^{\mathrm{ext}}$

$$\int_{-x_i}^{x_i} f := \int_0^{x_i} f - \int_0^{-x_i} f \quad \text{and} \quad \int_{-x_i\cdot\infty}^{x_i\cdot\infty} f := \lim_{r\to\infty}\int_{-q^{-2r}x_i}^{q^{-2r}x_i} f$$

are defined. The last definition is only formal so far, because the image of a power series is no longer a power series (coefficients might be infinite sums themselves), and we have no notion of convergence. One cannot find a convergence set because one cannot give nonzero values to noncommuting variables. This issue can be solved in different ways, so that we can give a meaning to equalities as well. The ideas here are based on the approaches of Kempf and Majid in Ref. 3 and of Koornwinder in Ref. 5 who treated the one-dimensional case. His approach was by means of extension of a suitable representation of $\hat{V}(R)^{\mathrm{ext}}$ and the search of a family of eigenvectors for which the integral would have a convergent eigenvalue. We approach the problem not for a single infinite integral, but for the n-dimensional integral $I(f) := \int_{-x_n\cdot\infty}^{x_n\cdot\infty}\cdots\int_{-x_i\cdot\infty}^{x_1\cdot\infty} f$ which is formally

$$(1-q^2)^n\sum_{k_n=-\infty}^{\infty}\cdots\sum_{k_1=-\infty}^{\infty}\sum_{\varepsilon\in\{\pm1\}^n} q^{2|K|+\binom{n}{2}}x_1\cdots x_n f(\varepsilon_1 q^{2k_1}x_1,\ldots,\varepsilon_n q^{2k_n}x_n).$$

As we said, $I(f)$ is not an element of $\hat{V}(R)^{\mathrm{ext}}$, in general, since the coefficients with respect to the elements of the basis are not always definite. In order to define integrability, we fix an action of $\hat{V}(R)$ on the space of power series in the n commuting variables $z_1,\ldots,z_n$ with complex coefficients. This representation corresponds with the choice of a normal form for the monomials in $\hat{V}(R)$. The representation, denoted by $\triangleright$, for monomials in the x_i's acting on monomials in the z_j's, is given by

$$x_1^{e_1}\cdots x_n^{e_n}\triangleright z_1^{h_1}\cdots z_n^{h_n} = q^{\Sigma_{i=1}^n e_j H_i} z_1^{h_1+e_1}\cdots z_n^{h_n+e_n},$$

and can be extended linearly to an action of $\hat{V}(R)$ on formal power series in the z_j's. We can restrict the space on which we act by taking the space V of power series which are absolutely convergent in a neighborhood of zero. This makes sense because the z_i's commute with each other. We see that this space is invariant under the action of $\hat{V}(R)$. Moreover, we see that we can extend the representation of $\hat{V}(R)$ on V to a representation of the class C given by the power series f in the x_i's such that $f\triangleright 1\in V$. Indeed one can see that

(A) $\forall f=f(\underline{x})\in C$ and $\forall g=g(\underline{z})\in V$ it holds that $(f\triangleright g)(\underline{z})$ belongs to V because the associated series of absolute values is majorized by the product in V of series of absolute values associated to $f\triangleright 1$ and g.

(B) $\forall f$ and $g\in C$, their product $fg\in C$ because $(fg)\triangleright 1 = f\triangleright(g\triangleright 1)\in V$ by property (A). Moreover,

(C) $\forall f\in C$, $f\triangleright 1=0 \Leftrightarrow f\equiv 0$.

From now on we write $f_{.1}$ for $f\triangleright 1$, for any expression $f(\underline{x})$ for which the action on 1 makes sense. We would like to extend the representation now to the formal expressions of type $I(f)$ for $f\in C$. This does not always make sense, hence we have to add further conditions. Let us take the subclass C' of C given by the series in C such that

(a) $f_{.1}$ can be continued analytically on $\mathbf{R}^n+iU$ for some open neighborhood U of 0 in $\mathbf{R}^n$;

(b) $f_{.1}$ is absolutely q^2-integrable for every $\underline{z}\in\mathbf{R}^n$ for which every $z_j\neq 0$, i.e.,

$$\sum_{k_1=-\infty}^{\infty}\cdots\sum_{k_n=-\infty}^{\infty}\sum_{\varepsilon\in\{\pm1\}^n} q^{2|K|}|z_1|\cdots|z_n||f_{\cdot 1}(\varepsilon_1 q^{2k_1}z_1,\ldots,\varepsilon_n q^{2k_n}z_n)|<\infty$$

for $\underline{z}$ outside the standard hyperplanes.

The class C' will be the class of integrable power series. For those series we can compute $I(f)_{\cdot 1}$, and this turns out to be

$$(I(f))_{\cdot 1}=(1-q^2)^n\left[\sum_{k_n=-\infty}^{\infty}\cdots\sum_{k_1=-\infty}^{\infty}\sum_{\varepsilon} q^{2|K|+\binom{n}{2}}x_1\cdots x_n f(\varepsilon_1 q^{2k_1}x_1,\ldots,\varepsilon_n q^{2k_n}x_n)\right]\triangleright 1$$
$$=\int_{-q^{n-1}z_1\cdot\infty}^{q^{n-1}z_1\cdot\infty}\cdots\int_{-q^{n-i}z_i\cdot\infty}^{q^{n-i}z_i\cdot\infty}\cdots\int_{-z_n\cdot\infty}^{z_n\cdot\infty} f_{\cdot 1}(t_1,\ldots,t_n)\,d_{q^2}t_n\cdots d_{q^2}t_1,$$

where the q^2-integral in the last line is the q^2-Jackson integral in n variables obtained by iterating (8.11) in Ref. 5. It is clear that if $f\in C'$, then $(I(f))_{\cdot 1}$ converges whenever $z_j\neq 0$ for every j.

Because of (C), two objects in C are equal if and only if they act in the same way on 1. Following this philosophy, we say that two q^2-integrals of objects in C' are equal if and only if they act in the same way on 1. This will be our tool to show equalities then.

The first purpose is to show translation invariance of the operator I in a less formal way than in Ref. 3 where this appeared first. For this we need an extra assumption on the elements in C', since we have to use Taylor's series, hence partial derivatives. We consider $f\in C'$ satisfying the following.

(c) For some $\eta>0$ there exists for each $J\in(\mathbf{Z}_{\geq 0})^n$ some constant C_J such that

$$|(D_{1,q^2}^{j_1}\cdots D_{n,q^2}^{j_n}f_{\cdot 1})(z_1,\ldots,z_n)|\leq C_J\prod_{k=1}^{n}(1+|z_k|^2)^{-(1+\eta)}$$

if $\underline{z}\in\mathbf{R}^n$, where D_{j,q^2} denotes the standard q^2-Jackson partial derivative with respect to z_j.

For an f in C', condition (c) implies that all the Jackson derivatives of $f_{\cdot 1}$ are absolutely q^2-integrable for all $\underline{z}$ in a neighborhood of 0 minus the intersection with the standard hyperplanes.

One sees immediately that $(\partial_i f)_{\cdot 1}=(D_{i,q^2}f_{\cdot 1})(qz_1,\ldots,qz_{i-1},z_i\ldots,z_n)$. We show now that it makes sense to compute $I(\partial_n^{j_n}\cdots\partial_1^{j_1}f)\triangleright 1$, and that this is equal to zero whenever $(j_1,\ldots,j_n)\neq(0,\ldots,0)$. We write

$$F_{\cdot 1}^{J}(z_1,\ldots,z_n):=(\partial_n^{j_n}\cdots\partial_1^{j_1}f)\triangleright 1=(D_{1,q^2}^{j_1}\cdots D_{n,q^2}^{j_n}(f_{\cdot 1}))(q^{J^1}z_1,\ldots,q^{J^n}z_n).$$

Hence $F_{\cdot 1}^{J}(\underline{z})\in V$ if f satisfies condition (c). Moreover, $F_{\cdot 1}^{J}$ is absolutely q^2-Jackson integrable if and only if for every choice of $(h_1,\ldots,h_n)\in\{\pm 1\}^n$

$$\sum_{\underline{\varepsilon}\in\{\pm1\}^n}\sum_{k_i=0}^{\infty}\cdots\sum_{k_n=0}^{\infty} q^{2K\cdot H}|z_1\cdots z_n||F_{\cdot 1}^{J}(q^{2k_1h_1}\varepsilon_1 z_1,\ldots,q^{2k_nh_n}\varepsilon_n z_n)|$$

has a positive radius of convergence. Hence, if condition (c) holds for $f(\underline{x})$, and since the above sums are of the form

$$\sum_{\underline{\varepsilon}}\sum_{k_1=0}^{\infty}\cdots\sum_{k_n=0}^{\infty} q^{2K\cdot H}|z_1\cdots z_n(D_{1,q^2}^{j_1}\cdots D_{n,q^2}^{j_n}f_{\cdot 1})(q^{2k_1h_1+J^1}\varepsilon_1 z_1,\ldots,q^{2k_nh_n+J^n}\varepsilon_n z_n)|$$
$$\leq C_J|z_1\cdots z_n|\sum_{\underline{\varepsilon}}\sum_{k_1=0}^{\infty}\cdots\sum_{k_n=0}^{\infty} q^{2K\cdot H}\prod_{r=1}^{n}(1+|q^{2k_rh_r}z_r|^2)^{-(1+\eta)}$$

that converges since $q\in(0,1)$, one sees that the $F^{J}_{\cdot 1}$ are q^2-integrable.

Moreover, for every $\gamma=(\gamma_1,\dots,\gamma_n)$ with γ_j in $\mathbf{R}-\{0\}$ we have

$$q^{\Sigma_{k=1}^n J^k}\int_{-\gamma_1\cdot\infty}^{\gamma_1\cdot\infty}\cdots\int_{-\gamma_n\cdot\infty}^{\gamma_n\cdot\infty}F^{J}_{\cdot 1}(t_1,\dots,t_n)d_{q^2}t_n\cdots d_{q^2}t_1$$

$$=\int_{-q^{J^1}\gamma_1\cdot\infty}^{q^{J^1}\gamma_1\cdot\infty}\cdots\int_{-q^{J^n}\gamma_n\cdot\infty}^{q^{J^n}\gamma_n\cdot\infty}D^{j_1}_{1,q^2}\cdots D^{j_n}_{n,q^2}f_{\cdot 1}(t_1,\dots,t_n)d_{q^2}t_n\cdots d_{q^2}t_1=0.$$

The proof is as in the one-dimensional case (see Ref. 5). Hence we can state the following.

Lemma 3.2: Let $f\in C'$ *satisfy condition* (c). *Then for every* $J\neq\underline{0}$ *there holds* $(I(\partial_n^{j_n}\cdots\partial_1^{j_1}f))\triangleright 1\equiv 0$, *so that we can conclude that* $I(\partial_n^{j_n}\cdots\partial_1^{j_1}f)=0$.

Proof: One has

$$(I(\partial_n^{j_n}\cdots\partial_1^{j_1}f))_{\cdot 1}=\int_{-q^{n-1}z_1\cdot\infty}^{q^{n-1}z_1\cdot\infty}\cdots\int_{-q^{n-i}z_i\cdot\infty}^{q^{n-i}z_i\cdot\infty}\cdots\int_{-z_n\cdot\infty}^{z_n\cdot\infty}F^{J}_{\cdot 1}(t_1,\dots,t_n)d_{q^2}t_n\cdots d_{q^2}t_1=0.$$

□

Proposition 3.3: Let $f\in C'$ *satisfy* (c). *Then* $(\mathrm{id}\otimes I)\Delta(f)=1\otimes(If)$.

Proof: By the braided Taylor formula we have

$$(\mathrm{id}\otimes I)(\Delta f)=\sum_{j_1,\dots,j_n\geqslant 0}\frac{y_1^{j_1}\cdots y_n^{j_n}}{[j_1]_{q^2}!\cdots[j_n]_{q^2}!}I(\partial_n^{j_n}\cdots\partial_1^{j_1}f),$$

but this is by Lemma 3.2 equal to the term with all j_k's equal to 0. □

Proposition 3.4: Let $f(\underline{x})\in C$ *such that* (c) *holds for every* $\eta>0$. Then the statement of Proposition 3.3 holds for every polynomial $p(\underline{x})$ times $f(\underline{x})$.

Proof: It is not restrictive to assume that $p(\underline{x})$ is a monomial. By property (A) there follows that also $p(\underline{x})f(\underline{x})\in C$. We have to check that condition. (c) holds for every element of the form $x_1^{e_1}\cdots x_n^{e_n}f(\underline{x})$. One sees immediately that

$$(x_1^{e_1}\cdots x_n^{e_n}f(\underline{x}))\triangleright 1=f_{\cdot 1}(q^{E^1}z_1,\dots,q^{E^n}z_n)z_1^{e_1}\cdots z_n^{e_n}\in V$$

if $f(\underline{x})\in C$, so that $(x_1^{e_1}\cdots x_n^{e_n}f(\underline{x}))_{\cdot 1}$ makes sense.

Condition (c) is on the q^2-Jackson partial derivatives on commuting variables, and it holds for $x_1^{e_1}\cdots x_n^{e_n}f(\underline{x})$ as a consequence of the fact that for two functions $a(\underline{z})$ and $b(\underline{z})$ and for any $j=1,\dots,n$,

$$D_{j,q^2}(a(\underline{z})b(\underline{z}))=(D_{j,q^2}a(\underline{z}))(b(z_1,\dots,z_{j-1},q^2z_j,z_{j+1},\dots,z_n))+a(\underline{z})D_{j,q^2}(b(\underline{z})).$$

Then, by Proposition 3.3 we have the statement. □

We have a description of a class of power series in the x_i's for which integration makes sense, although we are not able so far to make a complete classification of integrable functions. The same problem is treated in Ref. 6 for the one-dimensional case. Still, what we have is enough to allow computations in the following case.

Example 1: The q^2-Gaussian $g_{q^2}(\underline{x})$ is

$$g_{q^2}(\underline{x}):=e_{q^4}(-\underline{x}\cdot\underline{x})=e_{q^4}\left(-\sum_{j=1}^n x_j^2\right)=\prod_{j=1}^n e_{q^4}(-x_j^2),$$

where the last equality holds because of Proposition 3.1 in Ref. 5 (this result was already in Ref. 12) and the product in the above formula is taken with increasing order on the variables. It satisfies conditions (a), (b), and (c) for every $\eta>0$. This is a consequence of the one-dimensional case (see Ref. 5) and the fact that

$$(g_{q^2}(\underline{x}))_{.1}(\underline{z})=\prod_{j=1}^{n} e_{q^4}(-z_j^2), \quad \partial_j(g_{q^2}(\underline{x}))=-\frac{x_j}{(1-q^2)}g_{q^2}(\underline{x})$$

and that

$$D^{j_1}_{q^2,1}\cdots D^{j_n}_{q^2,n}\left(e_{q^4}\left(-\sum_{k=1}^{n} z_k^2\right)\right)=p(\underline{z})e_{q^4}\left(-\sum_{k=1}^{n} z_k^2\right),$$

where $p(\underline{z})$ is a polynomial in the z_j's.

It follows then that also elements of the form

$$x_1^{a_1}g_{q^2}(x_1)\cdots x_n^{a_n}g_{q^2}(x_n)=x_1^{a_1}\cdots x_n^{a_n}e_{q^4}\left(-\sum_j (q^{-A^j}x_j)^2\right)$$

satisfy condition (c), so that for every $p_j(x_j)\in\hat{V}(R)$ we can integrate every element of the form $p_1(x_1)g_{q^2}(x_1)\cdots p_n(x_n)g_{q^2}(x_n)$.

In particular, for $f(\underline{x})=g_{q^2}(x_1)x_1^{a_1}\cdots g_{q^2}(x_n)x_n^{a_n}$ one can compute

$$\begin{aligned}(I(f))_{.1}|_{z=\gamma}&=\int_{-q^{n-1}\gamma_1\cdot\infty}^{q^{n-1}\gamma_1\cdot\infty}\cdots\int_{-q^{n-i}\gamma_i\cdot\infty}^{q^{n-i}\gamma_i\cdot\infty}\cdots\int_{-\gamma_n\cdot\infty}^{\gamma_n\cdot\infty} f_{.1}(t_1,\ldots.t_n)d_{q^2}t_n\cdots d_{q^2}t_1\\
&=\prod_{j=1}^{n}\left(\int_{-q^{n-j}\gamma_j\cdot\infty}^{q^{n-j}\gamma_j\cdot\infty} e_{q^4}(-t_j^2)t_j^{a_j}d_{q^2}t_j\right)\\
&=\begin{cases}\prod_{j=1}^{n}(c_{q^2}(\gamma_j q^{n-j})q^{-a_j^2/2}(q^2;q^4)_{a_j/2}), & \text{if } a_j \text{ even } \forall_j,\\ 0, & \text{otherwise,}\end{cases}\end{aligned}$$

where

$$c_{q^2}(\gamma)=\frac{2(1-q^2)(q^4,-q^2\gamma^2,-q^2\gamma^{-2};q^4)_\infty}{(-\gamma^2,-q^4\gamma^{-2},q^2;q^4)_\infty}$$

as in formula (8.15) in Ref. 5. In particular,

$$(I(f))_{.1}|_{z=\gamma}=q^{-\Sigma_j a_j^2/2}\prod_j (q^2;q^4)_{a_j/2}(I(g_{q^2}(\underline{x})))|_{z=\gamma}$$

if all a_j's are even, and 0 otherwise, hence we can conclude that

$$I(f)=\begin{cases}q^{-\Sigma_j a_j^2/2}\prod_j (q^2;q^4)_{a_j/2}(I(g_{q^2}(\underline{x}))), & \text{if } a_j \text{ is even } \forall_j,\\ 0, & \text{otherwise.}\end{cases}$$

The observation that the integral of the Gaussian times a monomial is equal to a constant times the integral of the Gaussian appeared in Refs. 1 and 3 first. In our case we have to deal with the shift and this depends on the choice of our global integral. ♠

Remark: The reader may wonder whether we could have chosen another realization of $\hat{V}(R)^{\mathrm{ext}}$ and of the integrals of elements of $\hat{V}(R)^{\mathrm{ext}}$ other than $f \triangleright 1$. Of course one might consider a different representation, or a different choice of the normal form. The advantages of a representation associated to the choice of a normal form is the fact that it is enough to test operators on 1 to state an equivalence in $\hat{V}(R)$. The advantage of the particular normal form that we have chosen is based on the fact that C is closed under product, hence we have a map from formal expressions in $x_1 \ldots, x_n$ to formal expressions in the $z_1, \ldots, z_n$ such that on rather big subspaces it comes exactly from an algebra homomorphism. If we had chosen another normal form, we could no longer extend the representation π of $\hat{V}(R)$ on V to a representation of the subset S of $\hat{V}(R)^{\mathrm{ext}}$ such that $\pi(S)(1) \subset V$. Take, for instance, $n=2$, and the representation of $\hat{V}(R)$ on $\mathbf{R}[[z,w]]$ given by $\pi(x_1)(f(z,w)) = zf(z,q^{-1}w)$ and $\pi(x_2)(f(z,w)w) = f(z,w)$. This is the representation associated to the choice of the normal form with x_2 preceding x_1. ♠

Then $a = \Sigma_{k=0}^{\infty} x_1^k$ and $b = \Sigma_{l=0}^{\infty} x_2^l$ belong to S, but ab does not belong to S since $\pi(\Sigma_{k=0}^{\infty} x_1^k)(\pi(\Sigma_{l=0}^{\infty} x_2^l)(1)) = \Sigma_{k,l=0}^{\infty} q^{-kl} z^k \omega^l \notin V$.

IV. LATTICE INTEGRABILITY

In the previous section we saw a definition of integrable series in $\hat{V}(R)^{\mathrm{ext}}$. Unfortunately, the above method fails for another analog of the Gaussian we would like to q^2-integrate, namely the q^2-Gaussian

$$G_{q^2}(\underline{x}) := E_{q^4}(-\underline{x} \cdot \underline{x}) = E_{q^4}\left(-\sum_{j=1}^{n} x_j^2\right) = E_{q^4}(-x_n^2) \cdots E_{q^4}(-x_1^2).$$

In Ref. 5, Sec. 9, it is also shown that $G_{q^2}(\underline{x})$ does not satisfy condition (c), nor condition (b) for $n=1$. On the other hand, it is also shown there that for a given choice of a q^2-lattice of the form $\{\pm q^{2k}\gamma | k \in \mathbf{Z}\}$, namely for $\gamma=1$, $(I(G_{q^2}(x_1)))_{\cdot 1}|_{z_1=1}$ is absolutely convergent. Hence one can introduce a weaker version of integrability in $\hat{V}(R)^{\mathrm{ext}}$, which we will call *''lattice integrability,''* requiring for an $f(\underline{x})$ such that $f_{\cdot 1}$ is entire that there is a q^2-lattice $L(\gamma) = \{(\pm \gamma_1 q^{2k_1}, \ldots, \pm \gamma_n q^{2k_n}) | K \in \mathbf{Z}^n\}$ in $\mathbf{R}^n_{\neq 0}$ such that the expression $(I(f))_{\cdot 1}|_{z \in L(\gamma)}$ is absolutely convergent. Of course if a generalized function is q^2-integrable, then it is lattice integrable for every choice of a lattice. One can easily see that the power series $E_{q^4}(-x_1^2) \cdots E_{q^4}(-x_n^2)$ is lattice integrable for $\gamma = (q^{n-1}, \ldots, q^{n-j}, \ldots, 1)$. Unfortunately, this power series is not the q^2-Gaussian $G_{q^2}(\underline{x})$ that we wanted to integrate, for $n \geqslant 2$. Besides, we can show that already for $n=2$, $G_{q^2}(\underline{x})$ is not lattice integrable although it is entire.

(Counter)example 1: Let us consider $G_{q^2}(\underline{x})$ for $n=2$. We write for simplicity $x_1 = x$ and $x_2 = y$, and $z_1 = z$, $z_2 = w$. Then

$$G_{q^2}(\underline{x}) = \sum_{k,l=0}^{\infty} \frac{(-1)^{k+l} q^{2k^2+2l^2-2k-2l} y^{2l} x^{2k}}{(q^4;q^4)_l (q^4;q^4)_k} = \sum_{l=0}^{\infty} \frac{(-1)^l q^{2l^2-2l} E_{q^4}(-(q^{2l}x)^2) y^{2l}}{(q^4;q^4)_l},$$

hence

$$\begin{aligned} G_{q^2}(\underline{x})_{\cdot 1} &= \sum_{k,l=0}^{\infty} \frac{(-1)^{k+l} q^{2k^2+2l^2-2k-2l} q^{4kl} z^{2k} w^{2l}}{(q^4;q^4)_l (q^4;q^4)_k} \\ &= \sum_{l=0}^{\infty} \frac{(-1)^l q^{2l^2-2l} E_{q^4}(-(q^{2l}z)^2) w^{2l}}{(q^4;q^4)_l} \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k q^{2k^2-2k} E_{q^4}(-(q^{2k}w)^2) z^{2l}}{(q^4;q^4)_k}, \end{aligned}$$

which is entire since it is majorized by $E_{q^4}(|w|^2)E_{q^4}(|z|^2)$. Now we wonder whether this expression is lattice integrable or not. In order to have that, we would need that for some $\gamma=(\gamma_1,\gamma_2)$,

$$\int_{-q\gamma_1\cdot\infty}^{q\gamma_1\cdot\infty}\int_{-\gamma_2\cdot\infty}^{\gamma_2\cdot\infty}|(G_{q^2})_{.1}(t_1,t_2)|\,d_{q^2}\underline{t}<\infty.$$

For this we would need that

$$\sum_{h_1=-\infty}^{\infty}\sum_{h_2=-\infty}^{\infty}q^{2|H|}\left|\sum_{l=0}^{\infty}\frac{(-1)^l q^{2l^2-2l}E_{q^4}(-(q^{2l+2h_1}q\gamma_1)^2)(q^{2h_2}\gamma_2)^{2l}}{(q^4;q^4)_l}\right|<\infty,$$

therefore we have to look at the limit for $h_j\to-\infty$ of the summands. Clearly by the discussion in Sec. 9 of Ref. 5, we see that we would need to have $\gamma_1=q$. With a similar reasoning we see that γ_2 must be equal to 1. Now, for general z and w we have

$$\begin{aligned}(G_{q^2}(\underline{x})_{.1})(z,w)&=\sum_{l=0}^{\infty}\frac{(-1)^l q^{2l^2-2l}w^{2l}(q^{4l}z^2;q^4)_\infty}{(q^4;q^4)_l}\\&=(z^2;q^4)_\infty\sum_{l=0}^{\infty}\frac{(-1)^l q^{2l^2-2l}w^{2l}}{(q^4;q^4)_l(z^2;q^4)_l}\\&=(z^2;q^4)_\infty{}_1\phi_1(0;z^2;q^4,w^2),\end{aligned}$$

which is the q^4 version of the q-Bessel function described in Ref. 9. For $(z,w)=(q^{2-2r},q^{2s})$ with $r\geqslant 0$ and s any integer, we have, by the estimates (2.6) and the following extimates in Ref. 9, that

$$|(G_{q^2}(\underline{x})_{.1})(q^{2-2r},q^{2s})|=q^{2r(r-1)}q^{4rs}(q^{4r+4};q^4)_\infty|{}_1\phi_1(0;q^{4r+4};q^4,q^{4r+4s})|.$$

For $r\to\infty$ and $s=-r$ this behaves like $q^{2r(r-1)-4r^2}\to\infty$. Hence $G_{q^2}(\underline{x})$ is not lattice integrable. ♠

Remarks: An analog of the symmetry (2.2) for q-Bessel functions in Ref. 9 holds in our case, namely,

$$E_{q^4}(-x_1^2)\,{}_1\phi_1(0;x_1^2;q^4,x_2^2)=G_{q^2}(\underline{x})={}_1\phi_1(0;x_2^2;q^4,x_1^2)E_{q^4}(-x_2^2),$$

once we agree that in ${}_1\phi_1$ every time we have a product of type $x_2^1/(x_1^2;q^4)_l$, the terms in x_1 have to be taken *before* the terms in x_2. Hence, in general, one has

$$E_q(-x_1-x_2)=E_q(-x_1)\,{}_1\phi_1(0;x_1;q,x_2)={}_1\phi_1(0;x_2;q,x_1)E_q(-x_2)$$

with the above meaning for ${}_1\phi_1$ in noncommuting variables.

Another equality involving a ${}_1\phi_1$ and exponentials in q-commuting variables is obtained by writing $E_q(-x_1)E_q(-x_2)$ as

$$(x_2;q)_\infty\sum_{l=0}^{\infty}\frac{(-1)^l q^{(1/2)(l^2-l)}(q^{-1}x_2;q)_l x_1^l}{(q;q)_l}$$

and using (3.12) in Ref. 5 with $x_1=-y$ and $x_2=-x$. Then one obtains

$$\sum_{l=0}^{\infty}\frac{(-1)^l q^{(1/2)(l^2-l)}x_1^l(x_2;q)_l}{(q;q)_l}=E_q(x_1x_2)E_q(-x_1),$$

where the sum on the left-hand side can be considered as a ${}_1\phi_1$ in noncommuting variables once assumed that x_1 always precedes x_2 in products. These facts were pointed out to me by T. Koornwinder. ♠

On the other hand, one can easily check that for every $A=(a_1,\ldots,a_n)$ in $\mathbf{R}^n_{>0}$ and every $E=(e_1,\ldots,e_n)$, then $x_1^{e_1}\cdots x_n^{e_n}E_{q^4}(-(a_1x_1)^2)\cdots E_{q^4}(-(a_nx_n)^2)$ is lattice integrable in the q^2-lattice generated by γ where $\gamma_j=a_j^{-1}q^{n-j+E^j}$. Unfortunately, lattice integrability carries a lot of technical work with it whenever one wants to prove anything like translation invariance, for instance. This is a consequence of the fact that, in order to state that the integral of $\partial_n^{e_n}\cdots\partial_1^{e_1}f(\underline{x})$ is zero, one needs to keep track of the lattice in which this series is integrable, which in general is not the same as the lattice in which $f(\underline{x})$ is integrable, unless e_j is even for every j. Indeed, consider $n=2$ and $f(\underline{x})=E_{q^4}(-x_1^2)E_{q^4}(-x_2^2)$. Then, $f(\underline{x})$ is integrable for $(\gamma_1,\gamma_2)=(q,1)$ while $\partial_2(f(\underline{x}))=-E_{q^4}(-q^2x_1^2)[x_2/(1-q^2)]E_{q^4}(-q^4x_2^2)$ is integrable for $(\gamma_1,\gamma_2)=(1,1)$. However, since ''morally'' the integral of a function which is odd in a variable is zero, we might as well *define* the integral of every odd function to be zero by changing the definition of the integral. Namely, we define the new integral I' to be the integral of the even part of the series $f(\underline{x})$. We formalize this definition.

Let $f(\underline{x})$ be any formal power series in the x_j's. We want to decompose it in 2^n series depending on the parity with respect to each variable. Let

$$\Pi_j^{\pm}:\hat{V}(R)^{\text{ext}}\to\hat{V}(R)^{\text{ext}}f(\underline{x})\mapsto\tfrac{1}{2}(f(\underline{x})\pm f(x_1,\ldots,x_{j-1},-x_j,x_{j+1},\ldots,x_n))$$

for every j and for any choice of $\pm$. This makes sense formally, and makes sense even concretely for the series in C. Clearly those operators commute; they are projections on the space of power series that are even (resp. odd) in the jth variable, so that $\Pi_j^+\Pi_j^-=0$ for every j. We define then for every choice of β in $\{\pm\}^n$ the operators $\Pi_\beta:\hat{V}(R)\to\hat{V}(R)$ as $(\Pi_1^{\beta_1})\circ\cdots\circ(\Pi_n^{\beta_n})$. They are all projections on their image E_β, and clearly the decomposition of the space of power series in the x_i's descends to a decomposition of the space C in 2^n spaces that we will call C_β. We also write $V^\beta:=C_\beta\triangleright 1$. We will denote $\Pi_{(+,\ldots,+)}$ by Π_0 for simplicity.

In particular, $\Pi_0 f(\underline{x})=2^{-n}\Sigma_{\varepsilon\in\{\pm1\}^n}f(\varepsilon_1x_1,\ldots,\varepsilon_nx_n)$ is even in every variable, and we define the integral I' to be the composition $I\circ\Pi_0$.

Remarks: Since we work in characteristic zero, $I'f$ is also formally equal to

$$\int_0^{x_n\cdot\infty}\cdots\int_0^{x_1\cdot\infty}\sum_{\varepsilon\in\{\pm\}^n}f(\varepsilon_1x_1,\ldots,\varepsilon_nx_n).$$

Clearly the class of I' integrable series is bigger than the class of I integrable series, since all odd series are integrable and their integral is zero. Since I' integrability of a series $f(\underline{x})$ coincides with I integrability of $\Pi_0 f(\underline{x})$, if $f(\underline{x})$ is a series which is even in all the variables, then $f(\underline{x})$ is I integrable $\Leftrightarrow f(\underline{x})$ is I' integrable since $f(\underline{x})=\Pi_0 f(\underline{x})$. ♠

One can also introduce lattice I' integrability. Again, for series in $C_{(+,\ldots,+)}$, lattice integrability and lattice I' integrability trivially coincide, and for a generic $f(\underline{x})$, lattice I' integrability trivially coincides with lattice I integrability of $\Pi_0 f(\underline{x})$ in the same lattice.

We can provide generalizations of Lemma 3.2, and Propositions 3.3 and 3.4 by introducing condition (c′) for an f such that $\Pi_0 f\in C'$:

(c′) For some $\eta>0$, there exists for each $J=(j_1,\ldots,j_n)\in\mathbf{Z}^n_{\geqslant 0}$ and $\beta\in\{\pm\}^n$ such that j_k is even (resp. odd) if $\beta_k=+$ (resp. $-$), some constant K_J such that

$$|(D_{1,q^2}^{j_1}\cdots D_{n,q^2}^{j_n}(\Pi_\beta f)_{\cdot 1})(z_1,\ldots,z_n)|\leqslant K_J\prod_{k=1}^{n}(1+|z_k|^2)^{-(1+\eta)}$$

if $\underline{z}\in\mathbf{R}^n$.

Then we have the following Lemma

Lemma 4.1: Let $f\in C'$ *satisfy condition* (c′). *Then for every* $J\neq \underline{0}$ *there holds* $(I'(\partial_n^{j_n}\cdots\partial_1^{j_1}f))\triangleright 1\equiv 0$, *so that we can conclude that* $I'(\partial_n^{j_n}\cdots\partial_1^{j_1}f)=0$.

Proof: $I'(\partial_n^{j_n}\cdots\partial_1^{j_1}f)=I(\partial_n^{j_n}\cdots\partial_1^{j_1}\Pi_\beta f)$ for β related to J as in condition (c′). □

Proposition 4.2: Let $f\in C'$ *satisfy* (c′). *Then* $(\mathrm{id}\otimes I')\Delta(f)=1\otimes(I'f)$. □

Proposition 4.3: Let $f(\underline{x})\in C$ *such that* (c′) *holds for every* $\eta>0$. *Then the statement of Proposition 3.2 holds for every polynomial* $p(\underline{x})$ *times* $f(\underline{x})$. □

We also have another invariance property that is analogous to the classical property (for $n=1$):

$$\int_{-\infty}^{\infty}\frac{1}{2}(f(x)+f(-x))\,dx=\int_{-\infty}^{\infty}\frac{1}{2}(f(x+y)+f(x-y))\,dx.$$

Proposition 4.4: Let $f(\underline{x})\in C'$ *satisfy* (c′). *Then* $(\mathrm{id}\otimes I)(\Pi_0\otimes\mathrm{id})\Delta(f)=1\otimes(I'f)$. *If* $f(\underline{x})$ *satisfies condition* (c′) *for every* $\eta>0$, *then the statement is true for every series of the form* $x_1^{e_1}\cdots x_n^{e_n}f(\underline{x})$.

Proof: The proof uses Taylor's formula with summation only on even j_k's. □

Observe that for even j_k's

$$\begin{aligned}(I(\partial_n^{j_n}\cdots\partial_1^{j_1}f))_{\cdot 1}&=(I'(\partial_n^{j_n}\cdots\partial_1^{j_1}f))_{\cdot 1}\\ &=q^{-\Sigma_k J^k}\int_{-q^{n-1}z_1\cdot\infty}^{q^{n-1}z_1\cdot\infty}\cdots\int_{-z_n\cdot\infty}^{z_n\cdot\infty}D_{1,q^2}^{j_1}\cdots D_{n,q^2}^{j_n}f_{\cdot 1}(t_1,\ldots,t_n)\,d_{q^2}t_n\cdots d_{q^2}t_1,\end{aligned}$$

hence the proposition above is interesting also because it can be proved for lattice integrability with simple changes in the hypothesis and in the proof. This reads as follows. Let $\gamma=(\gamma_1,\ldots,\gamma_n)\in\mathbf{R}^n$. We define the following spaces,

$$C(\gamma)=\{f(\underline{x})\in\hat{V}(R)^{\mathrm{ext}}|f_{\cdot 1}|_{z=\gamma}\ \text{is absolutely convergent}\}$$

and $C(q^{2K},\gamma)$ as the space of $f(\underline{x})\in C(\gamma)$ such that $f_{\cdot 1}$ can be continued analytically on a domain containing the q^2-lattice $L(\gamma)$ generated by γ. Clearly $C(\gamma)$ is closed with respect to the multiplication, hence it acts on the space V_γ of power series in commuting variables $z_1,\ldots,z_n$ that are absolutely convergent for $z=\gamma$, hence on a polydisc with polyradius $(|\gamma_1|,\ldots,|\gamma_n|)$. Let $f(\underline{x})$ be a series in $C(q^{2K}\gamma)$ for a given γ. Then it makes sense to investigate $I'(f(\underline{x}))_{\cdot 1}$ at $z_j=q^{n-j}\gamma_j$ and if this expression is absolutely convergent, then we say that $f(\underline{x})$ is lattice integrable. Actually, we would only need $\Pi_0(f)\in C(q^{2K},\gamma)$, but, since we want to compute integrals of products, we keep the restriction on $f(\underline{x})$.

Consider now the lattice version of condition (c):

(c″) Let $f(\underline{x})\in C(q^{2K},\gamma)$ be such that for every J with even entries $j_1,\ldots,j_n$, the Jackson partial derivatives $D_{1,q^2}^{j_1}\cdots D_{n,q^2}^{j_n}f_{\cdot 1}$ exist on the lattice $L(\gamma)$, and are such that

$$|(D_{1,q^2}^{j_1}\cdots D_{n,q^2}^{j_n}(f_{\cdot 1}))(q^{\pm 2k_1}\gamma_1,\ldots,q^{\pm 2k_n}\gamma_n)|=O(q^{2(1+\eta)K_-})$$

for $k_i\to\infty$, for some $\eta>0$, where K_- is the sum of the k_j's appearing with the minus sign.

We introduce the equivalence relation $\sim_\gamma$ between two expressions $f(\underline{x})$ and $g(\underline{x})$ belonging to $C(q^{2K},\gamma)$ as follows:

$$f(\underline{x})\sim_\gamma g(\underline{x})\Leftrightarrow f_{\cdot 1}(\underline{z})=g_{\cdot 1}(\underline{z}),\forall \underline{z}\in L(\gamma).$$

Proposition 4.5: Let $f(\underline{x})$ *satisfy condition* (c′) *for a given* γ, *and let* γ' *denote the n-tuple* $(q^{n-1}\gamma_1,\ldots,q^{n-j}\gamma_j,\ldots,\gamma_n)$. *Then we have the following.*

(i) *For every* J *such that every* j_k *is even,* $I'(\partial_n^{j_n}\cdots\partial_1^{j_1}f)\sim_{\gamma'}=0$.

(ii) $(\mathrm{id}\otimes I)(\Pi_0\otimes\mathrm{id})\Delta(f)\sim_{\gamma'}1\otimes(I'f)$.

Moreover, if f(x̲) satisfies condition (c″) *for every* η *and for every* $J\in\mathbf{Z}_{\geqslant 0}^n$, *then for every monomial* $x_1^{e_1}\cdots x_n^{e_n}$ we have the following.

(iii) $I'(\partial_n^{2j_n}\cdots\partial_1^{2j_1}(x_1^{e_1}\cdots x_n^{e_n}f(\underline{x})))\sim_{\gamma''}0$ *for every J, where* $\gamma''_j=q^{n-j+E^j}\gamma_j$.
(iv) $(\mathrm{id}\otimes I)(\Pi_0\otimes\mathrm{id})\Delta(x_1^{e_1}\cdots x_n^{e_n}f(\underline{x}))\sim_{\gamma''}1\otimes I'(x_1^{e_1}\cdots x_n^{e_n}f(\underline{x}))$ *where* γ'' *is as above.*

Proof: Statements (i) and (ii) are clear by the remark after the proof of Proposition 4.4. In order to prove (iii) we recall that

$$q^{4\Sigma_k J^k}I'(\partial_n^{2j_n}\cdots\partial_1^{2j_1}(x_1^{e_1}\cdots x_n^{e_n}f(\underline{x})))_{\cdot 1}$$
$$=\int_{-q^{n-1}z_1\cdot\infty}^{q^{n-1}z_1\cdot\infty}\cdots\int_{-z_n\cdot\infty}^{z_n\cdot\infty}D_{1,q^2}^{2j_1}\cdots D_{n,q^2}^{2j_n}(t_1^{e_1}\cdots t_n^{e_n}f_{\cdot 1}(q^{E^1}t_1,\ldots,q^{E^n}t_n))\,d_{q^2}t,$$

hence for $\underline{z}=\gamma''$ this expression converges, and it converges to zero. By invariance under q^2-shifts of the Jackson integral we get the statement. Statement (iv) follows from statement (iii). □

Remark: Observe that in the proof of (iii) in Proposition 4.5 the lattice in which we compute the equality depends only on the parity of the e_j's and that it is enough to be able to keep under control the partial Jackson derivatives of $(P_\beta f)_{\cdot 1}$ with $\beta_j=+(\mathrm{resp.}-)$ if e_j is even (resp. odd).♠

One may check that $E_{q^4}(-x_1^2)\cdots E_{q^4}(-x_n^2)$ satisfies all conditions of Proposition 4.5. Computations are left to the reader.

V. LATTICE ORDER INTEGRABILITY

We are still left with the problem that the q^2-Gaussian $G_{q^2}(\underline{x})$ is not lattice integrable, even with respect to I'. We have to weaken again our condition and introduce the concept of *lattice order integrability*. To simplify notation, we use analogs of I' instead of I. What we do is repeatedly apply a one-dimensional integral with respect to a noncommutative variable, say x_j. If this expression ''has a meaning'' (i.e., this expression applied to 1 converges after evaluation at $z_j=\gamma_j$), then we will identify it with a power series in noncommuting variables, in one variable less, and we are allowed to go further and repeat the procedure. Namely:

Definition 5.1: A formal power series $f(\underline{x})\in C$ *is said to be lattice order integrable (l.o. integrable) if there is an ordering of 1,...,n, denoted by the corresponding permutation* $\sigma\in S_n$, *and an n-tuple* $\gamma\in\mathbf{R}^n_{>0}$ *such that for every* $j\in\{1,...,n\}$ *the expression* $\int_{\sigma(j)}(I_{\sigma(j-1)}\cdots I_{\sigma(1)}f)$ *is entire, where* $\int_{\sigma(k)}g$ *and* $I_{\sigma(k)}g$ *are defined inductively as follows. For a formal power series f in* $\{x_1,\ldots,x_n\}-\{x_{\sigma(1)},\ldots,x_{\sigma(k-1)}\}$, $\int_{\sigma(k)}f$ *is the formal expression in the commuting variables* $\{z_1,\ldots,z_n\}-\{z_{\sigma(1)},\ldots,z_{\sigma(k)}\}$ *defined as*

$$\left(\int_{\sigma(k)}f\right)(\underline{z}):=\left(\int_{-x_{\sigma(k)}\cdot\infty}^{x_{\sigma(k)}\cdot\infty}\prod_0 f\right)_{\cdot 1}\Bigg|_{z_{\sigma(k)}=\gamma_{\sigma(k)}}.$$

If $\int_{\sigma(k)}f$ *is entire,* $I_{\sigma(k)}f$ *will denote the unique power series in the noncommuting indeterminates* $\{x_1,\ldots,x_n\}-\{x_{\sigma(1)},\ldots,x_{\sigma(k)}\}$ *such that* $(\int_{\sigma(1)}f)=(I_{\sigma(1)}f)_{\cdot 1}$. *If* $f(\underline{x})$ *is l.o. integrable, we define the constant* $I''_{(\sigma,\gamma)}f:=\int_{\sigma(n)}(I_{\sigma(n-1)}\cdots I_{\sigma(1)}f)$ *to be the lattice order integral of* $f(\underline{x})$ *associated to the order* σ *and the lattice L(γ).*

Clearly there is quite a difference between I and $I''_{(\sigma,\gamma)}$ since I maps formal power series to formal expressions in $x_1,\ldots,x_n$ while $I''_{(\sigma,\gamma)}$ maps l.o. integrable power series to constants. We will see later what the relation is between the two maps, on the space where they are both defined. We will also see in the examples that even if a power series is l.o. integrable for every order, it could still not be lattice integrable.

Observe that by definition of $\int_l$, power series that are odd in some variables are automatically defined to be l.o. integrable and that the integral will be zero for every choice of σ. For this reason, we will only investigate lattice order integrability for even power series. We can state a few results about lattice order integrability.

Proposition 5.2: Let $f(\underline{x})$ *be an even element of* $\hat{V}(R)^{\mathrm{ext}}$ *such that, for some* $\tau\in S_n$, *and for some power series in one indeterminate* $f_1,\ldots,f_n$, *we can write* $f(\underline{x})=f_{\rho(1)}(x_{\rho(1)})\cdots$

$f\rho(n)(x\rho(n))$ where $\rho=\tau^{-1}$. If $f(\underline{x})$ is l.o. integrable and every $\int_l f\neq 0$, then each f_j (viewed as a power series in one variable) is entire and lattice integrable. Conversely, if each f_j (viewed as a power series in one variable) is entire and lattice integrable, then $f(\underline{x})$ is l.o. integrable. In this case, $f(\underline{x})$ is lattice order integrable for every order σ and a suitable lattice depending on σ. Moreover, one has

$$I''_{(\sigma,\gamma)}(f(\underline{x}))=q^{\mathrm{l}(\sigma)+\mathrm{l}(\tau)}\prod_{j=1}^{n}\int_{-\gamma_j\cdot\infty}^{\gamma_j\cdot\infty}(f_j)_{\cdot 1}(t_j)\,d_{q^2}t_j\,,$$

where l *denotes the usual length of a permutation.*

Proof: ($\Rightarrow$) Suppose that $f(\underline{x})$ is as in the hypothesis, and that each f_j is entire and lattice integrable for a given $\tilde{\gamma}_j$. We write $f_j(x_j)=\Sigma_k c_{jk}x_j^k$ for every j. For an n-tuple K and for $p\in\{1,...,n\}$, we will also write

$$K_{\tau,p}:=\sum_{\substack{j>p\\ \tau(j)<\tau(p)}}k_j\quad\text{and}\quad K^{\tau,p}:=\sum_{\substack{j<p\\ \tau(j)>\tau(p)}}k_j\,.$$

We fix a σ. Then for $\sigma(1)=l$ and $\gamma_l=\tilde{\gamma}_l$ one has

$$\begin{aligned}\left(\int_l f\right)&(z_1,\ldots,z_{l-1},z_{l+1},\ldots,z_n)\\ &=2(1-q^2)\sum_{h=-\infty}^{\infty}q^{2h}\gamma_l\sum_{K'}c_{1k_1}\cdots\hat{c}_{lk_n}\cdots c_{nk_n}(q^{-1}z_1)^{k_1}\cdots(q^{-1}z_{l-1})^{k_{l-1}}\\ &\quad\times z_{l+1}^{k_{l+1}}\cdots z_n^{k_n}q^{\Sigma_{j\neq l}k_jK'_{\tau,j}}f_l(q^{2h+K'_{\tau,p}+K'^{\tau,p}}\gamma_l),\end{aligned}$$

where K' is the $(n-1)$-tuple obtained by K by deleting $\sigma(1)=l$, and

$$K'_{\tau,p}:=\sum_{\substack{l\neq j>p\\ \tau(j)<\tau(p)}}k_j\quad\text{and}\quad K^{\tau,p}:=\sum_{\substack{l\neq j<p\\ \tau(j)>\tau(p)}}k_j\,.$$

The last equality holds because

$$\begin{aligned}\sum_p k_pK_{\tau,p}&=k_lK_{\tau,l}+\sum_{p>l}k_pK_{\tau,p}+\sum_{p<l}k_pK_{\tau,p}\\ &=k_lK_{\tau,l}+\sum_{p>l}k_pK'_{\tau,p}+\sum_{p<l}k_pK'_{\tau,p}+\sum_{\substack{p<l\\ \tau(l)<\tau(p)}}k_pk_l\\ &=k_l(K_{\tau,l}+K^{\tau,l})+\sum_{p\neq l}k_pK'_{\tau,p}\,.\end{aligned}$$

By convergence of the q^2-Jackson integral of $f_{l\cdot 1}$, together with the fact that the other f_k's are entire and the fact that $K'_{\tau,p}+K'^{\tau,p}$ is an even number because the f_j's are even, one can invert the order of summation in the above sum, using dominated convergence. One gets

$$\left(\int_l f\right)(z_1,\dots,z_{l-1},z_{l+1},\dots,z_n)$$
$$=\sum_{K'} q^{\Sigma_{j\neq l}k_jK'_{\tau,j}}c_{1k_1}\cdots c_{nk_n}(q^{-1}z_1)^{k_1}\cdots(q^{-1}z_{l-1})^{k_{l-1}}z_{l+1}^{k_{l+1}}\cdots z_n^{k_n}$$
$$\times q^{-K'_{\tau,p}-K'^{\tau,p}}\int_{-q^{K'_{\tau,p}+K'^{\tau,p}}\gamma_l\cdot\infty}^{q^{K'_{\tau,p}+K'^{\tau,p}}\gamma_l\cdot\infty}(f_l)_{\cdot 1}(t_l)\,d_{q^2}t_l\,.$$

The above power series is entire since all the f_j's are, and one finds that

$$I_{\sigma(1)}f=\left(\int_{-\gamma_l\cdot\infty}^{\gamma_l\cdot\infty}(f_l)_{\cdot 1}(t_l)\,d_{q^2}t_l\right)\prod_{\substack{j\\ \rho(j)\neq l}} f_{\rho(j)}(q^{\eta_{\rho(j)}+\theta_{\rho(j)}}x_{\rho(j)})$$

with

$$\eta_k=\begin{cases}-1, & \text{if } k<\sigma(1),\\ 0, & \text{if } k>\sigma(1),\end{cases}\qquad \theta_k=\begin{cases}-1, & \text{if } k<l \text{ and } \tau(k)>\tau(l),\\ 1, & \text{if } k>l \text{ and } \tau(k)<\tau(l),\\ 0, & \text{otherwise.}\end{cases}$$

Therefore, we are again in the hypothesis of the proposition, but in the case $(n-1)$. Since the statement in one dimension is obvious, lattice order integrability is proved, considering a shifted lattice. For every new step we make, the argument of the f_j that still has to be integrated will be shifted by powers of q. If one goes through computations, one finds that the exponential of q in the shift of the argument of f_r with $r=\sigma(s)$ is

$$-\Sigma_s(\sigma,\tau)=-[\#\{j<s|\sigma(s)<\sigma(j)\}+\#\{j<s|(\sigma(j)-\sigma(s))(\tau\sigma(j)-\tau\sigma(s))<0\}],$$

hence the right lattice to integrate is the one defined by $\gamma_{\sigma(s)}=\tilde{\gamma}_{\sigma(s)}q^{\Sigma_s(\sigma,\tau)}$. In this setting, the integral will be the product of the q^2-Jackson integrals of the $f_{j\cdot 1}$'s multiplied by a power of q with exponent

$$\sum_{s=1}^n \Sigma_s(\sigma,\tau)=l(\sigma)+\sum_{s=1}^n \#\{j<s|(\sigma(j)-\sigma(s))(\tau\sigma(j)-\tau\sigma(s))<0\}=\mathrm{l}(\sigma)+\mathrm{l}(\tau),$$

since the second term in the sum is equal to the cardinality of

$$\{j,s|j<s\}\cap(\{j,s|\sigma(j)<\sigma(s),\tau\sigma(j)>\tau\sigma(s)\}\cup\{j,s|\sigma(s)<\sigma(j),\tau\sigma(s)>\tau\sigma(j)\}).$$

For the converse of the statement, one sees that if $f(\underline{x})$ can be written as a product of one-dimensional power series, those series have to be entire, and if there is a σ such that $f(\underline{x})$ is lattice order integrable, this means that $f_{\sigma(1)}$ is lattice integrable on $q^{2k_1}\gamma_{\sigma(1)}$, and so on, for the following f_j's, with shifted argument. By the $\Rightarrow$ part, we see that lattice order integrability has to hold for every σ'. □

Example 1: By the above proposition, for every $a_j\neq 0$ and for every $e_j\in\mathbf{Z}_{\geqslant 0}$ the formal power series

$$f(\underline{x})=x_n^{e_n}E_{q^4}(-a_n^2x_n^2)\cdots x_1^{e_1}E_{q^4}(-a_1^2x_1^2)$$

is l.o. integrable for every σ and $\gamma_{\sigma(k)}=a_{\sigma(k)}^{-1}q^{(k-1)+\#\{j<k|\sigma(k)<\sigma(j)\}}$, since in this case $\tau(k)=n-k+1$. In particular, $G_{q^2}(q^2\underline{x})$ and all products of type $G_{q^2}(q^2\underline{x})x_1^{e_1}\cdots x_n^{e_n}$ are l.o. integrable for every choice of the order σ. One has

$$I''_{(\sigma,\gamma)}f=\begin{cases} b_{q^2}^n\left[\prod_j (q^2;q^4)_{f_j}\right]q^{2n+2|E|}q^{\binom{n}{2}}q^{l(\sigma)}\prod_{j=1}^n a_j^{-1-e_j} & \text{if } e_j=2f_j \text{ for every } j,\\ 0, \quad \text{otherwise}, \end{cases}$$

where $b_{q^2}=(1-q^2)(q^2,-q^2,-1;q^2)_\infty$ and the result follows by Ref. 5. In particular we observe that the result depends on the choice of σ only in a straightforward way and that $L(\gamma)$ does not depend on E, but only on σ and the a_j's. Therefore one may consider the relation between $I''_{(\sigma,\gamma)}f$ and $I''_{(\sigma,\gamma)}(E_{q^4}(-\Sigma_k a_k^2x_k^2))$. One immediately sees that if all the e_j's are even,

$$I''_{(\sigma,\gamma)}(f)=\frac{(\Pi_j(q^2;q^4)_{f_j})q^{2|E|}}{(\Pi_j a_j^{e_j})}I''_{(\sigma,\gamma)}\left(E_{q^4}\left(-\sum_k a_k^2x_k^2\right)\right).$$

We say in this case (and whenever an equivalence of integrals $I''_{(\sigma,\gamma)}$ holds, with the same σ and γ on both sides) that $I(x_n^{e_n}E_{q^4}(-a_n^2x_n^2)\cdots x_1^{e_1}E_{q^4}(-a_1^2x_1^2))$ is ''*weakly equivalent*'' to $I(E_{q^4x}(-\Sigma_k a_k^2x_k^2))$. In particular, one has weak equivalence of the expression $I(x_n^{e_n}\times E_{q^4}(-q^2x_n^2)\cdots x_1^{e_1}E_{q^4}(-q^2x_1^2))$ and the expression $\Pi_j(q^2;q^4)_{f_j}I(G_{q^2}(q^2\underline{x}))$. We also want to point out that the above $f(\underline{x})$ is an example of the fact that one can have l.o. integrability for every order and still not have lattice integrability. ♠

Properties and Remarks:

(a) It is easy to check that if $f(\underline{x})$ is l.o. integrable for the order σ and the lattice $L(\gamma)$. Then, for every n-tuple of nonzero real numbers $(a_1,\ldots,a_n)$, the power series $f_A(\underline{x})=f(a_1x_1,\ldots,a_nx_n)$ is also l.o. integrable for the same order σ and for γ replaced by $\tilde{\gamma}$ where $\tilde{\gamma}_j=a_j^{-1}\gamma_j$ for every j. Then one has equivalence of the numbers $I''_{(\sigma,\gamma)}(f)=(a_1\cdots a_n)I''_{(\sigma,\tilde{\gamma})}(f_A)$.

(b) It is also obvious that if $f(\underline{x})$ is l.o. integrable for σ and γ. Then the resulting $I''_{(\sigma,\gamma)}(f)$ is invariant under shifts of each γ_j by even powers of q.

(c) In the definition of lattice order integrability the requirement on the $\int_{\sigma(k)}f$'s to be entire for every k can be weakened to analyticity. The weaker version of the definition is left to the reader. ♠

□

By the discussion above, one can conclude that for well-behaved series [by this we mean series satisfying condition (c), (c′), etc.] also the integral $I''_{(\sigma,\gamma)}$ is invariant under translation.

Remarks: The whole construction of lattice order integrability may look artificial, and it may seem to be a definition that is useful only in a noncommutative setting. However, this is not the case. One can define a similar concept of integrability also for power series in commutative variables.

VI. THE BRAIDED FOURIER TRANSFORM

Now we have all ingredients for the introduction of braided Fourier transforms on a subspace of $\hat{V}(R)^{\text{ext}}$. We introduce two transforms, related to each other by a shift in the arguments and the application of the antipode to one of them. As we already said, the first time that a Fourier transform for this kind of algebra appeared was in Ref. 3, from which we took inspiration. One of the goals of this section is to provide n-dimensional analogs to formulas (8.19)–(8.21), hence to Theorem 8.1 in Ref. 5. The difference with Ref. 5 is that in our version, the algebra $\hat{V}(R)^{\text{ext}}\otimes V(R)^{\text{ext}}$ has the braided product $(m\otimes m)(\text{id}\otimes\Psi\otimes\text{id})$ instead of the ordinary one, although in normal form his formulas and ours for $n=1$ coincide. The difference with Ref. 3 lies mainly in the fact that our integral is not bosonic [i.e., it does not have trivial braiding with elements of the algebras $\hat{V}(R)^{\text{ext}}$ and $V(R)^{\text{ext}}$]. The use of the antipode appears also in Ref. 4 where the case of finite-dimensional braided groups is treated. The following transforms behave nicely with respect to a convolution product and with respect to the action of $V(R)$ on $\hat{V}(R)^{\mathrm{ext}}$. They also respect various classical properties of the Fourier transform. These facts are developed in Refs. 3 and 13.

We say that an element $f(\underline{x})$ of $\hat{V}(R)^{\mathrm{ext}}$ is of class $\mathcal{I}$ if $f(\underline{x})x_1^{e_1}\cdots x_n^{e_n}$ is I' integrable for every monomial $x_1^{e_1}\cdots x_n^{e_n}$. We say that it is of class $\mathcal{I}_{(\sigma,\gamma)}$ if for every monomial $x_1^{e_1}\cdots x_n^{e_n}$, the power series $f(\underline{x})x_1^{e_1}\cdots x_n^{e_n}$ is lattice order integrable for σ and γ.

Again, we do not provide a complete classification of $\mathcal{I}$, but we give a class for which this makes sense, which is big enough to reach our goal. Indeed, power series satisfying condition (c) of Sec. III belong to $\mathcal{I}$, hence products of $e_{q^2}(-x_j^2)$ and polynomials belong to $\mathcal{I}$ provided that for every $j\in\{1,...,n\}$, the one-dimensional q^2-Gaussian $e_{q^2}(-x_j^2)$ appears in the product.

Definition 6.1: The braided Fourier transforms F and F_S are defined on the class $\mathcal{I}$ and they have images in $\hat{V}(R)^{\mathrm{ext}}\otimes V(R)^{\mathrm{ext}}$. They are given by

$$F:=(I'\otimes\mathrm{id})(m\otimes\mathrm{id})\left(\mathrm{id}\otimes\exp\left(x\left|\frac{i}{(1-q^2)}(\partial_1,\dots,q^{-j+1}\partial_j,\dots,q^{-n+1}\partial_n)\right)\right.\right),$$

$$F_S:=(I'\otimes S)(m\otimes\mathrm{id})\left(\mathrm{id}\otimes\exp\left(x\left|\frac{iq^2}{(1-q^2)}(q^{n-1}\partial_1,\dots,q^{n-j}\partial_j,\dots,\partial_n)\right)\right.\right).$$

For an $f(\underline{x})\in\mathcal{I}\cap C_\beta$ one has that

$$F(f(\underline{x})):=(I'\otimes\mathrm{id})\left(f(\underline{x})E_{q^2}\left(i\sum_{j=1}^{n}x_j\otimes q^{-(j-1)}\partial_j\right)\right)$$

$$=\sum_{\substack{e_1,\dots,e_n\\A(e_j)=\beta_j}}I(f(\underline{x})x_1^{e_1}\cdots x_n^{e_n})\otimes\frac{i^{|E|}q^{-\Sigma_jE^j}}{\Pi_{j=1}^n(q^2;q^2)_{e_j}}\partial_n^{e_n}\cdots\partial_1^{e_1}$$

and

$$F_S(f(\underline{x})):=(I'\otimes S)\left(f(\underline{x})E_{q^2}\left(iq^2\sum_{j=1}^{n}x_j\otimes q^{(n-j)}\partial_j\right)\right)$$

$$=\sum_{\substack{e_1,\dots,e_n\\A(e_j)=\beta_j}}I(f(\underline{x})x_1^{e_1}\cdots x_n^{e_n})\otimes\frac{(-i)^{|E|}q^{|E|^2-\Sigma_je_jE_j+|E|+\Sigma_jE_j}}{\Pi_{j=1}^n(q^2;q^2)_{e_j}}\partial_1^{e_1}\cdots\partial_n^{e_n},$$

where $A(e_j)=+(\mathrm{resp.}-)$ if e_j is even (resp. odd). Here we used that $S(\partial_n^{e_n}\cdots\partial_1^{e_1})=(-1)^{|E|}q^{|E|^2-|E|-\Sigma_je_jE_j}\partial_1^{e_1}\cdots\partial_n^{e_n}$. It is clear that the second components in the tensor product of $F(f(\underline{x}))$ and of $F_S(f(\underline{x}))$ will also have parity β. In order to provide formulas analogous to (8.21) and (8.19) in Ref. 5, we need to compute F_S for

$$M(\underline{x},A)=e_{q^4}(-x_1^2)x_1^{a_1}\cdots e_{q^4}(-x_n^2)x_n^{a_n}=x_1^{a_1}\cdots x_n^{a_n}e_{q^4}\left(-\sum_j(q^{-A_j}x_j)^2\right)$$

and for $H(\underline{x},A)=e_{q^4}(-x_1^2)\tilde{h}_{a_1}(x_1;q^2)\cdots e_{q^4}(-x_n^2)\tilde{h}_{a_n}(x_n;q^2)$ where the $\tilde{h}_{a_j}$'s are the *discrete q-Hermite II polynomials* (see Ref. 14 and references therein) that are defined by

$$\tilde{h}_l(z;q):=z^l{}_2\phi_1(q^{-n},q^{-n+1};0;q^2,-q^2z^{-2})=(q;q)_l\sum_{k=0}^{[1/2]}\frac{(-1)^kq^{-2kl+k(2k+1)}{}_zl-2k}{(q^2;q^2)_k(q;q)_{l-2k}}.$$

Both $M(\underline{x},A)$ and $H(\underline{x},A)$ satisfy condition (c) of Sec. III, so that the transform in defined on both series. We first compute the transform F_S on a generic $f(\underline{x})$. In order to give a meaning to the transform we apply the realization map π_γ sending a power series $g(\underline{x})$ to $g_{\cdot 1}(\underline{z})$, followed by evaluation at $\underline{z}=\gamma$, to the first component of $F_S(f(\underline{x}))$. By the assumption that $f(\underline{x})\in\mathcal{I}$ we know that this is well defined so that $(\pi_\gamma\otimes\mathrm{id})F_S(f(\underline{x}))$ is a genuine power series in the noncommuting ∂_j's. By the computations in the previous section, one obtains, for an $f(\underline{x})\in C_\beta\cap\mathcal{I}$,

$$(\pi_\gamma\otimes\mathrm{id})(F_S(f(\underline{x})))=\sum_{\substack{e_1,\ldots,e_n\\ A(e_j)=\beta_j}}\frac{(-i)^{|E|}q^{\Sigma_j(e_j^2+e_j)}}{\Pi_{k=1}^n(q^2;q^2)_{e_k}}\times\left(\int_{-\gamma_1 q^{n-1+E_1}\cdot\infty}^{\gamma_1 q^{n-1+E_1}\cdot\infty}\cdots\int_{-\gamma_n q^{E_n}\cdot\infty}^{\gamma_n q^{E_n}\cdot\infty}f_{\cdot 1}(\underline{t})t_1^{e_1}\cdots t_n^{e_n}d_{q^2}\underline{t}\right)\partial_1^{e_1}\cdots\partial_n^{e_n}.$$

By invariance of the q^2-integral we see that the integration bounds do not depend on E, but only on the parity of its components, hence they only depend on β. In particular, for $f(\underline{x})=f_1(x_1)\cdots f_n(x_n)\in\mathcal{I}\cap C_\beta$ one has that

$$(\pi_\gamma\otimes\mathrm{id})F_S(f(\underline{x}))=\prod_{k=1}^n\left[\sum_{A(e_k)=\beta_k}\frac{(-i)^{e_k}q^{e_k^2+e_k}\partial_k^{e_k}}{(q^2;q^2)_{e_k}}\left(\int_{-q^{B(\beta)_k+n-k}\gamma_k\cdot\infty}^{q^{B(\beta)_k+n-k}\gamma_k\cdot\infty}[(f_k)_{\cdot 1}(t_k)]t_k^{e_k}d_{q^2}t_k\right)\right],$$

where the product is taken in *increasing* order and $B(\beta)=(b(\beta)_1,\ldots,b(\beta)_n)$ is the n-tuple $\{0,1\}^n$ such that the kth entry is 0 (resp. 1) if β_k is even (resp. odd) and $B(\beta)_k=\Sigma_{j=1}^{k-1}b(\beta)_j$ as usual.

Hence we come to an n-dimensional version of formula (8.21) in Ref. 5. Let $M(\underline{x},A)$ as above. We remind the reader that in this case $\beta_j=+(\text{resp.}-)$ if a_j is even (resp. odd). Then we use the one-dimensional case to obtain our result. Indeed, expanding in power series the left-hand side of (8.21) and using q^2 instead of q one has

$$c_{q^2}(\gamma)q^{-a^2-a}i^a h_a(t;q^2)E_{q^4}(-q^4t^2)=\sum_{\substack{k\geqslant 0\\ k+a\ \text{even}}}\frac{i^k q^{k^2+k}}{(q^2;q^2)_k}\left(\int_{-\gamma\cdot\infty}^{\gamma\cdot\infty}x^{a+k}e_{q^4}(-x^2)\,d_{q^2}x\right)t^k$$
$$=\sum_{\substack{k\geqslant 0\\ k+a\ \text{even}}}\frac{i^k q^{k^2+k}}{(q^2;q^2)_k}c_{q^2}(\gamma)q^{-(a+k)^2/2}(q^2;q^4)_{(a+k)/2}t^k.$$

Using the above formula we obtain

$$(\pi_\gamma\otimes\mathrm{id})F_S(M(\underline{x},A))=\prod_{j=1}^n\left[c_{q^2}(q^{n-j+B(\beta)_j}\gamma_j)\times\sum_{\substack{e_j\geqslant 0\\ e_j+a_j\ \text{even}}}\left(\frac{(-i)^{e_j}q^{e_j^2+e_j}}{(q^2;q^2)_{e_j}}q^{-(e_j+a_j)^2/2}(q^2;q^4)_{e_j+a_j/2}\right)\partial_j^{e_j}\right]$$
$$=(-i)^{|A|}\left[\prod_{j=1}^n c_{q^2}(q^{n-j+B(\beta)_j}\gamma_j)\right]q^{\Sigma_j a_j(1-a_j)}\times\prod_{k=1}^n[E_{q^4}(-q^4\partial_k^2)h_{a_k}(\partial_k;q^2)],\tag{1}$$

where

(i) the product is taken in *increasing* order, and

(ii) $h_l(z;q^2)$ is the discrete q^2-*Hermite I* polynomial of degree l (see Ref. 14 and references therein) and is defined as

$$h_l(z;q^2):=z^l{}_2\phi_0(q^{-2l},q^{2-2l};;q^4,q^{4l-2}z^{-2})=(q^2;q^2)_l\sum_{k=0}^{[l/2]}\frac{(-1)^k q^{2k(k-1)}z^{l-2k}}{(q^4;q^4)_k(q^2;q^2)_{l-2k}}.$$

We observe that the only part of $(\pi_\gamma\otimes\mathrm{id})F_S(M(\underline{x},A))$ involving the γ_j's is the coefficient, equal to $[\int_{-q^{B(\beta)_n x_n}\cdot\infty}^{q^{B(\beta)_n x_n}\cdot\infty}\cdots\int_{-q^{B(\beta)_1 x_1}\cdot\infty}^{q^{B(\beta)_1 x_1}\cdot\infty}g_{q^2}]_{\cdot 1}|_{\underline{z}=\gamma}$, i.e., it is a shifted integral of the Gaussian $g_{q^2}(\underline{x})$ where the shift only depends on the parity of the function $M(\underline{x},A)$, i.e., only on the parity of the a_j's. So, we conclude that

$$F_S(e_{q^4}(-x_1^2)x_1^{a_1}\cdots e_{q^4}(-x_n^2)x_n^{a_n})=\left[\int_{-q^{B(\beta)_n x_n}\cdot\infty}^{q^{B(\beta)_n x_n}\cdot\infty}\cdots\int_{-q^{B(\beta)_1 x_1}\cdot\infty}^{q^{B(\beta)_1 x_1}\cdot\infty}g_{q^2}\right]$$
$$\otimes(-i)^{|A|}q^{\Sigma_j a_j(1-a_j)}\prod_{k=1}^{n}\,[E_{q^4}(-q^4\partial_k^2)h_{a_k}(\partial_k;q^2)]. \qquad (2)$$

The above result gives the analog of the classical reciprocity between Gaussians times a monomial and rescaled Gaussians times a Hermite polynomial under the Fourier transform in $\mathbf{R}^n$. From the above result we derive an analog of formula (8.19) in Ref. 5, for $H(\underline{x},A)$ defined above. $H(\underline{x},A)$ is also contained in one of the subspaces C_β since each $\tilde{h}_a(x_j;q^2)$ has constant parity. We obtain

$$F_S(e_{q^4}(-x_1^2)\tilde{h}_{a_1}(x_1;q^2)\cdots e_{q^4}(-x_n^2)\tilde{h}_{a_n}(x_n;q^2))$$
$$=\left[\int_{-q^{B(\beta)_n x_n}\cdot\infty}^{q^{B(\beta)_n x_n}\cdot\infty}\cdots\int_{-q^{B(\beta)_1 x_1}\cdot\infty}^{q^{B(\beta)_1 x_1}\cdot\infty}g_{q^2}\right]$$
$$\otimes\prod_{k=1}^{n}\left[(q^2;q^2)_{a_k}\sum_{S_k=0}^{[1/2a_k]}\frac{(-i)^{a_k}q^{-a_k^2+a_k}}{(q^4;q^4)_{s_k}(q^2;q^2)_{a_k-2s_k}}h_{a_k-2s_k}(\partial_k;q^2)E_{q^4}(-q^4\partial_k^2)\right]$$
$$=\left[\int_{-q^{B(\beta)_n x_n}\cdot\infty}^{q^{B(\beta)_n x_n}\cdot\infty}\cdots\int_{-q^{B(\beta)_1 x_1}\cdot\infty}^{q^{B(\beta)_1 x_1}\cdot\infty}g_{q^2}\right](-i)^{|A|}q^{\Sigma_k(a_k-a_k^2)}\otimes\prod_{j=1}^{n}\,(\partial_j^{a_j}E_{q^4}(-q^4\partial_j^2)), \qquad (3)$$

where the last equality follows from (10) in Ref. 5 and the product is taken in increasing order.

Observe that for well-behaved functions, and for an n-tuple $A=(a_1,\ldots,a_n)$ of nonzero real numbers, the braided Fourier transform of $f(a_1x_1,\ldots,a_nx_n)$ can be obtained by the braided Fourier transform of $f(\underline{x})$. More precisely, $(\pi_\gamma\otimes\mathrm{id})F_S(f(a_1x_1,\ldots,a_nx_n))=(\Pi_j a_j)^{-1}(\pi_{\tilde{\gamma}}\otimes\mathrm{id})\times(F_S(f(\underline{x})))(a_1^{-1}\partial_1,\ldots,a_n^{-1}\partial_n)$, where $\tilde{\gamma}$ denotes the n-tuple obtained by γ multiplying each component γ_j by a_j. Clearly, similar results holds for $(\pi_\gamma\otimes\mathrm{id})F$, hence they hold for F_S and for F.

Now we want to compute the braided Fourier transform for monomials or polynomials times a q^2-Gaussian of type $G_{q^2}(\underline{x})$. We cannot use the same definition since $G_{q^2}(\underline{x})$ is not even lattice integrable. Therefore, we introduce a weaker notion of braided Fourier transform.

Definition 6.2: The ''weak'' braided Fourier transforms $F''(\sigma,\gamma)$ and $F''_S(\sigma,\gamma)$ are defined on the class $\mathcal{I}_{(\sigma,\gamma)}$. They map this class to $V(R)^{\mathrm{ext}}$ and they are defined as

$$F''(\sigma,\gamma):=(I''_{(\sigma,\gamma)}\otimes\mathrm{id})(m\otimes\mathrm{id})\left(\mathrm{id}\otimes\exp\left(x\left|\frac{i}{(1-q^2)}(\partial_1,\ldots,q^{1-j}\partial_j,\ldots,q^{1-n}\partial_n)\right)\right)\right),$$

$$F''_S(\sigma,\gamma):=(I''_{(\sigma,\gamma)}\otimes S)(m\otimes\mathrm{id})\left(\mathrm{id}\otimes\exp\left(x\left|\frac{iq^2}{(1-q^2)}(q^{n-1}\partial_1,\ldots,q^{n-j}\partial_j,\ldots,\partial_n)\right)\right)\right).$$

We will use this new notion in order to derive an n-dimensional version of (8.20) in Ref. 5. Namely, we will derive the transform $F''(\sigma,\gamma)$ of the formal power series $N(\underline{x},A) = E_{q^4}(-q^4x_n^2)x_n^{a_n}\cdots E_{q^4}(-q^4x_1^2)x_1^{a_1}$ for given positive integers $a_1,\ldots,a_n$. One has to compute $I''_{(\sigma,\gamma)}(N(\underline{x},A)x_1^{e_1}\cdots x_n^{e_n})$ for every E with each $e_j\equiv a_j \pmod 2$, for some fixed σ and γ. This makes sense by the computations in Sec. V because $N(\underline{x},A)x_1^{e_1}\cdots x_n^{e_n}$ is equal to

$$q^{-\Sigma_j E^j(e_j+a_j)}E_{q^4}(-q^{4-2E^n}x_n^2)x_n^{e_j+a_j}\cdots E_{q^4}(-q^{4-2E^1}x_1^2)x_1^{e_1+a_1},$$

which is l.o. integrable for every σ, with $\gamma_{\sigma(k)} = q^{(k-1)+A^{\sigma(k)}+\#\{j<k|\sigma(j)>\sigma(k)\}}$. In particular, since we showed that the resulting integrals differ only by a factor $q^{l(\sigma)}$, we compute it only for $\sigma=\mathrm{id}$ and $\gamma_k = q^{k-1+A^k}$. Then, denoting by $N'(\underline{x},A)$ the power series obtained by $N(\underline{x},A)$ by multiplying the argument of the $E_{q^4}(-x_j^2)$ by q^{-E^j} for every j, one has

$$F''(\sigma,\gamma)(N(\underline{x},A)) = \sum_{\substack{e_1,\ldots,e_n\\ e_j+a_j=2h_j}} \frac{i^{|E|}q^{-\Sigma_j E_j - \Sigma_j E^j(e_j+a_j)}}{\Pi_j(q^2;q^2)_{e_j}}(I''_{(\mathrm{id},\gamma)}(N'(\underline{x},A+E)))\partial_n^{e_n}\cdots\partial_1^{e_1}.$$

Hence

$$\begin{aligned}F''(\sigma,\gamma)(N(\underline{x},A)) &= (-1)^{|A|}I''_{(\mathrm{id},\tilde{\gamma})}\left(E_{q^4}\left(-\sum_j q^4x_j^2\right)\right)\prod_{j=n}^{1}\left[\sum_{e_j+a_j=2h_j}\frac{(-i)^{e_j}(q^2;q^4)_{h_j}}{\Pi_j(q^2;q^2)_{e_j}}\partial_j^{e_j}\right]\\ &= q^{\Sigma_j(a_j^2-a_j)}i^{|A|}b_{q^2}^n q^{\binom{n}{2}}\left(\prod_{j=n}^{1}\tilde{h}_{a_j}(\partial_j;q^2)e_{q^4}(-\partial_j^2)\right).\end{aligned} \tag{4}$$

Here $\tilde{\gamma}$ denotes the n-tuple such that $\tilde{\gamma}_k = q^{A^k}\gamma_k$ and the product is taken in decreasing order. For the last equality we used (9.15) and (8.20) in Ref. 5.

Using the definition of the h_{a_j}'s, formula (4) above, and formula (8.17) in Ref. 5 one also gets the following result:

$$\begin{aligned}&F''(\sigma,\gamma)(E_{q^4}(-q^4x_n^2)h_{a_n}(x_n;q^2)\cdots E_{q^4}(-q^4x_1^2)h_{a_1}(x_1;q^2))\\ &\quad= i^{|A|}q^{\Sigma_j(a_j^2-a_j)}I''_{(\sigma,\tilde{\gamma})}\left(E_{q^4}\left(-q^4\sum_j x_j^2\right)\right)\prod_{j=n}^{1}\partial_j^{a_j}e_{q^4}(-\partial_j^2),\end{aligned} \tag{5}$$

where the product is taken in *decreasing* order and $\tilde{\gamma}$ is given by $\tilde{\gamma}_k = q^{A^k}\gamma_k$ for every k as before.

VII. INTEGRAL ON $V(R)^{\mathrm{ext}}$ AND INVERSE TRANSFORM

We provide now an inverse for the braided Fourier transforms, at least on the subspaces of the image of $\mathcal{I}$ and $\mathcal{I}_{(\sigma,\chi)}$. In order to do this we need also the integral on $V(R)^{\mathrm{ext}}$. Since there is a symmetry between $\hat{V}(R)^{\mathrm{ext}}$ and $V(R)^{\mathrm{ext}}$, one can simply repeat the definitions and computations keeping in mind that whenever we had a left action involving $\hat{V}(R)^{\mathrm{ext}}$, we will need a right action in the case of $V(R)^{\mathrm{ext}}$. We will only provide the necessary formulas, while the properties and the proofs of similar statements as those of Secs. III–V are left to the reader. We observe that all the results in this Section can be achieved both by direct computation or by using the symmetry $\psi:\hat{V}(R)^{\mathrm{ext}}\to V(R)^{\mathrm{ext}}$ defined in Sec. II.

Similarly as for $\hat{V}(R)^{\mathrm{ext}}$, $\hat{V}(R)^{\mathrm{ext}}$ acts on the right on $V(R)^{\mathrm{ext}}$ by braided partial differentiation. For a monomial $\partial_n^{e_n}\cdots\partial_1^{e_1}$, and for $j\in\{1,\ldots,n\}$ we have

$$(\partial_n^{e_n}\cdots\partial_1^{e_1})\leftharpoonup x_j = [e_j]_{q^2}\partial_n^{e_n}\cdots\partial_{j+1}^{e_{j+1}}\partial_j^{e_j-1}(q\partial_{j-1})^{e_{j-1}}\cdots(q\partial_1)^{e_1}.$$

There holds a right version of Taylor's formula, namely,

$$\Delta(g(\underline{\partial}))=g(\Delta(\underline{\partial}))=g(\underline{\partial})\leftharpoondown\left(\sum_{e_1,\ldots,e_n\geqslant 0}\frac{x_1^{e_1}\cdots x_n^{e_n}\otimes\partial_n^{e_n}\cdots\partial_1^{e_1}}{[e_1]_{q^2}!\cdots[e_n]_{q^2}!}\right).$$

The (indefinite) q^2-integral acting from the right is

$$g(\underline{\partial})\leftharpoondown\int_0^{\partial_i}:=(1-q^2)\sum_{k=0}^{\infty}g(\partial_n,\ldots,q^{2k}\partial_i,q^{-2}\partial_{i-1},\ldots,q^{-2}\partial_1)q^{2k}\partial_i,$$

and again as in the case of $\hat{V}(R)^{\text{ext}}$, one can define $\int_0^{a\partial_i}$ for a nonzero constant a. The global integral is formally obtained as the limit for all $r_j\to\infty$ of

$$g(\underline{\partial})\leftharpoondown\int_{-q^{-2r_1}\partial_1}^{q^{-2r_1}\partial_1}\cdots\int_{-q^{-2r_n}\partial_n}^{q^{-2r_n}\partial_n},$$

where

$$g(\underline{\partial})\leftharpoondown\int_{-a\partial_j}^{a\partial_j}:=\left(g(\underline{\partial})\leftharpoondown\int_0^{a\partial_j}\right)-\left(g(\underline{\partial})\leftharpoondown\int_0^{-a\partial_j}\right)$$

for every a. Hence we have formally

$$g(\underline{\partial})\leftharpoondown\int^{\underline{\partial}}:=g(\underline{\partial})\leftharpoondown\int_{-\partial_1\cdot\infty}^{\partial_1\cdot\infty}\cdots\int_{-\partial_n\cdot\infty}^{\partial_n\cdot\infty}$$

$$=(1-q^2)^n\sum_{\underline{\varepsilon}\in\{\pm1\}^n}\sum_{k_n=-\infty}^{\infty}\cdots\sum_{k_1=-\infty}^{\infty}g(\varepsilon_n q^{2k_n}\partial_n,\ldots,\varepsilon_1 q^{2k_1}\partial_1)q^{2|K|}\partial_1\cdots\partial_n.$$

As in Sec. III, one can define an action of $V(R)$ on the power series in the n commuting indeterminates $z_1,\ldots,z_n$, in order to give a meaning to the integral. One can use this action to define integrability, lattice integrability, and lattice order integrability as in Secs. IV and V, and we leave this to the reader. The action will be the right regular action after the choice of a normal form. It is denoted by $\triangleleft$ and it is defined on monomials as

$$(z_1^{k_1}\cdots z_n^{k_n})\triangleleft\partial_n^{e_n}\cdots\partial_1^{e_1}:=q^{\Sigma_j K_j e_j}z^{e_1+k_1}\cdots z_n^{e_n+k_n}=z_1^{k_1}(q^{K_1}z_1)^{e_1}\cdots z_n^{k_n}(q^{K_n}z_n)^{e_n}.$$

For simplicity we will denote $1\triangleleft g(\underline{\partial}):={}_{1.}g(\underline{z})$ for every expression $g(\underline{\partial})$ for which the action on 1 makes sense.

As in Sec. IV we construct projections $P_j^{\pm}$ defined on $V(R)$ and $V(R)^{\text{ext}}$ for every $j=1,\ldots,n$ and every choice of $+$ or $-$ as follows:

$$P_j^{\pm}:V(R)^{\text{ext}}\to V(R)^{\text{ext}}g(\underline{\partial})\mapsto\tfrac{1}{2}[g(\underline{\partial})\pm g(\ldots,-\partial_j,\ldots)].$$

Again, the $P_j^{\pm}$'s commute with each other, they are projections on the subspaces of $V(R)^{\text{ext}}$ consisting of even (resp. odd) elements with respect to the jth variable, and $P_j^{+}P_j^{-}=0$ for every j. Then, for every $\beta\in\{-,+\}^n$ we define P_β as the composite $P_1^{\beta_1}\circ\cdots\circ P_n^{\beta_n}$. The P_β's are all projections on their image G_β and clearly the decomposition $V(R)^{\text{ext}}=\oplus_\beta G_\beta$ corresponds to the decomposition of $\mathbf{C}[[\underline{z}]]$ in series that are either odd or even in each variable after applying the action on 1. We write P_0 for $P_{(+,\ldots,+)}$. We can define again the integral J' defined by $J'g:=(P_0 g)\leftharpoondown\int^{\underline{\partial}}$.

In particular one can check that for a $g(\underline{\partial})$ for which this makes sense one has

$$_{1\cdot}\left((P_0(g(\underline{\partial})))\leftharpoonup\int^{\underline{\partial}}\right)=\int_{-z_1q^{n-1}\cdot\infty}^{z_1q^{n-1}\cdot\infty}\cdots\int_{-z_nq^{n-n}\cdot\infty}^{z_nq^{n-n}\cdot\infty}{}_{1\cdot}(P_0(g))(\underline{t})\,d_{q^2}\underline{t}$$

and

$$_{1\cdot}(g\leftharpoonup D_1^{j_1}\cdots D_n^{j_n})=(D_{q^2}^{J}(1\cdot g))(q^{J^1}z_1,\ldots,q^{J^n}z_n),$$

We also observe that under the ''symmetry'' $\psi:\hat{V}(R)^{\mathrm{ext}}\to V(R)^{\mathrm{ext}}$ mapping x_j to $q^{-j+(1/2)(n-1)}\partial_{n-j+1}$ one has

$$\psi(If(x_1,\ldots,x_n))=q^{-n}\left[(\psi(f(\underline{x}))(\underline{\partial}))\leftharpoonup\int_{-q^{-n+1}\partial_1\cdot\infty}^{q^{-n+1}\partial_1\cdot\infty}\cdots\int_{-q^{-n+2j-1}\partial_j\cdot\infty}^{q^{-n+2j-1}\partial_j\cdot\infty}\cdots\int_{-q^{n-1}\partial_n\cdot\infty}^{q^{n-1}\partial_n\cdot\infty}\right].\qquad(6)$$

One defines the integral $J''_{(\sigma,\gamma)}$ for l.o. integrable power series as the right-handed version of $\hat{I}''_{(\sigma,\gamma)}$. We will need $J''_{(\sigma,\gamma)}(E_{q^4}(-q^4\partial_1^2)\partial_1^{a_1},\ldots,E_{q^4}(-q^4\partial_n^2)\partial_n^{a_n})$. One checks as for the case of $\hat{V}(R)^{\mathrm{ext}}$ that the integrand is actually is l.o. integrable for every σ for a suitable choice of γ, and that the results differ only by a factor $q^{\mathrm{l}(\sigma)}$. In particular, for $\sigma=\mathrm{id}$ one needs $\gamma_k=q^{(k-1)+A^k}$ and one gets

$$J''_{(\sigma,\gamma)}(E_{q^4}(-q^4\partial_1^2)\partial_1^{a_1},\ldots,E_{q^4}(-q^4\partial_n^2)\partial_n^{a_n})=\begin{cases}q^{\binom{n}{2}}b_{q^2}^{n}\prod_j(q^2;q^4)_{b_j}, & \text{if } a_j=2b_j\,\forall j,\\ 0, & \text{otherwise.}\end{cases}$$

Hence,

$$J''_{(\sigma,\gamma)}(E_{q^4}(-q^4\partial_1^2)\partial_1^{a_1}\ldots,E_{q^4}(-q^4\partial_n^2)\partial_n^{a_n})=\left(J''_{(\sigma,\gamma)}\left(E_{q^4}\left(-q^4\sum_j\partial_j^2\right)\right)\right)\prod_j(q^2;q^2)_{b_j}$$

if $a_j=2b_j$ for every j.

Properties like right invariance of the integral, nullity of the integral of the partial derivative of a power series, etc., can be proved as in Sec. III–V.

We introduce now an inversion formula for F_S and F and their weak analogs. We will use the symmetry between $\hat{V}(R)^{\mathrm{ext}}$ and $V(R)^{\mathrm{ext}}$ and the results at the end of the previous section in order to provide an analog of Theorem 8.1 in Ref. 5. An inversion formula for the braided Fourier transform is to be found in Ref. 3, but the authors had the hypothesis that the integral is ''bosonic,'' i.e., it has a trivial braiding with $\hat{V}(R)$, or $V(R)$ for $n\geqslant 2$, which is not our case as the reader can easily check (see also Ref. 15 for a few remarks about this property of the integral). Moreover, the element *vol* in Ref. 3 is not necessarily convergent.

We say that a power series $g(\underline{\partial})$ in $V(R)^{\mathrm{ext}}$ is *of class* $\mathcal{J}$ if every monomial times $g(\underline{\partial})$ is J' integrable. This is possible, for instance, if $g(\underline{\partial})$ satisfies conditions similar to condition (c) of Sec. III. We say that $g(\underline{\partial})$ is *of class* $\mathcal{J}_{(\sigma,\gamma)}$ if there is an order σ and a lattice $L(\gamma)$ such that every monomial times $g(\underline{\partial})$ is lattice order integrable for σ and γ.

Definition 7.1: We define the linear maps G, $G_S:\mathcal{J}\to\hat{V}(R)^{\mathrm{ext}}\otimes V(R)^{\mathrm{ext}}$ *by*

$$G:=\left[(\mathrm{id}\otimes m)\left(\exp\left(x\left|\frac{i}{(1-q^2)}(\partial_1,\ldots,q^{-j+1}\partial_j,\ldots,q^{-n+1}\partial_n)\right.\right)\otimes\mathrm{id}\right)\right]\leftharpoonup(\mathrm{id}\otimes J'),$$

$$G_S:=\left[(S\otimes m)\left(\exp\left(x\left|\frac{iq^2}{(1-q^2)}(q^{n-1}\partial_1,\ldots,q^{n-j}\partial_j,\ldots,\partial_n)\right.\right)\otimes\mathrm{id}\right)\right]\leftharpoonup(\mathrm{id}\otimes J').$$

For instance, for G_S, the transform of a given $g(\underline{\partial})\in\mathcal{J}$ of fixed parity β will be

$$G_S(g(\underline{\partial}))=\left((S\otimes\mathrm{id})\left(E_{q^2}\left(iq^2\sum_j q^{(n-j)}x_j\otimes\partial_j\right)\right)g(\underline{\partial})\right)\leftharpoondown(\mathrm{id}\otimes J')$$

$$=\sum_{\substack{e_1,\dots,e_n\\ A(e_j)=\beta_j}}\frac{(-i)^{|E|}q^{|E|^2+|E|}q^{\Sigma_j(n-j)e_j}}{\Pi_j(q^2;q^2)_{e_j}}x_1^{e_1}\cdots x_n^{e_n}\otimes(\partial_n^{e_n}\cdots\partial_1^{e_1}g(\underline{\partial})\leftharpoondown(\mathrm{id}\otimes J')).$$

If $\tau: V(R)^{\mathrm{ext}}\otimes\hat{V}(R)^{\mathrm{ext}}\to\hat{V}(R)^{\mathrm{ext}}\otimes V(R)^{\mathrm{ext}}$ denotes the usual flip operator, putting $c_j=q^{-j+(1/2)(n-1)}$ for every $j=1,...,n$ we observe after some computations that, for $f(\underline{x})=f_1(x_1)\cdots f_n(x_n)$ with f_j of parity β_j, there holds

$$\tau(\psi\otimes\psi^{-1})F_S(f(\underline{x}))=\sum_{\substack{e_1,\dots,e_n\\ A(e_j)=\beta_j}}\frac{(-i)^{|E|}q^{|E|^2+|E|+\Sigma_jE^j}}{\Pi_j(q^2;q^2)_{e_j}}(c_1^{-1}x_1)^{e_n}\cdots(c_j^{-1}x_j)^{e_{n-j+1}}\cdots(c_n^{-1}x_n)^{e_1}$$

$$\otimes[\partial_n^{e_1}\cdots\partial_1^{e_n}f_1(\partial_n)\cdots f_n(\partial_1)]\leftharpoondown\int_{-q^{-n+1+E_n-E_1}c_n\partial_1\cdot\infty}^{q^{-n+1+E_n-E^1}c_n\partial_1\cdot\infty}\cdots\int_{-c_1q^{n-1+E_1-E^n}\partial_n\cdot\infty}^{c_1q^{n-1+E_1-E^n}\partial_n\cdot\infty}.$$

Hence, we see that the formal expression of $\tau(\psi\otimes\psi^{-1})F_S(f(\underline{x}))$ coincides with the formal expression of

$$G_S(f_1(\partial_n)\cdots f_n(\partial_1))((c_1^{-1}x_1,\dots,c_n^{-1}x_n)\otimes(q^{n-1+E_1-E^1}c_1\partial_n,\dots,q^{E_n-E^n+1-n}c_n\partial_1)).$$

In the same way one shows that the formal expression of $\tau(\psi\otimes\psi^{-1})F(f(\underline{x}))$ coincides with

$$G(f_1(\partial_n)\cdots f_n(\partial_1))((c_1^{-1}x_1,\dots,c_n^{-1}x_n)\otimes(q^{n-1+E_1-E^1}c_1\partial_n,\dots,q^{E_n-E^n+1-n}c_n\partial_1)).$$

We can use the symmetry between F_S and G_S, together with formula (2), in order to compute $G_S(\partial_n^{a_n}e_{q^4}(-\partial_n^2)\cdots\partial_1^{a_1}e_{q^4}(-\partial_1^2))$ for given positive integers $a_1,\dots,a_n$ of parity, respectively, $\beta_1,\dots,\beta_n$. This symmetry tells us that

$$G_S(e_{q^4}(-\partial_n^2)\partial_n^{a_n}\cdots e_{q^4}(-\partial_1^2)\partial_1^{a_1})=(L_\psi\otimes L'_{\psi,\beta})[\tau(\psi\otimes\psi^{-1})\tau F_S(e_{q^4}(-x_1^2)x_1^{a_n}\cdots e_{q^4}(-x_n^2)x_n^{a_1})],$$

where L_ψ is the shift operator mapping x_j to c_jx_j and $L'_{\psi,\beta}$ is the shift operator mapping ∂_j to $q^{n-2j+1+B(\beta)^j-B(\beta)_j}c_{n-j+1}^{-1}\partial_j$. Then the above expression is equal to

$$(-i)^{|A|}q^{\Sigma_j(a_j-a_j^2)}\prod_{j=n}^{1}[E_{q^4}(-q^4x_j^2)h_{a_j}(x_j;q^2)]\otimes L'_{\psi,\beta}\psi\int_{-q^{B(\beta)^1}x_n\cdot\infty}^{q^{B(\beta)^1}x_n\cdot\infty}\cdots\int_{-q^{B(\beta)^n}x_1\cdot\infty}^{q^{B(\beta)^n}x_1\cdot\infty}g_{q^2}$$

$$=(-i)^{|A|}q^{\Sigma_j(a_j-a_j^2)}\prod_{j=n}^{1}[E_{q^4}(-q^4x_j^2)h_{a_j}(x_j;q^2)]$$

$$\otimes q^{-n}L'_{\psi,\beta}\left(g_{q^2}(c_1\partial_n,\dots,c_n\partial_1)\leftharpoondown\int_{-q^{B(\beta)^1-n+1}\partial_1\cdot\infty}^{q^{B(\beta)^1-n+1}\partial_1\cdot\infty}\cdots\int_{-q^{B(\beta)^n+n-1}\partial_n\cdot\infty}^{q^{B(\beta)^n+n-1}\partial_n\cdot\infty}\right).$$

Hence we can conclude that

$$G_S(e_{q^4}(-\partial_n^2)\partial_n^{a_n}\cdots e_{q^4}(-\partial_1^2)\partial_1^{a_1})=(-i)^{|A|}q^{\Sigma_j(a_j-a_j^2)}\prod_{j=n}^{1}[E_{q^4}(-q^4x_j^2)h_{a_j}(x_j;q^2)]$$

$$\otimes\left(g_{q^2}\leftharpoondown\int_{-q^{B(\beta)_1}\partial_1\cdot\infty}^{q^{B(\beta)_1}\partial_1\cdot\infty}\cdots\int_{-q^{B(\beta)_n}\partial_n\cdot\infty}^{q^{B(\beta)_n}\partial_n\cdot\infty}\right), \tag{7}$$

where the second component of the tensor product clearly depends only on the parity of the a_j's. Formula (7) can also be obtained by direct computation. Using the definition of the $\tilde{h}_m(z;q^2)$, with the same relation as before between the a_j's and the β_j's, one obtains

$$G_S(\tilde{h}_{a_n}(\partial_n;q^2)e_{q^4}(-\partial_n^2)\cdots\tilde{h}_{a_1}(\partial_1;q^2)e_{q^4}(-\partial_1^2))$$
$$=(-i)^{|A|}q^{\Sigma_j(a_j-a_j^2)}\prod_{j=n}^{1}[E_{q^4}(-q^4x_j^2)x_j^{a_j}]\otimes\left(g_{q^2}\leftharpoonup\int_{-q^{B(\beta)_1}\partial_1\cdot\infty}^{q^{B(\beta)_1}\partial_1\cdot\infty}\cdots\int_{-q^{B(\beta)_n}\partial_n\cdot\infty}^{q^{B(\beta)_n}\partial_n\cdot\infty}\right),\quad(8)$$

which is the $V(R)^{\mathrm{ext}}$ version of (3). By these results we can conclude the following:

Proposition 7.2: Let $\beta=(\beta_1,\dots,\beta_n)\in\{\pm 1\}^n$, σ *be any order, and* γ *be the n-tuple with components given by* $\gamma_{\sigma(k)}=q^{B(\beta)^{\sigma(k)}+k-1+\#\{j<k|\sigma(j)>\sigma(k)\}}$. *Then on power series in* $\hat{V}(R)^{\mathrm{ext}}$ *of type* $E_{q^4}(-q^4x_n^2)p_n(x_n)\cdots E_{q^4}(-q^4x_1^2)p_1(x_1)$, *where the* p_i*'s are polynomials of fixed parity* $\beta_1,\dots,\beta_n$, *there holds*

$$G_S\circ F''(\sigma,\gamma)=\mathrm{id}\otimes(I''_{(\sigma,\tilde{\gamma})}G_{q^2}(q^2\underline{x}))\left(g_{q^2}\leftharpoonup\int_{-q^{B(\beta)_1}\partial_1\cdot\infty}^{q^{B(\beta)_1}\partial_1\cdot\infty}\cdots\int_{-q^{B(\beta)_n}\partial_n\cdot\infty}^{q^{B(\beta)_n}\partial_n\cdot\infty}\right),$$

where $\tilde{\gamma}$ *is such that* $\tilde{\gamma}_k=\gamma_k q^{B(\beta)_k}$ *for every k. Therefore for power series in* $V(R)^{\mathrm{ext}}$ *of the form* $w_n(\partial_n)e_{q^4}(-\partial_n^2)\cdots w_1(\partial_1)e_{q^4}(-\partial_1^2)$, *where the* w_j*'s are polynomials of fixed parity* $\beta_1,\dots,\beta_n$, *one has*

$$(F''(\sigma,\gamma)\otimes\mathrm{id})G_S=\mathrm{id}\otimes(I''_{(\sigma,\tilde{\gamma})}G_{q^2}(q^2\underline{x}))\left(g_{q^2}\leftharpoonup\int_{-q^{B(\beta)_1}\partial_1\cdot\infty}^{q^{B(\beta)_1}\partial_1\cdot\infty}\cdots\int_{-q^{B(\beta)_n}\partial_n\cdot\infty}^{q^{B(\beta)_n}\partial_n\cdot\infty}\right)$$

with $\tilde{\gamma}$ *as before.*

Proof: It follows by (5) and (7). □

A slightly more general version of this proposition holds by considering a proper γ and q^2-Gaussians where the argument is multiplied by a nonzero constant.

We also want to consider another inverse transform, the weak transform inverting (2) and (3).

Definition 7.3: The ''weak'' transforms $G''(\sigma,\gamma)$ *and* $G''_S(\sigma,\gamma)$ *map* $\mathcal{J}_{(\sigma,\gamma)}$ to $\hat{V}(R)^{\mathrm{ext}}$ *and are defined as*

$$G''(\sigma,\gamma)\coloneqq\left[(\mathrm{id}\otimes m)\left(\exp\left(x\left|\frac{i}{(1-q^2)}(\partial_1,\dots,q^{-j+1}\partial_j,\dots,q^{-n+1}\partial_n)\right)\otimes\mathrm{id}\right)\right]\leftharpoonup(\mathrm{id}\otimes J''_{(\sigma,\gamma)}),$$

$$G''_S(\sigma,\gamma)\coloneqq\left[(S\otimes m)\left(\exp\left(x\left|\frac{iq^2}{(1-q^2)}(q^{n-1}\partial_1,\dots,q^{n-j}\partial_j,\dots,\partial_n)\right)\otimes\mathrm{id}\right)\right]\leftharpoonup(\mathrm{id}\otimes J''_{(\sigma,\gamma)}).$$

As in formulas (4) and (5), one finds, for $\sigma=\mathrm{id}$ and $\gamma_k=q^{A^k+k-1}$ (and similarly for different σ's),

$$G''_{(\mathrm{id},\gamma)}(E_{q^4}(-q^4\partial_1^2)\partial_1^{a_1}\cdots E_{q^4}(-q^4\partial_n^2)\partial_n^{a_n})=q^{\Sigma_j(a_j^2-a_j)}i^{|A|}\left(\prod_{j=1}^{n}\tilde{h}_{a_j}(x_j;q^2)e_{q^4}(-x_j^2)\right)$$
$$\times[J''_{(\mathrm{id},\tilde{\gamma})}(G_{q^2}(q^2\underline{\partial}))]\quad(9)$$

and

$$G''_{(\mathrm{id},\gamma)}(E_{q^4}(-q^4\partial_1^2)h_{a_1}(\partial_1)\cdots E_{q^4}(-q^4\partial_n^2)h_{a_n}(\partial_n))$$
$$=q^{\Sigma_j(a_j^2-a_j)}i^{|A|}\left(\prod_{j=1}^{n}x_j^{a_j}e_{q^4}(-x_j^2)\right)[J''_{(id,\tilde{\gamma})}(G_{q^2}(q^2\underline{\partial}))],\quad(10)$$

where in both formulas $\tilde{\gamma}_k = q^{k-1}$ for every k and the product is taken in increasing order. These formulas can be obtained by using the symmetry or by direct computation. One has the second inversion property.

Proposition 7.4: Let $\beta = (\beta_1, \ldots, \beta_n) \in \{\pm 1\}^n$, σ *be any order, and* γ *be the n-tuple with components given by* $\gamma_{\sigma(k)} = q^{B(\beta)^{\sigma(k)} + k - 1 + \#\{j<k | \sigma(j) > \sigma(k)\}}$. *Then on power series in* $V(R)^{\mathrm{ext}}$ *of type* $E_{q^4}(-q^4 \partial_1^2) p_1(\partial_1) \cdots E_{q^4}(-q^4 \partial_n^2) p_n(\partial_n)$, *where the* p_j*'s are polynomials of fixed parity* $\beta_1, \ldots, \beta_n$, *there holds*

$$F_S \circ G''(\sigma,\gamma) = \left[\int_{-q^{B(\beta)_n} x_n \cdot \infty}^{q^{B(\beta)_n} x_n \cdot \infty} \cdots \int_{-q^{B(\beta)_1} x_1 \cdot \infty}^{q^{B(\beta)_1} x_1 \cdot \infty} g_{q^2} \right] (J''_{(\sigma,\tilde{\gamma})}(G_{q^2}(q^2 \underline{\partial}))) \otimes \mathrm{id},$$

where $\tilde{\gamma}_{\sigma(k)} = \gamma_{\sigma(k)} q^{B(\beta)^{\sigma(k)}}$. *Therefore for power series in* $\hat{V}(R)^{\mathrm{ext}}$ *having the form* $w_1(x_1) \times e_{q^4}(-x_1^2) \cdots w_n(x_n) e_{q^4}(-x_n^2)$ *where the* w_j*'s are polynomials of fixed parity* $\beta_1, \ldots, \beta_n$, *one has*

$$(\mathrm{id} \otimes G''(\sigma,\gamma)) F_S = \left[\int_{-q^{B(\beta)_n} x_n \cdot \infty}^{q^{B(\beta)_n} x_n \cdot \infty} \cdots \int_{-q^{B(\beta)_1} x_1 \cdot \infty}^{q^{B(\beta)_1} x_1 \cdot \infty} g_{q^2} \right] (J''_{(\sigma,\tilde{\gamma})}(G_{q^2}(q^2 \underline{\partial}))) \otimes \mathrm{id}$$

with $\tilde{\gamma}$ *as before.*

Proof: It follows by (2), (3), (9), and (10). □

One observes that in this case the Plancherel measure is always a product of integrals of q^2-Gaussians, but the integration bounds depend on the parity of the power series. So these transforms could also be seen as sine and cosine transforms (see for this also Ref. 9 where the q-sine and q-cosine transforms in commuting variables are defined).

Remark: The break in symmetry appearing in q^2-integration is a phenomenon that has recently been observed in Ref. 7. Sometimes this lack of symmetry can be avoided, for instance if the generalized function $f(\underline{x})$ that we want to integrate (and transform) is lattice integrable in $L(\gamma)$ and $L(q\gamma)$. In this case we could replace q^2-integration by q-integration since $\int_{-\gamma\cdot\infty}^{\gamma\cdot\infty} f_{\cdot 1}(t)\, d_q t = \int_{-\gamma\cdot\infty}^{\gamma\cdot\infty} f_{\cdot 1}(t)\, d_{q^2} t + \int_{-q\gamma\cdot\infty}^{q\gamma\cdot\infty} f_{\cdot 1}(t)\, d_{q^2} t$. The new defined integral will be again invariant under translation. The q-integral of a q^2-Gaussians $g_{q^2}(\underline{x})$ times a monomial will be similar to the q^2-integral of the same expression. Using the q-integral in the definition of $(\pi_\gamma \otimes \mathrm{id}) F_S$ will provide results similar to formula (1) but with $\Pi_j c_{q^2}(q^{n-j+B(\beta)_j} \gamma_j)$ replaced by $\Pi_j (c_{q^2}(\gamma_j) + c_{q^2}(q\gamma_j))$. The result will be therefore independent of the parity of the a_j's. However, this approach cannot be used for $G_{q^2}(\underline{x})$, since we have seen that there is only one q^2-lattice for which $G_{q^2}(\underline{x})$ is lattice order integrable. ♠

ACKNOWLEDGMENTS

The author would like to thank Professor T. H. Koornwinder for many discussions, and for his useful comments and remarks.

This research was financially made possible by the NWO-Stichting SWON, under Project No. 610-06-100.

[1] S. Majid, *Foundations of Quantum Group Theory* (Cambridge U. P., Cambridge, England, 1995).
[2] M. Hashimoto and T. Hayashi, ''Quantum multilinear algebra,'' Tohoku Math. J. **44**, 471–521 (1992).
[3] A. Kempf and S. Majid, ''Algebraic q-integration and Fourier theory on quantum and braided spaces,'' J. Math. Phys. **35**, 6802–6837 (1994).
[4] C. Chryssomalakos, ''Remarks on quantum integration,'' Commun. Math. Phys. **184**, 1–25 (1997).
[5] T. H. Koornwinder, ''Special functions and q-commuting variables,'' in *Special Functions, q-series and Related Topics*, edited by M. E. H. Ismail, D. R. Masson, and M. Rahman, Fields Institute Communications 14, AMS (1997), pp. 131–166.
[6] M. Olshanetsky and V. Rogov, ''The q-Fourier transform of q-distributions,'' IHES preprint q-alg 9712055 (1997).
[7] B. L. Cerchiai, R. Hinterding, J. Madore, and J. Wess, ''A Calculus Based on a q-deformed Heisenberg Algebra,'' QA/9809160 (1998).
[8] T. Masuda, K. Mimachi, Y. Nakagami, M. Noumi, and K. Ueno, ''Representations of the quantum group $SU_q(2)$ and the little q-Jacobi polynomials,'' J. Funct. Anal. **99**, 357–387 (1991).

[9] T. H. Koornwinder and R. F. Swarttouw, ‘‘On *q*-analogues of the Fourier and Hankel transforms,’’ Trans. Am. Math. Soc. **333**, 445–461 (1992).
[10] V. V. Lyubashenko and S. Majid, ‘‘Braided groups and quantum Fourier transform,’’ J. Alg. **166**, 506–528 (1994).
[11] S. Majid, ‘‘Free braided differential calculus, braided binomials theorem and the braided exponential map,’’ J. Math. Phys. **34**, 4843–4856 (1993).
[12] M. P. Schützenberger, ‘‘Une interpétation de certaines solutions de l'équation fonctionelle: $F(x+y)=F(x)F(y)$,’’ C. R. Acad. Sci. Paris **236**, 352–353 (1953).
[13] G. Carnovale and T. Koornwinder, ‘‘A *q*-analogue of convolution on the line’’ (preprint).
[14] R. Koekoek and R. Swarttouw, ‘‘The Askey-scheme of hypergeometric orthogonal polynomials and its *q*-analogue,’’ Revised Version of Report 94-05 Delft University of Technology, Faculty TWI (1996).
[15] C. Chryssomalakos and B. Zumino, ‘‘Translations, integrals and Fourier transforms in the quantum plane,’’ in *Salamfestschrift*, Proceedings of the ‘‘Conference on Highlights of Particles and condensed matter physics,’’ edited by A. Ali, J. Ellis, and S. Randjbar-Daemi (ICTP, Trieste, Italy, 1993).

Two-parameter differential calculus on the *h*-superplane

Salih Çelik[a)]
Mimar Sinan University, Department of Mathematics, 80690 Besiktas, Istanbul, Turkey

Sultan A. Çelik
Yildiz Technical University, Department of Mathematics, Sisli, Istanbul, Turkey

(Received 16 April 1999; accepted for publication 3 June 1999)

We introduce a noncommutative differential calculus on the two-parameter *h*-superplane via a contraction of the (p,q)-superplane. We manifestly show that the differential calculus is covariant under $GL_{h_1,h_2}(1|1)$ transformations. We also give a two-parameter deformation of the $(1+1)$-dimensional phase space algebra. [S0022-2488(99)04010-4]

I. INTRODUCTION

Quantum groups are a generalization of the concept of classical groups. The theory of quantum groups has become an important branch of mathematical physics and a new branch of mathematics. An approach to obtain the quantum groups is to identify the elements of a quantum group with the linear transformations of a space with noncommuting coordinates. It is known, from the work of Woronowicz,[1] that one can define a consistent differential calculus on the noncommutative space of a quantum group. Thus quantum group is a concrete example of noncommutative differential geometry.[2]

During the past few years, Wess–Zumino[3] have developed a differential calculus on the quantum (hyper)plane which is covariant under the action of the quantum group $GL_q(n)$. The natural extension of their scheme to superspace[4] was given by Soni[5] and the two-parameter differential calculus on the superplane has been worked out by Chung.[6] A differential calculus on the *h*-plane was given by Karimipour[7] and the two-parameter analog was introduced by Aghamohammadi.[8]

In this paper we construct a two-parameter differential calculus on the quantum *h*-superplane using the methods of Ref. 9. The paper is organized as follows: in Sec. II we obtain the (h_1,h_2)-superplanes via a contraction from the (p,q)-superplanes. We define derivatives and differentials on the (h_1,h_2)-superplane of noncommuting coordinates and give their commutation rules. In Sec. III we manifestly show that the differential calculus is covariant under the action of the quantum supergroup $GL_{h_1,h_2}(1|1)$ of Ref. 10. We give a two-parameter deformation of the $(1+1)$-dimensional phase space algebra in Sec. IV, and in the following section we show that the (p,q)-deformed superoscillator algebra satisfies the undeformed super-oscillator algebra when objects are transformed into new objects such that they are singular for certain values of the deformation parameters.

II. DIFFERENTIAL CALCULUS ON THE *h*-SUPERPLANE

In this work we denote (p,q)-deformed objects by primed quantities. Unprimed quantities represent transformed coordinates. As usual, we shall always assume that even (bosonic) objects commute with everything and odd (Grassmann) objects anticommute among themselves. Before discussing the two-parameter differential calculus on the *h*-superplane we give some notations and useful formulas in the following section. This first section closely follows the approach of Ref. 10.

[a)]Electronic mail: scelik@fened.msu.edu.tr

A. Quantum *h*-superplane

Quantum superplane is an associative coordinate algebra $\mathcal{A}_q$ equipped with a set $\{x', \theta'\}$ of generators x', θ'. The commutation relations of the generators is defined by[4]

$$x'\theta' - q\theta' x' = 0. \tag{1a}$$

$$\theta'^2 = 0, \tag{1b}$$

where q is a nonzero complex deformation parameter. The coordinates neither commute nor anticommute unless $q \to \pm 1$, respectively. In this work we shall use the limits $p \to 1, q \to 1$ to make a contraction.

We now introduce new coordinates x and θ, in terms of x' and θ', by[10]

$$\begin{aligned} x &= x' - \frac{h_1}{p-1}\theta', \\ \theta &= -\frac{h_2}{q-1}x' + \left(1 - \frac{h_1 h_2}{(p-1)(q-1)}\right)\theta'. \end{aligned} \tag{2}$$

Using relation (1), it is easy to verify that

$$x\theta = q\theta x + h_2 x^2, \tag{3a}$$

where the new deformation parameter h_2 commutes with the coordinate x and anticommutes with the coordinate θ. Similarly, from (1a) one obtains

$$\theta^2 = -h_2 \theta x. \tag{3b}$$

where

$$h_1 h_2 = -h_2 h_1, \quad h_1^2 = 0 = h_2^2. \tag{4}$$

That is, the new deformation parameters h_1 and h_2 are odd (Grassmann) numbers which anticommute. Note that although in the $p \to 1$, $q \to 1$ limits the transformation (2) is ill behaved, the resulting commutation relations are well defined.

The relations (3) define a new deformation,[11] which we called the h_2-deformation of the algebra of coordinate functions on the Manin superplane generated by x and θ in the limit $q \to 1$, and will be denoted by $\mathcal{A}_{h_2}$.

Differential calculus on the quantum superplane $\mathcal{A}_{h_2}$ requires the introduction of differentials $\mathrm{d}x, \mathrm{d}\theta$. The complete framework also includes the commutation relations of these differentials with the coordinates and derivatives.

B. Relations between coordinates and differentials

To establish a noncommutative differential calculus on the quantum superplane $\mathcal{A}_{h_2}$, we assume that the commutation relations between the coordinates and their differentials have the following form:

$$\begin{aligned} x'\mathrm{d}x' &= A\,\mathrm{d}x' x', \\ x'\mathrm{d}\theta' &= C_{11}\mathrm{d}\theta' x' + C_{12}\mathrm{d}x'\theta', \\ \theta'\mathrm{d}x' &= C_{21}\mathrm{d}x'\theta' + C_{22}\mathrm{d}\theta' x', \\ \theta'\mathrm{d}\theta' &= B\,\mathrm{d}\theta'\theta'. \end{aligned} \tag{5}$$

Now we would like to transform these relations to unprimed quantities to determine the coefficients *A*, *B*, and C_{ij}. We first introduce the exterior differential d.

The exterior differential d is an operator which gives the mapping from the coordinates to the differentials

$$\mathsf{d}: Z^i \rightarrow \mathsf{d}Z^i, \tag{6}$$

where $Z^1=x$, $Z^2=\theta$, $\mathsf{d}Z^1=\mathsf{d}x$, and $\mathsf{d}Z^2=\mathsf{d}\theta$. We demand that the exterior differential d has to satisfy two properties; the nilpotency

$$\mathsf{d}^2=0, \tag{7}$$

and the graded Leibniz rule

$$\mathsf{d}(FG)=(\mathsf{d}F)G+(-1)^{\hat{F}}F(\mathsf{d}G), \tag{8}$$

where $\hat{F}$ is the Grassmann degree of *F*, that is, $\hat{F}=0$ for even variables and $\hat{F}=1$ for odd variables. We wish to substitute into (5) the differentials $\mathsf{d}x'$ and $\mathsf{d}\theta'$ together with the coordinates x' and θ'. The deformation parameters h_1 and h_2 are both odd numbers and the exterior differential d is also odd. Therefore the action of the exterior differential d on αu is defined by

$$\mathsf{d}(\alpha u)=(-1)^{\hat{\alpha}}\alpha \mathsf{d}u, \tag{9}$$

where α is a number (even or odd) and *u* is a coordinate of superplane. So we can write from (2)

$$\begin{aligned} \mathsf{d}x &= \mathsf{d}x' + \frac{h_1}{p-1}\mathsf{d}\theta', \\ \mathsf{d}\theta &= \frac{h_2}{q-1}\mathsf{d}x' + \left(1-\frac{h_1h_2}{(p-1)(q-1)}\right)\mathsf{d}\theta'. \end{aligned} \tag{10}$$

Note that if we consider *x* and θ as functions of two variables (say x' and θ') and differentiate (2), as usual, then we do not obtain the expressions in (10). To obtain (10) one must take the differential from the left in Eq. (2). In the Appendix, we explain this in detail.

We now substitute (2) and (10) into (5) which are not explicitly written here. It will be calculate the coefficients *A*, *B* and C_{ij}. We first assume that

$$\mathsf{d}x'\mathsf{d}\theta' = p^{-1}\mathsf{d}\theta'\mathsf{d}x', \quad (\mathsf{d}x')^2=0. \tag{11}$$

Then we have

$$\mathsf{d}\theta\mathsf{d}x = p\mathsf{d}x\mathsf{d}\theta - h_1(\mathsf{d}\theta)^2, \tag{12a}$$

and

$$(\mathsf{d}x)^2 = h_1\mathsf{d}x\mathsf{d}\theta. \tag{12b}$$

Consequently, the coefficients are determined as follows:

$$\begin{aligned} &A \ \text{ undetermined}, \quad B=1, \\ &C_{11}=q, \quad C_{12}=pq-1, \quad C_{21}=0, \quad C_{22}=-p. \end{aligned} \tag{13}$$

Here we shall choose *A* equal to *pq* since the relations are then well defined.

C. Relations of derivatives and coordinates

In this section we shall define the derivatives and find the commutation relations of derivatives with coordinates and the commutation relations between derivatives. We first introduce the matrix[10]

$$g=\begin{pmatrix} 1+h_1h_2/(p-1)(q-1) & h_1/(p-1) \\ h_2/(q-1) & 1 \end{pmatrix}. \tag{14}$$

It is easy to verify that the matrix g is a supermatrix. Thus we can write the transformation in (2) of the form

$$Z'=gZ, \quad Z'=\begin{pmatrix} x' \\ \theta' \end{pmatrix}. \tag{15}$$

Let us denote the partial derivatives with respect to x' and θ' by

$$\partial_{x'}=\frac{\partial}{\partial x'}, \quad \partial_{\theta'}=\frac{\partial}{\partial \theta'},$$

respectively. The transformation law of the partial derivatives is then defined by

$$\partial'=(g^{\mathrm{st}})^{-1}\partial, \quad \partial=\begin{pmatrix} \partial_x \\ \partial_\theta \end{pmatrix}, \tag{16}$$

where g^{st} denotes the supertranspose of g. Explicitly

$$\partial_{x'}=\partial_x-\frac{h_2}{q-1}\partial_\theta, \quad \partial_{\theta'}=\frac{h_1}{p-1}\partial_x+\left(1-\frac{h_1h_2}{(p-1)(q-1)}\right)\partial_\theta. \tag{17}$$

Note that, when one demands the validity of the chain rule, to obtain the expressions in (17) it must be assumed that the derivatives act from the left on the transformed variables. This case will also be explained in detail in the Appendix.

We know that the exterior differential d is defined by

$$\mathsf{d}=\mathsf{d}x'\,\partial_{x'}+\mathsf{d}\theta'\,\partial_{\theta'}. \tag{18a}$$

Substituting (10) and (17) into (18a) one obtains

$$\mathsf{d}=\mathsf{d}x\,\partial_x+\mathsf{d}\theta\,\partial_\theta, \tag{18b}$$

that is, d preserves its form. So, since

$$\mathsf{d}F(x,\theta)=\mathsf{d}x\,\partial_xF+\mathsf{d}\theta\,\partial_\theta F, \tag{19}$$

for any function F, replacing F with xF and θF we get the following relations:

$$\begin{aligned}
\partial_x x&=1+pqx\partial_x+h_1\theta\partial_x+h_2x\partial_\theta+h_1h_2(x\partial_x+\theta\partial_\theta)+(pq-1)\theta\partial_\theta,\\
\partial_x\theta&=p\theta\partial_x-ph_2(x\partial_x+\theta\partial_\theta),\\
\partial_\theta x&=qx\partial_\theta-qh_1(x\partial_x+\theta\partial_\theta),\\
\partial_\theta\theta&=1-\theta\partial_\theta+h_1\theta\partial_x+h_2x\partial_\theta+h_1h_2(x\partial_x+\theta\partial_\theta).
\end{aligned} \tag{20}$$

We now find the commutation rules between derivatives. These rules can be easily obtained by using the nilpotency of the exterior differential. Thus we write

$$0=\mathrm{d}^2=\mathrm{d}x\,\mathrm{d}\theta(p\partial_x\partial_\theta-\partial_\theta\partial_x+h_1\partial_x^2)+(\mathrm{d}\theta)^2(\partial_\theta^2-h_1\partial_x\partial_\theta),$$

which says that

$$\partial_\theta\partial_x=p\partial_x\partial_\theta+h_1\partial_x^2,\quad \partial_\theta^2=h_1\partial_x\partial_\theta. \tag{21}$$

The complete framework of the differential calculus requires commutation relations of the differentials with derivatives.

D. Relations of differentials with derivatives

Finally we shall find the commutation relations between differentials and derivatives. We assume that they have the following form in terms of primed quantities:

$$\begin{aligned}
\partial_{x'}\mathrm{d}x'&=A_{11}\mathrm{d}x'\partial_{x'}+A_{12}\mathrm{d}\theta'\partial_{\theta'},\\
\partial_{x'}\mathrm{d}\theta'&=A_{21}\mathrm{d}\theta'\partial_{x'}+A_{22}\mathrm{d}x'\partial_{\theta'},\\
\partial_{\theta'}\mathrm{d}x'&=B_{11}\mathrm{d}x'\partial_{\theta'}+B_{12}\mathrm{d}\theta'\partial_{x'},\\
\partial_{\theta'}\mathrm{d}\theta'&=B_{21}\mathrm{d}\theta'\partial_{\theta'}+B_{22}\mathrm{d}x'\partial_{x'}.
\end{aligned}\tag{22}$$

Substituting (10) and (17) into (22) and using

$$\mathrm{d}(\mathrm{d}x)=-(\mathrm{d}x)\mathrm{d},\quad \mathrm{d}(\mathrm{d}\theta)=(\mathrm{d}\theta)\mathrm{d}, \tag{23a}$$

and the relation

$$\partial_i(X^j\mathrm{d}X^k)=\delta^i{}_j\delta^k{}_l\mathrm{d}X^k, \tag{23b}$$

where $\partial_1=\partial_x$ and $\partial_2=\partial_\theta$, we determine the coefficients A_{ij} and B_{ij}. So one has

$$\begin{aligned}
\partial_x\mathrm{d}x&=pq\,\mathrm{d}x\partial_x+h_1\mathrm{d}\theta\partial_x-h_2\mathrm{d}x\partial_\theta+h_1h_2(\mathrm{d}x\partial_x+\mathrm{d}\theta\partial_\theta)+(pq-1)\mathrm{d}\theta\partial_\theta,\\
\partial_x\mathrm{d}\theta&=p\,\mathrm{d}\theta\partial_x+ph_2(\mathrm{d}x\partial_x+\mathrm{d}\theta\partial_\theta),\\
\partial_\theta\mathrm{d}x&=-q\,\mathrm{d}x\partial_\theta-qh_1(\mathrm{d}x\partial_x+\mathrm{d}\theta\partial_\theta),\\
\partial_\theta\mathrm{d}\theta&=\mathrm{d}\theta\partial_\theta-h_1\mathrm{d}\theta\partial_x+h_2\mathrm{d}x\partial_\theta+h_1h_2(\mathrm{d}x\partial_x+\mathrm{d}\theta\partial_\theta).
\end{aligned}\tag{24}$$

E. Algebra of one-forms

In this section we shall define two one-forms using the generators of $\mathcal{A}$ and find the commutation relations of one-forms.

If we call them w and u then one can define them as follows:

$$w=\mathrm{d}x\,x^{-1},\quad u=\mathrm{d}\theta\,x^{-1}-\mathrm{d}x\,x^{-1}\theta x^{-1}. \tag{25}$$

We denote the algebra of one-forms generated by two elements w and u by Ω. The generators of the algebra Ω with the generators of $\mathcal{A}$ satisfy the following relations:

$$\begin{aligned}
xw&=wx-h_1ux,\quad \theta w=-w\theta+h_1u\theta,\\
xu&=ux,\quad \theta u=u\theta-h_2(w\theta+ux).
\end{aligned}\tag{26}$$

The commutation rules of the generators of Ω are

$$w^2=0, \quad wu=uw. \tag{27}$$

Using (18b) and (25), if we define the operators T and ∇ as

$$T=x\partial_x+\theta\partial_\theta, \quad \nabla=x\partial_\theta, \tag{28}$$

then we have

$$T\nabla=\nabla T, \quad \nabla^2=0, \tag{29}$$

as a subalgebra of gl(1|1).

The action of T and ∇ on the generators x and θ is

$$\begin{aligned} Tx&=x+xT, \quad \nabla x=x\nabla-h_1xT,\\ T\theta&=\theta+\theta T, \quad \nabla\theta=x-\theta\nabla+h_1\theta T. \end{aligned} \tag{30}$$

III. THE SUPERGROUP $GL_{h_1,h_2}(1|1)$ AND COVARIANCE

It is well known that the quantum supergroup $GL_{p,q}(1|1)$ acts as a linear transformation on the quantum superplane, preserves (1) and the dual relations

$$\varphi'^2=0, \quad \varphi'y'-p^{-1}y'\varphi'=0. \tag{31}$$

In extending this property of covariance under the coaction of $GL_{p,q}(1|1)$, from the superplane to its calculus, it will be assumed that the deformed group structure implies and is implied by invariance of the intermediary relations (5) under linear transformations of the quantum superplane. In the present work, this will be applied to the (h_1,h_2)-deformed superplane.

In this section we would like to discuss the meaning of covariance in a graded version of noncommutative differential calculus of Wess–Zumino[3] for the two-parameter case. Before proceeding, we define the dual quantum h-superplane.

To define the dual quantum superplane, we interpret the differentials $\mathsf{d}x$ and $\mathsf{d}\theta$, as the coordinates of the dual superplane, as follows:

$$\mathsf{d}x=\varphi, \quad \mathsf{d}\theta=y. \tag{32}$$

Now the quantum dual h-superplane generated by y, φ with the relations (12) in the limit $p\rightarrow 1$ will be denoted by $\mathsf{d}\mathcal{A}_{h_1}$. If we assume that $\mathcal{A}_{h_2}$ and $\mathsf{d}\mathcal{A}_{h_1}$ have to be covariant under the coaction,

$$\delta(x)=a\otimes x+\beta\otimes\theta, \quad \delta(\theta)=\gamma\otimes x+d\otimes\theta, \tag{33a}$$

$$\delta(\mathsf{d}Z)=(\tau\otimes\mathsf{d})\,\delta(Z), \quad \tau(u)=(-1)^{\hat{u}}u, \tag{33b}$$

and that β, γ anticommute with θ, φ, h_1, and h_2 we get the corresponding (h_1,h_2)-deformation of the supergroup $GL(1|1)$ as a quantum matrix supergroup $GL_{h_1,h_2}(1|1)$ generated by a, β, γ, d with the relations[10]

$$a\beta=\beta a-h_1(a^2-\beta\gamma-ad),\quad d\beta=\beta d+h_1(d^2+\beta\gamma-da),$$
$$a\gamma=\gamma a+h_2(a^2+\gamma\beta-ad),\quad d\gamma=\gamma d-h_2(d^2-\gamma\beta-da),$$
$$\beta^2=h_1\beta(a-d),\quad \gamma^2=h_2\gamma(d-a), \tag{34}$$
$$\beta\gamma=-\gamma\beta+(h_1\gamma-h_2\beta)(a-d),$$
$$ad=da+h_1(a-d)\gamma+h_2\beta(a-d),$$

where

$$\mathcal{D}=ad^{-1}-\beta d^{-1}\gamma d^{-1}=d^{-1}a-d^{-1}\beta d^{-1}\gamma.$$

The two-parameter differential calculus on the quantum superplane is explicitly as follows: The commutation relations of variables and their differentials are

$$x\theta=\theta x+h_2x^2,\quad \theta^2=-h_2\theta x,$$
$$\varphi y=y\varphi+h_1y^2,\quad \varphi^2=h_1\varphi y. \tag{35}$$

Note that the last two relations of (35) are obtained from (14) and (15). However they can also be obtained from (31) with the limits $p\to 1, q\to 1$.

The commutation relations between variables and derivatives are

$$\partial_x x=1+x\partial_x-h_1\theta\partial_x+h_2x\partial_\theta+h_1h_2(x\partial_x+\theta\partial_\theta),$$
$$\partial_x\theta=\theta\partial_x-h_2(x\partial_x+\theta\partial_\theta),$$
$$\partial_\theta x=x\partial_\theta-h_1(x\partial_x+\theta\partial_\theta), \tag{36}$$
$$\partial_\theta\theta=1-\theta\partial_\theta-h_1\theta\partial_x+h_2x\partial_\theta+h_1h_2(x\partial_x+\theta\partial_\theta),$$

and those among the derivatives are

$$\partial_x\partial_\theta=\partial_\theta\partial_x-h_1\partial_x^2,\quad \partial_\theta^2=h_1\partial_\theta\partial_x. \tag{37}$$

The commutation relations of variables with their differentials are

$$x\varphi=\varphi x+h_1(\varphi\theta-yx)+h_1h_2\varphi x,$$
$$xy=yx-h_1y\theta-h_2\varphi x+h_1h_2\varphi\theta,$$
$$\theta\varphi=-\varphi\theta+h_1y\theta-h_2\varphi x-h_1h_2yx, \tag{38}$$
$$\theta y=y\theta-h_2(\varphi\theta+yx)-h_1h_2y\theta.$$

The commutation relations between derivatives and differentials are

$$\partial_x\varphi=\varphi\partial_x+h_1y\partial_x-h_2\varphi\partial_\theta+h_1h_2(\varphi\partial_x+y\partial_\theta),$$
$$\partial_x y=y\partial_x+h_2(\varphi\partial_x+y\partial_\theta),$$
$$\partial_\theta\varphi=-\varphi\partial_\theta-h_1(\varphi\partial_x+y\partial_\theta), \tag{39}$$
$$\partial_\theta y=y\partial_\theta-h_1y\partial_x+h_2\varphi\partial_\theta+h_1h_2(\varphi\partial_x+y\partial_\theta).$$

Note that this calculus goes back to those of Ref. 9 when $h_1=0$ and $h_2=h$. This calculus is slightly different from Ref. 10. The reason for this difference is the use of commutation relations of the dual *exterior* superplane in Ref. 10 instead of the dual superplane in this work.

We now discuss the covariance of the differential calculus. The covariance here means that all the relations between coordinates x,θ, differentials $\mathrm{d}x,\mathrm{d}\theta$ and derivatives ∂_x, ∂_θ, etc. must preserve their form when one changes the coordinates by

$$x\to ax+\beta\theta, \quad \theta\to\gamma x+d\theta, \tag{40}$$

where the matrix $T=\begin{pmatrix} a & \beta \\ \gamma & d \end{pmatrix}$ is an element of the quantum supergroup $GL_{h_1,h_2}(1|1)$ acting on the quantum superplane. We must change the differentials by

$$\mathrm{d}x\to a\mathrm{d}x-\beta\mathrm{d}\theta, \quad \mathrm{d}\theta\to-\gamma\mathrm{d}x+d\mathrm{d}\theta, \tag{41}$$

since the odd objects anticommute among themselves. Covariance can be maintained if one defines the transformation law of the partial derivatives as follows:

$$\begin{aligned} \partial_x &\to (a^{-1}-a^{-1}\gamma d^{-1}\beta a^{-1})\partial_x - a^{-1}\gamma d^{-1}\partial_\theta, \\ \partial_\theta &\to (d^{-1}-d^{-1}\beta a^{-1}\gamma d^{-1})\partial_\theta + d^{-1}\beta a^{-1}\partial_x. \end{aligned} \tag{42}$$

IV. A TWO-PARAMETER DEFORMATION OF CLASSICAL PHASE SPACE

We shall now give a two-parameter deformation of the $(1+1)$-dimensional classical phase space. We denote the algebra (35)–(37) generated by coordinates x,θ and the derivatives ∂_x and ∂_θ by $\mathcal{B}_{h_1,h_2}$. It is interesting to note that simply identifying ∂_x and ∂_θ with ip_x and p_θ is not compatible with the hermiticity of coordinates and momenta. To identify ∂_x and ∂_θ with the momenta ip_x and p_θ, one must take care of the Hermiticity of the coordinates and momenta. To this end, we first define the Hermitian conjugation of the coordinates x and θ, respectively, as

$$x^+=(1+2h_1h_2)x+2h_1\theta, \quad \theta^+=(1-2h_1h_2)\theta+2h_2x. \tag{43}$$

It is then easy to see that the Hermiticity of x^+ and θ^+ impose some condition on the deformation parameters, i.e., h_1 is a real parameter and h_2 is a pure imaginary parameter,

$$\overline{h_1}=h_1, \quad \overline{h_2}=-h_2, \tag{44}$$

where the bar denotes complex conjugation. In this case, the Hermitian conjugation of the derivatives ∂_x and ∂_θ are

$$\partial_x^+=-(1+2h_1h_2)\partial_x+2h_2\partial_\theta, \quad \partial_\theta^+=(1-2h_1h_2)\partial_\theta+2h_1\partial_x. \tag{45}$$

In the $h_1\to0, h_2\to0$ limits the definitions (43) and (45) go back to those of the classical case.

The relations (35)–(37) are now invariant under the transformations (43) and (45). The above involution allows us to define the Hermitian operators,

$$\hat{x}=(1+h_1h_2)x+h_1\theta, \quad \hat{\theta}=(1-h_1h_2)\theta+h_2x, \tag{46}$$

and, as bosonic and fermionic momenta,

$$\hat{p}_x=i[(1+h_1h_2)\partial_x-h_2\partial_\theta], \quad \hat{p}_\theta=(1-h_1h_2)\partial_\theta+h_1\partial_x. \tag{47}$$

The final form of the (h_1,h_2)-deformed phase space algebra is

$$
\begin{gathered}
\hat{x}\hat{\theta}=\hat{\theta}\hat{x}+h_2\hat{x}^2, \quad \hat{\theta}^2=-h_2\hat{\theta}\hat{x},\\
\hat{p}_x\hat{p}_\theta=\hat{p}_\theta\hat{p}_x+ih_1\hat{p}_x^2, \quad \hat{p}_\theta^2=-ih_1\hat{p}_x\hat{p}_\theta,\\
\hat{p}_x\hat{x}=i+\hat{x}\hat{p}_x+ih_2\hat{x}\hat{p}_\theta-h_1\hat{\theta}\hat{p}_x+h_1h_2(1+\hat{x}\hat{p}_x+i\hat{\theta}\hat{p}_\theta),\\
\hat{p}_x\hat{\theta}=\hat{\theta}\hat{p}_x-h_2(\hat{x}\hat{p}_x+i\hat{\theta}\hat{p}_\theta),\\
\hat{p}_\theta\hat{x}=\hat{x}\hat{p}_\theta+h_1(i\hat{x}\hat{p}_x-\hat{\theta}\hat{p}_\theta),\\
\hat{p}_\theta\hat{\theta}=1-\hat{\theta}\hat{p}_\theta+h_2\hat{x}\hat{p}_\theta+ih_1\hat{\theta}\hat{p}_x-h_1h_2(1+i\hat{x}\hat{p}_x-\hat{\theta}\hat{p}_\theta).
\end{gathered}
\tag{48}
$$

This gives a (h_1,h_2)-deformed phase space algebra which may be used to study the $(1+1)$-dimensional quantum phase space.

Note that we can derive a deformed super-Clifford algebra from the phase space algebra as follows: suppose that we define gamma matrices

$$
\gamma^1\equiv\hat{p}_\theta, \quad \gamma^2\equiv\hat{\theta}, \quad c^1\equiv\hat{p}_x, \quad c^2\equiv\hat{x}. \tag{49}
$$

Then, they satisfy the super-Clifford algebra,

$$
\begin{gathered}
c^1c^2=c^2c^1-h_1\gamma^2c^1+i(1+h_2c^2\gamma^1)+h_1h_2(1+\gamma^2\gamma^1+c^2c^1),\\
c^1\gamma^2=\gamma^2c^1-h_2(c^2c^1+i\gamma^2\gamma^1),\\
\gamma^1c^1=c^1\gamma^1-ih_1(c^1)^2, \quad \gamma^1c^2=c^2\gamma^1-h_1(\gamma^2\gamma^1-ic^2c^1),\\
\gamma^1\gamma^2=1-\gamma^2\gamma^1+ih_1\gamma^2c^1+h_2c^2\gamma^1-h_1h_2(1+c^2c^1-\gamma^2\gamma^1),\\
(\gamma^1)^2=-ih_1c^1\gamma^1, \quad (\gamma^2)^2=-h_2\gamma^2c^2,\\
\gamma^2c^2=c^2\gamma^2-h_2(c^2)^2.
\end{gathered}
\tag{50}
$$

V. A COMMENT ON SUPEROSCILLATORS

We know that introducing one ''bosonic'' and one ''fermionic'' oscillator, A and B, respectively, and making the usual identification

$$
x'\leftrightarrow A^+, \quad \theta'\leftrightarrow B^+, \quad \partial_{x'}\leftrightarrow A, \quad \partial_{\theta'}\leftrightarrow B. \tag{51}
$$

one constructs the quantum superoscillator algebra which is covariant under the quantum supergroup $GL_{p,q}(1|1)$. Under identification (2) and (17) one has

$$
\begin{gathered}
x\leftrightarrow A^+-\frac{h_1}{p-1}B^+, \quad \partial_x\leftrightarrow\left(1+\frac{h_1h_2}{(p-1)(q-1)}\right)A+\frac{h_2}{q-1}B,\\
\theta\leftrightarrow\left(1-\frac{h_1h_2}{(p-1)(q-1)}\right)B^+-\frac{h_2}{q-1}A^+, \quad \partial_\theta\leftrightarrow B-\frac{h_1}{p-1}A,
\end{gathered}
\tag{52}
$$

where

$$
\overline{p}=q. \tag{53}
$$

Substituting (52) into (20) and (3), surprisingly all (h_1,h_2)-dependence cancels and one obtains the usual (p,q)-deformed superoscillator algebra[12]

$$AA^{+}=1+pqA^{+}A+(pq-1)B^{+}B,$$
$$BB^{+}=1-B^{+}B, \quad B^{2}=0=B^{+2},$$
$$AB^{+}=pB^{+}A, \quad AB=p^{-1}BA, \tag{54}$$
$$A^{+}B=q^{-1}BA^{+}, \quad A^{+}B^{+}=qB^{+}A^{+}.$$

In the $p\to 1, q\to 1$ limits, we get an undeformed superoscillator algebra.

APPENDIX: NONVALIDITY OF THE ACT FROM THE RIGHT

In this appendix, we show that a two-parameter covariant differential calculus on the quantum h-superplane can be constructed only if the derivatives and differentials act from the left.

Consider the change of coordinates which is given by (2)

$$x'=\left(1+\frac{h_1h_2}{(p-1)(q-1)}\right)x+\frac{h_1}{p-1}\,\theta, \quad \theta'=\theta+\frac{h_2}{q-1}\,x. \tag{A1}$$

If we interpret the symbols $\mathsf{d}x$ and $\mathsf{d}\theta$ as differentials acting from the right and demanding the validity of the chain rule, we have

$$\mathsf{d}x=\mathsf{d}x'\frac{\partial x}{\partial x'}+\mathsf{d}\theta'\frac{\partial x}{\partial \theta'}=\mathsf{d}x'+\mathsf{d}\theta'\left(-\frac{h_1}{p-1}\right)=\mathsf{d}x'-\frac{h_1}{p-1}\,\mathsf{d}\theta' \tag{A2}$$

and

$$\mathsf{d}\theta=\frac{h_2}{q-1}\,\mathsf{d}x'+\left(1+\frac{h_1h_2}{(p-1)(q-1)}\right)\mathsf{d}\theta'. \tag{A3}$$

Therefore, for example,

$$\mathsf{d}x'=\left(1+\frac{h_1h_2}{(p-1)(q-1)}\right)\mathsf{d}x+\frac{h_1}{p-1}\,\mathsf{d}\theta, \tag{A4}$$

so that

$$\text{RHS of (A4)}=\left(1+\frac{h_1h_2}{(p-1)(q-1)}\right)\mathsf{d}x'-\frac{h_1}{p-1}\,\mathsf{d}\theta'+\frac{h_1}{p-1}\,\mathsf{d}\theta'+\frac{h_1h_2}{(p-1)(q-1)}\,\mathsf{d}x'\neq \mathsf{d}x'.$$

Similarly, if we write, from the chain rule,

$$\partial_x=\left(1+\frac{h_1h_2}{(p-1)(q-1)}\right)\partial_{x'}+\frac{h_2}{q-1}\,\partial_{\theta'}\,, \tag{A5}$$

and

$$\partial_\theta=\frac{h_1}{p-1}\,\partial_{x'}+\partial_{\theta'}\,, \tag{A6}$$

then

$$\partial_{x'}=\partial_x-\frac{h_2}{q-1}\,\partial_\theta, \quad \partial_{\theta'}=-\frac{h_1}{p-1}\,\partial_x+\left(1-\frac{h_1h_2}{(p-1)(q-1)}\right)\partial_\theta, \tag{A7}$$

so that, for example

$$\text{RHS of (A6)}=\left(1-2\,\frac{h_1h_2}{(p-1)(q-1)}\right)\partial_\theta\neq\partial_\theta .$$

This asymmetry between right and left derivative and differential for transformed variables stems from the matrix g in (14) which off-diagonal elements are odd. That is, g is a supermatrix and so the supertranspose must be used.

ACKNOWLEDGMENT

This work was supported in part by T.B.T.A.K. the Turkish Scientific and Technical Research Council.

[1] S. L. Woronowicz, Commun. Math. Phys. **122**, 125 (1989).
[2] A. Connes, *Noncommutative Differential Geometry* (Academic, London, 1994).
[3] J. Wess and B. Zumino, Nucl. Phys. B, Proc. Suppl. **18**, 302 (1990).
[4] Yu I. Manin, Commun. Math. Phys. **123**, 163 (1989).
[5] S. Soni, J. Phys. A **24**, L459 (1991).
[6] W. S. Chung, J. Math. Phys. **35**, 2485 (1994).
[7] V. Karimipour, Lett. Math. Phys. **30**, 87 (1994).
[8] A. Aghamohammadi, Mod. Phys. Lett. A **8**, 2607 (1993).
[9] S. Çelik, S. A. Çelik, and M. Arik, J. Math. Phys. **39**, 3426 (1998).
[10] S. Çelik, J. Math. Phys. **42**, 299 (1997).
[11] L. Dabrowski and P. Parashar, Lett. Math. Phys. **38**, 331 (1996).
[12] S. Vokos, J. Math. Phys. **32**, 2979 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Higher-order BRST and anti-BRST operators and cohomology for compact Lie algebras

C. Chryssomalakos[a] and J. A. de Azcárraga[b]
Departamento de Física Teórica, Universidad de Valencia and IFIC, Centro Mixto Universidad de Valencia–CSIC, E-46100 Burjassot (Valencia), Spain

A. J. Macfarlane[c]
Department of Applied Mathematics and Theoretical Physics, Silver St., Cambridge, CB3 9EW, United Kingdom

J. C. Pérez Bueno[d]
Departamento de Física Teórica, Universidad de Valencia and IFIC, Centro Mixto Universidad de Valencia-CSIC, E-46100 Burjassot (Valencia), Spain

(Received 29 January 1999; accepted for publication 16 April 1999)

After defining cohomologically higher-order BRST and anti-BRST operators for a compact simple algebra $\mathcal{G}$, the associated higher-order Laplacians are introduced and the corresponding supersymmetry algebra Σ is analyzed. These operators act on the states generated by a set of fermionic ghost fields transforming under the adjoint representation. In contrast with the standard case, for which the Laplacian is given by the quadratic Casimir, the higher-order Laplacians W are not, in general, given completely in terms of the Casimir–Racah operators, and may involve the ghost number operator. The higher-order version of the Hodge decomposition is exhibited. The example of $su(3)$ is worked out in detail, including the expression of its higher-order Laplacian W.
[S0022-2488(99)03010-8]

I. INTRODUCTION

BRST symmetry,[1,2] or ''quantum gauge invariance,'' has played an important role in the quantization of non-Abelian gauge theories. The nilpotency of the operator $\mathcal{Q}$ generating the global BRST symmetry implies that the renormalization of gauge theories involves cohomological aspects: the physical content of the theory belongs to the kernel of $\mathcal{Q}$, the physical (BRST-invariant) states being defined by BRST-cocycles modulo BRST-trivial ones (coboundaries). The inclusion of the BRST symmetry in the Batalin–Vilkovisky antibracket–antifield formalism (see Refs. 3, 4 for further references), itself of a rich geometrical structure,[3–9] and where the antifields are the sources of the BRST transformations, has made of BRST quantization the most powerful method for quantizing systems possessing gauge symmetries. In particular, it is indispensable for understanding the general structure of string amplitudes. It is thus interesting to explore its possible generalizations and their cohomological structure.

An essential ingredient of $\mathcal{Q}$ is what we shall denote here the BRST *operator*,

$$s=-\frac{1}{2}C_{ij}{}^{k}c^{i}c^{j}\frac{\partial}{\partial c^{k}} \quad i,j,k=1,...,r=\dim\mathcal{G}, \tag{1.1}$$

where the c^i are anticommuting Grassmann (or *ghost*) variables transforming under the adjoint representation of the (compact semisimple) Lie group G of Lie algebra $\mathcal{G}$. In Yang Mills theories

[a]Electronic mail address: chryss@lie.ific.uv.es
[b]Electronic mail: azcarrag@lie1.ific.uv.es
[c]Electronic mail: a.j.macfarlane@damtp.cam.ac.uk
[d]Electronic mail: pbueno@lie.ific.uv.es

the c's correspond to the ghost fields, and (1.1) above is just part of the generator of the BRST transformations for the gauge group G. This paper is devoted to the generalizations of (1.1) and its associated *anti*-BRST *operator* $\bar{s}$.[10–16]

Using a Euclidean metric to raise and lower indices,[17] $\bar{s}$ is given by

$$\bar{s}=\frac{1}{2}C_{ij}{}^{k}c_{k}\frac{\partial}{\partial c_i}\frac{\partial}{\partial c_j}. \tag{1.2}$$

The $s(\bar{s})$ operator increases (decreases) the ghost number by one. The BRST and anti-BRST operators may be used to construct a Laplacian,[18,12,13,19,14,16] $\Delta=\bar{s}s+s\bar{s}$; clearly, Δ does not change the ghost number. It turns out (see Refs. 12–14) that this operator is given by the (second-order) Casimir operator of $\mathcal{G}$.

A few years ago, van Holten[14] discussed the BRST complex, generated by the s_ρ and $\bar{s}_\rho$ operators,

$$s_\rho=c^i\rho(X_i)-\frac{1}{2}C_{ij}{}^{k}c^ic^j\frac{\partial}{\partial c^k},\quad \bar{s}_\rho=-\rho(X_i)\frac{\partial}{\partial c_i}+\frac{1}{2}C_{ij}{}^{k}c_k\frac{\partial}{\partial c_i}\frac{\partial}{\partial c_j}, \tag{1.3}$$

in connection with the cohomology of compact semisimple Lie algebras. They act on generic states of ghost number q of the form

$$\psi=\frac{1}{q!}\psi^A_{i_1\cdots i_q}c^{i_1}\cdots c^{i_q}\otimes e_A. \tag{1.4}$$

The operators in (1.3) differ from those in (1.1), (1.2) by the inclusion of the ρ term, where $\rho^A{}_B$ is a representation of the Lie algebra $\mathcal{G}$ on a vector space V with basis $\{e_A\}$, $A=1,\ldots,\dim V$. However (see the *Remark* in Sec. II), only the trivial representation case is interesting. For $\rho=0$ the generic states have the form

$$\psi=\frac{1}{q!}\psi_{i_1\cdots i_q}c^{i_1}\cdots c^{i_q}, \tag{1.5}$$

and we shall consider mostly this case. The operators s, $\bar{s}$ and the Laplacian Δ may be used to define s-closed, $\bar{s}$-closed (coclosed), and harmonic states. A state ψ [Eq. (1.5)] is called s-closed, $\bar{s}$-closed, or harmonic if $s\psi=0$, $\bar{s}\psi=0$, or $\Delta\psi=0$, respectively. In this way, and using the nilpotency of s and $\bar{s}$, one may introduce a Hodge decomposition for (1.5) as a sum of an s-closed, an $\bar{s}$-closed, and a harmonic state. The interesting fact is that, using the above Euclidean metric on $\mathcal{G}$, one may introduce a positive scalar product among states ψ', ψ, of ghost numbers q', q by

$$\langle\psi',\psi\rangle:=\frac{1}{q!}\psi'_{j_1\cdots j_q}\psi^{j_1\cdots j_q}\delta_{q'q}. \tag{1.6}$$

Using the Hodge $*$ operator for δ_{ij}, it follows that $s=(-1)^{r(q+1)}*\bar{s}*$ (on states of ghost number q), and that s and $\bar{s}$ are also adjoint to each other with respect to the scalar product (1.6). As a result, there is a complete analogy between the harmonic analysis of forms, in which d and $\delta=(-1)^{r(q+1)+1}*d*$ are adjoint to each other, and the Hodge-like decomposition of states ψ for the operators s, $\bar{s}$[14] (see also Sec. II). This follows from the fact that, due to their anticommuting character, the ghost variables c may be identified[20] with (say) the left-invariant one-forms on the group manifold G, so that the action of s on c determines the Maurer–Cartan (MC) equations.

The nilpotency of the BRST operators (1.1) or (1.3) results from the Jacobi identity satisfied by the structure constants $C_{ij}{}^{k}$ of $\mathcal{G}$. This identity can also be viewed as a three-cocycle condition on the fully antisymmetric C_{ijk}, which define a nontrivial three-cocycle for any semisimple $\mathcal{G}$. This observation indicates the existence of a generalization by using the higher-order cocycles for $\mathcal{G}$. The cohomology ring of all compact simple Lie algebras of rank l (for simplicity, we shall assume G simple henceforth) is generated by l (classes of) nontrivial primitive cocycles, associated with the l invariant, symmetric primitive polynomials of order m_s ($s=1,...,l$) which, in turn, define the l Casimir–Racah operators (Refs. 21–23; also see Ref. 24 and references therein). The different integers m_s depend (for $s\neq 1$) on the specific simple algebra considered. It has been shown in Ref. 25 that, associated to each cocycle of order $2m_s-1$ there exists a higher-order BRST operator s_{2m_s-2} carrying ghost number $2m_s-3$, defined by the coordinates $\Omega^j_{i_1\ldots i_{2m_s-2}}$ of the $(2m_s-1)$-cocycle on $\mathcal{G}$ [we shall also give in (3.17) the corresponding operator $s_{\rho(2m_s-2)}$ for the $\rho\neq 0$ case]. The $\Omega^j_{i_1\ldots i_{2m_s-2}}$ may also be understood as being the (fully antisymmetric) higher-order structure constants of a higher $(2m_s-2)$-order algebra,[25] for which the multibrackets have $(2m_s-2)$ entries. The standard (lowest, $s=1$) case corresponds to $m_1=2$, $\forall\ \mathcal{G}$ (the invariant is Cartan–Killing metric), to the three-cocycle C_{ijk} and to the ordinary Lie algebra bracket. The $(2m_s-2)$-brackets of these higher-order algebras satisfy a generalized Jacobi identity, which again follows from the fact that the higher-order structure constants define $(2m_s-1)$-cocycles for the Lie algebra cohomology. These $(2m_s-2)$-algebras constitute a particular example (in which only one coderivation survives) of the strongly homotopy algebras,[26] which have recently appeared in different physical theories that share common cohomological aspects, as in closed string field theory,[27,28] the higher order generalizations of the antibracket[29,30] and the Batalin–Vilkovisky complex (see Ref. 31 and references therein). Higher-order structure constants satisfying generalized Jacobi identities of the types considered in Ref. 25 [see (3.11) below] and Ref. 32 have also appeared in a natural way in the extended master equation in the presence of higher-order conservation laws.[9]

In Sec. III of this paper we introduce, together with the l general BRST operators for a simple Lie algebra, the corresponding l anti-BRST operators $\overline{s}_{2m_s-2}$ and their associated higher-order Laplacians. We show there that harmonic analysis may be carried out in general (the standard case in Ref. 14 corresponds to $m_s=m_1=2$), although the Laplacians do not, in general, correspond to the Casimir–Racah operators. Nevertheless, we shall show that s_{2m_s-2} and $\overline{s}_{2m_s-2}$ are related to each other by means of the Hodge $*$ operator, and that they are also adjoints of each other. After showing that the different higher-order BRST, anti-BRST, and Laplacian operators generate, for each value $s=1,...,l$, a supersymmetry algebra Σ_{m_s}, we discuss its representations. The example of $\mathcal{G}=su(3)$ is studied in full in Sec. IV, where we construct the general $su(3)$ states and show the $su(3)$ representations contained in the $\Sigma_{m_s}=\Sigma_2,\Sigma_4$ irreducible multiplets.

II. THE STANDARD BRST COMPLEX AND HARMONIC STATES

Let $\mathcal{G}$ be defined by

$$[X_i,X_j]=C_{ij}{}^kX_k,\quad i,j,k=1,..., r\equiv\dim\mathcal{G}, \tag{2.1}$$

where $\{X_i\}_{i=1}^r$ is a basis of $\mathcal{G}$. For instance, we may think of $\{X_i\}$ as a basis for the left-invariant (LI) vector fields $X_i^L(g)\equiv X_i(g)$ on the group manifold $G(X_i(g)\in\mathfrak{X}^L(G))$.

Let V be a vector space. In the Chevalley–Eilenberg formulation (CE)[33] of the Lie algebra cohomology, the space of q-dimensional cochains $C^q(\mathcal{G},V)$ is spanned by the V-valued skew-symmetric mappings,

$$\psi:\mathcal{G}\wedge\stackrel{q}{\cdots}\wedge\mathcal{G}\rightarrow V,\quad \psi(g)=\frac{1}{q!}\psi^A_{i_1\cdots i_q}\omega^{i_1}(g)\wedge\cdots\wedge\omega^{i_q}(g)\otimes e_A, \tag{2.2}$$

where the $\{\omega^i(g)\}$ form a basis of $\mathcal{G}^*$ (LI one-forms on G), dual to the basis $\{X_i\}$ of LI vector fields on G, and the index $A=1,...,\dim V$ labels the components in V. Let ρ be a representation of $\mathcal{G}$ on V ($\rho:\mathcal{G}\rightarrow\mathrm{End}(V)$). The action of the Lie algebra coboundary operator s_ρ, $s_\rho{}^2=0$, on the q-cochains ψ^A (1.4) is given by the following.

Definition 2.1: (Coboundary operator) The coboundary operator s_ρ: $C^q(\mathcal{G},V)\rightarrow C^{q+1}(\mathcal{G},V)$ is defined by

$$(s_\rho\psi)^A(X_1,\ldots,X_{q+1}) := \sum_{i=1}^{q+1} (-1)^{i+1}\rho(X_i)^A_{.B}(\psi^B(X_1,\ldots,\hat{X}_i,\ldots,X_{q+1})) + \sum_{\substack{j,k=1\\ j<k}}^{q+1} (-1)^{j+k}\psi^A([X_j,X_k], \quad X_1,\ldots,\hat{X}_j,\ldots,\hat{X}_k,\ldots,X_{q+1}). \quad (2.3)$$

The space of q-cocycles $Z^q_\rho(\mathcal{G},V)$ (i.e., Ker s) modulo the q-coboundaries $B^q_\rho(\mathcal{G},V)$ (i.e., Im s) defines the qth Lie algebra cohomology group $H^q_\rho(\mathcal{G},V)$.

Since we are assuming $\mathcal{G}$ semisimple, Whitehead's lemma states that, for ρ nontrivial,

$$H^q_\rho(\mathcal{G},V)=0, \quad \forall q \geqslant 0, \quad (2.4)$$

and we can restrict ourselves to $\rho=0$ cohomology for which the action of s_ρ reduces to the second term on the rhs of Eq. (2.3).

For the trivial representation, s acts on ψ (1.5) in the same manner as the exterior derivative d acts on LI forms. It is then clear that we may replace the $\{\omega^i(g)\}$ by the ghost variables $\{c^i\}$,

$$c^ic^j=-c^jc^i \quad \left(\{c^i,c^j\}=0, \quad \left\{c^i,\frac{\partial}{\partial c^j}\right\}=\delta^i_j\right), \quad i,j=1,\ldots,r, \quad (2.5)$$

and the space of q-cochains by polynomials of (ghost) number $q \leqslant r$. The BRST operator (1.1) $s=s_2$ (the subindex 2 is added for convenience; its meaning will become clear in Sec. III) may be taken as the coboundary operator for the $(\rho=0)$ Lie algebra cohomology.[20] Indeed, the relations,

$$s_2c^k=-\tfrac{1}{2}C_{ij}{}^kc^ic^j \quad [\text{or} \;\; s_2c=-\tfrac{1}{2}[c,c], \quad c=c^i\rho(X_i)], \quad (2.6)$$

reproduce the MC equations. As a result, the Lie algebra cohomology may be equivalently formulated in terms of skew-symmetric tensors on $\mathcal{G}$, LI forms on G, or polynomials in ghost space (see, e.g., Ref. 34).

In the sequel we shall introduce the Grassmann variables π_i to refer to the "partial derivative" $\partial/\partial c^i$, appropriate for using the "ghost representation" for the cochains/states. These two sets of variables (c^i,π_j) generate a Clifford-like algebra,[35] defined by

$$\{c_i,\pi_j\}=\delta_{ij}, \quad \{c_i,c_j\}=0=\{\pi_i,\pi_j\}. \quad (2.7)$$

The algebra (2.7) admits the (order reversing) involution $\overline{\cdot}: c_i \mapsto \overline{c_i}=\pi_i$, $\pi_i \mapsto \overline{\pi_i}=c_i$. The *anti*-BRST *operator* $\overline{s_2}$ is given by

$$\overline{\cdot}: s_2 \mapsto \overline{s_2}, \quad \overline{s_2}=\frac{1}{2}C_{ij}{}^kc_k\pi^i\pi^j \quad \left(=\frac{1}{2}C_{ij}{}^kc_k\frac{\partial}{\partial c_i}\frac{\partial}{\partial c_j}\right), \quad (2.8)$$

and it is also nilpotent. Denoting the space of the BRST q-cochains (1.5) by $C^q(\mathcal{G})$, it follows that

$$s_2: C^q(\mathcal{G})\rightarrow C^{q+1}(\mathcal{G}), \quad \overline{s_2}: C^q(\mathcal{G})\rightarrow C^{q-1}(\mathcal{G}). \quad (2.9)$$

The presence of a metric (δ_{ij}) on $\mathcal{G}$ allows us to introduce the $*$ operator $[*: C^q(\mathcal{G}) \rightarrow C^{r-q}(\mathcal{G})]$ in the standard way. On q-forms on G,

$$(*\psi)=\frac{1}{q!}\frac{1}{(r-q)!}\epsilon_{i_1\cdots i_r}\psi^{i_1\cdots i_q}\omega^{i_{q+1}}\wedge\cdots\wedge\omega^{i_r}. \quad (2.10)$$

and

$$*^2=(-1)^{q(r-q)}=(-1)^{q(r-1)}. \tag{2.11}$$

The scalar product of two LI q-forms on G, $\langle\cdot,\cdot\rangle : C^q(\mathcal{G})\otimes C^q(\mathcal{G})\to\mathbb{R}$ is then given by

$$\begin{aligned}\langle\psi',\psi\rangle:=\int_G \psi'\wedge*\psi &= \int_G \frac{1}{q!^2}\frac{1}{(r-q)!}\psi'_{j_1\cdots j_q}\epsilon_{i_1\cdots i_q j_{q+1}\cdots j_r}\psi^{i_1\cdots i_q}\epsilon^{j_1\cdots j_r}\omega^1\wedge\cdots\wedge\omega^r\\ &=\int_G\frac{1}{q!^2}\epsilon^{j_1\cdots j_q}_{i_1\cdots i_q}\psi'_{j_1\cdots j_q}\psi^{i_1\cdots i_q}\omega^1\wedge\cdots\wedge\omega^r\\ &=\frac{1}{q!}\psi'_{j_1\cdots j_q}\psi^{j_1\cdots j_q}\int_G\omega^1\wedge\cdots\wedge\omega^r,\end{aligned} \tag{2.12}$$

and, normalizing the (compact) group volume $\int_G\omega^1\wedge\cdots\wedge\omega^r$ to 1, reduces to (1.6). Clearly,[36]

$$\langle\psi',\psi\rangle=\langle\psi,\psi'\rangle,\quad \langle\psi,\psi\rangle>0\quad \forall\psi\neq 0. \tag{2.13}$$

The codifferential δ is introduced, as usual, as the adjoint of the exterior derivative d, i.e., for a $(q-1)$-form ψ',

$$\begin{aligned}\langle d\psi',\psi\rangle=\int_G d\psi'\wedge*\psi=(-1)^q\int_G\psi'\wedge d*\psi\\ =(-1)^{q+(q-1)(r-q+1)}\int_G\psi'\wedge*(*d*\psi)\\ \equiv\int_G\psi'\wedge*\delta\psi=\langle\psi',\delta\psi\rangle,\end{aligned} \tag{2.14}$$

so that

$$\delta=(-1)^{r(q+1)+1}*d*,\quad (d=(-1)^{r(q+1)}*\delta*),\quad \delta^2=0. \tag{2.15}$$

The correspondence $\omega^i(g)\leftrightarrow c^i$, $d\leftrightarrow s_2$ above allows us to translate all this into the BRST language. First one checks, on any BRST q-cochain (1.5), that the basic operators c^i and π^i are transformed by $*$, according to

$$\pi^i=(-1)^{r(q+1)}*c^i*,\quad c^i=(-1)^{r(q+1)+1}*\pi^i*, \tag{2.16}$$

so that

$$\begin{aligned}&*(c^{i_1}\cdots c^{i_{2k}})*=(-1)^{(r+1)q+k}\pi^{i_1}\cdots\pi^{i_{2k}},\quad *(c^{i_1}\cdots c^{i_{2k+1}})*=(-1)^{r(q+1)+k}\pi^{i_1}\cdots\pi^{i_{2k+1}},\\ &*(\pi^{i_1}\cdots\pi^{i_{2k}})*=(-1)^{(r+1)q+k}c^{i_1}\cdots c^{i_{2k}},\quad *(\pi^{i_1}\cdots\pi^{i_{2k+1}})*=(-1)^{r(q+1)+k+1}c^{i_1}\cdots c^{i_{2k+1}}.\end{aligned} \tag{2.17}$$

As a consequence of (2.16), one finds for $\psi'\in C^{q+1}(\mathcal{G})$, $\psi\in C^q(\mathcal{G})$,

$$\begin{aligned}\langle\psi',c^i\psi\rangle=\int_G\psi'\wedge*c^i\psi=(-1)^{q(r-q)}\int_G\psi'\wedge*c^i**\psi\\ =(-1)^q\int_G\psi'\wedge\pi^i*\psi\\ =\int_G\pi^i\psi'\wedge*\psi=\langle\pi^i\psi',\psi\rangle,\end{aligned} \tag{2.18}$$

using the fact that π^i is a graded derivative and that $\psi' \wedge * \psi \equiv 0$. Thus, c^i and π^i are adjoints to each other with respect to the inner product $\langle , \rangle$ or, in other words, the involution $\overline{\cdot}$ defines the adjoint with respect to $\langle , \rangle$. Thus, $s_2 \sim d$ and (2.15) lead to

$$\overline{s_2} = (-1)^{r(q+1)+1} * s_2 *, \tag{2.19}$$

since

$$\begin{aligned}\delta = (-1)^{r(q+1)+1} * d * \sim (-1)^{r(q+1)+1} * s_2 * \\ &= -(-1)^{r(q+1)+1} \tfrac{1}{2} C_{ij}{}^k * c^i c^j \pi_k * \\ &= -(-1)^{r(q+1)+1+(q-1)(r-q+1)+q(r-q)} \tfrac{1}{2} C_{ij}{}^k * c^i ** c^j ** \pi_k * \\ &= \tfrac{1}{2} C_{ij}{}^k \pi^i \pi^j c_k = \overline{s_2}.\end{aligned} \tag{2.20}$$

The anticommutator of the nilpotent operators s_2 and $\overline{s_2}$ defines the Laplacian $\Delta \equiv W_2$, $W_2 : C^q(\mathcal{G}) \to C^q(\mathcal{G})$,

$$W_2 := \{ s_2 , \overline{s_2} \} = (s_2 + \overline{s_2})^2. \tag{2.21}$$

The operators W_2, s_2, $\overline{s_2}$ generate the supersymmetry algebra Σ_2,

$$[s_2, W_2] = 0, \quad [\overline{s_2}, W_2] = 0, \quad \{ s_2, \overline{s_2} \} = W_2. \tag{2.22}$$

Σ_2 has the structure of a central extension of $(s_2, \overline{s_2})$ by W_2, the Laplacian being the central generator. The operator W_2 is invariant under the involution $\overline{\cdot}(W_2 = \overline{W_2})$ and commutes with $*$, since

$$\begin{aligned} * W_2 * = * (s_2 \overline{s_2} + \overline{s_2} s_2) * = (-1)^{(q-1)(r-q+1)} * (s_2 ** \overline{s_2} + \overline{s_2} ** s_2) * \\ &= (-1)^{(q-1)(r-q+1) + r(q-1) + 1 + rq} (\overline{s_2} s_2 + s_2 \overline{s_2}) \\ &= (-1)^{q(r-q)} W_2, \end{aligned} \tag{2.23}$$

which, with the help of (2.11), implies $[W_2, *] = 0$. Then, as in the standard Hodge theory on compact Riemannian manifolds, we have the following.

Lemma 2.1: A BRST cochain ψ is W_2 harmonic, $W_2 \psi = 0$, iff it is s_2 and $\overline{s_2}$ closed.

Proof: It is clear that if $s_2 \psi = 0 = \overline{s_2} \psi$, then $W_2 \psi = 0$. Now, if $W_2 \psi = 0$,

$$0 = \langle \psi, W_2 \psi \rangle = \langle \psi, (s_2 \overline{s_2} + \overline{s_2} s_2) \psi \rangle = \langle \overline{s_2} \psi, \overline{s_2} \psi \rangle + \langle s_2 \psi, s_2 \psi \rangle; \tag{2.24}$$

from (2.13) easily follows that both terms have to be zero, and hence $s_2 \psi = 0 = \overline{s_2} \psi$.

Theorem 2.1: Each BRST cochain ψ admits the Hodge decomposition,

$$\psi = s_2 \alpha + \overline{s_2} \beta + \gamma, \tag{2.25}$$

where γ is W_2 harmonic (the proof of Theorem 3.1 below includes this case).

To find the algebraic meaning of W_2, let us write the generators X_i on ghost space as

$$X_i \equiv - C_{ij}{}^k c^j \pi_k. \tag{2.26}$$

They act on BRST cochains in the same way as the Lie derivatives with respect to the LI vector fields on G act on LI forms on G:

$$X_i c^k = - C_{ij}{}^k c^j \tag{2.27}$$

[cf. $L_{X_i}\omega^k = -C_{ij}{}^k\omega^j$, in which $X_i \in \mathfrak{X}^L(G)$ and $\omega \in \mathfrak{X}^{*L}(G)$]. The X_i in (2.26) are in the adjoint representation of $\mathcal{G}$ and satisfy $\overline{X_i} = -X_i$ and $*X_i = X_i*$. *Invariant states* are those for which $X_i\psi = 0$, $i = 1,...,r$.

In terms of X_i, the operators s_2 and $\overline{s_2}$ may be written as

$$s_2 = \tfrac{1}{2}c^i X_i, \quad \overline{s_2} = -\tfrac{1}{2}\pi^j X_j, \tag{2.28}$$

Using the fact that c^i and π^j transform in the adjoint representation,

$$X_k c^i = -C_{kr}{}^i c^r, \quad X_k \pi^i = -C_{kr}{}^i \pi^r, \tag{2.29}$$

it is easy to see that[38]

$$W_2 = -\tfrac{1}{2}\mathcal{C}^{(2)} = -\tfrac{1}{2}\delta^{ij}X_i X_j, \tag{2.30}$$

i.e., the Laplace-type operator is proportional to the second-order Casimir operator of the algebra.

Remark: The expression for W_2 in Refs. 12–14 contains more terms due to the fact that these authors consider $\rho \neq 0$, in general. But, as noticed in Ref. 14, $\rho = 0$ is the only possibility if we restrict ourselves to *nontrivial* harmonic states. In fact, we prove here that this is a direct consequence of Whitehead's lemma (2.4). Let τ be the operator defined by its action on (V-valued) q-cochains ψ through

$$(\tau\psi)^A_{i_1 \cdots i_{q-1}} = k^{ij}\rho(X_i)^A_B \psi^B_{j i_1 \cdots i_{q-1}}. \tag{2.31}$$

It may be verified that

$$[(s_\rho\tau + \tau s_\rho)\psi]^A_{i_1 \cdots i_q} = \psi^B_{i_1 \cdots i_q}\mathcal{C}^{(2)}(\rho)^A_B, \tag{2.32}$$

where $\mathcal{C}^{(2)}(\rho)^A_B \equiv k^{ij}\rho(X_i)^A_C\rho(X_j)^C_B$ is the Casimir operator for the representation ρ, and hence proportional to δ^A_B. It then follows that for any $\rho \neq 0$ q-cocycle ψ $(s_\rho\psi = 0)$,

$$s_\rho(\tau\psi\mathcal{C}^{(2)}(\rho)^{-1}) \propto \psi, \tag{2.33}$$

i.e., ψ is a (trivially harmonic state) coboundary generated by a $(q-1)$-cochain proportional to $\tau\psi\mathcal{C}^{(2)}(\rho)^{-1}$, q.e.d. Hence, any nontrivial BRST-invariant state ψ $(s_2\psi = 0,\ \psi \neq s_2\varphi)$ is a $\mathcal{G}$ singlet and, as a consequence of Theorem 2.1, its class contains a unique W_2 harmonic representative.

From (2.30) we also deduce the following.

Lemma 2.2: A state ψ is W_2 harmonic iff it is invariant,[39] $X_i\psi = 0$.

Proof: If ψ is invariant, $W_2\psi = -\tfrac{1}{2}\delta^{ij}X_i X_j\psi = 0$. If ψ is W_2 harmonic,

$$0 = \langle\psi, W_2\psi\rangle = -\tfrac{1}{2}\langle\psi, \delta^{ij}X_i X_j\psi\rangle = \tfrac{1}{2}\delta^{ij}\langle X_i\psi, X_j\psi\rangle, \tag{2.34}$$

and $X_j\psi = 0$, since $\langle\,,\,\rangle$ is nondegenerate, q.e.d. In fact, if ψ is invariant, ψ is both s_2 and $\overline{s_2}$ closed by (2.28).

Corollary 2.1: Each nontrivial element in the cohomology ring $H^*(\mathcal{G})$ may be represented by an invariant state.

Proof: Let $\psi \in Z(\mathcal{G})$ be nontrivial. Hence, its decomposition has the form

$$\psi = s_2\alpha + \gamma. \tag{2.35}$$

Therefore $\psi - s_2\alpha$ is in the cohomology class of ψ and is harmonic (and hence invariant).

III. HIGHER-ORDER BRST AND ANTI-BRST OPERATORS

A. Invariant tensors

The considerations of the previous section rely on the nilpotent operator s_2 and its adjoint, both constructed out of the structure constants C_{ijk}. The latter determine a skew-symmetric tensor of order three, which can be seen as a third-order cocycle $C=C_{ijk}c^ic^jc^k$ and, additionally, is invariant under the action of the Lie algebra generators X_k. Indeed, acting on C with the X's, one gets a sum of three terms, in each of which one of the indices of C is transformed in the adjoint representation and the statement of invariance is equivalent to the Jacobi identity. Notice that we need not saturate every index of C_{ijk} with the same type of variable in order to get an invariant quantity—it suffices that each type of variable transforms in the adjoint representation [for example, s_2 in (1.1), which is also invariant, involves saturating two c's and one π].

The cohomology of simple Lie algebras contains, besides the three cocycle C above, other, higher-order skew-symmetric tensors with similar properties. As mentioned in the Introduction, any compact simple Lie algebra $\mathcal{G}$ of rank l has l primitive cocycles given by skew-symmetric tensors $\Omega^{(2m_s-1)}_{i_1 i_2 \cdots i_{2m_s-1}}$ $(s=1,...,l)$, associated to the l Casimir–Racah primitive invariants of rank m_s.[21–23,25] Their invariance is expressed by the equation

$$\sum_{j=1}^{2m_s-1} C_{bi_j}{}^{a}\Omega^{(2m_s-1)}_{i_1 i_2 \cdots \hat{i}_j a i_{j+1} \cdots i_{2m_s-1}}=0. \tag{3.1}$$

Due to the MC equations, the above relation implies that $\Omega^{(2m_s-1)}_{i_1 \cdots i_{2m_s-1}}\omega^{i_1}\cdots\omega^{i_{2m_s-1}}$ is a cocycle for the Lie algebra coboundary operator (2.3) for $\rho=0$ (in the language of forms, this is equivalent to saying that any bi-invariant form is closed, and hence a CE cocycle). The existence of these cocycles is related to the topology of the corresponding group manifold, in particular, to the odd-sphere product structure that the simple compact group manifolds have from the point of view of real homology (see, e.g., Refs. 33, 40–42, 34). We may use the correspondence $c^i \leftrightarrow \omega^i$ and the discussion after Theorem 2.1 to move freely from the CE approach to the BRST one here.

Let us consider for definiteness the case of $su(n)$, for which $m_1=2$, $m_2=3$,..., $m_l=n$, and there exist $l=(n-1)$ different primitive skew-symmetric tensors of rank 3,5,...,$2n-1$. Consider, for a given m, the $(2m-1)$-form,

$$\Omega^{(2m-1)}=\frac{1}{(2m-1)!}\mathrm{Tr}(\theta\wedge \overset{2m-1}{\cdots} \wedge\theta), \tag{3.2}$$

where $\theta\equiv\omega^iT_i$ and $T_i\in\mathcal{G}$ is in the defining representation of $su(n)$. Since $d\Omega^{(2m-1)}=0$, the coordinates of $\Omega^{(2m-1)}$ provide a $(2m-1)$-cocycle on $su(n)$. One can show (see, e.g., Ref. 24) that

$$\Omega^{(2m-1)}_{\rho i_2 \cdots i_{2m-2}\sigma}=\frac{1}{(2m-3)!}k_{\rho l_1 \cdots l_{m-1}}C_{j_2 j_3}{}^{l_1}\cdots C_{j_{2m-2}\sigma}{}^{l_{m-1}}\epsilon^{j_2\cdots j_{2m-2}}_{i_2\cdots i_{2m-2}}, \tag{3.3}$$

is a skew-symmetric tensor, where

$$k_{\rho l_1 \cdots l_{m-1}}=s\,\mathrm{Tr}(T_\rho T_{l_1}\cdots T_{l_{m-1}}) \tag{3.4}$$

is a symmetric invariant tensor given by the symmetrized trace of a product of m generators [its invariance can be expressed by an equation similar to (3.1)]. Symmetric invariant tensors $k_{i_1\cdots i_m}$ give rise to Casimir–Racah operators,

$$\mathcal{C}^{(m)} = k^{i_1 \cdots i_m} X_{i_1} \cdots X_{i_m}, \tag{3.5}$$

which commute with the generators; $\mathcal{C}^{(2)}$ is the standard quadratic Casimir operator.

B. Higher-order operators

The above family of cocycles Ω can be used to construct *higher-order* BRST *operators*.[25,32] To each invariant tensor of rank m_s corresponds a BRST operator s_{2m_s-2}, which in terms of the coordinates $\Omega_{i_1 \cdots i_{2m_s-2}}{}^{\sigma}$ of the $(2m_s-1)$-cocycle (3.3), is given by

$$s_{2m_s-2} = -\frac{1}{(2m_s-2)!} \Omega^{(2m_s-1)}_{i_1 \cdots i_{2m_s-2}}{}^{\sigma} c^{i_1} c^{i_2} \cdots c^{i_{2m_s-2}} \pi_\sigma . \tag{3.6}$$

These operators are particularly interesting in view of the property

$$\{s_{2m_s-2}, s_{2m_{s'}-2}\} = 0, \quad s, s' = 1,...,l; \tag{3.7}$$

i.e., they are nilpotent and anticommute (see Ref. 25 for a proof).

For each m_s, $s > 1$, we may look at s_{2m-2} (we shall often write m for m_s henceforth) as a *higher-order coboundary* operator, s_{2m-2}: $C^q(\mathcal{G}) \to C^{q+(2m-3)}(\mathcal{G})$. The analog of the MC equation (2.6) for s_{2m-2} is given by

$$s_{2m-2} c^a = -\frac{1}{(2m-2)!} \Omega^{(2m-1)}_{i_1 \cdots i_{2m-2}}{}^{a} c^{i_1} c^{i_2} \cdots c^{i_{2m-2}}, \tag{3.8}$$

which may also be written as

$$s_{2m-2} c = -\frac{1}{(2m-2)!} [c, \overset{2m-2}{\cdots}, c], \tag{3.9}$$

where $[c, \overset{2m-2}{\cdots}, c] := c^{i_1} \cdots c^{i_{2m_2}} [T_{i_1}, \ldots, T_{i_{2m-2}}]$ and the higher-order structure constants of the $(2m-2)$ bracket[25] are given by the $(2m-1)$ cocycle, i.e.,

$$[T_{i_1}, \ldots, T_{i_{2m-2}}] = \Omega^{(2m-1)}_{i_1 \cdots i_{2m-2}}{}^{a} T_a . \tag{3.10}$$

Using (3.9), the nilpotency of s_{2m-2} follows from the higher-order Jacobi identity,

$$s^2_{2m_2} c = -\frac{1}{(2m-2)!} \frac{1}{(2m-3)!} [c, \overset{2m-3}{\cdots}, c, [c, \overset{2m-2}{\cdots}, c]] = 0, \tag{3.11}$$

which the rhs of (3.11) satisfies as a consequence of $\Omega^{(2m-1)}_{i_1 \cdots i_{2m-2}}{}^{a}$ being a cocycle.

Moreover, for each $\mathcal{G}$ we may introduce the *complete BRST operator* **s**,[25]

$$\begin{aligned} \mathbf{s} = & -\frac{1}{2} C_{j_1 j_2}{}^{\sigma} c^{j_1} c^{j_2} \pi_\sigma - \cdots - \frac{1}{(2m_s-2)!} \Omega^{(2m_s-1)}_{j_1 \cdots j_{2m_s-2}}{}^{\sigma} c^{j_1} \cdots c^{j_{2m_s-2}} \pi_\sigma - \cdots \\ & - \frac{1}{(2m_l-2)!} \Omega^{(2m_l-1)}_{j_1 \cdots j_{2m_l-2}}{}^{\sigma} c^{j_1} \cdots c^{j_{2m_l-2}} \pi_\sigma \\ \equiv & \, s_2 + \cdots + s_{2m_s-2} + \cdots + s_{2m_l-2} . \end{aligned} \tag{3.12}$$

This operator is nilpotent, and its terms have (except for some additional ones that break the generalized Jacobi identities, which are at the core of the nilpotency of s_{2m_s-2}), the same structure as those that appear in closed string theory[27] and lead to a strongly homotopy algebra.[26] In fact, the higher-order structure constants (which here have definite values and a geometrical meaning as higher-order cocycles of $\mathcal{G}$) correspond to the string correlation functions giving the string couplings. Since the expression for **s** in the homotopy Lie algebra that underlies closed string theory already includes a term of the form $f^{\sigma}_{j_1} c^{j_1} \pi_\sigma$, f nilpotent, $\mathbf{s}^2=0$ is not satisfied [as it is for (3.12)] by means of a sum of independently satisfied Jacobi identities, and, in particular, the $C_{ij}{}^k$ do not satisfy the Jacobi identity and hence do not define a Lie algebra.

For each s_{2m-2} we now introduce its adjoint *anti*-BRST *operator* $\overline{s}_{2m-2}$,

$$\overline{s}_{2m-2} = -\frac{1}{(2m-2)!} \Omega^{(2m-1)}_{i_1 \cdots i_{2m-2}}{}^{\sigma} c_\sigma \pi^{i_{2m-2}} \cdots \pi^{i_1}$$
$$= -\frac{(-1)^{m-1}}{(2m-2)!} \Omega^{(2m-1)}_{i_1 \cdots i_{2m-2}}{}^{\sigma} c_\sigma \pi^{i_1} \cdots \pi^{i_{2m-2}}. \tag{3.13}$$

Each pair $(s_{2m-2}, \overline{s}_{2m-2})$ allows us to construct a *higher-order* Laplacian W_{2m-2},

$$W_{2m-2} = (s_{2m-2} + \overline{s}_{2m-2})^2 = s_{2m-2}\overline{s}_{2m-2} + \overline{s}_{2m-2} s_{2m-2}. \tag{3.14}$$

Clearly, s_{2m-2}, $\overline{s}_{2m-2}$, and W_{2m-2} all commute with the generators X_i, and we have the following.

Lemma 3.1: For each $s=1,...,l$, the higher-order BRST and anti-BRST operators s_{2m_s-2} and $\overline{s}_{2m_s-2}$, together with their associated Laplacian W_{2m_s-2} define the superalgebra Σ_{m_s},

$$[s_{2m_s-2}, W_{2m_s-2}] = 0, \quad [\overline{s}_{2m_s-2}, W_{2m_s-2}] = 0, \quad \{s_{2m_s-2}, \overline{s}_{2m_s-2}\} = W_{2m_s-2}, \tag{3.15}$$

which has a central extension structure.

For $s=1$, $m_1=2$, $W_2=\Delta$, and the above expressions reproduce (1.1), (2.6), (2.8), and (2.22).

The BRST (anti-BRST) operator s_{2m-2} ($\overline{s}_{2m-2}$), acting on a monomial in the c's, raises (lowers) its ghost number by $2m-3$ while W_{2m-2} leaves the ghost number invariant and is self-adjoint. We notice that all terms in s_{2m-2} ($\overline{s}_{2m-2}$) contain one $(2m-2)\pi$, and that the term with the maximum number of π's in W_{2m-2} contains (at most) $2m-2$ of them. This is so because the two terms with $2m-1$ π's (from $s_{2m-2}\overline{s}_{2m-2}$, $\overline{s}_{2m-2}s_{2m-2}$) cancel, as one can verify. The BRST operator s_{2m-2} annihilates all states of ghost number $q > r-2m+2$, as well as zeroth-order states. Similarly, $\overline{s}_{2m-2}$ annihilates states of ghost number $q < 2m-2$ and the top state $c_1 \cdots c_r$. It follows that zero- and r-ghost number states are both W_{2m-2} harmonic.

Let us establish now the relation between the $\overline{\cdot}$ operation [the adjoint with respect to the inner product in (2.12)] and the conjugation by the Hodge $*$ operator, as these apply to s_{2m-2}.

Lemma 3.2: The following equalities hold on any state (BRST cochain) of ghost number q,

$$\overline{s}_{2m-2} = (-1)^{r(q+1)+1} * s_{2m-2} *, \quad s_{2m-2} = (-1)^{r(q+1)} * \overline{s}_{2m-2} *, \quad s_{2m-2} * = (-1)^q * \overline{s}_{2m-2}; \tag{3.16}$$

notice that the sign factors do not depend on m and hence they coincide with those of (2.15). The proof is straightforward, using (2.11), (2.17), where care should be taken to substitute the ghost numbers actually ''seen'' by the operators.

Although we shall not use them here we also introduce, for the sake of completeness, the V-valued higher-order coboundary operators $s_{\rho_{(2m_s-2)}}$ for a nontrivial representation $\rho \in \mathrm{End}\, V$ of the $(2m_s-2)$-algebra. They are given by

$$s_{\rho(2m_s-2)} = c^{i_1} \cdots c^{i_{2m_s-3}} \rho(X_{i_1}) \cdots \rho(X_{i_{2m_s-3}}) - \frac{1}{(2m_s-2)!} \Omega^{(2m_s-1)}_{i_1 \cdots i_{2m_s-2}}{}^{\sigma} c^{i_1} c^{i_2} \cdots c^{i_{2m_s-2}} \pi_\sigma. \tag{3.17}$$

It may be seen that the nilpotency of $s_{\rho_{(2m_s-2)}}$ is guaranteed by the fact that the skew-symmetric product of $(2m_s-2)$ ρ's, which defines the multibracket $(2m_s-2)$-algebra for an appropriate ρ [Eq. (3.10) with $T\rightarrow\rho$], satisfies the corresponding generalized Jacobi identity as before.

C. Higher-order Hodge decomposition and representations of Σ_{m_s}

Let us now look at the irreducible representations of (3.15), which have the same structure as the supersymmetry algebra. Since W_{2m-2} commutes with $s_{2m-2}, \overline{s}_{2m-2}$, each multiplet of states will have a fixed W_{2m-2} eigenvalue. Let us call γ a W_{2m-2}-*harmonic state* iff $W_{2m-2}\gamma=0$. Then Lemma 2.1 transcribes trivially to the present higher-order case so that γ is harmonic iff it is s_{2m-2} and $\overline{s}_{2m-2}$-closed. Hence, a harmonic state γ is a singlet of Σ_m. We may also extend Theorem 2.1 and prove the following.

Theorem 3.1: (*Higher-order Hodge decomposition*) Each BRST cochain ψ admits a unique decomposition,

$$\psi=s_{2m-2}\alpha+\overline{s}_{2m-2}\beta+\gamma, \tag{3.18}$$

where γ is W_{2m-2}-harmonic.

Proof: We denote by $\mathcal{S}$ the space of all states (i.e., skew-symmetric polynomials in the c's), $\mathcal{K}_{2m-2}$ the kernel of W_{2m-2} (W_{2m-2}-harmonic space) and $\mathcal{K}^{\perp}_{2m-2}$ the complement of $\mathcal{K}_{2m-2}$ in $\mathcal{S}$. Let $P_W^{(0)}$ be the projector from $\mathcal{S}$ to $\mathcal{K}_{2m-2}$. Let $\psi\in\mathcal{S}$; then, $(1-P_W^{(0)})\psi$ lies in $\mathcal{K}^{\perp}_{2m-2}$. However, since the restriction of W_{2m-2} to $\mathcal{K}^{\perp}_{2m-2}$ is invertible, there exists a unique ϕ in $\mathcal{K}^{\perp}_{2m-2}$ such that $(1-P_W^{(0)})\psi=W_{2m-2}\phi$, from which we get

$$\psi=W_{2m-2}\phi+P_W^{(0)}\psi=s_{2m-2}(\overline{s}_{2m-2}\phi)+\overline{s}_{2m-2}(s_{2m-2}\phi)+P_W^{(0)}\psi, \tag{3.19}$$

which provides the desired decomposition of ψ with $\alpha=\overline{s}_{2m-2}\phi$, $\beta=s_{2m-2}\phi$, and $\gamma=P_W^{(0)}\psi$, q.e.d.

To complete the analysis of the irreducible representations of Σ, consider now an eigenstate χ of W_{2m-2} for nonzero (and hence positive) eigenvalue w, $W_{2m-2}\chi=w\chi$, $w>0$. This gives rise to the states

$$\phi\equiv s_{2m-2}\chi,\quad \psi\equiv\overline{s}_{2m-2}\chi,\quad \sigma\equiv s_{2m-2}\overline{s}_{2m-2}\chi. \tag{3.20}$$

Further applications of s_{2m-2} or $\overline{s}_{2m-2}$ produce linear combinations of the above states, for example, $\overline{s}_{2m-2}s_{2m-2}\chi=W_{2m-2}\chi-s_{2m-2}\overline{s}_{2m-2}\chi=w\chi-\sigma$, etc. The quartet $\{\chi,\phi,\psi,\sigma\}$ collapses to a doublet if either $s_{2m-2}\chi=0$ or $\overline{s}_{2m-2}\chi=0$. In this case, χ is the Clifford vacuum and s_{2m-2}, or $\overline{s}_{2m-2}$, respectively, plays the role of the annihilation operator. Let χ be neither s_{2m-2} nor $\overline{s}_{2m-2}$-closed. The state σ of (3.20) is, by construction, s_{2m-2}-closed. Then, we can always choose a linear combination of χ and σ that is $\overline{s}_{2m-2}$ closed. Indeed, for the $\{\chi,\phi,\psi,\sigma\}$ of (3.20), we easily compute

$$\|\phi\|^2+\|\psi\|^2=\langle s_{2m-2}\chi,s_{2m-2}\chi\rangle+\langle\overline{s}_{2m-2}\chi,\overline{s}_{2m-2}\chi\rangle=\langle\chi,W_{2m-2}\chi\rangle=w, \tag{3.21}$$

where we have taken $\|\chi\|^2\equiv\langle\chi,\chi\rangle=1$. Setting

$$q=\sqrt{w},\quad \|\phi\|=q\sin\theta,\quad \|\psi\|=q\cos\theta, \tag{3.22}$$

we find that the following linear combinations:

$$\chi'=\frac{1}{q\sin\theta}(q\chi-q^{-1}\sigma),\quad \sigma'=\frac{1}{q^2\cos\theta}\sigma,\quad \phi'=\frac{1}{q\sin\theta}\phi,\quad \psi'=\frac{1}{q\cos\theta}\psi, \tag{3.23}$$

form an orthonormal set, with the doublet $\{\chi',\phi'\}$ satisfying

$$\overline{s}_{2m-2}\chi'=0, \quad s_{2m-2}\chi'=q\phi'; \quad s_{2m-2}\phi'=0, \quad \overline{s}_{2m-2}\phi'=q\chi', \tag{3.24}$$

and similarly for $\{\psi',\sigma'\}$, i.e., the two doublets decouple. Notice that θ in (3.22), $0<\theta<\pi/2$, is the angle between χ and σ in the $\chi-\sigma$ plane:

$$\langle\chi,\sigma\rangle=q^2\cos^2\theta=\|\chi\|\cdot\|\sigma\|\cos\theta. \tag{3.25}$$

Once in the primed basis of (3.23), $s_{2m-2}\overline{s}_{2m-2}\xi$ (where ξ stands for any of the four primed states above) is equal to either $w\xi$ or 0, and hence, $s_{2m-2}\overline{s}_{2m-2}$ commutes with all operators that commute with W_{2m-2} (similarly for $\overline{s}_{2m-2}s_{2m-2}$). Thus, the representation of the different superalgebras Σ_m fall, in all cases, into singlets (harmonic states) and pairs of doublets. Singlets and doublets here are the trivial analogs of the ''short'' (massless) and ''long'' (massive) multiplets of the standard supersymmetry algebra.

Owing to the particular importance of harmonicity, we investigate the relation between the kernel of a higher-order Laplacian and that of $W_2\propto\mathcal{C}^{(2)}$. To this end, we rewrite s_{2m-2} as (Greek indices below also range in $1,...,r\equiv\dim\mathcal{G}$),

$$\begin{aligned}
s_{2m-2}&=-\frac{1}{(2m-2)!}\Omega^{(2m-1)}_{i_1\cdots i_{2m-2}}{}^{\sigma}c^{i_1}\cdots c^{i_{2m-2}}\pi_\sigma\\
&=-\frac{1}{(2m-2)!}k_{j_1\cdots j_{m-1}}{}^{\sigma}C_{\rho i_2}{}^{j_1}\cdots C_{i_{2m-3}i_{2m-2}}{}^{j_{m-1}}c^{\rho}c^{i_2}\cdots c^{i_{2m-2}}\pi_\sigma\\
&=\frac{1}{(2m-2)!}\left(\sum_{r=2}^{m-1}k_{i_2j_2\cdots\hat{j}_r\alpha\cdots j_{m-1}}{}^{\sigma}C_{\rho j_r}{}^{\alpha}C_{i_3i_4}{}^{j_2}\cdots C_{i_{2r-1}i_{2r}}{}^{j_r}\cdots C_{i_{2m-3}i_{2m-2}}{}^{j_{m-1}}\right.\\
&\qquad\left.+k_{i_2j_2\cdots j_{m-1}}{}^{\alpha}C_{\alpha\rho}{}^{\sigma}C_{i_3i_4}{}^{j_2}\cdots C_{i_{2m-3}i_{2m-2}}{}^{j_{m-1}}\right)c^{\rho}c^{i_2}\cdots c^{i_{2m-2}}\pi_\sigma\\
&=\frac{1}{(2m-2)!}k_{\beta j_1\cdots j_{m-2}}{}^{\alpha}C_{i_1i_2}{}^{j_1}\cdots C_{i_{2m-5}i_{2m-4}}{}^{j_{m-2}}c^{\beta}c^{i_1}\cdots c^{i_{2m-4}}X_\alpha,
\end{aligned}\tag{3.26}$$

where the invariance of $k_{j_1\cdots j_{m-1}}{}^{\sigma}$ has been used in the third line and the Jacobi identity in the last equality. Similarly, $\overline{s}_{2m-2}$ may be written as

$$\overline{s}_{2m-2}=-(-1)^m\frac{1}{(2m-2)!}k_{\beta j_1\cdots j_{m-2}}{}^{\alpha}C_{i_1i_2}{}^{j_1}\cdots C_{i_{2m-5}i_{2m-4}}{}^{j_{m-2}}\pi^{\beta}\pi^{i_1}\cdots\pi^{i_{2m-4}}X_\alpha. \tag{3.27}$$

This proves the following.

Lemma 3.3: The higher-order BRST and anti-BRST operators s_{2m-2}, $\overline{s}_{2m-2}$ may be written as

$$s_{2m-2}=\Omega^{\alpha}X_\alpha, \quad \overline{s}_{2m-2}=-\overline{\Omega}^{\alpha}X_\alpha, \tag{3.28}$$

where

$$\Omega^{\alpha}\equiv\frac{1}{(2m-2)!}k_{\beta j_1\cdots j_{m-2}}{}^{\alpha}C_{i_1i_2}{}^{j_1}\cdots C_{i_{2m-5}i_{2m-4}}{}^{j_{m-2}}c^{\beta}c^{i_1}\cdots c^{i_{2m-4}}, \tag{3.29}$$

$$\overline{\Omega}^{\alpha}\equiv(-1)^m\frac{1}{(2m-2)!}k_{\beta j_1\cdots j_{m-2}}{}^{\alpha}C_{i_1i_2}{}^{j_1}\cdots C_{i_{2m-5}i_{2m-4}}{}^{j_{m-2}}\pi^{\beta}\pi^{i_1}\cdots\pi^{i_{2m-4}}. \tag{3.30}$$

For $m=2$ one gets $\Omega^{\alpha}=\frac{1}{2}c^{\alpha},\overline{\Omega}^{\alpha}=\frac{1}{2}\pi^{\alpha}$ and the expression (2.28) for s_2, $\overline{s_2}$ is recovered. As a W_2-harmonic state is invariant, the above relation shows that the kernel of W_2 is contained in the kernel of W_{2m_s-2} for all m_s, $s=2,...,l$. It follows from (3.28) that an invariant state is s_{2m-2} and $\overline{s}_{2m-2}$-closed, and hence the following lemma.

Lemma 3.4: Every invariant state is W_{2m-2} harmonic.

In the particular realization of the Σ_{m_s} algebra (3.15) in terms of ghosts and antighosts given in (3.6), (3.13), W_{2m_s-2} is ghost number preserving and commutes with the Lie algebra generators X_i. There exists therefore a basis of the c's in which W_{2m_s-2}, $s=1,...,l$ is diagonal. For a fixed ghost number q, the $\binom{r}{q}$ independent monomials $c_{i_1}\cdots c_{i_q}$ transform as the fully antisymmetric part of the qth tensor power of the adjoint representation of $\mathcal{G}$. This antisymmetric part is $\mathcal{G}$-reducible and W_{2m_s-2} will have a fixed eigenvalue in each of its irreducible components (which will change, in general, when going from one irreducible representation to another of the same or different ghost number). s_{2m_s-2} and $\overline{s}_{2m_s-2}$ connect states belonging to the same $\mathcal{G}$-irreducible representation and with the same W_{2m_s-2}-eigenvalue (but of different ghost number), and such states will fall into one of the Σ_{m_s} multiplets (singlets or doublets) discussed above. As the generators X_i commute with $*$, the $\mathcal{G}$-irreducible representation decomposition pattern will be symmetrical under $q\rightarrow r-q$.

The l Casimir–Racah operators $\mathcal{C}^{(m_s)}$, take fixed eigenvalues within each $\mathcal{G}$-irreducible component, which is uniquely labeled by them. The same irreducible representation may appear more than once, with equal or with different ghost numbers; the Casimirs will not distinguish among these different copies of the same irreducible representation. As mentioned, $W_2\equiv\Delta=-\frac{1}{2}\mathcal{C}^{(2)}$ [Eq. (2.30)]. An important question that naturally arises is whether W_{2m_s-2} also reduces, for $s>2$, to some higher-order Casimir–Racah operator or, more generally, to a sum of products of them. Since W_{2m_s-2} commutes with the X's [realized in ghost space via (2.26)] an equivalent question is whether it belongs to the universal enveloping algebra $\mathcal{U}(\mathcal{G})$ of $\mathcal{G}$. The answer is negative and we address this point in the next section, working out in full the $su(3)$ case.

IV. THE CASE OF $su(3)$

We opt here for mild departures from our previous conventions: the generators will now be chosen Hermitian so as to work with real eigenvalues and the normalization of all operators is such that fractional eigenvalues are avoided.

A. Invariant tensors and operators

The $su(3)$ algebra $[T_i,T_j]=if_{ij}{}^kT_k$, $i=1,...,8$, is determined by the nonzero structure constants $f_{ij}{}^k$ that are reproduced for convenience (Table I). These are also the coordinates of the $su(3)$ three-cocycle. The well-known third-order symmetric tensor d_{ijk} (Table II) gives the third-order Casimir–Racah operator. From d_{ijk} and (3.3) one finds the $su(3)$ five-cocycle coordinates[24] (Table III).

The Casimirs C_2 and C_3,

$$C_2=T^iT_i,\quad C_3=d^{ijk}T_iT_jT_k, \tag{4.1}$$

are related to the operators (3.5) simply by

$$C_2=-\mathcal{C}^{(2)},\quad C_3=-i\mathcal{C}^{(3)}. \tag{4.2}$$

TABLE I. Nonzero structure constants for $su(3)$.

$f_{123}=1$	$f_{147}=1/2$	$f_{156}=-1/2$
$f_{246}=1/2$	$f_{257}=1/2$	$f_{345}=1/2$
$f_{367}=-1/2$	$f_{458}=\sqrt{3}/2$	$f_{678}=\sqrt{3}/2$

TABLE II. Nonzero-components of the symmetric invariant d_{ijk} for $su(3)$.

$d_{118}=1/\sqrt{3}$	$d_{228}=1/\sqrt{3}$	$d_{338}=1/\sqrt{3}$	$d_{888}=-1/\sqrt{3}$
$d_{448}=-1/(2\sqrt{3})$	$d_{558}=-1/(2\sqrt{3})$	$d_{668}=-1/(2\sqrt{3})$	$d_{778}=-1/(2\sqrt{3})$
$d_{146}=1/2$	$d_{157}=1/2$	$d_{247}=-1/2$	$d_{256}=1/2$
$d_{344}=1/2$	$d_{355}=1/2$	$d_{366}=-1/2$	$d_{377}=-1/2$

The antisymmetric cocycles, on the other hand, give rise to the BRST and anti-BRST operators s_2, $\overline{s_2}$, s_4, $\overline{s_4}$ [see also (3.13)],

$$s_2=-\frac{1}{2}f_{ij}{}^{k}c^ic^j\pi_k, \quad s_4=-\frac{1}{4!}\Omega_{i_1i_2i_3i_4}{}^{\sigma}c^{i_1}c^{i_2}c^{i_3}c^{i_4}\pi_\sigma; \tag{4.3}$$

$$s_2{}^2=s_4{}^2=0=\overline{s_2}^2=\overline{s_4}^2, \quad s_2s_4+s_4s_2=0=\overline{s_2}\,\overline{s_4}+\overline{s_4}\,\overline{s_2}. \tag{4.4}$$

The corresponding Laplacians are

$$W_2=(s_2+\overline{s_2})^2=s_2\overline{s_2}+\overline{s_2}s_2, \quad W_4=(s_4+\overline{s_4})^2=s_4\overline{s_4}+\overline{s_4}s_4, \tag{4.5}$$

and satisfy, in addition to (2.22),

$$[W_2,(s_4,\overline{s_4},W_4)]=0, \quad [W_4,s_4]=0=[W_4,\overline{s_4}]. \tag{4.6}$$

Notice that W_4 does not commute with s_2 or $\overline{s_2}$ (but, being invariant, it does commute with W_2). As W_2 is proportional to C_2, we only refer to the latter in the sequel. Also, to avoid fractional eigenvalues, we define $W\equiv 4!W_4$.

B. Decomposition into irreducible representations

In general, a monomial in the c's of ghost number q transforms in $\mathbf{8}^{\wedge q}$, the part of the qth tensor power of $su(3)$ adjoint representation that is totally antisymmetric in the q factors. The reduction of $\mathbf{8}^{\wedge q}$ into irreducible representations of $su(3)$ can be achieved by a variety of methods. One way, which gives results useful in our analysis below, employs conventional tensor methods. We first quote the results:

$$\begin{aligned}
\mathbf{8}^{\wedge 0}&=\mathbf{1}=\mathbf{8}^{\wedge 8}\\
\mathbf{8}^{\wedge 1}&=\mathbf{8}=\mathbf{8}^{\wedge 7}\\
\mathbf{8}^{\wedge 2}&=\mathbf{8}+\mathbf{10}+\overline{\mathbf{10}}=\mathbf{8}^{\wedge 6}\\
\mathbf{8}^{\wedge 3}&=\mathbf{1}+\mathbf{8}+\mathbf{10}+\overline{\mathbf{10}}+\mathbf{27}=\mathbf{8}^{\wedge 5}\\
\mathbf{8}^{\wedge 4}&=2\times\mathbf{8}+2\times\mathbf{27},
\end{aligned} \tag{4.7}$$

noting the symmetry $\mathbf{8}^{\wedge q}=\mathbf{8}^{\wedge(r-q)}$, and then describe the tensorial method of developing the results in a fully explicit form.

We may refer to $su(3)$ irreducible representations either by dimension, or else in highest weight $\{\lambda_1,\lambda_2\}$ notation. In the latter notation $\{1,0\}$ and $\{0,1\}$ denote the "quark" and "anti-quark" representations $\mathbf{3}$ and $\overline{\mathbf{3}}$ each of dimension 3, and $\{\lambda_1,\lambda_2\}$ denotes the representation whose highest weight is $\mathbf{w}(\lambda_1,\lambda_2)=\lambda_1\mathbf{w}(1,0)+\lambda_2\mathbf{w}(0,1)$, where $\mathbf{w}(1,0)$, $\mathbf{w}(0,1)$ are the weights of $\mathbf{3}$, $\overline{\mathbf{3}}$, respectively. The representation $\{\lambda_1,\lambda_2\}$ has dimension

TABLE III. Nonzero coordinates of the $su(3)$ five-cocycle.

$\Omega_{12345}=1/4$	$\Omega_{12367}=1/4$	$\Omega_{12458}=\sqrt{3}/12$
$\Omega_{12678}=-\sqrt{3}/12$	$\Omega_{13468}=-\sqrt{3}/12$	$\Omega_{13578}=-\sqrt{3}/12$
$\Omega_{23478}=\sqrt{3}/12$	$\Omega_{23568}=-\sqrt{3}/12$	$\Omega_{45678}=-\sqrt{3}/6$

$$d(\lambda_1,\lambda_2)=\tfrac{1}{2}(\lambda_1+1)(\lambda_2+1)(\lambda_1+\lambda_2+2), \tag{4.8}$$

and Casimir operators (4.1) whose eigenvalues are[43]

$$C_2(\lambda_1,\lambda_2)=\tfrac{1}{3}(\lambda_1^2+\lambda_1\lambda_2+\lambda_2^2)+\lambda_1+\lambda_2, \tag{4.9}$$

$$C_3(\lambda_1,\lambda_2)=\tfrac{1}{18}(\lambda_1-\lambda_2)(\lambda_1+2\lambda_2+3)(2\lambda_1+\lambda_2+3)=-C_3(\lambda_2,\lambda_1). \tag{4.10}$$

Since $C_3(\lambda_1,\lambda_2)=-C_3(\lambda_2,\lambda_1)$, C_3 vanishes for all self-conjugate $(\lambda_1=\lambda_2)$ irreducible representations. The results for the representations ρ that occur in (4.7) are given in the table,

$$\begin{array}{ccc} \dim\rho & \{\lambda_1,\lambda_2\} & (C_2,C_3) \\ \mathbf{1} & \{0,0\} & (0,0) \\ \mathbf{8} & \{1,1\} & (3,0) \\ \mathbf{10} & \{3,0\} & (6,9) \\ \overline{\mathbf{10}} & \{0,3\} & (6,-9) \\ \mathbf{27} & \{2,2\} & (8,0) \end{array}\;. \tag{4.11}$$

Turning to the tensor analysis of tensors spanned, for $0\leqslant q\leqslant 8$, by the monomials $c_{i_1}\cdots c_{i_q}$, we start with the case $q=1$, where c_i describes the basis of the $su(3)$ adjoint representation, i.e., an octet. In the case $q=2$,

$$d_i=f_{ijk}c_jc_k \tag{4.12}$$

describes an independent octet, the only one available since $d_{ijk}c_jc_k\equiv 0$. The remaining tensor, irreducible over the field $\mathbb{R}$ is

$$c_ic_j-\tfrac{1}{3}f_{ijk}d_k=(\mathbf{20}_2)_{ij}, \tag{4.13}$$

for which $f_{ijk}(\mathbf{20}_2)_{ij}=0$ by construction. The notation implies that it has 20 components, agreeing with the simple count $\binom{8}{2}-8$. To reduce it into separate $\mathbf{10}$ and $\overline{\mathbf{10}}$ pieces can be done only over the field of complex numbers, but this is not needed here.[44] We may also write

$$c_ic_j=(c_ic_j-\tfrac{1}{3}f_{ijk}d_k)+\tfrac{1}{3}f_{ijk}d_k=(P_{20})_{ij,pq}c_pc_q+(P_8)_{ij,pq}c_pc_q, \tag{4.14}$$

where the projectors are given by

$$\begin{aligned}(P_{20})_{ij,pq}&=\tfrac{1}{2}(\delta_{ip}\delta_{jq}-\delta_{iq}\delta_{jp})-\tfrac{1}{3}f_{ijl}f_{lpq},\\ (P_8)_{ij,pq}&=\tfrac{1}{3}f_{ijl}f_{lpq}.\end{aligned} \tag{4.15}$$

The projection properties and orthogonality can be checked using well-known properties of $su(3)$ f tensors,[24,45] etc. Also, we have trivially,

$$P_{20}+P_8=U,\quad U_{ij,pq}=\tfrac{1}{2}(\delta_{ip}\delta_{jq}-\delta_{iq}\delta_{jp}), \tag{4.16}$$

where U is the relevant form of the unit operator in the ghost number $q=2$ space spanned by the antisymmetric tensors $c_i c_j$. Direct calculations on the explicit form for d_k given by (4.12) and for $(\mathbf{20}_2)_{ij}$ by (4.13) show that these have the C_2 eigenvalues 3 and 6 in (4.11).

The space spanned at $q=3$ by the tensor components $c_i c_j c_k$ gives rise easily to the singlet (0,0),

$$Y=f_{ijk}c_i c_j c_k = c_i d_i, \tag{4.17}$$

and the $su(3)$ octet,

$$e_i = d_{ijk} c_j d_k. \tag{4.18}$$

This is the only $q=3$ octet, since

$$\xi_i = f_{ijk} c_j d_k = 0 \tag{4.19}$$

follows from the definition (4.12) of d_k and the Jacobi identity for the f tensor. We note, however, that $\xi_i=0$ is a set of eight nonempty verifiable identities among various trilinears $c_i c_j c_k$. To build other irreducible tensors, it is natural to look at the tensors

$$c_i d_j - c_j d_i, \tag{4.20}$$

$$c_i d_j + c_j d_i, \tag{4.21}$$

with *a priori* 28 and 36 components. The former (4.20) is irreducible and defines $(\mathbf{20}_3)_{ij}$ as it stands, because $\xi_i=0$ yield eight identities automatically satisfied by its components. It is also not hard to check that the C_2 eigenvalue is 6. The latter (4.21) is not irreducible, but by extracting the scalar (4.17) and the octet (4.18), we find the irreducible tensor of $27=36-1-8$ components,

$$(\mathbf{27}_3)_{ij} = c_i d_j + c_j d_i - \tfrac{1}{4}\delta_{ij} Y - \tfrac{6}{5} d_{ijk} e_k. \tag{4.22}$$

It is easy to see that contracting with δ_{ij} and d_{ijk} gives zero as irreducibility requires. It is hard, needing a good selection of $su(3)$ such f- and d-tensor identities, as found in Ref. 24, to prove that C_2 indeed has eigenvalue 8 for $(\mathbf{27}_3)_{ij}$. We could turn results (4.17), (4.18), (4.20), and (4.22) into the form

$$c_i c_j c_k = \sum_R P^R_{ijk,pqr} c_p c_q c_r, \tag{4.23}$$

involving a complete set of orthogonal projectors for $R=1$, 8, 20, and 27.

Since the case at $q=4$ involves repetitions, it is best at this point to review the situation regarding octets. At $q=1$, 2, 3, we have

$$c_i, \quad d_i, \quad e_i, \tag{4.24}$$

and no others. At $q=4$, we find

$$f_i = d_{ijk} d_j d_k = \Omega_{i i_1 i_2 i_3 i_4} c_{i_1} c_{i_2} c_{i_3} c_{i_4}, \tag{4.25}$$

but $f_{ijk} d_j d_k \equiv 0$. A second octet that can be checked easily to be linearly independent of f_i is $Y c_i$. We may build other $q=4$ octets, but these will not give anything new, since, e.g., we can prove the results

$$d_{ijk} c_j e_k = -\tfrac{2}{3} c_i Y, \quad f_{ijk} c_j e_k = f_i. \tag{4.26}$$

It is thus now obvious that the complete family of octets can be presented as

$$\begin{array}{ccccccc} q=1 & 2 & 3 & 4 & 5 & 6 & 7 \\ c_i & d_i & e_i & f_i & & & \\ & & & Yc_i & Yd_i & Ye_i & Yf_i. \end{array} \tag{4.27}$$

As a check, we find that Yf_i (for example) is as expected,

$$Yf_1 \sim c_2c_3c_4c_5c_6c_7c_8 \sim *c_1. \tag{4.28}$$

An alternative but equivalent treatment would employ certain duals of f_i, e_i, d_i, c_i in place of Yc_i, Yd_i, Ye_i, Yf_i in (4.27). Of course, for the last case, we have just proved the easy bit of the equivalence. In fact, the use of duals in explicit work is much less convenient than the choice used in (4.27). To indicate this, and to do something instructive in its own right, we make explicit the dual relation of f_i and Yc_i (see also Ref. 24, Sec. 8).

To make contact with a dual to the octet f_i of (4.25) replace $c_{i_1}c_{i_2}c_{i_3}c_{i_4}$ there by $\epsilon_{i_1i_2i_3i_4j_1j_2j_3j_4}c_{j_1}c_{j_2}c_{j_3}c_{j_4}$ to reach

$$p_i = \Omega_{ii_1i_2i_3i_4}\epsilon_{i_1i_2i_3i_4j_1j_2j_3j_4}c_{j_1}c_{j_2}c_{j_3}c_{j_4}, \tag{4.29}$$

which clearly belongs to $\mathbf{8}^{\wedge 4}$. To relate p_i to Yc_i, we need the identity

$$\frac{1}{4!}\Omega_{ii_1i_2i_3i_4}\epsilon_{i_1i_2i_3i_4j_1j_2j_3j_4} = \frac{2}{\sqrt{3}}\delta_{i[j_1}f_{j_2j_3j_4]}, \tag{4.30}$$

in which the divisor 4! on the left is actually matched by one implicit in our definition of square antisymmetrization brackets on the right. Identity (4.30) allows us to prove

$$p_i = \frac{4!\cdot 2}{\sqrt{2}}c_iY = -\frac{4!\cdot 2}{\sqrt{3}}Yc_i, \tag{4.31}$$

as expected.

The contraction $i=j_1$ of (4.30) gives

$$\frac{1}{4!}\Omega_{i_1i_2i_3i_4i_5}\epsilon_{i_1i_2i_3i_4i_5j_1j_2j_3} = \frac{5}{2\sqrt{3}}f_{j_1j_2j_3}, \tag{4.32}$$

which is an evident and easily checked analog of the result,

$$\frac{1}{3!}f_{j_1j_2j_3}\epsilon_{j_1j_2j_3i_1i_2i_3i_4i_5} = -2\sqrt{3}\Omega_{i_1i_2i_3i_4i_5}, \tag{4.33}$$

given in Ref. 24, Eq. (8.14). The latter is a contraction of the more useful identity,

$$\frac{1}{3!}f_{ij_1j_2}\epsilon_{j_1j_2i_1i_2i_3i_4i_5i_6} = -4\sqrt{3}\delta_{i[i_1}\Omega_{i_2i_3i_4i_5i_6]}. \tag{4.34}$$

This may be used, as we used (4.30), to reach, e.g., eventually the dual relationship of d_i to Ye_i.

While the above tensorial analysis provides an explicit construction from first principles of all the entries of (4.7), the use of s_2 and s_4 expedites explicit work. For example, since $[s_2,X_i]=0=[s_4,X_i]$, s_2 and s_4 also commute with C_2 and C_3. Thus, s_2 (for example) either raises the ghost number of a tensor by one, leaving its $su(3)$ nature unaltered or else annihilates it. Thus $s_2c_i \sim d_i$, $s_2d_i=0$. Similarly, $s_4c_i \sim f_i$ and $\overline{s_2}s_4f_i \sim e_i$. Likewise,

$$s_2(\mathbf{20}_2)_{ij}=s_2(c_ic_j-\tfrac{1}{3}f_{ijk}d_k)\sim d_ic_j-c_id_j=-(\mathbf{20}_3)_{ij}\,, \tag{4.35}$$

since $s_2d_i=0$, which confirms what has been seen to hold above.

Further, we might expect $s_2(\mathbf{27}_3)_{ij}$ to yield one of the required $(\mathbf{27}_4)_{ij}$. Indeed $s_2d_i=0$, $s_2Y=0$, and $s_2e_k=f_k$ allow us to write

$$s_2(\mathbf{27}_3)_{ij}=d_id_j-\tfrac{3}{5}d_{ijk}f_k\equiv(\mathbf{27}_4)_{ij}\,. \tag{4.36}$$

A second 27-tensor in $\mathbf{8}^{\wedge 4}$ that is linearly independent of $(\mathbf{27}_4)_{ij}$ of (4.36) is suggested immediately by duality arguments. One replaces $c_{i_1}c_{i_2}c_{i_3}c_{i_4}$ in $d_id_j=f_{ii_1i_2}f_{ji_3i_4}c_{i_1}c_{i_2}c_{i_3}c_{i_4}$, etc. by $\epsilon_{i_1i_2i_3i_4j_1j_2j_3j_4}c_{j_1}c_{j_2}c_{j_3}c_{j_4}$. We thereby reach a tensor $(\mathbf{27}'_4)_{ij}$ that is plainly linearly independent of $(\mathbf{27}_4)_{ij}$. It turns out to be proportional to

$$(\mathbf{27}'_4)_{ij}=c_ie_j+c_je_i-\tfrac{4}{5}d_{ijk}Yc_k\,, \tag{4.37}$$

which can be seen to satisfy $d_{ijl}(27'_4)_{ij}=0$, using (4.26), as well as $(27'_4)_{ii}=0$, so that it is irreducible, with 27 components. Further, $(27_5)_{ij}$ can now be written down explicitly by the action of s_2 on $(27'_4)_{ij}$. No systematic work on projectors for $q=4$ has been done.

C. The Laplacian *W*

From the analysis of Sec. IV B of the $su(3)$ representations contained in $\mathbf{8}^{\wedge q}$, $0\leqslant q\leqslant 8$, where q is the ghost number, it can be seen that the states φ of the system are labeled by the eigenvalues of the ghost number operator $Q=c^i\pi_i$, $Q\varphi=q\varphi$, and of the $su(3)$ Casimirs C_2 and C_3 that label states within each $su(3)$ representation. Since W commutes with Q, C_2, C_3, we expect it to have well-defined eigenvalues on all the states of the system, and we might further expect it to be defined as a specific function of Q, C_2, C_3.

Some progress can be made analytically to compute W eigenvalues. For example, for c_i and d_i given by (4.12), which describe $q=1$, $q=2$ octets, we may compute directly from (4.5) the results

$$Wc_i=5c_i\,,\quad Wd_i=0. \tag{4.38}$$

These calculations, the latter already nontrivial, depend, among other things, on the identities

$$\begin{aligned}\Omega_{i_1i_2i_3i_4p}\Omega_{i_1i_2i_3i_4q}&=5\,\delta_{pq}\,,\\ \Omega_{i_1i_2i_3ab}\Omega_{i_1i_2i_3pq}&=\tfrac{1}{2}(\delta_{ap}\delta_{bq}-\delta_{aq}\delta_{bp}+f_{abi}f_{pqi}),\end{aligned} \tag{4.39}$$

of which only the first follows from the definition of Ω easily. Note also that since W distinguishes between different octets, Eq. (4.38), it cannot be a pure function of the Casimirs: it depends also on Q, which does not belong to the $\mathcal{U}(su(3))$ enveloping algebra.

The results of Sec. IV A also allow the minimal polynomials for C_2 and C_3 to be deduced. These are

$$C_2(C_2-3)(C_2-6)(C_2-8)=0, \tag{4.40}$$

$$C_3(C_3+9)(C_3-9)=0, \tag{4.41}$$

and the orthogonal projectors on the various eigenspaces for C_2 and C_3 are

$$P_0^{(2)} = -\tfrac{1}{144}(C_2-3)(C_2-6)(C_2-8), \quad P_0^{(3)} = -\tfrac{1}{81}(C_3+9)(C_3-9),$$
$$P_3^{(2)} = \tfrac{1}{45}C_2(C_2-6)(C_2-8), \quad P_{-9}^{(3)} = \tfrac{1}{162}C_3(C_3-9),$$
$$P_6^{(2)} = -\tfrac{1}{36}C_2(C_2-3)(C_2-8), \quad P_9^{(3)} = \tfrac{1}{162}C_3(C_3+9), \tag{4.42}$$
$$P_8^{(2)} = \tfrac{1}{80}C_2(C_2-3)(C_2-6).$$

Further progress by analytic methods soon becomes difficult, and we have made use of FORM.[46] This enables us first to compute all W eigenvalues, discussed below, and to find the following identities:

$$C_2C_3 = 6C_3,$$
$$C_3{}^2 = -\tfrac{9}{4}C_2(C_2-3)(C_2-8),$$
$$C_2W = 3W - \tfrac{1}{2}C_2(C_2-3)(C_2-8), \tag{4.43}$$
$$W^2 = 5W + \tfrac{2}{27}C_3{}^2.$$

These results allow the recovery of (4.40), (4.41) as a mild check on our procedures, and the deduction of the minimal polynomial of W,

$$W(W-5)(W-6) = 0, \tag{4.44}$$

which comprises, as it should, all the eigenvalues of W found in practice. Also, the orthogonal projectors onto the eigenspaces of W are

$$P_0^{(W)} = \tfrac{1}{30}(W-5)(W-6), \quad P_5^{(W)} = -\tfrac{1}{5}W(W-6), \quad P_6^{(W)} = \tfrac{1}{6}W(W-5). \tag{4.45}$$

Various useful inferences can be made regarding eigenspaces. For example, alongside the previous result $\ker C_2 \subseteq \ker W$, we have $\ker W \subseteq \ker C_3$. Also,

$$P_9^{(3)} + P_{-9}^{(3)} = P_6^{(2)} = P_6^{(W)},$$
$$P_0^{(2)}P_0^{(W)} = P_0^{(2)},$$
$$P_8^{(2)}P_0^{(W)} = P_8^{(2)}, \tag{4.46}$$
$$P_0^{(W)} + P_5^{(W)} = P_0^{(3)}.$$

So far no explicit expression for W in terms of Q, C_2, C_3 is at hand. The major complication in the pattern of the W-eigenvalues of the $su(3)$ representations in $\mathbf{8}^{\wedge q}$ concerns the octets. For these, the ghost number $q=1$, 4, 4, 7 octets have eigenvalue $W=5$ and the $q=2$, 3, 5, 6 octets $W=0$. This suggests the use of Lagrangian interpolation to define a function,

$$f(q) = \tfrac{1}{360}[(q-4)^2 - 10(q-7)(q-1)](q-2)(q-3)(q-5)(q-6), \tag{4.47}$$

which equals 1 at $q=1$, 4, 4, 7 and 0 at $q=2$, 3, 5, 6, so that for these values, $f(q)^2 = f(q)$. It is then immediate to see that the formula

$$W = \tfrac{1}{9}C_2(C_2-6)(C_2-8)f(Q) - \tfrac{1}{6}C_2(C_2-3)(C_2-8) = 5P_3^{(2)}f(Q) + 6P_6^{(2)}, \tag{4.48}$$

correctly predicts the W-eigenvalues of all states. The last two equations of (4.43) also follow directly, using projector properties, $f(Q)^2 = f(Q)$, and the second equation of (4.43). We should

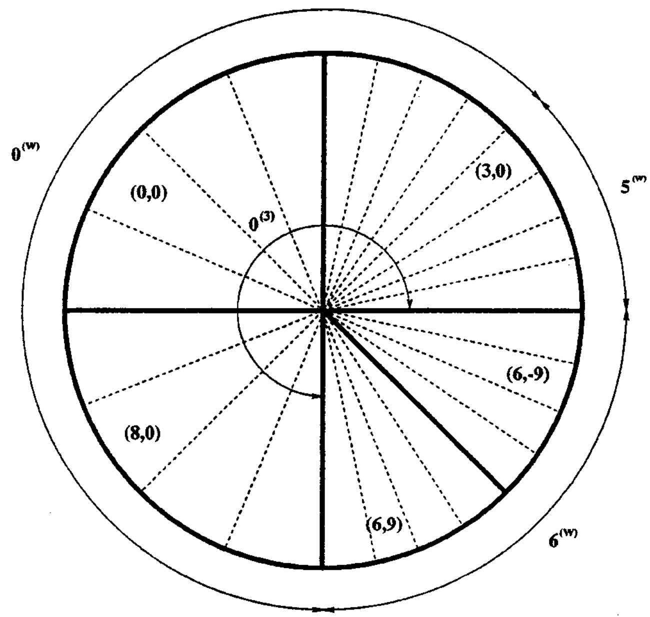

FIG. 1. The spectrum of the Casimirs (C_2,C_3) and of $W_4 \propto W$.

stress here that $f(Q)^2=f(Q)$ holds only for $q\in\{1,2,...,7\}$, whereas the allowed range of values of q is {0, 1,..., 8}. But this does not matter for (4.48), because although $f(Q)$ is finite $(=-27)$ at $q=0$ and $q=8$, $C_2=0$ for the $q=0$ and $q=8$ states. Finally, FORM confirms that W defined by (4.48), and $W=4!W_4$ given by (4.5) are equal as operators.

A final remark about the form of (4.47), (4.48) is in order here. Observing that $f(Q)=f(8-Q)$, one is led to write (4.47) in terms of $u:=Q(8-Q)$, finding

$$f(Q)=F(u):=\tfrac{1}{40}(u-6)(u-12)(u-15) \tag{4.49}$$

(a form that can also be directly derived by Lagrangian interpolation). A different approach is to start from the minimal polynomial for u,

$$u(u-7)(u-12)(u-15)(u-16)=0, \tag{4.50}$$

and write down directly a function $\tilde{F}(u)$, with $\tilde{F}(u)=0$ at $u=0$, 12, 15 and $\tilde{F}(u)=1$ at $u=7$, 16,

$$\tilde{F}(u)=P_7^{(u)}+P_{16}^{(u)}, \tag{4.51}$$

the projectors $P_\lambda^{(u)}$ being defined in the standard way from (4.50). Using this $\tilde{F}(u)$ in place of $F(u)=f(Q)$ in (4.48) also gives correctly W—the difference $\tilde{F}-F$ is annihilated by $P_3^{(2)}$. We note incidentally that the operators $M_{ij}:=c_i\pi_j-c_j\pi_i$, $i<j$, generate the algebra spin(8), the quadratic Casimir of which is proportional to u (since $M_{ij}M_{ij}=-2u$), i.e., (4.48) gives W in terms of the quadratic Casimirs of $su(3)$ and spin(8).

The results of the previous analysis may be summarized in the diagram of Fig. 1 representing the spectra of C_2, C_3, and W. The solid circular disk represents all the $2^8=256$ states available in $\oplus_{i=0}^{8}\mathbf{8}^{\wedge i}$. The four quadrants represent the four eigenspaces (0, 3, 6, 8) of C_2 while the numbers in parentheses are the eigenvalues of C_2, C_3 valid in each disk segment (bordered by solid black lines). Dashed lines within each disk segment separate multiple copies of the same irreducible representation, corresponding, in general, to states of a different ghost number. The arcs outside the disk specify the three (0,5,6) W-eigenspaces. We summarize its key features.

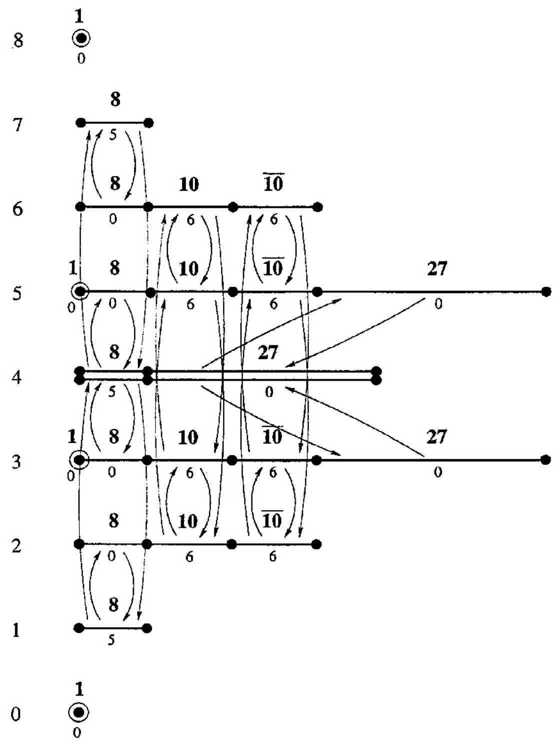

FIG. 2. Decomposition of $\oplus_{i=0}^{8}\mathbf{8}^{\wedge i}$ into irreps of $su(3)$ and the action of $s_2(\overline{s_2})$ and $s_4(\overline{s_4})$ on them (notice that $[s_2(\overline{s_2}),W]\neq 0$). The eigenvalue of W appears under each irrep.

(i) The eigenspace $\mathrm{Im}\,P_6^{(2)}$, equal to $\mathrm{Im}\,P_6^{(W)}$, is split into two parts with the same number of states, labeled by the C_3 eigenvalues 9 and -9. With the help of (4.11) we recognize these, respectively, as the **10** and $\overline{\mathbf{10}}$ $su(3)$ representations, each of which appears four times, with ghost numbers 2, 3, 5, and 6.

(ii) The (0, 0) subspace (i.e., $\ker C_2 = \mathcal{K}_2$), contains four invariant states [the $su(3)$ singlets **1**], with ghost numbers 0, 3, 5, and 8. All of them are W harmonic as well, i.e., Σ_4 singlets.

(iii) (8, 0) (=**27**) appears four times, with ghost numbers 3, 5, and 4 (twice) and is also W harmonic.

(iv) (3, 0) (=**8**) appears eight times, with ghost numbers 1 to 7 (twice for 4).

The diagram shows that $\mathcal{K}_2 + \mathrm{Im}\,P_8^{(2)} \subset \mathcal{K}_4 \subset \ker C_3$. $\mathcal{K}_4$ contains half of the copies of the adjoint representation, the rest belonging to the W eigenvalue 5.

To look now at the representations of Σ_2 and Σ_4 in (2.22) and (3.15), it is convenient to depict the $su(3)$ representations as in the diagram of Fig. 2 and to analyze there the role played by s_2, s_4 and their adjoints $\overline{s_2}$, $\overline{s_4}$ in interconnecting them. Each *straight line* segment in this diagram represents an irreducible representation. All segments in the same line correspond to the same ghost number, while singlets are represented by circles. The arrows between segments depict the action of the s's [solid (black) lines for s_2, $\overline{s_2}$, dotted (gray) lines for s_4, $\overline{s_4}$]; the number below each segment is the W eigenvalue of the irreducible representation. In the following we denote, e.g., by $\mathbf{10}_3$ the irreducible representation **10** with ghost number 3, while a further superscript $u(l)$ [for upper (lower)] distinguishes between the two ghost number 4 representations $\mathbf{8}_4$'s and $\mathbf{27}_4$'s. We point out the following.

(i) Referring to multiplets of the superalgebra (3.15) quartets-turned-into-pairs-of-doublets, according to the remark of Sec. III C, appear three times. For Σ_2, the quartet $\{\mathbf{8}_4^u, \mathbf{8}_5, \mathbf{8}_3, \mathbf{8}_4^l\}$ actually consists of the pair of doublets $\{\mathbf{8}_5, \mathbf{8}_4^u\}$, $\{\mathbf{8}_4^l, \mathbf{8}_3\}$—a similar pattern is exhibited by the **27**'s in the same orders as well as by the Σ_4 quartet $\{\mathbf{8}_4^u, \mathbf{8}_7, \mathbf{8}_1, \mathbf{8}_4^l\}$. The degeneracy seen in the

$q=4$ line is then resolved by noting that s_2 annihilates one octet and $\overline{s_2}$ the other (and similarly for the **27**).

(ii) Besides the above ''split quartets,'' we also have the Σ_2 doublets $\{\mathbf{8}_2,\mathbf{8}_1\}$, $\{\mathbf{10}_3,\mathbf{10}_2\}$, $\{\overline{\mathbf{10}}_3,\overline{\mathbf{10}}_2\}$, the Σ_4 doublets $\{\mathbf{10}_5,\mathbf{10}_2\}$, $\{\overline{\mathbf{10}}_5,\overline{\mathbf{10}}_2\}$, and their $*$ images $\{\mathbf{8}_7,\mathbf{8}_6\}$, $\{\mathbf{10}_6,\mathbf{10}_5\}$, $\{\overline{\mathbf{10}}_6,\overline{\mathbf{10}}_5\}$ and $\{\mathbf{10}_6,\mathbf{10}_3\}$, $\{\overline{\mathbf{10}}_6,\overline{\mathbf{10}}_3\}$, respectively. Notice that W changes eigenvalue within all Σ_2 doublets involving **8**'s, reflecting its failure to commute with s_2.

(iii) The $su(3)$ (and hence Σ_2, Σ_4) singlet $\mathbf{1}_0$ is simply the constant monomial 1, while $\mathbf{1}_3$ is the three-cocycle $f_{ijk}c^ic^jc^k=Y$. The other two singlets are the ''top form'' $c_1\cdots c_8$ and the five-cocycle $\Omega_{i_1\cdots i_5}c^{i_1}\cdots c^{i_5}$, $*$ images of the first two, respectively.

(iv) $\mathbf{8}_1$ consists of the eight c^k's. This is ''lifted'' by s_2 to give $\mathbf{8}_2\sim\{f_{ij}{}^kc^ic^j\}$ and by s_4, giving $\mathbf{8}_4^l\sim\{\Omega_{i_1i_2i_3i_4}{}^kc^{i_1}c^{i_2}c^{i_3}c^{i_4}\}$. $\mathbf{8}_3$ is the image of $\mathbf{8}_1$ under $\overline{s_2}s_4$, i.e., $\mathbf{8}_3 \sim\{f_{i_1ab}\Omega_{i_2i_3}{}^{abk}c^{i_1}c^{i_2}c^{i_3}\}$. The $q\to r-q$ symmetry accounts for the rest of the **8**'s. Notice that $\mathbf{8}_2$, $\mathbf{8}_3$ cannot be lifted by s_4 since they are W-harmonic.

V. CONCLUDING REMARKS

We have introduced and studied in this paper the supersymmetry algebra generated by the higher-order BRST operators. The central term in the algebra is given, in the standard (lowest-order) case, by the (quadratic) Casimir. As shown explicitly by the expression of W_4 for the algebra $\mathcal{G}=su(3)$, the higher-order Laplacians may involve the ghost number operator, which, unlike the Casimir–Racah operators, lies outside the enveloping algebra $\mathcal{U}(\mathcal{G})$. Thus, the fact that $\Delta\in\mathcal{U}(\mathcal{G})$ in the standard case is rather exceptional.

We wish to conclude with a purely mathematical remark. Using the correspondence $c_i\leftrightarrow$LI forms on the group manifold, the standard BRST operator s_2 may be identified with the exterior derivative d acting on forms. The basic properties of d, $d:\wedge^q\to\wedge^{q+1}$, $d^2=0$ (and of the codifferential δ) may be extended by introducing generalized operators $\tilde{d}$ in two different ways. One is by replacing the exterior differential by a higher-order nilpotent endomorphism $\tilde{d}'$ satisfying $(\tilde{d}')^k=0$, to study the associated generalized homology, etc.[47,48] This approach is reminiscent of the one used to generalize ordinary supersymmetry to fractional supersymmetry (for a review with earlier references see Ref. 49). The second one replaces d by a pth-order differential, $\tilde{d}$, p odd, satisfying $\tilde{d}_p:\wedge^q\to\wedge^{q+p}$, $\tilde{d}_p^2=0$, and it is this second point of view that corresponds to the analysis presented in this paper. In fact, the higher BRST operator s_{2m-3} may be considered as an explicit construction of this differential for $p=(2m-3)$, which acts *on LI* forms on the group manifold by translating (3.8) using the above correspondence. $\tilde{d}_p$ is an odd operator satisfying

$$\tilde{d}_p(\alpha\wedge\beta)=(\tilde{d}_p\alpha)\wedge\beta+(-1)^n\alpha\wedge(\tilde{d}_p\beta), \tag{5.1}$$

where n is the order of the LI form α. We recall that, using the (standard) product between manifolds and forms (given by $\langle\mathcal{M},\alpha\rangle=\int_{\mathcal{M}}\alpha$) one can define the adjoint ∂ of the exterior derivative d. Acting on manifolds, ∂ reduces their dimension by one, is nilpotent, and admits the interpretation as a boundary operator. Using an analogous procedure, one might think of defining the adjoint $\tilde{\partial}_p$ of $\tilde{d}_p$ as an operator. Acting on manifolds, it would reduce their dimension by p, being also nilpotent, and the question would arise whether it, too, admits a simple topological interpretation. One might also ask further, whether an analog of Stokes' theorem could be formulated along these lines or whether the spectrum of the higher-order Laplacians studied here provides topological information about the underlying manifold. We do not know whether these mathematical constructions involving $\tilde{d}_p$ can be carried through, in general.

To conclude we would like to stress that the cohomological properties used in this paper are also relevant in other related fields, although it may not be directly apparent. They determine and classify, for instance, the local conserved charges, in principal, chiral models (see Refs. 50 and references therein), and are also important in W algebras (see, for instance, Refs. 51–54), where BRST-type techniques, and hence Lie algebra cohomology, are relevant.

ACKNOWLEDGMENTS

This research is partially supported by a research Grant No. PB96-0756 from the DGICYT, Spain. C. C. wishes to thank the Spanish Ministry of Education and Culture for a postdoctoral fellowship and J. C. P. B. wishes to thank the Spanish MEC and CSIC for an FPI grant. A. J. M. wishes to thank the University of Valencia for hospitality. The research of A. J. M. is supported in part by a grant from PPARC, United Kingdom.

[1] C. Becchi, A. Rouet, and R. Stora, ''Renormalization and gauge theories,'' Ann. Phys. (N.Y.) **98**, 287–321 (1976).
[2] I. V. Tyutin, ''Gauge invariance in field theory and in statistical mechanics,'' Lebedev preprint FIAN #39 (unpublished).
[3] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, 1992).
[4] J. Gomis, J. Paris, and S. Samuel, ''Antibracket, antifields and gauge-theory quantization,'' Phys. Rep. **259**, 1–145 (1995).
[5] E. Witten, ''A note on the antibracket formalism,'' Mod. Phys. Lett. A **5**, 487–494 (1990).
[6] O. M. Khurdaverian and A. P. Nersessian, ''On the geometry of the Batalin–Vilkovisky formalism,'' Mod. Phys. Lett. A **8**, 2377–2385 (1993).
[7] M. Alexandrov, M. Kontsevich, A. Schwarz, and O. Zaboronsky, ''The geometry of the master equation and topological quantum field theory,'' Int. J. Mod. Phys. A **12**, 1405–1430 (1997).
[8] A. Nersessian, ''Antibracket and non-Abelian equivariant cohomology,'' Mod. Phys. Lett. A **10**, 3043–3050 (1995).
[9] F. Brandt, M. Henneaux, and A. Wilch, ''Extended antifield formalism,'' Nucl. Phys. B **510**, 640–656 (1998).
[10] L. Álvarez-Gaumé and L. Baulieu, ''The two quantum symmetries associated with a classical symmetry,'' Nucl. Phys. B **212**, 255–267 (1983).
[11] D. McMullan, ''Use of ghosts in the description of physical states,'' Phys. Rev. D **33**, 2501–2503 (1986).
[12] J.-L. Gervais, ''Group theoretic approach to the string field theory action,'' Nucl. Phys. B **276**, 339–348 (1986).
[13] I. Bars and S. Yankielowicz, ''Becchi–Rouet–Stora–Tyutin symmetry, differential geometry, and string field theory,'' Phys. Rev. D **35**, 3878–3889 (1987).
[14] J. W. van Holten, ''The BRST complex and the cohomology of compact Lie algebras,'' Nucl. Phys. B **339**, 158–176 (1990).
[15] L. Baulieu, ''On the symmetries of topological quantum field theories,'' Int. J. Mod. Phys. A **50**, 4483–4500 (1995).
[16] H. S. Yang and B.-H. Lee, ''Lie algebra cohomology and group structure of gauge theories,'' J. Math. Phys. **37**, 6106–6120 (1996).
[17] The Killing tensor k_{ij} is proportional to δ_{ij} since $\mathcal{G}$ is compact and semisimple.
[18] T. Kawai, ''Remarks on a class of BRST operators,'' Phys. Lett. B **168**, 355–359 (1986).
[19] J. M. Figueroa-O'Farrill and T. Kimura, ''Some remarks on BRST cohomology, Kähler algebra and 'no-ghost' theorems,'' ITP-SP-88-34, unpublished.
[20] L. Bonora and P. Cotta-Ramusino, ''Some remarks on BRST transformations, anomalies and the cohomology of the Lie algebra of the group gauge transformations,'' Commun. Math. Phys. **87**, 589–603 (1983).
[21] G. Racah, ''Sulla caratterizzazione delle rappresentazioni irreducibili dei gruppi semi-simplici di Lie,'' Lincei-Rend. Sc. fis. mat. e nat. **VIII**, 108–112 (1950), Princeton lectures, CERN-61-8 [reprinted in Ergeb. Exact Naturwiss. **37**, 28–84 (1965) (Springer-Verlag, Berlin, 1965)].
[22] I. M. Gel'fand, ''The center of an infinitesimal group ring,'' Mat. Sb. **26**, 103–112 (1950) (english translation by Los Alamos Sci. Lab., ACE-TR-6133, 1963).
[23] A. M. Perelomov and V. S. Popov, ''Casimir operators for semisimple groups,'' Math. USSR Izv. **2**, 1313–1335 (1968).
[24] J. A. de Azcárraga, A. J. Macfarlane, A. J. Mountain, and J. C. Pérez Bueno, ''Invariant tensors for simple groups,'' Nucl. Phys. B **510**, 657–687 (1998).
[25] J. A. de Azcárraga and J. C. Pérez Bueno, ''Higher-order simple Lie algebras,'' Commun. Math. Phys. **184**, 669–681 (1997).
[26] T. Lada and J. Stasheff, ''Introduction to SH Lie algebras for physicists,'' Int. J. Theor. Phys. **32**, 1087–1103 (1993).
[27] E. Witten and B. Zwiebach, ''Algebraic structures and differential geometry in two-dimensional string theory,'' Nucl. Phys. B **377**, 55–112 (1992).
[28] B. H. Lian and G. J. Zuckerman, ''New perspectives on the BRST-algebraic structure of string theory,'' Commun. Math. Phys. **154**, 613–646 (1993).
[29] K. Bering, P. H. Damgaard, and J. Alfaro, ''Algebra of higher antibrackets,'' Nucl. Phys. B **478**, 459–503 (1996).
[30] F. Akman, ''On some generalizations of Batalin–Villkovisky algebras,'' J. Pure Appl. Algebra **120**, 105–141 (1997).
[31] J. Stasheff, ''The (secret?) homological algebra of the Batalin–Vilkovisky approach,'' hepth/9712157, to appear in *Proceedings of the Conference on Secondary Calculus and Cohomological Physics,* Moscow, August 1997.
[32] J. A. De Azcárraga, J. M. Izquierdo, and J. C. Pérez Bueno, ''An introduction to some novel applications of Lie algebra cohomology in mathematics and physics,'' physics/9803046, to appear in the *Proceedings of the VI Fall Workshop on Geometry and Physics*, Salamanca, September 1997.
[33] C. Chevalley and S. Eilenberg, ''Cohomology theory of Lie groups and Lie algebras,'' Trans. Am. Math. Soc. **63**, 85–124 (1948).
[34] J. A. De Azcárraga and J. M. Izquierdo, *Lie Groups, Lie Algebras, Cohomology and Some Applications in Physics* (Cambridge University Press, Cambridge, 1995).
[35] The algebra (2.7) can be represented by a Clifford algebra (see, e.g., Ref. 14). Namely, if we define $c_i = \frac{1}{2}(\gamma_i$

$-i\gamma_{i+r})$, $\pi_i=\frac{1}{2}(\gamma_i+i\gamma_{i+r})$, $i=1,...,r$, where the γ's are the generators of a $2r$-dimensional Clifford algebra, then c_i and π_i verify the relations (2.7).

[36] Using the c's to write ψ [Eq. (1.5)], rather than the ω's of (2.12), one might introduce a Berezin[37] integral measure to define $\langle\psi',\psi\rangle$ above as $\int dc^1\cdots dc^r\psi'^\dagger\psi$[14] for states ψ' and ψ of ghost numbers q and $r-q$, respectively. However, this leads to a product that is not positive definite[14] and, moreover, does not have the natural geometrical interpretation above.

[37] F. A. Berezin, ''Differential forms on supermanifolds,'' Sov. J. Nucl. Phys. **30**, 605–609 (1979); *Introduction to Superanalysis* (Reidel, Dordrecht, 1987).

[38] Taking advantage of the Cartan formalism by means of the equivalences $s_2\sim d$, $\overline{s_2}\sim(-1/2)L_j i_j$ (where i_j indicates the inner product) and $X_j\sim L_j$, we may rapidly find $W_2\sim(-1/2)[dL_j i_j+L_j i_j d]=(-1/2)L_j[d i_j+i_j d]=(-1/2)L_j L_j$.

[39] The CE analog to the BRST q-cochains, the LI q-forms on G, automatically satisfy $L_{X^R}\psi=0$, since the RI vector fields X^R on G generate the left transformations. The invariance under the right transformations ($L_X\psi=0$, where X is a LI vector field, or $X\psi=0$ in the BRST formulation) is an additional condition. Thus, invariance above really means bi-invariance (under the left and right group translations) in the CE formulation of Lie algebra cohomology.

[40] W. V. D. Hodge, *The Theory and Applications of Harmonic Integrals* (Cambridge University Press, Cambridge, 1941).
[41] H. Samelson, ''Topology of Lie groups,'' Bull. Am. Math. Soc. **57**, 2–37 (1952).
[42] W. Greub, S. Halperin, and R. Vanstone, *Connections, Curvature and Cohomology* (Academic, New York, 1976), Vol. III.
[43] L. C. Biedenharn, Phys. Lett. **3**, 69–70 (1963).
[44] A. Sudhery, Ph.D. thesis, Cambridge University, 1969.
[45] A. J. Macfarlane, A. Sudbery, and P. H. Weisz, ''On Gell–Mann's λ matrices, *d*- and *f*-tensors, octets, and parametrisations of SU(3),'' Commun. Math. Phys. **11**, 77 (1968).
[46] J. A. M. Vermaseren, Symbolic Manipulation with FORM version 2: Tutorial and Reference Manual (CAN, 1992).
[47] M. Dubois-Violette, ''Generalized spaces with $d^N=0$ and the q-differential calculus,'' Czech. J. Phys. **46**, 1227–1233 (1997).
[48] R. Kerner, ''Z_n-graded differential calculus,'' Czech. J. Phys. **47**, 33–40 (1997).
[49] R. S. Dunne, A. J. Macfarlane, J. A. De Azcárraga, and J. C. Pérez Bueno, ''Geometrical foundations of fractional supersymmetry,'' Int. J. Mod. Phys. A **19**, 3275–3305 (1997).
[50] J. M. Evans, M. Hassan, N. J. MacKay, and A. J. Mountain, ''Local conserved charges and principal chiral models,'' hep-th/9902008.
[51] F. A. Bais, P. Bouwknegt, M. Surridge, and K. Schoutens, ''Extensions of the Virasoro algebra constructed from Kac–Moody algebras using higher order Casimir invariants,'' Nucl. Phys. B **304**, 348–370 (1988).
[52] E. Bergshoeff, A. Bilal, and K. S. Stelle, ''*W*-symmetries: gauging and symmetry,'' Mod. Phys. Lett. A **6**, 4959–4984 (1991).
[53] H. Lu, C. N. Pope, X. J. Wang, and K. W. Xu, ''The complete cohomology of the W_3 string,'' Class. Quantum Grav. **11**, 967–981 (1994).
[54] P. Bouwknegt, J. McCarthy, and K. Pilch, *The W_3 Algebra* (Springer-Verlag, Berlin, 1996).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Product integral formalism and non-Abelian Stokes theorem

Robert L. Karp[a)] and Freydoon Mansouri[b)]
Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221

Jung S. Rno[c)]
Department of Physics, University of Cincinnati-RWC, Cincinnati, Ohio 45236

(Received 18 March 1999; accepted for publication 14 July 1999)

We make use of the properties of product integrals to obtain a surface product integral representation for the Wilson loop operator. The result can be interpreted as the non-Abelian version of Stokes' theorem. © *1999 American Institute of Physics.*[S0022-2488(99)04610-1]

I. INTRODUCTION

Our main purpose of this work is to make use of product integrals to give two unambiguous proofs of the non-Abelian version of Stokes' theorem. The product integral formalism has been used extensively in the theory of differential equations and of matrix-valued functions.[1] In the latter context, it has a built-in feature for keeping track of the *order* of the matrix-valued functions involved. As a result, product integrals are ideally suited for the description of *path ordered* quantities such as holonomies. Moreover, since the theory is well developed independently of particular applications, we can be confident that the properties of such path ordered quantities that we establish using this method are correct and unambiguous.

Among the important advantages of the product integral representation of the path-dependent exponential of a matrix-valued function is that in such a framework the Banach space structure of the corresponding matrix-valued functions is already built into the formalism. In particular, for a closed path enclosing an orientable two-surface, this will permit a surface representation of such operators. Based on the central role of Stokes' theorem in physics and in mathematics, it is not surprising that the non-Abelian version of this theorem has attracted a good deal of attention in the physics literature.[2–12] The central features of the earlier attempts[2–8] have been reviewed and improved upon in a recent work.[9] Other recent works on the non-Abelian Stokes theorem[10–12] focus on specific problems such as confinement,[11] zigzag symmetry[12] suggested by Polyakov,[13] etc. With one exception,[9] the authors of these works seem to have been unaware of a 1927 work in the mathematical literature by Schlessinger[14] that bear strongly on the content of this theorem. Schlesinger's work dealt with integrals of matrix-valued functions and their ordering problems. This amounts to establishing the non-Abelian Stokes theorem in two (target space) dimensions. By an appropriate extension and reinterpretation of his results, we show using the product integral approach that this theorem is valid in any target space dimension.

This work is organized as follows: To make this manuscript self-contained, we review in Sec. II the main features of product integration[1] and state without proof a number of theorems that will be used in the proof of the non-Abelian Stokes theorem. In Sec. III, we deal with path ordered exponentials of matrix-valued functions that can be expressed as product integrals and turn to the proof of the non-Abelian Stokes theorem for orientable surfaces. In Sec. IV, we give a variant of this proof. In Sec. V, we explicitly demonstrate the gauge covariance of the results obtained in Secs. III and IV. Finally, Sec. V is devoted to some additional remarks.

[a)]Electronic mail: karp@physics.uc.edu
[b)]Electronic mail: mansouri@uc.edu
[c)]Electronic mail: rno@uc.edu

II. SOME PROPERTIES OF PRODUCT INTEGRALS

One of the initial motivations for the introduction of product integrals was[1] to solve differential equations of the type

$$Y'(s)=A(s)Y(s). \tag{1}$$

In this expression, $Y(s)$ is an n-dimensional vector, $A(s)$ is a matrix-valued function, and a prime indicates differentiation. So, for two real numbers a and b, the problem is to obtain $Y(b)$ given $Y(a)$. To deal with this problem, we make a partition $P=\{s_0,s_1,\ldots,s_n\}$ of the interval $[a,b]$. Let $\Delta s_k=s_k-s_{k-1}$ for $k=1,\ldots,n$, and set $a=s_0$, $b=s_n$. Then, solving the differential equation in each subinterval, we can write approximately[1]

$$Y(b)\approx\prod_{k=1}^{n} e^{A(s_k)\Delta s_k}Y(a)\equiv\Pi_P(A)Y(a). \tag{2}$$

Since $A(s)$ is matrix valued, the order in this product is important. Let $\mu(P)$ be the length of the longest Δs_k in the partition P. Then, as $\mu(P)\to 0$, we get

$$Y(b)=\lim_{\mu(P)\to 0}\Pi_P(A)Y(a). \tag{3}$$

The limit is clearly valid for all $Y(a)$.

The limit of the ordered product on the right-hand side of Eq. (3) is the fundamental expression in the definition of a product integral.[1] It is formally defined, in an obvious notation, as

$$\prod_{a}^{x} e^{A(s)ds}=\lim_{\mu(P)\to 0}\Pi_P(A)\equiv F(x,a). \tag{4}$$

It is easy to see that $F(x,a)$ satisfies the differential equation

$$\frac{d}{dx}F(x,a)=A(x)F(x,a), \tag{5}$$

with $F(a,a)=1$. The corresponding integral equation is

$$F(x,a)=1+\int_a^x ds\, A(s)F(s,a). \tag{6}$$

Clearly, $F(a,a)=1$, and $F(x,a)$ is unique.

Consider now some of the properties of the product integral matrices. For each $x\in[a,b]$ the product integral is nonsingular, and its determinant is given by

$$\det\left(\prod_{a}^{x} e^{A(s)ds}\right)=e^{\int_a^x \mathrm{tr}\,A(s)ds}. \tag{7}$$

In analogy with the additive property of ordinary integrals, product integrals have the multiplicative property, or the composition rule,

$$\prod_{z}^{x} e^{A(s)ds}=\prod_{y}^{x} e^{A(s)ds}\prod_{z}^{y} e^{A(s)ds}, \tag{8}$$

where $x,y,z\epsilon[a,b]$ and $z\leqslant y\leqslant x$. The result is independent of the choice of y and any further decomposition of the products on the right-hand side.

Derivatives with respect to the end points are given by

$$\frac{\partial}{\partial x}\prod_{y}^{x} e^{A(s)ds}=A(x)\prod_{y}^{x} e^{A(s)ds}, \tag{9}$$

and

$$\frac{\partial}{\partial y}\prod_{y}^{x} e^{A(s)ds}=-\prod_{y}^{x} e^{A(s)ds}A(y). \tag{10}$$

One of the fundamental features associated with a connection is the notion of parallel transport. To see how it can be formulated in product integral formalism, consider a map $P:[a,b]\rightarrow \mathbf{C}_{n\times n}$, which is continuously differentiable. Then $P(x)$ is an indefinite product integral if for a given $A(s)$,

$$P(x)=\prod_{a}^{x} e^{A(s)ds}P(a). \tag{11}$$

Next, we define an operation known as L operation that is like the logarithmic derivative operation on non-singular functions. Let

$$LP(x)=P'(x)P^{-1}(x), \tag{12}$$

where the prime indicates differentiation. Then, from Eq. (11) it follows that

$$(LP)(x)=A(x).$$

One of the byproducts of this operation is that

$$L(PQ)(x)=LP(x)+P(x)(LQ(x))P^{-1}(x). \tag{13}$$

The L operation is a crucial ingredient in establishing the analog of the fundamental theorem of calculus for product integrals. With the map P as defined above, this theorem states that

$$\prod_{a}^{x} e^{(LP)(s)ds}=P(x)P^{-1}(a). \tag{14}$$

From the results given above, it follows that P is a solution of the initial value problem,

$$P'(x)=(LP(x))P(x). \tag{15}$$

With the unique solution given by Eq. (11), this establishes the fundamental theorem of product integration. Just as in ordinary integration, the knowledge of simple product integrals can be used to evaluate more complicated product integrals. For example, one can prove the *sum rule* for product integrals:

$$\prod_{a}^{x} e^{[A(s)+B(s)]ds}=P(x)\prod_{a}^{x} e^{P^{-1}(s)B(s)P(s)ds}. \tag{16}$$

Finally, we state two other important properties of product integrals that will be used in the sequel. One is the *similarity theorem,* which states that

$$P(x)\prod_{a}^{x} e^{B(s)ds}P^{-1}(a)=\prod_{a}^{x} e^{[LP(s)+P(s)B(s)P^{-1}(s)]ds}. \tag{17}$$

The other property is differentiation with respect to a parameter. Let

$$P(x,y;\lambda)=\prod_{y}^{x} e^{A(s;\lambda)ds}, \tag{18}$$

where λ is a parameter. Then the differentiation rule with respect to this parameter is given by

$$\frac{\partial}{\partial\lambda}P(x,y;\lambda)=\int_{y}^{x} ds\, P(x,s;\lambda)\left(\frac{\partial}{\partial\lambda}A(s;\lambda)\right)P(s,y;\lambda). \tag{19}$$

III. THE NON-ABELIAN STOKES THEOREM

To provide the background for using the product integral formalism of Sec. II to prove the non-Abelian Stokes theorem, we begin with a statement of the problem as it arises in a physical context. Let M be an n-dimensional manifold representing the space–time (target space). Let A be a (connection) one-form on M. When M is a differentiable manifold, we can choose a local basis dx^{μ}, $\mu=1,...,n$, and express A in terms of its components:

$$A(x)=A_{\mu}(x)dx^{\mu}.$$

We take A to have values in the Lie algebra, or a representation thereof, of a Lie group. Then, with T_k, $k=1,...,m$, representing the generators of the Lie group, the components of A can be written as

$$A_{\mu}(x)=A_{\mu}^{k}(x)T_k.$$

With these preliminaries, we can express the path ordered phase factor of the non-Abelian gauge theories[15–18] in the form

$$W_{ab}(C)=\mathcal{P}e^{\int_e^b A},$$

where $\mathcal{P}$ indicates path ordering, and C is a path in M. When the path C is closed, the corresponding holonomy operator becomes

$$W(C)=\mathcal{P}e^{\oint A}. \tag{20}$$

The path C in M can be described in terms of an intrinsic parameter σ, so that for points x^{μ} of M, which lie on the path C, we have $x^{\mu}=x^{\mu}(\sigma)$. One can then write

$$A_{\mu}(x(\sigma))dx^{\mu}=A(\sigma)d\sigma,$$

where

$$A(\sigma)\equiv A^{\mu}(x(\sigma))\frac{dx^{\mu}(\sigma)}{d\sigma}.$$

It is the quantity $A(\sigma)$, and the variations thereof, which we will identify with the matrix-valued functions of the product integral formalism.

Let us next consider the loop operator. For simplicity, we assume that M has trivial first homology group with integer coefficients, i.e., $H_1(M,\mathbf{Z})=0$. This ensures that the loop may be taken to be the boundary of a two-dimensional surface Σ in M. More explicitly, we take the two-surface to be an orientable submanifold of M. It will be convenient to describe the properties of the two-surface in terms of its intrinsic parameters σ and τ or σ^a, $a=0,1$. So, for the points of the manifold M, which lie on Σ, we have $x=x(\sigma,\tau)$. The components of the one-form A on Σ can be obtained by means of the vielbeins (by the standard pullback construction):

$$v_a^{\mu}=\partial_a x^{\mu}(\sigma).$$

Thus, we get

$$A_a = v_a^\mu A_\mu .$$

The curvature two-form F of the connection A is given by

$$F = dA + A\wedge A = \tfrac{1}{2} F_{\mu\nu}\, dx^\mu \wedge dx^\nu .$$

The components of F on Σ can again be obtained by means of the vielbeins:

$$F_{ab} = v_a^\mu v_b^\nu F_{\mu\nu} .$$

We want to express the loop operator in terms of the product integral definition in a specific way. To achieve this, we begin with the definition of the path ordered phase factor in terms of a product integral. Consider the continuous map $A:[s_0,s_1]\rightarrow \mathbf{C}_{n\times n}$, where $[s_0,s_1]$ is a real interval. Then, we define the non-Abelian phase factor given above in terms of a product integral as follows:

$$\mathcal{P} e^{\int_{s_0}^{s_1} A(s)ds} \equiv \prod_{s_0}^{s_1} e^{A(s)ds}.$$

In particular, anticipating that we will identify the closed path C over which the Wilson loop is defined with the boundary of a two-surface, it is convenient to work from the beginning with the matrix-valued functions $A(\sigma,\tau)$. This means that our expression for the path ordered phase factor will depend on a parameter. That is, let

$$A:[\sigma_0,\sigma_1]\times[\tau_0,\tau_1]\rightarrow \mathbf{C}_{n\times n}, \tag{21}$$

where $[\sigma_0,\sigma_1]$ and $[\tau_0,\tau_1]$ are real intervals on the two surface Σ and hence in M. Then, we define the path ordered phase factor,

$$P(\sigma,\sigma_0;\tau) = \prod_{\sigma_0}^{\sigma} e^{A_1(\sigma';\tau)d\sigma'} \equiv \mathcal{P} e^{\int_{\sigma_0}^{\sigma} A_1(\sigma';\tau)d\sigma'}. \tag{22}$$

In this expression, $\mathcal{P}$ indicates path ordering with respect to σ as defined by the product integral, while τ is a parameter. To be able to describe such an operator for a closed path, we similarly define the path-dependent operator,

$$Q(\sigma;\tau,\tau_0) = \prod_{\tau_0}^{\tau} e^{A_0(\sigma;\tau')d\tau'} \equiv \mathcal{P} e^{\int_{\tau_0}^{\tau} A_0(\sigma;\tau')d\tau'}. \tag{23}$$

In this case, the path ordering is with respect to τ, and σ is a parameter.

To prove the non-Abelian version of the Stokes theorem, we want to make use of product integration techniques to express the holonomy loop operator as an integral over a two-dimensional surface bounded by the corresponding loop. In terms of the intrinsic coordinates of such a surface, we can write this loop operator in the form

$$W(C) = \mathcal{P} e^{\oint A_\alpha\, d\sigma^\alpha}, \tag{24}$$

where, as mentioned above,

$$\sigma^a = (\tau,\sigma); \quad a = (0,1). \tag{25}$$

The expression for the loop operator depends only on the homotopy class of paths in M to which the closed path C belongs. We can, therefore, parametrize the path C in any convenient manner consistent with its homotopy class. In particular, we can break up the path into piecewise continuous segments along which either σ or τ remains constant. The composition rule for product integrals given by Eq. (8) ensures that this breakup of the closed loop into a number of segments does not depend on the intermediate points on the closed path, which are used for this purpose. So, we write the closed loop operator as

$$W = W_4 W_3 W_2 W_1 . \tag{26}$$

In this expression, W_k, $k=1,...,4$, are Wilson lines such that $\tau=$const along W_1 and W_3, and $\sigma=$const along W_2 and W_4. We emphasize that the expressions $\sigma=$const and $\tau=$const represent arbitrary curves.

To see the advantage of parametrizing the closed path in this manner, consider the exponent of Eq. (24):

$$A_a \, d\sigma^a = A_0 \, d\tau + A_1 \, d\sigma. \tag{27}$$

Along each segment, one or the other of the terms on the right-hand side vanishes. For example, along the segment $[\sigma_0, \sigma]$, we have $\tau' = \tau_0 =$const. Recalling Eqs. (22) and (23), we get for the segments W_1 and W_2, respectively,

$$W_1 = \prod_{\sigma_0}^{\sigma} e^{A_1(\sigma';\tau_0)d\sigma'} \equiv \mathcal{P} e^{\int_{\sigma_0}^{\sigma} A_1(\sigma';\tau_0)d\sigma'} = P(\sigma,\sigma_0;\tau_0), \tag{28}$$

and

$$W_2 = \prod_{\tau_0}^{\tau} e^{A_0(\sigma;\tau')d\tau'} \equiv \mathcal{P} e^{\int_{\tau_0}^{\tau} A_0(\sigma;\tau')d\tau'} = Q(\sigma;\tau,\tau_0). \tag{29}$$

When the two-surface Σ requires more than one coordinate patch to cover it, the connections in different coordinate patches must be related to each other in their overlap region by transition functions.[16] Then, the decomposition given in Eq. (26) must be suitably augmented to take this complication into account. The product integral representation of the path ordered phase factor and the composition rule for product integrals given by Eq. (8) will still make it possible to describe the corresponding loop operator as a composite product integral. For definiteness, we will confine ourselves to the representation given by Eq. (26).

It is convenient for later purposes to define two composite Wilson line operators U and T according to

$$U(\sigma,\tau) = Q(\sigma;\tau,\tau_0) P(\sigma,\sigma_0;\tau), \tag{30}$$

$$T(\sigma;\tau) = P(\sigma,\sigma_0;\tau) Q(\sigma_0;\tau,\tau_0). \tag{31}$$

Using the first of these, we have

$$W_2 W_1 = U(\sigma,\tau). \tag{32}$$

Similarly, we have, for W_3 and W_4,

$$W_3 = P^{-1}(\sigma,\sigma_0;\tau), \tag{33}$$

and

$$W_4 = Q^{-1}(\sigma_0;\tau,\tau_0). \tag{34}$$

From Eq. (31), it follows that

$$W_4 W_3 = T^{-1}(\sigma,\tau). \tag{35}$$

Appealing again to Eq. (8) for the composition of product integrals, it is clear that this expression for the Wilson loop operator is independent of the choice of the point (σ,τ). In terms of the quantities T and U, the closed loop operator will take the compact form

$$W=T^{-1}(\sigma;\tau)U(\sigma;\tau). \tag{36}$$

As a first step in the proof of the non-Abelian Stokes theorem, we obtain the action of the L-derivative operator on W:

$$L_\tau W=L_\tau[T^{-1}(\sigma,\tau)Q(\sigma;\tau,\tau_0)P(\sigma,\sigma_0;\tau_0)]. \tag{37}$$

Using the definition of the L operation given by Eq. (12), noting that $P(\sigma,\sigma_0;\tau_0)$ is independent of τ, and carrying out the L operations on the right-hand side (rhs), we get

$$\begin{aligned}L_\tau W=L_\tau T^{-1}(\sigma,\tau)+T^{-1}(\sigma,\tau)[L_\tau Q(\sigma;\tau,\tau_0)\\ +Q(\sigma;\tau,\tau_0)(L_\tau P(\sigma,\sigma_0;\tau_0))Q^{-1}(\sigma;\tau,\tau_0)]T(\sigma,\tau).\end{aligned} \tag{38}$$

Simplifying this expression by means of Eqs. (12) and (13), we end up with

$$L_\tau W=T^{-1}(\sigma,\tau)[A_0(\sigma,\tau)-L_\tau T(\sigma,\tau)]T(\sigma,\tau). \tag{39}$$

Next, we prove the analog of Eq. (14), which applies to an elementary product integral, for the composite loop operator defined by Eqs. (26) and (36).

Theorem 1: *The loop operator given by Eq. (36) can be expressed in the form*

$$W=\prod_{\tau_0}^{\tau} e^{T^{-1}(\sigma,\tau')[A_0(\sigma,\tau')-L_\tau T(\sigma,\tau')]T(\sigma,\tau')d\tau'}. \tag{40}$$

To prove this theorem, first we note from the definition of the L operation that the right-hand side (rhs) of this equation can be written as

$$\text{rhs}=\prod_{\tau_0}^{\tau} e^{[T^{-1}(\sigma;\tau')A_0(\sigma;\tau')T(\sigma;\tau')-T^{-1}(\sigma;\tau')(\partial/\partial\tau')T(\sigma;\tau')]d\tau'}. \tag{41}$$

Noting that $-T^{-1}\,\partial_\tau T=L_\tau T$, we can use the similarity theorem given by Eq. (17) to obtain

$$\text{rhs}=T^{-1}(\sigma;\tau)\prod_{\tau_0}^{\tau} e^{A_0(\sigma;\tau')d\tau'}T(\sigma;\tau_0). \tag{42}$$

Moreover, making use of the defining Eq. (23), we get

$$\text{rhs}=T^{-1}(\sigma;\tau)Q(\sigma;\tau,\tau_0)P(\sigma,\sigma_0;\tau_0)Q(\sigma;\tau_0,\tau_0)=T^{-1}(\sigma;\tau)U(\sigma;\tau). \tag{43}$$

The last line is clearly the expression for W given by Eq. (36).

Finally, we want to express the quantity W in yet another form, which we state as the following.

Theorem 2: *The loop operator defined in Eq. (36) can be expressed as a surface integral of the field strength:*

$$W=\prod_{\tau_0}^{\tau} e^{\int_{\sigma_0}^{\sigma}T^{-1}(\sigma';\tau')F_{01}(\sigma';\tau')T(\sigma';\tau')d\sigma'\,d\tau'}, \tag{44}$$

where F_{01} is the 0–1 component of the non-Abelian field strength.

To prove this theorem, we note that

$$\frac{\partial}{\partial\sigma}[T^{-1}(\sigma,\tau)A_0(\sigma,\tau)T(\sigma,\tau)]=T^{-1}(\sigma,\tau)[\partial_\sigma A_0(\sigma,\tau)+[A_0(\sigma,\tau),A_1(\sigma,\tau)]]T(\sigma,\tau). \tag{45}$$

Moreover,

$$\frac{\partial}{\partial\sigma}\{T^{-1}(\sigma,\tau)(L_\tau T(\sigma,\tau))T(\sigma,\tau)\}=T^{-1}(\sigma,\tau)\partial_\tau A_1(\sigma,\tau)T(\sigma,\tau). \tag{46}$$

It then follows that

$$\begin{aligned}\frac{\partial}{\partial\sigma}(\{T^{-1}(\sigma,\tau)[A_0(\sigma,\tau)-L_\tau T(\sigma,\tau)]\}T(\sigma,\tau))\\ &=T^{-1}(\sigma,\tau)\left[\frac{\partial}{\partial\sigma}A_0(\sigma,\tau)-\frac{\partial}{\partial\tau}A_1(\sigma,\tau)+[A_0(\sigma,\tau),A_1(\sigma,\tau)]\right]T(\sigma,\tau)\\ &=T^{-1}(\sigma,\tau)F_{01}(\sigma,\tau)T(\sigma,\tau).\end{aligned} \tag{47}$$

The last step follows from the definition of the field strength in terms of the connection given above:

$$F_{01}=\frac{\partial}{\partial\sigma}A_0(\sigma,\tau)-\frac{\partial}{\partial\tau}A_1(\sigma,\tau)+[A_0(\sigma,\tau),A_1(\sigma,\tau)]. \tag{48}$$

Integrating Eq. (47) with respect to σ, we get

$$T^{-1}(\sigma,\tau)[A_0(\sigma,\tau)-L_\tau T(\sigma,\tau)]T(\sigma,\tau)=\int_{\sigma_0}^{\sigma}T^{-1}(\sigma';\tau')F_{01}(\sigma';\tau')T(\sigma';\tau')d\sigma'\,d\tau'. \tag{49}$$

We thus arrive at the surface integral representation of the loop operator:[19]

$$W=\prod_{\tau_0}^{\tau}e^{\int_{\sigma_0}^{\sigma}T^{-1}(\sigma';\tau')F_{01}(\sigma';\tau')T(\sigma';\tau')d\sigma'\,d\tau'}. \tag{50}$$

We note that in this expression the ordering of the operators is defined with respect to τ, whereas σ is a parameter. Recalling the antisymmetry of the components of the field strength, we can rewrite this expression in terms of path ordered exponentials familiar from the physics literature:

$$W=\mathcal{P}_\tau e^{(1/2)\int_\Sigma d\sigma^{ab}\,T^{-1}(\sigma;\tau)F_{ab}(\sigma;\tau)T(\sigma;\tau)}, \tag{51}$$

where $d\sigma^{ab}$ is the area element of the two-surface. Despite appearances, it must be remembered that σ and τ play very different roles in this expression.

IV. A SECOND PROOF

To illustrate the power and the flexibility of the product integral formalism, we give here a variant of the previous proof for the non-Abelian Stokes theorem. This time the proof makes essential use of the nontrivial relation (19) for product integrals. We start with the form of W given in Eq. (36) and take its derivatives with respect to τ:

$$\frac{\partial W}{\partial \tau} = \partial_\tau Q^{-1}(\sigma_0;\tau,\tau_0)P^{-1}(\sigma,\sigma_0;\tau)Q(\sigma;\tau,\tau_0)P(\sigma,\sigma_0;\tau_0)$$
$$+Q^{-1}(\sigma_0;\tau,\tau_0)\partial_\tau P^{-1}(\sigma,\sigma_0;\tau)Q(\sigma;\tau,\tau_0)P(\sigma,\sigma_0;\tau_0)$$
$$+Q^{-1}(\sigma_0;\tau,\tau_0)P^{-1}(\sigma,\sigma_0;\tau)\partial_\tau Q(\sigma;\tau,\tau_0)P(\sigma,\sigma_0;\tau_0). \tag{52}$$

Here, we have made use of the fact that $P(\sigma,\sigma_0;\tau_0)$ is independent of τ. Applying Eq. (12) to W and using Eq. (31), we get

$$L_\tau W = \frac{\partial W}{\partial \tau} W^{-1} = T^{-1}(\sigma;\tau)[A_0(\sigma;\tau) - P(\sigma,\sigma_0;\tau)A_0(\sigma_0;\tau)P^{-1}(\sigma,\sigma_0;\tau)$$
$$-\partial_\tau P(\sigma,\sigma_0;\tau)P^{-1}(\sigma,\sigma_0;\tau)]T(\sigma;\tau). \tag{53}$$

Now we can use Eq. (19) to evaluate the derivative of the product integral with respect to the parameter τ:

$$\partial_\tau P(\sigma,\sigma_0;\tau) = \int_{\sigma_0}^{\sigma} d\sigma'\, P(\sigma,\sigma';\tau)\partial_\tau A_1(\sigma';\tau)P(\sigma',\sigma_0;\tau). \tag{54}$$

Then, after some simple manipulations using the defining equations for the various terms in Eq. (53), we get, setting $P(\sigma,\sigma_0;\tau) = P(\tau)$,

$$T^{-1}(\sigma;\tau)\partial_\tau P(\tau)P^{-1}(\tau)T(\sigma;\tau) = \int_{\sigma_0}^{\sigma} d\sigma'\, T^{-1}(\sigma';\tau)\partial_\tau A_1(\sigma';\tau)T(\sigma';\tau). \tag{55}$$

Using Eq. (9) and the fact that $P(\sigma_0,\sigma_0;\tau) = 1$, we can write the rest of Eq. (53) as an integral too:

$$T^{-1}(\sigma;\tau)[A_0(\sigma;\tau) - P(\sigma,\sigma_0;\tau)A_0(\sigma_0;\tau)P^{-1}(\sigma,\sigma_0;\tau)]T(\sigma;\tau)$$
$$= Q^{-1}(\sigma_0;\tau,\tau_0)[P^{-1}(\sigma,\sigma_0;\tau)A_0(\sigma;\tau)P(\sigma,\sigma_0;\tau) - A_0(\sigma_0)]Q(\sigma_0;\tau,\tau_0)$$
$$= \int_{\sigma_0}^{\sigma} d\sigma'\, P^{-1}(\sigma',\sigma_0;\tau)(\partial_\tau A_0(\sigma',\tau) + [A_0(\sigma',\tau), A1(\sigma',\tau)])P(\sigma',\sigma_0;\tau). \tag{56}$$

Combining Eqs. (53), (55), and (56), we obtain

$$L_\tau W = \frac{\partial W}{\partial \tau} W^{-1} = \int_{\sigma_0}^{\sigma} d\sigma'\, T^{-1}(\sigma',\tau)F_{01}(\sigma',\tau)T(\sigma',\tau). \tag{57}$$

Finally, recalling Eq. (14), we are immediately led to Eq. (50), which was obtained by the previous method of proof.

There are two reasons for the relative simplicity of this proof over the one that was given in the previous section. One is due to the use of differentiation with respect to a parameter according to Eq. (19). The other is due to the use of Eq. (14) for the composite operator W. In the first proof, the use of this theorem for W was not assumed. Its justification for using it in the second proof lies in the composition law for product integrals given by Eq. (8).

V. GAUGE COVARIANCE OF THE RESULT

As a consistency check, we must show that the surface representation given by Eq. (50) is gauge covariant. To this end, it will be recalled that under a gauge transformation, the components of the connection, i.e., the gauge potentials, transform according to[20]

$$A_\mu(x)\rightarrow g(x)A_\mu(x)g^{-1}(x)-g(x)\partial_\mu g(x)^{-1}. \tag{58}$$

The components of the field strength (curvature) transform covariantly:

$$F_{\mu\nu}(x)\rightarrow g(x)F_{\mu\nu}(x)g^{-1}(x). \tag{59}$$

From these, it follows that[20]

$$\mathcal{P}e^{\int_a^b A_\mu(x)dx^\mu}\rightarrow g(b)(\mathcal{P}e^{\int_a^b A_\mu(x)dx^\mu})g^{-1}(a). \tag{60}$$

Equivalently, from its definition (22) in terms of product integrals, it is easy to show that the gauge transform of the quantity $P(\sigma,\sigma_0;\tau)$ has the form

$$g(\sigma;\tau)g^{-1}(\sigma_0;\tau)\prod_{\sigma_0}^{\sigma} e^{g(\sigma_0;\tau)A_1(\sigma';\tau)g^{-1}(\sigma_0;\tau)}. \tag{61}$$

To show the gauge covariance of the surface representation, we need to know how the operator $T(\sigma,\tau)$ transforms under gauge transformations. To this end, we note that the Wilson line $Q(\sigma;\tau,\tau_0)$ given by Eq. (23) transforms as

$$Q(\sigma;\tau,\tau_0)=\prod_{\tau_0}^{\tau} e^{A_0(\sigma;\tau')d\tau'}\rightarrow g(\sigma;\tau)Q(\sigma;\tau,\tau_0)g^{-1}(\sigma;\tau_0). \tag{62}$$

Then, the gauge transform of the composite operator $T(\sigma,\tau)$ given by Eq. (31) follows immediately:

$$T(\sigma;\tau)=P(\sigma,\sigma_0;\tau)Q(\sigma_0;\tau,\tau_0)\rightarrow g(\sigma;\tau)T(\sigma;\tau)g^{-1}(\sigma_0;\tau_0). \tag{63}$$

From the above results, it is straightforward to show that the surface integral representation of Wilson loop transforms as

$$W\rightarrow\prod_{\tau_0}^{\tau} e^{g(\sigma_0;\tau_0)(\int_{\sigma_0}^{\sigma}T^{-1}(\sigma';\tau')F_{01}(\sigma';\tau')T(\sigma';\tau')dt')g^{-1}(\sigma_0;\tau_0)}. \tag{64}$$

It follows from the composition rule (8) that the constant factors in the exponent factorize, so that under gauge transformations the surface representation of transforms covariantly, i.e.,

$$W\rightarrow g(\sigma_0;\tau_0)\prod_{\tau_0}^{\tau} e^{\int_{\sigma_0}^{\sigma}T^{-1}(\sigma';\tau')F_{01}(\sigma';\tau')T(\sigma';\tau')dt'}g^{-1}(\sigma_0;\tau_0). \tag{65}$$

We view this result as a nontrivial confirmation of our proofs.

VI. CONCLUDING REMARKS

We have provided two proofs of the non-Abelian Stokes theorem using the product integral method. An immediate question that comes to mind is whether there is a supersymmetric generalization of this theorem. Given the important developments in supersymmetric gauge theories in recent years, this question is not merely of academic interest. To explore this possibility using the product integral method, it is necessary to generalize this method to encompass Grassmann-valued operators. It turns out that such a generalization is indeed possible.[21] Further developments of this subject will be reported in a forthcoming work.

ACKNOWLEDGMENTS

This work was supported in part by the Department of Energy under Contract No. DOE-FGO2-84ER40153. The hospitality of the Aspen Center for Physics in the Summer of 1998 is also gratefully acknowledged. We are also grateful to Dr. M. Awada for valuable input at the initial stages of this work.

[1] J. D. Dollard and C. N. Friedman, Product Integration (Addison-Wesley, New York, 1979).
[2] M. B. Halpern, ''Field strength and dual variable formulations of gauge theory,'' Phys. Rev. D **19**, 517 (1979).
[3] Y. Aref'eva, ''Non-Abelian Stokes formula,'' Theor. Math. Phys. **43**, 353 (1980).
[4] P. M. Fishbane, S. Gasiorowicz, and P. Kaus, ''Stokes' Theorems for non-Abelian fields,'' Phys. Rev. D **24**, 2324 (1981).
[5] N. E. Bralic, ''Exact computation of loop averages in two-dimensional Yang–Mills theory,'' Phys. Rev. D **22**, 3090 (1980).
[6] L. Diosi, ''Comments on the non-Abelian Stokes theorem,'' Phys. Rev. D **27**, 2552 (1983).
[7] Y. A. Simonov, ''Cluster expansion, non-Abelian Stokes theorem, and magnetic monopoles,'' Sov. J. Nucl. Phys. **50**, 134 (1989); Y. A. Simonov, ''Topological charges and convergence of the cluster expansion,'' Sov. J. Nucl. Phys. **50**, 310 (1989).
[8] B. Broda, in *Advanced Electromagnetism*, edited by T. Barrett and D. Grimes (World Scientific, Singapore, 1995).
[9] M. Hirayama and S. Matsubara, ''Stokes theorem for loop variables of non-Abelian gauge field,'' Prog. Theor. Phys. **99**, 691 (1998); M. Hirayama, M. Kanno, M. Ueno, and H. Yamakoshi, ''Non-Abelian Stokes theorem for loop variables associated with nontrivial loops,'' hep-th/9806098.
[10] D. Diakonov and V. Petrov, ''Non-Abelian Stokes theorem and quark–monopole interaction,'' hep-th/9606104.
[11] F. A. Lunev, ''Pure bosonic worldline path integral representation for fermionic determinants, non-Abelian Stokes theorem, and quasiclassical approximation in QCD,'' Nucl. Phys. B **494**, 433 (1997).
[12] E. Alvarez, C. Gomez, and T. Ortin, ''String representation of Wilson loops,'' hep-th/9806075.
[13] A. M. Polyakov, ''String theory and quark confinement,'' Nucl. Phys. (Proc. Suppl.) **68**, 1 (1997); A. M. Polyakov, ''The wall of the cave,'' hep-th/9809057.
[14] L. Schlessinger, Math. Ann. **99**, 413 (1927).
[15] S. Mandelstam, ''Gauge independent formulation of electrodynamics,'' Ann. Phys. (N.Y.) **19**, 1 (1962).
[16] T. T. Wu and C. N. Yang, ''Concept Of non-integrable phase factors and global formulation of gauge fields,'' Phys. Rev. D **12**, 3845 (1975).
[17] K. G. Wilson, ''Confinement of quarks,'' Phys. Rev. D **10**, 2445 (1974).
[18] A. M. Polyakov, ''Compact gauge fields and the infrared catastrophe,'' Phys. Lett. B **59**, 82 (1975).
[19] A preliminary version of this work was presented at the *International Conference on Orbis Scientiae 1998*, Fort Lauderdale, Florida, 17–21 December 1998.
[20] Our conventions follow closely M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, New York, 1995).
[21] R. L. Karp and F. Mansouri, University of Cincinnati preprint UCTP102.99, and in preparation.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Functional determinants from Wronski Green functions

H. Kleinert[a] and A. Chervyakov[b]
Freie Universität Berlin, Institut für Theoretische Physik, Arnimallee14, D-14195 Berlin, Germany

(Received 24 March 1998; accepted for publication 17 June 1998)

A general technique is developed for calculating functional determinants of second-order differential operators with Dirichlet, periodic, and antiperiodic boundary conditions, without the knowledge of spectral properties. As an example, we give explicit formulas for a harmonic oscillator with an arbitrary time-dependent frequency, where our result is a generalization of the Gel'fand–Yaglom famous formula for Dirichlet boundary conditions. Our technique is based on the Wronski's construction of Green functions, which does not require spectral knowledge. Our final formula expresses the ratios of functional determinants in terms of an ordinary 2×2 determinant of a constant matrix constructed from two linearly independent solutions of the homogeneous differential equations associated with second-order differential operators. For ratios of determinants encountered in semiclassical fluctuations around a classical solution, the result can further be expressed in terms of the classical solution. Special properties of operators with a zero mode are exhibited. © *1999 American Institute of Physics.* [S0022-2488(99)02609-2]

I. INTRODUCTION

Gaussian path integrals appear in many physical problems, for instance, in all semiclassical approximations of fluctuating systems. They typically require the evaluation of a functional determinant of a second-order differential operator.[1] For Dirichlet boundary conditions, a first general solution of this problem was given by Gel'fand and Yaglom,[2] based on the lattice approximation to path integrals in the continuum limit. Their result was expressed in terms of a simple differential equation for the functional determinant. Unfortunately, the Gel'fand–Yaglom method becomes rather complicated for periodic and antiperiodic boundary condition relevant in quantum statistics (see Sec. 2.12 in Ref. 1). In recent work,[3,4] functional determinants of second-order differential operators were calculated via a reduction to the simpler first-order formalism. Divergences were removed by zeta-function regularization,[5] which has the disadvantage of a physical quantity depending unnecessarily on the analytic properties of generalized zeta functions. Moreover, the first-order formalism makes the treatment of zero modes of the second-order differential operator quite cumbersome.[6]

In this paper we present a systematic method for finding *ratio* functional determinants of second-order differential operators with Dirichlet, periodic, and antiperiodic boundary conditions. By focusing our attention upon ratios instead of the determinants themselves, we avoid the need of regularization. The main virtue of our method is that it takes advantage of Wronski's elegant construction method for Green functions. This permits us to reduce the functional determinant of a second-order differential operator to an ordinary determinant of a constant 2×2 matrix formed from solutions of the homogeneous differential equation associated with the differential operator.

For semiclassical fluctuations around a classical solution, our final result can be expressed entirely in terms of a classical trajectory. Furthermore, for fluctuation operators with a zero mode, a case encountered in many semiclassical calculations, we find a universal expression for the ratio of determinants without the zero mode.

[a]Electronic mail: kleinert@physik.fu-berlin.de
[b]On leave from LCTA, JINR, Dubna, Russia.

II. BASIC RELATIONS

The typical fluctuation action arising in semiclassical approximations has a quadratic Lagrangian of the form

$$L_{\rm fl}=\frac{M}{2}[\dot{x}^2-\Omega^2(t)x^2]. \tag{1}$$

Physically, this Lagrangian describes a harmonic oscillator with a time-dependent frequency $\Omega(t)$. The path integral for such a system was studied in several papers.[7,8] For such an oscillator, both the quantum mechanical propagator and the thermal partition function contain a phase factor $\exp[i\mathcal{A}_{\rm cl}(x)]$ and are multiplied by a pre-exponential fluctuation factor proportional to

$$F(t_b,t_a)\sim\left(\frac{\operatorname{Det}K_1}{\operatorname{Det}\tilde{K}}\right)^{-1/2}, \tag{2}$$

where $K_1=-\partial_t^2-\Omega^2(t)\equiv K_0-\Omega^2(t)$ is the kernel of the second variation of action $\mathcal{A}[x])$ along the classical path $x_{\rm cl}(t)$. The linear operator K_1 acts on the space of twice differentiable functions $y(t)=\delta x(t)$ on an interval $t\in[t_a,t_b]$ with appropriate boundary conditions. These are Dirichlet boundary conditions $y(t_a)=y(t_b)=0$ in the quantum-mechanical case, and periodic (antiperiodic) $y(t_b)=\pm y(t_a),\dot{y}(t_b)=\pm\dot{y}(t_a)$ in the quantum statistical case. In these two cases the operator $\tilde{K}$ may be chosen as K_0 or $K_0-\omega_0^2$, respectively, where ω_0 is a time-independent oscillator frequency. The ratio of determinants (2) arises naturally from the normalization of the path integral and is well defined.[1] Furthermore, this ratio has the Fredholm property

$$\frac{\operatorname{Det}K_1}{\operatorname{Det}\tilde{K}}=\operatorname{Det}\tilde{K}^{-1}K_1, \tag{3}$$

with a multiplicative anomaly being equal to unity.[9] Since the operator $\tilde{K}^{-1}K_1$ is of the form $I+B$, with B an operator of the trace class, it has a well-defined determinant, even without any regularization.

To calculate $F(t_b,t_a)$, we introduce a one-parameter family of operators K_g depending linearly on the parameter g:$K_g=K_0-g\Omega^2(t)$, $0\leqslant g\leqslant 1$. The above property (3) allows us to make use of the well-known formula $\log\operatorname{Det}\tilde{K}^{-1}K_g=\operatorname{Tr}\log\tilde{K}^{-1}K_g$ to relate the g derivative of the logarithm of the ratio (2) to the trace of the Green function of the operator K_g as follows:

$$\partial_g\log\operatorname{Det}\tilde{K}^{-1}K_g=-\operatorname{Tr}\Omega^2(t)G_g(t,t'), \tag{4}$$

the Green function being defined by

$$G_g(t,t')=[-\partial_t^2-g\Omega^2(t)]^{-1}\delta(t-t'). \tag{5}$$

Another proof of Eq. (4) can be found in Ref. 3. Note that formula (4) holds for the operator K_g by itself rather than the ratio $K^{-1}K_g$, provided we regularize the divergent trace on the right-hand side with the help of a generalized zeta functions $\zeta_{K_g}(s)=\Sigma_i\lambda^{-s}$, where the sum runs over all eigenvalues of the operator K_g. The sum converges for sufficinetly large s and may be defined for smaller s by analytic continuation (see Ref. 5). The zeta function yields the functional determinant itself via the formula $\operatorname{Det}K_g=\exp[-\zeta'_{K_g}(0)]$.

By integrating (4), we obtain for the ratio of functional determinants (3):

$$\operatorname{Det}\tilde{K}^{-1}K_g=C\exp\left\{-\int_0^g dg'\int_{t_a}^{t_b}dt\,\Omega^2(t)G_{g'}(t,t)\right\}, \tag{6}$$

where $C = \mathrm{Det}\,\tilde{K}^{-1}K_0$ is a g-independent constant. This is our basic formula to be supplemented by an appropriate boundary condition to Eq. (5) for the Green function, as we shall now discuss in detail.

III. DIRICHLET BOUNDARY CONDITIONS

A general solution of Eq. (5) may be expressed in terms of advanced or retarded Green functions as follows:

$$G_g^-(t,t') = G_g^+(t',t) = \Theta_{tt'} \cdot f_g(t,t'), \tag{7}$$

where $\Theta_{tt'} = \Theta(t-t')$ is Heaviside's function and $f_g(t,t')$ is a combination,

$$f_g(t,t') = \frac{1}{W_g}[\eta_g(t)\xi_g(t') - \xi_g(t)\eta_g(t')], \tag{8}$$

of two linearly independent solutions $\eta_g(t)$ and $\xi_g(t)$ of the homogeneous equation,

$$[-\partial_t^2 - g\Omega^2(t)]h_g(t) = 0. \tag{9}$$

The constant W_g is the time-independent Wronski determinant $W_g = \eta_g\dot{\xi}_g - \dot{\eta}_g\xi_g$. The solution (7) is not unique since it leaves room for an additional general solution of the homogeneous equation (9) with an arbitrary coefficient. This freedom is removed by appropriate boundary conditions. Consider first the quantum mechanical case, which requires Dirichlet boundary conditions $y(t_b) = y(t_a) = 0$ for the eigenfunctions $y(t)$ of K_1, implying for the Green function the boundary conditions

$$\begin{aligned} G_g(t_a,t') &= 0, \quad t \leqslant t', \\ G_g(t',t_b) &= 0, \quad t' \leqslant t. \end{aligned} \tag{10}$$

The operator $\tilde{K}$ in the ratio (2) is equal to K_0, and the constant C in Eq. (6) is unity. After imposing (10), the Green function is uniquely given by Wronski's formula:

$$G_g(t,t') = \frac{\Theta_{tt'} f_g(t',t_a) f_g(t_b,t) + \Theta_{t't} f_g(t,t_a) f_g(t_b,t')}{f_g(t_a,t_b)}, \tag{11}$$

where

$$f_g(t_a,t_b) = \frac{\mathrm{Det}\,\Lambda_g}{W_g} \neq 0, \tag{12}$$

with Λ being a constant (2×2)-matrix,

$$\Lambda = \begin{pmatrix} \eta_a & \xi_a \\ \eta_b & \xi_b \end{pmatrix}, \tag{13}$$

formed from the solutions $\eta_g(t)$ and $\xi_g(t)$ at arbitrary $g \neq 1$. Note that these solutions are restricted only to the condition (12). The result is unique and well defined, assuming the absence of a zero mode $\xi(t)$ of the operator K_1 with Dirichlet boundary conditions $\xi_a = \xi_b = 0$. Such a mode would cause problems, since according to (10), the Wronski determinant W would vanish at the initial point, and thus for all t.

Excluding zero modes, we obtain from (8):

$$\mathrm{Tr}\,\Omega^2(t)G_g(t,t')=\frac{1}{f_g(t_a,t_b)}\int_{t_a}^{t_b}dt\,\Omega^2(t)f_g(t,t_a)f_g(t_b,t). \tag{14}$$

To perform the time integral on the right-hand side, we make use of the identity

$$\Omega^2(t)\xi(t)\eta(t)=\partial_t[\dot{\eta}_g(t)\partial_g\xi_g(t)-\eta_g(t)\partial_g\dot{\xi}_g(t)]. \tag{15}$$

This follows from Eq. (9) for $\eta_g(t)$, and an analogous equation for $\xi_{\tilde{g}}(t)$, after multiplying the first by $\xi_{\tilde{g}}(t)$ and the second by $\eta_g(t)$, and taking their difference. In the limit $\tilde{g}\to g$, we obtain (15) from the linear term in $\tilde{g}-g$. Inserting (15) into (14), we see that

$$\mathrm{Tr}\,\Omega^2(t)G_g(t,t')=-\partial_g\log\left(\frac{\mathrm{Det}\,\Lambda_g}{W_g}\right). \tag{16}$$

Substituting (16) into (6), we find

$$\mathrm{Det}\,K_0^{-1}K_1=\frac{\mathrm{Det}\,\Lambda_1}{W_1}\Big/\frac{\mathrm{Det}\,\Lambda_0}{W_0}, \tag{17}$$

where $\mathrm{Det}\,\Lambda_0/W_0=t_b-t_a$. In a time-sliced quantum mechanical path integral, where the time slices have the width ϵ, the determinant of $K_0=-\partial_t^2$ is finite and has the value[1]

$$\mathrm{Det}\,K_0=(t_b-t_a)/\epsilon, \tag{18}$$

so that we obtain

$$\mathrm{Det}\,K_1=D(t_b)/\epsilon, \tag{19}$$

where $D(t)=[\eta_1(t_a)\xi_1(t)-\eta_1(t)\xi_1(t_a)]/W_1$ is a solution of the Gel'fand–Yaglom initial value problem (see Sec. 2.7 in Ref. 1):

$$K_1D(t)=0;\quad D(t_a)=0,\quad \dot{D}(t_a)=1. \tag{20}$$

For a harmonic oscillator with a time-dependent frequency $\Omega(t)$, it is convenient to relate the set of two independent solutions $\eta_1(t)$ and $\xi_1(t)$ to the classical path $x_{\mathrm{cl}}(t)=x_a\xi(t)+x_b\eta(t)$ satisfying the end point conditions $x_{\mathrm{cl}}(t_a)=x_a$ and $x_{\mathrm{cl}}(t_b)=x_b$. Since this construction satisfies $\eta_a=\xi_b=0$, $\eta_b=\xi_a=1$, and $W=\dot{\xi}_b=-\dot{\eta}_a$, the explicit solution being

$$\xi(t)=\frac{\partial x_{\mathrm{cl}}(t)}{\partial x_a}=\frac{q(t)q_b\sin\omega_0(\tau_b-\tau)}{q_aq_b\sin\omega_0(\tau_b-\tau_a)},$$
$$\eta(t)=\frac{\partial x_{\mathrm{cl}}(t)}{\partial x_b}=\frac{q(t)q_a\sin\omega_0(\tau-\tau_a)}{q_aq_b\sin\omega_0(\tau_b-\tau_a)}. \tag{21}$$

They are parametrized by two functions $\tau(t)$ and $q(t)$ satisfying the constraint

$$\omega_0\dot{\tau}q^2=1, \tag{22}$$

where ω_0 is an arbitrary constant frequency. The function $q(t)$ satisfies the Ermakov–Pinney equation,[10]

$$\ddot{q}+\Omega^2(t)q-q^{-3}=0. \tag{23}$$

Inserting (21) into (17), we obtain for the harmonic oscillator with a time-dependent frequency $\Omega(t)$ the ratio of functional determinants,

$$\mathrm{Det}\, K_0^{-1} K_1 = \frac{q_a q_b \sin \omega_0(\tau_b - \tau_a)}{(t_b - t_a)}, \tag{24}$$

where subscripts a and b indicate evaluation at $t = t_a$ and $t = t_b$, respectively. We check this representation by expressing the right-hand side in terms of the classical action $\mathcal{A}_{\mathrm{cl}}(x)$. With the same normalization as in (18), this yields the well-known one-dimensional Van-Vleck formula,

$$\mathrm{Det}\, K_1 = -M[\partial^2 \mathcal{A}_{\mathrm{cl}}(x_a, x_b; t_b - t_a)/\partial x_a\, \partial x_b]^{-1}. \tag{25}$$

To conclude this section we note that the ratio (17) can easily be extended to the stochastic case, where the final position of the trajectory $x(t)$ remains unspecified. To this end we consider Eqs. (14) and (15) with a variable upper time $t' \geqslant t \geqslant t_a$. Then the eigenvalues of the operator $K_0^{-1} K_1$ become functions of t' with a phase factor produced by each passage through a focal point.

IV. PERIODIC AND ANTIPERIODIC BOUNDARY CONDITIONS

Consider now periodic (antiperiodic) boundary conditions $y(t_b) = \pm y(t_a)$, $\dot{y}(t_b) = \pm \dot{y}(t_a)$ for the eigenfunctions $y(t)$ of the operator K_1 and for the Green function $G^{p}_{a}(t,t')$:

$$\begin{aligned} G^{p}_{a}(t_b, t') &= \pm G^{p}_{a}(t_a, t'), \\ \dot{G}^{p}_{a}(t_b, t') &= \pm \dot{G}^{p}_{a}(t_a, t'), \end{aligned} \tag{26}$$

where $T = t_b - t_a$ is the period. In both cases, the frequency $\Omega(t)$ and Dirac's δ function in Eq. (5) are also assumed to be periodic (antiperiodic) with the same period. The general solution of Eq. (5) satisfying the boundary conditions (26) is constructed by adding to (7) an expression of the same type as before, using the same homogeneous solutions $\eta_g(t)$ and $\xi_g(t)$. The result has the form

$$G^{p}_{a\,g}(t,t') = G_g(t,t') \mp \frac{[f_g(t,t_a) \pm f_g(t_b,t)][f_g(t',t_a) \pm f_g(t_b,t')]}{\Delta^{p}_{a} \cdot f_g(t_a,t_b)}, \tag{27}$$

with the condition

$$\Delta^{p}_{a} = \frac{\mathrm{Det}\, \bar{\Lambda}^{p}_{a\,g}}{W_g} \neq 0, \tag{28}$$

where $\bar{\Lambda}^{p}_{a}$ are now the (2×2)-constant matrices,

$$\bar{\Lambda}^{p}_{a} = \begin{pmatrix} (\eta_b \mp \eta_a) & (\xi_b \mp \xi_a) \\ (\dot{\eta}_b \mp \dot{\eta}_a) & (\dot{\xi}_b \mp \dot{\xi}_a), \end{pmatrix}, \tag{29}$$

evaluated at $g \neq 1$. In analogy to Eq. (16) we now find the formula

$$\mathrm{Tr}\, \Omega^2(t) G^{p}_{a\,g}(t,t') = -\partial_g \log\left(\frac{\mathrm{Det}\, \bar{\Lambda}^{p}_{a\,g}}{W_g} \right). \tag{30}$$

Substituting this into (6) and setting $g=1$, we obtain the ratio of the functional determinants for periodic boundary conditions,

$$\mathrm{Det}\,\tilde{K}^{-1}K_1=\frac{\mathrm{Det}\,\Lambda_1^p}{W_1}\Bigg/ 4\sin^2\frac{\omega_0(t_b-t_a)}{2}. \tag{31}$$

Here $\mathrm{Det}\,\tilde{K}=\mathrm{Det}(-\partial_t^2-\omega_0^2)$ is the fluctuation determinant of the harmonic oscillator, which in the same normalization as in (18) is equal to

$$\mathrm{Det}\,\tilde{K}=4\sin^2\frac{\omega_0(t_b-t_a)}{2}, \tag{32}$$

and thus the formula

$$\mathrm{Det}\,K_1=\frac{(\eta_b-\eta_a)(\dot{\xi}_b-\dot{\xi}_a)-(\xi_b-\xi_a)(\dot{\eta}_b-\dot{\eta}_a)}{W}, \tag{33}$$

the right-hand side being evaluated at $g=1$. For antiperiodic boundary conditions, the analogous expressions are

$$\mathrm{Det}\,\tilde{K}_1^{-1}=\frac{\mathrm{Det}\,\Lambda_1^a}{W_1}\Bigg/ 4\cos^2\frac{\omega_0(t_b-t_a)}{2}, \tag{34}$$

$$\mathrm{Det}\,K_1=\frac{(\eta_b+\eta_a)(\dot{\xi}_b+\dot{\xi}_a)-(\xi_b+\xi_a)(\dot{\eta}_b+\dot{\eta}_a)}{W}. \tag{35}$$

For a harmonic oscillator with a time-dependent frequency $\Omega(t)$, we use again the representation (21) for $\xi(t)$ and $\eta(t)$ in terms of the functions $p(t)$ and $q(t)$, which, in addition to (22) and (23), have the following properties: the function $p(t)$ is periodic and even,

$$p(t+T)=p(t),\quad p(-t)=p(t), \tag{36}$$

so that $p_b=p_a$, whereas the function $q(t)$ satisfies

$$q(t+T)=q(t)+q_b,\quad q_a=0, \tag{37}$$

where $T\equiv(t_b-t_a)$. Inserting now the solutions (21) into (31) and (34), we find the ratio of functional determinants for a harmonic oscillator with a time-dependent frequency $\Omega(t)$ with periodic boundary conditions,

$$\mathrm{Det}\,\tilde{K}^{-1}K_1=4\sin^2\frac{\omega_0 q_b}{2}\Bigg/ 4\sin^2\frac{\omega_0 t}{2}, \tag{38}$$

and with antiperiodic boundary conditions,

$$\mathrm{Det}\,\tilde{K}^{-1}K_1=4\cos^2\frac{\omega_0 q_b}{2}\Bigg/ 4\cos^2\frac{\omega_0 t}{2}. \tag{39}$$

Only formula (24) for the Dirichlet boundary condition has been known in the literature (see Refs. 7 and 8). The periodic and antiperiodic formulas (38) and (39) are new, although they have had predecssors on the lattice.[11] Moreover, our new derivation has the advantage of employing only Wronski's simple construction method for Green functions. The general expressions for the functional determinants (19), (33), and (35) are form invariant under an arbitrary changes $(\eta,\xi)\to(\tilde{\eta},\tilde{\xi})$ of the basic set $\eta(t)$ and $\xi(t)$ of two independent solutions of the homogeneous equation (9).

V. TREATMENT OF ZERO MODE

Contrary to the case of a harmonic oscillator with a time-dependent frequency $\Omega(t)$, in which the operator K_1 has no zero mode, consider now the situation where such a zero mode is present. It is typically encountered for fluctuations around localized classical solutions $x_{\rm cl}(t)$ such as solitons or instantons,[1] where the frequency $\Omega(t)$ in the operator K_1 has the special form $\Omega^2(t) = V''(x_{\rm cl}(t))/M$. The fluctuation Lagrangian (1) arises from the second-order variation around $x(t)=x_{\rm cl}(t)$ of a Lagrangian $L=\dot{x}^2/2-V(x)$, which has no explicit time dependence, so that $x_{\rm cl}(t)$ and $x_{\rm cl}(t+t_0)$ are equally good solutions of $\ddot{x}=-V'(x)$, which makes $\dot{x}_{\rm cl}(t)$ a zero mode of K_1, caused by translation invariance along the time axis.[12]

For simplicity, we shall assume the presence of only a single zero mode, which we choose as one of two independent solutions of the homogeneous differential equation, say $\eta(t)$. For Dirichlet boundary conditions, we call this a Dirichlet zero mode with the properties

$$\eta_b=0, \quad \eta_a=0. \tag{40}$$

For periodic and antiperiodic boundary conditions, the zero mode satisfies

$$\eta_b\mp\eta_a=0, \quad \dot{\eta}_b\mp\dot{\eta}_a=0, \tag{41}$$

respectively. As pointed out earlier, the Wronski construction for evaluating ratios of functional determinants is not applicable here since the conditions (12) and (28) are violated as a consequence of (40) and (41). In order to enforce (12) and (28), we modify the boundary conditions (40) and (41) by a small regulator parameter $\epsilon>0$, and determine new independent solutions $\eta^\epsilon(t)$ and $\xi^\epsilon(t)$ with $\eta^\epsilon(t)\to\eta(t)$ and $\xi^\epsilon(t)\to\xi(t)$ for $\epsilon\to 0$. The specific form of regularized boundary conditions will be irrelevant, requiring only the condition for the regularized Wronskian,

$$W^\epsilon=\eta^\epsilon\dot{\xi}^\epsilon-\dot{\eta}^\epsilon\xi^\epsilon=W, \tag{42}$$

being independent of ϵ. One may imagine such a regularization as a simple shifting of limiting end points of the time interval (t_a,t_b), provided that the operator K_1 remains nonsingular. Since the conditions (12) and (28) are now satisfied, the Wronski construction and our modification for periodic and antiperiodic boundary conditions provides us directly with the regularized determinants $\mathrm{Det}\,K_1^\epsilon$, which tend to small values in the limit $\epsilon\to 0$. In terms of $\eta^\epsilon(t)$ and $\xi^\epsilon(t)$, these determinants are given explicitly by the expressions (19), (33), and (35) for Dirichlet, periodic, and antiperiodic boundary conditions, respectively. We now remove the zero mode from the determinant using the standard methods.[13] The determinant without such a zero mode is defined by

$$\mathrm{Det}'\,K_1=\lim_{\varepsilon\to 0}\frac{\mathrm{Det}\,K_1^\varepsilon}{\delta\lambda^\varepsilon}, \tag{43}$$

where $\delta\lambda^\varepsilon$ is a small eigenvalue associated with the eigenfunction $\eta^\varepsilon(t)+\delta\eta^\varepsilon(t)$,

$$K_1\,\delta\eta^\varepsilon=\delta\lambda^\varepsilon\,\eta^\varepsilon, \tag{44}$$

with the limit $\delta\lambda^\varepsilon\to 0$ for $\varepsilon\to 0$. To first order in ε, it follows from (44) that

$$(\delta\eta^\varepsilon\,\dot{\eta}^\varepsilon-\delta\dot{\eta}^\varepsilon\,\eta^\varepsilon)\big|_{t_a}^{t_b}\approx\delta\lambda^\varepsilon\int_{t_a}^{t_b}dt\,\eta^2(t)\equiv\delta\lambda^\varepsilon\langle\eta|\eta\rangle. \tag{45}$$

Substituting (19) and (45) into (43), we obtain the functional determinant without a zero mode for the operator K_1 with Dirichlet boundary conditions,

$$\mathrm{Det}'\, K_1 = -\lim_{\varepsilon\to 0} \frac{\xi_a^\varepsilon \xi_b^\varepsilon}{W^2} \langle \eta | \eta \rangle = -\frac{\langle \eta | \eta \rangle}{\dot{\eta}_b \dot{\eta}_a}. \tag{46}$$

For the operator K_1 with periodic and antiperiodic boundary conditions, the analogous calculation extracting the zero mode from the functional determinant gives

$$\mathrm{Det}'\, K_1 = \pm \lim_{\varepsilon\to 0} \frac{(\xi_a^\varepsilon \dot{\xi}_b^\varepsilon - \xi_b^\varepsilon \dot{\xi}_a^\varepsilon)}{W^2} \langle \eta | \eta \rangle = -\frac{(\xi_b \mp \xi_a)\langle \eta | \eta \rangle}{\eta_b W}. \tag{47}$$

These formulas are useful for semiclassical calculations of path integrals processing nontrivial classical solutions such as solitons or instantons.[1] Note that our expressions (46) and (47) for the functional determinants are independent of the specific choice of regularization.

Note added. Further useful formulas, in particular an extremely efficient one for the fluctuation determinant of a solution of the equation of motion, can be found in recent detailed paper by H. Kleinert and A. Chervyakov, ''Simple explicit formulas for Gaussian path integrals with time-dependent frequencies,'' quantph/9803016, Phys. Lett. **A245**, 345 (1998).

ACKNOWLEDGMENTS

We thank A. Pelster for discussions. The work was supported by Deutsche Forschungsgemeinschaft under Grant No. Kl 256/28-1 and the Russian Foundation of Fundamental Research under Grant No. 97-01-00745.

[1] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, and Polymer Physics*, 2nd ed. (World Scientific, Singapore, 1995).
[2] I. M. Gel'fand and A. M. Yaglom, J. Math. Phys. **1**, 48 (1960).
[3] S. Levit and U. Smilansky, Proc. Am. Math. Soc. **65**, 299 (1977).
[4] M. Lesch and J. Tolksdorf, dg-ga/9707022v1, 1997; R. Forman, Invent. Math. **88**, 447 (1987); Commun. Math. Phys. **147**, 485 (1992); D. Burghelea, L. Friedlander, and T. Kappeler, Commun. Math. Phys. **138**, 1 (1991); Proc. Am. Math. Soc. **123**, 3027 (1995).
[5] D. B. Ray, Adv. Math. **4**, 109 (1970); D. B. Ray and I. M. Singer, *ibid.* **7**, 145 (1971); Ann. Math. **98**, 154 (1973).
[6] A. J. McKane and M. B. Tarlie, J. Phys. A **28**, 6931 (1995).
[7] B. K. Cheng, J. Math. Phys. **25**, 1804 (1984); **27**, 217 (1986).
[8] D. C. Khandekar and S. V. Lawande, Phys. Rep. **137**, 115 (1986).
[9] M. Kontsevich and S. Vishik, hep-th/945035, 1994; D. C. Khandekar and S. V. Lawande, J. Math. Phys. **20**, 1870 (1979); J. Rezende, J. Math. Phys. **25**, 3264 (1984).
[10] A. K. Common, E. Hesameddini, and M. Musette, J. Phys. A **29**, 6343 (1996).
[11] See Sec. 2.12 of Ref. 1.
[12] See Sec. 17.3 of Ref. 1.
[13] See Sec. 17.5 in 1, or Ref. 6.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

The inhomogeneous quantum groups from differential calculi with classical dimension

M. Lagraa and N. Touhami
Laboratoire de Physique Théorique, Université d'Oran Es-Sénia, 31100 Algérie, France

(Received 19 November 1997; accepted for publication 12 March 1999)

From the bicovariant first-order differential calculus on inhomogeneous Hopf algebra $\mathcal{B}$ we construct the set of right-invariant Maurer–Cartan one-forms considered as a right-invariant basis of a bicovariant $\mathcal{B}$-bimodule over which we develop the Woronowicz general theory of differential calculus on quantum groups. In this formalism, we introduce suitable functionals on $\mathcal{B}$ which control the inhomogeneous commutation rules. In particular, we find that the homogeneous part of commutation rules between the translations and those between the generators of the homogeneous part of $\mathcal{B}$ and translations are controlled by different R-matrices satisfying characteristic equations.
[S0022-2488(99)00110-3]

I. INTRODUCTION

The Poincaré group plays a fundamental role in physics. It is intrinsically connected to the geometry of the space-time on which the physical systems are described and could be invariant under its action. Then, it is especially interesting to study the noncommutative version of the Poincaré group from which one can hope to obtain new insight on the underlying space–time geometry as, for example, an improved ultraviolet in quantum field theories or a description of symmetries in a future quantum Einstein–Cartan gravity.

The construction of quantum Poincaré group and quantum space–times has already been considered by several authors.[1–8] These constructions start either from the existence of an R-matrix and the consistency of the commutation rules between the different elements of the generators of the Poincaré group and those of the quantum space–times or from the projection of homogeneous quantum groups. In certain works this consistency is only obtained by introducing an extra generator.[4,6]

Recently, in the work Podles and Woronowicz[9] inhomogeneous commutation rules for an inhomogeneous Hopf algebra without dilatation have been constructed. This construction is based on the existence of a bicovariant submodule of the inhomogeneous algebra regarded as a left module (over the homogeneous part) generated by translations and leads to the same R-matrix, subject to the condition $R^2 = I^{\otimes 2}$, which controls the homogeneous parts of the different commutation rules.

It is well known that the usual way of constructing differential calculus on quantum groups consists first of constructing the quantum group, and then defining the bicovariant calculi, using right ad-invariant ideals. In such a case, for some deformations the basic one-forms span the space with dimension exceeding the group dimension or the number of its Lie algebra generators. In this paper, we present a quite different way based on the inverse order of construction: we introduce first the right-invariant differential calculus by assuming Cartan–Maurer types of formulas, and then show which classes of deformations of inhomogeneous symmetries are required for bicovariance of an assumed differential calculus. Using the fact that the homogeneous part of the inhomogeneous quantum group is a Hopf subalgebra, we obtain inhomogeneous commutation rules between elements of the inhomogeneous Hopf algebra $\mathcal{B}$. In particular, the homogeneous parts of the commutation rules are controlled by different R-matrices satisfying characteristic equations. This formalism eliminates from the beginning the mismatch of the space of the basic one-forms and the group dimensions and gives an extension of the results presented in Ref. 9 and the condition of the existence of differential calculus, with the classical dimension of Ref. 10.

The present paper is organized as follows: In Sec. II, we recall some basic notions about the inhomogeneous Hopf algebra and construct a right-invariant basis of the bimodule over this algebra. This right-invariant basis allows us to generalize, in Sec. III, the Woronowicz proposal[11] to the inhomogeneous Hopf algebra by introducing suitable functionals over $\mathcal{B}$. The study of these functionals leads us to construct inhomogeneous commutation rules between the elements of $\mathcal{B}$ whose homogeneous parts are controlled by different R-matrices satisfying characteristic equations. Finally, we investigate in Sec. IV the consistency conditions between these functionals and the different commutation rules of the inhomogeneous quantum group. Throughout this paper we use Einstein's convention (sum over repeated indices) and for the simplicity of calculations we define $v\top w\in M_{NN}(\mathcal{B})$ by $(v\top w)^{nm}_{kl}=v^n_k w^m_l$, $n,m,k,l\in 1,...,N$, for any v and $w\in M_N(\mathcal{B})$.

II. DIFFERENTIAL CALCULUS ON INHOMOGENEOUS QUANTUM GROUPS

In this section, we start by recalling some basic background about inhomogeneous Hopf algebra $\mathcal{B}$ and the covariant first-order differential calculus to construct a bicovariant bimodule Γ of one-forms over $\mathcal{B}$ and a basis of the vector space $\Gamma_{\rm inv}\subset\Gamma$ of all right-invariant elements of Γ.

An inhomogeneous quantum group $\mathcal{G}$ is built from a quantum group $\mathcal{H}$ and translations described by elements p^n, $n=1,...,N$ corresponding to an irreducible representation Λ^n_m of $\mathcal{H}$. The corresponding Hopf algebras are treated as algebras of functions on quantum groups, Poly$(\mathcal{G})=\mathcal{B}$ and Poly$(\mathcal{H})=\mathcal{A}$. More precisely, following Ref. 9, one defines $\mathcal{B}$ as follows.

(1) An abstract unital Hopf algebra generated by $\mathcal{A}$ and elements p^n such that $I_{\mathcal{B}}=I_{\mathcal{A}}=I$.
(2) $\mathcal{A}$ is a Hopf subalgebra of $\mathcal{B}$.
(3) $\mathcal{P}=\left(\begin{smallmatrix}\Lambda & p\\ 0 & I\end{smallmatrix}\right)$ is a representation of $\mathcal{G}$.
(4) There exists $n\in 1,...,N$ such that $p^n\notin\mathcal{A}$.
(5) $\Gamma_p\mathcal{A}\subset\Gamma_p$, where $\Gamma_p=\mathcal{A}X+\mathcal{A}$, $X=\text{span}\,(p^n,n=1,...,N)$.

By virtue of (2) and (3) above, $\mathcal{B}$ is endowed with the following linear maps:

(i) The coaction (algebra homomorphism) $\Delta:\mathcal{B}\to\mathcal{B}\otimes\mathcal{B}$,

$$\Delta(a)=a_{(1)}\otimes a_{(2)},\quad a\in\mathcal{B},$$

satisfyies the coassociativity condition

$$(\Delta\otimes\text{id})\Delta(a)=(\text{id}\otimes\Delta)\Delta(a)=a_{(1)}\otimes a_{(2)}\otimes a_{(3)}.$$

Due to the representation $\mathcal{P}$ of $\mathcal{G}$ the coaction acts on the generators as

$$\Delta(\Lambda^n_m)=\Lambda^n_k\otimes\Lambda^k_m,$$

$$\Delta(p^n)=\Lambda^n_k\otimes p^k+p^n\otimes I.$$

(ii) The counit (character) $\varepsilon:\mathcal{B}\to\mathcal{C}$ satisfies

$$(\varepsilon\otimes\text{id})\Delta(a)=(\text{id}\otimes\varepsilon)\Delta(a)=a,\quad a\in\mathcal{B},$$

acts on the generators as

$$\varepsilon(\Lambda^n_m)=\delta^n_m\quad\text{and}\quad\varepsilon(p^n)=0.$$

(iii) The antipode (algebra antihomomorphism) $S:\mathcal{B}\to\mathcal{B}$ satisfies

$$m\circ(S\otimes\text{id})\Delta(a)=m\circ(\text{id}\otimes S)\Delta(a)=I_{\mathcal{B}}\varepsilon(a),\quad a\in\mathcal{B},$$

where $m:\mathcal{B}\otimes\mathcal{B}\rightarrow\mathcal{B}$ is the multiplication map. It acts on the generators as

$$S(\Lambda^n_k)\Lambda^k_m=\varepsilon(\Lambda^n_m)=\delta^n_m=\Lambda^n_k S(\Lambda^k_m),$$

$$S(\Lambda^n_k)p^k+S(p^n)=\varepsilon(p^n)=0=\Lambda^n_k S(p^k)+p^n.$$

One says[11] that (Γ,d) is a first-order differential calculus on the inhomogeneous Hopf algebra $(\mathcal{B},\Delta,S,\varepsilon)$ if $d:\mathcal{B}\rightarrow\Gamma$ is a linear map obeying the Leibniz rule, $d(ab)=(da)b+a(db)$ for any $a,b\in\mathcal{B}$. Γ is a bimodule over $\mathcal{B}$ and every element of Γ is of the form $\Sigma_k a_k db_k$, where a_k, $b_k\in\mathcal{B}$. We say that (Γ,d) is left covariant if for any a_k, $b_k\in\mathcal{B}$

$$\sum_k a_k db_k=0\Leftrightarrow\sum_k \Delta(a_k)(\mathrm{id}\otimes d)\Delta(b_k)=0$$

and is right covariant if

$$\sum_k a_k db_k=0\Leftrightarrow\sum_k \Delta(a_k)(d\otimes\mathrm{id})\Delta(b_k)=0.$$

(Γ,d) is bicovariant if it is left and right covariant. This notion of covariant differential calculus leads to the left-coaction Δ_L and right-coaction Δ_R which are bimodule homomorphisms

$$\Delta_{L(R)}(a\rho b)=\Delta(a)\Delta_{L(R)}(\rho)\Delta(b),\quad a,b\in\mathcal{B},\rho\in\Gamma, \tag{1}$$

and satisfy

$$\Delta_L d=(\mathrm{id}\otimes d)\Delta,\quad \Delta_R d=(d\otimes\mathrm{id})\Delta,$$
$$(\varepsilon\otimes\mathrm{id})\Delta_L=\mathrm{id}\quad\text{and}\quad(\mathrm{id}\otimes\varepsilon)\Delta_R=\mathrm{id}. \tag{2}$$

For a bicovariant bimodule, the following condition is satisfied:

$$(\Delta_L\otimes\mathrm{id})\Delta_R=(\mathrm{id}\otimes\Delta_R)\Delta_L\,.$$

Since, $\mathcal{B}=\mathrm{Poly}(\mathcal{G})=\mathrm{Poly}(\Lambda,p)$, $d\Lambda^n_m$ and dp^n generate Γ as a bimodule of one-forms over $\mathcal{B}$. We therefore have the following.

Lemma (2.1):

$$\Theta^n_{Rm}=d\Lambda^n_k S(\Lambda^k_m)\quad\text{and}\quad\Pi^n_R=dp^n-\Theta^n_k p^k \tag{3}$$

form a basis of the vector space $\Gamma_{\mathrm{inv}}\subset\Gamma$ of all right-invariant elements of Γ.

Proof: Since p^n correspond to an irreducible representation Λ of $\mathcal{H}$, they are linearly independent, and so are $S(p^n)=-S(\Lambda^n_m)p^m$. Then $d\Lambda^n_m$ and dp^n or $d\Lambda^n_m$ and $dS(p^n)$ form a basis of the one-form vector space which generates Γ as a $\mathcal{B}$-bimodule from which we can choose an another basis $d(\Lambda^n_k)S(\Lambda^k_m)=\Theta^n_{Rm}$ and $\Lambda^n_k dS(p^k)=p^n-\Theta^n_{Rm}p^m=\Pi^n_R$. To show the right invariance, it suffices to apply (1), (2), and the properties of the Hopf algebra structure of $\mathcal{B}$ to get

$$\Delta_R(\Theta^n_{Rm})=d\Lambda^n_l S(\Lambda^p_m)\otimes\Lambda^l_k S(\Lambda^k_p)=\Theta^n_{Rm}\otimes I, \tag{4}$$

$$\begin{aligned}\Delta_R(\Pi^n_R)&=d\Lambda^n_k\otimes p^k+dp^n\otimes I-\Theta^n_{Rm}\Lambda^m_k\otimes p^k-\Theta^n_{Rm}p^m\otimes I\\&=\Theta^n_{Rm}\Lambda^m_k\otimes p^k+(dp^n-\Theta^n_{Rm}p^m)\otimes I-\Theta^n_{Rm}\Lambda^m_k\otimes p^k=\Pi^n_R\otimes I,\end{aligned} \tag{5}$$

$$\Delta_L(\Theta^n_{Rm})=\Lambda^n_l S(\Lambda^k_m)\otimes\Theta^l_{Rk}, \tag{6}$$

$$\Delta_L(\Pi^n_R)=\Lambda^n_k\otimes dp^k-\Lambda^n_k S(\Lambda^l_m)\Lambda^m_p\otimes\Theta^k_{Rl}p^p-\Lambda^n_k S(\Lambda^l_m)p^m\otimes\Theta^k_{Rl}$$
$$=\Lambda^n_k\otimes(dp^k-\Theta^k_{Rl}p^l)+\Lambda^n_k S(p^l)\otimes\Theta^k_{Rl}=\Lambda^n_k\otimes\Pi^k_R+\Lambda^n_k S(p^l)\otimes\Theta^k_{Rl}. \tag{7}$$

Q.E.D.

Remark (2.1): Note that we can construct from Π^n_R and Θ^n_{Rm} a left-invariant basis as

$$\Theta^n_{Lm}=S(\Lambda^n_k S(\Lambda^l_m))\Theta^k_{Rl}, \tag{8}$$

$$\Pi^n_L=S(\Lambda^n_m)\Pi^m_R+S(\Lambda^n_m S(p^k))\Theta^m_{Rk}, \tag{9}$$

satisfying

$$\Delta_L(\Theta^n_{Lm})=I\otimes\Theta^n_{Lm} \quad \text{and} \quad \Delta_L(\Pi^n_L)=I\otimes\Pi^n_L,$$

$$\Delta_R(\Theta^n_{Lm})=\Theta^k_{Ll}\otimes S(\Lambda^n_k S(\Lambda^l_m)),$$

$$\Delta_R(\Pi^n_L)=\Pi^k_L\otimes S(\Lambda^n_k)+\Theta^k_{Ll}\otimes S(\Lambda^n_k S(p^l)).$$

The bicovariance condition of these bases can be checked by direct computation. In the following, we consider the bases $\Theta^n_{Rm}=\Theta^n_m$ and $\Pi^n_R=\Pi^n$.

III. COMMUTATION RULES FOR INHOMOGENEOUS HOPF ALGEBRAS

Now, we are ready to study the commutation rules for inhomogeneous Hopf algebras by generalizing the formalism of the bicovariant bimodule theory of Ref. 11 to our bimodule presented in the previous section. Since Π^n and Θ^n_m form a right-invariant basis, we are in the case where we can apply the different stages of Theorem (2.3) of Ref. 11 to state that there exist linear functionals f^n_m, f^{nk}_m, f^n_{km}, and $f^{nl}_{mk}\in\mathcal{B}'$ such that

$$\Pi^n a=(a\star f^n_k)\Pi^k+(a\star f^{nl}_k)\Theta^k_l, \tag{10}$$

$$\Theta^n_m a=(a\star f^n_{mk})\Pi^k+(a\star f^{nl}_{mk})\Theta^k_l, \tag{11}$$

and

$$b\Pi^n=\Pi^k(b\star f^n_k\circ S)+\Theta^k_l(b\star f^{nl}_k\circ S),$$

$$b\Theta^n_m=\Pi^k(b\star f^n_{mk}\circ S)+\Theta^k_l(b\star f^{nl}_{mk}\circ S)$$

where the convolution product of a functional $f\in\mathcal{B}'$ and an element a of $\mathcal{B}$ is defined as $a\star f=(f\otimes\mathrm{id})\Delta(a)$. From (10) and (11), one deduces

$$f^n_k(I)=\delta^n_k, \quad f^{nl}_{mk}(I)=\delta^n_k\delta^l_m, \quad f^n_{mk}(I)=0, \quad f^{nl}_k(I)=0 \tag{12}$$

by setting $a=I$, and

$$\Pi^n ab=(ab\star f^n_k)\Pi^k+(ab\star f^{nl}_k)\Theta^k_l$$
$$=((a\star f^n_l)(b\star f^l_k)+(a\star f^{np}_q)(b\star f^q_{pk}))\Pi^k+((a\star f^n_q)(b\star f^{ql}_k)+(a\star f^{np}_q)(b\star f^{ql}_{pk}))\Theta^k_l,$$

$$\Theta^n_m ab=(ab\star f^n_{mk})\Pi^k+(ab\star f^{nl}_{mk})\Theta^k_l$$
$$=((a\star f^n_{mq})(b\star f^q_k)+(a\star f^{nq}_{mp})(b\star f^p_{qk}))\Pi^k+((a\star f^n_{mq})(b\star f^{ql}_k)+(a\star f^{np}_{mq})(b\star f^{ql}_{pk}))\Theta^k_l,$$

for any $a,b\in\mathcal{B}$. Comparing the coefficients multiplying Π and Θ and then applying ε, we get

$$f^n_m(ab)=f^n_k(a)f^k_m(b)+f^{nk}_l(a)f^l_{km}(b), \tag{13}$$

$$f^{nk}_{ml}(ab)=f^{np}_{mq}(a)f^{qk}_{pl}(b)+f^n_{mq}(a)f^{qk}_l(b), \tag{14}$$

$$f^n_{ml}(ab)=f^{np}_{mq}(a)f^q_{pl}(b)+f^n_{mk}(a)f^k_l(b), \tag{15}$$

$$f^{nk}_l(ab)=f^n_q(a)f^{qk}_l(b)+f^{np}_q(a)f^{qk}_{pl}(b). \tag{16}$$

Applying now Δ_L on both sides of (10) and (11), we obtain, respectively,

$$\begin{aligned}\Lambda^n_p a_{(1)}\otimes\Pi^p a_{(2)}+\Lambda^n_p S(p^q)a_{(1)}\otimes\Theta^p_q a_{(2)}\\ =f^n_p(a_{(1)})a_{(2)}\Lambda^p_k\otimes a_{(3)}\Pi^k+(f^{nq}_p(a_{(1)})a_{(2)}\Lambda^p_k S(\Lambda^l_q)+f^n_p(a_{(1)})a_{(2)}\Lambda^p_k S(p^l))\otimes a_{(3)}\Theta^k_l\end{aligned} \tag{17}$$

and

$$\begin{aligned}\Lambda^n_q S(\Lambda^p_m)a_{(1)}\otimes\Theta^q_p a_{(2)}=f^n_{mp}(a_{(1)})a_{(2)}\Lambda^p_k\otimes a_{(3)}\Pi^k+(f^n_{mp}(a_{(1)})a_{(2)}\Lambda^p_k S(p^l)\\ +f^{nq}_{mp}(a_{(1)})a_{(2)}\Lambda^p_k S(\Lambda^l_q))\otimes a_{(3)}\Theta^k_l.\end{aligned} \tag{18}$$

By virtue of (10) and (11), the left-hand sides of previous equations can be rewritten respectively as

$$\begin{aligned}(\Lambda^n_p a_{(1)}f^p_k(a_{(2)})+\Lambda^n_p S(p^q)f^p_{qk}(a_{(2)}))\otimes a_{(3)}\Pi^k\\ +(\Lambda^n_p a_{(1)}f^{pl}_k(a_{(2)})+\Lambda^n_p S(p^q)a_{(1)}f^{pl}_{qk}(a_{(2)}))\otimes a_{(3)}\Theta^k_l\end{aligned}$$

and

$$\Lambda^n_q S(\Lambda^p_m)a_{(1)}f^q_{pk}(a_{(2)})\otimes a_{(3)}\Pi^k+\Lambda^n_q S(\Lambda^p_m)a_{(1)}f^{ql}_{pk}(a_{(2)})\otimes a_{(3)}\Theta^k_l.$$

Substituting the first expression into the left-hand side of (17) and the second into the left-hand side of (18), comparing the coefficients multiplying $I\otimes\Pi$ and $I\otimes\Theta$, and then applying $I\otimes\varepsilon$, we obtain

$$\Lambda^n_k(f^k_m\star a)+\Lambda^n_k S(p^l)(f^k_{lm}\star a)=(a\star f^n_k)\Lambda^k_m, \tag{19}$$

$$\Lambda^n_k(f^{kl}_m\star a)+\Lambda^n_k S(p^q)(f^{kl}_{qm}\star a)=(a\star f^n_k)\Lambda^k_m S(p^l)+(a\star f^{nk}_q)\Lambda^q_m S(\Lambda^l_k), \tag{20}$$

$$\Lambda^n_k S(\Lambda^q_m)(f^k_{ql}\star a)=(a\star f^n_{mk})\Lambda^k_l, \tag{21}$$

$$\Lambda^n_k S(\Lambda^q_m)(f^{kp}_{ql}\star a)=(a\star f^n_{mk})\Lambda^k_l S(p^p)+(a\star f^{nq}_{mk})\Lambda^k_l S(\Lambda^p_q), \tag{22}$$

for any $a\in\mathcal{B}$. In the following, we assume that

$$f^{nk}_{ml}=\tilde{f}^k_m\star f^n_l,\quad f^{nk}_m=\tilde{\eta}^k\star f^n_m,\quad \text{and}\ f^n_{km}=\tilde{\eta}_k\star f^n_m \tag{23}$$

where $\tilde{\eta}^n$, $\tilde{\eta}_n$ and $\tilde{f}^n_m\in\mathcal{B}'$ and the convolution product of two functionals $\in\mathcal{B}'$ is defined as $(f_1\star f_2)(a)=(f_1\otimes f_2)\Delta(a)$ for any $a\in\mathcal{B}$. The substitution of (23) into (12) gives

$$\tilde{\eta}^n(I)=0,\quad \tilde{\eta}_n(I)=0,\quad \text{and}\ \tilde{f}^n_m(I)=\delta^n_m, \tag{24}$$

and the substitution of (23) into (13)–(16) and (19) gives, respectively,

$$f^n_m(ab)=(\varepsilon(a_{(1)})\varepsilon(b_{(1)})+\tilde{\eta}^k(a_{(1)})\tilde{\eta}_k(b_{(1)}))f^n_l(a_{(2)})f^l_m(b_{(2)}), \tag{25}$$

$$f^{nk}_{ml}(ab)=\tilde{f}^{k}_{m}(a_{(1)}b_{(1)})f^{n}_{l}(a_{(2)}b_{(2)})=(\tilde{f}^{p}_{m}(a_{(1)})\tilde{f}^{k}_{p}(b_{(1)})+\tilde{\eta}_{m}(a_{(1)})\tilde{\eta}^{k}(b_{(1)}))f^{n}_{q}(a_{(2)})f^{q}_{l}(b_{(2)}), \tag{26}$$

$$f^{n}_{km}(ab)=\tilde{\eta}_{k}(a_{(1)}b_{(1)})f^{n}_{m}(a_{(2)})b_{(2)})=(\tilde{f}^{p}_{k}(a_{(1)})\tilde{\eta}_{p}(b_{(1)})+\tilde{\eta}_{k}(a_{(1)})\varepsilon(b_{(1)}))f^{n}_{q}(a_{(2)})f^{q}_{m}(b_{(2)}), \tag{27}$$

$$f^{nk}_{m}(ab)=\tilde{\eta}^{k}(a_{(1)}b_{(1)})f^{n}_{m}(a_{(2)}b_{(2)})=(\varepsilon(a_{(1)})\tilde{\eta}^{k}(b_{(1)})+\tilde{\eta}^{p}(a_{(1)})\tilde{f}^{k}_{p}(b_{(1)}))f^{n}_{q}(a_{(2)})f^{q}_{m}(b_{(2)}), \tag{28}$$

and

$$\Lambda^{n}_{k}(f^{k}_{m}\star a)+\Lambda^{n}_{k}S(p^{l})(\tilde{\eta}_{l}\star f^{k}_{m}\star a)=(a\star f^{n}_{k})\Lambda^{k}_{m}. \tag{29}$$

Using now the associativity of the convolution product and making the substitution of (29) into the right-hand side of (20)–(22), we get

$$\tilde{\eta}^{i}\star f^{k}_{j}\star a+S(p^{l})(\tilde{f}^{i}_{l}\star f^{k}_{j}\star a)=(f^{k}_{j}\star a)S(p^{i})+(f^{k}_{j}\star a\star\tilde{\eta}^{l})S(\Lambda^{i}_{l})+S(p^{l})(\tilde{\eta}_{l}\star f^{k}_{j}\star a)S(p^{i})+S(p^{p})\times(\tilde{\eta}_{p}\star f^{k}_{j}\star a\star\tilde{\eta}^{l})S(\Lambda^{i}_{l}), \tag{30}$$

$$S(\Lambda^{k}_{m})(\tilde{\eta}_{k}\star f^{n}_{l}\star a)=(f^{n}_{l}\star a\star\tilde{\eta}_{m})+S(p^{k})(\tilde{\eta}_{k}\star f^{n}_{l}\star a\star\tilde{\eta}_{m}), \tag{31}$$

and

$$S(\Lambda^{l}_{m})(\tilde{f}^{i}_{l}\star f^{k}_{j}\star a)=S(\Lambda^{l}_{m})(\tilde{\eta}_{l}\star f^{k}_{j}\star a)S(p^{i})+(f^{k}_{j}\star a\star\tilde{f}^{l}_{m})S(\Lambda^{i}_{l})+S(p^{q})(\tilde{\eta}_{q}\star f^{k}_{j}\star a\star\tilde{f}^{l}_{m})S(\Lambda^{i}_{l}), \tag{32}$$

where we have multiplied from the left by $S(\Lambda^{p}_{n})$ and we have used (21) before (29) to obtain (32). We see from these relations that for any $a\in\mathcal{A}$, (30) is a commutation rule between elements of $\mathcal{A}$ and p^n and (29), (31), and (32) involve p^n in the commutation rules between elements of $\mathcal{A}$. Demanding $\mathcal{A}$ to be a Hopf subalgebra of $\mathcal{B}$ [condition (2)] leads to the following.

Proposition (3.1): $\mathcal{A}$ is a Hopf subalgebra of $\mathcal{B}$ iff

$$\tilde{\eta}_{n}(a)=0, \quad a\in\mathcal{A}. \tag{33}$$

Proof: For $\mathcal{A}$ to be a Hopf subalgebra of $\mathcal{B}$, (29) and (32) must involve commutation rules between elements of $\mathcal{A}$ only. This is satisfied if the term $\Lambda^{n}_{k}S(p^{l})(\tilde{\eta}_{l}\star f^{k}_{m}\star a)$ of (29) vanishes. Multiplying it from the left by $S(\Lambda^{i}_{n})$ and using the fact that $S(p^{l})$ are linearly independent and generate with the unity the bimodule Γ_p [see condition (5)], one obtains $\tilde{\eta}_{l}\star f^{i}_{m}\star a=0$ which permits us to write, for any a and $b\in\mathcal{A}$, (25) under the form

$$f^{n}_{m}(ab)=f^{n}_{k}(a)f^{k}_{m}(b)\Rightarrow f^{n}_{k}\star f^{k}_{m}\circ S=f^{n}_{k}\circ S\star f^{k}_{m}=\delta^{n}_{m}\varepsilon. \tag{34}$$

Replacing a by $f^{m}_{k}\circ S\star a$ into $\tilde{\eta}_{l}\star f^{i}_{m}\star a=0$ and then acting ε, we get $\tilde{\eta}_{l}(a)=0$ for any $a\in\mathcal{A}$ which implies also that Eq. (31) is trivial and the term of (32) which involves p^n in the commutation rules between elements of $\mathcal{A}$ vanishes. Conversely, it is easy to see that if $\tilde{\eta}_{l}(a)=0$ for any $a\in\mathcal{A}$, the relations (29) and (32) involve commutation rules between elements of $\mathcal{A}$ only and, therefore, $\mathcal{A}$ is a Hopf subalgebra of $\mathcal{B}$. Q.E.D.

By virtue of this proposition, one deduces certain results which will be used in the following.

(1) Equation (11) can be written as

$$\Theta^{n}_{m}a=(a\star f^{nl}_{mk})\Theta^{k}_{l}, \quad a\in\mathcal{A},$$

which shows that the set of the right-invariant one-forms Θ^n_m generates a subimodule, $\Gamma_{\mathcal{A}}\subset\Gamma$, over $\mathcal{A}$ and for $\tilde{\eta}_n(p^k)=0$ it generates a $\mathcal{B}$-subimodule of Γ. Since it transforms according to an adjoint representation Λ of $\mathcal{H}$ (6), the assumption $f^{nl}_{mk}=\tilde{f}^l_m\star f^n_k$ is justified.[12]

(2) For any $a\in\mathcal{A}$, (26) can be written as

$$\tilde{f}^k_m(a_{(1)}b_{(1)})f^n_l(a_{(2)}b_{(2)})=\tilde{f}^p_m(a_{(1)})\tilde{f}^k_p(b_{(1)})f^n_q(a_{(2)})f^q_l(b_{(2)})=\tilde{f}^p_m(a_{(1)})\tilde{f}^k_p(b_{(1)})f^n_l(a_{(2)}b_{(2)}),$$

where we have used (34) to get the right-hand side. Multiplying both sides from the right by $f^l_r(S(a_{(3)}b_{(3)}))$ and using again (34), we get

$$\tilde{f}^n_m(ab)=\tilde{f}^k_m(a)\tilde{f}^n_k(b),\quad a,b\in\mathcal{A}. \tag{35}$$

Following similar considerations, one obtains from (28)

$$\tilde{\eta}^n(ab)=\varepsilon(a)\tilde{\eta}^n(b)+\tilde{\eta}^m(a)\tilde{f}^n_m(b),\quad a,b\in\mathcal{A}, \tag{36}$$

which leads to

$$\begin{aligned}\tilde{\eta}^l(\Lambda^n_kS(\Lambda^k_m)=\tilde{\eta}^l(\delta^n_m)=0=\tilde{\eta}^l(S(\Lambda^n_m))+\tilde{\eta}^q(\Lambda^n_k)\tilde{f}^l_q(S(\Lambda^k_m)),\\ \tilde{\eta}^l(S(\Lambda^n_k)\Lambda^k_m)=\tilde{\eta}^l(\delta^n_m)=0=\tilde{\eta}^l(\Lambda^n_m)=\tilde{\eta}^q(S(\Lambda^n_k))\tilde{f}^l_q(\Lambda^k_m).\end{aligned} \tag{37}$$

Now replacing ab by $-S(\Lambda^a_b)p^b=S(p^a)$ into (28) and using (37), one obtains

$$\tilde{\eta}^n(S(p^a))=-\tilde{\eta}^n(p^a)-\tilde{\eta}^k(S(\Lambda^a_b))\tilde{f}^n_k(p^b). \tag{38}$$

On the other hand, by using (33), (27) gives

$$(\tilde{\eta}_k\star f^n_m)(S(p^a))=\tilde{\eta}_k(S(p^c))f^n_m(S(\Lambda^a_c))=-(\tilde{\eta}_k\star f^n_m)(S(\Lambda^a_b)p^b)=-\tilde{f}^p_k(S(\Lambda^c_b))\tilde{\eta}_p(p^b)f^n_m(S(\Lambda^a_c)),$$

which can be multiplied from the right by $f^m_l(\Lambda^c_d)$ to give

$$\tilde{\eta}_n(S(p^a))=-\tilde{f}^k_n(S(\Lambda^a_b))\tilde{\eta}_k(p^b). \tag{39}$$

A similar computation gives respectively from (25) and (26)

$$f^n_m(p^a)=-f^n_m(\Lambda^a_bS(p^b))=-f^n_k(\Lambda^a_b)f^k_m(S(p^b))-\tilde{\eta}^k(\Lambda^a_b)\tilde{\eta}_k(S(p^b))\delta^n_m \tag{40}$$

and

$$\tilde{f}^n_m(p^a)=-\tilde{f}^k_m(\Lambda^a_b)\tilde{f}^n_k(S(p^b))+\tilde{\eta}^k(\Lambda^c_b)\tilde{\eta}_k(S(p^b))\tilde{f}^n_m(\Lambda^a_c). \tag{41}$$

We are now ready to investigate the different commutation rules of the inhomogeneous Hopf algebra $\mathcal{B}$.

First, due to Proposition (3.1), (29) reduces to

$$\Lambda^n_k(f^k_m\star a)=(a\star f^n_k)\Lambda^k_m \tag{42}$$

for any $a\in\mathcal{A}$ and gives

$$\Lambda^n_mp^a=f^n_k(\Lambda^a_b)p^b\Lambda^k_m+f^n_k(p^a)\Lambda^k_m-\Lambda^n_k\Lambda^a_bf^k_m(p^b)-\Lambda^n_mS(p^l)\Lambda^a_b\tilde{\eta}_l(p^b) \tag{43}$$

for $a=p^a$.

Second, replacing $a\in\mathcal{B}$ by $S(a)$ into (30)–(32) and then acting on both sides S^{-1}, we get, respectively,

$$(a\star f^k_m\circ S\star \tilde{\eta}^l\circ S)+(a\star f^k_m\circ S\star \tilde{f}^l_q\circ S)p^q$$
$$=p^l(a\star f^k_m\circ S)+p^l(a\star f^k_m\circ S\star \tilde{\eta}_q\circ S)p^q+\Lambda^l_p(\tilde{\eta}^p\circ S\star a\star f^k_m\circ S)$$
$$+\Lambda^l_p(\tilde{\eta}^p\circ S\star a\star f^k_m\circ S\star \tilde{\eta}_q\circ S)p^q, \tag{44}$$

$$(a\star f^n_l\circ S\star \tilde{\eta}_k\circ S)\Lambda^k_m=\tilde{\eta}_m\circ S\star a\star f^n_l\circ S+(\tilde{\eta}_m\circ S\star a\star f^n_l\circ S\star \tilde{\eta}_q\circ S)p^q, \tag{45}$$

and

$$(a\star f^k_j\circ S\star \tilde{f}^i_l\circ S)\Lambda^l_m=p^i(a\star f^k_j\circ S\star \tilde{\eta}_l\circ S)\Lambda^l_m+\Lambda^i_l(\tilde{f}^l_m\circ S\star a\star f^k_j\circ S)+\Lambda^i_l(\tilde{f}^l_m\circ S\star a\star f^k_j\circ S\star \tilde{\eta}_q\circ S)p^q. \tag{46}$$

Due to Proposition (3.1), (45) is trivial for $a\in\mathcal{A}$ and, by replacing a by $(a\star f^n_k)\in\mathcal{A}$, the equations (46) and (44) give, respectively,

$$\Lambda^n_k(\tilde{f}^k_m\circ S\star a)=(a\star \tilde{f}^n_k\circ S)\Lambda^k_m \tag{47}$$

and

$$p^n a=(a\star \tilde{f}^n_k\circ S)p^k+a\star \tilde{\eta}^n\circ S-\Lambda^n_k(\tilde{\eta}^k\circ S\star a). \tag{48}$$

For $a=p^a$, the equations (45) and (46) give, respectively,

$$\Lambda^k_m\tilde{\eta}_k(S(p^a))=\tilde{\eta}_m(S(p^b))\Lambda^a_b \tag{49}$$

and

$$\Lambda^n_m p^b(\delta^a_b+\tilde{\eta}_b(S(p^a)))=(\tilde{f}^n_k(S(\Lambda^a_b))-\delta^n_b\tilde{\eta}_k(S(p^a)))p^b\Lambda^k_m+\tilde{f}^n_k(S(p^a))\Lambda^k_m-\Lambda^n_k\Lambda^a_b\tilde{f}^k_m(S(p^b)). \tag{50}$$

To compare (50) with (43), we have to compute the fourth term of the right-hand side of (43), which can be written as

$$-\Lambda^n_m S(p^l)\Lambda^a_c\tilde{\eta}_l(p^c)=\Lambda^n_m S(\Lambda^l_k)p^k\Lambda^a_c\tilde{\eta}_l(p^c)$$
$$=-\Lambda^n_m p^l\tilde{\eta}_l(p^a)+\Lambda^n_m\tilde{\eta}^k(S(\Lambda^a_b))\tilde{\eta}_k(p^b)-\Lambda^n_m\Lambda^a_b\tilde{\eta}^k(S(\Lambda^b_c))\tilde{\eta}_k(p^c), \tag{51}$$

where we have used (39),

$$p^k\Lambda^a_c=\tilde{f}^k_l(S(\Lambda^a_b))\Lambda^b_c p^l+\tilde{\eta}^k(S(\Lambda^a_b))\Lambda^b_c-\Lambda^k_l\Lambda^a_b\tilde{\eta}^l(S(\Lambda^b_c)), \tag{52}$$

obtained from (48) by setting $a=\Lambda^a_c$, and

$$S(\Lambda^k_m)\Lambda^a_c\tilde{\eta}_k(p^c)=\tilde{\eta}_m(p^a), \tag{53}$$

obtained from (21) by setting $a=p^a$ and by using (23) and Proposition (3.1).

Replacing the fourth term of the right-hand side of (43) by (51), we get

$$\Lambda^n_m p^b(\delta^a_b+\tilde{\eta}_b(S(p^a)))=f^n_k(\Lambda^a_b)p^b\Lambda^k_m+(f^n_k(p^a)+\delta^n_k\tilde{\eta}^p(S(\Lambda^a_b))\tilde{\eta}_p(p^b))\Lambda^k_m$$
$$-\Lambda^n_k\Lambda^a_b(f^k_m(p^b)+\delta^k_m\tilde{\eta}^p(S(\Lambda^b_c))\tilde{\eta}_p(p^c)), \tag{54}$$

which can be written, after the make of use of (37), (39), and (40), under the form

$$\Lambda^n_m p^b(\delta^a_b+\tilde{\eta}_b(S(p^a)))=f^n_k(\Lambda^a_c)p^c\Lambda^k_m-f^n_l(\Lambda^a_b)f^l_k(S(p^b))\Lambda^k_m+\Lambda^n_k\Lambda^a_b f^k_l(\Lambda^b_c)f^l_m(S(p^c)). \tag{55}$$

Therefore, by virtue of the condition (5) of the inhomogeneous Hopf algebra definition, we can deduce from (55) and (50) that

$$\tilde{f}^n_k(S(\Lambda^a_b))=f^n_k(\Lambda^a_b)+\delta^n_b\tilde{\eta}_k(S(p^a)) \tag{56}$$

and

$$\tilde{f}^n_k(S(p^a))=-f^n_l(\Lambda^a_b)f^l_k(S(p^b))=f^n_k(p^a)+\delta^n_k\tilde{\eta}^l(S(\Lambda^a_b))\tilde{\eta}_l(p^b), \tag{57}$$

where we have used (40) to get the second relation. Let us note that, for $a=\Lambda^a_b$, the equations (47) and (42) are consistent in virtue of (56) and (49).

Now, multiplying both sides of (55) from the left by $f(S(\Lambda))$, we get

$$p^n\Lambda^k_m=f^k_q(S(\Lambda^n_p))(\delta^p_l+\tilde{\eta}_l(S(p^p)))\Lambda^q_m p^l+f^k_q(S(p^n))\Lambda^q_m-\Lambda^n_q\Lambda^k_p f^p_m(S(p^q)), \tag{58}$$

where we have used (42). Comparing the latter equation with (52), we obtain

$$\tilde{f}^n_l(S(\Lambda^k_q))=f^k_q(S(\Lambda^n_p))(\delta^p_l+\tilde{\eta}_l(S(p^p))) \tag{59}$$

and

$$\tilde{\eta}^n(S(\Lambda^k_q))=f^k_q(S(p^n)). \tag{60}$$

Substituting (56) into (59), we get

$$f^n_l(\Lambda^k_q)+\delta^n_q\tilde{\eta}_l(S(p^k))=f^k_q(S(\Lambda^n_p))(\delta^p_l+\tilde{\eta}^l(S(p^p))), \tag{61}$$

which can be written, after multiplying both sides by $f^m_k(\Lambda^t_n)$, as

$$(f^m_k(\Lambda^t_n)-\delta^m_n\delta^t_k)(f^n_l(\Lambda^k_q)+\delta^n_q\delta^k_l+\delta^n_q\tilde{\eta}_l(S(p^k)))=0 \tag{62}$$

or, using again (56),

$$(\tilde{f}^m_k(S(\Lambda^t_n))-\delta^m_n\delta^t_k-\delta^m_n\tilde{\eta}_k(S(p^t)))(\tilde{f}^n_l(S(\Lambda^k_q))+\delta^n_q\delta^k_l)=0, \tag{63}$$

which are the characteristic equations for the matrices $R^{mt}_{nk}=f^m_k(\Lambda^t_n)$ and $\tilde{R}^{mt}_{nk}=\tilde{f}^m_k(S(\Lambda^t_n))$. To have the commutation rules between the translations, it suffices to replace a by p^a into (44) to get after some straightforward computation

$$\begin{aligned}p^n p^m=&(\tilde{f}^n_k(S(\Lambda^m_l))-\delta^n_l\tilde{\eta}_k(S(p^m)))p^l p^k+\tilde{\eta}^n(S(\Lambda^m_k))p^k+\tilde{f}^n_k(S(p^m))p^k\\&+\tilde{\eta}^n(S(p^m))-\Lambda^n_k\Lambda^m_l\tilde{\eta}^k(S(p^l))\\=&f^n_k(\Lambda^m_l)p^l p^k-(f^n_k(\Lambda^m_l)-\delta^n_l\delta^m_k)\tilde{\eta}^l(S(\Lambda^k_p))p^p\\&+\tilde{\eta}^n(S(p^m))-\Lambda^n_k\Lambda^m_l\tilde{\eta}^k(S(p^l)),\end{aligned} \tag{64}$$

where we have used (56), (57), and (60). With the assumption (23), we have the following from the previous considerations.

Theorem (3.1): Let $\mathcal{B}$ an inhomogeneous Hopf algebra satisfying conditions (1)–(5) on which there exists a bicovariant differential calculus. Then there exist functionals f^n_m, $\tilde{f}^n_m$, $\tilde{\eta}^n$, and $\tilde{\eta}_m\in\mathcal{B}'$ satisfying (24), (25)–(28), (33), (56), (57), (59), and (60) such that the relations (42), (48), and (64) are satisfied.

IV. CONSISTENCY CONDITIONS

Here we shall continue the study of the consistency conditions between the different functionals and the commutation rules of the generators of $\mathcal{G}$. In the following we set $R^{kn}_{ml}=f^k_l(\Lambda^n_m)$, $\widetilde{R}^{kn}_{ml}=\widetilde{f}^k_l(S(\Lambda^n_m))$, $Q^n_k=\widetilde{\eta}_k(S(p^n))$, $Z^{nm}_k=\widetilde{\eta}^n(S(\Lambda^m_k)=f^m_k(S(p^n))$ [see (60)], $\widetilde{Z}^{nm}_k=\widetilde{f}^n_k(S(p^m))=-R^{nm}_{pq}Z^{pq}_k$ [see (57)], and $T^{nm}=\widetilde{\eta}^n(S(p^m))$. We start this study as follows.

Lemma (4.1):

$$Q^n_m=\lambda\,\delta^n_m,\quad \lambda\in\mathcal{C},\quad \lambda\neq-1. \tag{65}$$

Proof: Applying f^a_b and $\widetilde{f}^a_b\circ S$ on both sides of (49), we obtain respectively $(I\otimes Q)R=R(Q\otimes I)$ and $(I\otimes Q)\widetilde{R}=\widetilde{R}(Q\otimes I)$, implying, due to (56) $(\widetilde{R}=R+I\otimes Q)$, $I\otimes Q^2=Q\otimes Q$ which solution is of the form (65). For $\lambda=-1$, we see from (62) that the R-matrix is not invertible. Q.E.D.

Applying f^a_b on both sides of (42) and $\widetilde{f}^a_b\circ S$ on both sides of (47) for $a=\Lambda^p_q$ and using (34) and (35), we get the Yang–Baxter equations for the matrices R,

$$(I\otimes R)(R\otimes I)(I\otimes R)=(R\otimes I)(I\otimes R)(R\otimes I), \tag{66}$$

and $\widetilde{R}$,

$$(I\otimes\widetilde{R})(\widetilde{R}\otimes I)(I\otimes\widetilde{R})=(\widetilde{R}\otimes I)(I\otimes\widetilde{R})(\widetilde{R}\otimes I). \tag{67}$$

Since the $\widetilde{R}$-matrix is tied to the R-matrix by (56), it is necessary to check the consistency between both Yang–Baxter equations. Applying f^a_b on both sides of (47) and $\widetilde{f}^a_b\circ S$ on both sides of (42) for $a=\Lambda^p_q$, we obtain

$$(R\otimes I)(I\otimes R)(\widetilde{R}\otimes I)=(I\otimes\widetilde{R})(R\otimes I)(I\otimes R) \tag{68}$$

and

$$(\widetilde{R}\otimes I)(I\otimes\widetilde{R})(R\otimes I)=(I\otimes R)(\widetilde{R}\otimes I)(I\otimes\widetilde{R}). \tag{69}$$

Replacing (56) on both sides of (68) and (69) and using (66), we obtain, respectively,

$$(R\otimes I)(I\otimes R)(I\otimes Q\otimes I)=(I\otimes I\otimes Q)(R\otimes I)(I\otimes R),$$

and

$$\begin{aligned}&(R\otimes I)(I\otimes I\otimes Q)(R\otimes I)+(I\otimes Q\otimes I)(I\otimes R)(R\otimes I)+(I\otimes Q\otimes I)(I\otimes I\otimes Q)(R\otimes I)\\&\quad=(I\otimes R)(R\otimes I)(I\otimes I\otimes Q)+(I\otimes R)(I\otimes Q\otimes I)(I\otimes R)+(I\otimes R)(I\otimes Q\otimes I)(I\otimes I\otimes Q).\end{aligned}$$

By virtue of (65) the first equation is satisfied and, using (62) $[R^2=-R(I\otimes Q)+I\otimes I+I\otimes Q]$, we can see that the second equation is also satisfied. Now, replacing the third factor of the left-hand side and the first factor of the right-hand side of (67) by (56) and using (69), we get

$$(\widetilde{R}\otimes I)(I\otimes\widetilde{R})(I\otimes Q\otimes I)=(I\otimes I\otimes Q)(\widetilde{R}\otimes I)(I\otimes\widetilde{R}),$$

which is satisfied in virtue of (65). Therefore (66)–(69) are equivalent. The action of $\widetilde{\eta}^a\circ S$ on both sides of (49) gives

$$ZQ=(I\otimes Q)Z, \tag{70}$$

and on (42) and (47) gives, for $a=\Lambda^p_q$,

$$(Z\otimes I)R+(\widetilde{R}\otimes I)(I\otimes Z)R=(I\otimes R)(Z\otimes I)+(I\otimes R)(\widetilde{R}\otimes I)(I\otimes Z) \tag{71}$$

and

$$(Z\otimes I)\tilde{R}+(\tilde{R}\otimes I)(I\otimes Z)\tilde{R}=(I\otimes\tilde{R})(Z\otimes I)+(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes Z), \qquad (72)$$

where we have used (36). Due to (65), the equation (70) is satisfied and if we use (56) in (72) and compare with (71), we obtain

$$(Z\otimes I)(I\otimes Q)+(\tilde{R}\otimes I)(I\otimes Z)(I\otimes Q)=(I\otimes I\otimes Q)(Z\otimes I)+(I\otimes I\otimes Q)(\tilde{R}\otimes I)(I\otimes Z),$$

which is satisfied by virtue of (65) or (70). Therefore, the relation (71) implies (72).

We pass now to the action of the different functionals on the commutation rules between the generators of $\mathcal{A}$ and the translations. Applying $\tilde{\eta}_a\circ S$ on both sides of (52) and using the relations

$$\tilde{\eta}_a(aS(p^n))=\tilde{f}_a^b(a)\,\tilde{\eta}_b(S(p^n))$$

and

$$\tilde{\eta}_a(S(p^n)a)=\tilde{\eta}_a(S(p^n))\varepsilon(a)$$

obtained from (27) for any $a\in\mathcal{A}$, we get $(Q\otimes I)\tilde{R}=\tilde{R}(I\otimes Q)$, which is satisfied due to (65). Applying $f_b^a\circ S$ and $\tilde{f}_b^a\circ S$ on both sides of (52), we get, respectively,

$$\begin{aligned}&(I\otimes R^{-1})(Z\otimes I)+(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes Q\otimes I)(I\otimes Z)\\&\quad=(\tilde{R}\otimes I)(I\otimes Z)R^{-1}+(Z\otimes I)R^{-1}-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes Z)\end{aligned} \qquad (73)$$

and

$$\begin{aligned}&(\tilde{Z}\otimes I)\tilde{R}-(I\otimes Z)\tilde{R}-(I\otimes Q\otimes I)(I\otimes Z)\tilde{R}\\&\quad=(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes\tilde{Z})+(I\otimes\tilde{R})(I\otimes I\otimes Q)(Z\otimes I)-(\tilde{R}\otimes I)(I\otimes\tilde{R})(Z\otimes I),\end{aligned} \qquad (74)$$

where we have used (25) to have the first equation and

$$\tilde{f}_b^a(S(p^n)a)=\tilde{f}_b^c(S(p^n))\tilde{f}_c^a(a)+\tilde{\eta}_b(S(p^n))\,\tilde{\eta}^a(a)$$

and

$$\tilde{f}_b^a(aS(p^n))=\tilde{f}_b^c(a)\tilde{f}_c^a(S(p^n))-\tilde{f}_b^a(a_{(1)})\,\tilde{\eta}^d(a_{(2)})\,\tilde{\eta}_d(S(p^n)),$$

obtained from (26) for any $a\in\mathcal{A}$, to have the second equation. Multiplying both sides of (71) from the left by $I\otimes R^{-1}$ and from the right by R^{-1} and using (59) $[\tilde{R}=R^{-1}+R^{-1}(I\otimes Q)]$, we retrieve the equation (73). To investigate the equation (74), we have to multiply both sides of (72) from the left by $-(R\otimes I)$ to get

$$(\tilde{Z}\otimes I)\tilde{R}-(I\otimes Z)\tilde{R}-(I\otimes Q\otimes I)(I\otimes Z)\tilde{R}=-(R\otimes I)(I\otimes\tilde{R})(Z\otimes I)-(R\otimes I)(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes Z),$$

where we have used $\tilde{Z}=-RZ$ and $R\tilde{R}=I\otimes I+I\otimes Q$. Replacing R by $\tilde{R}-I\otimes Q$ on the right-hand side and using (67) and again $\tilde{Z}=-RZ$, we retrieve (74). Therefore, (71) implies (73) and (74).

Finally the action of $\tilde{\eta}^a\circ S$ on both sides of (52) gives

$$(R\otimes I-I\otimes I\otimes I)((I\otimes Z)Z-(Z\otimes I)Z)+T\otimes I-(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T)=0, \qquad (75)$$

where we have used the relations

$$\tilde{\eta}^a(S(ap^n))=\tilde{\eta}^b(S(p^n))\tilde{f}_b^a(S(a))$$

and

$$\tilde{\eta}^a(S(p^n a))=\varepsilon(S(a))\tilde{\eta}^a(S(p^n))+\tilde{\eta}^b(S(a))\tilde{f}^a_b(S(p^n))-\tilde{\eta}^a(S(a_{(2)}))\tilde{\eta}^b(S(a_{(1)}))\tilde{\eta}_b(S(p^n))$$

obtained from (28) for any $a\in\mathcal{A}$.

We consider now the commutation rules between the translations. From

$$\tilde{\eta}_a(S(p^m)S(p^n))=\tilde{f}^b_a(S(p^m))\tilde{\eta}_b(S(p^n))-\tilde{\eta}_a(S(p^k))\tilde{\eta}^b(S(\Lambda^m_k))\tilde{\eta}_b(S(p^n)),$$

obtained from (27), we can see that the action of $\tilde{\eta}_a\circ S$ on both sides of (64) gives

$$(R-I\otimes I)((Q\otimes I)\tilde{Z}-(Q\otimes I)ZQ-ZQ)=0.$$

From (65) and $\tilde{Z}=-RZ$ we can rewrite the left-hand side of this equation as $-(R-I\otimes I)(R+I\otimes I+I\otimes Q)ZQ$, which vanishes by virtue of (62). Applying $f^a_b\circ S$, $\tilde{f}^a_b\circ S$, and $\tilde{\eta}^a\circ S$ on both sides of (64), we get, respectively,

$$\begin{aligned}(R\otimes I-I\otimes I\otimes I)((I\otimes Z)Z-(Z\otimes I)Z+(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes Q\otimes I)(I\otimes T))\\ +(T\otimes I)-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes T)=0,\end{aligned}\tag{76}$$

$$\begin{aligned}(I\otimes R-I\otimes I\otimes I)((\tilde{Z}\otimes I)\tilde{Z}+(T\otimes I)Q-(I\otimes Q\otimes I)(I\otimes Z)\tilde{Z}-(I\otimes Q\otimes I)(I\otimes T)\\ -(I\otimes Z)\tilde{Z})+(I\otimes T)-(\tilde{R}\otimes I)(I\otimes\tilde{R})(T\otimes I)=0,\end{aligned}\tag{77}$$

and

$$(I\otimes R-I\otimes I\otimes I)((\tilde{Z}\otimes I)T-(I\otimes\tilde{Z})T)-(Z\otimes I)T-(\tilde{R}\otimes I)(I\otimes Z)T=0,\tag{78}$$

where we have used (25) to have (76),

$$\begin{aligned}\tilde{f}^a_b(S(p^m)S(p^n))=\tilde{f}^c_b(S(p^m))\tilde{f}^a_c(S(p^n))+\tilde{\eta}_b(S(p^m))\tilde{\eta}^a(S(p^n))\\ -\tilde{f}^a_b(S(p^k))\tilde{\eta}^c(S(\Lambda^m_k))\tilde{\eta}_c(S(p^n))-\delta^a_b\tilde{\eta}^c(S(p^m))\tilde{\eta}_c(S(p^n))\end{aligned}$$

obtained from (26) to have (77), and

$$\tilde{\eta}^a(S(p^m)S(p^n))=\tilde{\eta}^b(S(p^m))\tilde{f}^a_b(S(p^n))-\tilde{\eta}^a(S(p^k))\tilde{\eta}^b(S(\Lambda^m_k))\tilde{\eta}_b(S(p^n))$$

obtained from (28) to have (78). The consistency of (75)–(78) requires the following.

Lemma (4.2): The relations (75)–(78) are consistent if

$$T=-\tilde{R}T,\quad \lambda\neq 0,\quad -2.\tag{79}$$

Proof: Multiplying both sides of (75) and (76) from the left by $(\tilde{R}\otimes I+I\otimes I\otimes I)$ and both sides of (77) and (78) from the left by $(I\otimes\tilde{R}+I\otimes I\otimes I)$ and using the relation $(\tilde{R}+I\otimes I)(R-I\otimes I)=0$, we obtain the following consistency conditions:

$$(\tilde{R}\otimes I+I\otimes I\otimes I)(T\otimes I-(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T))=0,\tag{80}$$

$$(\tilde{R}\otimes I+I\otimes I\otimes I)(T\otimes I-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes T))=0,\tag{81}$$

$$(I\otimes\tilde{R}+I\otimes I\otimes I)(I\otimes T-(\tilde{R}\otimes I)(I\otimes\tilde{R})(T\otimes I))=0,\tag{82}$$

and

$$(I\otimes\tilde{R}+I\otimes I\otimes I)((Z\otimes I)T+(\tilde{R}\otimes I)(I\otimes Z)T)=0. \tag{83}$$

From (80) and (81), it follows

$$(\tilde{R}\otimes I+I\otimes I\otimes I)((I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T))-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes T))=0, \tag{84}$$

and from (67) and $(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes\tilde{R})=(\tilde{R}\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)$ obtained by multiplying both sides of (68) from the left and from the right by $(I\otimes R^{-1})(R^{-1}\otimes I)$, we can rewrite (84) as

$$((I\otimes\tilde{R})(\tilde{R}\otimes I)-(I\otimes R^{-1})(R^{-1}\otimes I))(I\otimes(\tilde{R}T+T))=0.$$

Multiplying it by $(R\otimes I)(I\otimes R)$ and using $R\tilde{R}=I\otimes I+I\otimes Q$, we obtain $(I\otimes I\otimes Q(2+Q))(I\otimes(\tilde{R}T+T))=0$, which is equivalent, by virtue of (65), to $\lambda(2+\lambda)(\tilde{R}+I\otimes I)T=0$ leading to (79) for $\lambda\neq 0$ or -2. From (79) and (67) we may see that the equation (82) is satisfied and it is also easy to see that the equation (83) is obtained by multiplying both sides of (72) from the right by T and by using (79). Q.E.D.

Remark (4.1): Note that for $\lambda=-2$ the characteristic equation (62) reduces to $(R-I\otimes I)^2=0$ showing that R does not possess negative eigenvalues and, therefore, in this case, we cannot construct from it antisymmetric products. [In the following we shall assume the relation (79) for $\lambda=-2$.]

Remark (4.2): For the case $\lambda=0$ we have $\tilde{R}=R=R^{-1}$ from which we see that the relations (80) and (81) are identical and cannot imply (79). However, by multiplying both sides of (64) by $\tilde{R}+I\otimes I$ and by using (62) we can deduce in this case the following consistency condition,

$$C^{nm}-\Lambda^n_k\Lambda^m_l C^{kl}=0, \tag{85}$$

where we have used (47) for $a=\Lambda$ and $C^{nm}=(\tilde{R}^{nm}_{kl}+\delta^n_k\delta^m_l)$. In fact, by applying $f^a_b\circ S$ and $\tilde{f}^a_b$ on both sides of (85), we get, respectively,

$$(\tilde{R}\otimes I+I\otimes I\otimes I)(T\otimes I)-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes\tilde{R}+I\otimes I\otimes I)(I\otimes T)=0$$

and

$$(I\otimes\tilde{R}+I\otimes I\otimes I)(I\otimes T)-(\tilde{R}\otimes I)(I\otimes\tilde{R})(\tilde{R}\otimes I+I\otimes I\otimes I)(T\otimes I)=0,$$

which are identical to the relations (81) and (82) by virtue of (67). To get the relation (83), we apply $\tilde{\eta}^a\circ S$ on both sides of (85) to have

$$(Z\otimes I)(\tilde{R}+I\otimes I)T+(\tilde{R}\otimes I)(I\otimes Z)(\tilde{R}+I\otimes I)T=0,$$

where we have used (24) and (36). By virtue of (72) we can see that this relation is equivalent to (83). Therefore, for $\lambda=0$ the consistency condition is not $RT=-T$ as assumed in remark (3.11) of Ref. 9 but (85).

Comparing now (75) with (76), we obtain

$$(R\otimes I-I\otimes I\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes Q\otimes I)(I\otimes T)-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes T)$$
$$=-(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T). \tag{86}$$

Replacing R by $\tilde{R}-I\otimes Q$ and using $(\tilde{R}\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)=(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes\tilde{R})$, (65) and (79), we can write the left-hand side as

$$-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes I\otimes I+I\otimes I\otimes Q+I\otimes Q\otimes I+I\otimes I\otimes Q^2)(I\otimes T),$$

which can be identified as the right-hand side of (86) by using (59) and (65). Therefore, (75) is equivalent to (76).

Remark (4.3): For $\lambda=0$ the relations (75) and (76) are identical, but for $\lambda=-2$ we must assume the relation (79) to have the equivalency between (75) and (76).

To investigate the equation (77), we multiply both sides of (71) from the left by $(R\otimes I)$ and from the right by Z to obtain

$$(\tilde{Z}\otimes I)\tilde{Z}+(I\otimes\tilde{R})(I\otimes\tilde{Z})\tilde{Z}=(R\otimes I)(I\otimes R)(Z\otimes I)Z+(R\otimes I)(I\otimes R)(\tilde{R}\otimes I)(I\otimes Z)Z,$$

where we have used $\tilde{Z}=-RZ$ and $R\tilde{R}=I\otimes I+I\otimes Q=\tilde{R}R=I\otimes I+Q\otimes I$ [due to (62) and (65)]. Now, multiplying both sides from the left by $I\otimes R-I\otimes I\otimes I$, we obtain

$$\begin{aligned}&(I\otimes R-I\otimes I\otimes I)((\tilde{Z}\otimes I)\tilde{Z}-(I\otimes\tilde{Z})\tilde{Z})\\&\quad=(I\otimes R-I\otimes I\otimes I)(R\otimes I)(I\otimes R)((Z\otimes I)Z+(\tilde{R}\otimes I)(I\otimes Z)Z),\end{aligned}$$

where we have used (68) and $(R-I\otimes I)\tilde{R}=-(R-I\otimes I)$. Taking into account this equation, (77) can be written as

$$\begin{aligned}&(I\otimes R-I\otimes I\otimes I)((R\otimes I)(I\otimes R)((Z\otimes I)Z+(\tilde{R}\otimes I)(I\otimes Z)Z)+(I\otimes I\otimes Q)(T\otimes I)\\&\quad-(I\otimes Q\otimes I)(I\otimes T))+I\otimes T-(\tilde{R}\otimes I)(I\otimes\tilde{R})(T\otimes I)=0.\end{aligned}$$

Multiplying both sides from the left by $-(I\otimes R^{-1})(R^{-1}\otimes I)$ and using $(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes R)=(R\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)$, we obtain

$$\begin{aligned}&(R\otimes I-I\otimes I\otimes I)((I\otimes Z)Z-(Z\otimes I)Z-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes I\otimes Q)(T\otimes I)+(I\otimes R^{-1})(R^{-1}\otimes I)\\&\quad\times(I\otimes Q\otimes I)(I\otimes T))-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes T)+(I\otimes R^{-1})(R^{-1}\otimes I)(\tilde{R}\otimes I)(I\otimes\tilde{R})(T\otimes I)\\&\quad=0=-(R\otimes I-I\otimes I\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes I\otimes Q)(T\otimes I)-T\otimes I\\&\qquad+(I\otimes R^{-1})(R^{-1}\otimes I)(\tilde{R}\otimes I)(I\otimes\tilde{R})(T\otimes I),\end{aligned}$$

where we have used (76) to have the right-hand side which gives, by using (56) into the third factor of the last term,

$$\begin{aligned}&-(R\otimes I-I\otimes I\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes I\otimes Q)(T\otimes I)+(I\otimes R^{-1})(I\otimes I\otimes Q)(T\otimes I)\\&\quad+(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes\tilde{R})(I\otimes I\otimes Q)(T\otimes I)=0.\end{aligned}$$

For $\lambda=0$ this relation is trivial, and for $\lambda\neq 0$ we use (79) and (59) into the second term and (56) into the third term to obtain

$$\begin{aligned}&-(R\otimes I-I\otimes I\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes I\otimes Q)(T\otimes I)-(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes I\otimes Q)(T\otimes I)\\&\quad+(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes R)(I\otimes I\otimes Q)(T\otimes I)=0,\end{aligned}$$

which is satisfied by virtue of $(I\otimes R^{-1})(R^{-1}\otimes I)(I\otimes R)=(R\otimes I)(I\otimes R^{-1})(R^{-1}\otimes I)$. Therefore, (77) is also equivalent to (75) and (76).

Now, we consider the consistency of the commutation rules (64) and the braiding properties of the product of three generators of $\mathcal{B}$. After the insertion of (64) into the first term of the right-hand side of (64) itself, we get

$$\begin{aligned}p\mathsf{T}p&=R^2(p\mathsf{T}p)-(R+I\otimes I)(R-I\otimes I)Zp+(R+I\otimes I)T-(\Lambda\mathsf{T}\Lambda)(R+I\otimes I)T\\&=p\mathsf{T}p-(I\otimes Q)((R-I\otimes I)((p\mathsf{T}p)-Zp)+T-(\Lambda\mathsf{T}\Lambda)T),\end{aligned}$$

where we have used into the first right-hand side (42) for $a=\Lambda$ to have the last term and (62), (56), and (79) to have the second right-hand side. We observe from (64) that the second term of the second right-hand side vanishes, implying the consistency of these commutation rules. For the product of three translations, we have

$$\begin{aligned} p\mathsf{T}p\mathsf{T}p &= (p\mathsf{T}p)\mathsf{T}p \\ &= (R\otimes I)(p\mathsf{T}(p\mathsf{T}p)) + r\mathsf{T}p \\ &= (R\otimes I)((I\otimes R)((p\mathsf{T}p)\mathsf{T}p) + p\mathsf{T}r) + r\mathsf{T}p \\ &= (R\otimes I)(I\otimes R)(R\otimes I)(p\mathsf{T}p\mathsf{T}p) + (R\otimes I)(I\otimes R)(r\mathsf{T}p) + (R\otimes I)(p\mathsf{T}r) + r\mathsf{T}p, \end{aligned}$$

where $r=-(R-I\otimes I)Zp+T-(\Lambda\mathsf{T}\Lambda)T$. On the other hand,

$$\begin{aligned} p\mathsf{T}p\mathsf{T}p = p\mathsf{T}(p\mathsf{T}p) &= (I\otimes R)((p\mathsf{T}p)\mathsf{T}p) + p\mathsf{T}r \\ &= (I\otimes R)((R\otimes I)(p\mathsf{T}(p\mathsf{T}p)) + r\mathsf{T}p) + p\mathsf{T}r \\ &= (I\otimes R)(R\otimes I)(I\otimes R)(p\mathsf{T}p\mathsf{T}p) + (I\otimes R)(R\otimes I)(p\mathsf{T}r) \\ &\quad + (I\otimes R)(r\mathsf{T}p) + p\mathsf{T}r. \end{aligned}$$

The Yang–Baxter equation (66) implies

$$A(r\mathsf{T}p) = B(p\mathsf{T}r), \tag{87}$$

where

$$A = (R\otimes I)(I\otimes R) - I\otimes R + I\otimes I\otimes I,$$

$$B = (I\otimes R)(R\otimes I) - R\otimes I + I\otimes I\otimes I.$$

Replacing on both sides of (87) r by $-(R-I\otimes I)Zp+T-(\Lambda\mathsf{T}\Lambda)T$, we obtain

$$\begin{aligned} &-A(R\otimes I - I\otimes I\otimes I)(Z\otimes I)(p\mathsf{T}p) + A(T\otimes I)p - A(\Lambda\mathsf{T}\Lambda\mathsf{T}p)T \\ &\quad = -B(I\otimes R - I\otimes I\otimes I)(I\otimes Z)(p\mathsf{T}p) + B(I\otimes T)p - B(p\mathsf{T}\Lambda\mathsf{T}\Lambda)T. \end{aligned} \tag{88}$$

Using (52) $[p\mathsf{T}\Lambda = \tilde{R}(\Lambda\mathsf{T}p) + Z\Lambda - (\Lambda\mathsf{T}\Lambda)Z]$, we can rewrite the third term of the right-hand side of (88) as

$$\begin{aligned} B(p\mathsf{T}\Lambda\mathsf{T}\Lambda)T &= B(\tilde{R}\otimes I)(\Lambda\mathsf{T}p\mathsf{T}\Lambda)T + B(Z\otimes I)(\Lambda\mathsf{T}\Lambda)T - B(\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)T \\ &= B(\tilde{R}\otimes I)(I\otimes\tilde{R})(\Lambda\mathsf{T}\Lambda\mathsf{T}p)T + B(\tilde{R}\otimes I)(I\otimes Z)(\Lambda\mathsf{T}\Lambda)T \\ &\quad - B(\tilde{R}\otimes I)(\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(I\otimes Z)T + B(Z\otimes I)(\Lambda\mathsf{T}\Lambda)T - B(\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(Z\otimes I)T. \end{aligned} \tag{89}$$

A straightforward computation shows that

$$B(\tilde{R}\otimes I)(I\otimes\tilde{R}) = A + (I\otimes I\otimes Q)(\tilde{R}\otimes I + I\otimes I\otimes I), \tag{90}$$

where we have used $R\tilde{R} = I\otimes I + I\otimes Q$ and (56). Multiplying both sides of (90) from the right by $(\Lambda\mathsf{T}\Lambda\mathsf{T}p)T$ and using (47) for $a=\Lambda$ and (79), we get

$$A(\Lambda\mathsf{T}\Lambda\mathsf{T}p)T = B(\tilde{R}\otimes I)(I\otimes\tilde{R})(\Lambda\mathsf{T}\Lambda\mathsf{T}p)T.$$

We see from (90) that for $\lambda=0$ this relation is satisfied without using (79). On the other hand, by using (42) and (47) for $a=\Lambda$, we can rewrite the third and the fifth terms of the right-hand side of (89) as

$$(\Lambda\top\Lambda\top\Lambda)B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))T.$$

Using $\tilde{Z}=-RZ$, $R\tilde{R}=\tilde{R}R=I\otimes I+I\otimes Q$, and (56) we get

$$B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z)T=-(I\otimes R-I\otimes I\otimes I)(\tilde{Z}\otimes I-I\otimes\tilde{Z})T+(Z\otimes I)T+(\tilde{R}\otimes I)(I\otimes Z)T, \tag{91}$$

which vanishes by virtue of (78). Therefore (88) reduces to

$$\begin{aligned}&-A(R\otimes I-I\otimes I\otimes I)(Z\otimes I)(p\top p)+A(T\otimes I)p\\&\quad=-B(I\otimes R-I\otimes I\otimes I)(I\otimes Z)(p\top p)+B(I\otimes T)p-B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(\Lambda\top\Lambda)T.\end{aligned} \tag{92}$$

Using now

$$A(R\otimes I-I\otimes I\otimes I)=B(I\otimes R-I\otimes I\otimes I)=\Lambda_3,$$

we may rewrite (92) as

$$\Lambda_3(Z\otimes I-I\otimes Z)(p\top p)-A(T\otimes I)p+B(I\otimes T)p-B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(\Lambda\top\Lambda)T=0. \tag{93}$$

From the properties of Λ_3 and $(R-I\otimes I)\tilde{R}=-(R-I\otimes I)$, we have

$$\begin{aligned}\Lambda_3(Z\otimes I-I\otimes Z)(p\top p)&=\Lambda_3(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(p\top p)\\&=B(I\otimes R)(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(p\top p)-B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(p\top p).\end{aligned}$$

Using (64) in the second term of the right-hand side, we obtain

$$\begin{aligned}&B(I\otimes R)(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(p\top p)-B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))R(p\top p)\\&\quad+B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(R-I\otimes I)Zp-B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))T\\&\quad+B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(\Lambda\top\Lambda)T.\end{aligned}$$

The first and the second terms vanish by virtue of (71) and the fourth term vanishes by virtue of (91). Therefore, (93) reduces to

$$+B(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))(R-I\otimes I)Zp-A(T\otimes I)p+B(I\otimes T)p,$$

which can be written, after making of use of (71), as

$$\begin{aligned}&+B(I\otimes R-I\otimes I\otimes I)(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))Zp-A(T\otimes I)p+B(I\otimes T)p\\&\quad=\Lambda_3(Z\otimes I+(\tilde{R}\otimes I)(I\otimes Z))Zp-A(T\otimes I)p+B(I\otimes T)p\\&\quad=A(R\otimes I-I\otimes I\otimes I)((Z\otimes I)Z-(I\otimes Z)Z)p-A(T\otimes I)p+B(I\otimes T)p,\end{aligned}$$

which reduces, due to (75), to

$$-A(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T)p+B(I\otimes T)p,$$

which vanishes according to the same calculation of (90).

The same procedure is applied to $p\mathsf{T}p\mathsf{T}\Lambda$ to have

$$
\begin{aligned}
p\mathsf{T}p\mathsf{T}\Lambda = p\mathsf{T}(p\mathsf{T}\Lambda) = {} & (I\otimes\tilde{R})((p\mathsf{T}\Lambda)\mathsf{T}p) + (I\otimes Z)(p\mathsf{T}\Lambda) - ((p\mathsf{T}\Lambda)\mathsf{T}\Lambda)Z \\
= {} & (I\otimes\tilde{R})(\tilde{R}\otimes I)(\Lambda\mathsf{T}(p\mathsf{T}p)) + (I\otimes\tilde{R})(Z\otimes I)(\Lambda\mathsf{T}p) - (I\otimes\tilde{R})(\Lambda\mathsf{T}\Lambda\mathsf{T}p)Z \\
& + (I\otimes Z)\tilde{R}(\Lambda\mathsf{T}p) + (I\otimes Z)Z\Lambda - (I\otimes Z)(\Lambda\mathsf{T}\Lambda)Z - (\tilde{R}\otimes I)(\Lambda\mathsf{T}(p\mathsf{T}\Lambda))Z \\
& - (Z\otimes I)(\Lambda\mathsf{T}\Lambda)Z + (\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(Z\otimes I)Z \\
= {} & (I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes R)(\Lambda\mathsf{T}p\mathsf{T}p) - (I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes R - I\otimes I\otimes I)(I\otimes Z) \\
& \times(\Lambda\mathsf{T}p) + (I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T)\Lambda - (I\otimes\tilde{R})(\tilde{R}\otimes I)(\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(I\otimes T) \\
& + (I\otimes\tilde{R})(Z\otimes I)(\Lambda\mathsf{T}p) - (I\otimes\tilde{R})(\Lambda\mathsf{T}\Lambda\mathsf{T}p)Z + (I\otimes Z)\tilde{R}(\Lambda\mathsf{T}p) \\
& + (I\otimes Z)Z\Lambda - (I\otimes Z)(\Lambda\mathsf{T}\Lambda)Z - (\tilde{R}\otimes I)(I\otimes\tilde{R})(\Lambda\mathsf{T}\Lambda\mathsf{T}p)Z - (\tilde{R}\otimes I) \\
& \times(I\otimes Z)(\Lambda\mathsf{T}\Lambda)Z + (\tilde{R}\otimes I)(\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(I\otimes Z)Z - (Z\otimes I)(\Lambda\mathsf{T}\Lambda)Z \\
& + (\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(Z\otimes I)Z.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
p\mathsf{T}p\mathsf{T}\Lambda = (p\mathsf{T}p)\mathsf{T}\Lambda = {} & (R\otimes I)(p\mathsf{T}(p\mathsf{T}\Lambda)) - (R\otimes I - I\otimes I\otimes I)(Z\otimes I)(p\mathsf{T}\Lambda) \\
& + (T\otimes I)\Lambda - (\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(T\otimes I) \\
= {} & (R\otimes I)(I\otimes\tilde{R})((p\mathsf{T}\Lambda)\mathsf{T}p) + (R\otimes I)(I\otimes Z)(P\mathsf{T}\Lambda) - (R\otimes I) \\
& \times((p\mathsf{T}\Lambda)\mathsf{T}\Lambda)Z - (R\otimes I - I\otimes I\otimes I)(Z\otimes I)\tilde{R}(\Lambda\mathsf{T}p) - (R\otimes I - I \\
& \otimes I\otimes I)(Z\otimes I)Z\Lambda + (R\otimes I - I\otimes I\otimes I)(Z\otimes I)(\Lambda\mathsf{T}\Lambda)Z + (T\otimes I)\Lambda \\
& - (\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(T\otimes I) \\
= {} & (R\otimes I)(I\otimes\tilde{R})(\tilde{R}\otimes I)(\Lambda\mathsf{T}p\mathsf{T}p) + (R\otimes I)(I\otimes\tilde{R})(Z\otimes I)(\Lambda\mathsf{T}p) \\
& - (R\otimes I)(I\otimes\tilde{R})(\Lambda\mathsf{T}\Lambda\mathsf{T}p)Z + (R\otimes I) \\
& \times(I\otimes Z)\tilde{R}(\Lambda\mathsf{T}p) + (R\otimes I)(I\otimes Z)Z\Lambda - (R\otimes I)(I\otimes Z)(\Lambda\mathsf{T}\Lambda)Z \\
& - (R\otimes I)(\tilde{R}\otimes I)(\Lambda\mathsf{T}p\mathsf{T}\Lambda)Z - (R\otimes I)(Z\otimes I)(\Lambda\mathsf{T}\Lambda)Z + (R\otimes I) \\
& \times(\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(Z\otimes I)Z - (R\otimes I - I\otimes I\otimes I)(Z\otimes I)\tilde{R}(\Lambda\mathsf{T}p) \\
& - (R\otimes I - I\otimes I\otimes I)(Z\otimes I)Z\Lambda + (R\otimes I - I\otimes I\otimes I)(Z\otimes I) \\
& \times(\Lambda\mathsf{T}\Lambda)Z + (T\otimes I)\Lambda - (\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(T\otimes I).
\end{aligned}
$$

Using again (52), we can replace $(R\otimes I)(\tilde{R}\otimes I)(\Lambda\mathsf{T}p\mathsf{T}\Lambda)Z$ by

$$
\begin{aligned}
& (R\otimes I)(\tilde{R}\otimes I)(I\otimes\tilde{R})(\Lambda\mathsf{T}\Lambda\mathsf{T}p)Z + (R\otimes I)(\tilde{R}\otimes I)(I\otimes Z)(\Lambda\mathsf{T}\Lambda)Z \\
& - (R\otimes I)(\tilde{R}\otimes I)(\Lambda\mathsf{T}\Lambda\mathsf{T}\Lambda)(I\otimes Z)Z.
\end{aligned}
$$

We remark that (69) implies the equality between the coefficients multiplying $\Lambda\mathsf{T}p\mathsf{T}p$.

For the coefficient multiplying $\Lambda\mathsf{T}p$, we must have the following relation,

$$(R\otimes I)(I\otimes\tilde{R})(Z\otimes I)+(R\otimes I)(I\otimes Z)\tilde{R}-(R\otimes I-I\otimes I\otimes I)(Z\otimes I)\tilde{R}$$
$$=-(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes R-I\otimes I\otimes I)(I\otimes Z)+(I\otimes\tilde{R})(Z\otimes I)+(I\otimes Z)\tilde{R},$$

which can be written under the form

$$(R\otimes I-I\otimes I\otimes I)((I\otimes\tilde{R})(Z\otimes I)-(Z\otimes I)\tilde{R}+(I\otimes Z)\tilde{R}+(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes Z))=0$$

where we have used (69). By using $(R-I\otimes I)\tilde{R}=-(R-I\otimes I)$ we can replace the third term of the second factor by $-(\tilde{R}\otimes I)(I\otimes Z)\tilde{R}$ to see that this relation is satisfied by virtue of (72). It is also easy to see that the coefficients multiplying Λ are precisely the relation (75) and those multiplying $(\Lambda\top\Lambda)Z$ and $(\Lambda\top\Lambda\top p)Z$ are equal in virtue of $R\tilde{R}=I\otimes I+I\otimes Q$ and (56). Using (42) and (47) for $a=\Lambda$, we obtain for the coefficients multiplying $\Lambda\top\Lambda\top\Lambda$ the same relation as (75) yielding the consistency of the braiding of $p\top p\top\Lambda$.

Finally by using $\Lambda\top\Lambda=R(\Lambda\top\Lambda)R^{-1}$, obtained from (42) for $a=\Lambda$, and (71) we can show by a similar way the consistency of the braiding of $p\top\Lambda\top\Lambda$.

Then from the results of this section, we have the following.

Theorem (4.1): Let $\mathcal{G}$ be an inhomogeneous quantum group, as defined in Sec. II, with the following commutation rules

$$R(\Lambda\top\Lambda)=(\Lambda\top\Lambda)R,$$

$$p\top\Lambda=\tilde{R}(\Lambda\top p)+Z\Lambda-(\Lambda\top\Lambda)Z,$$

$$p\top p=R(p\top p)-(R-I\otimes I)Zp+T-(\Lambda\top\Lambda)T.$$

Then,

$$\tilde{R}=R+I\otimes Q=R^{-1}+R^{-1}(I\otimes Q),$$

$$Q=\lambda I,\quad \lambda\in C,\quad \lambda\neq-1,$$

$$T=-\tilde{R}T\quad\text{for }\lambda\neq0\quad\text{or}\quad(\tilde{R}+I\otimes I)T-(\Lambda\top\Lambda)(\tilde{R}+I\otimes I)T=0\quad\text{for }\lambda=0,$$

$$(R\otimes I)(I\otimes R)(R\otimes I)=(I\otimes R)(R\otimes I)(I\otimes R),$$

$$(Z\otimes I)R+(\tilde{R}\otimes I)(I\otimes Z)R=(I\otimes R)(Z\otimes I)+(I\otimes R)(\tilde{R}\otimes I)(I\otimes Z),$$

$$(R\otimes I-I\otimes I\otimes I)\quad((I\otimes Z)Z-(Z\otimes I)Z)+T\otimes I-(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T)=0$$

and

$$(I\otimes R)-I\otimes I\otimes I)((\tilde{Z}\otimes I)T-(I\otimes\tilde{Z})T)-(Z\otimes I)T-(\tilde{R}\otimes I)(I\otimes Z)T=0$$

with

$$\tilde{Z}=-RZ.$$

V. DISCUSSIONS AND CONCLUSIONS

We end this paper by noticing the following.

(1) Let us recall that (10) and (11) imply that f_l^n, f_{kl}^{nm}, f_l^{nm}, and f_{kl}^n are linear functionals on $\mathcal{B}$, but not $\tilde{f}_k^m$, $\tilde{\eta}^m$, and $\tilde{\eta}_k$. Then their action on both sides of inhomogeneous commutation relations (48) and (64) is not justified, but a tedious and straightforward computation shows that they give the same results that those obtained directly by using f_{kl}^{nm}, f_l^{nm}, and f_{kl}^n.

(2) From (35), (36), and (24), we see that $\mathcal{A}\ni a\rightarrow\tilde{\rho}(a)=\left(\begin{smallmatrix}\tilde{f}(a) & \tilde{\eta}(a)\\ 0 & \varepsilon(a)\end{smallmatrix}\right)\in M_N(C)$ is a unital anti-homomorphism.

(3) Replacing a and $b\in\mathcal{A}$, respectively, by $S(a)$ and $S(b)$ into (36) and by setting $\eta^k=\tilde{\eta}^k\circ S$, we get

$$\eta^n(ab)=\eta^n(a)\varepsilon(b)+\tilde{f}^n_m(S(a))\eta^m(b),\quad a,b\in\mathcal{A}. \tag{94}$$

Then, although the formalism presented above is quite different from those of Ref. 9, we arrive at similar results for the commutation rules between the elements of $\mathcal{A}$ and the translations (48) and (94). These formulas become identical with those of Ref. 9 if $\tilde{f}^n_m(S(a))=f^n_m(a)$ for any $a\in\mathcal{A}$. The latter condition is satisfied for $\tilde{\eta}_n(p^a)=0$ and it is true, as a consequence of (56) and (57), for any $a\in\mathcal{B}$. In this case, (34), (35), and (94) are also true for any a and $b\in\mathcal{B}$ and (94) can be combined with (34) to see that $\mathcal{A}\ni a\rightarrow\rho(a)=\left(\begin{smallmatrix}f(a) & \eta(a)\\ 0 & \varepsilon(a)\end{smallmatrix}\right)\in M_N(C)$ is a unital homomorphism. In this case the R-matrix is subject to a strong constraint $R^2=I^{\otimes 2}$ as in Ref. 9.

(4) Using (85) and setting $T=-2T'$ we can rewrite, for $\lambda=0$, the commutation rules between the translations as

$$(R^{nm}_{kl}-\delta^n_k\delta^m_l)(p^kp^l-Z^{kl}_qp^q+T'^{kl}-\Lambda^k_q\Lambda^l_pT'^{qp})=0,$$

which are identical to those of Ref. 9. In this case one can also see that (75) reduces to the condition of the existence of covariant differential calculus on a quantum Minkowski space found in Ref. 10.

ACKNOWLEDGMENTS

We are grateful to M. Dubois-Violette, A. B. Hammou, and A. E. K. Yanallah for valuable discussions

[1] O. Ogiesvestsky, W. B. Schmidke, J. Wess, and B. Zumino, Commun. Math. Phys. **150**, 495 (1992).
[2] S. Majid, J. Math. Phys. **34**, 2045 (1993).
[3] M. Chaichian and A. P. Demichev, Phys. Lett. B **304**, 220 (1993).
[4] V. K. Dobrev, J. Phys. A **26**, 1317 (1993).
[5] J. Lukierski, A. Nowicki, H. Ruegg, and V. N. Tolstoy, Phys. Lett. B **264**, 331 (1991).
[6] J. Lukierski and A. Nowicki, Phys. Lett. B **279**, 299 (1992); M. Schliecker, W. Weich, and R. Weixler, Z. Phys. C **53**, 79 (1992); Lett. Math. Phys. **27**, 217 (1993); L. Castellani, *ibid.* **30**, 233 (1994).
[7] S. Zakrzewski, J. Phys. A **27**, 2075 (1994).
[8] L. Castellani, Commun. Math. Phys. **171**, 383 (1995); P. Aschieri and L. Castellani, Int. J. Mod. Phys. A **11**, 4513 (1996).
[9] P. Podles and S. L. Woronowicz, Commun. Math. Phys. **185**, 325 (1997), **178**, 61 (1996); ''Inhomogeneous quantum groups,'' in Proceedings of First Caribbean School of Mathematics and Physics in Guadeloupe (1993); M. Chaichian and A. Demichev, *Quantum Groups* (World, Scientific, Singapore, 1996).
[10] P. Podles, Commun. Math. Phys. **181**, 569 (1996).
[11] S. L. Woronowicz, Commun. Math. Phys. **122**, 125 (1989).
[12] U. Carow-Watamura, M. Schliecker, S. Watamura, and W. Weich, Commun. Math. Phys. **142**, 605 (1991); M. Lagraa, Int. J. Mod. Phys. A **11**, 699 (1996).

Bäcklund transformation and its superposition principle of a Blaszak–Marciniak four-field lattice

Wen-Xiu Ma
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, People's Republic of China

Xing-Biao Hu[a)]
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China and State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific Engineering Computing, Academia Sinica, P.O. Box 2719, Beijing 100080, People's Republic of China

Si-Ming Zhu
Department of Mathematics, Zhongshan University, Guangzhou 510275, People's Republic of China

Yong-Tang Wu
Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong, People's Republic of China

(Received 2 December 1998; accepted for publication 2 June 1999)

A four-field lattice furnished by Blaszak and Marciniak is transformed into bilinear form upon introducing two auxiliary independent variables. A Bäcklund transformation in bilinear form is found for the lattice and the corresponding nonlinear superposition formula is rigorously established. As a consequence, soliton solutions to the lattice are derived. © *1999 American Institute of Physics.* [S0022-2488(99)02811-X]

I. INTRODUCTION

Recently, Blaszak and Marciniak have derived several three-field and four-field lattices[1] as an application of r-matrix formalism to the algebra of shift operators. Two examples are

$$a_t(n)=c(n+1)-c(n-1), \tag{1}$$

$$b_t(n)=a(n-1)c(n-1)-a(n)c(n), \tag{2}$$

$$c_t(n)=c(n)(b(n)-b(n+1)), \tag{3}$$

and

$$u_t(n)=u(n)(v(n)-v(n-1)), \tag{4}$$

$$v_t(n)=w(n)u(n+1)-u(n)w(n-1), \tag{5}$$

$$w_t(n)=q(n)u(n+2)-u(n)q(n-1), \tag{6}$$

$$q_t(n)=u(n+3)-u(n), \tag{7}$$

both of which have Abelian symmetry algebras of infinite dimensions. Recently, an integrable symplectic map connected with (1)–(3) and its hierarchy was obtained by Wu and Geng[2] and

[a)]Electronic mail: hxb@lsec.cc.ac.cn

master symmetries were presented by the discrete zero curvature equation by one of the authors (Ma) and Fuchssteiner.[3] Moreover, (1)–(3) were transformed into bilinear equations by introducing an auxiliary independent variable, and thus a Bäcklund transformation and its nonlinear superposition formula were established by one of the authors (X.-B.H.) and Zhu.[4] As a result, soliton solutions to Eqs. (1)–(3) were found.

However, to the best of our knowledge, so far there have not been any results on solutions of the lattice (4)–(7) in the literature. In this paper, we would like to present a way to construct solutions to the Eqs. (4)–(7) by establishing a Bäcklund transformation and its nonlinear superposition formula for the lattice (4)–(7). As an application of the obtained results, soliton solutions of (4)–(7) are derived. The basic tool used in this paper is Hirota's bilinear formalism. As usual, the crucial step of using Hirota's method is to transform the system of equations under consideration into bilinear form, which is far from algorithmic and often highly technical. In Sec. II, through a long computation, we will show a way of deriving a bilinear form for the lattice (4)–(7). Then a Bäcklund transformation in bilinear form will be presented in Sec. III and the corresponding nonlinear superposition formula will be established in Sec. IV. Finally in Sec. V, a conclusion will be given. Some bilinear operator identities, which are fundamental and necessary for our discussion, are listed in Appendix A.

II. BILINEAR FORM

In this section, we want to derive a bilinear form for the lattice (4)–(7). To that end, let us make

$$u(n)=\frac{f(n+1)f(n-1)}{f^2(n)},\qquad v(n)=\left(\ln\frac{f(n+1)}{f(n)}\right)_t. \tag{8}$$

Our choice of the above-mentioned transformation comes from an observation that the first Eq. (4) of the lattice can be transformed into the following form:

$$(\ln u(n))_t=v(n)-v(n-1).$$

Furthermore let us introduce an auxiliary independent variable z such that

$$(D_tD_z-2e^{D_n}+2)f(n)\cdot f(n)=0, \tag{9}$$

where Hirota's bilinear differential operator $D_x^mD_t^k$, bilinear difference operator $\exp(\delta D_n)$ and bilinear differential-difference operator $D_x^mD_t^k\exp(\delta D_n)$ are defined as follows:[5–9]

$$D_x^mD_t^k a\cdot b\equiv\left(\frac{\partial}{\partial x}-\frac{\partial}{\partial x'}\right)^m\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial t'}\right)^k a(x,t)b(x',t')|_{x'=x,t'=t},$$

$$\exp(\delta D_n)a(n)\cdot b(n)\equiv\exp\left[\delta\left(\frac{\partial}{\partial n}-\frac{\partial}{\partial n'}\right)\right]a(n)b(n')|_{n'=n}=a(n+\delta)b(n-\delta),$$

$$D_x^mD_t^k\exp(\delta D_n)a(n)\cdot b(n)\equiv\left(\frac{\partial}{\partial x}-\frac{\partial}{\partial x'}\right)^m\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial t'}\right)^k a(n+\delta,x,t)b(n-\delta,x',t')|_{x'=x,t'=t}.$$

Then from (7) to (9), we know that

$$q_t(n)=\frac{f(n+4)f(n+2)}{f^2(n+3)}-\frac{f(n+1)f(n-1)}{f^2(n)}=\frac{\partial^2}{\partial t\partial z}\ln\frac{f(n+3)}{f(n)},$$

which implies that we can choose

$$q(n)=\frac{\partial}{\partial z}\ln\frac{f(n+3)}{f(n)}. \tag{10}$$

From here it is clear that the introduction of auxiliary variable z makes it easy to solve $q(n)$ in terms of f without containing any explicit integral. That is also our motivation for introducing the auxiliary variable z. Substituting (8) and (10) into Eq. (5) allows us to take

$$(\ln f(n+1))_{tt}=w(n)u(n+1),$$

from which it follows that

$$w(n)=\frac{1}{2}\,\frac{D_t^2 f(n+1)\cdot f(n+1)}{f(n+2)f(n)}. \tag{11}$$

Furthermore from (6) we have

$$\frac{1}{2}\,\frac{D_t(D_t^2 f(n+1)\cdot f(n+1))\cdot f(n+2)f(n)}{f^2(n+2)f^2(n)}=\frac{f(n+1)D_z f(n+3)\cdot f(n)}{f^2(n+2)f(n)} - \frac{f(n+1)D_z f(n+2)\cdot f(n-1)}{f^2(n)f(n+2)}$$

or

$$\begin{aligned}&\tfrac{1}{2}D_t(D_t^2 f(n)\cdot f(n))\cdot(e^{D_n}f(n)\cdot f(n))\\&\quad=2\sinh(\tfrac{1}{2}D_n)(D_z e^{(3/2)D_n}f(n)\cdot f(n))\cdot(e^{(1/2)D_n}f(n)\cdot f(n)).\end{aligned} \tag{12}$$

By use of (9) and (A1)–(A3), we can compute that

$$\begin{aligned}D_t(D_t^2 f(n)\cdot f(n))\cdot(e^{D_n}f(n)\cdot f(n))&=D_t(D_t^2 f(n)\cdot f(n))\cdot f^2(n)+\tfrac{1}{6}D_t(D_t^3 D_z f(n)\cdot f(n))\cdot f^2(n)\\&\quad-\tfrac{2}{3}\sinh(\tfrac{1}{2}D_n)[(D_t^3 e^{(1/2)D_n}f(n)\cdot f(n))\cdot(e^{(1/2)D_n}f(n)\cdot f(n))\\&\quad+3(D_t e^{(1/2)D_n}f(n)\cdot f(n))\cdot(D_t^2 e^{(1/2)D_n}f(n)\cdot f(n))].\end{aligned} \tag{13}$$

Now from (12) and by use of (13), (A4), and (A5) we can have the following relation:

$$\begin{aligned}&2\sinh(\tfrac{1}{2}D_n)(D_z e^{(3/2)D_n}f(n)\cdot f(n))\cdot(e^{(1/2)D_n}f(n)\cdot f(n))\\&\quad=\tfrac{3}{8}D_t(D_t^2 f(n)\cdot f(n))\cdot f^2(n)+\tfrac{1}{16}D_t(D_t^3 D_z f(n)\cdot f(n))\cdot f^2(n)\\&\qquad+\tfrac{1}{8}D_t(D_t^2 e^{D_n}f(n)\cdot f(n))\cdot f^2(n)-\tfrac{1}{2}\sinh(\tfrac{1}{2}D_n)(D_t^3 e^{(1/2)D_n}f(n)\cdot f(n))\\&\qquad\cdot(e^{(1/2)D_n}f(n)\cdot f(n))+\tfrac{1}{2}\{\sinh(\tfrac{1}{2}D_n)(D_t D_y e^{(1/2)D_n}f(n)\cdot f(n))\\&\qquad\cdot(e^{(1/2)D_n}f(n)\cdot f(n))-\tfrac{1}{2}D_t(D_y e^{D_n}f(n)\cdot f(n))\cdot f^2(n)\},\end{aligned} \tag{14}$$

where we have introduced another auxiliary independent variable y such that

$$D_t^2 e^{(1/2)D_n}f(n)\cdot f(n)=D_y e^{(1/2)D_n}f(n)\cdot f(n).$$

Finally, (14) can be decoupled into the following two bilinear equations:

$$(D_t^2 e^{D_n}+3D_t^2+\tfrac{1}{2}D_t^3 D_z-2D_y e^{D_n})f(n)\cdot f(n)=0,$$

$$(D_t^3 e^{(1/2)D_n} + 4D_z e^{(3/2)D_n} - D_y D_t e^{(1/2)D_n}) f(n)\cdot f(n) = 0.$$

To sum up, we obtain the following bilinear form for the lattice (4)–(7):

$$(D_t D_z - 2e^{D_n} + 2) f(n)\cdot f(n) = 0, \tag{15}$$

$$D_t^2 e^{(1/2)D_n} f(n)\cdot f(n) = D_y e^{(1/2)D_n} f(n)\cdot f(n), \tag{16}$$

$$(D_t^2 e^{D_n} + 3D_t^2 + \tfrac{1}{2} D_t^3 D_z - 2D_y e^{D_n}) f(n)\cdot f(n) = 0, \tag{17}$$

$$(D_t^3 e^{(1/2)D_n} + 4D_z e^{(3/2)D_n} - D_y D_t e^{(1/2)D_n}) f(n)\cdot f(n) = 0. \tag{18}$$

III. BÄCKLUND TRANSFORMATION

In this section, we derive a bilinear Bäcklund transformation (BT) for the system of the bilinear Eqs. (15)–(18). The concrete form of BT is presented as follows:

Theorem 1: The bilinear system of Eqs. (15)–(18) has the following Bäcklund transformation:

$$(D_z + \lambda^{-1} e^{-D_n} + \mu) f(n)\cdot g(n) = 0, \tag{19}$$

$$(D_t e^{-(1/2)D_n} - \lambda e^{(1/2)D_n} + \gamma e^{-(1/2)D_n}) f(n)\cdot g(n) = 0, \tag{20}$$

$$(D_y e^{-(1/2)D_n} - \lambda D_t e^{(1/2)D_n} - \lambda\gamma e^{(1/2)D_n} + \omega e^{-(1/2)D_n}) f(n)\cdot g(n) = 0, \tag{21}$$

$$\begin{aligned}(\lambda^{-1} D_t^3 e^{-(1/2)D_n} &+ D_t^2 e^{(1/2)D_n} + 2\gamma D_t e^{(1/2)D_n} + 4\gamma^2 e^{(1/2)D_n} + 3\lambda^{-1}\gamma D_t^2 e^{-(1/2)D_n} - 2D_y e^{(1/2)D_n} \\ &- 2\omega e^{(1/2)D_n} + \nu e^{-(1/2)D_n}) f(n)\cdot g(n) = 0,\end{aligned} \tag{22}$$

$$\begin{aligned}(2D_t^3 e^{-D_n} &+ 2D_t D_y e^{-D_n} + 6\gamma D_t^2 e^{-D_n} + \tfrac{1}{2}\lambda\nu e^{-D_n} + 8\lambda^{-1} e^{-2D_n} + 2\omega D_t e^{-D_n} + 2\gamma D_y e^{-D_n} \\ &+ 4\lambda^2 D_t e^{D_n} + 6\gamma^2 D_t e^{-D_n} + 2\omega\gamma e^{-D_n} + 3\gamma^3 e^{-D_n} + \theta e^{D_n}) f(n)\cdot g(n) = 0,\end{aligned} \tag{23}$$

where $\lambda,\ \mu,\ \gamma,\ \omega,\ \nu$, and θ are arbitrary constants.

Proof: Let $f(n)$ be a solution of Eqs. (15)–(18). What we need to prove is that the function $g(n)$ satisfying (19)–(23) is another solution of Eqs. (15)–(18), i.e.,

$$P_1 \equiv (D_z D_t - 2e^{D_n} + 2) g(n)\cdot g(n) = 0,$$

$$P_2 \equiv (D_t^2 e^{(1/2)D_n} - D_y e^{(1/2)D_n}) g(n)\cdot g(n) = 0,$$

$$P_3 \equiv (D_t^2 e^{D_n} + 3D_t^2 + \tfrac{1}{2} D_t^3 D_z - 2D_y e^{D_n}) g(n)\cdot g(n) = 0,$$

$$P_4 \equiv (D_t^3 e^{(1/2)D_n} + 4D_z e^{(3/2)D_n} - D_y D_t e^{(1/2)D_n}) g(n)\cdot g(n) = 0.$$

In fact, a similar deduction to that in Refs. 10 and 11 can give rise to $P_1=0$, $P_2=0$. Thus we focus on showing that $P_3=0$ and $P_4=0$. Let us first prove that $P_3=0$. Making use of (A6) and (A7) can give

$$\begin{aligned}-2P_3 f^2(n) = 2[&(D_t^2 e^{D_n} + 3D_t^2 + \tfrac{1}{2} D_t^3 D_z - 2D_y e^{D_n}) f\cdot f] g^2 \\ &- 2[(D_t^2 e^{D_n} + 3D_t^2 + \tfrac{1}{2} D_t^3 D_z - 2D_y e^{D_n}) g\cdot g] f^2 + 3[(D_z D_t - 2e^{D_n} + 2) f\cdot f] \\ &\times (D_t^2 g\cdot g) - 3(D_t^2 f\cdot f)[(D_z D_t - 2e^{D_n} + 2) g\cdot g]\end{aligned}$$

$$
\begin{aligned}
&= I_1 - 8\sinh(\tfrac{1}{2}D_n)[(D_y e^{(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) - (e^{(1/2)D_n} f\cdot g)\\
&\quad \cdot(D_y e^{-(1/2)D_n} f\cdot g)] - 4D_t\cosh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g)\\
&\quad -(e^{(1/2)D_n} f\cdot g)\cdot(D_t e^{-(1/2)D_n} f\cdot g)] - 4\sinh(\tfrac{1}{2}D_n)[(D_t^2 e^{(1/2)D_n} f\cdot g)\\
&\quad \cdot(e^{-(1/2)D_n} f\cdot g) - 2(D_t e^{(1/2)D_n} f\cdot g)\cdot(D_t e^{-(1/2)D_n} f\cdot g)\\
&\quad +(e^{(1/2)D_n} f\cdot g)\cdot(D_t^2 e^{-(1/2)D_n} f\cdot g)],
\end{aligned}
\tag{24}
$$

where the function I_1 is defined by

$$
\begin{aligned}
I_1 &\equiv [(D_t^3 D_z + 6D_t^2 + 6D_t^2 e^{D_n}) f\cdot f] g^2 - f^2 (D_t^3 D_z + 6D_t^2 + 6D_t^2 e^{D_n}) g\cdot g\\
&\quad + 3[(D_z D_t - 2e^{D_n} + 2) f\cdot f](D_t^2 g\cdot g) - 3(D_t^2 f\cdot f)[(D_z D_t - 2e^{D_n} + 2) g\cdot g].
\end{aligned}
$$

Using a similar deduction to that in Ref. 12, we have

$$
\begin{aligned}
I_1 &= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)(D_t^3 e^{-(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) + 12\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{(1/2)D_n} f\cdot g)\\
&\quad \cdot(e^{-(1/2)D_n} f\cdot g) + 24\gamma\sinh(\tfrac{1}{2}D_n)(D_t e^{(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) + 24\gamma^2\sinh(\tfrac{1}{2}D_n)\\
&\quad \times(e^{(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) + 12\lambda^{-1}\gamma\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{-(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g).
\end{aligned}
$$

Thus by using (A8)–(A11), (20), and (21) equality (24) can be further reduced to the following:

$$
\begin{aligned}
-2P_3 f^2(n) &= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)(D_t^3 e^{-(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) + 4\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{(1/2)D_n} f\cdot g)\\
&\quad \cdot(e^{-(1/2)D_n} f\cdot g) + 8\gamma\sinh(\tfrac{1}{2}D_n)(D_t e^{(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) + 16\gamma^2\sinh(\tfrac{1}{2}D_n)\\
&\quad \times(e^{(1/2D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) + 12\lambda^{-1}\gamma\sinh(\tfrac{1}{2}D_n)(D_t^2 e^{-(1/2)D_n} f\cdot g)\\
&\quad \cdot(e^{-(1/2)D_n} f\cdot g) - 8\sinh(\tfrac{1}{2}D_n)(D_y e^{(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g)\\
&\quad - 8\omega\sinh(\tfrac{1}{2}D_n)(e^{(1/2)D_n} f\cdot g)\cdot(e^{-(1/2)D_n} f\cdot g) = 0,
\end{aligned}
$$

which implies that $P_3 = 0$.

Second, let us prove that $P_4 = 0$. By using (A12)–(A25) and (19)–(22), we can deduce the following relation

$$
\begin{aligned}
-P_4 e^{(3/2)D_n} f\cdot f &= [(D_t^3 e^{(1/2)D_n} + 4D_z e^{(3/2)D_n} - D_y D_t e^{(1/2)D_n}) f\cdot f][e^{(3/2)D_n} g\cdot g] - [e^{(3/2)D_n} f\cdot f]\\
&\quad \times[(D_t^3 e^{(1/2)D_n} + 4D_z e^{(3/2)D_n} - D_y D_t e^{(1/2)D_n}) g\cdot g]\\
&= \sinh(\tfrac{1}{2}D_n)(e^{D_n} f\cdot g)\cdot[(2D_t^3 e^{-D_n} + 2D_t D_y e^{-D_n} + 6\gamma D_t^2 e^{-D_n} + \tfrac{1}{2}\lambda\nu e^{-D_n}\\
&\quad + 8\lambda^{-1} e^{-2D_n} + 2\omega D_t e^{-D_n} + 2\gamma D_y e^{-D_n} + 4\lambda^2 D_t e^{D_n} + 6\gamma^2 D_t e^{-D_n}\\
&\quad + 2\omega\gamma e^{-D_n}) f\cdot g] - 3\gamma^2\sinh(D_n)(e^{(1/2)D_n} f\cdot g)\cdot(-\gamma e^{-(1/2)D_n} f\cdot g)\\
&= \sinh(\tfrac{1}{2}D_n)(e^{D_n} f\cdot g)\cdot[(2D_t^3 e^{-D_n} + 2D_t D_y e^{-D_n} + 6\gamma D_t^2 e^{-D_n} + \tfrac{1}{2}\lambda\nu e^{-D_n}\\
&\quad + 8\lambda^{-1} e^{-2D_n} + 2\omega D_t e^{-D_n} + 2\gamma D_y e^{-D_n} + 4\lambda^2 D_t e^{D_n} + 6\gamma^2 D_t e^{-D_n} + 2\omega\gamma e^{-D_n}\\
&\quad + 3\gamma^3 e^{-D_n}) f\cdot g] = 0.
\end{aligned}
$$

Thus the proof of Theorem 1 is completed.

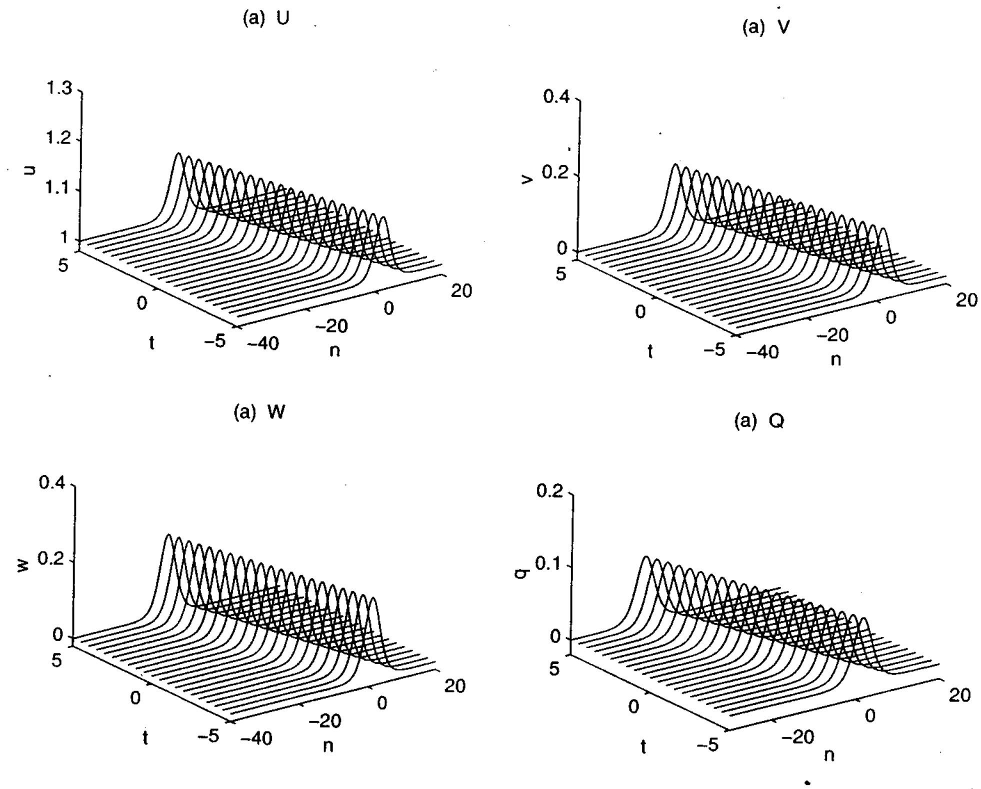

FIG. 1. 1-soliton solution of the lattice (4)–(7).

By using the BT given by (19)–(23), we can easily obtain the following solution from the trivial solution $f(n)=1$:

$$g(n)=1+\exp(\eta), \quad \eta=pn+qt+rz+sy+\eta^0 \tag{25}$$

with

$$\lambda^4=e^p(1+e^p+e^{2p}), \quad \mu=-\lambda^{-1}, \quad \gamma=\lambda, \quad \omega=\lambda^2, \quad \nu=-2\lambda^2, \quad \theta=-4\lambda^3-8\lambda^{-1},$$

where p is an arbitrary constant, $q=\lambda(1-e^{-p})$, $r=\lambda^{-1}(e^p-1)$, $s=\lambda^2(1-e^{-2p})$ and η^0 is an arbitrary constant. Therefore the corresponding one-soliton solution of the lattice (4)–(7) is

$$\begin{aligned} u(n)&=\frac{g(n+1)g(n-1)}{g^2(n)}, \quad v(n)=\left(\ln\frac{g(n+1)}{g(n)}\right)_t, \\ q(n)&=\left(\ln\frac{g(n+3)}{g(n)}\right)_z, \quad w(n)=\frac{1}{2}\,\frac{D_t^2 g(n+1)\cdot g(n+1)}{g(n+2)g(n)}, \end{aligned} \tag{26}$$

with $g(n)$ being given by (25). The plot of the solution (26) is shown in Fig. 1, where we choose $p=0.7$, $\lambda\approx 1.94$, $\eta^0=0$, $z=1$, $y=1$.

IV. NONLINEAR SUPERPOSITION FORMULA

In the following, we shall simply denote, without confusion, $f(n,t)$ by $f(n)$ or f. The corresponding nonlinear superposition formula is shown in the following theorem.

Theorem 2: Let f_0 be a solution of Eqs. (15)–(18). Suppose that f_i $(i=1,2)$ are two other solutions of (15)–(18) which are related to f_0 under the BT(19)–(23) with parameters $(\lambda_i,\mu_i,\gamma_i,\omega_i,\nu_i,\theta_i)$, i.e., $f_0 \rightarrow^{(\lambda_i,\mu_i,\gamma_i,\omega_i,\nu_i.\theta_i)} f_i$ $(i=1,2)$, where $\lambda_1\lambda_2 \neq 0$, $f_j \neq 0$ $(j=0,1,2)$. Then f_{12} defined by

$$\exp(-\tfrac{1}{2}D_n) f_0 \cdot f_{12} = k[\lambda_1 \exp(-\tfrac{1}{2}D_n) - \lambda_2 \exp(\tfrac{1}{2}D_n)] f_1 \cdot f_2, \tag{27}$$

where k is a nonzero constant, is a new solution which is related to f_1 and f_2 under the BT(19)-(23) with parameters $(\lambda_2,\mu_2,\gamma_2,\omega_2,\nu_2,\theta_2)$ and $(\lambda_1,\mu_1,\gamma_1,\omega_1,\nu_1,\theta_1)$, respectively.

In order to prove Theorem 2, we first establish some basic lemmas. In what follows, we always assume that the hypotheses of Theorem 2 are satisfied and f_{12} is determined by (27). Besides we set

$$J_i(n) \equiv (D_t e^{-(1/2)D_n} - \lambda_i e^{(1/2)D_n} + \gamma_i e^{-(1/2)D_n}) f_0(n) \cdot f_i(n), \quad i=1,2$$

$$K_i(n) \equiv (D_y e^{-(1/2)D_n} - \lambda_i D_t e^{(1/2)D_n} - \lambda_i \gamma_i e^{(1/2)D_n} + \omega_i e^{-(1/2)D_n}) f_0(n) \cdot f_i(n), \quad i=1,2$$

$$L_i(n) \equiv (\lambda_i^{-1} D_t^3 e^{-(1/2)D_n} + D_t^2 e^{(1/2)D_n} + 2\gamma_i D_t e^{(1/2)D_n} + 4\gamma_i^2 e^{(1/2)D_n} + 3\lambda_i^{-1}\gamma_i D_t^2 e^{-(1/2)D_n}$$
$$- 2D_y e^{(1/2)D_n} - 2\omega_i e^{(1/2)D_n} + \nu_i e^{-(1/2)D_n}) f_0(n) \cdot f_i(n), \quad i=1,2.$$

Lemma 1: The bilinear relations hold:

$$(D_z + \lambda_2^{-1} e^{-D_n} + \mu_2) f_1 \cdot f_{12} = 0, \tag{28}$$

$$(D_z + \lambda_1^{-1} e^{-D_n} + \mu_1) f_2 \cdot f_{12} = 0, \tag{29}$$

$$(D_t e^{-(1/2)D_n} - \lambda_2 e^{(1/2)D_n} + \gamma_2 e^{-(1/2)D_n}) f_1 \cdot f_{12} = 0, \tag{30}$$

$$(D_t e^{-(1/2)D_n} - \lambda_1 e^{(1/2)D_n} + \gamma_1 e^{-(1/2)D_n}) f_2 \cdot f_{12} = 0, \tag{31}$$

$$-D_z f_1 \cdot f_2 + (\mu_1 - \mu_2) f_1 f_2 - \frac{1}{k\lambda_1\lambda_2} e^{-D_n} f_0 \cdot f_{12} = 0, \tag{32}$$

$$(\lambda_2 D_t e^{(1/2)D_n} + \lambda_1 D_t e^{-(1/2)D_n} - 2\lambda_2\gamma_1 e^{(1/2)D_n} + 2\lambda_1\gamma_2 e^{-(1/2)D_n}) f_1 \cdot f_2 + \frac{1}{k} D_t e^{-(1/2)D_n} f_0 \cdot f_{12} = 0. \tag{33}$$

Proof: (28)–(33) can be proved similarly as in Refs. 10 and 11.

Lemma 2: The bilinear relations hold:

$$-D_t f_1 \cdot f_2 + (\gamma_1 - \gamma_2) f_1 f_2 - \frac{1}{k} f_0 f_{12} = 0, \tag{34}$$

$$D_y e^{-(1/2)D_n} f_1 \cdot f_{12} = (\lambda_2 D_t e^{(1/2)D_n} + \lambda_2\gamma_2 e^{(1/2)D_n} - \omega_2 e^{-(1/2)D_n}) f_1 \cdot f_{12}, \tag{35}$$

$$D_y e^{-(1/2)D_n} f_2 \cdot f_{12} = (\lambda_1 D_t e^{(1/2)D_n} + \lambda_1\gamma_1 e^{(1/2)D_n} - \omega_1 e^{-(1/2)D_n}) f_2 \cdot f_{12}. \tag{36}$$

Proof: First, according to the hypotheses of Theorem 2, we have

$$J_1(n) f_2(n+\tfrac{1}{2}) - J_2(n) f_1(n+\tfrac{1}{2}) = 0,$$

from which, by use of (A26) and (27), it follows that (34) holds. Next, since f_1 and f_2 are two solutions of (15)–(18), we have

$$[(D_t^2e^{(1/2)D_n}-D_ye^{(1/2)D_n})f_1\cdot f_1](e^{(1/2)D_n}f_2\cdot f_2) \\ -(e^{(1/2)D_n}f_1\cdot f_1)[(D_t^2e^{(1/2)D_n}-D_ye^{(1/2)D_n})f_2\cdot f_2]=0,$$

which can be rewritten as

$$-\frac{1}{k\lambda_2}(e^{-(1/2)D_n}f_0\cdot f_2)[(D_ye^{-(1/2)D_n}-\lambda_2D_te^{(1/2)D_n}-\lambda_2\gamma_2e^{(1/2)D_n}+\omega_2e^{-(1/2)D_n})f_1\cdot f_{12}]=0$$

by use of (A2), (A27)–(A30), (27), (30), and (34). Therefore (35) holds. Similarly we can prove that (36) also holds.

Lemma 3: The bilinear relations hold:

$$(\lambda_2^{-1}D_t^3e^{-(1/2)D_n}+D_t^2e^{(1/2)D_n}+2\gamma_2D_te^{(1/2)D_n}+4\gamma_2^2e^{(1/2)D_n}+3\lambda_2^{-1}\gamma_2D_t^2e^{-(1/2)D_n} \\ -2D_ye^{(1/2)D_n}-2\omega_2e^{(1/2)D_n}+\nu_2e^{-(1/2)D_n})f_1\cdot f_{12}=0, \tag{37}$$

$$(\lambda_1^{-1}D_t^3e^{-(1/2)D_n}+D_t^2e^{(1/2)D_n}+2\gamma_1D_te^{(1/2)D_n}+4\gamma_1^2e^{(1/2)D_n}+3\lambda_1^{-1}\gamma_1D_t^2e^{-(1/2)D_n} \\ -2D_ye^{(1/2)D_n}-2\omega_1e^{(1/2)D_n}+\nu_1e^{-(1/2)D_n})f_2\cdot f_{12}=0. \tag{38}$$

Proof: Since f_1 and f_2 are two solutions of Eqs. (15)–(18), we have

$$[(D_t^3D_z+6D_t^2+2D_t^2e^{D_n}-4D_ye^{D_n})f_1\cdot f_1]f_2^2-[(D_t^3D_z+6D_t^2+2D_t^2e^{D_n}-4D_ye^{D_n})f_2\cdot f_2]f_1^2 \\ +3[(D_zD_t-2e^{D_n}+2)f_1\cdot f_1](D_t^2f_2\cdot f_2)-3(D_t^2f_1\cdot f_1)[(D_zD_t-2e^{D_n}+2)f_2\cdot f_2]=0.$$

On the other hand, using a similar deduction as in Ref. 12, we obtain that

$$[(D_t^3D_z+6D_t^2+6D_t^2e^{D_n})f_1\cdot f_1]f_2^2-[(D_t^3D_z+6D_t^2+6D_t^2e^{D_n})f_2\cdot f_2]f_1^2 \\ +3[(D_zD_t-2e^{D_n}+2)f_1\cdot f_1](D_t^2f_2\cdot f_2)-3(D_t^2f_1\cdot f_1)[(D_zD_t-2e^{D_n}+2)f_2\cdot f_2] \\ =-\frac{2}{k\lambda_1\lambda_2}e^{-(1/2)D_n}\{[(D_t^3e^{-(1/2)D_n}+3\lambda_2D_t^2e^{(1/2)D_n}+3\gamma_2D_t^2e^{-(1/2)D_n}+6\lambda_2\gamma_2^2e^{(1/2)D_n} \\ +6\lambda_2\gamma_2D_te^{-(1/2)D_n})f_0\cdot f_2]\cdot(e^{-(1/2)D_n}f_1\cdot f_{12})-(e^{-(1/2)D_n}f_0\cdot f_2)\cdot[(D_t^3e^{-(1/2)D_n} \\ +3\lambda_2D_t^2e^{(1/2)D_n}+3\gamma_2D_t^2e^{(1/2)D_n}+6\lambda_2\gamma_2^2e^{(1/2)D_n}+6\lambda_2\gamma_2D_te^{-(1/2)D_n})f_1\cdot f_{12}]\}.$$

Thus by using (A6), (A7), (A31)–(A33), (27), and (34), we have

$$\frac{2}{k\lambda_1}e^{-(1/2)D_n}(e^{-(1/2)D_n}f_0\cdot f_2)\cdot[(\lambda_2^{-1}D_t^3e^{-(1/2)D_n}+D_t^2e^{(1/2)D_n}+2\gamma_2D_te^{(1/2)D_n}+4\gamma_2^2e^{(1/2)D_n} \\ +3\lambda_2^{-1}\gamma_2D_t^2e^{-(1/2)D_n}-2D_ye^{(1/2)D_n}-2\omega_2e^{(1/2)D_n}+\nu_2e^{-(1/2)D_n})f_1\cdot f_{12}]=0,$$

which implies that (37) holds. Similarly we can show that (38) holds.

Lemma 4: The bilinear relation holds:

$$(\lambda_1D_t^2e^{-(1/2)D_n}-\lambda_2D_t^2e^{(1/2)D_n}+4\lambda_2\gamma_1D_te^{(1/2)D_n}+4\lambda_1\gamma_2D_te^{-(1/2)D_n}-4\lambda_2\gamma_1^2e^{(1/2)D_n} \\ +4\lambda_1\gamma_2^2e^{-(1/2)D_n})f_1\cdot f_2-\frac{1}{k}D_t^2e^{-(1/2)D_n}f_0\cdot f_{12}+\frac{4}{k}\lambda_1\lambda_2e^{(1/2)D_n}f_0\cdot f_{12}=0. \tag{39}$$

Proof: According to the hypotheses of Theorem 2, we have

$$\frac{1}{\lambda_1}[J_1(n)]_t f_2\left(n-\frac{1}{2}\right)-\frac{1}{\lambda_2}[J_2(n)]_t f_1\left(n-\frac{1}{2}\right)-\frac{\gamma_1}{\lambda_1}J_1(n)f_2\left(n-\frac{1}{2}\right)+\frac{\gamma_2}{\lambda_2}J_2(n)f_1\left(n-\frac{1}{2}\right)=0,$$

which implies that (39) holds by use of (27), (33), and (34).

Lemma 5: The bilinear relation holds:

$$\begin{aligned}&(\lambda_1 D_y e^{-(1/2)D_n}+\lambda_2 D_y e^{(1/2)D_n}+2\lambda_1\omega_2 e^{-(1/2)D_n}-2\lambda_2\omega_1 e^{(1/2)D_n})f_1\cdot f_2\\&\quad+\left(\frac{1}{k}D_y e^{-(1/2)D_n}+\frac{2}{k}\lambda_1\lambda_2 e^{(1/2)D_n}\right)f_0\cdot f_{12}=0.\end{aligned} \tag{40}$$

Proof: Based on the hypotheses of Theorem 2, we have

$$\lambda_2 K_1(n)f_2(n-\tfrac{1}{2})-\lambda_1 K_2(n)f_1(n-\tfrac{1}{2})=0,$$

which implies that (40) holds upon using (27) and (33).

Lemma 6: The bilinear relation holds:

$$D_y f_1\cdot f_2+(\omega_2-\omega_1)f_1 f_2+\frac{1}{k}D_t f_0\cdot f_{12}+\frac{1}{k}(\gamma_1+\gamma_2)f_0 f_{12}=0. \tag{41}$$

Proof: According to the hypotheses of Theorem 2, we obtain

$$K_1(n)f_2(n+\tfrac{1}{2})-K_2(n)f_1(n+\tfrac{1}{2})=0,$$

which lead to (41) upon taking into account (27) and (33).

Lemma 7: The bilinear relation holds:

$$\begin{aligned}&-\frac{1}{k}D_t^3 e^{-(1/2)D_n}f_0\cdot f_{12}+\frac{4\lambda_1\lambda_2}{k}D_t e^{(1/2)D_n}f_0\cdot f_{12}-\frac{8\lambda_1\lambda_2}{k}(\gamma_1+\gamma_2)e^{(1/2)D_n}f_0\cdot f_{12}\\&\quad+6\lambda_2\gamma_1 D_t^2 e^{(1/2)D_n}f_1\cdot f_2-6\lambda_1\gamma_2 D_t^2 e^{-(1/2)D_n}f_1\cdot f_2+2\lambda_1\lambda_2\nu_1 e^{(1/2)D_n}f_1\cdot f_2\\&\quad-2\lambda_1\lambda_2\nu_2 e^{-(1/2)D_n}f_1\cdot f_2-12\lambda_1\gamma_2^2 D_t e^{-(1/2)D_n}f_1\cdot f_2-12\lambda_2\gamma_1^2 D_t e^{(1/2)D_n}f_1\cdot f_2\\&\quad+12\lambda_2\gamma_1^3 e^{(1/2)D_n}f_1\cdot f_2-12\lambda_1\gamma_2^3 e^{-(1/2)D_n}f_1\cdot f_2\\&\quad-\lambda_1 D_t^3 e^{-(1/2)D_n}f_1\cdot f_2-\lambda_2 D_t^3 e^{(1/2)D_n}f_1\cdot f_2=0.\end{aligned} \tag{42}$$

Proof: It follows from the hypotheses of Theorem 2 that

$$\begin{aligned}&\frac{1}{4}L_1(n)f_2\left(n-\frac{1}{2}\right)-\frac{1}{4}L_2(n)f_1\left(n-\frac{1}{2}\right)+\frac{3}{2}\frac{\gamma_1^2}{\lambda_1}J_1(n)f_2\left(n-\frac{1}{2}\right)-\frac{3}{2}\frac{\gamma_2^2}{\lambda_2}J_2(n)f_1\left(n-\frac{1}{2}\right)\\&\quad-\frac{3}{2}\frac{\gamma_1}{\lambda_1}[J_1(n)]_t f_2\left(n-\frac{1}{2}\right)+\frac{3}{2}\frac{\gamma_2}{\lambda_2}[J_2(n)]_t f_1\left(n-\frac{1}{2}\right)+\frac{3}{4}\lambda_1^{-1}[J_1(n)]_{tt}f_2\left(n-\frac{1}{2}\right)\\&\quad-\tfrac{3}{4}\lambda_2^{-1}[J_2(n)]_{tt}f_1(n-\tfrac{1}{2})=0.\end{aligned}$$

This tells that (42) holds by using (27), (33), (A34), and (A35).

Lemma 8: The bilinear relation holds:

$$\begin{aligned}&[-D_y D_t e^{-(1/2)D_n}+2\lambda_1\lambda_2 D_t e^{(1/2)D_n}+4\lambda_1\lambda_2(\gamma_1+\gamma_2)e^{(1/2)D_n}]f_0\cdot f_{12}+k(2\lambda_2\omega_1 D_t e^{(1/2)D_n}\\&\quad+2\lambda_1\omega_2 D_t e^{-(1/2)D_n}+2\lambda_2\gamma_1 D_y e^{(1/2)D_n}+2\lambda_1\gamma_2 D_y e^{-(1/2)D_n}-\lambda_2 D_y D_t e^{(1/2)D_n}\\&\quad+\lambda_1 D_y D_t e^{-(1/2)D_n}-4\lambda_2\gamma_1\omega_1 e^{(1/2)D_n}+4\lambda_1\gamma_2\omega_2 e^{-(1/2)D_n})f_1\cdot f_2=0.\end{aligned} \tag{43}$$

Proof: According to the hypotheses of Theorem 2, we have

$$\lambda_1^{-1}[K_1(n)]_t f_2\left(n-\frac{1}{2}\right)-\lambda_2^{-1}[K_2(n)]_t f_1\left(n-\frac{1}{2}\right)-\frac{\gamma_1}{\lambda_1}K_1(n)f_2\left(n-\frac{1}{2}\right)$$
$$+\frac{\gamma_2}{\lambda_2}K_2(n)f_1\left(n-\frac{1}{2}\right)=0,$$

which shows that (43) holds by using (27), (33), (A36), and (A37).

Lemma 9: The bilinear relation holds:

$$2\gamma_2\sinh(D_n)(e^{(1/2)D_n}f_0\cdot f_{12})\cdot(e^{(1/2)D_n}f_1\cdot f_2)-D_t\cosh(D_n)(e^{(1/2)D_n}f_0\cdot f_{12})\cdot(e^{(1/2)D_n}f_1\cdot f_2)$$
$$+\sinh(D_n)[(D_te^{(1/2)D_n}f_0\cdot f_{12})\cdot(e^{(1/2)D_n}f_1\cdot f_2)+(e^{(1/2)D_n}f_0\cdot f_{12})\cdot(D_te^{(1/2)D_n}f_1\cdot f_2)]$$
$$=-\frac{\lambda_2}{\lambda_1}e^{-(1/2)D_n}[(e^{D_n}f_0\cdot f_2)\cdot(D_te^{D_n}f_1\cdot f_{12})-(D_te^{D_n}f_0\cdot f_2)\cdot(e^{D_n}f_1\cdot f_{12})]. \tag{44}$$

Proof. On the one hand, by use of (A38), (A39), and (30), we can have

$$2\gamma_2\sinh(D_n)(e^{(1/2)D_n}f_0\cdot f_{12})\cdot(e^{(1/2)D_n}f_1\cdot f_2)-D_t\cosh(D_n)(e^{(1/2)D_n}f_0\cdot f_{12})\cdot(e^{(1/2)D_n}f_1\cdot f_2)$$
$$+\sinh(D_n)[(D_te^{(1/2)D_n}f_0\cdot f_{12})\cdot(e^{(1/2)D_n}f_1\cdot f_2)+(e^{(1/2)D_n}f_0\cdot f_{12})\cdot(D_te^{(1/2)D_n}f_1\cdot f_2)]$$
$$=\lambda_2[(e^{(3/2)D_n}f_0\cdot f_2)(e^{(1/2)D_n}f_1\cdot f_{12})-(e^{(1/2)D_n}f_0\cdot f_2)(e^{(3/2)D_n}f_1\cdot f_{12}).$$

On the other hand, by use of (A40) and (30), we can have

$$e^{-(1/2)D_n}[(e^{D_n}f_0\cdot f_2)\cdot(D_te^{D_n}f_1\cdot f_{12})-(D_te^{D_n}f_0\cdot f_2)\cdot(e^{D_n}f_1\cdot f_{12}]$$
$$=-\lambda_1[(e^{(3/2)D_n}f_0\cdot f_2)(e^{(1/2)D_n}f_1\cdot f_{12})-(e^{(1/2)D_n}f_0\cdot f_2)(e^{(3/2)D_n}f_1\cdot f_{12})].$$

Therefore combining these two equalities leads to the required equality (44).

We now turn to the proof of Theorem 2. Based on Lemma 1, Lemma 2, and Lemma 3, it suffices to show that

$$(2D_t^3e^{-D_n}+2D_tD_ye^{-D_n}+6\gamma_2D_t^2e^{-D_n}+\tfrac{1}{2}\lambda_2\nu_2e^{-D_n}+8\lambda_2^{-1}e^{-2D_n}+2\omega_2D_te^{-D_n}+2\gamma_2D_ye^{-D_n}$$
$$+4\lambda_2^2D_te^{D_n}+6\gamma_2^2D_te^{-D_n}+2\omega_2\gamma_2e^{-D_n}+3\gamma_2^3e^{-D_n}+\theta_2e^{D_n})f_1\cdot f_{12}=0, \tag{45}$$

$$(2D_t^3e^{-D_n}+2D_tD_ye^{-D_n}+6\gamma_1D_t^2e^{-D_n}+\tfrac{1}{2}\lambda_1\nu_1e^{-D_n}+8\lambda_1^{-1}e^{-2D_n}+2\omega_1D_te^{-D_n}+2\gamma_1D_ye^{-D_n}$$
$$+4\lambda_1^2D_te^{D_n}+6\gamma_1^2D_te^{-D_n}+2\omega_1\gamma_1e^{-D_n}+3\gamma_1^3e^{-D_n}+\theta_1e^{D_n})f_2\cdot f_{12}=0. \tag{46}$$

Since f_1 and f_2 are two solutions of Eqs. (15)–(18), we have

$$[(D_t^3e^{(1/2)D_n}+4D_ze^{(3/2)D_n}-D_yD_te^{(1/2)D_n})f_1\cdot f_1][e^{(3/2)D_n}f_2\cdot f_2]$$
$$-[(D_t^3e^{(1/2)D_n}+4D_ze^{(3/2)D_n}-D_yD_te^{(1/2)D_n})f_2\cdot f_2][e^{(3/2)D_n}f_1\cdot f_1]=0. \tag{47}$$

Taking advantage of (A12)–(A14), (A41)–(A43), (27), (32)–(34), and (39)–(44), we can rewrite (47) as

$$-\frac{1}{k\lambda_1}e^{-(1/2)D_n}(e^{D_n}f_0\cdot f_2)\cdot\left[\left(D_t^3e^{-D_n}+D_yD_te^{-D_n}+3\gamma_2D_t^2e^{-D_n}+\frac{1}{4}\lambda_2\nu_2e^{-D_n}\right.\right.$$
$$+4\lambda_2^{-1}e^{-2D_n}+\omega_2D_te^{-D_n}+\gamma_2D_ye^{-D_n}+2\lambda_2^2D_te^{D_n}+3\gamma_2^2D_te^{-D_n}+\omega_2\gamma_2e^{-D_n}$$
$$\left.\left.+\frac{3}{2}\gamma_2^3e^{-D_n}+\frac{1}{2}\;\theta_2e^{D_n}\right)f_1\cdot f_{12}\right]=0, \tag{48}$$

which implies that (45) holds. Similarly we can prove that (46) also holds. Therefore we have completed the proof of theorem 2.

As an application of the nonlinear superposition formula (27), we can construct soliton solutions of the Blaszak–Marciniak lattice of the Eqs. (15)–(18). Choose for example $f_0=1,k=1/\lambda_1-\lambda_2$. It is easily verified that

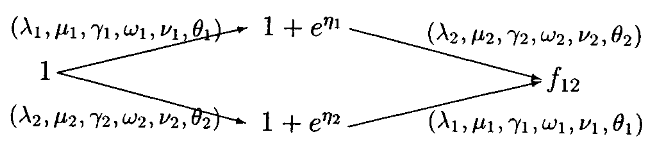

where f_{12} is given by

$$f_{12}=1+\frac{\lambda_1e^{-p_1}-\lambda_2}{\lambda_1-\lambda_2}e^{\eta_1}+\frac{\lambda_1-\lambda_2e^{-p_2}}{\lambda_1-\lambda_2}e^{\eta_2}+\frac{\lambda_1e^{-p_1}-\lambda_2e^{-p_2}}{\lambda_1-\lambda_2}e^{\eta_1+\eta_2} \tag{49}$$

with

$$\eta_i=p_in+q_it+r_iz+s_iy+\eta_i^0,\quad q_i=\lambda_i(1-e^{-p_i}),\quad r_i=\lambda_i^{-1}(e^{p_i}-1),\quad s_i=\lambda_i^2(1-e^{-2p_i})$$

and

$$\lambda_i^4=e^{p_i}(1+e^{p_i}+e^{2p_i}),\quad \mu_i=-\lambda_i^{-1},\quad \gamma_i=\lambda_i,\quad \omega_i=\lambda_i^2,\quad \nu_i=-2\lambda_i^2,\quad \theta_i=-4\lambda_i^3-8\lambda_i^{-1}$$

in which the p_i $(i=1,2)$ are arbitrary constants. Thus the corresponding 2-soliton solution of (4)–(7) is

$$u(n)=\frac{f_{12}(n+1)f_{12}(n-1)}{f_{12}^2(n)},\quad v(n)=\left(\ln\frac{f_{12}(n+1)}{f_{12}(n)}\right)_t$$
$$q(n)=\left(\ln\frac{f_{12}(n+3)}{f_{12}(n)}\right)_z,\quad w(n)=\frac{1}{2}\,\frac{D_t^2f_{12}(n+1)\cdot f_{12}(n+1)}{f_{12}(n+2)f_{12}(n)} \tag{50}$$

with $f_{12}(n)$ being given by (49). The plot of (50) is shown in Fig. 2 where we chose $p_1=1.2$, $p_2=1.42$, $\lambda_1\approx-2.67$, $\lambda_2\approx3.10$, $\eta_1^0=\eta_2^0=0$, $z=1$, $y=1$.

In general, along this line, we can generate multisoliton solutions for the Blaszak–Marciniak lattice (4)–(7) successively.

V. CONCLUSION

By introducing two auxiliary variables, a four-field lattice introduced by Blaszak and Marciniak[1] is transformed into Hirota's bilinear form. The transformation of the dependent variables are given by (8), (10), and (11). A bilinear Bäcklund transformation (Theorem 1) is found and its corresponding nonlinear superposition formula (Theorem 2) is rigorously proved. As a

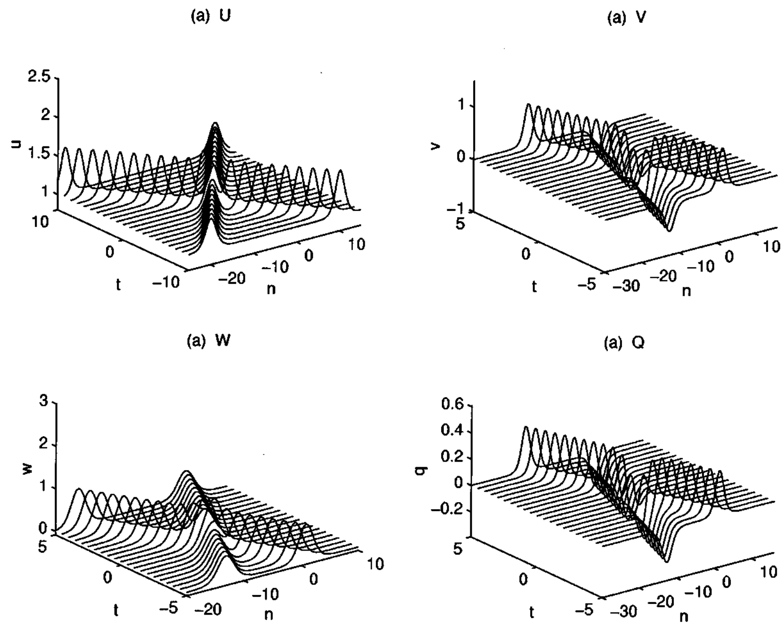

FIG. 2. 2-soliton solution of the lattice (4)–(7).

consequence, one-soliton and two-soliton solutions to the lattice are constructed. In principle, the resulted nonlinear superposition formula guarantees the existence of multisoliton solutions and tells us how to construct them explicitly.

ACKNOWLEDGMENTS

This work was supported by the City University of Hong Kong, the Research Grants Council of Hong Kong, the National Natural Science Foundation of China (Project No. 19871082) and the Chinese Academy of Sciences.

APPENDIX: HIROTA BILINEAR OPERATOR IDENTITIES

The following bilinear operator identities hold for arbitrary functions a,b,c and d:

$$D_t^3(D_zD_ta\cdot a)\cdot a^2=D_t[(D_t^3D_za\cdot a)\cdot a^2+3(D_tD_za\cdot a)\cdot(D_t^2a\cdot a)], \tag{A1}$$

$$D_t^{2n+1}a\cdot a=0,\quad n=0,1,2,\ldots, \tag{A2}$$

$$\begin{aligned}D_t^3(e^{D_n}a\cdot a)\cdot a^2=2\sinh(\tfrac{1}{2}D_n)[(D_t^3e^{(1/2)D_n}a\cdot a)\cdot(e^{(1/2)D_n}a\cdot a)\\+3(D_te^{(1/2)D_n}a\cdot a)\cdot(D_t^2e^{(1/2)D_n}a\cdot a)],\end{aligned} \tag{A3}$$

$$\begin{aligned}&D_t[(D_t^2e^{D_n}a\cdot a)\cdot a^2+(e^{D_n}a\cdot a)\cdot(D_t^2a\cdot a)]\\&\quad=2\sinh(\tfrac{1}{2}D_n)[(D_t^3e^{(1/2)D_n}a\cdot a)\cdot(e^{(1/2)D_n}a\cdot a)+(D_t^2e^{(1/2)D_n}a\cdot a)\cdot(D_te^{(1/2)D_n}a\cdot a)],\end{aligned} \tag{A4}$$

$$\begin{aligned}&2\sinh(\tfrac{1}{2}D_n)[(D_yD_te^{(1/2)D_n}a\cdot a)\cdot(e^{(1/2)D_n}a\cdot a)+(D_te^{(1/2)D_n}a\cdot a)\cdot(D_ye^{(1/2)D_n}a\cdot a)]\\&\quad=D_t(D_ye^{D_n}a\cdot a)\cdot a^2,\end{aligned} \tag{A5}$$

$$(D_y e^{D_n} a\cdot a)b^2 - a^2 D_y e^{D_n} b\cdot b = 2\sinh(\tfrac{1}{2}D_n)[(D_y e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) - (e^{(1/2)D_n} a\cdot b)\cdot(D_y e^{-(1/2)D_n} a\cdot b)] = 2D_y \cosh(\tfrac{1}{2}D_n)(e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b), \tag{A6}$$

$$(D_t^2 e^{D_n} a\cdot a)b^2 - a^2 D_t^2 e^{D_n} b\cdot b = D_t\cosh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) - (e^{(1/2)D_n} a\cdot b)\cdot(D_t e^{-(1/2)D_n} a\cdot b)] + \sinh(\tfrac{1}{2}D_n)[(D_t^2 e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) - 2(D_t e^{(1/2)D_n} a\cdot b)\cdot(D_t e^{-(1/2)D_n} a\cdot b) + (e^{(1/2)D_n} a\cdot b)\cdot(D_t^2 e^{-(1/2)D_n} a\cdot b)] = D_t\cosh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) - (e^{(1/2)D_n} a\cdot b)\cdot(D_t e^{-(1/2)D_n} a\cdot b)] + D_t^2 \sinh((1/2)D_n)(e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b), \tag{A7}$$

$$D_t\cosh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) + (e^{(1/2)D_n} a\cdot b)\cdot(D_t e^{-(1/2)D_n} a\cdot b)] = \sinh(\tfrac{1}{2}D_n)[(D_t^2 e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) - (e^{(1/2)D_n} a\cdot b)\cdot(D_t^2 e^{-(1/2)D_n} a\cdot b)], \tag{A8}$$

$$D_t\cosh(\tfrac{1}{2}D_n)a\cdot a = 0, \tag{A9}$$

$$\sinh(\tfrac{1}{2}D_n)a\cdot a = 0, \tag{A10}$$

$$D_t\cosh(\tfrac{1}{2}D_n)(e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) = \sinh(\tfrac{1}{2}D_n)[(D_t e^{(1/2)D_n} a\cdot b)\cdot(e^{-(1/2)D_n} a\cdot b) - (e^{(1/2)D_n} a\cdot b)\cdot(D_t e^{-(1/2)D_n} a\cdot b)], \tag{A11}$$

$$(D_t^3 e^{(1/2)D_n} a\cdot a)(e^{(3/2)D_n} b\cdot b) - (e^{(3/2)D_n} a\cdot a)(D_t^3 e^{(1/2)D_n} b\cdot b) = 2\sinh(\tfrac{1}{2}D_n)(e^{D_n} a\cdot b)\cdot(D_t^3 e^{-D_n} a\cdot b) + \tfrac{1}{2}\sinh(D_n)[(D_t^3 e^{-(1/2)D_n} a\cdot b)\cdot(e^{(1/2)D_n} a\cdot b) + 3(D_t e^{-(1/2)D_n} a\cdot b)\cdot(D_t^2 e^{(1/2)D_n} a\cdot b)] - 3D_t\cosh(D_n)\times(D_t e^{-(1/2)D_n} a\cdot b)\cdot(D_t e^{(1/2)D_n} a\cdot b) + \tfrac{3}{2}D_t^2\sinh(D_n)(D_t e^{-(1/2)D_n} a\cdot b)\cdot(e^{(1/2)D_n} a\cdot b) = \tfrac{1}{4}D_t^3\cosh(D_n)(e^{-(1/2)D_n} a\cdot b)\cdot(e^{(1/2)D_n} a\cdot b) + \tfrac{3}{4}D_t^2\sinh(D_n)\times[(D_t e^{-(1/2)D_n} a\cdot b)\cdot(e^{(1/2)D_n} a\cdot b) - (e^{-(1/2)D_n} a\cdot b)\cdot(D_t e^{(1/2)D_n} a\cdot b)] + \tfrac{3}{4}D_t\cosh(D_n)[(D_t^2 e^{-(1/2)D_n} a\cdot b)\cdot(e^{(1/2)D_n} a\cdot b) + (e^{-(1/2)D_n} a\cdot b)\cdot(D_t^2 e^{(1/2)D_n} a\cdot b) - 2(D_t e^{-(1/2)D_n} a\cdot b)\cdot(D_t e^{(1/2)D_n} a\cdot b)] + \tfrac{1}{4}\sinh(D_n)[(D_t^3 e^{-(1/2)D_n} a\cdot b)\cdot(e^{(1/2)D_n} a\cdot b) + 3(D_t e^{-(1/2)D_n} a\cdot b)\cdot(D_t^2 e^{(1/2)D_n} a\cdot b) - (e^{-(1/2)D_n} a\cdot b)\cdot(D_t^3 e^{(1/2)D_n} a\cdot b) - 3(D_t^2 e^{-(1/2)D_n} a\cdot b)\cdot(D_t e^{(1/2)D_n} a\cdot b)], \tag{A12}$$

$$(D_z e^{(3/2)D_n} a\cdot a)(e^{(3/2)D_n} b\cdot b) - (e^{(3/2)D_n} a\cdot a)(D_z e^{(3/2)D_n} b\cdot b) = 2\sinh(\tfrac{3}{2}D_n)(D_z a\cdot b)\cdot ab, \tag{A13}$$

$$(D_yD_te^{(1/2)D_n}a\cdot a)(e^{(3/2)D_n}b\cdot b)-(e^{(3/2)D_n}a\cdot a)(D_yD_te^{(1/2)D_n}b\cdot b)$$
$$=-2\sinh(\tfrac{1}{2}D_n)(e^{D_n}a\cdot b)\cdot(D_tD_ye^{-D_n}a\cdot b)-\sinh(D_n)[(D_ye^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}a\cdot b)$$
$$+(D_te^{-(1/2)D_n}a\cdot b)\cdot(D_ye^{(1/2)D_n}a\cdot b)]+D_y\cosh(D_n)(D_te^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)$$
$$+D_t\cosh(D_n)(D_ye^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)=\tfrac{1}{2}D_yD_t\sinh(D_n)(e^{-(1/2)D_n}a\cdot b)$$
$$\cdot(e^{(1/2)D_n}a\cdot b)+\tfrac{1}{2}D_t\cosh(D_n)[(D_ye^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)-(e^{-(1/2)D_n}a\cdot b)$$
$$\cdot(D_ye^{(1/2)D_n}a\cdot b)]+\tfrac{1}{2}D_y\cosh(D_n)[(D_te^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)-(e^{-(1/2)D_n}a\cdot b)$$
$$\cdot(D_te^{(1/2)D_n}a\cdot b)]+\tfrac{1}{2}\sinh(D_n)[(D_yD_te^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)+(e^{-(1/2)D_n}a\cdot b)$$
$$\cdot(D_yD_te^{(1/2)D_n}a\cdot b)-(D_ye^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}a\cdot b)-(D_te^{-(1/2)D_n}a\cdot b)$$
$$\cdot(D_ye^{(1/2)D_n}a\cdot b)], \qquad \text{(A14)}$$

$$2\sinh(D_n)(D_t^2e^{(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)$$
$$=D_t[(D_te^{(3/2)D_n}a\cdot b)\cdot(e^{-(1/2)D_n}a\cdot b)+(e^{(3/2)D_n}a\cdot b)\cdot(D_te^{-(1/2)D_n}a\cdot b)], \qquad \text{(A15)}$$

$$2D_t\cosh(D_n)(D_te^{(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)$$
$$=D_t[(D_te^{(3/2)D_n}a\cdot b)\cdot(e^{-(1/2)D_n}a\cdot b)-(e^{(3/2)D_n}a\cdot b)\cdot(D_te^{-(1/2)D_n}a\cdot b)], \qquad \text{(A16)}$$

$$\sinh(D_n)[(D_t^2e^{(1/2)D_n}a\cdot b)\cdot(e^{-(1/2)D_n}a\cdot b)+(e^{(1/2)D_n}a\cdot b)\cdot(D_t^2e^{-(1/2)D_n}a\cdot b)$$
$$+2(D_te^{(1/2)D_n}a\cdot b)\cdot(D_te^{-(1/2)D_n}a\cdot b)]$$
$$=\sinh(\tfrac{1}{2}D_n)[(D_t^2e^{D_n}a\cdot b)\cdot(e^{-D_n}a\cdot b)+(e^{D_n}a\cdot b)\cdot(D_t^2e^{-D_n}a\cdot b)+2(D_te^{D_n}a\cdot b)$$
$$\cdot(D_te^{-D_n}a\cdot b)], \qquad \text{(A17)}$$

$$D_t\cosh(D_n)[(D_te^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)1+(e^{-(1/2)D_n}a\cdot b)\cdot(D_te^{-(1/2)D_n}a\cdot b)]$$
$$=-\sinh(\tfrac{1}{2}D_n)[(D_t^2e^{D_n}a\cdot b)\cdot(e^{-D_n}a\cdot b)-(e^{D_n}a\cdot b)\cdot(D_t^2e^{-D_n}a\cdot b)], \qquad \text{(A18)}$$

$$D_t^2\sinh(D_n)(e^{-q(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)$$
$$=-\sinh(\tfrac{1}{2}D_n)[(D_t^2e^{D_n}a\cdot b)\cdot(e^{-D_n}a\cdot b)-2(D_te^{D_n}a\cdot b)\cdot(D_te^{-D_n}a\cdot b)$$
$$+(e^{D_n}a\cdot b)\cdot(D_t^2e^{-D_n}a\cdot b)], \qquad \text{(A19)}$$

$$\sinh(D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)=-\sinh(\tfrac{1}{2}D_n)(e^{D_n}a\cdot b)\cdot(e^{-D_n}a\cdot b), \qquad \text{(A20)}$$

$$\sinh(\tfrac{3}{2}D_n)(e^{-D_n}a\cdot b)\cdot ab=-\sinh(\tfrac{1}{2}D_n)(e^{D_n}a\cdot b)\cdot(e^{-2D_n}a\cdot b), \qquad \text{(A21)}$$

$$\sinh(D_n)[(D_te^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)+(e^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}a\cdot b)]$$
$$=-\sinh(\tfrac{1}{2}D_n)[(D_te^{D_n}a\cdot b)\cdot(e^{-D_n}a\cdot b)+(e^{D_n}a\cdot b)\cdot(D_te^{-D_n}a\cdot b)], \qquad \text{(A22)}$$

$$D_t\cosh(D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)$$
$$=-\sinh(\tfrac{1}{2}D_n)[(D_te^{D_n}a\cdot b)\cdot(e^{-D_n}a\cdot b)-(e^{D_n}a\cdot b)\cdot(D_te^{-D_n}a\cdot b)], \qquad \text{(A23)}$$

$$D_t(e^{(3/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b)=-2\sinh(\tfrac{1}{2}D_n)(e^{D_n}a\cdot b)\cdot(D_te^{D_n}a\cdot b), \qquad \text{(A24)}$$

$$D_t(e^{(3/2)D_n}a\cdot b)\cdot(e^{-(1/2)D_n}a\cdot b)=2\sinh(\tfrac{1}{2}D_n)(D_t e^{(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}a\cdot b), \tag{A25}$$

$$(D_t a\cdot b)c-(D_t a\cdot c)b=-aD_t b\cdot c, \tag{A26}$$

$$(D_t^2 e^{(1/2)D_n}a\cdot a)(e^{(1/2)D_n}b\cdot b)-(e^{(1/2)D_n}a\cdot a)(D_t^2 e^{(1/2)D_n}b\cdot b)=2D_t\cosh(\tfrac{1}{2}D_n)(D_t a\cdot b)\cdot ab, \tag{A27}$$

$$\begin{aligned}&(D_y e^{(1/2)D_n}a\cdot a)(e^{(1/2)D_n}b\cdot b)-(e^{(1/2)D_n}a\cdot a)(D_y e^{(1/2)D_n}b\cdot b)\\&=D_y(e^{(1/2)D_n}a\cdot b)\cdot(e^{-(1/2)D_n}a\cdot b),\end{aligned} \tag{A28}$$

$$\begin{aligned}2D_t\cosh(\tfrac{1}{2}D_n)ab\cdot cd=&(D_t e^{(1/2)D_n}a\cdot d)(e^{-(1/2)D_n}c\cdot b)-(e^{(1/2)D_n}a\cdot d)(D_t e^{-(1/2)D_n}c\cdot b)\\&+(D_t e^{-(1/2)D_n}a\cdot d)(e^{(1/2)D_n}c\cdot b)-(e^{-(1/2)D_n}a\cdot d)(D_t e^{(1/2)D_n}c\cdot b),\end{aligned} \tag{A29}$$

$$\begin{aligned}D_y(e^{-(1/2)D_n}a\cdot b)\cdot(e^{-(1/2)D_n}c\cdot d)=&(D_y e^{-(1/2)D_n}a\cdot d)(e^{-(1/2)D_n}c\cdot b)-(e^{-(1/2)D_n}a\cdot d)\\&\times(D_y e^{-(1/2)D_n}c\cdot b),\end{aligned} \tag{A30}$$

$$\begin{aligned}&2D_y\cosh((\tfrac{1}{2}D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\&=e^{-(1/2)D_n}[(D_y e^{(1/2)D_n}a\cdot d)\cdot(e^{-(1/2)D_n}c\cdot b)-(e^{(1/2)D_n}a\cdot d)\cdot(D_y e^{-(1/2)D_n}c\cdot b)\\&\quad+(D_y e^{-(1/2)D_n}a\cdot d)\cdot(e^{(1/2)D_n}c\cdot b)-(e^{-(1/2)D_n}a\cdot d)\cdot(D_y e^{(1/2)D_n}c\cdot b)],\end{aligned} \tag{A31}$$

$$\begin{aligned}&2D_t^2\sinh(\tfrac{1}{2}D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\&=e^{-(1/2)D_n}[(D_t^2 e^{(1/2)D_n}a\cdot d)\cdot(e^{-(1/2)D_n}c\cdot b)+(e^{(1/2)D_n}a\cdot d)\cdot(D_t^2 e^{-(1/2)D_n}c\cdot b)\\&\quad-2(D_t e^{(1/2)D_n}a\cdot d)\cdot(D_t e^{-(1/2)D_n}c\cdot b)-(D_t^2 e^{-(1/2)D_n}a\cdot d)\cdot(e^{(1/2)D_n}c\cdot b)\\&\quad-(e^{-(1/2)D_n}a\cdot d)\cdot(D_t^2 e^{(1/2)D_n}c\cdot b)+2(D_t e^{-(1/2)D_n}a\cdot d)\cdot(D_t e^{(1/2)D_n}c\cdot b)],\end{aligned} \tag{A32}$$

$$\begin{aligned}&2D_t\cosh(\tfrac{1}{2}D_n)[(D_t e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)+(e^{-(1/2)D_n}a\cdot b)\cdot(D_t e^{(1/2)D_n}c\cdot d)]\\&=e^{-(1/2)D_n}[(D_t^2 e^{(1/2)D_n}a\cdot d)\cdot(e^{-(1/2)D_n}c\cdot b)-(e^{(1/2)D_n}a\cdot d)\cdot(D_t^2 e^{-(1/2)D_n}c\cdot b)\\&\quad+(D_t^2 e^{-(1/2)D_n}a\cdot d)\cdot(e^{(1/2)D_n}c\cdot b)-(e^{-(1/2)D_n}a\cdot d)\cdot(D_t^2 e^{(1/2)D_n}c\cdot b)],\end{aligned} \tag{A33}$$

$$a_{ttt}b=\tfrac{1}{8}[D_t^3 a\cdot b+(ab)_{ttt}+3(D_t a\cdot b)_{tt}+3(D_t^2 a\cdot b)_t], \tag{A34}$$

$$ab_{ttt}=\tfrac{1}{8}[-D_t^3 a\cdot b+(ab)_{ttt}-3(D_t a\cdot b)_{tt}+3(D_t^2 a\cdot b)_t], \tag{A35}$$

$$a_{yt}b=\tfrac{1}{4}[D_y D_t a\cdot b+(ab)_{yt}+(D_y a\cdot b)_t+(D_t a\cdot b)_y], \tag{A36}$$

$$ab_{yt}=\tfrac{1}{4}[D_y D_t a\cdot b+(ab)_{yt}-(D_y a\cdot b)_t-(D_t a\cdot b)_y], \tag{A37}$$

$$\begin{aligned}2\sinh(D_n)(e^{(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)=&(e^{(3/2)D_n}a\cdot d)(e^{-(1/2)D_n}c\cdot b)-(e^{-(1/2)D_n}a\cdot d)\\&\times(e^{(3/2)D_n}c\cdot b),\end{aligned} \tag{A38}$$

$$\begin{aligned}&\sinh(D_n)[(D_te^{(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)+(e^{(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}c\cdot d)]\\&\quad-D_t\cosh(D_n)(e^{(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\&\quad=(e^{(3/2)D_n}a\cdot d)(D_te^{-(1/2)D_n}c\cdot b)-(D_te^{-(1/2)D_n}a\cdot d)(e^{(3/2)D_n}c\cdot b),\end{aligned}\tag{A39}$$

$$\begin{aligned}&e^{-(1/2)D_n}[(e^{D_n}a\cdot b)\cdot(D_te^{D_n}c\cdot d)-(D_te^{D_n}a\cdot b)\cdot(e^{D_n}c\cdot d)]\\&\quad=-e^{D_n}[(D_te^{-(1/2)D_n}a\cdot c)\cdot(e^{(1/2)D_n}d\cdot b)-(e^{-(1/2)D_n}a\cdot c)\cdot(D_te^{(1/2)D_n}d\cdot b)],\end{aligned}\tag{A40}$$

$$\begin{aligned}&2D_t\cosh(D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)-2\sinh(D_n)[(D_te^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\&\quad+(e^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}c\cdot d)]\\&\quad=2e^{-(1/2)D_n}[(D_te^{-D_n}a\cdot d)\cdot(e^{D_n}c\cdot b)-(e^{D_n}a\cdot d)\cdot(D_te^{-D_n}c\cdot b)],\end{aligned}\tag{A41}$$

$$\begin{aligned}&\tfrac{1}{4}D_t^3\cosh(D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)-\tfrac{3}{4}D_t^2\sinh(D_n)[(D_te^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\&\quad+(e^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}c\cdot d)]+\tfrac{3}{4}D_t\cosh(D_n)[(D_t^2e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)\\&\quad+(e^{-(1/2)D_n}a\cdot b)\cdot(D_t^2e^{(1/2)D_n}c\cdot d)+2(D_te^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}c\cdot d)]\\&\quad-\tfrac{1}{4}\sinh(D_n)[(D_t^3e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)+(e^{-(1/2)D_n}a\cdot b)\cdot(D_t^3e^{(1/2)D_n}c\cdot d)\\&\quad+3(D_te^{-(1/2)D_n}a\cdot b)\cdot(D_t^2e^{(1/2)D_n}c\cdot d)+3(D_t^2e^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}c\cdot d)]\\&\quad=-e^{-(1/2)D_n}[(e^{D_n}a\cdot d)\cdot(D_t^3e^{-D_n}c\cdot b)-(D_t^3e^{-D_n}a\cdot d)\cdot(e^{D_n}c\cdot b)],\end{aligned}\tag{A42}$$

$$\begin{aligned}&-\tfrac{1}{2}D_yD_t\sinh(D_n)(e^{-(1/2)D_n}a\cdot b)\cdot(e^{(1/2)D_n}c\cdot d)+\tfrac{1}{2}D_y\cosh(D_n)[(D_te^{-(1/2)D_n}a\cdot b)\\&\quad\cdot(e^{(1/2)D_n}c\cdot d)+(e^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}c\cdot d)]+\tfrac{1}{2}D_t\cosh(D_n)[(D_ye^{-(1/2)D_n}a\cdot b)\\&\quad\cdot(e^{(1/2)D_n}c\cdot d)+(e^{-(1/2)D_n}a\cdot b)\cdot(D_ye^{(1/2)D_n}c\cdot d)]-\tfrac{1}{2}\sinh(D_n)[(D_yD_te^{-(1/2)D_n}a\cdot b)\\&\quad\cdot(e^{(1/2)D_n}c\cdot d)+(e^{-(1/2)D_n}a\cdot b)\cdot(D_yD_te^{(1/2)D_n}c\cdot d)+(D_ye^{-(1/2)D_n}a\cdot b)\cdot(D_te^{(1/2)D_n}c\cdot d)\\&\quad+(D_te^{-(1/2)D_n}a\cdot b)\cdot(D_ye^{(1/2)D_n}c\cdot d)]\\&\quad=-e^{-(1/2)D_n}[(e^{D_n}a\cdot d)\cdot(D_yD_te^{-D_n}c\cdot b)-(D_yD_te^{-D_n}a\cdot d)\cdot(e^{D_n}c\cdot b)].\end{aligned}\tag{A43}$$

[1] M. Blaszak and K. Marciniak, J. Math. Phys. **35**, 4661 (1994).
[2] Y. T. Wu and X. G. Geng, J. Math. Phys. **37**, 2338 (1996).
[3] W. X. Ma and B. Fuchssteiner, Algebraic structure of zero curvature equations and master symmetries of discrete evolution equations, solv-int/9809009 (1998)
[4] X. B. Hu and Z. N. Zhu, J. Math. Phys. **39**, 4766 (1998).
[5] R. Hirota, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey, (Springer, Berlin, 1980).
[6] R. Hirota and J. Satsuma, Suppl. Prog. Theor. Phys. **59**, 64 (1976).
[7] Y. Matsuno, *Bilinear Transformation Method* (Academic, New York, 1984).
[8] J. J. C. Nimmo, in *Soliton Theory, A Survey of Results*, edited by A. P. Fordy (Manchester University Press, Manchester, 1990).
[9] J. Hietarinta, J. Math. Phys. **28**, 1732 (1987).
[10] X. B. Hu, J. Phys. A **27**, 201 (1994).
[11] X. B. Hu and Z. N. Zhu, J. Phys. A **31**, 4755 (1998).
[12] X. B. Hu, Y. T. Wu, and X. G. Geng, J. Math. Phys. **40**, 2001 (1999).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Some finite dimensional indecomposable representations of $E(2)$

Joe Repka
Department of Mathematics, University of Toronto, Toronto, Ontario M5S 3G3, Canada

Hubert de Guise
Centre de Recherches Mathématiques, Université de Montréal, C.P. 6128 Succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada

(Received 21 June 1999; accepted for publication 19 July 1999)

We describe the construction of some finite dimensional nonunitary representations of $E(2)$, the Lie group of Euclidean transformations in the plane. Some properties of these representations are also discussed, with emphasis on indecomposable representations. © *1999 American Institute of Physics.* [S0022-2488(99)02711-5]

I. INTRODUCTION

The group $E(2)$ of Euclidean transformations in two dimensions is the noncompact semidirect product group $[\mathbb{R}^2]\mathrm{SO}(2)$, which consists of Abelian translations in the plane together with rotations. Its unitary irreducible representations (unirreps) are either one-dimensional representations or infinite dimensional representations which can be constructed in the standard way by induction.[1] Much less is known about the finite dimensional, nonunitary representations of $E(2)$, the prototype of which is the ''natural'' representation

$$\pi:(R(\theta),x,y)\mapsto\begin{pmatrix}\cos\theta & -\sin\theta & x\\ \sin\theta & \cos\theta & y\\ 0 & 0 & 1\end{pmatrix} \tag{1}$$

in terms of 3×3 matrices, where $R(\theta)$ is the SO(2) rotation parametrized by the angle θ, and (x,y) is a vector describing the translation part of the transformation.

The representation of Eq. (1) was obtained in the familiar way from a 2×2 representation of SO(2), which is extended to a 3×3 matrix by addition of an extra line and an extra column with appropriate entries to account for the translation part of $E(2)$. This representation is not irreducible, but it is indecomposable.

It is the objective of this paper to present an explicit method of obtaining some finite-dimensional indecomposable representations of $E(2)$.

One can verify, using Eq. (1), the composition rule for $E(2)$ elements,

$$(R(\theta_1),x_1,y_1)\cdot(R(\theta_2),x_2,y_2)=(R(\theta_1+\theta_2),x_1+x_2\cos\theta_1-y_2\sin\theta_1,y_1+x_2\sin\theta_1+y_2\cos\theta_1). \tag{2}$$

From this composition rule, we can write a general element $(R(\theta),x,y)$ as the product $(1,x,y)\cdot(R(\theta),0,0)$, where $(1,0,0)=(R(\theta=0),0,0)$ is the unit element.

Throughout this paper, we will use complex coordinates, with $z=x+iy$. We can then obtain the 2×2 representations

$$\pi:(R(\theta),x,y)\equiv(R(\theta),z)\mapsto\begin{pmatrix}e^{i\theta} & z\\ 0 & 1\end{pmatrix},\quad \tilde{\pi}:(R(\theta),z)\mapsto\begin{pmatrix}1 & 0\\ \bar{z} & e^{-i\theta}\end{pmatrix}, \tag{3}$$

where $(R(\theta),z)$ now denotes an element of $E(2)$, and where the bar denotes complex conjugation. The composition rule now reads

$$(R(\theta_1),z_1)\cdot(R(\theta_2),z_2)=(R(\theta_1+\theta_2),z_1+z_2e^{i\theta_1}). \tag{4}$$

The full transformation in real space can be obtained from the real and imaginary parts of the complex transformation.

A motivation for our work is that $E(2)\sim[\mathbb{R}^2]\mathrm{SO}(2)$ represents the simplest nontrivial example of a semidirect product group, a family very useful in physics as it contains, amongst others, the rigid rotor group $[\mathbb{R}^5]\mathrm{SO}(3)$ of nuclear and molecular physics and the Poincaré group $[\mathbb{R}^4]\mathrm{SO}(3,1)$ of spacetime translations and boosts.

The starting point of our method is the Lie algebra $e(2)$ of the group $E(2)$. (We will jump freely between the algebra $e(2)$ and the group $E(2)$; all representations of $e(2)$ discussed here can be integrated to representations of $E(2)$.) Thus, suppose that $\pi(R(\theta),z)$ is a representation of $E(2)$ on a finite-dimensional space V. (It is a slight abuse of notation to write $\pi(R(\theta),z)$ because the representations will, in general, depend on both z and $\bar{z}$. However, this shorthand notation causes no problem. Technically speaking, we are thinking of z as an element of the complex plane, regarded as a *real* Lie group, not a complex Lie group.) Then, V decomposes into weight subspaces according to the action of SO(2),

$$V=\oplus W_k,$$

where

$$W_k=\{v\in V:\pi(R(\theta),0)v=e^{ik\theta}v\}, \tag{5}$$

where $k\in\mathbb{Z}$ so that $\pi(R(\theta+2\pi),z)=\pi(R(\theta),z)$ for representations of $E(2)$. We denote by

$$l_0=-i\frac{\partial}{\partial\theta}\pi(R(\theta),z)|_{\theta=z=0},\quad p_+=\frac{\partial}{\partial z}\pi(R(\theta),z)|_{\theta=z=0},\quad p_-=\frac{\partial}{\partial\bar{z}}\pi(R(\theta),z)|_{\theta=z=0}, \tag{6}$$

a basis for the $e(2)$ algebra, with nonzero commutation relations given by

$$[p_+,p_-]=0,\quad [l_0,p_\pm]=\pm p_\pm. \tag{7}$$

The elements p_+ and p_- are, respectively, ''raising'' and ''lowering'' operators, in the sense that

$$p_+W_k\subseteq W_{k+1},\quad p_-W_k\subseteq W_{k-1}. \tag{8}$$

In particular, for finite dimensional representations, they are nilpotent.

We have found that a useful and compact way of describing a representation of the $e(2)$ algebra is to display the result of Eqs. (7) and (8) in a graphical or diagrammatic form. We derive in Sec. II the rules for constructing representations of $e(2)$ that have no weight multiplicity. The tensor product of two such representations is simply obtained by combining their respective graphs in an appropriate way, as shown in Sec. II C. The resulting graph describes a representation of $e(2)$ which may or may not be decomposable; the problem of decomposing a tensor product turns out to be highly nontrivial, and we present in Sec. VIII some results on this issue.

A feature of tensor product representations and of certain other representations that we will present is that they typically contain indecomposable submodules with nontrivial weight multiplicities. One should recall that, thus far, the bulk of the results for $E(2)$ have dealt with unitary infinite dimensional representations, obtained either by induction or by the method of contraction,[2,3] where one considers representations of $E(2)$ as appropriate limits of representations of SU(2); in both cases, the weight multiplicity is never greater than 1. For the finite dimensional case, some of our representations can be thought of as smooth deformations of SU(2) representations. More generally, representations with trivial weight multiplicities are best accommodated inside the formalism of graded contractions[4] of SU(2), where the grading subgroup is the continuous subgroup $SO(2)\subset E(2)$. However, it is clear that the contraction of an SU(2) irrep cannot possibly yield a representation of $E(2)$ with nontrivial weight multiplicities. The possibility of constructing indecomposable modules containing arbitrarily high weight multiplicities is therefore, to our knowledge, completely new.

The representations of $E(2)$ that we construct belong to an identifiable family which, we think, is likely to contain many representations useful in physics. To illustrate this point, we give, in Sec. V, some explicit realizations of our representations. Moreover, the graphical method behind our results can certainly be adapted to more complicated semidirect product groups.[5]

II. STRING REPRESENTATIONS

In this section we discuss representations with weight multiplicities equal to 1, i.e., representations V for which, in the notation of (5), $\dim(W_k)\leqslant 1$, for all k. For such a representation, we let M and N be, respectively, the maximum and minimum nontrivial weights.

A. Some lemmas

Lemma 0: Every one-dimensional representation of $E(2)$ is of the form

$$\chi_k:(R(\theta),z)\mapsto e^{ik\theta}, \tag{9}$$

for some $k\in\mathbb{Z}$.

Proof: The translation subgroup T, i.e., the subgroup consisting of all elements of the form $(R(0),z)$, is the commutator subgroup of $E(2)$. So any one-dimensional representation of $E(2)$ must factor through the quotient $E(2)/T$, which is isomorphic to SO(2); the one-dimensional representations of SO(2) are of the specified form. □

Lemma 1: Let $0\neq|\varphi_k\rangle$ be an arbitrary vector in the one-dimensional subspace $W_k\subset V$. Then, at least one of $p_+|\varphi_k\rangle$ and $p_-(p_+|\varphi_k\rangle)$ must be zero for $[p_+,p_-]=0$ to be satisfied.

Proof: The raising and lowering operators p_+ and p_- are nilpotent and, since they commute, so is their product, the so(2)-invariant operator p_+p_-. The restriction of p_+p_- to any W_k subspace is therefore nilpotent. The only nilpotent operator on a one-dimensional space is the zero operator. If the W_k subspaces are all one-dimensional, this shows that $p_+p_-=0$ on V.

For an alternate, more explicit, proof, let $p_+p_-|\varphi_k\rangle=\alpha_k|\varphi_k\rangle$, where α_k is a proportionality constant. This holds since the subspace W_k is one-dimensional and p_+p_- is a weight-preserving operator. Since the representation is finite dimensional, there exists n such that $(p_+p_-)^n|\varphi_k\rangle=(p_+)^n(p_-)^n|\varphi_k\rangle=\alpha_k^n|\varphi_k\rangle=0$, from which it follows that $\alpha_k=0$. □

Proposition 1: If we specify on which subspaces W_m the raising and lowering operators are zero and on which they are nonzero, subject to the condition in lemma 1, this determines a unique representation of $e(2)$. The resulting representation is indecomposable if and only if p_+W_m and p_-W_{m+1} are not both zero for any m with $N\leqslant m<M$.

Proof: Because of the condition, we can choose a basis $\{|\varphi_m\rangle$ s.t. $l_0|\varphi_m\rangle=m|\varphi_m\rangle\}$ of eigenstates of l_0, with $|\varphi_m\rangle\in W_m$, for each m, and such that for each $m\in\{N,\ldots,M-1\}$, precisely one of the following holds:

(i) $p_+|\varphi_m\rangle=|\varphi_{m+1}\rangle$ and $p_-|\varphi_{m+1}\rangle=0$, which we represent by

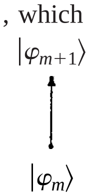

(ii) $p_+|\varphi_m\rangle=0$ and $p_-|\varphi_{m+1}\rangle=|\varphi_m\rangle$, with graph,

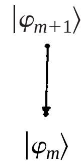

or

(iii) $p_+|\varphi_m\rangle=0$ and $p_-|\varphi_{m+1}\rangle=0$, i.e., there is no arrow between $|\varphi_{m+1}\rangle$ and $|\varphi_m\rangle$,

$$\begin{array}{c}|\varphi_m\rangle\\ \vdots\\ |\varphi_{m+1}\rangle\end{array}.$$

Relative to the basis $\{|\varphi_M\rangle,|\varphi_{M-1}\rangle,\ldots,|\varphi_{N+1}\rangle,|\varphi_N\rangle\}$, so(2) acts diagonally, p_+ is represented by a matrix which is zero except for a 1 immediately above the diagonal corresponding to each m for which possibility (i) above holds, and p_- is represented by a matrix which is zero except for a 1 immediately below the diagonal corresponding to each $m+1$ for which possibility (ii) above holds.

Clearly the matrices for p_+ and p_- commute. The remaining commutators $[l_0,p_\pm]=\pm p_\pm$ are satisfied since, for instance, $(l_0p_+-p_+l_0)|\varphi_m\rangle=l_0|\varphi_{m+1}\rangle-mp_+|\varphi_m\rangle=|\varphi_{m+1}\rangle=p_+|\varphi_m\rangle$ by construction.

If $p_+W_k=0=p_-W_{k+1}$, then

$$V=(\oplus_{m\leq k}W_m)\oplus(\oplus_{m>k}W_m) \tag{10}$$

is an $e(2)$-decomposition. Conversely, suppose $V=U\oplus U'$ is an $e(2)$-decomposition but that condition (iii) above does not hold for any $k\in\{N,N+1,...,M-1\}$. We can assume there exists $m\in\{N,N+1,...,M-1\}$ such that $W_m\subseteq U$, $W_{m+1}\subseteq U'$. Then either $p_+(W_m)$ or $p_-(W_{m+1})$ is nonzero. Since U and U' are both $e(2)$-spaces, this shows $U\cap U'\neq\{0\}$, a contradiction. □

Representations with weight multiplicities all equal to 1 will be called string representations.

B. String representations in graphical form

To a representation thus constructed, we can associate a graph as a mnemonic device to remember which of the conditions (i), (ii) or (iii) hold between two neighboring weight subspaces W_m and W_{m+1} by drawing an up arrow from W_m to W_{m+1} when (i) applies, a down arrow from W_{m+1} to W_m when (ii) applies, and no arrow when (iii) occurs; subgraphs of the type

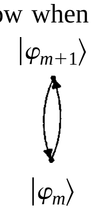

for which $p_+|\varphi_m\rangle\neq0$ and $p_-|\varphi_{m+1}\rangle\neq0$, cannot occur.

To obtain a representation of $E(2)$ relative to the chosen basis, we start by exponentiating separately the diagonal matrix of l_0 to obtain the image of $(R(\theta),0)\in$ SO(2), and the off-diagonal matrix elements of the generators of translations p_+ and p_- to obtain $(1,z)$. The element $(R(\theta),z)$ is then constructed from the matrix multiplication of $(1,z)\cdot(R(\theta),0)$.

For instance, the ''raising string'' representation of $e(2)$, with a graph consisting only of up arrows, exponentiates to the $E(2)$ representation,

$$\Rightarrow \pi_1:(R(\theta),z)\mapsto \begin{pmatrix} e^{iM\theta} & e^{i(M-1)\theta}z & e^{i(M-2)\theta}\frac{1}{2}z^2 & \dots & \frac{e^{i(N+1)\theta}}{(M-N+1)!}z^{M-N+1} & \frac{e^{iN\theta}}{(M-N)!}z^{M-N} \\ 0 & e^{i(M-1)\theta} & e^{i(M-2)\theta}z & \dots & \frac{e^{i(N+1)\theta}}{(M-N+2)!}z^{M-N+2} & \frac{e^{iN\theta}}{(M-N+1)!}z^{M-N+1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & e^{i(N+1)\theta} & e^{iN\theta}z \\ 0 & 0 & 0 & \dots & 0 & e^{iN\theta} \end{pmatrix}, \tag{11}$$

containing the SO(2) unirreps $M,M-1,\dots,N$ each with multiplicity 1. It is indecomposable.

The ''lowering string'' representation

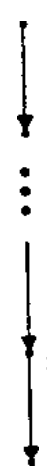

$$\Rightarrow \pi_2:(R(\theta),z)\mapsto \begin{pmatrix} e^{iM\theta} & 0 & 0 & \dots & 0 & 0 \\ e^{iM\theta}\bar{z} & e^{i(M-1)\theta} & 0 & \dots & 0 & 0 \\ e^{iM\theta}\frac{1}{2}\bar{z}^2 & e^{i(M-1)\theta}\bar{z} & e^{i(M-2)\theta} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \frac{e^{iM\theta}}{(M-N-1)!}\bar{z}^{M-N-1} & \frac{e^{i(M-1)\theta}}{(M-N-2)!}\bar{z}^{M-N-2} & \dots & \dots & e^{i(N+1)\theta} & 0 \\ \frac{e^{iM\theta}}{(M-N)!}\bar{z}^{M-N} & \frac{e^{i(M-1)\theta}}{(M-N-1)!}\bar{z}^{M-N-1} & \dots & \dots & e^{i(N+1)\theta}\bar{z} & e^{iN\theta} \end{pmatrix} \tag{12}$$

contains the SO(2) unirreps $M,M-1,\dots,N$ each with multiplicity 1, and nontrivial lowering operators between each pair of adjacent SO(2) subspaces. It is also indecomposable.

The five-dimensional representation with graph

$$\Rightarrow \pi_3:(R(\theta),z)\mapsto \begin{pmatrix} e^{2i\theta} & 0 & 0 & 0 & 0 \\ e^{2i\theta}\bar{z} & e^{i\theta} & z & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & e^{-i\theta} & e^{-2i\theta}z \\ 0 & 0 & 0 & 0 & e^{-2i\theta} \end{pmatrix} \tag{13}$$

is decomposable into two subspaces $V_1\oplus V_2$, containing respectively the SO(2) irreps 2,1,0 and $-1,-2$.

The three-dimensional representation

$$\Rightarrow \pi_4:(R(\theta),z)\mapsto \begin{pmatrix} e^{i\theta} & z & 0 \\ 0 & 1 & 0 \\ 0 & \bar{z} & e^{-i\theta} \end{pmatrix} \tag{14}$$

is indecomposable and equivalent to the ''natural'' representation of Eq. (1).

Note that, if π is an $E(2)$ representation containing the SO(2) irreps $M,M-1,\dots,N$, then $\chi_k\cdot\pi$ is another (inequivalent) representation containing the SO(2) irreps $M+k,M-1+k,\dots,N+k$.

C. Tensor product of two strings

Finally, it is also easy to represent the tensor product of two string representations in a graphical way. Thus, if V_1 and V_2 are two representations of $E(2)$ spanned, respectively, by $\{v_i, i=m_1, m_1-1, \ldots n_1\}$ and $\{w_j, j=m_2, m_2-1, \ldots, n_2\}$, then a basis for the tensor product representation $V_1 \otimes V_2$ is given by the points $v_i \otimes w_j$ having coordinates (i,j) on a two-dimensional grid. The arrows between point $v_i \otimes w_j$ and $v_k \otimes w_l$ are determined from the action of the $e(2)$ elements on v_i or w_j. Thus, for instance, consider the following tensor product:

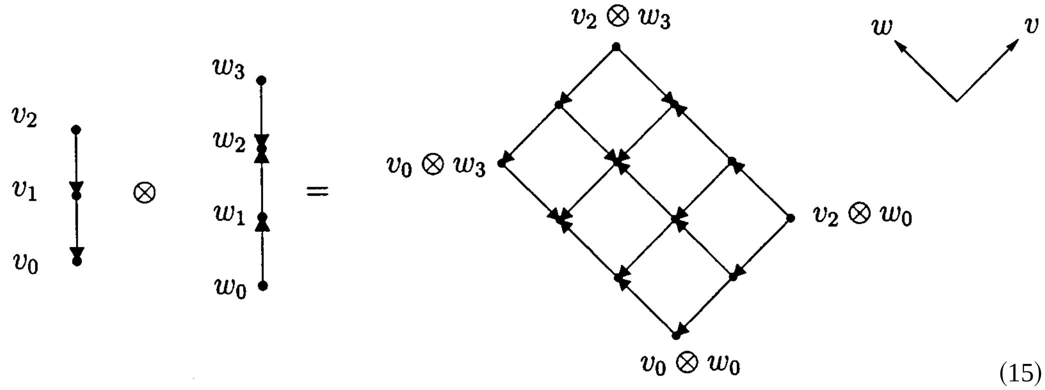

(15)

where the final two-dimensional graph has been tilted so that states with the same weight occur at the same horizontal height. (The ''corner'' states on the graph have been explicitly indicated.)

III. PARALLELOGRAM REPRESENTATIONS

A. The parallelogram representation as tensor product

Consider the representation

$$\pi_5 : (R(\theta), z) \mapsto \begin{pmatrix} 1 & 0 \\ \bar{z} & e^{-i\theta} \end{pmatrix} \otimes \begin{pmatrix} e^{i\theta} & z \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{i\theta} & z & 0 & 0 \\ 0 & 1 & 0 & 0 \\ e^{i\theta}\bar{z} & z\bar{z} & 1 & e^{-i\theta} z \\ 0 & \bar{z} & 0 & e^{-i\theta} \end{pmatrix}, \tag{16}$$

which is obtained from the tensor product of the two-dimensional lowering string representation and the two-dimensional raising string representation, with graph

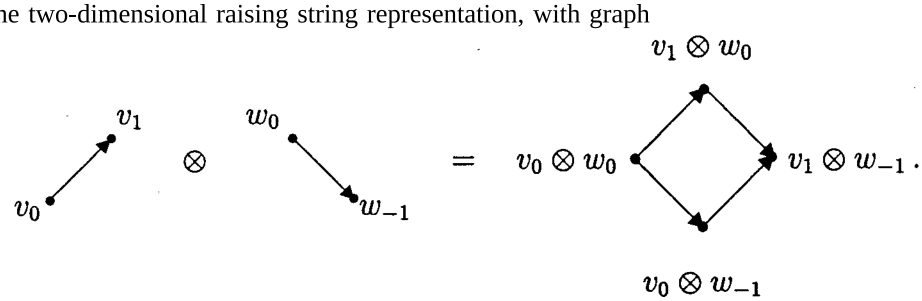

(17)

Claim: The representation π_5 is indecomposable.

Proof: Otherwise suppose $V = U \oplus U'$ is a nontrivial decomposition, and that one of the two subspaces, say U, contains a vector v in the two-dimensional subspace W_0 of weight 0 of the form $v = \alpha_1(v_0 \otimes w_0) + \alpha_2(v_1 \otimes w_{-1})$, with $\alpha_1 \neq 0$. By acting with $p_+ p_-$, we find $p_+ p_- v = \alpha_1(v_1 \otimes w_{-1})$ must also be in U. Thus, $v_1 \otimes w_{-1} \in W_0$ is also in U (since $\alpha_1 \neq 0$) and so $v - \alpha_2(v_1 \otimes w_{-1}) = \alpha_1(v_0 \otimes w_0) \in U$. Since $p_+(v_0 \otimes w_0) = v_1 \otimes w_0 \in U$ and $p_-(v_0 \otimes w_0) = v_0 \otimes w_{-1} \in U$ as well, we find that $U = V, U' = \{0\}$, a contradiction. □

This generalizes to larger representations. For $M \geqslant 0$, let V_M be the ''raising-string'' representation with lowest weight 0 and highest weight M. It has a weight basis consisting of the lowest weight vector v_0 of weight 0, and nonzero weight vectors $v_k = (p_+)^k v_0$, for $k = 1,...,M$, with $p_+(v_M) = 0$; the dimension is $M+1$. Similarly, for $N \geqslant 0$, let V_{-N} be the ''lowering-string'' representation with lowest weight $-N$ and highest weight 0. It has a weight basis consisting of the highest weight vector w_0 of weight 0, and nonzero weight vectors $w_{-k} = (p_-)^k w_0$, for $k = 1,...,N$, with $p_-(w_{-N}) = 0$; the dimension is $N+1$. The action of p_- on V_M and the action of p_+ on V_{-N} are both trivial.

The ''parallelogram'' representation $V_{M,-N}$ is the tensor product $V_{M,-N} = V_M \otimes V_{-N}$; it has lowest weight $-N$, highest weight M, and dimension $(M+1)(N+1)$. Since $v_k \otimes w_{-l} = (p_+)^k (p_-)^l v_0 \otimes w_0 = (p_-)^l (p_+)^k v_0 \otimes w_0$, we see that $V_{M,-N}$ is generated by the weight vector $v_0 \otimes w_0$, which we call the ''initial vector.'' Twisting by the character $\chi_r : (R(\theta), z) \mapsto e^{ir\theta}$, $r \in \mathbb{Z}$, gives a parallelogram representation $V_{M,-N;r} = \chi_r \otimes V_{M,-N}$ with lowest weight $r-N$, highest weight $r+M$, and dimension $(M+1)(N+1)$; it is generated by the ''initial vector'' $\chi_r \otimes v_0 \otimes w_0$.

For instance, the graph

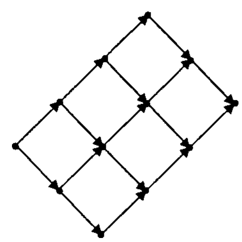

(18)

is associated with the $V_{3,-2;r}$ parallelogram representation.

Lemma 2: The parallelogram representation $V_{M,-N;r}$ is indecomposable.

Proof: Let $V_{M,-N;r} = U \oplus U'$ be a decomposition. Choose a basis T whose first vector is $\phi_0 = \chi_r \otimes v_0 \otimes w_0$ and whose first d_r vectors $\phi_0, p_+p_-\phi_0, (p_+p_-)^2\phi_0, \ldots, (p_+p_-)^{d_r-1}\phi_0$ span the weight subspace W_r (of dimension d_r). Writing vectors in terms of this basis, we have that the initial vector is $\phi_0 = \chi_r \otimes v_0 \otimes w_0 = (1,0,0,\ldots)^T$. Let $v = (\alpha_1, \alpha_2, \ldots, \alpha_{d_r}, 0, \ldots)^T, \alpha_1 \neq 0$, be an otherwise arbitrary vector in W_r. Assume that $v \in U$, and consider

$$
\begin{aligned}
p_+p_- v &= (0, \alpha_1, \alpha_2, \ldots, \alpha_{d_r-1}, 0, \ldots)^T \in U, \\
(p_+p_-)^2 v &= (0,0,\alpha_1, \alpha_2, \ldots, \alpha_{d_r-2}, 0, \ldots)^T \in U, \\
&\vdots \\
(p_+p_-)^{d_r-3} v &= (0, \ldots, \alpha_1, \alpha_2, \alpha_3, 0, \ldots)^T \in U, \\
(p_+p_-)^{d_r-2} v &= (0, \ldots, \alpha_1, \alpha_2, 0, \ldots)^T \in U, \\
(p_+p_-)^{d_r-1} v &= (0, \ldots, 0, \alpha_1, 0, \ldots)^T \in U.
\end{aligned} \tag{19}
$$

Thus, the d_rth basis vector

$$
v_{d_r} = \frac{1}{\alpha_1}(p_+p_-)^{d_r-1}(\chi_r \otimes v_0 \otimes w_0) = (0, \ldots, 1, 0, \ldots)^T \in U \tag{20}
$$

(since $\alpha_1 \neq 0$ by assumption). If this vector is in U, then the vector

$$v_{d_r-1}=\frac{1}{\alpha_1}((p_+p_-)^{d_r-2}v-\alpha_2 v_{d_r})=(0,\ldots,1,0,0,\ldots)^T\in U \tag{21}$$

as well, and so is

$$v_{d_r-2}=\frac{1}{\alpha_1}((p_+p_-)^{d_r-3}v-\alpha_2 v_{d_r-1}-\alpha_3 v_{d_r})=(0,\ldots,1,0,0,0,\ldots)^T\in U, \tag{22}$$

and so forth until one shows that all basis vectors in W_r are in U. In effect, this argument is based on the observation that the matrix of the restriction to each weight space of the operator p_+p_- is a triangular matrix.

In particular, the initial vector $\chi_r\otimes v_0\otimes w_0$ is in U and, since an arbitrary basis state $\chi_r\otimes v_s\otimes w_{-q}\in V$ can be obtained as $(p_+)^s(p_-)^q(\chi_r\otimes v_0\otimes w_0)$, it follows that $U=V$ and $U'=\{0\}$, which shows that V is indecomposable. □

In this way, we can construct indecomposable representations with arbitrarily high weight multiplicities.

B. Subrepresentations of the parallelogram

Observation: A vector $X_{k,-l,r}=\chi_r\otimes v_k\otimes w_{-l}=(p_+)^k(p_-)^l(\chi_r\otimes v_0\otimes w_0)$ generates a subrepresentation of $V_{M,-N;r}$ that is isomorphic to $V_{M-k,l-N;r+k-l}$. An example of this is given in Eq. (23), where, in $V_{3,-2;0}$, a 2×2 parallelogram subrepresentation $V_{1,-1;1}$ is generated by $X_{2,1,0}=(p_+)^2p_-(v_0\otimes w_0)$,

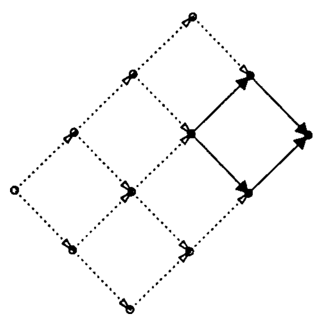

(23)

Proposition 2: Every subrepresentation of a parallelogram representation $V_{M,-N;r}$ can be expressed as the subrepresentation generated by finitely many ''initial vectors'' of the form $(p_+)^k(p_-)^l(\chi_r\otimes v_0\otimes w_0)$, for suitable k, l. The subrepresentations of $V_{M,-N;r}$ are all indecomposable.

Proof: First, note that the weight subspace $W_s\subset V_{M,-N;r}$ is spanned by vectors of the form $\{(p_+)^k(p_-)^l(\chi_r\otimes v_0\otimes w_0)|k-l=s-r\}$. The action of p_+p_- on this basis is to map

$$p_+p_-:(p_+)^k(p_-)^l(\chi_r\otimes v_0\otimes w_0)\mapsto(p_+)^{k+1}(p_-)^{l+1}(\chi_r\otimes v_0\otimes w_0). \tag{24}$$

Now, suppose $U\subset V_{M,-N;r}$ is an indecomposable subrepresentation with nontrivial intersection with the weight space W_s. Choose the minimal k such that U contains a nonzero vector $v\in W_s$ of the form

$$v=\sum_{i\geq 0}\alpha_i(p_+)^{\tilde{k}+i}(p_-)^{\tilde{l}+i}(\chi_r\otimes v_0\otimes w_0),\ \alpha_0\neq 0,\ \tilde{k}-\tilde{l}=s-r. \tag{25}$$

Then, by the p_+p_- argument of Lemma 2, all vectors of the form

$$(p_+)^{\tilde{k}+i}(p_-)^{\tilde{l}+i}(\chi_r\otimes v_0\otimes w_0),\ i\geqslant 0, \tag{26}$$

must be in U, and $(p_+)^{\tilde{k}}(p_-)^{\tilde{l}}(\chi_r\otimes v_0\otimes w_0)$ is an initial vector for the subspace of weight s in U.

Do this for each weight s of U and discard redundant choices (vectors as above that are contained in the subspace generated by another such vector). The remaining finitely many ''initial vectors'' generate the whole space U. Note that the initial vectors have distinct weights.

Now suppose $U=U_1\oplus U_2$ is a decomposition. Fix one of the initial vectors described above. Then one of U_1, U_2, say U_1, contains a weight vector with nonzero overlap with this chosen initial vector. But then U_1 actually contains that initial vector, because of the triangularity of the operator p_+p_-, as above. If there is only one initial vector, then we are done, by Lemma 2 and the previous Observation. Otherwise, notice that once U_1 contains $(p_+)^k(p_-)^l(\chi_r\otimes v_0\otimes v_0)$, it must also contain the ''final vector'' $(p_+)^m(p_-)^n(\chi_r\otimes v_0\otimes v_0)=(p_+)^{m-k}\times(p_-)^{n-l}((p_+)^k(p_-)^l(\chi_r\otimes v_0\otimes v_0))$. But then U_1 must contain all the initial vectors, and hence all of U, so U is indecomposable. □

Corollary: A basis for any subrepresentation of the parallelogram representation $V_{m,-n;r}$ *is given by vectors of the form* $(p_+)^k(p_-)^l(\chi_r\otimes v_0\otimes v_0)$ *contained in this subrepresentation.*

C. Quotient representations

Consider the representation associated with the graph

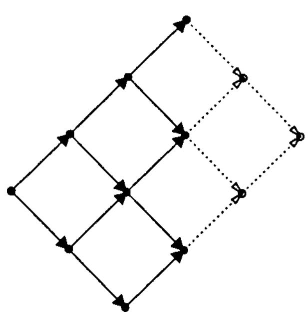

$$\Rightarrow \pi_6:(R(\theta),z)\mapsto \begin{pmatrix} e^{4i\theta} & ze^{3i\theta} & \frac{1}{2}z^2e^{2i\theta} & 0 & \frac{1}{6}z^3e^{i\theta} & 0 & 0 & 0 & 0 \\ 0 & e^{3i\theta} & ze^{2i\theta} & 0 & \frac{1}{2}z^2e^{i\theta} & 0 & 0 & 0 & 0 \\ 0 & 0 & e^{2i\theta} & 0 & ze^{i\theta} & 0 & 0 & 0 & 0 \\ 0 & \bar{z}e^{3i\theta} & 0 & e^{2i\theta} & 0 & ze^{i\theta} & \frac{1}{2}z^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & e^{i\theta} & 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{z}e^{2i\theta} & 0 & 0 & e^{i\theta} & z & 0 & 0 \\ 0 & 0 & 0 & 0 & \bar{z}e^{i\theta} & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2}z^2e^{2i\theta} & 0 & 0 & ze^{i\theta} & 0 & 1 & ze^{-i\theta} \\ 0 & 0 & 0 & 0 & \frac{1}{2}\bar{z}^2e^{i\theta} & 0 & \bar{z} & 0 & e^{-i\theta} \end{pmatrix}. \tag{27}$$

This representation can be constructed by starting from the parallelogram $V_{3,-2;0}$, and by removing (or, equivalently, setting to 0) each line and column corresponding to a node in the subrepresentation with initial vectors $v_3\otimes w_{-1}$ and $v_2\otimes w_{-2}$. It is therefore a quotient of the parallelogram $V_{3,-2;r}$.

Lemma 3: Quotient representations of a parallelogram are indecomposable.

Proof: Such a quotient representation has a single initial vector, $v=v_0\otimes w_0$. Let $V=V_1\oplus V_2$ be a decomposition. By the p_+p_- triangularity argument, the whole weight space containing v must be in either V_1 or V_2; without loss of generality, we assume that it is in V_1, and thus v itself is in V_1. Since everything in V can be obtained from $v=v_0\otimes w_0$ by acting with $e(2)$ generators, it follows that $V_1=V, V_2=\{0\}$, i.e., V is indecomposable. □

It is also possible to take quotients of subrepresentations of a parallelogram, as shown on the left-hand side of Eq. (28), or even to make a string representation out of a quotient of a subparallelogram, as shown on the right-hand side of that figure.

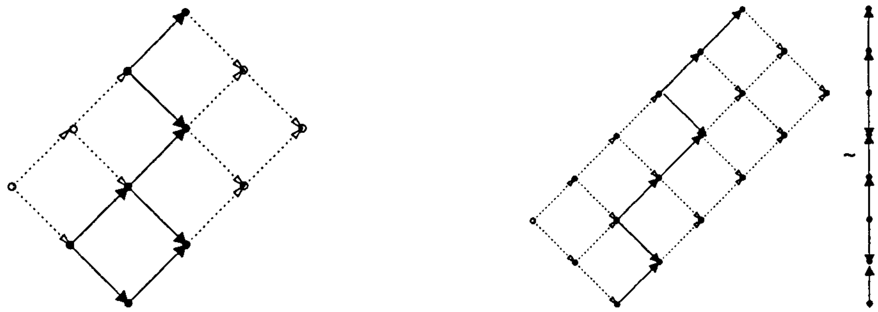

(28)

Thus, we can claim that

Proposition 3: A subquotient of a subrepresentation of the parallelogram $V_{m,-n;r}$ (i.e., the quotient of two subrepresentations of the parallelogram $V_{m,-n;r}$) is indecomposable, provided that its graph is connected.

Proof: Let i,j uniquely label a minimal set of initial vectors $X_{i,j,r}=(p_+)^i(p_-)^j(\chi_r)\otimes v_0 \otimes w_0)$ of weight $r+i-j$ in such a representation V. Because initial vectors cannot have the same weight, the elements in the set $\{X_{i,j,k}\}$ can be ordered by increasing weight. These initial vectors are also those of a subrepresentation of the parallelogram $V_{m,-n;r}$. Let $V=V_1\oplus V_2$. Using the p_+p_- triangularity argument, the vectors generated from an initial vector all belong to the same subspace, either V_1 or V_2, depending on whether the initial vector is in V_1 or V_2. However, the subspaces generated by two consecutive initial vectors must have a nontrivial intersection, for otherwise the graph would be disconnected (into the vectors in or above the higher subspace and the vectors in or below the lower subspace).

By the p_+p_- triangularity argument, each weight space in this intersection must belong to only one of V_1 or V_2. Thus, the subspaces generated by two consecutive initial vectors must belong to the same subspace, and, continuing this way, all subspaces must belong to the same subspace, say V_1. This means that V_2 is empty and we are done. □

The last noteworthy result on quotients of subparallelograms is as follows:

Proposition 4: Given a (connected) string representation, there exists a subquotient of a parallelogram which is isomorphic to this string.

Proof: In the string representation, let m be the number of up arrows, n the number of down arrows, and construct $V_{m,-n;r}$, with r adjusted so that the highest weight of $V_{m,-n;r}$ is equal to the highest weight of the string, i.e., $m+r$. If the topmost arrow of the string is a down arrow then the highest weight state of the representation, $\chi_r\otimes v_m\otimes w_0$ is the heaviest initial vector of the subrepresentation. Otherwise, the initial vectors are ''sources'' of the type

unless the bottom-most arrow of the string is an up arrow, in which case the lightest initial vector is the lowest weight vector of the representation, $\chi_r\otimes v_0\otimes w_{-N}$. The heaviest initial vector, for instance, is always of the form $\chi_r\otimes v_l\otimes w_0$. Next, we note that terminal vectors are either highest or lowest weight vectors or ''sinks'' of the type

If the topmost arrow in the string is a down arrow, then $\chi_r\otimes v_n\otimes w_0$ is an initial vector; if the bottommost arrow is an up arrow, then $\chi_r \otimes v_0 \otimes w_{-m}$ is an initial vector. For each ''source,'' as described above, V has an initial vector $\chi_r \otimes v_k \otimes w_{-l}$, where $m-k$ is the number of up arrows above the source and $n-l$ is the number of down arrows below it.

Having identified the initial vectors, one then constructs the corresponding subrepresentation V of $V_{m,-n;r}$. Consider now the subrepresentation $V' \subset V$ obtained by applying the operator $p_+ p_-$ to V, i.e., $V' = p_+ p_-(V)$. The original string is the subquotient V/V'. □

IV. GLUING

Not all representations need be quotients or subquotients of parallelograms. For instance, consider the representation

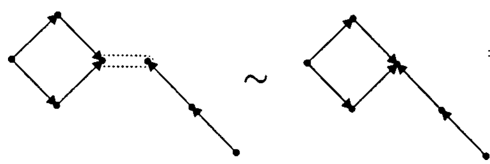

$$\Rightarrow \pi_7 : (R(\theta), z) \mapsto \begin{pmatrix} e^{i\theta} & z & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ e^{i\theta}\overline{z} & z\overline{z} & 1 & e^{-i\theta}z & e^{-i\theta}z & e^{-2i\theta}\frac{1}{2}z^2 \\ 0 & \overline{z} & 0 & e^{-i\theta} & 0 & 0 \\ 0 & 0 & 0 & 0 & e^{-i\theta} & e^{-2i\theta}z \\ 0 & 0 & 0 & 0 & 0 & e^{-2i\theta} \end{pmatrix} \tag{29}$$

with $\dim(W_1)=1$, $\dim(W_0)=\dim(W_{-1})=2$, $\dim(W_{-2})=1$. It is constructed by identifying the terminal vector of the 2×2 parallelogram representation of Eq. (17) with the terminal node of an extra two-element raising string ending at the terminal node, i.e., by ''gluing'' the parallelogram and the raising string at one node in the manner indicated by the graph.

Claim: The representation π_7 is indecomposable.

Proof: Otherwise suppose $V = U \oplus U'$ is a nontrivial decomposition. We can assume U contains the weight space W_{-2}. In this case, it also contains the standard basis vectors $\hat{e}_5 = (0,0,0,0,1,0)^T$ and $\hat{e}_4 = (0,0,0,1,0,0)^T$. But this last vector is the ''terminal node'' of the subspace which is isomorphic to π_5, and the argument given in connection to this representation shows that either U or U' must contain all of this subspace. Since we have already seen that U must contain e_4, it must contain all of the π_5 subspace, and therefore must be all of V. □

This can be generalized to other more complicated examples. For instance, the representation

π_8 :

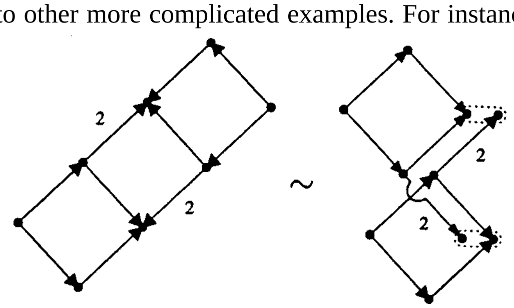

(30)

can be realized as shown as a two-point gluing. Note that the factors of 2 in the graph indicate that the corresponding matrix element is of strength 2 rather than 1, as has been thus far assumed throughout this paper. The representation π_8 can also be shown to be indecomposable.

Not all gluings are indecomposable. The representation of Fig. (53), which is obtained by gluing together two strings, is decomposable, as will be seen in Sec. VII.

V. EXPLICIT REALIZATIONS

In this section, we would like to exhibit some of the more interesting amongst the many explicit realizations of the $e(2)$ algebra, along with examples of the kind of representations that we have so far described.

Nonunitary representations occur in physics mostly as representations carried by tensor operators. Thus, if $\hat{T}_i^\lambda$ is the ith component of a tensor operator $\hat{T}^\lambda$ transforming by the representation λ, and if $\mathcal{O}$ is a representation of an element of $e(2)$ by linear operators, then the action of $\mathcal{O}$ on $\hat{T}^\lambda$ is given by

$$\mathcal{O}:\hat{T}_i^\lambda \mapsto [\mathcal{O},\hat{T}_i^\lambda] = \sum_j \alpha_{ij}\hat{T}_j^\lambda, \tag{31}$$

where α_{ij} are, in general, complex coefficients.

A. The adjoint representation

Let,

$$p_+ \mapsto \tfrac{1}{2}e^{-i\theta}\left(\frac{\partial}{\partial r} - \frac{i}{r}\frac{\partial}{\partial\theta}\right),\ p_- \mapsto \tfrac{1}{2}e^{i\theta}\left(\frac{\partial}{\partial r} + \frac{i}{r}\frac{\partial}{\partial\theta}\right),\ l_0 \mapsto i\frac{\partial}{\partial\theta}. \tag{32}$$

If we let these operators act on one another, we obtain the adjoint representation of $e(2)$, with the graph given in Eq. (14).

B. Raising string or lowering string representations

Let

$$p_+ \mapsto e^{i\theta}\frac{\partial}{\partial r},\ p_- \mapsto 0,\ l_0 \mapsto -i\frac{\partial}{\partial\theta}, \tag{33}$$

and consider the set of polynomials $\{e^{ik\theta}r^q, e^{i(k+1)\theta}r^{q-1}, e^{i(k+2)\theta}r^{q-2}, \ldots, e^{i(k+q)\theta}\}$, with $k \in \mathbb{Z}, q \in \mathbb{Z}^+$.

Using Eq. (31) and the realization of Eq. (33), one sees that these polynomials carry a representation equivalent to a raising string representation of dimension $q+1$ with lowest weight k.

Similarly, the realization

$$p_+ \mapsto 0,\ p_- \mapsto e^{-i\theta}\frac{\partial}{\partial r},\ l_0 \mapsto -i\frac{\partial}{\partial\theta}, \tag{34}$$

action on the polynomials $\{e^{in\theta}r^q, e^{i(n-1)\theta}r^{q-1}, \ldots, e^{i(n-q)\theta}\}$, with $n \in \mathbb{Z}, q \in \mathbb{Z}^+$, is a lowering string representation of dimension $q+1$ with lowest weight $n-q$.

We obtain a less trivial example if we let the operators of Eq. (32) act on the polynomials $\{r^t e^{it\theta}, r^{t-1}e^{i(t-1)\theta}, \ldots, 1\}$, which yields a raising string representation with highest weight vector $r^t e^{it\theta}$ of weight t.

In a similar way, the polynomials $\{r^t e^{-it\theta}, r^{t-1}e^{-i(t-1)\theta}, \ldots, 1\}$ span a lowering string representation of $E(2)$ under the action of the operators of Eq. (32).

C. Parallelograms

Consider the three operators

$$p_+ = \frac{\partial^2}{\partial x^2} + 2i\frac{\partial^2}{\partial x\partial y} - \frac{\partial^2}{\partial y^2} = \frac{\partial^2}{\partial \overline{w}^2},$$

$$p_-=\frac{\partial^2}{\partial x^2}-2i\frac{\partial^2}{\partial x\partial y}-\frac{\partial^2}{\partial y^2}=\frac{\partial^2}{\partial w^2}, \tag{35}$$

$$l_0=-\frac{i}{2}\left(x\frac{\partial}{\partial y}-y\frac{\partial}{\partial x}\right)=w\frac{\partial}{\partial w}-\bar{w}\frac{\partial}{\partial \bar{w}}. \tag{36}$$

With the identification $p_\pm \leftrightarrow \hat{Q}_{\pm 2}, l_0 \leftrightarrow \frac{1}{2}\hat{l}_z$, these are easily recognized as forming an $e(2)$ subalgebra of $[\mathbb{R}]^5$so(3), the algebra of the rigid rotor.

We now consider the operators

$$\hat{X}_{LM}=\sqrt{\frac{2^{L+M}(L+M)!(L-M)!L!}{(2L)!}}\sum_t \frac{\omega^t\tilde{\omega}^{t-M}z^{L+M-2t}}{2^t t!(t-M)!(L+M-2t)!}, \tag{37}$$

which can be recognized as proportional to the spherical harmonics

$$Y_{LM}(\theta,\varphi)=\sqrt{\frac{(2L+1)!}{4\pi 2^L L!}}\hat{X}_{LM} \tag{38}$$

if we make the identifications

$$\omega=\frac{-e^{i\varphi}\sin\theta}{\sqrt{2}},\quad z=\cos\theta,\quad \tilde{\omega}=\frac{e^{-i\varphi}\sin\theta}{\sqrt{2}}. \tag{39}$$

In general, we have

$$[p_\pm,\hat{X}_{LM}]=\alpha_{LM}\hat{X}_{L-2,M\pm 2}, \tag{40}$$

with

$$\alpha_{LM}=\langle \hat{X}_{L-2,M\pm 2}|p_\pm|\hat{X}_{LM}\rangle, \tag{41}$$

with the upper and lower sign valid for $p_\pm$, respectively, and where the number $\langle \hat{X}_{L-2,M\pm 2}|p_\pm|\hat{X}_{LM}\rangle$ is computed using the standard boson inner product. With these we can build a variety of parallelograms. (Note that the $e(2)$ weight is $\frac{1}{2}M$.)

Thus, for instance, the set of nine operators $\{\hat{X}_{80},\hat{X}_{6,\pm 2},\hat{X}_{4,\pm 4},\hat{X}_{40},\hat{X}_{2,\pm 2},\hat{X}_{00}\}$ are the components of a tensor which carries the parallelogram representation $V_{2,-2,0}$ with

$$\langle \hat{X}_{L-2,M\pm 2}|p_\pm|\hat{X}_{LM}\rangle=\sqrt{\frac{(2L+1)L(L-1)}{2L-3}}(L,M;2,\pm 2|L-2,M\pm 2), \tag{42}$$

where $(L,M;2,\pm 2|L-2,M\pm 2)$ is an SO(3) Clebsch–Gordan coefficient.

Similarly, the set $\{\hat{X}_{71},\hat{X}_{51},\hat{X}_{5,-3},\hat{X}_{33},\hat{X}_{3,-1},X_{11}\}$ are the six components of a tensor which carries the parallelogram representation $V_{2,-1;1}$.

Subparallelograms of these or of any parallelogram are obtained by simply removing from the original parallelogram all the basis polynomials not contained in the subrepresentations.

It is also possible to obtain parallelograms from the realization of Eq. (32). Repeated action of these operators on the initial vector $r^p e^{in\theta}$ produce the parallelogram representation $V_{(1/2)(p+n),-(1/2)(p-n);n}$. Note that, since $\frac{1}{2}(p+n)$ and $\frac{1}{2}(p-n)$ must be integers, $p+n$ must be even for the representation to remain finite dimensional.

The realization of Eq. (32) is a special case of the more general realization

$$p_+ \mapsto r^x e^{i(x-1)\theta}\left(\frac{\partial}{\partial r} - \frac{i}{r}\frac{\partial}{\partial \theta}\right), p_- \mapsto r^{-y} e^{i(y+1)\theta}\left(\frac{\partial}{\partial r} + \frac{i}{r}\frac{\partial}{\partial \theta}\right),$$

$$l_0 \mapsto i\frac{(x-y-2)}{2(x-1)(y+1)}\frac{\partial}{\partial \theta} + \frac{(x+y)}{2(x-1)(y+1)} r \frac{\partial}{\partial r}, \tag{43}$$

from which we can extract various parallelograms. (The exponents m,n of the initial vector $r^m e^{in\theta}$ must satisfy some conditions if the representation is to remain finite dimensional.)

D. Quotients and strings

Consider now the realization

$$p_+ \mapsto e^{i\theta}\frac{\partial^2}{\partial r^2},\ p_- \mapsto e^{-i\theta}\frac{\partial}{\partial r},\ l_0 = -i\frac{\partial}{\partial \theta}. \tag{44}$$

Acting on the initial vector r^4 to generate a parallelogram, and removing from this parallelogram the basis states r^4 and $r^3 e^{-i\theta}$, we obtain the string representation

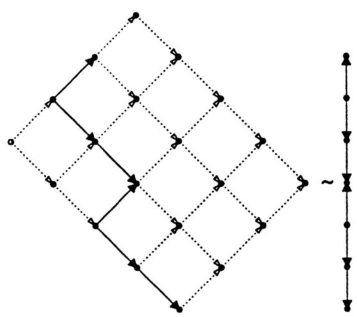

$$\Rightarrow \pi_9:(R(\theta),w) \mapsto \begin{pmatrix} e^{-4i\theta} & \bar{w}e^{-3i\theta} & \frac{1}{2}\bar{w}^2 e^{-2i\theta} & 0 & 0 & 0 & 0 \\ 0 & e^{-3i\theta} & \bar{w}e^{-2i\theta} & 0 & 0 & 0 & 0 \\ 0 & 0 & e^{-2i\theta} & 0 & 0 & 0 & 0 \\ 0 & 0 & we^{-2i\theta} & e^{-i\theta} & \bar{w} & \frac{1}{2}\bar{w}^2 e^{i\theta} & 0 \\ 0 & 0 & 0 & 0 & 1 & \bar{w}e^{i\theta} & 0 \\ 0 & 0 & 0 & 0 & 0 & e^{i\theta} & 0 \\ 0 & 0 & 0 & 0 & 0 & we^{i\theta} & e^{2i\theta} \end{pmatrix}, \tag{45}$$

spanned by the basis states $\{e^{2i\theta}, e^{i\theta}r^2, r, e^{-i\theta}, e^{-2i\theta}r^2, e^{-3i\theta}r, e^{-4i\theta}\}$.

A nice feature of this string is that it represents an example of a $\mathbb{Z}_3$ graded-contraction of su(2) into $e(2)$, as discussed in Ref. 4. The grading is generated by the SO(2) subgroup $e^{(2\pi/3)l_0}$. The carrier space V decomposes into $V = V_0 \oplus V_1 \oplus V_2$, which contain the basis states

$$V_0 = \{r, e^{-3i\theta}r\},\ V_1 = \{e^{i\theta}r^2, e^{-2i\theta}r^2\},\ V_2 = \{e^{2i\theta}, e^{-i\theta}, e^{-4i\theta}\}, \tag{46}$$

while the basis elements of Eq. (44) have grades 1, 2, and 0, respectively.

VI. TENSORING TWO RAISING OR TWO LOWERING STRINGS

In Sec. III A, we considered the tensoring of a raising and a lowering string representation. The resulting parallelogram, as well as all its subrepresentations and ''connected'' subquotients, were shown to be indecomposable.

The next simplest example of tensor product is the product of two raising or two lowering string representations. This yields a tensor product representation that is always decomposable into a sum of raising or lowering string representations. The decomposition is closely related to the decomposition of SU(2) tensor products.

In this section, we discuss only the tensor product of two raising string representations since the case of two lowering strings is handled in the exact same way.

A. Raising strings as contractions of su(2) irreps

Consider the su(2) algebra spanned by the operators $\{L_+, L_-, L_0\}$, with the usual nonzero commutation relations

$$[L_0, L_\pm] = \pm L_\pm, \quad [L_+, L_-] = 2L_0. \tag{47}$$

If we now rescale the su(2) generators to

$$L_- \rightarrow \mathcal{L}_- = \epsilon L_-, \quad L_+ \rightarrow \mathcal{L}_+ = L_+, \quad L_0 \rightarrow \mathcal{L}_0 = L_0, \tag{48}$$

express the su(2) commutation relations in terms of the $\mathcal{L}$ generators, and take the limit as $\epsilon \rightarrow 0$, we see that

$$\begin{aligned} [\mathcal{L}_+, \mathcal{L}_-] &= \lim_{\epsilon\rightarrow 0} \epsilon [L_+, L_-] = \lim_{\epsilon\rightarrow 0} \epsilon 2 L_0 = \lim_{\epsilon\rightarrow 0} 2\epsilon \mathcal{L}_0 = 0, \\ [\mathcal{L}_0, \mathcal{L}_-] &= \lim_{\epsilon\rightarrow 0} \epsilon [L_0, L_-] = \lim_{\epsilon\rightarrow 0} -\epsilon L_- = -\mathcal{L}_-, \end{aligned} \tag{49}$$

with the commutator $[\mathcal{L}_+, \mathcal{L}_0]$ remaining unchanged from the su(2) commutator. We recover the algebra $e(2)$ by the identification $\mathcal{L}_\pm = p_\pm, \mathcal{L}_0 = l_0$, but, because the generators L_- and L_+ have been treated asymmetrically, the resulting representation cannot be unitary.

This is an example of a contraction.[2] The more familiar example of rescaling, where L_- and L_+ are treated on the same footing and both multiplied by the scale factor ε, leads to unitary infinite dimensional representations. Coming back to Eq. (49), suppose that we are given a standard unitary representation Γ_j of the su(2) algebra, of dimension, say, $2j+1$. The effect of taking the $\epsilon \rightarrow 0$ limit of the asymmetric scaling is to leave matrix elements of $\mathcal{L}_+ = L_+$ and $\mathcal{L}_0 = L_0$ unchanged, while setting to 0 the matrix elements of $\mathcal{L}_- = \epsilon L_-$. The resulting representation, where $p_- = \mathcal{L}_- = 0$ everywhere but where $p_+ = \mathcal{L}_+$ acts by raising the weight of the su(2) states, is clearly equivalent to a $(2j+1)$-dimensional raising string representation γ_j. The equivalence relation just rescales the nonzero su(2) matrix elements of L_+ to the standard $e(2)$ matrix elements of p_+, which are 1. Thus, we can write, for a raising string representation

$$\gamma_j = \lim_{\epsilon\rightarrow 0} \Gamma_j. \tag{50}$$

The lowest weight of γ_j is $-j$. A general raising string representation $\gamma_{j;k}$ of dimension $2j+1$ with lowest weight $-j+k$ can be obtained by twisting γ_j by a character χ_k.

(Note that one can obtain a lowering string representation from an su(2) representation by scaling $L_+ \rightarrow \mathcal{L}_+ = \epsilon L_+$ and leaving the other two generators unchanged.)

The advantage of introducing limits in such a fashion is that one can then think of raising string representations as smooth deformations of su(2) representations.

B. Decomposing tensor products of two raising strings

First, recall that the tensor product $j_1 \otimes j_2$ of two su(2) representations j_1 and j_2 of dimensions $2j_1+1$ and $2j_2+1$, respectively, decomposes into a sum of su(2) representations of dimension $2j+1$, with possible values of j given by $|j_1-j_2|, |j_1-j_2|+1, |j_1-j_2|+2, \ldots, j_1+j_2$, and where each value of j allowed by the inequality occurs exactly once.

Proposition 5: The tensor product of two raising string representations $\gamma_{j_1;r_1} \otimes \gamma_{j_2;r_2}$ *of dimensions* $2j_1+1$ *and* $2j_2+1$, *respectively, decomposes into a sum of raising string representations. The representations occuring in this tensor product have the dimensions* $2j+1$, *where* j *takes the possible values* $|j_1-j_2|, |j_1-j_2|+1, |j_1-j_2|+2, \ldots, j_1+j_2$, *and where each value of* j *allowed by the inequality occurs exactly once.*

Proof: Write

$$\begin{aligned}\gamma_{j_1;r_1}\otimes\gamma_{j_2;r_2}&=\chi_{r_1}\otimes(\lim_{\epsilon\to 0}\Gamma_{j_1})\otimes\chi_{r_2}\otimes(\lim_{\epsilon\to 0}\Gamma_{j_2})\\&=\chi_{r_1+r_2}\otimes(\lim_{\epsilon\to 0}\Gamma_{j_1}\otimes\Gamma_{j_2})\\&=\chi_{r_1+r_2}\otimes\left(\lim_{\epsilon\to 0}\sum_j\ \Gamma_j\right),\quad |j_1-j_2|\le j\le j_1+j_2,\\&=\chi_{r_1+r_2}\otimes\left(\sum_j\ \gamma_j\right),\\&=\sum_j\ \gamma_{j;r_1+r_2}.\end{aligned}\tag{51}$$

□

Note that, because the limiting process by which we transform the su(2) irrep into an $e(2)$ representation is smooth, i.e., because it is possible to define a sequence of su(2) representations parametrized by ϵ such that the limit when $\epsilon\to 0$ of this sequence corresponds to a raising string representation, it is possible to interchange the process of taking the limit with the process of taking the tensor product.

One may further remark that this limiting process cannot be used to analyze parallelogram representations, since those correspond to tensor products of a raising and a lowering string representations, i.e., a tensor product of ''different'' contractions.

VII. ACYCLIC REPRESENTATIONS AS SUMS OF STRINGS

The decomposition of the tensor product of two raising or two lowering representations is a special case of a more general theorem regarding ''acyclic'' representations, i.e., representations containing no ''cycles'' of the form

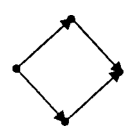

An algebraic characterization of such acyclic representations is that the operator p_+p_- is 0 everywhere. The main result of this section is that acyclic representations are always decomposable into sums of string representations.

Definition: A finite-dimensional representation V of $E(2)$ is said to be *acyclic* if $p_+p_-=0$ on V.

Clearly a string is acyclic, by Proposition 1. For that matter, a direct sum of strings must be acyclic. The goal of this section is to prove the converse: that any acyclic representation must be a direct sum of strings. We begin by establishing some machinery.

We need first a concept which is not restricted to acyclic representations, that of a ''chain.'' We define a ''chain'' to be a finite sequence of strictly increasing weight vectors $v_r, v_{r+1},\dots,v_s$ in V, where each v_j has weight j, and such that for each $r\le j<s$, either $p_+(v_j)=v_{j+1}$ or $p_-(v_{j+1})=v_j$.

Thus, for instance, in the 2×2 parallelogram representation of Eq. (17), which is not acyclic, there are infinitely many chains, each containing $v_1\otimes w_0$ and $v_0\otimes w_{-1}$ but each containing as middle element an otherwise arbitrary nonzero linear combination of $v_0\otimes w_0$ and $v_1\otimes w_{-1}$.

Specializing now to acyclic representations, we see that the condition $p_+p_-=0$ implies that it is not possible to have both $p_+(v_j)\neq 0$ *and* $p_-(v_{j+1})\neq 0$. A chain is ''maximal'' if it cannot be extended by including additional vectors (necessarily at the top or the bottom). The space spanned by the vectors in a maximal chain is a subrepresentation of V, necessarily a string.

Begin by observing that any weight vector can always be embedded in a maximal chain, i.e., it is always possible to find a maximal chain which contains a specified weight vector. Suppose that we have already defined N maximal chains,

$$
\begin{aligned}
&v^1_{r_1}, v^1_{r_1+1}, \ldots, v^1_{s_1};\\
&v^2_{r_2}, v^2_{r_2+1}, \ldots, v^2_{s_2};\\
&\ldots\\
&\ldots\\
&v^N_{r_N}, v^N_{r_N+1}, \ldots, v^N_{s_N}.
\end{aligned}
\tag{52}
$$

Note that we have displayed the chains horizontally rather than vertically for convenience. The chains need not all start nor end at the same weight nor all have the same length. Suppose that they are ''fully independent,'' in the sense that the vectors $\{v^m_i\}$ form a linearly independent set. Note that if v is any vector in the ''span'' of the N chains, i.e., the space spanned by the vectors $\{v^m_i\}$, then $p_+(v)$ and $p_-(v)$ are also in the span of these chains, again because of the assumption that $p_+p_-=0$.

As an inductive step, we have to show how to find $N+1$ fully independent maximal chains in V. It will be fruitful to illustrate the various steps of the induction with the following example.

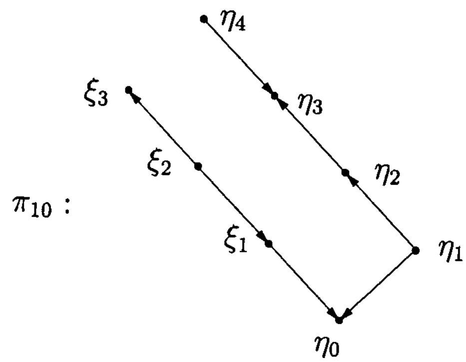

(53)

This representation is found by gluing the string representations containing $\{\eta_4,\eta_3,\eta_2,\eta_1,\eta_0\}$ and $\{\xi_3,\xi_2,\xi_1,\eta_0\}$ at the common node η_0. In the example, we assume that $N=1$ and that the first maximal chain in π_{10} contains $\{v^1_4=\eta_4, v^1_3=\eta_3, v^1_2=\eta_2, v^1_1=\eta_1, v^1_0=\eta_0\}$.

Let k be the highest weight for which the kth weight space is not spanned by the vectors $\{v^m_k : m=1,...,N\}$ that lie in the kth weight space of the first N chains. (In π_{10}, we have $k=3$.)

If there is a vector v in the kth weight space which satisfies $p_+(v)=0$ but which is not in the span of the $\{v^m_k\}$, we will choose it as the top vector in a new chain. Because the $(k+1)$th weight space is by construction contained in the span of the N chains, the action of p_- on vectors in the $(k+1)$th weight space must also take them into the span of these N chains, so that v cannot be in the image of p_-.

The other possibility is that every v in the kth weight space which is not in the span of the $\{v^m_k\}$ satisfies $p_+(v)\neq 0$. Choose such a v. But the intersection of the $(k+1)$th weight space with the image of p_+ is spanned by the vectors $\{p_+(v^m_k)\}$. It is therefore possible to find a (nonzero) vector of the form $v'=v-\Sigma c_m v^m_k$ satisfying $p_+(v')=0$; since v' is not in the span of the $\{v^m_k\}$, this is a contradiction.

In the example, this first step could yield $v^2_3=\xi_3$ as the top state of the second chain in π_{10}.

Having found the top vector for a new chain, $v^{N+1}_{s_{N+1}}$, with $s_{N+1}=k$, we can attempt to construct the rest of the chain inductively by extending it from below.

Suppose we have constructed a chain $v^{N+1}_{s_{N+1}}, \ldots, v^{N+1}_{l+1}, v^{N+1}_l$ so that it and the original N maximal chains form a fully independent set, in the sense defined above.

If $p_-(v^{N+1}_l)=0$, and if v^{N+1}_l is not in the image of p_+, then we have constructed a maximal chain, with $r_{N+1}=l$, completing the inductive step. In all other cases, it necessary to perform an induction to extend the chain.

In general, the situation is that the induction leaves us with the possibility of constructing $N+1$ fully independent chains, none of which can be extended at the ''top,'' meaning that for each $m=1,...,N+1$, $p_+(v^m_{s_m})=0$ and $v^m_{s_m}$ is not in the image of p_-. At least one of the chains is nonmaximal; we will assume that the $(N+1)$th chain is nonmaximal, with lowest weight l, and that l is the maximum of the lowest weights of the nonmaximal chains. (Note that, although the $(N+1)$th chain has so far only been extended as low as weight l, it is, by assumption, not maximal and can therefore certainly be extended below l.) Let L be the number of nonmaximal chains with lowest weight l.

We will produce $N+1$ chains, none of which can be extended at the top, so that the number of nonmaximal chains with lowest weight l is strictly less than L, and so that none of the chains are nonmaximal with lowest weight greater than l. Induction on L will then allow us to construct $N+1$ chains so that any nonmaximal chains among them have lowest weight below l, and, continuing in this way, we can produce $N+1$ maximal chains.

If $p_-(v^{N+1}_l)=0$ and $v^{N+1}_l=p_+(v'')$, for some weight vector v'' of weight $l-1$, then choose such a v'' and let $v^{N+1}_{l-1}=v''$. Note that the resulting chain and the chains labeled 1 to N still form a fully independent set, because v^{N+1}_{l-1} cannot possibly be in the span of the chains labeled 1 to N, since $p_+(v^{N+1}_{l-1})=v^{N+1}_l$ is not in those N chains. Extending the $(N+1)$th chain by adding v^{N+1}_{l-1} gives a chain which has lowest weight $l-1$, which reduces the number of nonmaximal chains with lowest weight l and completes the inductive step in this case.

If $p_-(v^{N+1}_l)$ is nonzero but linearly independent of $\{v^1_{l-1}, v^2_{l-1}, \ldots, v^N_{l-1}\}$, then let $v^{N+1}_{l-1}=p_-(v^{N+1}_l)$. The resulting $(N+1)$th chain and the original N maximal chains still form a fully independent set, and the $(N+1)$th chain has lowest weight $l-1$, which completes the inductive step in this case.

In the example of π_{10}, these two situations occur. Starting with the top vector ξ_3, we see that $p_-\xi_3=0$ but that $p_+\xi_2=\xi_3$, so that the chain containing ξ_3 can be extended to include ξ_2. Since $p_-\xi_2=\xi_1$, which is linearly independent of η_1, the vector of weight 1 which is in the first maximal chain, we can again extend the chain containing $\{\xi_3,\xi_2\}$ to include ξ_1. However, $p_-\xi_1=\eta_0$, which is already in the first maximal chain. We therefore have $l=1$. The second chain contains $\{\xi_3,\xi_2,\xi_1\}$, and it is not maximal.

It could be, (as in the case with ξ_1 in π_1), that $p_-(v^{N+1}_l)$ is in the span of $\{v^1_{l-1}, v^2_{l-1}, \ldots, v^N_{l-1}\}$. Suppose

$$p_-(v^{N+1}_l)=\sum_{m=1}^{N} a_m v^m_{l-1}=\sum_{m=1}^{N} a_m p_-(v^m_l). \tag{54}$$

(The last equality holds because a linear combination of vectors from the N chains that is in the image of p_- must be the image under p_- of a linear combination of vectors from the N chains, because of our assumption that $p_+p_-=0$.) Then let $a_{N+1}=-1$ and consider the vector

$$u_l=-v^{N+1}_l+\sum_{m=1}^{N} a_m v^m_l=\sum_{m=1}^{N+1} a_m v^m_l. \tag{55}$$

It satisfies $p_-(u_l)=0$. Note that, although v^{N+1}_l is not in the image of p_+, this is not necessarily true of u_l. Reordering if necessary, we can assume that $u_l=\Sigma_{m=1}^{N_0} a_m v_l^m$, with $1\leqslant N_0\leqslant N+1$ and $a_m\neq 0$, $p_-(v_l^m)\neq 0$, for $1\leqslant m\leqslant N_0$ (i.e., we reorder so that the first m coefficients in u_l are nonzero.)

In the example of π_{10}, Eq. (55) produces the vector $u_1=-\xi_1+\eta_1$, which is not in the first (maximal) nor in the second (nonmaximal) chain. There is no reordering necessary as u_1 is a linear combination of the two vectors in the first and second chains, i.e., $a_1,a_2\neq 0$, so that $N_0=2=N$.

The situation is as follows. We have $N+1$ ''bottom'' vectors; u_l and the bottom vectors of the original N maximal chains. We have $N+1$ top vectors and the subspace spanned by the vectors in the $N+1$ chains, but no chains containing u_l. It is then a matter of reorganizing the states in the subspace so as to replace one of the existing $N+1$ chains with one that contains u_l. First we construct a chain containing u_l. There is a unique integer $\mu_1\geqslant l$ so that $(p_+)^{\mu_1-l}(u_l)\neq 0$ and $(p_+)^{\mu_1-l+1}(u_l)=0$; let $u_i=(p_+)^{i-l}(u_l)$, for $l\leqslant i\leqslant\mu_1$.

In the example of π_{10}, this integer is $\mu_1=3$ since it is possible to act $\mu_1-l=3-1=2$ times on $u_1=-\xi_1+\eta_1$ before getting $(p_+)^3(-\xi_1+\eta_1)=0$. Thus we have $u_1=-\xi_1+\eta_1, u_2=\eta_2, u_3=\eta_3$.

Reordering the chains if necessary, we can assume that $u_{\mu_1}=\Sigma_{m=1}^{N_1} a_m v_{\mu_1}^m$, with $a_m\neq 0$ whenever $1\leqslant m\leqslant N_1$, for some $N_1\leqslant N_0$. (In π_1, we have $N_1=1$ since u_3 can be expressed as a linear combination of a single vector.)

If u_{μ_1} is not in the image of p_-, then the chain $u_l,\ldots,u_{\mu_1}$ cannot be extended further at the top; it is a ''raising chain.'' If u_{μ_1} is in the image of p_-, then it is possible to add a ''lowering chain'' above it. Indeed, in this case there is some $\nu_1\geqslant 1$ so that

$$(p_-)^{i-\mu_1}\left(\sum_{m=1}^{N_1} a_m v_i^m\right)=u_{\mu_1}, \tag{56}$$

for all i with $\mu_1\leqslant i\leqslant\nu_1$, and so that $\Sigma_{m=1}^{N_1} a_m v_{\nu_1}^m$ is not in the image of p_-. For all i with $\mu_1\leqslant i\leqslant\nu_1$, let $u_i=\Sigma_{m=1}^{N_1} a_m v_i^m$. These vectors form a lowering chain.

In the example of π_{10}, we have $\nu_1=4$ since $p_-\eta_4=\eta_3=u_3$. The chain now contains $u_1=-\xi_1+\eta_1$, $u_2=\eta_2$, $u_3=\eta_3$, $u_4=\eta_4$.

We continue in the same way, finding positive integers $N_0\geqslant N_1\geqslant N_2\geqslant\cdots\geqslant N_t$ and integers $l=\mu_0=\nu_0\leqslant\mu_1<\nu_1<\mu_2<\nu_2<\cdots<\mu_t\leqslant\nu_t$; making a suitable rearrangement of the $N+1$ chains; and for each $1\leqslant j\leqslant t$, and for each i with $\nu_{j-1}\leqslant i\leqslant\mu_j$, letting

$$u_i=(p_+)^{i-\nu_{j-1}}(u_{\nu_{j-1}}), \tag{57}$$

and for each i with $\mu_j\leqslant i\leqslant\nu_j$, letting

$$u_i=\sum_{m=1}^{N_j} a_m v_i^m, \tag{58}$$

so that for each i with $\mu_j<i\leqslant\nu_j$,

$$p_-(u_i)=u_{i-1}. \tag{59}$$

In this way we have constructed a lowering chain between each μ_j and ν_j. Because

$$p_+(u_i)=u_{i+1}, \tag{60}$$

for each i with $\nu_{j-1}\leqslant i<\mu_j$, $j=1,\ldots,t$, there is a raising chain between each ν_{j-1} and μ_j.

In the example of π_{10}, we have $t=1$, as there is only one raising and one lowering chain to be glued to the vector $u_1=-\xi_1+\eta_1$. Our process therefore stops at $\nu_1=4$.

We can continue the construction until we reach a point where $p_+(u_{\nu_t})=0$, so the chain cannot continue up to higher weights. Note that for

$$u_{\nu_t}=\sum_{m=1}^{N_t} a_m v_{\nu_t}^m \tag{61}$$

to be in the kernel of p_+, it must be true that $p_+(v_{\nu_t}^m)=0$, for each $1\leqslant m\leqslant N_t$. And if u_{ν_t} is not in the image of p_-, then at least one of the vectors $v_{\nu_t}^m$, for $1\leqslant m\leqslant N_t$, must not be in the image of p_-. Renumbering yet again, if necessary, we can assume that this is true for $m=1$, which means that ν_t is the highest weight of the chain $\{v_i^1\}$.

In the example of π_{10}, we have $u_4=\eta_4$ as the top vector; η_4 is in the kernel of p_+ and not in the image of p_-. Thus, with the renumbering, the chain $\{u_4,u_3,u_2,u_1\}$ becomes the first chain.

The vectors $u_l,u_{l+1},\ldots,u_{\nu_t}$ form a chain. They all lie in the span of the linearly independent vectors making up the $N+1$ chains $\{v_i^m\}$, for $m=1,...,N+1$. Since $p_-(u_l)=0$, either u_l is not in the image of p_+, in which case the chain $\{u_l,u_{l+1},\ldots,u_{\nu_t}\}$ is maximal, or $u_l=p_+(u_{l-1})$, for some weight vector u_{l-1} of weight $l-1$. It can be added to produce a longer chain $\{u_{l-1},u_l,u_{l+1},\ldots,u_{\nu_t}\}$. Observe that u_{l-1} cannot be in the space spanned by the chains $\{v_i^1\},\ldots,\{v_i^{N+1}\}$. We have constructed a chain $\{u_i\}$; in one case it is maximal, and in the other it has lowest weight $l-1$.

In the example of π_{10}, the chain is maximal since $u_1=-\xi_1+\eta_1$ is not in the image of p_+.

Since every one of the vectors u_i, for $i\geqslant l$, contains a nonzero component in what is now labeled as the first chain $\{v_i^1\}$, we can replace the chain $\{v_i^1\}$ with the chain $\{u_i\}$ and the resulting $N+1$ chains will still be fully independent.

At this point there are different possibilities. One is that the chain labeled $\{v_i^1\}$ that was removed was the original nonmaximal chain $\{v_i^{N+1}\}$, and it has just been replaced by the chain $\{u_i\}$. This does not apply to π_{10}; the original second chain is not identical to the newly constructed maximal chain.

Otherwise, the original nonmaximal chain $\{v_i^{N+1}\}$ with lowest weight l is still present. In this case we can change its label back to $v_l^{N+1},\ldots,v_{s_{N+1}}^{N+1}$. But now $p_-(v_l^{N+1})$ is not in the span of the vectors in the other N chains, since $p_-(v_l^1)$ is no longer present in the other N chains. This means that the $(N+1)$th chain can be extended to include $p_-(v_l^{N+1})$ as its ''bottom'' vector.

In the example of π_{10}, this is what happens. The original second chain contained $\{\xi_3,\xi_2,\xi_1\}$. Since $p_-\xi_1=\eta_0$, which is no longer in the span of the newly constructed first maximal chain, we can extend this second chain to include η_0.

In either case, the number of nonmaximal chains with lowest weight l has been reduced.

In the example of π_{10}, we now restart the induction with $\{\eta_4,\eta_3,\eta_2,-\xi_1+\eta_1\}$ as the first maximal chain and $\{\xi_3,\xi_2,\xi_1,\eta_0\}$ as the second chain, we find that now $k=3$, $N=1$ but $l=0$. However, $p_-\eta_0=0$ and η_0 is not in the image of p_+. Thus, the second chain is maximal as it is. This concludes the inductions: we have found the decomposition of π_{10} as the sum of two strings.

Theorem 1: A finite-dimensional representation V of $E(2)$ is acyclic if and only if V is a direct sum of indecomposable representations, in each of which the weight spaces are all of dimension 1.

Proof: ($\Leftarrow$) Trivial.

($\Rightarrow$) The above argument shows that, given N fully independent maximal chains that do not span all of V, it is possible to construct $N+1$ fully independent maximal chains.

In attempting this construction, we may reach a situation where there are $N+1$ fully independent chains, but not all of them are maximal. We let l be the maximum of the lowest weights of the nonmaximal strings. The inductive step described above reduces the number of nonmaximal strings with lowest weight l. Repeated application of this procedure will eventually reduce l, the maximum of the lowest weights of the nonmaximal strings.

Continuing with an induction on l, we can eventually eliminate all the nonmaximal chains, producing $N+1$ fully independent maximal chains, as required.

Then, since the number of fully independent chains is certainly bounded by the dimension of the whole space V, we will eventually be able to construct enough chains that they span the whole space. □

VIII. DISCUSSION AND CONCLUSION

In this paper, we have described numerous finite dimensional indecomposable representations of $E(2)$ by means of a method which encapsulates in graphical form all the necessary information to explicitly construct, up to a character, a representation.

The basic type representation is the string, in which all the weight subspaces are of dimension one. Using lemma 1 and proposition 1, we can associate to a string representation a graph, from which it is easy to determine if the representation is decomposable or not. In an indecomposable string representation, the ‘‘strength’’ of the $e(2)$ matrix element connecting two states is irrelevant; all indecomposable strings representations are equivalent to representations for which this matrix element is 1.

We have been successful in showing the indecomposability of another very important class of representations, the parallelograms and all their subrepresentations and quotients. Parallelograms and their subrepresentations may contain nontrivial weight multiplicities, an unusual feature for representations of $E(2)$. We have also shown how acyclic representations can be decomposed into sums of string representations.

The problem of decomposing a general graph containing nontrivial weight multiplicities arising either per se or as from the tensor product of two general string representations is difficult.

Consider for instance the acyclic graph

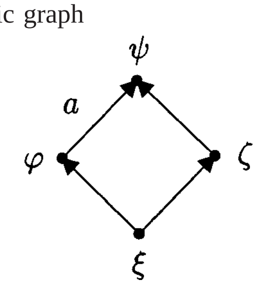

(62)

with basis states $\{\xi,\varphi,\zeta,\psi\}$, in which $p_+\varphi=a\psi$, with all other nonzero matrix elements being 1. When $a=-1$, the representation decomposes into a sum containing two (inequivalent) two-dimensional subrepresentations.

When $a\neq-1$, however, this can be decomposed into a sum of a three-dimensional and a one-dimensional string. The special case where $a=1$ corresponds to a tensor product of the two-dimensional raising string with itself.

The decomposability of some graphs can be understood in terms of representations of S_n, the permutation group of n objects. Unfortunately, arguments based on the permutation group are of limited use because (i) the S_n-invariant subspaces may themselves decompose further (for instance, in the tensor product of a three-dimensional raising string with itself, the six-dimensional subspace that carries the fully symmetric representation of S_2 can be divided into a five-dimensional and a one-dimensional indecomposable raising string), (ii) experience has shown that the problem of deciding if a given graph (string or otherwise) can be obtained by tensoring n copies of a given string is nontrivial.

There is, however, one case which we would like to mention. Consider the tensor product of an indecomposable string V, with a basis of weight vectors $v_l,\ldots,v_m$ with $l\leqslant m$, with another indecomposable string V' With weight vectors $v'_{-m},\ldots,v'_{-l}$ such that $p_+v'_{-k-1}=v'_{-k}$ if and only if $p_+v_k=v_{k+1}$ and $p_-v'_{-k}=v_{-k-1}$ if and only if $p_-v_{k+1}=v_k$. The tensor product $V\otimes V'$ is decomposable into two parts, one of which is the one-dimensional indecomposable representation with basis vector $v=\Sigma_{k=l}{}^m(-1)^k v_k\otimes v'_{-k}$, because V' occurs when we tensor together $(l-m-1)$ copies of V.

The simplest example of this is found in (63).

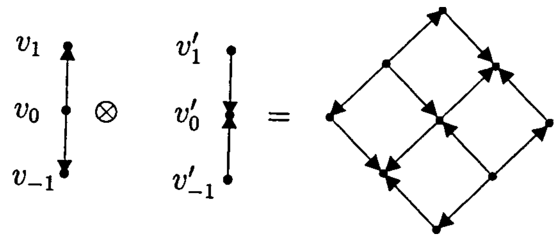

(63)

In this example, $V\otimes V'$ decomposes into an eight-dimensional representation isomorphic to (30) and a one-dimensional subspace $v=v_0\otimes v'_0-v_1\otimes v'_{-1}+v_2\otimes v'_{-2}$. Note that, obviously, the weights of V' could be shifted up or down and the tensor product would still be decomposable. What is important is the relative positions of the arrows, not the actual weights.

This family of decomposable tensor products can be related to the symmetric group as follows. It can be shown that, if the dimension of V is d, then the $(d-1)$-fold tensor product of V with itself contains, up to a character, V' in the S_{d-1}-invariant subspace labeled by a Young tableau containing a single column of $d-1$ boxes.

In the example of (63), V is of dimension $d=3$, and a basis for the three-dimensional representation of $V\otimes V$ which carries the S_2 representation labeled by 1 column of 2 boxes is given by

$$v'_{-2}=\chi_{-3}\otimes\det\begin{vmatrix} w_1 & x_1 \\ w_0 & x_0\end{vmatrix},\quad v'_{-1}=\chi_{-3}\otimes\det\begin{vmatrix} w_2 & x_2 \\ w_0 & x_0\end{vmatrix},\quad v'_0=\chi_{-3}\otimes\det\begin{vmatrix} w_2 & x_2 \\ w_1 & x_1\end{vmatrix}, \tag{64}$$

where $w_i, x_j, i,j=0,1,2$ are basis states for the first and second copy of V in the tensor product $V\otimes V$, respectively. The representation V' can be reconstructed if we observe that the nonzero matrix elements of $p_\pm$ are given by

$$\begin{aligned} p_+v'_{-2}&=\chi_{-3}\otimes\det\begin{vmatrix} p_+w_1 & p_+x_1 \\ w_0 & x_0\end{vmatrix}+\det\begin{vmatrix} w_1 & x_1 \\ p_+w_0 & p_+x_0\end{vmatrix}=\chi_{-3}\otimes\det\begin{vmatrix} w_2 & x_2 \\ w_0 & x_0\end{vmatrix}=v'_{-1},\\ p_-v'_0&=\chi_{-3}\otimes\det\begin{vmatrix} p_-w_2 & p_-x_2 \\ w_1 & x_1\end{vmatrix}+\chi_{-3}\otimes\det\begin{vmatrix} w_2 & x_2 \\ p_-w_1 & p_-x_1\end{vmatrix}=\chi_{-3}\otimes\det\begin{vmatrix} w_2 & x_2 \\ w_0 & x_0\end{vmatrix}=v'_0. \end{aligned} \tag{65}$$

From this, it can be seen how the decomposition of $V\otimes V'$ is related to the action of symmetric group on $(V)^d$, and why the scalar v is alternating in nature.

Finally, even if all the examples of decomposable tensor products discussed in this paper can ultimately be related to the symmetric group, we believe that there very likely exist decomposable graphs with nontrivial weight multiplicities which are unrelated to S_n. We have, unfortunately, been unable to isolate a provable conjecture on this matter. The low-dimensional examples of this section are sufficiently complex to illustrate the difficulty of the general problem.

In a subsequent publication, we will investigate the role of gluings of the type found in (30) in the construction of finite dimensional representations.

There is no doubt that results similar to Lemma 1 and Proposition 1 can be extended to other groups,[5] in particular within the context of graded contractions.[4] It is also reasonable to expect that the method can be generalized to the construction of finite dimensional representations of other semidirect product groups. In that regard, one should observe that the operator p_+p_- is in fact, the $e(2)$ Casimir operator, so that one way of generalizing the concept of string representations to other groups is to require that the appropriate Casimir be 0. It remains to see how other concepts, such as parallelograms, can be generalized to other examples.

ACKNOWLEDGMENTS

We are grateful to J. Patera and M. de Montigny for numerous discussion throughout this work. H.d.G. would like to thank Fonds FCAR of the Québec Government for additional financial support. This work was supported in part by N.S.E.R.C. of Canada.

[1] G. W. Mackey, *Induced Representations of Groups and Quantum Mechanics* (Benjamin, New York, 1968).
[2] I. E. Segal, Duke Math. J. **18**, 221–265 (1951); M. Levy-Nahas, J. Math. Phys. **8**, 1211–1222 (1967).
[3] R. Gilmore, *Lie Groups, Lie Algebras, and Some of their Applications* (Wiley, New York, 1974), Chap. 10.
[4] R. V. Moody and J. Patera, J. Phys. A **24**, 2227–2257 (1991).
[5] Hubert de Guise and Joe Repka, ‘‘Graphical methods for graded contractions,’’ (in preparation).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Level-one representations and vertex operators of quantum affine superalgebra $U_q[gl(\hat{N}|N)]$

Yao-Zhong Zhang
Department of Mathematics, University of Queensland, Brisbane, Queensland 4072, Australia

(Received 8 March 1999; accepted for publication 9 August 1999)

Level-one representations of the quantum affine superalgebra $U_q[gl(\hat{N}|N)]$ associated with the appropriate nonstandard system of simple roots and q-vertex operators (intertwining operators) associated with the level-one modules are constructed explicitly in terms of free bosonic fields. © *1999 American Institute of Physics.* [S0022-2488(99)02111-8]

I. INTRODUCTION

The algebraic analysis approach[1,2] based on quantum affine algebra symmetries enables one not only to solve massive or off-critical integrable models directly in the thermodynamic limit but also to compute their correlation functions[3] and form factors[4] in the form of integrals by applying the techniques similar to those used so successfully in the critical cases (see, for example, Ref. 5). The key components behind this method are infinite dimensional highest weight representations of the quantum affine algebras and the corresponding q-vertex operators[6] which are intertwiners of these representations. As in the critical cases, this procedure requires the explicit construction of the highest weight representations and vertex operators in terms of free bosonic fields.

By now, the level-one representations and vertex operators have been constructed in terms of free bosons for most quantum affine bosonic algebras (see, e.g., Refs. 7–12). In contrast, much less has been known for the case of quantum affine superalgebras. For the type I quantum affine superalgebra $U_q[gl(\hat{M}|N)]$, $M\neq N$, the level-one representations and vertex operators have been investigated in Ref. 13 (see Ref. 14 for a level-k free boson realization of $U_q[sl(\hat{2}|1)]$). In particular, the level-one irreducible highest weight representations of $U_q[gl(\hat{2}|1)]$ were studied in some details and the corresponding characters were derived.[13] These representations have been re-examined and used to compute the correlation functions of the q-deformed supersymmetric $t-J$ model in Ref. 15.

So far in the literature, the very interesting case of $M=N$ has been largely ignored. The only exception is[16] where the special case of $M=N=2$ was treated and the type I vertex operators involving infinite dimensional evaluation (or level-zero) representations were also constructed for this special case. By contrast, we shall consider the general $M=N$ case and investigate both type I and type II vertex operators with respect to finite dimensional evaluation modules. The $M=N$ case is interesting since it seems to us that $U_q[gl(\hat{N}|N)]$ is the only untwisted superalgebra which has a nonstandard system where all simple roots are odd or fermionic. It also seems to be the only superalgebra where a vertex type quasi-Hopf twistor can be constructed[17] and thus the corresponding elliptic quantum supergroup $\mathcal{A}_{q,p}[gl(\hat{N}|N)]$ can be introduced.

In this paper, we construct a level-one representation of $U_q[gl(\hat{N}|N)]$ by bosonizing the Drinfeld generators. We also construct the vertex operators associated with the level-one representations in terms of the free bosonic fields.

It should be pointed out that the $M=N$ case, treated in this paper, is more complicated than the $M\neq N$ case considered in Ref. 13. For the $M=N$ case, the Cartan matrix is degenerate, i.e., not invertible. As is well known, the invertibility of Cartan matrix is essential in the construction of vertex operators. We overcome this problem by enlarging the Cartan subalgebra in an appropriate way. Moreover, we shall work in a nonstandard system of simple roots, in contrast to Ref. 13 in which a standard system was used. Then our method is a generalization and modification of that used in Ref. 13. We also remark that the ideas in the present paper is applicable to the study of higher level vertex operator representations of quantum affine superalgebras.

The layout of this paper is the following. In Sec. II, we describe the Drinfeld realization[18] of $U_q[gl(\hat{N}|N)]$ in the nonstandard system of simple roots and determine the ''main terms'' [19] in the coproduct formulas of the Drinfeld generators. In Sec. III, we derive the $2N$-dimensional evaluation (or level-zero) representations of $U_q[gl(\hat{N}|N)]$. In Sec. IV, we investigate the bosonization of $U_q[gl(\hat{N}|N)]$ and construct an explicit level-one representation in terms of free bosonic fields. Section V is devoted to the study of the bosonization of the level-one vertex operators.

II. QUANTUM AFFINE SUPERALGEBRA $U_q[gl(\hat{N}|N)]$

As is well-known, a given Kac–Moody superalgebra[20] allows many inequivalent systems of simple roots. A system of simple roots is called distinguished if it has minimal odd roots. Let $\{\alpha_i, i=0,1,...,2N-1\}$ denote a chosen set of simple roots of the affine superalgebra $gl(\hat{N}|N)$. Let (,) be a fixed invariant bilinear form on the root space. Let $\mathcal{H}$ be the Cartan subalgebra and throughout we identify the dual $\mathcal{H}^*$ with $\mathcal{H}$ via (,). As is shown in Ref. 17, $gl(\hat{N}|N)$ has a simple root system in which all simple roots are odd (or fermionic). This system can be constructed from the distinguished simple root system by using the ''extended'' Weyl operation[21] repeatedly. We have the following simple roots, all of which are odd (or fermionic)

$$\alpha_0=\delta-\epsilon_1+\epsilon_{2N},$$
$$\alpha_l=\epsilon_l-\epsilon_{l+1},\quad l=1,2,...,2N-1 \tag{II.1}$$

with δ, $\{\epsilon_k\}_{k=1}^{2N}$ satisfying

$$(\delta,\delta)=(\delta,\epsilon_k)=0,\quad (\epsilon_k,\epsilon_{k'})=(-1)^{k+1}\delta_{kk'}. \tag{II.2}$$

Such a simple root system is usually called non-standard. The generalized symmetric Cartan matrix $(a_{ii'})$ takes the form

$$a_{01}\equiv(\alpha_0,\alpha_1)=-1,\quad a_{0,2N-1}\equiv(\alpha_0,\alpha_{2N-1})=1,$$
$$a_{ll'}\equiv(\alpha_l,\alpha_{l'})=(-1)^{l+1}(\delta_{l,l'-1}-\delta_{l,l'+1}),\quad l,l'=1,2,...,2N-1. \tag{II.3}$$

This Cartan matrix is degenerate. To obtain a nondegenerate Cartan matrix, we extend Ref. 22 $\mathcal{H}$ by adding to it the element

$$\alpha_{2N}=\sum_{k=1}^{2N}\epsilon_k. \tag{II.4}$$

In the following, we denote by $\tilde{\mathcal{H}}$ the extended Cartan subalgebra and by $\tilde{\mathcal{H}}^*$ the dual of $\tilde{\mathcal{H}}$. The enlarged Cartan matrix has the following extra matrix elements:

$$a_{2N,2N}\equiv(\alpha_{2N},\alpha_{2N})=0,\quad a_{i,2N}\equiv(\alpha_i,\alpha_{2N})=2\cdot(-1)^{i+1}. \tag{II.5}$$

Let $\{h_0,h_1,\ldots,h_{2N},d\}$ be a basis of $\tilde{\mathcal{H}}$, where h_{2N} is the element in $\tilde{\mathcal{H}}$ corresponding to α_{2N} and d is the usual derivation operator. We shall write $h_i=\alpha_i$ $(i=0,1,...,2N)$ with α_i given by (II.1), (II.4). Let $\{\Lambda_0,\Lambda_1,\ldots,\Lambda_{2N},c\}$ be the dual basis with Λ_j being fundamental weights and c the canonical central element. We have[17]

$$\Lambda_{2N}=\frac{1}{2N}\sum_{k=1}^{2N}(-1)^{k+1}\epsilon_k,$$

$$\Lambda_i=d+\sum_{k=1}^{i}(-k)^{k+1}\epsilon_k-\frac{i}{2N}\sum_{k=1}^{2N}(-1)^{k+1}\epsilon_k, \tag{II.6}$$

where $i=0,1,...,2N-1$.

The quantum affine superalgebra $U_q[gl(\hat{N}|N)]$ is a quantum (or q-) deformation of the universal enveloping algebra of $gl(\hat{N}|N)$ and is generated by the Chevalley generators $\{e_i, f_i\ q^{h_j}, d|i=0,1,...,2N-1, j=0,1,...,2N\}$. The $\mathbf{Z}_2$-grading of the Chevalley generators is $[e_i]=[f_i]=1$, $i=0,1,...,2N-1$ and zero otherwise. The defining relations are

$$hh'=h'h, \quad \forall h\in\tilde{H},$$

$$q^{h_j}e_iq^{-h_j}=q^{a_{ij}}e_i, \quad [d,e_i]=\delta_{i0}e_i,$$

$$q^{h_j}f_iq^{-h_j}=q^{-a_{ij}}f_i, \quad [d,f_i]=-\delta_{i0}f_i,$$

$$[e_i,f_{i'}]=\delta_{ii'}\frac{q^{h_i}-q^{-h_i}}{q-q^{-1}},$$

$$[e_i,e_{i'}]=[f_i,f_{i'}]=0, \text{ for } a_{ii'}=0,$$

$$[[e_0,e_1]_{q^{-1}},[e_0,e_{2N-1}]_q]=0,$$

$$[[e_l,e_{l-1}]_{q^{(-1)^l}},[e_l,e_{l+1}]_{q^{(-1)^{l+1}}}]=0,$$

$$[[e_{2N-1},e_{2N-2}]_{q^{-1}},[e_{2N-1},e_0]_q]=0,$$

$$[[f_0,f_1]_{q^{-1}},[f_0,f_{2N-1}]_q]=0,$$

$$[[f_l,f_{l-1}]_{q^{(-1)^l}},[f_l,f_{l+1}]_{q^{(-1)^{l+1}}}]=0,$$

$$[[f_{2N-1},f_{2N-2}]_{q^{-1}},[f_{2N-1},f_0]_q]=0, \quad l=1,2,...,2N-2. \tag{II.7}$$

Here and throughout, $[a,b]_x\equiv ab-(-1)^{[a][b]}xba$ and $[a,b]\equiv[a,b]_1$. The fourth order q-Serre relations are obtained by using Yamane's Dynkin diagram procedure.[23]

$U_q[gl(\hat{N}|N)]$ is a $\mathbf{Z}_2$-graded quasitriangular Hopf algebra endowed with the following coproduct Δ, counit ϵ and antipode S:

$$\Delta(h)=h\otimes 1+1\otimes h,$$

$$\Delta(e_i)=e_i\otimes 1+q^{h_i}\otimes e_i, \quad \Delta(f_i)=f_i\otimes q^{-h_i}+1\otimes f_i,$$

$$\epsilon(e_i)=\epsilon(f_i)=\epsilon(h)=0,$$

$$S(e_i)=-q^{-h_i}e_i, \quad S(f_i)=-f_iq^{h_i}, \quad S(h)=-h, \tag{II.8}$$

where $i=0,1,...,2N-1$ and $h\in\tilde{\mathcal{H}}$. Notice that the antipode S is a $\mathbf{Z}_2$-graded algebra antihomomorphism. Namely, for any homogeneous elements $a,b\in U_q[gl(\hat{N}|N)]$ $S(ab)=(-1)^{[a][b]}S(b)S(a)$, which extends to inhomogeneous elements through linearity. Moreover,

$$S^2(a)=q^{-2\rho}aq^{2\rho}, \quad \forall a\in U_q[gl(\hat{N}|N)], \tag{II.9}$$

where ρ is an element in $\tilde{\mathcal{H}}$ such that $(\rho,\alpha_i)=(\alpha_i,\alpha_i)/2$ for any simple root α_i, $i=0,1,2,...,2N-1$. Explicitly,

$$\rho=\frac{1}{2}\sum_{k=1}^{2N}(-1)^k\epsilon_k, \tag{II.10}$$

which coincides with $\overline{\rho}$, the half-sum of positive roots of $gl(N|N)$ in the present simple root system. The multiplication rule on the tensor products is $\mathbf{Z}_2$-graded: $(a\otimes b)(a'\otimes b')=(-1)^{[b][a']}(aa'\otimes bb')$ for any homogeneous elements $a,b,a',b'\in U_q[gl(\hat{N}|N)]$. We can also introduce the element in $\tilde{\mathcal{H}}$,

$$\tilde{\rho}=\sum_{i=0}^{2N-1}\Lambda_i+\xi N\Lambda_{2N}, \tag{II.11}$$

which gives the principal gradation

$$[\tilde{\rho},e_i]=e_i,\quad [\tilde{\rho},f_i]=-f_i,\quad i=0,1,...,2N-1. \tag{II.12}$$

In (II.11), ξ is an arbitrary constant.

$U_q[gl(\hat{N}|N)]$ can also be realized in terms of the Drinfeld generators[18] $\{X_m^{\pm,i},\ H_n^j,\ q^{\pm H_0^j},\ c\ d|m\in\mathbf{Z},\ n\in\mathbf{Z}-\{0\},\ i=1,2,...,2N-1,\ j=1,2,...,2N\}$. The $\mathbf{Z}_2$-grading of the Drinfeld generators is given by $[X_m^{\pm,i}]=1$ for all $i=1,...2N-1$, $m\in\mathbf{Z}$ and $[H_n^j]=[H_0^j]=[c]=[d]=0$ for all $j=1,...,2N$, $n\in\mathbf{Z}-\{0\}$. The relations satisfied by the Drinfeld generators reads (see Refs. 23, 24 for the Drinfeld realization of $U_q[gl(\hat{N}|N)]$ in the distinguished system of simple roots)

$$\begin{gathered}
[c,a]=[d,H_0^j]=[H_0^j,H_n^{j'}]=0,\quad \forall a\in U_q[gl(\hat{N}|N)],\\
q^{H_0^j}X_n^{\pm,i}q^{-H_0^j}=q^{\pm a_{ij}}X_n^{\pm,i},\quad [d,X_n^{\pm,i}]=nX_n^{\pm,i},\ [d,H_n^j]=nH_n^j,\\
[H_n^j,H_m^{j'}]=\delta_{n+m,0}\frac{[a_{jj'}n]_q[nc]_q}{n},\quad [H_n^j,X_m^{\pm,i}]=\pm\frac{[a_{ij}n]_q}{n}X_{n+m}^{\pm,i}q^{\mp|n|c/2},\\
[X_n^{+,i},X_m^{-,i'}]=\frac{\delta_{ii'}}{q-q^{-1}}(q^{(c/2)(n-m)}\psi_{n+m}^{+,i}-q^{-(c/2)(n-m)}\psi_{n+m}^{-,i}),\\
[X_n^{\pm,i},X_m^{\pm,i'}]=0\ \text{ for }\ a_{ii'}=0,\quad [X_{n+1}^{\pm,i},X_m^{\pm,i'}]_{q^{\pm a_{ii'}}}-[X_{m+1}^{\pm,i'},X_n^{\pm,i}]_{q^{\pm a_{ii'}}}=0,\\
[[X_m^{\pm,l},X_{m'}^{\pm,l-1}]_{q^{(-1)^l}},[X_n^{\pm,l},X_{n'}^{\pm,l+1}]_{q^{(-1)^{l+1}}}]\\
+[[X_n^{\pm,l},X_m^{\pm,l-1}]_{q^{(-1)^l}},[X_m^{\pm,l},X_{n'}^{\pm,l+1}]_{q^{(-1)^{l+1}}}]=0,\quad l=2,...,2N-2,
\end{gathered} \tag{II.13}$$

where $[x]_q=(q^x-q^{-x})/(q-q^{-1})$ and $\psi_n^{\pm,j}$ are related to $H_{\pm n}^j$ by relations

$$\sum_{n\in\mathbf{Z}}\psi_n^{\pm,j}z^{-n}=q^{\pm H_0^j}\exp\left(\pm(q-q^{-1})\sum_{n>0}H_{\pm n}^jz^{\mp n}\right). \tag{II.14}$$

The following relations can be proved by induction:

$$H_n^j=\frac{1}{q-q^{-1}}\sum_{p_1+2p_2+\cdots+\cup_n=n}\frac{(-1)^{\Sigma p_i-1}(\Sigma p_i-1)!}{p_1!\cdots p_n!}(q^{-H_0^j}\psi_1^{+,j})^{p_1}\cdots(q^{-H_0^j}\psi_n^{+,j})^{p_n},$$

$$H^j_{-n}=\frac{1}{q^{-1}-q}\sum_{p_1+2p_2+\cdots+np_n=n}\frac{(-1)^{\Sigma p_i-1}(\Sigma p_i-1)!}{p_1!\cdots p_n!}(q^{H^j_0}\psi^{-,j}_{-1})^{p_1}\cdots(q^{H^j_0}\psi^{-,j}_{-n})^{p_n}. \tag{II.15}$$

The Chevalley generators are related to the Drinfeld generators by the formulas,

$$h_i=H^i_0,\quad e_i=X^{+,i}_0,\quad f_i=X^{-,i}_0,\quad i=1,2,...,2N-1,$$

$$h_{2N}=H^{2N}_0,\quad h_0=c-\sum_{k=1}^{2N-1}H^k_0,$$

$$e_0=[X^{-,2N-1}_0,[X^{-,2N-2}_0,\ldots,[X^{-,3}_0,[X^{-,2}_0,X^{-,1}_1]_q]_{q^{-1}}\cdots]_q]_{q^{-1}}q^{-\Sigma_{k=1}^{2N-1}H^k_0},$$

$$f_0=(-1)^N q^{\Sigma_{k=1}^{2N-1}H^k_0}[[\cdots[[X^{+,1}_{-1},X^{+,2}_0]_{q^{-1}},X^{+,3}_0]_q,\ldots,X^{+,2N-2}_0]_{q^{-1}},X^{+,2N-1}_0]_q. \tag{II.16}$$

The coproduct of the Drinfeld generators is not known in full. However, for our purpose it suffices to derive the ''main terms''[19] in the coproduct formulas. We have

Proposition 1: For $m\in\mathbf{Z}_{\geqslant 0}$, $n\in\mathbf{Z}_{>0}$ *and* $i=1,2,...,2N-1$,

$$\Delta(X^{+,i}_m)=X^{+,i}_m\otimes q^{mc}+q^{H^i_0+2mc}\otimes X^{+,i}_m+\sum_{k=0}^{m-1}q^{(1/2)(m+3k)c}\psi^{+,i}_{m-k}\otimes q^{(m-k)c}X^{+,i}_k \bmod N_-\otimes N^2_+,$$

$$\Delta(X^{+,i}_{-n})=X^{+,i}_{-n}\otimes q^{-nc}+q^{-H^i_0}\otimes X^{+,i}_{-n}+\sum_{k=1}^{n-1}q^{(1/2)(n-k)c}\psi^{-,i}_{k-n}\otimes q^{(k-n)c}X^{+,i}_{-n} \bmod N_-\otimes N^2_+,$$

$$\Delta(X^{-,i}_n)=X^{-,i}_n\otimes q^{H^i_0}+q^{nc}\otimes X^{-,i}_n+\sum_{k=1}^{n-1}q^{(n-k)c}X^{-,i}_k\otimes q^{(1/2)(k-n)c}\psi^{+,i}_{n-k} \bmod N^2_-\otimes N_+,$$

$$\Delta(X^{-,i}_{-m})=X^{-,i}_{-m}\otimes q^{-H^i_0-2mc}+q^{-mc}\otimes X^{-,i}_{-m}$$
$$+\sum_{k=0}^{m-1}q^{(k-m)c}X^{-,i}_{-k}\otimes q^{-(1/2)(m+3k)c}\psi^{-,i}_{k-m} \bmod N^2_-\otimes N_+,$$

$$\Delta(H^i_n)=H^i_n\otimes q^{(1/2)nc}+q^{(3/2)nc}\otimes H^i_n \bmod N_-\otimes N_+,$$

$$\Delta(H^i_{-n})=H^i_{-n}\otimes q^{-(3/2)nc}+q^{-(1/2)nc}\otimes H^i_{-n} \bmod N_-\otimes N_+, \tag{II.17}$$

where $N_\pm$ *and* $N^2_\pm$ *are the left ideals generated by* $X^{\pm,k}_l$ *and* $X^{\pm,k}_l X^{\pm,k'}_{l'}$, k, $k'=1,...,2N-1$; l, $l'\in\mathbf{Z}$, *respectively.*

Remark: (i) We do not write down the formulas for $\Delta(H^{2N}_{\pm n})$ because they are not needed in this paper. $\Delta(H^{2N}_{\pm n})$ can be determined by requiring that Δ preserves the commutation relations (II.13). (ii) Modulo $N_+\otimes N_-+N_-\otimes N_+$, the elements $\psi^{\pm,i}_{\pm n}$ $(n\geqslant 0)$ are grouplike,

$$\Delta(\psi^{+,i}_n)=\sum_{k=0}^{n}q^{(3/2)kc}\psi^{+,i}_{n-k}\otimes q^{(1/2)(n-k)c}\psi^{+,i}_k,$$

$$\Delta(\psi^{-,i}_{-n})=\sum_{k=0}^{n}q^{-(1/2)kc}\psi^{-,i}_{k-n}\otimes q^{(3/2)(k-n)c}\psi^{-,i}_{-k}. \tag{II.18}$$

Define the Drinfeld currents or generating functions,

$$X^{\pm,i}(z)=\sum_{n\in\mathbf{Z}} X_n^{\pm,i} z^{-n-1},\quad \psi^{\pm,j}(z)=\sum_{n\in\mathbf{Z}} \psi_n^{\pm,j} z^{-n}. \tag{II.19}$$

In terms of these currents, (II.13) read

$$\psi^{\pm,j}(z)\psi^{\pm,j'}(w)=\psi^{\pm,j'}(w)\psi^{\pm,j}(z),$$

$$\psi^{+,j}(z)\psi^{-,j'}(w)=\frac{(z-wq^{c+a_{jj'}})(z-wq^{-c-a_{jj'}})}{(z-wq^{c-a_{jj'}})(z-wq^{-c+a_{jj'}})}\psi^{-j'}(w)\psi^{+,j}(z),$$

$$\psi^{+,j}(z)X^{\pm,i}(w)=q^{\pm a_{ij}}\frac{z-wq^{\mp(c/2)\mp a_{ij}}}{z-wq^{\mp(c/2)\pm a_{ij}}}X^{\pm,i}(w)\psi^{+,j}(z),$$

$$\psi^{-,j}(z)X^{\pm,i}(w)=q^{\pm a_{ij}}\frac{z-wq^{\pm(c/2)\mp a_{ij}}}{z-wq^{\pm(c/2)\pm a_{ij}}}X^{\pm,i}(w)\psi^{-,j}(z),$$

$$[X^{+,i}(z),X^{-,i'}(w)]=\frac{\delta_{ii'}}{(q-q^{-1})zw}\left(\delta\left(\frac{w}{z}q^{c}\right)\psi^{+,i}(wq^{(c/2)})-\delta\left(\frac{w}{z}q^{-c}\right)\psi^{-,i}(wq^{-(c/2)})\right),$$

$$X^{\pm,i}(z)X^{\pm,i'}(w)+X^{\pm,i'}(w)X^{\pm,i}(z)=0,\ \text{ for }\ a_{ii'}=0,$$

$$(z-wq^{\pm a_{ii'}})X^{\pm,i}(z)X^{\pm,i'}(w)+(zq^{\pm a_{ii'}}-w)X^{\pm,i'}(w)X^{\pm,i}(z)=0,$$

$$\{[[X^{\pm,l}(z_1),X^{\pm,l-1}(z)]_{q^{(-1)^l}},[X^{\pm,l}(z_2),X^{\pm,l+1}(w)]_{q^{(-1)^{l+1}}}]\}+\{z_1\leftrightarrow z_2\}=0,\ \ l=2,3,...,2N-2. \tag{II.20}$$

These current commutation relations can be derived from the super version[24,25] of the RS algebra[26] by means of the Gauss decomposition technique of Ding and Frenkel.[27]

III. LEVEL-ZERO REPRESENTATION

We consider the evaluation representation V_z of $U_q[gl(\widehat{N|N})]$, where V is an $2N$-dimensional graded vector space with basis vectors $\{v_1,v_2,\ldots,v_{2N}\}$. The $\mathbf{Z}_2$-grading of the basis vectors is chosen to be $[v_j]=[(-1)^j+1]/2$. Let $e_{j,j'}$ be the $2N\times 2N$ matrices satisfying $(e_{j,j'})_{kk'}=\delta_{jk}\delta_{j'k'}$ or equivalently $e_{i,j}v_k=\delta_{jk}v_i$, (which implies that for any operator A its matrix elements $A_{j,i}$ are defined by $Av_i=A_{j,i}v_j$). In the homogeneous gradation, the Chevalley generators on V_z are represented by

$$e_i=e_{i,i+1},\quad f_i=(-1)^{i+1}e_{i+1,i},\quad i=1,2,...,2N-1,$$

$$h_i=(-1)^{i+1}(e_{i,i}+e_{i+1,i+1}),\quad h_{2N}=\sum_{k=1}^{2N}(-1)^{k+1}e_{k,k},$$

$$e_0=ze_{2N,1},\quad f_0=-z^{-1}e_{1,2N},\quad h_0=-e_{1,1}-e_{2N,2N}. \tag{III.1}$$

Let V^{*S} be the left dual module of V, defined by

$$(a\cdot v^*)(v)=(-1)^{[a][v^*]}v^*(S(a)v),\quad \forall a\in U_q[gl(\widehat{N|N})],\quad v\in V, v^*\in V^*. \tag{III.2}$$

Namely, the representations on V^{*S} are given by

$$\pi_{V^{*S}}(a)=\pi_V(S(a))^{\mathrm{st}},\quad \forall a\in U_q[gl(\widehat{N|N})], \tag{III.3}$$

where st denotes the supertransposition defined by $(A_{i,j})^{\rm st}=(-1)^{[j]([i]+[j])}A_{j,i}$. Note that in general $((A_{i,j})^{\rm st})^{\rm st}=(-1)^{[A]}A_{i,j}\neq A_{i,j}$. Let V_z^{*S} be the $2N$-dimensional evaluation module corresponding to V^{*S}. On V_z^{*S}, the Chevalley generators are represented by

$$e_i=-(-1)^i q^{(-1)^i} e_{i+1,i},\quad f_i=-q^{(-1)^{i+1}} e_{i,i+1},\quad i=1,2,...,2N-1,$$

$$h_i=(-1)^i(e_{i,i}+e_{i+1,i+1}),\quad h_{2N}=\sum_{k=1}^{2N}(-1)^k e_{k,k},$$

$$e_0=zqe_{1,2N},\quad f_0=z^{-1}q^{-1}e_{2N,1},\quad h_0=e_{1,1}+e_{2N,2N}. \qquad \text{(III.4)}$$

Proposition 2: The Drinfeld generators are represented on V_z by

$$H_m^i=(-1)^{i+1}\frac{[m]_q}{m}q^{(-1)^i m}(q^{x_i}z)^m(e_{i,i}+e_{i+1,i+1}),$$

$$H_m^{2N}=z^m\frac{[2m]_q}{m}\left[-q^m\sum_{l=1}^N e_{2l,2l}+\sum_{l=1}^N (y+(l-1)(1-q^m))(e_{2l-1,2l-1}+e_{2l,2l})\right],$$

$$H_0^i=(-1)^{i+1}(e_{i,i}+e_{i+1,i+1}),\quad H_0^{2N}=\sum_{k=1}^{2N}(-1)^{k+1}e_{k,k},$$

$$X_m^{+,i}=(q^{x_i}z)^m e_{i,i+1},\quad X_m^{-,i}=(-1)^{i+1}(q^{x_i}z)^m e_{i+1,i}, \qquad \text{(III.5)}$$

and on V_z^{*S} by

$$H_m^i=(-1)^i\frac{[m]_q}{m}q^{(-1)^{i+1}m}(q^{-x_i z})^m(e_{i,i}+e_{i+1,i+1}),$$

$$H_m^{2N}=-z^m\frac{[2m]_q}{m}\left[-q^{-m}\sum_{l=1}^N e_{2l,2l}+\sum_{l=1}^N (-y^*+(l-1)(1-q^{-m}))(e_{2l-1,2l-1}+e_{2l,2l})\right],$$

$$H_0^i=(-1)^i(e_{i,i}+e_{i+1,i+1}),\quad H_0^{2N}=\sum_{k=1}^{2N}(-1)^k e_{k,k},$$

$$X_m^{+,i}=-(-1)^i q^{(-1)^i}(q^{-x_{i_z}})^m e_{i+1,i},\quad X_m^{-,i}=-q^{(-1)^{i+1}}(q^{-x_{i_z}})^m e_{i,i+1}, \qquad \text{(III.6)}$$

where $i=1,...,2N-1$, $x_i=\Sigma_{l=1}^i(-1)^{l+1}=[(-1)^{i+1}+1]/2$ and y, y^* are arbitrary constants.

IV. FREE BOSON REALIZATION AT LEVEL ONE

We use the notations similar to those in Refs. 9 and 13. Let us introduce bosonic oscillators $\{A_n^j,\ c_n^l,\ Q_{A^j},\ Q_{c^l}|n\in\mathbf{Z},\ j=1,2,...,2N,\ l=1,2,...,N\}$ which satisfy the commutation relations

$$[A_n^j,A_m^{j'}]=\delta_{n+m,0}\frac{[a_{jj'}n]_q[n]_q}{n},\quad [A_0^j,Q_{A^{j'}}]=a_{jj'},$$

$$[c_n^l,c_m^{l'}]=\delta_{ll'}\delta_{n+m,0}\frac{[n]_q^2}{n},\quad [c_0^l,Q_{c^{l'}}]=\delta_{ll'}. \qquad \text{(IV.1)}$$

The remaining commutation relations are zero. Introduce the currents

$$H^j(z;\kappa)=Q_{A^j}+A_0^j\ln z-\sum_{n\neq 0}\frac{A_n^j}{[n]_q}q^{\kappa|n|}z^{-n},$$

$$c^l(z)=Q_{c^l}+c_0^l\ln z-\sum_{n\neq 0}\frac{c_n^l}{[n]_q}z^{-n} \tag{IV.2}$$

and set

$$H_\pm^j(z)=H^j(q^{+(1/2)}z;-\tfrac{1}{2})-H^j(q^{\mp(1/2)}z;\tfrac{1}{2})=\pm(q-q^{-1})\sum_{n>0}A_{\pm n}^j z^{\mp n}\pm A_0^i\ln q. \tag{IV.3}$$

We make a basis transformation and express A_n^j and Q_{A^j} in terms of a new set of bosonic oscillators $\{a_n^j,\ Q_{a^j}|j=1,2,...,2N\}$ as

$$A_n^i=(-1)^{i+1}(a_n^i+a_n^{i+1}),\ \ A_n^{2N}=\frac{q^n+q^{-n}}{2}\sum_{l=1}^{2N}(-1)^{l+1}a_n^l,$$

$$Q_{A^i}=(-1)^{i+1}(Q_{a^i}+Q_{a^{i+1}}),\ \ Q_{A^{2N}}=\sum_{l=1}^{2N}(-1)^{l+1}Q_{a^l}, \tag{IV.4}$$

where $i=1,...,2N-1$ and $\{a_n^j,Q_{a^j}\}$ satisfy the commutation relations

$$[a_n^j,a_m^{j'}]=(-1)^{j+1}\delta_{jj'}\delta_{n+m,0}\frac{[n]_q^2}{n},\quad [a_0^j,Q_{a^{j'}}]=(-1)^{j+1}\delta_{jj'}. \tag{IV.5}$$

Now we state our main result in this section on the free boson realization of $U_q[gl(\widehat{N|N})]$ at level one.

Theorem 1: *The Drinfeld generators of $U_q[gl(\widehat{N|N})]$ at level one are realized by the free boson fields as*

$$c=1,\quad \psi^{\pm,j}(z)=e^{H_\pm^j(z)},\quad j=1,2,...,2N,$$

$$X^{\pm,i}(z)=:e^{\pm H^i(z;\mp 1/2)}Y^{\pm,i}(z):F^{\pm,i},\quad i=1,2,...,2N-1, \tag{IV.6}$$

where

$$F^{\pm,2k-1}=\prod_{l=1}^{k-1}e^{\pm\sqrt{-1}\pi a_0^{2l-1}},\quad F^{\pm,2k}=\prod_{l=1}^{k}e^{\mp\sqrt{-1}\pi a_0^{2l-1}},$$

$$Y^{+,2k-1}(z)=e^{c^k(z)},$$

$$Y^{-,2k-1}(z)=\frac{1}{z(q-q^{-1})}(e^{-c^k(qz)}-e^{-c^k(q^{-1}z)}),$$

$$Y^{+,2k}(z)=Y^{-,2k-1}(z)=\frac{1}{z(q-q^{-1})}(e^{-c^k(qz)}-e^{-c^k(q^{-1}z)}),$$

$$Y^{-,2k}(z)=-Y^{+,2k-1}(z)=-e^{c^k(z)},\quad k=1,2,...,N. \tag{IV.7}$$

Proof: We prove this theorem by checking that they satisfy the defining relations (II.20) of $U_q[gl(\widehat{N|N})]$ with $c=1$. It is easily seen that the first two relations in (II.20) are true by construction. The third and fourth ones follow from the definition of $X^{\pm,i}(z)$ and the commutativity between a_n^j and c_n^l. So we only need to check the last three relations in (II.20).

We write

$$Z^{\pm,i}(z) = :e^{\pm H^i(z;\mp 1/2)}:F^{\pm,i}. \tag{IV.8}$$

It is easily shown that

$$Z^{+,i}(z)Z^{+,i'}(w) = \begin{cases} :Z^{+,i}(z)Z^{+,i'}(w): & \text{for } a_{ii'}=0 \text{ and } i \leqslant i', \\ -:Z^{+,i}(z)Z^{+,i'}(w): & \text{for } a_{ii'}=0 \text{ and } i > i', \\ (z-q^{-1}w):Z^{+,i}(z)Z^{+,i'}(w): & \text{for } a_{ii'}=1 \text{ and } i < i', \\ -(z-q^{-1}w):Z^{+,i}(z)Z^{+,i'}(w): & \text{for } a_{ii'}=1 \text{ and } i > i', \\ (z-q^{-1}w)^{-1}:Z^{+,i}(z)Z^{+,i'}(w): & \text{for } a_{ii'}=-1, \end{cases} \tag{IV.9}$$

$$Z^{-,i}(z)Z^{-,i'}(w) = \begin{cases} :Z^{-,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=0 \text{ and } i \leqslant i', \\ -:Z^{-,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=0 \text{ and } i > i' \\ (z-qw):Z^{-,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=1 \text{ and } i < i', \\ -(z-qw):Z^{-,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=1 \text{ and } i > i', \\ (z-qw)^{-1}:Z^{-,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=-1, \end{cases} \tag{IV.10}$$

$$Z^{+,i}(z)Z^{-,i'}(w) = \begin{cases} :Z^{+,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=0 \text{ and } i \leqslant i', \\ -:Z^{+,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=0 \text{ and } i > i', \\ (z-w)^{-1}:Z^{+,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=1 \text{ and } i < i', \\ -(z-w)^{-1}:Z^{+,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=1 \text{ and } i > i', \\ (z-w):Z^{+,i}(z)Z^{-,i'}(w): & \text{for } a_{ii'}=-1. \end{cases} \tag{IV.11}$$

We have similar formulas for $Z^{+,i'}(w)Z^{+,i}(z)$, $Z^{-,i'}(w)Z^{-,i}(z)$, and $Z^{-,i'}(w)Z^{+,i}(z)$.

We now compute operator products $Y^{+,i}(z)Y^{+,i'}(w)$ and $Y^{+,i}(z)Y^{-,i'}(w)$. It is easily seen from the definition of $Y^{\pm,i}(z)$ that the nontrivial products are those corresponding to $i=i'$ and $a_{ii'}=1$. Note that $a_{ii'}=1$ whenever $i=2k-1$, $i'=2k$ (or $i=2k$, $i'=2k-1$) where $k=1,2,...,N-1$. The corresponding operator products are

$$Y^{+,2k-1}(z)Y^{+,2k}(w) = \frac{1}{w(q-q^{-1})}\left(\frac{:e^{c^k(z)}e^{-c^k(qw)}:}{z-qw} - \frac{:e^{c^k(z)}e^{-c^k(q^{-1}w)}:}{z-q^{-1}w}\right),$$

$$Y^{-,2k-1}(z)Y^{-,2k}(w) = -\frac{1}{z(q-q^{-1})}\left(\frac{:e^{-c^k(qz)}e^{c^k(w)}:}{qz-w} - \frac{:e^{-c^k(q^{-1}z)}e^{c^k(w)}:}{q^{-1}z-w}\right),$$

$$Y^{+,2k-1}(z)Y^{-,2k}(w) = -(z-w):Y^{+,2k-1}(z)Y^{-,2k}(w):$$

$$\begin{aligned} Y^{+,2k}(z)Y^{-,2k-1}(w) = \frac{1}{zw(q-q^{-1})}\big(&q(z-w):e^{-c^k(qz)}e^{-c^k(qw)}: \\ &-q^{-1}(z-w):e^{-c^k(q^{-1}z)}e^{-c^k(q^{-1}w)}: \\ &-(qz-q^{-1}w):e^{-c^k(qz)}e^{-c^k(q^{-1}w)}: \\ &-(q^{-1}z-qw):e^{-c^k(q^{-1}z)}e^{-c^k(qw)}:\big). \end{aligned} \tag{IV.12}$$

Since $Y^{+,2k}(z)=Y^{-,2k-1}(z)$ and $Y^{-,2k}(z)=-Y^{+,2k-1}(z)$, the products $Y^{+,i}(z)Y^{+,i}(w)$ and $Y^{\pm,i}(z)Y^{\mp,i}(w)$ can be deduced from (IV.12). For example,

$$Y^{+,2k}(z)Y^{-,2k}(w)=-Y^{+,2k}(z)Y^{+,2k-1}(w)=\frac{1}{z(q-q^{-1})}\left(\frac{:e^{-c^k(q^{-1}z)}e^{c^k(w)}:}{q^{-1}z-w}-\frac{:e^{-c^k(qz)}e^{c^k(w)}:}{qz-w}\right). \tag{IV.13}$$

By means of (IV.9), (IV.11), (IV.10), and (IV.12) we can show that the last three relations in (II.20) are satisfied by (IV.7). For instance,

$$\begin{aligned}[X^{+,2k}(z),Z^{-,2k-1}(w)]&=-\frac{1}{zw(q-q^{-1})^2}:Z^{+,2k}(z)Z^{-2,k-1}(w):\left(\frac{1}{z-w}+\frac{1}{w-z}\right)\\&\quad\times((qz-q^{-1}w):e^{-c^k(qz)}e^{-c^k(q^{-1}w)}:+(q^{-1}z-qw):e^{-c^k(q^{-1}z)}e^{-c^k(qw)}:)\\&=-\frac{1}{z^2w(q-q^{-1})^2}:Z^{+,2k}(z)Z^{-,2k-1}(w):\delta\left(\frac{w}{z}\right)\\&\quad\times((qz-q^{-1}w):e^{-c^k(qz)}e^{-c^k(q^{-1}w)}:+(q^{-1}z-qw):e^{-c^k(q^{-1}z)}e^{-c^k(qw)}:)\\&=0.\end{aligned} \tag{IV.14}$$

V. BOSONIZATION OF LEVEL-ONE VERTEX OPERATORS

In this section, we study the level-one vertex operators[6] of $U_q[gl(\widehat{N|N})]$. Let $V(\lambda)$ be the highest weight $U_q[gl(\widehat{N|N})]$-module with the highest weight λ. Consider the following intertwiners of $U_q[gl(\widehat{N|N})]$-modules:[2]

$$\Phi_\lambda^{\mu V}(z):\quad V(\lambda)\to V(\mu)\otimes V_z, \tag{V.1}$$

$$\Phi_\lambda^{\mu V^*}(z):\quad V(\lambda)\to V(\mu)\otimes V_z^{*S}, \tag{V.2}$$

$$\Psi_\lambda^{V\mu}(z):\quad V(\lambda)\to V_z\otimes V(\mu), \tag{V.3}$$

$$\Psi_\lambda^{V^*\mu}(z):\quad V(\lambda)\to V_z^{*S}\otimes V(\mu). \tag{V.4}$$

They are intertwiners in the sense that for any $x\in U_q[gl(\widehat{N|N})]$,

$$\Xi(z)\cdot x=\Delta(x)\cdot\Xi(z),\quad \Xi(z)=\Phi_\lambda^{\mu V}(z),\ \Phi_\lambda^{\mu V^*}(z),\ \Psi_\lambda^{V\mu}(z),\ \Psi_\lambda^{V^*\mu}(z). \tag{V.5}$$

These intertwiners are even operators, that is their gradings are $[\Phi_\lambda^{\mu V}(z)]=[\Phi_\lambda^{\mu V^*}(z)]=[\Psi_\lambda^{V\mu}(z)]=[\Psi_\lambda^{V^*\mu}(z)]=0$. According to Ref. 2, $\Phi_\lambda^{\mu V}(z)$ ($\Phi_\lambda^{\mu V^*}(z)$) is called type I (dual) vertex operator and $\Psi_\lambda^{V\mu}(z)$ ($\Psi_\lambda^{V^*\mu}(z)$) type II (dual) vertex operator.

We expand the vertex operators as[2]

$$\Phi_\lambda^{\mu V}(z)=\sum_{j=1}^{2N}\Phi_{\lambda,j}^{\mu V}(z)\otimes v_j,\quad \Phi_\lambda^{\mu V^*}(z)=\sum_{j=1}^{2N}\Phi_{\lambda,j}^{\mu V^*}(z)\otimes v_j^*,$$

$$\Psi_\lambda^{V\mu}(z)=\sum_{j=1}^{2N}v_j\otimes\Psi_{\lambda,j}^{V\mu}(z),\quad \Psi_\lambda^{V^*\mu}(z)=\sum_{j=1}^{2N}v_j^*\otimes\Psi_{\lambda,j}^{V^*\mu}(z). \tag{V.6}$$

Then the intertwining property (V.5) reads in terms of components,

$$\sum \Phi^{\mu V}_{\lambda,j}(z)x\otimes v_j(-1)^{[v_j][x]}=\sum x_{(1)}\Phi^{\mu V}_{\lambda,j}(z)\otimes x_{(2)}v_j(-1)^{[v_j][x_{(2)}]},$$

$$\sum \Phi^{\mu V^*}_{\lambda,j}(z)x\otimes v^*_j(-1)^{[v^*_j][x]}=\sum x_{(1)}\Phi^{\mu V^*}_{\lambda,j}(z)\otimes x_{(2)}v^*_j(-1)^{[v^*_j][x_{(2)}]},$$

$$\sum v_j\otimes\Psi^{V\mu}_{\lambda,j}(z)x=\sum x_{(1)}v_j\otimes x_{(2)}\Psi^{V\mu}_{\lambda,j}(z)(-1)^{[v_j][x_{(2)}]},$$

$$\sum v^*_j\otimes\Psi^{V^*\mu}_{\lambda,j}(z)x=\sum x_{(1)}v^*_j\otimes x_{(2)}\Psi^{V^*\mu}_{\lambda,j}(z)(-1)^{[v^*_j][x_{(2)}]}, \tag{V.7}$$

where we have used the notation $\Delta(x)=\Sigma_x x_{(1)}\otimes x_{(2)}$ and the fact that the vertex operators are even which implies $[\Phi^{\mu V}_{\lambda,j}(z)]=[\Phi^{\mu V^*}_{\lambda,j}(z)]=[\Psi^{V\mu}_{\lambda,j}(z)]=[\Psi^{V^*\mu}_{\lambda,j}(z)]=[v_j]=[(-1)^j+1]/2$.

Introduce the even operators $\phi(z)$, $\phi^*(z)$, $\psi(z)$, and $\psi^*(z)$,

$$\phi(z)=\sum_{j=1}^{2N}\phi_j(z)\otimes v_j,\quad \phi^*(z)=\sum_{j=1}^{2N}\phi^*_j(z)\otimes v^*_j,$$

$$\psi(z)=\sum_{j=1}^{2N}v_j\otimes\psi_j(z),\quad \psi^*(z)=\sum_{j=1}^{2N}v^*_j\otimes\psi^*_j(z). \tag{V.8}$$

The grading of the components is given by $[\phi_j(z)]=[\phi^*_j(z)]=[\psi_j(z)]=[\psi^*_j(z)]=[(-1)^j+1]/2$. Now we state

Proposition 3: Assume that the operators $\phi(z)$, $\phi^*(z)$, $\psi(z)$, $\psi^*(z)$ satisfy the intertwining relations (V.7). Then the operators $\phi(z)$ and $\psi(z)$ with respect to V_z are determined by the components $\phi_{2N}(z)$ and $\psi_1(z)$, respectively. With respect to V^{*S}_z, the operators $\phi^*(z)$ and $\psi^*(z)$ are determined by $\phi^*_1(z)$ and $\psi^*_{2N}(z)$, respectively. More explicitly, we have for $l=1,2,...,2N-1$,

$$(-1)^l\phi_l(z)=[\phi_{l+1}(z),f_l]_{q^{(-1)^l}},$$

$$[\phi_l(z),f_l]_{q^{(-1)^l}}=0,$$

$$[\phi_k(z),f_l]=0,\quad k\neq l,l+1, \tag{V.9}$$

$$q^{(-1)^{l+1}}\phi^*_{l+1}(z)=[\phi^*_l(z),f_l]_{q^{(-1)^{l+1}}},$$

$$[\phi^*_{l+1}(z),f_l]_{q^{(-1)^{l+1}}}=0,$$

$$[\phi^*_k(z),f_l]=0,\quad k\neq l,l+1, \tag{V.10}$$

$$\psi_{l+1}(z)=[\psi_l(z),e_l]_{q^{(-1)^{l+1}}},$$

$$[\psi_{l+1}(z),e_l]_{q^{(-1)^{l+1}}}=0,$$

$$[\psi_k(z),e_l]=0,\quad k\neq l,l+1, \tag{V.11}$$

$$(-1)^{l+1}q^{(-1)^l}\psi^*_l(z)=[\psi^*_{l+1}(z),e_l]_{q^{(-1)^l}},$$

$$[\psi^*_l(z),e_l]_{q^{(-1)^l}}=0,$$

$$[\psi_k^*(z),e_l]=0,\quad k\neq l,l+1. \tag{V.12}$$

Next we determine the relations of the components $\phi_{2N}(z)$, $\phi_1^*(z)$, $\psi_1(z)$, $\psi_{2N}^*(z)$ and the Drinfeld generators. By means of proposition 1 and the intertwining relations, we have

Proposition 4: For $\phi(z)$ *associated with* V_z,

$$\begin{aligned}
&[\phi_{2N}(z),X^{+,i}(w)]=0,\\
&q^{h_i}\phi_{2N}(z)q^{-h_i}=q^{-\delta_{i,2N-1}}\phi_{2N}(z),\\
&[H_n^i,\phi_{2N}(z)]=-\delta_{i,2N-1}q^{(3/2)n}\frac{[n]_q}{n}z^n\phi_{2N}(z),\\
&[H_{-n}^i,\phi_{2N}(z)]=-\delta_{i,2N-1}q^{-(1/2)n}\frac{[n]_q}{n}z^{-n}\phi_{2N}(z);
\end{aligned} \tag{V.13}$$

for $\phi^*(z)$ *associated with* V_z^*,

$$\begin{aligned}
&[\phi_1^*(z),X^{+,i}(w)]=0,\\
&q^{h_i}\phi_1^*(z)q^{-h_i}=q^{\delta_{i,1}}\phi_1^*(z),\\
&[H_n^i,\phi_1^*(z)]=\delta_{i,1}q^{(3/2)n}\frac{[n]_q}{n}z^n\phi_1^*(z),\\
&[H_{-n}^i,\phi_1^*(z)]=\delta_{i,1}q^{-(1/2)n}\frac{[n]_q}{n}z^{-n}\phi_1^*(z);
\end{aligned} \tag{V.14}$$

for $\psi(z)$ *associated with* V_z,

$$\begin{aligned}
&[\psi_1(z),X^{-,i}(w)]=0,\\
&q^{h_i}\psi_1(z)q^{-h_i}=q^{-\delta_{i,1}}\psi_1(z),\\
&[H_n^i,\psi_1(z)]=-\delta_{i,1}q^{(1/2)n}\frac{[n]_q}{n}z^n\psi_1(z),\\
&[H_{-n}^i,\psi_1(z)]=-\delta_{i,1}q^{-(3/2)n}\frac{[n]_q}{n}z^{-n}\psi_1(z);
\end{aligned} \tag{V.15}$$

and for $\psi^*(z)$ *associated with* V_z^*,

$$\begin{aligned}
&[\psi_{2N}^*(z),X^{-,i}(w)]=0,\\
&q^{h_i}\psi_{2N}^*(z)q^{-h_i}=q^{\delta_{i,2N-1}}\psi_{2N}^*(z),\\
&[H_n^i,\psi_{2N}^*(z)]=\delta_{i,2N-1}q^{(1/2)n}\frac{[n]_q}{n}z^n\psi_{2N}^*(z),\\
&[H_{-n}^i,\psi_{2N}^*(z)]=\delta_{i,2N-1}q^{-(3/2)n}\frac{[n]_q}{n}z^{-n}\psi_{2N}^*(z).
\end{aligned} \tag{V.16}$$

In order to obtain bosonized expressions of the vertex operators, we introduce the following combinations of the bosonic oscillators

$$A_n^{*i}=\sum_{l=1}^{2N-1} a_{il}^{-1}A_n^l+\frac{2}{q^n+q^{-n}}a_{i,2N}^{-1}A_n^{2N},$$

$$A_0^{*i}=\sum_{l=1}^{2N} a_{il}^{-1}A_0^l,\quad Q_{A^i}^{*}=\sum_{l=1}^{2N} a_{il}^{-1}Q_{A^l},\quad i=1,2,\ldots,2N-1, \tag{V.17}$$

which satisfy the relations

$$[A_n^{*i},A_m^{i'}]=\delta_{ii'}\delta_{n+m,0}\frac{[n]_q^2}{n},$$

$$[A_n^{*i},A_m^{*i'}]=a_{ii'}^{-1}\delta_{n+m,0}\frac{[n]_q^2}{n},$$

$$[A_0^{*i},Q_{A^{i'}}]=\delta_{ii'},\quad [A_0',Q_{A^{i'}}^{*}]=\delta_{ii'},$$

$$[A_0^{*i},Q_{A^{i'}}^{*}]=a_{ii'}^{-1},\quad i,i'=1,2,\ldots,2N-1. \tag{V.18}$$

Introduce the currents,

$$H^{*,j}(z;\kappa)=Q_{A^j}^{*}+A_0^{*j}\ln z-\sum_{n\neq 0}\frac{A_n^{*j}}{[n]_q}q^{k|n|}z^{-n}. \tag{V.19}$$

Now we state our main theorem in this section on the bosonic realization of the operators $\phi(z)$, $\phi^{*}(z)$, $\psi(z)$, and $\psi^{*}(z)$ at level one. Thanks to the previous propositions, we only need to determine one component for each operator and the other components are represented by the integral of the currents.

Theorem 2: *The components* $\phi_{2N}(z)$, $\phi_1^{*}(z)$, $\psi_1(z)$, *and* $\psi_{2N}^{*}(z)$ *can be realized explicitly as follows:*

$$\phi_{2N}(z)=:e^{-H^{*,2N-1}(qz;(1/2))}e^{c^N(qz)}:e^{-\sqrt{-1}\pi a_0^1}\prod_{l=1}^{N-1}e^{-\sqrt{-1}\pi[(2N+l)/2N]a_0^{2l+1}},$$

$$\phi_1^{*}(z)=:e^{H^{*,1}(qz;(1/2))}:\prod_{l=1}^{N-1}e^{\sqrt{-1}\pi[(2N-l)/2N]a_0^{2l+1}},$$

$$\psi_1(z)=:e^{-H^{*,1}(qz;-(1/2))}:\prod_{l=1}^{N-1}e^{\sqrt{-1}\pi[(2N-l)/2N]a_0^{2l+1}},$$

$$\psi_{2N}^{*}(z)=\frac{1}{z(q-q^{-1})}:e^{H^{*,2N-1}(qz;-(1/2))}(e^{-c^N(q^2z)}-e^{-c^N(z)}):$$
$$\times e^{-\sqrt{-1}\pi a_0^1}\prod_{l=1}^{N-1}e^{-\sqrt{-1}\pi[(2N+1)/2N]a_0^{2l+1}}. \tag{V.20}$$

Proof: This theorem is proved by checking that the construction satisfies all the intertwining relations.

Remark: The following inverse elements of the extended Cartan matrix are needed to determine the cocycle factors appearing in above theorem:

$$a^{-1}_{2N-1,2l}=a^{-1}_{2N-1,2l+1}=-\frac{l}{N},\quad l=1,2,...,N-1,$$

$$a^{-1}_{2N-1,1}=0,\quad a^{-1}_{2N-1,2N}=\frac{1}{2N},$$

$$a^{-1}_{1,2l-1}=a^{-1}_{1,2l}=\frac{N-l}{N},\quad l=1,2,...,N-1,$$

$$a^{-1}_{1,2N-1}=0,\quad a^{-1}_{1,2N}=\frac{1}{2N}. \tag{V.21}$$

We are now in a position to state the following result:

Proposition 5: The vertex operators $\Phi_\lambda^{\mu V}(z)$, $\Phi_\lambda^{\mu V^*}(z)$, $\Psi_\lambda^{V\mu}(z)$, *and* $\Psi_\lambda^{V^*\mu}(z)$, *if they exist, have the same bosonized expressions as the operators* $\phi(z)$, $\phi^*(z)$, $\psi(z)$, *and* $\psi^*(z)$, *respectively*.

This proposition follows immediately from the fact that the formers and the latters obey the same intertwining properties. Identifying $\Phi_\lambda^{\mu V}(z)$, $\Phi_\lambda^{\mu V^*}(z)$, $\Psi_\lambda^{V\mu}(z)$, and $\Psi_\lambda^{V^*\mu}(z)$ with $\phi(z)$, $\phi^*(z)$, $\psi(z)$, and $\psi^*(z)$, respectively, then the bosonic realization of the vertex operators is easily seen to be given by propositions 3, 4 and theorem 2.

ACKNOWLEDGMENTS

The financial support from Australian Research Council through a Queen Elizabeth II Fellowship Grant is gratefully acknowledged.

[1] B. Davies, O. Foda, M. Jimbo, T. Miwa, and A. Nakayashiki, Commun. Math. Phys. **151**, 89 (1993).
[2] M. Jimbo and T. Miwa, *Algebraic Analysis of Solvable Lattice Models*, CBMS Regional Conference Series in Mathematics (AMS, Providence, RI, 1995), Vol. 85.
[3] V. E. Korepin, N. M. Bogoliubov, and A. G. Izergin, *Quantum Inverse Scattering Method and Correlation Functions* (Cambridge University Press, Cambridge, 1993).
[4] F. A. Smirnov, *Form Factors in Completely Integrable Models of Quantum Field Theory* (World Scientific, Singapore, 1992).
[5] *Conformal Invariance and Applications to Statistical Mechanics*, edited by C. Itzykson, H. Saleur, and J.-B. Zuber (World Scientific, Singapore, 1988).
[6] I. B. Frenkel and N. Yu Reshetikhin, Commun. Math. Phys. **146**, 1 (1992).
[7] I. B. Frenkel and N. Jing, Proc. Natl. Acad. Sci. USA **85**, 9373 (1988).
[8] D. Bernard, Lett. Math. Phys. **17**, 239 (1989).
[9] H. Awata, S. Odake, and J. Shiraishi, Commun. Math. Phys. **162**, 61 (1994).
[10] Y. Koyama, Commun. Math. Phys. **164**, 277 (1994).
[11] N. Jing, S.-J. Kang, and Y. Koyama, Commun. Math. Phys. **174**, 367 (1995).
[12] N. Jing and K. C. Misra, e-print q-alg/9701034.
[13] K. Kimura, J. Shiraishi, and J. Uchiyama, Commun. Math. Phys. **188**, 367 (1997).
[14] H. Awata, S. Odake, and J. Shiraishi, Lett. Math. Phys. **42**, 271 (1997).
[15] W.-L. Yang and Y.-Z. Zhang, Nucl. Phys. B **547**, 599 (1999).
[16] R. Gade, Nucl. Phys. B **500**, 547 (1997).
[17] Y.-Z. Zhang and M. D. Gould, e-print math.QA/9809156, J. Math. Phys. (in press).
[18] V. G. Drinfeld, Sov. Math. Dokl. **36**, 212 (1988).
[19] V. Chari and A. Pressley, Commun. Math. Phys. **142**, 261 (1991).
[20] V. G. Kac, Lect. Notes Math. **676**, 597 (1978).
[21] L. Frappat, A. Sciarrino, and P. Sorba, Commun. Math. Phys. **121**, 457 (1989).
[22] S. M. Khoroshkin and V. N. Tolstoy, Commun. Math. Phys. **141**, 599 (1991).
[23] H. Yamane, e-print q-alg/9603015.

[24] Y.-Z. Zhang, J. Phys. A **30**, 8325 (1997).
[25] M. D. Gould and Y.-Z. Zhang, Lett. Math. Phys. **44**, 291 (1998).
[26] N. Yu. Reshetikhin and M. A. Semenov-Tian-Shansky, Lett. Math. Phys. **19**, 133 (1990).
[27] J. Ding and I. B. Frenkel, Commun. Math. Phys. **156**, 277 (1993).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 11 NOVEMBER 1999

Erratum: "Analytical solution of the relativistic Coulomb problem with a hard core interaction for a one-dimensional spinless Salpeter equation" [J. Math. Phys. 40, 1119 (1999)]

F. Brau
Université de Mons-Hainaut, Place du Parc 20, B-7000 Mons, Belgium

(Received 15 June 1999; accepted for publication 7 July 1999)

[S0022-2488(99)01911-8]

The supposedly exact formula for the bound-state spectrum of the one-dimensional relativistic Coulomb problem with a hard core interaction, given in Eq. (33) of this paper, is incorrect. The first part of this paper (Sec. II) which contains the evaluation of the action of the operator $\sqrt{-d_x^2+m^2}$ on the functions $x^n e^{-\beta x}$ (using the integral representation of this operator obtained in this paper), is correct if we consider the whole x axis. Application to the Coulomb problem must involve normalizable wave functions. That is why we have only considered the positive part of the x axis leading to wave functions of the form: $P_n(x)e^{-\beta x}\theta(x)$, where $\theta(x)$ is the Heaviside function. Unfortunately, as a consequence of the nonlocality of the square root operator, the action of this operator on these functions leads to rather different expressions than the action on the functions $P_n(x)e^{-\beta x}$, as calculated in Sec. II, even if we consider the results of these actions only on the positive x axis. We have not obtained the correct spectrum and the correct wave functions. We regret that we did not point out this problem earlier.

0022-2488/99/40(11)/6125/1/$15.00

Erratum: ''Rest frame system for asymptotically flat space–times'' [J. Math. Phys. 39, 6631 (1998)]

Osvaldo M. Moreschi[a)] and Sergio Dain[b)]
FaMAF, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

(Received 11 June 1999; accepted for publication 12 July 1999)

There is an error in the original presentation of the theorem 6.1. The corrected Sec. VI is:

VI. GLOBAL EXISTENCE OF NICE SECTIONS NEAR THE STATIONARY CASE

The results appearing in the previous sections deal only with local properties of nice sections; in this section the issue of global existence will be tackled. In Ref. 2 it was proved that there exists a four-parameter family of nice sections in a stationary space–time. This suggests that when radiation is ''small'' one could probably find an analogous result. In what follows we will apply once more the implicit function theorem to prove this assertion.

Let $C^0(\mathcal{I}^+)$ represent the vector space of all bounded continuous functions on scri plus, with the norm $\|F\|_{C^0(\mathcal{I}^+)}=\sup|F(u,\zeta,\zeta)|$. The vector space $C^0(\mathcal{I}^+)$ defines a Banach space.

Let X, Y and Z represent the spaces defined in Sec. III, and let us denote with Ξ the space whose elements are of the form $\xi=(x,F)$, with $x\in X$, $F\in C^0(\mathcal{I}^+)$ and with norm $\|\xi\|=\|x\|_X+\|F\|_{C^0(\mathcal{I}^+)}$. The space Ξ is a Banach space. Consider now the map from (Ξ,Y) into Z defined by

$$\Phi(\xi,y)=Dy-\left(\int_{u_0}^{x+y}F(u,\zeta,\overline{\zeta})du+\Psi(u_0,\zeta,\overline{\zeta})+K(x+y,F;\zeta,\overline{\zeta})^3M(x+y,F)\right), \tag{71}$$

which coincide with the map $f(x,y)$ defined in Sec. III, but where now it is taken into account explicitly the dependence on the flux F. The Fréchet derivative of this map with respect to the variable y is

$$\Phi_y(\xi,y)\,\delta y=D\,\delta y-\left[F(u=\gamma=x+y,\zeta,\overline{\zeta})\,\delta y+K^3\left(\frac{4P_{\underline{a}}}{M}-3K\ l_{\underline{a}}(\zeta,\overline{\zeta})\right)\delta P^{\underline{a}}\right], \tag{72}$$

where $P_{\underline{a}}$, M, and K are evaluated at $u=\gamma=x+y$; in particular one has

$$P^{\underline{a}}(\gamma)=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\overline{\zeta})\int_{u_0}^{\gamma}F(u,\zeta,\overline{\zeta})du\ dS^2+P^{\underline{a}}(u=u_0) \tag{73}$$

and $\delta P^{\underline{a}}$ is given by

$$\delta P^{\underline{a}}=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\overline{\zeta})F(u=\gamma=x+y,\zeta,\overline{\zeta})\,\delta y\ dS^2. \tag{74}$$

Let us denote with ξ_0 the nonradiation case, that is, $\xi_0=(x_0,F=0)$; we have shown in Ref. 2 that there exists a relation $y_0(x_0)$ such that $\Phi(\xi_0,y_0)=0$, for any translation x_0; in other words,

---

[a)]Member of CONICET; electronic mail: moreschi@fis.uncor.edu
[b)]Fellowship holder from CONICOR; electronic mail: dain@fis.uncor.edu

in the absence of radiation $(F=0)$, the map $\gamma_0(x_0)=x_0+y_0(x_0)$ defines a four-parameter family of nice sections globally on scri. It is observed that $\delta P^{\underline{a}}(F=0)=0$; therefore

$$\Phi_y(\xi(x,F=0),y)\,\delta y=D\,\delta y, \tag{75}$$

and it is clear that the map $\delta z=\Phi_y(\xi(x,F=0),y)\,\delta y$ from Y onto Z is invertible; that is, it is an homeomorphism of Y onto Z.

The implicit function theorem implies that there exists a unique continuous mapping G defined in a neighborhood $U_0\subset C^0(\mathcal{I}^+)$ of $F=0$, $G:U_0\rightarrow Y$, parametrized by x_0, such that $G(\xi(x_0,F=0))=y_0$ and $\Phi(\xi(x_0,F),y)=0$ for $F\in U_0$.

Let us call *small radiation data* the one that satisfies $\dot{\sigma}\dot{\bar{\sigma}}=F\in U_0$ in the last result. Then, the continuity properties contained in the implicit function theorem and the fact that x lives in a four-dimensional vector space, imply the following result:

Theorem 6.1:

For small radiation data there exists a global four-parameter family of nice sections at scri.

ACKNOWLEDGMENT

We would like to thank A. Rendall for pointing out the error and for discussions.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Quantum description of rigidly or adiabatically constrained four-particle systems and supersymmetry

E. Baloïtcha and M. N. Hounkonnou[a)]
Unité de Recherche en Physique Théorique (URPT), Institut de Mathématiques et de Sciences Physiques (IMSP), B.P. 613 Porto-Novo, Bénin

(Received 18 May 1999; accepted for publication 13 July 1999)

A general formalism for the quantum description of many-body systems is developed, using the principal-axis hyperspherical parametrization of coordinates. This formalism is applied to four-particle systems for which the exact kinetic energy operator is derived using a model constraints and dynamical constraints. Then, using the supersymmetry and shape invariance approach, we obtained in a closed form the eigenvalues and eigenfunctions of a wide class of noncentral potentials for the adiabatically constrained four-particle systems. © *1999 American Institute of Physics.* [S0022-2488(99)00112-7]

I. INTRODUCTION

In the field of molecular and chemical physics, the parametrization (for the purpose of dynamical studies) of the configuration of N-atom system which is likely to undergo very large amplitude deformations is, in most cases, looked for in terms of coordinates among which the distances are as few as possible and, on the contrary, the angles as many as possible. This is so in the context of quantum mechanics and in view of variational calculations requiring the evaluation of matrix elements, because of a practical intrinsic weakness of the radial function basis which have to span the domain $[0,\infty[$ and are often of a delicate use for computing Hamiltonian matrix elements, whereas (owing to the cyclic character of the angular coordinates) there exist a variety of practicable analytical angular function basis sets, for instance the spherical harmonics basis for any pair of angles that are mathematically spherical.

Moreover, the configuration of the molecular system is usually thought of in terms of $3N-6$ internal coordinates. This is consistent with the description of the rotation of a very deformable system as that of a body-fixed (BF) frame of reference, i.e., a frame with its origin at the center of mass and its axes rotating in a prescribed manner when the atoms (or only a few of them) move. With respect to the space-fixed (SF) frame of reference (with origin at the center of mass and axes parallel to those of the laboratory-fixed (LF) frame), the BF frame is oriented by means of three Euler angles. Finally, the translation of the center of mass in the LF frame is generally eliminated.

It has been established[1–5] that an useful expression of the exact nonrelativistic quantum-mechanical Hamiltonian operator of a molecular system made up of N atoms viewed in a BF frame of reference (O,x,y,z) and described in size and shape by a complete set of internal coordinates $\{q^i;\ i=1,2,\ldots,3N-6\}$ is very complex and contains at least seven terms plus the potential energy term.

The $3N-6$ internal coordinates q^i are invariant under rotations; this means that the expressions $q^i=q^i(x^1,x^2,\ldots,x^\alpha,\ldots,x^{3N})$, where the x^α are the Cartesian BF-coordinates of the particles, are unchanged whatever orientation of the BF-frame. In most cases, such coordinates turn out to be either distances between two points, or planar angles between three points, or still dihedral atoms or centers of mass of groups of atoms as well. Next, the orientation of the BF-frame is supposed to be uniquely defined by means of the three so-called ''axial constraints.''[4,6]

[a)]Electronic mail: hounkon@syfed.bj.refer.org

Last of all, the Hamiltonian of the system expressed with respect to the particle coordinates in the BF-frame, is designed to operate on a wave function that is normalized according to the usual (i.e., Euclidean) normalization convention, i.e., the volume element $d\tau_1 = [g(\mathbf{q})]^{-1/2} dq^1 dq^2 \cdots dq^{3N-6} d\psi d\beta d\gamma$ (the parameters will be defined later) is to be used when calculating matrix elements.

The hyperspherical parametrizations constitute the most appropriate answer to the requirement of few distances and many angles among the internal coordinates describing molecular systems with possible very deformed configurations. Their main feature is that the length-type internal coordinates are reduced to a single one (the so-called hyperradius ρ), all the others being angles (with various possible definitions, see below) so that, the overall rotation being also considered, the system configuration and orientation are described by one distance (ρ) and $3N-4$ angles. Whatever the actual parametrization, the hyperradius definition is unique and invariant,[7]

$$\rho^2 = \sum_{\lambda=1}^{N-1} \mu_\lambda \rho_\lambda^2 = \sum_{\alpha=1}^{N} \sum_{g} |r_{\alpha g}|^2,$$

where, any set of $n=N-1$ Jacobi vectors $\{\rho_\lambda^2\}$ being associated with the N-atom system, μ_λ and ρ_λ are, respectively, the reduced mass and the length of one vector, μ is the totally symmetric characteristic mass, and $r_{\alpha g}$ denotes the component $(g=x,y,z)$ of the mass-weighted position vector $\mathbf{r}_\alpha$ of the αth particle viewed in the body-fixed frame (BF).

The analytical treatment of this system has been the subject of many investigations these last years.[3,4] Among the most remarkable works, we can quote the papers by Parker,[8] Kosloff *et al.*,[9] and other remarkable contributions.[10–18]

To treat four-particle systems, we subject the system to model constraints, e.g., either in rigidifying a few bonds or entire atomic groups or in adjusting the variations of some internal coordinates (called inactive) to those of a few coordinates considered the most important ones for the dynamics. Indeed imposing model constraints alters the dimensions of the configuration space, so each differential operator must be modified accordingly.[19] Moreover, when only internal deformations are considered, the molecule is studied for zero total angular momentum $\mathbf{J}$, a condition that is physically realistic since $\mathbf{J}$ is a constant of motion. One of the purposes of this paper is, for the adiabatically constrained four-particle systems, to show that noncentral potentials that are separable in their coordinates and may have up to nine parameters can be solved by the operator method by using the well known results for the various shape invariance potentials (SIP). More precisely, we show that the energy eigenvalues and eigenfunctions of noncentral but separable potentials in PAH coordinates can be simply written down by applying supersymmetry and shape invariance to $\theta_3, \phi, \theta,$ and ρ dependent potentials.

The paper is organized as follows: In Sec. II, we recall a general formalism of quantization of many-body systems. In Sec. III, we construct the quantum-mechanical Hamiltonian for four-particle systems in terms of the quasimomenta using principal axis hyperspherical (PAH) coordinates. In Sec. IV, we subject the four-particle to model constraints and derive the exact kinetic energy operator derived. In Sec. V, using the results for the known SIP, we obtain in a closed form the eigenvalues and the eigenfunctions for noncentral but separable potentials.

II. GENERAL FORMALISM OF THE MANY BODY SYSTEM QUANTIZATION

Let $\mathbf{q}=(q^1,q^2,\ldots,q^{3n})$, $n=N-1$ be a set of curvilinear coordinates well suited for the description of a dynamical system, and let $\mathbf{p}=(p_1,p_2,\ldots,p_{3n})$ be the corresponding conjugate momenta in classical mechanics. The kinetic energy T can always be expressed as a quadratic form of momenta,

$$T = \frac{1}{2} \sum_{i,j=1}^{3n} p_i g^{ij}(\mathbf{q}) p_j, \tag{2.1}$$

where

$$[g^{ij}]=\begin{bmatrix} \underline{S}(\mathbf{q}) & \underline{C}^{\dagger}(\mathbf{q}) \\ \underline{C}(\mathbf{q}) & \underline{I}(\mathbf{q}) \end{bmatrix}^{-1}=\begin{bmatrix} \underline{\Sigma} & \underline{\Gamma}^{\dagger} \\ \underline{\Gamma} & \underline{\mu} \end{bmatrix} \tag{2.2}$$

with $\underline{S}(\mathbf{q})$ the $(3N-6)$-dimensional symmetric matrix of elements,

$$S_{ij}=\sum_{\alpha=1}^{N}[\partial_{q_i}r_{\alpha x}\partial_{q_j}r_{\alpha x}+\partial_{q_i}r_{\alpha y}\partial_{q_j}r_{\alpha y}+\partial_{q_i}r_{\alpha z}\partial_{q_j}r_{\alpha z}]\quad (i,j=1,2,\ldots,3N-6); \tag{2.3}$$

$\underline{C}(\mathbf{q})$ the $3\times(3N-6)$ Coriolis matrix of elements,

$$C_{gj}=\sum_{\alpha=1}^{N}[r_{\alpha g'}\partial_{q_j}r_{\alpha g''}-r_{\alpha g''}\partial_{q_j}r_{\alpha g'}] \tag{2.4}$$

$$(j=1,2,\ldots,3N-6;g,g',g''\ \text{is an even permutation of}\ x,y,z),$$

and $\underline{I}(\mathbf{q})$ being the usual inertia tensor represented by the three-dimensional symmetric matrix of elements,

$$I_{gg'}=\sum_{\alpha=1}^{N}\left[\sum_{g''}r_{\alpha g''}^{2}\delta_{gg'}-r_{\alpha g}r_{\alpha g'}\right]\quad (g,g'=x,y,z). \tag{2.5}$$

Now, one is always at liberty to use a *nonsingular* $3n$-dimensional matrix $\underline{B}(\mathbf{q})$, depending only on the coordinates, to transform the momenta into quasimomenta, namely,

$$p_K=\sum_{i=1}^{3n}[\underline{B}(\mathbf{q})]_K^i p_i\qquad (K=1,2,\ldots,3n). \tag{2.6}$$

The determinant of $\underline{B}(\mathbf{q})$ is denoted by

$$t^{-1}=\mathrm{Det}(\underline{B}(\mathbf{q})). \tag{2.7}$$

T can be rewritten as

$$T=\frac{1}{2}\sum_{K,L=1}^{3n}p_K g^{KL}(\mathbf{q})p_L \tag{2.8}$$

with

$$g^{KL}(\mathbf{q})=\sum_{i,j=1}^{3n}[\underline{B}^{-1}(\mathbf{q})]_i^K g^{ij}(\mathbf{q})[\underline{B}^{-1}(\mathbf{q})]_j^L\qquad (K,L=1,2,\ldots,3n) \tag{2.9}$$

so that, if $g(\mathbf{q})=\mathrm{Det}([g^{ij}(\mathbf{q})])$ and $\tilde{g}(\mathbf{q})=\mathrm{Det}([g^{KL}(\mathbf{q})])$, then

$$g(\mathbf{q})=\tilde{g}(\mathbf{q})t^{-2}(\mathbf{q}). \tag{2.10}$$

In quantum mechanics, let p_i be replaced by the Hermitian operator $\hat{p}_i$. Operators $\hat{q}$ and $\hat{p}$ act in the standard Hilbert space $\mathcal{H}$ and must satisfy the Heisenberg commutation relations

$$[\hat{q}^i,\hat{p}_j]=i\hbar\,\delta_j^i,\quad [\hat{q},\hat{I}_d]=[\hat{p},\hat{I}_d]=0.$$

Here $\hat{I}_d$ is the identity operator, $\hbar$ is Planck's constant, and the bracket means the commutator $[\hat{A},\hat{B}]=\hat{A}\hat{B}-\hat{B}\hat{A}$.

$$\hat{p}_i=-i\hbar\frac{1}{g^{-1/4}}\partial_{q_i}g^{-1/4}\quad (i=1,2,\ldots,3n). \tag{2.11}$$

In keeping, for the quasimomentum operators, with the same definition as above

$$\hat{p}_K=\sum_{i=1}^{3n}[\underline{B}(\mathbf{q})]^i_K\hat{p}_i\quad (K=1,2,\ldots,3n). \tag{2.12}$$

As long as the Euclidean volume element of the configuration space

$$d\tau_1=dx^1dx^2dx^3\cdots dx^{3N}=[g(\mathbf{q})]^{-1/2}dq^1dq^2\cdots dq^{3N-6}d\psi\,d\beta\,d\gamma$$

is used for the calculation of the matrix elements as integrals, the differential kinetic energy operator must be written as

$$\hat{T}=\frac{1}{2}\sum_{K,L=1}^{3n}\hat{p}_K g^{KL}(\mathbf{q})\hat{p}_L+V_D(\mathbf{q}) \tag{2.13}$$

$$=-\frac{\hbar^2}{2}\sum_{K,L=1}^{3n}\left[\frac{1}{g^{-1/4}}\partial_{q^K}g^{KL}\partial_{q^L}g^{-1/4}-\left(\frac{1}{g^{-1/4}}\partial_{q^K}g^{KL}\partial_{q^L}g^{-1/4}\right)\right] \tag{2.14}$$

$$=-\frac{\hbar^2}{2}\sum_{K,L=1}^{3n}\left[\frac{1}{g^{-1/2}}\partial_{q^K}g^{KL}g^{-1/2}\partial_{q^L}\right] \tag{2.15}$$

$$=-\frac{\hbar^2}{2}\sum_{K,L=1}^{3n}\left[g^{KL}\partial_{q^K}\partial_{q^L}+\left(\frac{1}{g^{-1/2}}\partial_{q^K}g^{KL}g^{-1/2}\right)\partial_{q^L}\right] \tag{2.16}$$

$$=-\frac{\hbar}{2}\sum_{K,L=1}^{3n}[g^{KL}\partial_{q^K}\partial_{q^L}+\Gamma^L_1\partial_{q^L}]. \tag{2.17}$$

The four expressions (2.13)–(2.16) call for several definitions and a few comments.

(i) Since the kinetic energy operator must be purely differential (in Euclidean normalization), whereas

$$\sum_{K,L=1}^{3n}\left[\frac{1}{g^{-1/4}}\partial_{q^K}g^{KL}\partial_{q^L}g^{-1/4}\right]$$

only is not; $V_D(\mathbf{q})$ the so-called ''extra potential'' (term discussed in Ref. 14, Appendix 2), which is a pure multiplicative operator, appears to make up the T differential.

(ii) When an operator appears in an expression in parentheses, it is always assumed not to operate beyond these parentheses.

If the classical Hamiltonian is known in the form

$$H=T(\mathbf{q,p})+V(\mathbf{q}), \tag{2.18}$$

where $T(\mathbf{q,p})$ is written as in Eq. (2.8), its quantization simply amounts to rewriting it as in Eq. (2.13) by means of the quantities in Eqs. (2.11) that are to be explicitly calculated.

III. QUANTUM-MECHANICAL PAH HAMILTONIAN FOR THE UNCONSTRAINED FOUR-PARTICLE SYSTEM IN TERMS OF THE QUASIMOMENTA

In this case, the set of the PAH coordinates is $(\rho,\theta,\phi,\theta_1,\theta_2,\theta_3,\psi,\beta,\gamma)$, where ρ is the hyperradius; θ and ϕ the spherical coordinates to parametrize the mass-weighted gyration radii; ψ,β, and γ the three usual Eulerian angles to orient the (BF) with respect to the (SF); θ_1,θ_2, and θ_3 are the so-called hyperangles.[10,11,15] We deal with $n=N-1=3$ in all the above expressions (2.1)–(2.18). So we have

$$[g^{KL}]=\begin{pmatrix} \underline{s}(\mathbf{q}) & \underline{0} & \underline{0} \\ \underline{0} & \frac{1}{\rho^2}\underline{b}(\mathbf{q}) & \frac{1}{\rho^2}\underline{d}(\mathbf{q}) \\ \underline{0} & \frac{1}{\rho^2}\underline{d}(\mathbf{q}) & \frac{1}{\rho^2}\underline{b}(\mathbf{q}) \end{pmatrix}, \tag{3.19}$$

where each element in the matrix above denotes a 3×3 diagonal matrix, respectively,

$$\underline{s}(\mathbf{q})=\begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\rho^2} & 0 \\ 0 & 0 & \frac{1}{\rho^2\sin^2\theta} \end{pmatrix}, \tag{3.20}$$

$$\underline{b}(\mathbf{q})=\begin{pmatrix} \frac{\cos^2\theta+\sin^2\theta\sin^2\phi}{(\cos^2\theta-\sin^2\theta\sin^2\phi)^2} & 0 & 0 \\ 0 & \frac{\cos^2\theta+\sin^2\theta\cos^2\phi}{(\cos^2\theta-\sin^2\theta\cos^2\phi)^2} & 0 \\ 0 & 0 & \frac{1}{\sin^2\theta\cos^2 2\phi} \end{pmatrix}, \tag{3.21}$$

and

$$\underline{d}(\mathbf{q})=\begin{pmatrix} \frac{\sin 2\theta\sin\phi}{(\cos^2\theta-\sin^2\theta\sin^2\phi)^2} & 0 & 0 \\ 0 & \frac{\sin 2\theta\cos\phi}{(\cos^2\theta-\sin^2\theta\sin^2\phi)^2} & 0 \\ 0 & 0 & \frac{\sin 2\phi}{\sin^2\theta\cos^2 2\phi} \end{pmatrix}. \tag{3.22}$$

ρ,θ, and ϕ are the three spherical coordinates which allow the parametrization of the three mass-weighted gyration radii of the system. The following result is obtained:

$$\tilde{g}(\mathbf{q})=\frac{16}{[\rho^8\sin^3\theta\cos 2\phi(4\cos^2\theta\cos 2\theta+\sin^4\theta\sin^2 2\phi)]^2}. \tag{3.23}$$

Here, the number of the Jacobi vectors describing the internal conformation of the system is $n=3$.

In addition, we have

$$\begin{pmatrix} \hat{K}_x \\ \hat{K}_y \\ \hat{K}_z \end{pmatrix} = -i\hbar \underline{\tilde{B}}(\theta_1,\theta_2) \begin{pmatrix} \partial_{\theta_1} \\ \dfrac{1}{\sqrt{\sin\theta_2}} \partial_{\theta_3} \sqrt{\sin\theta_2} \\ \partial_{\theta_3} \end{pmatrix}, \tag{3.24}$$

where the 3×3-dimensional matrix $\underline{\tilde{B}}(\theta_1,\theta_2)$, according to relation (2.6), is defined by

$$\underline{\tilde{B}} = \underline{\omega}^{*-1},$$

where $\underline{\omega}^{*-1}$ is still the 3×3 matrix, appears in the expression of $\underline{B}(\mathbf{q})$, the overall matrix transforming the momenta into quasimomenta, as follows:

$$\underline{B}(\mathbf{q})_{9\times9} = \begin{pmatrix} \underline{1}_{3\times3} & \underline{0}_{3\times3} & \underline{0}_{3\times3} \\ \underline{0}_{3\times3} & \underline{\tilde{B}}(\theta_1,\theta_2)_{3\times3} & \underline{0}_{3\times3} \\ \underline{0}_{3\times3} & \underline{0}_{3\times3} & \underline{\omega}^{*-1}(\gamma,\beta)_{3\times3} \end{pmatrix}. \tag{3.25}$$

This corresponds to the fact that

(i) The operators in the left hand column vector of Eq. (3.24) are actually the quasimomentum operators to be used;

(ii) The 3×3 unit matrix in the upper of $\underline{B}(\mathbf{q})$ in Eq. (3.25) is for

$$\hat{p}_\rho = -i\hbar\,(1/\rho^4)\,\partial_\rho \rho^4, \quad \hat{p}_\theta = -i\hbar\,(1/\sqrt{g_\theta})\,\partial_\theta \sqrt{g_\theta}, \quad \hat{p}_\phi = -i\hbar\,\partial_\phi,$$

with

$$g_\theta = \sin^3\theta(4\cos^2\theta\cos 2\theta + \sin^4\theta\sin^2 2\phi),$$

which are unaffected by the transformation, and

(iii) $\underline{\omega}^{*-1}(\gamma,\beta)$ in the lower right corner is for the overall-rotation angular-momentum vector operator definition,

$$\underline{\omega}^{*-1}(\theta_1,\theta_2) = \begin{pmatrix} -\sin\theta_1\cot\theta_2 & \cos\theta_1 & \sin\theta_1/\sin\theta_2 \\ -\cos\theta_1\cot\theta_2 & -\sin\theta_1 & \cos\theta_1/\sin\theta_2 \\ 1 & 0 & 0 \end{pmatrix}. \tag{3.26}$$

This matrix stands for the transformation of angular (quasi) velocities into angular (quasi) momenta, namely,

$$\begin{pmatrix} \hat{K}_x \\ \hat{K}_y \\ \hat{K}_z \end{pmatrix} = -i\hbar \underline{\omega}^{*-1}(\theta_1,\theta_2) \begin{pmatrix} \partial_{\theta_1} \\ \dfrac{1}{\sqrt{\sin\theta_2}} \partial_{\theta_2} \sqrt{\sin\theta_2} \\ \partial_{\theta_3} \end{pmatrix} \tag{3.27}$$

as well as

$$\begin{pmatrix} \hat{J}_x \\ \hat{J}_y \\ \hat{J}_z \end{pmatrix} = -i\hbar \underline{\omega}^{*-1}(\gamma,\beta) \begin{pmatrix} \partial\gamma \\ \frac{1}{\sqrt{\sin\beta}}\partial_\beta\sqrt{\sin\beta} \\ \partial_\psi \end{pmatrix}. \tag{3.28}$$

$\hat{\mathbf{K}}$ is the so-called pseudo-angular momentum vector operator,[10,12] which concerns the internal motion of the system (the hyperangles θ_1, θ_2, and θ_3 are actually internal coordinates), whereas $\hat{\mathbf{J}}$ is the usual total angular momentum operator for the overall rotation of the system (Eulerian angles ψ,β,γ). This results in

$$\begin{gathered} \hat{J}_z = -i\hbar\partial_\gamma \\ \hat{\mathbf{J}}^2 = -\hbar^2\left[\frac{1}{\sin^2\beta}(\partial_\gamma^2+\partial_\psi^2-2\cos\beta\partial_\psi\partial_\gamma)+\partial_\beta^2+\cot\beta\partial_\beta\right] \end{gathered} \tag{3.29}$$

and similar expressions for $\hat{\mathbf{K}}$, by substituting θ_1, θ_2, and θ_3 for γ,β, and ψ.

Let us come back to the quantization problem. Since $\underline{B}$ is block triangular, we have

$$t^{-1} = \mathrm{Det}(\underline{\omega}^{*-1}(\theta_1,\theta_2))\mathrm{Det}(\underline{\omega}^{*-1}(\gamma,\beta)) \tag{3.30}$$

so that

$$|t^{-1}| = \frac{1}{\sin\beta\sin\theta_2} \tag{3.31}$$

and Eqs. (2.10) and (3.23) yield

$$[g(\mathbf{q})]^{-1/2} = \rho^8\sin^3\theta\cos 2\phi(4\cos^2\theta\cos 2\theta+\sin^4\theta\sin^2 2\phi)\sin\theta_2\sin\beta. \tag{3.32}$$

g clearly does not depend on $\theta_1, \theta_3, \gamma$, and ψ.

Now, for Euclidean volume element of the configuration space,

$$d\tau_1 = [g(\mathbf{q})]^{-1/2} d\rho\, d\theta\, d\phi\, d\theta_1\, d\theta_2\, d\theta_3\, d\psi\, d\beta\, d\gamma.$$

The PAH Hamiltonian is written as

$$\hat{H}_1 = \hat{T}_1 + V(\rho,\theta,\phi,\theta_1,\theta_2,\theta_3). \tag{3.33}$$

Equations (2.16) and (3.32) yield

$$\hat{T}_1 = -\frac{\hbar^2}{2}\left[\partial_\rho^2+\frac{8}{2\rho}\partial_\rho\right]+\frac{[\hat{\Lambda}_8^2]_1}{2\rho^2} \tag{3.34}$$

$$= -\frac{\hbar^2}{2}\hat{\Lambda}_9, \tag{3.35}$$

$\hat{\Lambda}_9$ being the Laplacian. $[\hat{\Lambda}_8^2]_1$ is the so-called ''grand-angular-momentum'' operator, for the eight-dimensional hypersphere embedded into the nine-dimensional configuration space of the 3 mass-weighted Jacobi vectors describing the system, after separation of the center of mass. In applying the quantization rules given in section above, $[\hat{\Lambda}_8^2]_1$ is written as

$$[\hat{\Lambda}_8^2]_1 = -\hbar^2\left[\partial_\theta^2 + \Lambda_\theta^*\partial_\theta + \frac{1}{\sin^2\theta}[\partial_\phi^2 + \Lambda_\phi^*\partial_\phi]\right] + \sum_{g=x,y,z}[b_{gg}(\hat{J}_g^2 + \hat{K}_g^2) + 2d_{gg}\hat{J}_g\hat{K}_g], \tag{3.36}$$

where the b_{gg} and d_{gg}'s depend on θ and ϕ only [they are given in Eqs. (3.21) and (3.22), respectively] and Λ_θ^* and Λ_ϕ^* are given by

$$\Lambda_\theta^* = 3\cot\theta - 2\sin 2\theta \frac{4\cos^2\theta + 2\cos 2\theta - \sin^2\theta\sin^2 2\phi}{4\cos^2\theta\cos 2\theta + \sin^4\theta\sin^2 2\phi};$$

$$\Lambda_\phi^* = 2\left[-\tan 2\phi + \frac{\sin^4\theta\sin 4\phi}{4\cos^2\theta\cos 2\theta + \sin^4\theta\sin^2 2\phi}\right]. \tag{3.37}$$

Obviously, the part of $\hat{T}_1$ that is linear in $\partial\theta_1, \partial\theta_2, \partial\theta_3$ as it comes out of Eq. (2.16), i.e., the $\Gamma^L(\mathbf{q})$ coefficients in Eq. (2.18), has been entirely re-expressed in terms of the quasimomentum operators $\hat{\mathbf{K}}_g$ $(g=x,y,z)$.

IV. RIGIDLY OR ADIABATICALLY CONSTRAINED FOR THE FOUR-PARTICLE SYSTEM

In order to avoid untractable calculations, we decide to freeze m of $3N-6$ internal coordinates, in other words if we decide to subject the system to m model constraints, it is convenient to split up the set of the $3N-6$ coordinates $\mathbf{q}$ as follows: $\mathbf{q}=(\mathbf{q}',\mathbf{q}'')$ where $\mathbf{q}'=(q^1,q^2,\ldots,q^n)$ is the set of $n=3N-6-m$ coordinates that remain free and $\mathbf{q}''=(q^{n+1},q^{n+2},\ldots,q^{3N-6})$ is the set of the m coordinates that are frozen, so that imposing the m model constraints comes down to putting $\mathbf{q}''=\mathbf{q}''_0$ (and hence $\dot{\mathbf{q}}''=0$) with $\mathbf{q}''_0$ a set of m constants. The partition of $\underline{\Sigma}$ into $\underline{\Sigma}'$, $\underline{\Sigma}''$, and $\underline{\sigma}$, and that of $\underline{\Gamma}$ into $\underline{\Gamma}'$ and $\underline{\Gamma}''$, result from the splitting of $\mathbf{q}$ into $\mathbf{q}'$ and $\mathbf{q}''$. The constraint-free Hamiltonian expression of T [Eq. (2.1)] can be written also as Eq. (4.38),[19,20]

$$T = \frac{1}{2}[\mathbf{p}'|\mathbf{p}''|\mathbf{J}]\begin{bmatrix}\underline{\Sigma}' & \underline{\sigma}^\dagger & \underline{\Gamma}'^\dagger \\ \underline{\sigma} & \underline{\Sigma}'' & \underline{\Gamma}''^\dagger \\ \underline{\Gamma}' & \underline{\Gamma}'' & \underline{\mu}\end{bmatrix}\begin{bmatrix}\mathbf{p}'^\dagger \\ \mathbf{p}'^\dagger \\ \mathbf{J}^\dagger\end{bmatrix}, \tag{4.38}$$

where $\mathbf{p}'$ and $\mathbf{p}''$ denote the quasimomenta conjugate to the internal coordinates $\mathbf{q}'$, $\mathbf{q}''$, and $\mathbf{J}$ the total angular momentum.

(a) Distinguishing carefully the independent dynamic variables $\mathbf{p}$ and $\mathbf{J}$ in the Hamiltonian approach, cf. Eq. (2.1), from their expressions as linear combinations of $\dot{\mathbf{q}}'$ and ω, the latter being the independent dynamic variables in the Lagrangian frame work, will turn out to be an important point.

(b) Imposing model constraints amounts to cancelling components of $\dot{\mathbf{q}}$ (other than ω), which equally means suppressing the corresponding lines and columns of matrix $[g_{ij}]$ (inverse of $[g^{ij}]$ in relation (2.2)); whereas forcing the dynamical constraint $\mathbf{J}=0$ amounts to suppressing the last three lines and columns in matrix $[g^{ij}]$.

Let us apply the dynamic constraint $\mathbf{J}=0$ and the model constraints $\{\dot{\theta}_1=\dot{\theta}_2=0\}$ to our four-particle system, putting $\mathbf{q}'=(\rho,\theta,\ \phi,\ \theta_3)$. The square matrix in relation (4.38) yields [see also (3.19)],

$$\underline{\Sigma}' = [g^{KL}] = [g^{KL}]_{\text{constraint}}, \tag{4.39}$$

where $\underline{\Sigma}'$ is the diagonal matrix obtained by suppressing the 4th, 5th, and the three last columns and rows of the matrix in relation (3.19). So the elementary volume of the constrained space configuration is $d\tau_1 = d\tau_1|_{\text{constraint}} = \rho^3\sin^2\theta\cos 2\phi d\rho\, d\theta\, d\phi\, d\theta_3$.

The constrained Hamiltonian writes

$$\hat{H}=\hat{H}|_{\text{constraint}}=-\frac{\hbar^2}{2}\left[\left(\partial_\rho^2+\frac{3}{\rho}\partial_\rho\right)+\frac{1}{\rho^2\sin^2\theta}\partial_\theta\sin^2\theta\partial_\theta\right.$$
$$\left.+\frac{1}{\rho^2\sin^2\theta}(\partial_\phi^2-2\tan 2\phi\partial_\phi)+\frac{1}{\rho^2\sin^2\theta\cos^2 2\phi}\partial_{\theta_3}^2\right]. \tag{4.40}$$

V. SUPERSYMMETRY CONSTRAINED FOUR-PARTICLE SYSTEM

In this section, we first explain how supersymmetric shape-invariant potentials may be used to solve a problem of constrained four-particle system. The potential of this problem in PAH coordinates is denoted by $V(\rho,\theta,\phi,\theta_3)$, and the problem is formally equivalent to that of one particle in a four-dimensional hyperspherical coordinate.[21] It is well known that further separation of the variables $\rho,\theta,\phi,$ and θ_3 in classical (or quantum) equations of motion takes place[22] for the form

$$V(\rho,\theta,\phi,\theta_3)=V(\rho)+\frac{V(\theta)}{\rho^2}+\frac{V(\phi)}{\rho^2\sin^2\theta}+\frac{V(\theta_3)}{\rho^2\sin^2\theta\cos^2 2\phi}, \tag{5.41}$$

where $V(\rho),V(\theta),V(\phi)$, and $V(\theta_3)$ are arbitrary functions of their argument. For such a form, the problem is immediately integrable and the four constants of motion are easily obtained. The Schrödinger equation for the eigenvalue problem is ($\hbar=1,\rho^2\rightarrow 2\rho^2$),

$$[\hat{H}|_{\text{con}}+V(\rho,\theta,\phi,\theta_3)]\Psi_{nn_1n_2n_3}=E_{nn_1n_2n_3}\Psi_{nn_1n_2n_3}. \tag{5.42}$$

Writing the eigenfunction $\Psi(\rho,\theta,\phi,\theta_3)$ as

$$\Psi(\rho,\theta,\phi,\theta_3)=\frac{U(\rho)}{\rho^{3/2}}\times\frac{H(\theta)}{\sin\theta}\times\frac{Q(\phi)}{\cos^{1/2}2\phi}\times K(\theta_3), \tag{5.43}$$

after some algebra, we obtain

$$-\frac{d^2}{d\theta_3^2}K(\theta_3)+V(\theta_3)K(\theta_3)=k^2K(\theta_3), \tag{5.44}$$

$$-\frac{d^2}{d\phi^2}Q(\phi)+[V(\phi)+(k^2-1)\sec^2 2\phi]Q(\phi)=m^2Q(\phi), \tag{5.45}$$

$$-\frac{d^2}{d\theta^2}H(\theta)+[V(\theta)+(m^2-1)\csc^2\theta]H(\theta)=l^2H(\theta), \tag{5.46}$$

$$-\frac{d^2}{d\rho^2}U(\rho)+\left[V(\rho)+\frac{(l^2-\frac{1}{4})}{\rho^2}\right]U(\rho)=E_{nn_1n_2n_3}U(\rho). \tag{5.47}$$

The four constants of motion are the eigenvalues $E_{nn_1n_2n_3},l^2,m^2$, and k^2.

At this stage, we use the well known result of supersymmetry that for shape-invariant potentials, the solutions may be obtained algebraically.[23–25] To appreciate what is meant by shape-invariance, consider a ''super potential'' $W(u)$, where the variable u may stand for ρ,θ,ϕ, or θ_3. The supersymmetric partners $V_-(u)$ and $V_+(u)$ are (W^2-W') and (W^2+W'), respectively, the prime denoting a differentiation with respect to u. If the pair $V_\pm$ are of the same shape, but differ only in the parameters which appear in them, they are said to be shape invariant. More precisely, if the partner potentials $V_\pm(u,a_1)$ satisfy the condition[26,27]

$$V_+(u,a_1)=V_-(u,a_2)+\Re(a_1), \tag{5.48}$$

where a_1 is a set of parameters, a_2 is a function of a_1 (say $a_2=f(a_1)$) and the remainder $\Re(a_1)$ is independent of u, $V_\pm$ are said to be shape invariant. In such a case, the energy spectrum of the Hamiltonian with the potential V_- is given by

$$E_n^{(-)}(a_1)=\sum_{j=1}^{n}\Re(a_j), \quad E_0^{(-)}(a_1)=0 \tag{5.49}$$

with $a_j=f^{j-1}(a_1)$, i.e., the function f applies $(j-1)$-times. Subsequently, it has been shown that the corresponding eigenfunction can be obtained algebraically.[25,28,29]

Take

$$V(\rho,\theta,\phi,\theta_3)=\frac{\Omega^2}{4}\rho^2+\frac{\delta}{\rho^2}+\frac{C}{\rho^2\sin^2\theta}+\frac{D}{\rho^2\cos^2\theta}+\frac{F}{\rho^2\sin^2\theta\sin^2 2\phi}+\frac{G}{\rho^2\sin^2\theta\cos^2 2\phi}$$
$$+\frac{I}{\rho^2\sin^2\theta\cos^2 2\phi\sin^2\mu\theta_3}+\frac{J}{\rho^2\sin^2\theta\cos^2 2\phi\cos^2\mu\theta_3}, \tag{5.50}$$

with nine parameters $\Omega,\delta,C,D,F,G,I,J,$ and μ. Comparing Eqs. (5.41) and Eqs. (5.50), we obtain

$$V(\theta_3)=I\csc^2\mu\theta_3+J\sec^2\mu\theta_3 \tag{5.51}$$

and if the superpotential W is chosen as

$$W=A\tan\mu\theta_3-B\cot\mu\theta_3, \quad A,B>0, \tag{5.52}$$

then, the corresponding (shape invariant) potentials $V_\mp$ are

$$V_\mp=-(A+B)^2+A(A\mp\mu)\sec^2\mu\theta_3+B(B\mp\mu)\csc^2\mu\theta_3. \tag{5.53}$$

Hence,

$$V_+(\theta_3,A,B,\mu)=V_-(\theta_3,A+\mu,B+\mu,\mu)+(A+B+2\mu)^2-(A+B)^2. \tag{5.54}$$

On comparing Eqs. (5.51) and (5.53), we see that, discounting the overall constant $-(A+B)^2$, $V(\theta_3)$ and V_- are identical if $I=B(B-\mu)$ and $J=A(A-\mu)$.

Thus, the energy eigenvalues k^2 of Eq. (5.44) with $V(\theta_3)=V_-(\theta_3)$ follow from Eqs. (5.44), (5.48), and (5.54),

$$k^2=(A+B+2n\mu)^2 \quad n=0,1,2,\dots . \tag{5.55}$$

The corresponding eigenfunctions are given by[27,28]

$$K(\theta_3)=(\sin\mu\theta_3)^{B/\mu}(\cos\mu\theta_3)^{A/\mu}P_n^{(B/\mu)-(1/2),(A/\mu)-(1/2)}(\cos 2\mu\theta_3), \tag{5.56}$$

where $P_n^{\alpha,\beta}$ is the Jacobi polynomial.

Using the same algebra procedure as before, the eigenvalues m^2 and l^2 with their eigenfunctions are found to be, respectively,

$$m^2=(A_1+B_1+4n_1+1)^2, \tag{5.57}$$

$$Q(\phi)=(\sin 2\phi)^{B/2}(\cos 2\phi)^{(A_1/2)+(1/2)}P_{n_1}^{(B/_1/2)-(1/2),(A_1/2)}(\cos 4\phi), \tag{5.58}$$

where $G+k^2=A_1^2,\ F=B_1(B_1-2),\ n_1=0,1,2,\ldots,$ and

$$l^2=(A_2+B_2+2n_2+1)^2, \tag{5.59}$$

$$H(\theta)=(\sin\theta)^{B_2+1}(\cos\theta)^{A_2}P_{n_1}^{B_2+(1/2),A_2-(1/2)}(\cos 2\theta), \tag{5.60}$$

where $C+m^2=B_2^2, D=A_2(A_2-1), n_2=0,1,2,\ldots.$

Choosing the superpotential

$$W=A_3\rho-\frac{(B_3+1)}{\rho}, \tag{5.61}$$

the corresponding shape invariant potential V_- is given by

$$V_-(\rho)=A_3^2\rho^2+\frac{B_3(B_3+1)}{\rho^2}-A_3(2B_3+3). \tag{5.62}$$

Some algebra shows that

$$V_+(\rho,A_3,B_3)=V_-(\rho,A_3,B_3+1)-A_3(2B_3+3)+A_3(2B_3+1). \tag{5.63}$$

The energy eigenvalues of the potential in Eq. (5.50) is $((\omega^2/4)=A_2^2,\ \delta+l^2=(B_3+1/2)^2)$,

$$E_{n,n_1,n_2,n_3}=(2n_3+B_3+\tfrac{3}{2})\omega, \tag{5.64}$$

while the corresponding hyperradial part of the eigenfunction is[30]

$$U(\rho)=\rho^{B_3+1}\exp\left[-\frac{\omega\rho^2}{4}\right]L_{n_3}^{B_3+(1/2)}\left(\frac{1}{2}\omega\rho^2\right). \tag{5.65}$$

L_α^β is the Laguerre polynomials.

Thus the total eigenfunction for the noncentral potential Eq. (5.50) is given by the form (5.43), with $U(\rho),H(\theta),Q(\phi),$ and $K(\theta_3)$ given by Eqs. (5.65), (5.60), (5.58), and (5.56), respectively.

The energy eigenvalues are expressed in terms of the nine parameters,

$$E_{n,n_1,n_2,n_3}=[(2n_3+1)+(\delta+l^2)^{1/2}]\omega, \tag{5.66}$$

$$l^2=\left|\left(2n_2+\frac{3}{2}\right)+\sqrt{D+\frac{1}{4}}+\left\{C+\left[2(2n_1+1)+\sqrt{1+F}+\left(G+\left(\sqrt{I+\frac{\mu^2}{4}}+\sqrt{J+\frac{\mu^2}{4}}+(2n+1)\mu\right)^2\right)^{1/2}\right]^2\right\}^{1/2}\right|^2. \tag{5.67}$$

To conclude, let us point out that the four-body constrained systems involved here lead to a solvable potential with nine parameters. This is remarkable. Recall that the most general solvable central potentials of Natanzon-type[25,31] are built with six parameters while the noncentral solvable potential reconstruction needs seven parameters in three dimensions.

ACKNOWLEDGMENTS

The authors would like to thank the Ministère de la Coopération Française (France) and the SCAC (Cotonou) for the financial support of this work, under the Projet CAMPUS, reference No. 94016400.

[1] X. Chapuisat, A. Nauts, and J. P. Brunet, Mol. Phys. **72**, 1 (1991).
[2] A. Nauts and X. Chapuisat, Mol. Phys. **55**, 1287 (1985).
[3] F. Lequéré, Ph.D. thesis, Université de Paris-Sud, 1991.
[4] X. Chapuisat, A. Belafhal, and A. Nauts, J. Mol. Spectrosc. **149**, 274 (1991).
[5] M. N. Hounkonnou, thesis of Doctorat ès Sciences, Université Catholique de Louvain, Institut de Physique Théorique (FYMA), Belgium, 1992.
[6] E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations* (1955).
[7] Eckart, Phys. Rev. **46**, 383 (1934).
[8] J. V. Lill, G. A. Parker, and J. C. Light, Chem. Lett. **89**, 483 (1982).
[9] D. Kosloff and R. Kosloff, J. Comput. Phys. **52**, 35 (1983); J. Chem. Phys. **79**, 1823 (1983).
[10] X. Chapuisat and A. Nauts, Phys. Rev. A **44**, 1328 (1991).
[11] X. Chapuisat, Phys. Rev. A **45**, 4277 (1992).
[12] X. Chapuisat, A. Nauts, and J. P. Brunet, Chem. Phys. Lett. **136**, 153 (1987).
[13] G. O. Sørensen, Top. Curr. Chem. **82**, 97 (1979).
[14] F. Lequére and C. Leforestier, J. Chem. Phys. **94**, 1118 (1991).
[15] N. J. Vilenkin, *Special Functions and the Theory of Group Representation* (AMS Translations, Providence, R.I., 1968), Vol. 22.
[16] H. C. Longuet-Higgins, Mol. Phys. **6**, 445 (1963).
[17] J. G. Freiy, Mol. Phys. **70**, 145 (1990).
[18] J. Avery, B. L. Christensen-Daslgaard, P. S. Larsen, and S. Hengyi, Int. J. Quantum Chem., Symp. **18**, 321 (1984).
[19] X. Chapuisat and A. Nauts, Mol. Phys. **91**, 47 (1997).
[20] F. Gatti, Y. Justum, M. Menou, A. Nauts, and X. Chapuisat, J. Mol. Spectrosc. **181**, 403 (1997).
[21] J. E. Kilpatrick and S. Y. Larsen, Few-Body Syst. **3**, 75 (1987).
[22] L. Landau and E. M. Lifshitz, *Mechanics* (Pergammon, Oxford, 1960, 1976).
[23] L. Infeld and T. E. , Theor. Math. Phys. **23**, 21 (1951).
[24] L. Gendenshtein, JETP Lett. **38**, 356 (1983).
[25] A. Kare and R. K. Baduri, Institute of Physics, Bhubaneswar preprint IP-BBSR/93-65, 1993.
[26] F. Cooper, A. Khare, and U. Sukhatme, *Supersymmetry and Quantum Mechanics* (1998).
[27] A. Kare and R. K. Baduri, J. Phys. A **27**, 2213 (1994).
[28] R. Dutt, A. Khare, and U. P. Sukhatme, Phys. Lett. B **181**, 295 (1986); Am. J. Phys. **56**, 163 (1988).
[29] G. Levai, J. Phys. A **22**, 689 (1989).
[30] F. Calogero, J. Math. Phys. **10**, 2191 (1969).
[31] F. Cooper, J. N. Ginocchio, and A. Khare, Phys. Rev. D **36**, 2458 (1987).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Coulomb wave functions with complex values of the variable and the parameters

Aleksander Dzieciol, Staffan Yngve, and Per Olof Fröman
Department of Theoretical Physics, University of Uppsala, Box 803, S-751 08 Uppsala, Sweden

(Received 8 June 1998; accepted for publication 22 June 1999)

The motivation for the present paper lies in the fact that the literature concerning the Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$ is a jungle in which it may be hard to find a safe way when one needs general formulas for the Coulomb wave functions with complex values of the variable ρ and the parameters L and η. For the Coulomb wave functions and certain linear combinations of these functions we discuss the connection with the Whittaker function, the Coulomb phase shift, Wronskians, reflection formulas $(L\to -L-1)$, integral representations, series expansions, circuital relations $(\rho\to\rho e^{\pm i\pi})$ and asymptotic formulas on a Riemann surface for the variable ρ. The parameters L and η are allowed to assume complex values. © *1999 American Institute of Physics.* [S0022-2488(99)01710-7]

I. INTRODUCTION

The literature on Coulomb wave functions is very extensive and difficult to survey. The presentation of Yost, Breit, and Wheeler,[1] which contains the first general treatment, was later followed by the review article by Hull and Breit.[2] For particular purposes one has found it convenient to introduce a large number of functions that are solutions of the Coulomb differential equation.[3–9] For the standard Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$ in Abramowitz and Stegun,[10] one, in general, assumes the variable ρ to be real, the angular momentum quantum number L to be a non-negative integer, and the parameter η to be real. However, in theoretical physics there appear problems in which one needs Coulomb wave functions with complex values of the variable ρ and the parameters L and η. This is the case, for instance, in Regge pole theory, where the angular momentum is complex, in scattering theory, when the energy is complex, and in quantum defect theory for closed channels. In the existing literature it is, however, hard or impossible to find general formulas for the standard Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$ with complex values of the variable ρ and the parameters L and η. For such values of the parameters the analytic properties of the Coulomb phase shift $\sigma_L(\eta)$ make the discussion of these Coulomb wave functions much more complicated than the discussion of the related Whittaker function, and it is therefore desirable to present the result of such a discussion in a surveyable way.

Our purpose in the present paper is to present and collect general formulas for the standard Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$, which are solutions of the differential equation

$$\frac{d^2\psi}{d\rho^2}+\left(1-\frac{2\eta}{\rho}-\frac{L(L+1)}{\rho^2}\right)\psi=0, \tag{1.1}$$

and certain linear combinations of these functions, viz.,

$$\psi_\pm(L,\eta,\rho)=G_L(\eta,\rho)\pm iF_L(\eta,\rho), \tag{1.2}$$

when the variable ρ and the parameters L and η are allowed to assume complex values without unnecessary restrictions. In Sec. II we express the functions $\psi_+(L,\eta,\rho)$ and $\psi_-(L,\eta,\rho)$ in terms of the Whittaker function. For the Coulomb phase shift $\sigma_L(\eta)$, which appears in these formulas, we investigate the rather complicated limiting properties when L and η are complex and $\eta\to 0$, and we also present the behavior of $\sigma_L(\eta)$ when the quantities $L+1\pm i\eta$ are both large. Wronskian relations for the functions in (1.2) are also presented. In Sec. III we express $\psi_\pm(-L-1,\eta,\rho)$ in terms of $\psi_\pm(L,\eta,\rho)$, and we also express $F_{-L-1}(\eta,\rho)$ and $G_{-L-1}(\eta,\rho)$ in terms of $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$. In these formulas there appears a quantity $x(L,\eta)$, the rather complicated limiting properties of which as $\eta\to 0$ we investigate. From the definition of $x(L,\eta)$ we first derive an alternative formula for $x(L,\eta)$ and then formulas for $\partial x(L,\eta)/\partial L$ and $\partial x(L,\eta)/\partial\eta$. When $2L$ is an integer and $\eta\neq 0$, we obtain a simple formula for $x(L,\eta)$, with the aid of which we can simplify the previously mentioned formulas that express $\psi_\pm(-L-1,\eta,\rho)$ in terms of $\psi_\pm(L,\eta,\rho)$, as well as the formulas that express $F_{-L-1}(\eta,\rho)$ and $G_{-L-1}(\eta,\rho)$ in terms of $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$. In Sec. IV we present integral representations for $\psi_\pm(L,\eta,\rho)$ and $F_L(\eta,\rho)$. In Sec. V, $F_L(\eta,\rho)$ is expanded in powers of ρ, and from this formula we obtain a formula for $\partial \ln F_L(\eta,\rho)/\partial L$ when $2L+1$ is *not* a negative integer. Simplified formulas obtained when $L=-\frac{1}{2}$ are also given. Formulas for $G_L(\eta,\rho)$ are presented in Sec. VI. The cases when $2L$ is not an integer and when $2L$ is an integer (in particular, when $L=-\frac{1}{2}$) are discussed separately. In Sec. VII, $\psi_+(L,\eta,\rho)$ and $\psi_-(L,\eta,\rho)$ are expressed in terms of $\psi_+(L,-\eta,\rho e^{+i\pi})$ and $\psi_-(L,-\eta,\rho e^{+i\pi})$, as well as in terms of $\psi_+(L,-\eta,\rho e^{-i\pi})$ and $\psi_-(L,-\eta,\rho e^{-i\pi})$, and $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$ are expressed in terms of $F_L(-\eta,\rho e^{+i\pi})$ and $G_L(-\eta,\rho e^{+i\pi})$, as well as in terms of $F_L(-\eta,\rho e^{-i\pi})$ and $G_L(-\eta,\rho e^{-i\pi})$. General asymptotic formulas for $\psi_\pm(L,\eta,\rho)$, $F_L(\eta,\rho)$, and $G_L(\eta,\rho)$, valid for fixed complex values of L and η, when ρ tends to infinity, while $\arg\rho$ has an arbitrary fixed value, are given in Sec. VIII.

II. COULOMB WAVE FUNCTIONS EXPRESSED IN TERMS OF THE WHITTAKER FUNCTION, COULOMB PHASE SHIFT, AND WRONSKIANS

One can express $\psi_+(L,\eta,\rho)$ and $\psi_-(L,\eta,\rho)$ in terms of the Whittaker function as follows:

$$\psi_+(L,\eta,\rho)=\exp\{i[\sigma_L(\eta)-(L+i\eta)\pi/2]\}W_{-i\eta,L+1/2}(2\rho e^{-i\pi/2}), \tag{2.1a}$$

$$\psi_-(L,\eta,\rho)=\exp\{-i[\sigma_L(\eta)-(L-i\eta)\pi/2]\}W_{i\eta,L+1/2}(2\rho e^{+i\pi/2}), \tag{2.1b}$$

where $\sigma_L(\eta)$ is the Coulomb phase shift, defined as

$$\sigma_L(\eta)=\frac{1}{2i}[\ln\Gamma(L+1+i\eta)-\ln\Gamma(L+1-i\eta)], \tag{2.2}$$

and the logarithms $\ln\Gamma(L+1\pm i\eta)=\ln|\Gamma(L+1\pm i\eta)|+i\arg\Gamma(L+1\pm i\eta)$ are uniquely determined by the requirements that

$$L+1\pm i\eta\neq 0,\quad -\pi<\arg(L+1\pm i\eta)<\pi, \tag{2.3a}$$

$$\ln\Gamma(L+1\pm i\eta)\quad \text{is real when } L+1\pm i\eta \text{ is positive.} \tag{2.3b}$$

When L is *not* a negative number $\leqslant -1$, we obtain from (2.2),

$$\sigma_L(0)=0,\quad \text{except when } L \text{ is real and } \leqslant -1. \tag{2.4a}$$

When L is a real number <-1 but *not* an integer, we realize that

$$\lim_{\eta\to 0}\ln\Gamma(L+1\pm i\eta)-\ln|\Gamma(L+1)|=\mp[-L]\pi i,\quad \text{when } \operatorname{Re}\eta>0, \tag{2.5a}$$

$$\lim_{\eta\to 0}\ln\Gamma(L+1\pm i\eta)-\ln|\Gamma(L+1)|=\pm[-L]\pi i,\quad \text{when } \operatorname{Re}\eta<0, \tag{2.5b}$$

where $[-L]$ denotes the integer part of $-L$, and hence we obtain from (2.2),

$$\lim_{\eta\to 0}\sigma_L(\eta)=\begin{cases}-[-L]\pi, & \text{when Re }\eta>0,\\ +[-L]\pi, & \text{when Re }\eta<0.\end{cases}$$

$$\text{when } L \text{ is a negative number } <-1 \text{ but } \textit{not} \text{ an integer.} \tag{2.4b}$$

When L is a negative integer, and η is sufficiently small, we have the approximate formula

$$\Gamma(L+1\pm i\eta)=\frac{(-1)^{-L-1}+O(\eta)}{(-L-1)!(\pm i\eta)}, \tag{2.6}$$

with the aid of which we obtain

$$\lim_{\eta\to 0}\{\ln\Gamma(L+1\pm i\eta)+\ln[(-L-1)!|\eta|]\}$$
$$=i[\pm(L+\tfrac{1}{2})\pi+\pi/2-\arg(i\eta)],\quad \text{when } 0<\arg(i\eta)<\pi, \tag{2.7a}$$

$$\lim_{\eta\to 0}\{\ln\Gamma(L+1\pm i\eta)+\ln[(-L-1)!|\eta|]\}$$
$$=i[\mp(L+\tfrac{1}{2})\pi-\pi/2-\arg(i\eta)],\quad \text{when } -\pi<\arg(i\eta)<0, \tag{2.7b}$$

and hence we obtain from (2.2),

$$\lim_{\eta\to 0}\sigma_L(\eta)=\begin{cases}+(L+\tfrac{1}{2})\pi, & \text{when Re }\eta>0,\\ -(L+\tfrac{1}{2})\pi, & \text{when Re }\eta<0,\end{cases}\quad L=\text{negative integer.} \tag{2.4c}$$

Since the functions $\ln\Gamma(L+1\pm i\eta)$ are uniquely defined when the conditions (2.3a) and (2.3b) are fulfilled, one obtains from (2.2) the formula

$$\sigma_L(-\eta)=-\sigma_L(\eta), \tag{2.8}$$

which is consistent with (2.4a)–(2.4c). From (2.2), it also follows that

$$\exp[\pm i\sigma_L(\eta)]=\left(\frac{\exp[\frac{1}{2}\ln\Gamma(L+1+i\eta)]}{\exp[\frac{1}{2}\ln\Gamma(L+1-i\eta)]}\right)^{\pm 1}=\left(\frac{\Gamma(L+1+i\eta)}{\Gamma(L+1-i\eta)}\right)^{\pm 1/2}, \tag{2.9}$$

$$\frac{\exp[+i\sigma_L(\eta)]}{\Gamma(L+1+i\eta)}=\frac{\exp[-i\sigma_L(\eta)]}{\Gamma(L+1-i\eta)}=\exp[-\tfrac{1}{2}\ln\Gamma(L+1+i\eta)]\exp[-\tfrac{1}{2}\ln\Gamma(L+1-i\eta)]$$
$$=[\Gamma(L+1+i\eta)\Gamma(L+1-i\eta)]^{-1/2}, \tag{2.10}$$

the expressions in the last members of (2.9) and (2.10) being only shorthand notations for the precise expressions in the penultimate members.

We shall now investigate the behavior of $\sigma_L(\eta)$ when the quantities $L+1\pm i\eta$ are both large. Replacing (2.3a) by the more restrictive requirements that

$$|L+1\pm i\eta|\gg 1,\quad -\pi+\epsilon\leqslant\arg(L+1\pm i\eta)\leqslant\pi-\epsilon, \tag{2.3a$'$}$$

where ϵ is a fixed, arbitrarily small, positive number, we have the asymptotic expansion given by Eq. (5) on p. 32 in Ref. 11, i.e., with our notation,

$$\ln\Gamma(L+1\pm i\eta)\sim(L+\tfrac{1}{2}\pm i\eta)\ln(L+\tfrac{1}{2}\pm i\eta)-(L+\tfrac{1}{2}\pm i\eta)+\ln(2\pi)^{1/2}$$
$$-\frac{1}{24(L+\frac{1}{2}\pm i\eta)}+\cdots,\quad |L+\tfrac{1}{2}\pm i\eta|\gg 1, \tag{2.11}$$

where, for defining the appropriate branch of $\ln(L+\frac{1}{2}\pm i\eta)$, one has to choose $\ln(L+\frac{1}{2}\pm i\eta)$ to be close to $\ln(L+1\pm i\eta)$ when $L+\frac{1}{2}\pm i\eta$ is large. Using (2.11), we obtain from (2.2) the formula

$$\sigma_L(\eta)\sim\frac{1}{2i}[(L+\tfrac{1}{2}+i\eta)\ln(L+\tfrac{1}{2}+i\eta)-(L+\tfrac{1}{2}-i\eta)\ln(L+\tfrac{1}{2}-i\eta)]-\eta,$$
$$|L+\tfrac{1}{2}\pm i\eta|\gg 1. \tag{2.12}$$

In the particular situation when L and η are real, we can write (2.12) as

$$\sigma_L(\eta)\sim(L+\tfrac{1}{2})\arctan\frac{\eta}{L+\frac{1}{2}}+\eta\ln[(L+\tfrac{1}{2})^2+\eta^2]^{1/2}-\eta,$$
$$L \text{ and } \eta \text{ are real, } L+\tfrac{1}{2}\pm i\eta|\gg 1. \tag{2.13}$$

Using Eq. (34b) in Sec. 2 of Ref. 12 and (2.1a) and (2.1b) in the present paper, one finds that

$$\psi_+(L,\eta,\rho)\frac{d}{d\rho}\psi_-(L,\eta,\rho)-\psi_-(L,\eta,\rho)\frac{d}{d\rho}\psi_+(L,\eta,\rho)=-2i. \tag{2.14}$$

From (1.2) and (2.14) we get

$$G_L(\eta,\rho)\frac{d}{d\rho}F_L(\eta,\rho)-F_L(\eta,\rho)\frac{d}{d\rho}G_L(\eta,\rho)=1, \tag{2.15}$$

and hence

$$\frac{d}{d\rho}\frac{G_L(\eta,\rho)}{F_L(\eta,\rho)}=-\frac{1}{[F_L(\eta,\rho)]^2}. \tag{2.16}$$

From this formula, it follows that

$$\frac{G_L(\eta,\rho)}{F_L(\eta,\rho)}=-\int_{\rho_0}^{\rho}\frac{d\rho}{[F_L(\eta,\rho)]^2}, \tag{2.17}$$

where, however, the constant lower limit of integration ρ_0 is unknown.

In the next section we shall relate the Coulomb wave functions with the parameters $-L-1$ and η to the Coulomb wave functions with the parameters L and η. To that purpose we also need the phase shift $\sigma_{-L-1}(\eta)$ for the unique definition of which we, in analogy to (2.3a) and (2.3b), require that

$$-L\pm i\eta\neq 0,\quad -\pi<\arg(-L\pm i\eta)<\pi, \tag{2.18a}$$

$$\ln\Gamma(-L\pm i\eta)\quad \text{is real when} -L\pm i\eta \text{ is positive.} \tag{2.18b}$$

Throughout the next section we shall thus assume that the conditions (2.3a), (2.3b) and (2.18a), (2.18b) are fulfilled simultaneously. As we shall now show, the requirements (2.3a) and (2.18a) can be reformulated in a more concrete way.

Since $|(L+1+i\eta)-(-L-i\eta)|=2|L+\frac{1}{2}+i\eta|$ and $[(L+1+i\eta)+(-L-i\eta)]/2=\frac{1}{2}$, the distance in a complex plane between the points $L+1+i\eta$ and $-L-i\eta$ is $2|L+\frac{1}{2}+i\eta|$, and the middle point of the straight line segment joining these two points is the point $\frac{1}{2}$. When $|L+\frac{1}{2}+i\eta|<\frac{1}{2}$, the conditions (2.3a) with the upper signs, as well as the conditions (2.18a) with the lower signs, are thus automatically fulfilled. When $|L+\frac{1}{2}+i\eta|\geqslant\frac{1}{2}$ these conditions are fulfilled if either $-\pi<\arg(L+1+i\eta)<0$ or $0<\arg(L+1+i\eta)<\pi$. By replacing η by $-\eta$ we obtain similar conditions under which (2.3a) with the lower signs and (2.18a) with the upper signs are fulfilled.

III. REFLECTION FORMULAS ($L\to-L-1$)

Throughout this section we shall assume that the conditions stated at the end of the previous section, implying the simultaneous validity of (2.3a), (2.3b) and (2.18a), (2.18b), are fulfilled. Then the Coulomb phase shifts $\sigma_L(\eta)$ and $\sigma_{-L-1}(\eta)$ are both defined.

Replacing in (2.1a) and (2.1b) L by $-L-1$, we obtain

$$\psi_+(-L-1,\eta,\rho)=\exp\{i[\sigma_{-L-1}(\eta)+(L+1-i\eta)\pi/2]\}W_{-i\eta,-L-1/2}(2\rho e^{-i\pi/2}), \quad (3.1a)$$

$$\psi_-(-L-1,\eta,\rho)=\exp\{-i[\sigma_{-L-1}(\eta)+(L+1+i\eta)\pi/2]\}W_{i\eta,-L-1/2}(2\rho e^{+i\pi/2}). \quad (3.1b)$$

Since, according to Eq. (18a) in Sec. 2 of Ref. 12, the Whittaker function $W_{\kappa,\mu/2}(z)$ is an even function of the parameter μ, we obtain from (2.1a) and (3.1a),

$$\psi_+(-L-1,\eta,\rho)=\exp[ix(L,\eta)]\psi_+(L,\eta,\rho), \quad (3.2a)$$

and from (2.1b) and (3.1b),

$$\psi_-(-L-1,\eta,\rho)=\exp[-ix(L,\eta)]\psi_-(L,\eta,\rho), \quad (3.2b)$$

where the phase $x(L,\eta)$ is defined as

$$x(L,\eta)=(L+\tfrac{1}{2})\pi+\sigma_{-L-1}(\eta)-\sigma_L(\eta), \quad (3.3)$$

and thus has the property that

$$x(-L-1,\eta)=-x(L,\eta). \quad (3.4)$$

From (1.2) (3.2a), and (3.2b), we obtain

$$F_{-L-1}(\eta,\rho)=\cos x(L,\eta)F_L(\eta,\rho)+\sin x(L,\eta)G_L(\eta,\rho), \quad (3.5a)$$

$$G_{-L-1}(\eta,\rho)=-\sin x(L,\eta)F_L(\eta,\rho)+\cos x(L,\eta)G_L(\eta,\rho). \quad (3.5b)$$

Next, we shall derive formulas for $x(L,\eta)$ in the limit when $\eta\to0$. We remark in advance that the resulting formulas (3.6a)–(3.6d) fulfill (3.4).

When L is *not* a real number $\leqslant-1$ or $\geqslant0$, we obtain from (3.3) with the aid of (2.4a) and the corresponding formula with L replaced by $-L-1$,

$$x(L,0)=(L+\tfrac{1}{2})\pi, \quad \text{except when } L \text{ is real and } \leqslant-1 \text{ or } \geqslant0. \quad (3.6a)$$

When L is a real number <-1 but *not* an integer, we have $\sigma_{-L-1}(0)=0$ according to (2.4a) with L replaced by $-L-1$, and we obtain $\lim_{\eta\to0}\sigma_L(\eta)$ from (2.4b). From (3.3), we then obtain

$$\lim_{\eta\to0}x(L,\eta)=\begin{cases}(L+\frac{1}{2}+[-L])\pi, & \text{when Re }\eta>0,\\ (L+\frac{1}{2}-[-L])\pi, & \text{when Re }\eta<0,\end{cases}$$
$$\text{when } L \text{ is a real number } <-1 \text{ but } \textit{not} \text{ an integer}. \quad (3.6b)$$

When L is a real number >0 but *not* an integer, we have $\sigma_L(0)=0$ according to (2.4a), and we obtain $\lim_{\eta\to0}\sigma_{-L-1}(\eta)$ from (2.4b) with L replaced by $-L-1$. From (3.3) we then obtain

$$\lim_{\eta\to 0} x(L,\eta)=\begin{cases}(L+\frac{1}{2}-[L+1])\pi, & \text{when Re }\eta>0,\\ (L+\frac{1}{2}+[L+1])\pi, & \text{when Re }\eta<0,\end{cases}$$

$$\text{when } L \text{ is a real number } >0 \text{ but } \textit{not} \text{ an integer.} \tag{3.6c}$$

When L is an integer (positive, negative, or zero), we use (2.4a), (2.4c), and also these formulas with L replaced by $-L-1$, to obtain from (3.3),

$$\lim_{\eta\to 0} x(L,\eta)=\begin{cases}0, & \text{when Re }\eta>0,\\ (2L+1)\pi, & \text{when Re }\eta<0,\end{cases}\quad L=\text{integer.} \tag{3.6d}$$

Replacing in (2.2) L by $-L-1$, we get

$$\sigma_{-L-1}(\eta)=\frac{1}{2i}[\ln\Gamma(-L+i\eta)-\ln\Gamma(-L-i\eta)], \tag{3.7}$$

where the branches of the logarithms are to be chosen such that $\ln\Gamma(-L+i\eta)$ is a real number when $-L+i\eta$ is a positive number, and $\ln\Gamma(-L-i\eta)$ is a real number when $-L-i\eta$ is a positive number. With the aid of (2.2) and (3.7), we obtain

$$\exp\{2i[\sigma_{-L-1}(\eta)-\sigma_L(\eta)]\}=\frac{\Gamma(-L+i\eta)\Gamma(L+1-i\eta)}{\Gamma(-L-i\eta)\Gamma(L+1+i\eta)}, \tag{3.8}$$

and hence with the use of the reflection formula for the gamma function,

$$\exp\{2i[\sigma_{-L-1}(\eta)-\sigma_L(\eta)]\}=\frac{\sin[\pi(L+1+i\eta)]}{\sin[\pi(L+1-i\eta)]}. \tag{3.9}$$

Using (3.9), we obtain from (3.3),

$$\exp[2ix(L,\eta)]=\exp[(2L+1)\pi i]\frac{\sin[\pi(L+1+i\eta)]}{\sin[\pi(L+1-i\eta)]}, \tag{3.10}$$

and thus

$$x(L,\eta)=(L+\tfrac{1}{2})\pi+\frac{1}{2i}\ln\frac{\sin[\pi(L+1+i\eta)]}{\sin[\pi(L+1-i\eta)]}, \tag{3.11}$$

where according to (3.6a) the branch of the logarithm is to be chosen such that the logarithm is equal to zero when $\eta=0$ and L is a real number between -1 and 0. We can also write (3.10) as

$$\exp[2ix(L,\eta)]=\frac{1-\exp[+2\pi i(L+i\eta)]}{1-\exp[-2\pi i(L-i\eta)]}. \tag{3.12}$$

By partial logarithmic differentiation of (3.12) with respect to L, we obtain the formula

$$\begin{aligned}\frac{\partial x(L,\eta)}{\partial L}&=\frac{2\pi[1-\cos(2\pi L)\exp(2\pi\eta)]}{1-2\cos(2\pi L)\exp(2\pi\eta)+\exp(4\pi\eta)}\\ &=\frac{2\pi[1-\cos(2\pi L)\exp(2\pi\eta)]}{[1-\cos(2\pi L)\exp(2\pi\eta)]^2+\sin^2(2\pi L)\exp(4\pi\eta)},\end{aligned} \tag{3.13a}$$

and by partial logarithmic differentiation of (3.12) with respect to η, we obtain the formula

$$\frac{\partial x(L,\eta)}{\partial \eta}=\frac{2\pi \sin(2\pi L)\exp(2\pi\eta)}{1-2\cos(2\pi L)\exp(2\pi\eta)+\exp(4\pi\eta)}$$
$$=\frac{2\pi \sin(2\pi L)\exp(2\pi\eta)}{[1-\cos(2\pi L)\exp(2\pi\eta)]^2+\sin^2(2\pi L)\exp(4\pi\eta)}. \tag{3.13b}$$

When $2L$ is an integer and Re $\eta\neq 0$, we obtain from (3.13a),

$$\frac{\partial x(L,\eta)}{\partial L}=\frac{2\pi}{1-(-1)^{2L}\exp(2\pi\eta)},\quad \text{Re }\eta\neq 0,\ 2L=\text{integer}, \tag{3.14}$$

and from (3.13b), $\partial x(L,\eta)/\partial\eta=0$, and hence, when arg η is kept constant,

$$x(L,\eta)=\lim_{\eta\to 0} x(L,\eta),\quad \text{Re }\eta\neq 0,\ 2L=\text{integer}. \tag{3.15}$$

From this formula it follows, with the use of (3.6a)–(3.6d), that

$$x(L,\eta)=\begin{cases}0 & \text{when Re }\eta>0,\\ (2L+1)\pi, & \text{when Re }\eta<0,\end{cases}\quad 2L=\text{integer}. \tag{3.16}$$

Using (3.16), we obtain from (3.2a) and (3.2b),

$$\psi_{\pm}(-L-1,\eta,\rho)=\begin{cases}\psi_{\pm}(L,\eta,\rho), & \text{when Re }\eta>0,\\ (-1)^{2L+1}\psi_{\pm}(L,\eta,\rho), & \text{when Re }\eta<0,\end{cases}\quad 2L=\text{integer}, \tag{3.17}$$

and from (3.5a) and (3.5b),

$$F_{-L-1}(\eta,\rho)=\begin{cases}F_L(\eta,\rho), & \text{when Re }\eta>0,\\ (-1)^{2L+1}F_L(\eta,\rho), & \text{when Re }\eta<0,\end{cases}\quad 2L=\text{integer}, \tag{3.18a}$$

$$G_{-L-1}(\eta,\rho)=\begin{cases}G_L(\eta,\rho), & \text{when Re }\eta>0,\\ (-1)^{2L+1}G_L(\eta,\rho), & \text{when Re }\eta<0,\end{cases}\quad 2L=\text{integer}. \tag{3.18b}$$

From (3.18a) and (3.18b) one obtains the formula

$$\frac{G_{-L-1}(\eta,\rho)}{F_{-L-1}(\eta,\rho)}=\frac{G_L(\eta,\rho)}{F_L(\eta,\rho)},\quad \text{when Re }\eta\neq 0,\quad 2L=\text{integer}, \tag{3.19}$$

which can also be obtained from (3.5a), (3.5b) and (3.16).

IV. INTEGRAL REPRESENTATIONS

According to Eq. (5) in Sec. 5 of Ref. 12, we have for the Whittaker function the integral representation

$$W_{\kappa,\mu/2}(z)=\frac{z^{\kappa}e^{-z/2}}{\Gamma((\mu+1)/2-\kappa)}\int_0^{\infty}\exp(-w)w^{(\mu-1)/2-\kappa}\left(1+\frac{w}{z}\right)^{(\mu-1)/2+\kappa}dw,$$
$$z\neq 0,\quad -\pi<\arg z<\pi, \tag{4.1}$$

which is valid when $\text{Re}[(\mu+1)/2-\kappa]>0$; otherwise the integral does not converge at $w=0$. The integration in (4.1) is to be performed along the positive part of the real w axis with $\arg w=0$ and $\arg(1+w/z)\to -\arg z$ as $w\to\infty$.

With the aid of (4.1) we obtain from (2.1a),

$$\psi_+(L,\eta,\rho)=\frac{\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}}{\Gamma(L+1+i\eta)}$$
$$\times\int_0^\infty \exp(-w)w^{L+i\eta}\left(1+\frac{we^{+i\pi/2}}{2\rho}\right)^{L-i\eta}dw,$$
$$\rho\neq 0,\quad -\pi/2<\arg\rho<3\pi/2,\quad \mathrm{Re}(L+1+i\eta)>0, \tag{4.2a}$$

and from (2.1b),

$$\psi_-(L,\eta,\rho)=\frac{\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}}{\Gamma(L+1-i\eta)}$$
$$\times\int_0^\infty \exp(-w)w^{L-i\eta}\left(1+\frac{we^{-i\pi/2}}{2\rho}\right)^{L+i\eta}dw,$$
$$\rho\neq 0,\quad -3\pi/2<\arg\rho<\pi/2,\quad \mathrm{Re}(L+1-i\eta)>0. \tag{4.2b}$$

The integrations in (4.2a) and (4.2b) are to be performed along the positive part of the real w axis with $\arg w=0$ and $\arg[1+we^{i\pi/2}/(2\rho)]\to\pi/2-\arg\rho$ as $w\to\infty$ in (4.2a), but $\arg[1+we^{-i\pi/2}/(2\rho)]\to-\pi/2-\arg\rho$ as $w\to\infty$ in (4.2b). For the sake of clarity, we remark that $\mathrm{Re}(L+1\pm i\eta)>0$ implies that the conditions (2.3a) are fulfilled. Referring to (2.2), we emphasize that always when the Coulomb phase shift $\sigma_L(\eta)$ or a Coulomb wave function (including ψ_+ and ψ_-) with the parameters L and η appears, it is tacitly assumed that the conditions (2.3a) and (2.3b) are fulfilled. Therefore we have, in general, not found it necessary to repeat these conditions, and also, in the following, we shall assume that these conditions (even if they are not stated explicitly) are fulfilled for formulas involving $\sigma_L(\eta)$ or Coulomb wave functions with the parameters L and η.

When $L+i\eta$ and $L-i\eta$, respectively, is not an integer, the integrand in (4.2a) and (4.2b), respectively, has a branch point at $w=0$, and we can rewrite the integral by using a contour that encircles the point $w=0$. Thus, if we introduce a contour of integration Γ, which will be defined below, and use the reflection formula for the gamma function, we can rewrite (4.2a) as

$$\psi_+(L,\eta,\rho)=-\frac{1}{2\pi i}\exp\{+i[\rho-\eta\ln(2\rho)-3L\pi/2-i\eta\pi+\sigma_L(\eta)]\}\Gamma(-L-i\eta)$$
$$\times\int_\Gamma \exp(-w)w^{L+i\eta}\left(1+\frac{we^{+i\pi/2}}{2\rho}\right)^{L-i\eta}dw,$$
$$\rho\neq 0,\quad -\pi/2<\arg\rho<3\pi/2,\quad L+i\eta\neq \text{integer}, \tag{4.3a}$$

and (4.2b) as

$$\psi_-(L,\eta,\rho)=-\frac{1}{2\pi i}\exp\{-i[\rho-\eta\ln(2\rho)+L\pi/2-i\eta\pi+\sigma_L(\eta)]\}\Gamma(-L+i\eta)$$
$$\times\int_\Gamma \exp(-w)w^{L-i\eta}\left(1+\frac{we^{-i\pi/2}}{2\rho}\right)^{L+i\eta}dw,$$
$$\rho\neq 0,\quad -3\pi/2<\arg\rho<\pi/2,\quad L-i\eta\neq \text{integer}, \tag{4.3b}$$

where Γ is a contour of integration that begins at infinity with $\arg w=0$, encircles the origin in the positive sense without encircling the point $2i\rho$ for (4.3a) and point $-2i\rho$ for (4.3b), and ends at infinity with $\arg w=2\pi$; see Fig. 1. The conditions $\mathrm{Re}(L+1\pm i\eta)>0$ in (4.2a) and (4.2b) do not apply to (4.3a) and (4.3b). With the aid of the reflection formula for the gamma function and (2.10), the formulas (4.2a), (4.2b) and (4.3a), (4.3b) can be rewritten in alternative ways.

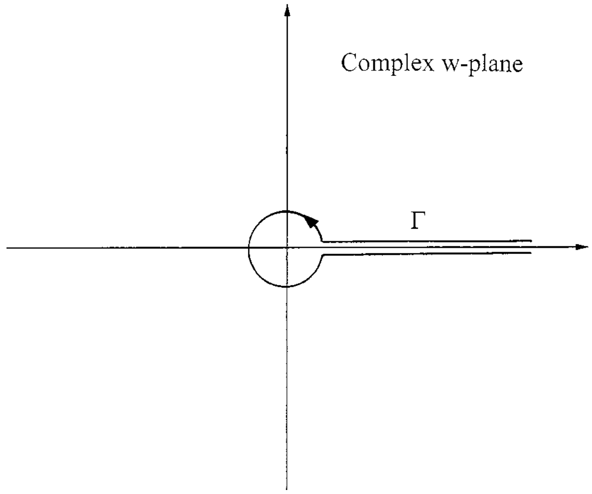

FIG. 1. The contour Γ, which must not encircle the points $+2i\rho$ and $-2i\rho$.

Introducing instead of w a new integration variable s by putting in (4.2a),

$$w=e^{i\pi/2}\rho(1+s),\quad \rho\neq 0,\quad \arg(1+s)=-\pi/2-\arg\rho$$

$$1+we^{i\pi/2}/(2\rho)=(1-s)/2,\quad \arg(1-s)\rightarrow\pi/2-\arg\rho\quad \text{as}\ \ s\rightarrow\infty, \tag{4.4a}$$

and in (4.2b),

$$w=e^{-i\pi/2}\rho(1-s),\quad \rho\neq 0,\quad \arg(1-s)=\pi/2-\arg\rho,$$

$$1+we^{-i\pi/2}/(2\rho)=(1+s)/2,\quad \arg(1+s)\rightarrow-\pi/2-\arg\rho\quad \text{as}\ \ s\rightarrow\infty, \tag{4.4b}$$

we get

$$\psi_{+}(L,\eta,\rho)=\frac{2i\exp[i\sigma_L(\eta)-\eta\pi/2]}{\Gamma(L+1+i\eta)}(\rho/2)^{L+i}$$
$$\times\int_{-1}^{+\infty e^{-i\pi/2}/\rho}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds,$$
$$\rho\neq 0,\quad -\pi/2<\arg\rho<3\pi/2,\quad \mathrm{Re}(L+1+i\eta)>0, \tag{4.5a}$$

and, when (2.10) is also used,

$$\psi_{-}(L,\eta,\rho)=\frac{2i\exp[i\sigma_L(\eta)-\eta\pi/2]}{\Gamma(L+1+i\eta)}(\rho/2)^{L+1}$$
$$\times\int_{1}^{+\infty e^{-i\pi/2}/\rho}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds,$$
$$\rho\neq 0,\quad -3\pi/2<\arg\rho<\pi/2,\quad \mathrm{Re}(L+1-i\eta)>0. \tag{4.5b}$$

Using (1.2), (4.5a), and (4.5b), we obtain

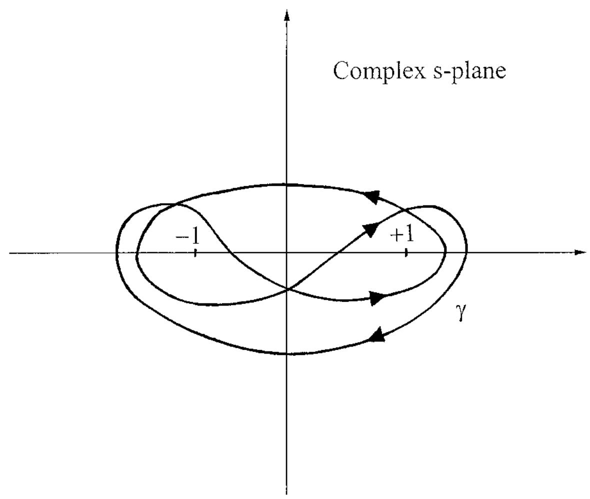

FIG. 2. The contour γ.

$$F_L(\eta,\rho)=\frac{\exp[i\sigma_L(\eta)-\eta\pi/2]}{\Gamma(L+1+i\eta)}(\rho/2)^{L+1}\int_{-1}^{+1}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds,$$

$$\operatorname{Re}(L+1+i\eta)>0,\quad \operatorname{Re}(L+1-i\eta)>0. \tag{4.6}$$

Although (4.6) has been obtained from (4.5a) and (4.5b), which are simultaneously valid only when $\rho\neq0$ and $-\pi/2<\arg\rho<\pi/2$, one realizes by means of analytic continuation that (4.6) is valid for any value of $\arg\rho$ (when $\rho\neq0$) and also for $\rho=0$. With the aid of (2.10), the formula (4.6) can be written in alternative ways.

Assuming that $L+i\eta\neq$ integer and $L-i\eta\neq$ integer, so that the integrand in (4.6) has branch points at $s=-1$ and $s=+1$, we let γ be a contour of integration that starts at a point on the real s axis between $s=-1$ and $s=+1$, where $\arg(1+s)=\arg(1-s)=0$, encircles successively, on a Riemann surface, $s=+1$ in the positive sense, $s=-1$ in the positive sense, $s=+1$ in the negative sense, $s=-1$ in the negative sense, and returns to the starting point; see Fig. 2. When we make this contour as small as possible, it coincides with the real s axis between $s=-1$ and $s=+1$, except for infinitesimally small circles around $s=-1$ and $s=+1$, and then we find that

$$\begin{aligned}
&\int_\gamma \exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds\\
&=\int_{-1}^{+1}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds\\
&+\exp[2\pi i(L-i\eta)]\int_{+1}^{-1}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds\\
&+\exp[2\pi i(L+i\eta)+2\pi i(L-i\eta)]\int_{-1}^{+1}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds\\
&+\exp[2\pi i(L+i\eta)]\int_{+1}^{-1}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds\\
&=\{1-\exp[2\pi i(L+i\eta)]\}\{1-\exp[2\pi i(L-i\eta)]\}\\
&\quad\times\int_{-1}^{+1}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds,
\end{aligned}$$

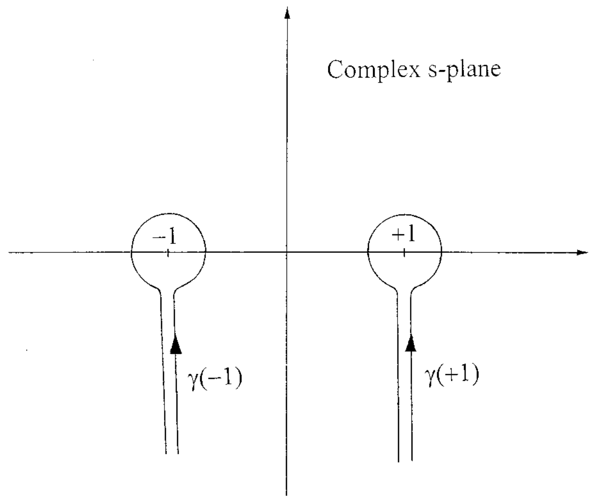

FIG. 3. The contours $\gamma(-1)$ and $\gamma(+1)$, drawn for the special case when arg $\rho=0$.

where $\arg(1+s)=\arg(1-s)=0$ in each integral with the limits of integration -1 and $+1$. Using this formula, we can write (4.6) as

$$F_L(\eta,\rho)=\frac{\exp[i\sigma_L(\eta)-\eta\pi/2]}{\Gamma(L+1+i\eta)\{1-\exp[2\pi i(L+i\eta)]\}\{1-\exp[2\pi i(L-i\eta)]\}}\times(\rho/2)^{L+1}\int_\gamma \exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds,$$

$$L+i\eta\neq\text{integer},\quad L-i\eta\neq\text{integer}. \tag{4.7}$$

Note that although the derivation of (4.7) was made under the conditions $\mathrm{Re}(L+1+i\eta)>0$, $\mathrm{Re}(L+1-i\eta)>0$ and $L+i\eta\neq$ integer, $L-i\eta\neq$ integer (the last two conditions ensure that the contour of integration γ cannot be shrunk to a point), we can now disregard the first two conditions. The formula (4.7) is still valid on account of analytic continuation.

Introducing in (4.3a) instead of w a new integration variable s by putting in analogy to (4.4a),

$$w=e^{i\pi/2}\rho(1+s),\quad \rho\neq 0,\quad \arg(1+s)=-\pi/2-\arg\rho+\arg w,$$

$$1+we^{i\pi/2}/(2\rho)=(1-s)/2,$$

$$\arg(1-s)\to\pi/2-\arg\rho \text{ at infinity, where the integration starts and stops,} \tag{4.8a}$$

we get

$$\psi_+(L,\eta,\rho)=-\frac{1}{\pi}\exp[i\sigma_L(\eta)-\eta\pi/2-i(L+i\eta)\pi]\Gamma(-L-i\eta)(\rho/2)^{L+1}\times\int_{\gamma(-1)}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds,$$

$$\rho\neq 0,\quad -\pi/2<\arg\rho<3\pi/2,\quad L+i\eta\neq\text{integer}, \tag{4.9a}$$

where $\gamma(-1)$ is a contour of integration that starts at infinity with $\arg(1+s)=-\pi/2-\arg\rho$ and $\arg(1-s)=\pi/2-\arg\rho$, encircles the point $s=-1$ in the positive sense without encircling the point $s=+1$, and returns to infinity with $\arg(1+s)=3\pi/2-\arg\rho$ and $\arg(1-s)=\pi/2-\arg\rho$; see Fig. 3. Recalling (4.8a) and the restrictions on $\arg\rho$ in (4.9a), we see that $-2\pi<\arg(1+s)<0$ and $-\pi<\arg(1-s)<\pi$ at the point at infinity where the integration starts, and $0<\arg(1+s)<2\pi$ and $-\pi<\arg(1-s)<\pi$ at the point at infinity where the integration stops.

Introducing in (4.3b) instead of w a new integration variable s by putting in analogy to (4.4b),

$$w=e^{-i\pi/2}\rho(1-s),\quad \rho\neq 0,\quad \arg(1-s)=\pi/2-\arg\rho+\arg w,$$

$$1+we^{-i\pi/2}/(2\rho)=(1+s)/2,$$

$$\arg(1+s)\rightarrow-\pi/2-\arg\rho \text{ at infinity, where the integration starts and stops,} \tag{4.8b}$$

we get

$$\psi_-(L,\eta,\rho)=-\frac{1}{\pi}\exp[-i\sigma_L(\eta)-\eta\pi/2-i(L-i\eta)\pi]\Gamma(-L+i\eta)(\rho/2)^{L+1}$$
$$\times\int_{\gamma(+1)}\exp(-i\rho s)(1+s)^{L+i\eta}(1-s)^{L-i\eta}\,ds,$$
$$\rho\neq 0,\quad -3\pi/2<\arg\rho<\pi/2,\quad L-i\eta\neq\text{integer}, \tag{4.9b}$$

where $\gamma(+1)$ is a contour of integration that starts at infinity with $\arg(1-s)=\pi/2-\arg\rho$ and $\arg(1+s)=-\pi/2-\arg\rho$, encircles the point $s=+1$ in the positive sense without encircling the point $s=-1$, and returns to infinity with $\arg(1-s)=5\pi/2-\arg\rho$ and $\arg(1+s)=-\pi/2-\arg\rho$; see Fig. 3. Recalling (4.8b) and the restrictions on $\arg\rho$ in (4.9b), we see that $0<\arg(1-s)<2\pi$ and $-\pi<\arg(1+s)<\pi$ at the point at infinity where the integration starts, and $2\pi<\arg(1-s)<4\pi$ and $-\pi<\arg(1+s)<\pi$ at the point at infinity where the integration stops.

The formula (4.7) can also be derived from (4.9a) and (4.9b), but that derivation is much more complicated than the derivation given previously in the present paper.

V. POWER SERIES FOR $F_L(\eta,\rho)$

Expanding in (4.6) the exponential $\exp(-i\rho s)$ in a power series and substituting $s=\pm(2u-1)$, where either the upper or the lower sign can be used, we obtain

$$F_L(\eta,\rho)=\frac{\exp[i\sigma_L(\eta)-\eta\pi/2]}{2\Gamma(L+1+i\eta)}$$
$$\times(2\rho)^{L+1}\sum_{k=0}^{\infty}\frac{(\pm i\rho)^k}{k!}\sum_{m=0}^{k}\binom{k}{m}(-2)^m\int_0^1 u^{m+L\pm i\eta}(1-u)^{L\mp i\eta}\,du,$$
$$\text{Re}(L+1+i\eta)>0,\quad \text{Re}(L+1-i\eta)>0, \tag{5.1}$$

where the integration is performed along the real u axis with $\arg u=\arg(1-u)=0$. In a similar way, as (4.7) was obtained from (4.6) we obtain from (5.1),

$$F_L(\eta,\rho)=\frac{\exp[i\sigma_L(\eta)-\eta\pi/2]}{2\Gamma(L+1+i\eta)\{1-\exp[2\pi i(L+i\eta)]\}\{1-\exp[2\pi i(L-i\eta)]\}}$$
$$\times(2\rho)^{L+1}\sum_{k=0}^{\infty}\frac{(\pm i\rho)^k}{k!}\sum_{m=0}^{k}\binom{k}{m}(-2)^m\int u^{m+L\pm i\eta}(1-u)^{L\mp i\eta}\,du,$$

$$L+i\eta\neq\text{integer},\quad L-i\eta\neq\text{integer},\tag{5.2}$$

where the integration is performed along a contour that starts at a point on the real u axis between $u=0$ and $u=+1$, where $\arg u=\arg(1-u)=0$, encircles successively, on a Riemann surface, $u=+1$ in the positive sense, $u=0$ in the positive sense, $u=+1$ in the negative sense, $u=0$ in the negative sense, and returns to the starting point. This contour of integration is thus analogous to the contour of integration γ in Fig. 2. Recalling the definition of the beta function and the expression for the beta function in terms of gamma functions, we have

$$\int_0^1 u^{m+L\pm i\eta}(1-u)^{L\mp i\eta}\,du=B(m+L+1\pm i\eta,L+1\mp i\eta)$$
$$=\frac{\Gamma(m+L+1\pm i\eta)\Gamma(L+1\mp i\eta)}{\Gamma(m+2L+2)},$$
$$\mathrm{Re}(L+1+i\eta)>0,\quad \mathrm{Re}(L+1-i\eta)>0,\tag{5.3}$$

where the integration is performed along the real u axis, and

$$\frac{\int u^{m+L\pm i\eta}(1-u)^{L\mp i\eta}\,du}{\{1-\exp[2\pi i(L+i\eta)]\}\{1-\exp[2\pi i(L-i\eta)]\}}$$
$$=B(m+L+1\pm i\eta,L+1\mp i\eta)=\frac{\Gamma(m+L+1\pm i\eta)\Gamma(L+1\mp i\eta)}{\Gamma(m+2L+2)},$$
$$L+i\eta\neq\text{integer},\quad L-i\eta\neq\text{integer},\tag{5.4}$$

where the contour of integration is the same as in (5.2). From (5.1), (5.3), and (2.2), or from (5.2), (5.3), and (2.2), we obtain

$$F_L(\eta,\rho)=\exp[\mp i\sigma_L(\eta)-\eta\pi/2](2\rho)^{L+1}\sum_{k=0}^{\infty}\frac{(\pm i\rho)^k}{k!}\sum_{m=0}^{k}\binom{k}{m}(-2)^m\frac{\Gamma(m+L+1\pm i\eta)}{2\Gamma(m+2L+2)},\tag{5.5}$$

where one can use either the upper or the lower signs. Since it follows from (2.3a) that $L\pm i\eta$ must not be a negative integer, the gamma functions $\Gamma(m+L+1\pm i\eta)$ cannot have poles for allowed values of the parameters L and η. The conditions in (5.1), (5.3) and (5.2), (5.4), which are associated with the integral representations used, do not apply anymore. With the use of (2.9) or (2.10), we can write (5.5) as

$$F_L(\eta,\rho)=\frac{[\Gamma(L+1+i\eta)\Gamma(L+1-i\eta)]^{1/2}}{2\exp(\eta\pi/2)}(2\rho)^{L+1}\bar{\Phi}_L(\eta,\rho),\tag{5.6}$$

with

$$\bar{\Phi}_L(\eta,\rho)=\frac{1}{\Gamma(L+1\pm i\eta)}\sum_{k=0}^{\infty}\frac{(\pm i\rho)^k}{k!}\sum_{m=0}^{k}\binom{k}{m}(-2)^m\frac{\Gamma(m+L+1\pm i\eta)}{\Gamma(m+2L+2)}.\tag{5.7}$$

For the precise definition of the square root in (5.6) we refer to the explanation below (2.10). As regards (5.7), we note that $\Gamma(L+1\pm i\eta)$ cannot be equal to zero and is finite because of (2.3a). When L, η, and ρ are real, the function $\bar{\Phi}_L(\eta,\rho)$ is according to (5.7) given by either of two complex conjugate expressions, and hence $\bar{\Phi}_L(\eta,\rho)$ must in that case be real. Our formulas (5.6) and (5.7) are to be compared with formulas in subsections 14.1.4–14.1.7 in Ref. 10, which apply when L is a nonnegative integer and η is real.

When $2L+1$ is *not* a negative integer, the terms in (5.7) corresponding to $k\leqslant -2L-2 = 2|L|-2$ vanish. The first nonvanishing term in the series (5.7) is then proportional to $\rho^{2|L|-1}$.

When $2L+1$ is *not* a negative integer, $\Gamma(2L+2)$ is finite and different from zero, and the term corresponding to $k=0$ in (5.7) is finite and does not vanish. In this case it is convenient to rewrite (5.6) along with (5.7) as follows:

$$F_L(\eta,\rho)=\frac{[\Gamma(L+1+i\eta)\Gamma(L+1-i\eta)]^{1/2}}{2\exp(\eta\pi/2)\Gamma(2L+2)}(2\rho)^{L+1}\sum_{k=0}^{\infty} c_{L,k}(\eta)\rho^k,\quad c_{L,0}(\eta)=1,$$

$$2L+1\neq \text{negative integer}, \tag{5.8}$$

where

$$c_{L,k}(\eta)=\frac{(\pm i)^k}{k!}\sum_{m=0}^{k}\binom{k}{m}(-2)^m\frac{\Gamma(2L+2)\Gamma(m+L+1\pm i\eta)}{\Gamma(m+2L+2)\Gamma(L+1\pm i\eta)},\quad 2L+1\neq \text{negative integer}. \tag{5.9}$$

In (5.9) one can use either the upper or the lower signs, and thus it is seen that

$$c_{L,k}(\eta)=(-1)^k c_{L,k}(-\eta). \tag{5.10}$$

Inserting $\psi=F_L(\eta,\rho)$ with the expression (5.8) for $F_L(\eta,\rho)$ into (1.1), and recalling that $c_{L,0}(\eta)=1$, we obtain the formulas

$$c_{L,0}(\eta)=1, \tag{5.11a}$$

$$c_{L,1}(\eta)=\eta/(L+1), \tag{5.11b}$$

$$k(k+2L+1)c_{L,k}(\eta)=2\eta c_{L,k-1}(\eta)-c_{L,k-2}(\eta),\quad k\geqslant 2, \tag{5.11c}$$

which one can use instead of (5.9) to calculate the coefficients $c_{L,k}(\eta)$. When L and η are real, the coefficients $c_{L,k}(\eta)$ are according to (5.9) given by either of two complex conjugate expressions, and hence $c_{L,k}(\eta)$ must in that case be real. This also follows from (5.11a)–(5.11c). In the still more particular case when L is a non-negative integer and η is real, (5.8) along with (5.11a)–(5.11c) agrees with the power series for $F_L(\eta,\rho)$ given in subsections 14.1.4–14.1.7 in Ref. 10, our $c_{L,k}(\eta)$ being the same as their $A^L_{L+1+k}(\eta)$.

From (5.8), it follows that

$$\frac{\partial}{\partial L}\ln F_L(\eta,\rho)=\ln(2\rho)-\frac{2\Gamma'(2L+2)}{\Gamma(2L+2)}+\frac{\Gamma'(L+1+i\eta)}{2\Gamma(L+1+i\eta)}+\frac{\Gamma'(L+1-i\eta)}{2\Gamma(L+1-i\eta)}$$
$$+\frac{\Sigma_{k=0}^{\infty}(\partial c_{L,k}/\partial L)\rho^k}{\Sigma_{k=0}^{\infty}c_{L,k}\rho^k},\quad 2L+1\neq \text{negative integer}. \tag{5.12}$$

To obtain the coefficients $\partial c_{L,k}/\partial L$ in (5.12) we differentiate (5.11a)–(5.11c) partially with respect to L, getting

$$\frac{\partial c_{L,0}}{\partial L}=0, \tag{5.13a}$$

$$\frac{\partial c_{L,1}}{\partial L}=-\frac{\eta}{(L+1)^2}, \tag{5.13b}$$

$$k(k+2L+1)\frac{\partial c_{L,k}}{\partial L}=2\eta\frac{\partial c_{L,k-1}}{\partial L}-\frac{\partial c_{L,k-2}}{\partial L}-2kc_{L,k}, \quad k\geqslant 2. \tag{5.13c}$$

Of special interest, when $2L+1$ is not a negative integer, is the particular case $L=-\frac{1}{2}$. For $L=-\frac{1}{2}$ we obtain from (5.8), with the aid of the reflection formula for the gamma function,

$$F_{-1/2}(\eta,\rho)=\left(\frac{\pi}{\exp(2\pi\eta)+1}\right)^{1/2}\rho^{1/2}\sum_{k=0}^{\infty}c_{-1/2,k}\rho^k, \quad c_{-1/2,0}=1, \tag{5.14}$$

where the square root is positive for real values of η, and from (5.12),

$$\left(\frac{\partial}{\partial L}\ln F_L(\eta,\rho)\right)_{L=-1/2}=\ln(2\rho)+2\gamma+\frac{\Gamma'(\frac{1}{2}+i\eta)}{2\Gamma(\frac{1}{2}+i\eta)}+\frac{\Gamma'(\frac{1}{2}-i\eta)}{2\Gamma(\frac{1}{2}-i\eta)}+\frac{\Sigma_{k=0}^{\infty}(\partial c_{L,k}/\partial L)_{L=-1/2}\rho^k}{\Sigma_{k=0}^{\infty}c_{-1/2,k}\rho^k}, \tag{5.15}$$

where $\gamma=-\Gamma'(1)/\Gamma(1)=-\Gamma'(1)=0.5772...$ is Euler's constant. Defining

$$d_0=-\left(\ln 2+2\gamma+\frac{\Gamma'(\frac{1}{2}+i\eta)}{2\Gamma(\frac{1}{2}+i\eta)}+\frac{\Gamma'(\frac{1}{2}-i\eta)}{2\Gamma(\frac{1}{2}-i\eta)}\right), \tag{5.16}$$

and expanding [cf. (5.11a), (5.13a), and (5.13b)]

$$\frac{\Sigma_{k=0}^{\infty}(\partial c_{L,k}/\partial L)_{L=-1/2}\rho^k}{\Sigma_{k=0}^{\infty}c_{-1/2,k}\rho^k}=-\sum_{k=1}^{\infty}d_k\rho^k=-(4\eta\rho+\cdots), \tag{5.17}$$

we can write (5.15) as

$$\left(\frac{\partial}{\partial L}\ln F_L(\eta,\rho)\right)_{L=-1/2}=-\left(\ln(1/\rho)+\sum_{k=0}^{\infty}d_k\rho^k\right)=[\ln(1/\rho)+4\eta\rho+\cdots]. \tag{5.18}$$

VI. FORMULAS FOR $G_L(\eta,\rho)$

Solving the reflection formula (3.5a) with respect to $G_L(\eta,\rho)$, we obtain

$$G_L(\eta,\rho)=\frac{F_{-L-1}(\eta,\rho)-\cos x(L,\eta)F_L(\eta,\rho)}{\sin x(L,\eta)}. \tag{6.1}$$

This formula can be directly used only when $\sin x(L,\eta)\neq 0$, and from (3.12) it follows that $\sin x(L,\eta)\neq 0$ when and only when $2L$ is not an integer (positive, negative, or zero), while η is unspecified. When $2L$ is not an integer, one can thus use (6.1) directly, and by inserting into this formula (5.8) and the formula that one obtains from (5.8) by replacing L by $-L-1$, one obtains $G_L(\eta,\rho)$ as the sum of two series in which there is no term containing $\ln\rho$. Since the series (5.8) for $F_L(\eta,\rho)$ contains the factor $(2\rho)^{L+1}$, and the corresponding series for $F_{-L-1}(\eta,\rho)$ contains the factor $(2\rho)^{-L}$, and since $2L$ is *not* an integer, one obtains $G_L(\eta,\rho)$ as the sum of two series that cannot be combined into one series.

When $2L$ is an integer, the denominator of (6.1) is equal to zero according to (3.16), and the numerator of (6.1) is equal to zero according to (3.16) and (3.18a). To use (6.1) in such a particular case, one applies l'Hospital's rule, getting, since according to (3.16) the phase $x(L,\eta)$ is an integer multiple of π,

$$G_L(\eta,\rho)=\frac{\dfrac{\partial F_{-L-1}(\eta,\rho)}{\partial L}-\cos x(L,\eta)\,\dfrac{\partial F_L(\eta,\rho)}{\partial L}}{\cos x(L,\eta)\,\dfrac{\partial x(L,\eta)}{\partial L}},\quad 2L=\text{integer}. \tag{6.2}$$

With the use of (3.14) and (3.16) we obtain from (6.2),

$$G_L(\eta,\rho)=\frac{(-1)^{2L}\exp(2\pi\eta)-1}{2\pi}\times\begin{cases}\left(\dfrac{\partial F_L(\eta,\rho)}{\partial L}+\dfrac{\partial F_{-L-1}(\eta,\rho)}{\partial(-L-1)}\right), & \text{when Re }\eta>0,\\ \left(\dfrac{\partial F_L(\eta,\rho)}{\partial L}+(-1)^{2L+1}\dfrac{\partial F_{-L-1}(\eta,\rho)}{\partial(-L-1)}\right), & \text{when Re }\eta<0,\end{cases}\quad 2L=\text{integer}. \tag{6.3}$$

From (6.3) we obtain with the use of (3.18a) the formula

$$\frac{G_L(\eta,\rho)}{F_L(\eta,\rho)}=\frac{(-1)^{2L}\exp(2\pi\eta)-1}{2\pi}\left(\frac{\partial\ln F_L(\eta,\rho)}{\partial L}+\frac{\partial\ln F_{-L-1}(\eta,\rho)}{\partial(-L-1)}\right),$$

$$\text{Re }\eta\neq 0,\quad 2L=\text{integer}, \tag{6.4}$$

from which one easily obtains (3.19) in an alternative way.

When $2L=-1$ we obtain from (6.4),

$$\frac{G_{-1/2}(\eta,\rho)}{F_{-1/2}(\eta,\rho)}=-\frac{\exp(2\pi\eta)+1}{\pi}\left(\frac{\partial\ln F_L(\eta,\rho)}{\partial L}\right)_{L=-1/2}. \tag{6.5}$$

Inserting (5.18) into (6.5), we obtain

$$\frac{G_{-1/2}(\eta,\rho)}{F_{-1/2}(\eta,\rho)}=\frac{\exp(2\pi\eta)+1}{\pi}\left(\ln(1/\rho)+\sum_{k=0}^{\infty}d_k\rho^k\right)=\frac{\exp(2\pi\eta)+1}{\pi}[\ln(1/\rho)+4\eta\rho+\cdots]. \tag{6.6}$$

When $2L$ is an integer $\neq -1$ we first note that because of (3.19) it is sufficient to consider the case when $L\geqslant 0$. Then we write (6.4) as

$$\frac{G_L(\eta,\rho)}{F_L(\eta,\rho)}=\frac{(-1)^{2L}\exp(2\pi\eta)-1}{2\pi}\left\{2\ln\rho+\frac{\partial\ln[\rho^{-L-1}F_L(\eta,\rho)]}{\partial L}+\rho^{-(2L+1)}\left[\rho^{(2L+1)}\frac{\partial\ln[\rho^L F_{-L-1}(\eta,\rho)]}{\partial(-L-1)}\right]\right\}, \tag{6.7}$$

and point out that from (5.6) and (5.7) it follows that $\partial\ln[\rho^{-L-1}F_L(\eta,\rho)]/\partial L$ as well as $\rho^{(2L+1)}\partial\ln[\rho^L F_{-L-1}(\eta,\rho)]/\partial(-L-1)$ can be expanded in a power series of ρ with the constant term $\neq 0$. The formula thus obtained from (6.7) is related to formulas in subsections 14.1.14–14.1.20 in Ref. 10.

VII. CIRCUITAL RELATIONS

To obtain formulas for the Coulomb wave functions for arbitrary values of arg ρ, it is convenient to use circuital relations. For the Coulomb wave function $F_L(\eta,\rho)$ with complex values of ρ, and allowed complex values of L and η, one obtains from (5.8) and (5.10) the circuital relation

$$F_L(\eta,\rho) = -\exp[+i\pi(L+i\eta)]F_L(-\eta,\rho e^{-i\pi}), \tag{7.1a}$$

which remains valid if i is replaced by $-i$. For the linear combinations (1.2) of the Coulomb wave functions with arbitrary complex values of ρ, and allowed complex values of L and η, one obtains from (2.1a), (2.1b), and (2.8) the circuital relation

$$\psi_+(L,\eta,\rho) = \exp[-i\pi(L+i\eta)]\psi_-(L,-\eta,\rho e^{-i\pi}). \tag{7.2a}$$

Using (1.2), (7.1a), and (7.2a), we obtain

$$\begin{aligned}\psi_-(L,\eta,\rho) &= \exp[+i\pi(L+i\eta)]\psi_+(L,-\eta,\rho e^{-i\pi})\\ &\quad -\{\exp[+i\pi(L+i\eta)]-\exp[-i\pi(L+i\eta)]\}\psi_-(L,-\eta,\rho e^{-i\pi}).\end{aligned} \tag{7.2b}$$

Using (1.2), (7.1a), and (7.2a), we then obtain

$$\begin{aligned}G_L(\eta,\rho) &= \exp[-i\pi(L+i\eta)]G_L(-\eta,\rho e^{-i\pi})\\ &\quad +i\{\exp[+i\pi(L+i\eta)]-\exp[-i\pi(L+i\eta)]\}F_L(-\eta,\rho e^{-i\pi}).\end{aligned} \tag{7.1b}$$

Now we obtain from (7.1a),

$$F_L(\eta,\rho) = -\exp[-i\pi(L-i\eta)]F_L(-\eta,\rho e^{+i\pi}), \tag{7.3a}$$

which like (7.1a) remains valid when i is replaced by $-i$, from (7.1a) and (7.1b),

$$\begin{aligned}G_L(\eta,\rho) &= \exp[+i\pi(L-i\eta)]G_L(-\eta,\rho e^{+i\pi})\\ &\quad +i\{\exp[+i\pi(L-i\eta)]-\exp[-i\pi(L-i\eta)]\}F_L(-\eta,\rho e^{+i\pi}),\end{aligned} \tag{7.3b}$$

from (7.2a) and (7.2b),

$$\begin{aligned}\psi_+(L,\eta,\rho) &= \{\exp[+i\pi(L-i\eta)]-\exp[-i\pi(L-i\eta)]\}\psi_+(L,-\eta,\rho e^{+i\pi})\\ &\quad +\exp[-i\pi(L-i\eta)]\psi_-(L,-\eta,\rho e^{+i\pi}),\end{aligned} \tag{7.4a}$$

and from (7.2a),

$$\psi_-(L,\eta,\rho) = \exp[+i\pi(L-i\eta)]\psi_+(L,-\eta,\rho e^{+i\pi}). \tag{7.4b}$$

For the Weber functions, which are related to Coulomb wave functions with $L=-\frac{1}{4}$, circuital relations were obtained by P. O. Fröman, Karlsson, and Yngve (unpublished report included in Yngve's thesis 1972) who noticed that the circuital relations for these functions given in Ref. 10, subsections 19.18.4 and 19.18.5, and in Ref. 13, Eqs. (281) and (282), have to be corrected by the replacement of ϕ_2 by $ia\pi/2$ and of $[\Gamma(1/2-ia)/\Gamma(1/2+ia)]^{1/2}$ by $\exp(a\pi/2)$.

VIII. ASYMPTOTIC FORMULAS

We recall that when considering a Coulomb wave function with the parameters L and η, or a linear combination $\psi_\pm$ of two such Coulomb wave functions, we always assume that the conditions (2.3a) and (2.3b) are fulfilled.

In Sec. IV we have, for the sake of simplicity, used fixed contours of integration in the formulas (4.1), (4.2a), (4.2b), (4.3a), and (4.3b). We can, however, rotate these contours of integration as long as the integrals remain convergent, and thus extend the region of validity for arg ρ. After the contours in (4.2a), (4.2b) and (4.3a), (4.3b) have been rotated, we introduce the expansions

$$\left(1+\frac{we^{\pm i\pi/2}}{2\rho}\right)^{L\mp i\eta}=\sum_{k=0}^{\infty}\binom{L\mp i\eta}{k}\left(\frac{we^{\pm i\pi/2}}{2\rho}\right)^{k}, \tag{8.1}$$

and change the order of summation and integration to get the asymptotic expansions

$$\psi_{+}(L,\eta,\rho)\sim\frac{\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}}{\Gamma(L+1+i\eta)}\times\sum_{k=0}^{\infty}\binom{L-i\eta}{k}\left(\frac{-1}{2i\rho}\right)^{k}\int_0^{\infty}\exp(-w)w^{L+i\eta+k}\,dw,$$

$$-\pi<\arg\rho<2\pi,\quad \mathrm{Re}(L+1+i\eta)>0, \tag{8.2a}$$

$$\psi_{-}(L,\eta,\rho)\sim\frac{\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}}{\Gamma(L+1-i\eta)}\times\sum_{k=0}^{\infty}\binom{L+i\eta}{k}\left(\frac{1}{2i\rho}\right)^{k}\int_0^{\infty}\exp(-w)w^{L-i\eta+k}\,dw,$$

$$-2\pi<\arg\rho<\pi,\quad \mathrm{Re}(L+1-i\eta)>0, \tag{8.2b}$$

and

$$\psi_{+}(L,\eta,\rho)\sim-\frac{1}{2\pi i}\exp\{+i[\rho-\eta\ln(2\rho)-3L\pi/2-i\eta\pi+\sigma_L(\eta)]\}\Gamma(-L-i\eta)\times\sum_{k=0}^{\infty}\binom{L-i\eta}{k}\left(\frac{-1}{2i\rho}\right)^{k}\int_{\Gamma}\exp(-w)w^{L+i\eta+k}\,dw,$$

$$-\pi<\arg\rho<2\pi,\quad L+i\eta\neq\text{integer}, \tag{8.3a}$$

$$\psi_{-}(L,\eta,\rho)\sim-\frac{1}{2\pi i}\exp\{-i[\rho-\eta\ln(2\rho)+L\pi/2-i\eta\pi+\sigma_L(\eta)]\}\Gamma(-L+i\eta)\times\sum_{k=0}^{\infty}\binom{L+i\eta}{k}\left(\frac{1}{2i\rho}\right)^{k}\int_{\Gamma}\exp(-w)w^{L-i\eta+k}\,dw,$$

$$-2\pi<\arg\rho<\pi,\quad L-i\eta\neq\text{integer}. \tag{8.3b}$$

Note that we started with the rotated contours, but in (8.2a), (8.2b) and (8.3a), (8.3b), where the regions of arg ρ for the validity of the formulas are extended, we could, according to Cauchy's integral theorem, use the fixed contours again. This is possible only *after* we have made the expansions (8.1). Since

$$\int_0^{\infty}\exp(-w)w^{L\pm i\eta+k}\,dw=\Gamma(L+1\pm i\eta+k), \tag{8.4}$$

we obtain from (8.2a) and (8.2b),

$$\psi_+(L,\eta,\rho)\sim\frac{\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}}{\Gamma(L+1+i\eta)}\sum_{k=0}^{\infty}\binom{L-i\eta}{k}\left(\frac{-1}{2i\rho}\right)^k\Gamma(L+1+i\eta+k),$$

$$-\pi<\arg\rho<2\pi, \tag{8.5a}$$

$$\psi_-(L,\eta,\rho)\sim\frac{\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}}{\Gamma(L+1-i\eta)}\sum_{k=0}^{\infty}\binom{L+i\eta}{k}\left(\frac{1}{2i\rho}\right)^k\Gamma(L+1-i\eta+k),$$

$$-2\pi<\arg\rho<\pi, \tag{8.5b}$$

and since, as a consequence of (8.4),

$$\int_\Gamma \exp(-w)w^{L\pm i\eta+k}\,dw=(-1)^{k+1}\frac{2\pi i\exp[(L\pm i\eta)\pi i]}{\Gamma(-L\mp i\eta-k)}, \tag{8.6}$$

we obtain from (8.3a) and (8.3b),

$$\psi_+(L,\eta,\rho)\sim\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}$$
$$\times\sum_{k=0}^{\infty}\binom{L-i\eta}{k}\left(\frac{1}{2i\rho}\right)^k\frac{\Gamma(-L-i\eta)}{\Gamma(-L-i\eta-k)},$$

$$-\pi<\arg\rho<2\pi, \tag{8.7a}$$

$$\psi_-(L,\eta,\rho)\sim\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L(\eta)]\}$$
$$\times\sum_{k=0}^{\infty}\binom{L+i\eta}{k}\left(\frac{-1}{2i\rho}\right)^k\frac{\Gamma(-L+i\eta)}{\Gamma(-L+i\eta-k)},$$

$$-2\pi<\arg\rho<\pi. \tag{8.7b}$$

With the aid of the reflection formula for the gamma function, one can easily show that (8.5a) and (8.5b) are the same as (8.7a) and (8.7b), respectively. The conditions (2.3a) are, of course, assumed to be fulfilled, but the limitations on the parameters L and η associated with the integral representations used in Sec. IV do not appear in (8.5a), (8.5b) and (8.7a), (8.7b).

Keeping in (8.5a), (8.5b) or (8.7a), (8.7b) only the term corresponding to $k=0$, we obtain the asymptotic formulas (8.8b), (8.8b′) corresponding to $\nu=0$ (to be given below), and from these formulas we obtain with the aid of the circuital relations (7.2a), (7.2b) and (7.4a), (7.4b) the asymptotic formulas (8.8a), (8.8a′), (8.8b), (8.8b′), and (8.8c), (8.8c′) for an arbitrary integer ν. From the asymptotic formulas (8.8a)–(8.8c′), where ν is an arbitrary integer, and ε is a fixed, arbitrarily small, positive number, one obtains with the aid of (1.2) the asymptotic formulas (8.9a)–(8.9c′), the consistency of which can be checked with the aid of the circuital relations (7.1a), (7.1b) and (7.3a), (7.3b). The consistency of the asymptotic formulas given below can also be checked with the aid of the reflection formulas (3.2a), (3.2b) and (3.5a), (3.5b). The asymptotic formulas below, in which, for the sake of simplicity, we have written σ_L instead of $\sigma_L(\eta)$, are valid when the parameters L and η are kept fixed, while ρ is sufficiently large. We also remark that one can replace $\exp[\pm i\sigma_L(\eta)]$ by the last member of (2.9) with the branch of the square root chosen as described below (2.9) and (2.10).

The asymptotic formulas (8.8a)–(8.8c′) and (8.9a)–(8.9c′) are, in general, valid in intervals for $\arg\rho$ of the extension $2\pi-2\varepsilon$. However, for $\nu=0$ (8.8b) and (8.8c) give the same asymptotic formula for ψ_+, which is thus valid for $-\pi+\varepsilon<\arg\rho<2\pi-\varepsilon$, and for $\nu=0$ (8.8a′) and (8.8b′) give the same asymptotic formula for ψ_-, which is thus valid for $-2\pi+\varepsilon<\arg\rho<\pi-\varepsilon$. For the particular value $\nu=0$ the asymptotic formulas (8.8b), (8.8c) and (8.8a′), (8.8b′) are thus valid in intervals for $\arg\rho$ of the extension $3\pi-2\varepsilon$, as is also seen from (8.5a), (8.5b) and (8.7a), (8.7b).

When $(2\nu-2)\pi+\epsilon<\arg\rho<2\nu\pi-\epsilon$ *we have*

$$\psi_+\sim\frac{\exp[-2(\nu-1)\pi\eta]}{\sin(2\pi L)}\{\sin(2\nu\pi L)-\sin[2(\nu-1)\pi L]\exp(-2\pi\eta)\}$$
$$\times\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp(+2\nu\pi\eta)}{\sin(2\pi L)}\sin[2(\nu-1)\pi L]\{1-\exp[-2\pi i(L-i\eta)]\}$$
$$\times\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\} \tag{8.8a}$$

and

$$\psi_-\sim\frac{\exp[-2(\nu-1)\pi\eta]}{\sin(2\pi L)}\sin(2\nu\pi L)\{1-\exp[2\pi i(L+i\eta)]\}\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[+2(\nu-1)\pi\eta]}{\sin(2\pi L)}\{\sin[2(\nu-1)\pi L]\exp(2\pi\eta)-\sin(2\nu\pi L)\}$$
$$\times\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}, \tag{8.8a$'$}$$

and hence

$$F_L(\eta,\rho)\sim\frac{\exp[+2\nu\pi i(L+i\eta)]}{2i}\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[+2(\nu-1)\pi i(L-i\eta)]}{2i}\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}, \tag{8.9a}$$

and

$$G_L(\eta,\rho)\sim\frac{\exp[-2(\nu-1)\pi\eta]}{2\sin(2\pi L)}(\sin(2\nu\pi L)\{2-\exp[2\pi i(L+i\eta)]\}$$
$$-\sin[2(\nu-1)\pi L]\exp(-2\pi\eta))\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp(+2\nu\pi\eta)}{2\sin(2\pi L)}(\sin[2(\nu-1)\pi L]\{2-\exp[-2\pi i(L-i\eta)]\}$$
$$-\sin(2\nu\pi L)\exp(-2\pi\eta))\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}. \tag{8.9a$'$}$$

When $(2\nu-1)\pi+\epsilon<\arg\rho<(2\nu+1)\pi-\epsilon$, *we have*

$$\psi_+\sim\frac{\exp[-2(\nu-1)\pi\eta]}{\sin(2\pi L)}\{\sin(2\nu\pi L)-\sin[2(\nu-1)\pi L]\exp(-2\pi\eta)\}$$
$$\times\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[+2(\nu+1)\pi\eta]}{\sin(2\pi L)}\sin(2\nu\pi L)\{1-\exp[-2\pi i(L-i\eta)]\}$$
$$\times\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\} \tag{8.8b}$$

and

$$\psi_- \sim \frac{\exp[-2(\nu-1)\pi\eta]}{\sin(2\pi L)}\sin(2\nu\pi L)\{1-\exp[2\pi i(L+i\eta)]\}$$
$$\times\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp(+2\nu\pi\eta)}{\sin(2\pi L)}\{\sin(2\nu\pi L)\exp(2\pi\eta)-\sin[2(\nu+1)\pi L]\}$$
$$\times\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}, \tag{8.8b$'$}$$

and hence

$$F_L(\eta,\rho)\sim\frac{\exp[2\nu\pi i(L+i\eta)]}{2i}\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[2\nu\pi i(L-i\eta)]}{2i}\exp\{-i[\rho-\eta\ln(2\pi)-L\pi/2+\sigma_L]\} \tag{8.9b}$$

and

$$G_L(\eta,\rho)\sim\frac{\exp[-2(\nu-1)\pi\eta]}{2\sin(2\pi L)}(\sin(2\nu\pi L)\{2-\exp[2\pi i(L+i\eta)]\}$$
$$-\sin[2(\nu-1)\pi L]\exp(-2\pi\eta))\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[2(\nu+1)\pi\eta]}{2\sin(2\pi L)}(\sin(2\nu\pi L)\{2-\exp[-2\pi i(L-i\eta)]\}$$
$$-\sin[2(\nu+1)\pi L]\exp(-2\pi\eta))\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}. \tag{8.9b$'$}$$

When $2\nu\pi+\epsilon<\arg\rho<(2\nu+2)\pi-\epsilon$, *we have*

$$\psi_+\sim\frac{\exp(-2\nu\pi\eta)}{\sin(2\pi L)}\{\sin[2(\nu+1)\pi L]-\sin(2\nu\pi L)\exp(-2\pi\eta)\}$$
$$\times\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[+2(\nu+1)\pi\eta]}{\sin(2\pi L)}\sin(2\nu\pi L)\{1-\exp[-2\pi i(L-i\eta)]\}$$
$$\times\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\} \tag{8.8c}$$

and

$$\psi_-\sim\frac{\exp(-2\nu\pi\eta)}{\sin(2\pi L)}\sin[2(\nu+1)\pi L]\{1-\exp[+2\pi i(L+i\eta)]\}$$
$$\times\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp(+2\nu\pi\eta)}{\sin(2\pi L)}\{\sin(2\nu\pi L)\exp(2\pi\eta)-\sin[2(\nu+1)\pi L]\}$$
$$\times\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}, \tag{8.8c$'$}$$

and hence

$$F_L(\eta,\rho)\sim\frac{\exp[2(\nu+1)\pi i(L+i\eta)]}{2i}\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[2\nu\pi i(L-i\eta)]}{2i}\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\} \tag{8.9c}$$

and

$$G_L(\eta,\rho)\sim\frac{\exp(-2\nu\pi\eta)}{2\sin(2\pi L)}(\sin[2(\nu+1)\pi L]\{2-\exp[2\pi i(L+i\eta)]\}$$
$$-\sin(2\nu\pi L)\exp(-2\pi\eta))\exp\{+i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}$$
$$-\frac{\exp[2(\nu+1)\pi\eta]}{2\sin(2\pi L)}(\sin(2\nu\pi L)\{2-\exp[-2\pi i(L-i\eta)]\}$$
$$-\sin[2(\nu+1)\pi L]\exp(-2\pi\eta))\exp\{-i[\rho-\eta\ln(2\rho)-L\pi/2+\sigma_L]\}. \tag{8.9c$'$}$$

ACKNOWLEDGMENTS

In the spring of 1995 two of the authors (Fröman and Yngve) were invited to participate in the program *Exponential Asymptotics* at the Isaac Newton Institute for Mathematical Sciences in Cambridge, England. These authors wish to express their gratitude to the Isaac Newton Institute for a most fruitful time in an extremely stimulating scientific atmosphere. They are also much indebted to Professor Frank Olver and other participants for fruitful discussions.

[1] F. L. Yost, J. A. Wheeler, and G. Breit, Phys. Rev. **49**, 174 (1936).
[2] M. H. Hull, Jr., and G. Breit, *Coulomb Wave Functions*, Encyclopedia of Physics, edited by S. Flügge, 1959, Vol. XLI/1, pp. 408–465.
[3] M. J. Seaton, Mon. Not. R. Astron. Soc. **118**, 504 (1958).
[4] A. R. Curtis, *Coulomb Wave Functions*, Royal Society Mathematical Tables II (Cambridge University Press, Cambridge, 1964).
[5] M. J. Seaton, Proc. Phys. Soc. London **88**, 801 (1966).
[6] W. Eissner, H. Nussbaumer, H. E. Saraph, and M. J. Seaton, J. Phys. B **2**, 341 (1969).
[7] M. J. Seaton, J. Phys. B **11**, 4067 (1978).
[8] M. J. Seaton, Comput. Phys. Commun. **25**, 87 (1982).
[9] J. Humblet, Ann. Phys. (N.Y.) **155**, 461 (1984).
[10] M. Abramowitz and I. A. Stegun (editors), in *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematics Series 55, Washington, 1964.
[11] Y. L. Luke, *The Special Functions and their Approximations* (Academic Press, New York, 1969), Vol. I.
[12] H. Buchholz, *The Confluent Hypergeometric Function*, Springer Tracts in Natural Philosophy (Springer-Verlag, Berlin, 1969), Vol. 15.
[13] National Physical Laboratory, *Tables of Weber Parabolic Cylinder Functions*, computed by Scientific Computing Service Ltd., Mathematical Introduction by J. C. P. Miller, Her Majesty's Stationery Office, London, 1955.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Phase-integral formulas for Coulomb wave functions with complex values of the variable and the parameters

Aleksander Dzieciol, Staffan Yngve, and Per Olof Fröman
Department of Theoretical Physics, University of Uppsala, Box 803, S-751 08 Uppsala, Sweden

(Received 8 June 1998; accepted for publication 22 June 1999)

Phase-integral formulas for the Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$ and certain linear combinations of these functions, with complex values of the variable ρ and the parameters L and η, are obtained explicitly up to the fifth order of the phase-integral approximation for two different choices of the base function.
 [S0022-2488(99)00710-0]

I. INTRODUCTION

The central role played by Coulomb wave functions in different fields of theoretical physics made the need for easily tractable analytic approximate formulas apparent already in the early days of quantum mechanics. The WKB method was a powerful tool, which, however, soon revealed certain deficiencies that have remained unremedied in the literature on Coulomb wave functions. Already, Kramers[1] had found that to obtain correct results for a quantal particle in an attractive Coulomb field by means of the ordinary WKB method in first-order approximation it is necessary to replace $L(L+1)$ by $(L+1/2)^2$, where L is the angular momentum quantum number. When applying that method to hydrogenic atoms, Young and Uhlenbeck[2] also found that the same replacement is necessary to obtain the correct wave function close to the origin and the correct Balmer formula. In their well-known and frequently quoted article Yost, Wheeler, and Breit[3] pointed out that the WKB method gives only a crude approximation of the Coulomb wave functions, frequently with errors of an unknown amount, and that the replacement of $L(L+1)$ by $(L+1/2)^2$ does not always improve formulas obtained by means of the ordinary WKB method. Langer[4] gave a new argument for the replacement in question. Bartlett, Rice, and Good, Jr.,[5] confirmed that the applicability of the ordinary WKB method to the differential equation for the Coulomb wave functions is uncertain in the neighborhood of the origin, and that it would be desirable to have a more definite method, which did not require an adjustment of the approximate solution. In their comprehensive review article on Coulomb wave functions, M. H. Hull, Jr., and Breit[6] state that Langer's above-mentioned argument cannot be regarded as more than an indication of likelihood regarding the applicability of the approximation.

The literature on Coulomb wave functions is to a large extent restricted to the consideration of real values of the variable ρ and the parameters L and η. Thus, one does not find general formulas that apply to complex values of the variable ρ and the parameters L and η in standard handbooks; see, for instance, Abramowitz and Stegun.[7] However, in theoretical physics there often appear problems in which one uses Coulomb wave functions with complex values of the variable and the parameters. This is the case, for instance, in Regge pole theory, where the angular momentum is complex, in scattering theory, when the energy is complex, and in quantum defect theory for closed channels.

Accurate approximations for special functions of mathematical physics can be obtained by means of the phase-integral method in which one uses a phase-integral approximation of arbitrary order generated from an unspecified base function. This method has its origin in the WKB method, but further development has yielded the just mentioned phase-integral method, which in essential respects differs from the WKB method; see Chap. 1 in a book by Fröman and Fröman[8] and a paper by Dammert and P. O. Fröman.[9] The phase-integral method is freed from the previously mentioned deficiencies of the WKB method, and, when a conveniently high order of the phase-integral approximation is used, it yields formulas of very high accuracy within their regions of validity. An application of the phase-integral method to special functions has been made in a paper by P. O. Fröman, Karlsson, and Yngve,[10] in which phase-integral formulas for Bessel functions were obtained.

Our purpose in the present paper is to use the phase-integral method to derive phase-integral formulas up to the fifth-order approximation for the Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$, and certain linear combinations of these functions, which are solutions of the differential equation,

$$d^2\psi/d\rho^2+[1-2\eta/\rho-L(L+1)/\rho^2]\psi=0. \tag{1.1}$$

The phase-integral formulas obtained are very general, since we allow the variable ρ and the parameters L and η to assume complex values without unnecessary restrictions. To make the formulas valid for a large range of parameter values, we consider two essentially different choices of the base function, which are appropriate when L is sufficiently large, and when L is sufficiently small while η is sufficiently large, respectively. The phase-integral formulas are valid close to $\rho=0$ only for the one of the choices of the base-function, which is appropriate when L is sufficiently large.

II. PHASE-INTEGRAL APPROXIMATION GENERATED FROM AN UNSPECIFIED BASE FUNCTION

In this section we consider the Schrödinger-like differential equation,

$$d^2\psi/d\rho^2+R(\rho)\psi=0, \tag{2.1}$$

where for the moment we let $R(\rho)$ be an unspecified analytic function of the complex variable ρ. As described in more detail by Fröman and Fröman,[8] Chap. 1, Sec. 1.3.1, we introduce into (2.1) a ''small'' book-keeping parameter λ that is finally put equal to unity. Thus we get the auxiliary differential equation,

$$d^2\psi/d\rho^2+[Q^2(\rho)/\lambda^2+R(\rho)-Q^2(\rho)]\psi=0, \tag{2.2}$$

which goes over into (2.1) when $\lambda=1$. The function $Q(\rho)$ is the unspecified base function from which the phase-integral approximation is generated. The auxiliary differential equation (2.2) has two linearly independent solutions of the form

$$f_1(\rho)=q^{-1/2}(\rho)\exp[+iw(\rho)], \tag{2.3a}$$

$$f_2(\rho)=q^{-1/2}(\rho)\exp[-iw(\rho)], \tag{2.3b}$$

where

$$w(\rho)=\int^{\rho}q(\rho)d\rho. \tag{2.4}$$

Inserting (2.3a) or (2.3b) for ψ into (2.2), we obtain

$$q^{+1/2}d^2q^{-1/2}/d\rho^2-q^2+Q^2(\rho)/\lambda^2+R(\rho)-Q^2(\rho)=0. \tag{2.5}$$

Introducing instead of ρ the variable

$$\zeta=\int^{\rho}Q(\rho)d\rho, \tag{2.6}$$

we can write (2.5) in the form

$$1-\left(\frac{q\lambda}{Q(\rho)}\right)^2+\epsilon_0\lambda^2+\left(\frac{q\lambda}{Q(\rho)}\right)^{+1/2}\frac{d^2}{d\zeta^2}\left(\frac{q\lambda}{Q(\rho)}\right)^{-1/2}\lambda^2=0, \tag{2.7}$$

where

$$\epsilon_0=Q^{-3/2}(\rho)d^2Q^{-1/2}(\rho)/d\rho^2+R(\rho)/Q^2(\rho)-1. \tag{2.8}$$

To obtain a formal solution of (2.7), we put

$$q\lambda/Q(\rho)=\sum_{n=0}^{\infty} Y_{2n}\lambda^{2n}, \tag{2.9}$$

where Y_0 is assumed to be different from zero, and Y_{2n} $(n=0,1,2,...)$ are independent of λ. Inserting the expansion (2.9) into (2.7), expanding the left-hand member in powers of λ, and putting the coefficient of each power of λ equal to zero, we get $Y_0=\pm 1$ and a recurrence formula, from which one can successively obtain the functions $Y_2, Y_4, Y_6,\ldots,$ each one of which can be expressed in terms of ϵ_0, defined by (2.8), and derivatives of ϵ_0 with respect to ζ. Since we have both $+$ and $-$ in the exponents of (2.3a) and (2.3b), it is no restriction to choose $Y_0=1$. The first few functions Y_{2n} are then

$$Y_0=1, \tag{2.10a}$$

$$Y_2=\frac{1}{2}\epsilon_0, \tag{2.10b}$$

$$Y_4=-\frac{1}{8}\left(\epsilon_0^2+\frac{d^2\epsilon_0}{d\zeta^2}\right). \tag{2.10c}$$

The choice of the unspecified base function $Q(\rho)$ shows itself only in the expressions (2.6) and (2.8) for ζ and ϵ_0, respectively, which depend explicitly on $R(\rho)$ and $Q(\rho)$, while the expressions for the functions Y_{2n}, which are given in terms of ϵ_0 and derivatives of ϵ_0 with respect to ζ, do not depend explicitly on $R(\rho)$ and the choice of the base function $Q(\rho)$. The expressions for the functions Y_{2n} can therefore be determined once and for all. We also remark that at the zeros and poles of $Q^2(\rho)$ the functions $Q(\rho)$ and $Q^{-1/2}(\rho)$ may have branch points, whereas the functions ϵ_0, Y_{2n} and $q(\rho)/Q(\rho)$ are all single valued.

Truncating the infinite series in (2.9) at $n=N$, we obtain

$$q(\rho)=Q(\rho)\sum_{n=1}^{N} Y_{2n}\lambda^{2n-1}. \tag{2.11}$$

Inserting (2.11) into (2.3a), (2.3b), and (2.4), and putting $\lambda=1$, we get the phase-integral approximation of the order $2N+1$, generated from the base function $Q(\rho)$, which is an approximate solution of the original differential equation (2.1). For $N>0$ the function $q(\rho)$ has poles at the transition zeros and simple zeros in the neighborhood of each transition zero.[11] In the first order the phase-integral approximation is the same as the usual JWKB approximation if $Q(\rho)=R^{1/2}(\rho)$, but in higher orders it differs in essential respects from the JWKB approximation of corresponding order; see Chap. 1 in Refs. 8 and 9.

In the present paper we would prefer to choose as the fixed lower limit of integration in (2.4) and (2.6) either a first-order zero or a first-order pole of $Q^2(\rho)$. Unfortunately, this can be done in (2.4) only for the first order of the phase-integral approximation, since for the higher-order approximations the integral would be divergent. In this situation we replace (2.4) and (2.6) by

$$w(\rho)=\int_{(t)}^{\rho} q(\rho)d\rho=\frac{1}{2}\int_{\Gamma_t(\rho)} q(\rho)d\rho, \tag{2.12a}$$

$$\zeta(\rho)=\int_t^{\rho} Q(\rho)d\rho, \tag{2.12b}$$

where t is the first-order zero or first-order pole in question. The integral in the second member of (2.12a) is a shorthand notation, introduced by Fröman *et al.*,[12] pp. 158–161, for the contour integral in the last member of (2.12a), where $\Gamma_t(\rho)$ is a path of integration that starts at the point corresponding to ρ on a Riemann sheet adjacent to the complex ρ plane under consideration, encircles t in the positive or in the negative sense, and ends at ρ. It is immaterial for the value of the contour integral whether the path of integration encircles t in the positive or in the negative sense, but the terminal point must be the point ρ in the complex ρ plane under consideration. In the first order of the phase-integral approximation, the contour integral can be replaced by an ordinary integral from t to ρ, i.e.,

$$\int_{(t)}^{\rho} q(\rho)d\rho=\int_t^{\rho} Q(\rho)d\rho=\zeta(\rho), \quad \text{first-order approximation.} \tag{2.13}$$

In the present paper we shall consider the differential equation for the Coulomb wave functions, which one obtains by putting in (2.1), with well-known notations,

$$R(\rho)=1-2\eta/\rho-L(L+1)/\rho^2, \tag{2.14}$$

and we shall choose the base function to be

$$Q(\rho)=(1-2\eta/\rho-\Lambda^2/\rho^2)^{1/2}, \tag{2.15}$$

where Λ is a conveniently chosen complex constant. From experience with the phase-integral method it is known that when L is sufficiently large, it is convenient to choose $\Lambda=L+\frac{1}{2}$, whereas when L is sufficiently small and η is sufficiently large, it is convenient to choose $\Lambda=0$. We restrict ourselves to cases when $\rho^2Q^2(\rho)$ has simple zeros by assuming that $\eta^2+\Lambda^2\neq 0$. We introduce in the complex ρ plane a cut that joins the roots $\eta\pm(\eta^2+\Lambda^2)^{1/2}$ of the equation $\rho^2Q^2(\rho)=0$. When $\Lambda\neq 0$, we extend this cut by a cut from one of these roots to the origin. As a result $Q^{-1/2}(\rho)$, and hence $q^{-1/2}(\rho)$ is a single valued function of ρ, whether Λ is equal to zero or different from zero. The phase of $Q^{-1/2}(\rho)$ is chosen such that $Q^{-1/2}(\rho)\to 1$ as $\rho\to\infty$. To make the integrals in (2.12a) and (2.12b) single valued, we replace the complex ρ plane by a Riemann surface with the above-mentioned cut on each Riemann sheet. Since the point t is assumed to be a simple zero of $\rho^2Q^2(\rho)$, it is either a first-order zero or a first-order pole of $Q^2(\rho)$.

We now define

$$\delta=\lim_{\rho\to\infty}\left(\int_{(t)}^{\rho} q(\rho)d\rho-[\rho-\eta\ln(2\rho)]\right), \tag{2.16}$$

where the path of integration is the same as in (2.12a). The branch of $\ln(2\rho)$ is determined by the value of $\ln(2t)$, which will be prescribed later. The value of δ depends on L, η, Λ, and the order $2N+1$ of the phase-integral approximation. It also depends on the choice of the transition point t. Inserting (2.11) with $\lambda=1$ into (2.16), we obtain

$$\delta=\sum_{n=0}^{N}\delta^{(2n+1)}, \tag{2.17}$$

with

$$\delta^{(1)} = \lim_{\rho\to\infty}\left(\int_t^{\rho} Q(\rho)d\rho - [\rho - \eta\ln(2\rho)]\right), \tag{2.18a}$$

$$\delta^{(2n+1)} = \int_{(t)}^{\infty} Y_{2n}Q(\rho)d\rho, \quad n>0. \tag{2.18b}$$

As already mentioned, the phase of $Q^{-1/2}(\rho)$ is chosen such that $Q^{-1/2}(\rho)$ tends to unity as ρ tends to infinity, and the case $\Lambda = \pm i\eta$ is excluded, since we assume $\rho^2 Q^2(\rho)$ to have two *simple* zeros; see (2.15). We thus assume that $\eta^2+\Lambda^2\neq 0$.

The essential problem in the present paper is the calculation of $\delta^{(2n+1)}$. This is problematical for $n=0$, since the expression for the integral in (2.18a) contains a logarithm, the appropriate branch of which must be clearly specified when ρ moves away from t. For $n>0$ there are no real difficulties but only the question of evaluating rather complicated integrals.

III. CALCULATION OF $\delta^{(2n+1)}$

The integral in (2.18a) can be evaluated in explicit analytical form. For the corresponding indefinite integral one obtains with the aid of (2.15) the formula

$$\int^{\rho} Q(\rho)d\rho = \rho Q(\rho) - \eta\ln\{\rho[1+Q(\rho)]-\eta\} + \frac{\Lambda}{2i}\ln\frac{\Lambda Q(\rho)+i(\eta+\Lambda^2/\rho)}{\Lambda Q(\rho)-i(\eta+\Lambda^2/\rho)}. \tag{3.1}$$

With the aid of (2.15) one also obtains the identities

$$\{\rho[1+Q(\rho)]-\eta\}\{\rho[1-Q(\rho)]-\eta\} = \eta^2+\Lambda^2, \tag{3.2a}$$

$$[\Lambda Q(\rho)+i(\eta+\Lambda^2/\rho)][\Lambda Q(\rho)-i(\eta+\Lambda^2/\rho)] = (\eta^2+\Lambda^2). \tag{3.2b}$$

Since $\eta^2+\Lambda^2\neq 0$ it follows from (3.2a) and (3.2b) that no one of the logarithms in (3.1) has a branch point off the cuts of the Riemann surface for the complex variable ρ. The logarithms in (3.1) are thus single-valued when ρ does not encircle the cut, but the first logarithm in (3.1) changes when ρ moves from one Riemann sheet to another. To obtain the definite integral in (2.18a), we choose the constant lower limit of integration in (3.1) to be one of the two roots of the equation $\rho^2 Q^2(\rho)=0$, i.e., from (2.15),

$$t = \eta \pm (\eta^2+\Lambda^2)^{1/2}, \tag{3.3}$$

where we choose $(\eta^2+\Lambda^2)^{1/2}=\eta$ when $\Lambda=0$ and $(\eta^2+\Lambda^2)^{1/2}=\Lambda$ when $\eta=0$. From (3.1) we now obtain the formula

$$\int_t^{\rho} Q(\rho)d\rho = \rho Q(\rho) - \eta\ln\frac{\rho[1+Q(\rho)]-\eta}{t-\eta} + \Lambda\left(\frac{1}{2i}\ln\frac{\Lambda Q(\rho)+i(\eta+\Lambda^2/\rho)}{\Lambda Q(\rho)-i(\eta+\Lambda^2/\rho)} - \frac{\pi}{2}\right), \tag{3.4}$$

where for $\rho=t$ the first logarithm is equal to zero and the second logarithm (when $\Lambda\neq 0$) is equal to πi; note that $t\neq 0$ and $Q(t)=0$ when $\Lambda\neq 0$. Since the logarithms in (3.4) have no branch points off the cuts of the Riemann surface, they are uniquely determined on that surface.

To treat the first logarithm in (3.4), we want to find a curve on which that logarithm is real and increases from 0 to $+\infty$ as ρ moves from t to ∞. To this purpose we put

$$\frac{\rho[1+Q(\rho)]-\eta}{t-\eta} = r, \tag{3.5}$$

where it is required that r is real and $\geqslant 1$. From (3.5) we obtain

$$Q(\rho)=\frac{r(t-\eta)+\eta}{\rho}-1. \tag{3.6}$$

Inserting (2.15) into (3.6) and squaring the resulting equation, and noting that $\eta^2+\Lambda^2=(t-\eta)^2$ according to (3.3), we obtain after some calculations,

$$\rho=\tfrac{1}{2}(r+1/r)(t-\eta)+\eta. \tag{3.7}$$

From (3.7) it is seen that as r increases from 1 to $+\infty$, ρ moves from t to ∞ along the straight line that emerges from t in the direction opposite to the direction from t to η. By means of (3.7) and (3.3) one finds that the point $\rho=0$ may lie on this straight line when Λ/η is purely imaginary and to its absolute value smaller than unity. When $\Lambda=0$ the straight line may emerge from but not pass through the point $\rho=0$. When $\Lambda\neq 0$ and ρ is sufficiently small, the left-hand member of (3.5) is a single valued function of ρ that does not change sign when ρ changes sign, and there appear therefore no difficulties if the straight line passes through $\rho=0$. Since $Q(\rho)\to 1$ as $\rho\to\infty$, and because of (3.7), we obtain from (3.5),

$$\ln\frac{\rho[1+Q(\rho)]-\eta}{t-\eta}\sim\ln(2\rho)-\ln(t-\eta),\quad \rho\to\infty, \tag{3.8}$$

where $\mathrm{Im}\ln(2\rho)=\mathrm{Im}\ln(t-\eta)=\arg(t-\eta)$ when $\arg\rho=\arg(t-\eta)$.

To discuss the second logarithm in (3.4), we consider first the particular case when $\Lambda\neq 0$ and $\eta=0$, and we choose the sign of Λ such that $t=\Lambda$; see (3.3). With $\eta=0$ we obtain from (2.15),

$$\rho=\frac{\Lambda}{(1-Q^2)^{1/2}},\quad \eta=0, \tag{3.9}$$

where we choose the sign of the square root such that $\rho=\Lambda=t$ when $Q(\rho)=0$. With the aid of (3.9) we can write the second logarithm in (3.4) with $\eta=0$ as

$$\ln\frac{Q(\rho)+i[1-Q^2(\rho)]^{1/2}}{Q(\rho)-i[1-Q^2(\rho)]^{1/2}},\quad \eta=0. \tag{3.10}$$

This logarithm is equal to πi when $\rho=t$, i.e., when $Q(\rho)=0$, since the second logarithm in (3.4) is equal to πi when $\Lambda\neq 0$ and $\rho=t$. Letting ρ move from t to ∞ in such a way that $Q(\rho)$ is real and increases monotonically from 0 to 1, which according to (3.9) means that $\arg\rho=\arg\Lambda$, we find from (3.10) that the value of the logarithm at $\rho=+\infty\exp(i\arg\Lambda)$ is equal to 0. When $\rho\to+\infty\exp(i\arg\Lambda)$ we therefore have for the second logarithm in (3.4) the expression

$$\ln\frac{\Lambda+i\eta}{\Lambda-i\eta}\to 0,\quad \text{as}\quad \eta\to 0. \tag{3.11}$$

This must be true for any value of $\arg\rho$, since in the limit $\rho\to\infty$ the logarithm is independent of ρ.

In the limit $\rho\to\infty$ we obtain from (3.4) with the aid of (2.15), (3.8), and (3.11) the formula

$$\int_t^\rho Q(\rho)d\rho\sim\rho-\eta\ln(2\rho)+\eta[\ln(t-\eta)-1]+\Lambda\left(\frac{1}{2i}\ln\frac{\Lambda+i\eta}{\Lambda-i\eta}-\frac{\pi}{2}\right), \tag{3.12}$$

where $\mathrm{Im}\ln(2\rho)=\mathrm{Im}\ln(t-\eta)=\arg(t-\eta)$ when $\arg\rho=\arg(t-\eta)$, and the last logarithm in (3.12) tends to 0 as $\eta\to 0$.

Inserting (3.12) into (2.18a), we obtain the formula

$$\delta^{(1)}=\eta[\ln(t-\eta)-1]+\Lambda\left(\frac{1}{2i}\ln\frac{\Lambda+i\eta}{\Lambda-i\eta}-\frac{\pi}{2}\right),\quad \Lambda=0 \quad\text{or}\quad \Lambda\neq 0. \tag{3.13}$$

It is seen that $\delta^{(1)}$ depends on the choice of the transition point t. In the particular situation when Λ and η are real, and we use the upper sign in (3.3), we can write (3.13) as

$$\begin{aligned}\delta^{(1)}&=\eta[\ln(t-\eta)-1]+\Lambda[\arctan(\eta/\Lambda)-\pi/2]\\ &=\eta[\ln(\eta^2+\Lambda^2)^{1/2}-1]+\Lambda[\arctan(\eta/\Lambda)-\pi/2],\quad \Lambda=0\quad\text{or}\quad\Lambda\neq 0.\end{aligned} \tag{3.13$'$}$$

As regards the higher-order contributions $\delta^{(2n+1)}$ $(n>0)$, which one obtains from (2.18b), we shall not present the derivation, but only give the expressions for $\delta^{(3)}$ and $\delta^{(5)}$.

When $\Lambda=0$, one obtains

$$\delta^{(3)}=-\frac{1+6L(L+1)}{12\eta}=\frac{1-3(2L+1)^2}{24\eta},\quad \Lambda=0, \tag{3.14a}$$

$$\delta^{(5)}=\frac{1-30[L(L+1)]^2}{360\eta^3}=\frac{7-30(2L+1)^2+15(2L+1)^4}{2880\eta^3},\quad \Lambda=0. \tag{3.15a}$$

It is seen that these quantities are independent of the choice of the transition point t.

If in Eqs. (6.3.44a)–(6.3.44c) in Ref. 13 we replace l by L and η_0 by η and put $\eta_2=\eta_4=0$, we obtain from Eq. (6.3.44a) in Ref. 13 our formula (3.13′), and we obtain from Eqs. (6.3.44b) and (6.3.44c) in Ref. 13 our formulas (3.14a) and (3.15a).

When $\Lambda\neq 0$, one obtains

$$\delta^{(3)}=\frac{(L+\frac{1}{2})^2-\Lambda^2}{2\Lambda}\left(\frac{1}{2i}\ln\frac{\Lambda+i\eta}{\Lambda-i\eta}-\pi/2\right)+\frac{\eta}{24(\Lambda^2+\eta^2)},\quad \Lambda\neq 0, \tag{3.14b}$$

$$\begin{aligned}\delta^{(5)}=&-\frac{[(L+\frac{1}{2})^2-\Lambda^2]^2}{8\Lambda^3}\left(\frac{1}{2i}\ln\frac{\Lambda+i\eta}{\Lambda-i\eta}-\pi/2\right)-\frac{\eta}{2(\Lambda^2+\eta^2)}\left(\frac{(L+\frac{1}{2})^2-\Lambda^2}{2\Lambda}\right)^2\\ &-\frac{\eta[(L+\frac{1}{2})^2-\Lambda^2]}{24(\Lambda^2+\eta^2)^2}-\frac{7\eta}{960(\Lambda^2+\eta^2)^2}+\frac{7\eta^3}{720(\Lambda^2+\eta^2)^3},\quad \Lambda\neq 0.\end{aligned} \tag{3.15b}$$

It is seen that $\delta^{(3)}$ and $\delta^{(5)}$ do not depend on the choice of the transition point t. In the particular situation when Λ and η are real, we can write (3.14b) and (3.15b) as

$$\delta^{(3)}=\frac{(L+\frac{1}{2})^2-\Lambda^2}{2\Lambda}[\arctan(\eta/\Lambda)-\pi/2]+\frac{\eta}{24(\Lambda^2+\eta^2)},\quad \Lambda\neq 0, \tag{3.14b$'$}$$

$$\begin{aligned}\delta^{(5)}=&-\frac{[(L+\frac{1}{2})^2-\Lambda^2]^2}{8\Lambda^3}[\arctan(\eta/\Lambda)-\pi/2]-\frac{\eta}{2(\Lambda^2+\eta^2)}\left(\frac{(L+\frac{1}{2})^2-\Lambda^2}{2\Lambda}\right)^2\\ &-\frac{\eta[(L+\frac{1}{2})^2-\Lambda^2]}{24(\Lambda^2+\eta^2)^2}-\frac{7\eta}{960(\Lambda^2+\eta^2)^2}+\frac{7\eta^3}{720(\Lambda^2+\eta^2)^3},\quad \Lambda\neq 0.\end{aligned} \tag{3.15b$'$}$$

If in Eqs. (6.3.35a)–(6.3.35c) in Ref. 13, we replace ξ_0 by Λ and η_0 by η, and use for ξ_2 and ξ_4 Eqs. (6.3.34a) and (6.3.34b) in Ref. 13 with l replaced by L, and put $\eta_2=\eta_4=0$, we obtain (3.13′), (3.14a′), and (3.15b′).

When $L \neq -\frac{1}{2}$ and one wants the phase-integral functions to be approximate solutions of the differential equation for the Coulomb wave functions also close to $\rho=0$, one chooses $\Lambda = L+\frac{1}{2}$. When Λ is chosen in this way, the formulas (3.14b), (3.14b′) and (3.15b), (3.15b′) are simplified considerably.

IV. PHASE-INTEGRAL FORMULAS FOR COULOMB WAVE FUNCTIONS

According to (2.3a), (2.3b), (2.12a), (2.15), and (2.16), we have, when $\rho \to \infty$ for fixed L, η and Λ,

$$f_1(\rho) \sim \exp\{+i[\rho - \eta \ln(2\rho) - L\pi/2 + \sigma_L(\eta) - \Delta]\}, \quad \rho \to \infty, \tag{4.1a}$$

$$f_2(\rho) \sim \exp\{-i[\rho - \eta \ln(2\rho) - L\pi/2 + \sigma_L(\eta) - \Delta]\}, \quad \rho \to \infty, \tag{4.1b}$$

where [cf. Eq. (6.3.26) in Ref. 13]

$$\Delta = \sigma_L(\eta) - L\pi/2 - \delta. \tag{4.2}$$

According to Eqs. (8.8a–c′) and (8.9a–c′) in Ref. 14, the Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$, as well as the functions $\psi_\pm(L,\eta,\rho)$ defined by Eq. (1.2) in Ref. 14, are asymptotically, for fixed values of L and η and large values of ρ, linear combinations of $\exp\{\pm i[\rho - \eta \ln(2\rho) - L\pi/2 + \sigma_L(\eta)]\}$. For sufficiently large values of ρ and fixed values of L, η, and Λ, one can, according to (4.1a) and (4.1b), replace $\exp\{+i[\rho - \eta \ln(2\rho) - L\pi/2 + \sigma_L(\eta)]\}$ by $\exp(+i\Delta) f_1(\rho)$ and $\exp\{-i[\rho - \eta \ln(2\rho) - L\pi/2 + \sigma_L(\eta)]\}$ by $\exp(-i\Delta) f_2(\rho)$. Equations (8.8a–c′) and (8.9a–c′) in Ref. 14, which give asymptotic expressions for $\psi_\pm(L,\eta,\rho)$, $F_L(\eta,\rho)$, and $G_L(\eta,\rho)$ when $\rho \to \infty$ for fixed L and η, can therefore be written as follows.

When $(2\nu-2)\pi + \epsilon < \arg \rho < 2\nu\pi - \epsilon$, *we get*

$$\psi_+ \sim \frac{\exp[-2(\nu-1)\pi\eta + i\Delta]}{\sin(2\pi L)} \{\sin(2\nu\pi L) - \sin[2(\nu-1)\pi L]\exp(-2\pi\eta)\} f_1(\rho) - \frac{\exp(+2\nu\pi\eta - i\Delta)}{\sin(2\pi L)} \sin[2(\nu-1)\pi L]\{1 - \exp[-2\pi i(L - i\eta)]\} f_2(\rho), \tag{4.3a}$$

$$\psi_- \sim \frac{\exp[-2(\nu-1)\pi\eta + i\Delta]}{\sin(2\pi L)} \sin(2\nu\pi L)\{1 - \exp[2\pi i(L + i\eta)]\} f_1(\rho) - \frac{\exp[+2(\nu-1)\pi\eta - i\Delta]}{\sin(2\pi L)} \{\sin[2(\nu-1)\pi L]\exp(2\pi\eta) - \sin(2\nu\pi L)\} f_2(\rho), \tag{4.3a′}$$

$$F_L(\eta,\rho) \sim \frac{\exp[+2\nu\pi i(L + i\eta) + i\Delta]}{2i} f_1(\rho) - \frac{\exp[+2(\nu-1)\pi i(L - i\eta) - i\Delta]}{2i} f_2(\rho), \tag{4.4a}$$

$$G_L(\eta,\rho) \sim \frac{\exp[-2(\nu-1)\pi\eta + i\Delta]}{2\sin(2\pi L)} (\sin(2\nu\pi L)\{2 - \exp[2\pi i(L + i\eta)]\} - \sin[2(\nu-1)\pi L]\exp(-2\pi\eta)) f_1(\rho) - \frac{\exp(+2\nu\pi\eta - i\Delta)}{2\sin(2\pi L)} (\sin[2(\nu-1)\pi L]\{2 - \exp[-2\pi i(L - i\eta)]\} - \sin(2\nu\pi L)\exp(-2\pi\eta)) f_2(\rho). \tag{4.4a′}$$

When $(2\nu-1)\pi+\epsilon<\arg\rho<(2\nu+1)\pi-\epsilon$, *we get*

$$\psi_{+}\sim\frac{\exp[-2(\nu-1)\pi\eta+i\Delta]}{\sin(2\pi L)}\{\sin(2\nu\pi L)-\sin[2(\nu-1)\pi L]\exp(-2\pi\eta)\}f_1(\rho)$$
$$-\frac{\exp[+2(\nu+1)\pi\eta-i\Delta]}{\sin(2\pi L)}\sin(2\nu\pi L)\{1-\exp[-2\pi i(L-i\eta)]\}f_2(\rho), \tag{4.3b}$$

$$\psi_{-}\sim\frac{\exp[-2(\nu-1)\pi\eta+i\Delta]}{\sin(2\pi L)}\sin(2\nu\pi L)\{1-\exp[2\pi i(L+i\eta)]\}f_1(\rho)$$
$$-\frac{\exp(+2\nu\pi\eta-i\Delta)}{\sin(2\pi L)}\{\sin(2\nu\pi L)\exp(2\pi\eta)-\sin[2(\nu+1)\pi L]\}f_2(\rho), \tag{4.3b$'$}$$

$$F_L(\eta,\rho)\sim\frac{\exp[2\nu\pi i(L+i\eta)+i\Delta]}{2i}f_1(\rho)-\frac{\exp[2\nu\pi i(L-i\eta)-i\Delta]}{2i}f_2(\rho), \tag{4.4b}$$

$$G_L(\eta,\rho)\sim\frac{\exp[-2(\nu-1)\pi\eta+i\Delta]}{2\sin(2\pi L)}(\sin(2\nu\pi L)\{2-\exp[2\pi i(L+i\eta)]\}$$
$$-\sin[2(\nu-1)\pi L]\exp(-2\pi\eta))f_1(\rho)$$
$$-\frac{\exp[2(\nu+1)\pi\eta-i\Delta]}{2\sin(2\pi L)}(\sin(2\nu\pi L)\{2-\exp[-2\pi i(L-i\eta)]\}$$
$$-\sin[2(\nu+1)\pi L]\exp(-2\pi\eta))f_2(\rho). \tag{4.4b$'$}$$

When $2\nu\pi+\epsilon<\arg\rho<(2\nu+2)\pi-\epsilon$, *we get*

$$\psi_{+}\sim\frac{\exp(-2\nu\pi\eta+i\Delta)}{\sin(2\pi L)}\{\sin[2(\nu+1)\pi L]-\sin(2\nu\pi L)\exp(-2\pi\eta)\}f_1(\rho)$$
$$-\frac{\exp[+2(\nu+1)\pi\eta-i\Delta]}{\sin(2\pi L)}\sin(2\nu\pi L)\{1-\exp[-2\pi i(L-i\eta)]\}f_2(\rho), \tag{4.3c}$$

$$\psi_{-}\sim\frac{\exp(-2\nu\pi\eta+i\Delta)}{\sin(2\pi L)}\sin[2(\nu+1)\pi L]\{1-\exp[+2\pi i(L+i\eta)]\}f_1(\rho)$$
$$-\frac{\exp(+2\nu\pi\eta-i\Delta)}{\sin(2\pi L)}\{\sin(2\nu\pi L)\exp(2\pi\eta)-\sin[2(\nu+1)\pi L]\}f_2(\rho), \tag{4.3c$'$}$$

$$F_L(\eta,\rho)\sim\frac{\exp[2(\nu+1)\pi i(L+i\eta)+i\Delta]}{2i}f_1(\rho)-\frac{\exp[2\nu\pi i(L-i\eta)-i\Delta]}{2i}f_2(\rho), \tag{4.4c}$$

$$G_L(\eta,\rho)\sim\frac{\exp(-2\nu\pi\eta+i\Delta)}{2\sin(2\pi L)}(\sin[2(\nu+1)\pi L]\{2-\exp[2\pi i(L+i\eta)]\}$$
$$-\sin(2\nu\pi L)\exp(-2\pi\eta))f_1(\rho)$$
$$-\frac{\exp[2(\nu+1)\pi\eta-i\Delta]}{2\sin(2\pi L)}(\sin(2\nu\pi L)\{2-\exp[-2\pi i(L-i\eta)]\}$$
$$-\sin[2(\nu+1)\pi L]\exp(-2\pi\eta))f_2(\rho). \tag{4.4c$'$}$$

The asymptotic formulas (8.8a–c′) and (8.9a–c′) in Ref. 14, from which the phase-integral formulas (4.3a–c′) and (4.4a–c′) in the present paper have been derived, are valid only when, for fixed values of L and η, ρ tends to infinity. The phase-integral formulas (4.3a–c′) and (4.4a–c′) in the present paper, however, may remain valid when, still for fixed values of L and η, the variable ρ moves from infinity along convenient paths of monotonicity, provided that ρ does not come too close to a transition point. These formulas have thus a considerably larger region of validity on the Riemann surface for the complex variable ρ than the formulas (8.8a-c′) and (8.9a–c′) in Ref. 14. To determine in detail the regions of validity for (4.3a–c′) and (4.4a–c′) and the significance of each one of the two terms in these formulas, one uses the basic estimates (4.3a)–(4.3d) in Ref. 15, which apply to a path on which the absolute value of $\exp[iw(\rho)]$ increases monotonically (in the nonstrict sense) from the initial point to the final point, and the corresponding basic estimates for a path on which the absolute value of $\exp[-iw(\rho)]$ increases monotonically (in the nonstrict sense) from the initial point to the final point.

For large values of ρ the phase-integral formulas (4.3a–c′) and (4.4a–c′), like the asymptotic formulas (8.8a–c′) and (8.9a–c′) in Ref. 14, are, in general, valid in intervals for $\arg\rho$ of the extension $2\pi-2\epsilon$, but for the reasons mentioned in Sec. VIII of Ref. 14, the phase-integral formula (4.3b,c) with $\nu=0$ for ψ_+ is valid for $-\pi+\epsilon<\arg\rho<2\pi-\epsilon$, and the phase-integral formula (4.3a′,b′) with $\nu=0$ for ψ_- is valid for $-2\pi+\epsilon<\arg\rho<\pi-\epsilon$. For the particular value $\nu=0$, the phase-integral formulas (4.3b,c) and (4.3a′,b′) for ψ_+ and ψ_-, respectively, are thus valid in intervals for $\arg\rho$ of the extension $3\pi-2\epsilon$, when ρ is large.

We recall that on the Riemann surface for the complex variable ρ we have introduced the cut described below (2.15) in Sec. II. We also remark that because of (4.2) the quantities exp $(\pm i\sigma_L)$, which depend on both L and η, appear in (4.3a–c′) and (4.4a–c′), and that these quantities can be replaced by the expressions given in Eq. (2.9) in Ref. 14.

V. CONCLUSIONS

Up to the fifth-order approximation phase-integral formulas are given, in (4.4a–c′) for the Coulomb wave functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$, and in (4.3a–c′) for the linear combinations $\psi_\pm(L,\eta,\rho)=G_L(\eta,\rho)\pm iF_L(\eta,\rho)$. The quantity Δ in these phase-integral formulas, which is given by (4.2), (2.17), (2.18a), and (2.18b), depends on the choice of the base function $Q(\rho)$ and the order $2N+1$ of the phase-integral approximation. The base function $Q(\rho)$ is given by (2.15), where Λ may be either equal zero (which is appropriate when L is sufficiently small, while η is sufficiently large) or different from zero (which is appropriate when L is sufficiently large). The phase of $Q^{-1/2}(\rho)$ is chosen such that $Q^{-1/2}(\rho)\rightarrow 1$ as $\rho\rightarrow\infty$, the restriction $\eta^2+\Lambda^2\neq 0$ is introduced on the parameters L and η, and certain cuts are introduced, as explained between (2.15) and (2.16). The phase-integral approximation is valid close to $\rho=0$ only when $\Lambda=L+\frac{1}{2}\neq 0$. The quantities $\delta^{(2n+1)}$ in (2.17) are for $2n+1=1$, 3, and 5 given by (3.13), (3.13′), (3.14a), and (3.15a) when $\Lambda=0$, and by (3.13), (3.13′), (3.14b), (3.14b′), (3.15b), and (3.15b′) when $\Lambda\neq 0$. For the choice $\Lambda=L+\frac{1}{2}$ the formulas (3.14b), (3.14b′) and (3.15b), (3.15b′) become considerably simplified.

ACKNOWLEDGMENTS

In the spring of 1995 two of the authors (Fröman and Yngve) were invited to participate in the program Exponential Asymptotics at the Isaac Newton Institute for Mathematical Sciences in Cambridge, England. These authors wish to express their gratitude to the Isaac Newton Institute for a most fruitful time in an extremely stimulating scientific atmosphere. They are also much indebted to Professor Frank Olver and other participants for fruitful discussions.

[1] H. A. Kramers, Z. Phys. **39**, 828 (1926).
[2] L. A. Young and G. E. Uhlenbeck, Phys. Rev. **36**, 1154 (1930).
[3] F. L. Yost, J. A. Wheeler, and G. Breit, Phys. Rev. **49**, 174 (1936).
[4] R. E. Langer, Phys. Rev. **51**, 669 (1937).
[5] R. H. Bartlett, M. H. Rice, and R. H. Good, Jr., Ann. Phys. (N.Y.) **2**, 372 (1957).

[6] M. H. Hull, Jr. and G. Breit, *Encyclopedia of Physics*, edited by S. Flügge, XLI/1, 1959, pp. 408–465.
[7] M. Abramowitz and I. A. Stegun (editors), *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematics Series 55, Washington, 1964.
[8] N. Fröman and P. O. Fröman, *Phase-Integral Method Allowing Nearlying Transition Points*, with adjoined papers by A. Dzieciol, N. Fröman, P. O. Fröman, A. Hökback, S. Linnæus, B. Lundborg, and E. Walles, Springer Tracts in Natural Philosophy, edited by C. Truesdell (Springer-Verlag New York, 1996), Vol. 40.
[9] Ö. Dammert and P. O. Fröman, J. Math. Phys. **21**, 1683 (1980).
[10] P. O. Fröman, F. Karlsson, and S. Yngve, J. Math. Phys. **27**, 2738 (1986).
[11] N. Fröman, Ann. Phys. (N.Y.) **61**, 451 (1970).
[12] N. Fröman, P. O. Fröman, and B. Lundborg, Math. Proc. Cambridge Philos. Soc. **104**, 153 (1988).
[13] N. Fröman, P. O. Fröman, and S. Linnæus, adjoined paper (Chap. 6) in Ref. 8.
[14] A. Dzieciol, S. Yngve, and P. O. Fröman, J. Math. Phys. **40**, 6145 (1999).
[15] N. Fröman and P. O. Fröman, *JWKB Approximation, Contributions to the Theory* (North-Holland, Amsterdam, 1965); Russian translation (MIR, Moscow, 1967).

Collective dynamics of solitons and inequivalent quantizations

J. P. Garrahan[a)]
Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón I, Ciudad Universitaria, 1428 Buenos Aires, Argentina

M. Kruczenski[b)]
Departamento de Física, TANDAR, Comisión Nacional de Energía Atómica, Av. Libertador 8250, 1429 Buenos Aires, Argentina

(Received 13 April 1998; accepted for publication 7 September 1999)

The collective dynamics of solitons with a coset space G/H as moduli space is studied. It is shown that the collective band for a vibrational state is given by the inequivalent coset space quantization corresponding to the representation of H carried by the vibration. To leading order the collective dynamics is free motion in G/H coupled to background gauge fields determined by the vibrational state. © *1999 American Institute of Physics.* [S0022-2488(99)01412-7]

I. INTRODUCTION

Solitons arise as static finite energy solutions to the equations of motion of nonlinear field theories. In general, a given soliton depends upon a set of parameters or moduli, and is a point in the manifold of solutions of equal energy, or moduli space. In many cases this manifold is simply an homogeneous or coset space G/H, where G is the group of symmetries of the action and $H \subset G$ is the symmetry of the solitonic solution.

Around a soliton there are two kinds of quantum excitations. The first corresponds to collective motion in the moduli space. The second are vibrational (intrinsic) excitations out of it. If the energy for the collective excitations is much lower than that for the vibrational ones the low energy spectrum can be approximately described by collective bands associated with each vibrational state. These bands can be described by an effective quantum mechanical problem given by the motion of a particle in the moduli space. However, as is well known from molecular and nuclear rotational bands, different vibrational states may have different collective bands. It is the purpose of this paper to show that in the case when the moduli space of the soliton is a coset space G/H a simple description of the collective bands of vibrational states can be given in terms of inequivalent coset space quantizations introduced by Mackey,[1] and more recently studied by Landsman and Linden[2] and McMullan and Tsutsui,[3] among others.

Since the soliton is invariant under the subgroup H, vibrational excitations fit into irreducible representations (irreps) of H. Below we show that the collective band corresponding to a vibrational state in a representation χ of H realizes a representation of G induced by χ. This representation of G is reducible, and when it is broken into irreducible representations the whole collective band is obtained. This is equivalent to saying that the collective band for a vibrational state is given by the inequivalent quantization of G/H corresponding to the irrep χ of H carried by the vibration. In this way we find that collective motion is a physical example of the inequivalent coset space quantizations.

The lowest energy collective band of the soliton is that of the ground state of the vibrations. If the ground state is in the trivial representation of H, our results provide nothing new for

[a)]Present address: Theoretical Physics, Department of Physics, University of Oxford, 1 Keble Road, Oxford, OX1 3NP, United Kingdom. Electronic mail: garrahan@df.uba.ar
[b)]Present address: Institutionen för Teoretisk Fysik, Box 803, S-751 08 Uppsala, Sweden.

determining this band, which is obtained from the usual (trivial) quantization of free motion in G/H. However, although the soliton is invariant under H, the quantum ground state need not be in the trivial representation because of topological factors which may arise when adiabatically performing the operations of H. For example, in the SU (3) Skyrmion the Wess–Zumino–Witten term[4] contributes with an extra phase. The same happens with the SU (2) Skyrmion[5] when it is quantized as a fermion. For the $B=1$ Skyrmion the result is the well-known fact that the ground state has spin an isospin 1/2.

The outline of this paper is as follows: In Sec. II we introduce a model[6] which generically describes a soliton model. The action is invariant under a group G of transformation of the fields but the soliton only under a subgroup H, so that the moduli space is a coset space G/H. In Sec. III we summarize Mackey's[1] approach to the quantization on a coset space, and the more recent results by Landsman and Linden[2] and McMullan and Tsutsui.[3] Section IV is the main section of this paper. There we demonstrate the relation between the wave functions and energies of the collective bands of vibrational states and the corresponding ones of the inequivalent coset space quantizations. Finally, in Sec. V we give our conclusions. We also discuss an example of a possible application of our results.

II. THE MODEL

Let us consider a general bosonic field theory in $d+1$ dimensions. The fields $\phi^s(\mathbf{x},t)$ are real and take their values in an internal space $\mathcal{T}$. The variables $\mathbf{x}$ and t are space and time coordinates, and $s=1...\dim(\mathcal{T})$ is an internal index. At any given time the fields $\phi^s(\mathbf{x},t)$ determine a configuration of the system, given by a map from R^d (or a compactification of it) into $\mathcal{T}$. The set of these maps corresponds to the configuration space $\mathcal{C}$ of the theory. The action is taken to be quadratic in the time derivatives of the fields, namely,

$$S=\int dt d^d x\left(\frac{1}{2}G_{st}\dot{\phi}^s(\mathbf{x},t)\dot{\phi}^t(\mathbf{x},t)-V\right), \tag{1}$$

where sums over repeated indices are implicit and,

$$G_{st}=G_{st}[\phi(\mathbf{x})]=G_{st}(\phi^{s_1},\partial_{i_1}\phi^{s_2},\ldots,\partial_{i_1\cdots i_n}\phi^{s_n}),$$

$$V=V[\phi(\mathbf{x})]=V(\phi^{s_1},\partial_{i_1}\phi^{s_2},\ldots,\partial_{i_1\cdots i_n}\phi^{s_n}),$$

are functions of the fields and a finite number of its spatial derivatives. For example if the action corresponds to a nonlinear σ-model,

$$S=\int dt d^d x\left(\frac{1}{2}g_{st}(\phi)\partial^\mu\phi^s(\mathbf{x},t)\partial_\mu\phi^t(\mathbf{x},t)-v(\phi)\right), \tag{2}$$

the functionals G_{st} and V are given by,

$$G_{st}=g_{st}(\phi),\quad V=g_{st}(\phi)\boldsymbol{\partial}\phi^s\boldsymbol{\partial}\phi^t+v(\phi), \tag{3}$$

that is, the spatial derivatives are included in V. In other cases, for example the Skyrme model,[5] the function G_{st} depend also on the spatial derivatives of the field. A useful way of thinking about the system is to consider it as a particle moving in the infinite dimensional space of configurations $\mathcal{C}$ (i.e., of maps from R^d into $\mathcal{T}$) with a potential V and a metric given by $g_{st}(\mathbf{x},\mathbf{y})=G_{st}\delta^d(\mathbf{x}-\mathbf{y})$.

The equations of motion following from the action (1) have static solutions which satisfy

$$\frac{\delta}{\delta\phi^s(\mathbf{x})}\int d^d y V[\phi(\mathbf{y})]=\frac{\partial V}{\partial\phi^s(\mathbf{x})}-\partial_{i_1}\frac{\partial V}{\partial(\partial_{i_1}\phi^s(\mathbf{x}))}+\cdots+(-)^n\partial_{i_1\cdots i_n}\frac{\partial V}{\partial(\partial_{i_1\cdots i_n}\phi^s(\mathbf{x}))}=0. \tag{4}$$

For our present purpose we assume that Eqs. (4) have soliton solutions. We consider the case in which the action S is invariant under an unbroken finite-dimensional compact group G of transformations of the fields,

$$\phi^s(\mathbf{x}) \mapsto R_g^s[\phi(\mathbf{x})], \quad g \in G. \tag{5}$$

This group usually includes spatial rotations and internal transformations of the fields. The invariance of the action under these transformations is expressed as,

$$G_{st}[\phi(\mathbf{x})]\delta^3(x-y) = \int d^dz \frac{\delta R_g^u[\phi(\mathbf{z})]}{\delta\phi^s(\mathbf{x})} \frac{\partial R_g^v[\phi(\mathbf{z})]}{\delta\phi^t(\mathbf{y})} G_{uv}[\phi(\mathbf{z})], \tag{6}$$

$$\int d^dx V[\phi(x)] = \int d^dx V[R_g[\phi(x)]]. \tag{7}$$

The first condition can be rephrased saying that the transformations are isometries of the space $\mathcal{C}$.

Let us consider a soliton solution $\tilde{\phi}(\mathbf{x})$ of (4), i.e., a minimum of the potential $\int V[\phi]d^dx$. In general, $\tilde{\phi}(\mathbf{x})$ is only invariant under a subgroup H of the symmetry group G of the action. Due to the invariance of the potential under G, any configuration $R_g[\tilde{\phi}(\mathbf{x})]$ obtained by acting with G on $\tilde{\phi}(\mathbf{x})$ is also a solution. The whole family of solutions, known as the moduli space of the soliton, is thus given by the orbit of $\tilde{\phi}(\mathbf{x})$ under G, and is therefore an homogeneous or coset space,

$$\mathcal{M} = O_G(\tilde{\phi}) = G/H. \tag{8}$$

It is possible that there exist other zero modes not related to symmetries of the action, but we shall ignore that possibility here. Consider now fluctuations around the soliton, $\tilde{\phi}(\mathbf{x}) \to \tilde{\phi}(\mathbf{x}) + \varphi(\mathbf{x},t)$. The linearized equations for the fluctuations read,

$$\tilde{G}_{st}\ddot{\varphi}^t(\mathbf{x},t) + \int d^dy K_{st}(\mathbf{x},\mathbf{y})\varphi^t(\mathbf{y},t) = 0, \tag{9}$$

where $\tilde{G}_{st}$ is evaluated on the solution, $\tilde{G}_{st} = G_{st}[\tilde{\phi}(\mathbf{x})]$ and,

$$K_{st}(\mathbf{x},\mathbf{y}) = \left.\frac{\delta^2 V}{\delta\phi^s(\mathbf{x})\delta\phi^t(\mathbf{y})}\right|_{\phi=\tilde{\phi}}. \tag{10}$$

The fluctuations can be understood as infinitesimal vectors in the tangent space to $\mathcal{C}$ at the point $\tilde{\phi}(\mathbf{x})$. The (infinite) set of those normal to the moduli space $\mathcal{M}$ are massive excitations, since $\mathcal{M}$ is the ''valley'' of minima of the static energy. They satisfy,

$$\int d^dy K_{st}(\mathbf{x},\mathbf{y})\varphi^t(\mathbf{y},t) - \omega_n^2 \tilde{G}_{st}\psi_n^t(\mathbf{x}) = 0, \tag{11}$$

where n labels the modes, $\omega_n \neq 0$, and $\psi_n^s(\mathbf{x})$ are the corresponding normalized eigenfunctions, $\int d^dx \tilde{G}_{st}\psi_n^s(\mathbf{x})\psi_m^t(\mathbf{x}) = \delta_{nm}$. The fluctuations tangent to $\mathcal{M}$ are massless, and are usually known as zero modes. They satisfy,

$$\int d^dy K_{st}(\mathbf{x},\mathbf{y})\varphi^t(\mathbf{y},t)\psi_\alpha^t(\mathbf{y}) = 0, \tag{12}$$

where $\alpha=1,...,\dim(G/H)$ label the zero modes, which are given by the nonvanishing infinitesimal transformations (5) of the soliton, $\psi_\alpha^s(\mathbf{x})=\delta_\alpha\tilde{\phi}^s(\mathbf{x})$. The norm of the zero modes corresponds to the inertia tensor of the soliton,

$$\mathcal{I}_{\alpha\beta}=\int d^d x\,\tilde{G}_{st}\psi^s_\alpha(\mathbf{x})\psi^t_\beta(\mathbf{x}). \tag{13}$$

The massive and zero modes satisfy the completeness relation,

$$\tilde{G}_{st}[\delta^{nm}\psi^s_n(\mathbf{x})\psi^t_m(\mathbf{y})+\mathcal{I}^{\alpha\beta}\psi^s_\alpha(\mathbf{x})\psi^s_\beta(\mathbf{y})]=\delta^d(x-y), \tag{14}$$

where $\mathcal{I}^{\alpha\beta}$ is the inverse of the inertia tensor.

Quantum-mechanically, the massive modes correspond to vibrations of the soliton, while the zero modes to collective motion in $\mathcal{M}$. We are interested in the case in which the collective energy is much smaller than the vibrational one. Schematically, this is given by

$$\frac{\hbar^2}{\mathcal{I}}\ll\hbar\omega\Rightarrow\frac{\hbar}{\mathcal{I}\omega}\ll 1, \tag{15}$$

where $\mathcal{I}$ is the inertia for the collective motion and ω the frequency for the intrinsic excitations. In this regime it is natural to treat the collective motion exactly and the vibrations in perturbation theory. The relation (15) implies that the small parameter is proportional to $\hbar$, and therefore the perturbative expansion is an expansion in loops (in powers of $\hbar$). In the following $\hbar=1$, so that the expansion parameter is $1/\mathcal{I}\omega$.

In order to proceed with the quantization around the soliton $\tilde{\phi}(\mathbf{x})$ it is necessary to treat differently the massive vibrations from the zero modes. One possible approach, which will prove useful to our purposes, is to introduce collective coordinates by performing a transformation of the fields $\phi^s(\mathbf{x},t)$ with time dependent parameters,[7–10,6]

$$\phi^s(\mathbf{x},t)\mapsto R^s_{g(\alpha)}[\phi(\mathbf{x},t)], \tag{16}$$

where $\alpha^a(t)$ $(a=1,...,\dim(G))$ parameterize G. The action (1) is not symmetric under this time dependent transformations, and it changes to,

$$S\mapsto\int dt\,d^d x\left(\frac{1}{2}G_{st}D_0\phi^s(\mathbf{x},t)D_0\phi^t(\mathbf{x},t)-V\right). \tag{17}$$

The covariant time derivatives are given by,

$$D_0\phi^s(\mathbf{x},t)=\dot{\phi}^s(\mathbf{x},t)+\dot{\alpha}^a(t)\zeta^b_a(\alpha(t))\delta_b\phi^s(\mathbf{x},t), \tag{18}$$

where $\zeta^b_a(\alpha)$ are components of the left-invariant Cartan–Maurer one-form $g^{-1}dg=d\alpha^a\zeta^b_a(\alpha)T_b$, with T_a being the infinitesimal group generators, and $\delta_a\phi^s(\mathbf{x},t)$ correspond to the infinitesimal transformations (5). The transformed action (17) is invariant under gauge transformations, i.e., time dependent transformations of the fields and the collective coordinates,

$$\delta_\varepsilon\phi^s(\mathbf{x},t)=\varepsilon^a(t)\delta_a\phi^s(\mathbf{x},t),\quad \delta_\varepsilon\alpha^a(t)=-\varepsilon^b(t)\Theta^a_b(\alpha(t)),\quad (\Theta=\zeta^{-1}). \tag{19}$$

The transformed action (17) can be understood as describing the problem from an arbitrary moving frame of reference, its motion given by the collective coordinates.

The classical Hamiltonian corresponding to the action (17) is given by

$$\mathcal{H}=\int d^d x\left(\frac{1}{2}G^{st}[\phi(\mathbf{x},t)]\pi_s(\mathbf{x},t)\pi_t(\mathbf{x},t)+V[\phi(\mathbf{x},t)]\right)-\lambda_a(t)\Phi_a(t), \tag{20}$$

where $\pi_s(\mathbf{x},t)$ are the canonical momenta conjugate to $\phi^s(\mathbf{x},t)$, i.e., $\{\pi_s(\mathbf{x},t),\phi^t(\mathbf{y},t)\} = -\delta_s^t\delta^d(x-y)$, and $\lambda_a(t)$ are Lagrange multipliers which impose the first class constraints $\Phi_a = J_a - I_a$, which generate the gauge transformations (19).[11] We display also the explicit dependence of G_{st} and V on the fields. The operators J_a are the generators for the infinitesimal transformations of the fields,

$$J_a = \int d^d x \, \pi_s(\mathbf{x}) \, \delta_a \phi^s(\mathbf{x}), \tag{21}$$

and I_a are the right generators for the collective coordinates, $I_a = \Theta_a^b \mathcal{P}_b$, where $\mathcal{P}_a$ are the conjugate momenta to the collective coordinates, $\{\alpha^a, \mathcal{P}_b\} = \delta_b^a$. While the gauge symmetry (19) acts from the right on the collective coordinates, the original symmetry G acts from the left. The operators that generate the left infinitesimal transformations of the collective coordinates are correspondingly given by $L_a = \hat{\Theta}_a^b \mathcal{P}_b$, where $\hat{\zeta}_a^b(\alpha)$ are components of the right-invariant Cartan–Maurer one-form $dgg^{-1} = d\alpha^a \hat{\zeta}_a^b(\alpha) T_b$, and $\hat{\Theta} = \hat{\zeta}^{-1}$.

The gauge algebra $\mathbf{g}$, generated by the constraints $\Phi_a = J_a - I_a$, can be decomposed into $\mathbf{g} = \mathbf{h} + \mathbf{p}$, where the subalgebra $\mathbf{h}$ is generated by $\Phi_i = J_i - I_i$ $(i = 1,...,\dim(H))$, while the complement $\mathbf{p}$ is generated by $\Phi_\alpha = J_\alpha - I_\alpha$ $(\alpha = \dim(H)+1,...,\dim(G))$. The operators J_i generate the subgroup of transformations of the fields that leave the soliton $\tilde{\phi}(\mathbf{x})$ invariant, while J_α generate the transformations which change it. Since H is a subgroup, the operators J_i close under commutation, $[J_i, J_j] = iC_{ij}^k J_k$, where C_{ab}^c are the structure constants of G. Furthermore, as G is compact the generators J_α can be chosen in such a way that

$$[J_i, J_\alpha] = iC_{i\alpha}^\beta J_\beta, \tag{22}$$

which shows that J_α transform in a representation (possibly reducible) of H.

As is usual in this kind of problems, the idea is to replace the zero mode excitations of the fields $\phi(\mathbf{x},t)$ by the collective coordinates $\alpha(t)$, by means of a suitable gauge fixing of action (17). It is important to note that the number of zero modes is equal to the dimension of G/H, while the number of collective variables is $\dim(G)$. By gauge fixing the constraints Φ_α we will eliminate the zero modes in favor of the collective coordinates α^α on G/H. We will make use of the remaining H-gauge symmetry to prove that the effective collective dynamics of the soliton can be understood in terms of inequivalent coset space quantizations.

III. INEQUIVALENT COSET SPACE QUANTIZATIONS

In this section we review briefly the approach developed by Mackey to the quantization of a system whose configuration space is a coset space G/H.[1] This must be distinguished from the quantization of G/H as a phase space, which is discussed for example in Ref. 12 (and references therein). It was shown by Mackey that when the configuration space is a coset space G/H there are many different quantizations not equivalent to each other by unitary transformations, which are labeled by the unitary irreps χ of the subgroup H. The wave functions in a given inequivalent quantization are vector valued, taking values in the representation space V_χ. They can be obtained from vector valued functions in G which satisfy (χ-equivariant condition),

$$f_\mu(gh) = \pi_{\nu\mu}^\chi(h) f_\nu(g), \tag{23}$$

where $g \in G$, $h \in H$, and $\pi^\chi(h)$ are the matrices of the representation χ of H. The left regular representation of G is defined on functions of G acting as,

$$f(g) \overset{g_1}{\rightarrow} f(g_1^{-1} g). \tag{24}$$

Equation (23) is invariant under this action. Therefore, the set of functions $f_\mu(g)$ satisfying condition (23) transforms under the left action of G in a representation called the representation of G induced by χ. See Refs. 1 and 13 for the definition and properties of induced representations.

Given a local section σ:$G/H\rightarrow G$ every element $g\in G$ can be uniquely written as,

$$g=\sigma(\xi(g))h_\sigma(g), \tag{25}$$

where $\xi\in G/H$ and $h\in H$. This allows to define a one-to-one correspondence between the set of functions $f_\mu(g)$ and functions $F_\mu:G/H\rightarrow V_\chi$ such that,

$$f_\mu(g)=\pi^\chi_{\nu\mu}(h(g))F_\nu(\xi(g)). \tag{26}$$

The functions $F_\mu(\xi)$ transform under G in the induced representation as,

$$F_\mu(\xi)\overset{g}{\rightarrow}\pi^\chi_{\nu\mu}(h_\sigma(g^{-1}\sigma(\xi)))F_\nu(g^{-1}\xi), \tag{27}$$

where $g^{-1}\xi=\xi(g^{-1}\sigma(\xi))$ defines the action of G on G/H.

Landsman and Linden[2] studied the dynamical consequences of the inequivalent quantizations for the motion of a particle in G/H. They discovered that in the nontrivial quantum sectors the particle couples to a background gauge field A_α, known as the H-connection, which takes values in the representation of the sub-algebra $\pi^\chi(\mathbf{h})$. The Hamiltonian is given by,

$$\mathcal{H}=-\tfrac{1}{2}g^{\alpha\beta}(\nabla_\alpha+A_\alpha)(\partial_\beta+A_\beta), \tag{28}$$

where ∇_α is the covariant derivative constructed out of the metric $g_{\alpha\beta}$ on G/H. Due to the H-connection the Hamiltonian is matrix valued (in the trivial representation of H it reduces to minus one-half the Laplacian $-\frac{1}{2}\Delta_{G/H}=-\frac{1}{2}g^{\alpha\beta}\nabla_\alpha\partial_\beta$).

McMullan and Tsutsui[3] developed a different approach to the inequivalent quantizations. Using the fact that in Eq. (25) $h\in H$ can be further decomposed as $h=r(h)s(h)$, where s belongs to the Cartan subgroup of H, they showed that instead of Mackey's functions $f_\mu(g)$ one can use its highest weight component f_χ evaluated at $g=\sigma r$ (i.e., $s=1$). The condition (23) implies that these functions are annihilated by the raising operators $E_{\varphi>0}$ in the Chevalley basis $\{H_\alpha,E_\varphi\}$ of $\mathbf{h}$. In this case the wave functions are scalars, which allows for a simpler definition of the corresponding path integral.

IV. COLLECTIVE DYNAMICS

In this section we will prove that the collective band associated with an intrinsic vibrational state can be obtained from the inequivalent quantization corresponding to the representation of the subgroup H carried by the vibration. To see this we consider the canonical quantization of the action (17) introduced in Sec. II. This action has a G-gauge invariance due to the introduction of the collective coordinates as additional variables. Gauge fixing the G/H part of the gauge symmetry allows to eliminate the zero modes in favor of the collective coordinates. The remaining H-gauge invariance can be treated with the Dirac method of imposing the constraints on the wave functions.[11] In our case this becomes Mackey's condition (23) for the collective wave functions associated to a given vibrational state. Alternatively, the collective coordinates which parameterize H can be eliminated, fixing the H-gauge symmetry. In this case the wave functions become $F_\mu(x)$ of Eq. (27) and the Hamiltonian that of Landsman and Linden.

A. Elimination of the zero modes

We start by eliminating the zero modes in favor of collective coordinates on G/H.

Let us expand the fields in terms of fluctuations around the soliton. The fluctuations can be written as linear combinations of the normal modes (11) and (12),

$$\phi^s(\mathbf{x},t)=\tilde{\phi}^s(\mathbf{x})+\psi_n^s(\mathbf{x})q_n(t)+\psi_\alpha^s(\mathbf{x})q_\alpha(t), \tag{29}$$

where the time dependent coefficients $q_A(t)$ (A stands both for n and α) are now the dynamical degrees of freedom. The conjugate momenta to the fields read,

$$\pi_s(\mathbf{x},t)=\tilde{G}_{st}[\psi_n^t(\mathbf{x})p_n(t)+\mathcal{I}^{\alpha\beta}\psi_\alpha^t(\mathbf{x})p_\beta(t)], \tag{30}$$

where p_A are conjugate to q_A. Inserting these expressions into the Hamiltonian (20), and making use of the equations for the normal modes (11), (12) and their orthogonality relations, we obtain up to quadratic order in the fluctuations,

$$\mathcal{H}=E_{\text{clas}}+\tfrac{1}{2}(p_n^2+\omega_n^2q_n^2)+\tfrac{1}{2}\mathcal{I}^{\alpha\beta}p_\alpha p_\beta-\lambda_a\Phi_a+O(q^3,pq^2). \tag{31}$$

The first term is the classical soliton energy $E_{\text{clas}}=\int d^dxV[\tilde{\phi}(\mathbf{x})]$. The second are the (infinite) harmonic oscillators corresponding to the vibrations of the soliton. The third term are the zero modes, and is purely kinetic. Anharmonic terms have been omitted.

Expanding similarly the constraints we get,

$$\Phi_\alpha=p_\alpha+(D_\alpha)_{AB}p_Aq_B-I_\alpha+O(pq^2), \tag{32}$$

$$\Phi_i=(D_i)_{AB}p_Aq_B-I_i+O(pq^2), \tag{33}$$

where we have kept up to quadratic terms in the fluctuations but the collective operators are treated exactly. The matrices $(D_a)_{AB}$ are defined as,

$$(D_a)_{AB}=\int d^dx\tilde{G}_{st}\psi_A^s(\mathbf{x})\delta_a\psi_B^t(\mathbf{x}), \tag{34}$$

where $\delta_a\psi_A^t(\mathbf{x})$ are the infinitesimal transformations (5) of the eigenfunctions, and correspond to the representation of G under which the fluctuations transform.

We want to eliminate the degrees of freedom associated with the zero energy fluctuations (q_α,p_α). We achieve this by choosing to fix the gauge invariance generated by Φ_α the following gauge fixing conditions,

$$q_\alpha=0, \quad (\alpha=1,...,\dim(G/H)), \tag{35}$$

which satisfy, $\{q_\alpha,\Phi_\beta\}\approx\delta_{\alpha\beta}+(D_\beta)_{\alpha n}q_n+O(q^2)$, where $\approx$ indicates evaluated where (35) holds. This gauge condition *does not* fix the H-gauge invariance generated by Φ_i since, $\{q_\alpha,\Phi_i\}\approx 0$, where we have used the fact that massive and zero fluctuations do not mix under H and therefore $(D_i)_{\alpha n}=0$.

We now replace the Poisson brackets by Dirac brackets,[11]

$$\{A,B\}_D=\{A,B\}-\{A,q_\alpha\}(\delta_{\alpha\beta}-D_{\beta\alpha n}q_n)\{\Phi_\beta,B\}, \tag{36}$$

in order to treat the gauge conditions and constraints as operator identities. We are then able to solve p_α from the equations $\Phi_\alpha=0$,

$$p_\alpha=I_\alpha-(D_\alpha)_{nm}P_nq_m+O(pq^2,Iq). \tag{37}$$

Replacing (q_α,p_α) by Eqs. (35), (37) in the Hamiltonian (31) we obtain,

$$\mathcal{H}=E_{\text{clas}}+\tfrac{1}{2}(p_n^2+\omega_n^2q_n^2)+\tfrac{1}{2}\mathcal{I}^{\alpha\beta}I_\alpha I_\beta-\mathcal{I}^{\alpha\beta}I_\alpha(D_\beta)_{nm}p_nq_m-\lambda_i(t)\Phi_i(t)+O(q^3,pq^2,Ipq^2). \tag{38}$$

The zero modes have been completely eliminated. Instead we have a kinetic term for the collective coordinates and a coupling between the collective coordinates and the vibrations. Anharmonic terms in the vibrations and higher order vibration-collective couplings have been omitted.

B. Collective energy eigenfunctions

The Hamiltonian (38) describes the dynamics of the system in the 1-loop approximation for the vibrations, while the collective motion is treated exactly. To this order, the spectrum corresponds to vibrational states and on top of each state a collective band. There is still the *H*-gauge symmetry.

Since the soliton is invariant under the subgroup *H*, the vibrations split into irreducible representations (irreps) of *H*. Consider now a given vibrational state classified by the irrep χ of *H*, and lets derive the *effective* collective dynamics for this state from the Hamiltonian (38). In other words, we restrict to energies which can only excite collective modes, but not any other vibrational ones. Taking into account this restriction, the wave functions for a state in the collective band has the general form,

$$\psi(q,\alpha)=\varphi_\mu(q)f_\mu(\alpha). \tag{39}$$

The functions $\varphi_\mu(q)$ $(\mu=1,...,\dim(\chi))$ form a basis for the irrep χ of the vibration, while the collective functions $f_\mu(\alpha)$ are arbitrary and have to be determined. We still have to take into account the *H*-gauge invariance. As mentioned above, to deal with it we apply Dirac's method of imposing the constraints on the wave functions.[11] This restricts the wave functions (39) to satisfy $\Phi_i\psi(q,\alpha)=0\forall_i$, which imposes on the collective functions the conditions,

$$I_i f_\mu(\alpha)=i(T_i)^\chi_{\nu\mu}f_\nu(\alpha)\quad \forall_i . \tag{40}$$

Here $(T_i)^\chi_{\mu\nu}$ are the infinitesimal generators of *H* in the χ representation. This is the infinitesimal version of Eq. (23). This condition also holds for the discrete transformations $h\in H$, which together with (40) imply for all $h\in H$,

$$R_h f_\mu(\alpha)=\pi^\chi_{\nu\mu}(h)f_\nu(\alpha), \tag{41}$$

where R_h stands for the right action of h on the group element parameterized by α, $g(\alpha)\to g(\alpha)h$. Therefore, the collective functions $f_\mu(\alpha)$ satisfy Mackey's condition (23), i.e., they transform under the left action of *G* in the representation of *G* induced by the representation χ of *H*.[13] In other words, the states in the collective band are those of the inequivalent quantization of G/H given by the representation χ of *H* carried by the intrinsic state.

In order to diagonalize the Hamiltonian (38) we can make use of the original symmetry *G*. The induced representation of *G* under which the collective functions in (39) transform is reducible. It can be broken into irreducible components. In fact, by the Peter–Weyl theorem[13] the collective functions $f_\mu(\alpha)$ can be rewritten as,

$$f_\mu(\alpha)=C^{IMN}_\mu D^I_{MN}(\alpha), \tag{42}$$

where the sum is over all irreps *I* of *G* defined by the matrices $D^I_{MN}(\alpha)$. Under the left action of *G* each term of the sum transforms in the corresponding representation *I* of *G*. Under the right action of *H* each term in (42) transforms in the representation *I* of *G* considered as a representation of *H*. This representation of *H* is in general reducible and can be broken in irreducible pieces, $I|_H=\chi_1+\cdots+\chi_n$, where χ_i are irreps of *H*. In this decomposition, the representation χ carried by the vibrations appears a number of times we denote by $d_{I\chi}$, that is, $I|_H=d_{I\chi}\chi+$other irreps. Condition (41) implies that the linear combination (42) must be restricted to

$$f_\mu(\alpha)=C^{IMN}_k D^I_{M\mu_k}(\alpha), \tag{43}$$

where k runs over the $d_{I\chi}$ representations $\chi_k=\chi$, and μ_k indicates the μth component of irrep χ_k. This means that each representation I of G appears in the collective band a number of times equal to $d_{I\chi}$.

The collective band of the vibration is given by the restriction of the Hamiltonian (38) to the subspace of functions (39), which satisfy the condition (41). The first two terms in (38) are constant throughout the band, and the last one vanishes on physical states. The collective Hamiltonian reads,

$$\mathcal{H}_{\text{coll}}=\text{const}+\tfrac{1}{2}\mathcal{I}^{\alpha\beta}I_\alpha I_\beta-\mathcal{I}^{\alpha\beta}I_\alpha M_\beta, \tag{44}$$

where M_α stand for the restriction of the operators $(D_\alpha)_{nm}p_n q_m$ to the subspace span by the vibrational functions $\varphi_\mu(q)$, that is, $M_\alpha=\langle\varphi_\mu|(D_\alpha)_{nm}p_n q_m|\varphi_\nu\rangle$. The collective Hamiltonian is invariant under the action of G by the left, and so does not mix different representations I. Therefore, it can be diagonalized in subspaces of dimension $d_{I\chi}$. It is easy to check that it also commutes with the simultaneous action of H on the intrinsic states (J_i) and on the collective coordinates from the right (I_i). This ensures that it preserves the physical condition $\Phi_i=0$ which the states ψ of Eq. (39) satisfy.

The diagonalization of $\mathcal{H}_{\text{coll}}$ in each subspace of dimension $d_{I\chi}>1$ must be performed case by case. However, some more information can be obtained using the H-invariance. The restricted matrices M_α no longer satisfy the G algebra, but they still satisfy,

$$[J_i,M_\alpha]=iC_{i\alpha}^{\beta}M_\beta, \tag{45}$$

which means that they transform under H in the same representation as J_α. By the Wigner–Eckart theorem in the group H, they are determined by Clebsch–Gordan coefficients up to as many independent constants as irreps of H are contained in this representation. The number of irreps also gives the number of independent inertia moments in $\mathcal{I}_{\alpha\beta}$. This is as far as we can go using symmetry arguments in the general case.

If G/H is a symmetric space a considerable simplification arises. In this case, there is an involutive automorphism $\tau:\mathbf{g}\to\mathbf{g}$ $(\tau^2=id_{\mathbf{g}})$ under which the generators satisfy,[13]

$$\tau J_i\tau=J_i, \tag{46}$$

$$\tau J_\alpha\tau=-J_\alpha. \tag{47}$$

We assume that τ acts on the fields $\phi^s(\mathbf{x},t)$ as $\tau:\phi^s(\mathbf{x},t)\mapsto\tau^s[\phi(\mathbf{x},t)]$ with $\tau^s[\tilde{\phi}(\mathbf{x})]=\tilde{\phi}(\mathbf{x})$. Equation (46) states that τ commutes with J_i, so a representation of H has a definite eigenvalue of $\tau(\pm1)$, and therefore by (47) we have $M_\alpha=0$. Another simplification is that the representation of H under which J_α transform is irreducible, so there is only one moment of inertia $\mathcal{I}$. The collective Hamiltonian becomes,

$$\mathcal{H}_{\text{coll}}=\frac{1}{2\mathcal{I}}I_\alpha^2=\frac{1}{2\mathcal{I}}(I_a^2-I_i^2), \tag{48}$$

and the $d_{I\chi}$ states,

$$\psi_{IM\chi}(q,\alpha)=\varphi_\mu(q)D^I_{M\mu_k}(\alpha)\quad(k=1,...,d_{I\chi}), \tag{49}$$

are degenerate eigenstates, since I_a^2 and I_i^2 are Casimirs of G and H, respectively.

C. Landsman and Linden Hamiltonian

In the previous section we found the collective wavefunctions by treating the H-gauge invariance by means of the Dirac method of imposing the constraints on the states. However, the dynamical consequences of the Mackey condition (41) on the collective functions are best understood by fixing the H-gauge symmetry.

Let us make the natural gauge choice of setting the collective coordinates related to motion on the subgroup $\alpha^i=0$. This fixes the H-gauge symmetry generated by Φ_i, i.e., $\{\alpha^i,\Phi_j\}\approx\delta^i_j$. Proceeding as before, we eliminate the pairs $(\alpha^i,\mathcal{P}_i)$ by solving $\mathcal{P}_i$ from the equations $\Phi_i=0$. This gives

$$\mathcal{P}_i=(D_i)_{nm}p_nq_m, \tag{50}$$

where we have used that $\Theta^j_i(0)=\delta^i_j$. Since the vibrations fall into irreps of H, if we restrict to a given vibrational state classified by χ, the above equation reduces to,

$$\mathcal{P}_i=\langle\varphi_\mu|(D_i)_{nm}p_nq_m|\varphi_\nu\rangle=-i(T_i)^\chi_{\mu\nu}, \tag{51}$$

where $(T_i)^\chi_{\mu\nu}$ are the generators of H in the irrep χ. Replacing in the collective operators I_α we get,

$$I_\alpha=\Theta^\beta_\alpha(\xi)\mathcal{P}_\beta+i\Theta^\beta_\alpha(\xi)\zeta^i_\beta(\xi)(T_i)^\chi_{\mu\nu}, \tag{52}$$

where $\xi\in G/H_0$ (i.e., the coordinates $\alpha^{a=1,\ldots,\dim(G/H)}$), being H_0 the identity component of H. In this gauge the collective functions f_μ become the functions $F_\mu(\xi)$ of Sec. III, and the collective Hamiltonian (44) becomes,

$$\mathcal{H}_{\text{coll}}=\text{const}-\tfrac{1}{2}g^{\alpha\beta}(\nabla_\alpha+A_\alpha)(\partial_\beta+A_\beta)-\tfrac{1}{2}g^{\alpha\beta}[(\nabla_\alpha+A_\alpha)B_\beta+B_\alpha(\partial_\beta+A_\beta)]. \tag{53}$$

The second term corresponds to Landsman–Linden Hamiltonian (28), with the metric on G/H_0 given by $g_{\alpha\beta}(\xi)=\mathcal{I}_{\gamma\delta}\zeta^\gamma_\alpha(\xi)\zeta^\delta_\beta(\xi)$, and the H_0-connection by,

$$(A_\alpha)_{\mu\nu}=-i\zeta^i_\alpha(T_i)^\chi_{\mu\nu}. \tag{54}$$

The third term in (53) gives the coupling to an extra background field,

$$(B_\alpha)_{\mu\nu}=-i\zeta^\beta_\alpha(M_\beta)_{\mu\nu}, \tag{55}$$

which comes from the ''Coriolis'' terms in (44). In general, for real representations, this is an SO(dimχ) connection, and is similar to the induced connections studied in Ref. 14.

The H_0-connection ensures that Hamiltonian (53) applied to functions independent of α^i gives the same result as Hamiltonian (44) acting on functions over G which satisfy Mackey condition (40). When $H_0\neq H$ the functions $F_\mu(\xi)$ are still restricted by condition (41) for the discrete elements of H. This restriction can be lifted including a pure gauge connection which associates a discrete element of H with each nontrivial path in $\Pi_1(G/H)$ (holonomy factors). The relation between inequivalent quantizations and holonomy factors in the path integral is discussed in Ref. 15.

V. CONCLUSIONS

We have discussed the collective bands of intrinsic states found when quantizing around a soliton solution with moduli space isomorphic to G/H. The result is that the collective band of an intrinsic vibrational state realizes an inequivalent coset space quantization given by the representation of H under which the intrinsic state transforms. The collective Hamiltonian is that of Landsman and Linden,[2] which describes free motion on G/H coupled to a background H-gauge field. Besides, there may be other background gauge fields coming from the Coriolis terms. The extra degrees of freedom associated with the nontrivial quantizations are given by the intrinsic coordinates. In this way, we have given a physical example of the inequivalent quantizations studied in Refs. 1–3.

Some of the consequences for the collective quantization of a soliton with moduli space G/H are the following. The states in the collective band of a vibration carrying the representation χ of H are classified by the irreducible representations I of the symmetry group G. Each irrep I appears in the band as many times as χ is contained in its decomposition into irreps of H. This also gives to lowest order the dimension of the collective Hamiltonian matrix within this subspace, from which the energies of the states are obtained. Determining the collective bands is analogous to calculating the rotational spectra of polyatomic molecules,[16] and our results can be understood as a generalization of this problem to a general symmetry group.

This work may be of interest for obtaining the spins and isospins bands of the ground state of multiskyrmions[17,18] and their excited states, which have been found for topological charges $B=2$ and $B=4$ by Barnes *et al.*[19,20] The symmetry group of the Skyrme model is the direct product of the spatial and isospatial rotations and the combined parity (without considering spatial translations) $G=\mathrm{SO}(3)_S\times\mathrm{SO}(3)_I\times P$. As described in Ref. 21, the symmetry group of a Skyrmion $H\subset G$ is given by pairs $(h,D(h))$, where h is an element of the spatial group $\mathrm{O}(3)_S$, and $D(h)$ an element of the isospin group $\mathrm{O}(3)_I$. The mapping $D:h\mapsto D(h)$ is a three-dimensional real representation of H, and $\det h=\det D(h)$. For example, for the $B=2$ case $H=D_\infty$ (not considering parity). The ground state is in the non trivial one-dimensional representation Σ^- if the Skyrmion is quantized as a fermion. Therefore, the lowest allowed state of the band are $(I=1,\ S=0)$ or $(I=0,\ S=1)$, since they are the lowest irreps of G containing Σ^- in its decomposition. Similarly, the other states of the band can be obtained. The $B=2$ case has already been considered in Ref. 22, but we expect the more systematic treatment presented here will be useful in more complicated situations, as for $B>2$, where the subgroup H is discrete.[23]

ACKNOWLEDGMENTS

We are grateful to N. S. Manton, D. R. Bes, B. J. Schroers, and N. N. Scoccola for enlightening discussions. J.P.G. has been supported by Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Partial support from Fundación Antorchas is acknowledged.

[1] G. W. Mackey, *Induced Representations of Groups and Quantum Mechanics* (Benjamin, New York, 1969).
[2] N. P. Landsman and N. Linden, Nucl. Phys. B **365**, 121 (1991).
[3] D. McMullan and I. Tsutsui, Ann. Phys. **237**, 269 (1995), hep-th/9308027.
[4] E. Witten, Nucl. Phys. B **223**, 422 (1983); **223**, 433 (1983).
[5] T. H. R. Skyrme, Proc. R. Soc. London **260**, 127 (1961); Nucl. Phys. **31**, 556 (1962).
[6] J. P. Garrahan, M. Kruczenski, and D. R. Bes, Phys. Rev. D **53**, 7176 (1996), hep-th/9603041.
[7] A. Hosoya and K. Kikkawa, Nucl. Phys. B **101**, 271 (1975).
[8] D. R. Bes and J. Kurchan, *The Treatment of Collective Coordinates in Many-Body Systems, Lecture Notes in Physics* (World Scientific, Singapore, 1990), Vol. 34.
[9] J. Alfaro and P. H. Damgaard, Ann. Phys. **202**, 398 (1990).
[10] H. J. W. Müller-Kirsten and J.-z. Zhang, Phys. Rev. D **50**, 6531 (1994); Phys. Lett. B **339**, 65 (1994).
[11] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, NJ, 1992).
[12] A. Perelomov, *Generalized Coherent States and Their Applications* (Springer-Verlag, Berlin, 1986).
[13] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (World Scientific, Singapore, 1986).
[14] P. Maraner, J. Phys. A **28**, 2939 (1995), hep-th/9409080.
[15] S. Tanimura and I. Tsutsui, Ann. Phys. **258**, 137 (1997), hep-th/9609089.
[16] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics, Nonrelativistic Theory* (Pergamon, Oxford, 1965).
[17] E. Braaten, S. Townsend, and L. Carson, Phys. Lett. B **391**, 150 (1990).
[18] R. A. Battye and P. M. Sutcliffe, Phys. Rev. Lett. **79**, 363 (1997), hep-th/9702081.
[19] C. Barnes, W. K. Baskerville, and N. Turok, Phys. Lett. B **411**, 180 (1997), hep-th/9704028.
[20] C. Barnes, W. K. Baskerville, and N. Turok, Phys. Rev. Lett. **79**, 367 (1997), hep-th/9704012.
[21] R. A. Leese and N. S. Manton, Nucl. Phys. A **572**, 575 (1994).
[22] R. A. Leese, N. S. Manton, and B. J. Schroers, Nucl. Phys. **442**, 228 (1995), hep-th/9502405.
[23] P. Irwin, ''Zero mode quantization of multi-Skyrmions,'' hep-th/9804142.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Ground states of a model in nonrelativistic quantum electrodynamics. I

Fumio Hiroshima[a)]
TU-München, Zentrum Mathematik, Arcis Strasse 21, D-80290 München, Germany

(Received 29 March 1999; accepted for publication 6 July 1999)

The system of a one charged nonrelativistic particle with external potentials minimally coupled to a massless quantized radiation field is considered. An ultraviolet cutoff is imposed on the quantized radiation field and the charged particle has spin 1/2. The class of external potentials considered in this paper contains *Coulomb potentials*. It is shown that the ground states of the system exist provided that a coupling constant is in a region. © *1999 American Institute of Physics.* [S0022-2488(99)03111-4]

I. INTRODUCTION

In this paper and Ref. 1 we are concerned with the ground states of a model in nonrelativistic quantum electrodynamics. In the present paper we prove the existence of the ground states of the model which describes the system of a nonrelativistic spin 1/2 particle minimally coupled to a massless quantized radiation field with external potentials (the Pauli–Fierz model[2]). The class of the external potentials considered in this paper includes Coulomb potentials. The quantized radiation field is quantized in the Coulomb gauge and has an ultraviolet cutoff. The Pauli–Fierz model interprets the physical phenomenon "Lamb shift."[3–7]

In the past few decades, several articles have been devoted to the study of the Pauli–Fierz model and has provoked a great deal of controversy.[8–14] A great deal of effect has been made on showing the existence of ground states of related models.[11,14–19]

In the previous paper[14] we have proved the existence of the ground states of the Pauli–Fierz model with external potentials V such as $V=x^2+V_0$, where V_0 is infinitesimally small with respect to the Laplacian. Hence this derives us to a question whether the ground states of the system with Coulomb potentials exist or not.[13] The goal of this paper is to prove the existence of the ground states of the Pauli–Fierz model with a general class of external potentials containing Coulomb potentials.

We shall show the strategy of this paper. A good place to start is to survey Ref. 14. Let H be the Hamiltonian of the system, which is a self-adjoint operator acting in a Hilbert space $\mathcal{H}$. One first introduces artificial mass $m>0$ of photon and define H_m. Second, we define H_m^a by lattice approximation of H_m with lattice length $2\pi/a$.[11,14,16,17,20] We call the bottom of the spectrum of self-adjoint operator T by the ground state energy of T. It is established that the spectrum of H_m^a is purely discrete in the neighborhood of its ground state energy. Then the ground states of H have been constructed in Ref. 14 through (1) a norm resolvent convergence of H_m^a to H_m as $a\to\infty$ (thus one knows that the ground states of H_m exist); (2) nonzero weak limit of a ground state of H_m as $m\to 0$ (the limit is just a ground state of H). To show (1) in Ref. 14 it plays an important role that particles are confined by external potentials such as $V=x^2+V_0$. Then in our case it seems, however, to be difficult to show (1). In order to avoid such sufferings we shall find a family of unitary transformations U_a. That goes to the very heart of solving our problem. We shall show that, instead of H_m^a, the operator $U_aH_m^aU_a^{-1}$ converges to an operator which is unitarily equivalent to H_m as $a\to\infty$ in the norm resolvent sense. Then we see that H_m has a ground state. Finally in the similar way as that of Ref. 14 we prove the existence of the ground states of H.

[a)]Electronic mail: hiro@mathematik.tu-muenchen.de

This paper is organized as follows: In Sec. II, we define the Pauli–Fierz model and we give Hypotheses 1, 2, and 3. Section III A is devoted to defining the lattice approximated Hamiltonians, gives Hypothesis 4 and considers the spectrum in the neighborhood of their ground state energy. In Sec. III B, we introduce a family of unitary transformations. Section III C offers the key to proving existence of ground states; we shall show convergence of operators in a norm resolvent sense. In Sec. III D we give Hypothesis 5 and some binding. In Sec. III E we give Hypothesis 6 and state the main theorem (Theorem 3.14). In Sec. III F we show a weak convergence of ground states.

II. DEFINITION OF A HAMILTONIAN

In this section we give a quick sketch of definition of the Pauli–Fierz model and carefully show some inequalities.

A. Quantized radiation fields

The Boson Fock space $\mathcal{F}$ over $W:=L^2(\mathbb{R}^3)\oplus L^2(\mathbb{R}^3)$ is defined by $\mathcal{F}:=\oplus_{N=0}^{\infty}[\otimes_s^N W]$, where $\otimes_s^N W$, $N\geqslant 1$, denotes the N-fold symmetric tensor product of W and $\otimes_s^0 W:=\mathbb{C}$. We denote by $\mathcal{F}_{\rm fin}$ the finite particle subspace (Ref. 21, X.7). The bare vacuum vector $\Omega\in\mathcal{F}$ is defined by $\Omega:=\{1,0,0,...\}$. We denote by $(\cdot,\cdot)_{\mathcal{K}}$ the scalar product on Hilbert space $\mathcal{K}$ and by $\|\cdot\|_{\mathcal{K}}$ the associated norm. Scalar product $(f,g)_{\mathcal{K}}$ is linear in g and antilinear in f. Unless confusion arising we omit $\mathcal{K}$ in $\|\cdot\|_{\mathcal{K}}$ and $(\cdot,\cdot)_{\mathcal{K}}$, respectively. We denote by $a^\dagger(F)$ and $a(F)$, $F\in W$, the smeared creation operator and the smeared annihilation operator defined on $\mathcal{F}_{\rm fin}$, respectively. In particular, for notational convenience, we define $a^{\dagger r}$ and a^r, $r=1,2$, by $a^{\dagger 1}(f):=a^\dagger(f\oplus 0)$, $a^{\dagger 2}(f):=a^\dagger(0\oplus f)$ and $a^1(f):=a(f\oplus 0)$, $a^2(f):=a(0\oplus f)$, $f\in L^2(\mathbb{R}^3)$, respectively. They satisfy canonical commutation relations on $\mathcal{F}_{\rm fin}$,

$$[a^r(f),a^{\dagger s}(g)]=\delta_{rs}(\overline{f},g), \tag{2.1}$$

$$[a^r(f),a^s(g)]=[a^{\dagger r}(f),a^{\dagger s}(g)]=0,\quad r,s=1,2, \tag{2.2}$$

$$(a^r(f)\Psi,\Phi)=(\Psi,a^{\dagger r}(\overline{f})\Phi),\quad \Phi,\Psi\in\mathcal{F}_{\rm fin}, \tag{2.3}$$

where $\overline{f}$ is the complex conjugate of f. We formally write $a^{\sharp r}(f)=\int a^{\sharp r}(k)f(k)dk$, $r=1,2$, where $a^{\sharp r}$ denotes a^r or $a^{\dagger r}$. For a contraction operator T on W, the second quantization of T is denoted by $\Gamma(T)$ (Ref. 21, X.7). In particular, for a nonnegative self-adjoint operator h in W, $\Gamma(e^{-th}),t\geqslant 0$, is a strongly continuous symmetric one-parameter semigroup on $\mathcal{F}$. Hence we can define a nonnegative self-adjoint operator $d\Gamma(h)$ in $\mathcal{F}$ by $\Gamma(e^{-th})=e^{-td\Gamma(h)},t\geqslant 0$. Note that $d\Gamma(h)\Omega=0$. The free Hamiltonian H_f in $\mathcal{F}$ is defined by $H_f:=d\Gamma(\omega\oplus\omega)$, where ω is the multiplication operator by $\omega(k):=|k|$ ($|k|$ is the three-dimensional Euclidean norm of k). Besides the number operator N is defined by $N:=d\Gamma(I\oplus I)$. We set

$$\mathcal{M}:=\{f:\text{Borel measureable}\,|\,\|f\|_n:=\|\omega^{n/2}f\|_{L^2(R^3)}<\infty\}.$$

The following relative bounds are well known:

$$\|a^{\dagger r}(f)\Psi\|\leqslant\|f\|_{-1}\|H_f^{1/2}\Psi\|+\|f\|_0\|\Psi\|,\quad f\in\mathcal{M}_{-1}\cap\mathcal{M}_0, \tag{2.4}$$

$$\|a^r(f)\Psi\|\leqslant\|f\|_{-1}\|H_f^{1/2}\Psi\|,\quad f\in\mathcal{M}_{-1}\cap\mathcal{M}_0, \tag{2.5}$$

$$\|a^{\dagger r}(f)\Phi\|\leqslant\|f\|_0\|(N+I)^{1/2}\Phi\|,\quad f\in\mathcal{M}_0, \tag{2.6}$$

$$\|a^r(f)\Phi\|\leqslant\|f\|_0\|N^{1/2}\Phi\|,\quad f\in\mathcal{M}_0, \tag{2.7}$$

where $\Psi\in D(H_f^{1/2})$ and $\Phi\in D(N^{1/2})$. With an ultraviolet cutoff ρ, we define a quantized radiation field $A_\mu(x)$ and a quantized magnetic field $B_\mu(x),\mu=1,2,3$, by

$$A_\mu(x) := A_\mu(\rho,x) := \sum_{r=1}^{2} \int \frac{dk}{\sqrt{2\omega(k)}} \{\rho(k) e_\mu^r(k) e^{-ikx} a^{\dagger r}(k) + \rho(-k) e_\mu^r(k) e^{ikx} a^r(k)\},$$

$$B_\mu(x) := B_\mu(\rho,x)$$
$$:= -i \sum_{r=1}^{2} \int \frac{dk}{\sqrt{2\omega(k)}} \{\rho(k)(k\times e^r(k))_\mu e^{-ikx} a^{\dagger r}(k) + \rho(-k)(-k\times e^r(k))_\mu e^{ikx} a^r(k)\},$$

where $e^r = (e_1^r, e_2^r, e_3^r)$, $r = 1, 2$, are polarization vectors which satisfy $e^r(k) e^s(k) = \delta_{rs}$, $ke^r(k) = 0$, $a.e. k \in \mathbb{R}^3$. We set $A_\mu := A_\mu(\rho) := A_\mu(\rho,0)$ and $B_\mu := B_\mu(\rho) = B_\mu(\rho,0)$, $\mu = 1,...,d$.

B. Definition of a Hamiltonian

The Hilbert space $\mathcal{H}$ of the system is given by

$$\mathcal{H} := L^2(\mathbb{R}^3;\mathbb{C}^2) \otimes \mathcal{F} := \mathcal{H}_p \otimes \mathcal{F}.$$

We work in units which is $c = \hbar = 1$ and put the mass of the particle one. Then the Hamiltonian H of the system is defined by

$$H = (1/2)\{\boldsymbol{\sigma}\cdot(\mathbf{P}\otimes I - e\mathbf{A}(x))\}^2 + V\otimes I + I \otimes H_f,$$

where $\mathbf{A}(x) := (A_1(x), A_2(x), A_3(x))$, $\mathbf{P} := (P_1, P_2, P_3) := (-i\nabla_{x_1}, -i\nabla_{x_2}, -i\nabla_{x_3})$, V is an external potential, and $\boldsymbol{\sigma} := (\sigma_1, \sigma_2, \sigma_3)$ defined by

$$\sigma_1 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 := \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

We immediately see that on $\mathcal{E} := C_0^\infty(\mathbb{R}^3)\hat{\otimes}\mathcal{F}_{\rm fin}$,

$$H = H_p \otimes I + I \otimes H_f + eH_I + e^2 H_{II} + eH_{III},$$

where $H_p := -(1/2)\Delta + V$, $H_I := -(\mathbf{P}\otimes I)\cdot\mathbf{A}(x)$, $H_{II} := (1/2)\mathbf{A}(x)\cdot\mathbf{A}(x)$, $H_{III} := (1/2)\boldsymbol{\sigma}\cdot\mathbf{B}(x)$. Unless confusion arising we abbreviate both $X\otimes I$ and $I\otimes X$ as X in what follows. We introduce Hypotheses 1 and 2.

Hypothesis 1: The ultraviolet cutoff ρ is such that $\rho \in \mathcal{M}_{-2}\cap\mathcal{M}_{-1}\cap\mathcal{M}_0\cap\mathcal{M}_1$ and that $\overline{\rho(k)} = \rho(-k)$.

Hypothesis 2: The external potential V is such that H_p is essentially self-adjoint on $C_0^\infty(\mathbb{R}^3)$ and that there exist constants a and b such that

$$\|-(1/2)\Delta f\| \leq a\|H_p f\| + b\|f\|^2, \quad f \in C_0^\infty(\mathbb{R}^3). \tag{2.8}$$

Moreover there exists $\Sigma \leq \infty$ such that $(-\infty, \Sigma] \cap \sigma(H_p) \subset \sigma_d(H_p)$, where $\sigma(T)$ and $\sigma_d(T)$ denote the spectrum and the discrete spectrum of T, respectively.

Note that (2.8) implies that, with the same a as in (2.8),

$$(f, -(1/2)\Delta f) \leq a(f, H_p f) + b'\|f\|, \quad f \in C_0^\infty(\mathbb{R}^3),$$

where b' is a positive constant. We fix a, b, b' and Σ throughout this paper. We have two standard examples in mind. One choice is (1) $V \in L^2(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$ with isolated ground state energy, e.g., $V(x) = -1/|x|$. The other is (2) $V(x)\to\infty$ as $|x|\to\infty$, e.g., $V(x) = |x|^2$. For the first case (1), the essential spectrum of H_p has continuum edge, i.e., $\Sigma < \infty$. For the latter case (2), H_p has purely discrete spectrum, i.e., one can take $\Sigma = \infty$. Let

$$r_n := \|\rho\|_n.$$

We have, by (2.4) and (2.5), for $\Psi \in \mathcal{E}$,

$$\|H_I\Psi\| \leq A\|H_f\Psi\| + B\|H_p\Psi\| + C\|\Psi\|, \tag{2.9}$$

$$\|H_{II}\Psi\| \leq D\|H_f\Psi\| + E\|\Psi\|, \tag{2.10}$$

$$\|H_{III}\Psi\| \leq F\|H_f\Psi\| + G\|\Psi\|, \tag{2.11}$$

where

$$\begin{gathered} A := 6r_{-2}, \quad B := 6a(2r_{-2}+r_{-1}), \quad C := 6b(2r_{-2}+r_{-1}) + 3r_{-1}, \\ D = E := 6(r_{-2}+r_{-1})^2 + 3k(r_{-2}+r_{-1})(r_0+r_1), \\ F := 6r_0/\sqrt{2}, \;\; G := 3r_1/\sqrt{2}, \quad k := (2\pi)^{-5/2}\int_0^\infty \sqrt{\lambda}/(\lambda+1)\,d\lambda. \end{gathered} \tag{2.12}$$

We define

$$e_0 := \frac{-(A+B+F) + \sqrt{(A+B+F)^2 + 4D}}{2D}.$$

We set $g[T] := \inf \sigma(T)$. In particular we set $\Sigma_0 := g[H_p]$. Hypothesis 3 is as follows:

Hypothesis 3: The coupling constant e is such that $0 \leq e < e_0$.

Proposition 2.1 (Ref. 22): We assume that Hypotheses 1, 2, and 3 hold. Then H is self-adjoint on $D(H_p)\cap D(H_f)$, bounded below and essentially self-adjoint on any core of $H_0 := H_p + H_f$.

Proof: Note that $\|H_f\Psi\| \leq \|H_0\Psi\| + |\Sigma_0|\|\Psi\|$ and $\|H_p\Psi\| \leq \|H_0\Psi\| + 2|\Sigma_0|\|\Psi\|$. Then, by (2.9), (2.10), and (2.11), we see that for $\Psi \in \mathcal{E}$,

$$\|H_I\Psi\| \leq (A+B)\|H_0\Psi\| + C\|\Psi\| + (A+2B)|\Sigma_0|\|\Psi\|, \tag{2.13}$$

$$\|H_{II}\Psi\| \leq D\|H_0\Psi\| + E\|\Psi\| + D|\Sigma_0|\|\Psi\|, \tag{2.14}$$

$$\|H_{III}\Psi\| \leq F\|H_0\Psi\| + G\|\Psi\| + F|\Sigma_0|\|\Psi\|. \tag{2.15}$$

Hence Kato–Rellich Theorem (Ref. 21, Theorem X.12) yields the desired results. □

Remark 2.2: Essential self-adjointness of H for arbitrary coupling constant $e \in \mathbb{R}$ is studied in Ref. 23.

We shall estimate the ground state energy of H. Let $\Psi \in \mathcal{E}$. We have

$$\begin{aligned} |(\Psi, H_I\Psi)| &\leq A'(\Psi, H_f\Psi) + B'(\Psi, H_p\Psi) + C'\|\Psi\|^2 \\ &\leq (A'+B')(\Psi, H_0\Psi) + (C' - A'\Sigma_0)\|\Psi\|^2, \end{aligned} \tag{2.16}$$

$$|(\Psi, H_{II}\Psi)| \leq D'(\Psi, H_f\Psi) + E'\|\Psi\|^2 \leq D'(\Psi, H_0\Psi) + (E' - D'\Sigma_0)\|\Psi\|^2, \tag{2.17}$$

$$|(\Psi, H_{III}\Psi)| \leq F'(\Psi, H_f\Psi) + G'\|\Psi\|^2 \leq F'(\Psi, H_0\Psi) + (G' - F'\Sigma_0)\|\Psi\|^2, \tag{2.18}$$

where

$$\begin{gathered} A' := 6r_{-2}/\sqrt{2}, \quad B' := 6a\sqrt{2}r_{-2}, \quad C' := 6b\sqrt{2}r_{-2}, \quad D' := 3(r_{-2}+r_{-1})r_{-2}, \\ E' := (3/2)(r_{-2}+r_{-1})r_{-1}, \quad F' = G' = 6r_0/(2\sqrt{2}). \end{gathered} \tag{2.19}$$

Hence we see that

$$(1-eA'-eB'-e^2D'-eF')(\Psi,H_0\Psi)+(eA'+e^2D'+eF')\Sigma_0\|\Psi\|^2 \\ -(eC'+e^2E'+eG')\|\Psi\|^2\leqslant(\Psi,H\Psi). \tag{2.20}$$

Moreover we have

$$(\Psi,H\Psi)\leqslant(1+eA'+eB'+e^2D'+eF')(\Psi,H_0\Psi) \\ +(eC'+e^2E'+eG')\|\Psi\|^2-(eA'+e^2D'+eF')\Sigma_0\|\Psi\|^2. \tag{2.21}$$

Since H is essentially self-adjoint on any core of H_0 for coupling constant $0\leqslant e<e_0$, in view of (2.20) and (2.21), we get, for $0\leqslant e<e_0$,

$$|g[H]-\Sigma_0|\leqslant\Delta(\rho), \tag{2.22}$$

where

$$\Delta(\rho):=e(B'\Sigma_0+C')+e^2E'+eG' \\ =e6\sqrt{2}r_{-2}(a\Sigma_0+b')+e^23(2r_{-2}+r_{-1})r_{-1}/2+e3r_0/\sqrt{2}.$$

Note that $a\Sigma_0+b'\geqslant 0$.

III. EXISTENCE OF GROUND STATES

Before we come on to a closer analysis of the existence of the ground states of H, let us pause here to look briefly at our strategy. We shall introduce three parameters m, a, Λ, and define self-adjoint operators $H^a_{m\Lambda}$, $H_{m\Lambda}$, and H_m, respectively. In addition to this we construct unitary operators U and U_a in $\mathcal{H}$, and define $\hat{H}_{m\Lambda}:=UH_{m\Lambda}U^{-1}$ and $\hat{H}^a_{m\Lambda}:=U_aH^a_{m\Lambda}U_a^{-1}$, respectively.

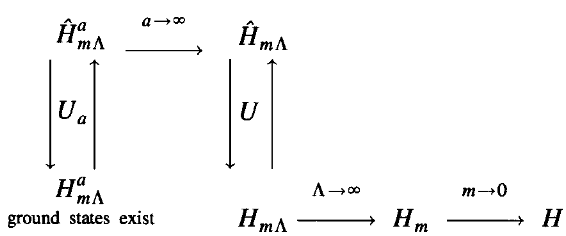

First, we shall show that $H^a_{m\Lambda}$ has purely discrete spectrum in the neighborhood of $g[H^a_{m\Lambda}]$. Second, we shall show the norm resolvent convergence of $\hat{H}^a_{m\Lambda}$ to $\hat{H}_{m\Lambda}$ as $a\to\infty$. Then by unitary equivalences between $H^a_{m\Lambda}$, and $\hat{H}^a_{m\Lambda}$, and between $H_{m\Lambda}$ and $\hat{H}_{m\Lambda}$, we can see that $H_{m\Lambda}$ has also purely discrete spectrum in the neighborhood of $g[H_{m\Lambda}]$. (See diagram 1, where $a\to\infty$ and $\Lambda\to\infty$ are in the sense of norm resolvent and $m\to 0$ in the sense of strong resolvent.) Finally the similar arguments on $\Lambda\to\infty$ and $m\to 0$ as those in Ref. 16 lead us to the final goal; we show the existence of the ground states of H. Remark that it is *difficult* to show the norm resolvent convergence of $H^a_{m\Lambda}\to H_{m\Lambda}$ as $a\to\infty$ directly with potentials V in Hypothesis 2.

A. Lattice approximation

Lattice approximation is essentially contained in Ref. 20; to make this paper as self-contained as possible we include a version of the proof here for the convenience of the reader. Define the set of lattice points by $\Gamma_a:=\{k=(k_1,k_2,k_3)|k_\mu=2\pi n_\mu/a,n_\mu\in\mathbb{Z},\mu=1,2,3\}$. Let $l_2(\Gamma_a)$ be the set of l_2 sequences over Γ_a. We define $\mathcal{F}_a$ as $\mathcal{F}_a:=\oplus_{n=0}^\infty\otimes_s^n[l_2(\Gamma_a)\oplus l_2(\Gamma_a)]$. By the map $l_2(\Gamma_a)\ni\{a_l\}_{l\in\Gamma_a}\mapsto(a/2\pi)\Sigma_{l\in\Gamma_a}a_l\chi_{\Gamma(l,a)}(\cdot)\in L^2(\mathbb{R}^3)$, we identify $l_2(\Gamma_a)$ with a subspace in $L^2(\mathbb{R}^3)$, where $\chi_{\Gamma(l,a)}(\cdot),l\in\Gamma_a$, is the characteristic function of

$$\Gamma(l,a):=[l_1,l_1+2\pi/a)\times[l_2,l_2+2\pi/a)\times[l_3,l_3+2\pi/a)\subset\mathbb{R}^3.$$

Then we identify $l_2(\Gamma_a)\oplus l_2(\Gamma_a)$ with the subspace of W and regard $\mathcal{F}_a$ as the closed subspace of $\mathcal{F}$. For $0<m$ and $0<a$, we define nonnegative self-adjoint operators by

$$H_{fm}^a := d\Gamma((\omega_a+m)\oplus(\omega_a+m)),$$

$$H_{fm} := d\Gamma((\omega+m)\oplus(\omega+m)),$$

where $\omega_a(k):=\omega(k_a)$ and $k_a:=(k_{1,a},k_{2,a},k_{3,a})$ is a step function, i.e., $k_{\mu,a}=2\pi n/a$, if $k_\mu \in[2\pi n/a,2\pi n/a+2\pi/a), n\in\mathbb{Z}$. The operator H_{fm}^a is reduced by $\mathcal{F}_a$ and $H_{fm}^a\lceil_{F_a}$ has purely discrete spectrum. We define

$$A_\mu^a(x):=\sum_{r=1}^2 \int dk \sum_{l\in\Gamma_a} X_{\Gamma(l,a)}(k)\left\{\frac{\rho(l)e^{-ilx}e_\mu^r(l)}{\sqrt{2\omega(l)}}a^{\dagger r}(k)+\frac{\rho(-l)e^{ilx}e_\mu^r(l)}{\sqrt{2\omega(l)}}a^r(k)\right\},$$

$$B_\mu^a(x):=-i\sum_{r=1}^2 \int dk \sum_{l\in\Gamma_a} X_{\Gamma(l,a)}(k) \times\left\{\frac{\rho(l)e^{-ilx}(l\times e^r(l))_\mu}{\sqrt{2\omega(l)}}a^{\dagger r}(k)+\frac{\rho(-l)e^{ilx}(-l\times e^r(l))_\mu}{\sqrt{2\omega(l)}}a^r(k)\right\}.$$

Thus H_m^a and H_m are given by

$$H_m^a := H_p+H_{fm}^a+eH_I^a+e^2H_{II}^a+eH_{III}^a,$$

$$H_m := H_p+H_{fm}+eH_I+e^2H_{II}+eH_{III},$$

where H_I^a, H_{II}^a, and H_{III}^a are defined with $A_\mu(x)$ and $B_\mu(x)$ in H_I, H_{II}, and H_{III} replaced by $A_\mu^a(x)$ and $B_\mu^a(x)$, respectively. Let X_Λ be the characteristic function of $\{k\in\mathbb{R}^3 \| k|<\Lambda\}$. We set $\rho_\Lambda:=X_\Lambda\rho$. We define $H_{m\Lambda}^a$ by H_m^a with ρ replaced by ρ_Λ. We set $A_\mu^a:=A_\mu^a(\rho,0)$, $B_\mu^a:=B_\mu^a(\rho,0)$, $A_{\Lambda\mu}^a:=A_\mu^a(\rho_\Lambda,0)$, $B_{\Lambda\mu}^a:=B_\mu^a(\rho_\Lambda,0)$, $A_{\Lambda\mu}:=A_\mu(\rho_\Lambda,0)$ and $B_{\Lambda\mu}:=B_\mu(\rho_\Lambda,0)$.

Lemma 3.1: We assume that Hypotheses 1, 2, and 3 hold. Then (1) H_m is self-adjoint on $\mathcal{D}_m:=D(H_p)\cap D(H_{fm})$, bounded below and essentially self-adjoint on any core of

$$H_{0,m}:=H_p+H_{fm};$$

(2) for sufficiently large a, $H_{m\Lambda}^a$ is self-adjoint on $\mathcal{D}_m$, bounded below and essentially self-adjoint on any core of

$$H_{0,m}^a:=H_p+H_{fm}^a.$$

Proof: Proof is the same as that of Proposition 2.1. We omit it. □

Let A', B', C', D', E', F', and G' be in (2.19), and $A'_{a\Lambda}$, $B'_{a\Lambda}$, $C'_{a\Lambda}$, $D'_{a\Lambda}$, $E'_{a\Lambda}$, $F'_{a\Lambda}$, and $G'_{a\Lambda}$, are defined by A', B', C', D', E', F', G' with ρ replaced by $\rho_{a\Lambda}:=\Sigma_{l\in\Gamma_l}X_{\Gamma(l,a)}(\cdot)\rho_\Lambda(l)$, respectively.

Lemma 3.2: Assume that Hypotheses 1, 2, and 3 hold. Let $\Psi\in\mathcal{E}$. Then the following inequalities hold:

$$(1)\quad (1-eA'-eB'-e^2D'-eF')(\Psi,H_{0,m}\Psi)+(eA'+e^2D'+eF')\Sigma_0\|\Psi\|^2 -(eC'+e^2E'+eG')\|\Psi\|^2\leqslant(\Psi,H_m\Psi); \tag{3.1}$$

$$(2)\quad (\Psi,H_m\Psi)\leqslant(1+eA'+eB'+e^2D'+eF')(\Psi,H_{0,m}\Psi) +(eC'+e^2E'+eG')\|\Psi\|^2-(eA'+e^2D'+eF')\Sigma_0\|\Psi\|^2; \tag{3.2}$$

(3) $(1-eA'_{a\Lambda}-eB'_{a\Lambda}-e^2D'_{a\Lambda}-eF'_{a\Lambda})(\Psi,H^a_{0,m}\Psi)+(eA'_{a\Lambda}+e^2D'_{a\Lambda}+eF'_{a\Lambda})\Sigma_0\|\Psi\|^2$

$$-(eC'_{a\Lambda}+e^2E'_{a\Lambda}+eG'_{a\Lambda})\|\Psi\|^2\leqslant(\Psi,H^a_{m\Lambda}\Psi); \tag{3.3}$$

(4) $(\Psi,H^a_{m\Lambda}\Psi)\leqslant(1+eA_{a\Lambda}+eB_{a\Lambda}+e^2D_{a\Lambda}+eF_{a\Lambda})(\Psi,H^a_{0,m}\Psi)$

$$+(eC_{a\Lambda}+e^2E_{a\Lambda}+eG_{a\Lambda})\|\Psi\|^2-(eA_{a\Lambda}+e^2D_{a\Lambda}+eF_{a\Lambda})\Sigma_0\|\Psi\|^2. \tag{3.4}$$

In particular,

$$|\Sigma_0-g[H_m]|\leqslant\Delta(\rho), \tag{3.5}$$

$$|\Sigma_0-g[H^a_{m\Lambda}]|\leqslant\Delta(\rho_{a\Lambda}). \tag{3.6}$$

Proof: Note that for sufficient large a, $(\Psi,H_0\Psi)\leqslant(\Psi,H^a_{0,m}\Psi)$. Then by (2.16), (2.17), and (2.18), one sees (3.3) and (3.4). (3.1) and (3.2) are proved quite similarly. □

We introduce Hypothesis 4:

Hypothesis 4: The coupling constant e is such that $2\Delta(\rho)/(1-eB')<\Sigma-\Sigma_0$.

Lemma 3.3: We assume that Hypotheses 1, 2, 3, and 4 hold. Let a be sufficiently large and m be such that

$$0<m\leqslant(1-eB')(\Sigma-\Sigma_0)-2\Delta(\rho). \tag{3.7}$$

Then $[g[H^a_{m\Lambda}],g[H^a_{m\Lambda}]+m)\subset\sigma_d(H^a_{m\Lambda})$.

Remark 3.4: By (3.6) and a limiting argument, for sufficiently large a, (3.7) implies that

$$0<(1-eB'_{a\Lambda})(\Sigma-\Sigma_0)+(\Sigma_0-g[H^a_{m\Lambda}])-\Delta(\rho_{a\Lambda})-m. \tag{3.8}$$

Proof: For notational brevity, we put $H_a:=H^a_{m\Lambda}$ and $g_a:=g[H^a_{m\Lambda}]$. Moreover set $\overline{H_p}:=H_p-\Sigma_0$. Let $E_T(\mathcal{B})$ be the spectral projection of operator T on Borel set $\mathcal{B}\subset\mathbb{R}$ and $\dim E_T(\mathcal{B})$ the dimension of the range of $E_T(\mathcal{B})$. Put

$$\mathcal{H}=\mathcal{H}_p\otimes[\mathcal{F}_a\oplus\mathcal{F}_a^\perp]=[\mathcal{H}_p\otimes\mathcal{F}_a]\oplus[\mathcal{H}_p\otimes\mathcal{F}_a^\perp]:=\mathcal{H}_1\oplus\mathcal{H}_2.$$

We see that H_a is reduced by $\mathcal{H}_1$ (Ref. 14, Lemma 3.7). By (2.16), (2.17), and (2.18), we have

$$(\Psi,H_a\Psi)\geqslant(\Psi,(\beta H_f+\overline{\alpha H_p}-\Delta(\rho_{a\Lambda})+\Sigma_0)\Psi), \tag{3.9}$$

where $\alpha:=1-eB'_{a\Lambda}$, $\beta:=1-(eA'_{a\Lambda}+e^2D'_{a\Lambda}+eF'_{a\Lambda})$. Note that by Hypothesis 3, $\alpha>0$ and $\beta>0$. We have, for $\Psi_1\in\mathcal{H}_1\cap D(H_a)$, by (3.9),

$$(\Psi_1,(H_a-m-g_a)\Psi_1)$$

$$\geqslant(\Psi_1,E_{\overline{H_p}}([0,\Sigma-\Sigma_0))\otimes(\beta H^a_{fm}-(m+g_a-\Sigma_0+\Delta(\rho_{a\Lambda})))\Psi_1)$$

$$+\{\alpha(\Sigma-\Sigma_0)-(m+g_a-\Sigma_0+\Delta(\rho_{a\Lambda}))\}(\Psi_1,E_{\overline{H_p}}([\Sigma-\Sigma_0,\infty))\Psi_1) \tag{3.10}$$

$$+(\Psi_1,E_{\overline{H_p}}([\Sigma-\Sigma_0,\infty))\otimes\beta H^a_{fm}\Psi_1). \tag{3.11}$$

By virtue of (3.8), (3.10) is non-negative, and (3.11) is also non-negative. Therefore we have

$$\mathrm{Ran}\,E_{H_a}([g_a,g_a+m))\lceil_{\mathcal{H}_1}\subset\mathrm{Ran}\,E_{\overline{H_p}}([0,\Sigma-\Sigma_0))\otimes E_{H^a_{fm}}([0,(m+g_a-\Sigma_0+\Delta(\rho_{a\Lambda}))/\beta))\lceil_{\mathcal{H}_1}.$$

Note that $g_a+\Delta(\rho_{a\Lambda})\geqslant\Sigma_0$. Thus it follows that

$$\dim E_{H_a}([g_a,g_a+m))\lceil_{\mathcal{H}_1}\leqslant \dim E_{\overline{H_p}}([0,\Sigma-\Sigma_0))\cdot \dim E_{H^a_{fm}}([0,(m+g_a-\Sigma_0+\Delta(\rho_{a\Lambda}))/\beta)).$$

On the other hand, we have $(\Psi_2,H_a\Psi_2)\geqslant(g_a+m)\|\Psi_2\|^2$ (Ref. 6, Lemma 3.7). Thus we obtain $\mathrm{Ran}\,E_{H_a}([g_a,g_a+m))=\mathrm{Ran}\,E_{H_a}([g_a,g_a+m))\lceil_{\mathcal{H}_1}$. Hence it follows that

$$\dim E_{H_a}([g_a,g_a+m))\leqslant \dim E_{H_a}([g_a,g_a+m))\lceil_{\mathcal{H}_1}<\infty.$$

Thus proof is complete. □

B. Unitary transformations

We define unitary transformations U and U_a by

$$U:=\exp\left(i\sum_{\mu=1}^{3} x_\mu\otimes d\Gamma(k_\mu)\right),\quad U_a:=\exp\left(i\sum_{\mu=1}^{3} x_\mu\otimes d\Gamma(k_{\mu,a})\right),\quad a>0.$$

Define

$$\mathcal{O}:=\mathcal{L}\{F(x)a^{\dagger r1}(e^{-i\theta_1(k)x}f_1)\cdots a^{\dagger r_n}(e^{-i\theta_n(k)x}f_n)\Omega,\Omega\,|\,F\in C_0^\infty(\mathbb{R}^3),f_j\in L_c^2(\mathbb{R}^3),$$
$$\theta_j(k)\in\mathcal{L}\{k,k_a\},r_j=1,2,\quad j=1,2,...,n,n\in N\},$$

where $\mathcal{L}\{\cdots\}$ denotes the finite linear hull of the vectors in $\{\cdots\}$ and $L_c^2(\mathbb{R}^3)$ is the set of functions in $L^2(\mathbb{R}^3)$ with compact support. Since $\mathcal{O}$ is a core for both $H_{0,m}$ and $H^a_{0,m}$, by Lemma 3.1, $\mathcal{O}$ is a core for both $H_{m\Lambda}$ and $H^a_{m\Lambda}$. Note that operators $U,U_a,d\Gamma(k_\mu)$, $d\Gamma(k_{\mu,a})$, H_{fm}, H^a_{fm}, and $a^{\sharp r}(e^{-i\theta(k)x}f)$, $\theta(k)\in\mathcal{L}\{k,k_a\}$, leave $\mathcal{O}$ invariant.

Lemma 3.5: The unitary operators U and U_a satisfy that, on $\mathcal{O}$,

$$UH_{m\Lambda}U^{-1}=(1/2)\{\boldsymbol{\sigma}\cdot(\mathbf{P}_T-e\mathbf{A}_\Lambda)\}^2+H_{fm}+V, \tag{3.12}$$

$$U_aH^a_{m\Lambda}U_a^{-1}=(1/2)\{\boldsymbol{\sigma}\cdot(\mathbf{P}^a_T-e\mathbf{A}^a_\Lambda)\}^2+H^a_{fm}+V, \tag{3.13}$$

where $\mathbf{P}_T:=(P_{T1},P_{T2},P_{T3})$ *and* $\mathbf{P}^a_T:=(P^a_{T1},P^a_{T2},P^a_{T3})$ *with* $P_{T\mu}:=P_\mu-d\Gamma(k_\mu)$ *and* $P^a_{T\mu}:=P_\mu-d\Gamma(k_{\mu,a})$, $\mu=1,2,3$. *Moreover* U_a *leaves* $\mathcal{H}_1$ *invariant*.

Proof: It is easily seen that, on $\mathcal{O}$, $U\mathbf{P}U^{-1}=\mathbf{P}_T$, $U_a\mathbf{P}U_a^{-1}=\mathbf{P}^a_T$, $Ua^r(e^{ikx}f)U^{-1}=a^r(f)$, $Ua^{\dagger r}(e^{-ikx}f)U^{-1}=a^{\dagger r}(f)$, $U_aa^r(e^{ik_ax}f)U_a^{-1}=a^r(f)$, $U_aa^{\dagger r}(e^{-ik_ax}f)U_a^{-1}=a^{\dagger r}(f)$, $UH_{fm}U^{-1}=H_{fm}$, $U_aH^a_{fm}U_a^{-1}=H^a_{fm}$. Thus (3.12) and (3.13) follow. Let $\Psi=F\otimes a^{\dagger r_1}(f_1)\cdots a^{\dagger r_n}(f_n)\Omega$, $F\in C_0^\infty(\mathbb{R}^3)$, $f_j\in l_2(\Gamma_a)$, $j=1,...,n$. The finite linear hull of the form of Ψ's is dense in $\mathcal{H}_1$. Since $e^{-ik_ax}f_j\in l_2(\Gamma_a)$, $j=1,...,n$, for each $x\in\mathbb{R}^3$, we have

$$UF\otimes a^{\dagger r_1}(f_1)\cdots a^{\dagger r_n}(f_n)\Omega=Fa^{\dagger r_1}(e^{-ik_ax}f_1)\cdots a^{\dagger r_n}(e^{-ik_ax}f_n)\Omega\in\mathcal{H}_1.$$

Thus U leaves $\mathcal{H}_1$ invariant. □

We write the right-hand sides of (3.12) and (3.13) as $\hat{H}_{m\Lambda}$ and $\hat{H}^a_{m\Lambda}$, respectively.

Lemma 3.6: The operators $\hat{H}_{m\Lambda}$ *and* $\hat{H}^a_{m\Lambda}$ *are essentially self-adjoint on* $\mathcal{O}$. *The unitary operator U maps* $D(\hat{H}_{m\Lambda})$ *onto* $D(H_{m\Lambda})$ *with* (3.12) *and* U_a *maps* $D(\hat{H}^a_{m\Lambda})$ *onto* $D(H^a_{m\Lambda})$ *with* (3.13).

Proof: Since $\mathcal{O}$ is a core for $H_{m\Lambda}$ and U maps $\mathcal{O}$ onto itself, $UH_{m\Lambda}U^{-1}\lceil_{\mathcal{O}}$ is essentially self-adjoint, which implies that $\hat{H}_{m\Lambda}\lceil_{\mathcal{O}}$ is essentially self-adjoint. Thus U maps $D(\hat{H}_{m\Lambda})$ onto $D(H_{m\Lambda})$ with (3.12). For the case of $\hat{H}^a_{m\Lambda}$, $H^a_{m\Lambda}$, and U_a, it is similarly proved. □

We define $H_T:=\mathbf{P}_T\cdot\mathbf{P}_T+V+H_{fm}$ and $H^a_T:=\mathbf{P}^a_T\cdot\mathbf{P}^a_T+V+H^a_{fm}$. Then $\hat{H}^a_{m\Lambda}=H^a_T+\mathbf{Q}_a$ and $\hat{H}_{m\Lambda}=H_T+\mathbf{Q}$, where $\mathbf{Q}_a:=-e\mathbf{P}^a_T\cdot\mathbf{A}^a_\Lambda+(e^2/2)\mathbf{A}^a_\Lambda\cdot\mathbf{A}^a_\Lambda+(e/2)\boldsymbol{\sigma}\cdot\mathbf{B}^a_\Lambda$ and $\mathbf{Q}:=-e\mathbf{P}^a_T\cdot\mathbf{A}_\Lambda+(e^2/2)\mathbf{A}_\Lambda\cdot\mathbf{A}_\Lambda+(e/2)\boldsymbol{\sigma}\cdot\mathbf{B}_\Lambda$.

Lemma 3.7: Assume that Hypotheses 1, 2, and 3 hold. Then, for sufficiently large a, there exist positive constant γ which is independent of sufficiently large a so that, for $\Psi\in\mathcal{O}$,

$$\|H_T\Psi\|\le\gamma(\|\hat{H}_{m\Lambda}\Psi\|+\|\Psi\|), \tag{3.14}$$

$$\|H_T^a\Psi\|\le\gamma(\|\hat{H}_{m\Lambda}^a\Psi\|+\|\Psi\|). \tag{3.15}$$

Proof: One sees that there exists constants $\alpha<1$ and $\beta\ge 0$ which are independent of large a so that $\|(eH_I^a+e^2H_{II}^a+eH_{III}^a)\Psi\|\le\alpha\|H_{0,m}^a\Psi\|+\beta\|\Psi\|$ for $\Psi\in\mathcal{O}$. Since U_a leaves $\mathcal{O}$ invariant, $\|\mathbf{Q}_a\Psi\|\le\alpha\|H_T^a\Psi\|+\beta\|\Psi\|$ for $\Psi\in\mathcal{O}$. Thus it follows that $\|H_T^a\Psi\|\le(1-\alpha)^{-1}\|\hat{H}_{m\Lambda}^a\Psi\|+\beta(1-\alpha)^{-1}\|\Psi\|$, which implies (3.15). (3.14) is proved similarly. □

C. Existence of ground states with artificial mass

The purpose of this subsection is to obtain the following lemma:

Lemma 3.8: We assume that Hypotheses 1, 2, and 3 hold. Then, for $z\in\mathbb{C}\setminus\mathbb{R}$,

$$\lim_{a\to\infty}(\hat{H}_{m\Lambda}^a-z)^{-1}=(\hat{H}_{m\Lambda}-z)^{-1} \tag{3.16}$$

in the uniform norm.

Before proving Lemma 3.8, we prepare a lemma. We have $0\le k_\mu-k_{\mu,a}\le 2\pi/a$, $\mu=1,2,3$. Then it follows that for $\Psi\in D(H_{fm})$,

$$\|(d\Gamma(k_\mu)-d\Gamma(k_{\mu,a}))\Psi\|\le\frac{2\pi}{ma}\|H_{fm}\Psi\|,\quad \mu=1,2,3. \tag{3.17}$$

Lemma 3.9: Let $\Psi\in\mathcal{O}$. Then there exists a positive constant α so that

$$\|P_{T\mu}^a\Psi\|\le\alpha(\|H_T^a\Psi\|+\|\Psi\|),\quad \mu=1,2,3,$$

$$\|H_f^{1/2}\Psi\|\le\alpha(\|H_T\Psi\|+\|\Psi\|),\quad \|H_f^{1/2}\Psi\|\le\alpha(|vH_T^a\Psi\|+\|\Psi\|),$$

$$\|H_{fm}\Psi\|\le\alpha(\|H_T\Psi\|+\|\Psi\|),\quad \|H_{fm}\Psi\|\le\alpha(\|H_T^a\Psi\|+\|\Psi\|).$$

Proof: One sees that $\|P_\mu\Psi\|\le\alpha(\|H_{0,m}^a\Psi\|+\|\Psi\|)$. Since U_a leaves $\mathcal{O}$ invariant, the first inequality holds. The other inequalities are handled similarly.

Proof of Lemma 3.8:

We set $\mathbf{R}:=(\hat{H}_{m\Lambda}-z)^{-1}$ and $\mathbf{R}_a:=(\hat{H}_{m\Lambda}^a-z)^{-1}$. We have

$$\mathbf{R}-\mathbf{R}_a=\sum_{\mu=1}^{3}(I_\mu^a+eII_\mu^a+e^2III_\mu^a+eIV_\mu^a)+R^a,$$

where

$$I_\mu^a:=\mathbf{R}_a((P_{T\mu}^a)^2-(P_{T\mu})^2)\mathbf{R},\quad II_\mu^a:=\mathbf{R}_a(P_{T\mu}^aA_{\Lambda\mu}^a-P_{T\mu}A_{\Lambda\mu})\mathbf{R},$$

$$III_\mu^a:=\mathbf{R}_a((A_{\Lambda\mu}^a)^2-(A_{\Lambda\mu})^2)\mathbf{R},\quad IV_\mu^a:=\mathbf{R}_a(\sigma_\mu B_{\Lambda\mu}^a-\sigma_\mu B_{\Lambda\mu})\mathbf{R},$$

$$R^a:=\mathbf{R}_a(H_{fm}^a-H_{fm})\mathbf{R}.$$

We shall check that I_μ^a, II_μ^a, III_μ^a, IV_μ^a, and R^a go to zero as $a\to\infty$, respectively. We also denote the norm of operators in $\mathcal{H}$ by $\|\cdot\|$. For $f\in\mathcal{M}_n$, we define $R_n(f):=\|f-\sum_{l\in\Gamma_a}f(l)X_{\Gamma(l,a)}\|_n$. Note that, if f has a compact support, then $R_n(f)$ goes to zero as $a\to\infty$. We put

$$X_n := \max_{\mu=1,2,3,r=1,2} R_n\left(\frac{\rho_\Lambda e^r_\mu}{\sqrt{2\omega}}\right), \quad Y_n := \max_{\mu=1,2,3,r=1,2} R_n\left(\frac{\rho_\Lambda (k\times e^r)_\mu}{\sqrt{2\omega}}\right), \quad Z_n := \max_{\mu=1,2,3,r=1,2} \left\|\frac{\rho_\Lambda e^r_\mu}{\sqrt{2\omega}}\right\|_n .$$

Define

$$\alpha := \max\left\{\begin{array}{ll} \|H_{fm}(H_T^a - z^*)^{-1}\|, & \|H_{fm}(H_T - z^*)^{-1}\| \\ \|H_f^{1/2}(H_T^a - z^*)^{-1}\|, & \|H_f^{1/2}(H_T - z^*)^{-1}\| \\ \|P^a_{T\mu}(H_T^a - z^*)^{-1}\| & \end{array}\right\}, \quad \beta := \max\{\|H_T^a \mathbf{R}_a^\sharp\|, \|H_T \mathbf{R}^\sharp\|\},$$

where $\mathbf{R}^\sharp$ (resp. $\mathbf{R}_a^\sharp$) denotes $\mathbf{R}$ or $\mathbf{R}^*$ (resp. $\mathbf{R}_a$ or $\mathbf{R}_a^*$) and z^*, z or $\overline{z}.$ The existence of α and β is ensured by Lemmas 3.7 and 3.9. Note that $\|(H_{fm} - H_{fm}^a)\Psi\| \leqslant 2C_a(1-C_a)^{-1}\|H_{fm}\Psi\|$ for $\Psi \in D(H_{fm})$, where $C_a := \sqrt{3}(\pi/a)(1/(2m)+1)$ (Ref. 16, Lemmas 3.1 and 3.6). Thus

$$\|\mathbf{R}^a\| \leqslant \frac{1}{|\mathfrak{F}z|}\,\frac{2C_a}{1-C_a}\,\alpha\beta. \tag{3.18}$$

We see that, by (3.17),

$$\begin{aligned} |(\Psi, I_\mu^a \Phi)| &\leqslant |(P^a_{T\mu}\mathbf{R}_a^*\Psi, (P^a_{T\mu} - P_{T\mu})\mathbf{R}\Phi)| + |((P^a_{T\mu} - P_{T\mu})\mathbf{R}_a^*\Psi, P_{T\mu}\mathbf{R}\Phi)| \\ &\leqslant \frac{2\pi}{ma}\{\|P^a_{T\mu}(H_T^a - \overline{z})^{-1}(H_T^a - \overline{z})\mathbf{R}_a^*\Psi\|\|H_{fm}(H_T - z)^{-1}(H_T - z)\mathbf{R}\Phi\| \\ &\quad + \|H_{fm}(H_T^a - \overline{z})^{-1}(H_T^a - \overline{z})\mathbf{R}_a^*\Psi\|\|P_{T\mu}(H_T - z)^{-1}(H_T - z)\mathbf{R}\Phi\|\} \\ &\leqslant \frac{2\cdot 2\pi}{ma}(\alpha\beta)^2\|\Psi\|\|\Phi\|. \end{aligned} \tag{3.19}$$

We see that by Lemma 4.10,

$$\begin{aligned} |(\Psi, II_\mu^a \Phi)| &\leqslant \|P^a_{T\mu}\mathbf{R}_a^*\Psi\|\|(A^a_{A_{\Lambda\mu}} - A_{\Lambda\mu})\mathbf{R}\Phi\| + \|(P^a_{T\mu} - P_{T\mu})\mathbf{R}_a^*\Psi\|\|A_{\Lambda\mu}\mathbf{R}\Phi\| \\ &\leqslant 2\|P^a_{T\mu}(H_T^a - \overline{z})^{-1}(H_T^a - \overline{z})\mathbf{R}_a^*\Psi\| \\ &\quad \times\{2X_{-1}\|H_f^{1/2}(H_T - z)^{-1}(H_T - z)\mathbf{R}\Phi\| + X_0\|\mathbf{R}\Phi\|\} \\ &\quad + \left(\frac{2\cdot 2\pi}{ma}\right)\|H_{fm}(H_T^a - \overline{z})^{-1}(H_T^a - \overline{z})\mathbf{R}_a^*\Psi\| \\ &\quad \times\{2Z_{-1}\|H_f^{1/2}(H_T - z)^{-1}(H_T - z)\mathbf{R}\Phi\| + Z_0\|\mathbf{R}\Phi\|\} \\ &\leqslant 2\left\{\alpha\beta\left(2\alpha\beta X_{-1} + \frac{X_0}{|\mathfrak{F}z|}\right) + \frac{2\pi}{ma}\alpha\beta\left(2\alpha\beta Z_{-1} + \frac{Z_0}{|\mathfrak{F}z|}\right)\right\}\|\Psi\|\|\Phi\|. \end{aligned} \tag{3.20}$$

We have

$$III_\mu^a = (1/2)\mathbf{R}_a\{(A^a_{\Lambda\mu} - A_{\Lambda\mu})A_{\Lambda\mu} + A^a_{\Lambda\mu}(A^a_{\Lambda\mu} - A_{\Lambda\mu})\}\mathbf{R}.$$

The first term in III_μ^a can be estimated as follows:

$$\begin{aligned} &|((A^a_{\Lambda\mu} - A_{\Lambda\mu})\mathbf{R}_a^*\Psi, A_{\Lambda\mu}\mathbf{R}\Phi)| \\ &\quad \leqslant 2\{X_{-1}\|H_f^{1/2}(H_T^a - \overline{z})^{-1}(H_T^a - \overline{z})\mathbf{R}_a^*\Psi\| + X_0\|\mathbf{R}_a^*\Psi\|\} \\ &\qquad \times\{2Z_{-1}\|H_f^{1/2}(H_T - z)^{-1}(H_T - z)\mathbf{R}\Phi\| + Z_0\|\mathbf{R}\Phi\|\} \end{aligned}$$

$$\leqslant 2\left(2\alpha A X_{-1}+\frac{X_0}{|\Im z|}\right)\left(2\alpha\beta Z_{-1}+\frac{Z_0}{|\Im z|}\right)\|\Psi\|\|\Psi\|. \tag{3.21}$$

The second term in III_μ^a are similarly handled. We see that

$$|(\Psi, IV_\mu^a\Phi)|\leqslant 2\|\mathbf{R}_a^*\Psi\|\{2Y_{-1}\|H_f^{1/2}(H_T-z)^{-1}(H_T-z)\mathbf{R}\Phi\|+Y_0\|\mathbf{R}\Phi\|\}$$
$$\leqslant \frac{2}{|\Im z|}\left(2\alpha\beta Y_{-1}+\frac{Y_0}{|\Im z|}\right)\|\Psi\|\|\Phi\|. \tag{3.22}$$

Thus, by (3.18), (3.19), (3.20), (3.21), and (3.22), R^a, I_μ^a, II_μ^a, III_μ^a, and IV_μ^a, $\mu=1,2,3$, go to zero, as $a\to\infty$, respectively. Hence (3.16) follows.

Lemma 3.10: We assume that Hypotheses 1, 2, 3, and 4 hold and that m satisfies (3.7). *Then* $[g[H_m],g[H_m]+m)\subset\sigma_d(H_m)$. *In particular, the ground states of* H_m *exist.*

Proof: By Lemma 3.3 and (3.13), $[g[\hat{H}_{m\Lambda}^a],g[\hat{H}_{m\Lambda}^a]+m)\subset\sigma_d(\hat{H}_{m\Lambda}^a)$. Lemma 3.9 and (Ref. 24, Lemma 4.6) yield that $[g[\hat{H}_{m\Lambda}],g[\hat{H}_{m\Lambda}]+m)\subset\sigma_d(\hat{H}_{m\Lambda})$, which implies that, by (3.12), $[g[H_{m\Lambda}],g[H_{m\Lambda}]+m)\subset\sigma_d(H_{m\Lambda})$. Since $\|\rho_\Lambda-\rho\|_n\to 0$ as $\Lambda\to\infty$ for $n=-2,-1$, $\lim_{L\to\infty}(H_{m\Lambda}-z)^{-1}=(H_m-z)^{-1}$ in the uniform norm. Then $[g[H_m],g[H_m]+m)\subset\sigma_d(H_m)$. □

D. Binding

Let Ω_m be a ground state of H_m. Define $T:=K\otimes E_{H_f}(\{0\})$, where K is a projection on $L^2(\mathbb{R}^3)$. In this subsection, we estimate both $\|N^{1/2}\Omega_m\|$ and $|(T\Omega_m,(eH_I+e^2H_{II}+eH_{III})\Omega_m)|$. Notation A,B,C,D,E,F,G are in (2.12). Define

$$A'':=A/2, B'':=6ar_{-2}, C'':=6br_{-2}, C'':=6br_{-2}, D'':=E'':=D/2=E/2, F'':=G'':=F/2=G/2.$$

Let

$$P:=e(A+B+F)+e^2D,\quad Q_1:=e(C+G)+e^2E,\quad Q_2:=e(A+2B+F)+e^2D,$$

$$P'':=e(A''+B''+F'')+e^2D'',\quad Q_1'':=e(C''+G'')+e^2E'',\quad Q_2'':=e(A''+2B''+F'')+e^2D''.$$

We define $\mathcal{B}:=\max\{|\Sigma_0-\Delta(\rho)|,|\Sigma_0+\Delta(\rho)|\}$. By (3.5), $|g[H_m]|\leqslant\mathcal{B}$. We have

$$\|P_\mu^2\Omega_m\|\leqslant\|P_\mu^2(H_0-i)^{-1}\|\|(H_0-i)(H_m-i)^{-1}\|\|(H_m-i)\Omega_m\|$$
$$\leqslant\|P_\mu^2(H_0-i)^{-1}\|\|(H_0-i)(H_m-i)^{-1}\|(\mathcal{B}+1)\|\Omega_m\|.$$

Since $\|P_\mu^2\Psi\|\leqslant 2a\|H_p\Psi\|+2b\|\Psi\|\leqslant 2a\|H_0\Psi\|+(4a|\Sigma_0|+2b)\|\Psi\|$ for $\Psi\in\mathcal{E}$,

$$\|P_\mu^2(H_0-i)^{-1}\|\leqslant 2a+2b+4a|\Sigma_0|. \tag{3.23}$$

Moreover since $\|H_0\Psi\|\leqslant\|H_m\Psi\|+e\|H_I\Psi\|+e^2\|H_{II}\Psi\|+e\|H_{III}\Psi\|+2|\Sigma_0|\|\Psi\|$ for $\Psi\in\mathcal{E}$,

$$\|H_0\Psi\|\leqslant\frac{1}{1-P}\{\|H_m\Psi\|+(Q_1+(Q_2+2)|\Sigma_0|)\|\Psi\|\}. \tag{3.24}$$

Hence it follows that

$$\|(H_0-i)(H_m-i)^{-1}\|\leqslant\frac{1}{1-P}+\frac{Q_1+(Q_2+2)|\Sigma_0|}{1-P}+1. \tag{3.25}$$

We define $\mathcal{A}:=(2-P+Q_1+(Q_2+2)|\Sigma_0|)(2a+2b+4a|\Sigma_0|)(1-P)^{-1}$. Thus, by (3.23) and (3.25), $\|P_\mu^2\Omega_m\|\leqslant\mathcal{A}(\mathcal{B}+1)\|\Omega_m\|$. In particular,

$$\|P_\mu\Omega_m\|\leq\sqrt{\mathcal{A}(\mathcal{B}+1)}\|\Omega_m\|. \tag{3.26}$$

We introduce Hypothesis 5:

Hypothesis 5: The ultraviolet cutoff ρ holds that $r_{-3}<\infty$ and the coupling constant e satisfies that

$$r_{-3}\leq\frac{1-3er_{-1}/\sqrt{2}}{6e(\sqrt{\mathcal{A}(\mathcal{B}+1)/2}+3er_{-1}/2)}.$$

Lemma 3.11: We assume that Hypotheses 1, 2, 3, 4, and 5 hold. Then $\|N^{1/2}\Omega_m\|\leq\delta(e)<1$, *where* $\delta(e)$ *is a positive constant.*

Proof: By (2.6), (2.7), and (3.26), we have

$$\|N^{1/2}\Omega_m\|^2\leq\frac{3\cdot 2e}{\sqrt{2}}r_{-3}\|N^{1/2}\Omega_m\|\sqrt{\mathcal{A}(\mathcal{B}+1)}\|\Omega_m\|$$

$$+\frac{3\cdot 2e^2}{2}r_{-3}r_{-1}\|N^{1/2}\Omega_m\|(\|N^{1/2}\Omega_m\|+\|(N+I)^{1/2}\Omega_m\|)+\frac{3\cdot 2e}{2\sqrt{2}}r_{-1}\|N^{1/2}\Omega_m\|\|Q_m\|.$$

Hence from Hypothesis 5 it follows that

$$\frac{\|N^{1/2}\Omega_m\|}{\|\Omega_m\|}\leq\frac{6er_{-3}(\sqrt{\mathcal{A}(\mathcal{B}+1)/2}+er_{-1}/2)+3e/\sqrt{2}r_{-1}}{1-6e^2r_{-3}r_{-1}}:=\delta(e)<1.$$

Thus lemma follows. □

Next we shall estimate $|(T\Omega_m,(eH_I+e^2H_{II}+eH_{III})\Omega_m)|$. Note that $a^r(f)Q\Psi=0$ for $f\in L^2(\mathbb{R}^3)$. We have

$$\frac{|(T\Omega_m,H_I\Omega_m)|}{\|T\Omega_m\|}\leq(A''+B'')\|H_0\Omega_m\|+C'\|\Omega_m\|+(A''+2B'')|\Sigma_0|\|\Omega_m\|,$$

$$\frac{|(T\Omega_m,H_{II}\Omega_m)|}{\|T\Omega_m\|}\leq D''\|H_0\Omega_m\|+E''\|\Omega_m\|+D''|\Sigma_0|\|\Omega_m\|,$$

$$\frac{|(T\Omega_m,H_{III}\Omega_m)|}{\|T\Omega_m\|}\leq F''\|H_0\Omega_m\|+G''\|\Omega_m\|+F''|\Sigma_0|\|\Omega_m\|.$$

From (3.24) it follows that

$$\frac{|(T\Omega_m,(eH_I+e^2H_{II}+eH_{III})\Omega_m)|}{\|T\Omega_m\|}\leq\frac{P''}{1-P}\{\mathcal{B}+Q_1+(Q_2+2)|\Sigma_0|)\}+Q_1''+Q_2''|\Sigma_0|. \tag{3.27}$$

We put the right-hand side of (3.27) by $\mathcal{C}$.

E. Overlap and existence of ground states

Without loss of generality, we can assume that $\|\Omega_m\|=1$. Then we can find a subsequence $\{m_j\}_{j=1}^\infty$ so that $\Omega_0:=w-\lim_{j\to\infty}\Omega_{m_j}$ exists.

Lemma 3.12: We assume that Hypotheses 1, 2, and 3 hold. Then $\lim_{m\to 0}g[H_m]=g[H]$.

Proof: See Ref. 14, Lemma 3.13: □

We introduce Hypothesis 6:

Hypothesis 6: The coupling constant e is such that $\Sigma-\Sigma_0>\Delta(\rho)+\mathcal{C}/\sqrt{1-\delta(e)^2}$. Note that Hypothesis 6 implies that $\Sigma-g[H_m]>0$.

Lemma 3.13: We assume that Hypotheses 1, 2, 3, 4, 5, and 6 hold. Then $\Omega_0 \neq 0$.

Proof: Put $P := E_{H_p}([\Sigma_0, \Sigma)) \otimes E_{H_f}(\{0\})$, $Q := E_{H_p}([\Sigma, \infty)) \otimes E_{H_f}(\{0\})$. Then we have $(\Omega_m, P\Omega_m) \geqslant 1 - \|N^{1/2}\Omega_m\|^2 - (\Omega_m, Q\Omega_m)$. Since $(\Omega_m, Q(H_p - g[H_m])\Omega_m) = -(\Omega_m, Q(eH_I + e^2 H_{II} + eH_{III})\Omega_m)$, (3.27) implies that $(\Sigma - g[H_m])(\Omega_m, Q\Omega_m) \leqslant \mathcal{C}\|Q\Omega_m\|$. Hence,

$$(\Omega_m, P\Omega_m) \geqslant 1 - \delta(e)^2 - (\mathcal{C}/\Sigma - g[H_m]))^2. \tag{3.28}$$

Taking the subsequence $\{m_j\}_{j=1}^{\infty}$, we see that both sides of (3.28) converge with

$$(\Omega_0, P\Omega_0) \geqslant 1 - \delta(e)^2 - (\mathcal{C}(\Sigma - g[H]))^2. \tag{3.29}$$

The right-hand side of (3.29) is positive by Hypotheses 5 and 6. Thus lemma follows. □

We state the main theorem.

Theorem 3.14: *We assume that Hypotheses 1, 2, 3, 4, 5, and 6 hold. Then* Ω_0 *is a ground state of H. In addition to this, we assume that* H_p *has a unique ground state* ϕ. *Then,*

$$\lim_{e \to 0} \|\phi \otimes \Omega - \Omega_0\| = 0.$$

Proof: By Lemmas 3.12 and 3.13, the theorem follows. □

F. Convergence of ground states

Define the Hamiltonian with ''g-factor'' as follows:

$$H_g := (1/2)(\mathbf{P} - e\mathbf{A}(x))^2 + V + H_f + (g/2)\boldsymbol{\sigma} \cdot \mathbf{B}(x).$$

Since $0 \leqslant g \leqslant 1$, under Hypotheses 1, 2, 3, 4, 5, and 6, their ground states exist. We denote a normalized ground state of H_g by $\Omega(g); \|\Omega(g)\| = 1$. Then there exists a subsequence $\{g_j\}_{j=1}^{\infty}$ so that $\tilde{\Omega} := w - \lim_{j \to \infty} \Omega(g_j)$ exists.

Theorem 3.15: *We assume that Hypotheses 1, 2, 3, 4, 5, and 6 hold, and the ground state* $\Omega(0)$ *is a unique ground state of* $H_{g=0}$. *Then* $\tilde{\Omega} = \theta\Omega(0)$, *where* $0 < |\theta| \leqslant 1$.

Proof: It is easily seen that H_g converges to $H_{g=0}$ as $g \to 0$ in the sense of the norm resolvent. Thus $g[H_g] \to g[H_{g=0}]$ as $g \to 0$. We define $\tilde{\mathcal{A}}$, $\tilde{\mathcal{B}}$, and $\tilde{\mathcal{C}}$ by $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ with F, G, F', G', F'', and G'' replaced by zero, respectively. Let $P = E_{H_p}([\Sigma_0, \Sigma)) \otimes E_{H_f}(\{0\})$. From (3.29) it follows that

$$(\tilde{\Omega}, P\tilde{\Omega}) \geqslant 1 - \left\{ \frac{6er_{-3}(\sqrt{\tilde{\mathcal{A}}(\tilde{\mathcal{B}}+1)/2} + er_{-1}/2)}{1 - 6e^2 r_{-3} r_{-1}} \right\}^2 - \left(\frac{\tilde{\mathcal{C}}}{\Sigma - g[H_{g=0}]} \right)^2 > 0.$$

Hence $\tilde{\Omega} \neq 0$. Then $\tilde{\Omega}$ is a ground state of $H_{g=0}$. Thus, the theorem follows. □

Remark 3.16: In the case of $g > 1$, *taking the coupling constant e sufficiently small, we can see that ground states of* H_g *exists and Theorem 3.15 holds.*

Remark 3.17: In the case of $\Sigma = \infty$, *we do not need Hypotheses 4 and 6 in Theorems 3.14 and 3.15.*

Note added in proof: After the completion of this paper, by V. Bach we have learned that Hypothesis 5 (infrared cutoff) is not needed[25] (cf. Ref. 26).

ACKNOWLEDGMENTS

I thank H. Spohn for the hospitality of TU-München and for pointing out some improvements of the first version of this manuscript. I also thank V. Bach for helpful comments. This work is supported in part by the Japan Society for the Promotion of Science (JSPS).

[1] F. Hiroshima, ''Ground states of a model in nonrelativistic quantum electrodynamics. II,'' J. Math. Phys. (to be published).
[2] W. Pauli and M. Fierz, ''Zur Theorie der Emission langwelliger Lichtquanten,'' Nuovo Cimento **15**, 167–188 (1938).
[3] H. A. Bethe, ''The electromagnetic shift of energy levels,'' Phys. Rev. **72**, 339–342 (1947).
[4] T. A. Welton, ''Some observable effects of the quantum mechanical fluctuations of the electromagnetic field,'' Phys. Rev. **74**, 1157–1169 (1948).
[5] A. Arai, ''An asymptotic analysis and its application to the nonrelativistic limit of the Pauli–Fierz and a spin-boson model,'' J. Math. Phys. **31**, 2653–2663 (1990).
[6] F. Hiroshima, ''Scaling limit of a model in quantum electrodynamics,'' J. Math. Phys. **34**, 4478–4518 (1993).
[7] F. Hiroshima, ''Scaling limit of a model in quantum electrodynamics with many nonrelativistic particles,'' Rev. Math. Phys. **9**, 201–225 (1997).
[8] A. Arai, ''Self-adjointness and spectrum of Hamiltonians in nonrelativistic quantum electrodynamics,'' J. Math. Phys. **22**, 534–537 (1981).
[9] A. Arai, ''On a model of a harmonic oscillator coupled to a quantized, massless scalar field I,'' J. Math. Phys. **22**, 2539–2548 (1981).
[10] A. Arai, ''Rigorous theory of spectra and radiation for a model in quantum electrodynamics,'' J. Math. Phys. **24**, 1896–1910 (1983).
[11] V. Bach, J. Fröhlich, and I. M. Sigal, ''Quantum electrodynamics of confined nonrelativistic particles,'' Adv. Math. **137**, 299–395 (1998).
[12] C. Fefferman, ''On electrons and nuclei in a magnetic field,'' Adv. Math. **124**, 100–153 (1996).
[13] C. Fefferman, J. Fröhlich, and G. M. Graf, ''Stability of ultraviolet-cutoff quantum electrodynamics with nonrelativistic matter,'' Commun. Math. Phys. **190**, 309–330 (1997).
[14] F. Hiroshima, ''Ground states and spectrum of quantum electrodynamics of nonrelativistic particles'' (preprint).
[15] A. Amman, ''Ground states of a spin-boson models,'' Ann. Phys. **208**, 414–448 (1991).
[16] A. Arai and M. Hirokawa, ''On the existence and uniqueness of the ground states of a generalized spin-boson model,'' J. Funct. Anal. **151**, 455–503 (1997).
[17] J. Fröhlich, ''Existence of dressed one electron states in a class of persistent models,'' Fortschr. Phys. **22**, 159–198 (1974).
[18] H. Spohn, ''Ground state(s) of the spin-boson Hamiltonian,'' Commun. Math. Phys. **123**, 277–304 (1989).
[19] H. Spohn, ''Ground state of quantum particle coupled to a scalar boson field,'' Lett. Math. Phys. **44**, 9–16 (1998).
[20] J. Glimm and A. Jaffe, ''The $\lambda(\phi^4)_2$ quantum field theory without cutoffs II. The field operators and approximate vacuum,'' Ann. Math. **91**, 362–401 (1970).
[21] M. Reed and B. Simon, *Method of Modern Mathematical Physics* (Academic, New York, 1974), Vol. II.
[22] T. Okamoto and K. Yajima, ''Complex scaling technique in nonrelativistic massive QED,'' Ann. Inst. Henri Poincaré **42**, 311–327 (1985).
[23] F. Hiroshima, ''Essential self-adjointness of translation-invariant quantum field models for arbitrary coupling constants'' (preprint).
[24] B. Simon and Høegh-Krohn, ''Hypercontractive semigroup and two-dimensional self-coupled Bose fields,'' J. Funct. Anal. **9**, 121–180 (1972).
[25] V. Bach, J. Fröhlich, and I. M. Sigal, ''Spectral analysis for systems of atoms and molecules coupled to the quantized radiation field'' (preprint).
[26] A. Arai, M. Hirokawa, and F. Hiroshima, ''On the absence of eigenvectors of Hamiltonians in a class of massless quantum field models without infrared cutoff,'' J. Funct. Anal. (to be published).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Spherically symmetric solutions of the SU(N) Skyrme models

T. Ioannidou[a)]
Institute of Mathematics, University of Kent at Canterbury, Canterbury CT2 7NF, United Kingdom

B. Piette[b)] and W. J. Zakrzewski[c)]
Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, United Kingdom

(Received 4 May 1999; accepted for publication 19 July 1999)

Recently we have presented an ansatz which allows us to construct skyrmion fields from the harmonic maps of S^2 to CP^{N-1}. In this paper we examine this construction in detail and use it to construct, in an explicit form, new static spherically symmetric solutions of the SU(N) Skyrme models. We also discuss some properties of these solutions.
[S0022-2488(99)02612-2]

I. INTRODUCTION

The Skyrme model presents an opportunity to understand nuclear physics as a low energy limit of quantum chromodynamics (QCD). The model was initially proposed as a theory of strong interactions of hadrons,[1] but recently, it was shown to be the low energy limit of QCD in the large N_c limit.[2] Since then further work has suggested that topologically nontrivial solutions of this model, known as skyrmions, can be identified with classical ground states of light nuclei. However, a thorough understanding of the structure and dynamics of multiskyrmion configurations is required before a more qualitative assessment of the validity of this application of the model can be made.

The SU(N) Skyrme model involves fields which take values in SU(N); i.e., are described by SU(N) valued functions of $\vec{x}$ and t. Its static solutions correspond to field configurations describing multiskyrmions. In this paper we construct new solutions for fields whose energy density is spherically symmetric.

Multiskyrmions are stationary points (maxima or saddle points) of the static energy functional, which is given in topological charge units by

$$E=\frac{1}{12\pi^2}\int_{R^3}\left\{-\frac{1}{2}\mathrm{tr}(\partial_i UU^-)^2-\frac{1}{16}\mathrm{tr}[\partial_i UU^{-1},\partial_j UU^{-1}]^2\right\}d^3\vec{x}, \tag{1}$$

where $U(\vec{x})\in \mathrm{SU}(N)$.

In this case multiskyrmions are solutions of the equation

$$\partial_i(\partial_i UU^{-1}-\tfrac{1}{4}[\partial_j UU^{-1},[\partial_j UU^{-1},\partial_i UU^{-1}]])=0. \tag{2}$$

We have, for simplicity, set the mass terms to zero. This has been done for convenience, since the conventional mass terms introduce only small changes and, as we will see later, affect only profile functions. Therefore, all our discussion can be easily generalized to include such mass terms.

a)Electronic mail: T.Ioannidou@ukc.ac.uk
b)Electronic mail: B.M.A.G.Piette@durham.ac.uk
c)Electronic mail: W.J.Zakrzewski@durham.ac.uk

Finiteness of the energy functional requires that $U(\vec{x})$ approaches a constant matrix at spatial infinity, which can be chosen to be the identity matrix by a global SU(N) transformation. So, without any loss of generality, we can impose the following boundary condition on U: $U \to I$ as $|\vec{x}| \to \infty$.

Since $U \to I$ as $|\vec{x}| \to \infty$ is a mapping from $S^3 \to \mathrm{SU}(N)$, it can be classified by the third homotopy group $\pi_3(\mathrm{SU}(N)) \equiv Z$ or, equivalently, by the integer valued winding number

$$B = \frac{1}{24\pi^2} \int_{R^3} \epsilon_{ijk} \operatorname{tr}(\partial_i U U^{-1} \partial_j U U^{-1} \partial_k U U^{-1}) d^3\vec{x}, \tag{3}$$

which is a topological invariant. This winding number classifies the solitonic sectors in the model, and as Skyrme has argued,[1] $B(U)$ may be identified with the baryon number of the field configuration.

Up to now most of the studies involving the Skyrme model have concentrated on the SU(2) version of the model and its embeddings into SU(N). The simplest nontrivial classical *solution* involves a single skyrmion ($B=1$) and has already been discussed by Skyrme.[1] The energy density of this solution is radially symmetric and, as a result, using the so-called hedgehog ansatz one can reduce (2) to an ordinary differential equation, which then has to be solved numerically. Many solutions with $B>1$ of the SU(2) model have also been computed numerically and, in all cases, the solutions are very symmetrical (cf. Battye, and Sutcliffe,[3] and references therein). However, since the model is not integrable, with few exceptions, explicit solutions (even) for spherical symmetric SU(N) skyrmions are not known.

The first example of a *nonembedded* solution for a higher group was the SO(3) soliton, corresponding to a bound system of two skyrmions, which was found by Balachandran *et al.*,[4] Another solution, with a large SU(3) strangeness content, was found by Kopeliovich, Schwesinger, and Stern.[5] However, all other known multiskyrmion configurations seem to be the embeddings of the solutions of the SU(2) model.

Recently, we have shown[6] how to construct low energy states of the SU(N) model by using CP^{N-1} harmonic maps. Our discussion involved only one projector. In this paper, we extend our method to more projectors. We show that, for the SU(N) model, when we take $N-1$ projectors which lead to spherically symmetric energy densities, the full equations of the model separate and the problem of finding exact solutions is reduced to having to solve $N-1$ coupled nonlinear ordinary differential equations for $N-1$ profile functions. This way we obtain a whole family of new spherical symmetric multibaryon solutions of the SU(N) models. Our solutions include the SU(3) dibaryon configuration of Balachandran *et al.*[4] and the nontopological SU(3) four baryon configuration of Ref. 6. Our solutions do not correspond to the global minima of the energy density as the SU(2) embedded solutions with the same baryon number have energies smaller than the energy of the solutions we present here.

II. HARMONIC MAPS

In Ref. 6 we generalized the SU(2) ansatz of Houghton, Manton, and Sutcliffe[7] to SU(N). This generalization involved rewriting the expression of Houghton and co-workers as a projector from S^2 to CP^{N-1}. It gave us a new way of interpreting old results and of deriving expressions for the low energy SU(N) field configurations which are *not* simple embeddings of SU(2) fields. In particular, the energy distributions exhibit very different symmetries from those of the embeddings. The method also gave us a new solution of the SU(3) model, which lies in the topologically trivial sector of the model (i.e., it has zero baryon number) and so, obviously, is not stable.

The method of Ref. 6 can be generalized further, to involve more projectors. In fact, we can exploit here some ideas taken from the theory of harmonic maps of $S^2 \to CP^{N-1}$,[8,9] since they play an important role in our construction.

Recall (cf. Ref. 9) that in N-dimensional space there is a ''natural'' set of projectors: $S^2 \to CP^{N-1}$ maps, which are constructed as follows.

Write each projector P as

$$P(V)=\frac{V\otimes V^{\dagger}}{|V|^{2}}, \tag{4}$$

where V is a N-component complex vector of two variables ξ and $\bar{\xi}$ which locally parameterize S^2. In terms of the more familiar θ and φ, they are given by $\xi=\tan(\theta/2)e^{i\varphi}$. The first projector is obtained by taking $V=f(\xi)$, i.e., an analytic vector of ξ; while the other projectors are obtained from the original V by differentiation and Gramm–Schmidt orthogonalization. If we define an operator P_+ by its action on any vector $v\in C^{N9}$ as

$$P_{+}v=\partial_{\xi}v-v\,\frac{v^{\dagger}\partial_{\xi}v}{|v|^{2}}, \tag{5}$$

then the further vectors $P_+^k v$ can be defined by induction: $P_+^k v=P_+(P_+^{k-1}v)$.

Therefore, in general, we can consider projectors P_k of the form (4) corresponding to the family of vectors $V\equiv V_k=P_+^k f$ [for $f=f(\xi)$] as

$$P_k=P(P_+^k f),\quad k=0,...,N-1, \tag{6}$$

where, due to the orthogonality of the projectors, we have $\Sigma_{k=0}^{N-1}P_k=1$.

The orthogonality properties of our projectors follow from the following properties of vectors $P_+^k f$ which hold when f is holomorphic:

$$(P_+^k f)^{\dagger}P_+^l f=0,\quad k\neq l, \tag{7}$$

$$\partial_{\bar{\xi}}(P_+^k f)=-P_+^{k-1}f\,\frac{|P_+^k f|^2}{|P_+^{k-1}f|^2},\quad \partial_{\xi}\left(\frac{P_+^{k-1}f}{|P_+^{k-1}f|^2}\right)=\frac{P_+^k f}{|P_+^{k-1}f|^2}. \tag{8}$$

Note that, for SU(N), the last projector P_{N-1} in the sequence corresponds to an antianalytic vector; (i.e., the components of $V_{N-1}=P_+^{N-1}f$, up to an irrelevant overall factor which cancels in the projector, are functions of only $\bar{\xi}$).

Our new SU(N) generalisation of Ref. 6 involves the introduction of $N-1$ projectors, i.e.,

$$\begin{aligned}U&=\exp\left\{ig_0\left(P_0-\frac{I}{N}\right)+ig_1\left(P_1-\frac{I}{N}\right)+\cdots+ig_{N-2}\left(P_{N-2}-\frac{I}{N}\right)\right\}\\&=e^{-ig_0/N}(I+A_0P_0)e^{-ig_1/N}(I+A_1P_1)\cdots e^{-ig_{N-2}/N}(I+A_{N-1}P_{N-2}),\end{aligned} \tag{9}$$

where $g_k=g_k(r)$, for $k=0,...,N-2$, are the profile functions and $A_k=e^{ig_k}-1$. Note that the projector P_{N-1} is not included in Eq. (9) since it is the linear combination of the others. (Our previous ansatz given in Ref. 6 corresponds to putting all the profile functions, but the first one, equal to zero.)

The spherically symmetric maps into CP^{N-1} are given by

$$f=(f_0,f_1,\ldots,f_{N-1})^t,$$

where

$$f_k=\xi^k\sqrt{C_{k+1}^{N-1}}, \tag{10}$$

where C_{k+1}^{N-1} denote the binomial coefficients. Furthermore, as we prove in the Appendix, the modulus of the corresponding vector $P_+^k f$ for f of the above-mentioned form is

$$|P_+^k f|^2=\alpha(1+|\xi|^2)^{N-2k-1}, \tag{11}$$

where α depends on N and k.

III. CONSTRUCTING THE SKYRMION SOLUTIONS

In this section we construct a family of exact spherical symmetric solutions of the SU(N) Skyrme models. In fact, we show that for each SU(N) model the Skyrme field involving $N-1$ projectors leads to an exact solution involving $N-1$ profile functions.

A. Skyrme equations

The Skyrme equations (2), when rewritten in spherical coordinates, take the form:

$$\partial_r\left[r^2R_r+\frac{1}{4}\left(A_{\theta r\theta}+\frac{1}{\sin^2\theta}A_{\varphi r\varphi}\right)\right]+\frac{1}{\sin\theta}\partial_\theta\left[\sin\theta\left\{R_\theta+\frac{1}{4}\left(A_{r\theta r}+\frac{1}{r^2\sin^2\theta}A_{\varphi\theta\varphi}\right)\right\}\right]$$
$$+\frac{1}{\sin^2\theta}\partial_\varphi\left[R_\varphi+\frac{1}{4}\left(A_{r\varphi r}+\frac{1}{r^2}A_{\theta\varphi\theta}\right)\right]=0, \tag{12}$$

where $R_i=U^{-1}U_i$ and $A_{\alpha\beta\gamma}\equiv[R_\alpha,[R_\beta,R_\gamma]]$.

It is easy to see, using (9), that

$$R_r=i\sum_{j=0}^{N-2}\dot{g}_j\left(P_j-\frac{I}{N}\right), \tag{13}$$

where $\dot{g}_j(r)$ denotes the derivative of $g_j(r)$ with respect to its argument; and that, in terms of the holomorphic variables ξ and $\bar{\xi}$,

$$\begin{aligned}R_\xi&=\exp\left[\left(-i\sum_{k=0}^{N-2}g_kP_k\right)\right]\partial_\xi\left[\exp\left(i\sum_{i=0}^{N-2}g_iP_i\right)\right]\\&=\left[1+\sum_{k=0}^{N-2}(e^{-ig_k}-1)P_k\right]\left[\sum_{l=0}^{N-2}(e^{ig_l}-1)P_{l\xi}\right]\\&=\sum_{i=1}^{N-1}[e^{i(g_i-g_{i-1})}-1]\frac{V_iV_{i-1}^\dagger}{|V_{i-1}|^2},\end{aligned} \tag{14}$$

where the last line follows from the identity $\exp[(-i\Sigma_{k=0}^{N-2}g_kP_k)]=1+\Sigma_{k=0}^{N-2}(e^{-ig_k}-1)P_k$. Here, $g_{N-1}=0$ and $R_{\bar{\xi}}=-(R_\xi)^\dagger$.

Next we note that

$$\partial_\theta=\frac{1+|\xi|^2}{2\sqrt{|\xi|^2}}(\xi\partial_\xi+\bar{\xi}\partial_{\bar{\xi}}),\quad \partial_\varphi=i(\xi\partial_\xi-\bar{\xi}\partial_{\bar{\xi}}), \tag{15}$$

and rewrite all the terms in (12) in terms of R_ξ, $R_{\bar{\xi}}$, and R_τ and their commutators, i.e.,

$$\partial_\theta(\sin\theta R_\theta)+\frac{1}{\sin\theta}\partial_\phi R_\phi=(1+|\xi|^2)\sqrt{|\xi|^2}((R_\xi)_{\bar{\xi}}+(R_{\bar{\xi}})_\xi), \tag{16}$$

$$A_{\theta r\theta}+\frac{1}{\sin^2\theta}A_{\varphi r\varphi}=\frac{(1+|\xi|^2)^2}{2}\{[R_{\bar{\xi}},[R_r,R_\xi]]+[R_\xi,[R_r,R_{\bar{\xi}}]]\}, \tag{17}$$

$$\sin\theta\partial_\theta(\sin\theta A_{r\theta r})+\partial_\phi(A_{r\varphi r})=2|\xi|^2([R_r,[R_\xi,R_r]]_{\bar{\xi}}+[R_r,[R_{\bar{\xi}},R_r]]_\xi), \tag{18}$$

$$\begin{aligned}\partial_\theta\left(\frac{A_{\varphi\theta\varphi}}{\sin\theta}\right)+\frac{1}{\sin\theta}\partial_\phi(A_{\theta\varphi\theta})=\frac{(1+|\xi|^2)\sqrt{|\xi|^2}}{2}&[\partial_{\bar{\xi}}((1+|\xi|^2)^2[R_\xi,[R_\xi,R_{\bar{\xi}}]])\\&-\partial_\xi((1+|\xi|^2)^2[R_{\bar{\xi}},[R_\xi,R_{\bar{\xi}}]])].\end{aligned} \tag{19}$$

Thus Eq. (12), when rewritten in the holomorphic variables, becomes

$$\partial_r\left[r^2R_r+\frac{(1+|\xi|^2)^2}{8}([R_{\bar\xi},[R_r,R_\xi]]+[R_\xi,[R_r,R_{\bar\xi}]])\right]+\frac{(1+|\xi|^2)^2}{2}((R_{\bar\xi})_\xi+(R_\xi)_{\bar\xi})$$
$$+\frac{(1+|\xi|^2)^3}{8r^2}(\xi[R_\xi,[R_\xi,R_{\bar\xi}]]-\bar\xi[R_{\bar\xi},[R_\xi,R_{\bar\xi}]])+\frac{(1+|\xi|^2)^4}{16r^2}([R_\xi,[R_\xi,R_{\bar\xi}]]_{\bar\xi}$$
$$-[R_{\bar\xi},[R_\xi,R_{\bar\xi}]]_\xi)+\frac{(1+|\xi|^2)^2}{8}([R_r,[R_{\bar\xi},R_r]]_\xi+[R_r,[R_\xi,R_r]]_{\bar\xi})=0. \tag{20}$$

Using (9) we observe that

$$[R_\xi,R_{\bar\xi}]=2P_0\frac{|V_1|^2}{|V_0|^2}(1-\cos(g_1-g_0))-2P_{N-1}\frac{|V_{N-1}|^2}{|V_{N-2}|^2}(1-\cos(g_{N-2}))$$
$$+2\sum_{i=1}^{N-2}P_i\left[\frac{|V_{i+1}|^2}{|V_i|^2}(1-\cos(g_{i+1}-g_i))-\frac{|V_i|^2}{|V_{i-1}|^2}(1-\cos(g_i-g_{i-1}))\right], \tag{21}$$

$$[R_\xi,[R_\xi,R_{\bar\xi}]]=\sum_{i=1}^{N-1}\left(\mu_i\frac{|V_{i+2}|^2}{|V_{i+1}|^2}+\nu_i\frac{|V_{i+1}|^2}{|V_i|^2}+\rho_i\frac{|V_i|^2}{|V_{i-1}|^2}\right)\frac{V_iV_{i-1}^\dagger}{|V_{i-1}|^2}, \tag{22}$$

$$[R_r,[R_{\bar\xi},R_r]]=\sum_{i=1}^{N-1}s_i\frac{V_{i-1}V_i^\dagger}{|V_{i-1}|^2}, \tag{23}$$

where μ, ν, and ρ are functions of $g_k(r)$, only; while s_i are functions of $g_k(r)$ and their derivatives.

Since $V_k=P_+^kf$, one can show that $|V_i|^2/|V_{i-1}|^2\propto(1+|\xi|^2)^{-2}$; while $\partial_\xi(1+|\xi|^2)^{-2}=-2\bar\xi(1+|\xi|^2)^{-3}$ and thus, the derivative terms involving $[R_\xi,[R_\xi,R_{\bar\xi}]]$ in (20) cancel leaving us with derivatives of $V_iV_{i-1}^\dagger/|V_{i-1}|^2$—which are proportional to $\Sigma_{i=1}^{N-1}(1+|\xi|^2)^{-2}(P_i-P_{i-1})h_i$, where h_i involve functions of $g_k(r)$ [due to (8)], multiplied by terms of the form $|V_i|^2/|V_{i-1}|^2$. So the factors $(1+|\xi|^2)^{-4}$ in (20) cancel—leaving us with a sum of differences of two successive projectors multiplied by functions dependent only on r.

Following the above argument and using the properties of R_r, etc., one can show that the terms $[R_r,[R_\xi,R_r]]_{\bar\xi}$ in (20) are proportional to $\Sigma_{i=1}^{N-1}S_i(1+|\xi|^2)^{-2}(P_i-P_{i-1})$, where S_i are functions of $g_k(r)$ and their derivatives—leaving us, once again, with a sum of differences of two successive projectors multiplied by functions dependent only on r.

Finally, the contribution of the terms $(1+|\xi|^2)R_{\xi\bar\xi}$ is given by $\Sigma_{i=1}^{N-1}(P_i-P_{i-1})H_i$, where H_i are only functions of g_k; while the commutators in (17) are equal to a sum of projectors multiplied by $(1+|\xi|^2)^{-2}$, which cancel out in (20). In addition, $\partial_r(r^2R_r)=i\Sigma_{i=0}^{N-2}(P_i-I/N)(2r\dot g_i+r^2\ddot g_i)$.

We note that, for our choice of the vectors V_k, all the dependence on ξ and $\bar\xi$ in (20) resides only in the projectors (the rest of it cancels out). The terms involving $\partial_r(r^2R_r)$ give us expressions involving $I/N-P_i$ while all the other terms give us expressions involving P_i-P_{i-1}. Although N projectors arise in (20), the projector P_{N-1} can be re-expressed in terms of the previous ones—giving $N-1$ factors involving the harmonic maps P_i-I/N (for $i=0,...,N-2$). To satisfy (20) the coefficients of such factors have to vanish leaving us with $N-1$ equations for the $N-1$ profile functions g_i. Hence, if these equations have solutions then they correspond to exact solutions of the SU(N) Skyrme models.

Notice that (10) implies that these solutions have a covariant spherical symmetry. A general rotation of the sphere realized by the Möbius transformation $z\rightarrow\tilde z=(az+b)/(\bar a-\bar bz)$ where $|a|^2+|b|^2=1$. It is easy to show that $\tilde f(z)=[1/(\bar a-\bar bz)^{N-1}]Af$ where A is a matrix depending only on a and b and their complex conjugates. Moreover $|Af(z)|^2=|\tilde{f}(\bar{a}-\bar{b}z)^{N-1}|^2=(|z|^2+1)^{N-1}=|f(z)|^2$ proving that A is an SU(N) matrix and that a rotation of the sphere in R^3 is equivalent to a global gauge transformation on f.

We must now show that $P_+^k f \to A P_+^k f$ under the Möbius transformation. To do this we use the fact that $P_+^k f=(1-P_0-P_1-\cdots-P_{k-1})\partial_z^k f$. As $\tilde{\partial}_z=(\bar{a}-\bar{b}z)^2\partial_z$ we can write

$$\tilde{\partial}_z^k \tilde{f}=(\bar{a}-\bar{b}z)^2\partial_z\left(\frac{Af}{(\bar{a}-\bar{b}z)^k}\right), \tag{24}$$

implying that

$$P_+^k \tilde{f}=(1-\tilde{P}_0-\tilde{P}_1-\cdots-\tilde{P}_{k-1})\frac{1}{(\bar{a}-\bar{b}z)^{N-2k}}A\partial z^k f. \tag{25}$$

By induction, this shows that the P_k all transform under the same unitary conjugation and that the Möbius transformation is equivalent to a global SU (N) gauge transformation on (9).

The $N-1$ equations for the profile functions can be obtained either from (20)—which is a hard task; or from the variation of the energy (1)—using (9) and integrating out ξ and $\bar{\xi}$ variables. Clearly, the two methods give the same equations.

Let us stress that our procedure hinges on having $N-1$ profile functions and on the very special form of our vectors V_k. Had we taken other vectors V_k, we would have obtained some ξ and $\bar{\xi}$ dependence outside the projectors; while had we taken less than $N-1$ profile functions and projectors we would have obtained too many equations for our functions. It is only in the case of $N-1$ projectors that we get the right number of equations.

B. Energy dependence on profile functions

The energy (1), when written in the holomorphic variables, becomes

$$E=-\frac{i}{12\pi^2}\int r^2\,dr\,d\xi\,d\bar{\xi}\,tr\left(\frac{1}{(1+|\xi|^2)^2}R_r^2+\frac{1}{r^2}|R_\xi|^2+\frac{1}{4r^2}[R_r,R_\xi][R_r,R_{\bar{\xi}}]-\frac{(1+|\xi|^2)^2}{16r^4}[R_{\bar{\xi}},R_\xi]^2\right). \tag{26}$$

Using (13) and (14) we find that

$$\mathrm{tr}\,R_r^2=\frac{1}{N}\left(\sum_{i=0}^{N-2}\dot{g}_i\right)^2-\sum_{i=0}^{N-2}\dot{g}_i^2, \tag{27}$$

$$\mathrm{tr}|R_\xi|^2=-2\sum_{i=1}^{N-1}B_i, \tag{28}$$

$$\mathrm{tr}([R_r,R_\xi][R_r,R_{\bar{\xi}}])=-2\sum_{k=1}^{N-1}B_k(\dot{g}_k-\dot{g}_{k-1})^2, \tag{29}$$

$$\mathrm{tr}[R_{\bar{\xi}},R_\xi]^2=4\left(B_1^2+\sum_{i=1}^{N-2}(B_i-B_{i+1})^2+B_{N-1}^2\right), \tag{30}$$

where $B_i=|V_i|^2/|V_{i-1}|^2(1-\cos(g_i-g_{i-1}))$ and $[R_{\bar{\xi}},R_\xi]=2\Sigma_{l=1}^{N-1}(P_{l-1}-P_l)B_l$.

Since $|V_k|^2/|V_{k-1}|^2=k(N-k)(1+|\xi|^2)^{-2}$ (see the Appendix) all terms in (26) have a factor $(1+|\xi|^2)^{-2}$ and the integration over ξ and $\bar{\xi}$ is a topological constant, i.e., $i\int d\xi\, d\bar{\xi}(1+|\xi|^2)^{-2}=2\pi$. Thus we get

$$E=\frac{1}{6\pi}\int r^2\,dr\Bigg\{-\frac{1}{N}\left(\sum_{i=0}^{N-2}\dot{g}_i\right)^2+\sum_{i=0}^{N-2}\dot{g}_i^2+\frac{1}{2r^2}\sum_{k=1}^{N-1}(\dot{g}_k-\dot{g}_{k-1})^2D_k$$
$$+\frac{2}{r^2}\sum_{k=1}^{N-1}D_k+\frac{1}{4r^4}\left(D_1^2+\sum_{k=1}^{N-2}(D_k-D_{k+1})^2+D_{N-1}^2\right)\Bigg\}, \tag{31}$$

where $D_k=k(N-k)(1-\cos(g_k-g_{k-1}))$.

Let us, for simplicity, take $F_k=g_k-g_{k+1},(k=0,...,N-2)$ with $F_{N-2}=g_{N-2}$. Then, the variation of the integrand of the energy $\tilde{E}$ with respect to the functions $\dot{F}_l$ (for $l=0,...,N-2$) is

$$\frac{\partial\tilde{E}}{\partial\dot{F}_l}=\left[-\frac{2(l+1)}{N}\sum_{i=0}^{N-2}(i+1)\dot{F}_i+2\sum_{k=0}^{l}\left(\sum_{i=k}^{N-2}\dot{F}_i\right)+\frac{1}{r^2}\dot{F}_lD_{l+1}\right]r^2, \tag{32}$$

where $D_k=k(N-k)(1-\cos F_{k-1})$.

Therefore, the equations of motion for the functions F_i, and thus for the profile functions, are

$$-\frac{2(l+1)}{N}\sum_{i=0}^{N-2}(i+1)\ddot{F}_i+2\sum_{k=0}^{l}\sum_{i=k}^{N-2}\ddot{F}_i+\frac{1}{r^2}\ddot{F}_l(l+1)(N-l-1)(1-\cos F_l)$$
$$+\frac{1}{2r^2}\dot{F}_l^2(l+1)(N-l-1)\sin F_l+\frac{2}{r}\left(-\frac{2(l+1)}{N}\sum_{i=0}^{N-2}(i+1)\dot{F}_i+2\sum_{k=0}^{l}\left(\sum_{i=k}^{N-2}\dot{F}_i\right)\right)$$
$$-\frac{2}{r^2}(l+1)(N-l-1)\sin F_l-\frac{1}{r^4}(l+1)^2(N-l-1)^2(1-\cos F_l)\sin F_l$$
$$+\frac{1}{2r^4}(l+1)(N-l-1)\sin F_l[l(N-l)(1-\cos F_{l-1})$$
$$+(l+2)(N-l-2)(1-\cos F_{l+1})]=0. \tag{33}$$

Equation (33) can be solved numerically by imposing the appropriate boundary conditions on the profile functions. To do this we have to specialize to a particular model, i.e., for specific N and diagonalize the terms involving the second derivatives. The simplest cases: $N=2$, $N=3$, and $N=4$ involve 1, 2, or 3 functions and will be discussed in the next sections.

IV. TOPOLOGICAL PROPERTIES AND SYMMETRIES

Before we discuss special cases, let us first investigate the general topological properties of our fields.

The topological charge (3), which in many applications of the Skyrme model is identified with the baryon number, is given by

$$B=\frac{1}{8\pi^2}\int dr\,d\xi\,d\bar{\xi}\,\mathrm{tr}(R_r[R_{\bar{\xi}},R_\xi]), \tag{34}$$

when written in the complex coordinates.

Due to (21) and (13) the terms involving $\dot{g}_i/N$ in R_r after taking the trace cancel and the expression for the baryon number simplifies to

$$B=-\frac{i}{4\pi^2}\int dr\,d\xi\,d\bar{\xi}\sum_{i=0}^{N-2}\dot{F}_i(1-\cos F_i)\frac{|V_{i+1}|^2}{|V_i|^2}$$
$$=\frac{1}{2\pi}\int dr\sum_{i=0}^{N-2}\dot{F}_i(1-\cos F_i)(i+1)(N-i-1)$$
$$=\frac{1}{2\pi}\sum_{i=0}^{N-2}(i+1)(N-i-1)(F_i-\sin F_i)_{r=0}^{r=\infty}. \tag{35}$$

As $g_i(\infty)=0$ (required for the finiteness of the energy) the only contributions to the topological charge come from $F_i(0)$.

The $N-1$ equations for the profile functions and their differences given in (33) have many symmetries. These symmetries can be used to derive special skyrmion solutions which involve a smaller number of profile functions and projectors.

The main symmetry of our expressions, are the independent interchanges

$$F_k\leftrightarrow F_{N-k-2} \quad \text{for } k=0,...,N-2. \tag{36}$$

This symmetry follows from the fact the terms $D_k=k(N-k)(1-\cos F_{k-1})$ which appear in the energy are symmetric under the interchange: $D_k\leftrightarrow D_{N-k}$ when $F_{k-1}\leftrightarrow F_{N-k-1}$. In addition, all the other terms in the energy also exhibit this symmetry since they are combinations of F_i and their derivatives.

V. SPHERICAL SKYRMIONS

The simplest case corresponds to the SU (2) spherically symmetric skyrmion. This is the solution which was found 30 years ago by Skyrme and is usually presented in terms of the well-known hedgehog ansatz.

A. SU (3) Skyrme model

In this case $N=3$ and we have two functions: F_0 and F_1. The energy density $\mathcal{E}$, such that $E=(6\pi)^{-1}\int\mathcal{E}r^2\,dr$, is given by

$$\mathcal{E}=\frac{2}{3}(\dot{F}_0^2+\dot{F}_1^2+\dot{F}_0\dot{F}_1)+\frac{1}{r^2}[(\dot{F}_0^2+4)(1-\cos F_0)+(\dot{F}_1^2+4)(1-\cos F_1)]$$
$$+\frac{2}{r^4}[(1-\cos F_0)^2-(1-\cos F_0)(1-\cos F_1)+(1-\cos F_1)^2], \tag{37}$$

and the equations for the profile functions are

$$\ddot{F}_0\left(1+\frac{3}{2r^2}(1-\cos F_0)\right)+\frac{\ddot{F}_1}{2}+\frac{2\dot{F}_0+\dot{F}_1}{r}+\frac{3\sin F_0}{4r^2}\left[\dot{F}_0^2-4-\frac{4(1-\cos F_0)}{r^2}+\frac{2(1-\cos F_1)}{r^2}\right]=0,$$
$$\ddot{F}_1\left(1+\frac{3}{2r^2}(1-\cos F_1)\right)+\frac{\ddot{F}_0}{2}+\frac{2\dot{F}_1+\dot{F}_0}{r}+\frac{3\sin F_1}{4r^2}\left[\dot{F}_1^2-4-\frac{4(1-\cos F_1)}{r^2}+\frac{2(1-\cos F_0)}{r^2}\right]=0. \tag{38}$$

These equations can be solved numerically when the right boundary conditions have been imposed.

However, by letting $F_0=F_1=F$ (i.e., using the symmetry) they reduce to

$$\ddot{F}\left(1+\frac{1-\cos F}{r^2}\right)+\frac{2\dot{F}}{r}+\frac{\sin F}{2r^2}\left[\dot{F}^2-4-\frac{2(1-\cos F)}{r^2}\right]=0. \tag{39}$$

Equation (39) coincides with the equation for the profile function of a single SU (2) skyrmion. Here we note that as $F_0(0)=F_1(0)=2\pi$ the topological charge of our solution is four. Thus the energy of this configuration, which corresponds to four skyrmions is $E_{B=4}=4\times 1.232$, i.e., is exactly four times the energy of one skyrmion. We see that we have a static solution corresponding to four noninteracting skyrmions, placed on top of each other in such a way that their energy density is spherically symmetric.

In addition, there is a further symmetry which allows us to set $F_0=-F_1=G$. In this case the equations reduce to

$$\ddot{G}\left(\frac{1}{2}+\frac{3}{2r^2}(1-\cos G)\right)+\frac{\dot{G}}{r}+\frac{3\sin G}{4r^2}\left[\dot{G}^2-4-\frac{2(1\text{-}\cos G)}{r^2}\right]=0. \tag{40}$$

Let us note that, since $F_0=g_0-g_1$ and $F_1=g_1$, this case corresponds to $g_0=0$ and thus, the field (9) involves only one projector, namely P_1. This solution is the topologically trivial solution discussed in Ref. 6 and its energy is 3.861.

Finally, the Balachandran *et al.* skyrmion solution can be obtained from (38) by imposing the following boundary conditions: $g_0(0)=2\pi$, $g_1(0)=0$, and $g_0(\infty)=0$, $g_1(\infty)=0$; its energy is $E_{B=2}=2.3764$.

B. SU (4) Skyrme model

In this case the energy density becomes

$$\begin{aligned}\mathcal{E}=&\frac{1}{4}(3\dot{F}_0^2+4\dot{F}_1^2+3\dot{F}_2^2+4\dot{F}_0\dot{F}_1+4\dot{F}_1\dot{F}_2+2\dot{F}_0\dot{F}_2)+\frac{1}{2r^2}[3(\dot{F}_0^2+4)(1-\cos F_0)\\&+4(\dot{F}_1^2+4)(1-\cos F_1)+3(\dot{F}_2^2+4)(1-\cos F_2)]+\frac{1}{2r^4}\{9(1-\cos F_0)^2+16(1-\cos F_1)^2\\&+9(1-\cos F_2)^2-12(1-\cos F_0)(1-\cos F_1)-12(1-\cos F_1)(1-\cos F_2)\},\end{aligned} \tag{41}$$

while the equations for F_0, F_1, and F_2 are more complicated:

$$\begin{aligned}&\ddot{F}_0\left(1+\frac{2(1-\cos F_0)}{r^2}\right)+\frac{2\ddot{F}_1+\ddot{F}_2}{3}+\frac{3\dot{F}_0+4\dot{F}_1+2\dot{F}_2}{3r}\\&\quad+\frac{\sin F_0}{r^2}\left[\dot{F}_0^2-4-\frac{6(1-\cos F_0)}{r^2}+\frac{4(1-\cos F_1)}{r^2}\right]=0,\\&\ddot{F}_1\left(1+\frac{2(1-\cos F_1)}{r^2}\right)+\frac{\ddot{F}_0+\ddot{F}_2}{2}+\frac{2\dot{F}_1+\dot{F}_0+\dot{F}_2}{r}\\&\quad+\frac{\sin F_1}{r^2}\left[\dot{F}_1^2-4-\frac{8(1-\cos F_1)}{r^2}+\frac{3(1-\cos F_0)}{r^2}+\frac{3(1-\cos F_2)}{r^2}\right]=0,\\&\ddot{F}_2\left(1+\frac{2(1-\cos F_2)}{r^2}\right)+\frac{\ddot{F}_0+2\ddot{F}_1}{3}+\frac{6\dot{F}_2+2\dot{F}_0+4\dot{F}_1}{3r}\\&\quad+\frac{\sin F_2}{r^2}\left[\dot{F}_2^2-4-\frac{6(1-\cos F_2)}{r^2}+\frac{4(1-\cos F_1)}{r^2}\right]=0.\end{aligned} \tag{42}$$

These equations have the previously mentioned symmetry $F_k \leftrightarrow F_{N-k-2}$, which allows us to set $F_0 = F_2$ by keeping F_1 arbitrary.

In addition, letting $F_0 = F_1 = F_2 = F$ the above-mentioned system reduces to Eq. (39) and therefore, the configuration which consists of ten skyrmions (as $B = [3F_0(0) + 4F_1(0) + 3F_2(0)]/2\pi = 10$) is $E_{B=10} = 10 \times 1.232$, i.e., is exactly ten times the energy of one skyrmion. Once again we have a solution describing many skyrmions, which are noninteracting and whose energy density is spherically symmetric.

In addition, letting $F_0 = -F_2 = G$ we spot that when $F_1 = 0$, we have a solution of the form

$$\ddot{G}\left(1 + \frac{3(1-\cos G)}{r^2}\right) + \frac{2\dot{G}}{r} + \frac{3\sin G}{2r^2}\left[\dot{G}^2 - 4 - \frac{6(1-\cos G)}{r^2}\right] = 0, \tag{43}$$

which corresponds to a nontopological solution, i.e., its baryon number is zero; however, the corresponding configuration consists of three skyrmions and three antiskyrmions. [Recall that the profile functions are $g_0 = 0$ and $g_1 = g_2$, i.e., the field (9) involves only one projector of rank two—namely $P_1 + P_2$.]

In general, however, the solutions depend on more functions. We can always assume that the functions F_i go to zero at infinity, so the topological charge of a solution is determined, using (35), by the boundary value of each F_i at the origin. When each of these values is positive the solution is a mixture of skyrmions. When some F_i's take positive and some F_i's take negative values at the origin the solution corresponds to a mixture of skyrmions and antiskyrmions. This is very similar to what happens in the two-dimensional Grasmannian sigma model.[8]

We have solved numerically Eq. (42) taking all combinations, modulo the exchange of F_0 and F_2, of 0, 2π, and -2π for the value of F_i at the origin. The results are summarized as follows.

$F_0(0)$	$F_1(0)$	$F_2(0)$	B	Energy	E/baryon	SU(2) En
2π	0	0	3	3.518	1.173	3.429
0	2π	0	4	4.788	1.197	4.464
2π	0	2π	6	7.225 53	1.204	6.654
2π	2π	0	7	8.452 19	1.207	7.7693
2π	2π	2π	10	12.32	1.232	⋯
2π	-2π	2π	6-4	8.852	0.8852	⋯
2π	2π	-2π	7-3	9.896	0.9896	⋯
2π	0	-2π	3-3	6.634 22	1.106	⋯
-2π	2π	0	4-3	6.614 78	0.945	⋯

The first five solutions are bound states of skyrmions with energies larger than the energies of the corresponding SU (2) solutions.[3] Moreover, the excitation energy of the first two solutions is very small. As previously mentioned, the energy of the $B = 10$ solution is exactly ten times the energy of a single skyrmion solution. These solutions are all spherically symmetric and thus they are all more symmetrical than the corresponding SU (2) solutions.

The last four solutions are bound states of skyrmions and antiskyrmions. Although their energies are very small, we know that these solutions must be unstable.

VI. CONCLUSIONS

In this paper we have shown how to construct radially symmetric solutions of the SU (N) Skyrme models. In the general case these solutions depend on $N-1$ profile functions which have to be determined numerically. In some cases we can exploit symmetries of our expressions and reduce the number of functions. Thus in the case of SU (3) we can recover the topologically trivial solution discussed in Ref. 6.

We have not discussed the derived solutions in much detail. Their properties and their relation to physics deserve further study and these topics are currently under investigation.

ACKNOWLEDGMENT

We thank C. Houghton, V. Kopeliovich, and P. Sutcliffe for their interest.

APPENDIX: MODULUS OF $|P^k{}_+ f|^2$

Here we prove Eq. (11) for f given by (10). This result can be proved by induction. Note that, the modulus of the N-dimensional vector f is $|f|^2=(1+|\xi|^2)^{N-1}$ and therefore, $|P_+f|^2|f|^2=|f|^2|\partial_+ f|^2-|f^\dagger\partial_+ f|^2=(N-1)(1+|\xi|^2)^{N-3}|f|^2$, therefore, $|P_+f|^2=(N-1)(1+|\xi|^2)^{N-3}$.

As

$$P_+^{k+1}f=\partial_\xi P_+^k f-P_+^k f\,\frac{(P_+^k f)^\dagger\partial_\xi P_+^k f}{|P_+^k f|^2}, \tag{A1}$$

using (8) we get

$$|P_+^{k+1}f|^2|P_+^k f|^2=|\partial_\xi P_+^k f|^2|P_+^k f|^2-|\partial_\xi|P_+^k f|^2|^2, \tag{A2}$$

$$\partial_\xi\partial_{\bar\xi}|P_+^k f|^2=|\partial_\xi P_+^k f|^2+(P_+^k f)^\dagger\partial_{\bar\xi}\partial_\xi P_+^k f, \tag{A3}$$

$$(P_+^k f)^\dagger\partial_\xi\partial_{\bar\xi}(P_+^k f)=-(P_+^k f)^\dagger\partial_\xi(P_+^{k-1}f)\,\frac{|P_+^k f|^2}{|P_+^{k-1}f|^2}=-\frac{|P_+^k f|^4}{|P_+^{k-1}f|^2}, \tag{A4}$$

which finally, leads to

$$|P_+^{k+1}f|^2=\partial_\xi\partial_{\bar\xi}|P_+^k f|^2+\frac{|P_+^k f|^4}{|P_+^{k-1}f|^2}-\frac{|\partial_\xi|P_+^k f|^2|^2}{|P_+^k f|^2}. \tag{A5}$$

Therefore if $|P_+^k f|^2=\alpha(1+|\xi|^2)^{N-2k-1}$ and $|P_+^{k-1}f|^2=\beta(1+|\xi|^2)^{N-2k\pm1}$ then

$$|P_+^{k+1}f|^2=\gamma(1+|\xi|^2)^{N-2k-3}, \tag{A6}$$

where $\gamma=\alpha(N-2k-1)+\alpha^2/\beta$. To find γ we again use induction, recalling that the coefficients of the two lowest terms in the sequence are 1 and $N-1$, respectively. Then it is easily seen that

$$|P_+^k f|^2=k!(N-1)(N-2)\cdots(N-k)(1+|\xi|^2)^{N-2k-1}, \tag{A7}$$

in which the last term in the sequence corresponds to $k=N-1$.

[1] T. H. R. Skyrme, Nucl. Phys. **31**, 556 (1962).
[2] E. Witten, Nucl. Phys. B **223**, 422 (1983).
[3] R. A. Battye and P. M. Sutcliffe, Phys. Rev. Lett. **79**, 363 (1997).
[4] A. P. Balachandran, A. Barducci, F. Lizzi, V. G. J. Rodgers, and A. Stern, Phys. Rev. Lett. **52**, 887 (1984).
[5] V. B. Kopeliovich, B. E. Schwesinger, and B. E. Stern, JETP Lett. **62**, 185 (1995).
[6] T. Ioannidou, B. Piette, and W. J. Zakrzewski, ''Low Energy States in the SU(N) Skyrme Model,'' hep-th/9811071 1999.
[7] C. J. Houghton, N. S. Manton, and P. M. Sutcliffe, Nucl. Phys. B **510**, 507 (1998).
[8] W. J. Zakrzewski, *Low Dimensional Sigma Models* (IOP, Bristol, 1989).
[9] A. Din and W. J. Zakrzewski, Nucl. Phys. B **174**, 397 (1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Two algebraic properties of thermal quantum field theories

Christian D. Jäkel[a)]
Dipartimento di Matematica, Via della Ricerca Scientifica, Università di Roma ''Tor Vergata,'' I-00133 Roma, Italy

(Received 26 April 1999; accepted for publication 23 August 1999)

We establish the Schlieder and the Borchers property for thermal field theories. In addition, we provide some information on the commutation and localization properties of projection operators.
[S0022-2488(99)00312-6]

I. INTRODUCTION

Recently, the author has explored the general structure of thermal field theories in some detail. Here we would like to fill in a gap which concerns two basic results, namely, the Borchers and the Schlieder property. Both results will look familiar to the experts and even the proofs which we will present here are more or less standard (see, e.g., Ref. 1 for a convenient collection of the corresponding results in the vacuum sector). However, a close inspection shows that the fundamental differences between thermal and vacuum QFT are clearly reflected in slightly different assumptions and consequences. For completeness we add a list of properties (due to Florig and Summers[2]) which are all equivalent to the Schlieder property.

In the algebraic formulation (as described in the monograph by Haag)[3] a QFT is cast into an inclusion preserving map

$$\mathcal{O} \rightarrow \mathcal{A}(\mathcal{O}) \tag{1}$$

which assigns to any open bounded region $\mathcal{O}$ in Minkowski space $\mathbb{R}^4$ a unital C^*-algebra $\mathcal{A}(\mathcal{O})$. The Hermitian elements of the *abstract* C^*-algebra $\mathcal{A}(\mathcal{O})$ are interpreted as the observables which can be measured at times and locations in $\mathcal{O}$. The physical states are described by positive, linear, and normalized functionals. By the GNS-construction, any state ω on $\mathcal{A}$ gives rise to a Hilbert space $\mathcal{H}_\omega$ and a representation π_ω together with a cyclic vector Ω_ω, such that

$$\omega(a) = (\Omega_\omega, \pi_\omega(a)\Omega_\omega) \quad \forall a \in \mathcal{A} = \overline{\bigcup_{\mathcal{O}\subset \mathbb{R}^4} \mathcal{A}(\mathcal{O})}^{C^*}. \tag{2}$$

The representation π_ω automatically determines the values of certain macroscopic observables in all states, which are normal (A linear functional on $\mathcal{A}$ is said to be normal relative to π_ω, if it is continuous with respect to the ultraweak topology determined by π_ω. Since normal states differ only locally from ω, various global physical situations will manifest themselves in unitarily inequivalent GNS-representations.) w.r.t. π_ω (these are exactly those states which can be specified by density matrices $\rho \in \mathcal{B}(\mathcal{H}_\omega)$, $\rho > 0$, $\mathrm{Tr}\rho = 1$). Thus a state is specified macroscopically by a representation and microscopically by a density matrix.

The relevant states describing thermal equilibrium, the so-called KMS-states, will soon be distinguished within the set of all time-invariant normalized, positive linear functionals of $\mathcal{A}$ by their stability properties with respect to timelike translations. Since the associated GNS-representations will allow a unitary implementation of the time-evolution, we will take advantage

[a)]Present address: Institut f. theor. Physik, Universität Innsbruck, A-6020 Innsbruck, Austria. Electronic mail: cjaekel@esi.ac.at

of it and formulate our problems in the better developed Hilbert space setting. But before we do so, we should mention that the net $\mathcal{O}\to\mathcal{A}(\mathcal{O})$ satisfies a number of properties which do not depend on the representation:

(i) The net $\mathcal{O}\to\mathcal{A}(\mathcal{O})$ is isotonous, i.e., there exists a unital embedding

$$\mathcal{A}(\mathcal{O}_1)\hookrightarrow\mathcal{A}(\mathcal{O}_2) \qquad \text{if} \quad \mathcal{O}_1\subset\mathcal{O}_2. \tag{3}$$

This property, called isotony, allows us to consider the quasilocal algebra $\mathcal{A}$ which is defined in (2) as the C^*-inductive limit of the local algebras. The elements of $\mathcal{A}$ are called quasilocal observables; they can be approximated in norm topology by strictly local elements; the total energy, total charge, etc., are considered as unobservable; these quantities refer to infinitely extended regions and cannot be controlled by local measurements.

(ii) Observables localized in spacelike separated space–time regions commute,

$$\mathcal{A}(\mathcal{O}_1)\subset\mathcal{A}^c(\mathcal{O}_2) \qquad \text{if} \quad \mathcal{O}_1\subset\mathcal{O}_2'. \tag{4}$$

Here $\mathcal{O}'$ denotes the spacelike complement of $\mathcal{O}$ and $\mathcal{A}^c(\mathcal{O})$ denotes the set of operators in $\mathcal{A}$ which commute with all operators in $\mathcal{A}(\mathcal{O})$.

(iii) The space–time symmetry of Minkowski space manifests itself in the existence of a representation

$$\alpha:(\Lambda,x)\mapsto\alpha_{\Lambda,x}\in\mathrm{Aut}(\mathcal{A}), \quad (\Lambda,x)\in\mathcal{P}_+^\uparrow, \tag{5}$$

of the (orthochronous) Poincaré group $\mathcal{P}_+^\uparrow$. Lorentz-transformations Λ and space–time translations x act geometrically,

$$\alpha_{\Lambda,x}(\mathcal{A}(\mathcal{O}))=\mathcal{A}(\Lambda\mathcal{O}+x) \quad \forall(\Lambda,x)\in\mathcal{P}_+^\uparrow. \tag{6}$$

For the present letter we may restrict our list of assumptions to include only the (strongly continuous) one-parameter subgroup of time-translations $\tau:\mathbb{R}\to\mathrm{Aut}(\mathcal{A})$. Of course, it acts geometrically, i.e.,

$$\tau_t(\mathcal{A}(\mathcal{O}))=\mathcal{A}(\mathcal{O}+te) \quad \forall t\in\mathbb{R}. \tag{7}$$

Here e is a unit vector denoting the time direction with respect to a given Lorentz-frame.

Remark: Let $h\in L^1(\mathbb{R},dt)$ such that the Fourier-transform $\tilde{h}$ of h has compact support. If the group of automorphisms $\tau:\mathbb{R}\to\mathrm{Aut}(\mathcal{A})$ is strongly continuous, then the Bochner integral

$$a_h=\int dt\,h(x)\,\tau_t(a), \quad a\in\mathcal{A}, \tag{8}$$

exists in $\mathcal{A}$ and defines an entire analytic element for the time-translations. Recall that $b\in\mathcal{A}$ is called an analytic element for the group τ, if there exists some $\lambda>0$, a strip $S(-\lambda,\lambda):=\{z\in\mathbb{C}:|\mathfrak{J}z|<\lambda\}$ and a function $g:S(-\lambda,\lambda)\to\mathcal{A}$ such that

(a) $g(t)=\tau_t(b)$ for all $t\in\mathbb{R}$;

(b) $z\mapsto\omega(g(z))$ is analytic for all states ω over $\mathcal{A}$.

An analytic element is called entire analytic, if $\lambda=\infty$; i.e., $S(-\lambda,\lambda)=\mathbb{C}$. $\mathcal{A}_\tau$ will denote the set of entire analytic elements for the group τ.

We now turn to thermal equilibrium states. Kubo[4] and subsequently Martin and Schwinger[5] found out that the Green's functions of finite-volume Gibbs-states satisfy an auxiliary boundary condition in the complex plane with respect to the time-evolution. The crucial step was to recognize that the so-called KMS-condition not only characterizes the finite-volume Gibbs-states but remains valid in the thermodynamic limit.[6] Nowadays, the KMS-condition is commonly accepted as the appropriate criterion for equilibrium of finite and infinite systems. However, only recently Buchholz and Junglas have shown that a large class of relativistic models admits KMS-states.[7]

Definition: A state ω_β over $\mathcal{A}$ is called a (τ, β)-KMS-state for some $\beta\in\mathbb{R}\cup\{\pm\infty\}$, if

$$\omega_\beta(a\,\tau_{i\beta}(b))=\omega_\beta(ba) \tag{9}$$

for all a, b in a norm dense, τ-invariant*-subalgebra of $\mathcal{A}_\tau$.

The GNS-representation $(\pi_\beta, \mathcal{H}_\beta, \Omega_\beta)$ of the net $\mathcal{O} \to \mathcal{A}(\mathcal{O})$ associated with a KMS-state ω_β assigns to any $\mathcal{O} \subset \mathbb{R}^4$ a von Neumann algebra

$$\mathcal{R}_\beta(\mathcal{O}) = \pi_\beta(\mathcal{A}(\mathcal{O}))''. \tag{10}$$

$\mathcal{R}_\beta := \pi_\beta(\mathcal{A})''$ possesses a cyclic (due to the GNS-construction) and separating (due to the KMS-condition) vector, namely, Ω_β.

The general analysis of KMS-states (see, e.g., Ref. 8) extends a number of results well known in classical ergodic theory to the noncommutative case. For instance, the set of KMS-states for any fixed $\beta > 0$ is a weak-* compact, convex set. In fact, an arbitrary KMS-state can be represented in a unique manner as a convex superposition of extremal KMS-states. (A KMS-state is called extremal, if it cannot be decomposed into other KMS-states.) Moreover, KMS-states can be distinguished within the set of all (physical) states from first principles. In a number of pioneering articles it has been demonstrated that the extremal KMS-states of an infinitely extended medium change continuously as the Hamiltonian is perturbed slightly. This condition characterizes the extremal KMS-states; they are precisely those states which are distinguished among (possible other) stationary states by the fact that they turn continuously into the unperturbed states as a certain family of perturbations tends to zero.[9,10] The same condition may also be interpreted as an adiabatic invariance.[11] Extremal KMS-states return to their original form at the end of a procedure in which the dynamical law is changed by a local perturbation which is slowly switched on and, as $t \to \infty$ slowly switched off again. A further important characteristic of KMS-states is their passivity,[12] which is the requirement that the energy of the system at time t can only have increased if the Hamiltonian depends on the time and has returned to its initial form at time t. This condition is just the second law of thermodynamics; it fixes the sign of β and means that no energy can be removed from a KMS-state having $\beta > 0$, just as a periodic process can extract no energy from the ground state.

Although the representation independent aspects of the map $\mathcal{O} \to \mathcal{A}(\mathcal{O})$ clearly deserve attention [it seems that the abstract operator algebraic formulation is inevitable for the description of nonequilibrium situations in which also the macroscopic observables (e.g., the specific heat, the mean magnetization, etc.) will change in the course of time], we will now specify our results and assumptions in the more restrictive Hilbert space framework. To be precise, in the remainder of this letter we will consider a thermal field theory (TFT), specified by a von Neumann algebra $\mathcal{R}_\beta$ with a cyclic and separating vector Ω_β and a net of subalgebras

$$\mathcal{O} \to \mathcal{R}_\beta(\mathcal{O}) \tag{11}$$

subject to the following conditions:

(i) The subalgebras associated with spacelike separated space–time regions commute, i.e.,
$$\mathcal{R}_\beta(\mathcal{O}_1) \subset \mathcal{R}_\beta(\mathcal{O}_2)' \qquad \text{if} \quad \mathcal{O}_1 \subset \mathcal{O}_2'. \tag{12}$$
Note that $\mathcal{R}_\beta(\mathcal{O})'$ denotes the commutant of $\mathcal{R}_\beta(\mathcal{O})$ in $\mathcal{B}(\mathcal{H}_\beta)$; i.e., $\mathcal{R}_\beta(\mathcal{O})'$ includes both $\mathcal{R}_\beta'$ and $\mathcal{R}_\beta(\mathcal{O}')$ as subalgebras.

(ii) The time-evolution is unitarily implemented by the modular group $t \mapsto \Delta^{it}$ (see, e.g., Ref. 8) associated with the pair $(\mathcal{R}_\beta, \Omega_\beta)$, i.e.,
$$\Delta = e^{-\beta H_\beta} \quad \text{and} \quad \Delta^{-it/\beta} \mathcal{R}_\beta(\mathcal{O}) \Delta^{it/\beta} = \mathcal{R}_\beta(\mathcal{O} + te) \quad \forall t \in \mathbb{R}. \tag{13}$$
Here H_β denotes the generator of the time-evolution and e is the unit vector denoting the time direction w.r.t. the distinguished rest frame.

(iii) $\mathcal{H}_\beta$ is separable and Ω_β is the unique—up to a phase—time-invariant vector in $\mathcal{H}_\beta$.

(iv) Ω_β is cyclic for $\mathcal{R}_\beta(\mathcal{O})$, where $\mathcal{O}$ is any open subset of $\mathbb{R}^4$, i.e.,
$$\overline{\mathcal{R}_\beta(\mathcal{O}) \Omega_\beta} = \mathcal{H}_\beta. \tag{14}$$
Ω_β shares the "Reeh–Schlieder property" (14) with the following dense set of vectors,
$$\mathcal{D}_\tau = \{ \Psi = \pi_\beta(a) \Omega_\beta \in \mathcal{H}_\beta : a, a^{-1} \in \mathcal{A}_\tau \} \subset \mathcal{R}_\beta \Omega_\beta. \tag{15}$$

We will show that under these assumptions the following statements are valid:

(i) (Schlieder property). Given two open space–time regions $\mathcal{O}$, $\hat{\mathcal{O}}$ and some $\delta > 0$ such that

$$\mathcal{O} + te \subset \mathcal{O} \qquad \text{for} \quad |t| < \delta, \tag{16}$$

the Schlieder property holds for the algebras $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$. (Recall that a pair of von Neumann algebras $\mathcal{M}$ and $\mathcal{N}$ satisfies the Schlieder property iff $0 \neq M \in \mathcal{M}$ and $0 \neq N \in \mathcal{N}$ implies that $MN \neq 0$.)

(ii) (Borchers property; also called Property B). Given a nonzero projection operator $E \in \mathcal{R}_\beta(\mathcal{O})$, where $\mathcal{O} \subset \mathbb{R}^4$ is bounded, there exists a partial isometry V in the von Neumann algebra $\mathcal{R}_\beta(\hat{\mathcal{O}})$, associated with a slightly larger region $\hat{\mathcal{O}}$, such that $V^*V = \mathbb{1}$ and $VV^* = E$. One writes

$$E \sim \mathbb{1}_{\mathrm{mod}\, \mathcal{R}_\beta(\mathcal{O})}. \tag{17}$$

Recall that a factor $\mathcal{M}$ is called type III, if $E \sim \mathbb{1}_{\mathrm{mod}\mathcal{M}}$ for all self-adjoint projections $E \in \mathcal{M}$. Thus $\mathcal{R}_\beta(\mathcal{O})$ is ''almost'' a factor of type III.

Remark: If the pair $(\mathcal{A}, \tau)$ is asymptotically Abelian in time, i.e.,

$$\lim_{t \to \infty} \| [a, \tau_t(b)] \| = 0 \quad \forall a, b \in \mathcal{A} \tag{18}$$

and ω_β is an extremal KMS state, then Ω_β is the unique—up to a phase—time-invariant vector in $\mathcal{H}_\beta$. The Reeh–Schlieder property (iv) can be derived from the relativistic KMS-condition[13] provided the net $\mathcal{O} \to \mathcal{R}_\beta(\mathcal{O})$ satisfies additivity.[14] [The net $\mathcal{O} \to \mathcal{R}_\beta(\mathcal{O})$ is called additive if $\cup_i \mathcal{O}_i = \mathcal{O} \Rightarrow \vee_i \mathcal{R}_\beta(\mathcal{O}_i) = \mathcal{R}_\beta(\mathcal{O})$. Here $\mathcal{R}_1 \vee \mathcal{R}_2$ denotes the von Neumann algebra generated by the algebras $\mathcal{R}_1$ and $\mathcal{R}_2$.] Junglas[15] has shown that the Reeh–Schlieder property of Ω_β also follows from the standard KMS-condition, if ω_β is locally normal w.r.t. the vacuum representation. Note that we do *not* require that there exists a group of unitary operators in $\mathcal{B}(\mathcal{H}_\beta)$ which implements spacelike translations, since spatial translation invariance may be spontaneously broken in a KMS-state.

II. TWO BASIC PROPERTIES OF TFTs

We start with the analogon of a result of Borchers (see Ref. 16).

Lemma II.1: Let $E \in \mathcal{R}_\beta$, $\|E\| = 1$ *and let* $F = F^* = F^2 \in \mathcal{B}(\mathcal{H}_\beta)$ *be a projection operator such that*

$$[e^{-itH_\beta} F e^{itH_\beta}, E] = 0 \quad \forall |t| < \delta. \tag{19}$$

Then $FE = 0$ *implies* $F e^{itH_\beta} E = 0$ *for all* $t \in \mathbb{R}$.

Proof: It is sufficient to show that

$$(\Phi, e^{-itH_\beta} F e^{itH_\beta} E \Psi) = 0 \quad \forall t \in \mathbb{R} \tag{20}$$

for the dense set of vectors $\Phi, \Psi \in \mathcal{D}_\tau$ introduced in (15). By construction, the vectors in $\mathcal{D}_\tau$ are entire analytic for the energy, i.e., $\mathcal{D}_\tau \subset \mathcal{D}(e^{-zH_\beta})$ for all $z \in \mathbb{C}$. Due to the KMS-relation,

$$\mathcal{R}_\beta \Omega_\beta \subset \mathcal{D}(e^{-\lambda H_\beta}) \quad \forall 0 \leq \lambda \leq \beta/2. \tag{21}$$

Thus the function

$$z \mapsto (e^{i\bar{z}H_\beta}\Phi, F e^{izH_\beta} E \Psi) \tag{22}$$

is analytic in the strip $0 < \Im z < \beta/2$, while the function

$$z \mapsto (e^{i\bar{z}H_\beta} E^* \Phi, F e^{izH_\beta} \Psi) \tag{23}$$

is analytic in the strip $-\beta/2<\mathfrak{J}z<0$. Both functions are bounded and analytic and have continuous boundary values for $\mathfrak{J}z\searrow 0$ and $\mathfrak{J}z\nearrow 0$, respectively. Now (19) implies

$$\lim_{\mathfrak{J}z\searrow 0}(e^{i\bar{z}H_\beta}\Phi,Fe^{izH_\beta}E\Psi)=\lim_{\mathfrak{J}z\nearrow 0}(e^{i\bar{z}H_\beta}E^*\Phi,Fe^{izH_\beta}\Psi)\quad\forall|t|<\delta. \tag{24}$$

Using the Edge-of-the-Wedge Theorem[17] one concludes that there exists a function

$$f_{E,F}:G_\delta\to\mathbb{C} \tag{25}$$

which is analytic on the doubly cut strip

$$G_\delta=\{z\in\mathbb{C}:-\beta/2<\mathfrak{J}z<\beta/2\}\setminus\{z\in\mathbb{C}:\mathfrak{J}z=0,|\mathfrak{R}z|\geqslant\delta\} \tag{26}$$

and satisfies

$$f_{E,F}(z)=\left\{\begin{matrix}(e^{i\bar{z}H_\beta}\Phi,Fe^{izH_\beta}E\Psi)\\(e^{i\bar{z}H_\beta}E^*\Phi,Fe^{izH_\beta}\Psi)\end{matrix}\right\}\quad\text{for}\quad\left\{\begin{matrix}0<\mathfrak{J}z<\beta/2\\-\beta/2<\mathfrak{J}z<0\end{matrix}\right\}. \tag{27}$$

Continuity and $FE=0$ imply $f_{E,F}(0)=0$. According to Lagrange's theorem $f_{E,F}(z)$ vanishes identically if 0 is a zero of infinite order. This follows from the original arguments of Borchers; put

$$t_j^{(i)}:=\frac{\delta j}{2in},\quad i\in\mathbb{N},\quad j=1,...,n. \tag{28}$$

Now set

$$f^+_{t_1^{(i)},\dots,t_n^{(i)}}(z):=(e^{i\bar{z}H_\beta}\Phi,F(t_1^{(i)})\cdots F(t_n^{(i)})e^{izH_\beta}E\Psi)\qquad\text{for}\quad 0\leqslant\mathfrak{J}z\leqslant\beta/2 \tag{29}$$

and

$$f^-_{t_1^{(i)},\dots,t_n^{(i)}}(z):=(e^{i\bar{z}H_\beta}E^*\Phi,F(t_1^{(i)})\cdots F(t_n^{(i)})e^{izH_\beta}\Psi)\qquad\text{for}\quad -\beta/2\leqslant\mathfrak{J}z\leqslant 0, \tag{30}$$

where

$$F(t_j^{(i)}):=e^{-it_j^{(i)}H_\beta}Fe^{it_j^{(i)}H_\beta}. \tag{31}$$

Both functions are analytic in the interior of their domains and bounded and continuous at the boundary. Since $|t_j^{(i)}|\leqslant\delta/2$ for all $i\in\mathbb{N}$ and $j\in\{1,...,n\}$ implies that

$$[E,e^{-itH_\beta}F(t_1^{(i)})\cdots F(t_n^{(i)})e^{itH_\beta}]=0\quad\forall|\mathfrak{R}z|<\delta/2, \tag{32}$$

the boundary values for $\mathfrak{J}z\searrow 0$ resp. $\mathfrak{J}z\nearrow 0$ coincide for $|\mathfrak{R}z|<\delta/2$. Applying the Edge-of-the-Wedge Theorem[17] one concludes that f^+ and f^- are the restrictions to the upper (resp. lower) half of the doubly cut strip $\mathcal{G}_{\delta/2}$ of a function

$$f_{t_1^{(i)},\dots,t_n^{(i)}}(z):=\left\{\begin{matrix}f^+_{t_1^{(i)},\dots,t_n^{(i)}}(z)\\f^-_{t_1^{(i)},\dots,t_n^{(i)}}(z)\end{matrix}\right\}\quad\text{for}\quad\left\{\begin{matrix}0<\mathfrak{J}z<\beta/2\\-\beta/2<\mathfrak{J}z<0\end{matrix}\right\}, \tag{33}$$

which is defined and analytic for $z\in\mathcal{G}_{\delta/2}$. The function $f_{t_1^{(i)},\dots,t_n^{(i)}}(z)$ has continuous boundary values for $z\to\partial\mathcal{G}_{\delta/2}$. Since Φ and Ψ are elements of $\mathcal{D}_\tau$ (see (15)), there exist two operators A, $B\in\pi_\beta(\mathcal{A}_\tau)$ such that $\Phi=A\Omega_\beta$ and $\Psi=B\Omega_\beta$. Applying the maximum modulus principle we obtain the following estimate:

$$\begin{aligned}|f_{t_1^{(i)},\dots,t_n^{(i)}}(z)| &\leq \sup_{w\in\partial\mathcal{G}_\epsilon} |f_{t_1^{(i)},\dots,t_n^{(i)}}(w)| \\ &\leq \max\{\|e^{(\beta/2)H_\beta}A\Omega_\beta\|\|B\|,\|A\|\|e^{(\beta/2)H_\beta}B\Omega_\beta\|\} \\ &\leq \|A\|\|B\| =: M_{\Phi,\Psi} \quad \forall z\in\mathcal{G}_{\delta/2}. \end{aligned} \tag{34}$$

For example,

$$\begin{aligned}\sup_{\{w\in\mathbb{C}:\ \mathrm{Im}\, w=\beta/2\}} f_{t_1^{(i)},\dots,t_n^{(i)}}(w) &= \sup_{s\in\mathbb{R}}(e^{(is+(\beta/2))H_\beta}A\Omega_\beta, F(t_1^{(i)})\cdots F(t_n^{(i)})e^{isH_\beta}J^2e^{-(\beta/2)H_\beta}EB\Omega_\beta) \\ &= \sup_{s\in\mathbb{R}}(e^{(\beta/2)H_\beta}A\Omega_\beta, e^{-isH_\beta}F(t_1^{(i)})\cdots F(t_n^{(i)})e^{isH_\beta}JB^*E^*\Omega_\beta) \\ &\leq \|e^{(\beta/2)H_\beta}A\Omega_\beta\|\|B\| \leq \|J\|\|A^*\Omega_\beta\|\|B\| \leq \|A\|\|B\|. \end{aligned} \tag{35}$$

Here we used $\|\Omega_\beta\|=1$, $\|E\|=1$, $\|F\|=1$ and $\|e^{-isH_\beta}\|=1$ for all $s\in\mathbb{R}$. We emphasize that at this point also the specific properties of a KMS state are used; in the last line of (35) we made use of the modular conjugation J associated with the pair $(\mathcal{R}_\beta,\Omega_\beta)$ (see, e.g., Ref. 8 for a general account on modular theory). By assumption $FE=0$, hence

$$f_{t_1^{(i)},\dots,t_n^{(i)}}(-t_j^{(i)})=0. \tag{36}$$

We conclude that inside the circle $|z|<\delta/2$ each of the functions $f_{t_1^{(i)},\dots,t_n^{(i)}}(z)$ possesses n zeros for pairwise different values of $t_j^{(i)}$. Thus all of the functions

$$\frac{f_{t_1^{(i)},\dots,t_n^{(i)}}(z)}{\Pi_{j=1}^n(z+t_j^{(i)})}, \quad i\in\mathbb{N}, \tag{37}$$

are analytic in the open disk $D_{\delta/2}$ of radius $\delta/2$ centered at the origin. Note that by definition $D_{\delta/2}\subset\mathcal{G}_{\delta/2}$. Yet the number of zeros does not change in the limit $i\to\infty$ and consequently, for $i>1$,

$$\left|\frac{f_{t_1^{(i)},\dots,t_n^{(i)}}(z)}{\Pi_{j=1}^n(z+t_j^{(i)})}\right| \leq \sup_{w\in\partial D_{\delta/2}} \frac{|f_{t_1^{(i)},\dots,t_n^{(i)}}(w)|}{\Pi_{j=1}^n|w+t_j^{(i)}|} \leq M_{\Phi,\Psi}\cdot\left(\frac{4}{\delta}\right)^n \quad \forall z\in D_{\delta/2}. \tag{38}$$

In the last inequality we used $|w+t_j^{(i)}|\geq\big||w|-|t_j^{(i)}|\big|$ and $|w|=\delta/2$ together with $|t_j^{(i)}|<\delta/4$ for $i>1$ and $j\in\{1,\dots,n\}$. Hence,

$$|f_{t_1^{(i)},\dots,t_n^{(i)}}(z)| \leq M_{\Phi,\Psi}\cdot\left(\frac{4}{\delta}\right)^n\prod_{j=1}^n |z+t_j^{(i)}| \leq \mathrm{const}\cdot|z|^n \quad \text{as } i\to\infty \quad \forall z\in D_{\delta/2}. \tag{39}$$

Because of $F^2=F$, it is obvious that $f_{0,\dots,0}(z)=f_{E,F}(z)$. The map $t\mapsto e^{itH_\beta}$ is strongly continuous, thus

$$|f_{t_1^{(i)},\dots,t_n^{(i)}}(z)-f_{0,\dots,0}(z)|\to 0 \quad \text{as } i\to\infty, \tag{40}$$

uniformly in $z\in G_{\delta/2}$. Thus

$$|f_{E,F}(z)|\leq \mathrm{const}\cdot|z|^n \quad \forall z\in D_{\delta/2}. \tag{41}$$

Hence 0 is a zero of nth order. Since $n\in\mathbb{N}$ was arbitrary, we conclude that $f_{E,F}(z)$ vanishes identically for all choices of Φ, $\Psi\in\mathcal{D}_\tau$. □

As a consequence of assumption (iii) ω_β is mixing and therefore the next lemma is more or less obvious.

Lemma II.2: Let $E, F \in \mathcal{B}(\mathcal{H}_\beta)$ *be two projection operators and assume that*

$$F e^{itH_\beta} E = 0 \quad \forall t \in \mathbb{R}. \tag{42}$$

It follows that $E\Omega_\beta \neq 0$ *implies* $F\Omega_\beta = 0$.

Proof: By assumption, Ω_β is the unique—up to a phase—normalized eigenvector for the discrete eigenvalue {0}, thus (Ref. 8)

$$0 = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt (\Omega_\beta, E e^{itH_\beta} F \Omega_\beta) = (\Omega_\beta, E\Omega_\beta)(\Omega_\beta, F\Omega_\beta) = \|E\Omega_\beta\|^2 \cdot \|F\Omega_\beta\|^2. \tag{43}$$

Hence, if $E\Omega_\beta \neq 0$, then $F\Omega_\beta = 0$. □

We add a result whose analogon in the vacuum sector is due to Schlieder (see Ref. 18, p. 220).

Corollary II.3: Let $E \in \mathcal{R}_\beta$ *be a nonzero projection and* $\xi \in \mathcal{H}_\beta$ *an arbitrary nonzero vector. It follows that the set of points*

$$\{t \in \mathbb{R} : E e^{itH_\beta} \xi \neq 0\} \tag{44}$$

is dense in $\mathbb{R}$.

Proof: For $t \in \mathbb{R}$ fixed, the set of vectors

$$\mathcal{H}_t := \{\Psi \in \mathcal{H}_\beta : E e^{iH_\beta t} \Psi = 0\} \tag{45}$$

is a closed subspace of $\mathcal{H}_\beta$. We set

$$\mathcal{H}_{]-\infty,\infty[} := \bigcap_{-\infty<t<\infty} \mathcal{H}_t . \tag{46}$$

By construction, $\mathcal{H}_{]-\infty,\infty[}$ is the intersection of closed subspaces and therefore $\mathcal{H}_{]-\infty,\infty[}$ itself is a closed subspace of $\mathcal{H}_\beta$. Let P denote the projection onto $\mathcal{H}_{]-\infty,\infty[}$. Clearly,

$$E e^{itH_\beta} P = 0 \quad \forall t \in \mathbb{R}. \tag{47}$$

Now $0 \neq E \in \mathcal{R}_\beta$ implies $E\Omega_\beta \neq 0$. Therefore Lemma II.2 implies $P\Omega_\beta = 0$. Since $e^{-iH_\beta{}^t} E e^{iH_\beta{}^t} \in \mathcal{R}_\beta$ for all $t \in \mathbb{R}$, we conclude that

$$e^{-iH_\beta t} E e^{itH_\beta} D P = 0 \quad \forall D \in \mathcal{R}_\beta', \quad \forall t \in \mathbb{R}. \tag{48}$$

Let $\hat{P}$ denote the projection onto the closed linear subspace $\overline{\mathcal{R}_\beta' \mathcal{H}_{]-\infty,\infty[}}$. Clearly,

$$E e^{itH_\beta} \hat{P} = 0 \quad \forall t \in \mathbb{R}. \tag{49}$$

But by definition P is the maximal projection such that $E e^{iH_\beta{}^t} P = 0$ holds for all $t \in \mathbb{R}$. It follows that $\hat{P} \leq P$. On the other hand

$$\mathcal{H}_{]-\infty,\infty[} \subset \overline{\mathcal{R}_\beta' \mathcal{H}_{-\infty,\infty[}} \tag{50}$$

implies $P \leq \hat{P}$, thus $P = \hat{P}$. We conclude that $P \in \mathcal{R}_\beta$. Since Ω_β is separable for $\mathcal{R}_\beta$, $P\Omega_\beta = 0$ implies $P = 0$. It follows that for any vector $\xi \in \mathcal{H}_\beta$ there exists some $t \in \mathbb{R}$ such that

$$E e^{itH_\beta} \xi \neq 0. \tag{51}$$

Now consider the projection P_ξ onto the one-dimensional subspace $\mathbb{C} \cdot \xi$ and assume there exists some $\delta > 0$ such that

$$Ee^{itH_\beta}P_\xi=0 \quad \forall|t-s|<\delta \tag{52}$$

with $s\in\mathbb{R}$ fixed. Set $P_\xi(s)=e^{isH_\beta}P_\xi e^{-isH_\beta}$. Then $P_\xi(s)=P_\xi(s)^2=P_\xi(s)^*$ is a projection and

$$Ee^{itH_\beta}P_\xi(s)=0 \quad \forall|t|<\delta. \tag{53}$$

Lemma II.1 implies that

$$Ee^{itH_\beta}P_\xi=0 \quad \forall t\in\mathbb{R}, \tag{54}$$

in contradiction to (51). Thus the set $\{t\in\mathbb{R}:Ee^{itH_\beta}\xi\neq 0\}$ does not contain any open intervall. Consequently, it is dense in $\mathbb{R}$. □

Theorem II.4: *(Schlieder property). Let $\mathcal{O}$ and $\hat{\mathcal{O}}$ denote two open (not necessarily bounded) space–time regions such that*

$$\mathcal{O}+te\subset\hat{\mathcal{O}} \quad \forall|t|<\delta, \quad \delta>0. \tag{55}$$

It follows that $0\neq A\in\mathcal{R}_\beta(\mathcal{O})$ and $0\neq B\in\mathcal{R}_\beta(\hat{\mathcal{O}})'$ implies $AB\neq 0$.

Proof: Let $A\in\mathcal{R}_\beta(\mathcal{O})$ and $B\in\mathcal{R}_\beta(\hat{\mathcal{O}})'$. We have to show that $AB=0$ implies $A=0$ or $B=0$. If either A or B is unitary, then $AB=0$ implies $B=0$ or $A=0$. Thus only the case when both A and B are not unitary has to be considered in detail. In this case one of the expressions A^*A or AA^* is unequal to $\mathbb{1}$. The same is true for B^*B or BB^*. Without loss of generality we assume that $A^*A\neq\mathbb{1}$ and $BB^*\neq\mathbb{1}$. With A^*A also the spectral projections of A^*A belong to $\mathcal{R}_\beta(\mathcal{O})$ and with BB^* also the spectral projections of BB^* belong to $\mathcal{R}_\beta(\hat{\mathcal{O}})'$. Thus

$$A^*ABB^*=0 \tag{56}$$

implies $FE=0$ for all spectral projections $E\in\mathcal{R}_\beta(\mathcal{O})$, $F\in\mathcal{R}_\beta(\hat{\mathcal{O}})'$ from the spectral resolution of A^*A and BB^*, respectively. Since $E\in\mathcal{R}_\beta(\mathcal{O})$ and $F\in\mathcal{R}_\beta(\hat{\mathcal{O}})'$ we find

$$[e^{-itH_\beta}Fe^{itH_\beta},E]=0 \quad \forall|t|<\delta. \tag{57}$$

Consequently, Lemma II.1 implies

$$Fe^{itH_\beta}E=0 \quad \forall t\in\mathbb{R} \tag{58}$$

and from Lemma II.2 it follows that $E\Omega_\beta=0$ or $F\Omega_\beta=0$. Finally, Ω_β is separating for both $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$. Thus $E=0$ or $F=0$.

Remark: The Schlieder property implies that $\mathcal{R}_\beta(\mathcal{O})$ is almost a factor, namely,

$$\mathcal{R}_\beta(\mathcal{O})\cap\mathcal{R}_\beta(\hat{\mathcal{O}})'=\mathbb{C}\cdot\mathbb{1}. \tag{59}$$

This can be seen as follows: assume

$$\mathcal{R}_\beta(\mathcal{O})\cap\mathcal{R}_\beta(\hat{\mathcal{O}})'\neq\mathbb{C}\cdot\mathbb{1}. \tag{60}$$

It follows that there exists a nontrivial projection P such that both

$$P\in\mathcal{R}_\beta(\mathcal{O})\cap\mathcal{R}_\beta(\hat{\mathcal{O}})' \quad \text{and} \quad (\mathbb{1}-P)\in\mathcal{R}_\beta(\mathcal{O})\cap\mathcal{R}_\beta(\hat{\mathcal{O}})'. \tag{61}$$

Set $A=P$ and $B=(\mathbb{1}-P)$. The Schlieder property implies $P=0$ or $\mathbb{1}-P=0$, in contradiction to the assumption that P is a nontrivial projection.

The Schlieder property is a first step towards the ‘‘statistical independence’’ of $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$. In fact, several precise conditions for ‘‘statistical independence’’ have been proposed; an overview can be found in Ref. 19. Florig and Summers have collected a list of properties which are equivalent to the Schlieder property.[2]

Corollary II.5 (Florig and Summers): Assume that $\mathcal{H}_\beta$ is separable. Let $\mathcal{O}$, $\hat{\mathcal{O}}$ denote a pair of space–time regions such that the closure of the open region $\mathcal{O}$ is contained in the interior of $\hat{\mathcal{O}}$. It follows that

(i) $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ are C^*-independent, i.e., for every state ω_1 on $\mathcal{R}_\beta(\mathcal{O})$ and every state ω_2 on $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ there exists a state ω on $\mathcal{R}_\beta$ such that $\omega_{|\mathcal{R}_{\beta(\mathcal{O})}}=\omega_1$ and $\omega_{|\mathcal{R}_{\beta(\hat{\mathcal{O}})'}}=\omega_2$.

(ii) $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ are W^*-independent, i.e., for every normal state ω_1 on $\mathcal{R}_\beta(\mathcal{O})$ and every normal state ω_2 on $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ there exists a normal state ω on $\mathcal{R}_\beta$ such that $\omega_{|\mathcal{R}_{\beta(\mathcal{O})}}=\omega_1$ and $\omega_{|\mathcal{R}_{\beta(\hat{\mathcal{O}})'}}=\omega_2$.

(iii) For any nonzero vectors $\Phi,\Psi\in\mathcal{H}_\beta$ there exists $A'\in\mathcal{R}_\beta(\mathcal{O})'$ and $B'\in\mathcal{R}_\beta(\hat{\mathcal{O}})$ such that $A'\Phi=B'\Psi\neq 0$.

(iv) The ordered pair $(\mathcal{R}_\beta(\mathcal{O}),\mathcal{R}_\beta(\hat{\mathcal{O}})')$ is strictly local; i.e., for any nonzero projection $E\in\mathcal{R}_\beta(\mathcal{O})$ and any state $\omega\in\mathcal{R}_\beta(\hat{\mathcal{O}})'_*$ there exists a state $\phi\in(\mathcal{R}_\beta(\mathcal{O})\vee\mathcal{R}_\beta(\hat{\mathcal{O}})')_*$ such that $\phi(E)=1$ and $\phi_{|\mathcal{R}_{\beta(\hat{\mathcal{O}})'}}=\omega$.

(v) For any nontrivial projections $E\in\mathcal{R}_\beta(\mathcal{O})$, $F\in\mathcal{R}_\beta(\hat{\mathcal{O}})'$ and $\lambda,\mu\in[0,1]$ there exists a state ϕ on $\mathcal{B}(\mathcal{H}_\beta)$ such that $\phi(E)=\lambda$ and $\phi(F)=\mu$.

(vi) $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ are statistically independent in the sense of Haag and Kastler; i.e., for every state ω_1 on $\mathcal{R}_\beta(\mathcal{O})$ and every state ω_2 on $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ there exists a state ω on $\mathcal{R}$ such that

$$\omega(AB)=\omega_1(A)\omega_2(B) \tag{62}$$

for all $A\in\mathcal{R}_\beta(\mathcal{O})$ and all $B\in\mathcal{R}_\beta(\hat{\mathcal{O}})'$.

(vii) $\|AB\|=\|A\|\|B\|$ for all $A\in\mathcal{R}_\beta(\mathcal{O})$ and all $B\in\mathcal{R}_\beta(\hat{\mathcal{O}})'$.

(viii) The von Neumann algebras $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ are algebraically independent; i.e., given two arbitrary sets $\{A_i:\ i=1,...,m\}$ and $\{B_j:j=1,...,n\}$ of linear independent elements of $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$, respectively, the collection $\{A_iB_j:i=1,...,m;\ j=1,...,n\}$ is linearly independent in $\mathcal{R}_\beta(\mathcal{O})\odot\mathcal{R}_\beta(\hat{\mathcal{O}})'$.

(ix) The map $\eta:(\mathcal{R}_\beta(\mathcal{O}),\mathcal{R}_\beta(\hat{\mathcal{O}})')\to\mathcal{R}_\beta(\mathcal{O})\odot\mathcal{R}_\beta(\hat{\mathcal{O}})'$ defined by

$$\eta(A,B)=A\otimes B,\quad A\in\mathcal{R}_\beta(\mathcal{O}),\quad B\in\mathcal{R}_\beta(\hat{\mathcal{O}})', \tag{63}$$

is an isomorphism continuous in the minimal C^*-cross norm on the algebraic tensor product $\mathcal{R}_\beta(\mathcal{O})\odot\mathcal{R}_\beta(\hat{\mathcal{O}})'$ and can therefore be continuously extended to a surjective homomorphism $\overline{\eta}:\mathcal{R}_\beta(\mathcal{O})\vee\mathcal{R}_\beta(\hat{\mathcal{O}})'\to\mathcal{R}_\beta(\mathcal{O})\otimes\mathcal{R}_\beta(\hat{\mathcal{O}})'$.

Remark: If $\mathcal{R}_\beta(\mathcal{O})$ is a factor of type III acting on a separable Hilbert space, then Corollary II.5 remains valid, if we replace $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ by $\mathcal{R}_\beta(\mathcal{O})'$. It is remarkable that for such a pair all normal partial states have normal extensions, none of which is allowed to be a product state, and also all partial states have extensions to product states, none of which can be normal.

Theorem II.6: *(Borchers property). Let $\mathcal{O}$ and $\hat{\mathcal{O}}$ denote two open and bounded space–time regions such that*

$$\mathcal{O}+te\subset\hat{\mathcal{O}}\quad\forall|t|<\delta,\quad\delta>0. \tag{64}$$

Given a nonzero projection $E\in\mathcal{R}_\beta(\mathcal{O})$, there exists a partial isometry $V\in\mathcal{R}_\beta(\hat{\mathcal{O}})$ such that $V^*V=\mathbb{1}$ and $VV^*=E$.

Proof: Once the Schlieder property is proven for $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$, the Borchers property follows by standard arguments (see Refs. 16, 1 for the corresponding result in the vacuum sector). We present them here for the sake of completeness only. By assumption the spacelike complement $\hat{\mathcal{O}}'$ of $\hat{\mathcal{O}}$ is not empty. Thus any vector $\Phi \in \mathcal{D}_\tau$ is cyclic for $\mathcal{R}_\beta(\hat{\mathcal{O}})' \supset \mathcal{R}_\beta(\hat{\mathcal{O}}')$. We show that $E\Phi$ is separating for $\mathcal{R}_\beta(\hat{\mathcal{O}})'$; choose a region $\mathcal{O}_0$ such that

$$\mathcal{O}_0 \subset \mathcal{O}' \cap \hat{\mathcal{O}} \tag{65}$$

and consider some $B \in \mathcal{R}_\beta(\hat{\mathcal{O}})'$ such that $BE\Phi = 0$. Locality implies that $BEC\Phi = 0$ for any $C \in \mathcal{R}_\beta(\mathcal{O}_0)$. By the Reeh–Schlieder property the set $\{C\Phi : C \in \mathcal{R}_\beta(\mathcal{O}_0)\}$ is dense in $\mathcal{H}_\beta$ and therefore $BE = 0$. Now the Schlieder property for $\mathcal{R}_\beta(\mathcal{O})$ and $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ implies $B = 0$, since by assumption $E \neq 0$. We conclude that $E\Phi$ is separating for $\mathcal{R}_\beta(\hat{\mathcal{O}})'$. Hence, the normal state

$$B \mapsto (E\Phi, BE\Phi) \tag{66}$$

is faithful on $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ and there exists a vector $\Psi \in \mathcal{H}_\beta$ cyclic for $\mathcal{R}_\beta(\hat{\mathcal{O}})'$ such that

$$(E\Phi, B^*BE\Phi) = (\Psi, B^*B\Psi) \quad \forall B \in \mathcal{R}_\beta(\hat{\mathcal{O}})'. \tag{67}$$

It follows that $V : \mathcal{H}_\beta \to \mathcal{H}_\beta$, given by

$$VB\Psi = BE\Phi \quad \forall B \in \mathcal{R}_\beta(\hat{\mathcal{O}})', \tag{68}$$

defines an isometry. Both Φ and Ψ are cyclic for $\mathcal{R}_\beta(\hat{\mathcal{O}})'$, thus V is densely defined and its range spans $E\mathcal{H}_\beta$. Moreover,

$$CVB\Psi = CBE\Phi = VCB\Psi \quad \forall C \in \mathcal{R}_\beta(\hat{\mathcal{O}})'. \tag{69}$$

Thus V commutes with all $C \in \mathcal{R}_\beta(\hat{\mathcal{O}})'$ on the dense set $\{B\Psi : B \in \mathcal{R}_\beta(\hat{\mathcal{O}})'\} \subset \mathcal{H}_\beta$ and therefore $V \in \mathcal{R}_\beta(\hat{\mathcal{O}})$. □

Remark: The Borchers property has interesting consequences for the actual preparation of states: Given an arbitrary state ω on $\mathcal{R}_\beta(\mathcal{O}) \vee \mathcal{R}_\beta(\hat{\mathcal{O}})'$, we set

$$\omega_V(C) := \omega(V^*CV) \quad \forall C \in \mathcal{R}_\beta(\mathcal{O}) V \mathcal{R}(\hat{\mathcal{O}})'. \tag{70}$$

It follows that

$$\omega_V(E) = \omega(V^*VV^*V) = \mathbb{1} \quad \text{and} \quad \omega_V(\mathbb{1} - E) = 0. \tag{71}$$

Moreover,

$$\omega_V(B) = \omega(V^*BV) = \omega(B) \quad \forall B \in \mathcal{R}(\hat{\mathcal{O}})'. \tag{72}$$

This demonstrates that the Borchers property allows us to prepare a state ω_V which satisfies the properties (71) and (72) by a strictly local operation. We emphasize that the state given remains completely unchanged in the spatial complement of $\hat{\mathcal{O}}$. This is a remarkable difference to the usual collapse of the wave-function type of preparation.

ACKNOWLEDGMENTS

Kind hospitality of the Erwin Schrödinger Institut (ESI) Vienna and the Dipartimento di Matematica, Universitá di Roma ''Tor Vergata'' is gratefully acknowledged. This work was financed by a fellowship of the Operator Algebras Network, EC TMR-Programme.

[1] C. D'Antoni, *Technical Properties of the Quasilocal Algebra, The Algebraic Theory of Superselection Sectors. Introduction and Recent Results*, in *Proceedings of the Convegno Internationale Algebraic Theory of Superselection Sectors and Field Theory*, Palermo, November 23–30, 1989, edited by D. Kastler (World Scientific, Singapore, 1990).
[2] M. Florig and S. J. Summers, ''On the statistical independence of algebras of observables,'' J. Math. Phys. **38/3**, 1318–1328 (1997).
[3] R. Haag, *Local Quantum Physics: Fields, Particles, Algebras* (Springer-Verlag, Berlin, 1992).
[4] R. Kubo, ''Statistical mechanical theory of irreversible prozesses. I,'' J. Math. Soc. Jpn. **12**, 570–586 (1957).
[5] P. C. Martin and J. Schwinger, ''Theory of many-particle systems. I,'' Phys. Rev. **115/6**, 1342–1373 (1959).
[6] R. Haag, N. M. Hugenholtz, and M. Winnink, ''On the equilibrium states in quantum statistical mechanics,'' Commun. Math. Phys. **5**, 215–236 (1967).
[7] D. Buchholz and P. Junglas, ''On the existence of equilibrium states in local quantum field theory,'' Commun. Math. Phys. **121**, 255–270 (1989).
[8] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I, II* (Springer-Verlag, New York, 1981).
[9] R. Haag, D. Kastler, and E. B. Trych-Pohlmeyer, ''Stability and equilibrium states,'' Commun. Math. Phys. **38**, 173–193 (1974).
[10] R. Haag and E. B. Trych-Pohlmeyer, ''Stability properties of equilibrium states,'' Commun. Math. Phys. **56**, 213–224 (1977).
[11] H. Narnhofer and W. Thirring, ''Adiabatic theorem in quantum statistical mechanics,'' Phys. Rev. A **26/6**, 3646–3652 (1982).
[12] W. Pusz and S. L. Woronowicz, ''Passive states and KMS states for general quantum systems,'' Commun. Math. Phys. **58**, 273–290 (1978).
[13] J. Bros and D. Buchholz, ''Towards a relativistic KMS-condition,'' Nucl. Phys. B **429**, 291–318 (1994).
[14] C. D. Jäkel, ''The Reeh–Schlieder property for thermal field theories,'' hep-th/9904049, to appear in J. Math. Phys.
[15] P. Junglas, ''Thermodynamisches Gleichgewicht und Energiespektrum in der Quantenfeldtheorie,'' dissertation, Hamburg, 1987.
[16] H. J. Borchers, ''A remark on a theorem of B. Misra,'' Commun. Math. Phys. **4**, 315–223 (1967).
[17] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics and All That* (Benjamin, New York, 1964).
[18] S. Schlieder, ''Einige Bemerkungen über Projektionsoperatoren,'' Commun. Math. Phys. **13**, 216–225 (1969).
[19] S. J. Summers, ''On the statistical independence of algebras of observables,'' Rev. Math. Phys. **2/2**, 201–247 (1990).

Nonrelativistic quantum Hamiltonian for Lorentz violation

V. Alan Kostelecký and Charles D. Lane
Physics Department, Indiana University, Bloomington, Indiana 47405

(Received 18 May 1999; accepted for publication 30 August 1999)

A method is presented for deriving the nonrelativistic quantum Hamiltonian of a free massive fermion from the relativistic Lagrangian of the Lorentz-violating standard-model extension. It permits the extraction of terms at arbitrary order in a Foldy–Wouthuysen expansion in inverse powers of the mass. The quantum particle Hamiltonian is obtained and its nonrelativistic limit is given explicitly to third order. © *1999 American Institute of Physics.* [S0022-2488(99)01712-0]

I. INTRODUCTION

Establishing the physical relevance of a Lagrangian in relativistic quantum field theory often requires a determination of its nonrelativistic content. The Foldy–Wouthuysen (FW) transformation[1] provides a systematic approach to understanding the low-energy effects of certain theories. Given the relativistic quantum Hamiltonian for a theory of massive four-component fermions, the nonrelativistic quantum Hamiltonian for the corresponding two-component particle can be derived in an expansion in inverse powers of the fermion mass.

In this work, we use generalized FW methods to investigate the quantum particle Hamiltonian that describes the physics of a free massive two-component fermion emerging from the relativistic Lagrangian of the Lorentz-violating standard-model extension.[2] This standard-model extension is based on the idea of spontaneous Lorentz breaking in an underlying theory[3] and has been used for various investigations placing constraints on possible violations of Lorentz symmetry,[2,4–15] several of which depend crucially on the nonrelativistic physics of free massive fermions. In these investigations, specific terms in the nonrelativistic Hamiltonian have been derived as needed, but a full treatment has been lacking. Here, we provide a systematic approach that permits extraction of the relevant terms in the nonrelativistic Hamiltonian at arbitrary order in the FW approximation. We obtain the quantum particle Hamiltonian and provide explicitly the form of the nonrelativistic Hamiltonian to third order. Our results are directly relevant to recent analyses of muon and clock-comparison experiments[14,15] and are expected to have substantial impact on further studies of the physical implications of the standard-model extension.

The general form of the relativistic Lagrangian for a free spin-$\frac{1}{2}$ Dirac fermion ψ of mass m in the standard-model extension is[2]

$$\mathcal{L}=\tfrac{1}{2}i\overline{\psi}(\gamma_\nu+c_{\mu\nu}\gamma^\mu+d_{\mu\nu}\gamma_5\gamma^\mu+e_\nu+if_\nu\gamma_5+\tfrac{1}{2}g_{\lambda\mu\nu}\sigma^{\lambda\mu})\overleftrightarrow{\partial}^\nu\psi$$
$$-\overline{\psi}(m+a_\mu\gamma^\mu+b_\mu\gamma_5\gamma^\mu+\tfrac{1}{2}H_{\mu\nu}\sigma^{\mu\nu})\psi. \tag{1}$$

This is a generalization of the usual relativistic Lagrangian for a free massive Dirac fermion. The Dirac matrices $\{1,\gamma_5,\gamma^\mu,\gamma_5\gamma^\mu,\sigma^{\mu\nu}\}$ have conventional properties, and the Minkowski metric $\eta_{\mu\nu}$ has signature -2. The parameters a_μ, b_μ, $c_{\mu\nu}$, $d_{\mu\nu}$, e_μ, f_μ, $g_{\lambda\mu\nu}$, and $H_{\mu\nu}$ control the extent of Lorentz violation in the theory. In a given observer inertial frame, they can be regarded as fixed real Lorentz vectors or tensors. Note that $H_{\mu\nu}$ can be taken as antisymmetric, $c_{\mu\nu}$ and $d_{\mu\nu}$ as traceless, and $g_{\lambda\mu\nu}$ as antisymmetric in the first two indices. Since Lorentz symmetry is known to be valid to high precision, any nonzero parameters in nature would need to be minuscule. We therefore restrict our attention in this work to terms linear in these parameters.

In Sec. II, the relativistic particle–antiparticle Hamiltonian H corresponding to the Lagrangian (1) is obtained. Some basic information about our procedure for extracting its FW form is discussed in Sec. III, together with our definition of the relevant FW sequence. Features of this sequence are derived in Sec. IV, and the quantum particle Hamiltonian and its nonrelativistic limit to third order are explicitly presented in Sec. V.

II. RELATIVISTIC QUANTUM HAMILTONIAN

The first step in deriving low-energy effects of the Lorentz-violating terms is to obtain the relativistic Hamiltonian H associated with the Lagrangian (1). However, methods for direct construction of H are inadequate because Eq. (1) contains couplings involving time derivatives. For example, applying the Euler–Lagrange equations to $\mathcal{L}$ and solving for H from the equation $i\partial_0\psi=H\psi$ results in a non-Hermitian Hamiltonian and a corresponding nonunitary time evolution.

One method of bypassing this technical difficulty is to perform a field redefinition $\psi=A\chi$ in the Langrangian,[5] with A chosen such that the dependence of the Lagrangian on $\partial_0\chi$ is just that of the usual Dirac Lagrangian. Then, the wave function associated with χ evolves conventionally in time. The field redefinition leaves unchanged the physics, while it causes the time-derivative couplings to be replaced by extra terms in the Lagrangian.

To implement this idea, we write the Lagrangian (1) in the forms

$$\mathcal{L}\equiv\tfrac{1}{2}i\overline{\psi}\Gamma_\nu\overleftrightarrow{\partial}^\nu\psi-\overline{\psi}M\psi=\tfrac{1}{2}i\overline{\chi}\gamma_0\overleftrightarrow{\partial}^0\chi+\tfrac{1}{2}i\overline{\chi}(\overline{A}\Gamma_jA)\overleftrightarrow{\partial}^j\chi-\overline{\chi}(\overline{A}MA)\chi, \tag{2}$$

where Γ_ν and M are defined according to the correspondence with Eq. (1), and $\overline{\psi}=\overline{\chi}\overline{A}$ with $\overline{A}:=\gamma^0A^\dagger\gamma^0$. In the second expression the Lorentz indices are separated into timelike and spacelike Cartesian components, $\mu\equiv 0$ and $j=1, 2, 3$, with summation on repeated indices understood.

The choice

$$A=1-\tfrac{1}{2}\gamma^0(\Gamma_0-\gamma_0), \quad \overline{A}=1-\tfrac{1}{2}(\Gamma_0-\gamma_0)\gamma^0 \tag{3}$$

implements the equality (2) to linear order in the parameters for Lorentz violation. Derivation of the relativistic Hamiltonian H can then proceed through the Euler–Lagrange equations, which take the form of a modified Dirac equation:

$$(i\overline{A}\Gamma_\mu A\partial^\mu-\overline{A}MA)\chi=0. \tag{4}$$

We find

$$H=-\gamma^0\overline{A}\Gamma_jAp^j+\gamma^0\overline{A}MA, \tag{5}$$

where the three-momentum of the particle is denoted p_j, and H obeys the equation $i\partial_0\chi=H\chi$.

Explicitly, the relativistic Hamiltonian can be written

$$H=m(\gamma^0+\mathcal{P}_0+\mathcal{O}_0+\mathcal{E}_0), \tag{6}$$

where

$$\begin{aligned}
m\mathcal{P}_0&:=-p_j\gamma^0\gamma^j,\\
m\mathcal{O}_0&:=[-b_0+(d_{0j}+d_{j0})p^j]\gamma_5+[a_j-(c_{jk}-c_{00}\eta_{jk})p^k]\gamma^0\gamma^j+if_jp^j\gamma_5\gamma^0\\
&\quad+i[H_{0j}+(g_{j0k}+g_{jk0})p^k]\gamma^j,\\
m\mathcal{E}_0&:=[a_0-(c_{0j}+c_{j0})p^j-me_0]+[-b_j+(d_{jk}-d_{00}\eta_{jk})p^k-\tfrac{1}{2}m\epsilon^{klm}\eta_{jm}g_{kl0}]\gamma_5\gamma^0\gamma^j\\
&\quad-[mc_{00}+e_jp^j]\gamma^0-[\tfrac{1}{2}\epsilon^{klm}\eta_{jm}H_{kl}+md_{j0}-\epsilon^{lmn}\eta_{jn}(\tfrac{1}{2}g_{lmk}-\eta_{km}g_{l00})p^k]\gamma_5\gamma^j.
\end{aligned} \tag{7}$$

In these expressions, the totally antisymmetric rotation tensor ϵ^{jkl} satisfies $\epsilon_{123}=+1$ and $\epsilon^{jkl}=-\epsilon_{jkl}$. The particular decomposition of H into the four terms in (6) is chosen for later convenience.

As an aside, we remark that the relativistic Hamiltonian is also readily found if the theory (1) is extended to include a minimal coupling to a U(1) gauge field A_μ. It suffices to replace the partial derivative $i\partial_\mu$ in Eq. (1) with the covariant derivative $iD_\mu := i\partial_\mu - qA_\mu$, where q is the particle charge. The relativistic Hamiltonian then has the same form as in Eqs. (6) and (7), except that all occurrences of p_j must be replaced with $\pi_j := p_j - qA_j$ and the term qA_0 must be added to Eq. (6). The resulting Hamiltonian is relevant, for example, for studies of Lorentz-violating effects in quantum electrodynamics.

III. DEFINITION OF THE FW SEQUENCE

In the strict nonrelativistic limit, the lower two components of the relativistic wave function χ are negligible, so the upper two components of χ suffice to determine the nonrelativistic particle behavior. However, more generally the Dirac equation couples the upper and lower components of χ. The object of the Foldy–Wouthuysen procedure is to find a (momentum-dependent) unitary transformation

$$H \mapsto \tilde{H} := e^{iS} H e^{-iS} = \exp[\mathrm{ad}(iS)]H, \tag{8}$$

where $\mathrm{ad}(X)Y := [X,Y]$, such that $\tilde{H}$ is 2×2 block diagonal. This therefore decouples the upper and lower components of the FW-transformed wave function $\phi := e^{iS}\chi$. Requiring hermiticity of S ensures that e^{iS} is unitary. It follows that $\tilde{H}$ is Hermitian and that both Hamiltonians H and $\tilde{H}$ describe the same physics. The FW transformation amounts to a unitary rotation in the Hilbert space of the free-particle states that preserves the dominance of the upper two components of the wave function. The quantum particle Hamiltonian h_{rel} and the nonrelativistic limit h we seek are given by the leading 2×2 block of $\tilde{H}$.

Solving directly for $\tilde{H}$ would be of interest but is challenging in the general case. Instead, we present a method that allows approximation of $\tilde{H}$ to arbitrary accuracy in an expansion in powers of $|\vec{p}|/m$. The basic idea is to apply a succession of transformations of the type (8), chosen so that each iteration of the transformed Hamiltonian has a smaller block off-diagonal part than the previous one. The exact FW transformation is the limit of this sequence. Although more direct approaches can yield a low-order approximation to h without the use of our method, the results derived here permit straightforward calculation of h_{rel} and of h to any desired order.

For definiteness in what follows, we work within the Dirac-Pauli representation of the Dirac matrices, for which

$$\gamma^0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma^j = \begin{pmatrix} 0 & \sigma^j \\ -\sigma^j & 0 \end{pmatrix},$$

where σ^j are the usual Pauli matrices. We define a matrix to be *even* if it is block diagonal and *odd* if it is block off-diagonal. Any 4×4 matrix X can be uniquely written as the sum of an even part and an odd part, $X = \mathrm{even}(X) + \mathrm{odd}(X)$, where $\mathrm{odd}(X) = \frac{1}{2}\gamma^0[\gamma^0, X]$ and $\mathrm{even}(X) = \frac{1}{2}\gamma^0\{\gamma^0, X\}$.

We seek a sequence of FW transformations such that the odd part of the Hamiltonian progressively decreases in some suitable matrix norm, such as $\|A\| := \max_{a,b}\{|A_{ab}|\}$ for $a,b=1,2,3,4$. In the remainder of this section, an appropriate sequence $\{H_n\}$ of Hamiltonians is introduced. For each n, we also introduce a parameter t_n that turns out to provide a measure of the size of $\mathrm{odd}(H_n)$. We show in Sec. IV that with our definition for the FW sequence roughly N iterations are needed to arrive at a nonrelativistic Hamiltonian that is even to order $(|\vec{p}|/m)^{(3^N-1)}$.

To start the FW sequence, choose

$$H_0 = m_0(\gamma^0 + \mathcal{P}_0 + \mathcal{O}_0 + \mathcal{E}_0), \tag{9}$$

where $m_0 := m$ and the terms $\mathcal{P}_0$, $\mathcal{O}_0$, and $\mathcal{E}_0$ are defined in Eq. (7). This decomposition of H_0 into four parts has the following useful properties: (i) $\mathcal{P}_0$ and $\mathcal{O}_0$ are odd; (ii) $\mathcal{E}_0$ is even; (iii) $\mathcal{O}_0$ and $\mathcal{E}_0$ are first order in parameters for Lorentz violation, so products of these quantities can be neglected; and (iv) $\mathcal{P}_0^2$ is proportional to the 4×4 Dirac identity matrix, with proportionality coefficient $t_0^2 = |\vec{p}|^2/m^2$. We choose for the initial FW transformation the Hermitian matrix S_0 defined by

$$iS_0 := \frac{1}{2m_0}\gamma^0[\mathrm{odd}(H_0)] = \tfrac{1}{2}\gamma^0(\mathcal{P}_0 + \mathcal{O}_0). \tag{10}$$

This choice ensures that the odd part of $\exp[\mathrm{ad}(iS_0)]H_0$ is smaller than the odd part of H_0.

Our FW sequence is then defined iteratively by

$$H_{n+1} := e^{iS_n} H_n e^{-iS_n} = \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{[iS_n,[iS_n,\cdots[iS_n}_{k \text{ commutations with } iS_n},H_0]\cdots]] = \exp[\mathrm{ad}(iS_n)]H_n \tag{11}$$

and

$$iS_{n+1} := \frac{1}{2m_{n+1}}\gamma^0[\mathrm{odd}(H_{n+1})]. \tag{12}$$

Note that

$$H_{n+1} = \left\{ \prod_{k=0}^{n} \exp[\mathrm{ad}(iS_k)] \right\} H_0, \tag{13}$$

where the product represents map composition.

In Sec. IV, we find that each H_{n+1} can be written in the form

$$H_{n+1} = m_{n+1}(\gamma^0 + \mathcal{P}_{n+1} + \mathcal{O}_{n+1} + \mathcal{E}_{n+1}), \tag{14}$$

where the decomposition has the following useful properties: (i) $\mathcal{P}_{n+1}$ and $\mathcal{O}_{n+1}$ are odd; (ii) $\mathcal{E}_{n+1}$ is even; (iii) $\mathcal{O}_{n+1}$ and $\mathcal{E}_{n+1}$ are first order in parameters for Lorentz violation; and (iv) $\mathcal{P}_{n+1}^2$ is proportional to the identity matrix, with proportionality coefficient t_{n+1}^2 determined by t_n^2. The existence of a decomposition of the form (14) for arbitrary n, as well as the case (9) above, is a key feature making it feasible to calculate the quantum particle Hamiltonian.

IV. CALCULATION OF THE FW SEQUENCE

To calculate the FW sequence defined in Sec. III, the explicit form is needed of the operator $\exp[\mathrm{ad}(iS_n)]$ connecting H_n to H_{n+1} according to Eq. (11). Although $\mathrm{ad}(iS_n)H_n$ can be obtained directly using the properties of the Dirac matrices, calculation of $\exp[\mathrm{ad}(iS_j)]H_n$ is more challenging because it is defined by an infinite series. To address this issue, we adopt the following approach: regard $\mathrm{ad}(iS_n)$ as a linear map on a suitable vector space $\mathcal{V}_n$ containing both H_n and H_{n+1}, and find a matrix expression of this map that can be exponentiated.

The first step in implementing this approach is to define $\mathcal{V}_n$ for each n. It is convenient to introduce $\mathcal{V}_n$ as the span of a set of basis vectors $\mathcal{B}_n$, defined in terms of the operators $\gamma^0, \mathcal{P}_n, \mathcal{O}_n, \mathcal{E}_n$ determining H_n together with the particular combinations of these four operators that determine $\mathrm{ad}(iS_n)H_n$ and thus also H_{n+1}. For each n, we define the ordered set

$$\mathcal{B}_n := \{\gamma^0, \mathcal{P}_n, \mathcal{O}_n, \mathcal{P}_n\{\mathcal{P}_n,\mathcal{O}_n\}, \gamma^0[\mathcal{P}_n,\mathcal{E}_n], \mathcal{E}_n, \gamma^0\{\mathcal{P}_n,\mathcal{O}_n\}, \mathcal{P}_n[\mathcal{P}_n,\mathcal{E}_n]\}. \tag{15}$$

The eight-dimensional vector space $\mathcal{V}_n$ is formally defined as the real span of this set, so the elements of $\mathcal{B}_n$ by definition form a (linearly independent) basis. One advantage of this vector space is its relatively small dimensionality, which makes it susceptible to practical calculation. We can thus specify a vector $V \in \mathcal{V}_n$ by eight components $V_1, \ldots, V_8$:

$$V := V_1 \gamma^0 + V_2 \mathcal{P}_n + V_3 \mathcal{O}_n + V_4 \mathcal{P}_n \{\mathcal{P}_n, \mathcal{O}_n\} + V_5 \gamma^0 [\mathcal{P}_n, \mathcal{E}_n] + V_6 \mathcal{E}_n + V_7 \gamma^0 \{\mathcal{P}_n, \mathcal{O}_n\} + V_8 \mathcal{P}_n [\mathcal{P}_n, \mathcal{E}_n] \leftrightarrow (V_1, \ldots, V_8). \tag{16}$$

For example, $H_n \leftrightarrow m_n(1,1,1,0,0,1,0,0)$.

The reader is warned to avoid confusing the properties of the elements (15) as a basis for the vector space $\mathcal{V}_n$ with their possible relationships when viewed as operators on the Hilbert space of wave functions. For example, the calculations below hold even if certain basis elements vanish as operators. Note also that for different n the corresponding vector spaces $\mathcal{V}_n$ differ *a priori*. However, since both $H_n \in \mathcal{V}_n$ and $H_{n+1} \in \mathcal{V}_n$, the vector space $\mathcal{V}_n$ is invariant under the action of $\exp[\mathrm{ad}(iS_n)]$, which means $\mathcal{V}_n \supseteq \mathcal{V}_{n+1}$ for all n.

With the above notation, we can present the results of a direct calculation of $\mathrm{ad}(iS_n)V$ for $V \in \mathcal{V}_n$ performed using the properties of the Dirac matrices:

$$\mathrm{ad}(iS_n)V \leftrightarrow (t_n^2 V_2, -V_1, -V_1, -V_7, \tfrac{1}{2}V_6 + t_n^2 V^8, 0, \tfrac{1}{2}V_2 + \tfrac{1}{2}V_3 + t_n^2 V_4, -V_5). \tag{17}$$

In this expression, t_n^2 is determined iteratively from t_{n-1}^2 through the relation

$$t_{n+1}^2 = \left(\frac{\cos t_n - \dfrac{\sin t_n}{t_n}}{\cos t_n + t_n \sin t_n} \right)^2 t_n^2. \tag{18}$$

Here and in what follows, we define functions of t_n through their power-series expressions. All relevant functions of t_n implicitly involve only powers of t_n^2 (and hence powers of $t_0^2 = |\vec{p}|^2/m^2$), so it suffices to define t_n^2. Note that $t_{n+1} \sim t_n^3$ to leading order in t_n, so $t_n \sim t_0^{(3^n)}$. This means that t_n rapidly approaches zero if $t_0 \ll 1$, which ultimately is the reason for the rapid convergence of our FW sequence.

With respect to the basis $\mathcal{B}_n$, the matrix map of $\mathrm{ad}(iS_n)$ can be extracted from Eq. (17) and is given by

$$\mathrm{ad}(iS_n) \leftrightarrow \begin{pmatrix} 0 & t_n^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \tfrac{1}{2} & 0 & t_n^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \tfrac{1}{2} & \tfrac{1}{2} & t_n^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix}. \tag{19}$$

The exponential of this matrix can be found in closed form, but its detailed expression is unimportant. It can be used to calculate $\exp[\mathrm{ad}(iS_n)]H_n$, which allows us to express H_{n+1} in terms of H_n according to Eq. (14) with

$$m_{n+1}=(\cos t_n+t_n\sin t_n)m_n\,,\quad m_{n+1}\mathcal{P}_{n+1}=\left(\cos t_n-\frac{\sin t_n}{t_n}\right)m_n\mathcal{P}_n\,,$$

$$m_{n+1}\mathcal{O}_{n+1}=\left(\cos t_n-\frac{\sin t_n}{t_n}\right)m_n\mathcal{O}_n+\frac{1}{2t_n^2}\left(\frac{\sin t_n}{t_n}-t_n\sin t_n-\cos t_n\right)m_n\mathcal{P}_n\{\mathcal{P}_n\,,\mathcal{O}_n\}$$
$$+\frac{\sin t_n}{2t_n}m_n\gamma^0[\mathcal{P}_n\,,\mathcal{E}_n], \tag{20}$$

$$m_{n+1}\mathcal{E}_{n+1}=m_n\mathcal{E}_n+(\tfrac{1}{2}\cos t_n)m_n\gamma^0\{\mathcal{P}_n\,,\mathcal{O}_n\}+\left(\frac{\cos t_n-1}{2t_n^2}\right)m_n\mathcal{P}_n[\mathcal{P}_n\,,\mathcal{E}_n].$$

A measure of the convergence of the FW sequence can be introduced using $t_n\sim t_0^{(3^n)}$. In terms of a suitable matrix norm, $\|\mathrm{odd}(H_{n+1})\|\sim t_n^2\|\mathrm{odd}(H_n)\|+t_n\|\mathcal{E}_0\|\sim t_0^{2(3^n)}\|\mathrm{odd}(H_n)\|+t_0^{(3^n)}\|\mathcal{E}_0\|$. Thus, as n grows $\|\mathrm{odd}(H_n)\|$ rapidly approaches zero as $(|\vec{p}|/m)^{3^n}$. Even a relatively small value of n can therefore produce a good approximation to the quantum particle Hamiltonian.

V. NONRELATIVISTIC QUANTUM HAMILTONIAN

The quantum particle Hamiltonian h_{rel} and its nonrelativistic quantum limit h are generated in the limit of the FW sequence studied in the last section. Next, we demonstrate how to obtain these using simple matrix multiplication, and we explicitly present h_{rel} and h to order t_0^3.

The calculation at the kth-iteration level in the FW sequence requires obtaining the composite map $\Pi_{n=0}^k\exp[\mathrm{ad}(iS_n)]$. For each n in the FW sequence, the matrix $\mathrm{ad}(iS_n)$ and the action of $\exp[\mathrm{ad}(iS_n)]$ are given with respect to the basis $\mathcal{B}_n$. Since in general the vector space $\mathcal{V}_n$ varies with n, immediate calculation of $\Pi_{n=0}^k\exp[\mathrm{ad}(iS_n)]$ by matrix multiplication is inappropriate. Instead, we first obtain the components of each matrix with respect to the special basis $\mathcal{B}_0$. Ordinary matrix multiplication can then be used to derive $\Pi_{n=0}^k\exp[\mathrm{ad}(iS_n)]$.

The matrix for each map $\exp[\mathrm{ad}(iS_n)]$ can be expressed in terms of t_n. Explicitly, the nonzero entries for $\exp[\mathrm{ad}(iS_n)]$ with respect to the basis $\mathcal{B}_0$ are:

$$\exp[\mathrm{ad}(iS_n)]\leftrightarrow\begin{pmatrix} c_n+t_n^2s_n & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & c_n-s_n & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & c_n-s_n & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & (s_n-t_n^2s_n-c_n)/2t_0^2 & -t_n^2s_n & 0 & 0 & -t_ns_n/t_0 & 0\\ 0 & 0 & 0 & 0 & -t_n^2s_n & t_ns_n/2t_0 & 0 & t_0t_ns_n\\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & t_nc_n/2t_0 & t_0t_nc_n & 0 & 0 & c_n & 0\\ 0 & 0 & 0 & 0 & -t_nc_n/t_0 & (c_n-1)/2t_0^2 & 0 & c_n \end{pmatrix}, \tag{21}$$

where we have defined $c_n:=\cos t_n\approx 1-\frac{1}{2}t_n^2$, and $s_n:=\sin t_n/t_n\approx 1-\frac{1}{6}t_n^2$.

Since $t_n\to 0$ as $n\to\infty$, it follows that $\exp[\mathrm{ad}(iS_n)]$ becomes a diagonal matrix with entries (1, 0, 0, 0, 0, 1, 1, 1) in this limit. The product $\Pi_{n=0}^k\exp[\mathrm{ad}(iS_n)]$ therefore converges as $k\to\infty$, so the limiting FW sequence giving the quantum particle Hamiltonian indeed exists. It can be shown that

$$\prod_{j=0}^{\infty}\exp[\mathrm{ad}(iS_j)]\leftrightarrow\begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & \frac{1}{2\gamma} & [t_0^2] & 0 & 0 & [1-\frac{1}{2}t_0^2] & 0\\ 0 & 0 & 0 & 0 & \left[-1+\frac{1}{2}t_0^2\right] & -\frac{1}{2\gamma(\gamma+1)} & 0 & \left[1-\frac{1}{2}t_0^2\right] \end{pmatrix}, \tag{22}$$

where $\gamma := \sqrt{1+t_0^2}$ is the usual relativistic gamma factor. Since the limiting FW Hamiltonian is obtained by applying this matrix to $H_0\leftrightarrow(1,1,1,0,0,1,0,0)$, the entries in brackets are irrelevant and so we have evaluated them only to order t_0^2. Thus, we find the limiting FW Hamiltonian to be

$$\tilde{H}=\gamma m_0\gamma^0+m_0\mathcal{E}_0+\frac{m_0}{2\gamma}\gamma^0\{\mathcal{P}_0,\mathcal{O}_0\}-\frac{m_0}{2\gamma(\gamma+1)}\mathcal{P}_0[\mathcal{P}_0,\mathcal{E}_0], \tag{23}$$

an expression that is accurate to all orders in t_0. Substitution from Eq. (7) yields the explicit form

$$\begin{aligned}\tilde{H}=\gamma m\gamma^0&+\left\{a_0-me_0-m(c_{0j}+c_{j0})\frac{p^j}{m}\right\}+\left\{-\frac{mc_{00}}{\gamma}+(a_j-me_j)\frac{p^j}{\gamma m}-m(c_{jk}-\eta_{jk}c_{00})\frac{p^jp^k}{\gamma m^2}\right\}\gamma^0\\ &+\Bigg\{-(md_{j0}+\tfrac{1}{2}\epsilon^{kl}{}_jH_{kl})+\left[-\frac{b_0\eta_{jk}}{\gamma}+m\epsilon^{lm}{}_j(\tfrac{1}{2}g_{lmk}-\eta_{km}g_{l00})\right]\frac{p^k}{m}\\ &\quad+\left[m(d_{0l}+d_{l0})-\frac{(\gamma-1)m^2}{p^2}(md_{l0}+\tfrac{1}{2}\epsilon^{mn}{}_lH_{mn})\right]\eta_{jk}\frac{p^lp^k}{\gamma m^2}\\ &\quad+\left[\frac{(\gamma-1)m^2}{2p^2}m\epsilon^{nq}{}_lg_{nqk}\right]\eta_{jm}\frac{p^kp^lp^m}{\gamma m^3}\Bigg\}\gamma_5\gamma^j\\ &+\Bigg\{\left[-b_j-\frac{1}{2}m\epsilon^{kl}{}_jg_{kl0}\right]\frac{1}{\gamma}+[\epsilon^l{}_{kj}H_{0l}+m(d_{jk}-\eta_{jk}d_{00})]\frac{p^k}{\gamma m}\\ &\quad+\left[m\epsilon^m{}_{lj}(g_{m0k}+g_{mk0})+\frac{(\gamma-1)m^2}{p^2}\eta_{jl}\left(b_k+\frac{1}{2}m\epsilon^{mn}{}_kg_{mn0}\right)\right]\frac{p^kp^l}{\gamma m^2}\\ &\quad+\left[-\frac{(\gamma-1)m^2}{p^2}m(d_{kl}-\eta_{kl}d_{00})\right]\eta_{jm}\frac{p^kp^lp^m}{\gamma m^3}\Bigg\}\gamma_5\gamma^0\gamma^j.\end{aligned} \tag{24}$$

This equation gives the FW form of the relativistic quantum Hamiltonian for a four-component fermion.

Certain limiting forms of Eq. (24) are directly relevant to experiment. For applications involving relativistic two-component particles, such as the analysis of muon storage-ring experiments,[14] it suffices to retain only the upper left block $h_{\rm rel}$ of $\tilde{H}$:

$$
\begin{aligned}
h_{\mathrm{rel}}=\gamma m&+\left(a_0-\frac{mc_{00}}{\gamma}-me_0\right)+[a_j-\gamma m(c_{0j}+c_{j0})-me_j]\frac{p^j}{\gamma m}\\
&-m(c_{jk}-\eta_{jk}c_{00})\frac{p^jp^k}{\gamma m^2}+\left\{\left[-\frac{1}{\gamma}b_j+md_{j0}+\frac{1}{2}\epsilon^{kl}{}_jH_{kl}-\frac{1}{2\gamma}m\epsilon^{kl}{}_jg_{kl0}\right]\right.\\
&+\left[\eta_{jk}b_0+m(d_{jk}-\eta_{jk}d_{00})+\epsilon^l{}_{kj}H_{0l}-\gamma m\epsilon^{lm}{}_j\left(\frac{1}{2}g_{lmk}-\eta_{km}g_{l00}\right)\right]\frac{p^k}{\gamma m}\\
&+\left[\frac{(\gamma-1)m^2}{p^2}\left(b_k+md_{k0}+\frac{1}{2}\epsilon^{mn}{}_kH_{mn}+\frac{1}{2}m\epsilon^{mn}{}_kg_{mn0}\right)\eta_{jl}\right.\\
&\qquad\left.-m(d_{0k}+d_{k0})\eta_{jl}+m\epsilon^m{}_{lj}(g_{m0k}+g_{mk0})\right]\frac{p^kp^l}{\gamma m^2}\\
&\qquad\left.+\frac{(\gamma-1)m^2}{p^2}\left[-m(d_{kl}-\eta_{kl}d_{00})-\frac{1}{2}m\epsilon^{nq}{}_lg_{nqk}\right]\eta_{jm}\frac{p^kp^lp^m}{\gamma m^3}\right\}\sigma^j.
\end{aligned}
\tag{25}
$$

This is the quantum particle Hamiltonian associated with the original Lorentz-violating theory.

For many low-energy applications, including analyses of high-precision atomic experiments,[5–8,15] only nonrelativistic and subleading relativistic terms in the quantum particle Hamiltonian are needed. To third order in $|\vec{p}|/m$, the nonrelativistic quantum Hamiltonian h for the two-component fermion is

$$
\begin{aligned}
h=m&+\frac{p^2}{2m}+(a_0-mc_{00}-me_0)+\left(-b_j+md_{j0}-\frac{1}{2}m\epsilon_{jkl}g_{kl0}+\frac{1}{2}\epsilon_{jkl}H_{kl}\right)\sigma^j\\
&+[-a_j+m(c_{0j}+c_{j0})+me_j]\frac{p_j}{m}\\
&+\left[b_0\delta_{jk}-m(d_{kj}+d_{00}\delta_{jk})-m\epsilon_{klm}\left(\frac{1}{2}g_{mlj}+g_{m00}\delta_{jl}\right)-\epsilon_{jkl}H_{l0}\right]\frac{p_j}{m}\sigma^k\\
&+\left[m\left(-c_{jk}-\frac{1}{2}c_{00}\delta_{jk}\right)\right]\frac{p_jp_k}{m^2}\\
&+\left\{\left[m(d_{0j}+d_{j0})-\frac{1}{2}\left(b_j+md_{j0}+\frac{1}{2}m\epsilon_{jmn}g_{mn0}+\frac{1}{2}\epsilon_{jmn}H_{mn}\right)\right]\delta_{kl}\right.\\
&\qquad\left.+\frac{1}{2}\left(b_l+\frac{1}{2}m\epsilon_{lmn}g_{mn0}\right)\delta_{jk}-m\epsilon_{jlm}(g_{m0k}+g_{mk0})\right\}\frac{p_jp_k}{m^2}\sigma^l+\frac{1}{2}(a_j\delta_{kl}-me_j\delta_{kl})\frac{p_jp_kp_l}{m^3}\\
&\qquad+\frac{1}{2}\left[(-b_0\delta_{jm}+md_{mj}+\epsilon_{jmn}H_{n0})\delta_{kl}+\left(-md_{jk}-\frac{1}{2}m\epsilon_{knp}g_{npj}\right)\delta_{lm}\right]\frac{p_jp_kp_l\sigma^m}{m^3}.
\end{aligned}
\tag{26}
$$

Note that the form of Eq. (23) includes all even elements of the basis set $\mathcal{B}_0$. This means that all possible combinations of the parameters for Lorentz violation are already contained in Eq. (26). Higher-order corrections to the nonrelativistic Hamiltonian involve only products of these combinations with powers of $|\vec{p}|^2/m^2$. One interesting implication of this result is that nonrelativistic experiments with single free fermions (or fermions in weak external fields) can at most be sensitive to the particular linear combinations of parameters for Lorentz violation appearing in Eq. (26). Disentangling individual parameters requires a different class of experiment.

As a final remark, note that our methods can also be used to obtain the nonrelativistic quantum Hamiltonian $\overline{h}$ for the antifermion. The result for $\overline{h}$ can be expressed in the same form as Eq. (26), with the substitutions $a_\mu \to \overline{a}_\mu = -a_\mu$, $b_\mu \to \overline{b}_\mu = +b_\mu$, $c_{\mu\nu} \to \overline{c}_{\mu\nu} = +c_{\mu\nu}$, $d_{\mu\nu} \to \overline{d}_{\mu\nu} = -d_{\mu\nu}$, $e_\mu \to \overline{e}_\mu = -e_\mu$, $f_\mu \to \overline{f}_\mu = -f_\mu$, $g_{\lambda\mu\nu} \to \overline{g}_{\lambda\mu\nu} = +g_{\lambda\mu\nu}$, $H_{\mu\nu} \to \overline{H}_{\mu\nu} = -H_{\mu\nu}$. This result is useful for experiments testing Lorentz symmetry with antimatter.

ACKNOWLEDGMENTS

We thank R. Bluhm for discussions. This work was supported in part by the United States Department of Energy under Grant No. DE-FG02-91ER40661.

[1] L. L. Foldy and S. A. Wouthuysen, Phys. Rev. **78**, 29 (1950).
[2] D. Colladay and V. A. Kostelecký, Phys. Rev. D **55**, 6760 (1997); **58**, 116002 (1998).
[3] V. A. Kostelecký and S. Samuel, Phys. Rev. Lett. **63**, 224 (1989); **66**, 1811 (1991); Phys. Rev. D **39**, 683 (1989); **40**, 1886 (1989); V. A. Kostelecký and R. Potting, Nucl. Phys. B **359**, 545 (1991); Phys. Lett. B **381**, 89 (1996).
[4] Some recent developments can be found in *CPT and Lorentz Symmetry*, edited by V. A. Kostelecký (World Scientific, Singapore, 1999).
[5] R. Bluhm, V. A. Kostelecký, and N. Russell, Phys. Rev. Lett. **79**, 1432 (1997); Phys. Rev. D **57**, 3932 (1998); Phys. Rev. Lett. **82**, 2254 (1999).
[6] G. Gabrielse *et al.*, Phys. Rev. Lett. **82**, 3198 (1999).
[7] H. Dehmelt *et al.*, to appear in Phys. Rev. Lett.
[8] R. Mittleman, I. Ioannou, and H. Dehmelt, in Ref. 4; R. Mittleman *et al.*, Phys. Rev. Lett. **83**, 2116 (1999).
[9] V. A. Kostelecký and R. Potting, in *Gamma Ray-Neutrino Cosmology and Planck Scale Physics*, edited by D. B. Cline (World Scientific, Singapore, 1993) (hep-th/9211116); Phys. Rev. D **51**, 3923 (1995); D. Colladay and V. A. Kostelecký, Phys. Lett. B **344**, 259 (1995); Phys. Rev. D **52**, 6224 (1995); V. A. Kostelecký and R. Van Kooten, *ibid.* **54**, 5585 (1996); V. A. Kostelecký, Phys. Rev. Lett. **80**, 1818 (1998); R. Jackiw and V. A. Kostelecký, *ibid.* **82**, 3572 (1999); V. A. Kostelecký, to appear in Phys. Rev. D.
[10] KTeV Collaboration, presented by Y. B. Hsiung at the KAON 99 Conference, Chicago, 1999.
[11] OPAL Collaboration, R. Ackerstaff *et al.*, Z. Phys. C **76**, 401 (1997); DELPHI Collaboration, M. Feindt *et al.* (unpublished).
[12] O. Bertolami *et al.*, Phys. Lett. B **395**, 178 (1997).
[13] S. Coleman and S. Glashow, Phys. Rev. D **59**, 116008 (1999).
[14] R. Bluhm, V. A. Kostelecký, and C. D. Lane (unpublished).
[15] V. A. Kostelecký and C. D. Lane, to appear in Phys. Rev. D.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Casimir energy of a ball and cylinder in the zeta function technique

G. Lambiase[a)]
Dipartimento di Scienze Fisiche ''E.R. Caianiello,'' Universitá di Salerno, 84081, Baronissi (SA), Italy and Istituto Nazionale di Fisica Nucleare, Sez. Napoli, Italy

V. V. Nesterenko[b)]
Bogoliubov Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, Dubna, 141980, Russia

M. Bordag[c)]
Universität Leipzig, Institute für Theoretical Physik, Augustusplatz 10, 04109 Leipzig, Germany

(Received 22 December 1998; accepted for publication 10 March 1999)

A simple method is proposed to construct the spectral zeta functions required for calculating the electromagnetic vacuum energy with boundary conditions given on a sphere or on an infinite cylinder. When calculating the Casimir energy in this approach no exact divergencies appear and no renormalization is needed. The starting point of the consideration is the representation of the zeta functions in terms of contour integral, further the uniform asymptotic expansion of the Bessel function is essentially used. After the analytic continuation, needed for calculating the Casimir energy, the zeta functions are presented as infinite series containing the Riemann zeta function with rapidly falling down terms. The spectral zeta functions are constructed exactly for a material ball and infinite cylinder placed in a uniform endless medium under the condition that the velocity of light does not change when crossing the interface. As a special case, perfectly conducting spherical and cylindrical shells are also considered in the same line. In this approach one succeeds, specifically, in justifying, in mathematically rigorous way, the appearance of the contribution to the Casimir energy for cylinder which is proportional to $\ln(2\pi)$. © *1999 American Institute of Physics.* [S0022-2488(99)04507-7]

I. INTRODUCTION

A considerable achievement in theoretical investigations of the Casimir effect[1,2] was its calculation for massive fields (scalar and spinor) with boundary conditions on a sphere.[3,4] The various divergent contributions had been discussed in detail from the point of view of the general theory of adiabatic expansions (resp. heat kernel expansion). In a subsequent paper[5] it was clarified in which cases the calculation of the Casimir energy, after the proper renormalization, yields a meaningful (unique) result and in which not independently of the regularization used. For a massive field a well defined result can be obtained in any case using the normalization condition proposed there. Instead, for a massless field the heat kernel coefficient a_2 must vanish in order to allow for a meaningful calculation of the Casimir energy. For instance, this is the case for a material body characterized by a polarizability and a permittivity when the speeds of light inside and outside are the same or their difference is small. The vanishing of a_2 for the Dirichlet and Robin boundary conditions (and as a consequence for the conductor and bag boundary conditions) when taking inside and outside contributions together made Boyer's[6] and all subsequent calcula-

[a)]Electronic mail: lambiase@physics.unisa.it
[b)]Electronic mail: nestr@thsun1.jinr.ru
[c)]Electronic mail: Michael.Bordag@itp.uni-leipzig.de

tions possible and meaningful. When using a clever regularization (like the zeta functional one[7,8]) it is even possible to avoid the appearance of divergencies other than that in the Minkowski space contribution at all.

Practically every problem on calculation of the Casimir energy (or force) has been considered multiply with employment of more and more effective and elaborated mathematical methods. For example, the first calculation of the Casimir energy of a perfectly conducting spherical shell[6] carried out by Boyer in 1968 has required computer calculations during 3 years.[9] Later this problem was considered in many papers.[10–12] By making use of the modern methods[13] it can be solved practically without numerical calculations (with a precision of a few percent). It requires only the application of the uniform asymptotic expansion for the Bessel functions.

In recent papers[14,15] the Casimir energy of a compact ball[16] and infinite cylinder has been calculated by making use of the mode-by-mode summation technique. In these problems two sums appear, over the roots of radial frequency equation at fixed value of angular momentum and then over angular momentum. The either of these sums is divergent. In papers[14,15] for each of these summation a separate regularization has been used. The first summation was carried out by applying the contour integration with subsequent subtraction of the contribution of an infinite homogeneous space. The second sum was evaluated by making use of the Riemann ζ function technique. However the procedure of analytic continuation, required by rigorous approach, has not been considered there.

The present paper pursues the aim to eliminate the minor points of preceding considerations, i.e., the Casimir energy for two configurations mentioned above will be calculated by the rigorous ζ function techniques, and the analytic continuation of the relevant spectral ζ functions will be carried out exactly. An essential advantage of this regularization procedure is that no manifestly divergent expressions arise in its framework, and it gives a final finite result without any subtractions (renormalizations).

The layout of this paper is as follows. In Sec. II, the spectral zeta function is constructed for a compact ball placed in uniform endless medium when the light velocity is the same inside and outside the ball. As a special case the zeta function for perfectly conducting spherical shell is also considered. In Sec. III the spectral functions for infinite cylinder are constructed under the same conditions. These results provide a firm footing for the previous calculations of the Casimir energy for given boundary conditions by making use of a ''naive'' zeta function method. In Sec. IV the obtained results are shortly discussed.

II. CASIMIR ENERGY OF A COMPACT BALL UNDER THE CONDITION $\epsilon\mu=1$

In the ζ function method[7,8] the Casimir energy E_C is defined in the following way. Let ω_p's be the eigenfrequencies of the quantum field system under the influence of the boundary conditions, and let $\overline{\omega_p}$'s be the same frequencies when the boundaries are removed. By making use of this spectrum one defines the ζ function for the problem in hand,

$$\zeta(s)=\sum_{\{p\}}(\omega_p^{-s}-\overline{\omega_p}^{-s}). \tag{2.1}$$

Here the summation (or integration) should be done over all the quantum numbers $\{p\}$ specifying the spectrum. The parameter s is considered at first to belong to region of the complex plane s where the sum (2.1) exists. Further the analytic continuation of (2.1) to the point $s=-1$ should be constructed. After that one puts

$$E_C=\tfrac{1}{2}\zeta(s=-1). \tag{2.2}$$

Let us consider a solid ball of radius a, consisting of a material which is characterized by permittivity ϵ_1 and permeability μ_1. The ball is assumed to be placed in an infinite medium with permittivity ϵ_2 and permeability μ_2. The eigenfrequencies of electromagnetic field for this configuration are determined by the frequency equation for the TE modes,[17]

$$\Delta_l^{\mathrm{TE}}(a\omega)\equiv\sqrt{\epsilon_1\mu_2}\,\tilde{s}_l'(k_1a)\tilde{e}_l(k_2a)-\sqrt{\epsilon_2\mu_1}\,\tilde{s}_l(k_1a)\tilde{e}_l'(k_2a)=0, \tag{2.3}$$

and the analogous equation for the TM modes

$$\Delta_l^{\mathrm{TM}}(a\omega)\equiv\sqrt{\epsilon_2\mu_1}\,\tilde{s}_l'(k_1a)\tilde{e}_l(k_2a)-\sqrt{\epsilon_1\mu_2}\,\tilde{s}_l(k_1a)\tilde{e}_l'(k_2a)=0, \tag{2.4}$$

where $k_i=\sqrt{\epsilon_i\mu_i}\omega,\ i=1,2$ are the wave numbers inside and outside the ball, respectively. Here $\tilde{s}_l(x)$ and $\tilde{e}_l(x)$ are the Riccati–Bessel functions,

$$\tilde{s}_l(x)=\sqrt{\frac{\pi}{2x}}J_{l+1/2}(x), \quad \tilde{e}_l(x)=\sqrt{\frac{\pi}{2x}}H^{(1)}_{l+1/2}(x), \tag{2.5}$$

and prime stands for the differentiation with respect to their arguments, k_1a or k_2a. The orbital momentum l in Eqs. (2.3) and (2.4) assumes the values 1,2,... .

As usual when one is dealing with an analytic continuation, it is convenient to represent the sum (2.1) in terms of the contour integral

$$\zeta_C(s)=\sum_{l=1}^{\infty}\frac{2l+1}{2\pi i}\lim_{\mu\to 0}\oint_C dz(z^2+\mu^2)^{-s/2}\frac{d}{dz}\ln\frac{\Delta_l^{\mathrm{TE}}(az)\Delta_l^{\mathrm{TM}}(az)}{\Delta_l^{\mathrm{TE}}(\infty)\Delta_l^{\mathrm{TM}}(\infty)}, \tag{2.6}$$

where the contour C surrounds, counterclockwise, the roots of the frequency equations in the right half-plane. For brevity we write in (2.6) simply $\Delta_l(\infty)$ instead of $\lim_{a\to\infty}\Delta_l(az)$. Transition to the complex frequencies z in Eq. (2.6) is accomplished by introducing the unphysical photon mass μ,

$$\omega\to(z^2+\mu^2)^{s/2}|_{\mu\to 0}. \tag{2.7}$$

Extension to the complex z-plane of the frequency equations $\Delta_l^{\mathrm{TE}}(az)$ and $\Delta_l^{\mathrm{TM}}(az)$ should be done in the following way. In the upper (lower) half-plane the Hankel functions of the first (second) kind $H_\nu^{(2)}(az)(H_\nu^{(1)}(az))$ must be used.[18] Location of the roots of Eqs. (2.3) and (2.4) enables one to deform the contour C into a segment of the imaginary axis $(-i\Lambda,i\Lambda)$ and a semicircle of radius Λ in the right half-plane. When Λ tends to infinity the contribution along the semicircle into $\zeta_{\mathrm{ball}}(s)$ vanishes because the argument of the logarithmic function in the integrand tends in this case to 1. As a result we obtain

$$\zeta_{\mathrm{ball}}(s)=-a^s\sum_{l=1}^{\infty}\frac{(2l+1)}{2\pi i}\lim_{\mu\to 0}\int_{-i\infty}^{+i\infty}dz(z^2+\mu^2)^{-s/2}\frac{d}{dz}\ln\frac{\Delta_l^{\mathrm{TE}}(z)\Delta_l^{\mathrm{TM}}(z)}{\Delta_l^{\mathrm{TE}}(\infty)\Delta_l^{\mathrm{TM}}(\infty)}. \tag{2.8}$$

Now we impose the condition that the velocity of light inside and outside the ball is the same

$$\epsilon_1\mu_1=\epsilon_2\mu_2=c^{-2}. \tag{2.9}$$

Under this assumption the argument of the logarithm in (2.8) can be simplified considerably[14] with the result

$$\zeta_{\mathrm{ball}}(s)=\left(\frac{c}{a}\right)^{-s}\sum_{l=1}^{\infty}(2l+1)\frac{\sin(\pi s/2)}{\pi}\int_0^{\infty}dy\,y^{-s}\frac{d}{dy}\ln[1-\xi^2\sigma_l^2(y)], \tag{2.10}$$

where

$$\xi=\frac{\epsilon_1-\epsilon_2}{\epsilon_1+\epsilon_2}, \quad \sigma_l(y)=\frac{d}{dy}[s_l(y)e_l(y)]. \tag{2.11}$$

Here $s_l(y)$ and $e_l(y)$ are the modified Riccati–Bessel functions,

$$s_l(x)=\sqrt{\frac{\pi x}{2}}I_\nu(x),\quad e_l(x)=\sqrt{\frac{2x}{\pi}}K_\nu(x),\quad \nu=l+1/2. \tag{2.12}$$

More details concerning the contour integral representation of the spectral ζ functions can be found in Refs. 3, 19–21.

Further the analytic continuation of Eq. (2.10) is accomplished by expressing the sum over l in terms of the Riemann ζ function. This cannot be done in a closed form. Making use of the uniform asymptotic expansion (UAE) for the Bessel functions in increase powers of $1/\nu$ enables one to construct the analytic continuation looked for in the form of the series, the terms of which are expressed through the Riemann ζ function. The problem of the convergence of this series does not arise because its terms go down very fast.

We demonstrate this keeping only two terms in the UAE for the product of the Bessel functions $I_\nu(\nu z)K_\nu(\nu z)$,[22]

$$I_\nu(\nu z)K_\nu(\nu z)\simeq\frac{t}{2\nu}\left[1+\frac{t^2(1-6t^2+t^4)}{8\nu^2}+\cdots\right],\quad t=\frac{1}{\sqrt{1+z^2}}. \tag{2.13}$$

After changing the integration variable $y=\nu z$ in Eq. (2.10) we substitute (2.13) into this formula and expand the logarithm function up to the order ν^{-3} keeping at the same time only the terms linear in ξ^2. The last assumption is not principal. It is introduced for simplicity and in order to have possibility of a direct comparison with the results of Ref. 14. Thus we have

$$\frac{d}{dz}\ln\left\{1-\xi^2\left[\frac{d}{dz}(zI_\nu(\nu z)K_\nu(\nu z))\right]^2\right\}$$
$$=\frac{3}{2}\frac{\xi^2}{\nu^2}zt^8+\frac{\xi^2}{16\nu^4}zt^8(-12+216t^2-600t^4+420t^6)+O(\nu^{-6}). \tag{2.14}$$

Integration over z can be done by making use of the formula

$$\int_0^\infty z^{-\alpha-1}t^\beta dz=\frac{1}{2}\frac{\Gamma\left(\frac{\alpha+\beta}{2}\right)\Gamma\left(-\frac{\alpha}{2}\right)}{\Gamma\left(\frac{\beta}{2}\right)}. \tag{2.15}$$

Also the properties of the Γ function,

$$\Gamma(z)\Gamma(1-z)=\frac{\pi}{\sin\pi z},\quad \Gamma(1+z)=z\Gamma(z) \tag{2.16}$$

prove to be useful. After simple calculations we arrive at the result

$$\zeta_{\text{ball}}(s)\simeq\frac{\xi^2}{32}\left(\frac{c}{a}\right)^{-s}s(2+s)(4+s)\left(\sum_{l=1}^\infty \nu^{-1-s}+p(s)\sum_{l=1}^\infty \nu^{-3-s}+\cdots\right),\quad \nu=l+1/2, \tag{2.17}$$

where

$$p(s)=-\frac{1}{2}\left[1-\frac{9}{4}(6+s)+\frac{5}{8}(6+s)(8+s)-\frac{7}{192}(6+s)(8+s)(10+s)\right]. \tag{2.18}$$

The zeta function $\zeta_{\text{ball}}(s)$ represented in the form (2.17) is defined for $\text{Re}\,s>0$ due to the first sum over l. This term corresponds to the order $1/\nu$ in the uniform asymptotic expansion (2.13). The second sum in (2.17), defined at $\text{Re}\,s>-2$, has been generated by the term $\sim 1/\nu^3$ in Eq. (2.13).

It is clear that the terms of order $1/\nu^{2k+1}$ in (2.13) will give rise to the singularity of $\zeta_{\mathrm{ball}}(s)$ at the points $s=-2k,\ k=0,1,2,\ldots$. Due to the multipliers in front of the square brackets in (2.17) the first three singularities are really the indefinitenesses like $0\cdot\infty$.

The analytic continuation of Eqs. (2.17), (2.18) into the region $\mathrm{Re}\, s\leqslant 0$ can be accomplished by expressing the sums over angular momentum l through the Riemann ζ function according to the formula,[23]

$$\sum_{l=1}^{\infty}\nu^{-s}=(2^s-1)\zeta(s)-2^s,\quad \nu=l+1/2. \tag{2.19}$$

As a result one gets

$$\zeta_{\mathrm{ball}}(s)\simeq\frac{\xi^2}{32}\left(\frac{c}{a}\right)^{-s}s(2+s)(4+s)\{(2^{1+s}-1)\zeta(1+s) -2^{1+s}+p(s)[(2^{3+s}-1)\zeta(1+s)-2^{3+s}]+\cdots\}. \tag{2.20}$$

The singularities in Eq. (2.17) are transformed in (2.20) into the poles of the Riemann ζ functions at the points $s=2k,\ k=0,1,2,\ldots$,

$$\begin{aligned}\zeta(1+s)&\simeq\frac{1}{s}+\gamma+\cdots,\quad s\to 0,\\ \zeta(3+s)&\simeq\frac{1}{s+2}+\gamma+\cdots,\quad s\to -2,\\ \zeta(5+s)&\simeq\frac{1}{s+4}+\gamma+\cdots,\quad s\to -4,\end{aligned} \tag{2.21}$$

where γ is the Euler constant. The first three poles are annihilated by the multipliers in front of the curly brackets in Eq. (2.20). The first surviving singularity (simple pole) appears only at the point $s=-6$. Thus the formula (2.20) affords the required analytic continuation of the function $\zeta_{\mathrm{ball}}(s)$ into the region $\mathrm{Re}\, s<0$. In view of Eq. (2.2) we are interested in the point $s=-1$, where $\zeta_{\mathrm{ball}}(s)$ given by (2.20) is regular,

$$\zeta_{\mathrm{ball}}(-1)=\frac{3\xi^2c}{32a}\left[1+\frac{9}{128}\left(\frac{\pi^2}{2}-4\right)+\cdots\right]. \tag{2.22}$$

It is exactly the first two terms in Eq. (3.10) of Ref. 14. The procedure of analytic continuation presented above can be extended in a straightforward way to the arbitrary order of the uniform asymptotic expansion (2.13). Certainly in this case analytical calculations should be done by making use of MATHEMATICA or MAPLE.

The problem under consideration with $\xi=1$ is of a special interest because in this case it gives the Casimir energy of a perfectly conducting spherical shell. As it was noted above, this configuration has been considered by many authors. We present here the basic formulas which afford the analytical continuation of the corresponding spectral ζ function. We again content ourselves with two terms in the UAE (2.13). It is impossible to put simply $\xi=1$ in the next formula (2.14). One has to do the expansion here anew keeping all the terms $\sim 1/\nu^4$. This gives

$$\frac{d}{dz}\ln\left\{1-\left[\frac{d}{dz}(zI_\nu(\nu z)K_\nu(\nu z))\right]^2\right\}=\left[\frac{3}{2\nu^2}zt^8+\frac{3}{4\nu^4}zt^8(-1+18t^2-50t^4+35t^6)+O(\nu^{-6})\right]. \tag{2.23}$$

After integration and elementary simplifications we arrive at the following result for the spectral zeta function in hand,

$$\zeta_{\text{shell}}(s) \simeq \frac{1}{32a^{-s}} s(2+s)(4+s)\left[\sum_{l=1}^{\infty} \nu^{-1-s} + q(s)\sum_{l=1}^{\infty} \nu^{-3-s} + \cdots\right], \tag{2.24}$$

where

$$q(s) = \frac{1}{3840}(480+868s+504s^2+71s^3). \tag{2.25}$$

Obviously formula (2.23) has the same singularities as Eq. (2.17), i.e., it is defined for $\mathrm{Re}\, s>0$. The analytic continuation is accomplished by making use of Eq. (2.19),

$$\zeta_{\text{shell}}(s) \simeq \frac{1}{32a^{-s}} s(2+s)(4+s)\{(2^{1+s}-1)\zeta(1+s) - 2^{1+s} + q(s)[(2^{3+s}-1)\zeta(3+s) - 2^{3+s}] + \cdots\}. \tag{2.26}$$

The nearest singularity in this formula is simple pole at $s=-6$. As above it is originated in the term $\sim 1/\nu^7$ in the UAE (2.13). At the point $s=-1$ the spectral zeta function $\zeta_{\text{shell}}(s)$ is regular and gives the following value for the Casimir energy of a perfectly conducting spherical shell,

$$E_{\text{shell}}(-1) = \frac{1}{2}\zeta_{\text{shell}}(-1) = \frac{3}{64a}\left[1 - \frac{3}{256}\left(\frac{\pi^2}{2}-4\right) + \cdots\right] = \frac{1}{a}0.046361\ldots. \tag{2.27}$$

Without considering the analytic continuation and by not carrying out the analysis of the singularities in the complex s plane this result has been obtained in Ref. 13.

Undoubtedly, the calculation of the Casimir energy of a nonmagnetic dielectric ball ($\epsilon_1\mu_1 \neq \epsilon_2\mu_2$) by a rigorous ζ function method is also of a special interest. However, in this case the very definition of the spectral zeta function should probably be changed in order to incorporate the contact terms which seem to be essential in this problem.[24–26]

III. VACUUM ENERGY OF ELECTROMAGNETIC FIELD WITH BOUNDARY CONDITIONS GIVEN ON AN INFINITE CYLINDER

Calculation of the Casimir energy of an infinite cylinder[27,15] proves to be a more involved problem in comparison with that for a sphere. In this section the spectral zeta function $\zeta_{\text{cyl}}(s)$, for this configuration will be constructed, its analytical continuation into the left half-plane of the complex variable s will be carried out, and relevant singularities will be analyzed.

Thus we are considering an infinite cylinder of radius a which is placed in a uniform unbounded medium. The permittivity and the permeability of the material making up the cylinder are ϵ_1 and μ_1, respectively, and those for the surrounding medium are ϵ_2 and μ_2. We assume again that the condition (2.9) is fulfilled. In this case the electromagnetic oscillations can again be divided into the transverse-electric (TE) modes and transverse-magnetic (TM) modes. In terms of the cylindrical coordinates (r,θ,z) the eigenfunctions of the given boundary value problem contain the multiplier

$$\exp(\pm i\omega t + ik_z z + in\theta) \tag{3.1}$$

and their dependence on r is described by the Bassel functions J_n for $r<a$ and by the Hankel functions $H_n^{(1)}$ or $H_n^{(2)}$ for $r>a$. The frequencies of TE- and TM-modes are determined, respectively, by the equations[17]

$$\Delta_n^{\text{TE}}(\lambda,a) \equiv \lambda a[\mu_1 J_n'(\lambda a)H_n(\lambda a) - \mu_1 J_n(\lambda a)H_n(\lambda a)'] = 0, \tag{3.2}$$

$$\Delta_n^{\mathrm{TM}}(\lambda,a)\equiv\lambda a[\epsilon_1 J_n'(\lambda a)H_n(\lambda a)-\epsilon_1 J_n(\lambda a)H_n(\lambda a)']=0, \tag{3.3}$$

where $n=0,\pm1,\pm2,\ldots$. Here λ is the eigenvalue of the corresponding transverse (membranelike) boundary value problem

$$\lambda^2=\frac{\omega^2}{c^2}-k_z^2. \tag{3.4}$$

In a complete analogy with the preceding section we define the Casimir energy per unit length of the cylinder through the spectral zeta function

$$E_{\mathrm{cyl}}=\tfrac{1}{2}\zeta_{\mathrm{cyl}}(-1). \tag{3.5}$$

Let λ_{nr} be the roots of the frequency Eqs. (3.2) and (3.3), then the function $\zeta_{\mathrm{cyl}}(s)$ is introduced in the following way:

$$\zeta_{\mathrm{cyl}}(s)=c^{-s}\int_{-\infty}^{+\infty}\frac{dk_z}{2\pi}\sum_{n,r}[(\lambda_{nr}(a)+k_z^2)^{-s/2}-(\lambda_{nr}(\infty)+k_z^2)^{-s/2}]. \tag{3.6}$$

In terms of the contour integral it can be represented in the form

$$\zeta_{\mathrm{cyl}}(s)=\frac{c^{-s}}{2\pi i}\int_{-\infty}^{+\infty}\frac{dk_z}{2\pi}\sum_{n=-\infty}^{+\infty}\oint_C(\lambda^2+k_z^2)^{-s/2}d_\lambda\ln\frac{\Delta_n^{\mathrm{TE}}(\lambda a)\Delta_n^{\mathrm{TM}}(\lambda a)}{\Delta_n^{\mathrm{TE}}(\infty)\Delta_n^{\mathrm{TM}}(\infty)}. \tag{3.7}$$

Again we can take the contour C to consist of the imaginary axis $(+i\infty,-i\infty)$ closed by a semicircle of an infinitely large radius in the right half-plane. Continuation of the expressions $\Delta_n^{\mathrm{TE}}(\lambda a)$ and $\Delta_n^{\mathrm{TM}}(\lambda a)$ into the complex plane λ should be done in the same way as in the preceding section, i.e., by using $H_n^{(1)}(\lambda)$ for $\mathrm{Im}\,\lambda<0$ and $H_n^{(2)}(\lambda)$ for $\mathrm{Im}\,\lambda>0$. On the semicircle the argument of the logarithm in Eq. (3.7) tends to 1. As a result this part of the contour C does not give any contribution into the zeta function $\zeta_{\mathrm{cyl}}(s)$. When integrating along the imaginary axis we choose the branch line of the function $\phi(\lambda)=(\lambda^2+k_z^2)^{-s/2}$ to run between $-ik_z$ and $+ik_z$, where $k_z=+\sqrt{k_z^2}>0$ and use further that branch of this function which assumes real values when $|y|<k_z$, where $y=\mathrm{Im}\,\lambda$. More precisely we have

$$\phi(iy)=\begin{cases} e^{-i\pi s/2}(y^2-k_z^2)^{-s/2}, & y>k_z, \\ (k_z^2-y^2)^{-s/2}, & |y|<k_z, \\ e^{i\pi s/2}(y^2-k_z^2)^{-s/2}, & y<-k_z. \end{cases} \tag{3.8}$$

Employment of the Hankel functions $H_n^{(1)}(\lambda)$ and $H_n^{(2)}(\lambda)$ by extending the expressions $\Delta_n^{\mathrm{TE}}(\lambda)$ and $\Delta_n^{\mathrm{TM}}(\lambda)$ into the complex plane λ, as it was noted above, gives rise to the argument of the logarithm function depending only on y^2 on the imaginary axis. It means that the derivative of the logarithm is an odd function of y. As a result the segment of the imaginary axis $(-ik_z,+ik_z)$ gives zero, and Eq. (3.7) acquires the form

$$\zeta_{\mathrm{cyl}}(s)=\frac{c^{-s}}{\pi^2}\sin\frac{\pi s}{2}\sum_{n=-\infty}^{+\infty}\int_0^\infty dk_z\int_{k_z}^\infty(y^2-k_z^2)^{-s/2}d_y\ln\frac{\Delta^{\mathrm{TE}}(iay)_n\Delta_n^{\mathrm{TM}}(iay)}{\Delta_n^{\mathrm{TE}}(i\infty)\Delta_n^{\mathrm{TM}}(i\infty)}. \tag{3.9}$$

Changing the order of integration of k_z and y and taking into account the value of the integral,

$$\int_0^y dk_z(y^2-k_z^2)^{-s/2}=\frac{\sqrt{\pi}}{2}y^{1-s}\frac{\Gamma\left(1-\frac{s}{2}\right)}{\Gamma\left(\frac{3-s}{2}\right)},\quad \mathrm{Re}\,s<2, \tag{3.10}$$

we obtain after the substitution $ay \to y$,

$$\zeta_{\rm cyl}(s)=\frac{1}{2\sqrt{\pi}a\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{3-s}{2}\right)}\left(\frac{c}{a}\right)^{-s}\sum_{n=-\infty}^{+\infty}\int_0^\infty dy y^{1-s}\frac{d}{dy}\ln[1-\xi^2\mu_n^2(y)], \tag{3.11}$$

where

$$\mu_n(y)=y(I_n(y)K_n(y))', \quad \xi=\frac{\epsilon_1-\epsilon_2}{\epsilon_1+\epsilon_2}. \tag{3.12}$$

We shall again content ourselves with the first two terms in the uniform asymptotic expansion (2.13) and take into account only the terms linear in ξ^2. In this approximation, upon changing the integration variable $y=nz$, $n=\pm 1,\pm 2,...$, we have

$$\ln\left\{1-\xi^2\left[z\frac{d}{dz}(I_n(nz)K_n(nz))\right]^2\right\}=-\xi^2\frac{z^4t^6}{4n^2}\left[1+\frac{t^2}{4n^2}(3-30t^2+35t^4)+O(n^{-4})\right]. \tag{3.13}$$

Now we substitute (3.13) into all the terms in (3.11) with $n\neq 0$. The term with $n=0$ in this sum will be treated by subtracting and adding to the logarithmic function the quantity

$$-\frac{\xi^2}{4}\frac{y^4}{(1+y^2)^3}. \tag{3.14}$$

As a result the zeta function $\zeta_{\rm cyl}(s)$ can be presented now as the sum of three terms,

$$\xi_{\rm cyl}(s)=Z_1(s)+Z_2(s)+Z_3(s), \tag{3.15}$$

where

$$Z_1(s)=\frac{\left(\frac{c}{a}\right)^{-s}}{2\sqrt{\pi}a\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{3-s}{2}\right)}\int_0^\infty dy y^{1-s}\frac{d}{dy}\left\{\ln[1-\xi^2\mu_0^2(y)]+\frac{\xi^2}{4}y^4t^6\right\}, \tag{3.16}$$

$$Z_2(s)=-\xi^2\left(\frac{c}{a}\right)^{-s}\frac{2\Sigma_{n=1}^{+\infty}n^{-s-1}+1}{8\sqrt{\pi}a\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{3-s}{2}\right)}\int_0^\infty dz z^{1-s}\frac{d}{dz}(z^4t^6), \tag{3.17}$$

$$Z_3(s)=-\xi^2\frac{2\left(\frac{c}{a}\right)^{-s}\Sigma_{n=1}^{+\infty}n^{-3-s}}{32\sqrt{\pi}a\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{3-s}{2}\right)}\int_0^\infty dz z^{1-s}\frac{d}{dz}[z^4t^8(3-30t^2+35t^4)]. \tag{3.18}$$

In these equations $Z_1(s)$ has accumulated the term with $n=0$ from Eq. (3.11) subtracted by (3.14); $Z_2(s)$ involves the contribution of the term of order $1/n^2$ in expansion (3.13) and the added expression (3.14); $Z_3(s)$ is generated by the terms of order $1/n^4$ in the expansion (3.13).

Taking into account that

$$\mu_0^2(y)|_{y\to 0}\to 1 \quad \text{and} \quad \mu_0^2(y)|_{y\to\infty}\to\frac{1}{4y^2}+\frac{3}{16y^4}, \tag{3.19}$$

the integration by parts in Eq. (3.16) can be done for $-3<\mathrm{Re}\, s<1$ with the result

$$Z_1(s)=\frac{s-1}{2\sqrt{\pi}a\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{3-s}{2}\right)}\left(\frac{c}{a}\right)^{-s}\int_0^\infty dy\, y^{-s}\left\{\ln[1-\xi^2\mu_0^2(y)]+\frac{\xi^2}{4}y^4t^6(y)\right\}. \tag{3.20}$$

With allowance for (3.19) one infers easily that the function $Z_1(s)$ is an analytic function of the complex variable s in the region $-3<\mathrm{Re}\,s<1$. In the linear order of ξ^2 it reduces to

$$Z_1^{\mathrm{lin}}(s)=\xi^2\frac{s-1}{2\sqrt{\pi}a\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{3-s}{2}\right)}\left(\frac{c}{a}\right)^{-s}\int_0^\infty dy\, y^{-s}\left[\frac{y^2}{4(1+y^2)^3}-\mu_0^2(y)\right]. \tag{3.21}$$

This function is also analytic in the region $-3<\mathrm{Re}\,s<1$. Integration in Eq. (3.17) can be accomplished exactly by making use of the formula

$$\int_0^\infty dz\, z^{1-s}\frac{d}{dz}(z^4t^6)=\frac{s-1}{4}\Gamma\left(\frac{1+s}{2}\right)\Gamma\left(\frac{5-s}{2}\right), \quad -1<\mathrm{Re}\,s<5. \tag{3.22}$$

This gives for $Z_2(s)$ in (3.17),

$$Z_2(s)=\xi^2\left(\frac{c}{a}\right)^{-s}\frac{(1-s)(3-s)}{64\sqrt{\pi}a}\left(2\sum_{n=1}^\infty n^{-s-1}+1\right)\frac{\Gamma\left(\frac{1+s}{2}\right)}{\Gamma\left(\frac{s}{2}\right)}. \tag{3.23}$$

In view of the sum over n in (3.23) the function $Z_2(s)$ is defined only for $\mathrm{Re}\,s>0$.

For simplicity we apply in Eq. (3.18) the integration by parts which is correct for $-3<\mathrm{Re}\,s<2$ and leads to the result,

$$Z_3(s)=\xi^2\left(\frac{c}{a}\right)^{-s}\frac{(1-s)(3-s)(7s^2-4s-27)}{6144\sqrt{\pi}a}\frac{\Gamma\left(\frac{3+s}{2}\right)}{\Gamma\left(\frac{s}{2}\right)}\sum_{n=1}^\infty n^{-s-3}. \tag{3.24}$$

Again the sum over n in (3.24) gives the restriction $\mathrm{Re}\,s>-2$ for definition of the function $Z_3(s)$.

Thus the spectral zeta function $\zeta_{\mathrm{cyl}}(s)$ in the linear approximation with respect to ξ^2 and with allowance for the first two terms in the UAE (3.13) is given by

$$\zeta_{\mathrm{cyl}}(s)=Z_1^{\mathrm{lin}}(s)+Z_2(s)+Z_3(s), \tag{3.25}$$

where the Z's are presented in Eqs. (3.21), (3.23), and (3.24), respectively. Summing up all the restrictions on the complex variable s which have been imposed when deriving Eqs. (3.21), (3.23), and (3.24), we infer that $\zeta_{\mathrm{cyl}}(s)$ is defined in the strip $0<\mathrm{Re}\,s<1$. In order to continue these equations into the surroundings of the point $s=-1$, it is sufficient to express the sum in Eq. (3.23) in terms of the Riemann ζ function and consider the right-hand side of Eq. (3.22) as an analytic continuation of its left-hand side to this region,

$$Z_2(s)=\xi^2\left(\frac{c}{a}\right)^{-s}\frac{(1-s)(3-s)}{64\sqrt{\pi}a}[2\zeta(s+1)+1]\frac{\Gamma\left(\frac{1+s}{2}\right)}{\Gamma\left(\frac{s}{2}\right)}. \tag{3.26}$$

It is left now to take the limit $s\to -1$ in Eqs. (3.21), (3.23), and (3.26). A special care should be paid when calculating this limit in (3.26) in view of the poles of the function $\Gamma((1+s)/2)$ at this point. Using the values

$$\zeta(0)=-\tfrac{1}{2},\quad \zeta'(0)=-\tfrac{1}{2}\ln 2\pi,\quad \Gamma'(1)=-\gamma, \tag{3.27}$$

one derives

$$\lim_{s\to -1}[2\zeta(1+s)+1]\Gamma\left(\frac{1+s}{2}\right)=\lim_{s\to -1}[2\zeta(0)+2\zeta'(0)(1+s)+O((1+s)^2)+1]$$
$$\times\left[\frac{2}{1+s}-\gamma+O(1+s)\right]=-2\ln(2\pi). \tag{3.28}$$

With allowance for this we obtain from (3.26)

$$Z_2(-1)=\frac{c\xi^2}{2\pi a^2}\frac{1}{4}\ln(2\pi). \tag{3.29}$$

The appearance of the finite term proportional to $\ln(2\pi)$ is remarkable for the problem under consideration. It is derived here in a consistent way by making use of an analytic continuation of the relevant spectral zeta function. In Ref. 15 it was obtained in a more transparent though not rigorous way.

Gathering together Eqs. (3.21), (3.24) with $s=-1$ and Eq. (3.29) we have

$$\xi_{\text{cyl}}(-1)=\frac{c\xi^2}{2\pi a^2}\left\{\int_0^\infty y d_y\left[\frac{y^4}{4(1+y^2)^3}-\mu_0^2(y)\right]+\frac{1}{48}\sum_{n=1}^{+\infty}\frac{1}{n^2}+\frac{1}{4}\ln(2\pi)\right\}$$
$$=\frac{c\xi^2}{2\pi a^2}(-0.490878+0.34269+0.459469)=\frac{c\xi^2}{2\pi a^2}0.002860. \tag{3.30}$$

This result is not the final answer in the problem in hand. The point is that in view of severe cancellations in (3.30) the contribution of the next term in the UAE (3.13) proves to be essential. Its account gives[15]

$$\zeta_{\text{cyl}}(-1)=0. \tag{3.31}$$

Thus the Casimir energy of a compact cylinder possessing the same speed of light inside and outside proves to be zero. The consideration presented in this section can be extended to the next term of order $\sim 1/n^6$ in the UAE (3.13) in a straightforward way. Therefore we shall not present here these rather cumbersome expressions.[28]

Now we address the consideration of a special case when $\xi=1$. It corresponds to a perfectly conducting cylindrical shell.[15] Instead of the expansion (3.13) we have

$$\ln\left\{1-\left[z\frac{d}{dz}(I_n(nz)K_n(nz))\right]^2\right\}=-\frac{z^4t^6}{4n^2}\left[1+\frac{t^2}{4n^2}\left(3-30t^2+35t^4+\frac{1}{2}z^4t^4\right)+O(n^{-4})\right]. \tag{3.32}$$

Proceeding in the same way as above we obtain for the spectral zeta function concerned

$$\zeta_{\text{cyl}}^{\text{shell}}(s)=Z_1(s)+Z_2(s)+Z_3(s), \tag{3.33}$$

where $Z_1(s)$ is given by Eq. (3.20) with $c=1$, $\xi=1$, $Z_2(s)$ is the same as in Eq. (3.26) with $c=1$, and $Z_3(s)$ now is

$$Z_3(s)=\frac{(1-s)(3-s)(71s^2-52s-235)}{61440\sqrt{\pi}a^{1-s}}\frac{\Gamma\left(\frac{3+s}{2}\right)}{\Gamma\left(\frac{s}{2}\right)}\sum_{n=1}^{+\infty} n^{-3-s}. \tag{3.34}$$

At the point $s=-1$ it has the value

$$\zeta_{\rm cyl}^{\rm shell}(-1)=\frac{1}{2\pi a^2}(-0.6517)+\frac{1}{2\pi a^2}\frac{7}{480}\sum_{n=1}^{+\infty}\frac{1}{n^2}+\frac{1}{8\pi a^2}\ln(2\pi)$$
$$=\frac{1}{2\pi a^2}(-0.6517+0.0240+0.4595)=-\frac{1}{a^2}0.0268. \tag{3.35}$$

This exactly reproduces the contribution of the first two terms in calculations of the Casimir energy for cylindrical shell in Ref. 15. With higher accuracy this energy is given by[27]

$$E_{\rm cyl}^{\rm shell}=-\frac{1}{a^2}0.01356. \tag{3.36}$$

In a recent paper[29] the vacuum energy of a perfectly conducting cylindrical surface has been calculated to much higher accuracy by making use of another version of the zeta function technique. By integrating over dk_z directly in Eq. (3.6) the authors reduced this problem to investigation of the zeta function for circle, which has been considered earlier by introducing the partial wave zeta functions for interior and exterior region separately. In this respect our approach dealing only with one spectral zeta function for given boundary conditions proves to be more simple and straightforward.

IV. CONCLUSION

The method for constructing the spectral zeta functions developed here proceeds from the contour integral representation with a subsequent employment of the uniform asymptotic expansions for the Bessel functions. Upon an analytic continuation the zeta functions prove to be presented as (infinite) series over the Riemann ζ functions with rapidly decreasing terms [see, for example, Eqs. (2.20) and (2.26)].

We did not pursue here the goal of obtaining high accuracy when calculating the Casimir energy. In fact we seek to present the consideration in such a form that no manifest divergencies appear. An obvious advantage of the regularization method in hand does not need any renormalization.

By treating the boundary condition given on an infinite cylinder, we have clearly demonstrated the importance of a consistent analytic continuation of the relevant spectral zeta function, in contrast to identifying simply the sum of the type $\Sigma_{n=1}^{\infty}n^{-s}$ with the Riemann ζ function, in order to involve correctly the contributions to the Casimir energy proportional to $\ln(2\pi)$.

Consideration in this framework of the same configuration of vacuum electromagnetic field but with different velocities of light inside and outside the boundaries probably will demand the modification of the definition of the spectral zeta functions for incorporating in a proper way the contact terms important in this case.[24–26]

ACKNOWLEDGMENTS

This work has been completed during the stay of one of the authors (V.V.N.) at Salerno University. It is a pleasant duty for him to thank Professor G. Scarpetta, Dr. G. Lambiase, and Dr. A. Feoli for kind hospitality. The work was accomplished by partial financial support of the Heisenberg–Landau Program and by the fund MURST ex 60% and ex 40% DPR 382/80.

[1] G. Plunien, B. Müller, and W. Greiner, Phys. Rep. **134**, 664 (1987).
[2] V. M. Mostepanenko and N. N. Trunov, Sov. Phys. Usp. **31**, 965 (1988).
[3] M. Bordag, E. Elizalde, K. Kirsten, and S. Leseduarte, Phys. Rev. D **56**, 4896 (1997).
[4] E. Elizalde, M. Bordag, and K. Kirsten, J. Phys. A **31**, 1743 (1998).
[5] M. Bordag, K. Kirsten, and D. Vassilevich, Phys. Rev. D **59**, 085011 (1999).
[6] T. H. Boyer, Phys. Rev. **174**, 1764 (1968).
[7] E. Elizalde, S. D. Odintsov, A. Romeo, A. A. Bytsenko, and S. Zerbini, *Zeta Regularization Techniques with Applications* (World Scientific, Singapore, 1994).
[8] E. Elizalde, *Ten Physical Applications of Spectral Zeta Functions* (Springer, Berlin, 1995).
[9] K. A. Milton, ''Julian Schwinger and the Casimir effect: The reality of zero-point energy,'' in *The Casimir Effect 50 Years Later*, Proceedings of the Fourth Workshop on Quantum Field Theory under the Influence of External Conditions, Leipzig, 14–18 September 1998, edited by M. Bordag (World Scientific, Singapore, 1999), pp. 20–36.
[10] B. Davies, J. Math. Phys. **13**, 1324 (1972).
[11] R. Balian and B. D. Duplantier, Ann. Phys. (N.Y.) **112**, 165 (1978).
[12] K. A. Milton, L. L. DeRaad, Jr., and J. Schwinger, Ann. Phys. (N.Y.) **115**, 388 (1978).
[13] V. V. Nesterenko and I. G. Pirozhenko, Phys. Rev. D **57**, 1284 (1998).
[14] I. Brevik, V. V. Nesterenko, and I. G. Pirozhenko, J. Phys. A **31**, 8661 (1998).
[15] K. A. Milton, A. V. Nesterenko, and V. V. Nesterenko, Phys. Rev. D **59**, 105009 (1999).
[16] Interest in calculating the Casimir energy of a material ball (or spherical cavity) in an infinite medium is caused, to some extent, by recent attempts to elucidate the mechanism of sonoluminescence and, in particular, by Schwinger's idea to use here the dynamical Casimir effect (Ref. 9).
[17] J. A. Stratton, *Electromagnetic Theory* (McGraw–Hill, New York, 1941).
[18] If we are considering only outgoing waves it is natural to use the solutions of the Maxwell equations which are finite (or at least do not grow infinitely) in the future. Such solutions should be proportional to $\exp(-i\omega t)H_\nu^{(1)}(\omega r)$ when ω lies in the lower half-plane. For the upper half-plane ω, the solutions describing outgoing waves and finite in the future should contain the factor $\exp(i\omega t)H_\nu^{(2)}(\omega r)$.
[19] M. Bordag, J. Phys. A **28**, 755 (1995).
[20] E. Elizalde, S. Leseduarte, and A. Romeo, J. Phys. A **26**, 2409 (1993).
[21] S. Leseduarte and A. Romeo, J. Phys. A **27**, 2483 (1994).
[22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards Appl. Math. Ser. 55 (U.S. GPO, Washington, D.C., 1964) (Dover, New York, reprinted 1972).
[23] I. S. Gradshteijn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 5th ed. (Academic, New York, 1994).
[24] K. A. Milton and Y. J. Ng, Phys. Rev. E **55**, 4207 (1997); **57**, 5504 (1998).
[25] I. Brevik and V. N. Marachevsky, Phys. Rev. D **60**, 085006 (1999).
[26] I. Brevik, V. N. Marachevsky, and K. A. Milton, Phys. Rev. Lett. **82**, 3948 (1999).
[27] L. L. DeRaad, Jr. and K. Milton, Ann. Phys. (N.Y.) **136**, 229 (1981).
[28] More precisely, in Ref. 15 not only the next term in the UAE (3.13) has been taken into account but also the first 5 terms in the sum in (3.18) were taken exactly instead of using their asymptotics n^{-3-s}. This turns out to be essential for reaching the required precision. However this point is not critical for our consideration because $Z_3(s)$ does not need analytic continuation.
[29] P. Gosdzinski and A. Romeo, Phys. Lett. B **441**, 265 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Quasilinearization method and its verification on exactly solvable models in quantum mechanics

V. B. Mandelzweig[a)]
Physics Department, Hebrew University, Jerusalem 91904, Israel

(Received 10 June 1999; accepted for publication 7 September 1999)

The proof of the convergence of the quasilinearization method of Bellman and Kalaba, whose origin lies in the theory of linear programming, is extended to large and infinite domains and to singular functionals in order to enable the application of the method to physical problems. This powerful method approximates solution of nonlinear differential equations by treating the nonlinear terms as a perturbation about the linear ones, and is not based, unlike perturbation theories, on existence of some kind of small parameter. The general properties of the method, particularly its uniform and quadratic convergence, which often also is monotonic, are analyzed and verified on exactly solvable models in quantum mechanics. Namely, application of the method to scattering length calculations in the variable phase method shows that each approximation of the method sums many orders of the perturbation theory and that the method reproduces properly the singular structure of the exact solutions. The method provides final and reasonable answers for infinite values of the coupling constant and is able to handle even super singular potentials for which each term of the perturbation theory is infinite and the perturbation expansion does not exist. [S0022-2488(99)01812-5]

I. INTRODUCTION

Most problems of physics are not solvable exactly and therefore should be tackled with the help of analytical or numerical approximation methods. In quantum mechanics and quantum field theory over the years many such methods were developed, from perturbation theories, Wentzel–Kramers–Brillouin (WKB) approach and Monte Carlo simulations to lattice computations, strong coupling approximation, 1/N expansion, and so on. The purpose of this paper is to apply to quantum mechanical problems an additional very powerful approximation technique called the quasilinearization method (QLM), whose origin lies in the theory of linear programming. The method, whose iterations are carefully constructed to yield rapid quadratic convergence and often monotonicity, was developed around 30 years ago by Bellman and Kalaba to solve a wide variety of nonlinear ordinary and partial differential equations or their systems arising in such different physics, engineering, and biology problems as orbit determination, detection of periodicities, radiative transfer, and cardiology.[1,2] The modern developments and applications of the method to different fields are given in Ref. 3. QLM, however, was never systematically studied or extensively applied in quantum physics though references to it could be found in well-known monographs[4,5] dealing with the variable phase approach to potential scattering as well as in a few scattered research papers.[6–9] This could be explained by the fact that convergence of the method has been proved only under rather restrictive conditions[1,2] which generally are not fulfilled in physical applications, such as, for example, a rather small domain of variables or forces which are finite everywhere in the domain (see the following). A goal of this work is to reformulate the proof of the convergence for more realistic physical conditions of infinite domains and forces which

[a)]Electronic-mail: victor@vms.huji.ac.il

could be singular at certain points of the domains. We also show how to deal with solutions which themselves could be infinite at certain values of variable such as, for example, scattering amplitudes at values corresponding to bound state energies, etc.

Since this is our first paper on the subject, in order to make presentation as simple and short as possible, we limit ourselves to the case of the first-order nonlinear ordinary differential equation in one variable. Physically this covers the quantum mechanics of one particle in a central field since in this case the Schrödinger equation for a wave function could be rewritten as the Ricatti equation for its logarithmic derivative.

Our proof of the convergence of the quasilinearization method for a general nonlinear ordinary or partial nth order differential equation in N-dimensional space could be formulated along the same lines and will be given in a subsequent article.

The paper is arranged as follows. In Sec. II we present the main ideas and conditions of convergence of the quasilinearization approach, formulated by Bellman and Kalaba[1,2] for the case of the first-order nonlinear ordinary differential equation in one variable, and modify their proof in order to meet the physical reality of infinite interval of the variable or the possibility of singular potentials. In order to highlight the power of the method in Sec. III we consider examples of different singular and nonsingular, attractive and repulsive potentials $V(r)$ for which the nonlinear first-order ordinary differential equation

$$\frac{da(r)}{dr} = -V(r)(r+a(r))^2, \quad a(0)=0, \tag{1.1}$$

for an S-wave scattering length $a_0 = a(\infty)$, obtained in variable phase approach,[4,5] can be solved exactly and compare the iterations obtained by the Bellman–Kalaba linearization method with exact solutions and with the usual perturbation theory. Our results, advantages of the method, and its possible future applications are discussed in Sec. IV.

II. THE QUASILINEARIZATION METHOD (QLM)

The aim of QLM is to obtain the solution $v(z)$ of a nonlinear first-order differential equation

$$\frac{dv(z)}{dz} = g(v(z),z) \tag{2.1}$$

with the boundary condition $v(a)=c$ as a limit of a sequence of linear differential equations. This goal is easily understandable in view of the fact that there is no useful technique of presenting the general solution of Eq. (2.1) in terms of a finite set of particular solutions as in a linear case where, as a result of the superposition property, the equation could be solved analytically or numerically in a convenient fashion. In addition, the sequence should be constructed in such a way as to obtain quadratic convergence and, if possible, monotonicity.

The shift of the coordinate $z=x+a$ and of the solution itself $u(x)=v(x+a)-c$ reduces Eq. (1) to the canonical form[10]

$$\frac{du(x)}{dx} = f(u(x),x), \quad u(0)=0, \tag{2.2}$$

where $f(u(x),x) \equiv g(u(x)+c,x+a)$.

The QLM prescription[1,2] determines the $n+1$ iterative approximation $u_{n+1}(x)$ to the solution of Eq. (2.2) as a solution of

$$u'_{n+1}(x) = f(u_n,x) + (u_{n+1}(x)-u_n(x)) f_u(u_n,x), \quad u_{n+1}(0)=0, \tag{2.3}$$

where the function $f_u(u,x)=\partial f(u,x)/\partial u$ is a functional derivative of a functional $f(u(x),x)$. If one defines m as an upper limit of a maximum of absolute values of the functional and its first and second functional derivatives

$$\max(|f(u(x),x)|,|f_u(u,x)|,|\tfrac{1}{2}f_{uu}(u,x)|)\leqslant m<\infty, \tag{2.4}$$

one can prove that the sequence of iterations $u_n(x)$, $n=1,2,\ldots$ converges *uniformly and quadratically* on the interval $[0,b]$ to solution $u(x)$ of Eq. (2.2) for bm sufficiently small. Indeed, introducing the metric $\|g\|$ of the function $g(x)$ as a maximum of the function on the interval $[0,b]$,

$$\|g\|=\max|g(x)|,0\leqslant x\leqslant b, \tag{2.5}$$

and introducing notations $\Delta u_{n+1}(x)=u(x)-u_n(x)$, $\delta u_{n+1}(x)=u_{n+1}(x)-u_n(x)$ one proves[1,2] the following inequalities:

$$\|\Delta u_{n+1}\|\leqslant k\|\Delta u_n\|^2, \tag{2.6}$$

$$\|\delta u_{n+1}\|\leqslant k\|\delta u_n\|^2, \tag{2.7}$$

$$k=\frac{bm}{1-bm}, \tag{2.8}$$

which establish the uniform quadratic convergence of sequence $u_n(x)$ on $[0,b]$ for sufficiently small bm. A simple induction of Eq. (2.7) shows[2] that $\delta u_{n+1}(x)$ for an arbitrary $l<n$ satisfies the inequality

$$\|\delta u_{n+1}\|\leqslant(k\|\delta u_{l+1}\|)^{2^{n-l}}/k, \tag{2.9}$$

or for $l=0$,

$$\|\delta u_{n+1}\|\leqslant(k\|\delta u_1\|)^{2^n}/k. \tag{2.10}$$

The convergence depends therefore upon the quantity $q_1=k\|u_1-u_0\|$, where zero iteration $u_0(x)$ satisfies the condition $u_0(0)=0$ and is chosen from physical and mathematical considerations. In view of Eq. (2.8) the convergence is reached if bm is sufficiently small. However, from Eq. (2.9) it follows that for the convergence it is sufficient that just one of the quantities $q_m=k\|\delta u_{m+1}\|$ will be small enough. Consequently, one can always hope[2] that even if the first convergent coefficient q_1 is large a well chosen initial approximation u_0 results in a smallness of at least one of the convergence coefficients q_m, $m>1$, which enables a rapid convergence of the iteration series for $n>m$.

One can prove in addition[1,2] that the convergence is monotonic from below (above), if functional $f(u(x),x)$ is strictly convex (concave), that is if the second functional derivative $f_{uu}(u,x)$ in interval $[0,b]$ exists and is strictly positive (negative).

The QLM treats the nonlinear terms as a perturbation about the linear ones[1,2] and is not based, unlike perturbation theories, on the existence of some kind of small parameter. In the proof of Bellman and Kalaba, a small parameter, bm, however, does appear sort of through the back door. The requirement of small bm is unfortunately too restrictive in most physical problems where m and b are often large or infinite, since x normally changes in an infinite domain and many potentials are infinite at some points in the domain. For example, in the case of the variable phase equation—Eq. (1.1), since most of the realistic forces, like Yukawa, Coulomb, van der Waals, or hard core potentials, are infinite at origin, a function

$$f(a(x),x)=-V(x)(x+a(x))^2 \tag{2.11}$$

or its first

$$f_a(a(x),x) = -2V(x)(x + a(x)) \tag{2.12}$$

or second

$$f_{aa}(a(x),x) = -2V(x) \tag{2.13}$$

functional derivatives, are infinite at the origin. This means $m = \infty$, which is a zero convergence interval. However it has been well known for a long time[4,5,11] that a first approximation of QLM gives finite and reasonable results even for super singular $1/r^n$, $n \geq 4$ potentials for which all the terms of the usual perturbation theory are strongly divergent. It indicates that the condition bm being small may be too restrictive and should be relaxed.

Our goal now is to modernize the proof of uniform quadratic convergence of QLM so the requirement of smallness of an interval for large m as well as the requirement of m being finite is removed. Let us subtract from both sides of Eq. (2.2) a term $h(w(x),x)u(x)$, where $w(x)$ and $h(w(x),x)$ are some arbitrary function and functional, respectively, which we chose later. We obtain

$$\frac{du(x)}{dx} - h(w(x),x)u(x) = f(u(x),x) - h(w(x),x)u(x), \quad u(0) = 0. \tag{2.14}$$

The integral form of Eq. (2.14) is

$$u(x) = \int_0^x ds(f(u(s),s) - h(w(s),s)u(s)) \exp \int_s^x dt h(w(t),t), \tag{2.15}$$

or, in case of nonzero boundary condition $u(0) = c$,

$$u(x) = c \exp \int_0^x dt h(w(t),t) + \int_0^x ds(f(u(s),s) - h(w(s),s)u(s)) \exp \int_s^x dt h(w(t),t), \tag{2.16}$$

which can be checked easily by a simple differentiation.

We consider three different forms of function $w(x)$ and its functional $h(w(x),x)$:

$$h(w(x),x)) \equiv 0, \tag{2.17}$$

$$h(w(x),x) = f_w(w(x),x), \quad w(x) \equiv 0, \tag{2.18}$$

$$h(w(x),x) = f_w(w(x),x), \quad w(x) \equiv u(x). \tag{2.19}$$

We can now define the iteration scheme by setting the function $u(x)$ on the right equal to its nth approximation $u_n(x)$ and obtaining the $(n+1)$th approximation on the left-hand side. The zero approximation $u_0(x)$ is chosen from some mathematical or physical considerations and satisfies the boundary condition $u_0(0) = 0$. We get three different iteration schemes, corresponding to Eqs. (2.17)–(2.19), respectively:

$$u_{n+1}(x) = \int_0^x ds(f(u_n(s),s), \tag{2.20}$$

$$u_{n+1}(x) = \int_0^x ds(f(u_n(s),s) - f_u(0,s)u_n(s)) \exp \int_s^x dt f_u(0,t), \tag{2.21}$$

and

$$u_{n+1}(x)=\int_0^x ds(f(u_n(s),s)-f_u(u_n(s),s)u_n(s))\exp\int_s^x dt f_u(u_n(t),t). \tag{2.22}$$

In case of nonzero boundary condition $u(0)=c$ the iteration sequence should be slightly modified. For example, in this case, according to Eq. (2.16), Eq. (2.22) has a somewhat different form, namely

$$u_{n+1}(x)=c\exp\int_0^x dt\, f_u(u_n(t),t)+\int_0^x ds(f(u_n(s),s)-f_u(u_n(s),s)u_n(s))\exp\int_s^x dt\, f_u(u_n(t),t). \tag{2.23}$$

Let us concentrate in the beginning on Eq. (2.22), which, being the solution of Eq. (2.3), displays the iteration sequence of the QLM. The subtraction of Eq. (2.3) for n and $n-1$ gives a similar differential equation for the difference $\delta u_{n+1}(x)=u_{n+1}(x)-u_n(x)$:

$$\delta u'_{n+1}(x)=f(u_n(x),x)-f(u_{n-1}(x),x)+\delta u_{n+1}(x)f_u(u_n(x),x)-\delta u_n(x)f_u(u_{n-1}(x),x),$$

$$\delta u_{n+1}(0)=0. \tag{2.24}$$

By use of the mean value theorem[12] one can write

$$f(u_n(x),x)=f(u_{n-1}(x),x)+\delta u_n(x)f_u(u_{n-1}(x),x)+\tfrac{1}{2}f_{uu}(\overline{u_n}(x),x)\,\delta u_n^2(x), \tag{2.25}$$

where $\overline{u_n}(x)$ lies between $u_n(x)$ and $u_{n-1}(x)$. As a result Eq. (2.24) could be written as

$$\delta u'_{n+1}(x)-\delta u_{n+1}(x)f_u(u_n(x),x)=\tfrac{1}{2}f_{uu}(\overline{u_n}(x))\,\delta u_n^2(x), \tag{2.26}$$

whose solution has a form

$$\delta u_{n+1}(x)=\frac{1}{2}\int_0^x ds\, f_{uu}(\overline{u_n}(s),s)\,\delta u_n^2(s)\exp\int_s^x dt\, f_u(u_n(t),t). \tag{2.27}$$

Obviously,

$$|\delta u_{n+1}(x)|\leqslant\frac{1}{2}\int_0^x ds|f_{uu}(\overline{u_n}(s),s)||\delta u_n(s)|^2\exp\int_s^x dt\, f_u(u_n(t),t)$$

$$\leqslant k_n(x)\cdot|\delta u_n(\bar{x})|^2\leqslant k_n(b)\cdot\|\delta u_n\|^2. \tag{2.28}$$

Here $\bar{x}$ is the point on the interval $[0,x]$ where $|\delta u_n(x)|$ is maximal,

$$k_n(x)=\frac{1}{2}\int_0^x ds|f_{uu}(\overline{u_n}(s),s)|\exp\int_s^x dt\, f_u(u_n(t),t), \tag{2.29}$$

and positiveness of the integrand in Eq. (2.29) as well as definition (2.5) are used. Since Eq. (2.28) is correct for any x in the interval $[0,b]$, it is correct also for a value of $x\in[0,b]$ for which $|\delta u_{n+1}(x)|$ reaches its maximal value. This gives

$$\|\delta u_{n+1}\|\leqslant k_n(b)\cdot\|\delta u_n\|^2. \tag{2.30}$$

Let us assume the boundness of the first two functional derivatives of $f(u(x),x)$, that is the existence of bounding functions $F(x)$ and $G(x)$ which for any u and x satisfy

$$f_u(u(x),x))\leqslant F(x), \quad |f_{uu}(u(x),x)|\leqslant G(x). \tag{2.31}$$

In this case $k_n(b) \leqslant k(b)$, where

$$k(b)=\frac{1}{2}\int_0^b ds\, G(s)\exp\int_s^b dt\, F(t), \tag{2.32}$$

and Eq. (2.30) could be written in the form

$$\|\delta u_{n+1}\| \leqslant k(b)\cdot\|\delta u_n\|^2, \tag{2.33}$$

which is identical to Eq. (2.7) but with $k=k(b)$ instead of k given by Eq. (2.8). We can reproduce the results of Bellman and Kalaba[1,2] by following their bounding restriction Eq. (2.4) and setting $F(x)=m$, $G(x)=2m$. In this case the integrals in Eq. (2.32) could be easily calculated and give $k(b)=(1-e^{-mb})/e^{-mb}$, which for small mb reduces to the expression for k given by Eq. (2.8). However, as we will see in different examples in Sec. III, $k(b)$ given by Eq. (2.32), unlike k given by Eq. (2.8), could be sufficiently small also for an infinite interval length b and for singular functions $G(x)$ and $F(x)$. This means that the quantity $q_1(b)$,

$$q_1(b)=k(b)\|u_1-u_0\|, \tag{2.34}$$

which is responsible for the convergence [see the discussion after Eq. (2.10)] could be less than unity and thus assure the convergence even in this case. As was pointed out there, the rapid convergence is actually enough that an initial guess for zero iteration is sufficiently good to ensure the smallness of just one of the convergence coefficients $q_m(b)=k(b)\|u_{m+1}-u_m\|$.

With the uniform quadratic convergence of the sequence $u_n(x)$ for the intervals $[0,b]$ in which at least one of the convergence coefficients $q_m(b)<1$ now proven, one can conclude from Eq. (2.27), that in addition for strictly convex (concave) functionals $f(u(x),x)$ the difference $u_{n+1}(x)-u_n(x)$ is strictly positive (negative), which establishes the monotonicity of the convergence from below (above), respectively, on this interval.

If $F(x)$ is a sign-definite function and $G(x)=|F(x)|$, the integral in Eq. (2.32) could be taken explicitly and produces a simple expression for $k(b)$,

$$k(b)=\frac{1}{2}\left|\exp\int_0^b dt\, F(t)-1\right|. \tag{2.35}$$

The subtraction of Eq. (2.3) from Eq. (2.2) gives

$$\Delta u'_{n+1}(x)=f(u,x)-f(u_n(x),x)+\Delta u_{n+1}(x) f_u(u_n(x),x)-\Delta u_n(x) f_u(u_n(x),x),$$

$$\Delta u_{n+1}(0)=0, \tag{2.36}$$

which is similar to Eq. (2.24)—the starting point for our derivation of Eq. (2.33). The derivation along the same lines, starting from Eq. (2.36), gives the analog of Eq. (2.6) with k changed to $k(b)$:

$$\|\Delta u_{n+1}\| \leqslant k(b)\cdot\|\Delta u_n\|^2. \tag{2.37}$$

Equation (2.31) again confirms the uniform quadratic convergence of the sequence u_n to a solution $u(x)$ of Eq. (2.2). One can show in exactly the same fashion as before that for strictly convex (concave) functionals $f(u(x),x)$ difference Δu_{n+1} is strictly positive (negative), proving in this case the monotonic convergence to a limiting function u from below (above), respectively.

In case the solution $u(x)$ and, respectively, its iterations $u_n(x)$ are going to infinity at some points on interval $[0,b]$, Eq. (2.22) could become meaningless. To deal with it, it is necessary to regularize Eq. (2.2), that is reformulate it in terms of a new function $v(x)$ which is finite, as, for example, to change to function $v(x)=1/u(x)$ for $|u(x)|>1$, the prescription which is used in the present work, or to set $u(x)=\tan v(x)$ as it was suggested in Refs. 13 and 14. The corresponding nonlinear equations for $v(x)$ have the form

$$\frac{dv(x)}{dx}=-v(x)^2 f\left(\frac{1}{v(x)},x\right),\quad v(0)=u(c),\quad |u(c)|=1, \tag{2.38}$$

and

$$\frac{dv(x)}{dx}=\cos^2 v(x) f(\tan v(x),x),\quad v(0)=0, \tag{2.39}$$

respectively.

Let us now turn our attention to the iteration sequences given by Eqs. (2.20) and (2.21). These successive approximation schemes were considered by Picard[15] and Calogero, Babikov, and Fluegge (CBF),[4,5,11] respectively. The quadratic convergence, reached in QLM, is based on a specific choice of function $w(x)$ and its functional $h(w(x),x)$ given by Eq. (2.19) which, in view of the mean value theorem of Eq. (2.25), assures cancellation of the first power of $\delta u_n(x)$ and $\Delta u_n(x)$ in recurrence relations of Eqs. (2.24) and (2.36), respectively. Such cancellation will not happen for the Picard and CBF choices of $w(x)$ and $h(w(x),x)$, given by Eqs. (2.17) and (2.18). One obtains, therefore, for these approximation schemes the usual inequality characteristic of the first-order convergence

$$\|\delta u_{n+1}\|<p\|\delta u_n\|, \tag{2.40}$$

where p is a correspondent convergence coefficient. This leads, instead of the very rapid 2^n-power type of convergence, displayed in Eqs. (2.33) and (2.37), to the much slower geometric convergence

$$\|\delta u_{n+1}\|<p^n\|\delta u_1\|. \tag{2.41}$$

III. QLM SCATTERING LENGTH CALCULATIONS AND THEIR COMPARISON WITH THE PERTURBATION THEORY AND EXACT SOLUTIONS

In Sec. II we proved that the QLM successive approximations to the exact solution $u(x)$ of Eq. (2.22) given by Eq. (2.2) converge quadratically and uniformly on interval $[0,b]$, where b is found from the requirement that one of the convergent coefficients $q_m(b)$ defined in a paragraph following Eq. (2.34) is less than unity. In addition for strictly convex (concave) functionals $f(u(x),x)$ the convergence to a limiting function u is monotonic from below (above), respectively.

In order to highlight the power of the method in this section we consider examples of different singular and nonsingular, attractive and repulsive potentials for which the nonlinear first-order ordinary differential equation for an S-wave scattering length, Eq. (1.1), obtained in variable phase approach[4,5] could be solved exactly. We will compare the iterations obtained by the Bellman–Kalaba quasilinearization method (QLM) with exact solutions and with the usual perturbation theory.

A. Square well potential

1. Repulsive square well

Let us start from the repulsive square well potential

$$V(r)=\frac{\lambda}{R^2}\Theta(R-r), \tag{3.1}$$

where $\Theta(R-r)$ is the Heavyside function and λ is a potential strength, which for now is assumed to be positive. The change of variables to the dimensionless variable $x=\sqrt{\lambda}(r/R)$ and dimensionless function $A(x)=\sqrt{\lambda}[a(xR/\sqrt{\lambda})]/R$ allows one to express Eq. (1.1) for $x\leqslant x_0$, $x_0=\sqrt{\lambda}$ in a form

$$\frac{dA(x)}{dx}=-(x+A(x))^2, \quad A(0)=0. \tag{3.2}$$

For $x>x_0$ $A(x)$ is a constant equal to the dimensionless scattering length $A_0=\sqrt{\lambda}(a_0/R)$, the scattering length itself being $a_0\equiv a(R)$. A further change of the function to $u(x)=x+A(x)$ gives a familiar equation for the hyperbolic tangent,

$$\frac{du(x)}{dx}=1-u^2(x), \quad u(0)=0. \tag{3.3}$$

The exact variable scattering length $a(r)$ for the repulsive square well potential is therefore

$$a(r)=\frac{R}{\sqrt{\lambda}}\tanh\left(\sqrt{\lambda}\frac{r}{R}\right)-r, \tag{3.4}$$

while the scattering length is given by

$$a_0=R\left(\frac{\tanh\sqrt{\lambda}}{\sqrt{\lambda}}-1\right)\equiv R\left(\frac{\tanh x_0}{x_0}-1\right). \tag{3.5}$$

Here we use the Calogero definition of the scattering length[4]

$$a_0=\lim_{k\to 0}\frac{\tan\delta(k)}{k}, \tag{3.6}$$

δ is a scattering phase, which is different in sign from the definition used in most publications.

Before considering the QLM, let us turn to the usual perturbation theory. Displaying explicitly the dependence of the potential on the coupling constant $V(r)=\lambda v(r)$ and expanding $a(r)$ in powers of λ, one obtains from Eq. (1.1):

$$\sum_{k=1}^{\infty}\lambda^k a_k'(r)=-\lambda v(r)\left(r+\sum_{n=1}^{\infty}\lambda^n a_n(r)\right)^2. \tag{3.7}$$

Comparisons of coefficients before the powers of λ gives the recurrence relation

$$a_k'(r)=-v(r)\left(r^2\delta_{k1}+2ra_{k-1}(r)+\sum_{n=1}^{k-2}a_{k-n-1}(r)\cdot a_n(r)\right), \quad k=1,2,3\ldots \tag{3.8}$$

The successive integrations of Eq. (3.8) produce the expansion $a(r)$ in the powers of the coupling constant. The first three terms of the perturbation expansion of the variable scattering length, for example, are

$$a_1(r)=-\int_0^r ds\, s^2 v(s),$$

$$a_2(r)=-\int_0^r ds\, 2s v(s) a_1(s), \tag{3.9}$$

$$a_3(r) = -\int_0^r ds\, v(s)(2sa_2(s) + a_1^2(s)),$$

and so on. For $u(x)$ this expansion gives

$$u(x) = x - \frac{1}{3}x^3 + \frac{2}{15}x^5 - \frac{17}{315}x^7 + \frac{62}{2835}x^9 - \frac{1382}{155\,925}x^{11} + \frac{21\,844}{6\,081\,075}x^{13} - \frac{929\,569}{638\,512\,875}x^{15} + \frac{6\,404\,582}{10\,854\,718\,875}x^{17} + O(x^{19}). \tag{3.10}$$

These series, of course, could also be obtained by using the power series expansion of tanh(x). The power expansion of scattering length is given by Eqs. (3.10) and (3.5), the latter can be written in the form $a_0 = R([u(x_0)/x_0] - 1)$.

Let us consider now the approximate QLM solutions of Eq. (3.3), choosing as a zero approximation a solution of this equation for a very small x: $u_0(x) = x$. The recurrence relation (2.22) now has the form

$$u_{n+1}(x) = \int_0^x ds(1 + u_n^2(s))\exp\left(-2\int_s^x dt\, u_n(t)\right), \tag{3.11}$$

while the nth approximation to the scattering length is given by

$$a_{0,n} = R\left(\frac{u_n(x_0)}{x_0} - 1\right). \tag{3.12}$$

The substitution of the zero iteration in Eq. (3.11) leads to a first-order approximation,

$$u_1(x) = -i\frac{\sqrt{\pi}}{4}\mathrm{erf}(ix)e^{-x^2} + \frac{x}{2}, \tag{3.13}$$

where erf(x) is the error function.[16] Expansion of (3.13) in power series enables a comparison with perturbation series (3.10),

$$u_1(x) = x - \frac{1}{3}x^3 + \frac{2}{15}x^5 - \frac{4}{105}x^7 + \frac{8}{945}x^9 + O(x^{11}), \tag{3.14}$$

which shows that the first approximation reproduces exactly three terms of the perturbation series, that is two more terms than was given correctly by the zero QLM approximation $u_0(x) = x$. This improvement of the representation of the perturbation series not by one, but by two powers of λ is, of course, precisely what one should expect from the quadratic convergence. In addition, the fourth term is also mostly correct being $-\frac{12}{315}$ vis-a-vis exact $-\frac{17}{315}$. The second iteration $u_2(x)$ could not be calculated analytically, but could be computed numerically or expressed by power series expansion with the help of symbolic computation program.[17] The latter gives

$$u_2(x) = x - \frac{1}{3}x^3 + \frac{2}{15}x^5 - \frac{17}{315}x^7 + \frac{62}{2835}x^9 - \frac{1382}{155\,925}x^{11} + \frac{21\,844}{6\,081\,075}x^{13} - \frac{918\,844}{638\,512\,875}x^{15} + \frac{39\,944}{70\,945\,875}x^{17} + O(x^{19}). \tag{3.15}$$

One can see that the second iteration of QLM reproduces correctly the first seven terms of the perturbation series, an improvement by 4 powers of λ compare with previous QLM approximation

$u_1(x)$. In addition, the eighth and ninth terms of the power series expansion of $u_2(x)$ are very close to their precise values in perturbation theory, being $-\frac{918\,844}{638\,512\,875}$ and, 5.63×10^{-4} vis-a-vis exact values $-\frac{929\,569}{638\,512\,875}$ and 5.90×10^{-4}, respectively.

Aside from the fact that already first QLM approximations sum many orders of the usual perturbation theory, the QLM iterations, unlike the perturbation series, have meaning also for a large or even infinite values of coupling constant. Indeed, for $\lambda\to\infty$ any term of the perturbation series is infinite. Even for a finite moderately large potential strength $\lambda\geqslant2.5$ perturbation expansion (3.10) diverges since the power series expansion of the hyperbolic tangent of x_0 converges[16] only for $x_0<\pi/2$, that is for $\lambda<\pi^2/4$. On the other side, the QLM approximations to the scattering length are finite. The first QLM approximation to scattering length (3.13) in view of an asymptotic expression

$$\mathrm{erf}(z)\simeq\left(1-\frac{e^{-z^2}}{\sqrt{\pi}z}\right) \tag{3.16}$$

for $|z|\to\infty$[16] shows that the scattering length in this approximation equals $-R/2$, a reasonable approximation to exact value $a_0=-R$. The computation of the scattering length in the second QLM approximation gives again a finite and improved result $a_0=-\frac{3}{4}R$.

To tackle more rigorously the question of convergence of the iteration series for dimensionless scattering length $A_{0,n}=a_{0,n}/R$ given by Eqs. (3.11) and (3.12) to exact result $A_0=a_0/R$ let us turn to the convergence condition demanding the smallness of convergence coefficient (2.34), which in this case is given by

$$q_1(b)=k(b)\|a_{0,1}-a_{0,0}\|=k(b)\left\|\frac{u_1(x)-u_0(x)}{x}\right\|=k(b)\cdot\max_{0\leqslant x\leqslant b}\left|\frac{u_1(x)}{x}-1\right|. \tag{3.17}$$

To calculate $q_1(b)$ one first has to estimate $k(b)$ using, for example, Eq. (2.35). From Eq. (3.3) and the boundary condition there follows $u(-x)=-u(x)$. We can consider therefore only positive branch of the solution whose extremum is reached when $u'(x)=1-u^2(x)=0$, that is when $u(x)=1$. This means that $0\leqslant u(x)\leqslant1$. Since the first and second functional derivatives of $f(u(x),x)=1-u^2(x)$ equal $-2u(x)$ and -2, respectively, one can set $F(x)=-2$ and $G(x)=|F(x)|=2$, which gives

$$k(b)=\tfrac{1}{2}|e^{-2b}-1|\leqslant\tfrac{1}{2}. \tag{3.18}$$

In view of the fact that, due to the properties[16] of the error function $|u_1(x)/x-1|\leqslant\frac{1}{2}$ for all positive x, one obtains that $q_1(b)\leqslant\frac{1}{4}$ for all values of b. Thus the convergence of QLM approximations Eq. (3.11), and therefore $a_{0,n}$, given by Eq. (3.12), to the exact scattering length a_0 in case of the repulsive square well is uniform and quadratic for all values of x_0, that is for all values of coupling constant λ.

2. Attractive square well

The same conclusions are correct also for the attractive square well potential the equations for which are obtained by changing λ to $-\lambda$. The equation for $u(x)$ now has a form

$$\frac{du(x)}{dx}=1+u^2(x),\quad u(0)=0. \tag{3.19}$$

Its solution is

$$u(x)=\tan x \tag{3.20}$$

and the scattering length is given by

$$a_0=R\left(\frac{\tan\sqrt{\lambda}}{\sqrt{\lambda}}-1\right)\equiv R\left(\frac{u(x_0)}{x_0}-1\right). \tag{3.21}$$

The QLM subsequent approximations are obtained with the help of recursion equations

$$u_{n+1}(x)=\int_0^x ds(1-u_n^2(s))\exp\left(2\int_s^x dt\, u_n(t)\right). \tag{3.22}$$

Choosing the zero QLM approximation as before in form $u_0(x)=x$ leads to first QLM approximation,

$$u_1(x)=\frac{\sqrt{\pi}}{4}\operatorname{erf}(x)e^{x^2}+\frac{x}{2}. \tag{3.23}$$

Now there is, however, an additional difficulty, since exact scattering length $a_0(x_0)$ is a singular function of $x_0=\sqrt{\lambda}$ and becomes infinite at values of the coupling constant corresponding to zero bound state energies $\lambda=((2n+1)\,\pi/2)^2$. This finds reflection in the fact that $u_1(x_0)$ is increasing very fast for x_0 around $\pi/2$. To deal with it let us, in accordance with the discussion in Sec. II, regularize Eq. (3.19), that is to rewrite it for $|u(x)|>1$ in terms of a new function

$$v(x)=\frac{1}{u(x)}. \tag{3.24}$$

Defining c as a singular point where $u(c)=\infty$ one obtains, according to Eq. (2.38), the following nonlinear equation for $v(x)$:

$$\frac{dv(x)}{dx}=-(1+v(x)^2), \quad v(c)=\frac{1}{u(c)}=0. \tag{3.25}$$

In view of Eq. (3.19) a solution of Eq. (3.25) is $v(x)=u(c-x)$. Equation (3.24) then gives

$$u(x)=\frac{1}{u(c-x)}. \tag{3.26}$$

Setting $x=c/2$ allows us to write

$$u^2\left(\frac{c}{2}\right)=1 \tag{3.27}$$

for constant c. Since the solution of Eq. (3.19) should be an odd function of x,

$$u(-x)=-u(x), \tag{3.28}$$

it is enough to choose only a positive branch of Eq. (3.27), that is

$$u\left(\frac{c}{2}\right)=1. \tag{3.29}$$

From Eqs. (3.26) and (3.28) follows the $2c$ periodicity of solution $u(x)$: $u(x+2c)=1/u(c-(x+2c))=-1/u(x+c)=-u(c-(c+x))=u(x)$. Thus it is enough to find a solution only on the interval $(0,2c)$. We can now formulate the following result.

The nth QLM approximation $U_n(x)$ to the solution of Eq. (3.19) on the interval $[0,2c_n]$, which is able to properly describe a singularity, is given by

$$U_n(x)=u_n(x)\Theta\left(\frac{c_n}{2}-x\right)\Theta(x)+\frac{1}{u_n(c_n-x)}\Theta\left(x-\frac{c_n}{2}\right)\Theta\left(\frac{3c_n}{2}-x\right)$$
$$+u_n(x-2c_n)\Theta\left(x-\frac{3c_n}{2}\right)\Theta(2c_n-x), \tag{3.30}$$

where the nth QLM approximation $u_n(x)$ on interval $(0,c_n/2)$ is found with the help of recurrence relations Eq. (3.22) and the nth approximate value c_n of c is given by

$$u_n\left(\frac{c_n}{2}\right)=1. \tag{3.31}$$

Computation of $c_n/2$ shows that the differences between the exact value $c=\pi/2$ and approximate values c_n are very small already for the first and second QLM iterations, namely $(c_1-\pi/2)$ and $(c_2-\pi/2)$ are 0.005 29 and 0.000 001 32, the errors of 0.5% and 10^{-4}%, respectively. Since the nth QLM approximation, Eq. (3.30), has a pole at $x_0=c_n$, $\lambda=c_n^2$ gives a value of potential strength corresponding to a zero energy bound state. One sees that the QLM description of such state is extremely accurate already in the first and especially in the second approximations.

To prove the uniform quadratic convergence of the QLM iterations it is enough, in view of Eqs. (3.28) and (3.30) to consider $u_n(x)$ only on intervals $(0,c_n/2)$ which are very close to interval $(0,\pi/4)$. Since the first and second functional derivatives of the left-hand side of Eq. (3.19) are $2u(x)$ and 2, respectively, and $|u(x)|\leqslant 1$, one can chose $F(x)=G(x)=2$ and use Eq. (2.35), which produces a simple expression for $k(b)$,

$$k(b)=\tfrac{1}{2}(e^{2b}-1). \tag{3.32}$$

This leads to the following result for $q_1(b)$:

$$q_1(b)=\frac{1}{2}(e^{2b}-1)\left(\frac{\sqrt{\pi}}{4}\operatorname{erf}(b)e^{b^2}-\frac{b}{2}\right). \tag{3.33}$$

A simple computation shows that $0<q_1(b)<1$ for $0<b<0.92$, which proves the uniform quadratic convergence of the QLM iterations on even larger interval (0,0.92) than interval $(0,\pi/4)$ and thus the convergence of the sequence $U_n(x_0)$ to the exact solution $\tan x_0$ on the interval $(0,2c_n)$ $\approx(0,\pi)$. In view of its $2c_n\approx\pi$ periodicity the nth QLM approximation $U_n(x_0)$ converges therefore to the exact solution for all x_0, that is for all values of the coupling constant λ.

The extremely fast convergence of QLM approximations given by Eq. (3.30) is evident from the ratios of the first [Eq. (3.23)] and second [Eq. (3.22) for n=1] QLM iterations to the exact solution (3.20), which are shown in Figs. 1 and 2, respectively.

B. δ-function potential

In case of the δ-function potential

$$V(r)=\frac{\lambda}{R}\delta(r-R), \tag{3.34}$$

Eq. (1.1) for the scattering length has the form

$$A'(x)=\lambda(x-A(x))^2\delta(x-1)\equiv\lambda(1-A(x))^2\delta(x-1),\quad A(0)=0, \tag{3.35}$$

where $x=r/R$ and $A(x)=-a(r)/R$ are dimensionless variable and variable scattering length, respectively; note, that in Eq. (3.35) $A(x)$ could not be set equal $A(1)$, since $A(x)$ is discontinu-

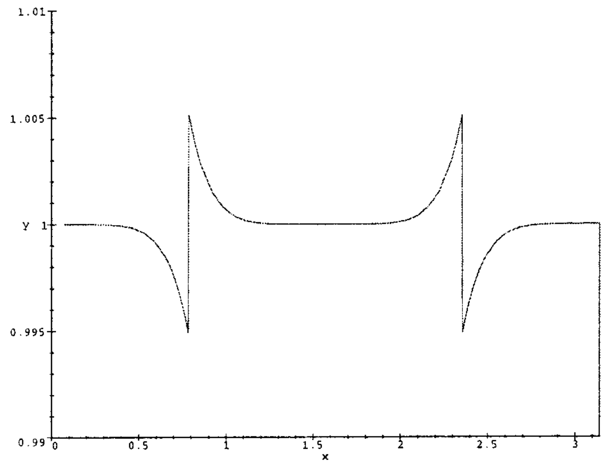

FIG. 1. The ratio of the first QLM iteration to the exact solution for the attractive square well as a function of the potential strength λ (axis x).

ous at $x=1$, its derivative being proportional to the δ-function. Introduction of a new function $y(x)=\lambda\Theta(x-1)$, $y(0)=0$, $y(\infty)=\lambda$ with a derivative $dy(x)=\lambda\,\delta(x-1)dx$ reduces Eq. (3.35) to the form

$$\frac{dA(y)}{dy}=(1-A(y))^2, \quad A(y)_{y=0}=0. \tag{3.36}$$

A solution of Eq. (3.36) is

$$A(y)=\frac{y}{1+y}. \tag{3.37}$$

An exact solution of Eq. (1.1) for the δ-potential thus is given by $a(r)=-RA(y)\equiv-R[\lambda\Theta(r-R)]/[1+\lambda\Theta(r-R)]$. The scattering length a_0 equals $a(r)_{r=\infty}\equiv-R\lambda/(1+\lambda)$. It is is singular at $\lambda=-1$, reflecting the existence of the zero energy bound state for the unit potential strength.

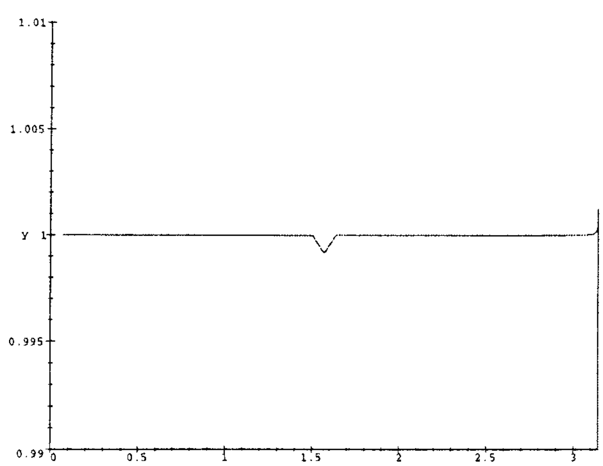

FIG. 2. Same as in Fig. 1, but for the second QLM approximation.

1. Repulsive δ-function potential

Let us now consider QLM approximations to the exact solution (3.37) in the case of the repulsive δ-function potential, $\lambda>0$. According to Eq. (2.22) they are given by the following iteration sequence:

$$A_{n+1}(y)=\int_0^y ds(1-A_n^2(s))\exp\left(-2\int_s^y dt(1-A_n(t)\right), \tag{3.38}$$

since the functional derivative of the right-hand part of Eq. (3.36) equals $-2(1-A(y))$. The introduction of the nth approximation $u_n(y)=1-A_n(y)$ to a function $u(y)=1-A(y)=1/(1+y)$ helps to write recurrence relationship (3.38) in a simpler form:

$$u_{n+1}(y)=\exp\left(-2\int_0^y dt\, u_n(t)\right)+\int_0^y ds\, u_n^2(s))\exp\left(-2\int_s^y dt\, u_n(t)\right), \tag{3.39}$$

which coincides with the QLM iteration scheme (2.23) for Eq. (3.36), rewritten with the help of the function $u(x)=1-A(x)$ as

$$u'(x)=-u(x)^2, \quad u(0)=1, \tag{3.40}$$

Since for $x=\infty$ $y=\lambda$, $u_n(\lambda)$ gives the nth approximation to $u(\lambda)=1-A_0(\lambda)=1/(1+\lambda)$ where $A_0(\lambda)$ is the exact dimensionless scattering length.

Let us chose as a zero approximation $u_0(y)\equiv u(0)=1$. The substitution in Eq. (3.39) for $n=0$ gives

$$u_1(y)=\tfrac{1}{2}(1+e^{-2y}). \tag{3.41}$$

One can see that already the first approximation $u_1(\lambda)$ for $\lambda\to\infty$ is finite and equals $\frac{1}{2}$, which gives a value of $\frac{1}{2}$ for the approximate dimensionless scattering length vis-a-vis the exact value $A_0=1$. Each term in the perturbation series for $u(\lambda)$,

$$u(\lambda)=\sum_{m=0}^{\infty}(-\lambda)^m, \tag{3.42}$$

in this case is infinite while the perturbation expansion itself is divergent already for $|\lambda|\geqslant 1$. The comparison of perturbation expansion (3.42) with the perturbative expansion of the first QLM approximation (3.41),

$$u_1(\lambda)=\tfrac{1}{2}(1+e^{-2\lambda})=1-\lambda+\lambda^2-\tfrac{2}{3}\lambda^3+\tfrac{1}{3}\lambda^4-\tfrac{2}{15}\lambda^5+O(\lambda^6), \tag{3.43}$$

shows that in this approximation the perturbation series is correct up to the fourth term. The next, second approximation also could be calculated analytically with the help of symbolic computation program[17] and gives the rather cumbersome expression

$$u_2(\lambda)=-1/4\,\frac{-2\sqrt{e^{e^{-2\lambda}}}-e^{1/2-\lambda}+\sqrt{2\pi}\,\mathrm{erf}\left(\dfrac{e^{-\lambda}}{\sqrt{2}}\right)\sqrt{e^{e^{-2\lambda}}}e^{1/2}-e^{1/2+\lambda}-\sqrt{2\pi}\,\mathrm{erf}\left(\dfrac{1}{\sqrt{2}}\right)e^{1/2}\sqrt{e^{e^{-2\lambda}}}}{\sqrt{e^{e^{-2\lambda}}}e^{1/2(-e^{-2\lambda}+2\lambda+1)}}. \tag{3.44}$$

For $\lambda\to\infty$ the largest term both in the numerator and denominator is $e^{1/2+\lambda}$. Therefore $u_2(\infty)=\frac{1}{4}$, which corresponds to the second QLM approximation to A_0 being $\frac{3}{4}$, a significant improvement compared with the result, obtained in this limit in the first QLM approximation (3.41). The computation of the power series expansion yields

$$u_2(\lambda)=1-\lambda+\lambda^2-\lambda^3+\lambda^4-\lambda^5+\lambda^6-\frac{62}{63}\lambda^7+\frac{79}{84}\lambda^8-\frac{4931}{5670}\lambda^9+O(\lambda^{10}). \tag{3.45}$$

The perturbation series in the second QLM approximation is given correctly up to the seventh term, while the coefficients of the eighth and ninth terms are different only by $\frac{1}{63}$ and $\frac{5}{84}$, that is by 1.6% and 6%, respectively.

Analytic calculation of the third QLM approximation seems impossible but the power series expansion could be evaluated with the help of the same program,[17] which yields

$$\begin{aligned}u_3(\lambda)=&1-\lambda+\lambda^2-\lambda^3+\lambda^4-\lambda^5+\lambda^6-\lambda^7+\lambda^8-\lambda^9+\lambda^{10}-\lambda^{11}+\lambda^{12}-\lambda^{13}+\lambda^{14}\\&-\frac{59\,534}{59\,535}\lambda^{15}+\frac{1\,904\,891}{1\,905\,120}\lambda^{16}-\frac{12\,139\,457}{12\,145\,140}\lambda^{17}\\&+\frac{161\,721\,779}{161\,935\,200}\lambda^{18}-\frac{113\,880\,892\,943}{114\,225\,041\,700}\lambda^{19}+O(\lambda^{20}).\end{aligned} \tag{3.46}$$

In the third QLM approximation the first 15 terms of the perturbation series are given exactly while the next 5 terms have coefficients extremely close to being exact.

Summing up, the number of the terms given precisely in the zero, first, second, and third QLM approximations equals 1, 3, 7, and 15, increasing by 2, 2^2 and 2^3, respectively, that is according to geometric progression with $q=2$, exactly as one should expect from the quadratic law of the convergence. The number N_n of perturbation series terms reproduced exactly in the nth QLM approximation is therefore

$$N_n=\sum_{k=0}^{n} q^k=\frac{q^{n+1}-1}{q-1}=2^{n+1}-1 \tag{3.47}$$

and for larger n approximately doubles with n increasing by each unit. For example, the sixth QLM approximation reproduces exactly $2^7-1=127$ terms of the perturbation expansion, while the twelfth approximation reproduces already $2^{13}-1=8191$ terms, and so on.

The numerical computation of $u_3(\infty)$ gives 0.125, corresponding to $A_0=0.875$, a finite and gratifying result.

Comparison of the first three QLM approximations $u_n(\lambda)$, $n=1,2,3$ with exact solution $u(\lambda)=1/(1+\lambda)$ and its perturbation expansion (3.42) containing 15 terms (up to λ^{14}, inclusively) for the δ-function potential with the potential strength λ changing in the interval (0,10) is shown graphically in Fig. 3. One can see that each subsequent QLM approximation reproduces the exact solution better than the previous one up to infinite values of the coupling constant, while even the 15th-order perturbation theory is not able to describe the exact solution adequately beyond $\lambda=1$.

To prove the uniform quadratic convergence of QLM iterations let us note that the first and second functional derivatives of the left hand side of Eq. (3.40) are $-2u(x)$ and -2, respectively, exactly as in the case of the repulsive square well which was discussed earlier. The extremal value of $u(x)$, reached when $u'(x)=-u^2(x)=0$, is, obviously, zero, which, in view of boundary condition $u(0)=1$, means $0\leqslant u(x)\leqslant 1$. This allows one to choose the same functions $F(x)=-2$, $G(x)=2$ as for the repulsive square well, and consequently results in the same expression 3.18 for $k(b)$. Since it follows from Fig. 3 that the maximal difference between zero and first QLM approximations $\|u_1(x)-u_1(x)\|$ equals $\frac{1}{2}$, one obtains as before $q_1(b)\leqslant\frac{1}{4}$, which proves the uniform quadratic convergence of the QLM iterations for all values of b. This means that the

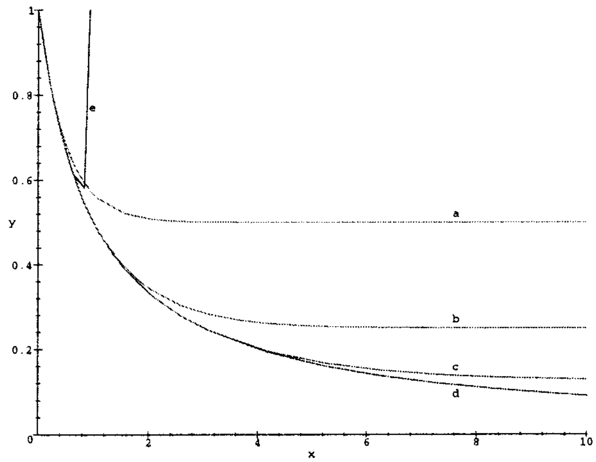

FIG. 3. Comparison of first three QLM approximations $u_n(\lambda), n=1,2,3$, curves a, b, and c, respectively, with exact solution $u(\lambda)=1/(1+\lambda)$, curve d, and its perturbation expansion (3.42), curve e, containing 15 terms (up to λ^{14}, inclusively), for the δ-function potential with the potential strength λ (axis x) changing in the interval (0,10).

convergence of subsequent QLM approximations to the exact scattering length for the repulsive δ-function potential is uniform and quadratic for all values of coupling constant λ, including very large and infinite ones.

2. Attractive δ-function potential

For negative λ the subsequent approximations $u_n(\lambda)$ start to increase very rapidly with $|\lambda|$ as one can see, for example, from analytic expressions (3.41) and (3.44). According to discussion before Eq. (2.38) we have to switch in this case in Eq. (3.40) to a new function $v(x)=1/u(x)$, which thus satisfies the trivial equation $v'(x)=1$ with a boundary condition $v(0)=1$. The QLM solution of this equation in the nth approximation, calculated from Eq. (2.23), is $v_n(x)=1+x$ or $u_n(x)=1/(1+x)$ for any n, which means that this form of the equation for the attractive δ-function potential generates an exact solution in any QLM approximation and there is no need for further investigation.

C. Inverse square potential

Let us consider now the inverse square potential

$$V(r)=\frac{\lambda}{r^2}\Theta(R-r), \tag{3.48}$$

where λ is the dimensionless coupling constant. As is well known,[18] this potential produces a fall to the center in case of $\lambda<-\frac{1}{4}$. For $r\leqslant R$, Eq. (1.1) for the scattering length could be written in the form

$$A'(x)=-\lambda\left(1+\frac{A(x)}{x}\right)^2, \quad A(0)=0, \tag{3.49}$$

where $x=r/R$ and $A(x)=a(r)/R$ are the dimensionless variable and variable scattering length, respectively; for $x>1$, $A(x)\equiv A(1)$ is a constant and represents the dimensionless scattering length A_0. Looking for a solution in the form $A(x)=x\alpha(x)$, we obtain for $\alpha(x)$,

$$\alpha'(x)=-\frac{1}{x}[\alpha(x)+\lambda(1+\alpha(x))^2. \tag{3.50}$$

Note, that in this equation boundary condition $\alpha(0)=0$ is not necessary: $\alpha(x)$ could be any function regular at $x=0$ so that condition $A(0)=0$ is satisfied. Setting $\alpha(x)=\text{constant}\equiv A_0$ gives an algebraic equation $A_0=-\lambda(1+A_0)^2$ whose solution is given by $A_0=-1-1/2\lambda(1\pm\sqrt{1+4\lambda})$. Since for $\lambda\to 0$ there should be no scattering only solution with the minus sign before the square root should be chosen, since only for this solution $A_0\to 0$ when $\lambda\to 0$. Setting for convenience $g=4\lambda$ we finally obtain

$$A_0=-1-\frac{2}{g}(1-\sqrt{1+g}). \tag{3.51}$$

The solution has a singularity, namely a branch point, at $g=-1$, that is at $\lambda=-\frac{1}{4}$. The singularity marks the beginning of interval $-\infty<\lambda<-\frac{1}{4}$ where a fall to the center takes place[18] and the expression for the scattering length becomes complex, its real and imaginary parts for $g<-1$ are given by

$$\operatorname{Re}A_0=-1-\frac{2}{g},\quad \operatorname{Im}A_0=\frac{2}{g}\sqrt{-1-g}. \tag{3.52}$$

Note that in view of our definition (3.6) of the scattering length one has to chose $\operatorname{Im}A_0\geqslant 0$.[18] The perturbation series for the scattering length could be obtained by expansion of the square root in Eq. (3.51) in the power series which gives

$$\begin{aligned}
A_0=&-\frac{1}{4}g+\frac{1}{8}g^2-\frac{5}{64}g^3+\frac{7}{128}g^4-\frac{21}{512}g^5+\frac{33}{1024}g^6-\frac{429}{16\,384}g^7\\
&+\frac{715}{32\,768}g^8\frac{2431}{131\,072}g^9+\frac{4199}{262\,144}g^{10}-\frac{29\,393}{2\,097\,152}g^{11}+\frac{52\,003}{4\,194\,304}g^{12}\\
&-\frac{185\,725}{16\,777\,216}g^{13}+\frac{334\,305}{33\,554\,432}g^{14}-\frac{9\,694\,845}{1\,073\,741\,824}g^{15}+\frac{17\,678\,835}{2\,147\,483\,648}g^{16}\\
&-\frac{64\,822\,395}{8\,859\,934\,592}g^{17}+\frac{119\,409\,675}{17\,179\,869\,184}g^{18}-\frac{883\,631\,595}{137\,438\,953\,472}g^{19}\\
&+\frac{1\,641\,030\,105}{274\,877\,906\,944}g^{20}-\frac{6\,116\,566\,755}{1\,099\,511\,627\,776}g^{21}+\frac{11\,435\,320\,455}{2\,199\,023\,255\,552}g^{22}\\
&-\frac{171\,529\,806\,825}{35\,184\,372\,088\,832}g^{23}+\frac{322\,476\,036\,831}{70\,368\,744\,177\,664}g^{24}-\frac{1\,215\,486\,600\,363}{281\,474\,976\,710\,656}g^{25}\\
&+\frac{2\,295\,919\,134\,019}{562\,949\,953\,421\,312}g^{26}-\frac{17\,383\,387\,729\,001}{4\,503\,599\,627\,370\,496}g^{27}\\
&+\frac{32\,968\,493\,968\,795}{9\,007\,199\,254\,740\,992}g^{28}-\frac{125\,280\,277\,081\,421}{36\,028\,797\,018\,963\,968}g^{29}\\
&+\frac{238\,436\,656\,380\,769}{72\,057\,594\,037\,927\,936}g^{30}-\frac{14\,544\,636\,039\,226\,909}{4\,611\,686\,018\,427\,387\,904}g^{31}\\
&+\frac{27\,767\,032\,438\,524\,099}{9\,223\,372\,036\,854\,775\,808}g^{32}+O(g^{33}).
\end{aligned} \tag{3.53}$$

The expansion is convergent[16] for $|g|<1$, that is for $|\lambda|<\frac{1}{4}$.

Let us now turn our attention to QLM approximations and their convergence. The QLM iterations sequences are easiest to find by considering differential form 2.3 of Eq. (2.22) which could be written as

$$\alpha'_{n+1}(x) = -\frac{1}{x}\left[\frac{g}{4}(1-\alpha_n^2(x)) + \alpha_{n+1}(x)\left(1+\frac{g}{2}(1+\alpha_n(x))\right)\right]. \tag{3.54}$$

The assumption that $\alpha_n(x)$ are constant functions, $\alpha_n(x)\equiv c_n$, immediately establish the QLM recurrence relationship

$$c_{n+1} = -g\frac{1-c_n^2}{4+2g(1+c_n)}. \tag{3.55}$$

Note that since $c_{n+1}\to 0$ when $g\to 0$ each approximation to the scattering amplitude vanishes for $g=0$ as it should be since in the absence of the potential there is no scattering. The convergence of the QLM iteration sequence to the exact solution (3.51) is obvious. Indeed, for $n\to\infty$, Eq. (3.55) is

$$c_\infty = -g\frac{1-c_\infty^2}{4+2g(1+c_\infty)}, \tag{3.56}$$

whose solution vanishing for $g\to 0$ is given by the expression for A_0 in Eq. (3.51). The QLM approximation c_n to the dimensionless scattering length for an infinite n therefore indeed is $c_\infty \equiv A_0$ as we wanted to show.

The explicit calculation of the first few QLM approximations, starting from the usual initial guess $c_0=0$ gives

$$c_1 = -\frac{g}{4+2g}, \tag{3.57}$$

$$c_2 = -1/4\frac{(16+16g+3g^2)g}{(8+8g+g^2)(2+g)}, \tag{3.58}$$

$$c_3 = -1/8\frac{(4096+12\,288g+14\,080g^2+7680g^3+2016g^4+224g^5+7g^6)g}{(128+256g+160g^2+32g^3+g^4)(2+g)(8+8g+g^2)}. \tag{3.59}$$

These expressions, unlike that of the perturbation theory, give finite values also for $g>1$ or even for $g=\infty$, where the first, second, and third QLM approximations give $-\frac{1}{2}$, $-\frac{3}{4}$, $-\frac{7}{8}$ vis-a-vis the exact value $A_0=-1$; the fourth approximation, not given here because of its cumbersome form, results in $-\frac{15}{16}$, and so on. The convergence of these values is from above in agreement with the law of convergence for the concave functions proved in Sec. II, since the second functional derivative $-\lambda/x^2$ of the right-hand side of Eq. (3.49) is negative for the repulsive potential.

The expansion of the QLM approximations in the power series in the coupling constant shows as in previous examples that each QLM iteration sums exactly many perturbation series terms, whose number is given by Eq. (3.47). One obtains:

$$c_0 = 0, \tag{3.60}$$

$$c_1 = -\frac{1}{4}g + \frac{1}{8}g^2 - \frac{1}{16}g^3 + \frac{1}{32}g^4 - \frac{1}{64}g^5 + O(g^6), \tag{3.61}$$

$$c_2 = -\frac{1}{4}g + \frac{1}{8}g^2 - \frac{5}{64}g^3 + \frac{7}{128}g^4 - \frac{21}{512}g^5 + \frac{33}{1024}g^6 - \frac{107}{4096}g^7 + \frac{177}{8192}g^8 - \frac{593}{32\,768}g^9 + O(g^{10}), \tag{3.62}$$

$$c_3 = -\frac{1}{4}g + \frac{1}{8}g^2 - \frac{5}{64}g^3 + \frac{7}{128}g^4 - \frac{21}{512}g^5 + \frac{33}{1024}g^6 - \frac{429}{16\,384}g^7 + \frac{715}{32\,768}g^8 - \frac{2431}{131\,072}g^9 + \frac{4199}{262\,144}g^{10} - \frac{29\,393}{2\,097\,152}g^{11} + \frac{52\,003}{4\,194\,304}g^{12} - \frac{185\,725}{16\,777\,216}g^{13} + \frac{334\,305}{33\,554\,432}g^{14} - \frac{2\,423\,711}{268\,435\,456}g^{15} + \frac{4\,419\,705}{536\,870\,912}g^{16} - \frac{16\,205\,537}{2\,147\,483\,648}g^{17} + \frac{29\,852\,049}{4\,294\,967\,296}g^{18} - \frac{220\,900\,693}{34\,359\,738\,368}g^{19} + O(g^{20}), \tag{3.63}$$

$$c_4 = -\frac{1}{4}g + \frac{1}{8}g^2 - \frac{5}{64}g^3 + \frac{7}{128}g^4 - \frac{21}{512}g^5 + \frac{33}{1024}g^6 - \frac{429}{16\,384}g^7 + \frac{715}{32\,768}g^8 - \frac{2431}{131\,072}g^9 + \frac{4199}{262\,144}g^{10} - \frac{29\,393}{2\,097\,152}g^{11} + \frac{52\,003}{4\,194\,304}g^{12}\frac{185\,725}{16\,777\,216}g^{13} + \frac{334\,305}{33\,554\,432}g^{14} - \frac{9\,694\,845}{1\,073\,741\,824}g^{15} + \frac{17\,678\,835}{2\,147\,483\,648}g^{16} - \frac{64\,822\,395}{8\,589\,934\,592}g^{17} + \frac{119\,409\,675}{17\,179\,869\,184}g^{18} - \frac{883\,631\,595}{137\,438\,953\,472}g^{19} + \frac{1\,641\,030\,105}{274\,877\,906\,944}g^{20} - \frac{6\,116\,566\,755}{1\,099\,511\,627\,776}g^{21} + \frac{11\,435\,320\,455}{2\,199\,023\,255\,552}g^{22} - \frac{171\,529\,806\,825}{35\,184\,372\,088\,832}g^{23} + \frac{322\,476\,036\,831}{70\,368\,744\,177\,664}g^{24} - \frac{1\,215\,486\,600\,363}{281\,474\,976\,710\,656}g^{25} + \frac{2\,295\,919\,134\,019}{562\,949\,953\,421\,312}g^{26} - \frac{17\,383\,387\,729\,001}{4\,503\,599\,627\,370\,496}g^{27} + \frac{32\,968\,493\,968\,795}{9\,007\,199\,254\,740\,992}g^{28} - \frac{125\,280\,277\,081\,421}{36\,028\,797\,018\,963\,968}g^{29} + \frac{238\,436\,656\,380\,769}{72\,057\,594\,037\,927\,936}g^{30} - \frac{3\,636\,159\,009\,806\,727}{1\,152\,921\,504\,606\,846\,976}g^{31} + \frac{6\,941\,758\,109\,631\,017}{2\,305\,843\,009\,213\,693\,952}g^{32} + O(g^{33}). \tag{3.64}$$

Comparison of Eqs. (3.60)–(3.64) with Eq. (3.53) shows that the QLM iterations with $n = 0,1,2,3,4$ reproduce exactly 1,3,7,15,31 terms of the perturbation series, respectively, in exact agreement with Eq. (3.47), while the next few terms have coefficients extremely close to being exact. The number of terms given precisely by the zero, first, second, third and fourth QLM approximations is increasing by 2, 2^2, 2^3 and 2^4, exactly as we saw earlier in the case of the δ-function potential and in precise agreement with the quadratic law of the convergence, proved in Sec. II. Due to simplicity of the algebraic recurrence relations (3.55) Eq. (3.47) for number N_n of the perturbation series terms given precisely by the nth QLM approximation could be checked for higher QLM approximations. For example, in Sec. III B on the example of the repulsive δ-potential we concluded that $N_6=127$. The simple calculation using a symbolic manipulation program[17] shows immediately that it is precisely the same for the inverse square potential. Indeed, the first seven nonzero terms of the expansion in powers of g of difference A_0-c_6 between exact scattering length Eq. (3.51) and its sixth QLM approximation are

$$-\frac{1}{28\,948\,022\,309\,329\,048\,855\,892\,746\,252\,171\,976\,963\,317\,496\,166\,410\,141\,009\,864\,396\,001\,978\,282\,409\,984}g^{127}$$

$$+\frac{127}{57\,896\,044\,618\,658\,097\,711\,785\,492\,504\,343\,953\,926\,634\,992\,332\,820\,282\,019\,728\,792\,003\,956\,564\,819\,968}g^{128},$$

$$-\frac{16319}{231\,584\,178\,474\,632\,390\,847\,141\,970\,017\,375\,815\,706\,539\,969\,331\,281\,128\,078\,915\,168\,015\,826\,259\,279\,872}g^{129},$$

$$+\frac{707\,135}{463\,168\,356\,949\,264\,781\,694\,283\,940\,034\,751\,631\,413\,079\,938\,662\,562\,256\,157\,830\,336\,031\,652\,518\,559\,744}g^{130},$$

$$-\frac{92\,988\,123}{3\,705\,346\,855\,594\,118\,253\,554\,271\,520\,278\,013\,051\,304\,639\,509\,300\,498\,049\,262\,642\,688\,253\,220\,148\,477\,952}g^{131},$$

$$+\frac{2\,473\,622\,041}{7\,410\,693\,711\,188\,236\,507\,108\,543\,040\,556\,026\,102\,609\,279\,018\,600\,996\,098\,525\,285\,376\,506\,440\,296\,955\,904}g^{132},$$

$$-\frac{110\,916\,205\,323}{29\,642\,774\,844\,752\,946\,028\,434\,172\,162\,224\,104\,410\,437\,116\,074\,403\,984\,394\,101\,141\,506\,025\,761\,187\,823\,616}g^{133}, \tag{3.65}$$

exactly as one expects from Eq. (3.47). In addition, one can see that the next terms of the perturbation series are also reproduced extremely well, their difference with the precise terms being infinitesimally small. Namely, the coefficient of 127th power of g is about 3.45×10^{-76}, the coefficient of 128th power is about 2.19×10^{-74}, and so on.

For the attractive potential expressions (3.57)–(3.59) become singular, with the number of zeros of denominators increasing with each iteration. This, of course, is a reflection of the fact that the exact scattering length A_0 has a branch point at $g=-1$ and a cut line along the real axis between $g=-1$ and $g=-\infty$. When n is increasing, the poles are getting closer and closer to each other and fuse together at $n=\infty$, where, as we saw earlier, the exact amplitude and its singularity are reproduced.

To handle the singularities one can try, as we have discussed earlier, to consider instead of the function $\alpha(x)$ a new function $\beta(x)$ such that $\alpha(x)=1/\beta(x)$. Substitution of Eq. (3.65) into Eq. (3.50) leads to

$$\beta'(x)=\frac{1}{x}[\beta(x)+\lambda(1+\beta(x))^2, \tag{3.66}$$

which is different from Eq. (3.50) only by the sign of the right-hand side. The QLM iterations sequence is found as before by considering differential form (2.3) of Eq. (2.22):

$$\beta'_{n+1}(x)=\frac{1}{x}\left[\frac{g}{4}(1-\beta_n^2(x))+\beta_{n+1}(x)\left(1+\frac{g}{2}(1+\beta_n(x))\right)\right], \tag{3.67}$$

which leads under a previous assumption of β_n being a constant function, $\beta_n\equiv c_n$, to exactly the same QLM recurrence relations (3.55). Again, the convergence of the QLM series follows from the fact that at $n\to\infty$ we have the same Eq. (3.56), as before, with only distinction—since now the scattering amplitude in the limit $n=\infty$ is given by $1/\beta_\infty$, one should take a solution of this equation which is going to infinity at $g\to0$ rather than to zero. Such solution is given by

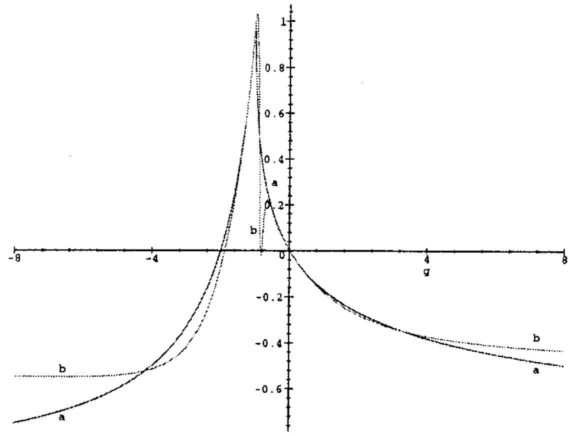

FIG. 4. Comparison of real parts of the exact scattering length (curve a) and of the second QLM approximation to it (curve b) for inverse square potential, $|g|\leqslant 8$.

$$\beta_\infty=-1-\frac{2}{g}(1+\sqrt{1+g}). \tag{3.68}$$

The $n=\infty$ QLM approximation to the scattering length A_0 thus equals to

$$\frac{1}{\beta_\infty}=\frac{1}{-1-\frac{2}{g}(1+\sqrt{1+g})}$$

$$\equiv-1-\frac{2}{g}(1-\sqrt{1+g}), \tag{3.69}$$

which indeed coincides exactly with expression (3.51) for A_0.

Since the change to $\beta_n(x)=1/\alpha_n(x)$ does not give anything new, the only way to avoid the singularities in the case of attractive potential seems therefore to use the fact that the zero approximation could be an arbitrary, not necessarily real, number, and to choose c_0 as a complex number with a positive imaginary part of the same order as a real part. The necessity of choosing c_0 complex in the case of the attractive potential follows also from the fact that in this case the fall to the center happens. The inelastic cross section for zero energies, determined by the imaginary part of the S-wave scattering length,[18] could not therefore be zero; however, from recurrence relations (3.55) it is obvious that unless the initial guess c_0 is a complex number, all subsequent QLM approximations are real.

Comparison of real and imaginary parts of the scattering length with those calculated in the second and third QLM approximations for an arbitrary initial guess $\alpha_0=1+i$ and for coupling constant values $|\lambda|\leqslant 2$ ($|g|\leqslant 8$) is shown in Figs. 4–7. One can see that already for the second QLM iteration the agreement between the exact scattering length and the QLM approximation to it is quite good. It improves visibly for the next QLM iteration. For the fourth and next iterations the distinction between exact and approximate scattering length is difficult to see and therefore the correspondent graphs are not shown.

D. Inverse quartic potential

Our next and last example is the inverse quartic potential of radius ρ,

$$V(r)=\lambda\frac{R^2}{r^4}\Theta(\rho-r), \tag{3.70}$$

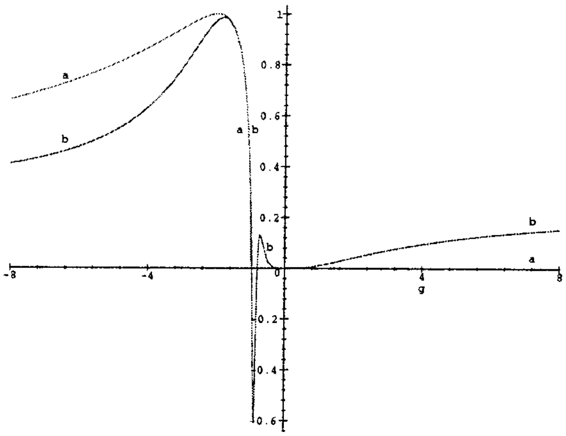

FIG. 5. Comparison of imaginary parts of the exact scattering length (curve a) and of the second QLM approximation to it (curve b) for inverse square potential, $|g|\leqslant 8$.

where λ is a dimensionless coupling constant. For $r\leqslant\rho$ the equation for a variable scattering length $a(r)$ is given by

$$\frac{da(r)}{dr}=-\lambda\frac{R^2}{r^4}(r+a(r))^2,\quad a(0)=0, \tag{3.71}$$

while the scattering length a_0 equals $a(\rho)$. Introduction of the dimensionless function $\alpha(x)=a(r)/R\sqrt{\lambda}$ and of the dimensionless variable $x=r/R\sqrt{\lambda}$ reduces Eq. (3.71) to

$$\frac{d\alpha(x)}{dx}=-\frac{1}{x^4}(x+\alpha(x))^2,\quad \alpha(0)=0, \tag{3.72}$$

whose solution is obviously given by

$$\alpha(x)=-\frac{x}{1+x}. \tag{3.73}$$

Since for $r=\rho$ $x=x_0=\rho/R\sqrt{\lambda}$, and $\alpha(\rho/R\sqrt{\lambda})=-\rho/(\rho+R\sqrt{\lambda})$, the dimensionless scattering length $A_0=a_0/R$ is given by

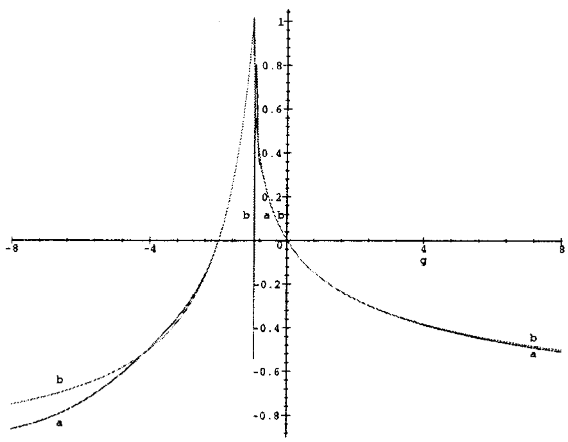

FIG. 6. Same as in Fig. 4, but in the third QLM approximation.

$$A_0=\sqrt{\lambda}\,\alpha\left(\frac{\rho}{R\sqrt{\lambda}}\right)=-\frac{\rho\sqrt{\lambda}}{\rho+R\sqrt{\lambda}}\equiv-\frac{\rho}{R}\,\frac{y_0}{1+y_0},\quad y_0=\frac{1}{x_0}=\frac{R}{\rho}\sqrt{\lambda}, \tag{3.74}$$

which coincides with that found in Ref. 19 and also in Refs. 4, 11, and 14, where ρ is set to ∞. The solution has a singularity, namely a branch point, at $\lambda=0$. The singularity marks the beginning of interval $-\infty<\lambda<0$, where potential is attractive and a fall to the center takes place[18,20] and where the expression for the scattering length A_0 becomes complex. We consider therefore only repulsive potentials.

As always, let us start from the perturbation expansion (3.7), whose coefficients $a_n(r)$ are calculated from recurrence relations (3.8). In view of a strong singularity of the potential at the origin from Eq. (3.9) it follows, however, that first coefficient $a_1(r)$, and therefore all the others, are infinite. The perturbation expansion does not exist, which, of course, is a consequence of the branch point singularity of the scattering length directly at $\lambda=0$.

There is, however, no problems with the QLM approximations to solution (3.73), whose iteration sequence in this case is given by

$$\alpha_{n+1}(x)=-\int_0^x\frac{ds}{s^4}[s^2-\alpha_n^2(s)]\exp\left(-2\int_s^x\frac{dt}{t^4}(t+\alpha_n(t))\right). \tag{3.75}$$

Starting from the usual initial guess $\alpha_0(x)=0$ one easily computes[4,11] the first iteration $\alpha_1(x)=-\exp(1/x^2\int_0^x ds/s^2\,e^{-1/s^2})$, which substitution $s=1/\sqrt{t}$ reduces to a form $\alpha_1(x)=-\frac{1}{2}\exp(1/x^2\int_{1/x^2}^\infty dt/t^{1/2}\,e^{-t})\equiv-\sqrt{\pi}/2\,e^{1/x^2}\,\mathrm{erfc}(1/x)$. The dimensionless scattering length in the first QLM approximation therefore is

$$A_{0,1}=\sqrt{\lambda}\,\alpha_1(x_0)=-\frac{\sqrt{\pi}}{2}\,\frac{\rho}{R}\,y_0e^{y_0^2}\,\mathrm{erfc}(y_0), \tag{3.76}$$

where $\mathrm{erfc}(x)=1-\mathrm{erf}(x)$ is the complementary error function.[16] One can see that $A_{0,1}$, like A_0, is a function of y_0 and therefore has the same branch point singularity at $\lambda=0$ as the exact scattering length.

To obtain a higher QLM approximation it is convenient to remove a fourth power of variable in the denominators of the integrals in Eq. (3.75). To do this, let us introduce a new variable $y=1/x$ and a new function $\beta(x)=\alpha(1/x)$ with a boundary condition $\beta(\infty)=\alpha(0)=0$. Equation (3.72) then has the form

$$\frac{d\beta(y)}{dy}=(1+y\beta(y))^2,\quad \beta(\infty)=0, \tag{3.77}$$

whose exact solution is given by

$$\beta(y)=-\frac{1}{(1+y)}, \tag{3.78}$$

while Eq. (3.75) is written as

$$\beta_{n+1}(y)=-\int_y^\infty ds[1-s^2\beta_n^2(s)]\exp\left(2\int_s^y dt\,t(1+\beta_n(t))\right). \tag{3.79}$$

The nth QLM approximation to $\alpha(x)$ is $\alpha_n(x)=\beta_n(1/x)\equiv\beta_n(y)$ so that $\alpha(x_0)\equiv\beta(y_0)$ and the initial guess could be chosen $\beta_0(x)\equiv\beta(\infty)=0$. In the long ranged case $\rho=\infty$, $y_0=0$, and, since $\mathrm{erfc}(0)=1$, $A_{0,1}=-\sqrt{\pi}/2\sqrt{\lambda}=-0.886\sqrt{\lambda}$, a very good approximation[11] to exact value $A_0=-\sqrt{\lambda}$. The next QLM approximation, which can be computed numerically, results in $A_{0,2}=-0.988\sqrt{\lambda}$, a precision of 1.2%. For comparison, calculation of A_0 in the second CBF approxi-

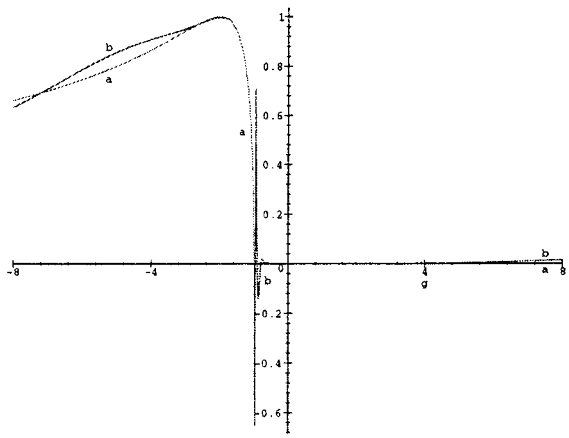

FIG. 7. Same as in Fig. 5, but in the third QLM approximation.

mation, Eq. (2.21), which in this case reduces to a simple minded linearization approach,[11] based on neglecting a nonlinear term, gives $A_{0,2}=-0.954\sqrt{\lambda}$, precision of only 4.6%. In the infinite coupling finite range case y_0 is very large and one can use asymptotic expression (3.16), which gives $A_{0,1}=-\frac{1}{2}\rho/R$, a reasonable first approximation to exact value $A_0=-\rho/R$. In the next, second QLM approximation, the numerical computation yelds $A_{0,2}=-\frac{3}{4}\rho/R$, a significant improvement.

The uniform quadratic convergence of QLM sequence (3.75) is proved, according to Eq. (2.34), by showing that the first convergence coefficient $q_1(b)$ is less than unity. In our case, we have chosen $\alpha_0(x)\equiv 0$ and therefore $q_1(b)=k(b)\|\alpha_1(x)\|\leqslant k(b)$, since in view of the properties of the complimentary error function[16] $|\alpha_1(x)|\leqslant 1$ for all x.

To estimate $k(b)$ we have to know $G(x)$ and $F(x)$, which have to satisfy inequalities (2.31). Since the first and the second derivatives of the right-hand side of Eq. (3.72) are $-2/x^4(x+\alpha(x))$ and $-2/x^4$, respectively, and since the scattering length for the repulsive potential has no poles and should be finite, $\alpha(x)\leqslant M$ where M is some positive constant, and therefore one can choose $G(x)=F(x)\equiv F_1(x)=2M/x^4$. For small $(x\leqslant\epsilon)$, where $\epsilon\ll 1$ is some small but finite number, looking for $\alpha(x)$ in the form $\alpha(x)\sim ax+bx^2$ and substituting in Eq. (3.72) gives $a=-1$, $b=1$. The first functional derivative for small x therefore equals $-2/x^2$ and one can choose in this case negative boundary function $F(x)\equiv F_2(x)=-2/x^2$ and $G(x)=|F_2(x)|$.

Separating in Eq. (2.35) for $k(b)$ smaller and larger values of x gives

$$k(b)=\frac{1}{2}\left|C\exp\int_0^{\epsilon}dt\,F_2(t)-1\right|, \tag{3.80}$$

where $C=\exp\int_{\epsilon}^{b}dt\,F_1(t)$ is a finite constant even for infinite interval $b=\infty$. Since $\exp(\int_0^{\epsilon}dtF_2(t))=\exp(-2\int_0^{b}dt\,1/t^2)=\exp(2/t\,|_{+0}^{\epsilon})=0$, Eq. (3.80) gives $k(b)\equiv\frac{1}{2}$ und thus $q_1(b)\leqslant\frac{1}{2}$, which proves uniform quadratic convergence of QLM iteration sequence $\alpha_n(x_0)$, Eq. (3.75), on the whole interval $(0,\infty)$, that is for all values of coupling constant λ, including large and infinite ones.

IV. CONCLUSION

Summing up, we have reformulated the proof of the convergence of the quasilinearization method (QLM) of Bellman and Kalaba[1,2] by removing unnecessary restrictive conditions generally not fulfilled in physical applications, and have generalized the proof for large or infinite domains of variables and for functionals which are singular at some points in the domain. We also have shown how to deal with solutions which are infinite at certain values of variable such as, for example, scattering amplitudes at values corresponding to bound state energies, etc.

In order to make presentation as simple and short as possible, we have limited ourselves here to the case of the first-order nonlinear ordinary differential equation in one variable, which physically covers the quantum mechanics of one particle in a central field (in this case the Schrödinger equation for a wave function could be rewritten as the nonlinear Ricatti equation for its logarithmic derivative) though the same modernization of the Bellman and Kalaba proof[1,2] could be provided also for a general nonlinear ordinary or partial nth order differential equations in the N-dimensional space.

In order to highlight the power of the method in Sec. III we have considered examples of different singular and nonsingular, attractive and repulsive potentials $V(r)$ for which the nonlinear first-order ordinary differential equation (1.1) for the S-wave scattering length $a_0=a(\infty)$ obtained in variable phase approach[4,5] could be solved exactly and have compared the iterations obtained by the Bellman–Kalaba linearization method with exact solutions and with the usual perturbation theory. The results could be summed up as follows.

(i) The sequence $u_n(x),\ n=1,2,\ldots$ of QLM iterations, Eq. (2.22), converges *uniformly and quadratically* to solution $u(x)$ of Eq. (2.2). For the convergence on the interval $[0,b]$ is enough that an initial guess for zero iteration is sufficiently good to ensure the smallness of just one of convergence coefficients $q_m(b)=k(b)\|u_{m+1}-u_m\|$, where $k(b)$ is given by Eq. (2.32) or Eq. (2.35). In addition, for strictly convex (concave) functionals $f(u(x),x)$ difference $u_{n+1}(x)-u_n(x)$ is strictly positive (negative), which establishes *the monotonicity* of the convergence from below (above), respectively, on this interval.

(ii) The QLM treats the nonlinear terms as a perturbation about the linear ones[1,2] and is not based, unlike perturbation theories, on the existence of some kind of small parameter. As a result, it is able to handle, unlike the perturbation theory, large or even infinite values of the coupling constant.

(iii) Comparison of QLM with the perturbation theory shows that each QLM iteration reproduces and sums many orders of perturbation theory exactly and in addition many more orders approximately. Namely, in agreement with the quadratic pattern of the convergence, the number N_n of the terms of the perturbation series, reproduced exactly in the nth QLM approximations, equals $2^{n+1}-1$, and approximately the same number of terms is reproduced approximately. The number of the exactly reproduced terms thus doubles with each subsequent QLM approximation, and reaches, for example, 127 terms in the sixth QLM approximation, 8191 terms in the twelfth QLM approximation and so on.

(iv) QLM handles without any problems not only singular potentials, like the inverse squared potential, for which the perturbation theory is divergent outside a narrow interval of the values of the coupling constant, but even super singular potentials, like reverse quartic potential, for which perturbation series are not existent at all, since their calculation leads to infinities in each order of the perturbation expansion.

(v) As we saw in all of our examples, QLM easily handles different singularities, like poles or branch point singularities, reproducing correct type of the singularity already in first iterations.

(vi) Although the analytic calculations of third and higher approximation in QLM, like in the usual perturbation theory, in most cases (excluding inverse square potential) seem impossible, the simplicity of the QLM iterational sequence Eq. (2.22) (which, unlike perturbation theory, containes no sums on intermediate energy states) assures nonproblematic numerical calculation of higher order iterations, while the extremely fast convergence of the method allows accurate estimate of the solution after relatively small number of iterations.

In view of all this, since most equations of physics, from classical mechanics to quantum field theory, are either not linear or could be transformed to a nonlinear form, the quasilinear method may turn out to be extremely useful and in many cases more advantageous than the perturbation theory or its different modifications, like expansion in inverse powers of the coupling constant, $1/N$ expansion, etc.

ACKNOWLEDGMENTS

This research was supported by the Israeli Science Foundation founded by The Israeli Academy of Sciences and Humanities. The author is indebted to Dr. A. Gal for useful discussions and constructive comments on this article.

[1] R. Kalaba, J. Math. Mech. **8**, 519 (1959).
[2] R.E. Bellman and R.E. Kalaba, *Quasilinearization and Nonlinear Boundary-Value Problems* (Elsevier, New York, 1965).
[3] V. Lakshmikantham and A.S. Vatsala, *Generalized Quasilinearization for Nonlinear Problems*, Mathematics and its Applications Vol. 440 (Kluwer, Dordrecht, 1998).
[4] F. Calogero, *Variable Phase Approach to Potential Scattering* (Academic, New York, 1965).
[5] V.V. Babikov, *Method Fazovych Funkzii v Kvantovoi Mechanike (Variable Method of Phase Functions in Quantum Mechanics)* (Nauka, Moscow, 1968).
[6] A.A. Adrianov, M.I. Ioffe, and F. Cannata, Mod. Phys. Lett. A **11**, 1417 (1996).
[7] M. Jameel, J. Phys. A **21**, 1719 (1988).
[8] K. Raghunathan and R. Vasudevan, J. Phys. A **20**, 839 (1987).
[9] M.A. Hooshyar and M. Razavy, Nuovo Cimento B **75**, 65 (1983).
[10] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, (Springer, Berlin 1937), Vols. I and II.
[11] S. Fluegge, *Practical Quantum Mechanics* (Springer, New York 1974), Vols. I and II.
[12] V. Volterra, *Theory of Functionals* (Blackie and Son, London, 1931).
[13] R.F. Dashen, Nuovo Cimento **28**, 229 (1963); J. Math. Phys. **4**, 388 (1963).
[14] V.V. Babikov, Sov. Phys. Usp. **10**, 271 (1967).
[15] E. Picard, J. de Math. **6**, 145 (1890).
[16] I.S. Gradsteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York 1994).
[17] B.W. Char *et al.*, *Maple V Library Reference Manual and Computer Program* (Springer, New York, 1991).
[18] L.D. Landau and E.M. Lifshitz, *Quantum Mechanics* (Pergamon, New York 1977).
[19] F. Calogero, Phys. Rev. **135**, B693 (1964).
[20] R.G. Newton, *Scattering Theory of Waves and Particles* (Springer, New York, 1982).

Picard–Fuchs equations and Whitham hierarchy in $N=2$ supersymmetric SU$(r+1)$ Yang–Mills theory

Yűji Ohta
Research Institute for Mathematical Sciences, Kyoto University, Sakyoku, Kyoto 606, Japan

(Received 6 July 1999; accepted for publication 4 August 1999)

In general, Whitham dynamics involves infinitely many parameters called Whitham times, but in the context of $N=2$ supersymmetric Yang–Mills theory it can be regarded as a finite system by restricting the number of Whitham times appropriately. For example, in the case of SU$(r+1)$ gauge theory without hypermultiplets, there are r Whitham times and they play an essential role in the theory. In this situation, the generating meromorphic one-form of the Whitham hierarchy on the Seiberg–Witten curve is represented by a finite linear combination of meromorphic one-forms associated with these Whitham times, but it turns out that there are various differential relations among these differentials. Since these relations can be written only in terms of the Seiberg–Witten one-form, their consistency conditions are found to give the Picard–Fuchs equations for the Seiberg–Witten periods.
 [S0022-2488(99)03411-8]

I. INTRODUCTION

Thanks to the study of electromagnetic duality initiated by Seiberg and Witten,[1] the prepotential of the low energy effective action of $N=2$ supersymmetric Yang–Mills theory was turned out to be viewed as a function on a complex projective space having singularities when the masses of charged particles vanish. This complex projective space can be identified with the moduli space of a Riemann surface determined by several physical requirements, thus the effective theory can be considered to be controlled by the geometry of moduli space of a Riemann surface.[2–16] According to this observation, since the effective coupling constants of the theory are interpreted as the period matrix of a Riemann surface, determining the period matrix from the calculation of periods becomes equivalent to evaluating effective coupling constants. It is interesting that the instanton contributions to prepotential[17] can be obtained from the evaluation of periods and the prepotentials obtained in this way[18–26] are known to be consistent with the instanton calculus.[26–33] In these studies, the method based on Picard–Fuchs equations[18–25,34–38] played a crucial role.

However, on the one hand, the theory of prepotential often shows unexpected aspects behind the effective theory. For example, it is known that the Seiberg–Witten solutions can be understood in the framework of Whitham theory.[39,40] Gorsky *et al.*[41] noticed that the Whitham dynamics in $N=2$ Yang–Mills theory could be written essentially by only a finite number of Whitham times and found that the second-order derivatives of prepotential over the Whitham times could be represented by an elliptic function associated with the Seiberg–Witten curve.

However, we can further learn more aspects of the Whitham hierarchy in gauge theory from the basic idea of Gorsky *et al.*[41] For instance, note that since the number of time variables of the hierarchy is restricted to be finite the generating meromorphic one-form of the Whitham hierarchy is represented by a finite linear combination of meromorphic one-forms associated with these Whitham times. Then we can expect that there must be closed differential relations among these meromorphic differentials associated with Whitham times. In fact, a detailed study supports this observation and the aim of the paper is to show the consequence of these relations, especially, a connection to Picard–Fuchs equations for the Seiberg–Witten periods.

The paper is organized as follows. In Sec. II, we briefly summarize the Whitham dynamics in SU$(r+1)$ gauge theory. In addition, following Gorsky *et al.*,[41] we consider the situation where the number of Whitham times is finite. Since the meromorphic one-forms on a Seiberg–Witten curve consisting of the Whitham hierarchy must always be represented by simply a linear combination of Abelian differentials, we can expect the existence of differential relations among these meromorphic one-forms. In Sec. III, it is shown that such relations can in fact be found and as a result Picard–Fuchs equations for the Seiberg–Witten periods are available from this view point. It should be noted that the generating meromorphic differential of the Whitham hierarchy can be written in terms of the Seiberg–Witten one-form. This indicates that it is sufficient to consider only the Seiberg–Witten periods in order to calculate the periods of the Whitham hierarchy. In Sec. IV, it is shown that the SU(3) Picard–Fuchs equations for the Seiberg–Witten periods can be obtained from the Picard–Fuchs equations with Whitham times for the periods of the Whitham hierarchy by considering the specialization condition to the Seiberg–Witten model. Section V is a brief summary.

II. WHITHAM HIERARCHY IN GAUGE THEORY

In this section, we briefly sketch the relation between the Seiberg–Witten solution and Whitham dynamics in the context of $N=2$ supersymmetric Yang–Mills theory.[39–41]

A. Seiberg–Witten solution

To begin with, let us recall that the Seiberg–Witten curve in SU$(r+1)$ gauge theory without matter hypermultiplets[2–4] is given by the characteristic equation

$$\det[x-L(\omega)]=0 \tag{2.1}$$

of the Lax operator $L(\omega)$ for Toda chain with $r+1$ sites,[39] where x is the eigenvalue of $L(\omega)$ and ω is the spectral parameter. (2.1) can be rewritten in the form of spectral curve

$$P(x)=\Lambda_{\mathrm{SU}(r+1)}^{r+1}\left(\omega+\frac{1}{\omega}\right), \tag{2.2}$$

where $\Lambda_{\mathrm{SU}(r+1)}$ is the dynamical mass parameter and

$$P(x):=x^{r+1}-\sum_{i=2}^{r+1} u_i x^{r+1-i} \tag{2.3}$$

represents the simple singularity of type A_r with moduli u_i. This spectral curve (2.2) can be further rewritten in the familiar hyperelliptic form[2–4]

$$y^2=P^2-4\Lambda^2, \tag{2.4}$$

where $\Lambda:=\Lambda_{\mathrm{SU}(r+1)}^{r+1}$ and we have introduced

$$y:=\Lambda_{\mathrm{SU}(r+1)}^{r+1}\left(\omega-\frac{1}{\omega}\right). \tag{2.5}$$

Note that the hyperelliptic curve (2.4) is a Riemann surface of genus r.

For a study of Riemann surface, it is often useful to consider the periods of Abelian differentials over the one-cycles on the surface. In the case at hand, we can take $2r$ one-cycles (A_i,B_i) $(i=1,\ldots,r)$ on (2.4) as a canonical basis (B_i are symplectic duals of A_i), which can be expressed by using the branching points of (2.4).

On the other hand, in order to interpret the components of a period matrix constructed from periods of Abelian differentials as the effective coupling constants, the combination of Abelian differentials must be fixed uniquely up to total derivatives. In addition, in general, there are three kinds of Abelian differentials on a Riemann surface, but that of the third kind is not required here because we are considering a pure gauge theory. Therefore, the expected meromorphic differential one-form is expressed by the Abelian differentials of the first and second kinds, and the one satisfying these requirements is called the Seiberg–Witten differential $dS_{\rm SW}$, given by

$$dS_{\rm SW} := x\frac{d\omega}{\omega} = x\frac{\partial_x P}{y}dx, \tag{2.6}$$

where we have ignored the numerical normalization for simplicity, and then the Seiberg–Witten periods are given by the loop integrals over the canonical cycles

$$a_i := \oint_{A_i} dS_{\rm SW}, \quad a_{D_i} := \oint_{B_i} dS_{\rm SW}. \tag{2.7}$$

Note that $dS_{\rm SW}$ can be viewed as the canonical one-form of the integrable system. In this way, we can see the relation between the Seiberg–Witten solution and integrable system.

B. Whitham hierarchy

We have seen that the Seiberg–Witten solution has a connection to the integrable system, but it can also be viewed as part of the Whitham theory of solitons on a Riemann surface.

To see this, let us recall that in general Whitham theory consists of the following three ingredients:[42]

(1) Riemann surface of genus g.
(2) Punctures on the surface.
(3) Existence of local coordinates near the punctures.

Gorsky *et al.*[41] noticed that the meromorphic differentials of the second kind $d\Omega_n$ of $(n+1)$th order punctures $(n>0)$ on a Riemann surface were defined up to a linear combination of g holomorphic differentials $d\omega_i$ and considered how to fix this combination by taking two basic requirements. The first one was to require

$$\oint_{A_i} d\Omega_n = 0 \tag{2.8}$$

and the second one was to introduce new meromorphic differentials $d\hat{\Omega}_n$, which enjoy the property that their differentiations over the moduli coincide with holomorphic differentials.

According to their result,[41] the differential

$$dS := \sum_{n=1}^{\infty} T_n d\hat{\Omega}_n = \sum_{i=1}^{g} \alpha_i d\omega_i + \sum_{n=1}^{\infty} T_n d\Omega_n \tag{2.9}$$

with infinitely many parameters T_n called Whitham times is found to be the expected solution which is suitable for applications to gauge theory. For this new meromorphic differential dS, the periods

$$\alpha_i := \oint_{A_i} dS, \quad \alpha_{D_i} := \oint_{B_i} dS \tag{2.10}$$

can be defined in a natural way.

Next, in order to make contact with the Seiberg–Witten solution, Gorsky *et al.*[41] regarded the Riemann surface used here as the Seiberg–Witten hyperelliptic curve (2.4).

In such a situation, they found that the Whitham hierarchy could actually be written only by first r time variables and gave an explicit expression of dS. In particular, in the case of SU$(r+1)$ gauge theory, n is restricted to $n<r+1$. Namely, in this situation, the periods (2.10) reduce to

$$\alpha_i=\sum_{n=1}^{r} T_n \oint_{A_i} d\hat{\Omega}_n, \quad \alpha_{D_i}=\sum_{n=1}^{r} T_n \oint_{B_i} d\hat{\Omega}_n \tag{2.11}$$

and $d\hat{\Omega}_n$ are given by

$$d\hat{\Omega}_n=R_n\frac{d\omega}{\omega}, \quad R_n:=P_+^{n/(r+1)}. \tag{2.12}$$

In this expression, $P_+^{n/(r+1)}$ means the non-negative terms in the expansion of $P^{n/(r+1)}$ for a large x, and in general, $P^{n/(r+1)}$ in SU$(r+1)$ gauge theory is easily found to give

$$P^{n/(r+1)}=x^n-\frac{n}{r+1}u_2x^{n-2}-\frac{n}{r+1}u_3x^{n-3}-\frac{n}{r+1}\left[u_4+\frac{u_2^2}{2}\left(1-\frac{n}{r+1}\right)\right]x^{n-4}-\cdots. \tag{2.13}$$

Note that the periods are now represented by a finite linear combination of $d\hat{\Omega}_n$ because we are only considering for the $n<r+1$ case. In addition, from (2.12), it is immediately seen that the Seiberg–Witten solution is recovered at the point

$$(T_1,T_2,T_3,\ldots,T_r)=(1,0,0,\ldots,0). \tag{2.14}$$

In fact, we find $d\hat{\Omega}_1=dS_{\mathrm{SW}}$. Of course, in this case, we have $dS=dS_{\mathrm{SW}}$.

III. PICARD–FUCHS STRUCTURE BEHIND WHITHAM HIERARCHY

A. Relations among meromorphic differentials

We have seen that dS is represented by a linear combination of $d\hat{\Omega}_n$ and also that $d\hat{\Omega}_1=dS_{\mathrm{SW}}$. Then, are $d\hat{\Omega}_n$ for $n\neq 1$ related to dS_{SW}? If we can find any relation among them, the role of the Seiberg–Witten solution in Whitham dynamics will be clarified.

To find an answer to this question, let us notice that any meromorphic differential on a Riemann surface must always be written in terms of the basis of Abelian differentials on the surface. Of course, this must be true also for $d\hat{\Omega}_n$ for all n. Therefore, if we consider a differentiation of $d\hat{\Omega}_n$ over moduli, it will ultimately be represented by a linear combination of various $d\hat{\Omega}_n$ and their derivatives. However, actually, in the case of the Seiberg–Witten Riemann surface, we can show that the derivatives of $d\hat{\Omega}_n$ for $n>1$ are obtained from the Seiberg–Witten differential $d\hat{\Omega}_1$. Thus as a result, we can conclude that dS is generated from $d\hat{\Omega}_1$ and accordingly the periods of dS can be directly determined through the Seiberg–Witten periods themselves.

To see this more concretely, let us consider the case of $d\hat{\Omega}_2$ as an example. Since the differentiations of $d\hat{\Omega}_2$ over moduli are

$$\frac{\partial d\hat{\Omega}_2}{\partial u_i}=\frac{dx}{y}\left[-2\,\delta_{2,i}\,x^r+\frac{\delta_{2,i}}{r+1}\sum_{j=2}^{r+1}(r+1-j)u_jx^{r-j}+2x^{r+2-i}\right], \tag{3.1}$$

where $\delta_{i,j}$ are Kronecker's delta symbols, and those for $d\hat{\Omega}_1$ are

$$\frac{\partial d\hat{\Omega}_1}{\partial u_i}=\frac{x^{r+1-i}}{y}dx, \tag{3.2}$$

it is easy to see that

$$\frac{\partial d\hat{\Omega}_2}{\partial u_2}=\frac{2}{r+1}\sum_{i=2}^{r+1}(r+1-i)u_i\frac{\partial d\hat{\Omega}_1}{\partial u_{i+1}},\quad \frac{\partial d\hat{\Omega}_2}{\partial u_i}=2\frac{\partial d\hat{\Omega}_1}{\partial u_{i-1}}\quad (i\neq 2). \tag{3.3}$$

Note that in the derivation of (3.1) and (3.2) we have used the general formulas

$$\frac{\partial d\hat{\Omega}_n}{\partial u_i}=\frac{dx}{y}[\partial_{u_i}R_n\cdot\partial_x P-\partial_x R_n\cdot\partial_{u_i}P]+d\left(\frac{R_n\partial_{u_i}P}{y}\right). \tag{3.4}$$

In a similar way, we can obtain differential relations between $d\hat{\Omega}_n$ for $n>1$ and $d\hat{\Omega}_1$, but we omit the derivations for them and show only the result for $n=3$ and 4 cases here.

For $d\hat{\Omega}_3$:

$$\begin{aligned}
\frac{\partial d\hat{\Omega}_3}{\partial u_2}&=-\frac{3}{r+1}\left[u_2\frac{\partial d\hat{\Omega}_1}{\partial u_2}-\sum_{i=2}^{r+1}(r+1-i)u_i\frac{\partial d\hat{\Omega}_1}{\partial u_i}\right],\\
\frac{\partial d\hat{\Omega}_3}{\partial u_3}&=-\frac{3}{r+1}\left[u_2\frac{\partial d\hat{\Omega}_1}{\partial u_3}-\sum_{i=2}^{r+1}(r+1-i)u_i\frac{\partial d\hat{\Omega}_1}{\partial u_{i+1}}\right],\\
\frac{\partial d\hat{\Omega}_3}{\partial u_i}&=3\left[\frac{\partial d\hat{\Omega}_1}{\partial u_{i-2}}-\frac{u_2}{r+1}\frac{\partial d\hat{\Omega}_1}{\partial u_i}\right]\quad (i\neq 2,3).
\end{aligned} \tag{3.5}$$

For $d\hat{\Omega}_4$:

$$\begin{aligned}
\frac{\partial d\hat{\Omega}_4}{\partial u_2}&=-\frac{4}{r+1}\left[u_3\frac{\partial d\hat{\Omega}_1}{\partial u_2}-\sum_{i=3}^{r+1}(r+1-i)u_i\frac{\partial d\hat{\Omega}_1}{\partial u_{i-1}}-\frac{r-3}{r+1}u_2\sum_{i=2}^{r+1}(r+1-i)u_i\frac{\partial d\hat{\Omega}_1}{\partial u_{i+1}}\right],\\
\frac{\partial d\hat{\Omega}_4}{\partial u_3}&=-\frac{4}{r+1}\left[2u_2\frac{\partial d\hat{\Omega}_1}{\partial u_2}+u_3\frac{\partial d\hat{\Omega}_1}{\partial u_3}-\sum_{i=2}^{r+1}(r+1-i)u_i\frac{\partial d\hat{\Omega}_1}{\partial u_i}\right],\\
\frac{\partial d\hat{\Omega}_4}{\partial u_4}&=-\frac{4}{r+1}\left[2u_2\frac{\partial d\hat{\Omega}_1}{\partial u_3}+u_3\frac{\partial d\hat{\Omega}_1}{\partial u_4}-\sum_{i=2}^{r+1}(r+1-i)u_i\frac{\partial d\hat{\Omega}_1}{\partial u_{i+1}}\right],\\
\frac{\partial d\hat{\Omega}_4}{\partial u_i}&=4\left[\frac{\partial d\hat{\Omega}_1}{\partial u_{i-3}}-\frac{2u_2}{r+1}\frac{\partial d\hat{\Omega}_1}{\partial u_{i-1}}-\frac{u_3}{r+1}\frac{\partial d\hat{\Omega}_1}{\partial u_i}\right]\quad (i\neq 2,3,4).
\end{aligned} \tag{3.6}$$

B. Picard–Fuchs equations from Whitham hierarchy

If the derivatives of $d\hat{\Omega}_n$ over moduli for $n>1$ are eliminated from relations (3.3), (3.5), and (3.6) by using differentiations, the equations satisfied by $d\hat{\Omega}_1$ will be obtained. Furthermore, since $d\hat{\Omega}_1=dS_{\mathrm{SW}}$, we can identify such equations as Picard–Fuchs equations for Seiberg–Witten periods.

To see this, it is enough to consider the cross derivatives of $d\hat{\Omega}_n$. For example, for $d\hat{\Omega}_2$, from $[\partial_2\partial_i-\partial_i\partial_2]d\hat{\Omega}_2=0$, where $\partial_i:=\partial/\partial u_i$, we get

$$\left[(r+1)\partial_2\partial_{i-1}-(r+1-i)\partial_{i+1}-\sum_{j=2}^{r+1}(r+1-j)u_j\partial_i\partial_{j+1}\right]d\hat{\Omega}_1=0 \quad (i\neq 2), \tag{3.7}$$

which are the Picard–Fuchs equations obtained by several authors.[26,37] For other $d\hat{\Omega}_n$, we can construct similar equations and, in fact, we can obtain the ''hierarchy'' of Picard–Fuchs equations as follows:

$$\left[\sum_{i=2}^{r+1}(r+1-i)u_i(\partial_3\partial_i-\partial_2\partial_{i+1})\right]d\hat{\Omega}_1=0,$$

$$\left[(r+1-i)\partial_{i+1}-(r+1)\partial_3\partial_{i-2}+\sum_{j=2}^{r+1}(r+1-j)u_j\partial_i\partial_{j+1}\right]d\hat{\Omega}_1=0 \quad (i\neq 2,3),$$

$$\left[(r+1)\partial_2\partial_{i-2}-(r+2-i)\partial_i-\sum_{j=2}^{r+1}(r+1-j)u_j\partial_i\partial_j\right]d\hat{\Omega}_1=0 \quad (i\neq 2,3),$$

$$\Bigg[2(r+1)u_2\partial_2^2+(r-3)(r-2)u_2\partial_4+(r+1)\sum_{i=3}^{r+1}(r+1-i)u_i\partial_{i-1}\partial_3+(r-3)u_2 \times\sum_{i=2}^{r+1}(r+1-i)u_i\partial_{i+1}\partial_3-(r+1)\sum_{i=2}^{r+1}(r+1-i)u_i\partial_i\partial_2\Bigg]d\hat{\Omega}_1=0,$$

$$\Bigg[2(r+1)u_2\partial_2\partial_3+(r-3)^2u_2\partial_5+(r+1)\sum_{i=3}^{r+1}(r+1-i)u_i\partial_{i-1}\partial_4-(r+1) \times\sum_{i=2}^{r+1}(r+1-i)u_i\partial_{i+1}\partial_2+(r-3)u_2\sum_{i=2}^{r+1}(r+1-i)u_i\partial_{i+1}\partial_4\Bigg]d\hat{\Omega}_1=0, \tag{3.8}$$

$$\Bigg[(r+1)^2\partial_2\partial_{i-3}-2(r+1)u_2\partial_{i-1}\partial_2-(r+1)(r+3-i)\partial_{i-1}-(r-3)(r+1-i)u_2\partial_{i+1} -(r+1)\sum_{j=3}^{r+1}(r+1-j)u_j\partial_i\partial_{j-1}-(r-3)u_2\sum_{j=2}^{r+1}(r+1-j)u_j\partial_i\partial_{j+1}\Bigg]d\hat{\Omega}_1=0 \quad (i\neq 2,3,4),$$

$$\left[2u_2(\partial_2\partial_4-\partial_3^2)-\sum_{i=2}^{r+1}(r+1-i)u_i(\partial_4\partial_i-\partial_3\partial_{i+1})\right]d\hat{\Omega}_1=0,$$

$$\left[2u_2(\partial_3\partial_i-\partial_4\partial_{i-1})+(r+1)\partial_4\partial_{i-3}-(r+1-i)\partial_{i+1}-\sum_{j=2}^{r+1}(r+1-j)u_j\partial_i\partial_{j+1}\right]d\hat{\Omega}_1=0 \quad (i\neq 2,3,4),$$

$$\left[2u_2(\partial_i\partial_2-\partial_3\partial_{i-1})+(r+1)\partial_3\partial_{i-3}-(r+2-i)\partial_i-\sum_{j=2}^{r+1}(r+1-j)u_j\partial_i\partial_j\right]d\hat{\Omega}_1=0 \quad (i\neq 2,3,4).$$

Note that the equations in (3.8) are all second-order equations and in some cases we can simplify them by using $(\partial_i\partial_j-\partial_p\partial_q)d\hat{\Omega}_1=0$, where $i+j=p+q$.[26,37]

C. Picard–Fuchs equations as a complete system

Of course, as a complete Picard–Fuchs system, it is not necessary to consider all equations in (3.7) and (3.8). In general, since there are r moduli parameters in the $SU(r+1)$ gauge theory, it is sufficient to extract at least r independent equations from them.

To see this, let us notice the equations in (3.7). Since the number of the equations is $r-1$, one more equation is necessary. However, we cannot obtain the expected equation from (3.8) because the equations presented there do not have the instanton corrections. If the instanton correction terms are not included in any one of Picard–Fuchs equations, the prepotential obtained from them will not show the instanton corrections precisely. Therefore, we require that the remaining one must include instanton terms.

Actually, such an equation was recognized by Ito and Yang[43] as the scaling relation. There, the Picard–Fuchs system was realized by two kinds of equations, one of which is the Gauss–Manin system and the other is the scaling relation. Since the Gauss–Manin system does not involve instanton corrections, the situation looks like ours. Therefore, also for our case, the scaling relation may be used as the remaining Picard–Fuchs equation.

For this, let us consider the Eulerian operator

$$\mathcal{E}:=\sum_{i=2}^{r+1} iu_i\partial_i+(r+1)\Lambda\partial_\Lambda, \tag{3.9}$$

which acts as

$$\mathcal{E}d\hat{\Omega}_n=nd\hat{\Omega}_n \tag{3.10}$$

for all $n>0$. (3.10) indicates that the degree of $d\hat{\Omega}_n$ is n. Realizing (3.10) as an equation only in terms of moduli derivatives can be easily accomplished by considering the squaring equation $(\mathcal{E}-n)^2d\hat{\Omega}_n=0$.[37,43]

In this way, we can associate r independent Picard–Fuchs equations for $d\hat{\Omega}_1$.

IV. PICARD–FUCHS EQUATIONS WITH WHITHAM TIMES

A. The SU(3) Picard–Fuchs equations

Next, let us consider Picard–Fuchs equations for the periods (α_i,α_{D_i}) of the Whitham hierarchy. In the case of $r=1$, the resulting Picard–Fuchs equation takes the same form with the usual one[18] up to rescaling of T_1. For this reason, we do not discuss this case, and instead, let us consider the $r=2$ case in order to find a nontrivial example of Picard–Fuchs equations with Whitham times.

In this case, the Picard–Fuchs equations with the Whitham times are found to be in the form

$$\mathcal{L}_j(\alpha_i,\alpha_{D_i})=0, \tag{4.1}$$

where

$$
\begin{aligned}
\mathcal{L}_1=&\frac{54u^2vT_1T_2-(108\Lambda^2-4u^3-27v^2)(uT_2^2-3T_1^2)}{u(-3T_1^2+4uT_2^2)}\partial_u^2\\
&-\frac{3[(-108\Lambda^2+4u^3+27v^2)T_1T_2+8uv(uT_2^2-3T_1^2)]}{2(3T_1^2-4uT_2^2)}\partial_u\partial_v\\
&+\frac{T_2[8(108\Lambda^2-4u^3-27v^2)(-3T_1^2+uT_2^2)T_2-9uvT_1(15T_1^2+28uT_2^2)]}{2u(3T_1^2-4uT_2^2)^2}\partial_u\\
&+\frac{2(108\Lambda^2-4u^3-27v^2)(3T_1^2-uT_2^2)T_1T_2+3uv(9T_1^4+27uT_1^2T_2^2-4u^2T_2^4)}{u(3T_1^2-4uT_2^2)^2}\partial_v+1,
\end{aligned}
\tag{4.2}
$$

$$
\begin{aligned}
\mathcal{L}_2=&\frac{-54u^2vT_1T_2+(108\Lambda^2-4u^3-27v^2)(uT_2^2-3T_1^2)}{3(3T_1^2-4uT_2^2)}\partial_v^2\\
&-\frac{3[8uv(uT_2^2-3T_1^2)+(-108\Lambda^2+4u^3+27v^2)T_1T_2]}{2(3T_1^2-4uT_2^2)}\partial_u\partial_v\\
&-\frac{45vT_1T_2}{2(3T_1^2-4uT_2^2)}\partial_u+\frac{3v(3T_1^2+uT_2^2)}{3T_1^2-4uT_2^2}\partial_v+1.
\end{aligned}
$$

Though the derivation of Picard–Fuchs equations for other higher r is straightforward, the result turns out to be too lengthy and complicated, so we do not consider these cases in this paper.

B. Specializations of SU(3) Picard–Fuchs equations

It may be instructive to see specializations of (4.2). With the help of (2.14), it is straightforward to make sure that the equations in (4.2) yield the usual SU(3) Picard–Fuchs equations[18] $\mathcal{L}_j(a_i,a_{D_i})=0$ for the Seiberg–Witten periods, which can be identified with Appell's F_4 system,[44–47]

$$
\begin{aligned}
&\mathcal{L}_1\to L_1:=(4u^3+27v^2-108\Lambda^2)\partial_u^2+12u^2v\,\partial_u\partial_v+3uv\,\partial_v+u,\\
&\mathcal{L}_2\to L_2:=(4u^3+27v^2-108\Lambda^2)\partial_v^2+36uv\,\partial_u\partial_v+9v\,\partial_v+3.
\end{aligned}
\tag{4.3}
$$

Note that the consistency condition of (4.3) leads to $[u\partial_v^2-3\partial_u^2](a_i,a_{D_i})=0$, which coincides with (3.7) for $r=2$.

On the other hand, from (4.2) with $(T_1,T_2)=(0,1)$, we can also consider Picard–Fuchs equations $\hat{L}_j(\oint_{A_i}d\hat{\Omega}_2,\oint_{B_i}d\hat{\Omega}_2)=0$ for the periods of $d\hat{\Omega}_2$, where

$$
\begin{aligned}
&\mathcal{L}_1\to\hat{L}_1:=u(4u^3+27v^2-108\Lambda^2)\partial_u^2+12u^3v\,\partial_u\partial_v-(4u^3+27v^2-108\Lambda^2)\partial_u-3u^2v\,\partial_v+4u^2,\\
&\mathcal{L}_2\to\hat{L}_2:=(4u^3+27v^2-108\Lambda^2)\partial_v^2+36uv\,\partial_u\partial_v-9v\,\partial_v+12.
\end{aligned}
\tag{4.4}
$$

From (4.4), we can obtain a relation like that from (4.3), but the same equation is also available from (3.3), provided $\partial d\hat{\Omega}_1/\partial u_i$ are eliminated from (3.3).

Finally, note that we have

$$\hat{L}_j(\alpha_i,\alpha_{D_i})=T_1\hat{L}_j\left(\oint_{A_i}d\hat{\Omega}_1,\ \oint_{B_i}d\hat{\Omega}_1\right),\quad L_j(\alpha_i,\alpha_{D_i})=T_2L_j\left(\oint_{A_i}d\hat{\Omega}_2,\ \oint_{B_i}d\hat{\Omega}_2\right)\qquad(4.5)$$

from (2.9), (4.3), and (4.4).

V. SUMMARY

In this paper, we have discussed the SU$(r+1)$ gauge theory in the standpoint of Whitham dynamics and realized $r-1$ Picard–Fuchs equations for Seiberg–Witten periods as consistency equations among meromorphic differentials associated with Whitham times. In addition, we have used the scaling relation as the remaining independent equation in order to include the instanton corrections. Though the generalization to other cases except SU$(r+1)$ group is straightforward, the case of exceptional gauge groups would be interesting because there are two types of Seiberg–Witten curves in these gauge theories.[7,11–16] In particular, it may be interesting to know how the differences of physics expected from these two curves[12,16,24,25] are reflected in the Whitham theory and the Picard–Fuchs structure behind it.

Of course, our construction of Picard–Fuchs equations may provide helpful information not only for these cases but also when we consider the relation among flat coordinates,[48,49] Witten–Dijkgraaf–Verlinde–Verlinde equations,[50–55] and the Whitham hierarchy.[39–41] We are now planning a discussion with respect to this point.

[1] N. Seiberg and E. Witten, Nucl. Phys. B **431**, 484 (1994); **435**, 129 (1994).
[2] A. Klemm, W. Lerche, S. Yankielowicz, and S. Theisen, Phys. Lett. B **344**, 169 (1995).
[3] P. C. Argyres and A. E. Faraggi, Phys. Rev. Lett. **74**, 3931 (1995).
[4] A. Hanany and Y. Oz, Nucl. Phys. B **452**, 283 (1995).
[5] J. A. Minahan and D. Nemeschansky, Nucl. Phys. B **464**, 3 (1996).
[6] U. H. Danielsson and B. Sundborg, Phys. Lett. B **358**, 273 (1995).
[7] U. H. Danielsson and B. Sundborg, Phys. Lett. B **370**, 83 (1996).
[8] A. Hanany, Nucl. Phys. B **466**, 85 (1996).
[9] A. Brandhuber and K. Landsteiner, Phys. Lett. B **358**, 73 (1995).
[10] P. C. Argyres, M. R. Plesser, and A. D. Shapere, Phys. Rev. Lett. **75**, 1699 (1995).
[11] M. Alishahiha, F. Ardalan, and F. Mansouri, Phys. Lett. B **381**, 446 (1996).
[12] K. Landsteiner, J. M. Pierre, and S. B. Giddings, Phys. Rev. D **55**, 2367 (1997).
[13] M. R. Abolhasani, M. Alishahiha, and A. M. Ghezelbash, Nucl. Phys. B **480**, 279 (1996).
[14] E. J. Martinec and N. P. Warner, Nucl. Phys. B **459**, 97 (1996).
[15] W. Lerche and N. P. Warner, Phys. Lett. B **423**, 79 (1998).
[16] T. Eguchi and S.-K. Yang, Phys. Lett. B **394**, 315 (1997).
[17] N. Seiberg, Phys. Lett. B **206**, 75 (1988).
[18] A. Klemm, W. Lerche, and S. Theisen, Int. J. Mod. Phys. A **11**, 1929 (1996).
[19] K. Ito and S.-K. Yang, Phys. Lett. B **366**, 165 (1996).
[20] Y. Ohta, J. Math. Phys. **37**, 6074 (1996); **38**, 682 (1997).
[21] M. Matone, Phys. Lett. B **357**, 342 (1995).
[22] H. Ewen, K. Förger, and S. Theisen, Nucl. Phys. B **485**, 63 (1997).
[23] H. Ewen and K. Förger, Int. J. Mod. Phys. A **12**, 4725 (1997).
[24] K. Ito, Phys. Lett. B **406**, 54 (1997).
[25] A. M. Ghezelbash, Phys. Lett. B **423**, 87 (1998).
[26] K. Ito and N. Sasakura, Nucl. Phys. B **484**, 141 (1997).
[27] D. Finnell and P. Pouliot, Nucl. Phys. B **453**, 225 (1995).
[28] K. Ito and N. Sasakura, Phys. Lett. B **382**, 95 (1996).
[29] M. J. Slater, Phys. Lett. B **403**, 57 (1997).
[30] N. Dorey, V. V. Khoze, and M. P. Mattis, Phys. Lett. B **388**, 324 (1996); Phys. Rev. D **54**, 2921 (1996); **54**, 7832 (1996).
[31] T. Harano and M. Sato, Nucl. Phys. B **484**, 167 (1997).
[32] H. Aoyama, T. Harano, M. Sato, and S. Wada, Phys. Lett. B **388**, 331 (1996).
[33] A. Yung, Nucl. Phys. B **485**, 38 (1997).
[34] S. Ryang, Phys. Lett. B **365**, 113 (1996).
[35] A. M. Ghezelbash, A. Shafiekhani, and M. R. Abolhasani, Mod. Phys. Lett. A **13**, 527 (1998).
[36] J. M. Isidro, A. Mukherjee, J. P. Nunes, and H. J. Schnitzer, Nucl. Phys. B **492**, 647 (1997); **502**, 363 (1997); Int. J. Mod. Phys. A **13**, 233 (1998).
[37] M. Alishahiha, Phys. Lett. B **398**, 100 (1997); **418**, 317 (1998).
[38] Y. Ohta, J. Math. Phys. **40**, 3211 (1999).
[39] A. Gorsky, I. Krichever, A. Marshakov, A. Mironov, and A. Morozov, Phys. Lett. B **355**, 466 (1995).
[40] T. Nakatsu and K. Takasaki, Mod. Phys. Lett. A **11**, 157 (1996).

[41] A. Gorsky, A. Marshakov, A. Mironov, and A. Morozov, Nucl. Phys. B **527**, 690 (1998).
[42] I. M. Krichever, Commun. Math. Phys. **143**, 415 (1992).
[43] K. Ito and S.-K. Yang, Phys. Lett. B **415**, 45 (1997).
[44] A. Erdélyi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1966).
[45] T. Takano, Funkcial. Ekvac. **23**, 97 (1980).
[46] M. Kato, Funkcial. Ekvac. **38**, 243 (1995).
[47] J. Kaneko, Tokyo J. Math. **4**, 35 (1981).
[48] K. Ito, C.-S. Xiong, and S.-K. Yang, Phys. Lett. B **441**, 155 (1998).
[49] K. Ito and S.-K. Yang, Int. J. Mod. Phys. A **13**, 5373 (1998).
[50] E. Witten, Nucl. Phys. B **340**, 281 (1990).
[51] R. Dijkgraaf and E. Witten, Nucl. Phys. B **342**, 486 (1990).
[52] R. Dijkgraaf, H. Verlinde, and E. Verlinde, Nucl. Phys. B **352**, 59 (1991).
[53] A. Marshakov, A. Mironov, and A. Morozov, Phys. Lett. B **389**, 43 (1996); Mod. Phys. Lett. A **12**, 773 (1997).
[54] A. Marshakov, A. Mironov, and A. Morozov, ''More evidence for the WDVV equations in $N=2$ SUSY Yang-Mills theories,'' hep-th/9701123.
[55] K. Ito and S.-K. Yang, Phys. Lett. B **433**, 56 (1998).

Wigner–Weyl correspondence and semiclassical quantization in spherical coordinates

Bill Poirier[a)]
The James Franck Institute, The University of Chicago, Chicago, Illinois 60637

(Received 20 January 1999; accepted for publication 7 September 1999)

The Wigner–Weyl quantum-to-classical correspondence rule is nonunique with respect to coordinate choice. This ambiguity can be exploited to improve the accuracy of semiclassical approximations. For instance, the well-known Langer modification was recently derived by applying a coordinate transformation to the radial Schrödinger equation prior to using the Wigner–Weyl rule—albeit only by presuming exact quantum solutions for all nonradial degrees of freedom [J. J. Morehead, J. Math. Phys. **36**, 5431 (1995)]. In this paper, the full classical Hamiltonian is derived in all degrees of freedom, using a (hyper)spherical coordinate Wigner–Weyl correspondence with a Langer-like modification of polar angles. For central force Hamiltonians, the new result is radially equivalent to that of Langer, and to the standard Cartesian form. The new correspondence is superior with respect to all angular momentum operators however, in that the resultant semiclassical eigenvalues are exact—a desirable goal, evidently achieved here for the first time. © *1999 American Institute of Physics.* [S0022-2488(99)01912-X]

I. INTRODUCTION

When viewed as mathematical structures, classical and quantum mechanics are in many respects very similar. In both cases, physical observables may be regarded as elements of a unital Lie $*$-algebra.[1] Moreover, in each case there is an important class of algebra automorphisms—the canonical (unitary) transformations—which induces an equivalence relation on the algebraic elements. Physically, this reflects the fact that canonically (unitarily) equivalent observables are indistinguishable from one another—these observables can even be regarded as characterizing ''the same'' physical system, if the transformation is interpreted as a passive change of coordinates.

Despite similarities, though, the classical and quantum algebras are *not* isomorphic. Thus, although a one-to-one correspondence can be established between the two, it is impossible to preserve the equivalence relation described above. The upshot is that a given quantum operator $\hat{A}$ can be mapped to physically distinct classical observables $A,A',A'',\ldots,$ depending on the choice of coordinates. This is a basic theoretical fact, which has thus far received surprisingly little attention; the issue is addressed in some detail in Sec. II. Indeed, even the correspondence rule itself[2–7]—discovered independently by Weyl[2] and Wigner[3] in the early days of the quantum theory—is usually omitted from the standard texts.

In any event, this paper concerns itself with semiclassical approximations, and how they are affected by the choice of correspondence coordinates. All semiclassical methods are canonically invariant in the following sense: Once the classical observable has been established, quantities such as Wentzel–Kramers–Brillouin (WKB) eigenvalues are the same in all canonical coordinate systems. However, because the Wigner–Weyl (WW) rule itself is *non*invariant, these quantities do vary with the coordinates used to obtain the correspondence. It follows that in this sense, some coordinate systems are better than others, vis-à-vis the accuracy of the resultant WKB approximations. It is the author's contention that a basic understanding of these preferred coordinate

a)Electronic mail: billp@rainbow.uchicago.edu

systems is a worthwhile pursuit—with respect to improving semiclassical calculations, and perhaps also in terms of providing physical insight.

The central force Hamiltonians $H = T + V(r)$ (Sec. III) and their generalizations with respect to dimensionality and form (respectively, Secs. IV and V) constitute the particular focus of the present work. These systems have always occupied a special role in physics, in part because of their prevalence in nature, but also because they lend themselves readily to analytical solution. It is well-known that the three-dimensional central force problem decouples, if spherical—rather than Cartesian—coordinates are used. The former are closely related to the action-angle variables, and are therefore more fundamentally associated with the Hamiltonian than the latter—despite the simplicity of the Cartesian expressions.

If the theories of mechanics admit any favoritism at all with respect to coordinates, then surely the action-angle variables are the preferred choice. It is therefore natural to ask whether proximity to action-angle coordinates is related to the efficacy of the resultant WW correspondence. This paper attempts to provide at least a partial answer to this question. More specifically, we determine whether the WKB approximation, as applied to the classical central force Hamiltonians obtained from a correspondence in spherical coordinates, is better or worse than that obtained from the standard correspondence in Cartesian coordinates.

The question of choosing correspondence coordinates appropriate for central force Hamiltonians has been addressed previously, albeit in a context of reduced dimensionality. If the quantum Hamiltonian is expressed in spherical coordinates, and the angular momentum operator $\hat{L}^2$ replaced with its exact quantum eigenvalues, then the one-dimensional radial Schrödinger equation results. This equation is often tackled semiclassically; in this context, it is well known that the results are improved dramatically under the ''Langer modification''[8–11]—which amounts to adding an extra $+1/(8mr^2)$ term to the classical Hamiltonian (in units where $\hbar = 1$, as will be used throughout this paper).

Over the years, many justifications for the introduction of Langer's extra term have been provided—beginning with that of Langer himself,[8] who believed it necessary to retain a differential equation with a Schrödinger-like form. In more recent years, the intrinsic Gaussian curvature of the constrained configuration space[12–16] has come to be regarded as the essential rationale. These treatments do not focus on the issue of correspondence coordinates; implicitly however, they presume a WW correspondence in the radial coordinate r. On the other hand, if the WW correspondence is obtained in the coordinate $x = \ln r$ rather than in r itself, then the ''extra'' term is present right from the start.

Evidently, this line of reasoning was first proposed in a paper by Morehead,[17] which shall be referred to extensively in this work. It should be noted that Langer himself made use of the same logarithmic coordinate transformation as Morehead, but not in any canonically meaningful way. Langer's procedure is related to the Podolsky transformation,[18] which can be regarded as a special case of the WW correspondence rule—though it is doubtful that Langer was aware of this. Unlike the WW rule however, the Podolsky procedure is inherently limited to observables that are quadratic in the momenta. Consequently, Langer's original derivation applies only to $T + V(r)$ Hamiltonians, whereas Morehead's approach allows the modification to be generalized for arbitrary radial operators.

The approach of this paper, in turn, can be considered a generalization of Morehead's work. The notion of the coordinate dependence of the WW correspondence rule is used throughout, as is Langer's modification—whose appropriate extension to polar angles is investigated herein. The present method is a fully WKB treatment however—i.e., one in which all of the degrees of freedom are treated semiclassically. In contrast, Morehead uses a mixed method, for which all but the radial degree of freedom are treated quantum mechanically. The mixed method is therefore only applicable to radial Hamiltonians, for which the two approaches will be shown to be essentially identical. However, the present method can also be applied to quantities that depend on the spherical angles, such as angular momenta. Indeed, WKB results for $\hat{L}^2$ and its N-dimensional generalizations shall be presented, which reproduce the quantum eigenvalues and degeneracies *exactly*. To the author's best knowledge, there is no other semiclassical treatment in the current literature that also does so.

II. THEORETICAL BACKGROUND

In this section, we review the Wigner–Weyl (WW) correspondence rule, the Podolsky kinetic energy transformation, Morehead's derivation of the Langer modification, and other theoretical aspects pertinent to the subsequent discussion. We begin with the basic task of specifying the exact correspondence between quantum operators and classical observables, for one-dimensional systems.

This fundamental problem has been with us since the inception of quantum mechanics. The conventional wisdom states that (1) there is no difficulty for kinetic-plus-potential ($T+V$) Hamiltonians, and (2) the more general case with position-momentum cross terms is an inherently many-to-one correspondence—presumably because multiplicative ordering matters in quantum mechanics, but not in classical mechanics. This seemingly innocuous description is intuitively sensible, and can even be found in standard texts.[19] Nevertheless, such a viewpoint is essentially *incorrect on both points*.

The cross terms dilemma was effectively resolved with the advent of the correspondence rules of Weyl and Wigner. Weyl's approach is to allow just one possible ordering of $\hat{q}$ and $\hat{p}$ operators to correspond to each classical monomial $q^m p^n$,[20] where q and p are canonically conjugate position and momentum variables. Quantum operators which are properly ''Weyl-ordered'' thus correspond to classical observables—the ''Weyl symbols''—in a manifestly unique manner.[2,6] Wigner's method, on the other hand, converts the position matrix representation of the operator to a mixed position-momentum representation, using a Fourier-like transformation.[3,5] That the end result of the Wigner prescription is identical to the Weyl symbol—if (q,p) are true canonical coordinates, and for legitimate polynomial expansions—was shown by Moyal.[4]

The WW rule certainly provides a one-to-one correspondence between quantum operators and classical Weyl symbols, and would appear to remove all arbitrariness from the correspondence procedure. As mentioned in Sec. I, however, a more subtle ambiguity still remains with respect to the property of canonical (unitary) invariance. It is in this sense that the WW correspondence fails to be unique; for one could always choose to express the initial Hamiltonian in a different coordinate system, thereby obtaining a canonically inequivalent correspondence. Moreover, *no* Hamiltonians are safe from this kind of ambiguity—not even those of the standard $T+V$ form. The only class of observables whose WW correspondence is always canonically invariant are the canonical coordinates themselves.

It is helpful to express these ideas in more precise mathematical language. Let $\hat{A}=A(\hat{q},\hat{p})$ be a Hermitian operator, expressed in terms of $\hat{q}$ and $\hat{p}$, the generators of the quantum algebra. These generators are presumed to be a conjugate pair of position and momentum operators, satisfying the canonical commutation relation

$$\hat{q}\hat{p}-\hat{p}\hat{q}=i\hat{I}. \tag{2.1}$$

Because the algebraic relation of Eq. (2.1) holds, there can be many different algebraic expressions $A(\hat{q},\hat{p})$ that result in the same operator $\hat{A}$. For convenience, we choose $A(\hat{q},\hat{p})$ to be Weyl-ordered. Thus, if ''$W[\ ]$'' denotes the WW mapping from quantum operators to classical Weyl symbols, then $W[\hat{A}]=A(q,p)$. The real-valued classical function $A(q,p)$ is the Weyl symbol corresponding to $\hat{A}$, *as obtained in the coordinates* $(\hat{q},\hat{p})$.

Let us consider a unitary transformation $\hat{U}$ that has the following effect on the generators:

$$\hat{U}^\dagger\hat{q}\hat{U}=\hat{Q}=Q(\hat{q},\hat{p}), \tag{2.2}$$

$$\hat{U}^\dagger\hat{p}\hat{U}=\hat{P}=P(\hat{q},\hat{p}), \tag{2.3}$$

where the algebraic expressions $Q(\hat{q},\hat{p})$, etc., are once again presumed to be Weyl-ordered. The new operators $\hat{Q}$ and $\hat{P}$ satisfy Eq. (2.1), and are therefore canonically conjugate in the quantum sense. In addition, it can be shown that the corresponding Weyl symbols $Q(q,p)$ and $P(q,p)$ satisfy the classical Poisson bracket—thereby comprising a canonically conjugate pair in the classical sense as well. We can therefore establish a one-to-one correspondence between quantum unitary transformations $\hat{U}$, and classical canonical transformations—designated $\mathcal{C}_U$. It can be shown, moreover, that this correspondence preserves the group properties of the transformations, so that $\hat{U}^{-1}$ corresponds to $\mathcal{C}_U^{-1}$, etc.

Under the passive unitary transformation $\hat{U}^{-1}$, the algebraic expression for $\hat{A}$ becomes

$$\hat{A}=A(q(\hat{Q},\hat{P}),p(\hat{Q},\hat{P})). \tag{2.4}$$

Under the corresponding canonical transformation $\mathcal{C}_U^{-1}$, the Weyl symbol $A(q,p)$ becomes $A(q(Q,P),p(Q,P))$. However, the right-hand side of Eq. (2.4) is *not*, in general, Weyl-ordered in $\hat{Q}$ and $\hat{P}$. Therefore,

$$A(q(Q,P),p(Q,P))\neq A'(Q,P), \tag{2.5}$$

where $A'(Q,P)$ is the Weyl symbol for $\hat{A}$ that actually is obtained in the coordinates $(\hat{Q},\hat{P})$. Therefore, the Weyl symbols $A(q,p)$ and $A'(Q,P)$ are not in general canonically equivalent.

In the preceding analysis, and throughout this paper, we rely on the notion of ''true canonical coordinates,'' which can be defined algebraically as the set of Hermitian elements $\hat{q}$ for which there exists a (conjugate) Hermitian element $\hat{p}$ such that Eq. (2.1) is satisfied. Not all Hermitian operators fit the bill—implying that the simple practice of assigning $\hat{p}\rightarrow -i\partial/\partial q$ is not always valid. For example, the spectrum of a true canonical coordinate must be Cartesian—i.e., nondegenerate, real, and ranging continuously from $-\infty$ to $+\infty$. This follows from Eq. (2.1), and from the fact that the translation operator $\exp(iq_0\hat{p})$ is unitary for all real q_0. In this strict sense, therefore, all bounded coordinates such as radii and angles, are not considered truly canonical.

A similar situation also holds in the classical realm. Consider the action-angle transformation

$$\phi=-\arctan(p/q), \quad J=(q^2+p^2)/2. \tag{2.6}$$

The coordinates (ϕ,J) are *almost* truly canonical, but not quite—because Eq. (2.6) is ill-defined at the origin (although the classical Poisson bracket is satisfied everywhere else in phase space). This disqualifies ϕ and J from being true canonical coordinates, although they are obviously still useful; we shall refer to such coordinates as being ''canonical-like.''

This distinction is relevant for the Hamiltonians of this paper, for which ''hyperspherical coordinates''—the N-dimensional generalization of (r,θ,ϕ)—shall be employed. These coordinates—also called ''ultraspherical''[21–23]—are significant, because the Hamiltonian and other constants of the motion become weakly separable when expressed in hyperspherical form (Appendix). Weakly separable basis representations have previously been investigated for other quantum applications.[24–26] Among other advantages, such representations necessarily involve product—though not necessarily direct product—basis sets.

Hyperspherical coordinates are not true canonical coordinates, because the transformation from the original Cartesian system is ill-defined at the origin. The standard WW rule does not apply in such cases; therefore there is some question as to how to establish a correspondence. Our provisional resolution shall be to retain Weyl ordering in the canonical-like conjugate pairs such as $(\hat{\phi},\hat{p}_\phi)$, with the usual conjugate momenta assignments, i.e., $(N=3)$,

$$\hat{p}_r\rightarrow -i\frac{\partial}{\partial r}, \quad \hat{p}_\theta\rightarrow -i\frac{\partial}{\partial \theta}, \quad \hat{p}_\phi\rightarrow -i\frac{\partial}{\partial \phi}. \tag{2.7}$$

There is empirical evidence suggesting that this rule is appropriate, at least for localized operators, i.e., those that depend polynomially on the hyperspherical momenta (Sec. V).[17,27] We presume this to be the case for the Hamiltonians considered in this paper.

There is another, less fundamental context in which Eq. (2.7) may not hold—even for true canonical coordinates. In general, differential quantum operator assignments are meaningless, unless the associated Jacobian weight factor that appears in the orthonormality relations is also specified. In particular, Eq. (2.7) applies only if the Jacobian weight factor is unity.[18] We shall refer to differential operator expressions as being in ''canonical(-like) form'' if the associated weight factor is unity over all of configuration space (except for subspaces of measure zero). Although the standard Cartesian differential expressions are almost always in canonical form, nontrivial weight factors can accrue, for instance, when the chain rule is used to effect a coordinate transformation.

The chain rule is usually invoked in the context of point transformations, i.e., canonical transformations for which the new positions are functions of the old positions only. Although the chain rule is perfectly valid for this, the Podolsky transformation—where applicable—is usually preferred, because it automatically preserves the canonical-like form of the differential expressions. It is for this reason that the Podolsky transformation is closely related to the WW correspondence in the new coordinates. In fact, by Weyl-ordering the Podolsky-transformed kinetic energy, and then transforming classically back to the original coordinates, one obtains the original Weyl symbol plus an additional ''quantum correction'' term. Quantum corrections are mass- and $\hbar$-dependent functions of position, that appear in the classical expressions when cross terms are introduced by the coordinate transformation.[28,29] They can be interpreted as a measure of the lack of canonical invariance associated with the WW correspondence for a particular observable.

It is as a quantum correction term, for instance, that the Langer modification arises in the point transformation of the one-dimensional radial Schrödinger equation, from the coordinate r to the coordinate $x=\ln r$. This is, at any rate, the essence of Morehead's argument.[17] His analysis does not tell us, however, why this *particular* coordinate should be singled out for the correspondence. Given that any other choice would result in a different classical Hamiltonian, it remains a mystery as to why x happens to be so beneficial for the subsequent semiclassical analysis (designated ''method I''). Most likely, this question cannot be properly addressed unless a fully WKB treatment is used. The method of this paper does in fact constitute such a treatment; accordingly, some tentative arguments are presented in Sec. V—but the question is still very much an open one.

As it happens, Langer's classical Hamiltonian has also been obtained using an alternate, fully WKB method. In this second approach—which does *not* employ any modification—the WW correspondence is obtained in the original Cartesian coordinates, and then solved semiclassically using Einstein–Brillouin–Keller (EBK) quantization. This second approach shall be referred to as ''method II,'' in contradistinction to the mixed, spherical coordinate method I of the preceding paragraph. Evidently, there is no improvement to be gained from a Langer-like modification, if Cartesian coordinates are used to begin with. This suggests that method I's success may be related to the fact that $x=\ln r$ is more Cartesian-like than r (Sec. V)—an idea that will motivate a similar transformation of the polar angle coordinates in Sec. III and IV.

Given that the Cartesian correspondence (method II) already yields Langer's result for the $N=3$ central force Hamiltonians, and also that the Cartesian forms are the simplest because they do not involve quantum corrections, one may well ask whether there is any point to considering other coordinates. There are, however, several important reasons for doing so. The Cartesian correspondence does not, for instance, yield very good WKB results for the angular momentum operators such as $\hat{L}^2$. For radial operators, moreover, the advantages of method II do not extend to more general Hamiltonian forms, as Morehead discovered. He found in fact, that method I is generally preferable to method II. The comparison is a bit unfair though, because the mixed method I assumes exact eigenvalues for the angular momentum operators, whereas the fully WKB method II does not. We can, however, legitimately compare method II to the fully WKB method III introduced in this paper, which is closely related to method I.

III. THREE-DIMENSIONAL $T+V(r)$ HAMILTONIANS

The nonrelativistic quantum central force Hamiltonian of arbitrary dimensionality is given in Cartesian coordinates as

$$\hat{H}=\frac{\hat{p}_1^2+\cdots+\hat{p}_N^2}{2m}+V(\hat{r}), \tag{3.1}$$

where $\hat{r}^2=\hat{x}_1^2+\cdots+\hat{x}_N^2$, and $N=3$ for the standard three-dimensional case. For our purposes, the quantum Hamiltonian is *always* equivalent to Eq. (3.1), regardless of the particular coordinate system in which it might be expressed.

In contrast, the Weyl symbol of $\hat{H}$, and all associated canonically invariant quantities (e.g., WKB eigenvalues) depend on the correspondence coordinates; but once the classical symbol H is obtained, any canonical transformation can be applied without affecting those quantities. We shall often find it convenient to use one set of coordinates to obtain the correspondence, and another to express the Hamiltonian. To avoid confusion, *all explicit references and equation labels henceforth refer to the correspondence coordinates*; the expression coordinates are generally self-evident from the context. Thus, the ''Cartesian Weyl symbol'' that corresponds to the $N=3$ Hamiltonian of Eq. (3.1) can be given as follows:

$$H=\frac{p_r^2}{2m}+\left(\frac{1}{2mr^2}\right)\left(p_\theta^2+\frac{p_\phi^2}{\sin^2\theta}\right)+V(r) \qquad \text{(Cartesian)}. \tag{3.2}$$

Although we have chosen to express this symbol in spherical coordinates, the label ''(Cartesian)'' denotes the coordinates that were actually used for the correspondence.

In dealing with multidimensional Hamiltonians, it is useful both classically and quantally to obtain the other independent constants of the motion. In the present case, these are of course found in the angular momentum operators $\hat{L}^2$ and $\hat{L}_z$. The Cartesian Weyl symbols for these operators are as follows:

$$L^2=p_\theta^2+\frac{p_\phi^2}{\sin^2\theta}, \quad L_z=p_\phi \qquad \text{(Cartesian)}. \tag{3.3}$$

The Cartesian symbols for $\hat{H}$, $\hat{L}^2$, and $\hat{L}_z$ are very straightforward, even when expressed in spherical coordinates, because they do not involve quantum correction terms.

On the other hand, when the WW correspondence is obtained directly in spherical coordinates, potential-like quantum correction terms arise (Sec. II). The L_z symbol is unaffected, but L^2 becomes

$$L^2=p_\theta^2+\frac{p_\phi^2}{\sin^2\theta}-\frac{1}{2}-\frac{1}{4}\cot^2\theta \qquad \text{(spherical)}, \tag{3.4}$$

with a corresponding change in the symbol for H. In principle, the p_r^2 term in Eq. (3.2) might have also gained a quantum correction, but this turns out to be zero when $N=3$.

Thus far, we have not improved matters much by switching the correspondence from Cartesian to spherical coordinates. The new, negatively singular contribution to L^2 is clearly unphysical, and the corresponding WKB eigenvalues are less accurate than the Cartesian ones. In fact, they are not even properly degenerate, as is easy to demonstrate. First, solve the L_z problem semiclassically; then, substitute the resultant eigenvalues $L_z\to m\in\mathbb{Z}$ into the classical L^2 expression, and solve the resultant θ problem semiclassically. Labeling the resultant eigenstates via the index $n_\theta\in\{0,1,...\}$, the WKB eigenvalues of L^2 are obtained as a function of m and n_θ. If this function depends only on $\ell=n_\theta+|m|$, then the degeneracy pattern is correct; this is true for Eq. (3.3), but not for Eq. (3.4).

The situation can be improved somewhat by using $\mu = \cos\theta$ rather than θ for the polar correspondence, as was explored by the author in a recent paper.[30] This yields the much simpler quantum correction of $-1/2$. Moreover, the WKB eigenvalues are now properly degenerate, and—in comparison with the true quantum results—are seen to be too *small* by $-1/4$ (in contrast with the Cartesian values, which are well known to be too large by this amount).

Before proceeding, it is worth examining the ''μ-spherical'' L^2 symbol as expressed in the μ-spherical coordinates themselves.[30]

$$L^2 = (1-\mu^2)p_\mu^2 + \frac{p_\phi^2}{(1-\mu^2)} - \frac{1}{2} \quad (\mu\text{-spherical}). \tag{3.5}$$

Because of the altered form of the kinetic energy, Eq. (3.5) is no longer Schrödinger-like. Nevertheless, Eq. (3.5) bears a strong resemblance to the (free-particle) radial Hamiltonian as considered by Langer, in that there is a positive centrifugal barrier in μ, which diverges in an inverse-squared fashion as either end point is approached. This similarity shall soon be exploited to develop a Langer-like modification for polar angles.

In the present state of affairs, the WKB eigenvalues for $\hat{L}^2$ are still too low; our task is thus to determine whether it is expedient to improve the situation via yet another coordinate transformation. In doing so, let us reiterate the primary motivation for considering spherical coordinates—that they separate the full Hamiltonian problem into three one-dimensional problems, and are therefore closely related to the action-angle coordinates. Because H is only weakly separable in (r,θ,ϕ) however, the one-dimensional problems must be solved in a specific order, as discussed in the Appendix.

If separability is in fact the principal criterion for distinguishing the most successful correspondence coordinates, then the spherical coordinates are by no means the only choice that fits the bill. It has been shown[25] that any arbitrary canonical transformation acting on each of the spherical coordinates *independently*, will yield a new coordinate system that preserves the weak separability property (Appendix). Any *modified* spherical coordinates obtained in this fashion should therefore also be under consideration. For the present application let us restrict ourselves to point transformations, but otherwise allow the mapping $(r,\theta,\phi) \to (x,\beta,\alpha)$ given by

$$x = X(r), \quad \beta = B(\theta), \quad \alpha = A(\phi) \tag{3.6}$$

to be completely arbitrary.

As for specifying the ''best'' transformations $X(r)$, $B(\theta)$, and $A(\phi)$, this is somewhat difficult to justify *a priori*. Motivated by past successes however, we shall remain as close to Langer as possible. The choice for $X(r)$ should evidently be that of Langer and Morehead, i.e., $x = \ln r$. With this choice, the current approach (method III) becomes closely related to method I; the only difference between the two is that method I requires the exact solution of the $\hat{L}^2$ problem, whereas this is dealt with semiclassically in method III.

Without explaining precisely *why* the Langer/Morehead transformation works as well as it does for the radial coordinate, let us consider something similar for the angular coordinates θ and ϕ. In the case of ϕ, L_z is the conjugate momentum itself, implying that the WW correspondence is canonically invariant in this case (Sec. II). However, it is L_z^2 rather than L_z that appears in Eq. (3.4); given that the WKB eigenvalues of $\hat{L}_z^2$ are already exact in the ϕ correspondence, this coordinate is left alone. No modification is required here—perhaps because the azimuthal coordinate, though finite, is unbounded (Sec. V).

On the other hand, the spherical L^2 symbol is an ideal candidate for a Langer-like modification in θ. In the Langer modification proper, the singular bound at $r=0$ is extended to $x=-\infty$. In the case of θ, there are two singular bounds instead of one; nevertheless, a logarithmic transformation that extends β to $\pm\infty$ can still be defined—one simply deals with $\beta \leqslant 0$ and $\beta \geqslant 0$ separately. The question arises as to whether θ or μ should be used in the exponential relation. We

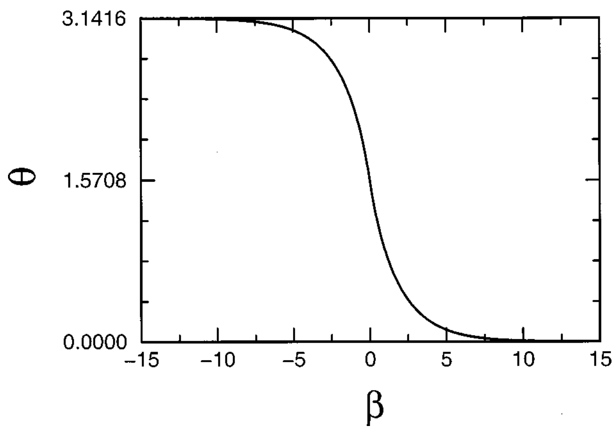

FIG. 1. Langer-like transformation from the unbounded, Cartesian-like coordinate β to the doubly bounded, polar angle coordinate θ. There is a discontinuity in the second derivative at $\beta=0$.

choose μ, because the centrifugal singularity is more like that of the radial problem, and also because the WKB eigenvalues of Eq. (3.5) exhibit the physically correct degeneracies. The piecewise transformations are as follows:

$$\begin{aligned} \mu &= e^{\beta}-1 \quad (\mu,\beta\leqslant 0), \\ \mu &= 1-e^{-\beta} \quad (\mu,\beta\geqslant 0). \end{aligned} \tag{3.7}$$

The above-mentioned transformation relations are a primary result of this paper, and are believed to comprise the correct generalization of Langer's modification for polar angle coordinates. A composite graph is given in Fig. 1. Despite the awkward discontinuity in the second derivative at $\beta=0$, it is unlikely that any other point transformation is as effective as Eq. (3.7), insofar as semiclassical accuracy is concerned. This issue is examined more closely in Sec. V.

Having specified the transformations $A(\phi)$, $B(\theta)$, and $X(r)$, it remains to determine the modifications to the L^2 and H Weyl symbols that result when the WW correspondence is obtained in the new, modified spherical coordinates. Using the standard Podolsky expression with Weyl ordering (Sec. II), it can be shown that the quantum correction to the spherical p_θ^2 symbol of Eq. (3.4) is

$$\tfrac{1}{4}(g'(\beta))^2,$$

where

$$g(\beta)=1/\theta'(\beta), \quad \theta=\theta(\beta)=B^{-1}(\beta). \tag{3.8}$$

In the $\beta\geqslant 0$ region for instance, where $\theta=\arccos(1-e^{-\beta})$, the correction is found to be

$$\mathrm{QC}(\theta\to\beta)=\tfrac{1}{4}\csc^2\theta. \tag{3.9}$$

It is easy to verify that Eq. (3.9) is also correct for $\beta\leqslant 0$. In any event, when the above result is combined with Eq. (3.4), the total quantum correction for L^2 [as compared to the Cartesian symbol of Eq. (3.3)] becomes simply $-1/4$. The WKB eigenvalues of L^2 under the new correspondence are therefore *exact*. Moreover, the degeneracy pattern is also correct, in that

$$L^2_{m,n_\theta}=\ell(\ell+1), \quad \ell=n_\theta+|m|. \tag{3.10}$$

Moving on to H, it is immediately clear from the exactness of the L^2 eigenvalues that the new radial Hamiltonian must be identical to the standard Langer-modified form of method I. More explicitly, the $-1/(8mr^2)$ correction from the L^2 contribution is exactly cancelled by the $+1/(8mr^2)$ Langer modification term from $\hat{p}_r^2$, thus resulting in a classical H symbol that is equivalent to the Cartesian symbol of method II. The coordinates (x,β,ϕ) shall be referred to as ''modified'' or ''logarithmic'' spherical coordinates; the associated WW correspondence and subsequent WKB quantization comprise the new method III.

The accuracy of all three methods, with respect to semiclassical approximations for three-dimensional central force Hamiltonians, is therefore identical. In method III however, we now have a fully WKB approach—incorporating the standard Langer modification explicitly—against which direct comparisons with method II can be legitimately made. Despite the elegance of the Cartesian symbols of method II, method III is evidently superior, for the WKB eigenvalues of L^2 as determined by method II are only approximate.

IV. *N*-DIMENSIONAL $T+V(r)$ HAMILTONIANS

For $N>3$, the generalization of the spherical coordinates are the hyperspherical coordinates, discussed in Sec. II. These are appropriate for our purposes, because they render the N-dimensional $T+V(r)$ Hamiltonians of Eq. (3.1) weakly separable, regardless of the dimensionality (Appendix). We shall continue to refer to these as ''central force Hamiltonians,'' though not in the sense of a many-body system with central force pair potentials. In any event, one still has an azimuthal angle ϕ and a radial coordinate r; but arranged in between these two are now $(N-2)$ polar angles θ_k, with $1\leq k\leq(N-2)$.

Associated with each polar angle coordinate θ_k is a constant of the motion $\hat{I}^{(k)}$, where our use of the index k is slightly different from that of Morehead,[17] and very different from that of Alcarás.[23] For convenience, let us define a zeroth constant $\hat{I}^{(0)}=\hat{p}_\phi^2$. The collection of $(N-1)$ constants $\hat{I}^{(k)}$ are in involution (commute) with the Hamiltonian $\hat{H}$, and with each other. They are conveniently expressed in terms of certain angular momentum quantities, with $\hat{I}^{(N-2)}$ being equivalent to $\hat{L}^2$.

Being weakly separable, the N-dimensional Hamiltonian problem is equivalent to a collection of N one-dimensional problems, one for each of the hyperspherical coordinates—but these one-dimensional problems must be solved in the proper order, starting with the $\hat{I}^{(0)}$ problem in ϕ, and working up to the generalized radial Schrödinger equation in r. Using hyperspherical coordinates, Morehead found that methods I and II can be generalized in a straightforward manner for arbitrary dimensionality N; moreover, he found the two to be equivalent with respect to N-dimensional central force Hamiltonians. Our task in this section is therefore to generalize method III in analogous fashion.

We must first work out the explicit Weyl symbols for the $\hat{I}^{(k)}$ operators in the various coordinate systems. The operators themselves are conveniently expressed in the Cartesian coordinates $(\hat{x}_i,\hat{p}_i)$ as sums of the squared generalized angular momentum operators[23,31]

$$\hat{L}_{ij}^2=(\hat{x}_i\hat{p}_j-\hat{x}_j\hat{p}_i)^2. \tag{4.1}$$

Despite the quadraticity in the momenta, the Cartesian Weyl symbols of the $\hat{L}_{ij}^2$'s do not have quantum corrections. In other words, the operator ordering in Eq. (4.1) is equivalent to Weyl ordering, so that the corresponding symbols are obtained by simply removing the carets from Eq. (4.1).

In any event, when the above-mentioned Cartesian Weyl symbols are expressed in hyperspherical coordinates, the following recursive relations are obtained:[17,23]

$$I^{(0)}=p_\phi^2,$$

$$I^{(1)} = p_{\theta_1}^2 + I^{(0)} \csc^2 \theta_1,$$
$$\vdots$$
$$I^{(k)} = p_{\theta_k}^2 + I^{(k-1)} \csc^2 \theta_k,$$
$$\vdots$$
$$I^{(N-2)} = p_{\theta_{N-2}}^2 + I^{(N-3)} \csc^2 \theta_{N-2} \qquad \text{(Cartesian)}. \tag{4.2}$$

The above-mentioned Cartesian relations are undeniably elegant, in that every $I^{(k)}$ is recursively equivalent to every other. Moreover, a successive WKB analysis of each of the $I^{(k)}$ in turn, yields the correct degeneracy structure to the pattern of eigenvalues. This pattern is extremely degenerate; for each $I^{(k)}$, the allowed quantum numbers ℓ_k are given by $\ell_k = n_1 + \cdots + n_k + |m|$, where each of the n_i is a non-negative integral index, as in Eq. (3.10).

Despite the correct degeneracy, however, the WKB eigenvalues themselves are quantitatively far from exact in the Cartesian correspondence. Moreover, the errors become *quadratically* worse with increasing k, and are least accurate for the most physically relevant of the constants, $\hat{I}^{(N-2)} = \hat{L}^2$. More specifically, the Cartesian WKB eigenvalues are given as

$$I_{\ell_k}^{(k)} = \left(\ell_k + \frac{k}{2} \right)^2 = \ell_k^2 + k\ell_k + \frac{k^2}{4} \qquad \text{(Cartesian)}, \tag{4.3}$$

whereas the exact quantum results are

$$I_{\ell_k}^{(k)} = \ell_k(\ell_k + k) = \ell_k^2 + k\ell_k \qquad \text{(exact quantum)}. \tag{4.4}$$

The Cartesian WKB eigenvalues are therefore too large, by $k^2/4$.

Now on to the hyperspherical correspondence. It is natural to deal with the recursive relations of Eq. (4.2) by induction. Assuming, therefore, that the hyperspherical $I^{(k-1)}$ symbol is known, then the Eq. (4.2) relation must also hold for the hyperspherical $I^{(k)}$ symbol, apart from a potential-like quantum correction term in θ_k. In other words,

$$I^{(k)} = p_{\theta_k}^2 + \frac{I^{(k-1)}}{\sin^2 \theta_k} + \mathrm{QC}_k(\theta_k) \qquad \text{(hyperspherical)}. \tag{4.5}$$

It is necessary to derive an explicit expression for the $\mathrm{QC}_k(\theta_k)$. The most straightforward way to obtain these is to apply the Podolsky procedure with Weyl ordering to the Cartesian $\hat{L}_{ij}^2$'s of Eq. (4.1). However, it is more convenient simply to use the known differential forms of these operators in hyperspherical coordinates. This will involve another kind of transformation however, from standard to canonical-like form (Sec. II).

The standard differential forms of the $\hat{I}^{(k)}$ operators are usually expressed in the ''μ-hyperspherical'' coordinates $(r, \mu_{1 \leqslant k \leqslant (N-2)}, \phi)$, where $\mu_k = \cos \theta_k$. From the standard ultraspherical equation[32] we find

$$\hat{I}_0^{(k)} = -(1-\mu_k^2) \frac{d^2}{d\mu_k^2} + (k+1)\mu_k \frac{d}{d\mu_k} \qquad \text{(standard form)}, \tag{4.6}$$

where $\hat{I}_0^{(k)} = \hat{I}^{(k)} - \hat{I}^{(k-1)} \csc^2 \hat{\theta}_k$. The standard form of Eq. (4.6) is associated with a nontrivial Jacobian weight factor that depends on the μ-hyperspherical coordinates. The Eq. (4.6) differential expression is therefore not in canonical-like form, in that $\hat{p}_{\mu_k}$ is not equivalent to $-i d/d\mu_k$ (Sec. II).

In addition to converting Eq. (4.6) into canonical-like form, we also transform from μ-hyperspherical to hyperspherical coordinates. Note that all of these manipulations are strictly quantum mechanical, and as yet have nothing to do with the WW correspondence. The final step will be to properly Weyl order the new expression for $\hat{I}_0^{(k)}$ in $\hat{\theta}_k$ and $\hat{p}_{\theta_k}$, so that the WW correspondence in these coordinates becomes manifest.

Using the chain rule to transform Eq. (4.6) to the coordinate θ_k yields

$$\hat{I}_0^{(k)} = -\frac{d^2}{d\theta_k^2} - k\cot\theta_k \frac{d}{d\theta_k} \qquad \text{(standard form)}, \tag{4.7}$$

where we have not yet taken the Jacobian weight factor into account. The original weight factor[21] was $(\sin\theta_k)^{k-1}$; however, the transformation from μ_k to θ_k introduces an extra $\sin\theta_k$ factor, resulting in the new weight factor

$$w_k(\theta_k) = \sin^k \theta_k . \tag{4.8}$$

A canonical-like form for the differential $\hat{I}_0^{(k)}$ expression of Eq. (4.7) is easily obtained by substituting the differentials with the correct expressions for the $\hat{p}_{\theta_k}$. These are as follows:[33]

$$\hat{p}_{\theta_k} = -i w_k^{-1/2} \frac{d}{d\theta_k} w_k^{1/2} = -i\frac{d}{d\theta_k} - i\frac{k}{2}\cot\theta_k \qquad \text{(standard form)}. \tag{4.9}$$

Substituting Eq. (4.9) into Eq. (4.7) yields

$$\hat{I}_0^{(k)} = \hat{p}_{\theta_k}^2 + \frac{k^2}{4}\cot^2\hat{\theta}_k - i\frac{k}{2}(\cot\hat{\theta}_k \hat{p}_{\theta_k} - \hat{p}_{\theta_k}\cot\hat{\theta}_k). \tag{4.10}$$

Although $\hat{\theta}_k$ and $\hat{p}_{\theta_k}$ are not true canonical coordinates, we nevertheless proceed using Eq. (2.1) to evaluate the last term in Eq. (4.10)—which in any case is a localized operator (Sec. II). The last term thereby becomes $-k/(2\sin^2\hat{\theta}_k)$.

Thus far we have been working with $\hat{I}_0^{(k)}$, for which $\hat{I}^{(k-1)}$ is in effect taken to be zero. By induction, however, we can restore the full $\hat{I}^{(k)}$ expression for any k, resulting at last in the following recursive, canonical-like relation:

$$\hat{I}^{(k)} = \hat{p}_{\theta_k}^2 + \frac{\hat{I}^{(k-1)}}{\sin^2\hat{\theta}_k} + \frac{k^2}{4}\cot^2\hat{\theta}_k - \frac{k/2}{\sin^2\hat{\theta}_k} \qquad \text{(canonical-like form)}. \tag{4.11}$$

Since Eq. (4.11) is already manifestly Weyl ordered in the $\{\hat{\theta}_{i\le k}, \hat{p}_{\theta_{i\le k}}\}$, the hyperspherical Weyl symbols are once again obtained by simply removing the carets. Note that the $(k+2=N=3)$ result is consistent with Eq. (3.4).

The quantum correction terms in the caretless version of Eq. (4.11) are k dependent, so that the various θ_k problems are no longer recursively equivalent to each other. Nevertheless, in the generalization of method III, let us employ the same exponential transformation for each of the $(N-2)$ polar coordinates. In other words, the modified hyperspherical (MHS) coordinates become $(x, \beta_{N-2}, \ldots, \beta_1, \phi)$, where β_k is given in terms of μ_k via Eq. (3.7).

In switching the WW correspondence from the hyperspherical to the modified hyperspherical coordinates, it is only the $\hat{p}_{\theta_k}^2$ term in Eq. (4.11) that contributes to the modification. This is because the coordinate change is a point transformation of the positions only. Moreover, since all of the $\hat{p}_{\theta_k}^2$ terms look alike, the modification must take the same form for each value of k, i.e., that of Eq. (3.9).

Incorporating the above modification, the total quantum correction for the MHS Weyl symbol for $\hat{I}^{(k)}$ (again, as compared to the Cartesian symbols) can be written as follows:

$$\mathrm{QC}_k(\text{Cartesian}\rightarrow\text{MHS}) = \frac{(k-1)^2/4}{\sin^2\theta_k} - \frac{k^2}{4}. \tag{4.12}$$

We shall prove by induction that the WKB eigenvalues for the $I^{(k)}$, obtained using the above-mentioned corrections, are exact for all k.

Let us assume that the WKB eigenvalues for $\hat{I}^{(k-1)}$ are exact in the MHS correspondence. If so, then we can replace the MHS symbol $I^{(k-1)}$ with one of its exact eigenvalues, in the MHS expression for $I^{(k)}$. In other words, we make the substitution

$$I^{(k-1)} \rightarrow I^{(k-1)}_{\ell_{k-1}} = \ell^2_{k-1} + (k-1)\ell_{k-1}. \tag{4.13}$$

By combining like terms in the MHS expression for $I^{(k)}$, we obtain an expression that defines the one-dimensional classical problem in θ_k:

$$I^{(k)}(\theta_k, p_{\theta_k}) = p^2_{\theta_k} + \frac{(\ell_{k-1} + (k-1)/2)^2}{\sin^2\theta_k} - \frac{k^2}{4} \qquad (\text{MHS}). \tag{4.14}$$

Now in view of Eqs. (4.3) and (4.4), it is clear that WKB quantization of Eq. (4.14) yields exact eigenvalues for the corresponding $\hat{I}^{(k)}$ operators. This is, of course, provided that the WKB eigenvalues are exact for the MHS symbol for $\hat{I}^{(k-1)}$. The WKB eigenvalues of the MHS symbol $I^{(0)}$ are the same as in Sec. III, and are therefore known to be exact. By induction then, the fully WKB eigenvalues of the $I^{(k)}$ observables in the MHS correspondence are also exact, for all values of k.

The degeneracy pattern of the WKB eigenvalues is also correct. This is due to the perfect square form of the numerator in the second term on the right-hand side of Eq. (4.14). Such a form is required to ensure that the eigenvalues of $I^{(k)}$ are degenerate in $l_k = n_k + l_{k-1}$. Of course, the true quantum eigenvalues are *not* perfect (half-)integral squares, but quantities which are smaller by a k-dependent constant. Thus, the quantum correction terms of Eq. (4.12) perform a delicate ''balancing act,'' wherein the first correction term compensates for the deviation from squareness of the $I^{(k-1)}$ contribution, and the second term introduces the correct deviation for the overall $I^{(k)}$ expression.

It should be emphasized that such a situation would be impossible under the standard Cartesian correspondence. Using a Cartesian approach, one would have to substitute *either* the fully WKB values *or* the exact quantum values for $I^{(k-1)}$ in Eq. (4.2). The former choice (method II) would yield the right degeneracy pattern, but highly inaccurate eigenvalues. More accurate results would be obtained with the latter choice (similar to method I), but these would no longer be properly degenerate. Obtaining both accuracy and degeneracy is evidently nontrivial.

It remains only to examine the radial coordinate WKB problem, as specified by the central force Hamiltonian $\hat{H}$ itself [Eq. (3.1)] under the MHS correspondence. Comparing this to the other two methods is trivial; since the MHS WKB eigenvalues for $\hat{I}^{(N-2)} = \hat{L}^2$ are exact, then method III must once again be equivalent to method I, and also to method II. It follows, for instance,[17] that the WKB wave functions for the radial states will have the correct asymptotic behavior as $r \rightarrow 0$.

To summarize the results thus far, we find that for all $T + V(r)$ Hamiltonians of arbitrary dimensionality, all three semiclassical methods yield identical results. However, only methods II and III are fully WKB treatments, admitting of a legitimate comparison. Although the radial coordinate WKB problems are the same for all three methods, method III is the only one that yields exact eigenvalues for the physically relevant Casimir-invariant quantities $\hat{I}^{(k)}$, associated with the angular hyperspherical coordinates.

V. ARBITRARY SPHERICALLY SYMMETRIC OBSERVABLES

In this section, we extend the analysis to arbitrary spherically symmetric observables, not necessarily of the Eq. (3.1) form. As noted by Morehead, the ability to generalize the Langer procedure to incorporate non-Schrödinger-like operators is a direct consequence of the WW correspondence interpretation, and would certainly not be possible using Langer's original theory. Put another way, Weyl ordering is more general than the Podolsky method, because it is not limited to polynomial forms in the momenta of order two or less.

Regardless of the total dimensionality N, a spherically symmetric operator $\hat{A}$ can be written as

$$\hat{A}=A(\hat{r},\hat{p}_r;\hat{I}^{(N-2)}) \quad \text{with} \quad \hat{I}^{(N-2)}=\hat{L}^2. \tag{5.1}$$

Thus, $\hat{A}$ depends only on $\hat{r}$, $\hat{p}_r$, and $\hat{L}^2$. Once again, the $\hat{I}^{(k)}$ are all constants of the motion, so that $\hat{A}$ is weakly separable in the hyperspherical coordinates. For such operators, Morehead found that methods I and II are *not* equivalent to each other.[17] It is therefore natural to ask which of the two should perform better, with respect to the accuracy of WKB results. Without offering a general proof, he suggested that method I is better, a claim which he supported with several specific results.

For example, Morehead showed that for all $\hat{A}$'s that are localized operators (differential operators in his language), method I yields exact asymptotic behavior for the radial WKB wave functions in the limit $r\rightarrow 0$. He also showed by constructing an explicit counterexample that this is false for method II. As for the accuracy of the WKB eigenvalues, this is very difficult to characterize in general. However, for the following simple non-Schrödinger-like test case

$$\hat{A}=\frac{1}{2}(\hat{r}\hat{p}^2+\hat{p}^2\hat{r})+\hat{r}, \quad \hat{p}^2=\hat{p}_r^2+\frac{\hat{L}^2}{\hat{r}^2}, \tag{5.2}$$

Morehead showed that method I yields exact eigenvalues, whereas the eigenvalues of method II are inexact.

Insofar as method III is concerned, it is a trivial matter to show that III is equivalent to I if $\hat{A}$ is linear in $\hat{L}^2$—again, because the MHS WW correspondence yields exact WKB eigenvalues for $\hat{L}^2$. The $\hat{A}$ of Eq. (5.2) is a linear $\hat{L}^2$ operator, for example. For such operators, method III should perhaps be considered a derivation of method I—it is in any event probably the best justification that one can provide for comparing the radial Langer-modified WKB method (I) to the Cartesian approach (method II).

On the other hand, it is impossible for *any* fully semiclassical method to be equivalent to method I in the more general case where $\hat{A}$ may depend *non*linearly on $\hat{L}^2$. This is due to the fact that the WW correspondence fails to preserve algebraic relationships, regardless of the coordinate system. Consequently, $W[f(\hat{L}^2)]\neq f(W[\hat{L}^2])$ generally, so that the WKB eigenvalues of $f(\hat{L}^2)$ are not necessarily exact. Nevertheless, the situation characterizing the present method III is probably the best that can be expected of a fully WKB approach.

There is a sense in which we can meaningfully extend the correspondence between methods I and III still further, however. Mathematically, an arbitrary weakly separable operator in hyperspherical coordinates (Appendix) is given by

$$\hat{A}=A(\hat{r},\hat{p}_r;\hat{I}^{(N-2)},\ldots,\hat{I}^{(0)}). \tag{5.3}$$

If we restrict ourselves to weakly separable operators which depend no more than linearly on each of the $\hat{I}^{(k)}$'s individually, then methods I and III are once again equivalent.

Equation (5.3) is not necessarily spherically symmetric. Nevertheless, we can still solve the radial problem independently of the other coordinates, for it is weak separability that is mathematically required in this regard, not rotational invariance. Even in just three dimensions, the expanded class of Eq. (5.3) contains some physically relevant operators that would not otherwise be included, such as

$$\hat{A}=B(\hat{L}_x^2+\hat{L}_y^2)+C\hat{L}_z^2, \tag{5.4}$$

reminiscent of the symmetric top free rotor.

In all cases such as Eq. (5.4) (or even $\hat{A}=\hat{L}^2+\hat{I}^{(0)}\hat{I}^{(1)}\hat{I}^{(2)}$) where $\hat{A}$ is independent of the radial coordinate, the WKB eigenvalues of method III are exact. In the more general case involving $\hat{r}$ and $\hat{p}_r$, the advantages of method I/III over method II, as discussed in Sec. IV, still apply. In particular, the eigenvalue spectra of the former are expected to be more accurate, and the asymptotic behavior of the radial wave functions is exact.

Having presented some of the advantages that the MHS correspondence provides, as well as a fairly broad class of observables which evidently benefit from such a treatment, some possible explanations shall now be offered as to why this approach seems to work so well. As Morehead put it however,[17] ''there still appears to be no general proof.'' The arguments sketched in the remainder of this section should therefore be considered somewhat speculative.

Putting off consideration of the Langer-like modifications for the moment, it seems likely that the success of a correspondence obtained in some kind of hyperspherical-*like* coordinate system is probably *due* to the fact that the Hamiltonians considered are separable when expressed in such coordinates. Presumably, the closer one can get to action-angle coordinates, the better—provided, of course, that the system is integrable.

One indication that a rationale such as this might be correct is that the WW correspondence in action-angle coordinates actually does preserve the algebraic structure, due to the mutual commutativity of the action operators $\hat{J}_k$ (Appendix). In other words,

$$W[H(\hat{J}_1,\ldots,\hat{J}_N)]=H(J_1,\ldots,J_N), \tag{5.5}$$

presuming of course that the Weyl ordering rule still holds, at least for a reasonable class of such $H(\hat{J}_1,\ldots,\hat{J}_N)$'s. In any case, this is significant because it implies that if the $\hat{J}_i$ are quantized exactly, then so would be all Hamiltonians obtained from them.

On the other hand, it is not entirely clear whether one can define a general WW correspondence rule in action-angle coordinates, and if so, for which observables conventional Weyl ordering would still apply. After all, the action-angle coordinates are not truly canonical, only canonical-like, as discussed in Sec. II. Even in hyperspherical coordinates, the validity of using the Weyl ordering rules for nonlocalized operators has been called into question. The author shares Morehead's view that a generalized new correspondence rule, for use with canonical-like coordinates such as radii and angles, is probably in order.

At present however, we have only the WW correspondence rule at our disposal; its use with canonical-like coordinates should be regarded as increasingly suspect, the further removed one is from a Cartesian system. The present application (i.e., weakly separable Hamiltonians in hyperspherical-like coordinates) should perhaps be viewed as a happy medium between Cartesian and action-angle coordinates. Because of the centrifugal potentials, the nonazimuthal coordinates are effectively somewhat Cartesian; and we can evidently get away with using Weyl ordering, at least for localized Hamiltonians.

As for the advantages of the Langer and related modifications, the improvement is likely related to the fact that the new coordinates are more like true canonical coordinates—ranging from $-\infty$ to $+\infty$, etc. Qualitative supporting evidence is provided by the following fact: Any Eq. (3.6) transformation which extends the coordinate bounds to $\pm\infty$ gives rise to a positive quantum correction term. This results in corrections to the WKB eigenvalues which are also positive, i.e., in the direction of the true eigenvalues.

There are, however, many such transformations available; and we have yet to account for why the specific exponential transforms of Langer and Eq. (3.7)—which yield exact results for the kinetic energies $\{\hat{I}^{(0)},\ldots,\hat{I}^{(N-2)},\hat{T}_r\}$ in the WKB approximation—work as well as they do. Most likely this is related to the specific form of these observables, rather than to some intrinsically special property of the exponential function. It is nevertheless worth investigating how ''special'' these transformations really are.

There is one obvious generalization of the Langer and Eq. (3.7) transformations that does not affect the semiclassical results at all. The scale transformation $\beta \to a\beta$ with a a real constant, results in the same quantum correction to p_θ^2 as for $a=1$. Thus, we can alter the length scale arbitrarily in Eq. (3.7), by replacing β with $a\beta$. Apart from this small generalization however, there appears to be little room for modification. As discussed in Sec. IV, there is a very delicate balance required of the quantum corrections in the polar angle problems.

Even for the simplest case of $N=3$, the author has been unable to find any transformation other than Eq. (3.7) that yields exact results for $\hat{L}^2$—including the seemingly more ''natural'' choices $\theta = \arccos(-\tanh\beta)$ and $\theta = \arctan(\beta) + \pi/2$. It is extremely unlikely that there are *any* other transformations that yield the right quantum correction potentials for *all* values of k; although a rigorous proof would involve a careful analysis of the appropriate nonlinear differential equation implied by Eqs. (3.8) and (4.12).

VI. CONCLUSIONS

This paper builds upon the earlier work of Langer[8] and Morehead.[17] In particular, the lack of canonical symmetry of the Wigner–Weyl correspondence rule has been exploited to improve the accuracy of semiclassical quantization, via a change of coordinates involving all N degrees of freedom. This has led to the somewhat surprising result that the Cartesian correspondence is *not* ideal from a semiclassical perspective, despite the elegance of the Cartesian expressions.

In its stead, a modified (hyper)spherical correspondence is proposed, which has been shown to be better than the Cartesian correspondence in several respects, for both Hamiltonian and angular momentum observables. The WKB eigenvalue spectra of the angular momentum-related quantities $\hat{I}^{(k)}$ was shown to be exact, for instance. Although a complete explanation for this success is still lacking, the weak separability of the observables in a hyperspherical-like coordinate system is almost certainly a key factor, as is the transformation from polar to Cartesian-like coordinates induced by Eq. (3.7).

Being a fully WKB method that treats all degrees of freedom semiclassically, the modified hyperspherical method III presented in this paper can be legitimately compared with the standard Cartesian method II. This situation is unlike that of previous spherical coordinate approaches, such as Morehead's method I, which have tended to be mixed methods. Some of the latter rely on geometric curvature or theory-of-constraints arguments to justify the introduction of the Langer modification.[12–16] It should perhaps be noted that the present work seems to question the essential significance of such arguments, in that the entire, unconstrained configuration space is employed here. Moreover, it is not clear that these other approaches would generalize as successfully beyond the three-dimensional case—though this has yet to be investigated, to the author's best knowledge. In contrast, the success of the modified hyperspherical method has been verified for arbitrary dimensionality, and for an extended class of Hamiltonians.

It was also demonstrated in a fully WKB manner, that for this extended class, the radial WKB wave functions exhibit the correct asymptotic behavior as $r \to 0$. In the future, the author would like to apply a similar analysis (i.e., the method of comparison equations[10,17]) to the WKB wave functions associated with the various polar angles for $N>3$, the so-called ''hyperspherical harmonics.''[21,23] It is also hoped that some progress will be made toward an improved correspondence rule for canonical-like coordinates. Such a rule would presumably extend the benefits of the modified (hyper)spherical correspondence to *non*localized observables, and might also enable a correspondence directly in the action-angle coordinates.

ACKNOWLEDGMENTS

The author wishes to thank Professor John C. Light for his support, and also Dr. X. Sun for bringing the work of J. J. Morehead to his attention. This work was supported, in part, by NSF Grant. No. CHE-9634440

APPENDIX: WEAKLY SEPARABLE HAMILTONIANS

The notion of weak separability in quantum mechanics has been examined in a series of earlier papers.[24–26] This appendix investigates a form of weak separability for which each degree of freedom comprises its own coordinate tier.[25] The present analysis is presented from a classical perspective, although all results could apply equally well to quantum mechanics.

A Hamiltonian H is said to be ''weakly separable'' in the N coordinates ξ_i (with $1 \leqslant i \leqslant N$), provided there exist N constants of the motion F_i such that

$$H = H(F_1, \ldots, F_N), \quad F_i = F_i(\xi_i, p_i, F_1, \ldots, F_{i-1}), \tag{A1}$$

with p_i canonically conjugate to ξ_i. The F_i comprise a cascading, ordered sequence of constants of the motion, in involution with the Hamiltonian and with each other. They cannot in general be considered action coordinates, although their existence implies that H is integrable.

In fact, the solution of each F_i problem—treated as a one-dimensional observable in (ξ_i, p_i) parametrized by the $F_{j<i}$—gives rise to the transformation to the true action-angle coordinates (ϕ_i, J_i). The explicit transformation is given by

$$\begin{aligned} \phi_i(\xi_1, p_1 \ldots, \xi_i, p_i) &= \phi_i^{(F_1, \ldots, F_{i-1})}(\xi_i, p_i), \\ J_i(\xi_1, p_1 \ldots, \xi_i, p_i) &= J_i^{(F_1, \ldots, F_{i-1})}(\xi_i, p_i), \end{aligned} \tag{A2}$$

Where $J_i^{(F_1, \ldots, F_{i-1})}(\xi_i, p_i)$ is the parametrized action associated with the one-dimensional F_i problem, etc.

In contrast to the weakly separable situation, H is said to be ''strongly separable'' in the ξ_i coordinates if

$$H = H(F_1, \ldots, F_N), \quad F_i = F_i(\xi_i, p_i). \tag{A3}$$

The Eq. (A3) definition of strong separability is slightly more general than what has been used in the past.[25]

As in the weakly separable case, a strongly separable system can be solved completely by treating each of the F_i's as independent, one-dimensional systems. The indexing of the various F_i problems is arbitrary however, unlike that of the weakly separable case. As a result, the strongly separable F_i problems can be solved in any order, whereas in the weakly separable case, we must proceed in sequence from F_1 to F_N, due to the fact that the $F_{j<i}$ appear as parameters in the F_i problem.

The ordering requirement on weakly separable systems is not, in practice, a serious limitation; for all intents and purposes, a weakly separable representation is as convenient as a strongly separable one. As a constraint upon possible coordinate systems in which to express a particular H however, the weakly separable form is much preferred over the strongly separable form, because it is *far* less restrictive than the latter. There is, in general, an enormous variety of coordinate systems in which the expression of a given Hamiltonian is weakly separable. For the strongly-separable case, however, one is in effect constrained to using action-angle coordinates only.

More accurately, each of the action-angle coordinate pairs (ϕ_i, J_i), when viewed as functions of the strongly separable coordinates, must depend only on the single pair (ξ_i, p_i). In the weakly separable case however, the corresponding limitation is much less severe: (ϕ_i, J_i) may depend arbitrarily on the entire set of coordinates $\{\xi_1, p_1, \ldots, \xi_i, p_i\}$, for each value of i. Incidentally, these restrictions imply that both strong separability and weak separability are preserved by canonical transformations that affect each (ξ_i, p_i) pair independently.

A weakly separable treatment is particularly useful for the Hamiltonians considered in this paper, because only a point transformation of the original Cartesian positions is required to obtain a weakly separable expression for H. In contrast, any strongly separable expression would require a more general canonical transformation involving both positions and momenta. The subsequent modification via Eq. (3.6), in addition to preserving weak separability, also preserves the point transformation property.

[1] A. Ram, in *Geometric Analysis and Lie Theory in Mathematics and Physics*, edited by A. L. Carey and M. K. Murray (Cambridge University Press, Cambridge, 1998), Chap. 2.
[2] H. Weyl, Z. Phys. **46**, 1 (1928).
[3] E. Wigner, Phys. Rev. **40**, 749 (1932).
[4] J. E. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).
[5] M. Hillery, R. F. O'Connell, M. O. Scully, and E. P. Wigner, Phys. Lett. **106**, 121 (1984).
[6] R. G. Littlejohn, Phys. Lett. **138**, 193 (1986).
[7] S. W. McDonald, Phys. Lett. **158**, 337 (1989).
[8] R. E. Langer, Phys. Rev. **51**, 669 (1937).
[9] S. C. Miller and R. H. Good, Phys. Rev. **91**, 174 (1953).
[10] M. V. Berry and K. V. Mount, Rep. Prog. Phys. **35**, 315 (1972).
[11] M. V. Berry and A. M. Ozorio de Almeida, J. Phys. A **6**, 1451 (1973).
[12] B. S. Dewitt, Rev. Mod. Phys. **29**, 3774 (1957).
[13] K. S. Cheng, J. Math. Phys. **23**, 1723 (1972).
[14] L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981).
[15] M. K. Ali, Can. J. Phys. **74**, 255 (1996).
[16] L. Kaplan, N. T. Maitra, and E. J. Heller, Phys. Rev. A **56**, 2592 (1997).
[17] J. J. Morehead, J. Math. Phys. **36**, 5431 (1995).
[18] B. Podolsky, Phys. Rev. **32**, 812 (1928).
[19] A. M. Ozorio de Almeida, in *Hamiltonian Systems: Chaos and Quantization* (Cambridge University Press, Cambridge, 1988), Chap. 7.
[20] For most monomials, there are actually two distinct Weyl orderings. However, these can be shown to result in equivalent quantum operators.
[21] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972), Chap. 22.
[22] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, in *Bateman Manuscript Project* (Addison–Wesley, Reading, MA, 1962), Vol. 2, Chap. 12.
[23] J. A. C. Alcarás and P. Leal Ferreira, J. Math. Phys. **6**, 578 (1965).
[24] B. Poirier and W. H. Miller, Chem. Phys. Lett. **265**, 77 (1997).
[25] B. Poirier, Phys. Rev. A **56**, 120 (1997).
[26] B. Poirier, J. Chem. Phys. **108**, 5216 (1998).
[27] B. Lesche and T. H. Seligman, J. Phys. A **19**, 91 (1986).
[28] J. K. G. Watson, Mol. Phys. **15**, 479 (1968).
[29] H. Essén, Am. J. Phys. **46**, 983 (1978).
[30] B. Poirier, Phys. Rev. E (submitted).
[31] F. T. Smith, Phys. Rev. **120**, 1058 (1960).
[32] G. Arfken, in *Mathematical Methods for Physicists* (Academic, New York, 1985), p. 735.
[33] J. K. G. Watson, Mol. Phys. **19**, 465 (1970) (Appendix).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

On separable Schrödinger equations

Renat Zhdanov[a] and Alexander Zhalij[b]
Institute of Mathematics of the Academy of Sciences of Ukraine, Tereshchenkivska Street 3, 252004 Kyiv, Ukraine

(Received 18 November 1998; accepted for publication 2 August 1999)

We classify $(1+3)$-dimensional Schrödinger equations for a particle interacting with the electromagnetic field that are solvable by the method of separation of variables. As a result, we get 11 classes of the vector potentials of the electromagnetic field $A(t,\vec{x})=(A_0(t,\vec{x}),\vec{A}(t,\vec{x}))$ providing separability of the corresponding Schrödinger equations. It is established, in particular, that the necessary condition for the Schrödinger equation to be separable is that the magnetic field must be independent of the spatial variables. Next, we prove that any Schrödinger equation admitting variable separation into second-order ordinary differential equations can be reduced to one of the 11 separable Schrödinger equations mentioned above and carry out variable separation in the latter. Furthermore, we apply the results obtained for separating variables in the Hamilton–Jacobi equation. © *1999 American Institute of Physics.* [S0022-2488(99)02211-2]

I. INTRODUCTION

Being invented by Fourier and Euler long ago, the method of separation of variables is still the most powerful and efficient one for integrating linear partial differential equations (PDEs). This is especially the case for PDEs having variable coefficients, where the standard Fourier transform is no longer applicable. Moreover, this method proves to be a useful tool for constructing particular solutions of some nonlinear partial differential equations such as the nonlinear Laplace,[1] wave,[2,3] and heat conductivity equations[4–6] in $(1+1)$ dimensions.

The principal object of study in the present paper is a problem of the separation of variables in the Schrödinger equation (SE) for a particle interacting with the electromagnetic field,

$$(p_0-p_ap_a)\,\psi(t,\vec{x})=0. \tag{1}$$

Here we use the notations

$$p_0=i\,\frac{\partial}{\partial t}-eA_0(t,\vec{x}),\quad p_a=i\,\frac{\partial}{\partial x_a}-eA_a(t,\vec{x}),\quad a=1,2,3,$$

where $A=(A_0,A_1,A_2,A_3)$ is the vector potential of the electromagnetic field, $e=$const. Hereafter the summation over the repeated indices from 1 to 3 is understood.

Böcher is believed to be the first to obtain in Ref. 7 a systematic classification of coordinate systems enabling separability of the three-dimensional stationary SE,

$$(-\Delta+E)\,\psi(\vec{x})=0, \tag{2}$$

where Δ is the Laplacian in three dimensions.

He has shown that Eq. (2) is separable via the separation *Ansatz*,

$$\psi(\vec{x})=\varphi_1(\omega_1(\vec{x}))\varphi_2(\omega_2(\vec{x}))\varphi_3(\omega_3(\vec{x})), \tag{3}$$

[a]Electronic-mail: renat@imath.kiev.ua
[b]Electronic-mail: zhaliy@imath.kiev.ua

in 11 inequivalent coordinate systems $\omega_1(\vec{x}),\omega_2(\vec{x}),\omega_3(\vec{x})$. We list these following the famous paper by Eisenhart,[8] where a rigorous geometric derivation of the corresponding results is given. Note that the coordinate systems are given in the implicit form $x_a=z_a(\omega_1,\omega_2,\omega_3)$, $a=1,2,3$.

(1) Cartesian coordinate system,

$$z_1=\omega_1,\quad z_2=\omega_2,\quad z_3=\omega_3,\qquad \omega_1,\omega_2,\omega_3\in\mathbf{R}.$$

(2) Cylindrical coordinate system,

$$z_1=e^{\omega_1}\cos\omega_2,\quad z_2=e^{\omega_1}\sin\omega_2,\quad z_3=\omega_3,\qquad 0\leqslant\omega_2<2\pi,\quad \omega_1,\omega_3\in\mathbf{R}.$$

(3) Parabolic cylindrical coordinate system,

$$z_1=(\omega_1^2-\omega_2^2)/2,\quad z_2=\omega_1\omega_2,\quad z_3=\omega_3,\qquad \omega_1>0,\quad \omega_2,\omega_3\in\mathbf{R}.$$

(4) Elliptic cylindrical coordinate system,

$$z_1=a\cosh\omega_1\cos\omega_2,\quad z_2=a\sinh\omega_1\sin\omega_2,\quad z_3=\omega_3,\qquad \omega_1>0,\quad -\pi<\omega_2\leqslant\pi,\quad \omega_3\in\mathbf{R},\quad a>0.$$

(5) Spherical coordinate system,

$$z_1=\omega_1^{-1}\operatorname{sech}\omega_2\cos\omega_3,\quad z_2=\omega_1^{-1}\operatorname{sech}\omega_2\sin\omega_3,\quad z_3=\omega_1^{-1}\tanh\omega_2,\qquad \omega_1>0,\quad \omega_2\in\mathbf{R},\quad 0\leqslant\omega_3<2\pi.$$

(6) Prolate spheroidal coordinate system,

$$z_1=a\operatorname{cosech}\omega_1\operatorname{sech}\omega_2\cos\omega_3,\quad a>0,\quad z_2=a\operatorname{cosech}\omega_1\operatorname{sech}\omega_2\sin\omega_3,\quad z_3=a\coth\omega_1\tanh\omega_2,\qquad \omega_1>0,\quad \omega_2\in\mathbf{R},\quad 0\leqslant\omega_3<2\pi. \tag{4}$$

(7) Oblate spheroidal coordinate system,

$$z_1=a\operatorname{cosec}\omega_1\operatorname{sech}\omega_2\cos\omega_3,\quad a>0,\quad z_2=a\operatorname{cosec}\omega_1\operatorname{sech}\omega_2\sin\omega_3,\quad z_3=a\cot\omega_1\tanh\omega_2,\qquad 0<\omega_1<\pi/2,\quad \omega_2\in\mathbf{R},\quad 0\leqslant\omega_3<2\pi.$$

(8) Parabolic coordinate system,

$$z_1=e^{\omega_1+\omega_2}\cos\omega_3,\quad z_2=e^{\omega_1+\omega_2}\sin\omega_3,\quad z_3=(e^{2\omega_1}-e^{2\omega_2})/2,\qquad \omega_1,\omega_2\in\mathbf{R},\quad 0\leqslant\omega_3\leqslant 2\pi.$$

(9) Paraboloidal coordinate system,

$$z_1=2a\cosh\omega_1\cos\omega_2\sinh\omega_3,\quad a>0,\quad z_2=2a\sinh\omega_1\sin\omega_2\cosh\omega_3,\quad z_3=a(\cosh 2\omega_1+\cos 2\omega_2-\cosh 2\omega_3)/2,\qquad \omega_1,\omega_3\in\mathbf{R},\quad 0\leqslant\omega_2<\pi.$$

(10) Ellipsoidal coordinate system,

$$z_1=a\frac{1}{\operatorname{sn}(\omega_1,k)}\operatorname{dn}(\omega_2,k')\operatorname{sn}(\omega_3,k),\quad a>0,$$

$$z_2=a\frac{\operatorname{dn}(\omega_1,k)}{\operatorname{sn}(\omega_1,k)}\operatorname{cn}(\omega_2,k')\operatorname{cn}(\omega_3,k),$$

$$z_3 = a\frac{\mathrm{cn}(\omega_1,k)}{\mathrm{sn}(\omega_1,k)}\mathrm{sn}(\omega_2,k')\mathrm{dn}(\omega_3,k),$$
$$0<\omega_1<K,\quad -K'\leqslant\omega_2\leqslant K',\quad 0\leqslant\omega_3\leqslant 4K.$$

(11) Conical coordinate system,

$$z_1=\omega_1^{-1}\mathrm{dn}(\omega_2,k')\mathrm{sn}(\omega_3,k),$$
$$z_2=\omega_1^{-1}\mathrm{cn}(\omega_2,k')\mathrm{cn}(\omega_3,k),$$
$$z_3=\omega_1^{-1}\mathrm{sn}(\omega_2,k')\mathrm{dn}(\omega_3,k),$$
$$\omega_1>0,\quad -K'\leqslant\omega_2\leqslant K',\quad 0\leqslant\omega_3\leqslant 4K.$$

Here we use the usual notations for the trigonometric, hyperbolic and Jacobi elliptic functions, $k(0<k<1)$ being the modulus of the latter and $k'=(1-k^2)^{1/2}$.

By the evident reasons, the coordinate systems 1, 2–4, and 5–11 from the above list are called completely split, partially split, and nonsplit, correspondingly.

Note that the above list differs slightly from the one presented in Ref. 8, since we have rearranged the coordinate systems $\omega_1,\omega_2,\omega_3$ in such a way that the relations

$$\Delta\omega_a=0,\quad a=1,2,3 \tag{5}$$

hold for all the cases 1–11.

Next, for each of the coordinate systems, Eisenhart[9] determined the form of the potential $V(\vec{x})$ that permits the separation of variables.

The Eisenhart's technique has been applied by Olevskii[10] for separating variables in the Laplace–Beltrami operator in the spaces of constant curvature (see, also Ref. 11).

Smorodinsky and Winternitz with co-workers[12,13] started a systematic study of potentials for which the stationary SE in two and three dimensions admits the separation of variables in two or more coordinate systems (so-called superintegrable potentials). The classification of these potentials has been completed by Evans.[14]

In the mid 1970s a series of papers by Miller and Kalnins appears, where a symmetry approach to variable separation has been developed. This approach is based on the well-known fact that a solution with separated variables is a common eigenfunction of first- or second-order differential operators, which commute each with another and with the operator of an equation under consideration. Further details and an extensive list of references can be found in the monograph[15] (also see, the review by Koornwinder[16]). Boyer, Kalnins, and Miller have obtained a systematic solution of the variable separation in the time-dependent $(1+2)$-dimensional free SE within the framework of the symmetry approach.[17] Later on, Boyer has described all coordinate systems providing separability of the $(1+2)$-dimensional SE having the potential $V(x_1,x_2)=\alpha/x_1^2+\beta/x_2^2$.[18] Reid has completely solved the problem of variable separation in the three-dimensional time-dependent SE for a free particle.[19]

Independently, the symmetry approach to the separation of variables in the equations of quantum mechanics and quantum field theory was developed by Shapovalov, who was the first to give a systematic treatment of the problem of variable separation in the Dirac equation using its non-Lie symmetry[20] and by Bagrov with collaborators (see, Ref. 21 and references therein). Shapovalov and Sukhomlin[22] have obtained some separable SEs of the form (1), however, their results are not complete. Let us also mention the papers[23,24] where the physical aspects of the problem of the separation of variables in some $(1+3)$- and $(1+1)$-dimensional SE with time-dependent potentials are studied.

In the present paper we are mainly devoted to the inverse problem of variable separation in SE (1), namely, to one of the classifying PDEs of the form (1) that can be solved by the method of separation of variables. Clearly, to be able to handle this problem efficiently, we need a precise algorithmic definition of what the separation of variables is. We have suggested a possible definition of the separation of variables applicable both to linear[25,26] and nonlinear[27] PDEs, which enables developing an efficient approach to solving classification problems for

$(1+2)$-dimensional time-dependent Schrödinger equations having time-independent[26,27] and time-dependent scalar potentials.[28] Recently, we have obtained an exhaustive classification of separable Schrödinger equations (1) in $(1+2)$ dimensions with $A_3=0$.[29]

In the present paper we solve completely the problem of variable separation in SE (1) into second-order ordinary differential equations in a sense that we obtain all possible forms of the vector potential $A(t,\vec{x})=(A_0(t,\vec{x}),A_1(t,\vec{x}),A_2(t,\vec{x}),A_3(t,\vec{x}))$ providing separability of (1). Furthermore, we construct inequivalent coordinate systems enabling us to separate variables in the corresponding SEs and carry out variable separation.

The paper has the following structure. In the second section we solve the classification problem for SE (1) and obtain all vector functions $A(t,\vec{x})$ such that (1) is separable. What is more, we consider briefly the problem of the separation of variables in SE having the generalized Coulomb potential and construct all the possible coordinate systems providing its separability. The third section is devoted to the application of the results of Sec. II to separate variables in the $(1+3)$-dimensional Hamilton–Jacobi equation. In a concluding section we indicate some further applications of the results obtained in the paper.

II. SEPARATION OF VARIABLES IN THE SCHRÖDINGER EQUATION

With all the variety of approaches to a separation of variables in PDEs, one can notice the three generic principles respected by all of them, namely, the following.

(1) Representation of a solution to be found in a separated (factorized) form via several functions of one variable.
(2) Requirement that the above mentioned functions of one variable should satisfy some ordinary differential equations.
(3) Dependence of the so found solution on several arbitrary (continuous or discrete) parameters, called spectral parameters or separation constants.

By a proper formalizing the above features we have formulated in Ref. 27 an algorithm for variable separation in linear PDEs. Below we apply this algorithm in order to classify separable SEs of the form (1).

We say that SE (1) is separable in a coordinate system t, $\omega_a=\omega_a(t,\vec{x})$, $a=1,2,3$ if the separation Ansatz,

$$\psi(t,\vec{x})=Q(t,\vec{x})\varphi_0(t)\prod_{a=1}^{3}\varphi_a(\omega_a(t,\vec{x}),\vec{\lambda}), \tag{6}$$

reduces PDE (1) to four ordinary differential equations for the functions φ_μ, $(\mu=0,1,2,3)$,

$$\varphi_0'=U_0(t,\varphi_0;\vec{\lambda}),\qquad \varphi_a''=U_a(\omega_a,\varphi_a,\varphi_a';\vec{\lambda}). \tag{7}$$

Here $U_0,\dots,U_3$ are some smooth functions of the indicated variables, $\vec{\lambda}=(\lambda_1,\lambda_2,\lambda_3)\in\Lambda=\{\text{an open domain in } \mathbf{R}^3\}$ are separation constants and, what is more,

$$\operatorname{rank}\left\|\begin{matrix}\frac{\partial U_0}{\partial\lambda_1} & \frac{\partial U_0}{\partial\lambda_2} & \frac{\partial U_0}{\partial\lambda_3}\\ \frac{\partial U_1}{\partial\lambda_1} & \frac{\partial U_1}{\partial\lambda_2} & \frac{\partial U_1}{\partial\lambda_3}\\ \frac{\partial U_2}{\partial\lambda_1} & \frac{\partial U_2}{\partial\lambda_2} & \frac{\partial U_2}{\partial\lambda_3}\\ \frac{\partial U_3}{\partial\lambda_1} & \frac{\partial U_3}{\partial\lambda_2} & \frac{\partial U_3}{\partial\lambda_3}\end{matrix}\right\|=3. \tag{8}$$

The above condition secures the essential dependence of a solution with separated variables on the separation constants $\vec{\lambda}$.

Formulas (6)–(8) form the input data of the method. The principal steps of the procedure of variable separation in SE (1) are as follows.

(1) We insert the *Ansatz* (6) into SE and express the derivatives $\varphi_0', \varphi_1'', \varphi_2'', \varphi_3''$ in terms of functions $\varphi_0, \varphi_1, \varphi_2, \varphi_3, \varphi_1', \varphi_2', \varphi_3'$ using Eqs. (7).
(2) We regard $\varphi_0, \varphi_1, \varphi_2, \varphi_3, \varphi_1', \varphi_2', \varphi_3', \lambda_1, \lambda_2, \lambda_3$ as the new independent variables $y_1, \ldots, y_{10}$. As the functions $Q, \omega_1, \omega_2, \omega_3$ are independent of the variables $y_1, \ldots, y_{10}$ we can split by these and get an overdetermined system of nonlinear partial differential equations for unknown functions $Q, \omega_1, \omega_2, \omega_3$.
(3) After solving the above system we get an exhaustive description of coordinate systems providing separability of SE.

Having performed the first two steps of the above algorithm, we arrive at the conclusion that the separation equations (7) are linear both in $\varphi_0, \ldots, \varphi_3$ and $\lambda_1, \lambda_2, \lambda_3$ [the principal reason for this is the fact that SE (1) is linear].

Next, we introduce an equivalence relation $\mathcal{E}$ on the set of all coordinate systems providing the separability of SE. We say that two coordinate systems $t, \omega_1, \omega_2, \omega_3$ and $\tilde{t}, \tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3$ are equivalent if the corresponding *Ansätze* (6) are transformed one into another by the group transformations from the Lie transformation group admitted by SE (1) with $A=0$; the invertible transformations of the form

$$t \to \tilde{t} = f_0(t), \quad \omega_i \to \tilde{\omega}_i = f_i(\omega_i), \tag{9}$$

$$Q \to \tilde{Q} = Q l_0(t) l_1(\omega_1) l_2(\omega_2) l_3(\omega_3), \tag{10}$$

where $f_0, \ldots, f_3, l_0, \ldots, l_3$ are some smooth functions and $i=1,2,3$.

This equivalence relation reflects the freedom in the choice of the functions $Q, \omega_1, \omega_2, \omega_3$ and separation constants $\lambda_1, \lambda_2, \lambda_3$ preserving the form of the separation *Ansatz* (6). It splits the set of all possible coordinate systems into equivalence classes. In a sequel, when presenting the lists of coordinate systems enabling us to separate variables in SE we will give only one representative for each equivalence class.

Within the equivalence relation $\mathcal{E}$ we can always choose the reduced equations (7) to be

$$i\varphi_0' = (T_0(t) - T_i(t)\lambda_i)\varphi_0, \quad \varphi_a'' = (F_{a0}(\omega_a) + F_{ai}(\omega_a)\lambda_i)\varphi_a, \tag{11}$$

where T_0, T_i, F_{a0}, F_{ai} are some smooth functions of the indicated variables, $a=1,2,3$. With this remark the system of nonlinear PDEs for unknown functions $Q, \omega_1, \omega_2, \omega_3$ takes the form

$$\frac{\partial \omega_i}{\partial x_a} \frac{\partial \omega_j}{\partial x_a} = 0, \quad i \neq j, \quad i,j=1,2,3; \tag{12}$$

$$\sum_{i=1}^{3} F_{ia}(\omega_i) \frac{\partial \omega_i}{\partial x_j} \frac{\partial \omega_i}{\partial x_j} = T_a(t), \quad a=1,2,3; \tag{13}$$

$$2\left(\frac{\partial Q}{\partial x_j} + ieQA_j\right)\frac{\partial \omega_a}{\partial x_j} + Q\left(i\frac{\partial \omega_a}{\partial t} + \Delta\omega_a\right) = 0, \quad a=1,2,3; \tag{14}$$

$$Q\sum_{i=1}^{3} F_{i0}(\omega_i) \frac{\partial \omega_i}{\partial x_j} \frac{\partial \omega_i}{\partial x_j} + i\frac{\partial Q}{\partial t} + \Delta Q + 2ieA_a\frac{\partial Q}{\partial x_a} + Q\left(T_0(t) + ie\frac{\partial A_a}{\partial x_a} - eA_0 - e^2 A_a A_a\right) = 0. \tag{15}$$

Thus, the problem of variable separation in SE reduces to integrating the system of ten nonlinear PDEs for four functions. What is more, some coefficients are arbitrary functions that should be determined while integrating Eqs. (12)–(15). We have succeeded in constructing the general solution of the later, which yields, in particular, all possible vector-potentials $A(t,\vec{x}) = (A_0(t,\vec{x}),\dots,A_3(t,\vec{x}))$ such that SE (1) is solvable by the method of separation of variables.

The integration procedure relies heavily upon the results on the separation of variables in the stationary SE (2). That is why we will briefly consider the principal steps of application of our approach for separating variables in Eq. (2) with $E=1$,

$$\Delta_3\psi-\psi=0. \tag{16}$$

Inserting the separation *Ansatz*,

$$\psi(\vec{x})=Q(\vec{x})\varphi_1(\omega_1(\vec{x}))\phi_2(\omega_2(\vec{x}))\varphi_3(\omega_3(\vec{x})),$$

into (16) and taking into account the relations

$$\varphi_a''=(F_{a1}(\omega_a)+F_{a2}(\omega_a)\lambda_1+F_{a3}(\omega_a)\lambda_2)\varphi_a, \quad a=1,2,3,$$

we get the system of nonlinear partial differential equations for the functions $Q,\omega_1,\omega_2,\omega_3$,

$$\begin{aligned}
&\text{(i)} \quad \frac{\partial\omega_i}{\partial x_a}\frac{\partial\omega_j}{\partial x_a}=0, \quad i\neq j, \quad i,j=1,2,3;\\
&\text{(ii)} \quad \sum_{i=1}^{3} F_{ia}(\omega_i)\frac{\partial\omega_i}{\partial x_j}\frac{\partial\omega_i}{\partial x_j}=0, \quad a=2,3;\\
&\text{(iii)} \quad 2\frac{\partial Q}{\partial x_j}\frac{\partial\omega_a}{\partial x_j}+Q\Delta\omega_a=0, \quad a=1,2,3;\\
&\text{(iv)} \quad Q\sum_{i=1}^{3} F_{i1}(\omega_i)\frac{\partial\omega_i}{\partial x_j}\frac{\partial\omega_i}{\partial x_j}+\Delta Q-Q=0.
\end{aligned} \tag{17}$$

Now we can utilize the classical results on variable separation in the stationary SE (16). According to Ref. 8 the general solution $\vec{\omega}=\vec{\omega}(\vec{x})$ of system (17) splits into 11 inequivalent classes whose representatives are given in (4).

As $\omega_1,\omega_2,\omega_3$ are functionally independent, the inequality

$$\det\left\|\frac{\partial\omega_i}{\partial x_a}\right\|_{i,a=1}^{3}\neq 0, \tag{18}$$

holds. Taking into account this fact and relations (5), we get from (iii) that $Q(\vec{x})=\text{const}$, whence it follows that without losing generality we may choose $Q(\vec{x})=1$. In view of this we rewrite system (17) in the following way:

$$\begin{aligned}
&\frac{\partial\omega_i}{\partial x_a}\frac{\partial\omega_j}{\partial x_a}=0, \quad i\neq j, \quad i,j=1,2,3;\\
&\sum_{i=1}^{3} F_{ij}(\omega_i)\omega_{ix_a}\omega_{ix_a}=\delta_{1j}, \quad j=1,2,3,
\end{aligned} \tag{19}$$

where δ_{1j} is the Kronecker symbol.

System (19) coincides with the system of Eqs. (12) and (13) under $T_1=1$, $T_2=T_3=0$. Next, in view of arbitrariness of the choice of the separation constants $\lambda_1,\lambda_2,\lambda_3$ we can always suppose that $T_1(t_0)=1$, $T_1(t_0)=T_2(t_0)=0$ for some $t_0\in\mathbf{R}$. Consequently, system (12), (13) with $t=t_0$ takes the form (19). This is a key point enabling us to use the results of Ref. 8 for integrating system (12), (13).

Lemma 1: The general solution $\vec{\omega}=\vec{\omega}(t,\vec{x})$ of the system of partial differential equations (12), (13) is given implicitly by the following formulas:

$$\vec{x}=\mathcal{T}(t)H(t)\vec{z}(\vec{\omega})+\vec{w}(t). \tag{20}$$

Here $\mathcal{T}(t)$ is the time-dependent 3×3 *orthogonal matrix*:

$$\mathcal{T}(t)=\begin{pmatrix} \cos\alpha\cos\beta-\sin\alpha\sin\beta\cos\gamma & -\cos\alpha\sin\beta-\sin\alpha\cos\beta\cos\gamma & \sin\alpha\sin\gamma \\ \sin\alpha\cos\beta+\cos\alpha\sin\beta\cos\gamma & -\sin\alpha\sin\beta+\cos\alpha\cos\beta\cos\gamma & -\cos\alpha\sin\gamma \\ \sin\beta\sin\gamma & \cos\beta\sin\gamma & \cos\gamma \end{pmatrix} \tag{21}$$

α,β,γ being arbitrary smooth functions of t;$\vec{z}=\vec{z}(\vec{\omega})$ is given by one of the 11 formulas from (4); *$H(t)$ is the* 3×3 *diagonal matrix,*

$$H(t)=\begin{pmatrix} h_1(t) & 0 & 0 \\ 0 & h_2(t) & 0 \\ 0 & 0 & h_3(t) \end{pmatrix}, \tag{22}$$

where (a) $h_1(t),h_2(t),h_3(t)$ are arbitrary smooth functions for the completely split coordinate system [case 1 from (4)], (b) $h_1(t)=h_2(t)$, $h_1(t)$, $h_3(t)$ being arbitrary smooth functions, for the partially split coordinate systems [cases 2–4 from (4)], (c) $h_1(t)=h_2(t)=h_3(t),h_1(t)$ being an arbitrary smooth function, for nonsplit coordinate systems [cases 5–11 from (4)], and $\vec{\omega}(t)$ stands for the vector column whose entries $\omega_1(t),\omega_2(t),\omega_3(t)$ are arbitrary smooth functions of t.

Proof: First we perform the hodograph transformation in the system of PDEs (12), (13),

$$t=t,\quad x_a=u_a(t,\omega_1,\omega_2,\omega_3),\quad a=1,2,3. \tag{23}$$

Direct computation shows that the following identities hold:

$$\begin{gathered}(1)\quad \omega_{ix_a}\omega_{jx_a}\equiv\frac{1}{\delta^2}(\Omega_{ik}\Omega_{jk}-\Omega_{ij}\Omega_{kk}),\quad (i,j,k)=\text{cycle}(1,2,3),\\ (2)\quad \omega_{ix_a}\omega_{ix_a}\equiv\frac{1}{\delta^2}(\Omega_{jj}\Omega_{kk}-\Omega_{jk}^2),\quad (i,j,k)=\text{cycle}(1,2,3),\end{gathered} \tag{24}$$

where

$$\Omega_{ij}=\frac{\partial u_a}{\partial\omega_i}\frac{\partial u_a}{\partial\omega_j},\qquad \delta=\det\left\|\frac{\partial u_a}{\partial\omega_b}\right\|_{a,b=1}^{3}\neq0.$$

Consequently, the initial system (12), (13), after being rewritten in the new variables, reads as

$$\begin{gathered}\frac{\partial u_a}{\partial\omega_i}\frac{\partial u_a}{\partial\omega_j}=0,\quad i\neq j,\quad i,j=1,2,3;\\ \sum_{(i,j,k)=\text{cycle}(1,2,3)}F_{ia}(\omega_i)\frac{\partial u_b}{\partial\omega_j}\frac{\partial u_b}{\partial\omega_j}\frac{\partial u_c}{\partial\omega_k}\frac{\partial u_c}{\partial\omega_k}=\delta^2T_a(t),\quad a=1,2,3.\end{gathered} \tag{25}$$

If we introduce three vectors,

$$\vec{v}_1=\frac{\partial\vec{u}}{\partial\omega_1},\quad \vec{v}_2=\frac{\partial\vec{u}}{\partial\omega_2},\quad \vec{v}_3=\frac{\partial\vec{u}}{\partial\omega_3},$$

then the first three equations of system (25) read as $\vec{v}_a\vec{v}_b=0$, $(a,b=1,2,3,\ a\neq b)$. Consequently $\vec{v}_1,\vec{v}_2,\vec{v}_3$ form an orthogonal system of vectors in the space $\mathbf{R}^3$. It is well known from the analytical geometry that the most general form of these vectors can be expressed via the Euler angles,

$$\frac{\partial\vec{u}}{\partial\omega_1}=R_1\begin{Bmatrix}\cos f_1\cos f_2-\sin f_1\sin f_2\cos f_3\\ \sin f_1\cos f_2+\cos f_1\sin f_2\cos f_3\\ \sin f_2\sin f_3\end{Bmatrix},$$

$$\frac{\partial\vec{u}}{\partial\omega_2}=R_2\begin{Bmatrix}-\cos f_1\sin f_2-\sin f_1\cos f_2\cos f_3\\ -\sin f_1\sin f_2+\cos f_1\cos f_2\cos f_3\\ \cos f_2\sin f_3\end{Bmatrix},\tag{26}$$

$$\frac{\partial\vec{u}}{\partial\omega_3}=R_3\begin{Bmatrix}\sin f_1\sin f_3\\ -\cos f_1\sin f_3\\ \cos f_3\end{Bmatrix},$$

where $f_1,\ldots,\ R_3$ are arbitrary smooth functions of $t,\ \omega_1,\omega_2,\omega_3$. The above formulas give the most general form of functions $\partial u_a/\partial\omega_b,\ a,b=1,2,3$ that satisfy the first three equations from (25). Next, inserting (26) into the remaining equations from (25) yields

$$\sum_{i=1}^{3}F_{ij}(\omega_i)R_i^{-2}=T_j(t),\quad j=1,2,3.\tag{27}$$

Thus, we have transformed system (25) to equivalent form (26), (27), where one should take into account the compatibility conditions for the system of PDEs (26),

$$\frac{\partial}{\partial\omega_k}\left(\frac{\partial u_i}{\partial\omega_j}\right)=\frac{\partial}{\partial\omega_j}\left(\frac{\partial u_i}{\partial\omega_k}\right),\quad j\neq k,\quad i,j,k=1,2,3.$$

Hence, we get the overdetermined system of nonlinear PDEs for the functions f_1,f_2,f_3,

$$\cos f_3\frac{\partial f_1}{\partial\omega_1}+\frac{\partial f_2}{\partial\omega_1}=-R_2^{-1}\frac{\partial R_1}{\partial\omega_2},$$

$$\cos f_3\frac{\partial f_1}{\partial\omega_2}+\frac{\partial f_2}{\partial\omega_2}=R_1^{-1}\frac{\partial R_2}{\partial\omega_1},$$

$$\cos f_3\frac{\partial f_1}{\partial\omega_3}+\frac{\partial f_2}{\partial\omega_3}=0,$$

$$\cos f_2\frac{\partial f_3}{\partial\omega_1}+\sin f_2\sin f_3\frac{\partial f_1}{\partial\omega_1}=0,$$

$$\cos f_2\frac{\partial f_3}{\partial\omega_2}+\sin f_2\sin f_3\frac{\partial f_1}{\partial\omega_2}=-R_3^{-1}\frac{\partial R_2}{\partial\omega_3},$$

$$\cos f_2 \frac{\partial f_3}{\partial \omega_3} + \sin f_2 \sin f_3 \frac{\partial f_1}{\partial \omega_3} = R_2^{-1} \frac{\partial R_3}{\partial \omega_2},$$

$$\sin f_2 \frac{\partial f_3}{\partial \omega_1} - \cos f_2 \sin f_3 \frac{\partial f_1}{\partial \omega_1} = -R_3^{-1} \frac{\partial R_1}{\partial \omega_3},$$

$$\sin f_2 \frac{\partial f_3}{\partial \omega_2} - \cos f_2 \sin f_3 \frac{\partial f_1}{\partial \omega_2} = 0,$$

$$\sin f_2 \frac{\partial f_3}{\partial \omega_3} - \cos f_2 \sin f_3 \frac{\partial f_1}{\partial \omega_3} = R_1^{-1} \frac{\partial R_3}{\partial \omega_1}.$$

While integrating the above system we have to differentiate between the two cases: (1) $\sin f_3 \neq 0$ and (2) $\sin f_3 = 0$.

Case 1. Suppose the condition $\sin f_3 \neq 0$ holds. Then we can solve the above system with respect to $\partial f_a / \partial \omega_b$, $(a,b=1,2,3)$ and get

$$\begin{aligned}
\frac{\partial f_1}{\partial \omega_1} &= R_3^{-1} \frac{\partial R_1}{\partial \omega_3} \cos f_2 \operatorname{cosec} f_3, \\
\frac{\partial f_1}{\partial \omega_2} &= -R_3^{-1} \frac{\partial R_2}{\partial \omega_3} \sin f_2 \operatorname{cosec} f_3, \\
\frac{\partial f_1}{\partial \omega_3} &= R_2^{-1} \frac{\partial R_3}{\partial \omega_2} \sin f_2 \operatorname{cosec} f_3 - R_1^{-1} \frac{\partial R_3}{\partial \omega_1} \cos f_2 \operatorname{cosec} f_3, \\
\frac{\partial f_2}{\partial \omega_1} &= -R_2^{-1} \frac{\partial R_1}{\partial \omega_2} - R_3^{-1} \frac{\partial R_1}{\partial \omega_3} \cos f_2 \cot f_3, \\
\frac{\partial f_2}{\partial \omega_2} &= R_1^{-1} \frac{\partial R_2}{\partial \omega_1} + R_3^{-1} \frac{\partial R_2}{\partial \omega_3} \sin f_2 \cot f_3, \\
\frac{\partial f_2}{\partial \omega_3} &= -R_2^{-1} \frac{\partial R_3}{\partial \omega_2} \sin f_2 \cot f_3 + R_1^{-1} \frac{\partial R_3}{\partial \omega_1} \cos f_2 \cot f_3, \\
\frac{\partial f_3}{\partial \omega_1} &= -R_3^{-1} \frac{\partial R_1}{\partial \omega_3} \sin f_2, \\
\frac{\partial f_3}{\partial \omega_2} &= -R_3^{-1} \frac{\partial R_2}{\partial \omega_3} \cos f_2, \\
\frac{\partial f_3}{\partial \omega_3} &= R_2^{-1} \frac{\partial R_3}{\partial \omega_2} \cos f_2 + R_1^{-1} \frac{\partial R_3}{\partial \omega_1} \sin f_2.
\end{aligned} \tag{28}$$

From the compatibility conditions of the above system of PDEs we get the system of nonlinear differential equations for R_1, R_2, R_3:

$$(1) \quad R_1 R_2 \frac{\partial^2 R_3}{\partial \omega_1 \partial \omega_2} - R_1 \frac{\partial R_2}{\partial \omega_1} \frac{\partial R_3}{\partial \omega_2} - R_2 \frac{\partial R_1}{\partial \omega_2} \frac{\partial R_3}{\partial \omega_1} = 0,$$

$$(2)\quad R_2R_3\frac{\partial^2 R_1}{\partial\omega_2\partial\omega_3}-R_2\frac{\partial R_1}{\partial\omega_3}\frac{\partial R_3}{\partial\omega_2}-R_3\frac{\partial R_1}{\partial\omega_2}\frac{\partial R_2}{\partial\omega_3}=0,$$

$$(3)\quad R_1R_3\frac{\partial^2 R_2}{\partial\omega_1\partial\omega_3}-R_3\frac{\partial R_1}{\partial\omega_3}\frac{\partial R_2}{\partial\omega_1}-R_1\frac{\partial R_2}{\partial\omega_3}\frac{\partial R_3}{\partial\omega_1}=0,$$

$$(4)\quad R_1^2R_2^2\frac{\partial R_2}{\partial\omega_3}\frac{\partial R_3}{\partial\omega_3}+R_1^2R_3^2\frac{\partial R_2}{\partial\omega_2}\frac{\partial R_3}{\partial\omega_2}-R_2^2R_3^2\frac{\partial R_2}{\partial\omega_1}\frac{\partial R_3}{\partial\omega_1}-R_1^2R_2^2R_3\frac{\partial^2 R_2}{\partial\omega_3\,\partial\omega_3}-R_1^2R_2R_3^2\frac{\partial^2 R_3}{\partial\omega_2\,\partial\omega_2}=0, \tag{29}$$

$$(5)\quad R_1^2R_2^2\frac{\partial R_1}{\partial\omega_3}\frac{\partial R_3}{\partial\omega_3}-R_1^2R_3^2\frac{\partial R_1}{\partial\omega_2}\frac{\partial R_3}{\partial\omega_2}+R_2^2R_3^2\frac{\partial R_1}{\partial\omega_1}\frac{\partial R_3}{\partial\omega_1}-R_1^2R_2^2R_3\frac{\partial^2 R_1}{\partial\omega_3\,\partial\omega_3}-R_1R_2^2R_3^2\frac{\partial^2 R_3}{\partial\omega_1\,\partial\omega_1}=0,$$

$$(6)\quad -R_1^2R_2^2\frac{\partial R_1}{\partial\omega_3}\frac{\partial R_2}{\partial\omega_3}+R_1^2R_3^2\frac{\partial R_1}{\partial\omega_2}\frac{\partial R_2}{\partial\omega_2}+R_2^2R_3^2\frac{\partial R_1}{\partial\omega_1}\frac{\partial R_2}{\partial\omega_1}-R_1^2R_2R_3^2\frac{\partial^2 R_1}{\partial\omega_2\,\partial\omega_2}$$

$$-R_1R_2^2R_3^2\frac{\partial^2 R_2}{\partial\omega_1\,\partial\omega_1}=0.$$

Remarkably, there is no need for a direct integrating of the system of nonlinear PDEs (29), since it has been solved by Eisenhart[8] under the assumption that R_1, R_2, R_3 are independent of t. The only thing to be done is to find out in which way the temporal variable t enters into the general solution of system (26)–(29) in the case under study.

As shown above, the problem of variable separations in SE (16) reduces to integrating system (19) and, what is more, the general solution of the latter is given within the equivalence relation (9) by one of the formulas from (4). Consequently, if we fix the temporal variable t to be equal to $t_0\in\mathbf{R}$, then after integrating equations (26)–(29) we get within the equivalence relation (9) formulas (4).

In view of this fact we can solve relations (25) under $t=t_0$ with respect to $F_{ij}(\omega_i)$ [note that F_{ij} are independent of t] for each class of functions $\vec{x}=\vec{z}(\vec{\omega})$ given in (4). The results of these calculations are presented below in the form of 3×3 Stäckel matrices $\mathcal{F}_1,\dots,\mathcal{F}_{11}$, whose (i,j)th entry is the corresponding function $F_{ij}(\omega_i)$. We give the canonical forms of the matrices $\mathcal{F}_1,\dots,\mathcal{F}_{11}$ up to the choice of separation constants λ_i, $i=1,2,3$ in (11):

$$\mathcal{F}_1=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\quad \mathcal{F}_2=\begin{pmatrix}e^{2\omega_1}&-1&0\\0&1&0\\0&0&1\end{pmatrix},$$

$$\mathcal{F}_3=\begin{pmatrix}\omega_1^2&-1&0\\\omega_2^2&1&0\\0&0&1\end{pmatrix},\quad \mathcal{F}_4=\begin{pmatrix}a^2\cosh^2\omega_1&1&0\\-a^2\cos^2\omega_2&-1&0\\0&0&1\end{pmatrix},$$

$$\mathcal{F}_5=\begin{pmatrix}\omega_1^{-4}&-\omega_1^{-2}&0\\0&\cosh^{-2}\omega_2&-1\\0&0&1\end{pmatrix},$$

$$\mathcal{F}_6=\begin{pmatrix}a^2\sinh^{-4}\omega_1&-\sinh^{-2}\omega_1&-1\\a^2\cosh^{-4}\omega_2&\cosh^{-2}\omega_2&-1\\0&0&1\end{pmatrix},$$

$$\mathcal{F}_7=\begin{pmatrix} a^2\sin^{-4}\omega_1 & -\sin^{-2}\omega_1 & 1\\ -a^2\cosh^{-4}\omega_2 & \cosh^{-2}\omega_2 & -1\\ 0 & 0 & 1\end{pmatrix},$$

$$\mathcal{F}_8=\begin{pmatrix} e^{4\omega_1} & -e^{2\omega_1} & -1\\ e^{4\omega_2} & e^{2\omega_2} & -1\\ 0 & 0 & 1\end{pmatrix}, \tag{30}$$

$$\mathcal{F}_9=\begin{pmatrix} a^2\cosh^2 2\omega_1 & -a\cosh 2\omega_1 & -1\\ -a^2\cos^2 2\omega_2 & a\cos 2\omega_2 & 1\\ a^2\cosh^2 2\omega_3 & a\cosh 2\omega_3 & -1\end{pmatrix},$$

$$\mathcal{F}_{10}=\begin{pmatrix} \dfrac{\mathrm{dn}^4(\omega_1,k)}{\mathrm{sn}^4(\omega_1,k)} & -\dfrac{\mathrm{dn}^2(\omega_1,k)}{\mathrm{sn}^2(\omega_1,k)} & 1\\ -k'^4\,\mathrm{cn}^4(\omega_2,k') & k'^2\,\mathrm{cn}^2(\omega_2,k') & -1\\ k^4\,\mathrm{cn}^4(\omega_3,k) & k^2\,\mathrm{cn}^2(\omega_3,k) & 1\end{pmatrix},$$

$$\mathcal{F}_{11}=\begin{pmatrix} \omega_1^{-4} & -\omega_1^{-2} & 0\\ 0 & k'^2\,\mathrm{cn}^2(\omega_2,k') & -1\\ 0 & k^2\,\mathrm{cn}^2(\omega_3,k) & 1\end{pmatrix}.$$

With the explicit forms of the functions F_{ij}, $(i,j=1,2,3)$ in hand, we can solve (27) with respect to R_1,R_2,R_3. Inserting the result obtained into (29) and splitting by $\omega_1,\omega_2,\omega_3$ yield the final forms of the functions R_1,R_2,R_3:

(1) $R_i^2=T_i^{-1},\ \ i=1,2,3;$

(2) $R_1^2=R_2^2=T_1^{-1}e^{2\omega_1},\quad R_3^2=T_3^{-1};$

(3) $R_1^2=R_2^2=T_1^{-1}(\omega_1^2+\omega_2^2),\quad R_3^2=T_3^{-1};$

(4) $R_1^2=R_2^2=T_1^{-1}a^2(\cosh 2\omega_1-\cos 2\omega_2),\quad R_3^2=T_3^{-1};$

(5) $R_1^2=T_1^{-1}\omega_1^{-4},\quad R_2^2=R_3^2=T_1^{-1}\omega_1^{-2}\cosh^{-2}\omega_2;$

(6) $R_1^2=T_1^{-1}a^2\sinh^{-2}\omega_1(\sinh^{-2}\omega_1+\cosh^{-2}\omega_2),$

$R_2^2=T_1^{-1}a^2\cosh^{-2}\omega_2(\sinh^{-2}\omega_1+\cosh^{-2}\omega_2),$

$R_3^2=T_1^{-1}a^2\sinh^{-2}\omega_1\cosh^{-2}\omega_2;$

(7) $R_1^2=T_1^{-1}a^2\sin^{-2}\omega_1(\sin^{-2}\omega_1-\cosh^{-2}\omega_2),$

$R_2^2=T_1^{-1}a^2\cosh^{-2}\omega_2(\sin^{-2}\omega_1-\cosh^{-2}\omega_2),$

$R_3^2=T_1^{-1}a^2\sin^{-2}\omega_1\cosh^{-2}\omega_2;$

(8) $R_1^2=T_1^{-1}e^{2\omega_1}(e^{2\omega_1}+e^{2\omega_2}),\quad R_2^2=T_1^{-1}e^{2\omega_2}(e^{2\omega_1}+e^{2\omega_2}),$

$R_3^2=T_1^{-1}e^{2(\omega_1+\omega_2)};$

$$(9)\quad R_1^2=T_1^{-1}a^2(\cosh 2\omega_1-\cos 2\omega_2)(\cosh 2\omega_1+\cosh 2\omega_3),$$

$$R_2^2=T_1^{-1}a^2(\cosh 2\omega_1-\cos 2\omega_2)(\cos 2\omega_2+\cosh 2\omega_3),$$

$$R_3^2=T_1^{-1}a^2(\cosh 2\omega_1+\cosh 2\omega_3)(\cos 2\omega_2+\cosh 2\omega_3);$$

$$(10)\quad R_1^2=T_1^{-1}\left(\frac{\mathrm{dn}^2(\omega_1,k)}{\mathrm{sn}^2(\omega_1,k)}-k'^2\,\mathrm{cn}^2(\omega_2,k')\right)\left(\frac{\mathrm{dn}^2(\omega_1,k)}{\mathrm{sn}^2(\omega_1,k)}+k^2\,\mathrm{cn}^2(\omega_3,k)\right),$$

$$R_2^2=T_1^{-1}\left(\frac{\mathrm{dn}^2(\omega_1,k)}{\mathrm{sn}^2(\omega_1,k)}-k'^2\,\mathrm{cn}^2(\omega_2,k')\right)(k'^2\,\mathrm{cn}^2(\omega_2,k')+k^2\,\mathrm{cn}^2(\omega_3,k)),$$

$$R_3^2=T_1^{-1}\left(\frac{\mathrm{dn}^2(\omega_1,k)}{\mathrm{sn}^2(\omega_1,k)}+k^2\,\mathrm{cn}^2(\omega_3,k)\right)(k'^2\,\mathrm{cn}^2(\omega_2,k')+k^2\,\mathrm{cn}^2(\omega_3,k));$$

$$(11)\quad R_1^2=T_1^{-1}\omega_1^{-4},\quad R_2^2=R_3^2=T_1^{-1}\omega_1^{-2}(k'^2\,\mathrm{cn}^2(\omega_2,k')+k^2\,\mathrm{cn}^2(\omega_3,k)). \tag{31}$$

Inserting the above expressions into (28), we see that the system obtained (we denote it temporarily as $\mathcal{S}$) does not depend explicitly on t. It is in involution and hence its general solution depends on three arbitrary functions of t. As a direct check shows the functions

$$f_1=\mathrm{arccot}(\cos\gamma\cot(\tilde f_1+\beta)+\sin\gamma\cot\tilde f_3\,\mathrm{cosec}(\tilde f_1+\beta))+\alpha,$$

$$f_2=\mathrm{arccot}(\cos\tilde f_3\cot(\tilde f_1+\beta)+\sin\tilde f_3\cot\gamma\,\mathrm{cosec}(\tilde f_1+\beta))+\tilde f_2,$$

$$f_3=\arccos(\cos\gamma\cos\tilde f_3-\sin\gamma\sin\tilde f_3\cos(\tilde f_1+\beta)), \tag{32}$$

where α, β, γ are arbitrary functions of t and $\tilde f_1,\tilde f_2,\tilde f_3$ are (time-independent) solutions of the system $\mathcal{S}$ under $t=t_0$, satisfy the system $\mathcal{S}$ identically. Since formulas (32) contain three arbitrary functions of t, they give the general solution of the system $\mathcal{S}$. Substituting (32) into (26) yields

$$\frac{\partial\vec u}{\partial\omega_1}=R_1\mathcal{T}(t)\left\{\begin{matrix}\cos\tilde f_1\cos\tilde f_2-\sin\tilde f_1\sin\tilde f_2\cos\tilde f_3\\ \sin\tilde f_1\cos\tilde f_2+\cos\tilde f_1\sin\tilde f_2\cos\tilde f_3\\ \sin\tilde f_2\sin f_3\end{matrix}\right\},$$

$$\frac{\partial\vec u}{\partial\omega_2}=R_2\mathcal{T}(t)\left\{\begin{matrix}-\cos\tilde f_1\sin\tilde f_2-\sin\tilde f_1\cos\tilde f_2\cos\tilde f_3\\ -\sin\tilde f_1\sin\tilde f_2+\cos\tilde f_1\cos\tilde f_2\cos\tilde f_3\\ \cos\tilde f_2\sin f_3\end{matrix}\right\}, \tag{33}$$

$$\frac{\partial\vec u}{\partial\omega_3}=R_3\mathcal{T}(t)\left\{\begin{matrix}\sin\tilde f_1\sin\tilde f_3\\ -\cos\tilde f_1\sin\tilde f_3\\ \cos\tilde f_3\end{matrix}\right\},$$

where the matrix $\mathcal{T}(t)$ is given by formula (21).

If we choose in the system obtained $t=t_0$, then its general solution is given within the equivalence relation $\mathcal{E}$ (9) by one of the formulas from (4). In view of this fact it is not difficult to integrate system (33) and thus get formulas (20), where $h_1(t),h_2(t),h_3(t)$ are expressed via $T_1(t),T_2(t),T_3(t)$,

$$
\begin{array}{ll}
(1) & T_i = h_i^{-2}, \quad i=1,2,3; \\
(2)-(4) & T_1 = h_1^{-2}, \quad T_2 = 0, \quad T_3 = h_3^{-2}; \\
(5)-(11) & T_1 = h_1^{-2}, \quad T_2 = T_3 = 0.
\end{array} \tag{34}
$$

Case 2. Let the relation $\sin f_3 = 0$ be valid. In this case we have an analog of system (28),

$$
\begin{gathered}
\frac{\partial g}{\partial \omega_1} = -R_2^{-1} \frac{\partial R_1}{\partial \omega_2}, \quad \frac{\partial g}{\partial \omega_2} = R_1^{-1} \frac{\partial R_2}{\partial \omega_1}, \\
\frac{\partial g}{\partial \omega_3} = 0, \quad \frac{\partial R_1}{\partial \omega_3} = 0, \quad \frac{\partial R_2}{\partial \omega_3} = 0, \quad \frac{\partial R_3}{\partial \omega_1} = 0, \quad \frac{\partial R_3}{\partial \omega_2} = 0,
\end{gathered} \tag{35}
$$

with $g = \pm f_1 + f_2$ and an analog of system (29),

$$
\begin{gathered}
R_1^2 R_3^2 \frac{\partial R_1}{\partial \omega_2} \frac{\partial R_2}{\partial \omega_2} + R_2^2 R_3^2 \frac{\partial R_1}{\partial \omega_1} \frac{\partial R_2}{\partial \omega_1} - R_1^2 R_2 R_3^2 \frac{\partial^2 R_1}{\partial \omega_2 \, \partial \omega_2} - R_1 R_2^2 R_3^2 \frac{\partial^2 R_2}{\partial \omega_1 \partial \omega_1} = 0, \\
\frac{\partial R_1}{\partial \omega_3} = 0, \quad \frac{\partial R_2}{\partial \omega_3} = 0, \quad \frac{\partial R_3}{\partial \omega_1} = 0, \quad \frac{\partial R_3}{\partial \omega_2} = 0.
\end{gathered} \tag{36}
$$

A system of PDEs (35), (36) is fairly simple and is easily integrated. As a result we get a particular case of (20) with $\sin \gamma = 0$. The lemma is proved. $\triangleright$

With this result in hand it is not difficult to integrate the remaining equations from the system under study. Indeed, Eqs. (14) and (15) may be treated as algebraic equations for the functions $A_j(t,\vec{x})$, $j=1,2,3$ and $A_0(t,\vec{x})$, correspondingly.

There are two different configurations of the electromagnetic field that should be considered separately. The first one is the case of a vanishing magnetic field $\vec{H} = \mathrm{rot}\, \vec{A}$, namely, the case when

$$
A_{ix_j} = A_{jx_i}, \quad i \neq j, \quad i,j = 1,2,3. \tag{37}
$$

Provided the above equality does not hold we have a nonvanishing magnetic field $\vec{H}$.

A. The case of nonvanishing magnetic field

Let us represent the complex-valued function Q in (6) as $Q = \exp(S_1 + iS_2)$, where S_1, S_2 are real-valued functions. Now, if we take into account that the components of the vector potential $A(t,\vec{x})$ and functions $\omega_1, \omega_2, \omega_3$ are real-valued functions, then after inserting Q into (14) with the use of (5) we can split the obtained equations into real and imaginary parts:

$$
\frac{\partial S_1}{\partial x_j} \frac{\partial \omega_a}{\partial x_j} = 0, \quad a = 1,2,3; \tag{38}
$$

$$
2\left(\frac{\partial S_2}{\partial x_j} + eA_j \right) \frac{\partial \omega_a}{\partial x_j} + \frac{\partial \omega_a}{\partial t} = 0, \quad a = 1,2,3. \tag{39}
$$

In view of (18) we get from (38) the relations $S_{1x_j} = 0$, $j = 1,2,3$. Hence, $S_1 = S_1(t)$ and within the equivalence relation $\mathcal{E}$ (10) we may choose S_1 to be equal to 0. Next, making use of the gauge invariance of SE (1) we get from (39) that within the equivalence relation $\mathcal{E}$ the equalities $S_{2x_i} = 0$, $i = 1,2,3$ hold, whence $S_2 = S_2(t)$. Again, in view of (10) we may put $S_2 = 0$, which means that the factor Q may be chosen as 1. With this result system (39) reduces to the system of three linear algebraic equations for the functions A_1, A_2, A_3,

$$\omega_{at} = -2e\,\omega_{ax_i} A_i\,, \quad a = 1,2,3.$$

The determinant of this system is not equal to zero due to (18). Consequently, it has a unique solution. Making in this solution the hodograph transformation (23), we get the following expressions for A_1, A_2, A_3:

$$\vec{A} = \frac{1}{2e}\,\frac{\partial \vec{u}(t,\vec{\omega})}{\partial t}.$$

Inserting into the above formula $\vec{x} = \vec{u}(t,\vec{\omega})$ from (20) yields the explicit forms of the spacelike components of the vector potential of the electromagnetic field $A(t,\vec{x})$,

$$\vec{A}(t,\vec{x}) = \frac{1}{2e}\left(\mathcal{M}(t)(\vec{x} - \vec{w}) + \dot{\vec{w}}\right). \tag{40}$$

Here we use the designation

$$\mathcal{M}(t) = \dot{\mathcal{T}}(t)\mathcal{T}^{-1}(t) + \mathcal{T}(t)\dot{H}(t)H^{-1}(t)\mathcal{T}^{-1}(t), \tag{41}$$

where $\mathcal{T}(t)$, $H(t)$ are variable 3×3 matrices defined by formulas (21) and (22), correspondingly, $\vec{w} = (w_1(t), w_2(t), w_3(t))^T$ and the dot over a symbol means differentiation with respect to t.

Given the form (40) of the spacelike components of the vector potential $A(t,\vec{x})$, the condition for the magnetic field $\vec{H}$ not to vanish means that at least one of the expressions,

$$\dot{\alpha} + \dot{\beta}\cos\gamma, \quad \dot{\beta}\cos\alpha\sin\gamma - \dot{\gamma}\sin\alpha, \quad \dot{\beta}\sin\alpha\sin\gamma + \dot{\gamma}\cos\alpha, \tag{42}$$

does not turn into zero. Hence, in view of the identity

$$\dot{\mathcal{T}}\mathcal{T}^{-1} = \begin{pmatrix} 0 & -(\dot{\alpha} + \dot{\beta}\cos\gamma) & -(\dot{\beta}\cos\alpha\sin\gamma - \dot{\gamma}\sin\alpha) \\ \dot{\alpha} + \dot{\beta}\cos\gamma & 0 & -(\dot{\beta}\sin\alpha\sin\gamma + \dot{\gamma}\cos\alpha) \\ \dot{\beta}\cos\alpha\sin\gamma - \dot{\gamma}\sin\alpha & \dot{\beta}\sin\alpha\sin\gamma + \dot{\gamma}\cos\alpha & 0 \end{pmatrix},$$

whose validity is checked by straightforward computation, we conclude that $\mathcal{T} \neq \mathrm{const}$. At last, solving (15) yields the explicit form of A_0,

$$A_0(t,\vec{x}) = \frac{1}{e}\left(\sum_{i=1}^{3} F_{i0}(\omega_i)\,\frac{\partial \omega_i}{\partial x_j}\,\frac{\partial \omega_i}{\partial x_j} + T_0(t)\right) + iA_{ax_a} - eA_aA_a\,.$$

As the function A_0 is real valued, we have to find the real part of the right-hand side of the above equality. Making use of (40) yields

$$2eA_{ax_a} = \sum_{i=1}^{3} \frac{\dot{h}_i}{h_i}.$$

This, in its turn, gives the imaginary part of $T_0 = T_0(t)$,

$$T_0 = \tilde{T}_0 - \frac{i}{2}\sum_{i=1}^{3}\frac{\dot{h}_i}{h_i}, \quad \mathrm{Im}\,\tilde{T}_0 = 0. \tag{43}$$

Finally, we get

$$A_0(t,\vec{x}) = \frac{1}{e}\left(\sum_{i=1}^{3} F_{i0}(\omega_i)\frac{\partial\omega_i}{\partial x_j}\frac{\partial\omega_i}{\partial x_j} + \tilde{T}_0(t)\right) - eA_aA_a, \tag{44}$$

where $F_{10}(\omega_1), F_{20}(\omega_2), F_{30}(\omega_3)$ are arbitrary smooth functions, A_1, A_2, A_3 are given by (40), and the functions $\omega_a = \omega_a(t,\vec{x})$, $a=1,2,3$ are defined implicitly by formulas (20)–(22) and (4). In order to keep an exposition of the results compact, we do not present here $A_0(t,\vec{x})$ in full detail, since the corresponding expressions are too cumbersome.

Thus we have proved the following assertion.

Theorem 1: *SE (1) for the case of nonvanishing magnetic field admits separation of variables if and only if the spacelike components A_1, A_2, A_3 of the vector potential of the electromagnetic field are linear in the spatial variables and given by (40) and, furthermore, the timelike component A_0 is given by (44).*

Summing up, we conclude that conditions (40) and (44) provide separability of SE for the case of a nonvanishing magnetic field. And what is more, the solutions with separated variables are of the form (6) with $Q=1$, where the functions $\omega_1(t,\vec{x}), \omega_2(t,\vec{x}), \omega_3(t,\vec{x})$ are given implicitly by formulas (20)–(22) and (4). In fact, we have the 11 classes of vector potentials $A(t,\vec{x})$ corresponding to the 11 classes of coordinate systems $\omega_a = \omega_a(t,\vec{x})$, $a=1,2,3$. SE (1) for each class of the functions $A_0(t,\vec{x}), \vec{A}(t,\vec{x})$ defined by (40) and (44) under arbitrary $\tilde{T}_0(t), F_{a0}(\omega_a)$, and fixed arbitrary functions $\alpha(t), \beta(t), \gamma(t), w_a(t), h_a(t)$, $a=1,2,3$ separates in exactly one coordinate system. The separation equations read as (11), where the coefficients F_{ai}, $a,i=1,2,3$ are the entries of the corresponding Stäckel matrices (30), functions T_a, $a=1,2,3$ are listed in (34), and the functions T_0, F_{a0}, $a=1,2,3$ are arbitrary smooth functions defining the form of the timelike component of the vector potential $A(t,\vec{x})$ [see (44)].

Note that proper specifying the functions $F_{a0}(\omega_a)$, $a=1,2,3$, may yield additional possibilities for variable separation in the corresponding SE. What we mean is that for some particular forms of the vector potential $A(t,\vec{x})$ (40), (44) there might exist several coordinate systems (20)–(22) enabling us to separate the corresponding SE. However, the detailed study of this problem goes beyond the scope of the present paper.

As an illustration the previous section, we consider the problem of separation of variables in SE (1), where the vector potential of the electromagnetic field is of the form

$$2e\vec{A} = \begin{pmatrix} 0 & -s_1(t) & -s_2(t) \\ s_1(t) & 0 & -s_3(t) \\ s_2(t) & s_3(t) & 0 \end{pmatrix}\vec{x}, \tag{45}$$

$$eA_0 = \frac{q}{|\vec{x}|} - \frac{1}{4}\left((s_1(t)x_2 + s_2(t)x_3)^2 + (s_1(t)x_1 - s_3(t)x_3)^2 + (s_2(t)x_1 + s_3(t)x_2)^2\right).$$

Here $q=\text{const}$ and

$$s_1(t) = \dot{\alpha}(t) + \dot{\beta}(t)\cos\gamma(t),$$

$$s_2(t) = \dot{\beta}(t)\cos\alpha(t)\sin\gamma(t) - \dot{\gamma}(t)\sin\alpha(t),$$

$$s_3(t) = \dot{\beta}(t)\sin\alpha(t)\sin\gamma(t) + \dot{\gamma}(t)\cos\alpha(t),$$

where $\alpha(t), \beta(t), \gamma(t)$ are arbitrary smooth functions. Evidently, choosing $\alpha(t)=\text{const}$, $\beta(t)=\text{const}$, and $\gamma(t)=\text{const}$ yields the standard Coulomb potential.

Making use of the results of Theorem 1, we conclude that SE with potential of the form (45) separates in four coordinate systems,

$$\vec{x} = \mathcal{T}(t)\vec{z},$$

where $\mathcal{T}$ is the time-dependent 3×3 orthogonal matrix (21), and $\vec{z}$ is one of the following coordinate systems:

(1) spherical [formula 5 from (4)],

(2) prolate spheroidal II [formula 6 from (4)], where one should replace z_3 with $z_3=a(\coth\omega_1\tanh\omega_2\pm1)$,

(3) parabolic [formula 8 from (4)],

(4) conical [formula 11 from (4)].

The separation equations (11) for these cases take the form

(1)

$$
\begin{aligned}
&i\varphi_0'=-\lambda_1\varphi_0,\\
&\varphi_1''=(\lambda_1\omega_1^{-4}-\lambda_2\omega_1^{-2}+q\omega_1^{-3})\varphi_1,\\
&\varphi_2''=(\lambda_2\operatorname{sech}^2\omega_2-\lambda_3)\varphi_2,\\
&\varphi_3''=\lambda_3\varphi_3.
\end{aligned}
$$

Integrating these equations yields a family of exact solutions of SE (1) with potential (45) that are products of the exponential, confluent hypergeometric,[30] and Legendre[30] functions:

(2)

$$
\begin{aligned}
&i\varphi_0'=-\lambda_1\varphi_0,\\
&\varphi_1''=(\lambda_1a^2\sinh^{-4}\omega_1-\lambda_2\sinh^{-2}\omega_1-\lambda_3+qa\cosh\omega_1\sinh^{-3}\omega_1)\varphi_1,\\
&\varphi_2''=(\lambda_1a^2\cosh^{-4}\omega_2+\lambda_2\cosh^{-2}\omega_2-\lambda_3\mp qa\sinh\omega_2\cosh^{-3}\omega_2)\varphi_2,\\
&\varphi_3''=\lambda_3\varphi_3.
\end{aligned}
$$

Integrating these equations yields a family of exact solutions of SE (1) with potential (45) in separated form that are products of the exponential and Coulomb spheroidal functions.[31]

(3)

$$
\begin{aligned}
&i\varphi_0'=-\lambda_1\varphi_0,\\
&\varphi_1''=(\lambda_1e^{4\omega_1}-\lambda_2e^{2\omega_1}-\lambda_3+2qe^{2\omega_1})\varphi_1,\\
&\varphi_2''=(\lambda_1e^{4\omega_2}+\lambda_2e^{2\omega_2}-\lambda_3)\varphi_1,\\
&\varphi_3''=\lambda_3\varphi_3.
\end{aligned}
$$

Integrating these equations yields a family of exact solutions of SE (1) with potential (45) that are products of the exponential and confluent hypergeometric functions.

(4)

$$
\begin{aligned}
&i\varphi_0'=-\lambda_1\varphi_0,\\
&\varphi_1''=(\lambda_1\omega_1^{-4}-\lambda_2\omega_1^{-2}+q\omega_1^{-3})\varphi_1,\\
&\varphi_2''=(\lambda_2k'^2\operatorname{cn}^2(\omega_2,k')-\lambda_3)\varphi_2,\\
&\varphi_3''=(\lambda_2k^2\operatorname{cn}^2(\omega_3,k)+\lambda_3)\varphi_3.
\end{aligned}
$$

Integrating these equations yields a family of exact solutions of SE (1) with potential (45) that are products of the exponential, confluent hypergeometric, and Lamé[30] functions.

Note that these families are parametrized both by integration constants and by the three continuous spectral parameters $\lambda_i, i=1,2,3$. Under appropriate initial and boundary conditions the latter become discrete, i.e., $\lambda_i=\lambda_{in}, n=1,2,3,...$, and we get a basis for expanding arbitrary smooth solutions of SE (1) with potential (45) in a properly chosen Hilbert space (for more details, see Ref. 15).

B. The case of vanishing magnetic field

Provided condition (37) holds true, we can, without loss of generality, put $\vec{A}=0$ at the expense of the gauge invariance of SE. Representing the complex-valued function Q in (6) as $Q=\exp(S_1+iS)$, inserting into (14) and splitting the obtained equations into real and imaginary parts analogously to what has been done for the case of a nonvanishing magnetic field we arrive at the conclusion that S_1 may be chosen to be equal to zero. Furthermore, the function $S=S(t,\vec{x})$ satisfy the over-determined system of linear partial differential equations

$$2\frac{\partial S}{\partial x_j}\frac{\partial \omega_a}{\partial x_j}+\frac{\partial \omega_a}{\partial t}=0, \quad a=1,2,3. \tag{46}$$

Solving it with respect to the derivatives $\partial S/\partial x_j$, $j=1,2,3$ [which is always possible due to (18)] and making in the relations obtained the hodograph transformation (23), yields

$$2\vec{\nabla}S=\mathcal{M}(t)(\vec{x}-\vec{w})+\dot{\vec{w}}, \tag{47}$$

where $\mathcal{M}(t)$ is the 3×3 matrix (41) and $\vec{w}=(w_1(t),w_2(t),w_3(t))^T$. The direct check shows that system (47) is compatible if and only if all the expressions given in (42) vanish identically. Hence, we conclude that the matrix $\mathcal{T}$ is constant. Utilizing the invariance of SE (1) with respect to the rotation group we can choose $\mathcal{T}=I$, where I is the unit 3×3 matrix. Given this condition, (20) takes the form

$$\vec{x}=H(t)\vec{z}(\vec{\omega})+\vec{w}(t), \tag{48}$$

and, furthermore,

$$\mathcal{M}(t)=\dot{H}(t)H^{-1}(t).$$

Next, integrating (47) and taking into account the equivalence relation (10) we get

$$S=\frac{1}{2}\sum_{i=1}^{3}\left(\frac{\dot{h}_i}{h_i}\left(\frac{x_i^2}{2}-w_ix_i\right)+\dot{w}_ix_i\right). \tag{49}$$

Substituting (49) into (15) yields the form of $A_0(t,\vec{x})$,

$$eA_0(t,\vec{x})=\sum_{i=1}^{3}F_{i0}(\omega_i)\frac{\partial\omega_i}{\partial x_j}\frac{\partial\omega_i}{\partial x_j}+\tilde{T}_0(t)-\frac{1}{4}\sum_{i=1}^{3}\left(\frac{\ddot{h}_i}{h_i}x_i^2+2\left(\ddot{w}_i-\frac{\ddot{h}_i}{h_i}w_i\right)x_i+\left(\dot{w}_i-\frac{\dot{h}_i}{h_i}w_i\right)^2\right), \tag{50}$$

the function $T_0(t)$ being given by (43).

Thus, we have proved the following assertion.

Theorem 2: *Given the restriction (37), SE (1) admits a separation of variables if and only if it is gauge equivalent to SE with* $\vec{A}=\vec{0}$ *and* A_0 *being given by (50).*

Consequently, the conditions $\vec{A}=\vec{0}$ and (50) provide separability of SE for the case of a vanishing magnetic field. Furthermore, the solutions with separated variables are of the form (6) with $Q=\exp(iS)$, where $S=S(t,\vec{x})$ is given by (49), the functions $\omega_1(t,\vec{x}),\omega_2(t,\vec{x}),\omega_3(t,\vec{x})$ are given implicitly by formulas (48), (22), and (4). Again, we have the 11 classes of vector potentials $A(t,\vec{x})$ corresponding to the 11 classes of coordinate systems $\omega_a=\omega_a(t,\vec{x})$, $a=1,2,3$. SE (1) for each class of the functions $A_0(t,\vec{x}),\vec{A}(t,\vec{x})=\vec{0}$ defined by (50) under arbitrary $\tilde{T}_0(t),F_{a0}(\omega_a)$, and fixed arbitrary functions $w_a(t),h_a(t)$, $a=1,2,3$ separates in exactly one coordinate system. The separation equations read as (11), where the coefficients F_{ai}, $a,i=1,2,3$ are the entries of the corresponding Stäckel matrices (30), functions T_a, $a=1,2,3$ are listed in (34) and the functions $T_0,F_{a0},a=1,2,3$ are arbitrary smooth functions defining the form of the timelike component of the vector potential $A(t,\vec{x})$ [see (50)].

If we fix the temporal variable t to be equal to $t_0\in\mathbf{R}$ in the above obtained results, then it is not difficult to classify separable stationary Schrödinger equations for a particle interacting with the electromagnetic field $A(\vec{x})=(A_0(\vec{x}),\vec{A}(\vec{x}))$,

$$(p_ap_a+eA_0+E)\psi(\vec{x})=0, \tag{51}$$

where $p_a=i\partial/\partial x_a-eA_a$, $a=1,2,3$, $e=\text{const}$, and E is a spectral parameter, and thus recover the classical result by Eisenhart.[9]

III. SEPARATION OF VARIABLES IN THE HAMILTON–JACOBI EQUATION

It is well known that there exists a deep connection between the separation of variables in the Schrödinger and Hamilton–Jacobi equations (see, e.g., Ref. 16). The Hamilton–Jacobi equation,

$$u_t+eA_0+(u_{x_a}+eA_a)(u_{x_a}+eA_a)=0, \tag{52}$$

separates in any coordinate system providing separability of the Schrödinger equations (1) and, what is more, the inverse assertion is not true. We will make use of this connection for the sake of classifying separable Hamilton–Jacobi equations.

First we fix the usual form of the separation *Ansatz* for the Hamilton–Jacobi equation,

$$u(t,\vec{x})=S(t,\vec{x})+\varphi_0(t)+\sum_{i=1}^{3}\varphi_i(\omega_i(t,\vec{x})), \tag{53}$$

and, furthermore, fix the form of the ordinary differential equations for $\varphi_0,\varphi_1,\varphi_2,\varphi_3$,

$$\varphi_0'=-T_0(t)-T_i(t)\lambda_i,\quad \varphi_a'=(-F_{a0}(\omega_a)+F_{ai}(\omega_a)\lambda_i)^{1/2}. \tag{54}$$

Now, inserting the *Ansatz* (53) into Eq. (52), eliminating the first derivatives of the functions $\varphi_0,\varphi_1,\varphi_2,\varphi_3$ with the use of the above equations and splitting by the variables $\varphi_0,\varphi_1,\varphi_2,\varphi_3,\lambda_1,\lambda_2,\lambda_3$, we arrive at the following system of nonlinear partial differential equations for the functions $S,\omega_1,\omega_2,\omega_3$:

$$\frac{\partial\omega_i}{\partial x_a}\frac{\partial\omega_j}{\partial x_a}=0,\quad i\neq j,\quad i,j=1,2,3;$$

$$\sum_{i=1}^{3}F_{ia}(\omega_i)\frac{\partial\omega_i}{\partial x_j}\frac{\partial\omega_i}{\partial x_j}=T_a(t),\quad a=1,2,3;$$

$$2\left(\frac{\partial S}{\partial x_j}+eA_j\right)\frac{\partial\omega_a}{\partial x_j}+\frac{\partial\omega_a}{\partial t}=0,\quad a=1,2,3;$$

$$-\sum_{i=1}^{3}F_{i0}(\omega_i)\frac{\partial\omega_i}{\partial x_j}\frac{\partial\omega_i}{\partial x_j}+\frac{\partial S}{\partial t}+2eA_a\frac{\partial S}{\partial x_a}+\frac{\partial S}{\partial x_a}\frac{\partial S}{\partial x_a}-T_0(t)+eA_0+e^2A_aA_a=0. \tag{55}$$

The general solution $\vec{\omega}=\vec{\omega}(t,\vec{x})$ of the first six equations of the above system [which coincide with Eqs. (12) and (13)] can be reduced with the help of an appropriate equivalence transformation $\mathcal{E}$ to such a form that it satisfies the Laplace equation (5).

It is not difficult to become convinced of the fact that making the change of variables,

$$Q(t,\vec{x})=\exp(iS(t,\vec{x})), \tag{56}$$

in (12)–(15) yields the system that coincides with (55) with an exception of the last equation, where an additional term $-i(\Delta S+eA_{ax_a})$ appears. As shown in the previous section, this term is a function of t only and is absorbed by T_0. Consequently, all the results on variable separation for SE apply to the case of the Hamilton–Jacobi equation (52) as well.

IV. CONCLUDING REMARKS

Theorems 1–2 give a complete solution of the problem of classification of SE's (1) and (51) that are solvable within the framework of the method of separation of variables. By appropriate reductions of these results we can get the results on the separation of variables in SE's for a particle interacting with the electromagnetic field in one[24] and two[26,27] spatial dimensions. For example, in order to recover the results of Ref. 27 one has to choose $\partial\psi/\partial x_3=0$ and consider the completely and partially split coordinate systems from the list (4).

It follows from Theorem 1 that the choice of magnetic fields $\vec{H}$ allowing for variable separation in the corresponding SE is very restricted. Namely, the magnetic field should be independent of spatial variables x_1,x_2,x_3 in order to provide the separability of SE (1) into three second-order ordinary differential equations. However, if we allow for separation equations to be of lower order, then additional possibilities for variable separation in SE arise. As an example we give the vector-potential,

$$A(t,\vec{x})=(A_0(\sqrt{x_1^2+x_2^2}),0,0A_3(\sqrt{x_1^2+x_2^2})),$$

where A_0,A_3 are arbitrary smooth functions. SE (1) with this vector potential separates in the cylindrical coordinate system $t,\omega_1=\ln(\sqrt{x_1^2+x_2^2})$, $\omega_2=\arctan(x_1/x_2)$, $\omega_3=x_3$ into two first-order and one second-order ordinary differential equations. The corresponding magnetic field $\vec{H}=\operatorname{rot}\vec{A}$ is evidently x dependent.

As mentioned in the Introduction, a possibility of variable separation in SE is intimately connected to its symmetry properties. Namely, solutions with separated variables are common eigenfunctions of three mutually commuting symmetry operators of SE. For all the cases of variable separation in SE (1) these operators can be constructed in explicit form, in analogy to what has been done in Ref. 26 for the $(1+2)$-dimensional case. They are expressed in terms of the coefficients of the separation equations (11). This fact enables application of the methods of the representation theory of Lie algebras for a further analysis of special functions arising as solutions of separation equations in the spirit of the famous Bateman's project [15].

The last remark is that the technique developed in the present paper can be directly applied in order to separate variables in the Pauli equation for a particle with spin $\frac{1}{2}$ moving in the electromagnetic field and in the Fokker–Planck equation with a constant diagonal diffusion matrix.

A study of the above-mentioned problems is in progress now and will be reported in our future publications.

[1] W. Miller, Jr. and Lee A. Rubel, J. Phys. A **26**, 1901 (1993).
[2] A. Osborne and A. E. G. Stuart, J. Math. Phys. **19**, 1573 (1978).
[3] R. Z. Zhdanov, J. Phys. A **27**, L291 (1994).
[4] V. A. Galaktionov, Diff. Int. Eq. **3**, 863 (1990).
[5] R. Z. Zhdanov, J. Phys. A **28**, 3841 (1995).
[6] P. J. Olver, Proc. R. Soc. London, Ser. A **444**, 509 (1994).
[7] M. Böcher, *Die Reihenentwickelungen der Potentialtheorie* (Leipzig, Teübner, 1894).
[8] L. P. Eisenhart, Ann. Math. **35**, 284 (1934).

[9] L. P. Eisenhart, Phys. Rev. **74**, 87 (1948).
[10] M. N. Olevskii, Mat. Sb. **27**, 379 (1950) (in Russian).
[11] E. G. Kalnins, *Separation of Variables in Riemannian Spaces of Constant Curvature* (Wiley, New York, 1986).
[12] P. Winternitz, J. A. Smorodinsky, M. Uhlir̆, and I. Fris̆, Yad. Fiz. **4**, 625 (1966) [Sov. J. Nucl. Phys. **4**, 444 (1966)].
[13] A. A. Makarov, J. A. Smorodinsky, Kh. Valiev, and P. Winternitz, Nuovo Cimento A **52**, 1061 (1967).
[14] N. W. Evans, Phys. Rev. A **41**, 5666 (1990).
[15] W. Miller, Jr., *Symmetry and Separation of Variables* (Addison-Wesley, Reading, MA, 1977).
[16] T. H. Koornwinder, Lect. Notes Math. **810**, 240 (1980).
[17] C. Boyer, E. G. Kalnins, and W. Miller, Jr., J. Math. Phys. **16**, 499 (1975).
[18] C. Boyer, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **7**, 230 (1976).
[19] G. J. Reid, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **17**, 646 (1986).
[20] V. N. Shapovalov and G. G. Ekle, *Algebraic Properties of the Dirac Equation* (Elista, Izdatelstvo Kalmytskogo Universiteta, 1972) (in Russian).
[21] V. G. Bagrov, D. G. Gitman, I. M. Ternov *et al.*, *Exact Solutions of Relativistic Wave Equations* (Novosibirsk, Nauka, 1982) (in Russian).
[22] V. N. Shapovalov and N. B. Sukhomlin, Izv. Vyssh. Uchebn. Zaved. Fiz. **12**, 100 (1974) [Sov. Phys. J. **17**, 1718 (1976)].
[23] V. A. Kostelecký, V. I. Man'ko, M. M. Nieto, and D. R. Truax, Phys. Rev. A **48**, 951 (1993).
[24] C. J. Efthimiou and D. Spector, Phys. Rev. A **49**, 2301 (1994).
[25] R. Z. Zhdanov, I. V. Revenko, and W. I. Fushchych, J. Phys. A **26**, 5959 (1993).
[26] R. Z. Zhdanov, I. V. Revenko, and W. I. Fushchych, J. Math. Phys. **36**, 5506 (1995).
[27] R. Z. Zhdanov, J. Math. Phys. **38**, 1197 (1997).
[28] R. Z. Zhdanov and M. V. Lutfullin, J. Math. Phys. **39**, 6454 (1998).
[29] R. Z. Zhdanov, preprint of the Erwin Schrödinger Institute for Mathematical Physics, No. 591, Vienna, Austria, 1998.
[30] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi, *Higher Transcedental Functions* (McGraw-Hill, New York, 1953, 1955).
[31] I. V. Komarov *et al.*, *Spheroidal and Coulomb Spheroidal Functions* (Nauka, Moskow, 1976) (in Russian).

On the explicit solutions of the elliptic Calogero system

L. Gavrilov[a)]
Laboratoire Emile Picard, CNRS UMR 5580, Université Paul Sabatier, 118, Route de Narbonne, 31062 Toulouse Cedex, France

A. M. Perelomov[b)]
Max-Planck-Institut für Mathematik, Vivatsgasse 7, 53111 Bonn, Germany

(Received 25 May 1999; accepted for publication 14 June 1999)

Let $q_1, q_2, \ldots, q_N$ be the coordinates of N particles on the circle, interacting with the integrable potential $\Sigma_{j<k}^{N} \wp(q_j - q_k)$, where $\wp$ is the Weierstrass elliptic function. We show that every symmetric elliptic function in $q_1, q_2, \ldots, q_N$ is a meromorphic function in time. We give explicit formulas for these functions in terms of genus $N-1$ theta functions.
[S0022-2488(99)01512-1]

I. INTRODUCTION

The elliptic Calogero system,[1]

$$\frac{d^2}{dt^2} q_i = -\sum_{j \neq i} \wp'(q_i - q_j), \quad i = 1,2,\ldots,N \tag{1.1}$$

is a canonical Hamiltonian system, describing the motion of N particles on the circle $S^1 = \mathbb{R}/\omega\mathbf{Z}$, $\omega \in \mathbb{R}$, with Hamiltonian (energy)

$$H = \frac{1}{2} \sum_{j=1}^{N} p_j^2 + \sum_{j<k}^{N} \wp(q_j - q_k), \tag{1.2}$$

where $\wp(q) = \wp(q|\omega, \omega')$ is the Weierstrass elliptic function

$$\wp(q|\omega,\omega') = \sum_{m,n \in \mathbf{Z}} (q + m\omega + n\omega')^{-2}, \quad \omega'/\omega \notin \mathbb{R}. \tag{1.3}$$

Denote by Γ_1 the elliptic curve $\mathbb{C}/\{2\omega\mathbf{Z} + 2\omega'\mathbf{Z}\}$ with period lattice generated by 2ω and $2\omega'$. The Hamiltonian H is invariant under the obvious action of the permutation group $\mathcal{S}_n$, so the phase space of the compexified system is the cotangent bundle $T^*(S^N\Gamma_1)$ of the Nth symmetric product $S^N\Gamma_1$.

It is known that this system has two Lax representations (Refs. 1, 2, also see Ref. 3 for details). The Lax operator L defines N integrals of motion $I_k(p,q) = k^{-1}\mathrm{tr}(L^k), k = 1,\ldots,N$. It was proved in Ref. 4 that these integrals are in involution and hence this system is completely integrable in the Jacobi–Liouville sense.[5,6]

The Krichever Lax pair has a spectral parameter. This means that the equations of motion of the system under consideration are equivalent to the matrix equation

$$i\dot{L}(\lambda) = [L(\lambda), M(\lambda)], \tag{1.4}$$

a)Electronic mail: gavrilov@picard.ups-tlse.fr.
b)On leave of absence from Institute for Theoretical and Experimental Physics, 117259 Moscow, Russia. Current electronic mail: perelomo@posta.unizar.es.

where $L(\lambda)=L(p,q;\lambda)$ and $M(\lambda)=M(p,q;\lambda)$ are two matrices of order N,

$$\{L(\lambda)\}_{jk}=p_j\,\delta_{jk}+i\,(1-\delta_{jk})\,\Phi(q_j-q_k,\lambda); \tag{1.5}$$

$$\{M(\lambda)\}_{jk}=\delta_{jk}\left(\sum_{l\neq j}\wp(q_j-q_l)-\wp(\lambda)\right)+(1-\delta_{jk})\,\Phi'(q_j-q_k,\lambda); \tag{1.6}$$

$$\Phi(q,\lambda)=\frac{\sigma(q-\lambda)}{\sigma(q)\,\sigma(\lambda)}\exp(\zeta(\lambda)\,q); \tag{1.7}$$

$$\sigma(q)=q\prod_{m,n}{}'\left(1-\frac{q}{\omega_{mn}}\right)\exp\left[\frac{q}{\omega_{mn}}+\frac{1}{2}\left(\frac{q}{\omega_{mn}}\right)^2\right], \tag{1.8}$$

$$\zeta(q)=\frac{\sigma'(q)}{\sigma(q)},\quad \omega_{mn}=m\omega+n\omega'.$$

As it was shown by Krichever,[2] the equations of motion may be ''linearized'' on the Jacobian of the spectral curve

$$\Gamma^N=\{(\lambda,\mu): f(\lambda,\mu)\equiv\det(L(\lambda)-\mu I)=0\}. \tag{1.9}$$

Namely, let

$$\theta(\mathbf{z}|B)=\sum_{\mathbf{N}\in\mathbf{Z}^N}e^{\pi i\langle\mathbf{N},B\mathbf{N}\rangle+2\pi i\langle\mathbf{N},\mathbf{z}\rangle},\quad \mathbf{z}\in\mathbb{C}^N \tag{1.10}$$

be the Riemann theta function with period matrix B, where

$$B=(B_{ij}),\quad B=B^t,\quad \operatorname{Im}B>0,\quad \langle x,y\rangle=\sum_j x_jy_j,\quad i,j=1,\ldots,N.$$

It has been shown by Krichever[2] that, if B is the period matrix of the curve Γ^N, then for suitable constant vectors $U,V,W\in\mathbb{C}^N$ and for a fixed parameter $t\in\mathbb{C}$, the equation

$$\theta(Uq+Vt+W)=0,\quad q\in\mathbb{C} \tag{1.11}$$

has exactly N solutions $q=q_j(t)$ on the Jacobian Jac (Γ^N) of the curve Γ^N. The functions $q_j(t)$ provide solutions of the elliptic Calogero system (1.1). The equation (1.11) for these solutions is, however, not explicit and seems to be not well understood.

The aim of the present paper is to give ''the effectivization'' of these formulas based on the projection method by Olshanetsky and Perelomov[7,8] of explicit integration of the equations of motion in the rational and the trigonometric cases, as well as on the algebro-geometric approach of Krichever.[9,2]

II. EXPLICIT SOLUTIONS

Let Γ_N be a genus N Riemann surface which is an N-sheeted covering of an elliptic curve Γ_1,

$$\Gamma_N\xrightarrow{\pi}\Gamma_1. \tag{2.1}$$

It follows from a theorem of Weierstrass (see, for example, Refs. 10, 11, 12, and 13, Theorem 7.4) that the period matrix of the curve Γ_N in a suitable basis has the form (I,B), where $I=\operatorname{diag}(1,1,\ldots,1)$, and

$$B=\begin{pmatrix} \frac{\tau}{N} & \frac{k}{N} & 0 & \dots & 0 \\ \frac{k}{N} & b_{22} & b_{23} & \dots & b_{2N} \\ 0 & b_{32} & b_{33} & \dots & b_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & b_{N2} & b_{N3} & \dots & b_{NN} \end{pmatrix} \tag{2.2}$$

for a suitable positive integer k. Consider the Riemann theta function $\theta(x,\mathbf{t})=\theta(x,\mathbf{t}|B)$, where $\mathbf{t}=(t_1,t_2,\dots,t_{N-1})$, $(x,\mathbf{t})\in\mathbb{C}^N$. We have

$$\theta(x+1,\mathbf{t})=\theta(x,\mathbf{t}),\theta(x+\tau,\mathbf{t})=e^{-2\pi iNx-\pi iN\tau}\theta(x,\mathbf{t}),i=\sqrt{-1} \tag{2.3}$$

and therefore for any fixed $\mathbf{t}$ the function $\theta(x,\mathbf{t})$ is an elliptic theta function of order N.[14] In particular it has exactly N zeros on $\Gamma_1=\mathbb{C}/\{\mathbf{Z}+\tau\mathbf{Z}\}$ which we denote by $x_i(\mathbf{t})$, $i=1,2,\dots,N$.

Lemma 2.1: *The following identity holds:*

$$\frac{\partial^2}{\partial x^2}\log\ \theta(x,\mathbf{t}|B)=\sum_{i=1}^{N}\wp(x-x_i(\mathbf{t})|\tau)+N\frac{\theta_1'''(0)}{3\theta_1'(0)},$$

where[15]

$$\theta_1(x|\tau)=\theta\begin{bmatrix}1/2\\1/2\end{bmatrix}(x|\tau).$$

Proof: The relations

$$\theta_1(x+1)=-\theta_1(x),\theta_1(x+\tau)=-e^{-2\pi ix-\pi i\tau}\theta_1(x) \tag{2.4}$$

compared to (2.3) imply that

$$\left(\frac{\theta(x,\mathbf{t})}{\prod_{i=1}^{N}\theta_1(x-x_i(\mathbf{t}))}\right)^2 \tag{2.5}$$

is a meromorphic function in x on Γ_1 which has no poles, and hence it is a constant (in x). It follows that

$$\frac{\partial^2}{\partial x^2}\log\frac{\theta(x,\mathbf{t})}{\prod_{i=1}^{N}\theta_1(x-x_i(\mathbf{t}))}\equiv 0.$$

Finally we use that

$$\wp(x)=-\frac{\partial^2}{\partial x^2}\log\sigma(x),\qquad \theta_1(x)=c\ \exp(\eta x^2)\ \sigma(x), \tag{2.6}$$

where

$$\eta=-\frac{\theta_1'''(0)}{6\theta_1'(0)}$$

and c is a suitable constant.[15] □

Theorem 2.2: *The Krichever curve Γ^N is an N-sheeted covering of an elliptic curve $\Gamma_1 = \mathbb{C}/\{\mathbf{Z}+\tau\mathbf{Z}\}$. There exists a canonical homology basis and a normalized basis of holomorphic one-forms on Γ^N, such that the corresponding period matrix of Γ_N takes the form (I,B), where $I=\mathrm{diag}(1,1,\ldots,1)$, and*

$$B=\begin{pmatrix} \frac{\tau}{N} & \frac{1}{N} & 0 & \ldots & 0 \\ \frac{1}{N} & b_{22} & b_{23} & \ldots & b_{2N} \\ 0 & b_{32} & b_{33} & \ldots & b_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & b_{N2} & b_{N3} & \ldots & b_{NN} \end{pmatrix}. \tag{2.7}$$

In the same basis the vectors U and V in (1.11) *read*

$$U=(1,0,\ldots,0), V=(0,V_2,\ldots,V_N). \tag{2.8}$$

A direct proof (without using the Weierstrass theorem) of the above Theorem will be given in the last section. From now on we make the convention that $2\omega=1$ so the period lattice of Γ_1 is

$$\mathbf{Z}+\tau\mathbf{Z}, \quad \tau=2\omega'/2\omega=2\omega'.$$

Corollary 2.3: The symmetric functions

$$f_k(t)=\sum_{i=1}^{N} \wp^{(k)}(q_i(t))$$

are meromorphic in t. Explicit formulas for them are obtained from Lemma 2.1,

$$f_0(t)=\frac{\partial^2}{\partial x^2}\log\,\theta(x,\mathbf{t})\big|_{x=0}-N\frac{\theta_1'''(0)}{3\theta_1'(0)},$$

$$f_k(t)=(-1)^k\frac{\partial^{k+2}}{\partial x^{k+2}}\log\,\theta(x,\mathbf{t})\big|_{x=0}, \quad k>0,$$

where

$$\mathbf{t}=(V_2t+W_2,V_3t+W_3,\ldots,V_Nt+W_N).$$

Our next construction is motivated by Refs. 7, 8, and 2. Let us define the function

$$F(x,t)=\prod_{j=1}^{N}\frac{\sigma(x-q_j(t))}{\sigma(x)\sigma(q_j(t))}=[\theta_1'(0)]^{-N}]\prod_{j=1}^{N}\frac{\theta_1(x-q_j(t))}{\theta_1(x)\theta_1(q_j(t))}, \quad \sum_{j=1}^{N} q_j(t)=0, \tag{2.9}$$

where

$$q_j(t), \quad t\in\mathbb{C}, \quad j=1,2,\ldots,N,$$

is a solution of the elliptic Calogero system.

Lemma 2.4: $F(x,t)$ is a meromorphic function in x on Γ_1 and meromorphic function in t on $\mathbb{C}$, explicitly given by

$$F(x,t)=[-\theta_1'(0)]^{-N}]\frac{\theta(Ux+Vt+W)}{\theta_1(x)^N\theta(Vt+W)}. \tag{2.10}$$

Proof: We already noted that the function (2.5) is a constant in *x*, and hence

$$\frac{\theta(x,\mathbf{t})}{\Pi_{i=1}^N\theta_1(x-x_i(\mathbf{t}))}\equiv\frac{\theta(0,\mathbf{t})}{\Pi_{i=1}^N\theta_1(-x_i(\mathbf{t}))}.$$

This combined with (2.8) gives

$$\frac{\Pi_{i=1}^N\theta_1(x-q_i(t))}{\Pi_{i=1}^N\theta_1(q_i(t))}=(-1)^N\frac{\theta(Ux+Vt+W)}{\theta(Vt+W)}.$$

□

The expansion of $F(x,t)$ on the basis of first order theta functions in x defines $(N-1)$ meromorphic functions in the variables $q_1,\ldots,q_N$ which are also meromorphic functions in t with only simple poles. Hence we can take them as new ''good'' variables. The expansion of $F(x,t)$ can be obtained by making use of the addition formulas for elliptic functions. In the case $N=2$, we have the following ''addition formula''[15]

$$F(x,t)=-\frac{\sigma(x-q)\,\sigma(x+q)}{\sigma^2(x)\sigma^2(q)}=\wp(x)-\wp(q), \tag{2.11}$$

which generalizes for arbitrary N in the following way

Lemma 2.5: For any $\mathbf{q}=(q_1,q_2,\ldots,q_N),x$, *such that* $\Sigma q_j=0$ *define*

$$F(x,\mathbf{q})=\prod_{j=1}^N\frac{\sigma(x-q_j)}{\sigma(x)\sigma(q_j)}, \tag{2.12}$$

$$\Delta(\mathbf{q})=(N-1)!\det\begin{vmatrix}1 & \wp(q_1) & \wp'(q_1) & \ldots & \wp^{(N-3)}(q_1)\\ 1 & \wp(q_2) & \wp'(q_2) & \ldots & \wp^{(N-3)}(q_2)\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & \wp(q_{N-1}) & \wp'(q_{N-1}) & \ldots & \wp^{(N-3)}(q_{N-1})\end{vmatrix}. \tag{2.13}$$

The following identity holds:

$$F(x,\mathbf{q})\Delta(\mathbf{q})\equiv\det\begin{vmatrix}1 & \wp(x) & \wp'(x) & \ldots & \wp^{(N-2)}(x)\\ 1 & \wp(q_1) & \wp'(q_1) & \ldots & \wp^{(N-2)}(q_1)\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & \wp(q_{N-1}) & \wp'(q_{N-1}) & \ldots & \wp^{(N-2)}(q_{N-1})\end{vmatrix}. \tag{2.14}$$

Remark: The substitution $x=q_N$ in (2.14) gives the following addition formula for the Weierstrass $\wp$-function:

$$\det\begin{vmatrix}1 & \wp(q_1) & \wp'(q_1) & \ldots & \wp^{(N-2)}(q_1)\\ 1 & \wp(q_2) & \wp'(q_2) & \ldots & \wp^{(N-2)}(q_2)\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & \wp(q_N) & \wp'(q_N) & \ldots & \wp^{(N-2)}(q_N)\end{vmatrix}\equiv 0. \tag{2.15}$$

Proof: For fixed $\mathbf{q}=(q_1,q_2,\ldots,q_N)$ the functions in the left and right-hand side of the identity (2.14) are meromorphic in x on the elliptic curve Γ_1. Both of them have a pole of order N at $x=0$ and simple zeros at $x=q_1,\ldots,q_{N-1}$. It follows that their ratio is a first order elliptic function, and hence a constant in x. To compute this constant we use that $\sigma(x)=x+\ldots$, $\wp(x)=1/x^2+\ldots$, and then compare the Laurent series of the two functions in a neighborhood of $x=0$. □

Note finally that if for fixed $\mathbf{q}$ and $\tilde{\mathbf{q}}$ holds $F(x,\mathbf{q})\equiv F(x,\tilde{\mathbf{q}})$, then up to a permutation $\mathbf{q}=\tilde{\mathbf{q}}$. Therefore there is a one-to-one correspondence between the coefficients of $\wp^k(x)$ in the expansion of $F(x,\mathbf{q})$, and the points of the $(N-1)$th symmetric power of the elliptic curve $\Gamma_1\setminus\{0\}$. In particular every meromorphic function on this symmetric power is a rational function in the above coefficients. This implies the following:

Corollary 2.6: Let $f(x)$ be a meromorphic function on the elliptic curve Γ_1, and let S be a symmetric rational function in $N-1$ variables. If $q_1(t),q_2(t),\ldots,q_N(t)$, $\Sigma q_i\equiv 0$ is a solution of the elliptic Calogero system, then $S(f(q_1(t),f(q_2(t)),\ldots,f_{N-1}(q_{N-1}(t)))$ is a meromorphic function in t.

The further analysis of the explicit formulas for the solutions of the elliptic Calogero system can be based on Lemma 2.4, Lemma 2.5, and the identity

$$F(x,t)\equiv F(x,\mathbf{q}(t)).$$

Consider the seemingly trivial case of two particles $(N=2)$. Let us give first an explanation of the Krichever formula (1.11) for the solutions $q_1(t)=-q_2(t)$. Put $q_1-q_2=q$ and $p_1=-p_2=p$. The Hamiltonian H becomes $H(p,q)=p^2+\wp(q)$, and the reduced Hamiltonian system is

$$\frac{d}{dt}q=2p,\ \ \frac{d}{dt}p=-\wp'(q),\ \ (q,p)\in T^*\Gamma_1. \tag{2.16}$$

The Lax matrix L is

$$L(\lambda)=\begin{pmatrix} p & i\Phi(q,\lambda) \\ i\Phi(-q,\lambda) & -p \end{pmatrix}$$

and the corresponding spectral polynomial

$$\det(L(\lambda)-\mu I)=\mu^2-p^2+\Phi(q,\lambda)\Phi(-q,\lambda)=\mu^2-p^2+\wp(\lambda)-\wp(q)=\mu^2+\wp(\lambda)-H(p,q)$$

defines a spectral curve

$$\Gamma_2=\{(\mu,\lambda):\mu^2+\wp(\lambda)=h\}.$$

Suppose that $h=H(p,q)$ is fixed in such a way, that the meromorphic function $\wp(\lambda)-h$ has two distinct zeros on Γ_1. The spectral curve Γ_2 is a double ramified covering over the elliptic curve Γ_1 with projection map $\pi:\Gamma_2\to\Gamma_1:(\mu,\lambda)\to\lambda$. It follows that Γ_2 is a genus two curve with holomorphic differentials

$$\omega_1=d\lambda,\quad \omega_2=\frac{d\lambda}{\mu}.$$

On the other hand Γ_2 is identified to the orbit

$$\{(p,q)\in T^*\Gamma_1=:H(p,q)=h\}$$

under the map

$$(p,q)\to(\mu,\lambda).$$

Consider further the embedding of the orbit Γ_2 into its Jacobian variety Jac(Γ_2)

$$\Gamma_2 \to \mathrm{Jac}(\Gamma_2): P \to \left(\int_{P_0}^{P} d\lambda, \int_{P_0}^{P} \frac{d\lambda}{\mu} \right). \tag{2.17}$$

By the Riemann theorem,[16] the curve $\Gamma_2 \subset \mathrm{Jac}(\Gamma_2)$ defines a divisor which coincides, up to addition of a constant, with the Riemann theta divisor $\Theta \subset \mathrm{Jac}(\Gamma_2)$ on the Jacobian variety $\mathrm{Jac}(\Gamma_2)$.

Let $(p(t),q(t))$ be a solution of the elliptic Calogero system, with initial condition $(p(t_0),q(t_0)) = P_0$. Taking into consideration that

$$\frac{d\lambda}{\mu} = 2\,dt, \quad d\lambda = dq, \quad (\lambda,\mu) \in \Gamma_2,$$

formula (2.17) takes the form

$$T^*\Gamma_1 = \mathbb{C} \times \Gamma_1 \ni (p(t),q(t)) \to (2t - 2t_0, q(t) - q(t_0)) \in \mathrm{Jac}(\Gamma_2). \tag{2.18}$$

It follows that there exist constant vectors $\mathbf{a},\mathbf{b},\mathbf{c} \in \mathbb{C}^2$ such that

$$\theta(\mathbf{a}q(t) + \mathbf{b}t + \mathbf{c}) \equiv 0. \tag{2.19}$$

Of course these constants depend on the choice of symplectic homology basis and the choice of normalized basis of holomorphic one-forms. Namely, let a,b be two loops on Γ_1, such that $\pi^{-1}(a) = \{a_1,a_2\}$, $\pi^{-1}(b) = \{b_1,b_2\}$, where a_i,b_j represent an integer symplectic homology basis on Γ_2: $a_i \circ b_j = \delta_{ij}$, $a_i \circ a_j = 0$, $b_i \circ b_j = 0$. Then,

$$\int_{a_1} d\lambda = \int_{a_2} d\lambda, \quad \int_{b_1} d\lambda = \int_{b_2} d\lambda,$$

$$\int_{a_1} \frac{d\lambda}{\mu} = -\int_{a_2} \frac{d\lambda}{\mu}, \quad \int_{b_1} \frac{d\lambda}{\mu} = -\int_{b_2} \frac{d\lambda}{\mu}.$$

If we define a new symplectic basis

$$\tilde{a}_1 = a_1 + a_2, \tilde{a}_2 = b_1 - b_2, \tilde{b}_1 = b_1, \tilde{b}_2 = a_2$$

and normalize the two holomorphic one-forms as

$$d\lambda \to \frac{d\lambda}{\int_{\tilde{a}} \pi^* d\lambda} = \frac{d\lambda}{2\int_a d\lambda}, \quad \frac{d\lambda}{\mu} \to \frac{d\lambda/\mu}{\int_{\tilde{a}_2} d\lambda/\mu},$$

then the period matrix of Γ_2 takes the form

$$\begin{pmatrix} 1 & 0 & \tau_1/2 & 1/2 \\ 0 & 1 & 1/2 & \tau_2/2 \end{pmatrix},$$

where

$$\tau_1 = \frac{\int_b d\lambda}{\int_a d\lambda}, \tau_2 = \frac{\int_{a_2} d\lambda/\mu}{\int_{b_1} d\lambda/\mu}.$$

This, together with (2.18) implies that

$$\mathbf{a}=\left(\frac{1}{\int_{\tilde{a}_1} d\lambda},0\right)=\left(\frac{1}{2\int_a d\lambda},0\right),\mathbf{b}=\left(0,\frac{1}{\int_{b_1} d\lambda/\mu}\right).$$

Finally the vector $\mathbf{c}$ is arbitrary and plays the role of initial condition. The function $F(x,t)$ defined in (2.9) takes the form

$$F(x,t)=-\frac{\sigma(x-q(t))\,\sigma(x+q(t))}{\sigma^2(x)\sigma^2(q(t))} \tag{2.20}$$

and hence[15,17]

$$F(x,t)=\wp(x)-\wp(t). \tag{2.21}$$

So the elliptic function $\wp(q|\omega,\omega')$, and also

$$\mathrm{sn}^2(q,k)\sim\frac{\theta_1^2(q|k)}{\theta_4^2(q|k)},\quad \mathrm{cn}^2(q,k)\sim\frac{\theta_2^2(q|k)}{\theta_4^2(q|k)},\quad \mathrm{dn}^2(q,k)\sim\frac{\theta_3^2(q|k)}{\theta_4^2(q|k)} \tag{2.22}$$

are ''good'' variables (in the sense that they are meromorphic in t). The equation of motion for them takes a very simple form. We get

$$\mathrm{sn}^2(q,k)=1-a^2+a^2\,\mathrm{sn}^2(\gamma t,\tilde{k}), \tag{2.23}$$

where

$$a^2=\frac{h-1}{h},\quad \gamma=2(h-k^2),\quad \tilde{k}^2=\frac{h-1}{h-k^2}\,k^2. \tag{2.24}$$

One can easily show that the even functions $\mathrm{cn}\,(q,k)$ and $\mathrm{dn}\,(q,k)$ (but not $\mathrm{sn}\,(q,k)$) are ''good'' variables and we get as in[3]

$$\mathrm{cn}\,(q,k)=\alpha\,\mathrm{cn}\,(\gamma t,\tilde{k}), \tag{2.25}$$

$$\mathrm{dn}\,(q,k)=\beta\,\mathrm{dn}\,(\gamma t,\tilde{k}),\ b=(k/\tilde{k})a. \tag{2.26}$$

III. REDUCTION OF THETA FUNCTIONS

The reduction theory was elaborated by Weierstrass (see, for example, Ref. 10) and Poincaré.[11,12] Consider first the case $N=2$. The Riemann theta function associated with the Riemann matrix (2.7) has the form,

$$\theta(z_1,z_2)=\sum_{n_i,n_j}\exp\{i\pi[B_{ij}n_in_j+2n_jz_j]\},\quad i,j=1,2, \tag{3.1}$$

where

$$B_{11}=\tau_1/2,\quad B_{22}=\tau_2/2,\quad B_{12}=B_{21}=1/2.$$

A straightforward computation gives

$$\theta(z_1,z_2)=\sum_{n_1,n_2}\exp\left\{i\pi\left[\tau_1\frac{n_1^2}{2}+n_1n_2+\tau_2\frac{n_2^2}{2}+2n_1z_1+2n_2z_2\right]\right\}$$
$$=\sum_{k_1,n_2\in\mathbf{Z}}\exp\{i\pi[2\tau_1k_1^2+4k_1z_1]\}\exp\left\{i\pi\left[\tau_2\frac{n_2^2}{2}+2n_2z_2\right]\right\}$$
$$+\sum_{k_1,n_2\in\mathbf{Z}}\exp\left\{i\pi\left[2\tau_1\left(k_1+\frac{1}{2}\right)^2+4\left(k_1+\frac{1}{2}\right)z_1\right]\right\}\exp\left\{i\pi\left[\tau_2\frac{n_2^2}{2}+2\left(n_2+\frac{1}{2}\right)z_2\right]\right\}$$
$$=\theta_3(2z_1|2\tau_1)\ \theta_3\left(z_2\left|\ \frac{\tau_2}{2}\right.\right)+\theta_2(2z_1|2\tau_1)\ \theta_4\left(z_2\left|\ \frac{\tau_2}{2}\right.\right),$$

where $\theta_1,\theta_2,\theta_3$, and θ_4 are defined by formulas,

$$\theta_1(z|\tau)=\theta\begin{bmatrix}1/2\\1/2\end{bmatrix}(z|\tau)=2q^{1/4}\sum_{n=1}^{\infty}(-1)^nq^{n(n+1)}\sin[(2n+1)\pi z];\tag{3.2}$$

$$\theta_2(z|\tau)=\theta\begin{bmatrix}1/2\\0\end{bmatrix}(z|\tau)=2q^{1/4}\sum_{n=1}^{\infty}q^{n(n+1)}\cos[(2n+1)\pi z];\tag{3.3}$$

$$\theta_3(z|\tau)=\theta\begin{bmatrix}0\\0\end{bmatrix}(z|\tau)=1+2\sum_{n=1}^{\infty}q^{n^2}\cos(2\pi nz);\ q=\exp(i\pi\tau);\tag{3.4}$$

$$\theta_4(z|\tau)=\theta\begin{bmatrix}0\\1/2\end{bmatrix}(z,\tau)=1+2\sum_{n=1}^{\infty}(-1)^nq^{n^2}\cos(2\pi nz).\tag{3.5}$$

So in this case, the equation $\theta(z_1,z_2)=0$ is equivalent either to

$$A\ \mathrm{dn}(2z_1|4\tau_1)\ \mathrm{dn}(z_2|\tau_2)+\mathrm{cn}(2z_1|4\tau_1)=0,\tag{3.6}$$

or to

$$A\ \mathrm{dn}(2z_2|4\tau_2)\ \mathrm{dn}(z_1|\tau_1)+\mathrm{cn}(2z_2|4\tau_2)=0,\tag{3.7}$$

where

$$A=\frac{\theta_3(0|4\tau_1)\ \theta_3(0|\tau_2)}{\theta_2(0|4\tau_1)\ \theta_4(0|\tau_2)}\tag{3.8}$$

or

$$\mathrm{dn}(z_1|\tau_1)=B\ \mathrm{dn}(2iz_2+K|\tilde{\tau}_2).\tag{3.9}$$

Let us give also a more symmetric form of the theta divisor for this case,

$$\mathrm{dn}(2z_1,k_1)\ \mathrm{dn}(2z_2,k_2)+\mathrm{dn}(2z_1,k_1)\ \mathrm{cn}(2z_2,k_2)+\mathrm{cn}(2z_1,k_1)\ \mathrm{dn}(2z_2,k_2)$$
$$-\mathrm{cn}(2z_1,k_1)\ \mathrm{cn}(2z_2,k_2)=0.\tag{3.10}$$

Using the constraint $\theta(\mathbf{a}x+\mathbf{b}t+\mathbf{c})=0$ and taking $z_1=q$, $z_2=(1/2)K+i\gamma t$, we get once again (2.25) and (2.26).

Consider now the case of arbitrary N. Let $\theta(z_1,z_2,\ldots,z_N|B)$ be the Riemann theta function with period matrix as in Theorem 2.2. In a quite similar way we get

$$\theta(z_1,z_2,\ldots,z_N)=\sum_{j=0}^{N-1}\theta_j(z_1)\,\Theta_j(z_2,\ldots,z_N),\tag{3.11}$$

where

$$\theta_j(z_1)=\theta\begin{bmatrix} j/N \\ 0 \end{bmatrix}(Nz_1|N^2\tau_1),\tag{3.12}$$

$$\Theta_j(z_2,\ldots,z_N)=\Theta\begin{bmatrix} 0 & 0 & \cdots & 0 \\ j/N & 0 & \cdots & 0 \end{bmatrix}(z_2,\ldots,z_N|\hat{B}).\tag{3.13}$$

In the above formula $\hat{B}$ is the right lower $(N-1)\times(N-1)$ minor of B (2.7), and the theta functions with fractional characteristics are defined, for example, in Refs. 19,18,14,13. A reduction formula similar to (3.11), but containing N^2 terms, can be found in Ref. 13, Corollary 7.3.

IV. GEOMETRY OF THE SPECTRAL CURVE

In this section we prove Theorem 2.2.

Let Γ_N be a genus N Riemann surface which is an N-sheeted covering of an elliptic curve Γ_1

$$\Gamma_N\xrightarrow{\pi}\Gamma_1.\tag{4.1}$$

Choose two loops a,b which generate the fundamental group $\pi_1(\Gamma_1,P)$, $P\in\Gamma_1$, and denote $\check{\Gamma}_1=\Gamma_1\setminus\{a\cup b\}$. Let us suppose for simplicity that the ramification points of the projection map π are distinct. Connect further these ramification points by non-intersecting arcs $\gamma_i\subset\check{\Gamma}_1$. The set $\pi^{-1}(\check{\Gamma}_1\setminus\cup_i\gamma_i)$ is a disjoint union of N "sheets." To reconstruct the topological covering (4.1) we have to indicate how the opposite borders of the cuts γ_i are glued, as well how the opposite borders of the (preimages of the) cuts a and b respectively are glued together. Thus there is only a finite number of topologically different coverings (4.1). It may be shown that the Krichever curve (1.9) is of genus at most N, and for generic (p_i,q_i) its genus is exactly N. The projection map π (4.1) is defined then by $\pi(\mu,\lambda)=\lambda$. From now on we shall always assume that (p_i,q_i) are generic. In the case when Γ_N is the genus N Krichever spectral curve (1.9), and Γ_1 is the elliptic curve with half periods ω,ω', the covering (4.1) has a number of special properties.

To prove (2.7) we shall need the following:

Proposition 4.1: Let Γ_N *be the Krichever curve* (1.9). *There exist loops* $a,b\in\pi_1(\Gamma_1,P)$ *such that, if* $\check{\Gamma}_1=\Gamma_1\setminus\{a\cup b\}$, $\partial\check{\Gamma}_1=a\circ b\circ a^{-1}\circ b^{-1}$, *then (i)* $\pi^{-1}(\check{\Gamma}_1)$ *is connected; (ii)* $\pi^{-1}(\partial\check{\Gamma}_1)$ *has exactly N connected components.*

On its hand the above proposition implies the following:

Proposition 4.2: There exists loops $a,b\in\pi_1(\Gamma_1,P)$, $P\in\Gamma_1$, *such that*

$$\pi^{-1}(a)=\{a_1,a_2,\ldots,a_N\},\pi^{-1}(b)=\{b_1,b_2,\ldots,b_N\},$$

where a_i,b_i *represent a symplectic homology basis of* $H_1(\Gamma_N,\mathbf{Z})$, $a_i\circ b_j=\delta_{ij}$.

Proof of (2.7) assuming Proposition 4.2: Let $d\lambda$ be the holomorphic one-form on Γ_1. Then the pullback $\pi^*d\lambda$ of $d\lambda$ is a holomorphic one-form on Γ_N and we have

$$\int_{a_i}\pi^*d\lambda=\int_a d\lambda,\int_{b_i}\pi^*d\lambda=\int_b d\lambda.$$

Choose the following new integer homology basis of Γ_N:

$$\tilde{a}_1=a_1+a_2+\ldots a_N,\quad \tilde{b}_1=b_1,$$

$$\tilde{a}_2 = Nb_1 - b_1 - b_2 - \ldots - b_N, \quad \tilde{b}_2 = a_2,$$

and

$$\tilde{a}_i = b_i - b_1, \quad \tilde{b}_i = a_2 - a_i, \quad i = 3, \ldots, N.$$

This is also a symplectic basis of $H_1(\Gamma_N, \mathbf{Z})$, as

$$\sum_{i=1}^{N} \tilde{a}_i \wedge \tilde{b}_i = \sum_{i=1}^{N} a_i \wedge b_i.$$

Let $\omega_1, \omega_2, \ldots, \omega_N$ be a basis of holomorphic one-forms on Γ_N, such that

$$\omega_1 = \frac{d\lambda}{\int_{\tilde{a}_1} d\lambda}, \int_{\tilde{a}_i} \omega_j = \delta_{ij}.$$

Then $B = (\int_{\tilde{b}_j} \omega_i)_{i,j}^{N,N}$ is a symmetric matrix with positive definite imaginary part, such that

$$\int_{\tilde{b}_1} \omega_1 = \frac{\tau}{N}, \int_{\tilde{b}_2} \omega_1 = \frac{1}{N}, \int_{\tilde{b}_i} \omega_1 = 0, \ i \geqslant 3$$

which completes the proof of 2.7.

Proof of Proposition 4.1: First of all let us note that if the claim holds for some Krichever curve, then it holds for any Krichever curve. Indeed, the space of all such curves is parameterized by $\mathbb{C}^{N-1}$ (the first integrals of the integrable Hamiltonian system (1.4)) and hence it is connected. Let us fix a generic point (p_i, q_i), $i = 1, 2, \ldots, N$. It is enough to prove now our proposition for at least one pair of half-periods ω, ω', for example for $|\omega|, |\omega'| \sim \infty$.

Let us represent $\check{\Gamma}_1 \subset \mathbb{C} = \mathbb{P}^1 \setminus \infty$ as the interior of the period parallelogram formed by 2ω and $2\omega'$. When $|\omega| \to \infty$, $|\omega'| \to \infty$, the boundary. $\partial \check{\Gamma}_1 = a \circ b \circ a^{-1} \circ b^{-1}$ tends to $\infty \in \mathbb{P}^1$, and $\check{\Gamma}_1$ tends to $\check{\Gamma}_1^{\infty} = \mathbb{C}$. In a similar way we define the ''limit'' curve $\check{\Gamma}_N^{\infty}$ which is explicitly described in the following way. When $|\omega| \to \infty$, $|\omega'| \to \infty$, then on any compact set the Weierstrass functions $\sigma(q), \zeta(q), \wp(q)$ tend to $q, 1/q, 1/q^2$ respectively, and hence the function $\Phi(q,\lambda)$ tends to

$$\frac{q - \lambda}{q\lambda} \exp(q/\lambda).$$

Denote the corresponding ''limit'' Lax matrix (1.5) by $L^{\infty}(\lambda)$. The curve $\check{\Gamma}_N^{\infty}$ is the affine curve

$$\{(\lambda, \mu) : \det(L^{\infty}(\lambda) - \mu I_N) = 0\}$$

completed with N distinct points corresponding to $\lambda = 0$. The last holds true if and only if the ramification points of the projection map π (4.1) tend to some values different from $\lambda = 0$ (it is easy to check that this is a generic condition on (p_i, q_i)). We shall also suppose that these values are different from $\lambda = \infty$ (another generic condition). Under these restrictions one may prove (as in Ref. 2) that $\check{\Gamma}_N^{\infty}$ is a Riemann sphere, with N punctures (the preimages of $\lambda = \infty$). We obtain thus a map $\pi: \mathbb{P}^1 \to \mathbb{P}^1$ with $2N-2$ ramification points different from $\lambda = 0, \infty$. The fact that $\pi^{-1}(\mathbb{C})$ is connected implies the part (i) of the proposition, and the fact that $\pi^{-1}(\infty)$ is a disjoint union of N points implies (ii).

Proof of Proposition 4.2: Let us represent $\check{\Gamma}_N$ by a graph with N vertices. A vertex corresponds to a sheet (see the beginning of this section), an edge connects two vertices if and only if the corresponding sheets have a common ramification point. Proposition 4.1 (i) implies that the graph is connected, and (ii) that each sheet contains an even number of ramification points. As the total number of ramification points is $2N-2$ and each point belongs to exactly two sheets, then in addition the graph of $\check{\Gamma}_N$ is simply connected.

Consider now the punctured curve

$$\tilde{\Gamma}_1=\check{\Gamma}_1\setminus\cup_i R_i,$$

where $R_i, i=1,\ldots,2N-2$ are the ramification points of π. The fundamental group $\pi_1(\tilde{\Gamma}_1,P)$ has a natural representation in the permutation group S_n. Namely, when a point $Q\in\Gamma_1$ makes one turn along a loop $a\in\pi_1(\tilde{\Gamma}_1,P)$, the set $\pi^{-1}(P)=\cup_{i=1}^N P_i$ is transformed to itself. If the loops a and b induce the identity permutation, then $\pi^{-1}(a)$, $\pi^{-1}(b)$ are disjoint unions of N loops with obvious intersections, which implies Proposition 4.2. If not, we shall modify a and b.

Let $c\in\pi_1(\tilde{\Gamma}_1,P)$ be a loop which makes one turn around some ramification point of π. Then c induces a permutation which exchanges the two sheets containing the ramification point. As the graph of $\check{\Gamma}_N$ is connected then all such transpositions generate the permutation group S_n. Thus for suitable c the loop $a\circ c$ induces the identity permutation. It remains to substitute $a\to a\circ c$ and to note that $a=a\circ c$ in $\pi_1(\Gamma_1,P)$.

Proof of (2.8) (compare to Ref. 13, Theorem 7.14): Let $0\in\Gamma_1$ be the pole of $\wp(z)$. We denote

$$\pi^{-1}(0)=\{\infty_1,\infty_2,\ldots,\infty_N\},\quad \infty_i\in\Gamma_N.$$

In a neighborhood of each point ∞_i on the Krichever curve $\{(\lambda,\mu):f(\lambda,\mu)=0\}$ the meromorphic function μ has the following Laurent expansion:[9]

$$\mu=-\frac{1}{\lambda}+O(1),\quad i=1,2,\ldots,N-1,$$

$$\mu=\frac{N-1}{\lambda}+O(1).$$

It follows that if

$$\omega_j=f_j(P)d\lambda,\quad P=(\lambda,\mu)\in\Gamma_N$$

is a differential of first kind (i.e., holomorphic) on Γ_N, then $\mu\omega_j$ is a differential of third kind with simple poles at ∞_i. The sum of the residues of $\mu\omega_j$ is equal to

$$\sum_{i=1}^{N-1} f_j(\infty_i)-(N-1)f_j(\infty_N)=0. \tag{4.2}$$

Let Ω be a differential of second kind on Γ_N with a single pole at ∞_N. Such is for example the differential

$$\frac{\mu^2-\wp(\lambda)}{\dfrac{\partial f}{\partial\mu}(\lambda,\mu)}d\lambda.$$

If moreover Ω is normalized as

$$\int_{\tilde{a}_i}\Omega=0,$$

then it is well known that the vector V is collinear to

$$\left(\int_{\tilde{b}_1} \Omega, \int_{\tilde{b}_2} \Omega, \dots, \int_{\tilde{b}_N} \Omega \right)$$

(see, for example, Ref. 13). Equivalently, if we apply the reciprocity law to the differentials of second and first kind Ω, ω_i, we get that V is colinear to

$$(f_1(\infty_N), f_2(\infty_N), \dots, f_N(\infty_N)).$$

On the other hand

$$\tilde{a}_1 = a_1 + a_2 + \dots + a_N = \pi^{-1}(a)$$

and hence

$$\int_{\tilde{a}_1} \omega_i = \sum_{k=1}^{N} \int_a f_i(\lambda, \mu_k) d\lambda,$$

where $(\lambda, \mu_k) \in \Gamma_N$ are the N preimages of $\lambda \in \Gamma_1$. It is clear that $\Sigma_{k=1}^N f_i(\lambda, \mu_k)$ is a single-valued function on Γ_1. As ω_i is a holomorphic differential on Γ_N and $d\lambda$ is the holomorphic differential on Γ_1, then $\Sigma_{k=1}^N f_i(\lambda, \mu_k)$ is a holomorphic function on Γ_1 and hence a constant. As ω_i is a normalized basis of holomorphic forms, then $\int_{\tilde{a}_1} \omega_i = 0$ for $i \geqslant 2$, and hence

$$\sum_{k=1}^{N} \int_a f_i(\lambda, \mu_k) d\lambda = \sum_{k=1}^{N} f_i(\lambda, \mu_k) \int_a d\lambda \equiv 0, \quad (\lambda, \mu) \in \Gamma_N, \quad i \geqslant 2.$$

Therefore,

$$\sum_{k=1}^{N} f_i(\lambda, \mu_k) \equiv 0, \quad i \geqslant 2,$$

which combined with (4.2) implies that

$$f_i(\infty_N) = 0, \quad i \geqslant 2$$

and hence the vector V is colinear to (1,0, . . . ,0). In fact V is equal to this vector, because $q_i(t) \in \Gamma_1 = \mathbb{C}/\{\mathbf{Z} + \tau \mathbf{Z}\}$. Finally, we may always suppose that $U = (0, U_2, \dots, U_N)$. Indeed the Calogero system (1.1) is invariant under the translation

$$q_i \rightarrow q_i - V_1 t.$$

□

ACKNOWLEDGMENTS

We are grateful to B.A. Dubrovin for the useful remarks. One of authors (A.M.P.) would like to thank the Laboratoire Emile Picard at the University of Toulouse III and Max-Planck-Institut für Mathematik, Bonn for the hospitality.

[1] F. Calogero, ''Exactly solvable one-dimensional many-body problems,'' Lett. Nuovo Cimento **13**, 411–415 (1975).
[2] I. M. Krichever, ''Elliptic solutions of the Kadomtsev–Petviashvili equation and integrable systems of particles,'' Funct. Anal. Appl. **14**, 282–290 (1980).
[3] A. M. Perelomov, *Integrable Systems of Classical Mechanics and Lie Algebras* (Birkhäuser, Berlin, 1990), Vol. I.
[4] A. M. Perelomov, ''Completely integrable classical systems connected with semisimple Lie algebras,'' Lett. Math. Phys. **1**, 531–540 (1977).
[5] C. Jacobi, ''Vorlesungen über Dynamik,'' which Jacobi gave at 1842–1843 at Königsberg University; these lectures were edited by Clebsch and published in Berlin, Reimer, 1866 .

[6] J. Liouville, ''Note sur les équations de la dynamique,'' J. Math. Pures Appl. **20**, 137–138 (1855).
[7] M. A. Olshanetsky and A. M. Perelomov, ''Geodesic flows on symmetric spaces of zero curvature and explicit solution of the generalized Calogero model for the classical case,'' Funct. Anal. Appl. **10**, 237–239 (1976).
[8] M. A. Olshanetsky and A. M. Perelomov, ''Explicit solutions of some completely integrable Hamiltonian systems,'' Funct. Anal. Appl. **11**, 66–68 (1977).
[9] I. M. Krichever, ''Rational solutions of the Kadomtsev–Petviashvili equation and integrable systems of *n* particles on line,'' Funct. Anal. Appl. **12**, 59–61 (1978).
[10] S. Kowalevski, ''Über die Reduction einer bestimmten Klasse Abel'scher Integrale 3[ten] Ranges auf elliptische Integrale,'' Acta Math. **4**, 392–414 (1884).
[11] H. Poincaré, ''Sur la réduction des intégrales abéliennes,'' Bull. Soc. Math. France **12**, 124–143 (1884).
[12] H. Poincaré, ''Sur les fonctions abéliennes,'' Am. J. Math. **8**, 289–342 (1886).
[13] B. A. Belokolos, A. I. Bobenko, V. Z. Enol'skii, A. R. Its, and V. B. Matveev, *Algebro-Geometric Approach to Nonlinear Integrable Equations* (Springer, Berlin, 1994).
[14] B. A. Dubrovin, ''Theta functions and nonlinear equations,'' Russian Math. Surveys **36**, 11–80 (1981).
[15] H. Bateman and A. Erdelyi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1955), Vol. 3.
[16] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1978).
[17] Whittaker and Watson, *Course of Modern Analysis* (Cambridge University Press, Cambridge, 1927).
[18] A. Krazer, *Lehrbuch der Thetafunctionen* (Teubner, Leipzig, 1903).
[19] S. Koizumi, ''The equations defining Abelian varieties and modular forms,'' Math. Ann. **242**, 127–145 (1976).

SU(N) skyrmions and harmonic maps

T. Ioannidou[a]
Institute of Mathematics, University of Kent at Canterbury, Canterbury CT2 7NF, United Kingdom

B. Piette[b] and W. J. Zakrzewski[c]
Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, United Kingdom

(Received 6 July 1999; accepted for publication 14 September 1999)

Harmonic maps from S^2 to CP^{N-1} are introduced to construct low-energy configurations of the SU(N) Skyrme model. We show that one of such maps gives an exact, topologically trivial, solution of the SU(3) model. We study various properties of these maps and show that, in general, their energies are only a little higher than the energies of the corresponding SU(2) embeddings. Moreover, we show that the baryon and energy densities of the SU(3) configurations with baryon number $B=3-6$ are more symmetrical than their SU(2) analogs, thus suggesting that there exist solutions of the model with these symmetries. We also show that any SU(2) solution embedded into the SU(4) Skyrme model becomes a topologically trivial solution of this model.
[S0022-2488(99)02712-7]

I. INTRODUCTION

The Skyrme model is now well established as an effective classical theory used to describe nuclei[1,2] for which the field, which describes pions and other pseudoscalar mesons, is valued in SU(N). To have finite-energy configurations, one must require that the field $U(\vec{x},t)$ goes to a constant matrix, say I, at spatial infinity: $U\to I$ as $|\vec{x}|\to\infty$. This effectively compactifies the three-dimensional Euclidean space into S^3 and hence implies that the field configurations of the Skyrme model can be considered as maps from S^3 into SU(N).

In the static limit, the energy of the Skyrme model with a mass term is

$$E=\frac{1}{12\pi^2}\int_{R^3}\left\{-\frac{1}{2}\,\mathrm{tr}(\partial_i U\,U^{-1})^2-\frac{1}{16}\,\mathrm{tr}[\partial_i U\,U^{-1},\partial_j U\,U^{-1}]^2-\frac{m_\pi^2}{2}\,\mathrm{tr}(U^{-1}+U-2I)\right\}d^3\vec{x}, \tag{1}$$

when the energy is expressed in the same units as the baryon number. In this case, the static fields U obey the equation

$$\partial_i\left(\partial_i U\,U^{-1}-\frac{1}{4}[\partial_j U\,U^{-1},[\partial_j U\,U^{-1},\partial_i U\,U^{-1}]]\right)-\frac{m_\pi^2}{4}(U-U^{-1})=0. \tag{2}$$

As the third homotopy class of SU(N) is Z, every field configuration is characterized by an integer:

$$B=\frac{1}{24\pi^2}\int_{R^3}\epsilon_{ijk}\,\mathrm{tr}(\partial_i U\,U^{-1}\,\partial_j U\,U^{-1}\,\partial_k U\,U^{-1})d^3\vec{x}, \tag{3}$$

[a]Electronic mail: T.Ioannidou@ukc.ac.uk
[b]Electronic mail: B.M.A.G.Piette@durham.ac.uk
[c]Electronic mail: W.J.Zakrzewski@durham.ac.uk

which is interpreted as the baryon number;[2,3] therefore, the lowest-energy state in the $B=1$ sector can be identified with the (classical) nucleon. From the mathematical point of view these field configurations represent three-dimensional topological solitons.

Up to now most of the studies involving the Skyrme model have concentrated on its SU(2) version and its embeddings into SU(N). The simplest nontrivial classical solution involves a single skyrmion ($B=1$) and has already been discussed by Skyrme.[1] The energy density of this solution is radially symmetric and, as a result, using the so-called hedgehog ansatz one can reduce (2) to an ordinary differential equation that can then be solved numerically. Many solutions with $B>1$ have also been computed numerically and in all cases the energy densities of these solutions are very symmetrical and exhibit interesting patterns (cf. Battye *et al.*[4] and references therein).

Last year, Houghton *et al.*[5] showed that using rational maps from S^2 to S^2 one can easily construct field configurations for the SU(2) model that are close to being solutions of the model: they have energies slightly higher than the energies of the exact solutions found numerically, but the symmetries of the baryon and energy densities are the same. When these configurations are used as initial conditions in a relaxation program, the fields do not change much as they evolve toward the exact solutions.

However, so far, very little has been done for the SU(N) model when $N>2$—except the obvious SU(2) embeddings. The first example of a *nonembedding* configuration for a higher group was the $SO(3)$ soliton, which corresponds to a bound system of two skyrmions, and that was found using the chiral field ansatz by Balachandran *et al.*[6] Another configuration, with a large SU(3) strangeness content, was found by Kopeliovich *et al.*[7] However, all other known skyrmion configurations seem to have been the embeddings of the solutions of the SU(2) model.

In the following sections, we construct baryon configurations for the SU(N)($N>2$) Skyrme model, with energies slightly higher than the corresponding SU(2) embbedings—using harmonic maps from S^2 to CP^{N-1}. However, the difference between the two energies, in most cases, is less than the energy of a pion. In most cases, the configurations are more symmetrical than their SU(2) embeddings, implying that the SU(N) Skyrme model has solutions with a larger symmetry than those of the SU(2) one. These solutions might have energies higher than the SU(2) embeddings and might be unstable, but our construction provides an upper bound on their energies. As these solutions have an excitation energy smaller than the pion mass, they will have to be considered in any attempt to quantize the low-energy modes of the Skyrme model.

II. SU(2) EMBEDDINGS

Any solution of the SU(N') model is automatically a solution of any SU(N) model when $N'<N$—as a simple embedding. The energy, the baryon number, and all other properties are unchanged by this operation, and so such embeddings have the same properties as the original fields. However, less obvious embeddings exist in which new fields have different properties from the original ones. In particular, it can be can shown that any solution of the SU(2) model generates a solution of the SU(4) one.

A special feature of the SU(2) field is that it can be written as

$$U=\vec{\phi}\cdot\vec{\tau}, \tag{4}$$

where $\vec{\tau}=(1,i\sigma_1,i\sigma_2,i\sigma_3)$ and the σ's stand for the Pauli matrices. The unitarity of U requires that $\vec{\phi}\cdot\vec{\phi}=1$. Introducing the notation that $\vec{\phi}_i=\partial_i\vec{\phi}$ and $\vec{\phi}_{ij}=\partial_i\,\partial_j\vec{\phi}$, it is easy to see that the energy density of this SU(2) field is

$$\mathcal{E}_2=(\vec{\phi}_i\cdot\vec{\phi}_i)+\tfrac{1}{2}[(\vec{\phi}_i\cdot\vec{\phi}_i)^2-(\vec{\phi}_i\cdot\vec{\phi}_j)^2], \tag{5}$$

and the equations of motion that follow from (2) are

$$\vec{\phi}_{ii}+\vec{\phi}_{ii}(\vec{\phi}_j\cdot\vec{\phi}_j)+\vec{\phi}_i(\vec{\phi}_{ij}\cdot\vec{\phi}_j)-\vec{\phi}_{ij}(\vec{\phi}_i\cdot\vec{\phi}_j)-\vec{\phi}_j(\vec{\phi}_{ii}\cdot\vec{\phi}_j)-\vec{\phi}(\vec{\phi}\cdot\vec{\phi}_{ii})$$
$$-\vec{\phi}[(\vec{\phi}\cdot\vec{\phi}_{ii})(\vec{\phi}_j\cdot\vec{\phi}_j)-(\vec{\phi}\cdot\vec{\phi}_{ij})(\vec{\phi}_i\cdot\vec{\phi}_j)]=0. \tag{6}$$

A SU(4) field can be constructed out of any S^3 field $\vec{\phi}$ as

$$U=U_0(I-2\vec{\phi}\otimes\vec{\phi}^\dagger), \tag{7}$$

where U_0 is a constant matrix. Note that, as can easily be seen by direct expansion of (7), $\det U=-\det U_0$ and so by choosing U_0 to be a constant matrix of determinant -1 we see that U is unitary. In this case, $U^{-1}=(I-2\vec{\phi}\otimes\vec{\phi}^\dagger)U_0^{-1}$ and the U_0's cancel in (2).

To derive the equation that the field $\vec{\phi}$ must satisfy so that (7) is a solution of (2), with $m_\pi=0$, we note that $\vec{\phi}\cdot\vec{\phi}=1$ gives

$$\partial_j U\, U^{-1}=2\vec{\phi}_j\otimes\vec{\phi}^\dagger-2\vec{\phi}\otimes\vec{\phi}_j^\dagger, \tag{8}$$

and so (2) becomes

$$2\partial_i[\vec{\phi}\otimes\vec{\phi}_j^\dagger(\vec{\phi}_j\cdot\vec{\phi}_i)-\vec{\phi}\otimes\vec{\phi}_i^\dagger(\vec{\phi}_j\cdot\vec{\phi}_j)+\vec{\phi}_i\otimes\vec{\phi}^\dagger(\vec{\phi}_j\cdot\vec{\phi}_j)-\vec{\phi}_j\otimes\vec{\phi}^\dagger(\vec{\phi}_i\cdot\vec{\phi}_j)]+\partial_i(\vec{\phi}_i\otimes\vec{\phi}^\dagger-\vec{\phi}\otimes\vec{\phi}_i^\dagger)$$
$$=0. \tag{9}$$

It can easily be shown that Eqs. (9) and (6) are equivalent, implying that $\vec{\phi}$ satisfies the equation of the SU(2) Skyrme model (9). Thus, any solution of the SU(2) model can be transformed into a solution of the SU(4) model by the embedding (7).

The solutions obtained this way are topologically trivial since their baryon density vanishes identically. Moreover, their energy density is four times larger than the corresponding energy of the original SU(2) field (5), i.e., $\mathcal{E}_4=4\mathcal{E}_2$.

This suggests that these particular SU(4) solutions may be interpreted as states corresponding to $2B$ skyrmions and $2B$ antiskyrmions, where B is the baryon number of the original SU(2) solution. Incidentally, a similar situation arises in two-dimensions, where any B solitonic solution of the CP^1 model gives a topologically trivial solution of the CP^2 model, which can be interpreted as a bound state of $2B$ solitons and $2B$ antisolitons.[8]

III. HARMONIC MAPS

The idea of the Houghton *et al.* construction[5] is to separate the radial and the angular dependence of the fields by using an appropriate ansatz. Using the polar coordinates (r,θ,ϕ) in $\mathbf{R}^3$, our SU(N) generalization of the SU(2) Houghton *et al.*'s rational map ansatz is

$$U(r,\theta,\phi)=e^{2ig(r)(P-I/N)}=e^{-2ig(r)/N}(I+(e^{2ig}-1)P), \tag{10}$$

where P is a $N\times N$ Hermitian projector that depends only on the angular variables (θ, ϕ) and $g(r)$ is the radial profile function. Note that the matrix P is a harmonic map from S^2 into CP^{N-1}. Hence it is convenient, rather than using the polar coordinates, to map the sphere onto the complex plane via a stereographic projection and, instead of θ and ϕ, use the complex coordinate ξ and its conjugate. Thus, P can be written as

$$P(V)=\frac{V\otimes V^\dagger}{|V|^2}, \tag{11}$$

where V is an N-component complex vector (dependent on ξ and $\bar{\xi}$).

For (10) to be well defined at the origin, the radial profile function $g(r)$ has to satisfy $g(0)=\pi$ while the boundary value $U\rightarrow I$ at $r=\infty$ requires that $g(\infty)=0$. An attractive feature of the ansatz (10) is that it leads to a simple expression for the energy density that can be successively minimized with respect to the parameters of the projector P and then with respect to the shape of the profile function $g(r)$. This procedure is then expected to give good approximations to multi-skyrmion field configurations.

Moreover, we will show that this method not only allows us to find such field configurations but also gives us an exact *nontopological* solution of the SU(3) Skyrme model. We will also present some upper bounds on the energy of some multiskyrmion field configurations in the SU(N) model (with radially symmetric energy density distribution). In what follows, we restrict our attention to the case $m_\pi = 0$.

To find an exact solution of the SU(3) model, we put (10) into (2) and obtain

$$\begin{aligned}
&-\frac{2i}{r^2}\partial_r(g_r r^2)\left(\frac{1}{N}-P\right)-\frac{i}{2r^2}\partial_r(g_r|A|^2)(1+|\xi|^2)^2([P_{\bar\xi},PP_\xi]+[P_\xi,PP_{\bar\xi}])\\
&+\frac{\bar{A}|A|^2}{16r^4}(1+|\xi|^2)^2\{\partial_{\bar\xi}((1+|\xi|^2)^2[P_\xi,[P_{\bar\xi}P_\xi]])-\partial_\xi((1+|\xi|^2)^2[P_{\bar\xi},[P_{\bar\xi},P_\xi]])\}\\
&+\frac{|A|^4}{16r^4}(1+|\xi|^2)^2\{\partial_{\bar\xi}((1+|\xi|^2)^2[P_\xi P,[P_{\bar\xi},P_\xi]])-\partial_\xi((1+|\xi|^2)^2[P_{\bar\xi}P,[P_{\bar\xi},P_\xi]])\}\\
&+\frac{1}{r^2}(1+g_r^2)(1+|\xi|^2)^2\left[\bar{A}P_{\xi\bar\xi}+\frac{|A|^2}{2}(\partial_\xi(P_{\bar\xi}P)+\partial_{\bar\xi}(P_\xi P))\right]=0,
\end{aligned}\tag{12}$$

where $A=e^{-2ig}-1$.

Moreover, the energy (1) simplifies to

$$E=\frac{1}{3\pi}\int dr\left(A_N g_r^2 r^2+2\mathcal{N}\sin^2 g(1+g_r^2)+\mathcal{I}\frac{\sin^4 g}{r^2}\right),\tag{13}$$

where

$$A_N=\frac{2}{N}(N-1),\tag{14}$$

$$\mathcal{N}=\frac{i}{2\pi}\int d\xi\, d\bar\xi\, \mathrm{tr}(|\partial_\xi P|^2),\tag{15}$$

$$\mathcal{I}=\frac{i}{4\pi}\int d\xi\, d\bar\xi(1+|\xi|^2)^2\,\mathrm{tr}([\partial_\xi P,\partial_{\bar\xi}P]^2).\tag{16}$$

As the integrals $\mathcal{N}$ and $\mathcal{I}$ in (13) are independent of r, we can minimize (13) by first minimizing $\mathcal{N}$ and $\mathcal{I}$ as functions of P and then with respect to the profile function g.

However, since $\mathcal{N}$ is the expression for the energy of the two-dimensional Euclidean CP^2 sigma model, all classical solutions contain the so-called self-dual solutions, instantons or holomorphic maps from S^2 into CP^{N-1}—introduced in Ref. 9 and are given by the projector P of the form (11) with $V=f(\xi)$. In this case, the energy $\mathcal{N}$ is the degree of f, i.e., the degree of the highest-order polynomial in ξ among the components of f after all their common factors have been canceled out.

By a Bogomolny-type argument it can be shown that

$$\begin{aligned}
E&=\frac{1}{3\pi}\int dr\left[\left(g_r r\sqrt{A_N}+\sqrt{\mathcal{I}}\frac{\sin^2 g}{r}\right)^2+2N\sin^2 g(1+g_r)^2-2g_r\sin^2 g(2N+\sqrt{A_N\mathcal{I}})\right]\\
&\geqslant\frac{1}{3}(2N+\sqrt{A_N\mathcal{I}}).
\end{aligned}\tag{17}$$

Finally, the baryon number for this ansatz is

$$B=\frac{i}{\pi^2}\int d\xi\, d\bar{\xi}\,\mathrm{tr}(P[\partial_\xi P,\partial_{\bar{\xi}}P])\int_0^\infty dr\,\sin^2 g\, g_r=\frac{i}{2\pi}\int d\xi\, d\bar{\xi}\,\mathrm{tr}(P[\partial_{\bar{\xi}}P,\partial_\xi P]), \tag{18}$$

which is the topological charge of the two-dimensional CP^{N-1} sigma model. In this calculation we have used the boundary conditions on the profile function g.

In the next two sections, we will show that this ansatz gives us interesting low-energy field configurations of the SU(N) Skyrme model, which are not the SU(2) embeddings. To minimize E we will, first of all, fix the baryon number $\mathcal{N}=B$ of the configurations in which we are interested. We will then minimize $\mathcal{I}$ over all maps of degree B and then derive a second-order differential equation for g by minimizing the energy (13), treating $\mathcal{N}$ and $\mathcal{I}$ as parameters.

IV. SU(3) EXACT SOLUTION

When the projector P is analytic, i.e., is of the form

$$P_0=P(f)=\frac{ff^\dagger}{|f|^2}, \tag{19}$$

where f is a holomorphic vector (i.e., whose entries are functions only of ξ) then P_0 satisfies the equation

$$P_0\,\partial_\xi P_0=0,\quad \partial_\xi P_0\,P_0=\partial_\xi P_0, \tag{20}$$

i.e., the self-dual equations of the two-dimensional CP^{N-1} sigma models.[8]

Following Ref. 10, we define a new operator P_+, by its action on any vector $v\in C^N$, as

$$P_+v=\partial_\xi v-\frac{v(v^\dagger\,\partial_\xi v)}{|v|^2}, \tag{21}$$

and then define further vectors $P_+^k v$ by induction: $P_+^k v=P_+(P_+^{k-1}v)$.

Next, we note the following useful properties of $P_+^k f$ when f is holomorphic:

$$(P_+^k f)^\dagger P_+^l f=0,\quad k\neq l, \tag{22}$$

$$\partial_{\bar{\xi}}(P_+^k f)=-P_+^{k-1}f\,\frac{|P_+^k f|^2}{|P_+^{k-1}f|^2},\quad \partial_\xi\left(\frac{P_+^{k-1}f}{|P_+^{k-1}f|^2}\right)=\frac{P_+^k f}{|P_+^{k-1}f|^2}. \tag{23}$$

These properties either follow directly from the definition of the P_+ operator or are very easy to prove.[8]

It is also convenient to define projectors corresponding to the family of $P_+^k f$ vectors:

$$P_k=P(P_+^k f),\quad k=0,\ldots,N-1, \tag{24}$$

which always satisfy the relation $\Sigma_{k=0}^{N-1}P_k=I$.

Taking $P=P_k$, for a given k, and using the above properties we observe that all the terms in (12), except the first one, can be gathered into one term if and only if

$$\frac{|P_+f|^2}{|f|^2}\equiv\frac{\mathcal{K}}{(1+|\xi|^2)^2}, \tag{25}$$

where $\mathcal{K}$ is a constant. Moreover, in the SU(2) case, the projectors P_0 and P_1 satisfy the relation $P_0+P_1=I$, and for $f=(1,\xi)^t$ all the terms in (12) are proportional to one common matrix—thus giving a second-order differential equation for the profile function g. This means that the Skyrme field (10), in the case when g satisfies its equation, is an *exact* solution of Eq. (12). A little thought shows that this is the well-known hedgehog solution.

Unfortunately, this discussion does not generalize to higher SU(N) groups—with one exception. Note that in the SU(3) model, if we take $P(V)=P_1$ and use the fact that $P_0+P_1+P_2=I$, all the matrix terms in Eq. (12) become proportional to each other, leading to a second order differential equation for the profile function, if and only if

$$\frac{|P_+^2 f|^2}{|P_+ f|^2}+\frac{|P_+ f|^2}{|f|^2}\equiv\frac{\tilde{\mathcal{K}}}{(1+|\xi^2|)^2}, \tag{26}$$

where $\tilde{\mathcal{K}}$ is a constant. This last condition is satisfied if

$$f=(1,\sqrt{2}\xi,\xi^2)^t. \tag{27}$$

Thus, by taking $P=P_1$ for f of the form (27), and requiring g to satisfy the equation

$$g_{rr}\left(\frac{1}{3}+2\,\frac{\sin^2 g}{r^2}\right)+\frac{2}{3}\,\frac{g_r}{r}+\frac{\sin 2g}{r^2}\left(g_r^2-1-\frac{\sin^2 g}{r^2}\right)=0, \tag{28}$$

we see that (10) is an *exact* solution of the SU(3) model.

For this solution, the parameters in the energy density can be evaluated analytically; we have $A_N=\frac{4}{3}$, $\mathcal{N}=4$, $\mathcal{I}=4$, and the total energy is found to be $E=3.861$.

To understand what this solution corresponds to, we calculate the topological charge of this configuration and get

$$B=\frac{i}{\pi}\int d\xi\, d\bar{\xi}\left(\frac{|P_+^2 f|^2}{|P_+ f|^2}-\frac{|P_+ f|^2}{|f|^2}\right), \tag{29}$$

which due to the conditions (22), (23), and (26) is identically zero.

Although the baryon density is identically zero, the solution itself is nontrivial. This follows from the fact that the CP^2 sigma model harmonic map P_1 corresponds to a mixture of two solitons and two antisolitons. Thus, it seems reasonable to interpret this solution as describing a bound state of two skyrmions and two antiskyrmions and as such to be unstable, i.e., correspond to a saddle point of the energy. However, let us emphasize once again; this field configuration is a genuine (although topologically trivial) solution of the SU(3) Skyrme model.

It is easy to see that this new field configuration has an energy density distribution shaped like a shell (i.e., is radially symmetric). To see this, note that for this solution, $\mathrm{tr}(|\partial_\xi P|^2)$ and $\mathrm{tr}([\partial_\xi P,\partial_{\bar{\xi}} P]^2)$ that appear in (15) and (16), are proportional to $(1+|\xi|^2)^{-2}$ and $(1+|\xi|^2)^{-4}$, respectively; demonstrating this symmetry. The radial energy density of this solution is given in Fig. 1, and one sees that, indeed, it corresponds to a shell.

V. SU(N) RADIALLY SYMMETRIC SKYRME FIELDS

In general, our method does not give us further solutions, but it is a matter of simple algebra to show that the condition (25) is true for any $N\geqslant 2$ when the modulus of the vector f is some power of $(1+|\xi|^2)$; i.e., for

$$f=(f_0,\ldots,f_i,\ldots,f_{N-1})^t:\quad f_i=\xi^i\sqrt{C_{i+1}^{N-1}}, \tag{30}$$

where C_{i+1}^{N-1} denotes the binomial coefficients. Note that in this case, the constant $\mathcal{K}$ in (25) is equal to the degree of the vector f: i.e., $\mathcal{K}=n$.

Using the condition (25) the integrals involving $P=P_0$ in the energy (13) can be evaluated analytically,

$$\mathcal{N}=n,\quad \mathcal{I}=(N-1)^2=n^2. \tag{31}$$

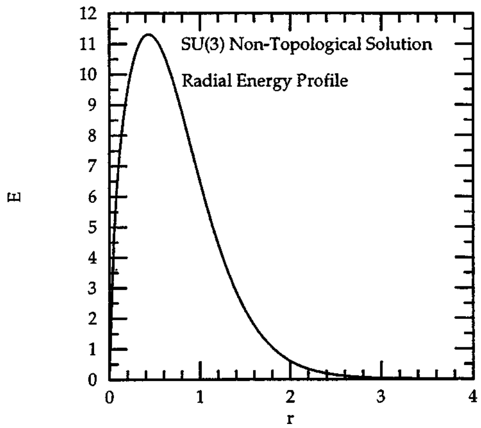

FIG. 1. Energy profile of the nontopological solution.

Due to the analyticity of the projector P_0, the baryon number of this field is $B=n$, i.e., the degree of f.

When $m_\pi \neq 0$ the energy (13) for the ansatz (10) becomes

$$E=\frac{1}{3\pi}\int dr\left\{A_N g_r^2 r^2+2\mathcal{N}\sin^2 g(1+g_r^2)+\mathcal{I}\frac{\sin^4 g}{r^2}\right.$$
$$\left.+m_\pi^2 r^2\left[(N-1)\left(1-\cos\left(\frac{2g}{N}\right)\right)+1-\cos\left(\frac{(N-1)2g}{N}\right)\right]\right\}. \tag{32}$$

Minimizing (32) given (31) leads to the following equation for the profile function:

$$g_{rr}\left(A_N+2n\frac{\sin^2 g}{r^2}\right)+2A_N\frac{g_r}{r}+\frac{\sin 2g}{r^2}\left(n(g_r^2-1)-n^2\frac{\sin^2 g}{r^2}\right)$$
$$-m_\pi^2\left(\frac{N-1}{N}\right)\left[\sin\left(\frac{2g}{N}\right)+\sin\left(\frac{(N-1)2g}{N}\right)\right]=0, \tag{33}$$

where A_N is given by (14).

Solving (33) to determine g and then calculating the energy of the configuration we find that, for small m_π, the energy for these configurations is a little higher than the energy of the SU(2) embedded ansatz with the same baryon number B when the mass is zero. However, when the mass increases, the picture changes.

We have looked at field configurations corresponding to $B=2-4$ for the SU(2) embeddings and for the SU(N) spherical symmetric fields (30) where $N=B+1$, and studied the dependence of their energies on m_π. In all cases at low values of the mass the embeddings have lower energies while as the mass increases the energies increase. However, as the embedding energies increase faster for all low B there is a value of m_π above which the embedding energy is higher. This value of m_π is quite large and it increases with the increase of B.

The results are summarized in Table I, which gives values of the energy for different values of the mass; while in Fig. 2, we present the plots of the dependence on m_π of the energies for the embeddings and for the radially symmetric fields (30). Note that for low values of m_π the energy per skyrmion of the harmonic ansatz configuration is lower than the energy of a single skyrmion—demonstrating the existence of bound states. For larger values of m_π our new field configurations become less massive than the embeddings and for even larger values of m_π either the bound states do not exist or the approximation through harmonic maps becomes unreliable.

TABLE I. Mass dependence of the energy for the radially symmetric configurations in the SU(2)–SU(5) models.

m_π	SU(2) $E_{B=1}$	SU(2) $E_{B=2}$	SU(3) $E_{B=2}$	SU(2) $E_{B=3}$	SU(4) $E_{B=3}$	SU(2) $E_{B=4}$	SU(5) $E_{B=4}$
0	1.232	2.416	2.444	3.553	3.644	4.546	4.838
0.2	1.247	2.444	2.472	3.594	3.683	4.597	4.886
1	1.416	2.795	2.808	4.125	4.172	5.270	5.520
2.23	1.693	3.381	3.370	5.021	5.006	6.419	6.615
7	2.510	5.101	5.030	7.634	7.478	9.776	9.880
30	4.783	9.836	9.633	14.793	14.339	18.971	18.948

Note that, although the Skyrme field ansatz (30) is axially symmetric the energy density is radially symmetric. In fact, a rotation by an angle α in the xy plane is equivalent to a unitary transformation $U \rightarrow A^{-1}UA$, where $A = e^{-iN(N-1)\alpha/2}\,\mathrm{diag}(1,e^{i\alpha},e^{2i\alpha},\ldots,e^{(N-1)i\alpha})$. Since the $B>2$ SU(2) skyrmion solutions have only discrete symmetries, the SU(N) skyrmion solutions for $N \geqslant 4$ are more symmetrical than the corresponding SU(2) embeddings. However, the $B=2$ SU(2) solution is already axially symmetric and so the symmetry arguments do not help us in comparing the SU(3) and SU(2) solutions.

The energy of the exact $B=3$ SU(2) skyrmion solution (for $m_\pi=0.2$) is $E=3.470$. From Table I, we see that the energy difference between this SU(2) solution and our $B=3$ SU(4) configuration is, approximately, equal to the pion mass. However, the energy of the actual axially symmetric $B=3$ SU(4) solution is clearly lower than that of our configuration, leading to an even smaller energy difference. Also, the energy of the exact $B=4$ SU(2) skyrmion (for the same value of the pion mass) is $E=4.522$. In this case, the energy gap with our $B=4$ SU(5) configuration is nearly twice as large as the pion mass; i.e., it is less clear whether the energy gap between the exact axially symmetric solution and the SU(2) embedding remains lower than the pion mass.

VI. SU(3) CASE

In this section, we restrict our attention to the SU(3) model with $m_\pi=0$ and construct low-energy states with a baryon number from one up to six. From now on, $N=3$ and so A_N, given by (14), becomes $A_N=\frac{4}{3}$.

A. General discussion

As in the previous section, we minimize (13) by first minimizing the integrals $\mathcal{N}$ and $\mathcal{I}$ as functions of P and then minimizing (13) with respect to the profile function g. Once again, $\mathcal{N}$ is

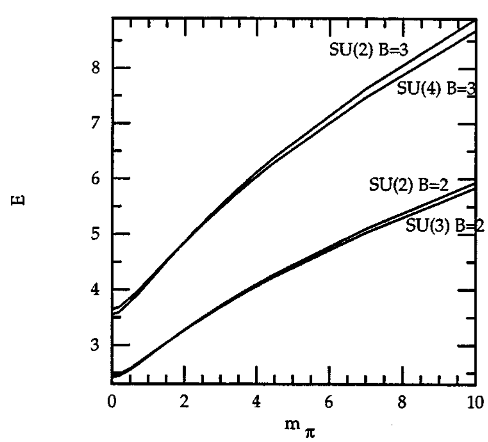

FIG. 2. Mass dependence of the SU(2) and SU($B+1$) harmonic map configurations for $B=2-3$.

minimized by the so-called self-dual solutions of the Euclidean CP^2 sigma model. They are given by (20), where f is any polynomial holomorphic vector $f(\xi)$ and their energy $\mathcal{N}$ is given by the degree of f.

Next, we note that the angular part of the baryon charge (18) coincides with the expression for the topological charge of the CP^2 sigma model, and so simplifies to

$$B=\frac{1}{8\pi}\int dS(1+|\xi|^2)^2\frac{|P_+f|^2}{|f|^2}, \tag{34}$$

where $dS\equiv\sin\theta\,d\theta\,d\phi=2i(1+|\xi|^2)^{-2}\,d\xi\,d\bar{\xi}$.

To minimize (13) for a configuration with a given baryon number B, we take $f(\xi)$ to be a holomorphic vector of degree B, which, by construction, minimizes $\mathcal{N}$. First we use the global SU(3) invariance of the model to reduce the number of parameters to the moduli space of the two-dimensional sigma model to

$$f=\begin{pmatrix}\xi^B+a_{B-1}\xi^{B-1}+\cdots+a_1\xi\\ b_{B-1}\xi^{B-1}+b_{B-2}\xi^{B-2}+\cdots b_1\xi+b_0\\ c_{B-2}\xi^{B-2}+c_{B-3}\xi^{B-3}+\cdots+c_1\xi+c_0\end{pmatrix}, \tag{35}$$

where all the coefficients are complex except b_{B-1}, which can be taken to be real. Then we substitute (19) for f of the form (35) into $\mathcal{I}$ and minimize numerically the integral with respect to all the coefficients. Finally, treating $\mathcal{N}=n$ and $\mathcal{I}$ as two fixed parameters, we minimize (13) by solving the resultant equation for g:

$$g_{rr}\left(1+2\mathcal{N}\frac{3\sin^2 g}{4r^2}\right)+2\frac{g_r}{r}+\frac{3\sin 2g}{4r^2}\left[\mathcal{N}(g_r^2-1)-\mathcal{I}\frac{\sin^2 g}{r^2}\right]=0. \tag{36}$$

An interesting feature of the SU(2) multiskyrmion solutions is the shape of surfaces of constant energy or baryon density. In fact, the energy and the baryon densities of the skyrmion solutions look very similar. For the baryon density these surfaces look like hollow shell-like structures with holes in it, while for the energy densities the holes are partly filled in and so are represented by local minima.[4]

In order to investigate the situation for our SU(3) field configurations, we have to look at the components of f given in (35) and study their effects on the density (34). Writing $f=(K,L,M)^t$, where K, L, and M are polynomials of degree B, $B-1$, and $B-2$, respectively, the integrand of (18) takes the form

$$\mathcal{B}=g_r\sin^2 g(1+|\xi|^2)^2\frac{|K_\xi L-L_\xi K|^2+|K_\xi M-M_\xi K|^2+|M_\xi L-L_\xi M|^2}{(|K|^2+|L|^2+|M|^2)^2}. \tag{37}$$

Note that the integrand of (37) is a scalar with respect to U(3) transformations applied to the vector f. Hence, any modifications of f that can be interpreted as such U(3) transformations are symmetries of (37).

The radial factor $g_r\sin^2 g$ in (37) indicates that if the angular part of the density vanishes, the baryon density will have radial holes going from the origin to infinity. For the density to vanish at some point we must require that the three factors in the numerator of (37) must vanish together, i.e., must have a common root. This is true when the three polynomials, $R_1=K_\xi L-L_\xi K$, $R_2=K_\xi M-M_\xi K$, and $R_3=M_\xi L-L_\xi K$, have a common factor. However, these polynomials have $2(B-1)$, $2B-3$, and $2(B-2)$ roots, respectively; with, in addition, a possible root at infinity (i.e., the south pole of the sphere). By counting powers we see that the density does not vanish at $\xi=\infty$ unless L is a polynomial of degree less than $B-1$.

From this we conclude that the baryon density can have, at most, $2B-3$ holes but, in general, it is likely to have fewer holes if any. Of course, when some terms in (37) vanish, the expression may (but does not have to) have a local minimum. Note that this is in contrast with the SU(2)

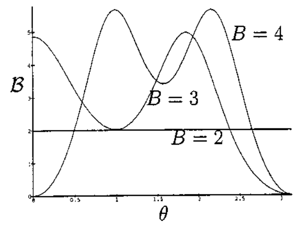

FIG. 3. The θ angular dependence of the baryon density (34).

configurations of Houghton *et al.*,[5] which always have $2(B-1)$ holes. In the SU(2) case, the vector f has only two components and so there is only one factor in the numerator of the baryon density, which thus has $2(B-1)$ zeros.

B. Specific fields

In this section we present the detailed form of harmonic maps that are used in the construction of the SU(3) skyrmion field ansatze.

First of all, the $B=1$ case, as discussed in Sec. V, is the SU(2) embedded skyrmion (i.e., the hedgehog ansatz). Next, we discuss field configurations for $B=2-6$. In each case, having found the map that minimizes $\mathcal{I}$, we solve numerically (36) and determine the corresponding profile function g. In Fig. 3 we present the θ angular dependence of the baryon densities for $B=2-4$ (no ϕ dependence). In Fig. 4, we present plots of surfaces of constant baryon density for $B=3-6$. The values we have chosen are, respectively, 0.3 times the maximum value of the topological density. (In all the graphs that follow, we always express the constant value for the curve as a fraction of the maximum density value.) The energy and baryon density for all the fields has the same symmetry and a virtually indistinguishable shape.

For $B=2$, using the ansatz (35), we have minimized $\mathcal{I}$ numerically and have found f to agree with the ansatz presented in Sec. V, i.e., to be given by

$$f=(\xi^2,\sqrt{2}\xi,1)^t. \tag{38}$$

For this field configuration $|P_+f|^2/|f|^2=2/(1+|\xi|^2)^2$ and, hence, as shown in Fig. 3, the baryon and energy density are independent of the polar angles on the sphere. Thus, the energy density of the $B=2$ field represents a shell.

The field (38) is axially symmetric, i.e., a rotation of an angle α in the xy plane is equivalent to the isorotation $U\rightarrow A^{-1}UA$, where $A=\mathrm{diag}(e^{-i\alpha},1,e^{i\alpha})$. This symmetry is identical to the symmetry of the $B=2$ SU(2) solutions.

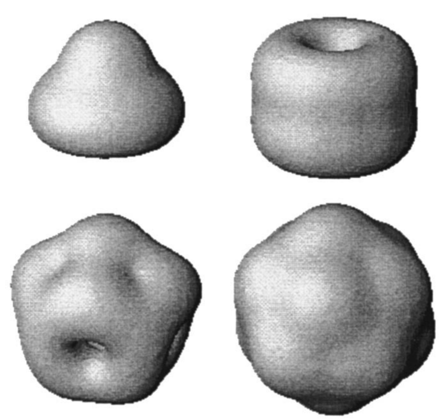

FIG. 4. Surfaces of constant baryon density for $B=3-6$.

For $B=3$, the numerical minimisation of $\mathcal{I}$ leads to the following expression for *f*:

$$f=(\xi^3,1.576\xi,\sqrt{2}^{-1})^t. \tag{39}$$

The baryon density of this configuration is axially symmetric and has the shape of a *torus with a sphere on top of it*. Since all the components of *f* are monomials, a transformation $\xi\to\xi'=\xi e^{i\alpha}$ for any α (i.e., a rotation around the *z*-axis), can be interpreted as the SU(3) transformation $U\to A^{-1}UA$, where $A=e^{-4/3i\alpha}\,\mathrm{diag}(e^{3i\alpha},e^{i\alpha},1)$. Hence, the configuration is axially symmetric—which proves the existence of an axially symmetric SU(3) solution with lower energy than the energy of this configuration (cf. Table III).

The baryon density for (39) does not vanish except when $|\xi|^2$ is infinite. This comes from the fact that the three terms in the numerator of (37) do not have common factors; however, as the second term of (39) is a polynomial of degree one, the baryon density vanishes for $\xi=\infty$. Indeed, we see in Fig. 4(a) that the density vanishes on the negative part of the *z*-axis ($\theta=\pi$).

For $B=4$ we find

$$f=(\xi^4,2.7191\xi^2,1)^t. \tag{40}$$

This configuration also leads to energy and baryon densities that are axially symmetric and have the shape of *two tori on top of each other close to the equator of the sphere* [see Fig. 4(b)]. These densities are invariant under a rotation around the *z*-axis, which is equivalent to the SU(3) transformation: $U\to A^{-1}UA$ where $A=\mathrm{diag}(e^{-2i\alpha},1,e^{-2i\alpha})$.

Note that the baryon density for (39) does vanish when ξ is zero or when its modulus $|\xi^2|$ is infinite. This happens as the three terms in the numerator of (37) have a single common factor at $\xi=0$, and the second term of *f* is a polynomial of degree two—implying once again that the baryon density vanishes when $\xi=\infty$. Indeed, this can be seen in Fig. 4(b); clearly the density vanishes along the *z*-axis ($\theta=0$ and $\theta=\pi$).

Let us mention that the energy densities of our configuration for $B=3-4$ are remarkably similar to the density of a SU(2) configuration corresponding to the scattering of three (four) skyrmions in an attractive channel,[11] and of three (four) monopoles.[12]

The holomorphic vector for $B=5$ is given by

$$f=(\xi^5-2.7\xi,2\xi^4+1,9/2\,\xi^3)^t, \tag{41}$$

and the baryon density resembles a structure consisting of *two deformed tori*, close to the equator, with an *additional ball* at the north pole of the angular sphere [see Fig. 4(c)]. Let us stress that the energy densities of the SU(2) embeddings have very different shapes and symmetries.

It is easy to check that the baryon density corresponding to the field in (41) does not have any holes. Note that by taking *f* in the form close to (41), i.e.,

$$f=\left(\xi^5+\frac{3C}{D}\xi,D\xi^4+C,E\xi^3\right)^t, \tag{42}$$

all the three terms in the numerator of (37) have zeros when

$$\xi^4=\frac{3C}{D}, \tag{43}$$

which would give four holes in the baryon density. So, since our field (41) is not very different from (42), our densities have minima; corresponding to the holes (43) partially filled in, by going from (42) to (41).

The holomorphic vector for $B=6$ is given by

$$f=(\xi^6+3\xi,1-3\xi^5,k\xi^3)^t, \tag{44}$$

TABLE II. Energy of SU(3) harmonic ansatz compared to the energy of the SU(2) harmonic ansatz and the energy of the SU(2) solutions obtained numerically.

B	$\mathcal{I}$(SU(3))	SU(3) En/Sk (Ans.)	SU(2) En/Sk (Ans.)	SU(2) En/Sk (Sol.)
1	1	1.232	1.232	1.232
2	4	1.222	1.208	1.171
3	10.65356	1.215	1.184	1.143
4	18.04501	1.184	1.137	1.116
5	27.26	1.164	1.147	1.116
6	37.33	1.1458	1.137	1.109

where k was found, numerically, to be 7.06. Once again, the baryon density of the field (44) does not have any holes but has regions where it is small but nonzero [see Fig. 4(d)]. This figure shows that this configuration has an *icosahedral symmetry* and this leads us to the conclusion that, modulo a SU(3) global transformation, (44) must be invariant under the following transformation:[13] $\xi \to \xi' = \xi e^{i2\pi/5}$ (i.e., a rotation by 72° around the z-axis); $\xi \to \xi' = -\xi^{-1}$ (which corresponds to $\theta \to \pi - \theta$ and $\phi \to \pi - \phi$) and $\xi \to \xi' = (\xi + b)(b\xi + 1)^{-1}$, where $b = 2\cos(2\pi/5) = (\sqrt{5}-1)/2$. This last transformation imposes a condition on k in (44): it is easy to see that the SU(3) transformation on f must be of the form $U = R/\det(R)^{1/3}$, with

$$R = \begin{pmatrix} 25+15a & 10+5a & (150+100a)k^{-1} \\ 10+5a & 25+15a & -(150+100a)k^{-1} \\ -k(3+2a) & k(3+2a) & 15+10a \end{pmatrix}, \tag{45}$$

where $a = -(1+\sqrt{5})/2$. Imposing the condition that the rows and columns of R are orthogonal to each other implies that $k = \sqrt{50} \approx 7.071$, which is within the precision of our numerical minimization program.

In Table II we present the energy values of the resulting Skyrme fields. All the numerical values of the energies are given in units of B and hence are close to unity. These values are then compared with the SU(2) skyrmion embeddings obtained using rational maps in Ref. 5. We see that both field configurations have similar values of energy, although the energies of the embeddings are marginally lower.

The symmetries of our $B=3-4$ and $B=6$ configurations are larger than the symmetry of the corresponding SU(2) embeddings. Therefore, solutions with the same symmetry as our configurations must exist and they will have lower energies than our fields. (Table II gives the upper bound on the energy of these solutions.) From Table III, it can be observed that in all three cases, the energy differences between our configurations and the corresponding SU(2) embeddings are lower or comparable to the pion mass.

This does not mean that the embeddings are closer (in shape of their energy densities) to the lowest-energy solutions of the SU(3) model. Our harmonic map fields and the embedded fields have very different symmetries, and as their energies are very close to each other it is pure

TABLE III. Energy of SU(2) multiskyrmion solutions for pion masses $m_\pi = 0$ and $m_\pi = 0.2$.

B	$E_{m_\pi=0}$	$E_{m_\pi=0.2}$
1	1.232	1.247
2	2.342	2.380
3	3.429	3.470
4	4.464	4.522
5	5.58	5.642
6	6.654	6.771

speculation to argue which are nearer (in their shapes) to the lowest-energy solutions. The true solutions may be close to either or neither of them, or it may be that the model has many different solutions, almost degenerate in energy. When the symmetries are different we do expect different solutions to exist, although some of them may not correspond to the local minima of the energy. This issue will be resolved only when detailed and reliable numerical studies of the SU(3) model have been performed.

VII. CONCLUSIONS

In this paper, we have discussed various static field configurations of the SU(N) Skyrme model. We have shown that, in addition to the obvious embeddings, any solution of the SU(2) model generates a solution of the SU(4) model. This solution is topologically trivial (i.e., zero baryon density) and its energy is equal to four times the energy of the original SU(2) solution.

Next, we have generalized the rational map ansatz of Houghton *et al.*[5] using harmonic maps and found another SU(3) exact spherically symmetric skyrmion solution. The baryon number of this solution is also zero and its energy is less than four (in topological units). We have argued that it represents a bound state of two skyrmions and two antiskyrmions.

Then we have presented topologically nontrivial field configurations of the SU(N) Skyrme model with radially symmetric energy densities—using harmonic maps. They correspond to $B=N-1$ skyrmions in SU(N) models. Although in the massless case their energies have turned out also to be above those of the SU(2) embeddings; when mass is added to the model, for sufficiently large masses, their energies can be lower than the energies of the embeddings.

We have also looked at various field configurations of the SU(3) model. The energy and baryon densities of these SU(3) fields exhibit shell-like structures; in all cases, except for $B=1$, they are different from the corresponding structures seen in the SU(2) model and are more symmetrical. Their energies are slightly higher but comparable to those of the embeddings. However, their different symmetry properties imply that there exist true solutions with these symmetries and with energies lower than our configurations. Thus, the model has at least two types of low-energy solutions. To know more about them, one will have to perform three-dimensional numerical simulations—this so far has not been done.

ACKNOWLEDGMENTS

The authors would like to thank V. Kopeliovich and J. Garraham for their help in finding a numerical mistake in the earlier report of this work (in the form of two papers). We also thank C. J. Houghton for his interest and correspondence about the symmetries of our field configurations.

[1] T. H. R. Skyrme, Nucl. Phys. **31**, 556 (1962).
[2] E. Witten, Nucl. Phys. B **223**, 422 (1983).
[3] T. H. R. Skyrme, Proc. R. Soc. London, Ser. A **260**, 127 (1961).
[4] R. A. Battye and P. M. Sutcliffe, Phys. Rev. Lett. **79**, 363 (1997).
[5] C. J. Houghton, N. S. Manton, and P. M. Sutcliffe, Nucl. Phys. B **510**, 507 (1998).
[6] A. P. Balachandran, A. Barducci, F. Lizzi, V. G. J. Rodgers, and A. Stern, Phys. Rev. Lett. **52**, 887 (1984).
[7] V. B. Kopeliovich, B. E. Schwesinger, and B. E. Stern, JETP Lett. **62**, 185–190 (1995).
[8] W. J. Zakrzewski, *Low Dimensional Sigma Models* (IOP, Bristol M, 1989).
[9] A. D'Adda, P. Di Vecchia, and M. Luscher, Nucl. Phys. B **146**, 63 (1980).
[10] A. Din and W. J. Zakrzewski, Nucl. Phys. B **174**, 397 (1980).
[11] R. A. Battye and P. M. Sutcliffe, in preparation.
[12] P. M. Sutcliffe, Nucl. Phys. **505**, 517 (1997).
[13] F. Klein, *The Icosahedron* (Dover, New York, 1956).

Quasi-Lagrangian systems of Newton equations

Stefan Rauch-Wojciechowski,[a)] Krzysztof Marciniak,[b)] and Hans Lundmark[c)]
Department of Mathematics, Linköping University, S-581 83 Linköping, Sweden

(Received 19 October 1998; accepted for publication 21 July 1999)

Systems of Newton equations of the form $\ddot{q}=-\frac{1}{2}A^{-1}(q)\nabla k$ with an integral of motion quadratic in velocities are studied. These equations generalize the potential case (when $A=I$, the identity matrix) and they admit a curious quasi-Lagrangian formulation which differs from the standard Lagrange equations by the plus sign between terms. A theory of such quasi-Lagrangian Newton (qLN) systems having two functionally independent integrals of motion is developed with focus on two-dimensional systems. Such systems admit a bi-Hamiltonian formulation and are proved to be completely integrable by embedding into five-dimensional integrable systems. They are characterized by a linear, second-order partial differential equation PDE which we call the fundamental equation. Fundamental equations are classified through linear pencils of matrices associated with qLN systems. The theory is illustrated by two classes of systems: separable potential systems and driven systems. New separation variables for driven systems are found. These variables are based on sets of nonconfocal conics. An effective criterion for existence of a qLN formulation of a given system is formulated and applied to dynamical systems of the Hénon–Heiles type. © *1999 American Institute of Physics.* [S0022-2488(99)00912-3]

I. INTRODUCTION

In this paper we introduce and study such systems of Newton equations $\ddot{q}=M(q)$ that can be generated as equations of the form

$$0=\frac{d}{dx}\frac{\partial E}{\partial \dot{q}}+\frac{\partial E}{\partial q}\equiv\delta^{+}E \tag{1.1}$$

by an energylike function quadratic in $\dot{q}$,

$$E(q,\dot{q})=\sum_{i,j=1}^{n}A_{ij}(q)\dot{q}_i\dot{q}_j+k(q)\equiv\dot{q}^tA\dot{q}+k(q), \tag{1.2}$$

where $A(q)$ is an $n\times n$ symmetric matrix with real entries $A_{ij}(q)$. Here and in what follows we use the standard mechanical notation $q=(q_1,\ldots,q_n)^t$, $\dot{q}=(\dot{q}_1,\ldots,\dot{q}_n)^t$, for position and velocity vectors (the superscript t denotes the transpose of a matrix), where $\dot{q}_k=(\partial/\partial x)q_k$, $k=1,...,n$, with $x\in\mathbf{R}$ being the independent (time) variable. By Newton equations we mean second-order ordinary differential equations (ODEs) of the form: acceleration $\ddot{q}$ is equal to the velocity independent force $M(q)$. The force M may be potential or not.

a)Electronic mail: strau@mai.liu.se
b)On leave of absence from Department of Physics, A. Mickiewicz University, Poznań, Poland. Electronic mail: krmar@mai.liu.se
c)Electronic mail: halun@mai.liu.se

The equations in (1.1) are called here *quasi-Lagrangian* (qL) equations since they differ from the Lagrange equations for $E(q,\dot{q})$ by sign between terms only. These equations are shortly denoted $0=\delta^+E=(\delta_1^+E,\ldots,\delta_n^+E)^t$ where

$$\delta_k^+E=\frac{d}{dx}\frac{\partial E}{\partial \dot{q}_k}+\frac{\partial E}{\partial q_k}.$$

The qL equations are not invariant with respect to arbitrary point transformation, but it can be easily shown that they remain invariant with respect to the affine change of variables $q=SQ+h$ where $Q=(Q_1,\ldots,Q_n)^t$ are the new variables and $S\in \mathrm{GL}(n),h\in \mathbf{R}^n$.

In the present article we shall mainly discuss quasi-Lagrangian sets of Newton equations (qLN) generated by a function E of the form (1.2) in the two-dimensional space of variables $q=(q_1,q_2)=(r,w)$. This class of equations (which seems to be completely new) is a very interesting class because of its rich differential-algebraic structure and also because it contains (as special cases) the well understood class of point-separable potential Newton equations $\ddot{q}=-\partial V(q)/\partial q$ and the class of nonpotential Newton equations of the triangular form $\ddot{r}=M_1(r,w)$, $\ddot{w}=M_2(w)$ which we shall call *driven* systems. The qLN systems are not necessarily Lagrangian and thus they do not have any straightforward Hamiltonian formulation.

In this paper we develop a theory of completely integrable sets of qLN equations characterized by the existence of two functionally independent integrals of motion quadratic in velocities: E as above and $F=\dot{q}^tB(q)\dot{q}+l(q)$. The existence of a second integral of motion has far-reaching consequences; it eventually leads to wide classes of completely integrable qLN systems.

Example 1.1: The function $E=r\dot{r}\dot{w}-w\dot{r}^2-\alpha wr^2+\frac{1}{2}dr^2+(w^2/2r^4)$ when inserted into (1.1) gives rise to

$$0=\begin{bmatrix}\frac{d}{dx}\frac{\partial E}{\partial \dot{r}}+\frac{\partial E}{\partial r}\\ \frac{d}{dx}\frac{\partial E}{\partial \dot{w}}+\frac{\partial E}{\partial w}\end{bmatrix}=\begin{bmatrix}-2w\left(\ddot{r}-\alpha r+\frac{w}{r^5}\right)+r(\ddot{w}-4\alpha w+d)\\ r\left(\ddot{r}-\alpha r+\frac{w}{r^5}\right)\end{bmatrix}$$
$$=\begin{bmatrix}-2w & r\\ r & 0\end{bmatrix}\begin{bmatrix}\ddot{r}-M_1(r,w)\\ \ddot{w}-M_2(w)\end{bmatrix}, \tag{1.3}$$

which is equivalent to a set of two Newton equations

$$\ddot{r}=\alpha r-\frac{w}{r^5}\equiv M_1(r,w),$$
$$\ddot{w}=4\alpha w-d\equiv M_2(w), \tag{1.4}$$

since the matrix

$$\begin{bmatrix}-2w & r\\ r & 0\end{bmatrix}$$

is nonsingular. We see that the operation $0=\delta^+E$ generates *linear combinations* of the Newton equations (1.4).

Equations (1.4) were discovered accidentally as a Newton parametrization of the second stationary flow of the Harry Dym hierarchy:[1]

$$0=(\tfrac{1}{4}\partial^3-\alpha\partial)(\alpha u^{-3/2}-\tfrac{5}{16}\dot{u}^2u^{-7/2}+\tfrac{1}{4}\ddot{u}u^{-5/2})=(\tfrac{1}{4}\partial^3-\alpha\partial)(-r^5\ddot{r}-\alpha r^6)$$

(here $\partial=\partial/\partial x$), where we substituted $u=r^{-4}$. The substitution $w=-r^5\ddot{r}+\alpha r^6$ gives the system (1.4). The particular feature of (1.4) is that it is a driven system: the equation for w can be solved independently and then the solution $w(x)$ drives the equation for r.

II. GENERAL PROPERTIES OF QUASI-LAGRANGIAN NEWTON SYSTEMS

Let us consider an n-dimensional qL system $0=\delta^+E$ with (quadratic in velocities) energylike function

$$E(q,\dot{q})=\sum_{i,j=1}^{n} A_{ij}(q)\dot{q}_i\dot{q}_j+k(q) \tag{2.1}$$

with a symmetric (which can be assumed without loss of generality) matrix $A(q)=A^t(q)$. We shall formulate the necessary and sufficient condition for the matrix $A(q)$ to make the equations $0=\delta^+E$ equivalent to the set of equations

$$0=\ddot{q}-M(q) \tag{2.2}$$

with a velocity independent force $M(q)=(M_1(q),\ldots,M_n(q))^t$.

Theorem 2.1: *For the function E given by (2.1) with a nonsingular matrix $A(q)$ the following conditions are equivalent:*

(1) *The equations $0=\delta^+E$ are equivalent to the set of Newton equations $\ddot{q}=M(q)$ with velocity independent forces $M=-\frac{1}{2}A^{-1}(q)\nabla k(q)$.*
(2) *The function E is an integral of motion for the qL system $0=\delta^+E$.*
(3) *The matrix elements $A_{ij}(q)$ satisfy the following set of ''cyclic'' differential equations:*

$$0=\partial_iA_{jk}(q)+\partial_jA_{ki}(q)+\partial_kA_{ij}(q) \quad \text{for all} \quad i,j,k=1,...,n. \tag{2.3}$$

Throughout the whole article the symbol ∇ denotes the gradient operator and $\partial_i=\partial/\partial q_i$. Later on we will also use the notation $\partial_{ij}=\partial^2/\partial q_i\partial q_j$.

Statement (2) of the above theorem explains the name ''energylike'' for the function E.

Proof: Let us calculate the ith equation in $0=\delta^+E$:

$$\begin{aligned}0=\delta_i^+E=\frac{d}{dx}\frac{\partial E}{\partial\dot{q}_i}+\frac{\partial E}{\partial q_i}&=\frac{d}{dx}\left(2\sum_j A_{ij}(q)\dot{q}_j\right)+\sum_{j,k}\partial_iA_{jk}(q)\dot{q}_j\dot{q}_k+\partial_ik\\&=2\sum_j A_{ij}(q)\ddot{q}_j+\partial_ik+\sum_{j,k}(\partial_iA_{jk}(q)+\partial_jA_{ki}(q)+\partial_kA_{ij}(q))\dot{q}_j\dot{q}_k.\end{aligned} \tag{2.4}$$

The last equality in (2.4) is due to the symmetry of $A(q)$. Thus, clearly, $2A\ddot{q}+\nabla k=0$ if and only if the equations (2.3) are satisfied and the equivalence of (1) and (3) is established.

Let us now calculate the total derivative of E with respect to *x:*

$$\begin{aligned}\dot{E}&=\sum_i\left(2\sum_j A_{ij}\ddot{q}_j+\partial_ik\right)\dot{q}_i+\sum_{i,j,k}\partial_kA_{ij}\dot{q}_i\dot{q}_j\dot{q}_k\\&=\sum_i\left(2\sum_j A_{ij}\ddot{q}_j+\partial_ik\right)\dot{q}_i+\frac{1}{3}\sum_{i,j,k}(\partial_iA_{jk}+\partial_jA_{ki}+\partial_kA_{ij})\dot{q}_i\dot{q}_j\dot{q}_k.\end{aligned} \tag{2.5}$$

The second term on the right-hand side of the above equation has been rewritten by renaming indices. It contains precisely the cyclic conditions (2.3). So, if one (and thus both) of the statements (1) and (3) are satisfied, then both terms in (2.5) vanish. On the other hand, if $\dot{E}=0$, then terms at different powers of $\dot{q}_i$ in (2.5) must be equal to zero, which implies both 1 and 3.

Remark 2.2: For $n=2$ the general solution of equations (2.3) can easily be found. It is

$$A_{11}(w)=aw^2+bw+\alpha,$$
$$2A_{12}(r,w)=-2arw-br-cw+\beta, \tag{2.6}$$
$$A_{22}(r)=ar^2+cr+\gamma,$$

with some real constants $a,b,c,\alpha,\beta,\gamma$. The corresponding qLN equations read explicitly as

$$0=\begin{bmatrix} \dfrac{d}{dx}\dfrac{\partial E}{\partial \dot r}+\dfrac{\partial E}{\partial r} \\ \dfrac{d}{dx}\dfrac{\partial E}{\partial \dot w}+\dfrac{\partial E}{\partial w}\end{bmatrix}=2\begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22}\end{bmatrix}\begin{bmatrix}\ddot r - M_1(r,w) \\ \ddot w - M_2(r,w)\end{bmatrix},$$

where

$$M_1(r,w)=\frac{1}{2\det(A)}\left(A_{12}\frac{\partial k}{\partial w}-A_{22}\frac{\partial k}{\partial r}\right),$$

$$M_2(r,w)=\frac{1}{2\det(A)}\left(A_{12}\frac{\partial k}{\partial r}-A_{11}\frac{\partial k}{\partial w}\right).$$

The remaining part of this work is mostly devoted to the case when qLN equations $\ddot q=-\frac{1}{2}A^{-1}\nabla k$ generated by E admit a *second* (quadratic in velocities) integral of motion $F(q,\dot q)=\Sigma_{i,j=1}^n B_{ij}(q)\dot q_i\dot q_j+l(q)\equiv\dot q^tB(q)\dot q+l(q)$ which is linearly, and therefore functionally, independent of E.

Theorem 2.3 (qLN systems with two integrals): *Let the qLN system of Newton equations,*

$$0=\delta^+E=2A(\ddot q+\tfrac{1}{2}A^{-1}\nabla k), \tag{2.7}$$

generated by the function $E(q,\dot q)=\dot q^tA(q)\dot q+k(q)$, admit a second, functionally independent quadratic integral of motion $F(q,\dot q)=\dot q^tB(q)\dot q+l(q)$. Then we have the following.

(1) The matrix $B(q)$ has the same structure as the matrix $A(q)$ in the sense that the coefficients $B_{ij}(q)$ of $B(q)$ satisfy the set of cyclic differential equations (2.3).

(2) *If* $\det(B)\neq 0$, *then*

$$A^{-1}\nabla k=B^{-1}\nabla l, \tag{2.8}$$

and so the qLN system $0=\delta^+F=2B(\ddot q+\frac{1}{2}B^{-1}\nabla l)$ generates the same Newton equations as E.

(3) *Any differentiable function $f(E,F)$ generates the same system of Newton equations [by $0=\delta^+f(E,F)$] as E does. In particular, any linear combination, $\lambda E+\mu F$ generates the same system of Newton equations.*

The statement (2) shows one of the peculiar features of qLN systems: all quadratic (in velocities) integrals of motion of a qLN system generate the same system (see also Sec. 151 in Ref. 2).

Proof: The requirement $\dot B=0$ yields [cf. (2.5)]

$$0=\sum_i\left(2\sum_j B_{ij}(q)\ddot q_j+\partial_i l\right)\dot q_i+\sum_{i,j,k}\partial_kB_{ij}\dot q_i\dot q_j\dot q_k$$
$$=\sum_i \dot q_i\left(2B\left(-\frac{1}{2}A^{-1}\nabla k\right)+\nabla l\right)_i+\sum_{i,j,k}\partial_kB_{ij}\dot q_i\dot q_j\dot q_k, \tag{2.9}$$

where the index i at the vector expression containing matrices B and A^{-1} denotes its ith component. The equality is satisfied identically with respect to $\dot{q}$ and so both sums must be separately equal to zero. It follows that B_{ij} satisfy the cyclic conditions $\partial_i B_{jk} + \text{cycl} = 0$ and that $2B(-\frac{1}{2}A^{-1}\nabla k) + \nabla l = 0$. The latter yields precisely the equation (2.8) since we assumed det(B) $\neq 0$. So the statements (1) and (2) are proved.

The operator δ^+ acts as differentiation on the algebra of constants of motion, so that

$$0 = \delta^+ f(E,F) = \frac{\partial f}{\partial E}\delta^+ E + \frac{\partial f}{\partial F}\delta^+ F = 2\left(\frac{\partial f}{\partial E}A + \frac{\partial f}{\partial F}B\right)(\ddot{q} - M)$$

(where $M = -\frac{1}{2}A^{-1}\nabla k = -\frac{1}{2}B^{-1}\nabla l$), which proves the statement (3) of the theorem. Q.E.D.

It is important to stress that the equation (2.8) is the necessary and sufficient condition for the equivalence of the qLN system (2.7) and the qLN system generated by $F = \dot{q}^t B(q)\dot{q} + l(q)$. This condition will be used later.

III. qLN EQUATIONS IN TWO DIMENSIONS

We shall from now on restrict our considerations to the case $n = 2$. We will use the notation $q = (q_1, q_2)^t = (r, w)^t$. The case of arbitrary n is studied in a separate paper.[3]

For $n = 2$ Theorem 2.3 contains two special cases which explain the connection of our theory with classical results[2] about separable potential Newton equations and with the class of driven systems where one of the Newton equations depends only on a single variable r or w and can be solved on its own.

Corollary 3.1: Assume that the Newton equations

$$\ddot{r} = M_1(r,w), \quad \ddot{w} = M_2(r,w) \tag{3.1}$$

generated by the integral $E = \dot{q}^t A(q)\dot{q} + k(q)$ *[with the matrix A given by (2.6)] as* $0 = \delta^+ E$ *have a potential force:* $M_1 = -\partial V/\partial r$, $M_2 = -\partial V/\partial w$. *Then the potential* $V(r,w)$ *satisfies the Bertrand–Darboux equation*[2]

$$\begin{aligned} 0 = {} & (V_{ww} - V_{rr})(-2arw - br - cw + \beta) + 2V_{rw}(aw^2 - ar^2 + bw - cr + \alpha - \gamma) \\ & + 3V_r(2aw + b) - 3V_w(2ar + c) \end{aligned} \tag{3.2}$$

(where the indices at V denote partial derivatives with respect to r and w) with the coefficients $a,b,c,\alpha,\beta,\gamma$ *being exactly the coefficients of the polynomials in entries of the matrix A as given by (2.6). This means that the Newton system (3.1) can be solved by separating variables in the related Hamilton–Jacobi equation (see Ref. 2).*

Proof: If M is potential, then, according to Theorem 2.3, $M = -\frac{1}{2}A^{-1}\nabla k = -\nabla V$ and so $\nabla k = 2A\nabla V$. The potential V exists provided that $\partial^2 k/\partial r \partial w = \partial^2 k/\partial w \partial r$. This yields exactly the Bertrand–Darboux equation (3.2) for V. Q.E.D.

Remark 3.2: The quantity $k(r,w)/\det(A)$ satisfies the same Bertrand–Darboux equation as the potential V. This result can be verified directly but it also follows from Theorem 4.1 in the next section.

Remark 3.3: Let us emphasize that the Hamiltonian system

$$\dot{r} = s, \quad \dot{w} = z, \quad \dot{s} = -\frac{\partial V}{\partial r}, \quad \dot{z} = -\frac{\partial V}{\partial w}$$

generated by a separable natural Hamiltonian $\mathrm{H} = \frac{1}{2}(s^2 + z^2) + V(r,w)$ can be reconstructed as the qLN system $0 = \delta^+ E = 2A(\ddot{q} + \frac{1}{2}A^{-1}\nabla k)$ from its second integral of motion E. This is easy to see, since the above Hamilton equations are equivalent to $\ddot{r} = -\partial V/\partial r, \ddot{w} = -\partial V/\partial w$.

The second class of equations satisfying the assumptions of Theorem 2.3 is the class of qLN systems of the form

$$\ddot{r} = M_1(r,w), \quad \ddot{w} = M_2(w), \tag{3.3}$$

which naturally generalizes the system in Example 1.1. Such systems are called *driven* since the equation for w can be solved independently and then $w(x)$ can be substituted into the equation for r. Observe that the second equation (and thus the whole system) admits an extra integral of motion of the form $F(w,\dot{w}) = \dot{w}^2/2 - \int M_2(w)\, dw$. The qLN system $0 = \delta^+ E$ attains the form (3.3) if and only if the second component M_2 of the force $-\frac{1}{2}A^{-1}\nabla k$ does not depend on *r:*

$$\frac{\partial}{\partial r}(A^{-1}\nabla k)_2 = 0. \tag{3.4}$$

Example 3.4: The qLN equations generated by the function

$$E = r\dot{r}\dot{w} - w\dot{r}^2 + k(r,w)$$

are driven [i.e., have the form (3.3)] provided that $k(r,w)$ satisfies the following second-order PDE:

$$0 = \frac{\partial}{\partial r}\left(\frac{1}{r}k_r + \frac{2w}{r^2}k_w\right),$$

which is a specialization of (3.4). The general solution of the above equation is

$$k(r,w) = f\left(\frac{r^2}{w}\right) + r^2 g(w)$$

with arbitrary twice differentiable functions f and g. The corresponding qLN system attains the form

$$\ddot{r} = -rg'(w) + \frac{r}{w^2}f'\left(\frac{r^2}{w}\right), \quad \ddot{w} = -2\frac{d}{dw}(wg(w))$$

and can be solved by quadratures (see Sec. VII). The second integral of motion of our system, $F = \dot{w}^2/2 - \int M_2(w)\, dw = \dot{w}^2/2 + 2wg(w)$, yields the matrix B

$$B = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

which is singular so F does not generate our system. However, any linear combination $\lambda E + \mu F$ of E and F (with both λ and $\mu \neq 0$) is another integral of motion with a nonsingular matrix $B' = \lambda A + \mu B$ and thus it generates the same driven system as E.

Existence of two functionally independent constants of motion does not automatically imply Liouville integrability since we also need a Hamiltonian formulation for our equations of motion. Our systems usually do not have a Lagrangian formulation and so they do not have the standard Hamiltonian formulation. On the other hand, the special system discussed in Example 1.1, being a stationary flow of the Harry Dym hierarchy, is expected to be integrable. The question thus arises if/when our qLN systems possess a nonstandard Hamiltonian formulation. In Sec. VI we shall demonstrate the existence of new Poisson structures for qLN systems and their close relationship with Poisson pencils for separable potentials. We shall also explain there when and in what sense our qLN systems are integrable.

IV. FUNDAMENTAL EQUATION

We shall now characterize those two-dimensional qLN systems which admit two (quadratic in velocities) functionally independent integrals of motion E and F, with the force $M=-\frac{1}{2}A^{-1}\nabla k=-\frac{1}{2}B^{-1}\nabla l$. We remind the reader that for $n=2$ we use the notation $q=(q_1,q_2)^t=(r,w)^t$.

Let us consider two symmetric 2×2 matrices $A(r,w)$ and $B(r,w)$ both satisfying the cyclic conditions (2.3). According to Remark 2.2 they must have the following structure,

$$A=\begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}, \quad B=\begin{bmatrix} B_{11} & B_{12} \\ B_{12} & B_{22} \end{bmatrix} \tag{4.1}$$

with the polynomial entries given by [cf. (2.6)]

$$\begin{aligned} A_{11}(w)&=a_1w^2+b_1w+\alpha_1, \\ 2A_{12}(r,w)&=-2a_1rw-b_1r-c_1w+\beta_1, \\ A_{22}(r)&=a_1r^2+c_1r+\gamma_1, \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} B_{11}(w)&=a_2w^2+b_2w+\alpha_2, \\ 2B_{12}(r,w)&=-2a_2rw-b_2r-c_2w+\beta_2, \\ B_{22}(r)&=a_2r^2+c_2r+\gamma_2, \end{aligned} \tag{4.3}$$

with some arbitrary real constants $a_1,\ldots,\gamma_2$.

Theorem 4.1 (fundamental equation): *Let*

$$\begin{bmatrix} \ddot r \\ \ddot w \end{bmatrix}=-\frac{1}{2}A^{-1}\nabla k=-\frac{1}{2}B^{-1}\nabla l, \tag{4.4}$$

with nonsingular 2×2 *matrices A, B given by (4.1), (4.2), and (4.3), be a set of qLN equations. Then the functions* $K_1=k/\det(A)$ *and* $K_2=l/\det(B)$ *both satisfy the same linear, second-order, partial differential equation*

$$\begin{aligned} 0=&2(A_{12}B_{22}-A_{22}B_{12})K_{rr}-2(A_{11}B_{22}-A_{22}B_{11})K_{rw}+2(A_{11}B_{12}-A_{12}B_{11})K_{ww} \\ &+3(A_{12}\partial_rB_{22}-B_{12}\partial_rA_{22}+A_{22}\partial_wB_{11}-B_{22}\partial_wA_{11})K_r-3(A_{11}\partial_rB_{22}-B_{11}\partial_rA_{22} \\ &+A_{12}\partial_wB_{11}-B_{12}\partial_wA_{11})K_w+3(\partial_rA_{22}\partial_wB_{11}-\partial_rB_{22}\partial_wA_{11})K, \end{aligned} \tag{4.5}$$

which explicitly reads

$$\begin{aligned} 0=&2K_{rr}[\gamma_2\beta_1-\gamma_1\beta_2+(b_2\gamma_1-\gamma_2b_1+\beta_1c_2-c_1\beta_2)r+(\gamma_1c_2-\gamma_2c_1)w+(b_2c_1-c_2b_1+a_2\beta_1 \\ &-a_1\beta_2)r^2+2(\gamma_1a_2-\gamma_2a_1)wr+(a_1b_2-a_2b_1)r^3+(a_2c_1-c_2a_1)wr^2]+4K_{rw}[\alpha_2\gamma_1-\alpha_1\gamma_2 \\ &+(\alpha_2c_1-\alpha_1c_2)r+(b_2\gamma_1-\gamma_2b_1)w+(\alpha_2a_1-\alpha_1a_2)r^2+(\gamma_1a_2-\gamma_2a_1)w^2+(b_2c_1-c_2b_1)rw \\ &+(a_1b_2-a_2b_1)wr^2+(a_2c_1-c_2a_1)rw^2]+2K_{ww}[\alpha_1\beta_2-\alpha_2\beta_1+(\alpha_2b_1-\alpha_1b_2)r \\ &+(\alpha_2c_1-\alpha_1c_2+b_1\beta_2-b_2\beta_1)w+(a_1\beta_2-a_2\beta_1+b_2c_1-c_2b_1)w^2+2(\alpha_2a_1-\alpha_1a_2)wr \\ &+(a_2c_1-c_2a_1)w^3+(a_1b_2-a_2b_1)rw^2]+3K_r[2b_2\gamma_1-2\gamma_2b_1+\beta_1c_2-c_1\beta_2 \\ &+(3b_2c_1-3c_2b_1+2a_2\beta_1-2a_1\beta_2)r+4(\gamma_1a_2-\gamma_2a_1)w+4(a_1b_2-a_2b_1)r^2 \end{aligned}$$

$$+4(a_2c_1-c_2a_1)rw]+3K_w[2\alpha_2c_1-2\alpha_1c_2+b_1\beta_2-b_2\beta_1+(2a_1\beta_2-2a_2\beta_1$$
$$+3b_2c_1-3c_2b_1)w+4(\alpha_2a_1-\alpha_1a_2)r+4(a_2c_1-c_2a_1)w^2+4(a_1b_2-a_2b_1)rw]$$
$$+6K[b_2c_1-c_2b_1+2(a_1b_2-a_2b_1)r+2(a_2c_1-c_2a_1)w] \tag{4.6}$$

with K denoting either K_1 *or* K_2 *and* $K_r=\partial K/\partial r$, $K_{rr}=\partial^2K/\partial r^2$ *and so on.*

Conversely, any solution $K_2(q)$ *of the equation (4.5) generates two different systems of qLN equations* $\ddot{q}=-\frac{1}{2}A^{-1}\nabla k_1=-\frac{1}{2}B^{-1}\nabla l_1$ *and* $\ddot{q}=-\frac{1}{2}A^{-1}\nabla k_2=-\frac{1}{2}B^{-1}\nabla l_2$*, where the functions* k_1,k_2,l_1,l_2 *are determined by the equations*

$$l_1=K_2\det(B),\quad \nabla k_1=AB^{-1}\nabla(K_2\det(B)),$$
$$k_2=K_2\det(A),\quad \nabla l_2=BA^{-1}\nabla(K_2\det(A)). \tag{4.7}$$

We will call the equation (4.5) the *fundamental equation* associated with the matrices A and B.

The fundamental equation plays a crucial role in our theory of qLN systems. Observe that it is invariant with respect to the transformation $A\mapsto\lambda A+\mu B$, $B\mapsto\lambda' A+\mu' B$, $(\lambda,\lambda',\mu,\mu'\in\mathbf{R})$ since the coefficients at every monomial in this equation are skew-symmetric in A and B. This is consistent with statement (3) of Theorem 2.3, which asserts that if any pair E,F of functions generates a qLN system, then the linear combinations $\lambda E+\mu F$ and $\lambda' E+\mu' F$ also generate the same system. This explains that the assumption of nonsingularity for both A and B is nonessential since if $\det(A)\neq 0$, a singular matrix B can always be substituted by an invertible matrix $B'=\lambda A+\mu B$. We shall investigate further properties of the fundamental equation in the next theorem.

Notice that in the second part of Theorem 4.1 one has to reconstruct l_2 and k_1 by integrating the expressions for ∇l_2 and for ∇k_1. This can always be done, as the above theorem implicitly states. Also, notice that in the fundamental equation (4.6) all terms of degree 4 and higher cancel so that the polynomial degree of coefficients in this equation is less than or equal to 3.

Proof (of Theorem 4.1): Our qLN system (4.4) is generated by either of the two functions $E(q,\dot{q})=\dot{q}^tA\dot{q}+k$ and $F(q,\dot{q})=\dot{q}^tB\dot{q}+l$ and so the condition (2.8), i.e., $A^{-1}\nabla k=B^{-1}\nabla l$, must be satisfied. This implies that $\nabla l=BA^{-1}\nabla k$. This equation for the function l has solutions if and only if its compatibility condition $l_{rw}=l_{wr}$ is satisfied. This yields a PDE for the function k which, after the substitution $k=K_1\det(A)$ and with use of the cyclic conditions (2.3), yields that K_1 satisfies equation (4.5). By inserting into this equation the explicit form of the polynomials $A_{11},\ldots,B_{22}$ we obtain (4.6). On the other hand, the condition (2.8) implies also $\nabla k=AB^{-1}\nabla l$, and its compatibility condition $k_{rw}=k_{wr}$ gives a PDE which in terms of $K_2=l/\det(B)$ must attain the form (4.5) with interchanged entries of A and B (since the equation $\nabla k=AB^{-1}\nabla l$ becomes $\nabla l=BA^{-1}\nabla k$ when one exchanges A,k and B,l). Due to the skew-symmetry of coefficients of the equation for K_1 with respect to the entries of matrices A,B [clearly seen from the form of (4.5)], the obtained equation for K_2 differs from the equation for K_1 by a minus sign on the right-hand side only. This proves that K_1 and K_2 both satisfy (4.5) (notice, however, that this does *not* imply $K_1=K_2$).

The existence of k_1 [i.e., the possibility of integrating the equations (4.7) in order to obtain k_1] follows from the fact that the condition $\partial^2k_1/\partial r\partial w=\partial^2k_1/\partial w\partial l$ together with $\nabla k_1=AB^{-1}\nabla(K_2\det(B))$ yields precisely the fundamental equation for K_2 which is satisfied due to assumptions. One can similarly prove the existence of l_2. The second statement of the theorem can now be proved by checking that both pairs k_1,l_1 and k_2,l_2 given by (4.7) satisfy the condition (2.8) and thus give rise to two systems of qLN equations. Q.E.D.

Remark 4.2: For $B(q)=\frac{1}{2}I$ (a 2×2 identity matrix) the equation (4.5) becomes the Bertrand–Darboux equation (3.2) characterizing all separable potentials since in this case $\ddot{q}=-\frac{1}{2}B^{-1}\nabla l=-\nabla l$ is a potential equation.

The next theorem shows that there exists a recursive relation between two different qLN systems constructed from a given solution $K_2(q)$ of the fundamental equation (4.5). This makes it possible to construct a doubly infinite sequence of qLN systems corresponding to a given fundamental equation.

Theorem 4.3 (recursion theorem): *Let k_1, l_1 and k_2, l_2 be two pairs of functions determined by a given solution K_2 of the fundamental equation (4.5) as in (4.7). Then these functions are related by the following linear algebraic equations:*

$$k_2 = l_1 \det(AB^{-1}), \quad l_2 = l_1 \operatorname{Tr}(AB^{-1}) - k_1 \tag{4.8}$$

(where Tr *denotes trace of matrix). Moreover, in the infinite sequence*

$$\begin{array}{ccccccccccc} & k_0 & & & & k_1 & & & & k_2 & & \\ & & & & \nearrow & & & & \nearrow & & & \nearrow \\ \cdots & \downarrow & & K_1 & & \downarrow & & K_2 & & \downarrow & & K_3 & & \cdots \\ & & \nearrow & & & & \nearrow & & & & \nearrow & \\ & l_0 & & & & l_1 & & & & l_2 & & \end{array} \tag{4.9}$$

of triples $(K_m, k_m, l_m), m \in \mathbf{Z}$, defined recursively by

$$k_m = l_{m-1} \det(AB^{-1}), \quad l_m = l_{m-1} \operatorname{Tr}(AB^{-1}) - k_{m-1} \tag{4.10}$$

and by

$$K_m = k_m / \det(A) = l_{m-1} / \det(B),$$

the functions k_m and l_m satisfy $A^{-1}\nabla k_m = B^{-1}\nabla l_m$ and thus they both determine the same (for a given m) qLN system $\ddot{q} = -\frac{1}{2}A^{-1}\nabla k_m = -\frac{1}{2}B^{-1}\nabla l_m$. All functions K_m satisfy the fundamental equation (4.5) and are related through the following two-step recursion:

$$K_{m+1} = K_m \operatorname{Tr}(AB^{-1}) - K_{m-1} \det(AB^{-1}). \tag{4.11}$$

The above recursion is reversible. The solution K_m placed between l_{m-1} and k_m determines both l_{m-1} and k_m. The recursion (4.11) is soluble. Namely, if we denote the eigenvalues of the matrix AB^{-1} by λ_1 and λ_2, then it can be proved that for the case $\lambda_1 \neq \lambda_2$ the solution of (4.11) is

$$K_m = \frac{1}{\lambda_1 - \lambda_2}(K_1 - \lambda_2 K_0)\lambda_1^m + \frac{1}{\lambda_1 - \lambda_2}(K_0\lambda_1 - K_1)\lambda_2^m,$$

while in the case $\lambda_1 = \lambda_2$ the solution of (4.11) becomes

$$K_m = K_0 \lambda_1^m + \left(\frac{K_1}{\lambda_1} - K_0\right) m \lambda_1^m.$$

In both cases K_0 and K_1 are two subsequent solutions of the fundamental equation in the sequence (4.9) which are related by

$$\nabla(K_1 \det(B)) = BA^{-1}\nabla(K_0 \det(A)).$$

In order to prove the recursion theorem we need the following lemma.

Lemma 4.4: Let $X = AB^{-1}$ with matrices A,B as above. Then

$$X^{-1}\nabla(\det(X)) = \nabla(\operatorname{Tr}(X)).$$

This lemma follows from the cyclic properties (2.3) of matrices A and B by a lengthy but straightforward calculation.

Proof (of the recursion theorem): Consider a solution K_2 of the fundamental equation and the functions $k_1, l_1; k_2, l_2$ defined by (4.7). Then obviously $k_2/\det(A)=l_1/\det(B)$, which immediately implies $k_2=l_1\det(AB^{-1})$. Let $X=AB^{-1}$. Then

$$\begin{aligned}\nabla l_2-\nabla(l_1\,\mathrm{Tr}\,(X)-k_1)&=X^{-1}\nabla(K_2\det(A))-\nabla(K_2\det(B)\,\mathrm{Tr}\,(X))+X\nabla(K_2\det(B))\\&=X^{-1}\nabla(K_2\det(A))-(\mathrm{Tr}\,(X)I-X)\nabla(K_2\det(B))-K_2\det(B)\nabla(\mathrm{Tr}\,(X))\\&=X^{-1}\nabla(K_2\det(A))-X^{-1}\det(X)\nabla(K_2\det(B))-K_2\det(B)\nabla(\mathrm{Tr}\,(X))\\&=K_2\det(B)(X^{-1}\nabla(\det(X))-\nabla(\mathrm{Tr}\,(X)))=0,\end{aligned}$$

where we used that $X^2-\mathrm{Tr}\,(X)X+\det(X)I=0$ as follows from the Cayley–Hamilton theorem. The last equality is due to Lemma 4.4 above. Thus $l_2=l_1\,\mathrm{Tr}\,(X)-k_1$ up to a nonessential additive constant. This proves the first assertion of the theorem.

If we now define the sequence $\{(k_m, l_m)\}$ via the recursive procedure (4.10), then a simple induction argument shows that each pair (k_m, l_m) satisfies the condition (2.8) and thus both k_m and l_m determine the same qLN system. Moreover, each $K_m=k_m/\det(A)=l_{m-1}/\det(B)$ is a solution of the fundamental equation as theorem (4.1) states. Finally, to obtain (4.11) it is enough to insert the formula $K_m=k_m/\det(A)=l_{m-1}/\det(B)$ into the second equation in (4.10). Q.E.D.

Example 4.5 (cf. Remark 4.2): For $B=\frac{1}{2}I$ (the potential case) the recursion (4.8) takes the form

$$k_2=4V_1\det(A),\quad V_2=2\,\mathrm{Tr}\,(A)V_1-k_1$$

with $V_1=l_1$. This is the separable case when (4.5) reduces to the Bertrand–Darboux equation. In the generic case, i.e., when $a\neq 0$ in (2.6), the matrix $A(q)$ can be reduced [with the use of affine transformations $q=SQ+h$ with $S\in \mathrm{GL}\,(2,\mathbf{R}), h\in\mathbf{R}^2$, see also Sec. V] to the form

$$A(q)=\begin{bmatrix}-q_2^2+\lambda_2 & q_1q_2\\ q_1q_2 & -q_1^2+\lambda_1\end{bmatrix}.$$

If we now start with the harmonic oscillator potential $V_1=\frac{1}{2}(q_1^2+q_2^2)$, then the condition $\nabla V_1=\frac{1}{2}A^{-1}\nabla k_1$ gives $k_1=\lambda_2q_1^2+\lambda_1q_2^2$ and the recursion formulas specify to

$$k_2=2(q_1^2+q_2^2)(\lambda_1\lambda_2-\lambda_2q_1^2-\lambda_1q_2^2),$$

$$V_2=\lambda_1q_1^2+\lambda_2q_2^2-(q_1^2+q_2^2)^2,$$

thus reproducing the potential of the Garnier system.[4] It can be shown that the above formulas prolongate to the $n=2$ case of the recursion for the Jacobi family of elliptic separable potentials.[5]

In order to explain the character of the recursion (4.9) more completely, let us consider instead of the pair (A,B) of cyclic matrices another pair $(A+\mu B,B)$ with $\mu\in\mathbf{R}$. As it can be shown (see below), this pair determines the same fundamental equation as the pair (A,B) does. By choosing a solution K_2 of the fundamental equation and the pair $(A+\mu B,B)$ we arrive at a different qLN system $\ddot{q}=M_\mu(q)=-\frac{1}{2}(A+\mu B)^{-1}\nabla(K_2\det(A+\mu B))$. It turns out that the force $M_\mu(q)$ is a linear combination of two neighboring forces in the sequence (4.9) generated by K_2.

Lemma 4.6: Let A and B be two 2×2 matrices satisfying the cyclic conditions (2.3) and let K be a solution of the fundamental equation associated with A and B. Let also $\mu\in\mathbf{R}$. Then

$$(A+\mu B)^{-1}\nabla(K\det(A+\mu B))=A^{-1}\nabla(K\det(A))+\mu B^{-1}\nabla(K\det(B)).$$

This lemma is a consequence of Lemma 4.4. It says that a solution of a given fundamental equation determines the force M (and so the system of qLN equations) up to *linear combinations* of two consecutive systems in the recursion (4.9).

As we have mentioned, the matrices A and B uniquely determine the fundamental equation. The choice of A,B which generate a given fundamental equation is, however, not unique since the pair $A' = \alpha A + \beta B,\ B' = \gamma A + \delta B$ determines the same equation. One can also ask to what extent a given fundamental equation determines the pair (A,B). The precise relationship between pairs (A,B) and the fundamental equation is explained in the following theorem.

Theorem 4.7: *Let (A,B) be a pair of linearly independent matrices A, B satisfying the cyclic conditions (2.3). Then there is a 1–1 relationship between the linear span $\{\lambda A + \mu B : \lambda, \mu \in \mathbf{R}\}$ of A and B and the fundamental equation (4.5), i.e.,*

(1) *any two linearly independent matrices $A' = \alpha A + \beta B,\ B' = \gamma A + \delta B$ determine the same fundamental equation as (A, B) does.*
(2) *If the pair (A',B') determines the same fundamental equation as (A,B) does, then the matrices A' and B' belong to the linear span $\{\lambda A + \mu B\}$ of A* and *B.*

Proof: An easy calculation shows that the fundamental equation associated with the matrices $A' = \alpha A + \beta B$ and $B' = \gamma A + \delta B$ differs from the fundamental equation associated with the matrices A and B by the multiplicative factor $\alpha\delta - \beta\gamma$ on the right-hand side, i.e., by the nonzero determinant of the transformation between (A,B) and (A',B'), and so it is, in fact, the same equation. This shows assertion (1) of the theorem.

Assume now that the equation (4.5) is associated with a pair (A,B). Consider the vector $\vec{X} = (X_1, X_2, X_3)^t \in \mathbf{R}^3$ of the coefficients of (4.5) at the highest derivatives K_{rr}, K_{rw}, K_{ww}, respectively. Then

$$X_1 = A_{12}B_{22} - A_{22}B_{12},$$
$$X_2 = A_{22}B_{11} - A_{11}B_{22},$$
$$X_3 = A_{11}B_{12} - A_{12}B_{11}, \tag{4.12}$$

or

$$\vec{X} = \vec{A} \times \vec{B}, \tag{4.13}$$

where $\vec{A} = (A_{11}, A_{12}, A_{22})^t$ and $\vec{B} = (B_{11}, B_{12}, B_{22})^t$ are three-dimensional vectors depending on r and w. Hence, for a fixed (r,w) both vectors $\vec{A}$ and $\vec{B}$ are orthogonal to $\vec{X}$. The coefficients at K_r, K_w, and K yield equations which are differential consequences of (4.12) and so they do not impose any additional restrictions on (A,B). Suppose now that there exist matrices A' and B' satisfying the cyclic condition (2.3) and associated with the same fundamental equation. This means that the equation (4.13) has another solution, i.e., that $\vec{X} = \vec{A}' \times \vec{B}'$, so that the vectors $\vec{A}'$ and $\vec{B}'$ are orthogonal to $\vec{X}$ and, in consequence, they are linear combinations of $\vec{A}$ and $\vec{B}$: $\vec{A}' = \alpha\vec{A} + \beta\vec{B},\ \vec{B}' = \gamma\vec{A} + \delta\vec{B}$ with some coefficients that may depend on r and w. For the corresponding matrices it immediately follows that

$$A' = \alpha A + \beta B, \quad B' = \gamma A + \delta B.$$

It remains to show that the coefficients α, β, γ, δ in fact do not depend on r nor w. This can be shown by inserting the explicit form (4.2) and (4.3) of entries of matrices A,B, A' and B' into (4.12). This shows assertion (2) of the theorem. Q.E.D.

V. AFFINE INEQUIVALENT FORMS OF FUNDAMENTAL EQUATION

In this section we are interested in characterizing all different types of two-dimensional qLN systems admitting two functionally independent integrals of motion E and F which are quadratic in velocities, i.e., systems of the form

$$\ddot{q}=M=-\tfrac{1}{2}A^{-1}\nabla k=-\tfrac{1}{2}B^{-1}\nabla l, \tag{5.1}$$

where M is the force of the system. Every such system is described by a pair of matrices $A(q),B(q)$ satisfying the cyclic conditions (2.3) and by a pair of functions $k(q)$, $l(q)$ satisfying $A^{-1}\nabla k=B^{-1}\nabla l$. We remind the reader that the functions $k/\det(A)$ and $l/\det(B)$ satisfy the same fundamental equation with the coefficients completely determined by the matrix elements of A and B.

Let us first consider how the qLN system $0=\delta^{+}E=2A(\ddot{q}+\tfrac{1}{2}A^{-1}\nabla k)$ transforms under the affine transformation of coordinates

$$q=SQ+h,\quad S\in \mathrm{GL}(2,\mathbf{R}),\quad h\epsilon\mathbf{R}^2, \tag{5.2}$$

where $Q=(Q_1,\dots,Q_n)^t$. It is easy to see that, under the affine transformation (5.2), the generating function E transforms as

$$E(q(Q),\dot{q}(\dot{Q}))=\dot{Q}^tS^tA(q(Q))SQ+k(q(Q)), \tag{5.3}$$

where $q(Q)=SQ+h$ and so $\dot{q}(\dot{Q})=S\dot{Q}$. It can be shown by a direct verification that the transformed matrix

$$A_Q(Q)=S^tA(q(Q))S \tag{5.4}$$

in (5.3) also satisfies the cyclic conditions (2.3) and therefore (5.3) generates a qLN system. This means that the qLN system $0=2A(\ddot{q}+\tfrac{1}{2}A^{-1}\nabla k)$ is indeed *invariant* with respect to the affine change of coordinates (5.2).

Let us now consider the system (5.1). Using (5.4) one can prove that the fundamental equation associated with the pair (A,B) of matrices is also invariant with respect to the affine transformations (5.2). This means that we can simplify this fundamental equation by performing an appropriate affine change of coordinates. However, Theorem 4.7 makes it possible to classify fundamental equations, and therefore the corresponding qLN systems, by classifying *pairs* of matrices (A,B). Instead of working with the coefficients of the fundamental equation we can thus work with linear spans $\{\lambda A+\mu B\}$ of A and B. Since the affine transformations do not change the polynomial degree of matrices A,B, the set of all linear spans of A and B can be divided into affine inequivalent classes corresponding to different polynomial degree of A and B. Each equivalence class will be represented by the algebraically simplest pair of matrices obtained by the use of affine transformations and linear combining of matrices (since the latter leave the fundamental equation unchanged, see above).

In order to be more precise we shall introduce some notation. By $A^{(i)}$ $(i=0,1,2)$ we will denote all matrices A which satisfy the cyclic conditions (2.3) and have the highest degree of polynomial entries equal to i. So, for example, the general form of matrices in the class $A^{(1)}$ is

$$\begin{bmatrix} bq_1+\alpha & -\tfrac{1}{2}bq_1-\tfrac{1}{2}cq_2+\tfrac{1}{2}\beta \\ * & cq_1+\gamma \end{bmatrix}$$

with arbitrary constants (parameters) b,c,α,β,γ. We will use the symbol $*$ to denote matrix elements determined by the symmetry of a given matrix. Moreover, by $[A^{(i)},B^{(j)}]$ $(i,j=0,1,2)$ we will denote the class of (nonordered) pairs (A,B) of linearly independent matrices A,B such that one of the matrices belongs to $A^{(i)}$ and the other to $B^{(j)}$. We have, of course, $[A^{(i)},B^{(j)}]=[A^{(j)},B^{(i)}]$ and so we have precisely six such classes. Obviously, all classes $[A^{(i)},B^{(j)}]$ are invariant with respect to the affine transformations (5.2). Notice that if $(A,B)\in[A^{(2)},B^{(2)}]$, we can kill the coefficient a_2 at the second degree monomials in B by subtraction $B\mapsto B-(a_2/a_1)A$ so every element of this class can be reduced to an element in $[A^{(2)},B^{(1)}]$. Thus we have to consider only five classes. It is easy to realize that the five classes $[A^{(i)},B^{(j)}]$ with $i\geqslant j,j<2$ are invariant with respect to the affine transformations (5.2) and with respect to taking linear combinations of pairs (A,B). These two operations can now be used to find for every class a simple representing pair (A,B) which has a minimal number of free parameters (a simple representative). Consider, for example, the class $[A^{(2)}, B^{(0)}]$. The general form of matrices belonging to this class is

$$A=\begin{bmatrix} a_1q_2^2+b_1b_2+\alpha_1 & -a_1q_1q_2-\frac{1}{2}b_1q_1-\frac{1}{2}c_1q_2+\beta_1/2 \\ * & a_1q_1^2+c_1q_1+\gamma_1 \end{bmatrix},$$
$$B=\begin{bmatrix} \alpha_2 & \frac{\beta_2}{2} \\ \frac{\beta_2}{2} & \gamma_2 \end{bmatrix}. \tag{5.5}$$

Translation by the vector $h=-(1/2a_1)(c_1,b_1)^t$ kills b_1 and c_1 in the matrix A. Since translations obviously preserve the form of B the above pair of matrices attains the form

$$A=\begin{bmatrix} a_1q_2^2+\alpha_1 & -a_1q_1q_2+\frac{\beta_1}{2} \\ * & a_1q_1^2+\gamma_1 \end{bmatrix}, \quad B=\begin{bmatrix} \alpha_2 & \frac{\beta_2}{2} \\ \frac{\beta_2}{2} & \gamma_2 \end{bmatrix}$$

with some new constants denoted by the same letters as in (5.5). Further, the transformation $A\mapsto A-(\beta_1/\beta_2)B$ kills the coefficient β_1 in A. In the case when $\beta_2=0$ we can still kill β_1 in A by the linear transformation $q=SQ$ with

$$S=\begin{bmatrix} \frac{-t\gamma_1-\beta_1/2}{\alpha_1+t\beta_1/2} & 1 \\ 1 & t \end{bmatrix},$$

where $t\in\mathbf{R}$ must be chosen so that $\alpha_1+t\beta_1/2\neq 0$ and $\det(S)\neq 0$, which can be always done. Finally, we can divide both matrices by a_1 and $2\alpha_2$, respectively. So, a simple representative of the class $[A^{(2)}, B^{(0)}]$ has the form

$$A=\begin{bmatrix} q_2^2+\alpha_1 & -q_1q_2 \\ * & q_1^2+\gamma_1 \end{bmatrix}, \quad B=\begin{bmatrix} \frac{1}{2} & \frac{\beta_2}{2} \\ \frac{\beta_2}{2} & \gamma_2 \end{bmatrix} \tag{5.6}$$

with four essential parameters. Both separable and driven systems belong to this class since $B=\frac{1}{2}I$ for separable systems and $B=\operatorname{diag}(\frac{1}{2},0)$ for driven systems.

We perform a similar reduction for each class $[A^{(i)}, B^{(j)}], i\geq j, j<2$. The results are presented below.

It is also easy to see that one can pass from one invariant class $[A^{(i)}, B^{(j)}]$ to another by specifying values of free parameters. For example, by setting $a_1=0$ we obtain $[A^{(1)}, B^{(1)}]$ from $[A^{(2)}, B^{(1)}]$; by setting $b_2=c_2=0$ we get $[A^{(1)}, B^{(0)}]$ and so on as shown in Fig. 1. This figure presents—for all classes $[A^{(i)}, B^{(j)}]$—complete results of simplification of a generic pair (A,B) belonging to each class with the use of linear combinations and affine transformations.

Below we list the form of the fundamental equation (4.6) corresponding to the simple representative pair (A,B) of each class as given in Fig. 1. We use the notation $K_i=\partial K/\partial q_i$, $K_{ij}=\partial^2K/\partial q_j\partial q_i$.

(a) For $[A^{(2)}, B^{(1)}]$,

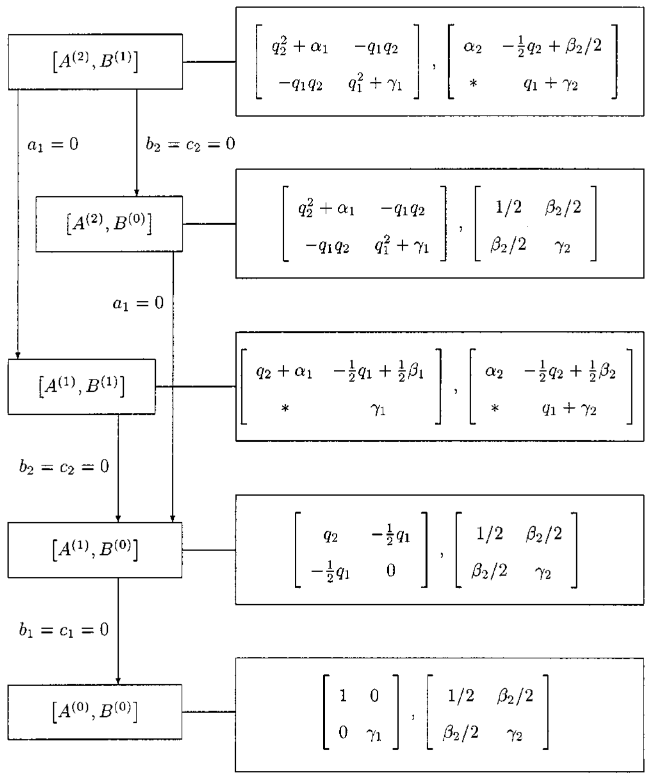

FIG. 1. Classification diagram.

$$0=K_{11}(-\gamma_1\beta_2+\gamma_1q_2-2\gamma_2q_1q_2-\beta_2q_1^2-q_1^2q_2)+2K_{12}(\alpha_2\gamma_1-\alpha_1\gamma_2-\alpha_1q_1-\gamma_2q_2^2+\alpha_2q_1^2$$
$$-q_1q_2^2)+K_{22}(\alpha_1\beta_2-\alpha_1q_2+2\alpha_2q_1q_2+\beta_2q_2^2-q_2^3)+3K_1(-\beta_2q_1-2\gamma_2q_2-2q_1q_2)$$
$$+3K_2(-\alpha_1+2\alpha_2q_1+\beta_2q_2-2q_2^2)-6q_2K.$$

(b) For $[A^{(2)},B^{(0)}]$,

$$0=K_{11}(-\gamma_1\beta_2-2\gamma_2q_1q_2-\beta_2q_1^2)+2K_{12}(\gamma_1-\alpha_1\gamma_2-\gamma_2q_2^2+q_1^2)$$
$$+K_{22}(\alpha_1\beta_2+2q_1q_2+\beta_2q_2^2)+3K_1(-\beta_2q_1-2\gamma_2q_2)+3K_2(2q_1+\beta_2w). \tag{5.7}$$

(c) For $[A^{(1)},B^{(1)}]$,

$$0=2K_{11}(\gamma_2\beta_1-\gamma_1\beta_2+(-\gamma_2+\beta_1)q_1+q_2-q_1^2)$$
$$+4K_{12}(\alpha_2\gamma_1-\alpha_1\gamma_2-\alpha_1q_1-\gamma_2q_2-q_1q_2)$$
$$+2K_{22}(\alpha_1\beta_2-\alpha_2\beta_1+\alpha_2q_1+(\beta_2-\alpha_1)q_2-q_2^2)$$
$$+3K_1(-2\gamma_2+\beta_1-3q_1)+3K_2(-2\alpha_1+\beta_2-3q_2)-6K.$$

(d) For $[A^{(1)},B^{(0)}]$,

$$0=2\gamma_2q_1K_{11}+4\gamma_2q_2K_{12}-2(q_1/2+\beta_2q_2)K_{22}+6\gamma_2K_1-3\beta_2K_2. \tag{5.8}$$

(e) For $[A^{(0)},B^{(0)}]$,

$$0=\gamma_1\beta_2K_{11}+2(\alpha_1\gamma_2-\alpha_2\gamma_1)K_{12}-\alpha_1\beta_2K_{22}. \tag{5.9}$$

What we present here is an illustrative characterization of different types of fundamental equations in terms of matrix pairs (A,B). This provides a good intuitive description of the world of qLN equations and helps to specify where two particular classes—separable potentials and driven systems—belong. An alternative way of classifying qLN equations with two quadratic integrals of motion is to simplify the fundamental equation (4.5) with the use of affine transformations as has been done for the Bertrand–Darboux equation[2] (see Example 5.1 below). This may amount to a similar picture as we have presented above, but the principles of simplification of the third-order polynomials at K_{rr}, K_{rw}, and K_{ww} are more difficult to discern. This is yet to be done.

Example 5.1: The classification of types of the Bertrand–Darboux equation with respect to Euclidean transformations leads to three forms of this equation which are separable in either elliptic, parabolic, or Cartesian coordinates. According to Corollary 3.1, if a potential two-dimensional Newton system

$$\ddot{q}_1 = -\frac{\partial V}{\partial q_1}, \quad \ddot{q}_2 = -\frac{\partial V}{\partial q_2}$$

with $V = V(q_1, q_2)$ possesses a second integral of motion of the form $E = \dot{q}^t A \dot{q} + k(q)$ with $A \in A^{(2)}$, then it has the qLN form $\ddot{q} = -\frac{1}{2}A^{-1}\nabla k = -\frac{1}{2}B^{-1}\nabla V$ with $B = \frac{1}{2}I$ where I is 2×2 identity matrix. Moreover, the potential V must satisfy the Bertrand–Darboux equation (3.2). This system belongs to the class $[A^{(2)}, B^{(0)}]$ with $\beta_2 = 0$ and with $\alpha_2 = \gamma_2 = \frac{1}{2}$. The corresponding fundamental equation is exactly the Bertrand–Darboux equation since in this case $K = V/\det(B) = 4V$. The simplification procedure described above does not alter the form of the matrix $B = \frac{1}{2}I$ and so the corresponding fundamental equation (5.7) attains the form

$$0 = (V_{22} - V_{11})q_1 q_2 + V_{12}(q_1^2 - q_2^2 + \gamma_1 - \alpha_1) - 6q_2 V_1 + 6q_1 V_2, \tag{5.10}$$

which has only one essential parameter $\gamma_1 - \alpha_1$. This form of the Bertrand–Darboux equation separates in the elliptic coordinates.[2] The specification $a_1 = 0$ reduces the class $[A^{(2)}, B^{(0)}]$ to $[A^{(1)}, B^{(0)}]$. The corresponding fundamental equation after the simplification procedure attains the form (5.8). In the case when $B = \frac{1}{2}I$, the final form of B (after simplification) will be exactly the same (i.e., with $\beta_2 = 0$, $\alpha_2 = \gamma_2 = \frac{1}{2}$), and so the fundamental equation (5.8) reads

$$0 = q_1(V_{11} - V_{22}) + 2q_2 V_{12} + 3V_1.$$

It does not contain any parameters now. This equation separates in the parabolic coordinates.[2] Further specification $b_1 = c_1 = 0$ leads to the class $[A^{(0)}, B^{(0)}]$ of constant symmetric matrices. The corresponding fundamental equation in the course of simplification attains the form (5.9) which in the case $B = \frac{1}{2}I$ (again, this form of B survives the simplification procedure—in this case just the diagonalization of A by a rotation) yields $V_{\xi\eta} = 0$, where ξ, η are coordinates which originate by a rotation of the Cartesian coordinates x,y. This is the case of the Bertrand–Darboux equation separable in (rotated) Cartesian coordinates.

The above example indicates that the fundamental equation plays the same role in the theory of qLN equations as the Bertrand–Darboux equation does in the theory of separable potential forces $M = -\partial V/\partial q$. For separable potentials the characteristic coordinates of the Bertrand–Darboux equation determine the coordinates of separation which makes it possible to solve the corresponding Newton equations by quadratures. In Sec. VII we prove a similar result for the class of two-dimensional driven systems by showing that the characteristic coordinates of the fundamental equation associated with a given driven system separate this system, i.e., that in these coordinates it is possible to integrate the system by quadratures. The question whether the characteristic coordinates of the fundamental equation separate general qLN systems admitting two integrals of motion remains to be investigated. We have here to do with a much broader theory depending on five essential parameters while the Bertrand–Darboux equation depends on one parameter only.

VI. HAMILTONIAN STRUCTURES AND COMPLETE INTEGRABILITY

In this section we will establish a Hamiltonian formulation of two-dimensional qLN systems and discuss their complete integrability. Let us consider first the qLN system $0=\delta^+ E = 2A(\ddot{q} + \frac{1}{2}A^{-1}\nabla k(q))$ generated by the function $E=\dot{q}^t A(q)\dot{q}+k(q)$, $q=(q_1,q_2)^t$ with the 2×2 matrix $A(q)$ satisfying the cyclic conditions (2.3). This system usually does not have any Lagrangian formulation and thus it does not have the standard Hamiltonian formulation. However, we can always embed this system in a Hamiltonian qLN system in the five-dimensional phase space of variables (q_1,q_2,p_1,p_2,d) as the following theorem states.

Theorem 6.1 (Hamiltonian form of qLN systems): *Let*

$$0=\ddot{q}+\tfrac{1}{2}A^{-1}(q)\nabla(k(q)+d\lambda\,\det(A(q))) \tag{6.1}$$

with $q=(q_1,q_2)^t$ *be the qLN system generated by*

$$\hat{E}=\dot{q}^t A(q)\dot{q}+k(q)+d\lambda\,\det(A(q))\equiv E+d\lambda\,\det(A)$$

with some constant λ *and with* $d\in\mathbf{R}$. *Let also* $\mathcal{M}$ *be the extended five-dimensional phase space of variables* (q_1,q_2,p_1,p_2,d) *with* $p_i=\dot{q}_i$, $i=1,2$. *Then the system* (6.1) *is equivalent to*

$$\begin{bmatrix}\dot{q}\\ \dot{p}\\ \dot{d}\end{bmatrix}=\left[\begin{array}{cc|c} 0 & -(\lambda/2)G(q) & p\\ (\lambda/2)G^t(q) & -(\lambda/2)F(q,p) & \hat{M}(q,d)\\ \hline -p^t & -\hat{M}^t(q,d) & 0\end{array}\right]\nabla_{\mathcal{M}}d\equiv\Pi_A\nabla_{\mathcal{M}}d, \tag{6.2}$$

where $\nabla_{\mathcal{M}}=(\partial/\partial q_1,\partial/\partial q_2,\partial/\partial p_1,\partial/\partial p_2,\partial/\partial d)^t$ *is the gradient operator in* $\mathcal{M}$ *and where the* 2×2 *matrices G and F and the vector* $\hat{M}$ *are given by*

$$G(q)=\det(A)A^{-1}=\begin{bmatrix} A_{22} & -A_{12}\\ -A_{12} & A_{11}\end{bmatrix},$$

$$F_{12}(q,p)=\frac{1}{2}\left(\frac{\partial A_{22}}{\partial q_1}p_2-\frac{\partial A_{11}}{\partial q_2}p_1\right),\quad F=-F^t,$$

$$\hat{M}(q,d)=M(q)-\tfrac{1}{2}d\lambda A^{-1}\nabla(\det(A))$$

with $M(q)=-\frac{1}{2}A^{-1}\nabla k$ *being the force of the qLN system* $0=\delta^+E$. *Moreover, the antisymmetric matrix* Π_A *is Poisson and so* (6.2) *is the Hamiltonian formulation of (6.1).*

Notice that the matrix G obtained above is symmetric due to the symmetry of A.

Proof: Since $\nabla_{\mathcal{M}}d=(0,0,0,0,1)^t$, the equation (6.2) yields $\dot{q}=p$, $\dot{p}=\hat{M}=-\frac{1}{2}A^{-1}\nabla(k+d\lambda\,\det(A))$, $\dot{d}=0$, i.e., it reproduces (6.1). The matrix Π_A is antisymmetric and it is straightforward to verify that it satisfies the Jacobi identity in the phase space $\mathcal{M}$. Q.E.D.

We remind the reader that the operator $\Pi: T^*\mathcal{M}\to T\mathcal{M}$ mapping fiberwise the cotangent bundle $T^*\mathcal{M}$ of $\mathcal{M}$ into the tangent bundle $T\mathcal{M}$ is Poisson if the bilinear mapping $\{\cdot,\cdot\}_\Pi: C^\infty(\mathcal{M})\times C^\infty(\mathcal{M})\to C^\infty(\mathcal{M})$ defined for any pair of functions $f,g:\mathcal{M}\to\mathbf{R}$ by

$$\{f,g\}_\Pi=\langle\nabla_{\mathcal{M}}f,\Pi\nabla_{\mathcal{M}}g\rangle$$

(where $\langle\cdot,\cdot\rangle$ is the dual map between cotangent and tangent spaces of $\mathcal{M}$) is a Poisson bracket.

Remark 6.2: In the hyperplane $d=0$ the solutions of (6.2) coincide with the solutions of $\ddot{q}=-\frac{1}{2}A^{-1}\nabla k$. Thus our original qLN system $\ddot{q}=-\frac{1}{2}A^{-1}\nabla k$ is in a natural way embedded in the Hamiltonian system (6.2).

Proposition 6.3: The function

$$\hat{E}=p^tA(q)p+k(q)+d\lambda\ \det(A(q))$$

is a Casimir function for the Poisson operator Π_A *in* (6.2), *that is,* $\Pi_A\nabla_{\mathcal{M}}\hat{E}=0$.

One can check this proposition by a direct verification.

A statement converse to the second statement of Theorem 6.1 also holds.

Theorem 6.4: *Let the antisymmetric matrix*

$$\Pi=\left[\begin{array}{cc|c} 0 & -(\lambda/2)G(q) & p \\ (\lambda/2)G^t(q) & -(\lambda/2)F(q,p) & \hat{M}(q,d) \\ \hline -p^t & -\hat{M}^t(q,d) & 0 \end{array}\right] \tag{6.3}$$

be a Poisson operator in the space of variables (q,p,d). *Then*

(1) $G(q)$ *must have the form*

$$G(q)=\begin{bmatrix} aq_1^2+cq_1+\gamma & aq_1q_2+\frac{b}{2}q_1+\frac{c}{2}q_2-\frac{\beta}{2} \\ * & aq_2^2+bq_2+\alpha \end{bmatrix} \tag{6.4}$$

(thus it is symmetric) with some constants $a,b,c,\alpha,\beta,\gamma$ *and so*

$$G=\begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix} \tag{6.5}$$

for some symmetric matrix $A(q)$ *satisfying the cyclic conditions (2.3). In other words,* $\Pi=\Pi_A$ *with* Π_A *defined in (6.2) and with*

$$A=\begin{bmatrix} G_{22} & -G_{12} \\ -G_{21} & G_{11} \end{bmatrix}.$$

(2) $F(q,p)$ *must have the form*

$$F_{12}(q,p)=\frac{1}{2}\left(\frac{\partial A_{22}}{\partial q_1}p_2-\frac{\partial A_{11}}{\partial q_2}p_1\right),\quad F=-F^t. \tag{6.6}$$

(3) $\hat{M}(q,d)$ *must have the form*

$$\hat{M}(q,d)=M(q)+d\lambda N(q),$$

where $-2AM(q)=\nabla k$ *for some function* $k(q)$, *so if* $\det(G)\neq 0$, *then* $M(q)=-\frac{1}{2}A^{-1}\nabla k$, *and where* $N(q)=-\frac{1}{2}A^{-1}\nabla(\det(A))$.

Proof: The conditions $\{\{q_i,q_j\}_\Pi,q_k\}_\Pi+\text{cycl}=0$ and $\{\{q_i,q_j\}_\Pi,p_k\}_\Pi+\text{cycl}=0$ (where ''cycl'' means the cyclic permutation of expressions) hold identically due to the block structure of Π. The condition $\{\{q_i,q_j\}_\Pi,d\}_\Pi+\text{cycl}=0$ yields the symmetry of G: $G=G^t$. Further,

$$0=\{\{q_i,p_j\}_\Pi,d\}_\Pi+\text{cycl}=-\frac{\lambda}{2}\left(p_1\frac{\partial G_{ij}}{\partial q_1}+p_2\frac{\partial G_{ij}}{\partial q_2}\right)+\frac{\lambda}{2}F_{ij}-p_i\frac{\partial \hat{M}_j}{\partial d}.$$

Let us denote the right-hand side of the above equality by $-(\lambda/2)R_{ij}$. Notice that $\partial\hat{M}_j/\partial d$ cannot depend on d, and so we have $\partial\hat{M}_j/\partial d=\lambda N_j(q)$ for some vector $N(q)=(N_1(q),N_2(q))^t$ which yields $\hat{M}(q,d)=M(q)+d\lambda N(q)$ for some vector $M(q)$. By taking linear combinations of the conditions $R_{ij}=0$ and using the symmetry of G and the antisymmetry of F we get the following sets of equations:

$$\frac{\partial G_{11}}{\partial q_2}=\frac{\partial G_{22}}{\partial q_1}=0, \quad \frac{\partial G_{11}}{\partial q_1}=2N_1, \quad \frac{\partial G_{22}}{\partial q_2}=2N_2, \tag{6.7}$$

$$\frac{\partial G_{12}}{\partial q_1}=N_2, \quad \frac{\partial G_{12}}{\partial q_2}=N_1, \tag{6.8}$$

$$F_{12}=p_2N_1-p_1N_2. \tag{6.9}$$

The equations (6.7) show that G_{11} and N_1 depend only on q_1 and that G_{22} and N_2 depend only on q_2. The equations (6.8) give $\partial N_1/\partial q_1=\partial^2 G_{12}/\partial q_1\partial q_2=\partial N_2/\partial q_2$ and so all terms in this expression must be equal to a constant a. Integration yields

$$N_1=aq_1+c/2, \quad N_2=aq_2+b/2, \tag{6.10}$$

where b and c are integration constants. Substituting (6.10) into (6.7) and (6.8) and integrating we get (6.4). If we now introduce the symmetric matrix A by the equality (6.5) and use (6.7) then (6.9) will attain the form (6.6).

It is straightforward to check that with the above forms of F and G the conditions $\{\{p_i,p_j\}_\Pi,p_k\}_\Pi+\mathrm{cycl}=0$ and $\{\{q_i,p_j\}_\Pi,p_k\}_\Pi+\mathrm{cycl}=0$ are satisfied identically.

Further, the condition $\{\{p_1,p_2\}_\Pi,d\}_\Pi+\mathrm{cycl}=0$ after some calculations attains the form

$$0=\frac{\partial}{\partial q_1}(G_{11}M_2-G_{21}M_1)-\frac{\partial}{\partial q_2}(G_{22}M_1-G_{12}M_2),$$

which means that in the vector

$$\begin{bmatrix} G_{22}M_1-G_{12}M_2 \\ -G_{21}M_1+G_{11}M_2 \end{bmatrix}=\begin{bmatrix} G_{22} & -G_{12} \\ -G_{21} & G_{11} \end{bmatrix}\begin{bmatrix} M_1 \\ M_2 \end{bmatrix}=AM,$$

the mixed derivatives of its components are equal and so this vector is equal to the gradient of some function $-\frac{1}{2}k(q)$, that is, $AM=-\frac{1}{2}\nabla k$ or $M=-\frac{1}{2}A^{-1}\nabla k$.

Finally, by direct calculation we verify that $N=-\frac{1}{2}A^{-1}\nabla(\det(A))$ and so statement (3) of the theorem is proved. Q.E.D.

Remark 6.5: This theorem generalizes the result of Ref. 6. In particular, if we assume $M=-\nabla V(q)$, then we recover the known second Poisson operator for separable potential systems.[7]

Notice that $\hat{M}$ is the force of the two-dimensional qLN system (6.1). This means that every Poisson operator of the form (6.3) is a Poisson operator for some qLN system of the form (6.1).

We are now in position to investigate complete integrability of qLN systems admitting two quadratic, functionally independent integrals of motion. Notice first that Theorem 6.4 provides us with an alternative way of characterizing qLN systems generated by a quadratic integral of motion *E:* by starting with a Poisson operator of the form (6.3) we arrive at qLN systems generated by the Hamiltonian $\mathrm{H}(q,p,d)=d$ which admit a quadratic integral E. In a similar way the following theorem characterizes all qLN systems admitting two independent quadratic integrals E,F.

Theorem 6.6 (Poisson pencil): *Consider the antisymmetric operator*

$$\Pi_\mu=\left[\begin{array}{cc|c} 0 & -(\lambda/2)G_\mu(q) & p \\ (\lambda/2)G_\mu^t(q) & -(\lambda/2)F_\mu(q,p) & M(q)+d\lambda N_\mu(q) \\ \hline -p^t & -M^t(q)-d\lambda N_\mu^t(q) & 0 \end{array}\right], \tag{6.11}$$

where

$$G_\mu=\begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11}\end{bmatrix}-\mu\begin{bmatrix} B_{22} & -B_{12} \\ -B_{21} & B_{11}\end{bmatrix}\equiv G_A-\mu G_B,$$

with both matrices A and B satisfying the cyclic conditions (2.3),

$$[F_\mu]_{12}=\frac{1}{2}\left(\frac{\partial(A_{22}-\mu B_{22})}{\partial q_1}p_2-\frac{\partial(A_{11}-\mu B_{11})}{\partial q_2}p_1\right)\equiv[F_A]_{12}-\mu[F_B]_{12}$$

(*with* $F=-F^t$, $F_A=-F_A^t$, $F_B=-F_B^t$) *and*

$$N_\mu=-\tfrac{1}{2}A^{-1}\nabla(\det(A))+\tfrac{1}{2}\mu B^{-1}\nabla(\det(B))\equiv N_A-\mu N_B.$$

Then Π_μ *is Poisson if and only if*

$$M(q)=-\tfrac{1}{2}A^{-1}\nabla k=-\tfrac{1}{2}B^{-1}\nabla l \tag{6.12}$$

for some functions $k(q)$ *and* $l(q)$. *Moreover, if we let*

$$\Pi_\mu=\Pi_1-\mu\Pi_2\equiv\left[\begin{array}{cc|c} 0 & -(\lambda/2)G_A & p \\ (\lambda/2)G_A^t & -(\lambda/2)F_A & M+d\lambda N_A \\ \hline -p^t & -M^t-d\lambda N_A^t & 0\end{array}\right]-\mu\left[\begin{array}{cc|c} 0 & -(\lambda/2)G_B & 0 \\ (\lambda/2)G_B^t & -(\lambda/2)F_B & d\lambda N_B \\ \hline 0 & -d\lambda N_B^t & 0\end{array}\right],$$

then both operators Π_1 *and* Π_2 *are Poisson and so* $\Pi_\mu=\Pi_1-\mu\Pi_2$ *is a Poisson pencil.*

Proof: According to the proof of Theorem 6.4 the matrix Π_μ satisfies all the Jacobi identities except possibly for $\{\{p_1,p_2\}_{\Pi_\mu},d\}_{\Pi_\mu}+\text{cycl}=0$, since Π_μ differs from $\Pi_{A-\mu B}=\Pi_A-\mu\Pi_B$ by the form of $M(q)$ only. Like in the proof of Theorem 6.4 we find that $\{\{p_1,p_2\}_{\Pi_\mu},d\}_{\Pi_\mu}+\text{cycl}=0$ yields that the mixed derivatives of the components of the vector $-2(A-\mu B)M$ are equal and so $-2(A-\mu B)M=\nabla(k-\mu l)$ for some functions $k(q)$ and $l(q)$. By comparing coefficients at different powers of μ we get $-2AM=\nabla k$ and $-2BM=\nabla l$ and thus $M=-\frac{1}{2}A^{-1}\nabla k=-\frac{1}{2}B^{-1}\nabla l$.

Further, $\Pi_1=\Pi_A$ in the notation of Theorem 6.4 so it is Poisson. Easy calculation shows that Π_2 is Poisson, too. Q.E.D.

The above theorem states that if $M(q)$ is the force of a qLN system admitting two functionally independent integrals of motion, then the matrix Π_μ is a Poisson pencil. We will establish its Casimir function, which will be a polynomial in μ. This will lead to a bi-Hamiltonian chain containing the qLN system (6.1). We will prove that this chain is completely integrable. In this way we will show that our original qLN system $\ddot{q}=-\frac{1}{2}A^{-1}\nabla k=-\frac{1}{2}B^{-1}\nabla l$ can be naturally embedded in a completely integrable bi-Hamiltonian system.

Proposition 6.7: Suppose that Π_μ *is Poisson, i.e., that (6.12) is satisfied. Then the function*

$$H_\mu=p^t(A-\mu B)p+k-\mu l+d\lambda\ \det(A-\mu B) \tag{6.13}$$

is a Casimir function for Π_μ, i.e., $\Pi_\mu\nabla H_\mu=0$.

Proof: This proposition is a consequence of Proposition 6.3. If we modify the matrix Π_A by substituting the matrix A by $A-\mu B$ and substituting k with $k-\mu l$ we obtain the matrix

$$\left[\begin{array}{cc|c} 0 & -\frac{\lambda}{2}G_\mu & p \\ \frac{\lambda}{2}G_\mu^t & -\frac{\lambda}{2}F_\mu & \widetilde{M}_\mu \\ \hline -p^t & -\widetilde{M}_\mu^t & 0 \end{array}\right], \tag{6.14}$$

where

$$\widetilde{M}_\mu=-\tfrac{1}{2}(A-\mu B)^{-1}\nabla(k-\mu l)-\tfrac{1}{2}d\lambda(A-\mu B)^{-1}\nabla(\det(A-\mu B)).$$

Due to Proposition 6.3 the function (6.13) is the Casimir of (6.14). However, (6.14) is, in fact, equal to Π_μ since it can be verified that $-\frac{1}{2}(A-\mu B)^{-1}\nabla(k-\mu l)=-\frac{1}{2}A^{-1}\nabla k=-\frac{1}{2}B^{-1}\nabla l$ and that $(A-\mu B)^{-1}\nabla(\det(A-\mu B))=A^{-1}\nabla(\det(A))-\mu B^{-1}\nabla(\det(B))$. Q.E.D.

Let us collect terms in H_μ at different powers of μ:

$$H_\mu=p^tAp+k+d\lambda\det(A)+\mu(-p^tBp-l-d\lambda Y)+\mu^2(d\lambda\det(B))\equiv\hat{E}+\mu\hat{F}+\mu^2\hat{H}$$

with $Y=B_{11}A_{22}+B_{22}A_{11}-2B_{12}A_{12}$. Then the above proposition gives

$$\begin{aligned} 0=\Pi_\mu\nabla H_\mu&=(\Pi_1-\mu\Pi_2)\nabla(\hat{E}+\mu\hat{F}+\mu^2\hat{H}) \\ &=\Pi_1\nabla\hat{E}+\mu(\Pi_1\nabla\hat{F}-\Pi_2\nabla\hat{E})+\mu^2(\Pi_1\nabla\hat{H}-\Pi_2\nabla\hat{F})-\mu^3\Pi_2\nabla\hat{H}. \end{aligned}$$

By equating to zero the coefficients at different powers of μ we obtain the following bi-Hamiltonian chain:

$$\begin{aligned} \Pi_1\nabla\hat{E}&=0, \\ \Pi_1\nabla\hat{F}&=\Pi_2\nabla\hat{E}, \\ \Pi_1\nabla\hat{H}&=\Pi_2\nabla\hat{F}, \\ 0&=\Pi_2\nabla\hat{H}. \end{aligned} \tag{6.15}$$

Theorem 6.8: *The bi-Hamiltonian chain (6.15) is completely integrable, i.e., both nontrivial bi-Hamiltonian vector fields,*

$$V_1=\Pi_1\nabla\hat{F}=\Pi_2\nabla\hat{E},\quad V_2=\Pi_1\nabla\hat{H}=\Pi_2\nabla\hat{F},$$

in (6.15) are completely integrable.

Proof (modification of the proof of Liouville–Arnold theorem[8]): Consider the two-dimensional manifold $\mathcal{N}=\{x\in\mathcal{M}:\hat{E}(x)=E_0,\hat{F}(x)=F_0,\hat{H}(x)=H_0\}$ in $\mathcal{M}$. Poisson brackets of all pairs of $\hat{E},\hat{F},\hat{H}$ induced by both structures Π_0 and Π_1 are equal to zero, since the functions $\hat{E},\hat{F},\hat{H}$ all belong to the same bi-Hamiltonian chain. For instance, $\{\hat{F},\hat{H}\}_{\Pi_1}=\langle\nabla\hat{F},\Pi_1\nabla\hat{H}\rangle=\langle\nabla\hat{F},\Pi_2\nabla\hat{F}\rangle=\{\hat{F},\hat{F}\}_{\Pi_2}=0$ with the second equality being a consequence of the bi-Hamiltonian structure of V_2. It follows that the Lie bracket $[V_1,V_2]$ of both vector fields V_1 and V_2 is equal to zero, $[V_1,V_2]=[\Pi_1\nabla\hat{F},\Pi_1\nabla\hat{H}]=0$, since the mappings $\Pi_i\nabla$ $(i=1,2)$ are Lie algebra homomorphisms between the Lie algebra of vector fields on $\mathcal{M}$ and the Lie algebra of all smooth functions on $\mathcal{M}$ with the Lie bracket defined by $[f_1,f_2]=\{f_1,f_2\}_{\Pi_i}$. Moreover, $\langle\nabla\hat{E},V_1\rangle=\langle\nabla\hat{E},\Pi_2\nabla\hat{E}\rangle$

$=\{\hat{E},\hat{E}\}_{\Pi_2}=0$ and similarly $\langle\nabla\hat{F},V_1\rangle=\langle\nabla\hat{H},V_2\rangle=0$, which proves that V_1 is tangent to $\mathcal{N}$. In the same way one can show that V_2 is also tangent to $\mathcal{N}$. Direct verification shows that V_1 and V_2 are linearly independent. We thus have a two-dimensional submanifold $\mathcal{N}$ in $\mathcal{M}$ equipped with a pair of linearly independent, commuting vector fields V_1 and V_2. We can now apply the construction of Liouville-Arnold[8] and conclude that both V_1 and V_2 are completely integrable. Q.E.D.

Corollary 6.9: The qLN system $\ddot{q}=M=-\frac{1}{2}A^{-1}\nabla k=-\frac{1}{2}B^{-1}\nabla l$ *with two linearly independent matrices A and B satisfying the cyclic conditions (2.3) is completely integrable in the sense that the trajectories of the system*

$$\begin{bmatrix}\dot{q}\\ \dot{p}\end{bmatrix}=\begin{bmatrix}p\\ M\end{bmatrix} \tag{6.16}$$

coincide on the hyperplane $d=0$ *with the trajectories of the completely integrable five-dimensional system*

$$\begin{bmatrix}\dot{q}\\ \dot{p}\\ \dot{d}\end{bmatrix}=V_2=\Pi_1\nabla\hat{H}=\Pi_2\nabla\hat{F}. \tag{6.17}$$

Proof: Consider the vector field V_2 from (6.15). Obviously

$$V_2=\Pi_1\nabla\hat{H}=\lambda\,\det(B)\Pi_1\nabla d+\lambda d\Pi_1\nabla(\det(B))$$

and so in the hyperplane $d=0$ we have

$$V_2|_{d=0}=\lambda\,\det(B)\begin{bmatrix}p\\ M\\ 0\end{bmatrix}, \tag{6.18}$$

which means that the hyperplane $d=0$ is invariant with respect to the action of the vector field V_2. The formula (6.18) also shows that in the hyperplane $d=0$ the vector field of the system (6.17) is parallel to the vector field of the system (6.16) and so their trajectories must coincide. Q.E.D.

Thus we have shown that the system (6.16) is embedded in the completely integrable bi-Hamiltonian system (6.17). The trajectories of (6.16) stay on the intersection of invariant manifolds for (6.17) with the hyperplane $d=0$. Also, since we can now solve the system (6.17) by quadratures the time evolution of the coefficient $\lambda\,\det(B)$ in (6.18) can be calculated which makes it possible to solve the system (6.16) by quadratures too.

VII. NEW TYPES OF SEPARATION VARIABLES FOR DRIVEN qLN SYSTEMS

In this section we study an important class of two-dimensional qLN equations called driven systems. We find for all such systems their separation variables and prove their integrability by quadratures. The variables of separation are of a completely new type: they consist of families of conics which are non-confocal in contrast with the classical separability theory for potential systems.

We remind the reader that we call a two-dimensional Newton system *driven* if one of the two differential equations depends on one variable only. By renaming the variables if necessary, we can always arrange for such a system to take the form

$$\begin{aligned}\ddot{q}_1&=M_1(q_1,q_2),\\ \ddot{q}_2&=M_2(q_2).\end{aligned} \tag{7.1}$$

The second equation can be solved on its own and its solution $q_2(x)$ then determines the equation for q_1, which explains the name ''driven.'' A driven system always has one integral of motion $F=\dot{q}_2^2/2-\int M_2\,dq_2$, obtained by integrating the second equation once, but, in general, there need not exist any others.

Here we shall consider driven systems that admit a quasi-Lagrangian formulation $\ddot{q}=-\frac{1}{2}A^{-1}\nabla k(q)$. Here, as usual, $A(q)$ is a nondegenerate 2×2 matrix satisfying the cyclic conditions (2.3), i.e., a matrix of the form

$$A=\begin{bmatrix} aq_2^2+bq_2+\alpha & -aq_1q_2-\frac{b}{2}q_1-\frac{c}{2}q_2+\frac{\beta}{2} \\ -aq_1q_2-\frac{b}{2}q_1-\frac{c}{2}q_2+\frac{\beta}{2} & aq_1^2+cq_1+\gamma \end{bmatrix}. \tag{7.2}$$

Such a system always has *two* functionally independent integrals of motion $E=\dot{q}^tA\dot{q}+k(q)$ and $F=\dot{q}_2^2/2-\int M_2\,dq_2$.

By examining the second component of the equation $\ddot{q}=-\frac{1}{2}A^{-1}\nabla k(q)$, we immediately see that a qLN system is driven iff

$$A_{12}\partial_1 k-A_{11}\partial_2 k=2\,\det(A)M_2(q_2), \tag{7.3}$$

for some function $M_2(q_2)$ depending on q_2 only. We can produce driven qLN systems with any given $M_2(q_2)$ and $A(q)$ by solving for $k(q)$ in this equation. The case $A_{11}=0$ is degenerate and will be treated separately later (see Remark 7.6), so we assume from now on that $A_{11}\neq 0$.

We start by introducing separation variables for (7.1) as characteristic coordinates for (7.3).

Definition 7.1: Define curvilinear coordinates $(u,v)=(u(q),v(q))$ as follows. Let u be a parameter indexing the family of characteristic curves of (7.3) given by

$$\dot{q}(x)=\begin{bmatrix} A_{12}(q(x)) \\ -A_{11}(q(x)) \end{bmatrix}, \tag{7.4}$$

and let $v=q_2$.

In other words, the curves given by (7.4) are the coordinate curves of constant u. For a given matrix A these curves can be explicitly calculated. In Theorem 7.7 we will describe these curves more explicitly. Let us just note for the moment that they are not parallel to the curves of constant v, because of the assumption $A_{11}\neq 0$. Thus the above description really defines a coordinate system (at least locally). There is some freedom in the choice of u, but this will not affect our results. By abuse of notation we will write $f(q_1,q_2)$ and $f(u,v)$ for the same function f expressed in different coordinate systems.

Lemma 7.2: The general solution of (7.3) is

$$k(u,v)=f(u)+D(u,v)g(v), \tag{7.5}$$

where f is an arbitrary function, $D=\det(A)$, and

$$g(q_2)=\frac{-2}{A_{11}(q_2)}\int M_2(q_2)\,dq_2.$$

Proof: Along each characteristic curve $q(x)$ given by (7.4) we can consider (7.3) as an ODE

$$\frac{d}{dx}k(q(x))=2D(q(x))M_2(q_2(x)),$$

with general solution

$$k(q(x))=D(q(x))g(q_2(x))+f,$$

where f is a constant of integration. This can be verified by direct differentiation; the cyclic conditions imply that

$$\frac{d}{dx}D(q(x))=\partial_1 D\dot{q}_1+\partial_2 D\dot{q}_2=\partial_1 DA_{12}+\partial_2 D(-A_{11})=-D\partial_2 A_{11},$$

and thus

$$\begin{aligned}\frac{d}{dx}k(q)=\frac{d}{dx}(D(q)g(q_2)+f)=\frac{dD}{dx}g+D\partial_2 g\dot{q}_2\\ =-D\partial_2 A_{11}g+D\partial_2 g(-A_{11})\\ =-D\partial_2(A_{11}g)=2DM_2.\end{aligned}$$

The constant of integration f can be different for different characteristic curves, so when we express the result in terms of u and v, f will depend on u (but not on v). Q.E.D.

Lemma 7.3: Equation (7.3) is equivalent, under the substitution $k=K\det(A)$, to the equation

$$A_{12}\partial_{11}K-A_{11}\partial_{12}K-\tfrac{3}{2}\partial_2 A_{11}\partial_1 K=0, \tag{7.6}$$

which is the fundamental equation (4.5) associated with the matrices A and

$$B=\begin{bmatrix}0 & 0\\ 0 & \tfrac{1}{2}\end{bmatrix}.$$

Proof: Equation (7.3) implies

$$\partial_1\left(\frac{-A_{12}\partial_1 k+A_{11}\partial_2 k}{\det(A)}\right)=0.$$

Conversely, this expression can be integrated to give (7.3), where $M_2(q_2)$ is an arbitrary function of integration. By substituting $k=K\det(A)$ and simplifying the resulting expression using the cyclic conditions one obtains (7.6). Comparison with the general expression for the fundamental equation in Theorem 4.1 proves the second statement of the lemma. Q.E.D.

Remarks 7.4: The fundamental equation (7.6) is hyperbolic. Its characteristic coordinates are precisely the coordinates (u,v) of Definition (7.1). The general solution is $K(u,v)=f(u)/D(u,v)+g(v)$, as can be seen by combining the above lemmas.

Let us turn to the question of how to integrate a driven qLN system. The solution $q_2(x)$ of the second equation can be found by quadrature from $F=\dot{q}_2^2/2-\int M_2\,dq_2$:

$$\int\frac{dq_2}{\sqrt{2F+2\int M_2 dq_2}}=\pm\int dx. \tag{7.7}$$

Inserting $q_2(x)$ and $\dot{q}_2(x)$ into

$$E=A_{11}(q_2)\dot{q}_1^2+2A_{12}(q_1,q_2)\dot{q}_1\dot{q}_2+A_{22}(q_1)\dot{q}_2^2+k(q_1,q_2)$$

would give a first-order ODE for $q_1(x)$, but there is no obvious way to solve this equation since the variables q_1 and x do not separate. We will now show how to proceed instead.

Theorem 7.5: *Every driven qLN system can be integrated by quadratures using the characteristic coordinates (u,v) of the fundamental equation (7.6) as separation variables.*

Proof: We use the notation of Lemma 7.2. Let the system be generated by $E=\dot{q}^tA\dot{q}+k(q)$ with $k(u,v)=f(u)+D(u,v)g(v)$. Since $v=q_2$, we can express F as $F=\frac{1}{2}(\dot{v}^2+A_{11}(v)g(v))$ and calculate $v(x)$ by quadrature, as above. Now note that since the curves of constant u by definition have tangent $\dot{q}=(A_{12},-A_{11})^t$, we must have $\nabla u=\rho(q)(A_{11},A_{12})^t$ for some function $\rho(q)$, whose exact form depends on the choice of u. This gives $\dot{u}=\partial_1 u\dot{q}_1+\partial_2 u\dot{q}_2=\rho(q)(A_{11}\dot{q}_1+A_{12}\dot{q}_2)$, and thus

$$\begin{aligned}\dot{u}^2=\rho^2A_{11}\left(A_{11}\dot{q}_1^2+2A_{12}\dot{q}_1\dot{q}_2+\frac{A_{12}^2}{A_{11}}\dot{q}_2^2\right)&=\rho^2A_{11}\left(E-A_{22}\dot{q}_2^2-k(q)+\frac{A_{12}^2}{A_{11}}\dot{q}_2^2\right)\\&=\rho^2A_{11}\left(E-\left(A_{22}-\frac{A_{12}^2}{A_{11}}\right)\dot{v}^2-f(u)-Dg(v)\right)\\&=\rho^2A_{11}\left(E-\frac{D}{A_{11}}(2F-A_{11}g(v))-f(u)-Dg(v)\right)\\&=\rho^2A_{11}\left(E-\frac{2D}{A_{11}}F-f(u)\right).\end{aligned}$$

In order to complete the proof, we will show that $\rho(u,v)=\phi(u)|A_{11}(v)|^{-3/2}$ and $D(u,v)/A_{11}(v)=\psi(u)$ for some functions ϕ and ψ, since this implies that the variables u and x separate. Explicitly, we can then find $u(x)$ from the quadrature

$$\int\frac{du}{\phi(u)\sqrt{E-2\psi(u)F-f(u)}}=\pm\int\frac{dx}{A_{11}(v(x))},\tag{7.8}$$

after which the inverse coordinate transformation gives us $q_1(x)$. Notice that for a given matrix A, the characteristic coordinates (u,v) can be calculated explicitly so that the function ρ and thus ϕ and ψ can be easily calculated and used in the quadrature (7.8) above. The theorem covers, however, all the cases at once without any need of calculating ρ explicitly.

To see that $\rho(u,v)=\phi(u)|A_{11}(v)|^{-3/2}$, note that $\partial_{12}u=\partial_{21}u$ implies that $\rho(q)$ satisfies the PDE

$$0=\partial_1(\rho A_{12})-\partial_2(\rho A_{11})=A_{12}\partial_1\rho-A_{11}\partial_2\rho-\tfrac{3}{2}\partial_2A_{11}\rho,$$

which has the same characteristic curves (7.4) as Eq. (7.3). Along such a curve we determine ρ by integrating

$$\frac{d}{dx}\rho(q(x))=\frac{3}{2}\partial_2A_{11}(q_2(x))\rho(q(x)),$$

which, taking into account $\dot{q}_2(x)=-A_{11}(q_2(x))$, gives

$$\rho(q(x))=\phi|A_{11}(q_2(x))|^{-3/2}.$$

The integration constant ϕ can be different on different characteristic curves, so changing to (u,v) coordinates we obtain

$$\rho(u,v)=\phi(u)|A_{11}(v)|^{-3/2},$$

as desired.

Finally, we calculate the total derivative of the function $\psi(q)=D(q)/A_{11}(q_2)$ along a characteristic curve:

$$\frac{d}{dx}\psi(q(x))=\partial_1\left(A_{22}-\frac{A_{12}^2}{A_{11}}\right)A_{12}-\partial_2\left(A_{22}-\frac{A_{12}^2}{A_{11}}\right)A_{11}.$$

Using the cyclic conditions, we find that this expression is identically zero. This implies that ψ is constant along the coordinate curves of constant u, i.e., $\psi=\psi(u)$. This completes the proof.

Q.E.D.

Remark 7.6: The degenerate case $A_{11}=0$ can be treated as follows. Since $a=b=\alpha=0$, the expression (7.2) for A reduces to

$$A=\begin{bmatrix} 0 & -\frac{c}{2}q_2+\frac{\beta}{2} \\ -\frac{c}{2}q_2+\frac{\beta}{2} & cq_1+\gamma \end{bmatrix} \tag{7.9}$$

and Eq. (7.3) reduces to

$$A_{12}\partial_1 k=2(-A_{12}^2)M_2(q_2),$$

with the general solution

$$k(q)=-2A_{12}(q_2)M_2(q_2)q_1+k_2(q_2).$$

We calculate $q_2(x)$ by quadrature as before. Inserting $q_2(x)$ and $\dot{q}_2(x)$ into $E=2A_{12}\dot{q}_1\dot{q}_2+A_{22}\dot{q}_2^2+k(q_1,q_2)$ yields in this case an equation of the form $\dot{q}_1(x)+\xi(x)q_1(x)=\eta(x)$, from which we can find $q_1(x)$ by quadrature.

Theorem 7.7: *The separation coordinates for driven qLN systems, i.e., the characteristic coordinates (u,v) of the fundamental equation (7.6), are of one of the following types, determined by the coefficients in the matrix A:*

(1) *fanlike hyperbolic, if $a\neq 0$ and $b^2/4-a\alpha=0$;*
(2) *axial hyperbolic, if $a\neq 0$ and $b^2/4-a\alpha<0$;*
(3) *two-point elliptic-hyperbolic, if $a\neq 0$ and $b^2/4-a\alpha>0$;*
(4) *one-point parabolic, if $a=0$ and $b\neq 0$; and*
(5) *parallel parabolic, if $a=0$ and $b=0$.*

Proof: We will compute explicitly the curves given by (7.4), which constitute the curves of constant u. (The curves of constant v are just horizontal lines, since $v=q_2$.) Inserting the explicit expression (7.2) for the matrix A into (7.4), we obtain

$$\begin{bmatrix}\dot{q}_1\\ \dot{q}_2\end{bmatrix}=-\begin{bmatrix} aq_1q_2+\frac{b}{2}q_1+\frac{c}{2}q_2-\frac{\beta}{2} \\ aq_2^2+bq_2+\alpha \end{bmatrix}. \tag{7.10}$$

When solving these equations, we distinguish four different cases, depending on the values of the parameters in A.

The case $a\neq 0$. By setting $r_1=aq_1+c/2$ and $r_2=aq_2+b/2$, which is just rescaling of the axes and translation of the origin, we transform (7.10) into

$$\begin{bmatrix}\dot r_1\\ \dot r_2\end{bmatrix}=\begin{bmatrix}-r_1r_2+C_1\\ -r_2^2+C_2\end{bmatrix},\quad \text{where}\begin{cases}C_1=bc/4+a\beta/2,\\ C_2=b^2/4-a\alpha.\end{cases} \tag{7.11}$$

Subcase $C_2=0$ *(type 1).* Either $r_2=0$, or $r_2=(x+D_1)^{-1}$ and $r_1=C_1(x+D_1)/2+D_2(x+D_1)^{-1}$, where D_1 and D_2 are constants of integration. Eliminating x and writing u instead of D_2, we obtain

$$r_1=\frac{C_1}{2r_2}+ur_2, \tag{7.12}$$

which represents a family of hyperbolas, each with asymptotes $r_2=0$ and $r_2=r_1/u$. The solution $r_2=0$ found above corresponds to the limiting cases $u\to\pm\infty$ (see Fig. 2).

Subcase $C_2\neq 0$ *(type 2 and 3).* The substitution $s_1=r_1-C_1r_2/C_2$, $s_2=r_2$ yields

$$\begin{bmatrix}\dot s_1\\ \dot s_2\end{bmatrix}=\begin{bmatrix}-s_1s_2\\ -s_2^2+C_2\end{bmatrix}, \tag{7.13}$$

and thus

$$\frac{ds_1}{ds_2}=\frac{\dot s_1}{\dot s_2}=\frac{s_2}{s_2^2-C_2}s_1,$$

resulting in

$$s_1^2=u^2|s_2^2-C_2|. \tag{7.14}$$

If $C_2<0$ (type 2), this represents in the s plane a family of hyperbolas centered around the s_1 axis, with asymptotes $s_2=\pm s_1/u$ and vertices $(\pm u\sqrt{-C_2},0)$ [Fig. 3(a)].

If $C_2>0$ (type 3), we obtain in the region $|s_2|>\sqrt{C_2}$ a family of hyperbolas with asymptotes $s_2=\pm s_1/u$ and vertices $(0,\pm\sqrt{C_2})$, and in the region $|s_2|<\sqrt{C_2}$ a family of ellipses with vertices

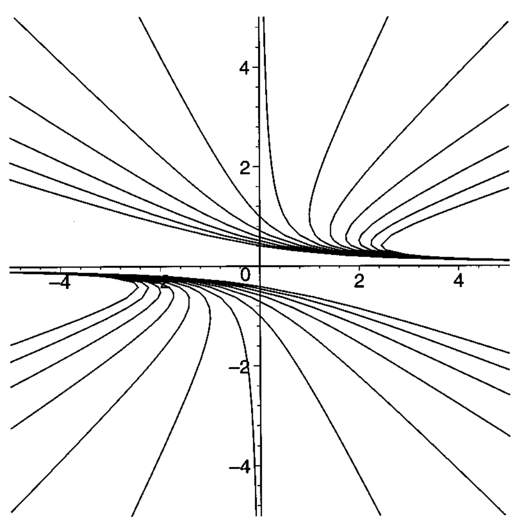

FIG. 2. Fan-like hyperbolic.

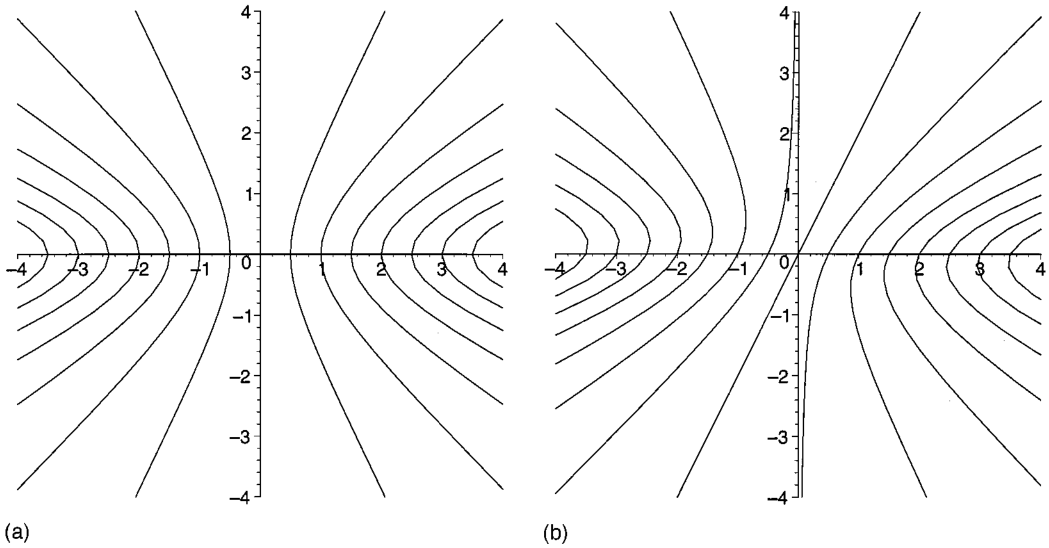

FIG. 3. (a) Axial-hyperbolic in the *s*-plane. (b) Axial-hyperbolic in the *r*-plane.

$(0,\pm\sqrt{C_2})$ and $(0,\pm u\sqrt{C_2})$. The corresponding curves in the r plane are obtained by a shear in the s_1 direction with factor C_1/C_2 [Figs. 3(b) and 4(b)]. They are still hyperbolas and ellipses, but not aligned parallel with the r axes.

The case $a=0$. *Subcase* $b\neq 0$ *(type 4)*. Translating the origin by $r_1=q_1-(\alpha c/b^2+\beta/b)$ and $r_2=q_2+\alpha/b$, we obtain

$$\begin{bmatrix}\dot r_1\\ \dot r_2\end{bmatrix}=\begin{bmatrix}-\dfrac{b}{2}r_1-\dfrac{c}{2}r_2\\ -br_2\end{bmatrix}, \tag{7.15}$$

which yields

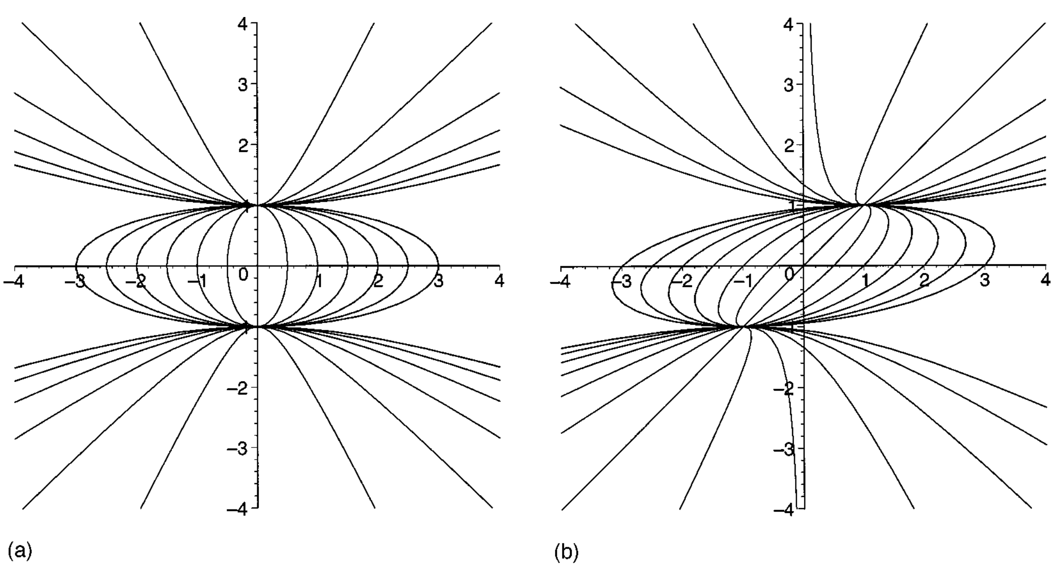

FIG. 4. (a) Two-point elliptic-hyperbolic in the *s*-plane. (b) Two-point elliptic-hyperbolic in the *r*-plane.

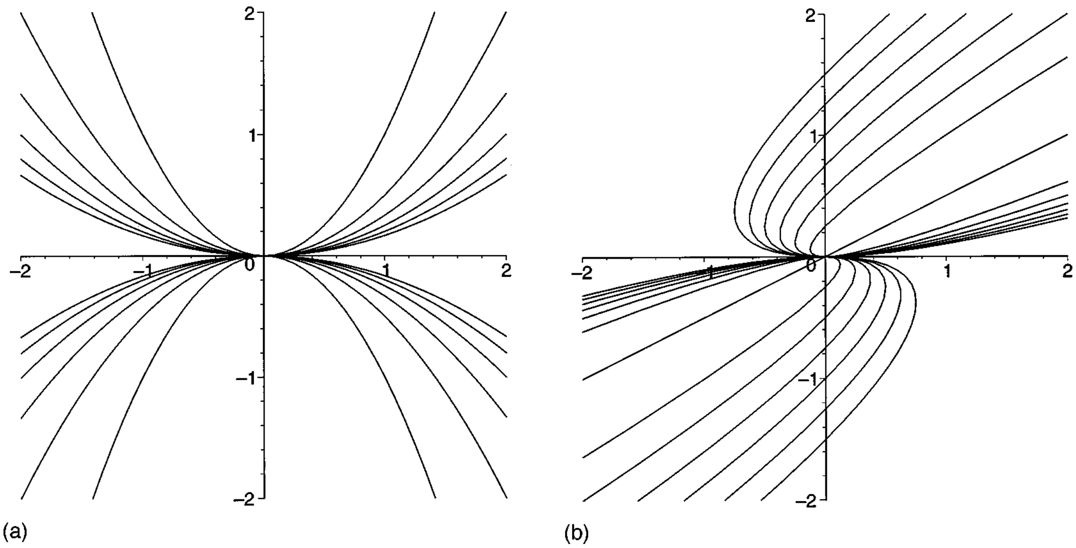

FIG. 5. (a) One-point parabolic in the s-plane. (b) One-point parabolic in the r plane.

$$\left(r_1-\frac{c}{b}r_2\right)^2=ur_2. \tag{7.16}$$

With $s_1=r_1-cr_2/b$, $s_2=r_2$, we obtain in the s plane a family of parabolas $s_2=s_1^2/u$ [Fig. 5(a)]. The corresponding curves in the r plane are parabolas obtained by a shear in the s_1 direction with factor c/b [Fig. 5(b)].

Subcase $b=0$ *(type 5)*. Here we can assume that $\alpha\neq 0$, or else we get the degenerate case $A_{11}=0$. A simple calculation shows that

$$q_1=-\frac{c}{4\alpha}q_2^2+\frac{\beta}{2\alpha}q_2+u, \tag{7.17}$$

which is a family of translated parabolas seen in Fig. 6 (or straight lines, if $c=0$). Q.E.D.

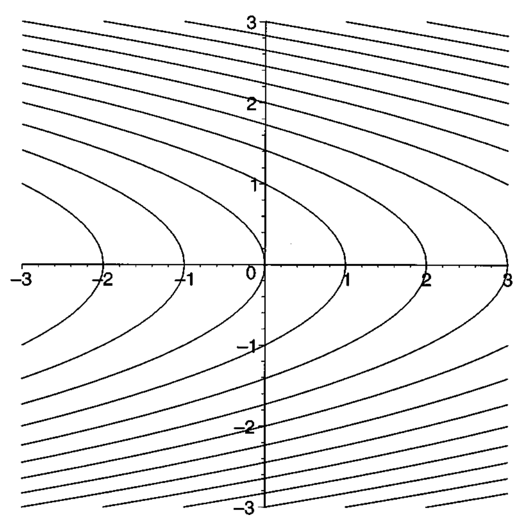

FIG. 6. Parallel-parabolic.

VIII. EXAMPLES AND APPLICATIONS

The notion of a qLN force $M(q)=-\frac{1}{2}A^{-1}(q)\nabla k(q)$ naturally generalizes the concept of a potential force $M(q)=-\nabla k(q)$, which is a special case. The qLN forces admit an integral of motion quadratic in velocities, which in the potential case becomes the energy integral. The function $k(q)$ may be called a ''quasi-potential'' of the force $M(q)$.

A given force is easy to test for the existence of a qLN formulation, provided that one knows the general form of the matrix $A(q)$ solving the cyclic condition (2.3). In two dimensions $A(q)$, given by

$$A(r,w)=\begin{bmatrix} aw^2+bw+\alpha & -arw-\frac{b}{2}r-\frac{c}{2}w+\frac{\beta}{2} \\ -arw-\frac{b}{2}r-\frac{c}{2}w+\frac{\beta}{2} & ar^2+cr+\gamma \end{bmatrix},$$

depends on six arbitrary parameters, and a qLN formulation exists provided that the mixed derivatives of $\nabla k(r,w)=-2A(r,w)M(r,w)$ are equal for some nonzero values of the parameters $a,b,c,\alpha,\beta,\gamma$. We thus have the following lemma:

Lemma 8.1: A given force $M(r,w)=(M_1(r,w),M_2(r,w))^t$ *admits a qLN formulation* $M(r,w)=-\frac{1}{2}A^{-1}(r,w)\nabla k(r,w)$ *if and only if there is a nontrivial solution A, with* det(*A*)≠0, *of the equation*

$$0=\frac{\partial}{\partial w}\left((aw^2+bw+\alpha)M_1+\left(-arw-\frac{b}{2}r-\frac{c}{2}w+\frac{\beta}{2}\right)M_2\right)$$
$$-\frac{\partial}{\partial r}\left(\left(-arw-\frac{b}{2}r-\frac{c}{2}w+\frac{\beta}{2}\right)M_1+(ar^2+cr+\gamma)M_2\right). \tag{8.1}$$

Lemma 8.2 (Criterion of integrability, $n=2$*): Equation (8.1) has a two-parameter family of solutions for* $A(r,w)$ *if and only if* $\ddot{q}=M(q)$ *admits two functionally independent integrals of motion E and F quadratic in velocities.*

Proof: If such $E=\dot{q}^tA\dot{q}+k$ and $F=\dot{q}^tB\dot{q}+l$ exist, then $\lambda E+\mu F=\dot{q}^t(\lambda A+\mu B)\dot{q}+(\lambda k+\mu l)$ is an integral of motion for all λ and μ, and thus $\lambda A+\mu B$ is a two-parameter solution of (8.1).

Conversely, if there is a two-parameter solution $D(\lambda,\mu)$ of (8.1), then there are linearly independent integrals E and F with $A=D(1,0)$ and $B=D(0,1)$. Q.E.D.

These two lemmas make it simple to test a given two-dimensional force for the existence of a qLN formulation, and to show integrability if a two-parameter family of solutions for A exists.

Example 8.3 (gH-H system): The generalized Hénon–Heiles (gH-H) system[9] is defined by the potential

$$V(q_1,q_2)=c_1q_1q_2^2-\frac{c_2q_1^3}{3}+\frac{c_0}{2q_2^2},\quad c_1,c_2\neq 0.$$

It is known to be integrable in three cases: the Korteweg–de Vries (KdV) case $6c_1+c_2=0$, the Sawada–Kotera (S-K) case $c_1+c_2=0$, and the Kaup–Kupershmidt (K-K) case $16c_1+c_2=0$. In the KdV case, and also in the S-K case, if $c_0=0$, the second integral of motion is quadratic in velocities, in the other cases it is quartic. The system appears naturally when integrating the equation $0=(\frac{1}{4}\partial^3+2c_1u\partial+c_1u_x)(\frac{1}{4}u_{xx}-\frac{1}{4}c_2u^2)$, which for the above cases corresponds to the stationary flow of the fifth-order KdV, S-K, and K-K soliton equations. This observation explained[10] the remarkable connection of the integrable cases of the gH-H system with soliton hierarchies.

We shall apply our criterion for existence of a qLN formulation to two Newton representations of the gH-H system; the original system in q variables

$$\ddot{q}_1 = -\frac{\partial V}{\partial q_1} = -c_1 q_2^2 + c_2 q_1^2, \qquad \ddot{q}_2 = -\frac{\partial V}{\partial q_2} = -2c_1 q_1 q_2 + \frac{c_0}{q_2^3}, \tag{8.2}$$

and another system in r variables

$$\ddot{r}_1 = r_2 + (c_1 + c_2) r_1^2, \qquad \ddot{r}_2 = c_3 - 10c_1 r_1 r_2 - 10c_1\left(c_1 + \frac{c_2}{3}\right) r_1^3, \tag{8.3}$$

which is equivalent[11] to the q system under the map $r_1 = q_1,\ r_2 = -c_1(q_1^2 + q_2^2),\ c = -4c_1[\frac{1}{2}(\dot{q}_1^2 + \dot{q}_2^2) + V(q_1, q_2)]$. The r system does not have any natural Lagrangian or Hamiltonian formulation and its integrability has previously been studied only through its equivalence with the original gH-H system. We will show here a more direct approach based on the qLN theory.

Beginning with the q system, we insert the right-hand side M from (8.2) into (8.1), identifying (q_1, q_2) with (r,w) as usual. Since the powers $r^i w^j$ are linearly independent, the coefficients at different powers must be individually zero. This gives, after some simplification, that $a = b = 0$, $\alpha = \gamma$ arbitrary, $(6c_1 + c_2)c = 0$, $(c_1 + c_2)\beta = 0$ and $c_0\beta = 0$. This means that we always have a solution $A = tI$ (of course, corresponding to the energy integral, since the system has a potential). Moreover, in two cases there exists a two-parameter solution; when $6c_1 + c_2 = 0$,

$$A = t\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + s\begin{bmatrix} 0 & -q_2 \\ -q_2 & 2q_1 \end{bmatrix},$$

and, when $c_1 + c_2 = c_0 = 0$,

$$A = t\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + s\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

So in this way we have recovered the KdV and S-K $(c_0 = 0)$ cases, while the K-K and S-K $(c_0 \neq 0)$ cases, with a quartic extra integral, fall outside of the qLN theory.

Performing the same procedure for the r system (8.3), we find that $a = 0$, $4c_1\gamma + b = 0$, $(31c_1 + c_2)c = (3c_1^2 + c_1 c_2)c = 0$, $2\alpha - 3c_3 c = 0$, and $(6c_1 + c_2)\beta = 0$. Since we have excluded the trivial case $c_1 = 0$, it follows that $c = \alpha = 0$, so that the solution is

$$A = t\begin{bmatrix} -2r_2 & r_1 \\ r_1 & 1/2c_1 \end{bmatrix}$$

except for the KdV case $6c_1 + c_2 = 0$ which admits a two-parameter solution

$$A = t\begin{bmatrix} -2r_2 & r_1 \\ r_1 & \dfrac{1}{c_1} \end{bmatrix} + s\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

This agrees with the known fact[11] that for the r system the second integral is quartic in velocities in the S-K and K-K cases and thus cannot be found by this method.

Suppose, however, that we had found the second integral in these cases by some other method. Then we would still have use for the qLN theory in proving the system's integrability, since the r parametrization admits the nonstandard Hamiltonian formulation (6.2) with $\lambda=1$ and c_3 playing the role of the fifth variable d. (This coincides with the Hamiltonian formulation that was found in Ref. 11 by transferring the standard Hamiltonian formulation from the q parametrization, except for naming the momenta in reverse order; here $p_i=\dot{r}_i$, while in that paper $s_1=\dot{r}_2$, $s_2=\dot{r}_1$.)

For example, in the S-K case ($c_1=-c_2=\frac{1}{2}$) we have

$$\begin{bmatrix} \dot{r}_1 \\ \dot{r}_2 \\ \dot{p}_1 \\ \dot{p}_2 \\ \dot{c}_3 \end{bmatrix} = \left[\begin{array}{cccc|c} 0 & 0 & -1/2 & -r_1/2 & p_1 \\ & 0 & -r_1/2 & r_2 & p_2 \\ & * & 0 & -p_1/2 & r_2 \\ & * & * & 0 & c_3-5r_1r_2-(5/3)r_1^3 \\ \hline * & * & * & * & 0 \end{array}\right] \nabla_{\mathcal{M}} c_3,$$

with the commuting integrals of motion $E=-2r_2p_1^2+2r_1p_1p_2+p_2^2+4r_1r_2^2+\frac{2}{3}r_1^3(r_1^2+5r_2)+c_3(-r_1^2-2r_2)$, which is a Casimir, the Hamiltonian c_3, and $F=\frac{3}{2}p_1^4-6r_1r_2p_1^2+(3r_1^2-r_2)p_1p_2+r_1p_2^2+\frac{7}{2}r_1^2r_2^2+\frac{10}{3}r_1^4r_2+\frac{5}{6}r_1^6+\frac{1}{3}r_2^3+(-2r_1r_2-r_1^3+\frac{3}{2}p_1^2)c_3$, which is quartic in momenta.

In order to give an impression of the wealth of different types of nonpotential Newton forces belonging to our theory, we will now examine solutions of the fundamental equation (4.5) for some specified pairs of matrices A and B. Any such solution corresponds to an integrable qLN system, and once one solution has been found, a whole family of solutions can be constructed using the recursion theorem 4.3.

Example 8.4 (One-dimensional complex motion): If we take A and B as

$$A=\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

then the fundamental equation reduces to the Laplace equation $K_{rr}+K_{ww}=0$. Given a solution $K(r,w)$, i.e., a harmonic function, we have $k=K\det(A)=-K$, so k is also harmonic. We find the corresponding l from the relation

$$\nabla l = BA^{-1}\nabla k = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \nabla k,$$

which is nothing but the Cauchy–Riemann equations for k and l, so l is the harmonic conjugate of k. The corresponding qLN system $\ddot{r}=-k_r/2$, $\ddot{w}=k_w/2$ can be integrated by introducing the complex variable $z=r+iw$ and the complex integral of motion $\mathcal{E}=E+iF=(\dot{r}^2-\dot{w}^2+k)+i(2\dot{r}\dot{w}+l)=(\dot{r}+i\dot{w})^2+(k+il)=\dot{z}^2+f(z)$, where $f(z)=k(z)+il(z)$ is analytic. We can now determine z, and thus r and w, from $\dot{z}=\pm\sqrt{\mathcal{E}-f(z)}$ by one complex quadrature.

Repeated application of the recursion formula (4.10) yields in this case the standard cycle of conjugate harmonic pairs $(k,l)\to(l,-k)\to(-k,-l)\to(-l,k)\to(k,l)$.

Example 8.5 (Fundamental equation separable in polar coordinates): Let

$$A=\begin{bmatrix} -2w & r \\ r & 0 \end{bmatrix}, \quad B=\begin{bmatrix} 0 & w \\ w & -2r \end{bmatrix}.$$

Then the fundamental equation becomes $0=2(r^2K_{rr}+2rwK_{rw}+w^2K_{ww})+9(rK_r+wK_w)+6K$, which in polar coordinates ($r=R\cos\phi$, $w=R\sin\phi$) transforms into $0=2R^2K_{RR}+9RK_R+6K$.

The general solution of this equation is $K(R,\phi)=f_0(\phi)R^{-2}+g_0(\phi)R^{-3/2}$, for some arbitrary functions f_0 and g_0. Changing back to r and w, we find that the general solution of the fundamental equation in this case is

$$K(r,w)=f_0\left(\arctan\frac{w}{r}\right)\frac{r^{-2}}{1+(w/r)^2}+g_0\left(\arctan\frac{w}{r}\right)\frac{r^{-3}}{(1+(w/r)^2)^{3/2}}=f\left(\frac{w}{r}\right)r^{-2}+g\left(\frac{w}{r}\right)r^{-3},$$

where f and g are arbitrary functions.

Finally, let us conclude with an example of a three-dimensional qLN system. A detailed treatment of higher-dimensional qLN systems is presented in a separate article.[3]

Example 8.6: The Newton system

$$\ddot{r}_1=-10r_1^2+4r_2,$$

$$\ddot{r}_2=-16r_1r_2+10r_1^3+4r_3,$$

$$\ddot{r}_3=-20r_1r_3-8r_2^2+30r_1^2r_2-15r_1^4+d$$

was found in Ref. 12 as a parametrization of the seventh-order stationary KdV flow. It has three integrals of motion

$$E_1=\dot{r}_1\dot{r}_3+\frac{\dot{r}_2^2}{2}+10r_1^2r_3-4r_2r_3+8r_1r_2^2-10r_1^3r_2+3r_1^5-dr_1,$$

$$E_2=r_3\dot{r}_1^2-r_1\dot{r}_2^2+r_2\dot{r}_1\dot{r}_2-\dot{r}_2\dot{r}_3-r_1\dot{r}_1\dot{r}_3+4r_1^2r_2^2+5r_1^4r_2-\frac{5}{2}r_1^6$$
$$-4r_2^3+2r_3^2-12r_1r_2r_3+\frac{dr_1^2}{2}+dr_2,$$

$$E_3=\frac{1}{8}(r_2^2\dot{r}_1^2+r_1^2\dot{r}_2^2+\dot{r}_3^2-(2r_1r_2+4r_3)\dot{r}_1\dot{r}_2+2r_1\dot{r}_2\dot{r}_3+2r_2\dot{r}_1\dot{r}_3)+r_1r_2^3-3r_1^3r_2^2$$
$$+\frac{5}{4}r_1^5r_2+2r_1r_3^2+\frac{5}{4}r_1^4r_3-r_1^2r_2r_3+r_2^2r_3-\frac{d}{4}(r_1r_2-r_3),$$

which all are quadratic in velocities. This means that the system is generated by any of them through the quasi-Lagrangian equations. From the velocity-dependent parts we find

$$\begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix},\quad \begin{bmatrix}2r_3&r_2&-r_1\\r_2&-2r_1&-1\\-r_1&-1&0\end{bmatrix},\quad \begin{bmatrix}r_2^2&-r_1r_2-2r_3&r_2\\-r_1r_2-2r_3&r_1^2&r_1\\r_2&r_1&1\end{bmatrix}$$

as examples of 3×3-matrices satisfying the cyclic conditions (2.3).

IX. CONCLUSIONS

In this article we have developed a new theory—the theory of quasi-Lagrangian Newton equations. It was originally inspired by interesting properties of the second stationary flow of the Harry Dym hierarchy, which led us to a broad theory which encompasses the classical separability theory but goes far beyond the classical results—the classical Bertrand–Darboux theory of separability for two-dimensional potential forces depends on one essential free parameter while our theory depends on five parameters.

The main part of this work has been focused on two-dimensional qLN systems which admit two integrals of motion E and F quadratic in velocities. These systems have only a nonstandard Hamiltonian formulation and are completely integrable by embedding into five-dimensional Liouville integrable systems. All such qLN systems are characterized by a single PDE called here the fundamental equation. We have shown that there is a one-to-one correspondence between fundamental equations and linear pencils $\lambda A + \mu B$ of matrices A and B. These linear pencils have been classified in Sec. V. In Sec. VII the class of driven systems has been studied in detail and new types of separation variables (non-confocal conics) have been found. We have also shown that any given force can be effectively tested for the existence of qLN formulation, which can further be used for unveiling its complete integrability and for solving the corresponding Newton equation. We have illustrated this by several examples including the generalized Hénon-Heiles system (Sec. VIII).

There are several natural directions of development of the theory of qLN systems. The n-dimensional versions of our main theorems on fundamental equation and on complete integrability have already been formulated and proved in Ref. 3.

The great wealth of different types of integrable Newton equations contained in the fundamental equation remains to be studied. Here we have only discussed two special cases: separable systems and driven systems. However, one of the most challenging questions yet to be answered is the existence of separation variables for the fundamental equation in its most general form. It can lead to new and interesting connections with the classical theory of separability of the Hamilton–Jacobi equation and of linear PDEs.

[1] S. Rauch-Wojciechowski, K. Marciniak, and M. Blaszak, ''Two Newton decompositions of stationary flows of KdV and Harry Dym hierarchies,'' Physica A **233**, 307–330 (1996).
[2] E. T. Whittaker, *A Treatise on the Analytic Dynamics of Particles and Rigid Bodies* (Cambridge U.P., Cambridge, 1959), Sec. 152.
[3] H. Lundmark, in *Integrable Nonconservative Newton Systems with Quadratic Integrals of Motion*, Linköping Studies in Science and Technology. Theses No. 756, Linköping, 1999.
[4] R. Garnier, Circolo Mat. Palermo **43**, 155 (1919).
[5] S. Wojciechowski, ''Integrable one-particle potentials related to the Neumann system and the Jacobi problem of geodesic motion on an ellipsoid,'' Phys. Lett. A **107**, 106–111 (1985).
[6] K. Marciniak and S. Rauch-Wojciechowski, ''Two families of nonstandard Poisson structures for Newton equations,'' J. Math. Phys. **39**, 5292–5306 (1998).
[7] S. Rauch-Wojciechowski, ''A bi-Hamiltonian formulation for separable potentials and its application to the Kepler problem and the Euler problem of two centers of gravitation,'' Phys. Lett. A **160**, 149–154 (1991).
[8] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, New York, 1989).
[9] M. Antonowicz and S. Rauch-Wojciechowski, ''Bi-Hamiltonian formulation of the Hénon-Heiles system and its multi-dimensional extensions,'' Phys. Lett. A **163**, 167–172 (1992).
[10] A. P. Fordy, ''The Hénon-Heiles system revisited,'' Physica D **52**, 204–210 (1991).
[11] M. Blaszak and S. Rauch-Wojciechowski, ''A generalized Hénon-Heiles system and related integrable Newton equations,'' J. Math. Phys. **35**, 1693–1709 (1994).
[12] S. Rauch-Wojciechowski, ''Newton representation for stationary flows of the KdV hierarchy,'' Phys. Lett. A **170**, 91–94 (1992).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Relativistic gas: Moment equations and maximum wave velocity

Guy Boillat[a] and Tommaso Ruggeri[b]
Department of Mathematics and Research Center of Applied Mathematics-(C.I.R.A.M.), University of Bologna, Via Saragozza 8, 40123 Bologna, Italy

(Received 16 August 1999; accepted for publication 7 September 1999)

For a rarefied relativistic gas we consider the N-moment equations associated with the relativistic Boltzmann–Chernikov equation and we require the compatibility with the entropy principle thus obtaining a closed symmetric hyperbolic system. This interesting form permits one to deduce a lower and an upper bound for the maximum velocity of a wave propagating in a monoatomic or a degenerate gas of fermions or bosons and to prove that when this number N increases this velocity tends to the speed of light. © *1999 American Institute of Physics.* [S0022-2488(99)02012-5]

I. INTRODUCTION

The thermo-mechanical evolution of a gas must be described in the relativistic context if the function $\gamma = mc^2/kT$ of the absolute temperature T is sufficiently small: $\gamma \ll 1$ (m, c, and k are, respectively, the atomic mass, the light velocity, and the Boltzmann constant). Relativistic fluid dynamics have possible applications in stellar physics and are appropriate for very hot gas, e.g., the white dwarfs, or for some degenerate gas in which the particle mass is very small, such as in the photon gas or in the neutrino gas.

As is well known in the case of rarefied gas we have two possible approaches; one in the context of phenomenological theory (continuum models), the second one using the relativistic kinetic theory. Ordinary relativistic thermodynamics was first developed by Eckart[1] and suffers from the paradox of infinite velocity due to the fact that the partial differential system is of parabolic type. Relativistists have shown a keen interest in the development of *extended thermodynamics*[2] that solves the paradox starting from the pioneering paper of Cattaneo.[3] The first approach of relativistic extended thermodynamics was given by Müller[4] and Israel[5] and was presented in a systematic manner by Liu, Müller, and Ruggeri.[6] This hyperbolic theory not only solves the paradox but also is in perfect agreement with the 14 moment equations deduced by Marle[7] from the Chernikov–Boltzmann kinetic equation.[8]

In Ref. 9 the closure procedure of extended thermodynamics[2] was applied to the N-moment system associated with both the classical and relativistic Boltzmann equation.

In particular it was proved that the compatibility with an entropy principle implies that the (truncated) distribution function f depends on a single variable χ that is a polynomial in the four-momentum with coefficients (*main field*) that are solutions of a symmetric hyperbolic system: the macroscopic field equations for the gas in nonequilibrium.

In wave problems the maximum wave speed plays a particular role (e.g., phase velocity in high frequencies, light scattering, shock waves, etc., see Ref. 2). Taking into account the special form of the symmetric hyperbolic system it was possible to deduce a lower bound estimate for the maximum velocity $\lambda_{\max}$ of a wave propagating in an equilibrium state. In the classical context this velocity increases without limit with N while in the relativistic case it was verified that the light speed c is an upper bound.[9] In Ref. 10 a lower bound was calculated also in the relativistic case

[a]Permanent address: Department of Applied Mathematics, University of Clermont, France; electronic mail: boillat@riemann.ing.unibo.it
[b]Electronic mail: ruggeri@lagrange.ing.unibo.it

for nondegenerate gas and it was proved that when the number of moments increases the maximum velocity does indeed tend to c. A review article on these subjects may be found in the recent survey of Müller.[11]

The aim of this paper is to extend this last result to a generic gas including also the interesting cases of degenerate Fermi–Dirac and Bose–Einstein gases.

The paper is organized in the following manner. In Sec. II we introduce the moment equations associated with the Chernikov–Boltzmann kinetic equation. In Sec. III we consider a finite number of moments and we close the system through the procedure of extended thermodynamics that is equivalent to the one obtained by the *maximum entropy principle*. In Sec. IV we deduce for all kinds of gases a lower and upper bound for the maximum wave speed, proving that when the number of moments tends to infinity this velocity tends to the light velocity.

II. BOLTZMANN–CHERNIKOV EQUATION AND MOMENTS

In the relativistic kinetic theory of a rarefied gas the phase density $f(x^\alpha,p^\alpha)$ $(\alpha=0,1,2,3)$ satisfies the Boltzmann–Chernikov equation[8,12]

$$p^\alpha\partial_\alpha f=Q,\quad \partial_\alpha=\partial/\partial x^\alpha, \tag{1}$$

in which x^α and p^α are the space–time coordinates and the four-momentum of an atom, respectively. We have $p_\alpha p^\alpha=(p^0)^2-\mathbf{p}^2=m^2c^2$, $\mathbf{p}^2=(p^1)^2+(p^2)^2+(p^3)^2$. The right-hand side of (1) is due to collision between the atoms.

Upon multiplication by $p^A(A=0,1,2,...)$ and integration the Boltzmann–Chernikov equation provides an infinite system of balance equations

$$\partial_\alpha F^{\alpha A}=g^A,\quad A=0,1,2,..., \tag{2}$$

for the moments $F^{\alpha A}$ and productions g^A given by

$$F^{\alpha A}(x^\beta)=\int p^\alpha p^A f\,dP,\quad g^A(x^\beta)=\int Qp^A f\,dP, \tag{3}$$

where A is a multi-index

$$p^A=\begin{cases}1\\ p^{\alpha_1}p^{\alpha_2}\cdots p^{\alpha_A},\end{cases}\quad F^{\alpha A}=\begin{cases}F^\alpha\\ F^{\alpha\alpha_1\ldots\alpha_A},\end{cases}\quad g^A=\begin{cases}0 & \text{for } A=0\\ g^{\alpha_1\cdots\alpha_A} & \text{for } A\geqslant 1\end{cases}$$

and $0\leqslant\alpha_1\leqslant\alpha_2\leqslant\ldots\leqslant\alpha_A\leqslant 3$.

The first five equations of (2) are the conservation laws of the number of particles and energy–momentum. According by the first five productions vanish: $g^A=0$ for $A=0,1$. The volume element of momentum space is given by $dP=d^3p/p^0=dp^1dp^2dp^3/p^0$ and the integrals—supposed convergent—are taken over the whole of $\mathbf{p}$ space.

On the other hand if we consider as momentum

$$h^\alpha=k\int p^\alpha\{(s^2-1+\ln f)f+s(1-sf)\ln(1-sf)\}dP, \tag{4}$$

where $s=0,+1,-1$ corresponds, respectively, to the nondegenerate gas, the Fermi gas, and the Bose gas, we obtain from (1) the supplementary inequality (H-Theorem)

$$\partial_\alpha h^\alpha=\Sigma\leqslant 0. \tag{5}$$

The previous expression represents the balance of entropy when we identify $-h^\alpha$ and $-\Sigma$ with the entropy four-vector and the entropy production, respectively.

III. FINITE SYSTEM OF MOMENT EQUATIONS AND CLOSURE

We consider now a finite number of moment equations (2) with the tensorial index $A=0,...,n$. In this case the main problem is the so-called *closure problem*. In fact the number of moments is larger than the number of moment equations and therefore constitutive equations are needed to determine the system. In Ref. 9 we proposed the closure procedure of extended thermodynamics[2] requiring the compatibility with an entropy principle that plays an important role in selecting the physical class of constitutive equation in thermomechanics. This consists in assuming that any differentiable solution of the finite moment equations (2) with $A=0,...,n$ is also a solution of the supplementary entropy law (5) which appears now as a constraint.

We recall that for a general quasilinear system of balance laws of type (2) the compatibility with the entropy law (5) implies the existence of a privileged field (*main field*) $u'_A(x^\alpha)$ and a four-vector $h'^\alpha(u'_A)$ (*generators*) such that[13,14]

$$F^{\alpha A}=\frac{\partial h'^\alpha}{\partial u'_A},\quad h^\alpha=\sum_{A=0}^{n} u'_A\frac{\partial h'^\alpha}{\partial u'_A}-h'^\alpha,\quad \Sigma=\sum_{A=0}^{n} u'_A g^A,\quad A=0,1,...,n. \tag{6}$$

In the present case from $(3)_1$ and $(6)_1$ we obtain

$$dh'^\alpha=\sum_{A=0}^{n} F^{\alpha A}du'_A=\sum_{A=0}^{n}\int p^\alpha p^A\, du'_A\, f\, dP. \tag{7}$$

It follows that f must be a function of the single variable

$$\chi=\sum_{A=0}^{n} u'_A p^A, \tag{8}$$

while

$$h'^\alpha=\int F(\chi)p^\alpha\, dP,\quad \frac{dF(\chi)}{d\chi}=f(\chi). \tag{9}$$

From $(6)_2$ we obtain the admissible entropy four-vector

$$h^\alpha=\int (\chi f(\chi)-F(\chi))p^\alpha\, dP. \tag{10}$$

The remaining condition is the so-called residual inequality that requires that the admissible productions $g^A(u'_B)$ are such that

$$\Sigma=\sum_{A=0}^{n} u'_A g^A(u'_B)\leqslant 0. \tag{11}$$

Therefore we have verified the following: *The compatibility of the truncated moment system with the entropy principle requires that the truncated distribution function f depend on a single variable χ given by (8), while the entropy four-vector and the productions satisfy (10) and the residual inequality (11).*

In Ref. 9 (see also Ref. 2, p. 218 sqq) it was shown that this closure procedure is equivalent to the one obtained by the so-called *maximum entropy principle* by which the truncated distribution function is chosen such that the entropy density thought of as a generic functional of $f: h^0=\int\Psi(f)d^3p$ is a maximum under the constraint that the moments $\int f p^A d^3p$ are assigned functions F^{0A}. In fact it is easy to verify that $\Psi(f)$ must be equal to $\chi f-F$ and the Lagrange multipliers in the maximization procedure are nothing but the main field variables u'_A. The first author that applied the idea of maximization of entropy in nonequilibrium thermodynamics was

Dreyer,[15] who started from the observation of Kogan[16] that the Grad's 13-moment density maximizes the entropy. The procedure of maximizing entropy was introduced in information theory and physics by Jaynes.[17,18]

By (6) the finite system of moments (2) becomes

$$\sum_{B=0}^{n} H^{\alpha AB}(u'_C)\partial_\alpha u'_B = g^A(u'_C), \quad A=0,1,...,n \tag{12}$$

with the symmetric matrices

$$H^{\alpha AB} = \int \frac{df}{d\chi} p^\alpha p^A p^B \, dP, \tag{13}$$

and $H^{\alpha AB}\xi_\alpha$ is positive definite for any timelike vector $(\xi_\alpha \xi^\alpha > 0)$ if $df/d\chi > 0$. This implies that the system of truncated moments (12) is [in the main field variables $u'_A(x^\alpha)$] a symmetric hyperbolic system (in the sense of Friedrichs) for which the Cauchy problem is well posed.[19]

The wave surface $\phi(x^\alpha)=0$ is solution of the characteristic equation

$$\det(H^{\alpha AB}\partial_\alpha \phi) = 0. \tag{14}$$

As a consequence, the four-gradient $\partial_\alpha \phi$ cannot be timelike and therefore the velocities of wave propagation cannot exceed the velocity of light, i.e., $\lambda_{\max} \leqslant c$. When the number of equations increases it has already been shown that the maximum wave velocity cannot decrease.[20] The question is now: does this velocity tends to c when n tends to infinity? This is the subject of the next paragraph.

We observe that since all of the indices in the main field $u'_{\alpha_1 \cdots \alpha_k}$ may assume the values 0, 1, 2, 3 and therefore,—following Refs. 9 and 10—the components $u'_{\alpha_1 \cdots \alpha_k}$ may be mapped into the variables u'_{pqrs} with $p+q+r+s=k(k=1,...,n)$, where p,q,r,s are the numbers of indices among $\alpha_1 \ldots \alpha_k$ which are equal to 0, 1, 2 or 3 respectively. Because p^0 is equal to $\sqrt{m^2c^2+\mathbf{p}^2}$, the first index p can take only two values: 0 and 1. Now, since the number of elements with the sum of three integers $q+r+s=k$ with $k=0,...,m$ is $(m+1)(m+2)(m+3)/6$, the number $N(n)$ of independent components of u'_{pqrs}-''and equations,''-up to order n is $N(n)=(n+1)(n+2)(n+3)/6 + n(n+1)(n+2)/6$. Therefore the maximum tensorial index n that appears in the moment system is related to the number N of independent moments and field equations through[10]

$$N(n) = \tfrac{1}{6}(n+1)(n+2)(2n+3), \quad N(n+1) = N(n) + (n+2)^2.$$

Assuming that (10) is of the form (4) follows

$$k\{f(s^2-1+\ln f) + s(1-sf)\ln(1-sf)\} = \chi f - F \tag{15}$$

from which (by differentiation)

$$f(\chi) = \frac{1}{e^{-\chi/k}+s}. \tag{16}$$

Summarizing we have: *The admissible truncated distribution function $f(\chi)$ such that the finite moment system (2) is compatible with the entropy principle (5) with entropy four-vector given by (4) is of the form (16) with χ given by (8) and u'_A solution of the symmetric hyperbolic system (12).*

IV. PROPAGATION IN AN EQUILIBRIUM STATE AND MAXIMUM WAVE VELOCITY

A thermodynamic equilibrium state is defined as the state for which the productions g^A vanish and the entropy production $-\Sigma$ reaches its minimum value, i.e., zero. The main field variables are privileged also in this context because it is possible to prove that all are zero[2,21] with exception of the first five:[14]

$$u' = \frac{G}{T}, \quad u'_\alpha = -\frac{u_\alpha}{T}, \tag{17}$$

where G and u_α are, respectively, the chemical potential and the four velocity. Therefore for any truncation index n the χ given by (8) reduces in the equilibrium state to $\chi/k = (G - u_\alpha p^\alpha)/(kT)$ and in the rest frame in which $u^i = 0$, $u^0 = c$, we have

$$\frac{\chi}{k} = -a - \gamma\sqrt{1 + \frac{\mathbf{p}^2}{m^2c^2}}, \tag{18}$$

with $a = -G/KT$, $\gamma = mc^2/kT$. In this case the distribution function (16) becomes the well-known Jüttner equilibrium distribution.[22] We recall that for a fermi gas a can assume all real values while for bosons $a + \gamma > 0$.

According to (14) the wave velocity λ in the direction of the normal $\mathbf{n}$ to the wave front is an eigenvalue of

$$\det(H^{iAB}n_i - \lambda H^{0AB}) = 0. \tag{19}$$

The matrix in (19) is negative semidefinite for the maximum eigenvalue, that is if we take $\mathbf{n} \equiv (1,0,0)$ the components of the matrix

$$H^{1AB} - \lambda_{\max} H^{0AB} = \int \frac{df}{d\chi} p^A p^B (p^1 - \lambda p^0)\, dP,$$

satisfy the inequalities $a_{ii}a_{jj} - a_{ij}^2 \geqslant 0$ so that choosing $p^A = (p^1)^n$, $p^B = (p^1)^{n-1}$, we have

$$\frac{\lambda_{\max}^2}{c^2} \int \frac{df}{d\chi} (p^1)^{2n} d^3p \int \frac{df}{d\chi} (p^1)^{2(n-1)} d^3p \geqslant \left(\int \frac{df}{d\chi} (p^1)^{2n} \frac{d^3p}{p^0} \right)^2 \tag{20}$$

since the integrals of odd functions vanish.

Introducing as in Ref. 10 the spherical coordinates $p^1 = mcr \sin\theta \cos\varphi$, $p^2 = mcr \sin\theta \sin\varphi$, $p^3 = mcr \cos\theta$ the previous inequality with (16) and (18) yields after some straightforward calculations

$$\frac{\lambda_{\max}^2}{c^2} \geqslant \frac{2n-1}{2n+1} \frac{J_{n+1}^2}{I_n I_{n+1}}, \tag{21}$$

where

$$I_n = \int_0^\infty \phi(r) r^{2n}\, dr, \quad J_n = \int_0^\infty \frac{\phi(r) r^{2n}}{\sqrt{1+r^2}}\, dr,$$

$$\phi(r) = \frac{e^{\psi(r)}}{(1 + s e^{\psi(r)})^2}, \quad \psi(r) = \frac{\chi}{k} = -a - \gamma\sqrt{1+r^2}.$$

Therefore we have: *For any type of gas including the degenerate gas of fermions and bosons the largest wave velocity satisfies the lower and upper bounds*:

$$\frac{2n-1}{2n+1}\frac{J_{n+1}^2}{I_n I_{n+1}} \leqslant \frac{\lambda_{\max}^2}{c^2} \leqslant 1. \tag{22}$$

Now our aim is to prove that when $n\to\infty$, $\lambda_{\max}\to c$. Integrating by parts

$$I_n = \int_0^\infty \phi(r)\frac{d(r^{2n+1})}{2n+1} = \frac{\gamma}{2n+1}\int_0^\infty \phi(r)\frac{1-se^{\psi}}{1+se^{\psi}}\frac{r^{2n+2}}{\sqrt{1+r^2}}dr,$$

and taking into account that

$$\frac{1-se^{\psi}}{1+se^{\psi}} \leqslant C, \quad C = \begin{cases} 1 & \text{for } s=0,1 \\ (1+e^{-(a+\gamma)})/(1-e^{-(a+\gamma)}) & \text{for } s=-1, \end{cases}$$

it follows

$$I_n \leqslant \frac{\gamma C}{2n+1}\int_0^\infty \frac{\phi r^{2n+2}}{\sqrt{1+r^2}}dr \Rightarrow J_{n+1} \geqslant \frac{2n+1}{\gamma C} I_n \tag{23}$$

and (21) becomes

$$\frac{\lambda_{\max}^2}{c^2} \geqslant \frac{2n-1}{2n+1}\left(\frac{2n+1}{\gamma C}\right)^2 \frac{I_n}{I_{n+1}}. \tag{24}$$

On the other hand for any constant μ,

$$0 \leqslant \int_0^\infty \phi\left(r^n\mu + \frac{r^{n+2}}{\sqrt{1+r^2}}\right)^2 dr \leqslant \mu^2 I_n + 2\mu J_{n+1} + I_{n+1},$$

which shows that

$$J_{n+1}^2 \leqslant I_n I_{n+1}$$

and by (23)

$$\left(\frac{2n+1}{\gamma C}\right)^2 \frac{I_n}{I_{n+1}} \leqslant 1. \tag{25}$$

Let M_s be the maximum of $\phi(r)$:

$$M_0 = e^{-(a+\gamma)}, \quad M_1 = \frac{1}{4}, \quad M_{-1} = \frac{e^{-(a+\gamma)}}{(1-e^{-(a+\gamma)})^2}$$

and consider now the series of functions

$$u_n(r) = \frac{1}{(n!)^2}\left(\frac{r\gamma C}{2}\right)^{2n}\phi(r) \leqslant \frac{M_s}{(n!)^2}\left(\frac{R\gamma C}{2}\right)^{2n},$$

which converges uniformly for all finite values of $r\leqslant R$ to the function $\phi(r)I_0(\gamma C r)$ where $I_0(x)$ is the Bessel function.

The integral

$$U_n(R) = \int_0^R u_n\, dr = \frac{1}{(n!)^2}\left(\frac{\gamma C}{2}\right)^{2n}\mathcal{I}_n(R), \quad \mathcal{I}_n(R) = \int_0^R \phi(r) r^{2n} dr, \quad I_n = \mathcal{I}_n(\infty)$$

exists and this series is convergent. We can choose R and n large enough so that the ratio

$$\frac{U_n(R)}{U_{n+1}(R)}=\left(\frac{2n+2}{2n+1}\right)^2\left(\frac{2n+1}{\gamma C}\right)^2\frac{\mathcal{I}_n(R)}{\mathcal{I}_{n+1}(R)}$$

is as close as we like to the left member of (25). Because of the convergence of the series it always exists an n as large as we wish such that this ratio is larger than or equal to 1. Therefore it is close to 1 and so, by (24), is $\lambda_{\max}/c$ which, on the other hand, is a nondecreasing function of n[20] and cannot exceed 1 and then must tend to 1 as n tends to infinity.

In conclusion *when the number of moments tends to infinity the maximum wave velocity tends for all kind of gases to the light velocity*.

These results are in perfect agreement with a recent numerical simulation given in Ref. 23 in which the maximum velocity was evaluated for increasing values of the truncation index n for different values of γ and a.

ACKNOWLEDGMENTS

This paper was supported by INdAM-GNFM (Gruppo Nazionale per la Fisica Matematica of the Instituto Nazionale di Alta Matematica), Ministero della Università e della Ricerca Scientifica e Technologica (MURST-fondi 40% and 60%), and by Istituto Nazionale di Fisica Nucleare (INFN).

[1] C. Eckart, ''The thermodynamics of irreversible processes. III. Relativistic theory of the simple fluid,'' Phys. Rev. **58**, 919 (1940).
[2] I. Müller and T. Ruggeri, *Rational Extended Thermodynamics*, Springer Tracts in Natural Philosophy, Vol. 37, 2nd ed. (Springer, New York, 1998).
[3] C. Cattaneo, ''Sulla conduzione del calore,'' Atti Sem. Mat. Fis. Univ. Modena **3**, 83 (1948–1949).
[4] I. Müller, ''Zur Ausbreitungsgeschwindigkeit von Störungen in kontinuierlichen Medien,'' dissertation Aachen, 1966.
[5] W. Israel, ''Nonstationary irreversible thermodynamics: A causal relativistic theory,'' Ann. Phys. (N.Y.) **100**, 341 (1976).
[6] I-Shih Lin, I. Müller, and T. Ruggeri, ''Relativistic thermodynamics of gases,'' Ann. Phys. (N.Y.) **169**, 191 (1986).
[7] C. Marle, ''Sur l'établissement des équations de l'hydrodynamique des fluides relativistes dissipatives,'' Ann. Inst. Henri Poincare **10**, 67 (1969).
[8] N. A. Chernikov, ''The relativistic gas in the gravitational field,'' Acta Phys. Pol. **23**, 629 (1963); ''Equilibrium distribution of the relativistic gas,'' **26**, 379 (1964); ''Microscopic foundation of relativistic hydrodynamics,'' **27**, 465 (1965).
[9] G. Boillat and T. Ruggeri, ''Moment equations in the kinetic theory of gases and wave velocities,'' Continuum Mech. Thermodyn. **9**, 205 (1997).
[10] G. Boillat and T. Ruggeri, ''Maximum wave velocity in the moment system of a relativistic gas,'' Continuum Mech. Thermodyn. **11**, 107 (1999).
[11] I. Müller, *Speeds of Propagation in Classical and Relativistic Extended Thermodynamics*, Living Reviews in Relativity (Max-Planck-Institute for Gravitational Physics, Albert Einstein Institute, Potsdam, 1999), http://www.livingreviews.org/Articles/Volume2/1999-1mueller/
[12] A. Lichnerowicz and R. Marrot, ''Propriétés statistiques des ensembles de particules en relativité restreinte,'' C. R. Acad. Sci. **210**, 759 (1940); ''Sur l'équation intégrodifférentielle de Boltzmann,'' **211**, 531 (1940); ''Le H-théorême et la notion de système isolé,'' **212**, 1074 (1941).
[13] G. Boillat, ''Sur l'existence et la recherche d'équations de conservation supplémentaires pour les systèmes hyperboliques,'' C. R. Acad. Sci. Paris **278A**, 909 (1974); *Nonlinear Fields and Waves in CIME Course Recent Mathematical Methods in Nonlinear Wave Propagation*, edited by T. Ruggeri, Lecture Notes in Mathematics Vol. 1640 (Springer, Berlin, 1996).
[14] T. Ruggeri and A. Strumia, ''Main field and convex covariant density for quasi-linear hyperbolic systems. Relativistic fluid dynamics,'' Ann. Inst. Henri Poincare, Sect. A **34**, 65 (1981).
[15] W. Dreyer, ''Maximisation of the Entropy in Non-Equilibrium,'' J. Phys. A **20**, 6505 (1987).
[16] M. N. Kogan, *Proceedings of the Fifth Symposium on Rarefied Gas Dynamics* (Academic, New York, 1967), Vol. 1, suppl. 4, pp. 359–368.
[17] E. T. Jaynes, ''Information theory and statistical mechanics I, II,'' Phys. Rev. **106**, 620 (1957); **108**, 171 (1957).
[18] J. N. Kapur, *Maximum-Entropy Models in Science and Engineering* (Wiley, New York, 1989); H. Gzyl, *The Method of Maximum Entropy* (World Scientific, Singapore, 1995).
[19] A. E. Fisher and J. E. Marsden, ''The Einstein evolution equations as a first order quasilinear symmetric hyperbolic system,'' Commun. Math. Phys. **28**, 1 (1972).
[20] G. Boillat and T. Ruggeri, ''Hyperbolic principal subsystems: entropy convexity and subcharacteristic conditions,'' Arch. Ration. Mech. Anal. **137**, 305 (1997).

[21] G. Boillat and T. Ruggeri, ''On the shock structure problem for hyperbolic system of balance laws and convex entropy,'' Continuum Mech. Thermodyn. **10**, 285 (1998).
[22] F. Jüttner, ''Das Maxwell'sche Gesetz der Geschwindigkeitsverteilung in der Relativtheorie,'' Ann. Phys. (Leipzig) **34**, 856 (1911); ''Die relativistische Quantentheorie des idealen Gases,'' Z. Phys. **47**, 542 (1928).
[23] F. Brini and T. Ruggeri, ''Maximum velocity for wave propagation in a relativistic gas,'' Continuum Mech. Thermodyn. **11**, 331 (1999).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Integral representations of thermodynamic 1PI Green's functions in the world-line formalism

Haru-Tada Sato[a)]
Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, D-69120 Heidelberg, Germany

(Received 5 March 1999; accepted for publication 18 March 1999)

The issue discussed is a thermodynamic version of the Bern–Kosower master amplitude formula, which contains all necessary one-loop Feynman diagrams. It is demonstrated how the master amplitude at finite values of temperature and chemical potential can be formulated within the framework of the world-line formalism. In particular we present an elegant method of how to introduce a chemical potential for a loop in the master formula. Various useful integral formulas for the master amplitude are then obtained. The nonanalytic property of the master formula is also derived in the zero temperature limit with the value of chemical potential kept finite. © *1999 American Institute of Physics.* [S0022-2488(99)00812-9]

I. INTRODUCTION

Based on the field theory limit of string theory (with infinite string tension limit), a very elegant method was invented several years ago: the Bern–Kosower (BK) rules to obtain one-loop gluon scattering amplitudes in a compact form[1] (see also Ref. 2). For example, five gluon scatterings are efficiently calculated by using these rules,[3] and various field theory limits have been studied along the line of the BK formalism: perturbative gravity,[4] super Yang–Mills theory,[5] bosonic string theory approach,[6] and multi-loop generalizations.[7–9]

The most important and conspicuous point in the BK formalism is that all Feynman diagrams are included in a single master integrand. In this formalism, we hence do not compute the loop integrals and the Dirac traces of respective Feynman diagrams. These facts can also be achieved by another approach, called the world-line formalism, which reformulates Feynman (field theory) amplitudes similar to string theory amplitudes; i.e., the field theory amplitudes can generally be obtained as a path integral average of vertex operators.[10–17]

In fact, there have been established many examples: ϕ^3 theory,[10,11] QED,[12,13] axial vector and pseudo-scalar couplings,[14,15] Yang–Mills theory[16,17] (see also Refs. 18 and 19), and more references can be found in Ref. 20. Both world-line and string theory methods help each other, and are useful to get an insight for solving mutually related problems each in its own way; for example, this viewpoint has been very useful for specifying pinching limits and corners of string moduli in the multi-loop analysis.[8] It is very interesting that these two methods, which differ entirely from the conventional Feynman diagram technique, improve the computational efficiency for obtaining Feynman amplitudes.

However, compared to these developments, their thermodynamic versions have not been studied very much[21–25] from the viewpoint of general formulation for constructing a master amplitude formula. In particular, there has been no general and convenient method of how to introduce a chemical potential for a loop in the world-line formalism. Before stepping in an unexplored calculation by using the formalism, it is important to establish a definite and universal foundation in the first place. In this paper, therefore, as a basic step on the thermodynamic generalization of world-line field theory, we present a universal and fundamental prescription of one-loop N-point amplitudes at finite values of temperature and chemical potential without help of

[a)]Electronic mail: sato@thphys.uni-heidelberg.de

any standard calculation. We shall study the amplitude of a particular form (the master formula), where the loop integration and the Matsubara summation are *a priori* finished and Feynman's parameter integrals are only left—laying emphasis on the point that we never mean the Feynman integrals as a single Feynman diagram, but as a sum of *all diagrams*. It is certainly nontrivial to introduce a chemical potential (as well as a temperature) with keeping this advantageous point pertaining to the master formula intact.

We address the following points in the matter of the thermodynamic generalization (at finite values of the temperature β^{-1} and the chemical potential μ for a loop). First, we present a formulation of the thermodynamic amplitudes along the same line as the nonthermodynamic world-line formalism. In particular, the way of introducing the chemical potential is a nontrivial problem. Since we do not introduce any idea of the continuous or discrete momentum integration/summation, we have to find an alternative to the shift of internal discrete momenta:

$$\omega_n \to \omega_n + i\mu. \tag{1.1}$$

This situation may be understood in the following way: In the standard method, the inverse temperature β is introduced by summing up all topological different S^1 paths along the zeroth component (imaginary time) direction. On the other hand, in the world-line formalism, this procedure modifies the path integral of a corresponding periodic world-line field $x^0(\sigma)$, $0\leqslant\sigma\leqslant 1$, into a summation of the path integrals with $x^0(\sigma)$ shifted by

$$x^0(\sigma) \to x^0(\sigma) + n\beta\sigma. \tag{1.2}$$

Although the radius β of S^1 and the world-line circumference (the unity) have nothing to do with each other, the shift (1.2) involves the world-line coordinate σ as well. In this sense, the way of prescribing the temperature and hence the chemical potential becomes different from the standard Matsubara formalism. To introduce the chemical potential, one may of course transform a world-line amplitude formula to the Feynman–Matsubara form, and then resume the world-line form after some efforts to apply the shift (1.1). However, such a calculation does not utilize the merits of the master formula at all, and there should be a more direct and transparent method to introduce μ within the world-line formalism (without tracing back and referring any internal loop calculation).

To this end, we propose a new rule, instead of (1.1), to introduce the internal chemical potential. It is simply done by applying the new shift procedure

$$\overline{\omega} \to \overline{\omega} + i\mu, \tag{1.3}$$

where $\overline{\omega}$ is an average of the zeroth components of external continuous/discrete momenta k_j^0, $j=1,2,...N$, with the summation weights σ_j (the local coordinates of the external legs on the closed loop). The prescription (1.3) is neither conceivable nor explicable from the standard method (1.1), because ω_n is the internal momentum while $\overline{\omega}$ concerns the external one. The parameter $\overline{\omega}$ will easily be identified in due course, if we adopt a statistical parameter to discern between boson and fermion loops when summing up the S^1 paths: This parameter dependence should vanish at the $\beta\to\infty$ limit as expected in the nonthermodynamic world-line formalism.

Another nontriviality in this formalism is how to derive a nonanalytic property at $\beta=\infty$ with finite μ from the master amplitude formula. Since the nonanalytic property can be derived from the $\beta=\infty$ limits of pure thermodynamic parts, we first have to separate a pure thermodynamic part $\tilde{\Gamma}_N^{\beta\mu}$ from a full thermodynamic amplitude $\Gamma_N^{\beta\mu}$. If the master integrand of $\Gamma_N^{\beta\mu}$ is composed only of the Jacobi Θ-function, the story is simple as expected. However, in a more complicated case like a photon scattering, the master integrand is not such a simple form but a product of a β- and μ-dependent operator $\mathcal{V}_{\beta\mu}$ and the Jacobi Θ-function part $\mathcal{K}_{\beta\mu}$. The pure thermodynamic parts of these quantities, $\tilde{\mathcal{V}}_{\beta\mu}$ and $\tilde{\mathcal{K}}_{\beta\mu}$, can easily be separated from their original full quantities, but we show that the pure thermodynamic part $\tilde{\Gamma}_N^{\beta\mu}$ is nontrivially given by $\mathcal{V}_{\beta\mu}\times\tilde{\mathcal{K}}_{\beta\mu}$ against naive expectation. (The simple arithmetical splittings indicate one more contribution, however, we shall show that it vanishes.) Separating the $\tilde{\Gamma}_N^{\beta\mu}$ in this way, we then analyze the nonanalytic property of the master amplitude at zero temperature.

This paper is organized as follows. In Sec. II, we explain our notations, definitions, and the general structure of master amplitudes at zero temperature. In Sec. III, we derive a set of general formulas for a master amplitude at finite β and μ parts by parts: path integral normalization and (scalar) kinematical factor in Sec. III A, and reduced kinematical factor in Sec. III b. In Sec. III c, we show a general master formula for the purely thermodynamic part of the full amplitude $\Gamma_N^{\beta\mu}$, and also prove the above statement, i.e., $\tilde{\Gamma}_N^{\beta\mu} \sim \mathcal{V}_{\beta\mu} \times \tilde{K}_{\beta\mu}$. The replacement rule (1.3) is verified in Appendices A–C from the viewpoints of both Feynman rule's calculation and the world-line formalism. In Secs. IV and V, for arbitrary N, we derive various integral formulas by analyzing two kinds of Θ-function representations, and check their consistency. Several explicit results are also presented up to $N=5$. In Sec. VI, we derive the nonanalytic property of the master formula in the zero temperature limit with the chemical potential kept finite. Section VII contains summary and conclusions.

II. NOTATIONS AND DEFINITIONS

First, we summarize the general structure of the one-loop N-point master amplitudes in the nonthermodynamic world-line formalism. (We assume that a particle change does not occur while circulating along the loop.) We refer the reader to Refs. 12, 16, and 20 for more details. The master amplitude of general form (for N external momenta k_j^μ, $j=1,2,...,N$, $\mu=0,1,...,D-1$) is written in terms of the closed path integrals of bosonic $x^\mu(\sigma)$ and fermionic $\psi^\mu(\sigma)$ world-line fields as follows:

$$\Gamma_N = \frac{1}{2}\int_0^\infty \frac{ds}{s} \oint \mathcal{D}x^\mu(\sigma)\mathcal{D}\psi^\mu(\sigma)\exp\left[-\int_0^1 \mathcal{L}(\sigma)\, d\sigma\right]\prod_{j=1}^N V_j \tag{2.1}$$

with the world-line Lagrangian and the vertex operators

$$\mathcal{L}(\sigma) = \frac{1}{4s}\left(\frac{\partial x^\mu(\sigma)}{\partial\sigma}\right)^2 + \frac{1}{2}\,\psi^\mu(\sigma)\frac{\partial}{\partial\sigma}\,\psi_\mu(\sigma) + sm^2, \tag{2.2}$$

$$V_j = s\int_0^1 d\sigma\, v_j[x(\sigma),\psi(\sigma)]e^{ik_j\cdot x(\sigma)}, \quad j=1,2,...,N, \tag{2.3}$$

where $k_j\cdot x$ stands for the Lorentz contraction, and we often omit the Lorentz indices as long as obvious. The zero mode integral of the bosonic path integral should be excluded.[18] The explicit form of v_j depends on what particle is inserted in the loop as an external leg; for example, $v_j=1$ for ϕ^3 theory, and is in Eq. (C2) for the photon vertex case. Note that our world-line coordinate σ is dimensionless, and is related to the standard notation[20] by scaling $\tau=s\sigma$. For the path integral average of a general quantity F,

$$\langle F(x,\psi)\rangle \equiv \mathcal{N}^{-1}\oint \mathcal{D}x\mathcal{D}\psi\exp\left[-\int_0^1 \mathcal{L}(\sigma)\, d\sigma\right]F(x,\psi), \tag{2.4}$$

one may use the Wick contractions with (the Euclidean metric is given by $g^{\mu\nu}=-\delta^{\mu\nu}$)

$$\langle x^\mu(\sigma_1)x^\nu(\sigma_2)\rangle = -sg^{\mu\nu}G(\sigma_1,\sigma_2), \tag{2.5}$$

$$\langle \psi^\mu(\sigma_1)\psi^\nu(\sigma_2)\rangle = \tfrac{1}{2}g^{\mu\nu}\,\mathrm{sign}\,(\sigma_1-\sigma_2), \tag{2.6}$$

where $\mathcal{N}$ is the path integral normalization

$$\mathcal{N}=\oint \mathcal{D}x\mathcal{D}\psi \exp\left[-\int_0^1 \mathcal{L}(\sigma)\, d\sigma\right]=e^{-sm^2}(4\pi s)^{-D/2}, \tag{2.7}$$

and G is the bosonic world-line correlator

$$G(\sigma_i,\sigma_j)=|\sigma_i-\sigma_j|-(\sigma_i-\sigma_j)^2\equiv G_{ij}. \tag{2.8}$$

Performing the path integrals (or Wick contractions), we arrive at the following master amplitude formula:

$$\Gamma_N=c\int_0^1 d\sigma_1\int_0^1 d\sigma_2\cdots\int_0^1 d\sigma_N\int_0^\infty ds\, s^{N-1}\mathcal{V}\times\mathcal{K}, \tag{2.9}$$

where the constant c depends on the particle circulating in a loop. (It is also related to an over-counting factor of σ-integration regions.[26]) For example,

$$c=\begin{cases}\frac{1}{2} & \text{for a (neutral) scalar loop}\\ -\frac{1}{2}\,\mathrm{tr}[1] & \text{for a fermion loop}\end{cases}, \tag{2.10}$$

where the $\mathrm{tr}[1]$ expresses the trace of a unit matrix in the D-dimensional gamma matrix space. For a gluon loop, c is no longer a simple constant.[16,17] The quantity $\mathcal{K}$, which we shall call the *kinematical factor* (multiplied by the normalization $\mathcal{N}$), is defined by the world-line path integral average of N scalar vertex operators[10,16]

$$\mathcal{K}\equiv\mathcal{N}\left\langle\prod_{j=1}^N e^{ik_jx(\sigma_j)}\right\rangle=\oint \mathcal{D}x^\mu\mathcal{D}\psi^\mu\left(\prod_{j=1}^N e^{ik_jx(\sigma_j)}\right)\exp\left[-\int_0^1\mathcal{L}(\sigma)\,d\sigma\right]. \tag{2.11}$$

The $\mathcal{V}$ stands for an effective vertex function, which is obtained by

$$\mathcal{V}=\frac{\langle\Pi_{j=1}^N v_j \exp[ik_j\cdot x(\sigma_j)]\rangle}{\langle\Pi_{j=1}^N \exp[ik_j\cdot x(\sigma_j)]\rangle}. \tag{2.12}$$

We shall refer to this quantity as the *vertex structure function/operator*, which corresponds to the quantities called the *reduced kinematical factor*[1] or the generating kinematical factor.[16] At this stage, the $\mathcal{V}$ is still a function of s and σ_j $(j=1,2,...,N)$, however, it will be generalized to an operator in the thermodynamic case.

If the vertex structure function can be expanded in the form

$$\mathcal{V}=\sum_{l\in\mathbb{Z}} a_l s^{l-N}\exp[-sb_l], \tag{2.13}$$

where the coefficients a_l and b_l may not depend on s, but on σ_j, and the amplitude (2.9) is obtained as the sum of the ''partial'' amplitudes

$$\mathcal{A}_l\overset{\text{def.}}{=}\int_0^\infty ds\, s^{l-1}\mathcal{K}=\int_0^\infty ds\, s^{l-1}\mathcal{N}\left\langle\prod_{j=1}^N e^{ik_j\cdot x(\sigma_j)}\right\rangle, \tag{2.14}$$

with shifting m^2 to m^2+b_l.

In the case of finite β, we assume the zeroth components of the external momenta k_j^μ to be the bosonic Matsubara frequencies

$$k_j^0=\omega_{k_j}=\frac{2\pi}{\beta}n_j, \tag{2.15}$$

as well as

$$k_j \cdot x \equiv \omega_{k_j} x^0 - \mathbf{k}_j \cdot \mathbf{x}. \tag{2.16}$$

In this paper, we shall not consider the fermionic external states, since the bosonic external states are only well formulated in the world-line formalism—however, one may formally generalize to the fermionic states. We use the notations for the counterparts of Eqs. (2.9), (2.11), (2.7), (2.12), and (2.14):

$$\Gamma_N^\beta, \quad \mathcal{K}_\beta, \quad \mathcal{N}_\beta, \quad \mathcal{V}_\beta, \quad \mathcal{A}_l^\beta. \tag{2.17}$$

In the case of finite μ and β, we denote

$$\Gamma_N^{\beta\mu}, \quad \mathcal{K}_{\beta\mu}, \quad \mathcal{N}_{\beta\mu}, \quad \mathcal{V}_{\beta\mu}, \quad \mathcal{A}_l^{\beta\mu}. \tag{2.18}$$

We explain how to define these quantities in the next section.

III. THE THERMODYNAMIC GENERALIZATION

A. The kinematical factor $\mathcal{K}_{\beta\mu}$ and the normalization $\mathcal{N}_{\beta\mu}$

We take a two-step procedure to introduce the temperature and the chemical potential. As the first step, we consider the $\mu=0$ case. Basically we follow the same method as discussed in Refs. 22–25, and we start with the following slightly general definition:

$$\Gamma_N^\beta \overset{\text{def.}}{=} \sum_{n=-\infty}^{\infty} e^{2n\pi i/\epsilon} \Gamma_N \big|_{x^0(\sigma)\to x^0(\sigma)+n\beta\sigma}, \tag{3.1}$$

where ϵ is a constant related to the statistics of the loop. For example, the $\epsilon=2$ case corresponds to a fermion loop, and the $\epsilon=\infty$ case to a bosonic loop. Without specifying the value of ϵ, we deal with both cases simultaneously (formally fractional statistics as well). The Γ_N on the right-hand side (rhs) of the above formula denotes the path integral representation (2.1), and the replacement (1.2) should be applied to the x^0 fields for all variables σ_j. In a nutshell, we have only to replace the bosonic path integral in the following way:

$$\oint \mathcal{D}x^\mu \to \oint_\beta \mathcal{D}x^\mu \equiv \sum_{n=-\infty}^{\infty} e^{2n\pi i/\epsilon} \oint_{x^0(\sigma)\to x^0(\sigma)+n\beta\sigma} \mathcal{D}x^\mu. \tag{3.2}$$

For simplicity, let us consider the $\mathcal{V}=1$ case, or the ''partial'' amplitude $\mathcal{A}_N$ [the $l=N$ term in (2.13)]. In this case, we have only to generalize the kinematical factor $\mathcal{K}$ to the finite temperature version $\mathcal{K}_\beta$;

$$\mathcal{K}_\beta \overset{\text{def.}}{=} \sum_{n=-\infty}^{\infty} e^{2n\pi i/\epsilon} \mathcal{N} \left\langle \prod_{j=1}^{N} e^{ik_j\cdot x(\sigma_j)} \right\rangle \Bigg|_{x^0(\sigma)\to x^0(\sigma)+n\beta\sigma} \tag{3.3}$$

$$= (4\pi s)^{-D/2} e^{-sM^2} \Theta_3\left(\frac{1}{\epsilon} + \frac{\beta\overline{\omega}}{2\pi}, i\frac{\beta^2}{4\pi s} \right), \tag{3.4}$$

where the definition of $\Theta_3(v,\tau)$ is

$$\Theta_3(v,r) = \sum_{n=-\infty}^{\infty} e^{n^2\tau\pi i} e^{2nv\pi i}, \tag{3.5}$$

and we have introduced the following two s-independent quantities:

$$M^2 = m^2 - \sum_{i<j,=1}^{N} k_i \cdot k_j G_{ij}, \tag{3.6}$$

$$\overline{\omega} = \sum_{j=1}^{N} \sigma_j \omega_{k_j}. \tag{3.7}$$

Here the two-point correlator (world-line Green's function) G_{ij}, given by Eq. (2.8), looks different from the one used in Ref. 22; however, the value of M^2 does not differ under the condition of momentum conservation regarding the external legs. The sign of M^2 seems not always to be positive in spite that $0 \leqslant G_{ij} \leqslant \frac{1}{2}$. We then assume $M^2 \geqslant 0$ in the following discussion, choosing the off-shell symmetric point (satisfying the momentum conservation constraint)

$$k_i k_j = \delta_{ij} k^2 + (\delta_{ij} - 1) \frac{1}{N-1} k^2. \tag{3.8}$$

For the $\mathcal{V} \neq 1$ case, we have to apply Eq. (3.2) to the $\mathcal{V}$ and $\mathcal{K}$ parts at the same time; however, the things are rather straightforward. We refer the reader to Appendix C for an example.

The second step is to introduce the chemical potential for the internal loop. One may do it with applying the Jacobi transformation to the quantity $\mathcal{K}_\beta$; i.e., rewriting Eq. (3.4) in such a way to revive the discrete summation over internal Matsubara frequencies ω_n, one may perform the replacement (1.1) (see Ref. 24 for more details). However, this is a roundabout way, since one has to make the Matsubara summation reappear in spite of dealing with the world-line formulation, where the integration and summation of the loop are already performed. Instead, we propose a much simpler and direct alternative method to steer clear of this problem. It can be promptly done by the replacement (1.3),

$$\overline{\omega} \to \Omega \equiv \overline{\omega} + i\mu. \tag{3.9}$$

This shift looks similar to (1.1); however, we stress that our shifting parameter is not the internal frequency but an average of external ones [cf. Eq. (3.7)]. In this sense, this prescription is nontrivial. For a rigorous reader, we put a justification of this replacement in Appendix A (the case of $\mathcal{V}=1$), and also refer to Appendix B for a more complicated case (the $\mathcal{V} \neq 1$ case). After all, the general path integral average for finite μ can be obtained in a couple of simple replacements:

$$\oint \mathcal{D}x^\mu \to \oint_\beta \mathcal{D}x^\mu \to \oint_{\beta\mu} \mathcal{D}x^\mu \equiv \oint_\beta \mathcal{D}x^\mu \big|_{\overline{\omega}\to\Omega}. \tag{3.10}$$

Note that the fermion path integral does not change by any means.

Let us write down the thermodynamic kinematical factor $\mathcal{K}_{\beta\mu}$ and the normalization factor $\mathcal{N}_{\beta\mu}$. Applying the above replacement to Eq. (3.4), we thereby yield the desired thermodynamic kinematical factor for finite β and μ:

$$\mathcal{K}_{\beta\mu} \overset{\text{def.}}{=} \text{Eq. (3.4)}\big|_{\overline{\omega}\to\Omega} \tag{3.11}$$

$$= (4\pi s)^{-D/2} e^{-sM^2} \Theta_3\left(i\frac{\beta\overline{\mu}}{2\pi}, i\frac{\beta^2}{4\pi s} \right), \tag{3.12}$$

where we have introduced the shorthand notation $\overline{\mu}$ by reason of analogy to the vacuum amplitude of a scalar loop

$$\overline{\mu} \equiv \mu - i\overline{\omega} - \frac{2\pi i}{\epsilon\beta} = -i\left(\Omega + \frac{2\pi}{\epsilon\beta} \right). \tag{3.13}$$

We refer to the representation (3.12) as the *first representation*. This representation with $\overline{\mu}=0$ ($\mu=\overline{\omega}=0$, $\epsilon=\infty$) is studied in Ref. 22. We also obtain another expression (which we shall call the *second representation*) through the Jacobi transformation,

$$\mathcal{K}_{\beta\mu}=\frac{1}{\beta}(4\pi s)^{(1-D)/2}e^{-s(M^2-\overline{\mu}^2)}\Theta_3\left(\frac{2s\overline{\mu}}{\beta},i\frac{4\pi s}{\beta^2}\right). \tag{3.14}$$

Note that for a fermion loop ($\epsilon=2$), the above first and the second Θ_3 representations become the Θ_4 and Θ_2 representations, respectively.

We can extract the thermodynamic quantity $\mathcal{N}_{\beta\mu}$ corresponding to the path integral normalization (2.7) from the kinematical factor $\mathcal{K}_{\beta\mu}$. If we rewrite Eq. (3.12) as

$$\mathcal{K}_{\beta\mu}=e^{-sm^2}(4\pi s)^{-D/2}\Theta_3\left(i\frac{\beta\overline{\mu}}{2\pi},i\frac{\beta^2}{4\pi s}\right)\exp\left[s\sum_{i<j}^{N}k_i\cdot k_jG_{ij}\right]$$
$$=\frac{1}{\beta}(4\pi s)^{(1-D)/2}e^{-s(m^2-\overline{\mu}^2)}\Theta_3\left(\frac{2s\overline{\mu}}{\beta},\frac{4\pi is}{\beta^2}\right)\left\langle\prod_{j=1}^{N}e^{ik_j\cdot x(\sigma_j)}\right\rangle, \tag{3.15}$$

the noncorrelator part can be regarded as an overall normalization to the N-point correlator of zero temperature type (the zeroth components k_j^0 are formally regarded as continuous variables here). Paraphrasing this fact in analogy to Eq. (2.11),

$$\mathcal{K}_{\beta\mu}\equiv\mathcal{N}_{\beta\mu}\left\langle\prod_{j=1}^{N}e^{ik_j\cdot x(\sigma_j)}\right\rangle=\oint_{\beta\mu}\mathcal{D}x\mathcal{D}\psi\exp\left[-\int_0^1\mathcal{L}(\sigma)d\sigma\right]\prod_{j=1}^{N}e^{ikj\cdot x(\sigma_j)}, \tag{3.16}$$

the thermodynamic version of the path integral normalization (for finite β and μ) is found to be

$$\mathcal{N}_{\beta\mu}=\frac{1}{\beta}(4\pi s)^{(1-D)/2}e^{-s(m^2-\overline{\mu}^2)}\Theta_3\left(\frac{2s\overline{\mu}}{\beta},\frac{4\pi is}{\beta^2}\right). \tag{3.17}$$

This is exactly the same normalization as assumed in Ref. 21 (taking a mass term inclusion into account).

B. The vertex structure operator $\mathcal{V}_{\beta\mu}$

In this subsection, we discuss the part of vertex structure ($\mathcal{V}\neq 1$). Because of the diversities of explicit forms of $\mathcal{V}$, there is no concrete formula such as Eq. (3.12). However, we can derive a general property of the pure thermodynamic part $\tilde{\mathcal{V}}_{\beta\mu}$. To explain this, we classify $\mathcal{V}$ into two categories by the criterion whether or not $\mathcal{V}$ contains local world-line variables σ_j.

First, let us start with the first category, the σ-independent $\mathcal{V}$ case. Apparently, $\mathcal{V}=1$ is the case. A nontrivial example of this category is the $\pi^0\to 2\gamma$ decay ($\epsilon=2$, $D=4$) without background field:[24]

$$c\mathcal{V}=-\mathrm{tr}[1]m\lambda e^2\epsilon_{\mu\nu\rho\sigma}\epsilon_1^\mu\epsilon_2^\nu k_1^\rho k_2^\sigma, \tag{3.18}$$

where λ is the pseudo-scalar coupling, m and e the (space–time) fermion mass and the quantum electrodynamics (QED) coupling constant. Another nontrivial example is the effective potential of a fermion loop ($\epsilon=2$, $N=0$) in a constant magnetic field in D dimensions.[21] In this case, $\mathcal{V}$ depends on the integration variable s,

$$c\mathcal{V}=-\tfrac{1}{2}\mathrm{tr}[1]sB\coth(sB), \tag{3.19}$$

and this is the case of the expansion (2.13), if we use

$$\coth(sB)=1+2\sum_{n=1}^{\infty} e^{-2nsB}. \tag{3.20}$$

With the shift $M^2\to M^2+2lB$, this case is essentially described by the ''partial'' amplitude $\mathcal{A}_1$ defined by Eq. (2.14). Now, at finite values of β and μ, we can, in principle, express any thermodynamic vertex structure function $\mathcal{V}_{\beta\mu}$ as

$$\mathcal{V}_{\beta\mu}=\mathcal{V}+\tilde{\mathcal{V}}_{\beta\mu}, \tag{3.21}$$

where $\tilde{\mathcal{V}}_{\beta\mu}$ denotes the purely thermodynamic part. However, in this category, as can be seen from the above examples, we simply have

$$\mathcal{V}_{\beta\mu}=\mathcal{V}, \quad \tilde{\mathcal{V}}_{\beta\mu}=0 \quad (\text{for } \sigma\text{-independent } \mathcal{V}), \tag{3.22}$$

because the $\mathcal{V}$'s of this category do not contain the local world-line variables σ_j, strictly speaking the G_{ij} which are generated by the bosonic field correlation from v_j. Recall that the β dependence only appears through the shift $x^0\to x^0+n\beta\sigma$ in the formula (3.1). The β dependence cannot be created from the quantity which does not contain an x-field correlation.

Second, let us consider the other category, the σ-dependent $\mathcal{V}$ case. In this category, we obtain a nonzero structure function $\tilde{\mathcal{V}}_{\beta\mu}$. We here consider the photon polarization case by way of example. In this case, the $\mathcal{V}$ is given by Eq. (C5) at zero temperature:[16]

$$\mathcal{V}=\epsilon_1\cdot\epsilon_2\frac{1}{s}\ddot{G}_{12}+\epsilon_1\cdot k_2\epsilon_2\cdot k_1(\dot{G}_{12})^2+\epsilon_1\cdot\epsilon_2 k_1\cdot k_2-\epsilon_1\cdot k_2\epsilon_2\cdot k_1, \tag{3.23}$$

where dots on G_{12} means the first and the second derivatives wrt the first argument of G_{12}. The thermodynamic generalization can be done by applying the formulas (3.1) and (3.9). For ease of presentation, we put the computational details in Appendix C, and a comparison with the Feynman diagram technique is also in Appendix B. For finite β, we derive Eq. (C17):

$$\mathcal{V}_\beta=\mathcal{V}-\epsilon_0^1\epsilon_0^2\left(\frac{1}{s}\frac{\partial}{\partial\bar{\omega}}\right)^2-\frac{1}{s}\frac{\partial}{\partial\bar{\omega}}(\epsilon_0^1\epsilon_2\cdot k_1-\epsilon_0^2\epsilon_1\cdot k_2)\dot{G}_{12}, \tag{3.24}$$

and then, shifting $\bar{\omega}\to\Omega$, we acquire the operator part given by Eq. (C19):

$$\tilde{\mathcal{V}}_{\beta\mu}=-\epsilon_0^1\epsilon_0^2\left(\frac{1}{s}\frac{\partial}{\partial\Omega}\right)^2-\frac{1}{s}\frac{\partial}{\partial\Omega}(\epsilon_0^1\epsilon_2\cdot k_1-\epsilon_0^2\epsilon_1\cdot k_2)\dot{G}_{12}. \tag{3.25}$$

Note that the origin of $\partial/\partial\Omega$ is the Wick contractions of the bosonic world-line fields furnished in the v_j parts of photon vertex operators [see Eqs. (C2) and (C10)], and it always happens if a vertex operator comprises a bosonic field in such a way.

We therefore conclude that if $\mathcal{V}$ includes G_{ij} or its derivatives, $\tilde{\mathcal{V}}_{\beta\mu}$ gives rise to a differential polynomial in $\partial/\partial\Omega$, and otherwise $\tilde{\mathcal{V}}_{\beta\mu}=0$. An important result following from this fact is

$$\tilde{\mathcal{V}}_{\beta\mu}\times\mathcal{K}=0 \quad (\text{for all } \mathcal{V}). \tag{3.26}$$

This is a model-independent result, and we shall make use of this relation in order to decouple the pure thermodynamic part $\tilde{\Gamma}_N^{\beta\mu}$ from the total N-point amplitude $\Gamma_N^{\beta\mu}$.

C. The master formulas $\tilde{\Gamma}_N^{\beta\mu}$ and $\tilde{\mathcal{A}}_l^{\beta\mu}$

Gathering the formulas obtained in the above sections, we compose the purely thermodynamic master amplitude $\tilde{\Gamma}_N^{\beta\mu}$, and define the thermodynamic ''partial'' amplitudes $\mathcal{A}_l^{\beta\mu}$. Applying the following formula to the first representation (3.12),

$$\Theta_3(v,\tau)=1+2\sum_{n=1}^{\infty} e^{n^2\tau\pi i}\cos(2n\pi v), \tag{3.27}$$

we separate the pure thermodynamic part $\tilde{\mathcal{K}}_{\beta\mu}$ from $\mathcal{K}_{\beta\mu}$ as

$$\mathcal{K}_{\beta\mu}=\mathcal{K}+\tilde{\mathcal{K}}_{\beta\mu}, \tag{3.28}$$

where

$$\tilde{\mathcal{K}}_{\beta\mu}=2(4\pi s)^{-D/2}e^{-sM^2}\sum_{n=1}^{\infty} e^{-n^2\beta^2/4s}\cosh(n\beta\overline{\mu}), \tag{3.29}$$

or, for the second representation (3.14),

$$\tilde{\mathcal{K}}_{\beta\mu}=\frac{2}{\beta}(4\pi s)^{(1-D)/2}e^{-s(M^2-\overline{\mu}^2)}\sum_{n=1}^{\infty} e^{-s(2n\pi/\beta)^2}\cos\left(\frac{4n\pi s\overline{\mu}}{\beta}\right). \tag{3.30}$$

For a given $\mathcal{V}_{\beta\mu}$, the thermodynamic master amplitude $\Gamma_N^{\beta\mu}$ is calculated as

$$\Gamma_N^{\beta\mu}\overset{\text{def.}}{=}c\int_0^1 d\sigma_1\int_0^1 d\sigma_2\cdots\int_0^1 d\sigma_N\int_0^{\infty} ds\, s^{N-1}\mathcal{V}_{\beta\mu}\times\mathcal{K}_{\beta\mu}. \tag{3.31}$$

Using the decompositions (3.21) and (3.28) with the general formula (3.26), we find

$$\Gamma_N^{\beta\mu}=\Gamma_N+\tilde{\Gamma}_N^{\beta\mu}, \tag{3.32}$$

where

$$\tilde{\Gamma}_N^{\beta\mu}\overset{\text{def.}}{=}c\int_0^1 d\sigma_1\int_0^1 d\sigma_2\cdots\int_0^1 d\sigma_N\int_0^{\infty} ds\, s^{N-1}\mathcal{V}_{\beta\mu}\times\tilde{\mathcal{K}}_{\beta\mu}. \tag{3.33}$$

Therefore we have only to separate the pure thermodynamic part of $\mathcal{K}_{\beta\mu}$ in order to obtain the purely thermodynamic part $\tilde{\Gamma}_N^{\beta\mu}$: no separation in the $\mathcal{V}_{\beta\mu}$ part at all. [Note the difference between Eqs. (3.31) and (3.33).]

As can be inferred from the examples in Sec. III B, a general expansion form of $\mathcal{V}_{\beta\mu}$ is

$$\mathcal{V}_{\beta\mu}=\sum_{l,n\in Z} a_{ln}s^{l-N}\frac{\partial^n}{\partial\Omega^n}\exp[-sb_l]. \tag{3.34}$$

where the coefficients a_{ln} and b_l may not depend on s, but on $\sigma_j(j=1,2,...,N)$. In analogy to Eq. (2.14), the relevant ''partial'' amplitudes for Eqs. (3.31) and (3.33) are defined as

$$\mathcal{A}_l^{\beta\mu}\overset{\text{def.}}{=}\int_0^{\infty} ds\, s^{l-1}\mathcal{K}_{\beta\mu}|_{m^2\to m^2+b_l}, \tag{3.35}$$

$$\tilde{\mathcal{A}}_l^{\beta\mu} \overset{\text{def.}}{=} \int_0^\infty ds\, s^{l-1} \tilde{\mathcal{K}}_{\beta\mu}|_{m^2 \to m^2 + b_l}, \tag{3.36}$$

and of course the following relation holds:

$$\mathcal{A}_l^{\beta\mu} = \mathcal{A}_l + \tilde{\mathcal{A}}_l^{\beta\mu}. \tag{3.37}$$

In the following sections, we present various integral representations for the pure thermodynamic parts $\tilde{\mathcal{A}}_l^{\beta\mu}$ of the N-point ''partial'' amplitudes without specifying any values of N, D, and ϵ. These parts are the essential quantities to analyze the zero temperature limits with revealing the nonanalyticity on μ.

IV. THE INTEGRAL FORMULAS FROM THE FIRST REPRESENTATION

In this section, we derive various integral formulas for $\tilde{\mathcal{A}}_l^{\beta\mu}$ based on the first representation (3.12). This representation is examined in the special case $l=1$ (with $N=0$ and $\epsilon=2$), and actually $(4\pi)^{D/2}\tilde{\mathcal{A}}_1^{\beta\mu}$ is the function $\mathcal{O}_\beta$ analyzed in Ref. 21:

$$\mathcal{O}_\beta(m) = 4\sum_{n=1}^{\infty} (-1)^n \cosh(n\beta\mu)\left(\frac{n\beta}{2m}\right)^{1-D/2} K_{D/2-1}(n\beta m). \tag{4.1}$$

We want to generalize this function to more generic l and ϵ cases, and it is convenient to mimic the computational technique of Ref. 21, introducing the parallel notation

$$\mathcal{O}_\beta^{(k)}(M) \overset{\text{def.}}{=} (4\pi)^{D/2}\tilde{\mathcal{A}}_l^{\beta\mu}; \quad k \equiv 2l+1-D. \tag{4.2}$$

Hereafter, for simplicity we set

$$b_l = 0, \tag{4.3}$$

and, for later convenience, we also define

$$d \equiv 3-k = D+2-2l. \tag{4.4}$$

From Eqs. (3.29) and (3.36), the pure thermodynamic part $\tilde{\mathcal{A}}_l^{\beta\mu}$ is now given by

$$\mathcal{O}_\beta^{(k)}(M) = 2(4\pi)^{D/2}\int_0^\infty ds\, s^{l-1}(4\pi s)^{-D/2} e^{-sM^2}\sum_{n=1}^{\infty} e^{-n^2\beta^2/4s}\cosh(n\beta\overline{\mu}). \tag{4.5}$$

Performing the s integration, we obtain

$$\mathcal{O}_\beta^{(k)}(M) = 4\sum_{n=1}^{\infty} \cosh(n\beta\overline{\mu})\left(\frac{n\beta}{2M}\right)^{1-d/2} K_{d/2-1}(n\beta M), \tag{4.6}$$

where $K_\nu(z)$ is the modified Bessel function of second kind. This is a generalized version (for arbitrary ϵ and l) of the above function $\mathcal{O}_\beta$. The aim is to obtain integral representations with performing the summation in Eq. (4.6) (for generic k or d). To this end, we first apply the following formula to Eq. (4.6):

$$K_\nu(z) = \frac{\sqrt{\pi}(z/2)^\nu}{\Gamma(\nu+\frac{1}{2})}\int_1^\infty e^{-zt}(t^2-1)^{\nu-1/2}\,dt; \quad \mathrm{Re}\,\nu > -\frac{1}{2}, \mathrm{Re}\,z > 0, \tag{4.7}$$

and note that the $k<2(d>1)$ and $k\geqslant 2(d\leqslant 1)$ cases involve different calculations. Here we put a few remarks. For the convergence of the summation over n, we have to assume

$$\mu<M \tag{4.8}$$

in order to satisfy the condition

$$|e^{-\beta(Mt\pm\overline{\mu})}|=e^{-\beta(Mt\pm\mu)}<1;\quad t>1. \tag{4.9}$$

The $k=1,0,-1(d=2,3,4)$ cases are calculated in Ref. 21, and the $k=3(d=0)$ case corresponds to Ref. 24, although they did not discuss this representation. (Note that their argument belongs to our second representation.)

(i) In the first case, $k<2(d>1)$, the condition on ν in the formula (4.7) is satisfied as it is; i.e., $\nu=d/2-1>-\frac{1}{2}$, and the calculation is almost parallel to Ref. 21. We then just write down the result

$$\mathcal{O}_\beta^{(k)}(M)=\frac{-2\sqrt{\pi}}{\Gamma((d-1)/2)}\left[\int_0^\infty \frac{(u^2+2Mu)^{(d-3)/2}}{1-e^{\beta(u+M+\overline{\mu})}}\,du+(\overline{\mu}\to-\overline{\mu})\right], \tag{4.10}$$

where we note that the minus sign of the denominator differs from the previous formula [Eq. (3.13) in Ref. 21]. For later convenience, let us derive another formula from this result. Performing the change of variable $u=M(\sqrt{p^2+1}-1)$, and applying the formula $\Gamma(z)\Gamma(1-z)=\pi/\sin \pi z$ with $z=(d-1)/2$, we derive

$$\mathcal{O}_\beta^{(k)}(M)=\frac{2}{\sqrt{\pi}}\,\Gamma\left(\frac{k}{2}\right)M^{1-k}\sin\left(\frac{2-k}{2}\,\pi\right)\int_0^\infty \frac{p^{1-k}(p^2+1)^{-1/2}}{e^{\beta X_+(p)}-1}\,dp+(\overline{\mu}\to-\overline{\mu}), \tag{4.11}$$

where we have introduced the compact notation

$$X_\pm(p)=M\sqrt{p^2+1}\pm\overline{\mu}. \tag{4.12}$$

(ii) In the second case, $2\leqslant k(d\leqslant 1)$, the calculation is similar to the case (i), excepting the point that we apply the formula (4.7) for $\nu=1-d/2$ instead of $\nu=d/2-1$ owing to the relation $K_\nu(z)=K_{-\nu}(z)$. Then the summation over n in Eq. (4.6) leads to the Lerch transcendental function:

$$\begin{aligned}\mathcal{O}_\beta^{(k)}(M)&=\alpha\int_1^\infty dt\,(t^2-1)^{(1-d)/2}e^{-\beta(Mt+\overline{\mu})}\Phi(e^{-\beta(Mt+\overline{\mu})},d-2,1)+(\overline{\mu}\to-\overline{\mu})\\&=\frac{\alpha\Gamma(3-d)}{2\pi i}\int_1^\infty dt\,(t^2-1)^{(1-d)/2}\int_\infty^{(0+)}\frac{(-z)^{d-3}dz}{1-e^{z+\beta(Mt+\mu)}}+(\overline{\mu}\to-\overline{\mu}),\end{aligned} \tag{4.13}$$

where just for conciseness we have defined the coefficient

$$\alpha=\frac{\sqrt{4\pi}}{\Gamma((3-d)/2)}\left(\frac{\beta}{2}\right)^{2-d}=\frac{\sqrt{4\pi}}{\Gamma(k/2)}\left(\frac{\beta}{2}\right)^{k-1}. \tag{4.14}$$

The z-integrand (4.13) possesses poles at $z=0$ and $z=-\beta Mt\pm\beta\overline{\mu}+2n\pi i$ $(n\in Z)$. Since $-\beta(Mt\pm\mu)$ is a negative value, there is no pole in the positive region on the real axis other than $z=0$ as long as $\beta\overline{\omega}+2\pi/\epsilon$ is not an integer. With the change of variable $t=\sqrt{p^2+1}$, and a replacement of one derivative $\partial/\partial z$ by $\partial/\partial p$, we have

$$\mathcal{O}(M)=\alpha\,\frac{(-1)^{-k}\Gamma(k)}{M\beta(k-1)!}\,\frac{\partial^{k-2}}{\partial z^{k-2}}\int_0^\infty p^{k-2}\frac{\partial}{\partial p}\left(\frac{1}{1-e^{z+\beta X_+}}\right)dp\big|_{z=0}+(\overline{\mu}\to-\overline{\mu}). \tag{4.15}$$

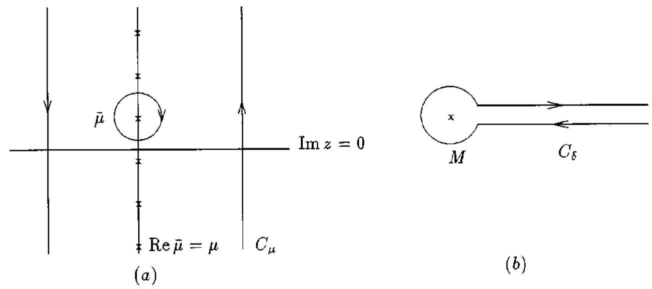

FIG. 1. The contours C_μ and C_δ. (a) The C_μ impounds an infinite number of poles at $z=\overline{\mu}+2n\pi i$, where n denotes nonzero integers. (b) Coming from $z=\infty+i\epsilon$ and going to $z=\infty-i\epsilon$, the C_δ impounds the pole at $z=M$.

Because of $k\geqslant 2$, the surface terms from a partial integral vanish in Eq. (4.15), and finally we find

$$\mathcal{O}_\beta^{(k)}(M)=2\Gamma\left(\frac{k+1}{2}\right)\frac{(-1)^{1-k}\beta^{k-2}(k-2)}{M(k-1)!}\frac{\partial^{k-2}}{\partial z^{k-2}}\int_0^\infty \frac{p^{k-3}}{1-e^{z+\beta X_+}}\,dp\Big|_{z=0}+(\overline{\mu}\to-\overline{\mu}). \tag{4.16}$$

More explicitly, we present the results for $k=2,3,4$ as follows:

$$\mathcal{O}_\beta^{(2)}(M)=-\frac{\sqrt{\pi}}{M}\left[\frac{1}{1-e^{\beta(M+\overline{\mu})}}+(\overline{\mu}\to-\overline{\mu})\right], \tag{4.17}$$

$$\mathcal{O}_\beta^{(3)}(M)=\frac{\beta}{M}\int_0^\infty \frac{e^{\beta X_+}}{(1-e^{\beta X_+})^2}\,dp+(\overline{\mu}\to-\overline{\mu}), \tag{4.18}$$

$$\mathcal{O}_\beta^{(4)}(M)=-\frac{\sqrt{\pi}}{2M^3}\left[\frac{1}{1-e^{\beta(M+\overline{\mu})}}-\frac{\beta M e^{\beta(M+\overline{\mu})}}{(1-e^{\beta(M+\overline{\mu})})^2}+(\overline{\mu}\to-\overline{\mu})\right]. \tag{4.19}$$

For even values of k, one may directly perform the summation (4.6) without using the integral representation (4.7), e.g., using $K_{1/2}(z)=\sqrt{\pi/2z}e^{-z}$ and $K_{3/2}(z)=\sqrt{\pi/2z}(1+z^{-1})e^{-z}$ for $k=2$ and 4. We will see in the next section that the results (4.11) and (4.17)–(4.19) can be reproduced from the second representation as well.

V. THE INTEGRAL FORMULAS FROM THE SECOND REPRESENTATION

In this section, based on the second representation (3.14) case, we derive some more formulas on $\mathcal{O}_\beta^{(k)}(M)$, with rederiving the results of the previous section. From Eqs. (3.30), (3.36), and (4.2), we have the following form of the pure thermodynamic part to start with:

$$\mathcal{O}_\beta^{(k)}(M)=\frac{(4\pi)^{1/2}}{\beta}\int_0^\infty ds\, s^{(k-2)/2}e^{-s(M^2-\overline{\mu}^2)}\sum_{n\in Z,\neq 0}e^{-s(2n\pi/\beta)^2}\cos(4n\pi s\overline{\mu}/\beta). \tag{5.1}$$

The summation can be converted to the integrals on the contour C_μ [see Fig. 1(a)]:

$$\mathcal{O}_\beta^{(k)}(M)=\frac{1}{2}(4\pi)^{1/2}\int_0^\infty ds\, s^{(k-2)/2}e^{-sM^2}\frac{1}{2\pi i}\int_{C_\mu}\frac{e^{sz^2}}{e^{\beta(z-\mu)}-1}\,dz+(\overline{\mu}\to-\overline{\mu}). \tag{5.2}$$

After performing the s integration, one may deform the contour C_μ into another one C_δ [Fig. 1(b)] in the same way as Ref. 24, thus obtaining

$$\mathcal{O}_\beta^{(k)}(M)=\sqrt{\pi}\Gamma\left(\frac{k}{2}\right)\frac{1}{2\pi i}\int_\infty^{(M+)}\frac{(M^2-z^2)^{-k/2}}{e^{\beta(z-\bar{\mu})}-1}\,dz+(\bar{\mu}\to-\bar{\mu}). \tag{5.3}$$

Parametrizing the circular part of C_δ (centered at M) by $z=M(1-\delta e^{i\phi})$, and using $p^2=z^2-1$ for the remaining parts of the contour, we arrive at

$$\mathcal{O}_\beta^{(k)}(M)=\frac{2}{\sqrt{\pi}}\Gamma\left(\frac{k}{2}\right)M^{1-k}\sin\left(\frac{2-k}{2}\pi\right)\times\left[\int_{\sqrt{2\delta}}^\infty\frac{p^{1-k}(p^2+1)^{-1/2}}{e^{\beta X_+}-1}\,dp+\frac{[1/(2-k)](2\delta)^{1-k/2}}{e^{\beta(M-\bar{\mu})}-1}+(\bar{\mu}\to-\bar{\mu})\right]. \tag{5.4}$$

Substituting $k=3$ and $\epsilon=2$, Eq. (5.4) reproduces the result (without background field) of Ref. 24.

Here are a few remarks on the computational difference between Ref. 24 and ours. We have extracted the pure thermodynamic part (5.1) from the beginning, and this fact is expressed by the circle at $\bar{\mu}$ in C_μ. On the other hand, in the method of Ref. 24, one needs an algebra to separate the nonthermodynamic part by using

$$\frac{-1}{e^{\beta(z-\mu)}-1}=1+\frac{1}{e^{-\beta(z-\mu)}-1}. \tag{5.5}$$

The first term on the rhs of Eq. (5.5) corresponds to the nonthermodynamic part, and the second term to the following expression instead of Eq. (5.2):

$$\mathcal{O}_{\beta\mu}^{(k)}(M)=(4\pi)^{1/2}\int_0^\infty ds\,s^{(k-2)/2}e^{-sM^2}\frac{1}{2\pi i}\times\left\{\int_{C_+}\frac{e^{s(z-i\bar{\omega})^2}}{e^{\beta(z-\bar{\mu}-i\bar{\omega})}-1}+\int_{C_-}\frac{e^{s(z-i\bar{\omega})^2}}{e^{-\beta(z-\bar{\mu}-i\bar{\omega})}-1}\right\}dz, \tag{5.6}$$

where the contours $C_\pm$ run from $\mu\pm o-i\infty$ to $\mu\pm o+i\infty$ for a small value of o. Although one can of course obtain the same result (5.4) from Eq. (5.6), it is not obvious at first glance that the C_- integral term corresponds to the $\bar{\mu}\to-\bar{\mu}$ term of Eq. (5.3).

Now, further results depend on the cases (i) and (ii). In the case (i), $k<2$, the second term in the square brackets on the rhs of Eq. (5.4) is proportional to $\delta^{1-k/2}$, and vanishes as $\delta\to 0$. Therefore Eq. (5.4) reproduces the first representation result (4.11).

In the case (ii), $k\geqslant 2$, the $\delta^{1-k/2}$ term diverges as $\delta\to 0$. We expect that it can be canceled with the lower surface term from the first integral in the square brackets on the rhs of Eq. (5.4). This can be seen by a partial integral with applying the formula

$$\frac{d}{dp}\left(\frac{\sqrt{p^2+1}}{p}\right)=\frac{1}{p^2\sqrt{p^2+1}} \tag{5.7}$$

to the p integration in Eq. (5.4) as follows:

$$\int_{\sqrt{2\delta}}^\infty\frac{p^{1-k}(p^2+1)^{-1/2}}{e^{\beta X_+}-1}\,dp=\frac{[1/(k-2)](2\delta)^{1-k/2}}{e^{\beta(M-\bar{\mu})}-1}+\frac{1}{2-k}\times\int_0^\infty p^{3-k}\left\{\frac{(p^2+1)^{-3/2}}{e^{\beta X_+}-1}+\frac{M\beta(p^2+1)^{-1}e^{\beta X_+}}{(e^{\beta X_+}-1)^2}\right\}dp. \tag{5.8}$$

The first term on the rhs of Eq. (5.8) cancels the divergence as expected, and we derive

$$\mathcal{O}_\beta^{(k)}(M)=\sqrt{4\pi}\Gamma\left(\frac{k}{2}\right)M^{1-k}\frac{\sin([(2-k)/2]\pi)}{(2-k)\pi}\times\left[\int_0^\infty p^{3-k}\left\{\frac{(p^2+1)^{-3/2}}{e^{\beta X_+}-1}+\frac{M\beta(p^2+1)^{-1}e^{\beta X_+}}{(e^{\beta X_+}-1)^2}\right\}dp+(\bar{\mu}\to-\bar{\mu})\right]. \tag{5.9}$$

Putting $t^2=p^2+1$ in Eq. (5.9), we finally obtain the following concise expression:

$$\mathcal{O}_\beta^{(k)}(M)=\sqrt{4\pi}\Gamma\left(\frac{k}{2}\right)M^{1-k}\frac{\sin([(2-k)/2]\pi)}{(2-k)\pi}\int_1^\infty\sqrt{t^2-1}^{2-k}\frac{\partial}{\partial t}\left[\frac{t^{-1}}{1-e^{\beta(Mt+\bar{\mu})}}\right]dt+(\bar{\mu}\to-\bar{\mu}), \tag{5.10}$$

which is an alternative representation of Eq. (4.16). As a result, we have derived two expressions in each case: Eqs. (4.10) and (4.11) in the case (i), and Eqs. (4.16) and (5.10) in the case (ii). In Eq. (4.16) we have to perform $k-2$ derivatives (after one integration), while just one integration in Eq. (5.10), whose integrand thus contains a common Boltzmann factor for all $k(\geqslant 2)$.

Let us check the consistency of Eq. (5.10) with the previous result (4.16) for $k=2,3,4$. The $k=2$ case (4.17) is immediate from the representation (5.10), however, let us handle these three cases simultaneously. Using $(p^2+1)^{-1}=1-p^2(p^2+1)^{-1}$, we first divide the second term in the curly brackets in Eq. (5.9), and then perform a partial integral on the rhs in the following quantity:

$$-\int_0^\infty p^{3-k}\frac{M\beta p^2(p^2+1)^{-1}e^{\beta X_+}}{(e^{\beta X_+}-1)^2}\,dp=\int_0^\infty p^{3-k}\frac{p}{\sqrt{p^2+1}}\frac{\partial}{\partial p}\left(\frac{1}{e^{\beta X_+}-1}\right)dp. \tag{5.11}$$

Changing the integration variable by $t^2=p^2+1$, we arrive at

$$\mathcal{O}_\beta^{(k)}(M)=\sqrt{4\pi}\Gamma\left(\frac{k}{2}\right)M^{1-k}\frac{\sin([(2-k)/2]\pi)}{(2-k)\pi}[I_k+(\bar{\mu}\to-\bar{\mu})], \tag{5.12}$$

where

$$I_k=\int_1^\infty\sqrt{t^2-1}^{2-k}t\left[\frac{k-3}{t}N(t)-N'(t)\right]dt \tag{5.13}$$

and $N'(t)\equiv\partial_t N(t)$ means the derivative of the function

$$N(t)=\frac{1}{e^{\beta(Mt-\bar{\mu})}-1}. \tag{5.14}$$

For the $k=2$ and 3 cases, this expression is sufficient to see the consistency in each. However, for the $k=4$ case, notice that we have to extract a singularity, which cancels a zero from the sine function in Eq. (5.10), and we hence perform a partial integral once more:

$$I_k=-\int_1^\infty\frac{1}{4-k}\sqrt{t^2-1}^{4-k}\frac{\partial}{\partial t}\left[(k-3)\frac{N(t)}{t}-N'(t)\right]dt,\quad 2\leqslant k<4, \tag{5.15}$$

where the surface terms can be dropped only when $k<4$ [also in Eq. (5.11)], and this is the reason why this formula is valid for $2\leqslant k<4(-1<d\leqslant 1)$. Substituting $k=2,3,4-\epsilon$ (with $\epsilon\to 0$), we verify that these expressions for I_k reproduce Eqs. (4.17)–(4.19). In this sense, Eq. (5.10) is equivalent to Eq. (4.16).

For further values of k, one should repeat the similar calculation for each interval between zero points of the sine function. For example, for $k=6$ and 8, we obtain

$$\mathcal{O}_\beta^{(6)}(M)=\frac{\pi}{16M^5}\,[3N(1)-3N'(1)+N''(1)], \tag{5.16}$$

$$\mathcal{O}_\beta^{(8)}(M)=\frac{\sqrt{\pi}}{8M^7}\,[15N(1)-15N'(1)+6N''(1)-N'''(1)]. \tag{5.17}$$

These pure thermodynamic ''partial'' amplitudes exactly correspond to the $N=4$ and 5 pure thermodynamic amplitudes in $D=3$ when $\mathcal{V}_{\beta\mu}=\mathcal{V}$ (the first category). In $D=4$, odd k integers are the similar cases. For the second category, for example, the N-photon amplitude in $D=4$, we need to combine the quantities from $k=-1$ to $2N-5$. These cases cannot get rid of nonintegrated quantities such as Eq. (4.18). After all, Eqs. (4.16) and (5.10) describe the general amplitude formulas which contain all Feynman diagrams.

VI. THE $\beta\to\infty$ LIMIT

In the above arguments, we have assumed $\mu<M$, and all $\beta\to\infty$ limits vanish because of it:

$$\mathcal{O}_\beta^{(k)}\xrightarrow{\beta\to\infty}0. \tag{6.1}$$

To obtain a nontrivial (nonzero) limit, we should remove this condition after all. In this sense, the μ dependence of the master amplitudes is nonanalytic. To see this, we need to transform the function $\mathcal{O}_\beta^{(k)}$ in a form indicating a Bose (or a Fermi) distribution, and we already derived this kind of representations in Secs. IV and V.

Let us begin with the case (i) $k<2$, in particular, the case of $k=0$ (which is also an odd dimensional case $D=2l+1$ by the way). From Eq. (4.10) with changing the variable $E=u+M$, we have

$$\mathcal{O}_\beta^{(0)}(M)=-2\sqrt{\pi}\int_M^\infty\frac{dE}{1\pm e^{\beta(E-\mu')}}+(\mu'\to-\mu'), \tag{6.2}$$

where the plus/minus signs correspond to the Fermi/Bose statistics, and we put

$$\mu'=\mu-i\overline{\omega}. \tag{6.3}$$

Equation (6.2) can be interpreted an effective action or total energy density with the chemical potential μ' and the mass $M(<\mu',E)$. Now, we understand in Eq. (6.2) that the $\mathcal{O}_\beta^{(0)}$ becomes zero as $\beta\to\infty$ if $\mu'<E$, while the $\mathcal{O}_\beta^{(0)}$ takes a nonzero value if $E<\mu'$. The similar arguments apply to the case of generic k value in the following way. Since we observe

$$e^{\beta(M-\overline{\mu}+u)}=e^{\beta(u+M-\mu)}e^{i\beta(\overline{\omega}+2\pi/\epsilon\beta)}\xrightarrow{\beta\to\infty}0\quad\text{for }u<\mu-M, \tag{6.4}$$

we have the finite upper boundary on the integral (4.10) at $u=\mu-M$, thus obtaining

$$\begin{aligned}\mathcal{O}_\infty^{(k)}(M)\equiv\lim_{\beta\to\infty}\mathcal{O}_\beta^{(k)}(M)&=\frac{-2\sqrt{\pi}}{\Gamma((2-k)/2)}\int_0^{\mu-M}(u^2+2Mu)^{-k/2}du\,\theta(\mu-M)\\&=\frac{-2\sqrt{\pi}}{\Gamma((4-k)/2)}(2M)^{-k/2}(\mu-M)^{(2-k)/2}F\left(\frac{k}{2},\frac{2-k}{2},\frac{4-k}{2};\frac{M-\mu}{2M}\right)\\&\quad\times\theta(\mu-M).\end{aligned} \tag{6.5}$$

The explicit results for $k=-1,0,1$ are

$$\mathcal{O}_\infty^{(1)}(M) = -2 \operatorname{arccosh}(\mu/M)\,\theta(\mu - M), \tag{6.6}$$

$$\mathcal{O}_\infty^{(0)}(M) = -2\sqrt{\pi}(\mu - M)\,\theta(\mu - M), \tag{6.7}$$

$$\mathcal{O}_\infty^{(-1)}(M) = [-2\mu\sqrt{\mu^2 - M^2} + 2M^2 \operatorname{arccosh}(\mu/M)]\,\theta(\mu - M). \tag{6.8}$$

In one of the $k=0$ cases ($D=3$, $N=0$, $l=1$, $\epsilon=2$),[21] Eq. (6.7) coincides with an exact result in Ref. 27.

The case (ii) $k \geqslant 2$ can also be estimated in the same way. Applying the similar treatment as above to the Boltzmann factor in Eq. (5.10), we then obtain the nonzero limit given by

$$\mathcal{O}_\infty^{(k)}(M) = -\sqrt{4\pi}\,\Gamma\left(\frac{k}{2}\right) M^{1-k} \frac{\sin([(2-k)/2]\pi)}{(2-k)\pi} \int_1^{\mu/M} (t^2 - 1)^{(2-k)/2} \frac{dt}{t^2}\,\theta(\mu - M). \tag{6.9}$$

Performing the t integration, some of the explicit results (the $k=2,3,4$ cases) can be shown as follows:

$$\mathcal{O}_\infty^{(2)}(M) = \frac{\sqrt{\pi}}{M}\left(\frac{M}{\mu} - 1\right)\theta(\mu - M), \tag{6.10}$$

$$\mathcal{O}_\infty^{(3)}(M) = -\frac{1}{\mu M}\sqrt{\left(\frac{\mu}{M}\right)^2 - 1}\,\theta(\mu - M), \tag{6.11}$$

$$\mathcal{O}_\infty^{(4)}(M) = -\frac{\sqrt{\pi}}{2M^3}\,\theta(\mu - M). \tag{6.12}$$

The calculations are rather straightforward for further k values, and the general formula (6.9) suffices. The ϵ dependence is gone away from the $\mathcal{O}_\infty^{(k)}$. This is natural because there is no difference in the kinematical factors (hence in ''partial'' amplitudes) between boson and fermion loops at zero temperature.

VII. CONCLUSIONS

In this paper, we studied the thermodynamic generalization of the one-loop amplitudes which can be cast into the BK master formula (2.9) at zero temperature (with zero chemical potential). We followed the two-step procedure: first, calculating the path integral S^1 summation (3.1) to introduce the temperature, and then performing the shift manipulation (1.3) to insert the chemical potential in a loop. This procedure has been applied parts by parts to the kinematical factor $\mathcal{K}$, the normalization factor $\mathcal{N}$, and the vertex structure function (reduced kinematical factor) $\mathcal{V}$, and we have derived the general formulas for these parts. From Eqs. (3.31), (3.34), and (3.35), the thermodynamic N-point amplitude of general form is thus summarized as

$$\Gamma_N^{\beta\mu} = c \sum_{l,n \in Z} a_{ln} \frac{\partial^n}{\partial \Omega^n} \int_0^1 d\sigma_1 \int_0^1 d\sigma_2 \cdots \int_0^1 d\sigma_N\, \mathcal{A}_l^{\beta\mu}. \tag{7.1}$$

One can check the validity of this master formula in various cases; the free effective potentials and the photon polarization in Appendix B, the effective potential in a constant magnetic field in Ref. 21, and the $\pi^0 \to 2\gamma$ decay in Ref. 24, etc.

The detail analyses on the building blocks $\mathcal{V}_{\beta\mu}$ and $\mathcal{K}_{\beta\mu}$ have given us useful information. We have realized that the total thermodynamic kinematical factor $\mathcal{K}_{\beta\mu}$ behaves as the thermodynamically generalized normalization to the zero-temperature-type correlator shown in Eq. (3.17). This is certainly a useful result and makes calculations for large N much simpler than ever; namely, we have only to attach the new normalization $\mathcal{N}_{\beta\mu}$ to the N-point scalar correlator (with switching

$k_j^0 \to \omega_{k_j}$). Another interesting point is that the pure thermodynamic part of a vertex structure function can be expressed in terms of either vanishing or Ω-differential operators, and this fact makes it possible to decouple the master formula (3.31) into $\Gamma_N^{\beta\mu} = \Gamma_N + \tilde{\Gamma}_N^{\beta\mu}$ in a nontrivial way [cf. Eqs. (3.31) and (3.33)]. It is worth noticing that we do not decouple the vertex structure function but the kinematical factor only. From Eqs. (3.33), (3.34), and (3.36), we then conclude that the pure thermodynamic master amplitude is simply given by

$$\tilde{\Gamma}_N^{\beta\mu} = c \sum_{l,n\in Z} a_{ln} \frac{\partial^n}{\partial\Omega^n} \int_0^1 d\sigma_1 \int_0^1 d\sigma_2 \cdots \int_0^1 d\sigma_N \, \tilde{\mathcal{A}}_l^{\beta\mu}. \tag{7.2}$$

Apart from Γ_N, the pure thermodynamic part is renormalization free, and hence Eq. (7.2) is the final form to apply our integral formulas derived in Secs. IV–VI. More explicitly, we have [from Eqs. (3.13) and (4.2)]

$$\tilde{\Gamma}_N^{\beta\mu} = c(4\pi)^{-D/2} \sum_{l,n\in Z} a_{ln}(-i)^n \frac{\partial^n}{\partial\bar{\mu}^n} \int_0^1 d\sigma_1 \int_0^1 d\sigma_2 \cdots \int_0^1 d\sigma_N \, \mathcal{O}_\beta^{(2l+1-D)}(M)\big|_{m^2\to m^2+b_l}, \tag{7.3}$$

and, in particular, for the $\beta\to\infty$ limit with $\mu\neq 0$

$$\lim_{\beta\to\infty} \tilde{\Gamma}_N^{\beta\mu} = c(4\pi)^{-D/2} \sum_{l\in Z} a_{l0} \int_0^1 d\sigma_1 \int_0^1 d\sigma_2 \cdots \int_0^1 d\sigma_N \, \mathcal{O}_\infty^{(2l+1-D)}(M)\big|_{m^2\to m^2+b_l}. \tag{7.4}$$

The thermodynamic master amplitudes (7.3) and (7.4) are provided with Eqs. (4.16), (5.10), and (6.9), and several explicit ''partial amplitudes'' $\mathcal{O}_\beta^{(k)}$ are also given up to $k=8$.

In the master formulas (7.2) and (7.3), the ''partial'' amplitudes $\tilde{\mathcal{A}}_l^{\beta\mu}$ or $\mathcal{O}_\beta^{(k)}$ (*s* integrals of the pure thermodynamic kinematical factor) are the fundamental computational blocks in our formalism, and in Secs. IV and V, we focused on some mathematical aspects and techniques how to calculate these integrals. We derived various representations and formulas on these integrals for arbitrary values of *N, D*, and ϵ. It is obvious that a variety of equivalent formulas makes it easy to find some relation and consistency with the results obtained by different methods. These should generally be related by analytic continuation, and a different analytic expression is certainly useful by definition; i.e., another formula can cover the region which cannot be reached in an original representation, and it also overcomes some defect of an inconvenient representation. For example, in Ref. 21, a dual rotation is used to obtain an electric gap equation from a magnetic one in a four-fermion model. We expect such a kind of utility when our formulas are applied in more explicit stages.

Although we could have possibly simplified our results furthermore through a certain technique,[22] our formulas were sufficiently convenient to extract the nonzero values of zero temperature limit with μ kept finite. For this purpose, it is necessary to have the integral representation, such as Eqs. (4.10) and (5.10), which clearly indicates a nonanalytic cut in its integrand at $\beta=\infty$. We also had to go beyond the condition $\mu<M$ in these representations; however, this might be justified by analytic continuation, and at least we know that the $k=0$ case coincides with the exact result.

We have examined the model-independent structure of the thermodynamic BK master formula with consulting several simple examples. The model dependence appears in the vertex structure functions, and hence one has to evaluate the vertex structure in the first place for explicit calculations in each model. This task is case by case and will become more lengthy as *N* increases for the cases involving more bosonic *x* fields in the vertex operator v_j parts such as photon, gluon, and pseudo-scalar particle. However, our systematic prescription is certainly promising to obtain the (pure) thermodynamic *N*-point amplitudes in a straightforward way as long as the $\mathcal{V}_{\beta\mu}$ belongs to the general form (3.34). Finally, we should not forget the advantage that the world-line formula encapsulates all necessary Feynman diagrams in a single integrand.

APPENDIX A: INSERTION OF CHEMICAL POTENTIAL

In this appendix, we give a brief explanation of the shift procedure (3.9). First consider the vacuum case $\overline{\omega}=0$. In this case, $M^2=m^2$ and

$$-\mathrm{Tr}\ln(\partial^2+m^2)=\frac{1}{\beta}\sum_n \int \frac{d^Dp}{(2\pi)^{D-1}}\int_0^\infty \frac{ds}{s}\, e^{-s(p^2+m^2)}. \tag{A1}$$

We usually perform the insertion of chemical potential in terms of the shift

$$p^2=p_0^2+\vec{p}^{\,2}\rightarrow(\omega_n+i\mu)^2+\vec{p}^{\,2} \tag{A2}$$

with the internal Matsubara frequencies

$$\omega_n=\frac{2\pi}{\beta}\left(n+\frac{1}{\epsilon}\right), \tag{A3}$$

and we have

$$\sum_n e^{-s(\omega_n+i\mu)^2}=e^{s\overline{\mu}_0^2}\Theta_3\left(\frac{2s\overline{\mu}_0}{\beta},is\frac{4\pi}{\beta^2}\right), \tag{A4}$$

where we have defined

$$\overline{\mu}_0=\mu-\frac{2\pi i}{\epsilon\beta}. \tag{A5}$$

Thus, we prove Eq. (3.14) for $N=0$:

$$\text{Eq. (A1)}=\frac{1}{\beta}\int_0^\infty\frac{ds}{s}\,(4\pi s)^{(1-D)/2}e^{s(\overline{\mu}_0^2-m^2)}\Theta_3\left(\frac{2s\overline{\mu}_0}{\beta},is\frac{4\pi}{\beta^2}\right). \tag{A6}$$

For further nonzero values of N, the proofs are straightforward, and we shall not explain the details anymore. For example, one can find the $N=2$ case in Appendix B.

Instead, we add a supplemental interpretation. Applying the transformation

$$\frac{\sqrt{4\pi s}}{\beta}\,e^{s\overline{\mu}_0^2}\Theta_3\left(\frac{2s\overline{\mu}_0}{\beta},is\frac{4\pi}{\beta^2}\right)=\Theta_3\left(\frac{i\beta\overline{\mu}_0}{2\pi},i\frac{\beta^2}{4\pi s}\right), \tag{A7}$$

we rewrite Eq. (A6) as

$$\text{Eq. (A6)}=\mathcal{A}_0^{\beta\mu}=\int_0^\infty\frac{ds}{s}\,(4\pi s)^{-D/2}e^{-sm^2}\Theta_3\left(i\frac{\beta\overline{\mu}_0}{2\pi},i\frac{\beta^2}{4\pi s}\right). \tag{A8}$$

Here $\overline{\mu}_0$ now appears only in the first argument of the Θ_3, and remember that $\overline{\omega}$ also appears in the same place as $\overline{\mu}_0$ does for $N\neq 0$. Taking account of exponent's additivity in the first argument of Θ_3 [*q.v.* Eq. (3.5)], one can imagine that the final result is given by the replacement of $\overline{\mu}_0$ with $\overline{\mu}_0-i\overline{\omega}(=\overline{\mu})$. This result coincides with the $\mathcal{K}_{\beta\mu}$ obtained by the shift manipulation (3.9) in $\mathcal{K}_\beta$.

APPENDIX B: PHOTON SELF-ENERGY PART FROM FEYNMAN RULE

In this appendix, we rearrange the $N=2$ Feynman amplitude of photon scattering into the world-line formula at finite β and μ. The photon self-energy part (in $D=4$) by the Feynman diagram technique is written in the form

$$\Pi_\beta = -\frac{1}{\beta}\sum_n \int \frac{d^3p}{(2\pi)^3}\frac{\mathrm{Tr}[\not\epsilon_1(\not p - \not k_1)\not\epsilon_2 \not p]}{p^2(p-k)^2}, \tag{B1}$$

where $p^\mu = (\omega_n, \vec{p})$, $k_i^\mu = (\omega_{k_i}, \vec{k}_i)$, and $\epsilon_i \equiv \epsilon_\mu^i$, $i=1,2$. We also make use of $k^\mu = k_1^\mu = -k_2^\mu$. The QED coupling e is set to be the unity for simplicity. For convenience, we decompose the Π_β into the following three parts:

$$\Pi_i = \frac{1}{\beta}\sum_n \int \frac{d^3p}{(2\pi)^3} W_i, \tag{B2}$$

in terms of

$$-\mathrm{Tr}\,[\not\epsilon_1(\not p - \not k)\not\epsilon_2 \not p] = W_1 + W_2 + W_3, \tag{B3}$$

where

$$W_1 = 2\epsilon_1\cdot\epsilon_2(p^2 + (p-k)^2), \tag{B4}$$

$$W_2 = 2(\epsilon_1\cdot\epsilon_2 k_1\cdot k_2 - \epsilon_1\cdot k_2 \epsilon_2\cdot k_1), \tag{B5}$$

$$W_3 = -2\epsilon_1\cdot(2p - k_1)\epsilon_2\cdot(2p-k_1). \tag{B6}$$

Applying the Feynman integral formula

$$\frac{1}{p^2(p-k)^2} = \int_0^\infty ds s \int_0^1 du e^{-sk^2(u-u^2)} e^{-s(p-(1-u)k)^2}, \tag{B7}$$

we easily rewrite Π_1 and Π_2 in the following forms:

$$\Pi_1 = \frac{4\epsilon_1\cdot\epsilon_2}{\beta}\sum_n \int_0^\infty \frac{ds s}{(4\pi s)^{3/2}}\int_0^1 du e^{-sk^2(u-u^2)} e^{-s(\omega_n + a)^2}\frac{1}{s}\,\delta(1-u), \tag{B8}$$

$$\Pi_2 = (\epsilon_1\cdot\epsilon_2 k_1\cdot k_2 - \epsilon_1\cdot k_2\epsilon_2\cdot k_1)\frac{2}{\beta}\sum_n \int_0^\infty \frac{ds s}{(4\pi s)^{3/2}}\int_0^1 du e^{-sk^2(u-u^2)} e^{-s(\omega_n+a)^2}, \tag{B9}$$

where

$$a = (1-u)\omega_k. \tag{B10}$$

The Π_3 can be rewritten in the similar way by using Eq. (B7) and shifting $\vec{p} \to \vec{p} + (1-u)\vec{k}$ in the p-integral. After some algebra, we have

$$\Pi_3 = \frac{2}{\beta}\sum \int_0^\infty \frac{ds}{(4\pi s)^{3/2}}\, s\int_0^1 du e^{-sk^2(u-u^2)} e^{-s(\omega_n+a)^2}$$
$$\times\left[\frac{2}{s}\,(\epsilon_0^1\epsilon_0^2 - \epsilon_1\cdot\epsilon_2) - \{2\epsilon_0^1(\omega_n + a) - (1-2u)\epsilon_1\cdot k\}\{2\epsilon_0^2(\omega_n+a) - (1-2u)\epsilon_2\cdot k\}\right]. \tag{B11}$$

Gathering Eqs. (B8), (B9), and (B11), we obtain

$$\Pi_\beta=\frac{2}{\beta}\sum_n \int_0^\infty \frac{dss}{(4\pi s)^{3/2}}\int_0^1 due^{-sk^2(u-u^2)}e^{-s(\omega_n+a)^2}\Bigg[\frac{2}{s}\,\epsilon_1\cdot\epsilon_2\{\delta(1-u)-1\}$$
$$+(\epsilon_1\cdot\epsilon_2 k_1\cdot k_2-\epsilon_1\cdot k_2\epsilon_2\cdot k_1)-(1-2u)^2\epsilon_1\cdot k\epsilon_2\cdot k-4\epsilon_0^1\epsilon_0^2\left\{(\omega_n+a)^2-\frac{1}{2s}\right\}$$
$$+2(1-2u)(\epsilon_0^1\epsilon_2\cdot k+\epsilon_0^2\epsilon_1\cdot k)(\omega_n+a)\Bigg]. \tag{B12}$$

Now, we can eliminate the forms $(\omega_n+a)^m, m=1,2$, in the summand of Eq. (B12) due to the following operations:

$$-\frac{1}{2s}\,\partial_a\sum_n e^{-s(\omega_n+a)^2}=\sum_n (\omega_n+a)e^{-s(\omega_n+a)^2}, \tag{B13}$$

$$\left(\frac{1}{2s}\,\partial_a\right)^2\sum_n e^{-s(\omega_n+a)^2}=\sum_n \left[(\omega_n+a)^2-\frac{1}{2s}\right]e^{-s(\omega_n+a)^2}. \tag{B14}$$

The a dependence of the summand exponential can be transformed into a linear exponent form by using the Jacobi transformation:

$$\sum_n e^{-s(\omega_n+a)^2}=e^{-sa^2}\Theta_2\left(\frac{2sia}{\beta},\frac{is4\pi}{\beta^2}\right)$$
$$=\frac{\beta}{\sqrt{4\pi s}}\,\Theta_4\left(\frac{\beta a}{2\pi},\frac{i\beta^2}{4\pi s}\right)=\frac{\beta}{\sqrt{4\pi s}}\sum_n (-1)^n e^{-n^2\beta^2/4s}e^{in\beta a}. \tag{B15}$$

Then we are allowed to perform the following replacement in the transformed summands,

$$\frac{\partial}{\partial a}\to in\beta, \tag{B16}$$

and we finally derive the formula

$$\Pi_\beta=2\int_0^\infty\frac{dss}{(4\pi s)^{4/2}}\int_0^1 due^{-sk^2(u-u^2)}\sum_n (-1)^n e^{-n^2\beta^2/4s}e^{in\beta a}\Bigg[\epsilon_0^1\epsilon_0^2\frac{n^2\beta^2}{s^2}+\frac{in\beta}{s}(2u-1)$$
$$\times(\epsilon_0^1\epsilon_2\cdot k_1-\epsilon_0^2\epsilon_1\cdot k_2)+\frac{2}{s}\,\epsilon_1\cdot\epsilon_2\{\delta(1-u)-1\}+(\epsilon_1\cdot\epsilon_2 k_1\cdot k_2-\epsilon_1\cdot k_2\epsilon_2\cdot k_1)$$
$$+(1-2u)^2\epsilon_1\cdot k_2\epsilon_2\cdot k_1\Bigg]. \tag{B17}$$

For $\mu\neq 0$, we have only to make the standard shift (1.1), and this causes the following change of the a defined in Eq. (B10):

$$a\to(1-u)\omega_k+i\mu. \tag{B18}$$

This modification exactly corresponds to the shift (3.9). Note that the a coincides with $\overline{\omega}$ with fixing $\sigma_1=1$.

APPENDIX C: WORLD-LINE METHOD FOR PHOTON SELF-ENERGY PART

In this appendix, we illustrate how we obtain the thermodynamic version of the vertex structure function in the case of $N=2$ photon scattering. In the world-line formalism, the photon self-energy part at zero temperature can be obtained from the formula

$$\Gamma_2=-\frac{1}{2}\,2^{D/2}\int_0^\infty \frac{ds}{s}\,e^{-sm^2}\oint \mathcal{D}x\mathcal{D}\psi\,\exp\left[-\int_0^s\left(\frac{1}{4}\,\dot{x}(\tau)+\frac{1}{2}\,\psi\dot{\psi}\right)d\tau\right]V_1V_2, \tag{C1}$$

where $V_j\,,j=1,2$, are the photon vertex operators

$$V_j=-ie\int_0^s d\tau_j\,(\epsilon_j\cdot\dot{x}+2i\psi\cdot\epsilon_j\psi\cdot k_j)(\tau_j)e^{ik_jx(\tau_j)}, \tag{C2}$$

and $\dot{x}=\partial_\tau x$, etc. In the following, we set the QED coupling $e=1$ as assumed in Appendix B. Here we follow the standard world-line notation τ, which is related to the main text notation σ by

$$\tau_j=s\sigma_j\,. \tag{C3}$$

Equation (C1) is known to become

$$\Gamma_2=-\frac{1}{2}\,2^{D/2}\int_0^\infty \frac{ds}{s}\int_0^s d\tau_1\int_0^s d\tau_2\mathcal{V}\times\mathcal{K} \tag{C4}$$

with the vertex structure function

$$\mathcal{V}=\epsilon_1\cdot\epsilon_2\ddot{G}_B^{12}+\epsilon_1\cdot k_2\epsilon_2\cdot k_1(\dot{G}_B^{12})^2+(\epsilon_1\cdot\epsilon_2k_1\cdot k_2-\epsilon_1\cdot k_2\epsilon_2\cdot k_1)(G_F^{12})^2, \tag{C5}$$

and the kinematical factor

$$\mathcal{K}=\mathcal{N}\langle e^{ik_1\cdot x(\tau_1)}e^{ik_2\cdot x(\tau_2)}\rangle=e^{-sm^2}\oint\oint\mathcal{D}x\exp\left[-\int_0^s\frac{1}{4}\,\dot{x}^2d\tau\right]\prod_{j=1}^{2}\,e^{ik_j\cdot x(\tau_j)}$$
$$=(4\pi s)^{-D/2}e^{-sm^2}e^{k_1\cdot k_2G_B^{12}}, \tag{C6}$$

where

$$G_B^{12}=|\tau_1-\tau_2|-\frac{(\tau_1-\tau_2)^2}{s},\quad G_F^{12}=\mathrm{sign}\,(\tau_1-\tau_2). \tag{C7}$$

The path integral normalizations are chosen to be

$$\oint\mathcal{D}x\exp\left[-\int_0^s\frac{1}{4}\,\dot{x}^2\,d\tau\right]=(4\pi s)^{-D/2},\quad \oint\mathcal{D}\psi\exp\left[-\int_0^s\psi\cdot\dot{\psi}\,dr\right]=1. \tag{C8}$$

It can be said that the kinematical factor is defined by the insertions of the ϕ^3 scalar vertex operators

$$V_j=\int_0^s d\tau_j\exp[ik_j\cdot x(\tau_j)],\quad j=1,2. \tag{C9}$$

According to the program presented in Sec. III, we are led to calculate the following path integral at finite temperature:

$$\Gamma_2^\beta = -\frac{1}{2}\,2^{D/2}(ie)^2\int_0^\infty \frac{ds}{s}\,e^{-sm^2}\oint \mathcal{D}x\mathcal{D}\psi \exp\left[-\int_0^s\left(\frac{1}{4}\,\dot{x}^2+\frac{1}{2}\,\psi\cdot\dot{\psi}\right)d\tau\right]$$
$$\times\left(\prod_{i=1}^2\int_0^s d\tau_i e^{ik_i x(\tau_i)}\right)\sum_n (-1)^n e^{-n^2\beta^2/4s}e^{in(\beta/s)(\tau_1-\tau_2)\omega_k}$$
$$\times\left(\epsilon_0^1\frac{n\beta}{s}+\epsilon^1\cdot\dot{x}+2i\psi\cdot\epsilon_1\psi\cdot k_1\right)(\tau_1)\left(\epsilon_0^2\frac{n\beta}{s}+\epsilon^2\cdot\dot{x}+2i\psi\cdot\epsilon_2\psi\cdot k_2\right)(\tau_2). \tag{C10}$$

Using the Wick contraction method with the correlators

$$\langle x^\mu(\tau_1)x^\nu(\tau_2)\rangle = -g^{\mu\nu}G_B^{12}, \quad \langle \psi^\mu(\tau_1)\psi^\nu(\tau_2)\rangle = \tfrac{1}{2}g^{\mu\nu}G_F^{12}, \tag{C11}$$

one can verify the coincidence of the Γ_2^β with the Π_β derived in Appendix B; we have arrived at the form

$$\Pi_\beta = \frac{1}{2}\,2^{D/2}\int_0^\infty \frac{ds}{s}\,e^{-sm^2}\left(\prod_{j=1}^2\int_0^s d\tau_j\right)\sum_n (-1)^n\mathcal{K}_\beta^{(n)}(\mathcal{V}+\tilde{\mathcal{V}}), \tag{C12}$$

where

$$\tilde{\mathcal{V}} = \epsilon_0^1\epsilon_0^2\frac{n^2\beta^2}{S^2}+\frac{in\beta}{s}\,(\epsilon_0^1\epsilon_2\cdot k_1-\epsilon_0^2\epsilon_1\cdot k_2)\dot{G}_B^{12}, \tag{C13}$$

and the $\mathcal{K}_\beta^{(n)}$ is the nth mode of the bosonic two point function at finite β defined by

$$\mathcal{K}_\beta^{(n)} = \oint \mathcal{D}x \exp\left[-\int_0^s (1/4)\dot{x}^2\right]\prod_{j=1}^2 e^{ik_j\cdot x(\tau_j)}\big|_{x^0\to x^0+n\beta\tau/s}$$
$$= e^{-n^2\beta^2/4s}e^{in\beta(\tau_1-\tau_2)\omega_k/s}(4\pi s)^{-D/2}e^{-k^2G_B^{12}}. \tag{C14}$$

Here Eq. (C12) with (C13) and (C14) reproduces the Feynman rule result (B17) with rescaling $\tau_i = s\sigma_i$ and fixing $\sigma_1 = 1$ with $\sigma_2 = u$. Note that $\dot{G}_B^{12}$ behaves as $2u-1$ in this respect.

Let us further rewrite the above $\tilde{\mathcal{V}}$ in an operator form suitable to Eq. (3.31). In the following, we shall not employ the fixing $\sigma_1 = 1$. As shown in Appendix B, we can replace $in\beta$ with $\partial/\partial\bar{\omega}$, where

$$\bar{\omega} = (\sigma_1-\sigma_2)\omega_k, \tag{C15}$$

and we therefore find

$$\Gamma_2^\beta = \frac{1}{2}\,2^{D/2}\int_0^\infty ds s\left(\prod_{i=1}^2\int_0^1 d\sigma_i\right)\mathcal{V}_\beta\times\mathcal{K}_\beta \tag{C16}$$

with having

$$\mathcal{V}_\beta = \mathcal{V}-\epsilon_0^1\epsilon_0^2\left(\frac{1}{s}\,\frac{\partial}{\partial\bar{\omega}}\right)^2-\frac{1}{s}\,\frac{\partial}{\partial\bar{\omega}}(\epsilon_0^1\epsilon_2\cdot k_1-\epsilon_0^2\epsilon_1\cdot k_2)\dot{G}_B^{12} \tag{C17}$$

and

$$\mathcal{K}_\beta = \sum_n (-1)^n\mathcal{K}_\beta^{(n)}e^{-sm^2}, \tag{C18}$$

where the change of variables $\tau_i = s\sigma_i$ in G_B^{12} should be understood. Since we already justified the shift (3.9) in the end of Appendix B, we can use the following parts for the finite μ case:

$$\mathcal{V}_{\beta_\mu} = \mathcal{V}_\beta\left[\frac{\partial}{\partial\overline{\omega}} \to \frac{\partial}{\partial\Omega}\right], \quad \mathcal{K}_{\beta_\mu} = \mathcal{K}_\beta[\,\overline{\omega} \to \Omega\,]. \tag{C19}$$

Substituting these in Eq. (C16), we obtain the two-point function $\Gamma_2^{\beta\mu}$, which therefore coincides with the corresponding Feynman rule result suggested at the end of Appendix B.

[1] Z. Bern and D. A. Kosower, Nucl. Phys. B **379**, 451 (1992).
[2] Z. Bern, L. Dixon, and D. A. Kosower, Annu. Rev. Nucl. Part. Sci. **46**, 109 (1996); Z. Bern, in *Theoretical Advanced Study Institute in High Energy Physics (TASI 92)*, edited by J. Harvey and J. Polchinski (World Scientific, Singapore, 1993).
[3] Z. Bern, L. Dixon, and D. A. Kosower, Phys. Rev. Lett. **70**, 2677 (1993).
[4] Z. Bern, D. C. Dunbar, and T. Shimada, Phys. Lett. B **312**, 277 (1993); D. C. Dunbar and P. S. Norridge, Nucl. Phys. B **433**, 181 (1995).
[5] Z. Bern, L. Dixon, D. C. Dunbar, and D. A. Kosower, Phys. Lett. B **394**, 105 (1997); Z. Bern, J. S. Rozowsky, and B. Yan, *ibid.* **401**, 273 (1997).
[6] P. Di Vecchia, A. Lerda, L. Magnea, and R. Marotta, Phys. Lett. B **351**, 445 (1995); P. Di Vecchia, A. Lerda, L. Magnea, R. Marotta, and R. Russo, Nucl. Phys. B **469**, 235 (1996).
[7] K. Roland, Phys. Lett. B **289**, 148 (1992); SISSA/ISAS 131-93-EP.
[8] K. Roland and H.-T. Sato, Nucl. Phys. B **480**, 99 (1996); **515**, 488 (1998).
[9] P. Di Vecchia, A. Lerda, L. Magnea, R. Marotta, and R. Russo, Phys. Lett. B **388**, 65 (1996).
[10] M. G. Schmidt and C. Schubert, Phys. Lett. B **331**, 69 (1994).
[11] H.-T. Sato, Phys. Lett. B **371**, 270 (1996).
[12] M. G. Schmidt and C. Schubert, Phys. Rev. D **53**, 2150 (1996).
[13] K. Daikouji, M. Shino, and Y. Sumino, Phys. Rev. D **53**, 4598 (1996).
[14] M. Mondragón, L. Nellen, M. G. Schmidt, and C. Schubert, Phys. Lett. B **351**, 200 (1995); **366**, 212 (1996); E. D'Hoker and D. G. Gagné, Nucl. Phys. B **467**, 272 (1996); **467**, 297 (1996).
[15] D. G. C. McKeon and C. Schubert, Phys. Lett. B **440**, 101 (1998).
[16] M. J. Strassler, Nucl. Phys. B **385**, 145 (1992).
[17] M. Reuter, M. G. Schmidt, and C. Schubert, Ann. Phys. (N.Y.) **259**, 313 (1997).
[18] A. M. Polyakov, *Gauge Fields and Strings* (Harwood, 1987).
[19] R. R. Metsaev and A. A. Tseytlin, Nucl. Phys. B **298**, 109 (1988); O. D. Andreev and A. A. Tseytlin, Phys. Lett. B **207**, 157 (1988).
[20] C. Schubert, Acta Phys. Pol. B **27**, 3965 (1996); PRINT-97-273.
[21] S. Kanemura, H.-T. Sato, and H. Tochimura, Nucl. Phys. B **517**, 567 (1998); H.-T. Sato, J. Math. Phys. **39**, 4540 (1998).
[22] D. G. C. McKeon and A. K. Rebhan, Phys. Rev. D **47**, 5487 (1993).
[23] J. Boorstein and D. Kutasov, Phys. Rev. D **51**, 7111 (1995).
[24] M. Haack and M. G. Schmidt, Eur. Phys. J. C **7**, 149 (1999).
[25] I. A. Shovkovy, Phys. Lett. B **441**, 313 (1998).
[26] M. G. Schmidt and H.-T. Sato, Nucl. Phys. B **524**, 742 (1998).
[27] J. O. Andersen and T. Haugset, Phys. Rev. D **51**, 3073 (1995); D. Cangemi and G. Dunne, Ann. Phys. (N.Y.) **249**, 582 (1996).

A field theory approach to Lindstedt series for hyperbolic tori in three time scales problems

G. Gallavotti[a)]
Università di Roma ''La Sapienza,'' Dipartimento di Fisica, Roma I-00185, Italy

G. Gentile[b)]
Università di Roma Tre, Dipartimento di Matematica, Roma I-00146, Italy

V. Mastropietro[c)]
Università di Roma ''Tor Vergata,'' Dipartimento di Matematica, Roma I-00139, Italy

(Received 25 March 1999; accepted for publication 8 April 1999)

Interacting systems consisting of two rotators and a pendulum are considered, in a case in which the uncoupled systems have three very different characteristic time scales. The abundance of unstable quasiperiodic motions in phase space is studied via the Lindstedt series regarded as a sum of Feynman graphs and studied with renormalization group techniques based on Eliasson's work on KAM tori. The result is a strong improvement, compared to our previous results, on the domain of validity of bounds that imply existence of invariant tori, large homoclinic angles, long heteroclinic chains, and drift-diffusion in phase space. © *1999 American Institute of Physics.* [S0022-2488(99)03310-1]

I. THE HAMILTONIAN SYSTEM

1.1. Let $(\varphi,\alpha_1,\alpha_2)=(\varphi,\boldsymbol{\alpha})\in\mathbb{T}^3$ be three angles (i.e., positions on circles); let $(I,A_1,A_2)=(I,\mathbf{A})\in\mathbb{R}^3$ be their conjugate momenta (or ''actions''). We consider the Hamiltonian function, depending on two parameters $\varepsilon,\eta>0$, defined by

$$\mathcal{H}=\eta^{1/2}\Omega_1A_1+\eta\frac{A_1^2}{2J}+\eta^{-1/2}\Omega_2A_2+\frac{I^2}{2J_0}+J_0g_0^2(\cos\varphi-1)+\varepsilon f(\varphi,\alpha_1,\alpha_2), \tag{1.1}$$

with f an *even* trigonometric polynomial of degree N, N_0 in $\boldsymbol{\alpha}$, φ, respectively; $\Omega_1,\Omega_2,J,J_0,g_0$ are positive constants.

This system describes two rotators (one anisochronous, labeled #1, and one isochronous, labeled #2) interacting with a pendulum which has its free (i.e., with $\varepsilon=0$) unstable equilibrium position at $I=0$, $\varphi=0$ and the stable one at $I=0$, $\varphi=\pi$. The scale of frequency of the pendulum is $O(1)$ in η; at the same time the two rotators rotate at constant speed $O(\eta^{1/2})$ and $O(\eta^{-1/2})$, respectively. Hence the system has three time scales; we assume $\eta<1$ so that the *slow* rotator is the #1 rotator.

The free motion admits invariant tori of dimension 2 (namely parameterized by $\mathbf{A}$, a constant vector, by $\boldsymbol{\alpha}$ arbitrary, and with $I=0$, $\varphi=0$) which are unstable and possess stable (labeled $+$) and unstable (labeled $-$) three-dimensional manifolds (parameterized by $\mathbf{A}$, the same constant vector, by $\boldsymbol{\alpha}$, φ arbitrary, and with $I=\pm J_0g_0\sqrt{2(1-\cos\varphi)}$).

We shall study properties that eventually hold when $\eta\to0$. It is well known (see Refs. 1 and 2, for instance) that if ε is small most of the unperturbed tori and their manifolds still exist, just a little deformed. This means that (under the condition stated below) there exist functions $\mathbf{U}^{\pm}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha})$

[a)]Electronic mail: giovanni@ipparco.roma1.infn.it
[b)]Electronic mail: gentile@ipparco.roma1.infn.it
[c)]Electronic mail: vieri@ipparco.roma1.infn.it

and $V^{\pm}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha})$ which are divisible by ε and analytic in $\boldsymbol{\alpha}$, φ, ε, for $\boldsymbol{\alpha}\in\mathbb{T}^2$, $|\varphi|<2\pi$, $|\varepsilon|<\varepsilon_0$, with ε_0 small enough, such that an initial datum starting on the (three-dimensional) surfaces W^{σ}_{ε}, $\sigma=\pm$, defined as

$$\mathbf{A}^{\sigma}(\varphi,\boldsymbol{\alpha})=\mathbf{A}'+\mathbf{U}^{\sigma}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha}),\quad I^{\sigma}(\varphi,\boldsymbol{\alpha})=\pm J_0 g_0\sqrt{2(1-\cos\varphi)}+V^{\sigma}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha}),\tag{1.2}$$

evolves, when the time $t\to\pm\infty$, tending to be confused with a quasiperiodic motion on a invariant torus $\mathcal{T}(\mathbf{A}')$, with rotation vector

$$\boldsymbol{\omega}'=(\omega_1',\omega_2'),\quad \omega_1'\overset{\text{def}}{=}\eta^{1/2}\Omega_1+\eta J^{-1}A_1',\quad \omega_2'\overset{\text{def}}{=}\eta^{-1/2}\Omega_2,\tag{1.3}$$

and furthermore such asymptotic motion takes place with $\mathbf{A}$ moving quasiperiodically *with average* $\mathbf{A}'$.

All this holds if $\boldsymbol{\omega}'$ *verifies the Diophantine condition*,

$$|\boldsymbol{\omega}'\cdot\boldsymbol{\nu}|>C|\boldsymbol{\nu}|^{-\tau},\quad \forall\boldsymbol{\nu}\in\mathbb{Z}^2\backslash\{\mathbf{0}\},\tag{1.4}$$

for $C,\tau>0$ (which may depend also on η). The values of ε for which we shall be able to prove the above will be so small that the part of the stable and unstable manifolds with $|\varphi|<\frac{3}{2}\pi$ *can be represented as a graph of* $\mathbf{A}$, I *over* $\boldsymbol{\alpha}$, φ. (Note that if $\varepsilon=0$ they are graphs over $\boldsymbol{\alpha}$, φ for $|\varphi|$ smaller than *any* prefixed quantity $<2\pi$.) Hence we look, since the beginning, for invariant tori which have the latter property.

The approach to the invariant tori, of the points that lie on their stable manifolds, will be exponential in the sense that their distances $d(t)$ to the tori will be such that

$$\lim_{t\to+\infty} t^{-1}\log d(t)^{-1}=\overline{g_0},\quad \overline{g_0}\equiv\overline{g_0}(\varepsilon)\overset{\text{def}}{=}(1+\Gamma(\varepsilon,g_0))g_0,\tag{1.5}$$

for a suitable analytic function $\Gamma(\varepsilon,g_0)$, divisible by ε. We shall call $\overline{g_0}$ the *Lyapunov exponent* of the torus (it will depend on ε as well as on the considered torus, i.e., on $\boldsymbol{\omega}'$ and on η). The exponent relative to the approach to the same torus along its unstable manifold (as $t\to-\infty$) will be the same, by time reversal symmetry defined below.

We fix throughout the paper τ ($\tau\geq 1$) and *we shall mainly study the dependence of* ε_0, *i.e., our estimate for the analyticity radius, as a function of* η: $\varepsilon_0=\varepsilon_0(C,\eta)$.

1.2. *Remark:* The relation (1.3) between the value of the average action and the rotation vector is nontrivial and it has been named in Refs. 3 and 4 (where it was pointed out) by saying that the tori of (1.1) are ''torsion free'' or ''twistless.'' It is a remarkable symmetry property of (1.1), see Refs. 3, 5, 6.

1.3. If $\varepsilon=0$ the stable and unstable manifolds coincide (because the pendulum separatrix is degenerate); it is a degeneracy that is lost when $\varepsilon\neq 0$ and generically the manifolds will have only pairwise isolated trajectories in common, called *homoclinic trajectories*.

Nevertheless time reversal symmetry and parity symmetry (the latter symmetry is due to the assumption of evenness of f) hold for (1.1). If S^t denotes the time evolution and the involution map i (composition of parity and time reversal) is defined by $i(\varphi,\boldsymbol{\alpha},I,\mathbf{A})=(2\pi-\varphi,-\boldsymbol{\alpha},I,\mathbf{A})$, then $iS^t=S^{-t}i$ and there are relations between the stable and unstable manifolds that are preserved even for $\varepsilon\neq 0$. Namely,

$$\begin{aligned}\mathbf{U}^{+}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha})&=\mathbf{U}^{-}_{\mathbf{A}'}(2\pi-\varphi,-\boldsymbol{\alpha}),\\ V^{+}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha})&=V^{-}_{\mathbf{A}'}(2\pi-\varphi,-\boldsymbol{\alpha}),\end{aligned}\tag{1.6}$$

where care must be exercised because the manifolds contain *two* points over each $\boldsymbol{\alpha}$, φ. (This is in fact already so for $\varepsilon=0$.) Hence if $\varphi\simeq\pi$ the relations in (1.6) concern points that lie on different connected manifolds; to understand what happens one should try a drawing taking into account that the above representations are considered only for $|\varphi|<\frac{3}{2}\pi$.

Looking at the manifolds at $\varphi=\pi$, *assuming their existence and that they are graphs above* $\boldsymbol{\alpha}$, φ for $|\varphi|<\frac{3}{2}\pi$, Eqs. (1.6) imply that, fixed $\mathbf{A}'$,

$$\mathbf{Q}(\alpha)\stackrel{\text{def}}{=}\mathbf{U}^{+}_{\mathbf{A}'}(\pi,\boldsymbol{\alpha})-\mathbf{U}^{-}_{\mathbf{A}'}(\pi,\boldsymbol{\alpha})=-\mathbf{Q}(-\boldsymbol{\alpha}),\tag{1.7}$$

so that $\mathbf{Q}(\mathbf{0})=\mathbf{0}$; but, in general, $\mathbf{Q}(\boldsymbol{\alpha})\neq\mathbf{0}$ for $\boldsymbol{\alpha}\neq\mathbf{0}$.

The function $\mathbf{Q}(\boldsymbol{\alpha})$ is called the *homoclinic splitting* (or simply *splitting*) *vector* at $\varphi=\pi$, and the determinant of the matrix with entries $\partial_{\alpha_i}Q_j(\mathbf{0})$ (splitting matrix) is called the *splitting*. One can more generally consider $\underline{Z}\equiv\underline{Z}(\varphi,\boldsymbol{\alpha})=(\mathbf{U}^{+}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha})-\mathbf{U}^{-}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha}),V^{+}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha})-V^{-}_{\mathbf{A}'}(\varphi,\boldsymbol{\alpha}))$ which would be the splitting vector at φ. Here and henceforth the vectors in $\mathbb{R}^l$ will be denoted with an underlined letter (while the boldface is used for vectors in $\mathbb{R}^{l-1}$); so far $l=3$, but shortly we shall consider $l\geqslant 3$. The function $\underline{Z}$ can be written as the gradient of a generating function Φ, i.e., $\underline{Z}=(\partial_\varphi\Phi,\partial_\alpha\Phi)$. This is a result due to Eliasson who points out that it follows immediately from the Lagrangian nature of the stable and unstable manifolds. It is a further symmetry property. (It can alternatively be easily seen from the explicit expressions for the stable and unstable manifolds equations derived in Ref. 3, which also provide a general algorithm for constructing the function Φ as a convergent series in ε for ε small; see Ref. 7.)

The symmetry of (1.1) (hence the consequent oddness of $\mathbf{Q}(\boldsymbol{\alpha})$) implies that there is one trajectory which swings through $\varphi=\pi$ when $\boldsymbol{\alpha}$ is exactly $\mathbf{0}$: it tends to the same invariant torus as $t\to\pm\infty$, provided the torus exists and its stable and unstable manifolds are graphs over $\boldsymbol{\alpha}$, φ over an interval of φ greater than $|\varphi|<\pi$.

In this paper we prove the following results:

1.4. **Theorem:** *Given the Hamiltonian (1.1), given constants s,* $\Omega>0$ *and given* $\eta>0$ *small enough, the following assertions hold:*

(1) *There are invariant tori with rotation vectors* $\boldsymbol{\omega}'$ *for all* $\boldsymbol{\omega}'$ *verifying the Diophantine condition (1.4) with constant* $C=C(\eta)=\Omega e^{-s\eta^{-1/2}}$ *and* $|\omega_1'|\in[\frac{1}{2}\Omega_1\eta^{1/2},2\Omega_1\eta^{1/2}]$.
(2) *Such tori exist for* $|\varepsilon|<\varepsilon_0=O(\eta^2)$ *and for* η *small enough.*
(3) *They can be parameterized by their average actions* $\mathbf{A}'$; *the angular velocity is then given by the rotation vector* $\boldsymbol{\omega}'\equiv(\Omega_1\eta^{1/2}+\eta J_1^{-1}A_1',\Omega_2\eta^{-1/2})$ *(i.e., they are "twistless") and the Lyapunov exponents have the form* $\overline{g_0}=(1+\Gamma(\varepsilon,g_0))g_0$, *with* $\Gamma(\varepsilon,g_0)$ *analytic in* ε *and divisible by* ε.
(4) *The parametric equations for such tori and for their stable and unstable manifolds ("whiskers") can be computed by a convergent perturbation series in powers of* ε *around the unperturbed tori with the same rotation vector and their corresponding stable and unstable manifolds*.
(5) *At the homoclinic intersection with* $\varphi=\pi$ *(existing by symmetry), between the stable manifold and the unstable manifold of each torus, the splitting is generically given by the Mel'nikov integral which is of order* $O(\varepsilon^2\eta^{-\beta}e^{-(\pi/2)g_0^{-1}\eta^{-(1/2)}})$, *for* ε *small enough, with* β *depending on the degree* N_0 *in* φ *of the perturbation f; one can take* $\beta=2N_0-1$ *and the asymptotic formula holds if* $|\varepsilon|<\eta^\zeta$, $\zeta=2(N_0+3)$ *and* η *is small enough*.

1.5. *Remark:* The novelty of the theorem is the "sharp" bound $\varepsilon_0=O(\eta^2)$. If we "only" require $\varepsilon_0=O(\eta^{(9/2)+})$ where $\frac{9}{2}+$ is any prefixed positive number $>\frac{9}{2}$ the result is proved in Ref. 6 (see also Refs. 2 or 5). The improvement is made possible by the *totally different technique* used (with respect to Ref. 6); a technique that has interest in its own right and, we think, beyond the result itself. In fact the proof of the last assertion of the theorem is the content of Ref. 8, and the values of the constants β and ζ are taken from Appendix A2 of Ref. 8.

1.6. Theorem 1.4 will be proved by a further extension of Eliasson's method,[9,4,10,5] for the KAM theorem. The following discussion will show the correctness of the intuition that ''new'' small divisors appear in the perturbation expansion *at orders spaced by* $O(\eta^{-1})$. So that the coupling constant is effectively $O(\varepsilon^{\eta^{-1}})$ and the analyticity condition is expected to be $\varepsilon^{\eta^{-1}}C(\eta)^{-q}$ small (for some $q>0$, determined as in the discussion in Remark 5.16, item (4), below). Hence the analyticity condition will be $\varepsilon C(\eta)^{-q\eta}$ small rather than the far worse $\varepsilon C(\eta)^{-q}$ small, that is implied directly from lemma 1 in Ref. 2 (where $q=6$ is an estimate).

In the one degree of freedom case the corresponding problem is studied in Ref. 11. It is a problem that arises naturally in the context of Nekhoroshev theory. In our case the rotation vector is not one-dimensional, so that the cancellations between resonances typical of small divisors problems,[9,3,10,5] have to be exploited in order to prove convergence of the perturbative series. The fact that the two components of the rotation vector (1.3) are so different in scale has the consequence that small divisors can appear only at high orders, so that the dependence of the radius convergence on the Diophantine constant $C(\eta)$ is highly improvable with respect the ''naïve'' one, as explained above; the proof of such an assertion is the subject of the present paper (as, in the weaker form, already of Ref. 6).

1.7. The paper is organized as follows. In Secs. II, III, IV the formalism is concisely illustrated and the graphic representations of the whiskers in terms of tree graphs is exhibited (for systems more general than (1.1); see (2.1) below). The analysis is brief but self-contained, with references to Refs. 3, 5 only given for further insight and details. The basic formalism is in Sec. III, then we work out in Sec. III two specific examples to explain the origin of the graphical interpretation, and in Sec. IV we set up the general Feynman rules for evaluating the equations of the whiskers (and the splitting vector as a particular case) as a sum of quantities that can be elementarily evaluated. In Sec. V bounds are derived, assuring the convergence of the perturbative series defining the whiskers in the more general system in (2.1) below and leading to Theorem 1.4, when restricted to the Hamiltonian (1.1).

The bounds are derived along the lines of Refs. 3, 5; the main part is the derivation of the bounds for the part of the expansion corresponding to what we call the contributions due to ''trees without leaves;'' this is done fully and self consistently in Sec. V and in the related appendices. Once the bounds on the contributions from trees without leaves are established, *which is the real difficulty*, the same analysis can be applied to bound the other contributions. Since this is simply reduced, without any further technical problems, to the case of contributions from the simpler trees with no leaves we do not repeat this part of the discussion which is done in Ref. 5 following the corresponding analysis done in Refs. 3, 10.

In Appendix A we relegate some technical details, while Appendix C concerns the cancellation analysis of Ref. 5, needed in order to treat the small divisors problems, with more details, with respect to the quoted paper. An original technical part is also in Appendix B and deals with the improvement of the dependence of the convergence radius on the Diophantine constant $C(\eta)$.

We do not comment here on the obvious relevance of the above results for the theory of Arnold diffusion, see Refs. 6 and 12.

II. LINDSTEDT SERIES FOR WHISKERED TORI

We use the formalism of Ref. 5. It would be pointless to repeat here the technical work required to motivate the necessity or usefulness of the notations, and we cannot imagine that any reader may have interest in the matter that follows unless he has some experience with Eliasson's method, as exposed in Ref. 9 and complemented in Refs. 3, 4, 10, 5. The references to Refs. 3, 5 are given only to point at places where further details on the motivations of the assertions can be found.

The following analysis innovates Ref. 5 in Sec. V because of the extension of Siegel–Bryuno's bound described in Appendix B below; this section and the next two provide a *self-contained* description of the graphical algorithm exploited in Sec. V and Appendix B.

2.1. In the following we shall consider a Hamiltonian (''Thirring model'') more general than the one in (1.1), i.e., a Hamiltonian which couples a pendulum with $l-1$ rotators via a perturbation f_1 which is always an *even trigonometric polynomial,*

$$\mathcal{H}=\boldsymbol{\omega}\cdot\mathbf{A}+\frac{1}{2J}\mathbf{A}\cdot\mathbf{A}+\frac{I^2}{2J_0}+J_0g_0^2f_0(\varphi)+\varepsilon J_0g_0^2f_1(\varphi,\boldsymbol{\alpha})+J_0g_0^2\gamma(\varepsilon,g_0)f_0(\varphi), \tag{2.1}$$

where $(\boldsymbol{\alpha},\mathbf{A})\in\mathbb{T}^{l-1}\times\mathbb{R}^{l-1}$, $(\varphi,I)\in\mathbb{T}^1\times\mathbb{R}^1$, $J_0>0$, J is a diagonal matrix, with $0<\det J\leqslant+\infty$, and

$$f_1(\varphi,\boldsymbol{\alpha})=\sum_{\substack{n\in\mathbb{Z}\\|n|\leqslant N_0}}\sum_{\substack{\nu\in\mathbb{Z}^{l-1}\\|\nu|\leqslant N}}f^1_{\underline{\nu}}e^{i(\nu\cdot\alpha+n\varphi)},\quad f^1_{\underline{\nu}}=f^1_{-\underline{\nu}},$$
$$f_0(\varphi)=(\cos\varphi-1)=\sum_{\substack{|n|=1\\\nu=0}}f^0_{\underline{\nu}}e^{i(\nu\cdot\alpha+n\varphi)}, \tag{2.2}$$

with $\underline{\nu}=(\nu_0,\boldsymbol{\nu})\equiv(n,\boldsymbol{\nu})\in\mathbb{Z}\times\mathbb{Z}^{l-1}$ and $|\boldsymbol{\nu}|=\sum_{j=1}^{l-1}|\nu_j|$; we prefer to consider the Hamiltonian (2.1) with l arbitrary because the Lindstedt series analysis holds for any $l\geqslant1$. So that the value $l=3$ and the existence of three time scales will be used only to obtain the second bound in (5.13) below.

The last term in (2.1) could be put together with the free pendulum potential $J_0g_0^2(\cos\varphi-1)$ thus modifying the ''gravity acceleration'' g_0^2 into $(1+\gamma(\varepsilon,g_0))g_0^2$; the term with $\gamma(\varepsilon,g_0)=\sum_{k=1}^{\infty}\gamma_k(g_0)\varepsilon^k$ is added because we follow here the approach in Ref. 5. We show that, given $s,\Omega,\eta,\boldsymbol{\omega}'$, with η small enough and $\boldsymbol{\omega}'$ verifying the Diophantine condition in Theorem 1.4, then one can fix $\gamma(\varepsilon,g_0)$ so that, for $|\varepsilon|<O(\eta^2)$, there is an invariant torus with average (over time) action $\mathbf{A}'$, with the properties in Theorem 1.4 and with Lyapunov exponent g_0 and rotation $\boldsymbol{\omega}'\equiv\boldsymbol{\omega}+J^{-1}\mathbf{A}'$. In other words by adding a *counterterm* to the Hamiltonian (1.1) one gets a new Hamiltonian system, (2.1), with an invariant torus with rotation $\boldsymbol{\omega}'$ *and Lyapunov exponent exactly equal to the prefixed* g_0 (see also Sec. II, 2.7 below).

We further show that, fixed $\boldsymbol{\omega}'$, $\gamma(\varepsilon,g_0)$ is jointly analytic in ε,g_0, if g_0 varies near a prefixed $\overline{g_0}>0$. Going back to the original Hamiltonian (1.1) we *therefore* set $g_0^2=\overline{g}_0^2(1+\gamma(\varepsilon,\overline{g}_0))$ and we can invert the latter relation as $\overline{g}_0^2=(1+\Gamma(\varepsilon,g_0))g_0^2$ for ε small enough (this will mean, for $|\varepsilon|<O(\eta^2)$). Hence by interpreting g_0^2 in (1.1) as $\overline{g}_0^2(1+\gamma(\varepsilon,\overline{g}_0))$, so that (1.5) holds, we obtain Theorem 1.4 as a corollary of the above statements.

Of course a similar proof could be done without first fixing the Lyapunov exponent $\overline{g_0}$ and then inverting the relation between the ''dressed exponent'' $\overline{g_0}$ and the ''bare'' one g_0. But it is well known, from the analogous problem in renormalization theory, that it is wiser technically and conceptually to work, in perturbation theory, with prefixed physical quantities (i.e., dressed ones). The idea that perturbation theory would be simpler, in the technical estimates, is the key idea beyond[3,10] that is introduced in Ref. 5.

2.2. From now on let us denote by $\boldsymbol{\alpha}$ the initial value of the rotators angles (i.e., at time $t=0$). We define by $X_j^\sigma(t;\boldsymbol{\alpha})$, $j=0,...,2l-1$, the values of the variables at time t that are reached from initial data $X^\sigma(0;\boldsymbol{\alpha})=(\pi,\boldsymbol{\alpha},I^\sigma(0;\mathbf{0}),\mathbf{A}^\sigma(0;\mathbf{0}))$, with the given $\boldsymbol{\alpha}$, with $\varphi=\pi$ and with I, $\mathbf{A}$ such that $X^\sigma(0;\boldsymbol{\alpha})$ is on the stable $(\sigma=+)$ or unstable $(\sigma=-)$ manifolds of the invariant torus that we are searching for; the convention on the labels of X is that

$$X_0^\sigma=\varphi^\sigma;\quad X_j^\sigma=\alpha_j^\sigma,\quad\text{for } 1<j<l;$$
$$X_l^\sigma=I^\sigma;\quad X_j^\sigma=A_j^\sigma,\quad\text{for } l<j<2l. \tag{2.3}$$

All functions in (2.3) depend on t and $\boldsymbol{\alpha}$ (the symbols $I^\sigma(t;\boldsymbol{\alpha})$ and $\mathbf{A}^\sigma(t;\boldsymbol{\alpha})$ should not be confused with $I^\sigma(\varphi,\boldsymbol{\alpha})$ and $\mathbf{A}^\sigma(\varphi,\boldsymbol{\alpha})$ defined in (1.2); in the following no ambiguity can arise as the quantities I^σ and $\mathbf{A}^\sigma$ *will be used always with the meaning in (2.3)* and as functions of t, $\boldsymbol{\alpha}$).

This is a parameterization of the stable and unstable manifolds in terms of $\boldsymbol{\alpha}$, t where $\boldsymbol{\alpha}$ is the value of the angular coordinates at the moment in which $\varphi=\pi$, and t is the time elapsed since. The parameterization is different from the one in terms of $\boldsymbol{\alpha}$, φ in (1.2) unless, of course, it is $\varphi=\pi$ and correspondingly $t=0$. Hence the splitting vector (1.7) at $\varphi=\pi$ can also be written $Q_j(\boldsymbol{\alpha}) = X_j^+(0;\boldsymbol{\alpha}) - X_j^-(0,\boldsymbol{\alpha})$, $j=l+1,...,2l-1$. Note that we do not need to consider explicitly the splitting in the I-coordinates because, by energy conservation, they are functions of $\varphi, \boldsymbol{\alpha}, \mathbf{A}^{\pm}$.

Let $\mathbf{A}'$ be given and let $\boldsymbol{\omega}'$ in (1.3) be Diophantine with constants $C=C(\eta)$, $\tau>0$; see (1.4). We look for an invariant torus and for its stable and unstable manifolds with the property that the quasiperiodic rotation on the torus takes place at velocity $\boldsymbol{\omega}'$ and, *at the same time*, the action variables oscillate with an average position $\mathbf{A}'$.

Before proceeding we remark that the above *two* requirements may seem contradictory as there may seem to be no reason for being able to prescribe simultaneously the ''spectrum'' $\boldsymbol{\omega}'$ and the ''average action'' $\mathbf{A}'$ of the invariant tori. In fact this property of ''*twistless*'' motion on the tori or of ''*absence of torsion*'' is very remarkable (see Remark 1.2 and Ref. 3); it will appear as due to the special symmetries of the system (2.1) and to the separation of the energy into a quadratic part involving actions only and an angular part involving only the angles.

Note also that we could confine ourselves to study the torus with average position $\mathbf{A}'=\mathbf{0}$, as in Refs. 3, 5 because any torus can be reduced to that one through a trivial canonical transformation (a translation in the action variables). This explains why in the quoted papers only the torus covered with rotation vector $\boldsymbol{\omega}$ is explicitly considered; however in the following we consider also $\mathbf{A}'\neq\mathbf{0}$, as we are interested in showing the abundance of such tori in phase space (see the Remark 1.5).

The quantity $X_j^{\sigma}(t;\boldsymbol{\alpha})$ can be graphically represented as sum of *values* which can be associated with tree graphs, that we shall call ''Feynman graphs'' or ''trees'' *tout court*, see Fig. 1. The trees are partially ordered sets of points, called *nodes*, connected by unit lines, called *branches*, and they are ''oriented'' towards a point called *root*, which is reached by a single branch of the tree. Given two nodes v and w of a tree, we say that w precedes v ($w\leq v$) if there is a path connecting w to v, oriented from w to v. With an abuse of notations we shall sometimes consider a tree as the collection of its nodes, sometimes as the collection of its branches and sometimes as the collection of both nodes and branches. The root *will not* be considered a node.

A typical tree considered below can be drawn as in Fig. 1; the labels meaning and the caption of such a drawing (which has to be interpreted as a mathematical formula) will be elucidated in the coming sections.

The branch starting at the node v and linking it to the uniquely determined next node (or to the root), which we call v', will be denoted by λ_v: there is a unique correspondence between nodes and branches starting at them. We shall say that λ_v exits from v and enters v'; given a node v we shall say that a branch λ *pertains* to v if either λ enters v or λ exits from v; e.g., in Fig. 1 the line $v_1v_0\equiv\lambda_{v_1}$ ''exits'' v_1 and ''enters'' v_0, hence it pertains to both.

In Ref. 3 two expansions are considered for the functions $X_j^{\sigma}(t;\boldsymbol{\alpha})$ representing the stable and unstable manifolds; one of them is used to exhibit cancellations taking place at all orders in the sums that express the coefficients of the power series in ε of the splitting vector;[3,13,8] it is somewhat more involved than the other one that is convenient to just discuss convergence of the perturbation series for the splitting vector and that we shall use here. This is the reason why (as in Ref. 5) we shall not have trees whose lowest nodes carry a graphical decoration called *form factor*, or *fruit* in Refs. 3, 8. Nevertheless some of the nodes will still have a particular structure; to characterize them we introduce, below as in Ref. 5, the notion of ''*leaf*,'' which is related to the notion of fruit in Ref. 3, from which it differs (and it, even, differs slightly from the similar notion of leaf in Ref. 5), see below for the motivation of the name.

2.3. As mentioned, Fig. 1 has to be regarded as a mathematical formula expressing a function of the labels and of the topological structure of the trees. We now prepare the notation for the definition of ''value'' of a tree (following Ref. 5) (see Ref. 3 for a simpler case); the derivation is not difficult but somewhat long and unusual for the subject (the breakthrough work[9] still does not seem to be well known in its technical aspects!). We discuss it in detail not only for completeness

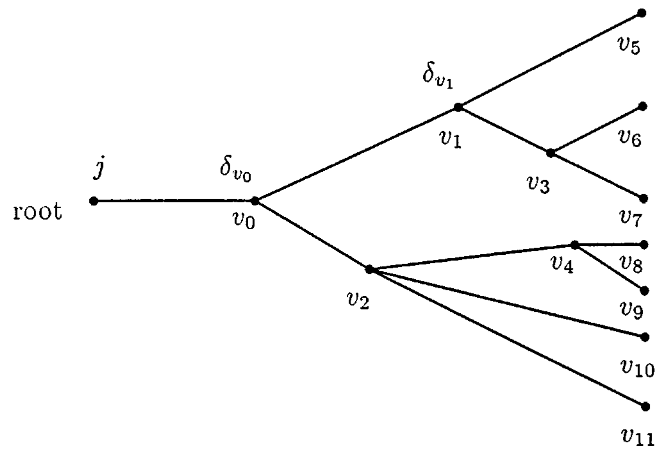

FIG. 1. A tree ϑ with $m=12$, and some labels. The line numbers, distinguishing the lines, and their orientation pointing at the root, are not shown. The lines length should be the same but it is drawn of arbitrary size. The nodes labels δ_v are indicated only for two nodes.

but in the attempt to clarify a construction that has generated quite a few new results starting from the work of Ref. 9, see Refs. 3, 8, 14.

Let us consider the unperturbed motion $X^0(t)\equiv(\varphi^0(t),\boldsymbol{\alpha}+\boldsymbol{\omega}'t,I^0(t),\mathbf{A}')$, where $(\varphi^0(t),I^0(t))$ is the separatrix motion, generated by the pendulum in (2.1) starting at $t=0$ in $\varphi=\pi$, $\mathbf{A}=\mathbf{A}'$, $I=-2J_0g_0$, so that $\varphi^0(t)=4\arctan e^{-g_0t}$. Let $X^\sigma(t;\alpha)$, $\sigma=\text{sign } t=\pm$, be the evolution, under the flow generated by (1.1), of the point on W^σ_ε which at time $t=0$ is $(\pi,\boldsymbol{\alpha},I^\sigma(\boldsymbol{\alpha},\pi),\mathbf{A}^\sigma(\boldsymbol{\alpha},\pi))$, see (1.2); let

$$X^\sigma(t)\equiv X^\sigma(t;\boldsymbol{\alpha})\equiv\sum_{h\geqslant0}X^{h\sigma}(t;\boldsymbol{\alpha})\varepsilon^h=\sum_{h\geqslant0}X^{h\sigma}(t)\varepsilon^h,\quad \sigma=\pm, \tag{2.4}$$

be the power series in ε of X^σ (which we want to show to be convergent for ε small); note that $X^{0\sigma}\equiv X^0$ is the unperturbed whisker. We shall often omit writing explicitly the $\boldsymbol{\alpha}$ variable among the arguments of various $\boldsymbol{\alpha}$-dependent functions, to simplify the notations, and we shall regard the two functions $X^{h\sigma}(t)$, as forming a single function $X^h(t)$, which is $X^{h+}(t)$ if $\sigma=+$, $t>0$, and $X^{h-}(t)$ if $\sigma=-$, $t<0$.

Components of X will be labeled j, $j=0,\ldots,2l-1$, consistently with (2.3), with the convention that $X_0\overset{\text{def}}{=}X_-$ describes the coordinate φ, $(X_j)_{j=1,\ldots,l-1}\overset{\text{def}}{=}\mathbf{X}_\downarrow$ describes the $\boldsymbol{\alpha}$ coordinates, $X_l\overset{\text{def}}{=}X_+$ describes the I coordinate and $(X_j)_{j=l+1,\ldots,2l-1}\overset{\text{def}}{=}\mathbf{X}_\uparrow$ describes the $\mathbf{A}$ coordinates,

$$X\overset{\text{def}}{=}(X_j)_{j=0,\ldots,2l-1}\overset{\text{def}}{=}(X_-,\mathbf{X}_\downarrow,X_+,\mathbf{X}_\uparrow), \tag{2.5}$$

i.e., we write first the angle and then the action components, first the pendulum and then the rotators. The $\uparrow$ (''up'') and $\downarrow$ (''down'') labels recall that the components with labels $\downarrow$ $(0<j<l)$ have ''lower'' index than the variables with labels $\uparrow$ $(l<j)$, which have a ''higher'' index (a mnemonically useful fact, on first reading at least).

Inserting (2.4) into the Hamilton equation associated with (2.1) we get that the coefficients $X^{h\sigma}(t)$, $h\geqslant1$, satisfy the hierarchy of linear equations

$$\frac{d}{dt}X^{h\sigma}(t)=L(t)X^{h\sigma}(t)+F^{h\sigma}(t), \tag{2.6}$$

with $F^{h\sigma}(t)$ a $2l$-vector and the $2l\times2l$-matrix $L(t)$ is

$$L(t)=\begin{pmatrix} 0 & \mathbf{0} & J_0^{-1} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & J^{-1} \\ g_0^2 J_0 \cos\varphi^0(t) & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 \end{pmatrix}. \tag{2.7}$$

For instance, $F^{1\sigma}(t)$ is a $2l$-vector with the first $0,...,l-1$ components vanishing (a consequence of the assumption that the perturbation only depends on the angular variables), with the lth component equal to $-J_0 g_0^2 \partial_\varphi f_1(\varphi^0(t),\boldsymbol{\alpha}+\boldsymbol{\omega}' t)+J_0 g_0^2 \gamma_1(g_0)\sin(\varphi^0(t))$ and with the remaining components equal to $-J_0 g_0^2 \partial_\alpha f_1(\varphi^0(t),\boldsymbol{\alpha}+\boldsymbol{\omega}' t)$.

In general $F^{h\sigma}$ depends upon $X^0,\dots,X^{h-1\sigma}$ *but not on* $X^{h\sigma}$. The entries of the $(2l\times 2l)$ matrix L have different meaning according to their position; the $\mathbf{0}$'s in the first and third row are $(l-1)$-(row)-vectors, the $\mathbf{0}$'s in the first and third column are $(l-1)$-(column)-vectors, and the 0's and J^{-1} in the second and fourth column are $(l-1)\times(l-1)$-matrices, while the 0's in the first and third columns are scalars (as J_0^{-1} is). The perturbed motions will be described by *dimensionless* quantities Ξ, Φ,

$$X_j^{h\sigma}=\Xi_j^{h\sigma},\quad 0\leqslant j\leqslant l-1,\quad X_j^{h\sigma}=J_0 g_0 \Xi_j^{h\sigma},\quad l\leqslant j\leqslant 2l-1,$$
$$\mathbf{F}_\uparrow^{h\sigma}=J_0 g_0^2 \boldsymbol{\Phi}_\uparrow^{h\sigma},\quad F_+^{h\sigma}=J_0 g_0^2 \Phi_+^{h\sigma}. \tag{2.8}$$

The simple form of the Hamiltonian equations for φ, $\boldsymbol{\alpha}$, namely $\dot\varphi=J_0^{-1}I$, $\dot{\boldsymbol{\alpha}}=\boldsymbol{\omega}+J^{-1}\mathbf{A}$ implies that $\boldsymbol{\Phi}_j^{h\sigma}=F_j^{h\sigma}\equiv 0$, for $j=0,...,l-1$. For instance,

$$\Phi^{1\sigma}=(0,\mathbf{0},-\partial_\varphi f_1(\varphi^0(t),\boldsymbol{\alpha}+\boldsymbol{\omega}' t)+\gamma_1(g_0)\sin(\varphi_0(t)),--\partial_\alpha f_1(\varphi^0(t),\boldsymbol{\alpha}+\boldsymbol{\omega}' t)). \tag{2.9}$$

Given the form of $L(t)$ and the vanishing of the first l components $F_-^{h\sigma}$, $\mathbf{F}_\downarrow^{h\sigma}$ of $F^{h\sigma}$, for $h\geqslant 1$, the above hierarchy of equations (determining the stable and unstable manifolds) takes the form

$$\frac{1}{g_0}\frac{d}{dt}\Xi_+^{h\sigma}=\cos\varphi^0\Xi_-^{h\sigma}+\Phi_+^{h\sigma},\quad \frac{1}{g_0}\frac{d}{dt}\boldsymbol{\Xi}_\uparrow^{h\sigma}=\boldsymbol{\Phi}_\uparrow^{h\sigma},$$
$$\frac{1}{g_0}\frac{d}{dt}\Xi_-^{h\sigma}=\Xi_+^{h\sigma},\quad \frac{1}{g_0}\frac{d}{dt}\boldsymbol{\Xi}_\downarrow^{h\sigma}=J_0 J^{-1}\boldsymbol{\Xi}_\uparrow^{h\sigma}. \tag{2.10}$$

And, for all $h\geqslant 1$, we can easily write (via Taylor expansion and order matching) the following formula for $\Phi^{h\sigma}$ in terms of the coefficients $\Xi^0,\dots,\Xi^{h-1\sigma}$ and of the derivatives of f_0 and $f_1\equiv f$, see (2.2). The first l components of $\boldsymbol{\Phi}^{h\sigma}$ vanish, as said above, $\Phi_-^{h\sigma}\equiv 0$, $\boldsymbol{\Phi}_\downarrow^{h\sigma}\equiv\mathbf{0}$, and

$$\boldsymbol{\Phi}_\uparrow^{h\sigma}=-\sum_{|\underline{m}|\geqslant 0}(\partial_\alpha f_1)_{\underline{m}}(\varphi^0,\boldsymbol{\alpha}+\boldsymbol{\omega}' t)\sum_{(h_j^i)_{\underline{m},h-1}}\prod_{i=0}^{l-1}\prod_{j=1}^{m_i}\Xi_i^{h_j^i\sigma},$$
$$\Phi_+^{h\sigma}\equiv-\sum_{|\underline{m}|\geqslant 2}(\partial_\varphi f_0(\varphi))_{\underline{m}}(\varphi^0)\sum_{(h_j^0)_{\underline{m},h}}\prod_{j=1}^{m_0}\Xi_-^{h_j^0\sigma}-\sum_{p=1}^{h}\sum_{|\underline{m}|\geqslant 0}\gamma_p(g_0)$$
$$\times(\partial_\varphi f_0(\varphi))_{\underline{m}}(\varphi^0)\sum_{(h_j^0)_{\underline{m},h-p}}\prod_{j=1}^{m_0}\Xi_-^{h_j^0\sigma}-\sum_{|\underline{m}|\geqslant 0}(\partial_\varphi f_1)_{\underline{m}}(\varphi^0,\boldsymbol{\alpha}$$
$$+\boldsymbol{\omega}' t)\sum_{(h_j^i)_{\underline{m},h-1}}\prod_{i=0}^{l-1}\prod_{j=1}^{m_i}\Xi_i^{h_j^i\sigma}, \tag{2.11}$$

where $(G)_{\underline{m}}(\cdot)$, with $G\in\{\partial_\varphi f_0,\partial_\alpha f_1,\partial_\varphi f_1\}$, and $(h_j^i)_{\underline{m},q}$, with $h_j^i\geqslant 1$, are defined as

$$
\begin{aligned}
&(G)_{\underline{m}}(\cdot)\equiv\left(\frac{\partial_{\varphi}^{m_0}\partial_{\alpha_1}^{m_1}\cdots\partial_{\alpha_{l-1}}^{m_{l-1}}G}{m_0!m_1!\cdots m_{l-1}!}\right)(\cdot),\\
&(h_j^i)_{\underline{m},q}\equiv(h_1^0,\ldots,h_{m_0}^0,h_1^1,\ldots,h_{m_1}^1,\ldots,h_1^{l-1},\ldots,h_{m_{l-1}}^{l-1}),\\
&\text{with}\quad\sum_{i=0}^{l-1}\sum_{j=1}^{m_i}h_j^i=q,
\end{aligned}
\tag{2.12}
$$

and $m_i\geqslant 0$, $m=(m_0,\ldots,m_{l-1})$, $|\underline{m}|=\Sigma_{i=0}^{l}m_i$. Note that the first two sums in the expression for $\Phi_+^{h\sigma}$ can only involve vectors $\underline{m}$ with $m_j=0$ if $j\geqslant 1$ (so that $|\underline{m}|=m_0$), because the function f_0, see (2.2), depends only on φ and not on $\boldsymbol{\alpha}$. The evolution of Ξ^h is determined by integrating (2.7), if the initial data are known. The $h=1$ case requires a suitable interpretation of the symbols, given explicitly by (2.9).

Elementary quadrability of the free pendulum equations on the separatrix leads to the following expression for the ''*Wronskian matrix*'' $W(t)$ of the separatrix motion for the pendulum appearing in (2.1), with initial data at $t=0$ given by $\varphi=\pi$, $I=-2g_0J_0$, i.e., $\Xi_+^0=-2$. The matrix

$$
W(t)=\begin{pmatrix} w_{00}(t) & w_{0l}(t)\\ w_{l0}(t) & w_{ll}(t)\end{pmatrix}
\tag{2.13}
$$

is defined to be the solution of the linearization of the free pendulum equation around the separatrix solution, with data $W(0)=1$ and with $J_0=1$ (*because we use dimensionless solutions* Ξ, see (2.10)),

$$
\begin{aligned}
&W(t)=\begin{pmatrix} \dfrac{1}{\cosh g_0t} & \dfrac{\bar{w}(t)}{4}\\ -\dfrac{\sinh g_0t}{\cosh^2 g_0t} & \left(1-\dfrac{\bar{w}(t)}{4}\dfrac{\sinh g_0t}{\cosh^2 g_0t}\right)\cosh g_0t\end{pmatrix},\\
&\bar{w}(t)\equiv\frac{2g_0t+\sinh 2g_0t}{\cosh g_0t}.
\end{aligned}
\tag{2.14}
$$

The evolution of the I,φ components, i.e., $\Xi_j^{h\sigma}$ with $j=0,l$ (also identified with the components with subscripts $\pm$, see (2.5)) can be determined from $W(t)$, by integrating (2.6) for the 0 and l components, to be

$$
\begin{pmatrix}\Xi_-^{h\sigma}\\ \Xi_+^{h\sigma}\end{pmatrix}=W(t)\begin{pmatrix}0\\ \Xi_+^{h\sigma}(0)\end{pmatrix}+W(t)\int_0^{g_0t}W^{-1}(\tau)\begin{pmatrix}0\\ \Phi_+^{h\sigma}(\tau)\end{pmatrix}dg_0\tau.
\tag{2.15}
$$

Thus, denoting by w_{ij} $(i,j=0,l)$ the entries of $W(t)$, (2.15) becomes, for $h\geqslant 1$,

$$
\begin{aligned}
&\Xi_-^{h\sigma}(t)=w_{0l}(t)\left(\Xi_+^{h\sigma}(0)+\int_0^{g_0t}w_{00}(\tau)\Phi_+^{h\sigma}(\tau)dg_0\tau\right)-w_{00}(t)\int_0^{g_0t}w_{0l}(\tau)\Phi_+^{h\sigma}(\tau)dg_0\tau,\\
&\Xi_+^{h\sigma}(t)=w_{ll}(t)\left(\Xi_+^{h\sigma}(0)+\int_0^{g_0t}w_{00}(\tau)\Phi_+^{h\sigma}(\tau)dg_0\tau\right)-w_{l0}(t)\int_0^{g_0t}w_{0l}(\tau)\Phi_+^{h\sigma}(\tau)dg_0\tau,
\end{aligned}
\tag{2.16}
$$

having used that $\Xi_-^{h,\sigma}(0)=0$ because the initial datum for φ is fixed and ε-independent. Likewise integration of Eqs. (2.11) for the $\uparrow,\downarrow$ components yields, for $h\geqslant 1$,

$$\Xi_{\downarrow}^{h\sigma}(t)=J^{-1}J_0\left[g_0t\left(\Xi_{\uparrow}^{h\sigma}(0)+\int_0^{g_0t}\Phi_{\uparrow}^{h\sigma}(\tau)dg_0\tau\right)-\int_0^{g_0t}g_0\tau\Phi_{\uparrow}^{h\sigma}(\tau)dg_0\tau\right],$$
$$\Xi_{\uparrow}^{h\sigma}(t)=\left(\Xi_{\uparrow}^{h\sigma}(0)+\int_0^{g_0t}\Phi_{\uparrow}^{h\sigma}(\tau)dg_0\tau\right), \tag{2.17}$$

having used that the $\Xi_{\downarrow}^{h\sigma}(0)\equiv\mathbf{0}$ because the initial datum for $\boldsymbol{\alpha}$ is fixed and ε-independent. Equations (2.16) and (2.17) can be used to find a reasonably simple algorithm to represent the whiskers equations to all orders $h\geqslant 1$ of the perturbation expansion.

2.4. The initial data in (2.16) and (2.17) have to be *determined by imposing that the solutions (to all orders) become quasiperiodic* as $t\rightarrow\sigma\infty$. This is quite easy and (as to be expected) this condition is simply that $\Xi_{+}^{h\sigma}(0)$, $\Xi_{\uparrow}^{h\sigma}(0)$ are determined by imposing that the integrals in parentheses become integrals between $\sigma\infty$ and t, i.e., $\Xi_{+}^{h\sigma}(0)=\int_{\sigma\infty}^{0}\cdots$ and $\Xi_{\uparrow}^{h\sigma}(0)=\int_{\sigma\infty}^{0}\cdots$; see below.

However the latter integrals are no longer necessarily convergent properly (a few examples suffice to see this); hence one has to go carefully through the process of imposing the correct asymptotic behavior in order to see what is the meaning to be given to such integrals $\int_{\sigma\infty}^{t}$. The analysis can be found in Refs. 3 and 5. The result is that all expressions under integral sign can be written as sums of functions that are rather special, namely,

$$M(t)=\sigma^{\chi}\frac{(\sigma g_0t)^j}{j!}e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}t-pg_0\sigma t}, \tag{2.18}$$

with $\chi,j,\boldsymbol{\nu},p$ integers and $p\geqslant -1$ (see below), so that one has

$$\Xi^{h\sigma}(t)=\sum_{\nu\in\mathbb{Z}^{l-1}}\sum_{p=-1}^{\infty}\tilde{\Xi}^{h\sigma}(\boldsymbol{\nu},p)e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}t-pg_0\sigma t},$$
$$\Phi^{h\sigma}(t)=\sum_{\nu\in\mathbb{Z}^{l-1}}\sum_{p=-1}^{\infty}\tilde{\Phi}^{h\sigma}(\boldsymbol{\nu},p)e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}t-pg_0\sigma t}, \tag{2.19}$$

where we explicitly write down only the dependence on $\boldsymbol{\nu}$ and p (clearly also the fixed constants like $J,J_0,g_0,\ldots$ enter).

The series turn out to be convergent for $\sigma t>0$; however their sums have *no singularity* at $t=0$ and can be analytically continued for $\sigma t<0$ (i.e., $x\geqslant 1$). More precisely the functions that one has to integrate are contained in an *algebra* $\hat{\mathcal{M}}$ on which the integration operations that we need can be given a meaning.

Definition (Ref. 3): Let $\hat{\mathcal{M}}$ *be the space of the functions of t which can be represented, for some* $k\geqslant 0$, *as*

$$M(t)=\sum_{j=0}^{k}\frac{(\sigma tg_0)^j}{j!}M_j^{\sigma}(x,\boldsymbol{\omega}t),\quad x\equiv e^{-\sigma g_0t},\quad \sigma=\operatorname{sign}t, \tag{2.20}$$

with $M_j^{\sigma}(x,\psi)$ *a trigonometric polynomial in* ψ *with coefficients holomorphic in the x-plane in the annulus* $0<|x|<1$, *with possible singularities, outside the open unit disk, in a closed cone centered at the origin, with axis of symmetry on the imaginary axis and half opening* $<\pi/2$, *and possible polar singularities at* $x=0$. *The smallest cone containing the singularities will be called the singularity cone of M.*

The proper interpretation of the improper integrals $\int_{\sigma\infty}^{g_0t}M(\tau)dg_0\tau$, which henceforth will be denoted by $⨍_{\sigma\infty}^{g_0t}M(\tau)dg_0\tau$, is simply the *residuum* at $R=0$ of the analytic function

$$\mathcal{I}_R M \stackrel{\text{def}}{=} \int_{\sigma\infty+i\theta}^{g_0 t} e^{-Rg_0\sigma z} M(z)\, dg_0 z, \tag{2.21}$$

(where θ is arbitrarily prefixed) which is defined and holomorphic for Re $R>0$ and large enough, i.e.,

$$\mathcal{I}M(t) \equiv \int_{\sigma\infty}^{g_0 t} dg_0\tau M(\tau) \stackrel{\text{def}}{=} \oint \frac{dR}{2\pi i R} \mathcal{I}_R M(t). \tag{2.22}$$

By linear extension this defines the integration of function in $\hat{\mathcal{M}}$ for $|x|<1$. The analyticity in x around $x=\pm 1$ and the remarks that $(d/dg_0 t)\mathcal{I}M(t) \equiv M(t)$, i.e., $\mathcal{I}M(t) \equiv \mathcal{I}M(t') + \int_{g_0 t'}^{g_0 t} dg_0 \tau M(\tau)$, so that $\mathcal{I}M(t)$ is a special primitive of $M(t)$ (at fixed σ), allow us to analytically continue the result of the integration to a function in $\hat{\mathcal{M}}$. The operator $\mathcal{I}$ maps the algebra $\hat{\mathcal{M}}$ into itself because one checks that on the monomial (2.18) one has

$$\mathcal{I}M(t) = \begin{cases} -g_0^{-1}\sigma^{\chi+1} e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}t - pg_0\sigma t} \displaystyle\sum_{h=0}^{j} \frac{(g_0\sigma t)^{j-h}}{(j-h)!} \frac{1}{(p - i\sigma g_0^{-1}\boldsymbol{\omega}'\cdot\boldsymbol{\nu})^{h+1}}, & \text{if } |p|+|\boldsymbol{\nu}|>0, \\ g_0^{-1}\sigma^{\chi+1} \dfrac{(\sigma g_0 t)^{j+1}}{(j+1)!}, & \text{otherwise}, \end{cases} \tag{2.23}$$

showing, in particular, that the radius of convergence in x of $\mathcal{I}M$, for a general M, is the same as that of M. But in general the singularities will not be polar, even when those of the M_j^σ's were such.

We shall see that the cases $|p|+|\boldsymbol{\nu}|=0$ do not enter in the discussion (a feature of the method of Ref. 5). The complete expression of $X^{h\sigma}(t)$ becomes

$$\begin{aligned} \Xi_-^{h\sigma}(t) &= w_{0l}(t)\mathcal{I}(w_{00}\Phi_+^{h\sigma})(t) - w_{00}(t)(\mathcal{I}(w_{0l}\Phi_+^{h\sigma})(t) - \mathcal{I}(w_{0l}\Phi_+^{h\sigma})(0^\sigma)) \stackrel{\text{def}}{=} \mathcal{O}(\Phi_+^{h\sigma})(t), \\ \boldsymbol{\Xi}_\downarrow^{h\sigma}(t) &= J^{-1}J_0(\mathcal{I}^2(\boldsymbol{\Phi}_\uparrow^{h\sigma})(t) - \mathcal{I}^2(\boldsymbol{\Phi}_\uparrow^{h\sigma})(0^\sigma)) \stackrel{\text{def}}{=} \overline{\mathcal{I}^2}(\boldsymbol{\Phi}_\uparrow^{h\sigma}(t)), \\ \Xi_+^{h\sigma}(t) &= w_{ll}(t)\mathcal{I}(w_{00}\Phi_+^{h\sigma})(t) - w_{l0}(t)(\mathcal{I}(w_{0l}\Phi_+^{h\sigma})(t) - \mathcal{I}(w_{0l}\Phi_+^{h\sigma})(0^\sigma)) \stackrel{\text{def}}{=} \mathcal{O}_+(\Phi_+^{h\sigma})(t), \\ \boldsymbol{\Xi}_\uparrow^{h\sigma}(t) &= \mathcal{I}(\boldsymbol{\Phi}_\uparrow^{h\sigma})(t), \end{aligned} \tag{2.24}$$

where $\mathcal{O},\mathcal{O}_+,\overline{\mathcal{I}^2}$ are implicitly defined here (and $\mathcal{I}^2$ is $\mathcal{I}$ applied twice); and $\Xi^{h\sigma}$, $\Phi^{h\sigma} \equiv (0,\mathbf{0},\Phi_+^{h\sigma},\boldsymbol{\Phi}_\uparrow^{h\sigma})$ are introduced in (2.9). While $\Xi^{h\sigma}$ has nonzero components over both the *angle* $(j=0,...,l-1)$ and over the *action* $(j=l,\ldots,2l-1)$ components, the $\Phi^{h\sigma}$ has, as already noted, only the action directions nonzero; the notation 0^σ means the limit as $t\to 0$ from the left $(\sigma=-)$ or from the right $(\sigma=+)$, but below we shall drop the superscript on 0 (always clear from he context because it is the same as the superscript σ of the functions $\Xi^{h\sigma}$). Furthermore, with the definitions (2.19) of $\tilde{\boldsymbol{\Phi}}_\uparrow^{h\sigma}(\boldsymbol{\nu},p)$ one finds also the property (with the notations in (2.1)),

$$\tilde{\boldsymbol{\Phi}}_\uparrow^{h\sigma}(\mathbf{0},0) = \mathbf{0}, \tag{2.25}$$

for all $h\geqslant 1$.

We shall repeatedly use that in order to compute $\Xi_j^{h\sigma}$ we only need $\Xi_{j'}^{h'\sigma}$ with $0 \leqslant j' < l$ (i.e., only $\Xi_+^{h'\sigma}$, $\Xi_\uparrow^{h'\sigma}$) and $h' < h$. This follows from (2.24) and (2.11); whether we want to compute an ''*action component*'' ($\Xi_j^{h\sigma}$, $j \geqslant l$) or an ''*angle component*'' ($\Xi_j^{h\sigma}$, $j < l$) of $\Xi^{h\sigma}$, we only need the angle components of lower orders, i.e., $\Xi_{j'}^{h'\sigma}$ with $h' < h$ and $j' < l$.

2.5. The linearity of the last of (2.24), together with (2.25) and the *t*-dependence of $\mathbf{\Phi}^{h\sigma}(t)$ in (2.19), implies that the prefixed value $\mathbf{A}'$ *has the interpretation of average action* of the quasiperiodic motion on the invariant torus to which the trajectories that we study asymptote; see the third statement in Theorem 1.4. This corresponds to the identity of Ref. 2 (see, in the latter reference, the first of (6.34) and its proof in Appendix (A12)) that follows from the symplectic structure of the equations of motion, according to a well known argument going back to Poincaré,[15] discussed also in Refs. 9,16. It is a property that generated the qualification of ''*twistless tori*'' given in Ref. 3 to such tori; the ''dispersion relation'' linking the frequencies to the average actions *does not change* or *is not twisted* when the perturbation is switched on. This is a property, established in the present context in (33) of Ref. 5, that can be ultimately traced back to the fact that in the above models the twist condition is not needed for establishing a KAM theorem.

2.6. By combining (2.24) and (2.11), (2.12) (and recalling (2.8)) the representation in terms of trees is immediate; the integrals in (2.24) and the lower order X^h in (2.11) become *recursively* multiple (improper) integrals over dummy ''time'' variables.

In this operation each function $(-\partial_\alpha f_1(\varphi^0(t), \boldsymbol{\alpha} + \boldsymbol{\omega}' t))_{\underline{m}}$ and $(-\partial_\varphi f_0(\varphi^0(t)))_{\underline{m}}$ is expanded as a linear combination of monomials $M(t)$ having the form $\sigma^\chi(\sigma g_0 t)^j (j!)^{-1} x^n e^{i\omega' \cdot \nu t}$ with $x = e^{-g_0 \sigma t}$; see (2.18).

The form of (2.11) shows that the integrations occur in a hierarchical order; *hence one can describe them by a tree*. The integrands can be identified by attaching to each node of the tree suitably many labels. We shall first illustrate the construction of the trees via two examples (in Sec. III, below); this can be useful in order to understand the general case (see also Refs. 3 and 5).

2.7. We shall establish, also recursively, that $\Xi^{h\sigma}$ will be expanded in monomial like (2.18) with $j=0$ and $p \geqslant 0$, so that at $t \to \pm\infty$ the quantities $\Xi^{h\sigma}$ will approach exponentially fast quasiperiodic functions describing the motion on the invariant torus. The approach will be proportional to $e^{-g_0|t|}$ or to a higher power of this quantity. This, together with the remark that at order 0 (i.e., on the unperturbed motion) the approach is precisely proportional to $e^{-g_0|t|}$ (in the I, φ coordinates), will imply that at least for $j = 0, l$ (and ''generically'' also for the other coordinates)

$$\lim_{t\to\infty} \frac{1}{\sigma t} \log|\Xi_j^\sigma(t)|^{-1} = g_0, \tag{2.26}$$

i.e., that the Lyapunov exponents of the torus are $\pm g_0$.

Before stating the general graphical rules to represent (2.24) in terms of explicitly performed integrals, we discuss in detail two examples; understanding them facilitates enormously, we think, the understanding of the general cases which will be exposed referring to the examples to make it more concrete.

III. TWO EXAMPLES OF THE TREES CONSTRUCTION

3.1. We discuss how to make more explicit (2.24) by performing two ''third order'' examples. The first order reduces trivially to the first order formulas (Mel'nikov integral); the second order is also a bit too simple and is left to the reader; the first two orders will be, of course, implicitly done below, because to compute the third order one needs the first and second, too.

To third order the last line in (2.24) gives $\mathbf{\Xi}_\uparrow^{3\sigma}(t) = \mathcal{I}(\mathbf{\Phi}_\uparrow^{3\sigma})(t)$, where $\mathbf{\Phi}_\uparrow^{3\sigma}$ can be expressed through the first equation in (2.11), so that, for $j = l+1, \ldots, 2l-1$, one has

$$\Phi_j^{3\sigma} = -\frac{1}{2}\partial_{\alpha_j}\partial_\varphi^2 f_1 \Xi_-^{1\sigma}\Xi_-^{1\sigma} - \partial_{\alpha_j}\partial_\varphi f_1 \Xi_-^{2\sigma} - \sum_{p=1}^{l-1}\partial_{\alpha_j}\partial_{\alpha_p}\partial_\varphi f_1 \Xi_p^{1\sigma}\Xi_-^{1\sigma}$$

$$-\frac{1}{2}\sum_{p,q=1}^{l-1}\partial_{\alpha_j}\partial_{\alpha_p}\partial_{\alpha_q} f_1 \Xi_p^{1\sigma}\Xi_q^{1\sigma} - \sum_{p=1}^{l-1}\partial_{\alpha_j}\partial_{\alpha_p} f_1 \Xi_p^{2\sigma}, \tag{3.1}$$

where $\Xi^{1\sigma}$ and $\Xi^{2\sigma}$ can be written by using once more (2.24) (the first two lines only as per the general remark in the last paragraph of Sec. II, 2.4).

We consider explicitly two contributions to $\boldsymbol{\Xi}_\uparrow^{3\sigma}(t)$. Recalling that $\sigma=+$ corresponds to the stable manifold and $\sigma=-$ to the unstable one, the first will be

$$\frac{1}{2}\int_{\sigma\infty}^{g_0 t} dg_0 \tau_{v_0}(-\partial_{\alpha_j\alpha_p\alpha_q} f_1(\varphi^0(\tau_{v_0}), \boldsymbol{\alpha}+\boldsymbol{\omega}'\tau_{v_0})\Xi_p^{1\sigma}(\tau_{v_0})\Xi_q^{1\sigma}(\tau_{v_0}), \tag{3.2}$$

arising from the fourth contribution in the r.h.s. of (3.1). The contribution (3.2) can be written more explicitly, by using again the expression for $\boldsymbol{\Xi}_\downarrow^{1\sigma}$ in (2.25), as

$$\frac{1}{2}\int_{\sigma\infty}^{g_0 t} dg_0 \tau_{v_0}(-\partial_{\alpha_j\alpha_p\alpha_q} f_1)(\tau_{v_0})\overline{\mathcal{I}^2}(-\partial_{\alpha_p} f_1(\tau_{v_1}))(\tau_{v_0})\overline{\mathcal{I}^2}(-\partial_{\alpha_p} f_1(\tau_{v_2}))(\tau_{v_0}), \tag{3.3}$$

where the $\overline{\mathcal{I}^2}$ operations involve, see (2.24), integrations over variables that we can call τ_{v_1}, τ_{v_2} and the derivatives of f_1 are evaluated at $(\varphi^0(\tau_{v_n}), \boldsymbol{\alpha}+\boldsymbol{\omega}'\tau_{v_n})$, $n=0,1,2$. Such variables have been indicated explicitly using the abbreviated notation (τ_{v_n}) and with a obvious abuses of notation (they should not appear at all, except τ_{v_0}, being dummy).

The second example is obtained by considering the contribution with $h_2^0=2$ from the first line of (2.11), i.e., the second contribution in the r.h.s. of (3.1),

$$\int_{\sigma\infty}^{g_0 t}(-\partial_{\alpha_j\varphi} f_1)(\tau_{v_0})\Xi_-^{2\sigma}(\tau_{v_0})d\tau_{v_0}, \tag{3.4}$$

still imagining the derivatives of f_1 evaluated at $(\varphi^0(\tau_{v_0}), \boldsymbol{\alpha}+\boldsymbol{\omega}'\tau_{v_0})$. This contribution will be the sum of several terms, because $\Xi_-^{2\sigma}(\tau_{v_0})$ has to be expressed by using (2.24) and (2.11). One of the (many) contributions will be

$$\frac{1}{2}\int_{\sigma\infty}^{g_0 t} dg_0 \tau_{v_0}(-\partial_{\alpha_j\varphi} f_1)(\tau_{v_0})\mathcal{O}(-\partial_\varphi^3 f_0(\tau_{v_1})\mathcal{O}(-\partial_\varphi f_1(\tau_{v_2}))(\tau_{v_1})$$
$$\times\mathcal{O}(-\partial_\varphi f_1(\tau_{v_3}))(\tau_{v_1}))(\tau_{v_0}), \tag{3.5}$$

where the $\mathcal{O}$ operations involve, see (2.24), integrations over variables that we can call τ_{v_1}, τ_{v_2}, τ_{v_3} and the derivatives of f_0, f_1 are evaluated at $(\varphi^0(\tau_{v_n}), \boldsymbol{\alpha}+\boldsymbol{\omega}'\tau_{v_n})$, $n=0,1,2,3$. Such variables have been indicated explicitly with the same abuse of notation as above; and the dependence on τ of the derivatives of f_0, f_1 has again been simply denoted by adding the symbol (τ_{v_n}) instead of the full argument $(\varphi^0(\tau_{v_n}), \boldsymbol{\alpha}+\boldsymbol{\omega}'\tau_{v_n})$.

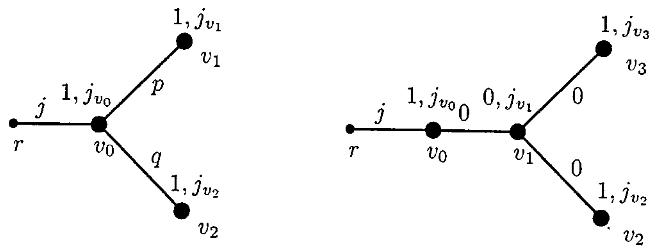

FIG. 2. Graphic representation of the contributions (3.3) and (3.5).

A complete representation of the above two contributions to $\Xi_j^{3\sigma}(t)$ is given, *with enormous notational simplification*, by the trees in Fig. 2, where the labels on the nodes v are denoted δ_v, j_v and those on the lines λ_v are denoted j_{λ_v}.

The label $\delta_v=0,1$ on the node v indicates selection of f_{δ_v}, i.e., of f_0 or f_1, the label j_v denotes a derivative with respect to φ if $j_v=l$ or with respect to α_{j_v} if $j_v=l+1,...,2l-1$. For the label j_{λ_v} associated with the branch λ_v following v, one has $j_{\lambda_v}=j_v-l$ for all v except for the highest node v_0, for which one has $j_{\lambda_{v_0}}=j_{v_0}$. In the examples above, (3.3) and (3.5) correspond, respectively, to the first figure in Fig. 2 with $j_{v_0}=j$, $j_{v_1}=p+l$, $j_{v_2}=q+l$ and to the second with $j_{v_1}=j_{v_2}=j_{v_3}=l$, $j_{v_0}=j$, (hence $j_{\lambda_{v_1}}=j_{\lambda_{v_2}}=j_{\lambda_{v_3}}=0$, $j_{v_0}=j$). In the examples the labels p,q correspond to $\partial_{\alpha_p},\partial_{\alpha_q}$ in (3.3).

3.2. *Remark:* The exception for the meaning of $j_{\lambda_{v_0}}$ is convenient, in the above cases, as the integration over τ_{v_0} differs from the others: the inner ones evaluate $\Xi_j^{h\sigma}$ for $j=0,...,l-1$, because the functions f_0,f_1 only depend on the angle variables (see the last paragraph in Sec. II); the last integral, however, evaluates in the examples a component of $\boldsymbol{\Xi}_{\uparrow}^{h\sigma}$ (which is labeled $j=l+1,...,2l-1$), but, in general, j can be any value $j=0,...,2l-1$. Note that this is not so for the inner labels j_λ which must be angle labels $j_\lambda=0,...,l-1$. So, in general, we shall have that the value of a tree with $j_{\lambda_{v_0}}=j$ contributes to $\Xi_j^{h\sigma}$.

IV. TREES AND FEYNMAN GRAPHS APPROACH TO WHISKERS CONSTRUCTION: THE GENERAL CASE

We now proceed to describe the general case.

4.1. To compute the splitting vector we only need to consider the variable t equal to 0. However we shall be also interested in $\Xi^{h\sigma}(\boldsymbol{\alpha},t)$ for $\sigma t>0$, for instance in order to study how fast the invariant torus is approached by the motions on its stable and unstable manifolds (to obtain its Lyapunov exponent). Hence it will be natural to attribute the label t to the root: this will also remind that the integral over τ_{v_0} has to be performed between $\sigma\infty$ and t, (the value $\sigma=-$ corresponds to the unstable manifold and the value $\sigma=+$ corresponds to the stable one). Since we shall *never* consider the stable manifold for $t>0$ or the unstable for $t<0$ the value of σ will be the same as that of the sign of t.

We shall be interested in computing not only $\mathbf{X}_{\uparrow}^{\sigma}(0;\boldsymbol{\alpha})-\mathbf{A}'$ (or $\mathbf{X}_{\uparrow}^{\sigma}(t;\boldsymbol{\alpha})-\mathbf{A}'$), as in Ref. 8, but, more generally, $X^{\sigma}(t;\boldsymbol{\alpha})-X^0(t;\boldsymbol{\alpha})$, with $\sigma=\text{sign}\,t$, (here X^0 denotes the unperturbed motion).

In general the rules to express $X^{\sigma}(t;\boldsymbol{\alpha})-X^0(t;\boldsymbol{\alpha})$ as sum of ''values'' associated with trees will be described now, assuming that the reader follows us by applying and checking them to the special cases (3.3), (3.5), illustrated in Fig. 2.

The reader might be helped in following the construction of the algorithm to express the stable and unstable manifolds below, by keeping in mind that we simply decompose the (quite involved and recursively defined by (2.24), (2.11)) expressions for the whiskers, so far obtained, *further*.

The purpose being of reducing their evaluation to *very elementary algebraic operations*; ultimately just products of simple factors associated with the nodes (and their labels) of a tree, that we shall call ''coupling constants,'' and of factors associated with the branches (and their labels), that we shall call ''propagators,'' each of which can be trivially evaluated and trivially bounded.

To each node we attach an *order label* $\delta_v = 0,1$, see Fig. 2, and a corresponding function f_{δ_v}; if a node v bears a label $\delta_v = 1$ the associated functions is f_1 and if it bears a label $\delta_v = 0$ it is f_0.

To each node v of a tree ϑ, see Fig. 1, we associate an integration *time variable* τ_v and an *integration operation*, which corresponds to $\overline{\mathcal{I}}^2$ or $\mathcal{O}$ if the node *is not the highest node* v_0 and to $\overline{\mathcal{I}}^2$ or $\mathcal{O}$ or $\mathcal{I}$ or $\mathcal{O}_+$ if the node v *is the highest*, i.e., $v = v_0$. This is so because in the first case (a ''lower node'') one must use the first two equations in (2.24) because in (2.11) only angle components of $X^{h\sigma}$ appear, while in the second case (that of the highest node) one can use all of (2.24) since we can evaluate either an angle coordinate $\Xi_j^{h\sigma}(\boldsymbol{\alpha},t)$, $j<l$, or an action coordinate, $j \geqslant l$.

When $v < v_0$ the choice between the two possibilities will be marked by an *action label* j_v associated with each node; if $j_v = l$, $v < v_0$, then we choose $\mathcal{O}$, if $j_v = l+1,...,2l-1$, $v < v_0$, we choose $\overline{\mathcal{I}}^2$.

When v is the highest node v_0, there are therefore more possibilities; to distinguish between them we use the *action label* j_{v_0} and the *branch label* $j_{\lambda_{v_0}}$, which can be equal either to j_{v_0} or to $j_{v_0} - l$. So when $v = v_0$ and $j_{v_0} = l$, we choose $\mathcal{O}$ if $j_{\lambda_{v_0}} = j_{v_0} - l = 0$ and $\mathcal{O}_+$ if $j_{\lambda_{v_0}} = j_{v_0} = l$, see (2.24), while when $v = v_0$ and $j_{v_0} > l$, we choose $\overline{\mathcal{I}}^2$ if $j_{\lambda_{v_0}} = j_{v_0} - l$ and $\mathcal{I}$ if $j_{\lambda_{v_0}} = j_{v_0}$, see (2.24).

As said in Remark 3.2, the meaning of the branch label is that a tree with $j_{\lambda_{v_0}} = j$ is a graphic representation of a ''contribution'' to $\Xi_j^{h\sigma}$. Therefore if $j_{\lambda_{v_0}} \geqslant l$ we call the branch an *action branch* and if $j_{\lambda_{v_0}} < l$ we call it an *angle branch*.

In the first of the figures in Fig. 2 integrals with respect to the nodes v_1, v_2 are of the type $\overline{\mathcal{I}}^2$. In the second the integrals over the τ_{v_n}, $n = 1,2,3$, are all of the type $\mathcal{O}$. In both cases the integrals over τ_{v_0} are of the form $\mathcal{I}$ because we fixed $j_{\lambda_{v_0}} = j > l$ to be an action label. We can associate a branch label j_{λ_v} also to the inner branches with $v < v_0$; however, in this case one has necessarily $j_{\lambda_v} = j_v - l$ because the inner branches necessarily represent angle components $\Xi_j^{h\sigma}$ with $j<l$, see (2.11). Hence no information is carried by such labels that we define only for uniformity of notation. The latter labels appear in Fig. 2 as *j*,*p*,*q* in the first tree and as *j*, 0, 0, 0 in the second one. The labels j_λ corresponding to the lines pertaining to a node v determine, as in the examples of Sec. III, which derivatives have to be taken of the function f_{δ_v} which is associated with v; each line λ_v with label j_{λ_v} corresponds to a derivative of f_{δ_v} with respect to φ if $j_{\lambda_v} = 0$ or to $\alpha_{j_{\lambda_v}}$ if $0 < j_{\lambda_v} < l$.

The integrations over the node times τ_v must be thought of as improper integrals, in the above sense, from $\sigma\infty$ to either $\tau_{v'}$ or 0 because (2.24) contains various integrals between such extremes. *It will be convenient to distinguish between such terms*.

This can be easily done by adding, on each node, a new label ρ_v also equal to 0, 1; if $\rho_v = 1$ this means, naturally, that in the evaluation of the integration operations relative to the node v we select the terms that correspond to integrations between $\sigma\infty$ and $\tau_{v'}$ while if $\rho_v = 0$ we select the integrations between $\sigma\infty$ and 0. We shall imagine that also the highest node carries a label ρ_{v_0} which is 1 necessarily if we consider only $\Xi_\uparrow(t)$ (because this implies that the function associated with the highest node must appear differentiated with respect to a α-component, see above and (2.11)), but which could be 0 for the other components of $\Xi(t)$. Recall also that in this case $\tau_{v'} \equiv t$, see (2.24).

We remark that the hierarchical structure of the integrations implies that if $\rho_v = 1$ and if v' is the node immediately following (in the direction of the root) v along the tree then one has $\tau_v > \tau_{v'}$ if $\sigma = +$ and $\tau_v < \tau_{v'}$ if $\sigma = -$, while τ_v, $\tau_{v'}$ have the same sign but are otherwise unrelated if $\rho_v = 0$; see (2.24) and check this in the examples.

Besides the labels already introduced also the labels $\rho_v=0,1$, just described but not shown in Fig. 2, should be imagined carried by each node.

Given a tree labeled as above we pick up the nodes v with $\rho_v=0$ which are closest to the root, and consider the subtrees having such nodes as highest nodes. We call each such subtree, i.e., each such node *together with the subtrees ending in it* (and its labels), a *leaf*. [This definition is slightly different from the one given in Ref. 5, where the leaf represents a collection of trees and, as explained below, is related to a resummation operation (see also comments in Sec. IV, 4.3 item (v), below, and (4.27)), that we do not consider here.] The name is natural if one imagines to enclose the part of the tree including the node v itself and half of the line λ_v into a circle (or, more pictorially, into a leaf shaped contour): hence, to whom tries the drawing, it will look like the *venations* of a leaf and the half line outside it will look like its *stalk*.

All nodes which do not belong to any leaves will be called *free nodes*; they carry, by construction, a label $\rho_v=1$, so that the corresponding time variables are hierarchically ordered from the lowest nodes up to the root; i.e., if $w<v$ then $\tau_w<\tau_v$ if $\sigma=+$ and $\tau_w>\tau_v$ if $\sigma=-$. Given a tree ϑ let us call ϑ_f the set of free nodes in ϑ, and call Θ_L the set of highest nodes of the leaves.

Each f_{δ_v} function, associated with the node v with order label δ_v, can be decomposed into its Fourier harmonics. This can be done graphically by adding to each node v a *mode* label $\underline{\nu}_v=(\nu_{0v},\boldsymbol{\nu}_v)\equiv(n_v,\boldsymbol{\nu}_v)\in\mathbb{Z}^l$, with $|\boldsymbol{\nu}_v|\leq N$ and $|n_v|\leq N_0$, that denotes the particular harmonic selected for the node v. If $(j_v,\delta_v,\rho_v,\underline{\nu}_v)$ are the labels of v we will associate with v the quantity $f^{\delta_v}_{\boldsymbol{\nu}_v}e^{i(\omega\cdot\nu_v\tau_v+n_v\varphi^0(\tau_v))}$ multiplied by appropriate products of factors in_v (one per φ-derivative) and $i\nu_{vj}$ (one per α_j-derivative, $j=j_{\lambda_v}$). If the mode labels $\underline{\nu}_v$ are specified for each v we shall define the *momentum* $\boldsymbol{\nu}(v)$ ''flowing'' on a branch λ_v as the sum of all the angle mode components $\boldsymbol{\nu}_w$ of the nodes w *preceding* the branch, with v included,

$$\boldsymbol{\nu}(v)\overset{\text{def}}{=}\sum_{w\in\vartheta,w\leq v}\boldsymbol{\nu}_w; \tag{4.1}$$

the momentum $\boldsymbol{\nu}(v_0)$ flowing through the root branch will be called the *total momentum* (of the tree).

We shall define also the *total free momentum* of the tree as the sum of the mode labels of the free nodes; more generally, for any free node v we can define the *free momentum* flowing through the branch λ_v as

$$\boldsymbol{\nu}_0(v)=\sum_{w\in\vartheta_f,w\leq v}\boldsymbol{\nu}_w. \tag{4.2}$$

For instance, in the above examples the two contributions (3.3), (3.5) (represented by Fig. 2) are decomposed into sums of several distinct contributions once the ρ_v and the mode labels ν_v are specified.

Likewise we can look at a leaf as a tree; the momentum $\boldsymbol{\nu}'$ flowing through its stalk will then be called the internal *leaf momentum*. Note that its value gives *no contribution* to the total free momentum of the tree to which the leaf belongs.

The free momenta will turn out to describe the harmonics of the time dependent quasiperiodic motion around the invariant tori, while the Fourier expansion modes of $X^{h\sigma}(t;\boldsymbol{\alpha})$ as a function of $\boldsymbol{\alpha}$ are related to the sum of the free momenta *and* of all the internal leaf momenta. This is an important difference: it is a property stressed in Ref. 3 where it is referred as ''quasiflatness,'' source of the main difficulties and interest in the theory of homoclinic splitting, see Refs. 3, 8, 6, 7.

4.2. The trees contributions of the examples of Sec. III will be sums over the various labels of ''values'' of trees decorated by more labels,

$$\frac{1}{2}\int_{\sigma\infty}^{g_0 t} dg_0 \tau_{v_0}(-i\nu_{v_0 j})(i\nu_{v_0 p})(i\nu_{v_0 q}) f^1_{\underline{\nu}_{v_0}} e^{i(\boldsymbol{\nu}_{v_0}\cdot\boldsymbol{\omega}'\tau_{v_0}+n_{v_0}\varphi^0(\tau_{v_0}))} \cdot \overline{\mathcal{I}}^2((-i\nu_{v_1 p}) f^1_{\underline{\nu}_{v_1}} e^{i(\boldsymbol{\nu}_{v_1}\cdot\boldsymbol{\omega}'\tau_{v_1}+n_{v_1}\varphi^0(\tau_{v_1}))})(\rho_{v_1}\tau_{v_0}) \cdot \overline{\mathcal{I}}^2((-i\nu_{v_2 q}) f^1_{\underline{\nu}_{v_2}} e^{i(\boldsymbol{\nu}_{v_2}\cdot\boldsymbol{\omega}'\tau_{v_2}+n_{v_2}\varphi^0(\tau_{v_2}))})(\rho_{v_2}\tau_{v_0}), \tag{4.3}$$

for (3.3) and

$$\frac{1}{2}\int_{\sigma\infty}^{g_0 t} dg_0 \tau_{v_0}(-i\nu_{v_0 j})(in_{v_0}) f^1_{\underline{\nu}_{v_0}} e^{i(\boldsymbol{\nu}_{v_0}\cdot\boldsymbol{\omega}'\tau_{v_0}+n_{v_0}\varphi^0(\tau_{v_0}))} \mathcal{O}((-in_{v_1}) f^0_{\underline{\nu}_{v_1}} e^{in_{v_1}\varphi^0(\tau_{v_1})} \cdot \mathcal{O}((-in_{v_2}) f^1_{\underline{\nu}_{v_2}} e^{i(\boldsymbol{\nu}_{v_2}\cdot\boldsymbol{\omega}'\tau_{v_2}+n_{v_2}\varphi^0(\tau_{v_2}))})(\rho_{v_2}\tau_{v_1}) \cdot \mathcal{O}((-in_{v_3}) f^1_{\underline{\nu}_{v_3}} e^{i(\boldsymbol{\nu}_{v_3}\cdot\boldsymbol{\omega}'\tau_{v_3}+n_{v_3}\varphi^0(\tau_{v_3}))})(\rho_{v_3}\tau_{v_1})), \tag{4.4}$$

for (3.5), with the conventions following (3.3) about the dummy integration variables.

The integration operations are still fairly involved, as it can be seen from (2.24) and from the expressions for $\overline{\mathcal{I}}^2$ and $\mathcal{O}$. With the above conventions for the dummy variables and noting that, for any function F in $\hat{\mathcal{M}}$,

$$\overline{\mathcal{I}}^2(F(\tau))(t) = J^{-1}J_0(\mathcal{I}(g_0(t-\tau)F(\tau))(t) - \mathcal{I}(g_0(t-\tau)F(\tau))(0)), \tag{4.5}$$

we see that the integration over the τ_v has (by (2.24)) one of the two forms, when $\rho_v=1$ and v' is not the root (so that $j_{\lambda_v}=j_v-l$),

$$(1)\quad \mathcal{I}((w_{0l}(\tau_{v'})w_{00}(\tau_v) - w_{00}(\tau_{v'})w_{0l}(\tau_v))\cdot e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(\tau_{v'}),\quad j_{\lambda_v}=0,$$

$$(2)\quad \mathcal{I}(g_0(\tau_{v'}-\tau_v)\cdot e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(\tau_{v'}),\quad 0<j_{\lambda_v}<l, \tag{4.6}$$

where $G_v(\tau_v)$ is a function that depends on the structure of the tree formed by the nodes preceding v and by the labels attached to the nodes. If $\rho_v=0$ it has one of the two forms

$$(1)\quad w_{00}(\tau_{v'})\mathcal{I}(w_{0l}(\tau_v)e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(0),\quad j_{\lambda_v}=0,$$

$$(2)\quad \mathcal{I}(g_0\tau_v e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(0),\quad 0<j_{\lambda_v}<l. \tag{4.7}$$

When v' is the root the operations involved in the evaluation of the τ_v-integral are slightly different *if* $j_{\lambda_{v_0}}=j_{v_0}$, i.e., if we are considering contributions to the action coordinates (if $j_{\lambda_{v_0}}=j_{v_0}-l$ we still have integrations of the form (4.6) or (4.7)). If $j_{\lambda_{v_0}}=j_{v_0}$ the integrations are particularly simple if we are interested in the evaluation of the splitting vector (1.7), that is $j_{v_0}>l$ and $t=0$; in such a case the last two of (2.24) are relevant and setting $v=v_0$ the integration over τ_{v_0} is the value for $\tau_{v'}$ of

$$(1)\quad \mathcal{I}(w_{00}(\tau_v)e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(0),\quad j_{\lambda_v}=l,$$

$$(2)\quad \mathcal{I}(e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(0),\quad j_{\lambda_v}>l, \tag{4.8}$$

because, if $\tau_{v'}=0$, one has $w_{ll}(0)=1$ and $w_{l0}(0)=0$; see (2.14) and the last two of (2.24).

More generally, if $\tau_{v_0'}=t\neq 0$, setting $v=v_0$ and $r=v_0'$, one defines for $\rho_{v_0}=1$,

$$
\begin{aligned}
&(1)\quad \mathcal{I}((w_{0l}(\tau_{v'})w_{00}(\tau_v)-w_{00}(\tau_{v'})w_{0l}(\tau_v))\cdot e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(\tau_{v'}),\quad j_{\lambda_v}=0,\\
&(2)\quad \mathcal{I}(g_0(\tau_{v'}-\tau_v)\cdot e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(\tau_{v'}),\quad 0<j_{\lambda_v}<l,\\
&(3)\quad \mathcal{I}((w_{ll}(\tau_{v'})w_{00}(\tau_v)-w_{l0}(\tau_{v'})w_{0l}(\tau_v))\cdot e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(\tau_{v'}),\quad j_{\lambda_v}=l,\\
&(4)\quad \mathcal{I}(e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(\tau_{v'}),\quad j_{\lambda_v}>l,
\end{aligned}
\tag{4.9}
$$

(see the last two relations in (2.24)) and for $\rho_{v_0}=0$

$$
\begin{aligned}
&(1)\quad w_{00}(\tau_{v'})\mathcal{I}(w_{0l}(\tau_v)e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(0),\quad j_{\lambda_v}=0,\\
&(2)\quad \mathcal{I}(g_0\tau_v e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(0),\quad 0<j_{\lambda_v}<l.\\
&(3)\quad w_{l0}(\tau_{v'})\mathcal{I}(w_{0l}(\tau_v)e^{i(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v+n_v\varphi^0(\tau_v))}G_v(\tau_v))(0),\quad j_{\lambda_v}=l,\\
&(4)\quad 0,\quad j_{\lambda_v}>l;
\end{aligned}
\tag{4.10}
$$

note that, for $\tau_{v'}=t=0$ and $j_{\lambda_{v_0}}\geqslant l$, (4.9) and (4.10), summed together, give (4.8).

Hence each node still describes a rather complicated set of operations: it is, therefore, convenient to consider separately the terms that appear in (4.6)–(4.10). This can be done by simply adding further labels at each node. To this end, looking at the integrals in (4.7) and (4.10), at $\rho_v=0$, and in (4.6) and (4.9), at $\rho_v=1$, we see that the following kernels are involved in the integrals:

$$
\begin{aligned}
&w^0_{j_{\lambda_v}}(\tau_{v'},\tau_v)=\begin{cases} w_{00}(\tau_{v'})w_{0l}(\tau_v), & v>v_0,\ j_v=l\rightarrow j_{\lambda_v}=0,\\ g_0\tau_v, & v>v_0,\ j_v>l\rightarrow 0<j_{\lambda_v}<l,\end{cases}\\
&w^0_{j_{\lambda_{v_0}}}(t,\tau_{v_0})=\begin{cases} w_{00}(t)w_{0l}(\tau_{v_0}), & j_{v_0}=l,\quad j_{\lambda_{v_0}}=0,\\ g_0\tau_{v_0}, & j_{v_0}>l,\quad 0<j_{\lambda_{v_0}}<l,\\ w_{l0}(t)w_{0l}(\tau_{v_0}), & j_{v_0}=l,\quad j_{\lambda_{v_0}}=l\\ 0, & j_{v_0}>l,\quad j_{\lambda_{v_0}}>l,\end{cases}\\
&w^1_{j_{\lambda_v}}(\tau_{v'},\tau_v)=\begin{cases} w_{0l}(\tau_{v'})w_{00}(\tau_v)-w_{00}(\tau_{v'})w_{0l}(\tau_v), & v>v_0,\ j_v=l\rightarrow j_{\lambda_v}=0,\\ g_0(\tau_{v'}-\tau_v), & v>v_0,\ j_v>l\rightarrow 0<j_{\lambda_v}<l,\end{cases}\\
&w^1_{j_{\lambda_{v_0}}}(t,\tau_{v_0})=\begin{cases} w_{0l}(t)w_{00}(\tau_{v_0})-w_{00}(t)w_{0l}(\tau_{v_0}), & j_{v_0}=l,\quad j_{\lambda_{v_0}}=0,\\ g_0(t-\tau_{v_0}), & j_{v_0}>l,\quad 0<j_{\lambda_{v_0}}<l,\\ w_{ll}(t)w_{00}(\tau_{v_0})-w_{l0}(t)w_{0l}(\tau_{v_0}), & j_{v_0}=l,\ j_{\lambda_{v_0}}=l,\\ 1, & j_{v_0}>l,\quad j_{\lambda_{v_0}}>l,\end{cases}
\end{aligned}
\tag{4.11}
$$

respectively appearing in (4.7) and (4.10), at $\rho_v=0$, and in (4.6) and (4.9), at $\rho_v=1$.

The function in (4.11) involving the Wronskian matrix elements can be computed from (2.14) and one finds, for instance, that the function in the seventh row on the r.h.s. is

$$w_{0l}(\tau_{v'})w_{00}(\tau_v)-w_{00}(\tau_{v'})w_{0l}(\tau_v)=\frac{1}{2}\left\{\frac{g_0(\tau_{v'}-\tau_v)}{\cosh g_0\tau_{v'}\cosh g_0\tau_v}+\frac{\sinh g_0\tau_{v'}}{\cosh g_0\tau_v}-\frac{\sinh g_0\tau_v}{\cosh g_0\tau_{v'}}\right\};\tag{4.12}$$

hence if we consider (4.6)–(4.10) we note that the integrals over τ_v involve functions that can be written, for $\rho=\rho_v$, $\tau=\tau_v$, $\tau'=\tau_{v'}$ and for suitable coefficients $c_j(\rho,\alpha,v)$, ($\rho=1$ if we consider (4.6), (4.9) and $\rho=0$ if we consider (4.7), (4.10)),

$$\sum_{\alpha=-1}^{2}T_\rho^{(\alpha)}(\rho\tau',\tau)Y^{(\alpha)}(\tau',\tau)c_j(\rho,\alpha,v),\tag{4.13}$$

where $Y^{(\alpha)}(\tau',\tau)$ are given, if $x=e^{-\sigma g_0\tau}$ and $x'=e^{-\sigma g_0\tau'}$, by

$$\begin{aligned}
Y^{(-1)}(\tau',\tau)&=\frac{1}{2}\frac{\sinh g_0\tau}{\cosh g_0\tau'}\exp[in\varphi^0(\tau)]=\sum_{k'=1}^{\infty}\sum_{k=-1}^{\infty}y_n^{(-1)}(k',k)x'^{k'}x^k,\\
Y^{(0)}(\tau',\tau)&=\frac{1}{2}\frac{\exp[in\varphi^0(\tau)]}{\cosh g_0\tau'\cosh g_0\tau}=\sum_{k'=1}^{\infty}\sum_{k=1}^{\infty}y_n^{(0)}(k',k)x'^{k'}x^k,\\
Y^{(1)}(\tau',\tau)&=\frac{1}{2}\frac{\sinh g_0\tau'}{\cosh g_0\tau}\exp[in\varphi^0(\tau)]=\sum_{k'=-1}^{\infty}\sum_{k=1}^{\infty}y_n^{(1)}(k',k)x'^{k'}x^k,\\
Y^{(2)}(\tau',\tau)&=\exp[in\varphi^0(\tau)]=\sum_{k=0}^{\infty}\tilde{y}_n^{(2)}(0,k)x^k,\quad k'\equiv 0;
\end{aligned}\tag{4.14}$$

(with k' odd in the first three functions), which define the coefficients $y_n^{(\alpha)}(k',k)$ for $\alpha=-1,0,1,2$ (it is easily checked that k' is *odd* in the first three relations) and we set, for $\alpha=-1,0,1,2,$

$$T_\rho^{(\alpha)}(\rho\tau',\tau)=\begin{cases}g_0(\tau'-\tau) & \text{if } \alpha \text{ is either 0 or 2 and } \rho=1,\\ g_0\tau & \text{if } \alpha \text{ is either 0 or 2 and } \rho=0,\\ 1 & \text{if } \alpha \text{ is either } -1 \text{ or } 1.\end{cases}\tag{4.15}$$

Likewise we shall set, defining the coefficients $\tilde{y}_n^{(\alpha)}(k',k)$, for $\alpha=-1,0,1$, and $\overline{y}_n^{(-1)}(k',k)$,

$$\begin{aligned}
\tilde{Y}^{(\alpha)}(\tau',\tau)&=-\tanh g_0\tau' Y^{(\alpha)}(\tau',\tau)\stackrel{\text{def}}{=}\sum_{k'=-\alpha}^{\infty}\sum_{k=\alpha}^{\infty}\tilde{y}_n^{(\alpha)}(k',k)x'^{k'}x^k,\quad \alpha=\pm 1,\\
\tilde{Y}^{(0)}(\tau',\tau)&=-\tanh g_0\tau' Y^{(0)}(\tau',\tau)\stackrel{\text{def}}{=}\sum_{k'=1}^{\infty}\sum_{k=1}^{\infty}\tilde{y}_n^{(0)}(k',k)x'^{k'}x^k,\\
\tilde{Y}^{(2)}(\tau',\tau)&=Y^{(2}(\tau',\tau)\stackrel{\text{def}}{=}\sum_{k=1}^{\infty}\tilde{y}_n^{(2)}(0,k)x^k,\\
\overline{Y}^{(1)}(\tau',\tau)&=\frac{\cosh g_0\tau'}{\cosh g_0\tau}\exp[in\varphi^0(\tau)]\stackrel{\text{def}}{=}\sum_{k'=-1}^{\infty}\sum_{k=1}^{\infty}\overline{y}_n^{(1)}(k',k)x'^{k'}x^k,\\
\tilde{T}_1^{(0)}(\tau',\tau)&=g_0(\tau'-\tau),\quad \tilde{T}_1^{(2)}(\tau',\tau)\equiv 1,\quad \tilde{T}_0^{(0)}(0,\tau)=T_0^{(0)},
\end{aligned}\tag{4.16}$$

where k' is odd except that for $\tilde{Y}^{(2)}(\tau',\tau)$; in all other cases the $T,\tilde{T},\overline{T}$-functions will be defined 1 (no matter which is the value of the labels that we attribute to them); this is done to uniformize the notation.

The label k will be called the *incoming hyperbolic mode* and k' the *outgoing hyperbolic mode* for reasons that become clear by contemplating (4.19) below.

In terms of (4.14)–(4.16) the functions (4.11) multiplied by $\exp[in_v\varphi^0(\tau_v)]$ can be expressed as in (4.13), thus defining implicitly the coefficients $c_j(\rho,\alpha,v)$ in (4.13),

$$
\begin{aligned}
& w^0_{j_{\lambda_v}}(\tau_{v'},\tau_v)\exp[in_v\varphi^0(\tau_v)] \\
&\quad = \begin{cases} T_0^{(0)}(0,\tau_v)Y^{(0)}(\tau_{v'},\tau_v)+Y^{(-1)}(\tau_{v'},\tau_v), & j_{\lambda_v}=j_v-l=0, \\ T_0^{(2)}(0,\tau_v)Y^{(2)}(\tau_{v'},\tau_v), & 0<j_{\lambda_v}=j_v-l<l, \end{cases} \\
& w^0_{j_{\lambda_{v_0}}}(t,\tau_{v_0})\exp[in_{v_0}\varphi^0(\tau_{v_0})] \\
&\quad = \begin{cases} T_0^{(0)}(0,\tau_{v_0})Y^{(0)}(t,\tau_{v_0})+Y^{(-1)}(t,\tau_{v_0}), & j_{\lambda_{v_0}}=j_{v_0}-l=0, \\ T_0^{(2)}(0,\tau_{v_0})Y^{(2)}(t,\tau_{v_0}), & 0<j_{\lambda_{v_0}}=j_{v_0}-l<l, \\ \tilde{T}_0^{(0)}(0,\tau_{v_0})\tilde{Y}^{(0)}(t,\tau_{v_0})+\tilde{Y}^{(-1)}(t,\tau_{v_0}), & j_{v_0}=j_{\lambda_{v_0}}=l, \\ 0, & j_{v_0}=j_{\lambda_{v_0}}>l, \end{cases} \\
& w^1_{j_{\lambda_v}}(\tau_{v'},\tau_v)\exp[in_v\varphi^0(\tau_v)] \\
&\quad = \begin{cases} T_1^{(0)}(\tau_{v'},\tau_v)Y^{(0)}(\tau_{v'},\tau_v)+Y^{(1)}(\tau_{v'},\tau_v) \\ \quad -Y^{(-1)}(\tau_{v'},\tau_v), & j_{\lambda_v}=j_v-l=0, \\ T_0^{(2)}(\tau_{v'},\tau_v)Y^{(2)}(\tau_{v'},\tau_v), & 0<j_{\lambda_v}=j_v-l<l, \end{cases} \\
& w^1_{j_{\lambda_{v_0}}}(t,\tau_{v_0})\exp[in_{v_0}\varphi^0(\tau_{v_0})] \\
&\quad = \begin{cases} T_1^{(0)}(t,\tau_{v_0})Y^{(0)}(t,\tau_{v_0})+Y^{(1)}(t,\tau_{v_0}) \\ \quad -Y^{(-1)}(t,\tau_{v_0}), & j_{\lambda_{v_0}}=j_{v_0}-l=0, \\ T_1^{(2)}(t,\tau_{v_0})Y^{(2)}(t,\tau_{v_0}), & 0<j_{\lambda_{v_0}}=j_{v_0}-l<l, \\ \tilde{T}_1^{(0)}(t,\tau_{v_0})\tilde{Y}^{(0)}(t,\tau_{v_0})+\tilde{Y}^{(1)}(t,\tau_{v_0})+ \\ \quad -\tilde{Y}^{(-1)}(t,\tau_{v_0}))+\overline{Y}^{(1)}(t,\tau_v), & j_{v_0}=j_{\lambda_{v_0}}=l, \\ \tilde{T}_1^{(2)}(t,\tau_{v_0})\tilde{Y}^{(2)}(t,\tau_{v_0}), & j_{v_0}=j_{\lambda_{v_0}}>l. \end{cases}
\end{aligned}
\tag{4.17}
$$

One could avoid introducing the $\tilde{T}$ functions as they are simply related to the T functions or are just identically 1: however it is convenient to introduce them to make the above formulas more symmetric and therefore easier to keep in mind while working with.

Finally, we define the coefficients $\xi_j(k',0)$ by the power series expansion,

$$\frac{1}{\cosh g_0\tau'}=\sum_{k'=1}^{\infty}\xi_l(k',0)x'^{k'},\quad k'\geqslant 1,\ \text{odd},$$
$$1=\xi_j(0,0),\quad j>l, \tag{4.18}$$

where $x'=e^{-\sigma g_0\tau'}$ and k' is odd, which occurs as coefficient $w_{00}(\tau')$ in (4.7) (when $\rho_v=0$, i.e., $v\in\Theta_L$).

The above definitions (taken from (42) and (45) in Ref. 5) suffice to discuss the whiskers (and therefore the splitting in the action variables).

The (4.13) allow us to introduce a ''relatively simple notation;'' we can add to each node a *badge* label $\alpha_v=(-1,0,1,2)$ that will distinguish which choice we make between the possibilities in (4.14) and (4.16) and two *hyperbolic mode* labels k'_v, k_v which select which particular term we choose in the sums in (4.14) and (4.16); they are integer numbers $\geqslant -1$. We shall not have to introduce labels to distinguish terms coming from the expansions of $Y^{(\alpha)}$, $\tilde{Y}^{(\alpha)}$, $\overline{Y^{(\alpha)}}$ bearing the same badge α because one can check that the labels α_v together with j_v and v itself uniquely determine which choice has to be made.

In terms of the latter labels we can define a *hyperbolic momentum* of a line λ_v as a label $p(v)\in\mathbb{Z}$ which will be the sum of all the hyperbolic modes of the nodes that precede v *plus* the incoming hyperbolic mode of the node v itself: this is the sum of the labels k_w associated with all free nodes $w\leqslant v$, with v included, and of the labels k'_w associated with all the *free* nodes $w<v$ or *highest* nodes of the leaves $w<v$, with v *not* included,

$$p(v)=k_v+\sum_{\substack{w\in\vartheta_f\\ w<v}}(k_w+k'_w)+\sum_{\substack{w\in\Theta_L\\ w<v}}k'_w. \tag{4.19}$$

A *very important property* is that $k_w+k'_w\geqslant 0$, by (4.14) and (4.16), and $k'_w\geqslant 0$ if $w\in\Theta_L$, by (4.18), so that $p(v)\geqslant -1$. Furthermore if $p(v)=0$ then *either* $k_v=-1$ and $k_w+k'_w=0$ for all $w<v$ except one single node $\tilde{w}<v$ (which is either in ϑ_f or Θ_L) for which $k_{\tilde{w}}+k'_{\tilde{w}}=1$, *or* $k_v=0$ and $k_w+k'_w=0$ for all $w<v$. If $p(v)=-1$ then $k_v=-1$ and $k_w+k'_w=0$ for all $w<v$.

In the above analysis we have not taken explicitly into account the possibility of contributions to $\Phi_+^{h\sigma}$ coming from the third line in (2.11), i.e., counterterm contributions. They are, of course, possible and they can be taken immediately into account in the graphical representation by considering the nodes with a label $\delta_v=0$ and adding to them a *counterterm label* κ_v, a non-negative integer. If $\kappa_v=0$ this will mean that the node represents a contribution from the first line of the definition of $\Phi_+^{h\sigma}$, i.e., a contribution that is unrelated to the counterterms, while if $\kappa_v\geqslant 1$ the node represents a contribution from the term with $p=\kappa_v$ in the second line contribution to $\Phi_+^{h\sigma}$ in (2.11).

4.3. The trees carry, at this point, quite a few decorating labels and each tree together with all its labels will represent a ''very simple'' contribution to the value of the hth order coefficient in the Taylor expansion in ε (at fixed η of course) of the $\Xi_{h\sigma}$ vector. Very simple means that the improper integrals that correspond to each term are very easy to evaluate explicitly and lead to a result that can be expressed as a product of factors determined by the labels of the tree and associated with the nodes and with the lines, see (4.30), below. We list here the set of labels that have been introduced,

j_v	action labels,
j_{λ_v}	branch labels,
δ_v	order labels,
ρ_v	leaf labels,
ν_v	mode labels,
$\boldsymbol{\nu}(v)$	momentum in the branch λ_v following v,

$\boldsymbol{\nu}_0(v)$	free momentum in the branch λ_v following v,
α_v	badge labels,
(k'_v,k_v)	hyperbolic mode labels,
$p(v)$	hyperbolic momentum in the branch λ_v following v,
κ_v	counterterm labels.

There are some constraints between the labels, which follow from the rules stated in Secs. IV, 4.1 and IV, 4.2 and from the choice of the counterterms (the latter will be discussed in Sec. IV, 4.5 below),

(1) one has $j_{\lambda_v}=j_v-l$ if $v<v_0$ and $j_{\lambda_v}=j_v$ or $j_{\lambda_v}=j_v-l$ if $v=v_0$ (see Sec. IV, 4.1);
(2) if $\rho_v=0$ then $\alpha_v\neq 1$, (see (4.17));
(3) if $j_{\lambda_v}\neq 0$, l, then $\alpha_v=2$, otherwise if $j_{\lambda_v}=0,l$, then α_v can be -1, 0, 1, (see (4.17));
(4) $\delta_v=0$ implies $j_v=l$ (by the $\boldsymbol{\alpha}$-independence of f_0);
(5) k_v, k'_v, $p(v)\geqslant -1$, (see (4.14), (4.16) and comment following (4.19));
(6) $(p(v),\boldsymbol{\nu}_0(v))\neq(0,\mathbf{0})$, see Remark 4.6 below.

In terms of such labels, given a decorated tree ϑ_0 with m_0 nodes and with highest node v_0, we can define the *value* of a *subtree* ϑ with m free nodes, highest node w (preceding the highest node v_0 of ϑ_0: $w\leqslant v_0$) and label $j_{\lambda_w}=0,...,l-1$, $\rho_w=0,1$. It will be given by the expression

$$\mathrm{Val}(\vartheta)=\left[\prod_{\substack{v\in\vartheta_f\\ v\leqslant w}}\int_{\sigma\infty}^{\rho_v g_0\tau_{v'}}dg_0\tau_v\mathcal{V}_v(\vartheta)\right]\left[\prod_{v\in\Theta_L}\mathcal{L}_v(\vartheta)\right]\left[\prod_{\substack{v\in\vartheta_f\\ \delta_v=0}}\gamma_{\kappa_v}(g_0)\right], \tag{4.20}$$

where the integration is the improper integration $\mathcal{I}$ (in the sense of Sec. II, 2.4), the tree ϑ consists of a ''free'' m-nodes tree ϑ_f with leaves attached to a (possibly empty) subset of the nodes of ϑ_f, and the following notation has been used:

(i) The coefficients $\mathcal{V}_v(\vartheta)$ and $\mathcal{L}_v(\vartheta)$ are described by the collection of labels enumerated above. They can be written, respectively, as

$$\mathcal{V}_v(\vartheta)=\overline{F}_{\underline{\nu}_v}\hat{T}^{(\alpha_v)}_{\rho_v}(\rho_v\tau_{v'},\tau_v)e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_v\tau_v}x_v^{k_v}\prod_{\substack{w\in\vartheta\\ w'=v}}x_v^{k'_w}(-1)^{\delta_{\alpha_v,-1}}\hat{y}^{(\alpha_v)}_{n_v}(k'_v,k_v), \tag{4.21}$$

and

$$\mathcal{L}_v(\vartheta)=\xi_{j_v}(k'_v,0)L^{h_v\sigma}_{j_v\boldsymbol{\nu}(v)}(\vartheta), \tag{4.22}$$

where $x_v=\exp[-\sigma g_0\tau_v]$ and $\hat{y}^{(\alpha_v)}_{n_v}$, $\hat{T}^{(\alpha_v)}_{\rho_v}$ are (see (4.17)) either

(a) $y^{(\alpha_v)}_{n_v}$, $T^{(\alpha_v)}_{\rho_v}$, if either $v<v_0$ or $v=v_0$ and $j_{\lambda_{v_0}}=j_{v_0}-l$, or
(b) $\tilde{y}^{(\alpha_v)}_{n_v}$ or $\overline{y}^{(1)}_{n_v}$ and $\tilde{T}^{(\alpha_v)}_{\rho_v}$ or $\overline{T}^{(\alpha_v)}_{\rho_v}$, if $v=v_0$ and $j_{\lambda_{v_0}}=j_{v_0}$.

Furthermore $\rho_v=1$ if $v<w$, while ρ_w can be either 0 or 1; j_{λ_w} can have any value $0,...,2l-1$ if $w=v_0$, in any other case $j_{\lambda_v}=0,...,l-1$ (see above). In (4.21)

$$\overline{F}_{\underline{\nu}_v}=\left(\frac{J_0}{J}\right)^{(1-\delta_{j_v,l})(1-\delta_{j_v,j_{\lambda_v}})}f^{\delta_v}_{\underline{\nu}_v}\left[(-i\nu_v)_{j_v-l}\prod_{\substack{w\in\vartheta\\ w'=v}}(i\nu_v)_{j_w-l}\right] \tag{4.23}$$

depends on the labels $(\delta,\underline{\nu},j)$ of the node v and of its predecessors w's (recall that by (2.2) $\underline{\nu}_v=(n_v,\boldsymbol{\nu}_v)$); in (4.22) the quantity $L^{h_v\sigma}_{j_v\boldsymbol{\nu}(v)}(\vartheta)$ is called the ''*value of the leaf*'' v of order h_v (see item (v) below for its definition). The matrix J is not, in general, a multiple of the identity and J_0J^{-1} will be interpreted as acting on the rotator components of $\underline{\nu}_v$ (and it will be 1 when raised to the power 0).

(i) For the purposes of the cancellations analysis performed in Appendix C, the exact form of a few coefficients among the $y_{n_v}^{(\alpha_v)}(k_v',k_v)$'s turns out to be essential, so that we list them here,

$$\begin{gathered} y_{n_v}^{(-1)}(2,-1)=0,\quad y_{n_v}^{(-1)}(1,-1)=\sigma/2,\quad y_{n_v}^{(-1)}(1,0)=2in_v,\\ y_{n_v}^{(1)}(0,1)=0,\quad y_{n_v}^{(1)}(-1,1)=\sigma/2,\quad y_{n_v}^{(1)}(-1,2)=2in_v,\\ y_{n_v}^{(2)}(1,0)=0,\quad y_{n_v}^{(2)}(0,0)=1,\quad y_{n_v}^{(2)}(0,1)=4in_v\sigma. \end{gathered}\tag{4.24}$$

The coefficients $\tilde{y}^{(-1)}(1,-1)$, $\tilde{y}^{(-1)}(1,0)$, $\tilde{y}^{(1)}(-1,1)$ and $\overline{y}^{(1)}(-1,2)$ are equal to the corresponding (i.e., with the same values of the labels k',k) $y^{(\alpha)}(k',k)$ coefficients.

(ii) The value of a leaf with highest node v in (4.22) is *not* the same as the value $L_{j_v\boldsymbol{\nu}(v)}^{h_v\sigma}$ in Ref. 5; this is because of the above mentioned change in notation (see the sixth item in Sec. IV 1). In Ref. 5 leaf values are defined as sums of the values of all leaves (in the sense we use now) with fixed order, action label and total momentum. Then the leaf value considered here, $L_{j_v\boldsymbol{\nu}(v)}^{h_v\sigma}(\vartheta)$, is a single contribution to the $L_{j_v\boldsymbol{\nu}(v)}^{h_v\sigma}$ of Ref. 5, and depends *only* on the part of the tree ϑ consisting of the nodes $w\leqslant v$; if we call ϑ_v such a subtree, we can write (*temporarily, just for the purposes of comparison*) the present definition of leaf value as $L_{j_v\boldsymbol{\nu}(v)}^{h_v\sigma}(\vartheta)\equiv\overline{L}_{j_v\boldsymbol{\nu}(v)}^{h_v\sigma}(\vartheta_v)$ (as it depends only on the labels of the subtree ϑ_v). In order to make a link between the different notations note that $L_{j\boldsymbol{\nu}}^{h\sigma}$ in Ref. 5 would be, with our present notations, just the sum

$$L_{j\boldsymbol{\nu}}^{h\sigma}\overset{\text{def}}{=}\sum_{\substack{\vartheta_{v_0}\in\mathcal{T}_{\boldsymbol{\nu},h}\\ j_{v_0}=j}}\overline{L}_{j\boldsymbol{\nu}}^{h\sigma}(\vartheta_{v_0}),\tag{4.25}$$

where ϑ_{v_0} is the part of the tree ϑ on which the leaf value really depends and $\mathcal{T}_{\boldsymbol{\nu},h}$ is defined after (4.27) below.

Coming back to our notations we define the *leaf value* $L_{j\boldsymbol{\nu}}^{h\sigma}(\vartheta)$, with $j=j_{v_0}$, (where the the first and third of (4.11) should be used), to be the value of a tree ϑ with $j_{\lambda_{v_0}}=j_{v_0}-l$ and $\rho_{v_0}=0$.

(iii) By construction (see (2.11) and corresponding comments), and if Θ_L is the set of highest nodes in the leaves, the total perturbation order k of ϑ is

$$k=\sum_{\substack{v\in\vartheta_f\\ \delta_v=0}}\kappa_v+\sum_{v\in\vartheta_f}\delta_v+\sum_{v\in\Theta_L}h_v=\sum_{\substack{v\in\vartheta\\ \delta_v=0}}\kappa_v+\sum_{v\in\vartheta}\delta_v,\quad m<2k.\tag{4.26}$$

(iv) Both the counterterms and the leaf values of a given perturbation order are recursively defined in terms of the same quantities with lower orders. In fact $\gamma_\kappa(g_0)$ admits a graphical representation as sum of tree values defined as in (4.20) with the difference that the integration operation corresponding to the highest node of the tree has to be suitably modified (see (4.32) below).

If Val(ϑ) is defined as in (4.20) (and in item (iv) above) then, by construction, one has

$$\Xi_j^{h\sigma}(t;\boldsymbol{\alpha})=\sum_{\boldsymbol{\nu}\in\mathbb{Z}^{l-1}}\Xi_{j\boldsymbol{\nu}}^{h\sigma}(t)e^{i\boldsymbol{\nu}\cdot\alpha},\quad \Xi_{j\boldsymbol{\nu}}^{h\sigma}(t)=\frac{1}{m_0!}\sum_{\substack{\vartheta_0\in\mathcal{T}_{\boldsymbol{\nu},h}\\ j_{\lambda_{v_0}}=j}}\text{Val}(\vartheta_0),\tag{4.27}$$

where $\mathcal{T}_{\boldsymbol{\nu},h}$ is the collection of all trees with total momentum $\boldsymbol{\nu}$ and order h. In (4.27) $m_0!^{-1}$ is a combinatorial factor, which depends on the way we count trees: the simplest is to think that the tree branches of ϑ are pairwise distinct and are distinguished by a label $1,2,\dots,m_0$, if m_0 is the number of nodes in the tree ϑ. In the latter case, which corresponds to our choice, the factor is simply $m_0!^{-1}$, see Refs. 3, 5, provided we regard as identical two trees that can be overlapped by pivoting the branches entering a node (rigidly, together with the subtree attached to them) around any node; as in Ref. 3 we shall call *numbered trees* the trees so counted.

4.4 *Remark:* Since the value of any leaf with highest node v depends only on the labels of the nodes $w \leqslant v$, Eq. (4.20) factorizes into a product of leaf values times a product of counterterms (whose value, so far arbitrary, has still to be specified and it will be, in the analysis between (4.31) and (4.32) when intervening compatibility requirements will dictate its value) times a factor

$$\left[\prod_{\substack{v\in\vartheta_f\\ v\leqslant w}} \int_{\sigma\mu}^{\rho_v g_0 \tau_{v'}} dg_0 \tau_v \mathcal{V}_v(\vartheta)\right]\left[\prod_{v\in\Theta_L} \xi_v(k'_v,0)\right], \tag{4.28}$$

which does not depend on the leaves.

4.5. The extra effort with respect to the approach without counterterms developed in Refs. 3 and 10, gives here (as in Ref. 5) a reward; few combinations of powers of the ‘‘times’’ τ_v appear in the integrand in (4.21). The time variables, by (4.21) and (4.14)–(4.17), appear only via exponentials like

$$e^{-\sigma(\tau_v-\tau_{v'})a} \quad \text{or} \quad (\tau_{v'}-\tau_v)e^{-\sigma(\tau_v-\tau_{v'})a}, \tag{4.29}$$

for some complex $a=(g_0 p(v)-i\sigma\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v))$, yielding, respectively, upon integration, a^{-1} or a^{-2}. Note also that by the hierarchical structure of the trees one has $\sigma(\tau_{v'}-\tau_v)\geqslant 0$.

This greatly simplifies the actual performance of the integration operations which, once one gets familiarity with the formalism, are trivial. One can say that the absence of high powers of the τ_v's is due to having *a priori* fixed the Lyapunov exponent g_0 by means of the counterterms (by contrast in Refs. 3, 10 arbitrary powers of τ_v appeared because g_0 is *not* fixed *a priori*).

Of course the triviality of the integrations is entirely due to the above *very fine* decomposition, into terms identified by labeled trees, of the more compact (but ‘‘difficult’’ to integrate) integrands appearing in (4.6)–(4.12) and in the middle terms in (4.14).

Once all the integration operations will have been performed, the tree value in (4.20) will become a product of ‘‘*factors*,’’ in complete analogy with what one is accustomed to find when defining *Feynman graphs* in Field Theory. The factors are associated with the nodes v and with the branches λ_v. The value of a tree ϑ will then be *defined* as

$$\begin{aligned}\mathrm{Val}(\vartheta)=e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v_0)t-\sigma g_0[k'_{v_0}+p(v_0)]t}&\left[\prod_{\substack{\lambda_v\in\vartheta\\ v\in\vartheta_f}}\left(-\frac{\sigma g_0}{g_0 p(v)-i\sigma\boldsymbol{\omega}'-\boldsymbol{\nu}_0(v)}\right)^{r_v}\right]\\ &\cdot\left[\prod_{v\in\vartheta_f}\overline{F}_{\underline{\nu}_v}(-1)^{\delta_{\alpha_v,-1}}y_{n_v}^{(\alpha_v)}(k'_v,k_v)\right]\\ &\cdot\left[\prod_{v\in\Theta_L}\xi_{j_v}(k'_v,0)L_{j_v\nu}^{h_v\sigma}(\vartheta)\right]\left[\prod_{\substack{v\in\vartheta_f\\ \delta_v=0}}\gamma_{\kappa_v}(g_0)\right],\end{aligned} \tag{4.30}$$

where r_v is either 1 or 2, and the case $(\boldsymbol{\nu}_0(v),p(v))=(\mathbf{0},0)$ has to be *excluded* for any node $v \in \vartheta_f$. This is not to claim that no trees with $\boldsymbol{\nu}_0(v)=\mathbf{0}$, $p(v)=0$ can be drawn by following the above rules; this is a *further rule* to impose on the labels in order that the analysis does not become contradictory requires fixing the function $\gamma(\varepsilon,g_0)$ conveniently; the consistence criterion determines $\gamma(\varepsilon,g_0)$ uniquely. This rule is a natural extension of the corresponding rule holding in perturbation theory of KAM tori, which was discussed by Lindstedt and Newcomb for the lowest orders of the perturbation expansions and which was proved to hold at all orders by Poincaré;[15] see the last paragraph in Sec. II, 2.4 above and the Remark 4.6, (c), below.

The values of the numbers r_v arise from the time variables integrals via the mechanism just illustrated above (whereby one either gets a^{-1} or a^{-2} from the integration of the functions (4.29)).

The factors $-(\sigma g_0)^{r_v}[g_0 p(v)-i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]^{-r_v}$, associated with the branches, will be called *propagators* or *small divisors*. The first name arises from the possible interpretation of the trees as Feynman graphs of a suitable field theory, see Ref. 17; the second name corresponds to the usual name given in Mechanics to such expressions generated by perturbation expansions.

It can be useful to write, if v_0 is the highest node of ϑ and $j_{\lambda_{v_0}}=0$,

$$\mathrm{Val}(\vartheta)=e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v_0)t-\sigma g_0[k'_{v_0}+p(v_0)]t}\left(-\frac{\sigma g_0}{g_0 p(v_0)-i\sigma\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v_0)}\right)^{r_{v_0}}\overline{\mathrm{Val}}(\vartheta), \tag{4.31}$$

so defining the quantity $\overline{\mathrm{Val}}(\vartheta)$ (this is a well known kind of operation on Feynman graphs, which associates with a graph another value gruesomely called the value of the *amputated* graph, amputated tree in our case). Moreover we can define $\overline{\mathrm{Val}}(\vartheta)$ also for $(\boldsymbol{\nu}_0(v_0),p(v_0))=(\mathbf{0},0)$ as no vanishing denominator appears in its expression. It is however clear that nodes with $(\boldsymbol{\nu}_0(v),p(v_0))=(\mathbf{0},0)$ must not appear at all in the trees that we consider, for (4.30) to make sense as it is written. This implies, not surprisingly, a consistence problem, namely, one has to check that the sum of all the $\overline{\mathrm{Val}}(\vartheta)$ over trees of a given order and with $(\boldsymbol{\nu}_0(v_0),p(v_0))=(\mathbf{0},0)$ cancel so that lines λ_v with $(\boldsymbol{\nu}_0(v),p(v_0))=(\mathbf{0},0)$ never appear, neither for $v=v_0$ nor for $v<v_0$.

The cancellation is made possible because we still have freedom to fix the counterterms and *their choice is in fact uniquely determined by the conditions that they be such that the needed cancellation takes place*. The quantities $\overline{\mathrm{Val}}(\vartheta)$ are convenient in order to find and to express the counterterms and also the ''resonance values'' introduced later (see Appendix C). One checks, see Appendix A, that the counterterms can be explicitly written, if $\mathcal{T}_{\boldsymbol{\nu},h}$ is the collection of all trees with total momentum $\boldsymbol{\nu}$ and order h, as

$$\gamma_k(g_0)=-\tfrac{1}{2}\sum^{*}_{\substack{\vartheta\in\mathcal{T}_{0,\kappa},\alpha_{v_0}=-1\\ p(v_0)=0,\nu_0(v_0)=0,k'_{v_0}=1}}\overline{\mathrm{Val}}(\vartheta), \tag{4.32}$$

and the $*$ means that the sum is further restricted so that the tree contains no leaves. This choice being simply imposed by the requirement that no contribution with $(\boldsymbol{\nu}_0(v_0),p(v_0))=(\mathbf{0},0)$ can arise for $\Xi_{-}^{h\sigma}(t)$; see Appendix A.

4.6 *Remarks:* (a) The presence of the counterterms will manifest itself not only through the elimination of the trees whose value would be meaningless if evaluated via (4.30) but also, and mainly, in the fact that the elements of the algebra $\hat{\mathcal{M}}$ met in the successive integrations have a special form (namely always like one of the (4.29)) which implies that the result of the improper integrals is *not* as complicated as one could fear from (2.24). This leads to the simple expression (4.32) (see Appendix A1 of Ref. 3 for what would otherwise happen without counterterms).

(b) From (4.31) one sees that if $j_{\lambda_{v_0}}>l$ it is natural to collect together the terms with $p(v_0)=0$: for them, since $j_{\lambda_{v_0}}>l$, in (4.30) one must have $k'_{v_0}+p(v_0)=0$ by the last of (4.14). Note also that by (2.26) the case $(\boldsymbol{\nu}_0(v),p(v_0))=(\mathbf{0},0)$ is excluded. If $j_{\lambda_{v_0}}\leqslant l$ we, likewise, collect the terms with $k'_{v_0}+p(v_0)=0$ and, for similar reasons the term with $k'_{v_0}+p(v_0)=-1$ cannot be present (see again (4.14) and (4.16), and use $p(v_0)\geqslant -1$ supplemented by the relations between the labels $p(v_0)$ and k'_{v_0} which will be exhibited in Sec. V 1).

Hence the cases with $p(v_0)+k'_{v_0}=-1$ are excluded by construction [the initial data $\Xi^{h\sigma}(0,\boldsymbol{\alpha})$ were determined precisely by imposing boundedness at $\sigma t=+\infty$, i.e., by imposing the absence of divergent terms in the expansion in powers of $x=e^{-g_0\sigma t}$ which would correspond to the terms with $p(v_0)=-1$] and we see that the sum of the values of the trees with $p(v_0)+k_{v'_0}=0$ give us the equations for the actions and the angles of the invariant torus to which the

TABLE I. Possible cases when $p(v)=0,-1$.

$p(v)$	k_v	k'_v	α_v	j_v
-1	-1	odd≥ 1	-1	l
0	-1	odd≥ 1	-1	l
0	0	≥ 1	-1	l
0	0	≥ 0	2	$>l$

whiskers considered are asymptotic; the terms with $p(v_0)+k_{v'_0}=0$ asymptote to quasiperiodic functions of $\boldsymbol{\omega}t$ so that replacing $\boldsymbol{\omega}t$ by $\boldsymbol{\psi}\in T^{l-1}$ one gets a parameterization of the points on the tori in terms of a point $\boldsymbol{\psi}\in T^{l-1}$ on a ''standard torus.''

And the terms with $k'_{v_0}+p(v_0)=1$ provide the leading corrections. Since such terms are present already to order 0 (as one sees from the expression of the pendulum separatrix) the distance between a point moving on the stable manifold of the torus and the torus itself will be proportional to $x=e^{-g_0\sigma t}$ as $\sigma t\to\infty$ so that g_0 has the interpretation of Lyapunov exponent of the invariant torus; see (2.26) in Sec. II, 2.6.

(c) Summarizing, *the case* $(\boldsymbol{\nu}_0(v),p(v))=(\mathbf{0},0)$ *has to be ruled out as a consequence of* (2.25) *and of* (4.32), *respectively, for the contributions to* $\Xi^{h\sigma}_{\uparrow}$ *and to* $\Xi^{h\sigma}_{+}$ *(see the last constraint listed at the beginning of Sec. IV, 4.3). All cases with* $k'_{v_0}+p(v_0)=-1$ *are also excluded.*

V. BOUNDS

We now discuss how to bound the value of a tree or of a sum of a small number of trees which we take for simplicity without leaves and without counterterms. The more general case will be eventually reduced, see below, to the one we consider here.

We shall discuss first how to bound values of trees without leaves and counterterms such that $p(v_0)=0$ if v_0 is the highest node; hence we shall consider trees, always without leaves and counterterms, with $p(v_0)=0$. At the end we shall see how the presence of leaves and counterterms modifies the analysis.

The following discussion is ''locally'' simple, but ''globally'' delicate and repeats that in Ref. 5, Sec. IV, the conclusions are also summarized in the Tables I–IV.

From (4.19) it follows that the hyperbolic momentum $p(v)$ is $p(v)\geq -1$ and, as remarked after (4.19), $p(v)=0$ can occur only in special cases; more precisely if $p(v)=0$, then k_v is either -1 or 0, and

(1) if $k_v=0$, all free nodes w preceding v (whether immediately or not) have $k'_w+k_w=0$, while
(2) if $k_v=-1$, all free nodes w preceding v have $k'_w+k_w=0$, *except* for a single node $\tilde{w}<v$ such that $k'_{\tilde{w}}+k_{\tilde{w}}=1$.

In the latter case we call $\mathcal{P}$ the path of nodes (i.e., the totally ordered set of nodes) which connect v to $\tilde{w}$, both extremes included (see also Ref. 5, Sec. IV).

Supposing $p(v)=-1,0$ and recalling that $\alpha_v<2$ implies k'_v odd, the expansions (4.14) impose that there are *very few possible choices* of the values of the hyperbolic modes at $w\leq v$;

TABLE II. Cases $p(v)=0$, $w\notin\mathcal{P}$.

α_w	-1	0	1	2
(k'_w,k_w)	$(1,-1)$	impossible	$(-1,1)$	(0,0)
p(w)	-1	impossible	1	0
j_w	l	impossible	l	$>l$

TABLE III. Cases $p(v)=0$, $w\in\mathcal{P}$, $w>\tilde{w}$.

α_w	-1	0	1	2
(k'_w,k_w)	$(1,-1)$	impossible	$(-1,1)$	$(0,0)$
$p(w)$	0	impossible	2	1
j_w	l	impossible	l	$>l$

(1) if there is no path or there is a path linking v to $\tilde{w}$ but $w\neq\tilde{w}$ and $w\notin\mathcal{P}$, then $k_w+k_{w'}=0$ and the cases $\alpha_w=1$, $\alpha_w=-1$ and $\alpha_w=2$ require, respectively, $(k'_w,k_w)=(-1,1)$, $(k'_w,k_w)=(1,-1)$ and $(k'_w,k_w)=(0,0)$: correspondingly $p(w)=1$, $p(w)=-1$ and $p(w)=0$. While, for $w\in\mathcal{P}$, the value of $p(w)$ ''increases by one unit,'' i.e., $p(w)=2$, $p(w)=0$ and $p(w)=1$ for $w\in\mathcal{P}$;
(2) if $w=\tilde{w}$ then $k_w+k_{w'}=1$ and the cases $\alpha_{\tilde{w}}=1$, $\alpha_{\tilde{w}}=-1$ and $\alpha_{\tilde{w}}=2$ require, respectively, $(k'_{\tilde{w}},k_{\tilde{w}})=(-1,2)$, $(k'_{\tilde{w}},k_{\tilde{w}})=(1,0)$ and $(k'_{\tilde{w}},k_{\tilde{w}})=(0,1)$ (correspondingly $p(\tilde{w})=2$, $p(\tilde{w})=0$ and $p(\tilde{w})=1$).

Note that in both cases $\alpha_w=0$ is not possible.

The above analysis covers both cases $v<v_0$ and $v=v_0$, as the functions in (4.14) and in (4.16) have the same dependence on τ (hence on k).

The latter properties have to be considered as a further restriction to impose on the tree labels, and play an essential role for the discussion of the cancellations.[5]

Moreover if $p(v)=0$, then

(1) if $k_v=0$ then v can be preceded only by leaves with the highest nodes w having $j_w>l$, because k'_w must be 0 in such a case, so that the second of (4.18) applies;
(2) if $k_v=-1$, then all the leaves again must have the highest node w with $j_w>l$, except at most one leaf with highest node $\tilde{w}$ with $j_{\tilde{w}}=l$ and $k'_{\tilde{w}}=1$.

We extend the definition of path also to the case $p(v)=0$, $k_v=0$, by setting $\mathcal{P}\overset{\text{def}}{=}\emptyset$ if $j_v=l$ and $\mathcal{P}\overset{\text{def}}{=}v$ if $j_v=l$, only for purposes of notational convenience (see below). This is consistent with the above tables and does not change them.

5.2. *Remark:* Note that, if a tree (or subtree) ϑ_0 with the highest node v_0 has a total hyperbolic momentum $p(v_0)=0$, then there is one and only one path $\mathcal{P}$, and, if $\mathcal{P}\neq\emptyset$, then $\mathcal{P}$ connects the node v_0 to some node $\tilde{w}<v_0$. In fact, if there is a path $\mathcal{P}\neq\emptyset$, the node $\tilde{w}$ is so defined that $k_{\tilde{w}}+k'_{\tilde{w}}=1$; then if $k_{v_0}=0$ there cannot be any of such nodes (and $\mathcal{P}=v_0$ in such a case), while if $k_{v_0}=-1$ there must be one and only one such node. This simply follows from the analysis in Sec. V, 5.1 by noting that all nodes $w<v_0$ except $\tilde{w}$ must have $k_w+k'_w=0$.

5.3. The small divisors can be really ''small'' only when $p(v)=0$: if $p(v)\neq 0$, they are bounded by $g_0^{-r_v}$, i.e., by a quantity of order $O(1)$. So one can consider all free nodes in the trees, among the ones having $p(v)=0$, which are closest to the root. All free nodes v between them and the root have propagators which are not small because $|p(v)|\geqslant 1$ (see also Ref. 5, p. 298).

Given a tree ϑ_0 with m free nodes, from each subtree ϑ ending in a node v_0 with $p(v_0)=0$ (here v_0 is some node of ϑ_0: it becomes the highest node of ϑ), one obtains contributions which can be naturally collected together (recall Remark 4.4) into a contribution to the tree value (see (4.30)) consisting of a factor

$$\prod_{v\leqslant v_0}\overline{F}_{\underline{\nu}_v}G_v[\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]y'_v, \tag{5.1}$$

(here the product is over all free nodes preceding v_0) times a product of counterterms $\gamma_{\kappa_v}(g_0)$ for

TABLE IV. Cases $p(v)=0$, $w=\tilde{w}$.

α_w	-1	0	1	2
(k'_w,k_w)	(1,0)	impossible	$(-1,2)$	(0,1)
$p(w)$	0	impossible	2	1
j_w	l	impossible	l	$>l$

$v\in\vartheta_f$ with $\delta_v=0$, times a products of factors $\xi_{j_v}(k'_v,0)L^{h_v\sigma}_{j_v\nu(v)}$, for each $v\in\Theta_L$; see (4.30). The vector $\boldsymbol{\nu}_0(v)$ is the free momentum (defined above; see (4.2)) flowing through the branch λ_v, the coefficients y'_v are related to the expansions (4.14) via

$$y'_v=\begin{cases}\frac{1}{2}[y^{(1)}_{n_v}(-1,1)+y^{(-1)}_{n_v}(1,-1)]=\sigma/2 & \text{if } v\in\mathcal{P},\alpha_v\in\{-1,1\},\\ (-1)^{\delta_{\alpha_v,-1}}y^{(\alpha_v)}_{n_v}(k_{v'},k_v) & \text{otherwise}\end{cases} \tag{5.2}$$

where $\mathcal{P}$ denotes the path in ϑ (there is always one such path, possibly the empty set, because we suppose $p(v_0)=0$); and $G_v[\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]$ is related to the propagator of the branch λ_v, and it will have the form

$$G_v[\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]=\begin{cases} g_0^2[i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]^{-2} & \text{if } v\notin\mathcal{P},\ j_v>l,\\ -\sigma g_0^2[g_0^2+(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v))^2]^{-1} & \text{if } v\notin\mathcal{P},\ j_v=l,\\ g_0^{2-\delta_{j_v,l}}[-\sigma(g_0p(v)-i\sigma\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v))]^{-(2-\delta_{j_v,l})} & \text{if } v\in\mathcal{P},\ \alpha_v\neq-1,\\ g_0[i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]^{-1} & \text{if } v\in\mathcal{P},\ \alpha_v=-1,\end{cases} \tag{5.3}$$

where $p(v)\neq0$ in the third line, because

(a) The first line is such because if $j_v>l$ one has necessarily $\alpha_v=2$, see Tables I–IV, and we have to integrate a function $g_0(\tau_{v'}-\tau_v)e^{in_v\varphi^0(\tau_v)}$ so that $k_v\geq0$; hence $k_v=p(v)=0$ and we have the second function in (4.29) to integrate.

(b) The second line is such because if $v\notin\mathcal{P}$, $j_v=l$ we have $w^1_l(\tau_{v'},\tau_v)e^{in_v\varphi^0(\tau_v)}$ which is a sum of three terms (see the third of (4.17)); the first has $k_v+k_{v'}\geq2$ so is excluded (recall that $p(v_0)=0$ and $v\leq v_0$); while the second only sees the contribution to $Y^{(1)}$ with $k'_v=-1$, $k_v=1$, see (4.14), and the third only contributes by the term with $k'_v=1$, $k_v=-1$ in $Y^{(-1)}$. In the two cases one has $p(v)=1$ or $p(v)=-1$, respectively; adding up together the latter two contributions and using the first of (5.2) to compute the sum of the coefficients we get

$$\frac{-\sigma g_0y^{(1)}_{n_v}(-1,1)}{g_0-i\sigma\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)}-\frac{-\sigma g_0y^{(-1)}_{n_v}(1,-1)}{-g_0-i\sigma\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)}=\frac{\sigma}{2}\,\frac{-2\sigma g_0^2}{g_0^2+(\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v))^2}, \tag{5.4}$$

as it can be read from the coefficients in the intermediate column of (4.24) and from $p(v)=\alpha_v=\pm1$.

(c) The third line of (5.3) is obtained by noting that, if $v\in\mathcal{P}$, $v>\tilde{w}$ one has $p(v)=1+k_v$, so that, if $j_v=l$ and $\alpha_v\neq-1$, then $p(v)>0$, see Table III; if $v=\tilde{w}$ and $\alpha_v\neq-1$, one has $p(v)\neq0$, see Table IV (note that $\alpha_v\neq2,0$ so that we have to consider the first integrand in (4.29)).

If $j_v>l$, then $\alpha_v=2$, and, by the Tables III and IV one has $k_{v'}=0$, $k_v\geq0$ and $k_v+k_{v'}=1$, so that $k_v=1$ and $p(v)=2$; while, if $v=w$, then $k_{v'}=0$, $k_v\geq0$ and $k'_v+k_v=1$ imply $k_v=1$, so that $p(v)=1$. So in both cases $p(v)\geq1$.

(d) The fourth line is found by looking at the Tables III and IV as follows: if $\alpha_v=-1$, $v\in\mathcal{P}$, $v>\tilde{w}$, one has $k_v+k'_v=0$, hence $k_v=-1$, $k_{v'}=1$ and $p(v)=0$; this happens only if $j_v=l$ so that we have to consider the first integral in (4.29) and we get the fourth relation.

This shows that the only trees that do not have a value tending to 0 as $t\to\sigma\infty$, i.e., are those with $p(v_0)+k'_{v_0}=0$ (all the others tend to 0 as a power of $x=e^{-g_0\sigma t}$), have propagators that are even functions of the momenta flowing in them. In fact the observation on the absence of paths preceding v_0 implies that only the first two propagators in (5.3) appear in such trees. Since, as already remarked, the trees with $p(v_0)+k'_{v_0}=0$ give the equations of the tori this is an interesting check that the tori equations so obtained at $t=+\infty$ and $t=-\infty$ do *coincide*. A similar analysis, and check, holds for the cases $j_{\lambda_{v_0}}\leq l$.

5.4. *Remark:* Collecting together the contributions from $\alpha_v=-1$ and $\alpha_v=1$, for $v\notin\mathcal{P}$, is a convenient operation and has nothing to do with the deeper resummations that imply the cancellations necessary for convergence estimates: the systematic use of this operation should be described by adding a label to the trees on the nodes $v\notin\mathcal{P}$ and replacing on the branches which give rise to one of the two propagators in (5.4) the α_v label by the new label (e.g., a $*$ label which indicates that we consider the sum of the values of a tree with $\alpha_v=1$ and one with $\alpha_v=-1$). We shall do this without explicitly mentioning the new label, to simplify the notation. Moreover we can no more associate a label $p(v)$ to a node of this kind, as two factors with different $p(v)$ label ($p(v)=\pm 1$ for $\alpha_v=\pm 1$) have been considered together; nevertheless we shall modify slightly the definition of $p(v)$ by setting $p(v)\overset{\text{def}}{=}$ in such a case (and letting it unchanged in all the other cases).

We shall continue to call $G_v[\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]$ a *propagator* as, for the purposes of the following analysis, only such modified version of the original propagators appearing in (4.30) plays a role.

5.5. Furthermore we define the *degree D* of a propagator to be $D=2$ if either $v\notin\mathcal{P}$ or $v\in\mathcal{P}$, $j_v>l$ (hence $\alpha_v\neq-1$), and $D=1$ otherwise (the constraint, see (4.3), $1\leq r_v\leq 2$ implies that the power to which the divisors appear raised is either 1 or 2); by extension we shall say that a branch λ has degree $D_\lambda=D$ if the corresponding propagator has degree D.

The coefficients $\overline{F}_{\underline{\nu}_v}$ and y'_v in (5.1) satisfy the bounds

$$|y'_v|\leq 4N,\quad \prod_{v\leq v_0}|\overline{F}_{\underline{\nu}_v}|\leq(\mathcal{C}N^2)^m, \tag{5.5}$$

for some constant $\mathcal{C}$ depending on the perturbation f_1 in (2.1); see (2.6), (2.13), and (2.18). For instance, one can take

$$\mathcal{C}=\max\{|J^{-1}|J_0,1\}\max_{|n|\leq N_0,|\nu|\leq N}|f_{\underline{\nu}}|; \tag{5.6}$$

see (4.23), where $|J^{-1}|$ is the maximum of the matrix elements of the (diagonal) matrix J^{-1}.

To bound the product in (5.1), we shall consider simultaneously the cases $k_{v_0}=0,-1$; if $k_{v_0}=0$ the path $\mathcal{P}$ is supposed to be reduced to a single node, v_0, or to the empty set, $\emptyset$, depending on the value of j_{v_0} (respectively, $j_{v_0}=l$, and $j_{v_0}>l$, see above).

What follows below and in Appendix B really goes beyond Ref. 5, although it constitutes a natural extension of it. From now now let us consider the case $l=3$ and the Hamiltonian (1.1).

We shall assume first a condition on the rotation vectors stronger than the Diophantine one, as done in Refs. 3, 18, 5, i.e., we suppose that they satisfy a *strong Diophantine condition*,

$$\begin{aligned}&(1)\quad C_0|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}|\geq|\boldsymbol{\nu}|^{-\tau},\quad \mathbf{0}\neq\boldsymbol{\nu}\in\mathbb{Z}^2,\quad C_0^{-1}=\eta^{-1/2}C(\eta),\\ &(2)\quad \min_{0\geq p\geq n}|C_0|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}|-2^p|\geq 2^{n+1},\ \text{ if }\ n\leq 0,\quad 0<|\boldsymbol{\nu}|\leq(2^{n+3})^{-1/\tau},\end{aligned} \tag{5.7}$$

where n, $p\in\mathbb{Z}$, $n\leq 0$, and

$$\boldsymbol{\omega}_0\equiv\eta^{-1/2}(\Omega_1+\eta^{1/2}J^{-1}A_1)^{-1}\boldsymbol{\omega}'=(1,\eta^{-1}(\Omega_1+\eta^{1/2}J^{-1}A_1)^{-1}\Omega_2), \tag{5.8}$$

so that $\boldsymbol{\omega}' \cdot \boldsymbol{\nu} = \eta^{1/2}(\Omega_1 + \eta^{1/2} J^{-1} A_1)\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}$. We suppose also that $A_1 \in [-\eta^{-1/2}R, \eta^{-1/2}R]$, with $R \leqslant J\Omega_1/2$, so that $\eta^{1/2}(\Omega_1 + \eta^{1/2}J^{-1}A_1) \geqslant \eta^{1/2}\Omega_1/2$.

If we write $\boldsymbol{\omega}' = (\eta^{1/2}\Omega_1 + \eta J^{-1}A_1', \eta^{-1/2}\Omega_2)$ then the measure of the set of A_1''s such that $\boldsymbol{\omega}_0$ verifies the strong Diophantine condition (5.7) has measure of size $O(C_0^{-1}\eta^{-3/2})$.

By reasoning as in Ref. 18, once the case of strong Diophantine vectors has been understood, it can be extended to cover also the case of the usual (weaker) Diophantine condition (expressed by (1) in (5.7) above). Alternatively one could follow the approach in Ref. 19 avoiding completely considering condition (2) in (5.7) and assuming only the ''usual'' condition (1) in (5.7). We shall not perform such an analysis (which can be easily adapted from the quoted papers), and we shall confine ourselves to the case of strongly Diophantine vectors. [Basically the argument is the following: The analysis that we present does not change if $2^p, 2^n$ are replaced by exponentials in another base q (larger than 1) or even if they are replaced by $\gamma(p), \gamma(n)$, where $\gamma(p)/q \xrightarrow[p\to-\infty]{}$, and if in the second of (5.7) we substitute $2^p, 2^{n+1}, 2^{n+3}$ by, respectively, $\gamma(p), \gamma(n+1), \gamma(n+3)$. One then proves a simple arithmetic lemma (see Ref. 18), whereby it follows that, if the first of (5.7) is verified and if $\gamma(p)$ is suitably chosen, then the second holds with $\gamma(p), \gamma(n+1), \gamma(n+3)$ replacing $2^p, 2^{n+1}, 2^{n+3}$.]

Keeping in mind that $C_0 = \eta^{-1/2} e^{+s\eta^{-1/2}}$ is enormous we shall say that

$$
\begin{aligned}
&(1) \quad G_v[\boldsymbol{\omega}' \cdot \boldsymbol{\nu}_0(v)] \quad \text{is on scale 1, if } C_0|\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}_0(v)| > C_0/4, \text{ or if } p(v) \neq 0;\\
&(2) \quad G_v[\boldsymbol{\omega}' \cdot \boldsymbol{\nu}_0(v)] \quad \text{is on scale 0, if } 1/2 < C_0|\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}_0(v)| \leqslant C_0/4;\\
&(3) \quad G_v[\boldsymbol{\omega}' \cdot \boldsymbol{\nu}_0(v)] \quad \text{is on scale } n \leqslant -1, \text{ if } 2^{n-1} < C_0|\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}_0(v)| \leqslant 2^n.
\end{aligned}
\tag{5.9}
$$

5.6. *Remark:* Note that in the above definition of scale the second and third cases can arise only if $p(v)=0$. The propagators on scale 1 can be bounded by 4^2 if $p(v)=0$ and by 1 if $p(v)\neq 0$. Note also that the definition of the scales $n=0$ and $n=1$ is different from Refs. 3 and 5: *this is an important modification*, exploited in Appendix B, useful in order to take advantage from the existence of different scale times.

5.7. As it is well known, (5.1) cannot usefully be bounded by just taking the absolute value of each factor and bounding the denominators by using the Diophantine condition. This is true not only if one wants to get the improved bounds that we are studying, but also if one, more modestly, wants to show convergence for ε small enough; this is a problem usually referred to as a ''*small divisors problem*.''

Useful bounds are nevertheless possible, as shown first in similar cases in Ref. 9, because one can collect the contributions from the various trees into pairwise disjoint (''*nonoverlapping*'') classes whose values add up to a quantity that verifies much better bounds than the individual elements of the same class. Each class $\mathcal{F}(\vartheta)$ will be determined by one of its elements ϑ called a *representative*. This means that there are important cancellations within each class.

The classes can be constructed by collecting trees which have the same *resonance structures*. The key notion of resonance is recalled below and the description of the classes will follow it.

Definition: A ''cluster'' T of scale n_T *will be a maximal connected set of branches with scales* $n > n_T$ *and with at least one branch of scale* n_T.

A free node v will be defined to be internal to T, $v \in T$, if at least one of the branches leading to it or coming from it, i.e., *pertaining to* v (as defined in Sec. IV), belongs to T; a leaf with highest node v' will be defined to be internal to the cluster T if $v' \in T$.

A branch λ_v is called *external* to T if it does not belong to T but it pertains to a node v internal to T, and it said to be entering T if the node v' following it is in T, exiting from T if $v \in T$ (note that an external branch of T is not any branch outside T). We define the *degree* D_T of a cluster T to be the degree of its exiting branch, and the *order* k_T of T to be given by the same formula as (4.26), with the extra constraint that the nodes are internal to T.

Definition: A ''resonance'' V will be a cluster with only two external branches λ_{v_0} *and* λ_{v_1} *carrying the same free momentum,* $\boldsymbol{\nu}_0(v_0) = \boldsymbol{\nu}_0(v_1)$ *and with order ''not too high,'' i.e.,*

$$k_V < \max\{N^{-1}2^{-(n_{\lambda_{v_0}}+3)/\tau},(\gamma N\eta)^{-1}\}, \quad \gamma \stackrel{\text{def}}{=} \frac{4\Omega_1}{\Omega_2}. \tag{5.10}$$

The branch exiting from a resonance will be called a resonant branch, and the scale $n_{\lambda_{v_0}}$ *of the two branches entering and exiting the resonance will be called the resonance-scale. The degree of the propagator of the exiting branch will be called the degree* D_V *of the resonance.*

Even though a node v either with $\delta_v=0$, $\kappa_v\geq 1$ or with $\delta_v=1$, $\boldsymbol{\nu}_v=\mathbf{0}$ is not a cluster in the above sense (because it does not consist of branches) we shall nevertheless regard it as a cluster when there are only one incoming branch and one exiting branch of equal scale. *Therefore, we shall also regard it as a resonance*, if

$$\kappa_v < \max\{N^{-1}2^{-(n_{\lambda_{v_0}}+3)/\tau},(\gamma N\eta)^{-1}\}, \quad \text{when} \quad \delta_v=0, \ k_v\geq 1,$$

$$1 < \max\{N^{-1}2^{-(n_{\lambda_{v_0}}+3)/\tau},(\gamma N\eta)^{-1}\}, \quad \text{when} \quad \delta_v=1, \ \boldsymbol{\nu}_v=\mathbf{0};$$

note that the restriction that if $\delta_v=0=\kappa_v$ there are at least two branches entering v implies that no node with $\delta_v=0=\kappa_v$ can be a resonance.

Definition: A resonance will be called ''strong'' if $p(v_0)=p(v_1)=0$.

All resonances on scale ≤ 0 are strong (see Remark 5.6).

5.8. It is important to note that a strong resonance of degree 2 is necessarily such that also the degree of the entering branch *must* be 2. No branch inside it can be of order 1 and no path can precede v_0. This is so because $D_{\lambda_{v_0}}=2$ implies $j_{v_0}>l$ (see the first of (5.3)) and $p(v_0)=0$ implies that $\alpha_{v_0}=2$, $k_{v_0}=0$, $k'_{v_0}=0$ so any path preceding v_0 would necessarily imply the contradiction $p(v_0)=1$. Also if $D_{\lambda_{v_1}}=1$, $D_{\lambda_{v_0}}=2$ one must have $j_{v_1}=l$ hence $k_{v_1}=0$ (otherwise $p(v_0)>0$) so that $k'_{v_1}=0$: but $\alpha_{v_1}<2$ and k'_{v_1} must be odd. The cases $D_{\lambda_{v_0}}=1$, $D_{\lambda_{v_1}}=1,2$ are both allowed.

5.9. Given a tree ϑ, let V be a resonance (if there are any) with entering branch λ_{v_1} of degree $D_{\lambda_{v_1}}=2$. Then consider the family of all trees which can be obtained from ϑ by detaching the part of the tree having λ_{v_1} as root branch and reattaching it to all the remaining nodes *internal to V but external to the resonances contained inside the cluster* V (if any); to the just defined set of trees we add all the trees obtained by reversing simultaneously the signs of the latter modes of the nodes (this can be done as the sum of the mode vectors $\boldsymbol{\nu}_w$ of such nodes, $w\in V$, vanishes). The set of all the so obtained trees will be denoted $\mathcal{F}_V(\vartheta)$.

The definition of resonance and the strong Diophantine condition insures that all the trees so constructed have a well defined value (*i.e.* no division by zero occurs in evaluating it with the above rules); see the Remark 5.10, (1), below.

If the entering branch λ_{v_1} of the resonance has degree $D_{\lambda_{v_1}}=1$ then also the exiting branch λ_{v_0} has degree $D_{\lambda_{v_0}}=1$, and we collect together with the considered tree also the tree which is obtained from ϑ through the following operation. Replace the resonance V with *a single node v* carrying labels $\delta_v=0$ and $\kappa_v=k_V$, if k_V is the order of the resonance. The set of all the so obtained trees will be denoted by $\mathcal{F}_V(\vartheta)$: the definition of the class $\mathcal{F}_V(\vartheta)$ will therefore depend on the degree of the branch entering V.

Then repeat the above operations for all resonances in ϑ. Thus a class $\mathcal{F}(\vartheta)$ has been constructed and the number of elements of $\mathcal{F}(\vartheta)$ is bounded by the product $\Pi_V 2\mathcal{N}_V$ of the numbers $\mathcal{N}_V$ of branches in each resonance V *which are not* contained inside inner resonances. The latter product is bounded by $\exp \Sigma_V 2\mathcal{N}_V \leq \exp 2m$; the $\mathcal{F}(\vartheta)$ can be obtained starting from any of its elements (which therefore we shall call *representatives* of the class): this is again a *consequence of the strong Diophantine condition*, see Ref. 5.

5.10. *Remarks:* (1) The strong Diophantine condition plays a role here that should be stressed. In fact one checks that because of it the scale of a branch inside a resonance *cannot* change too much, as one considers the different members of a given family. Not enough to change the sets of branches that belong to a given resonance and insures that the different families of trees *do not* *overlap*; for this reason the strong Diophantine condition leads to a simplification of the analysis (the simplification in the simpler case of the KAM theory). A simplification that is however not major (as explained informally in Ref. 3 and as shown in Ref. 18).

(2) To see how the above difficulty is bypassed by using the alternative approach of Refs. 20, 19, we refer to the conclusive comments in Ref. 20, Sec. III.

5.11. Consider trees with $p(v_0)=0$, if v_0 is the highest node of the tree; then the expression of each tree value contains a product like (5.1). As mentioned in the Introduction *we consider only trees without leaves*.

Since the leaf values factorize with respect the product (5.1), they can be dealt with separately, and no overlap arises with the cancellation mechanisms acting on the product (5.1), so that leaves can be easily taken into account; see Appendix C 3 (see also Refs. 3, 10, 5).

The counterterms can also be explicitly expanded in terms of tree values, according to (3.2), which again we can imagine to have no leaves (see however, the comments in Appendix C, 3 below).

The cancellation mechanisms described in Refs. 10, 5 (and recalled in Appendix C) lead to the bound (on a given family $\mathcal{F}(\vartheta)$) described above, in Sec. V, 5.9, see (5.1), (4.30), (4.23),

$$\left(\frac{1}{\eta^{1/2}}\right)^{2m}\left[(4N^3\mathcal{C}')^m 2^{4m}e^{2m}\prod_{n\leqslant 0}(C_0^{2N_n^2}2^{-2nN_n^2})(C_0^{N_n^1}2^{-nN_n^1})\right]\cdot$$

$$\cdot\left[\prod_{n\leqslant 0}\prod_{T,n_T=n}\prod_{i=1}^{m_T^1(n)}2^{(n-n_i+3)}\prod_{i=1}^{m_T^2(n)}2^{2(n-n_i+3)}\right], \tag{5.11}$$

where $\mathcal{C}'=\max\{(2g_0/\Omega_1)^2,4^2\}\mathcal{C}$, with $\mathcal{C}$ the dimensionless constant defined in (5.6); m is the number of nodes $v\geqslant v_0$; N_n^j is the number of propagators on scale n and of degree j in ϑ, which can be written as

$$N_n^j=\overline{N_n^j}+\sum_{\substack{T\\ n_T=n,D_T=j}}(-1)+\sum_{\substack{T\\ n_T=n}}m_T^j(n), \tag{5.12}$$

where $m_T^j(n)$ is the number of resonances on scale n and degree j (i.e., with entering branch having a propagator of degree j) contained inside the cluster T. The terms $\widetilde{N}_n^j$, $j=1,2$, which count the number of propagators *which do not correspond to resonant branches* plus the number of clusters on scale n and of degree j in ϑ, satisfy the bounds

$$\sum_{j=1}^{2}\overline{N_n^j}\leqslant 4mN2^{(n+3)/\tau},\qquad \sum_{n=-\infty}^{0}\sum_{j=1}^{2}\overline{N_n^j}\leqslant 4m\gamma N\eta, \tag{5.13}$$

(with $\gamma=4\Omega_1/\Omega_2$) which are proven in Appendix B. The first square bracket in (5.11) is the bound on the product of individual elements in the family $\mathcal{F}(\vartheta)$ times the bound on their number $\Pi_V 2\mathcal{N}_V<e^{2m}$, see above. The second square bracket term is the part coming from the maximum principle (in the form of Schwarz's lemma), applied to bound the sums of the tree values (''*resummations*'') over the classes $\mathcal{F}(\vartheta)$ introduced above; this is a *nontrivial product of small factors* that arise from the cancellations associated with the resummations, see Appendix C. In (5.11) n_i is the scale of the cluster V_i which is the ith resonance inside T, as in Ref. 5. The $\eta^{-m/2}$ arises as a lower bound on the small divisors of the form $\boldsymbol{\omega}'\cdot\boldsymbol{\nu}$ on scale $n=1$ (for $n=1$ we use the better bound $|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}|\geqslant 2^2\eta^{1/2}$).

5.12. *Remark:* The first bound (5.13) holds for all n and for all Hamiltonians of the form (2.1). On the contrary the second bound in (5.13) will follow from the fact that the rotation vector $\boldsymbol{\omega}_0$ has the form (5.8), with η small, and will be used to control the (huge) factors C_0 in (5.11).

5.13. Hence by substituting (5.12) and the first of (5.13) into (5.11) we see that, for $j=1,2$, the $m_T^j(n)$ is taken away by the first factor in $2^{jn}2^{-jn_i}$, while the remaining 2^{-jn_i} are compensated by the -1 before the $+m_T^j(n)$ in (5.11) taken from the factors with $T=V_i$ (note that there are always enough -1's), and therefore (5.11) is bounded by

$$\left(\frac{2}{\eta^{1/2}}\right)^{2m}(4N^3\mathcal{C}')^m e^m 2^{4m}2^{8m}C_0^{8m\gamma N\eta}\prod_{n=-\infty}^{0}2^{-8mNn2^{(n+3)/\tau}}, \tag{5.14}$$

because the product of the factors C_0 in (5.11) can be bounded by using the second of (5.13), since the product does not contain the $n=1$ factor. The last product in (5.14) is bounded by

$$\prod_{n=-\infty}^{0}2^{-8mNn2^{(n+3)/\tau}}\leqslant\exp\left[8mN2^{3/\tau}\log 2\sum_{p=1}^{\infty}p2^{-p/\tau}\right], \tag{5.15}$$

hence, by adding the remark that the perturbation degree k and the number of tree nodes m are related by $m<2k$, a bound on the sum over all the subtrees of order k with $p(v_0)=0$, $\boldsymbol{\nu}(v_0)=\boldsymbol{\nu}$ (recalling that the number of trees with m nodes is $<4^m m!$) is

$$\Delta_k\overset{\text{def}}{=}\left|\frac{1}{|\mathcal{F}(\vartheta)|}\sum_{\vartheta'\in\mathcal{F}(\vartheta)}\prod_{v\in\vartheta'}\overline{F}_{\underline{\nu}_v}G_v[\boldsymbol{\omega}'\cdot\boldsymbol{\nu}_0(v)]\tilde{y}_v\right|\leqslant B_0^{2k}\eta^{-2k}, \tag{5.16}$$

for some positive constant B_0. The normalization constant $|\mathcal{F}(\vartheta)|$ is introduced in order to avoid overcountings; in fact if $\vartheta'\in\mathcal{F}(\vartheta)$ then $\vartheta\in\mathcal{F}(\vartheta')$, so that, without dividing by $|\mathcal{F}(\vartheta)|$ in (5.16), each tree would be counted $|\mathcal{F}(\vartheta)|$ times.

If $C_0^{-1}=\eta^{-1/2}C(\eta)$ is chosen as in the statement of Theorem 1.4, an explicit calculation gives the bound on (5.11) of the form $(\eta^{-1/2})^{4k}B_0^k$, $k\geqslant 1$, and

$$B_0=2^{18}(4N^3\mathcal{C}')\exp\left[2+4\gamma N\eta\log\eta+8s\gamma N\eta^{1/2}+8N2^{3/\tau}\log 2\sum_{p=1}^{\infty}p2^{-p/\tau}\right], \tag{5.17}$$

which is bounded uniformly in η (for $\eta\leqslant 1$).

5.14. In the previous section trees with $p(v_0)=0$ have been considered; in particular only the contributions (5.1) arising from the value (4.30), once the corresponding tree has been deprived of leaves and counterterms, have been bounded and the bound (5.16) has been obtained through a suitable resummation operation. In such a case the sum over the labels (k_v',k_v) is trivial because the condition $p(v_0)=0$ imposes that only a few values (up to three per node) can be assumed by the hyperbolic mode labels; also the sum over the mode labels $\underline{\nu}_v$ cannot create any problems. In fact for any node v one has $|\underline{\nu}_v|\leqslant N$ and $|n_v|\leqslant N_0$ (see the eighth item in Sec. IV 1).

The cases $p(v_0)\neq 0$ as well as those involving graphs containing leaves or counterterms can be treated in the same manner as already done in Refs. 3, 5. We provide, in Appendix D, a quick description of the construction of the analyticity bound $\varepsilon_0=D^{-1}$ with

$$\varepsilon_0^{-1}=D=[B2^6l(2N+1)^{2l-1}(2N_0+1)]^2,\quad B=\max(B_0\eta^{-1},B_1), \tag{5.18}$$

and B_1 is a suitable numerical constant.

The part of Theorem 1.4 not concerning the connection between the average action $\mathbf{A}'$ and the rotation vector $\boldsymbol{\omega}'$ nor the splitting size follows.

5.15. Determining the exact splitting size (i.e., the leading behavior asymptotically as $\eta\to 0$ with $\varepsilon<B\eta^2$) is *not* trivial because of the existence of major cancellations in the evaluation of the determinant of the splitting matrix; however the analysis in Ref. 8 dealt with this question in detail; in the latter paper remarkable cancellations are exhibited and an exact formula for the splitting angles is derived (see (7.19) of Ref. 8).

One gets the results in the last item of Theorem 1.4 simply if Ref. 8 and the first part of Theorem 1.4 (to estimate the remainders) are used; then the claimed bounds on the splitting follow immediately (see Remark 1.5). In Ref. 6 an improvement of lemma 1 and lemma 1′ of Ref. 2 was used instead to control the density of tori in phase space (the lemmata in Ref. 2 were, as such, useless already in the case in Ref. 8 because they would require that ε be far smaller than the ε_0 of Theorem 1.4); see Ref. 6, where this is discussed in detail and differs from our case only because it relied on a theorem weaker than Theorem 1.4 above (as the radius of convergence estimate there is proportional to η to the power $\frac{9}{2}+$ rather than our 2).

5.16. *Remarks:*

(1) The bound (5.16) and the discussion in Sec. V 14 imply the convergence of the perturbative expansions for the parametric equations of the invariant tori (for the Hamiltonian (1.1)), if $|\varepsilon|<\varepsilon_0=O(\eta^2)$. This bound on the convergence radius should be compared with the value given by Ref. 6, which, for $\mathcal{N}=O(\eta^{-1/2})$, gives $\varepsilon_0=O(\eta^{9/2}/\log^2\eta^{-1})$. *As usual the Lindstedt series gives a much better estimate than the classical method* (i.e., an exponent 2 vs ~4.5). We do not see immediately how to improve substantially the classical estimate without important changes in the architecture of the proof of Ref. 6, although this should be possible; on the other hand, from the above analysis, $\varepsilon_0=O(\eta^2)$ might be close to an optimal result. If so it should be no surprise that our analysis is so delicate.

(2) In the Hamiltonian (1.1), (2.1) the polynomial dependence of the interaction on the rotators angles has very likely a purely technical motivation (as it simplifies the analysis) and could probably be relaxed into a more general analytical dependence, as in Ref. 13. On the contrary the hypothesis that the perturbation is a trigonometric polynomial of degree N_0 in φ is fundamental to get the correct asymptotic behavior, in order to apply the results in Ref. 8, where the dominance of Mel'nikov integral is proven *provided the perturbation is polynomially small in a power of* η^{N_0} (so that the results of Ref. 8 become meaningless for $N_0\to\infty$).

(3) A bound of the form (5.16) holds under the weaker condition that $C(\eta)\leqslant e^{-s\eta^{-a}}$, with $a\leqslant 1$ (see (5.17)).

(4) If q is defined as in Sec. I, 1.6 so that $|\varepsilon|C(\eta)^{q\eta}<1$ implies analyticity in ε, the above analysis gives that q can be taken $q=8\gamma N$. In general all the bounds found so far are not uniform in N; in order to deal with the analytical case in the frame of the exploited formalism one should bound the small divisors by using the results of Ref. 19 or Eliasson-Siegel's bound (see, for instance, Ref. 14), and use explicitly as in Ref. 13 the exponential decay in $\boldsymbol{\nu}$ of the Fourier coefficients $f^1_{\boldsymbol{\nu}}$.

(5) Note that we have convergence for $|\varepsilon|<O(\eta^2)$, while the asymptoticity of the splitting estimate follows only for $|\varepsilon|<O(\eta^\zeta)$ (with $\zeta=2(N_0+3)$), which is *much smaller* for $\eta\to 0$. Then the question is: What will be the asymptotics for ε small enough to be in the convergence domain but too large to be in the domain of the asymptotic result? There is some evidence that there is a critical value T_c such that, if $\varepsilon=\eta^T$, then for $T>T_c$ the asymptotic formula that we can prove only for $T>\zeta$ holds, but for $T=T_c$ it is modified remaining qualitatively of the same size $O(e^{-(1/2)\eta^{-(1/2)}})$ and for $T<T_c$ it becomes qualitatively different. The analogy with the critical point scaling phenomena seems to be substantial and, keeping in mind that the above theory can be interpreted as a field theory, see Ref. 17, one would say that the region $T>T_c$ is described by a trivial fixed point; a nontrivial fixed point describes the case $T=T_c$ and another ''low temperature'' fixed point describes the cases $T<T_c$. Evidence in this direction comes also from the theory of the standard map and its developments.[21,22,23]

ACKNOWLEDGMENTS

This work is part of the research program of the European Network on ''Stability and Universality in Classical Mechanics,'' No. ERBCHRXCT940460. Supported in part by CNR-GNFM and Rutgers University.

APPENDIX A: COUNTERTERMS

1. To order h one can write, by using (2.11), (2.24), and (4.11),

$$\Xi_-^{h\sigma}(t)=\int_{\sigma\infty}^{t} d\tau w_0^1(t,\tau)(\Phi_+^{h\sigma}(\tau;\gamma_1(g_0),\ldots,\gamma_{h-1}(g_0))+\gamma_h(g_0)\sin\varphi^0(\tau))+\int_{\sigma\infty}^{0} d\tau w_0^0(t,\tau)$$
$$\times(\Phi_+^{h\sigma}(\tau;\gamma_1(g_0),\ldots,\gamma_{h-1}(g_0))+\gamma_h(g_0)\sin\varphi^0(\tau)), \tag{A1}$$

where $\Phi_+^{h\sigma}(t;\gamma_1(g_0),\ldots,\gamma_{h-1}(g_0))$ takes into account all the contributions to $\Phi_+^{h\sigma}=(J_0g_0^2)^{-1}F_+^{h\sigma}$ except the only one explicitly depending on $\gamma_h(g_0)$, which is given by $\gamma_h(g_0)\sin\varphi^0(\tau)$.

We shall impose, recursively, that the contributions to the integrands in (A1) arising from amputated trees (see comments after (4.31)) with $(\boldsymbol{\nu}_0(v_0),p(v_0))=(\mathbf{0},0)$, *without leaves and without end nodes bearing a counterterm label* (see Remark A, 2 below), compensate exactly the contributions due to the trees with a single node representing a counterterm of order h (i.e., the terms with $\gamma_h(g_0)$ in (A1)). The first described type of contributions can be written

$$\mathcal{W}_1=w_l^1(t,\tau)\Phi_+^{h\sigma}(\tau;\gamma_1(g_0),\ldots,\gamma_{h-1}(g_0))\equiv\sum_{\nu\in\mathbb{Z}^{l-1}}\sum_{p=-1}^{\infty}\tilde{\mathcal{W}}_1(\boldsymbol{\nu},p)e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}t-pg_0\sigma t} \tag{A2}$$

in the first integral in (A1) and

$$\mathcal{W}_0=w_l^0(t,\tau)\Phi_+^{h\sigma}(\tau;\gamma_1(g_0),\ldots,\gamma_{h-1}(g_0))\equiv\sum_{p=-1}^{\infty}\sum_{\nu\in\mathbb{Z}^{l-1}}\tilde{\mathcal{W}}_0(\boldsymbol{\nu},p)e^{i\boldsymbol{\omega}'\cdot\boldsymbol{\nu}t-pg_0\sigma t} \tag{A3}$$

in the second one. Imposing that such contributions are canceled by the contributions with $p=0$ arising from $w_l^1(t,\tau)\gamma_h(g_0)\sin\varphi^0(\tau)$ and $w_l^0(t,\tau)\gamma_h(g_0)\sin\varphi^0(\tau)$ gives our prescription on how to fix $\gamma_h(g_0)$.

Since two integrals are involved (one for $\rho_{v_0}=1$ and one for $\rho_{v_0}=0$), the first time dependent and the second time independent, two conditions may seem to be required; however, note that $\sin\varphi^0(t)=2\sinh g_0t/\cosh^2 g_0t$ is expanded in odd powers of $x=e^{-\sigma g_0 t}$, hence the only terms appearing in (A2) and (A3) which can contribute to $p=0$ are, in both cases, those involving $y^{(-1)}(k'_{v_0},k_{v_0})$, with $k_{v_0}=-1$. This means that the contributions with $p=0$ arising from (A2) and (A3) are equal, so that no compatibility problem arises.

Note that an expression analogous to (A1) is obtained also for $\Xi_+^{h\sigma}(t)$; however the terms with $p=0$ have the same form as in (A1) (see (4.17) for $j_{\lambda_{v_0}}=l$ and for $j_{\lambda_{v_0}}=0$), so that if no term with $(\boldsymbol{\nu}_0(v_0),p(v_0)=(\mathbf{0},0))$ contributes to $\Xi_-^{h\sigma}(t)$, then also no term with $(\boldsymbol{\nu}_0(v_0),p(v_0))=(\mathbf{0},0)$ contributes to $\Xi_+^{h\sigma}(t)$.

2. Remark: It may seem strange that we exclude from the definition of the counterterms trees with leaves; in fact one can imagine to realize trees with $(\boldsymbol{\nu}(v_0),p(v_0))=(\mathbf{0},0)$ also by trees which contain leaves with a stalk bearing a label $j_w>l$ or $j_w=l$ and internal momenta $(\boldsymbol{\nu}',p')$. Such terms would give rise to $\boldsymbol{\alpha}$-dependent counterterms which of course are not allowed; however, it turns out that the sum over all contributions to tree values of trees with $(\boldsymbol{\nu}(v_0),p(v_0))=(\mathbf{0},0)$ from such trees cancel *exactly*; this is explained together with the other cancellations built in our algorithm, in Appendix C (see Appendix C, 5 in particular).

3. Let us consider the first integral in (A1). Corresponding to the node v_0 of each tree whose value contributes to $\Xi_-^{h\sigma}(t)$ there is a coefficient $\tilde{y}_{n_{v_0}}(k'_{v_0},k_{v_0})$, see (4.14), (4.17). Then from (A1) and the just formulated condition to impose we obtain

$$\sum_{\substack{\vartheta\in\mathcal{T}_{0,h},\alpha_{v_0}=-1\\ p(v_0)=0,k'_{v_0}=1}} \overline{\mathrm{Val}}(\vartheta)+\gamma_h(g_0)w_l^1(t,\tau)\sin\varphi_0(\tau)\Big|_{k'=1,p=0}=0, \tag{A4}$$

where the sum is over the set $\mathcal{T}_{0,h}$ of all trees of order h and momentum $\boldsymbol{\nu}(v_0)=\mathbf{0}$, with $\boldsymbol{\nu}_0(v_0)=\mathbf{0}$ (see Remark A, 2), $p(v_0)=0$, $j_{v_0}=l$ and $k'_{v_0}=1$; hence if $p(v_0)=0$, $j_{v_0}=l$, one must have $k_{v_0}=-1$, hence $\alpha_{v_0}=-1$ and $k'_{v_0}=1$ which is a possible case indeed.

A trivial calculation (just take into account that $y_{n_v}^{(-1)}(1,-1)=\sigma/2$ and $\sin\varphi^0(\tau)=4\sigma x+O(x^3)$) gives

$$w_l^1(t,\tau)\sin\varphi_0(\tau)|_{k'=1,p=0}=2, \tag{A5}$$

so that (4.32) follows; the above follows Ref. 5, p. 287.

APPENDIX B: (IMPROVED) RESONANT SIEGEL–BRYUNO'S BOUND

1. We follow the idea of Pöschel[24] (see also Refs. 3, 18, 5). In the discussion, we focus on the scale labels, so that it is quite irrelevant which value the $p(v)$'s, $v\in\vartheta$, assume, and therefore which resonances are strong and which are not.

Calling $N_n^*(\vartheta)$ the number of nonresonant branches carrying a scale label $\leqslant n$, in a tree ϑ with m nodes, we shall prove first that

$$N_n^*(\vartheta)\leqslant 2mE_n-1,\quad E_n\overset{\text{def}}{=}N2^{(3+n)/\tau},\quad n\leqslant 1, \tag{B1}$$

provided that $N_n^*(\vartheta)>0$, and

$$N_0^*(\vartheta)\leqslant 2m\gamma N\eta-1,\quad \gamma\overset{\text{def}}{=}4\Omega_1/\Omega_2, \tag{B2}$$

if $N_0^*(\vartheta)>0$.

Define, as in Sec. V, 5.7, $\boldsymbol{\omega}_0=(1,\eta^{-1}(\Omega_1+\eta^{1/2}J^{-1}A_1)^{-1}\Omega_2)$. Then $C_0|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}|>|\boldsymbol{\nu}|^{-\tau}$ for all $\mathbf{0}\neq\boldsymbol{\nu}\in\mathbb{Z}^{l-1}$; see (5.7). Assume also η so small that $C_0\geqslant 2$ (this is not restrictive as we are interested in $\eta\to 0$).

Set $E_n\equiv N2^{(n+3)/\tau}$ as in (B1). Note that if $m\leqslant E_n^{-1}$ one has $N_n^*(\vartheta)=0$. In fact $m\leqslant E_n^{-1}$ implies that, for all $v\in\vartheta$, $|\boldsymbol{\nu}_0(v)|\leqslant NE_n^{-1}$, i.e., $C_0|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}_0(v)|\geqslant(N^{-1}E_n)^\tau=2^{n+3}$, so that there are no clusters T with $n_T=n$. Note also that if $m\leqslant(\gamma N\eta)^{-1}$, with $\gamma=4\Omega_1/\Omega_2$, then $N_0^*(\vartheta)=0$, as $|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}_0(v)|\geqslant 1$ for all $v\in\vartheta$ in such a case.

2. Let us prove first the inequality (B1). If ϑ has the root branch either with scale $>n$, or with scale $\leqslant n$ and resonant, then calling $\vartheta_1,\vartheta_2,\ldots,\vartheta_k$ the subtrees of ϑ ending into the highest node v_0 of ϑ and with $m_j>E_n^{-1}$ nodes, $j=1,...,k$, one has $N_n^*(\vartheta)=N_n^*(\vartheta_1)+\cdots+N_n^*(\vartheta_k)$ and the statement is inductively implied from its validity for $m'<m$ and from the just proved fact that $N_n^*(\vartheta)=0$ if $m\leqslant E_n^{-1}$.

If the root branch is on scale $\leqslant n$ and nonresonant, one has $N_n^*(\vartheta)\leqslant 1+\Sigma_{i=1}^k N_n^*(\vartheta_i)$: if $k=0$ the statement is trivial, if $k\geqslant 2$ the statement is again inductively implied by its validity for $m'<m$. If $k=1$ one has $N_n^*(\vartheta)\leqslant 1+2m_1E_n-1$, hence we once more have a trivial case unless the order m_1 of ϑ_1 is $m_1>m-(2E_n)^{-1}$: but in the latter case we shall show that the root branch of ϑ_1 has scale $>n$.

Accepting the last statement (which will be proved below), one will obtain $N_n^*(\vartheta)=1+N_n^*(\vartheta_1)=1+N_n^*(\vartheta'_1)+\cdots+N_n^*(\vartheta'_{k'})$, where ϑ'_j's are the k' subtrees ending into the highest node of ϑ'_1 with orders $m'_j>E_n^{-1}$, $j=1,...,k'$. Going once more through the analysis the only nontrivial case is if $k'=1$ with the root branch of ϑ'_1 nonresonant; and in such case $N_n^*(\vartheta'_1)=N_n^*(\vartheta''_1)+\cdots+N_n(\vartheta''_{k''})$, etc., until we reach a trivial case or a tree of order $\leqslant m-(2E_n)^{-1}$.

It remains to check that if $m-m_1<(2E_n)^{-1}$ then the root branch of ϑ_1 has a scale $>n$. Let us proceed by *reductio ad absurdum*. Suppose that the root branch of ϑ_1 is on scale $\leqslant n$. Then $C_0|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}_0(v_0)|\leqslant 2^n$ and $C_0|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}_0(v_1)|\leqslant 2^n$, if v_1 is the highest node of ϑ_1. Hence $C_0|\boldsymbol{\omega}_0\cdot(\boldsymbol{\nu}_0(v)-\boldsymbol{\nu}_0(v_1))|<2^{n+1}$ (equality would imply violation of the strong Diophantine property, (5.7)), and the Diophantine condition implies that

$$|\boldsymbol{\nu}_0(v_0)-\boldsymbol{\nu}_0(v_1)|>2^{-(n+1)/\tau}\equiv\delta, \tag{B3}$$

because $\boldsymbol{\nu}_0(v_0)\neq\boldsymbol{\nu}_0(v_1)$ (the root branch of ϑ being supposed nonresonant). But $m-m_1<(2E_n)^{-1}$, so that $|\boldsymbol{\nu}_0(v_0)-\boldsymbol{\nu}_0(v_1)|<(2E_n)^{-1}N<2^{-1}2^{-(n+3)/\tau}=2^{-(1+2/\tau)}\delta<\delta$, which contradicts the inequality (B3).

3. Let us prove now (B2). If ϑ has the root branch either with scale 1, or with scale $\leqslant 0$ and resonant, then calling $\vartheta_1,\vartheta_2,\ldots,\vartheta_k$ the subtrees of ϑ ending into the highest node v_0 of ϑ and with $m_j>(\gamma N\eta)^{-1}$ nodes, $j=1,\ldots,k$, one has $N_0^*(\vartheta)=N_0^*(\vartheta_1)+\cdots+N_0^*(\vartheta_k)$ and the statement is inductively implied from its validity for $m'<m$ and from the fact that $N_0^*(\vartheta)=0$ if $m\leqslant(\gamma N\eta)^{-1}$.

If the root branch is on scale $\leqslant 0$ and nonresonant, one has $N_0^*(\vartheta)\leqslant 1+\Sigma_{i=1}^k N_0^*(\vartheta_i)$; if $k=0$ the statement is trivial, if $k\geqslant 2$ the statement is again inductively implied by its validity for any $m'<m$. If $k=1$ we once more have a trivial case unless the order m_1 of ϑ_1 is $m_1>m-(2\gamma N\eta)^{-1}$, but in the latter case the root branch of ϑ_1 has scale 1.

Accepting the last statement (which will be proved below), one will obtain $N_0^*(\vartheta)=1+N_0^*(\vartheta_1)=1+N_0^*(\vartheta_1')+\cdots+N_0^*(\vartheta_{k'}')$, where ϑ_j''s are the k' subtrees ending into the highest node of ϑ_1' with orders $m_j'>(2\gamma N\eta)^{-1}$. Going once more through the analysis the only nontrivial case is if $k'=1$ and in that case $N_0^*(\vartheta_1')=N_0^*(\vartheta_1'')+\cdots+N_0(\vartheta_{k''}'')$, etc., until we reach a trivial case or a tree of order $\leqslant m-(2\gamma N\eta)^{-1}$.

It remains to check that, if $m-m_1<(2\gamma N\eta)^{-1}$, then the root branch of ϑ_1 has scale 1. Suppose that the root branch of ϑ_1 is on scale $\leqslant 0$. Then $p(v_1)\neq 0$ and $|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}_0(v_0)|\leqslant 1/4$, $|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}_0(v_1)|\leqslant 1/4$, if v_1 is the highest node of ϑ_1, i.e.,

$$|\boldsymbol{\omega}_0\cdot(\boldsymbol{\nu}_0(v_0)-\boldsymbol{\nu}_0(v_1))|\leqslant 1/2. \tag{B4}$$

As the root branch of ϑ is supposed nonresonant, then $m-m_1<(2\gamma N\eta)^{-1}$ implies that $0<|\boldsymbol{\nu}_0(v_0)-\boldsymbol{\nu}_0(v_1)|<(2\gamma N\eta)^{-1}N=(2\gamma\eta)^{-1}$, so that one would have $|\boldsymbol{\omega}_0\cdot(\boldsymbol{\nu}(v_0)-\boldsymbol{\nu}(v_1))|\geqslant 1$, which is contradictory with the inequality (B4).

4. A similar induction can be used to prove that if the number of branches on scale n is $N_n(\vartheta)>0$ then the number $p_n(\vartheta)$ of clusters of scale n verifies the bound

$$p_n(\vartheta)\leqslant 2mN2^{(n+3)/\tau}-1. \tag{B5}$$

In fact this is true for $m\leqslant E_n^{-1}$, if E_n is defined as in Appendix B1. Otherwise, if the highest tree node v_0 is not in a cluster on scale n, one calls $\vartheta_1,\ldots,\vartheta_k$ the subtrees ending into v_0, and one has $p_n(\vartheta)=p_n(\vartheta_1)+\cdots+p_n(\vartheta_k)$, so that the statement follows by induction. If v_0 is in a cluster V of scale n, and $\vartheta_1,\ldots,\vartheta_k$ are the subtrees entering the cluster containing v_0 and with orders $m_j>E_n^{-1}$, one will find $p_n(\vartheta)=1+p_n(\vartheta_1)+\cdots+p_n(\vartheta_k)$. Again we can assume that $k=1$, the other cases being trivial. But in such case there will be only one branch entering the cluster V and it will have a propagator of scale $\leqslant n-1$. Therefore the cluster V must contain at least E_n^{-1} nodes. This means that $m_1\leqslant m-(2E_n)^{-1}$.

Finally, the bound

$$\sum_{n=-\infty}^{0} p_n(\vartheta)\leqslant 2m\gamma N\eta-1 \tag{B6}$$

is a trivial consequence of (B2).

5. Let $\overline{N}_n^* \leqslant N_n^*$ be the number of nonresonant branches on scale n. Then if $\overline{N}_n$ is the number of nonresonant branches *plus* the number of clusters on scale n, $\overline{N}_n^*$ verifies the bounds

$$\overline{N}_n^* = (\overline{N}_n^* + p_n) - p_n \equiv \overline{N}_n - p_n \leqslant 4mN2^{(n+3)/\tau} - 2 - \sum_{\substack{T \\ n_T = n}} (1) \leqslant 4mN2^{(n+3)/\tau} + \sum_{\substack{T \\ n_T = n}} (-1). \tag{B7}$$

This proves that (B1) and (B5) imply an inequality analogous to the first of (5.13); likewise one derives an inequality similar to the second of (5.13) by combining (B2) and (B6).

APPENDIX C: CANCELLATIONS BETWEEN RESONANCES

In this Appendix we recall briefly the cancellation mechanisms of Ref. 5. We provide this as a guide to the reader and as a tune up of a fine points of the analysis of Ref. 5 (the analysis in Appendix C 2 is given here in full details while in Ref. 5 it was left out).

1. Consider a tree ϑ with a strong resonance V of order k_V. Let λ_{v_0} and λ_{v_1} be, respectively, the exiting and entering branches of V. There are two possibilities; either the degree of the propagator corresponding to the branch exiting from V is $D_{\lambda_{v_0}} = 2$ or it is $D_{\lambda_{v_0}} = 1$ (equivalently the degree of the resonance is either $D_V = 2$ or $D_V = 1$).

Let us discuss first the case in which the degree D_V of the resonance is $D_V = 2$. Then $j_{v_0} > l$ (see (5.3) and the comments after the definition of strong resonance in Sec. V, 5.9) and, by following the notations of Sec. V, 5.1, we shall say that $\mathcal{P} = \emptyset$, i.e., there is no path $\mathcal{P}$ ending into v_0. It follows, from the properties of $\mathcal{P}$ discussed at the beginning of Sec. V, 5.1 above, that $p(v_1) = 0$ implies $j \equiv j_{v_1} > l$ and $D_{\lambda_{v_1}} = 2$ (see again (5.3)).

Consider all the trees belonging to the class $\mathcal{F}(\vartheta)$ which are obtained from ϑ by detaching the subtree having as branch root the entering branch λ_{v_1} of the resonance and attaching it to all the remaining nodes of V (see the definition of the class $\mathcal{F}_V(\vartheta)$ in Sec. 5 I). As a consequence of such an operation some of the branches internal to the resonance have changed the free momentum by an amount $\boldsymbol{\nu}_0(v_1)$, and if w is the node inside V to which the branch λ_{v_1} is attached and $j_{v_1} - l > 0$, then $\overline{F}_{\underline{\nu}_w}$ (see (4.23)), has the form of an even function of $\underline{\nu}_w$ times a factor $(i\nu_{wj})$.

We shall call the *resonance value* $\mathcal{R}_V$ the product of factors appearing in the definition of the tree value and relative only to the nodes and branches internal to the resonance V,

$$\mathcal{R}_V = \overline{F}_{\underline{\nu}_{v_0}} y'_{v_0} \prod_{\substack{v \in V \\ v < v_0}} \overline{F}_{\underline{\nu}_v} y'_v G_v[\boldsymbol{\omega}' \cdot \boldsymbol{\nu}_0(v)], \tag{C1}$$

and shall consider the resonance value as a function of the quantity $\mu \equiv \boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}_0(v_1)$.

Then for $\mu = 0$ a quantity proportional to $\Sigma_{w \in V} \boldsymbol{\nu}_{wj}$ is constructed, but such a quantity is vanishing by definition of resonance, as $j = j_{v_1} - l > 0$.

If we sum also on an overall change of signs of the mode labels of the nodes internal to the resonance (by following the definition of the class $\mathcal{F}(\vartheta)$ given in Sec. V, 5.9), we obtain a zero contribution also to first order in μ (here the even parity of the perturbation f is essential, see Refs. 3 and 5).

This can be seen by using the explicit form of the functions in (4.21), i.e., the coefficients listed in (4.24). Noting that in the present case *there cannot be any* $\mathcal{P}$ *inside* V the only propagators we can associate with the branches internal to V have the form of the two first terms of (5.3), so that, for $\mu = 0$, *they are even functions of the mode labels*. Moreover in such a case the analysis in Sec. V, 5.1 shows that $\alpha_v = -1$, $\alpha_v = 1$ and $\alpha_v = 2$ imply, respectively, $k'_v = -k_v = 1$, $k'_v = -k_v = -1$ and $k'_v = -k_v = 0$ (the case $\alpha_v = 0$ is not possible here), then no n_v labels appear in the coefficients $y_{n_v}^{(\alpha_v)}(k'_v, k_v)$ corresponding to the nodes $v \leqslant v_0$ (see the list of coefficients in (4.24)).

Therefore all the dependence on the n_v labels is through the factors $\overline{F}_{\underline{\nu}_v}$ in (4.23). This yields that there is an even number of the n_v (if there are any) corresponding to the nodes $v \in V$: two for each branch λ_v with $j_v = l$, by taking into account that $j_{v_0}, j_{v_1} > l$, so that no change is produced by the sign reversal (since, by the parity properties of the Hamiltonians (1.1) and (2.1), one has also $f^{\delta_v}_{\underline{\nu}_v} = f^{\delta_v}_{-\underline{\nu}_v}$). This means that the resonance value is an even function of μ.

2. Let us now consider the case in which the strong resonance is of degree $D_V = 1$ and the tree ϑ has no leaves inside V. In such a case $\alpha_{v_0} = -1$ and $j_{v_0} = l$, hence $D_{\lambda_{v_0}} = 1$ (see (5.3)), then a first order zero in μ will be enough. Moreover there is a $\mathcal{P}$ inside the resonance, we shall distinguish between the cases $v_1 \notin \mathcal{P}$ and $v_1 \in \mathcal{P}$.

Let us consider first the case $v_1 \notin \mathcal{P}$ (in particular this is the case when $\mathcal{P} = v_0$, $k_{v_0} = 0$, provided $k_V \geqslant 2$). In such a case $j_{v_1} > l$ and we can reason as above to obtain a first order zero. Note that in such a case there would be no cancellations between tree values of trees obtained by the sign reversal operation.

On the contrary, if $v_1 \in \mathcal{P}$, then $k_{v_0} = -1$, and one has also $\alpha_{v_1} = -1$ and $j_{v_1} = l$. In this case consider together with the tree ϑ also the tree ϑ' obtained from ϑ by performing the following operation (recall the definition of $\mathcal{F}_V(\vartheta)$): replace the resonance V with a single node v carrying labels $\delta_v = 0$ and $\kappa_v = k_V$ (if k_V is the order of the resonance V), then express the counterterm $\gamma_{\kappa_v}(g_0)$ associated with the node v in terms of trees. If ϑ_1 is the subtree having λ_{v_1} as a root branch, then the values of the two considered trees ϑ and ϑ' can be written, respectively, as $\mathrm{Val}(\vartheta) = A(\vartheta)\mathcal{R}_V \mathrm{Val}(\vartheta_1)$ and $\mathrm{Val}(\vartheta') = A(\vartheta)[\gamma_{k_V}(g_0)\sigma/2]\mathrm{Val}(\vartheta_1)$, where $\sigma/2 = y_v^{(-1)}(1, -1)$ and $A(\vartheta)$ takes into account the factors corresponding to all nodes *not* preceding v_0, and has the same value for both ϑ and ϑ'.

The resonance value $\mathcal{R}_V$, for $\mu = 0$, can be written as

$$\mathcal{R}_V = \overline{\mathrm{Val}}(\vartheta_0) i n_{v_1'}, \quad \text{for some} \quad \vartheta_0 \in \mathcal{T}_{0,k} \quad \text{with} \quad p(v_0) = 0, \quad k_{v_0} = -1, \tag{C2}$$

see the definitions (4.30) and (4.31) of tree value and the definition (C1) of resonance value; remember that we are considering resonances V with degree $D_V = 1$, so that $k_{v_0} = -1$ and, as a consequence, $k'_{v_0} \geqslant 1$; see (4.14). The counterterm $\gamma_\kappa(g_0)$ can be represented in terms of trees as in (4.32); note that, if the tree contributing to $\gamma_\kappa(g_0)$ has $k_{v_0} = -1$, the condition $\alpha_{v_0} = -1$ implies that such a tree has a node $w > v_0$ with $k_w + k'_w = 1$, while all the other nodes $v \neq w$ have $k_v + k'_v = 0$.

Among the contributions in (4.32) to $\gamma_{k_V}(g_0)$ there will be a quantity $\overline{\mathrm{Val}}(\vartheta_2)$, where ϑ_2 will have the same topological form of ϑ_0 in (C2) with the node w such that $k_w + k'_w = 1$ corresponding to the node $v'_1 \in V$; then we denote both nodes by w.

Then $\overline{\mathrm{Val}}(\vartheta_0)$ will be related to $\overline{\mathrm{Val}}(\vartheta_2)$ by

$$\overline{\mathrm{Val}}(\vartheta_2) = -\left[\frac{y^{(\alpha_v)}_{n_v}(k'_w, k_w)|_{k'_w + k_w = 1}}{y^{(\alpha_v)}_{n_v}(k'_w, k_w)|_{k'_w + k_w = 0}} \right] \overline{\mathrm{Val}}(\vartheta_0), \tag{C3}$$

so that a look at the coefficients listed in and after (4.24) shows that the factor in square brackets in (C3) (when it is not vanishing) is equal to $4 i n_w \sigma$. The quantity $\overline{\mathrm{Val}}(\vartheta_2)$, in order to contribute to $\gamma_{k_V}(g_0)\sigma/2$, has to be multiplied by a factor -4σ extra with respect to $\overline{\mathrm{Val}}(\vartheta_0)$, which, on the other hand, has to be multiplied by $i n_w$ in order to contribute to the resonance value $\mathcal{R}_V$ (see (4.32)).

Then, for $\mu = 0$, by summing the values of the two considered contributions one obtains

$$A(\vartheta)\left[\overline{\mathrm{Val}}(\vartheta_0) i n_w - \frac{1}{4\sigma} \overline{\mathrm{Val}}(\vartheta_2) \right] \mathrm{Val}(\vartheta_1), \tag{C4}$$

which is zero by (C3), so that a first order zero is obtained.

3. If there are leaves, nothing changes in the discussion of Appendix C 1, as $k_{v_0}=0$ implies that only leaves w with $j_w > l$ are possible, and $\xi_w(k'_w,0)\equiv 1$ in such a case (see (4.18)).

In Appendix C 2 when discussing the case $v_1 \in \mathcal{P}$, one has to take care of the case in which there is a leaf with highest node $\tilde{w}$ with $k'_w=1$ (such a leaf will be at the end of the path $\mathcal{P}$). In fact the resonances having as entering branch a branch of the path $\mathcal{P}$ cannot have any leaves with $k'_w=1$, while when considering the graphical representation for $\gamma_\kappa(g_0)$, there will be also contributions arising from trees containing a leaf; such contributions will be either of the form (4.32) with $\overline{\mathrm{Val}}(\vartheta_2)=\overline{\mathrm{Val}}(\vartheta_2) in_{v'_1}\xi_{v_1}(1,0)L^{h_1\sigma}_{l\nu(v_1)}(\vartheta_2)$, or of the form $\gamma_\kappa(g_0)=\gamma_{\kappa-h_1}(g_0) in_{v'_1}\xi_{v_1}$ $(1,0)L^{h_1\sigma}_{l\nu(v_1)}(\vartheta_2)$, where $h_1 \geqslant 1$, and ϑ'_2 is a suitable tree of order $k-h_1$. Then one realizes that the two contributions cancel exactly, so that no new case has to be discussed with respect to the analysis of Appendix C, 2.

4. The above discussion completes the proof of approximate cancellations of resonance values (i.e., of cancellations to first and second order, according to the degree of the resonant branch). The existence of cancellations, approximate to the first or second order, is all is needed to obtain the bound (5.16); the analysis continues exactly as in Ref. 5 and is based on simple analyticity arguments that allow us to exploit, via the maximum principle, the fact that in a resonance with momentum $\boldsymbol{\nu}$ the functions of $\mu=\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}$ that have been considered above have a zero in μ of order 1 or 2.

A complete analysis showing that the higher orders contributions (i.e., the part which does not cancel) can be performed as in Ref. 5, Appendix B, and the final result is given by the bound (5.16) in Sec. V, 5.13.

5. We shall show now that all contributions with $(\boldsymbol{\nu}_0(v_0),p(v_0))=(\mathbf{0},0)$ involved in the definition of the counterterms (see Appendix A) must have automatically also $\boldsymbol{\nu}(v_0)=\mathbf{0}$. The analysis performed in Sec. V, 5.1 shows that in order to have $p(v_0)=0$ (for $j_{v_0}=l$), there can be any number of leaves with highest nodes w such that $j_w > l$ and only one leaf w with $j_w=l$ (contributing, respectively, $k'_w=0$ and $k'_w=1$ to $p(v_0)$).

Each time a leaf with $j_w > l$ appears, if we sum together the values of all trees obtained by detaching the leaf with its stalk, then reattaching it to all the other nodes of ϑ_f, we obtain a vanishing contribution: simply by the cancellation mechanism described in Appendix C 1 (assuring there the first order zero), *which, now, is an exact cancellation as the leaf does not contribute to the free momenta of the branches of* ϑ_f, *so that it does not modify the propagators*. So we can suppose that no leaf with $j_w > l$ is possible in trees involved in the determination of the counterterms.

In the same way, if we have a tree ϑ having a leaf with $j_w=l$ and $h_w=h-h_1$ (for some h_1), we can reason as in Appendix C, 2 and consider, together with ϑ, also the tree formed by only one free node, carrying a counterterm label h_1 and bearing the same leaf as ϑ. The same cancellation mechanism described in Appendix C, 2 apply now; again the only difference is that now the cancellation is exact (by the same reason as before).

This shows that no tree with leaves can contribute to $(\boldsymbol{\nu}_0(v_0),p(v_0))=(\mathbf{0},0)$, so that for such trees one has $\boldsymbol{\nu}(v_0)=\boldsymbol{\nu}_0(v_0)=\mathbf{0}$. This, together with the analysis in Appendix A, 1 proves (A4).

APPENDIX D: GRAPHS WITH NONZERO TOTAL HYPERBOLIC MOMENTUM, WITH LEAVES OR WITH COUNTERTERMS

1. Consider first the cases $p(v_0)\neq 0$. In this case we consider the nodes $w < v_0$ with $p(w) =0$ and which are the nearest to v_0; by construction all nodes z between v_0 and the just singled out nodes have $p(z)\neq 0$. Let us denote by $\tilde{\vartheta}$ the set of such nodes z and k_1 the sum of their order labels. The subtrees having as root branches the branches exiting from the nodes w can be considered as trees of the kind of the previous sections (i.e., with $p(w)=0$), so that the integrations corresponding to their nodes can be performed and discussed as before and a bound

$B_0^{2k_0}\eta^{-2k_0}$ follows, if $k_0=k-k_1$. All the other nodes (in $\tilde{\vartheta}$) have $p(z)\neq 0$, so that no real small divisor appears (i.e., the propagators are trivially bounded by (1). The *only delicate point to discuss* concerns the sum over the hyperbolic mode labels, but this can be done as in Ref. 5, p. 292, (or in Ref. 3, item 7 in Appendix A), to which we refer for details beyond the summary that follows.

By noting that the Laurent expansion of each function (of x and x') appearing in (4.14) and (4.16) starts from $k\geqslant -1$ and $k'\geqslant =1$, we can denote by M_1 the maximum of all such functions (multiplied by $1/x$ and $1/x'$, respectively, when $k=-1$ and $k'=-1$) in a disk of radius $\lambda=1/2$. If M_2 is the coefficient of the term with $k_v=-1$ or $k_v'=-1$, set $M=\max\{M_1,M_2\}$.

Consider the tree value (4.20). If $\sigma t<1$, the first integral (corresponding to the highest node v_0) can be split into the sum of two integral, the first one from $\sigma\infty$ to $\sigma 1$ (here the value 1 is an arbitrarily chosen positive number) and the second one from $\sigma 1$ to t. Let us denote by $I_m(\vartheta)$ the first integral and $J_m(\vartheta)$ the second one, if m is the number of nodes in ϑ.

For the nodes $v\in\tilde{\vartheta}$, one has

(1) for each node the associated propagator is bounded by 1 (as $p(v)\neq 0$),
(2) $\Pi_{v\in\tilde{\vartheta}}|\overline{F}_{\nu_v}|\leqslant (CN^2)^{2k_1}$,
(3) for each node v one has $|y_{n_v}^{(\alpha_v)}(k_v,k_v')|\leqslant M2^{k_v+k_v'}$,
(4) the last integration (on τ_{v_0}) produces a factor $\exp[-k_{v_0}'+p(v_0)]=\exp[-\Sigma_{v\in\tilde{\vartheta}}(k_v+k_v')]$.

Then the sum over the hyperbolic mode labels can be performed and gives, for each node $v\in\tilde{\vartheta}$, a factor A^2, where

$$A=\sum_{k=-1}^{\infty}\left(\frac{2}{e}\right)^k=\frac{e^2}{2(e-2)}. \tag{D1}$$

The contribution to $I_m(\vartheta)$ arising from $\tilde{\vartheta}$ a bound $B_1^{2k_1}$ is obtained, $B_1=A^2\mathcal{C}N^2M$, so that, for $I_m(\vartheta)$ a bound

$$B^{2k},\quad B\geqslant\max\{B_0\eta^{-1},B_1\} \tag{D2}$$

is obtained (see (5.6) for the meaning of $\mathcal{C}$).

By taking into account the integral $J_m(\vartheta)$, one can perform a splitting of the integration domain for the integrals corresponding to the nodes immediately preceding v_0 (now for each such nodes the second integral is from $\sigma 1$ to τ_{v_0}), and, iterating such a splitting, one finds that $\mathrm{Val}(\vartheta)$ can be written as sum of at most 2^m terms each of which has the form (for some integer p),

$$\left[\prod_{v\in\vartheta^*}\int_{\sigma 1}^{\tau_{v'}}d\tau_v(\cdots)\right]I_{m_1}(\vartheta_1)\cdots I_{m_p}(\vartheta_p), \tag{D3}$$

where $\vartheta_1,\ldots,\vartheta_p$ are disjoint subtress of ϑ and ϑ^* is the set of the m_0 nodes in ϑ not belonging to any such subtrees and $m_1+\cdots+m_p+m_0=m$. The dots between the parentheses denote the product of the functions in

$$\prod_{w\in\vartheta}Y^{(\alpha_w)}(\tau_w',\tau_w) \tag{D4}$$

which depend on τ_v, for a given $v\in\vartheta^*$, and therefore is a quantity bounded by 1 (see (4.14) and (4.16)). Note that the functions Y have *no singularity* as functions of their arguments $x=e^{\sigma g\tau}$, $x'=e^{\sigma g\tau'}$, when $x,x'=1$ (or $\tau,\tau'=0$), even though the values $x,x'=1$ lie on the convergence circle (the singularities being at $x,x'=\pm i$). Furthermore each integration from $\sigma 1$ to $\tau_{v'}$, once the integrand has been bounded, gives 1, while the integrals $I_{m_j}(\vartheta_j)$, $j=1,...,p$ can be bounded as before.

Of course, if $\sigma t>1$, the discussion is easier as no splitting of the integration domains is needed.

So we can conclude that a final bound $(2B)^{2k}$ is obtained for Val(ϑ); so far neither leaves nor counterterms have been considered.

2. Introducing the leaves and the counterterms, one sees (recall Remark 4.4) that the value of any tree ϑ can be always be written as the product of a factor like (4.28) times the product of the counterterms and of the leaf values; each counterterm can be decomposed in turn as sum of values of amputated trees (see (4.32)). As each leaf and each amputated tree can contain other leaves and counterterms we can iterate such a decomposition procedure, until, at the end, the value of the tree ϑ, with highest node v_0, turns out to be given by the product of factorizing terms which (1) either are of the form (4.28), with $\rho_w=0$ for any subtree with highest node $w<v_0$ and with $\rho_w=0,1$ if $w=v_0$, (2) or differ from (4.28) simply because no integration is performed corresponding to the highest node.

The terms as in item (1) correspond to subtrees contributing to leaf values (for $w<v_0$) and to Val(ϑ) itself (for $w=v_0$), while the terms as in item (2) correspond to amputated trees contributing to counterterms. Then a natural decomposition of the tree ϑ into subtrees $\tilde{\vartheta}$ (amputated or not) follows: each of such subtree contains neither counterterms nor leaves (by construction). Furthermore the subtrees contributing to leaves are linked to nodes of some other subtrees through their stalks, while the amputated subtrees are not linked to any node (as there is no branch exiting from the highest node). To keep memory of the node to which the counterterm label is attached we can draw a hatched line connecting the amputated subtree to such a node.

So for each subtree $\tilde{\vartheta}$ (amputated or not) one can reason as above and a bound B^{m_0} is obtained, if B is the same constant as before and if $m_0\leqslant 2k_0$ is the number of free nodes of the subtree. For all of them the resummation described in Sec. V, 5.9 has to be performed, to bound the values of the subtrees ϑ_v, with $p(v)=0$, $v\in\tilde{\vartheta}$: such a resummation is taken into account by the constant B. By collecting together all bounds one obtains, for the (normalized) sum of the values of the all trees ϑ' generated by the resummations corresponding to the families $\mathcal{F}(\vartheta_v)$, a bound B^{2k}, if k is the order of ϑ.

Therefore we are left with the sum of all possible ways to arrange leaves and counterterms. The choice of the leaves is uniquely determined by the assignments of the labels ρ_v, $v\in\vartheta$, so that it gives a factor 2^m (recall that the number of nodes m is such that $m<2k$).

In the same way one can deal with the counterterms; simply one has to distinguish between solid and hatched lines, so that another factor 2^m is produced.

3. Then the sum over all the other labels can be performed, in the same way as for the contributions without leaves and without counterterms (see the beginning of this subsection). The sum over the hyperbolic modes has been taken into account by the constant B (see (D2)); moreover

(1) the sum over the mode labels is bounded by $(2N+1)^{m(l-1)}(2N_0+1)^m$,
(2) the sum over the angle labels is bounded by l^m,
(3) the sum over the order labels is bounded by 2^m,
(4) the sum over the badge labels is bounded by 2^m.

Therefore, by taking into account that the momenta and the hyperbolic momenta are uniquely determined by the mode labels and, respectively, the hyperbolic mode labels and that the sums over the leaf labels and counterterms labels have been already considered, we are left with the sum of (unlabeled) numbered trees (see comments after (4.27)): but these are no more than $2^{2m}m!$, so that, both for $X^{k\sigma}_{j\nu}(t)$ and $\gamma_k(g_0)$, a final bound D^k is obtained, for some constant D: in terms of B the constant D is given by $D=B2^6l(2N+1)^{(l-1)}(2N_0+1)$, i.e., (5.18). In particular one has that D is proportional to η^{-2}, as B is so.

Note that, as a matter of fact, we have bounded $\Xi_{j\nu}^{h\sigma}(t)$ in (4.27) by neglecting the constraint on j and $\boldsymbol{\nu}$. Therefore, by making use of the fact that the Fourier coefficients with $|\boldsymbol{\nu}|>h(N+N_0)$ vanish at order h as a consequence of the trigonometric assumption on the perturbation f_1, see (2.2), the bound $(2B)^{2k}$ is a bound both for the Fourier coefficients of $X^\sigma(t;\boldsymbol{\alpha})$ and for the function $X^\sigma(t;\boldsymbol{\alpha})$ itself.

[1] P. Holmes, J. Marsden, and J. Scheurle, ''Exponentially small splittings of separatrices with applications to KAM theory and degenerate bifurcations,'' Contemp. Math. **81**, 213–244 (1989).
[2] L. Chierchia and G. Gallavotti, ''Drift and diffusion in phase space,'' Ann. Inst. Henri Poincaré Sec. B **60**, 1–144 (1994); **68**, 135(E) (1998).
[3] G. Gallavotti, ''Twistless KAM tori, quasiflat homoclinic intersections, and other cancellations in the perturbation series of certain completely integrable Hamiltonian systems: A review,'' Rev. Math. Phys. **6**, 343–411 (1994).
[4] G. Gallavotti, ''Twistless KAM tori,'' Commun. Math. Phys. **164**, 145–156 (1994).
[5] G. Gentile, ''Whiskered tori with prefixed frequencies and Lyapunov spectrum,'' Dynamics and Stability of Systems **10**, 269–308 (1995).
[6] G. Gallavotti, G. Gentile, and V. Mastropietro, ''Hamilton–Jacobi equation, heteroclinic chains, and Arnol'd diffusion in three time scales systems,'' (submitted) (preprint in chao-dyn No. 9801004).
[7] G. Gallavotti, *Reminiscences on science at I.H.E.S: A problem on homoclinic theory and a brief review*, Publications Scientifiques de l'IHES, Special Volume for the 40th anniversary 99-117, edited by J.-P. Bourguignon and L. Michel, Paris, 1998 (preprint in chao-dyn/9804043).
[8] G. Gallavotti, G. Gentile, and V. Mastropietro, ''Separatrix splitting for systems with three degrees of freedom,'' Commun. Math. Phys. **202**, 197–236 (1999).
[9] L. H. Eliasson, Absolutely convergent series expansions for quasiperiodic motions, Mathematical Physics Electronic Journal **2**, (1996), http://mpej.unige.ch.
[10] G. Gentile, ''A proof of existence of whiskered tori with quasi flat homoclinic intersections in a class of almost integrable Hamiltonian systems,'' Forum Mathematicum **7**, 709–753 (1995).
[11] A. I. Neishtadt, ''The separation of motions in systems with rapidly rotating phase,'' J. Appl. Math. Mech. **48**, 133–139 (1984).
[12] G. Gallavotti, G. Gentile, and V. Mastropietro, ''Melnikov's approximation dominance: Some examples,'' Rev. Math. Phys. **11**, 451–461 (1999).
[13] G. Benettin, A. Carati, and G. Gallavotti, ''A rigorous implementation of the Jeans–Landau–Teller approximation for adiabatic invariants,'' Nonlinearity **10**, 479–507 (1997).
[14] F. Bonetto, G. Gallavotti, G. Gentile, and V. Mastropietro, ''Lindstedt series, ultraviolet divergences, and Moser's theorem,'' Ann. Scuola Normale Superiore di Pisa **26**, 545–593 (1998).
[15] H. Poincaré, *Les Méthode Nouvelles de la Mécanique Céleste* (Gauthier–Villard, Paris, 1899), Vol. III.
[16] L. Chierchia and E. Zehnder, ''Asymptotic expansions of quasiperiodic motions,'' Ann. della Scuola Normale Superiore di Pisa Cl. Sci. **16**(4), 245–258 (1989).
[17] G. Gallavotti, G. Gentile, and V. Mastropietro, ''Field theory and KAM tori,'' Math. Phys. Electronic Journal **1**, 1–9 (1995), http://mpej.unige.ch.
[18] G. Gallavotti and G. Gentile, ''Majorant series convergence for twistless KAM tori,'' Ergodic Theory and Dynamical Systems **15**, 1–69 (1995).
[19] G. Gentile and V. Mastropietro, ''Methods for the analysis of the Lindstedt series for KAM tori and renormalizability in classical mechanics: A review with some applications,'' Rev. Math. Phys. **8**, 393–444 (1996).
[20] G. Gentile and V. Mastropietro, ''KAM theorem revisited,'' Physica D **90**, 225–234 (1996).
[21] V. F. Lazutkin, ''Splitting of separatrices for the Chirikov's standard map,'' mp_arc@math.utexas.edu, #98-421 (annotated English translation of a 1984 paper).
[22] V. Gelfreich, ''Reference systems for splitting of separatrices,'' Nonlinearity **10**, 175–193 (1997).
[23] V. Gelfreich, ''Melnikov method and exponentially small splitting of separatrices,'' Physica D **101**, 227–248 (1997).
[24] J. Pöschel, ''Invariant manifolds of complex analytic mappings,'' Les Houches, XLIII, edited by K. Osterwalder and R. Stora (North–Holland, Amsterdam, 1986), Vol. II, pp. 949–964.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

On construction of recursion operators from Lax representation

Metin Gürses[a]
Department of Mathematics, Faculty of Sciences, Bilkent University, 06533 Ankara—Turkey

Atalay Karasu
Department of Physics, Faculty of Arts and Sciences, Middle East Technical University, 06531 Ankara—Turkey

Vladimir V. Sokolov
Landau Institute, Moscow, 117940 Russia

(Received 28 June 1999; accepted for publication 16 August 1999)

In this work we develop a general procedure for constructing the recursion operators for nonlinear integrable equations admitting Lax representation. Several new examples are given. In particular, we find the recursion operators for some KdV-type systems of integrable equations. © *1999 American Institute of Physics.* [S0022-2488(99)03212-0]

I. INTRODUCTION

It is well known that most of the integrable nonlinear partial differential equations,

$$u_t = F(t,x,u,u_x,\ldots,u_{nx}), \tag{1}$$

admit a Lax representation,

$$L_t = [A,L], \tag{2}$$

so that the inverse scattering method is applicable. The generalized symmetries[1] of (1) have also Lax representations with the same L operator,

$$L_{t_n} = [A_n, L], \quad n \geqslant 1. \tag{3}$$

The recursion operator $\mathcal{R}$, satisfying the equation (see Ref. 2)

$$\mathcal{R}_t + [D_{\mathrm{F}}, \mathcal{R}] = 0, \tag{4}$$

where D_{F} is the Frechét derivative of the function F, generates symmetries of (1) starting from the simplest ones. In general, $\mathcal{R}$ is a nonlocal operator (a pseudodifferential operator).

The construction of the recursion operator of a given integrable system (1) is not an easy task. Several works are devoted to this subject. Among these works, most of the authors use (4) for the construction of the recursion operator.[3–8] There are several difficulties in this direct approach. The main problems are the choices of the order of $\mathcal{R}$ and the structure of the nonlocal terms. This is an approach having no relation with the Lax representation (2).

On the other hand, some of the authors used Lax representation for this purpose. Most of these works are related to the squared eigenfunctions of the Lax operator[9–13] and are based on finding an eigenvalue equation for the squared eigenfunctions of the Lax operator. The operator corresponding to this eigenvalue equation turns out to be the adjoint of the recursion operator.

[a]Electronic mail: gurses@fen.bilkent.edu.tr

There is an alternative use of the Lax representation to construct recursion operators. This approach is based on the explicit construction of the A_n operators (3). It was first used by Symes,[14] Adler[15] (see also Dorfman–Fokas,[16] Fokas–Gel'fand[17]) and Antonowicz–Fordy.[18,19] Although these authors use the Lax representation in different ways, their approach is basically the same. Symes and Adler use the Gel'fand–Dikii[20] construction of the A_n operators. On the other hand, Antonowicz–Fordy determines these operators from integrability condition (3) and by using an ansatz for A_n. Their basic aim is to determine the Hamiltonian operators θ_1 and θ_2[21] of the equations under consideration. The recursion operator is simply given by $\mathcal{R}=\theta_2\theta_1^{-1}$. Their approach is based on some explicit formulas for coefficients of the A_n operator. This is necessary to find the Hamiltonian operators θ_1 and θ_2, and it seems that this approach is quite effective to determine the bi-Hamiltonian structure for the simple cases but it becomes more complicated when the L-operator has a sophisticated structure.

If one is interested only in the determination of the recursion operator $\mathcal{R}$, we shall show in this work that it is possible to succeed this without any concrete information of the coefficients of A_n operators. We use only an ansatz $\tilde{A}=\mathcal{P}A+R$ that relates A_n operators for different n. Here $\mathcal{P}$ is some operator that commutes with the L operator and R is the remainder.

We follow this basic idea, partially used by Symes,[14] Adler.[15] Shabat and Sokolov,[22] and establish an extremely simple, effective, and algorithmic method for the construction of recursion operators when the Lax representation (2) is given.[23]

In the next section we consider the case where L is a scalar operator. We first consider the case where L is a differential operator and then the case where it is a pseudodifferential operator. In each case we present our method, discuss the reductions, and give examples for illustrations. In Sec. III we consider Lax operator taking values in a Lie algebra. We give our method both for the general case and also for the reductions. We give one example for each case in the text. Several additional examples are given in the Appendices A, B, and C corresponding to all different cases.

II. SCALAR LAX REPRESENTATIONS

First we consider equations with the scalar Lax representations of the form

$$L_t=[A,L], \tag{5}$$

where L is, in general, a pseudodifferential operator of order m and A is a differential operator whose coefficients are functions of x and t.

The different choice of operators A for a given L leads to a hierarchy of nonlinear systems (3). It is well known that one can define operators A_n by the following formula:[20]

$$A_n=(L^{n/m})_+, \tag{6}$$

where $L^{n/m}$ is a pseudodifferential series of the form $L^{n/m}=\Sigma_{-\infty}^{n} v_i D^i$ and $(L^{n/m})_+=\Sigma_{i=0}^{n} v_i D^i$. Here v_i are some concrete functions depending on the coefficients of L and D is the total derivative with respect to x.

In Refs. 25 and 26 the relationships between the Kac–Moody algebras and special types of scalar differential and pseudodifferential operators L were established. All corresponding integrable systems are Hamiltonian ones. For most of them a second Hamiltonian structure is not known up to now.

In this section and Appendices A, B, and C we consider the simplest systems from Refs. 25 and 26 as examples and find their recursion operators. In the sequel these examples will be referred to as Drinfeld–Sokolov (DS) systems. It is interesting to note that in all these examples the order of the recursion operator is equal to the Coexter number of the corresponding Kac–Moody algebra.

A. Gel'fand–Dikii systems

In this section we shall consider the case where L is a differential operator,

$$L = D^m + u_{m-2}D^{m-2} + \cdots + u_0, \tag{7}$$

where u_i, $i=0,1,...,m-2$ are functions of x, t. In the framework of Ref. 25, this corresponds to the Kac–Moody algebras of the type $A^{(1)}_{m-1}$.

To show that (3) is equivalent to a system of $(m-1)$ evolution equations with respect to u_i one can use the following standard reasoning. Set

$$L^{n/m} = (L^{n/m})_+ + (L^{n/m})_-, \tag{8}$$

where $(L^{n/m})_+$ is the differential part of the series $L^{n/m}$ and $(L^{n/m})_-$ is a series of order$\leqslant -1$. Since $[L, L^{n/m}] = 0$ we have

$$[(L^{n/m})_+, L] = [L, (L^{n/m})_-]. \tag{9}$$

The left-hand side of (9) is a differential operator, but the right side is a series of order$\leqslant n-2$. Thus, both sides of (3) are differential operators of order$\leqslant n-2$ and it is equivalent to a system of evolution equations for the dependent variables u_i, $i=0,1,...,m-2$. This system can be obtained by comparing the coefficients of D^i, where $0,...,m-2$ in (3).

Since $L^{(n+m)/m} = LL^{n/m}$, then we have

$$A_{m+n} = (LL^{n/m})_+ = L(L^{n/m})_+ + (L(L^{n/m})_-)_+, \tag{10}$$

which leads directly to

$$L_{t_{n+m}} = [A_{n+m}, L] = LL_{t_n} + [(L(L^{n/m})_-)_+, L]. \tag{11}$$

The above equation (11) has been given also by Symes[14] (see also Adler's paper[15]). In his work Symes expressed the coefficients of the both parts of (11), in a rather complicated way, in terms of some finite set of coefficients of the resolvent for an L operator. That allows him to express $L_{t_{n+m}}$ in terms of L_{t_n}. This relation gives directly the recursion operator. He gave explicit formulas for the cases $m=2$ and $m=3$.

In this section we shall show that in order to construct the recursion operator it suffices to know only that

$$L_{t_{n+m}} = LL_{t_n} + [R_n, L]. \tag{12}$$

Obviously, it follows from the following.

Proposition 1: For any n,

$$A_{n+m} = LA_n + R_n, \tag{13}$$

where R_n is a differential operator of order $\leqslant m-1$.

Proof: The relation (13) coincides with (10) if we put

$$R_n = (L(L^{n/m})_-)_+. \tag{14}$$

Since $(L^{n/m})_-$ is a series of order $\leqslant -1$, then $\mathrm{ord}(R_n) \leqslant m-1$.

Remark 1: It follows from the formula

$$A_{n+m} = (L^{n/m}L)_+ = (L^{n/m})_+L + ((L^{n/m})_-L)_+, \tag{15}$$

that

$$A_{n+m}=A_nL+\overline{R_n}\,, \tag{16}$$

and

$$L_{t_{n+m}}=L_{t_n}L+[L,\overline{R_n}], \tag{17}$$

where $\overline{R_n}$ is a differential operator of order $\leqslant m-1$.

To find the recursion operator we can simply equate the coefficients of different powers of D in (12). It is easy to see that in this comparison of the coefficients of D^i, $i=2m-2,...,m-1$ we determine R_n in terms of the coefficients of operators L and L_{t_n}. It is important that the resulting formulas turn out to be linear in the coefficients of L_{t_n}. The remaining coefficients of D^i, $i=m-2,...,0$ in (12) give us the relation

$$\begin{pmatrix} u_0 \\ \cdot \\ \cdot \\ \cdot \\ u_{m-2} \end{pmatrix}_{t_{n+m}} =\mathcal{R}\begin{pmatrix} u_0 \\ \cdot \\ \cdot \\ \cdot \\ u_{m-2} \end{pmatrix}_{t_n}, \tag{18}$$

where $\mathcal{R}$ is a recursion operator. Instead of (12) one can use (17). The corresponding recursion operators coincide.

Example 1. KdV equation: The KdV equation,

$$u_t=\tfrac{1}{4}(u_{3x}+6uu_x), \tag{19}$$

has a Lax representation with

$$L=D^2+u, \quad A=(L^{3/2})_+\,. \tag{20}$$

Since in this case $L_{t_{n+2}}=u_{t_{n+2}}\equiv u_{n+2}$ and $L_{t_n}=u_{t_n}\equiv u_n$, the main relation (12) takes the form

$$u_{n+2}=(D^2+u)\cdot u_n+[R_n,L], \tag{21}$$

with $R_n=a_nD+b_n$.

Now if we equate successively to zero the coefficients of D^2, D, and D^0 in the above equation, we obtain

$$a_n=\tfrac{1}{2}D^{-1}(u_n), \quad b_n=\tfrac{3}{4}u_n\,,$$

and

$$u_{n+2}=(\tfrac{1}{4}D^2+u+\tfrac{1}{2}u_xD^{-1})u_n\,,$$

that gives the standard recursion operator for the KdV equation,

$$\mathcal{R}=\tfrac{1}{4}D^2+u+\tfrac{1}{2}u_xD^{-1}. \tag{22}$$

In the same way one can find a recursion operator for the Boussinesq equation (see Appendix A).

B. Symmetric and skew-symmetric reductions of a differential Lax operator

The standard reductions of the Gel'fand–Dikii systems are given by the conditions $L^*=L$ or $L^*=-L$. Here $*$ denotes the adjoint operation defined as follows. Let L be a differential operator,

$L=\Sigma a_i D^i$. Its adjoint L^* is given by $L^*=\Sigma(-D)^i \cdot a_i$. It is easy to see that if $L^*=L$ then $m=\mathrm{ord}(L)$ must be an even integer. For the case $L^*=-L$, it must be an odd integer.

It is well known that for both reductions all possible A_n are defined by (6), where n takes odd integer values. This condition provides that $(A_n)^*=-A_n$ that is necessary for (3) to be compatible.

If $L^*=L$, the formula $A_{n+m}=(LL^{n/m})_+=(L^{(n+m)/m})_+$ gives a correct A_n operator since $n+m$ is an odd integer. Thus, in this case Proposition 1 remains valid and the recursion operator can be found from (12) or (17).

On the other hand, if $L^*=-L$ then both integers m and n are odd and hence their sum $m+n$ is an even integer. This means that $(L^{(n+m)/m})_+$ cannot be taken as an A_n operator. In this (skew adjoint) case we must take

$$A_{n+2m}=(L^{(n+2m)/m})_+=(L^2L^{n/m})_+ ,$$

to find the recursion operator. Following the proof of Proposition 1 we obtain Proposition 2.

Proposition 2: If $L^*=-L$ then

$$A_{n+2m}=L^2A_n+R_n , \tag{23}$$

where $\mathrm{ord}(R_n)<2\,\mathrm{ord}(L)$. It follows from (23) that

$$L_{t_{n+2m}}=L^2L_{t_n}+[R_n ,L]. \tag{24}$$

Remark 2: Instead of (23) we can use the ansatz

$$A_{n+2m}=LA_nL+\tilde{R}_n , \tag{25}$$

or

$$A_{n+2m}=A_nL^2+\tilde{\tilde{R}}_n . \tag{26}$$

The recursion operators obtained by the utility of (23), (25), and (26) all coincide.

In the works,[25,26] more general reductions $L^\dagger=\pm L$ were also considered. Here $L^\dagger=KL^*K^{-1}$, where K is a given differential operator, such that LK^{-1} is a differential operator. In this general reductions, as well, possible A_n operators are given by (6), with n being an odd integer. Propositions 1 and 2 are valid for this general symmetric and skew-symmetric cases and hence one can use Eqs. (12), (24) accordingly to obtain the recursion operators.

Example 2. Kupershmidt equation: This equation,

$$u_t=u_{5x}+10uu_{3x}+25u_xu_{2x}+20u^2u_x , \tag{27}$$

has the Lax pair

$$L=D^3+2uD+u_x , \quad A=(L^{5/3})_+ . \tag{28}$$

In this case $L^*=-L$; therefore we use Eq. (24) with

$$\tilde{R}_n=a_nD^5+b_nD^4+c_nD^3+d_nD^2+e_nD+f_n . \tag{29}$$

By equating the coefficients of powers of D in (24), we obtain

$$a_n=\tfrac{2}{3}D^{-1}(u_n), \quad b_n=\tfrac{11}{3}u_n , \quad c_n=\tfrac{1}{9}(20uD^{-1}(u_n)+73u_{n,x}),$$

$$d_n=\tfrac{1}{3}(10u_xD^{-1}(u_n)+22uu_n+27u_{n,2x}),$$

$$e_n=\tfrac{1}{27}(70u_{2x}D^{-1}(u_n)-2D^{-1}(u_{2x}u_n)+40u^2D^{-1}(u_n)-8D^{-1}(u^2u_n)$$
$$+134u_{n,3x}+212uu_{n,x}+184u_xu_n),$$

$$f_{n,x}=\tfrac{1}{27}(20u_{4x}D^{-1}(u_n)+74u_{3x}u_n+126u_{2x}u_{n,x}+40uu_{2x}D^{-1}(u_n)+40u_x^2D^{-1}(u_n)$$
$$+136u_xu_{n,2x}+27uu_xu_n+28u_{n,5x}+64uu_{n,3x}+16u^2u_{n,x}),$$

and the recursion operator for the Kupershmidt equation:

$$\mathcal{R}=D^6+12uD^4+36u_xD^3+(49u_{2x}+36u^2)D^2+5(7u_{3x}+24uu_x)D+13u_{4x}+82uu_{2x}+69u_x^2$$
$$+32u^3+2u_xD^{-1}(u_{2x}+4u^2)+2(u_{5x}+10uu_{3x}+25u_xu_{2x}+20u^2u_x)D^{-1}. \tag{30}$$

C. Pseudodifferential Lax operator

In this section we generalize our scheme to the case of pseudodifferential Lax operators. The only difference is that in formulas like (13) and (23) the R_n operator also becomes a pseudodifferential operator.

It follows from these formulas that the structure of the nonlocal terms in R_n is, in general, similar to the nonlocal terms in L since A_{n+m} and A_n are differential operators.

For skew-symmetric case, A_n may be defined by either (23) or (25), or (26). In the pseudo-differential case they are not equivalent, in the sense that the nonlocal part of R_n depends on which ansatz we choose. For illustration, let us consider the case $L=MD^{-1}$, where M is a differential operator. The following lemma shows that if $L^\dagger=L$ or $L^\dagger=-L$, where

$$L^\dagger=DL^*D^{-1}, \tag{31}$$

then the formulas (13) and (25) are much suitable then (16), (23), and (26).

Lemma: Let $L^\dagger=\epsilon L$, where $\epsilon=\pm1$. Then

$$R_n=D^{m-1}+\cdots+a_0,\quad \text{for}\ \ \epsilon=1, \tag{32}$$

where R_n is defined by (13), and

$$\tilde{R}_n=D^{2m-1}+\cdots+a_{-1}D^{-1},\quad \text{for}\ \ \epsilon=-1, \tag{33}$$

where $\tilde{R}_n$ is defined by (25).

Proof: If $L=MD^{-1}$ then $L^\dagger=\epsilon L$ implies $M^*=-\epsilon M$. It is easy to show that $(L^{1/m})^\dagger=-L^{1/m}$. Hence $(L^{n/m})^\dagger=-L^{n/m}$ for an odd integer n. Define now a series K_n by

$$L^{n/m}=DK_n.$$

It is easy to prove that $K_n^*=K_n$. Since $K_n=(K_n)_++(K_n)_-$ and $(K_n)^*=K_n$, we have

$$(K_n)_+^*=(K_n)_+,\quad (K_n)_-^*=(K_n)_-.$$

From the last formula it follows that $\mathrm{ord}(K_n)_-\leqslant-2$, which leads to an important result,

$$A_n=(L^{n/m})_+=D(K_n)_+.$$

This implies that

$$LA_n=M(K_n)_+ \tag{34}$$

is a differential operator. Now using (34) in (13) and (25) for the cases $\epsilon=1$ and $\epsilon=-1$, respectively, we find the ansatz for A_n given by (32) and (33).

Example 3 ($\epsilon=-1$)*:* It is known that the KdV equation has, besides the standard Lax representation, the following Lax pair:

$$L=(D^2+u)D^{-1}, \quad A=(L^3)_+ . \tag{35}$$

The L operator satisfies the reduction $L^\dagger=-L$. According to the formula (33) we have

$$\tilde{R}_n=a_nD+b_n+c_nD^{-1}.$$

It follows from (25) that

$$a_n=D^{-1}(u_n), \quad b_n=u_n , \quad c_n=-u_{n,x}-uD^{-1}(u_n).$$

The remaining equation in (25) gives the recursion operator

$$\mathcal{R}=D^2+4u+2u_xD^{-1}. \tag{36}$$

Example 4 ($\epsilon=1$). *DSIII system:* The DSIII system[25,26] is given by

$$\begin{aligned} u_t&=-u_{3x}+6uu_x+6v_x , \\ v_t&=2v_{3x}-6uv_x . \end{aligned} \tag{37}$$

The nonlocal Lax representation for this system is

$$\begin{aligned} L&=(D^5-2uD^3-2D^3u-2Dw-2wD)D^{-1}, \\ A&=(L^{3/4})_+ , \end{aligned} \tag{38}$$

where $w=v-u_{2x}$. Since $L^\dagger=L$ we can use (32), which gives us

$$R_n=a_nD^3+b_nD^2+c_nD+d_n . \tag{39}$$

By equating the coefficients of the powers of D in (25), we obtain

$$a_n=D^{-1}(u_n), \quad b_n=4u_n ,$$

$$c_n=\tfrac{1}{2}(-6uD^{-1}(u_n)+11u_{n,x}+2D^{-1}(uu_n)+2D^{-1}(v_n)),$$

$$d_{n,x}=-\tfrac{1}{2}(6u_{2x}D^{-1}(u_n)+10u_xu_n-5u_{n,3x}+4uu_{n,x}-6v_{n,x}).$$

The recursion operator of the DSIII is found as

$$\mathcal{R}=\begin{pmatrix} \mathcal{R}_0^0 & \mathcal{R}_1^0 \\ \mathcal{R}_0^1 & \mathcal{R}_1^1 \end{pmatrix}, \tag{40}$$

with

$$\begin{aligned} \mathcal{R}_0^0&=D^4-8uD^2-12u_xD-8u_{2x}+16u^2+16v+(-2u_{3x}+12uu_x+12v_x)D^{-1}+4u_xD^{-1}u, \\ \mathcal{R}_1^0&=-10D^2+8u+4u_xD^{-1}, \\ \mathcal{R}_0^1&=10v_xD+12v_{2x}+(4v_{3x}-12uv_x)D^{-1}+4v_xD^{-1}u, \\ \mathcal{R}_1^1&=-4D^4+16uD^2+8u_xD+16v+4v_xD^{-1}. \end{aligned} \tag{41}$$

This recursion operator has recently been given in Ref. 6.

III. MATRIX *L* OPERATOR OF THE FIRST ORDER

In this section we demonstrate how our approach, given in the previous sections, can be generalized to the case where L is a matrix operator of the form

$$L=D_x+\lambda a+q(x,t). \tag{42}$$

A. General case

Let us consider the Lax operator (42), where q and a belong to a Lie algebra g and λ is the spectral parameter. The constant element a is supposed to be such that

$$g=\mathrm{Ker}(\mathrm{ad}_a)\oplus \mathrm{Im}(\mathrm{ad}_a). \tag{43}$$

First, let us recall the procedure[25] of constructing the A operators for the Lax operator (42).

Proposition 3: There exist unique series,

$$u=u_{-1}\lambda^{-1}+u_{-2}\lambda^{-2}+\cdots, \quad u_i\in \mathrm{Im}(\mathrm{ad}_a), \tag{44}$$

$$h=h_0+h_{-1}\lambda^{-1}+h_{-2}\lambda^{-2}+\cdots, \quad h_i\in \mathrm{Ker}(\mathrm{ad}_a), \tag{45}$$

such that

$$e^{\mathrm{ad}_u}(L)=L+[u,L]+\tfrac{1}{2}[u,[u,L]]+\cdots=D_x+a\lambda+h. \tag{46}$$

Let b be a constant element of g such that $[b,\mathrm{Ker}(\mathrm{ad}_a)]=\{0\}$. It follows from (45) that $[b\lambda^n,D_x+a\lambda+h]=0$. Hence $[\Phi_{b,n},L]=0$, where

$$\Phi_{b,n}=e^{-\mathrm{ad}_u}(b\lambda^n). \tag{47}$$

Then the corresponding A operator of the form

$$A_{b,n}=b\lambda^n+a_{n-1}\lambda^{n-1}+\cdots+a_0, \tag{48}$$

is defined by the formula

$$A_{b,n}=(\Phi_{b,n})_+, \tag{49}$$

where

$$\left(\Sigma_{-\infty}^{n}\alpha_i\lambda^i\right)_+=\Sigma_0^n\alpha_i\lambda^i. \tag{50}$$

According to (47),

$$\Phi_{b,n+1}=\lambda\Phi_{b,n}. \tag{51}$$

Hence

$$A_{b,n+1}=(\lambda\Phi_{b,n})_+=\lambda(\Phi_{b,n})_++(\lambda(\Phi_{b,n})_-)_+. \tag{52}$$

The last formula shows that

$$A_{b,n+1}=\lambda A_{b,n}+R_n, \quad R_n\in g, \tag{53}$$

where R_n does not depend on λ. Substituting (53) into the Lax equation $L_{t_{n+1}}=[A_{b,n+1},L]$, we get

$$L_{t_{n+1}}=\lambda L_{t_n}+[R_n,L]. \tag{54}$$

Using the ansatz (54), one can easily find the corresponding recursion operator.

Example 5: The system

$$u_t=-\tfrac{1}{2}u_{xx}+u^2v,$$
$$v_t=\tfrac{1}{2}v_{xx}-v^2u, \tag{55}$$

is equivalent to the nonlinear Schrödinger equation, has a Lax operator

$$L=D+\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\lambda+\begin{pmatrix}0 & u\\ v & 0\end{pmatrix}. \tag{56}$$

The Lie algebra g in this example coincides with $sl(2)$.

Using (54) with

$$R_n=\begin{pmatrix}a_n & b_n\\ c_n & -a_n\end{pmatrix},$$

we find that

$$a_n=\tfrac{1}{2}D^{-1}(vu_n+uv_n),$$
$$b_n=\tfrac{1}{2}u_n,\quad c_n=-\tfrac{1}{2}v_n,$$

and the recursion operator of the system (55) is given by

$$\mathcal{R}=\begin{pmatrix}-\tfrac{1}{2}D+uD^{-1}v & uD^{-1}u\\ -vD^{-1}v & \tfrac{1}{2}D-vD^{-1}u\end{pmatrix}. \tag{57}$$

B. Reductions in matrix case

In the general case considered in the previous section the A_n operators belong to the Lie algebra,

$$\mathfrak{a}_+=\{\Sigma_{i=0}^{\kappa}a_i\lambda^i,\ \ a_i\in g,\ \ \kappa\in Z_+\}, \tag{58}$$

that is a subalgebra of the Lie algebra,

$$\mathfrak{a}=\{\Sigma_{-\infty}^{\kappa}a_i\lambda^i,\ \ \alpha_i\in g,\ \ \kappa\in Z\}. \tag{59}$$

A standard σ reduction is defined by any automorphism σ of the Lie algebra g of finite order κ. Because $\sigma^\kappa=\mathrm{Id}$, the eigenvalues of σ are $\epsilon^i, i=0,...,\kappa-1$, where ϵ is a primitive κ root of unity.

Let g_i be an eigenspace corresponding to eigenvalue ϵ^i. Then the following reduction $a_j\in g_i$, where $i=j(\mathrm{mod}\ \kappa)$ in (58) and (59) is compatible with Eqs. (3). Note that according to this definition $a\in g_1$, and the potential $q(x,t)$ in (42) belongs to g_0 or, the same, satisfies $\sigma(q)=q$.

It is easy to see that, to satisfy such a reduction, we must use the ansatz

$$A_{b,n+\kappa}=\lambda^\kappa A_{b,n}+R_n, \tag{60}$$

where

$$R_n=r_{\kappa-1}\lambda^{\kappa-1}+\cdots+r_0,\quad r_i\in g_i. \tag{61}$$

Further generalizations are associated with modifications of sign ‘‘+’’ in (50), which corresponds to the simplest decomposition of algebra $\mathfrak{a}$ into the direct sum of two subalgebras,

$$\mathfrak{a}=\mathfrak{a}_+\oplus\mathfrak{a}_-\,, \tag{62}$$

where $\mathfrak{a}_+$ is given by (58) and

$$\mathfrak{a}_-=\{\Sigma_{-\infty}^{-1}a_i\lambda^i,\ \ a_i\in g\}. \tag{63}$$

The sign ''+'' in (50) is the projection of onto $\mathfrak{a}_+$ parallel to $\mathfrak{a}_-$. If we have a different decomposition (62), then the construction from Proposition 3 is also valid, but we have the following condition:

$$R_n\in\mathfrak{a}_+\cap\lambda\,\mathfrak{a}_-\,, \tag{64}$$

instead of $R_n\in g$. If we also have the σ reduction, we must use the most general ansatz (60), where

$$R_n\in\mathfrak{a}_+\cap\lambda^{\kappa}\mathfrak{a}_-\,. \tag{65}$$

Example 6: Let us consider the following equation:

$$u_t=\tfrac{1}{4}u_{xxx}-\tfrac{3}{8}u_{xx}u+\tfrac{3}{8}uu_{xx}-\tfrac{3}{8}uu_xu, \tag{66}$$

where u is a square matrix of arbitrary size, or more generally, u belongs to an arbitrary associative algebra $\mathcal{K}$. This equation has a Lax representation with

$$L=D+\begin{pmatrix}0&\mathbf{1}\\ \mathbf{1}&0\end{pmatrix}\lambda+\begin{pmatrix}u&0\\0&0\end{pmatrix}. \tag{67}$$

Here $\mathbf{1}$ is the unity of $\mathcal{K}$. The reduction (67) can be described as follows (see Ref. 27). The Lie algebra g is the algebra of all 2×2 matrices with entries belonging to $\mathcal{K}$. The automorphism σ is defined by

$$\sigma(X)=TXT^{-1}, \tag{68}$$

where

$$T=\begin{pmatrix}\mathbf{1}&0\\0&-\mathbf{1}\end{pmatrix}.$$

Obviously $\sigma^2=\mathrm{Id}$ and eigenvalues of σ are 1 and -1. The corresponding eigenspaces are

$$g_0=\begin{Bmatrix}*&0\\0&*\end{Bmatrix},\quad g_1=\begin{Bmatrix}0&*\\ *&0\end{Bmatrix}, \tag{69}$$

and therefore the coefficients a_i in (59) have the following structure:

$$a_{2j}=\begin{pmatrix}*&0\\0&*\end{pmatrix},\quad a_{2j+1}=\begin{pmatrix}0&*\\ *&0\end{pmatrix}. \tag{70}$$

The subalgebra $\mathfrak{a}_+$ is given by (58), where the coefficients have the structure (70) and, additionally,

$$a_0=\begin{pmatrix}*&0\\0&0\end{pmatrix}.$$

The subalgebra $\mathfrak{a}_-$ has the following form:

$$\mathfrak{a}_- = \Sigma_{-\infty}^{0} a_i \lambda^i, \tag{71}$$

where a_0 is of the form

$$a_0 = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}, \quad \alpha \in \mathcal{K}.$$

The A operator for (66) is given by formula $A = (\Phi_{a,3})_+$ [see (49)], where

$$a = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix},$$

and ''+'' means the projection onto $\mathfrak{a}_+$ parallel to $\mathfrak{a}_-$.

According to (65), R_n is of the form

$$R_n = \begin{pmatrix} a_n & 0 \\ 0 & a_n \end{pmatrix} \lambda^2 + \begin{pmatrix} 0 & b_n \\ c_n & 0 \end{pmatrix} \lambda + \begin{pmatrix} d_n & 0 \\ 0 & 0 \end{pmatrix}. \tag{72}$$

It follows from

$$L_{t_{n+2}} = \lambda^2 L_{t_n} + [R_n, L], \tag{73}$$

that

$$u_n - a_{n,x} + [a_n, u] + b_n - c_n = 0 \quad c_n - b_n - a_{n,x} = 0,$$

$$d_n - b_{n,x} - u b_n = 0, \quad d_n + c_{n,x} - c_n u = 0,$$

$$u_{n+2} = -d_{n,x} + [d_x, u].$$

Finding a_n, b_n, c_n, and d_n from this system, we obtain the following recursion operator:

$$\mathcal{R} = -(D + \mathrm{ad}_u)(-D + R_u)(2D + \mathrm{ad}_u)^{-1}(D + L_u)D(2D + \mathrm{ad}_u)^{-1}, \tag{74}$$

where R_u and L_u are the operators of right and left multiplications by u, respectively.

Note that in the commutative case (66) coincides with the modified KdV equation. It is easy to verify that (74) becomes the standard recursion operator of a modified KdV equation. All factors in (74) have to be regarded as operators acting on a (noncommutative) polynomial depending on $u, u_x, u_{xx}, \ldots$.

IV. CONCLUSION

In this work we devoted ourselves in the construction of recursion operators when the Lax representation is given. We have shown that our approach can be easily generalized to all cases where the L operator is a polynomial of λ. It would be interesting to generalize it for the cases of more complicated λ dependence of L as well as for the cases of $2+1$-dimensional equations, Toda-type lattices, and ordinary differential equations.

ACKNOWLEDGMENTS

We would like to thank Dr. Jing Ping Wang for reading the manuscript and pointing out some misprints. This work is partially supported by the Scientific and Technical Research Council of Turkey (TUBITAK). M. G. is a member of the Turkish Academy of Sciences (TUBA). V. S. is supported by Russian Foundation of Basic Researches (RFBR) Grant No. 99-01-00294 and INTAS.

APPENDIX A: EXAMPLE TO SEC. II A

The Boussinesq equation,

$$u_{tt} = -\tfrac{1}{3}(u_{4x} + 2(u^2)_{2x}), \tag{A1}$$

can be expressed in the form of a pair of first-order evolution equations,

$$u_t = v_x,$$
$$v_t = -\tfrac{1}{3}(u_{3x} + 8uu_x). \tag{A2}$$

This system has a Lax pair,

$$L = D^3 + 2uD + u_x + v, \quad A = (L^{2/3})_+. \tag{A3}$$

To construct the recursion operator for this system, we use Eq. (12) with the differential operator,

$$R_n = a_n D^2 + b_n D + c_n.$$

By equating the coefficients of the powers of D in (12), we find

$$a_n = \tfrac{2}{3}D^{-1}(u_n), \quad b_n = \tfrac{1}{3}(5u_n + D^{-1}(v_n)),$$
$$c_n = \tfrac{1}{9}(6v_n + 8uD^{-1}(u_n) + 10u_{n,x}),$$

and after that we obtain the recursion operator of the form (40) for (A2) with

$$\begin{aligned} \mathcal{R}_0^0 &= 3v + 2v_x D^{-1}, \\ \mathcal{R}_1^0 &= D^2 + 2u + u_x D^{-1}, \\ \mathcal{R}_0^1 &= -(\tfrac{1}{3}D^4 + \tfrac{10}{3}uD^2 + 5u_x D + 3u_{2x} + \tfrac{16}{3}u^2 + (\tfrac{2}{3}u_{3x} + \tfrac{16}{3}uu_x)D^{-1}), \\ \mathcal{R}_1^1 &= 3v + v_x D^{-1}. \end{aligned} \tag{A4}$$

APPENDIX B: EXAMPLES TO SEC. II B

1. Sawada–Kotera equation

The Lax pair for the Sawada–Kotera equation,[28]

$$u_t = u_{5x} + 5uu_{3x} + 5u_x u_{2x} + 5u^2 u_x, \tag{B1}$$

is given by

$$L = D^3 + uD, \quad A = (L^{5/3})_+. \tag{B2}$$

In this example, $L^\dagger = -L$, where $L^\dagger = D^{-1}L^* D$ and L is skew-symmetric, then we use (24). The operator $\tilde{R}_n$ has the same form as (29), with the coefficients given by

$$a_n = \tfrac{1}{3}D^{-1}(u_n), \quad b_n = \tfrac{5}{3}u_n, \quad c_n = \tfrac{1}{9}(5uD^{-1}(u_n) + 29u_{n,x}),$$
$$d_n = \tfrac{1}{9}(5u_x D^{-1}(u_n) + 14uu_n + 26u_{n,2x}),$$

$$e_n=\tfrac{1}{27}(10u_{2x}D^{-1}(u_n)-2D^{-1}(u_{2x}u_n)-D^{-1}(u^2u_n)+5u^2D^{-1}(u_n)$$
$$+28u_{n,3x}+32uu_{n,x}+32u_xu_n),$$
$$f_n=0.$$

The recursion operator is given as

$$\mathcal{R}=D^6+6uD^4+9u_xD^3+(9u^2+11u_{2x})D^2+(10u_{3x}+21uu_x)D+5u_{4x}+16uu_{2x}+6u_x^2+4u^3$$
$$+(u_{5x}+5uu_{3x}+5u_xu_{2x}+5u^2u_x)D^{-1}+u_xD^{-1}(u^2+2u_{2x}). \tag{B3}$$

2. DSI system

The DSI system,[25,26]

$$u_t=3vv_x,$$
$$v_t=2v_{3x}+2uv_x+vu_x, \tag{B4}$$

has a Lax representation with

$$L=[D^3+(u+v)D+\tfrac{1}{2}(u+v)_x][D^3+(u-v)D+\tfrac{1}{2}(u-v)_x],$$
$$A=(L^{1/2})_+. \tag{B5}$$

Here R_n is a differential operator of order 5, and since L is symmetric we again use Eq. (12). The expressions for the coefficients of the operator R_n are very long and complicated. Hence we do not display them here. We find that the recursion operator $\mathcal{R}$ of this system is of the form (40), where

$$\mathcal{R}_0^0=-4D^6-24uD^4-27u_xD^3+2(-49u_{2x}-18u^2+42v^2)D^2+10(-7u_{3x}-12uu_x+30vv_x)D$$
$$-26u_{4x}-82uu_{2x}-69u_x^2+222vv_x+141v_x^2-16u^3+48v^2u$$
$$+2(-2u_{5x}-10uu_{3x}-25u_xu_{2x}-10u^2u_x+15v^2u_x+30vv_{3x}+45v_xv_{2x}+30uvv_x)D^{-1}$$
$$+2u_xD^{-1}(3v^2-2u^2-u_{2x}),$$

$$\mathcal{R}_1^0=168vD^4+204vD^3+6(21v_{2x}+32uv)D^2+6(40vu_x+7v_{3x}+22uv_x)D$$
$$+6(13vu_{2x}+10u_xv_x+v_{4x}+5uv_{2x}+4vu^2+12v^3)+108vv_xD^{-1}v+2u_xD^{-1}(6uv+9v_{2x}), \tag{B6}$$
$$\mathcal{R}_0^1=56vD^4+268v_xD^3+2(243v_{2x}+32uv)D^2+2(36vu_x+219v_{3x}+106uv_x)D$$
$$+2(27vu_{2x}+92u_xv_x+99v_{ax}+99uv_{2x}+4vu^2+12v^3)+2(10vu_{3x}+35u_{2x}v_x+45u_xv_{2x}$$
$$+10uvu_x+18v_{5x}+30uv_{3x}+10u^2v_x+15v^2v_x)D^{-1}+2v_xD^{-1}(3v^2-2u^2-u_{2x}),$$

$$\mathcal{R}_1^1=108D^6+216uD^4+432u_xD^3+6(81u_{2x}+18u^2+22v^2)D^2+6(45u_{3x}+36uu_x+70vv_x)D$$
$$+3(18u_{4x}+18uu_{2x}+9u_x^2+98vv_{2x}+67v_x^2+32uv^2)+36(2v_{3x}+2v_xu+vu_x)D^{-1}v$$
$$+2v_xD^{-1}(6uv+9v_{2x}).$$

3. DSII system

The DSII system,[25,26]

$$u_t = 3v_x,$$
$$v_t = -2(v_{3x} + uv_x + vu_x), \tag{B7}$$

has a Lax representation with

$$L = (D^5 + uD^3 + D^3u + (v + \tfrac{1}{2}u^2)D + D(v + \tfrac{1}{2}u^2))D,$$
$$A = (L^{1/2})_+ . \tag{B8}$$

Since L is symmetric we again use Eq. (12). In this case the operator R_n is given as follows:

$$R_n = a_n D^5 + b_n D^4 + c_n D^3 + d_n D^2 + e_n D, \tag{B9}$$

where

$$a_n = \tfrac{1}{3}D^{-1}(u_n), \quad b_n = \tfrac{5}{3}u_n,$$
$$c_n = \tfrac{1}{9}[5uD^{-1}(u_n) + 3D^{-1}(v_n) + 29u_{n,x}],$$
$$d_n = \tfrac{1}{9}[5u_xD^{-1}(u_n) + 26u_{n,2x} + 14uu_n + 12v_n],$$
$$e_n = \tfrac{1}{27}[5(2u_{2x} + u^2 + 3v)D^{-1}(u_n) - 3D^{-1}(vu_n + uv_n) + 9uD^{-1}(v_n)$$
$$-2D^{-1}(u_{2x}u_n + \tfrac{1}{2}u^2u_n) + 54u_xu_n + 28u_{n,3x} + 32(uu_{n,x} - u_nu_x) + 42v_{n,x}].$$

The recursion operator (40) for the system can be found as[29]

$$\mathcal{R}_0^0 = -D^6 - 6uD^4 - 9u_xD^3 - (11u_{2x} + 9u^2 + 42v)D^2 + (-10u_{3x} - 21uu_x - 30v_x)D$$
$$-5u_{4x} - 16uu_{2x} - 6u_x^2 - 60v_{2x} - 4u^3 - 24vu + (-u_{5x} - 5uu_{3x} - 5u_xu_{2x}$$
$$-5u^2u_x - 15vu_x - 15v_{3x} - 15uv_x)D^{-1} - u_xD^{-1}(2u_{2x} + u^2 + 3v),$$
$$\mathcal{R}_1^0 = -42D^4 - 48uD^2 - 87u_xD - 6(7u_{2x} + u^2 - 6v) + 27v_xD^{-1} - 3u_xD^{-1}u,$$
$$\mathcal{R}_0^1 = 28vD^4 + 106v_xD^3 + (165v_{2x} + 32uv)D^2 + (54vu_x + 132v_{3x} + 74v_xu)D + 30vu_{2x} + 79u_xv_x$$
$$+54v_{4x} + 57uv_{2x} + 4u^2v - 24v^2 + (10vu_{3x} + 25v_xu_{2x} + 30u_xv_{2x} + 10uvu_x + 9v_{5x} + 15uv_{3x}$$
$$+5u^2v_x - 15vv_x)D^{-1} - v_xD^{-1}(3v + u^2 + 2u_{2x}),$$
$$\mathcal{R}_1^1 = 27D^6 + 54uD^4 + 135u_xD^3 + 3(54u_{2x} + 9u^2 - 22v)D^2 + 3(36u_{3x} + 27uu_x - 28v_x)D$$
$$+3(9u_{4x} + 9uu_{2x} + 9u_x^2 - 21v_{2x} - 16vu) - 18(v_{3x} + u_xv + v_xu)D^{-1} - 3v_xD^{-1}u. \tag{B10}$$

4. DSIV system

The DSIV system,[25,26] which is also known as the Hirota–Satsuma system,[30,31]

$$u_t = \tfrac{1}{2}u_{3x} + 3uu_x - 6vv_x,$$
$$v_t = -v_{3x} - 3uv_x, \tag{B11}$$

has Lax representation with

$$L = (D^2 + u + v)(D^2 + u - v), \quad A = (L^{3/4})_+ . \tag{B12}$$

Since the operator L is symmetric we use Eq. (12). In this case the operator R_n has the same form as (39), with coefficients given by

$$a_n=\tfrac{1}{2}D^{-1}(u_n), \quad b_n=\tfrac{7}{4}u_n-\tfrac{1}{2}v_n,$$

$$c_n=\tfrac{1}{8}[6uD^{-1}(u_n)+2D^{-1}(uu_n)-4D^{-1}(vv_n)+17u_{n,x}-12v_{n,x}],$$

$$d_{n,x}=\tfrac{1}{16}[6u_{2x}D^{-1}(u_n)-12v_{2x}D^{-1}(u_n)+30u_xu_n-8u_xv_n+24uu_{n,x}$$
$$+15u_{n,3x}-12v_xv_n-8uv_{n,x}-20vv_{n,x}-28v_{n,3x}].$$

The recursion operator (40) for the given system is

$$\begin{aligned}
\mathcal{R}_0^0&=\tfrac{1}{4}D^4+2uD^2+3u_xD+2u_{2x}+4(u^2-v^2)+(3uu_x-6vv_x+\tfrac{1}{2}u_{3x})D^{-1}+u_xD^{-1}u,\\
\mathcal{R}_1^0&=-5vD^2-4v_xD-v_{2x}-4uv-2u_xD^{-1}v,\\
\mathcal{R}_0^1&=-\tfrac{5}{2}v_xD-3v_{2x}-(v_{3x}+3uv_x)D^{-1}+v_xD^{-1}u,\\
\mathcal{R}_1^1&=-D^4-4uD^2-2u_xD-4v^2-2v_xD^{-1}v.
\end{aligned} \tag{B13}$$

5. $N=3$ Hirota–Satsuma system

This system is given by[29]

$$\begin{aligned}
u_t&=\tfrac{1}{4}u_{3x}+3uu_x+3(-v^2+w)_x,\\
v_t&=-\tfrac{1}{2}v_{3x}-3uv_x,\\
w_t&=-\tfrac{1}{2}w_{3x}-3uw_x.
\end{aligned} \tag{B14}$$

This is an example for the $N=3$ system that covers some other $N=2$ systems as special cases. For instance, letting $w=0$, we get DSIV and letting $v=0$ we get DSIII systems.

The corresponding Lax pair is

$$L=(D^2+2u-2v)(D^2+2u+2v)+4w, \quad A=(L^{3/4})_+. \tag{B15}$$

In this case the operator L is symmetric and hence R_n has the same form as (39), with the coefficients

$$a_n=D^{-1}(u_n), \quad b_n=\tfrac{7}{2}u_n+v_n,$$

$$c_n=\tfrac{1}{4}[12uD^{-1}(u_n)+4D^{-1}(uu_n+w_n-2vv_n)+17u_{n,x}+12v_{n,x}],$$

$$d_{n,x}=\tfrac{1}{8}[12u_{2x}D^{-1}(u_n)+24v_{2x}D^{-1}(u_n)+60u_xu_n+16u_xv_n+15u_{n,3x}+48uu_{n,x}+24v_xu_n$$
$$-40v_xv_n+20v_{n,3x}+16vv_{n,x}+20w_{n,x}].$$

The recursion operator is given by

$$\mathcal{R}=\begin{pmatrix} \mathcal{R}_0^0 & \mathcal{R}_1^0 & \mathcal{R}_2^0 \\ \mathcal{R}_0^1 & \mathcal{R}_1^1 & \mathcal{R}_2^1 \\ \mathcal{R}_0^2 & \mathcal{R}_1^2 & \mathcal{R}_2^2 \end{pmatrix}, \tag{B16}$$

where

$$\mathcal{R}_0^0 = \tfrac{1}{4}D^4 + 4uD^2 + 6u_xD + 4(u_{2x} + 4u^2 - 4v^2 + 4w) + 4(\tfrac{1}{4}u_{3x} + 3uu_x - 6vv_x + 3w_x)D^{-1} + 4u_xD^{-1}u,$$

$$\mathcal{R}_1^0 = -2(5vD^2 + 4v_xD + v_{2x} + 8uv + 4u_xD^{-1}v),$$

$$\mathcal{R}_2^0 = 5D^2 + 8u + 4u_xD^{-1},$$

$$\mathcal{R}_0^1 = -5u_xD - 6v_{2x} - 2(v_{3x} + 6v_xu)D^{-1} + 4v_xD^{-1}u,$$

$$\mathcal{R}_1^1 = -D^4 - 8uD^2 - 4u_xD + 8(8w - 2v^2) - 8v_xD^{-1}v - 8D^{-1}w_x,$$

$$\mathcal{R}_2^1 = 4(v_xD^{-1} + 2D^{-1}v_x),$$

$$\mathcal{R}_0^2 = -5w_xD - 6w_{2x} - 2(v_{3x} + 6w_xu)D^{-1} + 4w_xD^{-1}u.$$

$$\mathcal{R}_1^2 = -16vD^{-1}w_x - 8w_xD^{-1}v,$$

$$\mathcal{R}_2^2 = -D^4 - 8uD^2 - 4u_xD + 16(w - v^2) + 4w_xD^{-1} + 16vD^{-1}v_x. \tag{B17}$$

APPENDIX C: EXAMPLES TO SEC. III

1. Non-Abelian Schrödinger equation

This is the system given by

$$u_t = -\tfrac{1}{2}u_{xx} + uvu, \qquad v_t = \tfrac{1}{2}v_{xx} + vuv, \tag{C1}$$

where u and v belong to $\mathcal{K}$ (see Example 6 for the notations). The Lax operator of (C1) is given by

$$L = D + \begin{pmatrix} \mathbf{1} & 0 \\ 0 & -\mathbf{1} \end{pmatrix}\lambda + \begin{pmatrix} 0 & u \\ v & 0 \end{pmatrix}. \tag{C2}$$

The corresponding formula (54) reduces to

$$\begin{pmatrix} 0 & u_{n+1} \\ v_{n+1} & 0 \end{pmatrix} = \lambda \begin{pmatrix} 0 & u_n \\ v_n & 0 \end{pmatrix} + [R_n, L], \tag{C3}$$

where

$$R_n = \begin{pmatrix} a_n & b_n \\ c_n & -a_n \end{pmatrix}. \tag{C4}$$

The formula (C3) gives us both a_n, b_n, c_n and the recursion operator $\mathcal{R}$. They are given by

$$a_n = \tfrac{1}{2}D^{-1}(u_nv + uv_n), \quad b_n = \tfrac{1}{2}u_n, \quad c_n = -\tfrac{1}{2}u_n, \tag{C5}$$

$$\mathcal{R} = \frac{1}{2}\begin{pmatrix} -D + R_uD^{-1}R_v + L_uD^{-1}L_v & R_uD^{-1}L_u + L_uD^{-1}R_u \\ -L_vD^{-1}R_v - R_vD^{-1}L_v & D - R_vD^{-1}R_u - L_vD^{-1}L_u \end{pmatrix}. \tag{C6}$$

2. Non-Abelian modified KdV equation

The standard non-Abelian modified KdV equation is given by

$$u_t=\tfrac{1}{4}u_{xxx}-\tfrac{3}{4}u_xu^2-\tfrac{3}{4}u^2u_x. \tag{C7}$$

The Lax representation of this equation is given

$$L=D+\begin{pmatrix}0 & \mathbf{1}\\ \mathbf{1} & 0\end{pmatrix}\lambda+\begin{pmatrix}u & 0\\ 0 & -u\end{pmatrix}. \tag{C8}$$

The recursion operator $\mathcal{R}$ can be found from (60) and (61). In our case the automorphism σ is the same as in Example 6, and formulas (60) and (61) give us

$$\begin{pmatrix}0 & u_{n+1}\\ v_{n+1} & 0\end{pmatrix}=\lambda^2\begin{pmatrix}0 & u_n\\ v_n & 0\end{pmatrix}+[R_n,L], \tag{C9}$$

where

$$R_n=\begin{pmatrix}0 & a_n\\ b_n & 0\end{pmatrix}\lambda+\begin{pmatrix}c_n & 0\\ 0 & d_n\end{pmatrix}. \tag{C10}$$

Using (C9) we find a_n,b_n,c_n,d_n from the following:

$$b_n-a_n=u_n,\quad -a_{n,x}-a_nu-ua_n+c_n-d_n=0,$$

$$-b_{n,x}+b_nu+ub_n+d_n-c_n=0,\quad d_{n,x}+c_{n,x}=[c_n-d_n,u],$$

$$u_{n+1}=d_{n,x}+[d_n,u].$$

The resulting recursion operator is given by

$$\mathcal{R}=\tfrac{1}{4}(D-\mathrm{ad}_u\cdot D^{-1}\cdot \mathrm{ad}_u)(D-(L_u+R_u)D^{-1}(L_u+R_u)). \tag{C11}$$

[1] P. J. Oliver, *Applications of Lie Groups to Differential Equations*, Graduate Texts in Mathematics, 2nd ed. (Springer-Verlag, New York, 1993), Vol. 107.
[2] P. J. Olver, J. Math. Phys. **18**, 1212 (1977).
[3] M. Gürses and A. Karasu, J. Math. Phys. **36**, 3485 (1995).
[4] M. Gürses and A. Karasu, Phys. Lett. A **214**, 21 (1996).
[5] M. Gürses and A. Karasu, J. Math. Phys. **39**, 2103 (1998).
[6] M. Gürses and A. Karasu, Phys. Lett. A **251**, 247 (1999).
[7] S. I. Svinolupov, Theor. Math. Phys. **87**, 391 (1991).
[8] J. Krasil'shchik, Contemp. Math. **219**, 121 (1998).
[9] A. S. Fokas and R. L. Anderson, J. Math. Phys. **23**, 1066 (1982).
[10] A. S. Fokas, Stud. Appl. Math. **77**, 253 (1987).
[11] A. P. Fordy and J. Gibbons, J. Math. Phys. **22**, 1170 (1980).
[12] P. M. Santini and A. S. Fokas, Commun. Math. Phys. **115**, 375 (1988).
[13] A. S. Fokas and P. M. Santini, Commun. Math. Phys. **116**, 449 (1988).
[14] W. Symes, J. Math. Phys. **20**, 721 (1979).
[15] M. Adler, Invent. Math. **50**, 219 (1979).
[16] I. Ya. Dorfman and A. S. Fokas, J. Math. Phys. **33**, 2504 (1992).
[17] A. S. Fokas and I. M. Gel'fand, in *Important Developments in Soliton Theory*, Springer Series in Nonlinear Dynamics, edited by A. S. Fokas and V. E. Zakharov (Springer-Verlag, Berlin, 1993), pp. 259–282.
[18] M. Antonowicz and A. P. Fordy, in *Nonlinear Evolution Equations and Dynamical Systems (NEEDS'87)*, edited by J. Leon (World Scientific, Singapore, 1988), pp. 145–160.
[19] M. Antonowicz and A. P. Fordy, in *Soliton Theory: A Survey of Results*, edited by A. P. Fordy (Manchester University Press, Manchester, England, 1990). See also the related references therein.
[20] I. M. Gel'fand and L. A. Dikii, Funct. Anal. Appl. **10**, 13 (1976).
[21] F. Magri, J. Math. Phys. **19**, 1156 (1978).
[22] V. V. Sokolov, Sov. Math. Dokl. **30**, 44 (1984).

[23] Here we note that in 1980 that Shabat and Sokolov independently found the recursion operator for the Sawada–Kotera equation. This result was published in Ref. 24. In Ref. 22, Sokolov found the recursion operator for the Krichever–Novikov equation.
[24] N. H. Ibragimov, *Transformation Groups Applied to Mathematical Physics* (Reidel, Boston, 1985).
[25] V. G. Drinfeld and V. V. Sokolov, J. Sov. Math. **30**, 1975 (1985).
[26] V. G. Drinfeld and V. V. Sokolov, Proc. Sobolev Sem. Novosibirsk **2**, 5 (1981) (in Russian).
[27] I. Z. Golubchik and V. V. Sokolov, Theor. Math. Phys. **112**, 1097 (1997).
[28] K. Sawada and T. Kotera, Prog. Theor. Phys. **51**, 1355 (1974).
[29] J. Springael, X. B. Hu, and I. Loris, J. Phys. Soc. Jpn. **65**, 1222 (1996).
[30] J. Satsuma and R. Hirota, J. Phys. Soc. Jpn. **51**, 3390 (1982).
[31] G. Wilson, Phys. Lett. A **89**, 332 (1982).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Formal variable separation approach for nonintegrable models

Sen-yue Lou
Applied Physics Department of Shanghai Jiao Tong University, Shanghai 200030, People's Republic of China and Institute of Mathematical Physics, Ningbo University, Ningbo 315211, People's Republic of China

Li-Li Chen
Institute of Mathematical Physics, Ningbo University, Ningbo 315211, People's Republic of China

(Received 20 January 1999; accepted for publication 14 June 1999)

Using a formal variable separation approach, a nonlinear partial differential equation can be solved by solving ordinary different equations or even algebraic equations. Taking the KdV–Burgers and modified KdV–Burgers equations with background interaction as simple examples, some explicit solitary wave solutions which are induced by background source and nonlinearity or dispersion are obtained.
 [S0022-2488(99)00411-9]

I. INTRODUCTION

As the linear physics has been developed quite well and the real natural phenomena are nonlinear, the main interest of many scientists has been focused on the nonlinear science. To solve some types of special nonlinear problems, the so-called completely integrable models, many powerful methods like the inverse scattering transformation (IST),[1] Bäcklund and Dabourx transformations,[2] symmetry reduction,[3] bilinear approach,[4] standard[5] and extended[6] Painlevé analysis, etc. have been developed by various authors. In comparison with the linear case, it is known that IST is an extension of the Fourier transformation in the nonlinear case. In addition to the Fourier transformation, there is another powerful tool called the variable separation method in the linear case. However, there is little progress on obtaining some special solutions by means of a corresponding variable separation method in the nonlinear case. In Ref. 7, one of the present authors (Lou) and Lu have obtained some special solutions for the Davey–Stewartson equation by using a special variable separation procedure. A more systematic ''variable separation'' approach has been established by means of the symmetry constraints.[8–10] In that approach, although the independent variables of a reduced field have not totally been separated the field satisfies some lower-dimensional equations. However this type of procedure is used only for integrable models which possess Lax pairs. In this paper we try to extended the method to any models no matter they possess Lax pairs or not.

In the next section, we review the symmetry constraint approach of the KdV equation simply. The extended method is proposed to solve any nonlinear equation in Sec. III. Section IV is used to apply the method to a special nonintegrable model, KdV–Burgers (KdVB) and modified KdV–Burgers (MKdVB) equation. The last section is a short summary and discussion.

II. SYMMETRY CONSTRAINT APPROACH OF THE KDV EQUATION

It is known that the KdV equation

$$u_t-6uu_x-u_{xxx}=0 \tag{1}$$

possesses a Lax pair

$$\psi_{xx}+u\psi=\lambda\psi, \tag{2}$$

$$\psi_t = 2(u-\lambda)\psi_x - u_x\psi. \tag{3}$$

Equivalently, setting $\psi_x = \phi$, (2) and (3) can be written as

$$\begin{pmatrix} \psi \\ \phi \end{pmatrix}_x = \begin{pmatrix} 0 & 1 \\ \lambda - u & 0 \end{pmatrix} \begin{pmatrix} \psi \\ \phi \end{pmatrix}, \tag{4}$$

$$\begin{pmatrix} \psi \\ \phi \end{pmatrix}_t = \begin{pmatrix} -u_x & 2(u-\lambda) \\ -u_{xx} - 2(u-\lambda)^2 & u_x \end{pmatrix} \begin{pmatrix} \psi \\ \phi \end{pmatrix}. \tag{5}$$

A symmetry of the KdV equation (1) is defined as a solution of its linearized equation

$$\sigma_t - 6\sigma u_x - 6u\sigma_x - \sigma_{xxx} = 0, \tag{6}$$

that means (1) is form invariant under the transformation

$$u \to u + \epsilon\sigma, \tag{7}$$

where ϵ being a infinitesimal parameter.

It is known that $(\psi^2)_x$ and u_x are two typical symmetries of the KdV equation.[11] If we substitute the symmetry constraint condition

$$u_x - (\psi^2)_x = 0, \quad \text{i.e.,} \quad u = \psi^2 \tag{8}$$

into the Lax pair (4) and (5), we have

$$\begin{pmatrix} \psi \\ \phi \end{pmatrix}_x = \begin{pmatrix} \phi \\ \lambda\psi - \psi^3 \end{pmatrix} \equiv K_1, \tag{9}$$

$$\begin{pmatrix} \psi \\ \phi \end{pmatrix}_t = \begin{pmatrix} -2\lambda\phi \\ -2\lambda(\lambda\psi - \psi^3) \end{pmatrix} \equiv K_2. \tag{10}$$

Now, if we can get a solution of the lower dimensional Eqs. (9) and (10) then we can get a corresponding solution of the KdV equation at the same time from the symmetry constraint relation (8). It is clear that only one independent variable appears explicitly in Eqs. (9) and (10) though the dependent variables ψ and ϕ are the functions of the both independent variables x and t. Because ψ and ϕ are functions of two variables and only one independent variable appears in Eqs. (9) and (10), we may call this type of symmetry constraints as the ''formal variable separation approach.''

III. GENERAL FORMAL VARIABLE SEPARATION APPROACH

It is known that the procedure shown in the last section dependent on the existence of the Lax pair. In other words, the method is valid for completely integrable models. However, most of real physical models are not completely integrable. Can we get a hint to established a more general variable separation method to solve general nonlinear problems? Fortunately, the answer is positive!

In order to get some special solutions of a general $(n+1)$-dimensional N order PDE,

$$F(t, x_1, x_2, \ldots, x_n, u, u_{x_i}, u_{x_i x_j}, \ldots, u_{x_{i_1} x_{i_j} \cdots x_{i_N}}) \equiv F(u) = 0, \tag{11}$$

we may introduce a set of formally variable separated equations at first

$$\Psi_{x_i} = K_i, \quad i = 0,1,2,\ldots, \quad n, \quad x_0 \equiv t, \tag{12}$$

where $\Psi\equiv(\psi_1,\psi_2,\ldots,\psi_M)^T$ and $K_i\equiv K_i(\Psi)$ are M component matrix functions. The matrix functions K_i should be commuted to each other

$$[K_i,K_j]\equiv K_i'K_j-K_j'K_i\equiv\frac{\partial}{\partial\epsilon}(K_i(\Psi+\epsilon K_j)-K_j(\Psi+\epsilon K_i))\bigg|_{\epsilon=0} \tag{13}$$

because the compatibility conditions $\Psi_{x_ix_j}=\Psi_{x_jx_i}$. Finally one can suppose that the solution of (11) is related to Ψ by

$$u=U(\Psi) \tag{14}$$

and determine the functions U and K_i by substituting (12) and (14) into (11).

In the simple example given in the last section, $M=2$, $\Psi=(\psi,\phi)^T$, the compatibility condition is satisfied naturally because $K_2\sim K_1$ and the function U is given by the symmetry constraint condition (8) simply. The simplicity in the KdV equation is because of its integrability. In general case, it is insignificant to find some suitable K_i and U to give out some special exact solutions especially in nonintegrable case. For concreteness, we turn to treat a concrete example.

IV. EXPLICIT SOLUTIONS OF THE KdVB AND MKdVB EQUATION WITH BACKGROUND SOURCE

Usually, one treats the nonlinear equations describing nonlinear problems without background interactions for simplicity. Some authors have considered the case in which the background interaction is periodic[12] and nonperiodical.[13] The general KdVB and MKdVB with source have the form

$$u_t+\sum_{p=1}^{P}a_pu^pu_x+cu_{xxx}+\gamma u_{xx}=\sum_{q=0}^{Q}b_qu^q, \tag{15}$$

where the right-hand side of (15) is the source term and the summation term in the left-hand side is the usual nonlinear interaction of the MKdVB $(P>1)$ and the KdVB $(P=1)$ equations. The solitary wave solutions of three special cases, $\{\gamma=0,a_1=0,P=2,Q=3\}$, $\{P=1,Q=0\}$ and $\{c=0,P=1,Q=3\}$ have been given by Lan and Wang.[13]

Now, we treat the general MKdVB equation (15) by using the formal variable separation approach proposed in the last section. In this paper, we take only $M=1$ in (12), i.e., the formally variable separation equations have the forms

$$\Psi_x=K_1(\Psi),\quad \Psi_t=K_2(\Psi), \tag{16}$$

where $\Psi=\Psi(x,t)$ is a scalar function of $\{x,t\}$ and $K_i(\Psi)=K_i$, $i=1,2$ are functions of Ψ with the compatibility condition

$$[K_1,K_2]=0. \tag{17}$$

In the single component case, $K_2=\omega K_1$ with constant ω is an only possible solution of (17). Substituting (14) and (16) into (15) we know that (14) solves Eq. (15) if two functions K_1 and U satisfy the following ordinary differential equation:

$$\omega K_1U'+K_1U'\sum_{p=1}^{P}a_pU^p+cK_1(U'''K_1^2+3K_1K_1'U''$$
$$+K_1K_1''U'+K_1'^2U')+\gamma K_1(K_1U''+K_1'U')=\sum_{q=0}^{Q}b_qU^q. \tag{18}$$

In real application, for any fixed K_1, we can get a corresponding solution of (15) by solving (18). For instance, we can take K_1 as a polynomial function of Ψ,

$$K_1=\sum_{j=0}^{J} A_j\Psi^j \tag{19}$$

to get solitary wave solutions of (15). Equation (18) with (19) can always be solved by means of power series approach by setting

$$U=\sum_{j=0}^{\infty} U_j\Psi^j \tag{20}$$

and fixing the expansion coefficients from (18) recursively. If we take $J=1$ and $J=2$ in (19), (20) is equivalent to the usual exponential and tanh expansions used by many authors. In some special selections of a_p and b_q in (15), (20) may be truncated, then some closed explicit solitary wave solutions

$$U=\sum_{r=0}^{R} U_r\Psi^r \tag{21}$$

can be obtained. In the usual KdV and MKdV systems, the solitons are caused by the balance between the effects of the dispersion and the nonlinear interactions. Once the background interactions exist, the solitary waves may be caused in four ways, (i) by the balance between the effects of the dispersion $(c\neq 0)$ and the nonlinear interactions $(a_p\neq 0)$, (ii) by the balance between the effects of the dispersion and background interaction $(b_q\neq 0)$, (iii) by the balance between the effects of the background interactions and the nonlinear interactions, and (iv) by the balance of all three effects. Now we discuss the truncation conditions caused by balance among the linear dispersion, nonlinear interactions and background interactions.

Substituting (21) and (16) with (19) into (18) we know that the highest order terms of Ψ in the background, dispersion, and the nonlinear interactions read $b_QU_Q^R\Psi^{QR}$, $cA_J^3U_RR((R-1)(R-2)+3J(R-1)+J(2J-1))\Psi^{R+3J-3}$, and $RA_Ja_PU_R^{P+1}\Psi^{(P+1)R+J-1}$, respectively. From the above leading order terms, we know that (i) if a solitary wave is caused by the effects of the dispersion and the nonlinear interactions, the balance conditions are

$$\begin{gathered}(P+1)R+J-1=R+3J-3>QR,\\ RA_Ja_PU_R^{P+1}=-cA_J^3U_RR(((R-2)+3J)(R-1)+J(2J-1));\end{gathered} \tag{22}$$

(ii) if a solitary wave is caused by the effects of the dispersion and the background interactions, the balance conditions read

$$(P+1)R+J-1<R+3J-3=QR,\quad b_QU_Q^R=cA_J^3U_RR((R-1)((R-2)+3J)+J(2J-1)); \tag{23}$$

(iii) if a solitary wave is caused by the effects of the nonlinear interactions and the background interactions, the balance conditions become

$$(P+1)R+J-1=QR>R+3J-3,\quad b_QU_Q^R=RA_Ja_PU_R^{P+1} \tag{24}$$

and (iv) if a solitary wave is caused by all three effects, the balance conditions have the forms

$$\begin{gathered}(P+1)R+J-1=R+3J-3=QR,\\ b_QU_Q^R=cA_J^3U_RR((R-1)(R-2)+3J(R-1)+J(2J-1))+RA_Ja_PU_R^{P+1}.\end{gathered} \tag{25}$$

Finally, we list some special types of solitary wave solutions of (15) for $J=2$ in cases (22)–(25). The general solution of (16) with (19) and $J=2$ reads

$$\Psi=\frac{\psi_1\exp((\psi_2-\psi_1)A_2(x+\omega t+x_0))+\psi_2}{1+\exp((\psi_2-\psi_1)A_2(x+\omega t+x_0))}, \tag{26}$$

where x_0 is an arbitrary constant and

$$\psi_1=\frac{1}{2A_2}(\sqrt{A_1^2-4A_0A_2}-A_1),\quad \psi_2=\frac{1}{2A_2}(-\sqrt{A_1^2-4A_0A_2}-A_1). \tag{27}$$

A. Solitary waves caused by the balance between the dispersion and the nonlinear interactions

Substituting $J=2$ into (22), we know that the solitary wave solution with the form (21) and (26) exist for $(P+1)R+1=R+3>QR$. The simplest two cases read

$$\{R=1,P=2,Q\leqslant 3\},\ \text{and}\ \{R=2,P=1,Q\leqslant 2\}. \tag{28}$$

For the first subcase of (28), the solitary wave solution can be obtained by substituting (19) and (21) into (18). The final result shows us that if the model parameters satisfy a condition

$$\begin{aligned}&54b_3^5c^3-54c^2(a_2\gamma-3ca_1V)b_3^4+(18a_2^2c\gamma^2-27c^2(a_1^2+4\gamma a_1V+4Vcb_2)a_2)b_3^3\\&+((-9c^2b_1-2\gamma^3)a_2^3+9c(2\gamma V+a_1)(a_1\gamma+4cb_2)a_2^2-9a_2c^2a_1^3V)b_3^2\\&+((3cb_1\gamma-9c^2Vb_0)a_2^4-3c(3cVb_1a_1+4Vb_2\gamma^2+3cb_2^2+4b_2\gamma a_1)a_2^3\\&+18a_2^2c^2Vb_2a_1^2)b_3+3cb_2(b_2\gamma+3b_1cV)a_2^4-9c^2a_2^3Vb_2^2a_1=0,\end{aligned} \tag{29}$$

then (15) possesses a kink type solitary wave solution

$$u=u_0+\Psi,\quad u_0=-\frac{3b_3c-3A_1a_2c-\gamma a_2+3a_1Vc}{6ca_1V} \tag{30}$$

with

$$\omega=\frac{1}{A_2}\left(\frac{1}{3}A_0a_2^2+2a_2u_0A_1-a_1A_1-3b_3u_0+b_2-(a_1u_0+7cA_1^2+3\gamma A_1+a_2u_0^2)A_2\right), \tag{31}$$

$$A_0=\frac{V(b_0+b_3u_0^3+b_1u_0+b_2u_0^2)}{a_1A_1+2a_2u_0A_1+2\gamma VA_1+3b_3u_0+b_2+6cVA_1^2},\quad A_2=\pm\sqrt{\frac{-a_2}{6c}}\equiv V, \tag{32}$$

and A_1 being an arbitrary constant.

In the second case, the solitary waves have the form

$$u=u_0+u_1\Psi+u_2\Psi^2 \tag{33}$$

with Ψ being given by (26) while the parameters can be taken in four different ways:
(A)

$$\gamma=12b_2ca_1^{-1}, \tag{34}$$

$$u_0=\frac{b_1^2-4b_0b_2}{192b_2^4c}a_1^3-\frac{b_1}{2b_2}-\frac{3A_1c}{a_1^2}(a_1A_1+4b_2), \tag{35}$$

$$A_2=-\frac{1}{100a_1^2c^2A_0(\gamma a_1+18b_2c)}((25(5b_1-\gamma A_1^2+10b_2u_0)c^2+\gamma^3)a_1^3$$
$$+4b_2c(75cA_1(A_1c+\gamma)-4\gamma^2)a_1^2-4c^2b_2^2(-13\gamma+150A_1c)a_1-48b_2^3c^3), \quad (36)$$

$$\omega=-\left(u_0a_1+\left((8A_2A_0+A_1^2)c+\frac{6}{5}A_1\gamma-\frac{\gamma^2}{25c}\right)-\frac{2b_2}{25a_1}(30A_1c+23\gamma)+\frac{96}{25a_1^2}b_2^2c\right), \quad (37)$$

$$u_2=-\frac{12cA_2^2}{a_1},\ \ u_1=-\frac{12A_2}{5a_1^2}(-2b_2c+5cA_1a_1+\gamma a_1), \quad (38)$$

where A_1 is an arbitrary constant;

(B) In the second subcase, all the parameters can still expressed by (35)–(38) while (34) should be changed as

$$\gamma=2b_2ca_1^{-1}; \quad (39)$$

(C) If conditions (34) and (35) are replaced by

$$625c^2(-b_1^2+4b_0b_2)a_1^6+144b_2^2(\gamma a_1-2b_2c)^4=0 \quad (40)$$

and

$$u_0=-\frac{1}{50b_2ca_1^3}(6b_2(-(\gamma a_1-2b_2c)^2+10a_1c(\gamma a_1-2b_2c)A_1+25c^2A_1^2a_1^2)+25b_1ca_1^3), \quad (41)$$

we get the third type of solitary wave solution of (15) with the form (33);

(D) In the first three subcases, A_1 is a free parameter and there is a restriction condition on the model parameters. For the forth type of solitary wave solution of the form (33), there is no constraint on the model parameters if A_1 is fixed by

$$15625c^4(-b_1^2+4b_0b_2)a_1^6+(-25a_1^2c^2(\gamma a_1-12b_2c)A_1^2$$
$$+300a_1b_2c^2(\gamma a_1-2b_2c)A_1+(\gamma a_1-12b_2c)(\gamma a_1-2b_2c)^2)^2=0 \quad (42)$$

and u_0 is given by

$$u_0=-\frac{1}{250b_2c^2a_1^3}(-25a_1^2c^2(\gamma a_1-12b_2c)A_1^2+300a_1b_2c^2(\gamma a_1-2b_2c)A_1$$
$$+125b_1c^2a_1^3-48b_2^3c^3+\gamma^3a_1^3+52b_2^2c^2\gamma a_1-16b_2\gamma^2a_1^2c), \quad (43)$$

while the forms of (36)–(38) are remained to have the same forms though the parameters u_0 and A_1 in (36)–(38) should be replaced by (42) and (43).

B. Solitary wave solution caused by the balance between the dispersion and the background interactions

When $J=2$, the first condition of (23) becomes $(P+1)R+1<R+3=QR$. The simplest case is

$$R=1, \quad P=1, \quad Q=4. \quad (44)$$

In this case, if the model parameters are restricted by

$$2cVV_0^2+(\omega_0+cV_1^2+\gamma V_1)V_0+b_0=0, \quad (45)$$

where

$$V\equiv-\left(\frac{b_4}{6c}\right)^{1/3},\quad V_1\equiv-\frac{1}{12cV^2}(b_3+a_1V+2\gamma V^2), \tag{46}$$

$$V_0\equiv\frac{-b_1V+2\gamma V_1^2V+6cV_1^3V+a_1V_1^2+V_1b_2}{V(a_1+2\gamma V)}, \tag{47}$$

and

$$\omega_0\equiv-\frac{1}{V}(a_1V_1+3\gamma V_1V+8cV^2V_0+b_2+7cVV_1^2), \tag{48}$$

the form of the solitary wave solution is given by

$$u=u_0+u_1\Psi \tag{49}$$

with (26) and u_0 and u_1 being arbitrary while

$$A_2=Vu_1,\quad A_1=-\frac{1}{12cV^2}(b_3+a_1V+2\gamma V^2+4u_0b_4), \tag{50}$$

$$A_0=\frac{1}{u_1V(a_1+2\gamma V)}(-b_1V+2\gamma V_1^2V+6cV_1^3V+a_1V_1^2+V_1b_2-4b_4u_0^3V+(6a_1b_4-3b_3V)u_0^2+(3a_1b_3-2b_2V)u_0), \tag{51}$$

$$\omega=-\frac{1}{V}(a_1V_1+3\gamma V_1V+8cV^2V_0+b_2+7cVV_1^2+6b_4u_0^2+(3b_3+a_1V)u_0). \tag{52}$$

C. Solitary wave solution caused by the balance between the nonlinear interactions and the background interactions

The first condition of (24) reduces to $(P+1)R+1=QR>R+3$, for $J=2$. The simplest one case is

$$R=1,\quad P=3,\quad Q=5. \tag{53}$$

The corresponding solitary wave solution has the form of (49) also. Substituting (49) and (53) into (18), one can find that solitary wave of the form (49) with (26) exist for the following two restrictions among the model parameters that are satisfied:

$$b_0a_3-2cb_5C_0^2+\omega_0C_0a_3+cC_1^2C_0a_3+\gamma C_1C_0a_3=0, \tag{54}$$

$$-2\gamma b_5C_0+A_1C_0a_3+b_1a_3+\omega_0C_1a_3+\gamma C_1^2a_3-8cb_5C_0C_1+cC_1^3a_3=0, \tag{55}$$

where

$$C_1=\frac{1}{a_3^4}(6cb_5^3-b_4a_3^3+a_2b_5a_3^2), \tag{56}$$

$$C_0=\frac{1}{a^3}(-a_2C_1a_3^2-2\gamma b_5^2-b_3a_3^2+a_1b_5a_3-12cb_5^2C_1), \tag{57}$$

$$\omega_0=\frac{1}{a_3b_5}(a_1C_1a_3^2-3\gamma C_1b_5a_3-7cb_5C_1^2a_3+a_2C_0a_3^2+b_2a_3^2+8cb_5^2C_0). \tag{58}$$

The constants u_0 and u_1 are remains free again when A_2, A_1, A_0 and ω in (26) is given by

$$A_2=-b_5u_1/a_3, \tag{59}$$

$$A_1=(6cb_5^3-2u_0b_5a_3^3-b_4a_3^3+a_2b_5a_3^2)/a_3^4, \tag{60}$$

$$A_0=\frac{1}{a_3^3u_1}(b_5(a_1+2a_2u_0)a_3-3a_3^3u_0A_1-(a_2A_1+b_3+4b_4u_0+7b_5u_0^2)a_3^2-b_5^2(2\gamma+12cA_1)), \tag{61}$$

$$\omega=\frac{1}{a_2b_5}(3u_0(u_0A_1+u_1A_0)a_3^3+(6b_4u_0^2+3b_3u_0+2a_2u_0A_1+b_2+9b_5u_0^3+a_2u_1A_0+a_1A_1)a_3^2$$
$$+(-a_1u_0-3\gamma A_1-a_2u_0^2-7cA_1^2)b_5a_3+8cu_1b_5^2A_0). \tag{62}$$

D. Solitary wave caused by the effects of the nonlinear interactions, dispersion, and the background interactions

In $J=2$ case, the first condition of (25) is simplified to $(P+1)R+1=QR=R+3$ with the only positive integer solution

$$R=1,\quad P=2,\quad Q=4. \tag{63}$$

If the restriction condition

$$2cVC_0^2+(cC_1^2+\omega_0+\gamma C_1)C_0+b_0=0, \tag{64}$$

with

$$C_1=-\frac{b_3+2\gamma V^2+a_1V}{a_2+12cV^2}, \tag{65}$$

$$C_0=\frac{2\gamma C_1^2V-b_1V+6cC_1^3V+C_1b_2+a_1C_1^2}{2\gamma V^2+a_1V-a_2C_1}, \tag{66}$$

$$\omega_0=-\frac{1}{V}(a_1C_1+7cVC_1^2+b_2+3\gamma C_1V+a_2u_1C_0+8cu_1V^2C_0), \tag{67}$$

where V is a root of

$$6cV^3+a_2V+b_4=0, \tag{68}$$

we can obtain the solitary wave solution which are caused by all three effects in the form with two arbitrary constants u_0 and u_1 while A_2, A_1, A_0 and ω in (26) is expressed as

$$A_2=Vu_1, \tag{69}$$

$$A_1=-\frac{4b_4u_0+2a_2u_0V+b_3+2\gamma V^2+a_1V}{a_2+12cV^2}, \tag{70}$$

$$A_0=\frac{1}{u_1(2\gamma V^2+2a_2u_0V+a_1V-a_2A_1)}(6cA_1^3V+(2\gamma V+2A_2u_0+a_1)A_1^2$$
$$+(b_2+6b_4u_0^2+3b_3u_0)A_1-4b_4u_0^3V-b_1V-2b_2u_0V-3b_3u_0^2V), \tag{71}$$

$$\omega=-((a_2V+6b_4)u_0^2+(3b_3+a_1V+2a_2A_1)u_0$$
$$+(a_2+8cV^2)u_1A_0+7cVA_1^2+(a_1+3\gamma V)A_1+b_2)/V. \tag{72}$$

V. SUMMARY AND DISCUSSION

In linear physics, both the Fourier transformation approach and the variable separation method play a very important role. However, in nonlinear physics, both the traditional Fourier transformation approach and the variable separation method are not valid again. For some special types of nonlinear equations, the so-called completely integrable models, the inverse scattering transformation approach is considered as the extension of the Fourier transformation and the symmetry constraint approach or the nonlinearization of the Lax pair can be considered as a type of extension of the variable separation approach. Nevertheless, most of the problems in nonlinear physics are not completely integrable. Many mathematicians and physicists have been trying to find some powerful methods to solve the nonlinear partial differential equations which are not completely integrable models, but there is little progress in this direction except for the symmetry reduction and the perturbation theory basing on the inverse scattering transformation.

In this paper, we have extended the formal variable separation approach to a general form such that the method can be used to solve both the integrable and nonintegrable models. Using the method to a general modified KdV equation with background interactions, we know that once a solution of a formally variable separated system is obtained, a corresponding solution of the modified KdVB equation can be obtained by solving an ordinary differential equation. When we restrict the functions that appeared in the variable formally separated system as polynomial functions, the ordinary differential equation can be solved using the series method. In some special cases where some restricted conditions on the model parameters are satisfied, the series solution becomes a closed truncated form.

Once the background interactions exist for the KdVB and modified KdVB system, the solitary wave may be formed in some different ways. Usually, the soliton solution is caused by the balance between nonlinear interactions and dispersion. From our calculations, we know that the solitary wave solutions can be formed by the balance of any two of three effects: dispersion, nonlinear interactions, and background interactions. Some special explicit solitary wave solutions in four possible cases are listed in Sec. IV.

In this paper, we treat only the formally variable separation system (16) as one component for the KdVB MKdVB equation. In one component case for (16), the only allowed solution has the traveling wave form. To get nontraveling wave solutions, one has to treat the method with more components. More about the formally variable separation approach is worthy of further study.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China and the Natural Science Foundation of Zhejiang Province of China.

[1] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations, and Inverse Scattering, London Mathematical Society Lecture Note series 149* (Cambridge University Press, Cambridge, 1991).
[2] C. H. Gu, Lett. Math. Phys. **26**, 199 (1992).
[3] P. J. Olver, *Application of Lie Group to Differential Equations* (Springer, New York, 1986); P. A. Clarkson and M. D. Kruskal, J. Math. Phys. **30**, 2201 (1989); S-y Lou, H-y Ruan, D-f Chen, and W-z Chen, J. Phys. A **24**, 1455 (1991).
[4] R. Hirota, Phys. Rev. Lett. **27**, 1192 (1971).
[5] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[6] A. Pickering, J. Phys. A **26**, 4395 (1993); Lou S-y, Z. Naturforsch. **53a**, 251 (1998).
[7] S-y Lou and J-z Lu, J. Phys. A **29**, 4209 (1996).

[8] C-w Cao, Sci. China, Ser. A **33**, 528 (1990).
[9] Y-b Zeng, Phys. Lett. A **169**, 541 (1991); Y. Chen and Y-s Li, *ibid.* **157**, 22 (1991).
[10] B. Konopelchenko, J. Sidorenko, and W. Strampp, Phys. Lett. A **157**, 17 (1991).
[11] S-y Lou, J. Math. Phys. **35**, 2390 (1994).
[12] S. Takeno and S. Homma, J. Phys. Soc. Jpn. **55**, 2549 (1986); K. Konno, W. Komeyama, and Sanuki, *ibid.* **37**, 171 (1974).
[13] H. Lan and K. Wang, Phys. Lett. A **137**, 369 (1989).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

KP constraints from reduced multi-component hierarchies

R. Willox
Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8914, Japan and Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

I. Loris
Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

(Received 22 February 1999; accepted for publication 30 June 1999)

The (m-vector) k-constrained Kadomtsev–Petviashvili (KP) hierarchy is shown to be a ''pseudo''-reduction of the $(m+1)$-component KP hierarchy. To facilitate the implementation of this reduction on the level of the solutions, the typical multi-component KP solutions are mapped onto solutions of a Toda molecule-type equation from which (Wronskian and Grammian) solutions for the constrained KP hierarchy follow. The reduction of the associated linear systems is discussed and its importance for the choice of bilinear representation of the reduced systems is explained. © *1999 American Institute of Physics.* [S0022-2488(99)01310-9]

I. INTRODUCTION

Judging by the considerable body of work which has appeared on the topic of constrained KP systems, one would expect their properties to be quite well understood by now. Among the most important facts discovered (apart from the original formulation of the ''constrained hierarchies,'' see Refs. 1–3 for a full account of these) there are the bilinear representation of the reduced systems,[4,5] the interpretation of the constraint procedure in terms of the basic KP construction (be it on a pseudo-differential, Grassmanian,[6] or purely bilinear level,[4,7]), and, of course, the description of various classes of solutions.[8–12] The motivation for this paper however, stems from our belief that many (if not most) of these features are best appreciated in the framework of multi-component KP hierarchies. As it will turn out, the problem of constructing solutions plays a central role therein.

At this early point, in order to render the main object we wish to study slightly more tangible—the m-vector, k-constrained KP hierarchy—we shall characterize it in terms of the familiar KP tau functions: a necessary condition for a KP tau function $\tau(\mathbf{x})$ [depending on a set of coordinates $\mathbf{x}=(x_1,x_2,x_3,\dots)$] to be a tau function for the m-vector, k-constrained KP hierarchy is that its x_k-derivative τ_{x_k} is a superposition of m (quite particular) KP tau functions: $\tau_{x_k}=\sum_{i=1}^m\hat{\tau}_i$, the actual description of such a ''constrained'' tau function $\tau(\mathbf{x})$ and the corresponding $\hat{\tau}_i$'s being the main problem to be solved. Note that as one may always consider all $\hat{\tau}_i\equiv 0$, the standard k-reductions of the KP hierarchy are included here as well.

In the past a variety of solution methods for the constrained KP evolutions have been proposed but, undoubtedly, the most important and fruitful method which has been employed is that of so-called Darboux–Bäcklund transformations.[13] Actually, it was recognized early on that this construction—which works so well in the KP case—should also be adaptable to the specific problem of constructing ''constrained'' solutions. Several applications of Darboux–Bäcklund transformations have been reported in the mean-time: ordinary Darboux transformations have been used to construct Wronskian-type determinant solutions whereas binary-type Darboux transformations have been shown to yield Grammian solutions for the reduced hierarchies.[14,10] Going through the literature on this topic one feature of these methods immediately ''jumps'' out at the reader: in some cases[10,11] authors are discussing solutions generated through Darboux–Bäcklund transformations which appear to be more general than those treated in other accounts. The difference almost always resides in the particular choice of the *vacuum* solution for the constrained equations which is used as a *seed* in the recursive application of the transformations. However, the problem which invariably arises is how to construct an appropriate vacuum. Any recursion formula or solution expressed in terms of a postulated vacuum remains rather ''empty'' without an actual construction of at least one example of such a ''general'' vacuum. In other words, the problem of obtaining solutions ultimately boils down to constructing vacuum solutions, or, put another way, to proving the existence or nonexistence of nontrivial vacuum solutions. Along the course of the present discussion we shall comment on this problem.

An interesting property of the solutions of the constrained hierarchies obtainable through consecutive Darboux–Bäcklund transformations is that (at least in the simplest case of the scalar constraints, i.e., for $m=1$) the transformation chain can be viewed as a Toda-like system.[11] In this paper we wish to invert and at the same time strengthen this relation by presenting a discussion of the m-vector k-constrained KP hierarchy as a reduction of the multi (2, 3,...)-component KP hierarchy,[15] in which multi-dimensional generalizations of the (molecule-type) Toda chains take up a prominent place.

First of all, we believe this to be a natural description of the constrained systems as for example their bilinear representations or the interpretation of the constrained tau functions arise quite effortlessly from reductions of the multi-component KP hierarchies. As we shall try to advocate, another advantage of the present point of view is the relative ease with which Wronskian and Grammian solutions for the reduced hierarchies can be obtained. As mentioned above, although such solutions have been discussed by a number of authors (including the present ones),[8–11,14,16] we have the feeling that their origins have never been quite clear altogether. All these solutions exhibit a remarkable ''bi-directional'' structure which, at least for ordinary vacuum, appears to bear no trace of the constraints imposed on the original KP solutions; a situation quite opposed to that of the standard k-reductions of the KP hierarchy. We will show that this structure is inherited from the solutions of (what is essentially) the Toda molecule system. Furthermore, we shall make a case for the importance of the constructability of appropriate vacuum solutions by showing that the present reduction procedure does not give rise to solutions which genuinely refer to nontrivial vacuum. Although the procedure allows for the description of what would be general Darboux–Bäcklund orbits, it becomes immediately apparent that there is no hope whatsoever of obtaining anything (apart from minor adjustments of previously stated results) more general than the solutions already presented in Ref. 14.

In the next paragraph we introduce the construction of the multi-component KP hierarchy itself. The language in which the hierarchy is formulated is that of free charged fermion operators, essentially following the treatment given in Refs. 17 and 18. In Sec. II, we discuss the reductions we shall impose on the multi-component KP hierarchies in order to obtain constrained KP evolutions. These reductions are of a ''nonstandard'' type and we would like to refer to them as ''pseudo-reductions'' to set them apart from regular dimensional reductions. Here we use the qualification ''pseudo'' in the original sense of Hirota,[19] meaning that in the set of coordinates introduced for the original hierarchy the reduction is not quite a dimensional one, as (generically) only the dependence on a single variable will be eliminated by it. The effect of these reductions then proves to be such that they impose a constraint on that part of the original hierarchy which consists of the KP evolutions. The present paper is, however, not intended to give a full account of the effect of such reductions on the multi-component KP hierarchies themselves (as that would lead us too far from our immediate goal), but rather to study their effect on the KP evolutions. In an earlier paper[20] we discussed the origin (similar to the above) of a particular generalized k-constraint for the KP hierarchy, lying within the 2-D Toda lattice hierarchy. There, obtaining solutions proved to be rather effortless as they arose immediately from those for the Toda lattice itself. Here we shall see that the structure of the solutions of the multi-component KP hierarchy differs considerably from that of the solutions we shall find eventually for the reduced systems. In order to establish a clear link we shall first have to recast them into a form which is particularly well suited for imposing the reduction. All this will be explained in the fourth part of this paper. Finally, for the scalar constraints several comments on the available bilinear descriptions for the constrained KP hierarchies can be made. Starting from the reduced multi-component linear system it will become obvious that every reduced hierarchy essentially allows for three different bilinear descriptions. Each description has its own underlying linear formulation, well adapted to that particular case, all of them being related by rather trivial gauge transformations.

II. MULTI-COMPONENT KP HIERARCHIES

We begin by describing the *n*-component KP hierarchy (by and large following notations introduced by the Kyoto school[17,21,22] rather than those found in Ref. 15) in terms of charged (free) fermion creation-annihilation operators $\psi_i^{(\alpha)},\psi_j^{*(\beta)}(i,j\in\mathbf{Z}+\frac{1}{2};\ \alpha,\beta=1,...,n)$ carrying ''charges'' $+1$ and -1, respectively. The operators satisfy the anticommutation relations:

$$[\psi_i^{(\alpha)},\psi_j^{*(\beta)}]_+=\delta_{\alpha,\beta}\delta_{i+j,0};\quad [\psi_i^{(\alpha)},\psi_j^{(\beta)}]_+=[\psi_i^{*(\alpha)},\psi_j^{*(\beta)}]_+=0. \tag{2.1}$$

For the (usual) Fock representation of this algebra (and for its dual representation), one introduces cyclic vectors $|\mathrm{vac}\rangle$ (and $\langle\mathrm{vac}|$):

$$\begin{aligned}\psi_i^{(\alpha)}|\mathrm{vac}\rangle=\psi_i^{*(\alpha)}|\mathrm{vac}\rangle=0 \quad \text{for } i>0,\\ \langle\mathrm{vac}|\psi_i^{(\alpha)}=\langle\mathrm{vac}|\psi_i^{*(\alpha)}=0 \quad \text{for } i<0.\end{aligned} \tag{2.2}$$

For our purposes it will prove worthwhile to use the (invertible) *shift* operators S_α,

$$\begin{aligned}S_\alpha\psi_i^{(\alpha)}=\psi_{i-1}^{(\alpha)}S_\alpha,\quad S_\alpha\psi_i^{*(\alpha)}=\psi_{i+1}^{*(\alpha)}S_\alpha,\\ S_\alpha\psi_i^{(\beta)}=-\psi_i^{(\beta)}S_\alpha,\quad S_\alpha\psi_i^{*(\beta)}=-\psi_i^{*(\beta)}S_\alpha,\qquad \text{for } \alpha\neq\beta,\end{aligned} \tag{2.3}$$

to represent the (dual) Fock states

$$\begin{aligned}|m_1,m_2,\ldots,m_n\rangle\equiv S_1^{m_1}S_2^{m_2}\cdots S_n^{m_n}|\mathrm{vac}\rangle,\\ \langle m_1,m_2,\ldots,m_n|\equiv\langle\mathrm{vac}|S_n^{-m_n}\cdots S_2^{-m_2}S_1^{-m_1}\end{aligned} \tag{2.4}$$

of total charge $\Sigma_{i=1}^n m_i$ but referring to particular ''colors'' m_i (the m_i being integers). These states are orthonormal with respect to the usual pairing $\langle\mathrm{vac}|1|\mathrm{vac}\rangle=1$:

$$\langle m_1,m_2,\ldots,m_n|m_1',m_2',\ldots,m_n'\rangle=\delta_{m_1,m_1'}\delta_{m_2,m_2'}\cdots\delta_{m_n,m_n'}. \tag{2.5}$$

We also introduce formal operators $\psi^{(\alpha)}(p)$ and $\psi^{*(\alpha)}(q)$ depending on some parameters p and q:

$$\psi^{(\alpha)}(p)\equiv\sum_{j\in\mathbf{Z}+1/2}\psi_j^{(\alpha)}p^{-j-1/2},\quad \psi^{*(\alpha)}(q)\equiv\sum_{j\in\mathbf{Z}+1/2}\psi_j^{*(\alpha)}q^{-j-1/2}, \tag{2.6}$$

for which ''time'' evolutions with respect to n different sets of infinitely many variables $\mathbf{x}^{(\alpha)}\equiv(x_1^{(\alpha)},x_2^{(\alpha)},\ldots,x_l^{(\alpha)},\ldots)$ $(\alpha=1,...n)$ can be defined in terms of the Hamiltonian:

$$H(\mathbf{x})\equiv\sum_{\alpha=1}^{n}\sum_{l=1}^{\infty}x_l^{(\alpha)}H_l^{(\alpha)}\quad\text{with}\quad H_l^{(\alpha)}\equiv\sum_{j\in Z+1/2}:\psi_{-j}^{(\alpha)}\psi_{j+l}^{*(\alpha)}: \tag{2.7}$$

($::$ denotes the usual normal ordering $:\psi_i^{(\alpha)}\psi_j^{*(\alpha)}:=\psi_i^{(\alpha)}\psi_j^{*(\alpha)}-\langle\mathrm{vac}|\psi_i^{(\alpha)}\psi_j^{*(\alpha)}|\mathrm{vac}\rangle$). Note also that $H_l^{(\alpha)}|\mathrm{vac}\rangle=0$, $\forall l\geq 1$ and accordingly that $\exp(H(\mathbf{x}))|\mathrm{vac}\rangle=1$. The evolutions for $\psi^{(\alpha)}(p)$ and $\psi^{*(\alpha)}(q)$wrt the coordinates $\mathbf{x}$ are then given by

$$e^{H(\mathbf{x})}\psi^{(\alpha)}(p)e^{-H(\mathbf{x})}=\psi^{(\alpha)}(p)e^{\xi_\alpha(x,p)},\quad e^{H(\mathbf{x})}\psi^{*(\alpha)}(q)e^{-H(\mathbf{x})}=\psi^{*(\alpha)}(q)e^{-\xi_\alpha(\mathbf{x},q)}, \tag{2.8}$$

where $\xi_\alpha(\mathbf{x},k)=\Sigma_{l=1}^{\infty}x_l^{(\alpha)}k^l$. Making use of the time evolutions $g(\mathbf{x})=e^{H(\mathbf{x})}ge^{-H(\mathbf{x})}$ of an element g of (a suitable completion of the $\mathrm{GL}(\infty)$ group—generated by elements

$$\sum_{\alpha,\beta}\sum_{i,j} c_{ij}^{\alpha\beta}\psi^{(\alpha)}(p_i)\psi^{*(\beta)}(q_j)+c \tag{2.9}$$

of (a completion of) the $\mathrm{gl}(\infty)$ algebra[15]—the n-component tau functions $\tau^{\mathbf{m}}(\mathbf{x})$ are defined as expectation values:

$$\tau^{\mathbf{m}}(\mathbf{x})\equiv\langle\mathbf{m}|g(\mathbf{x})|vac\rangle \tag{2.10}$$

for zero-charge states with color $\mathbf{m}=(m_1,m_2,\ldots,m_n)$:

$$\langle\mathbf{m}|=\langle\mathrm{vac}|S_n^{-m_n}\cdots S_2^{-m_2}S_1^{-m_1},\quad \sum_{i=1}^{n} m_i=0. \tag{2.11}$$

These tau functions satisfy the well-known bilinear identity,[15,17] $\forall\mathbf{x},\mathbf{x}'$:

$$\sum_{\alpha=1}^{n}\mathrm{Res}_\lambda[(-1)^{\Sigma_{i=1}^{\alpha-1}(k_i+k_2')}\lambda^{k_\alpha-k_\alpha'-2}e^{\xi_\alpha(x-x',\lambda)}\tau^{\mathbf{k}-\boldsymbol{\delta}_\alpha}(\mathbf{x}-\boldsymbol{\epsilon}_\alpha(\lambda))\tau^{\mathbf{k}'+\boldsymbol{\delta}_\alpha}(\mathbf{x}'+\boldsymbol{\epsilon}_\alpha(\lambda))]=0 \tag{2.12}$$

for $\mathbf{k}$, $\mathbf{k}'$ subject to $\Sigma_{\alpha=1}^{n}k_\alpha-1=\Sigma_{\alpha=1}^{n}k_\alpha'+1=0$. The shifts $\boldsymbol{\delta}_\alpha$ and $\boldsymbol{\epsilon}_\alpha(\lambda)$ only affect the αth component:

$$\begin{gathered}\mathbf{k}\pm\boldsymbol{\delta}_\alpha=(k_1,\ldots,k_\alpha\pm1,\ldots,k_n),\\ \mathbf{x}\pm\boldsymbol{\epsilon}_\alpha(\lambda):\ \mathbf{x}^{(\alpha)}\rightarrow\left(x_1^{(\alpha)}\pm\frac{1}{\lambda},x_2^{(\alpha)}\pm\frac{1}{2\lambda^2},x_3^{(\alpha)}\pm\frac{1}{3\lambda^3},\cdots\right).\end{gathered} \tag{2.13}$$

At lowest order, we find that the bilinear identity (2.12) contains a system of coupled second-order equations[15] made up of $n(n-1)$ copies of Toda molecule equations[23,24] (expressed using the Hirota D-operators[25]),

$$D_{x_1^{(\beta)}}D_{x_1^{(\alpha)}}\tau^{\mathbf{k}}\cdot\tau^{\mathbf{k}}=2\tau^{\mathbf{k}+\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha}\tau^{\mathbf{k}-\boldsymbol{\delta}_\beta+\boldsymbol{\delta}_\alpha}\quad\forall\beta\neq\alpha, \tag{2.14}$$

and a higher-order system involving $x_2^{(\beta)}$-derivatives

$$D_{x_2^{(\beta)}}D_{x_1^{(\alpha)}}\tau^{\mathbf{k}}\cdot\tau^{\mathbf{k}}=2D_{x_1^{(\beta)}}\tau^{\mathbf{k}+\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha}\cdot\tau^{\mathbf{k}-\boldsymbol{\delta}_\beta+\boldsymbol{\delta}_\alpha}\quad\forall\beta\neq\alpha \tag{2.15}$$

[as will be explained later on, opposed to the ordinary A_∞ Toda lattice,[17,26] the lattices described by the equations (2.14) correspond to finite or at most semi-infinite chains of tau functions, hence the name ''molecule'']. An intrinsic property of the n-component KP hierarchy (which is crucial to our further purposes) is that a tau function $\tau^{\mathbf{n}}$ always satisfies the (one-component) KP hierarchy expressed in terms of some variable $\mathbf{x}^{(\beta)}$. This can be readily seen as the bilinear identity (2.12) collapses to the regular KP bilinear identity when $\mathbf{k}=\mathbf{n}+\boldsymbol{\delta}_\beta$ and $\mathbf{k}'=\mathbf{n}-\boldsymbol{\delta}_\beta$ (for some specific βth variable) with $\mathbf{x}^{(\alpha)}=\mathbf{x}^{(\alpha)\prime}\,\forall\alpha\neq\beta$.

Introducing the functions

$$V_{\mathbf{k}}^{(\beta)}(\mathbf{x};p)=\langle\mathbf{k}+\boldsymbol{\delta}_\beta|e^{H(\mathbf{x})}\psi^{(\beta)}(p)g|\mathrm{vac}\rangle=(-1)^{\Sigma_{i=1}^{\beta-1}k_i}\tau^{\mathbf{k}}(\mathbf{x}-\boldsymbol{\epsilon}_\beta(p))p^{k_\beta}e^{\xi_\beta(\mathbf{x},p)}, \tag{2.16}$$

we can write down a bilinear identity describing the linear problem which underlies the n-component KP evolutions ($\forall\mathbf{x},\mathbf{x}'$):

$$\sum_{\alpha=1}^{n} \mathrm{Res}_\lambda [\epsilon_{\alpha\beta} (-1)^{\delta_{\alpha,\beta}+\Sigma_{i=1}^{\alpha-1}(k_i+k_i')} \lambda^{k_\alpha - k_\alpha' - 2 - \delta_{\alpha,\beta}} e^{\xi_\alpha(\mathbf{x}-\mathbf{x}',\lambda)}$$
$$\times \tau^{\mathbf{k}-\boldsymbol{\delta}_\alpha}(\mathbf{x}-\boldsymbol{\epsilon}_\alpha(\lambda)) V^{(\beta)}_{\mathbf{k}'+\boldsymbol{\delta}_\alpha}(\mathbf{x}'+\boldsymbol{\epsilon}_\alpha(\lambda);p)] = -V^{(\beta)}_{\mathbf{k}-\boldsymbol{\delta}_\beta}(\mathbf{x};p)\, \tau^{\mathbf{k}'+\boldsymbol{\delta}_\beta}(\mathbf{x}'), \tag{2.17}$$

with

$$\epsilon_{\alpha\beta} = \begin{cases} -1, & \text{if } \alpha < \beta. \\ 1, & \text{if } \alpha \geqslant \beta. \end{cases} \tag{2.18}$$

The linear problem itself can then be expressed in terms of the *wavefunctions*

$$\Psi^{(\beta)}_{\mathbf{k}}(\mathbf{x};p) = \frac{V^{(\beta)}_{\mathbf{k}}(\mathbf{x};p)}{\tau^{\mathbf{k}}(\mathbf{x})} \quad \forall \beta = 1,...,n, \tag{2.19}$$

the lowest-order equations being ($\forall \alpha \neq \beta$)

$$[\Psi^{(\beta)}_{\mathbf{k}}(\mathbf{x};p)]_{x_1^{(\beta)}} = v^{(\beta,\alpha)}_{\mathbf{k}} \Psi^{(\beta)}_{\mathbf{k}}(\mathbf{x};p) + \epsilon_{\alpha\beta} \Psi^{(\beta)}_{\mathbf{k}+\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha}(\mathbf{x};p)$$
$$[\Psi^{(\beta)}_{\mathbf{k}}(\mathbf{x};p)]_{x_1^{(\alpha)}} = \epsilon_{\beta\alpha}\, e^{-\theta^{(\beta,\alpha)}_{\mathbf{k}}} \Psi^{(\beta)}_{\mathbf{k}-\boldsymbol{\delta}_\beta+\boldsymbol{\delta}_\alpha}(\mathbf{x};p) \tag{2.20}$$

(remark that, when considering, e.g., the $x^{(\beta)}$-variables, the functions $\Psi^{(\beta)}_{\mathbf{k}}(\mathbf{x};p)$ are nothing but standard wavefunctions for the KP hierarchy expressed in those coordinates).

In the above linear equations we made use of the variables

$$\theta^{(\beta,\alpha)}_{\mathbf{k}} = \log \frac{(\tau^{\mathbf{k}})^2}{\tau^{\mathbf{k}+\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha} \tau^{\mathbf{k}-\boldsymbol{\delta}_\beta+\boldsymbol{\delta}_\alpha}}, \tag{2.21}$$

$$v^{(\beta,\alpha)}_{\mathbf{k}} = \left[\log \frac{\tau^{\mathbf{k}+\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha}}{\tau^{\mathbf{k}}} \right]_{x_1^{(\beta)}}. \tag{2.22}$$

Obviously, the following relationship exists between these variables,

$$[\theta^{(\beta,\alpha)}_{\mathbf{k}}]_{x_1^{(\beta)}} = v^{(\beta,\alpha)}_{\mathbf{k}-\boldsymbol{\delta}_\beta+\boldsymbol{\delta}_\alpha} - v^{(\beta,\alpha)}_{\mathbf{k}}, \tag{2.23}$$

as a result of which we find the 2-D Toda-lattice-like nonlinear equations[26]

$$[\theta^{(\beta,\alpha)}_{\mathbf{k}}]_{x_1^{(\beta)},x_1^{(\alpha)}} = 2\exp(-\theta^{(\beta,\alpha)}_{\mathbf{k}}) - \exp(-\theta^{(\beta,\alpha)}_{\mathbf{k}+\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha}) - \exp(-\theta^{(\beta,\alpha)}_{\mathbf{k}-\boldsymbol{\delta}_\beta+\boldsymbol{\delta}_\alpha}) \tag{2.24}$$

as compatibility conditions of the linear system (2.20) [note that the ‘‘compatibility’’ conditions which the first equations in the system (2.20) obviously impose are rather involved and of higher order]. Formula (2.21) acts as the dependent variable transformation connecting the above nonlinear equations to the bilinear equations (2.14). A bilinear identity similar to (2.17), describing the adjoint linear problem in terms of adjoint wavefunctions

$$\Psi^{*(\beta)}_{\mathbf{k}}(\mathbf{x};q) = \frac{\langle \mathbf{k}-\boldsymbol{\delta}_\beta | e^{H(\mathbf{x})} \psi^{*(\beta)}(q) g | \mathrm{vac} \rangle}{\tau^{\mathbf{k}}(\mathbf{x})} \quad \forall \beta = 1,...,n, \tag{2.25}$$

will not be given here as its explicit form is of no real importance in what follows. Remark, however, that a ratio $\tau^{\mathbf{n}+\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\gamma}/\tau^{\mathbf{n}} (\gamma \neq \beta)$ satisfies the (one-component) KP linear problem expressed in terms of the $\mathbf{x}^{(\beta)}$-variables and the $\tau^{\mathbf{n}}$ tau function and that similarly the ratio $\tau^{\mathbf{n}-\boldsymbol{\delta}_\beta+\boldsymbol{\delta}_\gamma}/\tau^{\mathbf{n}}$ satisfies the adjoint linear problem for the KP hierarchy in the $\mathbf{x}^{(\beta)}$-variables. The former property can easily be proven by choosing $\mathbf{k}=\mathbf{n}+2\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\gamma$, $\mathbf{k}'=\mathbf{n}-\boldsymbol{\delta}_\beta$, and $\mathbf{x}^{(\alpha)}=\mathbf{x}^{(\alpha)'}$ for all $\alpha\neq\beta$ and some specific $\gamma\neq\beta$ in the bilinear identity (2.12), reducing it to a modified KP bilinear identity in the $\mathbf{x}^{(\beta)}$ variables.

Finally, let us turn to some important examples of multi-component tau functions. In general, these are constructed through (auto-) Bäcklund transformations performed on the group elements $g\in\mathrm{GL}(\infty)$:

$$g\to(c+\phi\phi^*)g, \tag{2.26}$$

with constant c and where ϕ and ϕ^* are generalized fermion operators

$$\phi=\sum_{\alpha=1}^{n}\mathrm{Res}_{p_\alpha}[h_\alpha(p_\alpha)\psi^{(\alpha)}(p_\alpha)],\quad \phi^*=\sum_{\alpha=1}^{n}\mathrm{Res}_{q_\alpha}[h_\alpha^*(q_\alpha)\psi^{*(\alpha)}(q_\alpha)], \tag{2.27}$$

defined in terms of certain ''densities'' h_α and h_α^*.

The corresponding tau functions then transform as[27]

$$\tau(\mathbf{x})=\langle\mathrm{vac}|g(\mathbf{x})|\mathrm{vac}\rangle\to\hat{\tau}(\mathbf{x})=\langle\mathrm{vac}|e^{H(\mathbf{x})}(c+\phi\phi^*)g|\mathrm{vac}\rangle\equiv\tau(\mathbf{x})\times\Omega(\mathbf{x}), \tag{2.28}$$

the resulting tau function being ''factorizable'' in terms of a so-called (multi-component) *eigenfunction potential* $\Omega(\mathbf{x})$:

$$\Omega(\mathbf{x})\equiv\sum_{\alpha,\beta}^{n}\mathrm{Res}_{p_\alpha,q_\beta}[h_\alpha(p_\alpha)h_\beta^*(q_\beta)\Omega^{(\alpha,\beta)}(\mathbf{x})], \tag{2.29}$$

$$\Omega^{(\alpha,\beta)}(\mathbf{x})=\langle\mathrm{vac}|e^{H(\mathbf{x})}\psi^{(\alpha)}(p_\alpha)\psi^{*(\beta)}(q_\beta)g|\mathrm{vac}\rangle/\tau(\mathbf{x}). \tag{2.30}$$

A basic explicit solution arises from repeated Bäcklund transformations (2.26) applied to the *vacuum* element $g^{\mathrm{vac}}=1$,

$$g_G=\prod_{i=1}^{N}(c_i+\phi_i\phi_i^*), \tag{2.31}$$

in terms of generalized operators as in (2.27),

$$\phi_i=\sum_{\alpha=1}^{n}\mathrm{Res}_{p_{\alpha,i}}[h_{\alpha,i}(p_{\alpha,i})\psi^{(\alpha)}(p_{\alpha,i})],\quad \phi_i^*=\sum_{\alpha=1}^{n}\mathrm{Res}_{q_{\alpha,i}}[h_{\alpha,i}^*(q_{\alpha,i})\psi^{*(\alpha)}(q_{\alpha,i})], \tag{2.32}$$

and by using Wick's theorem to calculate the corresponding tau function for the color $\mathbf{k}\equiv\mathbf{0}$:

$$\tau_G^0=\det[(m_{ij})_{N\times N}]. \tag{2.33}$$

The entries of this Grammian determinant are well-defined potentials m_{ij}:[7,27,28]

$$(m_{ij})_{x_1^{(\alpha)}}=\varphi_i^{(\alpha)}\varphi_j^{*(\alpha)}\quad\forall\alpha=1,\ldots,n, \tag{2.34}$$

$$(m_{ij})_{x_k^{(\alpha)}}=A_k^{(\alpha)}[\varphi_i^{(\alpha)},\varphi_j^{*(\alpha)}]\equiv\sum_{s=0}^{k-1}(-1)^s(\varphi_i^{(\alpha)})_{(k-1-s)x_i^{(\alpha)}}(\varphi_j^{*(\alpha)})_{sx_i^{(\alpha)}}, \tag{2.35}$$

$$(A_k^{(\alpha)}[\varphi_i^{(\alpha)},\varphi_j^{*(\alpha)}])_{x_l^{(\alpha)}}=(A_l^{(\alpha)}[\varphi_i^{(\alpha)},\varphi_j^{*(\alpha)}])_{x_k^{(\alpha)}}\quad\forall k,l. \tag{2.36}$$

The functions $\varphi_i^{(\alpha)}$ and $\varphi_j^{*(\alpha)}$ appearing in the determinant (2.33)

$$\varphi_i^{(\alpha)}=\mathrm{Res}_{p_{\alpha,i}}[h_{\alpha,i}(p_{\alpha,i})\epsilon^{\xi_\alpha(\mathbf{x},p_{\alpha,i})}],\quad \varphi_j^{*(\alpha)}=\mathrm{Res}_{q_{\alpha,j}}[h^*_{\alpha,j}(q_{\alpha,j})\epsilon^{-\xi_\alpha(\mathbf{x},q_{\alpha,j})}],\tag{2.37}$$

only depend on the $\mathbf{x}^{(\alpha)}$-coordinates and satisfy the dispersion relations ($\forall l=1,2,...$)

$$(\varphi_i^{(\alpha)})_{x_l^{(\alpha)}}=(\varphi_i^{(\alpha)})_{lx_1^{(\alpha)}},\quad (\varphi_j^{*(\alpha)})_{x_l^{(\alpha)}}=(-1)^{l+1}(\varphi_j^{*(\alpha)})_{lx_l^{(\alpha)}}.\tag{2.38}$$

The tau functions $\tau^{\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha}$ and $\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_\beta}$ linked to the tau function (2.33) by the lattice equations (2.14) can be easily calculated from the above group element g_G by what amounts to a Schlesinger transformation, e.g.,[29]

$$\mathcal{S}_{\beta\alpha}:\ g\to\epsilon_{\beta\alpha}S_\beta^{-1}S_\alpha g\ \Rightarrow\ \tau^{\boldsymbol{\delta}_\beta-\boldsymbol{\delta}_\alpha}=\epsilon_{\beta\alpha}\langle\mathrm{vac}|S_\beta^{-1}S_\alpha g_G(\mathbf{x})|\mathrm{vac}\rangle.\tag{2.39}$$

In fact, all tau functions of different color in the zero-charge section of the Fock space can be obtained by application of such transformations. Use of Wick's theorem then shows that, in general, these tau functions are representable as determinants of combined Wronskian/Grammian type:

$$H_{(\beta_1,\dots,\beta_l)}^{(\alpha_1\dots,\alpha_k)}=\begin{vmatrix}\mathbf{O}_{N_W\times N_W} & W^*_{(\alpha_1,\dots,\alpha_k)}\\ W_{(\beta_1,\dots,\beta_l)} & (m_{ij})_{N\times N}\end{vmatrix},\tag{2.40}$$

the indices $\{\alpha_1,\dots,\alpha_k,\beta_1,\dots,\beta_l\}$ forming a subset of the set $\{1,...,n\}$ (i.e., $k+l\leqslant n$) and where $N_W=\Sigma_{s=1}^k N_{\alpha_s}=\Sigma_{s=1}^l N_{\beta_s}\leqslant N$. The matrices $W_{(\beta_1,\dots,\beta_l)}$ and $W^*_{(\alpha_1,\dots,\alpha_k)}$ themselves exhibit a block structure,

$$W_{(\beta_1,\dots,\beta_l)}\equiv(\mathcal{W}_{N_{\beta_1}}(\varphi^{(\beta_1)})\cdots\mathcal{W}_{N_{\beta_l}}(\varphi^{(\beta_1)})),\quad W^*_{(\alpha_1,\dots,\alpha_k)}\equiv\begin{pmatrix}\mathcal{W}^*_{N_{\alpha_1}}(\varphi^{*(\alpha_1)})\\ \vdots\\ \mathcal{W}^*_{N_{\alpha_k}}(\varphi^{*(\alpha_k)})\end{pmatrix},\tag{2.41}$$

defined in terms of Wronskian-like matrices

$$\mathcal{W}_{N_\beta}(\varphi^{(\beta)})=\begin{pmatrix}\varphi_1^{(\beta)} & (\varphi_1^{(\beta)})_{x_1^{(\beta)}} & \cdots & (\varphi_1^{(\beta)})_{(N_\beta-1)x_1^{(\beta)}}\\ \vdots & \vdots & \ddots & \vdots\\ \varphi_N^{(\beta)} & (\varphi_N^{(\beta)})_{x_1^{(\beta)}} & \cdots & (\varphi_N^{(\beta)})_{(N_\beta-1)x_1^{(\beta)}}\end{pmatrix},\tag{2.42}$$

$$\mathcal{W}^*_{N_\alpha}(\varphi^{*(\alpha)})=\begin{pmatrix}\varphi_1^{*(\alpha)} & \cdots & \varphi_N^{*(\alpha)}\\ (\varphi_1^{*(\alpha)})_{x_1^{(\alpha)}} & \cdots & (\varphi_N^{*(\alpha)})_{x_1^{(\alpha)}}\\ \vdots & \ddots & \vdots\\ (\varphi_1^{*(\alpha)})_{(N_\alpha-1)x_1^{(\alpha)}} & \cdots & (\varphi_N^{*(\alpha)})_{(N_\alpha-1)x_1^{(\alpha)}}\end{pmatrix}.\tag{2.43}$$

The ''effect'' of the Schlesinger transformation $\mathcal{S}_{\beta\alpha}$ (2.39) on such a general determinant (2.40) then is to add a column (containing N_β^{th} $x_1^{(\beta)}$-derivatives) to the β-component block in W, while deleting the last row of the α-component block in W^*. A Schlesinger $\mathcal{S}_{\beta_i\beta_j}$ ($\mathcal{S}_{\alpha_i\alpha_j}$) acts by adding and deleting columns (rows) in the β_i and β_j (α_j and α_i!) blocks, respectively. Moreover, the determinant of the matrix $W_{(1,\dots,n)}$ itself is a Wronskian tau function for the multi-component KP hierarchy, constructed from a trivial element $g^{\mathrm{vac}}=1$ by means of the repeated action of so-called Darboux transformations:

$$gW=S_1^{-N_1}S_2^{-N_2}\cdots S_n^{-N_n}\prod_{i=1}^{N}\phi_i\ \Rightarrow\ |W_{(1,\dots,n)}|=\langle\mathrm{vac}|gW(\mathbf{x})|\mathrm{vac}\rangle\tag{2.44}$$

with $N=\Sigma_{i=1}^{n}N_i$, some N_i's possibly being zero. The Schlesinger transformation (2.39) acts on $W_{(1,\ldots,n)}$ as described above for $\mathcal{S}_{\beta_i\beta_j}$. Similarly, the determinant of $W^*_{(1,\ldots,n)}$ is a Wronskian tau function as well, this time obtained through repeated action of adjoint Darboux transformations:[27] $g\rightarrow S\phi^* g$. The Schlesinger action is as explained above for $\mathcal{S}_{\alpha_i\alpha_j}$.

III. PSEUDO-REDUCTIONS

In this section we shall consider a reduction procedure which describes the constrained KP hierarchies as part of an m-component KP hierarchy, for which m denotes the number of components ($m\geqslant 2$). As mentioned in the Introduction, this reduction procedure will turn out to be of a rather nonstandard type as it generically does not act as a dimensional reduction for the entire multi-component hierarchy, but only eliminates certain (lowest order) time evolutions in favor of other time variables. Because of this fact and because of the considerable notational complexity already apparent in the previous section, we shall first rewrite some of the relevant formulas in a more telling way.

In our discussion of reductions, we shall be primarily interested in their effect on the molecule Toda evolutions for tau functions (2.14) and their derivative nonlinear fields (2.24). Furthermore, as the ultimate goal is to describe a reduction of the KP hierarchy which only involves a single set of variables, we shall privilege the $\mathbf{x}^{(1)}$-variables by taking them to be the (final) KP variables whereby dropping the superscript altogether: $\mathbf{x}^{(1)}\rightarrow\mathbf{x}$. The other sets of $\mathbf{x}^{(\alpha)}$-variables in the multi-component hierarchy will be renamed as $\mathbf{y}^{(\alpha)}$ $(\alpha=2,\ldots,m)$. In this new notation, the $\mathbf{x}$-evolutions of a (multi-component) tau function $\tau(\mathbf{x},\mathbf{y}^{(\alpha)})$ are governed by the ''standard'' KP evolution equations, with respect to which one is permitted to think of the $\mathbf{y}^{(\alpha)}$-evolutions as being purely parametric. Also, as we only consider tau functions defined in the zero-charge section of the Fock space(s), we shall drop the explicit reference to the (basic) color $\mathbf{k}$ in the description of the tau functions, e.g., $\tau^{\mathbf{k}+\boldsymbol{\delta}_\gamma}\rightarrow\tau^{\boldsymbol{\delta}_\gamma}$.

In the following our chief interest will be the subset of bilinear equations among (2.14) describing the x_1-evolutions of the multi-component tau functions:

$$D_{x_1}D_{y_1^{(\alpha)}}\tau\cdot\tau=2\,\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}\quad\forall\,\alpha=2,\ldots,m, \tag{3.1}$$

the relevant linear problem being

$$\begin{aligned}&[\Psi(\lambda)]_{x_1}=v^{(1,\alpha)}\Psi(\lambda)+\Psi_{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}(\lambda),\\&[\Psi(\lambda)]_{y_1^{(\alpha)}}=-e^{-\theta^{(1,\alpha)}}\Psi_{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}(\lambda),\end{aligned} \tag{3.2}$$

where we have dropped the reference to the first component in $\Psi^{(1)}\rightarrow\Psi$, as well as the explicit dependence on the $\mathbf{x}$ and $\mathbf{y}^{(\alpha)}$ variables or on the $\mathbf{k}$-state in $\Psi(p),v^{(1,\alpha)}$ and $\theta^{(1,\alpha)}$. As for the functions $\Psi(\mathbf{x},\mathbf{y}^{(\alpha)};\lambda)$, it should be clear (cf. Sec. II) that they solve the standard KP linear problem expressed in the $\mathbf{x}$-coordinates.

Let us now, for some fixed integer k, impose the following reduction on the m-component tau functions:

$$\sum_{\alpha=2}^{m}\tau_{y_1^{(\alpha)}}=\tau_{x_k}. \tag{3.3}$$

In the excellent account of the multi-component KP hierarchy given in Ref. 15 it is shown that the case $k=1$ is a standard reduction for m-component hierarchies[30] as it results in subsequent identifications of the sort $\Sigma_{\alpha=2}^{m}\tau_{y_l^{(\alpha)}}=\tau_{x_l}$, $\forall\, l=1,2,\ldots$. The generic case, however, does not produce such subsequent identities for the $\mathbf{x}$- and $\mathbf{y}$- variables considered here (suitable coordinate transformations might alter this picture considerably, but these are a different topic altogether). Hence, we shall refer to the reduction (3.3) as a pseudo reduction. However, the effect of this reduction on the **x** dependence of the tau functions is dramatic, as it allows us to eliminate the $\mathbf{y}^{(\alpha)}$ dependence from the equations (3.1):

$$D_{x_1}D_{x_k}\tau\cdot\tau=2\sum_{\alpha=2}^{m}\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}, \tag{3.4}$$

thereby providing a constraint on the **x**-evolutions of the tau functions. If, after this reduction, we choose to regard the $\mathbf{y}^{(\alpha)}$-variables as mere parameters—the **x**-evolution being the only relevant one—we obtain a constraint on ordinary KP tau functions. Equation (3.4) can be rewritten as

$$(\log\tau)_{x_1x_k}=\tau^{-2}\sum_{\alpha=2}^{m}\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}=\sum_{\alpha=2}^{m}q_\alpha r_\alpha, \tag{3.5}$$

where the functions q_α and r_α are defined as the ratios

$$q_\alpha=\frac{\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}}{\tau},\qquad r_\alpha=\frac{\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}}{\tau}, \tag{3.6}$$

for which it was pointed out in the previous section that they correspond to solutions of the KP linear problem (q_α) and adjoint linear problem (r_α). If we now consider the one-component case of the eigenfunction potential (2.30) introduced in Sec. III,

$$\Omega(\mathbf{x})=\mathrm{Res}_{p,q}[\langle\mathrm{vac}|e^{H(\mathbf{x})}h(p)\psi(p)h^*(q)\psi^*(q)g|\mathrm{vac}\rangle]/\tau(\mathbf{x}), \tag{3.7}$$

we recover the standard KP eigenfunction potential. We shall not go into too much detail here (the interested reader is referred to Refs. 7, 31, 27 and 28 for a full account on eigenfunction potentials), but we do need to mention two basic properties which are of importance to our discussion. First, as $\Omega(\mathbf{x})$ is a well-defined potential wrt gradients in **x** [generalizing the one-component case of the potential in (2.34)], it can, for example, be defined (up to an arbitrary constant) in terms of its x_1 derivative,

$$[\Omega(\Phi,\Phi^*)]_{x_1}=\Phi(\mathbf{x})\Phi^*(\mathbf{x}), \tag{3.8}$$

where the functions $\Phi(\mathbf{x})$ and $\Phi^*(\mathbf{x})$ are general solutions to the KP linear and adjoint linear problems, respectively:

$$\Phi(\mathbf{x})=\mathrm{Res}_p[h(p)\langle 1|e^{H(\mathbf{x})}\psi(p)g|\mathrm{vac}\rangle]/\tau(\mathbf{x}), \tag{3.9}$$

$$\Phi^*(\mathbf{x})=\mathrm{Res}_q[h^*(q)\langle -1|e^{H(x)}\psi^*(q)g|\mathrm{vac}\rangle]/\tau(\mathbf{x}). \tag{3.10}$$

Second and most importantly, a shift (2.13) on the **x**-variables in $\Omega(\Phi,\Phi^*)$ is also completely defined in terms of its ''constituent'' functions $\Phi(\mathbf{x})$ and $\Phi^*(\mathbf{x})$:

$$\Omega(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))-\Omega(\mathbf{x})\equiv-\frac{1}{\lambda}\Phi(\mathbf{x})\Phi^*(\mathbf{x}-\boldsymbol{\epsilon}(\lambda)). \tag{3.11}$$

In other words, to each pair of functions (q_α, r_α) in (3.5) we can associate a KP eigenfunction potential $\Omega_\alpha(\mathbf{x})$

$$[\Omega_\alpha(\mathbf{x})]_{x_1}=q_\alpha r_\alpha\quad\forall\alpha=2,\ldots,m, \tag{3.12}$$

and hence the constraint (3.5) is equivalent to

$$(\log \tau)_{x_k} = \sum_{\alpha=2}^{m} \Omega_\alpha(q_\alpha, r_\alpha) + C(\mathbf{x}), \quad C_{x_1}=0. \tag{3.13}$$

Next we shall proceed to show that the function $C(\mathbf{x})$ is not only independent of x_1, but that it actually is a constant that can be absorbed into the definitions of the potentials Ω_α. The result, however, is that the pseudo-reduction (3.3) turns the Toda molecule equations (3.1) into a relation which, when imposed on KP tau functions, reduces the KP hierarchy to an $(m-1)$-vector k-constrained hierarchy (see Ref. 7 for a full discussion of the implications of this particular relation and the way it corresponds to other definitions of constrained hierarchies):

$$\tau_{x_k} = \sum_{\alpha=2}^{m} \tau \times \Omega_\alpha(q_\alpha, r_\alpha) = \sum_{\alpha=2}^{m} \hat{\tau}_\alpha \tag{3.14}$$

[that a product $\tau\Omega_\alpha(q_\alpha, r_\alpha)$ actually defines a new KP tau function $\hat{\tau}_\alpha$ should be clear from the defining property (2.28) of an eigenfunction potential].

Proving that $C_{x_l}=0(\forall l=1,2,...)$ is quite straightforward if one expresses the second equation in the linear system (3.2) in terms of tau functions,

$$\begin{aligned}
[\Psi(\lambda)]_{y_1^{(\alpha)}} &= -\frac{\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}}{\tau^2} \times \frac{V^{(1)}_{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}(\lambda)}{\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}} \\
&= -\frac{1}{\lambda}\frac{\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}}{\tau} \times \frac{\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))}{\tau} e^{\xi(\mathbf{x},\lambda)},
\end{aligned} \tag{3.15}$$

making use of the definitions (2.21), (2.19), and (2.16) (for the color $\mathbf{k}=\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1$), also allowing one to write

$$\begin{aligned}
[\log \Psi(\lambda)]_{y_1^\alpha} &= -\frac{1}{\lambda}\frac{\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}}{\tau} \times \frac{\tau^{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))}{\tau(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))} \\
\Leftrightarrow \left[\log \frac{\tau(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))}{\tau}\right]_{y_1^{(\alpha)}} &= -\frac{1}{\lambda} q_\alpha r_\alpha(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))
\end{aligned} \tag{3.16}$$

[remark that the exponential appearing in $\Psi(\lambda)$ only depends on $\mathbf{x}$]. If one now imposes the pseudo reduction (3.3) on this last equation, one obtains

$$\begin{aligned}
\left[\log \frac{\tau(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))}{\tau}\right]_{x_k} &= -\frac{1}{\lambda}\sum_{\alpha=2}^{m} q_\alpha r_\alpha(\mathbf{x}-\boldsymbol{\epsilon}(\lambda)) \\
&= \sum_{\alpha=2}^{m} [\Omega_\alpha(\mathbf{x}-\boldsymbol{\epsilon}(\lambda)) - \Omega_\alpha(\mathbf{x})].
\end{aligned} \tag{3.17}$$

due to property (3.11) of the eigenfunction potentials. Comparison with relation (3.13) (and a shifted version thereof) immediately shows that $C(\mathbf{x}-\boldsymbol{\epsilon}(\lambda))=C(\mathbf{x})$ and hence that C is actually a constant.

The effect of the pseudo-reduction on the linear problem for the KP hierarchy is also easily calculated. We immediately have that

$$[\Psi(\lambda)]_{x_k} = \sum_{\alpha=2}^{m} [\Psi(\lambda)]_{y_1^{(\alpha)}} + \lambda^k \Psi(\lambda), \tag{3.18}$$

which, due the linear equations (3.2), can be made independent of the $\mathbf{y}^{(\alpha)}$-variables:

$$[\Psi(\lambda)]_{x_k}-\lambda^k\Psi(\lambda)+\sum_{\alpha=2}^{m} e^{-\theta^{(1,\alpha)}}\Psi_{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}(\lambda)=0, \tag{3.19}$$

after which these can be regarded as mere parameters appearing in the functions $\Psi(\mathbf{x};\lambda)$. Finally, the above equation can be turned into a relation which contains no reference whatsoever to the multi-component nature of the functions involved and which can then be imposed as a constraint on the KP linear problem. It suffices to notice that because of the obvious identifications,

$$v^{(1,\alpha)}=[\log q_\alpha]_{x_1}, \quad v^{(1,\alpha)}_{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}=-[\log r_\alpha]_{x_1}, \quad \theta^{(1,\alpha)}=-\log q_\alpha r_\alpha, \tag{3.20}$$

the first set of linear equations in (3.2) contains two very special equations,

$$(q_\alpha\partial_{x_1}q_\alpha^{-1})[\Psi(\lambda)]=\Psi_{\boldsymbol{\delta}_1-\boldsymbol{\delta}_\alpha}(\lambda), \tag{3.21}$$

$$(r_\alpha^{-1}\partial_{x_1}r_\alpha)[\Psi_{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}(\lambda)]=\Psi(\lambda), \tag{3.22}$$

on account of which we have the property that

$$(q_\alpha\partial_{x_1}q_\alpha^{-1})[e^{-\theta^{(1,\alpha)}}\Psi_{\boldsymbol{\delta}_\alpha-\boldsymbol{\delta}_1}(\lambda)]=q_\alpha r_\alpha\Psi(\lambda). \tag{3.23}$$

Hence, in an appropriate pseudo-differential context where the ''inverse'' of relation (3.23) is meaningful, formula (3.19) can be cast into a pseudo-differential constraint on the KP wavefunctions:

$$[\Psi(\lambda)]_{x_k}+\sum_{\alpha=2}^{m} q_\alpha\partial^{-1}r_\alpha\Psi(\lambda)=\lambda^k\Psi(\lambda), \tag{3.24}$$

which is nothing but the spectral problem for the $(m-1)$-vector k-constrained KP hierarchy discussed in much detail in Refs. 3, 11 and 1 where it is used to define the constrained KP hierarchies. In the last section of this paper we shall discuss different ways of turning the constraint (3.19) into a purely differential spectral problem and connections of these to different bilinear forms for the reduced hierarchies.

IV. SOLUTIONS OF THE CONSTRAINED KP HIERARCHIES

We now come to the problem of constructing solutions for the $(m-1)$-vector k-constrained KP hierarchy, in light of the reduction procedure which was described in the previous chapter. However, taking the constraint (3.3) at face value, it is not at all clear how to implement it on explicit examples of multi-component tau functions in order to obtain a reduced tau function. The key to solving this problem lies with the Toda molecule equation (3.1). As is described in the literature for Wronskian solutions,[24] there exist two different types of Wronskian determinants solving the Toda molecule equation: those of multi-component type [as in (2.44)] and those of so-called *bidirectional* type.[32] Between these two types of solutions there exists a peculiar correspondence which is, so to speak, tailored to the pseudo-reduction we wish to impose. Similarly, for the hybrid Wronskian/Grammian determinants described in Sec. II, there also exists a transition from multi-component ones to ''bidirectional'' ones, with a slight abuse of language as will be seen shortly (no written account of this correspondence is known to the authors). Before starting, let us introduce a rather explicit notation for the multi-component tau functions we shall be discussing. Instead of referring to different colors in the labels of the tau functions as we did up to now, we shall refer to the explicit form of, e.g., the Wronskian determinant (2.44) by labeling tau functions with the number of columns in the determinant alotted to each component:

$$\tau_{\mathbf{N}} \equiv \tau_{(N_1, N_2, \ldots, N_m)} = |W_{(1,\ldots,m)}|, \tag{4.1}$$

or slightly more involved for hybrid determinants (2.40) of size N,

$$\tau_{\mathbf{N}} \equiv \tau_{(N_1, N_2, \ldots, N_m)} = \sigma\, H^{(\alpha_1, \ldots, \alpha_k)}_{(\beta_1, \ldots, \beta_l)}, \tag{4.2}$$

where the l-tuple $(\beta_1, \ldots, \beta_l)$ corresponds to positive indices N_i for the tau function, whereas the k-tuple $(\alpha_1, \ldots, \alpha_k)$ for the adjoint Wronskian part corresponds to *negative* indices N_i, the number of rows in each component block then being $|N_i|$ [the factor $\sigma = -\mathrm{sgn}(N_1)$ is included in (4.2) to take care of the change in sign induced by the Schlesinger transformation (2.39) when the index N_1 becomes strictly greater than 0]. At this point it should also be remarked that the above tau functions always give rise to *finite* Toda chains (as opposed to semi-infinite ones). Clearly, the Toda molecule equations (3.1) link tau functions related through the Schlesinger actions described at the end of Sec. II, i.e., $\tau_{(N_1 N_2, \ldots, N_m)}$ is connected to two other tau functions $\tau_{(N_1 \pm 1, \ldots, N_\alpha \mp 1, \ldots)}$ in a chain. This means that tau functions (4.1) in the Wronskian chains become trivially zero whenever one of the components is ''exhausted'' (i.e., after some N_α or $N_1 = 0$), the number of nonzero tau functions (the ''chain length'') being equal to $L_\alpha = N_1 + N_\alpha + 1$. The hybrid tau functions (4.2) become trivial whenever the size of the Wronskian blocks (N_W) reaches that of the Grammian part (N), $N_W = N$, as the determinant then factorizes into the product of two determinants, one of which is independent of x_1 while the other does not depend on x_α. The length of such a chain can also be expressed in terms of the tau function labels (which can be negative in this case) as $L_\alpha = (2N+1) - |N_1 + N_\alpha|$.

A. Bidirectional Wronskian determinants

We start by discussing the case of Wronskian solutions, which is by far the simpler one. In a paper by Hirota *et al.*[24] on the 2-D molecule Toda equation, it is explained how two-component solutions $\tau_{(N_1, N_2)}$ for the molecule equation can give rise to determinant solutions with a striking bidirectional structure (which were originally obtained by Leznov and Saveliev[32]),

$$\overleftrightarrow{\tau}_n = \begin{vmatrix} f & f_{x_1} & \cdots & f_{(n-1)x_1} \\ f_{y_1} & f_{x_1, y_1} & \cdots & f_{(n-1)x_1, y_1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{(n-1)y_1} & \cdots & \cdots & f_{(n-1)x_1, (n-1)y_1} \end{vmatrix}, \tag{4.3}$$

the function $f(\mathbf{x}, \mathbf{y})$ depending on both sets of coordinates $\mathbf{x}$ and $\mathbf{y}$; subscripts denote (multiple) derivatives. That $\overleftrightarrow{\tau}_n$ satisfies the two-component case of (3.1) is easily verified since the Toda molecule equation

$$D_{x_1} D_{y_1} \overleftrightarrow{\tau}_n \cdot \overleftrightarrow{\tau}_n = 2\, \overleftrightarrow{\tau}_{n+1} \overleftrightarrow{\tau}_{n-1} \tag{4.4}$$

immediately reduces to a simple Jacobi identity (alternatively, see Ref. 24 for a proof based on Lapace expansions of determinants).

As it is very instructive to describe the correspondence with the two-component tau functions $\tau_{(N_1, N_2)}$ in detail and as notations in the multi-component case are quite involved, we shall at first limit ourselves to this example; afterwards the two-component case can readily be extended to the general m-component one. Our starting point is the observation that the Toda equations (3.1) are invariant under gauge transformations $\tau \to \tau \times X(\mathbf{x}) Y(\mathbf{y}^{(\alpha)})$, combined with the obvious fact that two tau functions suffice to completely determine a particular Toda chain (τ_n and $\tau_{n\pm1}$ always unambiguously define $\tau_{n\mp1}$). Hence, setting the tau function $\overleftrightarrow{\tau}_0 = 1$ [which can be identified with $\tau_{(0,N)} = |\mathcal{W}_N(\varphi^{(2)})|$ up to a gauge transformation in $\mathbf{y}$], we can ''mimic'' the Toda chain 1, f, $f f_{x_1, y_1} - f_{x_1} f_{y_1}, \ldots$ using two-component Wronskians:

$$\overleftrightarrow{\tau}_0 \equiv 1, \quad \overleftrightarrow{\tau}_1 = f(\mathbf{x},\mathbf{y}) \equiv \frac{\tau_{(1,N-1)}}{\tau_{(0,N)}} \Rightarrow \overleftrightarrow{\tau}_2 = \frac{\tau_{(2,N-2)}}{\tau_{(0,N)}}, \quad \ldots . \tag{4.5}$$

As pointed out above, this identification is unambiguous and thus the above sequence of bidirectional Wronskians will have finite length as well. Actually, from the explicit form of the function f as defined above in terms of matrices (2.42),

$$f(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N} (-1)^{1+i} \varphi_i^{(1)} \frac{M_{iN}[\mathcal{W}_N(\varphi^{(2)})]}{|\mathcal{W}_N(\varphi^{(2)})|} \tag{4.6}$$

(the symbol $M_{ij}[A]$ stands for the minor of the ijth element in the matrix A), it can easily be calculated that $\overleftrightarrow{\tau}_N$ factorizes and therefore becomes trivial in that it only depends on the $\mathbf{x}$-coordinates: $\overleftrightarrow{\tau}_N = |\mathcal{W}_N(\varphi^{(1)})|$.

Now that we have expressed bidirectional tau functions for the Toda molecule equations in terms of two-component tau functions, we can immediately find solutions of the form (4.3) to the reduced Toda molecule equations (3.4). It suffices to notice that although it is not a two-component tau function itself, because of the reduction (3.3) the following identity still holds for $f(\mathbf{x},\mathbf{y})$:

$$f_{y_1} = \left(\frac{\tau_{(1,N-1)}}{\tau_{(0,N)}} \right)_{y_1} \overset{Eq.(3.3)}{=} \left(\frac{\tau_{(1,N-1)}}{\tau_{(0,N)}} \right)_{x_k} = f_{x_k}. \tag{4.7}$$

However, whereas the gauge transformation we applied in the sequence (4.5) alters the $\mathbf{y}$-dependence of $\tau_{(1,N)}$, it does not alter its $\mathbf{x}$-evolutions and thus we have that the function $f(\mathbf{x},\mathbf{y})$ as defined in (4.5) is still a KP tau function for the $\mathbf{x}$-coordinates. Moreover, being defined as the ratio of two two-component tau functions related by a Schlesinger transformation, $f(\mathbf{x},\mathbf{y})$ also solves the KP linear problem in $\mathbf{x}$ with respect to the tau function $\tau_{(0,N)}$ (which for these evolutions is a mere constant). Hence it follows that $f(\mathbf{x},\mathbf{y})$ is an arbitrary function of $\mathbf{x}$ apart from the requirement that it satisfies the (standard) dispersion relations which also appear in Wronskian solutions of the KP hierarchy:

$$f_{x_l} = f_{lx_1} 0 \quad \forall l = 1,2,\ldots . \tag{4.8}$$

Consequently, we have that the bidirectional Wronskian determinants,

$$\overleftrightarrow{\tau}_n = \begin{vmatrix} f & f_{x_1} & \cdots & f_{(n-1)x_1} \\ f_{x_k} & f_{x_1 x_k} & \cdots & f_{(n-1)x_1, x_k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{(n-1)x_k} & \cdots & \cdots & f_{(n-1)x_1,(n-1)x_k} \end{vmatrix}, \tag{4.9}$$

are KP tau functions which also satisfy the constraint (3.14). In other words, this means that we have constructed a Wronskian solution for the scalar $(m=1)$ k-constrained KP hierarchy. These solutions appear in several papers.[4,8,11,33] Obviously, the functions q and r associated to these tau functions in the formulation (3.14) of the scalar k-constraint are given by the ratios [cf. formula (3.6)]

$$q = \frac{\overleftrightarrow{\tau}_{n+1}}{\overleftrightarrow{\tau}_n}, \quad r = \frac{\overleftrightarrow{\tau}_{n-1}}{\overleftrightarrow{\tau}_n}. \tag{4.10}$$

In the general case of the m-component KP hierarchy, the construction of bidirectional Wronskians goes through in pretty much the same way. First, we have to remark that the Toda molecule equations (3.1) allow for more general Wronskian solutions of type (4.3). Assigning the symbol $\mathcal{B}_{n,N}(f)$ to the matrix appearing in (4.3) as

$$\mathcal{B}_{n_i,n}(f_i)\equiv\begin{pmatrix} (f_i) & (f_i)_{x_1} & \cdots & (f_i)_{(n-1)x_1} \\ (f_i)_{y_1^{(i+1)}} & (f_i)_{x1,y_1^{(i+1)}} & \cdots & (f_i)_{(n-1)x_1,y_1^{(i+1)}} \\ \vdots & \vdots & \ddots & \vdots \\ (f_i)_{(n_i-1)y_1^{(i+1)}} & \cdots & \cdots & (f_i)_{(n-1)x_1,(n_i-1)y_1^{(i+1)}} \end{pmatrix}, \tag{4.11}$$

we have that determinants ($n=\Sigma_{i=1}^{m-1}n_i$)

$$\overleftrightarrow{\tau}_{n_1,\ldots n_{m-1}}=\begin{vmatrix} \mathcal{B}_{n_1,n}(f_1) \\ \vdots \\ \mathcal{B}_{n_{m-1},n}(f_{m-1}) \end{vmatrix}, \tag{4.12}$$

with functions $f_i(\mathbf{x},\mathbf{y}^{(i+1)})$ $(i=1,...,m-1)$ only depending on $\mathbf{x}$ and $\mathbf{y}^{(i+1)}$ each, still satisfy the Toda molecule equations (3.1) [or (4.4) with $\mathbf{y}=\mathbf{y}^{(i+1)}$ if one prefers]. This can be seen by direct calculation, a verification which is very much the same as in case of equation (4.4) because the different blocks in the determinants only depend on a single set of $\mathbf{y}^{(l)}$ coordinates and each Toda equation therefore only involves a single one of those blocks. Actually, this rather trivial intertwining (or should one say ''lack of'') of the different components $\mathbf{y}^{(l)}$ in the determinants (4.12) is at the same time a disappointment as well as a blessing. What is disappointing is that it makes it clear straightaway that there is no chance whatsoever for a general m-component Wronskian tau function to be mapped onto the above determinants, for such a mapping would entangle the components in a nontrivial way. One rather has to look at it as follows: the ''lattice'' of Toda chains one can obtain by Schlesinger transformations starting from a generic (N-sized) m-component Wronskian (2.44) contains exactly $m-1$ suitable starting pairs $(\alpha=1,...,m-1)$,

$$\overleftrightarrow{\tau}_{0,...,0}=1 \qquad \overleftrightarrow{\tau}_{\ldots,0,n_\alpha=1,0,\ldots}=f_\alpha(\mathbf{x},\mathbf{y}^{(\alpha+1)})\equiv\frac{\tau_{(1,...,0,N_{\alpha+1}=N-1,0,...)}}{\tau_{(\ldots,0,N_{\alpha+1}=N,0,\ldots)}}, \tag{4.13}$$

suitable for constructing strains of determinants $\overleftrightarrow{\tau}_{\ldots,0,n_\alpha,0,\ldots}$ in exactly the same way as was done for the two-component case. As these involve $m-1$ different gauge transformations (and as these destroy the tau function nature of the resulting determinants), the different strains are not related to each other as real (m-component) Toda chains would be. They can, however, be assembled into a single determinant expression like (4.12), as is explained above. The upside of this lack of nontrivial intertwining is that the pseudo-reduction (3.3) still implies that $(f_\alpha)_{y_1^{(\alpha+1)}}=(f_\alpha)_{x_k}$, making the construction of solutions to the constraint equation (3.14) a matter of replacing $y_1^{(l)}$-derivatives in (4.12) by x_k-derivatives. Thus one obtains the Wronskian solutions to the $(m-1)$ vector k-constrained KP hierarchy which were also described in Refs. 8, 10 and 11. The functions q_α and r_α associated to the tau functions in the constraint (3.14) are of course given by the obvious generalizations of the two-component ones (4.10).

B. Grammian determinants

Let us now move on to the hybrid-type solutions (4.2) and the solutions these give rise to for the constrained KP hierarchies. Again, initially we shall limit ourselves to the two-component case after which the generalization to the m-component case will be straightforward. Consider a Grammian tau function made up of potentials $\omega_{ij}(\mathbf{x},\mathbf{y})$ defined by their x_1-derivatives

$$\overleftrightarrow{\tau}_{N,N}=|(\omega_{ij})_{N\times N}|, \quad (\omega_{ij})_{x_1}=f_{(i-1)y_1}f_j^*, \tag{4.14}$$

in terms of a function $f(\mathbf{x},\mathbf{y})$ and N functions $f_j^*(\mathbf{x},\mathbf{y})$ obeying the dispersion relations

$$f_{x_n}=f_{nx_1} \quad (f_j^*)_{x_n}=(-1)^{n+1}(f_j^*)_{nx_1}, \tag{4.15}$$

$$(f_j^*)_{y_1}=-\mu_j f_j^*. \tag{4.16}$$

As should be clear from the discussion in Sec. II, the $\omega_{ij}(\mathbf{x},\mathbf{y})$ are well-defined potentials in the $\mathbf{x}$-coordinates, determined up to some functions $F_{ij}(\mathbf{y})$ which we require here to be related to only N arbitrary functions $F_{1j}(\mathbf{y})$ by means of the recursion relations:

$$F_{i+1,j}(\mathbf{y})\equiv(F_{ij})_{y_1}+\mu_j F_{ij} \quad \forall_i=1,\ldots,N. \tag{4.17}$$

Remark that the above requirement is equivalent to imposing the following recurrence relation on the $\omega_{ij}(\mathbf{x},\mathbf{y})$:

$$(\omega_{ij})_{y_1}=\omega_{i+1,j}-\mu_j\omega_{ij}, \tag{4.18}$$

for N ''original'' potentials ω_{1i}, the functions F_{ij} being (unambiguously) defined by $\omega_{ij}[f \equiv 0; f_j^*\equiv 0]=F_{ij}$. Furthermore, in analogy to the matrices (2.42) and (2.43), we also define matrices

$$\overleftrightarrow{\mathcal{W}}_{N,n}(f)=\begin{pmatrix} f & f_{x_1} & \cdots & f_{(n-1)x_1} \\ f_{y_1} & f_{y_1,x_1} & \cdots & f_{y_1,(n-1)x_1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{(N-1)y_1} & f_{(N-1)y_1,x_1} & \cdots & f_{(N-1)y_1,(n-1)x_1} \end{pmatrix}, \tag{4.19}$$

$$\overleftrightarrow{\mathcal{W}}^*_{n,N}(f_j^*)=\begin{pmatrix} f_1^* & \cdots & f_N^* \\ (f_1^*)_{x_1} & \cdots & (f_N^*)_{x_1} \\ \vdots & \ddots & \vdots \\ (f_1^*)_{(n-1)x_1} & \cdots & (f_N^*)_{(n-1)x_1} \end{pmatrix}. \tag{4.20}$$

Because of relations (4.16) and (4.18), the following derivatives of $\overleftrightarrow{\tau}_N$ can be easily expressed in terms of these matrices ($\mu=\Sigma_{j=1}^N \mu_N$):

$$(\overleftrightarrow{\tau}_{N,N})_{x_1}=-\begin{vmatrix} (\omega_{ij})_{N\times N} & \overleftrightarrow{\mathcal{W}}_{N,1} \\ \overleftrightarrow{\mathcal{W}}^*_{1,N} & 0 \end{vmatrix} \quad (\overleftrightarrow{\tau}_{N,N})_{y_1}=-\mu\overleftrightarrow{\tau}_N+\begin{vmatrix} (\omega_{ij})_{(N-1)\times N} \\ (\omega_{N+1,1}\cdots\omega_{N+1,N}) \end{vmatrix}, \tag{4.21}$$

$$(\overleftrightarrow{\tau}_{N,N})_{x_1,y_1}=-\mu(\overleftrightarrow{\tau}_N)_{x_1}-\begin{vmatrix} (\omega_{ij})_{(N-1)\times N} & \overleftrightarrow{\mathcal{W}}_{N-1,1} \\ (\omega_{N+1,1}\cdots\omega_{N+1,N}) & f_{Ny_1} \\ \overleftrightarrow{\mathcal{W}}^*_{1,N} & 0 \end{vmatrix} \tag{4.22}$$

with the convention that the symbol $(\omega_{ij})_{M\times N}$ denotes the $M\times N$ matrix obtained from $(\omega_{ij})_{N\times N}$ by the obvious addition or deletion of rows of ω_{ij} [the entries in additional rows being of course defined in terms of the previous ones using (4.18)]. Then, it is straightforward to show (using a simple Jacobi determinant identity) that $\overleftrightarrow{\tau}_N$ satisfies the Toda molecule equation

$$\tfrac{1}{2}D_{x_1}D_{y_1}\overleftrightarrow{\tau}_{N,N}\cdot\overleftrightarrow{\tau}_{N,N}\equiv\overleftrightarrow{\tau}_{N,N}(\overleftrightarrow{\tau}_{N,N})_{x_1,y_1}-(\overleftrightarrow{\tau}_{N,N})_{x_1}(\overleftrightarrow{\tau}_{N,N})_{y_1}=\overleftrightarrow{\tau}_{N,N+1}\overleftrightarrow{\tau}_{N,N-1}, \tag{4.23}$$

with $\overleftrightarrow{\tau}_{N,M}$ (for $M\neq N$) defined as

$$\overleftrightarrow{\tau}_{N,M<N}=\begin{vmatrix}(\omega_{ij})_{M\times N}\\ \overleftrightarrow{\mathcal{W}}^{*}_{N-M,N}\end{vmatrix},\quad \overleftrightarrow{\tau}_{N,M>N}=|(\omega_{ij})_{M\times N}\quad \overleftrightarrow{\mathcal{W}}_{M,M-N}|. \tag{4.24}$$

In general, it can be shown that these determinants $\overleftrightarrow{\tau}_{N,M}$ also satisfy the above Toda molecule equation, the proof being a straightforward generalization of the preceding one. It should also be noted that these solutions again exhibit a certain ''bidirectionality'' which is crucial in these proofs.

We can now proceed to discuss the construction of such determinant solutions using two-component tau functions. Due to the dispersion relation (4.16) we have that $(\overleftrightarrow{\tau}_{N,0})_{y_1}=-\mu\overleftrightarrow{\tau}_{N,0}$ and hence that $\overleftrightarrow{\tau}_{N,0}$ can actually be used as the starting point of the chain of $\overleftrightarrow{\tau}_{N,M}$ we wish to describe. We therefore first have to find a suitable two-component partner for this determinant. This can be done by considering the tau function $\tau_{(-N,N)}\,(N\geqslant 0)$, as defined in (4.2), together with a gauge $Y(\mathbf{y})=|\mathcal{W}_N(\varphi^{(2)})|^{-1}\times|\mathcal{W}^*_N(\varphi^{*(2)})|\times V(-\mu_j)^{-1}$ where we have chosen the $\mathbf{y}$-component adjoint functions to be

$$\varphi^{*(2)}_j\equiv\exp(-\xi(\mathbf{y},\mu_j)); \tag{4.25}$$

$V(-\mu_j)$ is the Vandermonde determinant in $-\mu_1,\dots,-\mu_j$. Hence we obtain a solution $\tau_{(-N,N)}$ to the (two-component) Toda molecule equations (3.1):

$$\begin{vmatrix}\mathbf{O}_{N\times N} & \mathcal{W}^*_N(\varphi^{*(1)})\\ \mathcal{W}_N(\varphi^{(2)}) & (m_{ij})_{N\times N}\end{vmatrix}\times|\mathcal{W}_N(\varphi^{(2)})|^{-1}\prod_{j=1}^{N}e^{-\xi(y,\mu_j)}=\mathcal{W}^*_N(-\varphi^{*(1)})\prod_{j=1}^{N}e^{-\xi(y,\mu_j)}, \tag{4.26}$$

which matches a determinant of the form $\overleftrightarrow{\tau}_{N,0}$ by choosing

$$f^*_j(\mathbf{x},\mathbf{y})\equiv-\varphi^{*(1)}_j e^{-\xi(\mathbf{y},\mu_j)}\quad \forall_j=1,\dots,N. \tag{4.27}$$

It can readily be verified that these $f^*_j(\mathbf{x},\mathbf{y})$ indeed satisfy the dispersion relations (4.15) and (4.16).

The second determinant in the Toda chain (4.23) can be written as

$$\overleftrightarrow{\tau}_{N,1}=\sum_{j=1}^{N}(-1)^{1+j}\omega_{1j}M_{Nj}[\overleftrightarrow{\mathcal{W}}^*_{N,N}]. \tag{4.28}$$

On the other hand, a Laplace expansion of the determinant $\tau_{(1-N,N-1)}$ given by (4.2) yields the following expression [after gauge transformation $Y(\mathbf{y})$]:

$$\tau_{(1-N,N-1)}\times|\mathcal{W}_N(\varphi^{(2)})|^{-1}\prod_{j=1}^{N}e^{-\xi(\mathbf{y},\mu_j)}=|\mathcal{W}_N(\varphi^{(2)})|^{-1}\sum_{i=1}^{N}M_{Ni}[\overleftrightarrow{\mathcal{W}}^*_{N,N}]e^{-\xi(\mathbf{y},\mu_i)}\times\left(\sum_{j=1}^{N}m_{ji}M_{jN}[\mathcal{W}_N(\varphi^{(2)})]\right), \tag{4.29}$$

which suggests the identification $(\forall_i=1,\dots,N)$

$$\omega_{1i}=e^{-\xi(\mathbf{y},\mu_i)}|\mathcal{W}_N(\varphi^{(2)})|^{-1}\left(\sum_{j=1}^{N}m_{ji}M_{jN}[\mathcal{W}_N(\varphi^{(2)})]\right). \tag{4.30}$$

Then, the x_1-derivative of ω_{1i} allows for an easy determination of the function $f(\mathbf{x},\mathbf{y})$ (the potential m_{ji} is the only part in this expression which depends on the $\mathbf{x}$-variables):

$$(\omega_{1i})_{x_1}=ff_i^*=e^{-\xi(\mathbf{y},\mu_i)}|\mathcal{W}_N(\varphi^{(2)})|^{-1}\left(\sum_{j=1}^N \varphi_j^{(1)}\varphi_i^{*(1)}M_{jN}[\mathcal{W}_N(\varphi^{(2)})]\right)$$

$$\overset{Eq.(4.27)}{\Leftrightarrow} f(\mathbf{x},\mathbf{y})=-\sum_{j=1}^N \varphi_j^{(1)}\frac{M_{jN}[\mathcal{W}_N(\varphi^{(2)})]}{|\mathcal{W}_N(\varphi^{(2)})|}, \tag{4.31}$$

which is essentially the same solution to the KP linear problem as in case of the Wronskian determinants (4.6). Hence, we have succeeded in expressing the functions f, f_j^* and the potentials ω_{1j} in terms of two-component tau functions. Moreover, the nontrivial $\mathbf{y}$-contribution in the two-component potentials m_{ij} is responsible for the fact that the ω_{1j} are only defined up to some arbitrary functions $F_{1j}(\mathbf{y})$. The other determinant solutions $\overleftrightarrow{\tau}_{N,M}$ in the Toda chain can now be calculated explicitly from the two determinants $\overleftrightarrow{\tau}_{N,0}$ and $\overleftrightarrow{\tau}_{N,1}$ we just constructed. The fact that this chain is finite follows from the exact identification with the two-component Toda chain $\tau_{M-N,N-M}$, but is not entirely trivial if one attempts a proof directly on the determinants themselves.

Of course, since the function $f(\mathbf{x},\mathbf{y})$ is the same as before, we know that it behaves as $f_{y_1}=f_{x_k}$ under the pseudo-reduction (3.3) and that it solves the KP linear problem in $\mathbf{x}$. Similarly, because the functions $f_j^*(\mathbf{x},\mathbf{y})$ defined in (4.27) can be expressed as the product of two two-component (adjoint) Wronskian tau functions $|W^*|$, we have that the pseudo-reduction not only implies that $(f_j^*)_{y_1}=(f_j^*)_{x_k}$, but [due to (4.27)] also that

$$(f_j^*)_{x_k}=-\mu_j f_j^*. \tag{4.32}$$

Discarding now the $\mathbf{y}$ dependence as merely parametric, we can define the potentials ω_{ij} as pure (one-component) KP eigenfunction potentials by replacing the recurrence relation (4.18) by

$$(\omega_{ij})_{x_k}=\omega_{i+1,j}-\mu_j\omega_{ij}. \tag{4.33}$$

This implies that the ''arbitrary'' constants c_{ij} in the potentials ω_{ij} [defined by $(\omega_{ij})_{x_1}=f_{(i-1)x_k}f_j^*$] are not all arbitrary, but are recursively defined in terms of N initial ones; in particular $c_{i+1,j}=\mu_j c_{ij}$. Hence, with $\overleftrightarrow{\tau}_{N,N}=|(\omega_{ij})|$ we have constructed a particular example of a KP Grammian solution which also satisfies the constraint (3.4) and thus $\overleftrightarrow{\tau}_{N,N}$ is a Grammian solution to the scalar k-constrained KP hierarchy. Similar solutions have been discussed previously in Refs. 7, 14, and 10. The functions q and r appearing in the constraint (3.14) are given by

$$q=\frac{\overleftrightarrow{\tau}_{N,N+1}}{\overleftrightarrow{\tau}_{N,N}}, \quad r=\frac{\overleftrightarrow{\tau}_{N,N-1}}{\overleftrightarrow{\tau}_{N}}. \tag{4.34}$$

We shall only briefly sketch the extension to the m-component case. It suffices to note that there also exist solutions ($N=\Sigma_{i=1}^{m-1}N_i$) to the Toda molecule equations in the form

$$\overleftrightarrow{\tau}_{N,N_1,\ldots,N_{m-1}}=\begin{vmatrix} (\omega_{ij}^{(1)})_{N_1\times N} \\ \vdots \\ (\omega_{ij}^{(m-1)})_{N_{m-1}\times N}\end{vmatrix}, \tag{4.35}$$

$$(\omega_{ij}^{(l)})_{x_1}=[f_l(\mathbf{x},\mathbf{y}^{(l+1)})]_{(i-1)y_1^{(l+1)}}f_j^*(\mathbf{x},\mathbf{y}), \tag{4.36}$$

where [in contrast to the $m-1$ functions $f_l(\mathbf{x},\mathbf{y}^{(l+1)})$] the N functions $f_j^*(\mathbf{x},\mathbf{y})$ depend on $\mathbf{x}$ and on all of the $\mathbf{y}^\alpha$ coordinates, such that

$$(f_l)_{x_n}=(f_l)_{nx_1}, \quad (f_j^*)_{x_n}=(-1)^{n+1}(f_j^*)_{nx_1}, \tag{4.37}$$

$$(f_j^*)_{y_1^{(\alpha)}} = -\mu_j^{(\alpha)} f_j^* \quad \forall \alpha = 2,...,m \tag{4.38}$$

(the proof being identical to the two-component case). The determinants (4.24) associated to these solutions are the obvious extensions of the above ones; in case of $M<N$, the ''adding or deleting'' of rows of potentials are restricted to a particular block (i.e., a particular coordinate set). Again, the different blocks in these determinants lead, so to speak, separate lives, and can therefore only be linked (by different gauges) to sequences of m-component tau functions which refer to different starting points. Thus, repeating the above construction step-by-step for each starting point $\tau_{(-N,\ldots,0,N_\alpha=N,0,\ldots)}$ of chains of m-component tau functions (4.2), for particular gauges $|\mathcal{W}_N(\varphi^{(\alpha)})|^{-1}\times\prod_{\beta=2}^m|\mathcal{W}_N^*(\varphi^{*(\beta)})|V(-\mu_j^{(\beta)})^{-1}$ and for the choice of $\varphi^{*(\beta)} = \exp[-\xi(\mathbf{y}^{(\beta)},\mu_j^{(\beta)})]$ $(\beta=2,...,m)$, we find that the

$$f_j^*(x,y) = -\varphi_j^{*(1)} \prod_{\beta=2}^m e^{-\xi(\mathbf{y}^{(\beta)},\mu_j^{(\beta)})} \quad \forall j=1,...,N. \tag{4.39}$$

One then proceeds to $\tau_{(1-N,\ldots,0,N_\alpha=N-1,0,\ldots)}$, which yields $m-1$ different functions f_l and consequently the $m-1$ series of potentials $\omega_{1j}^{(l)}$ needed to define the determinants (4.35). Since the functions f_l only depend on a single set of $\mathbf{y}$-variables and as the functions f_j^* obey

$$(f_j^*)_{x_k} = -\mu_j f_j^* \quad \text{with} \quad \mu_j = \sum_{\alpha=2}^m \mu_j^{(\alpha)}, \tag{4.40}$$

the pseudo-reduction is performed in exactly the same way as in the two-component case: i.e., the $(m-1)$-vector k-constrained KP hierarchy possesses Grammian solutions of type (4.35) where, at each occurrence, $y_1^{(l)}$ is replaced by x_k.

C. Alternative dispersion relations

In all the above solutions, Wronskians and Grammians alike, the functions appearing in the determinants (or in the potentials) satisfy trivial versions of the KP linear problem like the dispersion relations (4.15). This is because the tau functions themselves always refer to a trivial (i.e., constant) ''vacuum'' when one thinks of them as resulting from Bäcklund–Darboux transformations (as was explained in Sec. II). Some authors,[11,10] however, present determinant solutions referring to nontrivial vacuum, i.e., defined in terms of functions which satisfy nontrivial KP linear problems. The bidirectional nature of these solutions then manifests itself through a (highly) nonlinear operator, taking over the role of the ∂_{x_k}-operator. We shall briefly describe such solutions for the case of Wronskian solutions and scalar k-constrained hierarchies. However, it will turn out that these particular solutions can always be reformulated such that they again refer to trivial vacuum.

We shall use a Toda chain of two-component *adjoint* Wronskian solutions:

$$\tau_{(N_1,N_2)} = |W_{(1,2)}^*|, \quad N=N_1+N_2. \tag{4.41}$$

Starting the Toda chain with the tau function $\tau_{(N,0)}$, we take this to be our ''vacuum:''

$$\tau^V \equiv \tau_{(N,0)} = |\mathcal{W}_N^*(\varphi^{*(1)})|. \tag{4.42}$$

Now, performing exactly the same construction as in case of the Wronskian solutions, the second tau function in the chain allows us to define a function

$$g(\mathbf{x},\mathbf{y}) \equiv \frac{\tau_{(N-1,1)}}{\tau^V} = \sum_{i=1}^N (-1)^{i+N} \frac{M_{Ni}[\mathcal{W}_N^*(\varphi^{*(1)})]}{|\mathcal{W}_N^*(\varphi^{*(1)})|} \varphi_i^{*(2)}, \tag{4.43}$$

eventually giving rise to a chain of bidirectional Wronskian solutions,

$$\tau^V \times \begin{vmatrix} g & g_{x_1} & \cdots & g_{(n-1)x_1} \\ g_{y_1} & g_{x_1,y_1} & \cdots & g_{(n-1)x_1,y_1} \\ \vdots & \vdots & \ddots & \vdots \\ g_{(n-1)y_1} & \cdots & \cdots & g_{(n-1)x_1,(n-1)y_1} \end{vmatrix} \quad n=1,2...,N, \tag{4.44}$$

to the Toda molecule equation. Remark that we did not perform any gauge transformations on the original two-component chain, so the initial tau function τ^V appears as a multiplicative factor in these solutions. The ratio which defines $g(\mathbf{x},\mathbf{y})$ is actually the ratio of two tau functions $\tau^{\boldsymbol{\delta}_1-\boldsymbol{\delta}_2}$ and $\tau(=\tau^V)$ related by a Schlesinger transformation S_{12} (see Sec. II) and hence $g(\mathbf{x},\mathbf{y})$ is a solution of the KP linear problem[34,35,27,31] expressed in terms of the tau function τ^V:

$$p_n(-\tilde{\partial}_{\mathbf{x}})\Psi = \Psi \times p_{n-1}(-\tilde{\partial}_{\mathbf{x}})(\log \tau^V)_{x_1} \quad \forall n \geqslant 2 \tag{4.45}$$

{the operators $p_n(\tilde{\theta})$ are defined in terms of the Schur polynomials: $\Sigma_{n=1}^{\infty} p_n(\tilde{\theta})\lambda^n = \exp[\Sigma_{l=1}^{\infty}\theta_l / l\, \lambda^l]$}. In addition to these equations, the function $g(\mathbf{x},\mathbf{y})$ satisfies an additional identity stemming from the fact that we require the vacuum tau function τ^V itself to satisfy the constraint (3.14); the product $\tau^V g$ being a two-component tau function, the pseudo-reduction then yields

$$(\tau^V g)_{y_1} = \tau^V g_{y_1} = \tau^V_{x_k} g + \tau^V g_{x_k} \Leftrightarrow g_{y_1} = g_{x_k} + g\Omega(q^V, r^V). \tag{4.46}$$

As for the functions q^V and r^V, since the vacuum tau function τ^V is a Wronskian determinant [expressed in terms of solutions $\varphi^{*(1)}(\mathbf{x})$ of a trivial, i.e., vacuum $=1$, KP linear problem] which also has to solve the constrained hierarchy, one is forced to put

$$\varphi_j^{*(1)} = f^*_{(j-1)x_k}, \quad q^V = \frac{M_{NN}[\mathcal{W}_N^*(\varphi_j^{*(1)})]}{\tau^V(\varphi_j^{*(1)})}, \quad r^V = \frac{|\mathcal{W}_{N+1}^*(\varphi_j^{*(1)})|}{\tau^V(\varphi_j^{*(1)})} \tag{4.47}$$

for some function $f^*: f^*_{x_n} = (-1)^{n+1} f^*_{nx_1}$. Furthermore, it is easily seen that the function $g(\mathbf{x},\mathbf{y})$ can only satisfy the constraint (4.46) in all generality if the summation in its definition (4.43) is restricted to a single term such that $g(\mathbf{x},\mathbf{y}) = F(\mathbf{y}) \times q^V$. Equation (4.46) can then be written as

$$\partial_{y_1}(g) = \partial_{x_k}(g) + q^V \Omega(g, r^V), \tag{4.48}$$

allowing us to implement the pseudo-reduction on the determinants (4.44). It suffices to replace all occurrences of y_1-derivatives by the operator $\mathcal{L}_k \cdot = \partial_{x_k} \cdot + q^V \Omega(\cdot, r^V)$— note that as q^V nor r^V depend on the $\mathbf{y}$-variables, we can identify $\partial^n_{y_1}$ with $\mathcal{L}_k^n$—and at the same time guaranteeing that the resultant determinant is of (one-component) KP type. This last statement follows from comparison of the operator $\mathcal{L}_k$ with the constraint on the KP linear problem (3.24), as is explained in Ref. 14 and also in Refs. 10 and 11, where solutions of this type are presented and to which we also refer the reader for the relevant details on the operator $\mathcal{L}_k$.

At face value it might seem that we truly have constructed solutions which refer to a nontrivial vacuum. It is, however, not difficult to see that all these solutions can be rearranged such that they take the form of mere Wronskian solutions in the functions f^*. The following reasoning proves this fact: for all practical matters considered, we can look upon the above solutions as being linked in a Toda-like manner (if not directly, then at least through the construction in terms of real Toda chains), one ''end'' of which is given by the determinants

$$T_0 \equiv \tau^V = |\mathcal{W}_N^*(f^*_{(j-1)x_k})|, \tag{4.49}$$

$$T_1 \equiv \tau^V g = M_{NN}[\mathcal{W}_N^*(f^*_{(j-1)x_k})] = |\mathcal{W}_{N-1}^*(f^*_{(j-1)x_k})|. \tag{4.50}$$

However, this sequence could just as easily have been obtained by starting from the ''opposite'' end of the same two-component Toda chain, i.e., at some $\tau_{(0,M\geqslant N)}$ rather than $\tau_{(N,0)}$. The resulting sequence is then a chain of bidirectional Wronskian determinants in terms of the function $f^*: T_M = 1, T_{M-1} = f^*, f^* f^*_{x_1,x_k} - f^*_{x_1} f^*_{x_k}, \ldots,$ eventually reaching T_1 and T_0 and beyond. Thus, although the solutions (4.44) appear to be referring to nontrivial vacuum, in practice they do not. Unfortunately, it seems that in all explicitly ''constructable'' cases a similar thing will happen, i.e., the determinants will always be such that they can be reexpressed in terms of functions solving trivial dispersion relations, rather than full-fledged linear problems. This fact is intimately related with the restriction to finite Toda molecule chains if one regards them as being constructed as part of m-component hierarchies. This, in the fermionic picture, is an immediate result of being restricted to constant elements as starting points for sequences of Bäcklund transformations for the elements of GL(∞) generating the tau functions.

V. THE SCALAR CONSTRAINTS: BILINEAR DESCRIPTIONS

In this last section we discuss some peculiar properties of the scalar k-constrained KP hierarchies, with regard to their natural bilinear description. When imposing the pseudo-reduction on the linear problem for the two-component hierarchy (see Sec. III) it becomes clear that in this particular case the reduced linear equation (3.19) can be freed of all reference to the different components in essentially three different ways. All three resulting linear constraints lead to a natural (and very different) bilinear representation of the reduced nonlinear systems, for which the linear constraints then act as spectral problems.

A first way to turn the (two-component case of) equation (3.19),

$$(\Psi)_{x_k} - \lambda^k \Psi + qr\Psi_{\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1} = 0, \tag{5.1}$$

into a (purely) differential spectral problem for the constrained KP hierarchies is by rather trivially introducing the auxiliary field $\Phi = r\Psi_{\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1}$, for which we have [due to (3.22)] that

$$\Phi_x = r\Psi. \tag{5.2}$$

Note that because of (the $m=2$ case of) the constraint (3.14), the product qr is defined in terms of a single tau function,

$$qr = (\log \tau)_{x,x_k}, \tag{5.3}$$

where we from now on denote x_1 by x for simplicity. Now, from the KP linear problem as given by formula (4.45), it can be seen that any x_k-derivative $(k \geqslant 2)$ of a solution Ψ can be expressed (recursively) in terms of its x-derivatives $\Psi_{nx} (n \leqslant k)$ and derivatives of the $k-1$ functions contained in the set $\mathcal{T} = \{(\log \tau)_{x_1 x_l} | l \leqslant k-1\}$. Hence we can always rewrite the scalar constraint (5.1)

$$(\Psi)_{x_k} + q\Phi = \lambda^k \Psi \tag{5.4}$$

as a first-order spectral problem (in x) for the $(k+1)$-vector $(\psi_1 = \Psi, \psi_2 = \Psi_x, \ldots, \psi_k = \Psi_{(k-1)x}, \psi_{k+1} = \Phi)^t$, with coefficients defined in terms of the $k+1$ functions in $\mathcal{T} \cup \{q,r\}$. The higher-order $(x_{n\neq k})$ evolutions of this vector are calculated from (5.2) and the KP linear problem (4.45) and all of them will have coefficients defined in terms of these same functions $\mathcal{T} \cup \{q,r\}$. The first example (at $k=1$) is of course the well-known Ablowitz-Kaup-Newell-Segur (AKNS) spectral problem (see, e.g., Ref. 36 for a list of interesting systems obtainable at various values of k):

$$\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}_x = \begin{pmatrix} \lambda & -q \\ r & 0 \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} \tag{5.5}$$

with lowest-order evolution

$$\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}_{x_2} = \begin{pmatrix} \lambda^2 + 2qr & -q_x - \lambda q \\ r_x - \lambda r & qr \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}. \tag{5.6}$$

The nonlinear fields q and r are representable as ratios of KP tau functions (denoted here as $q = \tau^*/\tau$ and $r = \tau_*/\tau$) for which it was mentioned in Sec. II that the pairs (τ^*, τ) and (τ_*, τ) satisfy the modified KP bilinear equations. Hence, as the constraint (3.14) also has a bilinear representation in these three tau functions,

$$D_x D_{x_k} \tau \cdot \tau = 2 \tau^* \tau_*, \tag{5.7}$$

we can give a bilinear description of the nonlinear evolution equations governed by the above linear problems by imposing the bilinear constraint (5.7) on the modified KP bilinear hierarchies for (τ^*, τ) and (τ_*, τ) (see, e.g., Ref. 17 for a list of bilinear equations for the modified KP hiearchy). In this way we describe the scalar k-constrained KP hierarchies in terms of only three tau functions τ, τ^*, and τ_*. For example, from the above example we recover the well-known bilinear form of the AKNS equations:

$$\begin{aligned} q_{x_2} &= q_{2x} + 2q^2 r, \\ r_{x_2} &= -r_{2x} - 2qr^2, \end{aligned} \quad \leftrightarrow \begin{cases} D_x^2 \tau \cdot \tau = 2 \tau^* \tau_*, \\ (D_{x_2} - D_x^2) \tau^* \cdot \tau = 0, \\ (D_{x_2} + D_x^2) \tau_* \cdot \tau = 0. \end{cases} \tag{5.8}$$

A second example (omitting the relevant linear problem) is that of $k=2$ which—up to a gauge and coordinate transformation—can be identified as the Yajima–Oikawa system[37] $[w = (\log \tau)_{2x}]$:

$$\begin{aligned} q_{x_2} &= q_{2x} + 2wq, \\ r_{x_2} &= -(r_{2x} + 2wr) \\ w_{x_2} &= (qr)_x, \end{aligned} \quad \leftrightarrow \begin{cases} D_x D_{x_2} \tau \cdot \tau = 2 \tau^* \tau_*, \\ (D_{x_2} - D_x^2) \tau^* \cdot \tau = 0, \\ (D_{x_2} + D_x^2) \tau_* \cdot \tau = 0. \end{cases} \tag{5.9}$$

Another way of obtaining differential spectral problems from (5.1) is by introducing a slightly different auxiliary field $\Phi = qr \Psi_{\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1}$, since in this case formula (3.22) tells us that

$$\Phi_x - (\log q)_x \Phi = qr \Psi. \tag{5.10}$$

This again allows us to rewrite the linear constraint (5.1) as a first-order spectral problem (in x) for the $(k+1)$-vector $(\psi_1 = \Psi, \psi_2 = \Psi_x, \ldots, \psi_k = \Psi_{(k-1)x}, \psi_{k+1} = \Phi)^t$, but this time with coefficients defined in terms of the $k+1$ functions in $\mathcal{T} \cup \{h \equiv \log q, (\log \tau)_{x, x_k} = qr\}$. As we now eliminated the explicit dependence on the r field (and hence on the τ_* tau function), we should look for a bilinear description of the resulting nonlinear systems in terms of only two tau functions: τ^* and τ. The answer lies with the x_k derivative of τ. In Sec. III it is explained that the constraint (3.14) actually tells us that the x_k derivative of τ is not only again a tau function (say, $\hat{\tau}$), but a very special one indeed, as it can be expressed using an eigenfunction potential. The element of $\mathrm{GL}(\infty)$ describing $\hat{\tau}$ is obtained from that for τ (let's call this one g) by a mere Bäcklund transformation on g:

$$\tau_{x_k} = \langle \mathrm{vac} | \hat{g}(x) | \mathrm{vac} \rangle \quad \text{with} \quad \hat{g} = (c + \phi \phi^*) g. \tag{5.11}$$

The tau function τ^* in its turn is linked to τ (and therefore to g) by the Darboux transformation,

$$\tilde{g} = S^{-1} \phi g \Rightarrow \tau^* \equiv \langle 1 | \Phi(\mathbf{x}) g(\mathbf{x}) | \mathrm{vac} \rangle = q \times \tau, \tag{5.12}$$

which is only a different way of saying that the pair (τ^*,τ) satisfies the modified KP bilinear equations (i.e., that q is a solution to the KP linear equations). However, since the (generalized) fermion operators are nilpotent, we immedately have that

$$S^{-1}\phi\hat{g}=S^{-1}\phi(c+\phi\phi^*)g=cS^{-1}\phi g\sim\tilde{g}, \tag{5.13}$$

and thus we see that the pair (τ^*,τ_{x_k}) also satisfies the modified KP bilinear equations. In other words, one can also give a bilinear description of the scalar k-constrained KP hierarchy by replacing the bilinear constraint (5.7) by an expression which states that both pairs (τ^*,τ) and (τ^*,τ_{x_k}) satisfy the modified KP bilinear hierarchies, i.e., using only two tau functions. The corresponding nonlinear equations are then expressible in the $k+1$ fields $\mathcal{T}\cup\{h,u\}$ with $h=\log\tau^*/\tau$ and $u\equiv(\log\tau)_{x,x_k}=qr$. These are, of course, related to the previous ones by simple dependent variable transformations, their linear problems being related by gauge transformations (as is clear from the construction). The most famous occurrence of this fact is that for the Broer–Kaup system,[38] which is obtained for the basic case $k=1$. As was described above, its spectral problem is

$$\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}_x=\begin{pmatrix}\lambda & -1\\ u & h_x\end{pmatrix}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}, \tag{5.14}$$

with lowest-order evolution

$$\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}_{x_2}=\begin{pmatrix}\lambda^2+u & -h_x-\lambda\\ \lambda u+h_xu-u_x & -h_{2x}-h_x^2-u\end{pmatrix}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}. \tag{5.15}$$

The compatibility condition of this system (Broer–Kaup) can be bilinearized as

$$\begin{array}{l} h_{x_2}=h_{2x}+h_x^2+2u,\\ u_{x_2}=(2h_xu-u_x)_x,\end{array}\leftrightarrow\begin{cases}(D_{x_2}-D_x^2)\tau^*\cdot\tau=0,\\ (D_{x_2}-D_x^2)\tau^*\cdot\tau_x=0.\end{cases} \tag{5.16}$$

Remark that for KP tau functions τ this last bilinear equation can be rewritten as $(D_{x_2}-D_x^2)\tau^*\cdot\tau_x=0\Leftrightarrow D_x(D_{x_2}-D_x^2)\tau^*\cdot\tau=0$, a property which extends to all the bilinear equations in the modified KP hierarchy. The correspondence between the AKNS hierarchy and the Broer–Kaup hierarchy has been discussed by a number of authors in the past;[19,39,35] the present construction has the nice feature of allowing one to dispense altogether with gauge transformations for ''energy dependent'' eigenvalue problems[40] or constraints for modified KP hierarchies[36] in order to appreciate the fact that both hierarchies describe exactly the same systems. It should be clear from the construction that such gauge transformations are a general feature of the (scalar) k-constrained KP hierarchies. In particular, for the Yajima–Oikawa system, one finds the following ''partner'' system,

$$\begin{array}{l} h_{x_2}=h_{2x}+h_x^2+2w,\\ w_{x_2}=u_x,\\ u_{x_2}=(2h_xu-u_x)_x,\end{array}\leftrightarrow\begin{cases}(D_{x_2}-D_x^2)\tau^*\cdot\tau=0,\\ D_{x_2}(D_{x_2}-D_x^2)\tau^*\cdot\tau=0'\end{cases} \tag{5.17}$$

providing an alternative bilinearization of the Yajima–Oikawa equations using only two tau functions [$h=\log\tau^*/\tau$, $w=(\log\tau)_{2x}$ and $u=(\log\tau)_{x,x_2}$].

Finally, there exists a third bilinear description of the k-constrained KP hierarchy, one in which only a single tau function plays a role. Using property (3.23), we immediately have the spectral equation

$$\left(\partial_x-\frac{q_x}{q}\right)[\Psi_{x_k}-\lambda^k\Psi]+qr\Psi=0. \tag{5.18}$$

Performing now the pseudo-reduction (3.3) on the (two-component version of the) higher-order evolution (2.15), we obtain

$$(\log\tau)_{x_2,x_k}=\tau^{-2}(\tau_x^*\tau_*-\tau^*\tau_{*x})=q_xr-qr_x, \tag{5.19}$$

which, together with (5.3), yields the following expression for h:

$$h\equiv\frac{q_x}{q}=\frac{(\log\tau)_{x_2,x_k}+(\log\tau)_{2x,x_k}}{2(\log\tau)_{x,x_k}}. \tag{5.20}$$

Accordingly, we can reexpress the spectral problem as

$$(\partial_x-h)[\Psi_{x_k}-\lambda^k\Psi]+u\Psi=0, \tag{5.21}$$

but where $u=(\log\tau)_{x,x_k}$ and h are defined in terms of only one tau function τ. The KP linear problem (4.45) allows us to rewrite this spectral problem in terms of a (scalar) $(k+1)$st-order differential operator in x, with coefficients defined in terms of the $k+1$ functions $\{(\log\tau)_{x_1,x_l}|l\leqslant k\}\cup\{(\log\tau)_{x_2,x_k}\}$. The compatibility conditions of this operator and the usual evolution operators for the KP linear problem yield nonlinear equations which can be bilinearized using a single tau function. The corresponding bilinear forms are readily obtained by realizing that the k-constraint imposes the requirement that x_k-derivatives of tau functions should again be KP tau functions, in other words, that not only τ should satisfy the KP bilinear equations but τ_{x_k} as well. Of course all these systems (at every value of k) are gauge equivalent with the ones we described earlier. For example, in the case $k=1$ the condition discussed above yields (at lowest order) the system of bilinear equations,

$$\begin{gathered}(4D_xD_{x_3}-D_x^4-3D_{x_2}^2)\tau\cdot\tau=0,\\ (4D_xD_{x_3}-D_x^4-3D_{x_2}^2)\tau_x\cdot\tau_x=0,\end{gathered} \tag{5.22}$$

the last of which can be recombined with the first one so as to give

$$D_x^2(4D_xD_{x_3}-D_x^4-3D_{x_2}^2)\tau\cdot\tau=0 \tag{5.23}$$

provided τ is a KP tau function. When expressed in the fields $(\log\tau)_{2x}$ and $(\log\tau)_{x,x_2}$, this bilinear system yields the first two members of what was identified as a *nonlocal Boussinesq* hierarchy,[41–43] which first appeared as an equation governing the amplitude (squared) of the nonlinear Schrödinger waves.[44] Obviously, this hierarchy is gauge equivalent to the AKNS system given earlier. The ''one-field'' bilinearization of the k-constrained KP hierarchies was originally proposed in Ref. 4 and was actually first discovered for the example of the nonlocal-Boussinesq system.[42]

As a final example we give (at $k=2$) the following single field bilinearization of the Yajima–Oikawa system:

$$\begin{gathered}(4D_xD_{x_3}-D_x^4-3D_{x_2}^2)\tau\cdot\tau=0,\\ D_{x_2}^2(4D_xD_{x_3}-D_x^4-3D_{x_2}^2)\tau\cdot\tau=0,\end{gathered} \tag{5.24}$$

VI. CONCLUSIONS

We feel the description of constrained KP hiearchies offers a number of appealing features, notably the ease with which the tau function interpretation of the constraints can be obtained. As an important consequence of this interpretation we explained how the scalar *k*-constrained KP hierarchy possesses three different bilinear representations, each corresponding in a natural way to different nonlinearizations whose mutual connections were always regarded as somewhat ''special.'' Here we showed these are an intrinsic feature of constrained hierarchies.

We also paid a lot of attention to the explicit construction of solutions. In particular, two types of solutions, Wronskians and hybrid Wronskian/Grammians, were constructed. We would like to point out that previous accounts on Grammian-type solutions somehow disregarded the nature of the ''arbitrary'' constants which appear in the entries of the Grammian determinants. As we showed in Sec. IV, these constants are not quite arbitrary, but exhibit a structure similar to that of the actual Grammian entries. As for the Wronskian solutions (4.12), it should be clear that these allow for a slight generalization in that one can add a number (say M) of ''*k*-reduced''-like Wronskian blocks. These would contain rows expressed using *k*-reduced solutions to the KP linear problem, meaning functions $f_i(\mathbf{x})$ $(i=1,...,M)$ which—besides the dispersion relations (4.8)—have to satisfy $(f_i)_{x_k}=c_i f_i$ for some constants c_i. Such extended Wronskians clearly solve the constraint (3.14) as they can be viewed as part of an $(m-1+M)$-vector *k*-constrained hierarchy, but one for which the functions q_j corresponding to these extra blocks are always identically zero (Wronskian determinants, solely expressed in terms of such functions, will solve the *k*-reduced KP hierarchy). It is interesting to note that very similar solutions arise when one adopts a vertex operator approach to constrained KP evolutions, as described in Ref. 12. It seems worthwhile to try and understand why exactly these particular solutions should appear in the vertex operator formalism.

Finally, there remains the (general) problem of whether or not one can construct examples of determinant solutions whose entries genuinely satisfy nontrivial KP linear problems (as opposed to solutions as those discussed in the latter part of Sec. IV). As we explained, this problem can be traced back to the choice of starting point in GL(∞) if one thinks of the determinants as being constructed through Bäcklund transformations applied to such a vacuum element (cf. Sec. II).

ACKNOWLEDGMENTS

Both authors are postdoctoral fellows at the Fund for Scientific Research (F.W.O.), Flanders (Belgium). One of the authors (RW) is the benificiary of a JSPS postdoctoral fellowship for foreign researchers in Japan and as such the present work was also partially supported by a Grant-in-Aid from the Japan Ministry of Education, Science and Culture. IL wishes to thank Professor J. Satsuma for the hospitality and support during his stay at the University of Tokyo at the occasion of which the present work was initiated. The authors also wish to acknowledge financial support extended within the framework of the ''Interuniversitary Poles of Attraction Programme, Contract nr P4/08—Belgian State.''

[1] A. Y. Orlov, ''Volterra operator algebra for zero curvature representation. Universality of KP,'' in *Nonlinear Processes in Physics* edited by A. S. Fokas, D. J. Kaup, A. C. Newell, and V. E. Zakharov (Springer Verlag, Berlin, 1991), pp. 126–131.
[2] Y. Cheng and Y. S. Li, ''The constraint of the Kadomtsev-Petviashvili hierarchy and its special solutions,'' Phys. Lett. A **157**, 22–26 (1991).
[3] Y. Cheng, ''Constraints of the Kadomtsev–Petviashvili hierarchy,'' J. Math. Phys. **33**, 3774–3782 (1992).
[4] I. Loris and R. Willox, ''Bilinear form and solutions of the *k*-constrained Kadomtsev-Petviasvili hierarchy,'' Inverse Probl. **13**, 411–420 (1997).
[5] Y. Cheng and Y. L. Zhang, ''Bilinear equations for the constrained KP hierarchy,'' Inverse Probl. **10**, L11–L17 (1994).
[6] G. F. Helminck and J. W. van de Leur, ''An analytic description of the vector constrained KP hierarchy,'' Commun. Math. Phys. **193**, 627–641 (1998).
[7] I. Loris and R. Willox, ''KP symmetry reductions and a generalized constraint,'' J. Phys. A **30**, 6925–6938 (1997).
[8] I. Loris and R. Willox, ''On solutions of constrained KP equations,'' J. Math. Phys. **38**, 283–291 (1997).
[9] I. Loris and R. Willox, ''On solutions of constrained KP equations: Grammians,'' J. Math. Phys. **38**, 5190–5197 (1997).

[10] L.-L. Chau, J.-C. Shaw, and M.-H. Tu, ''Solving the constrained KP hierarchy by gauge transformations,'' J. Math. Phys. **38**, 4128–4137 (1997).
[11] H. Aratyn, E. Nissimov, and S. Pacheva, ''Constrained KP hierarchies: Additional symmetries, Darboux-Bäcklund solutions and relations to multi-matrix models,'' Int. J. Mod. Phys. A **12**, 1265–1340 (1997).
[12] H. Aratyn, L. A. Ferreira, J. F. Gomes, and A. H. Zimmerman, ''Vertex Operators and Solitons of Constrained KP Hierarchies,'' in Supersymmetry and Integrable Models, edited by H. Aratyn, T. D. Imbo, W.-Y. Keung, and U. Sukhatme, Lecture Notes in Physics (Springer, Berlin, 1998), pp. 197–209.
[13] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons* (Springer Verlag, Berlin, 1991).
[14] R. Willox, I. Loris, and C. R. Gilson, ''Binary Darboux transformations for constrained KP hierarchies,'' Inverse Probl. **13**, 849–865 (1997).
[15] V. G. Kac and J. van de Leur, ''The *n*-component KP hierarchy and representation theory,'' in Important Developments in Soliton Theory, edited by A. S. Fokas and V. E. Zakharov (Springer, Berlin, 1993), pp. 302–343.
[16] W. Oevel and W. Strampp, ''Wronskian solutions of the constrained Kadomtsev–Petviashvili hierarchy,'' J. Math. Phys. **37**, 6213–6219 (1996).
[17] M. Jimbo and T. Miwa, ''Solitons and infinite dimensional Lie algebras,'' Publ. RIMS Kyoto Univ. **19**, 943–1001 (1983).
[18] T. Miwa, M. Jimbo, and E. Date, *Soliton no suri* (Iwanami, Tokyo, 1993) (in Japanese).
[19] R. Hirota, ''Classical Boussinesq equation is a reduction of the modified KP equation,'' J. Phys. Soc. Jpn. **54**, 2409–2415 (1985).
[20] R. Willox and I. Loris, ''An algebraic description of generalized *k*-constraints,'' J. Phys. A **32**, 2027–2036 (1999).
[21] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, ''Operator approach to the Kadomtsev–Petviashvili equation. Transformation groups for soliton equations III,'' J. Phys. Soc. Jpn. **50**, 3806–3812 (1981).
[22] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, ''Transformation groups for soliton equations. Euclidean Lie algebras and reductions of the KP hierarchy,'' Publ. RIMS Kyoto Univ. **18**, 1077–1110 (1982).
[23] R. Hirota, Y. Ohta, and J. Satsuma, ''Solutions of the Kadomtsev-Petviashvili Equation and the Two-dimensional Toda equation,'' J. Phys. Soc. Jpn. **57**, 1901–1904 (1988).
[24] R. Hirota, Y. Ohta, and J. Satsuma, ''Wronskian structure of solutions for soliton equations,'' Prog. Theor. Phys. Suppl. **94**, 59–72 (1988).
[25] R. Hirota, ''Direct methods in soliton theory,'' in Solitons, edited by R. K. Bullough and P. J. Caudrey (Springer Verlag, Berlin, 1980), pp. 157–176.
[26] A. V. Mikhailov, ''Integrability of a two-dimensional generalization of the Toda-chain,'' JETP Lett. **30**, 414–418 (1979).
[27] R. Willox, T. Tokihiro, I. Loris, and J. Satsuma, ''Fermionic approach to Darboux transformations,'' Inverse Probl. **14**, 745–762 (1998).
[28] W. Oevel, ''Darboux theorems and Wronskian formulas,'' Physica A **195**, 533–576 (1993).
[29] J. van de Leur, ''Schlesinger-Bäcklund transformations for the *N-component KP,'' J. Math. Phys.* **39**, 2833–2847 (1998).
[30] The case $k\geqslant 1$ appears in the 2nd edition of Ref. 15, unknown to the authors at the time.
[31] H. Aratyn, E. Nissimov, and S. Pacheva, ''Method of squared eigenfunction potentials in integrable hierarchies of KP type,'' Commun. Math. Phys. **193**, 493–525 (1998).
[32] A. N. Leznov and M. V. Saveliev, ''Theory of group representations and integration of nonlinear systems $x_{a,z\bar{z}} = \exp(kx)_a$,'' Physica D **3**, 62–72 (1981).
[33] H. Aratyn, E. Nissimov, and S. Pacheva, ''Virasoro symmetries of constrained KP hierarchies,'' Phys. Lett. A **228**, 164–175 (1997).
[34] M. Sato and Y. Sato, ''Soliton equations as dynamical systems on infinite dimensional Grassmanian manifolds,'' in ''Nonlinear PDE in Applied Science, US–Japan Seminar (Tokyo),'' Lect. Notes in Num. Appl. Anal. **5**, 259–271 (1982).
[35] Y. Cheng, W. Strampp, and B. Zhang, ''Constraints of the KP hierarchy and multilinear forms,'' Commun. Math. Phys. **168**, 117–135 (1995).
[36] B. Konopelchenko and W. Strampp, ''New reductions of the Kadomtsev–Petviashvili and two-dimensional Toda lattice hierarchies via symmetry constraints,'' J. Math. Phys. **33**, 3676–3686 (1992).
[37] N. Yajima and M. Oikawa, ''Formation and interaction of sonic-Langmuir solitons—inverse scattering method,'' Prog. Theor. Phys. **56**, 1719–1739 (1976).
[38] L. J. F. Broer, ''Approximate solutions for long water waves,'' Appl. Sci. Res. **31**, 377–395 (1975).
[39] R. L. Sachs, ''On the integrable variant of the Boussinesq system: Painlevé property, rational solutions, a related many-body system, and equivalence with the AKNS hierarchy,'' Physica D **30**, 1–27 (1988).
[40] M. Jaulent and I. Miodek, ''Connection between Zakharov-Shabat and Schrödinger-type inverse-scattering transformations,'' Lett. Nuovo Cimento **20**, 655–660 (1977).
[41] F. Lambert, I. Loris, J. Springael, and R. Willox, ''On a direct bilinearization method: Kaup's higher order water wave equation as a modified nonlocal Boussinesq equation,'' J. Phys. A **27**, 5325–5334 (1994).
[42] R. Willox and I. Loris, ''Bilinearization of the non-local Boussinesq equation,'' J. Phys. A **28**, 5963–5972 (1995).
[43] I. Loris and R. Willox, ''Soliton solutions of Wronskian form to the nonlocal Boussinesq equation,'' J. Phys. Soc. Jpn. **65**, 1222–1226 (1996).
[44] M. Boiti, C. Laddomada, and F. Pempinelli, ''An equivalent real form of the non-linear Schrödinger equation and the permuability for the Bäcklund transformation,'' Nuovo Cimento B **62**, 315–326 (1981).

Separation of variables for soliton equations via their binary constrained flows

Yunbo Zeng[a)]
Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China

Wen-Xiu Ma[b)]
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, People's Republic of China

(Received 2 June 1999; accepted for publication 7 September 1999)

Binary constrained flows of soliton equations admitting 2×2 Lax matrices have $2N$ degrees of freedom, which is twice as many degrees of freedom than in the case of monoconstrained flows. By using the normal method, their Lax matrices directly give rise to first N pairs of canonical separated variables for their separation of variables. We propose a new method to introduce the other N pairs of canonical separated variables and additional separated equations. The Jacobi inversion problems for binary constrained flows are established. Finally, the factorization of soliton equations by two commuting binary constrained flows and the separability of binary constrained flows enable us to construct the Jacobi inversion problems for some soliton hierarchies.
[S0022-2488(99)02112-X]

I. INTRODUCTION

The separation of variables is one of the most universal methods for solving completely integrable (classical and quantum) models. It has been applied successfully to the study of a large number of finite-dimensional integrable Hamiltonian systems (FDIHSs) (see, e.g., Refs. 1–12), as well as infinite-dimensional integrable Hamiltonian systems in the determination of finite-dimensional quasiperiodic solutions (see, e.g., Refs. 13–18). In many cases the separation of variables of integrable classical systems prepares the passage to the corresponding quantum systems. For the classical integrable systems subject to the inverse scattering method, the standard construction of the action-angle variables using the poles of the Baker–Akhiezer function is in fact equivalent to the separation of variables.[4]

For a FDIHS, let m denote the number of degrees of freedom, and P_i, $i=1,...,m$, be functionally independent integrals of motion in involution, the separation of variables means to construct m pairs of canonical separated variables v_k, u_k, $k=1,...,m$,[2–4]

$$\{u_k,u_l\}=\{v_k,v_l\}=0, \quad \{v_k,u_l\}=\delta_{kl}, \quad k,l=1,...,m, \tag{1.1}$$

and m functions f_k such that

$$f_k(u_k,v_k,P_1,\dots,P_m)=0, \quad k=1,...,m. \tag{1.2}$$

Equation (1.2) is called the separated equation, which gives rise to an explicit factorization of the Liouville tori.

For the FDIHSs with the Lax matrices admitting the r-matrices of the *XXX*, *XXZ* and *XYZ* type, there is a general approach to introduce canonical separated variables.[2–4,8] The correspond-

a)Electronic mail: yzeng@tsinghua.edu.cn
b)Electronic mail: mawx@cityu.edu.hk

ing separated equations enable us to express the generating function of canonical transformation in completely separated form as an Abelian integral on the associated invariant spectral curve. The resulted linearizing map is essentially the Abel map to the Jacobi variety of the spectral curve, thus providing a link, through purely Hamiltonian methods, with the algebro-geometric linearization methods given by Refs. 19–22.

An important feature of the separation of variables for a FDIHS is that the number of canonical separated variables u_k should be equal to the number m of degrees of freedom. In some cases, the number of u_k resulting from the normal method may be less than m and so some additional canonical separated variables should be introduced. So far very few models in these cases have been studied. These cases remain a challenging problem.[4] In recent years binary constrained flows of soliton hierarchies have attracted attention (see, e.g., Refs. 23–29), whose basic idea was described in Ref. 30. The degree of freedom for binary constrained flows admitting 2×2 Lax matrices is an even natural number usually denoted by $2N$. By using the method,[2–4,8] the Lax matrices allow one to directly introduce first N pairs of canonical separated variables $u_1,\ldots,u_N$ and $v_1,\ldots,v_N$. The other N pairs of canonical separated variables $u_{N+1},\ldots,u_{2N}$ and $v_{N+1},\ldots,v_{2N}$ may be constructed by the method in Ref. 3. In this paper we propose a new method for determining additional N pairs of canonical separated variables and separated equations for binary constrained flows. The main idea is to construct two functions $\tilde{B}(\lambda)$ and $\tilde{A}(\lambda)$ defining $u_{N+1},\ldots,u_{2N}$ by the set of zeros of $\tilde{B}(\lambda)$ and $v_{N+k}=\tilde{A}(u_{N+k})$. To keep the canonical conditions (1.1) and the requirement for the separated equation (1.2), it is found that certain commutator relations should be imposed on $\tilde{B}(\lambda)$, $\tilde{A}(\lambda)$ and $\tilde{A}(\lambda)$ has some link with the generating function of integrals of motion of binary constrained flows, which provides a way to construct the $\tilde{B}(\lambda)$ and $\tilde{A}(\lambda)$. In fact, we have to modify the original approach for introducing $u_1,\ldots,u_N$ and $v_1,\ldots,v_N$ so that $u_1,\ldots,u_{2N}$ and $v_1,\ldots,v_{2N}$ are canonical conjugated. Having produced the separation of variables, we further construct the Jacobi inversion problems for binary constrained flows. This method is somewhat different from that for introducing canonical variables presented in Ref. 31 and can be applied to more binary constrained flows.

Briefly, separation of variables can be characterized as a reduction of a multidimensional problem to a set of one-dimensional ones. The separation of variables of soliton equations in this paper contains two steps of separation of variables. The first step is to factorize $(1+1)$-dimensional soliton equations into two commuting x- and t-FDIHSs via binary constrained flows, namely the x and t dependencies of the soliton equations are separated by the x- and t-FDIHSs obtained from the x- and t-binary constrained flows. The second step is to produce separation of variables for the x- and t-FDIHDs by our method to be proposed later on. Finally, combining the factorization of soliton equations with the Jacobi inversion problems for x- and t-FDIHSs enables us to establish the Jacobi inversion problems for soliton equations. We will present the separation of variables for the Korteweg–de Vries (KdV) hierarchy, the AKNS hierarchy, and the Kaup–Newell hierarchy via their binary constrained flows. In fact, we employ our method in a little different way for those three cases.

In Sec. II, we recall the binary constrained flows of KdV hierarchy and present factorization of the KdV equations into the x- and t-binary constrained flows. By means of the Lax matrix $M=\begin{pmatrix} A(\lambda) & B(\lambda) \\ C(\lambda) & -A(\lambda) \end{pmatrix}$ for the binary constrained flows, the method in Refs. 2–4 and 8 introduce the first N pairs of canonical variables $u_1,\ldots,u_N$ by the set of zeros of $B(\lambda)$ and $v_k=2A(u_k)$. We propose a new method to construct two new functions $\tilde{B}(\lambda)$ and $\tilde{A}(\lambda)$ for introducing $u_{N+1},\ldots,u_{2N}$ by the set of zeros of $\tilde{B}(\lambda)$ and $v_{N+k}=\tilde{A}(u_{N+k})$. The construction of $\tilde{B}(\lambda)$ and $\tilde{A}(\lambda)$ is based on an observation that the canonical conditions (1.1) need certain commutator relations between $\tilde{A}(\lambda)$, $\tilde{B}(\lambda)$, and the requirement for the separated equation (1.2) links $\tilde{A}(\lambda)$ with another generating function of integrals of motion. To guarantee that $v_1,\ldots,v_{2N}$ and $u_1,\ldots,u_{2N}$ are canonical conjugated, we also have to modify the original way for introducing $u_1,\ldots,u_N$ and $v_1,\ldots,v_N$. Then we establish the Jacobi inversion problems for the x- and t-binary constrained flows. Finally, these Jacobi inversion problems together with the factorization of the KdV equations give rise to the

Jacobi inversion problems for the KdV equations. In Sec. III, the factorization of the Ablowitz-Kaup-Newell-Segur (AKNS) equations is given. Since $B(\lambda)$ for the binary constrained AKNS flows, unlike the $B(\lambda)$ for the binary constrained KdV flows, has only $N-1$ zeros, we have to modify the method proposed in Sec. II in order to find $2N$ pairs of canonical variables for the binary constrained AKNS flows. In Sec. IV, we present the factorization of the Kaup–Newell equations. Since the commutator relations of $A(\lambda)$, $B(\lambda)$, and $C(\lambda)$ for the binary constrained Kaup–Newell flows are quite different from those for both the binary constrained KdV flows and the binary constrained AKNS flows, we need to further modify the method in Secs. II and III in order to find the separation of variables for the Kaup–Newell equations. Finally some remarks are made in Sec. V.

II. SEPARATION OF VARIABLES FOR THE KdV EQUATIONS

In this section, we use the binary constrained flows of KdV hierarchy to illustrate our method of introducing canonical separated variables. Then we show how to produce the separation of variables for the KdV equations. To make the paper self-contained, we first briefly describe the binary constrained flows of the KdV hierarchy.[26]

A. Binary constrained flows of the KdV hierarchy

Let us start from the Schrödinger equation[32]

$$\phi_{xx}+(\lambda+u)\phi=0,$$

which can be rewritten as the following spectral problem:

$$\phi_x=U(u,\lambda)\phi, \quad U(u,\lambda)=\begin{pmatrix} 0 & 1 \\ -\lambda-u & 0 \end{pmatrix}, \quad \phi=\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}. \tag{2.1}$$

Its adjoint representation reads

$$V_x=[U,V]\equiv UV-VU. \tag{2.2}$$

Set

$$V=\sum_{i=0}^{\infty}\begin{pmatrix} a_i & b_i \\ c_i & -a_i \end{pmatrix}\lambda^{-i}. \tag{2.3}$$

Equation (2.2) yields

$$a_0=b_0=0, \quad c_0=-1, \quad a_1=0, \quad b_1=1, \quad c_1=-\tfrac{1}{2}u,$$

$$a_2=\tfrac{1}{4}u_x, \quad b_2=-\tfrac{1}{2}u, \quad c_2=\tfrac{1}{8}(u_{xx}+u^2),\ldots,$$

and in general

$$b_{k+1}=Lb_k=-\tfrac{1}{2}L^{k-1}u, \quad a_k=-\tfrac{1}{2}b_{k,x}, \tag{2.4a}$$

$$c_k=-\tfrac{1}{2}b_{k,xx}-b_{k+1}-b_k u, \quad k=1,2,..., \tag{2.4b}$$

where

$$L=-\tfrac{1}{4}\partial^2-u+\tfrac{1}{2}\partial^{-1}u_x, \quad \partial=\partial_x, \quad \partial^{-1}\partial=\partial\partial^{-1}=1.$$

Set

$$V^{(n)}(u,\lambda)=\sum_{i=0}^{n+1}\begin{pmatrix} a_i & b_i \\ c_i & -a_i \end{pmatrix}\lambda^{n+1-i}+\begin{pmatrix} 0 & 0 \\ b_{n+2} & 0 \end{pmatrix}, \tag{2.5}$$

and take the time evolution law of ϕ as

$$\phi_{t_n}=V^{(n)}(u,\lambda)\phi. \tag{2.6}$$

Then the compatibility condition of Eqs. (2.1) and (2.6) gives rise to the nth KdV equation which can be written as the infinite-dimensional Hamiltonian system

$$u_{t_n}=-2b_{n+2,x}=\partial L^n u=\partial\frac{\delta H_n}{\delta u}, \tag{2.7}$$

where the Hamiltonian H_n is given by

$$H_n=\frac{4b_{n+3}}{2n+3},\quad \frac{\delta H_n}{\delta u}=-2b_{n+2}.$$

The matrix V determined by (2.2) and (2.3) also satisfies the adjoint representation of (2.6),

$$V_{t_n}=[V^{(n)},V], \tag{2.8}$$

when u satisfies (2.7).

For $n=1$ we have

$$\phi_{t_1}=V^{(1)}(u,\lambda)\phi,\quad V^{(1)}=\begin{pmatrix}\frac{1}{4}u_x & \lambda-\frac{1}{2}u\\ -\lambda^2-\frac{1}{2}u\lambda+\frac{1}{4}u_{xx}+\frac{1}{2}u^2 & -\frac{1}{4}u_x\end{pmatrix}, \tag{2.9}$$

and Eq. (2.7) for $n=1$ is the well-known KdV equation

$$u_{t_1}=-\tfrac{1}{4}(u_{xxx}+6uu_x). \tag{2.10}$$

The adjoint spectral problem reads

$$\psi_x=-U^T(u,\lambda)\psi,\quad \psi\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}. \tag{2.11}$$

We have[26]

$$\frac{\delta\lambda}{\delta u}=\beta\,\mathrm{Tr}\left[\begin{pmatrix}\phi_1\psi_1 & \phi_1\psi_2\\ \phi_2\psi_1 & \phi_2\psi_2\end{pmatrix}\frac{\partial U(u,\lambda)}{\partial u}\right]=-\beta\psi_2\phi_1, \tag{2.12}$$

where β is some constant.

The binary x-constrained flows of the KdV hierarchy (2.7) consist of the equations obtained from the spectral problem (2.1) and the adjoint spectral problem (2.11) for N distinct real numbers λ_j and the restriction of the variational derivatives for the conserved quantities H_{k_0} (for any fixed k_0) and λ_j:

$$\Phi_{1,x}=\Phi_2,\quad \Phi_{2,x}=-\Lambda\Phi_1-u\Phi_1, \tag{2.13a}$$

$$\Psi_{1,x}=\Lambda\Psi_2+u\Psi_2,\quad \Psi_{2,x}=-\Psi_1, \tag{2.13b}$$

$$\frac{\delta H_{k_0}}{\delta u}-\beta^{-1}\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta u}=-2b_{k_0+2}+\langle\Psi_2,\Phi_1\rangle=0. \tag{2.13c}$$

Such a constraint (2.13c) has been recognized as a symmetry constraint.[25,26,30] Hereafter we denote the inner product in $\mathbf{R}^N$ by $\langle .,.\rangle$ and

$$\Phi_i=(\phi_{i1},\ldots,\phi_{iN})^T,\quad \Psi_i=(\psi_{i1},\ldots,\psi_{iN})^T,\quad i=1,2,\quad \Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_N).$$

For $k_0=0$, we have

$$b_2=-\tfrac{1}{2}u=\tfrac{1}{2}\langle\Psi_2,\Phi_1\rangle,$$

i.e.,

$$u=-\langle\Psi_2,\Phi_1\rangle. \tag{2.14}$$

By substituting (2.14) into (2.13a) and (2.13b), the first binary x-constrained flow becomes a finite-dimensional Hamiltonian system (FDHS)[26]

$$\Phi_{1x}=\frac{\partial F_1}{\partial\Psi_1},\quad \Phi_{2x}=\frac{\partial F_1}{\partial\Psi_2},\quad \Psi_{1x}=-\frac{\partial F_1}{\partial\Phi_1},\quad \Psi_{2x}=-\frac{\partial F_1}{\partial\Phi_2}, \tag{2.15}$$

with the Hamiltonian

$$F_1=\langle\Psi_1,\Phi_2\rangle-\langle\Lambda\Psi_2,\Phi_1\rangle+\tfrac{1}{2}\langle\Psi_2,\Phi_1\rangle^2.$$

The binary t_n-constrained flows of the KdV hierarchy (2.7) are defined by the replicas of (2.6) and its adjoint system for N distinct real number λ_j,

$$\begin{pmatrix}\phi_{1j}\\ \phi_{2j}\end{pmatrix}_{t_n}=V^{(n)}(u,\lambda_j)\begin{pmatrix}\phi_{1j}\\ \phi_{2j}\end{pmatrix},\quad \begin{pmatrix}\psi_{1j}\\ \psi_{2j}\end{pmatrix}_{t_n}=-(V^{(n)}(u,\lambda_j))^T\begin{pmatrix}\psi_{1j}\\ \psi_{2j}\end{pmatrix},\quad j=1,\ldots,N, \tag{2.16}$$

as well as the nth KdV equation itself (2.7) in the case of the higher-order constraint for $k_0\geqslant 1$. Under the constraint (2.14) and the x-FDHS (2.15), the binary t_1-constrained flow obtained from (2.16) with $V^{(1)}$ given by (2.9) can also be written as a t_1-FDHS,

$$\Phi_{1,t_1}=\frac{\partial F_2}{\partial\Psi_1},\quad \Phi_{2,t_1}=-\frac{\partial F_2}{\partial\Psi_2},\quad \Psi_{1,t_1}=-\frac{\partial F_2}{\partial\Phi_1},\quad \Psi_{2,t_1}=-\frac{\partial F_2}{\partial\Psi_2}, \tag{2.17}$$

with the Hamiltonian

$$F_2=-\langle\Lambda^2\Psi_2,\Phi_1\rangle+\langle\Lambda\Psi_1,\Phi_2\rangle+\tfrac{1}{2}\langle\Psi_2,\Phi_1\rangle\langle\Lambda\Psi_2,\Phi_1\rangle+\tfrac{1}{2}\langle\Psi_2,\Phi_1\rangle\langle\Psi_1,\Phi_2\rangle+\tfrac{1}{8}(\langle\Psi_2,\Phi_2\rangle$$
$$-\langle\Psi_1,\Phi_1\rangle)^2.$$

The Lax representation for the x-FDHS (2.15) and the t_1-FDHS (2.17) can be deduced from the adjoint representation (2.2) and (2.8) by using the method in[33,34]

$$M_x=[\widetilde{U},M],\quad M_{t_n}=[\widetilde{V}^{(n)},M], \tag{2.18}$$

where $\widetilde{U}$ and $\widetilde{V}^{(n)}$ are obtained from U and $V^{(n)}$ by a substitution of (2.14), and the Lax matrix M is given by

$$M=\begin{pmatrix}A(\lambda) & B(\lambda)\\ C(\lambda) & -A(\lambda)\end{pmatrix},$$

$$A(\lambda)=\frac{1}{4}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j}}{\lambda-\lambda_j},\quad B(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{1j}}{\lambda-\lambda_j}, \tag{2.19}$$

$$C(\lambda)=-\lambda+\frac{1}{2}\langle\Psi_2,\Phi_1\rangle+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{2j}}{\lambda-\lambda_j}.$$

Equation (2.18) implies that $\frac{1}{2}\mathrm{Tr}\,M^2(\lambda)=A^2(\lambda)+B(\lambda)C(\lambda)$ is the generating function of integrals of motion for (2.15) and (2.17). A straightforward calculation yields

$$A^2(\lambda)+B(\lambda)C(\lambda)\equiv P(\lambda)=-\lambda+\sum_{j=1}^{N}\left[\frac{P_j}{\lambda-\lambda_j}+\frac{P_{N+j}^2}{(\lambda-\lambda_j)^2}\right], \tag{2.20}$$

where P_j, $j=1,...,2N$, are $2N$ independent integrals of motion for the FDHSs (2.15) and (2.17),

$$P_j=\frac{1}{2}\psi_{1j}\phi_{2j}+\left(-\frac{1}{2}\lambda_j+\frac{1}{4}\langle\Psi_2,\Phi_1\rangle\right)\psi_{2j}\phi_{1j}$$
$$+\frac{1}{8}\sum_{k\neq j}\frac{1}{\lambda_j-\lambda_k}[(\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j})(\psi_{1k}\phi_{1k}-\psi_{2k}\phi_{2k})+4\psi_{1j}\phi_{2j}\psi_{2k}\phi_{1k}],$$
$$j=1,...,N \tag{2.21a}$$

$$P_{N+j}=\tfrac{1}{4}(\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}),\quad j=1,...,N. \tag{2.21b}$$

It is easy to verify that

$$F_1=2\sum_{j=1}^{N}P_j,\quad F_2=2\sum_{j=1}^{N}(\lambda_jP_j+P_{N+j}^2). \tag{2.22}$$

With respect to the standard Poisson bracket it is found that

$$\{A(\lambda),A(\mu)\}=\{B(\lambda),B(\mu)\}=0,\quad \{C(\lambda),C(\mu)\}=A(\mu)-A(\lambda), \tag{2.23a}$$

$$\{A(\lambda),B(\mu)\}=\frac{1}{2(\lambda-\mu)}[B(\mu)-B(\lambda)], \tag{2.23b}$$

$$\{A(\lambda),C(\mu)\}=\frac{1}{2(\lambda-\mu)}[C(\lambda)-C(\mu)], \tag{2.23c}$$

$$\{B(\lambda),C(\mu)\}=\frac{1}{\lambda-\mu}[A(\mu)-A(\lambda)]. \tag{2.23d}$$

It follows from (2.23) that

$$\{A^2(\lambda)+B(\lambda)C(\lambda),A^2(\mu)+B(\mu)C(\mu)\}=0,$$

which implies that P_j, $j=1,...,2N$, are in involution:

$$\{P_k,P_l\}=0,\quad k,l=1,...,2N.$$

Therefore the FDHSs (2.15) and (2.17) are integrable and commute with each other. The construction of (2.15) and (2.17) ensures that if $(\Psi_1,\Psi_2,\Phi_1,\Phi_2)$ satisfies the finite-dimensional integrable Hamiltonian systems (FDIHSs) (2.15) and (2.17) simultaneously, then u defined by (2.14) solves the KdV equation (2.10).

In general, by substituting (2.14) and using (2.15), the t_n-constrained flow (2.16) becomes a t_n-FDIHS and the nth KdV equation (2.7) is factorized by the x-FDIHS (2.15) and the t_n-FDIHS. Set

$$A^2(\lambda)+B(\lambda)C(\lambda)=\lambda\sum_{k=0}^{\infty}\tilde{F}_k\lambda^{-k}, \tag{2.24a}$$

where $\tilde{F}_k$, $k=1,2,...$, are also integrals of motion for both the x-FDIHSs (2.15) and the t_n-binary constrained flows (2.16). Comparing (2.24a) with (2.20), one gets

$$\tilde{F}_0=-1,\quad \tilde{F}_1=0,\quad \tilde{F}_k=\sum_{j=1}^{N}[\lambda_j^{k-2}P_j+(k-2)\lambda_j^{k-3}P_{N+j}^2],\quad k=2,3,\ldots. \tag{2.24b}$$

By employing the method in Refs. 34 and 35, the t_n-FDIHS obtained from the t_n-constrained flow (2.16) is found to be of the form

$$\Phi_{1,t_n}=\frac{\partial F_{n+1}}{\partial\Psi_1},\quad \Phi_{2,t_n}=\frac{\partial F_{n+1}}{\partial\Psi_2},\quad \Psi_{1,t_n}=-\frac{\partial F_{n+1}}{\partial\Phi_1},\quad \Psi_{2,t_n}=-\frac{\partial F_{n+1}}{\partial\Phi_2}, \tag{2.25a}$$

with the Hamiltonian

$$F_{n+1}=\sum_{m=0}^{n}\left(\frac{1}{2}\right)^{m-1}\frac{\alpha_m}{m+1}\sum_{l_1+\cdots+l_{m+1}=n+2}\tilde{F}_{l_1}\cdots\tilde{F}_{l_{m+1}}, \tag{2.25b}$$

where $l_1\geqslant 1,\ldots, l_{m+1}\geqslant 1$, $\alpha_0=1$, $\alpha_1=\frac{1}{2}$, $\alpha_2=\frac{3}{2}$, and[34,35]

$$\alpha_m=2\alpha_{m-1}+\sum_{l=1}^{m-2}\alpha_l\alpha_{m-l-1}-\frac{1}{2}\sum_{l=1}^{m-1}\alpha_l\alpha_{m-l},\quad m\geqslant 3. \tag{2.25c}$$

The nth KdV equation (2.7) is factorized by the x-FDIHS (2.15) and the t_n-FDIHS (2.25).

For example, for the second equation in the KdV hierarchy (2.7) with $n=2$,

$$u_{t_2}=\tfrac{1}{16}(u_{xxxx}+10uu_{xx}+5u_x^2+10u^3)_x, \tag{2.26}$$

the Hamiltonian F_3 for the t_2-FDIHS reads

$$F_3=2\tilde{F}_4+\frac{1}{2}\tilde{F}_2^2=2\sum_{j=1}^{N}(\lambda_j^2P_j+2\lambda_jP_{N+j}^2)+\frac{1}{2}\left(\sum_{j=1}^{N}P_j\right)^2. \tag{2.27}$$

Then the second KdV equation (2.26) is factorized by the x-FDIHS (2.15) and the t_2-FDIHS with the Hamiltonian F_3.

B. The separation of variables for the KdV equations

An effective way to introduce the separated variables v_k, u_k and to obtain the separated equations is to use the Lax matrix M and the generating function of integrals of motion. For the FDIHSs (2.15) and (2.17), we can define the first N pairs of the canonical variables u_k, v_k, $k=1,...,N$, by the method.[2,3,4,8] The commutator relations (2.23) and the generating function of integrals of motion (2.20) enable us to define $u_1,\ldots,u_N$ by the set of zeros of $B(\lambda)$,

$$B(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{1j}}{\lambda-\lambda_j}=\frac{R(\lambda)}{K(\lambda)}, \tag{2.28a}$$

where

$$R(\lambda)=\prod_{k=1}^{N}(\lambda-u_k),\quad K(\lambda)=\prod_{k=1}^{N}(\lambda-\lambda_k),$$

and $v_1,\ldots,v_N$ by

$$v_k=A_1(u_k), \quad k=1,...,N, \tag{2.28b}$$

where

$$A_1(\lambda)=2A(\lambda)=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j}}{\lambda-\lambda_j}.$$

As we will see later, the commutator relations (2.23) guarantee that $u_1,\ldots,u_N$ and $v_1,\ldots,v_N$ satisfy the canonical conditions (1.1). Then substituting u_k into (2.20) gives rise to the separated equations

$$v_k=A_1(u_k)=2\sqrt{P(u_k)}=2\sqrt{-u_k+\sum_{j=1}^{N}\left[\frac{P_j}{u_k-\lambda_j}+\frac{P_{N+j}^2}{(u_k-\lambda_j)^2}\right]}, \quad k=1,...,N.$$

Now the reason for taking our choice of $B(\lambda)$ and $A_1(\lambda)$ becomes apparent.

The FDIHSs (2.15) and (2.17) have $2N$ degrees of freedom, therefore we need to introduce the other N pairs of canonical variables v_k,u_k, $k=N+1,...,2N$. The main idea is to construct two suitable functions $\tilde{B}(\lambda),\tilde{A}(\lambda)$ in order to define $u_{N+1},\ldots,u_{2N}$ by the set of zeros of $\tilde{B}(\lambda)$ and $v_{N+1},\ldots,v_{2N}$ by $v_{N+k}=\tilde{A}(u_{N+k})$. The above mentioned way for introducing u_k,v_k, $k=1,...,N$ stimulates us to impose two requirements on $\tilde{B}(\lambda)$ and $\tilde{A}(\lambda)$ in order to construct them. First, the canonical conditions (1.1) require that $\tilde{B}(\lambda)$ and $\tilde{A}(\lambda)$ satisfy

$$\{\tilde{B}(\lambda),\tilde{B}(\mu)\}=\{\tilde{B}(\lambda),B(\mu)\}=\{\tilde{A}(\lambda),\tilde{A}(\mu)\}=\{\tilde{A}(\lambda),B(\mu)\}=\{\tilde{A}(\lambda),A_1(\mu)\}=0, \tag{2.29a}$$

$$\{\tilde{A}(\lambda),\tilde{B}(\mu)\}=\frac{1}{\lambda-\mu}[\tilde{B}(\mu)-\tilde{B}(\lambda)], \tag{2.29b}$$

$$\{A_1(\lambda),\tilde{B}(\mu)\}=0. \tag{2.29c}$$

The second requirement is that the equation $v_{N+k}=\tilde{A}(u_{N+k})$ should give rise to the separated equations. Notice that P_{N+j} given by (2.21b) are integrals of motion for the FDIHSs (2.15) and (2.17), we can construct another generating function $\tilde{A}(\lambda)$ of integrals of motion by

$$\tilde{A}(\lambda)=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}}{\lambda-\lambda_j}=2\sum_{j=1}^{N}\frac{P_{N+j}}{\lambda-\lambda_j}. \tag{2.30a}$$

We may use $\tilde{A}(\lambda)$ to define $v_{N+1},\ldots,v_{2N}$ since substituting u_{N+k} into Eq. (2.30a) immediately leads to the separated equations for v_{N+k} and u_{N+k}. It is easy to see that $\{\tilde{A}(\lambda),B(\mu)\}=\{\tilde{A}(\lambda),\tilde{A}(\mu)\}=\{\tilde{A}(\lambda),A_1(\mu)\}=0$. We look for $\tilde{B}(\lambda)$ in the form

$$\tilde{B}(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}(\delta_1\phi_{1j}^2+\delta_2\phi_{1j}\phi_{2j}+\delta_3\phi_{2j}^2).$$

By requiring $\tilde{B}(\lambda)$ to satisfy (2.29a) and (2.29b), one gets $\delta_1=1$, $\delta_2=\delta_3=0$, i.e.,

$$\tilde{B}(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\phi_{1j}^2}{\lambda-\lambda_j}. \tag{2.30b}$$

But $\tilde{B}(\lambda)$ does not fit (2.29c). In fact, we have

$$\{A_1(\lambda),\tilde{B}(\mu)\}=\frac{1}{\lambda-\mu}[\tilde{B}(\mu)-\tilde{B}(\lambda)]. \tag{2.31}$$

However, (2.29a), (2.29b), and (2.31) enable us to replace $A_1(\lambda)$ by $\overline{A}(\lambda)$,

$$\overline{A}(\lambda)\equiv A_1(\lambda)-\tilde{A}(\lambda)=-\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{2j}}{\lambda-\lambda_j}, \tag{2.32}$$

and we will redefine v_k by $v_k=\overline{A}(u_k)$.

Then a straightforward calculation shows that $\overline{B}(\lambda)=B(\lambda),\overline{A}(\lambda),\tilde{B}(\lambda),\tilde{A}(\lambda)$ satisfy the following required commutator relations:

$$\{\overline{B}(\lambda),\overline{B}(\mu)\}=\{\tilde{B}(\lambda),\tilde{B}(\mu)\}=\{\overline{A}(\lambda),\overline{A}(\mu)\}=\{\tilde{A}(\lambda),\tilde{A}(\mu)\}=0, \tag{2.33a}$$

$$\{\overline{B}(\lambda),\tilde{B}(\mu)\}=\{\overline{B}(\lambda),\tilde{A}(\mu)\}=\{\tilde{B}(\lambda),\overline{A}(\mu)\}=\{\overline{A}(\lambda),\tilde{A}(\mu)\}=0, \tag{2.33b}$$

$$\{\overline{A}(\lambda),\overline{B}(\mu)\}=\frac{1}{\lambda-\mu}[\overline{B}(\mu)-\overline{B}(\lambda)],\quad \{\tilde{A}(\lambda),\tilde{B}(\mu)\}=\frac{1}{\lambda-\mu}[\tilde{B}(\mu)-\tilde{B}(\lambda)]. \tag{2.33c}$$

We have the following proposition.

Proposition 1: Assume that $\lambda_j,\phi_{ij},\psi_{ij}\in\mathbf{R}$, $i=1,2$, $j=1,...,N$. *Introduce the separated variables* $u_1,\dots,u_{2N}$ *by the set of zeros of* $\overline{B}(\lambda)$ *and* $\tilde{B}(\lambda)$:

$$\overline{B}(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{1j}}{\lambda-\lambda_j}=\frac{R(\lambda)}{K(\lambda)}, \tag{2.34a}$$

$$\tilde{B}(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\phi_{1j}^2}{\lambda-\lambda_j}=\frac{\overline{R}(\lambda)}{K(\lambda)}, \tag{2.34b}$$

with

$$R(\lambda)=\prod_{k=1}^{N}(\lambda-u_k),\quad \overline{R}(\lambda)=\prod_{k=1}^{N}(\lambda-u_{N+k}),$$

and $v_1,\dots,v_{2N}$ *by*

$$v_k=\overline{A}(u_k)=A_1(u_k)-\tilde{A}(u_k)=-\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{2j}}{u_k-\lambda_j},\quad k=1,...,N, \tag{2.34c}$$

$$v_{N+k}=\tilde{A}(u_{N+k})=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}}{u_{N+k}-\lambda_j},\quad k=1,...,N. \tag{2.34d}$$

If $u_1,\dots,u_N$, *are single zeros of* $\overline{B}(\lambda)$, *then* $v_1,\dots,v_{2N}$ *and* $u_1,\dots,u_{2N}$ *are canonically conjugated, i.e., they satisfy* (1.1).

Proof: Notice that

$$\lim_{\lambda\to\lambda_j-0}\tilde{B}(\lambda)=-\infty,\qquad \lim_{\lambda\to\lambda_j+0}\tilde{B}(\lambda)=\infty,$$

it is easy to see that

$$u_{N+1}<\lambda_1<u_{N+2}<\lambda_2<\cdots<u_{2N}<\lambda_N. \tag{2.35}$$

We have $\overline{B}'(u_k)\neq 0$, $\widetilde{B}'(u_{N+k})\neq 0$. Hereafter the prime denotes the differentiation with respect to λ. In what follows, we take k, $l=1,...,N$. It follows from (2.33b) that

$$0=\{u_k,\widetilde{B}(u_{N+l})\}=\widetilde{B}'(u_{N+l})\{u_k,u_{N+l}\}+\{u_k,\widetilde{B}(\mu)\}|_{\mu=u_{N+l}},$$

$$0=\{\overline{B}(u_k),u_{N+l}\}=\overline{B'}(u_k)\{u_k,u_{N+l}\}+\{\overline{B}(\lambda),u_{N+l}\}|_{\lambda=u_k},$$

$$\begin{aligned}0&=\{\overline{B}(u_k),\widetilde{B}(u_{N+l})\}\\&=\overline{B}'(u_k)\widetilde{B}'(u_{N+l})\{u_k,u_{N+l}\}+\overline{B}'(u_k)\{u_k,\widetilde{B}(\mu)\}|_{\mu=u_{N+l}}\\&\quad+\widetilde{B}'(u_{N+l})\{\overline{B}(\lambda),u_{N+l}\}|_{\lambda=u_k}+\{\overline{B}(\lambda),\widetilde{B}(\mu)\}|_{\lambda=u_k,\mu=u_{N+l}}\\&=\overline{B'}(u_k)\widetilde{B}'(u_{N+l})\{u_k,u_{N+l}\}+\overline{B'}(u_k)\{u_k,\widetilde{B}(\mu)\}|_{\mu=u_{N+1}}+\widetilde{B}'(u_{N+l})\{\overline{B}(\lambda),u_{N+l}\}|_{\lambda=u_k},\end{aligned}$$

which together lead to $\{u_k,u_{N+l}\}=0$. Similarly, $\{u_k,u_l\}=0$, $\{u_{N+k},u_{N+l}\}=0$.

Using (2.33b), (2.33c), and the above-mentioned results, one gets

$$\begin{aligned}\{v_k,\overline{B}(\mu)\}|_{\mu=u_l}&=\{\overline{A}(u_k),\overline{B}(\mu)\}|_{\mu=u_l}\\&=\overline{A'}(u_k)\{u_k,\overline{B}(\mu)\}|_{\mu=u_l}+[\{\overline{A}(\lambda),\overline{B}(\mu)\}|_{\lambda=u_k}]|_{\mu=u_l},\\&=\overline{A'}(u_k)[\{u_k,\overline{B}(u_l)\}-\overline{B'}(u_l)\{u_k,u_l\}]+[\{\overline{A}(\lambda),\overline{B}(\mu)\}|_{\lambda=u_k}]|_{\mu=u_l}\\&=[\{\overline{A}(\lambda),\overline{B}(\mu)\}|_{\lambda=u_k}]|_{\mu=u_l}\\&=\frac{\overline{B}(\mu)-\overline{B}(u_k)}{u_k-\mu}|_{\mu=u_l}\\&=-\delta_{kl}\overline{B'}(u_k),\end{aligned}$$

and

$$0=\{v_k,\overline{B}(u_l)\}=\overline{B}(u_l)\{v_k,u_l\}+\{v_k,\overline{B}(\mu)\}|_{\mu=u_l},$$

then

$$\begin{aligned}\{v_k,u_l\}&=-\frac{1}{\overline{B'}(u_l)}\{v_k,\overline{B}(\mu)\}|_{\mu=u_l}\\&=\delta_{kl}\frac{\overline{B'}(u_k)}{\overline{B'}(u_l)}=\delta_{kl}.\end{aligned}$$

In the same way, one gets $\{v_{N+k},u_{N+l}\}=\delta_{kl}$. The following equalities;

$$\{v_k,u_{N+l}\}=\{\overline{A}(u_k),u_{N+l}\}=\{\overline{A}(\lambda),u_{N+l}\}|_{\lambda=u_k},$$

$$0=\{\overline{A}(\lambda),\widetilde{B}(u_{N+l})\}=\widetilde{B}'(u_{N+l})\{\overline{A}(\lambda),u_{N+l}\},$$

yield $\{v_k,u_{N+l}\}=0$ and similarly $\{v_{N+k},u_l\}=0$.

Finally,

$$
\begin{aligned}
\{v_k, v_{N+l}\} &= \{\overline{A}(u_k), \widetilde{A}(u_{N+l})\} \\
&= \overline{A}'(u_k)\{u_k, \widetilde{A}(\mu)\}|_{\mu=u_{N+l}} + \overline{A}'(u_{N+l})\{\overline{A}(\lambda), u_{N+l}\}|_{\lambda=u_k} \\
&= \overline{A}'(u_k)[\{u_k, v_{N+l}\} - \widetilde{A}'(u_{N+l})\{u_k, u_{N+l}\}] \\
&\quad + \widetilde{A}'(u_{N+l})[\{v_k, u_{N+l}\} - \overline{A}'(u_k)\{u_k, u_{N+l}\}] = 0,
\end{aligned}
$$

similarly

$$\{v_k, v_l\} = \overline{A}'(u_k)\{u_k, v_l\} + \widetilde{A}'(u_l)\{v_k, u_l\} = -\overline{A}'(u_k)\delta_{kl} + \overline{A}'(u_l)\delta_{kl} = 0,$$

and $\{v_{N+k}, v_{N+l}\} = 0$. This completes the proof.

It follows from (2.34a) and (2.34b) that

$$\psi_{2j}\phi_{1j} = 2\frac{R(\lambda_j)}{K'(\lambda_j)}, \quad \phi_{1j}^2 = 2\frac{\overline{R}(\lambda_j)}{K'(\lambda_j)}, \quad j = 1,\ldots,N,$$

or

$$\phi_{1j} = \sqrt{\frac{2\overline{R}(\lambda_j)}{K'(\lambda_j)}}, \quad \psi_{2j} = \frac{\sqrt{2}R(\lambda_j)}{\sqrt{\overline{R}(\lambda_j)K'(\lambda_j)}}, \quad j = 1,\ldots,N. \tag{2.36}$$

Also (2.34a) results

$$u = -\langle \Psi_2, \Phi_1 \rangle = 2\sum_{j=1}^{N} (u_j - \lambda_j). \tag{2.37}$$

We now present the separated equations. By substituting u_k into (2.20), u_{N+k} into (2.30a), and using (2.34), one gets the separated equations

$$
\begin{aligned}
v_k = A_1(u_k) - \widetilde{A}(u_k) &= 2\sqrt{P(u_k)} - \widetilde{A}(u_k) \\
&= 2\sqrt{-u_k + \sum_{j=1}^{N}\left[\frac{P_j}{u_k - \lambda_j} + \frac{P_{N+j}^2}{(u_k - \lambda_j)^2}\right]} - 2\sum_{j=1}^{N}\frac{P_{N+j}}{u_k - \lambda_j}, \\
&\quad k = 1,\ldots,N,
\end{aligned} \tag{2.38a}
$$

$$v_{N+k} = \widetilde{A}(u_{N+k}) = 2\sum_{j=1}^{N}\frac{P_{N+j}}{u_{N+k} - \lambda_j}, \quad k = 1,\ldots,N. \tag{2.38b}$$

Replacing v_k by the partial derivative $\partial S/\partial u_k$ of the generating function S of the canonical transformation and interpreting the P_i as integration constants, Eqs. (2.38a) and (2.38b) give rise to the Hamilton–Jacobi equations which are completely separable and may be integrated to give the completely separated solution

$$
\begin{aligned}
S(u_1,\ldots,u_{2N}) &= \sum_{k=1}^{N}\left[\int^{u_k}(2\sqrt{P(\lambda)} - \widetilde{A}(\lambda))d\lambda + \int^{u_{N+k}}\widetilde{A}(\lambda)d\lambda\right] \\
&= 2\sum_{k=1}^{N}\left[\int^{u_k}\sqrt{P(\lambda)}d\lambda - \sum_{i=1}^{N} P_{N+i}\ln\left|\frac{u_k - \lambda_i}{u_{N+k} - \lambda_i}\right|\right].
\end{aligned} \tag{2.39}
$$

The linearizing coordinates are then

$$Q_i=\frac{\partial S}{\partial P_i}=\sum_{k=1}^{N}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda,\quad i=1,...,N, \tag{2.40a}$$

$$Q_{N+i}=\frac{\partial S}{\partial P_{N+i}}=2\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right],\quad i=1,...,N. \tag{2.40b}$$

By using (2.22), the linear flow induced by (2.15) is then given by

$$Q_i=\gamma_i+x\frac{\partial F_1}{\partial P_i}=\gamma_i+2x,\quad Q_{N+i}=2\gamma_{N+i}+x\frac{\partial F_1}{\partial P_{N+i}}=2\gamma_{N+i},\quad i=1,...,N. \tag{2.41}$$

Hereafter γ_i, $i=1,...,2N$, are arbitrary constants. Combining Eqs. (2.40a) and (2.40b) with Eq. (2.41) leads to the Jacobi inversion problem for the FDIHS (2.15),

$$\sum_{k=1}^{N}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda=\gamma_i+2x,\quad i=1,...,N, \tag{2.42a}$$

$$\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]=\gamma_{N+i},\quad i=1,...,N. \tag{2.42b}$$

The ϕ_{1j}, ψ_{2j} and $\langle\Psi_2,\Phi_1\rangle$ defined by (2.36) and (2.37) are the symmetric functions of u_k, $k=1,...,2N$. If, by using the Jacobi inversion technique,[19] ϕ_{1j}, ψ_{2j}, and $\langle\Psi_2,\Phi_1\rangle$ can be obtained from (2.42), then ϕ_{2j},ψ_{1j} can be found from the first and the last equation in (2.15) by an algebraic calculation, respectively. The $(\phi_{1j},\phi_{2j},\psi_{1j},\psi_{2j})$ provides the solution to the FDIHS (2.15).

By using (2.22), the linear flow induced by (2.17) is then given by

$$Q_i=\overline{\gamma_i}+\frac{\partial F_2}{\partial P_i}t_1=\overline{\gamma_i}+2\lambda_i t_1,$$
$$Q_{N+i}=2\overline{\gamma}_{N+i}+\frac{\partial F_2}{\partial P_{N+i}}t_1=2\overline{\gamma}_{N+i}+4P_{N+i}t_1,\quad i=1,...,N, \tag{2.43}$$

where $\overline{\gamma_i}$ are arbitrary constants. Combining Eqs. (2.40a) and (2.40b) with Eq. (2.43) leads to the Jacobi inversion problem for the FDIHS (2.17),

$$\sum_{k=1}^{N}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda=\overline{\gamma_i}+2\lambda_i t_1,\quad i=1,...,N, \tag{2.44a}$$

$$\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]=\overline{\gamma}_{N+i}+2P_{N+i}t_1,\quad i=1,...,N. \tag{2.44b}$$

Finally, since the KdV equation (2.10) is factorized by the FDIHSs (2.15) and (2.17), combining Eqs. (2.42a) and (2.42b) with Eqs. (2.44a) and (2.44b) and using (2.37) give rise to the Jacobi inversion problem for the KdV equation (2.10),

$$\sum_{k=1}^{N}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda=\gamma_i+2x+2\lambda_i t_1,\quad i=1,...,N, \tag{2.45a}$$

$$\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]=\gamma_{N+i}+2P_{N+i}t_1,\quad i=1,...,N. \tag{2.45b}$$

Notice that u defined by (2.37) is the symmetric function of u_k, $k=1,...,N$. If, by using the Jacobi inversion technique,[19] u can be found in terms of Riemann theta functions by solving (2.45), then u provides the solution of the KdV equation (2.10).

In general, since the nth KdV equation (2.7) is factorized by the x-FDIHS (2.15) and the t_n-FDIHS (2.25), the above-mentioned procedure can be applied to find the Jacobi inversion problem for the nth KdV equation (2.7). We have the following proposition.

Proposition 2: The Jacobi inversion problem for the nth KdV equation (2.7) is given by

$$\sum_{k=1}^{N}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda=\gamma_i+2x+t_n\sum_{m=0}^{n}\left(\frac{1}{2}\right)^{m-1}\alpha_m\sum_{l_1+\cdots+l_{m+1}=n+2}\lambda_i^{l_{m+1}-2}\tilde{F}_{l_1}\cdots\tilde{F}_{l_m},$$
$$i=1,...,N, \tag{2.46a}$$

$$\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]$$
$$=\gamma_{N+i}+t_n\sum_{m=0}^{n}\left(\frac{1}{2}\right)^{m-2}\alpha_m\sum_{l_1+\cdots+l_{m+1}=n+2}(l_{m+1}-2)\lambda_i^{l_{m+1}-3}P_{N+i}\tilde{F}_{l_1}\cdots\tilde{F}_{l_m},$$
$$i=1,...,N, \tag{2.46b}$$

where $l_1\geqslant 1,..., l_{m+1}\geqslant 1$ *and* $\tilde{F}_{l_1},\ldots,\tilde{F}_{l_m}$, *are given by* (2.24b).

For example, by using (2.27), the Jacobi inversion problem for the second KdV equation (2.26) is given by

$$\sum_{k=1}^{N}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda=\gamma_i+2x+\left(2\lambda_i^2+\sum_{j=1}^{N}P_j\right)t_2,\quad i=1,...,N, \tag{2.47a}$$

$$\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right]=\gamma_{N+i}+4\lambda_iP_{N+i}t_2,\quad i=1,...,N. \tag{2.47b}$$

The u solved from the Jacobi inversion problem (2.47) provides the solution for the second KdV equation (2.26).

The Jacobi inversion problem for the KdV hierarchy in our case is somewhat different from that derived by means of the stationary equations of the KdV hierarchy,[36] since there is an additional term $-\ln|u_k-\lambda_i|+\ln|u_{N+k}-\lambda_i|$ in (2.46b).

III. THE SEPARATION OF VARIABLES FOR THE AKNS EQUATIONS

A. Binary constrained flows of the AKNS hierarchy

For the AKNS spectral problem[37]

$$\phi_x=U(u,\lambda)\phi,\quad U(u,\lambda)=\begin{pmatrix}-\lambda & q\\ r & \lambda\end{pmatrix},\quad \phi=\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\quad u=\begin{pmatrix}q\\ r\end{pmatrix}, \tag{3.1}$$

its adjoint representation (2.2) and (2.3) yield

$$a_0=-1,\quad b_0=c_0=0,\quad a_1=0,\quad b_1=q,\quad c_1=r,\quad a_2=\tfrac{1}{2}qr,\ldots,$$

and in general

$$\begin{pmatrix} c_{k+1} \\ b_{k+1} \end{pmatrix} = L \begin{pmatrix} c_k \\ b_k \end{pmatrix}, \quad a_k = \partial^{-1}(qc_k - rb_k), \quad k=1,2,..., \tag{3.2}$$

$$L = \frac{1}{2} \begin{pmatrix} \partial - 2r\partial^{-1}q & 2r\partial^{-1}r \\ -2q\partial^{-1}q & -\partial + 2q\partial^{-1}r \end{pmatrix}.$$

Take

$$\phi_{t_n} = V^{(n)}(u,\lambda)\phi, \quad V^{(n)}(u,\lambda) = \sum_{i=0}^{n} \begin{pmatrix} a_i & b_i \\ c_i & -a_i \end{pmatrix} \lambda^{n-i}. \tag{3.3}$$

Then the compatibility condition of Eqs. (3.1) and (3.3) gives rise to the AKNS hierarchy

$$u_{t_n} = \begin{pmatrix} q \\ r \end{pmatrix}_{t_n} = JL^n \begin{pmatrix} r \\ q \end{pmatrix} = J \frac{\delta H_{n+1}}{\delta u}, \quad n=1,2,..., \tag{3.4}$$

where the Hamiltonian H_n and the Hamiltonian operator J are given by

$$J = \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix}, \quad H_n = \frac{2a_{n+1}}{n+1}, \quad \begin{pmatrix} c_n \\ b_n \end{pmatrix} = \frac{\delta H_n}{\delta u}, \quad n=1,2,... .$$

For $n=2$ we have

$$\phi_{t_2} = V^{(2)}(u,\lambda)\phi, \quad V^{(2)} = \begin{pmatrix} -\lambda^2 + \frac{1}{2}qr & q\lambda - \frac{1}{2}q_x \\ r\lambda + \frac{1}{2}r_x & \lambda^2 - \frac{1}{2}qr \end{pmatrix}, \tag{3.5}$$

and the AKNS equation (3.4) for $n=2$ reads

$$q_{t_2} = -\tfrac{1}{2}q_{xx} + q^2 r, \quad r_{t_2} = \tfrac{1}{2}r_{xx} - r^2 q. \tag{3.6}$$

The adjoint AKNS spectral problem is Eq. (2.11) with U given by (3.1). We have[25]

$$\frac{\delta\lambda}{\delta u} = \begin{pmatrix} \dfrac{\delta\lambda}{\delta q} \\[2mm] \dfrac{\delta\lambda}{\delta r} \end{pmatrix} = \beta \,\mathrm{Tr}\left[\begin{pmatrix} \phi_1\psi_1 & \phi_1\psi_2 \\ \phi_2\psi_1 & \phi_2\psi_2 \end{pmatrix} \frac{\partial U(u,\lambda)}{\partial u} \right] = \beta \begin{pmatrix} \psi_1\phi_2 \\ \psi_2\phi_1 \end{pmatrix}. \tag{3.7}$$

The binary x-constrained flows of the AKNS hierarchy (3.4) are defined by[25,29]

$$\Phi_{1,x} = -\Lambda\Phi_1 + q\Phi_2, \quad \Phi_{2,x} = r\Phi_1 + \Lambda\Phi_2, \tag{3.8a}$$

$$\Psi_{1,x} = \Lambda\Psi_1 - r\Psi_2, \quad \Psi_{2,x} = -q\Psi_1 - \Lambda\Psi_2, \tag{3.8b}$$

$$\frac{\delta H_{k_0}}{\delta u} - \beta^{-1}\sum_{j=1}^{N} \frac{\delta\lambda_j}{\delta u} = \begin{pmatrix} c_{k_0} \\ b_{k_0} \end{pmatrix} - \begin{pmatrix} \langle\Psi_1,\Phi_2\rangle \\ \langle\Psi_2,\Phi_1\rangle \end{pmatrix} = 0. \tag{3.8c}$$

For $k_0 = 1$, we have

$$\begin{pmatrix} c_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} r \\ q \end{pmatrix} = \begin{pmatrix} \langle\Psi_1,\Phi_2\rangle \\ \langle\Psi_2,\Phi_1\rangle \end{pmatrix} = 0. \tag{3.9}$$

By substituting (3.9) into (3.8a) and (3.8b), the first binary x-constrained flow becomes a x-FDHS,[25]

$$\Phi_{1x}=\frac{\partial F_1}{\partial \Psi_1},\quad \Phi_{2x}=\frac{\partial F_1}{\partial \Psi_2},\quad \Psi_{1x}=-\frac{\partial F_1}{\partial \Phi_1},\quad \Psi_{2x}=-\frac{\partial F_1}{\partial \Phi_2},\tag{3.10}$$

with the Hamiltonian

$$F_1=\langle \Lambda\Psi_2,\Phi_2\rangle-\langle \Lambda\Psi_1,\Phi_1\rangle+\langle \Psi_2,\Phi_1\rangle\langle \Psi_1,\Phi_2\rangle.$$

Under the constraint (3.9) and the FDHS (3.10), the binary t_2-constrained flow obtained from (3.3) with $V^{(2)}$ given by (3.5) and its adjoint equation for N distinct real number λ_j can also be written as a t_2-FDHS,

$$\Phi_{1,t_2}=\frac{\partial F_2}{\partial \Psi_1},\quad \Phi_{2,t_2}=\frac{\partial F_2}{\partial \Psi_2},\quad \Psi_{1,t_2}=-\frac{\partial F_2}{\partial \Phi_1},\quad \Psi_{2,t_2}=-\frac{\partial F_2}{\partial \Phi_2},\tag{3.11}$$

with the Hamiltonian

$$\begin{aligned}F_2=&\langle \Lambda^2\Psi_2,\Phi_2\rangle-\langle \Lambda^2\Psi_1,\Phi_1\rangle+\langle \Psi_2,\Phi_1\rangle\langle \Lambda\Psi_1,\Phi_2\rangle+\langle \Lambda\Psi_2,\Phi_1\rangle\langle \Psi_1,\Phi_2\rangle\\&-\tfrac{1}{2}(\langle \Psi_2,\Phi_2\rangle-\langle \Psi_1,\Phi_1\rangle)\langle \Psi_2,\Phi_1\rangle\langle \Psi_1,\Phi_2\rangle.\end{aligned}$$

The Lax representation for the FDHSs (3.10) and (3.11) which can also be deduced from the adjoint representation (2.2) and (2.8) are presented by (2.18) with the entries of the Lax matrix M given by[29]

$$A(\lambda)=-1+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j}}{\lambda-\lambda_j},\tag{3.12a}$$

$$B(\lambda)=\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{1j}}{\lambda-\lambda_j},\quad C(\lambda)=\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{2j}}{\lambda-\lambda_j}.\tag{3.12b}$$

A straightforward calculation yields

$$A^2(\lambda)+B(\lambda)C(\lambda)\equiv P(\lambda)=1+\sum_{j=1}^{N}\left[\frac{P_j}{\lambda-\lambda_j}+\frac{P_{N+j}^2}{(\lambda-\lambda_j)^2}\right],\tag{3.13}$$

where P_j, $j=1,\ldots,2N$, are $2N$ independent integrals of motion for the FDHSs (3.10) and (3.11),

$$P_j=\psi_{2j}\phi_{2j}-\psi_{1j}\phi_{1j}+\frac{1}{2}\sum_{k\neq j}\frac{1}{\lambda_j-\lambda_k}[(\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j})(\psi_{1k}\phi_{1k}-\psi_{2k}\phi_{2k})+4\psi_{1j}\phi_{2j}\psi_{2k}\phi_{1k}],$$
$$j=1,\ldots,N,\tag{3.14a}$$

$$P_{N+j}=\tfrac{1}{2}(\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}),\quad j=1,\ldots,N.\tag{3.14b}$$

It is easy to verify that

$$F_1=\sum_{j=1}^{N}(\lambda_jP_j+P_{N+j}^2)-\frac{1}{4}\left(\sum_{j=1}^{N}P_j\right)^2,\tag{3.15a}$$

$$F_2=\sum_{j=1}^{N}(\lambda_j^2P_j+2\lambda_jP_{N+j}^2)-\frac{1}{2}\left(\sum_{j=1}^{N}P_j\right)\sum_{j=1}^{N}(\lambda_jP_j+P_{N+j}^2)+\frac{1}{8}\left(\sum_{j=1}^{N}P_j\right)^3.\tag{3.15b}$$

With respect to the standard Poisson bracket it is found that

$$\{A(\lambda),A(\mu)\}=\{B(\lambda),B(\mu)\}=\{C(\lambda),C(\mu)\}=0, \tag{3.16a}$$

$$\{A(\lambda),B(\mu)\}=\frac{1}{\lambda-\mu}[B(\mu)-B(\lambda)], \tag{3.16b}$$

$$\{A(\lambda),C(\mu)\}=\frac{1}{\lambda-\mu}[C(\lambda)-C(\mu)], \tag{3.16c}$$

$$\{B(\lambda),C(\mu)\}=\frac{2}{\lambda-\mu}[A(\mu)-A(\lambda)]. \tag{3.16d}$$

Then $\{A^2(\lambda)+B(\lambda)C(\lambda),A^2(\mu)+B(\mu)C(\mu)\}=0$ implies that P_j, $j=1,...,2N$, are in involution. The AKNS equation (3.6) is factorized by the x-FDIHS (3.10) and the t_2-FDIHS (3.11), namely, if $(\Psi_1,\Psi_2,\Phi_1,\Phi_2)$ satisfies the FDIHSs (3.10) and (3.11) simultaneously, then (q, r) given by (3.9) solves the AKNS equation (3.6). In general, the factorization of the nth AKNS equations (3.4) will be presented at the end of Sec. III B.

B. The separation of variables for the AKNS equations

In contrast with the $B(\lambda)$ in the Lax matrix M for the constrained KdV flows, the $B(\lambda)$ given by (3.12b) has only $N-1$ zeros, one has to define the canonical variables u_k, v_k, $k=1,...,2N$, in a little different way. The commutator relations (3.16) and the generating function of integrals of motion (3.13) enable us to introduce $u_1,\ldots,u_N$ by means of $B(\lambda)$ in the following way:

$$B(\lambda)=\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{1j}}{\lambda-\lambda_j}=e^{u_N}\frac{R(\lambda)}{K(\lambda)}, \tag{3.17a}$$

where

$$R(\lambda)=\prod_{k=1}^{N-1}(\lambda-u_k),\quad K(\lambda)=\prod_{k=1}^{N}(\lambda-\lambda_k),$$

and $v_1,\ldots,v_N$ by

$$v_k=A(u_k),\quad k=1,...,N-1,\quad v_N=\tfrac{1}{2}(\langle\Psi_1,\Phi_1\rangle-\langle\Psi_2,\Phi_2\rangle). \tag{3.17b}$$

Equation (3.17a) yields

$$u_N=\ln\langle\Psi_2,\Phi_1\rangle. \tag{3.17c}$$

Then it is easy to verify that

$$\{u_N,B(\mu)\}=\{v_N,A(\mu)\}=0,\quad \{v_N,u_N\}=1, \tag{3.18a}$$

$$\{u_N,A(\mu)\}=-\frac{B(\mu)}{\langle\Psi_2,\Phi_1\rangle},\quad \{v_N,B(\mu)\}=B(\mu). \tag{3.18b}$$

As we will show later, the commutator relations (3.16) and (3.18) guarantee that $u_1,\ldots,u_N$, $v_1,\ldots,v_N$ satisfy the canonical conditions (1.1).

We now need to construct two functions $\tilde{B}(\lambda)$, $\tilde{A}(\lambda)$ to define $u_{N+1},\ldots,u_{2N}$ by means of $\tilde{B}(\lambda)$ and $v_{N+1},\ldots,v_{2N}$ by $v_{N+k}=\tilde{A}(u_{N+k})$. By the exactly same argument as in Sec. II B, we construct $\tilde{A}(\lambda)$ by

$$\tilde{A}(\lambda)=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}}{\lambda-\lambda_j}=\sum_{j=1}^{N}\frac{P_{N+j}}{\lambda-\lambda_j}, \tag{3.19a}$$

since Eq. (3.19a) enables us to obtain immediately the separated equation (1.2) for v_{N+k} and u_{N+k}, and $\tilde{B}(\lambda)$ by

$$\tilde{B}(\lambda)=\sum_{j=1}^{N}\frac{\phi_{1j}^2}{\lambda-\lambda_j}. \tag{3.19b}$$

Then it is easy to verify that $A(\lambda),B(\lambda),\tilde{A}(\lambda),\tilde{B}(\mu)$ satisfy the commutator relations

$$\{\tilde{B}(\lambda),\tilde{B}(\mu)\}=\{\tilde{B}(\lambda),B(\mu)\}=\{\tilde{A}(\lambda),\tilde{A}(\mu)\}=\{\tilde{A}(\lambda),B(\mu)\}=\{\tilde{A}(\lambda),A(\mu)\}=0, \tag{3.20a}$$

$$\{\tilde{A}(\lambda),\tilde{B}(\mu)\}=\frac{1}{\lambda-\mu}[\tilde{B}(\mu)-\tilde{B}(\lambda)], \tag{3.20b}$$

$$\{A(\lambda),\tilde{B}(\mu)\}=\frac{1}{\lambda-\mu}[\tilde{B}(\mu)-\tilde{B}(\lambda)]. \tag{3.20c}$$

Relation (3.20c) does not fit the requirement for the canonical conditions (1.1). According to (3.20) and (3.16) we can replace $A(\lambda)$ by $\overline{A}(\lambda)$,

$$\overline{A}(\lambda)\equiv A(\lambda)-\tilde{A}(\lambda)=-1-\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{2j}}{\lambda-\lambda_j}, \tag{3.21}$$

namely, we redefine $v_1,\ldots,v_N$ by

$$v_k=\overline{A}(u_k)=A(u_k)-\tilde{A}(u_k)=-1-\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{2j}}{u_k-\lambda_j},\quad k=1,\ldots,N-1, \tag{3.22a}$$

$$v_N=-\langle\Psi_2,\Phi_2\rangle. \tag{3.22b}$$

We now define $u_{N+1},\ldots,u_{2N}$ by $\tilde{B}(\lambda)$ as follows:

$$\tilde{B}(\lambda)=\sum_{j=1}^{N}\frac{\phi_{1j}^2}{\lambda-\lambda_j}=e^{u_{2N}}\frac{\overline{R}(\lambda)}{K(\lambda)},\quad \overline{R}(\lambda)=\prod_{k=1}^{N-1}(\lambda-u_{N+k}), \tag{3.23a}$$

and $v_{N+1},\ldots,v_{2N}$ by

$$v_{N+k}=\tilde{A}(u_{N+k})=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}}{u_{N+k}-\lambda_j},\quad k=1,\ldots,N-1, \tag{3.23b}$$

$$v_{2N}=\tfrac{1}{2}(\langle\Psi_1,\Phi_1\rangle+\langle\Psi_2,\Phi_2\rangle). \tag{3.23c}$$

Equation (3.23a) leads to

$$u_{2N}=\ln\langle\Phi_1,\Phi_1\rangle. \tag{3.23d}$$

Then a straightforward calculation shows that $\overline{B}(\lambda)=B(\lambda)$, $\overline{A}(\lambda),\tilde{B}(\lambda),\tilde{A}(\lambda)$ satisfy the commutator relations (2.33) and

$$\{u_N,\overline{B}(\mu)\}=\{u_N,\tilde{B}(\mu)\}=\{u_N,\tilde{A}(\mu)\}=0,\quad \{u_N,\overline{A}(\mu)\}=-\frac{\overline{B}(\mu)}{\langle\Psi_2,\Phi_1\rangle},\tag{3.24a}$$

$$\{v_N,\overline{A}(\mu)\}=\{v_N,\tilde{B}(\mu)\}=\{v_N,\tilde{A}(\mu)\}=0,\quad \{v_N,\overline{B}(\mu)\}=\overline{B}(\mu),\tag{3.24b}$$

$$\{u_{2N},\overline{B}(\mu)\}=\{u_{2N},\overline{A}(\mu)\}=\{u_{2N},\tilde{B}(\mu)\}=0,\quad \{u_{2N},\tilde{A}(\mu)\}=-\frac{\tilde{B}(\mu)}{\langle\Psi_1,\Phi_1\rangle},\tag{3.24c}$$

$$\{v_{2N},\overline{B}(\mu)\}=\{v_{2N},\overline{A}(\mu)\}=\{v_{2N},\tilde{A}(\mu)\}=0,\quad \{v_{2N},\tilde{B}(\mu)\}=\tilde{B}(\mu),\tag{3.24d}$$

$$\{v_N,u_N\}=1,\quad \{v_{2N},u_{2N}\}=1,\tag{3.24e}$$

$$\{u_{2N},u_N\}=\{u_{2N},v_N\}=\{v_{2N},u_N\}=\{v_{2N},v_N\}=0.\tag{3.24f}$$

We have the following proposition.

Proposition 3: Assume that $\lambda_j,\ \phi_{ij},\ \psi_{ij}\in\mathbf{R},\ i=1,2,\ j=1,...,N$. *Introduce the separated variables* $u_1,\ldots,u_{2N}$ *by the* $\overline{B}(\lambda)$ *and* $\tilde{B}(\lambda)$:

$$\overline{B}(\lambda)=B(\lambda)=\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{1j}}{\lambda-\lambda_j}=e^{u_N}\frac{R(\lambda)}{K(\lambda)},\tag{3.25a}$$

$$\tilde{B}(\lambda)=\sum_{j=1}^{N}\frac{\phi_{1j}^2}{\lambda-\lambda_j}=e^{u_{2N}}\frac{\overline{R}(\lambda)}{K(\lambda)},\tag{3.25b}$$

with

$$R(\lambda)=\prod_{k=1}^{N-1}(\lambda-u_k),\quad \overline{R}(\lambda)=\prod_{k=1}^{N-1}(\lambda-u_{N+k}),$$

and $v_1,\ldots,v_{2N}$ *by*

$$v_k=\overline{A}(u_k)=A(u_k)=-\tilde{A}(u_k)=-1-\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{2j}}{u_k-\lambda_j},\quad k=1,...,N-1,\tag{3.25c}$$

$$v_N=-\langle\Psi_2,\Phi_2\rangle,\tag{3.25d}$$

$$v_{N+k}=\tilde{A}(u_{N+k})=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}}{u_{N+k}-\lambda_j},\quad k=1,...,N-1,\tag{3.25e}$$

$$v_{2N}=\tfrac{1}{2}(\langle\Psi_1,\Phi_1\rangle+\langle\Psi_2,\Phi_2\rangle).\tag{3.25f}$$

If $u_1,\ldots,u_N$, *are single zeros of* $\overline{B}(\lambda)$, *then* $v_1,\ldots,v_{2N}$ *and* $u_1,\ldots,u_{2N}$ *are canonically conjugated, i.e., they satisfy* (1.1).

Proof: By using the exactly same method as in the proof of proposition 1, the commutator relations (2.33) guarantee that $u_1,\ldots,u_{N-1},\ v_1,\ldots,v_{N-1}$ satisfy the canonical conditions (1.1). By using the similar method, for example, it is found from (3.24) that for $k=1,...,N-1$, we have

$$0=\{u_N,\overline{B}(u_k)\}=\overline{B}'(u_k)\{u_N,u_k\}+\{u_N,\overline{B}(\mu)\}|_{\mu=u_k}=\overline{B}'(u_k)\{u_N,u_k\},$$

$$\{u_N,v_k\}=\{u_N,\overline{A}(u_k)\}=\overline{A}'(u_k)\{u_N,u_k\}-\frac{\overline{B}(\mu)}{\langle\Psi_2,\Phi_1\rangle}\bigg|_{\mu=u_k}=0,$$

which gives rise to $\{u_N,u_k\}=\{u_N,v_k\}=0$ and so on. In this way we complete the proof.

It follows from (3.25a) and (3.25b) that

$$\psi_{2j}\phi_{1j}=e^{u_N}\frac{R(\lambda_j)}{K'(\lambda_j)},\quad \phi_{1j}^2=e^{u_{2N}}\frac{\overline{R}(\lambda_j)}{K'(\lambda_j)},\quad j=1,...,N,$$

or

$$\phi_{1j}=\sqrt{\frac{e^{u_{2N}}\overline{R}(\lambda_j)}{K'(\lambda_j)}},\quad \psi_{2j}=\frac{e^{u_N}R(\lambda_j)}{\sqrt{e^{u_{2N}}\overline{R}(\lambda_j)K'(\lambda_j)}},\quad j=1,..,N. \tag{3.26}$$

Equations (3.9) and (3.17c) result in

$$q=e^{u_N}. \tag{3.27}$$

We now present the separated equations. By substituting u_k into (3.13), u_{N+k} into (3.19a) and using (3.25c) and (3.25e), one gets the separated equations

$$\begin{aligned}v_k=A(u_k)-\tilde{A}(u_k)&=\sqrt{P(u_k)}-\tilde{A}(u_k)\\ &=\sqrt{1+\sum_{j=1}^{N}\left[\frac{P_j}{u_k-\lambda_j}+\frac{P_{N+j}}{(u_k-\lambda_j)^2}\right]}-\sum_{j=1}^{N}\frac{P_{N+j}}{u_k-\lambda_j},\quad k=1,...,N-1,\end{aligned} \tag{3.28a}$$

$$v_{N+k}=\tilde{A}(u_{N+k})=\sum_{j=1}^{N}\frac{P_{N+j}}{u_{N+k}-\lambda_j},\quad k=1,...,N-1. \tag{3.28b}$$

It is easy to see from (3.14) that

$$\langle\Psi_2,\Phi_2\rangle-\langle\Psi_1,\Phi_1\rangle=\sum_{i=1}^{N}P_i,\quad \langle\Psi_1,\Phi_1\rangle+\langle\Psi_2,\Phi_2\rangle=2\sum_{i=1}^{N}P_{N+i},$$

which together with (3.25d) and (3.25f) leads to

$$v_N=-\frac{1}{2}\sum_{i=1}^{N}(P_i+2P_{N+i}),\quad v_{2N}=\sum_{i=1}^{N}P_{N+i}. \tag{3.28c}$$

Replacing v_k by the partial derivative $\partial S/\partial u_k$ of the generating function S of the canonical transformation and interpreting the P_i as integration constants, Eqs. (3.28a)–(3.28c) may be integrated to give the generating function of the canonical transformation

$$S(u_1,\ldots,u_{2N})=\sum_{k=1}^{N-1}\left[\int^{u_k}(\sqrt{P(\lambda)}-\tilde{A}(\lambda))d\lambda+\int^{u_{N+k}}\tilde{A}(\lambda)d\lambda\right]$$

$$-\frac{1}{2}\sum_{i=1}^{N}(P_i+2P_{N+i})u_N+\sum_{i=1}^{N}P_{N+i}u_{2N}$$

$$=\sum_{k=1}^{N-1}\left[\int^{u_k}\sqrt{P(\lambda)}d\lambda-\sum_{i=1}^{N}P_{N+i}\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]$$

$$-\frac{1}{2}\sum_{i=1}^{N}(P_i+2P_{N+i})u_N+\sum_{i=1}^{N}P_{N+i}u_{2N}. \tag{3.29}$$

The linearizing coordinates are then

$$Q_i=\frac{\partial S}{\partial P_i}=\frac{1}{2}\sum_{k=1}^{N-1}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda-\frac{1}{2}u_N,\quad i=1,\ldots,N, \tag{3.30a}$$

$$Q_{N+i}=\frac{\partial S}{\partial P_{N+i}}=\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]-u_N+u_{2N},$$

$$i=1,\ldots,N. \tag{3.30b}$$

By using (3.15a), the linear flow induced by the FDIHS (3.10) together with Eqs. (3.30a) and (3.30b) leads to the Jacobi inversion problem for the FDIHS (3.10),

$$\sum_{k=1}^{N-1}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda-u_N=\gamma_i+\left(2\lambda_i-\sum_{k=1}^{N}P_k\right)x,\quad i=1,\ldots,N, \tag{3.31a}$$

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]-u_N+u_{2N}=\gamma_{N+i}+2P_{N+i}x,\quad i=1,\ldots,N. \tag{3.31b}$$

By using (3.15b), the linear flow induced by the FDIHS (3.11) and Eqs. (3.30a) and (3.30b) result in the Jacobi inversion problem for the FDIHS (3.11),

$$\sum_{k=1}^{N-1}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda-u_N=\overline{\gamma}_i+\left[2\lambda_i^2-\sum_{k=1}^{N}(\lambda_kP_k+\lambda_iP_k+P_{N+k}^2)+\frac{3}{4}\left(\sum_{k=1}^{N}P_k\right)^2\right]t_2,$$

$$i=1,\ldots,N, \tag{3.32a}$$

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]-u_N+u_{2N}$$

$$=\overline{\gamma}_{N+i}+P_{N+i}\left(4\lambda_i-\sum_{k=1}^{N}P_k\right)t_2,$$

$$i=1,\ldots,N. \tag{3.32b}$$

Then, since the AKNS equations (3.6) are factorized by the FDIHSs (3.10) and (3.11), combining Eqs. (3.31a) and (3.31b) with Eqs. (3.32a) and (3.32b) gives rise to the Jacobi inversion problem for the AKNS equations (3.6),

$$\sum_{k=1}^{N-1}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda-u_N$$
$$=\gamma_i+\left(2\lambda_i-\sum_{k=1}^{N}P_k\right)x+\left[2\lambda_i^2-\sum_{k=1}^{N}(\lambda_kP_k+\lambda_iP_k+P_{N+k}^2)+\frac{3}{4}\left(\sum_{k=1}^{N}P_k\right)^2\right]t_2,$$
$$i=1,...,N, \tag{3.33a}$$

$$\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]-u_N+u_{2N}$$
$$=\gamma_{N+i}+2P_{N+i}x+P_{N+i}\left(4\lambda_i-\sum_{k=1}^{N}P_k\right)t_2,$$
$$i=1,...,N. \tag{3.33b}$$

If ϕ_{1j}, ψ_{2j}, q defined by (3.26) and (3.27) can be solved from (3.33) by using the Jacobi inversion technique, then ϕ_{2j}, ψ_{1j} can be obtained from the first equation and the last equation in (3.10) by an algebraic calculation, respectively. Finally q and $r=\langle\Psi_1,\Phi_2\rangle$ provides the solution to the AKNS equations (3.6).

In general, in order to obtain the Jacobi inversion problem for the nth AKNS equations (3.4), we set

$$A^2(\lambda)+B(\lambda)C(\lambda)=\sum_{k=0}^{\infty}\tilde{F}_k\lambda^{-k}, \tag{3.34}$$

where $\tilde{F}_k$, $k=1,2,...$, are also integrals of motion for both the FDHS (3.10) and the t_n-binary constrained flow. Comparing (3.34) with (3.13), one gets

$$\tilde{F}_0=1,\quad \tilde{F}_k=\sum_{j=1}^{N}[\lambda_j^{k-1}P_j+(k-1)\lambda_j^{k-2}P_{N+j}^2],\quad k=1,2,... . \tag{3.35}$$

The nth AKNS equations (3.4) are factorized by the x-FDIHS (3.10) and the t_n-FDIHS with the Hamiltonian F_n given by[25]

$$F_n=2\sum_{m=0}^{n}\left(-\frac{1}{2}\right)^m\frac{\alpha_m}{m+1}\sum_{l_1+\cdots+l_{m+1}=n+1}\tilde{F}_{l_1}\cdots\tilde{F}_{l_{m+1}}, \tag{3.36}$$

where $l_1\geqslant 1,...,\ l_{m+1}\geqslant 1$, and α_m are given by (2.25c). In the same way, we have the following proposition.

Proposition 4: The Jacobi inversion problem for the nth AKNS equations (3.4) is of the form

$$\sum_{k=1}^{N-1}\int^{u_k}\frac{1}{(\lambda-\lambda_i)\sqrt{P(\lambda)}}d\lambda-u_N=\gamma_i+\left(2\lambda_i-\sum_{k=1}^{N}P_k\right)x$$
$$+2t_n\sum_{m=0}^{n}\left(-\frac{1}{2}\right)^m\alpha_m\sum_{l_1+\cdots+l_{m+1}=n+1}\lambda_i^{l_{m+1}-1}\tilde{F}_{l_1}\cdots\tilde{F}_{l_m},$$
$$i=1,...,N, \tag{3.37a}$$

$$\sum_{k=1}^{N}\left[\int^{u_k}\frac{P_{N+i}}{(\lambda-\lambda_i)^2\sqrt{P(\lambda)}}d\lambda-\ln\left|\frac{u_k-\lambda_i}{u_{N+k}-\lambda_i}\right|\right]-u_N+u_{2N}$$
$$=\gamma_{N+i}+2P_{N+i}x+4t_n\sum_{m=0}^{n}\left(-\frac{1}{2}\right)^m\alpha_m\sum_{l_1+\cdots+l_{m+1}=n+1}(l_{m+1}-1)\lambda_i^{l_{m+1}-2}P_{N+i}\tilde{F}_{l_1}\cdots\tilde{F}_{l_m},$$
$$i=1,...,N, \tag{3.37b}$$

where $l_1\geqslant 1,..., l_{m+1}\geqslant 1$, *and* $\tilde{F}_{l_1},\ldots\tilde{F}_{l_m}$, *are given by* (3.35).

IV. THE SEPARATION OF VARIABLES FOR THE KAUP–NEWELL EQUATIONS

A. Binary constrained flows of the Kaup–Newell hierarchy

For the Kaup–Newell spectral problem[38]

$$\phi_x=U(u,\lambda)\phi,\quad U(u,\lambda)=\begin{pmatrix}-\lambda^2 & q\lambda\\ r\lambda & \lambda^2\end{pmatrix},\quad \phi=\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\quad u=\begin{pmatrix}q\\ r\end{pmatrix}, \tag{4.1}$$

its adjoint representation (2.2) and (2.3) yields

$$a_0=1,\quad a_2=-\tfrac{1}{2}qr,\quad b_1=-q,\quad c_1=-r,\quad b_3=\tfrac{1}{2}(q^2r+q_x),\quad c_3=\tfrac{1}{2}(qr^2-r_x),\ldots,$$

and in general $a_{2k+1}=b_{2k}=c_{2k}=0$,

$$\begin{pmatrix}c_{2k+1}\\ b_{2k+1}\end{pmatrix}=L\begin{pmatrix}c_{2k-1}\\ b_{2k-1}\end{pmatrix},\quad a_{2k}=\frac{1}{2}\,\partial^{-1}(qc_{2k-1,x}+rb_{2k-1,x}),\quad k=1,2,..., \tag{4.2}$$

$$L=\tfrac{1}{2}\begin{pmatrix}\partial-r\partial^{-1}q\partial & -r\partial^{-1}r\partial\\ -q\partial^{-1}q\partial & -\partial-q\partial^{-1}r\partial\end{pmatrix}.$$

Take

$$\phi_{t_n}=V^{(n)}(u,\lambda)\phi,\quad V^{(n)}(u,\lambda)=\sum_{i=0}^{n-1}\begin{pmatrix}a_{2i}\lambda^{2n-2i} & b_{2i+1}\lambda^{2n-2i-1}\\ c_{2i+1}\lambda^{2n-2i-1} & -a_{2i}\lambda^{2n-2i}\end{pmatrix}. \tag{4.3}$$

Then the compatibility condition of Eqs. (4.1) and (4.3) gives rise to the Kaup–Newell hierarchy

$$u_{t_n}=\begin{pmatrix}q\\ r\end{pmatrix}_{t_n}=J\begin{pmatrix}c_{2n-1}\\ b_{2n-1}\end{pmatrix}=J\,\frac{\delta H_{2n-2}}{\delta u},\quad n=1,2,..., \tag{4.4}$$

where the Hamiltonian H_n and the Hamiltonian operator J are given by

$$J=\begin{pmatrix}0 & \partial\\ \partial & 0\end{pmatrix},\quad H_{2n}=\frac{4a_{2n+2}-rc_{2n+1}-qb_{2n+1}}{2n},\quad \begin{pmatrix}c_{2n+1}\\ b_{2n+1}\end{pmatrix}=\frac{\delta H_{2n}}{\delta u}.$$

For $n=2$ we have

$$\phi_{t_2}=V^{(2)}(u,\lambda)\phi,\quad V^{(2)}=\begin{pmatrix}\lambda^4-\tfrac{1}{2}qr\lambda^2 & -q\lambda^3+\tfrac{1}{2}(q^2r+q_x)\lambda\\ -r\lambda^3+\tfrac{1}{2}(qr^2-r_x)\lambda & -\lambda^4+\tfrac{1}{2}qr\lambda^2\end{pmatrix}, \tag{4.5}$$

and the coupled derivative nonlinear Schrödinger (CDNS) equations obtained from Eq. (4.4) for $n=2$ read

$$q_{t_2}=\tfrac{1}{2}q_{xx}+\tfrac{1}{2}(q^2r)_x\,,\quad r_{t_2}=-\tfrac{1}{2}r_{xx}+\tfrac{1}{2}(r^2q)_x\,. \tag{4.6}$$

The adjoint Kaup–Newell spectral problem is Eq. (2.11) with U given by (4.1). We have[25]

$$\frac{\delta\lambda}{\delta u}=\begin{pmatrix}\dfrac{\delta\lambda}{\delta q}\\ \dfrac{\delta\lambda}{\delta r}\end{pmatrix}=\beta\,\mathrm{Tr}\left[\begin{pmatrix}\phi_1\psi_1 & \phi_1\psi_2\\ \phi_2\psi_1 & \phi_2\psi_2\end{pmatrix}\frac{\partial U(u,\lambda)}{\partial u}\right]=\beta\begin{pmatrix}\lambda\,\psi_1\phi_2\\ \lambda\,\psi_2\phi_1\end{pmatrix}. \tag{4.7}$$

The binary x-constrained flows of the Kaup–Newell hierarchy (4.4) are defined by

$$\Phi_{1,x}=-\Lambda^2\Phi_1+q\Lambda\Phi_2,\quad \Phi_{2,x}=r\Lambda\Phi_1+\Lambda^2\Phi_2, \tag{4.8a}$$

$$\Psi_{1,x}=\Lambda^2\Psi_1-r\Lambda\Psi_2,\quad \Psi_{2,x}=-q\Lambda\Psi_1-\Lambda^2\Psi_2, \tag{4.8b}$$

$$\frac{\delta H_{k_0}}{\delta u}-\beta^{-1}\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta u}=\begin{pmatrix}c_{2k_0+1}\\ b_{2k_0+1}\end{pmatrix}-\frac{1}{2}\begin{pmatrix}\langle\Lambda\Psi_1,\Phi_2\rangle\\ \langle\Lambda\Psi_2,\Phi_1\rangle\end{pmatrix}=0. \tag{4.8c}$$

For $k_0=1$, we have

$$\begin{pmatrix}c_1\\ b_1\end{pmatrix}=-\begin{pmatrix}r\\ q\end{pmatrix}=\frac{1}{2}\begin{pmatrix}\langle\Lambda\Psi_1,\Phi_2\rangle\\ \langle\Lambda\Psi_2,\Phi_1\rangle\end{pmatrix}=0. \tag{4.9}$$

By substituting (4.9) into (4.8a) and (4.8b), the first binary x-constrained flow becomes a FDHS

$$\Phi_{1x}=\frac{\partial F_1}{\partial\Psi_1},\quad \Phi_{2x}=\frac{\partial F_1}{\partial\Psi_2},\quad \Psi_{1x}=-\frac{\partial F_1}{\partial\Phi_1},\quad \Psi_{2x}=-\frac{\partial F_1}{\partial\Phi_2}, \tag{4.10}$$

with the Hamiltonian

$$F_1=\langle\Lambda^2\Psi_2,\Phi_2\rangle-\langle\Lambda^2\Psi_1,\Phi_1\rangle-\tfrac{1}{2}\langle\Lambda\Psi_2,\Phi_1\rangle\langle\Lambda\Psi_1,\Phi_2\rangle.$$

Under the constraint (4.9) and the FDHS (4.10), the binary t_2-constrained flow obtained from (4.3) with $V^{(2)}$ given by (4.5) and its adjoint equation for N distinct real numbers λ_j can also be written as a FDHS,

$$\Phi_{1,t_2}=\frac{\partial F_2}{\partial\Psi_1},\quad \Phi_{2,t_2}=\frac{\partial F_2}{\partial\Psi_2},\quad \Psi_{1,t_2}=-\frac{\partial F_2}{\partial\Phi_1},\quad \Psi_{2,t_2}=-\frac{\partial F_2}{\partial\Phi_2}, \tag{4.11}$$

with the Hamiltonian

$$\begin{aligned}F_2=&-\langle\Lambda^4\Psi_2,\Phi_2\rangle+\langle\Lambda^4\Psi_1,\Phi_1\rangle+\tfrac{1}{2}\langle\Lambda\Psi_2,\Phi_1\rangle\langle\Lambda^3\Psi_1,\Phi_2\rangle+\tfrac{1}{2}\langle\Lambda^3\Psi_2,\Phi_1\rangle\langle\Lambda\Psi_1,\Phi_2\rangle\\ &-\tfrac{1}{32}\langle\Lambda\Psi_2,\Phi_1\rangle^2\langle\Lambda\Psi_1,\Phi_2\rangle^2+\tfrac{1}{8}(\langle\Lambda^2\Psi_2,\Phi_2\rangle-\langle\Lambda^2\Psi_1,\Phi_1\rangle)\langle\Lambda\Psi_2,\Phi_1\rangle\langle\Lambda\Psi_1,\Phi_2\rangle.\end{aligned}$$

The Lax representation for the FDHSs (4.10) and (4.11) are presented by (2.18) with the entries of the Lax matrix M given by

$$A(\lambda)=1+\frac{1}{4}\sum_{j=1}^{N}\frac{\lambda_j^2(\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j})}{\lambda^2-\lambda_j^2}, \tag{4.12a}$$

$$B(\lambda)=\frac{1}{2}\lambda\sum_{j=1}^{N}\frac{\lambda_j\psi_{2j}\phi_{1j}}{\lambda^2-\lambda_j^2},\quad C(\lambda)=\frac{1}{2}\lambda\sum_{j=1}^{N}\frac{\lambda_j\psi_{1j}\phi_{2j}}{\lambda^2-\lambda_j^2}. \tag{4.12b}$$

A straightforward calculation yields

$$A^2(\lambda)+B(\lambda)C(\lambda)\equiv P(\lambda)=1+\sum_{j=1}^{N}\left[\frac{P_j}{\lambda^2-\lambda_j^2}+\frac{\lambda_j^4P_{N+j}^2}{(\lambda^2-\lambda_j^2)^2}\right], \tag{4.13}$$

where P_j, $j=1,...,2N$, are $2N$ independent integrals of motion for the FDIHSs (4.10) and (4.11),

$$P_j=-\frac{1}{2}\lambda_j^2(\psi_{2j}\phi_{2j}-\psi_{1j}\phi_{1j})+\frac{1}{8}\langle\Lambda\Psi_2,\Phi_1\rangle\lambda_j\psi_{1j}\phi_{2j}+\frac{1}{8}\langle\Lambda\Psi_1,\Phi_2\rangle\lambda_j\psi_{2j}\phi_{1j}+\frac{1}{8}$$
$$\times\sum_{k\neq j}\frac{1}{\lambda_j^2-\lambda_k^2}[\lambda_j^2\lambda_k^2(\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j})(\psi_{1k}\phi_{1k}-\psi_{2k}\phi_{2k})+2\lambda_j\lambda_k(\lambda_j^2+\lambda_k^2)\psi_{1j}\phi_{2j}\psi_{2k}\phi_{1k}],$$
$$j=1,...,N, \tag{4.14a}$$

$$P_{N+j}=\tfrac{1}{4}(\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}),\quad j=1,...,N. \tag{4.14b}$$

It is easy to verify that

$$F_1=-2\sum_{j=1}^{N}P_j,\quad F_2=2\sum_{j=1}^{N}(\lambda_j^2P_j+\lambda_j^4P_{N+j}^2)-\frac{1}{2}\left(\sum_{j=1}^{N}P_j\right)^2, \tag{4.15a}$$

$$\langle\Psi_2,\Phi_2\rangle+\langle\Psi_1,\Phi_1\rangle=4\sum_{j=1}^{N}P_{N+j}. \tag{4.15b}$$

By inserting $\lambda=0$, (4.13) leads to

$$1+\frac{1}{4}(\langle\Psi_2,\Phi_2\rangle-\langle\Psi_1,\Phi_1\rangle)=\sqrt{P(0)}=\sqrt{1+\sum_{j=1}^{N}[-P_j\lambda_j^{-2}+P_{N+j}^2]}. \tag{4.16}$$

With respect to the standard Poisson bracket it is found that

$$\{A(\lambda),A(\mu)\}=\{B(\lambda),B(\mu)\}=\{C(\lambda),C(\mu)\}=0, \tag{4.17a}$$

$$\{A(\lambda),B(\mu)\}=\frac{\mu}{2(\lambda^2-\mu^2)}[\mu B(\mu)-\lambda B(\lambda)], \tag{4.17b}$$

$$\{A(\lambda),C(\mu)\}=\frac{\mu}{2(\lambda^2-\mu^2)}[\lambda C(\lambda)-\mu C(\mu)], \tag{4.17c}$$

$$\{B(\lambda),C(\mu)\}=\frac{\lambda\mu}{\lambda^2-\mu^2}[A(\mu)-A(\lambda)]. \tag{4.17d}$$

Then $\{A^2(\lambda)+B(\lambda)C(\lambda),\ A^2(\mu)+B(\mu)C(\mu)\}=0$ implies that P_j, $j=1,...,2N$, are in involution. The CDNS equations (4.6) are factorized by the x-FDIHS (4.10) and the t_2-FDIHS (4.11), namely, if $(\Psi_1,\Psi_2,\Phi_1,\Phi_2)$ satisfies the FDIHSs (4.10) and (4.11) simultaneously, then $(q,\ r)$ given by (4.9) solves the CDNS equations (4.6). The factorization of the nth Kaup–Newell equations (4.4) will be presented at the end of Sec. IV B.

B. The separation of variables for the Kaup–Newell equations

Since the commutator relations (4.17) are quite different from (2.23) and (3.16), we have to modify a little bit of the method presented in Secs. II and III. Let us denote $\tilde{\lambda}=\lambda^2$, $\tilde{\lambda}_j=\lambda_j^2$. The entries of the Lax matrix M given by (4.12) can be rewritten as

$$A(\tilde{\lambda})=1+\tfrac{1}{4}(\langle\Psi_2,\Phi_2\rangle-\langle\Psi_1,\Phi_1\rangle)+\tfrac{1}{2}\tilde{\lambda}A_1(\tilde{\lambda}),\quad B(\tilde{\lambda})=\tfrac{1}{2}\sqrt{\tilde{\lambda}}\overline{B}(\tilde{\lambda}), \tag{4.18a}$$

where

$$A_1(\overline{\lambda})=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}-\psi_{2j}\phi_{2j}}{\tilde{\lambda}-\tilde{\lambda}_j},\quad \overline{B}(\tilde{\lambda})=\sum_{j=1}^{N}\frac{\sqrt{\tilde{\lambda}_j}\psi_{2j}\phi_{1j}}{\tilde{\lambda}-\tilde{\lambda}_j}. \tag{4.18b}$$

It is easy to see that

$$\{A_1(\tilde{\lambda}),A_1(\tilde{\mu})\}=\{\overline{B}(\tilde{\lambda}),\overline{B}(\tilde{\mu})\}=0, \tag{4.19a}$$

$$\{A_1(\tilde{\lambda}),\overline{B}(\tilde{\mu})\}=\frac{1}{\tilde{\lambda}-\tilde{\mu}}[\overline{B}(\tilde{\mu})-\overline{B}(\tilde{\lambda})]. \tag{4.19b}$$

It follows from (4.16) and (4.18a) that

$$A(\tilde{\lambda})=\sqrt{1+\sum_{j=1}^{N}[-P_j\tilde{\lambda}_j^{-1}+P_{N+j}^2]}+\tfrac{1}{2}\tilde{\lambda}A_1(\tilde{\lambda}). \tag{4.19c}$$

The commutator relations (4.19) and the generating function of integrals of motion (4.13) enable us to introduce $u_1,\dots,u_N$ in the following way:

$$\overline{B}(\tilde{\lambda})=\sum_{j=1}^{N}\frac{\sqrt{\tilde{\lambda}_j}\psi_{2j}\phi_{1j}}{\tilde{\lambda}-\tilde{\lambda}_j}=e^{u_N}\frac{R(\tilde{\lambda})}{K(\tilde{\lambda})}, \tag{4.20}$$

where

$$R(\tilde{\lambda})=\prod_{k=1}^{N-1}(\tilde{\lambda}-u_k),\quad K(\tilde{\lambda})=\prod_{k=1}^{N}(\tilde{\lambda}-\tilde{\lambda}_k),$$

and $v_1,\dots,v_N$ by $A_1(\tilde{\lambda})$.

By the exactly same argument as in Secs. II and III, we construct $\tilde{A}(\tilde{\lambda})$ and $\tilde{B}(\tilde{\lambda})$ by

$$\tilde{A}(\tilde{\lambda})=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}}{\tilde{\lambda}-\tilde{\lambda}_j}=2\sum_{j=1}^{N}\frac{P_{N+j}}{\tilde{\lambda}-\tilde{\lambda}_j}, \tag{4.21}$$

$$\tilde{B}(\tilde{\lambda})=\sum_{j=1}^{N}\frac{\phi_{1j}^2}{\tilde{\lambda}-\tilde{\lambda}_j}, \tag{4.22}$$

and, for the same reason, we have to replace $A_1(\tilde{\lambda})$ by $\overline{A}(\tilde{\lambda})$,

$$\overline{A}(\tilde{\lambda})\equiv A_1(\tilde{\lambda})-\tilde{A}(\tilde{\lambda})=-\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{2j}}{\tilde{\lambda}-\tilde{\lambda}_j}. \tag{4.23}$$

Then we have the following proposition.

Proposition 5: Assume that $\lambda_j, \phi_{ij}, \psi_{ij} \in \mathbf{R}$, $i=1,2$, $j=1,...,N$. *Introduce the separated variables* $u_1,\ldots,u_{2N}$ *by the* $\overline{B}(\tilde{\lambda})$ *and* $\tilde{B}(\tilde{\lambda})$:

$$\overline{B}(\tilde{\lambda})=\sum_{j=1}^{N}\frac{\sqrt{\tilde{\lambda}_j}\psi_{2j}\phi_{1j}}{\tilde{\lambda}-\tilde{\lambda}_j}=e^{u_N}\frac{R(\tilde{\lambda})}{K(\tilde{\lambda})}, \tag{4.24a}$$

$$\tilde{B}(\tilde{\lambda})=\sum_{j=1}^{N}\frac{\phi_{1j}^2}{\tilde{\lambda}-\tilde{\lambda}_j}=e^{u_{2N}}\frac{\overline{R}(\tilde{\lambda})}{K(\tilde{\lambda})}, \tag{4.24b}$$

with

$$R(\tilde{\lambda})=\prod_{k=1}^{N-1}(\tilde{\lambda}-u_k), \quad \overline{R}(\tilde{\lambda})=\prod_{k=1}^{N-1}(\tilde{\lambda}-u_{N+k}),$$

and $v_1,\ldots,v_{2N}$ *by*

$$v_k=\overline{A}(u_k)=A_1(u_k)-\tilde{A}(u_k)=-\sum_{j=1}^{N}\frac{\psi_{2j}\phi_{2j}}{u_k-\tilde{\lambda}_j}, \quad k=1,...,N-1, \tag{4.24c}$$

$$v_N=-\langle\Psi_2,\Phi_2\rangle, \tag{4.24d}$$

$$v_{N+k}=\overline{A}(u_{N+k})=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\phi_{1j}+\psi_{2j}\phi_{2j}}{u_{N+k}-\tilde{\lambda}_j}, \quad k=1,...,N-1, \tag{4.24e}$$

$$v_{2N}=\tfrac{1}{2}(\langle\Psi_1,\Phi_1\rangle+\langle\Psi_2,\Phi_2\rangle). \tag{4.24f}$$

If $u_1,\ldots,u_N$, *are single zeros of* $\overline{B}(\lambda)$, *then* $v_1,\ldots,v_{2N}$ *and* $u_1,\ldots,u_{2N}$ *are canonically conjugated, i.e., they satisfy* (1.1).

Proof: It follows from (4.24a) and (4.24b) that

$$u_N=\ln\langle\Lambda\Psi_2,\Phi_1\rangle, \tag{4.25}$$

$$u_{2N}=\ln\langle\Psi_1,\Phi_1\rangle. \tag{4.26}$$

By a straightforward calculation, it is found that $\overline{B}(\tilde{\lambda})$, $\overline{A}(\tilde{\lambda})$, $\tilde{B}(\tilde{\lambda})$, $\tilde{A}(\tilde{\lambda})$ satisfy the commutator relations (2.33) with λ, μ replaced by $\tilde{\lambda}$, $\tilde{\mu}$, as well as the following commutator relations:

$$\{u_N,\overline{B}(\mu)\}=\{u_N,\tilde{B}(\mu)\}=\{u_N,\tilde{A}(\mu)\}=0, \quad \{u_N,\overline{A}(\mu)\}=-\frac{\overline{B}(\mu)}{\langle\Lambda\Psi_2,\Phi_1\rangle}, \tag{4.27a}$$

$$\{v_N,\overline{A}(\mu)\}=\{v_N,\tilde{B}(\mu)\}=\{v_N,\tilde{A}(\mu)\}=0, \quad \{v_N,\overline{B}(\mu)\}=\overline{B}(\mu), \tag{4.27b}$$

$$\{u_{2N},\overline{B}(\mu)\}=\{u_{2N},\overline{A}(\mu)\}=\{u_{2N},\tilde{B}(\mu)\}=0, \quad \{u_{2N},\tilde{A}(\mu)\}=-\frac{\tilde{B}(\mu)}{\langle\Psi_1,\Phi_1\rangle}, \tag{4.27c}$$

$$\{v_{2N},\overline{B}(\mu)\}=\{v_{2N},\overline{A}(\mu)\}=\{v_{2N},\tilde{A}(\mu)\}=0, \quad \{v_{2N},\tilde{B}(\mu)\}=\tilde{B}(\mu), \tag{4.27d}$$

$$\{v_N,u_N\}=1, \quad \{v_{2N},u_{2N}\}=1, \tag{4.27e}$$

$$\{u_{2N},u_N\}=\{u_{2N},v_N\}=\{v_{2N},u_N\}=\{v_{2N},v_N\}=0. \tag{4.27f}$$

Then in the exactly same way as for propositions 1 and 3, we can complete the proof.
It follows from (4.24a) and (4.24b) that

$$\lambda_j\psi_{2j}\phi_{1j}=e^{u_N}\frac{R(\lambda_j^2)}{K'(\lambda_j^2)},\quad \phi_{1j}^2=e^{u_{2N}}\frac{\overline{R}(\lambda_j^2)}{K'(\lambda_j^2)},\quad j=1,...,N, \tag{4.28}$$

or

$$\phi_{1j}=\sqrt{\frac{e^{u_{2N}}\overline{R}(\lambda_j^2)}{K'(\lambda_j^2)}},\quad \psi_{2j}=\frac{e^{u_N}R(\lambda_j^2)}{\lambda_j\sqrt{e^{u_{2N}}\overline{R}(\lambda_j^2)K'(\lambda_j^2)}},\quad j=1,...,N. \tag{4.29}$$

Equations (4.9) and (4.25) result in

$$q=-\tfrac{1}{2}e^{u_N}. \tag{4.30}$$

We now present the separated equations. By substituting u_k into (4.13), u_{N+k} into (4.21), and using (4.19c), (4.24c), and (4.24e), one gets the separated equations

$$v_k=A_1(u_k)-\widetilde{A}(u_k)=\frac{2}{u_k}[\sqrt{\widetilde{P}(u_k)}-\sqrt{P(0)}]-2\sum_{j=1}^{N}\frac{P_{N+j}}{u_k-\lambda_j^2},\quad k=1,...,N-1, \tag{4.31a}$$

$$v_{N+k}=\widetilde{A}(u_{N+k})=2\sum_{j=1}^{N}\frac{P_{N+j}}{u_{N+k}-\lambda_j^2},\quad k=1,...,N-1, \tag{4.31b}$$

where $P(0)$ are given by (4.16) and

$$\widetilde{P}(\tilde{\lambda})=1+\sum_{j=1}^{N}\left[\frac{P_j}{\tilde{\lambda}-\lambda_j^2}+\frac{\lambda_j^4P_{N+j}^2}{(\tilde{\lambda}-\lambda_j^2)^2}\right].$$

It follows from (4.15b), (4.16), (4.24d), and (4.24f) that

$$v_N=2-2\sqrt{P(0)}-2\sum_{i=1}^{N}P_{N+i},\quad v_{2N}=2\sum_{i=1}^{N}P_{N+i}. \tag{4.31c}$$

Replacing v_k by the partial derivative $\partial S/\partial u_k$ of the generating function S of the canonical transformation and interpreting the P_i as integration constants, Eqs. (4.31a)–(4.31c) may be integrated to give the generating function of the canonical transformation

$$\begin{aligned}S(u_1,\dots,u_{2N})=&\sum_{k=1}^{N-1}\left[\int^{u_k}\left(\frac{2}{\tilde{\lambda}}\sqrt{\widetilde{P}(\tilde{\lambda})}-\frac{2}{\tilde{\lambda}}\sqrt{P(0)}-\widetilde{A}(\tilde{\lambda})\right)d\tilde{\lambda}+\int^{u_{N+k}}\widetilde{A}(\tilde{\lambda})d\tilde{\lambda}\right]\\&+\left(2-2\sqrt{P(0)}-2\sum_{i=1}^{N}P_{N+i}\right)u_N+2\sum_{i=1}^{N}P_{N+i}u_{2N}\\=&\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{2}{\tilde{\lambda}}\sqrt{\widetilde{P}(\tilde{\lambda})}d\tilde{\lambda}-2\sqrt{P(0)}\ln|u_k|-2\sum_{i=1}^{N}P_{N+i}\ln\left|\frac{u_k-\lambda_i^2}{u_{N+k}-\lambda_i^2}\right|\right]\end{aligned}$$

$$+\left(2-2\sqrt{P(0)}-2\sum_{i=1}^{N} P_{N+i}\right)u_N+2\sum_{i=1}^{N} P_{N+i}u_{2N}.$$

The linearizing coordinates are then

$$Q_i=\frac{\partial S}{\partial P_i}=\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{1}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}+\frac{1}{\lambda_i^2\sqrt{P(0)}}\ln|u_k|\right]+\frac{1}{\lambda_i^2\sqrt{P(0)}}u_N,\quad i=1,\ldots,N, \tag{4.33a}$$

$$Q_{N+i}=\frac{\partial S}{\partial P_{N+i}}=\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{2\lambda_i^4P_{N+i}}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)^2\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}-\frac{2P_{N+i}}{\sqrt{P(0)}}\ln|u_k|\right.$$
$$\left.-2\ln\left|\frac{u_k-\lambda_i^2}{u_{N+k}-\lambda_i^2}\right|\right]-2\left(\frac{P_{N+i}}{\sqrt{P(0)}}+1\right)u_N+2u_{2N},\quad i=1,\ldots,N. \tag{4.33b}$$

By using (4.15a), the linear flow induced by (4.10) together with Eqs. (4.33a) and (4.33b) leads to the Jacobi inversion problem for the FDIHS (4.10),

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{1}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}+\frac{1}{\lambda_i^2\sqrt{P(0)}}\ln|u_k|\right]+\frac{1}{\lambda_i^2\sqrt{P(0)}}u_N=\gamma_i-2x,$$
$$i=1,\ldots,N, \tag{4.34a}$$

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{2\lambda_i^4P_{N+i}}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)^2\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}-\frac{2P_{N+i}}{\sqrt{P(0)}}\ln|u_k|\right]-2\ln\left|\frac{u_k-\lambda_i^2}{u_{N+k}-\lambda_i^2}\right|\Bigg]-2\left(\frac{P_{N+i}}{\sqrt{P(0)}}+1\right)u_N+2u_{2N}$$
$$=\gamma_{N+i},\quad i=1,\ldots,N. \tag{4.34b}$$

By using (4.15a), the linear flow induced by (4.11) and Eqs. (4.34a) and (4.34b) yield the Jacobi inversion problem for the FDIHS (4.11),

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{1}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}+\frac{1}{\lambda_i^2\sqrt{P(0)}}\ln|u_k|\right]+\frac{1}{\lambda_i^2\sqrt{P(0)}}u_N$$
$$=\overline{\gamma_i}+\left(2\lambda_i^2-\sum_{k=1}^{N}P_k\right)t_2,\quad i=1,\ldots,N, \tag{4.35a}$$

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{2\lambda_i^4P_{N+i}}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)^2\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}-\frac{2P_{N+i}}{\sqrt{P(0)}}\ln|u_k|\right]-2\ln\left|\frac{u_k-\lambda_i^2}{u_{N+k}-\lambda_i^2}\right|\Bigg]-2\left(\frac{P_{N+i}}{\sqrt{P(0)}}+1\right)u_N+2u_{2N}$$
$$=\overline{\gamma}_{N+i}+4\lambda_i^4P_{N+i}t_2,\quad i=1,\ldots,N. \tag{4.35b}$$

Finally, since the CDNS equations (4.6) are factorized by the FDIHS (4.10) and (4.11), combining Eqs. (4.34) and (4.35) gives rise to the Jacobi inversion problem for the CDNS equations (4.6),

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{1}{\tilde\lambda(\tilde\lambda-\lambda_i^2)\sqrt{\tilde P(\tilde\lambda)}}d\tilde\lambda+\frac{1}{\lambda_i^2\sqrt{P(0)}}\ln|u_k|\right]+\frac{1}{\lambda_i^2\sqrt{P(0)}}u_N$$

$$=\gamma_i-2x+\left(2\lambda_i^2-\sum_{k=1}^{N}P_k\right)t_2,\quad i=1,\dots,N, \tag{4.36a}$$

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{2\lambda_i^4P_{N+i}}{\tilde\lambda(\tilde\lambda-\lambda_i^2)^2\sqrt{\tilde P(\tilde\lambda)}}d\tilde\lambda-\frac{2P_{N+i}}{\sqrt{P(0)}}\ln|u_k|\right]-2\ln\left|\frac{u_k-\lambda_i^2}{u_{N+k}-\lambda_i^2}\right|\Bigg]-2\left(\frac{P_{N+i}}{\sqrt{P(0)}}+1\right)u_N+2u_{2N}$$

$$=\gamma_{N+i}+4\lambda_i^4P_{N+i}t_2,\quad i=1,\dots,N. \tag{4.36b}$$

If ϕ_{1j}, ψ_{2j}, q defined by (4.29) and (4.30) can be solved from (4.36) by using the Jacobi inversion technique, then ϕ_{2j}, ψ_{1j} can be obtained from the first equation and the last equation in (4.10), respectively. Finally q and $r=-\langle\Lambda\Psi_1,\Phi_2\rangle$ provides the solution to the CDNS equations (4.6).

In general, the above mentioned procedure can be applied to the whole Kaup–Newell hierarchy (4.4). Set

$$A^2(\lambda)+B(\lambda)C(\lambda)=\sum_{k=0}^{\infty}\tilde F_k\lambda^{-2k}, \tag{4.37a}$$

where $\tilde F_k$, $k=1,2,\dots$, are also integrals of motion for both the x-FDHSs (4.10) and the t_n-binary constrained flows (2.16). Comparing (4.37a) with (4.13), one gets

$$\tilde F_0=1,\quad \tilde F_k=\sum_{j=1}^{N}[\lambda_j^{2k-2}P_j+(k-1)\lambda_j^{2k}P_{N+j}^2],\quad k=1,2,\dots . \tag{4.37b}$$

By employing the method in Refs. 34 and 35, the t_n-FDIHS obtained from the t_n-constrained flow is of the form

$$\Phi_{1,t_n}=\frac{\partial F_n}{\partial\Psi_1},\quad \Phi_{2,t_n}=\frac{\partial F_n}{\partial\Psi_2},\quad \Psi_{1,t_n}=-\frac{\partial F_n}{\partial\Phi_1},\quad \Psi_{2,t_n}=-\frac{\partial F_n}{\partial\Phi_2}, \tag{4.38a}$$

with the Hamiltonian

$$F_n=2\sum_{m=0}^{n-1}\left(-\frac{1}{2}\right)^m\frac{\alpha_m}{m+1}\sum_{l_1+\cdots+l_{m+1}=n}\tilde F_{l_1}\cdots\tilde F_{l_{m+1}}, \tag{4.38b}$$

where $l_1\geqslant1,\dots,l_{m+1}\geqslant1$, and α_m are given by (2.25c). Since the nth Kaup–Newell equations (4.4) is factorized by the x-FDIHS (4.10) and the t_n-FDIHS (4.38). We have the following proposition.

Proposition 6: The Jacobi inversion problem for the nth Kaup–Newell equations (4.4) is given by

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{1}{\tilde\lambda(\tilde\lambda-\lambda_i^2)\sqrt{\tilde P(\tilde\lambda)}}d\tilde\lambda+\frac{1}{\lambda_i^2\sqrt{P(0)}}\ln|u_k|\right]+\frac{1}{\lambda_i^2\sqrt{P(0)}}u_N$$

$$=\gamma_i-2x+2t_n\sum_{m=0}^{n-1}(-\tfrac{1}{2})^m\alpha_m\sum_{l_1+\cdots l_{m+1}=n}\lambda_i^{2l_{m+1}-2}\tilde F_{l_1}\cdots\tilde F_{l_m},\quad i=1,\dots,N, \tag{4.39a}$$

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{2\lambda_i^4 P_{N+i}}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)^2\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}-\frac{2P_{N+i}}{\sqrt{P(0)}}\ln|u_k|\right]-2\ln\left|\frac{u_k-\lambda_i^2}{u_{N+k}-\lambda_i^2}\right|\Bigg]-2\left(\frac{P_{N+i}}{\sqrt{P(0)}}+1\right)u_N+2u_{2N}$$
$$=\gamma_{N+i}+4t_n\sum_{m=0}^{n-1}\left(-\frac{1}{2}\right)^m\alpha_m\sum_{l_1+\cdots l_{m+1}=n}(l_{m+1}-1)\lambda_i^{2l_{m+1}}P_{N+i}\tilde{F}_{l_1}\cdots\tilde{F}_{l_m},\quad i=1,\ldots,N, \tag{4.39b}$$

where $l_1\geqslant 1,\ldots,\ l_{m+1}\geqslant 1$, *and* $\tilde{F}_{l_1},\ldots\tilde{F}_{l_m}$, *are given by* (4.37b).

For example, the third equations in the Kaup–Newell hierarchy with $n=3$ are of the form

$$q_{t_3}=-\tfrac{1}{4}q_{xxx}-\tfrac{3}{8}(q^3r^2+2qrq_x)_x,\quad r_{t_3}=-\tfrac{1}{4}r_{xxx}-\tfrac{3}{8}(r^3q^2-2qrr_x)_x. \tag{4.40}$$

The Kaup–Newell equations (4.40) can be factorized by the x-FDIHS (4.10) and t_3-FDIHS with the Hamiltonian F_3 defined by

$$F_3=\sum_{j=1}^{N}(2\lambda_j^4P_j+4\lambda_j^6P_{N+j}^2)-\left[\sum_{j=1}^{N}(\lambda_j^2P_j+\lambda_j^4P_{N+j}^2)\right]\sum_{j=1}^{N}P_j+\frac{1}{4}\left(\sum_{j=1}^{N}P_j\right)^3. \tag{4.41}$$

The Jacobi inversion problem for Eq. (4.40) is given by

$$\sum_{k=1}^{N-1}\left[\int^{u_k}\frac{1}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}+\frac{1}{\lambda_i^2\sqrt{P(0)}}\ln|u_k|\right]+\frac{1}{\lambda_i^2\sqrt{P(0)}}u_N$$
$$=\gamma_i-2x+\left[2\lambda_i^4-\sum_{j=1}^{N}(\lambda_j^2P_j+\lambda_i^2P_j+\lambda_j^4P_{N+j}^2)+\frac{3}{4}\left(\sum_{j=1}^{N}P_j\right)^2\right]t_3,\quad i=1,\ldots,N.$$

$$\sum_{k=1}^{N-1}\left[\left[\int^{u_k}\frac{2\lambda_i^4P_{N+i}}{\tilde{\lambda}(\tilde{\lambda}-\lambda_i^2)^2\sqrt{\tilde{P}(\tilde{\lambda})}}d\tilde{\lambda}-\frac{2P_{N+i}}{\sqrt{P(0)}}\ln|u_k|\right]-2\ln\left|\frac{u_k-\lambda_i^2}{u_{N+k}-\lambda_i^2}\right|\right]$$
$$-2\left(\frac{P_{N+i}}{\sqrt{P(0)}}+1\right)u_N+2u_{2N}$$
$$=\gamma_{N+i}+\left[8\lambda_i^6P_{N+i}-2\lambda_i^4P_{N+j}\sum_{j=1}^{N}P_j\right]t_3,\quad i=1,\ldots,N.$$

V. CONCLUDING REMARKS

This paper proposed a new method to find the additional N pairs canonical separated variables for the separation of variables for binary constrained flows of soliton hierarchies. For a certain kind of integrable models, a general approach to the separation of variables was proposed in Refs. 2–4 by taking the poles of the properly normalized Baker–Akhiezer function and the corresponding eigenvalues of the Lax operator as separated variables. As pointed out in Ref. 4, there is no guarantee that the separated variables so constructed satisfy the canonical conditions (1.1). The method proposed in this paper is to start directly from the canonical conditions (1.1) and the requirement for the separated equations (1.2). We introduced $2N$ pairs of separated variables by means of four functions $\bar{B}(\lambda)$, $\bar{A}(\lambda)$, $\tilde{B}(\lambda)$, and $\tilde{A}(\lambda)$, which are constructed in such way that they satisfy certain commutator relations required by the canonical conditions (1.1) and $\bar{A}(\lambda)$ and $\tilde{A}(\lambda)$ are linked to the generating functions of the integrals of motion for the models. This method ensures that the separated variables are canonically conjugated. We produced two sets of separated equations directly from two generating functions of integrals of motion. It seems that the separated equations are intimately connected with the generating functions of integrals of motion.

The finite gap solutions or finite-dimensional quasiperiodic solutions for the KdV equation was studied in Ref. 36 by means of the stationary equation of the KdV hierarchy called the Lax–Novikov equation. By the standard Jacobi inversion technique,[19] the finite-dimensional quasiperiodic solution can be given in an explicit form in terms of the Riemann theta functions associated with the invariant spectral curve. The Jacobi inversion problem (2.45) for the KdV equation is somewhat different from that in Ref. 36 due to some additional terms. The Jacobi inversion problems for some binary constrained flows require the development of the standard Jacobi inversion technique in order to solve them explicitly.

ACKNOWLEDGMENTS

This work was supported by the City University of Hong Kong and the Research Grants Council of Hong Kong and the Chinese Basic Research Project ''Nonlinear Science.'' One of the authors (Y.B.Z.) wishes to express his gratitude to Department of Mathematics of the City University of Hong Kong for warm hospitality.

[1] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics*, 2nd ed. (Springer, New York, 1994).
[2] E. K. Sklyanin, ''Separation of variables in the Gaudin model,'' J. Sov. Math. **47**, 2473–2488 (1989).
[3] V. B. Kuznetsov, ''Quadrics on real Riemannian spaces of constant curvature: Separation of variables and connection with Gaudin magnet,'' J. Math. Phys. **33**, 3240–3254 (1992).
[4] E. K. Sklyanin, ''Separation of variables,'' Prog. Theor. Phys. Suppl. **118**, 35–60 (1995).
[5] O. Babelon and M. Talon, ''Separation of variables for the classical and quantum Neumann model,'' Nucl. Phys. B **379**, 321–339 (1992).
[6] J. C. Eilbeck, V. Z. Enol'skii, V. B. Kuznetsov, and A. V. Tsiganov, ''Linear *r*-matrix algebra for classical separable systems,'' J. Phys. A **27**, 567–578 (1994).
[7] E. G. Kalnins, V. B. Kuznetsov, and W. Miller, Jr., ''Quadrics on complex Riemannian spaces of constant curvature, separation on variables, and the Gaudin magnet,'' J. Math. Phys. **35**, 1710–1731 (1994).
[8] J. Harnad and P. Winternitz, ''Classical and quantum integrable systems in $\widetilde{gl}(2)^+$ and separation of variables,'' Commun. Math. Phys. **172**, 263–285 (1995).
[9] P. P. Kulish, S. Rauch-Wojciechowski, and A. V. Tsiganov, ''Stationary problems for equation of the KdV type and dynamical *r*-matrices,'' J. Math. Phys. **37**, 3463–3482 (1996).
[10] Y. B. Zeng, ''The separability and dynamical *r*-matrix for the constrained flows of the Jaulent–Miodek hierarchy,'' Phys. Lett. A **216**, 26–32 (1996).
[11] Y. B. Zeng, ''A family of separable Hamiltonian systems and their classical dynamical *r*-matrix,'' Inverse Probl. **12**, 1–13 (1996).
[12] Y. B. Zeng, ''Separation of variables for the constrained flows,'' J. Math. Phys. **38**, 321–329 (1997).
[13] M. R. Adams, J. Harnad, and J. Hurtubise, ''Darboux coordinates and Liouville–Arnold integration in loop algebras,'' Commun. Math. Phys. **155**, 385–413 (1993).
[14] M. R. Adams, J. Harnad, and J. Hurtubise, ''Liouville generating function for isospectral Hamiltonian flow in Loop algebras,'' in *Integrable and Superintegrable Systems*, edited by B. A. Kuperschmidt (World Scientific, Singapore, 1990).
[15] J. Harnad and M. A. Wisse, ''Isospectral flow in Loop algebras and quasiperiodic solution to the sine Gordon equation,'' J. Math. Phys. **34**, 3518–3526 (1993).
[16] M. A. Wisse, ''Darboux coordinates and isospectral Hamiltonian flow for the massive thirring model,'' Lett. Math. Phys. **28**, 287–294 (1993).
[17] Y. B. Zeng, ''Using factorization to solve soliton equation,'' J. Phys. A **30**, 3719–3724 (1997).
[18] Y. B. Zeng, ''The Jacobi inversion problem for soliton equations,'' J. Phys. Soc. Jpn. **66**, 2277–2282 (1997).
[19] B. A. Dubrovin, ''Theta functions and nonlinear equations,'' Russ. Math. Survey **36**, 11–92 (1981).
[20] I. M. Krichever and S. P. Novikov, ''Holomophic bundles over algebraic curves and nonlinear equations,'' Russ. Math. Surveys **32**, 53–79 (1980).
[21] M. Adler and P. van Moerbeke, ''Completely integrable systems, Euclidean Lie algebras and curve,'' Adv. Math. **38**, 267–317 (1980).
[22] M. Adler and P. van Moerbeke, ''Linearization of Hamiltonian systems, Jacobi variables and representation theory,'' Adv. Math. **38**, 318–379 (1980).
[23] O. Ragnisco and S. Rauch-Wojciechowski, ''Restricted flows of the AKNS hierarchy,'' Inverse Probl. **8**, 245–262 (1992).
[24] Y. B. Zeng, ''The higher-order constraint and integrable systems related to Boussinesq equation,'' Chin. Sci. Bull. **37**, 1937–1942 (1992).

[25] W. X. Ma and W. Strampp, ''An explicit symmetry constraint for the Lax pairs and the adjoint Lax pairs of AKNS systems,'' Phys. Lett. A **185**, 277–286 (1994).
[26] W. X. Ma, ''New finite-dimensional integrable systems by symmetry constraint of the KdV equations,'' J. Phys. Soc. Jpn. **64**, 1085–1091 (1995).
[27] W. X. Ma, B. Fuchssteiner, and W. Oevel, ''A 3×3 matrix spectral problem for AKNS hierarchy and its binary nonlinearization,'' Physica A **233**, 331–354 (1996).
[28] W. X. Ma, Q. Ding, W. G. Zhang, and B. Q. Lu, ''Binary nonlinearization of Lax pairs of Kaup–Newell soliton hierarchy,'' Nuovo Cimento B **111**, 1135–1149 (1996).
[29] Y. S. Li and W. X. Ma, ''Binary nonlinearization of AKNS spectral problem under higher-order symmetry constraints,'' Chaos Solitons Fractals (to be published).
[30] W. X. Ma and B. Fuchssteiner, ''Binary nonlinearization of Lax pairs,'' in *Nonlinear Physics*, edited by E. Alfinito, M. Boiti, L. Martina, and F. Pempinelli (World Scientific, Singapore, 1996), pp. 217–224.
[31] Y. B. Zeng and W. X. Ma, ''The construction of canonical separated variables for binary constrained AKNS flow,'' Physica A (to be published).
[32] A. C. Newell, *Solitons in Mathematics and Physics* (SIAM, Philadelphia, 1985).
[33] Y. B. Zeng and Y. S. Li, ''The deduction of the Lax representation for constrained flows from the adjoint representation,'' J. Phys. A **26**, L273–L278 (1993).
[34] Y. B. Zeng, ''New factorization of the Kaup–Newell hierarchy,'' Physica D **73**, 171–188 (1994).
[35] Y. B. Zeng, ''An approach to the deduction of the finite-dimensional integrability from the infinite-dimensional integrability,'' Phys. Lett. A **160**, 541–547 (1991).
[36] S. P. Novikov, ''A method for solving the periodic problem for the KdV equation and its generalizations,'' in *Solitons, Topics in Current Physics*, edited by R. Bullough and P. Caudrey (Springer, New York, 1980), pp. 325–338.
[37] M. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[38] D. J. Kaup and A. C. Newell, ''An exact solution for a derivative nonlinear Schrödinger equation,'' J. Math. Phys. **19**, 798–801 (1978).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Cosmological and spherically symmetric solutions with intersecting p-branes

V. D. Ivashchuk[a)] and V. N. Melnikov
Center for Gravitation and Fundamental Metrology, VNIIMS, 3-1 M. Ulyanovoy Str., Moscow, 117313, Russia

(Received 20 January 1999; accepted for publication 12 May 1999)

Multidimensional model describing the cosmological evolution and/or spherically symmetric configuration with $n+1$ Einstein spaces in the theory with several scalar fields and forms is considered. When electro-magnetic composite p-brane ansatz is adopted, n ''internal'' spaces are Ricci-flat, one space M_0 has a nonzero curvature, and all p-branes do not ''live'' in M_0, a class of exact solutions is obtained if certain block-orthogonality relations on p-brane vectors are imposed. A subclass of spherically symmetric solutions (containing nonextremal p-brane black holes) is considered. Post-Newtonian parameters are calculated. © *1999 American Institute of Physics.* [S0022-2488(99)00510-1]

I. INTRODUCTION

At present there exists a special interest to the so-called M- and F theories, etc.[1–4] These theories are ''supermembrane'' analogues of superstring models[5] in $D=11,12$, etc. The low-energy limit of these theories leads to models governed by the action

$$S=\int_M d^Dz\sqrt{|g|}\left\{R[g]-h_{\alpha\beta}g^{MN}\partial_M\varphi^\alpha\partial_N\varphi^\beta-\sum_{a\in\Delta}\frac{\theta_a}{n_a!}\exp[2\lambda_a(\varphi)](F^a)^2_g\right\}, \tag{1.1}$$

where $g=g_{MN}\,dz^M\otimes dz^N$ is the metric $(M,N=1,\dots,D)$, $\varphi=(\varphi^\alpha)\in\mathbb{R}^l$ is a vector from dilatonic scalar fields, $(h_{\alpha\beta})$ is a nondegenerate symmetric $l\times l$ matrix $(l\in\mathbb{N})$, $\theta_a=\pm 1$,

$$F^a=dA^a=\frac{1}{n_a!}F^a_{M_1\cdots M_{n_a}}dz^{M_1}\wedge\cdots\wedge dz^{M_{n_a}} \tag{1.2}$$

is a n_a-form $(n_a\geq 1)$ on a D-dimensional manifold M, Λ is cosmological constant and λ_a is a 1-form on $\mathbb{R}^l$: $\lambda_a(\varphi)=\lambda_{a\alpha}\varphi^\alpha$, $a\in\Delta$, $\alpha=1,\dots,l$. In (1.1) we denote $|g|=|\det(g_{MN})|$,

$$(F^a)^2_g=F^a_{M_1\cdots M_{n_a}}F^a_{N_1\cdots N_{n_a}}g^{M_1N_1}\cdots g^{M_{n_a}N_{n_a}}, \tag{1.3}$$

$a\in\Delta$, where Δ is some finite set. In the models with one time all $\theta_a=1$ when the signature of the metric is $(-1,+1,\dots,+1)$.

In Ref. 6 it was shown that after dimensional reduction on the manifold $M_*\times M_1\times\cdots\times M_n$ and when the composite p-brane ansatz is considered (for review see, for example, Refs. 6, 7, and 8) the problem is reduced to the gravitating self-interacting σ-model with certain constraints imposed. For electric p-branes see also Refs. 9, 10, and 11 (in Ref. 11 the composite electric case was considered). (For pure gravitational sector see Refs. 12 and 13) This representation may be considered as a powerful tool for obtaining different solutions with intersecting p-branes (analogs of membranes). In Refs. 6 and 14 the Majumdar–Papapetrou-type solutions were obtained (for the noncomposite electric case see Refs. 9 and 10 and for composite electric

[a)]Electronic mail: ivas@rgs.phys.msu.su

case see Ref. 11). These solutions correspond to Ricci-flat (M_i, g^i), $i=1,\ldots,n$, and were generalized also to the case of Einstein internal spaces.[6] Earlier some special classes of these solutions were also considered in Refs. 15–22. The obtained solutions take place, when certain orthogonality relations (on couplings parameters, dimensions of ''branes,'' total dimension) are imposed.

In cosmological (or spherically symmetric) case $M_* = \mathbb{R}$ and the problem is effectively reduced to Toda-like system with the Lagrangian[23]

$$L = \frac{1}{2}\bar{G}_{AB}\dot{x}^A\dot{x}^B - \sum_{s\in S_*} A_s \exp(2U_A^s x^A), \tag{1.4}$$

with the zero-energy constraint $E=0$, where $(\bar{G}_{AB})$ is a nondegenerate symmetric $N\times N$ matrix $(N=n+l)$, $A_s \neq 0$, $x=(x^A)\in\mathbb{R}^N$, $U^s=(U_A^s)\in\mathbb{R}^N$, $s\in S_*$. The considered cosmological model contains some stringy cosmological models (see, for example, Ref. 24). It may be obtained (at classical level) from multidimensional cosmological model with perfect fluid[25,26] as a special case.

The integrability of the Lagrange equations corresponding to (1.4) crucially depends upon the scalar products $(U^{s_1}, U^{s_2})=0$, $s_1, s_2 \in S_*$, where

$$(U,U') = \bar{G}^{AB} U_A U'_B, \tag{1.5}$$

$U,U' \in \mathbb{R}^N$, where $(\bar{G}^{AB}) = (\bar{G}_{AB})^{-1}$.

In the ''orthogonal'' case

$$(U^s, U^{s'}) = 0 \tag{1.6}$$

$s,s' \in S_*$, a class of cosmological and spherically symmetric solutions was obtained.[23] Special cases were also considered in Refs. 27–30. The solutions with the horizon were considered in details in Refs. 23, 31–35. We note that the ''orthogonality'' relation (1.6) is known in literature as a ''no-force'' condition.

In this paper we consider a more general ''block-orthogonal'' case:

$$S_* = S_1 \cup \ldots \cup S_k, \quad S_i \cap S_j = \emptyset, \quad i \neq j, \tag{1.7}$$

$S_i \neq \emptyset$, i.e., the set S is a union of k nonintersecting (nonempty) subsets $S_1, \ldots, S_k$, and

$$(U^s, U^{s'}) = 0 \tag{1.8}$$

for all $s\in S_i$, $s'\in S_j$, $i\neq j$; $i,j=1,\ldots,k$. According to (1.8) the set of vectors $(U^s, s\in S)$ has a block-orthogonal structure with respect to the scalar product (1.5): it splits into k mutually orthogonal blocks $(U^s, s\in S_i)$, $i=1,\ldots,k$. Here we find a class of special solutions to Lagrange equations corresponding to the Lagrangian (1.1) (see Appendix B). Using this result we find a family of cosmological and/or spherically symmetric solutions with composite electromagnetic p-branes (see Sec. IV) in the case when n ''internal'' spaces are Ricci-flat, one space M_0 has a nonzero curvature, and all p-branes do not ''live'' in M_0, if block-orthogonality relations (on p-brane vectors U^s) (1.7) and (1.8) are imposed. These solutions generalize the solutions from Ref. 23 with orthogonal set of vectors U^s. A special class of ''block-orthogonal'' solutions (with coinciding parameters ν_s inside blocks) was considered earlier in Ref. 36.

Recently a class of Majumdar–Papapetrou type solutions with p-branes was obtained in Ref. 14 (see also Ref. 37) for block-orthogonal case. For the solutions from Ref. 14 U^s vectors may be related to finite-dimensional Lie algebras or hyperbolic (Kac–Moody) algebras but not to affine (Kac–Moody) algebras.

We also note that there exists a large variety of p-brane Toda solutions (open or closed) when certain intersection rules are satisfied.[23] The scalar products $(U^s, U^{s'})$, $s,s' \in S_*$, in this case are governed by Cartan matrices of finite-dimensional or affine Lie algebras. For concrete exact solutions see also Refs. 27, 38, and 39.

In Sec. V we consider a subclass of spherically symmetric solutions. This subclass contains nonextremal p-brane black holes for special values of ''Kasner-like'' parameters. In Sec. VI we calculate post-Newtonian parameters β and γ (Eddington parameters) for the spherically symmetric solutions. These parameters may be useful for possible physical applications.

II. THE MODEL

The equations of motion corresponding to (1.1) have the following form:

$$R_{MN}-\tfrac{1}{2}g_{MN}R=T_{MN}, \tag{2.1}$$

$$\Delta[g]\varphi^{\alpha}-\sum_{a\in\Delta}\theta_a\frac{\lambda_a^{\alpha}}{n_a!}e^{2\lambda_a(\varphi)}(F^a)_g^2=0, \tag{2.2}$$

$$\nabla_{M_1}[g](e^{2\lambda_a(\varphi)}F^{a,M_1\cdots M_{n_a}})=0, \tag{2.3}$$

$a\in\Delta$; $\alpha=1,\ldots,l$. In (2.2) $\lambda_a^{\alpha}=h^{\alpha\beta}\lambda_{a\beta}$, where $(h^{\alpha\beta})$ is matrix inverse to $(h_{\alpha\beta})$. In (2.1)

$$T_{MN}=T_{MN}[\varphi,g]+\sum_{a\in\Delta}\theta_a e^{2\lambda_a(\varphi)}T_{MN}[F^a,g], \tag{2.4}$$

where

$$T_{MN}[\varphi,g]=h_{\alpha\beta}(\partial_M\varphi^{\alpha}\partial_N\varphi^{\beta}-\tfrac{1}{2}g_{MN}\partial_P\varphi^{\alpha}\partial^P\varphi^{\beta}), \tag{2.5}$$

$$T_{MN}[F^a,g]=\frac{1}{n_a!}\left[-\frac{1}{2}g_{MN}(F^a)_g^2+n_aF^a_{MM_2\ldots M_{n_a}}F_N^{a,M_2\ldots M_{n_a}}\right]. \tag{2.6}$$

In (2.2) and (2.3) $\Delta[g]$ and $\nabla[g]$ are Laplace–Beltrami and covariant derivative operators, respectively, corresponding to g.

Let us consider the manifold

$$M=\mathbb{R}\times M_0\times\cdots\times M_n \tag{2.7}$$

with the metric

$$g=w\mathrm{e}^{2\gamma(u)}du\otimes du+\sum_{i=0}^{n}\mathrm{e}^{2\phi^i(u)}g^i, \tag{2.8}$$

where $w=\pm1$, u is a distinguished coordinate which, by convention, will be called ''time;'' $g^i=g^i_{m_in_i}(y_i)dy_i^{m_i}\otimes dy_i^{n_i}$ is a metric on M_i satisfying the equation

$$R_{m_in_i}[g^i]=\xi_ig^i_{m_in_i}, \tag{2.9}$$

$m_i,n_i=1,\ldots,d_i$; $d_i=\dim M_i$, $\xi_i=\text{const}$, $i=0,\ldots,n$; $n\in\mathbb{N}$. Thus, (M_i,g^i) are Einstein spaces. The functions γ,ϕ^i: $(u_-,u_+)\to\mathbb{R}$ are smooth.

Remark: It is more correct to write in (2.8) $\hat{g}^i$ instead of g^i, where $\hat{g}^i=p_i^*g^i$ is the pullback of the metric g^i to the manifold M by the canonical projection: p_i: $M\to M_i$, $i=0,\ldots,n$. In what follows we omit ''hats'' for simplicity.

Each manifold M_i is assumed to be oriented and connected, $i=0,\ldots,n$. Then the volume d_i-form

$$\tau_i=\sqrt{|g^i(y_i)|}dy_i^1\wedge\cdots\wedge dy_i^{d_i}, \tag{2.10}$$

and the signature parameter

$$\varepsilon(i)=\text{sign}\,\det(g^i_{m_i n_i})=\pm 1 \tag{2.11}$$

are correctly defined for all $i=0,\ldots,n$.

Let

$$\Omega_0=\{\emptyset,\{0\},\{1\},\ldots,\{n\},\{0,1\},\ldots,\{0,1,\ldots,n\}\} \tag{2.12}$$

be a set of all subsets of

$$I_0\equiv\{0,\ldots,n\}. \tag{2.13}$$

Let $I=\{i_1,\ldots,i_k\}\in\Omega_0$, $i_1<\ldots<i_k$. We define a form

$$\tau(I)\equiv\tau_{i_1}\wedge\cdots\wedge\tau_{i_k}, \tag{2.14}$$

of rank

$$d(I)\equiv\sum_{i\in I} d_i, \tag{2.15}$$

and a corresponding p-brane submanifold

$$M_I\equiv M_{i_1}\times\cdots\times M_{i_k}, \tag{2.16}$$

where $p=d(I)-1$ $(\dim M_I=d(I))$. We also define ε-symbol

$$\varepsilon(I)\equiv\varepsilon(i_1)\cdots\varepsilon(i_k). \tag{2.17}$$

For $I=\emptyset$ we put $\tau(\emptyset)=\varepsilon(\emptyset)=1$, $d(\emptyset)=0$.

For fields of forms we adopt the following ‘‘composite electro-magnetic’’ ansatz

$$F^a=\sum_{I\in\Omega_{a,e}}\mathcal{F}^{(a,e,I)}+\sum_{J\in\Omega_{a,m}}\mathcal{F}^{(a,m,J)}, \tag{2.18}$$

where

$$\mathcal{F}^{(a,e,I)}=d\Phi^{(a,e,I)}\wedge\tau(I), \tag{2.19}$$

$$\mathcal{F}^{(a,m,J)}=\mathrm{e}^{-2\lambda_a(\varphi)}*(d\Phi^{(a,m,J)}\wedge\tau(J)), \tag{2.20}$$

$a\in\Delta$, $I\in\Omega_{a,e}$, $J\in\Omega_{a,m}$ and

$$\Omega_{a,e},\Omega_{a,m}\subset\Omega_0. \tag{2.21}$$

[For empty $\Omega_{a,v}=\emptyset$, $v=e,m$, we put $\Sigma_{\emptyset}=0$ in (2.18)]. In (2.20) $*=*[g]$ is the Hodge operator on (M,g).

For the potentials in (2.19) and (2.20) we put

$$\Phi^s=\Phi^s(u), \tag{2.22}$$

$s\in S$, where

$$S=S_e\sqcup S_m, \quad S_v\equiv\sqcup_{a\in\Delta}\{a\}\times\{v\}\times\Omega_{a,v}, \tag{2.23}$$

$v=e,m$. Here $\sqcup$ means the union of nonintersecting sets. The set S consists of elements $s=(a_s,v_s,I_s)$, where $a_s\in\Delta$, $v_s=e,m$, and $I_s\in\Omega_{a,v_s}$ are ''color,'' ''electro-magnetic,'' and ''brane'' indices, respectively.

For dilatonic scalar fields we put

$$\varphi^\alpha=\varphi^\alpha(u), \tag{2.24}$$

$\alpha=1,\ldots,l$.

From (2.19) and (2.20) we obtain the relations between dimensions of p-brane worldsheets and ranks of forms

$$d(I)=n_a-1, \quad I\in\Omega_{a,e}, \tag{2.25}$$

$$d(J)=D-n_a-1, \quad J\in\Omega_{a,m}, \tag{2.26}$$

in electric and magnetic cases, respectively.

III. LAGRANGE REPRESENTATION

Here, like in Ref. 23, we impose a restriction on p-brane configurations, or, equivalently, on $\Omega_{a,v}$. We assume that the energy momentum tensor (T_{MN}) has a block-diagonal structure [as it takes place for (g_{MN})]. Sufficient restrictions on $\Omega_{a,v}$ that guarantee a block-diagonality of (T_{MN}) are presented in Appendix A.

It follows from Ref. 6 (see Proposition 2 in Ref. 6) that the equations of motion (2.1)–(2.3) and the Bianchi identities

$$d\mathcal{F}^s=0, \quad s\in S \tag{3.1}$$

for the field configuration (2.8), (2.18)–(2.20), (2.22), (2.24) with the restrictions (A2), (A3) (from Appendix A) imposed are equivalent to equations of motion for Lagrange system with the Lagrangian

$$L=\frac{1}{2}\mathcal{N}\left\{G_{ij}\dot{\phi}^i\dot{\phi}^j+h_{\alpha\beta}\dot{\varphi}^\alpha\dot{\varphi}^\beta+\sum_{s\in S}\varepsilon_s\exp[-2U^s(\phi,\varphi)](\dot{\Phi}^s)^2-2\mathcal{N}^{-2}V(\phi)\right\}, \tag{3.2}$$

where $\dot{x}\equiv dx/du$,

$$V=V(\phi)=\frac{w}{2}\sum_{i=0}^{n}\xi_i d_i e^{-2\phi^i+2\gamma_0(\phi)} \tag{3.3}$$

is the potential with

$$\gamma_0(\phi)\equiv\sum_{i=0}^{n}d_i\phi^i, \tag{3.4}$$

and

$$\mathcal{N}=\exp(\gamma_0-\gamma)>0 \tag{3.5}$$

is the lapse function,

$$U^s=U^s(\phi,\varphi)=-\chi_s\lambda_{a_s}(\varphi)+\sum_{i\in I_s}d_i\phi^i, \tag{3.6}$$

$$\varepsilon_s = (-\varepsilon[g])^{(1-\chi_s)/2}\varepsilon(I_s)\theta_{a_s} \tag{3.7}$$

for $s=(a_s, v_s, I_s)\in S$, $\varepsilon[g]=\mathrm{sign}\det(g_{MN})$, [more explicitly (3.7) reads $\varepsilon_s=\varepsilon(I_s)\theta_{a_s}$ for $v_s=e$ and $\varepsilon_s=-\varepsilon[g]\varepsilon(I_s)\theta_{a_s}$, for $v_s=m$]

$$\chi_s=+1, \quad v_s=e, \tag{3.8}$$

$$\chi_s=-1, \quad v_s=m, \tag{3.9}$$

and

$$G_{ij}=d_i\delta_{ij}-d_id_j \tag{3.10}$$

are components of the ''pure cosmological'' minisupermetric; $i,j=0,\ldots,n$.[40]

Let $x=(x^A)=(\phi^i,\varphi^\alpha)$,

$$\bar{G}=\bar{G}_{AB}\,dx^A\otimes dx^B=G_{ij}\,d\phi^i\otimes d\phi^j+h_{\alpha\beta}\,d\varphi^\alpha\otimes d\varphi^\beta, \tag{3.11}$$

$$(\bar{G}_{AB})=\begin{pmatrix} G_{ij} & 0 \\ 0 & h_{\alpha\beta}\end{pmatrix}, \tag{3.12}$$

$U^s(x)=U^s_Ax^A$ is defined in (3.6) and

$$(U^s_A)=(d_i\delta_{iI_s},-\chi_s\lambda_{a_s\alpha}). \tag{3.13}$$

Here

$$\delta_{iI}\equiv\sum_{j\in I}\delta_{ij}=\begin{matrix}1, & i\in I\\ 0, & i\notin I\end{matrix} \tag{3.14}$$

is an indicator of i belonging to I. The potential (3.3) reads

$$V=\sum_{j=0}^{n}\frac{w}{2}\xi_j\,d_j\mathrm{e}^{2U^j(x)}, \tag{3.15}$$

where

$$U^j(x)=U^j_Ax^A=-\phi^j+\gamma_0(\phi), \tag{3.16}$$

$$(U^j_A)=(-\delta^j_i+d_i,0). \tag{3.17}$$

The integrability of the Lagrange system (3.2) depends upon the scalar products of co$-$vectors U^j, U^s corresponding to $\bar{G}$:

$$(U,U')=\bar{G}^{AB}U_AU'_B, \tag{3.18}$$

where

$$(\bar{G}^{AB})=\begin{pmatrix} G^{ij} & 0 \\ 0 & h^{\alpha\beta}\end{pmatrix} \tag{3.19}$$

is matrix inverse to (3.12). Here (as in Ref. 40)

$$G^{ij}=\frac{\delta^{ij}}{d_i}+\frac{1}{2-D}, \tag{3.20}$$

$i,j=0,\ldots,n$. These products have the following form[26,6]

$$(U^i,U^j)=\frac{\delta_{ij}}{d_j}-1, \tag{3.21}$$

$$(U^s,U^{s'})=d(I_s\cap I_{s'})+\frac{d(I_s)d(I_{s'})}{2-D}+\chi_s\chi_{s'}\lambda_{a\alpha}\lambda_{b\beta}h^{\alpha\beta}, \tag{3.22}$$

$$(U^s,U^i)=-\delta_{iI_s}, \tag{3.23}$$

where $s=(a_s,v_s,I_s)$, $s'=(a_{s'},v_{s'},I_{s'})\in S$.

IV. COSMOLOGICAL AND SPHERICALLY SYMMETRIC SOLUTIONS

Here we put the following restrictions on the parameters of the model

$$\textbf{(i)}\quad \xi_0\neq 0,\quad \xi_1=\cdots=\xi_n=0, \tag{4.1}$$

i.e., one space is curved and others are Ricci-flat,

$$\textbf{(ii)}\quad 0\notin I_s,\quad \forall s=(a_s,v_s,I_s)\in S, \tag{4.2}$$

i.e., all ''brane'' manifolds M_{I_s} [see (2.16)] do not contain M_0.

We also impose a block-orthogonality restriction on the set of vectors $(U^s, s\in S)$. Let

$$S=S_1\sqcup\cdots\sqcup S_k, \tag{4.3}$$

$S_i\neq\emptyset$, $i=1,\ldots,k$, and

$$\textbf{(iii)}\quad (U^s,U^{s'})=d(I_s\cap I_{s'})+\frac{d(I_s)d(I_{s'})}{2-D}+\chi_s\chi_{s'}\lambda_{a_s\alpha}\lambda_{a_{s'}\beta}h^{\alpha\beta}=0, \tag{4.4}$$

for all $s=(a_s,v_s,I_s)\in S_i$, $s'=(a_{s'},v_{s'},I_{s'})\in S_j$, $i\neq j$; $i,j=1,\ldots,k$. Relation (4.3) means that the set S is a union of k nonintersecting (nonempty) subsets $S_1,\ldots,S_k$. According to (4.4) the set of vectors $(U^s, s\in S)$ has a block-orthogonal structure with respect to the scalar product (3.18), i.e., it splits into k mutually orthogonal blocks $(U^s, s\in S_i)$, $i=1,\ldots,k$.

From **(i)** we get for the potential (3.15)

$$V=\tfrac{1}{2}w\xi_0 d_0 e^{2U^0(x)}, \tag{4.5}$$

where

$$(U^0,U^0)=\frac{1}{d_0}-1<0 \tag{4.6}$$

[see (3.21)].

From **(ii)** and (3.23) we get

$$(U^0,U^s)=0 \tag{4.7}$$

for all $s\in S$. Thus, the set of co-vectors U^0, $U^s, s\in S$ [belonging to dual space $(\mathbb{R}^{n+1+l})^*\simeq\mathbb{R}^{n+1+l}$] has also a block-orthogonal structure with respect to the scalar product (3.18).

Here we will integrate the Lagrange equations corresponding to the Lagrangian (3.2) and hence we will find classical exact solutions when the restrictions (A2) and (A3) from Appendix A are imposed.

First we integrate the ''Maxwell equations'' (for $s \in S_e$) and Bianchi identities (for $s \in S_m$):

$$\frac{d}{du}(\exp(-2U^s)\dot{\Phi}^s)=0 \Leftrightarrow \dot{\Phi}^s = Q_s \exp(2U^s), \tag{4.8}$$

where Q_s are constants, $s \in S$. We put

$$Q_s \neq 0, \tag{4.9}$$

for all $s \in S$.

Let us fix the time gauge as follows:

$$\gamma = \gamma_0, \quad \mathcal{N}=1, \tag{4.10}$$

i.e., the harmonic time gauge is used.

For fixed $Q=(Q_s, s \in S)$ the Lagrange equations for the Lagrangian (3.2) corresponding to $(x^A)=(\phi^i,\varphi^\alpha)$, when relations (4.10) and (4.8) are substituted, are equivalent to the Lagrange equations for the Lagrangian[23]

$$L_Q = \tfrac{1}{2}\bar{G}_{AB}\dot{x}^A\dot{x}^B - V_Q, \tag{4.11}$$

with the zero-energy constraint

$$E_Q = \tfrac{1}{2}\bar{G}_{AB}\dot{x}^A\dot{x}^B + V_Q = 0, \tag{4.12}$$

where

$$V_Q = V + \frac{1}{2}\sum_{s \in S} \varepsilon_s Q_s^2 \exp[2U^s(x)], \tag{4.13}$$

$(\bar{G}_{AB})$ and V are defined in (3.12) and (4.5), respectively.

When the conditions **(i)**–**(iii)** are satisfied exact solutions to Lagrange equations corresponding to (4.11) with the potential (4.13) and V from (4.5) could be readily obtained using the relations from Appendix B.

The solutions read

$$x^A(u) = -\frac{U^{0A}}{(U^0,U^0)}\ln|f_0(u)| - \sum_{s \in S}\eta_s \nu_s^2 U^{sA}\ln|f_s(u)| + c^A u + \bar{c}^A. \tag{4.14}$$

Functions f_0 and f_s in (4.14) are the following:

$$f_0(u) = |\xi_0(d_0-1)|^{1/2} s(u-u_0, w\xi_0, C_0) \tag{4.15}$$

$$= \left|\frac{\xi_0(d_0-1)}{C_0}\right|^{1/2} \mathrm{sh}(\sqrt{C_0}(u-u_0)), C_0>0, \xi_0 w>0; \tag{4.16}$$

$$= \left|\frac{\xi_0(d_0-1)}{C_0}\right|^{1/2} \sin(\sqrt{|C_0|}(u-u_0)), C_0<0, \xi_0 w>0; \tag{4.17}$$

$$= \left| \frac{\xi_0(d_0-1)}{C_0} \right|^{1/2} \mathrm{ch}(\sqrt{C_0}(u-u_0)), C_0>0, \xi_0 w<0; \tag{4.18}$$

$$= |\xi_0(d_0-1)|^{1/2}(u-u_0), C_0=0, \xi_0 w>0; \tag{4.19}$$

and

$$f_s(u) = \frac{|Q_s|}{|\nu_s|} s(u-u_s, -\eta_s\varepsilon_s, C_s) \tag{4.20}$$

$$= \frac{|Q_s|}{|\nu_s||C_s|^{1/2}} \mathrm{sh}(\sqrt{C_s}(u-u_s)), \quad C_s>0, \quad \eta_s\varepsilon_s<0; \tag{4.21}$$

$$= \frac{|Q_s|}{|\nu_s||C_s|^{1/2}} \sin(\sqrt{|C_s|}(u-u_s)), \quad C_s<0, \quad \eta_s\varepsilon_s<0; \tag{4.22}$$

$$= \frac{|Q_s|}{|\nu_s||C_s|^{1/2}} \mathrm{ch}(\sqrt{C_s}(u-u_s)), \quad C_s>0, \quad \eta_s\varepsilon_s>0; \tag{4.23}$$

$$\frac{|Q_s|}{|\nu_s|}(u-u_s), \quad C_s=0, \quad \eta_s\varepsilon_s<0; \tag{4.24}$$

where C_0, C_s, u_0, u_s are constants, $s \in S$. The function $s(u,\xi,C)$ is defined in Appendix B.

The parameters $\eta_s = \pm 1$, $\nu_s \neq 0$, $s \in S$, satisfy the relations

$$\sum_{s' \in S} (U^s, U^{s'}) \eta_{s'} \nu_{s'}^2 = 1, \tag{4.25}$$

for all $s \in S$, with scalar products $(U^s, U^{s'})$ defined in (3.22).

The constants C_s, u_s are coinciding inside blocks:

$$u_s = u_{s'}, \quad C_s = C_{s'}, \tag{4.26}$$

$s, s' \in S_i$, $i = 1, \ldots, k$ [see relation (B13) from Appendix B]. The ratios $\varepsilon_s Q_s^2/(\eta_s \nu_s^2)$ are also coinciding inside blocks, or, equivalently,

$$\varepsilon_s \eta_s = \varepsilon_{s'} \eta_{s'}, \tag{4.27}$$

$$\frac{Q_s^2}{\nu_s^2} = \frac{Q_{s'}^2}{\nu_{s'}^2}, \tag{4.28}$$

$s, s' \in S_i$, $i = 1, \ldots, k$. Here we used the relations (4.6) and (4.7).

The contravariant components $U^{rA} = \bar{G}^{AB} U_B^r$ are[23]

$$U^{0i} = -\frac{\delta_0^i}{d_0}, \quad U^{0\alpha} = 0, \tag{4.29}$$

$$U^{si} = G^{ij} U_j^s = \delta_{iI_s} - \frac{d(I_s)}{D-2}, \quad U^{s\alpha} = -\chi_s \lambda_{a_s}^{\alpha}. \tag{4.30}$$

Using (4.14), (4.6), (4.30), and (4.29) we obtain

$$\phi^i = \frac{\delta_0^i}{1-d_0} \ln|f_0| - \sum_{s \in S} \eta_s \nu_s^2 \left(\delta_{iI_s} - \frac{d(I_s)}{D-2} \right) \ln|f_s| + c^i u + \bar{c}^i \tag{4.31}$$

and

$$\varphi^\alpha = \sum_{s \in S} \eta_s \nu_s^2 \chi_s \lambda_{a_s}^\alpha \ln|f_s| + c^\alpha u + \bar{c}^\alpha, \tag{4.32}$$

$\alpha = 1, \dots, l$.

Vectors $c = (c^A)$ and $\bar{c} = (\bar{c}^A)$ satisfy the linear constraint relations [see (B20) in Appendix B]

$$U^0(c) = U_A^0 c^A = -c^0 + \sum_{j=0}^n d_j c^j = 0, \tag{4.33}$$

$$U^0(\bar{c}) = U_A^0 \bar{c}^A = -\bar{c}^0 + \sum_{j=0}^n d_j \bar{c}^j = 0, \tag{4.34}$$

$$U^s(c) = U_A^s c^A = \sum_{i \in I_s} d_i c^i - \chi_s \lambda_{a_s \alpha} c^\alpha = 0, \tag{4.35}$$

$$U^s(\bar{c}) = U_A^s \bar{c}^A = \sum_{i \in I_s} d_i \bar{c}^i - \chi_s \lambda_{a_s \alpha} \bar{c}^\alpha = 0, \tag{4.36}$$

$s \in S$. The (3.4) reads

$$\gamma_0(\phi) = \frac{d_0}{1-d_0} \ln|f_0| + \sum_{s \in S} \frac{d(I_s)}{D-2} \eta_s \nu_s^2 \ln|f_s| + c^0 u + \bar{c}^0. \tag{4.37}$$

The zero-energy constraint reads (see Appendix B)

$$E = E_0 + \sum_{s \in S} E_s + \frac{1}{2} \bar{G}_{AB} c^A c^B = 0, \tag{4.38}$$

where $E_0 = C_0 (U^0, U^0)^{-1}/2$, $E_s = C_s (\eta_s \nu_s^2)/2$. Using relations (3.10), (3.12), (4.6), and (4.33) we rewrite (4.38) as

$$C_0 \frac{d_0}{d_0 - 1} = \sum_{s \in S} C_s \nu_s^2 \eta_s + h_{\alpha\beta} c^\alpha c^\beta + \sum_{i=1}^n d_i (c^i)^2 + \frac{1}{d_0 - 1} \left(\sum_{i=1}^n d_i c^i \right)^2. \tag{4.39}$$

From relation

$$\exp(2U^s) = f_s^{-2}, \tag{4.40}$$

following from (4.4), (4.7), (4.14), (4.35), and (4.36) we get for electric-type forms (2.19)

$$\mathcal{F}^s = Q_s f_s^{-2} du \wedge \tau(I_s), \tag{4.41}$$

$s \in S_e$, and for magnetic-type forms (2.20)

$$\mathcal{F}^s = e^{-2\lambda_a(\varphi)} * [Q_s f_s^{-2} du \wedge \tau(I_s)] = \bar{Q}_s \tau(\bar{I}_s), \tag{4.42}$$

$s \in S_m$, where $\overline{Q_s} = Q_s \varepsilon(I_s)\mu(I_s)w$ and $\mu(I) = \pm 1$ is defined by the relation $\mu(I) du \wedge \tau(I_0) = \tau(\overline{I}) \wedge du \wedge \tau(I)$. The relation (4.42) follows from the formula (5.26) from Ref. 6 (for $\gamma = \gamma_0$).

Relations for the metric follows from (4.31) and (4.37)

$$g = \left(\prod_{s \in S} [f_s^2(u)]^{\eta_s d(I_s)\nu_s^2/(D-2)} \right) \Bigg\{ [f_0^2(u)]^{d_0/(1-d_0)} e^{2c^0 u + 2\overline{c^0}} [w\, du \otimes du + f_0^2(u) g^0] + \sum_{i=1}^{n} \left(\prod_{s \in S} [f_s^2(u)]^{-\eta_s \nu_s^2 \delta_{iI_s}} \right) e^{2c^i u + 2\overline{c^i}} g^i \Bigg\}. \tag{4.43}$$

Thus, here we obtained the ''block-orthogonal'' generalization of the solution from Ref. 23. This solution describes the evolution of $n+1$ spaces $(M_0, g_0), \ldots, (M_n, g_n)$, where (M_0, g_0) is an Einstein space of nonzero curvature, and (M_i, g^i) are ''internal'' Ricci-flat spaces, $i = 1, \ldots, n$; in the presence of several scalar fields and forms. The solution is presented by relations (4.32), (4.41)–(4.43) with the functions f_0, f_s defined in (4.15)–(4.24) and the relations on the parameters of solutions c^A, $\overline{c^A}$ $(A = i, \alpha)$, C_0, C_s, u_s, Q_s, η_s, ν_s $(s \in S)$ imposed in (4.25)–(4.28), (4.33)–(4.36), (4.39), respectively.

This solution describes a set of charged (by forms) overlapping p-branes $[p_s = d(I_s) - 1$, $s \in S]$ ''living'' on submanifolds (isomorphic to) M_{I_s} (2.16), where the sets I_s do not contain 0, i.e., all p-branes live in ''internal'' Ricci-flat spaces.

The solution is valid if the dimensions of p-branes and dilatonic coupling vector satisfy the relations (4.4). In ''orthogonal'' noncomposite case these solutions were considered in Refs. 29 and 28 (electric case) and Ref. 35 (electro-magnetic case). For $n = 1$ see also Refs. 24 and 27. In block-orthogonal (noncomposite) case a special class of solutions with ν_s^2 coinciding inside blocks was considered earlier in Ref. 36.

V. SPHERICALLY SYMMETRIC SOLUTIONS

Here we consider the spherically symmetric case

$$w = 1, \quad M_0 = S^{d_0}, \quad g^0 = d\Omega_{d_0}^2, \tag{5.1}$$

where $d\Omega_{d_0}^2$ is the canonical metric on a unit sphere S^{d_0}, $d \geqslant 2$. We also assume that

$$M_1 = \mathbb{R}, \quad g^1 = -dt \otimes dt \tag{5.2}$$

(here M_1 is a time manifold) and

$$1 \in I_s, \quad \forall s \in S, \tag{5.3}$$

i.e., all p-branes have a common time direction t.

For integration constants we put $\overline{c^A} = 0$,

$$c^A = \overline{\mu}(\overline{b^A} - b^A), \tag{5.4}$$

$$\overline{b^A} = \overline{\mu} \sum_{r \in S} \eta_r \nu_r^2 U^{rA} - \overline{\mu} \delta_1^A, \tag{5.5}$$

$$C_0 = \overline{\mu}^2, \tag{5.6}$$

$$C_s = \overline{\mu}^2 b_s^2, \quad b_s > 0, \tag{5.7}$$

where $\bar{\mu}>0$, $\bar{S}=\{0\}\cup S$ and $\eta_0\nu_0^2=(U^0,U^0)^{-1}$.

The only essential restrictions imposed are the inequalities $C_0,C_s>0$ that cut a subclass in the class of solutions from Sec. IV. This subclass contains nonextremal black hole solutions and its ''Kasner-like'' (non-black-hole) deformations. For extremal black hole solutions one should consider the special case $C_0=C_s=0$. (For extremal black hole solutions and its multicenter generalizations see Ref. 14.)

Due to (4.26) the parameters b_s, $s\in S$, are coinciding inside blocks:

$$b_s=b_{s'}, \tag{5.8}$$

$s,s'\in S_i$, $i=1,\dots,k$.

It may be verified that the restrictions (4.33) and (4.35) are satisfied identically if and only if

$$U^0(b)=U_A^0 b^A=-b^0+\sum_{j=0}^{n} d_j b^j=1, \tag{5.9}$$

$$U^s(b)=U_A^s b^A=\sum_{i\in I_s} d_i b^i-\chi_s\lambda_{a_s\alpha}b^\alpha=1, \tag{5.10}$$

$s\in S$. This follows from identities $U^0(\bar{b})=1$ and $U^s(\bar{b})=1$, $s\in S$.

Relation (4.39) reads

$$\sum_{s\in S}\eta_s\nu_s^2(b_s^2-1)+h_{\alpha\beta}b^\alpha b^\beta+\sum_{i=1}^{n} d_i(b^i)^2+\frac{1}{d_0-1}\left(\sum_{i=1}^{n} d_i b^i\right)^2=\frac{d_0}{d_0-1}, \tag{5.11}$$

where the relation [equivalent to (5.9)]

$$b^0=\frac{1}{1-d_0}\left[\sum_{j=1}^{n} d_j b^j-1\right], \tag{5.12}$$

is used.

Now we rewrite a solution (under restrictions imposed) in a so-called ''Kasner-type'' form that is more suitable for analyzing the behavior at large distances and for singling out the black hole solutions. For this reason we introduce a new radial variable $R=R(u)$ by relations

$$\exp(-2\bar{\mu}u)=1-\frac{2\mu}{R^{\bar{d}}},\quad \mu=\bar{\mu}/\bar{d}>0,\quad \bar{d}=d_0-1, \tag{5.13}$$

$u>0$, $R^{\bar{d}}>2\mu$. For the function

$$f_s(u)=\frac{|Q_s|}{2\bar{\mu}b_s|\nu_s|}[\exp(\bar{\mu}b_s(u-u_s))+\eta_s\varepsilon_s\exp(-\bar{\mu}b_s(u-u_s))] \tag{5.14}$$

we put the restriction $f_s(0)=1$, or, equivalently,

$$\exp(-\bar{\mu}b_s u_s)+\eta_s\varepsilon_s\exp(\bar{\mu}b_s u_s)=\frac{2\bar{\mu}b_s|\nu_s|}{|Q_s|}. \tag{5.15}$$

This restriction guarantees the asymptotical flatness of the $(2+d_0)$-dimensional section of the metric in the limit $R\to+\infty$ (or, when, $u\to+0$). It follows from (5.15) that $u_s<0$ for $\eta_s\varepsilon_s=-1$. In any case $f_s(u)>0$ for $u\geq 0$.

Then, solutions for the metric and scalar fields [see (4.32) and (4.43)] may be written as follows:

$$g=\left(\prod_{s\in S}\bar{H}_s^{2\eta_s d(I_s)\nu_s^2/(D-2)}\right)\left\{F^{b^0-1}dR\otimes dR+R^2F^{b^0}d\Omega^2_{d_0}\right.$$
$$\left.-\left(\prod_{s\in S}\bar{H}_s^{-2\eta_s\nu_s^2}\right)F^{b^1}dt\otimes dt+\sum_{i=2}^{n}\left(\prod_{s\in S}\bar{H}_s^{-2\eta_s\nu_s^2\delta_{iI_s}}\right)F^{b^i}g^i\right\}, \tag{5.16}$$

$$\varphi^\alpha=\sum_{s\in S}\eta_s\nu_s^2\chi_s\lambda_{a_s}^\alpha\ln\bar{H}_s+\frac{1}{2}b^\alpha\ln F, \tag{5.17}$$

where

$$F=1-\frac{2\mu}{R^{\bar{d}}}, \tag{5.18}$$

$$\bar{H}_s=\hat{H}_sF^{(1-b_s)/2}, \tag{5.19}$$

$$\hat{H}_s=1+\hat{P}_s\frac{(1-F^{b_s})}{2\mu b_s}, \tag{5.20}$$

$$\hat{P}_s=-\varepsilon_s\eta_sP_s, \tag{5.21}$$

$$P_s=\frac{|Q_s|}{\bar{d}|\nu_s|}\exp(\mu u_s)>0, \tag{5.22}$$

$s\in S$. Due to (4.26)–(4.28) parameters P_s and $\hat{P}_s$ are coinciding inside blocks:

$$P_s=P_{s'},\quad \hat{P}_s=\hat{P}_{s'}, \tag{5.23}$$

$s,s'\in S_i$, $i=1,\dots,k$. Parameters b_s are also coinciding inside blocks, see (5.8). Parameters b_s, b^i, b^α obey the relations (5.10)–(5.12).

The fields of forms are given by (2.19) and (2.20) with

$$\Phi^s=\frac{\nu_s}{H_s'}, \tag{5.24}$$

$$H_s'=\left[1-P_s'\hat{H}_s^{-1}\frac{(1-F^{b_s})}{2\mu b_s}\right]^{-1}, \tag{5.25}$$

$$P_s'=-\frac{Q_s}{\nu_s\bar{d}}, \tag{5.26}$$

$s\in S$. It follows from (5.15), (5.20), (5.21), and (5.26) that

$$(P_s')^2=P_s(\hat{P}_s+2b_s\mu)=-\varepsilon_s\eta_s\hat{P}_s(\hat{P}_s+2b_s\mu), \tag{5.27}$$

$s\in S$. This relation is self-consistent, i.e., its left- and right-hand sides have the same sign, since due to (5.15) and (5.22):

$$P_s < 2\mu b_s \tag{5.28}$$

for $\varepsilon_s \eta_s = +1$ and hence

$$\hat{P}_s > -2 b_s \mu, \tag{5.29}$$

for all $s \in S$.

Remark: The black hole solution from Ref. 14 may be obtained from our spherically-symmetric solutions (5.16)–(5.27) when

$$b^1 = b_s = 1, \quad b^i = b^\alpha = 0, \tag{5.30}$$

$s \in S$, $i = 0,2, \ldots ,n$; $\alpha = 1, \ldots ,l$. For details see Ref. 41.

We note that the obtained solution generalizes the solutions from Ref. 42 (in the pure gravitational case) and Ref. 43 (in the case of gravity with one scalar field).

VI. POST-NEWTONIAN APPROXIMATION

Let $d_0 = 2$. Here we consider the four-dimensional section of the metric (5.16)

$$g^{(4)} = U\{F^{b^0 - 1} dR \otimes dR + F^{b^0} R^2 d\Omega_2^2 - U_1 F^{b^1} dt \otimes dt\}, \tag{6.1}$$

where $F = 1 - (2\mu/R)$, and

$$U = \prod_{s \in S} \bar{H}_s^{2\eta_s d(I_s) \nu_s^2/(D-2)}, \tag{6.2}$$

$$U_1 = \prod_{s \in S} \bar{H}_s^{-2\eta_s \nu_s^2}, \tag{6.3}$$

$$U_i = \prod_{s \in S} \bar{H}_s^{-2\eta_s \nu_s^2 \delta_{iI_s}}, \quad i > 1, \tag{6.4}$$

$R > 2\mu$.

We may suppose that some real astrophysical objects (e.g., stars) are described by the four-dimensional ''physical'' metric (6.1), i.e., they are ''traces'' of extended multidimensional objects (charged *p*-branes).

Introducing a new radial variable ρ by the relation

$$R = \rho \left(1 + \frac{\mu}{2\rho} \right)^2, \tag{6.5}$$

$(\rho > \mu/2)$, we rewrite the metric (6.1) in a three-dimensional conformally flat form

$$g^{(4)} = U\left\{ -U_1 F^{b^1} dt \otimes dt + F^{b^0} \left(1 + \frac{\mu}{2\rho} \right)^4 \delta_{ij} dx^i \otimes dx^j \right\}, \tag{6.6}$$

$$F = \left(1 - \frac{\mu}{2\rho} \right)^2 \left(1 + \frac{\mu}{2\rho} \right)^{-2}, \tag{6.7}$$

where $\rho^2 = |x|^2 = \delta_{ij} x^i x^j$ $(i,j = 1,2,3)$.

For possible physical applications we should calculate the post-Newtonian parameters β and γ (Eddington parameters) using the following relations (see, for example, Ref. 44 and references therein)

$$g_{00}^{(4)} = -(1-2V+2\beta V^2) + O(V^3), \tag{6.8}$$

$$g_{ij}^{(4)} = \delta_{ij}(1+2\gamma V) + O(V^2), \tag{6.9}$$

$i,j=1,2,3$, where

$$V = \frac{GM}{\rho} \tag{6.10}$$

is the Newton's potential, G is the gravitational constant, M is the gravitational mass. From (6.6)–(6.10) we get

$$GM = \mu b^1 + \sum_{s\in S} \eta_s \nu_s^2 [\hat{P}_s + (b_s - 1)\mu]\left(1 - \frac{d(I_s)}{D-2}\right) \tag{6.11}$$

and for $GM \neq 0$

$$\beta - 1 = \frac{1}{2(GM)^2} \sum_{s\in S} \eta_s \nu_s^2 \hat{P}_s (\hat{P}_s + 2 b_s \mu)\left(1 - \frac{d(I_s)}{D-2}\right), \tag{6.12}$$

$$\gamma - 1 = -\frac{1}{GM}\left[\mu(b^0 + b^1 - 1) + \sum_{s\in S} \eta_s \nu_s^2 [\hat{P}_s + (b_s - 1)\mu]\left(1 - 2\frac{d(I_s)}{D-2}\right)\right]. \tag{6.13}$$

It follows from (5.27) and (6.12) and the inequalities $d(I_s) < D-2$ (for all $s \in S$) that the following inequalities take place:

$$\beta > 1, \text{ if all } \varepsilon_s = -1, \tag{6.14}$$

$$\beta < 1, \text{ if all } \varepsilon_s = +1. \tag{6.15}$$

There exists a large variety of configurations with $\beta = 1$ when the relation $\varepsilon_s = \text{const}$ is broken.

There exist also nontrivial p-brane configurations with $\gamma = 1$.

Proposition: Let the set of p-branes consist of several pairs of electric and magnetic branes. Let any such pair $(s, \bar{s} \in S)$ correspond to the same color index, i.e., $a_s = a_{\bar{s}}$, and $\hat{P}_s = \hat{P}_{\bar{s}}$, $b_s = b_{\bar{s}}$, $\eta_s \nu_s^2 = \eta_{\bar{s}} \nu_{\bar{s}}^2$. Then for $b^0 + b^1 = 1$ we get

$$\gamma = 1. \tag{6.16}$$

The proposition can be readily proved using the relation $d(I_s) + d(I_{\bar{s}}) = D-2$, following from (2.25) and (2.26).

Observational restrictions: The observations in the solar system give the tight constraints on the Eddington parameters[44]

$$\gamma = 1.000 \pm 0.002, \tag{6.17}$$

$$\beta = 0.9998 \pm 0.0006. \tag{6.18}$$

The first restriction is a result of the Viking time-delay experiment.[45] The second restriction follows from (6.17) and the analysis of the laser ranging data to the Moon. In this case a high precision test based on the calculation of the combination $(4\beta - \gamma - 3)$ appearing in the Nordtvedt effect[46] is used.[47] We note, that as it was pointed in Ref. 44 the ''classic'' tests of general relativity, i.e., the Mercury-perihelion and light deflection tests, are somewhat outdated.

For small enough $\hat{p}_s = \hat{P}_s / GM$, $b_s - 1$, $b^1 - 1$, b^i $(i > 1)$ of the same order we get $GM \sim \mu$ and hence

$$\beta-1\sim\sum_{s\in S}\eta_s\nu_s^2\hat{p}_s\left(1-\frac{d(I_s)}{D-2}\right), \tag{6.19}$$

$$\gamma-1\sim-b^0-b^1+1-\sum_{s\in S}\eta_s\nu_s^2[\hat{p}_s+b_s-1]\left(1-2\frac{d(I_s)}{D-2}\right), \tag{6.20}$$

i.e., $\beta-1$ and $\gamma-1$ are of the same order. Thus for small enough $\hat{p}_s$, b_s-1, b^1-1, b^i $(i>1)$ it is possible to fit the ''solar system'' restrictions (6.17) and (6.18).

There exists also another possibility to satisfy these restrictions.

One brane case: Let us consider a special case of one p-brane. In this case we have

$$\eta_s\nu_s^{-2}=d(I_s)\left(1-\frac{d(I_s)}{D-2}\right)+\lambda^2. \tag{6.21}$$

Relations (6.12), (6.13), and (6.21) imply that for large enough values of (dilatonic coupling constant squared) λ^2 and $b^0+b^1=1$ it is possible to perform the ''fine tuning'' the parameters (β,γ) near the point (1, 1) even if the parameters $\hat{P}_s$ are big.

VII. CONCLUSIONS

In this paper we obtained exact solutions to Einstein equations for the multidimensional cosmological model describing the evolution of n Ricci-flat spaces and one Einstein space M_0 of nonzero curvature in the presence of composite electro-magnetic p-branes. The solutions were obtained in the block-orthogonal case (4.4), when p-branes do not ''live'' in M_0. We also considered the spherically symmetric solutions containing nonextremal p-brane black holes.[36,14] The relations for post-Newtonian parameters β and γ are obtained.

ACKNOWLEDGMENTS

This work was supported in part by DFG Grant No. 436 RUS 113/236/O(R) and by the Russian Ministry for Science and Technology, Russian Fund for Basic Research, project N 98-02-16414 and project SEE. V.N.M. is grateful to the Dept. Math., University of Aegean for the kind hospitality during his stay in Karlovassi, Greece.

APPENDIX A: RESTRICTIONS ON *p*-BRANE CONFIGURATIONS

Restrictions on $\Omega_{a,v}$*:*[23] Let

$$w_1\equiv\{i|i\in\{0,\ldots,n\},d_i=1\}. \tag{A1}$$

The set w_1 describes all one-dimensional manifolds among M_i $(i\geqslant 0)$. We impose the following restrictions on the sets $\Omega_{a,v}$ (2.21):

$$W_{ij}(\Omega_{a,v})=\varnothing, \tag{A2}$$

$a\in\Delta;\ v=e,m;\ i,j\in w_1,\ i<j$ and

$$W_j^{(1)}(\Omega_{a,m},\Omega_{a,e})=\varnothing, \tag{A3}$$

$a\in\Delta;\ j\in w_1$. Here

$$W_{ij}(\Omega_*)\equiv\{(I,J)|I,J\in\Omega_*,\ \ I=\{i\}\sqcup(I\cap J),\quad J=\{j\}\sqcup(I\cap J)\}, \tag{A4}$$

$i,j\in w_1,\ i\neq j,\ \Omega_*\subset\Omega_0$ and

$$W_j^{(1)}(\Omega_{a,m},\Omega_{a,e}) \equiv \{(I,J) \in \Omega_{a,m} \times \Omega_{a,e} | \bar{I} = \{j\} \sqcup J\}, \tag{A5}$$

$j \in w_1$. In (A5)

$$\bar{I} \equiv I_0 \setminus I \tag{A6}$$

is ''dual'' set, ($I_0 = \{0,1,\ldots,n\}$).

The restrictions (A2) and (A3) are trivially satisfied when $n_1 \leqslant 1$ and $n_1 = 0$, respectively, where $n_1 = |w_1|$ is the number of one-dimensional manifolds among M_i. They are also satisfied in the noncomposite case when all $|\Omega_{a,v}| = 1$. For $n_1 \geqslant 2$ and $n_1 \geqslant 1$, respectively, these restrictions forbid certain pairs of two p-branes, corresponding to the same form $F^a, a \in \Delta$:

APPENDIX B: SOLUTIONS WITH BLOCK-ORTHOGONAL SET OF VECTORS

Let

$$L = \frac{1}{2}\langle \dot{x}, \dot{x} \rangle - \sum_{s \in S} A_s \exp(2\langle b_s, x \rangle) \tag{B1}$$

be a Lagrangian, defined on $V \times V$, where V is a n-dimensional vector space over $\mathbb{R}$, $A_s \neq 0$, $s \in S$; $S \neq \emptyset$, and $\langle \cdot, \cdot \rangle$ is a nondegenerate real-valued quadratic form on V. Let

$$S = S_1 \sqcup \cdots \sqcup S_k, \tag{B2}$$

all $S_i \neq \emptyset$, and

$$\langle b_s, b_{s'} \rangle = 0, \tag{B3}$$

for all $s \in S_i$, $s' \in S_j$, $i \neq j$; $i,j = 1,\ldots,k$.

Let us suppose that there exists a set $h_s \in \mathbb{R}$, $h_s \neq 0$, $s \in S$, such that

$$\sum_{s \in S} \langle b_s, b_{s'} \rangle h_{s'} = -1, \tag{B4}$$

for all $s \in S$, and

$$\frac{A_s}{h_s} = \frac{A_{s'}}{h_{s'}}, \tag{B5}$$

$s,s' \in S_i$, $i = 1,\ldots,k$, (the ratio A_s/h_s is constant inside S_i).

Then, the Euler–Lagrange equations for the Lagrangian (B1)

$$\ddot{x} + \sum_{s \in S} 2A_s b_s \exp(2\langle b_s, x \rangle) = 0, \tag{B6}$$

have the following special solutions

$$x(t) = \frac{1}{2} \sum_{s \in S} h_s b_s \ln\left[y_s^2(t) \left| \frac{2A_s}{h_s} \right| \right] + \alpha t + \beta, \tag{B7}$$

where $\alpha, \beta \in V$,

$$\langle \alpha, b_s \rangle = \langle \beta, b_s \rangle = 0, \tag{B8}$$

$s \in S$, and functions $y_s(t) \neq 0$ satisfy the equations

$$\frac{d}{dt}\left(y_s^{-1}\frac{dy_s}{dt}\right)=-\xi_s y_s^{-2}, \tag{B9}$$

with

$$\xi_s=\mathrm{sign}\left(\frac{A_s}{h_s}\right), \tag{B10}$$

$s\in S$, and coincide inside blocks:

$$y_s(t)=y_{s'}(t), \tag{B11}$$

$s,s'\in S_i$, $i=1,\ldots,k$. More explicitly

$$y_s(t)=s(t-t_s,\xi_s,C_s), \tag{B12}$$

where constants $t_s,C_s\in\mathbb{R}$ coincide inside blocks

$$t_s=t_{s'},\quad C_s=C_{s'}, \tag{B13}$$

$s,s'\in S_i$, $i=1,\ldots,k$, and

$$s(t,\xi,C)\equiv\frac{1}{\sqrt{C}}\mathrm{sh}(t\sqrt{C}),\xi=+1,\quad C>0; \tag{B14}$$

$$\frac{1}{\sqrt{-C}}\sin(t\sqrt{-C}),\xi=+1,\quad C<0; \tag{B15}$$

$$t,\xi=+1,\quad C=0; \tag{B16}$$

$$\frac{1}{\sqrt{C}}\mathrm{ch}(t\sqrt{C}),\xi=-1,\quad C>0. \tag{B17}$$

For the energy

$$E=\frac{1}{2}\langle\dot{x},\dot{x}\rangle+\sum_{s\in S}A_s\exp(2\langle b_s,x\rangle) \tag{B18}$$

corresponding to the solution (B7) we have

$$E=\frac{1}{2}\sum_{s\in S}C_s(-h_s)+\frac{1}{2}\langle\alpha,\alpha\rangle. \tag{B19}$$

For dual vectors $u^s\in V^*$ defined as $u^s(x)=\langle b_s,x\rangle$, $\forall x\in V$, we have $\langle u^s,u^l\rangle_*=\langle b_s,b_l\rangle$, where $\langle\cdot,\cdot\rangle_*$ is dual form on V^*. The orthogonality conditions (B8) read

$$u^s(\alpha)=u^s(\beta)=0, \tag{B20}$$

$s\in S$.

[1] C. Hull and P. Townsend, ''Unity of Superstring Dualities,'' Nucl. Phys. B **438**, 109 (1995). P. Horava and E. Witten, *ibid.* **460**, 506 (1996).
[2] J. M. Schwarz, ''Lectures on superstring and *M*-theory dualities,'' hep-th/9607201.
[3] M. J. Duff, ''*M*-theory (the theory formerly known as strings),'' hep-th/9608117.
[4] C. Vafa, ''Evidence for *F*-theory,'' hep-th/9602022; Nucl. Phys. B **469**, 403 (1996).

[5] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge Univ. Press, Cambridge, 1987).
[6] V. D. Ivashchuk and V. N. Melnikov, ''Sigma-model for generalized composite *p*-branes,'' hep-th/9705036; Class. Quantum Grav. **14**, 3001 (1997); Corrigenda, *ibid.* **15**, 3941 (1998).
[7] M. J. Duff, R. R. Khuri, and J. X. Lu, Phys. Rep. **259**, 213 (1995).
[8] K. S. Stelle, ''Lectures on Supergravity *p*-Branes,'' hep-th/9701088; hep-th/9608117.
[9] V. D. Ivashchuk and V. N. Melnikov, ''Interesting *p*-brane solutions in multidimensional gravity and *M*-theory,'' hep-th/9612089; Gravit. Cosmol. **2**, 297 (1996).
[10] V. D. Ivashchuk and V. N. Melnikov, Phys. Lett. B **403**, 23 (1997).
[11] V. D. Ivashchuk, M. Rainer, and V. N. Melnikov, ''Multidimensional sigma-models with composite electric *p*-branes,'' gr-qc/9705005; Gravit. Cosmol. **4**, 13 (1998).
[12] M. Rainer and A. Zhuk, Phys. Rev. D **54**, 6186 (1996).
[13] V. D. Ivashchuk and V. N. Melnikov, ''Multidimensional Gravity with Einstein Internal spaces,'' hep-th/9612054; Gravit. Cosmol. **2**, 211 (1996).
[14] V. D. Ivashchuk and V. N. Melnikov, ''Mudjumdar-Papapetrou type solutions in sigma-model and intersecting *p*-branes,'' hep-th/9702121, Class. Quantum Grav. (to be published).
[15] G. Papadopoulos and P. K. Townsend, Phys. Lett. B **380**, 273 (1996).
[16] A. A. Tseytlin, ''Harmonic superpositions of *M*-branes,'' hep-th/9604035; Nucl. Phys. B **475**, 149 (1996).
[17] J. P. Gauntlett, D. A. Kastor, and J. Traschen, ''Overlapping branes in *M*-theory,'' hep-th/9604179; Nucl. Phys. B **478**, 544 (1996).
[18] I. Ya. Aref'eva and A. I. Volovich, Composite *p*-branes in Diverse Dimensions, preprint SMI-19-96, hep-th/9611026; Class. Quantum Grav. **14**, 2990 (1997).
[19] E. Bergshoeff, M. de Roo, E. Eyras, B. Janssen, and J. P. van der Schaar, hep-th/9612095.
[20] I. Ya. Aref'eva and O. A. Rytchkov, ''Incidence matrix description of intersecting *p*-brane solutions,'' hep-th/9612236.
[21] R. Argurio, F. Englert, and L. Hourant, ''Intersection rules for *p*-branes,'' hep-th/9701042; Phys. Lett. B **398**, 2991 (1997).
[22] I. Ya. Aref'eva, M. G. Ivanov, and O. A. Rytchkov, ''Properties of intersecting *p*-branes in various dimensions,'' hep-th/9702077.
[23] V. D. Ivashchuk and V. N. Melnikov, ''Multidimensional classical and quantum cosmology with intersecting *p*-branes,'' hep-th/9708157; J. Math. Phys. **39**, 2866 (1998).
[24] H. Lü, S. Mukherji, C. N. Pope, and K.-W. Xu, ''Cosmological solutions in string theories,'' hep-th/9610107.
[25] V. D. Ivashchuk and V. N. Melnikov, ''Multidimensional cosmology with *m*-component perfect fluid,'' gr-qc/9403063; Int. J. Mod. Phys. D **3**, 795 (1994).
[26] V. R. Gavrilov, V. D. Ivashchuk, and V. N. Melnikov, ''Integrable pseudo-euclidean Toda-like systems in multidimensional cosmology with multicomponent perfect fluid,'' J. Math. Phys. **36**, 5829 (1995).
[27] H. Lü, C. N. Pope, and K. W. Xu, ''Liouville and Toda solitons in *M*-theory,'' hep-th/9604058.
[28] K. A. Bronnikov, M. A. Grebeniuk, V. D. Ivashchuk, and V. N. Melnikov, ''Integrable multi-dimensional cosmology for intersecting *p*-branes,'' Gravit. Cosmol. **3**, 105 (1997).
[29] M. A. Grebeniuk, V. D. Ivashchuk, and V. N. Melnikov, ''Integrable multidimensional quantum cosmology for intersecting *p*-branes,'' Grav. and Cosmol. **3**, 243 (1997); gr-qc/9708031.
[30] K. A. Bronnikov, U. Kasper, and M. Rainer, ''Intersecting electric and magnetic *p*-branes: Spherically symmetric solutions,'' gr-qc/9708058.
[31] M. Cvetic and A. Tseytlin, Nucl. Phys. B **478**, 181 (1996).
[32] N. Ohta and T. Shimizu, ''Non-extreme black holes from intersecting *M*-branes,'' hep-th/9701095.
[33] I. Ya. Aref'eva, M. G. Ivanov, and I. V. Volovich, ''Non-Extremal Intersecting *p*-Branes in Various Dimensions,'' hep-th/9702079; Phys. Lett. B **406**, 44 (1997).
[34] N. Ohta, ''Intersection rules for non-extreme *p*-branes,'' hep-th/9702164.
[35] K. A. Bronnikov, V. D. Ivashchuk, and V. N. Melnikov, ''The Reissner-Nordström problem for intersecting electric and magnetic *p*-branes,'' gr-qc/9710054; Gravit. Cosmol. **3**, 203 (1997).
[36] K. A. Bronnikov, ''Block-orthogonal brane systems, black holes and wormholes,'' hep-th/9710207; Gravit. Cosmol. **4**, 14 (1998).
[37] M. A. Grebeniuk and V. D. Ivashchuk, ''Solutions in Sigma Model and Intersecting *p*-branes Related to the Lie Algebras,'' hep-th/9805113; Phys. Lett. B **442**, 125 (1998).
[38] H. Lü, J. Maharana, S. Mukherji, and C. N. Pope, ''Cosmological solutions, *p*-branes and the Wheeler De Witt equation,'' hep-th/9707182.
[39] V. R. Gavrilov and V. N. Melnikov, Toda chains with type A_m Lie algebra for multidimensional classical cosmology with intersecting *p*-branes, in *Proceedings of International Seminar on Mathematical Cosmology, Potsdam 1998*, edited by M. Rainer and H.-J. Schmidt (World Scientific, Singapore, 1998), pp. 299–311; gr-qc/9807004.
[40] V. D. Ivashchuk, V. N. Melnikov, and A. I. Zhuk, Nuovo Cimento B **104**, 575 (1989).
[41] S. Cotsakis, V. D. Ivashchuk, and V. N. Melnikov, ''*P*-brane Black Holes and Post-Newtonian Approximation,'' gr-qc/9902148; Gravit. Cosmol. **5**, 52 (1999).
[42] S. B. Fadeev, V. D. Ivashchuk, and V. N. Melnikov, Phys. Lett. A **161**, 98 (1991).
[43] V. D. Ivashchuk and V. N. Melnikov, Gravit. Cosmol. **1**, 204 (1995).
[44] T. Damour, ''Gravitation, Experiment and Cosmology,'' gr-qc/9606079.
[45] R. D. Reasenberg, Astrophys. J. **234**, L219 (1979).
[46] K. Nordtvedt, Phys. Rev. **169**, 1017 (1968).
[47] J. O. Dickey *et al.*, Science **265**, 482 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Kostant's cubic Dirac operator of Lie superalgebras

Teparksorn Pengpan[a)]
Institute for Fundamental Theory, Department of Physics, University of Florida, Gainesville, Florida 32611

(Received 7 September 1999; accepted for publication 15 September 1999)

We extend equal rank embedding of reductive Lie algebras to that of basic Lie superalgebras. The Kac character formulas for equal rank embedding are derived in terms of subalgebras and Kostant's cubic Dirac operator for equal rank embedding of Lie superalgebras is constructed from both even and odd generators and their related structure constants.
[S0022-2488(99)03012-1]

I. INTRODUCTION

The study of some patterns,[1] connected with $N=1$ supergravity theory in 11 dimensions, has led to a recent understanding in terms of a Weyl character formula by GKRS,[2] based on equal rank embedding of reductive Lie algebras. By using the construction of equal rank embedding of reductive Lie algebras, all possible equal rank embeddings were catalogued, and nearly all supersymmetric multiplets of massless and massive particles that are already known in supersymmetric gauge field theories emerge as the lowest lines of the infinite tower multiplet spectra, some of them shown in detail in Ref. 1. Immediately after the appearance of the Weyl character formula for equal rank embedding as an index formula for the Dirac operator, Kostant moved the subject forward and related the multiplet spectrum to the kernel of a cubic Dirac operator, which he introduced around 30 years ago.[3]

In this paper, we extend the Gross–Kostant–Ramond–Sternberg (GKRS) Weyl character formula and Kostant's cubic Dirac operator for equal rank embedding of reductive Lie algebras to those of basic Lie superalgebras. In Sec. II, we give a brief review in GKRS's paper. A derivation of the Weyl character formula for an equal rank embedding is shown. In Sec. III, we extend the equal rank embedding construction of reductive Lie algebras shown in Sec. II to that of basic Lie superalgebras. The Kac character formulas are written in terms of equal rank subalgebras. In Sec. IV, we give a simple and explicit formulation for a typical representation of type I Lie superalgebras. In Sec. V, we build a multiplet of type I Lie superalgebras from that of reductive Lie algebras. In Sec. VI, Kostant's cubic Dirac operator is constructed for full Lie superalgebras and then for equal rank embeddings.

II. WEYL CHARACTER FORMULA AND EQUAL RANK EMBEDDINGS OF REDUCTIVE LIE ALGEBRAS

Let r be an equal rank subalgebra of reductive Lie algebras, g, and let **C** be of the order of C, the ratio of the Weyl group of g to that of r. The restricted conditions for this kind of equal rank embedding, $g\supset r$, are that (1) positive roots of g must contain those of r, i.e., $\Phi^+(g)\supset\Phi^+(r)$, and (2) the simple roots of g and r must be chosen consistently so that the positive Weyl chamber of r contains that of g. In the Cartan–Weyl basis, the c elements of the Weyl group in C acting on the sum of highest weight, λ, and the Weyl vector of g, ρ_g, and then after subtracting by the Weyl vector of r, ρ_r, generate **C** irreducible representations of r, called a **C** multiplet,

$$c\cdot\lambda := c(\lambda+\rho)_g-\rho_r\,.$$

[a)]Electronic mail: tpengpan@phys.ufl.edu

The Weyl character formula of the irreducible representation of g, V_λ, can be rewritten in terms of the irreducible representation of r, $U_{c\cdot\lambda}$, as follows:

$$\begin{aligned}\mathrm{ch}\,V_\lambda &= \frac{\Sigma_{w\in W(g)}\,\mathrm{sgn}(w)e^{w(\lambda+\rho_g)}}{\Sigma_{w\in W(g)}\,\mathrm{sgn}(w)e^{w\rho_g}}\\ &= \frac{\Sigma_{c\in C}\,\mathrm{sgn}(c)(\Sigma_{w_r\in W(r)}\,\mathrm{sgn}(w_r)e^{w_r(c\cdot\lambda+\rho_r)})}{(\Pi_{\phi\in\Phi^+(g/r)}(e^{\phi/2}-e^{-\phi/2}))(\Pi_{\phi\in\Phi+(r)}(e^{\phi/2}-e^{-\phi/2}))},\\ &= \frac{1}{\Delta}\sum_{c\in C}\,\mathrm{sgn}(c)\,\mathrm{ch}\,U_{c\cdot\lambda}, \end{aligned} \tag{1}$$

where Δ is the character difference of two spinor modules, S^+ and S^-, of $SO(p=g/r)$, i.e.,

$$\Delta := \prod_{\phi\in\Phi^+(g/r)} (e^{\phi/2}-e^{-\phi/2}) = \mathrm{ch}\,S^+ - \mathrm{ch}\,S^-.$$

The beauty of Eq. (1) is that it gives us

$$V_\lambda\otimes S^+ - V_\lambda\otimes S^- = \sum_{c\in C}\,\mathrm{sgn}(c)U_{c\cdot\lambda}. \tag{2}$$

Equation (2) should be viewed as an equation in the Grothendieck ring of r. The lhs is the character of the difference between two representations of the same dimension as the representation of g with highest weight λ, this representation restricted to r tensored with each of the half-spin representations of $\mathrm{SO}(p)$, where $p=\dim g-\dim r$, restricted to r. In the other words, the lhs is the algebraic index of the Dirac operator associated to λ and the two half-spin representations. The alternating sum on the rhs is just the dimension difference between the kernel and cokernel of the Dirac operator. All the representations on the rhs are inequivalent, and so the rhs is the end result of a lot of cancellation on the lhs, in short an index formula for the Dirac operator. One of the remarkable consequences of Eq. (2) is that, on the lhs, the order of difference of $V_\lambda\otimes S^+$ and $V_\lambda\otimes S^-$ is always equal to **C**, independent of λ, and, on the rhs, the multiplicity of each $U_{c\cdot\lambda}$ representation is exactly one.

Notice that **C** is equal to the Euler number, which is a topological invariant of the coset manifold, G/R, corresponding to exponentiation of g/r. From our catalog of equal rank embedding of reductive Lie algebras, we would like to give some examples of the coset manifolds, where supersymmetric multiplets appear to be in their lowest lines of the infinite tower multiplet spectra. $N=2$ hypermultiplet, $N=4$ vector multiplet, and $N=8$ supergravity multiplet, which undoubtedly emerge in the lowest lines of $\mathrm{SU}(N+1)\supset\mathrm{SU}(N)\times\mathrm{U}(1)$ and $\mathrm{SO}(N+2)\supset\mathrm{SO}(N)\times\mathrm{SO}(2)$ series live on the N-dimensional complex projective space and $(N+2)$-dimensional complex Grassmannian manifolds, respectively. If SO(2) or U(1) is viewed as the light-cone little group, these lowest line spectra are massless supermultiplets in four-dimensional space–time. Whereas $N=1$, $N=2$, $N=3$, and $N=4$ massive (massless) multiplets in four-dimensional (six-dimensional) space–time that emerge in the lowest lines of the $\mathrm{Sp}(2N+2)\supset\mathrm{Sp}(2N)\times\mathrm{Sp}(2)$ series live on N-dimensional quaternionic projective space. The last multiplet that we would like to mention is the $N=1$ massless (massive) supergravity triplet in 11-dimensional (10-dimensional) space–time. The triplet emerges from $F_4\supset\mathrm{SO}(9)$ and lives on a (16-dimensional) Cayley plane. All infinite tower multiplet spectra are a kernel of Kostant's cubic Dirac operator, $\not{\mathcal{K}}$,

$$\mathrm{Ker}(\not{\mathcal{K}}_\lambda^2) = \mathrm{Ker}(\not{\mathcal{K}}_\lambda) = \sum_{c\in C}\,\mathrm{sgn}(c)U_{c\cdot\lambda}.$$

For more analytical details on Kostant's operator, see Ref. 4.

III. KAC CHARACTER FORMULAS AND EQUAL RANK EMBEDDING OF BASIC LIE SUPERALGEBRAS

Now, we extend the results of reductive Lie algebras to Lie superalgebras with nondegenerate Killing form. According to Kac's classification,[5] there are two types of basic Lie superalgebras type I, which is $su(m|n)$ and $osp(2|2n)$ and type II, which is $osp(2m+1|2n)$, $osp(1|2n)$, $osp(2m|2n)$, $osp(4|2;\alpha)$, $F(4)$, and $G(3)$.

Let $g=g_{\mathrm{even}}\oplus g_{\mathrm{odd}}$ be the Lie superalgebras with the root system $\Phi=\Phi_{\mathrm{even}}\cup\Phi_{\mathrm{odd}}$. For type I, g_{even} is simple, i.e., $g_{\mathrm{even}}=g_0$, and, for type II, g_{even} can be graded into $g_2\oplus g_0\oplus g_{-2}$. While, for the odd part of both type I and II, odd generators can be graded into fermionic creation and annihilation ones, i.e. $g_{\mathrm{odd}}=g_1\oplus g_{-1}$. The Poincaré–Birkhoff–Witt theorem for Lie algebras can be applied to the case of Lie superalgebras with some extension.[6] This grading gives us a universal enveloping algebra, $\mathcal{U}(g)$, e.g., for type I,

$$\mathcal{U}(g)=\mathcal{U}(g_1)\otimes\mathcal{U}(g_0)\otimes\mathcal{U}(g_{-1}).$$

Define root subsystems, $\bar{\Phi}_{\mathrm{even}}$ and $\bar{\Phi}_{\mathrm{odd}}$, such that $\bar{\Phi}_{\mathrm{even}}=\{\alpha|\alpha/2\notin\Phi_{\mathrm{odd}}\}$ and $\bar{\Phi}_{\mathrm{odd}}=\{\beta|2\beta\notin\Phi_{\mathrm{even}}\}$. Since Φ_{even}, Φ_{odd}, $\bar{\Phi}_{\mathrm{even}}$, and $\bar{\Phi}_{\mathrm{odd}}$ are invariant under the action of the Weyl group of g_{even}. Hence, the Weyl group of g is equal to that of g_{even}, i.e., $W(g)=W(g_{\mathrm{even}})$. Define the Weyl vector of g to be one-half the sum of positive even roots minus one-half the sum of positive odd roots, i.e., $\rho=\rho_{\mathrm{even}}-\rho_{\mathrm{odd}}$. Let $V(\Lambda)$ be a representation of g with Λ as a highest weight in dual Cartan subalgebra. The highest weight representations of g are classified into typical and atypical. The representation is typical if, for $\forall\beta\in\bar{\Phi}^{+}_{\mathrm{odd}}(\Lambda+\rho,\beta)\neq 0$; otherwise, it is atypical. The typical Kac character and supercharacter formulas of $V(\Lambda)$ are defined, respectively, as

$$\mathrm{ch}\,V(\Lambda)=\frac{N_1}{N_0}\sum_{w\in W(g)}\mathrm{sgn}(w)e^{w(\Lambda+\rho)},\tag{3}$$

$$\mathrm{sch}\,V(\Lambda)=\frac{N_1'}{N_0}\sum_{w\in W(g)}\mathrm{sgn}(\bar{w})e^{w(\Lambda+\rho)},\tag{4}$$

where

$$N_0=\prod_{\alpha\in\Phi^{+}_{\mathrm{even}}}(e^{\alpha/2}-e^{-\alpha/2}),$$

$$N_1=\prod_{\beta_{+}\in\Phi^{+}_{\mathrm{odd}}}(e^{\beta/2}+e^{-\beta/2}),$$

and

$$N_1'=\prod_{\beta_{+}\in\Phi^{+}_{\mathrm{odd}}}(e^{\beta/2}-e^{-\beta/2}).$$

The sgn(w) and ($\bar{w}$) are sign change due to number of reflections with respect to Φ^{+}_{even} and $\bar{\Phi}^{+}_{\mathrm{even}}$.

In general, in an equal rank embedding of Lie superalgebras r in g with $\Phi^{+}(r)\subset\Phi^{+}(g)$ and **C** as an index of the Weyl group of r in the Weyl group of g, a **C** multiplet of r is obtained by

$$c\cdot\Lambda:=c(\Lambda+\rho)_g-\rho_r,$$

where $c\in C$. Under the condition that $\Phi^{+}_{\mathrm{even,odd}}(r)$ is invariant under the action of c, i.e., $c\cdot\Phi^{+}_{\mathrm{even,odd}}(r)=\Phi^{+}_{\mathrm{even,odd}}(r)$, the typical Kac character formula of g module $V(\Lambda)$ can be written in terms of the r module $U(c\cdot\Lambda)$ as follows:

$$\mathrm{ch}\,V(\Lambda)=\frac{\Pi_{\beta\in\Phi^{+}_{\mathrm{odd}}(g)}(e^{\beta/2}+e^{-\beta/2})}{\Pi_{\alpha\in\Phi^{+}_{\mathrm{even}}(g)}(e^{\alpha/2}+e^{-\alpha/2})}\sum_{w\in W(g)}\mathrm{sgn}(w)e^{w(\Lambda+\rho_g)}$$

$$=\left(\frac{\Pi_{\beta\in\Phi^{+}_{\mathrm{odd}}(g/r)}(e^{\beta/2}+e^{-\beta/2})}{\Pi_{\alpha\in\Phi^{+}_{\mathrm{even}}(g/r)}(e^{\alpha/2}-e^{-\alpha/2})}\right)\sum_{c\in C}\mathrm{sgn}\,(c)\mathrm{ch}\,U(c\cdot\Lambda). \tag{5}$$

Similarly done, the typical Kac supercharacter becomes

$$\mathrm{sch}\,V(\Lambda)=\left(\frac{\Pi_{\beta\in\Phi^{+}_{\mathrm{odd}}(g/r)}(e^{\beta/2}-e^{-\beta/2})}{\Pi_{\alpha\in\Phi^{+}_{\mathrm{even}}(g/r)}(e^{\alpha/2}-e^{-\alpha/2})}\right)\sum_{c\in C}\mathrm{sgn}(c)\mathrm{sch}\,U(c\cdot\Lambda). \tag{6}$$

For type I Lie superalgebras, there are both typical and atypical representations. For the typical representation, superdimension, $\mathrm{sdim}V(\Lambda)=\dim V_{\mathrm{even}}(\Lambda)-\dim V_{\mathrm{odd}}(\Lambda)$, is equal to zero. Whereas, $\mathrm{sdim}V(\Lambda)$ of the atypical representation is not. Every type I odd root is zero length and $\Phi^{+}_{\mathrm{even}}=\overline{\Phi}^{+}_{\mathrm{even}}$. Since ρ_{odd} is invariant under the action of the Weyl group, i.e., $w\rho_{\mathrm{odd}}=\rho_{\mathrm{odd}}$. The type I typical Kac character formula (3) can be written as

$$\mathrm{ch}\,V(\Lambda)=\prod_{\beta_+\in\Phi_1^+}(1+e^{-\beta})\mathrm{ch}\,V_0(\Lambda)=\prod_{\beta_-\in\Phi_1^-}(1+e^{\beta_-})\mathrm{ch}\,V_0(\Lambda). \tag{7}$$

Multiplying out the product factor on the rhs of Eq. (7), we obtain a Chern character of an exterior algebra over g_{-1},

$$\prod_{\beta-\in\Phi_1^-}(1+e^{\beta_-})=\sum_{n=0}^{n=\dim(\Phi^-_{-1})}\mathrm{ch}(\wedge^n g_{-1})=\mathrm{ch}(\wedge g_{-1}).$$

So, Eq. (7) simply becomes

$$\mathrm{ch}\,V(\Lambda)=\mathrm{ch}(\wedge g_{-1})\mathrm{ch}\,V_0(\Lambda). \tag{8}$$

Similarly, the type I typical Kac supercharacter can be shown to be

$$\mathrm{sch}\,V(\Lambda)=\prod_{\beta_-\in\Phi_1^-}(1-e^{\beta_-})\mathrm{ch}\,V_0(\Lambda)=\mathrm{sch}(\wedge g_{-1})\mathrm{ch}\,V_0(\Lambda). \tag{9}$$

On the other hand, since $\{g_{-1},g_{-1}\}=0$, the universal enveloping algebra over g_{-1}, $\mathcal{U}(g_{-1})$, is isomorphic to the exterior algebra over $g_{-1},\wedge(g_{-1})$. The g module $V(\Lambda)$ can be induced by applying the antisymmetric combinations of the g_{-1} generators on $V_0(\Lambda)$, i.e.,

$$V(\Lambda)=\wedge(g_{-1})\otimes V_0(\Lambda)\simeq\mathcal{U}(g_{-1})\otimes V_0(\Lambda). \tag{10}$$

The character and supercharacter of Eq. (10) are exactly Eq. (8) and Eq. (9).

In an equal rank embedding, $g\supset r$, of type I, which has **C** as an index of $W(r)$ in $W(g)$ and is restricted to the condition that $\Phi^{+}(g)\supset\Phi^{+}(r)$, Eq. (7) becomes

$$\mathrm{ch}\,V(\Lambda)=\frac{1}{\Delta}\prod_{\beta_-\in\Phi_1^-(g/r)}(1+e^{\beta_-})\sum_{c\in C}\mathrm{sgn}(c)\left(\prod_{\beta_-\in\Phi_1^-(r)}(1+e^{\beta_-})\mathrm{ch}\,U_0(c\cdot\Lambda)\right),$$

i.e.,

$$\mathrm{ch}\, V(\Lambda)(\mathrm{ch}\, S^{+}-\mathrm{ch}\, S^{-})=\mathrm{ch}(\wedge(g_{-1}/r_{-1}))\sum_{c\in C}\ \mathrm{sgn}(c)\mathrm{ch}\, U(c\cdot\Lambda). \tag{11}$$

Similarly done for the supercharacter, we obtain

$$\mathrm{sch}\, V(\Lambda)(\mathrm{ch}\, S^{+}-\mathrm{ch}\, S^{-})=\mathrm{sch}(\wedge(g_{-1}/r_{-1}))\sum_{c\in C}\ \mathrm{sgn}(c)\mathrm{sch}\, U(c\cdot\Lambda). \tag{12}$$

Equation (11) and Eq. (12) correspond to

$$V(\Lambda)\otimes S^{+}-V(\Lambda)\otimes S^{-}=\wedge(g_{-1}/r_{-1})\otimes\sum_{c\in C}\ \mathrm{sgn}(c)U(c\cdot\Lambda). \tag{13}$$

Now, Eq. (13) can also be derived explicitly from Eq. (10). By decomposing the g_{-1} basis such that $g_{-1}=(g_{-1}/r_{-1})\oplus r_{-1}$, there exists a map $(g_{-1}/r_{-1})\oplus r_{-1}\mapsto(g_{-1}/r_{-1})\otimes 1+1\otimes r_{-1}$ from $(g_{-1}/r_{-1})\oplus r_{-1}$ to $(g_{-1}/r_{-1})\otimes r_{-1}$, which extends uniquely to an isomorphism of exterior algebra,

$$\wedge(g_{-1}/r_{-1}\oplus r_{-1})\simeq\wedge(g_{-1}/r_{-1})\otimes\wedge(r_{-1}).$$

By substituting the above equation into Eq. (10) and tensoring on both sides by $(S^{+}-S^{-})$, we obtain

$$\begin{aligned}V(\Lambda)\otimes S^{+}-V(\Lambda)\otimes S^{-}&=\wedge(g_{-1}/r_{-1})\otimes\sum_{c\in C}\ \mathrm{sgn}(c)(\wedge(r_{-1})\otimes U_0(c\cdot\Lambda))\\&=\wedge(g_{-1}/r_{-1})\otimes\sum_{c\in C}\ \mathrm{sgn}(c)U(c\cdot\Lambda),\end{aligned}$$

which is exactly Eq. (13).

For $\mathrm{osp}(1|2n)$ of type II Lie superalgebras, every $\mathrm{osp}(1|2n)$ odd root has length equal to one-half the short positive one and

$$\rho_{\mathrm{osp}(1|2n)}=\rho_{\mathrm{even}}-\rho_{\mathrm{odd}}=(n-\tfrac{1}{2},n-\tfrac{3}{2},\ldots,\tfrac{1}{2}).$$

Furthermore, every $\mathrm{osp}(1|2n)$ representation is typical, but $\dim V_{\mathrm{even}}(\Lambda)\neq\dim V_{\mathrm{odd}}(\Lambda)$. Nevertheless, the Kac character and supercharacter formulas of an $\mathrm{osp}(1|2n)$ representation can be shown to be similar to that of Lie algebra, i.e.,

$$\mathrm{ch}\, V(\Lambda)=\frac{\Sigma_{w\in W}\,\mathrm{sgn}(w)e^{w(\Lambda+\rho)}}{\Sigma_{w\in W}\,\mathrm{sgn}(w)e^{w(\rho)}}, \tag{14}$$

$$\mathrm{sch}\, V(\Lambda)=\frac{\Sigma_{w\in W}\,\mathrm{sgn}(\overline{w})e^{w(\Lambda+\rho)}}{\Sigma_{w\in W}\,\mathrm{sgn}(\overline{w})e^{w(\rho)}}. \tag{15}$$

The equal rank embedding of $\mathrm{osp}(1|2m)\times\mathrm{osp}(1|2n-2m)$ in $\mathrm{osp}(1|2n)$ is the only possible type with the full subsuperalgebra. In this case, all odd generators of g are completely eaten by the r. Whereas, the even generators of g that are left form the orthogonal complement basis to the even basis of r under the Killing form of g. Under the restriction to an r module, Eq. (14) and Eq. (15) simply become

$$\mathrm{ch}\, V(\Lambda) = \frac{1}{\Delta} \sum_{c \in C} \mathrm{sgn}(c) \mathrm{ch}\, U(c \cdot \Lambda), \tag{16}$$

and

$$\mathrm{sch}\, V(\Lambda) = \frac{1}{\Delta} \sum_{c \in C} \mathrm{sgn}(c) \mathrm{sch}\, U(c \cdot \Lambda). \tag{17}$$

Both Eq. (16) and Eq. (17) correspond to

$$V(\Lambda) \otimes S^{+} - V(\Lambda) \otimes S^{-} = \sum_{c \in C} \mathrm{sgn}(c) U(c \cdot \Lambda). \tag{18}$$

For the rest of type II, we use Kac character and the supercharacter formulas of equal rank embedding, Eq. (5) and Eq. (6). If g_1 and g_{-1} are vector spaces of odd generators, there is a canonical linear map from $\wedge^a g_1 \otimes \wedge^b g_{-1}$ to $\wedge^{a+b}(g_1 \oplus g_{-1})$, which takes $((g_1)_1 \wedge \cdots \wedge (g_1)_a) \otimes ((g_{-1})_1 \wedge \cdots \wedge (g_{-1})_b)$ to $((g_1)_1 \wedge \cdots \wedge (g_1)_a \wedge (g_{-1})_1 \wedge \cdots \wedge (g_{-1})_b)$. This determines a linear isomorphism,

$$\begin{aligned} \wedge(g_1 \oplus g_{-1}) &\simeq \bigoplus_{a=0}^{N} (\wedge^{N-a} g_1 \otimes \wedge^a g_{-1}) \\ &\simeq \wedge g_1 \otimes \wedge g_{-1}. \end{aligned}$$

The prefactor N_1 of the Kac character formula is generally the character of the exterior algebra over the direct sum of g_1 and g_{-1} vector spaces,

$$\begin{aligned} \prod_{\beta \in \Phi_1^{+}} (e^{\beta/2} + e^{-\beta/2}) &= \prod_{\beta_{\pm} \in \Phi_1^{\pm}} (e^{\beta_{+}/2} + e^{\beta_{-}/2}) \\ &= \mathrm{ch} \wedge (g_1 \oplus g_{-1}). \end{aligned}$$

Finally, the type II typical Kac character and supercharacter can be written as

$$\mathrm{ch}\, V(\Lambda)(\mathrm{ch}\, S^{+} - \mathrm{ch}\, S^{-}) = \mathrm{ch} \wedge (g_1/r_1 \oplus g_{-1}/r_{-1}) \sum_{c \in C} \mathrm{sgn}(c) \mathrm{ch}\, U(c \cdot \Lambda), \tag{19}$$

and

$$\mathrm{sch}\, V(\Lambda)(\mathrm{ch}\, S^{+} - \mathrm{ch}\, S^{-}) = \mathrm{sch} \wedge (g_1/r_1 \oplus g_{-1}/r_{-1}) \sum_{c \in C} \mathrm{sgn}(c) \mathrm{sch}\, U(c \cdot \Lambda), \tag{20}$$

which correspond to

$$V(\Lambda) \otimes S^{+} - V(\Lambda) \otimes S^{-} = \wedge (g_1/r_1 \oplus g_{-1}/r_{-1}) \otimes \sum_{c \in C} \mathrm{sgn}(c) U(c \cdot \Lambda). \tag{21}$$

IV. REPRESENTATIONS OF $\wedge g_{-1}$ OF TYPE I LIE SUPERALGEBRAS

For type I Lie superalgebras with a nondegenerate Killing form, the typical representation is induced by applying g_{-1} generators on a g_0 module. Therefore, we need to know an explicit representation of $\wedge g_{-1}$ in terms of a g_0 module. Let N be a dimension of Φ_1^{-}. The exterior algebra over g_{-1} is

$$\wedge g_{-1}=\bigoplus_{k=0}^{N}\wedge^{k}g_{-1},$$

with dimension

$$\dim(\wedge g_{-1})=\bigoplus_{k=0}^{N}\dim(\wedge^{k}g_{-1})=\sum_{k=0}^{N}\binom{N}{k}=2^{N},$$

and superdimension

$$\mathrm{sdim}(\wedge g_{-1})=\bigoplus_{k=0}^{N}\mathrm{sdim}(\wedge^{k}g_{-1})=\sum_{k=0}^{N}(-1)^{k}\binom{N}{k}=0.$$

Let $Q^{\dagger}_{i=1,2,\ldots,N}$ be N completely antisymmetric fermionic generators that transform in the fundamental representation of g_0. The fermionic generators generate even and odd modules that are isomorphic to a direct sum of two spinor representations of $\mathrm{so}(2N)$. One of the spinor representations of $\mathrm{so}(2N)$ is the even module of $\wedge g_{-1}$, called the bosonic module, and the other is the odd module of $\wedge g_{-1}$, called the fermionic module. Let T^{+} be the bosonic module and T^{-} be the fermionic module of $\mathrm{so}(2N)$, such that

$$\begin{aligned}T^{+}&=\wedge^{0}g_{-1}\oplus\wedge^{2}g_{-1}\oplus\wedge^{4}g_{-1}\oplus\cdots\\&=\big((Q_i^{\dagger})^{0}\oplus(Q_i^{\dagger})^{2}\oplus(Q_i^{\dagger})^{4}\oplus\cdots\big)|1\rangle_0\\&\equiv\text{bosonic module},\end{aligned}$$

and

$$\begin{aligned}T^{-}&=\wedge^{1}g_{-1}\oplus\wedge^{3}g_{-1}\oplus\wedge^{5}g_{-1}\oplus\cdots\\&=\big((Q_i^{\dagger})^{1}\oplus(Q_i^{\dagger})^{3}\oplus(Q_i^{\dagger})^{5}\oplus\cdots\big)|1\rangle_0\\&\equiv\text{fermionic module}.\end{aligned}$$

With a restriction to the g_0 module, the type I typical representation is

$$\begin{aligned}V(\Lambda)&=\wedge g_{-1}\otimes V_0(\Lambda)\\&=(T^{+}\oplus T^{-})\otimes V_0(\Lambda).\end{aligned}$$

For $\mathrm{su}(m|n)$, the $\wedge g_{-1}$ representation is isomorphic to a direct sum of two spinor representations of $\mathrm{so}(2mn)$. With a restriction to $\mathrm{su}(m)\times\mathrm{su}(n)\times\mathrm{u}(1)$, $Q_i^{\dagger}$ transforms as $(m,n)_{-1}$, where -1 is a $u(1)$ charge.

Ex. 1 $\mathrm{su}(2|1)$*:*

$$\dim(g_{-1})=2^{2}=1+2+1;$$

$Q_i^{\dagger}\sim 2_{-1}$,

$$T^{+}=1_0\oplus 1_{-2},$$

$$T^{-}=2_{-1},$$

$$V_{\mathrm{su}(2|1)}(\Lambda)=(T^{+}\oplus T^{-})\otimes V_{\mathrm{su}(2)\times\mathrm{u}(1)}(\Lambda).$$

Ex. 2 $\mathrm{su}(3|1)$*:*

$$\dim(g_{-1})=2^3=1+3+3+1;$$

$Q_i^\dagger \sim 3_{-1}$

$$T^+=1_0\oplus\bar{3}_{-2},$$

$$T^-=3_{-1}\oplus 1_{-3},$$

$$V_{\mathrm{su}(3|1)}(\Lambda)=(T^+\oplus T^-)\otimes V_{\mathrm{su}(3)\times \mathrm{u}(1)}(\Lambda).$$

Ex. 3 su(3|2):

$$\dim(g_{-1})=2^6=1+6+15+20+15+6+1;$$

$Q_i^\dagger \sim (3,2)_{-1}$,

$$T^+=(1,1)_0\oplus(\bar{3},3)_{-2}\oplus(6,1)_{-2}\oplus(6,1)_{-4}\oplus(3,3)_{-4}\oplus(1,1)_{-6},$$

$$T^-=(3,2)_{-1}\oplus(8,2)_{-3}\oplus(4,1)_{-3}\oplus(\bar{3},2)_{-5},$$

$$V_{su(3|2)}(\Lambda)=(T^+\oplus T^-)\otimes V_{\mathrm{su}(3)\times \mathrm{su}(2)\times \mathrm{u}(1)}(\Lambda).$$

Ex. 4 su(4|2):

$$\dim(g_{-1})=2^8=1+8+28+56+70+56+28+8+1,$$

$Q_i^\dagger \sim (4,2)_{-1}$,

$$T^+=(1,1)_0\oplus(10,1)_{-2}\oplus(6,3)_{-2}\oplus(20',1)_{-4}\oplus(15,3)_{-4}\oplus(5,1)_{-4}\oplus\overline{(10,1)}_{-6}\oplus(6,3)_{-6}\oplus(1,1)_{-8},$$

$$T^-=(4,2)_{-1}\oplus(20,1)_{-3}\oplus(\bar{4},4)_{-3}\oplus\overline{(20,1)}_{-5}\oplus(4,4)_{-5}\oplus(\bar{4},2)_{-7},$$

$$V_{\mathrm{su}(4|2)}(\Lambda)=(T^+\oplus T^-)\otimes V_{\mathrm{su}(4)\times \mathrm{su}(2)\times \mathrm{u}(1)}(\Lambda).$$

For osp$(2|2n)$, $\wedge g_{-1}$ representation is a direct sum of two spinor representations of so$(4n)$. With restriction to sp$(2n)\times u(1)$, $Q_i^\dagger$ transforms as $(2n)_{-1}$ with -1 as a $u(1)$ charge.

Ex. 5 osp(2|4):

$$\dim(g_{-1})=2^4=1+4+6+4+1;$$

$Q_i^\dagger \sim 4_{-1}$,

$$T^+=1_0\oplus 5_{-2}\oplus 1_{-2}\oplus 1_{-4},$$

$$T^-=4_{-1}\oplus 4_{-3},$$

$$V_{\mathrm{osp}(2|4)}(\Lambda)=(T^+\oplus T^-)\otimes V_{\mathrm{sp}(4)\times u(1)}(\Lambda).$$

Ex. 6 osp(2|6):

$$\dim(g_{-1})=2^6=1+6+15+20+15+6+1;$$

$Q_i^\dagger \sim 6_{-1}$,

$$T^+=1_0\oplus 14_{-2}\oplus 1_{-2}\oplus 14_{-4}\oplus 1_{-4}\oplus 1_{-6},$$

$$T^- = 6_{-1} \oplus 14'_{-3} \oplus 6_{-3} \oplus 6_{-5},$$

$$V_{\mathrm{osp}(2|6)}(\Lambda) = (T^+ \oplus T^-) \otimes V_{\mathrm{sp}(6)\times \mathrm{u}(1)}(\Lambda).$$

V. BUILDING TYPE I C MULTIPLETS AND KOSTANT'S CUBIC DIRAC OPERATOR

For type I Lie superalgebras, there is a remarkable point we would like to mention. From the multiplet spectrum of equal rank embedding of reductive Lie algebras, $r_0 \subset g_0$, we can build the spectrum of the type I Lie superalgebras on top of them by simply tensoring them with a module generated by r_{-1} generators. The r_{-1} generators transform in the fundamental representation of r_0.

Define $\wedge(r_{-1}) = R^+ \oplus R^-$ such that

$$U(\Lambda) = \wedge(r_{-1}) \otimes U_0(\Lambda) = (R^+ \oplus R^-) \otimes U_0(\Lambda).$$

Tensoring on both side of Eq. (2) with $\wedge(r_{-1})$, we obtain

$$\wedge(r_{-1}) \otimes (V_0(\Lambda)\otimes S^+ - V_0(\Lambda)\otimes S^-) = \sum_{c\in C} \mathrm{sgn}(c)(\wedge(r_{-1})\otimes U_0(c\cdot\Lambda))$$

$$= \sum_{c\in C} \mathrm{sgn}(c) U(c\cdot \Lambda). \tag{22}$$

Under restriction to an r_0 module, whenever $V_0(\Lambda)$ on the lhs of Eq. (22) is $\mathrm{su}(n)\times \mathrm{u}(1)$, or $\mathrm{sp}(2n)\times \mathrm{u}(1)$, there is an emergence of a type I typical **C** multiplet on the rhs. Notice in the case that $U(c\cdot\Lambda)$ is $\mathrm{su}(m|n)$, the representations of $R^\pm$ are similar to those of $S^\pm$ except $u(1)$ values.

Let $g = r \oplus p$ be Lie superalgebras where p is the orthogonal complement to r under the nondegenerate Killing form of g. Let $p = p_{\mathrm{even}} \oplus p_{\mathrm{odd}} = p_0 \oplus p_1 \oplus p_{-1}$. Tensoring on both sides of Eq. (22) by $\wedge p_{-1}$, we get

$$V(\Lambda)\otimes S^+ - V(\Lambda)\otimes S^- = \wedge p_{-1} \otimes \sum_{c\in C} \mathrm{sgn}(c) U(c\cdot\Lambda), \tag{23}$$

which is exactly Eq. (13). Now, we need to get rid of $\wedge p_{-1}$ on the rhs of Eq. (23) by mapping it into identity. To do so, we do a tensor product on both sides of Eq. (23) by the following contraction, sometimes called an internal product, on exterior powers between vector space and its dual:

$$\wedge^N(p_{\mathrm{odd}}) = \wedge^N p_1 \otimes \wedge^N p_{-1} = \mathbf{1}.$$

Equation (23) becomes

$$(\wedge^N p_{\mathrm{odd}}) \otimes V(\Lambda) \otimes (S^+ - S^-) = \mathbf{1} \otimes \sum_{c\in C} \mathrm{sgn}(c) U(c\cdot\Lambda). \tag{24}$$

As already known, the character form of a Dirac operator, $\partial\!\!\!/ \in C(p)$, with a map $\partial\!\!\!/ : S^\pm \to S^\mp$, is

$$\mathrm{ch}(\partial\!\!\!/) = \mathrm{ch}(S^+) - \mathrm{ch}(S^-).$$

The character of Eq. (1) implies that there exists a Kostant's Dirac operator, $\mathcal{K}\!\!\!/ \in \mathcal{U}(g_{\mathrm{even}}) \otimes C(p_{\mathrm{even}})$, with a map

$$\mathcal{K}\!\!\!/_\lambda : V_\lambda \otimes S^\pm \to V_\lambda \otimes S^\mp. \tag{25}$$

Similarly, the character of Eq. (24) suggests that there exists the operator for Lie superalgebras, $\not{\mathcal{K}}_\Lambda \in \mathcal{U}(p_{\rm odd}) \otimes \mathcal{U}(g_{\rm even} \oplus g_{\rm odd}) \otimes C(p_{\rm even})$, with a map

$$\not{\mathcal{K}}_\Lambda : V(\Lambda) \otimes S^{\pm} \rightarrow V(\Lambda) \otimes S^{\mp}. \tag{26}$$

VI. KOSTANT's CUBIC DIRAC OPERATOR FOR AN EQUAL RANK EMBEDDING OF LIE SUPERALGEBRAS

Let L_i and F_a be even and odd generators, respectively, for Z_2-graded Lie superalgebras such that

$$[L_i, L_j] = f_{[ijk]} L_k,$$

$$\{F_a, F_b\} = f_{i(ab)} L_i,$$

and

$$[L_i, F_a] = f_{[ia]b} F_b,$$

where $i,j,k = 1,2,...,\dim(g_0)$ and $a,b = 1,2,...,\dim(g_1)$. The Kostant's cubic Dirac operator of Lie superalgebras is extended from that of Lie algebras by adding just two terms; a linear term in odd operators and a structure constant term, i.e.,

$$\not{\mathcal{K}}_\Lambda = \not{\mathcal{K}}_\Lambda^0 + \not{\mathcal{K}}_\Lambda^1, \tag{27}$$

where

$$\not{\mathcal{K}}_\Lambda^0 = \gamma_i L_i - \tfrac{1}{2} \gamma_{[ijk]} f_{[ijk]} \tag{28}$$

and

$$\not{\mathcal{K}}_\Lambda^1 = \alpha_a F_a - \tfrac{1}{2} \gamma_i \alpha_{(ab)} f_{i(ab)}. \tag{29}$$

The sum over all indices is assumed in the above equations, where $[\cdots]$ in the subscript is for the antisymmetric sum and $(\cdots)$ for the symmetric sum. Equation (28) is the Kostant' cubic Dirac operator for reductive Lie algebras. The γ matrices associated to even generators are normalized so that

$$\{\gamma_i, \gamma_j\} = \delta_{ij},$$

which gives

$$\{\gamma_{i'}, \gamma_{[ijk]}\} = \delta_{i'k} \gamma_{[ij]}.$$

The α matrices associated to odd generators are subjected to the following conditions.

(I) $\{\gamma_i, \alpha_a\} = 0$. This relation is consistent with the antisymmetric property of the product of even and odd generators.

(II) $\{\alpha_a, [\alpha_b, \alpha_c]\} + \{[\alpha_a, \alpha_c], \alpha_b\} = [\alpha_{(ab)}, \alpha_c] = 0$. This property is due to invariance of odd generators under the Killing form.

Squaring Eq. (28), we get

$$\begin{aligned}(\not{\mathcal{K}}_\Lambda^0)^2 &= \tfrac{1}{2} \gamma_{(ij)} \{L_i, L_j\} + \tfrac{1}{2} \gamma_{[ij]} [L_i, L_j] - \tfrac{1}{2} \{\gamma_{i'}, \gamma_{[ijk]}\} f_{[ijk]} L_{i'} + (\tfrac{1}{2} \gamma_{[ijk]} f_{[ijk]})^2 \\ &= \tfrac{1}{2} \gamma_{(ij)} \{L_i, L_j\} + (\tfrac{1}{2} \gamma_{[ijk]} f_{[ijk]})^2. \end{aligned} \tag{30}$$

Notice that the linear terms in even generators cancel each other. Thus, Eq. (30) is explicitly invariant under the action of even and odd generators. The first term of Eq. (30) is the quadratic Casimir operator of reductive Lie algebras,

$$C_2^0(\Lambda)=\tfrac{1}{2}\gamma_{(ij)}\{L_i,L_j\}. \tag{31}$$

Since ρ_0, one-half the sum of positive even roots can be identified as

$$\rho_0=\tfrac{1}{2}\gamma_{[ijk]}f_{[ijk]}. \tag{32}$$

The second term of Eq. (30) is the Freudenthal–de Vries' strange formula,

$$\begin{aligned}(\rho,\rho)_0&=(\tfrac{1}{2}\gamma_{[ijk]}f_{[ijk]})^2=\tfrac{1}{24}\ \dim(g_0)h_0^\vee(\theta,\theta)_0\\&=\tfrac{1}{24}\ \dim(g_0)C_2^{0\mathrm{ad}}.\end{aligned} \tag{33}$$

Where, in the above equation, $h^\vee$ is the dual Coxeter number, θ is the highest root, and C_2^{ad} is the quadratic Casimir value in the adjoint representation.

Squaring Eq. (29), we get

$$\begin{aligned}(\not{\mathcal{K}}_\Lambda^1)^2&=\tfrac{1}{2}\alpha_{[ab]}[F_a,F_b]+\tfrac{1}{2}\alpha_{(ab)}\{F_a,F_b\}-\tfrac{1}{2}\gamma_i[\alpha_{(ab)},\alpha_{a'}]f_{i[ab]}F_{a'}+(\tfrac{1}{2}\gamma_i\alpha_{(ab)}f_{i(ab)})^2\\&=\tfrac{1}{2}\alpha_{[ab]}[F_a,F_b]+\tfrac{1}{2}\alpha_{(ab)}f_{i(ab)}L_i+(\tfrac{1}{2}\gamma_i\alpha_{(ab)}f_{i(ab)})^2.\end{aligned} \tag{34}$$

In contrary to the even part, Eq. (34) by itself is not invariant due to the presence of a linear term in even generators. The linear term in L_i will be cancelled out by one of the the cross terms of the square of the combined even and odd Kostant's cubic Dirac operator,

$$\begin{aligned}(\not{\mathcal{K}}_\Lambda)^2=&(\not{\mathcal{K}}_\Lambda^0)^2+(\not{\mathcal{K}}_\Lambda^1)^2+\{\gamma_i,\alpha_a\}L_iF_a-\tfrac{1}{2}\{\gamma_{[ijk]},\alpha_a\}f_{[ijk]}F_a-\tfrac{1}{2}\{\gamma_i,\gamma_{i'}\}\alpha_{(ab)}f_{i'(ab)}L_i\\&-\tfrac{1}{2}\gamma_i[\alpha_{(ab)},\alpha_{a'}]f_{i(ab)}F_{a'}+\tfrac{1}{2}\{\gamma_{[ijk]},\gamma_{i'}\}\alpha_{(ab)}f_{[ijk]}f_{i'(ab)}\\&=\gamma_{(ij)}L_{(i}L_{j)}+\alpha_{[ab]}F_{[a}F_{b]}+(\tfrac{1}{2}\gamma_{[ijk]}f_{[ijk]})^2+(\tfrac{1}{2}\gamma_i\alpha_{(ab)}f_{i(ab)})^2+\tfrac{1}{2}\gamma_{[ij]}\alpha_{(ab)}f_{[ijk]}f_{k(ab)}.\end{aligned} \tag{35}$$

Since ρ_1, one-half the sum of positive odd roots, can be identified as

$$\rho_1=-\tfrac{1}{2}\gamma_i\alpha_{(ab)}f_{i(ab)}. \tag{36}$$

Recall that, for Lie superalgebras, $\rho=\rho_0-\rho_1$. Equation (35) can be simply written as

$$(\not{\mathcal{K}}_\Lambda)^2=C_2(\Lambda)+(\rho,\rho), \tag{37}$$

where

$$C_2(\Lambda)=\gamma_{(ij)}L_{(i}L_{j)}+\alpha_{[ab]}F_{[a}F_{b]}.$$

The Laplacian operator turns out to be invariant under the action of even and odd generators of g again. For Lie superalgebras, the generalization of the Freudenthal–de Vries strange formula still holds,[7]

$$(\rho,\rho)=\frac{h^\vee}{24}(\dim g_0-\dim g_1), \tag{38}$$

where $h^\vee$'s, the dual Coxeter numbers of g, are given in Table I.

In an equal rank embedding of Lie superalgebras, $g\supset r$, with $\Phi_g\supset\Phi_r$, let the even and odd generators, L_i and F_a, span the basis of g. According to the $g=r\oplus p$ decomposition, $L_{\hat{i}}$ and $F_{\hat{a}}$ span the basis of r and L_I and F_A span the basis of p, the orthogonal complement of r under the Killing form of g. The Kostant's cubic Dirac operator on $p=g/r$ is

$$\not{\mathcal{K}}_p=\not{\mathcal{K}}_g-\not{\mathcal{K}}_r=\gamma_IL_I+\alpha_AF_A-\tfrac{1}{2}\gamma_{[IJK]}f_{[IJK]}-\tfrac{1}{2}\gamma_I\alpha_{(AB)}f_{I(AB)}, \tag{39}$$

TABLE I. Dual Coxeter numbers of basic Lie superalgebras.

g	$h^\vee$
$\mathrm{su}(m\|n)$	$\|m-n\|$
$\mathrm{osp}(2\|2n)$	n
$\mathrm{osp}(2m+1\|n)$	$2(m-n)-1$ if $m>n, n-m+\frac{1}{2}$ if $m\leqslant n$
$\mathrm{osp}(2m\|n)$	$2(m-n-1)$ if $m\geqslant n, n-m+1$ if $m<n+1$
$\mathrm{osp}(4\|2;\alpha)$	0
$F(4)$	3
$G(3)$	2

where $f_{[IJK]}$ and $f_{I(AB)}$ are the structure constants of g that are not in r. Since, under the Killing form of g, r and p bases are orthogonal to each other,

$$\{\gamma_{\hat{i}},\gamma_I\}=\{\gamma_{\hat{i}},\alpha_A\}=\{\alpha_{\hat{a}},\gamma_I\}=\{\alpha_{\hat{a}},\alpha_A\}=0.$$

As a result, we have

$$\{\not{\mathcal{K}}_r,\not{\mathcal{K}}_p\}=0.$$

The square of Kostant's coset cubic Dirac operator simply is

$$(\not{\mathcal{K}}_p)^2=(\not{\mathcal{K}}_g)^2-(\not{\mathcal{K}}_r)^2=(C_2+(\rho,\rho))_g-(C_2+(\rho,\rho))_r\,. \tag{40}$$

Notice that both $(\not{\mathcal{K}}_g)^2$ and $(\not{\mathcal{K}}_r)^2$ commute with the generators of r. Hence, $(\not{\mathcal{K}}_p)^2$ is also invariant under the action of r.

In conclusion, we have derived the Kac character formulas and have constructed Kostant's cubic Dirac operator for equal rank embeddings of Lie superalgebras. In case of reductive Lie algebras, the coset space method of equal rank embeddings led to a Kazama–Suzuki model construction of a new class of unitary $N=2$ superconformal theories[8] and a subclass of the construction could be represented by Landau–Ginzburg models.[9] In case of Lie superalgebras, it deserves to be pursued whether there will be any relevance of the coset superspace method to a construction of any physical model.

ACKNOWLEDGMENTS

The author would like to thank Professor P. Ramond for invaluable supervision and Professor S. Sternberg for pointing out Ref. 6 and Ref. 7 into attention. This work is supported by the Ministry of University Affairs, Thailand.

[1] T. Pengpan and P. Ramond, Phys. Rep. **315**, 137 (1999).
[2] B. Gross, B. Kostant, P. Ramond, and S. Sternberg, Proc. Natl. Acad. Sci. USA **95**, 8441 (1998).
[3] B. Kostant, ''A cubic Dirac operator and the emergence of the Euler number multiplets of representations for equal rank subgroups,'' to be published in Duke Math. J.
[4] L. Brink and P. Ramond, hep-th/9908208.
[5] V. G. Kac, Adv. Math. **26**, 8 (1977); V. G. Kac, *Lectures Notes in Mathematics* (Springer-Verlag, Berlin, 1978), Vol. 676, pp. 597–626.
[6] L. Corwin, Y. Ne'eman, and S. Sternberg, Rev. Mod. Phys. **47**, 573 (1975).
[7] V. G. Kac and M. Wakimoto, in *Lie Theory and Geometry: In Honor of Bertram Kostant*, edited by J. Brylinski, R. Brylinski, V. Guillemin, V. Kac, and B. Kostant (Birkhauser, Boston, 1994), pp. 421–423.
[8] Y. Kazama and H. Suzuki, Nucl. Phys. **B321**, 232 (1989).
[9] W. Lerche, C. Vafa, and N. Warner, Nucl. Phys. **B324**, 427 (1989).

Cones and causal structures on topological and differentiable manifolds

M. Rainer[a)]
Center for Gravitational Physics and Geometry, 104 Davey Laboratory, The Penn State University, University Park, Pennsylvania 16802-6300 and Mathematische Physik I, Mathematisches Institut, Universität Potsdam, PF 601553, D-14415 Potsdam, Germany

(Received 4 June 1999; accepted for publication 14 July 1999)

General definitions for causal structures on manifolds of dimension $d+1>2$ are presented for the topological category and for any differentiable one. Locally, these are given as cone structures via local (pointwise) homeomorphic or diffeomorphic abstraction from the standard null cone variety in $\mathbb{R}^{d+1}$. Weak ($\mathscr{C}$) and strong ($\mathscr{C}^m$) local cone (LC) structures refer to the cone itself or a manifold thickening of the cone, respectively. After introducing cone (C-)causality, a causal complement with reasonable duality properties can be defined. The most common causal concepts of space–times are generalized to the present topological setting. A new notion of precausality precludes inner boundaries within future/past cones. LC-structures, C-causality, a topological causal complement, and precausality may be useful tools in conformal and background independent formulations of (algebraic) quantum field theory and quantum gravity. © *1999 American Institute of Physics.* [S0022-2488(99)00712-4]

I. INTRODUCTION

While classical general relativity usually employs a Lorentzian space–time metric, all genuine approaches to quantum gravity are free of such a metric background. This poses the question of whether there still exists a notion of structure which captures some essential features of light cones and their mutual relations in manifolds in a purely topological manner without a priori recursion to a Lorentzian metric or a conformal class of such metrics. Below we will see that the answer is positive.

It is a well known folk theorem that the causal structure on a Lorentzian manifold determines its metric up to conformal transformations. In Refs. 1 and 2 a path topology for strongly causal space–times was defined which then determined their differential, causal, and conformal structure. In Ref. 3 it was shown that the conformal class of a Lorentzian metric can be reconstructed from the characteristic surfaces of the manifold. Similarly, Ref. 4 gives a nice proof that the null cones determine the Lorentzian metric (modulo global sign) up to a conformal factor. All these previous results already indicate that the notion of a causal structure could exist indeed in a different and possibly more general setting than that of Lorentzian space–times. However, all the previously mentioned investigations in the literature assume a priori the existence of some undetermined Lorentzian metric and then show that it can be determined modulo conformal transformation uniquely by some other structure.

Motivated by the requirements on suitable structures for a theory of quantum gravity, in this paper new notions of causal structure are developed which do not assume a priori the existence of any (Lorentzian) metric or conformal metric but rather work on arbitrary topological and differential manifolds.

In Sec. II weak ($\mathscr{C}$) and strong ($\mathscr{C}^m$) local cone (LC) structures are defined on any topological (or differentiable) manifold M. These structures are given by continuous (or differentiable) fami-

[a)]Electronic mail: mrainer@rz.uni-potsdam.de

lies of pointwise homeomorphisms from the standard null cone variety in $\mathbb{R}^{d+1}$ or a manifold thickening thereof, respectively, into M. In the differentiable case it turns out that a strong LC structure implies the existence of a conformal Lorentzian metric, while a weak LC structure already implies its uniqueness should one exist. However, the metric resulting from a strong LC structure contains only pointwise information about the asymptotic structure of the cone at the vertex. Within a given manifold thickening of the cone at a given point of M, the cone in any neighborhood of the vertex need a priori not at all be related to the null structure spanned by the null geodesics of this metric. However, if such a relationship holds in some region, then all the cones in that region are consistent with each other and this way yield a notion of causality.

In Sec. III we provide precisely those definitions of causality which allow us to formulate the consistency of different strength for cones at different points, in some or any open region in M. Cone (C-) causality allows first of all the definition of a causal complement with reasonable properties. It enables us also to define in a topological (differentiable) manner spacelike, null, and timelike curves. We discuss C-causality also in the particular context of a fibration. Generalizations of the most common causality notions for space–times in purely topological terms are provided. In the case of Lorentzian manifolds these notions agree with the usual ones and they assume their usual hierarchy. Finally, precausality is defined as a notion which makes the future and the past of any cone homeomorphic to the future and the past of the standard cone $\mathcal{C}$ in $\mathbb{R}^{d+1}$, respectively.

The discussion points out some of the major open issues which require further investigation. It addresses also the issues of causal diffeomorphisms, foliations, and possible restrictions of the cone structure and causality from the manifold to an embedded graph therein, giving also a perspective for possible applications in conformal and background independent quantum field theories and quantum gravity.

Here and below a CAT manifold refers to a Hausdorff (T_2) space with a CAT structure, where CAT$=\mathcal{C}^0$ (the topological category) or CAT$\subset\mathcal{C}^1$ (any differentiable category). If CAT$\subset\mathcal{C}^1$, a CAT homeomorphism is a diffeomorphism and a CAT continuous map is a differentiable map. For differentiable categories we also define $\mathrm{CAT}_{-1}:=\mathcal{C}^r$ if CAT$=\mathcal{C}^{r+1}$, $\mathrm{CAT}_{-1}:=\mathcal{C}^\infty$ if CAT $=\mathcal{C}^\infty$, and $\mathrm{CAT}_{-1}:=\mathcal{C}^\omega$ if CAT$=\mathcal{C}^\omega$.

II. LOCAL CONE (LC) STRUCTURES OF MANIFOLDS

In this section we derive local notions of a cone structure on a topological $d+1$-dimensional manifold M (CAT$\subset\mathcal{C}^0$). Let

$$\mathscr{C}:=\{x\in\mathbb{R}^{d+1}:x_0^2=(x-x_0e_0)^2\},\mathscr{C}^+:=\{x\in\mathscr{C}:x_0\geqslant 0\},\mathscr{C}^-:=\{x\in\mathscr{C}:x_0\leqslant 0\} \tag{2.1}$$

be the standard (unbounded double) light cone, and the forward and backward subcones in $\mathbb{R}^{d+1}$, respectively.

The standard open interior and exterior of $\mathscr{C}$ is defined as

$$\mathcal{I}:=\{x\in\mathbb{R}^{d+1}:x_0^2>(x-x_0e_0)^2\},\mathcal{E}:=\{x\in\mathbb{R}^{d+1}:x_0^2<(x-x_0e_0)^2\}. \tag{2.2}$$

A *manifold thickening* with thickness $m>0$ is given as

$$\mathscr{C}^m=\{x\in\mathbb{R}^{d+1}:|x_0^2-(x-x_0e_0)^2|<m^2\}. \tag{2.3}$$

The characteristic topological data of the standard cone is encoded in the topological relations of all its manifold subspaces (which includes in particular also the singular vertex O) and among each other.

Typical (CAT) manifold subspaces of $\mathscr{C}$ are the standard future and past cones $\mathscr{C}^\pm$, and the standard light rays,

$$l(n):=\{x\in\mathcal{C}:x_0=(x,n)\}, \tag{2.4}$$

where $n \in S^{d-1} \subset p$ is a normal direction in the d-dimensional hyperplane $p := \{x \in \mathbb{R}^{d+1} : (x,y) = 0, \forall y \in a\}$ perpendicular to the cone axis $a := \{x \in \mathbb{R}^{d+1} : x = \lambda e_0, \lambda \in \mathbb{R}\}$.

The topological relations between all the CAT manifold subspaces of the cone are the natural data which will be required to be conserved under a homeomorphism of the cone as a topological space into the manifold M at any point p.

Let τ denote the closed sets of the manifold topology of $\mathscr{C} - O$. The set $\mathscr{C}$ can either inherit the induced topology τ_1 from $\mathbb{R}^{d+1}$ which is T_1 but not T_2 (Hausdorff) or it can be equipped with a more coarse subtopology defined in terms of closed sets as $\tau_2 := \{\{0\} \cup V : V \in \tau\} \cup \{V \in \tau\}$ which is Hausdorff. However, τ_2 places geometrically unnatural restrictions on possible submanifolds of $\mathscr{C}$. Hence, unless specified otherwise, $\mathscr{C}$ will be equipped with τ_1.

Definition 1: Let M be a CAT manifold. A (CAT) *(null) cone* at $p \in \operatorname{int} M$ is the homeomorphic image $\mathscr{C}_p := \phi_p \mathscr{C}$ of a homeomorphism of topological spaces $\phi_p : \mathscr{C} \to \mathscr{C}_p \subset M$ with $\phi_p(0) = p$, such that we have the following.

(i) Every (CAT) submanifold $N \subset \mathscr{C}$ is mapped (CAT) homeomorphically on a submanifold $\phi_p(N) \subset M$.

(ii) For any two submanifolds $N_1, N_2 \subset \mathscr{C}$ there exist homeomorphisms $\phi_p(N_1) \cap \phi_p(N_2) \cong N_1 \cap N_2$ and $\phi_p(N_1) \cup \phi_p(N_2) \cong N_1 \cup N_2$ of (CAT) manifolds if these are (CAT) manifolds and as topological spaces otherwise

(iii) If CAT $\subset \mathcal{C}^1$ then for any two CAT curves $c_1, c_2 :]-\epsilon, \epsilon[\to \mathscr{C}$ with $c_1(0) = c_2(0) = p$ it holds that $T_0 c_1 = T_0 c_2 \Rightarrow T_p(\phi_p \circ c_1)|_{]-\epsilon,\epsilon[} = T_p(\phi_p \circ c_2)|_{]-\epsilon,\epsilon[}$.

Condition (iii) says that in the differentiable case the well defined notion of transversality of intersections at the vertex is preserved by ϕ_p.

On each homeomorphic cone $\mathscr{C}_p$ at any $p \in \operatorname{int} M$, the topology τ_1 or τ_2 of $\mathscr{C}$ yields under ϕ_p likewise a non-Hausdorff T_1 topology $\phi_p(\tau_1)$ or a T_2 one $\phi_p(\tau_2)$. However, $\phi_p \circ \tau_2$ would unnaturally restrict the possible submanifolds of $\mathscr{C}$, while $\phi_p \circ \tau_1$ is consistent with the topology induced from M.

Definition 2: An *(ultraweak) cone structure* on M is an assignment $\operatorname{int} M \ni p \to \mathscr{C}_p$ of a cone at every $p \in \operatorname{int} M$.

A cone structure on M can, in general, be rather wild with cones at different points totally unrelated unless we impose a topological connection between the cones at different points. Most naturally the connection is provided by continuity of the family $\{\mathscr{C}_p\}$. This allows us to define local cone (LC) structures.

Definition 3: Let M be a CAT manifold. A *weak ($\mathscr{C}$) local cone (LC)-structure* on M is a cone structure which is (CAT) continuous, i.e., $\{p \to \mathscr{C}_p\}$ is a (CAT) continuous family.

Given a cone structure one wants to know first of all under which conditions, for given $p \in \operatorname{int} M$, an exterior and interior of the cone can be distinguished locally, i.e., for any open neighborhood $U \ni p$ within $(M - \mathscr{C}_p) \cap U$.

Proposition 1: Let $\forall p \in \operatorname{int} M$ exist open (CAT) submanifolds $\mathcal{I}_p$ and $\mathcal{E}_p$ such that the interior of M decomposes in the disjoint union $\operatorname{int} M = \mathscr{C}_p \mathring{\cup} \mathcal{I}_p \mathring{\cup} \mathcal{E}_p$.

(i) Then $\mathcal{I}_p$ and $\mathcal{E}_p$ can be topologically distinguished locally in any neighborhood of the vertex p if and only if for any neighborhood $U \ni p$ it holds $(\mathcal{I}_p|_U) \not\cong (\mathcal{E}_p|_U)$.

(ii) Given any neighborhood $U \ni p$ assume $\exists k \in \mathbb{N}_0 : \Pi_k(\mathcal{I}_p|_U) \neq \Pi_k(\mathcal{E}_p|_U)$. Then $\mathcal{I}_p$ and $\mathcal{E}_p$ can be topologically distinguished locally in any neighborhood of the vertex p.

Proof: (i) follows from $U - \mathscr{C}_p|_U = \mathcal{I}_p|_U \mathring{\cup} \mathcal{E}_p|_U$. (ii) holds because homotopy groups are topological invariants. □

Note that, although $\mathscr{C}_p = \phi_p(\mathscr{C})$, $\mathcal{I}$ and $\mathcal{E}$ here need not be homeomorphic to $\phi_p(\mathcal{I})$ and $\phi_p(\mathcal{E})$, respectively. The notion of precausality (see below in Sec. III) is set up to ensure $\mathcal{E}_p \cong \phi_p(\mathcal{E})$.

A weak LC structure at each point $p \in \operatorname{int} M$ defines a characteristic topological space $\mathscr{C}_p$ of codimension 1 which is Hausdorff everywhere but at p. In particular $\mathscr{C}_p$ does not contain any open $U \ni p$ from the manifold topology of M. However, stronger structures can be defined as follows.

Definition 4: Let M be a CAT manifold. A (CAT) *(manifold) thickened cone* of thickness

$m > 0$ at $p \in \text{int}\, M$ is the (CAT) homeomorphic image $\mathscr{C}_p^m := \phi_p \mathscr{C}^m$ of a (CAT) homeomorphism of manifolds $\phi_p : \mathscr{C} \to \mathscr{C}_p \subset M$ with $\phi_p(0) = p$.

Note that due to the manifold property the notion of a thickened cone is much more simple than that of a cone itself. It also clear that now the only consistent topology on $\mathscr{C} \subset \mathscr{C}_p$ is τ_1 and correspondingly $\phi_p(\tau_1)$ on $\mathscr{C}_p \subset \mathscr{C}_p^m$.

Definition 5: A *thickened cone structure* on M is an assignment $\text{int}M \ni p \mapsto \mathscr{C}_p^{m(p)}$ of a thickened cone at every $p \in \text{int}\, M$.

Note that in general the thickness m can vary from point to point in M. Here $m : M \to \mathbb{R}_+$ is an a priori not necessarily continuous function. However, an important case even more special than the continuous one is that of constant m.

Definition 6: A *homogeneously thickened cone structure* on M is a thickened cone structure $\text{int}M \ni p \to \mathscr{C}_p$ with constant thickness m.

Although homogeneity might be too restrictive, at least continuity of structures on M is a natural assumption in the topological category.

Definition 7: Let M be a CAT manifold. A *strong* ($\mathscr{C}^m$) *LC structure* on M is a (CAT) continuous family of (CAT) homeomorphisms $\phi_p : \mathscr{C}^m \to \mathscr{C}_p^{m(p)} \subset M$ with $\phi_p(0) = p$ and such that the thickness m is a CAT function on M.

In particular the conditions of (ii) in Proposition 1 apply for all manifolds of dimension $d + 1 > 2$ with a strong LC-structure, while a weak LC-structure at $p \in \text{int}\, M$ may not be able to distinguish $\mathcal{I}_p|_U$ and $\mathcal{E}_p|_U$ within any $U \ni p$.

Theorem 1: Let M carry a strong LC structure. At any $p \in \text{int}\, M$ there exists an open $U \ni p$ such that for $d := \dim M - 1 > 0$ it is $|\Pi_0(\mathcal{I}_p|_U)| = 2$ and $\Pi_{d-1}(\mathcal{E}_p|_U) = \Pi_{d-1}(S^{d-1})$; for $d > 1$ it is $\Pi_{d-1}(\mathcal{I}_p|_U) = 0$ and $|\Pi_0(\mathcal{E}_p|_U)| = 1$; for $d = 1$ it is $\Pi_{d-1}(\mathcal{I}_p|_U) = \Pi_{d-1}(\mathcal{E}_p|_U) = \Pi_0(S^0)$, i.e., $|\Pi_0(\mathcal{I}_p|_U)| = |\Pi_0(\mathcal{E}_p|_U)| = 2$, and in dimension $d = 0$ it is $\mathcal{I}_p = \mathcal{E}_p = \emptyset$.

Proof: For all $p \in \text{int}\, M$ the strong LC structure provides a thickened cone $\mathscr{C}_p^{m(p)}$. Since $m(p) > 0$, $\mathscr{C}_p^{m(p)}$ contains always a neighborhood $U \ni p$ homeomorphic to a neighborhood $\phi_p^{-1}(U) \ni 0$ of the standard cone which in any dimension has the desired properties. □

At any interior point $p \in \text{int}\, M$ the open exterior $\mathcal{E}_p$ and the open interior $\mathcal{I}_p$ of the cone $\mathscr{C}_p$ are locally topologically distinguishable for $d > 1$, indistinguishable for $d = 1$, and empty for $d = 0$. With a strong LC structure $\mathcal{I}_p|_U \neq \mathcal{E}_p|_U \forall U \ni p \Leftrightarrow d + 1 > 2$, whence locally in any neighborhood $U \ni p$ the interior and exterior of $\mathscr{C}_p \cap U$ at p in U has an intrinsic invariant meaning. $\mathscr{C}_p|_U$ divides $U - \mathcal{C}_p|_U$ in three open submanifolds: a noncontractable exterior $\mathcal{E}_p|_U$ plus two contractable connected components of $\mathcal{I} =: \mathcal{F}|_U \cup \mathcal{P}|_U$, the *local future* $\mathcal{F}_p|_U$ and the *local past* $\mathcal{P}_p|_U$ with $\partial \mathcal{F}_p|_U = \mathscr{C}_p^+|_U$ where $\mathscr{C}_p^+ := (\phi_p \mathscr{C}^+)$, and $\partial \mathcal{P}_p|_U = \mathscr{C}_p^-|_U$ where $\mathscr{C}_p^- := \phi_p \mathscr{C}^-$, respectively. This also raises the question of if and how $\mathcal{F}_p$ and $\mathcal{P}_p$ or their local restriction to $U \ni p$ can be distinguished. This problem is solved by a topological $\mathbb{Z}_2$ connection (see also Sec. III below).

Given a strong LC structure, a local (conformal) metric can always be proven to exist on any differentiable manifold M with CAT$\subset \mathscr{C}^1$. Within such CAT, any metric η on $\mathbb{R}^{d+1}$ can be restricted to $\mathscr{C}^m$ and pulled back pointwise along $(\phi_p)^{-1}$ to a metric g on $\mathscr{C}_p^{m(p)}$. The CAT continuity of the family $\{p \to \mathscr{C}_p^{m(p)}\}$ implies CAT$_{-1}$ continuity of the family $\{p \to g|_{\mathscr{C}_p^{m(p)}}\}$. So we can extract a CAT$_{-1}$ continuous metric $\{p \mapsto g_p\}$.

Here we are interested in particular only in Lorentzian metrics which are *locally compatible* with a (weak or strong) LC structure in the sense that $\eta_0(v,v) = 0 \Leftrightarrow v \in T_x N, \ \forall (x,v) \in TN$, with arbitrary submanifold $N \subset \mathscr{C} \subset \mathbb{R}^{d+1}$, and correspondingly with $\mathscr{C}_p \supset \phi_p(N) \cong N$, it holds that

$$g_p(V(p), V(p)) = 0 \Leftrightarrow V(p) \in T_p \phi_p(N), \forall p \in \text{int}\, M, \tag{2.5}$$

i.e., locally at any vertex the cone determines the characteristic null directions in the tangent space.

On the other hand, the cone structure poses an equivalence relation on Lorentzian metrics which are compatible with the LC structure. Given any such metric g, the corresponding equivalence class $[g]$ is the conformal class of g. We summarize the existence and uniqueness result as follows.

Proposition 2: Given a strong LC structure on a (CAT) manifold, (i) there always exist a (CAT_{-1}) Lorentzian metric g on M compatible with the LC structure.

(ii) the conformal class $[g]$ of LC compatible metrics is uniquely determined by the LC structure.

The existence of a conformal Lorentzian metric is guaranteed by a *strong* LC structure, but not by a weak one. However, since conditions 3.2 needs only the existence of the tangent bundle of $\mathscr{C}_p$, uniqueness is assured already by a differentiable *weak* LC structure.

Although at each $p \in \operatorname{int} M$ a CAT strong LC structure on M admits a conformal class $[g]$ of CAT_{-1} Lorentzian metrics g with characteristic directions in T_pM tangential to $\mathscr{C}_p$, away from the vertex p the cones of the LC structure need not at all be compatible with the null structure of any conformal metric $[g]$. This reflect the fact that, apart from its local vertex structure, a strong LC structure is still much more flexible than a conformal structure. For any $q \neq p$ the tangent directions given by $T_q\mathscr{C}_p$ need a priori not be related to tangent directions of null curves of g, since the cone (or its thickening) at p is in general unrelated to that at q. The need for compatibility conditions between cones at different points motivates the introduction of some of the causality structures in open regions of M introduced later in the following section.

III. CAUSALITY STRUCTURES ON MANIFOLDS

Given a (weak or strong) LC structure one wants to know first of all under which conditions, for given $p \in \operatorname{int} M$ an exterior and interior of the cone can be distinguished within the complement $M - \mathscr{C}_p$. This problem is the global analog of the local one which was answered by Proposition 1 and Theorem 1 above.

Proposition 3: Assume that at $p \in \operatorname{int} M$ there are open (CAT) submanifolds $\mathcal{I}_p$ and $\mathcal{E}_p$ such that the interior of M decomposes into the disjoint union $\operatorname{int} M = \mathscr{C}_p \mathring{\cup} \mathcal{I}_p \mathring{\cup} \mathcal{E}_p$. Assume $\exists k \in \mathbb{N}_0 : \Pi_k(\mathcal{I}_p) \neq \Pi_k(\mathcal{E}_p)$. Then $\mathcal{I}_p$ and $\mathcal{E}_p$ can be topologically distinguished.

Proof: $\operatorname{int} M - \mathscr{C}_p = \mathcal{I}_p \mathring{\cup} \mathcal{E}_p$, and homotopy groups are topological invariants. □

In particular the conditions of Proposition 3 apply for $d+1>2$ in particular to all manifolds with the following topological structure.

Example 1: Let in any dimension $d := \dim M - 1 > 0$ at any $p \in \operatorname{int} M$ be $|\Pi_0(\mathcal{I}_p)| = 2$ and $\Pi_{d-1}(\mathcal{E}_p) = \Pi_{d-1}(S^{d-1})$, for $d>1$ be $\Pi_{d-1}(\mathcal{I}_p) = 0$ and $|\Pi_0(\mathcal{E}_p)| = 1$, for $d=1$ be $\Pi_{d-1}(\mathcal{I}_p) = \Pi_{d-1}(\mathcal{E}_p) = \Pi_0(S^0)$, i.e., $|\Pi_0(\mathcal{I}_p)| = |\Pi_0(\mathcal{E}_p)| = 2$, and in dimension $d=0$ be $\mathcal{I}_p = \mathcal{E}_p = \emptyset$ at any $p \in \operatorname{int} M$. Then in particular $\mathcal{I}_p \not\cong \mathcal{E}_p \Leftrightarrow d+1>2$. The open exterior $\mathcal{E}_p$ and the open interior $\mathcal{I}_p$ of the cone $\mathscr{C}_p$ at any interior point $p \in \operatorname{int} M$ are topologically distinct for $d>1$, indistinguishable for $d=1$, and empty for $d=0$.

In the case of Example 1, $\mathscr{C}_p$ divides $M - \mathscr{C}_p$ in three open submanifolds, a noncontractible exterior $\mathcal{E}_p$, plus two contractible connected components of $\mathcal{I}_p =: \mathcal{F}_p \cup \mathcal{P}_p$, the future $\mathcal{F}_p$ and the past $\mathcal{P}_p$ with $\partial \mathcal{F}_p = \mathscr{C}_p^+ := \phi_p \mathscr{C}^+$ and $\partial \mathcal{P}_p = \mathscr{C}_p^- := \phi_p \mathscr{C}^-$, respectively. This raises also the question of if and how $\mathcal{F}_p$ and $\mathcal{P}_p$ can be distinguished.

Let M be differentiable and τ be any vector field $M \to TM$ such that at any $p \in \operatorname{int} M$ its orientation agrees with that of $\phi_p(a)$. Such a orientation vector field comes naturally along with a (CAT_{-1}) $\mathbb{Z}_2$-connection on M which allows us to compare the orientation $\tau(p)$ at different $p \in \operatorname{int} M$. Given a strong LC structure on M, the $\mathbb{Z}_2$-connection is in fact provided via continuity of $p \mapsto T_p \phi_p(a)$. Then τ is tangent to an integral curve segment through p from $\mathcal{P}_p$ to $\mathcal{F}_p$. In particular, $\mathcal{F}_p$ and $\mathcal{P}_p$ are distinguished from each other by a consistent τ-orientation on M.

If M is not differentiable, in order to distinguish continuously $\mathcal{P}_p$ from $\mathcal{F}_p$ on $\operatorname{int} M$ it remains just to impose a topological $\mathbb{Z}_2$-connection on $\operatorname{int} M$ a fortiori.

In order to obtain a more specific causal structure it remains to identify natural consistency conditions for the intersections of cones of different points. In order to define topologically timelike, lightlike, and spacelike relations, and a reasonable causal complement, we introduce the following causal consistency conditions on cones.

Definition 8: M is (locally) cone causal or C-causal in an open region U, if it carries a (weak or strong) LC structure and in U the following relations between different cones in $\operatorname{int} M$ hold.

(1) For $p \neq q$ one and only one of the following is true:

(i) q and p are causally timelike related, $q \ll p :\Leftrightarrow q \in \mathcal{F}_p \wedge p \in \mathcal{P}_q$ (or $p \ll q$);

(ii) q and p are causally lightlike related, $q \lhd p :\Leftrightarrow q \in \mathscr{C}_p^+ - \{p\} \wedge p \in \mathscr{C}_q^- - \{q\}$ (or $p \lhd q$);

(iii) q and p are causally unrelated, i.e., relatively spacelike to each other, $q \bowtie p :\Leftrightarrow q \in \mathcal{E}_p \wedge p \in \mathcal{E}_q$.

(2) Other cases (in particular nonsymmetric ones) do not occur. M is *locally C-causal*, if it is C-causal in any region $U \subset M$. M is *C-causal* if conditions (1) and (2) hold $\forall p \in \mathscr{C}$.

Let M be C-causal in U. Then, $q \ll p \Leftrightarrow \exists r : q \in \mathcal{P}_r \wedge p \in \mathcal{F}_r$, and $q \lhd p \Leftrightarrow \exists r : q \in \mathscr{C}_r^+ \wedge p \in \mathscr{C}_r^-$.

If an open curve $\mathbb{R} \ni s \mapsto c(s)$ or a closed curve $S^1 \ni s \mapsto c(s)$ is embedded in M, then in particular its image is $\mathrm{im}(c) \equiv c(\mathbb{R}) \cong \mathbb{R}$ or $\mathrm{im}(c) \equiv c(S^1) \cong S^1$, respectively, whence it is free of self-intersections. Such a curve is called *spacelike*: $\Leftrightarrow \forall p \equiv c(s) \in \mathrm{im}(c) \exists \epsilon : c|_{]s-\epsilon, s+\epsilon[\setminus \{s\}} \in \mathcal{E}_{c(s)}$, and *timelike*: $\Leftrightarrow \forall p \equiv c(s) \in \mathrm{im}(c) \exists \epsilon : c|_{]s-\epsilon, s+\epsilon[\setminus \{s\}} \in \mathcal{I}_{c(s)}$.

Note that C-causality of M forbids a multiple refolding intersection topology for any two cones. In particular along any timelike curve the future/past cones do not intersect, because otherwise there would exist points which are simultaneously timelike and lightlike related. Continuity then implies that future/past cones in fact foliate the part of M which they cover. Hence, if there exists a fibration $\mathbb{R} \hookrightarrow \mathrm{int}\, M \twoheadrightarrow \Sigma$, then C-causality implies that the future/past cones foliate in particular on any fiber. In fact, given a fibration, one could define also a weaker form of causality just by the foliating property of all future/past cones on each fiber. (Physically this situation corresponds to ultralocal classical clocks. Quantum uncertainty of the fiber would require us to take appropriate ensemble averages over some bundle of neighboring fibers which then contains in particular spacelike related vertices on the fibers of the bundle. Then the corresponding future or past cones intersect for sure, and even timelike related ones of different fibers *may* intersect!) C-causality, however, requires more, namely the future/past cones of *all* timelike related vertices should be nonintersecting, not only those in a particular fiber.

Therefore C-causality allows also a reasonable definition of a *causal complement*.

Definition 9: For any open set S in a C-causal manifold M the causal complement is defined as

$$S^{\perp} := \bigcap_{p \in \mathrm{cl} S} \mathcal{E}_p, \tag{3.1}$$

where $\mathrm{cl}S$ denotes the closure in the topology induced from the manifold. Although the causal complement is always open, it will in general not be a contractable region even if S itself is so.

Assume p and q are timelike related, $p \in \mathcal{P}_q$ and $q \in \mathcal{F}_p$. $\mathcal{K}_p^q := \mathcal{F}_p \cap \mathcal{P}_q$ is the bounded open double cone between p and q. Given any bounded open $\mathcal{K}$ such that $\exists p, q \in M : \mathcal{K} = \mathcal{F}_p \cap \mathcal{P}_q$, we set $i^+(\mathcal{K}) := \{q\}$, $i^-(\mathcal{K}) := \{p\}$, and $i^0(\mathcal{K}) := \mathscr{C}_p^+ \cap \mathscr{C}_q^-$. For any $\mathcal{K}_p^q \subset M$ let $\mathrm{cl_c}(\mathcal{K}_p^q)$ be its causal closure.

Since C-causality prohibits relative refolding of cones, it also ensures that $(\mathcal{K}_p^q)^{\perp\perp} = \mathcal{K}_p^q$, i.e. the causal complement is a duality operation on double cones.

The open double cones of a C-causal manifold M generate a topology, called the *double cone topology* which is a genuine generalization of the usual Alexandrov topology for strongly causal space–times. For strongly causal space–times the Alexandrov topology is equivalent to the manifold topology.[5,6] When M fails to be locally causal the double cone topology may be coarser than the manifold topology.

Let us discuss now possible natural relations that can appear between two double cones $\mathcal{K}_1$ and $\mathcal{K}_2$ of a C-causal manifold. First there is the case $\mathcal{K}_1 \cap \mathcal{K}_2 = \emptyset$ which corresponds to causally unrelated sets. For $\mathcal{K}_1 \cap \mathcal{K}_2 \neq \emptyset$, the intersection is such that $\mathcal{K}_1 \cup \mathcal{K}_2 - \mathcal{K}_1 \cap \mathcal{K}_2$ is either given by two disconnected pieces or it is connected. In the latter case we distinguish whether $\partial\mathcal{K}_1 \cap \partial\mathcal{K}_2$ is empty or not. It is in the former case that one of $\mathcal{K}_1$ and $\mathcal{K}_2$ will be contained in the other.

Local C-causality does a priori not preclude other more pathological possibilities. However, it is possible to define in a purely topological manner more refined causality notions.

Definition 10: Let M be a C-causal manifold.

(i) M is globally hyperbolic: $\Leftrightarrow \mathrm{cl_c}\mathcal{K}_p^q$ compact $\forall p, q \in M$

(ii) M is *causally simple*: $\Leftrightarrow \mathrm{cl}_c \mathcal{K}_p^q$ closed $\forall p,q \in M$

(iii) M is *causally continuous*: $\Leftrightarrow$ M is distinguishing and both $\mathcal{F}: p \mapsto \mathcal{F}_p$ and $\mathcal{P}: q \mapsto \mathcal{P}_q$ are continuous

(iv) M is *stably causal*: $\Leftrightarrow$ M admits $\mathcal{C}^0$ function $f: M \to \mathbb{R}$ strictly monotonously increasing along each future directed nonspacelike curve (global time function)

(v) M is *strongly causal*: $\Leftrightarrow$ the topology generated by $\{\mathcal{K}_p^q\}_{p,q \in M}$ is equivalent to the manifold topology of M

(vi) M is *distinguishing*: $\Leftrightarrow \mathcal{F}_p = \mathcal{F}_q \Rightarrow p = q \wedge \mathcal{P}_p = \mathcal{P}_q \Rightarrow p = q$

(vii) M is *causal*: $\Leftrightarrow$ every closed curve in M is not nonspacelike

(viii) M is *chronological*: $\Leftrightarrow$ every closed curve in M is not timelike

If a manifold carries a Lorentzian metric we saw in Sec. II above that this is locally compatible with a strong LC structure. Beyond that it is an interesting question under which conditions a Lorentzian metric is *compatible* with some LC structure in the sense that $\eta_x(v,v) = 0 \Leftrightarrow [x \in \mathscr{C} \Rightarrow v \in T\mathscr{C} := \cup_{y \in N \subset \mathcal{C}} T_y N]$ where $N \subset \mathcal{C}$ is a submanifold of $\mathscr{C}$ with codimension 0 when x is not the vertex of $\mathscr{C}$ and codimension 1 when it is the vertex, and, correspondingly,

$$g_q(V(q),V(q)) = 0 \Leftrightarrow \left[\forall p \in M : q \in \mathscr{C}_p \Rightarrow V(q) \in T\mathscr{C}_p := \bigcap_{r \in \mathcal{C}} T_r \phi_p(N) \right], \tag{3.2}$$

i.e., the cones are the characteristic surfaces of the Lorentzian metric. As pointed out above, this does not hold in general. However, one might search for sufficient and necessary causality conditions such that this compatibility holds. A systematic investigation of this point is beyond our present investigations. Let us rather assure the correspondence of the causality notions of Def. 10 to the usual ones in the case of a Lorentzian space–time.

Theorem 2: Let M carry a smooth Lorentzian metric g. Then the Lorentzian metric determines a C-causal structure. If a C-causal structure of M is related to a Lorentz metric, then the definitions (i)–(viii) agree with the standard definitions and the following chain of implications of properties of M holds: globally hyperbolic $\Rightarrow$ causally simple $\Rightarrow$ causally continuous $\Rightarrow$ stably causal $\Rightarrow$ strongly causal $\Rightarrow$ distinguishing $\Rightarrow$ causal $\Rightarrow$ chronological.

Proof: Given a smooth Lorentzian metric g the cones determined by the null structure $[g]$ respect the relations of Def. 8, because otherwise there would exist some singularities. For (v) in the case of Lorentzian manifolds see Ref. 6; for the other notions and for the chain of implications see Ref. 7. □

Finally let us define a condition which excludes the existence of singularities or internal boundaries within the future and past cones.

Definition 11: Let M carry a (weak or strong) LC structure.

(i) M is precausal in an open region $U \subset M$, if the $d+1$-parameter CAT family $\{\phi_p\}_{p \in U}$ of CAT homeomorphisms $\phi_p : \mathbb{R}^{d+1} \supset V \to U$ is such that at any $p \in U$ it is $\mathscr{C}_p|_U = \phi_p \mathscr{C}|_V$, and any CAT submanifold of $\mathscr{C}_p$ or $(M - \mathscr{C}_p) \cap U$ is a CAT homeomorphic image of $\mathscr{C}$ or $(\mathbb{R}^{d+1} - \mathscr{C}) \cap V$, respectively. M is *locally precausal* iff it is precausal in any open region $U \subset M$.

(ii) M is *precausal* if it is locally precausal such that in the CAT $d+1$-parameter family $\{\phi_p\}_{p \in U}$ any CAT homeomorphism extends also to a homeomorphism of the interior $\phi_p : \mathcal{E} \to \mathcal{E}_p$.

IV. DISCUSSION AND PERSPECTIVE

Above we presented topological definitions of local (i.e., pointwise) cone (LC) structures for a general topological or differentiable manifold M of dimension $d+1 > 2$ and notions of causality on M in a purely topological manner. It is remarkable that such definitions are possible, whence the usual recursion to a Lorentzian metric becomes redundant.

Proposition 1 gives criteria which locally distinguish the exterior and the interior of the cone at any point from each other. Proposition 3 and Example 1 provide concrete global topological conditions for M in order to allow the relative distinction of interior and exterior of all its cones.

Minkowski space is obviously a manifold which satisfies the conditions for topologically distinguished interior and exterior according to Example 1. It is, however, a priori not clear what for each given category CAT of manifolds is the largest class of manifolds with the topological structure described in Example 1.

We saw that a global consistent distinction between future and past cones requires just a topological $\mathbb{Z}_2$-connection. Note that, as an important possible application, the canonical approach to quantum gravity comes always along with such a connection. In fact the canonical configuration variables for oriented manifolds maybe there be chosen as $\mathrm{SO}(d+1)$-connections.

The presented LC structures, C-causality, and other our purely topological causality notions provide some alternative to the poset approach [8–11] for defining causality in quantum theories of quantum gravity. While that approach is based on a much weaker local notion of causality on sets, which essentially involves only a partial ordering, the present definition gives us the possibility of working with the local definition of causality on differentiable manifolds which still captures the essential notions for curves in a C-causal manifold to be lightlike, timelike or spacelike without the need of an underlying Lorentzian structure. For any set S in a C-causal manifold a topological notion of a causal complement $S^{\perp}$ is given by (3.1). Any double cone $\mathcal{K}$ in a C-causal manifold then has the duality property $\mathcal{K}^{\perp\perp}=\mathcal{K}$.

Some advantages of conformal invariance in the quantization in minisuperspace models of higher-dimensional Einstein gravity have been pointed out in Refs. 12, 13. In particular, factor ordering problems can be resolved uniquely this way. For a more general background independent quantum theory the restriction of local diffeomorphisms to those consistent with a causal structure, say, e.g., a LC structure, on the whole manifold might appear too restrictive. After all a strong LC structure implies already the existence of a conformal metric, whence diffeomorphisms may be restricted locally to conformal ones. Nevertheless note that even a strong LC structure is much more flexible than a conformal metric structure. The local cones of different vertices might refold away from their vertices with rather complicated intersection topologies while a CAT continuous conformal metric within its (regular!) domain does not admit refolding singularities of the characteristic surfaces, each of the which is spanned out by all the null geodesics passing through a given vertex. Of course refolding and the associated singularities should be a topic of further more systematic investigations elsewhere.

The canonical approach to field quantization usually employs a foliation $\Sigma\hookrightarrow \mathrm{int}\, M \twoheadrightarrow \mathbb{R}$. This raises the question of when this is consistent with a (C-)causal structure. This may roughly be answered as follows: A CAT foliation of M may be said to be consistent with a C-causal structure, if for any open set O in a CAT slice $\Sigma\subset M$ there exists a double cone $\mathcal{K}\subset \mathrm{int}\, M$ such that $i^0(\mathcal{K})\subset\partial(\mathcal{K}\cap\Sigma)$ (compare also Sec. III, below Def. 9).

Consider now such a double cone $\mathcal{K}$ in M with $O:=\mathcal{K}\cap\Sigma$ and $\partial O=i^0(\mathcal{K})$ and a diffeomorphism ϕ in M with pullbacks $\phi^{\Sigma}\in\mathrm{Diff}(\Sigma)$ to Σ and $\phi^{\mathcal{K}}\in\mathrm{Diff}(\Sigma)$ to $\mathcal{K}$. If $\phi(\mathcal{K})=\mathcal{K}$, it can be naturally identified with an element of $\mathrm{Diff}(\mathcal{K})$. ($\phi|_{M-K}=\mathrm{id}_{M-\mathcal{K}}$ is a sufficient but not necessary condition for that to be true.) If in addition $\phi(\Sigma)=\Sigma$ then also $\phi(O)=O$, and $\phi|_O$ is a diffeomorphism of O.

When M is homeomorphic to $\Sigma\times\mathbb{R}$ it is straightforward to extend the above from a single hypersurface $\Sigma\subset M$ to a foliation of M via a 1-parameter set of embeddings $\Sigma\hookrightarrow M$.

For a canonical approach to quantum gravity, one might want to work with a restriction of the causal structure to cones with their vertices on a given graph Γ within a slice Σ of a foliation. A given topological (differentiable) causal structure, selects particular causal homeomorphisms (diffeomorphisms) which preserve it. A strong LC structure on all of M already implies the existence of a conformal metric structure and a requirement of compatibility with that metric would reduce the local covariance group to local conformal diffeomorphisms. One might, however, also weaken the LC and causal structure of the manifold by considering in any leaf Σ of a given foliation only cones with vertices on $\Gamma\subset\Sigma$ instead on all of Σ. A natural choice for Γ is the dual graph of a triangulation. Then the cones have to CAT vary along the edges, but at least for $\mathrm{CAT}\supset C^{\infty}$ the cones at the vertices of the graph can be freely ascribed. Consequently, a geometry constructed on that basis will be invariant under diffeomorphisms much more general than conformal ones.

Let us, however, also emphasize that, although the existence of a local conformal metric is guaranteed by a strong LC structure, it is a priori not obvious that this metric should play any significant rôle. Then, however, also the need to restrict diffeomorphisms to those compatible with the conformal metric may be questioned. One might eventually expect that within some approach to quantum geometry a cone at a vertex $p \in O \subset \Sigma$ should be replaced by an appropriate average over cones with vertices within some region O of minimal Heisenberg uncertainty. Then the flexibility of the weak and strong LC structures makes them interesting concepts and potential ingredients for a possible definition of quantum causality too. Presently, however, this is still a matter of many speculations.

Classically, the existence of a local metric requires only the differentiable structure in an arbitrary small neighborhood of the vertex, and the defined LC structures fix the preferred null directions only locally at each vertex. With sufficiently strong notions of causality (e.g., C-causality above) the null structures of this metric may become consistent with the global structure of cones of the LC structure. Note that in the case of a given Lorentz metric null geodesics lie on cones, and with sufficiently strong causality, e.g. global hyperbolicity, these cones have to be consistent with respect to each other and under variation of the vertex without refolding into each other (i.e., in particular without conjugate points).

For Lorentzian manifolds there is a hierarchy of common notions of causality which have been generalized above. Provided our definitions of causality are sufficiently natural it should be possible to prove (at least parts) of this hierarchy in the more general topological setting. However, a complete investigation of the mutual relations between different topological causality concepts is beyond the scope and goal of the present paper.

It should be emphasized that the above was just brief demonstration of the possibility to introduce notions of cones and causality on CAT topological manifolds without a metric. In particular, weak and strong LC structures, C-causality, precausality, and some generalizations of the most common notions of causality have been obtained. However, the investigation is far from complete. It remains for future work to develop the topological approach to causal structures on manifolds further, to investigate better some of its implications, and last not least to demonstrate its applicability in background independent formulations of algebraic quantum field theory and quantum gravity.

ACKNOWLEDGMENTS

I wish to thank to O. Dreyer, M. Holman, R. Puzio, and M. Reisenberger for discussions, and to N. Guerras for continuous support.

[1] S. W. Hawking, A. R. King, and P. J. McCarthy, ''A new topology for curved space–time which incorporates the causal, differential, and conformal structures,'' J. Math. Phys. **17**, 174–181 (1976).
[2] D. B. Malament, ''The class of continuous timelike curves determines the topology of spacetime,'' J. Math. Phys. **18**, 1399–1404 (1977).
[3] S. Frittelli, C. Kozameh, and T. Newman, ''Lorentzian metrics from characteristic surfaces,'' J. Math. Phys. **36**, 4975–4983 (1995).
[4] P. Ehrlich, ''Null cones and pseudo-Riemannian metrics,'' Semigroup Forum **43**, 337–343 (1991).
[5] E. H. Kronheimer and R. Penrose, ''On the structure of causal spaces,'' Proc. Cambridge Philos. Soc. **63**, 481–501 (1967).
[6] R. Penrose, ''Techniques of differential topology in relativity,'' *Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics* (SIAM, Philadelphia, 1972), No. 7.
[7] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian Geometry* (Marcel Dekker, New York, 1996).
[8] L. Bombelli, J. Lee, D. A. Meyer, and R. D. Sorkin, ''Space–time as a causal set,'' Phys. Rev. Lett. **59**, 481–501 (1987).
[9] F. Markopoulou and L. Smolin, ''Quantum geometry with intrinsic local causality,'' Phys. Rev. D **58**, 084032 (1998).
[10] F. Markopoulou, ''Quantum causal histories,'' http://xxx.lanl.gov/abs/hepth/9904009.
[11] D. P. Rideout and R. D. Sorkin, ''A classical sequential growth dynamics for causal sets,'' http://xxx.lanl.gov/abs/gr-qc/9904062.
[12] M. Rainer, ''Conformal coupling and invariance in arbitrary dimensions,'' Int. J. Mod. Phys. D **4**, 397–416 (1995).
[13] M. Rainer, ''Conformal minisuperspace quantization and equivalence of multidimensional classical and quantum models,'' Grav. Cosmol. **1**, 121–130 (1995).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

On the limiting power of the set of knots generated by 1+1- and 2+1-braids

R. Bikbov
L. D. Landau Institute for Theoretical Physics, 117940, Moscow, Russia

S. Nechaev
L. D. Landau Institute for Theoretical Physics, 117940, Moscow, Russia and LPTMS, UMR 8626, CNRS-Université Paris Sud, 91405 Orsay Cedex, France

(Received 31 August 1998; accepted for publication 18 March 1999)

We estimate from above the set of knots, $\Omega(n,\mu)$, generated by closure of the n-string 1+1- and 2+1-dimensional braids of irreducible length $\mu(\mu \gg 1)$ in the limit $n \gg 1$. © *1999 American Institute of Physics.* [S0022-2488(99)04307-8]

I. INTRODUCTION

Besides the traditional fundamental topological issues concerning the construction of new topological invariants, investigation of homotopic classes and fibre bundles we mark a set of ajoint but much less studied problems. First of all, we mean the problem of so-called ''knot entropy'' calculation. Most generally it can be formulated as follows. Take the lattice $\mathbb{Z}^3$ embedded in the space $\mathbb{R}^3$. Let Ω be the ensemble of all possible closed non-self-intersecting N-step paths with one common fixed point on $\mathbb{Z}^3$; by ω_N we denote the particular configuration of the trajectory. The main question is: what is the fraction P_N of the trajectories $\omega_N \in \Omega$ belonging to some specific homotopic class characterized by the topological invariant Inv (we do not specify the way of defining the topological invariant). The distribution function $P\{\text{Inv}\}$ satisfies the obvious normalization condition $\Sigma_{\text{all } \omega_N \in \Omega} P_N\{\text{Inv}\} = 1$.

In the present paper we pay attention to the statistical problem concerning the estimation of the set $\Omega = \{\Omega^{(1)}, \Omega^{(2)}\}$ of knots generated by closure of braids embedded in 1+1- and 2+1-dimensions (see the definitions below).

The paper is organized as follows. Below we give the basic definitions of the standard 1+1-dimensional and 2+1-dimensional braid groups as well as formulate the basic results; Sec. II is devoted to the estimations of the sets $\Omega^{(1)}$ and $\Omega^{(2)}$ using the concept of 1+1- and 2+1-locally-free groups; while in the Conclusion we discuss in more details the corollaries following our consideration.

A. The basic definitions

(1) The 1+1-dimensional (''standard'') braid group $B^{(1)}_{n+1}$ of $n+1$ strings has n generators $\{\sigma_1, \sigma_2, \ldots, \sigma_n$ and their inverses$\}$ [see Fig. 1(a)] with the following relations:

$$\begin{aligned} \sigma_i\sigma_{i+1}\sigma_i &= \sigma_{i+1}\sigma_i\sigma_{i+1} \quad (1 \le i < n) \\ \sigma_i\sigma_j &= \sigma_j\sigma_i \quad (|i-j| \ge 2) \\ \sigma_i\sigma_i^{-1} &= \sigma_i^{-1}\sigma_i = e. \end{aligned} \tag{1}$$

(2) The 2+1-dimensional (''surface'') braid group $B^{(2)}_{n+1}$ can be defined in the following way (see, for instance Refs. 1, 2). Consider the two-dimensional lattice $\mathbb{Z}^2$ and take distinct points $P_1, P_2, \ldots, P_{n+1} \in \mathbb{Z}^2$. A 2+1-braid of $n+1$ strings on $\mathbb{Z}^2$ based at $\{P_1, P_2, \ldots, P_{n+1}\}$ is an $n+1$-tuple $b = (b_1, \ldots, b_{n+1})$ of paths, $b_i : [1,N] \to \mathbb{Z}^2$, such that

(i) $b_i(1) = P_i$ and $b_1(1) \in \{P_1, P_2, \ldots, P_{n+1}\} \forall i \in \{1, ... n+1\}$;

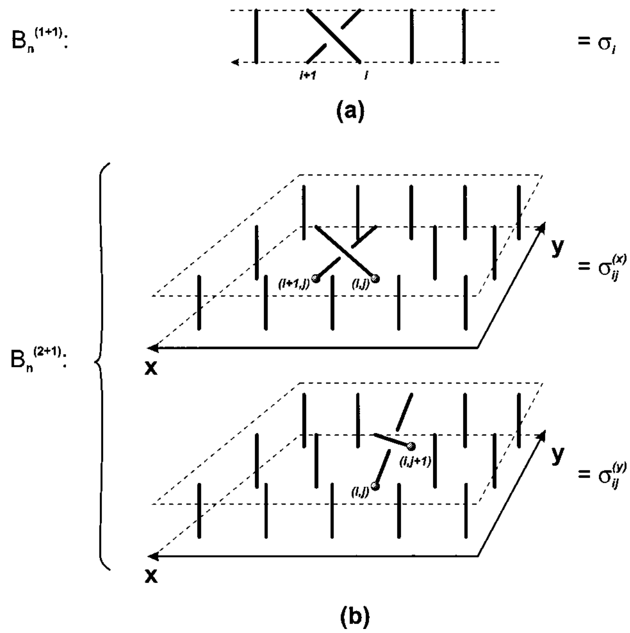

FIG. 1. Representation of (a) $1+1$-braid group generator, σ_i; (b) $2+1$-braid group generators $\sigma^x_{i,j}$ and $\sigma^y_{i,j}$.

(ii) $b_i(t) \neq b_j(t) \forall \{i,j\} \in \{1,...,n+1\},\ i \neq j;\ t \in [1,N]$.

The braid group $B^{(2)}_{n+1}$ on $\mathbb{Z}^2$ based at $\{P_1, P_2, \ldots, P_{n+1}\}$ is the group of homotopy classes of braids based at $\{P_1, P_2, \ldots, P_{n+1}\}$. The group $B^{(2)}_{n+1}$ has $2(n \times n)$ generators $\{(\sigma^{(x)}_{11}, \sigma^{(y)}_{11}), \ldots, (\sigma^{(x)}_{1n}, \sigma^{(y)}_{1n}); \ldots; (\sigma^{(x)}_{n1}, \sigma^{(y)}_{n1}), \ldots, (\sigma^{(x)}_{nn}, \sigma^{(y)}_{nn})$ and their inverses$\}$ [see Fig. 1(b)] with the following relations:

$$
\begin{gathered}
\sigma^{(x)}_{i,j}\sigma^{(x)}_{i+1,j}\sigma^{(x)}_{i,j} = \sigma^{(x)}_{i+1,j}\sigma^{(x)}_{i,j}\sigma^{(x)}_{i+1,j} \quad (1 \leqslant \{i,j\} \leqslant n) \\
\sigma^{(y)}_{i,j}\sigma^{(y)}_{i,j+1}\sigma^{(y)}_{i,j} = \sigma^{(y)}_{i,j+1}\sigma^{(y)}_{i,j}\sigma^{(y)}_{i,j+1} \quad (1 \leqslant \{i,j\} \leqslant n) \\
\sigma^{(x)}_{i,j}\sigma^{(y)}_{i,j}\sigma^{(x)}_{i,j} = \sigma^{(y)}_{i,j}\sigma^{(x)}_{i,j}\sigma^{(y)}_{i,j} \quad (1 \leqslant \{i,j\} \leqslant n) \\
\sigma^{(x)}_{i_1,j_1}\sigma^{(x)}_{i_2,j_2} = gj^{(x)}_{i_2,j_2}\sigma^{(x)}_{i_1,j_1} \quad (|i_1 - i_2| > 1 \ \text{ or } \ |j_1 - j_2| > 0) \\
\sigma^{(x)}_{i_1,j_1}\sigma^{(y)}_{i_2,j_2} = \sigma^{(y)}_{i_2,j_2}\sigma^{(x)}_{i_1,j_1} \quad (i_2 - i_1 \neq \{0,1\} \ \text{ or } \ j_1 - j_2 \neq \{0,1\}) \\
\sigma^{(x)}_{i,j}(\sigma^{(x)}_{i,j}))^{-1} = \sigma^{(y)}_{i,j}(\sigma^{(y)}_{i,j}))^{-1} = e.
\end{gathered} \tag{2}
$$

The braid groups $B^{(1)}_n$ and $B^{(2)}_n$ have the following general properties:

(a) Any arbitrary word written in terms of ''letters''—generators of the groups $B^{(1)}_n$ or $B^{(2)}_n$—gives a particular braid.

(b) The *length*, *N*, of the braid is the total number of used letters, while the *minimal irreducible length*, μ, hereafter referred to as the ''primitive length'' is the shortest noncontractible length of a particular braid which remains after applying of all possible group relations. Diagrammatically the braid can be represented as a set of crossed strings going from the top to the bottom appeared after subsequent gluing the braid generators.

(c) The closed braid appears after gluing the ''upper'' and the ''lower'' free ends of the braid on the cylinder.

B. The main results

Our basic results might be formulated in a geometrically clear way. Consider two sets of braids $\{B_n^{(1)}\}$ and $\{B_n^{(2)}\}$, embedded in $1+1$- and $2+1$-dimensions correspondingly. Let each particular braid has the primitive length μ and is represented by n strings.

Then

(1) The set $\Omega^{(1)}(n,\mu)$ of knots which can be generated by the standard braids of given irreducible length $\mu(\mu\geqslant 1)$ from the set $\{B_n^{(1)}\}$ ($n=\text{const}\geqslant 1$) is restricted from above by the value

$$\Omega^{(1)}(n,\mu)<\frac{32\pi^2}{\log^4 2}\frac{2^n}{n^3}7^{\mu-1}. \tag{3}$$

(2) The set $\Omega^{(2)}(n,\mu)$ of knots which can be generated by the surface braids of given irreducible length $\mu(\mu\geqslant 1)$ from the set $\{B_n^{(2)}\}$ ($n=\text{const}\geqslant 1$) is restricted from above by the value

$$\Omega^{(2)}(n,\mu)<\frac{32n^2}{\pi^2}\left(\frac{2n}{\log n}\right)^{\mu-1}. \tag{4}$$

(See the Conclusion for more detailed discussion of the results (3) and (4).)

II. COMBINATORICS OF WORDS

Any braid corresponds to some knot or link. The correspondence between braids and knots is not mutually single valued and each knot or link can be represented by infinite series of different braids. However, we can estimate from above the partition functions $\Omega^{(1)}(n,\mu)$ and $\Omega^{(2)}(n,\mu)$ of all possible knots generated by the ensemble of all $1+1$- and $2+1$-braids of primitive length μ using the following obvious fact. *The sets $\Omega^{(1)}(n,\mu)$ and $\Omega^{(2)}(n,\mu)$ are bounded from above by the number of all distinct words of the primitive length μ in $1+1$- and $2+1$-braid groups correspondingly.* Thus in what follows we are aimed in the estimation of the number of non-equivalent words in the standard and surface braid groups.

A. Definitions of 1+1- (''standard'') and 2+1- (''surface'') locally free groups

(1) Following the ideas of Vershik concerning the notion of the ''local groups''[3] and the papers,[4] where the concept of a ''locally free'' group was proposed at first in the topological context, let us define the group, $\mathcal{LF}_{n+1}^{(1)}$, which has n generators $\{f_1,\dots,f_n$ and their inverses$\}$ with the relations,

$$\begin{cases} f_j f_k = f_k f_j \ \text{ for } \ |j-k|\geqslant 2 \\ f_i f_i^{-1} = e. \end{cases} \tag{5}$$

We call the group with relations (5) the $1+1$-dimensional ''locally free group,'' because each pair of generators $(f_j, f_{j\pm 1})$ produces a free subgroup of the group $\mathcal{LF}_{n+1}^{(1)}$.

The group $\mathcal{LF}_{n+1}^{(1)}$ can be obtained from the braid group $B_{n+1}^{(1)}$ if we replace the braiding (''Yang–Baxter-type'') relations by the free ones. The geometrical interpretation of the generators of a group $\mathcal{LF}_{n+1}^{(1)}$ is shown in Fig. 2(a).

Apparently, in mathematical literature the notion similar to our ''locally free group'' appeared first in the paper[5] devoted to the investigation of the combinatorial properties of rearrangements of

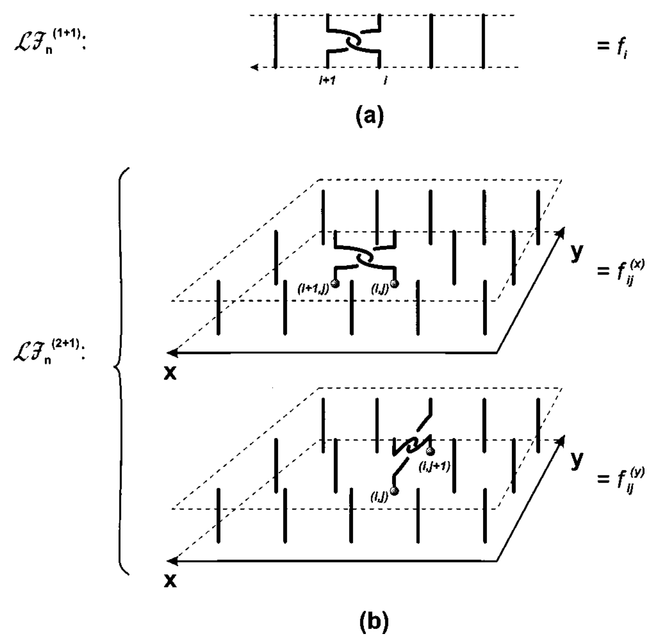

FIG. 2. Representation of (a) $1+1$-locally free group generator, f_i; (b) $2+1$-locally free group generators $f^x_{i,j}$ and $f^y_{i,j}$.

sequences, known also as ''partially commutative monoids'' (see Ref. 6, and references therein).

(2) The $2+1$-dimensional (''surface'') locally free group $\mathcal{LF}^{(2)}_{n+1}$ has $2(n\times n)$ generators $\{(f^{(x)}_{11},f^{(y)}_{11}),\dots,(f^{(x)}_{1n},f^{(y)}_{1n});\dots;(f^{(x)}_{n1},f^{(y)}_{n1}),\dots,(f^{(x)}_{nn},f^{(y)}_{nn})$ and their inverses$\}$ with the following relations:

$$f^{(x)}_{i_1,j_1}f^{(x)}_{i_2,j_2}=f^{(x)}_{i_2,j_2}f^{(x)}_{i_1,j_1}\quad(|j_1-j_2|>0\ \text{ or }\ |i_1-i_2|>1)$$

$$f^{(x)}_{i_1,j_1}f^{(y)}_{i_2,j_2}=f^{(y)}_{i_2,j_2}f^{(x)}_{i_1,j_1}\quad(i_2-i_1\neq\{0,1\}\ \text{ or }\ j_1-j_2\neq\{0,1\})\tag{6}$$

$$f^{(x)}_{i,j}(f^{(x)}_{i,j})^{-1}=f^{(y)}_{i,j}(f^{(y)}_{i,j})^{-1}=e.$$

Thus, we can construct the $2+1$-locally free group $\mathcal{LF}^{(2)}_{n+1}$ from the surface braid group $B^{(2)}_{n+1}$ if we replace the braiding relations of the neighboring generators by the ''full monodromy,'' i.e., by the free group relations [see Fig. 2(b)].

The following important properties of $1+1$- and $2+1$-locally free groups should be mentioned.

(i) By definition the locally free groups $\mathcal{LF}^{(1)}_{n+1}$ and $\mathcal{LF}^{(2)}_{n+1}$ have less relations than the braid groups $B^{(1)}_{n+1}$ and $B^{(2)}_{n+1}$ correspondingly. *Thus, the number of distinct words of the primitive length* μ *in the* $1+1$- *and* $2+1$-*braid groups is bounded from above by the number of distinct words of the primitive length* μ *in the* $1+1$- *and* $2+1$-*locally free groups.*

(ii) By construction (compare Figs. 1 and 2) the monodromy generators f_i ($i\in[1,n]$) of the group $\mathcal{LF}^{(1)}_{n+1}$ and $f^{(x,y)}_{i,j}$ ($\{i,j\}\in[1,n]$) of the group $\mathcal{LF}^{(2)}_{n+1}$ can be written as $f_i=(\sigma_i)^2$ ($i\in[1,n]$) and $f^{(x,y)}_{i,j}=(\sigma^{(x,y)}_{i,j})^2$ ($\{i,j\}\in[1,n]$), where σ_i and $\sigma^{(x,y)}_{i,j}$ are the generators of the groups $B^{(1)}_{n+1}$ and $B^{(2)}_{n+1}$ correspondingly. *Thus, the number of distinct words of the primitive length* 2μ *in the* $1+1$- *and* $2+1$-*braid groups is bounded from below by the* *number of distinct words of the primitive length μ in the $1+1$- and $2+1$-locally free groups.*

B. Computation of number of nonequivalent words in 1+1- and 2+1-locally free groups

We derive explicitly the expressions of the numbers $V^{(1)}(n,\mu)$ and $V^{(2)}(n,\mu)$ of all nonequivalent primitive words of length μ in the groups $\mathcal{LF}_{n+1}^{(1)}$ and $\mathcal{LF}_{n+1}^{(2)}$, respectively. Our computations are based on the so-called "normal order" representation of words proposed by Vershik in Ref. 7 (see also Ref. 4).

1. The group $\mathcal{LF}_{n+1}^{(1)}$

Let us represent each word W_p of irreducible length μ in the group $\mathcal{LF}_{n+1}^{(1)}$ in the "standard" form

$$W_p=(f_{\alpha_1})^{m_1}(f_{\alpha_2})^{m_2}\cdots(f_{\alpha_s})^{m_s}, \tag{7}$$

where $\Sigma_{i=1}^{s}|m_i|=\mu$ $(m_i\neq 0\,\forall i;\ 1\leqslant s\leqslant\mu)$ and the sequence of generators f_{α_i} in Eq. (7) *for all distinct* f_{α_i}, satisfies the following local rules[4] ("normal order" representation),

(i) If $f_{\alpha_i}=f_1$, then $f_{\alpha_{i+1}}=f_2$;

(ii) If $f_{\alpha_i}=f_k$ $(2\leqslant k\leqslant n-1)$, then $f_{\alpha_{i+1}}\in\{f_1,\ldots,f_{k-1},f_{k+1},\}$;

(iii) If $f_{\alpha_i}=f_n$, then $f_{\alpha_{i+1}}\in\{f_1,\ldots,f_{n-1}\}$.

The rules (i)–(iii) give the prescription how to encode and enumerate all distinct primitive words in the group $\mathcal{LF}_{n+1}^{(1)}$. If the sequence of generators in the primitive word W_p does not satisfy the rules (i)–(iii), we commute the generators in the word W_p until the normal order is restored. Hence, the normal order representation enables one to give the unique coding of all nonequivalent primitive words in our group.

Let $\theta_n(m)$ be the number of all distinct sequences of $m+1$ generators, $0\leqslant m\leqslant\mu-1$, satisfying the rules (i), (ii), (iii). The calculation of the number of distinct primitive words $V^{(1)}\times(n,\mu)$ of given primitive length μ is now straightforward,

$$V^{(1)}(n,\mu)=\sum_{m=0}^{\mu-1}2^{m+1}\binom{\mu-1}{m}\theta_n(m). \tag{8}$$

The combinatorial factor $2^{m+1}\binom{\mu-1}{m}$ in Eq. (8) is the number of all primitive words of length μ written in a normal order form for the *fixed sequence* of $m+1$ generators.

Our approach to the computation of $\theta_n(m)$ is based on the consideration of a "correlation function" $\theta_n(x,x_0,m)$ which is defined as the number of all distinct sequences of $m+1$ generators satisfying the rules (i), (ii), (iii), beginning with the generator f_{x_0} and ending with the generator f_x. It is easy to write an evolution equation for $\theta(x,m)\equiv\theta_n(x,x_0,m)$ with the "time" m,

$$\theta(x,m+1)=\theta(x-1,m)+\sum_{y=x+1}^{n}\theta(y,m). \tag{9}$$

This equation should be completed by initial and boundary conditions

$$\begin{gathered}\theta(x,0)=\delta_{x,x_0}\\ \theta(0,m)=\theta(n+1,m)=0.\end{gathered} \tag{10}$$

We solve the boundary problems (9)–(10) in the limit $n\gg 1$ supposing the periodical boundary conditions on the segment $[0,n+1]$. Namely, we have

$$\theta(x+1,m+1)-\theta(x,m+1)=\theta(x,m)-\theta(x-1,m)-\theta(x+1,m)$$
$$\theta(x,0)=\delta_{x_0,x} \tag{11}$$
$$\theta(0,m)=\theta(n+1,m)=0.$$

The substitution

$$\theta(n,m)=\sum_{k=1}^{n} A_k \lambda_k^m \alpha_k(x) \tag{12}$$

enables us to pass to the following recursion relations

$$(\lambda_k+1)\alpha_k(x+1)-(\lambda_k+1)\alpha_k(x)+\alpha_k(x-1)=0$$
$$\alpha_k(0)=\alpha_k(n+1)=0. \tag{13}$$

One can readily find the eigenvalues and eigenfunctions of (13),

$$\lambda_k=4\cos^2\frac{\pi k}{n+1}-1,$$
$$\alpha_k(x)=\frac{\sin\frac{\pi k x}{n+1}}{\left(2\cos\frac{\pi k}{n+1}\right)^x}, \quad k=1,...,n. \tag{14}$$

As the function (13) is not symmetric, the set of eigenfunctions is not orthogonal on the segment $[0,n+1]$ and it is difficult to ensure the initial condition. It is convenient to pass from (13) to symmetric problem. Consider a generating function

$$Z(x,s)=\sum_{m=0}^{\infty} s^m\theta(x,m). \tag{15}$$

The equation for the function $Z(x,s)$ reads

$$(s+1)Z(x+1,s)-(s+1)Z(x,s)+sZ(x-1,s)=\delta_{x,x_0-1}-\delta_{x,x_0}$$
$$Z(0,s)=Z(n+1,s)=0. \tag{16}$$

The last equation can be symmetrized via the substitution $Z(x,s)=A^x\varphi(x,s)$, where $A=\sqrt{s/(s+1)}$. Thus, we get

$$\varphi(x+1,s)-\frac{1}{A}\varphi(x,s)+\varphi(x-1,s)=\frac{A^{-x}}{\sqrt{s(s+1)}}(\delta_{x,x_0-1}-\delta_{x,x_0})$$
$$\varphi(0,s)=\varphi(n+1,s)=0. \tag{17}$$

Making use of the sin-Fourier transform, $f(k,s)=\sum_{x=1}^{n}\varphi(x,s)\sin[\pi kx/(n+1)]$, let us rewrite (17) in the form

$$\left(2\cos\frac{\pi k}{n+1}-\frac{1}{A}\right)f(k,s)=\frac{A^{-x_0}}{\sqrt{s(s+1)}}\left(A\sin\frac{\pi k(x_0-1)}{n+1}-\sin\frac{\pi k x_0}{n+1}\right).$$

The final explicit expression of the function $Z(x,s)$ reads as follows:

$$Z(x,s)=\frac{2}{(n+1)(s+1)}\left(\frac{s}{s+1}\right)^{(x-x_0)/2}\sum_{k=1}^{n}\frac{\sqrt{\frac{s+1}{s}}\sin\frac{\pi k x_0}{n+1}-\sin\frac{\pi k(x_0-1)}{n+1}}{\sqrt{\frac{s+1}{s}}-2\cos\frac{\pi k}{n+1}}\sin\frac{\pi k x}{n+1}.$$

Now we can restore the function $\theta(x,m)$ via contour integration

$$\theta(x,m)=\frac{1}{2\pi i}\oint_C\frac{Z(x,s)}{s^{m+1}}ds,$$

where the contour C surrounds the point $s=0$ and is displaced in the regularity area of the function $Z(s)\equiv Z(x,s)$. Hence

$$\theta(x,m)=-\sum_{s_k}\mathrm{Res}\left(\frac{Z(x,s_k)}{s_k^{m+1}}\right),$$

where s_k are the poles out of the regularity area,

$$s_k=\frac{1}{4\cos^2\frac{\pi k}{n+1}-1}$$

(compare to (14)). We are interested only in the asymptotic behavior $m\gg 1$ of the function $\theta_n(x,x_0,m)$ which is determined by the poles nearest to the origin, $s_1=s_n=\frac{1}{3}$ for $n\gg 1$. So we get

$$\theta_n(x,x_0,m)=\frac{4}{n+1}\sin\frac{\pi(x_0+1)}{n+1}\sin\frac{\pi x}{n+1}2^{x_0-x}3^m. \tag{18}$$

To find the function $\theta_n(m)$ we should sum up $\theta(x,x_0,m)$ over all x and x_0; $\theta_n(m)=\Sigma^n{}_{x,x_0=1}\theta_n(x,x_0,m)$. We obtain in the limit for $n=\text{const}\gg 1$ the following expression:

$$\theta_n(m)=\frac{16\pi^2}{\log^4 2}\frac{2^n}{n^3}3^m. \tag{19}$$

The whole number of nonequivalent words follows from Eq. (8) in the limits $n=\text{const}\gg 1$, $\mu\gg 1$,

$$V^{(1)}(n,\mu)=\frac{32\pi^2}{\log^4 2}\frac{2^n}{n^3}7^{\mu-1}. \tag{20}$$

2. The group $\mathcal{LF}_{n+1}^{(2)}$

It is convenient to enumerate the generators $f_{ij}^{(\alpha)}$, $i,j=1,...,n^2$, $\alpha=x,y$, ordering them in a sequence, $(f_{11}^{(x)},f_{11}^{(y)}),\ldots,(f_{1n}^{(x)},f_{1n}^{(y)}),\ldots,(f_{n1}^{(x)},f_{n1}^{(y)}),\ldots,\ (f_{nn}^{(x)},f_{nn}^{(y)})$. For any such sequence we define the ''normal order'' according to the prescriptions (i)–(iii). Let z be the serial number of the pair $(f_{ij}^{(x)},f_{ij}^{(y)})$. Consider the functions $a(z,m)$ and $b(z,m)$ defined as numbers of all distinct sequences of $m+1$ generators satisfying the rules (i)–(iii) and ending with $f_{ij}^{(x)}$ and $f_{ij}^{(y)}$, respectively. One can readily write the evolution equations for $a(z,m)$ and $b(z,m)$ similar to (9),

$$a(z,m+1)=a(z-1,m)+b(z-n,m)+b(z-n+1,m)+b(z,m)+\sum_{z'=z+1}^{n^2}(a(z',m)+b(z',m))$$
$$b(z,m+1)=b(z-n,m)+a(z-1,m)+a(z,m)+\sum_{z'=z+1}^{n^2}(a(z',m)+b(z',m)). \tag{21}$$

Analogous to the case of the group $\mathcal{LF}_{n+1}^{(1)}$ let us suppose in the limit $n\gg 1$ the periodical boundary conditions on the segment $[0,n^2+1]$. So we have

$$\begin{aligned}a(z+1,m+1)-a(z,m+1)&=b(z-n+2,m)-b(z-n,m)\\&\quad-b(z,m)+a(z,m)-a(z-1,m)-a(z+1,m),\\ b(z+1,m+1)-b(z,m+1)&=b(z-n+1,m)-b(z-n,m)-a(z-1,m)-b(z+1,m),\\ a(0,m)=b(0,m)&=a(n^2+1,m)=b(n^2+1,m)=0,\quad m=0,1,...,\\ a(z,0)=b(z,0)&=1,\quad z=1,...,n^2.\end{aligned} \tag{22}$$

The initial conditions differ from (11) because in (22) we do not fix the first generator in the sequence involved.

Assuming that $|a(z,m)-b(z,m)|\to 0(n\to\infty)$ uniformly for z and m, we may pass from (22) to a single closed equation for the function $b(z,m)$. (The self-consistency of this supposition we check at the end of our computations.) So, we get

$$\begin{aligned}b(z+1,m+1)-b(z,m+1)&=b(z-n+1,m)-b(z-n,m)-b(z-1,m)-b(z+1,m),\\ b(0,m)=b(n^2+1,m)&=0,\quad m=0,1,...,\\ b(z,0)&=1,\quad z=1,...,n^2.\end{aligned} \tag{23}$$

Performing the decomposition $b(z,m)=\sum_{k=1}^{n^2}B_k\lambda_k^m\beta_k(z)$, we arrive at the following boundary problem:

$$\begin{aligned}(\lambda_k+1)\beta_k(z+1)-\lambda_k\beta_k(z)+\beta_k(z-1)+\beta_k(z-n)-\beta_k(z-n+1)&=0\\ \beta_k(0)=\beta_k(n^2+1)&=0.\end{aligned} \tag{24}$$

Let us look for the solution of Eq. (24) in the form $\beta_k(z)=p_k^z\sin[\pi kz/(n^2+1)]$. Substituting this ansatz in (24) we obtain an equation for p_k as well as an expression for λ_k,

$$\begin{aligned}&\sin\frac{\pi k}{n^2+1}p_k^{n+1}+\sin\frac{2nk}{n^2+1}p_k^n-\sin\frac{\pi k}{n^2+1}p_k^{n-1}-\sin\frac{\pi kn}{n^2+1}p_k^2\\&+2\cos\frac{\pi k}{n^2+1}\sin\frac{\pi kn}{n^2+1}p_k-\sin\frac{\pi kn}{n^2+1}=0,\\ &\lambda_k=p_k^{-2}-1-p_k^{-n}\frac{\sin\frac{\pi k(n-1)}{n^2+1}}{\sin\frac{\pi k}{n^2+1}}+p_k^{-n-1}\frac{\sin\frac{\pi kn}{n^2+1}}{\sin\frac{\pi k}{n^2+1}}.\end{aligned} \tag{25}$$

In Eq. (25) each root $p_k^{(i)}$ corresponds to different values of $\lambda_k^{(i)}$. However, we are interested only in the asymptotic behavior of $b(z,m)(m\gg 1)$ determined by the largest value of $\lambda_k^{(i)}$ for $k=1$. (Compare to the case of the group $\mathcal{LF}_{n+1}^{(1)}$, Eq. (14).) Equation (25) at $k=1$ and $n\gg 1$ reads

$$\begin{aligned} &p_1^{n+1}+2p_1^n-p_1^{n-1}-n(p_1-1)^2=0,\\ &\lambda_1=p_1^{-2}-1+np_1^{-n-1}-(n-1)p_1^{-n}. \end{aligned} \tag{26}$$

One can easily check that the smallest positive root corresponding to the largest value of $\lambda_1^{(i)}$ is

$$p_1=1-\frac{\log n}{n}+o\left(\frac{\log n}{n}\right), \quad n=\text{const}\gg 1,$$

and

$$\lambda_1=\frac{n}{\log n}+o\left(\frac{n}{\log n}\right), \quad n=\text{const}\gg 1. \tag{27}$$

In the $1+1$-dimensional case the same value of λ_k was given by the right edge of the spectrum, but in $2+1$-dimensional case one can prove that Eq. (25) for any $i>1$ has no solution $\lambda_{n^2}^{(i)}$ growing as fast as λ_1. The coefficients B_k should be found from the initial condition

$$\sum_{k=1}^{n^2} B_k\beta_k(z)=1.$$

As $p_1\to 1$ $(n\gg 1)$ we can expect the set $\beta_k(z)$ to be orthogonal in the vicinity of the left edge of the spectrum, so B_1 is determined basically by the expression

$$B_1=\frac{\int_0^{n^2+1}\beta_1(z)dz}{\int_0^{n^2+1}\beta_1^2(z)dz}, \quad n\gg 1.$$

Now we have the following equation for the function for $b(z,m)$ in the limits $n=\text{const}\gg 1$ and $m\gg 1$,

$$b(z,m)=\frac{4}{\pi}\sin\frac{\pi z}{n^2+1}\left(\frac{n}{\log n}\right)^m. \tag{28}$$

If we suppose the equality $a(z,m)=b(z,m)$, where $b(z,m)$ is given by (28), it is easy to check that $a(z,m)$ and $b(z,m)$ really satisfy Eqs. (22) in the limits $n\gg 1$, $m\gg 1$. This fact proves our assumption about the behaviors of $a(z,m)$ and $b(z,m)$ in a self-consistent way.

The limiting expression of the function $\tilde{\theta}_n(m)$ which is defined similar to the function $\theta_n(m)$ in the $1+1$-dimensional case reads

$$\tilde{\theta}_n(m)=\sum_{z=1}^{n^2} a(z,m)+b(z,m)=\frac{16n^2}{\pi^2}\left(\frac{n}{\log n}\right)^m.$$

Thus, the asymptotics of the number of nonequivalent words of given irreducible length, μ in the limit $n=\text{const}\gg 1$, $\mu\gg 1$ is

$$V^{(2)}(n,\mu)=\sum_{m=0}^{\mu-1} 2^{m+1}\binom{\mu-1}{m}\tilde{\theta}_n(m)=\frac{32n^2}{\pi^2}\left(\frac{2n}{\log n}\right)^{\mu-1} \tag{29}$$

(compare to Eq. (20).)

III. CONCLUSION

The principal difference between the limiting behavior of the partition functions $V^{(1)}(n,\mu)$ and $V^{(2)}(n,\mu)$ (and, hence, between the upper boundaries of the sets $\Omega^{(1)}(n,\mu)$ and $\Omega^{(2)}(n,\mu)$) becomes at most illuminating in the limit $n=\text{const}\gg 1$ and $\mu\to\infty$ if we consider the following limit:

$$f_{1,2}=\left[\lim_{\mu\to\infty}\frac{\log V^{(1,2)}(n,\mu)}{\mu}\right]_{n=\text{const}\gg 1}. \tag{30}$$

Using Eqs. (20) and (29), we get

$$f_1=\log 7$$

$$f_2=\frac{2n}{\log n}.$$

Thus, we can conclude, that with the exponential accuracy in the limit $n=\text{const}\gg 1$ and $\mu\to\infty$ the set $\Omega^{(1)}(n,\mu)$ is bounded from above by the n-independent estimate, i.e., $\Omega^{(1)}(n,\mu)$ is ''representation-independent;'' while the set $\Omega^{(2)}(n,\mu)$ with the same accuracy and in the same limit depends strongly on the braid representation (i.e., on the number of strings, n).

Equations (20) and (29) enable us to make some conclusions about the structure of the graphs corresponding to the groups $\mathcal{LF}_n^{(1)}$ and $\mathcal{LF}_n^{(2)}$. These graphs can be viewed as follows. Take the *free* $1+1$- or $2+1$-groups, where all generators do not commute at all. The graphs of these groups have structures of $2n$- and $2n^2$-branching Cayley trees, where the number of distinct words of length μ is equal to

$$V_{\text{free}}^{(1)}(n,\mu)=2n(2n-1)^{\mu-1} \quad \text{for } 1+1-\text{free group}$$

$$V_{\text{free}}^{(2)}(n,\mu)=2n^2(2n^2-1)^{\mu-1} \quad \text{for } 2+1-\text{free group.}$$

The graphs corresponding to the groups $\mathcal{LF}_n^{(1,2)}$ can be constructed from the graphs of the free groups in accordance with the following recursion procedure:

(i) Take the root vertex of the free group graph and consider all vertices on the distance $\mu=2$. Identify those vertices which correspond to the equivalent words in groups $\mathcal{LF}_n^{(1,2)}$;

(ii) Repeat this procedure taking all vertices on the distance $\mu=(1,2,...)$ and ''gluing'' them on the distance $\mu+2$ according to the definition of the locally free groups.

By means of this procedure we raise a graph which in average has $z_{\text{eff}}^{(1,2)}-1$ distinct branches leading from the level μ to the level $\mu+1$. We may easily find the expressions of $z_{\text{eff}}^{(1,2)}$ using Eqs. (20) and (29). We have in the limit $n=\text{const}\gg 1$ and $\mu\to\infty$,

$$z_{\text{eff}}^{(1,2)}=\frac{V^{(1,2)}(n,\mu+1)}{V^{(1,2)}(n,\mu)}+1=\begin{cases} 8 & \text{for } 1+1-\text{locally free group} \\ \dfrac{2n}{\log n} & \text{for } 2+1-\text{locally free group.}\end{cases}$$

We see that the graph of the group $\mathcal{LF}_n^{(1)}$ coincides (in average) with ($z_{\text{eff}}=8$)-branching Cayley tree for any $n\gg 1$, while the effective coordinational number of the graph of the group $\mathcal{LF}_n^{(2)}$ depends on n and does not ''saturate'' for $n\gg 1$.

ACKNOWLEDGMENTS

We are very grateful to A. Comtet, J. Desbois, and A. M. Vershik for fruitful discussions and useful suggestions.

[1] L. Paris and D. Rolfsen, ''Geometric subgroups of surface braid groups,'' Preprint No. 115, 1997, Université de Bourgogne, Laboratoire de Topologie.
[2] J. Birman, ''Mapping class groups of surfaces,'' Contemp. Math. **78**, 13 (1988).
[3] A. M. Vershik, Topics in Algebra **26**, 467 (1990); Proc. Am. Math. Soc. **148**, 1 (1991).
[4] S. K. Nechaev, A. Yu. Grosberg, and A. M. Vershik, J. Phys. A **29**, 2411 (1996); J. Desbois and S. Nechaev, J. Stat. Phys. **88**, 201 (1997); J. Phys. A **31**, 2767 (1998); A. Comtet and S. Nechaev, *ibid.* **31**, 5609 (1998).
[5] P. Cartier and D. Foata, Lectural Notes in Mathematics (Springer, New York/Berlin, 1969), Vol. 85.
[6] G. X. Viennot, Lect. Notes Math. **1234**, 321 (1986).
[7] A. M. Vershik (private communication).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Graded differential geometry of graded matrix algebras

H. Grosse[a)]
Universität Wien, Institut für Theoretische Physik, Boltzmanngasse 5, A-1090 Wien, Austria

G. Reiter[b)]
Universität Wien, Institut für Theoretische Physik, Boltzmanngasse 5, A-1090 Wien, Austria and Technische Universität Graz, Institut für Theoretische Physik, Petersgasse 16, A-8010 Graz, Austria

(Received 1 June 1999; accepted for publication 16 August 1999)

We study the graded derivation-based noncommutative differential geometry of the $\mathbb{Z}_2$-graded algebra $\mathbb{M}(n|m)$ of complex $(n+m)\times(n+m)$-matrices with the ''usual block matrix grading'' (for $n\neq m$). Beside the (infinite-dimensional) algebra of graded forms, the graded Cartan calculus, graded symplectic structure, graded vector bundles, graded connections and curvature are introduced and investigated. In particular we prove the universality of the graded derivation-based first-order differential calculus and show that $\mathbb{M}(n|m)$ is a ''noncommutative graded manifold'' in a stricter sense: There is a natural body map and the cohomologies of $\mathbb{M}(n|m)$ and its body coincide (as in the case of ordinary graded manifolds). © *1999 American Institute of Physics.* [S0022-2488(99)03811-6]

I. INTRODUCTION

The basic idea of noncommutative geometry,[1] which is the formulation of differential geometric concepts on more general algebras than the algebras of $\mathcal{C}^\infty$ functions on differentiable manifolds, is at least conceptionally rooted in the fact that all the information about the differentiable manifold and its sheaf of differentiable functions is encoded in its algebra of global $\mathcal{C}^\infty$-functions such that differential geometry can be formulated in terms of the latter algebras. Although the $\mathbb{Z}_2$-graded algebra of global sections of the structure sheaf of a graded manifold is a ''baby-noncommutative geometry,'' the differential geometry of graded manifolds is treated and interpreted in the spirit of classical differential and algebraic geometry. So graded manifolds should not be seen as specific noncommutative geometries to which the general methods of noncommutative geometry applies, but rather as a conceptional starting point of a ''supergeneralization'' of noncommutative geometry. Because graded manifolds are completely determined by the $\mathbb{Z}_2$-graded algebra of global sections of their structure sheafs,[2] the natural class of objects to which such a generalization applies are $\mathbb{Z}_2$-graded real and complex algebras, respectively.

There exist already several articles and books in the literature dealing with various aspects of $\mathbb{Z}_2$-graded $\mathbb{C}$-algebras, supersymmetry, and noncommutative geometry. Examples, are Refs. 3 and 4, where notions such as cyclic cohomology and Fredholm modules are treated in the $\mathbb{Z}_2$-graded setting; Ref. 5, where supersymmetry is employed to establish metric, Kähler, and symplectic structures in noncommutative geometry; Ref. 6, where the concept of a spectral triple is extended to algebras which contain bosonic and fermionic degrees of freedom, and Refs. 7 and 8, where the possibility of generalizing matrix geometry to the $\mathbb{Z}_2$-graded framework is presented. Here we want to adopt a somewhat different point of view.

If $\mathcal{O}(X)$ is the $\mathbb{Z}_2$-graded algebra of global sections of the (complexified) structure sheaf of some graded manifold (complex) global graded vector fields on the graded manifold are by definition graded derivations of $\mathcal{O}(X)$. All global graded vector fields $\mathfrak{D}^g(X)$ from a complex Lie

a)Electronic mail: grosse@doppler.thp.univie.ac.at
b)Electronic mail: reiter@itp.tu-graz.ac.at

subsuperalgebra and a graded module over the graded center $\mathcal{Z}^g(\mathcal{O}(X))$. (Complex) global graded p-forms for $p \in \mathbb{N}$ are defined as p-fold $\mathcal{Z}^g(\mathcal{O}(X))$-graded-multilinear, graded-alternating maps from $\mathfrak{D}^g(X)$ to $\mathcal{O}(X)$ and one can form the $\mathbb{N}_0 \times \mathbb{Z}_2$-bigraded $\mathbb{C}$-vector space $\Omega^g(X)$ of global graded forms as the direct sum of all $\mathbb{Z}_2$-graded $\mathbb{C}$-vector spaces $\Omega^{g,p}(X)$ of global graded p forms. The graded wedge product as well as the whole graded Cartan calculus on $\Omega^g(X)$ can be introduced (see Refs. 2 and 9, for example) by employing only the facts that $\mathfrak{D}^g(X)$ is a $\mathbb{C}$-Lie superalgebra and $\mathbb{Z}_2$-graded $\mathcal{Z}^g(\mathcal{O}(X))$-module and that $\mathcal{O}(X)$ is a $\mathbb{Z}_2$-graded $\mathbb{C}$-algebra.

The important feature of the recipe for the construction of the graded deRham complex and the graded Cartan calculus formulated above is that it uses only the graded algebra structure of $\mathcal{O}(X)$. In particular it does not use the graded commutativity of $\mathcal{O}(X)$ and we can define on arbitrary $\mathbb{Z}_2$-graded $\mathbb{C}$ algebras noncommutative graded differential calculi.

What we have just described is mutatis mutandis the basic idea of the so-called derivation-based differential calculi[10–13] transposed to the $\mathbb{Z}_2$-graded setting. Such graded derivation-based differential calculi were investigated for arbitrary, but graded-commutative $\mathbb{Z}_2$-graded algebras in the framework of $\mathbb{Z}_2$-graded Lie–Cartan pairs.[3,4,14]

Motivated by the rich differential geometric structure of ordinary matrix algebras[10,11,15] and by our previous investigation of the fuzzy supersphere,[16,17] where each truncated supersphere was a graded matrix algebra in particular, we will investigate especially the differential calculus based on all graded derivations on the $\mathbb{Z}_2$-graded $\mathbb{C}$-algebra $\mathbb{M}(n|m)$ of complex $(n+m)\times(n+m)$-matrices $(n,m \in \mathbb{N}_0, n+m \in \mathbb{N})$, $\mathbb{Z}_2$-graded by declaring the vector subspace

$$\mathbb{M}(n|m)_{\bar{0}} := \left\{ M = \begin{pmatrix} M_1 & 0 \\ 0 & M_4 \end{pmatrix} \middle| M_1 \in \mathbb{M}(n), M_4 \in \mathbb{M}(m) \right\} \tag{1}$$

of $\mathbb{M}(n|m)$ as even and the vector subspace

$$\mathbb{M}(n|m)_{\bar{1}} := \left\{ M = \begin{pmatrix} 0 & M_2 \\ M_3 & 0 \end{pmatrix} \middle| M_2 \in \mathbb{M}(n,m), M_3 \in \mathbb{M}(m,n) \right\} \tag{2}$$

of $\mathbb{M}(n|m)$ as odd. Here $\mathbb{M}(n,m)$ and $\mathbb{M}(n)$ denote the vector space of $n \times m$, respectively, the algebra of $n \times n$-matrices and we will always assume $n \neq m$.

Section II is devoted to the precise definition of the graded derivation-based differential calculus on $\mathbb{M}(n|m)$ as described in Sec. I and its immediate consequences. The resulting complexes are nothing else than the complexes of Lie superalgebra cohomology with values in $\mathbb{M}(n|m)$ and typically infinite. The latter fact shows, in particular, that the complex is completely different from that proposed in Refs. 7 and 8.

In Sec. III we continue the investigation of the differential calculus using the facts that there exist graded-commutative homogeneous bases in the $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$-modules of all graded p-forms and that all graded derivations of $\mathbb{M}(n|m)$ are inner. In particular, we construct an invariant graded one-form, which determines the differential of graded matrices in terms of graded commutators (within the graded algebra of graded forms) and show that the first-order differential calculus is universal.

Associated with every graded manifold there exists an even, surjective algebra homomorphism β_X from the $\mathbb{Z}_2$-graded algebra of global sections of the (complexified) structure sheaf of the graded manifold $\mathcal{O}(X)$ to the algebra $\mathcal{C}^\infty(X)$ of (complex) $\mathcal{C}^\infty$-functions on its body manifold X. We call this map, which is the key to all further developments in graded manifold theory, the body map. In Sec. IV we show that there exists a natural noncommutative analog to the body map. It induces an isomorphism between the graded derivation-based cohomology of $\mathbb{M}(n|m)$ and the derivation-based cohomology of its body, such that the situation described by a theorem of Kostant[9] is generalized to the noncommutative case.

In Sec. V we study the graded symplectic geometry of $\mathbb{M}(n|m)$. As for ordinary matrix geometry,[10,11,18,19] which is included as special case, there exists a graded symplectic structure such that the induced graded Poisson bracket on $\mathbb{M}(n|m)$ is (i times) the graded commutator on $\mathbb{M}(n|m)$.

In Sec. VI we investigate the noncommutative generalization of graded vector bundles over graded manifolds. Graded vector bundles over a graded manifold $(X,\mathcal{O})$ are usually introduced as locally graded-free $\mathcal{O}$-modules.[2,9,20] In the spirit of noncommutative geometry[1,21] we concentrate on the module of global sections and introduced graded vector bundles over $\mathbb{M}(n|m)$ as $\mathbb{Z}_2$-graded, finitely generated (graded-projective) $\mathbb{M}(n|m)$-modules. Concepts like connections and curvature can be generalized to the $\mathbb{Z}_2$-graded noncommutative setting.

In addition we have included an appendix in which we analyze the associative product of supertrace-free, graded matrices. The results of this analysis are used for a minimality proof in Sec. III.

There will appear lots of $\mathbb{Z}_2$-graded objects. If the object is denoted by $\mathcal{A}$ its even part is denoted by $\mathcal{A}_{\bar{0}}$, its odd part by $\mathcal{A}_{\bar{1}}$. If a is some homogeneous element of such an object its degree will be denoted by $\overline{a}$. Speaking of grading in the context of an ungraded object, we mean that the object is endowed with its trivial graduation. If for some construction the $\mathbb{Z}_2$-grading is indicated by an index ''g'', we omit this index in the case of trivial graduation.

II. GRADED DERIVATION-BASED DIFFERENTIAL CALCULUS ON GRADED MATRIX ALGEBRAS

We will interpret the $\mathbb{C}$-Lie superalgebra and $\mathbb{Z}_2$-graded $\mathcal{Z}^g(\mathbb{M}(n|m))$-module $\mathfrak{Der}^g(\mathbb{M}(n|m))$ of all graded derivations of $\mathbb{M}(n|m)$ as ''noncommutative graded vector fields'' on $\mathbb{M}(n|m)$. Because $\mathbb{M}(n|m)$ is graded central,

$$\mathcal{Z}^g(\mathbb{M}(n|m)) = \mathcal{Z}^g(\mathbb{M}(n|m))_{\bar{0}} = \mathbb{C}1_{n+m} \cong \mathbb{C}, \tag{3}$$

the concept of graded $\mathcal{Z}^g(\mathbb{M}(n|m))$ multilinearity reduces to ordinary $\mathbb{C}$ multilinearity and we can employ ideas and results of Lie superalgebra cohomology (see Refs. 22 and 23) for the construction of the graded derivation-based differential calculus on $\mathbb{M}(n|m)$.

For every natural number $p \in \mathbb{N}$ let us denote by $\mathrm{Hom}^p(\mathfrak{Der}^g(\mathbb{M}(n|m));\mathbb{M}(n|m))$ the $\mathbb{Z}_2$-graded $\mathbb{C}$-vector space of all p-linear maps $\mathfrak{Der}^g(\mathbb{M}(n|m)) \times \dots_p \times \mathfrak{Der}^g(\mathbb{M}(n|m)) \to \mathbb{M}(n|m)$ and by $\mathfrak{S}_p$ the symmetric group of p letters. Introducing a commutation factor $\gamma_p : \mathfrak{S}_p \times \mathbb{Z}_2 \times \dots_p \times \mathbb{Z}_2 \to \{\pm 1\}$ via

$$\gamma_p(\sigma;\overline{\underline{\imath}_1},\dots,\overline{\underline{\imath}_p}) := \prod_{\substack{r,s=1,\cdots p;r<s \\ \sigma^{-1}(r)>\sigma^{-1}(s)}} (-1)^{\overline{\underline{\imath}_r}\,\overline{\underline{\imath}_s}}, \tag{4}$$

we can define a representation π of $\mathfrak{S}_p$ on $\mathrm{Hom}^p(\mathfrak{Der}^g(\mathbb{M}(n|m));\mathbb{M}(n|m))$ by

$$(\pi_\sigma \omega)(D_1,\dots,D_p) := \gamma_p(\sigma;\overline{D_1},\dots,\overline{D_p})\,\omega(D_{\sigma(1)},\dots,D_{\sigma(p)}) \tag{5}$$

for all $\omega \in \mathrm{Hom}^p(\mathfrak{Der}^g(\mathbb{M}(n|m));\mathbb{M}(n|m))$, all homogeneous $D_1,\dots,D_p \in \mathfrak{Der}^g(\mathbb{M}(n|m))$, and all $\sigma \in \mathfrak{S}_p$. Now by definition a p-linear map $\omega \in \mathrm{Hom}^p(\mathfrak{Der}^g(\mathbb{M}(n|m));\mathbb{M}(n|m))$ is called graded alternating if

$$\pi_\sigma \omega = \mathrm{sgn}\,\sigma\,\omega \tag{6}$$

is fulfilled for all $\sigma \in \mathfrak{S}_p$ and we interpret such maps as graded p-forms on $\mathbb{M}(n|m)$. All graded p-forms on $\mathbb{M}(n|m)$ form a graded vector subspace of $\mathrm{Hom}^p(\mathfrak{Der}^g(\mathbb{M}(n|m));\mathbb{M}(n|m))$, which we will denote by $\Omega^{g,p}(\mathbb{M}(n|m))$.

A general graded form on $\mathbb{M}(n|m)$ is an element of the direct sum

$$\Omega^g(\mathbb{M}(n|m)) := \bigoplus_{p\in N_0} \Omega^{g,p}(\mathbb{M}(n|m)), \tag{7}$$

where we set $\Omega^{g,0}(\mathbb{M}(n|m)) := \mathbb{M}(n|m)$. Employing the multiplicative structure of $\mathbb{M}(n|m)$ we can proceed exactly as in the case of graded manifolds[2,9] (respectively, graded Lie–Cartan pairs[3,4,14]) to introduce a graded wedge product on $\Omega^g(\mathbb{M}(n|m))$. So we first define for all $p,p' \in \mathbb{N}_0, \overline{\imath}, \overline{i'} \in \mathbb{Z}_2$ a bilinear map $\wedge : \Omega^{g,p}(\mathbb{M}(n|m))_{\overline{\imath}} \times \Omega^{g,p}(\mathbb{M}(n|m))_{\overline{i'}} \to \Omega^{g,p+p'}(\mathbb{M}(n|m))_{\overline{i+i'}}$ by

$$(\omega\wedge\omega')(D_1,\ldots,D_{p+p'}) := \frac{1}{p!p'!} \sum_{\sigma\in\mathfrak{S}_{p+p'}} \operatorname{sgn}\sigma\, \gamma_{p+p'}(\sigma;\overline{D}_1,\ldots,\overline{D}_{p+p'}) \cdot (-1)^{\overline{i'}\sum_{l=1}^{p}\overline{D}_{\sigma(l)}} \omega(D_{\sigma(1)},\ldots,D_{\sigma(p)})\,\omega'(D_{\sigma(p+1)},\ldots,D_{\sigma(p+p')}). \tag{8}$$

for all homogeneous $D_1,\ldots,D_{p+p'} \in \mathfrak{Der}^g(\mathbb{M}(n|m))$ and extend these by bilinearity to $\Omega^g(\mathbb{M}(n|m))$. With respect to it, $\Omega^g(\mathbb{M}(n|m))$ becomes a $\mathbb{N}_0\times\mathbb{Z}_2$-bigraded $\mathbb{C}$ algebra.

Via

$$(L_{D_0}\omega)(D_1,\ldots,D_p) := D_0(\omega(D_1,\ldots,D_p)) - \sum_{l=1}^{p} (-1)^{\overline{D}_0(\overline{\omega}+\sum_{l'=1}^{l-1}\overline{D}_{l'})} \omega(D_1,\ldots,[D_0,D_l]_g,\ldots,D_p), \tag{9}$$

$$(\iota_{D_1}\omega)(D_2,\ldots,D_p) := \omega(D_1,D_2,\ldots,D_p) \tag{10}$$

and

$$d\omega(D_0,\ldots,D_p) = \sum_{l=0}^{p} (-1)^{l+\overline{D}_l(\overline{\omega}+\sum_{l'=0}^{l-1}\overline{D}_{l'})} L_{D_l}(\omega(D_0,\ldots,\check{D}_l,\ldots,D_p)) + \sum_{0\le l<l'\le p} (-1)^{l'+\overline{D}_{l'}\sum_{l''=l+1}^{l'-1}\overline{D}_{l''}} \omega(D_0,\ldots,D_{l-1},[D_l,D_{l'}]_g,\ldots,\check{D}_{l'},\ldots,D_p) \tag{11}$$

for all homogeneous $D_0,D_1,\ldots,D_p \in \mathfrak{Der}^g(\mathbb{M}(n|m))$ and all homogeneous $\omega \in \Omega^{g,p}(\mathbb{M}(n|m))$ ($\vee$ denotes omission), one defines homogeneous endomorphisms $\Omega^g(\mathbb{M}(n|m)) \to \Omega^g(\mathbb{M}(n|m))$ of bidegree $(0,\overline{D}_0)$, $(-1,\overline{D}_0)$, and $(1,\overline{0})$ respectively. The assignments $D\mapsto\iota_D$ and $D\mapsto L_D$ extend to $\mathbb{C}$-linear maps $\mathfrak{Der}^g(\mathbb{M}(n|m)) \to \mathrm{End}(\Omega^g(\mathbb{M}(n|m)))$ and L is a graded representation of $\mathfrak{Der}^g(\mathbb{M}(n|m))$ in particular. Furthermore the relations

$$d\circ d = 0, \quad d\circ L_D = L_D\circ d, \tag{12}$$

as well as

$$\begin{aligned} &\iota_D\iota_{D'} + (-1)^{\overline{DD'}}\iota_{D'}\circ\iota_D = 0,\\ &(L_D\circ\iota_{D'} - \iota_{D'}\circ L_D)\omega = (-1)^{\overline{D}\overline{\omega}}\iota_{[D,D']_g}\omega,\\ &(\iota_D\circ d + d\circ\iota_D)\omega = (-1)^{\overline{D}\overline{\omega}}L_D\omega \end{aligned} \tag{13}$$

for all homogeneous $D,D' \in \mathfrak{Der}^g(\mathbb{M}(n|m))$ and all bihomogeneous $\omega \in \Omega^g(\mathbb{M}(n|m))$ are known from Lie superalgebra cohomology.[23]

By analogy with the case of graded manifolds we call d, L_D, and ι_D the exterior derivative, Lie derivative and interior product (with respect to a graded vector field $D \in \mathfrak{Der}^g(\mathbb{M}(n|m))$). (12) and (13) tell us that they fulfill exactly the same relations as in the ''graded-commutative case'', but this observation remains also true for the graded wedge product (8) of graded forms.

Proposition 1: The relations

$$L_D(\omega \wedge \omega') = (L_D\omega)\wedge \omega' + (-1)^{\bar{D}\bar{\omega}}\omega \wedge L_D\omega',$$

$$\iota_D(\omega\wedge\omega') = (-1)^{\bar{D}\bar{\omega}'}(\iota_D\omega)\wedge\omega' + (-1)^p \omega \wedge \iota_D\omega', \tag{14}$$

$$d(\omega\wedge\omega') = (d\omega)\wedge\omega' + (-1)^p\omega\wedge d\omega'$$

are fulfilled for all homogeneous $D \in \mathfrak{Der}^g(\mathbb{M}(n|m)), \omega \in \Omega^{g,p}(\mathbb{M}(n|m)), \omega' \in \Omega^{g,p'}(\mathbb{M}(n|m))$.

Proof: This can be shown exactly as in the case of graded manifolds. That is, one starts with a direct proof of the second relation and proofs the other equations inductively using the last two relations (13).

Because we interpret d as exterior derivative, the Lie superalgebra cohomology of $\mathfrak{Der}^g(\mathbb{M}(n|m))$ with values in $\mathbb{M}(n|m)$,

$$H(\mathbb{M}(n|m)) \equiv \bigoplus_{p\in N_0} H^p(\mathbb{M}(n|m)) := \frac{\ker d}{\operatorname{im} d}, \tag{15}$$

has to be seen as analogous to the graded deRham cohomology on graded manifolds. Via

$$[\omega]\wedge[\omega'] := [\omega\wedge\omega'], \tag{16}$$

the above graded derivation-based cohomology of $\mathbb{M}(n|m)$ becomes a $\mathbb{N}_0\times\mathbb{Z}_2$-bigraded $\mathbb{C}$-algebra and we will continue to study it in Sec. IV.

III. HOMOGENEOUS BASES AND THE CANONICAL GRADED ONE-FORM

Whereas the definitions and results of the preceding considerations apply to each $\mathbb{Z}_2$-graded, graded-central $\mathbb{C}$-algebra we shall now employ more specific properties of $\mathbb{M}(n|m)$. There will result similar formulas as in ''ordinary'' matrix geometry,[10,11,15] which is included as a special case.

The sets $\Omega^{g,p}_{\mathcal{Z}g}(\mathbb{M}(n|m))$ of graded p-forms with values in the graded center of $\mathbb{M}(n|m)$ form graded vector subspaces of $\Omega^{g,p}(\mathbb{M}(n|m))$ for all $p\in\mathbb{N}$ and one can introduce

$$\Omega^{g}_{\mathcal{Z}g}(\mathbb{M}(n|m)) := \bigoplus_{p\in N_0}\Omega^{g,p}_{\mathcal{Z}g}(\mathbb{M}(n|m)) \tag{17}$$

with $\Omega^{g,0}_{\mathcal{Z}g}(\mathbb{M}(n|m)) = \mathcal{Z}^g(\mathbb{M}(n|m))$. $\Omega^{g}_{\mathcal{Z}g}(\mathbb{M}(n|m))$ is a bigraded subalgebra of $\Omega^g(\mathbb{M}(n|m))$, whose product fulfills

$$\omega\wedge\omega' = (-1)^{pp' + \overline{\omega\omega'}}\omega'\wedge\omega \tag{18}$$

for all homogeneous $\omega\in\Omega^{g,p}_{\mathcal{Z}g}(\mathbb{M}(n|m)), \omega'\in\Omega^{g,p'}_{\mathcal{Z}g}(\mathbb{M}(n|m))$, and which is stable with respect to the whole Cartan calculus.

Now let us introduce a homogeneous basis $\{\partial_A\}_{A=1,\dots,n'+m'}$ of $\mathfrak{Der}^g(\mathbb{M}(n|m))$ with $\partial_1,\dots,\partial_{n'}\in\mathfrak{Der}^g(\mathbb{M}(n|m))_{\bar{0}}$, $\partial_{n'+1},\dots,\partial_{n'+m'}\in\mathfrak{Der}^g(\mathbb{M}(n|m))_{\bar{1}}$, where we set $n' := \dim_{\mathbb{C}}\mathfrak{Der}^g(\mathbb{M}(n|m))_{\bar{0}}$ and $m' := \dim_{\mathbb{C}}\mathfrak{Der}^g(\mathbb{M}(n|m))_{\bar{1}}$. If $\{\eta^A\}_{A=1,\dots,n'+m'}$ denotes the dual basis to $\{\partial_A\}_{A=1,\dots,n'+m'}$ we can introduce a homogeneous basis $\{\theta^A\}_{A=1,\dots,n'+m'}$ of $\Omega^{g,1}_{\mathcal{Z}g}(\mathbb{M}(n|m))$ by

$$\theta^A(D) := \eta^A(D) 1_{n+m} \tag{19}$$

for all $D \in \mathfrak{Der}^g(\mathbb{M}(n|m))$. Employing the standard isomorphisms between graded-alternating maps and the graded exterior algebra,[2,23] one deduces that

$$\{\theta^{A_1} \wedge \cdots \wedge \theta^{A_p} | (A_1, \ldots, A_p) \in \mathcal{J}_p^{n'|m'}\} \tag{20}$$

with

$$\mathcal{J}_p^{n'|m'} := \{(k_1, \ldots, k_{p'}, \alpha_{p'+1}, \ldots, \alpha_p) | 0 \leqslant p' \leqslant p; k_1, \ldots, k_{p'} = 1, \ldots, n'; \tag{21}$$

$$\alpha_{p'+1}, \ldots, \alpha_p = n'+1, \ldots, n'+m'; k_1 < k_2 < \cdots < k_{p'} < \alpha_{p'+1} \leqslant \cdots \leqslant \alpha_{p-1} \leqslant \alpha_p\}$$

is an homogeneous basis of $\Omega_{\mathcal{Z}g}^{g,p}(\mathbb{M}(n|m)), p \in \mathbb{N}$.

Because of

$$M \wedge \omega = (-1)^{\bar{M}\bar{\omega}} \omega \wedge M, \tag{22}$$

for all homogeneous $M \in \mathbb{M}(n|m)$ and all bihomogeneous $\omega \in \Omega_{\mathcal{Z}g}^g(\mathbb{M}(n|m))$, the $\mathbb{N}_0 \times \mathbb{Z}_2$-bigraded $\mathbb{C}$-algebras $\Omega^g(\mathbb{M}(n|m))$ and $\mathbb{M}(n|m) \hat{\otimes}_{\mathbb{C}} \Omega_{\mathcal{Z}g}^g(\mathbb{M}(n|m))$, where $\hat{\otimes}$ denotes the tensor product of $\mathbb{Z}_2$-graded algebras, are canonically isomorphic. In particular we can conclude:

Proposition 2: The $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$-bimodules $\Omega^{g,p}(\mathbb{M}(n|m))$ are graded-free for both multiplications and for all $p \in \mathbb{N}_0$. The set (20) determines a homogeneous basis of the left (right), $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$-module $\Omega^{g,p}(\mathbb{M}(n|m))$.

Consequently every $\omega \in \Omega^{g,p}(\mathbb{M}(n|m))$ can be written as

$$\omega = \sum_{(A_1, \ldots, A_p) \in \mathcal{J}_p^{n'|m'}} \omega_{A_1 \ldots A_p} \wedge \theta^{A_1} \wedge \cdots \wedge \theta^{A_p} \tag{23}$$

with unique coefficients $\omega_{A_1 \ldots, A_p} \in \mathbb{M}(n|m)$. Explicitly these coefficients are given by

$$\omega_{A_1 \ldots A_p} = (-1)^{(1/2)p''(p''-1)} \frac{1}{\Pi_{l=1}^{n'+m'} N_l!} \omega(\partial_{A_1}, \ldots, \partial_{A_p}), \tag{24}$$

where p'' is the number of entries in $(A_1, \ldots, A_p)$ greater than n' and N_l is the number of entries in $(A_1, \ldots, A_p)$ being equal to l.

In order to investigate graded derivations of $\mathbb{M}(n|m)$ (we include the case $n=m$ for the moment) let us denote by $\mathfrak{gl}(n|m)$ the (complex) general linear Lie superalgebra and by $\mathfrak{sl}(n|m)$ the (complex) special linear Lie superalgebra. The adjoint representation of $\mathfrak{gl}(n|m)$ is at the same time a Lie superalgebra homomorphism $\mathrm{ad}: \mathfrak{gl}(n|m) \to \mathfrak{Der}^g(\mathbb{M}(n|m))$ and, as we will see, the structure of $\mathfrak{Der}^g(\mathbb{M}(n|m))$ and its Lie subsuperalgebras is determined by this homomorphism.

Proposition 3: If $\mathfrak{L}$ is a Lie subsuperalgebra of $\mathfrak{gl}(n|m)$ then

$$\mathcal{L}^{\mathrm{ad}} := \mathrm{im}\, \mathrm{ad}|_{\mathcal{L}} \tag{25}$$

is a Lie *subsuperalgebra of $\mathfrak{Der}^g(\mathbb{M}(n|m))$. Conversely every Lie subsuperalgebra of $\mathfrak{Der}^g(\mathbb{M}(n|m))$ is of this form. There are two different cases.*

(i) *For $n \neq m$ the restriction of* ad *to $\mathfrak{sl}(n|m)$ is a Lie superalgebra isomorphism onto $\mathfrak{Der}^g(\mathbb{M}(n|m))$ and the various restrictions of* ad *induce a bijective correspondence between Lie subsuperalgebras of $\mathfrak{sl}(n|m)$ and Lie subsuperalgebras of $\mathfrak{Der}^g(\mathbb{M}(n|m))$.*

(ii) *For $n=m$ there is no Lie subsuperalgebra $\mathcal{L}$ of $\mathfrak{gl}(n|n)$ such that the restriction of* ad *to $\mathcal{L}$ becomes a Lie superalgebra isomorphism onto $\mathfrak{Der}^g(\mathbb{M}(n|m))$.*

Proof: An even graded derivation of $\mathbb{M}(n|m)$ is just an ordinary derivation of the $\mathbb{C}$-algebra $\mathbb{M}(n+m)$ and these are inner, because the first Hochschild cohomology group of $\mathbb{M}(n+m)$ with values in $\mathbb{M}(n+m)$ is trivial.[24] Introducing

$$\Gamma := \begin{pmatrix} 1_n & 0 \\ 0 & -1_m \end{pmatrix}$$

we find for some $D \in \mathfrak{Der}^g(\mathbb{M}(n|m))_1^-$ and all homogeneous $M \in \mathbb{M}(n|m)$,

$$DM = \mathrm{ad}(\tfrac{1}{2}(D\Gamma)\Gamma)(M),$$

from which we can conclude that D is inner. Consequently, if $\mathfrak{D}$ is a Lie subsuperalgebra of $\mathfrak{Der}^g(\mathbb{M}(n|m))$, then $\mathfrak{L} := \mathrm{ad}^{-1}(\mathfrak{D})$ is a Lie subsuperalgebra of $\mathfrak{gl}(n|m)$ with $\mathfrak{L}^{\mathrm{ad}} = \mathfrak{D}$. (i) and (ii) are consequences of $1_{n+m} \notin \mathfrak{sl}(n|m)$ for $n \neq m$, respectively, $1_{2n} \in [\mathfrak{gl}(n|n)_1^-, \mathfrak{gl}(n|n)_1^-]_g$. □

The ultimate reason for restricting our geometric investigation to the case $n \neq m$ lies in the existence of the Lie superalgebra isomorphism $\mathrm{ad}: \mathfrak{sl}(n|m) \to \mathfrak{Der}^g(\mathbb{M}(n|m))$. The elements of every homogeneous basis $\{\partial_A\}_{A=1,\ldots,n'+m'}$ of $\mathfrak{Der}^g(\mathbb{M}(n|m))$ are given by

$$\partial_A = \mathrm{ad}\, E_A, \tag{26}$$

where $\{E_A\}_{A=1,\ldots,n'+m'}$ is a homogeneous basis of $\mathfrak{sl}(n|m)$ and we have $n' = n^2 + m^2 - 1$, $m' = 2nm$ in particular. Moreover, the structure constants c_{AB}^C appearing in

$$[\partial_A, \partial_B]_g = \sum_{C=1}^{(n+m)^2-1} c_{AB}^C \partial_C \tag{27}$$

are the structure constants of the homogeneous $\mathfrak{sl}(n|m)$-basis $\{E_A\}_{A=1,\ldots,(n+m)^2-1}$ and one deduces the nice formulas

$$dE_A = -\sum_{B,C=1}^{(n+m)^2-1} c_{AB}^C E_C \wedge \theta^B \tag{28}$$

and

$$d\theta^A = \frac{1}{2} \sum_{B,C=1}^{(n+m)^2-1} c_{BC}^A \theta^C \wedge \theta^B. \tag{29}$$

The even graded one-form

$$\Theta := \sum_{A=1}^{(n+m)^2-1} E_A \wedge \theta^A \tag{30}$$

will be called the canonical graded one-form, because it plays a distinguished role.

Proposition 4: The definition of Θ is independent of the choice of the homogeneous basis of $\mathfrak{Der}^g(\mathbb{M}(n|m))$ and

$$\Theta(\mathrm{ad}\, M) = M - \frac{1}{n-m} \mathrm{Tr}_s(M) 1_{n+m} \tag{31}$$

for all $M \in \mathbb{M}(n|m)$. Θ is $(\mathfrak{Der}^g(\mathbb{M}(n|m))$-invariant and this property determines Θ up to constant multiples. Furthermore its exterior differential fulfills

$$d\Theta = \Theta \wedge \Theta \tag{32}$$

and the exterior differential of each $M \in \mathbb{M}(n|m)$ *can be expressed according to*

$$dM = [\Theta, M]_g \equiv \Theta \wedge M - M \wedge \Theta. \tag{33}$$

Proof: Beside the uniqueness statement only simple calculations are involved [for which one can use (28) and (29) advantageously]. The irreducibility of the adjoint representation of $\mathfrak{sl}(n|m)$[25] guarantees that

$$L_D \omega = 0, \quad \omega \in \Omega^{g,1}(\mathbb{M}(n|m)),$$

for all $D \in \mathfrak{Der}^g(\mathbb{M}(n|m))$ implies $\omega = c\Theta$, $c \in \mathbb{C}$. □

Finally we note that $\Omega^g(\mathbb{M}(n|m))$ is in a certain sense minimal (for the ungraded case see Refs. 12 and 15).

Proposition 5: (28) can be inverted according to

$$\theta^A = 4(n-m)^2 \sum_{B,C,D=1}^{(n+m)^2-1} (-1)^{\bar{E}_B \bar{E}_D} K^{AB} K^{CD} E_C E_B \wedge dE_D, \tag{34}$$

where K is the Killing form of $\mathfrak{sl}(n|m)$ *and* K^{AB} *denote the components of the inverse matrix of* $(K(E_A, E_B))$. *Consequently, if* Ω *is differential subalgebra of* $\Omega^g(\mathbb{M}(n|m))$ *containing* $\mathbb{M}(n|m)$, *then* $\Omega = \Omega^g(\mathbb{M}(n|m))$.

Proof: The minimality statement follows from (34) because of proposition 2. In order to show (34) one uses (28) and expands the threefold product of the basis elements E_A according to (A4). Using the results of proposition A (34) follows.

The second part of proposition 5 can be stated differently: The canonical even algebra homomorphisms from the (intrinsic) $\mathbb{Z}_2$-graded universal differential envelope of $\mathbb{M}(n|m)$ to $\Omega^g(\mathbb{M}(n|m))$ (see Refs. 3 and 26 for a precise definition) is onto. The restriction of this homomorphism to the corresponding first-order differential calculi is an isomorphism.

IV. COHOMOLOGY AND THE NONCOMMUTATIVE BODY MAP

We will call the even, surjective $\mathbb{C}$-linear map

$$\beta: \mathbb{M}(n|m) \to \mathbb{M}(\underline{n}) \quad \text{with } \underline{n} := \begin{cases} n & \text{if } n > m \\ m & \text{if } n < m, \end{cases} \tag{35}$$

defined by

$$\beta(M) \equiv \beta\left(\begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} \right) := \begin{cases} M_1 & \text{if } \underline{n} = n \\ M_4 & \text{if } \underline{n} = m \end{cases} \tag{36}$$

canonical body map of $\mathbb{M}(n|m)$. A justification for choosing this terminology will result from the investigation of its properties: They are completely analogous to the corresponding map of graded manifolds if one takes the noncommutativity of $\mathbb{M}(n|m)$ and its ''body'' $\mathbb{M}(\underline{n})$ appropriately into account. In order to distinguish between ''objects'' on $\mathbb{M}(n|m)$ and corresponding ''objects'' on the body we underline the latter.

The restriction of β to $\mathcal{Z}^g(\mathbb{M}(n|m))$ is an even algebra homomorphism onto $\mathcal{Z}(\mathbb{M}(\underline{n}))$ and by

$$\iota(\underline{M}) := \begin{cases} \begin{pmatrix} \underline{M} & 0 \\ 0 & \frac{1}{n}\operatorname{Tr}(\underline{M}) 1_m \end{pmatrix} & \text{if } \underline{n}=n \\ \begin{pmatrix} \frac{1}{m}\operatorname{Tr}(\underline{M}) 1_n & 0 \\ 0 & \underline{M} \end{pmatrix} & \text{if } \underline{n}=m \end{cases} \tag{37}$$

we can introduce an even, injective $\mathbb{C}$-linear map $\iota:\mathbb{M}(\underline{n})\to\mathbb{M}(n|m)$, which is right inverse to β on the one hand and whose restriction to $\mathcal{Z}(\mathbb{M}(\underline{n}))$ is an even algebra homomorphism into $\mathcal{Z}^g(\mathbb{M}(n|m))$ on the other hand.

Analogous to the body map of graded manifolds β induces a Lie algebra homomorphism $\hat{\beta}:\mathfrak{Der}^g(\mathbb{M}(n|m))_{\bar{0}}\to\mathfrak{Der}(\mathbb{M}(\underline{n}))$ via

$$\hat{\beta}(D)\beta(M) := \beta(DM) \tag{38}$$

for all $M\in\mathbb{M}(n|m)$. $\hat{\beta}$ is surjective because of

$$\hat{\beta}(\operatorname{ad} E) = \operatorname{ad}\beta(E) \tag{39}$$

for all $E\in\mathfrak{sl}(n|m)_{\bar{0}}$ and in addition $\hat{\iota}:\mathfrak{Der}(\mathbb{M}(\underline{n}))\to\mathfrak{Der}^g(\mathbb{M}(n|m))_{\bar{0}}$,

$$\hat{\iota}(\operatorname{ad}\underline{E}) := \operatorname{ad}\iota(\underline{E}) \tag{40}$$

is a Lie algebra homomorphism right-inverse to $\hat{\beta}$.
Now we can introduce even $\mathbb{C}$-linear maps $\beta^{(p)}:\Omega^{g,p}(\mathbb{M}(n|m))\to\Omega^p(\mathbb{M}(\underline{n}))$, $p\in\mathbb{N}$ by

$$(\beta^{(p)}(\omega))(\underline{D}_1,\ldots,\underline{D}_p) := \beta(\omega(\hat{\iota}(\underline{D}_1),\ldots,\hat{\iota}(\underline{D}_p))) \tag{41}$$

for all $\underline{D}_1,\ldots,\underline{D}_p\in\mathfrak{Der}(\mathbb{M}(\underline{n}))$. $\mathfrak{sl}(n|m)_{\bar{0}}$ is canonically isomorphic to $\mathfrak{sl}(n)\oplus\mathfrak{gl}(1)\oplus\mathfrak{sl}(m)$ and we can choose a homogeneous basis $\{E_A\}_{A=1,\ldots,(n+m)^2-1}$ of $\mathfrak{sl}(n|m)$ such that $E_1,\ldots,E_{\underline{n}^2-1}$ lie in the isomorphic copy of $\mathfrak{sl}(\underline{n})$ and $E_{\underline{n}^2},\ldots,E_{n^2+m^2-1}$ in the isomorphic copy of $\mathfrak{gl}(1)\oplus\mathfrak{sl}(\underline{m})$ with $\underline{m}:=\min\{n,m\}$. Then the elements $\beta(E_k) := \underline{E}_k$, $k=1,\ldots,\underline{n}^2-1$ form a basis of $\mathfrak{sl}(\underline{n})$. Denoting the elements of the basis of $\Omega^1(\mathbb{M}(\underline{n}))$ corresponding to $\{\underline{E}_k\}_{k=1,\ldots,\underline{n}^2-1}$ according to (19) and (26) by $\underline{\theta}^k$ the action of the maps $\beta^{(p)}$ can be described alternatively by

$$\begin{aligned} \beta^{(p)}(\omega) &\equiv \beta^{(p)}\left(\sum_{(A_1,\ldots,A_p)\in\mathcal{J}_p^{n^2+m^2-1|2nm}} \omega_{A_1,\ldots,A_p}\wedge\theta^{A_1}\wedge\cdots\wedge\theta^{A_p}\right) \\ &= \sum_{(k_1,\ldots,k_p)\in\mathcal{J}_p^{\underline{n}^2-1|0}} \beta(\omega_{k_1,\ldots,k_p})\wedge\underline{\theta}^{k_1}\wedge\cdots\wedge\underline{\theta}^{k_p}. \end{aligned} \tag{42}$$

If we set $\beta^{(0)}\equiv\beta$ the maps $\beta^{(p)}$, $p\in\mathbb{N}_0$, extend uniquely to a bihomogeneous, $\mathbb{C}$-linear map $\Omega^g(\mathbb{M}(n|m))\to\Omega(\mathbb{M}(\underline{n}))$ of bidegree $(0,\bar{0})$, which we again denote by β. Because of (42) β is onto and its restriction to $\Omega^g_{\mathcal{Z}^g}(\mathbb{M}(n|m))$ is a surjective homomorphism of bigraded $\mathbb{C}$-algebras onto $\Omega_{\mathcal{Z}}(\mathbb{M}(\underline{n}))$. Furthermore β fulfills

$$\beta\circ L_D = L_{\hat{\beta}(D)}\circ\beta \tag{43}$$

for all $D\in\mathfrak{Der}^g(\mathbb{M}(n|m))_{\bar{0}}$ as well as

$$\beta\circ d = d\circ\beta. \tag{44}$$

Consequently β induces a homomorphism $H(\beta):H(\mathbb{M}(n|m))\to H(\mathbb{M}(\underline{n}))$ of cohomologies in the usual way. Analogous to graded manifold theory[9] this map is an isomorphism.

Proposition 6: $H(\beta)$ is an isomorphism of bigraded $\mathbb{C}$-algebras, such that both cohomologies $H(\mathbb{M}(n|m))$ *and* $H(\mathbb{M}(n))$ *are isomorphic to the Lie algebra cohomology* $H(\mathfrak{sl}(n);\mathbb{C})$ *of* $\mathfrak{sl}(n)$ *with trivial coefficients.*

Proof: Using the results of Ref. 23 as well as $\mathfrak{Der}^g(\mathbb{M}(n|m))=\mathfrak{sl}(n|m)^{\mathrm{ad}}$ we find the sequence

$$\begin{aligned}H(\mathbb{M}(n|m))&\cong H(\mathfrak{sl}(n|m);\mathbb{M}(n|m))\\&\cong H(\mathfrak{sl}(n|m);\mathbb{C}1_{n+m})\oplus H(\mathfrak{sl}(n|m);\mathfrak{sl}(n|m))\\&\cong H(\mathfrak{sl}(n|m);\mathbb{C}1_{n+m})\cong H(\mathfrak{sl}(n|m;\mathbb{C})\end{aligned}$$

of natural isomorphisms between Lie superalgebra cohomologies. In particular we have $H(\mathbb{M}(n))\cong H(\mathfrak{sl}(\underline{n});\mathbb{C})$ (as $\mathbb{N}_0$-graded $\mathbb{C}$-algebra), which is well known from matrix geometry.[10–12] Combining the above-mentioned result with the calculations of the cohomology of $\mathfrak{sl}(n|m)$ with trivial coefficients[22,27] one can conclude that $H(\beta)$ is an isomorphism of bigraded $\mathbb{C}$-algebras.

V. NONCOMMUTATIVE GRADED SYMPLECTIC GEOMETRY

Generalizing the situation on graded manifolds[9] as well as the one of ordinary matrix algebras[10,11,18,19] we call an even, closed graded two-form $\omega\in\Omega^{g,2}(\mathbb{M}(n|m))$ graded symplectic structure on $\mathbb{M}(n|m)$, if the equation

$$\omega(D,D_M)=DM \tag{45}$$

for all $D\in\mathfrak{Der}^g(\mathbb{M}(n|m))$ possesses a unique solution $D_M\in\mathfrak{Der}^g(\mathbb{M}(n|m))$ for each $M\in\mathbb{M}(n|m)$. The graded vector fields $D_M\in\mathfrak{Der}^g(\mathbb{M}(n|m))$ are called Hamiltonian and the set of all graded Hamiltonian vector fields is denoted by $\mathfrak{Ham}^g(\omega)$.

If $\omega\in\Omega^{g,2}(\mathbb{M}(n|m))$ is a graded symplectic structure on $\mathbb{M}(n|m)$ the assignment $M\mapsto D_M$ defines an even $\mathbb{C}$-linear map $D^\omega:\mathbb{M}(n|m)\to\mathfrak{Ham}^g(\omega)\subseteq\mathfrak{Der}^g(\mathbb{M}(n|m))$ and one can conclude that (45) is equivalent to

$$\iota_{D_M}\omega+dM=0. \tag{46}$$

Using (13) we find

$$L_{D_M}\omega=0 \tag{47}$$

for all $D_M\in\mathfrak{Ham}^g(\omega)$, that is a graded symplectic structure on $\mathbb{M}(n|m)$ is—as usual—invariant with respect to all graded Hamiltonian vector fields.

Via

$$\{M,M'\}_g^\omega:=\omega(D_M,D_{M'}) \tag{48}$$

for all $M,M'\in\mathbb{M}(n|m)$ we can introduce a graded Poisson bracket, which has the analogous properties as its graded-commutative pendant.

Proposition 7: $(\mathbb{M}(n|m),\{\cdot,\cdot\}_g^\omega)$ *is a* $\mathbb{C}$*-Lie superalgebra and the graded Poisson bracket fulfills in addition*

$$\begin{aligned}\{M,M'M''\}_g^\omega&=\{M,M'\}_g^\omega M''+(-1)^{\overline{MM'}}M'\{M,M''\}_g^\omega,\\\{1_{n+m},M\}_g^\omega&=0\end{aligned} \tag{49}$$

for all homogeneous $M,M',M''\in\mathbb{M}(n|m)$. *Moreover, the map* $D^\omega:\mathbb{M}(n|m)\to\mathfrak{Ham}^g(\omega)$ *is a homomorphism of Lie superalgebras and*

$$\mathfrak{Ham}^g(\omega) = \mathfrak{Der}^g(\mathbb{M}(n|m)). \tag{50}$$

Proof: The properties of $\{\cdot,\cdot\}_g^\omega$ and of D^ω result from the defining properties of the graded symplectic structure ω. From the irreducibility of the adjoint representation of $\mathfrak{sl}(n|m)$ one can deduce $\ker D^\omega = \mathbb{C}1_{n+m}$ on the one hand and the injectivity of $D^\omega|_{\mathfrak{sl}(n|m)}$ on the other hand. Then (50) follows because of $\mathfrak{Der}^g(\mathbb{M}(n|m)) = \mathfrak{sl}(n|m)^{\mathrm{ad}}$. □

There exists an essentially unique graded symplectic structure on $\mathbb{M}(n|m)$.

Proposition 8: $d\Theta$ is a graded symplectic structure on $\mathbb{M}(n|m)$ and up to complex multiples it is the only one. The corresponding graded Poisson bracket is given by

$$\{M,M'\}_g^{d\Theta} = [M,M']_g \tag{51}$$

for all $M,M' \in \mathbb{M}(n|m)$.

Proof: The exact, even graded two-form $c\,d\Theta, c \in \mathbb{C}\setminus\{0\}$ induces via (45) a homomorphism $D^{c\,d\Theta}: \mathbb{M}(n|m) \to \mathfrak{Der}^g(\mathbb{M}(n|m))$,

$$D^{c\,d\Theta}(M) = \frac{1}{c}\,\mathrm{ad}\,M \tag{52}$$

of Lie superalgebras and the corresponding graded Poisson bracket is given by $\{M,M'\}_g^{c\,d\Theta} = c^{-1}[M,M']_g$. The uniqueness property is a consequence of proposition 3, (50), and Schur's Lemma. □

Consequently the extension of the body map β maps a graded symplectic structure ω onto a symplectic structure $\beta(\omega)$. Moreover one has

$$\hat{\beta}(D^\omega(M)) = D^{\beta(\omega)}(\beta(M)) \tag{53}$$

for all even graded (Hamiltonian) vector fields as well as

$$\beta(\{M,M'\}_g^\omega) = \{\beta(M),\beta(M')\}^{\beta(\omega)} \tag{54}$$

for the graded Poisson bracket of $M,M' \in \mathbb{M}(n|m)_{\bar{0}}$. That is, the relation between $\mathbb{M}(n|m)$ and its body is analogous to the one for graded symplectic manifolds and their respective underlying manifolds.

VI. GRADED VECTOR BUNDLES OVER GRADED MATRIX ALGEBRAS

As a synthesis of the definition of graded vector bundles over graded manifolds[2,9,20] and the idea how to introduce vector bundles in noncommutative geometry[1,21] we interpret left, $\mathbb{Z}_2$-graded, finitely generated (graded-projective) $\mathbb{M}(n|m)$ modules as graded vector bundles over $\mathbb{M}(n|m)$ and even $\mathbb{M}(n|m)$-module homomorphisms between such modules as graded vector bundle homomorphisms. We note that the specifying property of graded projectivity is redundant in the context of left, $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$-modules, because on the one hand graded-projective means $\mathbb{Z}_2$-graded plus projective[28] and on the other hand every left $\mathbb{M}(n+m)$-module is projective.[24]

Let us denote by $\mathbb{M}(n|m,r|s), r,s \in \mathbb{N}_0, r+s \in \mathbb{N}$, the $\mathbb{C}$-vector space $\mathbb{M}(n+m,r+s)$ together with the $\mathbb{Z}_2$-grading defined by

$$\begin{aligned}\mathbb{M}(n|m,r|s)_{\bar{0}} &:= \left\{ v = \begin{pmatrix} v_1 & 0 \\ 0 & v_4 \end{pmatrix} \middle| v_1 \in \mathbb{M}(n,r), v_4 \in \mathbb{M}(m,s) \right\},\\ \mathbb{M}(n|m,r|s)_{\bar{1}} &:= \left\{ v = \begin{pmatrix} 0 & v_2 \\ v_3 & 0 \end{pmatrix} \middle| v_2 \in \mathbb{M}(n,s), v_3 \in \mathbb{M}(m,r) \right\}.\end{aligned} \tag{55}$$

With respect to ordinary matrix multiplication $\mathbb{M}(n|m,r|s)$ becomes a left, $\mathbb{Z}_2$-graded, finitely generated $\mathbb{M}(n|m)$-module and these examples constitute essentially all graded vector bundles over $\mathbb{M}(n|m)$.

Proposition 9: If $\mathcal{V}$ is a graded vector bundle over $\mathbb{M}(n|m)$ then there exist unique numbers $r,s\in\mathbb{N}_0, r+s\in\mathbb{N}$ and a graded vector bundle isomorphism $\phi:\mathcal{V}\to\mathbb{M}(n|m,r|s)$. $\mathcal{V}$ is graded-free if and only if there are natural numbers $p,q\in\mathbb{N}_0, p+q\in\mathbb{N}$, such that

$$pn+qm=r,\quad pm+qn=s. \tag{56}$$

Proof: The existence of the isomorphisms are implied by the graded simplicity of $\mathbb{M}(n|m,1|0)$ and $\mathbb{M}(n|m,0|1)$ and the fact, that every left, $\mathbb{Z}_2$-graded, finitely generated $\mathbb{M}(n|m)$ module is the homomorphic image of a left, $\mathbb{Z}_2$-graded, graded-free $\mathbb{M}(n|m)$-module with homogeneous basis of suitable cardinality $p|q$. Because all $\mathbb{M}(n|m)$-module isomorphisms are $\mathbb{C}$-vector space isomorphisms in particular, the uniqueness statement and (56) follow. □

After this ''miniature-classification'' we develop graded differential geometry on a fixed graded vector bundle $\mathcal{V}$ generalizing the treatment of noncommutative geometry[1,10,13,21] on the one hand and the one of supergeometry[2] on the other hand.

So we first define the set $\Omega^g(\mathcal{V})$ of $\mathcal{V}$-valued graded forms according to

$$\Omega^g(\mathcal{V})\equiv\bigoplus_{p\in\mathbb{N}_0}\Omega^{g,p}(\mathcal{V}):=\Omega^g(\mathbb{M}(n|m))\hat{\otimes}_{M(n|m)}\mathcal{V}. \tag{57}$$

$\Omega^g(\mathcal{V})$ is a left $\mathbb{N}_0\times\mathbb{Z}_2$-bigraded $\Omega^g(\mathbb{M}(n|m))$-module in a natural way and each $\Omega^{g,p}(\mathcal{V}),p\in\mathbb{N}_0$, is a left, $\mathbb{Z}_2$-graded, finitely generated $\mathbb{M}(n|m)$ module, in particular. The product will again be denoted by $\wedge$.

A connection on $\mathcal{V}$ is an even $\mathbb{C}$-linear map $\nabla:\mathcal{V}\to\Omega^{g,1}(\mathcal{V})$ such that

$$\nabla(Mv)=dM\otimes v+M\wedge\nabla v \tag{58}$$

is fulfilled for all $M\in\mathbb{M}(n|m), v\in\mathcal{V}$. Connections always exist due to (graded) projectivity.

Proposition 10: Let $\mathcal{V}$ be a graded vector bundle over $\mathbb{M}(n|m)$. Then there exists a graded-free vector bundle $\mathcal{V}^{p|q}$ over $\mathbb{M}(n|m)$ with homogeneous basis $\{\epsilon_A|\epsilon_A\in\mathcal{V}^{p|q}_0,\ A=1,...,p;\epsilon_A\in\mathcal{V}^{p|q}_1,\ A=p+1,...,p+q\}$, $p,q\in\mathbb{N}_0, p+q\in\mathbb{N}$, together with an even, idempotent endomorphism $P:\mathcal{V}^{p|q}\to\mathcal{V}^{p|q}$ and an isomorphism $\varphi:\mathcal{V}\to\operatorname{im}P$ of $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$ modules. The map $\nabla_d:\mathcal{V}^{p|q}\to\Omega^{g,1}(\mathcal{V}^{p|q})$ defined by

$$\nabla_d(v)\equiv\nabla_d\left(\sum_{A=1}^{p+q}v^A\epsilon_A\right):=\sum_{A=1}^{p+q}dv^A\otimes\epsilon_A \tag{59}$$

is a connection on $\mathcal{V}^{p|q}$ and

$$\nabla_{Pd}:=\mathrm{Id}_{\Omega^{g,1}(\mathbb{M}(n|m))}\otimes\varphi^{-1}\circ\mathrm{Id}_{\Omega^{g,1}(\mathbb{M}(n|m))}\otimes P\circ\nabla_d\circ\varphi \tag{60}$$

is a connection on $\mathcal{V}$. A map $\nabla:\mathcal{V}\to\Omega^{g,1}(\mathcal{V})$ is a connection on $\mathcal{V}$ if and only if it is of the form

$$\nabla=\nabla_{Pd}+\alpha, \tag{61}$$

where $\alpha:\mathcal{V}\to\Omega^{g,1}(\mathcal{V})$ is an even homomorphism of $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$ modules.

Proof: Analogous to the ungraded case.[1,21] □

We note that the existence of connections on $\mathcal{V}$ can also be shown without using (graded) projectivity (see Refs. 10, 11, 13, 15 for the ungraded case): According to (33),

$$\nabla_\Theta(v):=\Theta\otimes v \tag{62}$$

for all $v\in\mathcal{V}$ defines a connection on $\mathcal{V}$.

Quite generally we will denote the $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$-bimodule of graded $(p+q)\times(p+q)$-matrices over a $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$-bimodule $\mathcal{B}$ with $\mathbb{M}(p|q;\mathcal{B})$. It is an $\mathbb{M}(n|m)$ bimodule in a natural way and $\mathbb{Z}_2$-graded by declaring those matrices with even diagonal entries and odd off-diagonal entries as even and those with odd diagonal entries and even off-diagonal entries as odd. Adopting the notation of the above proposition we introduce homogeneous generators

$$\eta_A := \varphi^{-1}\circ P(\epsilon_A) \tag{63}$$

of $\mathcal{V}$ as well as an even matrix $(P_A^B)\in\mathbb{M}(p|q;\mathbb{M}(n|m))_{\bar{0}}$ via

$$P(\epsilon_A) =: \sum_{B=1}^{p+q} P_A^B \epsilon_B . \tag{64}$$

Then

$$(\nabla-\nabla_{Pd})(\eta_A)=\alpha(\eta_A)=:\sum_{B=1}^{p+q} \alpha_A^B\otimes\eta_B . \tag{65}$$

establishes a bijective correspondence between the set of all connections on $\mathcal{V}$ and the set $P\mathbb{M}(p|q;\Omega^{g,1}(\mathbb{M}(n|m)))_{\bar{0}}P$, which consists of those $(\alpha_A^B)\in\mathbb{M}(p|q;\Omega^{g,1}(\mathbb{M}(n|m)))_{\bar{0}}$ fulfilling $\alpha_A^B=\sum_{C,D=1}^{p+q} P_A^C\wedge\alpha_C^D\wedge P_D^B$ (for $\mathcal{V}=\mathcal{V}^{p|q}$ set $P=\varphi=\mathrm{Id}_{\mathcal{V}^{p|q}}$). The graded one-forms α_A^B are called connection forms of the connection $\nabla=\nabla_{Pd}+\alpha$.

If ∇ is a connection on a graded vector bundle $\mathcal{V}$ we can introduce a $\mathbb{C}$-linear map $\Omega^g(\mathcal{V})\to\Omega^g(\mathcal{V})$, again denoted by ∇, via

$$\nabla(\omega\otimes v)=d\omega\otimes v+(-1)^p\omega\wedge\nabla v \tag{66}$$

for all $v\in\mathcal{V},\omega\in\Omega^{g,p}(\mathbb{M}(n|m)),p\in\mathbb{N}_0$. This homogeneous map of bidegree $(1,\bar{0})$ extends the original connection if we identify $\mathcal{V}$ with $\Omega^{g,0}(\mathcal{V})$. Moreover it fulfills

$$\nabla(\omega'\wedge\omega\otimes v)=d\omega'\wedge(\omega\otimes v)+(-1)^{p'}\omega'\wedge\nabla(\omega\otimes v) \tag{67}$$

for all $v\in\mathcal{V},\omega\in\Omega^{g,p}(\mathbb{M}(n|m)),\omega'\in\Omega^{g,p'}(\mathbb{M}(n|m)),p,p'\in\mathbb{N}_0$, and this property determines the extension of the connection uniquely.

The curvature of a connection ∇ on a graded vector bundle $\mathcal{V}$ is defined as

$$\nabla^2\equiv\nabla\circ\nabla:\mathcal{V}\to\Omega^{g,2}(\mathcal{V}). \tag{68}$$

It is an even homomorphism of $\mathbb{Z}_2$-graded $\mathbb{M}(n|m)$ modules and one can describe its action on an element $v=\sum_{A=1}^{p+q}\varphi(v)^A\eta_A\in\mathcal{V}$ according to

$$\nabla^2(v)=:\sum_{A,B=1}^{p+q}\varphi(v)^A\wedge R_A^B\otimes\eta_B \tag{69}$$

with a uniquely determined matrix $(R_A^B)\in P\mathbb{M}(p|q;\Omega^{g,2}(\mathbb{M}(n|m)))_{\bar{0}}P$. The graded two-forms R_A^B are called curvature forms and they can be expressed according to

$$R_A^B=-\sum_{C=1}^{p+q}\alpha_A^C\wedge\alpha_C^B+\sum_{C,D=1}^{p+q}(P_A^C\wedge d\alpha_C^D\wedge P_D^B-P_A^C\wedge dP_C^D\wedge dP_D^B) \tag{70}$$

in terms of the connection forms α_A^B of the connection. Moreover they have to fulfill the Bianchi identity

$$\sum_{C,D=1}^{p+q} P_A^C \wedge dR_C^D \wedge P_D^B - \sum_{C=1}^{p+q} (\alpha_A^C \wedge R_C^B - R_A^C \wedge \alpha_C^B) = 0. \tag{71}$$

Let us finally analyze the space of flat connections, that is the set of all connections with vanishing curvature. We will not do this in complete generality but only for a graded-free vector bundle $\mathcal{V}^{1|0}$ with an even basis element ϵ.

Proposition 11: A connection on $\mathcal{V}^{1|0}$ *is flat if and only if its connection form* $\alpha \in \Omega^{g,1}(\mathbb{M}(n|m))_0^-$ *is either given by*

$$\alpha = \Theta \tag{72}$$

or by

$$\alpha = \Theta - \sum_{A=1}^{(n+m)^2-1} f(E_A) \wedge \theta^A, \tag{73}$$

where $\{E_A\}$ *is the homogeneous basis of* $\mathfrak{sl}(n|m)$ *"corresponding" to* $\{\theta^A\}$ *and f is some automorphism of* $\mathfrak{sl}(n|m)$.

Proof: Let us introduce an even graded one-form $\rho = \Sigma_{A=1}^{(n+m)^2-1} \rho_A \wedge \theta^A$ according to $\alpha =: \Theta - \rho$. Using proposition 4 we find that the curvature form is given by

$$R = \frac{1}{2} \sum_{A,B=1}^{(n+m)^2-1} \Omega_{AB} \wedge \theta^A \wedge \theta^B \tag{74}$$

with

$$\Omega_{AB} = [\rho_b, \rho_A]_g - \sum_{C=1}^{(n+m)^2-1} C_{BA}^C \rho_C. \tag{75}$$

Because the vanishing of the curvature is equivalent to $\Omega_{AB} = 0$ the proposition follows from the simplicity of $\mathfrak{sl}(n|m)$. □

That is, we have the same situation as in ordinary matrix geometry:[10,11,13,15] There exist different "classes" of flat connections. Here "class" refers to the action of the group of automorphisms of the graded vector bundle on the space of connections, which can be introduced similar to the ungraded case. The connection ∇_d and the one associated with the connection form Θ will lie in different classes, because the latter is invariant. However, if one does not restrict the space of connections by a suitable compatibility requirement with respect to a graded Hermitian structure there will exist even more than two classes of flat connections.

VII. CONCLUDING REMARKS

We have developed the graded differential geometry of graded matrix algebras and shown that the results of matrix geometry[10,11,13,15] carry over to the $\mathbb{Z}_2$-graded setting. In addition we found a natural noncommutative analog of the body map, which allows us to view graded matrix geometries as true noncommutative generalizations of graded manifolds.

Whereas in ordinary differential geometry one integrates forms this is not true in supergeometry. Except from the before mentioned body map, which plays a central role in the global theory of Berezin integration,[29] we completely excluded the integral geometry of graded matrix algebras. We plan to treat this together with metric aspects in a separate work.

Beside its immediate application for the construction of (graded) differential calculi on fuzzy (super) manifolds[15,17] the developments of this article offer another perspective. The extension of space–time by matrix geometries led to interesting new gauge models. In particular the existence of different gauge orbits of flat connections in matrix geometry is the origin of the appearance of the Higgs effect.[10,13,15,30] The possibility of extending the structures of matrix geometry to $\mathbb{Z}_2$-graded matrix algebras suggests thinking about similar ''supersymmetric'' noncommutative extensions of space–time.

ACKNOWLEDGMENTS

The research of G.R. was supported by the ''Fonds zur Förderung der wissenschaftlichen Forschung (FWF)'' through Research Project No. P11783-PHY ''Quantum Field Theory on Noncommutative Manifolds.'' The authors would like to thank W. Bulla, J. Madore, P. Prešnajder, and L. Pittner for helpful discussions, the ''Fonds zur Förderung der wissenschaftlichen Forschung (FWF)'' for support funding, and the referee for useful comments.

APPENDIX: ASSOCIATIVE PRODUCT OF SUPERTRACE-FREE, GRADED MATRICES

Let $\{E_A | E_A \in \mathfrak{sl}(n|m)_0^-, A=1,...,n^2+m^2-1; E_A \in \mathfrak{sl}(n|m)_1^-, A=n^2+m^2,\ldots,(n+m)^2-1\}$ be a homogeneous basis of $\mathfrak{sl}(n|m), n\neq m$. The aim of this Appendix is to investigate the associative product of the homogeneous matrices E_A in a similar way as was done in Ref. 31 for trace-free, Hermitian matrices.

If we introduce a graded anticommutator

$$[M,M']_g^+ := MM' + (-1)^{\overline{MM'}} M'M \tag{A1}$$

of two homogeneous $M,M' \in \mathbb{M}(n|m)$ we find the relations

$$\begin{aligned} &[M,[M',M'']_g]_g - [M,[M',M'']_g^+]_g^+ + (-1)^{\bar{M}'\bar{M}''}[[M,M'']_g^+, M']_g^+ = 0,\\ &[[M,M']_g^+, M'']_g - [M,[M',M'']_g]_g^+ - (-1)^{\bar{M}'\bar{M}''}[[M,M'']_g, M']_g^+ = 0 \end{aligned} \tag{A2}$$

between the graded commutator and the graded anticommutator of homogeneous $M,M',M'' \in \mathbb{M}(n|m)$.

Because $\{E_A, 1_{n+m}\}_{A=1,\ldots,(n+m)^2-1}$ forms a homogeneous basis of $\mathbb{M}(n|m)$ the graded anticommutator of E_A and E_B can be written according to

$$[E_A, E_B]_g^+ = \sum_{C=1}^{(n+m)^2-1} d_{AB}^C E_C + g_{AB} 1_{n+m} \tag{A3}$$

with uniquely determined coefficients $d_{AB}^C, g_{AB} \in \mathbb{C}$. Then the associative product of E_A and E_B is given by

$$E_A E_B = \frac{1}{2} \sum_{C=1}^{(n+m)^2-1} (c_{AB}^C + d_{AB}^C) E_C + \frac{1}{2} g_{AB} 1_{n+m}. \tag{A4}$$

Independent of the specific choice of the homogeneous basis $\{E_A\}_{A=1,\ldots,(n+m)^2-1}$ there exist a lot of relations between the ''structure constants'' c_{AB}^C, d_{AB}^C, and g_{AB} which we summarize in

Proposition A: (i) c_{AB}^C *and* d_{AB}^C *vanish if* $\bar{E}_A + \bar{E}_B + \bar{E}_C = \bar{1}$ *and* g_{AB} *vanishes if* $\bar{E}_A + \bar{E}_B = \bar{1}$.

$$\text{(ii)} \qquad \sum_{B=1}^{(n+m)^2-1} (-1)^{\bar{E}_B} c_{AB}^B = 0, \qquad \sum_{B=1}^{(n+m)^2-1} (-1)^{\bar{E}_B} d_{AB}^B = 0. \tag{A5}$$

(iii) c_{ABC} *and* d_{ABC}, *defined via*

$$c_{ABC} := \sum_{D=1}^{(n+m)^2-1} c_{AB}^D g_{DC}, \quad d_{ABC} := \sum_{D=1}^{(n+m)^2-1} d_{AB}^D g_{DC}, \tag{A6}$$

are totally antisymmetric, respectively, totally symmetric in the $\mathbb{Z}_2$-graded sense.

(iv) $$\sum_{E=1}^{(n+m)^2-1} \{(-1)^{\bar{E}_A\bar{E}_C}C^E_{BC}C^D_{AE}+(-1)^{\bar{E}_B\bar{E}_A}c^E_{CA}c^D_{BE}+(-1)^{\bar{E}_C\bar{E}_B}c^E_{AB}c^D_{CE}\}=0,$$

$$\sum_{E=1}^{(n+m)^2-1} \{c^E_{BC}c^D_{AE}-d^E_{AB}d^D_{EC}+(-1)^{\bar{E}_A\bar{E}_C+\bar{E}_B\bar{E}_C}d^E_{CA}d^D_{EB}\}$$
$$+2(-1)^{\bar{E}_A\bar{E}_C+\bar{E}_B\bar{E}_C}g_{CA}\delta^D_B-2g_{AB}\delta^D_C=0, \tag{A7}$$

$$\sum_{E=1}^{(n+m)^2-1} \{d^E_{AB}c^D_{EC}-c^E_{BC}d^D_{AE}-(-1)^{\bar{E}_B\bar{E}_C}c^E_{AC}d^D_{EB}\}=0.$$

(v) *If $K_{AB}:=K(E_A,E_B)$, where K is the Killing form of $\mathfrak{sl}(n|m)$, then*

$$K_{AB}=(n-m)^2g_{AB}=\sum_{C,D=1}^{(n+m)^2-1}(-1)^{\bar{E}_C}c^C_{AD}c^D_{BC}=\frac{(n-m)^2}{(n-m)^2-4}\sum_{C,D=1}^{(n+m)^2-1}(-1)^{\bar{E}_C}d^C_{AD}d^D_{BC}. \tag{A8}$$

(vi) *Denoting by g^{AB} the components of the matrix inverse to (g_{AB}), then*

$$\sum_{B,C=1}^{(n+m)^2-1} g^{BC}c^A_{BC}=0, \qquad \sum_{B,C=1}^{(n+m)^2-1} g^{BC}d^A_{BC}=0. \tag{A9}$$

(vii) $$\sum_{C,D,E=1}^{(n+m)^2-1} g^{CD}c^A_{CE}c^E_{DB}=(n-m)^2\delta^A_B,$$

$$\sum_{C,D,E=1}^{(n+m)^2-1} g^{CD}c^A_{CE}d^E_{DB}=0, \tag{A10}$$

$$\sum_{C,D,E=1}^{(n+m)^2-1} g^{CD}d^A_{CE}d^E_{DB}=((n-m)^2-4)\delta^A_B.$$

(viii) $$\sum_{D,E,F,G=1}^{(n+m)^2-1}(-1)^{\bar{E}_A\bar{E}_E}g^{DE}c^F_{EB}c^G_{AF}c^C_{DG}=\frac{1}{2}(n-m)^2c^C_{AB},$$

$$\sum_{D,E,F,G=1}^{(n+m)^2-1}(-1)^{\bar{E}_A\bar{E}_E}g^{DE}c^F_{EB}c^G_{AF}c^C_{DG}=-\frac{1}{2}(n-m)^2c^C_{AB},$$

$$\sum_{D,E,F,G=1}^{(n+m)^2-1}(-1)^{\bar{E}_A\bar{E}_E}g^{DE}c^F_{EB}d^G_{AF}d^C_{DG}=-\frac{1}{2}((n-m)^2-4)c^C_{AB}, \tag{A11}$$

$$\sum_{D,E,F,G=1}^{(n+m)^2-1}(-1)^{\bar{E}_A\bar{E}_E}g^{DE}d^F_{EB}d^G_{AF}d^C_{DG}=\frac{1}{2}((n-m)^2-12)d^C_{AB}.$$

Proof: (i) is a reformulation of the homogeneity of $\{E_A\}$. The first line of (A8) as well as (iii) result from $K_{AB}=2(n-m)\mathrm{Tr}_s(E_AE_B)$. (ii) is a consequence of $\mathrm{Tr}_s(\mathrm{ad}\,E_A)=0$ and of

$\Sigma_{B,C}\,\mathrm{Tr}_s(g^{BC}[E_B,E_C]_g^+E_A)=0$. (iv) is a reformulation of the graded Jacobi identity and (A2). Using the second equation (A7) one deduces the second line of (A8). (vi) follows from (ii) and (iii). The left-hand side of the first equation (A10) is essentially the second-order Casimir operator of $\mathfrak{sl}(n|m)$ in the adjoint representation. The second part of (A10) follows from (iii), whereas the third equation is a consequence of the first part together with (iv) and (vi). The relations (viii) are results of calculations using (iii), (iv), (vi), and (vii). □

[1] A. Connes, *Noncommutative Geometry* (Academic, New York, 1994).
[2] C. Bartocci, U. Bruzzo, and D. Hernández Ruipérez, *The Geometry of Supermanifolds* (Kluwer, Dordrecht, 1991).
[3] D. Kastler, *Cyclic Cohomology Within the Differential Envelope* (Hermann, Paris, 1988).
[4] L. Pittner, *Algebraic Foundations of Non-Commutative Differential Geometry and Quantum Groups* (Springer, New York, 1996).
[5] J. Fröhlich, O. Grandjean, and A. Recknagel, Commun. Math. Phys. **203**, 119 (1999).
[6] W. Kalau and M. Walze, J. Geom. Phys. **22**, 77 (1997).
[7] M. Dubois-Violette, R. Kerner, and J. Madore, Class. Quantum Grav. **8**, 1077 (1991).
[8] R. Kerner, J. Geom. Phys. **11**, 325 (1993).
[9] B. Kostant, Lect. Notes Math. **570**, 177 (1977).
[10] A. E. F. Djemai, Int. J. Theor. Phys. **34**, 801 (1995).
[11] M. Dubois-Violette, R. Kerner, and J. Madore, J. Math. Phys. **31**, 316 (1990).
[12] M. Dubois-Violette, C. R. Acad. Sci., Ser. I: Math. **307**, 403 (1988).
[13] T. Masson, thesis, Orsay, 1995.
[14] A. Jadczyk and D. Kastler, Rep. Math. Phys. **25**, 1 (1987).
[15] J. Madore, *An Introduction to Noncommutative Differential Geometry and its Physical Applications* (Cambridge University Press, Cambridge, 1995).
[16] H. Grosse, C. Klimčik, and P. Prešnajder, Commun. Math. Phys. **185**, 155 (1997).
[17] H. Grosse and G. Reiter, J. Geom. Phys. **28**, 349 (1998).
[18] M. Dubois-Violette, in *Differential Geometric Methods in Theoretical Physics*, edited by C. Bartocci, U. Bruzzo, and R. Cianci (Springer, New York, 1991).
[19] M. Dubois-Violette, in *Geometry and Nature*, Contemporary Mathematics Vol. 203, edited by H. Nencka and J. P. Bourguignon (American Mathematical Society, Providence, 1997).
[20] D. Hernández Ruipérez and J. Muñoz Masqué, J. Math. Pures Appl. **63**, 283 (1984).
[21] J. M. Gracia-Bondia and J. C. Várilly, J. Geom. Phys. **12**, 223 (1993).
[22] D. B. Fuks, *Cohomology of Infinite-Dimensional Lie Algebras* (Consultants Bureau, New York, 1986).
[23] M. Scheunert and R. B. Zhang, J. Math. Phys. **39**, 5024 (1998).
[24] R. S. Pierce, *Associative Algebras* (Springer, New York, 1982).
[25] J. F. Cornwell, *Group Theory in Physics*, Supersymmetries and Infinite-Dimensional Algebras, Vol. III (Academic, New York, 1989).
[26] R. Coquereaux and D. Kastler, Pac. J. Math. **137**, 245 (1989).
[27] D. B. Fuks and D. A. Leites, C. R. Acad. Bulg. Sci. **37**, 1595 (1984).
[28] C. Năstăsecu and F. van Oystaeyen, *Graded Ring Theory* (North-Holland, Amsterdam, 1982).
[29] D. Hernández Ruipérez and J. Muñoz Masqué, C. R. Acad. Sci., Ser. I: Math. **301**, 915 (1985).
[30] M. Dubois-Violette, R. Kerner, and J. Madore, J. Math. Phys. **31**, 323 (1990).
[31] A. J. Macfarlane, A. Sudbery, and P. H. Weisz, Commun. Math. Phys. **11**, 77 (1968).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Even moments of a polymer chain distribution

Gary G. Hoffman[a)]
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 25 February 1999; accepted for publication 17 September 1999)

An algorithm is presented for the evaluation of the even moments of the end-to-end separation of a polymer chain. The only assumption made in the distribution function is that the energy of the chain depends only on the individual bond lengths of the chain. This assumption covers many of the commonly used distribution functions. © *1999 American Institute of Physics.* [S0022-2488(99)02512-8]

I. INTRODUCTION

Recent investigations[1] into the stress–strain behavior of polymer network solids have led to an expression for the stress tensor in the form of an expansion in moments of the distribution of end-to-end separations of a polymer chain. Practical implementation of the expression requires, specifically, the even moments for chain lengths ranging from one on up to arbitrary numbers of bonds. It was found that moments up to at least the eighth were needed to obtain a reasonable description of the stress–strain behavior. It was found, further, that the non-Gaussian nature of the distribution was a critical ingredient in the computations and that the contributions of the shorter chains were found to be important, especially for the larger strains. It was therefore desired to have a model distribution for polymer chain end-to-end separations that reasonably describes the molecular nature of the material and expressions for the associated moments that are accurate for all chain lengths. For sufficiently long chain lengths, the probability distribution is well approximated by a Gaussian chain distribution. In such a case, the moments could be profitably obtained from such a distribution or a correction to it. However, a Gaussian distribution would not be appropriate for the shorter chain lengths. Since the stress–strain studies involved chain lengths as short as one bond, more accurate expressions were needed. In this paper, an algorithm for deriving exact expressions for the moments is presented and specific expressions for the lower order moments are given.

While the moments can be evaluated by brute force, each order being considered in turn, the procedure becomes increasingly difficult for higher orders. In trying to recognize patterns and simplifications, a general procedure was developed that can be applied to arbitrary orders. With this procedure, closed form expressions were readily generated for the moments up to the twelfth order. It was further noted that the procedure is applicable to a more general class of distribution than the one originally considered. In the event that the evaluation of higher moments of such distributions is desired, the general procedure will be presented in this paper.

This procedure uses a diagrammatic approach and has some similarities to the treatment used for the cluster expansion of the two-body distribution function in the classical statistical mechanics of many-body systems[2] or the quantum mechanical treatment of many-fermion systems.[3] Although the spirit of the development is the same, however, the details are different.

II. MATHEMATICAL FORMULATION

Consider a polymer chain that contains N equivalent bonds. The direction and length of the jth bond is given by the bond vector, $\mathbf{Q}_j$, and the full set of N such vectors is represented by $\mathbf{Q}^N$. It is assumed that the energy of the polymer chain depends only on the lengths of the bonds. Mathematically, the energy of a polymer chain in a particular configuration is then given by

[a)]Permanent address: Department of Chemistry, Florida International University, Miami, FL 33199.

$$E(\mathbf{Q}^N)=E_0+\sum_{j=1}^{N} f(Q_j), \tag{1}$$

where E_0 is a reference energy associated with the relaxed chain, and $f(Q)$ is an arbitrary function of the vector magnitude, Q. Recent work[4] has focused specifically on a chain of harmonic springs, where $f(Q)=\frac{1}{2}H(Q-a)^2$. The equilibrium probability of end-to-end separations of the chain at absolute temperature, T, is then given by

$$P_{\mathrm{eq}}^{(N)}(r)=\frac{1}{Z_1^N}\int \delta\left(\mathbf{r}-\sum_{j=1}^{N}\mathbf{Q}_j\right)\exp\left[-\sum_{n=1}^{N} f(Q_n)/kT\right]d^3Q_1\cdots d^3Q_N, \tag{2}$$

where k is Boltzmann's constant and

$$Z_1=\int e^{-f(Q)/kT}d^3Q. \tag{3}$$

It is easily shown that the distribution (2) depends only on the magnitude of the separation and not its direction.

The objective of this paper is to evaluate the even moments of this distribution:

$$\langle r^{2j}\rangle_0\equiv\int r^{2j}P_{\mathrm{eq}}^{(N)}(r)d^3r. \tag{4}$$

The subscript ''0'' denotes the use of the equilibrium distribution, as opposed to the distribution for a system under a deformation. In the course of the development, averages over bond vectors will be needed and it is convenient to introduce the shorthand notation

$$\langle F(\mathbf{Q}^N)\rangle_Q\equiv\frac{1}{Z_1^N}\int F(\mathbf{Q}^N)\exp\left[-\sum_{n=1}^{N} f(Q_n)/kT\right]d^3Q_1\cdots d^3Q_N \tag{5}$$

for arbitrary functions of the bond vectors. The subscript ''Q'' is used to distinguish this average from those over r. Although the details vary with the choice for the function, $f(Q)$, it will be assumed that the specific averages

$$\langle Q^{2j}\rangle_Q\equiv\frac{1}{Z_1}\int Q^{2j}e^{-f(Q)/kT}d^3Q \tag{6}$$

are available by some other means. For instance, in the above-mentioned case of the chain of harmonic springs, it is readily worked out that

$$\langle Q^{2j}\rangle_Q=\left(\frac{a^2}{\epsilon}\right)^j\frac{(2j+2)!}{2}\,\frac{D_{-(2j+3)}(-\sqrt{\epsilon})}{D_{-3}(-\sqrt{\epsilon})}, \tag{7}$$

where $\epsilon=Ha^2/kT$ and $D_\nu(z)$ is a parabolic cylinder function.

If Eq. (2) is substituted into Eq. (4), the integrations over $\mathbf{r}$ can be evaluated immediately, and there results

$$\langle r^{2j}\rangle_0=\sum_{k_1=1}^{N}\cdots\sum_{k_{2j}=1}^{N}\langle(\mathbf{Q}_{k_1}\cdot\mathbf{Q}_{k_2})\cdots(\mathbf{Q}_{k_{2j-1}}\cdot\mathbf{Q}_{k_{2j}})\rangle_Q \tag{8}$$

Consider each term in the sum separately. The distribution function does not depend on the orientation of the bond vectors, so that the directions of the bond vectors are manifest solely in the combination of dot products in the average. In any one term, a specific vector may appear once, twice, or even more times. Since each vector is odd under inversion and the overall integral must be invariant under an inversion, the nonvanishing terms in the sum must have the indices paired up. The complication is that there are many distinct ways that the indices can be paired up, so that it is not valid to just pick j of the indices and then set them equal to the other j.

A schematic way to represent this constraint on the indices is with the formula

$$\langle r^{2j}\rangle_0=\sum_{k_1=1}^{N}\cdots\sum_{k_{2j}=1}^{N}\langle(\mathbf{Q}_{k_1}\cdot\mathbf{Q}_{k_2})\cdots(\mathbf{Q}_{k_{2j-1}}\cdot\mathbf{Q}_{k_{2j}})\times\{\delta_{k_1k_2}\cdots\delta_{k_{2j-1}k_{2j}}+\text{all distinct permutations}\}\rangle_Q. \tag{9}$$

The phrase ''all distinct permutations'' needs some elaboration. There is a total of $(2j)!$ possible permutations of the indices, but a large number of them yield identical terms. For instance, interchanging k_1 and k_2 in the first product of Krönecker deltas yields exactly the same combination of Krönecker deltas and really does not provide a distinct term in the average. It should not be included in the sum of ''distinct permutations.'' Similarly, the interchange of the pair k_1,k_2 with the pair k_3,k_4 in the first term yields an identical product of Krönecker deltas and should also not be included in the sum of distinct permutations. One may imagine a large number of such equivalences, not necessarily involving the first product of Krönecker deltas. Further complications arise when considering terms in which more than one pair are the same. Suppose that $k_1=k_3$, so that this implies that $k_1=k_2=k_3=k_4$ in the first term of the sum of Krönecker deltas. Any permutation of these four indices, even one that would have yielded distinct terms if $k_1\neq k_3$ were true, should not be viewed as a distinct permutation in this case.

A second thing to notice is that the distribution of bond vectors is symmetric with respect to all permutations of the N vectors. Any terms in the sum, therefore, which differ only by a permutation of the particular labels used on the vectors must have the same value. This allows the grouping of similar terms and an elimination of the sums over the indices. This also is not a trivial matter, but depends on the number of pairs of indices that are equal as well as how they are paired in the dot products. If only terms where j indices are different and where only the first product of Krönecker deltas in Eq. (9) is considered, there results

$$\begin{aligned}&\sum_{\substack{k_1=1\\k_1\neq k_3\neq\cdots\neq k_{2j-1}}}^{N}\cdots\sum_{k_{2j}=1}^{N}\langle(\mathbf{Q}_{k_1}\cdot\mathbf{Q}_{k_2})\cdots(\mathbf{Q}_{k_{2j-1}}\cdot\mathbf{Q}_{k_{2j}})\delta_{k_1k_2}\cdots\delta_{k_{2j-1}k_{2j}}\rangle_Q\\&=\sum_{\substack{k_1=1\\k_1\neq k_2\neq\cdots\neq k_j}}^{N}\cdots\sum_{k_j=1}^{N}\langle Q_{k_1}^2Q_{k_2}^2\cdots Q_{k_j}^2\rangle_Q\\&=N(N-1)\cdots(N-j+1)\langle Q^2\rangle_Q^j.\end{aligned} \tag{10}$$

Clearly, there are many more terms to consider and it is important to properly categorize terms as well as determine how many times they occur.

Before getting to the diagrammatic treatment, it is instructive to see how the first two even moments can be evaluated by brute force. For $j=1$, there are only two sums in Eq. (8), and it is clear that only terms with $k_2=k_1$ will contribute. Further, every term in the remaining sum has the same value so that

$$\langle r^2\rangle_0=N\langle Q^2\rangle_Q. \tag{11}$$

For $j=2$, there are four sums, and k_1 may equal either k_2, k_3, or k_4. Whichever one k_1 equals, the remaining two must then be equal to each other. In addition, if k_1 equals k_2, the dot product $\mathbf{Q}_{k_1}\cdot\mathbf{Q}_{k_1}$ eliminates the angular dependence of the vectors; the dot product of the other pair will likewise eliminate any angular dependence. On the other hand, if k_1 equals either k_3 or k_4 and the two pairs are distinct, there results $(\mathbf{Q}_{k_1}\cdot\mathbf{Q}_{k_2})^2$ and the angular dependence of the vectors must be taken into account. There is a further distinction in that the two pairs may be different or equal. Proceeding in detail, there results

$$\langle r^4\rangle_0=\sum_{\substack{k_1=1\\k_1\neq k_2}}^{N}\sum_{k_2=1}^{N}\langle Q_{k_1}^2Q_{k_2}^2\rangle_Q+2\sum_{\substack{k_1=1\\k_1\neq k_2}}^{N}\sum_{k_2=1}^{N}\langle(\mathbf{Q}_{k_1}\cdot\mathbf{Q}_{k_2})^2\rangle_Q+\sum_{k_1=1}^{N}\langle Q_{k_1}^4\rangle_Q$$
$$=N(N-1)\{\langle Q^2\rangle_Q^2+2\langle(\mathbf{Q}_1\cdot\mathbf{Q}_2)^2\rangle_Q\}+N\langle Q^4\rangle_Q. \quad (12)$$

There are three different types of terms and each has its own associated coefficient. The evaluation over the angles of the bond vectors will be taken up later.

The types of complications that arise are already evident for $j=2$, but the above-mentioned procedure illustrates the general approach that must be taken for an arbitrary term. First, the types of terms that can occur must be identified. The indices must all be paired, but terms differ in how many pairs of indices are equal and in how the indices are connected in the dot products. All distinct arrangements lead to different results and so, should be considered separately. The second task is to determine how many times each type of term occurs to get the correct associated coefficient. Finally, the terms must be evaluated and combined.

III. CATEGORIZATION OF TERMS

Equation (9) contains summations of a large number of terms, in general. The objective of categorizing them is to group together equivalent terms and evaluate them all together. This reduces the overall number of terms to consider. The $2j$ indices must all be paired, so that a first step is to group the indices together into j pairs. It is possible for all pairs to be distinct, or for any number of pairs to be the same. The types of terms can be further distinguished by how they are linked through the dot products.

The categorization is broken down into two steps, which will be called the primary and secondary categorizations. The primary categorization involves finding all the possible sets of ‘‘cluster’’ sizes. After pairing up the indices, clusters can be made that contain one or more equal pairs. All possible arrangements of cluster sizes can occur and must be treated separately. For instance, for $j=2$, there are two pairs and it is possible for the pairs to be different or the same. In the first case, there will be two clusters, both of size one. In the second case, there is just one cluster of size two. The number of possibilities, naturally, increases with j.

The secondary categorization determines how many distinct ways the pairs of indices can be linked given a set of cluster sizes. For instance, for $j=2$ and two clusters of size one, there are terms where each pair consists of a dot product with itself and other terms where a member of each pair has a dot product with a member of the other. The general procedure for categorizing the terms will now be presented and the specific case for $j=5$ will be used for illustrative purposes.

The primary categorization can be reduced to a recursive procedure that is easily programed on a computer. Each primary categorization is labeled by a set of cluster sizes. For a given j, the clusters can have sizes from 1 to j. Let g_m denote the number of clusters with size m. Each categorization is then labeled by a unique set of integers, $\{g_m,\ m=1,...,j\}$. It is only necessary that

$$\sum_{m=1}^{i} mg_m=j. \quad (13)$$

The following procedure can be followed to generate all possible sets.

(a) Begin with $g_1=j$, the rest of the g_m's being zero.
(b) Evaluate the quantity $J_k=\Sigma_{m=1}^{k} mg_m$, starting with $k=1$ and continuing to successively higher values of k until $J_k-(k+1)>0$.

TABLE I. Clusters and coefficients associated with $j=5$.

Set	g_1	g_2	g_3	g_4	g_5	Factor	N_l
1	5	0	0	0	0	$N(N-1)(N-2)(N-3)(N-4)$	945
2	3	1	0	0	0	$N(N-1)(N-2)(N-3)$	3150
3	1	2	0	0	0	$N(N-1)(N-2)$	1575
4	2	0	1	0	0	$N(N-1)(N-2)$	630
5	0	1	1	0	0	$N(N-1)$	210
6	1	0	0	1	0	$N(N-1)$	45
7	0	0	0	0	1	N	1

(c) With this value of k, set $g_m=0$ for all $m\leqslant k$, increase g_{k+1} by one, and set $g_1=J_k-(k+1)$.

(d) Perform steps (b) and (c) repeatedly until the final case, where $g_j=1$ and the rest are zero, is achieved.

For $j=5$, the seven sets of clusters listed in Table I are obtained.

The secondary categorization looks at how the indices are connected through the dot products. This is where the diagrams are helpful. Begin by constructing a representation for the primary categorizations; call these the primary diagrams. Each pair of indices is assigned a dot. If any pairs are equal, the corresponding dots are joined to form a ''star.'' A star is a filled-in object with the number of points representing the number of equal-index pairs. For generality, a dot may be viewed as a star with one point. Each pair of indices is then associated with a point on a star. For $j=5$, the sets of stars associated with all the primary categorizations is shown in Fig. 1. With practice, generation of these primary diagrams may be performed by inspection and is often easier than following the set of rules outlined in the previous paragraph. However, the set of rules is more amenable for writing a computer program to perform the task.

Once these primary diagrams are generated, a new symbol is introduced that represents the arrangement of the dot products. Each point on a star represents a pair of bond vectors. Each bond vector is connected with another through a dot product, so each point on a star can make two connections. These connections are represented in the diagrams by directed lines. Every point of a star in the diagram must have one directed line coming in and one going out. It is possible to have a point with a single directed line coming in and going out of it; this corresponds to a dot product of the vector with itself.

There are some complications that can arise, and so some care is needed to construct a set of rules for the full generation of allowable diagrams. First of all, to avoid cluttering up the diagram, single lines that leave and enter the same point are left out. Any unconnected points are assumed to represent dot products of the vector with itself. Second, a star with more than one point must be treated specially to avoid generating more than one equivalent diagram. Because all the points on a star represent the same vector, there is some arbitrariness in choosing where lines come in and

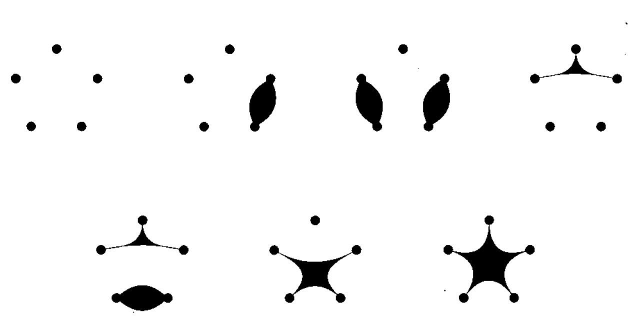

FIG. 1. The primary diagrams for $j=5$.

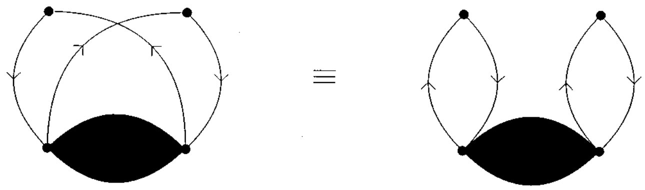

FIG. 2. Two equivalent diagrams, only one of which is acceptable.

where they leave. To avoid confusion, if a line comes into one point of a star from another star, a line must leave by the same point to another star. Lines should not connect two different points of the same star. Since such lines would represent the dot product of a vector with itself, they should be represented as points without any directed lines connected.

Another restriction needs to be made regarding stars of more than one point. Directed lines may be followed to form closed loops. Because of the equivalence of points on a star, it is possible to manipulate lines in a diagram so that a given loop contacts the star at only one point. To illustrate this, Fig. 2 shows two equivalent diagrams for $j=4$. In the first diagram, there is one loop that contacts the two-point star at two points. In the second, the lines have been rearranged so that there are two loops, each contacting the star at only one point. For any general set of equivalent diagrams, only one has closed loops that contact each star at no more than one point. The rule is therefore made that any loop of an acceptable diagram may contact a given star at no more than one point. With this rule, only the second diagram in Fig. 2 would be acceptable.

It must also be kept in mind that stars with the same number of points are equivalent. Diagrams which differ only by a rearrangement of such stars are equivalent and only one should be included.

A moment can be represented by a sum of diagrams. A set of rules for constructing the diagrams can be stated as follows.

(a) Given j, the set of all primary diagrams is generated as described above.

(b) Starting with each primary diagram, directed lines are added to construct all topologically distinct diagrams such that:

(i) a point may have no lines connected to it,

(ii) if a line enters a point, another line must leave it,

(iii) a line may not leave and enter the same point, and

(iv) the lines form closed loops which can connect to any star at no more than one point.

Following these rules, all distinct diagrams may be constructed. These represent the types of terms that will be encountered. As an example, all the diagrams associated with $j=5$ are shown in Fig. 3.

IV. EVALUATION OF THE DIAGRAMS

The moment is now expressed as a sum of diagrams. It only remains to evaluate these diagrams and combine them for the final result. To interpret the diagrams, label each star with an index. These represent labels for the bond vectors. The directed lines are used to form dot products between the vectors and unconnected points represent vectors that have dot products with themselves. The full list of terms for $j=5$ is rather long, so only the terms for $j=5$ and the first primary categorization, with $g_1=5$, will be presented. The bond vector averages that go with these diagrams are given in Table II. Aside from evaluating these averages, it is necessary to determine the coefficients that go with each term.

Each primary categorization has associated with it a certain number of distinct indices. For instance, set 1 in Table I contains five distinct indices, set 2 contains four, and so on. Denote the number of distinct indices in a given set by M. As already mentioned, terms which differ only in the indices used to label the vectors have the same value. Therefore, with a given primary categorization, the sum over all appropriate indices can be performed explicitly. For instance, for set

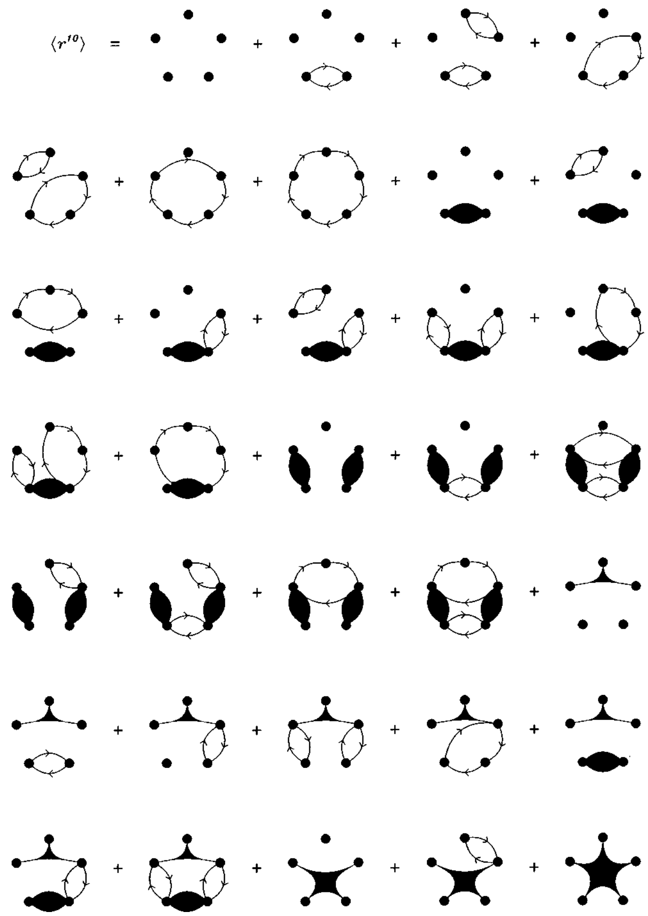

FIG. 3. The full set of diagrams for $j=5$.

1 in Table I, all sets of five distinct indices give the same set of terms corresponding to this primary categorization. It is then possible to extract one such set of terms—say for $k_1=1$, $k_2=2$, $k_3=3$, $k_4=4$, and $k_5=5$—and pull it out of the sums completely. It is then necessary to sum over all sets of five distinct indices, none of which may be equal, to arrive at the coefficient, $N(N-1)(N-2)(N-3)(N-4)$, which must multiply all diagrams associated with this primary categorization. More generally, with M different indices in a given categorization, the sums over indices yield a factor of $N!/(N-M)!$. The factors for $j=5$ are included in Table I.

A useful quantity for checking the generation of secondary diagrams is the number of terms associated with each primary categorization. This is associated with the number of distinct permutations of the delta functions in Eq. (9) that correspond to a given categorization. This is a

TABLE II. Evaluated terms for the $g_1=5$ cluster diagrams for $j=5$.

Term	Coefficient
$Q_1^2Q_2^2Q_3^2Q_4^2Q_5^2$	1
$Q_1^2Q_2^2Q_3^2(\mathbf{Q}_4\cdot\mathbf{Q}_5)^2$	20
$Q_1^2(\mathbf{Q}_2\cdot\mathbf{Q}_3)^2(\mathbf{Q}_4\cdot\mathbf{Q}_5)^2$	60
$Q_1^2Q_2^2(\mathbf{Q}_3\cdot\mathbf{Q}_4)(\mathbf{Q}_4\cdot\mathbf{Q}_5)(\mathbf{Q}_5\cdot\mathbf{Q}_3)$	80
$(\mathbf{Q}_1\cdot\mathbf{Q}_2)^2(\mathbf{Q}_3\cdot\mathbf{Q}_4)(\mathbf{Q}_4\cdot\mathbf{Q}_5)(\mathbf{Q}_5\cdot\mathbf{Q}_3)$	160
$Q_1^2(\mathbf{Q}_2\cdot\mathbf{Q}_3)(\mathbf{Q}_3\cdot\mathbf{Q}_4)(\mathbf{Q}_4\cdot\mathbf{Q}_5)(\mathbf{Q}_5\cdot\mathbf{Q}_2)$	240
$(\mathbf{Q}_1\cdot\mathbf{Q}_2)(\mathbf{Q}_2\cdot\mathbf{Q}_3)(\mathbf{Q}_3\cdot\mathbf{Q}_4)(\mathbf{Q}_4\cdot\mathbf{Q}_5)(\mathbf{Q}_5\cdot\mathbf{Q}_1)$	384

simple problem in combinatorics, being equal to the number of distinct permutations of the j pairs of integers in the list among j boxes, each being able to hold only two integers. For an arbitrary set l, this is given by

$$N_l = \frac{(2j)!}{\Pi_m (m!)^{g_m} g_m!} \tag{14}$$

For $j=5$, the values are included in Table I.

A coefficient is also associated with each diagram based on the secondary categorization. These coefficients represent the number of times such a term appears in the sum of Krönecker deltas. Determining these coefficients leads to exercises in combinatorics. The indices are viewed as pairs of labels that are to go into j boxes, each being able to hold two labels. Vectors that are connected by a dot product have their indices paired. It is necessary to determine how many distinct ways a particular set of pairings may be accomplished. It is important to keep in mind, though, that overall permutations of the labels has already been performed in the sums over the indices. If a simple permutation of labels leads to an indistinguishable term, this permutation must not be included in the coefficient.

For example, the first term in Table II contains the five pairings (11), (22), (33), (44), and (55). The only permutations that yield equivalent pairings are those where a pair is interchanged completely with another. There are $5!=120$ such permutations. However, note that all simple permutations of the five labels also lead to equivalent terms. Therefore, the value of 120 must be divided by the number of such permutations, i.e., 120, to get the correct coefficient of 1.

Consider, now, the second term in Table II. This has the five pairings (11), (22), (33), (45), and (45). One way to count the possible permutations is to note that the pair (11) has five possible positions. Once this placement is taken care of, the pair (22) has four possible positions. Then, the pair (33) has three possible positions. This then fixes where the two pairs (45) must go. The indices in each of these two final pairs may be interchanged, meaning there are four possible arrangements of them. Specifically, these are (45)(45), (45)(54), (54)(45), and (54)(54). Overall, then, there are $5\times4\times3\times4=240$ appropriate permutations. However, permutations of the labels 1, 2, and 3 among themselves lead to indistinguishable terms; there are six such permutations. Further, permutation of 4 and 5 lead to indistinguishable terms; there are two such permutations. No other permutations of the labels yield indistinguishable terms. Therefore, the result of 240 must be divided by $2\times6=12$ to give 20. The other terms can be worked out in an analogous way and the coefficients for the diagrams listed in Table II are given there.

As a check, it should be noted that the sum of coefficients for all the diagrams associated with a given primary categorization should equal the quantity, N_l, defined in Eq. (14). Note that the sum of coefficients in Table II equals 945, the number given in Table I for this primary categorization.

The general procedure has been given. The details for the specific cases can be worked out with little difficulty. The final step involves performing the averages over the dot products. This requires integrals over angles, which are readily performed, the procedure being outlined in the Appendix. For the first few moments, there results

$$\langle r^2\rangle_0 = N\langle Q^2\rangle_Q, \tag{15a}$$

$$\langle r^4\rangle_0 = \tfrac{5}{3}N(N-1)\langle Q^2\rangle_Q^2 + N\langle Q^4\rangle_Q, \tag{15b}$$

$$\langle r^6\rangle_0 = \tfrac{35}{9}N(N-1)(N-2)\langle Q^2\rangle_Q^3 + 7N(N-1)\langle Q^2\rangle_Q^2\langle Q^4\rangle_Q + N\langle Q^6\rangle_Q, \tag{15c}$$

$$\begin{aligned}\langle r^8\rangle_0 = {}& \tfrac{35}{3}N(N-1)(N-2)(N-3)\langle Q^2\rangle_Q^4 + 42N(N-1)(N-2)\langle Q^2\rangle_Q^2\langle Q^4\rangle_Q \\ & + \tfrac{63}{5}N(N-1)\langle Q^4\rangle_Q^2 + 12N(N-1)\langle Q^2\rangle_Q\langle Q^6\rangle_Q + N\langle Q^8\rangle_Q,\end{aligned} \tag{15d}$$

$$\langle r^{10}\rangle_0=\tfrac{385}{9}N(N-1)(N-2)(N-3)(N-4)\langle Q^2\rangle_Q^5+\tfrac{770}{3}N(N-1)(N-2)(N-3)\langle Q^2\rangle_Q^3\langle Q^4\rangle_Q$$
$$+231N(N-1)(N-2)\langle Q^2\rangle_Q\langle Q^4\rangle_Q^2+110N(N-1)(N-2)\langle Q^2\rangle_Q^2\langle Q^6\rangle_Q+66N(N-1)$$
$$\times\langle Q^4\rangle_Q\langle Q^6\rangle_Q+\tfrac{55}{3}N(N-1)\langle Q^2\rangle_Q\langle Q^8\rangle_Q+N\langle Q^{10}\rangle_Q, \tag{15e}$$

$$\langle r^{12}\rangle_0=\tfrac{5005}{27}N(N-1)(N-2)(N-3)(N-4)(N-5)\langle Q^2\rangle_Q^6+\tfrac{5005}{3}N(N-1)(N-2)(N-3)(N-4)$$
$$\times\langle Q^2\rangle_Q^4\langle Q^4\rangle_Q+3003N(N-1)(N-2)(N-3)\langle Q^2\rangle_Q^2\langle Q^4\rangle_Q^2+\tfrac{2860}{3}N(N-1)(N-2)(N$$
$$-3)\langle Q^2\rangle_Q^3\langle Q^6\rangle_Q+\tfrac{3003}{5}N(N-1)(N-2)\langle Q^4\rangle_Q^3+1716N(N-1)(N-2)$$
$$\times\langle Q^2\rangle_Q\langle Q^4\rangle_Q\langle Q^6\rangle_Q+\tfrac{715}{3}N(N-1)(N-2)\langle Q^2\rangle_Q^2\langle Q^8\rangle_Q+\frac{858}{7}N(N-1)\langle Q^6\rangle_Q^2$$
$$+143N(N-1)\langle Q^4\rangle_Q\langle Q^8\rangle_Q+26N(N-1)\langle Q^2\rangle_Q\langle Q^{10}\rangle_Q+N\langle Q^{12}\rangle_Q. \tag{15f}$$

Note that these expressions reduce to those for the Kramers chain[5] when $\langle Q^{2j}\rangle_Q=a^{2j}$.

V. CONCLUSIONS

A general procedure has been given for the closed form evaluation of the even moments of the distribution of end-to-end separations of a polymer chain. The only assumption made is that the energy of the polymer chain depends only on the lengths of the polymer bonds. The procedure can be applied to any order, although it is acknowledged that the procedure becomes increasingly tedious with higher orders. Nevertheless, the procedure is considerably easier than the brute force approach.

It may be noted that the terms in Eq. (15) to leading order in N are identical to what would be obtained from the Gaussian chain model:

$$\langle r^{2j}\rangle_G=\frac{(2j+1)!!}{3^j}N^j\langle Q^2\rangle_Q^j. \tag{16}$$

The expressions derived here would therefore converge to the Gaussian chain results as N gets larger, corrections being of the order $1/N$. This may be viewed simply as a statement of the central limit theorem.[6] A general expression for the distribution function of polymer chain end-to-end separations may be derived as a Gaussian chain distribution times a series expansion correction, successive terms being of higher orders in $1/N$.[7] The moments would therefore be expected to be expansions in $1/N$ as well, as is clear from the equations derived here. As noted in Sec. I, it was desired to have moments for all possible chain lengths. The expressions derived here were therefore found to be quite useful in the stress–strain studies.[1] The expansion derived in that work was truncated so that moments only up to the eighth were needed. If a truncation at the next higher order were desired, moments up to the twelfth would be needed, and the appropriate expressions are presented in closed form here. Higher orders, if needed, could be derived by the procedure outlined in this paper.

This procedure is restricted to a specific type of energy dependence of the polymer chain, and the possibility of generalization may be worth investigating. As it turns out, the procedure is unchanged if the energy depends not only on the length, but also on the direction of each bond, except that the evaluation of the angular integrals as described in the Appendix is more complicated. The restriction of bond angles and even dihedral angles would lead to more realistic distribution functions. The methods for evaluating the moments of such distributions with rigid bonds in the rotational isomeric state approximation[8,9] seem rather cumbersome and generally require numerical evaluation. The extension of these models to chains where bonds may bend and stretch away from their equilibrium values would be extremely difficult using these methods. However, there may be some hope for the diagrammatic approach presented here.

There is another intriguing possibility associated with this procedure that is worth mentioning. The diagrammatic approach used here has analogies with that used in cluster expansions of the two-body distribution function of many-body systems.[2,3] Not only does the diagrammatic treatment aid in the identification and evaluation of terms in the cluster expansions, it is also helpful in deriving relationships between the various types of terms that arise. These relationships allow a resummation of the expansion to infinite order resulting in either the hypernetted chain (HNC)[10] equations in the classical case or the fermion hypernetted chain (FHNC)[11,12] equations in the many-fermion case. The importance of these equations is realized when it is noted that the terms of the cluster expansion may either converge slowly or even diverge at low orders. The resummation removes these difficulties. The moments of the polymer equilibrium distribution function can be used in an expansion of the distribution function for a polymer network under strain[13] and subsequently in an expansion of the free energy density associated with the material. A resummation along the lines of the HNC or FHNC equations might prove valuable for an improved theory of polymer networks. The method to be presented may serve as a starting point for such a development.

ACKNOWLEDGMENTS

The author would like to thank the T-12 group of the Los Alamos National Laboratory for their hospitality and support of this work. This work was supported by the U.S. Department of Energy, Contract No. W-7405-ENG-36 under the ASCI program.

APPENDIX: EVALUATION OF THE ANGULAR INTEGRALS

The evaluation of an angular integral can be reduced to an exercise in tensor analysis. Since the distribution of vector bonds depends only on their magnitudes, the general average can be reduced to

$$\langle(\mathbf{Q}_{k_1}\cdot\mathbf{Q}_{k_2})\cdots(\mathbf{Q}_{k_{2j-1}}\cdot\mathbf{Q}_{k_{2j}})\rangle_Q=\langle Q_{k_1}Q_{k_2}\cdots Q_{k_2}\rangle_{jQ}\langle(\mathbf{u}_{k_1}\cdot\mathbf{u}_{k_2})\cdots(\mathbf{u}_{k_{2j-1}}\cdot\mathbf{u}_{k_{2j}})\rangle_a, \tag{A1}$$

where $\mathbf{u}_i$ is a unit vector in the direction of $\mathbf{Q}_i$ and the subscript ''a'' on the final averaged quantity indicates that the average is taken only over angles. A generic way to write this is

$$\langle F(\mathbf{u}^N)\rangle_a\equiv\frac{1}{(4\pi)^N}\int F(\mathbf{u}^N)d\Omega_1\cdots d\Omega_N \tag{A2}$$

for an arbitrary function of the unit vectors.

The functions of interest here are simply dot products of the unit vectors. In general, there will be unit vectors appearing as various powers and connected with each other through the dot products. The angular average then reduces to a product of averages over single unit vectors, which are then connected through tensor products. For instance, there is the simple average

$$\langle(\mathbf{u}_1\cdot\mathbf{u}_2)^2\rangle_a=\langle\mathbf{u}_1\mathbf{u}_1\rangle_a:\langle\mathbf{u}_2\mathbf{u}_2\rangle_a \tag{A3}$$

or, getting a little more complicated, there is

$$\langle(\mathbf{u}_1\cdot\mathbf{u}_3)^2(\mathbf{u}_2\cdot\mathbf{u}_3)^2\rangle_a=\langle\mathbf{u}_1\mathbf{u}_1\rangle_a\langle\mathbf{u}_2\mathbf{u}_2\rangle_a\vdots\langle\mathbf{u}_3\mathbf{u}_3\mathbf{u}_3\mathbf{u}_3\rangle_a. \tag{A4}$$

An important first step is therefore the evaluation of the average of a particular power of a unit vector.

An elegant procedure for performing these averages has been presented by Brenner.[14] The general results is

$$\langle \mathbf{u}^p \rangle_a = \begin{cases} 0 & \text{if } p \text{ is odd} \\ \dfrac{1}{(p+1)!} (\boldsymbol{\nabla})^p r^p & \text{if } p \text{ is even.} \end{cases} \tag{A5}$$

As already noted in the main part of the paper, only even powers should appear in these averages. The above result makes this explicit.

This formal result can be made a bit more convenient. First, note that the successive gradients reduce the power of the coordinates by one, so that the final result is independent of r. Two successive gradients, though, will yield a nonzero result only if a particular coordinate appears twice. This leads to a pairing of the coordinate indices. This pairing, though, can occur in many different ways. A careful analysis leads to

$$\langle u_{j_1} \cdots u_{j_{2n}} \rangle_a = \frac{1}{(2n+1)!!} \{ \delta_{j_1 j_2} \cdots \delta_{j_{2n-1} j_{2n}} + \text{all proper permutations} \}, \tag{A6}$$

where the indices in Eq. (A6) indicate specific coordinates of the unit vectors: u_1, u_2, and u_3. The phrase ''all proper permutations'' has a slightly different meaning than the phrase ''all distinct permutations'' used in the main part of the paper. All indices, $\{j_1,\ldots,j_{2n}\}$ are viewed as distinct and all distinct permutations of the first product of delta functions is included in the sum. In all, there are $(2n-1)!!$ such permutations. Several examples might help see how this works.

The average for $n=1$ is given explicitly by

$$\langle \mathbf{uu} \rangle_a = \tfrac{1}{3} \boldsymbol{\delta}, \tag{A7}$$

where $\boldsymbol{\delta}$ is the unit tensor. For $n=2$, there results

$$\langle \mathbf{uuuu} \rangle_a = \tfrac{1}{15} \{ \boldsymbol{\delta\delta} + \mathbf{I} + \mathbf{I}^\dagger \}, \tag{A8}$$

where $\mathbf{I}$ is a fourth-rank tensor with elements $I_{ijkl} = \delta_{ik}\delta_{jl}$. Similarly, the transpose has elements $I^\dagger_{ijkl} = \delta_{il}\delta_{jk}$. The result for $n=3$ requires more space, but is obtained in a straightforward way,

$$\begin{aligned}(\langle \mathbf{uuuuuu} \rangle_a)_{ijklmn} = \tfrac{1}{105} \{ & \delta_{ij}\delta_{kl}\delta_{mn} + \delta_{ij}\delta_{km}\delta_{ln} + \delta_{ij}\delta_{kn}\delta_{lm} + \delta_{ik}\delta_{jl}\delta_{mn} + \delta_{ik}\delta_{jm}\delta_{ln} \\ & + \delta_{ik}\delta_{jn}\delta_{lm} + \delta_{il}\delta_{jk}\delta_{mn} + \delta_{il}\delta_{jm}\delta_{kn} + \delta_{il}\delta_{jn}\delta_{km} + \delta_{im}\delta_{jk}\delta_{ln} \\ & + \delta_{im}\delta_{jl}\delta_{kn} + \delta_{im}\delta_{jn}\delta_{kl} + \delta_{in}\delta_{jk}\delta_{lm} + \delta_{in}\delta_{jl}\delta_{km} + \delta_{in}\delta_{jm}\delta_{kl} \}. \end{aligned} \tag{A9}$$

The results for higher orders should be apparent. The procedure is readily programmed on a computer.

These results can now be used, the appropriate summations over indices being performed to obtain the final results. Several examples can illustrate this. First, consider

$$\langle (\mathbf{Q}_1 \cdot \mathbf{Q}_2)^2 \rangle_Q = \langle Q_1^2 Q_2^2 \rangle_Q \langle \mathbf{u}_1 \mathbf{u}_1 \rangle_a : \langle \mathbf{u}_2 \mathbf{u}_2 \rangle_a = \tfrac{1}{9} \langle Q^2 \rangle_Q^2 \boldsymbol{\delta} : \boldsymbol{\delta} = \tfrac{1}{3} \langle Q^2 \rangle_Q^2. \tag{A10}$$

A more general result is given by

$$\langle (\mathbf{Q}_1 \cdot \mathbf{Q}_2)^{2n} \rangle_Q = \frac{1}{2n+1} \langle Q^n \rangle_Q^2. \tag{A11}$$

Another example is

$$\langle (\mathbf{Q}_1 \cdot \mathbf{Q}_2)(\mathbf{Q}_2 \cdot \mathbf{Q}_3)(\mathbf{Q}_3 \cdot \mathbf{Q}_1) \rangle_Q = \langle Q_1^2 Q_2^2 Q_3^2 \rangle_Q \, \mathrm{tr} \{ \tfrac{1}{3} \boldsymbol{\delta} \cdot \tfrac{1}{3} \boldsymbol{\delta} \cdot \tfrac{1}{3} \boldsymbol{\delta} \} = \tfrac{1}{9} \langle Q^2 \rangle_Q^3. \tag{A12}$$

Generalizing this,

$$\langle(\mathbf{Q}_1\cdot\mathbf{Q}_2)(\mathbf{Q}_2\cdot\mathbf{Q}_3)\cdots(\mathbf{Q}_n\cdot\mathbf{Q}_1)\rangle_Q=\frac{1}{3^{n-1}}\langle Q^2\rangle_Q^n. \tag{A13}$$

The procedure is readily applied to any of the angular averages that arise.

[1] G. G. Hoffman, J. Polymer Sci. (submitted).
[2] H. Ursell, Proc. Cambridge Philos. Soc. **23**, 685 (1927); J. Yvon, Actualités Sci. Ind. **203**, (1935). J. Mayer J. Chem. Phys. **5**, 67 (1937).
[3] M. Gaudin, J. Gillespie, and G. Ripka, Nucl. Phys. A **176**, 237 (1971).
[4] G. G. Hoffman, J. Phys. Chem. B **103**, 7167 (1999).
[5] P. J. Flory, *Statistical Mechanics of Chain Models* (Interscience, New York, 1969).
[6] See, for example, W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. (Wiley, New York, 1968), Vol. I.
[7] K. Nagai, J. Chem. Phys. **40**, 2818 (1964).
[8] P. J. Flory and R. J. Jernigan, J. Chem. Phys. **42**, 3509 (1965).
[9] K. Nagai, J. Chem. Phys. **48**, 5646 (1968).
[10] J. M. Van Leeuwen, J. Groeneveld, and J. deBoer, Physica (Amsterdam) **25**, 792 (1959).
[11] S. Fantoni and S. Rosati, Lett. Nuovo Cimento **10**, 545 (1974); Nuovo Cimento A **25**, 593 (1975).
[12] E. Krotscheck and M. L. Ristig, Phys. Lett. **48A**, 17 (1974); Nucl. Phys. A **242**, 389 (1975).
[13] K. Nagai, J. Chem. Phys. **38**, 924 (1963).
[14] H. Brenner, Chem. Eng. Sci. **19**, 519 (1964).

Zeros of the Jimbo, Miwa, Ueno tau function

John Palmer[a)]
Department of Mathematics, University of Arizona, Tucson, Arizona 85721

(Received 2 March 1999; accepted for publication 30 August 1999)

We introduce a family of local deformations for meromorphic connections on $\mathbf{P}^1$ in the neighborhood of a higher rank (simple) singularity. Following the scheme in Malgrange [*Mathematique et Physique*, Progress in Mathematics (Birkhäuser, Boston, 1983), Vol. 37, pp. 381–400, *ibid.*, pp. 401–426; *ibid.*, pp. 427–438] we use these local models to prove that the zeros of the tau function, introduced by Jimbo, Miwa, and Ueno in their pioneering work on ''Birkhoff'' deformations at irregular singular points [Physica D **2**, 306–352; **2**, 407–448 (1981); **4**, 26–46 (1983); Publ. RIMS Kyoto Univ. **17-2**, 703–721 (1981)], occur at precisely those points in the deformation space at which a certain Birkhoff–Riemann–Hilbert problem fails to have a solution. [S0022-2488(99)01112-3]

I. INTRODUCTION

A. The Riemann–Hilbert problem and monodromy preserving deformations

Suppose that $A(x)$ is a $p\times p$ matrix with entries that are rational functions of x on $\mathbf{P}^1$. Linear differential equations,

$$\frac{d\psi}{dx}=A(x)\psi, \tag{1.0}$$

arise in many important applications and have been studied intensively for more than 100 years. The nature of the singularities in the coefficient matrix $A(x)$ at its poles has a lot to do with the character of the solutions to (1.0). For example, in the neighborhood of a point a where $A(x)$ has but a simple pole, it is well known that the equation (1.0) has a fundamental solution, $\Psi(x)$, with polynomially limited growth as $x\rightarrow a$ (by which we mean that solutions are dominated near $x=a$ by $c|x-a|^{-N}$ where c and N are constants). Higher-order poles in $A(x)$ typically produce local fundamental solutions with more complicated exponential–polynomial growth near the pole. Fundamental solutions to the equation (1.0) are usually not single valued. If one has a fundamental solution $\Psi(x)$ defined for x in some small ball not containing any of the poles, $\{a_1,a_2,\ldots,a_n\}$, of $A(x)$, then it is possible to analytically continue $\Psi(x)$ along paths in $\mathbf{P}^1$ that avoid these poles. The analytic continuation depends only on the homotopy class of the path, so that the resulting function, $\Psi(x)$, is defined not on $\mathbf{P}^1$ but for x in the simply connected covering space, $\mathcal{R}(X)$, of $X:=\mathbf{P}^1\backslash\{a_1,a_2,\ldots,a_n\}$ with projection $\pi:\mathcal{R}(X)\rightarrow X$. If $x_0\in\mathcal{R}(X)$ and γ is a closed path in X with base point $\pi(x_0)$ then

$$\Psi(\gamma\cdot x_0)=\Psi(x_0)M(\gamma)^{-1},$$

where $x_0\rightarrow\gamma\cdot x_0$ is the natural action of the homotopy class of the path γ on $x_0\in\mathcal{R}(X)$, and $\gamma\rightarrow M(\gamma)$ is an n-dimensional representation of the fundamental group $\pi_1(X,\pi(x_0))$. Riemann's analysis of the solutions of the hypergeometric equation in terms of its *monodromy representation* $M(\cdot)$ led to the general problem of determining if any representation of the fundamental group of the punctured sphere could arise as the monodromy representation for a linear differential equation (1.0). It is clear from simple examples that there will not be a unique association between a

[a)]Electronic mail: Palmer@math.arizona.edu

representation of π_1 and a differential equation (1.0) without some further restriction on the differential equation. It is possibly most natural to look for a differential equation with *simple poles* to realize a given monodromy representation. It is not always possible to solve this problem,[1] but if one relaxes the condition on the differential equation to admit *regular singular points* then it is classical that the inverse monodromy problem always has a solution (a good discussion of the confusion about the status of the solution to this problem can also be found in Ref. 1). A pole a of $A(x)$ in (1.0) is said to be a regular singular point if fundamental solutions to (1.0) have, at worst, polynomial growth at a. Because the fundamental solutions are defined on the simply connected covering and paths that wind around the point a pick up powers of the local monodromy, the precise notion of polynomial growth requires some restriction to sectors in the covering space.[1] More important for us, however, is *a deformation* variant of the simple pole condition. Suppose that for some collection of points $\{a_1^0,a_2^0,\ldots,a_n^0\}$, a given representation of

$$\pi_1(\mathbf{P}^1\backslash\{a_1^0,a_2^0,\ldots,a_n^0\},a_0),$$

can be realized by a differential equation (1.0) with *simple poles* at a_j^0 $j=1,2,...,n$ and a fundamental solution $\Psi(x)$ normalized to $\Psi(a_0)=I$ at the base point a_0. In this case, we write $A^0(x)$ for the matrix coefficient in the differential equation (1.0) that realizes the appropriate monodromy representation. Since $A^0(x)$ has simple poles,

$$A^0(x)=\sum_{\nu=1}^{n}\frac{A_\nu^0}{x-a_\nu^0}.$$

The problem, first formulated by Schlesinger,[2] is to ask whether it is possible to deform the coefficients A_ν^0 in $A^0(x)$ as functions of the pole locations a_ν so that the differential equation (1.0) with coefficient matrix.

$$A(x)=\sum_{\nu=1}^{n}\frac{A_\nu(a)}{x-a_\nu}$$

realizes the *same* monodromy representation as as the differential equation with coefficient matrix (1.1) (we will be more precise about what this means later on). Note that we have written $a=(a_1,a_2,\ldots,a_n)$, and we want

$$A_\nu(a^0)=A_\nu(a_1^0,a_2^0,\ldots,a_n^0)=A_\nu^0.$$

When the point at infinity is a regular point and $a_0=\infty$, Schlesinger showed that if such a deformation exits the coefficients A_ν must satisfy a nonlinear system of differential equations,

$$dA_\mu=-\sum_{\nu\neq\mu}\frac{A_\mu A_\nu-A_\nu A_\mu}{a_\mu-a_\nu}d(a_\mu-a_\nu),$$

now called the Schlesinger equations. A modern treatment of the existence question can be found in Ref. 3. If we write $a\in\mathbf{C}^n$ then, as might be guessed from looking at the Schlesinger equations, it is important to remove the points at which $a_\mu=a_\nu$ from consideration. Let

$$\Delta_{\mu\nu}=\{a\in\mathbf{C}^n | a_\nu=a_\mu\},$$

and define

$$Z^n=\mathbf{C}^n\backslash\cup_{\nu\neq\mu}\Delta_{\mu\nu}.$$

As observed by Malgrange, an appropriate place to seek monodromy preserving deformations is the simply connected covering space $\mathcal{R}(Z^n)\to Z^n$. Because this same space will enter our considerations with a different significance later on, it will be useful at this point to introduce distinctive notation for elements in $\mathcal{R}(Z^n)$ and their projection onto Z^n. We will write $t\in\mathcal{R}(Z^n)$ and $a(t)\in Z^n$ for the projection of t on $Z^n\subset C^n$. We write $a_j(t)$ for the jth component of $a(t)$.

As pointed out in Refs. 3 and 4, the space $\mathcal{R}(Z^n)$ has a number of advantages as a deformation space. Not only is it simply connected but since Z^n is contractible[5] the long exact sequence associated with the fiber bundle projection,

$$a:\mathcal{R}(Z^n)\rightarrow Z^n,$$

shows that all the higher homotopy groups of $\mathcal{R}(Z^n)$ are also trivial and hence that $\mathcal{R}(Z^n)$ is contractible. Since Z^n is the complement of the zero set of an an analytic function on $\mathbf{C}^n$, it is a Stein space, and since $\mathcal{R}(Z^n)$ is an unramified covering of Z^n it too is a Stein space.[6] The consequent triviality of the sheaf cohomology $H^1(\mathcal{R}(Z^n),\mathcal{O}^*)$ plays a role in giving a global definition of tau functions following Ref. 4. Finally, although it will not be the principal focus of the deformations we consider in this paper, we indicate the crucial property used to construct the Schlesinger deformations considered in Refs. 3, 4. Let Y_k denote the subset of $\mathbf{P}^1\times\mathcal{R}(Z^n)$, given by

$$Y_k=\{(x,t)|x=a_k(t)\}.$$

Let

$$Y_\infty=\{(\infty,t)|t\in\mathcal{R}(Z^n)\},$$

and

$$Y=Y_\infty\cup Y_1\cup Y_2\cup\ldots\cup Y_n. \tag{1.1}$$

Then the property that is exploited in Ref. 3 to construct the Schlesinger deformations, is that for any choice of $t_0\in\mathcal{R}(Z^n)$, the injection

$$\mathbf{P}^1\backslash\{a_1^0,a_2^0,\ldots a_n^0,\infty\}\ni x\rightarrow(x,t^0)\in\mathbf{P}^1\times\mathcal{R}(Z^n)\backslash Y$$

induces an isomorphism of fundamental groups (where $a_j^0=a_j(t^0)$ is the jth coordinate of the projection). Using the correspondence between flat connections and representations of the fundamental group from Ref. 7, this allows one to prolong the original connection from $\mathbf{P}^1\backslash\{a_1^0,a_2^0,\ldots,a_n^0,\infty\}$ to $\mathbf{P}^1\times\mathcal{R}(Z^n)\backslash Y$ in a holonomy preserving fashion. The crux of the existence proof for the Schlesinger deformations is then to extend this connection to a neighborhood of Y so that it has *logarithmic poles* along Y. In Ref. 3 (and by a similar construction in Ref. 4) this is accomplished by exhibiting local deformations in a neighborhood of each Y_k and then proving that these can be fit together to provide a solution to the deformation problem if and only if a certain Fredholm integral equation has a solution. We will consider a related construction later on in this paper.

Some further analysis then leads to the existence of a holomorphic function τ defined on $\mathcal{R}(Z^n)$ whose 0 set is an exceptional set for the solution of the original deformation problem. In this paper we wish to use similar constructions to examine a somewhat different class of deformations. Our goal will be to show that the τ function introduced by Jimbo, Miwa, and Ueno[8] in a study of such deformations has a similar property. Namely, that the zeros of these tau functions are exceptional sets for the solution of a deformation problem.

B. Birkhoff deformations

The deformations we wish to consider are associated with Birkhoff's generalization of the Riemann–Hilbert problem.[9] To understand what is involved, it is useful to review the analysis of solutions to linear differential equations,

$$\frac{d\psi}{dx}=A(x)\psi,$$

with a rational matrix-valued coefficient $A(x)$ in a neighborhood of an irregular singular point. A helpful modern review of this subject can be found in Ref. 10. Suppose that $x=a$ is a pole of $A(x)$ of order $r+1$; then we have

$$A(x)=A_r(x-a)^{-r-1}+A_{r-1}(x-a)^{-r}+\cdots.$$

The integer r is called the rank of the singularity and for the moment we will confine our attention to the case $r\geqslant 1$. The theory of *formal* solutions to (1.0) is much simplified by one further assumption that we will make from now on. We require the following.

Standing Assumption (1.2). The coefficient, A_r, of the leading singularity in the Laurent expansion of $A(x)$ at $x=a$ has distinct eigenvalues.

This assumption, of course, guarantees that A_r can be diagonalized by a nonsingular matrix G,

$$G^{-1}A_rG=\Lambda_r,$$

where Λ_r is a diagonal matrix with distinct complex entries. In such circumstances (1.0) has a unique *formal* fundamental solution,

$$\hat{\Psi}(x,t)=G\hat{\alpha}(x)e^{H(x)},$$

where $\hat{\alpha}(x)$ is an $n\times n$ matrix-valued formal power series,

$$\hat{\alpha}(x)=I+\beta_1(x-a)^1+\beta_2(x-a)^2+\cdots$$

and

$$H(x)=\Lambda_r\frac{(x-a)^{-r}}{-r}+\Lambda_{r-1}\frac{(x-a)^{-r+1}}{-r+1}\cdots+\Lambda_1\frac{(x-a)^{-1}}{-1}+\Lambda_0\log(x-a),$$

where all the matrices Λ_k are diagonal with diagonal entries,

$$\Lambda_{k,j}:=(\Lambda_k)_{jj}.$$

The construction of this formal solution hinges on the inversion of $\mathrm{ad}(A_r)$ acting on the off-diagonal matrices. Since the eigenvalues of $\mathrm{ad}(A_r)$ acting on the off-diagonal matrices are differences of distinct eigenvalues for A_r, our standing assumption guarantees that this can be done. It is quite typical that the series for $\hat{\alpha}$ does not converge and this leads to some complication in making a connection between the formal solution to (1.0) and genuine solutions to this equation. Before we turn to this matter we mention a slightly different way of looking at this result that will make it simpler for the reader to connect this way of thinking with the developments in Refs. 3, 4, and 10.

Now write $\partial=\partial/\partial x$ and $\overline{\partial}=\partial/\partial\overline{x}$ and instead of the differential equation (1.0) one regards the connection

$$dx\otimes(\partial-A(x))+d\overline{x}\otimes\overline{\partial}, \tag{1.3}$$

on the trivial bundle

$$\mathbf{P}^1\times\mathbf{C}^p\rightarrow\mathbf{P}^1,$$

as the fundamental object. If $\{a_1,a_2,\dots,a_n\}$ is the set of poles for $A(x)$, then flat sections, ψ, for (1.3) defined locally in $\mathbf{P}^1\backslash\{a_1,a_2,\dots,a_n\}$ are solutions to the differential equation (1.0).

Gauge transformations (e.g., multiplication by smooth invertible matrix-valued functions of x) which are holomorphic (or meromorphic) in x do not change the $d\overline{x}\otimes\overline{\partial}$ part of the connection (at least away from the singularities) and the solution of the differential equation (1.0) is effectively accomplished by gauging $\partial - A(x)$ into diagonal form by such a gauge transformation. The formal gauge substitution,

$$\psi \leftarrow G\hat{\alpha}\psi,$$

can then be seen to reduce the connection (1.3) to the diagonal form

$$dx\otimes(\partial - h(x)) + d\overline{x}\otimes\overline{\partial},$$

where $h(x) := dH/dx$. A fundamental solution to the differential equation

$$(\partial - h(x))\psi = 0, \tag{1.4}$$

is given by

$$\Psi = e^{H(x)},$$

which is well defined modulo the possible appearance of a multivalued log term in $H(x)$. This accounts for the structure of the formal fundamental solution above but the relation between this formal solution and a genuine fundamental solution is complicated by Stokes' phenomena, which we will now describe.

We first adopt some notation and definitions from Ref. 10. Σ will denote a sector with a vertex at the origin, consisting of points $re^{i\theta}$ with $r>0$ and argument $\theta\in(a,b)$ with $0\leqslant a<b<2\pi$. For $\delta>0$, we write Σ_δ for the subset of Σ with $r<\delta$. If Σ and Σ' are two sectors we write $\Sigma'\subset\subset\Sigma$ if the bounding rays for Σ' are contained in Σ. An open set $\Omega\subset\Sigma$ is said to be asymptotic to the sector Σ, if for each $\Sigma'\subset\subset\Sigma$ we have $\Sigma'_\delta\subset\Omega$ for all sufficiently small δ. We introduce $\mathcal{A}(\Sigma)$, the complex algebra of germs of analytic functions defined on open sets asymptotic to Σ consisting of functions that are asymptotic to a formal meromorphic series. The asymptotic condition on $f\in\mathcal{A}(\Sigma)$ is understood to mean that there exists a formal series,

$$\hat{f}(x) \sim \sum_{k\geqslant m} f_k x^k,$$

with $m>-\infty$, so that for any $\Sigma'\subset\subset\Sigma$ and any integer M one has

$$f(x) = \sum_{k\geqslant m}^{M} f_k x^k + \mathcal{O}(|x|^{M+1}), \quad \text{as} \quad x\rightarrow 0 \quad \text{in} \quad \Sigma'.$$

The basic local existence result for solutions of (1.0) near a pole $x=a$ of $A(x)$ (and we take $a=0$ for convenience) with leading singularity diagonalized by G is that if a sector Σ is chosen appropriately, there is a function (or germ) $\alpha_\Sigma\in\mathcal{A}(\Sigma)$ that is asymptotic to $\hat{\alpha}$ with the property that the gauge transformation by $\alpha_\Sigma^{-1}G^{-1}$ reduces the differential equation (1.0) to the diagonal form (1.4) in an open set Ω asymptotic to Σ. For simplicity in the following local analysis, we will suppose that a trivialization has been chosen so that $G=I$. To describe the nonuniqueness for α_Σ, upon which the Stokes' phenomena hinges, we introduce

$$\alpha_{\Sigma,k} = k\text{th column of } \alpha_\Sigma,$$

and

$$H_k(x) = k\text{th diagonal entry of } H(x).$$

Another way to state the existence result mentioned above is that the n vector-valued functions,

$$\psi_k(x) = \alpha_{\Sigma,k}(x) e^{H_k(x)},$$

are independent solutions to (1.0) in some open set Ω asymptotic to the sector Σ. Because the functions $\alpha_{\Sigma,k}(x)$ are asymptotic to the power series, the ''growth'' of the functions $\psi_k(x)$ as $x \to 0$ is controlled by the exponential factors $e^{H_k(x)}$, which have absolute value $e^{\Re H_k(x)}$ (where $\Re x =$ real part of x). Now let

$$\Delta H_{jk}(x) = \Re(H_j(x) - H_k(x)).$$

The curves along which

$$\Delta H_{jk}(x) = 0, \tag{1.5}$$

play an important role in understanding the relationship between formal solutions and analytic solutions near the singularity at $x=0$. To get an idea of what such curves look like near 0, it is enough to consider the leading-order equivalent of (1.5). This is

$$\Re(\Lambda_{r,j} - \Lambda_{r,k}) x^{-r} = 0. \tag{1.6}$$

Since each difference $\Lambda_{r,j} - \Lambda_{r,k} \neq 0$, it follows that there are $2r$ rays emanating from 0 that satisfy (1.6), given by

$$\arg x = \frac{1}{r}\left(\arg(\Lambda_{r,j} - \Lambda_{r,k}) + \left(n + \frac{1}{2}\right)\pi\right), \quad \text{for } n = 0,...,2r-1.$$

These rays are called Stokes lines. Near $x=0$ the family of solutions to (1.5) consists of $2r$ curves, each asymptotic to one of the Stokes lines (1.6). The property of the Stokes lines that will be important for us is that any open sector that contains one of the Stokes lines (1.6) will contain points with $\Delta H_{jk} < 0$ and also points with $\Delta H_{jk} > 0$.

Now suppose that one crosses such a Stokes line going from $\Delta H_{jk} < 0$ to $\Delta H_{jk} > 0$. One moves from a region in which $\psi_k(x)$ dominates $\psi_j(x)$ as $x \to 0$ to a region in which this dominance relation is reversed. For the purpose of illustration assume that $\Delta H_{jk}(x) < 0$ for x in some truncation Σ_δ of Σ. Then for any constant c, one finds

$$\psi_k + c\psi_j = (\alpha_{\Sigma,k} + c\alpha_{\Sigma,j} e^{H_j - H_k}) e^{H_k},$$

and because $e^{H_j - H_k}$ is exponentially small in Σ_δ, it follows that

$$\alpha_{\Sigma,k} + c\alpha_{\Sigma,j} e^{H_j - H_k} \sim \hat{\alpha}_k,$$

where $\hat{\alpha}_k$ is the kth column of the formal series $\hat{\alpha}$. Thus, the less dominant solution ψ_j may be freely mixed in with ψ_k without affecting the asymptotics of $\alpha_{\Sigma,k}$. In this way one sees that genuine solutions with a given exponential behavior are not uniquely determined by the asymptotics of their ''power series component.'' One obvious way to ''cure'' this particular nonuniqueness would be to include a Stokes line from the family $\Delta H_{jk} = 0$ in the sector Σ. The exchange of dominance across the line makes it impossible to alter ψ_k by adding in multiples of ψ_j without changing the asymptotics of $\alpha_{\Sigma,k}$, and *vice versa*. In fact, if one includes *exactly one Stokes line from each of the families (1.6) for* $j<k$, then a simple argument[11] shows that a genuine fundamental solution,

$$\Psi = \alpha_\Sigma e^H,$$

in Σ_δ, is uniquely determined by the condition that the asymptotic expansion of α_Σ is given by $\hat{\alpha}$. What is more, there is also an existence result for such sectors that can be proved using a variant of the usual integral equation technique but employing different contours for the each of the matrix elements in the solution (see the references in Ref. 12). Following Ref. 13 we call such sectors *good sectors*. It is not hard to see that a punctured neighborhood of 0 can be covered by $2r$ (truncated) good sectors $\Sigma_{i,\delta}$ that we will take to be arranged in counterclockwise order, starting with $\Sigma_{1,\delta}$. Because each good sector contains exactly one Stokes line from each family (1.5), it follows that the intersections $\Sigma_{i,\delta}\cap\Sigma_{i+1,\delta}$ do not contain any Stokes lines. On this overlap the two local fundamental solutions $\alpha_{\Sigma_i}e^H$ and $\alpha_{\Sigma_{i+1}}e^H$ of (1.0) necessarily differ by a constant invertible $p\times p$ matrix $S_{i,i+1}$,

$$\alpha_{\Sigma_{i+1}}e^H=\alpha_{\Sigma_i}e^H S_{i,i+1}. \tag{1.7}$$

The matrices $S_{i,i+1}$ are called Stokes multipliers and must satisfy a triangularity property that we will now explain. Since $\Sigma_{i,\delta}\cap\Sigma_{i+1,\delta}$ does not contain any Stokes lines, it follows that for each fixed choice of (j,k) the quantity ΔH_{jk} is either always positive or always negative in $\Sigma_{i,\delta}\cap\Sigma_{i+1,\delta}$ (at least if δ is small enough so that the Stokes lines are good ''stand ins'' for the curves $\Delta H_{jk}=0$). Thus there is a fixed dominance ordering,

$$\Re\Lambda_{r,i_1}x^{-r}>\Re\Lambda_{r,i_2}x^{-r}>\cdots>\Re\Lambda_{r,i_n}x^{-r},$$

for $x\in\Sigma_{i,\delta}\cap\Sigma_{i+1,\delta}$, and some permutation $(i_1,i_2,\ldots,i_n)$ of 1,2,....,n. If we write (1.7) in matrix form relative to the ordered basis $\{e_{i_1},e_{i_2},\ldots,e_{i_n}\}$, then the matrix of $S_{i,i+1}$ relative to this ordered basis must be lower triangular with 1's on the diagonal in order that $S_{i,i+1}$ should only alter α_{Σ_i} by exponentially small terms. Another (basis-independent) version of the same observation is that

$$e^H S_{i,i+1}e^{-H}=I+O(|x|^N),\quad \text{for } x\to 0 \quad \text{in} \quad \Sigma_{i,\delta}\cap\Sigma_{i+1,\delta},$$

for some $\epsilon>0$ and all positive integers N. The relation (1.7) allows us to construct a fundamental solution Ψ to (1.0) in a neighborhood of $x=a$ by analytically continuing the fundamental solution $\alpha_{\Sigma_1}e^H$ from $\Sigma_{1,\delta}$ to $\Sigma_{2,\delta}$ to $\Sigma_{3,\delta}$ and etc. The result is

$$\Psi(x)=\begin{cases}\alpha_{\Sigma_1}(x)e^{H(x)}, & \text{for } x\in\Sigma_{1,\delta},\\ \alpha_{\Sigma_2}(x)e^{H(x)}S_{1,2}^{-1}, & \text{for } x\in\Sigma_{2,\delta},\\ \cdots, & \\ \alpha_{\Sigma_{2r}}(x)e^{H(x)}S_{1,2r}^{-1}, & \text{for } x\in\Sigma_{2r,\delta},\end{cases} \tag{1.8}$$

where we have written

$$S_{1,k}=S_{1,2}S_{2,3}\cdots S_{k-1,k},$$

and it is understood that the logarithmic term in $H(x)$ is analytically continued from Σ_1 to Σ_2 to $\cdots$ to Σ_{2r}. If we write $S_{2r,1}$ for the Stokes multiplier connecting Σ_{2r} with Σ_1 then it is not hard to see that the analytic continuation of $\alpha_{\Sigma_1}(x)e^{H(x)}$ around $x=a$ comes back to its original value multiplied on the right by

$$e^{H(e^{2\pi i}x)-H(x)}(S_{1,2r}S_{2r,1})^{-1}=e^{2\pi i\Lambda_0}(S_{1,2r}S_{2r,1})^{-1}. \tag{1.9}$$

The matrix (1.9) is thus the local monodromy for the fundamental solution (1.8). We will refer to the *exponent of formal monodromy* Λ_0, together with the Stokes multipliers $S_{i,i+1}$ as the generalized monodromy data for (1.0) at $x=a$. Roughly speaking the deformations of (1.0) we are interested in are those that fix the local generalized monodromy data at each of the singularities for (1.0) and fix the global monodromy for (1.0) but permit the formal expansion coefficients $\{\Lambda_r,\Lambda_{r-1},\ldots,\Lambda_1\}$ at each singularity to vary. The global monodromy is precisely the representation of the fundamental group described earlier and we will say more about this later on, following the presentation in Malgrange[3] and Helmink.[4] Generalizations of the Schlesinger deformations in which the location of the poles are varied are also quite interesting; however, the issues we wish to pursue have already been treated for these deformations in Refs. 3 and 4, and so for the moment we confine our attention to deformations of the local exponents $\{\Lambda_r,\Lambda_{r-1},\ldots,\Lambda_1\}$. Our strategy in studying these deformations will follow closely the developments in Ref. 14. In fact, in Ref. 14 Malgrange already proves existence results for such deformations in the irregular singular case. However, we did not understand how to make use of his results to establish the connection with the JMU tau function that is the principal object in this paper. Instead, we will adapt an integral equation technique from Flaschka and Newell[15] to produce local models for the desired deformation at each of the poles. These local deformations are then fit together by solving the same Toeplitz integral equations that one finds in Ref. 3. The tau function is introduced by identifying its log derivative as the connection one-form for a flat connection on an appropriate determinant bundle. A computation shows that this connection one-form differs from the JMU connection one-form by a regular term, and so the tau function we have introduced and the JMU tau functions have the same 0 set. We show that this 0 set is precisely the exceptional set for the existence of the deformations we are considering.

C. Local analysis and Stokes multipliers

It will be useful at this point to be a little more precise about the nature of the generalized local monodromy data that is to be ''fixed'' under the deformations that are of interest to us. These deformations concern the local model for our connections. Suppose that one starts with a holomorphic connection $\bar{\boldsymbol{\nabla}}^0$ defined on a trivial bundle over a punctured neighborhood of the point $a\in\mathbf{P}^1$ with a singularity of type r at a (*note*: we will put a bar over connections defined on subsets of $\mathbf{P}^1$ to distinguish them from the connections in many variables that will soon appear as deformations). Suppose that the connection satisfies our standing assumption and that the trivialization is chosen so that the leading singularity in the one-form for $\bar{\boldsymbol{\nabla}}^0$ has a diagonal matrix coefficient. Then $\bar{\boldsymbol{\nabla}}^0$ is formally gauge equivalent to the diagonal form,

$$d_x - d_x\mathbf{H}_0 - \Lambda_0\frac{dx}{x-a}, \tag{1.10}$$

where

$$\mathbf{H}_0=\sum_{j=1}^{r}\frac{\Lambda_j^0}{-j}(x-a)^{-j}, \tag{1.11}$$

where Λ_j^0 and Λ_0 are diagonal matrices. More precisely there is a formal gauge transformation,

$$\hat{\alpha}^0(x)=I+\beta_1^0(x-a)+\beta_2^0(x-a)^2+\cdots,$$

so that

$$\bar{\boldsymbol{\nabla}}^0=\hat{\alpha}^0\cdot\left[d_x-d_x\mathbf{H}_0-\Lambda_0\frac{dx}{x-a}\right],$$

where $\hat{\alpha}\cdot[X]=\hat{\alpha}[X]\hat{\alpha}^{-1}$. Note that it is actually the inverse of $\hat{\alpha}^0$ that reduces $\bar{\boldsymbol{\nabla}}^0$ to diagonal form. This is just how things work out if the relationship between $\hat{\alpha}^0$ and a fundamental solution to $\bar{\boldsymbol{\nabla}}^0$ is given along the lines explained above.

The local analytic equivalence class of $\bar{\boldsymbol{\nabla}}^0$ is determined by further data, which can be specified by choosing a covering of the punctured neighborhood of a by good sectors, $\Sigma_1,\Sigma_2,\ldots,\Sigma_{2r}$ (to somewhat unburden the notation we will write Σ_k for the truncated sector $\Sigma_{k,\delta}$ when the precise value of δ is not an issue). For simplicity in the following discussion we will always suppose that such a covering is obtained by first choosing a good sector Σ_1. The other sectors Σ_j are obtained from Σ_1 by rotating counterclockwise by $j\pi/r$ radians. The Stokes multipliers $S_{j,j+1}$ that connect local fundamental solutions with fixed asymptotics in the sectors Σ_j and Σ_{j+1} (with $\Sigma_{2r+1}=\Sigma_1$) then determine the local holomorphic equivalence class of $\bar{\nabla}^0$. The deformations we wish to focus on allow the local model

$$\bar{\nabla}_\Lambda := d_x - d_x\mathbf{H} - \Lambda_0 \frac{dx}{x-a}, \tag{1.12}$$

with

$$\mathbf{H} = \sum_{j=1}^{r} \frac{\Lambda_j}{-j}(x-a)^{-j}, \tag{1.13}$$

to vary in the sense that the diagonal coefficient matrices $\Lambda_r, \Lambda_{r-1}, \ldots, \Lambda_1$ vary in a neighborhood of $\Lambda_r^0, \Lambda_{r-1}^0, \ldots, \Lambda_1^0$ with Λ_r maintaining distinct eigenvalues and Λ_0 held fixed. In what follows Λ_0 will always denote a fixed diagonal $p\times p$ matrix.

The Stokes multipliers are to remain fixed under our deformations. However, since the Stokes multipliers are significant relative to some choice of a covering by good sectors $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_{2r}\}$ we must either fix this covering for all values of the deformation parameters (this is what is done in Refs. 11 and 8) or say how the choice of covering affects the notion of ''fixed'' Stokes multipliers. We follow Malgrange in adopting the second alternative. The choice of a good sector Σ_1 is nearly equivalent to the selection of a suitable collection of ''consecutive'' Stokes lines [one from each of the families (1.6)]. Once one has selected such a collection of Stokes lines any open sector that contains these Stokes lines could serve as a good sector. However, we wish to put a further limitation on our good sectors. No Stokes line should lie on the boundary of a good sector. The reason for this is that for a fixed connection, the Stokes multipliers associated with such a good sector are not stable under small rotations of the sector; Stokes lines can rotate in or out of the sector and this changes the associated Stokes multipliers. Another way of saying this is that the local analytic equivalence class associated with a choice of a covering by good sectors and associated Stokes multipliers is not stable under small rotations of the good sectors unless the sectors do not have Stokes lines on their boundary. We will say that a covering $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_{2r}\}$ of a punctured neighborhood of a by good sectors is *stable* if the good sector Σ_1 has no Stokes lines on its boundary.

We now define a configuration space, $\mathcal{C}$, for our local models (1.12):

$$\mathcal{C} := Z^p \times \mathbf{C}^p \times \cdots \times \mathbf{C}^p,$$

where there are $r-1$ factors $\mathbf{C}^p$. The first factor Z^p is, of course, the configuration space for the leading coefficient Λ_r, and the remaining factors $\mathbf{C}^p$ are associated with the Λ_j, $j=r-1,\ldots,1$. Following Malgrange, we now introduce a fiber bundle, $\mathcal{M}\to\mathcal{C}$. The fiber $\mathcal{M}_\Lambda$ over each point $\Lambda := (\Lambda_r, \Lambda_{r-1}, \ldots, \Lambda_1)$ in $\mathcal{C}$ is the moduli space of all holomorphic gauge equivalence classes of locally defined type r connections on the trivial bundle $\{x : |x-a| < \epsilon\} \times \mathbf{C}^p$ (for some $\epsilon > 0$) that are formally equivalent to the connection (1.12)–(1.13). Or put another way, if $\bar{\nabla}_1$ and $\bar{\nabla}_2$ are two type r connections at a and they are formally equivalent to $\bar{\nabla}_\Lambda$ via $\hat{\alpha}_1$ and $\hat{\alpha}_2$, then $\bar{\nabla}_1 \simeq \bar{\nabla}_2$ if $\hat{\alpha}_2^{-1}\hat{\alpha}_1$ is convergent in a neighborhood of a (see Ref. 14). Next, we want to show that

$$\mathcal{M} \xrightarrow{\pi} \mathcal{C}, \tag{1.14}$$

is a fiber bundle with a natural flat connection. This connection will play a role in providing a global significance for monodromy preserving deformations. We first discuss local trivializations for (1.14). Suppose that $\Lambda^0 \in \mathcal{C}$ and write $\bar{\nabla}^0$ for the connection (1.12) associated with Λ^0.

Suppose that $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_{2r}\}$ is a covering of a punctured neighborhood of a by stable good sectors for $\bar{\boldsymbol{\nabla}}^0$. Then we can find a neighborhood U of Λ^0 in $\mathcal{C}$ so that $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_{2r}\}$ is a stable covering at a for all $\bar{\boldsymbol{\nabla}}_\Lambda$ with $\Lambda\in U$ (in fact, it is clear that U can be taken to be of the form $U_r\times\mathbf{C}^p\times\ldots\times\mathbf{C}^p$, where U_r is a sufficiently small neighborhood of Λ_r^0 in Z^p). It is a result of Malgrange and Sibuya,[14,16] that for any $\Lambda\in U$ and suitable choice of Stokes multipliers $S_{j,j+1}$ there exists a type r connection at a with local model $\bar{\boldsymbol{\nabla}}_\Lambda$ and Stokes multipliers $S_{j,j+1}$ relative to the covering $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_{2r}\}$. This connection is not unique but its local holomorphic gauge equivalence class is unique. To be ''suitable'' the Stokes multipliers must be 1 on the diagonal and lower triangular relative to the dominance ordering of $\Re(\Lambda_{r,j}x)$ for $x\in\Sigma_j\cap\Sigma_{j+1}$. This identifies the fiber of $\mathcal{M}$ as $\mathbf{C}^{rp(p-1)}$, since there are $2r$ intersections $\Sigma_j\cap\Sigma_{j+1}$ and $p(p-1)/2$ arbitrary coefficients for each Stokes multiplier. The choice of a neighborhood U and a covering $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_{2r}\}$, which is stable for all $\boldsymbol{\nabla}_\Lambda$ with $\Lambda\in U$, thus produces a trivialization,

$$\pi^{-1}(U)\simeq U\times\mathbf{C}^{rp(p-1)}.$$

Note: The fiber $\mathcal{M}_\Lambda$ is more invariantly defined in Ref. 14, as $H^1(S^1,\mathrm{St}(\bar{\boldsymbol{\nabla}}_\Lambda))$, the first cohomology of the Stokes sheaf associated with the connection $\bar{\boldsymbol{\nabla}}_\Lambda$ (see also Ref. 10).

We will now show that $\mathcal{M}$ is a fiber bundle by determining that the transition maps between trivializations are given by polynomial diffeomorphisms in the fiber. Suppose that U' is a neighborhood in $\mathcal{C}$ with $\{\Sigma_1',\Sigma_2',\ldots,\Sigma_{2r}'\}$ a covering by sectors at a that is stable for all $\bar{\boldsymbol{\nabla}}_\Lambda$ with $\Lambda\in U'$. Now suppose that $\Lambda\in U\cap U'$. Then since $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_{2r}\}$ and $\{\Sigma_1',\Sigma_2',\ldots,\Sigma_{2r}'\}$ are both good stable coverings for $\bar{\boldsymbol{\nabla}}_\Lambda$ it follows that, up to small changes in the opening angle of Σ_1', which do not affect the Stokes multipliers, Σ_1' can be obtained from Σ_1 by rotating Σ_1 counterclockwise through an angle θ. To emphasize this we will write $\Sigma_j'=\Sigma_j^\theta$. Now let $\bar{\boldsymbol{\nabla}}$ denote a connection of type r at a that is formally equivalent to $\bar{\boldsymbol{\nabla}}_\Lambda$ [i.e., $\pi(\bar{\boldsymbol{\nabla}})=\Lambda$]. Suppose $\hat{\alpha}$ is a formal series with

$$\bar{\boldsymbol{\nabla}}=\hat{\alpha}\cdot[\bar{\boldsymbol{\nabla}}_\Lambda],$$

in the sense of formal series at a. Suppose that $\alpha_j^\theta\in\mathcal{A}(\Sigma_j^\theta)$ is asymptotic to $\hat{\alpha}$ and $\bar{\boldsymbol{\nabla}}=\alpha_j^\theta\cdot[\bar{\boldsymbol{\nabla}}_\Lambda]$ analytically in the sector Σ_j^θ. Then the Stokes multipliers $S_{j,j+1}^\theta$ are defined by,

$$\alpha_{j+1}^\theta(x)e^{H_\Lambda(x)}=\alpha_j^\theta(x)e^{H_\Lambda(x)}S_{j,j+1}^\theta, \tag{1.15}$$

where

$$H_\Lambda(x)=\sum_{j=1}^{r}\frac{\Lambda_j}{-j}(x-a)^{-j}+\Lambda_0\log(x-a). \tag{1.16}$$

There is a slight ambiguity in the definition of the Stokes multipliers in (1.15) associated with the choice of $x\to\log(x-a)$ in (1.16). To deal with this ambiguity we require that the data that specifies a local trivialization for $\mathcal{M}\to\mathcal{C}$ includes not only the choice of a stable good covering $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_{2r}\}$ but also a branch of the function $x\to\log(x-a)$ in the sector Σ_1. One may analytically continue this choice from Σ_1 to Σ_2 to Σ_3 and so on to fix a choice of log in (1.16) and render (1.15) an unambiguous defining relation for $S_{j,j+1}^\theta$. In the rank $r=1$ case there are only two sectors Σ_1 and Σ_2 and the intersection $\Sigma_1\cap\Sigma_2$ is disconnected. In this case we suppose that the analytic continuation from Σ_1 to Σ_2 is accomplished so that the function is smooth on counterclockwise oriented circles passing from Σ_1 into Σ_2.

In the special case $\theta=0$ we will simply write $\alpha_j^0=\alpha_j$ and $S_{j,j+1}^0=S_{j,j+1}$.

Now we turn to the proof that $\mathcal{M}\to\mathcal{C}$ is a fiber bundle. As noted above our trivializations depend on a choice of $\log(x-a)$ in Σ_1^θ. However, different choices will simply alter $S_{j,j+1}^\theta$ by conjugation with powers of $\exp(2\pi i\Lambda_0)$. Since this is a linear transformation in the fiber constant in the base, we may as well suppose that this choice of $\log(x-a)$ has been fixed in Σ_1^θ (and Σ_1). The Stokes multipliers (1.15) are thus well defined and the transition map we wish to compute takes $\{S_{j,j+1}\}$ to $\{S_{j,j+1}^\theta\}$. It is enough to compute the transition in the fiber for $0<\theta<\pi/r$ since a larger rotation may be realized as a composition of rotations satisfying this condition. Suppose now that $0<\theta<\pi/r$. Because the rotation θ is smaller than π/r, it follows that $\Sigma_j\cap\Sigma_j^\theta$ is not empty. We may thus compare the local fundamental solutions $\alpha_j e^{H_\Lambda}$ and $\alpha_j^\theta e^{H_\Lambda}$ on this intersection,

$$\alpha_j e^{H_\Lambda}=\alpha_j^\theta e^{H_\Lambda}S_j(\theta), \tag{1.17}$$

where $S_j(\theta)$ is a constant $p\times p$ matrix. Combining (1.16) and (1.17), one finds that

$$S_{j,j+1}^\theta=S_j(\theta)S_{j,j+1}S_{j+1}(\theta)^{-1}. \tag{1.18}$$

Thus, to find $S_{j,j+1}^\theta$ in terms of the Stokes multipliers $\{S_{k,k+1}\}$ it will suffice to determine $S_k(\theta)$ in terms of $\{S_{k,k+1}\}$.

By relabling the eigenvalues, $\Lambda_{r,j}j=1,...,p$, of Λ_r we may suppose that the dominance ordering in $\Sigma_j\cap\Sigma_{j+1}$ is the ‘‘natural’’ one,

$$\Re(\Lambda_{r,1}x)<\Re(\Lambda_{r,2}x)<\cdots<\Re(\Lambda_{r,p}x),$$

for $x\in\Sigma_j\cap\Sigma_{j+1}$. It is not hard to see what happens to this dominance ordering as one rotates $\Sigma_j\cap\Sigma_{j+1}$ counterclockwise. As $\Sigma_j\cap\Sigma_{j+1}$ crosses a ‘‘simple’’ Stokes line, say

$$\Re(\Lambda_{r,1}x)<\Re(\Lambda_{r,2}x)=\Re(\Lambda_{r,3}x)<\Re(\Lambda_{r,3}x),$$

then the dominance ordering permutes 2 and 3 leaving the rest of the indices in sequence. If one crosses a Stokes line with ‘‘higher multiplicity,’’ say

$$\Re(\Lambda_{r,1}x)<\Re(\Lambda_{r,2}x)=\Re(\lambda_{r,3}x)=\Re(\Lambda_{r,4}x)<\Re(\Lambda_{r,5}x),$$

then the (2, 3, 4) part of the ordering is inverted to (4, 3, 2).

For the purpose of illustration suppose that $p=5$ and that in going from $\Sigma_j\cap\Sigma_{j+1}$ to $\Sigma_j^\theta\cap\Sigma_{j+1}^\theta$ one passes through the simple Stokes line $\Re(\Lambda_{r,1}x)=\Re(\Lambda_{r,2}x)$ and the higher multiplicity Stokes line $\Re(\Lambda_{r,3}x)=\Re(\Lambda_{r,4}x)=\Re(\Lambda_{r,5}x)$. Then it is not hard to see that the Stokes multiplier for $\Sigma_j^\theta\cap\Sigma_{j+1}^\theta$ must have the following ‘‘triangularity,’’

$$S_{j,j+1}^\theta=\begin{bmatrix}1&*&0&0&0\\0&1&0&0&0\\ *&*&1&*&*\\ *&*&0&1&*\\ *&*&0&0&1\end{bmatrix}, \tag{1.19}$$

relative to the basis for $\mathbf{C}^5$ in which $S_{j,j+1}$ is lower triangular. The $*$'s represent possibly nonzero entries. Next consider (1.17). Since α_j and α_j^θ have the same asymptotics in the sector $\Sigma_j\cap\Sigma_j^\theta$, it follows that $S_j(\theta)$ must have 1's on the diagonal and cannot have nonzero off-diagonal elements for any of the pairs associated with Stokes lines in the intersection $\Sigma_j\cap\Sigma_j^\theta$. In the example (1.19) this means that the (l,m) matrix elements for $S_j(\theta)$ are zero for

$$(l,m)\in\{(k,1),(k,2),(1,k),(2,k):k=3,4,5\}.$$

Since $\theta < \pi/r$ it follows that $\Sigma_j \cap \Sigma_{j+1}$ is contained in $\Sigma_j \cap \Sigma_j^\theta$ (provided at least one Stokes line is crossed, which is the only interesting case). Thus one can also say of $S_j(\theta)$ that it is lower triangular with respect to the dominance ordering in $\Sigma_j \cap \Sigma_{j+1}$. Thus, for our example, the matrix of $S_j(\theta)$ must have the form

$$S_j(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ a & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & b & 1 & 0 \\ 0 & 0 & c & d & 1 \end{bmatrix}. \tag{1.20}$$

In a similar fashion $\Sigma_j^\theta \cap \Sigma_{j+1}^\theta$ is contained in $\Sigma_{j+1} \cap \Sigma_{j+1}^\theta$, and it follows that $S_{j+1}(\theta)$ must be lower triangular with respect to the dominance ordering for $\Sigma_j^\theta \cap \Sigma_{j+1}^\theta$. In our example, if we rewrite (1.17) in the form $S_{j,j+1}^\theta S_{j+1}(\theta) = S_j(\theta) S_{j,j+1}$ and make use of the lower triangularity of the product $S_{j,j+1}^\theta S_{j+1}(\theta)$ with respect to the dominance ordering in $\Sigma_j^\theta \cap \Sigma_{j+1}^\theta$. We find

$$\begin{bmatrix} 1 & * & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ * & * & 1 & * & * \\ * & * & 0 & 1 & * \\ * & * & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ a & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & b & 1 & 0 \\ 0 & 0 & c & d & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ * & 1 & 0 & 0 & 0 \\ * & * & 1 & 0 & 0 \\ * & * & * & 1 & 0 \\ * & * & * & * & 1 \end{bmatrix}. \tag{1.21}$$

The (2,1) matrix element of $S_j(\theta)$ can be determined simply by equating the (2,1) matrix elements on both sides of (1.21). One finds

$$a = S_j(\theta)_{2,1} = -(S_{j,j+1})_{2,1}.$$

The same thing can be done for the matrix elements b and d for $S_j(\theta)$ on the subdiagonal. For example,

$$d = S_j(\theta)_{5,2} = -(S_{j,j+1})_{5,2}.$$

Once one knows the subdiagonal elements one can move out to the diagonal below the subdiagonal. Equating (5,3) matrix elements for (1.21), one finds

$$0 = c + d(S_{j,j+1})_{4,3} + (S_{j,j+1})_{5,3}.$$

From the earlier relation for d one finds that c is a polynomial function of the matrix elements of $S_{j,j+1}$. Thus, the entries of $S_j(\theta)$ are polynomials in the entries for $S_{j,j+1}$ and it is clear that this is true quite generally and does not depend on anything special in our example. From (1.18) it then follows that $S_{j,j+1}^\theta$ is a polynomial in the entries of $S_{j,j+1}$ and $S_{j+1,j+2}$. Since this relation is invertible by construction it follows that the map from $\{S_{k,k+1}\}$ to $\{S_{k,k+1}^\theta\}$ is a polynomial diffeomorphism on $\mathbf{C}^{rp(p-1)}$.

We have seen that π: $\mathcal{M} \to \mathcal{C}$ is a fiber bundle and that there is a natural family of trivializations for this bundle that are related by polynomial diffeomorphisms in the fiber that are constant in the base. Recall that the vertical vectors in the tangent space to $\mathcal{M}$ at $p, T_p(\mathcal{M})$, are those killed by $d\pi_p$. A connection on $\mathcal{M}$ is determined by a one-form ω on $\mathcal{M}$ whose value $\omega_p(v)$ at a vector $v \in T_p(\mathcal{M})$ is a vertical vector in $T_p(\mathcal{M})$ at p; ω must have the further property that $\omega_p(v) = v$ if v is a vertical vector in $T_p(\mathcal{M})$. Our trivializations single out a flat connection on $\mathcal{M}$. If $\pi^{-1}(U)$ has trivialization $U \times \mathbf{C}^N$ and $f_1, f_2, \ldots, f_N$ are the natural coordinate functions on $\mathbf{C}^N$ then it is easy to see that

$$\omega = \sum_{k=1}^{N} \frac{\partial}{\partial f_k} df_k$$

defines a connection one-form independent of the choice of trivialization (among the special class of trivializations that we have been considering for $\mathcal{M}$ in which the fiber coordinates transform among themselves). The curve $\hat{\gamma}(t)$ in $\mathcal{M}$ is the horizontal lift of $\gamma(t) = \pi\hat{\gamma}(t)$ provided that $\omega(\hat{\gamma}'(t)) = 0$. Relative to one of our distinguished trivializations this translates into

$$\hat{\gamma}(t) = (\gamma(t), f),$$

where $f \in \mathbf{C}^N$ is constant in t. The curvature of this connection is clearly 0 and locally parallel sections look like

$$\sigma(x) = (x, f),$$

where f is independent of x. Now let $\mathcal{R}(\mathcal{C})$ denote the simply connected covering space of $\mathcal{C}$. As was the case for $\mathcal{R}(Z^n)$ above it will be convenient to introduce special notation for projection from $\mathcal{R}(\mathcal{C})$ to $\mathcal{C}$. We will write $\lambda \in \mathcal{R}(\mathcal{C})$ and $\Lambda(\lambda) \in \mathcal{C}$ for the projection of λ onto $\mathcal{C}$.

The bundle $\mathcal{M}$ over $\mathcal{C}$ with flat connection ω pulls back to a bundle over $\mathcal{R}(\mathcal{C})$ (which we will continue to denote by $\mathcal{M}$) with a flat connection (which we will continue to denote by ω). Because $\mathcal{R}(\mathcal{C})$ is simply connected, the bundle

$$\mathcal{M} \to \mathcal{R}(\mathcal{C})$$

has globally defined parallel sections. The existence of these global sections, which are the analogs of the local notion of ''constant Stokes multipliers'' in $\mathcal{C}$ is the reason for working here in $\mathcal{R}(\mathcal{C})$ instead of just locally in $\mathcal{C}$ (*note*: the theory of local systems explained in Ref. 10 also applies to $\mathcal{M}$ and could be used to bypass the introduction of a connection in the fiber bundle $\mathcal{M}$).

D. Local models for integrable deformations

Before we state our global result for $\mathcal{R}(\mathcal{C})$ deformations we introduce some notation from Ref. 14 concerning connections in several complex variables. Suppose that X is a complex analytic variety of dimension n, Y is a smooth hypersurface in X, E is a rank p complex vector bundle over X, $\boldsymbol{\nabla}$ is a holomorphic connection on $X \backslash Y$, and Ω is the one form for $\boldsymbol{\nabla}$ in a local frame for E.

Definition 1.22: For $r \geqslant 0$ an integer one says that $\boldsymbol{\nabla}$ is of type r along Y, if in a system of local coordinates $x_1, x_2, \ldots, x_n$ for X with Y locally defined by $x_1 = 0$ one has

$$\Omega = \sum_{j=1}^{n} M_j \, dx_j,$$

where M_1 has a pole of order $r+1$ in x_1, M_j for $j \geqslant 2$ has a pole of order r in x_1, and M_j is holomorphic in $x_2, x_3, \ldots, x_n$ for all j.

Suppose that

$$M_1 = \sum_{-r-1}^{\infty} M_{1,k} x_1^k.$$

Our standing assumption (1.2) for type $r \geqslant 1$ connections, translates into this situation as the assumption that $M_{1,-r-1}(x_2, x_3, \ldots, x_n)$ has distinct eigenvalues. We will refer to such connections as *simple* type r connections along Y. When $r = 0$ we will say that a type 0 connection is *simple* if $M_{1,-1}(x_2, x_3, \ldots, x_n)$ has distinct eigenvalues, which, in addition, *do not differ by integers*. Malgrange has shown how to develop the theory of Schlesinger deformations without this assumption, but it will be convenient for us to insist on it so that we may work with regular and irregular singular points in parallel. Propositions 1.23a and 1.23b below are the principal reason that this is possible and will permit us to present the ''unified'' formula for $d \log \tau$, which can be found in Ref. 8.

We recall some results of Malgrange for simple *integrable* type r connections. Suppose that one has a simple integrable type r connection ∇ along Y and that, as above, Y is locally defined by $x_1=0$. Let $y=(x_2,x_3,\ldots,x_n)$ denote the local coordinates on Y. Then one has the following local version of Proposition 1.3 in Ref. 14.

Proposition 1.23a: Suppose that (E,∇*) is a vector bundle with a simple integrable connection* ∇ *of type* $r\geqslant 1$ *along Y. There exists a local trivialization in a neighborhood of* $x=0$ *so that* $M_{1,-r-1}$ *is diagonal in this trivialization. Let* Ω *denote the connection form for* ∇ *in this trivialization. Then in a sufficiently small neighborhood* $|y|<\epsilon$ *there exists a formal power series* $\hat{\alpha}$ *in* x_1,

$$\hat{\alpha}(x)=I+\beta_1(y)x_1+\beta_2(y)x_1^2+\cdots,$$

with matrix coefficients $\beta_j(y)$, *which are holomorphic in y for* $|y|<\epsilon$, *with the property that*

$$\hat{\alpha}\cdot[\nabla_\Lambda]=d+\Omega,$$

where $\hat{\alpha}\cdot[\nabla_\Lambda]=\hat{\alpha}\,\nabla_\Lambda\hat{\alpha}^{-1}$ *is the formal gauge transformation by* $\hat{\alpha}$,

$$\nabla_\Lambda=d-d(\mathbf{H})-\Lambda_0\frac{dx_1}{x_1},$$

d is the exterior derivative in the x variables, Λ_0 *is a constant diagonal matrix,*

$$\mathbf{H}=\sum_{j=1}^{r}\Lambda_j(y)\frac{x_1^{-j}}{-j},$$

with $\Lambda_j(y)$ *a diagonal matrix with entries that are holomorphic functions of y, and* $\Lambda_r(y)$ *a diagonal matrix with distinct entries [note: in the statement of this Proposition* Λ *does not denote the projection* $\mathcal{R}(\mathcal{C})\rightarrow\mathcal{C}$ *described above].*

There is an analog of this result for simple type 0 connections, which is the principal reason we work with such connections.

Proposition 1.23b: Suppose that (E,∇*) is a vector bundle with a simple integrable type 0 connection* ∇ *along Y. There exists a local trivialization near* $x=0$ *so that* $M_{1,-1}$ *is diagonal. Let* Ω *denote the one form for* ∇ *in this trivialization. Then in a sufficiently small neighborhood of* $x=0$ *there exists a* $\mathrm{GL}(p,C)$*-valued holomorphic function,*

$$\alpha(x)=I+\beta_1(y)x_1+\beta_2(y)x_1^2+\cdots,$$

so that

$$\alpha\left(d-\frac{\Lambda_0}{x_1}dx_1\right)\alpha^{-1}=d+\Omega,$$

where Λ_0 *is a diagonal matrix that is independent of x.*

Remark: The conclusion of this theorem is simply that the connection ∇ is given by $d-(\Lambda_0/x_1)dx_1$ in a suitable local trivialization. The analogy with Proposition 1.23a is not so apparent in this formulation, however.

Proof (of Proposition 1.23b): Let $\Omega=\Sigma_{j=1}^{n}M_j\,dx_j$ denote the connection one-form for ∇ relative to some trivialization in a neighborhood of $x=0$. Suppose $M_{1,-1}(y)$ is the residue of M_1

at $x_1=0$. Then the matrix $M_{1,-1}(0)$ has distinct eigenvalues, and this remains true for $M_{1,-1}(y)$ for x in a sufficiently small neighborhood of 0. Because its eigenvalues are distinct one may diagonalize $M_{1,-1}(y)$ by a holomorphic similarity transformation,

$$Q(y)M_{1,-1}(y)Q(y)^{-1}=\Lambda_0(y),$$

where $Q(y)$ is holomorphic and $\Lambda_0(y)$ is diagonal. Making a gauge transformation by Q, one finds that the connection ∇ becomes

$$d-\frac{\Lambda_0(y)}{x_1}dx_1-\sum_{j=1}^{n} B^j(x)dx_j,$$

where the $B^j(x)$ is holomorphic in x. By assumption, the eigenvalues of $\Lambda_0(0)$ do not differ by integers and so the same is true for $\Lambda_0(y)$ if y remains in a sufficiently small neighborhood of 0. Now let $z=x_1$ and consider the connection

$$d_z-\frac{\Lambda_0(y)}{z}dz-B^1(z,y)dz,$$

depending on the parameter y. Since the eigenvalues of $\Lambda_0(y)$ are distinct and do not differ by integers, the standard construction of a fundamental solution [which depends on the inversion of $\mathrm{ad}(\Lambda_0)-nI$ where $n=1,2,...,$ is an integer] shows that one can find a gauge transformation $\beta(z,y)=I+O(z)$, which depends holomorphically on the parameters y and that transforms this connection to

$$d_z-\frac{\Lambda_0(y)}{z}dz.$$

The gauge transformation of the complete connection by this transform gives one

$$d-\frac{\Lambda_0(y)}{z}dz-\sum_{j=1}^{n-1} B^{j+1}(z,y)dy_j, \tag{1.23}$$

where $y_j=x_{j+1}$. Now we express the integrability of this connection $(d\Omega+\Omega\wedge\Omega=0)$ expanding $B^k(z,y)$ in powers of z,

$$B^k(z,y)=\sum_{j=0}^{\infty} B_j^k(y)z^j.$$

Equating the coefficients of z^{-1} in the integrability condition, we find that

$$\frac{\partial\Lambda_0(y)}{\partial y_j}=[B_0^j,\Lambda_0].$$

Since Λ_0 is diagonal the right-hand side vanishes on the diagonal and this shows that $\Lambda_0(y)=\Lambda_0=\text{const}$. The left-hand side vanishes off the diagonal and since the entries of Λ_0 are distinct this implies that B_0^j must be diagonal. Equating the coefficients of $z^j dz\wedge dy_j$ in the integrability condition, one finds

$$nB_n^j=[\Lambda_0,B_n^j], \quad \text{for } n=1,2...,$$

which implies that $B_n^j=0$ for $n=1,2,...$. Thus, the connection ∇ has the form

$$d_z - \frac{\Lambda_0}{z}dz + d_y - \sum_{j=1}^{n-1} B_0^{j+1}(y)dy_j .$$

The connection

$$d_y - \sum_{j=1}^{n-1} B_0^{j+1}(y)dy_j$$

is integrable and so a gauge transformation, $\gamma(y)$, in the y variables reduces this connection to d_y. Since $B_0^{j+1}(y)$ is diagonal this gauge transformation may be chosen to be diagonal and also may be chosen so that $\gamma(y)=I+O(y)$. Since γ is diagonal, it does not alter the dz part of the connection. Composing β and γ we get a gauge transformation $\alpha=I+O(z)$, which reduces (1.23) to

$$d - \frac{\Lambda_0}{z}dz,$$

and this finishes the proof. QED

Although we do not require the result until the next section it will be convenient here to recall Theorem 2.1 from Ref. 14. Suppose that $(E,\boldsymbol{\nabla})$ is a vector bundle with a connection $\boldsymbol{\nabla}$ that has a simple type r pole along the hypersurface Y defined by $x_1=0$. Let $y=(x_2,\ldots,x_n)$ denote the coordinates along Y. Suppose that in a neighborhood of a point $x^0\in Y$, with coordinates $x_1=0$ and $y=0$, and relative to some local trivialization of the bundle one has a formal isomorphism,

$$\boldsymbol{\nabla} = \hat{\alpha}\cdot[\boldsymbol{\nabla}_\Lambda],$$

where $\boldsymbol{\nabla}_\Lambda$ is the diagonal connection described in Proposition 1.23a and $\hat{\alpha}$ is the formal power series described in that same proposition. Suppose that Σ is a good stable sector (in the x_1 variable) for the connection $\boldsymbol{\nabla}$ restricted to $y=0$. Suppose that $\epsilon>0$ is chosen small enough so that for all y_1 with $|y_1|<\epsilon$, the sector Σ remains a good stable sector for the restriction of $\boldsymbol{\nabla}$ to $y=y_1$.

Then one has (Theorem 2.1 in Ref. 14).

Proposition 1.23c: There exists a uniquely determined invertible holomorphic map $\alpha_\Sigma \in \mathcal{A}(\Sigma_\epsilon\times|y|<\epsilon)$ *such that on* $\Sigma_\epsilon\times|y|<\epsilon$ *and in an appropriate trivialization for E, one has*

$$\boldsymbol{\nabla} = \alpha_\Sigma\cdot[\boldsymbol{\nabla}_\Lambda],$$

and such that the map α_Σ *extends* $\hat{\alpha}$ *in the sense that* α_Σ *has an asymptotic development along Y, which is equal to* $\hat{\alpha}$.

Remark: The consequence of this result that is of interest for us is that the formal isomorphism class $\boldsymbol{\nabla}_\Lambda$ together with the Stokes multipliers determine the local holomorphic equivalence class of a simple, integrable, type r connection. Two collections, α_{Σ_k} and α'_{Σ_k}, associated with the same good stable cover $\{\Sigma_1,\ldots,\Sigma_{2r}\}$ with the same Stokes multipliers, and the same asymptotics $\hat{\alpha}$ clearly differ by an invertible holomorphic map.

We are now prepared to state an existence result for a global version of a "Stokes multiplier preserving deformation," which is, however, local in the x_1 variable. The space we will work on is $D\times\mathcal{R}(\mathcal{C})$, where D is the unit disk about $x=a$ in $\mathbf{C}$ and the connection whose existence we wish to demonstrate is a simple integrable type r connection along $\{a\}\times\mathcal{R}(\mathcal{C})$, which has formal reduction to

$$\boldsymbol{\nabla}_\lambda = d - d(\mathbf{H}) - \Lambda_0\frac{dx}{x-a}, \tag{1.24}$$

where $d=d_x+d_\lambda$ is the exterior derivative in the $(x,\lambda)\in D\times\mathcal{R}(\mathcal{C})$ variables and

$$\mathbf{H}=\sum_{j=1}^{r}\Lambda_k(\lambda)\frac{(x-a)^{-j}}{-j}. \tag{1.25}$$

Note that $\nabla_\lambda=e^H de^{-H}$, is the gauge transform of the connection d by e^H, with H given by $\mathbf{H}+\Lambda_0\log(x-a)$. However, since H is singular at $x=a$ and multivalued this is not properly a global statement, but does make sense locally.

Theorem 1.26: *Suppose that σ is a parallel section for $\mathcal{M}\rightarrow\mathcal{R}(\mathcal{C})$. Let D denote the unit disk in* $\mathbf{C}$ *centered at a. Then there exists a holomorphic integrable connection* ∇ *defined on the trivial vector bundle,*

$$D\times\mathcal{R}(\mathcal{C})\times\mathbf{C}^p\rightarrow D\times\mathcal{R}(\mathcal{C}), \tag{1.27}$$

which has a singularity of type $r\geqslant 1$ along the hypersurface $Y=\{a\}\times\mathcal{R}(\mathcal{C})$ such that the restriction of the connection ∇ to $D\backslash\{a\}\times\mathcal{R}(\mathcal{C})$ is formally equivalent to the diagonal model ∇_λ defined in (1.24) and such that the the holomorphic equivalence class of the restriction of ∇ to $(D\backslash\{a\})\times\{\lambda\}$ is given by $\sigma(\lambda)$.

Proof: We first recall a result of Malgrange and Sibuya, for which one can also find a detailed proof in Ref. 17. Suppose that $\lambda\in\mathcal{R}(\mathcal{C})$ and $\sigma\in\mathcal{M}_\lambda$, where $\mathcal{M}_\lambda$ is the fiber in $\mathcal{M}$ over λ. Then on the trivial bundle $D\times\mathbf{C}^p\rightarrow D$ there exists a simple type r connection $\overline{\nabla}$ singular at $x=a$ in D, which has formal reduction to the diagonal model,

$$\overline{\nabla}_\lambda:=d_x-d_x\mathbf{H}(\lambda)-\Lambda_0\frac{dx}{x-a}, \tag{1.28}$$

and which has ''Stokes multipliers'' given by σ. In the version of this result that is proved in Ref. 17 the connection is shown to exist on a disk, D_δ, of small radius δ. It is not difficult to use the Birkhoff factorization theorem to produce a connection defined on $D\backslash\{a\}$ with the same properties. Suppose then that one has a connection $\overline{\nabla}$ defined on the trivial bundle over D_δ and satisfying the conditions above. There exists a connection $\overline{\nabla}_{\text{ext}}$ defined on the trivial bundle $\mathbf{C}\backslash\{a\}\times\mathbf{C}^p\rightarrow\mathbf{C}\backslash\{a\}$ with the same holonomy about $x=a$ as the connection $\overline{\nabla}$ (it is easy to produce such a connection with a logarithmic pole at $x=a$). Because the holonomy of $\overline{\nabla}$ and of $\overline{\nabla}_{\text{ext}}$ about $x=a$ are equal it follows that there is an annulus A containing the circle, $S_{\delta/2}$, of radius $\delta/2$ on which the two connections are gauge equivalent. Thus there exists a holomorphic map $g:A\rightarrow\text{GL}(p,\mathbf{C})$ such that

$$g\cdot[\overline{\nabla}]=\overline{\nabla}_{\text{ext}}.$$

The Birkhoff theorem gives us a factorization,

$$g=g_\infty^{-1}(x-a)^N g_0,$$

where g_0 is holomorphic and invertible in a neighborhood of $x=a$ containing the annulus A, g_∞ is holomorphic and invertible in a neighborhood of ∞ that contains the annulus, A, and N is a diagonal matrix with integer entries. The equality

$$g_0\cdot[\overline{\nabla}]=(x-a)^{-N}g_\infty\cdot[\overline{\nabla}_{\text{ext}}], \tag{1.29}$$

on the annulus A, shows that the connection $g_0\cdot[\overline{\nabla}]$ extends to the punctured unit disk, and since it is in the same local holomorphic equivalence class as $\overline{\nabla}$, we have finished the demonstration that we may work on the unit disk, D, rather than D_δ.

To complete the proof of Theorem 1.26 we proceed in two steps. First, we show that if we confine our attention to a sufficiently small neighborhood U of $\lambda \in \mathcal{R}(\mathcal{C})$, then we can find a connection $\boldsymbol{\nabla}_U$ defined over $D\backslash\{a\}\times U$ that satisfies the conclusions of Theorem 1.26, with σ_U the unique local flat section of $\mathcal{M}$ with $\sigma_U(\lambda)=\sigma$. We will prove this using a variant of the Flaschka–Newell integral equation to produce a ''perturbation'' of the connection $\overline{\boldsymbol{\nabla}}$. We defer the proof of this result to Proposition 1.35 below. The second step is to put together the ''local'' solutions $\boldsymbol{\nabla}_U$ to get something defined on all of $\mathcal{R}(\mathcal{C})$. We will now show how to do this.

Let $\lambda \rightarrow \sigma(\lambda)$ denote a flat section of $\mathcal{M}\rightarrow\mathcal{R}(\mathcal{C})$. For each point $\lambda\in\mathcal{R}(\mathcal{C})$ there exists a neighborhood $U(\lambda,\sigma)$ of λ in which the construction of Proposition 1.35 applies. Let $\mathcal{U}$ denote a subcollection of such open neighborhoods that is a covering for $\mathcal{R}(\mathcal{C})$ and for which $U\cap V$ is contractible for each pair U, $V\in\mathcal{U}$. We also suppose that each neighborhood $U\in\mathcal{U}$ is chosen sufficiently small so that for $\delta>0$ small enough there exists a sectorial covering $\{\Sigma_{1,\delta},\Sigma_{2,\delta},\ldots,\Sigma_{2r,\delta}\}$ of the punctured neighborhood $D_\delta\backslash\{a\}$ that is stable and good for the all connections,

$$d_x - d_x\mathbf{H}(\lambda) - \Lambda_0\frac{dx}{x-a},$$

with $\lambda\in U$. The construction of Proposition 1.35 shows that there exist holomorphic maps,

$$\alpha_k^U:\Sigma_{k,\delta}\times U\rightarrow \mathrm{GL}(p,\mathbf{C}),$$

so that

$$\alpha_k^U\cdot[\boldsymbol{\nabla}_\lambda]=\boldsymbol{\nabla}_U,\quad \text{on}\ \ \Sigma_{k,\delta}\times U.$$

Furthermore, the maps α_k^U are related to one another,

$$\alpha_{k+1}^U=\alpha_k^U S_{k,k+1}(x,\lambda),$$

where $S_{k,k+1}:\Sigma_{k,\delta}\cap\Sigma_{k+1,\delta}\times U\rightarrow\mathrm{GL}(p,\mathbf{C})$ is a gauge automorphism of $\boldsymbol{\nabla}_\lambda$ that is asymptotic to the identity *to all orders* at $x=a$. Recall that

$$S_{k,k+1}(x,\lambda)=e^H S_{k,k+1}e^{-H},$$

where $S_{k,k+1}$ is a constant matrix and e^H is well defined once a choice of $x\rightarrow\log(x-a)$ is made for $x\in\Sigma_{1,\delta}$. Suppose that U, $V\in\mathcal{U}$; then for $\lambda\in U\cap V$, the fact that $S_{k,k+1}$ does not depend on U implies that the collection of holomorphic maps,

$$\alpha_k^U(\alpha_k^V)^{-1}\quad\text{for}\ \ k=1,2,...,2r,$$

defines a holomorphic map, g_{UV}, in a punctured neighborhood, $D_\delta\backslash\{a\}\times U\cap V$, and the fact that $S_{k,k+1}(x,\lambda)$ is asymptotic to the identity to all orders in $(x-a)$ implies that g_{UV} asymptotic to a power series near $x=a$ that does not depend on the sector. This implies that g_{UV} is actually holomorphic on $D_\delta\times U\cap V$. By construction,

$$g_{UV}\cdot[\boldsymbol{\nabla}_V]=\boldsymbol{\nabla}_U,\quad\text{on}\ \ D_\delta\backslash\{a\}\times U\cap V. \tag{1.30}$$

This shows that the holonomy of the connection $\boldsymbol{\nabla}_U$ and the holonomy of the connection $\boldsymbol{\nabla}_V$ agree on $D\backslash\{a\}\times U\cap V$ (since $U\cap V$ is contractible the fundamental group of the product $D\backslash\{a\}\times U\cap V$ is determined by the first factor) and hence that they are holomorphically equivalent on all of $D\backslash\{a\}\times U\cap V$. The map g_{UV} has an invertible holomorphic extension to all of $D\times U\cap V$ with the property that $g(a,\lambda)=I$ for all $\lambda\in U\cap V$. It is not difficult to see that the gauge transformation g_{UV} is uniquely determined by (1.30) and this normalization. Because of this $g_{UV}g_{VW}=g_{UW}$. Thus, the collection $\{g_{UV}|U,V\in\mathcal{U}\}$ is a collection of transition functions for a holomorphic vector bundle over $D\times\mathcal{R}(\mathcal{C})$. But $D\times\mathcal{R}(\mathcal{C})$ is contractible since both factors are, and every bundle on $D\times\mathcal{R}(\mathcal{C})$, is thus topologically trivial. But $D\times\mathcal{R}(\mathcal{C})$ is a Stein space. It is a theorem of Grauert that on a Stein space every topologically trivial holomorphic bundle is holomorphically trivial.[18] Thus, the bundle defined by these transition functions is holomorphically trivial. Thus, there exists invertible holomorphic maps $g_U:D\times U\rightarrow \mathrm{GL}(p,\mathbf{C})$, so that

$$g_{UV}=g_U^{-1}g_V.$$

Equation (1.30) becomes

$$g_V\cdot[\boldsymbol{\nabla}_V]=g_U\cdot[\boldsymbol{\nabla}_U],$$

and we see that $g_U\cdot[\boldsymbol{\nabla}_U]$ defines a global connection on $D\backslash\{a\}\times\mathcal{R}(\mathcal{C})$ that satisfies the conditions of Theorem 1.26. QED

Now we turn to the proof of the perturbation result used in the proof of the preceeding theorem. Suppose that $\overline{\boldsymbol{\nabla}}^0$ is a simple type r connection on the trivial bundle,

$$D\times\mathbf{C}^p\rightarrow D,$$

with singularity at $x=a$ in D. For the purpose of a technical result later on, it will be convenient to choose a special trivialization in which to consider the connection $\overline{\boldsymbol{\nabla}}^0$. Choose $\rho<1$ and let A denote the annulus,

$$A=\{x:\rho<|x-a|<1\}\subset D.$$

Choose a point $q\in A$ and let M denote the $p\times p$ invertible matrix, which gives the holonomy of the connection $\overline{\boldsymbol{\nabla}}^0$ (with respect to the initial trivialization) along a counterclockwise oriented circle of radius $|q-a|$ about a. Let

$$m:=\frac{1}{2\pi i}\log M,$$

for some choice of a logarithm for M. The connection

$$\overline{\boldsymbol{\nabla}}^\infty:=d_x-\frac{m}{x-a},$$

defined on the trivial bundle $\mathbf{P}^1\times\mathbf{C}^p$, has regular singular points at 0 and ∞ and its restriction to A has the same holonomy representation as $\overline{\boldsymbol{\nabla}}^0$. Hence, there exists an invertible holomorphic map $g:A\rightarrow\mathrm{GL}(p,\mathbf{C})$ so that

$$g\cdot\overline{\boldsymbol{\nabla}}^0=\overline{\boldsymbol{\nabla}}^\infty,\quad \text{on } A.$$

Let $g=g_\infty^{-1}(x-a)^{-N}g_0$ denote the Birkhoff factorization of g, with g_0 holomorphic and invertible in D, g_∞ holomorphic and invertible in $\{x:|x-a|>\rho\}\cup\{\infty\}$, and N a diagonal matrix with integer entries. Thus,

$$g_0\cdot\overline{\boldsymbol{\nabla}}^0=(x-a)^N g_\infty\cdot\overline{\boldsymbol{\nabla}}^\infty,$$

and we see that by adjusting the trivialization on $D\times\mathbf{C}^p$ by g_0 we may suppose that our connection $\overline{\boldsymbol{\nabla}}^0$ extends to a connection on the trivial bundle $\mathbf{P}^1\times\mathbf{C}^p\rightarrow\mathbf{P}^1$ with a regular singular point at ∞ (the resulting connection may not be of a simple type at ∞, however). In what follows we suppose that we are looking at $\overline{\boldsymbol{\nabla}}^0$ in just such a trivialization.

Now suppose that the leading singularity in the connection one-form for $\bar{\boldsymbol{\nabla}}^0$ is diagonal and that for the formal power series,

$$\hat{\alpha}^0 = I + \beta_1^0(x-a) + \beta_2^0(x-a)^2 + \cdots,$$

we have

$$\bar{\boldsymbol{\nabla}}^0 = \hat{\alpha}^0 \cdot \left[d_x - d_x \mathbf{H}(\lambda^0) - \Lambda_0 \frac{dx}{x-a} \right], \tag{1.31}$$

where $\mathbf{H}$ is given by (1.25) and λ^0 is some fixed element in $\mathcal{R}(\mathcal{C})$ covering $\Lambda(\lambda^0)$. Note that the projection $\Lambda^0 = \Lambda(\lambda^0)$ is determined by $\bar{\boldsymbol{\nabla}}^0$ through (1.31). Let $\{\Sigma_{1,\delta}, \Sigma_{2,\delta}, \ldots, \Sigma_{2r,\delta}\}$ denote a stable good covering of a punctured neighborhood of $x=a$ for the diagonal connection $\bar{\boldsymbol{\nabla}}_{\lambda^0}$ [see (1.28)]. Let $\alpha_k^0 = \alpha_{\Sigma_k}^0 \in \mathcal{A}(\Sigma_k)$ be a holomorphic function whose asymptotics are given by $\hat{\alpha}^0$ and for which one has the analytical relation

$$\bar{\boldsymbol{\nabla}}^0 = \alpha_k^0 \cdot \left[d_x - d_x \mathbf{H}(\lambda^0) - \Lambda_0 \frac{dx}{x-a} \right], \tag{1.32}$$

in the sector $\Sigma_{k,\delta}$. Finally, suppose that on $\Sigma_{k,\delta} \cap \Sigma_{k+1,\delta}$, we have

$$\alpha_{k+1}^0 = \alpha_k^0 S_{k,k+1}(x, \lambda^0), \tag{1.33}$$

where

$$S_{k,k+1}(x,\lambda) = e^{H(\lambda)} S_{k,k+1}^0 e^{-H(\lambda)}. \tag{1.34}$$

Here H is given by (1.16) and a choice of $\log(x-a)$ is fixed for $x \in \Sigma_{1,\delta}$ to make H well defined. The Stokes multiplier $S_{k,k+1}^0$ is independent of x and λ^0. Let Ω^0 denote the connection form for $\bar{\boldsymbol{\nabla}}^0$ and let pr denote the natural projection,

$$D \times \mathcal{R}(\mathcal{C}) \to D.$$

Define

$$\boldsymbol{\nabla}^0 = d + \mathrm{pr}^* \, \Omega^0,$$

with $d = d_x + d_\lambda$, so that $\boldsymbol{\nabla}^0$ defines a connection on the trivial bundle,

$$D \times \mathcal{R}(\mathcal{C}) \times \mathbf{C}^p \to D \times \mathcal{R}(\mathcal{C}).$$

The following proposition demonstrates the existence of local ''Birkhoff deformations'' of the connection $\bar{\boldsymbol{\nabla}}^0$ in the space $D \times \mathcal{R}(\mathcal{C})$.

Proposition 1.35: For a sufficiently small neighborhood U of λ^0 there exists a simple integrable type r connection $\boldsymbol{\nabla}_U$ defined on the trivial bundle,

$$D \times U \times \mathbf{C}^p \to D \times U,$$

so that we have the following.

(i) $\boldsymbol{\nabla}_U$ is formally reducible to the diagonal form (1.28),

$$\boldsymbol{\nabla}_U = \hat{\alpha} \cdot [\boldsymbol{\nabla}_\lambda], \tag{1.36}$$

where

$$\hat{\alpha}=I+\beta_1(\lambda)(x-a)+\beta_2(\lambda)(x-a)^2+\cdots \tag{1.37}$$

is a formal power series with holomorphic matrix-valued coefficients $\beta_k(\lambda)$*. The sectors* $\Sigma_k \times U$ *are good stable sectors for* ∇_U*; there exist holomorphic maps,*

$$\alpha_k \in \mathcal{A}(\Sigma_k, U), \tag{1.38}$$

with asymptotics given by $\hat{\alpha}$*, so that on open sets asymptotic to* $\Sigma_k \times U$*.*

$$\nabla_U=\alpha_k\cdot[\nabla_\lambda], \tag{1.39}$$

(ii) ∇_U *is a ''Birkhoff deformation'' of* $\overline{\nabla}^0$*, in that the restriction of* ∇_U *to* $D\backslash\{a\}\times\{\lambda^0\}$ *is equivalent to* $\overline{\nabla}^0$ *and*

$$\alpha_{k+1}(x,\lambda)=\alpha_k(x,\lambda)S_{k,k+1}(x,\lambda), \tag{1.40}$$

on $\Sigma_{k,\delta}\cap\Sigma_{k+1,\delta}\times U$ *[see (1.34)].*

(iii) On the punctured neighborhood $D\backslash\{a\}\times U$*, the connections* ∇_U *and* ∇^0 *are gauge equivalent,*

$$\nabla_U=\Phi\cdot[\nabla^0], \tag{1.41}$$

by a gauge transformation $x\to\Phi(x,\lambda)$ *that is holomorphic in the exterior of the disk D and asymptotic to I as* $x\to\infty$*.*

(iv) If $\overline{\nabla}^0$ *extends to a connection on the trivial bundle* $\mathbf{P}^1\times\mathbf{C}^p\to\mathbf{P}^1$ *with a regular singular point at infinity; then*

$$d_\lambda \operatorname{Res}_{x=a} \operatorname{Tr}\{\hat{\alpha}^{-1}\, d_x\hat{\alpha}\, d_\lambda H(\lambda)\}=0. \tag{1.42}$$

Proof: To construct ∇_U through Eq. (1.39) it will suffice to construct functions α_k satisfying (1.40) with $\alpha_k(x,\lambda^0)=\alpha_k^0(x)$. Our strategy will be to look for solutions

$$\alpha_k(x,\lambda)=\varphi_k(x,\lambda)\,\alpha_k^0(x),$$

with $\varphi_k:\Sigma_{k,\delta}\times U\to \mathrm{GL}(p,\mathbf{C})$ a holomorphic map with appropriate asymptotics. Condition (1.40) for $\{\alpha_k\}$ translates into

$$\varphi_{k+1}(x,\lambda)=\varphi_k(x,\lambda)(I+\Delta s_{k,k+1}(x,\lambda)), \tag{1.43}$$

for (x,λ) in $\Sigma_{k,\delta}\cap\Sigma_{k+1,\delta}\times U$, where

$$I+\Delta s_{k,k+1}(x,\lambda)=\alpha_k^0(x)S_{k,k+1}(x,\lambda)S_{k,k+1}(x,\lambda^0)^{-1}\alpha_k^0(x)^{-1}.$$

The important property of $\Delta s_{k,k+1}$ for us is that for any fixed $\delta>0$ one can make $\Delta s_{k,k+1}$ as close to zero as one likes by choosing $\lambda\in U$, with U a sufficiently small neighborhood of λ^0.

The condition (1.40) [or its translation (1.43)] does not determine α_k uniquely. We will impose a further condition on α_k that will uniquely determine it. The extra condition is that the ''fundamental solutions'' $\Psi(x,\lambda)$ and $\Psi^0(x)$ associated to $\{\alpha_k\}$ and $\{\alpha_k^0\}$ by (1.8) differ on the circle of radius δ by a map that is holomorphic in the exterior of the circle of radius δ about $x=a$ and asymptotic to the identity at ∞. More precisely, $\Psi(x,\lambda)\Psi^0(x)^{-1}$ should be holomorphic in x outside the circle of radius δ and

$$\Psi(x,\lambda)\Psi^0(x)^{-1}=I+O(x^{-1}).$$

Note that $\Phi(x,\lambda)=\Psi(x,\lambda)\Psi^0(x)^{-1}$ will be the gauge transformation in (iii) of Proposition 1.35. We will now translate the conditions we have outlined into an integral equation for the functions

φ_k. It will be useful to begin by describing the pieces that make up the integration contours we will use. Let $\sigma_{k,k+1}$ denote an oriented ray segment that lies in the intersection $\Sigma_{k,\delta}\cap\Sigma_{k+1,\delta}$, joins the point a to the circle of radius δ about a, and separates the Stokes lines in the sector $\Sigma_{k,\delta}$ from the Stokes lines in the sector $\Sigma_{k+1,\delta}$ for all $\lambda\in U$ [Stokes lines are associated with the function $\Lambda_r(\lambda)$]. It is always possible to do this if δ is small enough and U is chosen to be a sufficiently small neighborhood of λ^0. Let γ_k denote the counterclockwise oriented segment of the circle of radius δ about a that joins the end point of $\sigma_{k-1,k}$ to the end point of $\sigma_{k,k+1}$. Finally, let σ_k denote the open wedge that is bounded by $\sigma_{k-1,k}$, γ_k, and $\sigma_{k,k+1}$. Then clearly $\sigma_k\subset\Sigma_{k,\delta}$, and the oriented boundary of σ_k is given by

$$\partial\sigma_k=\sigma_{k-1,k}+\gamma_k-\sigma_{k,k+1}.$$

If we now compare Ψ and Ψ^0 defined by (1.8) on the circle of radius δ about a we find

$$\Psi(x,\lambda)\Psi^0(x)^{-1}=\varphi_k(x,\lambda)(I+\Delta m_k(x,\lambda))\quad\text{for }x\in\gamma_k, \tag{1.44}$$

where $k=1,2,...,2r$, with

$$I+\Delta m_k(x,\lambda):=\alpha_k^0(x)e^{H(x,\lambda)-H(x,\lambda^0)}\alpha_k^0(x)^{-1}. \tag{1.45}$$

We have written (1.45) in the special form $I+\Delta m_k$ to emphasize the fact that for fixed δ the right-hand side of (1.45) can be made as close to the identity as one pleases by choosing $\lambda\in U$, with U a sufficiently small neighborhood of λ^0. Next we will obtain a system of integral equations for φ_k following Flaschka and Newell.[15] Suppose that $y\in\sigma_1$; then by Cauchy's theorem,

$$\varphi_1(y,\lambda)=\int_{\partial\sigma_1}\frac{\varphi_1(x,\lambda)}{x-y}\frac{dx}{2\pi i}=\int_{\sigma_{0,1}}\frac{\varphi_1(x,\lambda)}{x-y}\frac{dx}{2\pi i}+\int_{\gamma_1}\frac{\varphi_1(x,\lambda)}{x-y}\frac{dx}{2\pi i}-\int_{\sigma_{1,2}}\frac{\varphi_1(x,\lambda)}{x-y}\frac{dx}{2\pi i}. \tag{1.46}$$

On $\sigma_{1,2}$ we can use (1.43) to write

$$\varphi_1(x,\lambda)=\varphi_1(x,\lambda)-\varphi_2(x,\lambda)+\varphi_2(x,\lambda)=-\varphi_1\Delta s_{1,2}(x,\lambda)+\varphi_2(x,\lambda), \tag{1.47}$$

where for brevity we have written $\varphi_1\,\Delta s_{1,2}(x,\lambda)$ for $\varphi_1(x,\lambda)\Delta s_{1,2}(x,\lambda)$ (we will use this notation without further comment in what follows). Substituting this expression in the $\sigma_{1,2}$ integral in (1.46), one finds

$$\varphi_1(y,\lambda)=\int_{\sigma_{0,1}}\frac{\varphi_1(x,\lambda)}{x-y}\frac{dx}{2\pi i}+\int_{\gamma_1}\frac{\varphi_1(x,\lambda)}{x-y}\frac{dx}{2\pi i}+\int_{\sigma_{1,2}}\frac{\varphi_1\Delta s_{1,2}(x,\lambda)}{x-y}\frac{dx}{2\pi i}-\int_{\sigma_{1,2}}\frac{\varphi_2(x,\lambda)}{x-y}\frac{dx}{2\pi i}. \tag{1.48}$$

Since $\varphi_2(x,\lambda)$ is holomorphic in σ_2 and $y\in\sigma_1$, which is outside of σ_2 it follows that

$$-\int_{\sigma_{1,2}}\frac{\varphi_2(x,\lambda)}{x-y}\frac{dx}{2\pi i}=\int_{\gamma_2}\frac{\varphi_2(x,\lambda)}{x-y}\frac{dx}{2\pi i}-\int_{\sigma_{2,3}}\frac{\varphi_2(x,\lambda)}{x-y}\frac{dx}{2\pi i}. \tag{1.49}$$

We substitute (1.49) for the last integral to appear in (1.48). In the expression that results we observe that the integral,

$$\int_{\sigma_{2,3}}\frac{\varphi_2(x,\lambda)}{x-y}\frac{dx}{2\pi i}=-\int_{\sigma_{2,3}}\frac{\varphi_2\,\Delta s_{2,3}(x,\lambda)}{x-y}\frac{dx}{2\pi i}+\int_{\sigma_{2,3}}\frac{\varphi_3(x,\lambda)}{x-y}\frac{dx}{2\pi i},$$

and one may continue this last integral to γ_3 and $\sigma_{3,4}$ by Cauchy's theorem as above. Proceeding all the way around the circle in this fashion, one finds

$$\varphi_1(y,\lambda)-\sum_{k=1}^{2r}\left\{\int_{\gamma_k}\frac{\varphi_k(x,\lambda)}{x-y}\frac{dx}{2\pi i}+\int_{\sigma_{k,k+1}}\frac{\varphi_k\Delta s_{k,k+1}(x,\lambda)}{x-y}\frac{dx}{2\pi i}\right\}=0. \tag{1.50}$$

Now we formulate the condition that the right-hand side of (1.44) should be holomorphic in the exterior of the circle of radius δ about a and asymptotic to the identity at ∞. This can be expressed as

$$\sum_{k=1}^{2r}\int_{\gamma_k}\frac{\varphi_k(x,\lambda)(I+\Delta m_k(x,\lambda))}{x-y}\frac{dx}{2\pi i}=I. \tag{1.51}$$

Adding this result to the preceding equation, one finds

$$\varphi_1(y,\lambda)+K_1\varphi(y,\lambda)=I, \tag{1.52}$$

where

$$K_1\varphi(y,\lambda)=\sum_{k=1}^{2r}\left\{\int_{\gamma_k}\frac{\varphi_k\,\Delta m_k(x,\lambda)}{x-y}\frac{dx}{2\pi i}-\int_{\sigma_{k,k+1}}\frac{\varphi_k\,\Delta s_{k,k+1}(x,\lambda)}{x-y}\frac{dx}{2\pi i}\right\}, \tag{1.53}$$

and $\varphi=(\varphi_1,\varphi_2,\ldots,\varphi_{2r})$. We ''derived'' (1.52) with y chosen to be in the interior of σ_1. However, we now choose to think of (1.52) as an integral equation for the restriction of φ_1 to $\partial\sigma_1$. In this case the integral operator K_1 defined in (1.52) is understood to involve nontangential limits for y on $\partial\sigma_1$ from the interior of σ_1. There is was nothing special about φ_1 in the arguments above, and so we find that the vector φ satisfies the system of integral equations,

$$\varphi_k(y,\lambda)+K_k\varphi(y,\lambda)=I, \tag{1.54}$$

where the integral operators K_k are defined by the same formula as K_1 but the y variable that occurs in (1.53) takes values in $\partial\sigma_k$ (with the integral operator defined by nontangential limits from the interior). It might be instructive for the reader to compare the system (1.54) of singular integral equations with the standard translation of a Riemann–Hilbert problem into a singular integral equation to be found in (3.17) below.

It is well known that nontangential limits for the Cauchy kernel $(x-y)^{-1}$ determine a bounded operator on $L^2(\partial\sigma_k)$ in case both x and y are in $\partial\sigma_k$. It is not hard to see from this that the integral operator

$$K\varphi=(K_1\varphi,K_2\varphi,\ldots,K_{2r}\varphi)$$

is a bounded operator on

$$\mathcal{H}=L^2(\partial\sigma_1)\oplus L^2(\partial\sigma_2)\oplus\cdots\oplus L^2(\partial\sigma_{2r}).$$

Furthermore, it is clear that because $\Delta m_k(x,\lambda)$ and $\Delta s_{k,k+1}(x,\lambda)$ can be made uniformly small by choosing λ close enough to λ^0, the system of integral equations (1.54) has a unique solution in $\mathcal{H}$ provided the neighborhood U is small enough.

Next, we wish to show that the solution of (1.54) satisfies (1.43). Suppose then that φ is a solution to (1.54) in $\mathcal{H}$. One calculates that, for $y\in\sigma_{j,j+1}$,

$$\varphi_{j+1}(y,\lambda)-\varphi_j(y,\lambda)=K_j\varphi(y,\lambda)-K_{j+1}\varphi(y,\lambda)$$
$$=\int_{\sigma_{j,j+1}}\left\{\frac{\varphi_j\,\Delta s_{j,j+1}(x,\lambda)}{x-y(\sigma_j)}-\frac{\varphi_j\,\Delta s_{j,j+1}(x,\lambda)}{x-y(\sigma_{j+1})}\right\}\frac{dx}{2\pi i}, \tag{1.55}$$

where we have written $y(\sigma_k)$ for the boundary value on $\sigma_{j,j+1}$ taken from the interior of σ_k for $k=j$ and $k=j+1$. Since

$$\frac{1}{2\pi i}\left\{\frac{1}{x-y(\sigma_j)}-\frac{1}{x-y(\sigma_{j+1})}\right\}=\delta(x-y),\quad \text{for } y\in\sigma_{j,j+1},$$

it follows from (1.55) that

$$\varphi_{j+1}(y,\lambda)-\varphi_j(y,\lambda)=\varphi_j(y,\lambda)\Delta s_{j,j+1}(y,\lambda),\quad \text{for } y\in\sigma_{j,j+1}. \tag{1.56}$$

This is a ''boundary value'' version of (1.43), that we will now extend to a sectorial neighborhood of $\sigma_{j,j+1}$. As a simple consequence of satisfying the integral equation (1.54), we know that $\varphi_j(y,\lambda)$ for $y\in\sigma_{j,j+1}$ is the boundary value of a holomorphic function $\varphi_j(y,\lambda)$ for $y\in\sigma_j$. Also, $\varphi_{j+1}(y,\lambda)$ for $y\in\sigma_{j,j+1}$ is the boundary value of a holomorphic function $\varphi_{j+1}(y,\lambda)$ for $y\in\sigma_{j+1}$. Solving (1.56) for $\varphi_{j+1}(y,\lambda)$, we see this function has an analytic continuation into a sector containing $\sigma_{j,j+1}$, since Eq. (1.43) shows that the function $I+\Delta s_{j,j+1}(y,\lambda)$ (and its inverse) has an analytic continuation into such a sector. One can use Morera's theorem to show that the function obtained by gluing together L^2 boundary values along $\sigma_{j,j+1}$ is actually holomorphic in a neighborhood of $\sigma_{j,j+1}$. The same argument works for $\varphi_j(y,\lambda)$ and Eq. (1.56) extends to a sectorial neighborhood of $\sigma_{j,j+1}$. This is (1.43).

It is easy to see that the operator K depends analytically on λ in the operator norm on $\mathcal{H}$. It follows that the iterative solution to (1.54) produces an analytic function of λ with values in $\mathcal{H}$. We would like to know that the solution $\varphi_k(x,\lambda)$ is jointly analytic in x and λ in the usual sense. We know already that for fixed $\lambda\in U$, the function $\varphi_k(x,\lambda)$ is analytic in a sectorial neighborhood of σ_k. It is a theorem (Hartog's theorem) that separate analyticity in each argument implies joint analyticity; thus, it will suffice to show that for x_0 chosen in a suitable sectorial neighborhood of σ_k, the function $\lambda\to\varphi_k(x_0,\lambda)$ is holomorphic in λ. We can argue for this analyticity in the following way. If necessary, enlarge the opening of the wedge σ_k to produce a wedge σ_k' so that the point x_0 becomes an interior point of the region σ_k', and σ_k' remains inside the region of analyticity for $x\to\varphi_k(x,\lambda)$ $(\lambda\in U)$. Since we have confirmed that $\varphi(x,\lambda)$ satisfies (1.43) the arguments leading up to (1.53) now show that $\varphi(x,\lambda)$ satisfies the integral equation (1.53) with the contours adjusted to accommodate σ_k' (σ_{k-1} and σ_{k+1} are a bit smaller). In particular, $U\ni\lambda\to\varphi_k(\cdot,\lambda)$ is an analytic function of λ with values in $L^2(\partial\sigma_k')$. Cauchy's theorem shows that

$$\varphi_k(x_0,\lambda)=\frac{1}{2\pi i}\int_{\partial\sigma_k'}\frac{\varphi_k(x,\lambda)}{x-x_0}\,dx.$$

Since $x\to 1/(x-x_0)$ is in $L^2(\partial\sigma_k')$ the result is analytic in λ. This proves separate analyticity in λ.

The asymptotics for φ_j are obtained by substituting

$$\frac{1}{x-y}=\frac{1}{x}\sum_{n=0}^{N}\left(\frac{y}{x}\right)^n+y^{N+1}\frac{x^{-N-1}}{x-y}$$

into the integral equation (1.54) and noting that the functions

$$\varphi_j(x,\lambda)x^{-k}\,\Delta s_{j,j+1}(x,\lambda),$$

are integrable in x on $\sigma_{j,j+1}$ for all integers $k\geqslant 0$ with integrals that are analytic in λ.

This finishes the proof of (i), (ii). To establish (iii) note that we have shown that each solution φ_k to (1.54) extends to a holomorphic function in a sector containing σ_k, with controlled asymptotic behavior as $x \to a$. This is all that is needed to establish the analog of (1.50) for φ_k. Subtracting this from (1.54), one obtains the analog of (1.51) for φ_k. As noted above, this is an expression of the holomorphic character of the gauge transformation $\Phi(x,\lambda) := \Psi(x,\lambda)\Psi^0(x)^{-1}$ in the exterior of *D*, and this finishes the proof of (iii).

To establish (iv) (which will play an important role in a tau function calculation in Sec. III) write

$$\Omega = \Omega_x + \Omega_\lambda ,$$

for the one-form associated with ∇_U. Here Ω_x is the *dx* term in the one-form and Ω_λ is a sum,

$$\Omega_\lambda = \sum_k \Omega_{\lambda,k}\, d\lambda_k .$$

The $d\lambda$ component of the formal equivalence $\nabla_U = \hat{\alpha}\cdot[\nabla_\lambda]$ is

$$-d\hat{\alpha}\hat{\alpha}^{-1} - \hat{\alpha}\, dH\, \hat{\alpha}^{-1} = \Omega_\lambda ,$$

or

$$d\hat{\alpha} = -\hat{\alpha}\, dH - \Omega_\lambda \hat{\alpha}. \tag{1.57}$$

For simplicity, we write $d = d_\lambda$ and calculate

$$\begin{aligned} d\, \mathrm{Res}_{x=a} \mathrm{Tr}(\hat{\alpha}^{-1} d_x\hat{\alpha}\, dH) &= \mathrm{Res}_{x=a} \mathrm{Tr}(-\hat{\alpha}^{-1} d\hat{\alpha}\, \hat{\alpha}^{-1} d_x\hat{\alpha}\, dH - \hat{\alpha}^{-1} d_x d\hat{\alpha}\, dH) \\ &= \mathrm{Res}_{x=a} \mathrm{Tr}(d_x(dH)dH + \hat{\alpha}^{-1} d_x\Omega_\lambda\, \hat{\alpha}\, dH), \end{aligned}$$

where to get from the second to the the third line we substituted (1.57) for $d\hat{\alpha}$ did an obvious cancellation in the result and made use of the fact that $dH\, dH = 0$ since *dH* is diagonal. But

$$\mathrm{Res}_{x=a} \mathrm{Tr}(d_x(dH)dH) = 0,$$

since the Laurent series for $d_x(dH)dH$ begins with terms $C(x-a)^{-3}$, and so we find

$$d\, \mathrm{Res}_{x=a} \mathrm{Tr}(\hat{\alpha}^{-1} d_x\hat{\alpha}\, dH) = \mathrm{Res}_{x=a} \mathrm{Tr}(d_x\Omega_\lambda\, \hat{\alpha}\, dH\, \hat{\alpha}^{-1}). \tag{1.58}$$

A straightforward calculation now shows that

$$\mathrm{Res}_{x=a} \mathrm{Tr}(d_x(\hat{\alpha}\, dH\, \hat{\alpha}^{-1})\hat{\alpha}\, dH\, \hat{\alpha}^{-1}) = \mathrm{Res}_{x=a} \mathrm{Tr}(d_x(dH)dH) = 0. \tag{1.59}$$

In (1.59) we replace $\hat{\alpha}\, dH\, \hat{\alpha}^{-1}$ by $-d\hat{\alpha}\, \hat{\alpha}^{-1} - \Omega_\lambda$ from (1.57), and find

$$\mathrm{Res}_{x=a} \mathrm{Tr}(d_x(\Omega_\lambda)\Omega_\lambda) + 2\, \mathrm{Res}_{x=a} \mathrm{Tr}(d_x(\Omega_\lambda)d\hat{\alpha}\, \hat{\alpha}^{-1}) = 0, \tag{1.60}$$

where we made use of the fact that $d_x(d\hat{\alpha}\, \hat{\alpha}^{-1})d\hat{\alpha}\, \hat{\alpha}^{-1}$ is ''regular'' at $x=a$ and so has 0 residue, and that

$$\mathrm{Res}_{x=a}\, d_x\, \mathrm{Tr}(\Omega_\lambda\, d\hat{\alpha}\, \hat{\alpha}^{-1}) = 0.$$

Now substitute $-d\hat{\alpha}\, \hat{\alpha}^{-1} - \Omega_\lambda$ for $\hat{\alpha}\, dH\, \hat{\alpha}^{-1}$ in (1.58) and make use of (1.60), to get

$$d\, \mathrm{Res}_{x=a} \mathrm{Tr}(\hat{\alpha}^{-1} d_x\hat{\alpha}\, dH) = -\tfrac{1}{2}\mathrm{Res}_{x=a} \mathrm{Tr}(d_x(\Omega_\lambda)\Omega_\lambda). \tag{1.61}$$

Now for the first time we use the fact that $\overline{\nabla}^0$ is looked at in a trivialization in which it extends to a connection on $\mathbf{P}^1$ with an additional regular singularity at ∞. We see from this and (iii) that

$$\Omega_\lambda = -d\Phi\,\Phi^{-1},$$

is holomorphic in a neighborhood of ∞. Since $\mathrm{Tr}(d_x(\Omega_\lambda)\Omega_\lambda)$ is meromorphic on $\mathbf{P}^1$ with a single pole at $x=a$, it follows that the residue at this pole must be 0. With (1.61), this finishes the proof of (iv) (incidentally, the argument here follows the argument in Ref. 8 used to show that the Jimbo, Miwa, Ueno expression for $d\log\tau$ is closed). QED

II. THE VECTOR BUNDLE DEFORMATION OF MALGRANGE

A. Representations of the fundamental group and flat connections

In this section we are interested in constructing an integrable deformation of a connection on a bundle over $\mathbf{P}^1$ that is monodromy preserving and that respects the local character of the connection near its singular points.

The principal tool in the construction of this deformation away from the singular set is a correspondence between representations of the fundamental group and vector bundles with flat connections. More precisely, suppose that X is a connected complex manifold with base point x^0. Suppose that $E\to X$ is a complex vector bundle with a flat holomorphic connection ∇. Suppose that $\gamma:[0,1]\to X$ is a piecewise smooth curve in X and let $\mathcal{P}_{\nabla}(\gamma)$ denote parallel translation with respect to ∇ along γ. Then

$$\mathcal{P}\nabla(\gamma):E_{\gamma(0)}\to E_{\gamma(1)},$$

is a linear isomorphism between the fibers of E at the end points $\gamma(0)$ and $\gamma(1)$. Now suppose that γ is a piecewise smooth closed loop based at x^0 and let $g=[\gamma]$ denote the homotopy class of γ. Then

$$\rho(g) := \mathcal{P}_{\nabla}(\gamma)^{-1}, \tag{2.1}$$

defines a representation of $\pi_1(X,x^0)$ on E_{x^0}. The right-hand side depends only on the homotopy class of γ because the curvature of ∇ is zero. The equivalence class of the representation ρ actually determines the pair (E,∇) up to an isomorphism. Before we turn to the main theorem of this section we digress to sketch a construction that takes one from ρ to (E,∇). Let $\pi:\mathcal{R}(X)\to X$ denote the simply connected covering space of X and suppose that (E,∇) is a vector bundle with flat connection over X as above. The pull-back bundle $\pi^*(E)\to\mathcal{R}(X)$ is necessarily trivial since the base $\mathcal{R}(X)$ is simply connected. The natural projection $\tilde{\pi}:\pi^*(E)\to E$ is a local diffeomorphism, and since $d\tilde{\pi}_{\tilde{p}}$ is an isomorphism of tangent spaces that maps the vertical vectors in $T_{\tilde{p}}(\pi^*(E))$ bijectively onto the vertical vectors in $T_p(E)$ we may use $d\tilde{\pi}_{\tilde{p}}$ to lift the horizontal subspace in $T_p(E)$ that comes from ∇ to a horizontal subspace in $T_{\tilde{p}}(\pi^*(E))$. We write $\pi^*(\nabla)$ for the resulting connection on $\pi^*(E)$, and note that since the pull-back connection $\pi^*(\nabla)$ is related to ∇ by a local diffeomorphism, it is also a flat connection. We may thus produce a trivialization for $\pi^*(E)$ consisting of flat sections for $\pi^*(\nabla)$. Let $\tilde{x}^0$ denote a base point in $\mathcal{R}(X)$ such that $\pi(\tilde{x}^0)=x^0$. For $u\in\pi^*(E)_{\tilde{x}^0}$ let $\tilde{x}\to\mathcal{P}(\tilde{x},u)\in\pi^*(E)_{\tilde{x}}$ denote the parallel section of $\pi^*(E)$, which agrees with u at $\tilde{x}^0$. Writing $E_0 := \pi^*(E)_{\tilde{x}^0}$ (which is naturally isomorphic to E_{x^0}), we have

$$\mathcal{R}(X)\times E_0 \ni (x,u)\to\mathcal{P}(x,u)\in\pi^*(E),$$

is an isomorphism between the bundle $\pi^*(E)$ and the trivial bundle $\mathcal{R}(X)\times E_0$. If we compose this map $\mathcal{P}$ with the vector bundle projection,

$$\tilde{\pi}:\pi^*(E)\to E,$$

one obtains a vector bundle map,

$$\tilde{\pi}\mathcal{P}:R(X)\times E_0\to E, \tag{2.2}$$

which covers the projection $\pi:\mathcal{R}(X)\to X$.

The representation ρ from (2.1) determines a left action of $\pi_1(X,x^0)$ on $R(X)\times E_0$ given by

$$g\cdot(\tilde{x},u)=(g\cdot\tilde{x},\rho(g)u), \tag{2.3}$$

where $g\cdot\tilde{x}$ is just the usual action of π_1 on the simply connected covering space $\mathcal{R}(X)$. We will show that the quotient bundle $\pi_1(X,x^0)\backslash\mathcal{R}(X)\times E_0\to\pi_1(X,x^0)\backslash\mathcal{R}(X)$ is isomorphic to $E\to X$ as a vector bundle with connection.

We begin by showing that the map $\tilde{\pi}\mathcal{P}$ is equivariant for this action of $\pi_1(X,x^0)$. To see this suppose that $\tilde{x}=[\chi]$, where χ is a smooth curve joining x^0 to x in X. Let $\tilde{x}^0$ denote the class of the constant path starting and ending at x^0. Write $\tilde{\chi}$ for the lift of χ into $\mathcal{R}(X)$ with initial point $\tilde{x}^0$. Let $g=[\gamma]$ where γ is a smooth closed path in X based at x^0. Then one finds

$$\begin{aligned}\tilde{\pi}\mathcal{P}(g\cdot\tilde{x},\rho(g)u)&=\tilde{\pi}\mathcal{P}_{\pi^*(\nabla)}(\widetilde{\gamma\chi})\rho(g)u\\&=\mathcal{P}_{\nabla}(\gamma\chi)\rho(g)u\\&=\mathcal{P}_{\nabla}(\chi)\mathcal{P}_{\nabla}(\gamma)\rho(g)u=\mathcal{P}_{\nabla}(\chi)u=\tilde{\pi}\mathcal{P}(\tilde{x},u),\end{aligned}$$

which shows the equivariance of the map $\tilde{\pi}\mathcal{P}$. In the third and last equality we used the fact that $\tilde{\pi}\mathcal{P}_{\pi^*(\nabla)}(\tilde{\gamma})u=\mathcal{P}_{\nabla}(\gamma)u$, where $\tilde{\gamma}$ is a lift of γ and $u\in E_{\gamma(0)}$. This is obvious if the curve γ stays in a neighborhood U in X, which is evenly covered by the projection on X from $\mathcal{R}(X)$; the general result follows from the fact that parallel translation is an antihomomorphism under homotopy composition.

Since the vector bundle action (2.3) covers the standard left action of $\pi_1(X,x^0)$ on $\mathcal{R}(X)$, and since $\pi_1(X,x^0)\backslash\mathcal{R}(X)\simeq X$, one finds that

$$\pi_1(X,x^0)\backslash\mathcal{R}(X)\times E_0 := \mathcal{R}(X)\times_\rho E_0,$$

is a vector bundle over X isomorphic to E through the map induced by (2.2). Since the construction of the bundle

$$\mathcal{R}(X)\times_\rho E_0\to X \tag{2.4}$$

depends only on the representation ρ, it follows that this representation determines the bundle $E\to X$ up to isomorphism. In fact, the bundle (2.4) has a naturally defined flat connection ∇_ρ so that the map induced by (2.2) determines an isomorphism,

$$(\mathcal{R}(X)\times_\rho E_0,\nabla_\rho)\simeq(E,\nabla). \tag{2.5}$$

In order to define ∇_ρ we introduce a family of trivializations for the vector bundle,

$$\pi_\rho:\mathcal{R}(X)\times_\rho E_0\to X. \tag{2.6}$$

Let $\mathcal{F}$ denote a covering of X by open sets, with the following properties.

(1) If $U\in\mathcal{F}$, then U is evenly covered by $\pi:\mathcal{R}(X)\to X$. That is there exist disjoint sets $U_\alpha\subset X$ such that $\pi^{-1}(U)=\cup_\alpha U_\alpha$ and $\pi:U_\alpha\to U$ is a diffeomorphism for each α.

(2) If U, $V\in\mathcal{F}$ and $U\cap V\neq\phi$, then $U\cup V$ is evenly covered by π.

The existence of such a covering is an easy consequence of the fact that X has a metric topology. Now suppose that $U\in\mathcal{F}$ and $U_\alpha\subset\mathcal{R}(X)$ is such that $\pi:U_\alpha\to U$ is a diffeomorphism. We define a trivialization, $\phi(U_\alpha)$ of $\pi_\rho^{-1}(U)$, by sending each equivalence class,

$$[(\tilde{x},u)] := \{(y,v):(\tilde{y},v)=g\cdot(\tilde{x},u)\quad\text{for }\ g\in\pi_1(X,x^0)\},$$

in $\pi_\rho^{-1}(U)$ into the unique representative of the form $(\tilde{x},u)$ with $\tilde{x}\in U_\alpha$ and $u\in E_0$. Then

$$\phi(U_\alpha)[(\tilde{x},u)]:=(\pi(\tilde{x}),u)=(x,u)\in U\times E_0.$$

Now suppose that $U, V\in\mathcal{F}$ and $\pi:U_\alpha\to U$ and $\pi:V_\beta\to V$ are diffeomorphisms (note that $U=V$ is a possibility). Then if $U\cap V\neq\phi$, it follows that there exists a unique $g_{\alpha\beta}\in\pi_1(X,x^0)$ so that $U_\alpha\cap g_{\alpha\beta}V_\beta\neq\phi$. The existence of such a $g_{\alpha\beta}$ is trivial and uniqueness is equivalent to the assertion that if $U_\alpha\cap V_\beta\neq\phi$ and $U_\alpha\cap gV_\beta\neq\phi$ for some $g\in\pi_1(X,x^0)$ then $g=1$. Suppose then that $U_\alpha\cap V_\beta\neq\phi$ and $U_\alpha\cap gV_\beta\neq\phi$. Then since $U\cup V$ is evenly covered, we have $(U_\alpha\cup V_\beta)\cap g(U_\alpha\cup V_\beta)=\phi$ if $g\neq 1$. But evidently,

$$U_\alpha\cap gV_\beta\subset(U_\alpha\cup V_\beta)\cap g(U_\alpha\cup V_\beta)=\phi,$$

so $U_\alpha\cap gV_\beta=\phi$ if $g\neq 1$. Uniqueness follows. One may now easily compute

$$\phi(U_\alpha)\phi(V_\beta)^{-1}(x,u)=(x,\rho(g_{\alpha\beta})u). \tag{2.7}$$

Since these transition functions are constant in the base variables there is a globally defined flat connection $\boldsymbol{\nabla}_\rho$ on $\mathcal{R}(X)\times_\rho E_0$ that is obtained by gluing together the exterior derivative in the base variables defined in each of the trivializations $\phi(U_\alpha)$. In these trivializations it is not hard to check the isomorphism of connections (2.5).

To finish this account of the reconstruction of a bundle with a flat connection and prescribed holonomy it is still necessary to check that $(\mathcal{R}(X)\times_\rho E_0,\boldsymbol{\nabla}_\rho)$ has a holonomy at x^0 given by ρ. Suppose that γ is a piecewise smooth loop in X based at x^0 with $g=[\gamma]\in\pi_1(X,x^0)$. Let $\gamma^{-1}(\mathcal{F})$ be the open covering of the interval $[0, 1]$ by the inverse image of sets from $\mathcal{F}$ under γ. Let $\delta>0$ be a Lebesque number for this covering and suppose that $0=t_0<t_1\cdots<t_n=1$ is a partition of $[0, 1]$ with $t_{j+1}-t_j<\delta$ for all j. Then each curve segment $\{\gamma(t):t\in[t_j,t_{j+1}]\}$ lies inside some $U_j\in\mathcal{F}$. For each j one can find $U_{\alpha_j}\subset\mathcal{R}(X)$ so that $\pi:U_{\alpha_j}\to U_j$ is a diffeomorphism and one can also arrange that $U_{\alpha_{j+1}}\cap U_{\alpha_j}\neq\phi$ for $j=0,...,n-1$. Then parallel translation of a vector u in E_0 along γ is constant in the trivializations $\phi(U_{\alpha_j})$ for $j=0,...,n-1$. To compute the holonomy one must only compute what the vector u in the trivialization $\phi(U_{\alpha_{n-1}})$ looks like in the trivialization $\phi(U_{\alpha_0})$. However, it is clear from the construction that $gU_{\alpha_0}\cap U_{\alpha_{n-1}}\neq\phi$. Thus, the parallel transport of $u\in E_0$ along γ gives $\rho(g)^{-1}u\in E_0$. This finishes our sketch of the reconstruction of a vector bundle with flat connection (up to equivalence) from its holonomy representation. We now begin to explain the setting for Theorem 2.9 below.

B. The vector bundle deformation

Suppose that $\{a_1^0,a_2^0,\dots,a_n^0\}$ is a collection of n distinct points in $\mathbf{C}$. In this section we will construct a global deformation for a connection $\overline{\boldsymbol{\nabla}}^0$ defined on the trivial bundle $E^0:=\mathbf{P}^1\times\mathbf{C}^p$, with a simple type r_j singularity at a_j^0 and a regular point or simple type r_∞ singularity at ∞. For simplicity in stating results, when $\overline{\boldsymbol{\nabla}}^0$ has a regular point at ∞ we put $r_\infty=-1$ and say that $\overline{\boldsymbol{\nabla}}^0$ has a simple type 1 singularity.

Roughly speaking the deformation we consider will preserve the local type of the singularities for $\overline{\boldsymbol{\nabla}}^0$ and also the local and global monodromy data for the connection $\overline{\boldsymbol{\nabla}}^0$. We now make this more precise.

By relabling the points if necessary we may suppose that $r_j\geqslant 1$ for $j=1,2,...,m$ and $r_j=0$ for $j=m+1,m+2,...,n$. It could happen that $m=0$ or $m=n$. The point ∞ is somewhat special in this context since it does not contribute to the space of pole deformations, $\mathcal{R}(Z^n)$, which we defined in Sec. I. We will mention special considerations concerning ∞ when we encounter them.

For $j=1,...,m$ let

$$\mathcal{C}_j:=Z^p\times\mathbf{C}^p\times\cdots\times\mathbf{C}^p \quad \text{with } r_j-1 \quad \text{factors } \mathbf{C}^\rho,$$

denote the local configuration space at a_j^0, as described in Sec. I. Recall that each point in $\mathcal{C}_j$ corresponds to a formal equivalence class for a simple type r_j connection at a_j^0. Write $\mathcal{C}_\infty$ for the corresponding configuration space at ∞, defined if $r_\infty \geqslant 1$.

Write $\Lambda_j^0 \in \mathcal{C}_j$ for the data associated to $\overline{\boldsymbol{\nabla}}^0$ at a_j^0, for $j=1,...,m$. Recall that $\mathcal{R}(X)$ is just the simply connected cover of X, and define

$$\mathcal{D} := \mathcal{R}(Z^n) \times \prod_{j=1}^{m} \mathcal{R}(\mathcal{C}_j) \times \mathcal{R}(\mathcal{C}_\infty), \tag{2.8}$$

where the product is just the Cartesian product, and the final factor $\mathcal{R}(\mathcal{C}_\infty)$ only appears if $r_\infty \geqslant 1$.

The space $\mathcal{D}$ will serve as our ''deformation space.'' In the first factor, Z^n, is the space of pole locations and in each of the subsequent factors, $\mathcal{C}_j$, is the space of local formal equivalence classes at a_j^0. We must pass to the simply connected cover in $\mathcal{D}$ to guarantee global existence for the sort of deformation we are about to describe.

Let $\mathcal{M}_j \to \mathcal{C}_j$ denote the fiber bundle over $\mathcal{C}_j$ whose fiber over $\Lambda \in \mathcal{C}_j$ is the holomorphic equivalence class of connections formally equivalent to the diagonal model (1.12) associated to the base point Λ (defined in Sec. I). As in Sec. I we also write $\mathcal{M}_j \to \mathcal{R}(\mathcal{C}_j)$ for the pull-back of $\mathcal{M}_j$ under the projection $\mathcal{R}(\mathcal{C}_j) \to \mathcal{C}_j$. Let $\sigma_j^0 \in \mathcal{M}_j$ denote the point in the fiber associated to the class of $\overline{\boldsymbol{\nabla}}^0$ in a neighborhood of $x=a_j^0$ (σ_∞^0 is also defined if $r_\infty \geqslant 1$). Let $\lambda \to \sigma_j(\lambda)$ denote the unique flat section of $\mathcal{M}_j \to \mathcal{R}(\mathcal{C}_j)$ with $\sigma_j(\Lambda_j^0) = \sigma_j^0$ [$\sigma_\infty(\lambda)$ is defined in a similar fashion if $r_\infty \geqslant 1$].

Recall that Y_k is the subset of points $(x,t) \in \mathbf{P}^1 \times \mathcal{R}(Z^n)$ with $x=a_k(t)$, and for $j=1,2,...,n$ define

$$\mathcal{Y}_k = Y_k \times \prod_{j=1}^{m} \mathcal{R}(\mathcal{C}_j) \times \mathcal{R}(\mathcal{C}_\infty) \subset \mathbf{P}^1 \times \mathcal{D},$$

where the factor $\mathcal{R}(\mathcal{C}_\infty)$ is present only if $r_\infty \geqslant 1$. Let $t^0 \in \mathcal{R}(Z^n)$ denote a point in the covering space of Z^n such that $a_j(t^0) = a_j^0$, let $\lambda_j^0 \in \mathcal{R}(\mathcal{C}_j)$ denote a point in the covering space of $\mathcal{C}_j$ such that $\Lambda_j^0 = \Lambda_j(\lambda_j^0)$ [where Λ_j is the projection from $\mathcal{R}(\mathcal{C}_j)$ to $\mathcal{C}_j$] and write

$$\mathbf{P}^1 \ni x \to i(x) := (x, t^0, \lambda^0) \in \mathbf{P}^1 \times \mathcal{D},$$

where

$$\lambda^0 = (\lambda_1^0, \ldots, \lambda_m^0, \lambda_\infty^0),$$

and, as above, λ_∞^0 only occurs when $r_\infty \geqslant 1$.

The following theorem is due to Malgrange (Ref. 14, Theorem 3.1).

Theorem 2.9: *There exists a rank p holomorphic vector bundle* $E \to \mathbf{P}^1 \times \mathcal{D}$ *and an integrable connection* $\boldsymbol{\nabla}$ *on E with a simple type* r_j *singularity along* $\mathcal{Y}_j$ *for* $j=1,...,n$ *and a simple type* r_∞ *singularity along* $\mathcal{Y}_\infty$ *such that the restriction of* $(E, \boldsymbol{\nabla})$ *to* $\mathbf{P}^1 \times \{(t^0, \lambda^0)\}$ is equivalent to $(E^0, \overline{\boldsymbol{\nabla}}^0)$ [that is, $i^*(E,\boldsymbol{\nabla}) \simeq (E^0, \overline{\boldsymbol{\nabla}}^0)$]. *Furthermore, for* $j=1,...,m$ *the restriction of* $(E, \boldsymbol{\nabla})$ to $\mathbf{P}^1 \times \{(t,\lambda)\}$ *is formally equivalent to the model connection* $\overline{\boldsymbol{\nabla}}_{\Lambda_j(\lambda_j)}$ (1.12) *near* $x=a_j(t)$ *and is in the holomorphic equivalence class* $\sigma_j(\lambda_j) \in \mathcal{M}_j$.

Proof: We will prove this result as Malgrange does by first constructing the deformation in the complement of

$$\mathcal{Y} = \cup_{j=1}^{n} \mathcal{Y}_j \cup \mathcal{Y}_\infty \subset \mathbf{P}^1 \times \mathcal{D},$$

and then extending the connection $\overline{\nabla}^0$ to a tubular neighborhood $\mathcal{T}(\mathcal{Y}_j)(\mathcal{T}(\mathcal{Y}_\infty))$ of each singular set $\mathcal{Y}_j(\mathcal{Y}_\infty)$ so that it has the right local characteristics. In particular, one finds that the two constructions must be holomorphically equivalent on $\mathcal{T}(\mathcal{Y}_k)\backslash\mathcal{Y}_k$ and this equivalence allows one to define a bundle over all $\mathbf{P}^1\times D$ together with a connection that has the right global and local properties.

An important result for the construction of the deformation on $\mathbf{P}^1\times D\backslash\mathcal{Y}$ is the following observation of Malgrange. Choose some point $x^0\in\mathbf{C}$ so that $x^0\neq a_j(t^0)$ for all $j=1,...,n$. Define $p^0=(x^0,t^0,\lambda^0)$. Then the map

$$\mathbf{P}^1\backslash\{a_1^0,a_2^0,\ldots,a_n^0,\infty\}\ni x\rightarrow(x,t^0,\lambda^0)\in\mathbf{P}^1\times\mathcal{D}\backslash\mathcal{Y}, \tag{2.10}$$

induces an isomorphism of fundamental groups,

$$\pi_1(\mathbf{P}^1\backslash\{a_1^0,a_2^0,\ldots,a_n^0,\infty\},x^0)\simeq\pi_1(\mathbf{P}^1\times\mathcal{D}\backslash\mathcal{Y},p^0). \tag{2.11}$$

This is explained in both Refs. 3 and 4, where the deformation space does not include the factors $\mathcal{R}(\mathcal{C}_j)$. However, the product of these factors is simply connected so it does not influence the result. The holonomy of the connection $\overline{\nabla}^0$ at the base point x^0 determines a representation, ρ, of $\pi_1(\mathbf{P}^1\backslash\{a_1^0,a_2^0,\ldots,a_n^0,\infty\},x^0)$ on $\mathrm{GL}(p,\mathbf{C})$. The isomorphism (2.11) and the representation ρ determines a $\mathrm{GL}(p,\mathbf{C})$ representation of $\pi_1(\mathbf{P}^1\times\mathcal{D}\backslash\mathcal{Y},p^0)$ that we continue to denote by ρ. Associated with this representation is a vector bundle $E_\rho:=(\mathbf{P}^1\times\mathcal{D}\backslash\mathcal{Y})\times_\rho\mathbf{C}^p$ with connection ∇_ρ whose holonomy representation is given by ρ.

Next we turn to the construction of the local deformations. Suppose that $a=(a_1,a_2,\ldots,a_n)\in Z^n$ and let $\delta(a)$ denote the minimum of the distances, $\{|a_i-a_j|,|a_i^0-a_j^0|\}_{i\neq j}$ [the reason for insisting that $\delta(a)\leqslant\min\{|a_i^0-a_j^0|\}_{i\neq j}$ will appear below]. Let $D_j(a)$ denote the disk of radius $\delta(a)/3$ about the point a_j. Let $D_\infty(a)$ denote the open complement of the closed disk of radius $\delta(a)+\max_i\{|a_i|,|a_i^0|\}$. It is clear by construction that the disks $\{D_\infty(a),D_j(a),j=1,...,n\}$ are pairwise disjoint for each $a\in Z^n$. Define a tubular neighborhood, $\mathcal{T}(\mathcal{Y}_j)$, of $\mathcal{Y}_j$ by

$$\mathcal{T}(\mathcal{Y}_j)=\{(x,t,\lambda)\in\mathbf{P}^1\times\mathcal{D}|x\in D_j(a(t))\},$$

and a tubular neighborhood of $\mathcal{Y}_\infty$ by

$$\mathcal{T}(\mathcal{Y}_\infty):=\{(x,t,\lambda)\in\mathbf{P}^1\times\mathcal{D}|x\in D_\infty(a(t))\}.$$

Then the neighborhoods $\{\mathcal{T}(\mathcal{Y}_\infty),\mathcal{T}(\mathcal{Y}_j),j=1,...,n\}$ are pairwise disjoint. Following the scheme that can be found in Malgrange[3] we define a connection on the trivial bundle $\mathcal{T}(\mathcal{Y}_j)\times\mathbf{C}^p\rightarrow\mathcal{T}(\mathcal{Y}_j)$ by lifting the connection on the trivial bundle over $D_j(a^0)\times\mathcal{R}(\mathcal{C}_j)$ that one obtains from Theorem 1.26 above. Recall that $\lambda\rightarrow\sigma_j(\lambda)$ is the unique flat section of $\mathcal{M}_j\rightarrow\mathcal{R}(\mathcal{C}_j)$ with $\sigma_j(\Lambda_j^0)=\sigma_j^0$. For $j=1,...,m$ let ∇_j denote the integrable connection on the trivial bundle,

$$D_j(a^0)\times\mathcal{R}(\mathcal{C}_j)\times\mathbf{C}^p\rightarrow D_j(a^0)\times\mathcal{R}(\mathcal{C}_j),$$

with a simple type r_j singularity along $\{a_j^0\}\times\mathcal{R}(\mathcal{C}_j)$, whose existence is guaranteed by Theorem 1.26. This connection naturally extends to a connection on the trivial bundle,

$$D_j(a^0)\times\mathcal{R}(\mathcal{C})\times\mathbf{C}^p\rightarrow D_j(a^0)\times\mathcal{R}(\mathcal{C}),$$

by pulling back the connection one-form under the natural projection,

$$D_j(a_0)\times\mathcal{R}(\mathcal{C})\rightarrow D_j(a_0)\times\mathcal{R}(\mathcal{C}_j),$$

where we have written

$$\mathcal{R}(\mathcal{C}) := \prod_{j=1}^{m} \mathcal{R}(\mathcal{C}_j) \times \mathcal{R}(\mathcal{C}_\infty).$$

We continue to denote this connection by $\boldsymbol{\nabla}_j$. Now define a map,

$$\mathrm{pr}_j : \mathcal{T}(\mathcal{Y}_j) \to D_j(a^0) \times \mathcal{R}(\mathcal{C}), \tag{2.12}$$

by $\mathrm{pr}_j(x,t,\lambda) = (x - a_j(t) + a_j^0, \lambda)$ [the extra condition that $\delta(a) \leqslant \min\{|a_i^0 - a_j^0|\}_{i\neq j}$ now guarantees that $x - a_j(t) + a_j^0 \in D_j(a^0)$ for $x \in D_j(a^0)$. Let $\Omega_j(x,\lambda_j)$ denote the one-form for $\boldsymbol{\nabla}_j$ [and we have written $\Omega_j(x,\lambda_j)$ to emphasize the fact that Ω_j only depends on the variables (x,λ_j)],

$$\boldsymbol{\nabla}_j = d + \Omega_j(x,\lambda_j), \tag{2.13}$$

and define a connection (that we again call $\boldsymbol{\nabla}_j$!) on the trivial bundle $E_j := \mathcal{T}(\mathcal{Y}_j) \times \mathbf{C}^p \to \mathcal{T}(\mathcal{Y}_j)$ by

$$\boldsymbol{\nabla}_j := d + \mathrm{pr}_j^*(\Omega_j), \tag{2.14}$$

where d denotes the exterior derivative on $\mathcal{T}(\mathcal{Y}_j)$ (acting on $\mathbf{C}^p$-valued functions). It is easy to check that connection (2.14) is integrable and has a simple type r_j singularity along $\mathcal{Y}_j$ as a consequence of (2.13) being integrable with a simple type r_j singularity along $\{a_j^0\} \times \mathcal{R}(\mathcal{C})$. Now we wish to determine the holonomy for $\boldsymbol{\nabla}_j$ on $\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j$. The map

$$\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j \ni (x,t,\lambda) \to t \in \mathcal{R}(Z^n), \tag{2.15}$$

is surjective with fiber over t given by $D_j(a(t))\backslash\{a(t)\} \times \mathcal{R}(\mathcal{C})$ [which is homeomorphic to $D_j(a^0)\backslash\{a_j^0\} \times \mathcal{R}(\mathcal{C})$]. The first part of the homotopy exact sequence for this fiber bundle reads as

$$\to \pi_2(\mathcal{R}(Z^n)) \to \pi_1(\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j) \to \pi_1(D_j(a^0)\backslash\{a_j^0\}) \to \pi_1(\mathcal{R}(Z^n)) \to 0,$$

where we substituted $\pi_1(D_j(a^0)\backslash\{a_j^0\})$ for $\pi_1(D_j(a^0)\backslash\{a_j^0\} \times \mathcal{R}(\mathcal{C}))$. Thus, we have

$$\pi_1(\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j) \simeq \pi_1(D_j(a^0)\backslash\{a_j^0\}). \tag{2.16}$$

Let

$$(\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j)_{t=t^0} = \{(x,t^0,\lambda) \in \mathcal{T}(\mathcal{Y}_j)\backslash Y_j\} \simeq D_j(a^0)\backslash\{a_j^0\} \times \mathcal{R}(\mathcal{C}),$$

Then the restriction of pr_j,

$$\mathrm{pr}_j : (\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j)_{t=t^0} \to D_j(a^0)\backslash\{a_j^0\} \times \mathcal{R}(\mathcal{C}), \tag{2.17}$$

is essentially the identity. Since (2.16) shows that representatives of all the homotopy classes of curves in $\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j$ can be found among the loops that stay in the section $(\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j)_{t=t^0}$, it follows that the holonomy of the connection $\boldsymbol{\nabla}_j$ can be computed from its restriction to $(\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j)_{t=t^0}$. But the identification (2.16) shows that this holonomy is the same as the holonomy of the connection coming from Theorem 1.26 on the the space $D_j(a^0)\backslash\{a_j^0\} \times \mathcal{R}(\mathcal{C})$. Since $\pi_1(\mathcal{R}(\mathcal{C})) = 0$ we may compute this holonomy by restricting to $\lambda = \lambda^0$ where the connection agrees with $\overline{\boldsymbol{\nabla}}^0$ by construction.

We also need to construct a model connection in the tubular neighborhoods $\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j$ for $j = m+1,...,n$ and $\mathcal{T}(\mathcal{Y}_\infty)\backslash\mathcal{Y}_\infty$. In the first instance, we are looking for a connection on $\mathcal{T}(\mathcal{Y}_j)$ with a simple type 0 singularity along $\mathcal{Y}_j$, which is equivalent to $\overline{\boldsymbol{\nabla}}^0$ in a neighborhood of a_j^0. Since $\overline{\boldsymbol{\nabla}}^0$ is a simple type 0 connection, Proposition 1.25b shows that in a neighborhood of $x = a_j^0$ there exists a diagonal matrix Λ_j so that in the appropriate local trivialization about $x = a_j^0$, $\overline{\boldsymbol{\nabla}}^0$ is represented by

$$d_x+\Omega_j\,,$$

on the trivial bundle

$$D_j(a^0)\times\mathbf{C}^p\to D_j(a^0), \tag{2.18}$$

where

$$\Omega_j=-\frac{\Lambda_j d_x(x-a_j^0)}{x-a_j^0}.$$

Define a connection ∇_j on the trivial bundle over $\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j$ by

$$\nabla_j=d-\frac{\Lambda_j d(x-a_j(t))}{x-a_j(t)},$$

where $d=d_x+d_t+d_\lambda$ is the exterior derivative on $\mathcal{T}(\mathcal{Y}_j)$. It is easy to check that ∇_j is a simple integrable connection with type 0 singularity along $\mathcal{Y}_j$ that has a restriction to $D_j(a^0)\times\{(t^0,\lambda^0)\}$ that is equivalent to $\overline{\nabla}^0$ by construction. The same argument given above for $j=1,...,m$ applies here, and one sees that the connection ∇_j defined on the trivial bundle,

$$\mathcal{T}(\mathcal{Y}_j)\times\mathbf{C}^p\to\mathcal{T}(\mathcal{Y}_j),$$

has holonomy determined by its restriction to $t=t^0$ and $\lambda=\lambda^0$, where it is essentially $\overline{\nabla}^0$. To extend the connection to $\mathcal{T}(\mathcal{Y}_\infty)\backslash\mathcal{Y}_\infty$ so that it is regular along $\mathcal{Y}_\infty$, or has a logarithmic pole along $\mathcal{Y}_\infty$, or a higher rank simple singularity one proceeds as above with some simplification arising from the fact that ∞ adds no component to the space of pole deformations. One should use the local parameter $w=1/x$ with

$$\mathrm{pr}_\infty(w,t,\lambda)=(w,\lambda).$$

We leave the details to the reader.

What we have now is a bundle E_ρ over $\mathbf{P}^1\times\mathcal{D}\backslash\mathcal{Y}$, bundles E_j over $\mathcal{T}(\mathcal{Y}_j)$, and E_∞ over $\mathcal{T}(\mathcal{Y}_\infty)$ with bundle isomorphisms,

$$b_j:E_\rho|_{\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j}\to E_j|_{\mathcal{T}(\mathcal{Y}_j)\backslash\mathcal{Y}_j}, \tag{2.19}$$

and

$$b_\infty:E_\rho|_{\mathcal{T}(\mathcal{Y}_\infty)\backslash\mathcal{Y}_\infty}\to E_\infty|_{\mathcal{T}(\mathcal{Y}_\infty)\backslash\mathcal{Y}_\infty}, \tag{2.20}$$

which take the connection ∇_ρ into ∇_j and ∇_∞. We define a vector bundle E over $\mathbf{P}^1\times\mathcal{D}$ by forming the union $E_\rho\cup E_1\cup\cdots\cup E_n\cup E_\infty$ modulo the equivalence relations determined by (2.19) and (2.20). By construction the bundle E and the connection ∇ obtained by gluing together the connections ∇_ρ, ∇_j, and ∇_∞ have the properties asserted in the formulation of Theorem 2.9. This finishes the proof. QED

III. TAU FUNCTIONS

In this section we will first follow Helmink[4] to define a tau function associated with the deformation construction of Theorem 2.0 above. Let (E,∇) be the integrable deformation of $(E^0,\overline{\nabla}^0)$ constructed in Theorem 2.9. Recall that $\mathcal{D}=\mathcal{R}(Z^n)\times\mathcal{R}(\mathcal{C})$ and define

$$\Theta:=\{(t,\lambda)\in\mathcal{D},E|_{\mathbf{P}^1\times\{(t,\lambda)\}}\text{is nontrivial}\}.$$

Theorem 3.0: *There exists a nonvanishing holomorphic map $\tau:\mathcal{D}\rightarrow\mathbf{C}$ so that Θ is equal to the zero set of τ.*

This is basically a result of Malgrange[14] and Helmink[4] that we will sketch a proof of following the arguments for Proposition 3.2 in Ref. 4.

Sketch of Proof (more details can be found in Ref. 4): Choose $0<\rho_1<1<\rho_2$ and set

$$D_1 := \{x | x\in\mathbf{P}^1, |x|>\rho_1\}, D_2 := \{x | x\in\mathbf{P}^1, |x|<\rho_2\}.$$

Then $D_k\times\mathcal{D}$ for $k=1,2$ is a contractible Stein space, and hence $E|_{D_k\times\mathcal{D}}$ is holomorphically trivial for $k=1,2$. Let $\mathbf{f}:=(f_1,f_2,\ldots,f_p)$ be a row vector of sections for the restriction of E to $D_1\times\mathcal{D}$ that trivializes this restriction. Let $\mathbf{g}:=(g_1,g_2,\ldots,g_p)$ be a row vector of sections for the restriction of E to $D_2\times\mathcal{D}$ that trivializes this restriction. Then there exists a holomorphic map $S:D_1\cap D_2\times\mathcal{D}\rightarrow GL(p,\mathbf{C})$ so that

$$\mathbf{g}=\mathbf{f}S$$

on $D_1\cap D_2\times\mathcal{D}$. Write $S(x,t,\lambda)=S_{t,\lambda}(x)$. Then the restriction of E to $\mathbf{P}^1\times\{(t,\lambda)\}$ will be trivial if and only if there are holomorphic maps $S^-_{t,\lambda}:D_1\rightarrow \mathrm{GL}(p,\mathbf{C})$ and $S^+_{t,\lambda}:D_2\rightarrow \mathrm{GL}(p,\mathbf{C})$, so that for all $x\in D_1\cap D_2$,

$$S_{t,\lambda}(x)=S^-_{t,\lambda}(x)S^+_{t,\lambda}(x)^{-1}. \tag{3.1}$$

A p vector of sections that trivializes E is then given by the appropriate extension of

$$\mathbf{g}S^+_{t,\lambda}=\mathbf{f}S^-_{t,\lambda},$$

from $D_1\cap D_2$. Let S^1 denote the unit circle, write $\mathcal{H}:=L^2(S^1,\mathbf{C}^p)$, and let $\mathcal{H}_+$ be the closed subspace of $\mathcal{H}$ consisting of those functions that are boundary value functions holomorphic inside the unit disk. Let $\mathcal{H}_-=\mathcal{H}_+^\perp$. Suppose that $S^1\ni x\rightarrow S(x)$ is a smooth $p\times p$ matrix-valued function on the circle and let $\mathcal{H}\ni f\rightarrow Sf$ denote the associated multiplication operator on $\mathcal{H}$. Let

$$S=\begin{bmatrix} a(S) & b(S)\\ c(S) & d(S)\end{bmatrix},$$

denote the decomposition of S relative to the direct sum decomposition $\mathcal{H}=\mathcal{H}_+\oplus\mathcal{H}_-$. It is well known[1] that the factorization (3.1) exists if and only if $a(S_{t,\lambda}):\mathcal{H}_+\rightarrow\mathcal{H}_+$ is invertible. Since $(t,\lambda)\rightarrow a(S_{t,\lambda})$ is continuous in the uniform norm topology for $a(S_{t,\lambda})$ and $a(S)$ is known to be Fredholm if S is smooth, it follows that the index of $a(S_{t,\lambda})$ is independent of (t, λ). Since $a(S_{t^0,\lambda^0})$ is invertible by construction, it follows that the index of $a(S_{t,\lambda})$ is 0 for all $(t,\lambda)\in\mathcal{D}$. Fix (t, λ) for the moment. Then since $a(S_{t,\lambda})$ has index 0, there exists a finite rank operator $k:\mathcal{H}_+\rightarrow\mathcal{H}_+$ so that $k+a(S_{t,\lambda})$ is invertible. In fact, there exists a neighborhood $V_{t,\lambda}$ of (t, λ) so that for all $(s,\mu)\in V_{t,\lambda}$ the operator $q_{s,\mu}:=k+a(S_{s,\mu})$ is invertible. Note that $q_{s,\mu}$ is a parametrix for $a(S_{s,\mu})$ for $(s,\mu)\in V_{t,\lambda}$ in that

$$a(S_{s,\mu})q^{-1}_{s,\mu}=I+F_{s,\mu},$$

where $F_{s,\mu}$ is a finite rank operator that depends holomorphically on (s,μ). Now we show that it is possible to make a coherent choice of parametrices $q_{s,\mu}$ following the argument in Helmink.[4] Let $\{V_i\}$ be a locally finite covering of $\mathcal{D}$ with $q_i(s,\mu)$ a holomorphic parametrix for $a(S_{s,\mu})$ for all $(s,\mu)\in V_i$. For each i,j such that $V_i\cap V_j\neq\phi$ define $q_iq_j^{-1}=\phi_{ij}$. Note that each ϕ_{ij} is a finite rank perturbation of the identity. The maps $V_i\cap V_j\ni(s,\mu)\rightarrow\det\phi_{ij}(s,\mu)$ are the transition functions for a holomorphic line bundle on $\mathcal{D}$. Since $\mathcal{D}$ is a Stein space we have $H^1(\mathcal{D},\mathcal{O}^*)=0$, and it follows that this line bundle must be trivial. Hence there exist holomorphic maps $\tau_i:V_i\rightarrow\mathbf{C}^*$ so that $\tau_i^{-1}\tau_j = \det \phi_{ij}$. Now define $t_i : \mathcal{H}_+ \to \mathcal{H}_+$ by $t_i 1 = \tau_i$ and $t_i(e^{in\theta}) = e^{in\theta}$ for $n = 1,2,\ldots$. Now define $\mathbf{q}_i = t_i q_i$. Then $\mathbf{q}_i(s,\mu)$ remains a holomorphic parametrix for $a(S_{s,\mu})$ for $(s,\mu)\in V_i$ and since

$$\det \mathbf{q}_i \mathbf{q}_j^{-1} = \det t_i t_j^{-1} \det q_i q_j^{-1} = \tau_i \tau_j^{-1} \det \phi_{ij} = 1,$$

it follows that

$$\tau(s,\mu) := \det(a(S_{s,\mu})\mathbf{q}_i(s,\mu)^{-1}),$$

is a well-defined holomorphic function on all of $\mathcal{D}$ whose 0 set is equal to Θ. QED

Next, we will make a connection with the tau function defined by Jimbo, Miwa, and Ueno.[8] We follow Malgrange by computing the regularized logarithmic derivative of the determinant of a Toeplitz operator, the invertibility of which determines if $E|_{\mathbf{P}^1\times\{(t,\lambda)\}}$ is trivial or not. The bundle E is first realized in terms of a system of transition functions that relate local models for the connection ∇ in a neighborhood of the singularities to a model for the connection ∇ in a complement of a neighborhood of the singularities. We will spend some effort to choose the local models carefully [so that (3.7) is satisfied], even though this is not important for the calculation of the regularized logarithmic derivative of the Toeplitz operator. We do it because it will simplify a curvature calculation later on.

First fix a point $(t^0,\lambda^0)\in\mathcal{D}$. We will examine the restriction of E to $\mathbf{P}^1\times W$, where W is a suitably small neighborhood of (t^0,λ^0). We will cover $\mathbf{P}^1\times W$ by neighborhoods $B_j\times W$ that contain the singular sets $x = a_j(t)$ (including $x=\infty$ for $j=\infty$) and a complementary set $B_{\text{ex}}\times W$. We will choose trivializations for E over each of these sets and understand the bundle E in terms of the transition maps between these trivializations.

Now we turn to the identification of a suitable local model for the connection ∇ in a neighborhood of each singularity.

For each $j=1,\ldots,m$ choose a connected, simply connected product neighborhood $W_j = U_j \times V_j$ of (t_j^0,λ_j^0), with compact closure. For $j=m+1,\ldots,n (j=\infty)$ choose a connected, simply connected neighborhood of $t_j^0(\lambda_\infty^0)$ with compact closure. Let

$$W = \prod_{j=1}^{n} W_j \times W_\infty .$$

Let $D_j(a^0)$ denote the disk of radius $\delta(a^0)/3$ defined in Sec. II and let $D_j'(a^0)\subseteq D_j(a^0)$ denote the disk of radius $\delta(a^0)/6$. Choose a trivialization, $\mathbf{g}^j$, for the restriction $E|_{D_j(a^0)\times W}$ and suppose that relative to this trivialization the connection ∇ is given by

$$d + \Omega_j .$$

Let Ω_j^0 denote the restriction of Ω_j to $(t_j,\lambda_j) = (t_j^0,\lambda_j^0)$. As in the developments preceding Theorem 1.35, it is possible to adjust the trivialization $\mathbf{g}^j$ by a gauge transformation in the x variables alone so that $d_x + \Omega_j^0$ extends to a connection on the trivial bundle $\mathbf{P}^1\times\mathbf{C}^p\to\mathbf{P}^1$ with a simple type r singularity at a_j^0 and a regular singularity at ∞. In what follows we suppose that this has been done. For $j=1,\ldots,m$ choose V_j small enough so that the Birkhoff deformation of $d_x+\Omega_j^0$ constructed in Proposition 1.35 exists on $D_j(a^0)\times V_j$. Thus there exists an integrable connection,

$$d_x + d_{\lambda_j} + \Omega_j^{\text{loc}}, \tag{3.2}$$

which is a Birkhoff deformation of $d_x+\Omega_j^0$ in the sense of Proposition 1.35 (i) and (ii) and such that on $D_j(a^0)\backslash\{a_j^0\}\times V_j$ one has the gauge equivalence

$$d_x + d_{\lambda_j} + \Omega_j^{\text{loc}} = \varphi_j \cdot [d_x + d_{\lambda_j} + \Omega_j^0], \tag{3.3}$$

where $x \to \varphi_j(x,\lambda_j)$ is holomorphic in the punctured disk $D_j(a^0)\backslash\{a_j^0\}$ and asymptotic to the identity as $x\to\infty$. Now choose the neighborhood U_j of $\{t_j^0\}$ small enough so that the map p_j, defined by

$$D_j'(a^0)\times U_j\times V_j \ni (x,t,\lambda)\to p_j(x,t,\lambda):=(x-a_j(t)+a_j^0,\lambda),$$

maps into $D_j(a^0)\times V_j$. Let $d_j=d_x+d_{t_j}+d_\lambda$, where d_λ is the exterior derivative on $\Pi_{k=1}^n \mathcal{R}(\mathcal{C}_k)\times\mathcal{R}(\mathcal{C}_\infty)$ and define the connection

$$d_j+p_j^*\Omega_j^{\text{loc}}, \tag{3.4}$$

on the trivial $\mathbf{C}^p$ bundle over

$$X_j:=D_j'(a^0)\times W.$$

It is a consequence of Proposition 1.26c above that the connection (3.4) on the trivial bundle over X_j is equivalent to $\boldsymbol{\nabla}$ on $E|_{X_j}$ (they have the same formal reduction and Stokes multipliers). Possibly shrinking the open neighborhood U_j, we can ensure that there is an annular region,

$$A_j=\{x|\rho_j<|x-a_j^0|<\rho_j'\}\subset D_j'(a^0),$$

with the property that $x-a_j(t)\neq 0$ for $(x,t)\in A_j\times U_j$. The integrable connection (3.4) and the integrable connection,

$$d_j+\Omega_j^{\text{loc}},$$

have the same holonomy on $A_j\times W_j$. They are thus related by a holomorphic gauge transformation, ψ_j, defined on $A_j\times W_j$, so that

$$d_j+p_j^*\Omega_j^{\text{loc}}=\psi_j\cdot[d_j+\Omega_j^{\text{loc}}], \tag{3.5}$$

which can be normalized so that $\psi_j(x,t_j^0,\lambda_j^0)=I$ for $x\in A_j$. By choosing W_j small enough, we may ensure that ψ_j is a sufficiently small perturbation of the identity so that it has a ''canonical factorization'' $\psi_j=\psi_j^+\psi_j^-$, where $\psi_j^+(x,t_j,\lambda_j)$ is holomorphic for $|x-a_j^0|<\rho_j'$ and $\psi_j^-(x,t_j,\lambda_j)$ is holomorphic for $|x-a_j^0|>\rho_j$, with $\psi_j^-(\infty,t_j,\lambda_j)=I$. Now, define the gauge transform of $d_j+p_j^*\Omega_j^{\text{loc}}$ by $(\psi_j^+)^{-1}$,

$$\boldsymbol{\nabla}_j:=(\psi^+)^{-1}\cdot[d_j+p_j^*\Omega_j^{\text{loc}}]=\psi_j^-\cdot[d_j+\Omega_j^{\text{loc}}]. \tag{3.6}$$

Combining (3.6) with the extension of (3.3) to

$$d_j+\Omega_j^{\text{loc}}=\varphi_j\cdot[d_j+\Omega_j^0],$$

which follows from (3.3), since φ_j does not depend on t_j, or λ_k for $k\neq j$, we also have

$$\boldsymbol{\nabla}_j=\phi_j\cdot[d_j+\Omega_j^0], \tag{3.7}$$

where $\phi_j(x,t_j,\lambda_j)$ is holomorphic for $|x-a_j^0|>\rho_j$. The connection $\boldsymbol{\nabla}_j$ is equivalent to the restriction of $\boldsymbol{\nabla}$ to $E|_{X_j}$ and so is a good local model for $\boldsymbol{\nabla}$. Thus, we can choose a trivialization $\mathbf{f}^j$ for $E|_{X_j}$ so that in this trivialization $\boldsymbol{\nabla}$ is represented by $\boldsymbol{\nabla}_j$.

Now we wish to do something similar for $j=m+1,...,n$. W_j should be small enough so the map

$$D_j'(a^0)\times W_j\ni(x,t)\to p_j(x,t):=x-a_j(t)+a_j^0,$$

maps into $D_j(a^0)$. Choose a trivialization for $E|_{D_j(a^0)\times W}$ and write Ω_j for the connection one-form for ∇ in this trivialization. Thus, ∇ is represented by

$$d+\Omega_j,$$

in this trivialization.

Let Ω_j^0 denote the restriction of Ω_j to $(t,\lambda)=(t^0,\lambda^0)$. Again, using the same argument to be found in the preliminaries to Proposition 1.35, we may suppose that the trivialization of $E|_{D_j(a^0)\times W}$ has been chosen so that $d_x+\Omega_j^0$ extends to a meromorphic connection on the trivial bundle $\mathbf{P}^1\times\mathbf{C}^p\rightarrow\mathbf{P}^1$ with a simple type 0 singularity at $x=a$ and a regular singularity at $x=\infty$.

As above, consider the connection,

$$d_j+p_j^*\Omega_j^0.$$

One can choose W_j small enough so that there exists an annular region $A_j=\{x|\rho_j<|x-a_j^0|<\rho_j'\}$ contained in $D_j'(a^0)$ with the property that $x-a_j(t)\neq 0$ for $(x,t_j)\in A_j\times W_j$. The connection $d_j+p_j^*\Omega_j^0$ and $d_j+\Omega_j^0$ then have the same holonomy over $A_j\times W$ and so there exists a holomorphic gauge transformation ψ_j defined on $A_j\times W$ so that

$$d_j+p_j^*\Omega_j^0=\psi_j\cdot[d_j+\Omega_j^0].$$

One can normalize ψ_j so that $\psi_j(x,t^0,\lambda^0)=I$ and by choosing W sufficiently small we can guarantee that ψ_j has a canonical factorization $\psi_j^+\psi_j^-$ as above. We define

$$\nabla_j:=(\psi_j^+)^{-1}\cdot[d_j+p_j^*\Omega_j^0]=\psi_j^-\cdot[d_j+\Omega_j^0], \tag{3.8}$$

where $\psi_j^-(x,t,\lambda)$ is holomorphic for $|x-a_j^0|>\rho_j$ and asymptotic to the identity I as $x\rightarrow\infty$. As above, we can find a trivialization $\mathbf{f}^j$ for the restriction of E to $D_j(a^0)\times W$ so that ∇ is represented by ∇_j.

The local connection at infinity, ∇_∞, might be regular, or have a simple type r_∞ singularity. It does not matter much for the calculation we are going to do, but for definiteness we suppose $r_\infty\geqslant 1$. Then W_∞ is a neighborhood of λ_∞^0 (no pole deformation parameters). Proceeding in close analogy with the first case discussed above, we choose a trivialization $\mathbf{g}^\infty$ for E restricted to $D_\infty(a^0)\times W$ so that the connection ∇ is represented by

$$d_x+d_\lambda+\Omega_\infty.$$

We further suppose that the trivialization $\mathbf{g}^\infty$ has been chosen so that the restriction of this connection to $\lambda=\lambda^0$, given by $d_x+\Omega_\infty^0$, extends to a connection on the trivial bundle $\mathbf{P}^1\times\mathbf{C}^p\rightarrow\mathbf{P}^1$ with a simple type r_∞ singularity at ∞ and a regular singular point at 0.

Now let

$$d_x+d_{\lambda_\infty}+\Omega_\infty^{\mathrm{loc}}$$

denote the Birkhoff deformation of $d_x+\Omega_\infty^0$ constructed in Proposition 1.35 (with the slight alterations needed to locate the singularity at ∞). Let p_∞ denote the projection $(x,\lambda)\rightarrow(x,\lambda_\infty)$ and define

$$\nabla_\infty:=d_x+d_\lambda+p_\infty^*\Omega_\infty^{\mathrm{loc}}. \tag{3.9}$$

Choose ρ_∞ and ρ_∞' so that ∇_∞ and $d_x+d_\lambda+\Omega_\infty^0$ are holomorphically equivalent on the annulus,

$$A_\infty=\{x|\rho_\infty<x<\rho_\infty'\}\subset D_\infty(a^0),$$

by a gauge transformation ϕ_∞,

$$\nabla_\infty = \phi_\infty \cdot [d_x + d_\lambda + \Omega_\infty^0], \tag{3.10}$$

which can be chosen so that $\phi_\infty(x,\lambda)$ is holomorphic for $|x| < \rho'_\infty$.

In contrast to Theorem 2.9 we do *not* assume that the restriction

$$E|_{\mathbf{P}^1 \times \{t^0,\lambda^0\}}$$

is holomorphically trivial. Let $B_j(\rho) = \{x : |x - a_j^0| < \rho\}$ denote the open ball of radius ρ about a_j^0. Let $B_\infty(\rho) = \{x : |x| > \rho\}$, and define

$$B_j := B_j(\rho'_j),$$

$$B_\infty := B_\infty(\rho_\infty),$$

$$B := B_1(\rho_1) \cup B_2(\rho_2) \cup \cdots \cup B_n(\rho_n) \cup B_\infty(\rho'_\infty),$$

$$B_{\text{ex}} := \mathbf{P}^1 \backslash \overline{B}, \tag{3.11}$$

where $\overline{X}$ denotes the closure of X. Then

$$\{B_1, B_2, \ldots, B_n, B_\infty, B_{\text{ex}}\} \tag{3.12}$$

is an open covering of $\mathbf{P}^1$.

We now wish to show that there is a holomorphic trivialization of the bundle E over $B_{\text{ex}} \times W$. Since we have seen that Θ is the zero set of a holomorphic function, τ which is not identically 0, it follows that there is some $(t^1,\lambda^1) \in W$ so that $E|_{\mathbf{P}^1 \times \{(t^1,\lambda^1)\}}$ is holomorphically trivial. Since $B_{\text{ex}} \times W$ does not intersect any of the singular sets (3.2) or $\{\infty\} \times W$ it follows that the integrable connection ∇ on E is smooth over $B_{\text{ex}} \times W$, and since W is connected and simply connected one may integrate ∇ over W to extend the trivialization over $B_{\text{ex}} \times \{(t^1,\lambda^1)\}$ to a holomorphic trivialization of E over $B_{\text{ex}} \times W$. We wish to pick a trivialization so that the connection form for ∇ is particularly simple. Suppose that relative to some choice of a trivialization for the restriction of E to $\mathbf{P}^1 \times \{(t^1,\lambda^1)\}$, the connection ∇ is

$$d_x + \Omega_{\text{ex}}(x),$$

where $\Omega_{\text{ex}}(x)$ depends only on x. Write $d = d_x + d_t + d_\lambda$; then, since W is connected and simply connected it is easy to see that the connection,

$$d + \Omega_{\text{ex}}(x), \tag{3.13}$$

defined on the trivial bundle over $B_{\text{ex}} \times W$ with fiber $\mathbf{C}^p$ has the same holonomy representation as ∇ on the restriction of E to $B_{\text{ex}} \times W$. The holomorphic equivalence of these two connections implies that we can choose a trivialization $\mathbf{f}^{\text{ex}}$ for $E|_{B_{\text{ex}} \times W}$ so that relative to this trivialization the connection ∇ is given by (3.13). Observe that this connection form has no dt or $d\lambda$ components.

For each $j = 1,...,n$ there exists a holomorphic map $S^j : A_j \times W \to \text{GL}(p,\mathbf{C})$ so that

$$\mathbf{f}^{\text{ex}} = \mathbf{f}^j S^j, \tag{3.14}$$

and a holomorphic map $S^\infty : A_\infty \times W \to \text{GL}(p,\mathbf{C})$ so that

$$\mathbf{f}^{\text{ex}} = \mathbf{f}^\infty S^\infty, \tag{3.15}$$

where we think of each trivialization $\mathbf{f} = (f_1, \ldots, f_p)$ as a row vector of sections.

Define

$$B_{\text{in}} = \cup_j B_j \cup B_\infty.$$

Let S denote the holomorphic map from $B_{\mathrm{in}}\cap B_{\mathrm{ex}}\times W$ into $\mathrm{GL}(p,\mathbf{C})$ that restricts to S^j on $A_j\times W$ and S^∞ on $A_\infty\times W$. Then for $(t,\lambda)\in W$, $E|_{\mathbf{P}^1\times\{(t,\lambda)\}}$ will be holomorphically trivial if and only if there exists a factorization,

$$S(x,t,\lambda)=\Phi_{\mathrm{in}}(x,t,\lambda)^{-1}\Phi_{\mathrm{ex}}(x,t,\lambda), \tag{3.16}$$

where $x\to\Phi_{\mathrm{in}}(x,t,\lambda)$ is holomorphic and invertible in B_{in} and $x\to\Phi_{\mathrm{ex}}(x,t,\lambda)$ is holomorphic and invertible in B_{ex}. Let $C_j(C_\infty)$ denote a counterclockwise oriented circle contained in $A_j(A_\infty)$, and define the oriented curve,

$$C=C_\infty-C_1-C_2-\cdots-C_n.$$

Choose a point $z_0\in B_{\mathrm{ex}}$ with $z_0\notin\overline{B_{\mathrm{in}}}$ and define a projection P_{ex} on $L^2(C)$ by

$$P_{\mathrm{ex}}f(z)=\int_C f(x)\left\{\frac{1}{x-z}-\frac{1}{x-z_0}\right\}\frac{dx}{2\pi i},$$

where $z\in C$ is a nontangential limit from B_{ex}. This projects onto the subspace of $L^2(C)$, which has a holomorphic extension into the connected part of B_{ex} bounded by the curve C, with the further property that this holomorphic extension vanishes at $z=z_0$. We can make the factorization (3.16) unique (when it exists) by normalizing

$$\Phi_{\mathrm{ex}}(z_0,t,\lambda)=\mathrm{identity}=I.$$

Rewriting (3.16), we have

$$\Phi_{\mathrm{in}}S=\Phi_{\mathrm{ex}},$$

and writing $P_{\mathrm{in}}=I-P_{\mathrm{ex}}$ we find

$$P_{\mathrm{in}}(\Phi_{\mathrm{in}}S)=I. \tag{3.17}$$

Regarding this as an equation for the rows of Φ_{in}, the solution of this equation is equivalent to the existence of the factorization (3.16). Suppose that (3.16) has a solution Φ_{in}, Φ_{ex}. Then the Toeplitz operator T_S defined by

$$T_Sf=P_{\mathrm{in}}(fS),$$

where f is a row vector in the range of P_{in}, has an inverse,

$$T_S^{-1}g=P_{\mathrm{in}}(g\Phi_{\mathrm{ex}}^{-1})\Phi_{\mathrm{in}}.$$

Following Malgrange, we now calculate the regularized trace,

$$\omega:=\mathrm{Tr}(T_S^{-1}T_{dS}-T_{dS\,S^{-1}}), \tag{3.18}$$

where for brevity we write $d=d_{t,\lambda}$.

Note: because the multiplication operator in our Toeplitz operator is acting on the right it is $T_{RS}-T_ST_R$, which is compact when R and S are smooth.

We will eventually show that (3.18) differs from $d\log\tau$, defined above, by a regular term, and this will allow us to make a connection with the formula for the tau function given by Jimbo, Miwa, and Ueno in Ref. 8.

We compute

$$T_S^{-1}T_{dS}f=P_{\rm in}((f\,dS)\Phi_{\rm ex}^{-1})\Phi_{\rm in}=P_{\rm in}(f\,dS\Phi_{\rm ex}^{-1})\Phi_{\rm in}=P_{\rm in}\big(f(\Phi_{\rm in}^{-1}\,d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}$$
$$-\Phi_{\rm in}^{-1}\,d\Phi_{\rm in}\,\Phi_{\rm in}^{-1})\big)\Phi_{\rm in}, \tag{3.19}$$

and

$$T_{dS\,S^{-1}}f=P_{\rm in}(f\,dS\,S^{-1})=P_{\rm in}\big(f(\Phi_{\rm in}^{-1}\,d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}\Phi_{\rm in}-\Phi_{\rm in}^{-1}\,d\Phi_{\rm in})\big). \tag{3.20}$$

Now let

$$R_qf=fq,$$

denote right multiplication by q, and define

$$Q=\Phi_{\rm in}^{-1}\,d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}-\Phi_{\rm in}^{-1}\,d\Phi_{\rm in}\,\Phi_{\rm in}^{-1}.$$

Then from (3.19) and (3.20), one sees that

$$T_S^{-1}T_{dS}-T_{dS\,S^{-1}}=R_{\Phi_{\rm in}}P_{\rm in}R_Q-P_{\rm in}R_{\Phi_{\rm in}}R_Q=[R_{\Phi_{\rm in}},P_{\rm in}]R_Q=[P_{\rm ex},R_{\Phi_{\rm in}}]R_Q.$$

Thus, the trace of interest is

$$\mathrm{Tr}[P_{\rm ex},R_{\Phi_{\rm in}}]R_Q, \tag{3.21}$$

and one computes

$$[P_{\rm ex},R_{\Phi_{\rm in}}]R_Qf(z)=\int_C f(x)Q(x)(\Phi_{\rm in}(x)-\Phi_{\rm in}(z))\left\{\frac{1}{x-z}-\frac{1}{x-z_0}\right\}\frac{dx}{2\pi i}.$$

Writing $d_x\Phi=\Phi'\,dx$, the integral of the (finite-dimensional) trace over the diagonal is then

$$\int_C \mathrm{Tr}\big(Q(x)\Phi_{\rm in}'(x)\big)\frac{dx}{2\pi i}=\int_C \mathrm{Tr}(d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}\Phi_{\rm in}'\Phi_{\rm in}^{-1}-d\Phi_{\rm in}\,\Phi_{\rm in}^{-1}\Phi_{\rm in}'\Phi_{\rm in}^{-1})\frac{dx}{2\pi i}$$
$$=\int_C \mathrm{Tr}(d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}\Phi_{\rm in}'\Phi_{\rm in}^{-1})\frac{dx}{2\pi i}, \tag{3.22}$$

where the second equality follows from Cauchy's theorem and the fact that $\Phi_{\rm in}$ is holomorphic in $B_{\rm in}$. To make a connection between (3.22) and the JMU[8] formula for the log derivative of the tau function we replace $\Phi_{\rm in}$ and $\Phi_{\rm ex}$ in (3.22) with quantities more intimately related to the connection $\boldsymbol{\nabla}$. First, choose $j\in\{1,...,n,\infty\}$, and let $\boldsymbol{\nabla}_j$ denote the representation for $\boldsymbol{\nabla}$ in the trivialization $\mathbf{f}^j$. Let $\boldsymbol{\nabla}_{\rm ex}$ denote the representation of $\boldsymbol{\nabla}$ in the trivialization $\mathbf{f}^{\rm ex}$. Then, in the trivialization $\mathbf{f}^j\Phi_{\rm in}^{-1}=\mathbf{f}^{\rm ex}\Phi_{\rm ex}^{-1}$ for the restriction of E to $\mathbf{P}^1\times(W\backslash\Theta)$, we find that the connection $\boldsymbol{\nabla}$ is given by

$$\boldsymbol{\nabla}=\Phi_{\rm ex}\cdot[\boldsymbol{\nabla}_{\rm ex}]=\Phi_j\cdot[\boldsymbol{\nabla}_j], \tag{3.23}$$

on $A_j\times W\backslash\Theta$ (or $A_\infty\times W\backslash\Theta$ for $j=\infty$).

Now fix $j\in\{1,...,m,\infty\}$ and let $\hat\alpha_j$ be the formal power series near $x=a_j(t)$ (or ∞ if $j=\infty$) for which

$$\hat\alpha_j\cdot[\boldsymbol{\nabla}_{\lambda_j}]=\boldsymbol{\nabla}. \tag{3.24}$$

Let $\hat\alpha_j(\mathrm{loc})$ denote the formal power series near $x=a_j(t)$, such that

$$\hat\alpha_j(\mathrm{loc})\cdot[\boldsymbol{\nabla}_{\lambda_j}]=\boldsymbol{\nabla}_j. \tag{3.25}$$

Equating (3.24) with the first term of (3.23) and recalling that $\Omega_{\rm ex}$ does not have any dt or $d\lambda$ terms, one finds

$$d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}=d\,\hat{\alpha}_j\hat{\alpha}_j^{-1}-\hat{\alpha}_j\,dH_j\,\hat{\alpha}_j^{-1}, \tag{3.26}$$

which should be understood in the following sense. Replacing $\hat{\alpha}_j$ by $\alpha_{\Sigma,j}$ defined in a suitable sector Σ with vertex at $x=a_j(t)$, one finds a sectorial version of (3.26). This shows that the function $d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}$ that is analytic in an annular region about $x=a_j(t)$ extends holomorphically into the sector Σ. Since this is true for a collection of sectors that cover a punctured neighborhood of $x=a_j(t)$, it follows that $d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}$ is holomorphic in a punctured neighborhood of $x=a_j(t)$. Equation (3.26) may then be understood as an equality of formal Laurent series (which, in fact, converge since the left-hand side has a convergent Laurent series).

Applying $\hat{\Phi}_{\rm in}$ (the formal power series associated to $\Phi_{\rm in}$) to both sides of (3.25) and comparing the result with (3.23) and (3.24), one finds

$$\hat{\Phi}_{\rm in}\hat{\alpha}_j(\mathrm{loc})=\hat{\alpha}_j c_j\,,$$

where c_j is a diagonal constant matrix (the only gauge automorphisms of $\boldsymbol{\nabla}_{\lambda_j}$ are diagonal constants). Thus

$$\hat{\Phi}_{in}=\hat{\alpha}_j c_j\hat{\alpha}_j(\mathrm{loc})^{-1}, \tag{3.27}$$

for $j=1,2,...,m,\infty$. This is to be understood in the sense of a formal power series at $x=a_j(t)$.

For $j=m+1,...,n$ let α_j denote the local holomorphic gauge transformation constructed in Proposition 1.25b such that

$$\alpha_j\cdot\left[d_x+d-\frac{\Lambda_j}{x-a_j(t)}\,d(x-a_j(t))\right]=\boldsymbol{\nabla}, \tag{3.28}$$

where Λ_j is a constant diagonal matrix. Let $\alpha_j(\mathrm{loc})$ denote the local holomorphic gauge transformation so that

$$\alpha_j(\mathrm{loc})\cdot\left[d_x+d-\frac{\Lambda_j}{x-a_j(t)}\,d(x-a_j(t))\right]=\boldsymbol{\nabla}_j\,. \tag{3.29}$$

Comparing this with the first term in (3.23) and making use of the fact that $\Omega_{\rm ex}$ has no dt or $d\lambda$ components, one finds

$$d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}=d\,\alpha_j\,\alpha_j^{-1}+\alpha_j\frac{\Lambda_j d a_j(t)}{x-a_j(t)}\,\alpha_j^{-1}. \tag{3.30}$$

If we write $dH_j=\Lambda_j d(x-a_j(t))/(x-a_j(t))$ for $j=m+1,...,n$ (remember $d=d_t+d_\lambda$), then (3.30) can be written as

$$d\Phi_{\rm ex}\,\Phi_{\rm ex}^{-1}=d\,\alpha_j\,\alpha_j^{-1}-\alpha_j\,dH_j\,\alpha_j^{-1}, \tag{3.31}$$

which is analogous to (3.26).

Applying $\Phi_{\rm in}$ to both sides of (3.29) and comparing with (3.28), we find

$$\Phi_{\rm in}=\alpha_j c_j\alpha_j(\mathrm{loc})^{-1},\quad \text{for }\ j=m+1,...,n, \tag{3.32}$$

which is analogous to (3.27).

The formal power series expansion for (3.26) and the powers series expansion for (3.30) shows that the integral,

$$\int_{C_j} \mathrm{Tr}(d\Phi_{\mathrm{ex}}\, \Phi_{\mathrm{ex}}^{-1} \Phi_{\mathrm{in}}' \Phi_{\mathrm{in}}^{-1})\, \frac{dx}{2\pi i},$$

can be ''done'' by residues to get

$$\pm \mathrm{Res}_{x=a_j(t)}\, \mathrm{Tr}(d\Phi_{\mathrm{ex}}\, \Phi_{\mathrm{ex}}^{-1} \Phi_{\mathrm{in}}' \Phi_{\mathrm{in}}^{-1}), \tag{3.33}$$

with the $+$ choice for $j=1,...,n$ and the $-$ choice for $j=\infty$. We now substitute (3.26), (3.27), (3.31), and (3.32) into (3.33). The first term in (3.26) and (3.30) does not make a contribution to the formal residue in (3.28) since $d\hat{\alpha}_j\, \hat{\alpha}_j^{-1}$ and $\Phi_{\mathrm{in}}' \Phi_{\mathrm{in}}^{-1}$ both have formal power series expansions at $x=a_j(t)$. After some simplification (using $c_j^{-1}\, dH_j\, c_j = dH_j$), one finds

$$\int_{C_j} \mathrm{Tr}(d\Phi_{\mathrm{ex}}\, \Phi_{\mathrm{ex}}^{-1} \Phi_{\mathrm{in}}' \Phi_{\mathrm{in}}^{-})\, \frac{dx}{2\pi i} = \pm \mathrm{Res}_j\, \mathrm{Tr}(\hat{\alpha}_j^{-1} \hat{\alpha}_j'\, dH_j - \hat{\alpha}_j(\mathrm{loc})^{-1} \hat{\alpha}_j'(\mathrm{loc}) dH_j), \tag{3.34}$$

where Res_j is the residue at $x=a_j(t)$ or ∞ if $j=\infty$, and the sign $\pm$ is $+$ for $j=1,...,n$ and $-$ for $j=\infty$. Combining (3.22) with (3.34), one finds that

$$\mathrm{Tr}(T_S^{-1} T_{dS} - T_{dS\, S^{-1}}) = -\sum_j \mathrm{Res}_j\, \mathrm{Tr}(\hat{\alpha}_j^{-1} \hat{\alpha}_j'\, dH_j - \hat{\alpha}_j(\mathrm{loc})^{-1} \hat{\alpha}_j'(\mathrm{loc}) dH_j), \tag{3.35}$$

where the sum is over $j \in \{1,...,n,\infty\}$.

Next, we make use of the special choice we made for the local models $\boldsymbol{\nabla}_j$ that is reflected in (3.7), (3.8), and (3.10). Following the arguments for the proof of part (iv) of Proposition 1.35, we find that

$$d_{t,\lambda}\, \mathrm{Res}_j\, \mathrm{Tr}(\hat{\alpha}_j(\mathrm{loc})^{-1} \hat{\alpha}_j'(\mathrm{loc}) dH_j) = 0, \tag{3.36}$$

for $j=1,...,n,\infty$ follows from (3.7), (3.8), and (3.10) in the same way that (1.42) follows from (1.41).

We will now make use of (3.36) to show that the one form in (3.35) is closed off the singular set Θ. One easily computes

$$d\, \mathrm{Tr}(T_S^{-1} T_{dS} - T_{dS\, S^{-1}}) = -\frac{1}{2} \sum_{j>k} \mathrm{Tr}([T_S^{-1} T_{\partial_k S}, T_S^{-1} T_{\partial_j S}] + T_{[\partial_j S S^{-1}, \partial_k S\, S^{-1}]}) ds_j \wedge ds_k,$$

where $s := (t,\lambda)$ and $\partial_k = \partial/\partial s_k$. This last expression can be computed as in Malgrange,[3] and one finds

$$\frac{1}{2} \sum_{j>k} \int_C \frac{dx}{2\pi i} \mathrm{Tr}(\partial_j S\, S^{-1} (\partial_k S\, S^{-1})') ds_j \wedge ds_k. \tag{3.37}$$

If $(t^0,\lambda^0) \notin \Theta$, then we may take $(t^1,\lambda^1) = (t^0,\lambda^0)$ in the construction above. Note that Eq. (3.35) shows that $\mathrm{Tr}(T_S^{-1} T_{dS} - T_{dS\, S^{-1}})$ is actually independent of the choice of (t^1,λ^1), though it does depend on (t^0,λ^0) through $\hat{\alpha}_j(\mathrm{loc})$. With this choice for (t^1,λ^1), Eqs. (3.7), (3.8) and (3.10) take on added significance. In this case one may choose a global trivialization for $E|_{\mathbf{P}^1 \times \{(t^0,\lambda^0)\}}$. If Ω_j^0 denotes the one-form for $\boldsymbol{\nabla}$ relative to this trivialization, then (3.7), (3.8), and (3.9) show that the transition function S_j can be chosen so that it has a holomorphic continuation into the exterior of C_j in $\mathbf{P}^1$ (including $j=\infty$). Each of the integrals,

$$\int_{C_j} \frac{dx}{2\pi i} \mathrm{Tr}(\partial_l S_j\, S_j^{-1} (\partial_k S_j\, S_j^{-1})'),$$

then vanishes by Cauchy's theorem. This shows that if $(t^0,\lambda^0)\notin\Theta$, then

$$d\,\mathrm{Tr}(T_S^{-1}T_{dS}-T_{dS\,S^{-1}})=0, \tag{3.38}$$

at least for the special constructions associated with $(t^0,\lambda^0)=(t^1,\lambda^1)$. But we can now use (3.35) and (3.36) to show that $\mathrm{Tr}(T_S^{-1}T_{dS}-T_{dS\,S^{-1}})$ is closed even without this restriction. Define

$$\omega_{\mathrm{JMU}}:=-\sum_j \mathrm{Res}_j\,\mathrm{Tr}(\hat{\alpha}_j^{-1}\hat{\alpha}_j'\,dH_j),$$

which is the one-form introduced by Jimbo, Miwa, and Ueno in Ref. 4. Then (3.38), (3.35), and (3.36) together show that

$$d\omega_{\mathrm{JMU}}=0, \tag{3.39}$$

for $(t,\lambda)\notin\Theta$. This result and (3.36) [which is not restricted to the special choice $(t^0,\lambda^0)=(t^1,\lambda^1)$] then show that the right-hand side of (3.35) is closed, in general, and we conclude that $\mathrm{Tr}(T_S^{-1}T_{dS}-T_{dS\,S^{-1}})$ is closed even when (t^1,λ^1) is different from (t^0,λ^0). We are now prepared to state the principal result of this paper.

Theorem 3.40: *Suppose that $\bar{\nabla}^0$ is a connection on the trivial bundle, $\mathbf{P}^1\times\mathbf{C}^p\to\mathbf{P}^1$, with simple type r_j singularities at the distinct points $a_j\in\mathbf{P}^1$ for $j\in\{1,2,...,n,\infty\}$ with $a_\infty:=\infty$. Let (E,∇) denote the vector bundle with connection constructed as a deformation of $\bar{\nabla}^0$ in Theorem 2.9. Let Θ denote the set of $(t,\lambda)\in\mathcal{D}$ such that $E|_{\mathbf{P}^1\times\{(t,\lambda)\}}$ is not trivial. For $(t,\lambda)\notin\Theta$ let $\hat{\alpha}_j$ and α_j be defined as in (3.24) and (3.28). Then we have the following.*

(i) *The form defined by*

$$\omega_{\mathrm{JMU}}=-\sum_j \mathrm{Res}_j\,\mathrm{Tr}(\hat{\alpha}_j^{-1}\hat{\alpha}_j'\,dH_j), \tag{3.41}$$

is a closed one-form on $\mathcal{D}\backslash\Theta$ and there exists a holomorphic function τ_{JMU} on $\mathcal{D}$ such that

$$\omega_{\mathrm{JMU}}=d\log\tau_{\mathrm{JMU}}.$$

(ii) *The point $(t,\lambda)\in\mathcal{D}$ is a zero of τ_{JMU} if and only if $(t,\lambda)\in\Theta$.*

Proof: Let S denote the transition function defined by (3.14) and (3.15). Then, as in the proof of Theorem 3.0, one can find an invertible holomorphic parametrix,

$$W\ni(t,\lambda)\to q(t,\lambda),$$

so that

$$T_{S(t,\lambda)}q(t,\lambda)^{-1}=I+\text{trace class},$$

for $(t,\lambda)\in W$. Define

$$\tau_q(t,\lambda):=\det(T_{S(t,\lambda)}q(t,\lambda)^{-1}).$$

Then it is clear that $\tau_q(t,\lambda)=0$ if and only if $(t,\lambda)\in\Theta$. As above, write $d=d_t+d_\lambda$. Then the usual formula for the derivative of a determinant gives

$$d\log\tau_q=\mathrm{Tr}(T_S^{-1}T_{dS}-dq\,q^{-1}), \tag{3.42}$$

off the singular set Θ. Comparing this with ω defined by (3.18) above, we find that

$$\omega-d\log\tau_q=\mathrm{Tr}(dq\,q^{-1}-T_{dS\,S^{-1}}). \tag{3.43}$$

The left-hand side is a closed form on $W\backslash\Theta$ and the right-hand side is holomorphic on W. Thus, the right hand side of (3.43) is a closed form on W. Since W is simply connected, it follows that there exists a holomorphic function ϕ on W so that

$$\omega - d\log\tau_q = d\phi. \tag{3.44}$$

Consulting the definition (3.41) for ω_{JMU} and the result (3.35) for ω, one finds

$$\omega_{\mathrm{JMU}} - \omega = -\sum_j \mathrm{Res}_j\,\mathrm{Tr}(\hat{\alpha}_j(\mathrm{loc})^{-1}\hat{\alpha}_j'(\mathrm{loc})dH_j). \tag{3.45}$$

The right-hand side of (3.45) is a holomorphic one-form on W and by (3.36) it is closed. Thus, there exists a holomorphic function φ on W so that

$$\omega_{\mathrm{JMU}} - \omega = d\varphi. \tag{3.46}$$

Adding (3.44) and (3.46) and writing $\Phi = \phi + \varphi$, we find

$$\omega_{\mathrm{JMU}} = d\log\tau_q + d\Phi = d\log(e^{\Phi}\tau_q). \tag{3.47}$$

Thus, for some constant, c, we have

$$\tau_{\mathrm{JMU}} = c e^{\Phi}\tau_q\,.$$

From this it follows immediately that $\tau_{\mathrm{JMU}}(t,\lambda)=0$ if and only if $(t,\lambda)\in\Theta$ and this finishes the proof of the theorem. QED

ACKNOWLEDGMENTS

The author would like to thank Doug Pickrell for helpful conversations. This research was partially funded by National Science Foundation Grant No. DMS 9401594.

[1] D. V. Anosov and A. A. Bolibruch, ''The Riemann–Hilbert problem,'' *Aspects of Mathematics:* E (Vieweg and Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1994), Vol. 22.
[2] L. Schlesinger, ''Über eine klasse von differentialsystem beliebiger ordnung mit festen kritischen punkten,'' J. Reine Angew. Math. **141**, 96–145 (1912).
[3] B. Malgrange, ''Sur les déformations isomonodromiques, I. Singularités régulières,'' in Ref. 13, pp. 401–426.
[4] G. Helmink, ''Deformations of connections, the Riemann–Hilbert problem and τ-functions,'' in *Computational and Combinatorial Methods in System Theory*, Elsevier Science Publishers B.V., edited by C. I. Byrnes and A. Lindquist (North Holland, New York, 1986), pp. 75–89.
[5] E. Faddell and L. Neuwirth, ''Configuration spaces,'' Math. Scand. **10**, 111–118 (1962).
[6] H. Grauert and R. Remmert, *Theory of Stein Spaces* (Springer-Verlag, Berlin, 1979).
[7] P. Deligne, ''Equations differéntielles à points singulieres réguliers,'' *Springer Lecture Notes* (Springer-Verlag, Berlin, 1970), Vol. 163.
[8] M. Jimbo, T. Miwa, and K. Ueno, ''Monodromy preserving deformations of linear ordinary differential equations with rational coefficients I, General theory and tau function,'' Physica D **2**, 306–352 (1981); M. Jimbo and T. Miwa, ''Monodromy preserving deformations of linear ordinary differential equations II,'' *ibid.* **2**, 407–448 (1981); M. Jimbo and T. Miwa, ''Monodromy preserving deformations of linear ordinary differential equations III,'' *ibid.* **4**, 26–46 (1983); T. Miwa, ''Painlevé property of monodromy preserving equations and the analyticity of τ-function,'' Publ. RIMS Kyoto Univ. **17-2**, 703–721 (1981).
[9] G. D. Birkhoff, ''The generalized Riemann problem for linear differential equations,'' Proc. Am. Acad. Arts Sci. **49**, 531–568 (1913).
[10] V. S. Varadarajan, ''Linear meromorphic differential equations: a modern point of view,'' Bull. Am. Math. Soc. **33-1**, 1–41 (1996).
[11] K. Ueno, ''Monodromy preserving deformations of linear differential equations with irregular singular points,'' RIMS preprint 301, 1979.
[12] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations* (Interscience, New York, 1965).
[13] B. Malgrange, ''La classification des connections irrégulières à une variable.'' in Mathématique et Physique, Progress in Mathematics, edited by L. Boutet de Monvel, A. Douady, and J-L. Verdier (Birkhäuser, Boston, 1983), Vol. 37, pp. 381–400.
[14] B. Malgrange, ''Sur les déformations isomonodromiques, II. Singularités irrégulières,'' in Ref. 13, pp. 427–438.

[15] H. Flaschka and A. Newell, ''Monodromy and spectrum preserving deformations I,'' Commun. Math. Phys. **76**, 65–116 (1980).
[16] Y. Sibuya, ''Stokes phenomena,'' Bull. Am. Math. Soc. **83-5**, 1075–1077 (1977).
[17] D. G. Babbitt and V. S. Varadarajan, ''Local moduli for meromorphic differential equations,'' Astérisque **169–170**, 1–217 (1989).
[18] H. Grauert, ''Analytische Faserungen über holomorph-vollständigen Räumen,'' Math. Annalen **135**, 263–273 (1958).

A note on the Morse index theorem for geodesics between submanifolds in semi-Riemannian geometry

Paolo Piccione[a] and Daniel V. Tausk[b]
Departamento de Matemática, Universidade de São Paulo, Brazil

(Received 26 April 1999; accepted for publication 24 August 1999)

The computation of the index of the Hessian of the action functional in semi-Riemannian geometry at geodesics with two variable endpoints is reduced to the case of a *fixed* final endpoint. Using this observation, we give an elementary proof of the Morse index theorem for Riemannian geodesics with two variable endpoints, in the spirit of the original Morse proof. This approach reduces substantially the effort required in the proofs of the theorem given previously [Ann. Math. **73**(1), 49–86 (1961); J. Diff. Geom **12**, 567–581 (1977); Trans. Am. Math. Soc. **308**(1), 341–348 (1988)]. Exactly the same argument works also in the case of timelike geodesics between two submanifolds of a Lorentzian manifold. For the extension to the lightlike Lorentzian case, just minor changes are required and one obtains easily a proof of the focal index theorem previously presented [J. Geom. Phys. **6**(4), 657–670 (1989)].
[S0022-2488(99)01012-9]

I. INTRODUCTION

A geodesic in a semi-Riemannian manifold $(\mathcal{M},g)$ is a smooth curve $\gamma:[a,b]\mapsto\mathcal{M}$ that is a stationary point for the action functional $f(z)=\frac{1}{2}\int_a^b g(\dot z,\dot z)\,dt$ defined in the set of paths z joining two given points of $\mathcal{M}$. If $(\mathcal{M},g)$ is Riemannian, i.e., if g is positive definite, given one such critical point γ, the celebrated Morse index theorem relates some analytical properties of the second variation of f at γ with some geometrical properties of γ. More precisely, the *index* of Hess_f at γ, that gives the number of *essentially different* directions in which γ can be deformed to obtain a shorter curve, equals the number of conjugate points along γ counted with multiplicity, excluding the endpoints $\gamma(a)$ and $\gamma(b)$.

The index theorem opened a very active field of research for both geometers and analysts, and the original result of Morse was successively extended in several directions. Beem and Ehrlich extended the results to the case of timelike Lorentzian geodesics (see Ref. 1) and to the lightlike Lorentzian case.[1,2] The case of a Riemannian geodesic with endpoints variable in two submanifolds of $\mathcal{M}$ has been treated by several authors, including Ambrose, Bolton and Kalish, (see Refs. 3–5, see also Ref. 6); Kalish's result (Ref. 5) has successively been extended in Ref. 7 to the case that each end manifold lies at a focal point of the other. Following the approach of Ref. 5, Ehrlich and Kim have then proven in Ref. 8 the Morse index theorem for lightlike geodesics with endpoints varying on two spacelike submanifolds of a Lorentzian manifold. The case of spacelike geodesics in semi-Riemannian manifolds was treated by Helfer in Ref. 9, where an extension of the index theorem was proven in terms of the *Maslov* index of a curve, and by the introduction of a notion of *signature* for conjugate points. Edwards extended in Ref. 10 the Morse Index Theorem to the case of formally self-adjoint linear systems of ordinary differential equations (ODEs), and Smale proved in Ref. 11 a general version of the Index Theorem for strongly elliptic operators on a Riemannian manifold.

The key point in the original Morse's proof of the theorem was the introduction of a function

[a]Electronic mail: piccione@ime.usp.br
[b]Electronic mail: tausk@ime.usp.br

$i:[a,b]\mapsto\mathbb{N}$ that gives the index of the form I_t, which is the Hessian Hess_f restricted to the geodesic $\gamma|_{[a,t]}$. Using a suitable subdivision of the interval $[a,b]$ and some geometrical arguments (see Refs. 12 and 13) Morse proved that i is nondecreasing and left continuous, with discontinuities precisely at the conjugate points, and that the jump of i at each discontinuity point t_0 is given by the value of the multiplicity of the conjugate point $\gamma(t_0)$.

When passing to the case of variable endpoints, i.e., when one admits variations with curves having endpoints varying on two fixed submanifolds P and Q of $\mathcal{M}$, in which case a stationary point of f is a geodesic γ that is orthogonal to P and Q at its endpoints, some obstructions to the use of the original argument of Morse arise, due mainly to the fact that the restricted index form I_t does not detect the influence of the final manifold Q.

Ambrose[3] gave a proof of the Index Theorem that uses the subdivision argument by introducing a family Q_t of *localized end manifolds* along γ, constructed with the help of the geodesic flow of the normal bundle of P around $(\gamma(a),\dot\gamma(a))$. This construction leads to technical difficulties (see also Ref. 6), due to the fact that the submanifold Q_t may lose dimension and differentiability. The proof of Bolton[4] also uses a subdivision argument, and it avoids the introduction of the manifolds Q_t, but it employs a restricted index function which is no longer nondecreasing.

The passage to a restricted index function is avoided in Kalish's proof of the index theorem in the variable endpoints case (see Ref. 5). In this article, it is given an explicit direct sum decomposition of the space $\mathcal{H}^{(P,Q)}=B\oplus B_+^c\oplus B_-^c$ of vector fields along γ which are everywhere orthogonal to γ and tangent to P and Q respectively at $\gamma(a)$ and $\gamma(b)$. The index theorem is deduced with a study of the sign of the index form in each of the three spaces; the definition of such decomposition is not very natural, and the remaining calculations are rather involved.

Ehrlich and Kim[8] have adapted Kalish's proof to the case of lightlike Lorentzian geodesics, where a suitable quotient space is used, in analogy with the null Morse index theorem of Refs. 1 and 2.

The aim of this paper is to show that the proof of the Morse index theorem for geodesics with two variable endpoints is a simple adaptation of the classical proof for the fixed endpoints case, in the spirit of the original proof of Morse, which is well understood. To this goal, the key observation is that the case of a geodesic with final point varying on a submanifold Q can be deduced immediately from the case of a fixed final endpoint (see Theorem II.6) by using a natural splitting of the space $\mathcal{H}^{(P,Q)}$. Moreover, we emphasize that the case of causal (nonspacelike) Lorentzian geodesics is essentially analogous to the Riemannian case.

We try to keep all the statements and proofs of the paper at the maximum level of generality; in particular, we present an approach that unifies the Riemannian and the causal Lorentzian case, obtaining a proof of all the results for Riemannian and causal Lorentzian geodesics at the same time. In Remark II.7, among other things we observe that, in the Lorentzian lightlike case, the use of the quotient bundle employed in Refs. 1, 2, and 8 is not really essential for the computation of the (nonaugmented) index, which allows to give an easier statement of the focal index theorem. The possibility of circumventing the quotient bundle construction had already been emphasized in Ref. 14, where the author's idea was to consider vector fields which are perpendicular to the given lightlike geodesic but never tangential to it.

It is also important to observe that the result of Theorem II.6 applies to a great number of situations in semi-Riemannian geometry where the Morse index theorem may *not* work, such as, for instance, in the case of spacelike geodesics in stationary Lorentzian manifolds (see Remark II.8).

II. THE INDEX THEOREM

Let $(\mathcal{M},g)$ be a semi-Riemannian manifold, $m=\dim(\mathcal{M})$, $P\subset\mathcal{M}$ be a smooth submanifold of $\mathcal{M}$, and $\gamma:[a,b]\mapsto\mathcal{M}$ be a nonconstant geodesic in $\mathcal{M}$, with $\gamma(a)\in P$ and $\dot\gamma(a)\in T_{\gamma(a)}P^\perp$. We will say that γ is spacelike, timelike, or lightlike according to $g(\dot\gamma,\dot\gamma)$ positive, negative, or zero, respectively; by *causal* we will mean either timelike or lightlike.

Let ∇ denote the Levi-Civita connection of g and let $R(X,Y)=\nabla_X\nabla_Y-\nabla_Y\nabla_X-\nabla_{[X,Y]}$ be the curvature tensor of ∇; moreover, for all $p \in P$ and all $n \in T_pP^\perp$, let $\mathcal{S}_n^P$ be the second fundamental form of P in the orthogonal direction *n*, which is the following symmetric bilinear form on T_pP:

$$\mathcal{S}_n^P(v_1, v_2) = g(n, \nabla_{v_1} V_2),$$

where V_2 is any extension of v_2 to a vector field tangent to P. Observe that we are *not* in principle making any nondegeneracy assumption on P, but if the metric is nondegenerate on T_pP, then we can also define a linear map $\mathcal{S}_n^P: T_pP \mapsto T_pP$ such that $g(\mathcal{S}_n^P(v_1), v_2) = \mathcal{S}_n^P(v_1, v_2)$.

Given a (piecewise) smooth vector field V along γ, we denote by V' the covariant derivative of V along γ.

If $(\mathcal{M}, g)$ is Lorentzian, i.e., if the index of g is 1, and γ is timelike, we have that $T_{\gamma(a)}P$ is spacelike, in the sense that the restriction of g to $T_{\gamma(a)}P$ is positive definite. More in general, the restriction of the metric g to the orthogonal space $\dot\gamma(t)^\perp$ is positive definite for all $t \in [a,b]$. If γ is lightlike, the restriction of the metric to the orthogonal space is just positive semi-definite [having a one-dimensional kernel spanned by $\dot\gamma(t)$]. We will assume *henceforth* that $\dot\gamma(a) \notin T_{\gamma(a)}P$, so that, again, $T_{\gamma(a)}P$ is spacelike.

Let $\tilde{\mathcal{H}}^P$ denote the vector space of all piecewise smooth vector fields V along γ such that $V(a) \in T_{\gamma(a)}P$ and let $\mathcal{H}^P$ be the subspace of $\tilde{\mathcal{H}}^P$ consisting of those V such that $g(V, \dot\gamma) \equiv 0$ and $V(b) = 0$. Moreover, let $I^P: \tilde{\mathcal{H}}^P \times \tilde{\mathcal{H}}^P \mapsto \mathbb{R}$ be the symmetric bilinear form given by

$$I^P(V,W) = \int_a^b [g(V',W') + g(R(\dot\gamma, V)\dot\gamma, W)]\,dt - \mathcal{S}^P_{\dot\gamma(a)}(V(a), W(a)). \tag{1}$$

Observe that if the submanifold P consists of just one point, the term involving its second fundamental form $\mathcal{S}^P_{\dot\gamma(a)}$ in (1) disappears. In this case we will write just I instead of I^P.

We will be concerned with the *index* of I^P in $\mathcal{H}^P$, defined as follows. If $\mathcal{K}$ is a vector subspace of $\tilde{\mathcal{H}}^P$, then the index $i(I^P, \mathcal{K})$ of I^P in $\mathcal{K}$ is the number

$$\mathrm{ind}\,(I^P, \mathcal{K}) = \sup\{\dim(\mathcal{V}) : \mathcal{V} \text{ subspace of } \mathcal{K} \ with \ I^P|_{\mathcal{V}} < 0\},$$

and we set

$$\mathrm{ind}\,(I^P) = \mathrm{ind}\,(I^P, \mathcal{H}^P). \tag{2}$$

The number $\mathrm{ind}\,(I^P)$ will be called the *Morse index* of γ.

A Jacobi field along γ is a smooth vector field J satisfying the linear equation $J'' - R(\dot\gamma, J)\dot\gamma = 0$. We say that J is a P-Jacobi field if it satisfies in addition

$$J(a) \in T_{\gamma(a)}P \quad \text{and} \quad g(J'(a), w) + \mathcal{S}^P_{\dot\gamma(a)}(J(a), w) = 0, \quad \text{for all } w \in T_{\gamma(a)}P. \tag{3}$$

If the metric is nondegenerate on $T_{\gamma(a)}P$ we can rewrite the second condition in (3) as

$$J'(a) + \mathcal{S}^P_{\dot\gamma(a)}(J(a)) \in T_{\gamma(a)}P^\perp.$$

In this case, a simple counting argument shows that the dimension of the vector space of P-Jacobi fields along γ is precisely equal to m and that the dimension of P-Jacobi fields satisfying $g(J, \dot\gamma) = 0$ is equal to $m-1$ [for P-Jacobi fields the condition $g(J, \dot\gamma) = 0$ is equivalent to $g(J'(a), \dot\gamma(a)) = 0$]. Observe that if P is a point, then a P-Jacobi field is simply a Jacobi field J along γ such that $J(a) = 0$.

Two points $q_0 = \gamma(t_0)$ and $q_1 = \gamma(t_1)$, t_0, $t_1 \in [a,b]$, are said to be *conjugate* along γ if there exists a non-null Jacobi field J along γ with $J(t_0) = 0$ and $J(t_1) = 0$. A point $q_0 = \gamma(t_0), t_0 \in [a,b]$ is said to be a *P-focal point* along γ if there exists a non-null P-Jacobi field J along γ such that $J(t_0)=0$; the *geometrical multiplicity* $\mu^P(t_0)$ of a P-focal point $\gamma(t_0)$ is the dimension of the vector space of all P-Jacobi fields along γ that vanish at t_0. If $\gamma(t_0)$ is not P-focal, we set $\mu^P(t_0)=0$.

It is well known that, if γ is either Riemannian or causal Lorentzian, then the set of P-focal points along γ is discrete, hence finite. (As proved in Ref. 9, along a spacelike Lorentzian geodesic, or more in general along a semi-Riemannian geodesic the conjugate points may accumulate.) Namely, if $J_1,\ldots,J_m$ is a linear basis for the space of P-Jacobi fields along γ and $E_1,\ldots,E_m$ is a parallelly transported orthogonal basis along γ, then the smooth function $r(t)=\det(g(J_i,E_j))$ has only simple zeroes on $[a,b]$, i.e., zeroes of finite multiplicity, exactly at those points $t_0\in[a,b]$ such that $\gamma(t_0)$ is a P-focal point along γ [see for instance (Ref. 15, Ex. 8, p. 299)]. Similarly, for all $q_0=\gamma(t_0)$, the set of points q_1 that are conjugate to q_0 along γ is finite.

We are interested in the kernel of the restriction of I^P to $\mathcal{H}^P$. To this aim, we introduce the spaces $\mathcal{N}$ and $\mathcal{J}_0$ as follows:

$$\mathcal{N}=\{f\dot\gamma : f:[a,b]\mapsto\mathbb{R} \quad \text{piecewise smooth}, \quad f(a)=f(b)=0\};$$
$$\mathcal{J}_0=\{P\text{-Jacobi fields } J \text{ along } \gamma : J(b)=0\}. \tag{4}$$

If γ is lightlike we have $\mathcal{N}\subset\mathcal{H}^P$, and in fact $\mathcal{N}$ is contained in the kernel of I^P in $\mathcal{H}^P$, as follows directly from (1). We now compute this kernel in the case of Riemannian or causal Lorentzian geodesics.

Lemma II.1: Let $(\mathcal{M},g)$ be either Riemannian or Lorentzian; in the latter case assume that γ is causal. The kernel of the restriction of the bilinear form I^P to $\mathcal{H}^P$ is equal to $\mathcal{J}_0$ if $(\mathcal{M},g)$ is Riemannian or if $(\mathcal{M},g)$ is Lorentzian and γ is timelike. If γ is lightlike, this kernel is equal to $\mathcal{J}_0\oplus\mathcal{N}$.

Proof: It follows by standard integration by parts of (1) and simple arguments involving the Fundamental lemma of calculus of variations. □

The proof of the index theorem for Riemannian or causal Lorentzian geodesics with initial endpoint varying on a submanifold and fixed endpoint is a simple adaptation of the classical Morse proof of the index theorem in the case of fixed endpoints (see, for instance, Refs. 12 and 13). For the reader's convenience, we outline briefly such adaptation.

We start with the following:

Lemma II.2: Let $J_1,J_2,\ldots,J_n$ be any family of P-Jacobi fields (not necessarily linearly independent) and $\phi_1,\ldots,\phi_n,\psi_1,\ldots,\psi_n$ be piecewise smooth real functions on $[a,b]$. Then,

$$I^P\left(\sum_{i=1}^n \phi_i\cdot J_i,\sum_{j=1}^n \psi_j\cdot J_j\right)=\int_a^b g\left(\sum_{i=1}^n \phi_i'\cdot J_i,\sum_{j=1}^n \psi_j'\cdot J_j\right)dt + g\left(\sum_{i=1}^n \phi_i(b)\cdot J_i'(b),\sum_{j=1}^n \psi_j(b)\cdot J_j(b)\right). \tag{5}$$

Proof: It is a simple computation that uses the Jacobi equation, formulas (3), (1), and the fact that, for P-Jacobi fields J_i and J_j, one has $g(J_i',J_j)=g(J_i,J_j')$. □

For Riemannian or causal Lorentzian geodesics, the above Lemma gives immediately the following Corollary:

Corollary II.3: Let $(\mathcal{M},g)$ be either Riemannian or Lorentzian; in the latter case assume that γ is causal. Suppose there are no P-focal points along γ. Let $V,J\in\tilde{\mathcal{H}}^P$ be vector fields orthogonal to γ, with J a P-Jacobi field and such that $V(b)=J(b)$. Then $I^P(V,V)\geqslant I^P(J,J)$. In the Riemannian and timelike Lorentzian case equality holds if and only if $V=J$, and in the lightlike Lorentzian case it holds if and only if $V-J\in\mathcal{N}$.

Proof: By standard arguments, the absence of P-focal points along γ implies that V can be written as a linear combination of P-Jacobi fields with piecewise smooth coefficients. The conclusion follows directly from Lemma II.2. □

We give the following definition:

Definition II.4: A partition $a=t_0<t_1<\cdots<t_N=b$ of $[a,b]$ is said to be *normal* if the following conditions are satisfied:

(a) for all $i\geqslant 1$ and all $t\in]t_i,t_{i+1}]$, the point $\gamma(t)$ is not conjugate to $\gamma(t_i)$ along γ;
(b) (b) for all $t\in]t_0,t_1]$, the point $\gamma(t)$ is not P-focal along γ.

If γ is either Riemannian or causal Lorentzian, since the set of P-focal points along γ is finite, it is easy to see that there exists $\delta>0$ such that every partition $t_0,\ldots,t_N$ of $[a,b]$ with $t_{i+1}-t_i\leqslant\delta$ for all i is normal. Namely, in order to (b) be satisfied, one can take δ to be the Lebesgue number of a covering of γ by *totally normal neighborhoods* (see Ref. 13).

Given a normal partition, we define the following two subspaces of $\mathcal{H}^P$:

$$\mathcal{H}_0^P=\{V\in\mathcal{H}^P:V(t_i)=0,\forall i\geqslant 1\};$$
$$\mathcal{H}_J^P=\{V\in\mathcal{H}^P:V|_{[t_i,t_{i+1}]}\text{ is Jacobi }\forall i\geqslant 1,\text{ and }V|_{[t_0,t_1]}\text{ is }P\text{-Jacobi}\}. \tag{6}$$

Observe that there exists an isomorphism:

$$\phi:\mathcal{H}_J^P\mapsto\bigoplus_{i=1}^{N-1}\dot\gamma(t_i)^\perp \tag{7}$$

given by setting $\phi(V)=(V(t_1),V(t_2),\ldots,V(t_{N-1}))$. By using the isomorphism (7) one proves that $\mathcal{H}_0^P\cap\mathcal{H}_J^P=\{0\}$ and that $\mathcal{H}_0^P+\mathcal{H}_J^P=\mathcal{H}^P$, hence we have

$$\mathcal{H}_0^P\oplus\mathcal{H}_J^P=\mathcal{H}^P. \tag{8}$$

Using standard integration by parts in (1) one sees that the above decomposition is I^P-orthogonal; moreover, it follows directly from Corollary II.3 that I^P is non-negative on $\mathcal{H}_0^P$. Thus, the index of I^P on $\mathcal{H}^P$ is finite, and it equals the index of I^P on $\mathcal{H}_J^P$.

For the proof of the Morse index theorem for Riemannian or causal Lorentzian geodesics with variable initial point we will follow the outline of the proof of Ref. 13, Theorem 2.2, §11:

Theorem II.5: *Let $(\mathcal{M},g)$ be either Riemannian or Lorentzian, P a smooth submanifold of $\mathcal{M}$, and $\gamma:[a,b]\mapsto\mathcal{M}$ a geodesic [causal, if (M,g) is Lorentzian] with $\gamma(a)\in P$ and $\dot\gamma(a)\in T_{\gamma(a)}P^\perp\setminus T_{\gamma(a)}P$. Then,* $\operatorname{ind}(I^P)=\Sigma_{t_0\in]a,b[}\mu^P(t_0)<+\infty$.

Proof: Let I_t be the index form of the restricted geodesic $\gamma|_{[a,t]}$ and denote by $i(t)$ its index. By classical arguments, one sees that $i(t)=0$ for t near a, and that $i:]a,b[\mapsto\mathbb{N}$ is non decreasing. Let $\bar t\in]a,b[$ be fixed; in order to study the behavior of i near $\bar t$ we consider a normal partition $t_0=a<t_1<\ldots<t_N=b$, with $t_j<\bar t<t_{j+1}$ for some $j\geqslant 1$ (or $t_j<\bar t\leqslant t_{j+1}$ if $\bar t=b$). Replacing N with $j+1$ in (7), for $t\in]t_j,t_{j+1}[$ we obtain an isomorphism ϕ_t onto the *fixed* space $\Theta=\oplus_{i=1}^j\dot\gamma(t_i)^\perp$; we regard I_t as a family of symmetric bilinear forms on Θ. As we have already observed, the index of I_t on Θ equals $i(t)$; we also observe that I_t depends continuously on t.

Since negativity is an open condition, it follows that $i(t)\geqslant i(\bar t)$ for t near $\bar t$; the monotonicity of i implies then that $i(t)=i(\bar t)$ for $t<\bar t$ near $\bar t$.

If γ is Riemannian or timelike Lorentzian, Lemma II.1 implies that $\dim(\operatorname{Ker}(I_{\bar t}))=\mu^P(\bar t)$. If γ is lightlike Lorentzian, again by Lemma II.1, we see that the subspace $\phi_t(\mathcal{N}\cap\mathcal{H}_J^P)$ of Θ is contained in the kernel of I_t; this space consists of the j-tuples of vectors parallel to $\dot\gamma$, and therefore it does not depend on $t\in]t_j,t_{j+1}[$.

Since positivity is an open condition, then $i(t)\leqslant i(\bar t)+\mu^P(\bar t)$ for $t>\bar t$ near $\bar t$; to conclude the proof we now show the opposite inequality. For, a standard argument that uses Corollary II.3 shows that

$$I_t(\theta,\theta)\leqslant I_{\bar t}(\theta,\theta),\quad \forall t\in]\bar t,t_{j+1}[,\quad \forall\theta\in\Theta,$$

where the inequality is strict if $V=\phi_t^{-1}(\theta)$ is a nonzero P-Jacobi field such that $V(\overline{t})=0$. This concludes the proof. □

We now want to extend the Morse index theorem to the case of two variable endpoints. To this end, we now assume that P and Q are smooth submanifolds of $\mathcal{M}$, and that $\gamma:[a,b]\mapsto\mathcal{M}$ is a geodesic with $\gamma(a)\in P$, $\dot\gamma(a)\in T_{\gamma(a)}P^\perp$, $\gamma(b)\in Q$ and $\dot\gamma(b)\in T_{\gamma(b)}Q^\perp$.

We denote by $\mathcal{H}^{(P,Q)}$ the vector space of all piecewise smooth vector fields V along γ, with $g(V,\dot\gamma)\equiv 0$, $V(a)\in T_{\gamma(a)}P$, and $V(b)\in T_{\gamma(b)}Q$. Moreover, we will consider the symmetric bilinear form $I^{(P,Q)}$ on $\mathcal{H}^{(P,Q)}$, given by

$$I^{(P,Q)}(V,W)=I^P(V,W)+\mathcal{S}^Q_{\dot\gamma(b)}(V(b),W(b)). \tag{9}$$

Let $\mathcal{J}^Q$ denote the subspace of $\mathcal{H}^{(P,Q)}$ consisting of P-Jacobi fields, and let $\mathcal{A}$ be the symmetric bilinear form on $\mathcal{J}^Q$ obtained by the restriction of $I^{(P,Q)}$. Then, it is easily computed from (1) using integration by parts:

$$\mathcal{A}(J_1,J_2)=\mathcal{S}^Q_{\dot\gamma(b)}(J_1(b),J_2(b))+g(J_1'(b),J_2(b)),\quad J_1,J_2\in\mathcal{J}^Q.$$

Moreover, for $t\in[a,b]$, we introduce the space $\mathcal{J}[t]$:

$$\mathcal{J}[t]=\{J(t):J\ \text{ is }\ P\text{-Jacobi}\}\subset T_{\gamma(t)}\mathcal{M};$$

observe that, for $t\in]a,b],\gamma(t)$ is not P-focal if and only if $\mathcal{J}[t]=T_{\gamma(t)}\mathcal{M}$.

We can now state and prove the following extension of the Morse Index Theorem for geodesics between submanifolds:

Theorem II.6: *Let* $(\mathcal{M}, g)$ *be a semi-Riemannian manifold, P, Q submanifolds of* $\mathcal{M}$ and $\gamma:[a,b]\mapsto\mathcal{M}$ *be a geodesic such that* $\gamma(a)\in P$, $\dot\gamma(a)\in T_{\gamma(a)}P^\perp$, $\gamma(b)\in Q$, *and* $\dot\gamma(b)\in T_{\gamma(b)}Q^\perp$. *Assume that* $\mathcal{J}[b]\subseteq T_{\gamma(b)}Q$. *Let* $\mathcal{V}$ *be a subspace of* $\mathcal{H}^{(P,Q)}$ *that contains the space* $\mathcal{J}^Q$ *of P-Jacobi fields along* γ *in* $\mathcal{H}^{(P,Q)}$. *Then,* $\mathrm{ind}(I^{(P,Q)},\mathcal{V})=\mathrm{ind}(I_P,\mathcal{H}_P\cap\mathcal{V})+\mathrm{ind}(\mathcal{A},\mathcal{J}^Q)$.

Proof: The space $\mathcal{H}^P$ is given by the subspace of $\mathcal{H}^{(P,Q)}$ consisting of those vector fields V such that $V(b)=0$; moreover, the restriction of $I^{(P,Q)}$ to $\mathcal{H}^P$ is precisely I^P. Defining $\mathcal{J}_0$ as in formula (4), let $\mathcal{J}_1$ be any subspace of $\mathcal{J}^Q$ such that $\mathcal{J}^Q=\mathcal{J}_0\oplus\mathcal{J}_1$. Clearly, $\mathcal{H}^{(P,Q)}=\mathcal{H}^P\oplus\mathcal{J}_1$, because $\mathcal{J}[b]\supset T_{\gamma(b)}Q$; moreover, from (9) it follows immediately that this decomposition is $I^{(P,Q)}$-orthogonal, i.e., $I^{(P,Q)}(V,J)=0$ for all $V\in\mathcal{H}^P$ and all $J\in\mathcal{J}_1$. Since $\mathcal{V}$ contains $\mathcal{J}_1$, then $\mathcal{V}=(\mathcal{V}\cap\mathcal{H}^P)\oplus\mathcal{J}_1$. Hence, $\mathrm{ind}(I^{(P,Q)},\mathcal{V})=\mathrm{ind}(I^P,\mathcal{H}^P\cap\mathcal{V})+\mathrm{ind}(\mathcal{A},\mathcal{J}_1)$. To conclude the proof, we simply observe that $\mathrm{ind}(\mathcal{A},\mathcal{J}_1)=\mathrm{ind}(\mathcal{A},\mathcal{J})$, because $\mathcal{J}_0\subset\mathrm{Ker}(\mathcal{A})$. □

Remark II.7: If $(\mathcal{M}, g)$ is Riemannian and $\mathcal{V}=\mathcal{H}^{(P,Q)}$, then Theorems II.5 and II.6 give as a particular case the Index Theorem of Ref. 5, p. 342, and the older versions of the Morse Index Theorem presented in Refs. 3 and 4. In Ref. 8 the fact was briefly mentioned that results analogous to the Riemannian case could apply to the Lorentzian timelike case. As to the lightlike case, in Refs. 1, 2, and 8 the authors consider the index of $I^{(P,Q)}$ in the quotient space $\mathcal{H}^{(P,Q)}/\mathcal{N}$ [recall formula (4)]; in this situation, $\mathcal{N}$ is contained in the kernel of $I^{(P,Q)}$. By a simple linear algebra argument one proves that the index of a bilinear form in a quotient space by a subspace of its kernel is the same as the index of the form in the original space. Hence, Theorems II.5 and II.6 generalize the results of Refs. 1, 2, and 8. In Ref. 14 (2.2 Null Index Lemma) the quotient bundle construction had been avoided by taking into consideration only those vector fields along a null geodesic which are orthogonal but never tangential to it.

Remark II.8: The result of Theorem II.6 becomes significant when the subspace $\mathcal{V}$ of $\mathcal{H}^{(P,Q)}$ is chosen in such a way that $\mathrm{ind}(I^P,\mathcal{H}^P\cap\mathcal{V})$ is finite; observe that ind $(\mathcal{A}, \mathcal{J}^Q)$ is always finite. If one considers geodesics in semi-Riemannian manifolds with metric of index greater or equal to 2, or spacelike geodesics in Lorentzian manifolds, then $\mathrm{ind}(I^P,\mathcal{H}^P)$ is in general infinite (see Refs. 9 and 16 for further results in this direction). Nevertheless, the restriction to suitable subspaces may yield the finiteness of the index, and, possibly, weaker versions of the Morse Index Theorem may apply. For instance (see Ref. 17), let us consider the case of a stationary Lorentzian manifold $(\mathcal{M}, g)$, i.e., a Lorentzian manifold endowed with a timelike Killing vector field Y. Let γ be a spacelike geodesic. The Killing vector field Y induces the conservation law $g(\dot\gamma,Y)\equiv C_\gamma$ for all geodesic γ; then, one can consider only variations γ_s of γ such that $g(\dot\gamma_s,Y)\equiv C_s$, and the corresponding variational field $V=(d/ds)|_{s=0}\gamma_s$ belongs to the space

$$\mathcal{V}=\{V:\exists C_V\in\mathbb{R}\quad\text{such that}\quad g(V',Y)-g(V,Y')\equiv C_V\}.$$

It is a simple observation that $\mathcal{V}$ contains all the Jacobi fields along γ; moreover, using the Sobolev Embedding Theorem one proves that the bilinear form I^P is given by a self-adjoint operator T on the closure of $\mathcal{V}\cap\mathcal{H}^P$ in a suitable Sobolev space completion of $\mathcal{H}^P$, where T is a *compact perturbation* of the identity (see Ref. 18). Thus, $\mathcal{V}\cap\mathcal{H}^{(P,Q)}$ satisfies the hypothesis of Theorem II.6 and it is such that $\mathrm{ind}(I^P,\mathcal{V}\cap\mathcal{H}^P)$ is finite. The value of this index has been recently proven to be equal to the sum of the signature of the restriction of g to $T_{\gamma(a)}P$ with the *Maslov* index of γ (see Ref. 19 for details).

ACKNOWLEDGMENTS

The first author is partially sponsored by CNPq (Brazil), Processo 301410/95-0 (RN). The second author is sponsored by FAPESP (São Paulo, Brazil), Processo 98/12530-2.

[1] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian Geometry* (Marcel Dekker, New York, 1996).
[2] J. Beem and P. Ehrlich, ''A Morse Index Theorem for Null Geodesics,'' Duke Math. J. **46**, 561–569 (1979).
[3] W. Ambrose, ''The Index Theorem in Riemannian Geometry,'' Ann. Math. **73**(1), 49–86 (1961).
[4] J. Bolton, ''The Morse Index Theorem in the case of Two Variable Endpoints,'' J. Diff. Geom. **12**, 567–581 (1977).
[5] D. Kalish, ''The Morse Index Theorem where the Ends are Submanifolds,'' Trans. Am. Math. Soc. **308**(1), 341–348 (1988).
[6] T. Takahashi, ''Correction to The Index Theorem In Riemannian Geometry by W. Ambrose'' Ann. Math. **80**, 538–541 (1964).
[7] N. Hingston and D. Kalish, ''The Morse Index Theorem in the Degenerate Endmanifold Case,'' Proc. Am. Math. Soc. **118**(2), 663–668 (1993).
[8] P. E. Ehrlich and S. Kim, ''A Focal Index Theorem for Null Geodesics,'' J. Geom. Phys. **6**(4), 657–670 (1989).
[9] A. D. Helfer, ''Conjugate Points on Spacelike Geodesics or Pseudo-Self-Adjoint Morse-Sturm-Liouville Systems,'' Pac. J. Math. **164**(2), 321–340 (1994).
[10] H. M. Edwards, ''A Generalized Sturm Theorem,'' Ann. Math. **80**, 22–57 (1964).
[11] S. Smale, ''On the Morse Index Theorem,'' J. Math. Mech. **14**, 1049–1056 (1965).
[12] J. Milnor, *Morse Theory* (Princeton U. P., Princeton, 1969).
[13] M. do Carmo, *Riemannian Geometry* (Birkhäuser, Boston, 1992).
[14] S. G. Harris, ''A Triangle Comparison Theorem for Lorentz Manifolds,'' Indiana Univ. Math. J. **31**(3), 289–308 (1982).
[15] B. O'Neill, *Semi-Riemannian Geometry with Applications to Relativity* (Academic, New York, 1983).
[16] V. Benci, F. Giannoni, and A. Masiello, ''Some Properties of the Spectral Flow in Semiriemannian Geometry,'' J. Geom. Phys. **27**, 267–280 (1998).
[17] F. Giannoni and P. Piccione, ''An Intrinsic Approach to the Geodesical Connectedness of Stationary Lorentzian Manifolds,'' Commun. An. Geom. **7**(1), 157–197 (1999).
[18] A. Masiello, *Variational Methods in Lorentzian Geometry*, Pitman Research Notes in Mathematics Vol. 309 (Longman, London 1994).
[19] F. Giannoni, A. Masiello, P. Piccione, and D. Tausk, ''A Generalized Index Theorem for Morse-Sturm Systems and Applications to semi-Riemannian Geometry,'' preprint 1999 (LANL MATH. DG/9908056).

Another proof of the triple sum formula for Wigner $9j$-symbols

Hjalmar Rosengren[a)]
Centre for Mathematical Sciences, Lund University, Box 118, S-221 00 Sweden

(Received 5 April 1999; accepted for publication 23 August 1999)

We give a simple proof of the Ališauskas–Jucys triple sum formula for Wigner $9j$-symbols. © *1999 American Institute of Physics.* [S0022-2488(99)00412-0]

In Ref. 1 we gave a new simple proof of the triple sum formula for Wigner $9j$-symbols due to Ališauskas and Jucys.[2] In this note we give an even more simple proof. We will use the notation of Ref. 1. In particular, we will work with the Lie group SU(1, 1) (and its covering groups), though the same proof works in the physically more important case of SU(2).

We will denote the (unnormalized) Racah coefficients or $6j$-symbols by

$$R_{qr}^{tu}(\nu_1,\nu_2,\nu_3)=\langle(((1,1)_q,1)_r,(1,(1,1)_t)_u\rangle_{\mathcal{A}^{\nu_1}\otimes\mathcal{A}^{\nu_2}\otimes\mathcal{A}^{\nu_3}}$$

(where $q+r=t+u$) and use notation such as

$$x_{12}=x_1+x_2,\quad |x|=x_1+\cdots+x_n$$

for a vector $x=(x_1,\dots,x_n)$. As in Ref. 1 we denote Wigner $9j$-symbols by

$$W_{pqr}^{tuv}=W_{pqr}^{tuv}(\nu_1,\nu_2,\nu_3,\nu_4),$$

where $p+q+r=t+u+v=s$. It is not hard to derive the expression

$$W_{pqr}^{tuv}=\sum_{j=0}^{\min(p+r,t+v)}\frac{(-1)^{p+r+j+u}}{C_{p+r-j}^{\nu_2,\nu_{34}+2q}C_j^{\nu_1,\nu_{234}+2s-2j}C_{t+v-j}^{\nu_3,\nu_{24}+2u}}R_{pr}^{p+r-j,j}(\nu_1,\nu_2,\nu_{34}+2q)$$
$$\times R_{q,p+r-j}^{u,t+v-j}(\nu_3,\nu_4,\nu_2)R_{t+v-j,j}^{t,v}(\nu_{24}+2u,\nu_3,\nu_1);\tag{1}$$

in fact this (or rather the corresponding formula for SU(2)) is often taken as the definition of Wigner $9j$-symbols. We will prove the triple sum formula by inserting appropriate single sum expressions for the Racah coefficients and changing the order of summation.

There is a large number of expressions for Racah coefficients as terminating ${}_4F_3$ series, including

$$R_{qr}^{tu}(\nu_1,\nu_2,\nu_3)=(\nu_1)_u(\nu_2)_t(\nu_3)_r(\nu_{12}+q-1)_q(\nu_{23}+r+t)_q(|\nu|+2q+r-1)_r$$
$$\times\frac{r!u!}{(u-q)!}\,{}_4F_3\left(\left.\begin{matrix}-q,1-\nu_1-q,1+r,\nu_3+r\\1+u-q,\nu_{23}+r+t,2-\nu_{12}-2q\end{matrix}\right|1\right)$$
$$=(\nu_1)_u(\nu_2)_q(\nu_2)_t(\nu_3)_s(\nu_{12}+s)_q(|\nu|+2q+r-1)_r$$
$$\times r!\,{}_4F_3\left(\left.\begin{matrix}-q,1-\nu_3-t,\nu_2+t,\nu_{12}+q-1\\\nu_2,1-\nu_3-s,\nu_{12}+s\end{matrix}\right|1\right)$$
$$=(\nu_1)_q(\nu_2)_t(\nu_3)_t(\nu_{12}+q-1)_u(\nu_{23}+2t)_u(|\nu|+2q+r-1)_r$$

[a)]Electronic mail: hjalmar@maths.lth.se

$$\times\frac{(-1)^{q+u}q!t!}{(q-u)!}\,{}_4F_3\left(\left.\begin{matrix}-u,1-\nu_1-u,1+t,\nu_3+t\\ 1+q-u,\nu_{23}+2t,2-\nu_{12}-q-u\end{matrix}\right|1\right),$$

where $q+r=t+u=s$. In the first and third expression we use the ${}_4F_3$ notation somewhat formally; for instance the first ${}_4F_3$ should be interpreted as in

$$\frac{1}{(u-q)!}\,{}_4F_3\left(\left.\begin{matrix}-q,a,b,c\\ 1+u-q,d,e\end{matrix}\right|1\right)=\sum_{k=\max(0,q-u)}^{q}\frac{(-q)_k(a)_k(b)_k(c)_k}{k!(u-q+k)!(d)_k(e)_k},$$

where one may have $u<q$. Inserting these three expressions, in this order, in (1) gives

$$\begin{aligned}W_{pqr}^{tuv}=&\sum_{j=0}^{\min(p+r,t+v)}\frac{(-1)^{q+v}r!t!(\nu_1)_t(\nu_2)_{s-j}(\nu_3)_t(\nu_4)_q(\nu_4)_u(\nu_{12}+p-1)_p(\nu_{13}+2t)_v}{(\nu_{34})_{s-j}(\nu_{234}+2s-2j)_j(\nu_{234}+2u+t+v-j-1)_{t+v-j}}\\&\times(\nu_{34})_{2q+r}(\nu_{234}+2q+2r+p-j)_p(\nu_{234}+2u+t+v-j-1)_v(|\nu|+2p+2q+r-1)_r\\&\times\frac{1}{(j-p)!}\,{}_4F_3\left(\left.\begin{matrix}-p,1-\nu_1-p,1+r,\nu_{34}+2q+r\\ 1+j-p,\nu_{234}+2q+2r+p-j,2-\nu_{12}-2p\end{matrix}\right|1\right)\\&\times{}_4F_3\left(\left.\begin{matrix}-q,1-\nu_2-u,\nu_4+u,\nu_{34}+q-1\\ \nu_4,1-\nu_2-s+j,\nu_{34}+s-j\end{matrix}\right|1\right)\\&\times\frac{1}{(t-j)!}\,{}_4F_3\left(\left.\begin{matrix}-v,1-\nu_{24}-2u-v,1+t,\nu_1+t\\ 1+t-j,\nu_{13}+2t,2-\nu_{234}-2u-t-2v+j\end{matrix}\right|1\right).\end{aligned}$$

By Karlsson's summation formula[3] the second ${}_4F_3$ vanishes if $p+r<j\leqslant t+v$; therefore the condition $j\leqslant p+r$ is superfluous. Let us now expand the ${}_4F_3$ functions and rearrange the terms according to

$$\sum_{j=0}^{t+v}\sum_{k=\max(0,p-j)}^{p}\sum_{l=0}^{q}\sum_{m=\max(0,j-t)}^{v}C_{jklm}=\sum_{\substack{0\leqslant k\leqslant p,0\leqslant l\leqslant q,\\ 0\leqslant m\leqslant v,k+m\geqslant p-t}}\sum_{j=0}^{t-p+k+m}C_{t+m-j,klm}.$$

This gives

$$\begin{aligned}W_{pqr}^{tuv}=&(-1)^{q+v}r!t!(\nu_1)_t(\nu_2)_{u+v}(\nu_3)_t(\nu_4)_q(\nu_4)_u(\nu_{12}+p-1)_p(\nu_{13}+2t)_v(\nu_{34}+q+r)_q\\&\times(|\nu|+2p+2q+r-1)_r\\&\times\sum_{\substack{0\leqslant k\leqslant p,0\leqslant l\leqslant q,\\ 0\leqslant m\leqslant v,k+m\geqslant p-t}}\frac{1}{(\nu_{34}+q+r)_{k+l}(1-\nu_2-u-v)_{l+m}(t-p+k+m)!}\\&\times\frac{(-p)_k(1-\nu_1-p)_k(1+r)_k(\nu_{34}+2q+r)_k}{k!(2-\nu_{12}-2p)_k}\,\frac{(-q)_l(1-\nu_2-u)_l(\nu_4+u)_l(\nu_{34}+q-1)_l}{l!(\nu_4)_l}\\&\times\frac{(-v)_m(1-\nu_{24}-2u-v)_m(1+t)_m(\nu_1+t)_m}{m!(\nu_{13}+2t)_m}\\&\times\frac{(\nu_{34}+u+v+l-m)_{t-p+k+m}}{(\nu_{234}+2s-2t-2m)_{t-p+k+m}}\sum_{j=0}^{t-p+k+m}\frac{\nu_{234}+2s-2t-2m-1+2j}{\nu_{234}+2s-2t-2m-1}\\&\times\frac{(p-t-k-m)_j(\nu_{234}+2s-2t-2m-1)_j(\nu_2+u+v-l-m)_j}{j!(\nu_{34}+u+v+l-m)_j(\nu_{234}+2s-p-t+k-m)_j}\,(-1)^j,\end{aligned}$$

where, by the summation formula[4]

$$\sum_{k=0}^{n} \frac{a+2k}{a} \frac{(-n)_k(a)_k(b)_k}{k!(1+a-b)_k(1+a+n)_k}(-1)^k = {}_4F_3\left(\begin{matrix} a,1+\frac{1}{2}a,b,-n \\ \frac{1}{2}a,1+a-b,1+a+n \end{matrix} \middle| -1 \right) = \frac{(1+a)_n}{(1+a-b)_n},$$

the inner sum equals

$$\frac{(\nu_{234}+2s-2t-2m)_{t-p+k+m}}{(\nu_{34}+u+v+l-m)_{t-p+k+m}}.$$

This leads to the expression

$$\begin{aligned} W_{pqr}^{tuv} &= (-1)^{q+v} r!t!(\nu_1)_t(\nu_2)_{u+v}(\nu_3)_t(\nu_4)_q(\nu_4)_u(\nu_{12}+p-1)_p(\nu_{13}+2t)_v(\nu_{34}+q+r)_q \\ &\quad \times(|\nu|+2p+2q+r-1)_r \\ &\quad \times \sum_{\substack{0\leqslant k\leqslant p, 0\leqslant l\leqslant q, \\ 0\leqslant m\leqslant v, k+m\geqslant p-t}} \frac{1}{(\nu_{34}+q+r)_{k+l}(1-\nu_2-u-v)_{l+m}(t-p+k+m)!} \\ &\quad \times \frac{(-p)_k(1-\nu_1-p)_k(1+r)_k(\nu_{34}+2q+r)_k}{k!(2-\nu_{12}-2p)_k} \frac{(-q)_l(1-\nu_2-u)_l(\nu_4+u)_l(\nu_{34}+q-1)_l}{l!(\nu_4)_l} \\ &\quad \times \frac{(-v)_m(1-\nu_{24}-2u-v)_m(1+t)_m(\nu_1+t)_m}{m!(\nu_{13}+2t)_m} \end{aligned} \tag{2}$$

which is a version of the Ališauskas–Jucys formula.

Finally we remark that Corollary 3.2 of Ref. 1 has previously been obtained (in the case of the group SU(2)) by Masao Nomura.[5] We owe this reference to Sigitas Ališauskas.

[1] H. Rosengren, ''On the triple sum formula for Wigner 9*j*-symbols,'' J. Math. Phys. **39**, 6730–6744 (1998).
[2] S. J. Ališauskas and A. P. Jucys, ''Weight lowering operators and the multiplicity-free isoscalar factors of the group R_5,'' J. Math. Phys. **12**, 594–605 (1971).
[3] P. W. Karlsson, ''Hypergeometric functions with integral parameter differences,'' J. Math. Phys. **12**, 270–271 (1971).
[4] L. J. Slater, *Generalized Hypergeometric Functions* (Cambridge University Press, London, 1966), Eq. (III.11).
[5] M. Nomura, ''Description of the 6-*j* and the 9-*j* symbols in terms of small numbers of 3-*j* symbols,'' J. Phys. Soc. Jpn. **58**, 2677–2686 (1989).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Invariance groups of transformations of basic hypergeometric series

J. Van der Jeugt[a)]
Department of Applied Mathematics and Computer Science, University of Ghent, Krijgslaan 281-S9, B-9000 Gent, Belgium

K. Srinivasa Rao
Institute of Mathematical Sciences, CIT Campus, Madras 600113, India

(Received 26 April 1999; accepted for publication 31 August 1999)

We show that certain two-term transformation formulas between basic hypergeometric series can easily be described by means of invariance groups. For the transformations of nonterminating ${}_3\phi_2$ series, and those of terminating balanced ${}_4\phi_3$ series, these invariance groups are symmetric groups. For transformations of ${}_2\phi_1$ series the invariance group is the dihedral group of order 12. Transformations of terminating ${}_3\phi_2$ series are described by means of some subgroup of S_6, and finally the invariance group of transformations of very-well-poised nonterminating ${}_8\phi_7$ series is shown to be isomorphic to the Weyl group of a root system of type D_5.
 [S0022-2488(99)00512-5]

I. INTRODUCTION

This article deals with two-term transformation formulas for basic hypergeometric series. Our aim is to identify the (finite) invariance group structures underlying known transformation formulas. Examining these invariance groups is not only an interesting problem by itself but it enables a whole list of (known) transformation formulas to be summarized as elegant one-line statements!

Anyone working with (basic) hypergeometric series is bound to encounter sooner or later transformation formulas which express a series into another. A summary of the most commonly used transformation formulas can be found in Appendix III of Gasper and Rahman.[1] A required, specific transformation formula, not listed in this appendix, can be derived by iterative use of one or more of the given transformation formulas. This can be done by hand, or even simpler by a computer algebra package such as HYP or HYPQ that performs these transformations consecutively.[2] This process of iteration is not new. Already in 1923 Whipple[3] showed that by iterating Thomae's ${}_3F_2$ transformation formula, one obtains a set of 120 such series, and he tabulated the parameters of these 120 series. He did not, however, recognize any group structure behind these 120 series. In some way, this group comes naturally into the picture, once a transformation formula between two (basic) hypergeometric series is translated into a linear transformation between the set of parameters (and/or variables) of the two series. The application of two transformation formulas then translates into the composition (or product) of two linear transformations, and this is how a group structure emerges, (see, for example, Refs. 4 and 5). This group is then called the invariance group of the transformations of (basic) hypergeometric series.

For the transformations of nonterminating ${}_3\phi_2$ series, and those of terminating balanced ${}_4\phi_3$ series, it is established in Sec. II that the invariance groups are the symmetric groups, S_5 and S_6, respectively. This result is not surprising, and essentially the same as in the classical case of the transformations of the (ordinary) hypergeometric series of unit argument, ${}_3F_2(1)$ and ${}_4F_3(1)$, respectively.[4] Our aim in proving these results is twofold: that the results can be established

[a)]Research Associate of the Fund for Scientific Research—Flanders (Belgium). Electronic mail: Joris.VanderJeugt @rug.ac.be simply, *à la* Hardy,[6] and, more importantly, they pave the way for a study of other cases. In Sec. III, Heine's transformations for ${}_2\phi_1$ series are considered, and it is shown that the invariance group is the dihedral group of order 12, D_{12}. Next, transformations of terminating ${}_3\phi_2$ series are considered. The invariance group is a nonsimple group of order 72, which can be described in terms of extended symmetries of the hexagon. Finally, in Sec. V it is established that the invariance group of transformations of very-well-poised nonterminating ${}_8\phi_7$ series is WD_5, a group of signed permutations isomorphic to the Weyl group of a root system of type D_5.

Apart from the mathematical interest, our motivation for studying these invariance groups also stems from a number of examples illustrating the usefulness of the transformations of the ordinary and basic hypergeometric series in physics and quantum groups. Louck *et al.* used the invariance group related to the ${}_4F_3(1)$ series to discuss the symmetries of extended 6-j coefficients.[6,7] Rajeswari and Srinivasa Rao used Thomae transformations of the ${}_3F_2(1)$ to derive the well-known Wigner, Racah, and Majumdar forms of the 3-j angular momentum coefficient[8] of quantum mechanics from the Van der Waerden form. They also derived the q analog of the Van der Waerden form of the 3-j coefficient and the q analogs of the aforesaid Wigner, Racah, and Majumdar forms of the 3-j coefficient.[9] They showed[10] that the transformation theory for basic hypergeometric functions, ${}_3\phi_2$, provides the necessary framework to relate the explicit forms for the Clebsch–Gordan coefficient of the quantum group $\mathrm{SU}_q(2)$.

We end this Introduction by fixing some notation. For (basic) hypergeometric series, our notation is the standard one of Ref. 1. Throughout the paper, the base q satisfies $|q|<1$. The q-shifted factorial is

$$(a;q)_n=\begin{cases}1, & n=0,\\ \prod_{k=0}^{n-1}(1-aq^k), & n=1,2,...,\infty,\end{cases}$$

and $(a_1,\ldots,a_m;q)_n=(a_1;q)_n\cdots(a_m;q)_n$. The basic hypergeometric series appearing here are

$${}_{r+1}\phi_r\left[\begin{matrix}a_1,a_2,\ldots,a_{r+1}\\ b_1,b_2,\ldots,b_r\end{matrix};q,z\right]=\sum_{k=0}^{\infty}\frac{(a_1,a_2,\ldots,a_{r+1};q)_k}{(q,b_1,b_2,\ldots,b_r;q)_k}z^k,$$

and for $r=1$ we shall usually write ${}_2\phi_1(a_1,a_2;b_1;q,z)$.

It is also useful to establish some notation about permutations here. A permutation on n elements can be denoted by its cycle structure (see, e.g., Ref. 11, Chap. 1). For example, the permutation $p\in S_5$ with $p(1)=2$, $p(2)=3$, $p(3)=1$, $p(4)=5$, $p(5)=4$ is simply denoted by (123) (45). For the action of p on an array of n symbols, we use by abuse of notation again p; thus for the above example,

$$p\cdot(x_1,x_2,x_3,x_4,x_5)=p\cdot x=(x_2,x_3,x_1,x_5,x_4).$$

With a permutation, an $n\times n$ monomial matrix (this is a matrix with exactly one nonzero element in every row and column) is associated with nonzero entries equal to 1. For the above example, this matrix is

$$\begin{pmatrix}0&1&0&0&0\\0&0&1&0&0\\1&0&0&0&0\\0&0&0&0&1\\0&0&0&1&0\end{pmatrix};$$

the action $p\cdot x$ can then be interpreted as a multiplication of this permutation matrix with a column vector x. We will also encounter ''signed permutations;'' here the nonzero entries of the monomial matrices are ± 1. If, for example, the matrix of a signed permutation p is given by

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

then we define its action on an array of n variables by

$$p\cdot(x_1,x_2,x_3,x_4,x_5)=p\cdot x=\left(x_2,\frac{1}{x_3},x_1,\frac{1}{x_5},x_4\right).$$

II. TRANSFORMATIONS OF NONTERMINATING ${}_3\phi_2$ SERIES, AND SEARS' TRANSFORMATIONS OF TERMINATING BALANCED ${}_4\phi_3$ SERIES

Thomae[12] obtained a number of transformations for hypergeometric series of type ${}_3F_2$ with unit argument (see also Ref. 13, Sec. 3.2) through the calculus of finite differences. Whipple[3] introduced a new notation in order to simplify the numerous formulas obtained by Thomae, and showed there are 120 such series related to each other. Still, it was not recognized that an invariance group (the symmetric group S_5) can describe all these relations. To our knowledge, the existence of such an invariance group was first stated by Hardy (Ref. 14, footnote on p. 111). Being unaware of Hardy's result, this property was later rediscovered by Beyer *et al.*.[4]

Here, the q analogs of Thomae's transformations are considered. These transformations were derived by Sears (Ref. 15, Sec. 10), and two of them are given by (III.9) and (III. 10) of Ref. 1. It is easy to show that for these transformations of nonterminating ${}_3\phi_2$ series, the invariance group is S_5.

Proposition 1: The function

$$f(x)=f(x_1,x_2,x_3,x_4,x_5)=\left(\frac{x_1x_2x_3}{x_4x_5},x_4^2,x_5^2;q\right)_\infty \times {}_3\phi_2\left[\begin{matrix}\frac{x_1}{x_2x_3}x_4x_5,\frac{x_2}{x_1x_3}x_4x_5,\frac{x_3}{x_1x_2}x_4x_5\\ x_4^2,x_5^2\end{matrix};q,\frac{x_1x_2x_3}{x_4x_5}\right] \tag{1}$$

is symmetric in the variables (x_1,x_2,x_3,x_4,x_5).

Proof: Clearly, $f(x)$ is invariant under permutations of (x_1,x_2,x_3) [resp. (x_4,x_5)], since these do not change the factor in front of the ${}_3\phi_2$, and on the ${}_3\phi_2$ they have the effect of permuting numerator (resp. denominator) parameters. Next, consider the permutation $p=(14325)$ (in cycle notation: $x_1\to x_4\to x_3\to x_2\to x_5\to x_1$), which is a permutation of order 5. Upon relabeling the parameters of the ${}_3\phi_2$ in $f(x)$ by

$${}_3\phi_2\left[\begin{matrix}a,b,c\\ d,e\end{matrix};q,\frac{de}{abc}\right],$$

it is easy to see that equation (III.10) of Ref. 1 is equivalent to $f(x)=f(p\cdot x)$. So $f(x)$ is invariant under the permutation p, and under the transposition $x_4\leftrightarrow x_5$. However, since p is a permutation of order 5, the group generated by p and this transposition is the complete group of permutations on five elements (see, for example, Ref. 16, p. 5), i.e., the symmetric group S_5. □

For convergence of all the 120 series in (1) it is sufficient that all x_i satisfy $\alpha^{3/2}<|x_i|<\alpha$, with $\alpha<1$ a positive real number.

Next, we consider transformations of terminating balanced ${}_4\phi_3$ series. This is the q-analog of Bailey's transformations for terminating Saalschützian hypergeometric series of type ${}_4F_3$ with unit argument [Ref. 13, Sec. 7.2(1)]. All 720 transformations of this type were already obtained by Bailey (Ref. 13, Chap. VII), and the fact that the invariance group for these transformations is the symmetric group S_6 was established by Beyer *et al.*[4] For the corresponding basic series, the essential transformation was first given by Sears [Ref. 15, (8.3)] [see also (III.15) and (III.16) of Ref. 1]. Here again, it is not difficult to show that also for these transformations of terminating balanced ${}_4\phi_3$ series, the invariance group is the symmetric group S_6.

Proposition 2: Let $x_1,\ldots,x_6$ be six parameters satisfying

$$x_1x_2x_3x_4x_5x_6=q^{1-n} \tag{2}$$

for some non-negative integer n. Then the function

$$f(x)=f(x_1,x_2,x_3,x_4,x_5,x_6)=q^{\binom{n}{2}}(x_1x_2x_3x_4,x_1x_2x_3x_5,x_1x_2x_3x_6;q)_n/(x_1x_2x_3)^n$$
$$\times{}_4\phi_3\left[\begin{matrix}q^{-n},x_2x_3,x_1x_3,x_1x_2\\ x_1x_2x_3x_4,x_1x_2x_3x_5,x_1x_2x_3x_6\end{matrix};q,q\right] \tag{3}$$

is symmetric in the variables $(x_1,x_2,x_3,x_4,x_5,x_6)$.

Proof: Again $f(x)$ is obviously invariant under permutations of (x_1,x_2,x_3) [resp. (x_4,x_5,x_6)]. Next, consider the cyclic permutation $p=(164253)$, which is a permutation of order 6. Upon relabeling the parameters of the ${}_4\phi_3$ in $f(x)$ by

$${}_4\phi_3\left[\begin{matrix}q^{-n},a,b,c\\ d,e,f\end{matrix};q,q\right],$$

it is easy to see that equation (III.16) of Ref. 1 is equivalent to $f(x)=f(p\cdot x)$. A similar argument as in the previous proof then implies the current proposition, i.e., the invariance group is S_6. □

Note that in this context equation (III.15) of Ref. 1 corresponds to $f(x)=f(p\cdot x)$ with $p=(14)(23)(56)$.

III. HEINE'S TRANSFORMATIONS OF ${}_2\phi_1$ SERIES

Heine[17] showed that

$${}_2\phi_1(a,b;c;q,z)=\frac{(a,bz;q)_\infty}{(c,z;q)_\infty}\,{}_2\phi_1(c/a,z;bz;q,a); \tag{4}$$

see [Ref. 1, (III.1)–(III.3)] for this and similar transformations. In (4), $|z|<1$ and $|a|<1$ is required for convergence. When iterating this transformation, using the symmetry of the numerator parameters of the ${}_2\phi_1$, a set of 12 transformation formulas is obtained. This observation is not new, and can already be found in Ref. 15, Sec. 11. However, the identification of the invariance group behind these 12 transformations has never been made. Here, we shall show that the invariance group is the dihedral group D_{12} (sometimes also denoted by D_6), i.e., the group of symmetries of the hexagon. In terms of a hexagon with sides labeled $x_1,\ldots,x_6$, the group consists of all rotations and reflections that map the hexagon into itself. The hexagon and two of its symmetries are shown in the following picture:

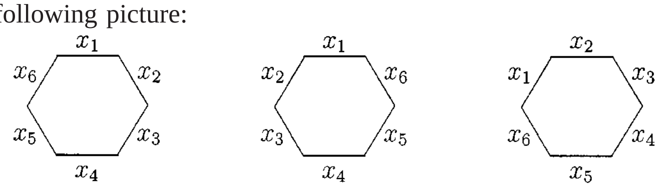

(5)

The second hexagon is obtained from the first one by a reflection about the vertical axis; the third one is obtained from the first one by a rotation through angle $\pi/3$. In all, there are 12 such symmetries. The group D_{12}, a subgroup of S_6, is generated by the two elements described above.

Proposition 3: The function

$$f(x)=f(x_1,x_2,x_3,x_4,x_5)=\left(x_1x_4,\frac{x_2x_6}{x_1};q\right)_\infty {}_2\phi_1\left(\frac{x_1x_3}{x_2},\frac{x_1x_5}{x_6};x_1x_4;q,\frac{x_2x_6}{x_1}\right) \tag{6}$$

is invariant under the dihedral group D_{12} acting on the variables $(x_1,x_2,x_3,x_4,x_5,x_6)$.

Proof: It is clear that the only effect of the permutation (26)(35) (in cycle notation), i.e., $x_2\leftrightarrow x_6$ and $x_3\leftrightarrow x_5$, on (6) is a transposition of the two numerator parameters in the ${}_2\phi_1$. Next, consider the permutation (123456), or explicitly

$$p\cdot(x_1,x_2,x_3,x_4,x_5,x_6)=(x_2,x_3,x_4,x_5,x_6,x_1). \tag{7}$$

Upon relabeling the parameters of the ${}_2\phi_1$ in $f(x)$ by ${}_2\phi_1(a,b;c;q,z)$, it is easy to see that (4) is equivalent to $f(x)=f(p\cdot x)$. The two permutations considered here correspond to the two symmetries of the hexagon described in (5), and they form a set of generators for the group D_{12}. □

For convergence of all the 12 series in (6) it is sufficient that all x_i satisfy $\alpha^2<|x_i|<\alpha$, with $\alpha<1$ a positive real number.

IV. TRANSFORMATIONS OF TERMINATING ${}_3\phi_2$ SERIES

Whipple[3] not only showed that there are 120 nonterminating ${}_3F_2$ series of unit argument, he also established that there are 72 terminating ${}_3F_2(1)$ series (see also Ref. 13, §3.9). In Ref. 5 the 72-element group associated with these series transformations was examined. The q-analogs of these transformation formulas were given by Sears.[15] The essential transformations are summarized in Ref. 1, (III.11)–(III.13), and others can be obtained from these by iteration. For example, from (III.11) and (III.13) one can deduce that

$${}_3\phi_2\left[\begin{matrix}q^{-n},b,c\\ d,e\end{matrix};q,q\right]=\frac{(c,de/bc;q)_n}{(d,e;q)_n}(b)^n{}_3\phi_2\left[\begin{matrix}q^{-n},d/c,e/c\\ de/bc,q^{1-n}/c\end{matrix};q,\frac{q}{b}\right]. \tag{8}$$

In the present case the group G associated with the transformations is somewhat more difficult to describe. By Cayley's and Lagrange's theorems,[18] it is obvious that the 72-element group G cannot be a subgroup of S_5 (since 72 is not a factor of 120), but it is a subgroup of S_6 generated by the two elements (24) and (123456). One way to see that the subgroup generated by these two elements has order 72 is by extending the 12 classical symmetries of the hexagon described in the previous section. The new transformations that are allowed on the hexagon are the interchange of two sides that are next to nearest neighbors, such as $x_1\leftrightarrow x_3$, $x_2\leftrightarrow x_4$, $x_3\leftrightarrow x_5$, etc. Superposed on the 12 classical symmetries, this enlarges the total number of allowed transformations to $12\times6=72$. The two generating elements of G, (24) and (123456), are described by the second and third hexagons in the following picture:

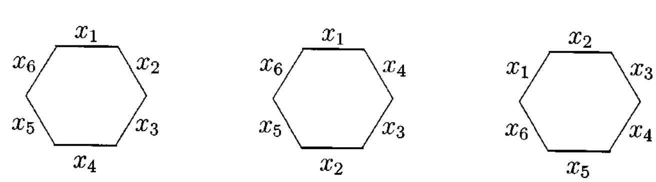

Another way to describe the elements of G is by calling x_2, x_4, and x_6 the even labels (or even sides), and x_1, x_3, and x_5 the odd labels (or odd sides) of the hexagon. Then the elements of G consist of the symmetries of the hexagon on which permutations of even labels (resp. odd labels) are superposed. As a consequence, G has a subgroup H consisting of permutations of (x_1,x_3,x_5) and of (x_2,x_4,x_6). Thus H is isomorphic to $S_3\times S_3$ and its order is 36.

The description of the invariance itself is also slightly more complicated in this case, mainly because the argument z of the ${}_3\phi_2$ is equal to q only in 36 cases, and different from q in the other 36 cases.

Proposition 4: Let $x_1,\dots,x_6$ be six parameters satisfying

$$x_1x_2x_3x_4x_5x_6=q^{1-n} \tag{10}$$

for some non-negative integer n. Consider the functions

$$\begin{aligned} f_0(x)&=f_0(x_1,x_2,x_3,x_4,x_5,x_6)\\ &=(x_1x_2x_3x_4,x_1x_2x_4x_5;q)_n/(x_1x_2x_4)^n\,{}_3\phi_2\left[\begin{matrix} q^{-n},x_1x_2,x_1x_4\\ x_1x_2x_3x_4,x_1x_2x_4x_5\end{matrix};q,q\right], \end{aligned} \tag{11}$$

$$\begin{aligned} f_1(x)&=f_1(x_1,x_2,x_3,x_4,x_5,x_6)\\ &=(-1)^nq^{\binom{n}{2}}(x_1x_2x_3x_4,x_1x_2x_4x_5;q)_n(x_6)^n\,{}_3\phi_2\left[\begin{matrix} q^{-n},x_1x_2,x_1x_4\\ x_1x_2x_3x_4,x_1x_2x_4x_5\end{matrix};q,\frac{q}{x_1x_6}\right]. \end{aligned} \tag{12}$$

Then $f_i(x)=f_i(p\cdot x)$ if $p\in H(i=0,1)$, and $f_0(x)=f_1(p\cdot x)$ (or $f_1(x)=f_0(p\cdot x)$) if $p\in G-H$.

Proof: The permutation (24) (in cycle notation) only transposes two of the numerator parameters in the ${}_3\phi_2$ of $f_0(x)$ or $f_1(x)$. Next, consider the permutation (123456), or explicitly

$$p\cdot(x_1,x_2,x_3,x_4,x_5,x_6)=(x_2,x_3,x_4,x_5,x_6,x_1). \tag{13}$$

This is an element of $G-H$, and it is straightforward to verify that (8) is equivalent to $f_0(x)=f_1(p\cdot x)$. The two permutations considered here generate G, hence G is the invariance group of the terminating ${}_3\phi_2$ series transformations. □

Consider the permutations (in cycle notation) $p_1=(13)(24)$ and $p_2=(1432)(56)$. Thus $p_1\in H$ and $p_2\in G-H$. As an extra verification, one can check that (III.11), (III.12), and (III.13) of Ref. 1 correspond to $f_0(x)=f_0(p_1\cdot x)$, $f_0(x)=f_1(p_2\cdot x)$, and $f_1(x)=f_0(p_2\cdot x)$, respectively.

V. TRANSFORMATIONS OF VERY-WELL-POISED ${}_8\phi_7$ SERIES

The last type of transformation considered here is for very-well-poised ${}_8\phi_7$ series. These transformations can be obtained as a limiting case of Bailey's terminating ${}_{10}\phi_9$ series transformations[19] [see also Ref. 1, (III.23) and (III.24)]. Such transformations are the q-analog of transformations of ''unrestricted well-poised ${}_7F_6$ series'' of unit argument (see Ref. 13, Sec. 7.5).

For our purposes, we define

$$\begin{aligned} &w(a;b,c,d,e,f)\\ &=\frac{(aq/b,aq/c,aq/d,aq/e,aq/f,a^2q^2/bcdef;q)_\infty}{(aq;q)_\infty}\\ &\times{}_8\phi_7\left[\begin{matrix} a, & q\sqrt{a}, & -q\sqrt{a}, & b, & c, & d, & e, & f\\ & \sqrt{a}, & -\sqrt{a}, & aq/b, & aq/c, & aq/d, & aq/e, & aq/f\end{matrix};q,\frac{a^2q^2}{bcdef}\right]. \end{aligned} \tag{14}$$

Clearly, w is invariant under permutations of (b,c,d,e,f), and hence S_5 is going to be a subgroup of the invariance group. In order to find the complete invariance group, we make again a relabelling as follows:

$$f(x_0,x_1,x_2,x_3,x_4,x_5)=w\left(x_0^3x_1x_2x_3x_4x_5/q;\frac{x_0x_2x_3x_4x_5}{x_1},\frac{x_0x_1x_3x_4x_5}{x_2},\right.$$
$$\left.\times\frac{x_0x_1x_2x_4x_5}{x_3},\frac{x_0x_1x_2x_3x_5}{x_4},\frac{x_0x_1x_2x_3x_4}{x_5}\right). \tag{15}$$

Thus permutations of (b,c,d,e,f) in w correspond to permutations of (x_1,x_2,x_3,x_4,x_5) in f. On examining the transformations given in Ref. 1, one finds that (III.23) is equivalent to

$$f(x_0,x_1,x_2,x_3,x_4,x_5)=f(x_0,x_1,x_2,x_3,1/x_5,1/x_4), \tag{16}$$

and that (III.24) is equivalent to

$$f(x_0,x_1,x_2,x_3,x_4,x_5)=f(x_0,1/x_2,1/x_3,1/x_4,1/x_5,x_1). \tag{17}$$

This implies that the invariance group will be a subgroup of the group WB_5 of signed permutations on five letters. In general, the group WB_n of signed permutations on n letters is the semidirect product of the symmetric group S_n operating as permutations on a standard basis e_i and $(\mathbf{Z}_2)^n$ operating as $e_i\rightarrow\pm e_i$, so its order is $n!2^n$. WB_n can also be seen as the group of $n\times n$ signed permutation matrices [cf. Sec. I, also for the action of such a signed permutation on the array (x_1,x_2,x_3,x_4,x_5)]. Since the invariance group is generated by the elements of S_5 and the elements corresponding to (16) and (17), we are dealing with a subgroup WD_5 of WB_5 consisting of those elements with an even number of (-1)'s in their matrix representation. Therefore the order of G is $5!2^4=1920$. We have used the notation WB_n and WD_n because WB_n is the Weyl group of a root system of type B_n (or C_n), and WD_n is the Weyl group of a root system of type D_n (see, e.g., Ref. 20).

Thus we have the following result:

Proposition 5: The function $f(x_0,x)=f(x_0,x_1,x_2,x_3,x_4,x_5)$ *satisfies* $f(x_0,x)=f(x_0,p\cdot x)$ *for every element p of* WB_5 *that has an even number of minus signs in its matrix representation. Hence the invariance group of the very-well-poised* ${}_8\phi_7$ *transformations is the group* WD_5.

For convergence of all the 1920 series it is sufficient that all $x_i (i=1,...,5)$ satisfy $1/\alpha<|x_i|<\alpha$ and $|x_0|<1/\alpha^5$ for some real number $\alpha>1$.

VI. SUMMARY AND REMARKS

The seminal remark of Hardy in his *Notes on Lecture VII*:[14]

''Formula (7.3.3) is equivalent to (1), §3.2, of Bailey's tract. It is an expression of the theorem that

$$\frac{1}{\Gamma(\beta_1)\Gamma(\beta_2)\Gamma(\beta_1+\beta_2-\alpha_1-\alpha_2-\alpha_3)}F\begin{pmatrix}\alpha_1,\alpha_2,\alpha_3\\ \beta_1,\beta_2\end{pmatrix}$$

is a symmetric function of the five arguments

$$\beta_1,\beta_2,\beta_1+\beta_2-\alpha_2-\alpha_3,\quad \beta_1+\beta_2-\alpha_3-\alpha_1,\quad \beta_1+\beta_2-\alpha_1-\alpha_2\text{''}$$

clearly implies a group theoretical interpretation for formula (7.3.3), which is a ${}_3F_2(1)$ transformation of Thomae:[12]

$$\frac{\Gamma(x+y+s+1)}{\Gamma(x+s+1)\Gamma(y+s+1)}F\begin{pmatrix}-a,-b,x+y+s+1\\ x+s+1,y+s+1\end{pmatrix}$$
$$=\frac{\Gamma(a+b+s+1)}{\Gamma(a+s+1)\Gamma(b+s+1)}F\begin{pmatrix}-x,-y,a+b+s+1\\ a+s+1,b+s+1\end{pmatrix}.$$

Forty-seven years later, Beyer *et al.*[4] rediscovered that Thomae's identity between two ${}_3F_2(1)$ hypergeometric series of unit argument, together with the trivial invariance under separate permutations of the numerator and denominator parameters, implies the symmetry group S_5 is an invariance group of this series. They also showed that Bailey's identity for Saalschützian ${}_4F_3(1)$ series has S_6 as its invariance group. Srinivasa Rao *et al.*[5] studied the group theory of terminating ${}_3F_2(1)$ series—a case not considered by Beyer *et al.*[4]—and found all the invariant subgroups of the 72-element group.

In this article, taking the cue from Hardy's remark, the group theoretical basis of well-known basic hypergeometric transformations has been studied, for the first time. We have constructed explicit functions $f(x)$, where x is a multi-dimensional parameter, expressed in terms of a basic hypergeometric series for the given transformation, and established the appropriate symmetry group. The transformations are then given by

$$f(x)=f(p\cdot x), \tag{18}$$

where p is any element of the symmetry group. The constructed functions (1), (3), and (6) together with (18) are succinct, quintessential one-line statements for Sears' nonterminating ${}_3\phi_2$, Sears' terminating balanced ${}_4\phi_3$, and Heine's ${}_2\phi_1$ transformations of basic hypergeometric series, respectively, which have as their corresponding invariance groups the symmetric groups S_5, S_6 and the dihedral group D_{12}.

In the case of the Sears' terminating ${}_3\phi_2$ transformations, the 72-element invariance group G (being a subgroup of S_6), a 36-element subgroup H of G, and two functions (instead of one) $f_0(x)$ and $f_1(x)$ were necessary to give a complete description. In terms of these all the transformations were generated by

$$f_i(x)=f_j(p\cdot x), \quad i,j=0,1,$$

where $i=j$ if $p\in H$ and $i\neq j$ if $p\in G-H$.

Finally, in the case of transformations of very-well-poised ${}_8\phi_7$ series, though the constructed function $f(x)$ is a function of six parameters, the invariance group is a subgroup WD_5 of WB_5, which is the Weyl group of a root system of type B_n (or, equivalently, the group of $n\times n$ signed permutation matrices).

ACKNOWLEDGMENTS

One of us (KSR) wishes to thank Professor G. Vanden Berghe and the Department of Applied Mathematics and Computer Science for excellent hospitality during the author's stay when most of this work was done.

[1] G. Gasper and M. Rahman, *Basic Hypergeometric Series* (Cambridge U.P., Cambridge, 1990).
[2] C. Krattenthaler, ''HYP and HYPQ: Mathematica packages for the manipulation of binomial sums and hypergeometric series, respectively q-binomial sums and basic hypergeometric series,'' J. Symbolic Comput. **20**, 737–744 (1995). Also available on http://radon.mat.univie.ac.at/People/kratt.
[3] F. J. W. Whipple, ''A group of generalized hypergeometric series: relations between 120 allied series of the type $F[a,b,c;d,e]$,'' Proc. London Math. Soc. (2) **23**, 104–114 (1925).
[4] W. A. Beyer, J. D. Louck, and P. R. Stein, ''Group theoretical basis of some identities for the generalized hypergeometric series,'' J. Math. Phys. **28**, 497–508 (1987).
[5] K. Srinivasa Rao, J. Van der Jeugt, J. Raynal, R. Jagannathan, and V. Rajeswari, ''Group theoretical basis for the terminating ${}_3F_2(1)$ series,'' J. Phys. A **25**, 861–876 (1992).
[6] J. D. Louck, W. A. Beyer, L. C. Biedenharn, and P. R. Stein, ''Symmetries of some hypergeometric series: implications for $3j$- and $6j$-coefficients,'' in *XV International Colloquium on Group Theoretical Methods in Physics*, edited by R. Gilmore (World Scientific, Singapore, 1987).
[7] For an overview of the relations between 3-j and 6-j coefficients and hypergeometric series, see L. C. Biedenharn and J. D. Louck, *The Racah-Wigner Algebra in Quantum Theory. Encyclopedia of Math. and its Applications, Vol. 9.* (Addison–Wesley, Reading, MA, 1981), Topic 11, p. 429.
[8] V. Rajeswari and K. Srinivasa Rao, ''Four sets of ${}_3F_2(1)$ functions, Hahn polynomials and recurrence relations for the 3-j coefficients,'' J. Phys. A **22**, 4113–4123 (1989).

[9] V. Rajeswari and K. Srinivasa Rao, ''Generalized basic hypergeometric functions and the q-analogues of 3-j and 6-j coefficients,'' J. Phys. A **24**, 3761–3780 (1991).
[10] V. Rajeswari and K. Srinivasa Rao, ''Note on the Explicit Forms for the Clebsch-Gordan Coefficient of the Quantum Group $SU_q(2)$,'' J. Phys. Soc. Jpn. **60**, 3583–3584 (1991).
[11] M. Hamermesh, *Group Theory, and its Application to Physical Problems* (Addison–Wesley, Reading, MA, 1962).
[12] J. Thomae, ''Über die Funktionen welche durch Reihen von der Form dargestellt werden: $1+pp'p''/1q'q''+\cdots$,'' J. Reine Angew. Math. **87** 26–73 (1879).
[13] W. N. Bailey, *Generalized Hypergeometric Series* (Cambridge U.P., Cambridge, 1935).
[14] G. H. Hardy, *Ramanujan. Twelve lectures on subjects suggested by his life and work* (Cambridge U.P., Cambridge, 1940).
[15] D. B. Sears, ''On the transformation theory of basic hypergeometric functions,'' Proc. London Math. Soc. (2) **53**, 158–180 (1951).
[16] G. James and A. Kerber, *The Representation Theory of the Symmetric Group, Encyclopedia of Mathematics and its Applications, Vol. 16* (Addison–Wesley, Reading, MA 1981).
[17] E. Heine, *Handbuch der Kugelfunctionen. Theorie und Anwendungen*, (Reimer, Berlin, 1878), Vol. 1.
[18] F. J. Budden, *The Fascination of Groups* (Cambridge U.P., Cambridge, 1972).
[19] W. N. Bailey, ''An identity involving Heine's basic hypergeometric series,'' J. London Math. Soc. **4**, 254–257 (1929).
[20] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory, Graduate Texts in Mathematics, Vol. 9* (Springer, Berlin, 1978).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Kählerian supermanifolds

S. Varsaie
Department of Mathematics, University of Tehran, Tehran, Iran

(Received 12 August 1998; accepted for publication 27 July 1999)

In this paper the concepts of almost Hermitian and Kählerian supermanifolds are introduced. Using the classical Newlander–Nirenberg theorm, it is shown that an almost complex structure J is integrable if it is compatible with Δ_L, the supermanifold analog of the Levi-Civita connection. It is proved that an almost Hermitian supermanifold is Kählerian if and only if Δ_L is compatible with J. [S0022-2488(99)00211-X]

I. INTRODUCTION

The concepts of differential geometry on a smooth manifold are based on the sheaf of germs of functions which is a sheaf of commutative algebras. Most of these concepts can be extended to a supermanifold which is a manifold and a sheaf of $\mathbb{Z}_2$-graded commutative algebras with certain local conditions. Although a unified theory of strong, weak, electromagnetic and gravitational intraction can be constructed in the language of supermanifolds (see Ref. 1, and references therein), from a purely mathematical point of view there exists certain open questions on the theory of supermanifolds. A main problem is that the natural definition of characteristic classes does not provide new invariants and this may be due to the lack of a suitable (co)homology theory which carries the new invariants (see the Introduction to Ref. 2). In commutative geometry, it is known that the space of harmonic forms represents cohomology groups. Therefore, it is natural to expect that an appropriate harmonicity for pseudodifferential forms can be related with a new cohomology theory.[3] As a first step in this direction, we introduce the concept of almost Hermitian supermanifolds. This leads to a new proof of integrability of an almost complex structure on almost Hermitian supermanifolds (see Theorem 4.4). Using this result a necessary and sufficient condition that an almost Hermitian supermanifold be Kählerian is obtained.

II. PRELIMINARIES

Let $\mathcal{A}$ be a sheaf of $\mathbb{Z}_2$-graded commutative (i.e., supercommutative) rings on a topological space M. Thus $\mathcal{A}=\mathcal{A}_0\oplus\mathcal{A}_1$ with $\mathcal{A}_i\mathcal{A}_j\subset\mathcal{A}_{(i+j)\text{mod }2}$, and on each open set U for $a\in\mathcal{A}_i(U)$, $b\in\mathcal{A}_j(U)$, we have $ab=(-1)^{ij}ba$. The index i is called the parity of a and is denoted by $\tilde{a}$. The pair $(M, \mathcal{A})$ is called a supercommutative ringed space.

A morphism of supercommutative ringed spaces $(M, \mathcal{A})$ to $(N, \mathcal{B})$ is a pair (τ, φ) of a continuous map $\tau:M\to N$ and a morphism $\varphi:\mathcal{B}\to\tau_*\mathcal{A}$ of sheaves of $\mathbb{Z}_2$-graded commutative rings on N.

A complex (real) supermanifold of dimension (m, n) is a supercommutative ringed space $(M, \mathcal{A})$ such that:

(i) $(M, \mathcal{A}/\mathcal{N})$ is an m-dimensional complex (respectively, real) manifold, where $\mathcal{N}$ is the sheaf of nilpotents of $\mathcal{A}$.
(ii) $\mathcal{E}=\mathcal{N}/\mathcal{N}^2$ is a locally free $\mathcal{A}/\mathcal{N}$-module of finite rank n.
(iii) $\mathcal{A}$ is locally isomorphic to the Grassman algebra $\wedge\mathcal{E}$.

Any open set U on which the ringed spaces in (iii) are isomorphic is said to be a splitting neighborhood. Let $x^1,\ldots,x^m$ be a system of coordinates of M on U and let $e^1,\ldots,e^n$ be a basis for the sections of $\mathcal{E}$ on U. The x^i's are called the even coordinates, and e^j's are called the odd coordinates. According to (i), $\mathcal{A}/\mathcal{N}$ will be denoted by $\mathcal{O}(M)$ in the complex case and by $C^\infty(M)$ in the real case in the sequel.

Let $\mathrm{Der}\,\mathcal{A}$ be the sheaf of derivations on $\mathcal{A}$. This is a $\mathbb{Z}_2$-graded $\mathcal{A}$-module and $\mathrm{Der}\,\mathcal{A} = (\mathrm{Der}\,\mathcal{A})_0 \oplus (\mathrm{Der}\,\mathcal{A})_1$. If (x^i, e^j) is a system of coordinates of $(M, \mathcal{A})$ on U, then $\mathrm{Der}\,\mathcal{A}$ is a free $\wedge\mathcal{E}$-module with basis $\{\partial/\partial x^i, \partial/\partial e^j\}$.

An almost complex structure on the real supermanifold $(M, \mathcal{A})$ is an even automorphism J on $\mathrm{Der}\,\mathcal{A}$ such that $J^2 = -1$. On the underlying real supermanifold of a complex supermanifold, there exists a natural almost complex structure (Ref. 4, step 1). Let (z^i, u^j) be a system of coordinates on the complex supermanifold $(M, \mathcal{A})$. With $z^i = x^i + \sqrt{-1}y^i$ and $u^j = v^j + \sqrt{-1}w^j$, let

$$J\left(\frac{\partial}{\partial x^i}\right) = \frac{\partial}{\partial y^i}, \quad J\left(\frac{\partial}{\partial y^i}\right) = -\frac{\partial}{\partial x^i},$$
$$J\left(\frac{\partial}{\partial v^j}\right) = \frac{\partial}{\partial w^j}, \quad J\left(\frac{\partial}{\partial w^j}\right) = -\frac{\partial}{\partial v^j}. \tag{1}$$

Obviously, J is an almost complex structure.

III. ALMOST HERMITIAN SUPERMANIFOLDS

In this section, the concept of a (Riemannian) supermetric is introduced. Due to the $\mathbb{Z}_2$-graded structure and hence the existence of zero divisors, in contrast with the commutative case, a supermetric cannot be commutative and positive definite. Instead, some other substitute conditions will be imposed.

Definition 3.1: By a supermetric on a real supermanifold $(M, \mathcal{A})$ we mean an $\mathcal{A}$-linear even map $g:\mathrm{Der}\,\mathcal{A}\otimes_{\mathcal{A}}\mathrm{Der}\,\mathcal{A}\to\mathcal{A}$ such that:

(i) $g(X\otimes Y) = (-1)^{\tilde{X}\tilde{Y}} g(Y\otimes X)$,
(ii) g induces a Riemannian metric on M, i.e., the map g_M with $g_M(\partial/\partial x^i, \partial/\partial x^k) = g(\partial/\partial x^i \otimes \partial/\partial x^k) \bmod \mathcal{N}$ is a Riemannian metric,
(iii) g induces a symplectic form on $\mathcal{E}$, i.e., the map $g_{\mathcal{E}}$ with $g_{\mathcal{E}}(e^j, e^l) = g(\partial/\partial e^j \otimes \partial/\partial e^l) \bmod \mathcal{N}$, is a symplectic form.

Definition 3.2: Let $(M, \mathcal{A})$ be a real supermanifold and let J be an almost complex structure on $(M, \mathcal{A})$. A supermetric g on $(M, \mathcal{A})$ is called a Hermitian supermetric if $g(JX\otimes JY) = g(X\otimes Y)$ for every $X, Y \in \mathrm{Der}\,\mathcal{A}$. Under these circumstances, $(M, \mathcal{A})$ is called an almost Hermitian supermanifold.

As in commutative case, any Hermitian supermetric g can be extended on $\mathrm{Der}\,\mathcal{A}\otimes\mathbb{C}$. Since J preserves g, on any splitting neighborhood U, g can be written as

$$g = dz^i d\overline{z}^k g_{ik} + dz^i d\overline{u}^j t_{ij} + du^j d\overline{z}^k s_{jk} + du^j d\overline{u}^l h_{jl}, \tag{2}$$

where $dz^i = dx^i + \sqrt{-1}J^*(dx^i)$, $du^j = de^j + \sqrt{-1}J^*(de^j)$ and $d\overline{z}^i$, $d\overline{u}^j$ are their conjugations, respectively, and (x^i, e^j) is a coordinate system on U and J^* is the automorphism induced by J on the dual of $\mathrm{Der}\,\mathcal{A}$.

Obviously every complex supermanifold admits a Hermitian supermetric. More precisely, by the existence of partition of unity for supermanifolds, it is sufficient to find a Hermitian supermetric locally. Let U be a spilitting neighborhood of a complex supermanifold $(M, \mathcal{A})$. If $z^i = x^i + \sqrt{-1}y^i$, $u^j = v^j + \sqrt{-1}w^j$ are coordinates on U then $J^*(dx^i) = dy^i$, $J^*(dv^j) = dw^j$, where J is the natural almost complex structure on the underlying real supermanifold of $(M, \mathcal{A})$ defined by (1). Thus $dz^i = dx^i + \sqrt{-1}dy^i$ and $du^j = dv^j + \sqrt{-1}dw^j$. On U we set

$$g_U = dz^i d\overline{z}^i + \sqrt{-1}du^j\, d\overline{u}^j,$$

then g_U gives a Hermitian supermetric.

Definition 3.3: Let g be a Hermitian supermetric. Set $\omega(X,Y)=g(X, JY)$ for all X, $Y \in \mathrm{Der}\,\mathcal{A}$. The supermetric g is called Kählerian if $d\omega=0$. By a Kählerian supermanifold, we mean a complex supermanifold equipped with a Kählerian supermetric.

On a splitting neighborhood U, g is given by (2), and hence on U,

$$\omega=\sqrt{-1}(dz^i\wedge d\overline{z^k}g_{ik}+dz^i\wedge d\overline{e^j}t_{ij}+de^j\wedge d\overline{z^k}s_{jk}+de^j\wedge d\overline{e^l}h_{jl}).$$

We now give an example of a Kählerian supermanifold.

Example 3.4: The complex projective superspace of dimension (m, n) is the supermanifold $\mathbb{P}_{m|n}=(\mathbb{P}_m,\wedge(\mathbb{C}^n\otimes_{\mathbb{C}}\mathcal{O}(-1)))$, where $\mathcal{O}(-1)$ is the sheaf of sections of the tautological line bundle over the complex projective space $\mathbb{P}_m$. Let $\mathbb{P}_m=\cup_{i=0}^m U_i$ where $U_i=\{[z^0,\ldots,z^m],z^i\neq 0\}$. If a system of coordinates of $\mathbb{P}_m$ on U_i is denoted by $z^k_{(i)}$ where $k\in\{0,1,...,m\}\backslash\{i\}$ and if $\{u_i^1,\ldots,u_i^n\}$ is a basis for sections of $\mathbb{C}^n\otimes_{\mathbb{C}}\mathcal{O}(-1)$ on U_i, then $(z^k_{(i)},u^j_{(i)})$ is a system of coordinates of $\mathbb{P}_{m|n}$ on U_i and the transition rules are $z^k_{(l)}=(z^l_{(i)})^{-1}z^k_{(i)}$ and $u^j_{(l)}=(z^l_{(i)})^{-1}u^j_{(i)}$ (Ref. 5, Chap. 4, Sec. 3).

In fact the Fubini–Study metric on $\mathbb{P}_m$ (see Ref. 6) can be naturally extended on $\mathbb{P}_{m|n}$. Let

$$f_i=\log\left(1+\sum_k z^k_{(i)}\overline{z^k_{(i)}}+\sqrt{-1}\sum_j u^j_{(i)}\overline{u^j_{(i)}}\right),$$

where $\log(1+t)=t-t^2/2+t^3/3-\cdots$. Since $u^l_{(i)}\cdot u^l_{(i)}=0$, it follows that

$$f_i=\log\left(1+\sum_k z^k_{(i)}\overline{z^k_{(i)}}\right)+\sqrt{-1}\left(\sum_j u^j_{(i)}\overline{u^j_{(i)}}\right)\left(\frac{1}{1+\Sigma_k z^k_{(i)}\overline{z^k_{(i)}}}\right). \tag{3}$$

Thus f_i is a well-defined section of $\mathbb{C}^n\otimes_{\mathbb{C}}\mathcal{O}(-1)$ on U_i and

$$f_i-f_k=\log|z^k_{(i)}|^2.$$

Therefore,

$$\partial\overline{\partial}f_i=\partial\overline{\partial}f_k. \tag{4}$$

Let $\omega=\partial\overline{\partial}f_i$. By (4) it follows that ω is a 1-1 form defined on $\mathbb{P}_{m|n}$. Since $d\omega=0$, the supermetric corresponding to ω is Kählerian.

Remark: Expression (3) first appeared in Ref. 7.

IV. CONNECTIONS

In this section the concept of a left connection on a supermanifold is given. It is shown that there exists a unique torsion free left connection on a supermanifold equipped with a supermetric. Then it is shown that an almost complex structure is integrable if it is parallel to Δ_L. Here Δ_L is the supermanifold analog of the Levi-Civita connection. Finally a necessary and sufficient condition that an almost Hermitian supermanifold is Kählerian will be presented.

Let $(M, \mathcal{A})$ be a real supermanifold and let $\mathcal{F}$ be a locally free sheaf of $\mathcal{A}$-modules over M. By $D_{\leqslant 1}$ we mean the $\mathcal{A}$-module of differential operators of rank $\leqslant 1$.

Definition 4.1: A left connection on $\mathcal{F}$ is a map $\Delta_L:D_{\leqslant 1}\otimes\mathcal{F}\to\mathcal{F}$ such that:

(i) $\Delta_L(a\otimes f)=af$,
(ii) $\Delta_L(Y\otimes af)=\Delta_L(Y\circ a\otimes f)$,
(iii) $\Delta_L(aY\otimes f)=a\Delta_L(Y\otimes f)$,

where $a\in\mathcal{A}(U)$, $Y\in\mathrm{Der}\,\mathcal{A}(U)$, and $f\in\mathcal{F}(U)$, and U is any open subset.

The natural left connection on $\mathcal{A}$ given by

$$Y\otimes a \mapsto Y(a)$$

will be denoted by d.

The torsion of a connection Δ_L is the map T:Der $\mathcal{A}\otimes$Der $\mathcal{A}\rightarrow$Der $\mathcal{A}$ defined by

$$T(X\otimes Y)=\Delta_L(X\otimes Y)-(-1)^{\tilde{X}\tilde{Y}}\Delta_L(Y\otimes X)-[X,Y].$$

Obviously, T is even and $\mathcal{A}$-linear.

Let g be a supermetric on $(M, \mathcal{A})$ and let Δ_L be a connection on Der $\mathcal{A}$ [in this case we say that Δ_L is a left connection on $(M, \mathcal{A})$]. Then Δ_L is said to be compatible with g if:

$$d(X\otimes g(Y\otimes Z))=g(\Delta_L(X\otimes Y)\otimes Z)+(-1)^{\tilde{X}\tilde{Y}}g(Y\otimes\Delta_L(X\otimes Z)). \tag{5}$$

Theorem 4.2: Let $(M, \mathcal{A})$ be a real supermanifold with supermetric g. There is a unique torsion free left connection on $(M, \mathcal{A})$ compatible with g.

Proof (The uniqueness): Let Δ_L be one such connection. By (5), for every $X,Y,Z\in$ Der $\mathcal{A}$ we have

$$g(\Delta_L(X\otimes Y)\otimes Z)=d(X\otimes g(Y\otimes Z))-(-1)^{\tilde{X}\tilde{Y}}g(Y\otimes\Delta_L(X\otimes Z)).$$

Since Δ_L is torsion free, it follows that

$$g(\Delta_L(X\otimes Y)\otimes Z)=d(X\otimes g(Y\otimes Z))-(-1)^{\tilde{X}\tilde{Y}}g(Y\otimes[X,Z])-(-1)^{\tilde{X}\tilde{Y}+\tilde{X}\tilde{Z}}g(Y\otimes\Delta_L(Z\otimes X)).$$

By (5) and similar substitution as above we get

$$\begin{aligned}g(\Delta_L(X\otimes Y)\otimes Z)=\tfrac{1}{2}[&d(X\otimes g(Y\otimes Z))-(-1)^{\tilde{X}\tilde{Z}+\tilde{Y}\tilde{Z}}d(Z\otimes g(X\otimes Y))\\&+(-1)^{\tilde{X}\tilde{Y}+\tilde{X}\tilde{Z}}d(Y\otimes g(Z\otimes X))-g(X\otimes[Y,Z])\\&+(-1)^{\tilde{X}\tilde{Z}+\tilde{Y}\tilde{Z}}g(Z\otimes[X,Y])+(-1)^{\tilde{X}\tilde{Y}+\tilde{X}\tilde{Z}}g(Y\otimes[Z,X])].\end{aligned} \tag{6}$$

The conditions (ii) and (iii) in Definition 3.1 imply that g is nondegenerate. Thus by (6), the uniqueness of Δ_L follows.

(*The existence*). Every $X\in$ Der $\mathcal{A}$ determines an element of Hom (Der $\mathcal{A}$, $\mathcal{A}$) which is denoted by the same X and is defined as

$$X(Y)=g(X\otimes Y).$$

This correspondence is 1-1 and onto. Since g is nondegenerate, (6) determines $\Delta_L(X\otimes Y)$ as a derivation on $\mathcal{A}$. With direct computation, it can be shown that $g(T(X,Y)\otimes Z)=0$ for every $Z\in$ Der $\mathcal{A}$. Again as g is nondegenerate, $T(X,Y)=0$. □

The connection Δ_L is the supermanifold analog of the Levi-Civita connection.

Definition 4.3: Let Δ_L be a left connection and let J be an almost complex structure on $(M, \mathcal{A})$. By compatibility of Δ_L and J we mean

$$\Delta_L(X\otimes JY)-J\Delta_L(X\otimes Y)=0$$

for every $X,Y\in$ Der $\mathcal{A}$.

By (6) it is obvious that the left connection corresponding to a Kählerian supermetric on a complex supermanifold is compatible with the natural almost complex structure. On the other hand, if the left connection corresponding to an almost Hermitian supermanifold given by Theorem 4.2 is compatible with the almost complex structure J, then the supermanifold has a complex structure, i.e., there exist local coordinates (x^i,e^j), $i=1,...,2m$; $j=1,...,2n$, such that $J(\partial/\partial x^i)=\partial/\partial x^{m+i}$, $J(\partial/\partial e^j)=\partial/\partial e^{n+j}$. In this case J is called integrable.

Theorem 4.4: Let $(M, \mathcal{A})$ be a real supermanifold with almost complex structure J and let g be a Hermitian supermetric. If Δ_L, the left connection corresponding to g, is compatible with J then J is integrable.

Proof: Let $\{U_\alpha\}$ be an open covering consisting of splitting neighborhoods on M and let $\varphi_\alpha:\mathcal{A}(U_\alpha)\to\wedge\mathcal{E}$ be the isomorphisms of $\mathbb{Z}_2$-graded algebras. If (x^i_α, e^j_α) is a system of coordinates on U_α, we set:

$$\Delta_M\frac{\partial}{\partial x^i_\alpha}:=\Delta_L\frac{\partial}{\partial x^i_\alpha}\bmod \mathcal{N}\left\langle\frac{\partial}{\partial x^k_\alpha}\right\rangle\oplus\left\langle\frac{\partial}{\partial e^j_\alpha}\right\rangle_{\mathcal{A}},$$

where $\langle\partial/\partial e^j_\alpha\rangle_{\mathcal{A}}$ means the $\mathcal{A}$-module generated by $\partial/\partial e^j_\alpha$. Obviously, Δ_M is compatible with the induced metric and the induced almost complex structure on M. Thus, by a well-known theorem on smooth manifolds,[8] the induced almost complex structure on M is integrable. Therefore, M is a complex manifold.

Similarly, it can be seen that P, the total space of the vector bundle $\mathcal{E}$, is also a complex manifold. Set

$$\Delta_P\frac{\partial}{\partial x^i_\alpha}=\Delta_L\frac{\partial}{\partial x^i_\alpha}\bmod \mathcal{N}\left\langle\frac{\partial}{\partial x^k_\alpha},\frac{\partial}{\partial e^l_\alpha}\right\rangle_{\substack{k=1,\dots,2m\\ l=1,\dots,2n}},$$

$$\Delta_P\frac{\partial}{\partial e^j_\alpha}=\Delta_L\frac{\partial}{\partial e^j_\alpha}\bmod \mathcal{N}\left\langle\frac{\partial}{\partial x^k_\alpha},\frac{\partial}{\partial e^l_\alpha}\right\rangle_{\substack{k=1,\dots,2m;\\ l=1,\dots,2n}}.$$

Then Δ_P is a well-defined connection on P compatible with the induced almost complex structure on P. Since Δ_P is torsion free, by Ref. 8, Corollary 3.5, P is a complex manifold. In particular, every point of P has a neighborhood on which there exist holomorphic coordinates z^t and u^s, $t=1,\dots,m;\ s=1,\dots,n$.

Let E_x be the fiber of the vector bundle $\mathcal{E}$ over $x\in M$. Identifying E_x with T_pE_x, where $\pi(p)=x$ and $\pi:P\to M$ is the projection map associated with $\mathcal{E}$, we can regard the derivation of the inclusion map $E_x\to P$ as a linear isomorphism $\rho_p:E_x\to V_p$ where V_p is the fiber of the vertical bundle over $p\in P$. Let κ be the vertical projection associated with the linear connection $\Delta_{\mathcal{E}}$ defined by

$$\Delta_{\mathcal{E}}e^j_\alpha=\Delta_L\frac{\partial}{\partial e^j_\alpha}\bmod \mathcal{A}\left\langle\frac{\partial}{\partial x^i_\alpha}\right\rangle_{i=1,\dots,2m}+\mathcal{N}\left\langle\frac{\partial}{\partial e^l_\alpha}\right\rangle_{l=1,\dots,2n}.$$

In order to obtain a section of $\mathcal{E}$, after applying mod in the above congruence, we need to substitute $\partial/\partial e^l_\alpha$ by e^l_α in the resulting expression. Observe that since the transition map on the intersection of any two splitting neighborhoods is a polynomial map in terms of anticommutative coordinates, the definition of $\Delta_{\mathcal{E}}$ is independent of the choice of sections e^j_α. Set:

$$v^s(x)=\rho_p^{-1}\circ\kappa\circ\frac{\partial}{\partial u^s}(o_x),\quad w^t(x)=\rho_p^{-1}\circ\kappa\circ\frac{\partial}{\partial z^t}(o_x),$$

where o_x is the null element of the vector space E_x. Obviously, $\{w^t, v^s\}$ is varied holomorphically, so that $\mathcal{E}$ is a holomorphic vector bundle.

Now consider the algebra of holomorphic sections of $\wedge\mathcal{E}(U_\alpha)\otimes\mathbb{C}$. If the inverse image of this algebra under the isomorphism $\varphi_\alpha\otimes i:\mathcal{A}(U_\alpha)\otimes\mathbb{C}\to\wedge\mathcal{E}(U_\alpha)\otimes\mathbb{C}$ is denoted by $\mathcal{B}(U_\alpha)$, then $(M, \mathcal{B})$ is a complex supermanifold and $(M, \mathcal{A})$ is the underlying real supermanifold. □

Remark: Theorem 4.4 can be obtained by invoking an extended Newlander–Nirenberg theorem proved by Mc Hugh.[9] Indeed, since Δ_L is torsion free and is compatible with g and J, straightforward computations reveal that

$$[JX,JY]=[X,Y]+J[JX,Y]+J[X,JY],$$

where $X,Y\in \mathrm{Der}\,\mathcal{A}$. Therefore by the extended Newlander–Nirenberg theorem for supermanifolds in Ref. 9, J is integrable. However, in our proof of Theorem 4.2 we have avoided the use of the extended Newlander–Nirenberg theorem.

Let $(M,\mathcal{A})$, g, J and Δ_L be as specified in Theorem 4.4 under the same hypotheses.

Theorem 4.5: $(M,\mathcal{A})$ is a Kählerian supermanifold.

Proof: It is sufficient to show that ω is closed, where ω is the associated 1-1 form to the Hermitian supermetric g. By Ref. 10, Theorem 4-3-7, $d\omega=0$ if and only if

$$(-1)^{\tilde{X}\tilde{Z}}d(X\otimes_{\omega}(Y,Z))+(-1)^{\tilde{X}\tilde{Y}}d(Y\otimes_{\omega}(Z,X))+(-1)^{\tilde{Y}\tilde{Z}}d(Z\otimes_{\omega}(X,Y))$$
$$=(-1)^{\tilde{X}\tilde{Z}}\omega([X,Y],Z)+(-1)^{\tilde{X}\tilde{Y}}\omega([Y,Z],X)+(-1)^{\tilde{Y}\tilde{Z}}\omega([Z,X],Y).$$

By the definition of ω and the properties of Δ_L, the validity of the last equation follows.

Theorem 4.6: Let $(M,\mathcal{A},g)$ be a Kählerian supermanifold as defined in Definition 3.3. Then Δ_L, the left connection associated to g, is compatible with J.

Proof: It is sufficient to show that

$$g(\Delta_L(X\otimes JY)-J\Delta_L(X\otimes Y),Z)=0 \tag{7}$$

for every $Z\in \mathrm{Der}\,\mathcal{A}$. By (6), adding and subtracting suitable terms, the left-hand side of the last equation is equal to

$$\tfrac{1}{2}[-d(X,\omega(Y,Z))+(-1)^{\tilde{X}\tilde{Y}}d(Y,\omega(X,Z))-(-1)^{\tilde{X}\tilde{Z}+\tilde{Y}\tilde{Z}}d(Z,\omega(X,Y))$$
$$+\omega([X,Y],Z)+(-1)^{\tilde{X}\tilde{Y}+\tilde{X}\tilde{Z}}\omega([Y,Z],X)-(-1)^{\tilde{Y}\tilde{Z}}\omega([X,Z],Y)]$$
$$+\tfrac{1}{2}[d(X,\omega(JY,JZ))-(-1)^{\tilde{X}\tilde{Y}}d(JY,\omega(X,JZ))+(-1)^{\tilde{X}\tilde{Z}+\tilde{Y}\tilde{Z}}d(JZ,\omega(X,JY))$$
$$-\omega([X,JY],JZ)-(-1)^{\tilde{X}\tilde{Y}+\tilde{X}\tilde{Z}}\omega([JY,JZ],X)+(-1)^{\tilde{Z}\tilde{Y}}\omega([X,JZ],JY)].$$

By Ref. 10, pp. 252 the first and the second brackets above are equal to $d\omega(X,Y,Z)$ and $-d\omega(X,JY,JZ)$, respectively. Thus (7) follows. □

ACKNOWLEDGMENT

The author would like to thank Professor Rahim Zaare-Nahandi for his encouragement and in particular, for his help on the final version of this paper.

[1] D. A. Leites, Russ. Math. Surveys **35**, 1–64 (1980).
[2] Yu. I. Manin and I. B. Penkov, *Lectures on Supermanifolds, Geometrical Methods of Conformal Groups* (World Scientific, Teaneck, 1989), pp. 3–13.
[3] F. A. Berezine, Sov. J. Nucl. Phys. **30**, 605–609 (1979).
[4] C. Haske and R. O. Wells, Jr., Duke Math. J. **54**, 493–500 (1987).
[5] Yu. I. Manin, *Gauge Field Theory and Complex Geometry* (Springer, Berlin, 1988).
[6] H. H. Wu, *The Equidimensional Theory of Holomorphic Curves*, Annals of Mathematical Studies, No. 64 (Princeton University Press, Princeton, NJ, 1970).
[7] M. M. Rezaii, Ph.D. thesis, Shahid Bahonar University, Kerman, Iran, 1991.
[8] S. Kobayashi and K. Numizu, *Foundation of Differential Geometry* (Wiley, New York, 1969), Vol. 2.
[9] A. Mc Hugh, J. Math. Phys. **30**, 1039–1042 (1989).
[10] B. Kostant, *Springer Lecture Notes in Mathematics* (Springer, New York, 1977), Vol. 510, pp. 177–306.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 40, NUMBER 12 DECEMBER 1999

Erratum: ‘‘Application of the discrete Wentzel–Kramers–Brillouin method to spin tunnelling’’ [J. Math. Phys. 39, 5166 (1998)]

Anupam Garg
Northwestern University, Department of Physics and Astronomy, Evanston, Illinois 60208

(Received 13 July 1999; accepted for publication 13 July 1999)

[S0022-2488(99)00212-1]

Dr. van Hemmen has pointed out to me that except for the minor extension to odd and even numbers of sites, the argument leading to Herring's formula, Eq. (3.11) of my paper, appears in the same form in Ref. 1. I regret the omission.

Secondly, condition (2.10) for the validity of the discrete WKB method is too weak, and should be replaced by

$$|dq(j)/dj| \leqslant \sin^2 q(j).$$

Lastly, the second of Eqs. (5.8) contains a sign error. The correct equation is

$$\cos \theta_1 = \cos \theta_0 + \sin^2 \theta_0 / 2S \cos \theta_0 + \cdots.$$

The conclusions are unaffected by either error.

[1] J. L. van Hemmen and A. Sütő, in *Quantum Tunneling of Magnetisation—QTM '94*, edited by L. Gunther and B. Barbara (Kluwer, Dordrecht, 1995).

